COURS DE DATA MINING

2 : COMPREHENSION DU METIER COMPREHENSION ET PREPARATION DES DONNEES

EPF - 4ème année - Option Ingénierie d'Affaires et de Projets
Bertrand LIAUDET

PHASE 1 : COMPREHENSION DU METIER

PROCESSUS du DATA MINING		
Acteurs	**Étapes**	**Phases**
Maître d'œuvre	Objectifs	1 : Compréhension du métier
	Données	2 : Compréhension des données
		3 : Préparation des données
	Traitements	4 : Modélisation
		5 : Évaluation de la modélisation
Maître d'ouvrage	Déploiement	6 : Déploiement des résultats de l'étude

Description de la phase de compréhension du métier

La phase de compréhension du métier consiste à établir clairement le cahier des charges du projet de data mining.

Cette phase est réalisé par le maître d'œuvre à partir de la demande du maître d'ouvrage, demande éventuellement exprimée dans un cahier des charges.

Cette phase est l'équivalent de la phase de spécifications fonctionnelles d'un projet de développement informatique.

Cette phase consiste à :

- Énoncer clairement les objectifs globaux du projet et les contraintes de l'entreprise.
- Traduire ces objectifs et ces contraintes en un problème de data mining.

  Il s'agit donc de formuler une recherche de corrélations, c'est-à-dire la recherche de règles du type : si A alors B.

  Pour cela, on commencera si nécessaire par produire un modèle conceptuel des données.
- Préparer une stratégie initiale pour atteindre ces objectifs.

La phase de compréhension métier amènera probablement le maître d'œuvre communiquer avec les experts du métier (le maître d'ouvrage).

Trois exemples

1 : Constructeur automobile : analyse des plaintes de clients

Contexte :

Un constructeur automobile veut étudier les plaintes de clients acheteur d'automobile.

Le constructeur fournit un système d'information sur la qualité qui contient les informations de plus de 7 millions de véhicules et qui a une taille de plus de 5 GO et d'une base de données contenant des informations sur les plaintes et sur les véhicules : caractéristiques de la fabrication, caractéristiques de la voiture (options, etc.), caractéristiques du garage vendeur.

Objectifs :

- Réduire les coûts associés aux plaintes.
- Améliorer la satisfaction des clients.

MCD

Garages <------Voitures <-------- Plaintes

Une plainte correspond à une voiture et une seule

Une voiture correspond à un garage et un seul

Donc, une plainte correspond à un garage et un seul

Problèmes de data mining :

- Peut-on établir une corrélation entre les voitures et les plaintes ? Et plus précisément, entre une ou plusieurs caractéristiques techniques de voiture et certains types de plaintes.
- Peut-on établir une corrélation entre les garages et les plaintes ? Et plus précisément certains types de plaintes.

Types de résultats qu'on pourrait obtenir :

Une combinaison donnée de spécifications techniques de la voiture augmente la probabilité de rencontrer un problème de câblage électrique (par exemple).

2 : Service social d'assistance juridique

Contexte :

Un service social gère des demandes d'assistance juridique.

Le service social dispose de plus de 380 000 cas d'assistance juridique.

Objectifs :

- Mieux comprendre le profil des gens qui demandent une aide juridique

- Trouver des liens entre le profil des gens et le type d'aide demandée.

Problèmes de data mining :

- Peut-on établir une corrélation entre le profil des gens et le type d'aide demandé ?
- Peut-on établir une corrélation entre le profil des gens et le résultat obtenu (aide financière ou pas).

Types de résultats qu'on pourrait obtenir :

Si profil = p alors type d'aide = a

3 : Crise économique dans le sud-est asiatique

Contexte :

La crise économique de 1997-1998 dans le sud-est asiatique a conduit à un grand nombre de dépôts de bilan. On veut étudier les raisons de ces dépôts de bilan.

On partira de données concernant les entreprises en dépôt de bilan pendant la période 97-98, mais aussi de la période 91-95 qui était une période à croissance stable.

Objectifs :

- Produire un modèle pour prédire le dépôt de bilan.

Problèmes de data mining :

- Générer des ensembles de règles pour éviter une formulation trop dogmatique (on utilisera pour cela les arbres de décision).

Types de résultats qu'on pourrait obtenir :

Si la productivité du capital (volume produit / volume du capital fixe) est supérieure à 19,65, alors on prédit le non-dépôt de bilan avec un seuil de confiance à 86 %.

PHASE 2 : COMPREHENSION DES DONNEES

<table>
<tr><th colspan="3">PROCESSUS du DATA MINING</th></tr>
<tr><th>Acteurs</th><th>Étapes</th><th>Phases</th></tr>
<tr><td rowspan="5">Maître d'œuvre</td><td>Objectifs</td><td>1 : Compréhension du métier</td></tr>
<tr><td rowspan="2">Données</td><td>2 : Compréhension des données</td></tr>
<tr><td>3 : Préparation des données</td></tr>
<tr><td rowspan="2">Traitements</td><td>4 : Modélisation</td></tr>
<tr><td>5 : Évaluation de la modélisation</td></tr>
<tr><td>Maître d'ouvrage</td><td colspan="2">Déploiement des résultats de l'étude</td></tr>
</table>

Introduction

Présentation générale

Les phases de compréhension et de préparation sont deux phases très importantes d'un projet de data mining. Elles se font en aller-retour. Pour préparer, il faut comprendre, pour comprendre il faut avoir préparé.

Ces deux phases vont occuper plus de la moitié du temps d'un projet de data mining.

Le principe de la compréhension des données est **qu'il faut comprendre toutes les valeurs des tableaux** sur lesquels on va travailler (un seul en général), ce qui passe par l'analyse de la répartition des valeurs dans chaque colonne, le repérage des valeurs anormales et des valeurs manquantes et l'analyse des corrélations les plus évidentes.

Ces analyses élémentaires ne posent pas de problèmes théoriques. Elles demandent uniquement du bon sens et de la rigueur.

On y ajoutera la recherche de corrélations moins évidentes.

La bonne réalisation des ces deux phases est nécessaires à une modélisation correcte.

Vocabulaire - notions de donnée, de variable, de type, de modèle

Globalement, on travaille sur des tableaux de données.

- Le tableau, c'est « ce dont on parle », c'est-à-dire le « concept » dont on parle. C'est une abstraction. Par exemple, un tableau de clients, de malades, etc.

  Rappelons qu'un concept (ou **notion**, ou **idée**) est une représentation mentale générale et abstraite d'un objet. Le concept est le résultat de l'opération de l'esprit qui fait qu'on place tel objet dans telle catégorie et non dans telle autre.

- Chaque ligne du tableau est un élément du tableau, c'est-à-dire un objet concret correspondant au concept abstrait dont on parle.
- Chaque colonne du tableau est un attribut du « concept ». On parle aussi de « propriété ». La colonne a un nom général qui est le nom de l'attribut pour le concept. Pour un objet concret, la colonne a une valeur particulière qui est la valeur particulière de l'attribut pour l'objet concret.

  En data mining (et en statistique), **les attributs des objets sont appelés : « variables ».**

  En data mining, **chaque valeur d'un objet concret est appelé : « donnée ».**
- **Un sous-ensemble de valeurs pour un ou plusieurs attributs donnés peut être appelé : « type », « classe », « catégorie » , « segment » ou encore « modalité »**

  Par exemple, « grand » et « petit » sont deux types (ou classe, ou catégorie, ou segment) de l'attribut « taille ».

  On parle de **« variable catégorielle »** par opposition aux **« variables numériques ».** Par exemple, si la variable (attribut) « taille » peut prendre deux valeurs possibles : « grand » et « petit », c'est une variable catégorielle. Si les valeurs de la variable « taille » sont données en cm, c'est une variable numérique.
- **Quand on fait de la prévision**, on travaille sur une variable particulière appelée **: « variable cible »** et sur un ensemble d'autres variables utiles pour la prédiction appelées : **« prédicteurs ».**

  Le principe général de la prédiction sera : si le ou les prédicteurs valent tant, alors la variable cible vaut tant.
- Les statisticiens construisent des **modèles**. Un modèle est un résumé global des relations entre variables permettant de comprendre des phénomènes (description, jugement) et d'émettre des prévisions (prédiction, raisonnement).

  Dans l'absolu, tous les modèles sont faux. Un modèle n'est pas une loi scientifique. Cependant, certains sont utiles.

Plan du chapitre : les étapes de la compréhension des données

1. Recueillir et préparer les données
2. Dictionnaire des variables.
3. Statistiques de base par variable. Remarque : qualité, statistiques et corrélations relèvent de ce qu'on appelle : **l'analyse exploratoire des données.**
4. Suppression de la clé primaire
5. Nettoyage des données : valeurs manquantes et aberrantes
6. Corrélations de base entre variables numériques : suppression des variables calculées.

Méthodes graphiques : histogrammes et nuages de point.

7. Corrélations entre variables catégorielles.
8. Corrélations entre variables numériques et variables catégorielles.
9. Corrélations multiples.
10. Dernières techniques : discrétisation, variables calculées, sous-ensembles

1 : Recueillir les données

Le recueil des données consiste à faire l'inventaire des données existantes et utilisables.

Ces données se présenteront sous forme de tables (comme les tables des bases de données ou des tableaux excel).

Les analyses de data mining s'appliquent à une seule table.

Si on a plusieurs tables, il faudra commencer par définir la ou les tables de travail en fusionnant éventuellement certaines tables entre elles (par des jointures classiques) pour pouvoir commencer à faire l'étude. Il est possible de faire des jointures avec les logiciels de data mining.

On reviendra sur cette partie dans la phase 3 sur la préparation des données.

2 : Dictionnaire des variables : signification et type

Présentation

On commence par établir un dictionnaire des variables (à peu près équivalent au dictionnaire des données des bases de données).

Il faut donc commencer par lister les variables (les attributs) et leurs caractéristiques. Éventuellement, on fera appel à des experts du domaine pour finir de rendre intelligibles les attributs et leurs valeurs.

On peut considérer 5 colonnes dans le dictionnaires :

- le numéro (référence de la variable),
- le nom de la variable dans le tableau de données,
- la signification de la variable,
- son type, en précisant si on a affaire à une **variable catégorielle (type énuméré)** ou pas
- une colonne pour préciser d'autres caractéristiques.

La signification d'une variable

Pour trouver la signification d'une variable, on peut faire appel à :

- Culture générale
- Documentation
- Comparaison avec les autres variables
- Les données de la variables (les valeurs prises par la variable)
- Expert du domaine

Distinction entre variable continue (numérique) et variable discrète (catégorielle)

On distingue essentiellement entre deux types de variable.

Variable continue (numérique)

Les variables numériques permettent de faire des calculs statistiques de base (moyenne, minimum, maximum, etc. Elles permettent aussi de faire des corrélations linéaires.

Variable discrète (catégorielle)

Les variables catégorielles permettent de définir des sous-ensembles par catégorie (valeur possible de la variables). Cela correspond à un type énuméré en informatique.

Une variable ni numérique, ni catégorielle ne présente aucun intérêt pour une étude de data mining (et de statistique en général). En effet, puisqu'on ne peut ni en tirer de connaissances statistiques, ni l'utiliser pour des regroupements, elle ne sert à rien. C'est typiquement le cas des clés primaires dans les bases de données. Mais aussi le nom des gens dans un tableau de clients.

Les différents types

Entier

Le type entier est en général **numérique**. Il peut être **catégoriel** si l'entier est un code de catégorie, ou si des regroupements sont possibles par valeur de l'entier (par exemple, l'âge en année entière).

Réel

Le type réel est **presque toujours numérique**, sauf si des regroupements sont possibles par valeurs du réel (cas plutôt rare).

Date

Le type date peut être **numérique et / ou catégorielle**. Il est catégoriel si des regroupements sont possibles par valeur de la date.

Chaîne de caractères

Le type chaîne de caractères est **toujours catégoriel**. On ne peut pas faire de calculs statistiques dessus.

La transformation des types

On peut être amené à changer le type des variables pour pouvoir leur appliquer certaines techniques d'investigation.

Particulièrement, on peut passer de variables numériques à des variables catégorielles.

Exemple 01 : fichier de données d'attrition

On prend l'exemple d'un fichier de données d'attrition.

L'attrition, c'est l'usure, c'est-à-dire le fait de rompre ou pas le contrat. Le score d'attrition est un cas d'étude de data mining très important. En effet, c'est un domaine dans lequel on a beaucoup de clients, relativement peu d'opérateurs, et dans lequel le coût de la communication pour trouver de nouveaux clients est très important. L'opérateur a donc intérêt à fidéliser ses clients. L'étude statistique des données clients va permettre de mieux cibler les attentes du client.

Le fichier contient 21 variables et 3333 objets.

On peut l'ouvrir avec un éditeur de texte ou sous excel ou directement avec un logiciel de data mining Excel permet de faire des tris, ce qui est l'opération de base sur les données.

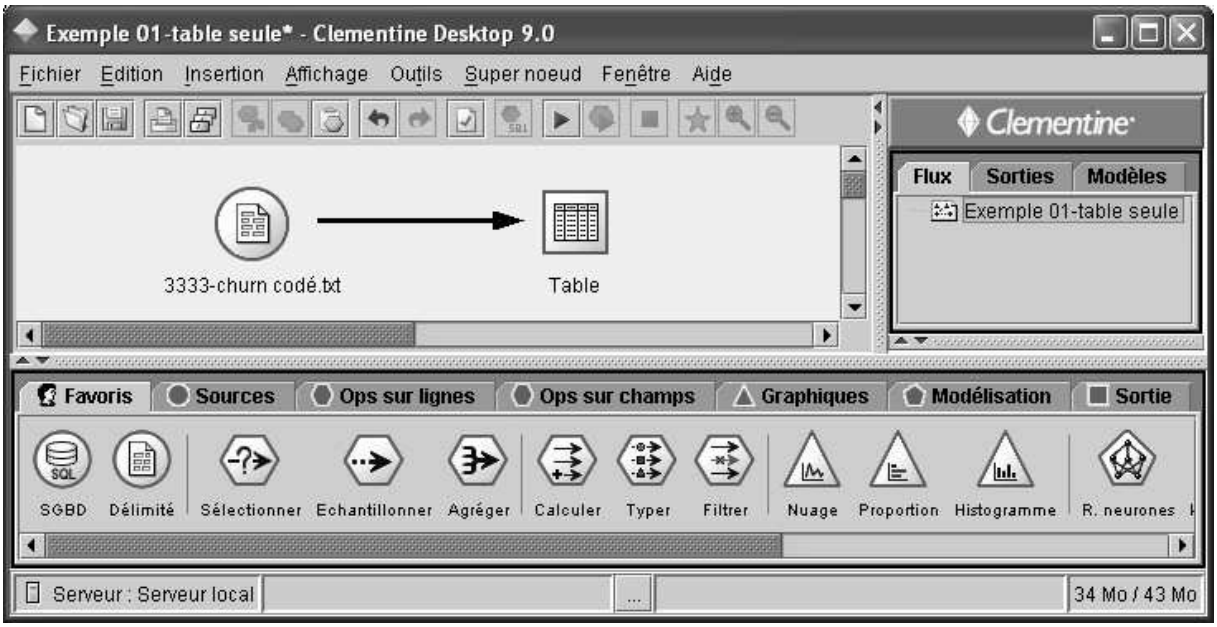
Dans tous les cas, le fichier d'origine doit être conservé intact.

Dictionnaire des données

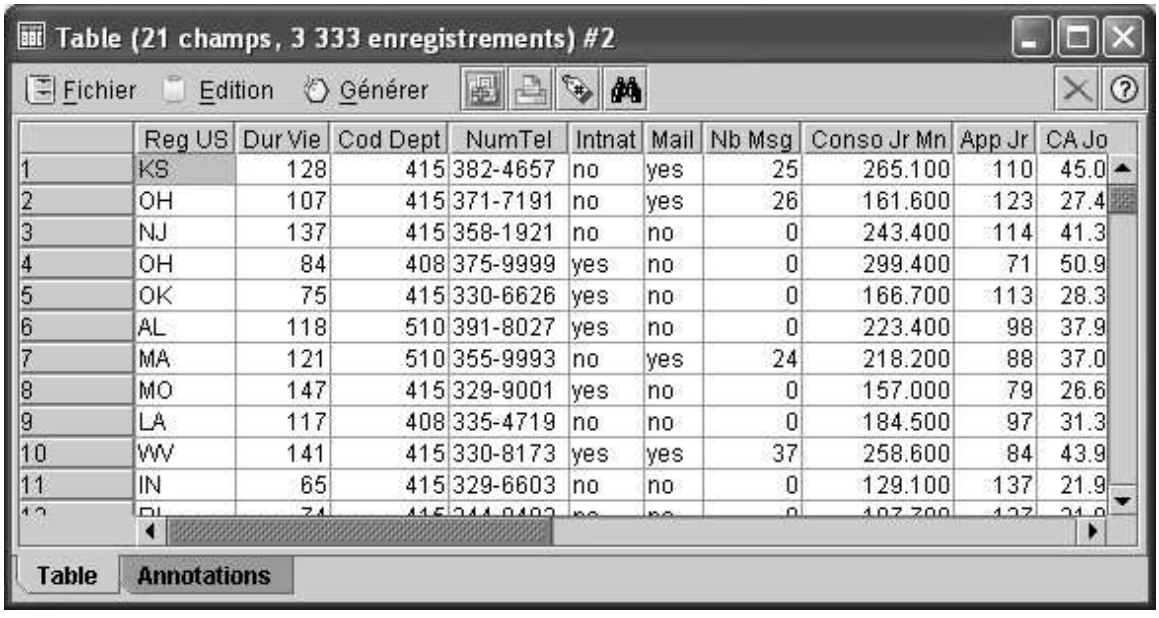
N°	Nom de la variable	Signification de la variable	Type de variable : Catégorielle ou autre	Autres caractéristiques
1	Reg US	Un code pour un État des États-Unis.	Catégorielle	2 caractères
2	Dur Vie	Durée de vie du compte	Entier	Unité à définir. Probablement la semaine.
3	Cod Dept	Numéro du département dans l'État (la Reg US)	Catégorielle	Entier
4	NumTel	Numéro de téléphone servant d'identifiant du client		
5	Intnat	Option pour les appels à l'international	Catégorielle : oui ou non	
6	Mail	Option pour la messagerie vocale	Catégorielle : oui ou non	
7	Nb Msg	Nombre de messages	Entier	
8	Conso Jr Mn	Consommation de la journée en minutes	Donnée continue : réel	
9	App Jr	Nombre d'appels dans la journée	Entier	
10	CA Jour	Chiffre d'affaires dans la journée	Donnée continue : réel	Fonction du nombre d'appels et de la durée de consommation
11	Conso Sr Mn	Consommation de la soirée en minutes	Donnée continue : réel	
12	App Sr	Nombre d'appels dans la soirée	Entier	
13	CA Soirée	Chiffre d'affaires dans la soirée	Donnée continue : réel	Cf. CA Jour

14	Conso Nt Mn	Consommation de la nuit en minutes	Donnée continue : réel	
15	App Nt	Nombre d'appels dans la nuit	Entier	
16	CA Nuit	Chiffre d'affaires dans la nuit	Donnée continue : réel	Cf. CA Jour
17	Conso Int Mn	Consommation à l'international en minutes	Donnée continue : réel	
18	App Int	Nombre d'appels à l'international	Entier	
19	CA Internat	Chiffre d'affaires à l'international	Donnée continue : réel	Cf. CA Jour
20	Serv Cli	Nombre d'appels au service client	Entier	
21	Churn?	Nous dit si le client a quitté la société ou pas	Catégorielle : oui ou non	

