Projet 5 :

Segmentez des clients d'un site e-commerce

Nom : TRABIS

Prénom : Mohamed

Table des matières

Introduction

Introduction

- **Contexte** :

  Olist est une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne.

  Cette entreprise souhaite que vous fournissiez à ses équipes d'e-commerce une **segmentation des clients** qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

  Vous devrez fournir à l'équipe marketing une description de votre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

Introduction

- **Mission** :

  La mission consiste à aider les équipes d'***Olist*** à comprendre les différents types d'utilisateurs. En utilisant des méthodes non supervisées pour regrouper des clients de profils similaires. Ces catégories pourront être utilisées par l'équipe Marketing pour mieux communiquer.

- **Informations complémentaires** :
  - 3 % des clients du fichier de données partagé ont réalisé plusieurs commandes.
  - La segmentation proposée doit être exploitable et facile d'utilisation par l'équipe Marketing, pour différencier les bons et moins bons clients en termes de commandes et de satisfaction

Préparation des données

Préparation des données- Évaluation et découverte des données

- **Base de données :**
  - Pour cette mission, ***Olist*** a fournit une base de données anonymisée comportant des informations sur l'historique des commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients depuis janvier 2017.
  - Les données clients sont partagées sous format de 9 fichiers CSV :

1. olist_geolocation_dataset.csv
2. olist_customers_dataset.csv
3. olist_order_items_dataset.csv
4. olist_order_payments_dataset.csv
5. olist_order_reviews_dataset.csv
6. olist_orders_dataset.csv
7. olist_products_dataset.csv
8. olist_sellers_dataset.csv
9. product_category_name_translation.csv

Préparation des données- Nettoyage et validation des données

- Les étapes effectuées pour le nettoyage et la validation des données :

  - Importer les 9 fichiers CSV.
  - Fusionner toutes les données dans une seule DataFrame.
  - Mutualiser les données.
  - Supprimer les colonnes inutiles
  - Remplir les valeurs manquantes

Préparation des données- Nettoyage et validation des données

- Suite à cette fusion nous avons 114092 lignes et 37 colonnes.

- Supprimer les colonnes qui contiennent des informations sur les dimensions :

  'product_name_lenght',' product_description_lenght', 'product_weight_g', 'product_length_cm', 'product_height_cm', 'product_width_cm'

- Ci-dessous les informations du DataFrame :

```
Data columns (total 33 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   order_id                       119143 non-null  object
 1   customer_id                    119143 non-null  object
 2   customer_unique_id             119143 non-null  object
 3   product_id                     118310 non-null  object
 4   seller_id                      118310 non-null  object
 5   review_id                      118146 non-null  object
 6   order_item_id                  118310 non-null  float64
 7   price                          118310 non-null  float64
 8   order_status                   119143 non-null  object
 9   customer_zip_code_prefix       119143 non-null  int64
 10  customer_city                  119143 non-null  object
 11  customer_state                 119143 non-null  object
 12  payment_sequential             119140 non-null  float64
 13  payment_type                   119140 non-null  object
 14  payment_installments           119140 non-null  float64
 15  payment_value                  119140 non-null  float64
 16  review_score                   118146 non-null  float64
 17  review_comment_title           13989 non-null   object
 18  review_comment_message         50245 non-null   object
 19  freight_value                  118310 non-null  float64
 20  product_category_name          116601 non-null  object
 21  product_photos_qty             116601 non-null  float64
 22  seller_zip_code_prefix         118310 non-null  float64
 23  seller_city                    118310 non-null  object
 24  seller_state                   118310 non-null  object
 25  order_purchase_timestamp       119143 non-null  object
 26  order_approved_at              118966 non-null  object
 27  order_delivered_carrier_date   117057 non-null  object
 28  order_delivered_customer_date  115722 non-null  object
 29  order_estimated_delivery_date  119143 non-null  object
 30  review_creation_date           118146 non-null  object
 31  review_answer_timestamp        118146 non-null  object
 32  shipping_limit_date            118310 non-null  object
```

Préparation des données- Nettoyage et validation des données

- Convertir les colonnes qui contiennent des dates en format '***datetime***' :

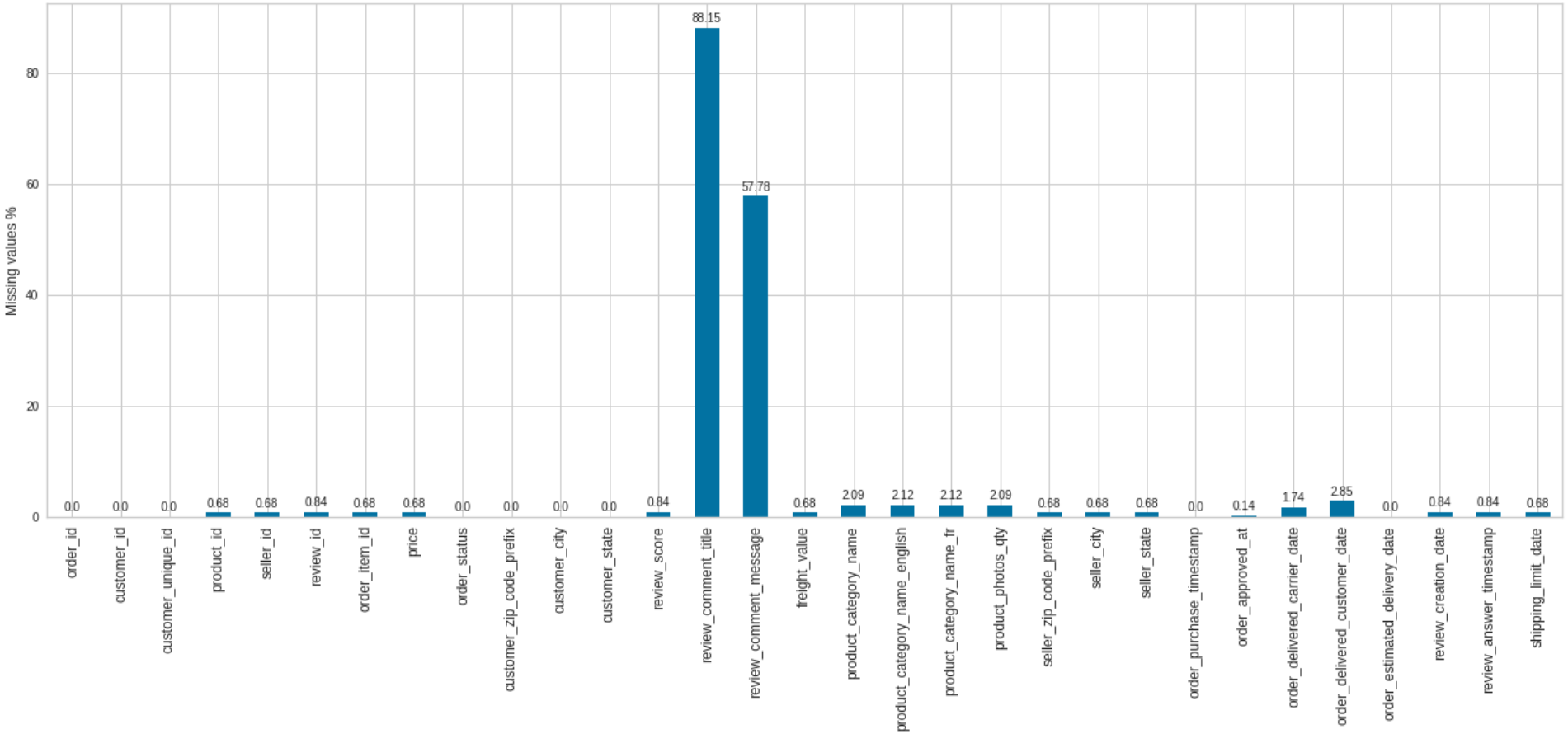
```
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   order_purchase_timestamp       119143 non-null  datetime64[ns]
 1   order_approved_at              118966 non-null  datetime64[ns]
 2   order_delivered_carrier_date   117057 non-null  datetime64[ns]
 3   order_delivered_customer_date  115722 non-null  datetime64[ns]
 4   order_estimated_delivery_date  119143 non-null  datetime64[ns]
 5   review_creation_date           118146 non-null  datetime64[ns]
 6   review_answer_timestamp        118146 non-null  datetime64[ns]
 7   shipping_limit_date            118310 non-null  datetime64[ns]
```

- Ci-dessous Les statistiques descriptives des données:

	order_item_id	price	customer_zip_code_prefix	review_score	freight_value	product_photos_qty	seller_zip_code_prefix
count	113314.00	113314.00	114092.00	113131.00	113314.00	111702.00	113314.00
mean	1.20	120.48	35105.23	4.02	19.98	2.21	24441.67
std	0.71	183.28	29868.30	1.40	15.78	1.72	27597.24
min	1.00	0.85	1003.00	1.00	0.00	1.00	1001.00
25%	1.00	39.90	11250.00	4.00	13.08	1.00	6429.00
50%	1.00	74.90	24320.00	5.00	16.26	1.00	13568.00
75%	1.00	134.90	59022.00	5.00	21.15	3.00	27930.00
max	21.00	6735.00	99990.00	5.00	409.68	20.00	99730.00

Préparation des données- Nettoyage et validation des données

- **Traiter les valeurs manquantes** :
  - Ci-dessous un graphique des valeurs manquantes de notre jeu de données :

