Segmentez des clients d’un site e-commerce

olist

Meltem YILDIRIM

16/08/2021

Sommaire

I. Problématique

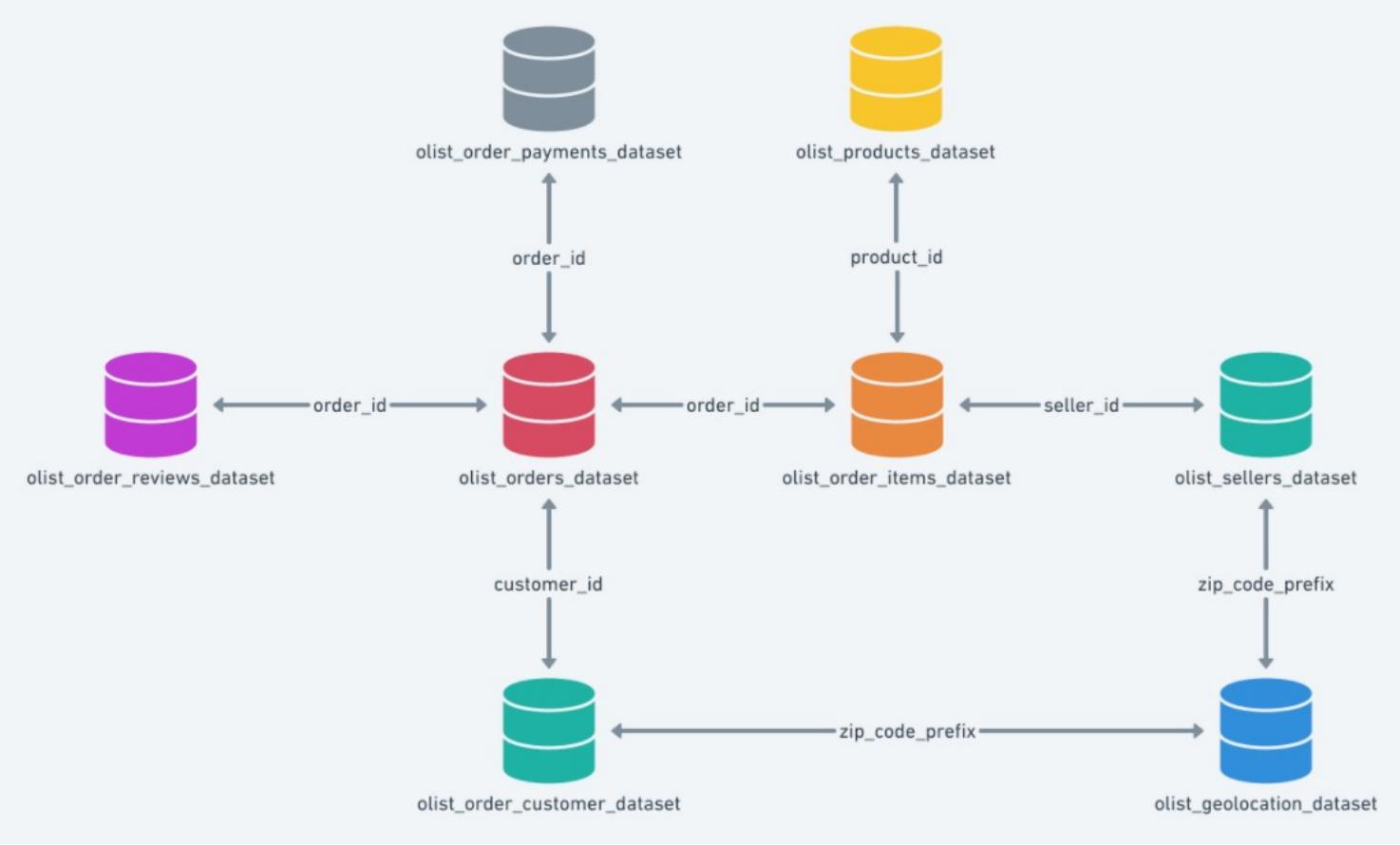
- **Missions :**
  - Fournir aux équipes d’e-commerce une segmentation des clients pour les campagnes de communication.
  - Comprendre les différents types d’utilisateurs.
  - Fournir à l’équipe marketing une description actionnable
  - Proposer un contrat de maintenance.

II. Interprétation de la problématique

- Problème d’apprentissage non-supervisé :
  - Choix des variables pour effectuer une classification ?
  - Description des clusters / Choix d’un modèle

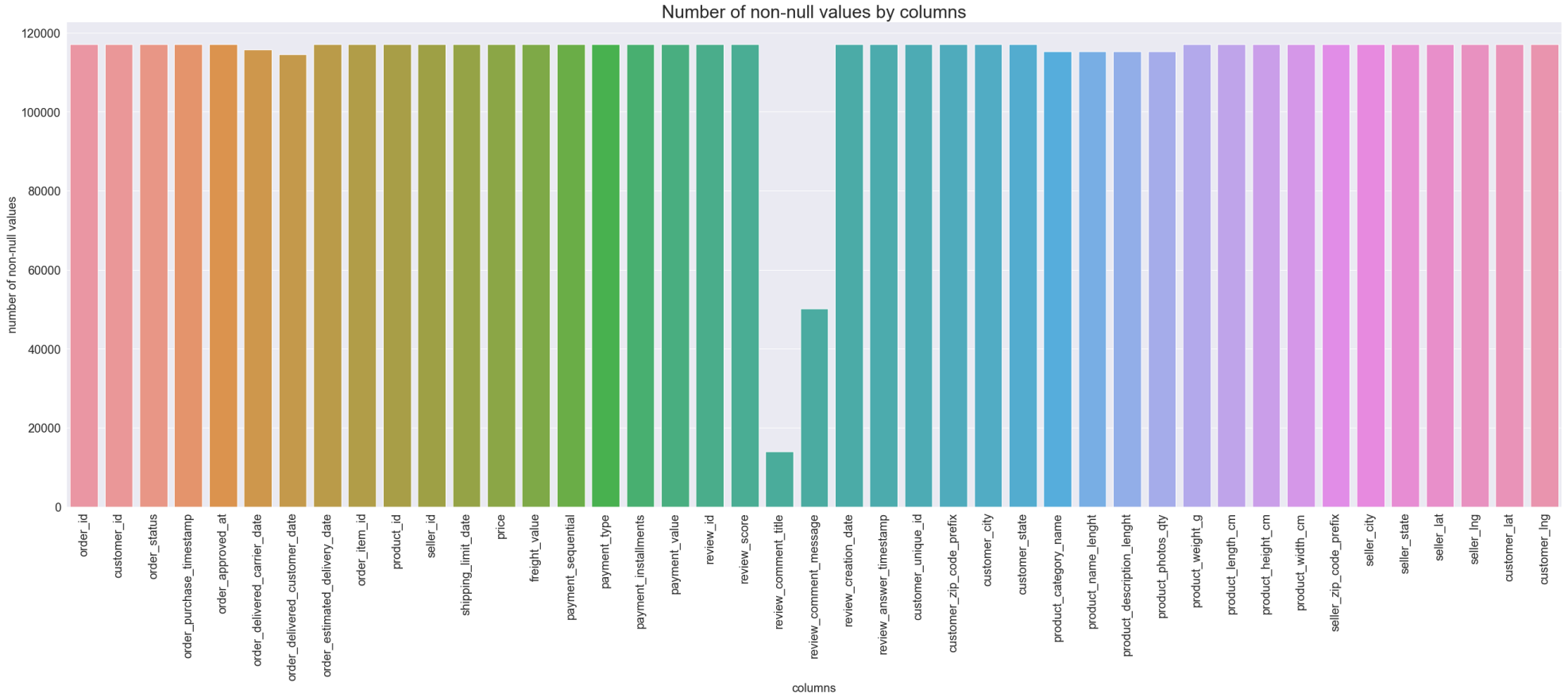
- Proposer un contrat de maintenance :
  - Tester le modèle sur les données du 3, 6 et 12 dernières mois et comparer avec le modèle testé sur l’ensemble des données.