Clementine

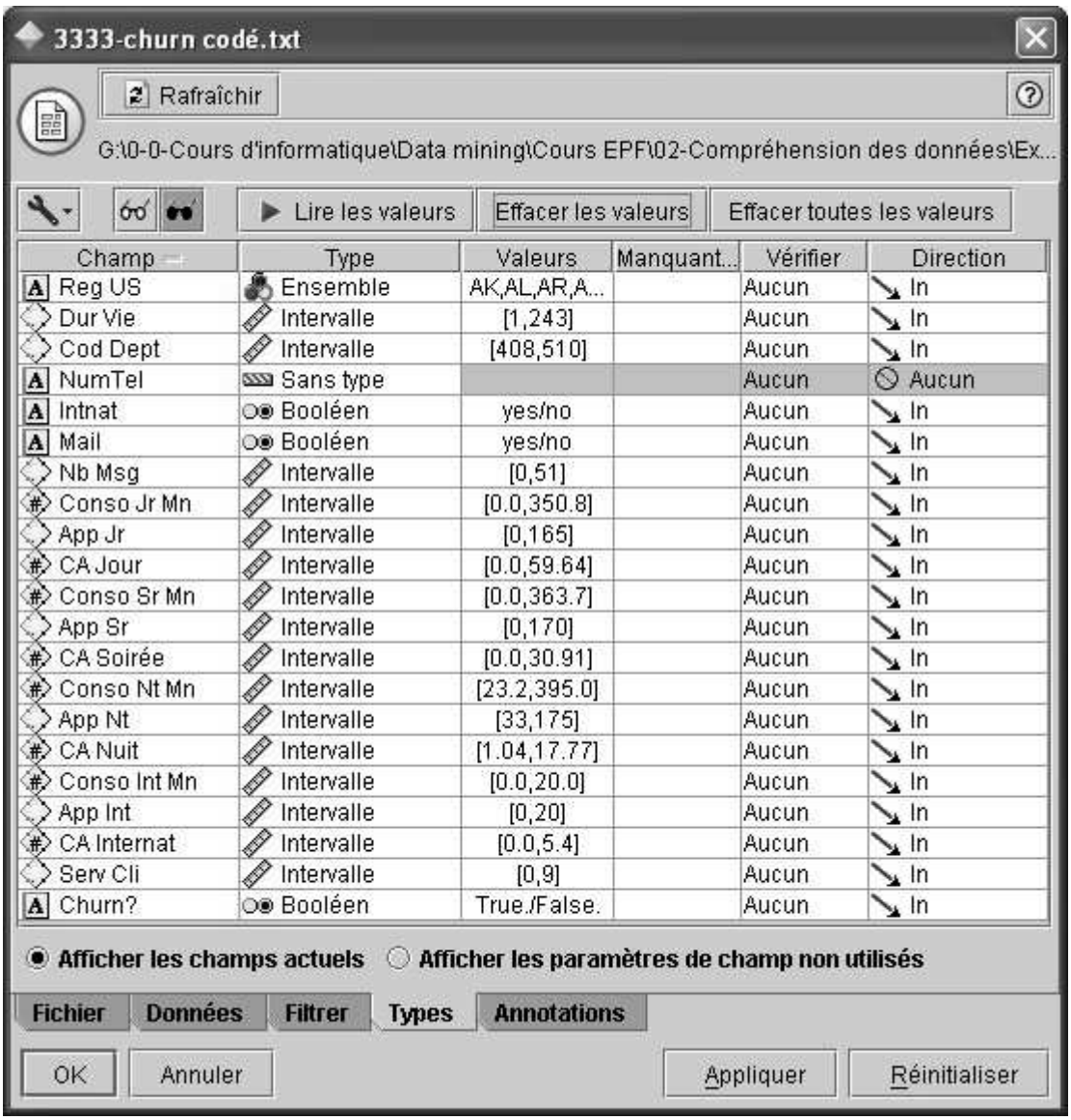


Flux constitué de deux nœuds



	Reg US	Dur Vie	Cod Dept	NumTel	Intnat	Mail	Nb Msg	Conso Jr Mn	App Jr	CA Jo
1	KS	128	415	382-4657	no	yes	25	265.100	110	45.0
2	OH	107	415	371-7191	no	yes	26	161.600	123	27.4
3	NJ	137	415	358-1921	no	no	0	243.400	114	41.3
4	OH	84	408	375-9999	yes	no	0	299.400	71	50.9
5	OK	75	415	330-6626	yes	no	0	166.700	113	28.3
6	AL	118	510	391-8027	yes	no	0	223.400	98	37.9
7	MA	121	510	355-9993	no	yes	24	218.200	88	37.0
8	MO	147	415	329-9001	yes	no	0	157.000	79	26.6
9	LA	117	408	335-4719	no	no	0	184.500	97	31.3
10	WV	141	415	330-8173	yes	yes	37	258.600	84	43.9
11	IN	65	415	329-6603	no	no	0	129.100	137	21.9

Résultat du flux : tableau de données



Liste des attributs du tableau de données

A ce niveau, le logiciel fournit le type, les valeurs possibles, le minimum et le maximum.

Le type intervalle est un type numérique.

On constate que le numéro de téléphone est « sans type » et qu'il n'y a pas de liste de valeurs proposée.

3 : Statistiques par variable

Examen des principales caractéristiques statistiques

Il s'agit de faire une investigation statistique des données pour se familiariser avec les données.

Il faut particulièrement s'intéresser à la répartition des valeurs et aux ordres de grandeurs.

On utilisera autant que possibles des méthodes graphiques.

Les outils logiciels sont plus ou moins performants pour obtenir rapidement les informations utiles.

Clementine

Audit des données de [20 champs] #5

Fichier Edition Générer

Champ	Graphique éch...	Type	Min	Max	Moyenne	Ecart-type	Asymétrie	Médiane d...	Mode	Unique	Valide
Appels au service client		Intervalle	0	9	1.563	1.315	1.091	1.000	--	--	3333
Chiffre daffaires de jour		Intervalle	0.000	59.640	30.562	9.259	-0.029	30.470	--	--	3333
Chiffre daffaires de linternational		Intervalle	0.000	5.400	2.765	0.754	-0.245	2.750	--	--	3333
Chiffre daffaires de nuit		Intervalle	1.040	17.770	9.039	2.276	0.009	9.115	--	--	3333
Chiffre daffaires de soirée		Intervalle	0.000	30.910	17.084	4.311	-0.024	17.270	--	--	3333
Churn?		Booléen	--	--	--	--	--	--	False.	2	3333
Code de département		Intervalle	408	510	437.182	42.371	1.127	415.000	--	--	3333
Conso de jour en minutes		Intervalle	0.000	350.800	179.775	54.467	-0.029	179.250	--	--	3333

Audit | Annotations

Résultats de l'analyse statistique

Il faut observer la répartition (l'histogramme) et les bilans statistiques.

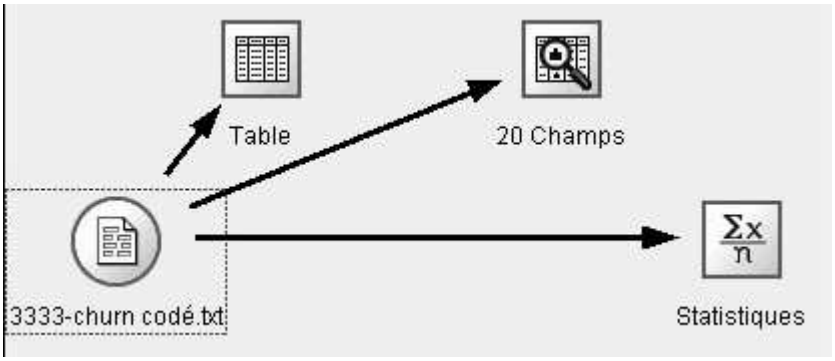


Schéma du flux Clémentine

Comptage	3333
Moyenne	101.065
Somme	336849
Min	1
Max	243
Intervalle	242
Variance	1585.800
Ecart-type	39.822
Erreur standard de la moyenne	0.690
Médiane	101.000
Mode	105.000

Statistiques détaillées

Attention

- ➔ La vitesse avec laquelle on produit les résultats (grâce aux logiciels qui permettent de faire les calculs facilement), ne doit pas faire oublier que l'observation des résultats est un travail qui demande du temps, de la précision et de la concentration. Il s'agit de se demander à chaque fois quel est le sens de l'information observée.
- ➔ Le but n'est pas de produire de l'information, mais bien de la comprendre pour pouvoir prendre une décision.

Rappels de statistique : tendance centrale et dispersion

La tendance centrale et la dispersion sont deux notions centrales de la statistique.

La tendance centrale

La moyenne, la médiane et le mode : ce sont trois informations qui caractérisent la tendance centrale d'un ensemble de données. Quand la moyenne, la médiane et le mode sont à peu près les mêmes, c'est un signe de la symétrie de la distribution des valeurs.

La médiane : c'est la valeur de la variable au 50^{e} centile (valeur qui sépare l'effectif total en deux parties égales). Ou encore, valeur du milieu lorsque les valeurs sont classées dans un ordre ascendant. La médiane est plus résistante que la moyenne à la présence de valeurs aberrantes.

Le mode : valeur de la variable qui apparaît le plus souvent dans la distribution. Si plusieurs valeurs distinctes apparaissent avec des fréquences égales, l'ensemble des variables ne possède pas de mode : on dit alors qu'il est plurimodal (bimodal, trimodal, etc.). Le mode peut s'appliquer pour les variables numériques ou catégorielles. Il n'est pas forcément lié à la valeur centrale.

La dispersion

La variance et l'écart-type : ce sont deux informations qui caractérisent la dispersion d'un ensemble de données.

La variance : moyenne des déviations au carré de chaque variable par rapport à la moyenne de l'ensemble des variables (somme((xi - moy(x))²) / (n-1)).

L'écart-type (déviation standard) : c'est la racine carrée de la variance. Il est aussi appelé : déviation standard. Il peut être interprété comme la distance « typique » (et pas moyenne) entre une variable et la moyenne.

Il existe beaucoup d'autres informations sur la dispersion : l'écart moyen, etc.

Remarque

Si les informations de tendance centrale sont clairement intuitives, celles de dispersion le sont nettement moins. L'écart moyen est l'information de dispersion la plus intuitive (moyenne des écarts par rapport à la moyenne). Cependant, ce n'est pas la plus utilisée.

Attention : la tendance centrale ne suffit pas à résumer une variable

Les mesures de valeurs centrales ne suffisent pas à résumer une variable efficacement. Deux variables peuvent avoir les mêmes valeurs centrales mais des répartitions très différentes :

Soit les 3 variables suivantes :

Variable 1	Variable 2	Variable 3
1	7	1
11	8	7
11	11	11
11	11	15
16	13	16

On obtient les résultats statistiques suivants :

Informations sur la tendance centrale			
	Variable 1	**Variable 2**	**Variable 3**
Moyenne	10.000	10.000	10.000
Somme	50	50	50
Médiane	11.000	11.000	11.000
Mode	11.000	11.000	Pluri-modale

Informations sur la dispersion			
	Variable 1	**Variable 2**	**Variable 2**
Ecart-type	5.477	2.449	6.164
Min	1	7	1
Max	16	13	16
Intervalle	15	6	15

On voit que la tendance centrale est la même pour les trois variables (notons que la somme est une information de la tendance centrale).

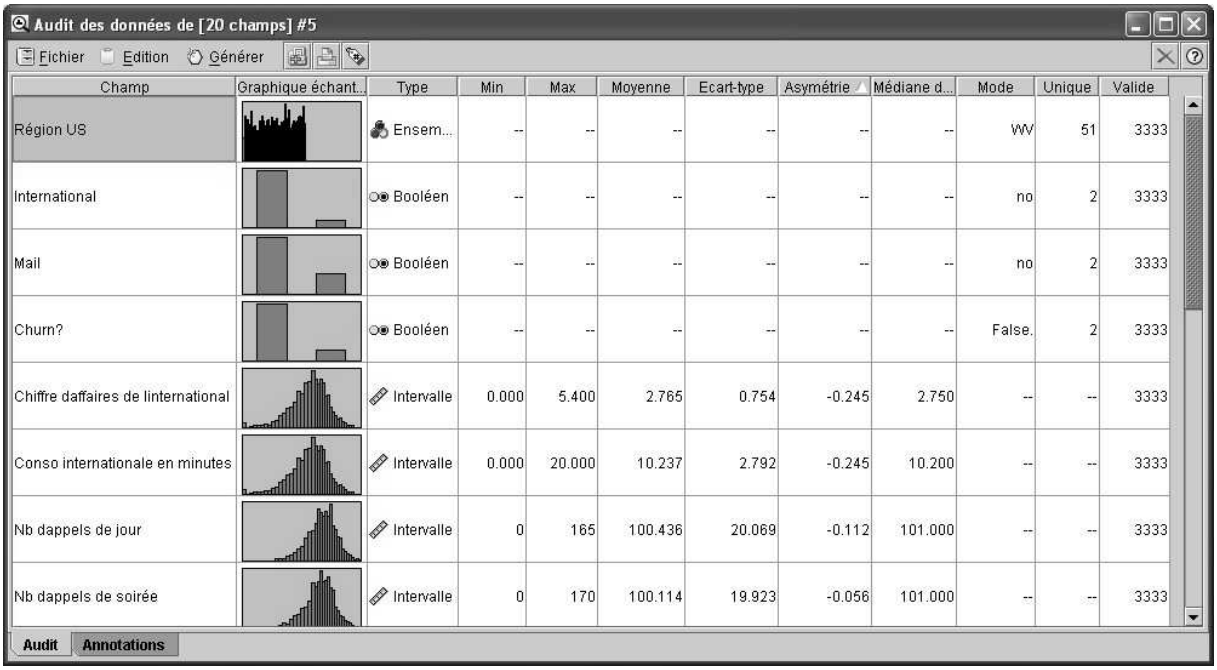
On voit par contre que la dispersion est plus marquée dans le premier cas que dans le deuxième : plus d'écart-type et plus d'intervalle.

Les dispersions du premier et du troisième cas semblent assez proches, pourtant, la répartition est quand même assez différente.

Conclusion : la tendance centrale et la dispersion ne suffisent pas à résumer les variables. Les graphiques sont aussi utiles. On peut aussi utiliser d'autres indicateurs de dispersion comme

Eléments d'interprétation

Premier exemple



Champ	Graphique échant...	Type	Min	Max	Moyenne	Ecart-type	Asymétrie	Médiane d...	Mode	Unique	Valide
Région US		Ensem...	--	--	--	--	--	--	WV	51	3333
International		Booléen	--	--	--	--	--	--	no	2	3333
Mail		Booléen	--	--	--	--	--	--	no	2	3333
Churn?		Booléen	--	--	--	--	--	--	False.	2	3333
Chiffre daffaires de linternational		Intervalle	0.000	5.400	2.765	0.754	-0.245	2.750	--	--	3333
Conso internationale en minutes		Intervalle	0.000	20.000	10.237	2.792	-0.245	10.200	--	--	3333
Nb dappels de jour		Intervalle	0	165	100.436	20.069	-0.112	101.000	--	--	3333
Nb dappels de soirée		Intervalle	0	170	100.114	19.923	-0.056	101.000	--	--	3333

A partir du tableau de résultats ci-dessus, on a intérêt à le trier par type ou par asymétrie.

Les deux grands types de variables : numériques et catégorielles conduisent à deux classes de résultats différentes.

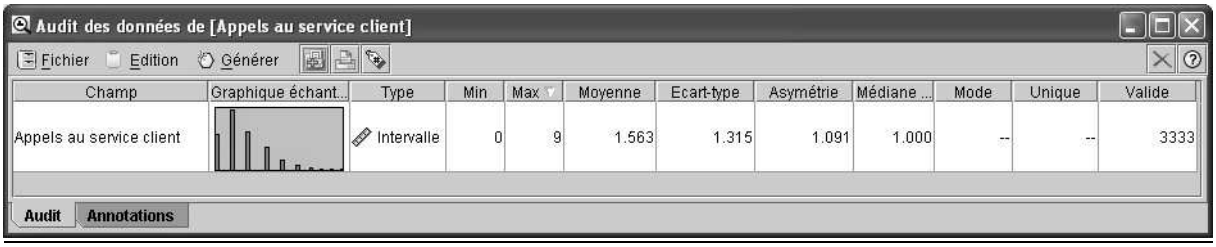
Unique concerne les variables catégorielles (discrètes). Ca donne le nombre de valeurs possibles dans la catégorie.

L'asymétrie concerne les variables numériques (continues). Une répartition parfaitement symétrique a une valeur d'asymétrie nulle. Une valeur négative signifie une asymétrie vers la gauche, une asymétrie positive signifie une asymétrie vers la droite.

Aide

En cliquant sur le « ? » en haut à droite de la fenêtre, Clementine fournit une aide pour l'interprétation et le paramétrage du résultat.

Cas des appels au service client



Les valeurs sont réparties de 0 à 9. Mais la médiane est à 1.0 : donc la moitié au moins des clients ont fait 0 ou 1 appel. Le mode est à 1 : c'est donc le plus présent. La moyenne est à 1,5 : donc la courbe n'est pas symétrique autour du 1 mais décalée vers le 9. On peut le constater sur l'histogramme.

La tendance centrale ne suffit pas à résumer une variable

Attention : les mesures de valeurs centrales ne suffisent pas à résumer une variable efficacement. Deux variables peuvent avoir les mêmes valeurs centrales mais des répartitions très différentes :

Considérons, par exemple, les 3 variables suivantes :

Variable 1	Variable 2	Variable 3
1	7	1
11	8	7
11	11	11
11	11	15
16	13	16

On obtient les résultats statistiques suivants :

Informations sur la tendance centrale			
	Variable 1	**Variable 2**	**Variable 3**
Moyenne	10.000	10.000	10.000
Somme	50	50	50
Médiane	11.000	11.000	11.000
Mode	11.000	11.000	Multiple

Informations sur la dispersion			
	Variable 1	**Variable 2**	**Variable 2**
Ecart-type	5.477	2.449	6.164
Min	1	7	1
Max	16	13	16
Intervalle	15	6	15

On voit que la tendance centrale est la même pour les trois variables (notons que la somme est une information de la tendance centrale).

On voit par contre que la dispersion est plus marquée dans le premier cas que dans le deuxième : plus d'écart-type et plus d'intervalle.

Les dispersions du premier et du troisième cas semblent assez proches, pourtant, la répartition est quand même assez différente.

Conclusion : la tendance centrale et la dispersion ne suffisent pas à résumer les variables. Les graphiques sont aussi utiles. On peut aussi utiliser d'autres indicateurs de dispersion comme l'écart interquartile ou la moyenne de l'écart absolu.

Histogramme et proportion

Histogrammes

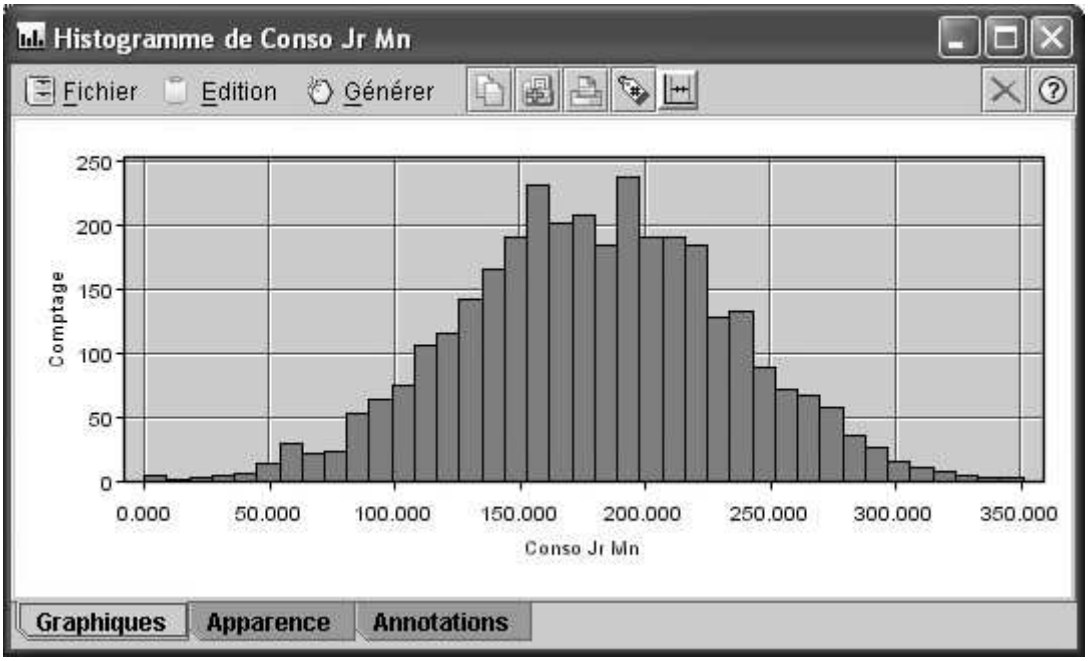
Les histogrammes concernent les **variables continues.**

L'histogramme compte le nombre d'occurrences pour une valeur donnée de la variable.