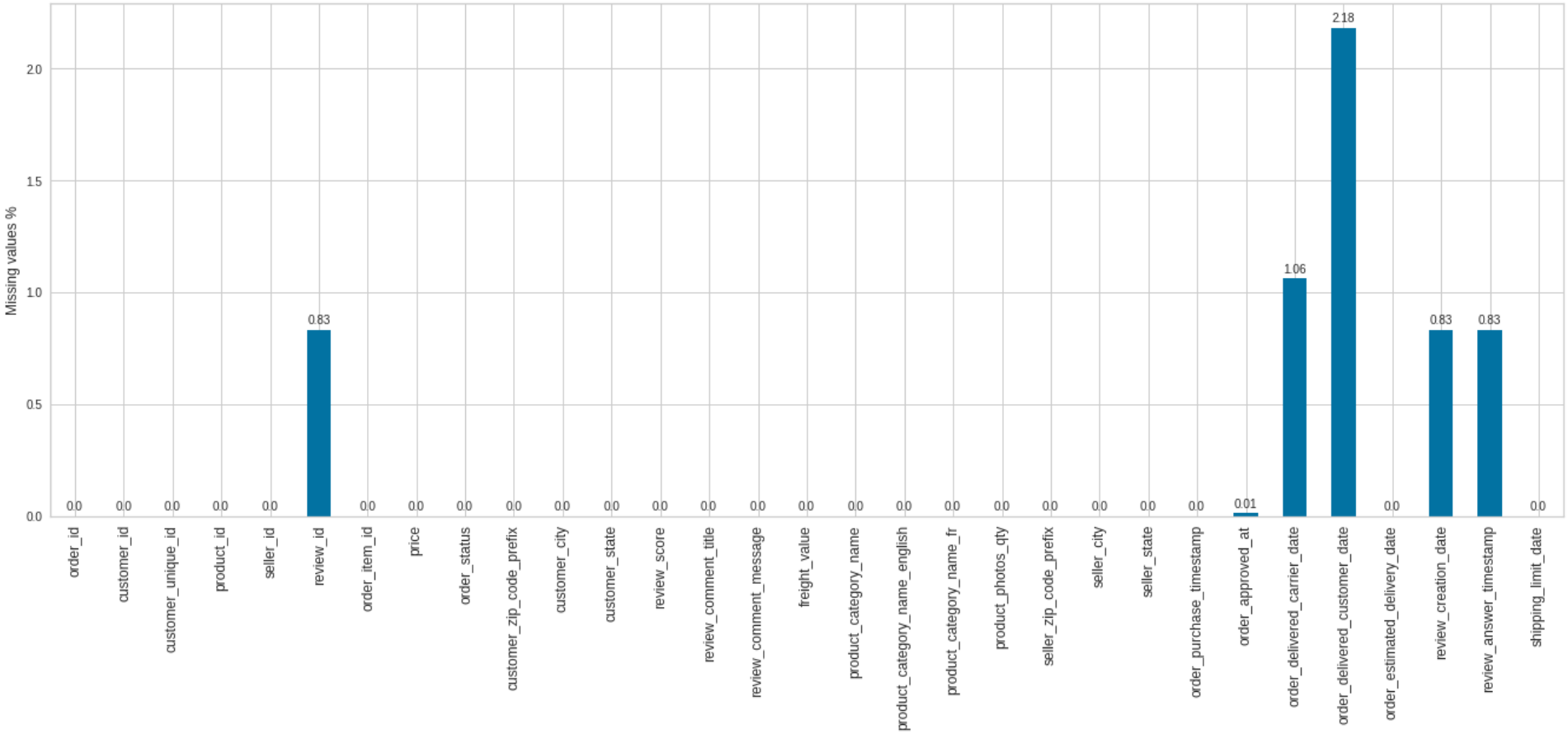


Préparation des données- Nettoyage et validation des données

- Remplir les valeurs manquantes des colonnes *'**review_comment_title***' et '***review_comment_message***‘ par '*Aucun titre de commentaire*‘ et *‘Aucun commentaire’*
- Remplir les valeurs manquantes de la colonne '***product_photos_qty***' par 0
- Remplir les valeurs manquantes de la colonne ***'product_category_name***‘ par '*Aucune categorie*‘
- Remplir les valeurs manquantes de la colonne ***'review_score***' par la moyenne de cette colonne.
- Supprimer les lignes avec le prix qui n’est pas renseigné (environ 0,7% des données)

Préparation des données- Nettoyage et validation des données

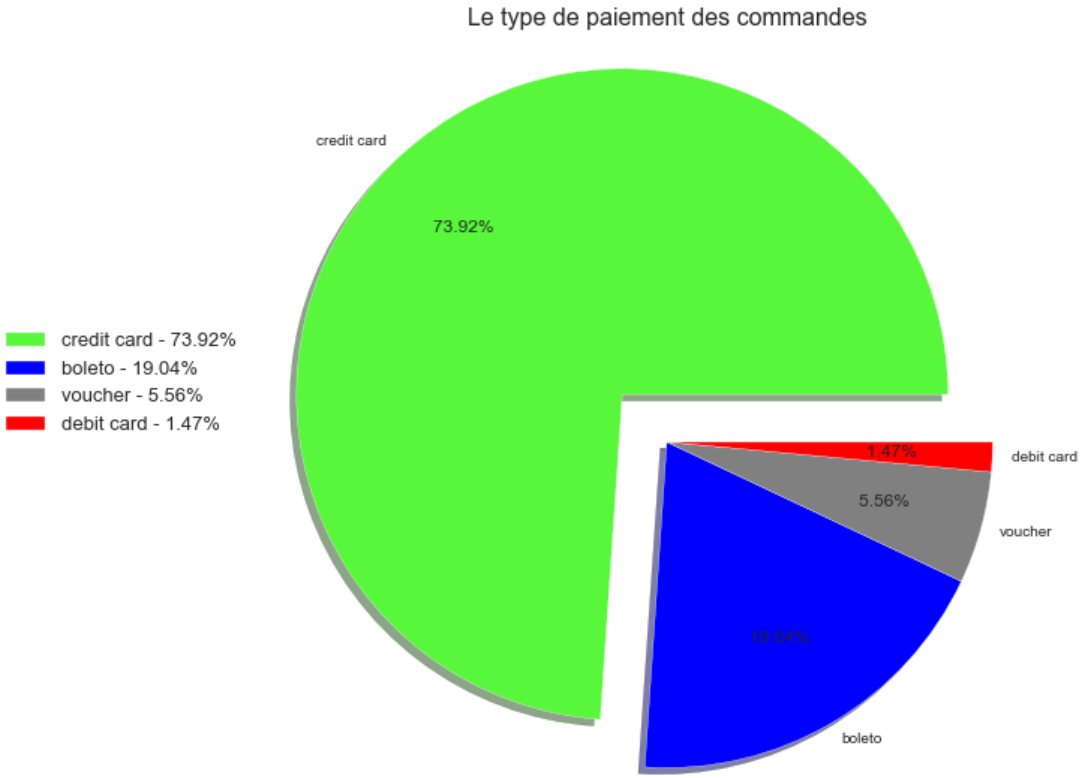
- Ci-dessous un graphique des valeurs manquantes de notre jeu de données après nettoyage :



Analyse exploratoire des données

Analyse exploratoire des données - Type de paiement

- Graphique de distribution par type de paiement :



- **Remarque :** La majorité des commandes (environ 74%) ont été payées par carte de crédit, et 5,5% avec un bon d'achat.

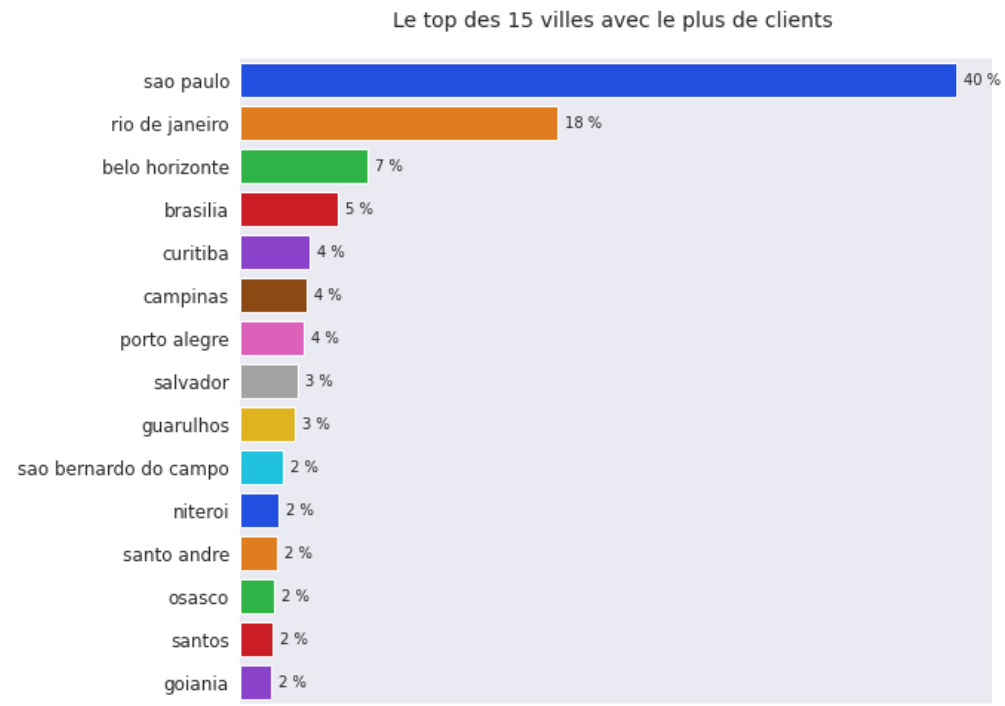
Analyse exploratoire des données - Statut des commandes

- Graphique du statut des commandes :

- **Remarque :** 97% des commandes ont été livrées, et 0,6% annulées.

Analyse exploratoire des données - Top 15 villes / états

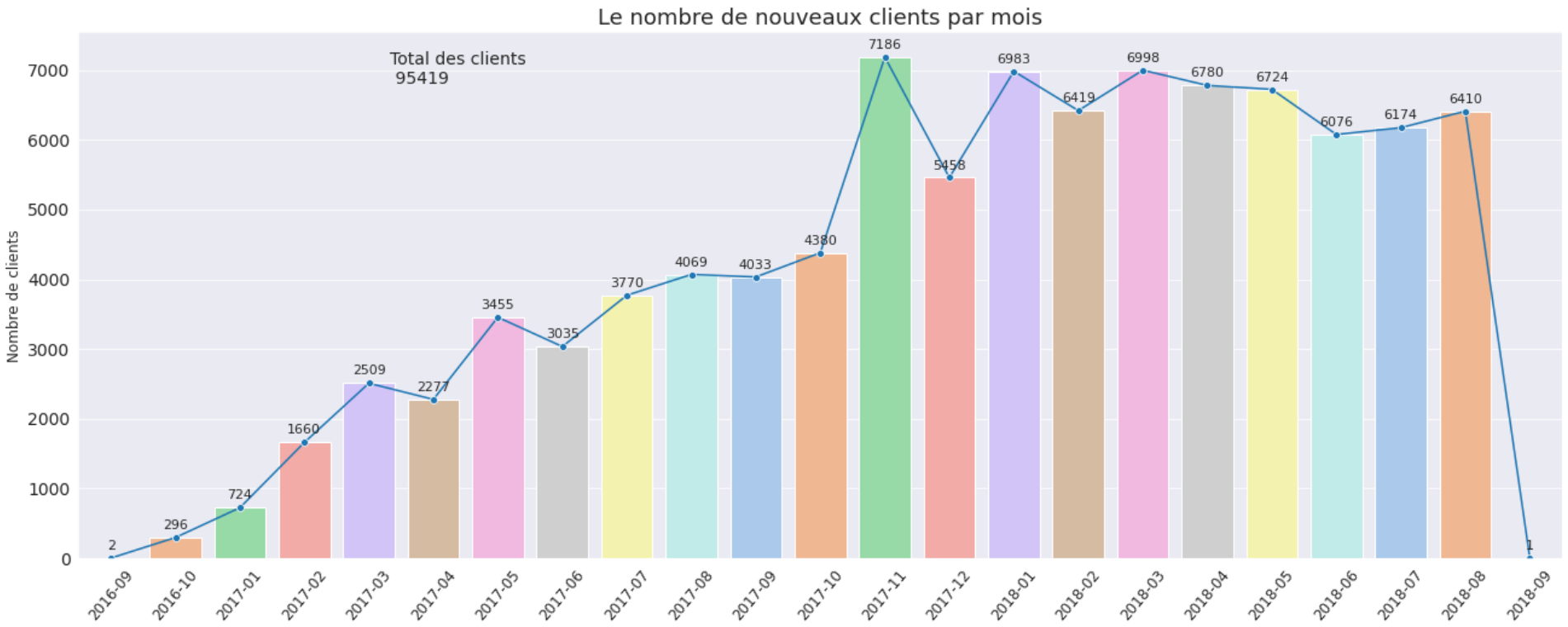
- Le top des 15 villes et états avec le plus grand nombre de clients :



- **Remarque :** Sao Paulo est à la tête des ville et état avec le plus grand nombre de clients.

