REPUBLIQUE DU BENIN

Ministère d'Etat chargé de l'Enseignement Supérieur et de la Recherche Scientifique

**********

UNIVERSITE D'ABOMEY-CALAVI

**********

Faculté des Sciences Economiques et de Gestion

**********

Travaux pratiques en Marketing Quantitatif

MASTER PROFESSIONNEL EN ECONOMETRIE ET STATISTIQUE APPLIQUEE.

THEME :

# Modélisation de la valeur client


***Réalisé par :***

1- **GBAGUIDI Gbènankpon Lionel**
2- **HOUNKANRIN-BESSAN Alex P.D**
3- **HOUEHA Théophile**

***Sous la Direction de :***

**M. SONDJO Dieudonnée**


*Année académique : 2020 – 2021*

# Table des matières

## Liste des tableaux :

## Liste des Figures :

## Résumé

Le site d'e-commerce Olist avec un peu plus de 3 000 vendeurs propose plus de 32 000 articles sur tout le territoire brésilien. Il n'a cessé de croître sa clientèle grâce à des politiques telles que le respect des délais de livraison et des campagnes publicitaires bien réussis. Cependant, un défi de taille que rencontre cette entreprise est sa capacité à fidéliser sa clientèle et à les inciter à procéder à de multiples achats. Au cœur de toute entreprise, une bonne relation client est un must qu'il se doit d'atteindre pour prospérer. Il est donc nécessaire à l'entreprise de mettre en place des stratégies marketing efficientes pour sa survie. En effet, un secteur d'activité comme le e-commerce est un secteur dynamique avec une lutte acharnée entre les principaux acteurs du marché.

# Introduction

La production des biens ou services par une firme est sous-jacente à l'écoulement de ces derniers sur les marchés. Ainsi, l'atteinte de cet objectif suscite la mise en œuvre de plusieurs outils, dont le marketing, qui permet de stimuler la décision d'achat d'un client potentiel. Il permet de développer une stratégie d'adaptation des organisations à des marchés concurrentiels, pour influencer en leur faveur des comportements des publics dont elles dépendent, par une offre dont la valeur perçue est durablement supérieure à celles du concurrent. Cet outil de persuasion qu'est le marketing englobe plusieurs variantes dont le marketing quantitatif qui a pour rôle d'optimiser le coût des campagnes publicitaires d'une entreprise ainsi que la détermination de la valeur vie client qui existe entre une entreprise et un client.

L'objectif de notre étude s'inscrit dans le cadre de ce dernier outil permettant de déterminer la valeur vie d'un client qui passe par l'estimation de la durée de vie d'un client dans une entreprise.

Autrement, il consiste à déterminer les modalités qui permettent de fidéliser les clients dans ses transactions avec une entreprise et de sonder si les clients sont plutôt satisfaits des services proposés.

Il s'agira de prendre connaissance de la base de données, de faire l'analyse exploratoire et enfin de procéder à la modélisation de la valeur vie du client.

# I. Connaissance des bases de données

## I.1- La description des bases de données

Nous avons un ensemble de données publiques de commerce électronique brésilien des commandes passées sur Olist Store. Les informations que nous disposons couvrent 2016 à 2018. Il s'agit de données commerciales réelles. Elles ont été anonymisées et les références aux entreprises et partenaires dans le texte de la revue ont été remplacées par les noms des grandes maisons de Game of Thrones. Nous disposons de neuf (09) bases de données présentées dans la figure ci-dessous. Ces bases de données sont reliées entre elles par des variables comme l'indique le schéma ci-après :

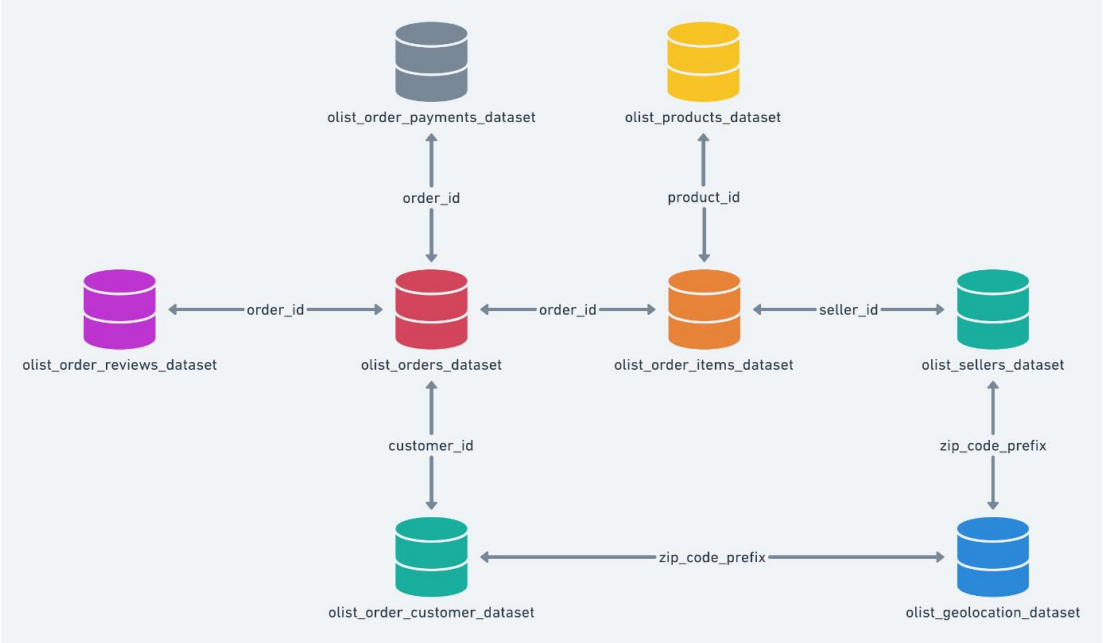



Figure 1 : Schéma de données

Voici, la description de chaque base donnée :

- **olist_order_customers_dataset**

Cette base de données renseigne sur les informations du client, elle contient cinq (05) variables qui sont :

- customer_id : l'identifiant du client généré au moment du lancement d'une commande ;
- customer_unique_id : l'identifiant unique du client,
- customer_zip_code_prefix : le préfixe du code zip du client

- customer_city : la ville du client
- customer_state : l'Etat du client

Elle est reliée à la base olist_orders_dataset par la variable customer_id et olist_geolocation_dataset par la variable zip_code_prefix.

❖ **olist_geolocation_dataset**

Cette base de données renseigne sur les codes préfixes pour chaque ville de chaque état, elle contient cinq (05) variables à savoir :

- geolocation_zip_code_pref : c'est une variable qui présente les différents préfixes des codes zip pour chaque ville de chaque État. Remarquons que les observations de cette variable sont identiques à celles de la variable customer_zip_code_prefix de la base précédente ;
- geolocation_lat : la latitude de la géolocalisation ;
- geolocation_lng : la longitude da la géolocalisation ;
- geolocation_city : la ville de chaque localisation ;
- geolocation_state : l'État de chaque localisation.

Ainsi, cette base est reliée à la base olist_sellers_dataset par la variable zip_code_prefix et olist_customer_dataset par la variable zip_code_prefix comme l'indique le graphique 1 ci-dessus.

❖ **olist_orders_dataset**

Cette base de données donne des renseignements sur les commandes du client, elle contient huit (08) variables à savoir :

- order_id : l'identifiant de la commande ;
- customer_id : l'identifiant du client ;
- order_status : le statut de la livraison de la commande qui prend sept (07) modalités(delivered, canceled, shipped, unavailable, invoiced, created, approved) ;
- order_purchase_timestamp : l'horodatage de la commande, c'est-à-dire, l'heure et la date à laquelle la commande a été lancée ;
- order_approved_at : l'heure à laquelle la commande a été approuvée ;
- order_delivered_carrier_date : la date de livraison de la commande par le transporteur ;
- order_delivered_customer_date : la date de livraison de la commande au client ;

- order_estimated_delivery_date : la date de livraison estimée de la commande.

Ainsi, elle est reliée à quatre bases à savoir : olist_order_payments_dataset par la variable order_id, olist_order_items_dataset par la variable order_id, olist_customer_dataset par la variable customer_id, et enfin olist_order_reviews_dataset par la variable order_id comme l'indique le graphique 1 ci-dessus.

❖ **olist_order_payments_dataset**

Cette base de données renseigne sur le paiement de la commande par le client, elle contient cinq (05) variables à savoir :

- order_id : l'identifiant de la commande ;
- payment_sequential : la séquence de paiement de la commande par le client ;
- payment_type : le type de paiement qui prend cinq (05) modalités (boleto, voucher, credit_card, debit_card, not_defined) ;
- payment_installments : acompte provisionnel, c'est à dire impôt sur la commande.
- payment_value : la valeur du montant de la commande ;

Ainsi, elle est reliée à la base olist_orders_dataset par la variable order_id comme l'indique le graphique 1 ci-dessus.

❖ **olist_products_dataset**

Cette base de données renseigne sur les produits de l'entreprise, elle contient neuf (09) variables à savoir :

- product_id : l'identifiant du produit ;
- product_category_name : le nom de la catégorie du produit ;
- product_name_lenght : la longueur du nom du produit ;
- product_description_lenght : la longueur de la description du produit ;
- product_photos_qty : la qualité de la photo du produit ;
- product_weight_g : le poids du produit (g) ;
- product_length_cm : la longueur du produit (cm) ;
- product_height_cm : la hauteur du produit (cm) ;
- product_width_cm : la profondeur du produit (cm).

Ainsi, elle est reliée à la base olist_order_items_dataset par la variable product_id comme l'indique le graphique 1 ci-dessus.

❖ **olist_sellers_dataset**

Cette base de données renseigne sur les vendeurs des articles, elle contient quatre (04) variables à savoir :

- seller_id : l'identifiant du vendeur ;
- seller_zip_code_prefix : le préfixe du code zip du vendeur ;
- seller_city : la ville du vendeur ;
- seller_state : l'État du vendeur

Ainsi, elle est reliée à la base olist_order_items_dataset par la variable seller_id comme l'indique le graphique 1 ci-dessus.

❖ **olist_order_reviews_dataset**

Cette base de données renseigne sur les commentaires des clients, elle contient cinq (05) variables à savoir :

- review_id : l'identifiant du commentaire ;
- order_id : l'identifiant de la commande ;
- review_comment_title : le titre du commentaire ;
- review_comment_message : le message contenu dans le commentaire ;
- review_creation_date : date de la création du commentaire ;
- review_answer_timestamp : l'horodatage de la réponse suite au commentaire

Ainsi, elle est reliée à la base olist_orders_dataset par la variable order_id comme l'indique le graphique 1 ci-dessus

❖ **product_category_name_translation**

Cette base de données renseigne sur la traduction des noms des produits en anglais, elle contient deux (02) variables à savoir :

- product_category_name : le nom de la catégorie du produit en portugais ;
- product_category_name_english : le nom de la catégorie du produit en anglais.

En effet, elle est reliée à la base olist_products_dataset par la variable product_category_name.

## I.2- Conception d'une base de données et nettoyage de la base de données

Au premier abord, nous avons joint la base olist_orders_dataset à la base olist_order_items_dataset que nous avons traitée avant la jointure. En effet, dans cette base, on remarque que certains numéros de commandes (order_id) sont répétés plusieurs fois ; dû aux faites que dans une commande lancée par un client plusieurs articles (product_id) peuvent y figurer. Ainsi chaque numéro de commande est répété selon le nombre d'articles qui figurent dans la commande. De plus, lorsque la quantité « N », d'article commandé est supérieure à 1, on obtient N répétition du numéro de la commande (order_id) et ceci par article contenu dans cette commande. Ainsi, au niveau de la variable order_item_id (quantité achetée), on obtient une suite de nombre allant de 1 à N pour chaque article. Pour régler ce problème, nous avons pris la valeur maximum de la variable order_items_id (quantité achetée) pour chaque article (product_id) contenu dans une commande (order_id). Remarquons également que certaines commandes (order_id) contenues dans la base olist_orders_dataset ne figurent pas dans la base olist_order_items_dataset. En effet le nombre de commandes (order_id) dans la base olist_orders_dataset est de 99441 tandis qu'il est de 98666 dans la base olist_order_items_dataset après avoir faire fi des répétitions de numéro de commandes contenues dans la base olist_order_items_dataset. Le véritable nombre de commandes est donc de 99441.

Après cette première jointure, nous obtenons une base nommée olist et nous procédons à de nouvelles jointures. La base Olist fut jointe aux bases olist_customer_dataset, olist_sellers_dataset et olist_products_dataset. Nous obtenons une grande base de données appelée « olist » qui est constituée de 29 variables, dont les variables ont déjà été décrites précédemment et de 103.200 observations.

La base olist_payments_dataset est particulière. Nous avons dû procéder à un traitement au préalable. Le paiement d'une commande est réglé parfois par tranche (variable payment_sequential à consulter) par le client et parfois par plusieurs modes de paiement (variable payment_type à consulter). Ainsi, la base nous renseigne sur le montant relatif de chaque mode de paiement d'une même commande. Pour avoir le montant réel de la commande (order_id), nous sommons tous ces montants pour chacune des commandes ainsi nous obtenons une seule valeur (qui sera contenu dans une variable nommée order_value, créer spécialement) pour chaque commande effectuée par le client. Nous obtenons ainsi une nouvelle base nommée Payment1. À partir de la base Olist et de la base Payment1 nous constituerons de petites bases

contenant des variables d'intérêts pour les différentes analyses que nous aurons à faire dans le cadre de cette étude.

## I.3- Création de nouvelles variables pertinentes pour l'analyse exploratoire

Dans la perspective d'analyser les performances de livraison et d'optimiser les délais de livraison, nous procédons à la création de nouvelles variables utiles pour notre analyse. Les variables créées à cette fin sont présentées ci-dessous :

### a) Order_waiting_time_approved

Elle renseigne sur la différence de jour entre la date où la commande a été lancé par le client et celle où le vendeur confirme être en mesure de fournir les articles. Elle est obtenue par la différence entre les variables **order_approved_at** et **order_purchase_timestamp.** Il s'agit du temps d'attente pour que la commande soit effectivement approuvée par le site Olist.

### b) Delivered_time (jours)

Elle renseigne sur le nombre de jours séparant la date estimée par le site Olist pour la livraison (order_estimated_delivery_date) et la date à laquelle la commande a été effectivement livrée (order_delivered_customer_date). Par exemple : une valeur de -15 jours stipule que la commande a été livrée 15 jours avant la date prévue. Tandis une valeur de 10 jours veut dire que la commande a été livrée 10 jours après la date prévue. Elle nous sera très utile dans l'analyse de la satisfaction. Notons que cet indicateur est propre aux commandes ayant la modalité **delivered** (livré) de la variable **order_status.**

### c) Waiting_time_shipping (jours)

Elle renseigne sur le nombre de jours séparant la date où la commande a été remise au transporteur pour livraison (order_delivered_carrier_date) et la date limite d'expédition (shipping_limite_date). Par exemple : une valeur de -5 (jours) stipule que la commande a été remise 5 jours avant la date limite, tandis qu'une valeur de 2 (jours)veut dire que la commande a été remise 2 jours après la date limite.