III. Nettoyage



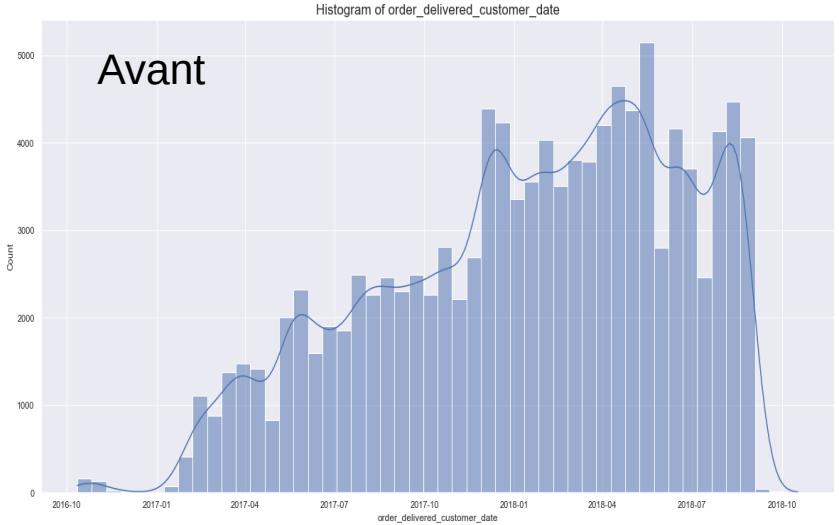
117029 lignes
43 colonnes

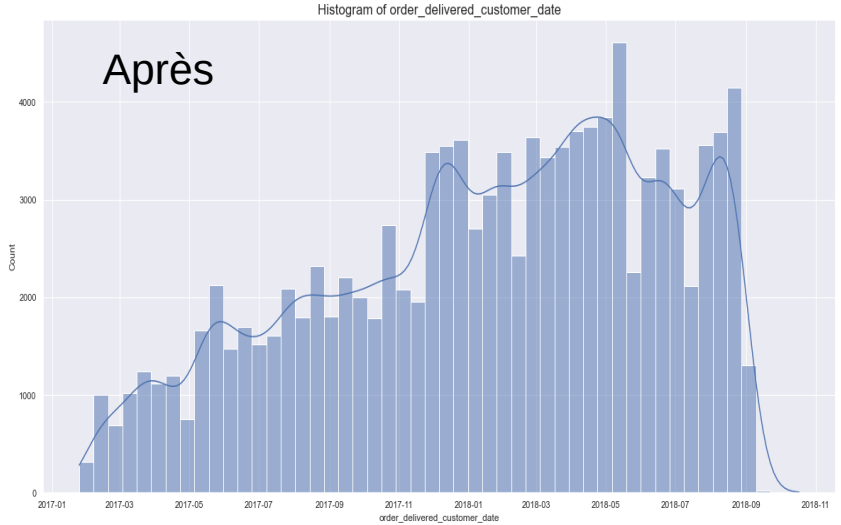
III. Nettoyage



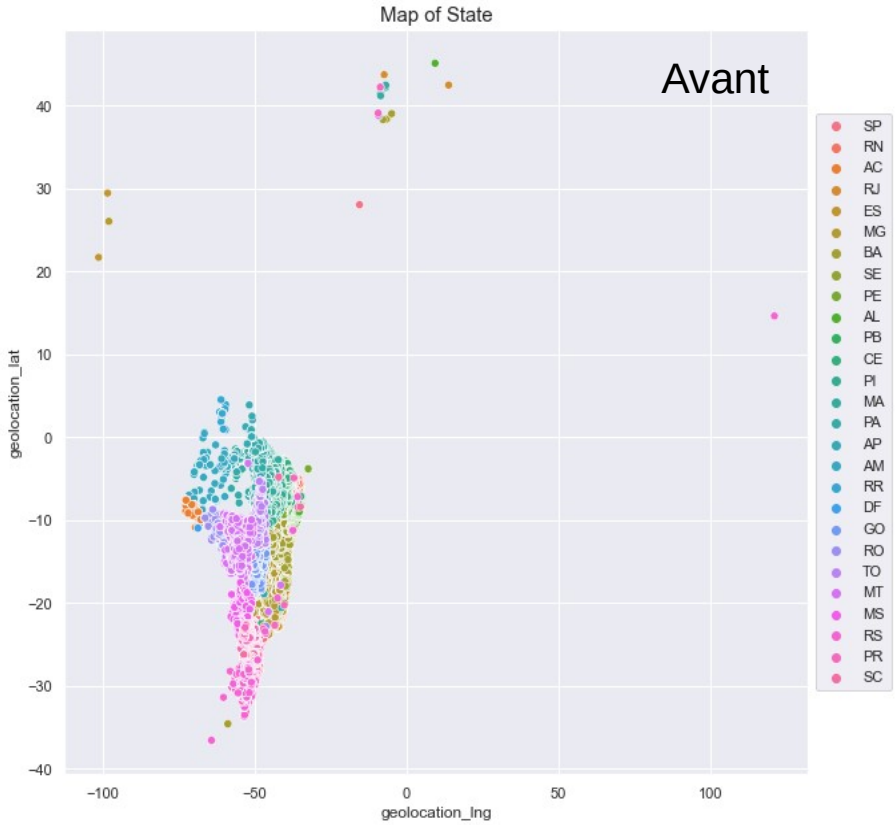
III. Nettoyage

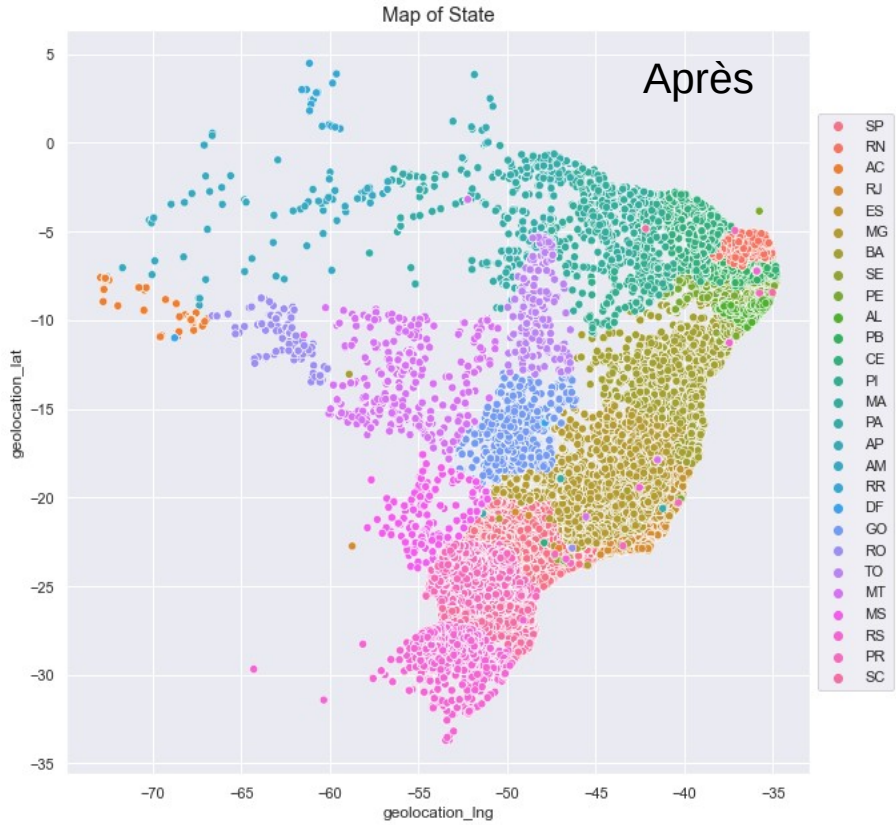
- Imputation des valeurs manquantes par la moyenne (product_name_lenght, product_description_lenght, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm) et par ‘no_information’ (product_category_name).
- Suppression des données pour dates antérieures à 2017.