Analyse exploratoire des données - Nouveaux Clients (par mois)

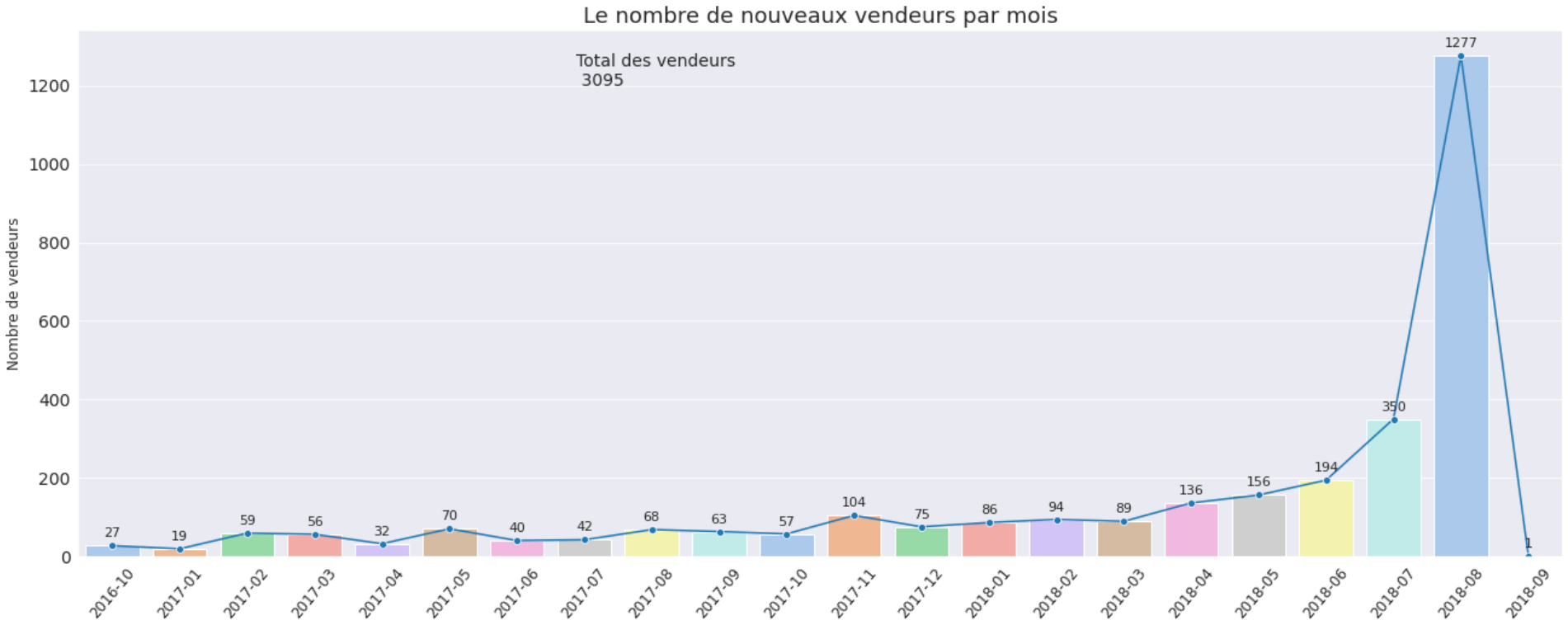
- Le nombre de nouveaux clients par mois :



- **Remarque :** En novembre 2017 « Olist » a enregistré 7186 nouveaux clients, et 6983 en janvier 2018.

Analyse exploratoire des données - Nouveaux vendeurs (par mois)

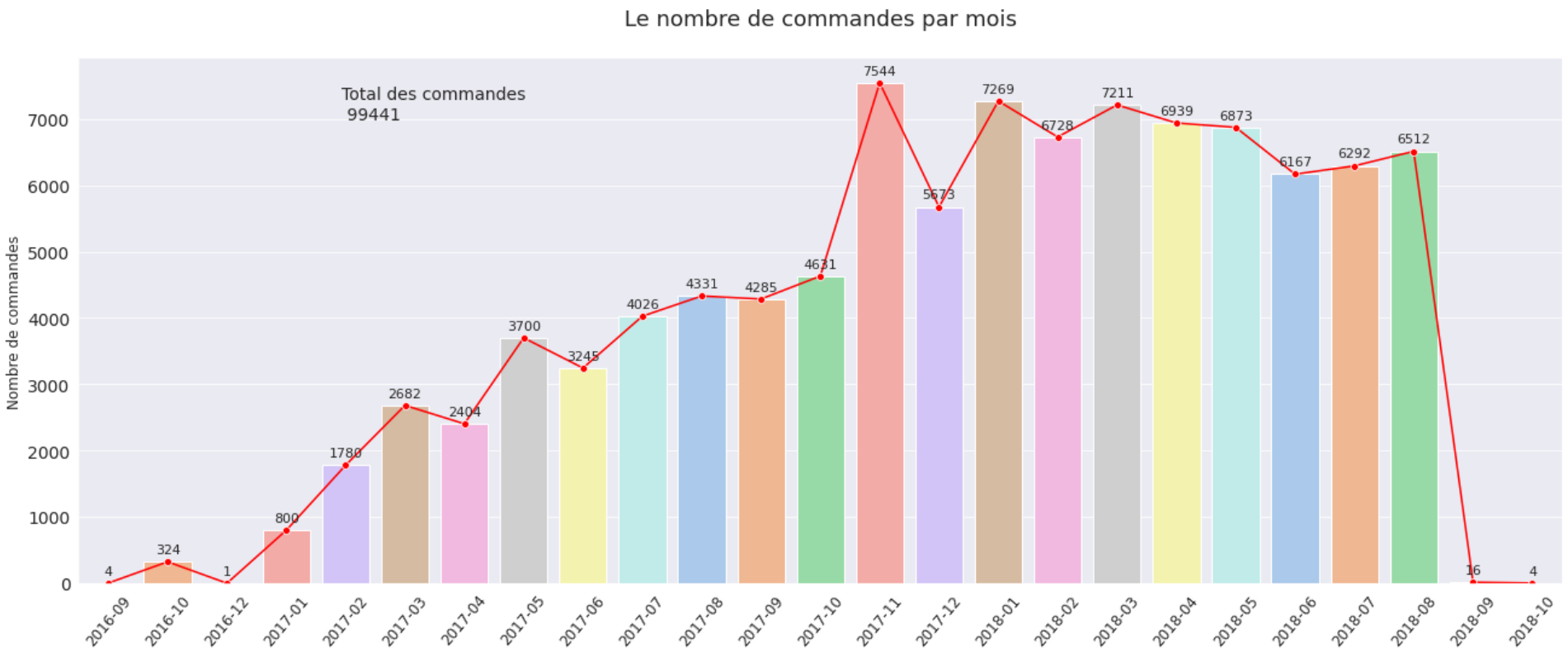
- Le nombre de nouveaux vendeurs par mois :



- **Remarque :** Le nombre de nouveaux vendeurs ne dépassaient pas les 104 vendeurs par mois pour l'année 2017, or ce nombre a atteint 1277 vendeurs seulement pour la période du mois d'août 2018.

Analyse exploratoire des données - Nombre de commandes (par mois)

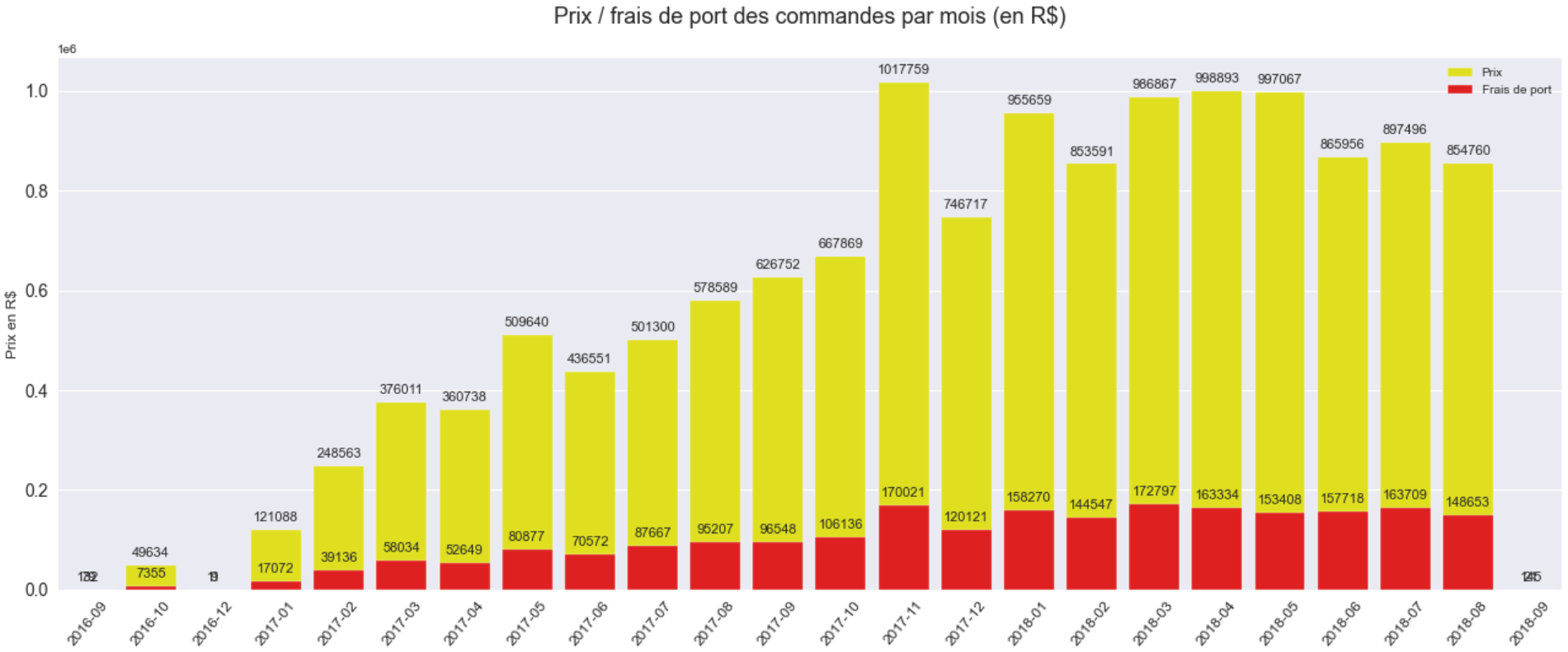
- Le nombre de commandes par mois :



- **Remarque :** On constate que le nombre de commandes a attient 7544 pour le mois de novembre 2017 suivi de 7269 pour le mois janvier 2018.

