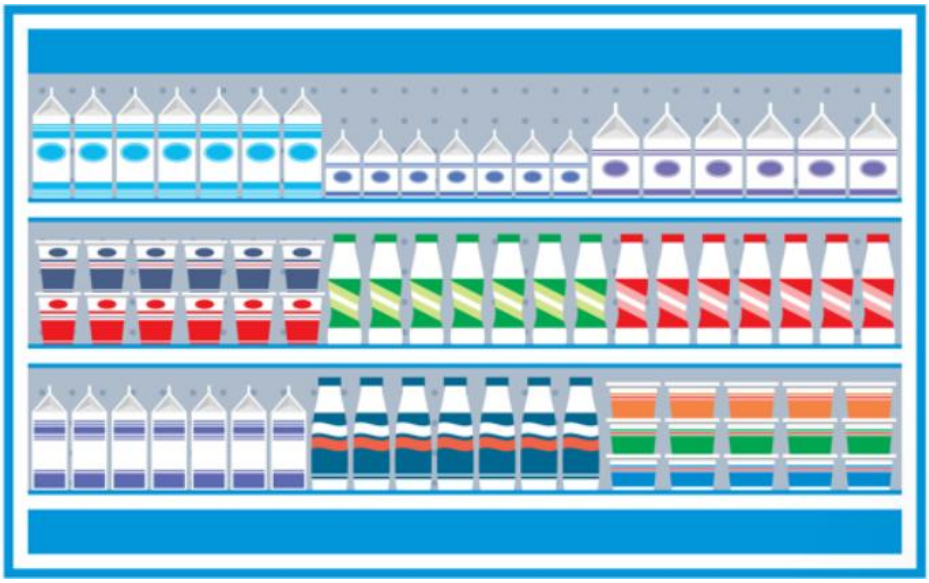
Université Mohammed V – Souissi
Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Projet Datamining

Segmentation des clients



Réalisé par :
BOUHALI Sara
TOUZANI Sara
BELAHMIDI Hind

Encadré par :

Mme BENBRAHIM Houda

Année Universitaire 2017 – 2018

Table des matières

Table des figures

Remerciements :

Au terme de ce travail, nous profitons de l'occasion pour remercier du fond du cœur tous ceux qui ont participé de près ou de loin à réalisation de ce projet et qui ont contribué à en faire une expérience enrichissante, nous nous adressons donc, en particulier nos sincères remerciements à notre encadrante Mme.Benbrahim Houda pour ses précieux conseils et pour nous avoir donné l'occasion de mettre en application nos connaissances et de les approfondir à travers ce projet. Nos remerciements vont aussi à Mme.Mountassir Asmaa pour sa disponibilité et son effort afin de nous éclaircir plus le sujet. En espérant que ce travail soit à la hauteur de vos attentes. Que tous ceux et celles qui ont contribué de près ou de loin à l'accomplissement de ce travail trouvent l'expression de nos remerciements les plus chaleureux. Cordialement.

Résumé

Le sujet de notre projet datamining consistait à « Segmenter les clients ». Notre travail s'est fait en plusieurs phases : Une phase d'analyse et une phase de réalisation. Grâce à l'aide de notre encadrante Mme BENBRAHIM Houda qui nous a supervisé tout au long du projet, nous avons appris, durant ces derniers mois, mener un des projets inhérents à notre branche d' « Informatique Décisionnelle » , à savoir , le projet de Datamining, en s'appuyant sur la méthode CRISP-DM. Nous vous présenterons donc tout au long de ce rapport, les étapes que nous avons suivies, les outils que nous avons utilisés pour réaliser notre projet et l'élaboration de différentes discussions sur les résultats obtenus.
Tout au long de la phase de réalisation, plus précisément, au niveau de compréhension des données, nous avons fait face à des situations délicates. Ainsi avons-nous multiplié nos efforts afin d'atteindre notre objectif en respectant la contrainte temporelle. Ce qui nous a permis de mettre en œuvre nos connaissances d'une manière efficace et efficiente.

Mots clés : Datamining, CRISP-DM.

Abstract

The subject of our datamining project was "Segmenting Customers". Our work was done in several phases: an analysis phase and a realization phase. Thanks to the help of our professor Ms. BENBRAHIM Houda, who supervised us throughout the project, we learned during these last months how to lead one of the projects inherent to our branch of "Business Intelligence", in other words, a Datamining project, using the CRISP-DM method. Throughout this report, we will present the steps we followed, the tools we used to carry out our project and the development of various discussions on the results obtained.

Throughout the implementation phase, more precisely, at the level of data understanding, we have faced some difficult situations. Thus, we have multiplied our efforts to reach our goal by respecting the time constraint. This project was an opportunity for us to implement our knowledge in an effective and efficient manner.

Key words: Datamining, CRISP-DM.

INTRODUCTION

Les entreprises exploitent de nos jours des volumes de données de plus en plus importants. Ces données permettent d'effectuer des analyses poussées à l'aide des techniques d'analyses classiques. Cependant, lorsque le volume de données devient trop conséquent, les traitements statistiques classiques atteignent leur limite et l'utilisation du datamining est alors envisageable. Ce dernier utilise des techniques statistiques traditionnelles comme la régression linéaire et logistique ou plus élaborées, telles que l'analyse multi variée, l'analyse en composante principale, les arbres décisionnels et les réseaux de neurones.
Nous manipulerons tout au long de notre analyse des données dites symboliques, permettant de résumer les données par des concepts plus larges (par exemple, on ne s'intéresse plus à un brand mais à une catégorie d'un produit). Cela permettra d'obtenir de nouvelles connaissances et d'aborder les problèmes sous un nouvel angle.
Notre étude se portera sur la segmentation des clients. Nous présenterons dans un premier temps le contexte général du projet. En seconde partie, nous entamerons l'analyse du sujet. Ensuite, nous présenterons les outils utilisés durant le développement du projet. Enfin, nous procéderons à l'application de la méthode CRISP-DM afin de générer un clustering des données de notre projet.

CHAPITRE

1 Contexte général du projet

Contexte général du projet

Contexte

Le marketing ciblé nécessite une bonne connaissance de son marché, et d'en procéder à une analyse suffisamment fine pour ajuster la nature et le contenu de ses offres.
La première étape de cette analyse repose sur les techniques dites de segmentation des clients. Segmenter sa clientèle consiste à diviser celle-ci selon des critères précis qui vont permettre de regrouper les acteurs composant cette clientèle en groupes d'individus (personnes ou organisations) de façon à ce que ces groupes soient les plus homogènes possible. Ainsi un segment de marché est un groupe dont les individus qui le composent présentent des caractéristiques proches, et sont donc susceptibles de répondre de manière relativement identique à une proposition commerciale.

Objectif du projet

Le travail demandé consiste à découper une population (clients, prospects) en groupes homogènes suivant multiple critères (Données socio-démographiques, comportement d'achat, ..), tel que les clients appartenant au même segment se ressemblent au maximum selon certains critères et les clients de segments différents sont différents le maximum possible.

Figure 1: Critères de segmentation des clients

Spécification du projet

Le projet doit se baser sur la méthode CRISP-DM -Cross Industry Standard Process for Data Mining- étant un Modèle de Processus de data mining qui décrit une approche communément utilisée par les experts en data mining pour résoudre les problèmes qui se posent à eux.

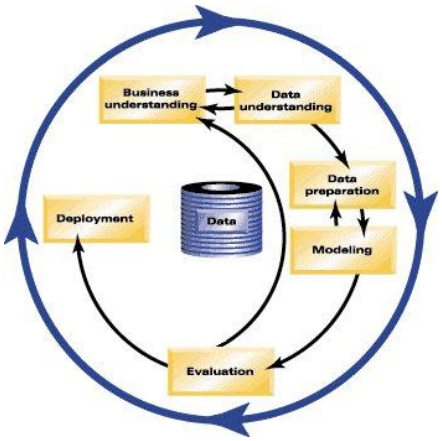


Figure 2: CRISP-DM

Conclusion :

Ce chapitre met en exerguele but crucial de notre projet ‘Segmentation des clients’, qui consiste à découper la clientèle en des groupes d'individus homogènes.

CHAPITRE

2 Analyse

Analyse

La **segmentation** consiste à découper une population en des groupes d'individus homogènes et que les groupes entre eux soient les plus différents possibles. La segmentation peut s'appliquer sur des bases de données clients pour construire des politiques marketing différenciées, des catalogues de produits pour mener des analyses de structure de gamme, des fichiers de personnel pour identifier les facteurs de motivation ou de performance.

Une bonne **segmentation** est une segmentation partagée par les acteurs qui permet la mise en œuvre de stratégie différenciée.

Une segmentation permet de mieux :

- **allouer les ressources de l'entreprise** sur les segments prioritaires (gains de ressources)
- **identifier les attentes** des différents segments et d'entreprendre une amélioration de l'offre
- **piloter le capital** client de l'entreprise en suivant l'évolution de la valeur des différents segments dans le temps.

La **segmentation** est donc :

- **un outil stratégique** pour améliorer l'offre,
- **un outil opérationnel** pour affecter les ressources,
- **un outil de pilotage** pour évaluer la performance des politiques mises en œuvre.

Les critères utilisés pour segmenter un marché sont nombreux, et peuvent être répartis selon trois dimensions : démographiques, géographiques, sociaux et économiques, de personnalité, d'avantages recherchés par les clients et enfin comportementaux.