### d) Product_capacity_cm3 ($cm^3$)

Elle renseigne sur le volume des articles. C'est le produit des variables **product_length_cm, product_height_cm** et **product_width_cm.** Nous avons procédé par la suite à la suppression des variables utilisées dans la création de ces nouvelles.

## I.4- Création d'indicateur

Nous avons créé un indicateur dénommé **respect_delai** qui prend la modalité « **delai respecte** » lorsque la variable **delivered_time** est inférieur ou égale à 0 et la modalité « **delai non respecte** » dans le cas contraire. La modalité « **delai respecte** » signifie donc que la commande a été livrée à temps donc avant la date estimée de livraison et modalité « **delai non respecte** » signifie qu'elle a été livrée plus tard que prévu. Notons que cet indicateur est propre aux commandes ayant la modalité **delivered** (livré) de la variable **order_status.**

# II. Analyse exploratoire

## II.1- Statistique descriptive

Dans cette partie, nous allons procéder à quelques analyses descriptives portant sur les variables de la base. La commande la plus ancienne de la base fut lancée le 04 septembre 2016 (**order_purchase_timestamp**) et la plus récente fut lancée le 17 octobre 2018. Notre base de données s'étale donc sur trois années : 2016, 2017 et 2018. Si nous considérons la base obtenue après jointure telle qu'elle est pour l'analyse exploratoire, nous serons confrontés à une difficulté lors de la segmentation RFM et du calcul de la valeur vie client observée et celle potentielle. En effet ces derniers sont généralement calculés sur une période de temps annuelle, trimestrielle, ou parfois même mensuelle, en fonction du secteur d'activité. Ainsi nous serons confrontés à un problème d'harmonie entre notre analyse exploratoire et la partie relative à la segmentation RFM et la modélisation de la valeur client potentiel. En effet, les conclusions obtenues dans la partie exploratoire nous serviront de socle pour la modélisation de la valeur client potentiel. Puisque la segmentation RFM et le calcul de la Customer lifetime Value nécessite le choix d'une période. Nous décidons de considérer une période d'un an.

Les commandes courent du 04 septembre 2016 au le 17 octobre 2018. Nous sommes amenés à définir deux périodes annuelles qui couvriront le temps entre ces deux dates. La première période couvrira du 04 septembre 2016 au 26 septembre 2017. La deuxième période quant à elle, débutera le 27 septembre 2017 au 17 octobre 2018. Nous remarquons que les deux périodes excèdent 1 an. Mais nous n'avons pas le choix puisque du 04 septembre 2016 au 17 octobre 2018, nous avons plus de 02 ans de différence entre ces deux dates. De plus nous

risquons d'omettre plusieurs commandes si nous sommes rigoureux sur les 1 an. Cependant nous avons délimité ces périodes pour qu'ils soient plus au moins égaux en nombre de jours.

Dans un souci d'harmonie, entre les différentes analyses, l'analyse exploratoire se fera sur chacune de ces périodes. Cela nous permettra de faire des comparaisons entre périodes. Rappelons que bien qu'une commande soit lancée dans la première période elle peut être livrée au client dans la deuxième période. Mais elle sera comptabilisée dans la première période. De plus un même individu peut se retrouver dans les deux périodes, mais sera considéré comme unique à chaque période. Le véritable objectif visé par la création de deux périodes d'études est le calcul du taux d'attrition. En effet son calcul, nécessite de connaitre les individus qui ont rompu contact avec l'enseigne (désabonner). Considérer toute la période allant du 04 septembre 2016 au 17 octobre 2018, sans subdivisions nous empêche d'identifier ces individus.

### II.1.1- Commandes

#### a) Order_waiting_time_approved

Il y a 99 441 commandes qui furent lancées en tout et pour tout du 04 septembre 2016 au 17 octobre 2018, et ceci par 96 096 clients. La répartition entre les périodes est de 27 062 commandes faites par 26 275 clients à la première période et de 72 379 commandes faites par 70 463 clients à la seconde période. Que ce soit à la première période ou à la deuxième période, un peu moins de 75% des commandes ont été approuvés par Olist le jour même où ils ont étés émis par les clients. Le plus grand écart maximal entre le jour ou une commande a été lancé et le jour où elle a été approuvée est de 188 jours (un plus de 9 mois après) pour la première période et de 33 jours pour la deuxième période. Notons qu'il existe pour la première et la deuxième période respectivement 55 et 105 commandes dont nous ne disposons pas d'informations sur le temps d'attente pour qu'elles soient effectivement approuvées par le site Olist. Cependant on remarque que la plupart de ces commandes ont étés annulées.

#### b) Waiting_time_shipping (jours)

On remarque que pour les deux périodes, plus de 75 % des commandes ont étés livrées par le vendeur, au transporteur de la compagnie avant la date limite pour expédition. Cependant, il existe des commandes ayant étés livrées au transporteur, au-delà de la date limite par le vendeur. Le plus grand dépassement pour la première et la deuxième période est respectivement de 58 jours (un peu moins de 2 mois) et 117 jours (plus de 3 mois après). Notons qu'il existe près de 772 et 1012 commandes respectivement pour la première et la deuxième période dont nous ne

disposons pas d'informations relatives à cette variable. On remarque que la plupart de ces commandes ont étés annulées.

### c) Delivered_time (jours) et delivered_time_indicator

Autant pour la première et la deuxième période, 75 % des commandes ont étés livrées au client avant la date prévue de livraison. Parmi ces commandes qui ont étés livrées très tôt avant la date prévue, le nombre de jours maximal entre la date de livraison et la date prévue pour la livraison est de 140 et 147 jours (un peu plus de 4 mois) respectivement pour la première et deuxième période. Parmi les commandes restantes, celles qui furent livrées au-delà de la date prévue pour la livraison sont un peu moins de 25% des commandes pour les deux périodes. Parmi ces commandes livrées tardivement, le nombre de jours maximal entre la date de livraison et la date prévue pour la livraison est de 181 et 188 jours (un peu plus de 5 mois) respectivement pour la première et deuxième période. Notons qu'il existe pour les deux périodes cumulées, 2966 commandes dont nous ne disposons pas d'informations relatives pour cette variable.

### d) Customer_State

Au cours de la première période, 27 062 commandes furent lancées et 72 379 commandes à la deuxième période. Ci-dessous la répartition des commandes par lieu de résidence (Etat) du client à l'origine de la commande au cours de la première période :

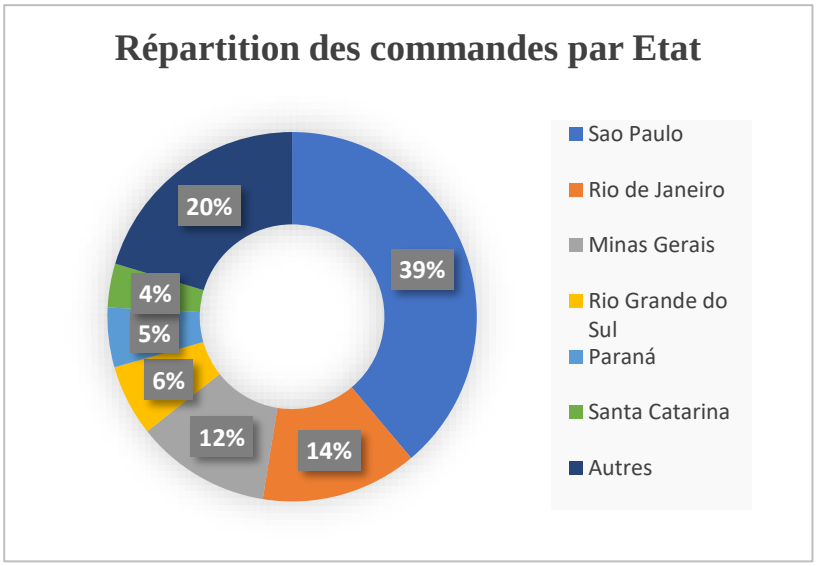



Figure 2 : Répartition des commandes par état pour la première période

À la lecture du graphique, on remarque que l'État de Sao Paulo détient 39% des commandes. Viennent après les états de Rio de Janeiro, de Minas Gerais, de Rio Grande do Sul, de Paraná et de Rappelons que nous avons 27 états au Brésil.

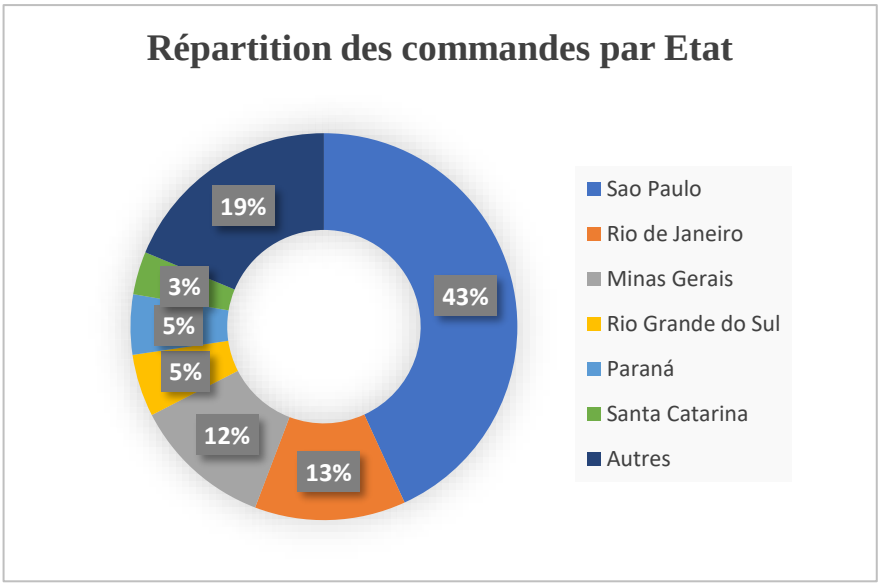



Figure 3 : Répartition des commandes par état pour la deuxième période

À la lecture du graphique, on remarque que la part du nombre de commandes de l'État de Sao Paulo a augmenté de 4 % pour atteindre 43% des commandes à la deuxième période. Celle des autres états a baissé d'un point. Rappelons que le nombre de commandes de cette période est 72 319.

### e) Order_status

Quelle que soit la période, la très grande majorité, des commandes ont étés livrées aux clients. Ils représentent un peu plus de 95 % des commandes. La répartition des commandes selon les différentes modalités de la variable **order_status** est la suivante :

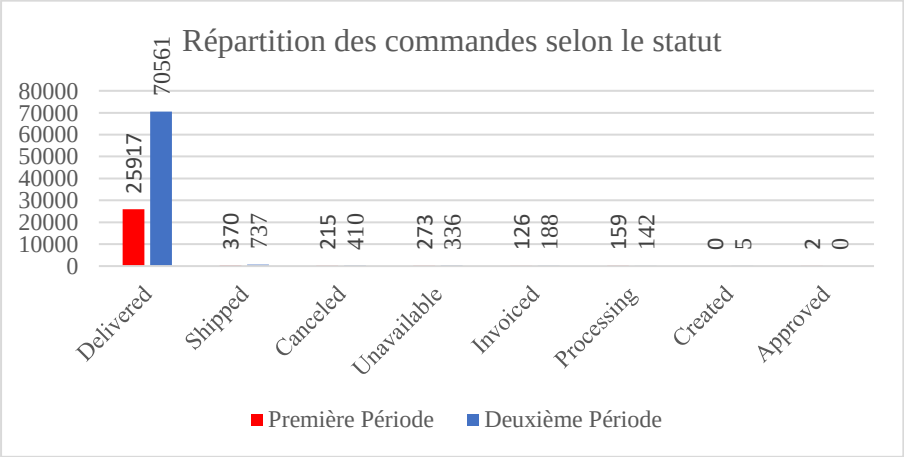



Figure 4 : Répartition des commandes selon le statut

### f) Payment_type

Lors du lancement d'une commande, il est demandé au client par quel mode (type) de paiement compte-t-il réglé sa commande. Il lui ai permis de régler avec un ou plusieurs modes de paiement à la fois. Le graphique ci-dessous présente le nombre de commandes auxquelles les clients ont eu recours à tel mode de paiement.

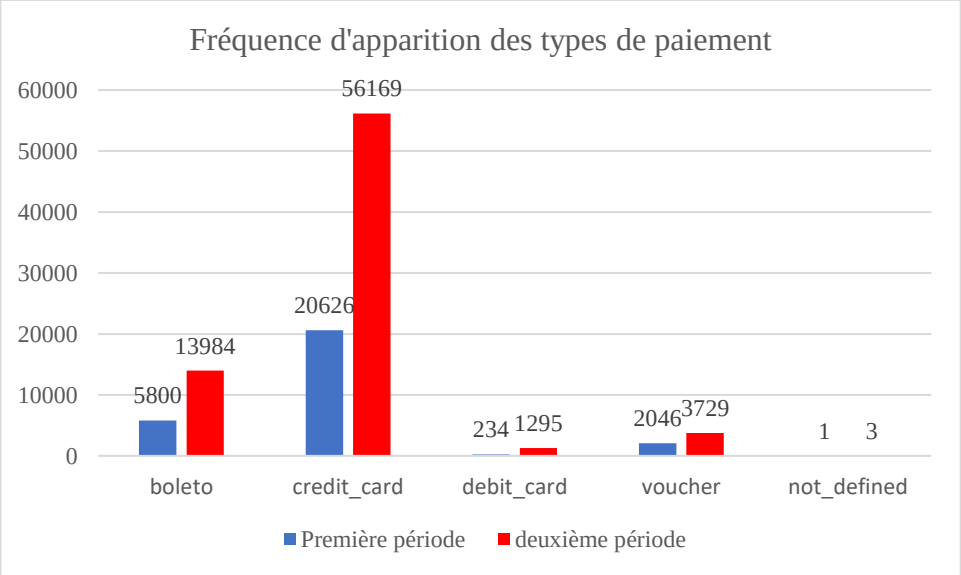



Figure 5 : Fréquence d'apparition des types de paiements

On remarque que le paiement par carte de crédit est le plus prisé parmi les clients sur les deux périodes. Le paiement en espèce est la deuxième forme de paiement auquel à recours la clientèle.