En cas de type réel, le type est discrétisé.

On peut aussi choisir de discrétiser un type entier.

Dans tous les cas, on peut choisir la taille de l'intervalle de compte.



Histogramme de la consommation par jour

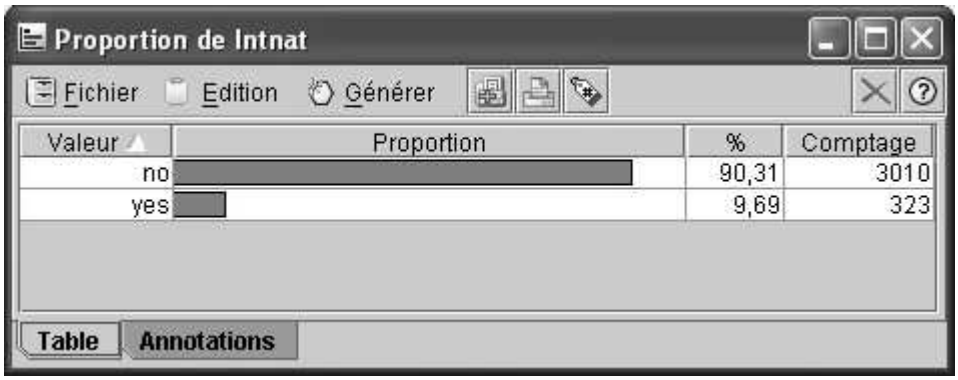
Proportions

Les proportions concernent les **variables catégorielles.**

La proportion compte le nombre d'occurrences pour une valeur donnée de la variable et la proportion de cette occurrence.

La différence avec l'histogramme, les valeurs possibles pour la variable considérée sont mises à la verticale. Le résultat présente le nombre et le pourcentage.

Etant donné que l'ordre des valeurs possible est rarement significatif, on ne retrouve pas de répartition sous la forme d'une courbe de Gauss.

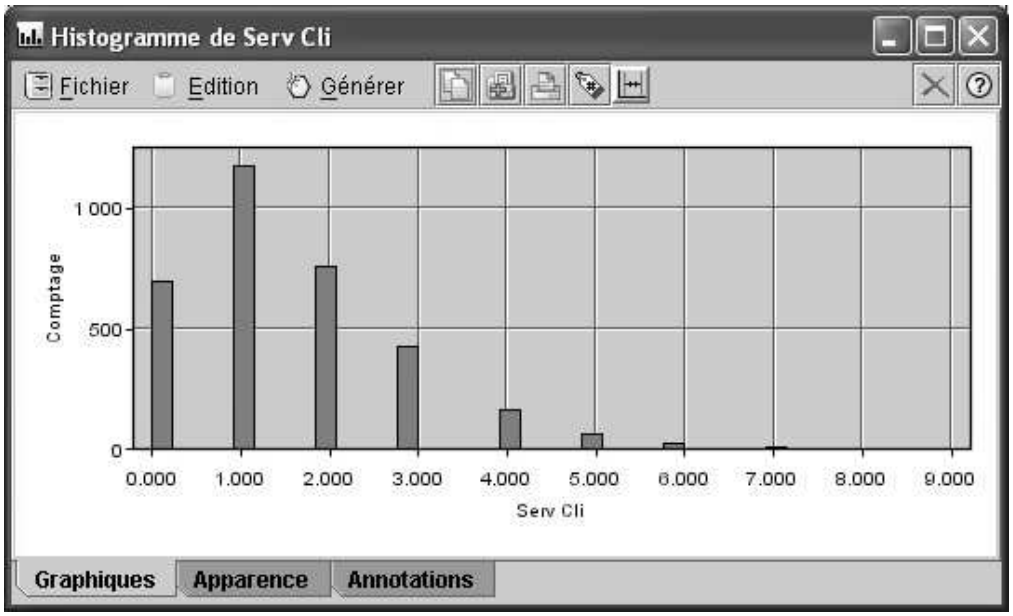


Proportion de contrats à l'international

Passage d'un histogramme à une proportion

On peut toujours discrétiser une variable numérique et la considérer ensuite comme une variable catégorielle.

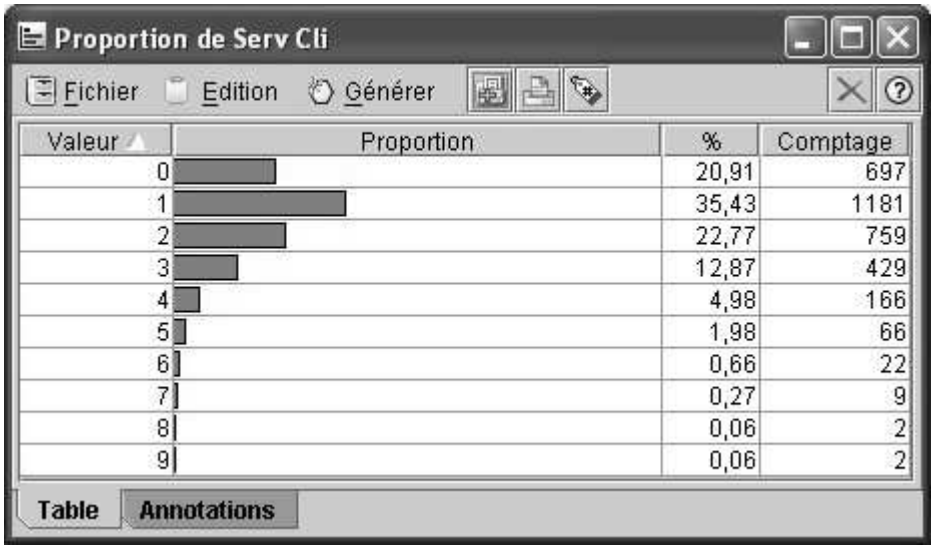
On peut aussi considérer les variables numérique entière comme des variables catégorielles.



Histogramme du nombre d'appels au service client

La variable « nombre d'appels au service client » est une variable numérique entière.

On peut changer son type et en faire une variable catégorielle dont les valeurs possibles vont de 0 à 9 :



Valeur	Proportion	%	Comptage
0		20,91	697
1		35,43	1181
2		22,77	759
3		12,87	429
4		4,98	166
5		1,98	66
6		0,66	22
7		0,27	9
8		0,06	2
9		0,06	2

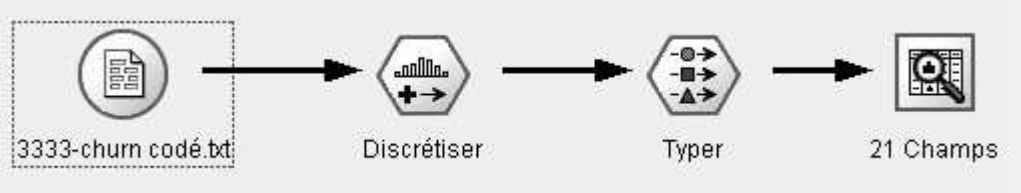
Proportion du nombre d'appels au service client

Clementine

L'audit des données de Clémentine permet d'accéder directement aux histogrammes et aux proportions.

Pour changer les types, il faudra utiliser le nœud « typer ».

Pour discrétiser un type, il faudra utiliser le nœud « discrétiser ».



Flux Clementine permettant de discrétiser et de typer.

4 : Suppression de la clé primaire

Présentation

Dans les bases de données relationnelles, le clé primaire est un attribut obligatoire et unique qui détermine tous les autres attributs de la table.

Les clés primaires sont des variables qui ne sont **ni numériques** (même si c'est un chiffre, car la moyenne de ces chiffres n'a pas de sens), **ni catégorielles** (pour qu'il y ait catégorie, il faut qu'il y ait plusieurs objets dans la catégorie).

Elles n'ont donc aucun intérêt pour le data mining puisqu'on ne peut ni en tirer de connaissances statistiques (la moyenne des numéros de référence ne veut rien dire), ni l'utiliser pour des regroupements puisque c'est la valeur qui permet de distinguer une ligne d'une autre dans la table.

Les clés primaires et tous les attributs obligatoires et uniques doivent être supprimés du modèle.

Notion de dépendance fonctionnelle

Dans la théorie des bases de données relationnelles, il existe la notion de dépendance fonctionnelle.

On dit qu'un attribut (ou variable) B dépend fonctionnellement d'un attribut A si, pour toute valeur de A, la valeur de B correspondante est déterminée.

La clé primaire est un cas typique de dépendance fonctionnelle.

Tous les attributs d'une table dépendent fonctionnellement de la clé primaire de la table.

Clé primaire ---> Attributs

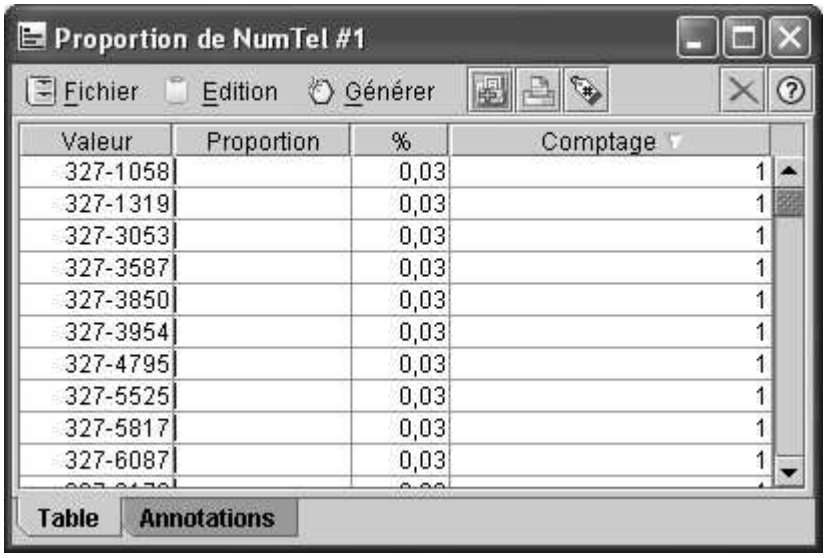
Application

Clementine

Clementine supprime automatiquement les variables clés primaires.

On peut forcer Clementine à garder les variables clé primaires en imposant leur type : un ensemble.

En observant la répartition des valeurs, on constate que chaque donnée de numéro de téléphone n'est représenté qu'une seule fois.



Proportion des valeurs de numéro de téléphone (comptage)

5 : Nettoyage des données : valeurs manquantes et aberrantes

Présentation

Les différents types d'anomalie

Dans les données qu'on récupère, il peut y avoir des anomalies. Elles sont de trois sortes :

- Les données manquantes : pour une variable donnée, certaines données peuvent être manquantes
- Les données aberrantes : pour une variable donnée, certaines données peuvent être aberrantes (trop grande ou trop petite, pas dans la fourchette des valeurs possibles).
- Les données conjointement incohérentes : les données d'une variable peuvent être incohérentes par rapport aux données d'une autre variable.

Il va falloir « nettoyer » le fichier de départ.

De l'importance d'un bon nettoyage

Quelques données incohérentes suffisent à fausser les résultats des algorithmes de modélisation et donc les prédictions qu'on peut en tirer. Il est donc absolument nécessaire de faire le nettoyage.

Exemple

Id Client	CP	Sexe	Revenu	Âge	Marié	Montant
1001	75000	M	75000	C	M	5000
1002	4000	F	-40000	40	V	4000
1003	92100		1000000	45	C	7000
1004	6260	M	50000	0	C	1000
1005	29000	F	99999	30	D	3000

D'après Des données à la connaissance, de Daniel T. Larose, p. 26

Que peut-on constater dans ce tableau ?

CP : 4000 ? 6260 ? 4000 est le code postal de Liège en Belgique. 06260 est le code postal de La Croix sur Roudoule. Que dire de 75000 : normalement, il faut préciser les arrondissements.

Sexe : il y a un champ manquant.

Revenu : -40000 ? Négatif ! 1000000 ? C'est beaucoup. 99999 : c'est très précis quand tous les autres sont par pas de 5000. Il y a aussi le problème de la monnaie.

Age : C ? 0 ?

Marié : que signifient les symboles ?

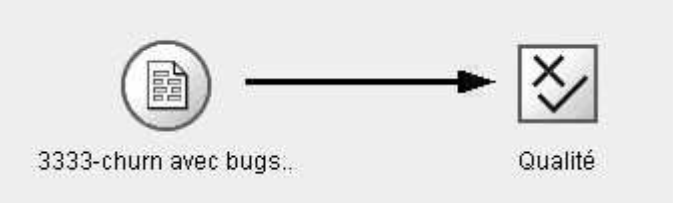
Montant : comme pour revenu, le problème de la monnaie.

Découvrir les données manquantes

Pour découvrir les données manquantes, il suffit de trier chaque variable. Les données manquantes apparaîtront soit au début, soit à la fin.

Clementine

Un nœud permet de faire l'analyse de la qualité : le nœud « qualité ».



Flux Clementine d'analyse de la qualité

On obtient les résultats suivants :

Qualité de [21 champs] #2

Fichier Edition Générer

Champ	% terminé(s)	Enregistrements valides	Valeur nulle	Chaîne vide	Blanc	Valeur non renseignée
App Int	100	3333	0	0	0	0
App Jr	100	3333	0	0	0	0
App Nt	100	3333	0	0	0	0
App Sr	100	3333	0	0	0	0
CA Internat	100	3333	0	0	0	0
CA Jour	100	3333	0	0	0	0
CA Nuit	100	3333	0	0	0	0
CA Soirée	100	3333	0	0	0	0
Churn?	100	3333	0	0	0	0
Cod Dept	100	3333	0	0	0	0
Conso Int Mn	100	3333	0	0	0	0
Conso Jr Mn	100	3333	0	0	0	0
Conso Nt Mn	100	3333	0	0	0	0
Conso Sr Mn	100	3333	0	0	0	0
Dur Vie	99,97	3332	1	0	0	0
Intnat	99,91	3330	0	3	3	0
Mail	100	3333	0	0	0	0
Nb Msg	100	3333	0	0	0	0
NumTel	100	3333	0	0	0	0
Reg US	100	3333	0	0	0	0
Serv Cli	100	3333	0	0	0	0

Qualité | Annotations

Découvrir les données aberrantes : méthodes graphiques

Principe

En observant les histogrammes de répartitions des données, on peut découvrir les données aberrantes.

Le principe empirique de la découverte consiste à rechercher des valeurs extrêmes : il faut regarder ce qui se passe aux extrémités de l'histogramme.

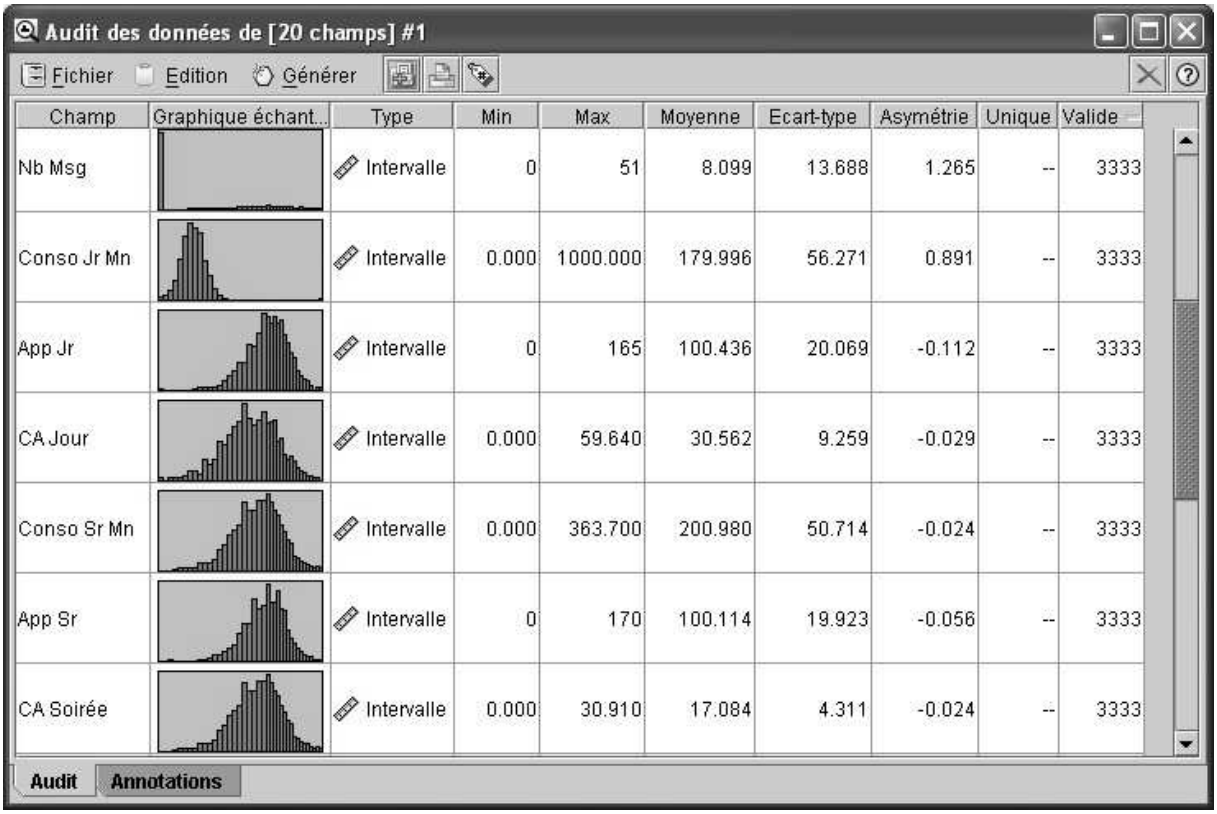
Le principe théorique est qu'une répartition suit le plus souvent une courbe de Gauss. Si on n'a pas une telle courbe, c'est suspect.

