

IBM SPSS Direct Marketing 21



Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 109.

Cette version s'applique à IBM® SPSS® Statistics 21 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Matériel sous licence - Propriété d'IBM

© Copyright IBM Corporation 1989, 2012.

Droits limités pour les utilisateurs au sein d'administrations américaines : utilisation, copie ou divulgation soumise au GSA ADP Schedule Contract avec IBM Corp.

Préface

IBM® SPSS® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Marketing direct fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Marketing direct doit être utilisé avec le système central SPSS Statistics auquel il est entièrement intégré.

A propos de IBM Business Analytics

Le logiciel IBM Business Analytics offre des informations complètes, cohérentes et précises permettant aux preneurs de décision d'améliorer leurs performances professionnelles. Un portefeuille complet de solutions de [business intelligence](#), d'[analyses prédictives](#), de [performance financière et de gestion de la stratégie](#), et d'[applications analytiques](#) permet une connaissance claire et immédiate et offre des possibilités d'actions sur les performances actuelles et la capacité de prédire les résultats futurs. En combinant des solutions du secteur, des pratiques prouvées et des services professionnels, les entreprises de toute taille peuvent générer la plus grande productivité, automatiser les décisions en toute confiance et apporter de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les entreprises à prédire des événements futurs et à agir de manière proactive en fonction de ces prédictions pour apporter de meilleurs résultats. Des clients dans les domaines commerciaux, gouvernementaux et académiques se servent de la technologie IBM SPSS comme d'un avantage concurrentiel pour attirer ou retenir des clients, tout en réduisant les risques liés à l'incertitude et à la fraude. En intégrant le logiciel IBM SPSS à leurs opérations quotidiennes, les entreprises peuvent effectuer des prévisions, et sont capables de diriger et d'automatiser leurs décisions afin d'atteindre leurs objectifs commerciaux et d'obtenir des avantages concurrentiels mesurables. Pour plus d'informations ou pour contacter un représentant, visitez le site <http://www.ibm.com/spss>.

Support technique

Un support technique est disponible pour les clients du service de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, visitez le site IBM Corp. à l'adresse <http://www.ibm.com/support>. Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Support technique pour les étudiants

Si vous êtes un étudiant qui utilise la version pour étudiant, personnel de l'éducation ou diplômé d'un produit logiciel IBM SPSS, veuillez consulter les pages [Solutions pour l'éducation](#) (<http://www.ibm.com/spss/rd/students/>) consacrées aux étudiants. Si vous êtes un étudiant utilisant une copie du logiciel IBM SPSS fournie par votre université, veuillez contacter le coordinateur des produits IBM SPSS de votre université.

Service clients

Si vous avez des questions concernant votre livraison ou votre compte, contactez votre bureau local. Veuillez préparer et conserver votre numéro de série à portée de main pour l'identification.

Séminaires de formation

IBM Corp. propose des séminaires de formation, publics et sur site. Tous les séminaires font appel à des ateliers de travaux pratiques. Ces séminaires seront proposés régulièrement dans les grandes villes. Pour plus d'informations sur ces séminaires, accédez au site <http://www.ibm.com/software/analytics/spss/training>.

Contenu

Partie I: Guide de l'utilisateur

1	<i>Marketing direct</i>	1
2	<i>Analyse RFM</i>	2
	Scores RFM de Données de transactions	3
	Scores RFM de Données de clients.	4
	Regroupement par casiers RFM	6
	Enregistrement des scores RFM de Données de transactions	9
	Enregistrement des Scores RFM de Données de clients	11
	Résultats RFM	13
3	<i>Classification</i>	15
	Paramètres.	18
4	<i>Profils des clients potentiels</i>	20
	Paramètres.	25
	Création d'un champ de réponse qualitatif	25
5	<i>Taux de réponse des codes postaux</i>	27
	Paramètres.	31
	Création d'un champ de réponse qualitatif	33
6	<i>Propension à acheter</i>	34
	Paramètres.	39
	Création d'un champ de réponse qualitatif	41

7 Test de package de contrôle 42

Partie II: Exemples

8 Analyse RFM à partir des données de transaction 46

Données de transaction 46
Exécution de l'analyse 46
Evaluation des résultats 48
Fusionner les données de score avec les données Client 50

9 Classification 53

Exécution de l'analyse 53
Résultats 55
Sélection d'enregistrements en fonction des classes 63
 Création d'un filtre dans le Viewer de modèles de classe 64
 Sélection d'enregistrements en fonction des valeurs du champ de la classe 66
Récapitulatif 69

10 Profils des clients potentiels 70

Analyse des données 70
Exécution de l'analyse 70
Résultats 73
Récapitulatif 75

11 Taux de réponse des codes postaux 76

Analyse des données 76
Exécution de l'analyse 76
Résultats 79
Récapitulatif 82

12 Propension à acheter **83**

Analyse des données	83
Création d'un modèle prédictif	83
Evaluation du modèle.	87
Application du modèle	88
Récapitulatif	94

13 Test de package de contrôle **95**

Exécution de l'analyse	95
Résultats	97
Récapitulatif	97

Annexes

A Fichiers d'exemple **98**

B Remarques **109**

Index **112**

Partie I: Guide de l'utilisateur

Marketing direct

L'option de Marketing direct fournit un ensemble d'outils conçus pour améliorer les résultats des campagnes de marketing direct, en identifiant les caractéristiques démographiques, d'achat et autres qui définissent différents groupes de consommateurs et qui ciblent des groupes spécifiques afin d'optimiser les taux de réponse positive.

Analyse RFM. Cette technique permet d'identifier les clients existants qui ont plus de chances de répondre à une nouvelle offre. [Pour plus d'informations, reportez-vous à la section Analyse RFM dans le chapitre 2 sur p. 2.](#)

Classification C'est un outil d'exploration conçu pour révéler des groupements naturels (ou classes) au sein de vos données. Par exemple, elle permet d'identifier différents groupes de clients en fonction de diverses caractéristiques démographiques et d'achat. [Pour plus d'informations, reportez-vous à la section Classification dans le chapitre 3 sur p. 15.](#)

Profils des clients potentiels. Cette technique utilise les résultats d'une campagne précédente ou d'une campagne test pour créer des profils descriptifs. Vous pouvez utiliser ces profils pour cibler des groupes de contacts spécifiques lors des futures campagnes. [Pour plus d'informations, reportez-vous à la section Profils des clients potentiels dans le chapitre 4 sur p. 20.](#)

Taux de réponse des codes postaux. Cette technique utilise les résultats de la campagne précédente pour calculer les taux de réponse par code postal. Ces taux peuvent être utilisés pour cibler des codes postaux spécifiques lors des futures campagnes. [Pour plus d'informations, reportez-vous à la section Taux de réponse des codes postaux dans le chapitre 5 sur p. 27.](#)

Propension à acheter. Cette technique utilise les résultats d'un envoi d'essai ou d'une campagne précédente pour générer des scores de propension. Les scores indiquent quels contacts sont les plus susceptibles de répondre. [Pour plus d'informations, reportez-vous à la section Propension à acheter dans le chapitre 6 sur p. 34.](#)

Test de package de contrôle. Cette technique compare des campagnes de marketing pour vérifier s'il existe une différence significative de l'efficacité pour différents packages et offres. [Pour plus d'informations, reportez-vous à la section Test de package de contrôle dans le chapitre 7 sur p. 42.](#)

Analyse RFM

L'analyse RFM est une technique permettant d'identifier des clients existants qui ont plus de chances de répondre à une nouvelle offre. Cette technique est couramment utilisée dans le marketing direct. L'analyse RFM repose sur la théorie simple suivante :

- Le facteur le plus important dans l'identification des clients qui ont le plus de chances de répondre à une nouvelle offre est la **récence**. Les clients qui ont acheté plus récemment ont plus de chances d'acheter à nouveau que ceux qui ont acheté dans un passé plus lointain.
- Le deuxième facteur le plus important est la **fréquence**. Les clients qui ont acheté davantage dans le passé ont plus de chances de répondre que ceux qui ont moins acheté.
- Le troisième facteur le plus important est le montant total dépensé, appelé **monétaire**. Les clients qui ont dépensé davantage (au total pour tous les achats) dans le passé ont plus de chances de répondre que ceux qui ont dépensé moins.

Fonctionnement de l'analyse RFM

- Les clients obtiennent un score de récence selon la date de l'achat le plus récent ou l'intervalle de temps depuis l'achat le plus récent. Ce score repose sur un simple classement des valeurs de récence en un petit nombre de modalités. Par exemple, si vous utilisez cinq modalités, les clients avec les dates d'achat les plus récentes reçoivent un classement de récence de 5, tandis que ceux avec des dates plus éloignées dans le passé reçoivent un classement de récence de 1.
- De manière semblable, les clients reçoivent ensuite un classement de fréquence, avec des valeurs supérieures représentant une fréquence d'achat supérieure. Par exemple, dans un schéma de classement à cinq modalités, les clients qui achètent le plus souvent reçoivent un classement de fréquence de 5.
- Finalement, les clients sont classés par valeur monétaire, avec les valeurs monétaires les plus élevées ayant le classement le plus haut. Toujours avec le même exemple, les clients qui ont dépensé le plus reçoivent un classement monétaire de 5.

Le résultat est quatre scores pour chaque client : score de récence, de fréquence, monétaire et RFM combiné, qui représente les trois scores individuels concaténés en une seule valeur. Les « meilleurs » clients (ceux qui ont le plus de chances de répondre à une offre) sont ceux ayant les scores RFM combinés les plus élevés. Par exemple, dans un classement à cinq modalités, il existe un total de 125 scores RFM combinés possibles, et celui le plus élevé est 555.

Analyse des données

- Si les lignes de données représentent des transactions (chaque ligne représente une seule transaction, et il peut y avoir plusieurs transactions pour chaque client), utilisez RFM de Transactions. [Pour plus d'informations, reportez-vous à la section Scores RFM de Données de transactions sur p. 3.](#)
- Si les lignes de données représentent des clients avec des informations récapitulatives pour toutes les transactions (avec des colonnes qui contiennent des valeurs pour le montant total dépensé, le nombre total de transactions, et la date de transaction la plus récente), utilisez

RFM de Données clients. Pour plus d'informations, reportez-vous à la section Scores RFM de Données de clients sur p. 4.

Figure 2-1
Données de transactions et de clients

Les lignes représentent des transactions			
ID	Gender	Date	Amount
1	Male	9/25/2005	21
2	Male	1/15/2006	297
4	Male	2/5/2006	249
4	Male	5/7/2005	172
6	Male	4/16/2005	164
6	Male	4/17/2005	286
7	Femal		
9	Male		
9	Male		
9	Male		
10	Femal		
10	Femal		

Les lignes représentent des clients et les récapitulatifs des transactions associées				
ID	Gender	Most Recent	Total Amount	Number of Purchases
1	Male	9/25/2005	21	1
2	Male	1/15/2006	297	1
4	Male	2/5/2006	421	2
6	Male	4/16/2005	450	2

Scores RFM de Données de transactions

Analyse des données

L'ensemble de données contient les variables qui comportent les informations suivantes :

- Une variable ou une combinaison de variables qui identifient chaque observation (client).
- Une variable avec la date de chaque transaction.
- Une variable avec la valeur monétaire de chaque transaction.

Figure 2-2
Données de transactions RFM

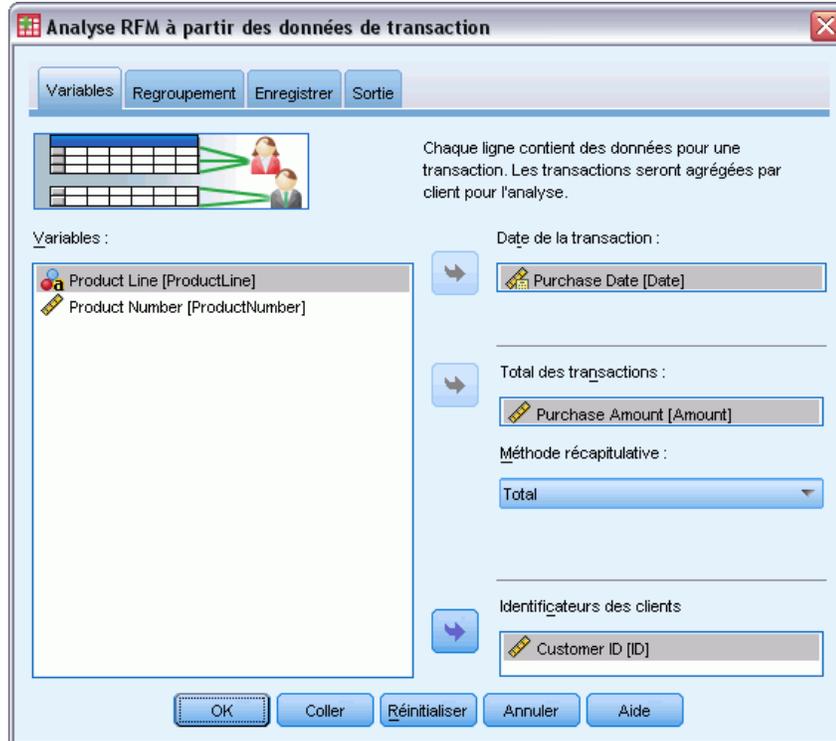
ID	Date	Amount
1	08/04/2005	129
1	10/25/2004	50
1	07/24/2004	118
1	07/24/2004	136
1	09/04/2006	52
2	09/23/2005	183
2	11/05/2004	24
2	11/10/2005	66
2	12/03/2004	77
3	06/04/2005	102
3	05/15/2005	131

Création de scores RFM de Données de transactions

- ▶ A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique
- ▶ Sélectionnez Aider à identifier mes meilleurs contacts (Analyse RFM) puis cliquez sur Poursuivre.

- Sélectionnez Données de transactions, puis cliquez sur Poursuivre.

Figure 2-3
Données de transactions, onglet Variables



- Sélectionnez la variable qui contient les dates de transactions.
- Sélectionnez la variable qui contient le montant monétaire de chaque transaction.
- Sélectionnez la méthode de récapitulatif des montants de transaction pour chaque client. Total (somme de toutes les transactions), moyenne, médiane ou maximum (montant de transaction le plus élevé).
- Sélectionnez la variable ou combinaison de variables qui identifient de façon unique chaque client. Par exemple, les observations peuvent être identifiées par un code d'identification unique ou une combinaison du nom et du prénom.

Scores RFM de Données de clients

Analyse des données

L'ensemble de données contient les variables qui comportent les informations suivantes :

- Date d'achat la plus récente ou intervalle de temps depuis la date d'achat la plus récente. Vous l'utilisez pour calculer les scores de récence.

- Nombre total d'achats. Vous l'utilisez pour calculer les scores de fréquence.
- Valeur monétaire récapitulative pour tous les achats. Vous l'utilisez pour calculer les scores monétaires. En général, il s'agit de la somme (du total) de tous les achats, mais peut être la moyenne, le maximum (montant le plus grand) ou autre mesure récapitulative.

Figure 2-4

Données de clients RFM

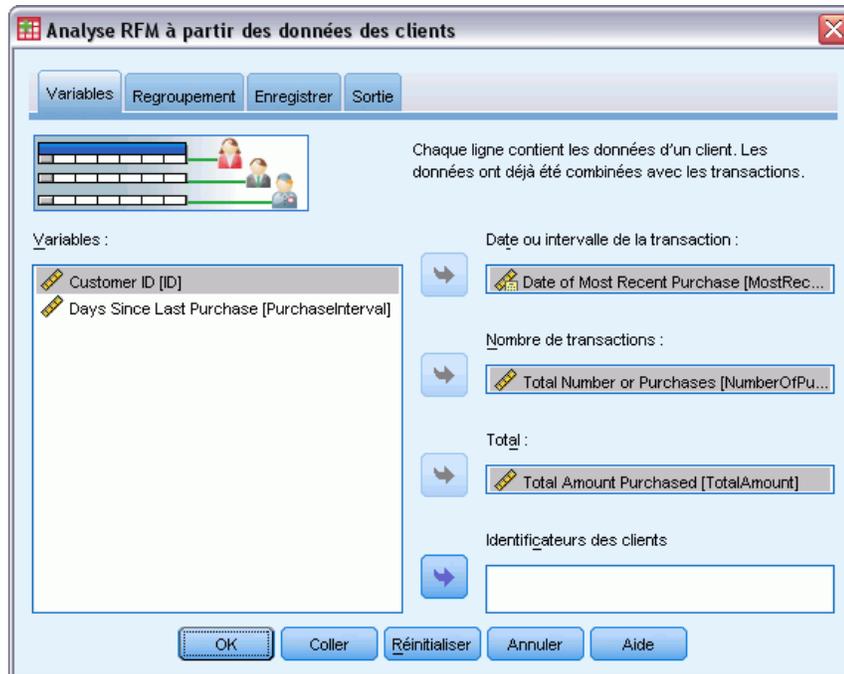
ID	TotalAmount	MostRecent	NumberOfPurchases
1	485.00	09/04/2006	5
2	350.00	11/10/2005	4
3	233.00	06/04/2005	2
4	936.00	08/18/2006	7
5	359.00	07/07/2006	3
6	249.00	07/16/2006	3
7	1089.00	02/15/2006	7
8	423.00	08/21/2006	4
9	689.00	08/31/2006	7
10	325.00	10/13/2005	3

Si vous souhaitez écrire des scores RFM dans un nouvel ensemble de données, l'ensemble de données actif doit aussi contenir une variable ou une combinaison de variables qui identifient chaque observation (client).

Création de scores RFM de Données de clients

- ▶ A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique
- ▶ Sélectionnez Aider à identifier mes meilleurs contacts (Analyse RFM) puis cliquez sur Poursuivre.
- ▶ Sélectionnez Données clients, puis cliquez sur Poursuivre.

Figure 2-5
Données clients , onglet Variables



- ▶ Sélectionnez la variable qui contient la date de transaction la plus récente ou un nombre qui représente un intervalle de temps depuis la transaction la plus récente.
- ▶ Sélectionnez la variable qui contient le nombre total de transactions pour chaque client.
- ▶ Sélectionnez la variable qui contient le montant monétaire récapitulatif de chaque client.
- ▶ Si vous souhaitez écrire des scores RFM dans un nouvel ensemble de données, sélectionnez la variable ou la combinaison de variables qui identifient de façon unique chaque client. Par exemple, les observations peuvent être identifiées par un code d'identification unique ou une combinaison du nom et du prénom.

Regroupement par casiers RFM

Le processus de regroupement d'un grand nombre de valeurs numériques en un petit nombre de modalités est parfois appelé **regroupement par casiers**. Dans l'analyse RFM, les casiers sont les modalités classées. L'onglet Regroupement par casiers permet de modifier la méthode utilisée pour affecter les valeurs de récence, de fréquence et monétaires à ces casiers.

Figure 2-6
Onglet Regroupement par casiers RFM

Analyse RFM à partir des données des clients

Variables Regroupement Enregistrer Sortie

Méthode de regroupement

Imbriquée
Les clients sont d'abord regroupés par récence. Les clients de chaque intervalle de récence sont ensuite regroupés par fréquence. Puis les clients de chaque intervalle de récence-fréquence sont ensuite regroupés par valeur monétaire.

Indépendante
Les clients sont regroupés indépendamment par récence, fréquence et valeur monétaire.

Nombre d'intervalles

Récence : 5
Fréquence : 5
Monétaire : 5

Ex-aequo

Affecter les ex-aequo au même intervalle
 Affecter les ex-aequo au hasard dans les intervalles adjacents

OK Coller Réinitialiser Annuler Aide

Méthode de regroupement

Imbriquée. Dans le regroupement par casier imbriqué, un rang simple est affecté à des valeurs de récence. Dans chaque rang de récence, les clients reçoivent ensuite un rang de fréquence, et dans chacun, les clients reçoivent un rang monétaire. Ceci tend à fournir une distribution plus équitable des scores RFM combinés, mais a pour inconvénient de compliquer l'interprétation des scores de rang de fréquence et monétaire. Par exemple, un rang de fréquence de 5 pour un client avec un rang de récence de 5 risque d'avoir une signification autre qu'un rang de fréquence de 5 pour un client avec un rang de récence de 4, puisque le rang de fréquence dépend de celui de récence.

Indépendant : Les rangs simples sont affectés aux valeurs de récence, de fréquence et monétaires. Ces trois rangs sont affectés de manière indépendante. L'interprétation de chacun des trois composants RFM est donc claire. Un score de fréquence de 5 pour un client a la même signification qu'un score de fréquence de 5 pour un autre client, quels que soient les scores de récence. Pour des échantillons plus petits, ceci a l'inconvénient d'entraîner une distribution moins équitable des scores RFM combinés.

Nombre d'intervalles

Le nombre de modalités (casiers) à utiliser pour chaque composant pour créer des scores RFM. Le nombre total de scores RFM combinés possibles est le produit des trois scores. Par exemple, 5 casiers de récence, 4 casiers de fréquence et 3 casiers monétaires créent un total de 60 scores RFM combinés possibles, allant de 111 à 543.

- La valeur par défaut est 5 pour chaque composant ; ce qui crée 125 scores RFM combinés possibles, allant de 111 à 555.
- Le nombre maximum de casiers autorisés pour chaque composant de score est neuf.

Ex-æquo

Un « ex æquo » est simplement deux ou plusieurs valeurs de récence, de fréquence ou monétaires égales. Dans l'idéal, vous souhaitez disposer d'environ le même nombre de clients dans chaque casier, mais un grand nombre de valeurs ex æquo peut affecter la distribution des casiers. Il existe deux moyens de gérer les ex æquo :

- **Affecter des ex æquo au même casier.** Cette méthode permet de toujours assigner des valeurs ex æquo au même casier, que ceci affecte ou non la distribution des casiers. Ceci offre une méthode de regroupement par casiers cohérente : Si la même valeur de récence est attribuée à deux clients, ils obtiendront toujours le même score de récence. Dans un cas extrême, toutefois, il se peut que 500 clients sur un total de 1000 fassent leurs achats les plus récents à la même date. Dans un classement à 5 casiers, 50 % des clients reçoivent ainsi un score de récence de 5, au lieu des 20 % souhaités.

Notez que la « cohérence » de la méthode imbriquée de regroupement par casiers est un peu plus complexe pour les scores monétaires et de fréquence, puisque ces derniers sont affectés dans les casiers de score de récence, et les scores monétaires sont affectés dans les casiers de score de fréquence. Par conséquent, deux clients ayant la même valeur de fréquence risquent de ne pas avoir le score de fréquence identique s'ils ne possèdent pas aussi le même score de récence, quel que soit le traitement des valeurs ex æquo.

- **Affectation aléatoire des ex æquo.** Ceci garantit une distribution équitable des casiers en affectant un facteur de variance aléatoire très faible avant tout classement. Ainsi, il n'existe aucune valeur ex æquo afin d'affecter les valeurs aux casiers classés. Ce processus n'a aucune incidence sur les valeurs d'origine. Il n'est utilisé que pour désambiguïser les ex-æquo. Tandis que ce processus génère une distribution équitable des casiers (environ le même nombre de clients dans chaque casier), il peut entraîner des scores complètement différents pour les clients avec des valeurs de récence, de fréquence, et/ou monétaires semblables voire identiques, surtout si le nombre total de clients est relativement faible et/ou le nombre d'ex-æquo élevé.

Table 2-1

Affecter des ex æquo au même casier et Affectation aléatoire des ex æquo

ID	Achat le plus récent (récence)	Affecter des ex æquo au même casier	Affectation aléatoire des ex æquo
1	10/29/2006	5	5
2	10/28/2006	4	4
3	10/28/2006	4	4
4	10/28/2006	4	5
5	10/28/2006	4	3
6	9/21/2006	3	3
7	9/21/2006	3	2
8	8/13/2006	2	2

ID	Achat le plus récent (récence)	Affecter des ex æquo au même casier	Affectation aléatoire des ex æquo
9	8/13/2006	2	1
10	6/20/2006	1	1

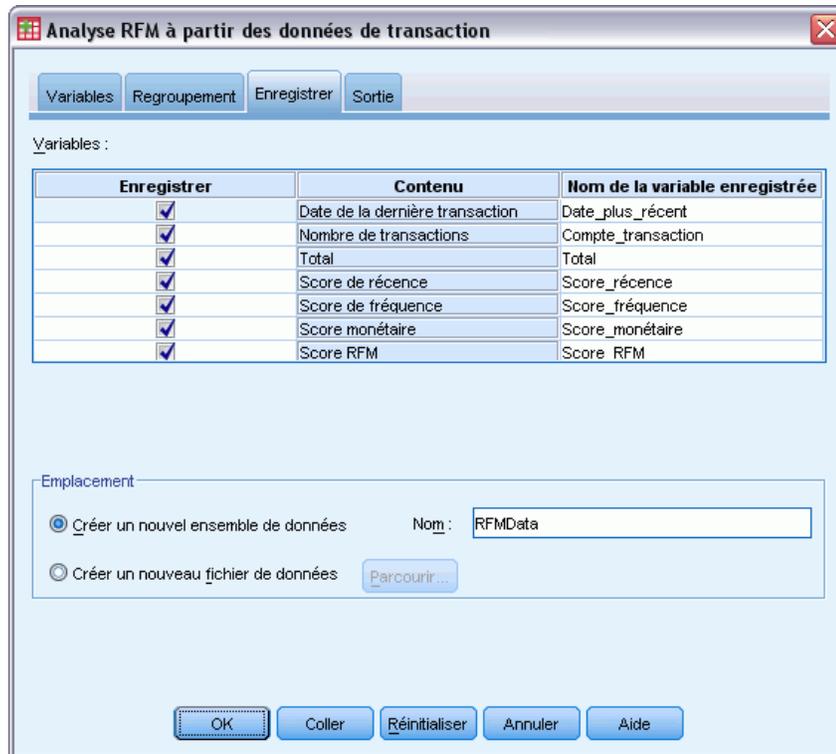
- Dans cet exemple, l'affectation d'ex-æquo au même casier entraîne une distribution inéquitable des casiers : 5 (10%), 4 (40%), 3 (20%), 2 (20%), 1 (10%).
- L'affectation aléatoire d'ex-æquo entraîne 20% dans chaque casier, mais pour obtenir ce résultat, les quatre observations avec une valeur de date de 28/10/2006 sont affectées à 3 casiers différents, et les 2 observations avec une valeur de date de 13/8/2006 sont aussi affectées à des casiers différents.

Notez que la manière dont les ex-æquo sont affectés à des casiers différents est entièrement aléatoire (dans les limites du résultat final étant un nombre égal d'observations dans chaque casier). Si vous avez calculé un deuxième ensemble de scores à l'aide de la même méthode, le classement d'une observation particulière avec une valeur ex-æquo peut changer. Par exemple, les rangs de récence de 5 et 3 pour les observations 4 et 5 respectivement, peuvent être changés la seconde fois.

Enregistrement des scores RFM de Données de transactions

RFM de Données de transactions permet de toujours créer un ensemble de données agrégé avec une seule ligne pour chaque client. L'onglet Enregistrer permet d'indiquer les scores et autres variables à enregistrer et l'emplacement de leur enregistrement.

Figure 2-7
Données de transaction, onglet Enregistrer



Variables

Les variables d'identification, qui identifient de façon unique chaque client, sont enregistrées automatiquement dans le nouvel ensemble de données. Les autres variables suivantes peuvent être enregistrées dans le nouvel ensemble de données :

- **Date de la transaction la plus récente pour chaque client.**
- **Nombre de transactions** Le nombre total de lignes de transaction pour chaque client.
- **Montant.** Le montant récapitulatif pour chaque client selon la méthode récapitulative sélectionnée dans l'onglet Variables.
- **Score de récence.** Le score assigné à chaque client selon la date de transaction la plus récente. Des scores plus élevés indiquent des dates de transaction plus récentes.
- **Score de fréquence.** Le score assigné à chaque client selon le nombre total de transactions. Des scores plus élevés indiquent davantage de transactions.
- **Score monétaire.** Le score affecté à chaque client selon la mesure récapitulative monétaire sélectionnée. Des scores plus élevés indiquent une valeur plus grande pour la mesure récapitulative monétaire.
- **Score RFM.** Les trois scores individuels combinés en une seule valeur : $(récence \times 100) + (fréquence \times 10) + monétaire$.