## II.1.2- Articles

Le site Olist propose 32 952 articles, cataloguer dans 74 catégories distinctes. Le nombre maximum d'un même article contenu dans une commande est de 21. Les prix des articles varient entre 0,85 UM et 6 735 UM, tandis que les frais de transport sont compris entre 0 UM et 409,68 UM. Cependant, il existe 775 articles dont nous ne connaissons pas ni le prix ni les frais de transport. Sur la base des informations connues, le prix moyen des articles est de 124,42 UM et les frais de transport moyen sont de 20,11 UM. Le poids maximal des articles est de 40 425 g soit 40,42 kg et le volume maximal est 296 208 $cm^3$. Un peu moins de 75 % des articles possèdent une description plutôt détaillée supérieure à 365 mots. Le maximum étant de 3 992 mots. La plupart des articles ont une image pour que le client ait un aperçu de ce qu'il achète. Cependant on ne dispose pas d'informations relatives sur la description (product_description_lenght) et le nombre de photos (product_photos_qty) pour 2 235 articles.

Sur l'ensemble des commandes, les catégories d'article les plus fréquemment achetées dans la première période sont :

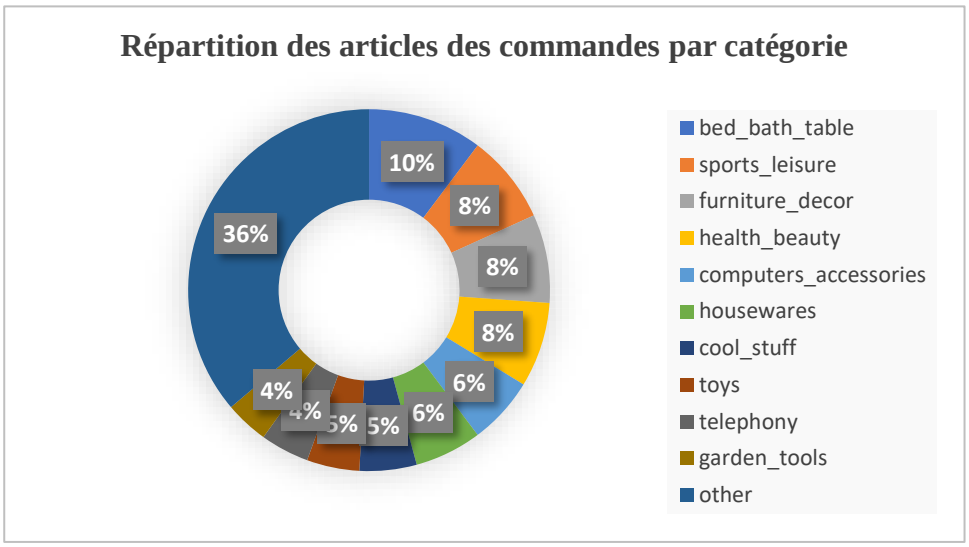



Figure 6 : Répartition des articles des commandes par catégorie au cours de la première période

À la lecture du graphique, on remarque que la catégorie Bed_bath_table (table bain de lit) est la plus fréquente. Viennent ensuite les catégories sports_leisure (loisir sportif), furniture_decor (meubles de décoration), health_beauty (soin de beauté), computers_accessories (accessoires d'ordinateurs). La modalité **other** contient toutes les 62 autres catégories restantes.

La répartition des commandes au cours de la deuxième période est présentée comme suit :

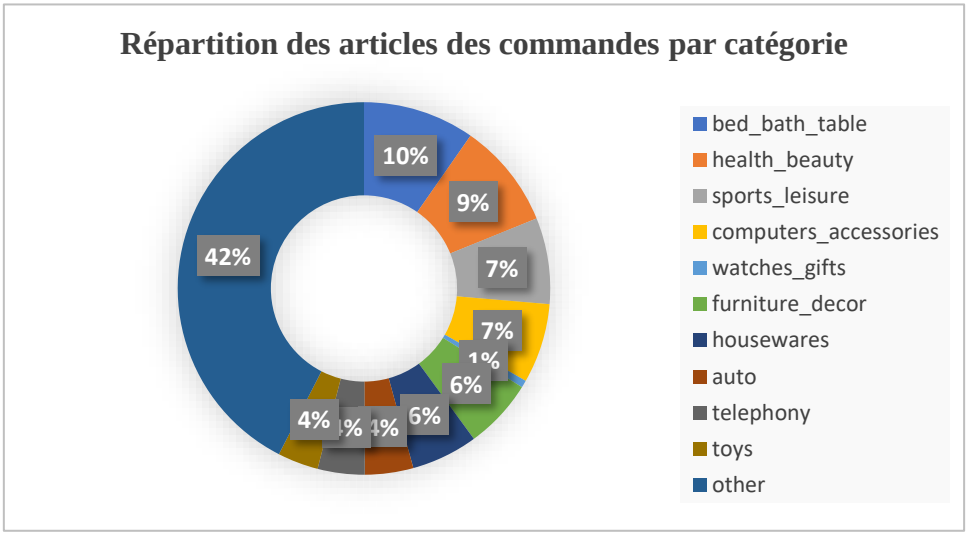



Figure 7 : Répartition des articles des commandes par catégorie au cours de la seconde période

À la lecture du graphique, on remarque que la catégorie Bed_bath_table (table bain de lit) demeure toujours la plus fréquente. Viennent ensuite les catégories health_beauty (soin de

beauté), sports_leisure (loisir sportif), computers_accessories (accessoires d'ordinateurs). La modalité **other** contient toutes les 62 autres catégories restantes.

## II.2- Classification

Dans cadre de cette section nous nous intéressons à trois méthodes de classification. En premier lieu, nous procédons à une analyse en composante principale. Il s'agira d'identifier les liaisons qui peuvent exister entre les variables. Nous nous intéressons au nuage des individus afin de dégager des groupes d'individus possibles. Les deux méthodes restantes de classification qui feront l'objet de cette section sont la méthode des k-means et la classification ascendante hiérarchique. À partir de ces méthodes, nous pourrons dégager des groupes d'individus que nous allons caractériser dans le cadre du profilage des clients.

Mais avant la mise en application de ces trois méthodes, nous avons conçu une table de travail. Elle renseigne sur les clients et contient les variables à savoir :

- nbr_order (le nombre de commandes faites par le client) ;
- montant_depense (montant total dépensé par le client) ;
- customer_state.

Nous ne disposons pas suffisamment d'informations sur les clients. Nous nous contenterons donc que de ces trois variables dont la variable qualitative sera mise en supplémentaire. Nous avons constaté que certaines commandes effectuées par un même individu ont étés livrées dans des états différents. Nous supposons donc que ces individus ont eu à déménager au fil du temps ou commandés pour une personne tierce. Nous décidons donc de considérer un état du Brésil au hasard comme leur lieu de résidence.

### II.2.1- Analyse en composantes principales

#### *A) Première Période*

En premier lieu nous avons procédé à une analyse des données mixtes. Cependant, l'inertie cumulée des dix premiers axes ne dépasse pas les 50 % de l'inertie totale. Nous décidons donc de procéder à l'analyse en composante principale normée, en mettant en supplémentaire la variable customer_state. Deux variables seront donc utilisées pour la formation des axes. Après réalisation de l'ACP. Nous obtenons deux valeurs propres relatives qui sont :

Tableau 1 : Valeurs propres pour l'ACP sur la première période

| | Valeur propre | Inertie | Inertie cumulée |
|---|---|---|---|
| Dim 1 | 1,1031973 | 55,15986 | 55,15986 |
| Dim 2 | 0,8968027 | 44,84014 | 100 |

On remarque que la somme des valeurs propres donne 2 le nombre de variables ayant servis à la formation des axes. Ci-dessous, le nuage des variables. On remarque que les deux variables sont bien représentées sur le plan factoriel. En effet, leur projection est sur le cercle de corrélations. L'écart entre les projections des deux variables montre qu'elles ne sont pas proches, c'est-à-dire corrélées.

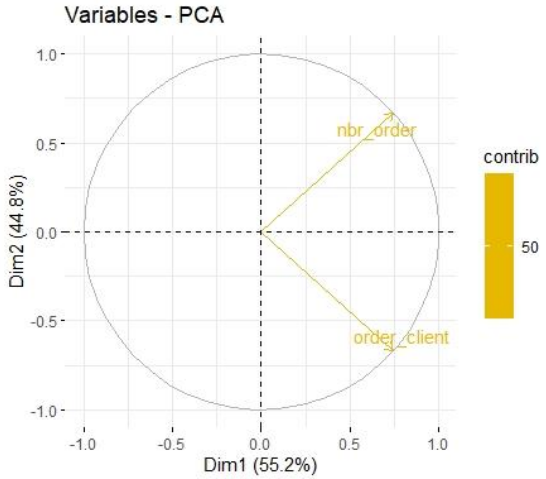



Figure 8 : Nuages des variables de l'ACP pour la première période

Le nuage des individus ci-dessous est d'une forme assez particulière et nous amène à conclure à l'existence de 3 ou 4 groupes naturels dans la population.

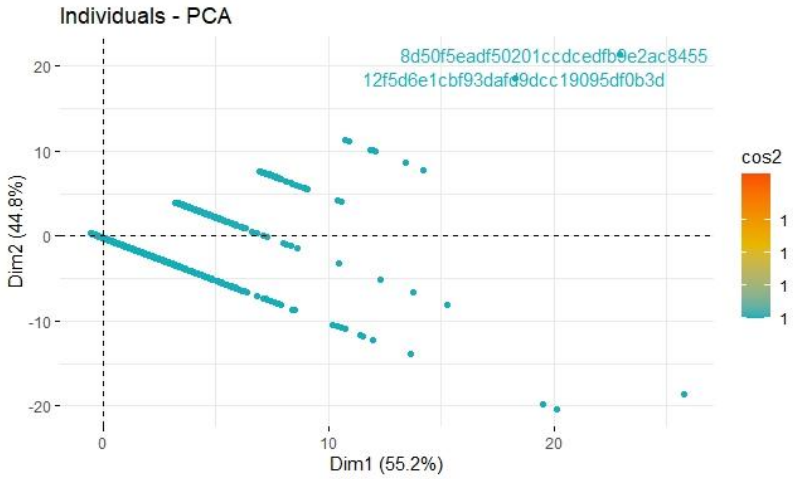



Figure 9 : Nuage des individus de l'ACP pour la première période

Le 4ème groupe composé des individus situés plus haut ne contient qu'une poignée d'individus dont la caractéristique principale est d'avoir acheté plus d'une fois au cours de cette période. L'identifiant de deux individus est présenté sur le graphique. Avec l'option **repel = TRUE** de la fonction **PCA** du package **factoextra**, nous demandons un nuage des individus sans l'affichage des identifiants. Malheureusement, nous recevons un message d'erreur stipulant que sur les 26 275 individus, seuls 26 273 furent traités avec cette option. Dans l'impossibilité de régler ce problème, nous nous sommes contentés de ce nuage des variables. La variable customer_state étant mise en supplémentaire devrait servir pour distinguer les individus dans le nuage des individus. Mais vu le nombre d'états qui est de 27 et la trop grande proximité des individus d'un même groupe, nous ne pourrons pas nous servir de cette option.

### *B) Deuxième période*

L'analyse en composante principale sur les clients de la deuxième période avec la variable qualitative mise en supplémentaire nous donne les résultats ci-après pour les valeurs propres :

Tableau 2 : Valeurs propres pour l'ACP sur la deuxième période

| | Valeur propre | Inertie | Inertie cumulée |
|---|---|---|---|
| Dim 1 | 1,1031973 | 55,15986 | 55,15986 |
| Dim 2 | 0,8968027 | 44,84014 | 100 |

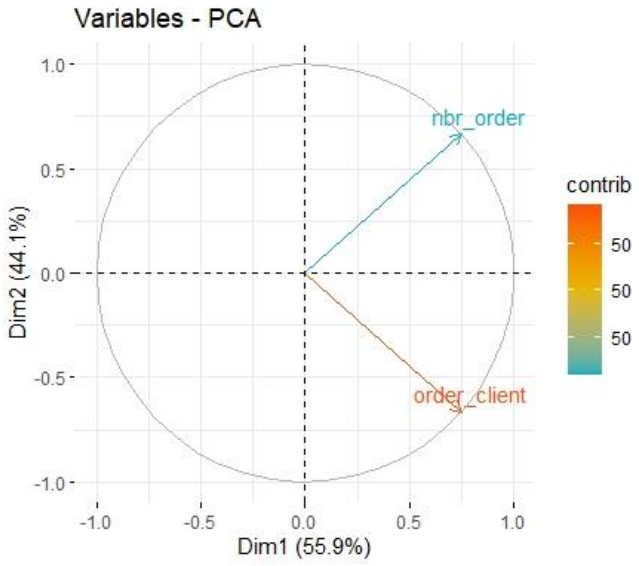



Figure 10 : Nuages des variables de l'ACP pour la seconde période

Ci-dessus, le nuage des variables. On remarque que les deux variables sont bien représentées sur le plan factoriel. En effet, leur projection est sur le cercle de corrélations et de plus nous n'avons que deux axes. L'écart entre les projections des deux variables montre qu'elles ne sont pas proches, c'est-à-dire corrélées.

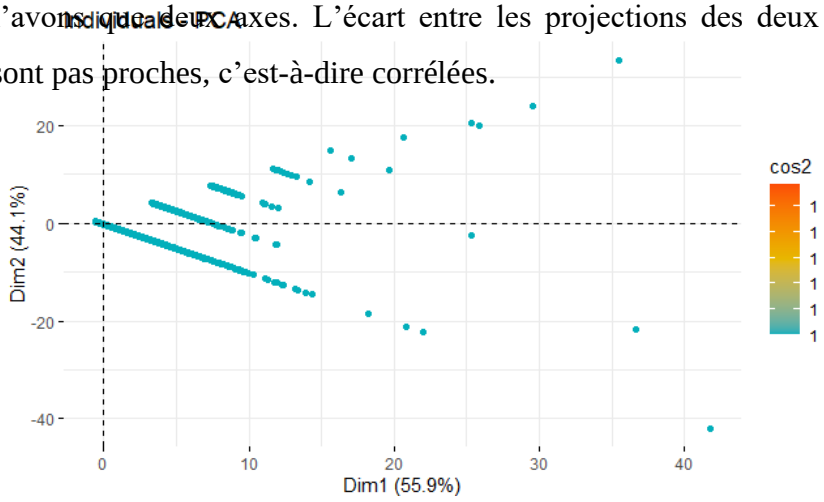



Figure 11 : Nuage des individus de l'ACP pour la deuxième période

Le nuage des individus que nous obtenons à partir de l'ACP nous amène à conclure l'existence de 3 ou 4 groupes comme dans la première période.