III. Nettoyage



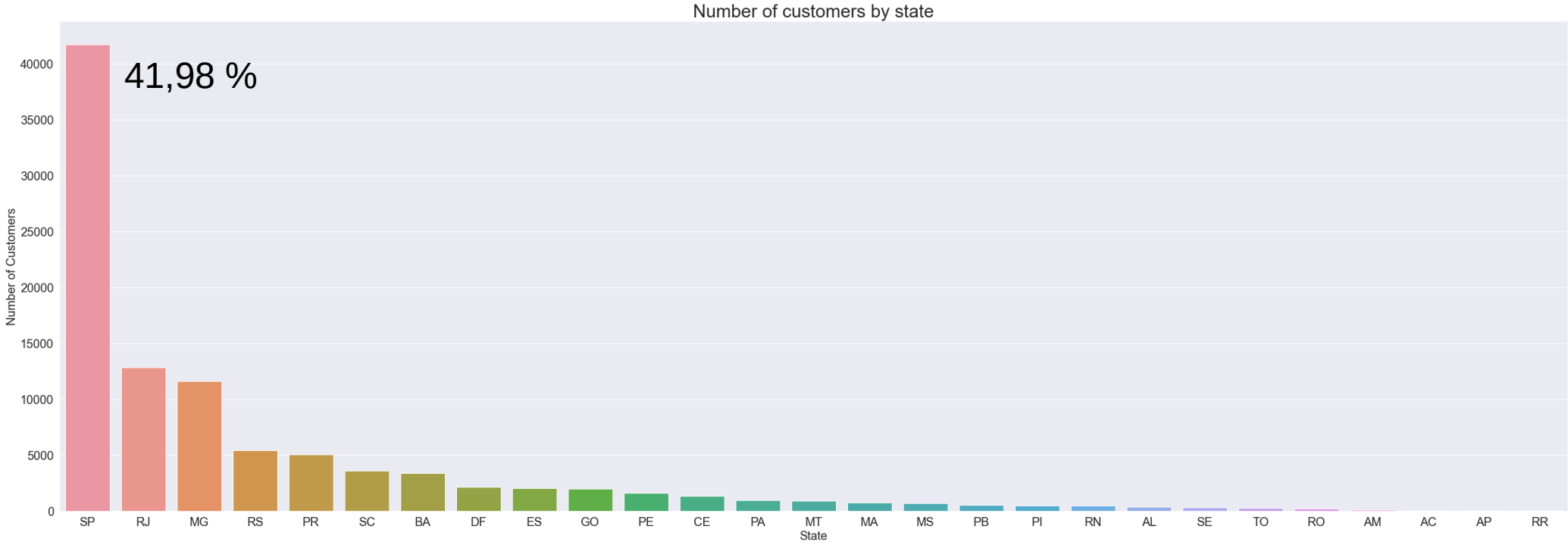


IV. Feature Engineering

Distance vendeurs – acheteur
Temps de livraison estimé
Temps de livraison
Score moyen du vendeur
Année, mois, jour, heure d’achat
Nombre de commandes par client
Prix total du produit (prix + fret)
Type de paiement (binaire)
Statut de la commande (binaire)
Volume du produit
Review (binaire)
State / mois d’achat / jour d’achat / catégorie de produit