Analyse exploratoire des données – Prix / Frais de port

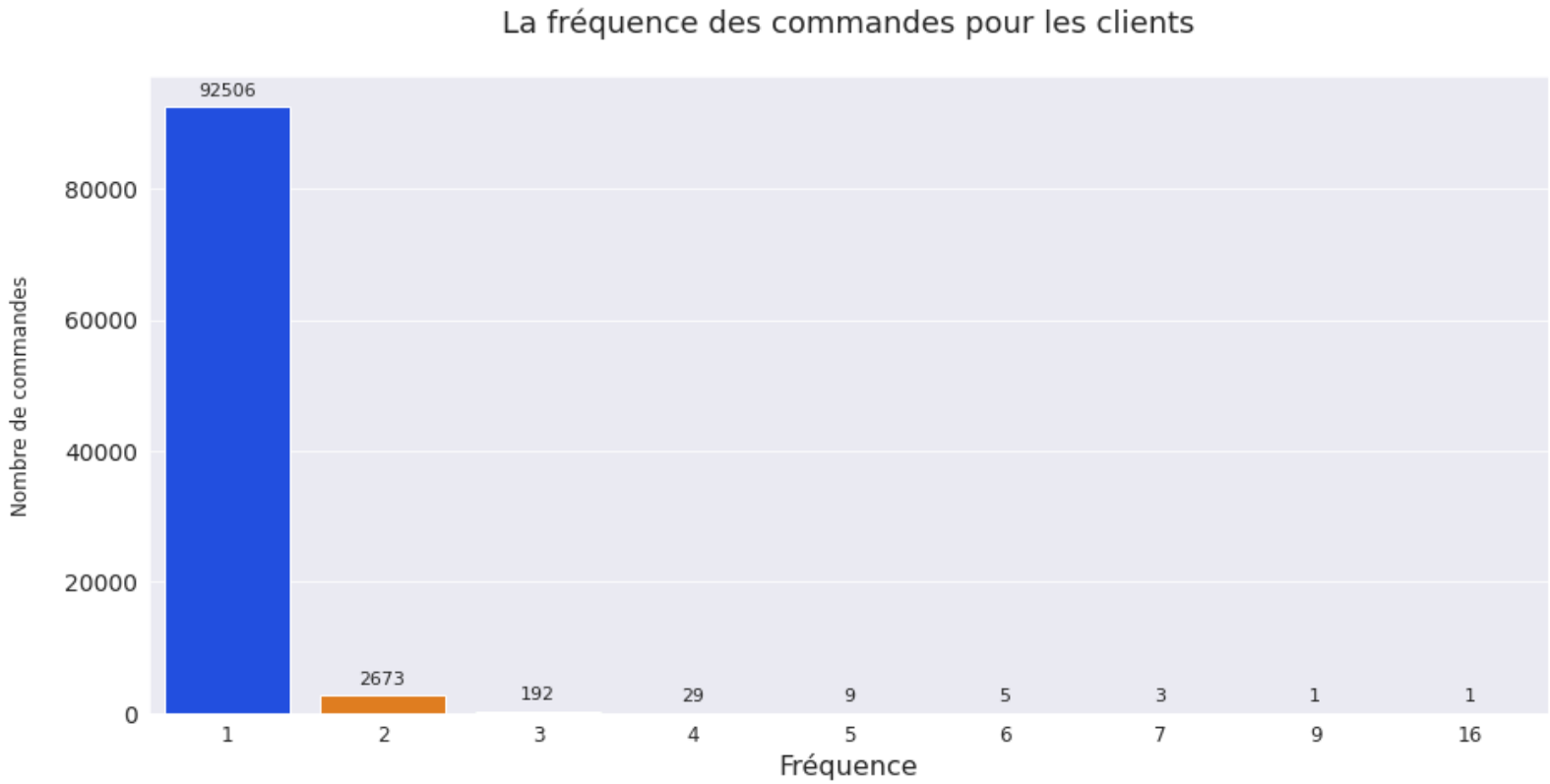
- Prix / frais de port des commandes par mois (en R$) :



- **Remarque :** Les frais de port représentent environ 15 à 20% du prix de la commande

Analyse exploratoire des données – Fréquence des commandes

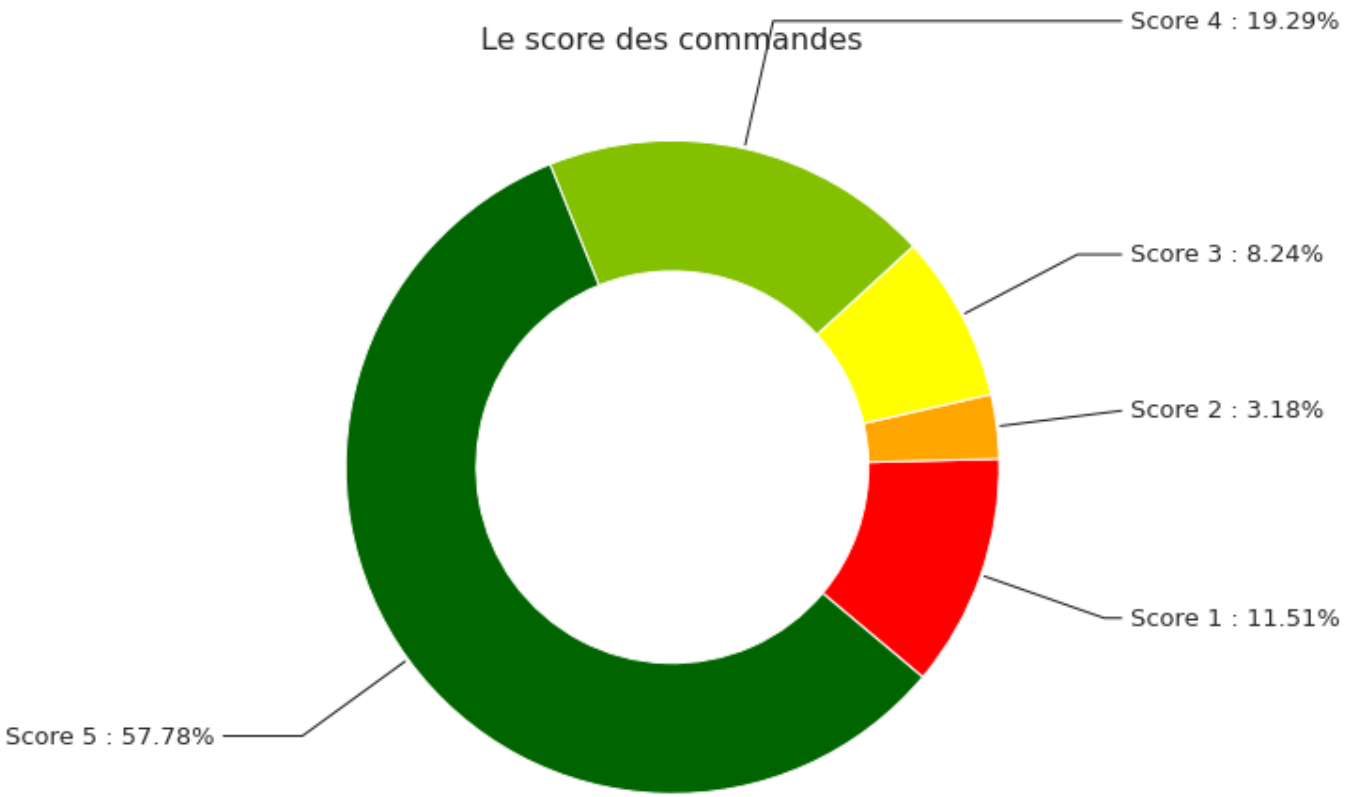
- La fréquence des commandes pour les clients



- **Remarque :** Une majorité des clients (environ 97%) n'ont effectués qu'une seul commande entre 09/2016 et 09/2018.