### II.2.2- Méthodes de la Classification hiérarchique Ascendante

De l'ACP nous avons retenu entre 3 et 4 groupes pour les individus de la première période. Dans cette section il s'agira de réaliser une classification ascendante hiérarchique afin de former les différents groupes à partir de la coupure du dendrogramme. La CAH sera réalisée uniquement sur les individus de la première période tandis que la méthode des K-Means se fera sur les individus de la deuxième période. Après réalisation de la CAH sur nos données, une analyse de la forme du dendrogramme pourra nous donner une indication sur le nombre de classes à retenir. Ci-dessous le dendrogramme :

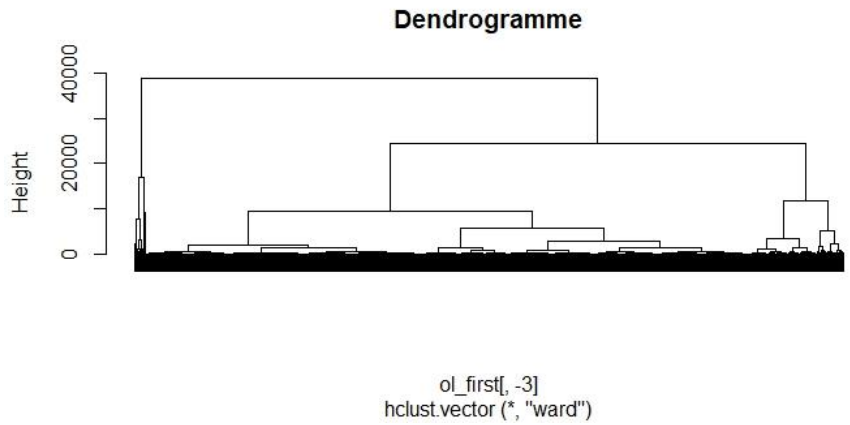



Figure 12 : Dendrogramme initial

À la lecture, on remarque que trois branches bien distinctes apparaissent sur l'arbre. Pour nous aider, nous pouvons représenter les **sauts d'inertie** du dendrogramme selon le nombre de classes retenues. Ci-dessous le graphique :

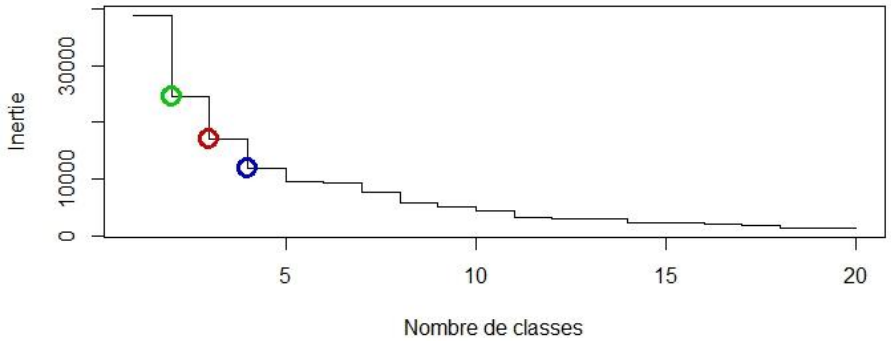



Figure 13 : Perte d'inertie de la CAH

On voit trois sauts assez nets à 2, 3 et 4 classes, que nous avons représentées ci-dessous respectivement en vert, en rouge et en bleu. On remarque qu'un découpage en trois classes minimise ce critère. Cependant, si l'on souhaite réaliser une analyse un peu plus fine, un nombre de classes plus élevé serait pertinent. Nous allons donc retenir un découpage en quatre classes. Nous pouvons donc procéder à la coupure du dendrogramme. Nous obtenons la répartition suivante :

Tableau 3 : Répartition des individus (clients) de la première période par groupe

| Groupe | Effectif | Fréquence (%) |
|---|---|---|
| 1 | 22436 | 85,38 |
| 2 | 3454 | 13,14 |
| 3 | 385 | 1,47 |

À la lecture du tableau, on serait tenté de regrouper les individus du groupe 3 et 4 en un même groupe. Les individus du groupe 4 peuvent être considérés comme la clientèle VIP du site Olist ; celle qui dépense énormément et vers laquelle on pourrait proposer davantage d'offre pour les amener à consommer plus. Une segmentation en 3 classes semble donc être pertinente.

### II.2.3- Méthode des k-Means

Nous avons décidé d'appliquer la méthode des k-means sur la base relative à la seconde période. Ce choix a été déterminé à cause du grand nombre d'individus de la classe qui est de 70 463 et des capacités limitées de nos ordinateurs. Une fois la méthode mise en œuvre, nous nous servons du graphique de la perte d'inertie intra classe pour choisir le nombre de classes approprié. Ci-dessous le graphique :

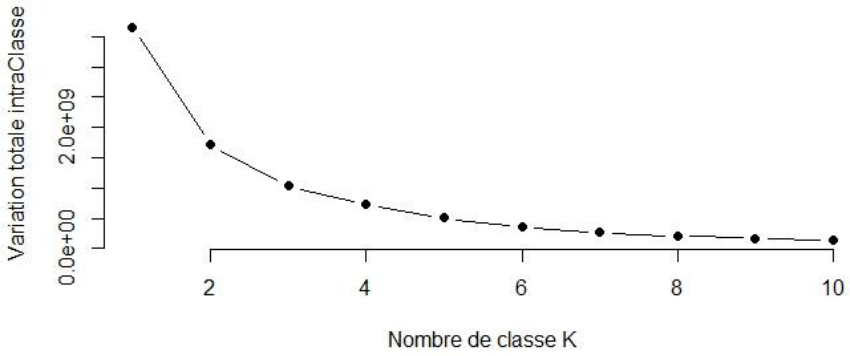



Figure 14: Perte d'inertie intra classe pour les K-Means

À la lecture du graphique, le nombre de classes idéal semble être 3. Cependant pour rester en harmonie avec les résultats de la CAH sur les individus de la première période, nous choisirons 3 groupes. Le nuage des individus avec les clusters représentés est présenté ci-dessous :

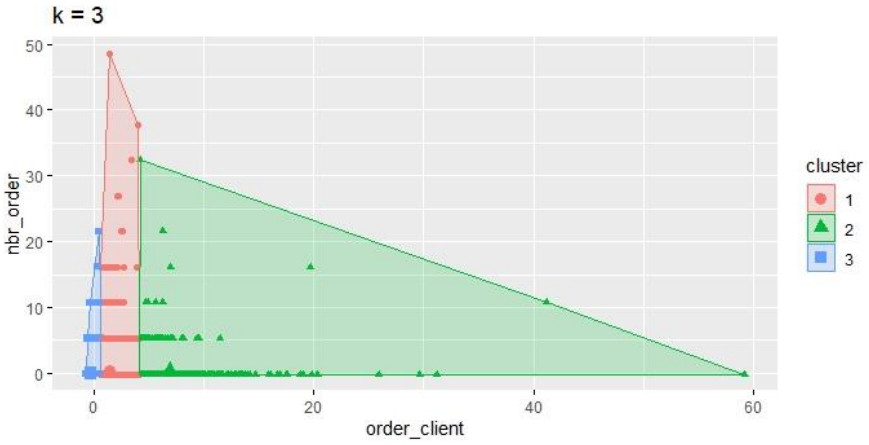



Figure 16 : Nuage des individus par la méthode des K-Means

Après regroupement nous obtenons la répartition suivante :

Tableau 4 : Répartition des individus (clients) de la deuxième période par groupe

| Groupe | Effectif | Fréquence (%) |
|---|---|---|
| 1 | 62855 | 89,20 |
| 2 | 6927 | 9,83 |
| 3 | 681 | 0,97 |

## II.3- Analyse géographique

Cette section se déclinera en trois sous-sections. Il s'agira dans un premier temps d'étudier la répartition géographique des clients sur le territoire brésilien, ensuite celle des vendeurs sur le territoire et enfin le chiffres d'affaires réalisées par l'entreprise dans chaque état du pays. Ces analyses se feront pour chaque période. Rappelons que nous avons constaté que certaines commandes effectuées par un même individu ont étés livrées dans des états différents. Nous supposons donc que ces individus ont eu à déménager au fil du temps ou commandés pour une personne tierce. Notons également que nous nous intéressons à toutes les commandes sauf celles annulées.

### II.3.1- Analyse géographique des clients

La figure ci-dessous nous montre la répartition des clients sur le territoire brésilien à la première et à la deuxième période. À la lecture du graphique, on constate que l'entreprise Olist a une clientèle présente sur tout le territoire brésilien. La plus grande concentration de sa clientèle est dans l'état de Sao Paulo avec plus de 30% et 40% des clients respectivement pour la première et la deuxième période. Les états de Rio de Janeiro, Minas Gerais complètent le podium quel que soit la période. Remarquons que le pourcentage de la clientèle résidant dans l'Etat de Sao Paulo a augmenté de la première à la deuxième période, tandis que celui des autres états semble avoir légèrement diminué.



Figure 17 : Répartition des clients par État

### II.3.2- Analyse géographique des vendeurs

La figure ci-dessous nous montre la répartition des vendeurs sur le territoire brésilien.

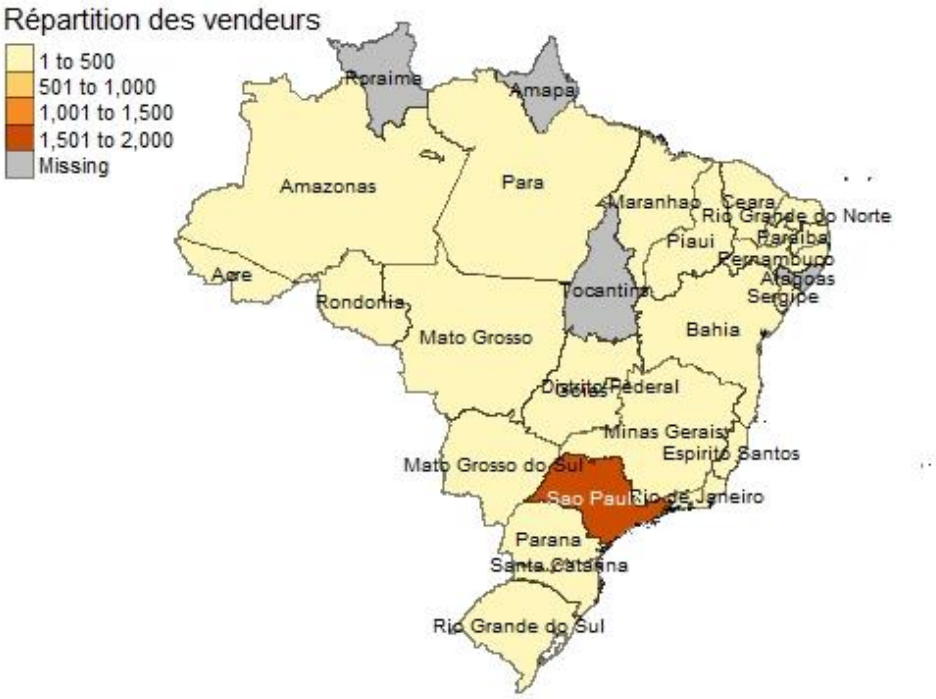



Figure 18 : Répartition des vendeurs par État

À la lecture du graphique, on constate que les vendeurs opérant pour le compte de l'entreprise Olist sont présents sur tout le territoire brésilien à l'exception de 4 états que sont : Roraima, Amapa, Tocantins et Alagoas. La plus grande concentration de vendeurs est dans l'état de Sao Paulo avec plus de 1500 vendeurs opérants.

### II.3.3- Analyse géographique du Chiffre d'affaires

La figure ci-dessous nous montre le Chiffre d'Affaires réalisé dans chaque Etat du territoire brésilien. À la lecture du graphique, on constate que l'état de Sao Paulo est celui où l'entreprise Olist a réalisé le plus de chiffres d'affaires et ceci, quelle que soit la période. Les états du Rio de Janeiro et de Minas Gerais complètent le podium aussi bien à la première qu'a la deuxième période. La grande concentration de client dans ces états explique en partie les conclusions tirées un peu plus haut. De plus la réalisation d'un chiffre d'affaires dans les états dépourvut de vendeurs à savoir : Roraima, Amapa, Tocantins et Alagoas ; signifie que les clients de ces états ont étés livrés par des vendeurs situés dans des états autres que le leur.

Ce qui certainement entraine des frais de transport plus importants qui pourraient décourager les clients de ces 4 états.

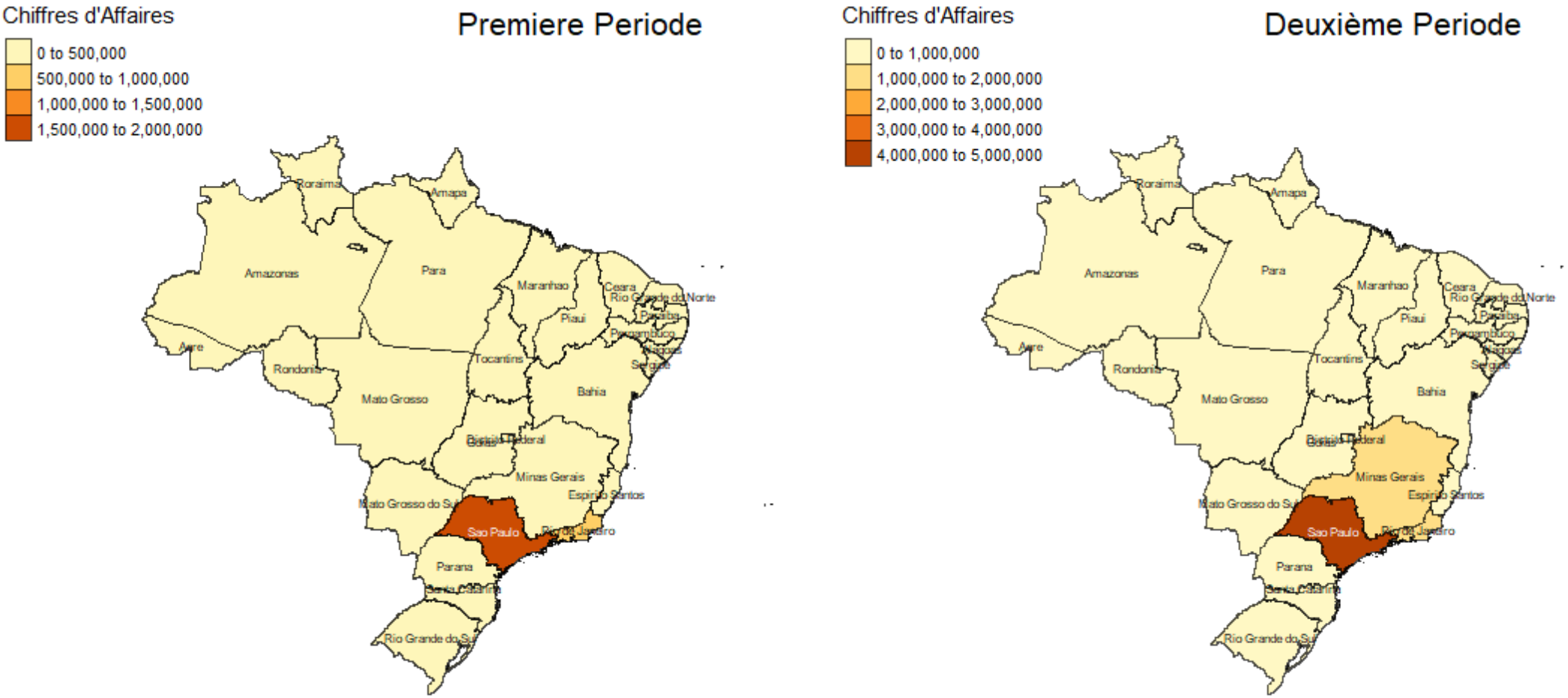



Figure 19 : Chiffres d'Affaires par État

On peut également déduire de la légende que le chiffre d'affaires réalisé par Olist au cours de la première période est inférieur à celle de la deuxième période. En effet le chiffre d'affaires réalisé sur la première période est de 4 305 305 U.M tandis que celui de la deuxième période est de 11 560 455 U.M Il s'agit d'une augmentation de 268,51 %. Ce qui est relativement énorme et compréhensible vu que le nombre de clients a presque triplé de la première à la deuxième période.