Sont suspects :

- Les valeurs isolées de part et d'autre de l'histogramme
- Les occurrences très faibles, quelle que soit leur position dans l'histogramme.
- Les histogrammes plats

Exemple

Soit la série d'histogramme suivante :

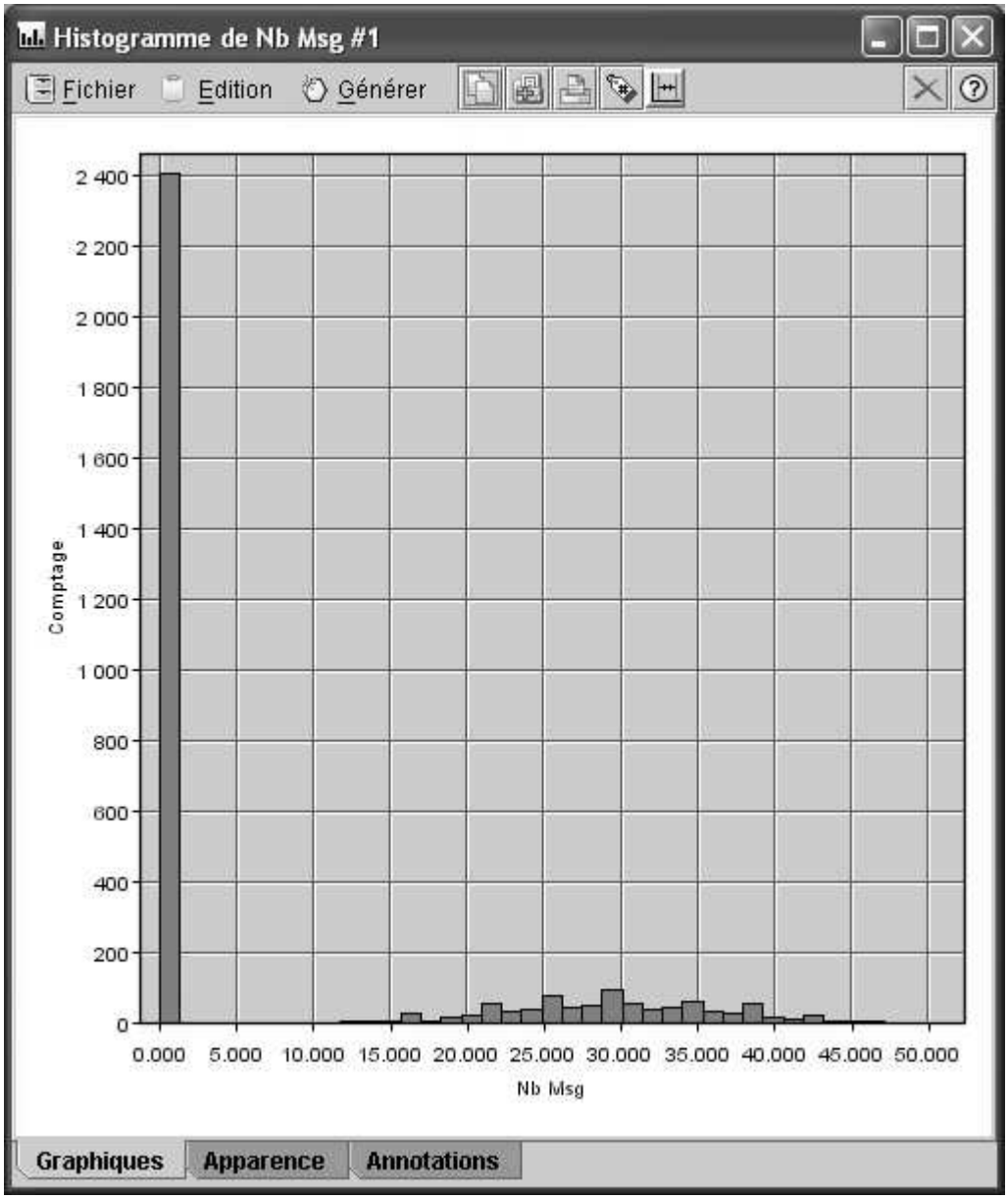


Champ	Graphique échant...	Type	Min	Max	Moyenne	Ecart-type	Asymétrie	Unique	Valide
Nb Msg		Intervalle	0	51	8.099	13.688	1.265	--	3333
Conso Jr Mn		Intervalle	0.000	1000.000	179.996	56.271	0.891	--	3333
App Jr		Intervalle	0	165	100.436	20.069	-0.112	--	3333
CA Jour		Intervalle	0.000	59.640	30.562	9.259	-0.029	--	3333
Conso Sr Mn		Intervalle	0.000	363.700	200.980	50.714	-0.024	--	3333
App Sr		Intervalle	0	170	100.114	19.923	-0.056	--	3333
CA Soirée		Intervalle	0.000	30.910	17.084	4.311	-0.024	--	3333

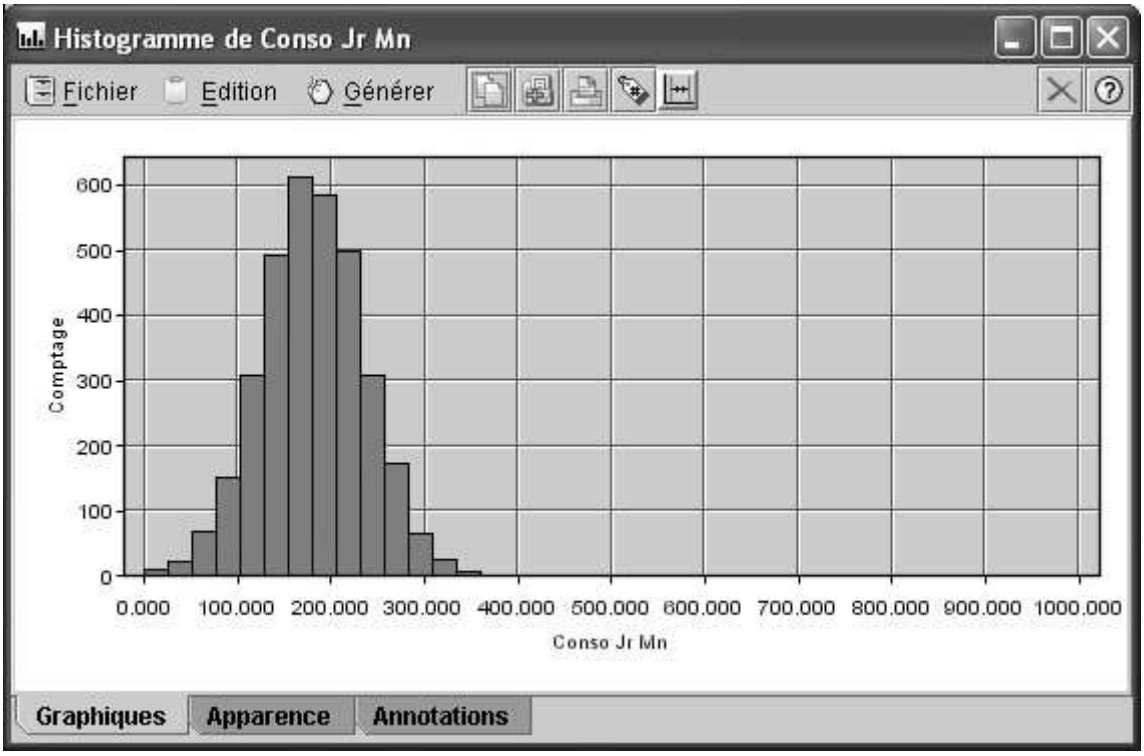
Dans cette série, on constate que :

- Il y a un pic à 0 pour le premier histogramme et une répartition peu lisible. L'asymétrie est forte
- Le deuxième histogramme est décalé à gauche. L'asymétrie est forte.
- Les 3ème et 6ème histogramme sont légèrement décalé à gauche. L'assymétrie est prononcée.
- Les autres histogrammes sont classiques. L'asymétrie est faible.

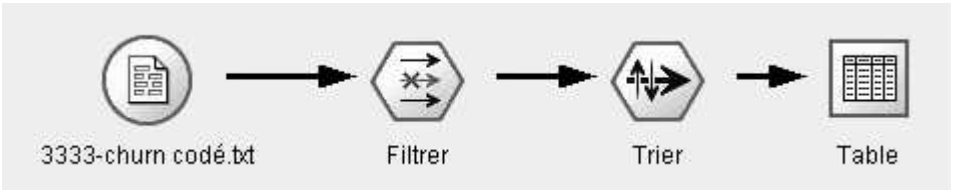
Dans le cas du premier histogramme (nb msg), en zoomant on constate qu'on a beaucoup de cas à 0 et une courbe de Gauss entre 10 et 50. D'un point de vue sémantique, c'est compréhensible : certains clients n'ont pas ou n'utilisent pas la messagerie ; pour ceux qui l'utilisent, la répartition suit une courbe de Gauss. Il n'y a donc pas de valeurs abérantes.



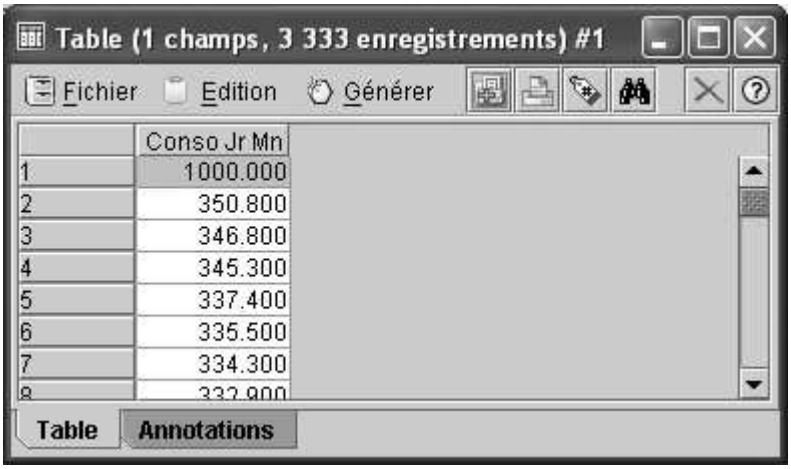
Dans le cas du second histogramme (conso jour mn), on constate la présence d'une courbe de Gauss classique et de valeurs allant jusqu'à 1000. Il semble donc bien que dans ce cas, on ait des valeurs aberrantes.



Pour préciser les choses, il faut afficher les valeurs de la variables triées en ordre décroissant.



Flux clémentine permettant de sélectionner une variable et de la trier



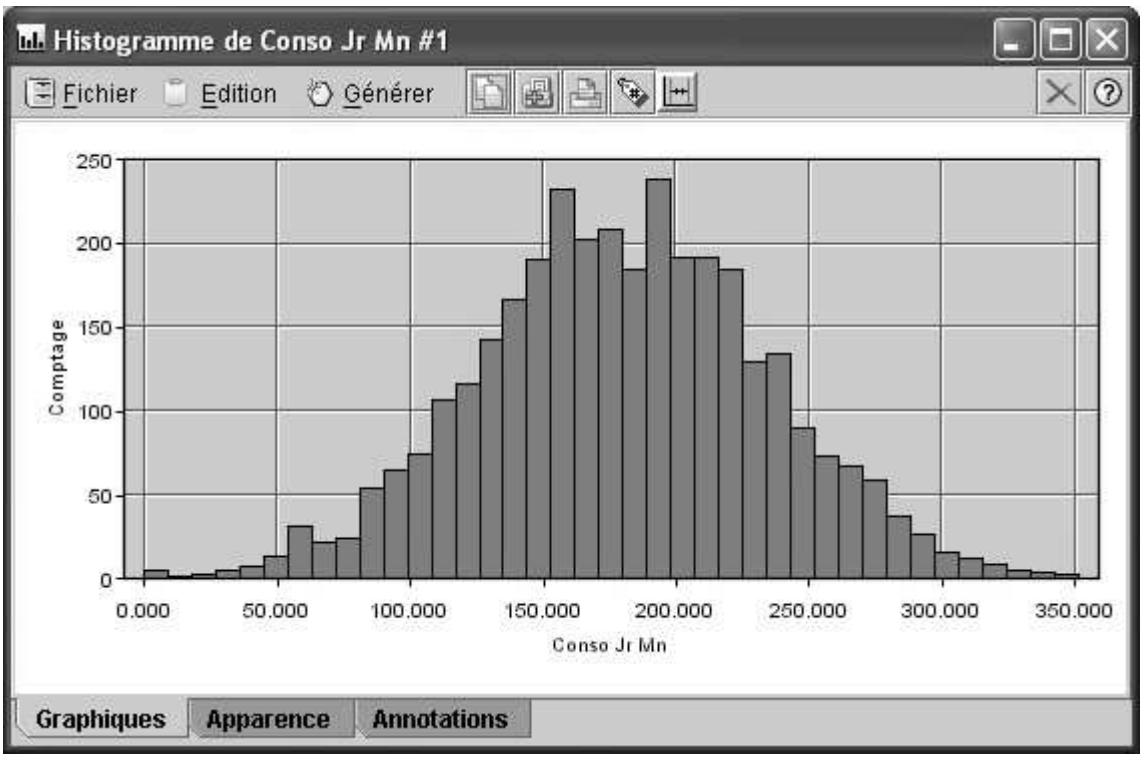
Valeurs de la variable, triées par ordre décroissant

On constate bien que 1000 est une valeur isolée. On peut faire l'hypothèse que c'est une valeur aberrante.

On peut supprimer l'individu correspondant de notre jeu de départ :

Flux clémentine permettant de sélectionner une variable et de la trier

On obtient alors l'histogramme suivant qui est totalement normalisé :



Que faire avec les données aberrantes ?

Trois solutions s'offrent à nous :

1 : On supprime l'individu. Le risque est de supprimer par la même occasion des valeurs utiles. Si la population est nombreuse, ce risque devient très faible.

2 : On remplace la valeur aberrante par une autre valeur : la valeur moyenne, une valeur prise au hasard dans la distribution des valeurs. Le risque est de produire une incohérence par rapport aux autres valeurs (aberration croisée).

3 : On réduit les valeurs possibles pour la valeur manquante du fait de l'existence de corrélations entre la variable à valeur manquante avec d'autres variables renseignées.

La meilleure solution est la 3ème si on peut arriver à un jeu de valeurs possibles très restreint, voir à une seule valeur possible.

Sinon, mieux vaut supprimer l'individu pour éviter de produire des aberrations croisées.

Découvrir les valeurs aberrantes : méthodes numériques

Test Z et triple écart-type

Le test Z consiste à échelonner les valeurs d'une variable en appliquant la formule suivante :

$$\text{test Z}(x) = (x - \text{moy}(X)) / \text{écart type}(X)$$

Cette formule donne des valeurs réparties autour de 0 (pour les x = moy (X) et permet de comparer l'écart de x par rapport à l'écart type[1].

Une règle empirique dit que :

Si abs (test Z (x)) > 3, alors x est probablement une valeur aberrante

C'est seulement une règle empirique, et donc une piste à explorer par d'autres moyens. C'est d'autant plus vrai que la moyenne et l'écart type sont deux indicateurs qui sont sensibles à la présence de valeurs aberrantes ! La méthode a donc un peu tendance à se mordre la queue !

La méthode des quartiles

C'est une méthode moins sensible à la présence de valeurs aberrantes.

La méthode consiste à trier la variable et à diviser ensuite l'ensemble des données en quatre ensembles contenant le quart des données.

Le premier quartile (Q1) : c'est le 25ème centile : la valeur de la variable au 25ème centile.

Le second quartile (Q2) : c'est le 50ème centile, c'est-à-dire la médiane.

Le troisième quartile (Q3) : est le 75ème centile.

L'amplitude interquartile (AIQ) : c'est Q3 – Q1

Une règle empirique dit que :

Si x < Q1 - 1,5 * AIQ, alors x est probablement une valeur aberrante.

Si x > Q3 + 1,5 * AIQ, alors x est probablement une valeur aberrante.

L'amplitude interquartile est une valeur peu sensible à la présence de valeurs aberrantes.

[1] L'écart-type, ou déviation standard, est égal à la racine carrée de la variance, qui est la moyenne des déviations au carré de chaque variable par rapport à la moyenne de l'ensemble des variables.

6 : Corrélations entre variables numériques : suppression des var. calculées

Principe général des corrélations

La notion de corrélation rejoint la notion de prévision de la modélisation.

Le principe général d'une corrélation, ou d'une prédiction, c'est :

Si une ou plusieurs variables valent tant, alors telle autre variable vaudra tant.

Les variables qui permettent la prédiction sont appelées « prédicteurs » ou « variables en entrée ».

La variable prédite est appelée « variable cible » ou « variable en sortie ».

La notion de corrélation rejoint la notion de dépendance fonctionnelle :

prédicteurs ---> variable cible

La corrélation la plus simple est la corrélation linéaire entre deux variables. Une simple équation de droite corrèle les deux variables.

variable cible = a * prédicteur + b

Principe des corrélations entre variables numériques

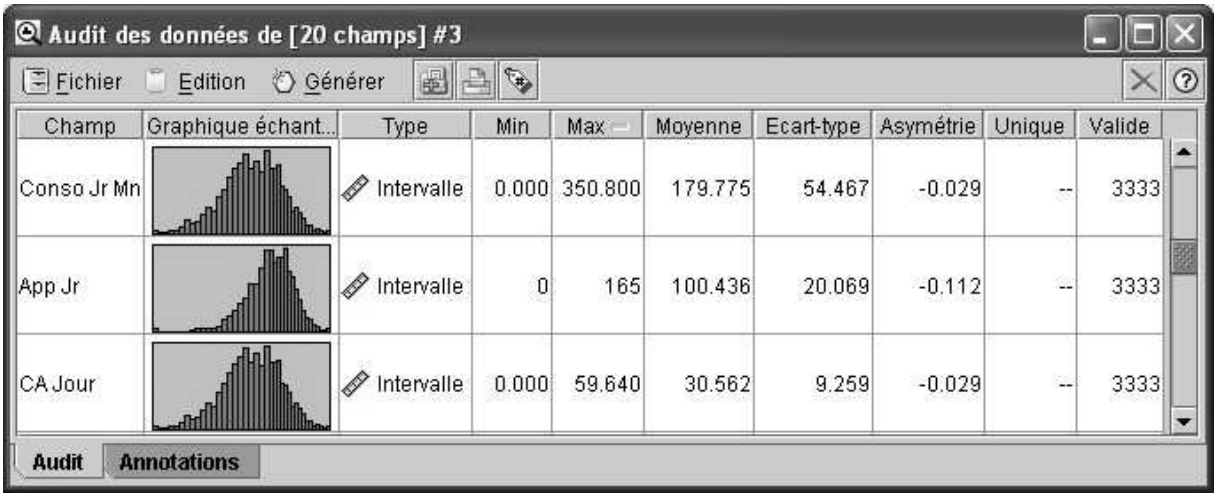
Trouver une corrélation entre deux variables numériques consiste à trouver l'équation de la courbe qui relie tous les données entre elles quand les variables sont classées en ordre croissant.

La corrélation linéaire correspond à l'équation la plus simple : une équation de droite du type aX + b = Y

Représentation graphique : en prenant une variable comme axe des X et une variable comme axe des Y, on peut placer chaque donnée sur le graphique et obtenir ainsi un « nuage de points » qui formera une ligne droite en cas de corrélation linéaire.

Répérer graphiquement les corrélations : les variables corrélées linéairement ont un histogramme de même forme. Elles ont aussi une asymétrie proche voir identique.

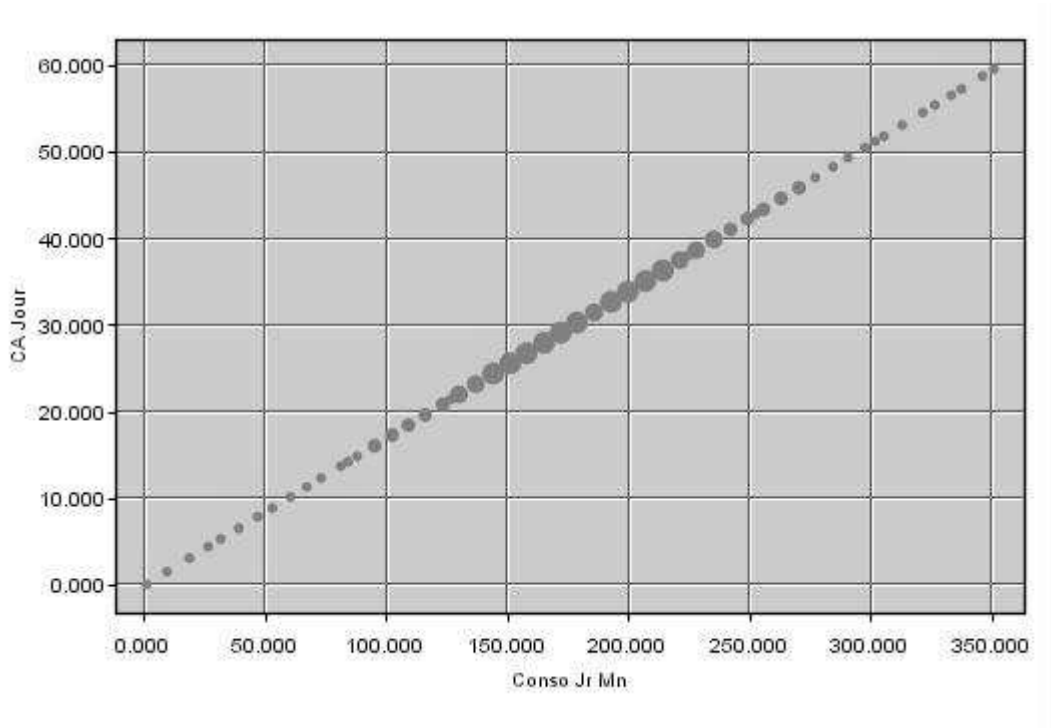
Exemple



Dans l'extrait d'audit ci-dessus, on constate que les histogrammes de « conso Jr Mn » et « CA Jour » sont très proches. On constate aussi que l'asymétrie est identique : -0.029.