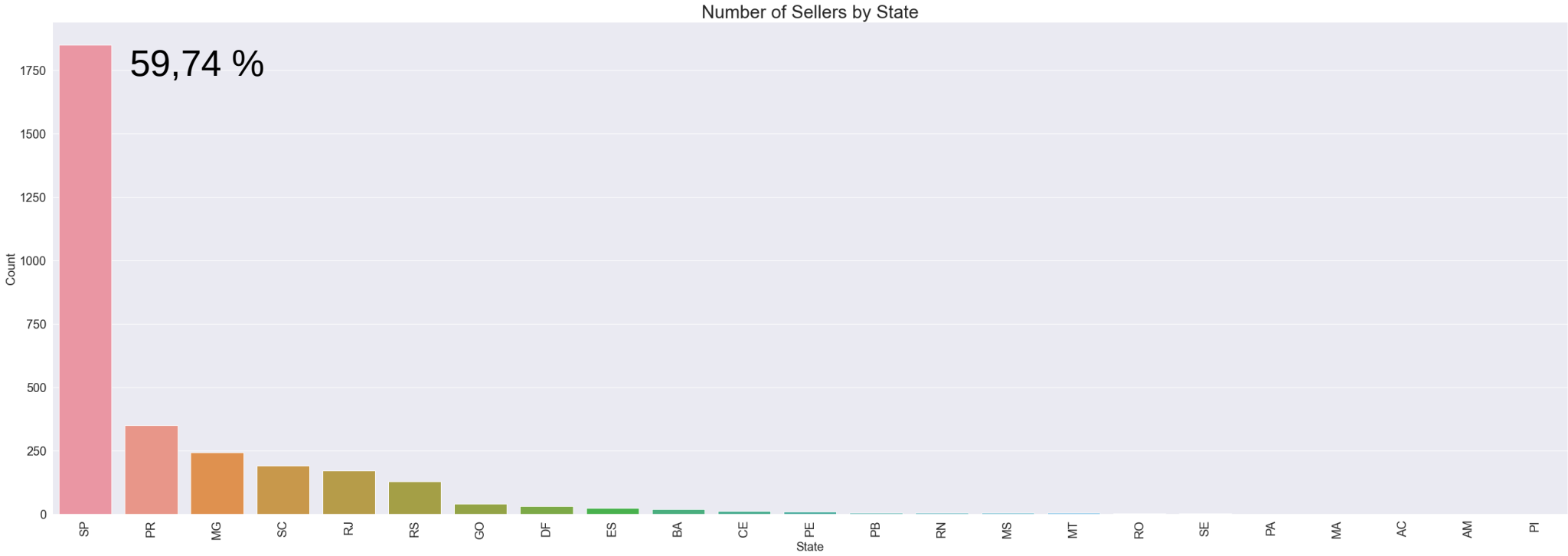
V. Exploration

96096 clients et 99441 commandes



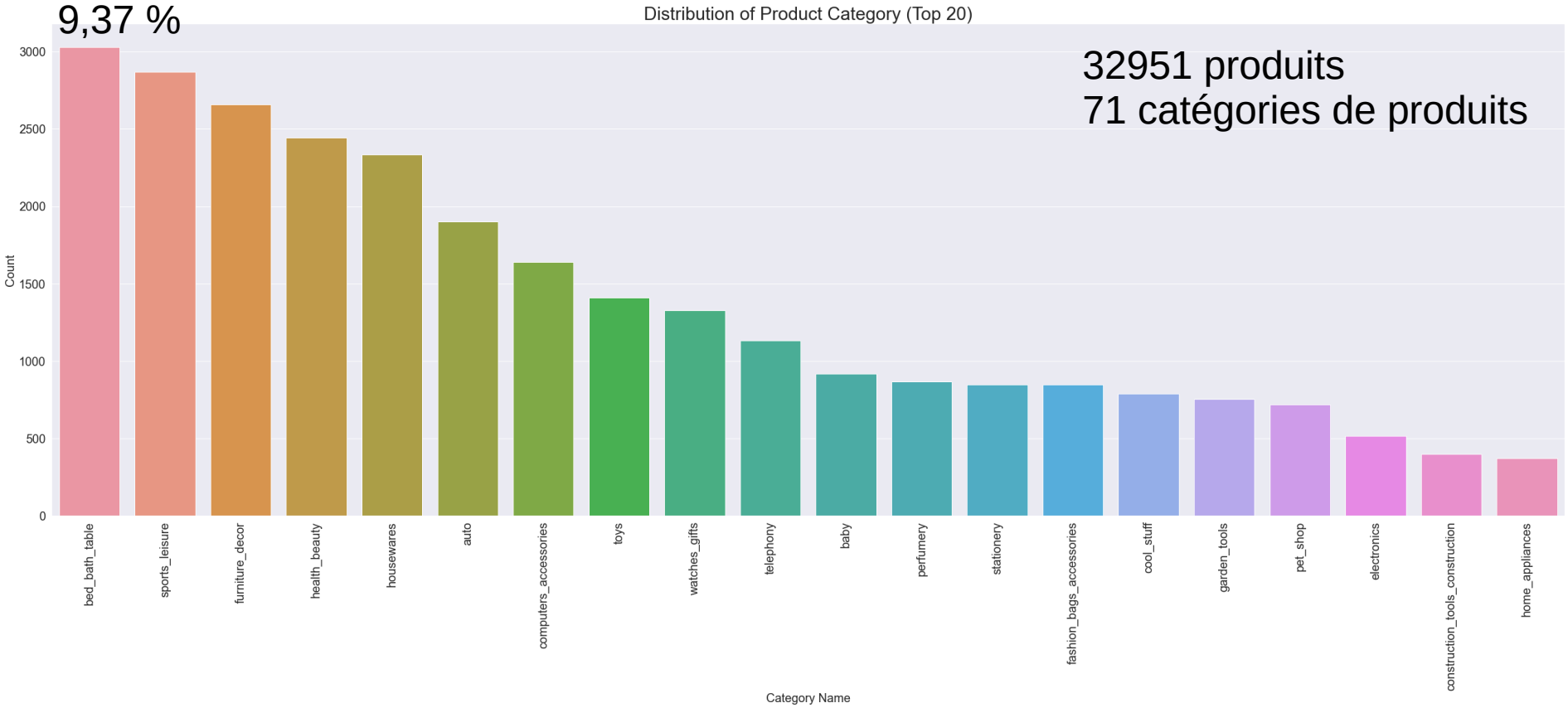
V. Exploration

3095 vendeurs



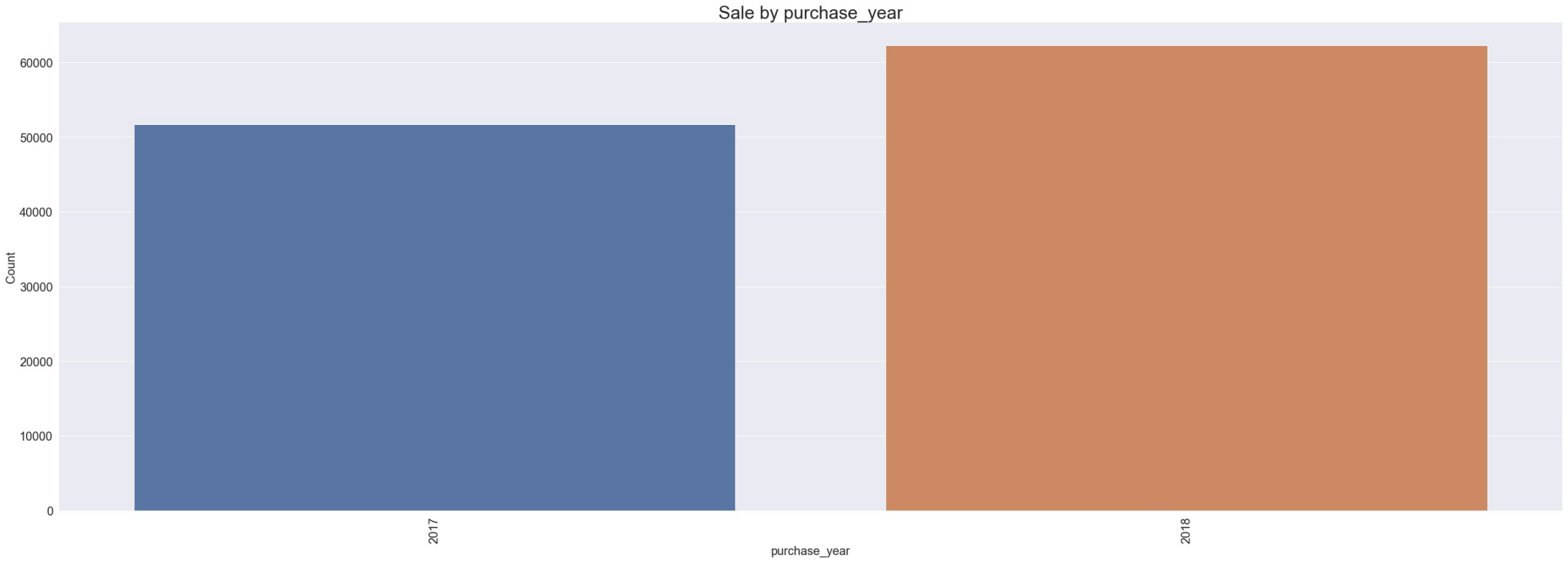
V. Exploration

9,37 %

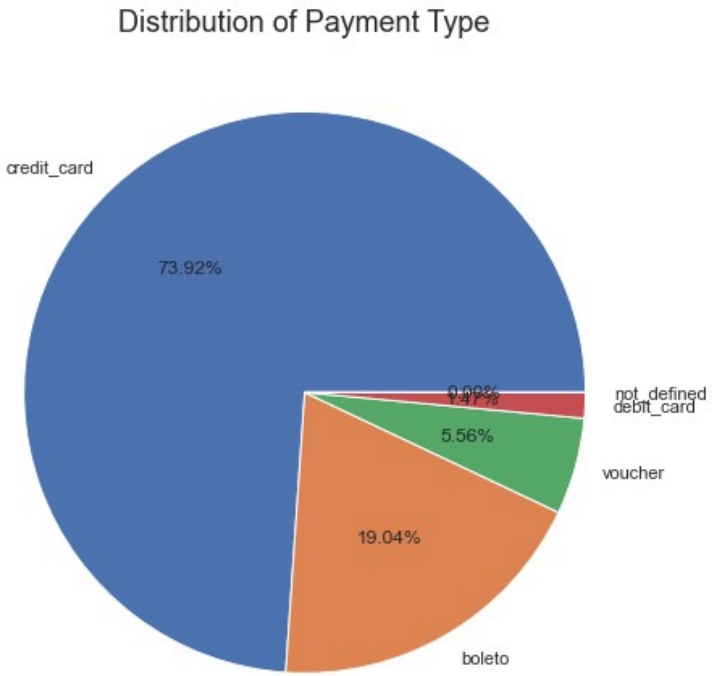


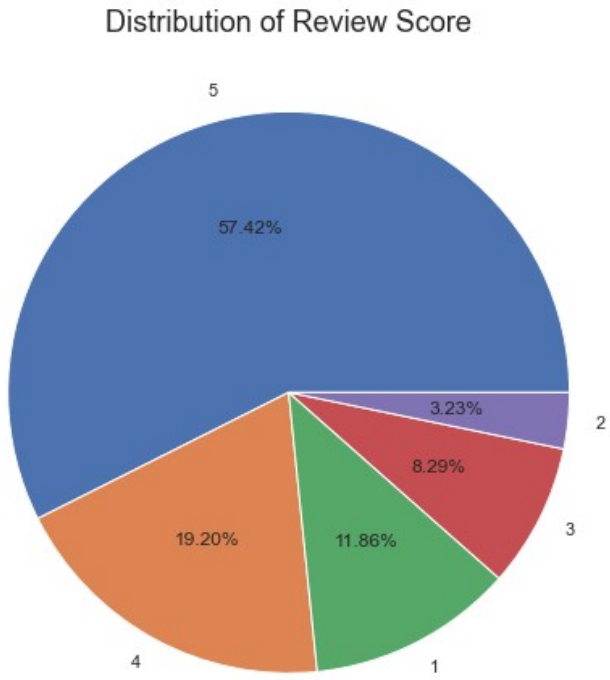
32951 produits
71 catégories de produits