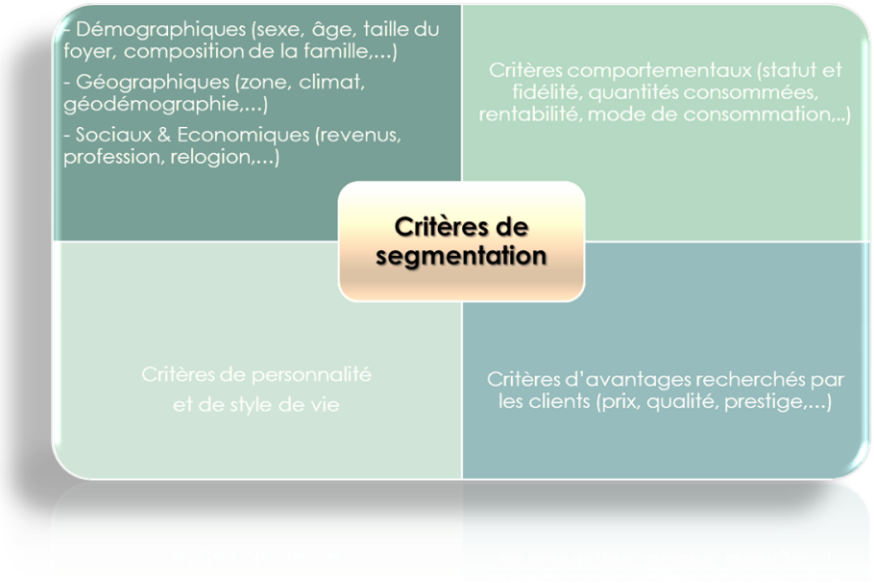


Figure 3:Classification des différents critères

Segmentation RFM :

La segmentation RFM ou méthode RFM est une méthode de segmentation principalement développée à l'origine pour les actions de marketing direct des véadistes(entreprise pratiquant la vente à distance ou VAD) et qui s'applique désormais également aux acteurs du e-commerce et du commerce traditionnel.

C'est un concept marketing qui permet d'évaluer le potentiel d'un client. Les méthodes liées à la segmentation RFM sont nombreuses, et dépendent de ce qui est recherché.

Méthode d'analyse de la qualité d'un client selon trois critères :

1. **La Récence**

La Récence, c'est le laps de temps écoulé depuis la dernière visite d'un client ou la dernière commande d'un client. Par exemple, nous pouvons distinguer 5 récences :

4 : dernière commande/visite dans les 3 derniers mois ;

3 : dernière commande/visite entre 3 et 6 mois ;

2 : dernière commande/visite entre 6 et 9 mois ;

1 : dernière commande/visite entre 9 et 12 mois ;

0 : aucune commande dans la dernière année.

Des trois paramètres de la segmentation RFM, c'est le plus important. L'hypothèse formulée ici est **« plus un client est venu récemment, plus la probabilité qu'il revienne est grande. »** Ceci est corrélé au secteur considéré, car si cette hypothèse est robuste pour la consommation immédiate (Alimentaire, habillement, sorties), pour le secteur immobilier, ou automobile, par exemple, plus le client est venu récemment, moins sa probabilité d'acheter à nouveau est faible.

2. **La Fréquence**

La fréquence correspond au nombre de fois où le client est venu sur la période concernée, cette période dépend du secteur considéré. 0 signifie que le client n'a pas commandé sur la période concernée, 1 qu'il a fait une commande, etc… Ici, l'hypothèse est **« Plus le client a commandé dans le passé, plus grande est la probabilité qu'il commande à nouveau »**

3. **Le Montant**

Quand on parle de montant, on utilise le montant moyen du panier du client sur une période. Par exemple, nous pouvons distinguer 4 groupes : 1 : panier moyen du client inférieur de 10% au montant moyen sur la période considérée, 2 : panier moyen du client compris entre 10% en dessous et 10% au-dessus du montant moyen sur la période considérée, 3 : panier moyen supérieur de 10% au montant moyen, et enfin 0 : pas d'achat sur la période

Potentiel Client

Ces trois paramètres, connus sous le sigle de RFM, permettent d'estimer le potentiel d'un client. Nous pouvons le retranscrire sous la forme d'un nombre à 3 chiffres. Le chiffre des centaines sera la récence, celui des dizaines correspondra à la fréquence, et enfin le chiffre des unités sera celui du montant, ou nous pouvons le retranscrire sous la forme d'un score :

Potentiel Score = Récence * pr + Fréquence * pf + Montant * pm

Avec pr, pf, et pm les poids accordés, respectivement à la récence, la fréquence et au montant.

Par exemple, dans le premier cas, nous prenons deux clients,

- l'un étant venu un mois auparavant, et ayant commandé 3 fois dans l'année pour un montant moyen d'achat correspondant au panier moyen de la période en cours. Son potentiel client sera de 432.
- L'autre étant venu 7 mois auparavant, et ayant commandé 5 fois pour un montant moyen d'achat inférieur de 10% au panier moyen de la période considérée. Son potentiel client sera de 251.

En attribuant à chacun des clients ce potentiel client, nous analyserons leur comportement et attribuerons à chacun le segment leur correspondant.

En croisant la récence et la fréquence, nous pouvons visualiser les segments de clients sur le tableau suivant :

<table>
<tr><td></td><td colspan="6">Fréquence</td></tr>
<tr><td></td><td></td><td>0</td><td>1</td><td>2</td><td>3</td><td>4+</td></tr>
<tr><td rowspan="5">Récence</td><td>0</td><td rowspan="5">Inscrits mais 0 cmdes</td><td colspan="4">Hors période (date dernière cmde > 1 an)</td></tr>
<tr><td>1</td><td colspan="2">A consolider</td><td colspan="2">« Gold »</td></tr>
<tr><td>2</td><td colspan="4" rowspan="2">Redynamiser</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td><td colspan="2">A relancer</td><td colspan="2">A reconquérir</td></tr>
</table>

Figure 4:Matrice Récence-Fréquence

Les personnes catégorisées « gold » sont des personnes consommant indépendamment des circonstances et des actions commerciales. Pour ces personnes, nous opterons pour une stratégie d'information.

Les personnes à consolider sont des clients ayant consommé récemment, mais peu. Nous menons donc des actions de promotions visant à augmenter la fréquence de leurs commandes.

Les personnes à relancer, ainsi que les personnes à reconquérir sont des personnes dont la date de dernière commande est supérieure à neuf mois, et avec qui il faudra reprendre contact par le biais de promotions ciblées.

La méthode RFM, favorisée par l'essor du marketing de base de données, est très utilisée par les entreprises de V.A.D. pour segmenter leur clientèle d'après les informations stockées dans leurs bases de données. Elle permet notamment de vérifier l'adéquation offre/ demande pour une période donnée, de définir la probabilité de l'occurrence d'un nouvel achat par catégorie de clients ou de provoquer celle-ci en intensifiant la publicité et/ou la promotion à un moment précis. En effet,la méthode RFM est aujourd'hui parfois remise en question quant à sa réelle pertinence pour la gestion efficace et rentable de la relation client.

Critiques de RFM :

Elle ne prend pas en compte la manière dont la volatilité passée d'un client quant à son comportement d'achat, peut affecter les modèles de consommation future de ce client, ou à quel point les modèles de différents clients peuvent être sélectivement affectés par le marché ou les forces macroéconomiques.

Etude comparative entre quelques algorithmes de segmentation :

- Le clustering hiérarchique ne peut pas gérer correctement les données volumineuses, mais K Cleansing peut le faire. En effet, la complexité temporelle de K Means est linéaire, c'est-à-dire O (n), alors que celle de clustering hiérarchique est quadratique, c'est-à-dire O (n2).
  La methode twostep traite les données de grandes tailles.
- Puisque nous commençons par un choix aléatoire de clusters dans le clustering K Means, les résultats produits en exécutant l'algorithme plusieurs fois peuvent différer. Alors que les résultats sont reproductibles dans la classification hiérarchique.
  Two-step de sa part, n'est pas reproductible puisque la première étape de cette méthode se base sur le K-Means.
- K Means fonctionne bien lorsque la forme des clusters est hyper sphérique (comme un cercle en 2D, une sphère en 3D).
  K-means nécessite une connaissance préalable de K le nombre de clusters que vous voulez avoir. Mais dans le clustering hiérarchique et en interprétant le dendrogramme, vous pouvez vous arrêter au nombre de clusters que vous trouvez approprié. Two steps nécessite une connaissance préalable de K puisque la première étape de cette méthode se base sur le K-Means.

Conclusion

Ce chapitre a été consacré à l'analyse du sujet en présentant la segmentation comme technique de datamining ainsi qu'une étude comparative entre quelques algorithmes de clustering.

CHAPITRE

3 Les outils de développement

Les outils de développement

IBM SPSS Modeler

IBM SPSS Modeler (*Statistical Package for the Social Sciences*) est une plateforme d'analyse prédictive qui vous aide à créer rapidement des modèles d'analyse prédictive et intégrer les éléments prédictifs aux décisions des individus, des équipes, des systèmes et l'entreprise. Il fournit toute une gamme d'algorithmes et de techniques d'analyse avancés, dont l'analytique textuelle, l'analyse d'entité, la gestion et l'optimisation des décisions, et fournit des connaissances quasiment en temps réel. Il vous permet de prendre régulièrement de bonnes décisions, depuis votre poste de travail ou dans les systèmes opérationnels.

Figure 5:IBM SPSS MODELER

Présentation de la méthode CRISP-DM

CRoss Industry Standard Process for Data mining

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter vos travaux de Data mining.

- En tant que méthodologie, CRISP-DM comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.
- En tant que modèle de processus, CRISP-DM offre un aperçu du cycle de vie du Data mining.

Figure 6: étapes du CRISP-DM

Le modèle de cycle de vie comporte six phases dotées de flèches indiquant les dépendances les plus importantes et les plus fréquentes entre les phases. La séquence des phases n'est pas strictement établie. De fait, les projets, pour la plupart, passent d'une phase à l'autre en fonction des besoins.

Adaptable, le modèle CRISP-DM peut être aisément personnalisé. Ainsi, si votre entreprise cherche à repérer un blanchiment d'argent, vous examinerez certainement une grande quantité de données sans objectif précis concernant la modélisation. Votre travail sera ciblé non sur la modélisation, mais sur l'exploration et la visualisation de données avec pour objectif de découvrir des configurations suspectes parmi les données financières. CRISP-DM vous permet de créer un modèle de Data mining adapté à vos besoins.

Dans une telle situation, les phases de modélisation, d'évaluation et de déploiement peuvent s'avérer d'un intérêt moindre que les phases de préparation et de compréhension des données. Toutefois, certaines des questions soulevées durant ces dernières phases sont tout de même à prendre en considération pour les planifications à long terme et les futurs objectifs de Data mining.