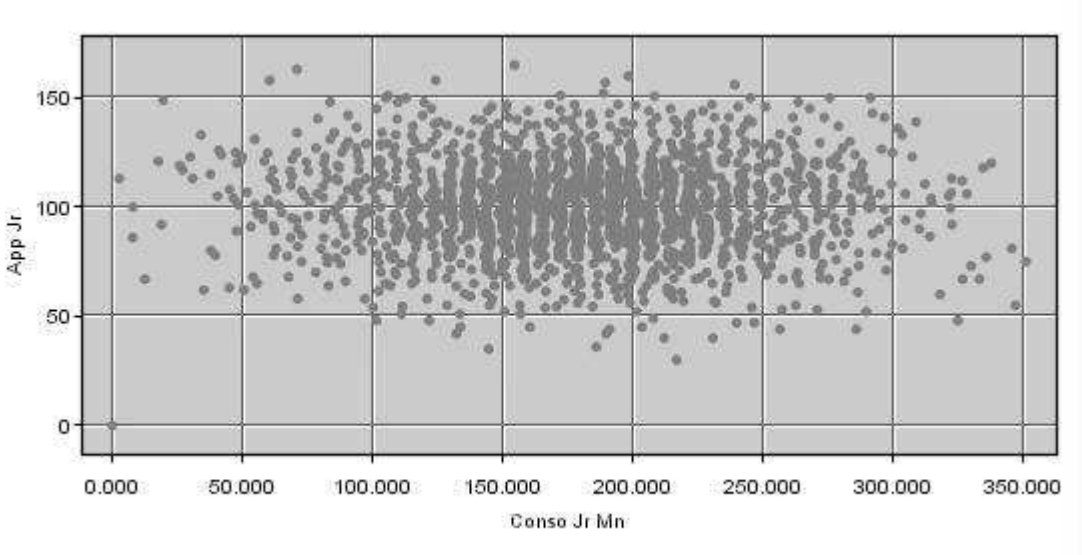
D'un point de vue sémantique, ça peut vouloir dire qu'il y a une corrélation linéaire entre la consommation et le chiffre d'affaire, ce qui semble assez normal.

On peut afficher le nuage de point entre les deux variables :

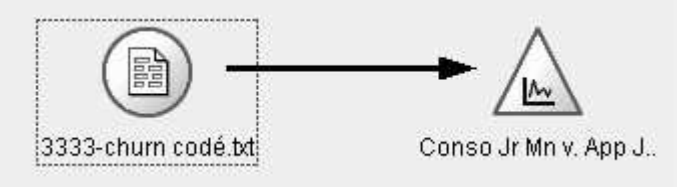


On obtient bien une droite.

En affichant le relation de points entre la Consommation par jour et le nombre d'appel par jour, on obtient le nuage suivant :



Il n'y a donc pas de corrélation linéaire entre ces deux variables.

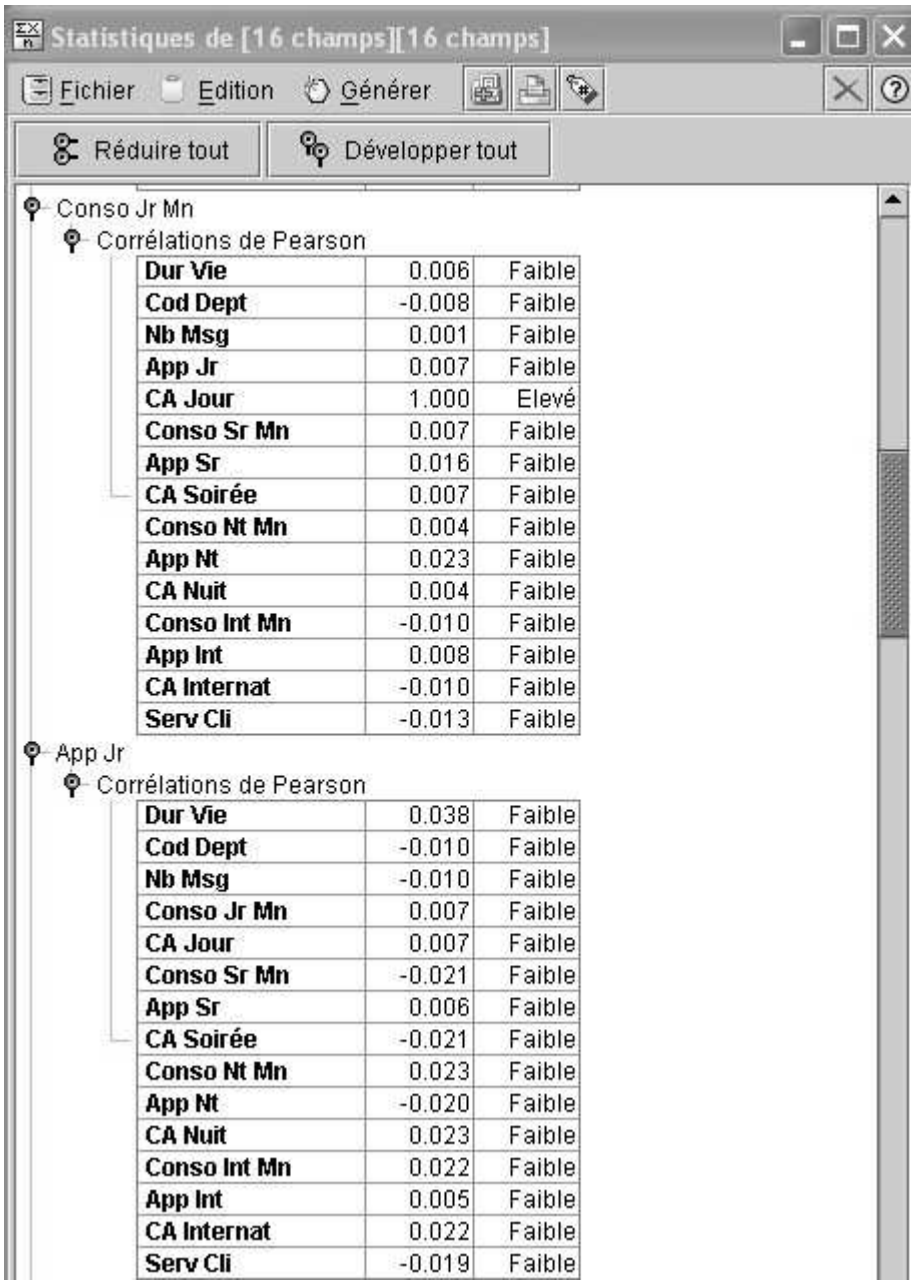


Flux clémentine d'affichage d'un nuage de points

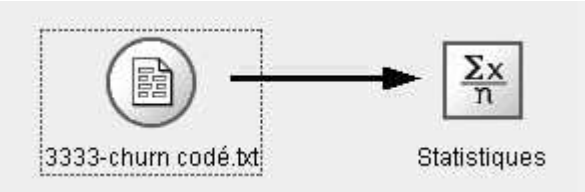
Analyse systématique

L'analyse intuitive n'est pas suffisante. On peut faire une analyse systématique de tous les couples de variables numériques possibles. Pour chaque couple, on calculera l'équation de droite de la régression linéaire.

Clementine



Le logiciel calcule le coefficient de corrélation. Plus il est proche de 1, plus il y a corrélation linéaire. Ici, on voit qu'il n'y a qu'un cas de corrélation linéaire.



Flux clémentine correspondant

Distinction entre corrélation mathématique et corrélation empirique

Corrélation mathématique

Une corrélation mathématique est une corrélation parfaite : la courbe reliant tous les points est parfaitement linéaire (par exemple, une droite). Une telle corrélation est rare dans les phénomènes naturels.

Une corrélation mathématique entre deux variables signifie qu'une variable est calculée artificiellement à partir d'une autre.

Dans notre exemple, le chiffre d'affaire par jour est calculé à partir de la consommation par jour.

Corrélation empirique

Une corrélation empirique est une corrélation imparfaite : la courbe reliant tous les points se rapproche d'une droite mais n'en est pas une.

Suppression des variables calculées

Toutes les variables mathématiquement corrélées à d'autres variables doivent être supprimée des modèles de data mining. En effet, du fait que ces variables dupliquent une information déjà existante, elles risquent de fausser le calcul des modèles de data mining et donc de conduire à des résultats erronés.

Quelle variable va-t-on garder ? La consommation ou le chiffre d'affaires ? Il vaut mieux garder la donnée première qui est la source de la seconde. Ici, c'est la consommation qui permet de générer le chiffre d'affaires à partir du coût et pas le contraire. On gardera donc la consommation.

Toutefois, ça dépend aussi de ce qu'on cherche à montrer dans la modélisation.

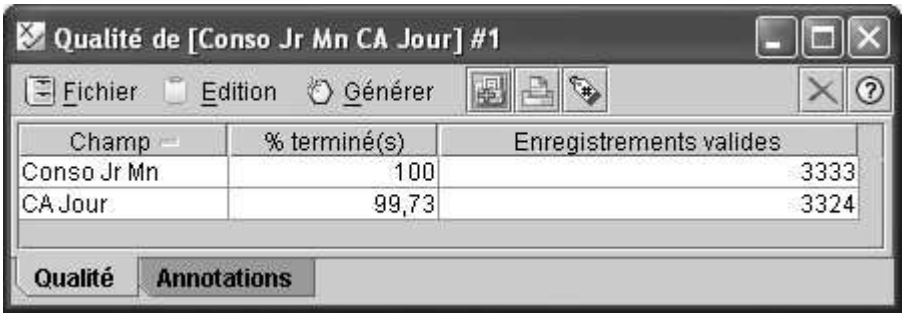
Calcul de l'équation de la droite

On peut calculer l'équation de droite correspondant à la corrélation linéaire.

Une telle équation est un modèle prédictif. On pourra, avec ce modèle, déterminer le Chiffre d'affaire (variable cible) à partir de la consommation (variable prédictive).

Clementine

On part de 9 données de « CA Jour » absentes :



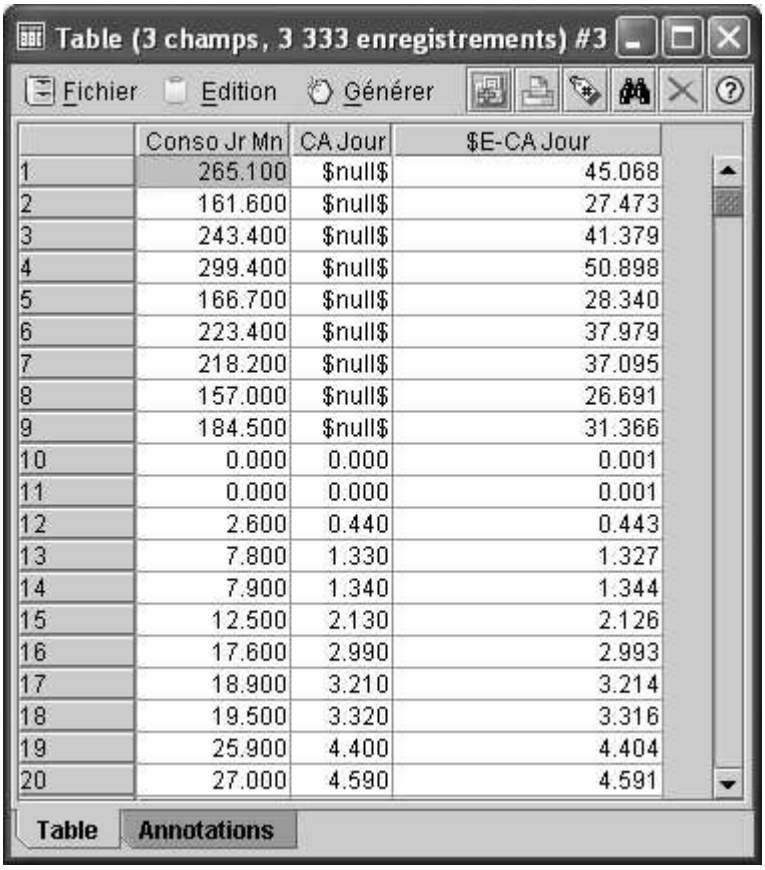
On calcule l'équation de régression linéaire :

CA Jour = 0,17 * Conso Jr Mn + 0,0006204

On pourra probablement se satisfaire de l'approximation suivante :

CA Jour = 0,17 * Conso Jr Mn

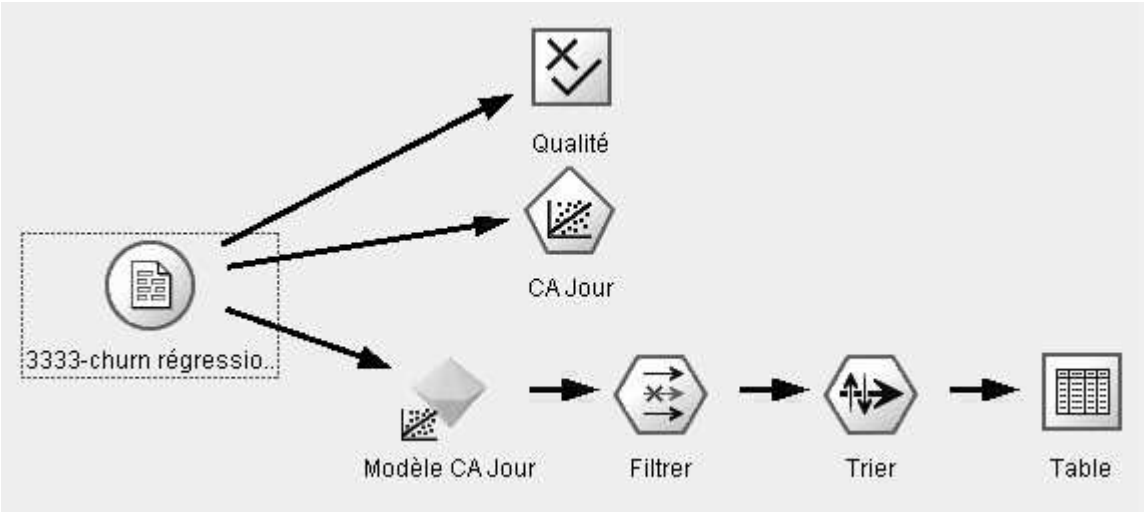
Cette équation est un modèle qui va permettre de calculer la valeur du CA pour les 9 valeurs non déterminées :



	Conso Jr Mn	CA Jour	$E-CA Jour
1	265.100	$null$	45.068
2	161.600	$null$	27.473
3	243.400	$null$	41.379
4	299.400	$null$	50.898
5	166.700	$null$	28.340
6	223.400	$null$	37.979
7	218.200	$null$	37.095
8	157.000	$null$	26.691
9	184.500	$null$	31.366
10	0.000	0.000	0.001
11	0.000	0.000	0.001
12	2.600	0.440	0.443
13	7.800	1.330	1.327
14	7.900	1.340	1.344
15	12.500	2.130	2.126
16	17.600	2.990	2.993
17	18.900	3.210	3.214
18	19.500	3.320	3.316
19	25.900	4.400	4.404
20	27.000	4.590	4.591

On constate qu'on retrouve les deux variables : « Conso Jr Min » et « CA Jour », mais aussi une variable calculée : « $E-CA Jour » : cette variable est calculée avec l'équation de droite proposée. On peut constater qu'on obtient une valeur pour les 9 valeurs non déterminées de « CA Jour », et que les valeurs déterminées de « CA Jour » correspondent à celles calculées par l'équation.

Bilan des flux clémentine :



:

Il y a un nœud « Qualité », un nœud « CA Jour » qui a permit de créer le modèle qui permet ensuite de créer un attribut calculé. On filtre pour n'avoir que Conso et CA, et on trie par CA pour faire apparaître en premier les CA non renseignés.

7 : Corrélations entre variables catégorielles

Présentation

On vient d'aborder les corrélations simples entre variables numériques.

Nous allons maintenant aborder les corrélations simples entre variables catégorielles.

On peut appliquer le même principe que pour les variables numériques : croiser chaque variable catégorielle avec toutes les autres variables catégorielles.

On peut aussi considérer une variable catégorielle particulière : la variable cible, c'est-à-dire celle sur laquelle on souhaite faire des prédictions.

Principe de la corrélation

Dans chaque catégorie de la variable prédictive, on va observer la répartition des catégories de la variable cible.

L'observation peut se faire graphiquement et / ou numériquement.

A noter qu'on ne peut pas faire de nuages de point, car pour cette méthode, il faut au moins un axe des X ordonné.

Exemples : la variable « churn » et les variables catégorielles

On reprend le fichier d'attrition.

Il y a 5 variables catégorielles : Code département, Région US, International, Mail et Churn.

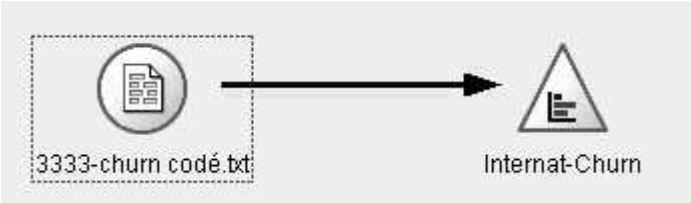
On élimine Région Us et Code Département (cf paragraphe suivant).

Il reste 3 variables catégorielles : International, Mail et Churn.

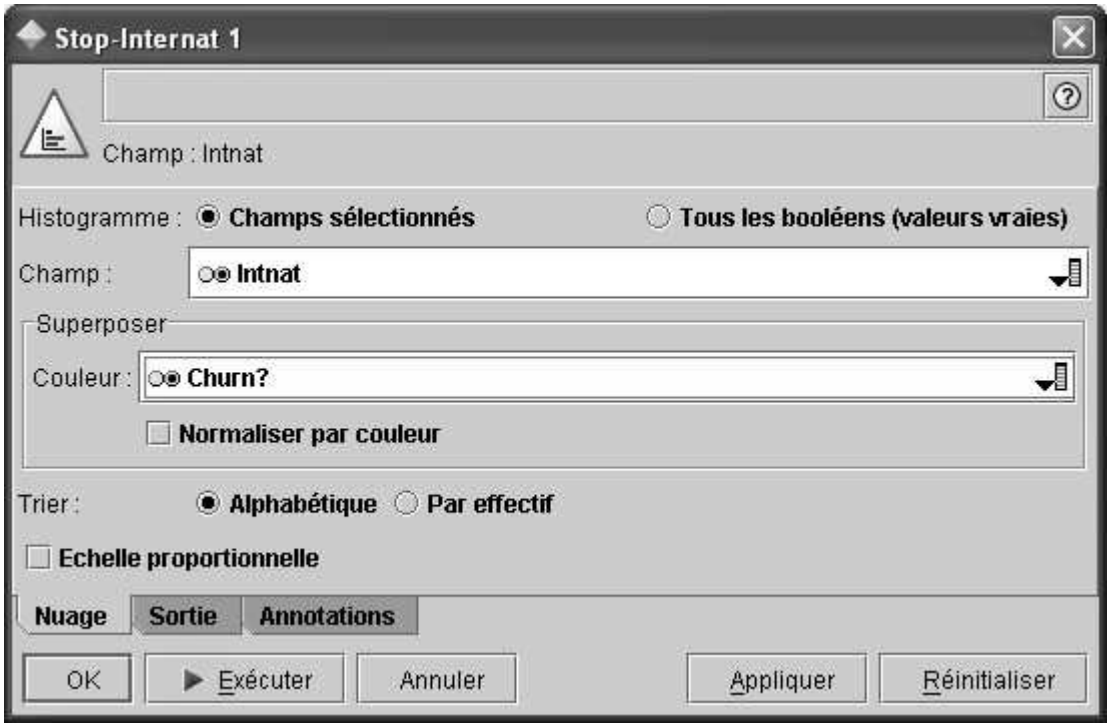
En faisant de Churn une variable cible, on a deux couples intéressants : (international – Churn) et (mail – Churn).