Analyse exploratoire des données - Score des commandes

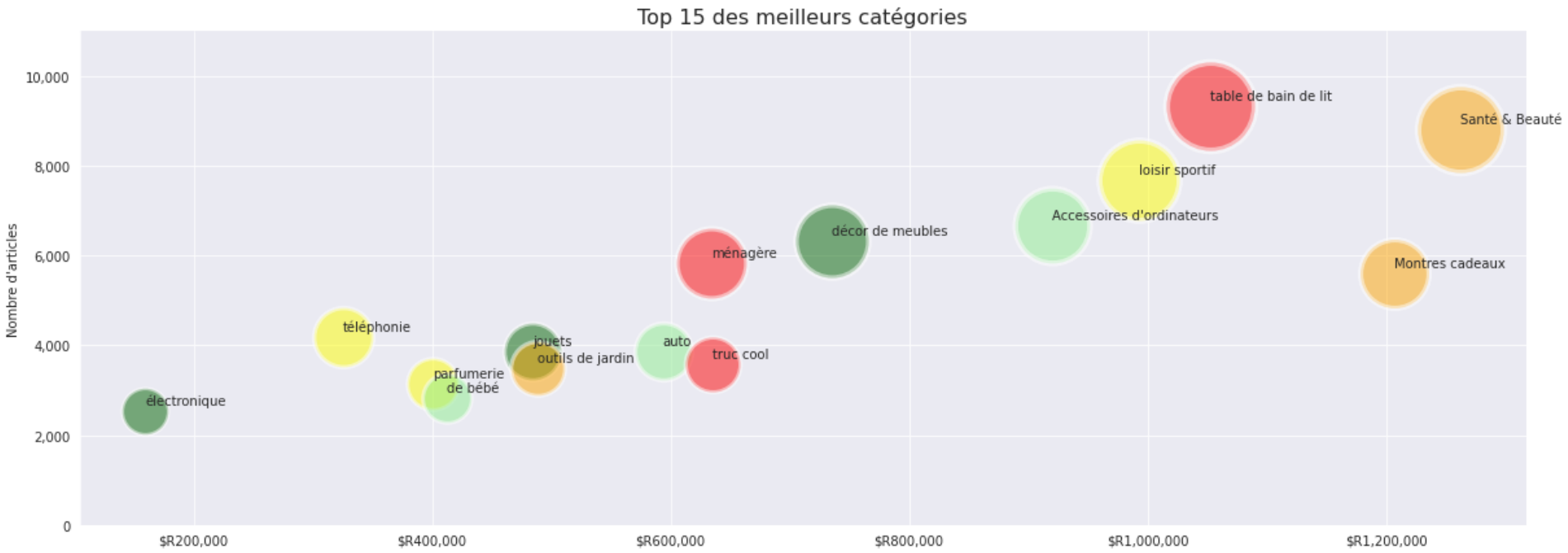
- Graphique des scores des commandes :



- **Remarque :** Nous avons 77% des commandes ont un score 4 et 5.

Analyse exploratoire des données - Catégorie

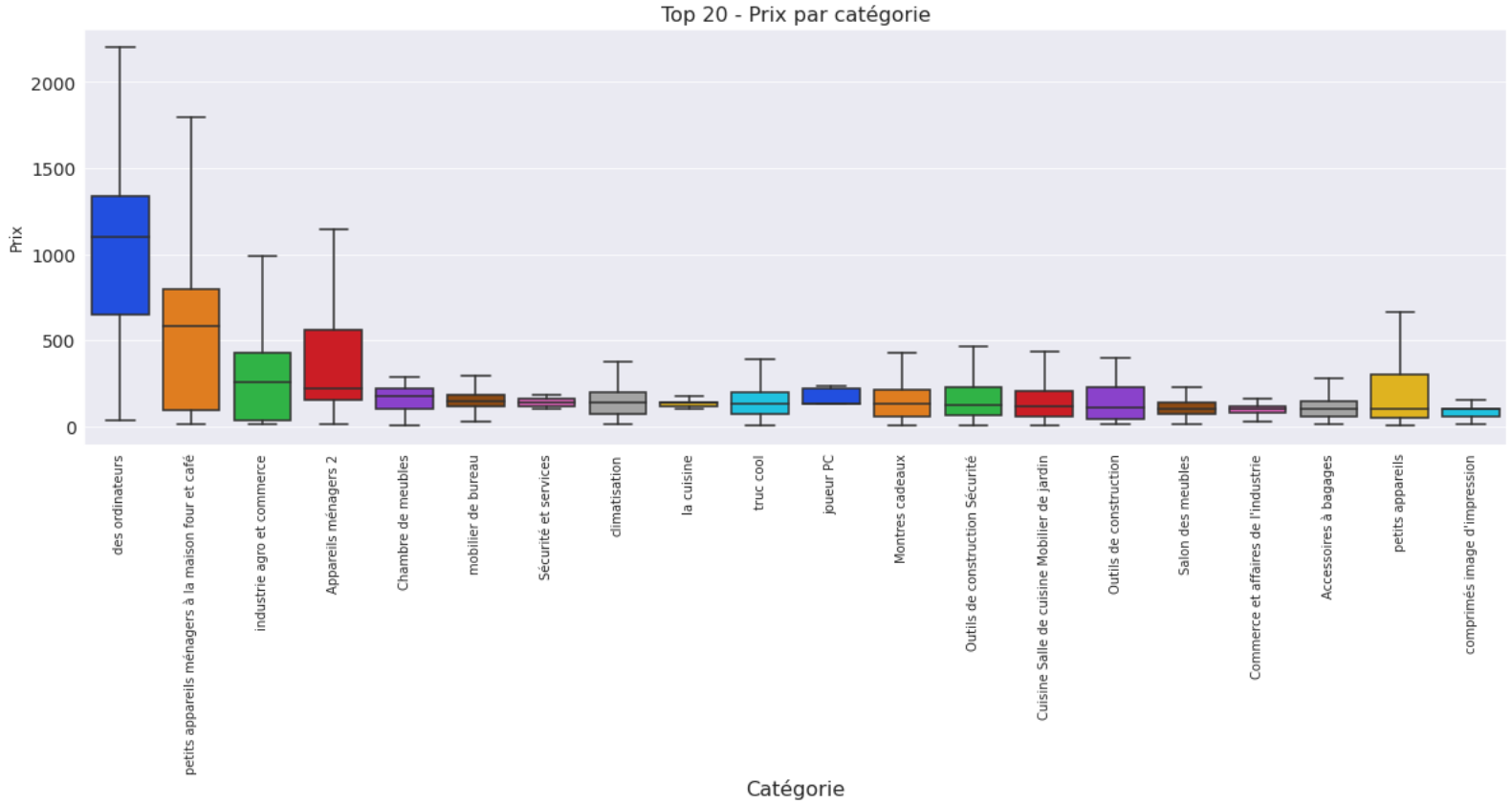
- Graphique des 15 meilleurs ventes selon la catégorie :



- **<u>Remarque :</u>** Les catégories santé & beauté et ameublement font parti des meilleurs ventes pour ‘olist’.

Analyse exploratoire des données - Catégorie

- Boxplot des top 20 des catégories par prix :



- **Remarque** : La catégorie avec un prix médian élevé est la catégorie «des ordinateurs».

Segmentation RFM

Segmentation RFM - Définition

- **Définition** :

  La segmentation RFM est un outil marketing permettant de regrouper les clients en fonction de caractéristiques communes afin de concentrer et commercialiser efficacement auprès de chaque groupe et maximiser la valeur de chaque client pour l'entreprise :

  - ✓ **Récence** : date du dernier achat ou dernier contact client
  - ✓ **Fréquence** : fréquence des achats sur une période de référence donnée
  - ✓ **Montant** : somme des achats cumulés sur cette période

- La segmentation de la clientèle a au moins deux objectifs principaux :

  1. Continuez à fournir le meilleur service aux meilleurs clients.
  2. Se focaliser sur les clients potentiels qui ont le même profil que les meilleurs clients.

Segmentation RFM - Calcul du RFM

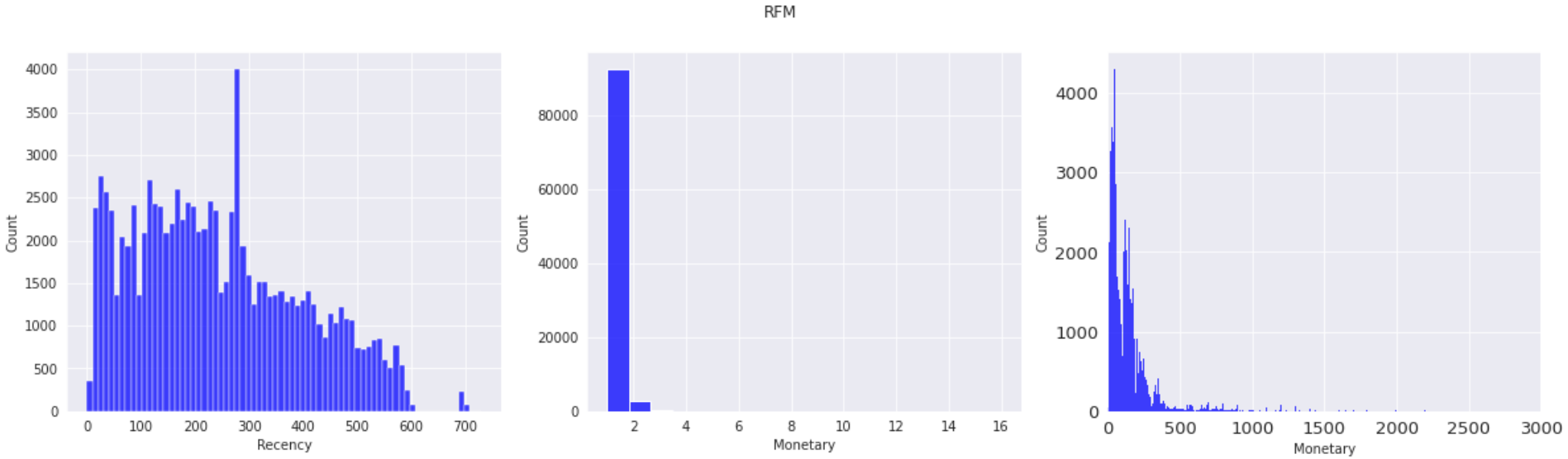
- Méthodologie du calcul RFM :
  - Calculer les trois facteurs (RFM) en regroupant les données par client, et en prenant La date du dernier achat qui correspond à *2018-09-03*.
  - Calculer le score RFM en divisant les clients en quatre segments (25% pour chaque segment), et attribuer des notes de 1 à 4 (du meilleur au pire) à chaque segment.

- La matrice de notation des segments « RFM » est la suivante:

Segments / Critères	S1	S2	S3	S4
Récence	< 119	119-224	224-353	> 353
Fréquence	> 1			<= 1
Monétaire	> 155.99 R$	89.9-155.99 R$	47.9-89.9 R$	< 47.9 R$
Note du segment	**1**	**2**	**3**	**4**

Segmentation RFM - Calcul du RFM

- Graphique de la distribution RFM :



- Remarque : On constate que la récence se situe entre 1 et 600 jours, la majorité des clients (environ 87000) ont commandé une seul fois.

Analyse en composantes principales « ACP »

ACP – Définition / Objectif

- **Définition** :

  Analyse en composantes principales « ACP » permet de transformer un ensemble de variables en un ensemble réduit de nouvelles variables, combinaisons linéaires des variables initiales. Les variables associées au coefficient le plus fort sont les plus informatives pour le clustering.

- **Objectif**:

  Le but d'une analyse en composantes principales est de trouver une nouvelle base orthonormée dans laquelle représenter nos données, telle que la variance des données selon ces nouveaux axes soit maximisée.

ACP – Calculer les composantes principales

- Les étapes effectuées pour calculer l'ACP :

  ✓Fusionner la segmentation RFM avec les données initiales.