Les Phases de CRISP-DM

Phases	Tâches
- Compréhension du problème	▪ Déterminer les objectifs commerciaux. ▪ Evaluer la situation. ▪ Déterminer les objectives du Data Mining. ▪ Produire d'un plan du projet.
- Compréhension des données	▪ Collecte des données initiales ▪ Description des données ▪ exploration des données ▪ Vérification de la qualité des données
- Préparation des données	▪ Sélection des données ▪ Nettoyage des données ▪ Construction de nouvelles données ▪ Intégration des données ▪ Formatage des données
- Modélisation	▪ Sélection des techniques de modélisation. ▪ Génération d'une conception de test. ▪ Création des modèles. ▪ Evaluation de modèles.
- Evaluation	▪ Evaluation de résultats ▪ Processus de révision ▪ Détermination des étapes suivantes
- Déploiement	▪ Planification du déploiement ▪ Planification de surveillance et maintenance ▪ Production de rapport final ▪ Exécution d'une révision de projet final

Conclusion

Ce chapitre a été consacré à la présentation des différents outils de segmentation.

CHAPITRE

4 Mise en Œuvre

Mise en œuvre

En ce qui concerne la mise en œuvre du projet, on procèdera moyennant la méthode CRISP-DM Cross-Industry Standard Process for Data Mining, qui nous a permis d'orienter notre travail de Data Mining, où nous allons décrire les différentes étapes constituant cette dernière :

Les étapes de CRISP-DM sont les suivantes :

Compréhension du problème :

Détermination des objectifs stratégiques, opérationnels et Datamining:

En perpétuelle mutation, le marketing doit aujourd'hui agir subtilement pour attirer l'attention d'un consommateur constamment sollicité, et souvent agacé de recevoir des offres qui ne lui correspondent pas.

A l'heure où les réactions de masse s'effacent peu à peu face à un individualisme de plus en plus affirmé, le **processus de ciblage** devient primordial non seulement en termes de **pertinence de l'offre**, mais aussi en termes de **maîtrise de son R.O.I(Return On Investment),** qui permet de comparer des investissements en prenant en compte l'argent investi et l'argent gagné (ou perdu). Il permet d'orienter ses choix en matière d'investissements pour choisir le plus rentable.

C'est dans ce cadre, où s'introduit la nécessité de regarder en détails les caractéristiques de la clientèlede l'entreprise : Qui sont-ils ? Comment consomment-ils ? Quel est le type de clients les plus rentables ? C'est une question intéressante dans la mesure où elle permet d'identifier quels sont les consommateurs similaires à ceux existants qui pourraient profiter de votre offre.

Compréhension des données :

Cette Base de Données offre près de 40 millions de lignes de données transactionnelles complètement anonymisées à plus de 30 000 acheteurs.

Tous les champs sont anonymisés et catégorisés pour protéger les informations sur les clients et les ventes.

Dictionnaire de données :

Historie

id - Identifiant unique représentant un client

chain - Un entier représentant une chaîne de magasins

offer - Un identifiant représentant une offre

market - Un identifiant représentant une région géographique

repeattrips - Le nombre de fois que le client a fait un achat répété

repeater - Un booléen

offerdate - Date à laquelle un client a reçu l'offre

Transactions

id - voir ci-dessus

chain - voir ci-dessus

dept - Un regroupement agrégé de la catégorie (par exemple, eau)

category - La catégorie de produit (par exemple, eau pétillante)

company - Identifiant de la société qui vend l'article

brand - Identifiant de la marque à laquelle appartient l'article

date - La date d'achat

productsize - Le montant de l'achat du produit (par exemple 16 oz d'eau)

productmeasure - Les unités de l'achat du produit (par exemple onces)

purchasequantity - Le nombre d'unités achetées

purchaseeamount - Le montant en dollars de l'achat

Offers

offer - voir ci-dessus

category - voir ci-dessus

quantity - Le nombre d'unités que l'on doit acheter pour obtenir la réduction

company - voir ci-dessus

offervalue - La valeur en dollars de l'offre

brand - voir ci-dessus

Flux de données

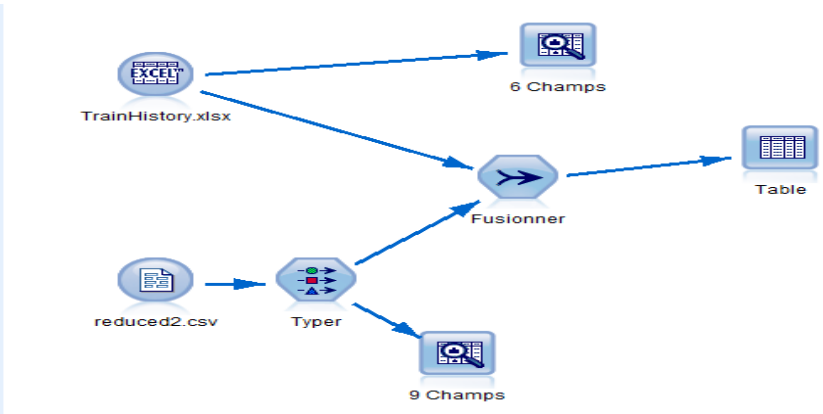


Figure 7: flux de données

Exploration des données :

	Champ	Graphique de l'échantillon	Mesure	Min.	Max.	Somme	Plage	Moyenne	Erreur standard moyenne	Ecart type	Ecart	Asymétrie	Erreur standard d'asymétrie	Kurtosis	Erreur standard Kurtosis	Médiane	Mode	Unique	Valide
1	chain		Nominale	2.000	526.000	--	--	--	--	--	--	--	--	--	--	--	21.000	130	14980932
2	offer		Nominale	1194044.000	1208503.000	--	--	--	--	--	--	--	--	--	--	--	1197502.000	24	14980932
3	market		Nominale	1.000	96.000	--	--	--	--	--	--	--	--	--	--	--	10.000	34	14980932
4	repeattrips		Continue	0.000	2124.000	255423531.000	2124.000	17.050	0.044	171.575	29437.959	12.083	0.001	145.154	0.001	0.000	0.000	--	14980932
5	repeater		Indicateur	--	--	--	--	--	--	--	--	--	--	--	--	--	f	2	14980932
6	offerdate		Continue	2013-03-01	2013-04-30	--	5184000.000	--	--	--	--	--	--	--	--	2013-04-05	2013-03-25	--	14980932
7	dept		Nominale	2	99	--	--	--	--	--	--	--	--	--	--	--	99	53	14980932
8	category		Nominale	201	9909	--	--	--	--	--	--	--	--	--	--	--	9909	203	14980932
9	brand		Nominale	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	14980932
10	date		Continue	2012-03-02	2013-04-29	--	36547200.000	--	--	--	--	--	--	--	--	2012-09-25	2013-02-02	--	14980932
11	productsize		Continue	0.000	1080.000	453921283.843	1080.000	30.300	0.011	44.300	1962.504	4.244	0.001	31.058	0.001	14.500	16.000	--	14980932
12	productmeasure		Nominale	--	--	--	--	--	--	--	--	--	--	--	--	--	OZ	6	14980932
13	purchasequantity		Continue	-17	2608	23770667	2625	1.587	0.001	4.162	17.326	111.121	0.001	28248.759	0.001	1	1	--	14980932
14	purchaseamount		Continue	-885.040	3248.980	82860793.258	4134.020	5.531	0.003	9.759	95.239	55.494	0.001	8752.922	0.001	3.960	3.990	--	14980932

Figure 8: audit des données de la table transaction

La figure ci-dessus représente l'ensemble des graphiques et des statistiques descriptives calculées pour chaque champ de la table transaction.

	Champ	Mesure	Valeurs extrêmes	Extrêmes	Action	Attribuer une entrée manquante	Méthode	% terminé	Enregistrements valides	Valeur nulle	Chaine vide	Blanc	Valeur non renseignée
1	chain	Nominale	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
2	offer	Nominale	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
3	market	Nominale	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
4	repeattrips	Continue	10527	97723	Aucun	Jamais	Fixe	100.000	14980932	0	0	0	0
5	repeater	Indicateur	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
6	offerdate	Continue	0	0	Aucun	Jamais	Fixe	100.000	14980932	0	0	0	0
7	dept	Nominale	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
8	category	Nominale	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
9	brand	Nominale	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
10	date	Continue	0	0	Aucun	Jamais	Fixe	100.000	14980932	0	0	0	0
11	productsize	Continue	274098	42935	Aucun	Jamais	Fixe	100.000	14980932	0	0	0	0
12	productmeasure	Nominale	--	--	--	Jamais	Fixe	100.000	14980932	0	0	0	0
13	purchasequantity	Continue	25905	28619	Aucun	Jamais	Fixe	100.000	14980932	0	0	0	0
14	purchaseamount	Continue	61126	45021	Aucun	Jamais	Fixe	100.000	14980932	0	0	0	0

Figure 9: audit des données de la table transaction -suite-

Cette table permet de voir les informations sur les champs notamment les valeurs extrêmes, chaine vide ou valeur non renseignée. En effet, nous pouvons remarquer qu'on a dans aucun cas une donnée manquante.

Valeur	Proportion	%	Comptage
1200579.000		0.44	65278
1200582.000		0.48	72072
1198271.000		1.0	149857
1198274.000		1.12	167562
1198273.000		1.13	168850
1204821.000		1.14	170231
1204822.000		1.14	170559
1194044.000		1.32	198376
1208503.000		1.5	224674
1198272.000		1.52	228304
1200578.000		1.96	294157
1198275.000		2.24	336258
1208252.000		2.44	364987
1200581.000		3.72	556685
1208501.000		3.96	592636
1199258.000		4.17	624605
1200988.000		4.26	638037
1199256.000		4.34	649701
1204576.000		4.76	713622
1203052.000		4.77	715312
1208251.000		12.25	1835672
1208329.000		17.75	2658417
1197502.000		22.5	3370105
1200584.000		0.1	14975

Figure 10: distribution de la variable offer

Cette figure nous permet de voir que les clients ayant un id de 1197502 et 1208329 sont les plus intéressés par les offres proposés par la company du fait qu'ils essaient d'en profiter le plus.

Valeur	Proportion	%	Comptage
f		63.71	9545065
t		36.29	5435867

Figure 11: distribution de la variable repeater

Parmi les clients de la table transaction on peut voir que ceux qui sont revenu aux magasins plus qu'une fois est plus petit (36%) que ceux qui les ont visité une seule fois (64%).

