

Shuai WANG

Rapport sur le stage effectué du 14/04/2014 au 26/09/2014

Dans la Société : IZICAP

à Sophia Antipolis

Pour l'obtention du Master Statistique Informatique et Techniques
Numériques

Le 12 Septembre 2014

De l'université Université Claude Bernard Lyon 1



Remerciements

Je tiens à remercier toute l'équipe d'IZICAP, pour son accueil, sa gentillesse et l'aide qu'elle m'a apportée durant ce stage et tout particulièrement mon maître de stage, Monsieur Jérôme SAUNIER, pour m'avoir aidé intégrer rapidement dans l'équipe, m'avoir accordé toute sa confiance et ses conseils dans les élaboration de ce rapport.

Je remercie sincèrement mon professeur et encadrant de stage, Monsieur François Wahl, pour sa gentillesse, ses conseils concernant les missions évoquées dans ce rapport et ses explications en statistiques.

Enfin, je souhaite remercier, toute l'équipe pédagogique de l'Université Lyon 1 et les professeurs de SITN pour avoir assuré ma formation, qui a servi à la partie théorique de ce rapport.

Résumé

Ce stage se déroule dans une nouvelle startup IZICAP à Sophia Antipolis, avec toutes les transactions de cartes bancaires passées dans une banque précise, nous avons créé un site www.twiing.com pour fonder trois ponts entre les commerces de proximité, les clients et ses cartes bancaires. On offre aux commerçants le marketing analytique, et on leur propose d'offrir à leurs clients les promotions avec ses cartes bancaires.

Sur le site, on offre le dashboard aux commerçants, ma mission principale est d'enrichir et remplir la base de données de notre site et faire un partitionnement de commerces ayant la même activité, par exemple la coiffure, afin que le commerçant puisse comparer son indicateur clé de performance avec celui des commerces similaires.

Ce rapport présentera tout d'abord l'entreprise et le domaine du stage, puis la présentation de 4 missions auxquelles j'ai participé. Enfin viendront la conclusion et les questions à améliorer.

Mots Clés : commerce de proximité, base de données, kmeans, arbre de décision

Table des matières

Remerciements.....	1
Résumé	2
1. Introduction	4
1.1. Izicap.....	4
1.2. Mission principale	5
1.3. Contexte.....	5
2. Présentation des missions.....	6
2.1. Web scraping pour les informations et les coordonnées géographiques de commerçants ...	6
2.2. Corriger et améliorer la base de données pour le dashboard de notre site.....	7
2.3. Automatiser la génération des « reporting » destinées à l'élaboration des recommandations	
12	
2.4. Partitionnement des commerçants des Alpes Maritime via les méthodes ACP et K-means et mise en place de règles d'affectation via un arbre de décision	13
2.4.1. Traitement des données	13
2.4.2. Introduction des indicateurs.....	13
2.4.3. La méthode de partitionnement : k-moyennes	14
2.4.4. La méthode de l'ACP (Analyse en composantes principales)	16
2.4.5. La méthode de l'arbre de décision	16
2.4.6. La méthodologie et les étapes	17
2.4.7. Les résultats	23
3. Conclusion et discussion	29
4. Questions	30
Annexe	31
Bibliographie.....	35

1. Introduction

1.1. Izicap

Izicap est une nouvelle start-up à la croisée des chemins entre l'univers du paiement et le marketing analytique. En partenariat avec les banques acquéreurs, Izicap exploite les gisements de données de paiements (on-line et off-line) et y associe un savoir-faire datamining afin d'optimiser les promotions des commerçants et de leur permettre de lisser l'activité dans le temps (yield).

A travers sa plateforme « card link », Izicap propose une solution unique en France permettant de transformer la carte de paiement des clients en un programme marketing et de fidélisation ultra-personnalisé.

Une solution innovante au service des commerces, des banques et des consommateurs.

Quant au client, la simple utilisation de sa carte bancaire, lui permettra de gagner de l'argent à chaque passage en caisse.

Izicap permet ainsi de transformer la carte de paiement du client en programme de fidélisation ultra-personnalisé, ultra-simple et 100% dématérialisé

Izicap procure aux institutions financières une capacité sans précédent pour fournir un service marketing personnalisé à leurs clients via multiples canaux.

Fidélisation

Proposez un nouveau service à valeur ajoutée à vos clientèles commerçants en leur permettant d'accéder aux dernières techniques de « Customer Base Management » et d'acquérir de nouveaux clients en les incitant à revenir régulièrement pour effectuer de nouveaux achats à travers un programme de récompense personnalisé et très ciblé.

Revenus

En dynamisant les ventes de vos commerçants tous les acteurs bénéficient de revenus supplémentaires
Commerçants – Banque - Clients

Data

Avec la plateforme Izicap, l'analyse des données des transactions de la carte bancaire permet aux banques de proposer à leurs clients des offres sur mesure.

(Ces informations viennent du site internet du [1].)

1.2. Mission principale

- Web scraping pour enrichir la base de données commerçants des codes NAF (Nomenclature d'activités française) et des coordonnées géographiques
- Améliorer la base de données, récrire les requêtes SQL pour le dashboard de notre site
- Automatiser la génération des reporting destinée à l'élaboration des recommandations
- Partitionnement des commerçants des Alpes Maritime via les méthodes ACP et K-means et mise en place de règles d'affectation via un arbre de décision

1.3. Contexte

Nous avons les transactions de cartes bancaires d'une Banque de la région PACA, ce sont les données de 2 614 commerçants de proximité avec 8 032 115 lignes de transactions, soit 355 restaurants traditionnels, 118 dans la restauration de type rapide et 111 coiffeurs, etc. Avec toutes ces données nous avons créé le site « Twiing », mon travail principal est pour ce site.

2. Présentation des missions

2.1. Web scraping pour les informations et les coordonnées géographiques de commerçants

On programme en R et obtient les informations comme le nom, le code NAF, le libellé NAF, l'adresse, la ville, le code postal et le nombre d'établissements de 2 421 commerçants sur le site de www.infogreffe.com [2]

Figure 1 : le site de www.infogreffe.com

The screenshot shows the infogreffe.com homepage with a search bar at the top. Below it, a card displays information for 'FERRARO JEAN-LOUIS'. The card includes the address '310 474 317 R.C.S. GRASSE Greffe du Tribunal de Commerce de GRASSE', a button to 'Surveiller cet établissement', a dropdown menu for 'EFFECTUER UNE FORMALITÉ' (set to 'Sélectionner'), and a link to 'NOUVELLE RECHERCHE AVANCÉE'.

INFORMATIONS SUR L'ENTREPRISE

Below the card, there's a navigation bar with tabs: IDENTITÉ (selected), ÉTABLISSEMENT(S), ACTES DÉPOSÉS, ANNONCES BODACC, and VOIR LES DOCUMENTS OFFICIELS.

ETABLISSEMENT PRINCIPAL

- 19 ROUTE DU PLAN
06130 GRASSE
- [Voir le plan](#)

SIRET

310 474 317 00115

ENSEIGNE

LE TILOULOU

ACTIVITÉ (CODE NAF)

5610A : Restauration traditionnelle

INSCRIPTION

Immatriculée le 07/02/2011.
Société dans le ressort du greffe de GRASSE depuis le 01/12/2010.
Siège social antérieur dans le ressort du greffe de CANNES
[Cliquez ici pour accéder aux informations de l'ancien siège](#)

DERNIERS CHIFFRES CLÉS

Les entreprises en nom personnel ne sont pas tenues de déposer leurs comptes au Greffe

Et grâce à API de google maps [3] et R, on s'occupe des coordonnées géographiques, comme longitude et latitude des commerces grâce à leur adresse. Le package utilisé : RCurl, RJSONIO, XML.

Figure 2 : le résultat

E12	A	B	C	E	H	I	J	formated_address
1	siret	siren	name	adresse2	lat	lng	location_type	
2	30186707300027	301867073 JEAN Robert Emile Armand		1 LA PLACETTE , SAINT-PAUL	43.8047802	7.4967595	RANGE_INTERPOLATED	1 La Placette, 06500 Ca:
3	30213805200014	302138052 BAR RESTAURANT DES FLEURS		QUARTIER SAINT AUGUSTIN, NICE	43.6785503	7.2277241	GEOMETRIC_CENTER	Avenue Saint-Augustin, I
4	30227728000038	302277280 NOEL CHRISTIAN		3 PLACE DE GAULLE , BIOT	43.5798247	7.0590392	RANGE_INTERPOLATED	3 Avenue du Général de
5	30399723300018	303997233 CLINIQUE DU CHEVEU		41 RUE HOTEL DES POSTES, NICE	43.6992665	7.2712249	ROOFTOP	41 Rue de l'Hôtel des P
6	31006908300015	310069083 BASTANTI ET FILS		LE SCHUSS, ISOLA 2000	42.61316	9.298677	GEOMETRIC_CENTER	Isola, 20246 Piève, Fran
7	31014724400019	310147244 LE SANT ANA		LE PORT , SAINT-LAURENT-DU-VAR	43.224088	6.573329	APPROXIMATE	Saint-Laurent, 83580 Ga
8	31047431700115	310474317 FERRARO JEAN-LOUIS		19 ROUTE DU PLAN, GRASSE	43.6526103	6.946918	RANGE_INTERPOLATED	19 Route du Plan, 0613
9	31055242700018	310552427 TOCADE		3 AVENUE AUGUSTE RENOIR , CAGNES-SUR-M	43.6626009	7.1494778	RANGE_INTERPOLATED	3 Avenue Auguste Renoi
10	31105017300062	311050173 NIAUDOT JEAN PIERRE LOUIS EUGENE		10-12 RUE DE FRANCE, NICE	43.6964717	7.2635995	RANGE_INTERPOLATED	12 Rue de France, 0600
11	31107009800018	311070098 SALAMI VIVIANE DENISE LEO		PLACE DE CAUCADE, NICE	43.6762057	7.2147296	APPROXIMATE	Cimetière de Caucade, :
12	31199431300048	311994313 NOTARIAN MARIE-CHRISTINE		33 AVENUE JEAN DE NOAILLES, CANNES	43.5534079	6.9997527	ROOFTOP	33 Avenue Jean de Noai
13	31214385200050	312143852 SIDDI REINE DARIO		9 AVENUE ST SYLVESTRE, NICE	43.7251291	7.2510001	ROOFTOP	9 Avenue de Saint-Sylv
14	31264290300081	312642903 ZIRAH ROBERT ALBERT		13 R DE LA BOUCHERIE, NICE	43.6978514	7.2775591	RANGE_INTERPOLATED	13 Rue de la Boucherie,
15	31276182800051	312761828 MOREAU BRIGITTE MARIE		28 BOULEVARD RENE CASSIN, NICE	43.6727942	7.2244552	ROOFTOP	28 Boulevard René Cass
16	31430811500026	314308115 SARL MAZAL SOCIETE D EXPLOITATION DE	48 BLD JEAN JAURES, NICE	43.6975192	7.2748458	ROOFTOP	48 Boulevard Jean Jaure	
17	31519067800013	315190678 LIONS MAITRE-TAILLEUR		2 AVENUE MALAUSSENA, NICE	43.7062043	7.2645597	RANGE_INTERPOLATED	2 Avenue Malaussena, C
18	31548243000032	315482430 BARRACO IGNACE		12 BOULEVARD CARNOT, GRASSE	43.6549488	6.920865	ROOFTOP	12 Boulevard Carnot, 06
19	31618699800001	316186998 HEREM		14 AV. LOUIS ROUX , ST LAURENT DU VAR	43.6721946	7.1898597	ROOFTOP	14 Boulevard Louis Rou
20	31868749800010	318687498 PIZZERIA TRAVESTERA		QUAI COURBET, VILLEFRANCHE SUR MER	43.703955	7.312384	GEOMETRIC_CENTER	Quai de l'Amiral Courbe
21	31899332600012	318993326 BRUN ET CIE		126 BLD DE CESSIONE, NICE	43.7207944	7.2516523	ROOFTOP	126 Boulevard de Cesso
22	31910205900056	319102059 FERREIRA DA SILVA MARGARITA		130 GRANDE RUE, GREOLIERES	43.7954661	6.9431754	ROOFTOP	130 Grande rue, 06620 (
23	32084144800048	320841448 MECHICHE ALAIN JACQUES		LIEUDIT LE VILLAGE, ST-ETIENNE-DE-TINEE	44.2915572	6.895233	APPROXIMATE	La Tinée, Mercantour N
24	32148360400013	321483604 GIUSTO		7 PL. DE LA REPUBLIQUE, TENDER	44.0873042	7.5935245	RANGE_INTERPOLATED	7 Place de la République
25	32177387100017	321773871 I CORSICA		1 RUE ANFOSSI . SAINT-I AURFNT-DU-VAR	43.6627561	7.1946558	RANGF INTERPOI ATFD	1 Rue léonard Anfossi.

2.2. Corriger et améliorer la base de données pour le dashboard de notre site

Au sujet du dashboard de notre site www.twiing.com, on a discuté avec notre informaticien et nous avons amélioré la base de données. On écrit les requêtes de SQL pour remplir 9 tableaux dans la Figure 11 Data Model pour le site « Twiing » (Pas encore mis en ligne).