Par défaut, toutes les variables disponibles sont comprises dans le nouvel ensemble de données ; ainsi, désélectionnez (sélectionnez) celles à exclure. Vous pouvez en option préciser vos propres noms de variables. Les noms de variables doivent être conformes aux règles standard de dénomination de variables.

Emplacement

RFM de Données de transactions permet de toujours créer un ensemble de données agrégé avec une seule ligne pour chaque client. Vous pouvez créer un nouvel ensemble de données dans la session en cours ou enregistrer les données de score RFM dans un fichier de données externe. Les noms des ensembles de données doivent être conformes aux règles standard de dénomination de variables. (Cette restriction ne s'applique pas aux noms de fichiers de données externes).

Enregistrement des Scores RFM de Données de clients

Pour les données de clients, vous pouvez ajouter les variables de score RFM à l'ensemble de données actif ou créer un ensemble de données qui contient les variables de scores sélectionnées. L'onglet Enregistrer permet d'indiquer les scores de variables à enregistrer et l'emplacement de leur enregistrement.

Figure 2-8
Données clients , onglet Enregistrer

Analyse RFM à partir des données des clients

Variables Regroupement **Enregistrer** Sortie

Noms de la variable enregistrée

Génère automatiquement des noms uniques

Noms personnalisés

Variables :

Enregistrer	Contenu	Nom de la variable enregistrée
<input checked="" type="checkbox"/>	Score de récence	Score_récence
<input checked="" type="checkbox"/>	Score de fréquence	Score_fréquence
<input checked="" type="checkbox"/>	Score monétaire	Score_monétaire
<input checked="" type="checkbox"/>	Score RFM	Score RFM

Emplacement

Ecrire dans l'ensemble de données actif

Créer un nouvel ensemble de données Nom :

Créer un nouveau fichier de données

Noms des variables enregistrées

- **Générer automatiquement des noms uniques.** Lors de l'ajout de variables de scores à l'ensemble de données actif, ceci garantit que les nouveaux noms de variables sont uniques. Ceci est particulièrement utile si vous souhaitez ajouter plusieurs ensembles de scores RFM différents (selon des critères différents)
- **Noms personnalisés.** Ceci permet d'affecter vos propres noms de variables aux variables de scores. Les noms de variables doivent être conformes aux règles standard de dénomination de variables.

Variables

Sélectionnez (cochez) les variables de scores à enregistrer :

- **Score de récence.** Le score affecté à chaque client selon la valeur de la variable Date de transaction ou Intervalle sélectionnée dans l'onglet Variables. Des scores élevés sont affectés à des dates plus récentes ou à des valeurs d'intervalles inférieures.
- **Score de fréquence.** Le score affecté à chaque client selon la valeur de la variable Nombre de transactions sélectionnée dans l'onglet Variables. Les scores les plus élevés sont affectés à des valeurs plus grandes.
- **Score monétaire.** Le score affecté à chaque client selon la valeur de la variable Montant sélectionnée dans l'onglet Variables. Les scores les plus élevés sont affectés à des valeurs plus grandes.
- **Score RFM.** Les trois scores individuels combinés en une seule valeur :
 $(récence*100)+(fréquence*10)+monétaire.$

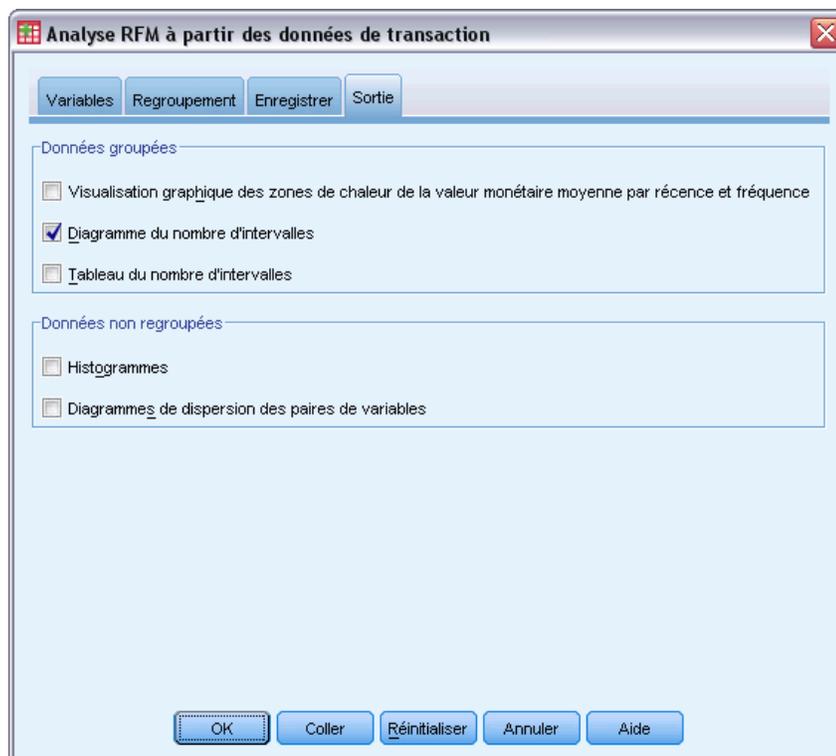
Emplacement

Pour les données de clients, trois alternatives sont possibles pour l'emplacement de l'enregistrement des nouveaux scores RFM :

- **Ensemble de données actif.** Les variables de scores RFM sélectionnées sont ajoutées à l'ensemble de données actif.
- **Nouvel ensemble de données.** Les variables de scores RFM sélectionnées et les variables d'identification qui identifient de façon unique chaque client (observation) sont écrites dans un nouvel ensemble de données de la session en cours. Les noms des ensembles de données doivent être conformes aux règles standard de dénomination de variables. Cette option n'est disponible que si vous sélectionnez une ou plusieurs variables Identificateur client de l'onglet Variables.
- **Fichier.** Les scores RFM sélectionnés et les variables d'identification qui identifient de façon unique chaque client (observation) sont enregistrés dans un fichier de données externes. Cette option n'est disponible que si vous sélectionnez une ou plusieurs variables Identificateur client de l'onglet Variables.

Résultats RFM

Figure 2-9
Onglet Résultats RFM



Données groupées

Les graphiques et les tableaux de données regroupées par casiers sont basés sur le calcul des scores de récence, de fréquence et monétaires.

Carte des zones de chaleur de la valeur monétaire moyenne par récence et par fréquence. La carte des zones de chaleur de la distribution monétaire moyenne indique la valeur monétaire moyenne pour les catégories définies par des scores de récence et de fréquence. Les zones les plus foncées indiquent une valeur monétaire moyenne plus élevée.

Diagramme des effectifs de casier. Le diagramme des effectifs de casier affiche la distribution des casiers pour la méthode de regroupement sélectionnée. Chaque bâton représente le nombre d'observations qui sera assigné à chaque score RFM combiné.

- Bien qu'une distribution régulière, avec des bâtons ayant tous (ou presque tous) approximativement la même taille, soit généralement souhaitée, il faut néanmoins s'attendre à une certaine variance lors de l'utilisation de la méthode de regroupement par défaut qui assigne les valeurs ex aequo au même casier.
- Des fluctuations extrêmes dans la distribution des casiers et/ou de nombreux casiers vides peuvent indiquer que vous devriez essayer une autre méthode de regroupement (casiers moins nombreux et/ou attribution aléatoire d'ex aequo) ou essayer de remettre en question l'adéquation de l'analyse RFM.

Tableau des effectifs de casier. Informations semblables à celles du diagramme des effectifs de casier, sauf qu'elles sont exprimées sous forme d'un tableau, avec des effectifs de casier dans chaque cellule.

Données non regroupées

Les graphiques et les tableaux de données non regroupées par casiers sont basés sur les variables d'origine utilisées pour créer les scores de récence, de fréquence et monétaires.

Histogrammes. Les histogrammes indiquent la distribution relative des valeurs pour les trois variables utilisées pour le calcul des scores de récence, de fréquence et monétaires. Ces histogrammes présentent plus souvent des distributions relativement asymétriques que des distributions normales ou symétriques.

L'axe horizontal de chaque histogramme est toujours ordonné des valeurs les moins élevées sur la gauche aux valeurs les plus élevées sur la droite. Cependant, avec la récence, l'interprétation du diagramme dépend du type de mesure de la récence : intervalle de date ou de temps. Pour les dates, les bâtons sur la gauche représentent des valeurs plus éloignées dans le temps (une date moins récente a une valeur inférieure à celle d'une date plus récente). Pour les intervalles de temps, les bâtons sur la gauche représentent des valeurs plus récentes (plus l'intervalle de temps est réduit, plus la transaction est récente).

Diagrammes de dispersion de paires de variables. Ces diagrammes de dispersion présentent la relation entre les trois variables utilisées pour calculer les scores de récence, de fréquence et monétaires.

Il est habituel d'observer des regroupements linéaires de points sur l'échelle de fréquence, car la fréquence représente souvent un intervalle de valeurs discrètes relativement réduit. Par exemple, si le nombre total des transactions ne dépasse pas 15, alors il n'existe que 15 valeurs de fréquence possibles (sauf si les transactions fractionnelles sont comptabilisées), alors qu'il existe des centaines de valeurs de récence et des milliers de valeurs monétaires possibles.

L'interprétation de l'axe de la récence dépend du type de mesure de la récence : intervalle de date ou de temps. Pour les dates, les points les plus proches de l'origine représentent des dates plus éloignées dans le temps. Pour les intervalles de temps, des points plus proches de l'origine représentent des valeurs plus récentes.

Classification

La classification est un outil d'exploration conçu pour révéler des groupements naturels (ou classes) au sein de données. Par exemple, elle permet d'identifier différents groupes de clients en fonction de diverses caractéristiques démographiques et d'achat.

Exemple : Les entreprises du domaine des produits de consommation et du commerce de détail utilisent régulièrement des techniques de classification des données qui décrivent les habitudes d'achat, le sexe, l'âge, le niveau de revenu, etc. de leurs clients. Ces sociétés adaptent leurs stratégies de marketing et de développement produit à chaque groupe de consommation afin d'augmenter les ventes et de développer la fidélité à la marque.

Remarques sur les données de la classification

Données. Cette procédure fonctionne avec des champs continus et qualitatifs. Chaque enregistrement (ligne) représente un client à classer et les champs (variables) représentent les attributs sur lesquels est basée la classification.

Ordre des enregistrements. Notez que les résultats peuvent dépendre de l'ordre des enregistrements. Pour réduire les effets de tri, classez les enregistrements de manière aléatoire. Vous pouvez exécuter l'analyse plusieurs fois, avec les enregistrements triés de manière aléatoire, afin de vérifier la stabilité d'une solution donnée.

Niveau de mesure. L'affectation d'un niveau de mesure correct est important car il affecte le calcul des résultats.

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.
- **Continu.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

Une icône en regard de chaque champ indique le niveau de mesure actuel.

	Numérique	Chaîne	Date	Heure
Echelle (continue).		n/a		
Ordinal				
Nominal				

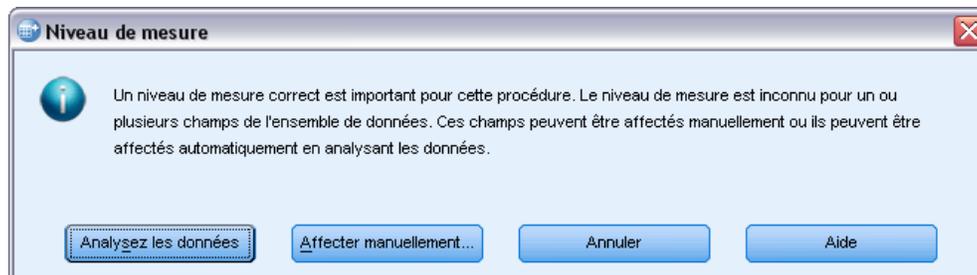
Vous pouvez modifier le niveau de mesure dans la vue des variables de l'éditeur de données ou utilisez la boîte de dialogue Définir les propriétés de variable pour suggérer un niveau de mesure approprié pour chaque champ .

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 3-1

Alerte du niveau de mesure



- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Pour obtenir une classification

A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique

- Sélectionnez Segmenter mes contacts en classes.

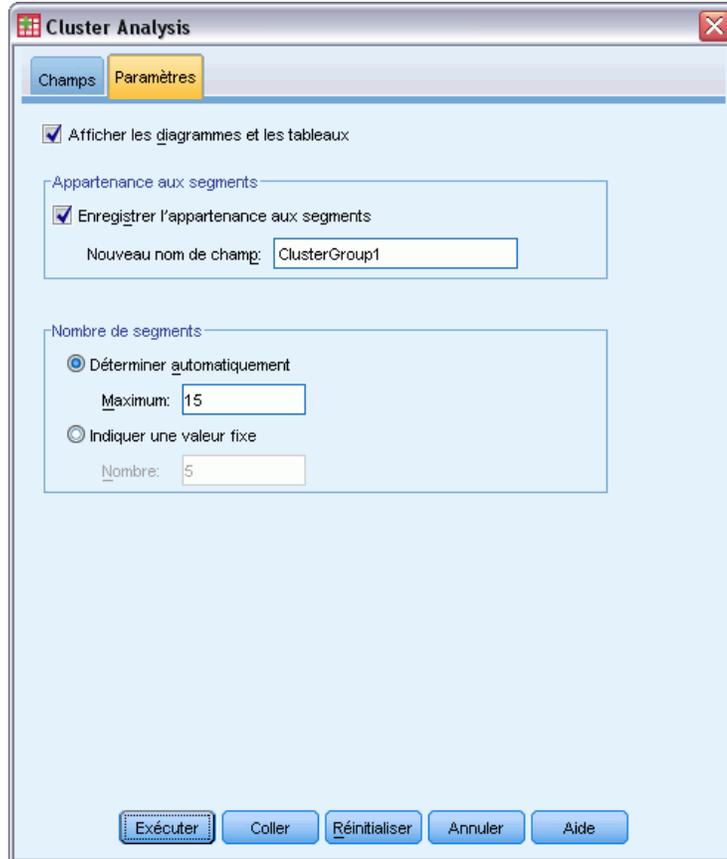
Figure 3-2
Onglet Champs de la classification



- Sélectionnez les champs qualitatifs (nominaux, ordinaux) et continus (d'échelle) que vous souhaitez utiliser pour créer des segments.
- Cliquez sur Exécuter pour exécuter la procédure.

Paramètres

Figure 3-3
Onglet Paramètres de la classification



L'onglet Paramètres vous permet d'afficher ou de masquer les diagrammes et tableaux qui décrivent les segments, d'enregistrer un nouveau champ qui identifie le segment (classe) pour chaque enregistrement de l'ensemble de données et de spécifier le nombre de segments à inclure dans la solution de classe.

Afficher les diagrammes et les tableaux. Permet d'afficher les tableaux et les diagrammes qui décrivent les segments.

Appartenance aux segments. Permet d'enregistrer un nouveau champ (variable) qui identifie le segment auquel appartient l'enregistrement.

- Les noms de champ doivent être conformes aux règles de dénomination IBM® SPSS® Statistics.
- Le nom de champ de l'appartenance au segment ne peut pas dupliquer un nom de champ qui existe déjà dans l'ensemble de données. Si vous exécutez cette procédure plusieurs fois sur le même ensemble de données, vous devrez spécifier un nom différent à chaque fois.

- **Nombre de segments.** Contrôle le nombre des segments.
- **Déterminer automatiquement.** Cette procédure déterminera automatiquement le meilleur nombre de segments, jusqu'au maximum autorisé.

Indiquer une valeur fixe. La procédure va produire le nombre de segments spécifié.

Profils des clients potentiels

Cette technique utilise les résultats d'une campagne précédente ou d'une campagne test pour créer des profils descriptifs. Vous pouvez utiliser ces profils pour cibler des groupes de contacts spécifiques lors des futures campagnes. Le champ Réponse indique qui a répondu à la campagne précédente ou à la campagne test. La liste Profils contient les caractéristiques à utiliser pour créer le profil.

Exemple : En fonction des résultats d'un envoi d'essai, la division de marketing direct d'une entreprise souhaite générer des profils des types de clients les plus susceptibles de répondre à une offre, sur la base d'informations démographiques.

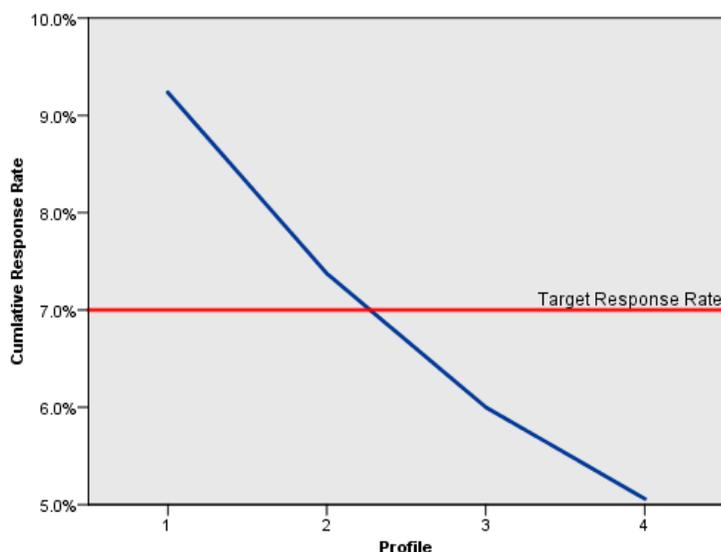
Résultats

Le résultat contient un tableau qui fournit une description de chaque groupe de profils et affiche les taux de réponse (pourcentage de réponses positives) et les taux de réponse cumulés, ainsi qu'un diagramme des taux de réponses cumulés. Si vous incluez un taux de réponse minimum cible, le tableau sera codé par couleurs pour indiquer les profils qui satisfont le taux de réponse cumulé minimum, et le diagramme contiendra une ligne de référence au niveau de la valeur du taux de réponse minimum spécifié.

Figure 4-1
Tableau et diagramme des taux de réponse

Response Rate Table				
Number	Profile			
	Description	Group Size	Response Rate	Cumulative Response Rate
1	Region = "West", "South", "East" Gender = "Female" Married = "No"	379	9.2%	9.2%
2	Region = "West", "South", "East" Gender = "Female" Married = "Yes"	299	5.0%	7.4%
3	Region = "West", "South", "East" Gender = "Male"	722	4.7%	6.0%
4	Region = "North"	517	2.5%	5.1%

Green: Meets target response rate.
Red: Does not meet target response rate.



Remarques sur les données de profil des clients potentiels

Champ de réponse. Le champ de réponse doit être nominal ou ordinal. Il peut s'agir d'une variable de type chaîne ou numérique. Si ce champ contient une valeur indiquant un nombre d'achats ou leur montant, vous devrez créer un nouveau champ dans lequel une valeur unique représente toutes les réponses positives. [Pour plus d'informations, reportez-vous à la section Création d'un champ de réponse qualitatif sur p. 25.](#)

Valeur de réponse positive. La valeur de réponse positive identifie les clients qui ont répondu positivement (par exemple, qui ont effectué un achat). Toutes les autres valeurs de réponse non manquantes sont supposées indiquer une réponse négative. S'il existe des étiquettes de valeurs définies pour le champ de réponse, ces étiquettes s'affichent dans la liste déroulante.

Créez des profils avec. Ces champs peuvent être de type nominal, ordinal ou continu (échelle). Il peut s'agir de variables de type chaîne ou numérique.

Niveau de mesure. L'affectation d'un niveau de mesure correct est important car il affecte le calcul des résultats.

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.
- **Continu.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

Une icône en regard de chaque champ indique le niveau de mesure actuel.

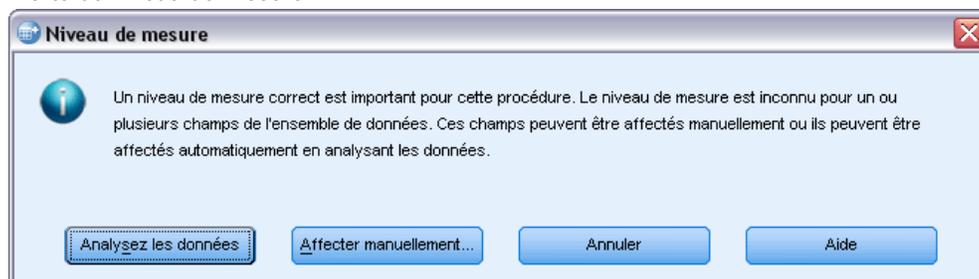
	Numérique	Chaîne	Date	Heure
Echelle (continue).		n/a		
Ordinal				
Nominal				

Vous pouvez modifier le niveau de mesure dans la vue des variables de l'éditeur de données ou utilisez la boîte de dialogue Définir les propriétés de variable pour suggérer un niveau de mesure approprié pour chaque champ .

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 4-2
Alerte du niveau de mesure



- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Pour obtenir les profils des clients potentiels

A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique

- Sélectionnez Générer les profils de mes contacts qui ont répondu à une offre.

Figure 4-3
Onglet Champs Profils des clients potentiels



- ▶ Sélectionnez le champ qui identifie les contacts ayant répondu à l'offre. Ce champ doit être nominal ou ordinal.
- ▶ Entrez la valeur qui indique une réponse positive. Si des valeurs disposent d'étiquettes de valeur définies, vous pouvez sélectionner l'étiquette de valeur dans la liste déroulante, et la valeur correspondante sera affichée.
- ▶ Sélectionnez les champs à utiliser pour créer les profils.
- ▶ Cliquez sur Exécuter pour exécuter la procédure.

Paramètres

Figure 4-4
Onglet Paramètres des profils des clients potentiels

Profils des clients potentiels

Champs Paramètres

Taille minimale du groupe: 100

Inclure des informations de seuil minimal de taux de réponse dans les résultats.

Specify target response rate (%): 7

Exécuter Coller Réinitialiser Annuler Aide

L'onglet Paramètres vous permet de contrôler la taille minimale des groupes de profils et inclut un seuil de taux de réponse minimum dans les résultats.

Taille minimale du groupe de profils. Chaque profil représente les caractéristiques partagées d'un groupe de contacts dans l'ensemble de données (par exemple, femmes de moins de 40 ans vivant dans l'Ouest). Par défaut, la plus petite taille de groupe est de 100. Les petites tailles de groupe permettent de révéler davantage de groupes, mais les tailles de groupe plus grandes peuvent fournir des résultats plus fiables. Il doit s'agir d'un nombre entier positif.

Inclure des informations de seuil minimal de taux de réponse dans les résultats. Les résultats contiennent un tableau qui affiche les taux de réponse (pourcentage de réponses positives) et les taux de réponse cumulés, ainsi qu'un diagramme des taux de réponses cumulés. Si vous saisissez un taux de réponse minimum cible, le tableau sera codé par couleurs pour indiquer les profils qui satisfont le taux de réponse cumulé minimum, et le diagramme contiendra une ligne de référence au niveau de la valeur du taux de réponse minimum spécifié. La valeur doit être supérieure à 0 et inférieure à 100.

Création d'un champ de réponse qualitatif

Le champ de réponse doit être qualitatif, avec une valeur représentant toutes les réponses positives. Toutes les autres valeurs non manquantes sont supposées indiquer une réponse négative. Si le champ de réponse représente une valeur continue (échelle), telle que le nombre d'achats ou leur

valeur monétaire, vous devez créer un nouveau champ qui assigne une valeur de réponse positive unique à toutes les réponses différentes de zéro.

- Si les réponses négatives sont enregistrées sous la forme d'un 0 (et non un blanc, ce qui est traité comme une valeur manquante), ceci peut être calculé avec la formule suivante :

$NewName = OldName > 0$

où *NewName* est le nom du nouveau champ et *OldName* est le nom du champ d'origine.

Il s'agit d'une expression logique qui assigne une valeur de 1 à toutes les valeurs non manquantes supérieures à 0, et 0 à toutes les valeurs non manquantes inférieures ou égales à 0.

- Si aucune valeur n'est enregistrée pour les réponses négatives, ces valeurs sont traitées comme manquantes, et la formule est légèrement plus compliquée :

$NewName = NOT(MISSING(OldName))$

Dans cette expression logique, toutes les valeurs de réponse non manquantes se voient attribuer une valeur de 1 et toutes les valeurs de réponse manquantes se voient attribuer une valeur de 0.

- Si vous ne pouvez pas faire la distinction entre les valeurs de réponse négatives (0) et les valeurs manquantes, le calcul d'une valeur de réponse précise est impossible. S'il existe relativement peu de valeurs véritablement manquantes, cela n'aura peut-être pas d'effet significatif sur les taux de réponse calculés. Toutefois, si de nombreuses valeurs sont manquantes, par exemple lorsque les informations de réponse sont enregistrées uniquement pour un petit échantillon de test de l'ensemble de données total, les taux de réponse calculés ne seront pas pertinents, car ils seront considérablement inférieurs aux vrais taux de réponse.

Pour créer un champ de réponse qualitatif

- ▶ A partir des menus, sélectionnez :
Transformer > Calculer la variable
- ▶ Pour la variable cible, entrez le nom du nouveau champ (variable).
- ▶ Si des réponses négatives sont enregistrées en tant que 0, pour l'expression numérique saisissez $OldName > 0$, où *OldName* est le nom du champ d'origine.
- ▶ Si des réponses négatives sont enregistrées en tant que valeur manquante (blanc), pour l'expression numérique saisissez $NOT(MISSING(OldName))$, où *OldName* est le nom du champ d'origine.

Taux de réponse des codes postaux

Cette technique utilise les résultats de la campagne précédente pour calculer les taux de réponse par code postal. Ces taux peuvent être utilisés pour cibler des codes postaux spécifiques lors des futures campagnes. Le champ Réponse indique qui a répondu à la campagne précédente. Le champ Code postal identifie le champ qui contient les codes postaux.

Exemple : En fonction des résultats d'un précédent mailing, la division marketing direct d'une entreprise génère des taux de réponse par code postal. Sur la base de plusieurs critères, tels que le taux de réponse minimum acceptable et/ou le nombre maximum de contacts à inclure dans le mailing, il est ensuite possible de cibler des codes postaux spécifiques.

Résultats

Le résultat de cette procédure inclut un nouvel ensemble de données qui contient les taux de réponse par code postal, ainsi qu'un tableau et un diagramme qui récapitulent les résultats par rang de décile (premiers 10%, 20%, etc.). Le tableau peut également faire l'objet d'un codage couleur, en fonction d'un taux de réponse cumulé minimum spécifié par l'utilisateur ou d'un nombre de contacts maximum.