  ✓Détecter et supprimer les valeurs aberrantes en utlisant l'algorithme « ***Isolation Forest*** »

  ✓Transformer les données en array numpy

  ✓Standardiser les données

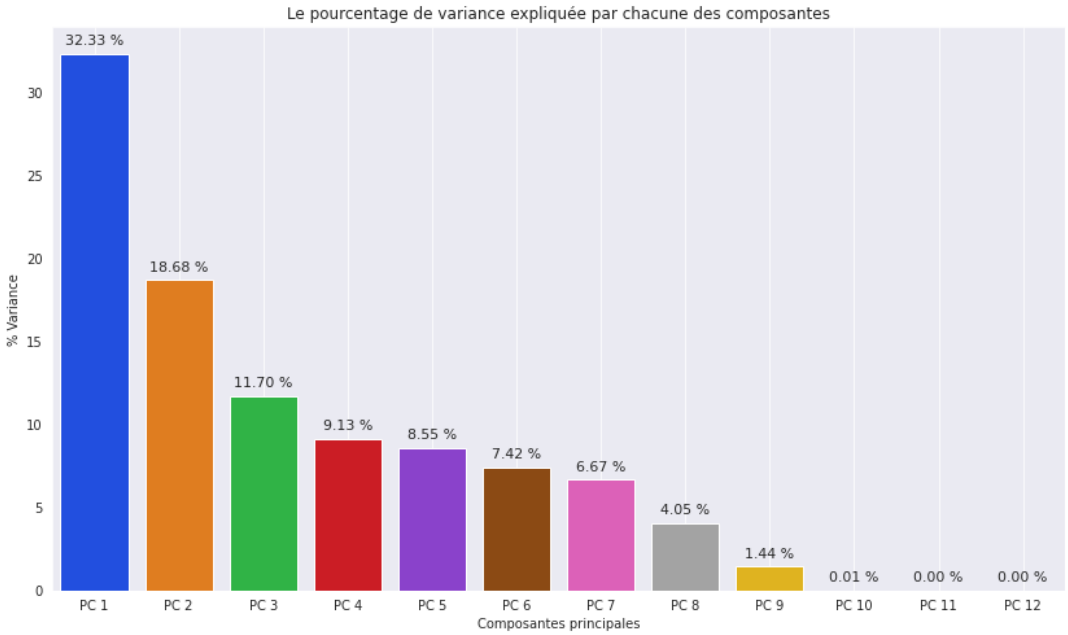
  ✓Calculer les composantes principales

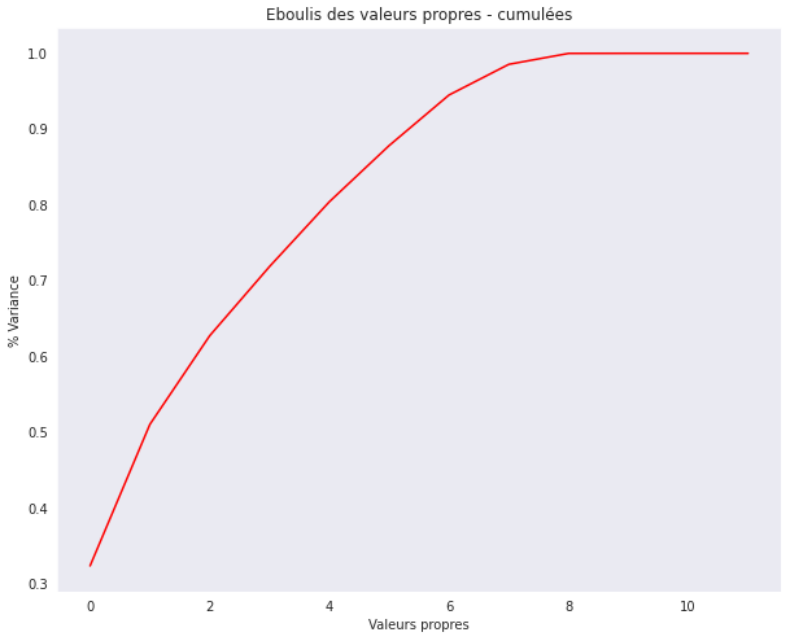
- Les colonnes de la base de données utilisées pour l'ACP :

  'price', 'review_score' ,'nb_order', 'freight_value', 'Recency', 'Frequency', 'Monetary', 'RFMScore', 'total_payment', 'nb_item', 'payment_sequential', 'nb_payment_installments'

ACP – Calculer les composantes principales

- Le pourcentage de variance expliquée par chacune des composantes:





- **Remarque :**

  On constate si on se fixe une proportion de variance expliquée de 94%, on peut se contenter de 7 composantes principales.

Segmentation K-Means

Segmentation K-Means - Définition

- **Définition :**

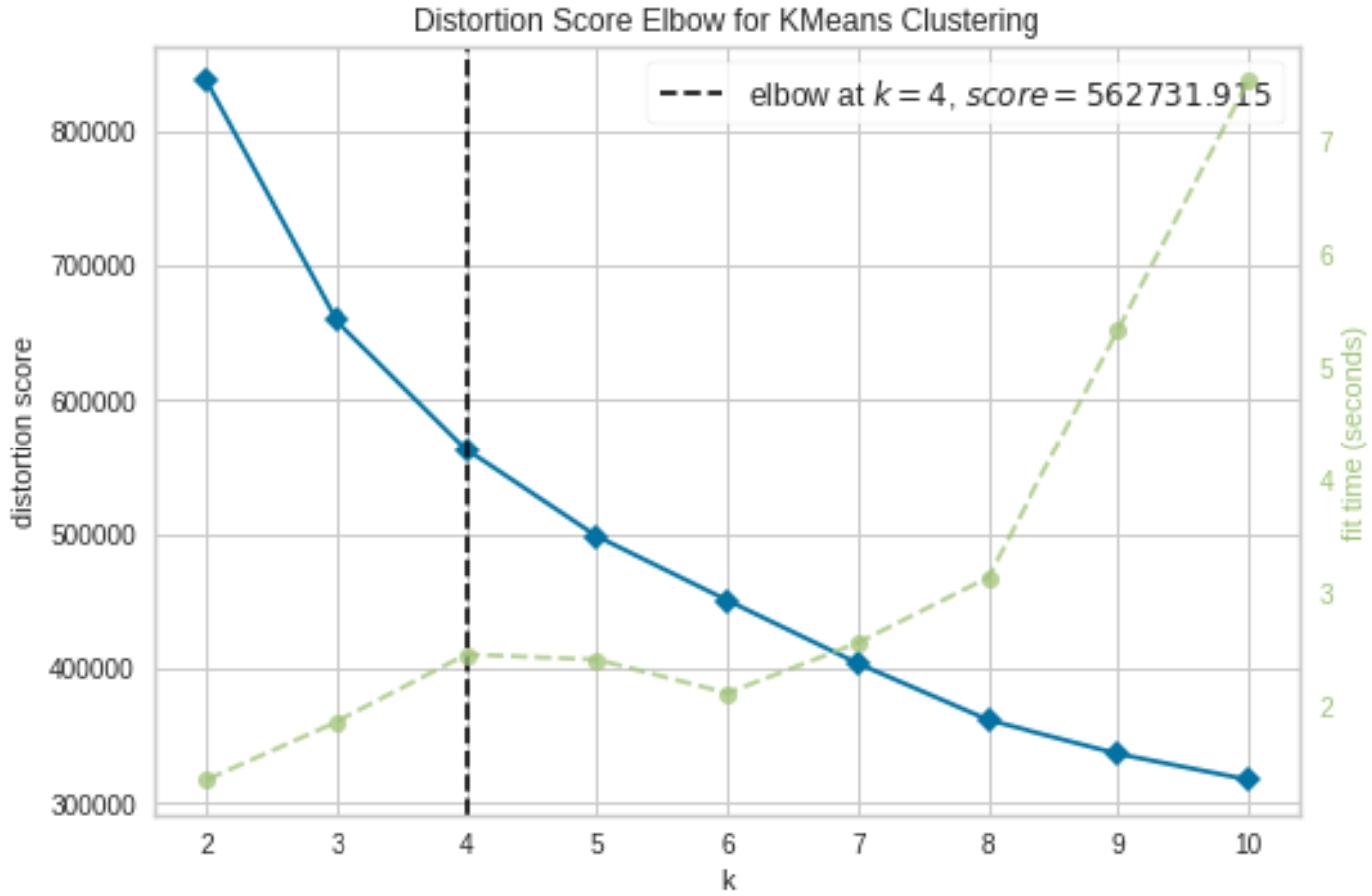
  K-means est un algorithme non supervisé de clustering non hiérarchique. Il permet de regrouper en K clusters distincts les observations du data set. Ainsi les données similaires se retrouveront dans un même cluster. Par ailleurs, une observation ne peut se retrouver que dans un cluster à la fois (exclusivité d'appartenance). Une même observation, ne pourra donc appartenir à deux clusters différents.

- **Le clustering :**

  Le clustering est une méthode d'apprentissage non supervisé. Ainsi, on n'essaie pas d'apprendre une relation de corrélation entre un ensemble de caractéristiques d'une observation et une valeur à prédire, comme c'est le cas pour l'apprentissage supervisé.