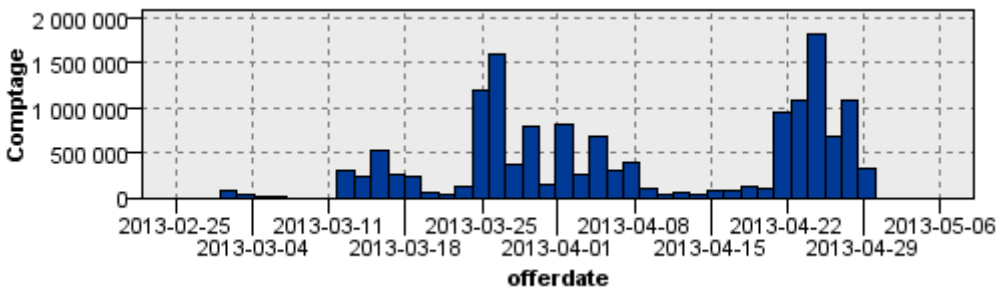
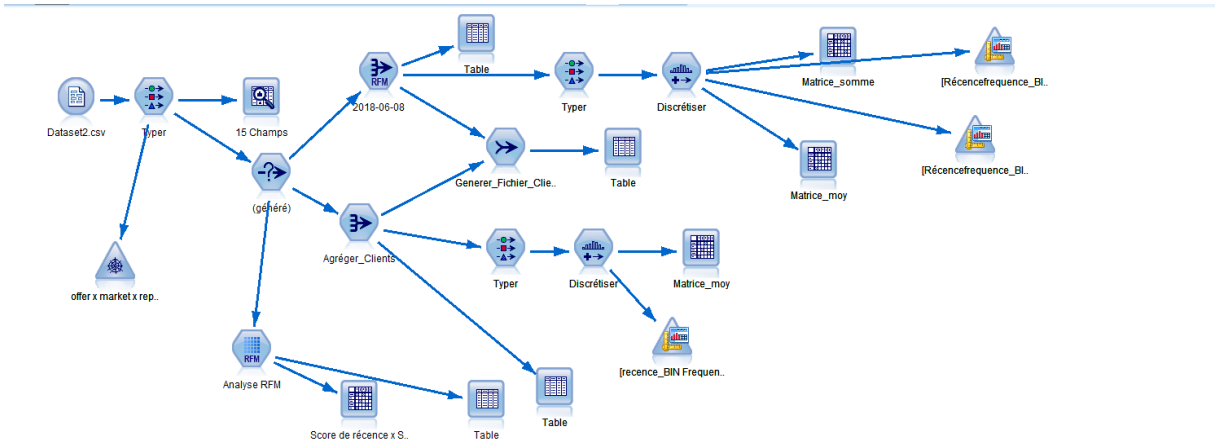


Figure 12: distribution de la variable offerdate

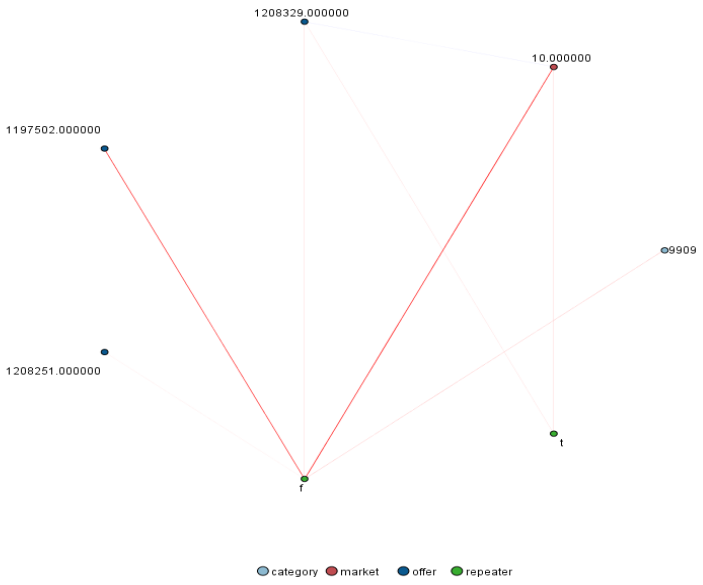
Cette figure nous permet de distinguer un fort intérêt vers les offres pendant la fin du mois de mars ainsi qu'entre le 22 avril au 29 avril. Ceci semble justifiable par le fait que les clients cherche les bons plans durant les périodes de fin de mois du moins en ce qui concerne les salariés.

Préparation des données :



Afin de preparer notre data afin de lui appliquer dessus un modèle nous permettant d'extraire de l'information pertinente nous avons procéder par effectuer du typage, et de l´audit pour détecter existence de valeurs manquantes ou des anomalies qu´on a éliminer par la suite à travers un nettoyage de données ; ainsi nous avons fait le choix de passer à travers des agrégations RFM tant qu'on est face à un concept de profilage de clients ainsi une de création de nouvelle variable associées à l'id de chaque client finalement on s'est trouvé avec un fichier contentent des variables continues décrivant chaque client

Relations intra variables :



Ce graphe ou réseau a un concept proche au règles d'associations par exemple celui-ci nous permet de voir que dans la plupart des transactions la plupart des clients ayant achetés le produit référé par l'id 1197502000000 dans le market référé par l'id 10000000 n'ont pas été satisfait de cet offre est ne l'ont pas acheté une autre fois .

Nettoyage de données :

La figure ci-dessous représente le résultat de l'audit effectué sur la table transaction concernant le champs repeattrips. On remarque un outlier de valeur 2124 qu'on a éliminé en guise d'éliminer le bruitage.

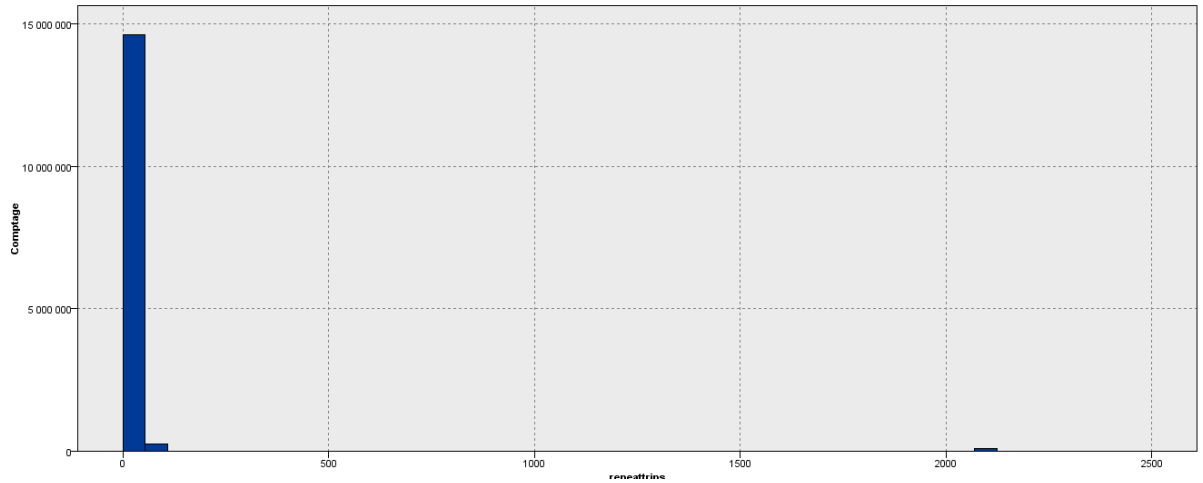


Figure 13: audit repeattrips

Construction de nouvelles données :

- Somme(repeattrips) pour une offre ;
- somme(repeattrips pour tous les offres) pour un client ;
- length of customership=date_dernierachat-date_premierachat;
- Capacité d'achat ;
- récence=date système – date dernier achat
- ancienneté= date système – date premier achat

La matrice RFM :

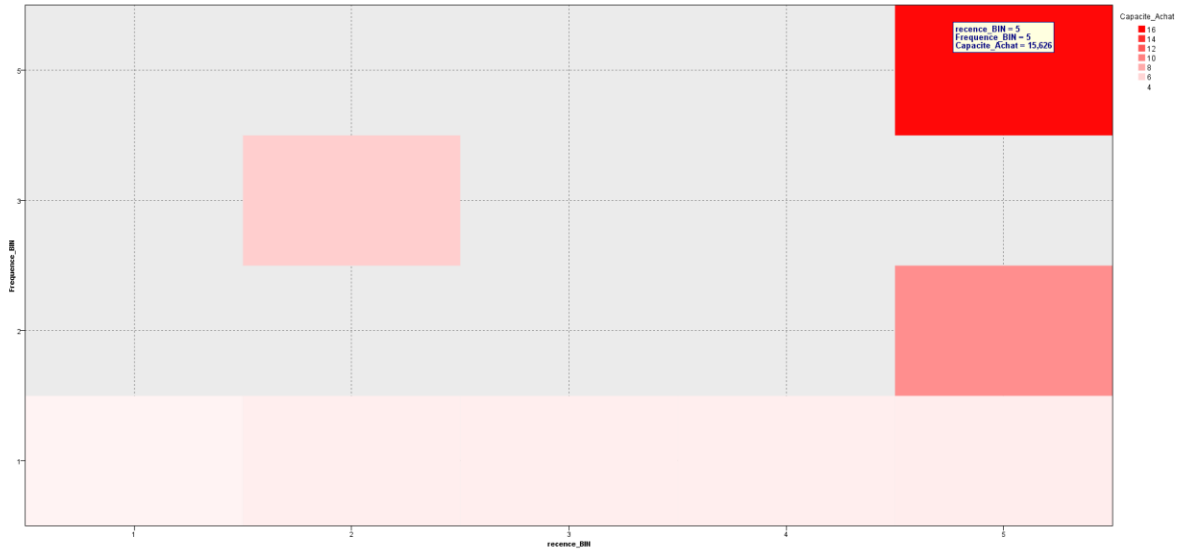


Figure 14: matrice RFM

Cette figure permet de voir quatre catégories principales des clients :

- **Première catégorie :**leurs capacités d'achat est minimale, leur récence s'étale sur tout l'axe de récence et leur fréquence d'achat (la variété des produits achetés) est minimale. Ces clients sont à garder puisqu'ils constituent la base de clientèle du magasin.
- **Deuxième catégorie :**leurs capacités d'achat est élevée par rapport à l'autre catégorie, leur récence est minimale c.à.d. qu'ils peuvent être des clients potentiels et leur fréquence d'achat (la variété des produits achetés) est plus ou moins élevée. Ces clients sont à fidéliser en créant des offres susceptibles de les intéresser.
- **Troisième catégorie :** leurs capacités d'achat est élevée par rapport aux catégories précédentes, leur récence est maximale c.à.d. que cela fait du temps depuis leurs dernière visite du magasin et leur fréquence d'achat est faible. Ces clients semblent ne plus être intéresser par le magasin ou bien ils étaient des clients one shoot, à titre d'instance des personnes qui ont décidé d'aménager leur maison une fois pour toute ou bien ils voulaient un seul article. Il n'est plus intéressant de les convaincre de revenir.
- **Quatrième catégorie :** leurs capacités d'achat est la plus élevée, leur récence, leur fréquence d'achat est aussi élevée néanmoins leur récence est élevée, cela fait du temps depuis leur dernière visite. Ces clients gold sont à refidéliser en créant des offres susceptibles de les intéresser puisqu'ils sont une plus-value au magasin.