- client_month : le sum_transactions et nb_transactions de client par mois.
Pour le tableau monthly_stats
- client_12_months : le sum_transactions, nb_transactions et date_first_transaction_12_months de clients sur 12 dernières mois par mois.
Pour les tableaux transactions_12_months et monthly_stats_12_months
- daily_stats : le sum_transactions et nb_transactions de commerçants par jour.
Pour la Figure 4 : activité au quotidien
- monthly_stats : le sum_transactions, nb_transactions, count_clients, best_CA, best_CA_date et avg_CA de commerçants par mois.
Pour la Figure 5 : Mes chiffres mensuels et la Figure 4 : Mon activité au quotidien
- monthly_stats_12_months : le sum_transactions, nb_transactions, count_clients, best_CA, best_CA_date, mono_CA, mono_nb_transactions, recurrent_CA et recurrent_nb_transactions de commerçant sur les 12 derniers mois, par mois.
Pour la Figure 8 : Monos/Fréquents, la Figure 10 : Des clients gagné/perdu et la Figure 4 : Mon activité au quotidien
- amount_12_months : l'amount_slice et nb_transactions de commerçants par tranche de montant (chaque 5€, de 0 à 200€).
Pour la Figure 7 : L'offre la plus vendue
- transactions_12_months : le nb_transactions_slice, nb_clients, nb_transactions et sum_transactions de commerçant par fréquence de clients.
Pour la Figure 9 : Clients Fréquents
- clients_12_months : le nb_earned et nb_lost de commerçant.
Pour le tableau transactions_12_months et monthly_stats_12_months
- Heatmap : Chiffre d'affaire par heure dans la semaine.
Pour la Figure 6 : Mes heures creuses

Figure 3 : Vue d'ensemble

Figure 4 : Mon activité au quotidien

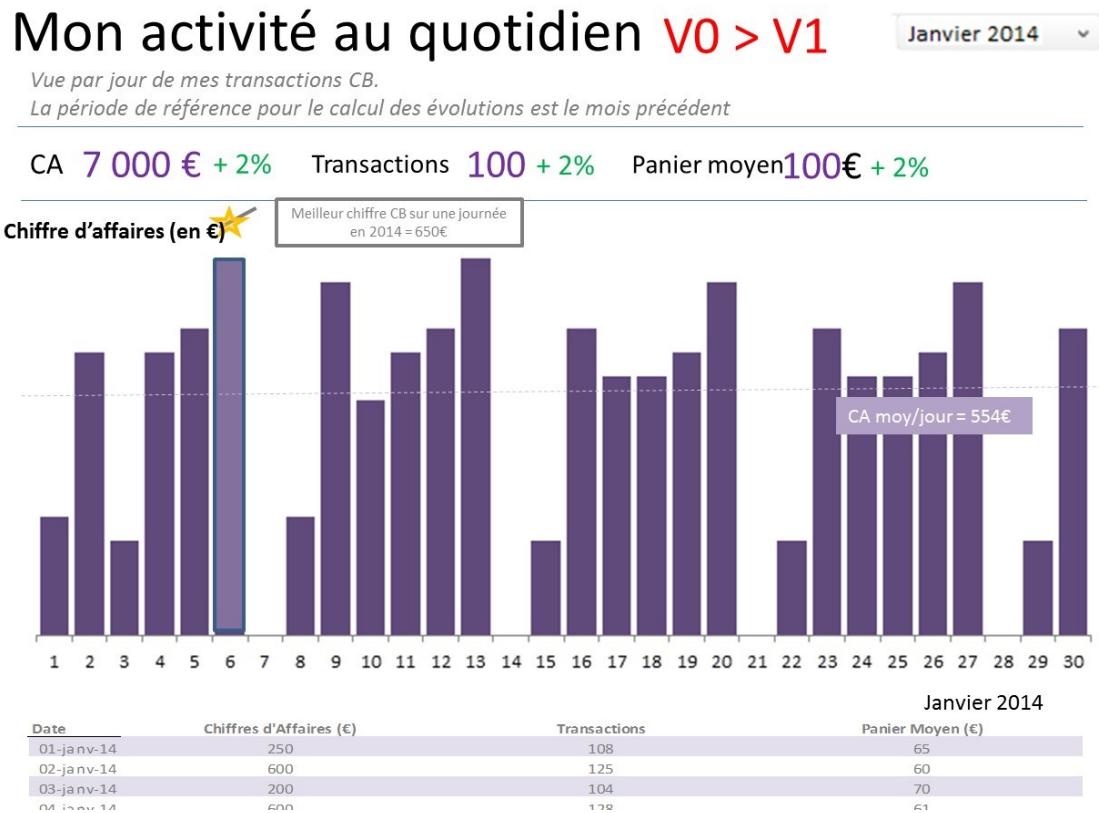


Figure 5 : Mes chiffres mensuels

Mes chiffres mensuels V0 > V1

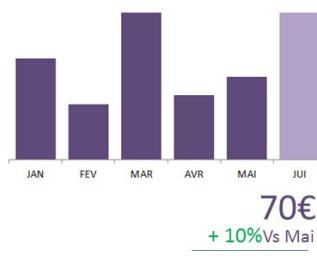
Vue agrégée par mois de mes transactions CB.
La période de référence pour le calcul des évolutions est l'année précédente

Janvier 2014

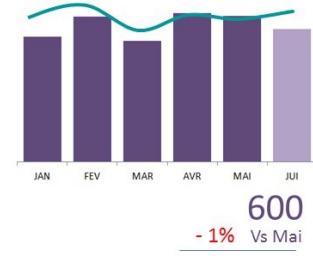
Chiffre d'affaires (en €)



Panier moyen (en €)



Nombre de clients uniques et de transactions



	CA			Panier Moyen			Transactions			Clients uniques		
	Nbre	Evol A-1	Evol M-1	Nbre	Evol A-1	Evol M-1	Nbre	Evol A-1	Evol M-1	Nbre	Evol A-1	Evol M-1
JAN	7 000			65,0			108			108		
FEV	7 500	2%	7%	60,0	2%	-8%	125	2%	16%	125	2%	16%
MAR	7 300	-5%	-3%	70,0	-5%	17%	104	-5%	-17%	104	-5%	-17%

Figure 6 : Mes heures creuses

Quelles sont mes heures creuses? V0 > V1

Vue par jour et par heure de mes transactions CB afin d'identifier les heures creuse et affiner mes horaires d'ouvertures, créer des happy hours....
Plus la cellule est foncée, plus la part du CA moyen par jour d'opération est élevée

Janvier 2014

+ 12 derniers mois
Affiner les « trous »

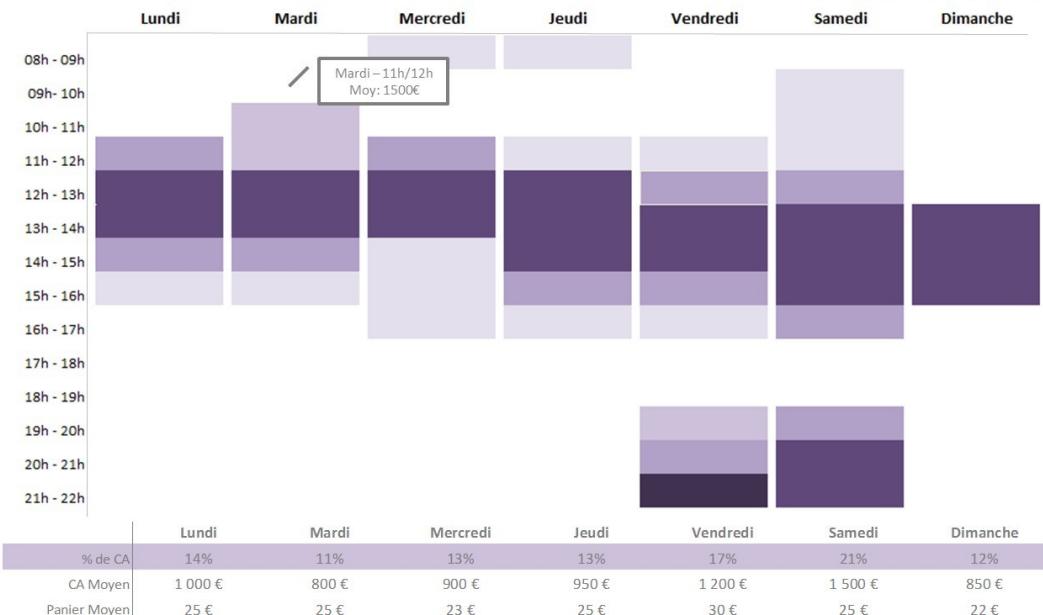


Figure 7 : L'offre la plus vendue

Quelle offre est la plus vendue? V1

Quels sont les prestations/offres qui pèsent le plus dans mon CA? Quel est mon positionnement par rapport au secteur?

Janvier 2014

Répartition de mes transactions par tranche de montant sur les 12 derniers mois

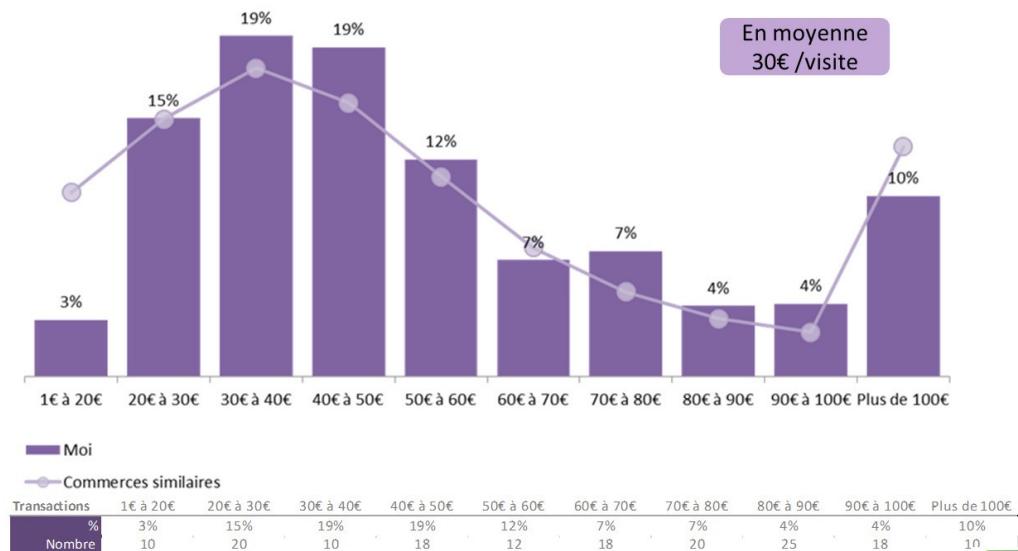


Figure 8 : Monos/Fréquents

Quelle est ma part de « Monos/Fréquents » V1

Est-ce dans la norme du secteur? Qui sont les clients qui pèsent le plus dans mon CA? Combien dépensent-ils en moyenne?