Indépendamment de la variable cible, on peut aussi regarder la corrélation entre international et mail : (international – mail).

Quelle corrélation y a-t-il entre le fait de stopper son abonnement (*churn*) et le fait d'avoir une option internationale ?



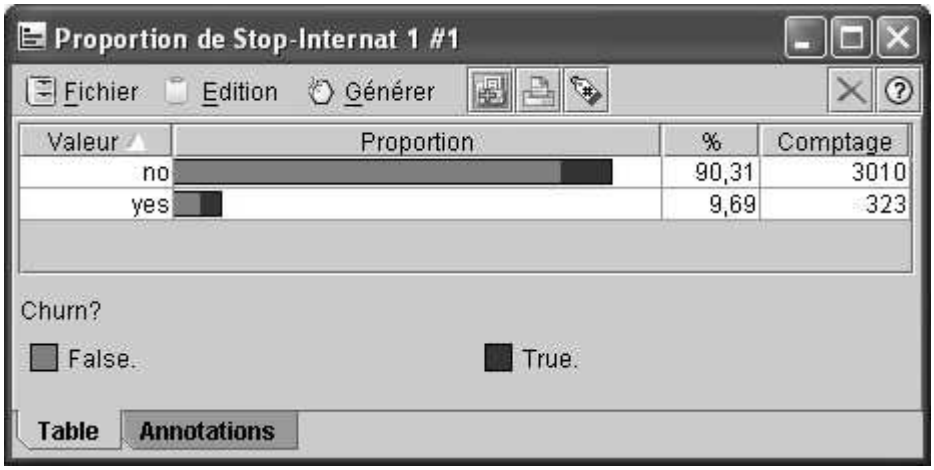
Flux clementine



Paramétrage du flux

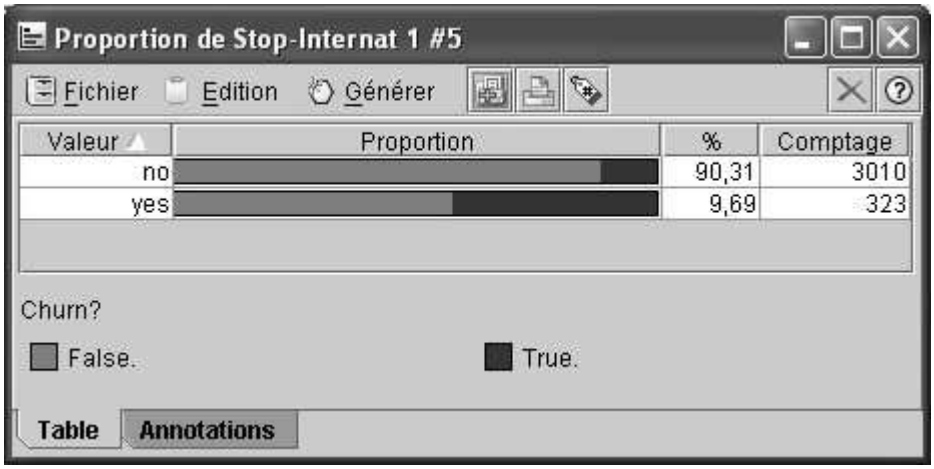
Le champ « Intnat » est la variable prédictive. On lui superpose la variable cible « Churn »

Tel quel, le résultat n'est pas très lisible. On ne voit pas clairement la différence entre la proportion de « Churn » à vrai pour « intnat » à « no » ou à « yes ».



Superposition de la variable cible dans la variable prédictive

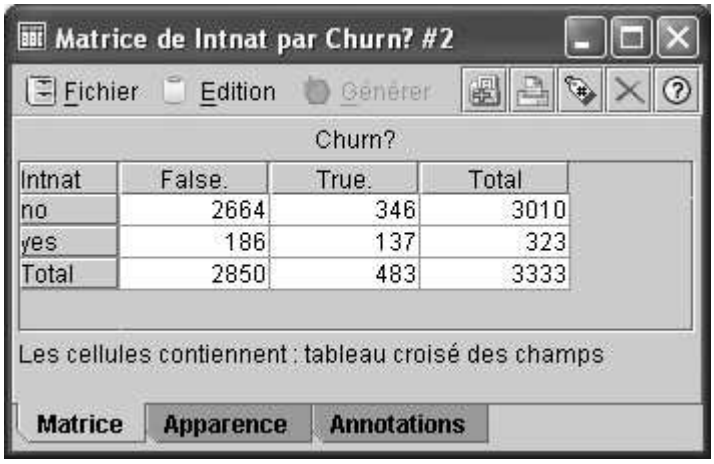
Pour améliorer la lisibilité, il faut normaliser par couleur au niveau du paramétrage :



Présentation avec normalisation par couleur

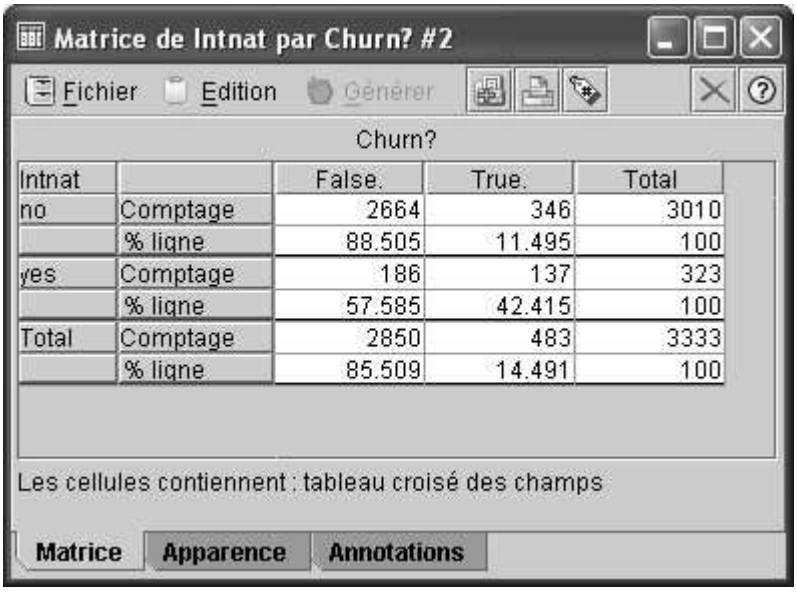
Avec cette présentation, on constate graphiquement qu'il y a beaucoup plus de gens qui arrêtent leur abonnement en cas d'option internationale.

On peut aussi produire un bilan chiffré :



On voit clairement que la proportion de « churn » est beaucoup plus grande pour l'international (137 sur 323) que sinon (346 sur 3010).

On peut préciser encore plus le bilan chiffré :



Bilan chiffré complet

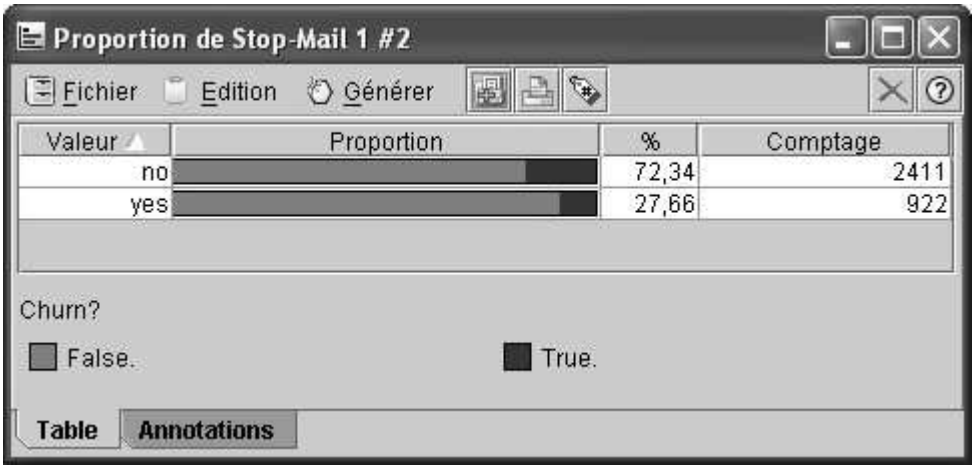
On a donc 42,4 % des gens qui stoppent leur contrat quand c'est un contrat international contre 14,5 % qui stoppe leur contrat quand ce n'est pas un contrat international.

C'est une information intéressante sur le plan stratégique : cela signifie que le contrat international est mal positionné sur le marché.

Cela concerne 137 clients sur 483 soit 28,4 % des « churn ».

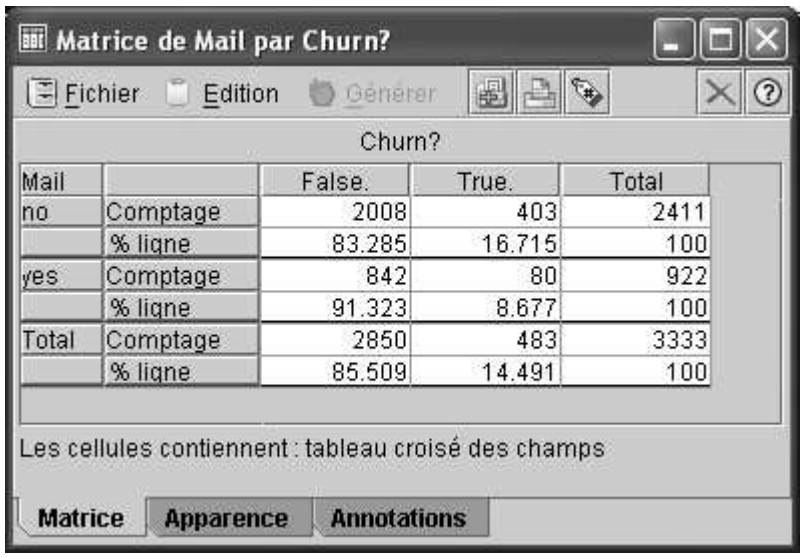
Monter au niveau de fidélité du national c'est progresser de (42,415-11,495)*323 = 100 clients.

<u>Quelle corrélation y a-t-il entre le fait de stopper son abonnement (*churn*) et le fait d'avoir une option messagerie ?</u>



Superposition normalisée du churn dans la variable « mail »

On constate qu'il y a plus de gens qui arrêtent leur abonnement quand ils n'ont pas l'option mail. C'est moins flagrant que dans le cas précédent, mais ça concerne plus de clients (2411 contre 323)



Bilan chiffré complet

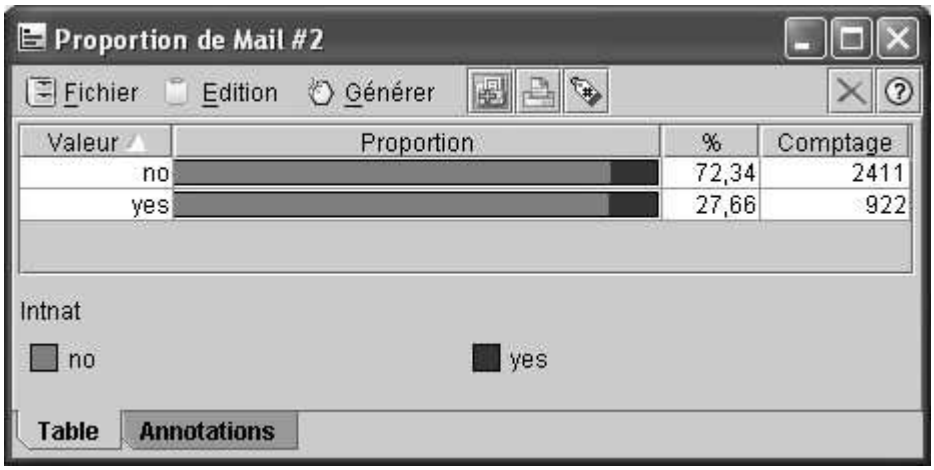
On a donc 16,7 % des gens qui stoppent leur contrat quand c'est un contrat sans mail contre 8,7 % qui stoppe leur contrat quand c'est un contrat avec mail.

C'est une information intéressante sur le plan stratégique : cela signifie que le contrat sans mail est mal positionné sur le marché.

Cela concerne 403 clients sur 483 soit 83,4 % des « churn ».

Monter au niveau de fidélité du mail c'est progresser de (16,715-8,677)*2411 = 194 clients.

<u>Quelle corrélation y a-t-il entre l'option internationale et l'option mail ?</u>



Superposition normalisée de la variable « intnat » dans la variable « mail »

On constate qu'il y a autant de gens abonnés à l'internationale dans la catégorie avec mail que dans la catégorie sans mail

Mail		no	yes	Total
no	Comptage	2180	231	2411
	% ligne	90.419	9.581	100
yes	Comptage	830	92	922
	% ligne	90.022	9.978	100
Total	Comptage	3010	323	3333
	% ligne	90.309	9.691	100

Bilan chiffré complet

Plus précisément, il y a 9,7% de gens abonnés à l'international, et 9,6% dans la catégorie sans mail, 10,0 % dans la catégorie sans mail.

Le couple (international – mail) est donc équivalent à une variable unique.

8 : Corrélations entre variables numériques et variables catégorielles

Présentation

Après les corrélations entre variables numériques et les corrélations entre variables catégorielles, nous allons maintenant aborder les corrélations simples entre variables numériques et variables catégorielles.

On peut appliquer le même principe que pour les variables numériques et les variables catégorielles : croiser toutes les variables entre elles.

On peut aussi considérer une variable catégorielle particulière : la variable cible, c'est-à-dire celle sur laquelle on souhaite faire des prédictions.

Principe de la corrélation

Dans chaque l'histogramme de la variable numérique qui sera la variable prédictive, on va observer la répartition des catégories de la variable cible.

L'observation se fait graphiquement.

En toute rigueur, on peut aussi faire un nuage de point puisqu'on a un axe des X ordonné (variable numérique). Toutefois, en général avec cette méthode, le résultat est difficile à interpréter.

Exemples : la variable « churn » et les variables numériques

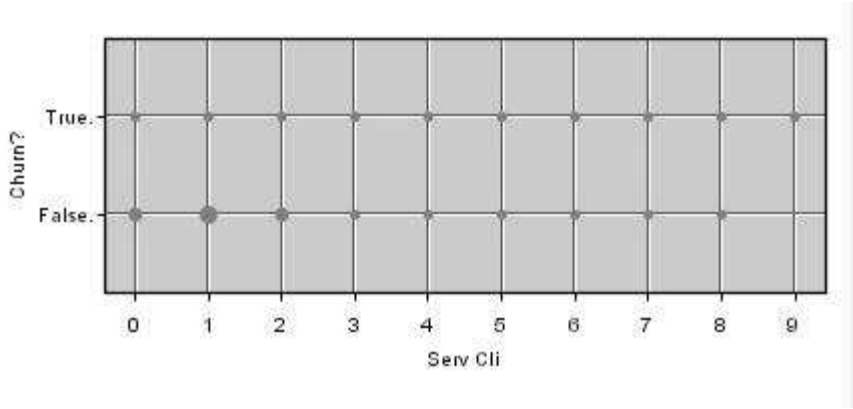
On reprend le fichier d'attrition.

En fait de Churn la variable cible. On examinera donc les couples de variables suivants :

- appels au service client - *churn*
- nb d'appels - *churn*
- consommation - *churn*
- nb msg - *churn*
- durée de vie - *churn*

<u>Quelle corrélation y a-t-il entre le fait de stopper son abonnement (*churn*) et le nombre d'appels au service client ?</u>

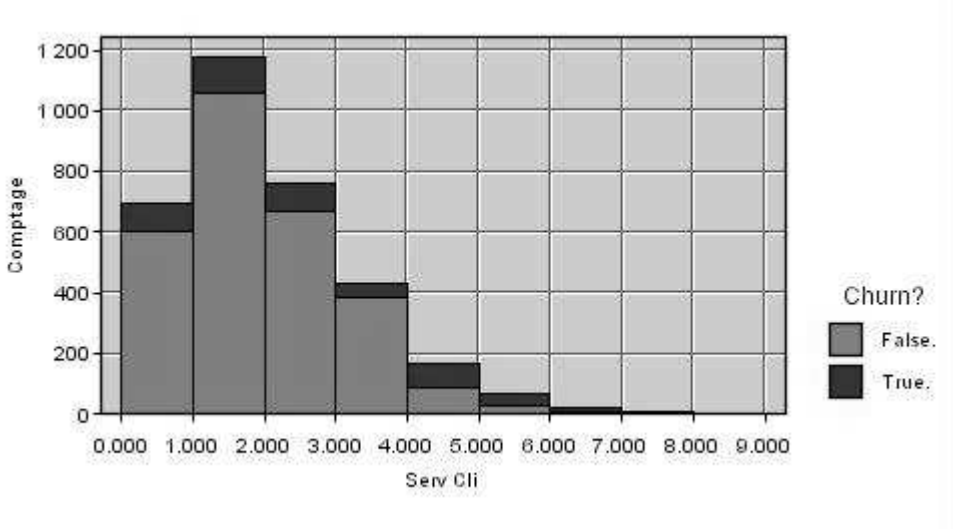
Méthode du nuage de points :



Le résultat n'est pas très interprétable. Faire un nuage de point avec une variable catégorielle sur deux valeurs (churn), ce n'est pas très lisible. Cependant, la taille du point donne une idée du nombre de données concernées.

Le point (False – 1) correspond à 1059 individus (on peut le voir en passant la souris sur le point).

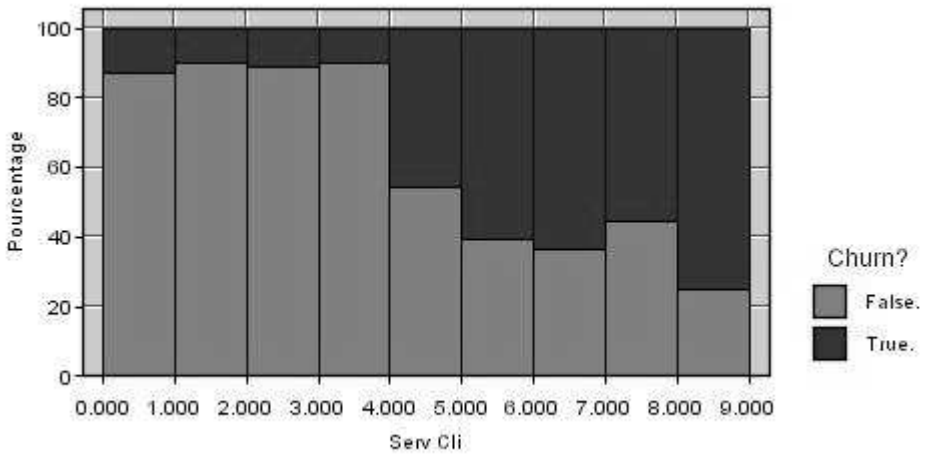
Méthode de l'histogramme :



Superposition de la variable cible dans la variable prédictive

Cette présentation n'est pas très lisible car on ne voit pas bien la proportion des « True » et des « False ».