V. Exploration



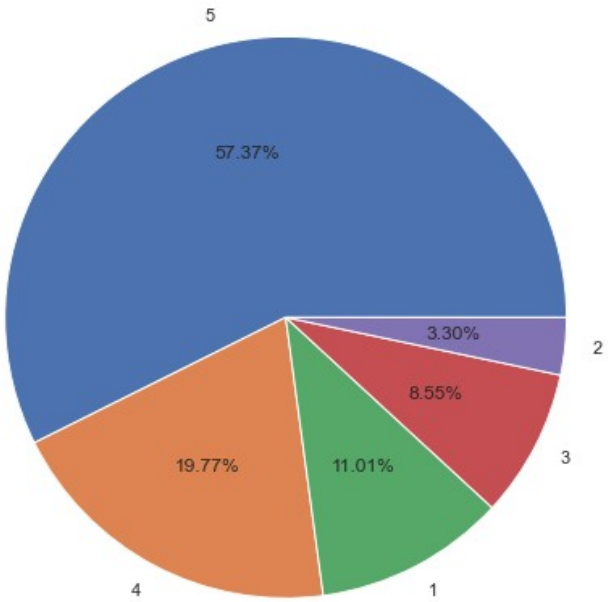
V. Exploration



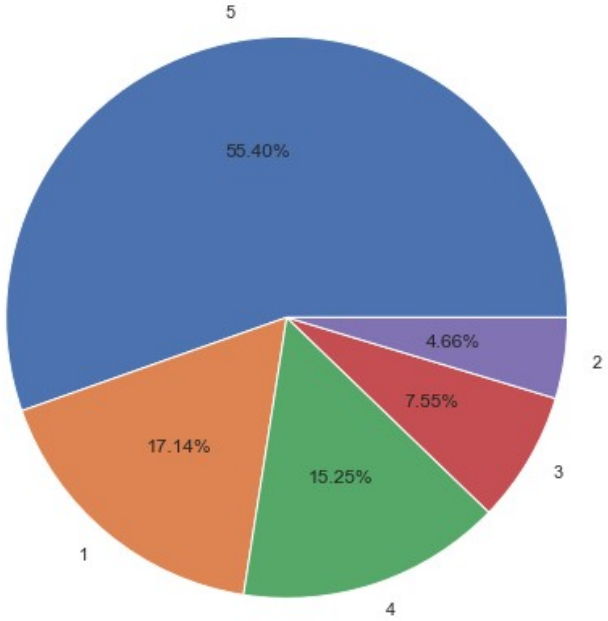


V. Exploration

Distribution of Reviews Without Comment Title

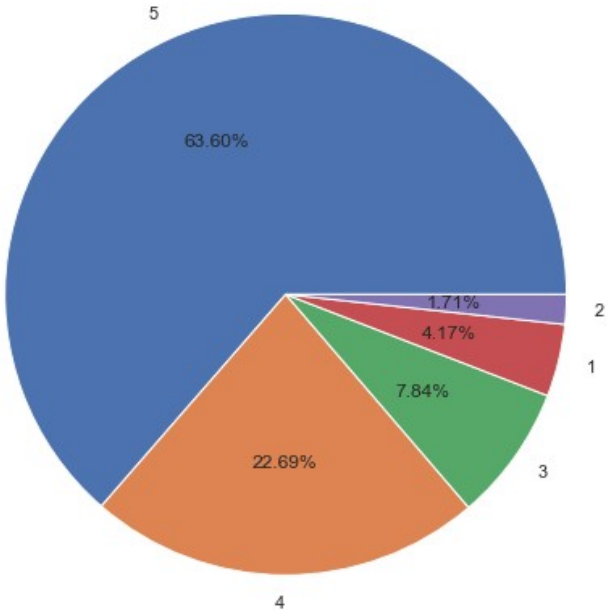


Distribution of Reviews With Comment Title

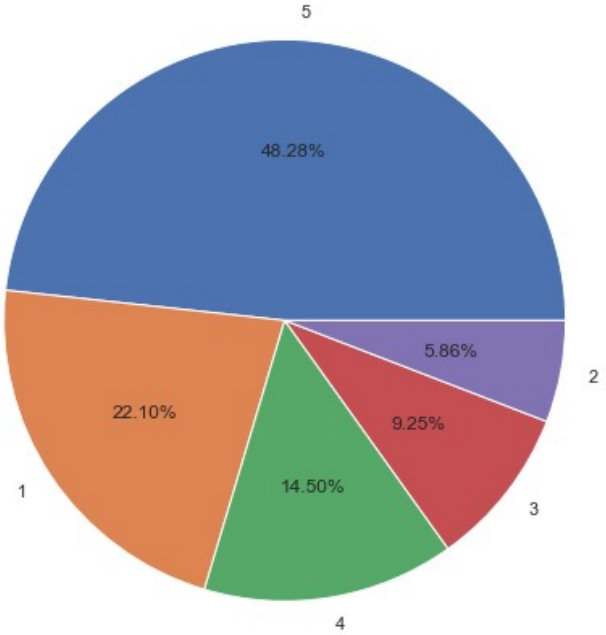


V. Exploration

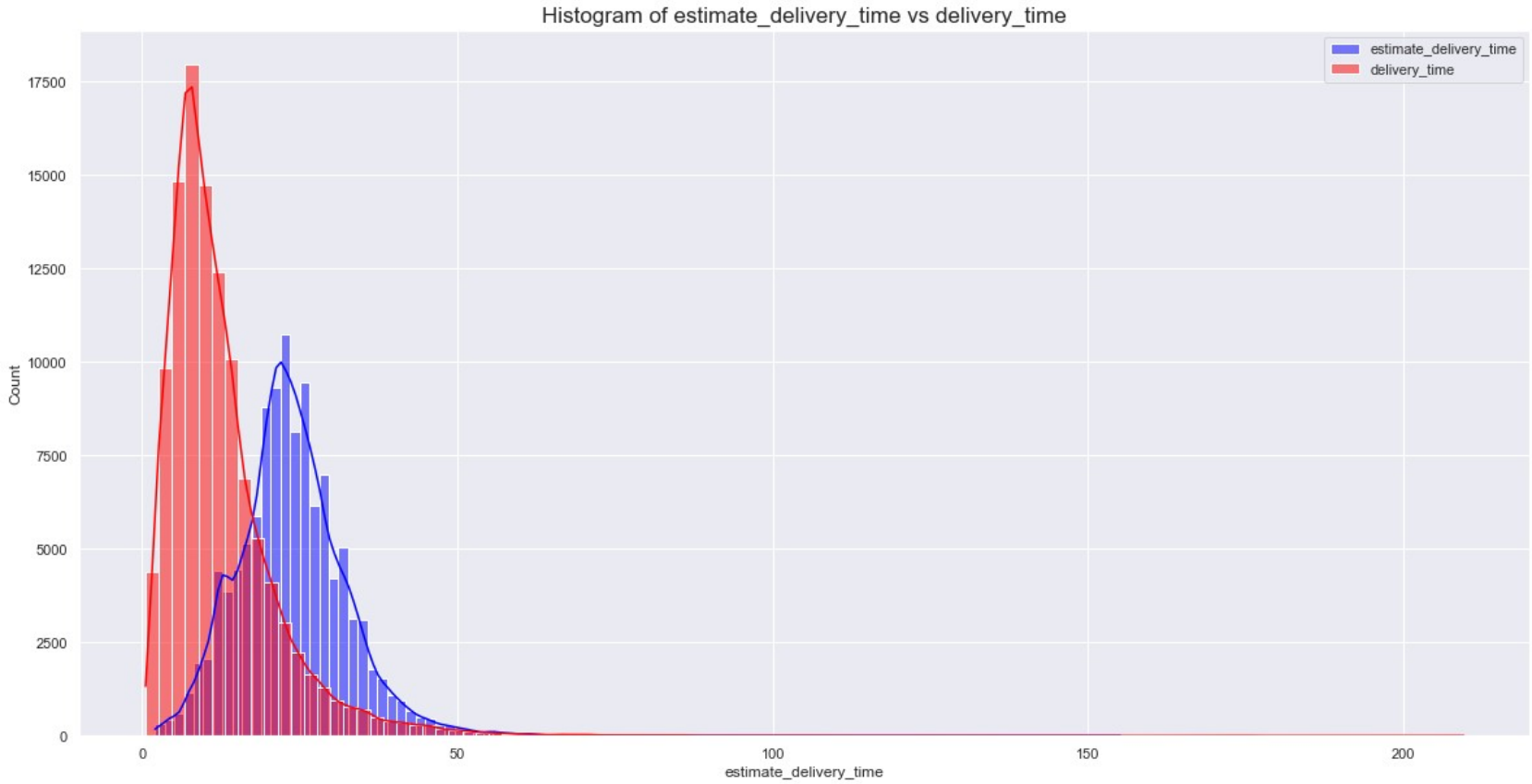
Distribution of Reviews Without Comment Message



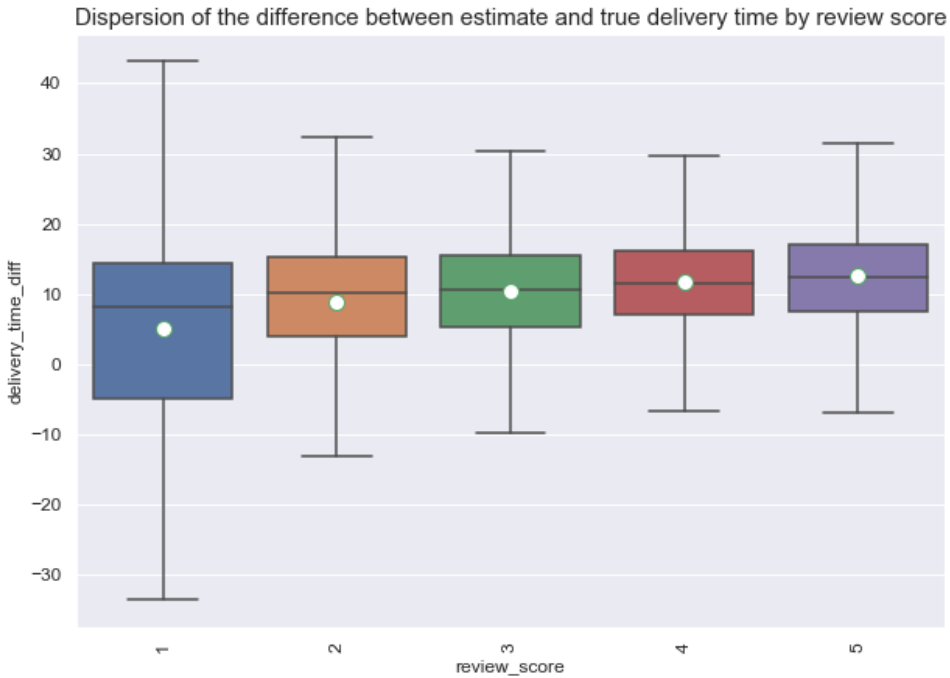
Distribution of Reviews With Comment Message