Janvier 2014

% de clients par fréquence de visite sur les 12 derniers mois

En moyenne 1,7 visite/an

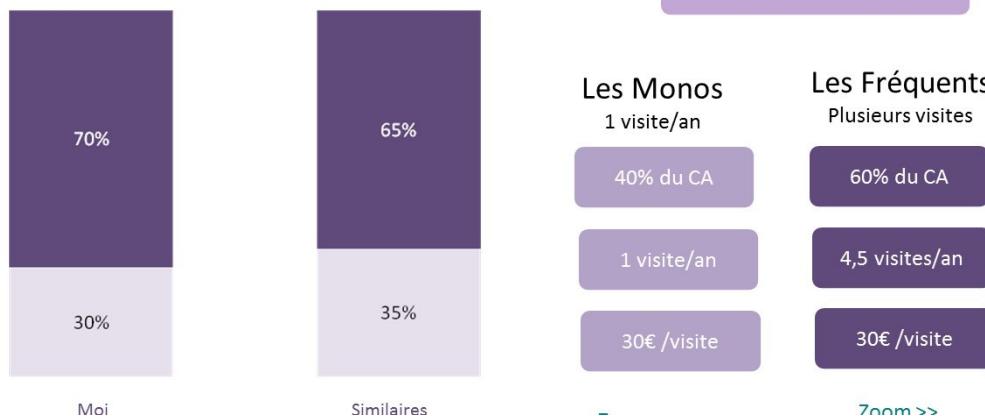


Figure 9 : Clients Fréquents

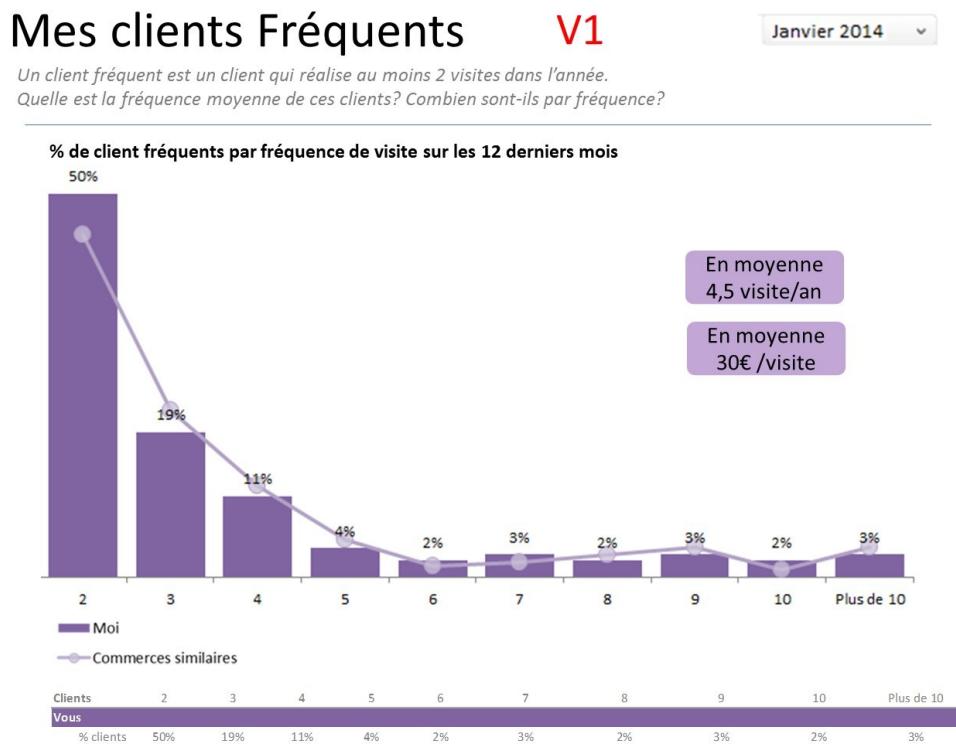


Figure 10 : Des clients gagné/perdu

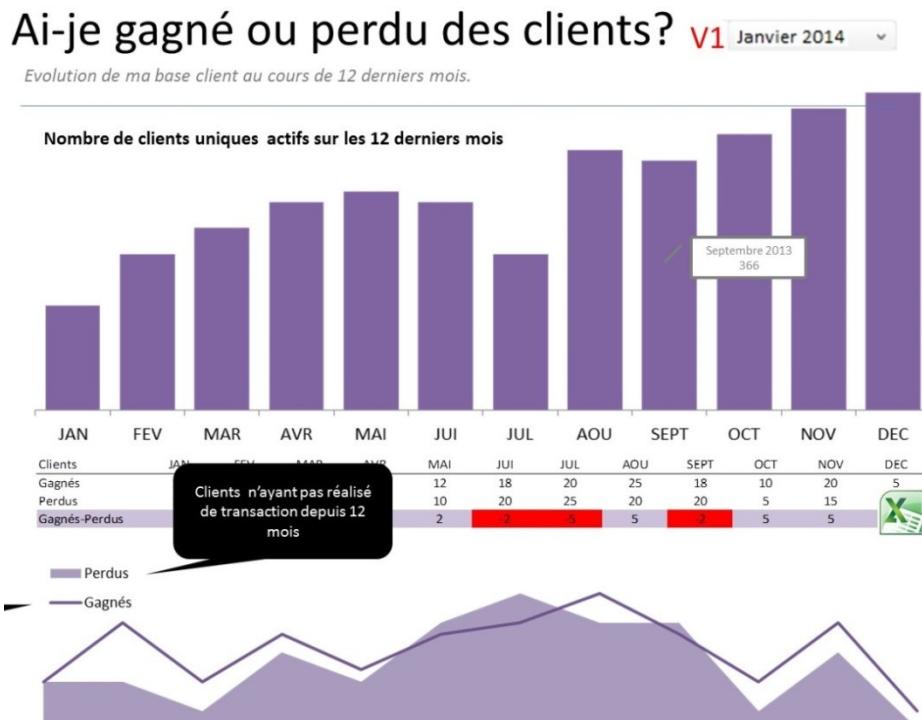
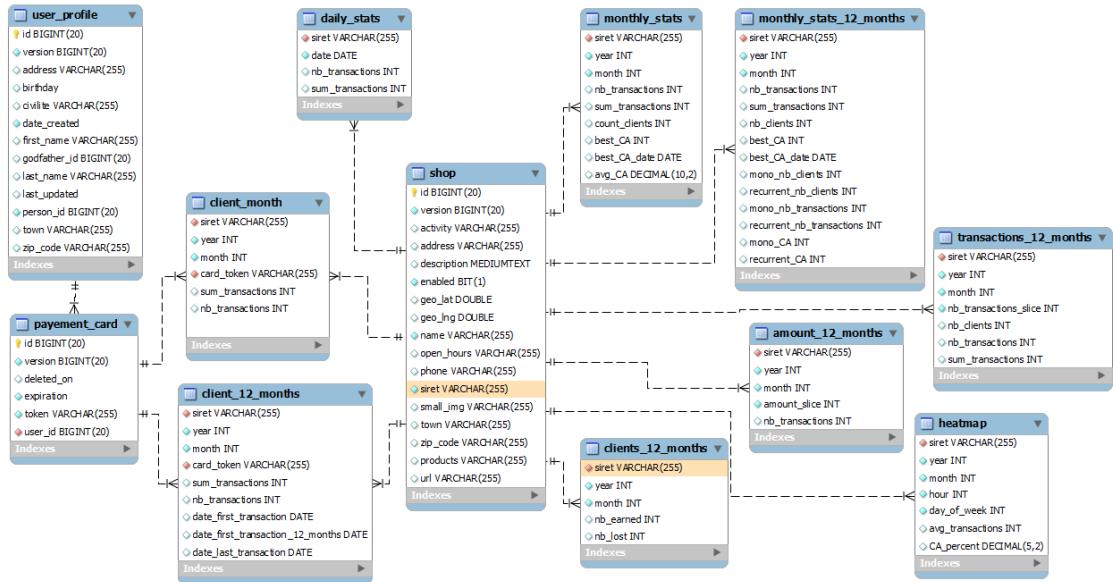


Figure 11 : Data Model pour le site



2.3. Automatiser la génération des « reporting » destinées à l'élaboration des recommandations

Rédaction et automatisation des reporting destinées à l'élaboration des recommandations après la définition du besoin avec le responsable marketing, pour faire la recommandation de notre service aux clients.

On calcule les données et trace les graphiques comme ci-dessus avec EXCEL et Power Point, et notre responsable de marketing fera la présentation de nos services aux nos clients potentiels.

2.4. Partitionnement des commerçants des Alpes Maritime via les méthodes ACP et K-means et mise en place de règles d'affectation via un arbre de décision

Pour le dashboard de notre site « Twiing », on veut comparer un commerce avec des commerces similaires, donc on veut classifier les commerçants du même code NAF. On calcule les indicateurs, le partitionne avec méthode kmeans et méthode ACP+kmeans. Avec les résultats, on a lancé l'algorithme arbre de décision pour mise en place son règles et prédire le cluster des futures commerces sans refaire la partition de tous commerces.

2.4.1. Traitement des données

A partir d'un fiche de 8 032 115 transactions bancaires, on utilise le logiciel R et le package data.table pour calculer les indicateurs qui serviront à faire la méthode de clustering. Comme les données sont assez grandes pour R, on trouve data.table est beaucoup plus rapide que data.frame pour l'agrégation des données et calculer les indicateurs [4]. On fait le traitement des données en 6 étapes :

1. Lire les données avec fonction read.fwf pour une fiche « .dat »
2. Transmettre le « class » de table au data.table
3. Mettre les variables dans les bonnes classes
4. Supprimer les doublons
5. Calculer deux tableaux « data.burt » et « chaque.client » pour calculer des indicateurs de commerces
6. Calculer les indicateurs de commerces pour les classifier

2.4.2. Introduction des indicateurs

Les indicateurs de commences:

id.siret : identité de commerçant

nbr.trans : nbr de transactions de ce commerçant

nbr.client : nbr de clients de ce commerçant

code.NAF : Nomenclature d'activités française

mois.durée : l'écart de mois entre la première transaction et la dernière

nbr.client.multi (récurrents) : nbr de clients qui ont effectué au moins 2 transactions avec ce commerçant

tx.client.multi : nbr.client.multi / nbr.client

freq.trans : nbr.trans / nbr.client
panier.moyen : Montant Total Transaction / Nombre de transaction
délai.moyen(en mois) : délai moyen entre deux transactions de chaque client
recence.moyen : moyen du délai entre la date de la dernière transaction de ce commerçant et la date de la dernière transaction de chaque client chez ce commerçant
sum.trans.houri: i = 1:24
nbr.trans.houri: i = 1:24
p.nbr.trans.houri: i = 1:24
pannie.moyen.houri: i = 1:24
sum.trans.month: month = 20133:20142
nbr.trans.month: month = 20133:20142
p.nbr.trans.month: month = 20133:20142
pannie.moyen.month: month = 20133:20142
sum.trans.weekday: weekday = lundi: dimanche
nbr.trans.weekday: weekday = lundi: dimanche
p.nbr.trans.weekday: weekday = lundi: dimanche
pannie.moyen.weekday: weekday = lundi: dimanche
sum.trans.par.mont.j: j = 0.20: 200
nbr.trans.par.mont.j: j = 0.20: 200
p.nbr.trans.par.mont.j: j = 0.20: 200
pannie.moyen.par.mont.j: j = 0.20: 200

2.4.3. La méthode de partitionnement : k-moyennes

L'algorithme des k-moyennes (ou K-means en anglais) est un algorithme de partitionnement de données relevant des statistiques et de l'apprentissage automatique (plus précisément de l'apprentissage non supervisé). C'est une méthode dont le but est de diviser des observations en K partitions (clusters) dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche. [5] [6]

C'est une technique d'apprentissage non-supervisé.

Algorithme : [6] [7]

Dans la méthode des "k-means", le choix des centres initiaux s'effectue sur la base d'un tirage aléatoire sans remise de k individus à partir de la population à classifier. La partition des classes est modifiée avec chaque affectation d'un individu i de I.

Les individus sont géométriquement représentés dans l'espace vectoriel P muni d'une distance notée d. L'algorithme de la méthode des "k-means" se déroule comme suit :

Etape -0-

1- On choisit par un tirage aléatoire sans remise k individus parmi n individus composant l'ensemble I. Ces k centres notés $\{C_1^0, C_2^0, \dots, C_k^0\}$ sont provisoires.

2- Chaque individu i de I est affecté à une classe et une seule. Chacune de ces classes est localisée par son centre. La procédure d'affectation est la suivante : i est affecté à la classe notée P_1^0 de centre C_1^0 si et seulement si $d(i, C_1^0) = \inf_{j \in \{1, \dots, k\}} \{d(i, C_j^0)\}$

Après avoir affecté tous les individus on obtient k classes notées $\{P_1^0, P_2^0, \dots, P_k^0\}$ de centres respectifs $\{C_1^0, C_2^0, \dots, C_k^0\}$.

Etape -1- En considérant les k classes obtenues à l'étape -0-, on calcule ses centres de gravité. On obtient donc k nouveaux centres notés $\{C_1^1, C_2^1, \dots, C_k^1\}$. On utilise la même règle d'affectation qu'à l'étape -0-, on obtient k nouveaux classes $\{P_1^1, P_2^1, \dots, P_k^1\}$ de centres respectifs $\{C_1^1, C_2^1, \dots, C_k^1\}$.

Etape -h- On détermine k nouveaux classes en calculant les centres de gravité des classes obtenues à l'étape (h-1). La règle d'affectation reste la même qu'à l'étape précédente et on obtient par la suite une nouvelle typologie de l'ensemble I : $\{P_1^h, P_2^h, \dots, P_k^h\}$ de centres respectifs $\{C_1^h, C_2^h, \dots, C_k^h\}$

L'arrêt de l'algorithme de la méthode des "k-means" se fait :

- Lorsque deux itérations successives conduisent à une même partition.
- Lorsqu'on fixe un critère d'arrêt tel que le nombre maximal d'itérations.

On utilise la fonction kmeans dans R, pour réaliser l'algorithme de Hartigan et Wong (1979). [8]

Usage:

```
kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
  "MacQueen"), trace=FALSE)
```

Arguments importants :

x: une matrice numérique

centers: le nombre de cluster

iter.max: le nombre maximum d'itérations à effectuer. Par défaut, kmeans() effectue 10 itérations; cela peut être insuffisant pour que l'algorithme converge ; dans ce cas, le message ci-dessous est affiché :

Warning message:

did not converge in 10 iterations

nstart: on peut exécuter plusieurs segmentations et considérer la meilleure en utilisant l'argument nstart
algorithm: L'algorithme de Hartigan et Wong (1979) est utilisé ici par défaut. [9]

2.4.4. La méthode de l'ACP (Analyse en composantes principales)

L'Analyse en composantes principales (ACP) est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites "corrélées" en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou « axes principaux ». Elles permettent au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Il s'agit d'une approche à la fois géométrique (les variables étant représentées dans un nouvel espace, selon des directions d'inertie maximale) et statistique (la recherche portant sur des axes indépendants expliquant au mieux la variabilité - la variance - des données). Lorsqu'on veut compresser un ensemble de N variables aléatoires, les n premiers axes de l'analyse en composantes principales sont un meilleur choix, du point de vue de l'inertie ou de la variance. [10]

J'ai utilisé la fonction prcomp dans R, pour réaliser l'algorithme ACP. [11]

2.4.5. La méthode de l'arbre de décision

Un arbre de décision est un outil d'aide à la décision qui représente la situation plus ou moins complexe que l'on représente sous la forme graphique d'un arbre de façon à faire apparaître à l'extrémité de chaque branche (ou feuille) les différents résultats possibles en fonction des décisions prises à chaque étape. Sa lisibilité, sa rapidité d'exécution et le peu d'hypothèses nécessaires à priori expliquent sa popularité actuelle. [12]

L'apprentissage par arbre de décision est une technique d'apprentissage supervisée: on utilise une ensemble de données pour lesquelles on connaît la valeur de la variable-cible afin de construire l'arbre, puis on extrapole les résultats à l'ensemble des données de test. [13]

On a utilisé la fonction rpart dans R avec le package rpart, pour réaliser l'algorithme de l'arbre de décision, CART (CLASSIFICATION AND REGRESSION TREES). [14]

Package : rpart

Description: construire l'arbre de décision

Usage: rpart(formula, data, weights, subset, na.action = na.rpart, method, model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...)

Arguments importants:

method: On utilise méthode = "class" qui signifie un arbre de classification

control: contrôle du paramètre de complexité.

Réglage de la complexité de l'arbre : plus l'arbre est complexe (beaucoup de noeuds), plus il va bien apprendre l'échantillon d'apprentissage, mais aussi plus il va faire de l'overfitting : adaptation uniquement à l'échantillon d'apprentissage, mais beaucoup d'erreurs sur un nouvel échantillon de test. La complexité doit donc être pénalisée, d'où un paramètre cp (complexity parameter) : plus cp est petit, plus l'arbre peut être grand (beaucoup de noeuds), plus il est grand, plus la complexité est pénalisée.

minsplit : control=rpart.control(minsplit=10) signifie le nombre minimum d'observations dans un nœud est 10 avant la division. [15] [16]

On va expliquer les résultats d'arbre de décision dans l'annexe avec les résultats de « coiffures ».

2.4.6. La méthodologie et les étapes

Après avoir regardé nos données, on a choisi de faire le clustering sur les données des services de restauration soit 563 services.