Il faut normaliser la représentation pour obtenir un graphique lisible :



Superposition normalisée du churn dans la variable « Serv Cli »

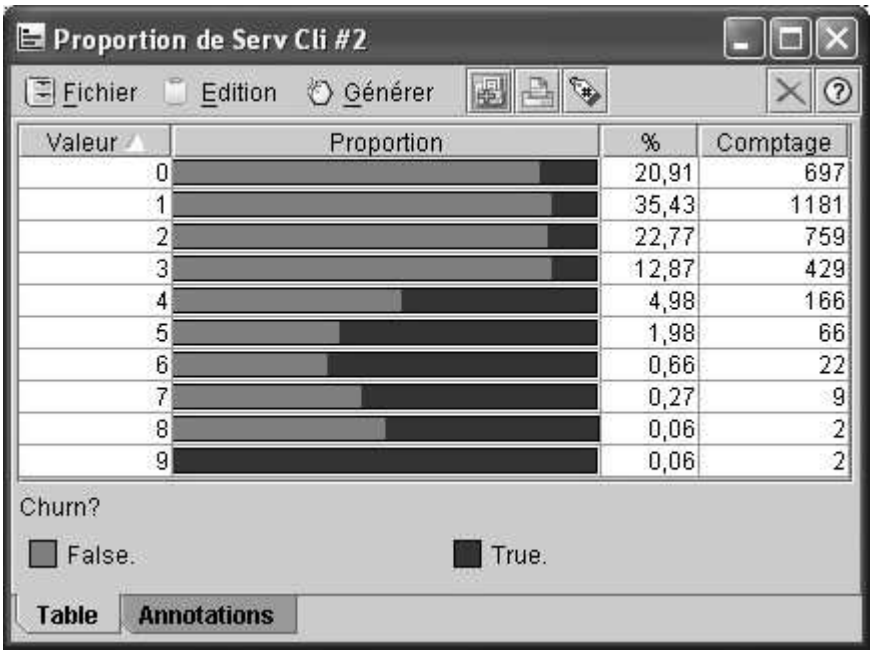
Sur ce graphique, l'axe des Y est un pourcentage et non plus un nombre d'individus comme précédemment.

On voit très clairement qu'à partir du 4ème appel au service client, on rentre dans une zone à fort risque de départ (de « churn »).

C'est une information très utile : on peut en conclure qu'il faut prendre un soin tout particulier des clients qui atteignent le 3ème et le 4ème appel/

Bilan chiffré :

Les schémas chiffrent assez mal les résultats. Pour avoir des résultats chiffrés, on peut utiliser le nœud « proportion ». Mais ce nœud ne s'applique qu'aux variables catégorielles. Il faut donc commencer par faire du nombre d'appels au service client une variable catégorielle. C'est possible simplement car c'est un entier : il suffit de changer son type. Quand on a affaire à un réel, il suffit de le disctétiser.



On retrouve le même schéma qu'avec l'histogramme avec en plus le nombre d'individus total par valeur du nombre d'appels au service client.

On peut aussi produire toutes les données chiffrées :

Matrice de Serv Cli par Churn?

Fichier Edition Générer

		Churn?		
Serv Cli		False.	True.	Total
0	Comptage	605	92	697
	% ligne	86.801	13.199	100
	% colonne	21.228	19.048	20.912
1	Comptage	1059	122	1181
	% ligne	89.670	10.330	100
	% colonne	37.158	25.259	35.434
2	Comptage	672	87	759
	% ligne	88.538	11.462	100
	% colonne	23.579	18.012	22.772
3	Comptage	385	44	429
	% ligne	89.744	10.256	100
	% colonne	13.509	9.110	12.871
4	Comptage	90	76	166
	% ligne	54.217	45.783	100
	% colonne	3.158	15.735	4.980
5	Comptage	26	40	66
	% ligne	39.394	60.606	100
	% colonne	0.912	8.282	1.980
6	Comptage	8	14	22
	% ligne	36.364	63.636	100
	% colonne	0.281	2.899	0.660
7	Comptage	4	5	9
	% ligne	44.444	55.556	100
	% colonne	0.140	1.035	0.270
8	Comptage	1	1	2
	% ligne	50.000	50.000	100
	% colonne	0.035	0.207	0.060
9	Comptage	0	2	2
	% ligne	0.000	100.000	100
	% colonne	0.000	0.414	0.060
Total	Comptage	2850	483	3333
	% ligne	85.509	14.491	100
	% colonne	100	100	100

Les cellules contiennent : tableau croisé des champs

Matrice Apparence Annotations

Interprétation :

Entre 0 et 3 appels au service client (ASC), on est entre 10 et 13 % de « churn » positif.

Entre 0 et 3 ASC, le taux de churn « positif » est un peu inférieur au taux de « churn » négatif.

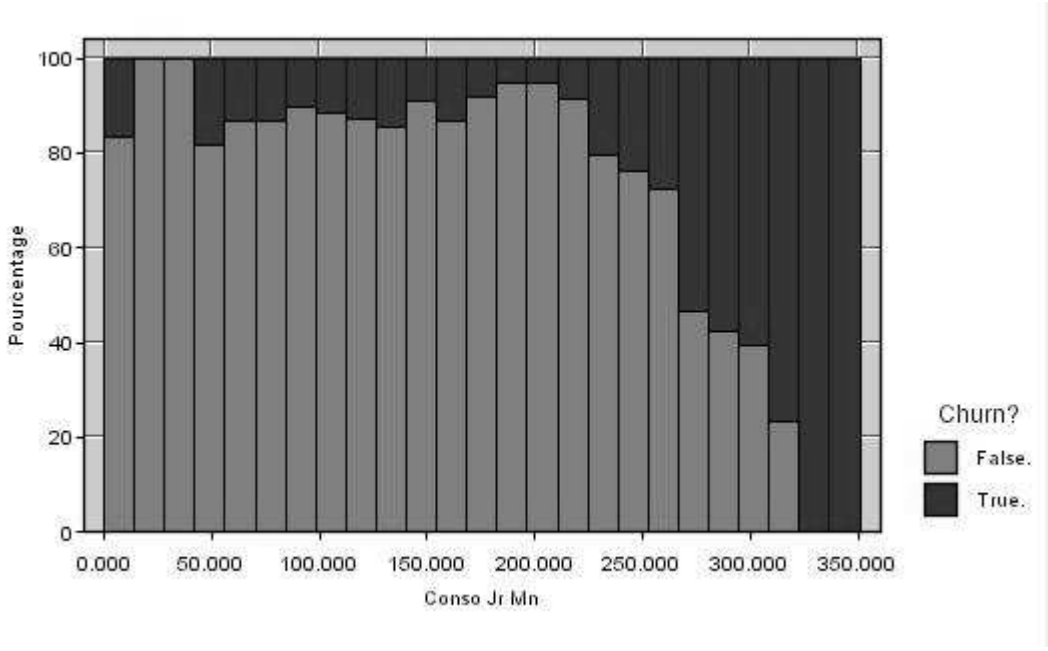
Entre 4 et 7 ASC, on est entre 45 et 63 % de « churn » positif.

Pour 8 et 9 ASC, ça ne concerne que 2 individus à chaque fois : on peut considérer cela comme trop peu significatif.

Remarque :

Finalement, le bilan chiffré ne nous apprend pas grand chose de plus que la lecture de l'histogramme.

<u>Quelle corrélation y a-t-il entre le fait de stopper son abonnement (*churn*) et la consommation par jour</u>



Superposition normalisée du churn dans la variable « Conso Jr Min »

Conclusion :

On a une forte corrélation entre le nombre d'appels au service client et le départ des clients.

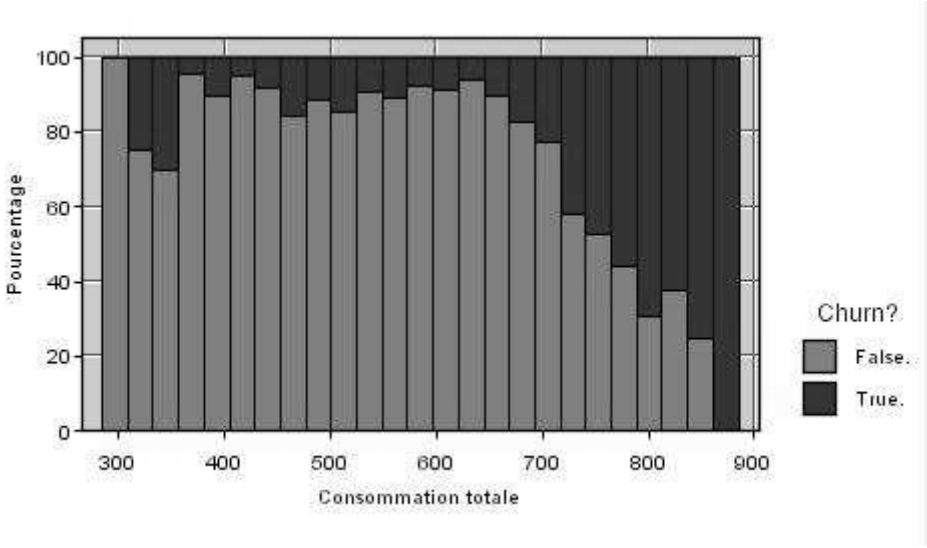
Au-delà de 225 minutes de consommation le jour, on augmente un peu les chances de départ du client.

Au-delà de 270 minutes de consommation le jour, on augmente massivement les chances de départ du client.

Cette situation veut probablement dire que l'offre commerciale n'est pas adaptée à ce type de consommation. Ce sera à vérifier avec le client et l'étude des offres concurrentes.

<u>Corrélation avec une variable calculée : la somme des consommation</u>

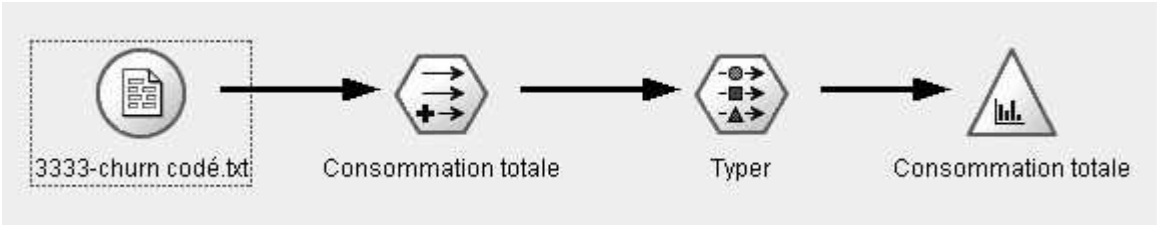
On peut créer une variable qui somme toutes les consommations et la corréler avec la variable cible « churn ».



Superposition normalisée du churn dans la variable calculée « Consommation totale »

Cet histogramme montre qu'à partir de 650, plus on consomme, plus on s'en va !

Cette information n'est pas forcément plus pertinente que la précédente. Dans les deux cas, il faut se référer à l'offre commerciale de l'entreprise et des concurrents pour, probablement, comprendre le phénomène.



Flux clémentine correpondant

Bilan de l'analyse : distinction entre cause et symptôme

La conclusion finale peut donc être que :

On a deux **causes** principales du départ des clients :

- L'augmentation de la consommation
- L'option internationale

Il y a un **symptôme** du départ des clients :

- Le passage à 4 appels au service client.

9 : Corrélations multiples

Présentation

Jusqu'à présent, on a abordé tous les cas possibles de corrélations simples, c'est-à-dire les corrélations avec une seule variable prédictive.

On peut aussi s'intéresser aux corrélations multiples, c'est-à-dire avec plusieurs variables prédictives.

Dans l'absolu, on peut passer en revue tous les n-uplet possibles de variables prédictives.

Dans la pratique, le problème sera de choisir les n-uplets pertinents.

Ce type d'analyse pourra être descriptif ou prédictif : on peut travailler avec ou sans variable cible.

Sans variable cible, on aboutira à des critères de classification.

Avec variable cible, on aboutira à un modèle de prédiction.

Entre variables catégorielles

Exemple

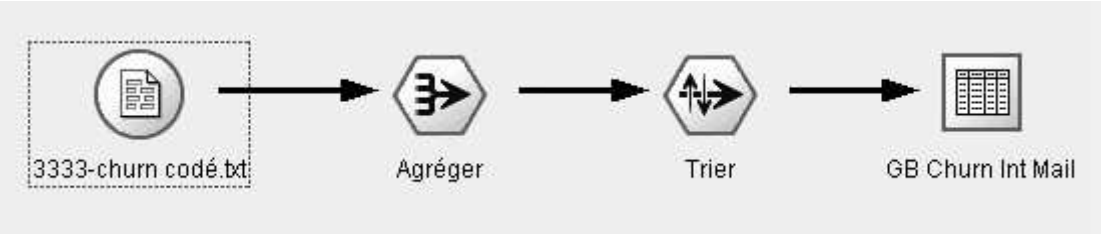
On va s'intéresser au triplet : (Churn - International - messagerie)

Ce qu'on souhaite faire, c'est un « Group by » multiple :

Select International, Messagerie, Churn count(*)

Group by Churn, International, Messagerie;

Clementine



GB Churn Int Mail (4 champs, 8 enre...

Fichie Editio Génére

	Intnat	Mail	Churn?	Effectif
1	no	no	False.	1878
2	no	no	True.	302
3	no	yes	False.	786
4	no	yes	True.	44
5	yes	no	False.	130
6	yes	no	True.	101
7	yes	yes	False.	56
8	yes	yes	True.	36

Table Annotations

Ce tableau n'est pas facile à interpréter.

On a intérêt à le passer sous excel pour faire des calculs de pourcentages.

Excel

Intnat	Mail	Churn?	Effectif	%	Bilan	
No	No	False.	1878	56,35%	86,15%	100,00%
No	Yes	False.	786	23,58%	94,70%	100,00%
yes	No	False.	130	3,90%	56,28%	100,00%
yes	Yes	False.	56	1,68%	60,87%	100,00%
No	No	True.	302	9,06%	13,85%	
No	Yes	True.	44	1,32%	5,30%	
yes	No	True.	101	3,03%	43,72%	
yes	Yes	True.	36	1,08%	39,13%	
			3333	100,00%		

La colonne « bilan » calcule le %age de churn de la catégorie : dans la catégorie (no – no), on a 86,15 % de chance de rester contre 13,85% de chance de partir.

Donc, c'est la catégorie (internat=yes, mail=no) qui présente le plus mauvais taux de fidélisation, suivie de près par la catégorie (internat=yes, mail=yes). On retrouve donc une information qu'on avait déjà trouvée : quand internat = yes, on a un mauvais taux de fidélisation, quelle que soit la situation côté mail.

Entre variables numériques et variable cible catégorielle

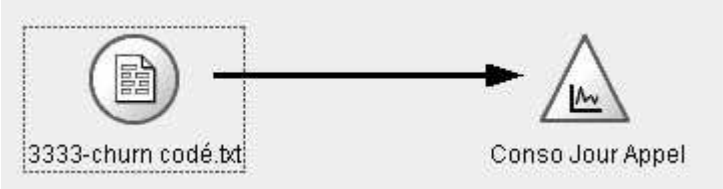
On peut regarder la répartition du « *churn* » sur des nuages de points en plan (2 variables croisées) ou dans l'espace (3 variables croisées).

Exemple

On va s'intéresser au triplet : (Churn – Serv Cli– Conso Jr Mn)

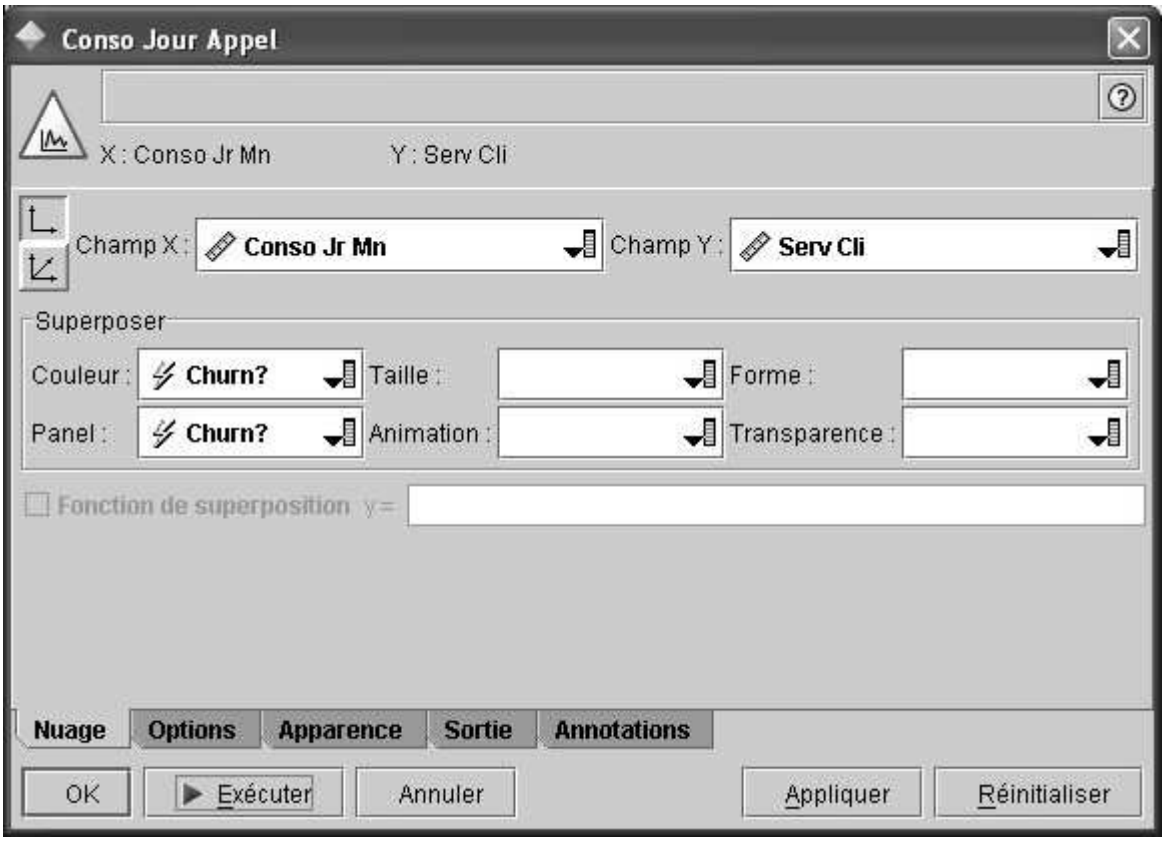
Nuages de points en plan

Clementine propose deux représentations : soit en regroupant toutes les valeurs de la variables cibles sur un même graphique, soit en faisant un graphique par valeur de la variable cible.