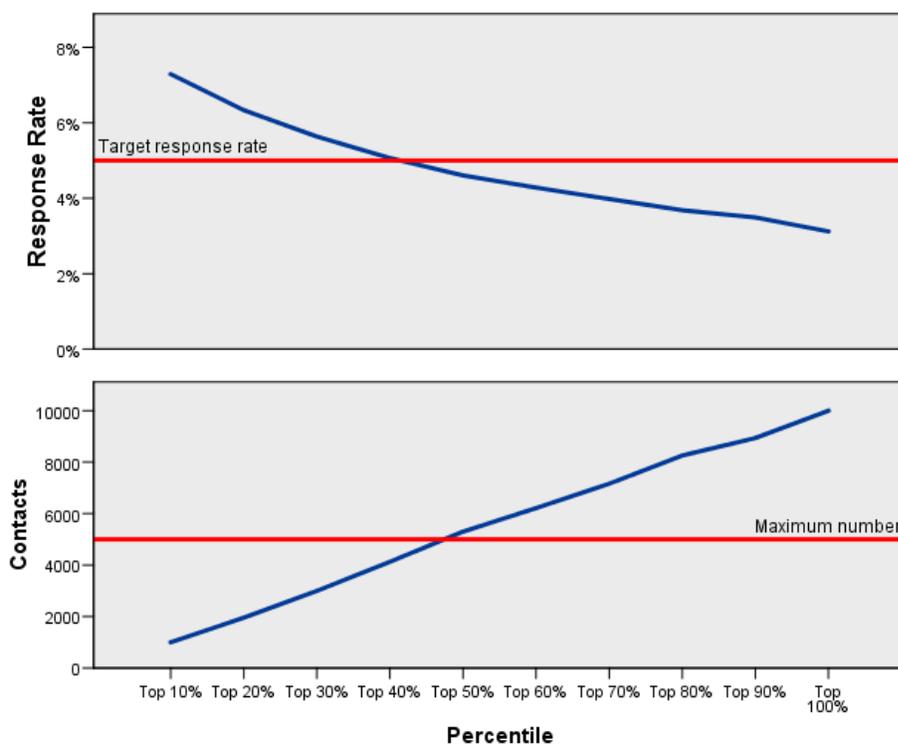
Figure 5-1
Ensemble de données avec taux de réponse par code postal

	PostalCode	ResponseRate	Responses	Contacts	Index	Rank
1	932	10.00%	4.00	40	3.60	1
2	098	8.82%	6.00	68	5.47	1
3	740	7.76%	9.00	116	8.30	1
4	100	7.69%	7.00	91	6.46	1
5	110	7.69%	5.00	65	4.62	1
6	954	7.55%	4.00	53	3.70	1
7	108	7.32%	6.00	82	5.56	1
8	107	7.04%	5.00	71	4.65	1
9	090	6.90%	4.00	58	3.72	1
10	966	6.90%	4.00	58	3.72	1
11	760	6.72%	8.00	119	7.46	1
12	113	6.25%	5.00	80	4.69	1

Figure 5-2
Tableau et diagramme récapitulatifs

Taux de réponse				
Percentile	Taux de réponse	Contacts	Cumulative Response Rate	Total Contacts
Top 10%	7.3	1001	7.3	1001
Top 20%	5.3	956	6.3	1957
Top 30%	4.3	1042	5.6	2999
Top 40%	3.5	1127	5.1	4126
Top 50%	3.0	1173	4.6	5299
Top 60%	2.4	914	4.3	6213
Top 70%	2.0	948	4.0	7161
Top 80%	1.7	1095	3.7	8256
Top 90%	1.2	680	3.5	8936
Top 100%	.0	1064	3.1	10000

Green Red Caption



Le nouvel ensemble de données contient les champs suivants :

- **Code postal.** Si les groupes de codes postaux reposent sur une partie seulement de la valeur complète, il s'agit de la valeur de cette partie du code postal. L'intitulé de ligne pour cette colonne dans le fichier Excel est le nom du champ de code postal de l'ensemble de données d'origine.
- **TauxDeRéponse.** Pourcentage de réponses positives dans chaque code postal.

- **Réponses.** Nombre de réponses positives dans chaque code postal.
- **Contacts.** Nombre total de contacts dans chaque code postal qui contiennent une valeur non manquante dans le champ de réponse.
- **Index.** Réponse « pondérée » basée sur la formule $N \times P \times (1-P)$, où N est le nombre de contacts, et P est le taux de réponse exprimé en proportion.
- **Rang.** Rang de décile (premiers 10%, 20% , etc.) des taux de réponse de code postal cumulés en ordre décroissant.

Analyse des données de taux de réponse des codes postaux

Champ de réponse. Le champ de réponse peut être une variable numérique ou une chaîne. Si ce champ contient une valeur indiquant un nombre d'achats ou leur valeur monétaire, vous devez créer un nouveau champ dans lequel une valeur unique représente toutes les réponses positives. [Pour plus d'informations, reportez-vous à la section Création d'un champ de réponse qualitatif sur p. 33.](#)

Valeur de réponse positive. La valeur de réponse positive identifie les clients qui ont répondu positivement (par exemple, qui ont effectué un achat). Toutes les autres valeurs de réponse non manquantes sont supposées indiquer une réponse négative. S'il existe des étiquettes de valeurs définies pour le champ de réponse, ces étiquettes s'affichent dans la liste déroulante.

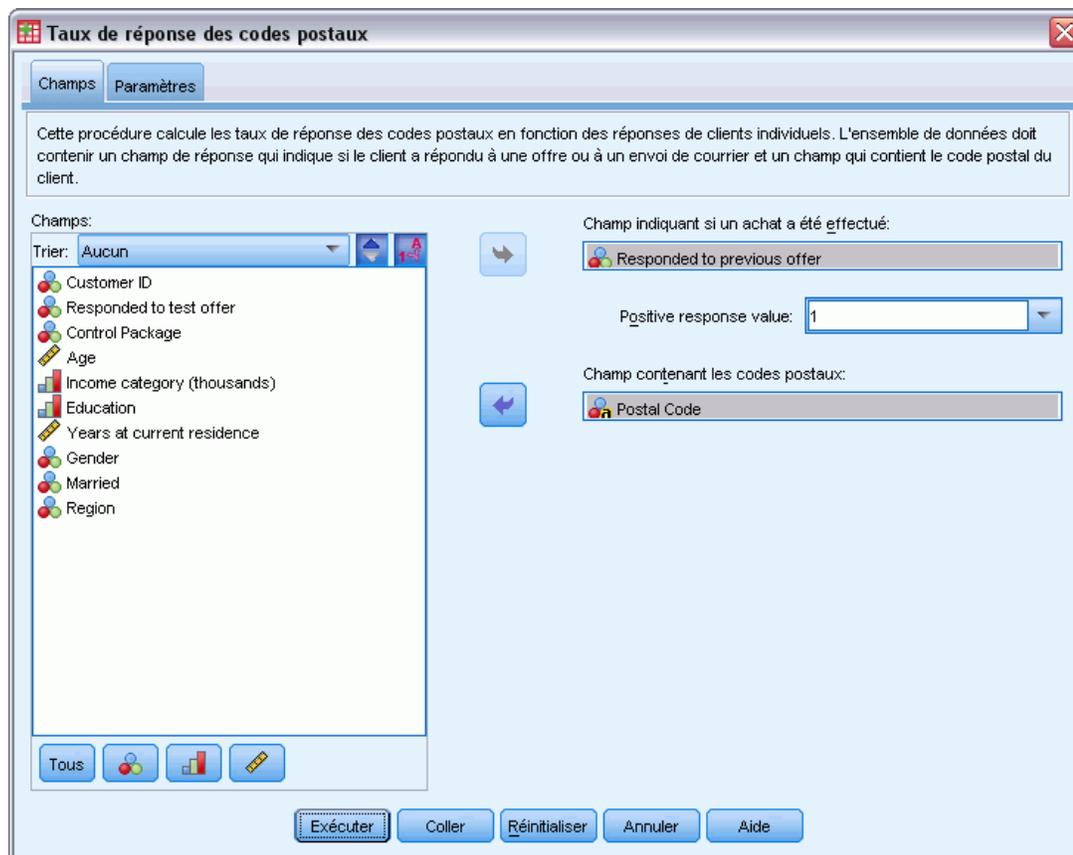
Champ Code postal. Le champ de code postal peut être une variable numérique ou une chaîne.

Pour obtenir les taux de réponse des codes postaux

A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique

- ▶ Sélectionnez Identifier les codes postaux avec les meilleurs taux de réponse.

Figure 5-3
Onglet Champs Taux de réponse des codes postaux



- ▶ Sélectionnez le champ qui identifie les contacts ayant répondu à l'offre.
- ▶ Entrez la valeur qui indique une réponse positive. Si des valeurs disposent d'étiquettes de valeur définies, vous pouvez sélectionner l'étiquette de valeur dans la liste déroulante, et la valeur correspondante sera affichée.
- ▶ Sélectionnez le champ qui contient le code postal.
- ▶ Cliquez sur Exécuter pour exécuter la procédure.

Eventuellement, vous pouvez aussi :

- Générer des taux de réponse en fonction des n premiers caractères ou chiffres du code postal au lieu de la valeur complète
- Enregistrer les résultats dans un fichier Excel
- Contrôler les options d'affichage des résultats

Paramètres

Figure 5-4
Onglet Paramètres des taux de réponse des codes postaux

Grouper les codes postaux selon

Ceci détermine la manière dont les enregistrements sont groupés pour calculer les taux de réponse. Par défaut, le code postal entier est utilisé, et tous les enregistrements avec le même code postal sont regroupés pour calculer le taux de réponse du groupe. Vous pouvez également grouper les enregistrements selon une partie seulement du code postal complet, composée des n premiers chiffres ou caractères. Vous pouvez par exemple grouper des enregistrements en fonction uniquement des 5 premiers caractères d'un code postal de 10 caractères, ou des 3 premiers caractères d'un code postal de 5 caractères. L'ensemble de données de sortie contiendra un enregistrement pour chaque groupe de codes postaux. Si vous saisissez une valeur, il doit s'agir d'un nombre entier positif.

format numérique du code postal

Si le champ de code postal est numérique et que vous souhaitez grouper des codes postaux selon les n premiers chiffres au lieu de la valeur entière, vous devez spécifier le nombre de chiffres dans la valeur d'origine. Le nombre de chiffres est le nombre de chiffres *maximum* possible dans le

code postal. Par exemple, si le champ de code postal contient un mélange de codes postaux à 5 chiffres et à 9 chiffres, vous devez indiquer 9 comme nombre de chiffres.

Remarque : Selon le format d'affichage, certains codes postaux à 5 chiffres peuvent apparaître avec seulement 4 chiffres, mais ils contiennent un zéro implicite à gauche.

Résultats

Outre le nouvel ensemble de données qui contient les taux de réponse par code postal, vous pouvez afficher un tableau et un graphique qui récapitulent les résultats par rang de décile (premiers 10%, 20%, etc.). Le tableau affiche les taux de réponse, les taux de réponse cumulés, le nombre d'enregistrements, et le nombre cumulé d'enregistrements dans chaque décile. Le diagramme affiche les taux de réponse cumulés et le nombre cumulé d'enregistrements dans chaque décile.

Taux de réponse acceptable minimum. Si vous saisissez un taux de réponse minimum cible ou une formule de calcul de point-mort, le tableau sera codé par couleurs pour indiquer les déciles qui satisfont le taux de réponse cumulé minimum, et le diagramme contiendra une ligne de référence au niveau de la valeur du taux de réponse minimum spécifié.

- **Taux de réponse cible.** Taux de réponse exprimé en pourcentage (pourcentage de réponses positives dans chaque groupe de codes postaux). La valeur doit être supérieure à 0 et inférieure à 100.
- **Calculer le point mort à partir de la formule.** Calculez le taux de réponse cumulé minimum en fonction de la formule : $(\text{Coût de l'envoi d'un package}/\text{Revenu net par réponse}) \times 100$. Les deux valeurs doivent être positives. Le résultat doit être une valeur supérieure à 0 et inférieure à 100. Par exemple, si le coût d'envoi d'un package est de 0,75 \$, et que le revenu net par réponse est de 56 \$, le taux de réponse minimum est : $(0,75/56) \times 100 = 1,34 \%$.

Nombre maximal de contacts. Si vous spécifiez un nombre maximal de contacts, le tableau sera codé par couleurs pour indiquer les déciles qui ne dépassent pas le nombre de contacts (enregistrements) cumulé maximal et le diagramme contiendra une ligne de référence au niveau de cette valeur.

- **Pourcentage de contacts.** Maximum exprimé en pourcentage. Par exemple, vous souhaitez peut-être connaître les déciles avec les taux de réponse les plus élevés qui ne contiennent pas plus de 50 % des contacts. La valeur doit être supérieure à 0 et inférieure à 100.
- **Nombre de contacts.** Maximum exprimé en nombre de contacts. Par exemple, si vous ne prévoyez pas d'envoyer plus de 10 000 packages, vous pouvez définir la valeur sur 10000. La valeur doit être un nombre entier positif (sans symbole de regroupement).

Si vous spécifiez un taux de réponse minimum acceptable et un nombre maximal de contacts, le codage couleur du tableau s'effectuera selon la condition satisfaite en premier.

Exporter dans Excel

Cette procédure crée automatiquement un nouvel ensemble de données qui contient les taux de réponse par code postal. Chaque enregistrement (ligne) de l'ensemble de données représente un code postal. Vous pouvez enregistrer automatiquement les mêmes informations dans un fichier Excel. Le fichier est enregistré au format Excel 97-2003.

Création d'un champ de réponse qualitatif

Le champ de réponse doit être qualitatif, avec une valeur représentant toutes les réponses positives. Toutes les autres valeurs non manquantes sont supposées indiquer une réponse négative. Si le champ de réponse représente une valeur continue (échelle), telle que le nombre d'achats ou leur valeur monétaire, vous devez créer un nouveau champ qui assigne une valeur de réponse positive unique à toutes les réponses différentes de zéro.

- Si les réponses négatives sont enregistrées sous la forme d'un 0 (et non un blanc, ce qui est traité comme une valeur manquante), ceci peut être calculé avec la formule suivante :

$NewName = OldName > 0$

où *NewName* est le nom du nouveau champ et *OldName* est le nom du champ d'origine.

Il s'agit d'une expression logique qui assigne une valeur de 1 à toutes les valeurs non manquantes supérieures à 0, et 0 à toutes les valeurs non manquantes inférieures ou égales à 0.

- Si aucune valeur n'est enregistrée pour les réponses négatives, ces valeurs sont traitées comme manquantes, et la formule est légèrement plus compliquée :

$NewName = NOT(MISSING(OldName))$

Dans cette expression logique, toutes les valeurs de réponse non manquantes se voient attribuer une valeur de 1 et toutes les valeurs de réponse manquantes se voient attribuer une valeur de 0.

- Si vous ne pouvez pas faire la distinction entre les valeurs de réponse négatives (0) et les valeurs manquantes, le calcul d'une valeur de réponse précise est impossible. S'il existe relativement peu de valeurs véritablement manquantes, cela n'aura peut-être pas d'effet significatif sur les taux de réponse calculés. Toutefois, si de nombreuses valeurs sont manquantes, par exemple lorsque les informations de réponse sont enregistrées uniquement pour un petit échantillon de test de l'ensemble de données total, les taux de réponse calculés ne seront pas pertinents, car ils seront considérablement inférieurs aux vrais taux de réponse.

Pour créer un champ de réponse qualitatif

- ▶ A partir des menus, sélectionnez :
Transformer > Calculer la variable
- ▶ Pour la variable cible, entrez le nom du nouveau champ (variable).
- ▶ Si des réponses négatives sont enregistrées en tant que 0, pour l'expression numérique saisissez $OldName > 0$, où *OldName* est le nom du champ d'origine.
- ▶ Si des réponses négatives sont enregistrées en tant que valeur manquante (blanc), pour l'expression numérique saisissez $NOT(MISSING(OldName))$, où *OldName* est le nom du champ d'origine.

Propension à acheter

La propension à acheter utilise les résultats d'un envoi d'essai ou d'une campagne précédente pour générer les scores. Les scores indiquent quels contacts sont les plus susceptibles de répondre. Le champ Réponse indique qui a répondu à l'envoi d'essai ou à la campagne précédente. Les champs Propension sont les caractéristiques à utiliser pour prévoir la probabilité que les contacts avec des caractéristiques similaires répondent.

Cette technique utilise la régression logistique binaire pour construire un modèle prédictif. Le processus de construction et d'application d'un modèle prédictif se fait en deux étapes de base :

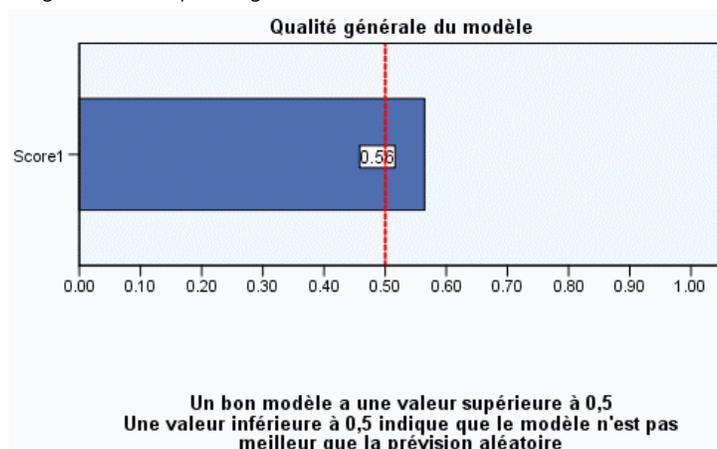
- ▶ La construction du modèle et l'enregistrement du fichier du modèle. Vous construisez le modèle en utilisant un ensemble de données pour lequel le résultat qui vous intéresse (souvent appelé la **cible**) est connu. Par exemple, si vous voulez construire un modèle pour prédire qui est susceptible de répondre à une campagne de publipostage, vous devez commencer par un ensemble de données qui indique déjà qui a répondu et qui n'a pas répondu. Par exemple, il pourrait s'agir des résultats d'un envoi d'essai à un petit groupe de clients ou d'informations sur les réponses à une campagne similaire par le passé.
- ▶ L'application de ce modèle à un ensemble de données différent (pour lequel le résultat qui vous intéresse n'est pas connu) pour prévoir les résultats.

Exemple : La division de marketing direct d'une entreprise utilise les résultats d'un envoi d'essai pour assigner des scores de propension au reste de leur base de données de contacts, à l'aide de diverses caractéristiques démographiques permettant d'identifier les contacts les plus susceptibles de répondre et d'effectuer un achat.

Résultats

Cette procédure crée automatiquement un nouveau champ dans l'ensemble de données qui contient des scores de propension pour les données de test et un fichier de modèle XML qui peut être utilisé pour évaluer d'autres ensembles de données. Le résultat du diagnostic facultatif comprend un diagramme de qualité générale du modèle et un tableau de classement qui compare les réponses prédites aux réponses réelles.

Figure 6-1
Diagramme de qualité générale du modèle



Remarques sur les données de propension à acheter

Champ de réponse. Le champ de réponse peut être une variable numérique ou une chaîne. Si ce champ contient une valeur indiquant un nombre d'achats ou leur valeur monétaire, vous devrez créer un nouveau champ dans lequel une valeur unique représente toutes les réponses positives. [Pour plus d'informations, reportez-vous à la section Création d'un champ de réponse qualitatif sur p. 41.](#)

Valeur de réponse positive. La valeur de réponse positive identifie les clients qui ont répondu positivement (par exemple, qui ont effectué un achat). Toutes les autres valeurs de réponse non manquantes sont supposées indiquer une réponse négative. S'il existe des étiquettes de valeurs définies pour le champ de réponse, ces étiquettes s'affichent dans la liste déroulante.

Prédire la propension avec. Les champs utilisés pour prédire la propension peuvent être une variable numérique ou une chaîne, et ils peuvent être de type nominal, ordinal ou continu (échelle). Toutefois il est important d'affecter le niveau de mesure approprié à tous les champs prédictifs.

Niveau de mesure. L'affectation d'un niveau de mesure correct est important car il affecte le calcul des résultats.

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores

d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.

- **Continu.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

Une icône en regard de chaque champ indique le niveau de mesure actuel.

	Numérique	Chaîne	Date	Heure
Echelle (continue).		n/a		
Ordinal				
Nominal				

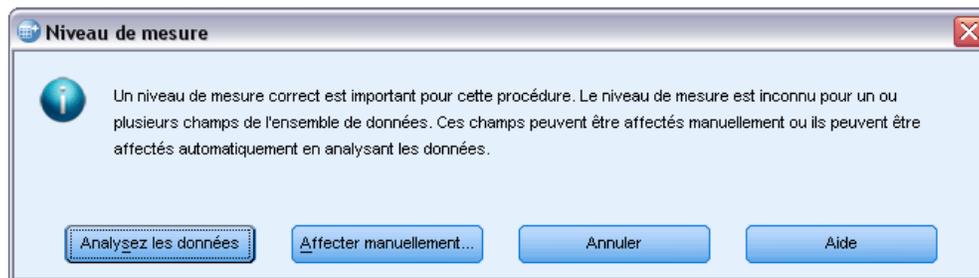
Vous pouvez modifier le niveau de mesure dans la vue des variables de l'éditeur de données ou utilisez la boîte de dialogue Définir les propriétés de variable pour suggérer un niveau de mesure approprié pour chaque champ .

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 6-2

Alerte du niveau de mesure



- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

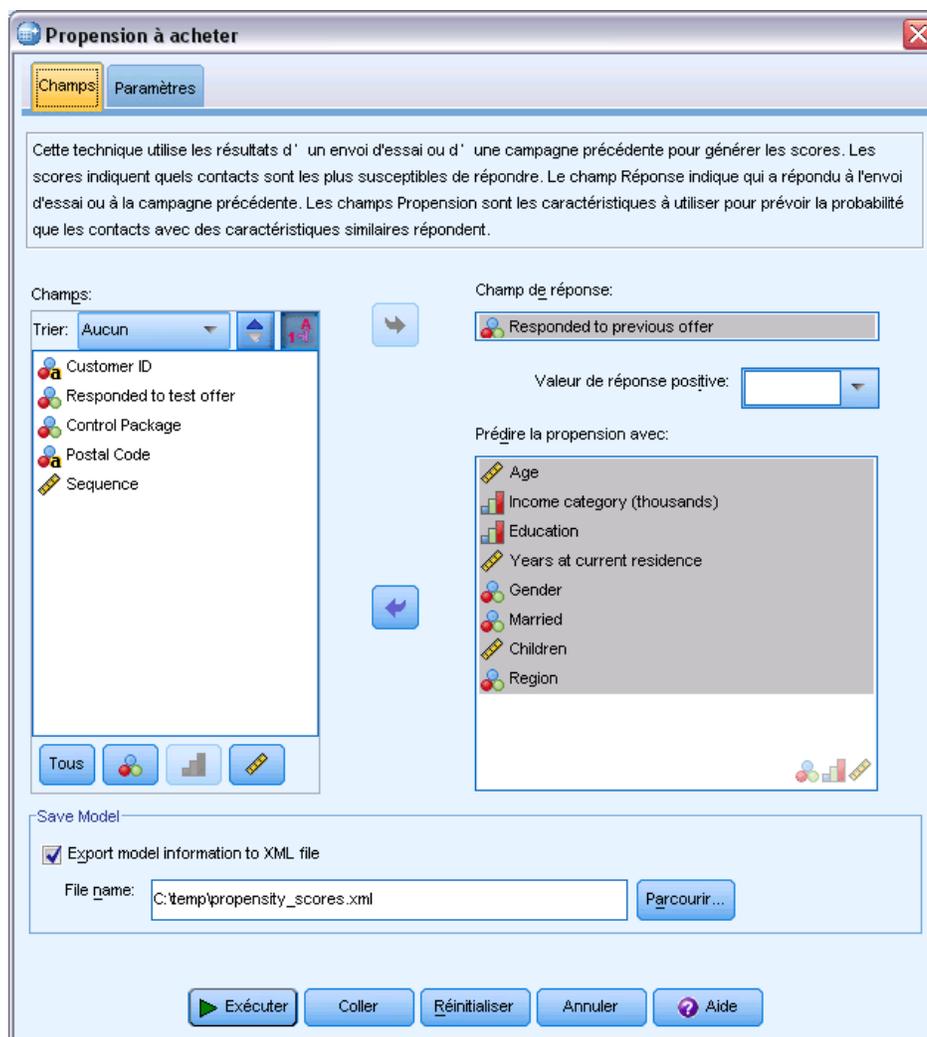
Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Pour obtenir les scores de propension à acheter

A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique

- Sélectionner Sélectionner les contacts les plus susceptibles d'acheter.

Figure 6-3
Onglet Champs Propension à acheter



- Sélectionnez le champ qui identifie les contacts ayant répondu à l'offre.

- ▶ Entrez la valeur qui indique une réponse positive. Si des valeurs disposent d'étiquettes de valeur définies, vous pouvez sélectionner l'étiquette de valeur dans la liste déroulante, et la valeur correspondante sera affichée.
- ▶ Sélectionnez les champs à utiliser pour prédire la propension.
Pour enregistrer un fichier de modèle XML afin d'évaluer d'autres fichiers de données :
- ▶ Sélectionnez (cochez) Exporter les informations du modèle dans un fichier XML.
- ▶ Saisissez un chemin de répertoire et un nom de fichier ou cliquez sur Parcourir pour accéder à l'emplacement où vous souhaitez enregistrer le fichier de modèle XML.
- ▶ Cliquez sur Exécuter pour exécuter la procédure.
Pour utiliser le fichier de modèle afin d'évaluer d'autres ensembles de données :
- ▶ Ouvrez l'ensemble de données que vous souhaitez évaluer.
- ▶ Utilisez l'Assistant d'évaluation pour appliquer le modèle à l'ensemble de données. A partir des menus, sélectionnez :
Utilitaires > Assistant d'évaluation.

Paramètres

Figure 6-4
Onglet Paramètres, Propension à acheter

Propension à acheter

Champs Paramètres

La validation de modèles

Vous pouvez valider le modèle utilisé pour générer les scores. Pour valider le modèle, vous devez diviser vos données en partitions. La partition de formation permet de former ou de construire le modèle. La partition de test permet de valider le modèle. Si vous souhaitez valider le modèle, cette technique attribuera automatiquement les enregistrements aux partitions.

Valider le modèle

Echantillon d'apprentissage taille de la partition (%):

Définir une valeur pour répliquer les résultats

Graine:

Résultat du diagnostic

Qualité générale du modèle

Tableau de classement

Probabilité minimale:

Nom et étiquette du champ de réponses recodé

Cette technique recode automatiquement le champ de réponses en un nouveau champ dans lequel 1 représente des réponses positives et 0 représente des réponses négatives.

Nouveau nom de champ:

Nouvelle étiquette de champ:

Enregistrer les résultats

Cette technique utilise les résultats d'un envoi d'essai ou d'une campagne précédente pour générer les scores. Les scores sont automatiquement enregistrés pour que vous puissiez les utiliser. Les autres contrôles de cet onglet offrent des contrôles supplémentaires sur ce qui a été enregistré.

Nouveau nom de champ pour les scores:

Exécuter Coller Réinitialiser Annuler Aide

La validation de modèles

La validation de modèles crée des groupes d'apprentissage et de test à des fins de diagnostic. Si vous sélectionnez le tableau de classement dans la section Résultat du diagnostic, le tableau sera divisé en deux sections, formation (sélectionnée) et test (non sélectionnée), à des fins de comparaison. Ne sélectionnez pas la validation du modèle à moins de sélectionner également le tableau de classement. Les scores sont basés sur le modèle généré à partir de l'échantillon d'apprentissage, qui contiendra toujours moins d'enregistrements que le nombre total d'enregistrements disponibles. Par exemple, la taille d'échantillon d'apprentissage par défaut est 50%, et il est possible qu'un modèle créé sur la moitié seulement des enregistrements disponibles ne soit pas aussi fiable qu'un modèle créé sur tous les enregistrements disponibles.

- **Echantillon d'apprentissage taille de la partition (%).** Spécifier le pourcentage d'enregistrements à affecter à l'échantillon d'apprentissage. Le reste des enregistrements avec des valeurs non manquantes dans le champ de réponse est affecté à l'échantillon de test. La valeur doit être supérieure à 0 et inférieure à 100.
- **Définir une valeur pour répliquer les résultats.** Les résultats étant affectés au hasard aux échantillons d'apprentissage et de test, chaque fois que vous exécutez la procédure, il est possible que obteniez des résultats différents, à moins de toujours spécifier la même valeur de départ pour le générateur de nombre aléatoire.

Résultat du diagnostic

Qualité générale du modèle. Affiche un diagramme en bâtons de la qualité générale du modèle, exprimée sous la forme d'une valeur comprise entre 0 et 1. Un bon modèle doit avoir une valeur supérieure à 0,5.

Tableau de classement : Affiche un tableau qui compare les réponses positives et négatives prédites aux réponses positives et négatives réelles. Le taux de précision globale peut fournir une certaine indication de la performance du modèle, mais le pourcentage de réponses positives prédites correctes peut s'avérer plus intéressant.

- **Probabilité minimale.** Affecte les enregistrements avec une valeur de score supérieure à la valeur spécifiée à la catégorie de réponses positives prédites dans le tableau de classement. Les scores générés par la procédure représentent la probabilité que le contact réponde de manière positive (par exemple, effectuer un achat). En règle générale, vous devez indiquer une valeur proche de votre taux de réponse cible minimum, exprimé sous forme de proportion. Par exemple, si vous êtes intéressé par un taux de réponse d'au moins 5%, indiquez 0,05. La valeur doit être supérieure à 0 et inférieure à 1.