Intégration des données :

Fusion des données

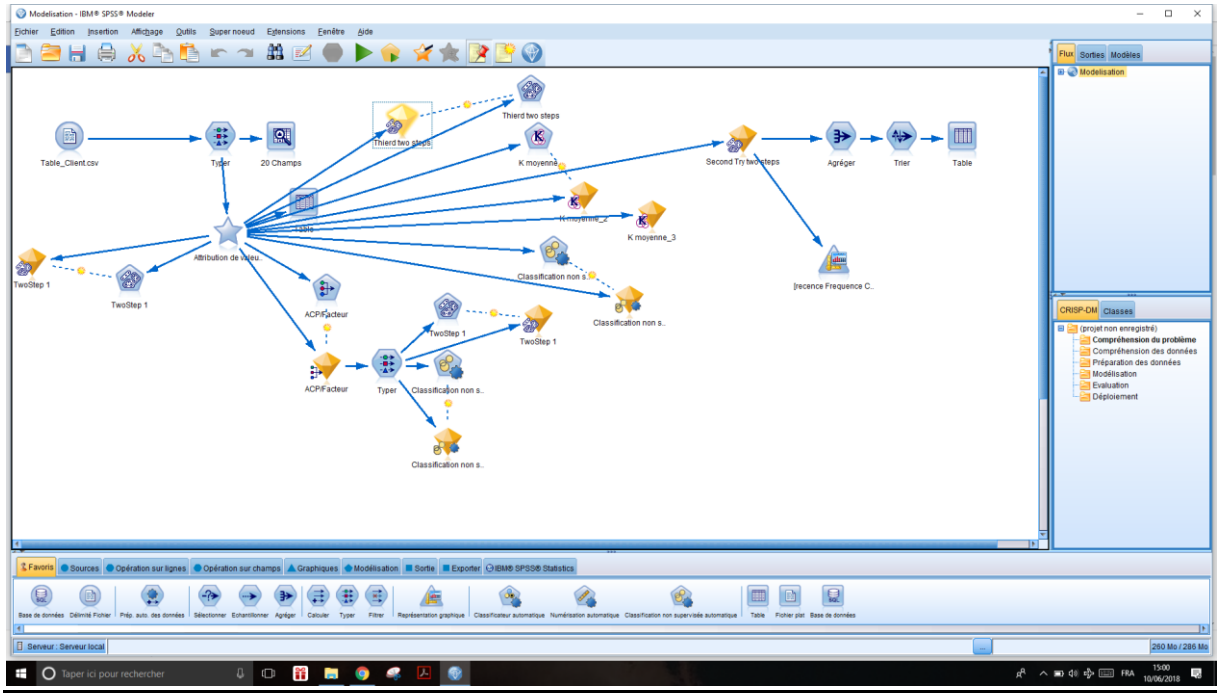
Nous avons fusionné les tables transactions et trainHistory et offre en guise de collecter davantage de variables ainsi que réduire la taille de la table transaction qui atteignait plus de 20Go en une base de données de taille 1.6 Go.

Modélisation :

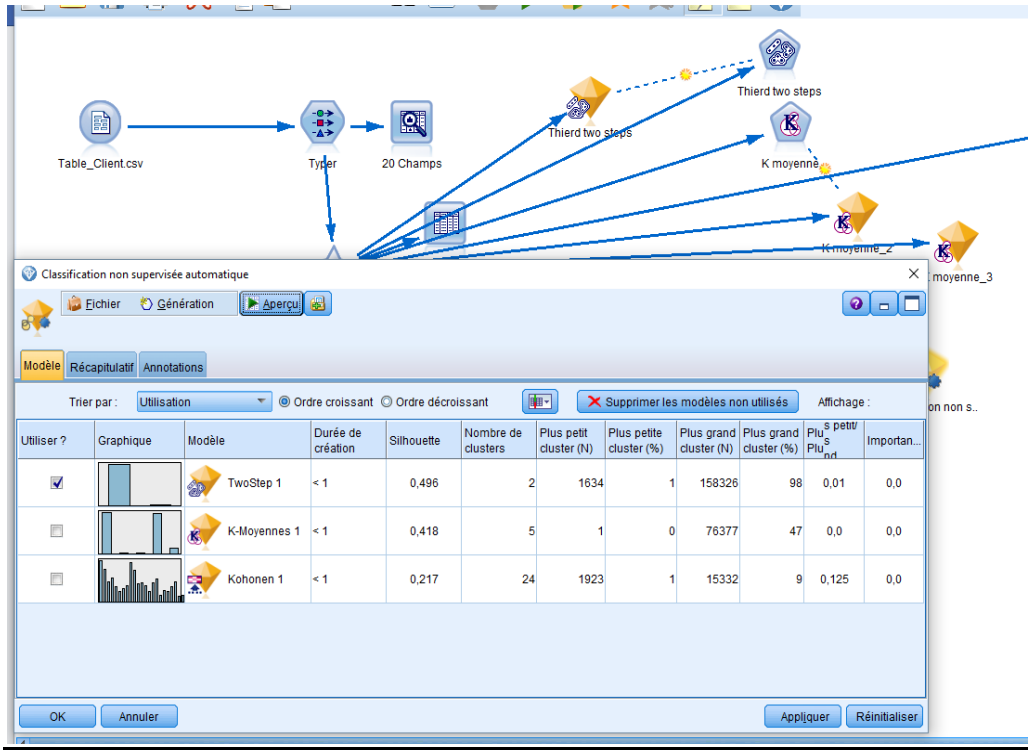
après que notre data est prête nous li appliquerons un algorithme le plus adéquat à notre situation vu que notre fichier Client exploité ne contient que des variables continues associés aux Id du client nous pourrons lui appliquer dessus tout algorithme que l'on désire à condition qu'il soit robuste face aux données massives comme est le cas de notre data.

Vu que l'algorithme du kohonen donne des solutions difficile à interpréter ; est qu'il se peut qu'il ne converge pas ,il nous restait d'utiliser soit l'algorithme de k-Means et d'un algorithme qui dérive de ce dernier l'algorithme de Two steps ; tout on prenant en considération que le k-Means trouve une difficulté à traiter des data nom sphérique pas come l'algorithme Two steps qui a fait preuve de détecter les similarités et de produire des clusters optimaux même lorsque notre data est un peu messy. On verra par la suite si c'est notre cas ou nom .

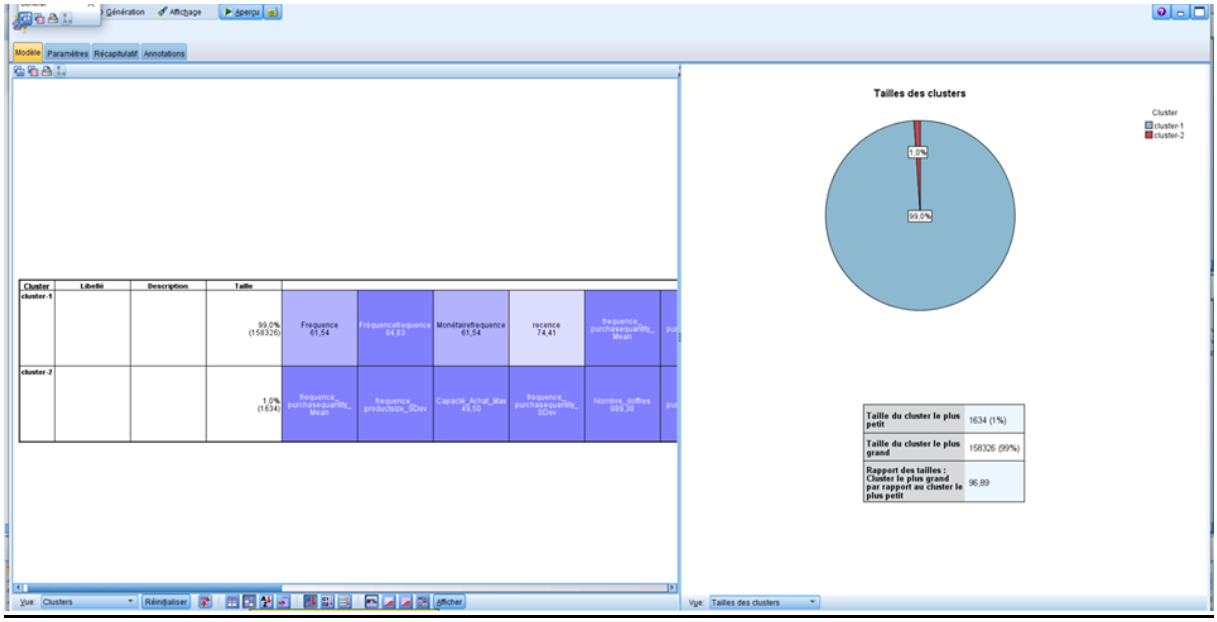
Bon voila notre démarche de travail :