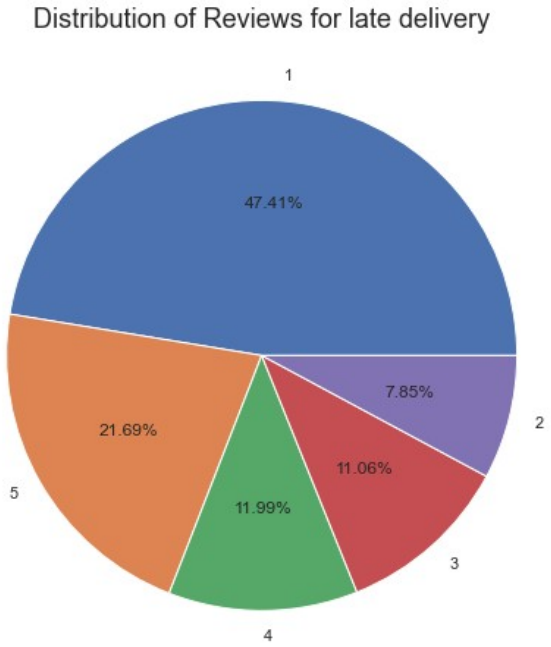


V. Exploration



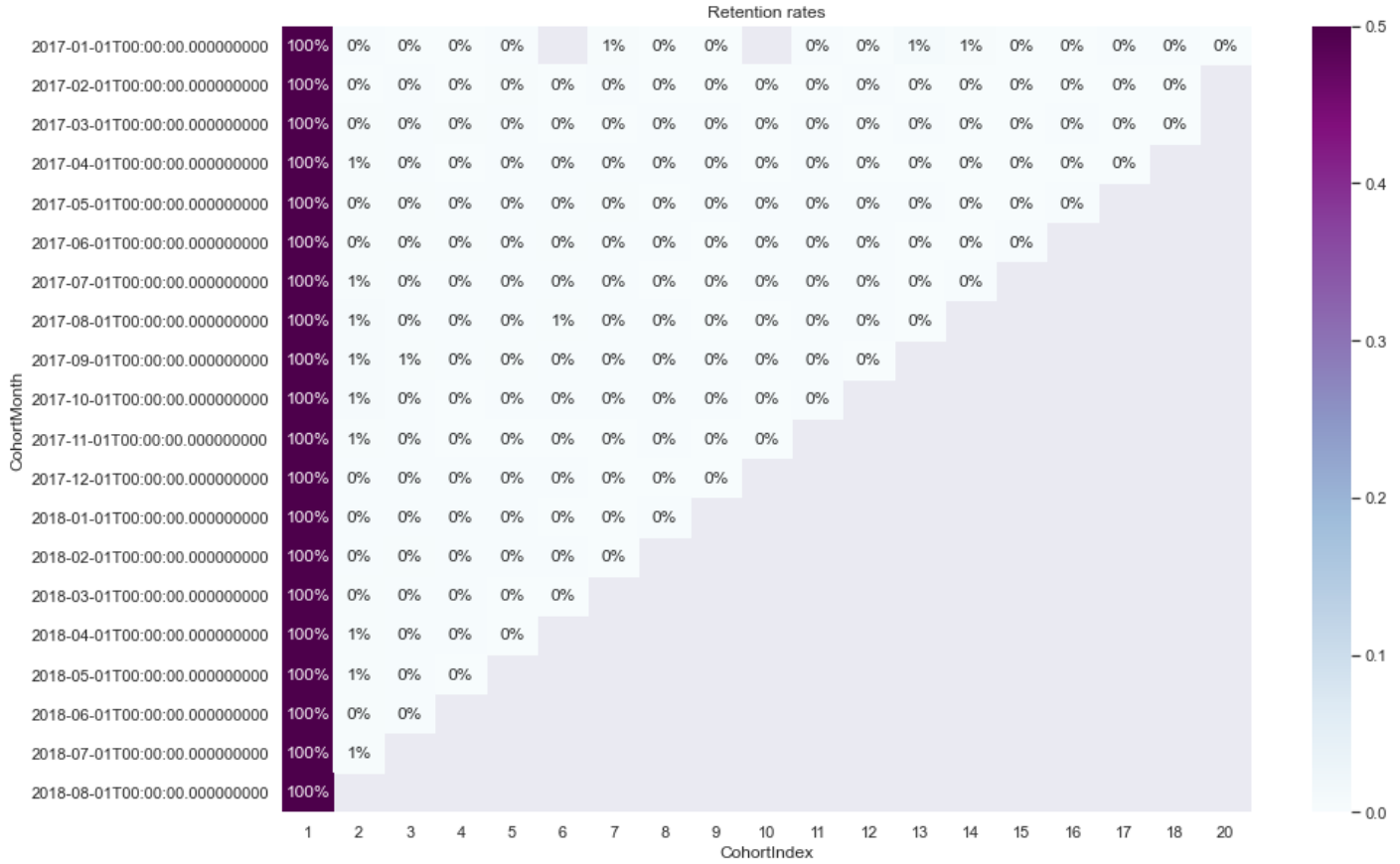
V. Exploration





V. Exploration

Mois du 1er achat



Rétention sur des mois

VI. Modélisation

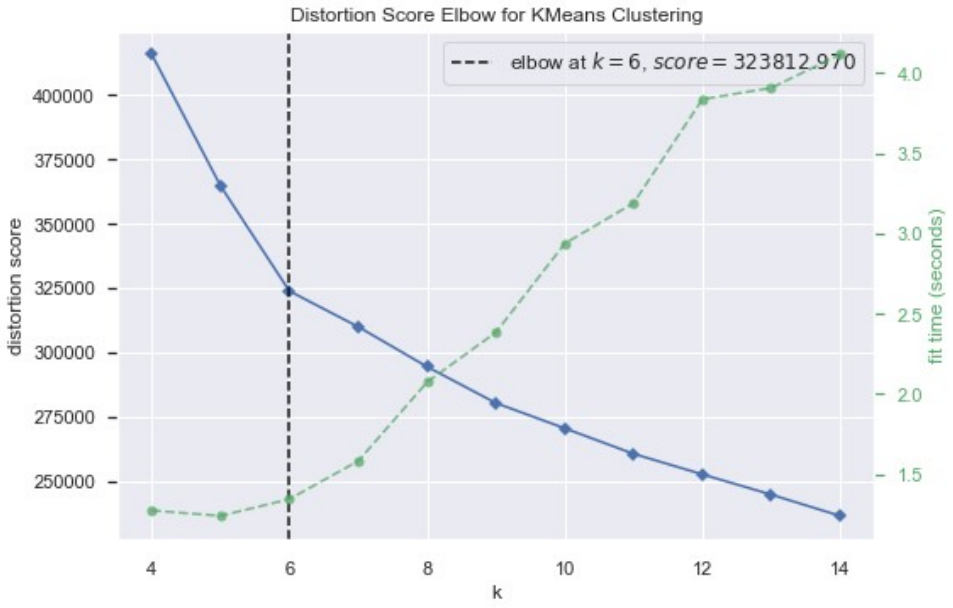
- KMeans, Gaussian Mixture et RFM
- Sélection des features
- ACP pour visualisation
- Kelbow pour le nombre de clusters
- Coefficient de silhouette pour la mesure de la qualité
- Explication des clusters

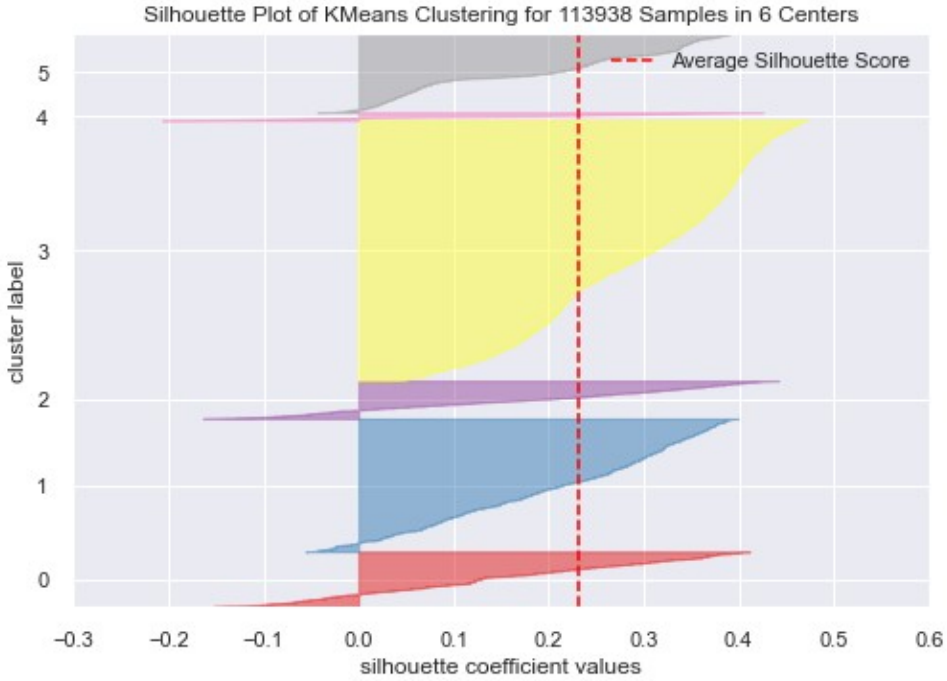
VI. Modélisation

Features	Modèle	Nombre de Clusters	Silhouette Score
Un maximum possible	KMeans	10	0,082
Un maximum possible	Gaussian Mixture	10	0,005
En lien avec les produits	KMeans	6	0,231
En lien avec les produits	Gaussian Mixture	6	0,044
En lien avec les habitudes d’achat du client	KMeans	8	0,124
En lien avec les habitudes d’achat du client	Gaussian Mixture	8	0,081