Nom et étiquette du champ de réponses recodé

Cette procédure recode automatiquement le champ de réponse en un nouveau champ dans lequel 1 représente des réponses positives et 0 représente des réponses négatives, et l'analyse est effectuée sur le champ recodé. Vous pouvez ignorer le nom et l'étiquette par défaut et fournir les vôtres. Les noms doivent être conformes aux règles de dénomination IBM® SPSS® Statistics.

Enregistrer les résultats

Un nouveau champ contenant les scores de propension est automatiquement enregistré dans l'ensemble de données d'origine. Les scores représentent la probabilité d'une réponse positive, exprimée sous forme de proportion.

- Les noms de champ doivent être conformes aux règles de dénomination SPSS Statistics.
- Le nom de champ ne peut pas dupliquer un nom de champ qui existe déjà dans l'ensemble de données. Si vous exécutez cette procédure plusieurs fois sur le même ensemble de données, vous devrez spécifier un nom différent à chaque fois.

Création d'un champ de réponse qualitatif

Le champ de réponse doit être qualitatif, avec une valeur représentant toutes les réponses positives. Toutes les autres valeurs non manquantes sont supposées indiquer une réponse négative. Si le champ de réponse représente une valeur continue (échelle), telle que le nombre d'achats ou leur valeur monétaire, vous devez créer un nouveau champ qui assigne une valeur de réponse positive unique à toutes les réponses différentes de zéro.

- Si les réponses négatives sont enregistrées sous la forme d'un 0 (et non un blanc, ce qui est traité comme une valeur manquante), ceci peut être calculé avec la formule suivante :

$NewName = OldName > 0$

où *NewName* est le nom du nouveau champ et *OldName* est le nom du champ d'origine.

Il s'agit d'une expression logique qui assigne une valeur de 1 à toutes les valeurs non manquantes supérieures à 0, et 0 à toutes les valeurs non manquantes inférieures ou égales à 0.

- Si aucune valeur n'est enregistrée pour les réponses négatives, ces valeurs sont traitées comme manquantes, et la formule est légèrement plus compliquée :

$NewName = NOT(MISSING(OldName))$

Dans cette expression logique, toutes les valeurs de réponse non manquantes se voient attribuer une valeur de 1 et toutes les valeurs de réponse manquantes se voient attribuer une valeur de 0.

- Si vous ne pouvez pas faire la distinction entre les valeurs de réponse négatives (0) et les valeurs manquantes, le calcul d'une valeur de réponse précise est impossible. S'il existe relativement peu de valeurs véritablement manquantes, cela n'aura peut-être pas d'effet significatif sur les taux de réponse calculés. Toutefois, si de nombreuses valeurs sont manquantes, par exemple lorsque les informations de réponse sont enregistrées uniquement pour un petit échantillon de test de l'ensemble de données total, les taux de réponse calculés ne seront pas pertinents, car ils seront considérablement inférieurs aux vrais taux de réponse.

Pour créer un champ de réponse qualitatif

- ▶ A partir des menus, sélectionnez :
Transformer > Calculer la variable
- ▶ Pour la variable cible, entrez le nom du nouveau champ (variable).
- ▶ Si des réponses négatives sont enregistrées en tant que 0, pour l'expression numérique saisissez $OldName > 0$, où *OldName* est le nom du champ d'origine.
- ▶ Si des réponses négatives sont enregistrées en tant que valeur manquante (blanc), pour l'expression numérique saisissez $NOT(MISSING(OldName))$, où *OldName* est le nom du champ d'origine.

Test de package de contrôle

Cette technique compare des campagnes de marketing pour vérifier s’il existe une différence significative de l’efficacité des différents packages et offres. L’efficacité de la campagne est mesurée en fonction des réponses Le champ de campagne identifie différentes campagnes, par exemple Offre A et Offre B. Le champ de réponses indique si un contact a répondu à la campagne. Sélectionnez le Montant d’achat lorsque la réponse est enregistrée comme un montant d’achat, par exemple “99,99”. Sélectionnez Répondre lorsque la réponse est simplement enregistrée comme une réponse positive ou négative, par exemple “Oui” ou “Non”.

Exemple : La division de marketing direct d’une société souhaite vérifier si un nouveau package peut générer plus de réponses positives que le package existant. Ils procèdent à un envoi d’essai pour déterminer si le nouveau package génère un taux de réponse significativement plus élevé. L’envoi d’essai consiste en un groupe de contrôle qui reçoit le package existant et un groupe de test qui reçoit le nouveau package. Les résultats des deux groupes sont comparés pour juger de la différence.

Résultats

Les résultats comprennent un tableau qui affiche les effectifs et les pourcentages des réponses positives et négatives de chaque groupe défini par le champ Campagne et un tableau qui identifie les groupes qui diffèrent entre eux de manière significative.

Figure 7-1
Résultats du Test de package de contrôle

	Control Package			
	Control		Test	
	Effectif	N % colonne	Effectif	N % colonne
Efficacité (1 =OUI 0=NON) 0	875	96,2%	945	93,8%
1	35	3,8%	62	6,2%

Il existe une différence statistiquement significative entre Control et Test.

Remarques et hypothèses concernant les données du test de package de contrôle

Champ Campagne. Le champ Campagne doit être qualitatif (nominal ou ordinal).

Champ de réponse d’efficacité. Si vous avez sélectionné Montant d’achat pour le champ d’efficacité, le champ doit être numérique, et le niveau de mesure doit être continu (échelle).

Si vous ne pouvez pas faire la distinction entre les valeurs de réponse négatives (pour le montant d’achat, une valeur de 0) et les valeurs manquantes, le calcul d’un taux de réponse précise est impossible. S’il existe relativement peu de valeurs véritablement manquantes, cela n’aura peut-être pas d’effet significatif sur les taux de réponse calculés. Toutefois, si de nombreuses valeurs sont

manquantes, par exemple lorsque les informations de réponse sont enregistrées uniquement pour un petit échantillon de test de l'ensemble de données total, les taux de réponse calculés ne seront pas pertinents, car ils seront considérablement inférieurs aux vrais taux de réponse.

Hypothèses : Cette procédure considère que les contacts ont été affectés de manière aléatoire à chaque groupe de la campagne. En d'autres termes, aucune affectation selon des caractéristiques démographiques, d'habitude d'achat ou autres n'a été effectuée et la probabilité d'être affecté à l'un des groupes est identique pour tous les contacts.

Pour obtenir un test de package de contrôle

A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique

- Sélectionnez Comparer l'efficacité des campagnes.

Figure 7-2
Boîte de dialogue Test de package de contrôle

- Sélectionnez le champ qui identifie à quel groupe de campagne chaque contact appartient (par exemple, offre A, offre B, etc.) Ce champ doit être nominal ou ordinal.

- Sélectionnez le champ qui indique l'efficacité des réponses.

Si le champ de réponse est un montant d'achat, il doit être numérique.

Si le champ de réponse indique simplement si la réponse du contact est positive ou négative, (par exemple "Oui" ou "Non"), sélectionnez Répondre et saisissez la valeur qui représente une réponse positive. Si des valeurs disposent d'étiquettes de valeur définies, vous pouvez sélectionner l'étiquette de valeur dans la liste déroulante, et la valeur correspondante sera affichée.

Un nouveau champ est créé automatiquement, dans lequel 1 représente des réponses positives et 0 représente des réponses négatives, et l'analyse est effectuée sur le nouveau champ. Vous pouvez ignorer le nom et l'étiquette par défaut et fournir les vôtres. Les noms doivent être conformes aux règles de dénomination IBM® SPSS® Statistics.

- Cliquez sur Exécuter pour exécuter la procédure.

Partie II: Exemples

Analyse RFM à partir des données de transaction

Dans un fichier de données de transactions, chaque ligne représente une transaction distincte, plutôt qu'un client distinct, et il peut exister plusieurs lignes de transactions pour chaque client. Cet exemple utilise le fichier de données *rfm_transactions.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)

Données de transaction

L'ensemble de données contient les variables qui comportent les informations suivantes :

- Une variable ou une combinaison de variables qui identifient chaque observation (client).
- Une variable avec la date de chaque transaction.
- Une variable avec la valeur monétaire de chaque transaction.

Figure 8-1
Données de transactions RFM

ID	Date	Amount
1	08/04/2005	129
1	10/25/2004	50
1	07/24/2004	118
1	07/24/2004	136
1	09/04/2006	52
2	09/23/2005	183
2	11/05/2004	24
2	11/10/2005	66
2	12/03/2004	77
3	06/04/2005	102
3	05/15/2005	131

Exécution de l'analyse

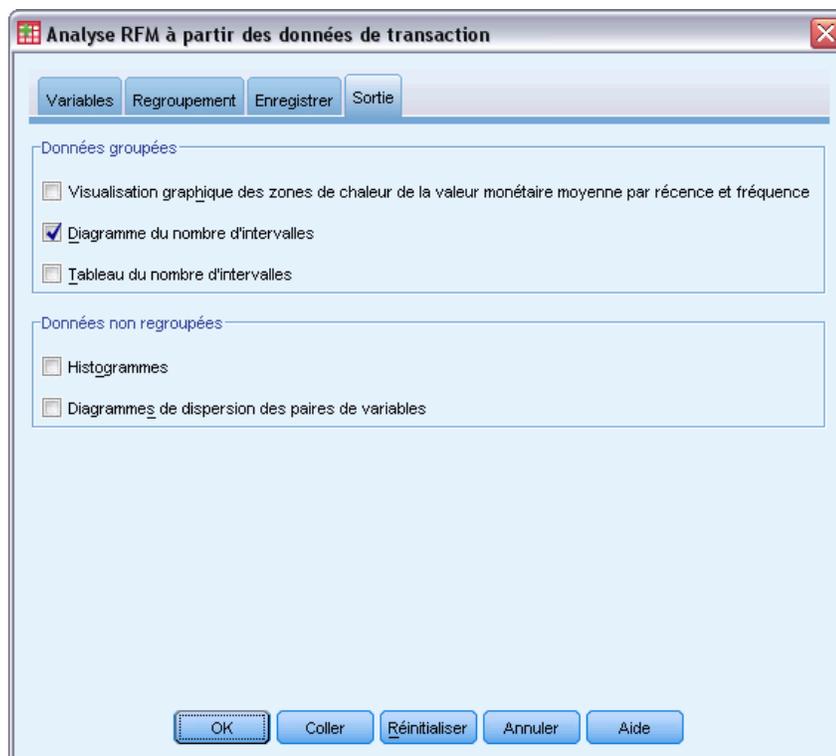
- ▶ Pour calculer les scores RFM, dans les menus, sélectionnez :
Marketing direct > Choisissez la technique
- ▶ Sélectionnez Aider à identifier mes meilleurs contacts (Analyse RFM) puis cliquez sur Poursuivre.
- ▶ Cliquez sur Données de transactions, puis cliquez sur Poursuivre.

Figure 8-2
RFM à partir de Transactions, onglet Variables

The screenshot shows a software dialog box titled "Analyse RFM à partir des données de transaction". It has four tabs: "Variables", "Regroupement", "Enregistrer", and "Sortie". The "Variables" tab is active. At the top left, there is a small grid icon and a diagram showing two rows of data being processed by two people. To the right of this diagram, a text box explains: "Chaque ligne contient des données pour une transaction. Les transactions seront agrégées par client pour l'analyse." Below this, there is a "Variables :" section with a list box containing "Product Line [ProductLine]" and "Product Number [ProductNumber]". To the right of this list box are three arrow buttons pointing right. The first arrow points to a text box labeled "Date de la transaction :" containing "Purchase Date [Date]". The second arrow points to a text box labeled "Total des transactions :" containing "Purchase Amount [Amount]". Below this is a dropdown menu labeled "Méthode récapitulative :" with "Total" selected. The third arrow points to a text box labeled "Identificateurs des clients" containing "Customer ID [ID]". At the bottom of the dialog are five buttons: "OK", "Coller", "Réinitialiser", "Annuler", and "Aide".

- ▶ Cliquez sur Réinitialiser pour effacer les paramètres précédents.
- ▶ Pour la Date de transaction, sélectionnez *Date d'achat [Date]*.
- ▶ Pour le Montant de transaction, sélectionnez *Montant d'achat [Amount]*.
- ▶ Pour la Méthode récapitulative, sélectionnez Total.
- ▶ Pour les Identificateurs client, sélectionnez *ID Client [ID]*.
- ▶ Cliquez ensuite sur l'onglet Résultats.

Figure 8-3
RFM pour Transactions, onglet Résultats



- ▶ Sélectionnez (cochez) Diagramme des effectifs de casiers.
- ▶ Cliquez ensuite sur OK pour exécuter la procédure.

Evaluation des résultats

Lorsque vous calculez les scores RFM à partir des données de transactions, un nouvel ensemble de données incluant les nouveaux scores RFM est créé.

Figure 8-4
RFM à partir d'ensembles de données de Transactions

ID	Date_plus_récent	Compte_transaction	Total	Score_récence	Score_fréquence	Score_monétaire	Score_RFM
1	05/17/2006	10	1313,00	2	3	5	235
2	09/21/2005	11	1230,00	1	5	4	154
3	08/11/2006	13	1194,00	3	5	2	352
4	05/24/2006	9	794,00	2	3	2	232
5	03/13/2005	3	278,00	1	1	2	112
6	07/28/2006	9	922,00	3	2	4	324
7	06/20/2006	11	961,00	2	4	2	242

Par défaut, cet ensemble de données inclut pour chaque client les informations suivantes :

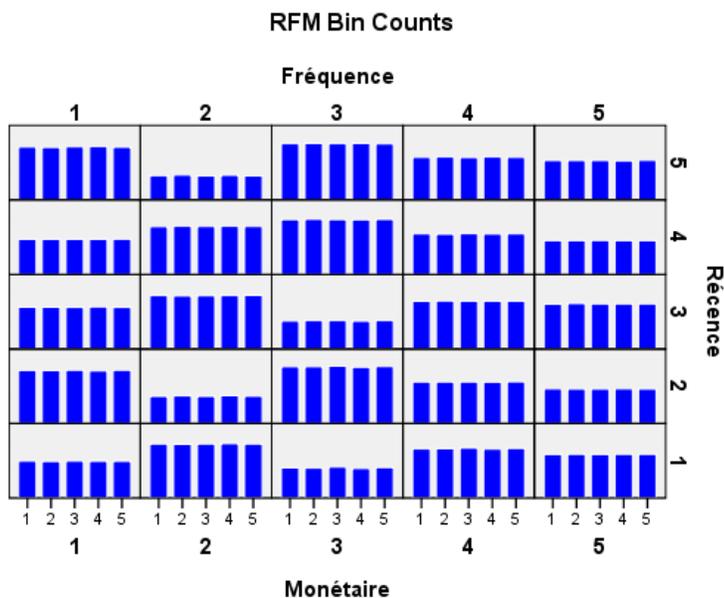
- Variable(s) d'ID client
- Date de la dernière transaction
- Nombre total de transactions.
- Montant de la transaction totale (total par défaut)
- Récence, Fréquence, Monétaire et scores RFM combinés

Le nouvel ensemble de données ne contient qu'une seule ligne (enregistrement) pour chaque client. Les données de transaction d'origine ont été agrégées par des valeurs des variables d'identificateur client. Les variables d'identificateur sont toujours présentes dans le nouvel ensemble de données, sans quoi il vous serait impossible de faire correspondre les scores RFM aux clients.

Le score RFM pour chaque client est simplement la concaténation des trois scores individuels, calculés comme : $(\text{récence} \times 100) + (\text{fréquence} \times 10) + \text{monétaire}$.

Le diagramme des effectifs de casier affiché dans la fenêtre du Viewer montre le nombre de clients pour chaque catégorie RFM.

Figure 8-5
Diagramme du nombre d'intervalles



Le recours à la méthode par défaut de cinq modalités de score pour chacune des trois composantes RFM génère 125 catégories de scores RFM possibles. Chaque bâton du diagramme représente le nombre de clients pour chaque catégorie RFM.

Idéalement, vous souhaiteriez une distribution relativement équitable des clients sur l'ensemble des catégories de scores RFM. En réalité, cela comporte généralement un degré de variation, tel que montré dans cet exemple. S'il existe trop de modalités vides, il se peut que vous deviez songer à changer de méthode de regroupement par casiers.

Il existe un certain nombre de stratégies pour traiter les distributions inéquitables de scores RFM, comme par exemple :

- Utiliser le regroupement par emboîtement plutôt que le regroupement par casiers indépendant.
- Réduire le nombre de catégories de score possible (casiers).
- Lorsqu'il existe un grand nombre de valeurs ex æquo, affecter des observations de façon aléatoire avec les mêmes scores à différentes catégories.

Pour plus d'informations, reportez-vous à la section [Regroupement par casiers RFM](#) dans le chapitre 2 sur p. 6.

Fusionner les données de score avec les données Client

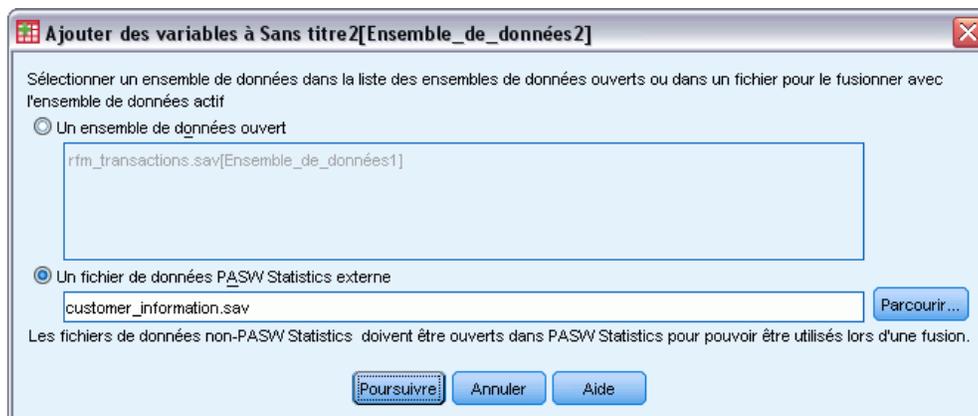
Maintenant que votre ensemble de données contient des scores RFM, vous devez les faire correspondre aux clients. Vous pouvez fusionner les scores avec le fichier de données de transaction, mais d'une manière plus générale, vous souhaiteriez fusionner les données de score avec un fichier de données, qui, tout comme l'ensemble de données de scores RFM, contient une ligne (enregistrement) pour chaque client — et contient également des informations telles que le nom et l'adresse du client.

Figure 8-6
Ensemble de données de scores RFM dans la vue Variable

Nom	Type	Largeur	Décimales	Etiquette	Valeurs
ID	Numérique	5	0	Customer ID	Aucun
Date_plus_récent	Date	10	0	Date de la dernière tr...	Aucun
Compte_transaction	Numérique	7	0	Nombre de transactio...	Aucun
Total	Numérique	8	2	Total	Aucun
Score_récence	Numérique	3	0	Score de récence	Aucun
Score fréquence	Numérique	3	0	Score de fréquence	Aucun
Score_monétaire	Numérique	3	0	Score monétaire	Aucun
Score_RFM	Numérique	3	0	Score RFM	Aucun

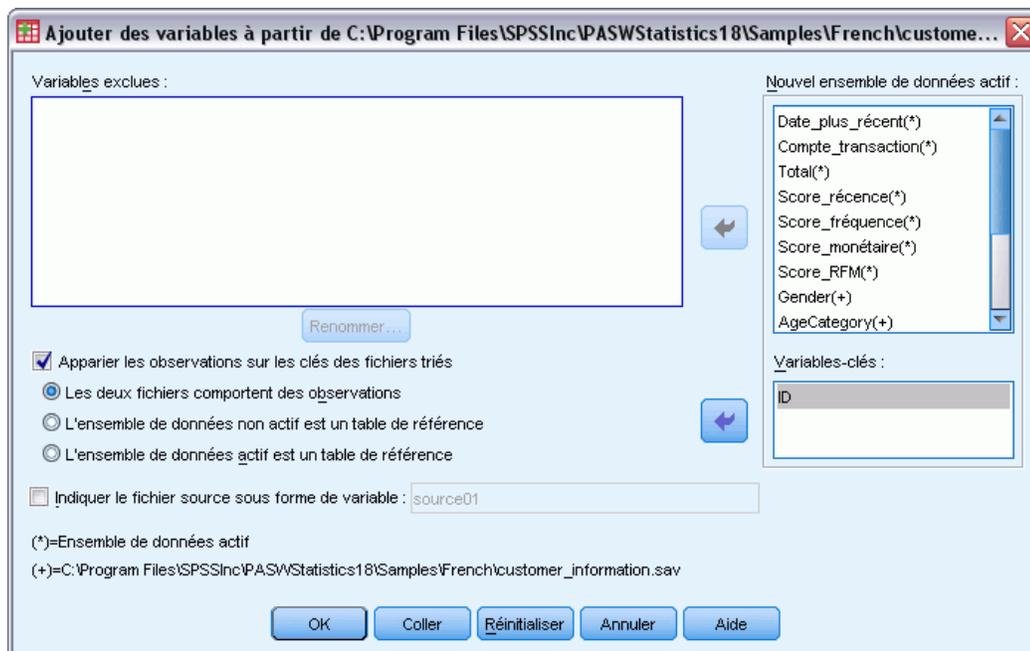
- ▶ Faire de l'ensemble de données qui contient les scores RFM l'ensemble de données actif. (Cliquer n'importe où dans la fenêtre de l'éditeur de données qui contient l'ensemble de données).
- ▶ A partir des menus, sélectionnez :
Données > Fusionner des fichiers > Ajouter des variables

Figure 8-7
Ajouter des variables, boîte de dialogue sélectionner les fichiers



- ▶ Sélectionnez Un fichier de données externe.
- ▶ Utilisez le bouton Parcourir pour naviguer dans le dossier *Echantillons* et sélectionnez *customer_information.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)
- ▶ Cliquez ensuite sur Poursuivre.

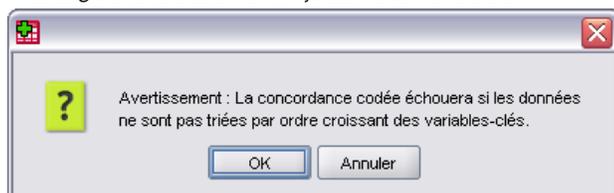
Figure 8-8
Ajouter des variables, boîte de dialogue sélectionner les variables



- ▶ Sélectionnez (cochez) Apparier les observations sur les clés des fichiers triés.
- ▶ Sélectionnez Les deux fichiers comportent des observations.

- ▶ Sélectionnez *ID* pour la liste des Clés d'appariement.
- ▶ Cliquez sur OK.

Figure 8-9
Message d'avertissement Ajouter des variables



Remarque : cet avertissement vous informe que les deux fichiers doivent être triés dans l'ordre croissant des clés d'appariement. Dans cet exemple, les deux fichiers sont déjà triés dans l'ordre croissant de la clé d'appariement, qui est la variable d'identificateur client que nous avons sélectionnée lors du calcul des scores RFM. Lorsque vous calculez des scores RFM à partir de données de transaction, un nouvel ensemble de données sera automatiquement trié dans l'ordre croissant de la ou des variable(s) d'identificateur client. Si vous modifiez l'ordre de tri de l'ensemble de données ou si le fichier de données avec lequel vous souhaitez fusionner l'ensemble de données de scores n'est pas trié dans cet ordre, vous devez d'abord trier les deux fichiers dans l'ordre croissant de la ou des variable(s) d'identificateur client.

- ▶ Cliquez sur OK pour fusionner les deux ensembles de données.

L'ensemble de données contenant les scores RFM contient également le nom et l'adresse du client, ainsi que que d'autres informations.

Figure 8-10
Ensemble de données fusionnés

Nom	Type	Largeur	Décimales	Etiquette	Valeurs
ID	Numérique	5	0	Customer ID	Aucun
Date_plus_récent	Date	10	0	Date de la dernière tr...	Aucun
Compte_transaction	Numérique	7	0	Nombre de transactio...	Aucun
Total	Numérique	8	2	Total	Aucun
Score_récence	Numérique	3	0	Score de récence	Aucun
Score_fréquence	Numérique	3	0	Score de fréquence	Aucun
Score_monétaire	Numérique	3	0	Score monétaire	Aucun
Score_RFM	Numérique	3	0	Score RFM	Aucun
TotalAmount	Numérique	8	2	Total Amount Purcha...	Aucun
MostRecent	Date	10	0	Date of Most Recent ...	Aucun
NumberOfPurchases	Numérique	7	0	Total Number or Purc...	Aucun
PurchaseInterval	Numérique	8	2	Days Since Last Pur...	Aucun

Classification

La classification est un outil d'exploration conçu pour révéler des groupements naturels (ou classes) au sein de données. Par exemple, elle permet d'identifier différents groupes de clients en fonction de diverses caractéristiques démographiques et d'achat.

Par exemple, le service de marketing direct d'une société souhaite identifier des regroupements démographiques dans sa base de données clients afin d'aider à la détermination de stratégies de campagnes de marketing et au développement de nouvelles offres de produits.

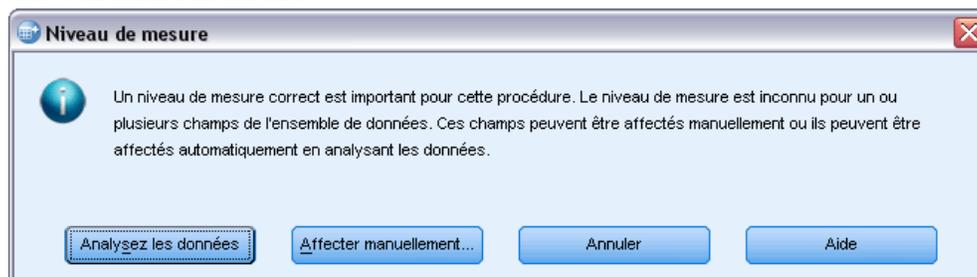
Ces informations sont regroupées dans le fichier *dmdata.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)

Exécution de l'analyse

- ▶ Pour exécuter une classification, à partir des menus, sélectionnez :
Marketing direct > Choisissez la technique
- ▶ Sélectionnez Segmenter mes contacts en classes et cliquez sur Continuer.

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 9-1
Alerte du niveau de mesure

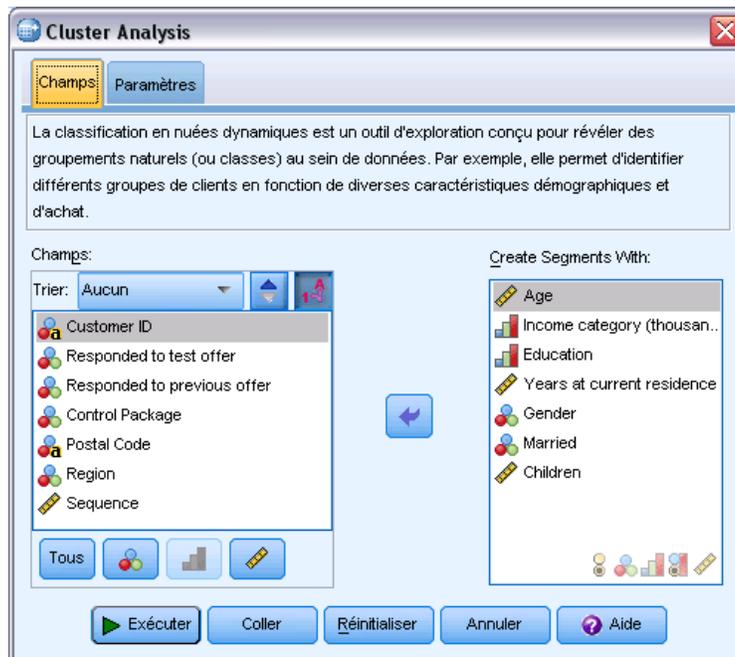


- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Dans ce fichier d'exemple, il n'existe pas de champs avec niveau de mesure inconnu et tous les champs ont le bon niveau de mesure ; par conséquent, aucune alerte de niveau de mesure ne doit apparaître.