Segmentation K-Means – Définir le nombre de clusters

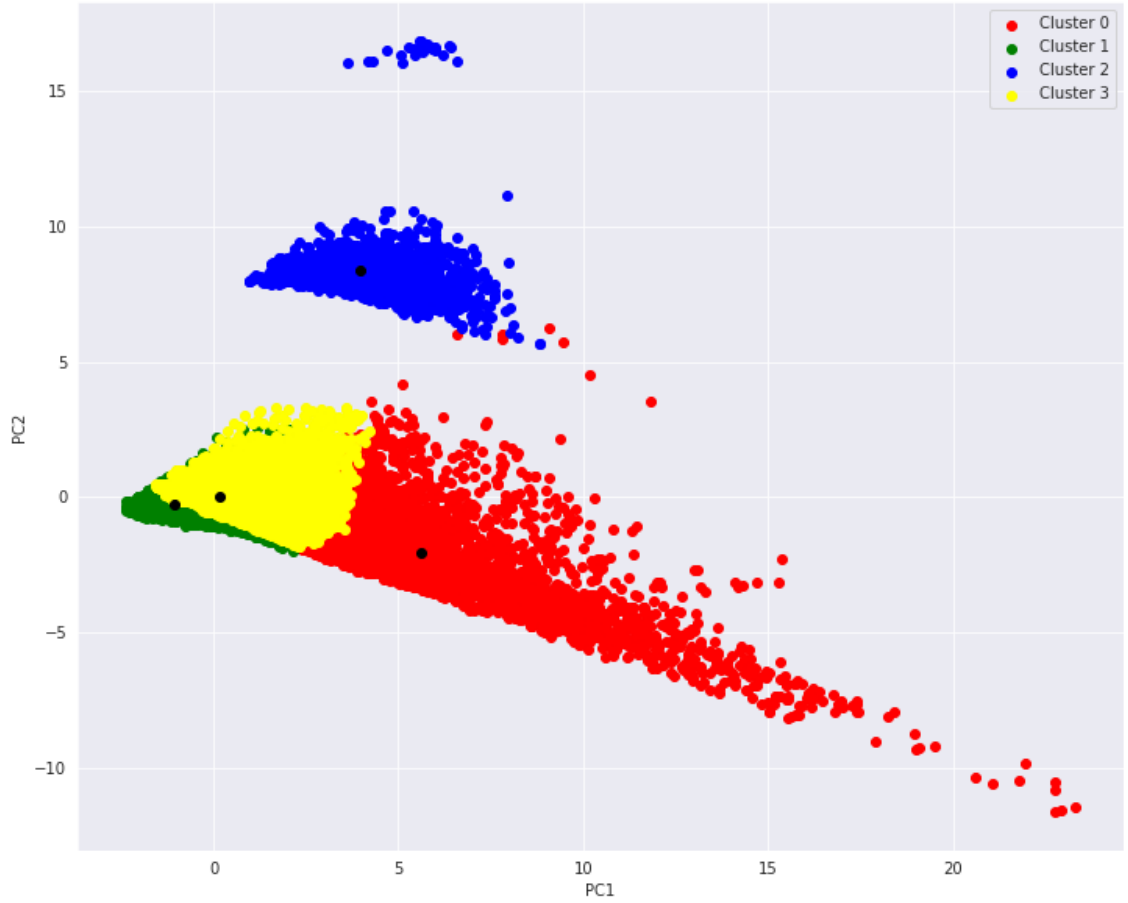
- Ci-dessous le graphique de la méthode « Elbow » :



- **Remarque :** Le nombre optimal de clusters pour les données est 4.

Segmentation K-Means – Analyse des clusters

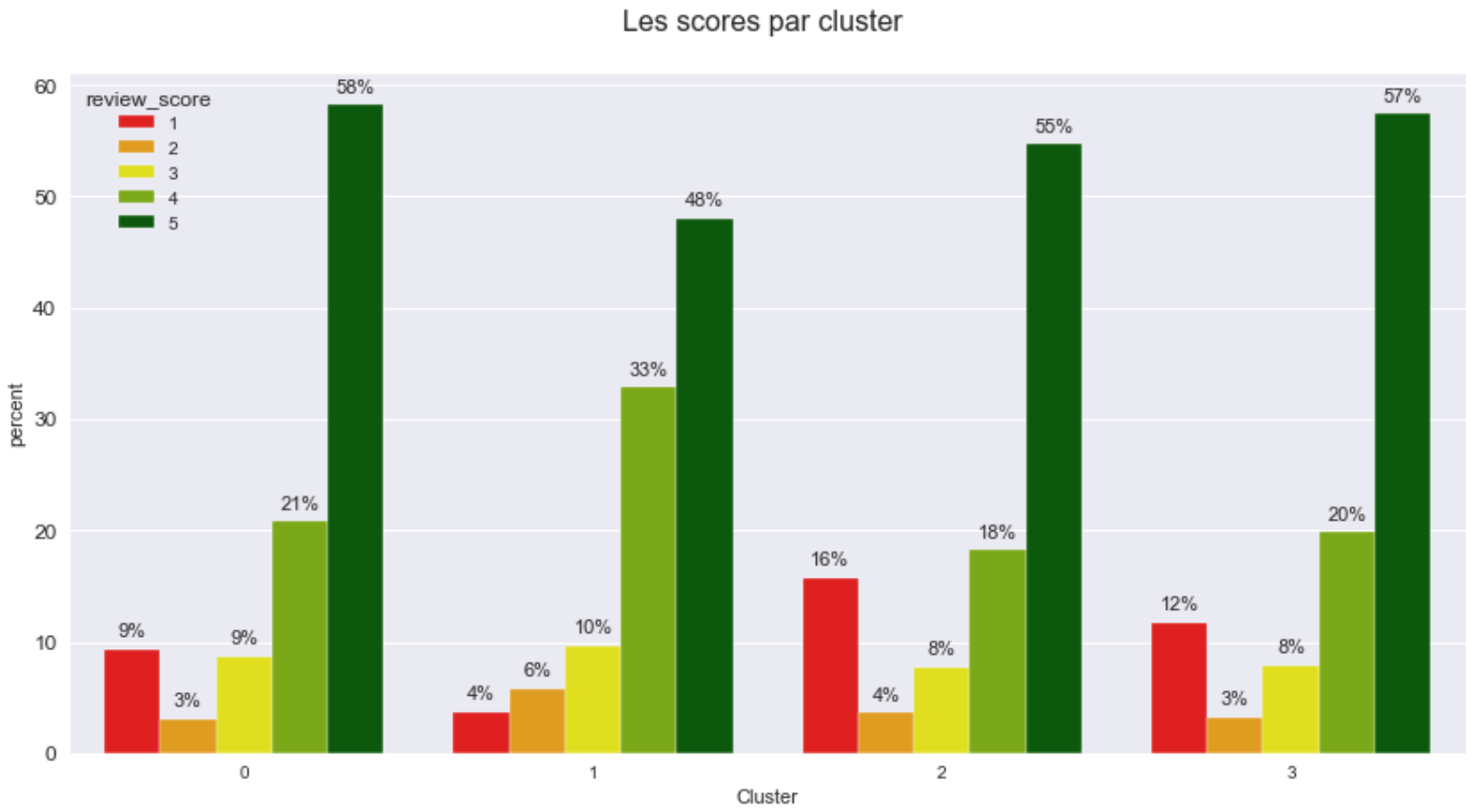
- Ci-dessous le graphique des clusters et des centroïdes:



- **Remarque :** Le cluster 2 est séparé des autres clusters

Segmentation K-Means – Analyse des clusters

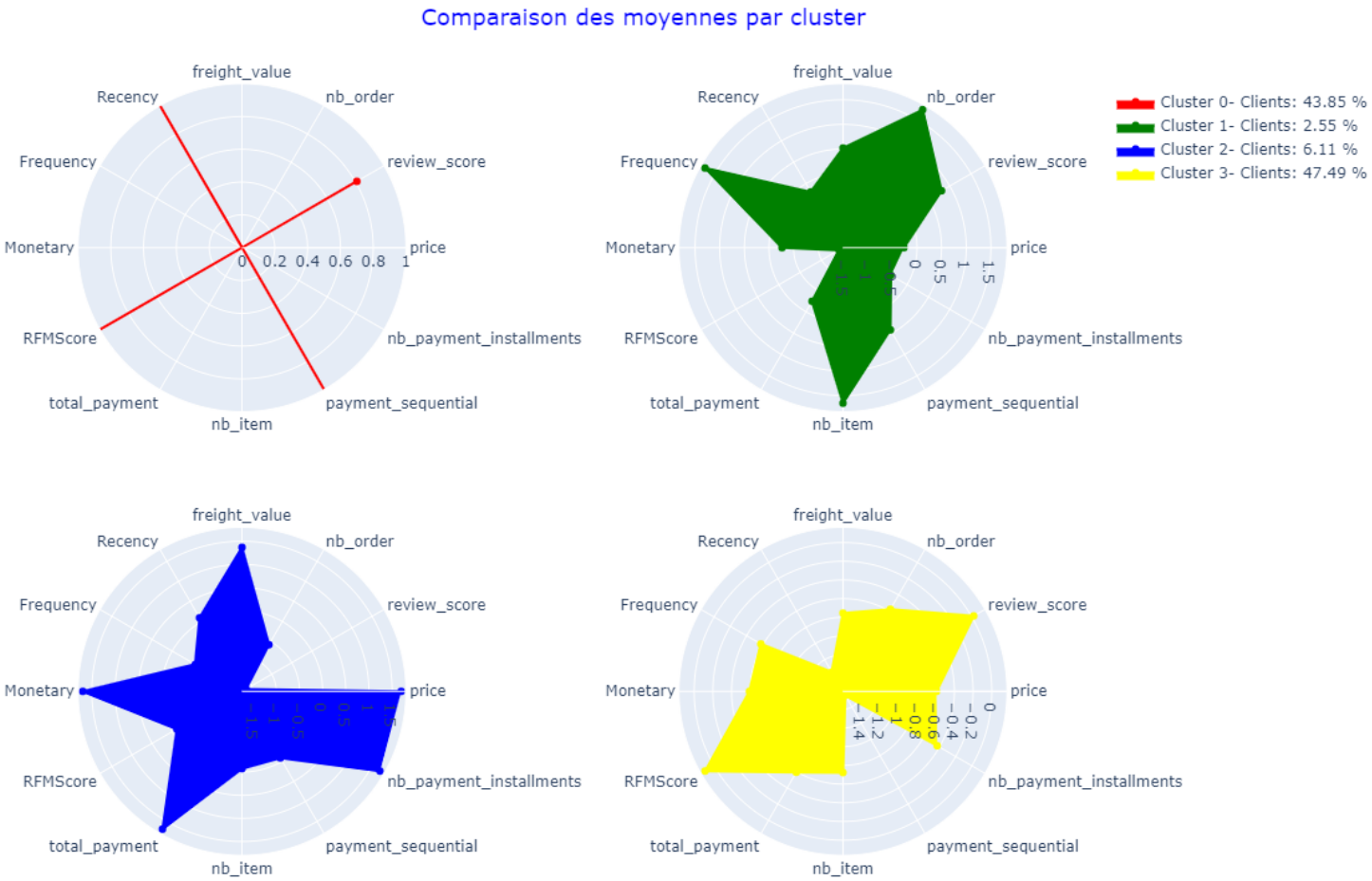
- Les scores par cluster:



- **Remarque :** Le cluster 2 a 18% d'avis négatif.

Segmentation K-Means – Analyse des clusters

- Graphique ***Radar*** pour visualiser la différence entre les clusters :



Segmentation K-Means – Analyse des clusters

- **Résultat de l'analyse :**

  Après avoir effectué notre premier clustering K-Means, nous avons 4 segments:

  - Cluster 0 – Classification « ***Bronze*** » : Ce segment de clients (43% des clients) a le montant total de dépenses le plus faible avec une récence élevée, un score de satisfaction au dessus de la moyenne et une utilisation de plusieurs modes paiement.

  - Cluster 1 – Classification « ***Silver***» : C'est le plus petit segment (environ 3%) il est composé de clients avec un nombre élevé de commandes et d'articles, une dépense dans la moyenne et une satisfaction légèrement au dessus de la moyenne.