## II.4- Analyse de la satisfaction et Textmining

### A) Première Période

Pour l'analyse de la satisfaction nous nous servons de la base reviews qui nous renseigne sur les notes (score allant de 1 à 5) données par les clients pour chaque commande. Toutes les commandes n'ont pas reçu de note, mais également certaines commandes ont jusqu'à trois notes. Pour les commandes ayant plusieurs notes nous décidons de conserver uniquement la dernière note donnée puisqu'elle exprime à priori l'idée que se fait le client de la prestation de service d'Olist. Sur les 27 062 commandes de cette période, nous ne disposons que de 27 019 notes attribuées à ces commandes. Ci-dessous l'histogramme des scores :

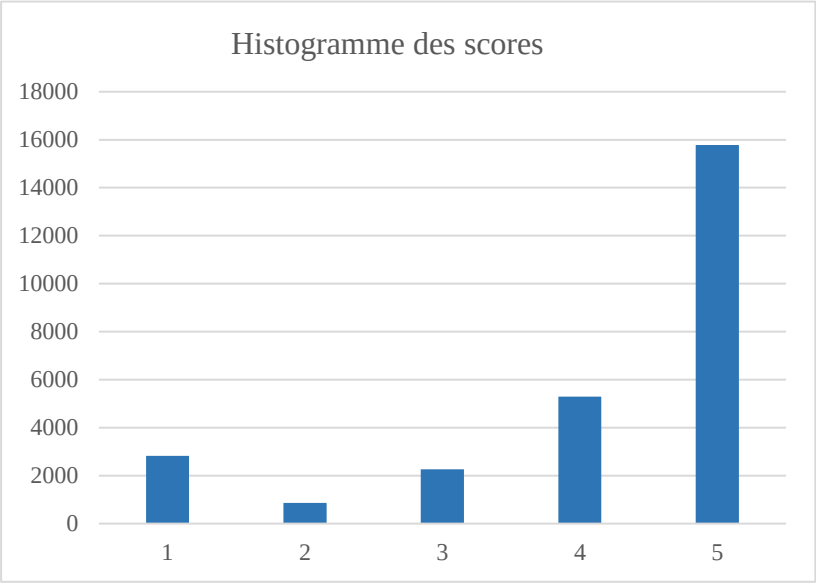



Figure 20 : Histogramme des scores - Première période

À la lecture de l'histogramme, la majorité des commandes ont reçu la note maximale. On remarque également que la majorité des commandes ayant reçu la note maximale de 5 sont des commandes ayant étés livrées avant la date de livraison estimée. La livraison d'une commande à temps pourrait donc être l'une des raisons qui expliquent une telle note.

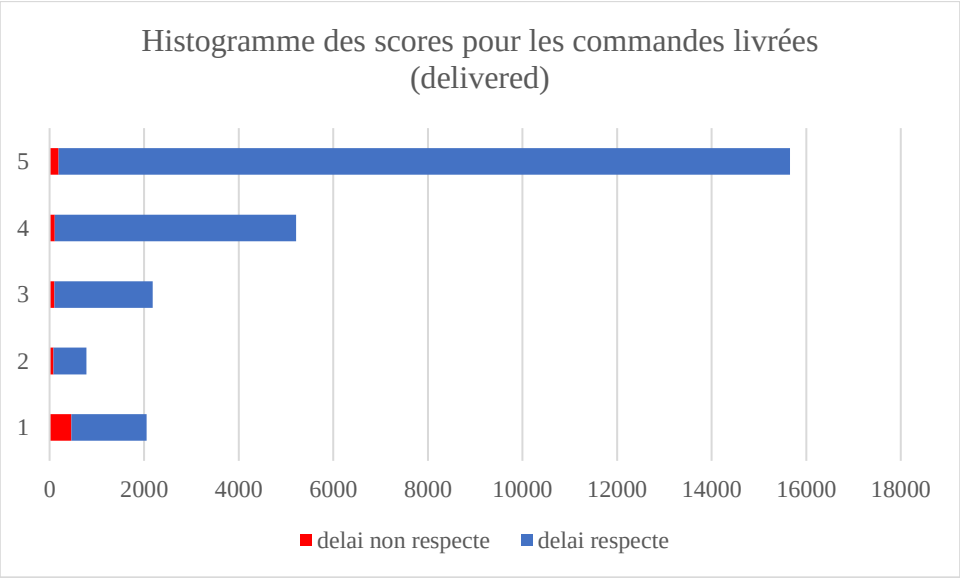



Figure 21 : Histogramme des scores pour les commandes livrées- Première période

**B) Deuxième Période**

Toutes les commandes de cette période ont étés notées par les clients. Les observations faites au niveau de la première période sont observées ici également. Il s'agit de commandes ayant plusieurs notes. Ci-dessous l'histogramme des scores.

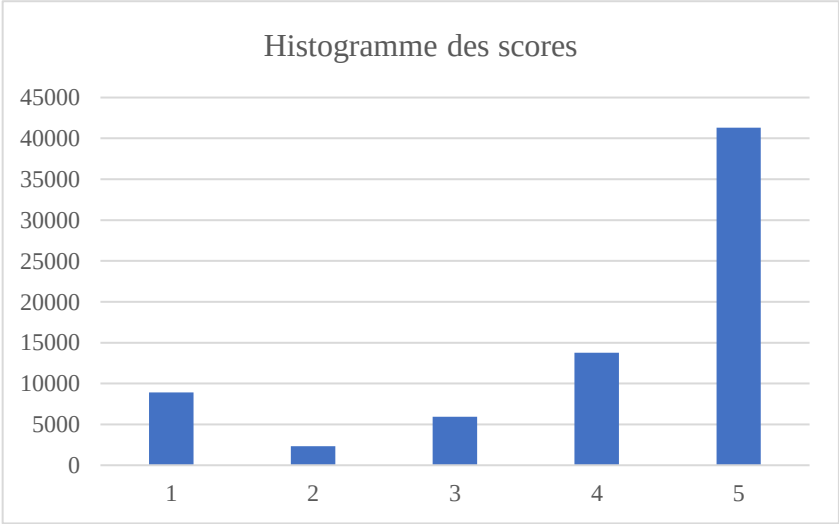



Figure 22 : Histogramme des scores - Deuxième période

À la lecture de l'histogramme, la majorité des commandes ont reçu la note maximale. On remarque également que la majorité des commandes ayant reçu la note maximale de 5 sont des commandes ayant étés livrées avant la date de livraison estimée (graphique ci-dessous). Nous pouvons tirer la même conclusion que celle de la première période.

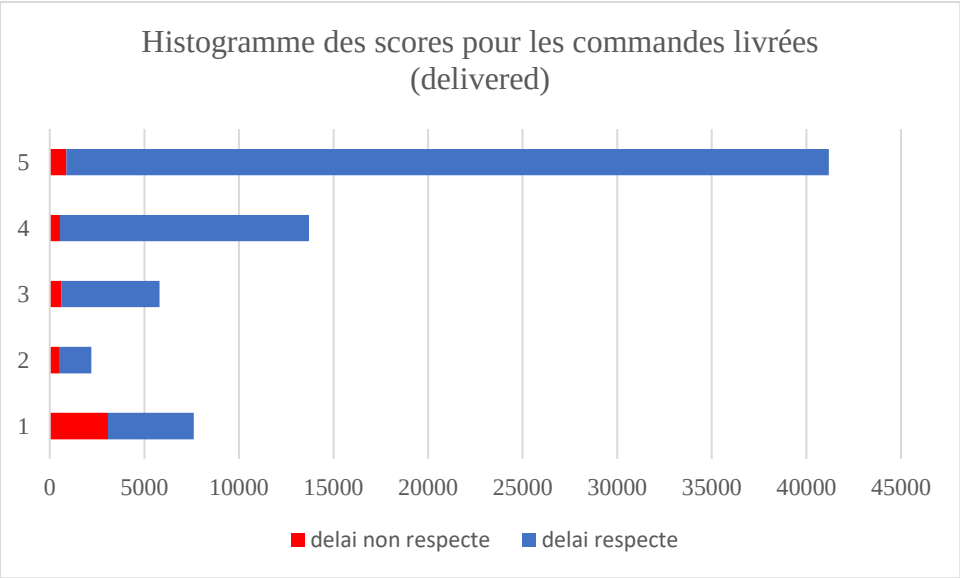



Figure 23 : Histogramme des scores pour les commandes livrées- Deuxième période

### II.4.3- Text mining

Le site Olist offre la possibilité à ses clients de faire des commentaires sur le service qui leur offert. Nous avons pu recueillir ces commentaires et procéder à une fouille de données textuelles. Ci-dessous le graphique présentant les 25 mots les plus fréquents employés par les clients aux deux périodes de nos données.

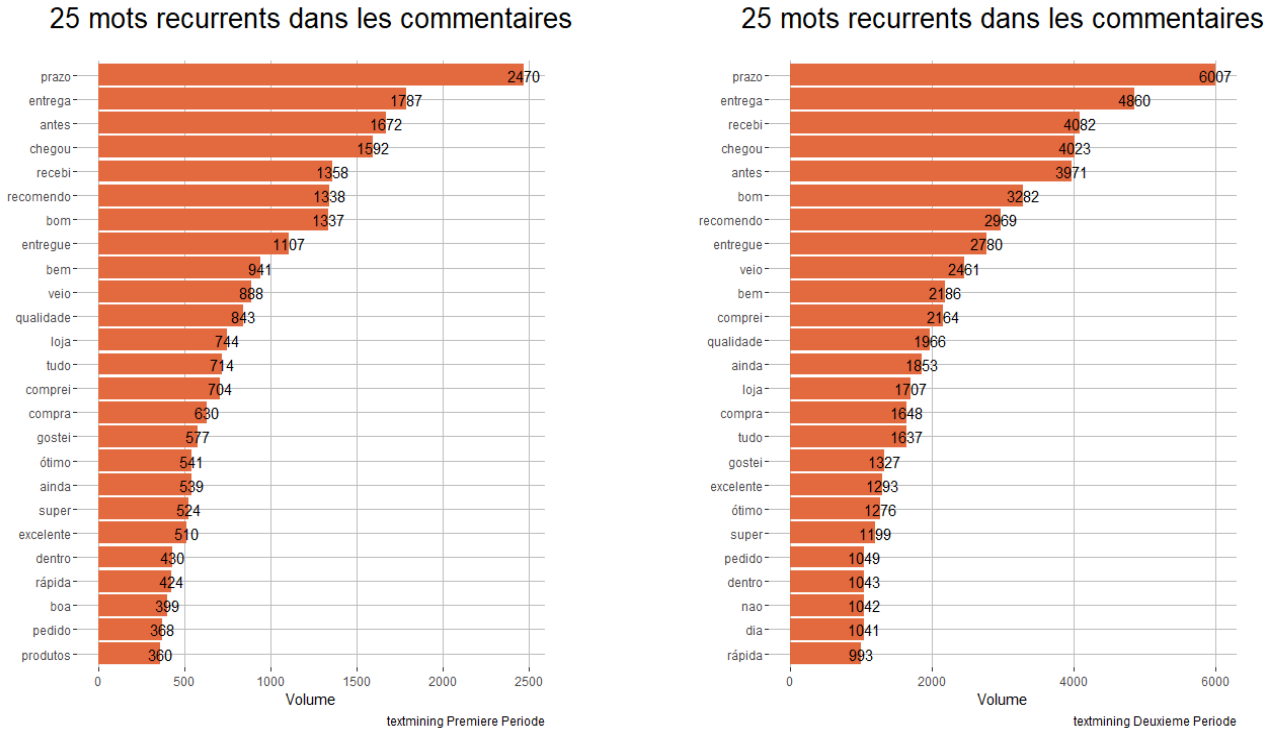



Figure 24 : Les 25 mots récurrents dans les commentaires à chaque période

Quel que soit la période étudiée, on remarque que le mot **prazo** (date limite) et le mot **entrega** (livré) ont étés le plus employés. On pourrait donc supposer l'intérêt particulier des clients par rapport aux délais de livraison estimés des produits qu'ils ont commandés. Aux annexes D et E, nous présentons le nuage des mots relative à chacune des périodes.

## II.5- Profil des clients

Dans cette section nous présentons les statistiques sommaires sur le montant total dépensé et le nombre d'achats des clients de chaque groupe et ceci sur chacune des périodes.
Ci-dessous, les statistiques sommaires :

Tableau 5 : Statistiques sommaires du montant dépensé par les clients / clusters

| Montant total dépensé par les clients (U.M) au cours de la première période | | | | | |
|---|---|---|---|---|---|
| Clusters | Minimum | Médiane | Moyenne | Maximum | Montant total |
| Groupe 1 | 10,1 | 89,4 | 103 | 251 | 2305591 |
| Groupe 2 | 251 | 360 | 420 | 963 | 1449626 |
| Groupe 3 | 967 | 1316 | 1559 | 7572 | 600187 |

Tableau 6 : Statistiques sommaires du nombre d'achats par clients / clusters

| Nombre d'achats au cours de la première période | | | | | |
|---|---|---|---|---|---|
| Clusters | Minimum | Médiane | Moyenne | Maximum | Écart type |
| Groupe 1 | 1 | 1 | 1,02 | 6 | 0,150 |
| Groupe 2 | 1 | 1 | 1,09 | 7 | 0,337 |
| Groupe 3 | 1 | 1 | 1,07 | 4 | 0,331 |

À la lecture des deux tableaux présentés ci-dessus, on remarque que le montant minimum le plus haut dépensé par un client est dans le groupe 3. Avec 385 individus dans ce groupe et un montant moyen de 1 559 UM dépensé par client, il s'agit du groupe avec la plus grande dépense par client. De plus c'est un groupe dont les clients achètent très peu. Les articles achetés par eux doivent être certainement, des articles de très grande valeur. Le qualificatif approprié de ce groupe est : **VIP.** Le mot **VIP** exprime : la clientèle la plus aisée. Le groupe 2 avec 3454 individus se classe deuxième en termes de dépense moyenne par client et achète relativement peu par rapport aux groupes 1. Les individus de ce groupe achètent des articles de valeurs moyennes. Les individus du groupe 1 complètent le podium. Ces derniers achètent un peu plus fréquemment que les autres et dépensent de petits montants à chaque achat. Le

groupe 2 est l'intermédiaire entre le groupe 3 et 1. La majorité des individus du groupe 3 vivent à Sao Paulo, Rio de Janeiro et Minas Gerais. Il s'agit des états les plus développés du pays.