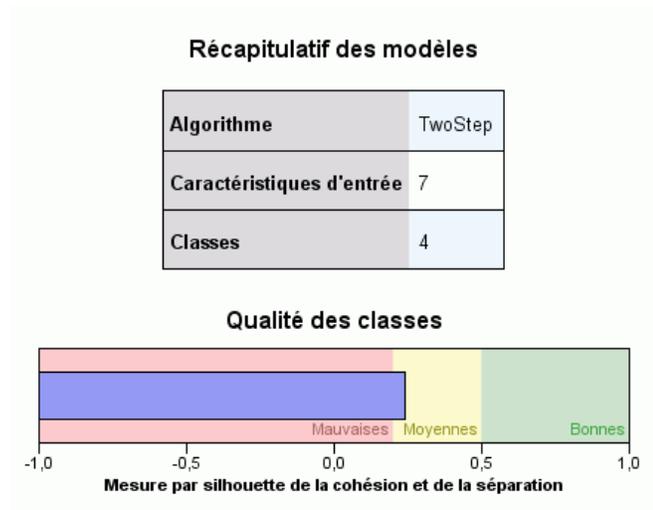
Figure 9-2
Classification, onglet Champs



- ▶ Sélectionnez les champs suivants pour créer des segments : *Âge, Catégorie de revenus, Education, Nombre d'années à l'adresse actuelle, Sexe, Marié et Enfants.*
- ▶ Cliquez sur Exécuter pour exécuter la procédure.

Résultats

Figure 9-3
Récapitulatif du modèle de classe



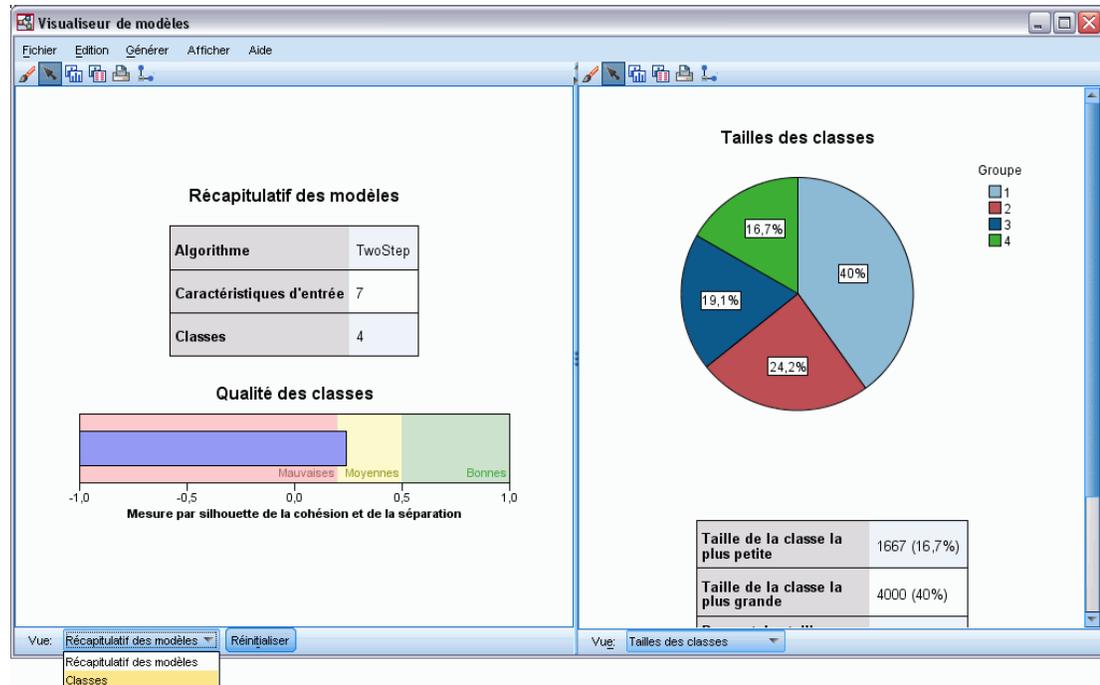
Les résultats sont affichés dans le Viewer de modèles de classe.

- Le récapitulatif du modèle indique que quatre classes ont été trouvées en fonction des sept caractéristiques (champs) d'entrée que vous avez sélectionnées.
- Le diagramme de qualité de classe indique que la qualité générale du modèle se trouve au centre de la plage « moyenne ».

- Double-cliquez sur la sortie du Viewer de modèles de classe pour activer le Viewer de modèle.

Figure 9-4

Viewer de modèles de classe activé



- Dans la liste déroulante Afficher, en bas de la fenêtre du Viewer de modèles de classe, sélectionnez Classes.

Figure 9-5
Affichage des classes

Groupe	1	2	3	4
Etiquette				
Description				
Taille	 40,0% (4000)	 24,2% (2424)	 19,1% (1909)	 16,7% (1667)
Caractéristiques	Age 50,30	Age 44,07	Age 39,05	Age 33,09
	Children 1,58	Children 1,29	Children 0,39	Children 0,12
	Gender Male (57,0%)	Gender Female (100,0%)	Gender Male (100,0%)	Gender Female (50,9%)
	Income category (thousands) 75+ (56,1%)	Income category (thousands) 50-74 (47,2%)	Income category (thousands) 75+ (34,8%)	Income category (thousands) <25 (100,0%)
	Married Yes (100,0%)	Married No (78,5%)	Married No (100,0%)	Married No (78,5%)
	Education Post-graduate (20,5%)	Education Post-graduate (20,5%)	Education College (21,1%)	Education Post-graduate (20,6%)
	Years at current residence 9,47	Years at current residence 9,51	Years at current residence 9,47	Years at current residence 9,42

La vue des classes affiche des informations sur les attributs de chaque classe.

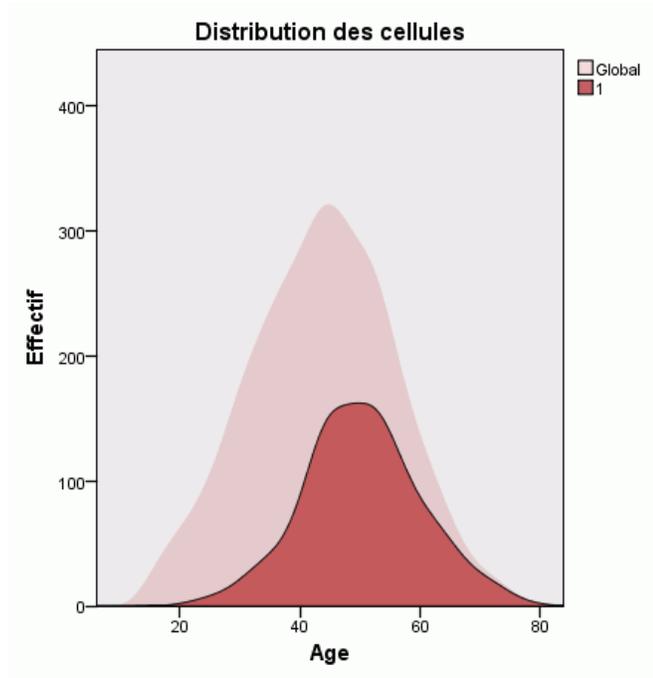
- Pour des champs continus (échelle), la valeur moyenne (moyenne) est affichée.
- Pour des champs qualitatifs (nominaux, ordinaux), le mode est affiché. Le mode est la catégorie comportant le plus grand nombre d'enregistrements. Dans cet exemple, chaque enregistrement est un client.
- Par défaut, les champs sont affichés dans l'ordre de leur importance générale par rapport au modèle. Dans cet exemple, *Âge* a la plus grande importance générale. Vous pouvez aussi trier des champs par importance au sein d'une classe ou par ordre alphabétique.

Si vous sélectionnez (cliquez) sur n'importe quelle cellule de la vue des classes, vous pouvez voir un diagramme qui résume les valeurs de ce champ pour cette classe.

- Par exemple, sélectionnez la cellule *Âge* pour la classe 1.

Figure 9-6

Histogramme de l'âge pour la classe 1

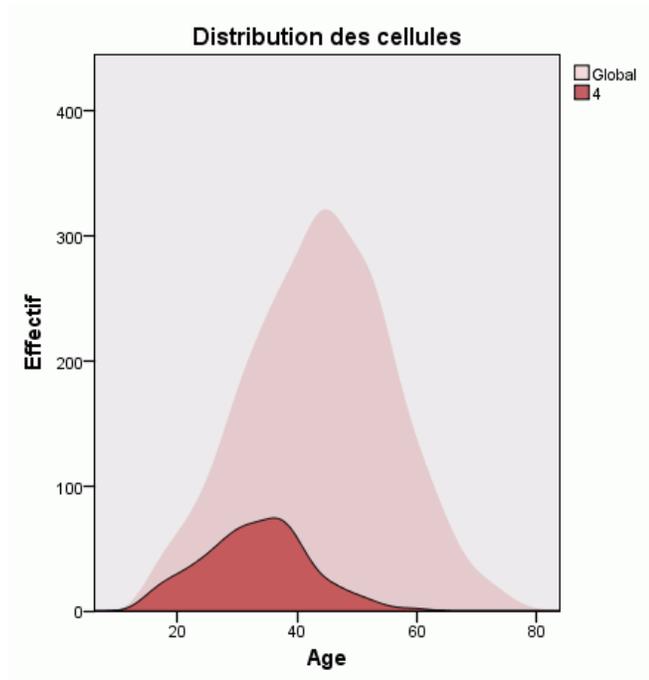


Pour des champs continus, un histogramme s'affiche. L'historgramme affiche la distribution des valeurs dans cette classe et la distribution générale des valeurs pour ce champ. L'historgramme indique que les clients de la classe 1 ont tendance à être un peu plus âgés.

- Sélectionnez la cellule *Âge* pour la classe 4 dans la vue des classes.

Figure 9-7

Histogramme de l'âge pour la classe 4

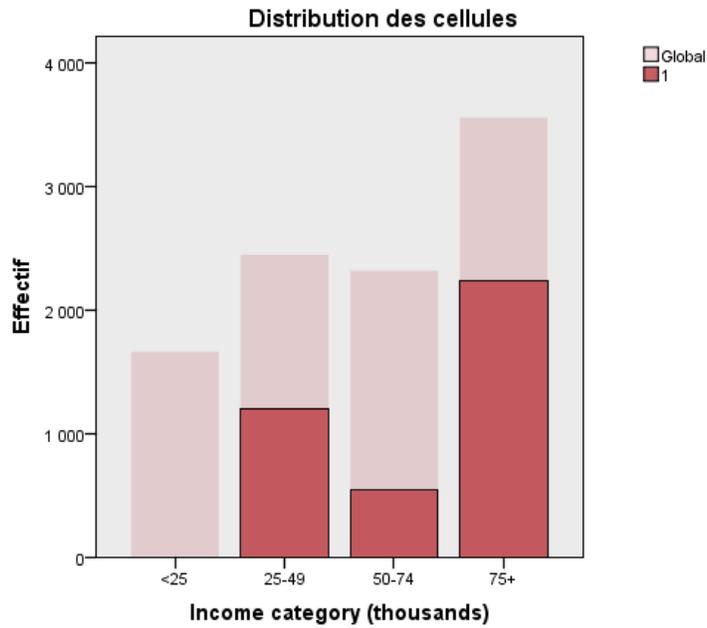


Par contraste avec la classe 1, les clients de la classe 4 ont tendance à être plus jeunes que la moyenne générale.

- Sélectionnez la cellule *Catégorie de revenus* pour la classe 1 dans la vue des classes.

Figure 9-8

Diagramme en bâtons de la catégorie de revenus pour la classe 1

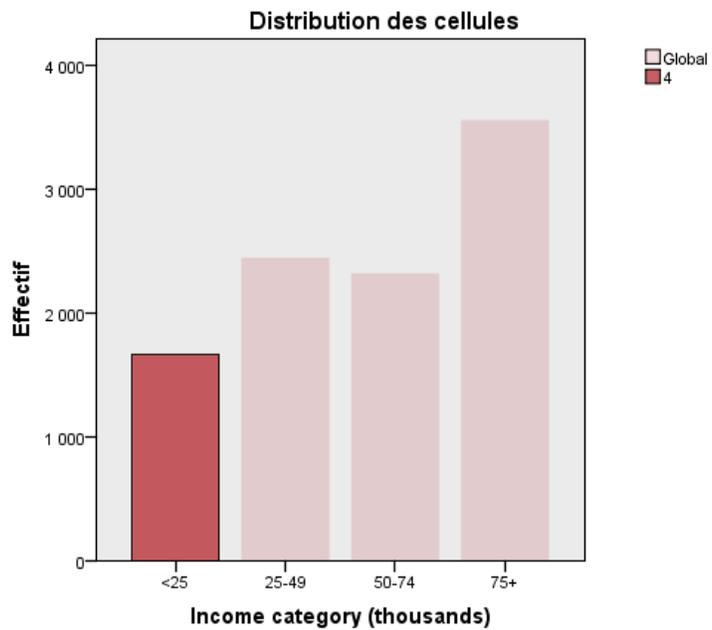


Pour des champs qualitatifs, un diagramme en bâtons s'affiche. La caractéristique la plus notable du diagramme en bâtons de la catégorie des revenus pour cette classe est l'absence totale de tout client dans la catégorie des revenus les plus bas.

- Sélectionnez la cellule *Catégorie de revenus* pour la classe 4 dans la vue des classes.

Figure 9-9

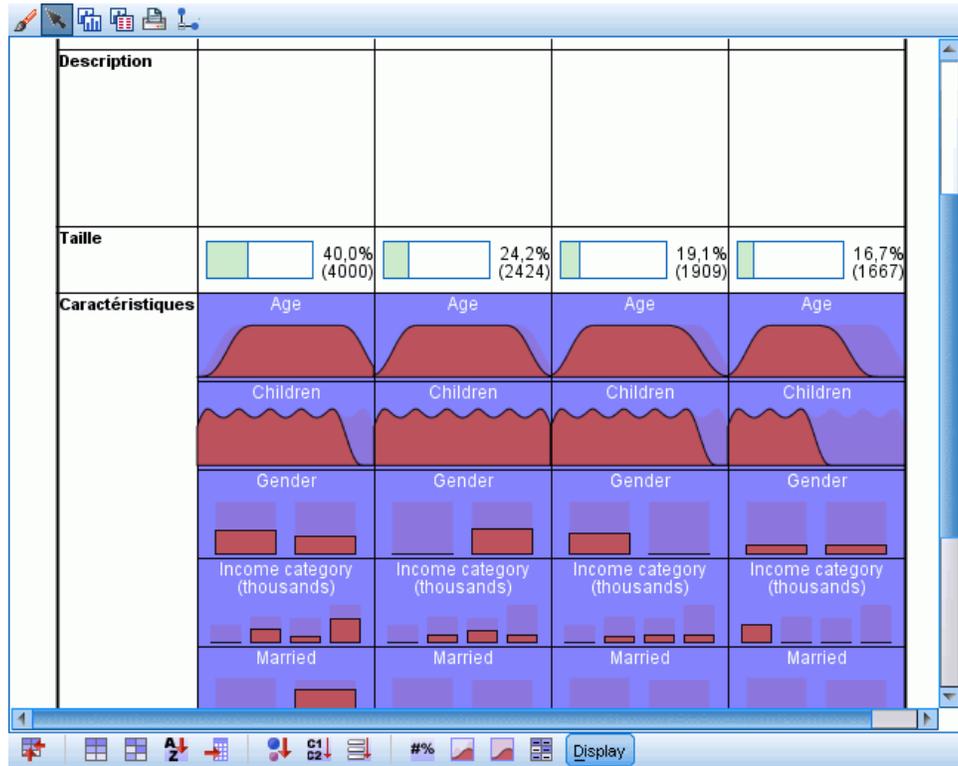
Diagramme en bâtons de la catégorie de revenus pour la classe 4



Par contraste avec la classe 1, tous les clients de la classe 4 font partie de la catégorie des revenus les plus bas.

En utilisant la barre d'outils qui se trouve en bas de la fenêtre du Viewer de modèles, vous pouvez aussi modifier la vue des classes afin d'afficher des diagrammes dans les cellules, ce qui facilite la comparaison rapide des distributions des valeurs entre les classes.

Figure 9-10
Diagrammes affichés dans la classe



En regardant la vue des classes et les informations supplémentaires fournies par les diagrammes de chaque cellule, vous pouvez voir quelques différences entre les classes :

- Les clients de la classe 1 tendent à être des personnes plus âgées, mariées, avec des enfants et des revenus supérieurs.
- Les clients de la classe 2 ont quelque peu tendance à être des mères célibataires avec des revenus modérés.
- Les clients de la classe 3 tendent à être de jeunes hommes célibataires sans enfants.
- Les clients de la classe 4 tendent à être de jeunes femmes célibataires sans enfants avec des revenus inférieurs.

Les cellules Description de la vue des classes sont des champs de texte que vous pouvez modifier afin d'ajouter les descriptions de chaque classe.

Figure 9-11

Vue des classes avec les descriptions de classe

Groupe	1	2	3	4
Étiquette				
Description	Older, married, have children, higher income	Older single mothers, moderate income	Younger single men, no children	Younger single women, no children, low income
Taille	 40,0% (4000)	 24,2% (2424)	 19,1% (1909)	 16,7% (1667)
Caractéristiques	Age 50,30	Age 44,07	Age 39,05	Age 33,09
	Children 1,58	Children 1,29	Children 0,39	Children 0,12
	Gender Male (57,0%)	Gender Female (100,0%)	Gender Male (100,0%)	Gender Female (50,9%)
	Income category (thousands) 75+ (56,1%)	Income category (thousands) 50-74 (47,2%)	Income category (thousands) 75+ (34,8%)	Income category (thousands) <25 (100,0%)
	Married Yes (100,0%)	Married No (78,5%)	Married No (100,0%)	Married No (78,5%)
	Education Post-graduate (20,5%)	Education Post-graduate (20,5%)	Education College (21,1%)	Education Post-graduate (20,6%)
	Years at current residence 9,47	Years at current residence 9,51	Years at current residence 9,47	Years at current residence 9,42

Sélection d'enregistrements en fonction des classes

Vous pouvez sélectionner des enregistrements en fonction de l'appartenance d'une classe de deux manières :

- Créez interactivement une condition de filtre dans le Viewer de modèles de classe.
- Utilisez les valeurs du champ de classe générées par la procédure pour spécifier les conditions de filtre ou de sélection.

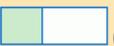
Création d'un filtre dans le Viewer de modèles de classe

Pour créer une condition de filtre qui sélectionne des enregistrements à partir de classes spécifiques dans le Viewer de modèles de classe :

- ▶ Activez (double-cliquez) sur le Viewer de modèles de classe.
- ▶ Dans la liste déroulante Afficher, en bas de la fenêtre du Viewer de modèles de classe, sélectionnez Classes.
- ▶ Cliquez sur le numéro de la classe que vous souhaitez positionner en haut de la vue des classes. Si vous souhaitez sélectionner plusieurs classes, cliquez sur chaque numéro de classe supplémentaire souhaitée tout en maintenant la touche Ctrl enfoncée.

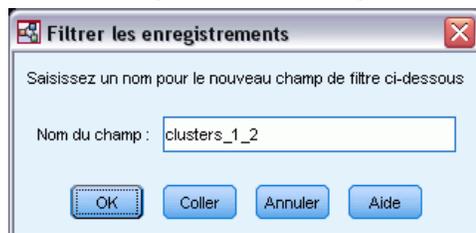
Figure 9-12

Classes sélectionnées dans la vue des classes

Groupe	1	2	3	4
Etiquette				
Description	Older, married, have children, higher income	Older single mothers, moderate income	Younger single men, no children	Younger single women, no children, low income
Taille	 40,0% (4000)	 24,2% (2424)	 19,1% (1909)	 16,7% (1667)
Caractéristiques	Age 50,30	Age 44,07	Age 39,05	Age 33,09
	Children 1,58	Children 1,29	Children 0,39	Children 0,12
	Gender Male (57,0%)	Gender Female (100,0%)	Gender Male (100,0%)	Gender Female (50,9%)
	Income category (thousands) 75+ (56,1%)	Income category (thousands) 50-74 (47,2%)	Income category (thousands) 75+ (34,8%)	Income category (thousands) <25 (100,0%)
	Married Yes (100,0%)	Married No (78,5%)	Married No (100,0%)	Married No (78,5%)
	Education Post-graduate (20,5%)	Education Post-graduate (20,5%)	Education College (21,1%)	Education Post-graduate (20,6%)
	Years at current residence 9,47	Years at current residence 9,51	Years at current residence 9,47	Years at current residence 9,42

- ▶ Dans les menus du Viewer de modèles de classe, sélectionnez :
Générer > Filtrer les enregistrements

Figure 9-13
Boîte de dialogue Filtrer les enregistrements



- Saisissez un nom dans le champ de filtre et cliquez sur OK. Les noms doivent être conformes aux règles de dénomination IBM® SPSS® Statistics.

Figure 9-14
Enregistrements filtrés dans l'éditeur de données

	ID	Married	Children	Region	ClusterGroup1	clusters_1_2
14	03623	No	0	West	3	.00
15	01353	No	0	West	3	.00
16	07055	No	0	West	3	.00
17	04455	No	0	West	2	1.00
18	07210	No	1	West	2	1.00
19	08054	No	0	West	4	.00
20	06937	No	0	West	4	.00
21	06512	No	0	West	4	.00
22	08315	No	0	West	4	.00
23	09676	No	3	West	2	1.00
24	09636	No	0	West	4	.00
25	08579	No	1	West	2	1.00
26	01480	No	1	West	2	1.00

Ceci crée un nouveau champ dans l'ensemble de données et filtre les enregistrements de l'ensemble de données en fonction des valeurs de ce champ.

- Les enregistrements ayant une valeur de 1 pour le champ du filtre seront inclus dans les analyses, les diagrammes et les rapports ultérieurs.
- Les enregistrements ayant une valeur de 0 pour le champ du filtre seront exclus.
- Les enregistrements exclus ne sont pas supprimés de l'ensemble de données. Ils sont conservés avec un indicateur d'état du filtre qui est affiché sous la forme d'une barre diagonale en travers du numéro d'enregistrement dans l'éditeur de données.

Sélection d'enregistrements en fonction des valeurs du champ de la classe

Par défaut, la classification crée un nouveau champ qui identifie le groupe de classe de chaque enregistrement. Le nom par défaut de ce champ est *ClusterGroupn*, où *n* est un entier qui représente un nom de champ unique.

Figure 9-15

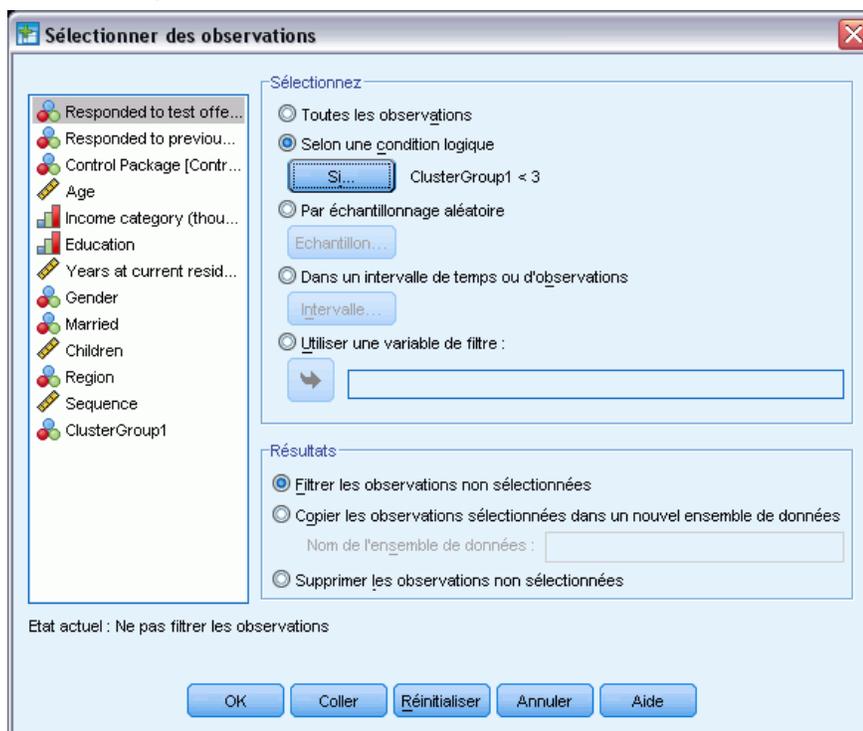
Champ de classe ajouté à un ensemble de données

	ID	Gender	Married	Children	Region	ClusterGroup1
1	01359	Female	No	0	West	4
2	06262	Female	No	1	West	2
3	08031	Male	No	0	West	3
4	01971	Male	No	0	West	4
5	09689	Male	No	0	West	3
6	06108	Male	No	1	West	3
7	09853	Male	No	0	West	3
8	06802	Male	No	0	West	4
9	07597	Male	No	0	West	3
10	03692	Male	No	1	West	3
11	00071	Male	No	0	West	4
12	00769	Male	No	0	West	3

Pour utiliser les valeurs du champ de classe afin de sélectionner des enregistrements dans des classes spécifiques :

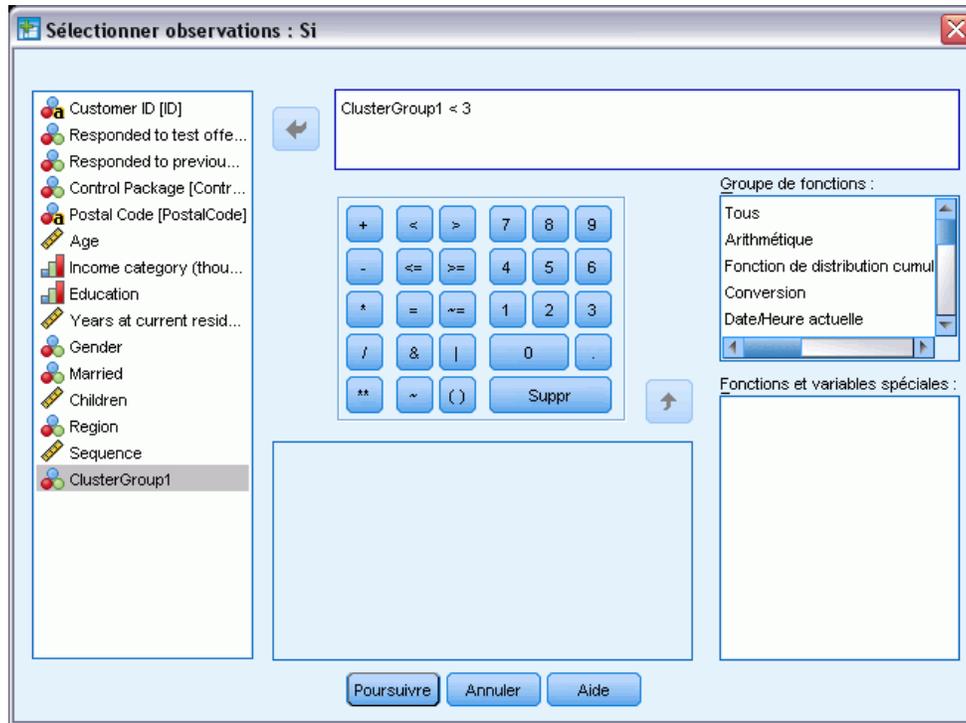
- A partir des menus, sélectionnez :
Données > Sélectionner des observations

Figure 9-16
Boîte de dialogue Sélectionner des observations



- Dans la boîte de dialogue Sélectionner des observations, sélectionnez Selon une condition logique et cliquez sur Si.

Figure 9-17
Sélectionner observations : Boîte de dialogue Si



- Saisissez la condition de sélection.

Par exemple, $\text{ClusterGroup1} < 3$ sélectionne tous les enregistrements dans les classes 1 et 2, et exclut les enregistrements des classes 3 et supérieures.

- Cliquez sur Poursuivre.