  - Cluster 2 – Classification « ***Gold***» : Ce petit segment (6% des clients) a le montant total de dépenses et frais de port le plus élevé, un score de satisfaction faible, et des paiements en plusieurs échéances.

  - Cluster 3 – Classification « ***Platinum*** » : C'est le segment qui contient le plus de clients (environ 48%), avec un montant de dépenses au dessus de la moyenne, des clients satisfait avec un score élevé, Un nombre de commandes au dessus de la moyenne.

Segmentation K-Means – Analyse des clusters « outliers »

- Les clients « outliers » ont une importance pour l'équipe marketing, pour cette raison il faut les classer dans les clusters précédemment générés.

- La segmentation K-Means des clients « outliers » a générée 4 clusters :
  - Cluster 0 – Classification « ***Bronze*** » : contient 453 clients.
  - Cluster 1 – Classification « ***Silver***» : contient 12 clients.
  - Cluster 2 – Classification « ***Gold***» : contient 42 clients.
  - Cluster 3 – Classification « ***Platinum*** » : contient 448 clients.

Analyse de stabilité

Analyse de stabilité – Etude de stabilité

- Etude de la stabilité de la segmentation:

  L'équipe marketing souhaiterait une recommandation de fréquence à laquelle la segmentation doit être mise à jour pour rester pertinente, afin de pouvoir effectuer un devis de contrat de maintenance.

- La notion de stabilité est définie par deux critères :

  - L'évolution du pourcentage d'appartenance à un cluster. On recommence la même opération aux mois suivants avec d'autres ensembles de données.

  - L'évolution de la répartition des classes ou des valeurs cibles au sein des clusters.

Analyse de stabilité - Méthodologie

- La méthodologie effectuer pour l'analyse de stabilité est la suivante :

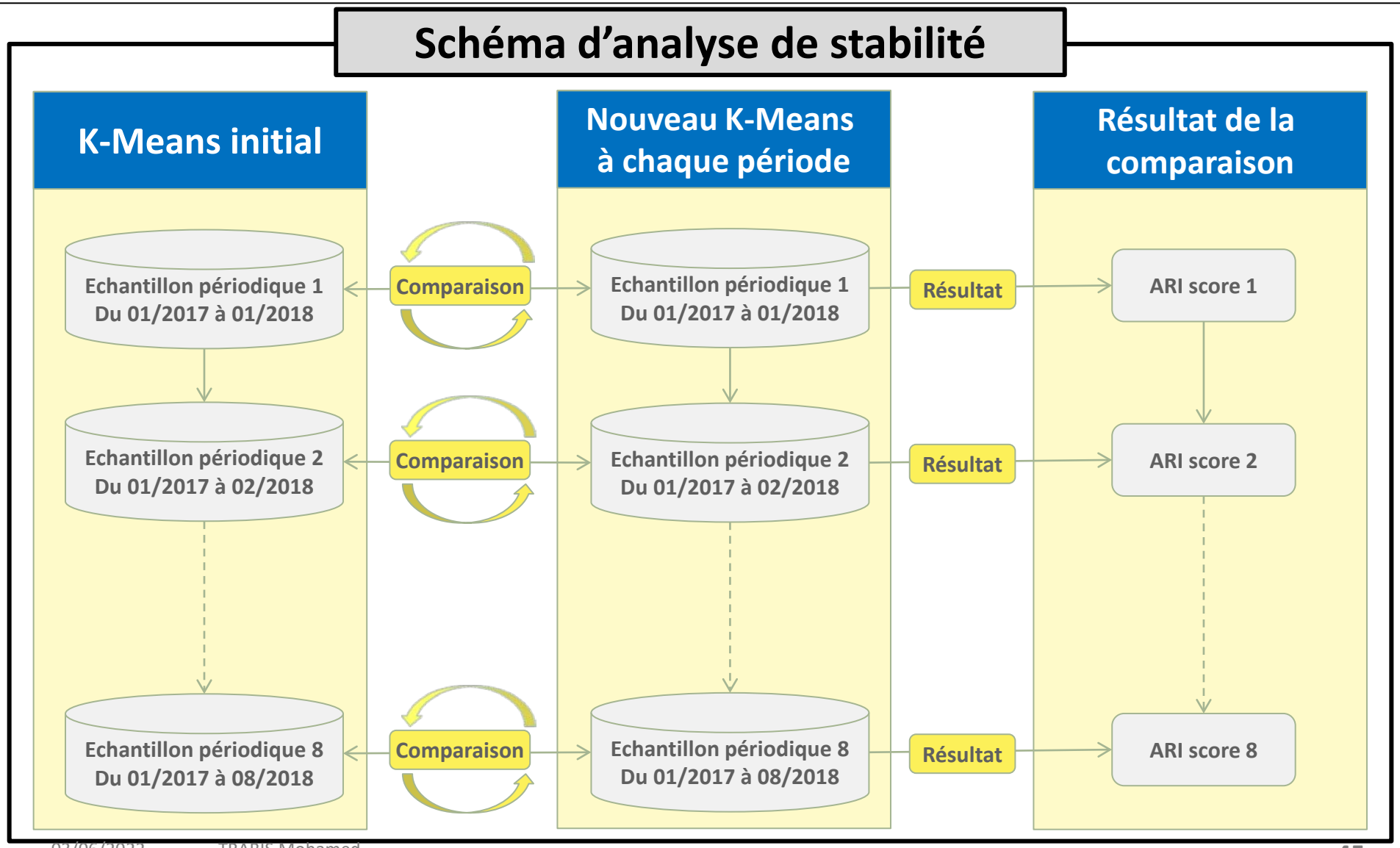
  ✓Définir les clients sur toute la période de janvier 2017 à août 2018 avec un delta d'un mois (le premier échantillon du janvier 2017 à janvier 2018).

  ✓Effectuer une prédiction pour chaque échantillon avec le K-Means initial.

  ✓Effectuer un nouveau K-Means sur chaque échantillon.

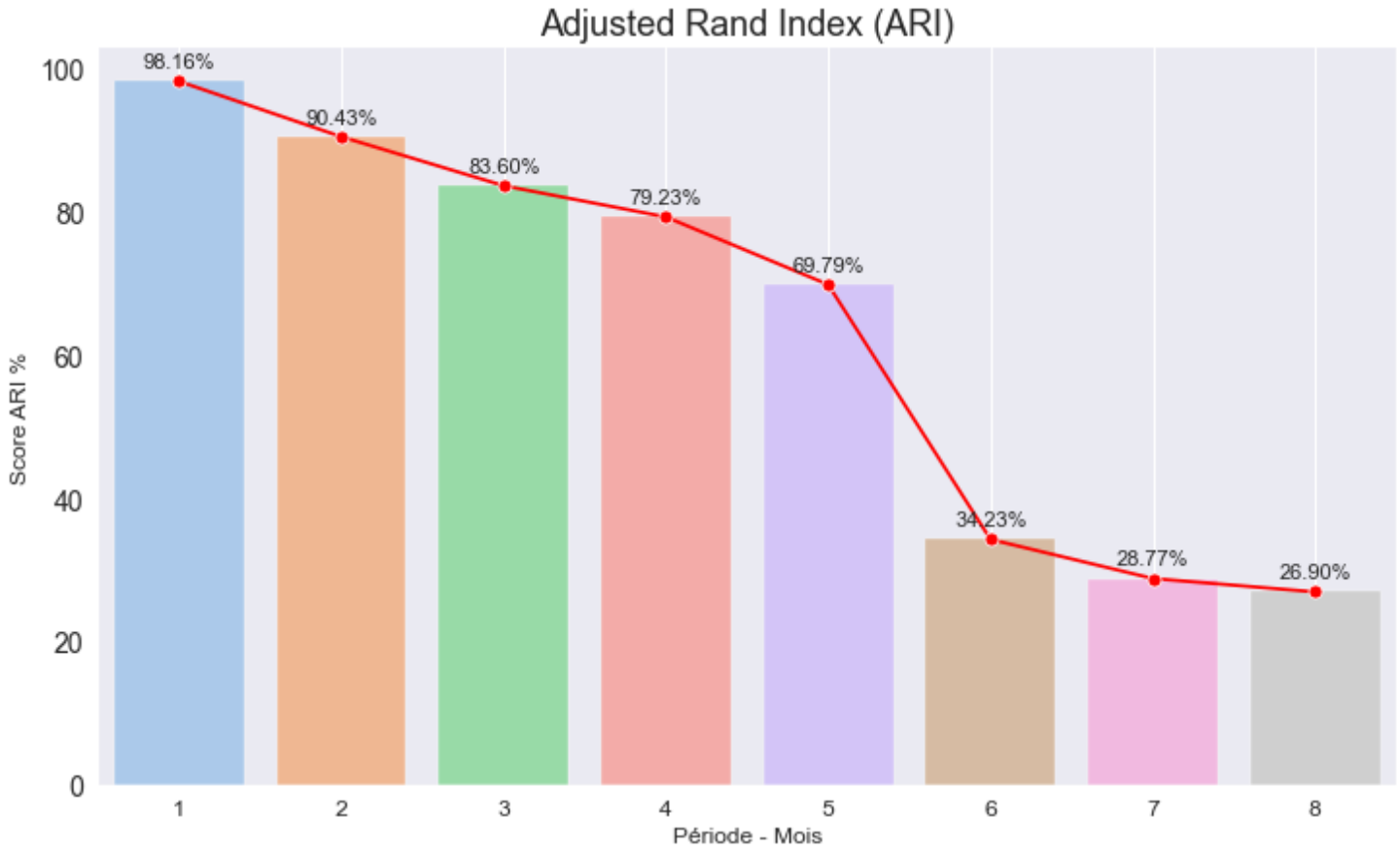
  ✓Comparer le clustering des clients du nouveau K-Means avec le clustering des clients du K-Means initial, en utilisant l'indice Rand « ***ARI* -Adjusted Rand Index**» pour mesurer la similarité entre les clusters.

Analyse de stabilité - Schéma



Analyse de stabilité - Score ARI

- Graphique de l'évolution du score « ***ARI*** » :



- **Remarque:** On constate une baisse significative du score ARI après 5 mois.

Analyse de stabilité - Conclusions

- **<u>Conclusions</u> :**

Pour que la segmentation des clients « ***Olist*** » reste pertinente il faut :

- ✓ Prévoir une maintenance de l'algorithme de segmentation tous les cinq mois.
- ✓ Effectuer une nouvelle segmentation à chaque maintenance.
- ✓ Analyser et tester la stabilité de la nouvelle segmentation.