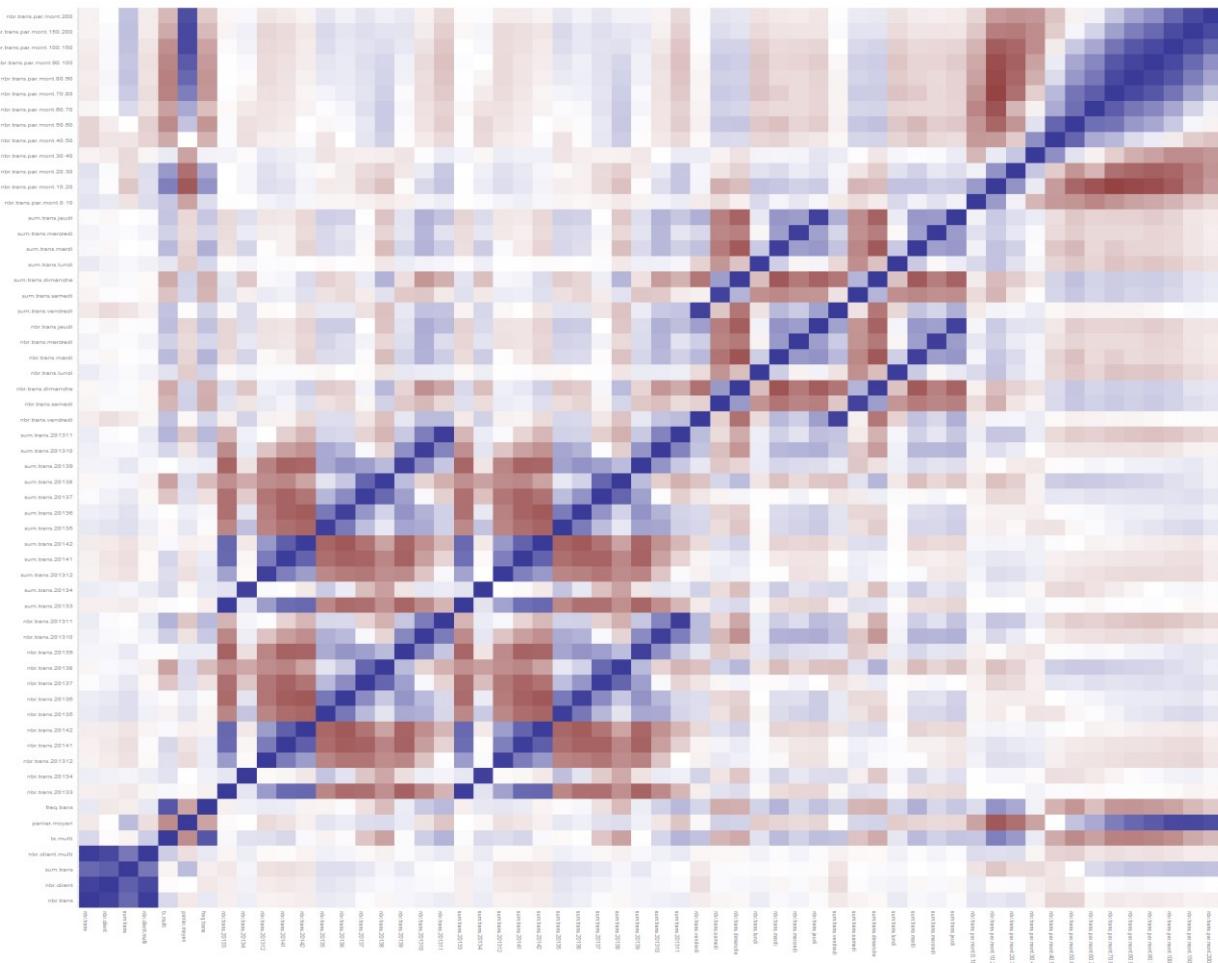
Description d'étape :

Le premier pas, on choisit les données allant du 01-03-2013 jusqu'au 28-02-2014 et on enlève les restaurants qui sont très petits ou trop grands. Il nous reste 363 individuels. Et on choisit les 58 variables suivants à lancer l'algorithme :

```
"nbr.trans"      "nbr.client""sum.trans"    "nbr.client.multi"      "tx.multi""panier.moyen"      "freq.trans"
"nbr.trans.20133"    "nbr.trans.20134"    "nbr.trans.201312"    "nbr.trans.20141"    "nbr.trans.20142"
"nbr.trans.20135"    "nbr.trans.20136"    "nbr.trans.20137"    "nbr.trans.20138"    "nbr.trans.20139"
"nbr.trans.201310"    "nbr.trans.201311"    "sum.trans.20133"    "sum.trans.20134"    "sum.trans.201312"
"sum.trans.20141"    "sum.trans.20142"    "sum.trans.20135"    "sum.trans.20136"    "sum.trans.20137"
"sum.trans.20138"    "sum.trans.20139"    "sum.trans.201310"    "sum.trans.201311"    "nbr.trans.vendredi"
"nbr.trans.samedi"    "nbr.trans.dimanche"    "nbr.trans.lundi"    "nbr.trans.mardi"    "nbr.trans.mercredi"
"nbr.trans.jeudi"    "sum.trans.vendredi"    "sum.trans.samedi"    "sum.trans.dimanche"    "sum.trans.lundi"
"sum.trans.mardi"    "sum.trans.mercredi"
```

Sa corrélation entre les variables est calculée et tracée ci-dessous :

Figure 12: heatmap de corrélation



On voit que la corrélation entre les variables est très logiques, par exemple, le nombre de transactions, le nombre de clients et le chiffre d'affaire sont très corrélés, le nombre de transactions de juillet est corrélé avec celui d'août.

Le deuxième pas, on lance le kmeans sur les données normalisées.

```
tous.vars.scale = scale(tous.vars.analyse)
```

```
tous.scale.km=kmeans(tous.vars.scale, centers=8, nstart = 25)
```

On choisit 8 centres et 25 groupes aléatoires.

Le troisième pas, on trace et sortit les résultats comme ci-dessous :

Figure 13 : Tracer le résultat de kmeans en 7 dimensions

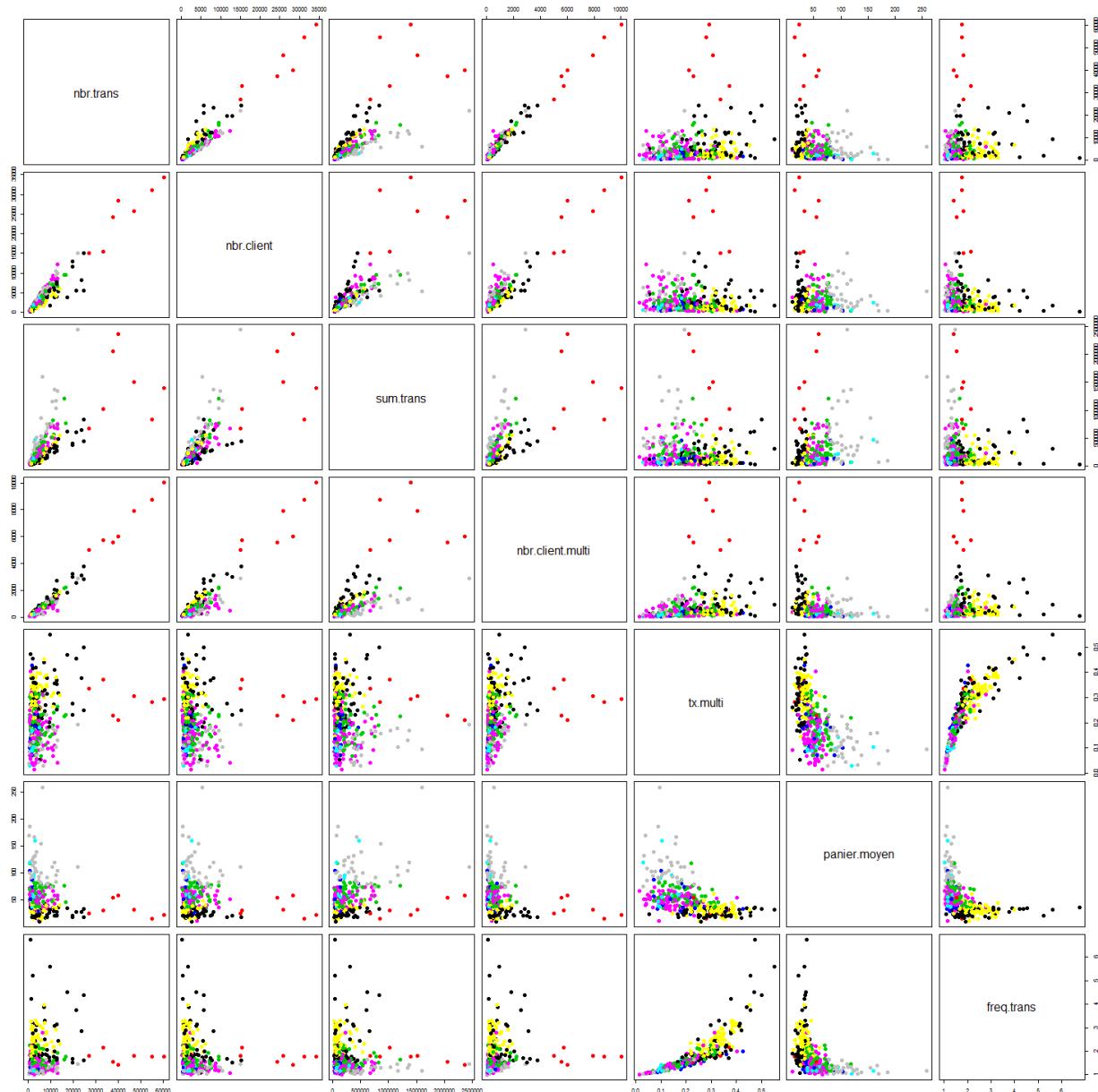
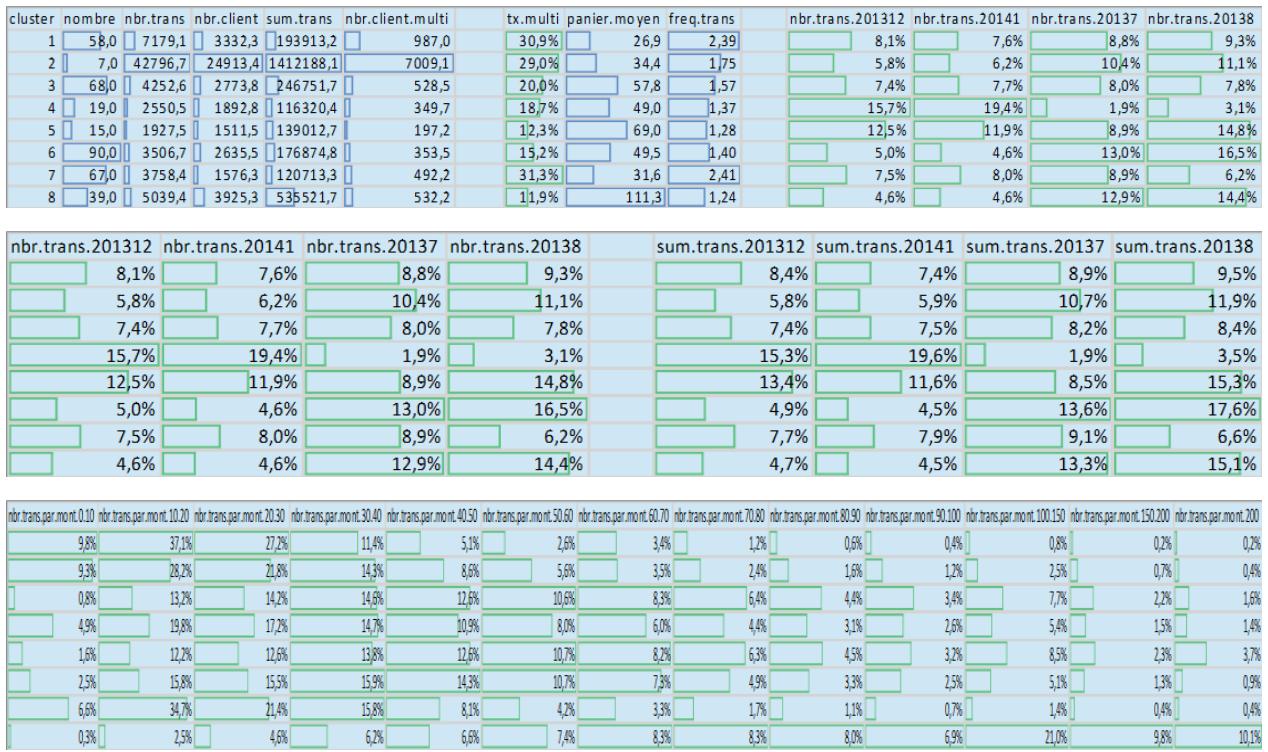
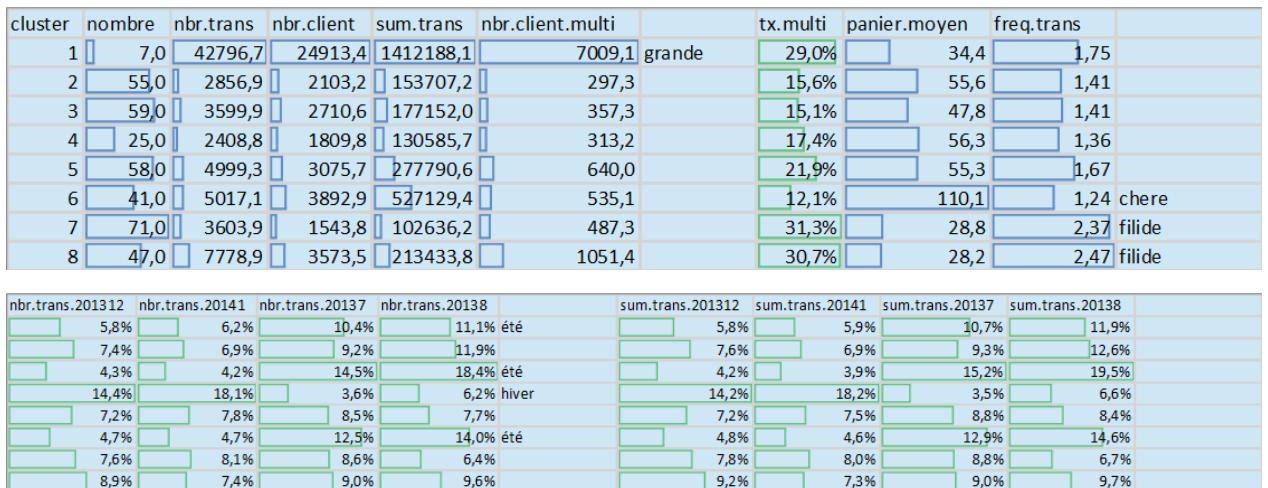