Dans la boîte de dialogue Sélectionner des observations, se trouvent plusieurs options concernant ce qu'il convient de faire avec les enregistrements sélectionnés et non sélectionnés :

Filtrez les observations exclues. Ceci crée un nouveau champ qui spécifie une condition de filtre. Les enregistrements exclus ne sont pas supprimés de l'ensemble de données. Ils sont conservés avec un indicateur d'état du filtre qui est affiché sous la forme d'une barre diagonale en travers du numéro d'enregistrement dans l'éditeur de données. Ceci est équivalent à une sélection interactive de classe dans le Viewer de modèles de classe.

Copiez les observations sélectionnées dans l'ensemble de données. Ceci crée un nouvel ensemble de données dans la session actuelle qui ne contient que les enregistrements qui satisfont à la condition du filtre. L'ensemble de données original n'est pas affecté.

Supprimez les observations exclues. Les enregistrements non sélectionnés sont supprimés de l'ensemble de données. Les enregistrements supprimés peuvent être récupérés uniquement en quittant le fichier sans enregistrer aucune modification puis en l'ouvrant à nouveau. La

suppression des observations est définitive si vous enregistrez les modifications apportées au fichier de données.

La boîte de dialogue Sélectionner des observations dispose aussi d'une option permettant d'utiliser une variable existante comme variable de filtre (champ). Si vous créez de manière interactive une condition de filtre dans le Viewer de modèles de classe et que vous enregistrez le champ de filtre généré avec l'ensemble de données, vous pouvez utiliser ce champ afin de filtrer les enregistrements lors de sessions ultérieures.

Récapitulatif

La classification est un outil d'exploration utile qui peut révéler des regroupements naturels (ou classes) au sein de données. Vous pouvez utiliser les informations de ces classes afin de déterminer des stratégies de campagne de marketing et développer de nouvelles offres de produits. Vous pouvez sélectionner des enregistrements en fonction de l'appartenance de classe pour une analyse plus poussée ou pour des campagnes de marketing ciblées.

Profils des clients potentiels

Les profils des clients potentiels utilisent les résultats d'une campagne précédente pour créer des profils descriptifs. Vous pouvez utiliser ces profils pour cibler des groupes de contacts spécifiques lors des futures campagnes. Par exemple, selon les résultats d'un envoi d'essai, la division de marketing direct d'une société souhaite générer les profils des types de personnes les plus susceptibles de répondre à un certain type d'offre, sur la base d'informations démographiques. Selon ces résultats, ils pourront alors déterminer les types de fichiers d'adresses à utiliser pour des offres similaires.

Par exemple, la division de marketing direct d'une société effectue un envoi d'essai à environ 20 % de l'ensemble des clients de sa base de données. Les résultats de cet envoi d'essai seront enregistrés dans un fichier de données qui contient aussi les caractéristiques démographiques de chaque client, y compris l'âge, le sexe, le statut marital et la région géographique. Les résultats sont enregistrés d'une manière simple (oui/non), en indiquant les clients ayant répondu à l'envoi d'essai (ayant effectué un achat) et ceux n'y ayant pas répondu.

Ces informations sont regroupées dans le fichier *dmdata.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)

Analyse des données

Le champ de réponse doit être qualitatif, avec une valeur représentant toutes les réponses positives. Toutes les autres valeurs non manquantes sont supposées indiquer une réponse négative. Si le champ de la réponse représente une valeur continue (échelle), tel qu'un nombre d'achats ou le montant des achats, vous devez créer un nouveau champ qui attribue une valeur unique de réponse positive à toutes les valeurs de réponse non nulles. [Pour plus d'informations, reportez-vous à la section Création d'un champ de réponse qualitatif dans le chapitre 4 sur p. 25.](#)

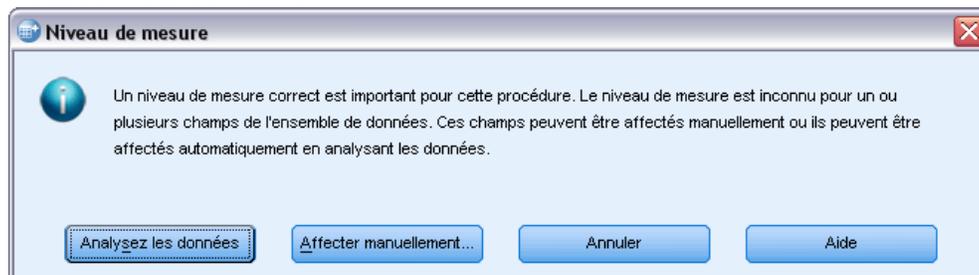
Exécution de l'analyse

- ▶ Pour exécuter une analyse des profils des clients potentiels, sélectionnez à partir des menus :
Marketing direct > Choisissez la technique

- ▶ Sélectionnez Générer les profils de mes contacts ayant répondu à une offre et cliquez sur Continuer.

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 10-1
Alerte du niveau de mesure



- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Dans ce fichier d'exemple, il n'existe pas de champs avec niveau de mesure inconnu et tous les champs ont le bon niveau de mesure ; par conséquent, aucune alerte de niveau de mesure ne doit apparaître.

Figure 10-2

Profils des clients potentiels, onglet Champs



- ▶ Pour le champ Réponse, sélectionnez *A répondu à une offre d'essai*.
- ▶ Comme valeur de réponse positive, sélectionnez *Oui* dans la liste déroulante. La valeur 1 est affichée dans le champ de texte car « Oui » est en fait une étiquette de valeur associée à la valeur enregistrée 1. (Si la valeur de réponse positive ne possède pas d'étiquette de valeur définie, vous pouvez simplement saisir la valeur dans le champ de texte.)
- ▶ Pour Créez des profils avec, sélectionnez *Âge, Catégorie de revenus, Formation, Nombre d'années à l'adresse actuelle, Sexe, Marié, Région* et *Enfants*.
- ▶ Cliquez sur l'onglet Paramètres.

Figure 10-3
Profils des clients potentiels, onglet Paramètres

Profils des clients potentiels

Champs Paramètres

Taille minimale du groupe: 100

Inclure des informations de seuil minimal de taux de réponse dans les résultats.

Specify target response rate (%): 7

Exécuter Coller Réinitialiser Annuler Aide

- ▶ Sélectionnez (cochez) Inclure des informations de seuil minimal de taux de réponse dans les résultats.
- ▶ Pour le taux de réponse cible, saisissez 7.
- ▶ Cliquez ensuite sur OK pour exécuter la procédure.

Résultats

Figure 10-4
Tableau des taux de réponse

Nombre	Description	Profil		
		Taille du groupe	Taux de réponse	Taux de réponses cumulé
1	Region = "West", "South", "East" Gender = "Female" Married = "No"	379	9.2%	9.2%
2	Region = "West", "South", "East" Gender = "Female" Married = "Yes"	299	5.0%	7.4%
3	Region = "West", "South", "East" Gender = "Male"	722	4.7%	6.0%
4	Region = "North"	517	2.5%	5.1%

Vert : Correspond au taux de réponse cible.
Rouge : Ne correspond pas au taux de réponse cible.

Le tableau des taux de réponse affiche des informations pour chaque groupe de profils identifié par la procédure.

- Les profils seront affichés dans l'ordre décroissant du taux de réponse.

- Le taux de réponse est le pourcentage de clients ayant répondu positivement (ayant effectué un achat).
- Le taux de réponse cumulé représente le taux de réponse combiné pour le groupe de profils actuel et tous les groupes précédents. Comme les profils sont affichés dans l'ordre décroissant du taux de réponse, cela signifie que le taux de réponse cumulé est un taux de réponse combiné pour le groupe de profils actuel plus tous les groupes de profils ayant un taux de réponse supérieur.
- La description de profils inclut les caractéristiques pour les seuls champs fournissant une contribution significative au modèle. Dans cet exemple, la région, le sexe et le statut marital sont inclus dans le modèle. Les champs restants — âge, revenu, formation et nombre d'années à l'adresse actuelle — ne sont pas inclus car ils n'apportent pas de contribution significative au modèle.
- La zone verte du tableau représente le groupe de profils ayant un taux de réponse cumulé égal ou supérieur au taux de réponse cible spécifié, qui est de 7 % dans cet exemple.
- La zone rouge du tableau représente le groupe de profils ayant un taux de réponse cumulé inférieur au taux de réponse cible spécifié.
- Le taux de réponse cumulé de la dernière ligne du tableau correspond au taux de réponse général ou moyen de tous les clients inclus dans l'envoi d'essai, puisqu'il s'agit du taux de réponse de tous les groupes de profils.

Les résultats affichés dans le tableau suggèrent que, si vous ciblez les femmes habitant à l'ouest, au sud et à l'est, vous devriez obtenir un taux de réponse légèrement supérieur au taux de réponse cible.

Remarquez cependant qu'il existe une différence importante entre les taux de réponse des femmes célibataires (9,2 %) et des femmes mariées (5,0 %) dans ces régions. Bien que le taux de réponse cumulé des deux groupes soit supérieur au taux de réponse cible, le taux de réponse de ce seul dernier groupe est, en fait, inférieur au taux de réponse cible, ce qui suggère que vous devriez rechercher d'autres caractéristiques qui pourraient améliorer le modèle.

Résultats intelligents

Figure 10-5
Résultats intelligents

Le taux de réponse le tableau affiche des informations pour chaque groupe de profils identifié par la procédure. La description de profils inclut les caractéristiques pour les seuls champs fournissant une contribution significative au modèle. Les champs n'apportant pas une contribution significative ne sont pas inclus.

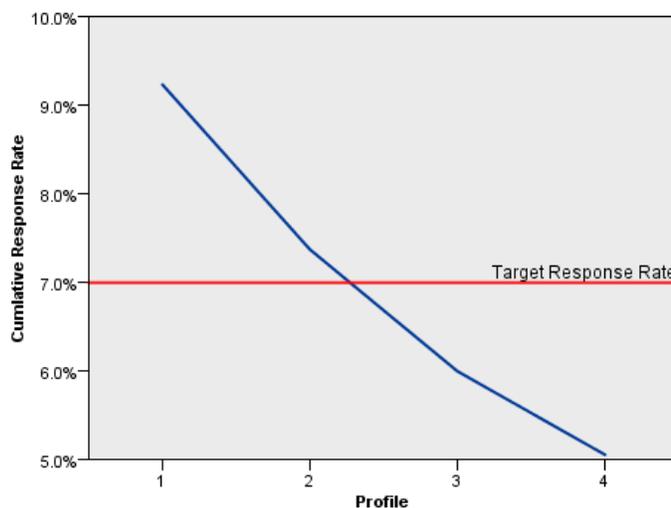
Les profils sont affichés dans l'ordre décroissant du taux de réponse. Le taux de réponse est le pourcentage de clients ayant répondu positivement (ayant effectué un achat).

Le taux de réponse cumulé est le taux de réponse combiné pour le groupe de profils actuel et tous les groupes de profils précédents. Comme les profils sont affichés dans l'ordre décroissant du taux de réponse, cela signifie que le taux de réponse cumulé est un taux de réponse combiné pour le groupe de profils actuel plus tous les groupes de profils ayant un taux de réponse supérieur.

Le taux de réponse cible spécifié est 7.00%. Les lignes vertes ont un taux de réponse cumulé supérieur à 7.00%, et les lignes rouges un taux de réponse cumulé inférieur à 7.00%. Bien que certains groupes de profils de la zone verte puissent présenter des taux de réponse inférieurs à 7.00%, le taux de réponse cumulé est toujours supérieur à 7.00%.

Le tableau est accompagné des « résultats intelligents » qui fournissent des informations générales sur la façon d'interpréter le tableau et des informations spécifiques sur les résultats du tableau.

Figure 10-6
Diagramme du taux de réponse cumulé



Le diagramme du taux de réponse cumulé est essentiellement une représentation visuelle des taux de réponse cumulés affichés dans le tableau. Comme les profils sont rapportés dans l'ordre décroissant du taux des réponses, la ligne du taux de réponse cumulé est toujours décroissante pour chaque profil ultérieur. Tout comme pour le tableau, le diagramme indique que le taux de réponse cumulé descend sous le taux de réponse cible entre les groupes de profils 2 et 3.

Récapitulatif

Pour cet envoi d'essai particulier, quatre groupes de profils ont été identifiés, et les résultats indiquent que les seules caractéristiques démographiques significatives qui semblent en relation avec le fait qu'une personne ait répondu ou non à l'offre sont le sexe, la région et le statut marital. Le groupe ayant le taux de réponse le plus élevé se compose des femmes célibataires vivant au sud, à l'est et à l'ouest. Après cela, les taux de réponse chutent rapidement, bien que l'inclusion des femmes mariées des mêmes régions produise encore un taux de réponse cumulé supérieur au taux de réponse cible.

Taux de réponse des codes postaux

Cette technique utilise les résultats de la campagne précédente pour calculer les taux de réponse par code postal. Ces taux peuvent être utilisés pour cibler des codes postaux spécifiques lors des futures campagnes.

Par exemple, en fonction des résultats d'un précédent publipostage, la division marketing direct d'une entreprise génère des taux de réponse par code postal. Sur la base de plusieurs critères, tels que le taux de réponse minimum acceptable et/ou le nombre maximum de contacts à inclure dans le mailing, il est ensuite possible de cibler des codes postaux spécifiques.

Ces informations sont regroupées dans le fichier *dmdata.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)

Analyse des données

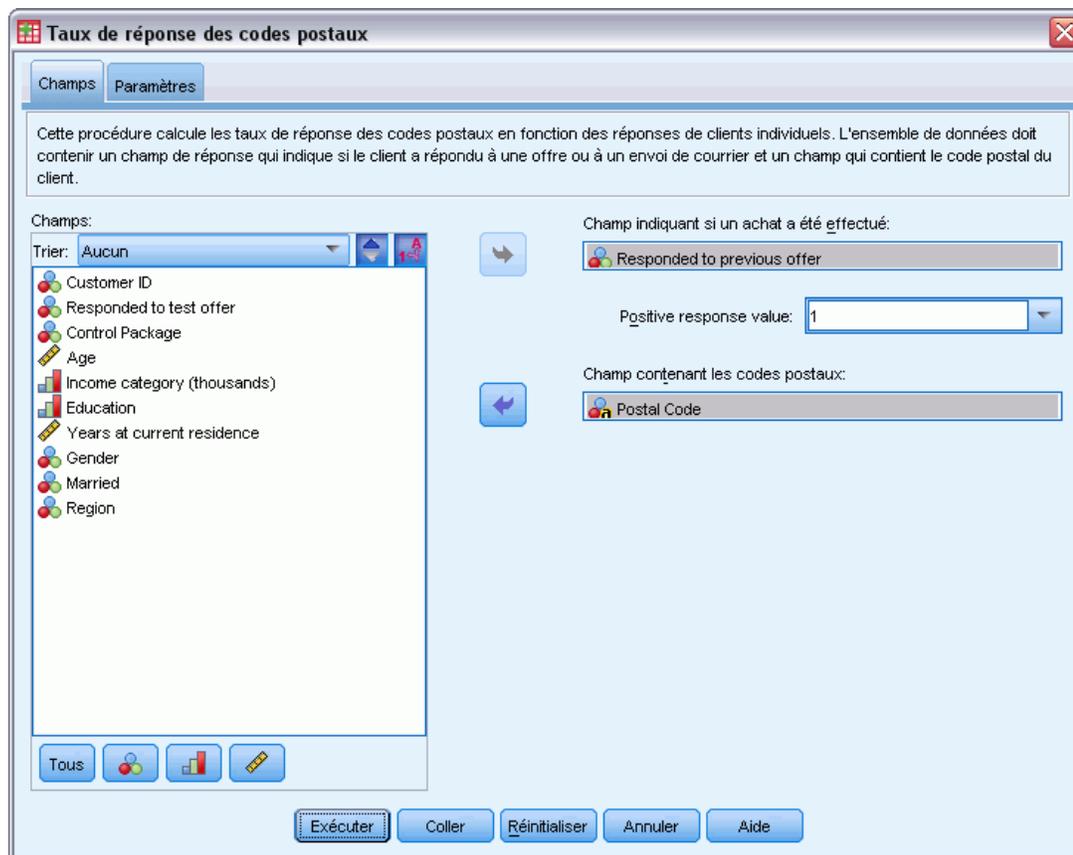
Le champ de réponse doit être qualitatif, avec une valeur représentant toutes les réponses positives. Toutes les autres valeurs non manquantes sont supposées indiquer une réponse négative. Si le champ de réponse représente une valeur continue (échelle), telle que le nombre d'achats ou leur valeur monétaire, vous devez créer un nouveau champ qui assigne une valeur de réponse positive unique à toutes les réponses différentes de zéro. [Pour plus d'informations, reportez-vous à la section Création d'un champ de réponse qualitatif dans le chapitre 5 sur p. 33.](#)

Exécution de l'analyse

- ▶ Pour calculer les taux de réponse des codes postaux, dans les menus sélectionnez :
Marketing direct > Choisissez la technique

- ▶ Sélectionnez Identifier les codes postaux avec les meilleurs taux de réponse et cliquez sur Poursuivre.

Figure 11-1
Onglet Champs - Taux de réponse des codes postaux



- ▶ Comme champ de réponse, sélectionnez *Réponse à une offre précédente*.
- ▶ Comme valeur de réponse positive, sélectionnez *Oui* dans la liste déroulante. La valeur 1 est affichée dans le champ de texte car « Oui » est en fait une étiquette de valeur associée à la valeur enregistrée 1. (Si la valeur de réponse positive ne possède pas d'étiquette de valeur définie, vous pouvez simplement saisir la valeur dans le champ de texte.)
- ▶ Comme Champ de code postal, sélectionnez *Code postal*.
- ▶ Cliquez sur l'onglet Paramètres.

Figure 11-2
Onglet Paramètres - Taux de réponse des codes postaux

The screenshot shows a dialog box titled "Taux de réponse des codes postaux" with a close button (X) in the top right corner. The dialog is divided into several sections:

- Champs / Paramètres:**
 - Grouper les codes postaux selon:** Radio buttons for "Totale", "First 3 digits or characters" (selected), "First 5 digits or characters", and "First N digits or characters". Below is a text input field "N:".
 - format numérique du code postal:** Radio buttons for "3 chiffres", "5 chiffres", "9 chiffres", and "Autre" (selected). Below is a text input field "Nombre de chiffres:".
- Résultat:**
 - "Taux de réponse et analyse de capacité"
 - Taux de réponse:** Radio buttons for "Calculer le taux de réponse moyen à partir des données", "Taux de réponse cible (%)" (selected), and "Calculer le taux de rentabilité à partir de la formule". Below are text input fields for "Coût de l'envoi d'un package:" and "Coût par réponse:".
 - Nombre maximal de contacts:** Radio buttons for "Tous les contacts", "Pourcentage de contacts" (with a text input field), and "Nombre de contacts" (selected, with a text input field containing "5000").
- Exporter dans Excel:** "Enregistrer les taux de réponse des codes postaux dans Excel". Below is a text input field "Nom du fichier:" and a "Parcourir..." button.

At the bottom of the dialog are five buttons: "Exécuter" (highlighted with a dashed border), "Coller", "Réinitialiser", "Annuler", and "Aide".

- ▶ Dans le groupe « Grouper les codes postaux selon », sélectionnez Les 3 premiers chiffres ou caractères. Ceci calcule les taux de réponse combinés pour tous les contacts qui ont des codes postaux commençant par les 3 mêmes premiers chiffres ou caractères. Par exemple, les 3 premiers chiffres ou caractères d'un code postal aux Etats-Unis représentent une même région qui est plus étendue que la région géographique définie par le code à 5 chiffres.
- ▶ Dans le groupe Résultats, cochez l'option Taux de réponse et analyse de capacité.
- ▶ Sélectionnez Taux de réponse cible et saisissez une valeur de 5.
- ▶ Sélectionnez Nombre de contacts et saisissez une valeur de 5000.
- ▶ Cliquez ensuite sur OK pour exécuter la procédure.

Résultats

Figure 11-3

Nouvel ensemble de données avec taux de réponse par code postal

The screenshot shows the PASW Statistics software interface. The main window displays a data table with the following columns: PostalCode, ResponseRate, Responses, Contacts, Index, and Rank. The data is sorted by ResponseRate in descending order. The first row (PostalCode 932) is highlighted in yellow. The status bar at the bottom indicates 'Le processeur PASW Statistics est prêt'.

	PostalCode	ResponseRate	Responses	Contacts	Index	Rank
1	932	10.00%	4.00	40	3.60	1
2	098	8.82%	6.00	68	5.47	1
3	740	7.76%	9.00	116	8.30	1
4	100	7.69%	7.00	91	6.46	1
5	110	7.69%	5.00	65	4.62	1
6	954	7.55%	4.00	53	3.70	1
7	108	7.32%	6.00	82	5.56	1
8	107	7.04%	5.00	71	4.65	1
9	090	6.90%	4.00	58	3.72	1
10	966	6.90%	4.00	58	3.72	1
11	760	6.72%	8.00	119	7.46	1
12	113	6.25%	5.00	80	4.69	1

Un nouvel ensemble de données est créé automatiquement. Ce nouvel ensemble de données ne contient qu'un seul enregistrement (ligne) pour chaque code postal. Dans cet exemple, chaque ligne contient des informations récapitulatives sur tous les codes postaux commençant par les 3 mêmes premiers chiffres ou caractères.

En plus du champ qui contient le code postal, le nouvel ensemble de données contient les champs suivants :

- **TauxDeRéponse.** Pourcentage de réponses positives dans chaque code postal. Les enregistrements sont triés automatiquement dans l'ordre décroissant des taux de réponses, les codes postaux ayant le taux le plus élevé de réponses apparaissant en haut.
- **Réponses.** Nombre de réponses positives dans chaque code postal.
- **Contacts.** Nombre total de contacts dans chaque code postal qui contiennent une valeur non manquante dans le champ de réponse.
- **Index.** Réponse « pondérée » basée sur la formule $N \times P \times (1-P)$, où N est le nombre de contacts, et P est le taux de réponse exprimé en proportion. Pour deux codes postaux possédant le même taux de réponse, cette formule affecte une valeur d'index supérieure au code postal ayant le plus grand nombre de contacts.
- **Rang.** Rang de décile (premiers 10%, 20% , etc.) des taux de réponse de code postal cumulés en ordre décroissant.

L'option Taux de réponse et analyse de capacité ayant été sélectionnée sur l'onglet Paramètres de la boîte de dialogue Taux de réponse des codes postaux, un tableau récapitulatif des taux de réponse et un diagramme s'affichent dans le Viewer.

Figure 11-4
Tableau des taux de réponse

Percentile	Taux de réponse	Contacts	Cumulative Response Rate	Total Contacts
Top 10%	7.3	1001	7.3	1001
Top 20%	5.3	956	6.3	1957
Top 30%	4.3	1042	5.6	2999
Top 40%	3.5	1127	5.1	4126
Top 50%	3.0	1173	4.6	5299
Top 60%	2.4	914	4.3	6213
Top 70%	2.0	948	4.0	7161
Top 80%	1.7	1095	3.7	8256
Top 90%	1.2	680	3.5	8936
Top 100%	.0	1064	3.1	10000

Vert : Correspond au taux de réponse cible.
Rouge : Ne correspond pas au taux de réponse cible.

Le tableau récapitule les résultats par rang de décile (premiers 10%, 20% , etc.) des taux de réponse dans l'ordre décroissant.

- Le taux de réponse cumulé est le pourcentage combiné des réponses positives de la ligne en cours et des lignes précédentes. Les résultats étant affichés dans l'ordre décroissant des taux de réponse, il s'agit du taux de réponse combiné du décile en cours et de tous les déciles comprenant un taux de réponse supérieur.
- Le tableau peut également faire l'objet d'un codage couleur, en fonction d'un taux de réponse cible et d'un nombre de contacts maximum spécifié par l'utilisateur. Les lignes dont le taux de réponse cumulé est égal ou supérieur à 5% et 5 000 ou moins de contacts cumulés sont colorées en vert. Le codage de couleur est basée sur la valeur seuil atteinte en premier. Dans cet exemple, les deux valeurs sont atteintes dans le même décile.

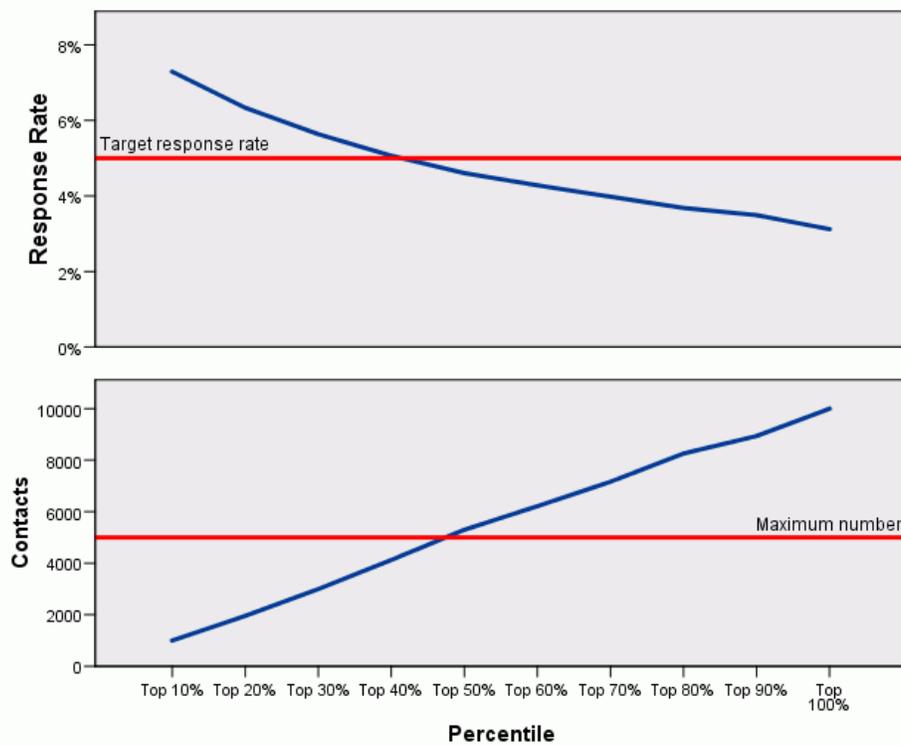
Figure 11-5
Résultats intelligents du tableau de taux de réponse

Le tableau du taux de réponses récapitule les résultats par rang de décile dans l'ordre décroissant (premiers 10%, 20% , etc.). Le taux de réponses cumulé est le pourcentage combiné des réponses positives de la ligne en cours et des lignes précédentes. Les résultats étant affichés dans l'ordre décroissant des taux de réponses, il s'agit du taux de réponse combiné du décile en cours et de tous les déciles comprenant un taux de réponses plus élevé. Le rang des déciles étant inclus dans le nouvel ensemble de données, vous pouvez identifier facilement les codes postaux qui correspondent à un certain taux de réponses cumulé. Le champ du nouvel ensemble de données identifiant le rang des déciles est appelé Rang, où 1=Premiers 10%, 2=Premiers 20%, etc.

Le taux de réponse minimal spécifié est 5.00%. Le nombre maximal de contacts spécifié est 5000. Le codage de couleur du tableau est basé sur la valeur seuil atteinte en premier. Les deux valeurs de seuil sont atteintes dans la même catégorie. Les lignes vertes ont un taux de réponse cumulé supérieur ou égal au taux de réponse minimum spécifié et un nombre de contacts cumulé inférieur ou égal au nombre de contacts maximum spécifié. Les lignes rouges ont un taux de réponse cumulé inférieur au taux de réponse minimum spécifié et un nombre de contacts cumulé supérieur au nombre de contacts maximum spécifié.

Le tableau est accompagné de texte qui fournit une description générale sur la lecture du tableau. Si vous avez spécifié un taux de réponse minimum ou un nombre de contacts maximum, il contient également une section décrivant la façon dont les résultats sont liés aux valeurs seuil spécifiées.

Figure 11-6
Diagramme du taux de réponse cumulé



Le diagramme du taux de réponse cumulé et du nombre de contacts cumulé dans chaque décile est une représentation visuelle des informations affichées dans le tableau du taux de réponse . Le seuil du taux de réponse cumulé minimum et du nombre maximum cumulé de contacts est atteint entre le 40e et le 50e centile.