On arrive aux mêmes conclusions lorsque nous nous intéressons à la deuxième période. En annexe A, les statistiques sommaires sur le montant total dépensé et le nombre d'achats des clients de la deuxième période. Cependant, ici c'est le groupe 3 (avec 681 individus) qui est le groupe **VIP**, avec 1 738 UM dépensé en moyenne par client. Il est suivi par ordre décroissant (en termes de dépense moyenne par client) du groupe 1 et du groupe 2. Au cours de la deuxième période, la majorité des individus du groupe 3 vivent à Sao Paulo, Rio de Janeiro et Minas Gerais ; états cités précédemment.

# III. Calcul de la valeur Client

## III.1- Segmentation RFM

Dans cette section, nous ferons une analyse RFM (récence, fréquence, montant) sur les données de la première et deuxième période afin de dégager des segments de marché. Ces différents segments seront comparés aux groupes issus de l'analyse exploratoire afin de dégager des similitudes ou dissemblances.

### III.1.1- Première période

La récence est le nombre de jours séparant la dernière date d'achat du client et la date d'étude. Pour date d'étude, nous avons pris le 31 décembre 2018. Cette même date sera utilisée à la période 2. Dans le calcul du montant dépensé par chaque client, nous avons omis les commandes annulées. De ce fait, le nombre de clients que nous avons est de 26 077 contrairement aux 26 275 clients dont nous disposons pour l'analyse exploratoire. Notons que pour le montant, il s'agit du montant moyen dépensé par le client. En d'autres termes :

$$Montant = \frac{Montant\ total\ dépensé\ par\ le\ client}{Nombre\ de\ commandes\ faites}$$

Les statistiques sommaires relatives aux différentes variables sont présentées dans le tableau suivant :

Tableau 7 : Statistique sommaire de la récence, fréquence, montant – Première période

| | Minimum | Médiane | Moyenne | Maximum | Écart type |
|---|---|---|---|---|---|
| Récence (jours) | 461 | 562 | 570(569,9) | 848 | 74,14 |
| Fréquence | 1 | 1 | 1(1,03) | 7 | 0,19 |
| Montant (U.M) | 10,07 | 103,11 | 160,85 | 6929 | 225,45 |

À la lecture du tableau on note que le montant maximum dépensé en moyenne par les clients est 6 929 U.M et le minimum est de 10,07 U.M. Le troisième quartile de la fréquence est de 1 ce qui signifie que sur cette période près de 75 % des clients ont achetés qu'une fois. Le client le plus récent a acheté il y a 461 jours et le plus ancien il y a 848 jours (date d'étude : 31 décembre 2018). L'histogramme de la récence, de la fréquence et du montant est présenté en annexe B. On remarque que les distributions ne sont pas gaussiennes. Pour le montant, nous procédons à une transformation logarithmique qui nous donne une distribution normale que voici :

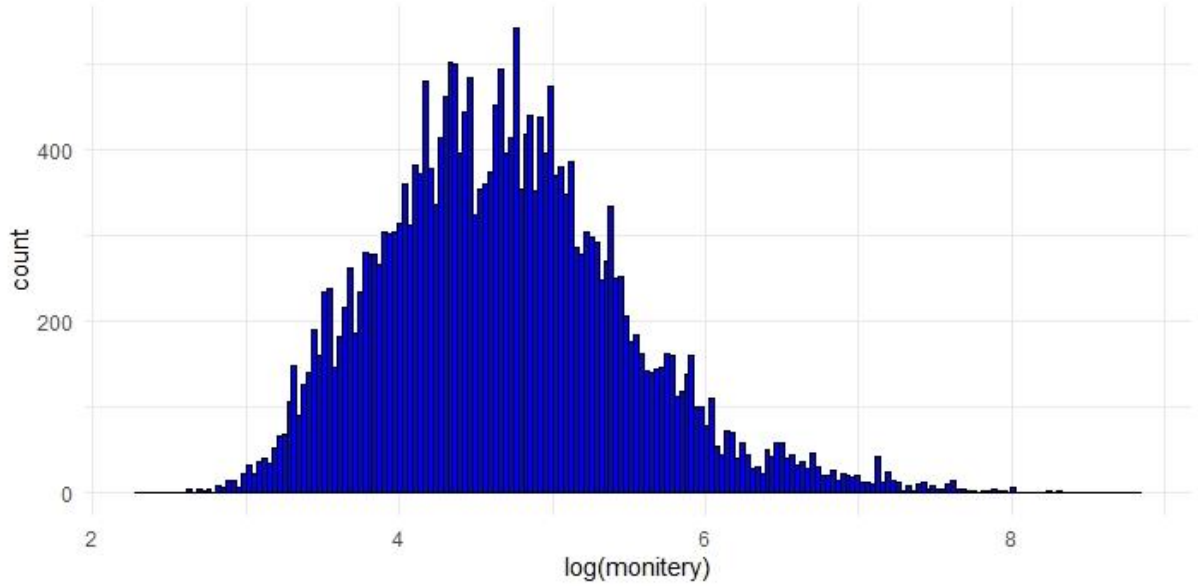



Figure 25 : Histogramme du log(montant) - Première période

Pour l'analyse RFM, nous avons attribué à la variable récence et le montant (monitery) un score allant de 1 à 4. Puisqu'un peu plus de 75 % des clients ont fait qu'un seul achat nous avons attribuer un score allant de 1 à 2 pour la variable fréquence et jugeons de lui attribuer un poids faible de 20%. Dans le calcul du score total, la récence et le montant ont un poids de 40 %. Ci-joint la répartition des individus dans leur groupe respectif.

Tableau 8 : Segmentation RFM - première période

| Clusters | Effectif | Fréquence (Moyenne) | Minimum Montant | Maximum Montant | CA (moyen) |
|---|---|---|---|---|---|
| Groupe 1 | 5184 | 1,02 | 13,68 | 136,23 | 325226,1167 |
| Groupe 2 | 15748 | 1,03 | 10,07 | 6929,31 | 2257184,229 |
| Groupe 3 | 5145 | 1,03 | 125,82 | 3048,27 | 1612172,042 |

À la lecture du tableau, on remarque que les individus du groupe 3 sont peu nombreux que ceux des autres groupent. Le groupe 3 avec un score de 4, concentre les individus qui ont procédé à des achats de montant supérieur. On pourrait dire que le groupe 3 de l'analyse RFM correspond au groupe VIP issu du profilage des clients. Il vient ensuite le groupe 2 avec un score de 3 suivis du groupe 3 (score 2). Notons que tous les groupes ont une fréquence d'achat moyenne presque identique. Puisque la segmentation RFM est une technique prisée des marketeurs pour segmenter leur clientèle, nous utiliserons ces groupes dans la suite de notre analyse en particulier le taux d'attrition.

### III.1.2- Deuxième période

Les statistiques sommaires relatives aux différentes variables sont présentées dans le tableau suivant :

Tableau 9 : Statistique sommaire de la récence, fréquence, montant – deuxième période

| | Minimum | Médiane | Moyenne | Maximum | Écart type |
|---|---|---|---|---|---|
| Récence (jours) | 119 | 288 | 288(287,7) | 460 | 93,34 |
| Fréquence | 1 | 1 | 1(1,026) | 10 | 0,18 |
| Montant (U.M) | 9,59 | 106,78 | 160,98 | 13664,08 | 218,83 |

À la lecture du tableau on note que le montant maximum dépensé en moyenne par les clients est 13 664,08 U.M et le minimum est de 9,59 U.M. Le troisième quartile de la Fréquence est de 1 ce qui signifie que sur cette période près de 75 % des clients ont achetés qu'une fois. Le client le plus récent a acheté il y a 119 jours et le plus ancien il y a 460 jours (date d'étude : 31 décembre 2018). L'histogramme de la récence, de la fréquence et du montant est présenté en annexe C. On remarque que les distributions ne sont pas gaussiennes. Pour le montant, nous procédons donc à une transformation logarithmique. Nous procédons par suite à l'analyse RFM sur les mêmes bases que précédemment. Cependant lors de cette analyse nous nous sommes

retrouvés avec 53 individus formant un 4ème groupe. Nous les avons réunis avec un groupe afin d'être dans les mêmes dispositions que précédemment. Ci-dessus la répartition :

Tableau 10 : Segmentation RFM - deuxième période

| Clusters | Effectif | Fréquence (Moyenne) | Minimum Montant | Maximum Montant | CA (Moyen) |
|---|---|---|---|---|---|
| Groupe 1 | 14288 | 1,09 | 611,63 | 198,41 | 929600,7733 |
| Groupe 2 | 41775 | 1,03 | 9,59 | 13664,08 | 5787858,503 |
| Groupe 3 | 14058 | 1,03 | 130,69 | 7274,88 | 4570324,174 |

À la lecture du tableau, on remarque que les individus du groupe 3 sont peu nombreux que ceux des autres groupent. Le groupe 3 avec un score de 4, est le groupe VIP issu du profilage des clients. Il vient ensuite le groupe 2 avec un score de 3 suivis du groupe 3 (score 2). Notons que tous les groupes ont une fréquence d'achat moyenne presque identique.

## III.2- Calcul de la valeur client observée

La valeur vie d'un client est élaborée sur la base de la durée de vie moyenne d'un client et du cours des consommations. Elle va également permettre de calculer le cout d'acquisition d'un client et de valoriser une entreprise à partir du portefeuille client.

Elle est obtenue à travers l'équation suivante :

$$Valeur\ Vie\ Client = (valeur\ moyenne\ des\ ventes)\ x\ (nombre\ moyen\ de\ ventes)\ x\ (durée\ moyenne\ de\ la\ relation\ client)$$

Il s'agira de calculer à chaque période la valeur vie client pour chacun des clients. Cette dernière dépend de la valeur moyenne des ventes obtenues déjà à travers la variable montant de la section précédente. On a aucune information précise sur la durée moyenne de la relation client. Tandis que le nombre moyen de ventes est le nombre d'achats, faites-en moyenne par l'individu par semaine ou par mois, etc. Nous ne disposons d'aucune information de la sorte. Cependant nous savons combien de fois un client achète dans une période. Nous utiliserons donc cette donnée comme le produit du nombre moyen de ventes et de la durée moyenne de la relation client. Nous présentons dans le tableau suivant quelques statistiques sommaires relatives à la valeur client observée.

Tableau 11 : Statistique sommaire sur la valeur client observée

| | Minimum | Médiane | Moyenne | Maximum | Ecart-type |
|---|---|---|---|---|---|
| Première période | 10,07 | 105,28 | 165,10 | 7571,63 | 233,7937 |
| Deuxième période | 9,59 | 108,30 | 164,86 | 13664,08 | 223,6628 |

- Toutes ces statistiques sont en unité monétaire excepté l'écart-type.

On note que le montant maximum dépensé par les clients à la première et deuxième période est respectivement de 7 571,63 U.M et 13 664,08 U.M avec un montant moyen de 165,10 U.M et 164,86 U.M. La moitié des clients ont dépensés à la première et deuxième période respectivement un montant inférieur à 105,28 U.M et 108,30 U.M.

## III.3- Modélisation du prix d'un produit

Dans cette section, il s'agira d'identifier les déterminants du prix d'un produit. En d'autres termes il ne s'agira pas de prévoir le prix d'un produit, mais plutôt d'identifier les sous-jacents.

Le site Olist propose 32 952 produits, cataloguer dans 74 catégories distinctes. Il y a près de 3096 vendeurs d'articles chez Olist. Tous ne proposent pas le même produit. On note également que deux vendeurs qui proposent un même produit peuvent résider dans des états différents et fixer des prix différents pour le même produit selon le client et le lieu de résidence de ce dernier. On remarque aussi qu'un vendeur peut vendre le même produit à des prix différents à deux clients vivants dans la même ville. De plus, bon nombre des produits proposés par l'entreprise n'ont qu'un unique vendeur. Un constat également fait est la variation des frais de transport au même titre que les prix des produits.

À partir des informations que nous disposons et des remarques faites, les variables indépendantes de notre modèle sont : **customer_city**, **seller_city**, **order_id,** (nombre de fois le produit est acheté dans une commande), **freight_value (frais de transport)**. La variable dépendante est le prix nommé « **price** ». Les variables de la base qui renseignent sur les caractéristiques du produit ne varient pas puisqu'il s'agit du même produit. Nous les ignorons donc.

Pour pouvoir procéder à la modélisation nous nous sommes intéressés au produit le plus vendu toutes périodes confondues. Il s'agit du produit ayant pour identifiant **99a4788cb24856965c36a24e339b6058** qui s'est retrouvé dans 467 commandes. Lors de la modélisation, nous nous sommes rendu compte qu'un unique vendeur propose ce produit et il est domicilié dans la ville de Ibitinga dans l'état de Sao Paulo. On fera donc fi de la variable seller_city. De plus sur les 467 observations relatives à ce produit, la variable customer_city contient 234 modalités différentes (234 villes). Ce qui est beaucoup. Nous décidons donc qu'au lieu de la variable customer_city nous utiliserons la variable customer_state avec 27 modalités. La variable price a principalement 6 valeurs : 74 ; 79,90 ; 83,79 ; 84 ; 86,90 ; 89,90.