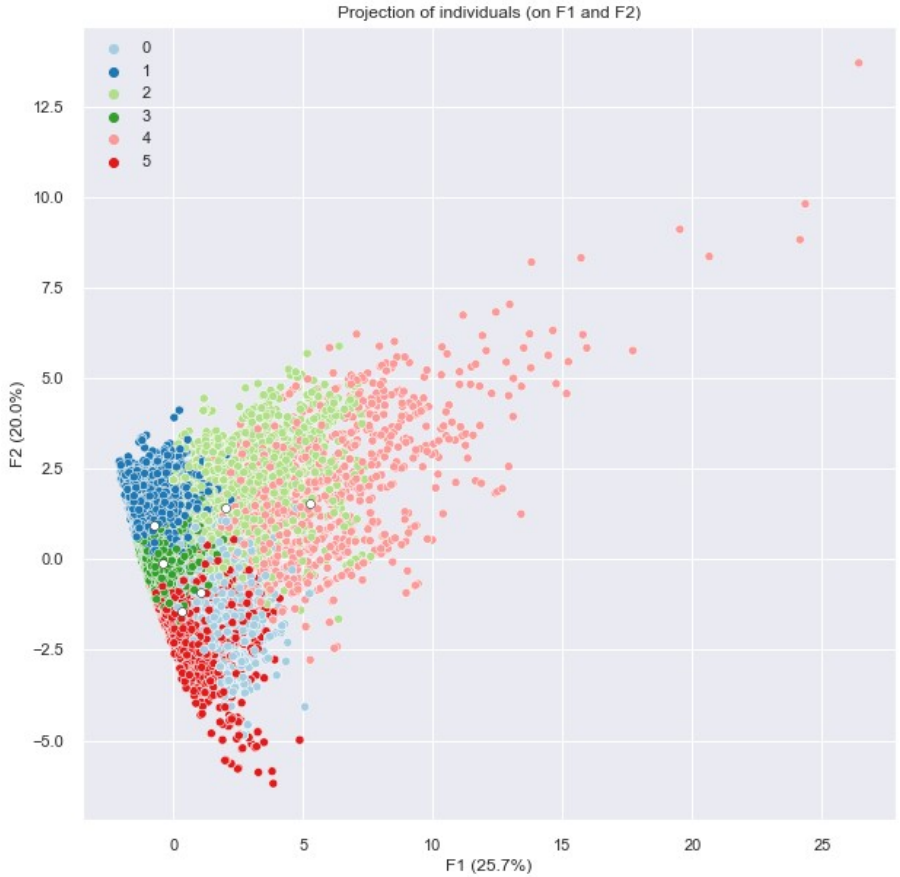
VI. Modélisation

KMeans avec feature en lien avec les produits

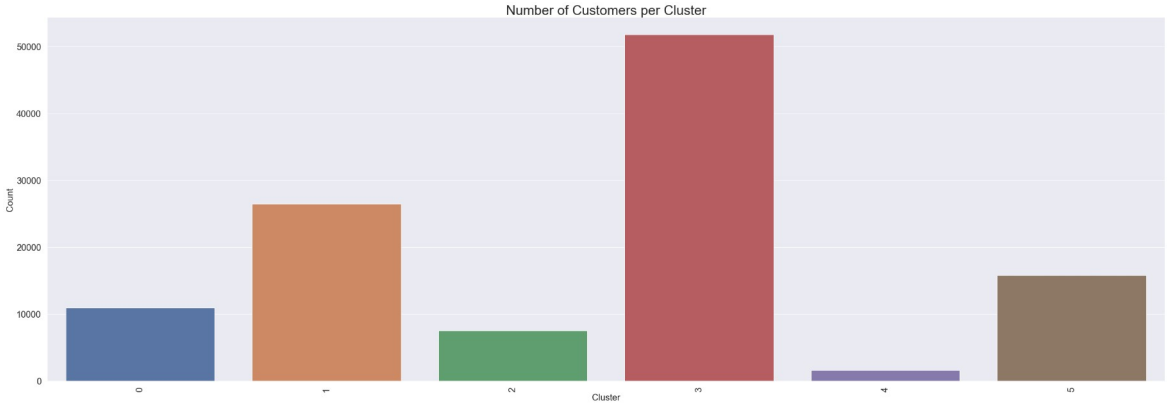




VI. Modélisation



KMeans avec feature en lien avec les produits

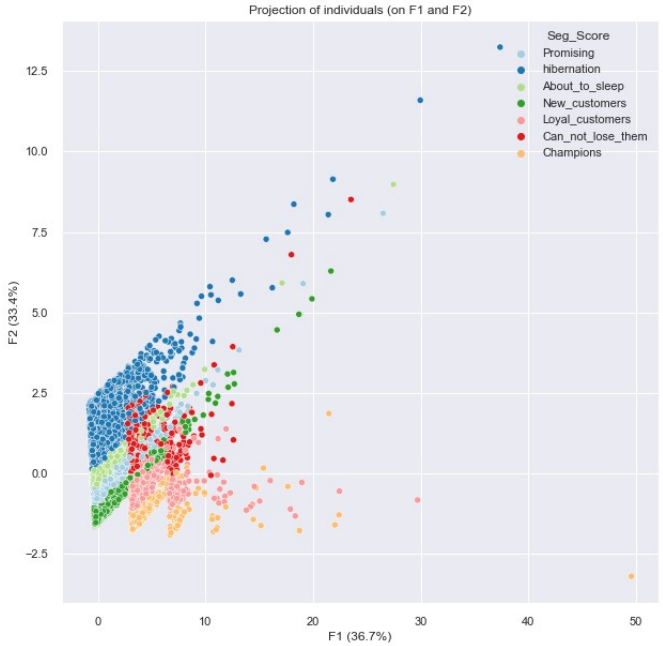


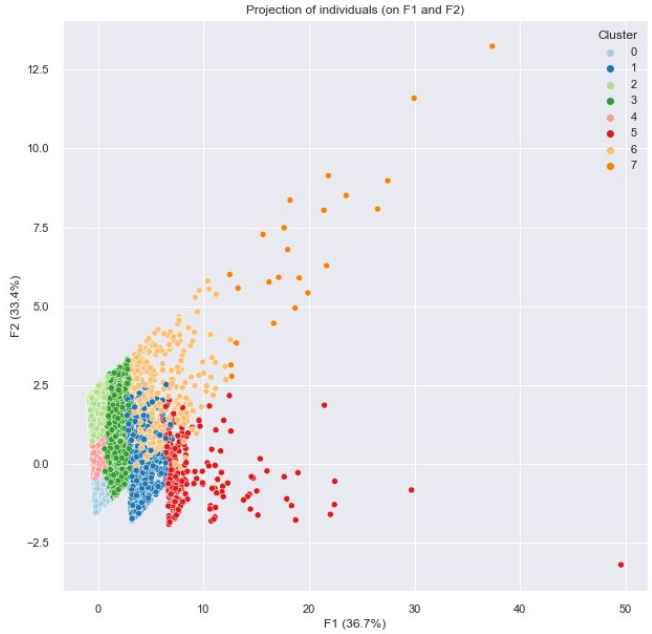
VI. Modélisation

Cluster	Description
0	Achète des produits avec une description longue.
1	Achète des produits avec un nom court.
2	Achète des produits lourds. Moins intéressé en produit électronique. Moins intéressé en produit de santé, soin, mode. Plus intéressé en produit maison. Moins intéressé en produit loisirs. Plus intéressé en produit d'office.
3	Plus intéressé en produit maison.
4	Dépense le plus. Achète des produits avec une description longue. Achète des produits lourds.
5	Achète des produits avec beaucoup de photos.