Figure 14 : Les centres de chaque cluster



Le quatrième pas, on fait un ACP avec les 58 variables de 363 observations, et on lance le kmeans avec les premiers 10 composants qui contiennent 82% d'informations, et on obtient :

Figure 15 : Les centres de chaque cluster



	nbr.trans.vendredi	nbr.trans.samedi	nbr.trans.dimanche		sum.trans.vendredi	sum.trans.samedi	sum.trans.dimanche	
	17,8%	17,7%	15,7%		17,7%	18,5%	16,9%	
	17,3%	22,1%	20,5% weekend		17,1%	22,7%	22,0%	
	15,5%	17,7%	17,7%		15,5%	18,1%	18,8%	
	13,8%	21,9%	26,1% weekend		14,7%	22,0%	24,7%	
	22,1%	16,7%	4,2%		22,5%	19,5%	5,0%	
	16,6%	18,5%	17,8%		16,8%	18,8%	18,7%	
	20,0%	10,4%	1,7%		20,3%	12,1%	2,1%	
	16,5%	18,5%	14,3%		16,6%	19,2%	14,8%	

nbr.trans.par.mont.0.10	28,2%	21,8%	14,3%	8,8%	5,6%	3,5%	2,4%	1,6%	1,2%	2,5%	0,7%	0,4%
	1,2%	12,4%	14,8%	15,4%	13,9%	11,6%	8,4%	5,9%	3,9%	2,9%	6,5%	1,7%
	3,1%	17,5%	15,9%	15,7%	13,8%	10,3%	7,1%	4,7%	3,1%	2,3%	4,9%	1,2%
	3,8%	16,9%	15,7%	14,1%	11,5%	9,0%	6,8%	5,2%	3,7%	3,0%	6,3%	1,9%
	0,8%	15,5%	14,8%	15,7%	12,7%	9,8%	7,5%	5,7%	4,0%	3,0%	7,0%	2,0%
	0,4%	2,5%	4,9%	6,3%	6,8%	7,5%	8,5%	8,4%	8,0%	6,9%	20,7%	9,8%
	9,1%	36,7%	22,2%	14,3%	7,1%	3,5%	2,7%	1,4%	0,8%	0,5%	1,0%	0,3%
	7,7%	36,3%	28,3%	11,9%	5,5%	2,7%	3,9%	1,3%	0,7%	0,5%	0,9%	0,2%
												0,3%

Figure 16 : Tracer le résultat de kmeans en 7 dimensions

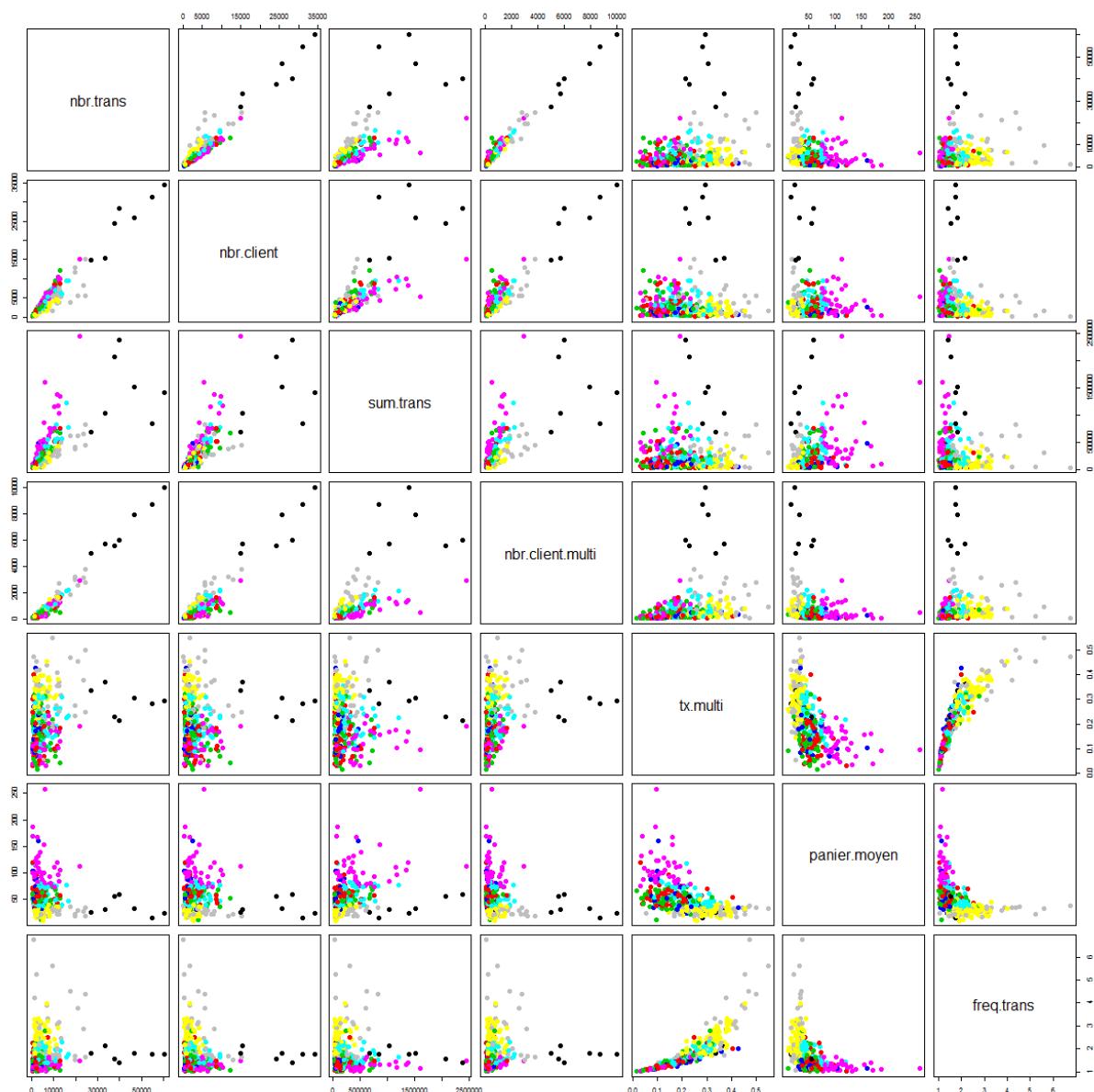
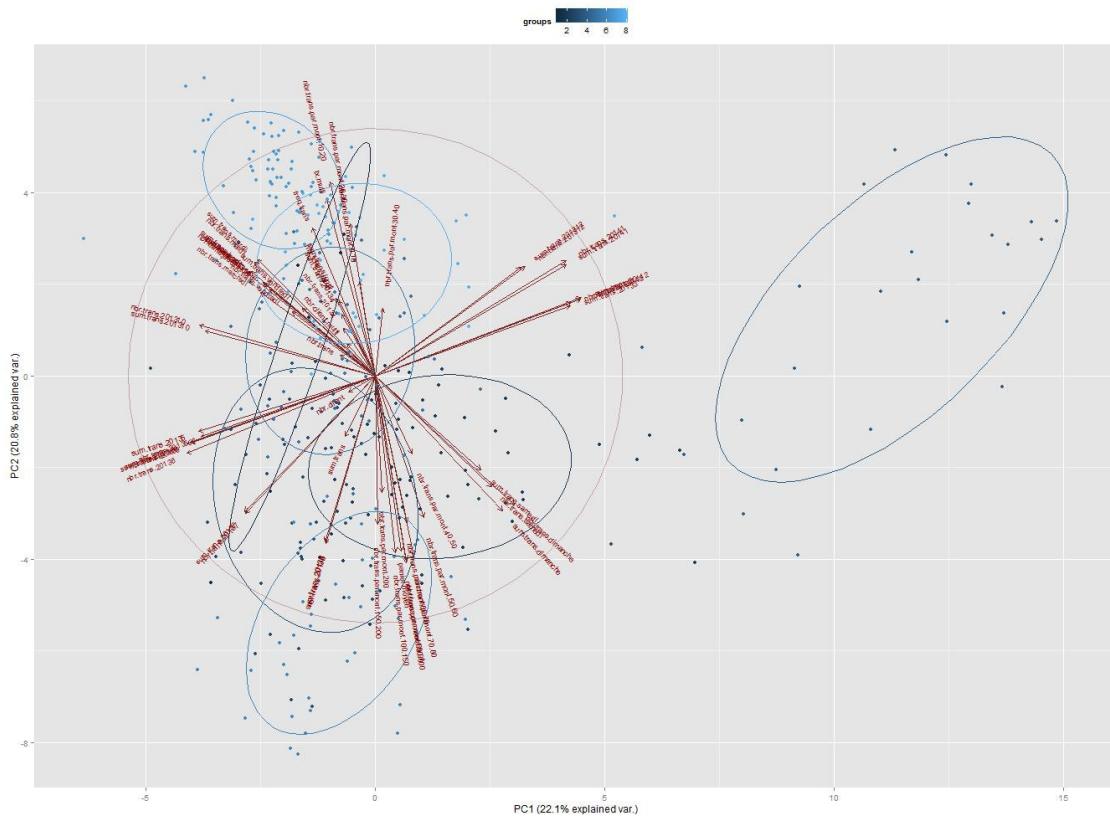


Figure 17 : Tracer le résultat de kmeans en 2 dimensions de composantes principales

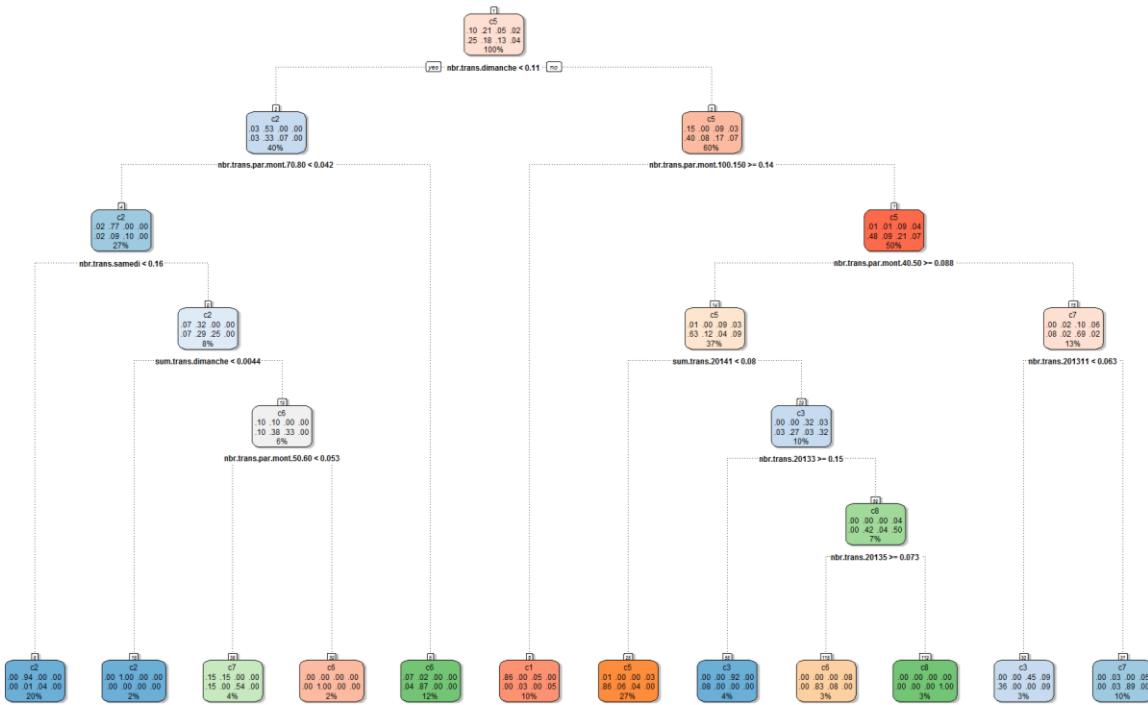


Le cinquième pas, on lance l'arbre de décision. On fait la régression avec 275 individus et on le teste avec les autres pour chaqu'une des deux méthodes.

Avec le résultat d'ACP, on obtient :

	c1	c2	c3	c4	c5	c6	c7	c8
c1	7	0	0	0	0	2	0	0
c2	0	16	0	0	0	2	3	0
c3	0	0	6	0	0	0	0	1
c4	0	0	0	1	0	0	0	0
c5	0	0	0	0	13	5	0	0
c6	0	0	0	0	0	15	1	0
c7	0	1	0	0	1	0	8	0
c8	0	0	1	0	0	1	2	2

Figure 18 : L'arbre de décision



2.4.7. Les résultats

Après la discussion avec mon maître de stage, Monsieur SAUNIER, on a essayé plusieurs combinaisons de variables, et a lancé toutes ces étapes à plusieurs reprises.