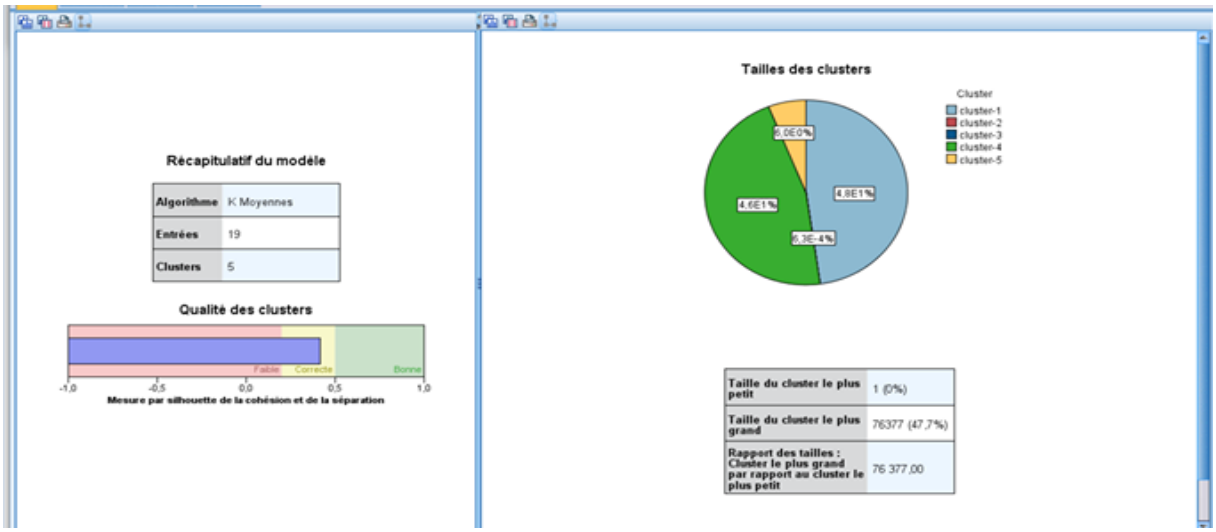
On a du passer à travers plusieurs itérations afin d'atteinre des résultats convenables où chaque fois notre critère de comparaison d'augmentation de la silouhaite et d'avoir un rapport de taille entre cluster qui n'est pas exagéré . On a commencé par appliquer un noud qui représente un système de recommandations de cluster à utiliser qui nous a recommandé d'utiliser l'algorithme de Two steps

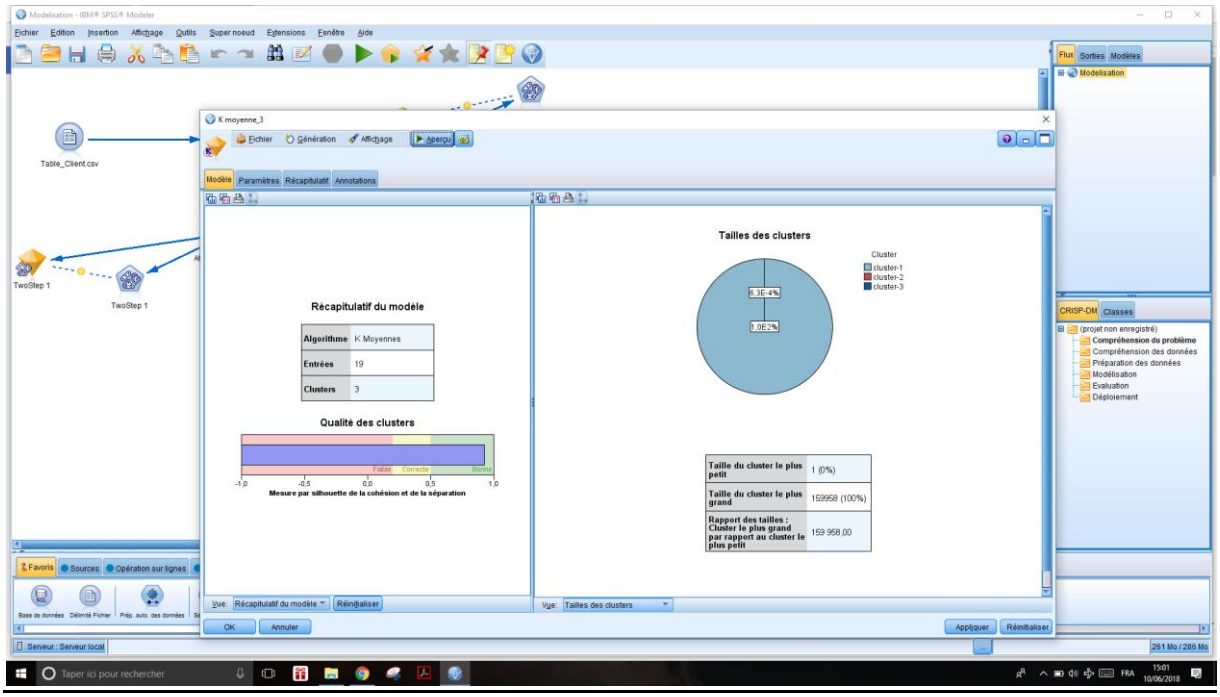


Notre premier algorithme modèle Two steps basé sur un nombre de clusters détectés d'une manière automatique nous à donner deux clusters avec une silouhaite de 0.496 qui est considérablement bonne or l'un des deux clusters (le bleu) l'emporte sur l'autre donc on ne pourrait pas prendre ce modèle



On ainsi tenter avec le k Means avec un k déterminé automatiquement et un k=3 or ceci n'a pas aboutis à de bons résultats nom plus car les résultats ne semblaient pas être raisonnable.





On a pensé ainsi de passer par appliquer l'ACP sur notre data afin de rendre les résultats plus homogène on a réduit ainsi le nombre d'axes à 6 axes (ainsi on a répéter la même démarche qu'avant) voici si dessous le résultat de l'application de l'ACP sur notre données continues du fichier client:

Qualités de représentation

	Initiales	Extraction
Récencefrequence	1,000	,989
Fréquencefrequence	1,000	,897
Monétairefrequence	1,000	,936
frequence_repeattrips_Mean	1,000	,998
frequence_repeattrips_Median	1,000	,998
frequence_productsize_Mean	1,000	,943
frequence_productsize_SDev	1,000	,717
frequence_productsize_Median	1,000	,569
frequence_purchasequantity_Mean	1,000	,901
frequence_purchasequantity_SDev	1,000	,737
frequence_purchasequantity_Median	1,000	,550
Nombre_doffres	1,000	,897
Frequence	1,000	,936
recence	1,000	,147
ancienete	1,000	,989
Capacite_Achat	1,000	,426
Capacté_Achat_Max	1,000	,697
Capacté_Achat_Min	1,000	,546

Méthode d'extraction : Analyse en composantes principales.

Variance totale expliquée

Composante	Valeurs propres initiales			Sommes extraites du carré des	
	Total	% de la variance	% cumulé	Total	% de la variance
1	3,861	21,449	21,449	3,861	21,449
2	2,793	15,518	36,967	2,793	15,518
3	2,263	12,574	49,542	2,263	12,574
4	1,990	11,058	60,599	1,990	11,058
5	1,800	10,000	70,599	1,800	10,000
6	1,164	6,466	77,066	1,164	6,466
7	,964	5,356	82,421		
8	,890	4,946	87,368		
9	,692	3,845	91,213		
10	,632	3,512	94,725		
11	,529	2,940	97,665		
12	,285	1,581	99,247		
13	,075	,416	99,663		
14	,061	,337	100,000		
15	5,161E-8	2,867E-7	100,000		
16	1,065E-17	5,916E-17	100,000		
17	-2,002E-16	-1,112E-15	100,000		
18	-2,769E-16	-1,539E-15	100,000		

Variance totale expliquée

Composante	Sommes ...	Sommes de rotation du carré des chargements		
	% cumulé	Total	% de la variance	% cumulé
1	21,449	2,911	16,171	16,171
2	36,967	2,511	13,948	30,119
3	49,542	2,351	13,064	43,182
4	60,599	2,058	11,432	54,614
5	70,599	2,032	11,290	65,904
6	77,066	2,009	11,162	77,066
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				

Matrice des composantes

	Composante					
	1	2	3	4	5	6
Nombre_doffres	,863					
Fréquencefrequence	,863					
Frequence	,755	-,351				,458
Monétairefrequence	,755	-,351				,458
Capacté_Achat_Max	,655					-,388
frequence_productsize_Mean		,679	-,417		,515	
frequence_productsize_SDev		,586	-,412		,418	
Capacite_Achat		,562				
frequence_productsize_Median		,510			,455	
frequence_purchasequantity_Mean	,302	,566	,600			
frequence_purchasequantity_SDev	,348	,490	,529		-,300	
frequence_purchasequantity_Median		,411	,487			
frequence_repeattrips_Mean			-,359	,794	-,306	
frequence_repeattrips_Median			-,359	,794	-,306	
Récencefrequence		-,318	,569	,394	,592	
ancienete		-,318	,569	,394	,592	
Capacté_Achat_Min	-,441					,507
recence						

Méthode d'extraction : Analyse en composantes principales.

Rotation de la matrice des composantes

	Composante					
	1	2	3	4	5	6
Frequence	,958					
Monétairefrequence	,958					
Nombre_doffres	,701			,636		
Fréquencefrequence	,701			,636		
frequence_purchasequantity_Mean		,949				
frequence_purchasequantity_SDev		,838				
frequence_purchasequantity_Median		,723				
Capacite_Achat		,518	,367			
frequence_productsize_Mean			,966			
frequence_productsize_SDev			,832			
frequence_productsize_Median			,744			
Capacté_Achat_Max		,321		,732		
Capacté_Achat_Min				-,728		
recence						
Récencefrequence					,990	
ancienete					,990	
frequence_repeattrips_Mean						,992
frequence_repeattrips_Median						,992

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation Kaiser.[a]