- Le diagramme affichant les taux de réponse cumulés dans l'ordre décroissant des rangs de déciles des taux de réponses, la ligne du taux de réponse cumulé diminue de plus en plus pour chaque décile.
- La ligne du nombre de contacts représentant le nombre cumulé de contacts, elle augmente au fur et à mesure.

Les informations du tableau et du diagramme vous apprennent que si vous souhaitez atteindre un taux de réponse de 5% au moins sans inclure plus de 5 000 contacts dans la campagne, vous devez vous concentrer sur les codes postaux des quatre déciles supérieurs. Le rang des déciles étant

inclus dans le nouvel ensemble de données, vous pouvez identifier facilement les codes postaux qui correspondent aux 40% supérieurs du tableau.

Figure 11-7
Nouvel ensemble de données

	PostalCode	ResponseRate	Responses	Contacts	Index	Rank
48	120	3.57%	3.00	84	2.89	Top 40%
49	965	3.57%	2.00	56	1.93	Top 40%
50	618	3.54%	4.00	113	3.86	Top 40%
51	603	3.53%	3.00	85	2.89	Top 40%
52	757	3.48%	4.00	115	3.86	Top 40%
53	948	3.39%	2.00	59	1.93	Top 40%
54	103	3.33%	3.00	90	2.90	Top 40%
55	608	3.33%	3.00	90	2.90	Top 40%
56	612	3.28%	4.00	122	3.87	Top 50%
57	762	3.23%	1.00	31	.97	Top 50%
58	933	3.23%	2.00	62	1.94	Top 50%

Remarque : Le *Rang* est enregistré comme une valeur entière comprise entre 1 et 10. Le champ possède des étiquettes de valeurs définies, avec 1= 10% , 2 = 20% supérieurs, etc. Vous pouvez afficher les valeurs des rangs ou les étiquettes des valeurs dans l'affichage des données de l'éditeur de données, en fonction de vos paramètres d'affichage.

Récapitulatif

La procédure des Taux de réponse des codes postaux utilise les résultats d'une campagne précédente pour calculer les taux de réponse par code postal. Ces taux peuvent être utilisés pour cibler des codes postaux spécifiques lors des futures campagnes. La procédure crée automatiquement un nouvel ensemble de données qui contient les taux de réponse de chaque code postal. Sur la base des informations du tableau des taux de réponse, du diagramme et des informations sur les rangs des déciles du nouvel ensemble de données, vous pouvez identifier les codes postaux qui correspondent à un certain taux de réponse cumulé minimum et/ou un certain nombre de contacts cumulé maximum.

Propension à acheter

La propension à acheter utilise les résultats d'un envoi d'essai ou d'une campagne précédente pour générer des scores de propension. Les scores indiquent les contacts qui sont le plus susceptibles de répondre en fonction de diverses caractéristiques sélectionnées.

Cette technique utilise la régression logistique binaire pour construire un modèle prédictif. Le processus de construction et d'application d'un modèle prédictif se fait en deux étapes de base :

- ▶ La construction du modèle et l'enregistrement du fichier du modèle. Vous construisez le modèle en utilisant un ensemble de données pour lequel le résultat qui vous intéresse (souvent appelé la **cible**) est connu. Par exemple, si vous voulez construire un modèle pour prédire qui est susceptible de répondre à une campagne de publipostage, vous devez commencer par un ensemble de données qui indique déjà qui a répondu et qui n'a pas répondu. Par exemple, il pourrait s'agir des résultats d'un envoi d'essai à un petit groupe de clients ou d'informations sur les réponses à une campagne similaire par le passé.
- ▶ L'application de ce modèle à un ensemble de données différent (pour lequel le résultat qui vous intéresse n'est pas connu) pour prévoir les résultats.

Cet exemple utilise deux fichiers de données : *dmdata2.sav* est utilisé pour construire le modèle, puis ce modèle est appliqué à *dmdata3.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)

Analyse des données

Le champ de réponse (la valeur de la cible) doit être qualitatif, avec une valeur représentant toutes les réponses positives. Toutes les autres valeurs non manquantes sont supposées indiquer une réponse négative. Si le champ de la réponse représente une valeur continue (échelle), tel qu'un nombre d'achats ou le montant des achats, vous devez créer un nouveau champ qui attribue une valeur unique de réponse positive à toutes les valeurs de réponse non nulles. [Pour plus d'informations, reportez-vous à la section Création d'un champ de réponse qualitatif dans le chapitre 6 sur p. 41.](#)

Création d'un modèle prédictif

- ▶ Ouvrez le fichier de données *dmdata2.sav*.

Ce fichier contient diverses caractéristiques des personnes ayant reçu l'envoi d'essai ainsi que des informations permettant de savoir si ces personnes ont répondu ou non à l'envoi. Ces informations sont enregistrées dans le champ (de variables) *A répondu*. Une valeur de 1 indique que le contact a répondu à l'envoi ; une valeur de 0 indique que le contact n'a pas répondu.

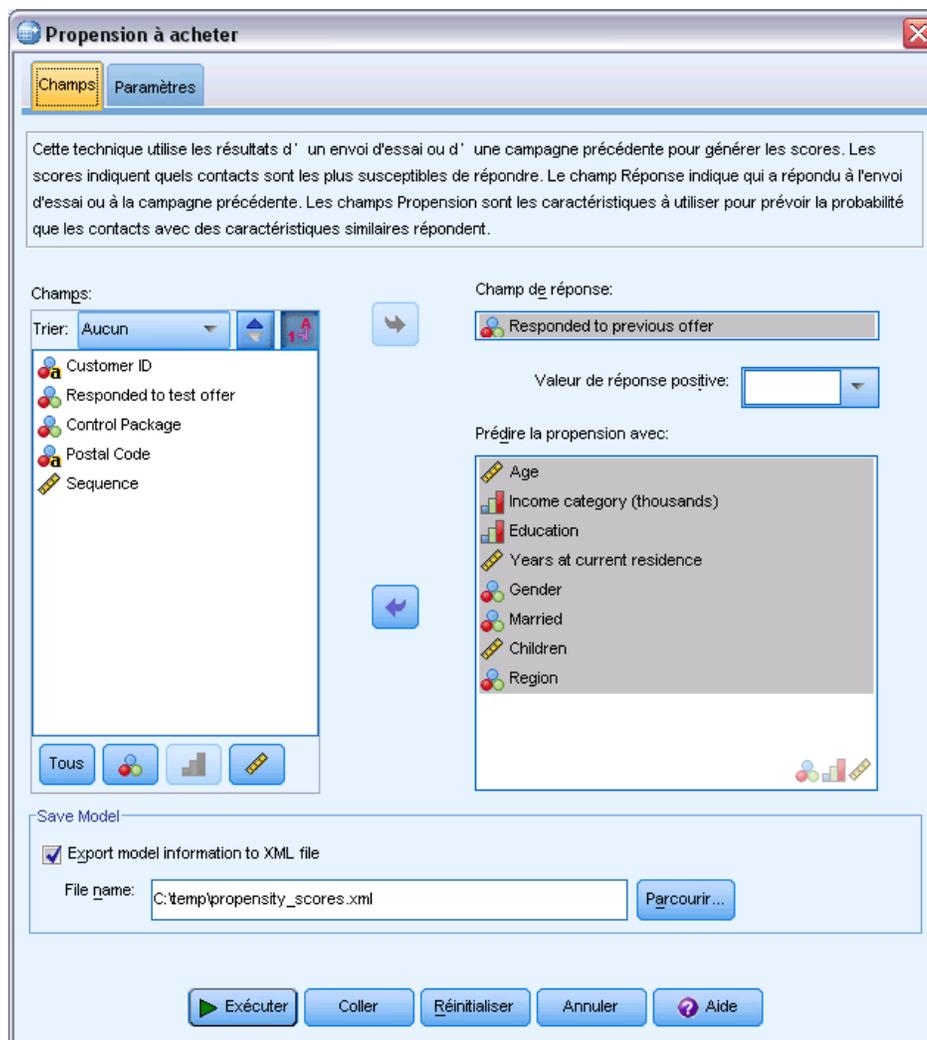
Figure 12-1

Contenu du fichier de données dans l'éditeur de données.

ID	Responded	Previous	ControlPackage	PostalCode	Age	Income	Education	Reside	Gender
03179	0	0	0	93640	38	3	4	11	1
03647	1	0	1	93760	27	2	5	14	1
01741	0	0	1	93850	52	1	5	12	1
05388	0	0	0	93900	66	3	4	10	1
01942	0	0	0	93900	41	4	5	11	1
06254	0	0	1	94120	48	4	4	12	1
02164	0	0	1	94130	46	1	5	9	1
02865	1	0	1	94150	37	4	3	10	1
03330	0	0	1	94270	43	3	2	13	1

- ▶ A partir des menus, sélectionnez :
Marketing direct > Choisissez la technique
- ▶ Sélectionnez Sélectionner les contacts les plus susceptibles d'acheter et cliquez sur Continuer.

Figure 12-2
Propension à acheter - onglet Champs



- ▶ Pour le champ Réponse, sélectionnez *A répondu à une offre d'essai*.
- ▶ Comme valeur de réponse positive, sélectionnez *Oui* dans la liste déroulante. La valeur 1 est affichée dans le champ de texte car « Oui » est en fait une étiquette de valeur associée à la valeur enregistrée 1. (Si la valeur de réponse positive ne possède pas d'étiquette de valeur définie, vous pouvez simplement saisir la valeur dans le champ de texte.)
- ▶ Pour Prédire la propension avec, sélectionnez *Âge*, *Catégorie de revenus*, *Éducation*, *Nombre d'années à l'adresse actuelle*, *Sexe*, *Marié*, *Région* et *Enfants*.
- ▶ Sélectionnez (cochez) Exporter les informations du modèle dans un fichier XML.
- ▶ Cliquez sur Parcourir, pour accéder à l'emplacement où vous souhaitez enregistrer le fichier et entrez un nom pour le fichier.
- ▶ Dans la boîte de dialogue Propension à acheter, cliquez sur l'onglet Paramètres.

Figure 12-3
Propension à acheter - onglet Paramètres

- ▶ Dans le groupe de validation du modèle, sélectionnez (cochez) Valider le modèle et Définir une valeur du générateur pour répliquer les résultats.
- ▶ Utilisez la taille d'échantillon d'apprentissage par défaut de 50 % et la valeur du générateur par défaut de 2 000 000.
- ▶ Dans le groupe Résultat du diagnostic, sélectionnez (cochez) Qualité générale du modèle et Tableau de classification.
- ▶ Pour la probabilité minimale, saisissez 0,05. En règle générale, vous devez indiquer une valeur proche de votre taux de réponse cible minimum, exprimé sous forme de proportion. Une valeur de 0,05 représente un taux de réponse de 5 %.
- ▶ Cliquez sur **Exécuter** pour exécuter la procédure et générer le modèle.

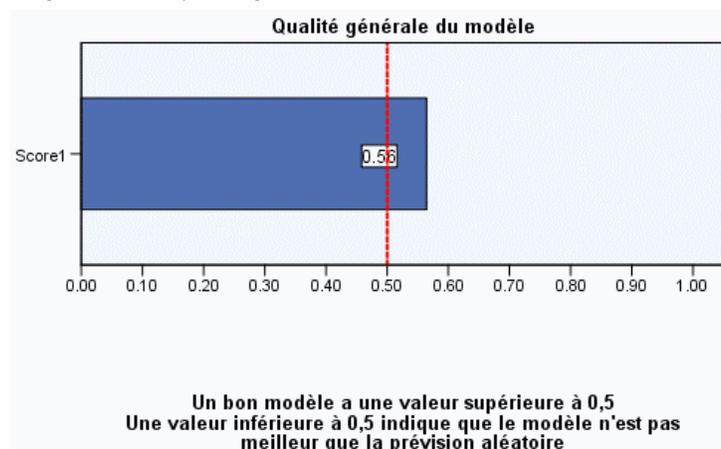
Evaluation du modèle

La Propension à acheter produit un diagramme de qualité générale du modèle et un tableau de classement qui peut être utilisé pour évaluer le modèle.

Le diagramme de qualité générale du modèle fournit une rapide indication visuelle de la qualité du modèle. En règle générale, la qualité générale du modèle doit être supérieure à 0,5.

Figure 12-4

Diagramme de qualité générale du modèle



Pour confirmer que le modèle peut être évalué, vous devez également examiner le tableau de classement.

Figure 12-5

Tableau de classification

		Tableau de classement					
		Prévisions					
		Echantillon d'apprentissage			Echantillon de test		
		Réponse recodée (1=Yes, 0=No)		Pourcentage correct	Réponse recodée (1=Yes, 0=No)		Pourcentage correct
Non	Oui	Non	Oui				
Observations							
Réponse recodée (1=Yes, 0=No)	Non	651	249	72.33	653	267	70.98
	Oui	19	20	51.28	36	22	37.93
Pourcentage global		2.84	7.43	71.46	5.22	7.61	69.02

Le tableau de classement compare les valeurs prédites du champ cible avec les valeurs réelles du champ cible. Le taux de précision globale peut fournir une certaine indication de la performance du modèle, mais le pourcentage de réponses positives prédites correctes peut s'avérer plus intéressant, si l'objectif est de construire un modèle qui identifie le groupe de contacts susceptibles de produire un taux de réponse positive égal ou supérieur au taux de réponse positive minimum spécifié.

Dans cet exemple, le tableau de classement se compose d'un **échantillon d'apprentissage** et d'un **échantillon de test**. L'échantillon de test est utilisé pour construire le modèle. Le modèle est ensuite appliqué à l'échantillon de test pour examiner les performances du modèle.

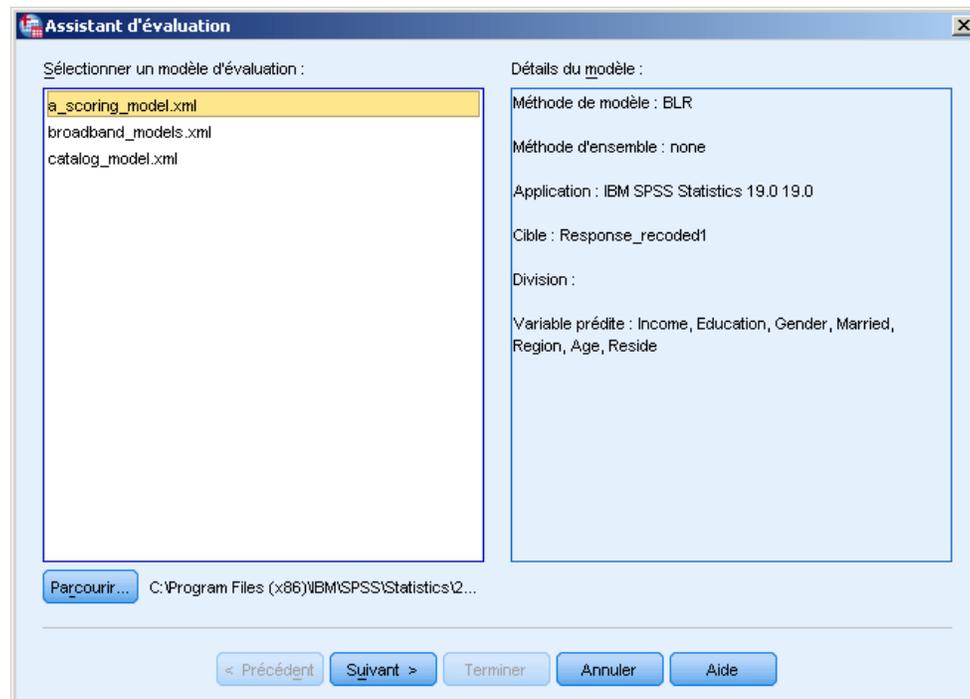
Le taux de réponse positive minimum spécifié était de 0,05 ou 5 %. Le tableau de classement affiche que le taux de classement des réponses positives correct est de 7,43 % dans l'échantillon d'apprentissage et de 7,61 % dans l'échantillon de test. Comme le taux de réponse de l'échantillon de test est supérieur à 5 %, ce modèle peut être capable d'identifier un groupe de contacts susceptibles de produire un taux de réponse supérieur à 5 %.

Application du modèle

- ▶ Ouvrez le fichier de données *dmdata3.sav*. Ce fichier de données contient des informations démographiques et d'autres types d'informations sur tous les contacts qui ne faisaient pas partie de l'envoi d'essai. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)
- ▶ Ouvrez l'assistant d'évaluation. Pour ouvrir l'assistant d'évaluation, à partir des menus, sélectionnez :
Utilitaires > Assistant d'évaluation

Figure 12-6

Assistant d'évaluation, Sélectionnez un modèle d'évaluation.



- ▶ Cliquez sur Parcourir pour vous rendre à l'emplacement où vous avez enregistré le fichier de modèle XML, puis cliquez sur Sélectionner dans la boîte de dialogue de navigation.

Tous les fichiers ayant une extension *.xml* ou *.zip* sont affichés dans l'assistant d'évaluation. Si le fichier sélectionné est reconnu comme un fichier de modèle valide, une description du modèle s'affiche.

- Sélectionnez le fichier de modèle XML que vous avez créé, puis cliquez sur Suivant.

Figure 12-7

Assistant d'évaluation : Faire correspondre les champs du modèle

Nom du modèle : propensity_scores
Type de modèle : Régression logistique binaire

Faire correspondre les champs du modèle à l'ensemble de données :

Champs de l'ensemble de don...	Champs du modèle	Rôle	Mesure	Type
IncomeCategory	Income	Variable prédite	Nominal	Numérique
Education	Education	Variable prédite	Nominal	Numérique
Gender	Gender	Variable prédite	Nominal	Numérique
Married	Married	Variable prédite	Nominal	Numérique
Region	Region	Variable prédite	Nominal	Numérique
Age	Age	Variable prédite	Continu	Numérique
Reside	Reside	Variable prédite	Continu	Numérique

Valeurs manquantes

Utiliser la valeur de substitution

Utiliser les valeurs manquantes par défaut

< Précédent Suivant > Terminer Annuler Aide

Afin de pouvoir être évalué, l'ensemble de données actif doit contenir des champs (variables) correspondant à toutes les valeurs prédites du modèle. Si le modèle contient aussi des champs fractionnés, l'ensemble de données doit aussi contenir des champs correspondant à tous les champs fractionnés du modèle.

- Par défaut, tous les champs de l'ensemble de données actif qui ont le même nom et le même type que les champs du modèle sont automatiquement associés.
- Utilisez la liste déroulante pour associer les champs de l'ensemble de données à ceux du modèle. Pour pouvoir associer les champs, le type de données de chaque champ doit être le même dans le modèle et dans l'ensemble de données.
- Il n'est pas possible de poursuivre l'assistant ou d'évaluer l'ensemble de données actif tant que toutes les valeurs prédites (et champs fractionnés, le cas échéant) du modèle n'ont pas été associées à celles de l'ensemble de données actif.

L'ensemble de données actif ne contient pas de champ nommé *Revenu*. Par conséquent, la cellule de la colonne Champs de l'ensemble de données qui correspond au champ de modèle *Revenu* est initialement vide. Vous devez sélectionner un champ équivalent à ce champ de modèle dans l'ensemble de données actif.

- Dans la liste déroulante de la colonne des champs de l'ensemble de données, dans la cellule vide correspondant à la ligne du champ de modèle *Revenu*, sélectionnez *IncomeCategory* (Catégories de revenus).

Remarque : En plus du nom et du type de champ, vous devez vous assurer que les valeurs de données réelles contenues dans l'ensemble de données en cours d'évaluation, sont bien enregistrées de la même façon que les valeurs de données de l'ensemble de données utilisé pour construire le modèle. Par exemple, si le modèle a été construit avec un champ *Revenu* qui divise le revenu en quatre catégories, et *IncomeCategory* (Catégoriederevenus) dans l'ensemble de données actif qui divise le revenu en six ou quatre catégories différentes, ces champs ne correspondent pas et les résultats d'évaluation ne seront pas fiables.

Cliquez sur **Suivant** pour continuer à l'étape suivante de l'assistant d'évaluation.

Figure 12-8

Assistant d'évaluation : Sélection des fonctions d'évaluation

Nom du modèle : propensity_scores
Type de modèle : BLR

Sélectionner les fonctions d'évaluation
Chaque fonction sélectionnée va créer un nouveau champ dans l'ensemble de données.

	Fonction	Noms de champs	Valeur
<input type="checkbox"/>	Probabilité de la modalité prédite	PredictedProbability	
<input checked="" type="checkbox"/>	Probabilité de la modalité sélectionnée	SelectedProbability	1
<input type="checkbox"/>	Confiance	Confidence	
<input type="checkbox"/>	Valeur prédite	PredictedValue	

< Précédent Suivant > Terminer Annuler Aide

Les fonctions d'évaluation sont les types d'« évaluations » disponibles pour le modèle sélectionné. Les fonctions d'évaluation disponibles dépendent du modèle. Pour le modèle logistique binaire utilisé dans cet exemple, les fonctions disponibles sont valeur prédite, probabilité de la valeur prédite, probabilité d'une valeur sélectionnée et confiance.

Dans cet exemple, nous nous sommes intéressés à la probabilité prédite d'une réponse positive à l'envoi ; nous voulons donc connaître la probabilité d'une valeur sélectionnée.

- ▶ Sélectionnez (cochez) Probabilité de la catégorie sélectionnée.
- ▶ Dans la colonne Valeur, sélectionnez 1 dans la liste déroulante. La liste des valeurs possibles pour la cible est définie dans le modèle, et se base sur les valeurs cibles du fichier de données utilisé pour construire le modèle.

- ▶ Désélectionnez (décochez) toutes les autres fonctions d'évaluation.
- ▶ Vous pouvez aussi attribuer un nom plus descriptif au nouveau champ qui contiendra les valeurs d'évaluation dans l'ensemble de données actif. Par exemple, *Probabilité_de_reponse*.
- ▶ Cliquez sur Terminer pour appliquer le modèle à l'ensemble de données actif.

Le nouveau champ qui contient la probabilité d'une réponse positive est ajouté à la fin de l'ensemble de données.

Figure 12-9

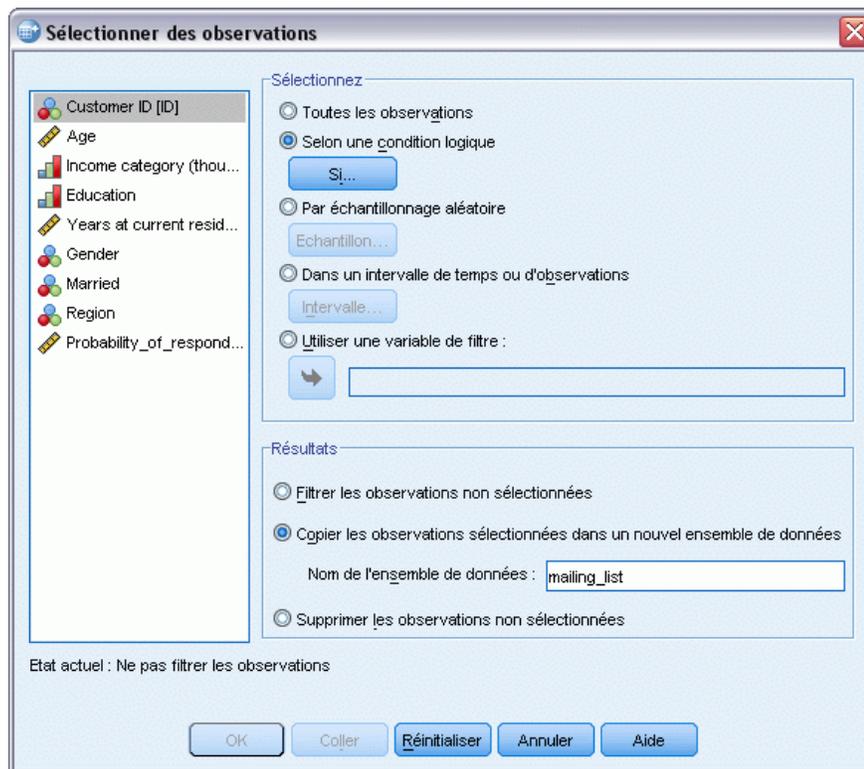
Ensemble de données contenant un nouveau champ de probabilité

Reside	Gender	Married	Region	Probability_of_responding
7	1	0	4	.04
9	0	0	4	.03
12	0	0	4	.03
8	0	0	4	.04
13	0	0	4	.07
10	0	0	4	.04
12	0	0	4	.03
15	0	0	4	.05
10	0	0	4	.05
14	0	0	4	.02
5	0	0	4	.12

Vous pouvez ensuite utiliser ce champ pour sélectionner le sous-ensemble des contacts susceptibles de produire un taux de réponses positives à un certain niveau ou supérieur à ce niveau. Par exemple, vous pouvez créer un nouvel ensemble de données qui contient le sous-ensemble des observations susceptibles de produire un taux de réponses positives d'au moins 5 %.

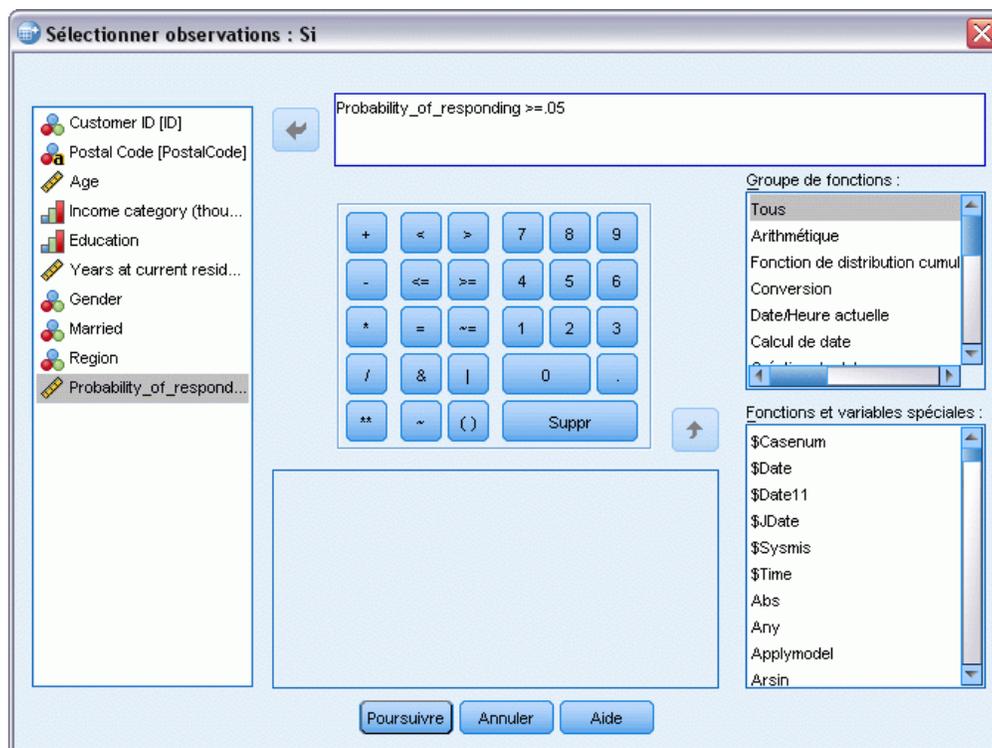
- A partir des menus, sélectionnez :
Données > Sélectionner des observations

Figure 12-10
Boîte de dialogue Sélectionner des observations



- Dans la boîte de dialogue Sélectionner des observations, sélectionnez Selon une condition logique et cliquez sur Si.

Figure 12-11
Sélectionner observations : Boîte de dialogue Si



- Dans la boîte de dialogue Sélectionner des observations : Si, entrez l'expression suivante :
Probabilité_de_réponse >= 0,05

Remarque : Si vous avez utilisé un nom différent pour le champ qui contient les valeurs de probabilité, entrez ce nom à la place de Probabilité_de_réponse. Le nom par défaut est *SelectedProbability*(ProbabilitéSélectionnée).