Au final, on a choisi les résultats d'analyse avec les 42 variables suivants :

"nbr.trans" "nbr.client" "sum.trans" "nbr.client.multi" "tx.client.multi" "panier.moyen" "freq.trans"
"p.nbr.trans.vendredi" "p.nbr.trans.samedi" "p.nbr.trans.dimanche" "p.nbr.trans.lundi" "p.nbr.trans.mardi"
"p.nbr.trans.mercredi" "p.nbr.trans.jeudi" "p.nbr.trans.par.mont.0.20" "p.nbr.trans.par.mont.20.40"
"p.nbr.trans.par.mont.40.60" "p.nbr.trans.par.mont.60.80" "p.nbr.trans.par.mont.80.100"
"p.nbr.trans.par.mont.100.150" "p.nbr.trans.par.mont.150.200" "p.nbr.trans.par.mont.200"
"p.nbr.trans.0.5" "p.nbr.trans.6.10" "p.nbr.trans.11.14" "p.nbr.trans.15.18" "p.nbr.trans.19.23"
"p.nbr.trans.20141.20133" "p.nbr.trans.20134.20136" "p.nbr.trans.20137.20139"
"p.nbr.trans.201310.201312" "pan.moy.20141.20133" "pan.moy.20134.20136" "pan.moy.20137.20139"
"pan.moy.201310.201312" "pan.moy.vendredi" "pan.moy.samedi" "pan.moy.dimanche" "pan.moy.lundi"
"pan.moy.mardi" "pan.moy.mercredi" "pan.moy.jeudi"

Et on obtient les résultats suivants :

Figure 19 : Tracer le résultat de kmeans en 7 dimensions

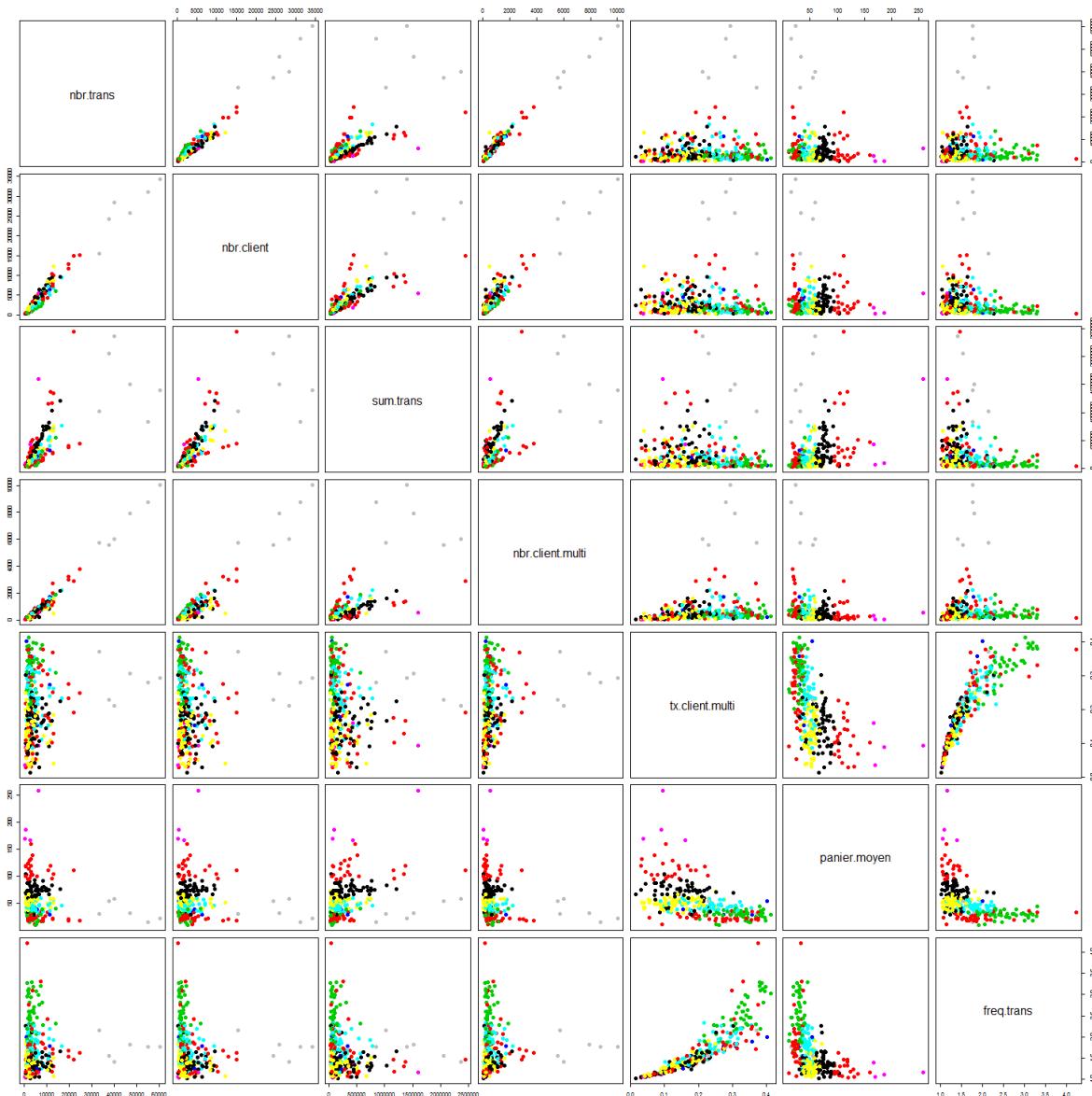


Figure 20 : Les centres de chaque cluster

cluster	nbr.cluster	nbr.trans	nbr.client	sum.trans	nbr.client.multi	tx.client.multi	panier.moyen	freq.trans	
1	46	5029,5	3727,0	381738,3	558,72	15,0%	75,7	1,37	
2	23	4735,9	3666,7	527151,4	492,91	11,6%	113,2	1,24	chere
3	45	3468,7	1420,6	98695,5	451,69	32,2%	28,9	2,50	fidèle
4	6	4662,3	2899,2	154262,7	686,50	28,0%	35,2	1,70	
5	59	4311,3	2555,6	196239,6	563,53	23,0%	44,5	1,74	
6	4	2401,8	1995,3	545229,2	213,50	9,6%	195,4	1,18	chere
7	64	2969,9	2291,2	153515,5	285,55	13,5%	51,1	1,34	
8	6	45422,8	26570,0	1534802,5	7339,33	28,2%	36,0	1,74	grand
9	15	2645,5	2033,8	132678,4	336,20	15,4%	54,3	1,29	
10	36	5889,2	3372,5	135381,7	865,83	26,4%	24,5	1,92	fidèle

p.nbr.trans.vendredi	p.nbr.trans.samedi	p.nbr.trans.dimanche	p.nbr.trans.lundi	p.nbr.trans.mardi	p.nbr.trans.mercredi	p.nbr.trans.jeudi	
17,7%	18,1%	13,8%	9,0%	13,6%	13,5%	14,2%	
16,4%	19,0%	20,3%	8,3%	11,3%	12,0%	12,7%	
20,4%	8,7%	1,0%	15,8%	18,3%	17,2%	18,7%	le semaine, pas chère
16,1%	25,2%	22,6%	8,7%	7,3%	9,3%	10,8%	le weekend, pas chère
20,7%	17,0%	7,4%	8,2%	14,9%	15,1%	16,7%	
20,8%	20,4%	9,4%	8,2%	13,3%	14,4%	13,4%	
16,2%	20,2%	19,4%	10,6%	10,4%	10,5%	12,6%	
16,1%	16,6%	15,5%	12,0%	12,7%	13,6%	13,5%	
12,9%	21,0%	26,6%	11,9%	9,0%	8,5%	10,2%	weekend
16,8%	17,0%	11,1%	12,6%	13,8%	14,1%	14,6%	pas chère

p.nbr.trans.par.mont.0.20	p.nbr.trans.par.mont.20.40	p.nbr.trans.par.mont.40.60	p.nbr.trans.par.mont.60.80	p.nbr.trans.par.mont.80.100	p.nbr.trans.par.mont.100.150	p.nbr.trans.par.mont.150.200	p.nbr.trans.par.mont.200
6,5%	18,8%	21,3%	19,8%	12,6%	13,4%	4,4%	3,1%
3,2%	11,2%	14,1%	15,7%	15,1%	21,6%	8,8%	10,2%
48,3%	33,2%	11,5%	3,7%	1,5%	1,1%	0,3%	0,4%
36,4%	37,0%	14,6%	6,0%	2,5%	2,5%	0,7%	0,4%
21,5%	35,1%	23,0%	10,6%	4,5%	3,7%	0,9%	0,6%
0,2%	1,8%	3,1%	4,4%	8,8%	29,1%	21,9%	30,8%
14,6%	31,5%	27,1%	13,3%	6,1%	5,2%	1,3%	0,9%
35,2%	35,6%	15,4%	6,6%	3,1%	2,8%	0,8%	0,5%
17,8%	30,7%	22,3%	12,8%	6,7%	6,3%	1,8%	1,6%
51,3%	36,7%	7,7%	2,8%	0,7%	0,5%	0,1%	0,1%

p.nbr.trans.0.5	p.nbr.trans.6.10	p.nbr.trans.11.14	p.nbr.trans.15.18	p.nbr.trans.19.23	
2,9%	0,4%	38,5%	7,6%	50,6%	
8,3%	2,9%	31,4%	10,8%	46,6%	
0,7%	2,0%	84,8%	5,0%	7,4%	midi
55,1%	0,1%	0,1%	1,0%	43,6%	le soir, la nuit
1,3%	0,6%	53,7%	5,0%	39,3%	
8,5%	0,0%	18,8%	6,8%	66,0%	le soir
2,2%	1,1%	44,5%	11,8%	40,4%	
1,2%	0,4%	37,7%	14,8%	45,9%	
0,9%	1,2%	76,3%	10,7%	11,0%	midi
1,6%	3,9%	41,3%	19,7%	33,5%	

p.nbr.trans.20141.20133	p.nbr.trans.20134.20136	p.nbr.trans.20137.20139	p.nbr.trans.201310.201312	
21,1%	27,0%	30,9%	21,0%	
20,8%	27,2%	34,5%	17,5%	
24,8%	25,6%	24,0%	25,6%	
29,2%	24,6%	23,8%	22,4%	l'hiver
23,3%	25,7%	27,2%	23,8%	
18,9%	28,4%	37,1%	15,6%	l'été
22,0%	24,0%	35,0%	19,0%	
20,8%	27,7%	30,6%	20,9%	
71,0%	8,0%	5,7%	15,2%	nouveau
22,6%	25,5%	27,3%	24,6%	

Les résultats avec méthode ACP+kmeans :

Figure 21 : Les centres de chaque cluster

cluster	V2	nbr.trans	nbr.client	sum.trans	nbr.client.multi	tx.client.multi	panier.moyen	freq.trans	
1	44	5884,3	3378,5	147316,6		856,23	25,7%	26,3	1,88 fidèle
2	6	45422,8	26570,0	1534802,5		7339,33	28,2%	36,0	1,74 grand,fidèle
3	46	3420,9	1405,2	97611,3		445,74	32,1%	29,1	2,48 fidèle, petit
4	16	2529,3	1944,6	126895,5		320,38	15,3%	54,1	1,29
5	54	4248,8	2518,8	201364,5		553,52	23,1%	46,3	1,75 fidèle
6	64	3003,7	2330,0	156023,1		284,02	13,2%	51,5	1,33
7	43	5158,0	3836,7	393817,5		575,44	14,9%	76,4	1,35
8	4	2401,8	1995,3	545229,2		213,50	9,6%	195,4	1,18 chère
9	6	2345,7	1623,8	111876,6		336,67	23,3%	60,1	1,54
10	21	5057,4	3912,7	564352,4		527,81	11,9%	113,7	1,24 chère

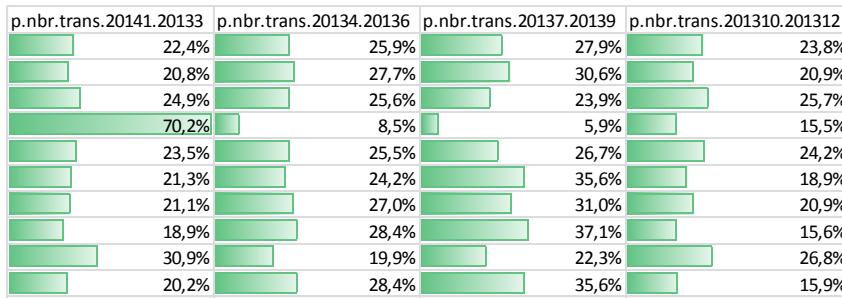
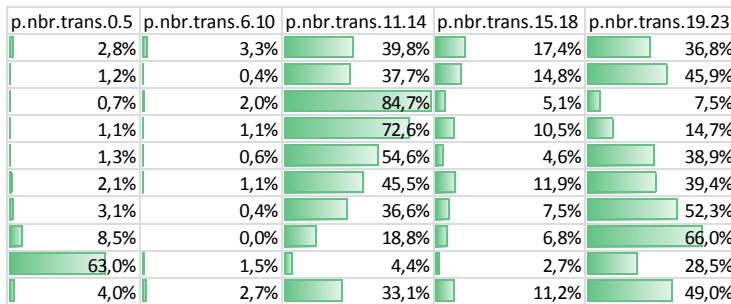
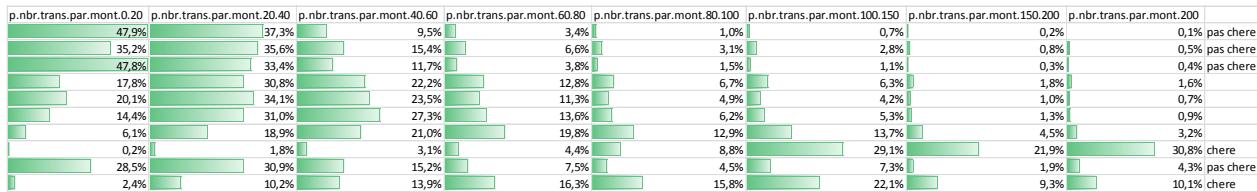
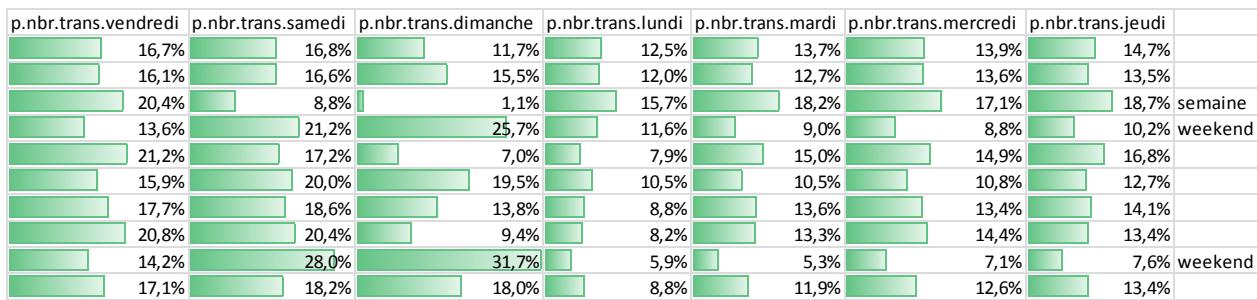