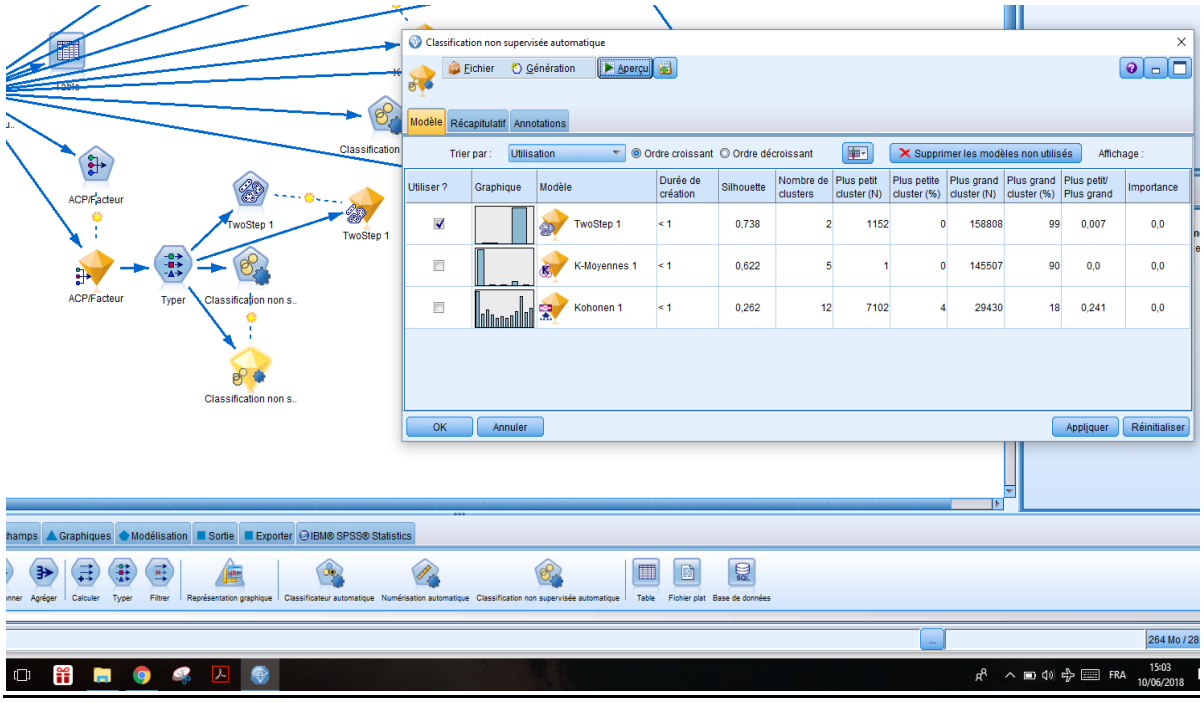
a. Convergence de la rotation dans 6 itérations.

Matrice de transformation des composantes

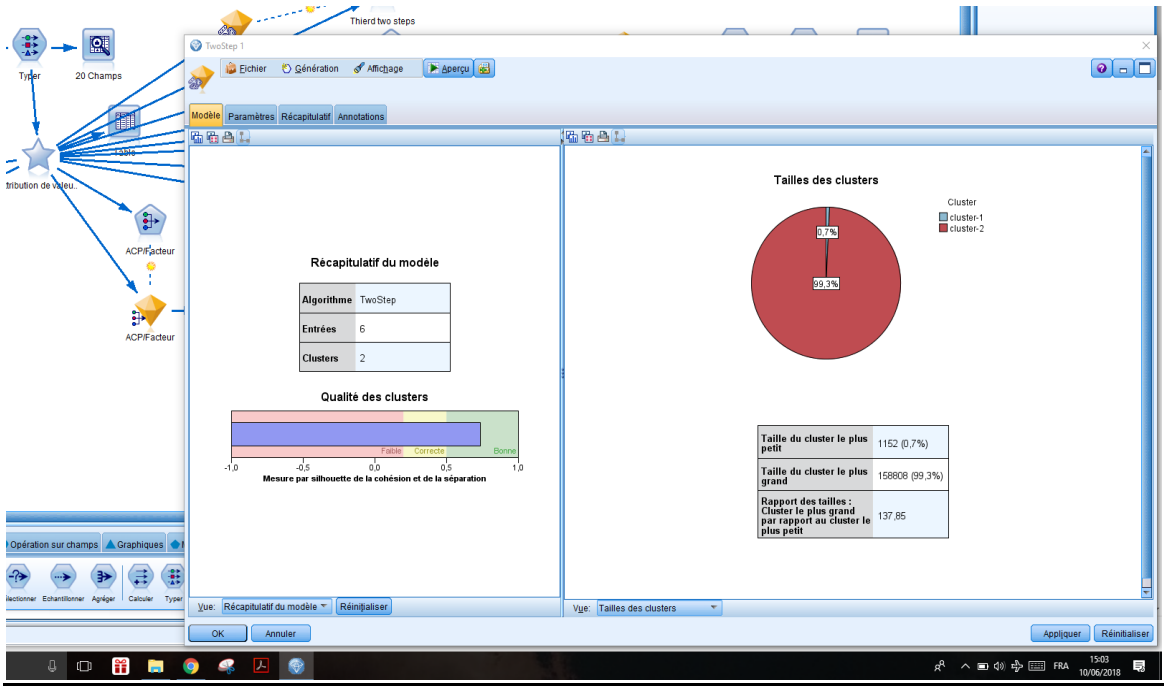
Composante	1	2	3	4	5	6
1	,747	,293	,153	,535	-,100	,193
2	-,370	,593	,651	-,015	-,282	-,090
3	-,039	,641	-,415	,047	,546	-,339
4	,020	,191	,110	-,381	,395	,806
5	,169	-,309	,607	-,004	,634	-,326
6	,524	,141	,011	-,753	-,235	-,289

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation Kaiser.

Et on a appliqué de nouveau la même démarche

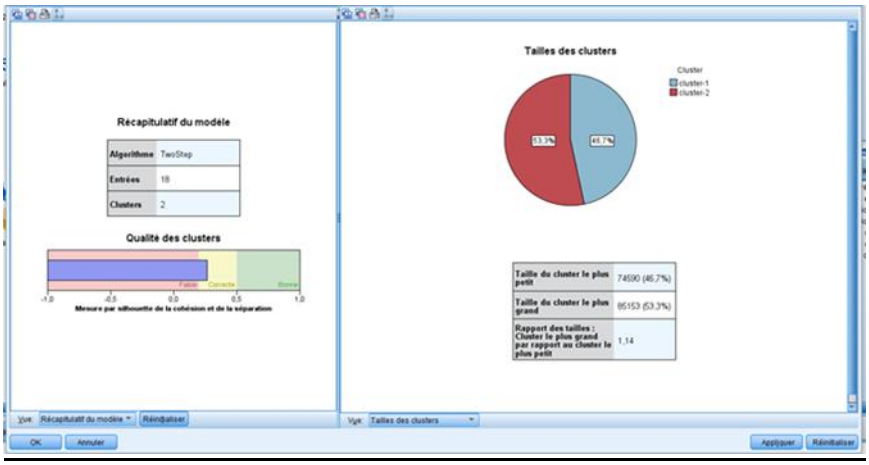


Ceci a permis d'améliorer le niveau de la silouhaite

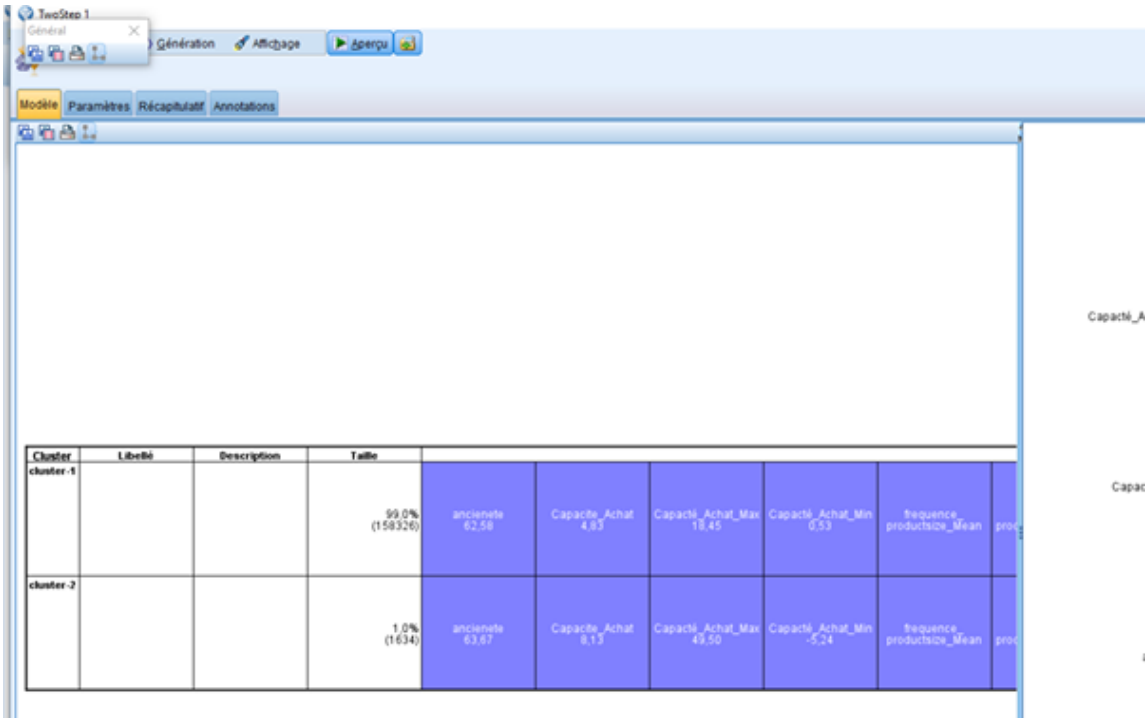


Or ceci à donner des résultats similaire à celle précédentes; et lorsqu'on a essayer de forcer d'avoir un nombre de cluster k=3 et plus donc utiliser une distance euclidienne cette fois le système bugger à chaque fois et nous retourner des erreurs .