## III.3.1- Analyse de corrélation

### A) Price et freight_value

Pour tester la corrélation entre les deux variables, nous procédons à la réalisation du nuage de point afin d'avoir une idée de la forme de la liaison

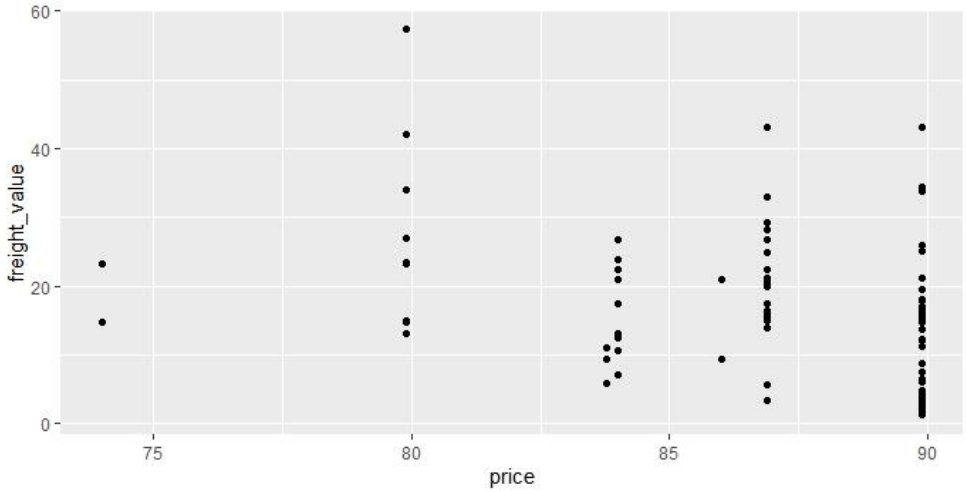



Figure 26 : Nuage de point entre price et freight_value

L'analyse du nuage de point nous montre une forme non linéaire et non monotone. On pourrait donc conclure l'inexistence d'une corrélation entre ces deux variables. Procédons au test de Kendal pour confirmer nos dires. Ce test est le plus approprié, car il a des ex æquo dans les variables. Les hypothèses du test sont :

$$\begin{cases} Hypothèse\ nulle : Les\ variables\ sont\ corrélés \\ Hypothèse\ alternative :\ les\ variables\ sont\ indépendants \end{cases}$$

Ci-dessous le résultat du test :

Tableau 12 : Test de Kendall

| Test de Kendall | |
|---|---|
| Tau de Kendall (τ) | -0,1978 |
| p-value | $1{,}088 \times 10^{-7}$ |
| Statistique de test | -5,3113 |

La valeur du tau de kendall τ, nous permet de conclure que les deux variables ne sont pas corrélées ni positivement ni négativement. En effet $|\tau| < 0{,}7$. Toutefois cette corrélation est presque nulle. En effet la statistique de test et la p-valeur associée qui teste la nullité de la pente entre les deux variables permettent de rejet l'hypothèse de nullité de τ.

Pour la modélisation, le modèle choisi est un modèle logistique avec pour fonction de lien le logit. Pour cela nous procédons à la scission en deux classes de la variable price. La première classe sera [74 ;82[ et la deuxième [82 ;90[. Nous sommes en présence de modalités ordonnées. Nous ferons donc une régression logistique ordinale après étude des liaisons en variables.

Ci-joint le boxplot de la variable freight_value par modalités de la variable price :

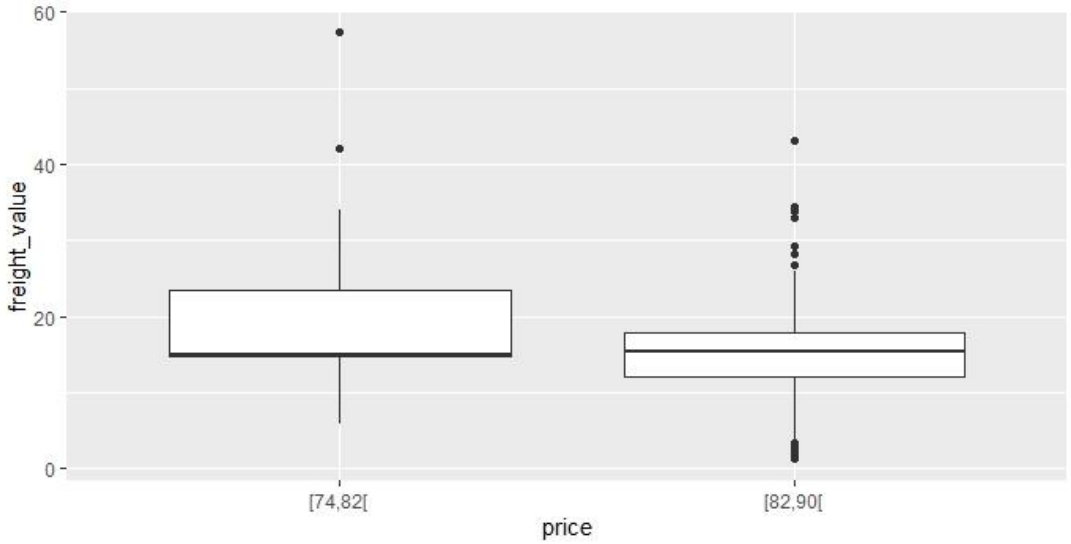



Figure 27 : Boxplot de la variable freight_value

À la lecture du graphique, il semble que la moyenne, le premier et le troisième quartile de la variable freight_value ne sont pas identiques entre les deux groupes. Nous procédons au test de Kruskal-Wallis afin de dégager si les deux échantillons proviennent de la même distribution. Notons que l'hypothèse de normalité (annexe F) n'étant pas respecté ici nous faisons un test non paramétrique sur la variable freight_value. Les hypothèses du test sont :

$$\begin{cases} Hypothèse\ nulle : les\ deux\ échantillons\ proviennent\ d'une\ même\ distribution \\ Hypothèse\ alternative : les\ deux\ échantillons\ proviennent\ de\ distributions\ différentes \end{cases}$$

Ci-dessous le résultat du test :

Tableau 13 : Test de Kruskal-Wallis

| Test de Kruskal-Wallis | |
|---|---|
| Chi-deux | 1,6823 |
| df | 1 |
| p-value | 0,1946 |
| Eta-carré | 0,00147 |

La valeur de la p-value, nous permet de rejeter l'hypothèse nulle. L'eta-carré, basé sur la statistique de Chi-deux du test, peut être utilisé comme mesure de la taille de l'effet du test Kruskal-Wallis. L'estimation eta-carré est comprise entre 0 et 1 ; la valeur multipliée par 100 indique le pourcentage de variance de la variable dépendante expliqué par la variable indépendante. Avec une valeur de 0,00147, on conclut à un petit effet de taille.

### B) Price et customer_state

Puisqu'il s'agit de deux variables qualitatives, nous procédons au test V de Cramer. Le V de Cramer nous donne l'intensité de la liaison entre les deux variables. Nous trouvons un V de Cramer égale à 0,1995. Nous sommes donc en présence d'une liaison faible.

### C) Modélisation

Nous procédons à la modélisation. Malgré la faible liaison entre les deux variables qualitatives nous intégrons la variable Customer_state dans le modèle logit ordinal. Malheureusement, nous recevons les messages d'avertissement suivants : la matrice Hessienne est un entier numérique. Les paramètres déterminés ne sont pas uniques. Les critères de convergence absolus sont obtenus, mais les critères de convergences relatives ne le sont pas. Nous supposons que cette erreur est peut-être due au grand nombre de modalités de la variable customer_unique _id que malheureusement nous ne pouvons procéder à un regroupement à l'état actuel. Nous nous contentons donc de la seule variable explicative restante. Les résultats de l'estimation sont :

```
formula: price ~ freight_value
data:    produit

 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  467  -175.93 355.85 5(0)  1.39e-09 2.7e+03

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
freight_value -0.05053    0.01945  -2.598  0.00939 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
                Estimate Std. Error z value
[74,82[|[82,90[  -2.7833     0.3748  -7.426
```

Figure 28 : Modèle logistique ordinale du prix d'un produit

La p-value étant inférieur à 5 % alors la variable freight_value est statistiquement significative. Puisque le coefficient associé à freight_value < 0, alors une augmentation des frais de transport d'une unité, diminue la probabilité que le prix du produit passe de [74 ; 82[ à [82 ; 90[ .

## III.4- Fréquence d'achat et Taux d'attrition

### III.4.1- Fréquence d'achat

À partir des données dont nous disposons, nous procédons au calcul de la fréquence d'achat. Il s'agit d'une fréquence moyenne d'achat sur chacune des périodes et par groupe (3 clusters). Nous aurons donc une fréquence moyenne d'achat relative à chacun des groupes de chaque période. La formule utilisée pour le calcul de cette fréquence moyenne est :

$$fréquence\ moyenne\ d'achat = \frac{nombre\ total\ de\ commande\ faites\ par\ les\ clients\ du\ groupe}{nombre\ de\ client\ dans\ le\ groupe}$$

### III.4.2- Taux d'attrition

L'objectif visé par la création de deux périodes d'études est le calcul du taux d'attrition. En effet ce dernier est obtenu par l'expression suivante :

$$Taux\ d'attrition = \frac{nombre\ de\ désabonnement}{nombre\ d'abonné\ en\ début\ de\ période}$$

Grâce à la séparation en deux périodes, nous pouvons estimer le nombre de désabonnements. Pour cela nous considérons qu'un individu ayant acheté à la première période s'est désabonné

si à la deuxième période il n'a procédé à aucun achat. C'est sur cette base que repose le calcul du taux d'attrition. Après quelques des données, on note que 638 individus de la première période ont procédé à un achat à la deuxième période. Le nombre de désabonnements est donc de 25 439. Ce qui représente 97,55 % des clients de la première période.

Pour procéder à la modélisation du taux d'attrition, nous recueillons les informations que nous disposons sur les clients susceptibles de nous aider. Il s'agit du groupe (analyse RFM) auquel appartient le client et de l'état de résidence (customer_state) ; la ville de résidence (customer_city) contient trop de modalités. Nous procédons au test du V de Cramer pour apprécier la liaison entre les variables indépendantes et la variable dépendante qui est le renouvèlement d'un achat à la deuxième période. Cette variable est nommée **renouvele**. Nous avons procédé à une codification de cette variable à travers la variable **renouvele_code** qui prend 1 si l'individu à procéder à un achat en deuxième période et 0 sinon. Le test du V de cramer nous donne :

Tableau 14 : Test V de Cramer pour le modèle du taux d'attrition

| Test V de Cramer entre la variable **renouvele** | |
|---|---|
| Groupe | 0.02344265 |
| Customer_state | 0.0311868 |

À la lecture des valeurs, on peut dire que la liaison entre les explicatifs et la variable explicative est très faible. Un modèle basé sur ces variables ne sera pertinent ni globalement significatif. Nous décidons donc de laisser tomber la modélisation du taux d'attrition par un modèle logit. Pour la détermination du taux d'attrition, nous procédons par calcul en nous servant de l'expression donnée plus haut. Le calcul se fera par groupe d'individu et nous aurons donc un taux d'attrition relatif à chaque groupe. N'ayant pas d'informations supplémentaires pour procéder au calcul du taux d'attrition pour la deuxième période, nous décidons d'attribuer la même valeur du taux d'attrition aux groupes des deux périodes ayant un même score. Ainsi le taux d'attrition des groupes est le suivant :

Tableau 15 : Taux d'attrition

| Taux d'attrition | |
|---|---|
| Groupe 1 | 0,982 |
| Groupe 2 | 0,974 |
| Groupe 3 | 0,972 |

Avec taux d'attrition de 0,982 , l'entreprise Olist perd en moyenne par période 98,2% de ces clients actifs du groupe 1.

## III.5- Valeur Client potentiel (CLV)

La valeur client potentiel est obtenue par l'expression suivante :

$$Valeur\ client\ potentielle = panier\ moyen\ par\ client * fréquence\ moyenne\ d'achat * duréee\ de\ la\ relation\ client$$

Le panier moyen par client est égal au montant moyen dépensé par le client dans une période. Il est déjà calculé à travers la variable montant (monitery) utilisée dans l'analyse RFM.

Quant à la durée de la relation client, il s'agit de l'inverse du taux d'attrition. Dans ce cas, pour un taux d'attrition de 0,982, la durée de la relation client est de 1,02. Alors les clients d'Olist vont rester peu plus d'un an en moyenne. Remarquons ici qu'un an correspond approximativement à la durée de nos périodes (première et deuxième), bien que ces périodes ne fassent pas exactement 1 an. Les raisons ont été expliquées un peu plus haut.

Nous procédons par la suite au calcul de la valeur client potentiel et après à l'écart entre cette dernière et la valeur client observé. Ci-dessus l'histogramme de la valeur client potentiel par groupe à la première période :

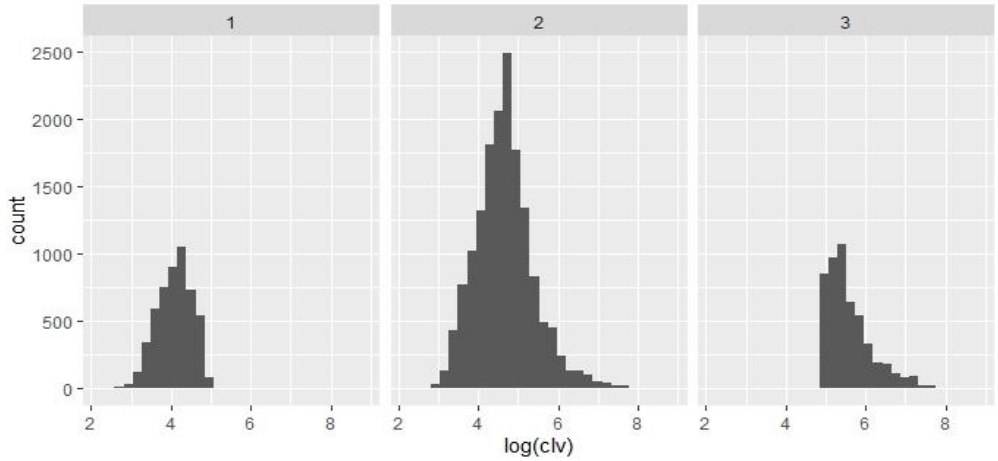



Figure 29 : Histogramme du log (CLV) par groupe

La valeur maximale de la valeur vie client de cette période pour les groupes 1,2 et 3 est respectivement de 142,0077 ; 7319,865 ; 3247,667. La valeur minimale de la valeur vie client à la même période pour les groupes 1, 2 et 3 est respectivement de 14,26018 ; 10,63757 ; 134,0556. En annexe G, est présenté l'histogramme du log de la valeur vie client à la deuxième période. Ci-dessous, l'histogramme de l'écart entre la valeur vie client observé et potentiel pour la première période :

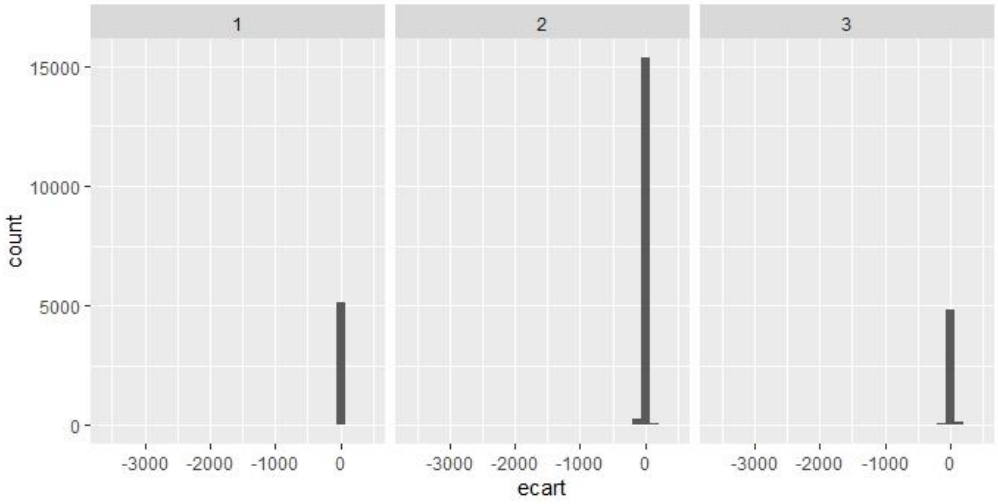



Figure 30 : Histogramme des écarts entre la CLV potentielle et observée

Les écarts sont pour la plupart concentrés autour de 0, signent que les valeurs clients potentiels sont assez proches des valeurs clients observées. En annexe H, l'histogramme des écarts entre les CLV potentielle et observée au cours de la deuxième période. Ici également nous arrivons aux mêmes conclusions.