Figure 22 : Tracer le résultat de kmeans en 6 dimensions

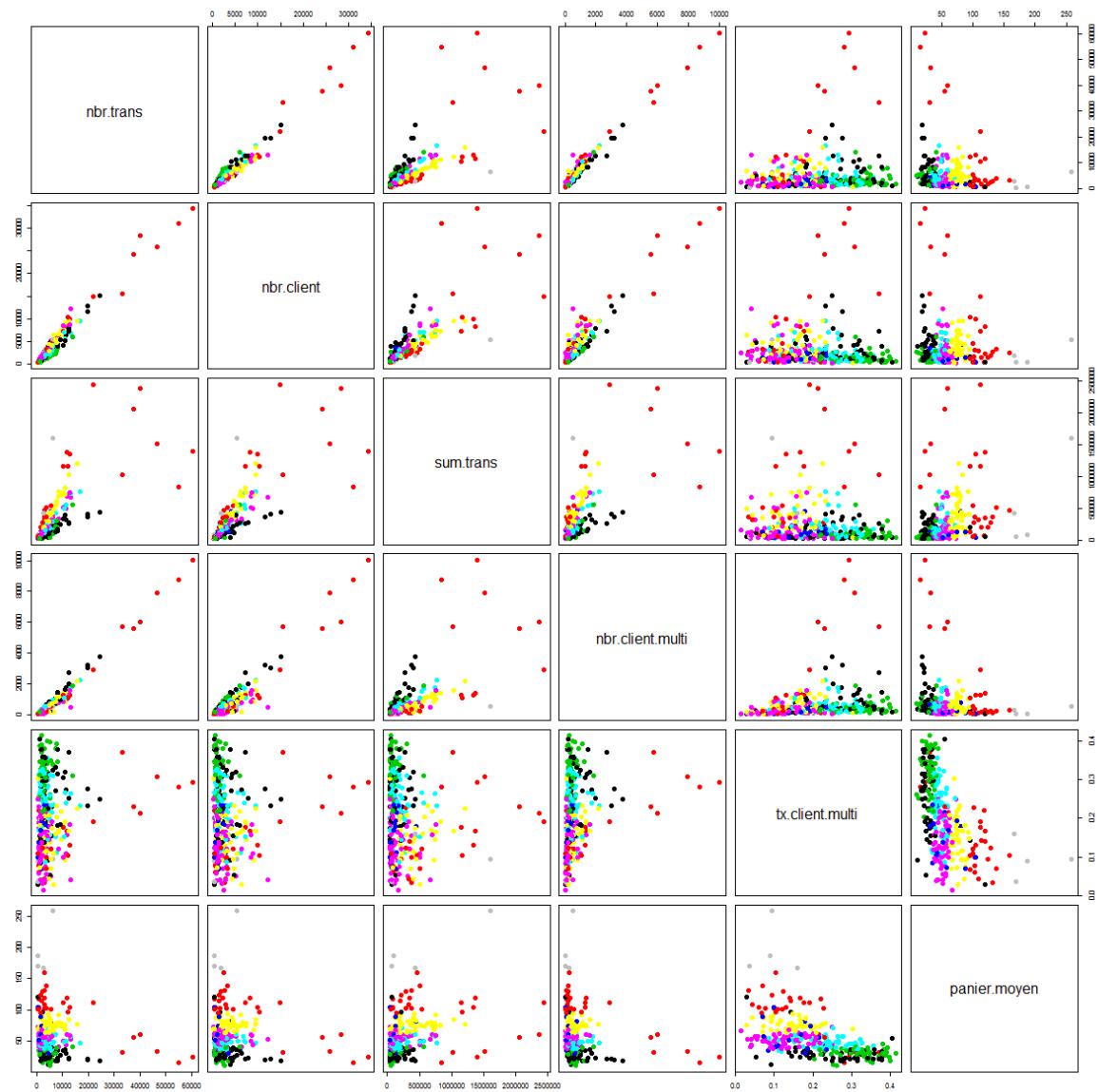
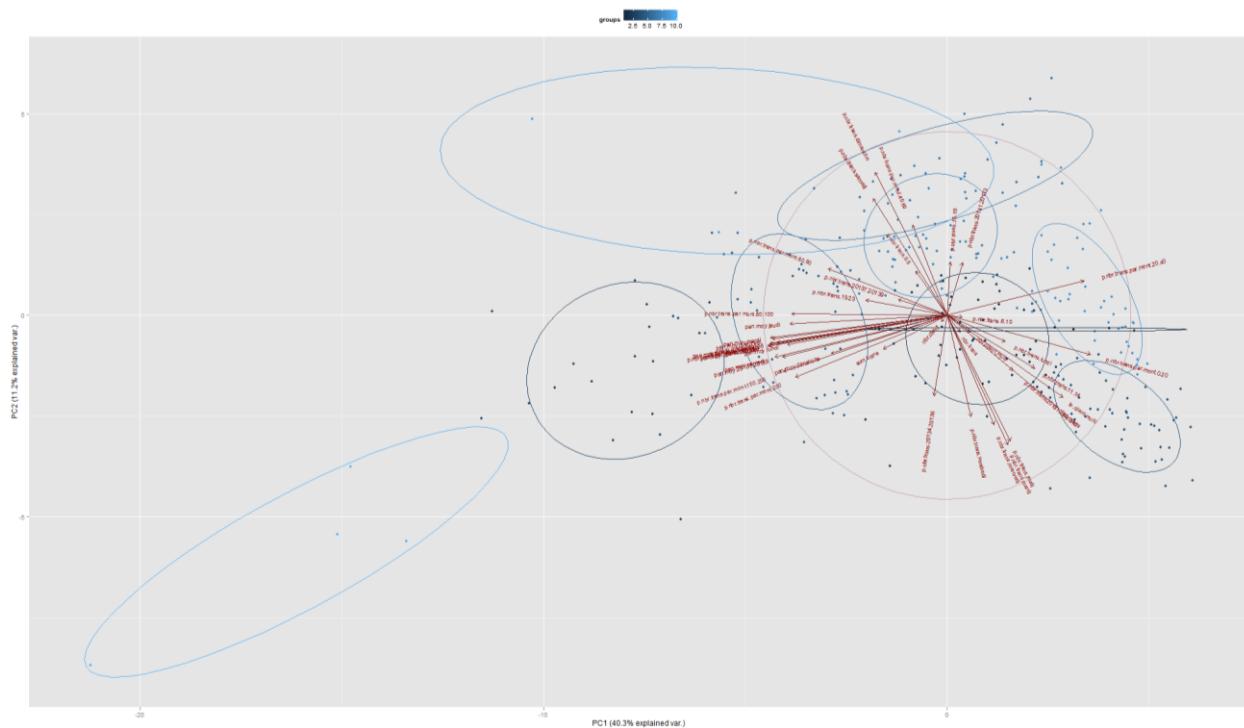


Figure 23 : Tracer le résultat de kmeans en 2 dimensions de composantes principales



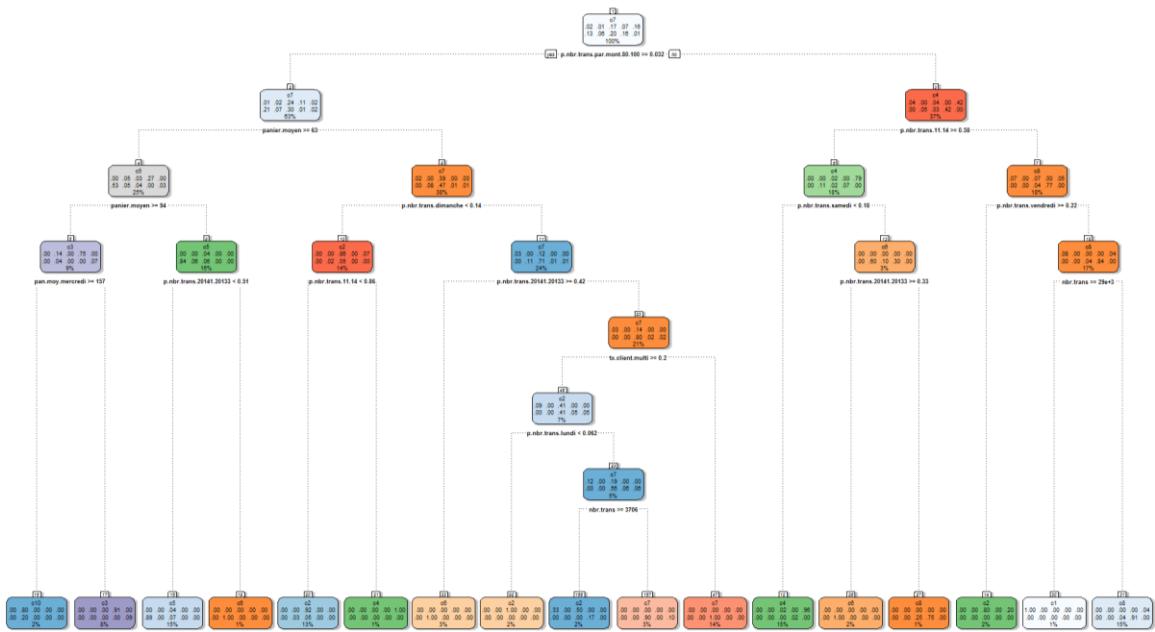
Pour l'arbre de décision, on forme l'arbre par 260 restaurants et le teste par 89, on a le résultat comme ci-dessous :

Quand on prends les paramètres comme : rpart.control(minsplit = 10, cp = 0.01), on obtient un bon résultat pour le prédiction :

	Predict									
	c1	c10	c2	c3	c4	c5	c6	c7	c8	c9
c1	1	0	0	0	0	0	0	0	0	0
c10	0	1	0	0	0	0	0	0	0	0
c2	0	0	11	0	0	0	0	1	0	0
c3	0	1	0	6	0	0	0	0	0	0
c4	0	0	3	0	10	0	0	0	1	0
c5	0	0	0	1	0	11	0	0	0	0
c6	0	0	0	0	0	1	3	0	0	0
c7	0	0	2	0	0	2	0	10	0	0
c8	0	0	0	0	1	0	0	1	10	0
c9	0	0	0	1	0	0	0	0	0	0

Et un arbre compliqué :

Figure 24 : L'arbre de décision



3. Conclusion et discussion

Le travail que j'ai réalisé durant ce stage m'a apporté des connaissances en marketing analytique pour les commerces de proximité, ainsi que la possibilité de mettre en pratique l'utilisation d'ACP, et de découvrir des méthodes de data mining comme kmeans et l'arbre de décision. J'ai travaillé sur la base de données et la création des tableaux et les graphiques à partir des données de transactions bancaires. Et j'ai aussi fait les programmes avec R et SQL et j'ai obtenu les nouvelles connaissances de HTML, XML, JSON et plusieurs packages de R.

Après l'observation de chaque résultat, on trouve qu'il n'y a pas beaucoup de différences entre les résultats de kmeans et ACP+kmeans. Dans l'article [17], on voit que « K-means Clustering via Principal Component Analysis » est une bonne méthode qui permet de diminuer le temps de calcul. Donc pour un grand nombre de données, comme le partitionnement de clients, il a l'avantage de trouver un bon résultat en moins de temps.

4. Questions

1, Est-ce qu'on peut choisir la racine de l'arbre pour faire un arbre de décision ?