Donc on a décidé de se contenter de forcer le système à avoir 3 cluster en se basant sur la visualisation de la matrice RFM expliquée avant . Cette fois ci on a eu une silouhaite acceptable



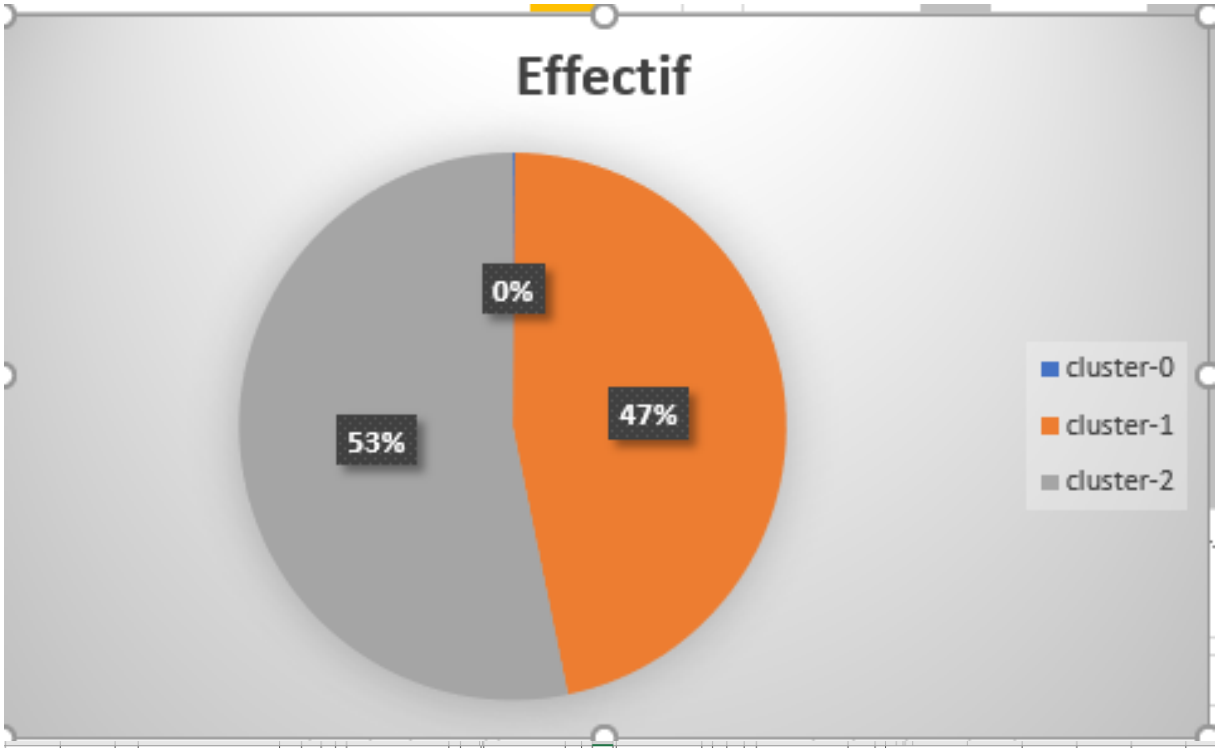
Dont nos clusters sont basés principalement sur l'indice de l'ancienneté et de capacité d'achat.



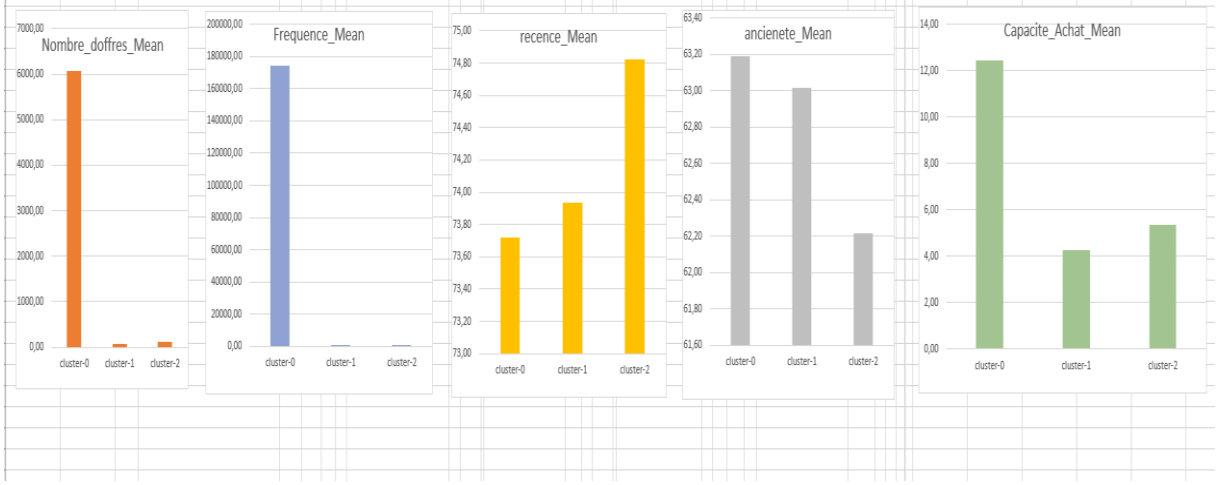
Evaluation :

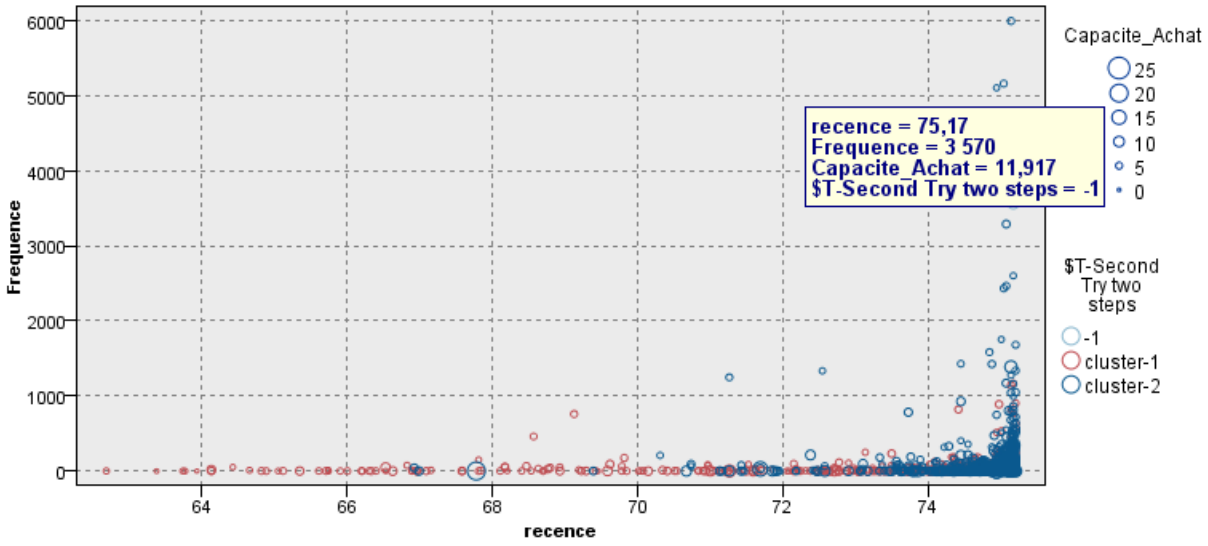
Cette phase consiste en gros à valider notre modèle par rapport au hypothèse prises au préalable à travers la matrice RFM ou comme le montre notre graphe de bulles des fréquences en fonction de la récence ; ainsi on a un chevauchement de données entre celle du cluster 1 et 2 ou comme on a prédit on a une classe qui s'étale sur l'axe de la récence qui ne se rendent pas comptent aussi fréquemment au magasin et qui ont une capacité d'achat basse une bleu qui a une valeur grande par

rapport à la récence c'est qu'il a longtemps qu'ils nous ont pas rendu visite mais qui achètent avec une fréquence moyenne et un troisième le gri peut être rare mais ayant une bonne capacité d'achat ils sont a la fois des anciens client et se rendent aussi récemment pour acheter des articles de nos magasins ce qu'ont va vous présenter à travers les graphes suivants :



ST-Second Tr	Effectif	Nom	Nombre_doffres_Mean	Nom	Nom	No	Fr	Frequence_Mean	Fr	Freq	recence_Mean	rec	re	rece	ancienete_Mean	ar	an	anc	Ca	Capacite_Achat_Mea	Capac	Ca	Capacite_Achat_Median
cluster-0	217,00	####	6071,86	1,00	###	##	#	174328,45	#	###	73,72	##	#	###	63,19	#	##	##	#	12,47	#####	#	#
cluster-1	74590,00	####	60,28	1,00	###	##	#	28,08	#	###	73,94	##	#	###	63,02	#	##	##	#	4,27	-1,01	#	#
cluster-2	85153,00	####	106,51	1,00	###	##	#	98,34	#	###	74,82	##	#	###	62,21	#	##	##	#	5,37	1,05	#	#





C'est vrai que pour notre cas dans la matrice on distinguer l'existence d'une quatrième classe; or il faut prendre toujours en compte qu'on a effectué un nettoyage de données à 2 reprises.

Déploiement :

Un modèle déployé pour développer la fidélisation de la clientèle parmi les clients les plus importants aura probablement besoin d'être réajusté une fois qu'un niveau particulier de fidélisation aura été atteint. Le modèle peut être modifié et réutilisé pour conserver les clients appartenant à un niveau inférieur mais néanmoins rentable, sur la pyramide des valeurs.

Il faut ainsi faire à chaque fois faire une mise à jour des données pour garantir une fiabilité des interprétations.

CONCLUSION

A travers ces différents chapitres que nous avons dévoilé, en première partie, le contexte du projet et sa finalité. Dans la deuxième partie, nous avons mis la lumière sur la phase d'étude et d'analyse en amont. Dans le troisième volet, nous avons défini les outils de développement utilisés pour réaliser le projet.

Enfin, nous avons entamé la mise en œuvre de l'application en ayant tenu compte de la solution proposée dans la partie analyse. Nous nous sommes basées également sur des « captures-écran » afin de concrétiser la démarche adoptée.
Par cet itinéraire parcouru, nous avons voulu présenter une succession logique et argumentée d'un rapport qui respecte les normes d'un écrit académique.

Quant à l'apport de ce projet, ce dernier nous a aidé à améliorer les connaissances nécessaires à l'élaboration des projets datamining suivant la démarche CRISP-DM et nous a donné l'opportunité d'apprécier le travail en groupe qui s'avère très enrichissant pour ce genre de projets.

Perspective

Le travail que nous avons réalisé est satisfaisant, toutefois pour améliorer le projet on peut ajouter les règles d'association pour adapter des offres au besoins des clients segmentés.

Webographie

http://www.spssmaroc.ma/datamining/crisp.php

http://www.mywebmarketing.fr/la-methode-rfm-segmentation-de-clients/

http://lms.inead.fr/courses/course_modules/TU9u4vIz2L6Gkocr0UMgZLXQsO2kUt3p/scorm/co/MSE1-C07-ciblage.html

https://fr.wikipedia.org/wiki/Cible_(marketing)

http://www.e-marketing.fr/Definitions-Glossaire/Marketing-cible-242230.htm#x8OHbpJcz00ewMLJ.97

https://solutions.lesechos.fr/com-marketing/c/mettre-place-strategie-de-communication-efficace-entreprise-4142/