## Conclusion

Le succès d'une entreprise se base sur sa capacité à offrir un service ou produit qui répond aux besoins de son segment de marché. Cependant une dimension très importante, de ce succès repose sur la relation client. Avec un taux d'attrition qui avoisine les 97 % pour chaque groupe de la clientèle, l'entreprise Olist fera face à une énorme proportion de clients perdus. À cela s'ajoute un taux de réachat très faible de moins de 25 %. À long terme ces difficultés si elles sont négligées, seront des freins à la croissance de l'entreprise. Pour éviter un tel scénario, l'entreprise devra mettre en place des stratégies efficientes.

Un autre aspect qui doit susciter l'attention de l'entreprise est l'inexistence de vendeurs dans certains états brésiliens ou est présent une partie de sa clientèle. Malgré tout il faut reconnaitre à l'entreprise deux points forts. L'un est sa capacité à attirer de nouveau client qui se traduit par une forte augmentation du nombre de clients de la première période (2016-2017) à la deuxième période (2017-2018) et l'autre qui est le respect des délais de livraison.

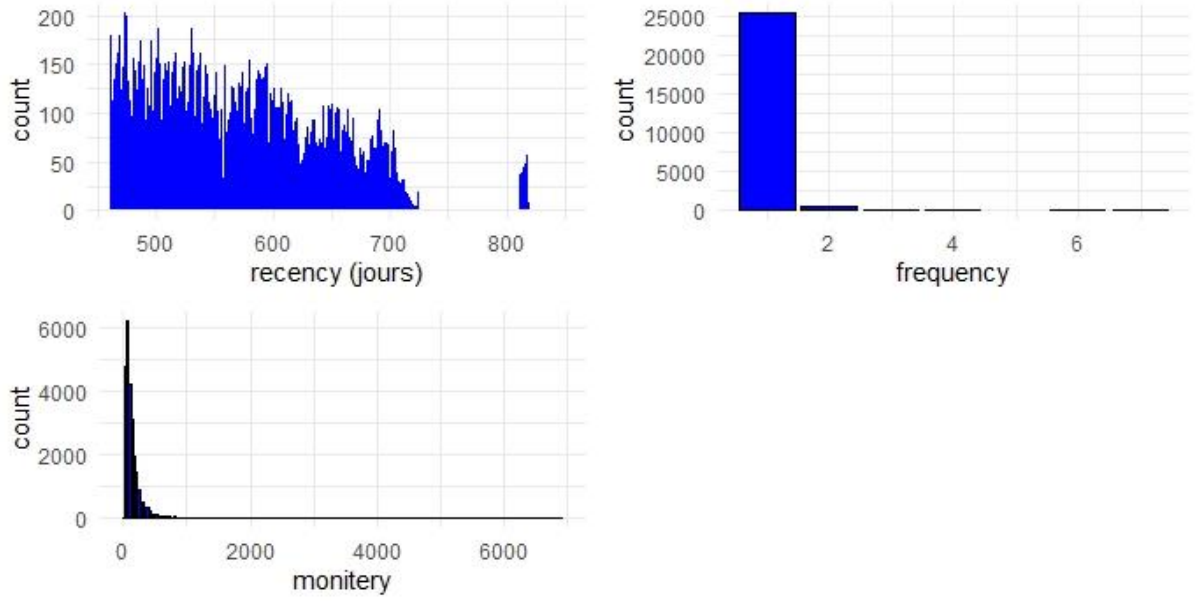
# Annexe

## A) Statistique sommaire pour le Profilage des clients de la deuxième période

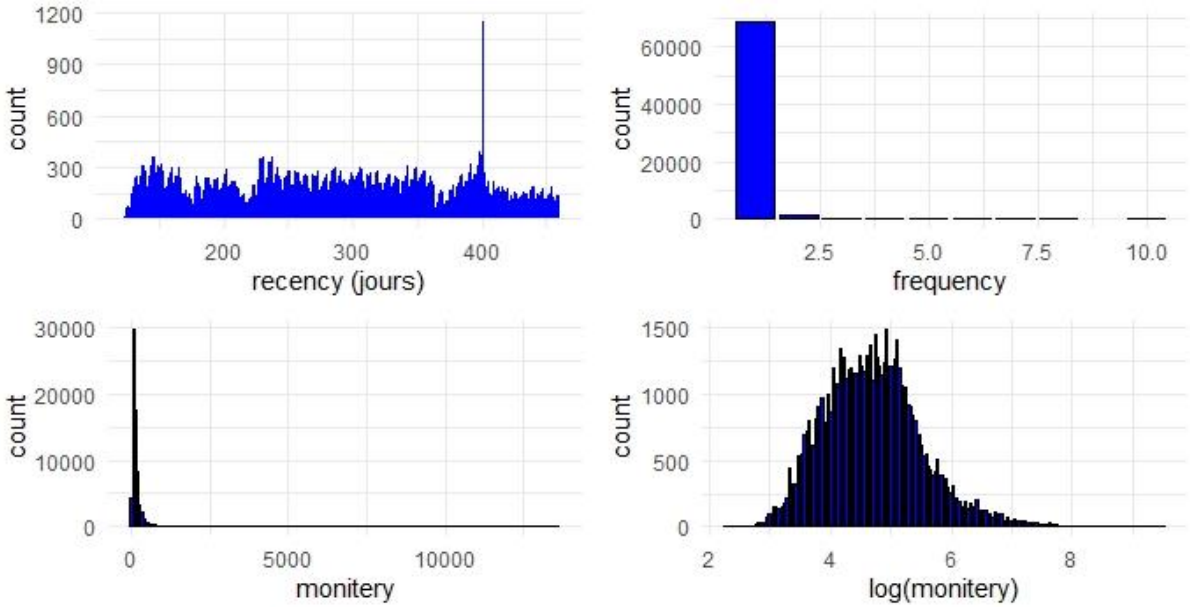
| Montant total dépensé par les clients (U.M) au cours de la deuxième période | | | | | |
|---|---|---|---|---|---|
| Clusters | Minimum | Médiane | Moyenne | Maximum | Montant total |
| Groupe 1 | 304 | 426 | 497 | 1117 | 3443603 |
| Groupe 2 | 0 | 97,2 | 112 | 304 | 1183702 |
| Groupe 3 | 1118 | 1493 | 1738 | 13664 | 7026306 |

| Clusters | Minimum | Médiane | Moyenne | Maximum |
|---|---|---|---|---|
| Groupe 1 | 1 | 1 | 1(1,01) | 10 |
| Groupe 2 | 1 | 1 | 1(1,09) | 5 |
| Groupe 3 | 1 | 1 | 1(1,10) | 7 |

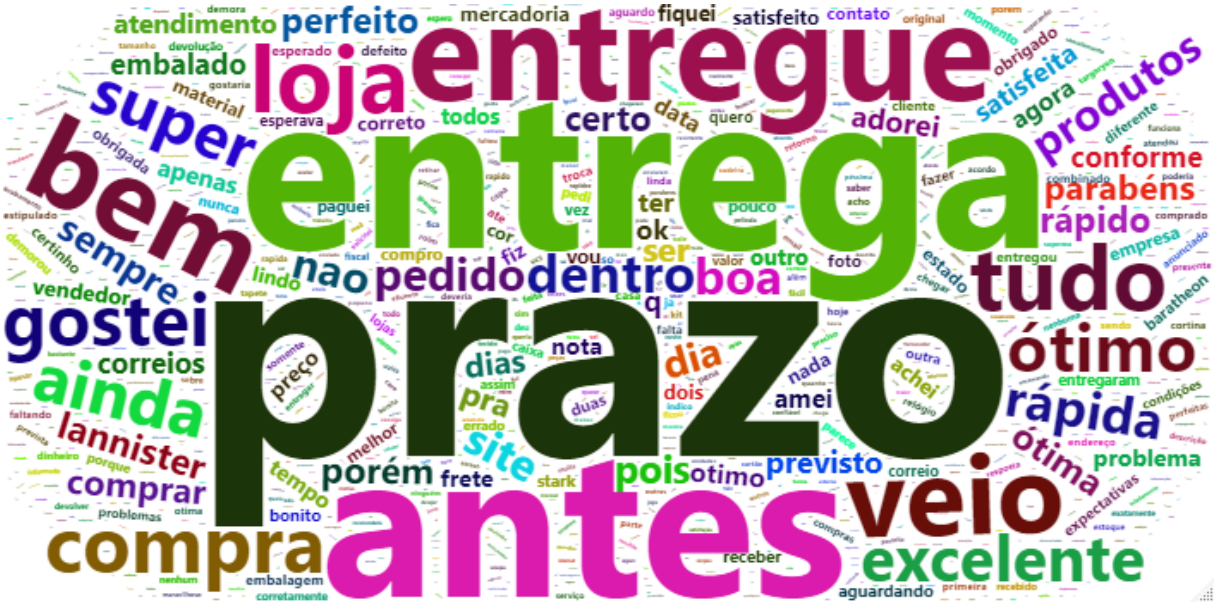
## B) L'histogramme de la récence, de la fréquence et du montant première période



**C) L'histogramme de la récence, de la fréquence et du montant première période**

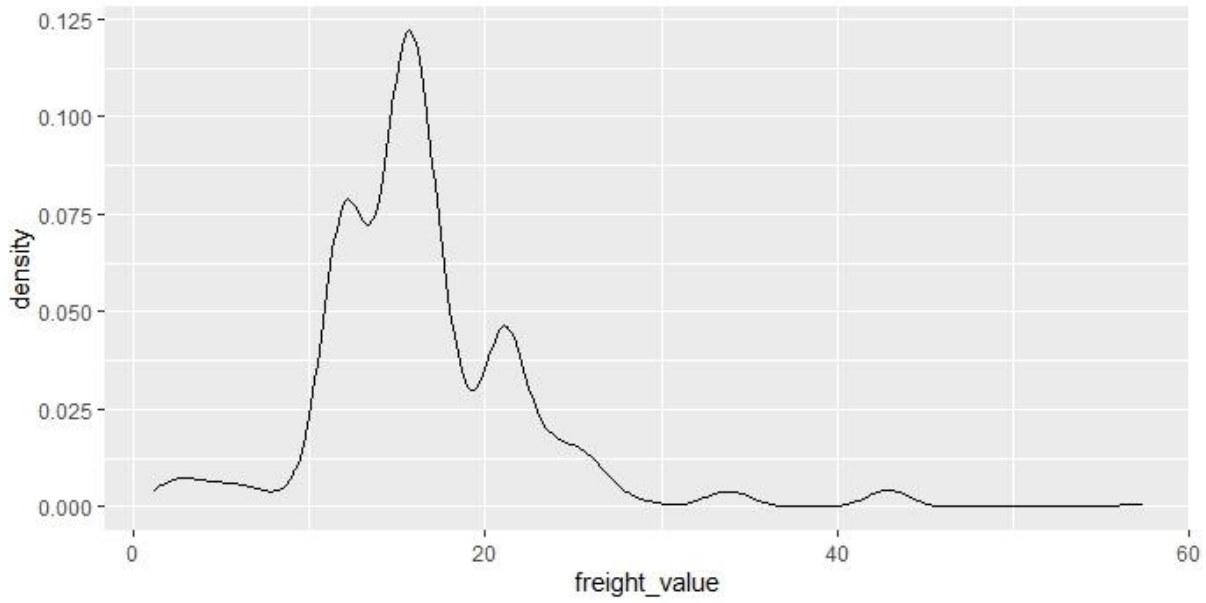



**D) Nuage des mots de la première période**

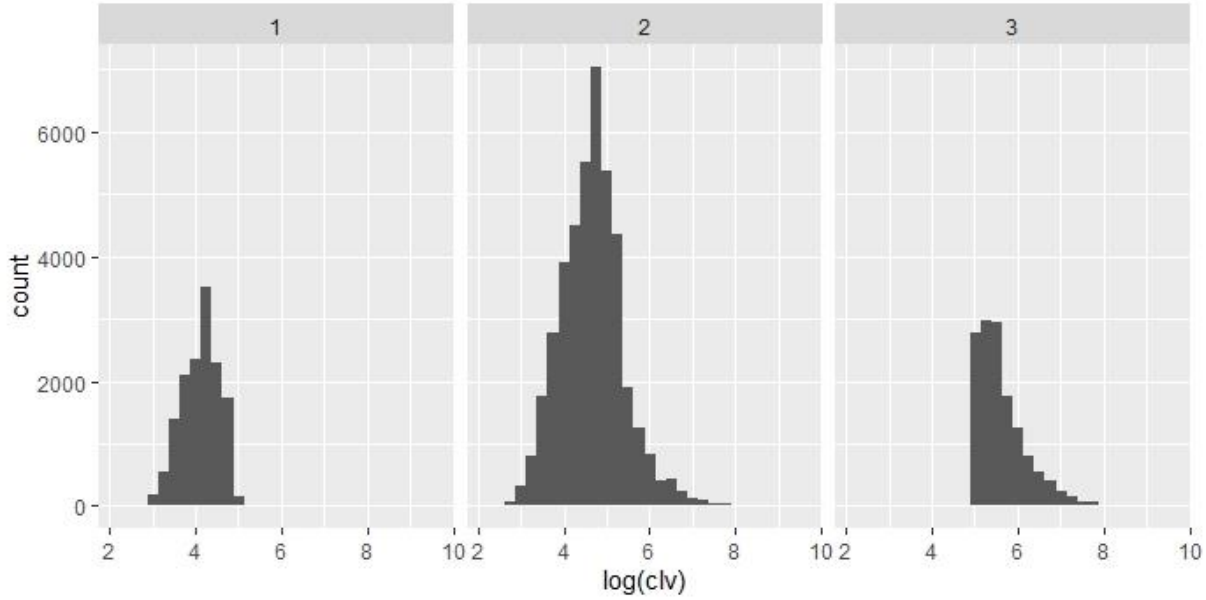


**E) Nuage des mots de la deuxième période**


**F) Densité de la variable freight_value**



**G) Histogramme du log de la valeur client potentiel à la deuxième période**




**H) Histogramme des écarts entre les CLV potentielle et observée au cours de la deuxième période**