- Cliquez sur Poursuivre.
- Dans la boîte de dialogue Sélectionner des observations, sélectionnez Copier les observations sélectionnées dans un nouvel ensemble de données et entrez un nom pour ce nouvel ensemble de données. Le nom des ensembles de données doit être conforme aux règles de dénomination de champ (de variables).
- Cliquez sur OK pour créer l'ensemble de données contenant les contacts sélectionnés.

Le nouvel ensemble de données contient uniquement les contacts ayant une probabilité de réponse positive d'au moins 5 %.

Figure 12-12

Nouvel ensemble de données avec des contacts sélectionnés

Reside	Gender	Married	Region	Probability_of_responding
13	0	0	4	.07
15	0	0	4	.05
10	0	0	4	.05
5	0	0	4	.12
7	0	0	4	.08
10	0	0	4	.10
15	1	0	4	.05
11	1	0	4	.08
9	1	0	4	.08
9	1	0	4	.05

Récapitulatif

La propension à acheter utilise les résultats d'un envoi d'essai ou d'une campagne précédente pour générer des scores de propension. Les scores indiquent les contacts qui sont le plus susceptibles de répondre en fonction de diverses caractéristiques sélectionnées. Cette technique construit un modèle prédictif qui peut être appliqué à un ensemble de données pour obtenir des scores de propension.

Test de package de contrôle

Cette technique compare des campagnes de marketing pour vérifier s'il existe une différence significative de l'efficacité des différents packages et offres. L'efficacité de la campagne est mesurée en fonction des réponses

Par exemple, le service de marketing direct d'une société souhaite vérifier si un nouveau package générera davantage de réponses positives que le package existant. Ils procèdent à un envoi d'essai pour déterminer si le nouveau package génère un taux de réponse significativement plus élevé. L'envoi d'essai consiste en un groupe de contrôle qui reçoit le package existant et un groupe de test qui reçoit le nouveau package. Les résultats des deux groupes sont comparés pour juger de la différence.

Ces informations sont regroupées dans le fichier *dmdata.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 98.](#)

Exécution de l'analyse

- ▶ Pour obtenir un test de package de contrôle, dans les menus, sélectionnez :
Marketing direct > Choisissez la technique
- ▶ Sélectionnez Comparer l'efficacité des campagnes (test de package de contrôle) et cliquez sur Continuer.

Figure 13-1
Test de package de contrôle, onglet Champs

- ▶ Dans le champ Campagne, sélectionnez *Package de contrôle*.
- ▶ Pour le champ Réponse d'efficacité, sélectionnez *A répondu à une offre d'essai*.
- ▶ Sélectionnez Répondre.
- ▶ Comme valeur de réponse positive, sélectionnez *Oui* dans la liste déroulante. La valeur 1 est affichée dans le champ de texte car « Oui » est en fait une étiquette de valeur associée à la valeur enregistrée 1. (Si la valeur de réponse positive ne possède pas d'étiquette de valeur définie, vous pouvez simplement saisir la valeur dans le champ de texte.)

Un nouveau champ est créé automatiquement, dans lequel 1 représente des réponses positives et 0 représente des réponses négatives, et l'analyse est effectuée sur le nouveau champ. Vous pouvez ignorer le nom et l'étiquette par défaut et fournir les vôtres. Pour cet exemple, nous utiliserons le nom de champ déjà fourni.

- ▶ Cliquez sur Exécuter pour exécuter la procédure.

Résultats

Figure 13-2
Résultats du Test de package de contrôle

		Control Package			
		Control		Test	
		Effectif	N % colonne	Effectif	N % colonne
Efficacité (1 =OUI 0=NON)	0	875	96,2%	945	93,8%
	1	35	3,8%	62	6,2%

Il existe une différence statistiquement significative entre Control et Test.

La sortie à partir de la procédure inclut un tableau qui affiche le nombre et les pourcentages de réponses positives et négatives de chaque groupe défini par le champ Campagne et un tableau qui indique si les taux de réponse des groupes diffèrent de manière significative les uns des autres.

- *Efficacité* est la version recodée du champ de réponse, où 1 représente les réponses positives et 0 les réponses négatives.
- Le taux de réponses positives pour le package de contrôle est de 3,8 %, alors que le taux de réponses positives pour le package de test est de 6,2 %.

La courte description sous le tableau indique que la différence entre les groupes est significative, ce qui signifie que le plus haut taux de réponse pour le package de test n'est probablement pas dû au hasard. Ce tableau de texte contiendra une comparaison pour chaque paire de groupes possible incluse dans l'analyse. Puisqu'il n'y a que deux groupes dans cet exemple, une seule comparaison est effectuée. S'il y a plus de cinq groupes, le tableau de description textuelle est remplacé par le tableau Comparaison des proportions de colonnes.

Récapitulatif

Le test de package de contrôle compare des campagnes de marketing pour vérifier s'il existe une différence significative de l'efficacité des différents packages et offres. Dans cet exemple, les réponses positives de 6,2 % au package de test étaient significativement supérieures aux taux de réponses positives de 3,8 % du package de contrôle. Ceci suggère que vous devriez utiliser le nouveau package au lieu de l'ancien, mais il peut y avoir d'autres facteurs que vous devriez prendre en considération tels que les coûts supplémentaires associés au nouveau package.

Fichiers d'exemple

Les fichiers d'exemple installés avec le produit figurent dans le sous-répertoire *Echantillons* du répertoire d'installation. Il existe un dossier distinct au sein du sous-répertoire *Echantillons* pour chacune des langues suivantes : Anglais, Français, Allemand, Italien, Japonais, Coréen, Polonais, Russe, Chinois simplifié, Espagnol et Chinois traditionnel.

Seuls quelques fichiers d'exemples sont disponibles dans toutes les langues. Si un fichier d'exemple n'est pas disponible dans une langue, le dossier de langue contient la version anglaise du fichier d'exemple.

Descriptions

Voici de brèves descriptions des fichiers d'exemple utilisés dans divers exemples à travers la documentation.

- **accidents.sav.** Ce fichier de données d'hypothèse concerne une société d'assurance qui étudie les facteurs de risque liés à l'âge et au sexe dans les accidents de la route survenant dans une région donnée. Chaque observation correspond à une classification croisée de la catégorie d'âge et du sexe.
- **adl.sav.** Ce fichier de données d'hypothèse concerne les mesures entreprises pour identifier les avantages d'un type de thérapie proposé aux patients qui ont subi une attaque cardiaque. Les médecins ont assigné de manière aléatoire les patients du sexe féminin ayant subi une attaque cardiaque à un groupe parmi deux groupes possibles. Le premier groupe a fait l'objet de la thérapie standard tandis que le second a bénéficié en plus d'une thérapie émotionnelle. Trois mois après les traitements, les capacités de chaque patient à effectuer les tâches ordinaires de la vie quotidienne ont été notées en tant que variables ordinales.
- **advert.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- **aflatoxin.sav.** Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'alfatoxine en parties par milliard (PPB).
- **anorectic.sav.** En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.

- **bankloan.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondent aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- **bankloan_binning.sav.** Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- **behavior.sav.** Dans un exemple classique, on a demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = « extrêmement approprié » et 9 = « extrêmement inapproprié ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.
- **behavior_ini.sav.** Ce fichier de données contient la configuration initiale d'une solution bidimensionnelle pour *behavior.sav*.
- **brakes.sav.** Ce fichier de données d'hypothèse concerne le contrôle qualité effectué dans une usine qui fabrique des freins à disque pour des voitures haut de gamme. Le fichier de données contient les mesures de diamètre de 16 disques de 8 machines de production. Le diamètre cible des freins est de 322 millimètres.
- **breakfast.sav.** Au cours d'une étude classique, on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1 = « aliment préféré » à 15 = « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ».
- **breakfast-overall.sav.** Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, « Préférence générale ».
- **broadband_1.sav.** Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- **broadband_2.sav.** Ce fichier de données est identique au fichier *broadband_1.sav* mais contient les données relatives à trois mois supplémentaires.
- **car_insurance_claims.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modélisé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- **car_sales.sav.** Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de *edmunds.com* et des sites des constructeurs.
- **car_sales_uprepared.sav.** Il s'agit d'une version modifiée de *car_sales.sav* qui n'inclut aucune version transformée des champs.

- **carpet.sav.** Dans un exemple courant, une société intéressée par la commercialisation d'un nouveau nettoyeur de tapis souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil.
- **carpet_prefs.sav.** Ce fichier de données repose sur le même exemple que celui décrit pour *carpet.sav*, mais contient les classements réels issus de chacun des 10 clients. On a demandé aux consommateurs de classer les 22 profils de produits, du préféré au moins intéressant. Les variables *PREF1* à *PREF22* contiennent les identificateurs des profils associés, tels qu'ils sont définis dans *carpet_plan.sav*.
- **catalog.sav.** Ce fichier de données contient des chiffres de ventes mensuelles hypothétiques relatifs à trois produits vendus par une entreprise de vente par correspondance. Les données relatives à cinq variables explicatives possibles sont également incluses.
- **catalog_seasfac.sav.** Ce fichier de données est identique à *catalog.sav* mais contient en plus un ensemble de facteurs saisonniers calculés à partir de la procédure de désaisonnalisation, ainsi que les variables de date correspondantes.
- **cellular.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un opérateur téléphonique pour réduire les taux de désabonnement. Des scores de propension au désabonnement sont attribués aux comptes, de 0 à 100. Les comptes ayant une note égale ou supérieure à 50 sont susceptibles de changer de fournisseur.
- **ceramics.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fabricant pour déterminer si un nouvel alliage haute qualité résiste mieux à la chaleur qu'un alliage standard. Chaque observation représente un test séparé de l'un des deux alliages ; le degré de chaleur auquel l'alliage ne résiste pas est enregistré.
- **cereal.sav.** Ce fichier de données d'hypothèse concerne un sondage de 880 personnes interrogées sur leurs préférences de petit-déjeuner et sur leur âge, leur sexe, leur situation familiale et leur mode de vie (actif ou non actif, selon qu'elles pratiquent une activité physique au moins deux fois par semaine). Chaque observation correspond à un répondant distinct.
- **clothing_defects.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de textile. Dans chaque lot produit à l'usine, les inspecteurs prélèvent un échantillon de vêtements et comptent le nombre de vêtements qui ne sont pas acceptables.
- **coffee.sav.** Ce fichier de données concerne l'image perçue de six marques de café frappé. Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées AA, BB, CC, DD, EE et FF à des fins de confidentialité.
- **contacts.sav.** Ce fichier de données d'hypothèse concerne les listes de contacts d'un groupe de représentants en informatique d'entreprise. Chaque contact est classé selon le service de l'entreprise où il travaille et le classement de son entreprise. Sont également enregistrés le

montant de la dernière vente effectuée, le temps passé depuis la dernière vente et la taille de l'entreprise du contact.

- **creditpromo.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un grand magasin pour évaluer l'efficacité d'une promotion récente de carte de crédit. A cette fin, 500 détenteurs de carte ont été sélectionnés au hasard. La moitié a reçu une publicité faisant la promotion d'un taux d'intérêt réduit sur les achats effectués dans les trois mois à venir. L'autre moitié a reçu une publicité saisonnière standard.
- **customer_dbase.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour utiliser les informations figurant dans sa banque de données et proposer des offres spéciales aux clients susceptibles d'être intéressés. Un sous-groupe de la base de clients a été sélectionné au hasard et a reçu des offres spéciales. Les réponses des clients ont été enregistrées.
- **customer_information.sav.** Un fichier de données d'hypothèse qui contient les informations postales du client, telles que le nom et l'adresse.
- **customer_subset.sav.** Un sous-ensemble de 80 observations de *customer_dbase.sav*.
- **debate.sav.** Ce fichier de données d'hypothèse concerne des réponses appariées à une enquête donnée aux participants à un débat politique avant et après le débat. Chaque observation représente un répondant distinct.
- **debate_aggregate.sav.** Il s'agit d'un fichier de données d'hypothèse qui rassemble les réponses dans le fichier *debate.sav*. Chaque observation correspond à une classification croisée de préférence avant et après le débat.
- **demo.sav.** Ce fichier de données d'hypothèse concerne une base de données clients achetée en vue de diffuser des offres mensuelles. Les données indiquent si le client a répondu ou non à l'offre et contiennent diverses informations démographiques.
- **demo_cs_1.sav.** Ce fichier de données d'hypothèse concerne la première mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à une ville différente. La région, la province, le quartier et la ville sont enregistrés.
- **demo_cs_2.sav.** Ce fichier de données d'hypothèse concerne la seconde mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à un ménage différent issu des villes sélectionnées à la première étape. La région, la province, le quartier, la ville, la sous-division et l'identification sont enregistrés. Les informations d'échantillonnage des deux premières étapes de la conception sont également incluses.
- **demo_cs.sav.** Ce fichier de données d'hypothèse concerne des informations d'enquête collectées via une méthode complexe d'échantillonnage. Chaque observation correspond à un ménage différent et diverses informations géographiques et d'échantillonnage sont enregistrées.
- **dmdata.sav.** Ceci est un fichier de données d'hypothèse qui contient des informations démographiques et des informations concernant les achats pour une entreprise de marketing direct. *dmdata2.sav* contient les informations pour un sous-ensemble de contacts qui ont reçu un envoi d'essai, et *dmdata3.sav* contient des informations sur les contacts restants qui n'ont pas reçu l'envoi d'essai.

- **dietstudy.sav.** Ce fichier de données d'hypothèse contient les résultats d'une étude portant sur le régime de Stillman. Chaque observation correspond à un sujet distinct et enregistre son poids en livres avant et après le régime, ainsi que ses niveaux de triglycérides en mg/100 ml.
- **dvdplayer.sav.** Ce fichier de données d'hypothèse concerne le développement d'un nouveau lecteur DVD. A l'aide d'un prototype, l'équipe de marketing a collecté des données de groupes spécifiques. Chaque observation correspond à un utilisateur interrogé et enregistre des informations démographiques sur cet utilisateur, ainsi que ses réponses aux questions portant sur le prototype.
- **german_credit.sav.** Ce fichier de données provient de l'ensemble de données « German credit » figurant dans le référentiel Machine Learning Databases de l'université de Californie, Irvine.
- **grocery_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *grocery_coupons.sav* dans lequel les achats hebdomadaires sont organisés par client distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, le montant dépensé enregistré est à présent la somme des montants dépensés au cours des quatre semaines de l'enquête.
- **grocery_coupons.sav.** Il s'agit d'un fichier de données d'hypothèse qui contient des données d'enquête collectées par une chaîne de magasins d'alimentation qui cherchent à déterminer les habitudes de consommation de ses clients. Chaque client est suivi pendant quatre semaines et chaque observation correspond à une semaine distincte. Les informations enregistrées concernent les endroits où le client effectue ses achats, la manière dont il les effectue, ainsi que les sommes dépensées en provisions au cours de cette semaine.
- **guttman.sav.** Bell a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).
- **health_funding.sav.** Ce fichier de données d'hypothèse concerne des données sur le financement des soins de santé (montant par groupe de 100 individus), les taux de maladie (taux par groupe de 10 000 individus) et les visites chez les prestataires de soins de santé (taux par groupe de 10 000 individus). Chaque observation représente une ville différente.
- **hivassay.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un laboratoire pharmaceutique pour développer une analyse rapide de détection d'infection HIV. L'analyse a pour résultat huit nuances de rouge, les nuances les plus marquées indiquant une plus forte probabilité d'infection. Un test en laboratoire a été effectué sur 2 000 échantillons de sang, la moitié de ces échantillons étant infectée par le virus HIV et l'autre moitié étant saine.
- **hourlywagedata.sav.** Ce fichier de données d'hypothèse concerne les salaires horaires d'infirmières occupant des postes administratifs et dans les services de soins, et affichant divers niveaux d'expérience.

- **insurance_claims.sav.** Il s'agit d'un fichier de données hypothétiques qui concerne une compagnie d'assurance souhaitant développer un modèle pour signaler des réclamations suspectes, potentiellement frauduleuses. Chaque observation correspond à une réclamation distincte.
- **insure.sav.** Ce fichier de données d'hypothèse concerne une compagnie d'assurance qui étudie les facteurs de risque indiquant si un client sera amené à déclarer un incident au cours d'un contrat d'assurance vie d'une durée de 10 ans. Chaque observation figurant dans le fichier de données représente deux contrats, l'un ayant enregistré une réclamation et l'autre non, appariés par âge et sexe.
- **judges.sav.** Ce fichier de données d'hypothèse concerne les scores attribués par des juges expérimentés (plus un juge enthousiaste) à 300 performances de gymnastique. Chaque ligne représente une performance distincte ; les juges ont examiné les mêmes performances.
- **kinship_dat.sav.** Rosenberg et Kim se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Ainsi, un total de six "sources" a été obtenu. Chaque source correspond à une matrice de proximité 15×15 , dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source.
- **kinship_ini.sav.** Ce fichier de données contient une configuration initiale d'une solution tridimensionnelle pour *kinship_dat.sav*.
- **kinship_var.sav.** Ce fichier de données contient les variables indépendantes *sexe*, *génér(ation)* et *degré* (de séparation) permettant d'interpréter les dimensions d'une solution pour *kinship_dat.sav*. Elles permettent en particulier de réduire l'espace de la solution à une combinaison linéaire de ces variables.
- **marketvalues.sav.** Ce fichier de données concerne les ventes de maisons dans un nouvel ensemble à Algonquin (Illinois) au cours des années 1999–2000. Ces ventes relèvent des archives publiques.
- **nhis2000_subset.sav.** Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des États-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage. Ce fichier de données contient un sous-groupe d'informations issues de l'enquête de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Fichier de données et documentation d'usage public. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/. Accès en 2003.
- **ozone.sav.** Les données incluent 330 observations portant sur six variables météorologiques pour prévoir la concentration d'ozone à partir des variables restantes. Des chercheurs précédents, , ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.

- **pain_medication.sav.** Ce fichier de données d'hypothèse contient les résultats d'un essai clinique d'un remède anti-inflammatoire traitant les douleurs de l'arthrite chronique. On cherche notamment à déterminer le temps nécessaire au médicament pour agir et les résultats qu'il permet d'obtenir par rapport à un médicament existant.
- **patient_los.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux de patients admis à l'hôpital pour suspicion d'infarctus du myocarde suspecté (ou « attaque cardiaque »). Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **patlos_sample.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux d'un échantillon de patients sous traitement thrombolytique après un infarctus du myocarde. Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **poll_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un enquêteur pour déterminer le niveau de soutien du public pour un projet de loi avant législature. Les observations correspondent à des électeurs enregistrés. Chaque observation enregistre le comté, la ville et le quartier où habite l'électeur.
- **poll_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des électeurs répertoriés dans le fichier *poll_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *poll_csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. Toutefois, ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS – Probability-Proportional-to-Size), il existe également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Les variables supplémentaires correspondant à la répartition démographique des électeurs et à leur opinion sur le projet de loi proposé ont été collectées et ajoutées au fichier de données une fois l'échantillon prélevé.
- **property_assess.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur au niveau du comté pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés vendues dans le comté au cours de l'année précédente. Chaque observation du fichier de données enregistre la ville où se trouve la propriété, l'évaluateur ayant visité la propriété pour la dernière fois, le temps écoulé depuis cette évaluation, l'évaluation effectuée à ce moment-là et la valeur de vente de la propriété.
- **property_assess_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur du gouvernement pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés de l'état. Chaque observation du fichier de données enregistre le comté, la ville et le quartier où se trouve la propriété, le temps écoulé depuis la dernière évaluation et l'évaluation alors effectuée.
- **property_assess_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des propriétés répertoriées dans le fichier *property_assess_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *property_assess_csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. La variable supplémentaire *Valeur courante* a été collectée et ajoutée au fichier de données une fois l'échantillon prélevé.

- **recidivism.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis, ainsi que le temps écoulé jusqu'à la seconde arrestation si elle s'est produite dans les deux années suivant la première.
- **recidivism_cs_sample.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste libéré suite à la première arrestation en juin 2003 et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis et les données relatives à la seconde arrestation, si elle a eu lieu avant fin juin 2006. Les récidivistes ont été choisis dans plusieurs départements échantillonnés conformément au plan d'échantillonnage spécifié dans *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Un fichier de données d'hypothèse qui contient les données de transaction d'achat, y compris la date d'achat, le/les élément(s) acheté(s) et le montant monétaire pour chaque transaction.
- **salesperformance.sav.** Ce fichier de données d'hypothèse concerne l'évaluation de deux nouveaux cours de formation en vente. Soixante employés, divisés en trois groupes, reçoivent chacun une formation standard. En outre, le groupe 2 suit une formation technique et le groupe 3 un didacticiel pratique. A l'issue du cours de formation, chaque employé est testé et sa note enregistrée. Chaque observation du fichier de données représente un stagiaire distinct et enregistre le groupe auquel il a été assigné et la note qu'il a obtenue au test.
- **satisf.sav.** Il s'agit d'un fichier de données d'hypothèse portant sur une enquête de satisfaction effectuée par une société de vente au détail au niveau de quatre magasins. Un total de 582 clients ont été interrogés et chaque observation représente la réponse d'un seul client.
- **screws.sav.** Ce fichier de données contient des informations sur les descriptives des vis, des boulons, des écrous et des clous..
- **shampoo_ph.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de produits capillaires. A intervalles réguliers, six lots de sortie distincts sont mesurés et leur pH enregistré. La plage cible est 4,5–5,5.
- **ships.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs et concernant les dommages causés à des cargos par les vagues. Les effectifs d'incidents peuvent être modélisés comme des incidents se produisant selon un taux de Poisson en fonction du type de navire, de la période de construction et de la période de service. Les mois de service totalisés pour chaque cellule du tableau formé par la classification croisée des facteurs fournissent les valeurs d'exposition au risque.
- **site.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour choisir de nouveaux sites pour le développement de ses activités. L'entreprise a fait appel à deux consultants pour évaluer séparément les sites. Ces consultants, en plus de fournir un rapport approfondi, ont classé chaque site comme constituant une éventualité « bonne », « moyenne » ou « faible ».

- **smokers.sav.** Ce fichier de données est extrait de l'étude National Household Survey of Drug Abuse de 1998 et constitue un échantillon de probabilité des ménages américains. (<http://dx.doi.org/10.3886/ICPSR02934>) Ainsi, la première étape dans l'analyse de ce fichier doit consister à pondérer les données pour refléter les tendances de population.
- **stocks.sav** Ce fichier de données hypothétiques contient le cours et le volume des actions pour un an.
- **stroke_clean.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois celle-ci purgée via des procédures de l'option Validation de données.
- **stroke_invalid.sav.** Ce fichier de données d'hypothèse concerne l'état initial d'une base de données médicales et comporte plusieurs erreurs de saisie de données.
- **stroke_survival.** Ce fichier de données d'hypothèse concerne les temps de survie de patients qui quittent un programme de rééducation à la suite d'un accident ischémique et rencontrent un certain nombre de problèmes. Après l'attaque, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. L'échantillon est tronqué à gauche car il n'inclut que les patients ayant survécu durant le programme de rééducation mis en place suite à une attaque.
- **stroke_valid.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois les valeurs vérifiées via la procédure Validation de données. Elle contient encore des observations anormales potentielles.
- **survey_sample.sav.** Ce fichier de données concerne des informations d'enquête dont des données démographiques et des mesures comportementales. Il est basé sur un sous-ensemble de variables de la 1998 NORC General Social Survey, bien que certaines valeurs de données aient été modifiées et que des variables supplémentaires fictives aient été ajoutées à titre de démonstration.
- **telco.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société de télécommunications pour réduire les taux de désabonnement de sa base de clients. Chaque observation correspond à un client distinct et enregistre diverses informations démographiques et d'utilisation de service.
- **telco_extra.sav.** Ce fichier de données est semblable au fichier de données *telco.sav* mais les variables de permanence et de dépenses des consommateurs transformées log ont été supprimées et remplacées par des variables de dépenses des consommateurs transformées log standardisées.
- **telco_missing.sav.** Ce fichier de données est un sous-ensemble du fichier de données *telco.sav* mais certaines des valeurs de données démographiques ont été remplacées par des valeurs manquantes.
- **testmarket.sav.** Ce fichier de données d'hypothèse concerne une chaîne de fast foods et ses plans marketing visant à ajouter un nouveau plat à son menu. Trois campagnes étant possibles pour promouvoir le nouveau produit, le nouveau plat est introduit sur des sites sur plusieurs marchés sélectionnés au hasard. Une promotion différente est effectuée sur chaque site et les ventes hebdomadaires du nouveau plat sont enregistrées pour les quatre premières semaines. Chaque observation correspond à un site-semaine distinct.
- **testmarket_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *testmarket.sav* dans lequel les ventes hebdomadaires sont organisées par site distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, les ventes

enregistrées sont à présent la somme des ventes réalisées au cours des quatre semaines de l'enquête.

- **tree_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_credit.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire.
- **tree_missing_data.sav** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire avec un grand nombre de valeurs manquantes.
- **tree_score_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_textdata.sav.** Ce fichier de données simples ne comporte que deux variables et vise essentiellement à indiquer l'état par défaut des variables avant affectation du niveau de mesure et des étiquettes de valeurs.
- **tv-survey.sav.** Ce fichier de données d'hypothèse concerne une enquête menée par un studio de télévision qui envisage de prolonger la diffusion d'un programme ou de l'arrêter. On a demandé à 906 personnes si elles regarderaient le programme dans diverses situations. Chaque ligne représente un répondant distinct et chaque colonne une situation distincte.
- **ulcer_recurrence.sav.** Ce fichier contient des informations partielles d'une enquête visant à comparer l'efficacité de deux thérapies de prévention de la récurrence des ulcères. Il fournit un bon exemple de données censurées par intervalle et a été présenté et analysé ailleurs .
- **ulcer_recurrence_recoded.sav.** Ce fichier réorganise les informations figurant dans le fichier *ulcer_recurrence.sav* pour que vous puissiez modéliser la probabilité d'événement pour chaque intervalle de l'enquête plutôt que la probabilité d'événement de fin d'enquête. Il a été présenté et analysé ailleurs .
- **verd1985.sav.** Ce fichier de données concerne une enquête . Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables présentant un intérêt sont divisées en trois ensembles. Le groupe 1 comprend l'âge et la *situation familiale*, le groupe 2 les *animaux domestiques* et la *presse*, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal domestique* est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple.
- **virus.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fournisseur de services Internet pour déterminer les effets d'un virus sur ses réseaux. Il a suivi le pourcentage (approximatif) de trafic de messages électroniques infectés par un virus sur ses réseaux sur la durée, de la découverte à la circonscription de la menace.
- **wheeze_steubenville.sav.** Il s'agit d'un sous-ensemble d'une enquête longitudinale des effets de la pollution de l'air sur la santé des enfants . Les données contiennent des mesures binaires répétées de l'état asthmatique d'enfants de la ville de Steubenville (Ohio), âgés de 7, 8, 9 et 10 ans, et indiquent si la mère fumait au cours de la première année de l'enquête.
- **workprog.sav.** Ce fichier de données d'hypothèse concerne un programme de l'administration visant à proposer de meilleurs postes aux personnes défavorisées. Un échantillon de participants potentiels au programme a ensuite été prélevé. Certains de ces participants ont

été sélectionnés au hasard pour participer au programme. Chaque observation représente un participant au programme distinct.

- **worldsales.sav** Ce fichier de données hypothétiques contient les revenus des ventes par continent et par produit.

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, États-Unis

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.

Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local : INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ÉTAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.

Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

Marques commerciales

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse <http://www.ibm.com/legal/copytrade.shtml>.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques déposées ou des marques commerciales de Adobe Systems Incorporated aux États-Unis et/ou dans d'autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux États-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux États-Unis et/ou dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux États-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux États-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux États-Unis et dans d'autres pays.

Ce produit utilise WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, <http://www.winwrap.com/>.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.



Index

classe, 15
classification , 15, 53
Classification, 15

fichiers d'exemple
 emplacement, 98

marques commerciales, 110
mentions légales, 109

profils des clients potentiels, 20, 70
propension à acheter, 34, 83

régression logistique , 34
Régression logistique, 83
RFM, 2, 9, 11, 13, 46
 données de clients, 4
 données de transactions, 3, 46
 regroupement par casiers, 6

taux de réponse des codes postaux, 27, 76
test de package de contrôle, 42, 95