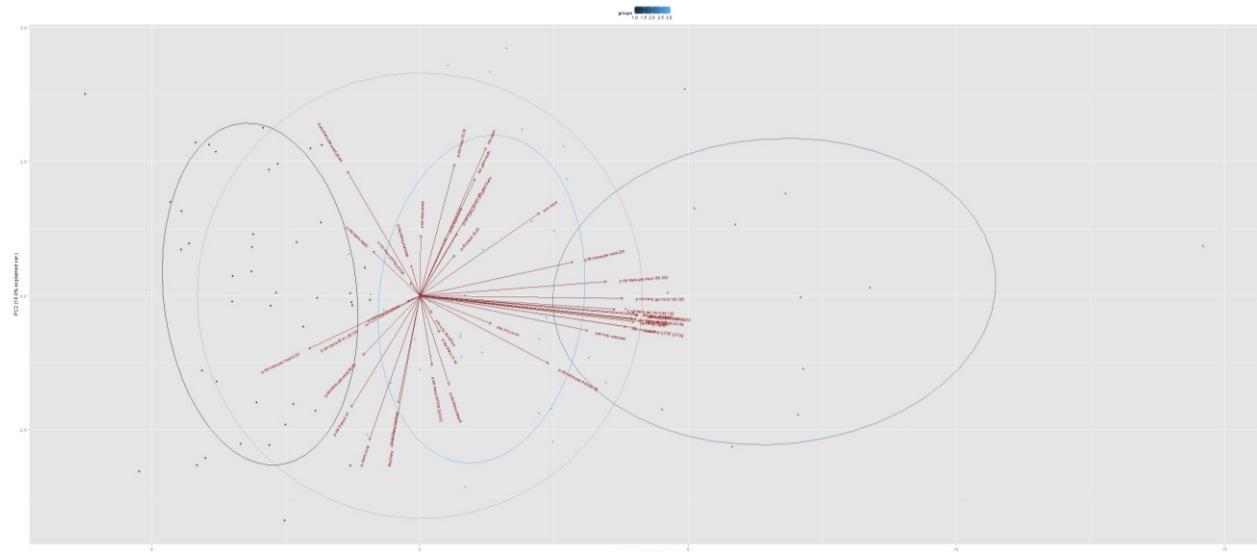
2, On a choisi la méthode kmeans, parce qu'elle converge bien pour des données quantitatives. Mais pour le résultat de cluster qu'on obtient, on le choisit par expérience et on le vérifie un par un manuellement.

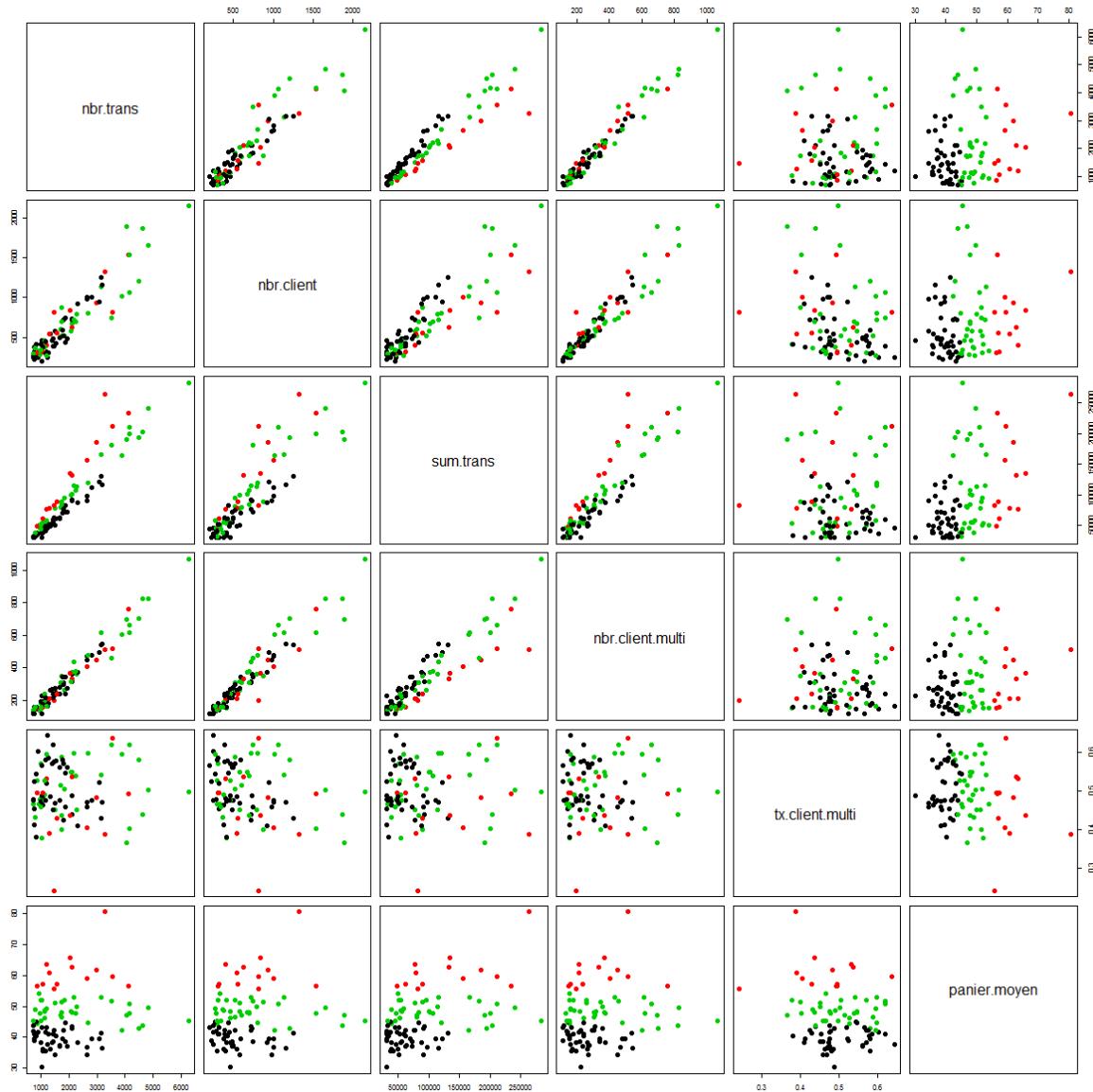
Y-a-t-il une méthode ou l'indicateur pour distinguer le bon cluster du mauvais ?

Annexe

Les résultats des coiffeurs :

cluster	V2	nbr.trans	nbr.client	sum.trans	nbr.client.multi	tx.client.multi	panier.moyen	freq.trans
1	42	1569,0	530,9	61696,8	260,60	50,5%	39,5	3,10
2	13	2168,2	768,5	134838,2	346,38	45,8%	61,3	2,88
3	34	2317,2	780,3	110946,7	388,76	50,5%	48,4	3,02
p.nbr.trans.6.10	16,9%	38,5%	43,2%	1,4%	23,3%	25,2%	24,5%	25,2%
	12,3%	39,7%	45,4%	2,6%	22,2%	25,5%	25,8%	25,0%
	15,4%	40,0%	42,7%	2,0%	22,8%	25,1%	25,3%	25,4%
p.nbr.trans.jeudi	16,6%	23,7%	26,0%	17,1%	15,1%	1,4%	0,1%	0,1%
	17,3%	23,3%	25,9%	15,8%	15,1%	2,5%	0,0%	0,0%
	15,4%	23,7%	27,1%	15,8%	13,5%	4,4%	0,0%	0,0%
p.nbr.trans.par.mont.0.20	22,7%	37,9%	23,7%	10,4%	3,7%	1,4%	0,1%	0,1%
	7,7%	30,5%	19,3%	18,0%	13,0%	8,7%	1,5%	1,3%
	20,3%	30,8%	19,4%	16,2%	7,3%	5,0%	0,7%	0,3%
pan.moy.jeudi	40,0	39,5	39,9	38,9	39,2	34,0	3,5	3,5
	62,0	61,9	61,3	60,5	60,5	49,2	6,8	6,8
	49,2	48,5	49,1	47,9	44,7	45,3	10,3	10,3





Pour l'arbre de décision, on a utilisé 60 salons de coiffure pour former l'arbre et l'a testé par 29, il en sort un bon résultat comme prédition :

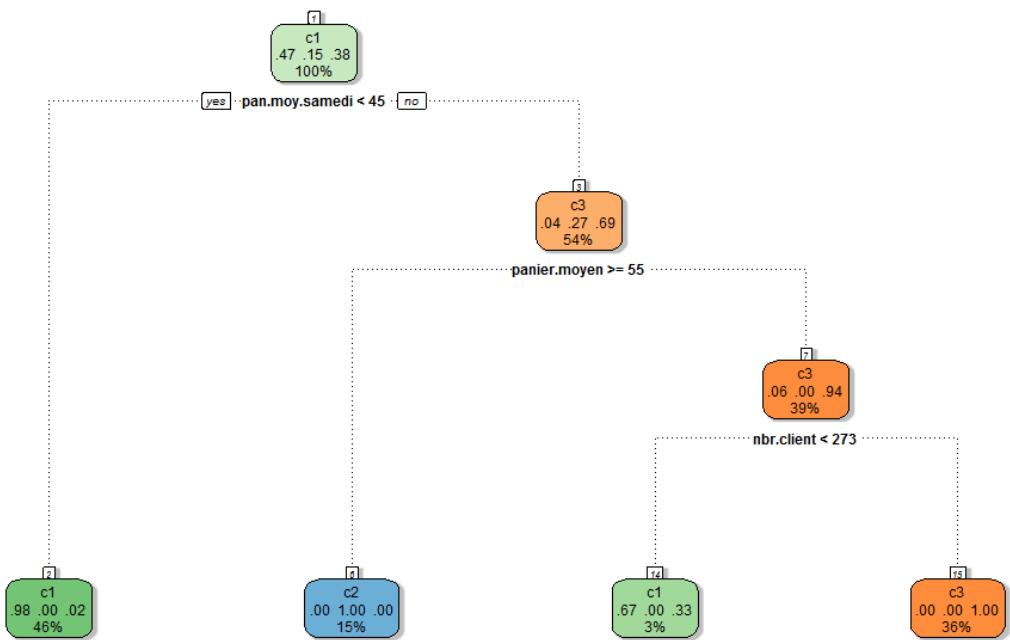
Predict :

c1 c2 c3

c1 13 0 1

c2 0 5 0

c3 1 0 9



Description d'arbre :

$n = 89$

node), split, n, loss, yval, (yprob)

* denotes terminal node

- 1) root 89 47 c1 (0.47191011 0.14606742 0.38202247)
- 2) pan.moy.samedi < 44.65692 41 1 c1 (0.97560976 0.00000000 0.02439024) *
- 3) pan.moy.samedi >= 44.65692 48 15 c3 (0.04166667 0.27083333 0.68750000)
- 6) panier.moyen >= 54.95 13 0 c2 (0.00000000 1.00000000 0.00000000) *
- 7) panier.moyen < 54.95 35 2 c3 (0.05714286 0.00000000 0.94285714)
- 14) nbr.client < 273 3 1 c1 (0.66666667 0.00000000 0.33333333) *
- 15) nbr.client >= 273 32 0 c3 (0.00000000 0.00000000 1.00000000) *

Explication :

A chaque noeud, on a :

Le numéro de noeud : 2 (noeuds de gauche et droite numérotés $2x$ et $2x + 1$ si père numéroté x).

le critère de split (ou root pour la racine) : pan.moy.samedi < 44.65692

le nombre total d'instances pour le noeud : 41

le nombre d'instances mal classées ($0 \Rightarrow$ toutes les instances sont bien prédites) : 1

la valeur prédite (donc majoritaire) de la variable à prédire : c1

entre parenthèses, les proportions d'instances bien et mal prédites : (0.97560976 0.00000000 0.02439024)

une '*' si c'est un noeud terminal. [16]

Bibliographie

Références

- [1] IZICAP, «www.izicap.com,» 2014. [En ligne]. Available: www.izicap.com.
- [2] Infogreffé, «www.infogreffé.com,» [En ligne]. Available: www.infogreffé.com. [Accès le 5 2014].
- [3] GOOGLE, «Google Maps API,» [En ligne]. Available: <https://developers.google.com/maps/?hl=FR>. [Accès le 2014].
- [4] D. Matthew, "Introduction to data.table," May 2013. [Online]. Available: http://datatable.r-forge.r-project.org/RFinance2013_Lightning.pdf. [Accessed 7 2014].
- [5] «Algorithme_des_k-moyennes,» 2014. [En ligne]. Available: http://fr.wikipedia.org/wiki/Algorithme_des_k-moyennes. [Accès le 2014].
- [6] J. A. H. a. M. A. Wong, «A K-Means Clustering Algorithm,» 1979. [En ligne]. Available: http://www.labri.fr/perso/bpinaud/userfiles/downloads/hartigan_1979_kmeans.pdf.
- [7] M. HADD, «Classification de la population en catégories socio-économiques : méthodologie et application pratique,» 1999. [En ligne]. Available: http://www.memoireonline.com/10/08/1603/m_classification-population-categories-socio-economiques-methodologie-application17.html.
- [8] R. Documentation, «K-Means Clustering,» [En ligne]. Available: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html>. [Accès le 8 2014].
- [9] P. Preux, «Segmentation par les k-moyennes,» [En ligne]. Available: <http://www.grappa.univ-lille3.fr/~ppreux/ensg/miashs/fouilleDeDonneesII/tp/k-moyennes/>. [Accès le 2014].
- [10] wikipedia, «Analyse en composantes principales,» [En ligne]. Available: http://fr.wikipedia.org/wiki/Analyse_en_composantes_principales. [Accès le 8 2014].
- [11] R. Documentation, «Principal Components Analysis,» [En ligne]. Available: <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html>. [Accès le 8 2014].
- [12] wikipedia, «Arbre de décision,» [En ligne]. Available: http://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision. [Accès le 8 2014].
- [13] wikipedia, «Arbre de décision (apprentissage),» [En ligne]. Available: [http://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision_\(apprentissage\)](http://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision_(apprentissage)). [Accès le 8 2014].
- [14] B. A. B. R. Terry Therneau, "Package 'rpart', " 2 july 2014. [Online]. Available: <http://cran.r-project.org/web/packages/rpart/rpart.pdf>. [Accessed 8 2014].

- [15] A. Duclert, «Arbres de décision (rpart),» 2013. [En ligne]. Available: <http://www.duclert.org/Aide-memoire-R/Apprentissage/Arbres-de-decision-rpart.php>. [Accès le 2014].
- [16] Quick-R, «Tree-Based Models,» [En ligne]. Available: <http://www.statmethods.net/advstats/cart.html>. [Accès le 8 2014].
- [17] X. H. Chris Ding, «K-means Clustering via Principal Component Analysis,» [En ligne]. Available: <http://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf>.