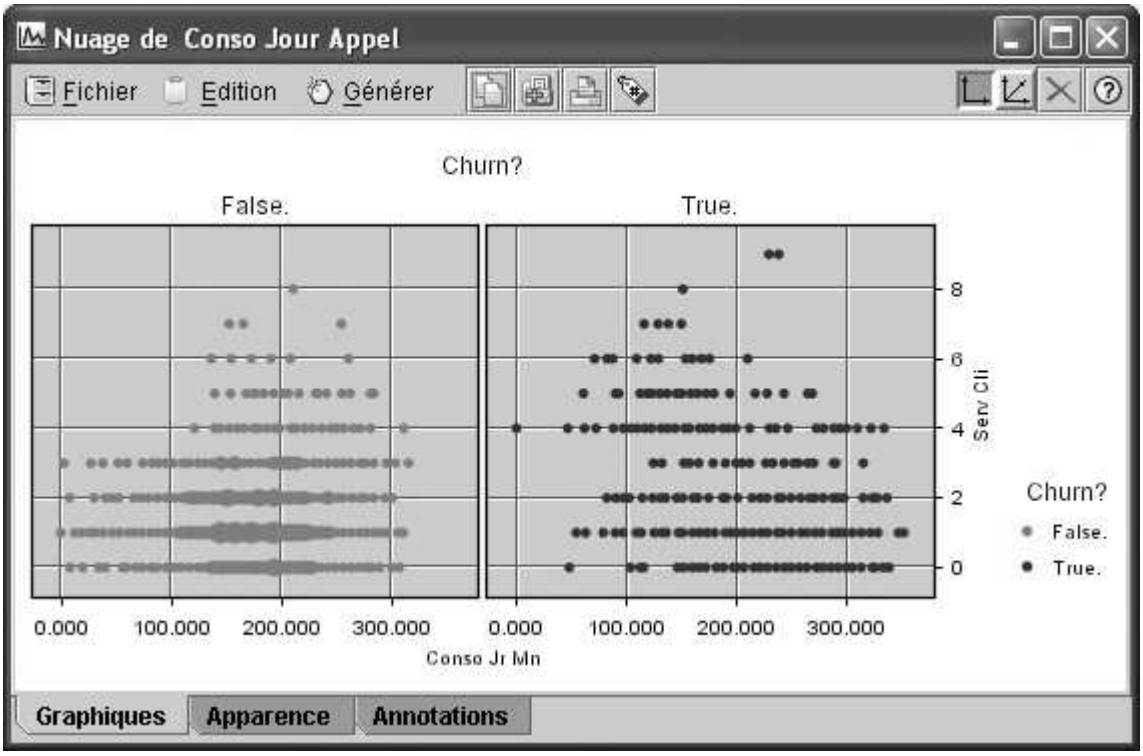


Flux clementine

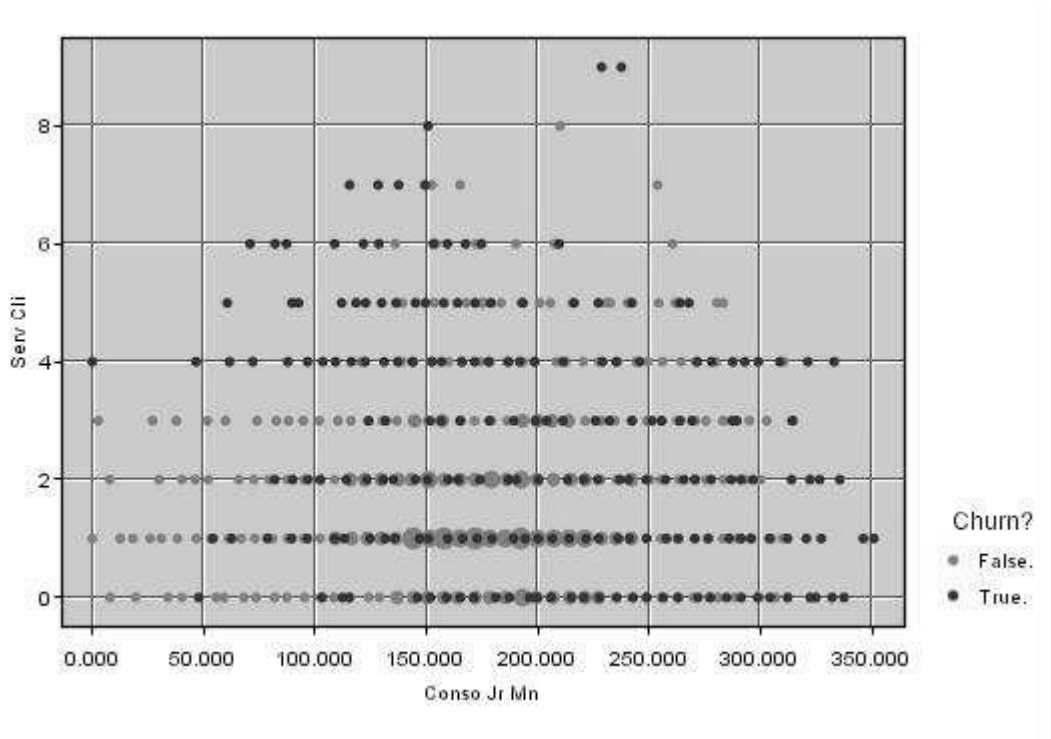
Paramétrage du nœud « Nuage de points » :



On a deux variables prédicteurs en entrée et une variable cible en sortie. En choisissant l'option panel, on aura un graphique par catégorie de la variable cible.



Version sur un seul shéma :



En observant la répartition du « *churn* » dans le nuage (consommation jour - appels service client), on constate que :

- Les petits consommateurs (< à 110) qui appellent beaucoup le service client (> 3) s'en vont quasiment tous !
- Les très gros consommateurs s'en vont tous.

Nuages de points en volume

C'est difficile à interpréter ! Il faut faire tourner le cube et essayer de repérer des groupes cohérents. Ca peut être parlant si on a des regroupements très cohérents.

10 : Les données corrélées incohérentes

Principe

Si deux variables sont corrélées, autrement dit s'il existe une dépendance fonctionnelle entre les deux variables, alors il peut arriver qu'il y ait des incohérences dans les couples de variables.

Ceci concerne surtout les dépendances fonctionnelles simples repérables par bon sens ou expertise métier.

Exemple

Dans le cas du fichier d'attrition, on a la dépendance fonctionnelle suivante :

CodDept (numéro de département dans l'Etat) ----> RegUs (Etat des Etats-Unis)

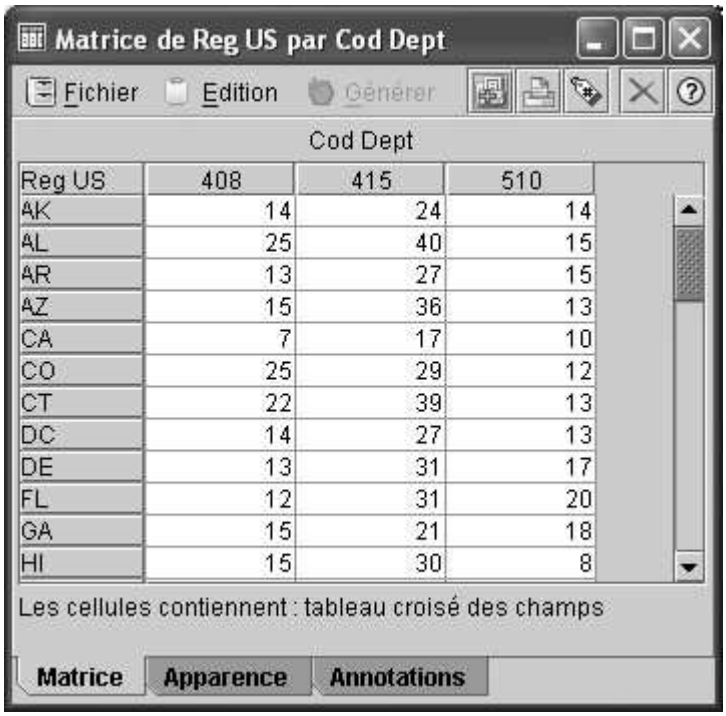
Donc, pour une valeur du CodDept, on doit toujours avoir la même valeur de RegUs.

Pour vérifier cela on peut :

- Trier le tableau par CodDept et par RegUs et vérifier visuellement
- Afficher une matrice qui compte le nombre d'occurrences de chaque couple de catégories, et vérifier la cohérence.

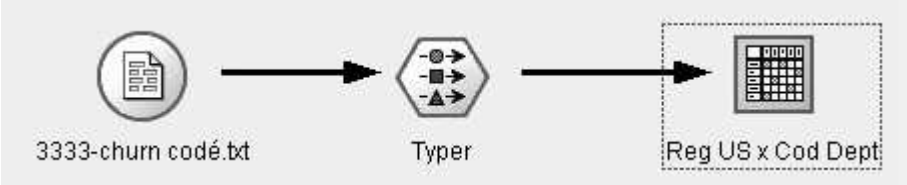
Clementine

On va afficher la matrice d'occurrences par couple de catégories



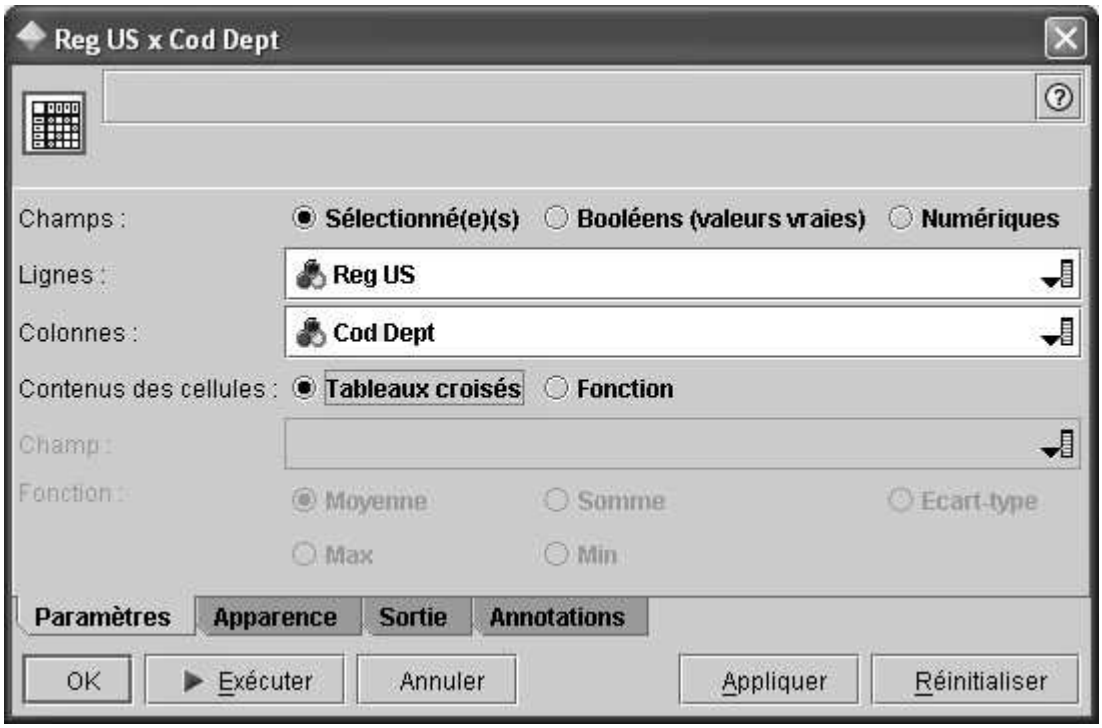
On constate d'une part qu'il y a très peu de valeur pour le CodDept, d'autre part que chaque valeur est couplée avec toutes les valeurs de Reg US.

La variable Cod Dept paraît donc clairement incohérente. On peut aussi se poser des questions sur la variable Reg Us. On a intérêt à faire appel à un expert du métier pour vérifier.



Flux clémentine de la matrice : il faut typer les attributs pour avoir deux variables catégorielles

Paramétrage de la matrice :



11 : Dernières techniques : discrétisation, variables calculées, sous-ensembles

Discrétisation des variables continues

Le principe est de passer de variables continues à des variables catégorielles (discrètes).

On a déjà abordé cette question dans différents exemples.

Le problème, c'est de définir les catégories : de 1 en 1, de 10 en 10, par paquets irréguliers.

La discrétisation peut se faire de façon intuitive (au hasard !)

Ce sont surtout les techniques de classification qui permettront de faire apparaître des catégories pour les variables.

Avec ces nouvelles catégories, on sera amené, éventuellement, à affiner la phase de compréhension des données.

Création de variables calculées

On peut être amené à créer des variables calculées si celle-ci sont utiles à une meilleure compréhension des données ou à une meilleure modélisation.

Dans l'exemple du fichier d'attrition, on a vu l'exemple de la variable « somme des consommation » qui fait la somme des consommation de jour, de soirée et de nuit.

Travailler sur des sous-ensembles

À partir de l'analyse graphique, on peut sélectionner des sous-ensembles pour travailler sur des populations homogènes.

Par exemple, on a vu qu'à partir de 4 appels au service client, on entre dans une nouvelle catégorie de clients avec un risque d'attrition beaucoup plus important.

Il pourrait donc être intéressant de diviser la population en deux et de refaire des analyses sur ces deux populations distinctes.

Cela dit, fondamentalement, la détermination de sous-ensemble relève de la partie classification de la modélisation.

Conclusion

L'analyse des données est le travail préparatoire à la modélisation.

C'est un travail lent, laborieux, méthodique et assez systématique qui produit à coup sûr des résultats importants : il faut s'y atteler sérieusement avant toute modélisation.

PHASE 3 : PREPARATION DES DONNEES

<table>
<tr><th colspan="3">PROCESSUS du DATA MINING</th></tr>
<tr><th>Acteurs</th><th>Étapes</th><th>Phases</th></tr>
<tr><td rowspan="5">Maître d'œuvre</td><td>Objectifs</td><td>1 : Compréhension du métier</td></tr>
<tr><td rowspan="2">Données</td><td>2 : Compréhension des données</td></tr>
<tr><td>3 : Préparation des données</td></tr>
<tr><td rowspan="2">Traitements</td><td>4 : Modélisation</td></tr>
<tr><td>5 : Évaluation de la modélisation</td></tr>
<tr><td>Maître d'ouvrage</td><td colspan="2">Déploiement des résultats de l'étude</td></tr>
</table>

Description de la phase de préparation des données

Introduction

Les phases de compréhension et de préparation sont deux phases très importantes d'un projet de data mining. Elles se font en aller-retour. Pour préparer, il faut comprendre, pour comprendre il faut avoir préparé.

Ces deux phases vont occuper plus de la moitié du temps d'un projet de data mining.

Contexte

Les analyses de data mining travaillent souvent sur des données qui n'ont pas été gérées depuis des années. De ce fait, les données nécessitent une phase de préparation.

Fondamentalement, la préparation précède l'analyse exploratoire et les premières recherches de corrélation.

Principe de la préparation

La préparation des données consiste à :

- Préparer, à partir des données brutes, l'ensemble final des données qui va être utilisé pour toutes les phases suivantes.

Cet objectif comporte plusieurs étapes :

- Fusionner si nécessaire plusieurs tables de données en une seule.
- Réaliser si nécessaire la suppression de certaines données.
- Réaliser si nécessaire la correction de certaines données.

- Réaliser si nécessaire la transformation du types des données
- Mettre de côté si nécessaire les variables inutiles
- Créer si nécessaire des variables calculées

On a déjà abordé certaines de ces étapes dans la partie « compréhension des données ».

Relations entre préparation et compréhension

Les phases de préparation et de compréhension sont liées circulairement :

- C'est parce qu'on comprend les données qu'on peut les fusionner, les sélectionner, les corriger ou les nettoyer.
- En fusionnant les données, en les corrigeant ou en les nettoyant, on est amené à réévaluer la compréhension des données.

La fusion des données

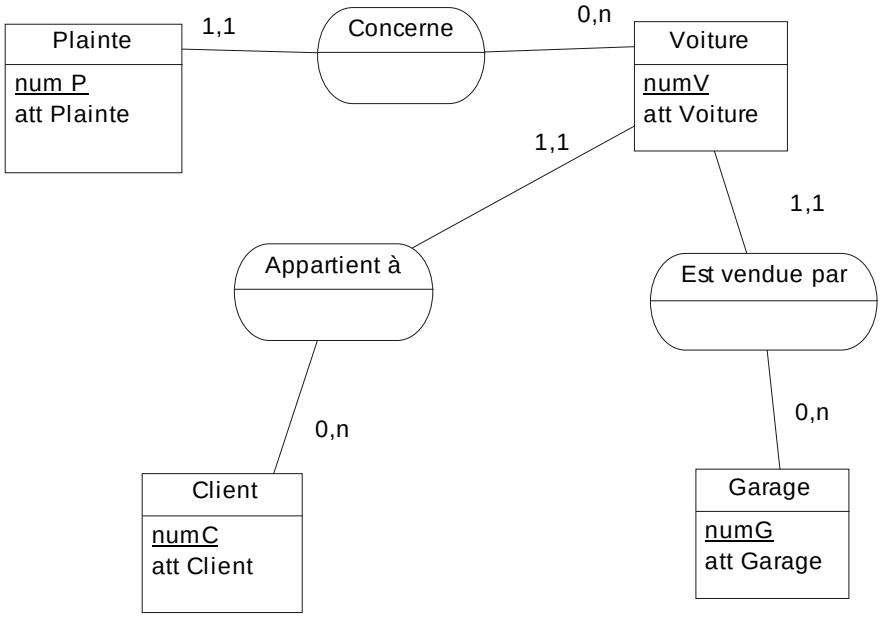
Principe de la fusion

Étant donné que les modèles de data mining s'appliquent à une seule table, la fusion des données consiste à fabriquer une seule table à partir de données réparties dans plusieurs tables.

Exemple de fusion

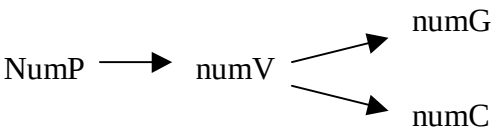
On reprend l'exemple n°1 de la compréhension du métier : un constructeur automobile veut étudier les plaintes de clients acheteur d'automobile.

Les données se trouvent dans quatre tables : une table des plaintes, une table des personnes, une table des voitures et une table des garages. Le MCD de la BD est le suivant :



L'objectif est d'établir des corrélations entre les voitures et les plaintes et entre les garages et les plaintes.

Pour déterminer la table à utiliser, on a intérêt à faire le graphe des dépendances fonctionnelles :



On partira donc de la table suivante :

Plainte (numP, attPlainte, #numV, attVoiture, #numG, attGarage, #numC, attClient)

Cette table obtenue par la jointure suivante :

```
Select numP, attPlainte, voiture.numV, attVoiture, client.numC, attClient, garage.numG, attGarage
From plainte, voiture, garage, client
Where plainte.numV = voiture.numV
And voiture.numG = garage.numG
And voiture.numC = client.numC;
```

Il est possible de faire des jointures avec les logiciels de data mining.

Les valeurs aberrantes

Présentation

Les valeurs aberrantes (*outlier* en anglais) sont des valeurs extrêmes.

Certaines méthodes statistiques seront sensibles à ces valeurs : il faut donc, autant que possibles, les repérer et les corriger ou les éliminer.

Repérage des valeurs aberrantes

On a déjà abordé le repérage de ces valeurs dans le chapitre sur la compréhension des données.

On peut utiliser des méthodes numériques ou des méthodes graphiques (tri, histogrammes, nuages de points).

Correction des valeurs aberrantes

Deux options se présentent :

1) On a affaire à des cas individuels de valeurs aberrantes.

Trois solutions s'offrent à nous :

1 : On supprime l'individu. Le risque est de supprimer par la même occasion des valeurs utiles. Si la population est nombreuse, ce risque devient très faible.

2 : On remplace la valeur aberrante par une autre valeur : la valeur moyenne, une valeur prise au hasard dans la distribution des valeurs. Le risque est de produire une incohérence par rapport aux autres valeurs (aberration croisée).

3 : On réduit les valeurs possibles pour la valeur manquante du fait de l'existence de corrélations entre la variable à valeur manquante avec d'autres variables renseignées.

La meilleure solution est la 3ème si on peut arriver à un jeu de valeurs possibles très restreint, voir à une seule valeur possible.

Sinon, mieux vaut supprimer l'individu pour éviter de produire des aberrations croisées.

2) On a affaire à un cas général de valeurs aberrantes pour une ou plusieurs variables données.

Dans ce cas, on élimine les variables concernées. C'était le cas des variables États et Département dans le fichier d'attrition traité en exemple.

Le problème des données manquantes

Les données manquantes sont un cas particulier de valeurs aberrantes.

Le repérage est facile à faire : il suffit en général de trier la variable pour trouver les données manquantes au début ou à la fin.

Pour corriger les données manquantes, on utilise la même méthode que pour les données à valeurs aberrantes.