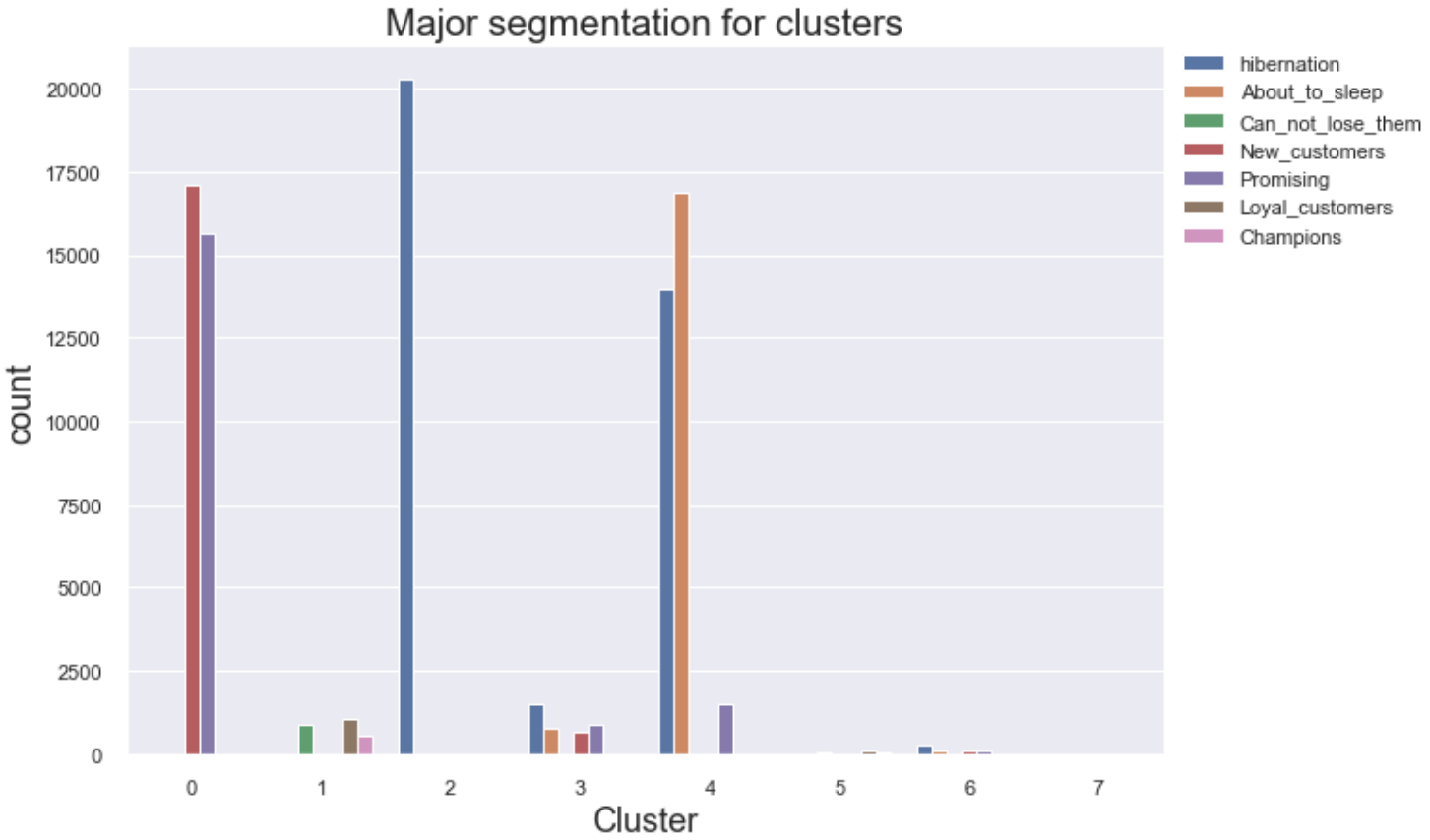
VI. Modélisation

- RFM → Manuel (7 clusters) et KMeans (8 clusters)
- Silhouette score : 0,456 (KMeans)



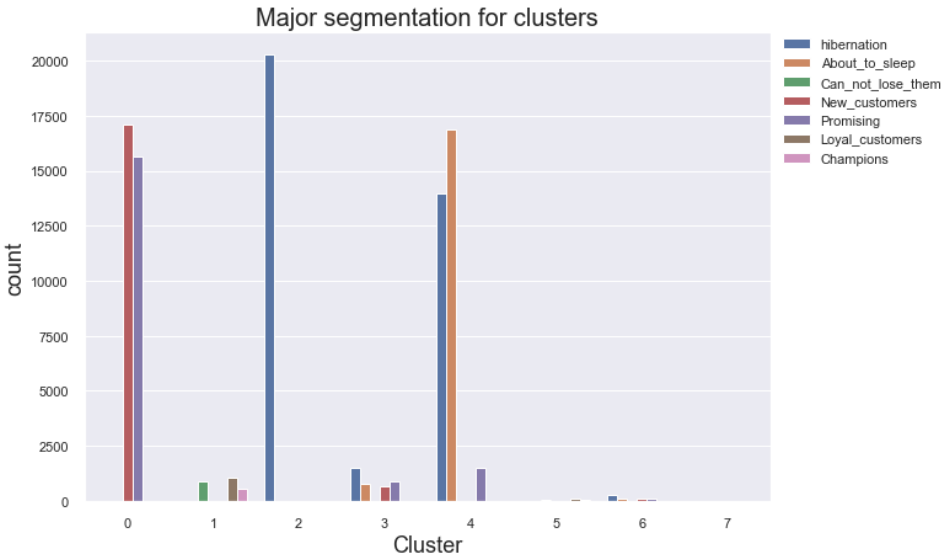


VI. Modélisation



VI. Modélisation

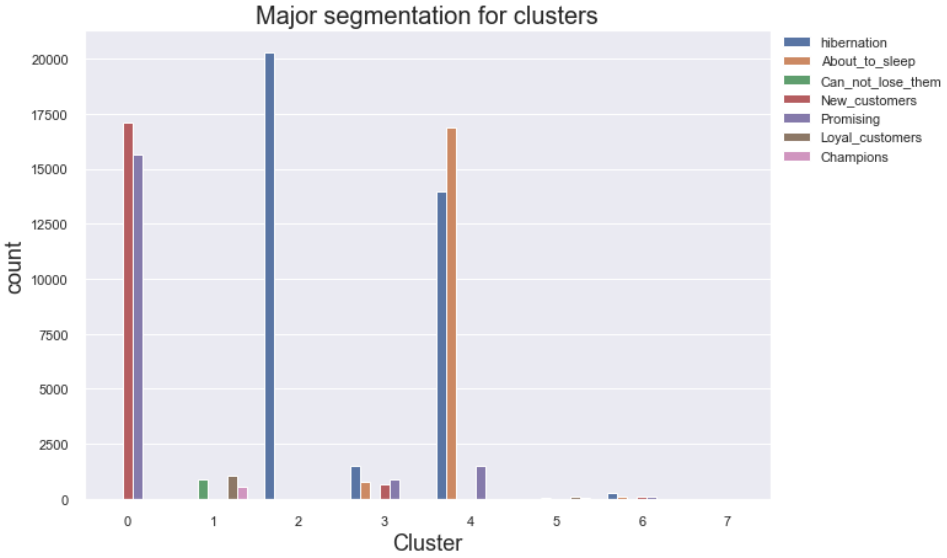
Tous les données



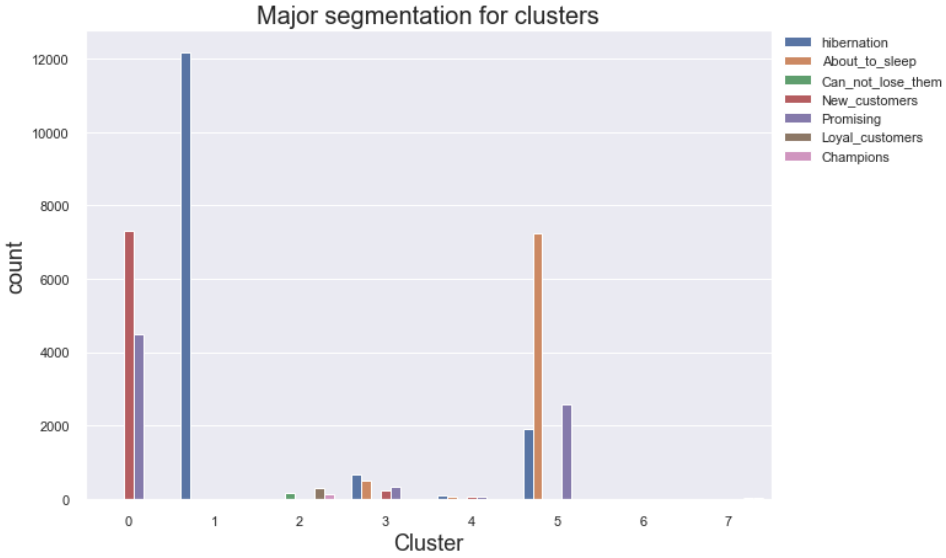
3 derniers mois (silhouette score = 0,401)

VI. Modélisation

Tous les données



6 derniers mois (silhouette score = 0,462)



VII. Conclusion

- KMeans
- RFM : clusters simples, intelligibles, mais certains clusters très vides (inutiles)
- Avec autres features : clusters plus complexes, peu intelligibles
- Amélioration :
  - Plus de données sur les clients (sexe, âge, nombre de visite sur le site, produits visités / produits intérêts…)
- Contrat de maintenance : arbitraire