PROJET N°4: SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

BOURBON Vicente

Sommaire

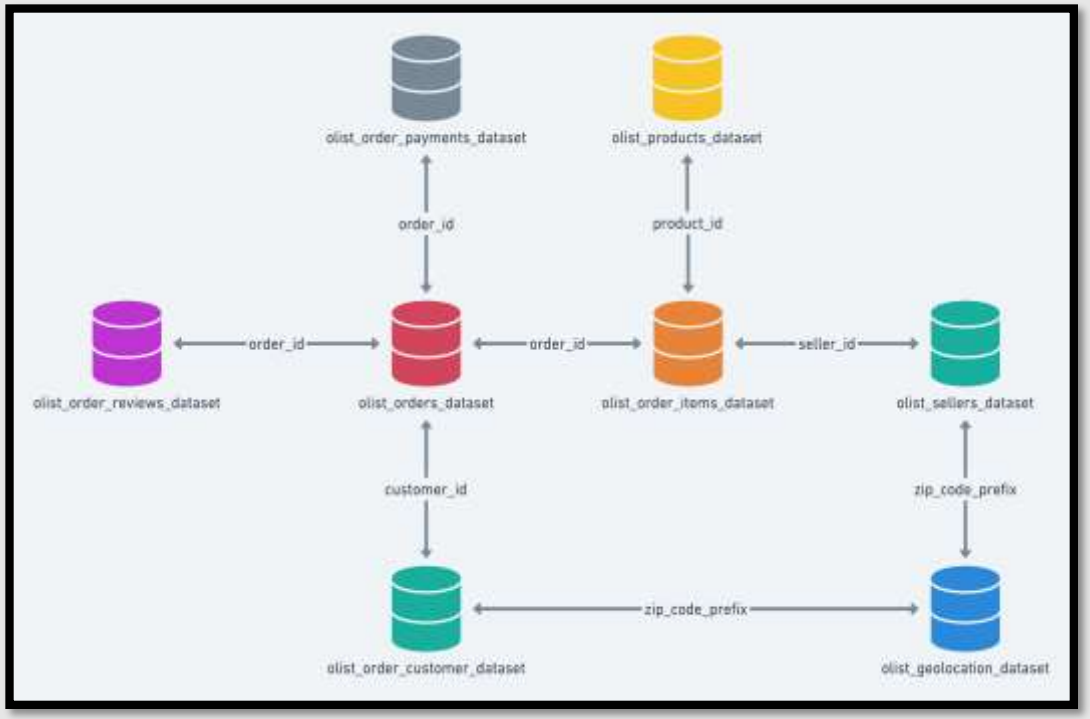
- Problématique
- Cleaning
- Feature Engineering
- Exploration
- Modélisations du problème
- Choix du modèle
- Simulation de maintenance du modèle

I. Problématique

- Olist: entreprise brésilienne de vente en ligne
- Segmentation des clients
  - Comprendre les différents types d'utilisateurs
  - Utilisation facile pour les équipes marketing
- Contrat de maintenance de la segmentation
- Mise à disposition d'une base de données anonymisée

II. Cleaning

- Union des jeux de données
- Valeurs manquantes
  - Implémentation par valeur moyenne et valeur la plus fréquente
  - Suppression de variables
- Doublons
  - Lignes identiques
  - Commandes identiques
- Outliers
  - Quantités non négatives



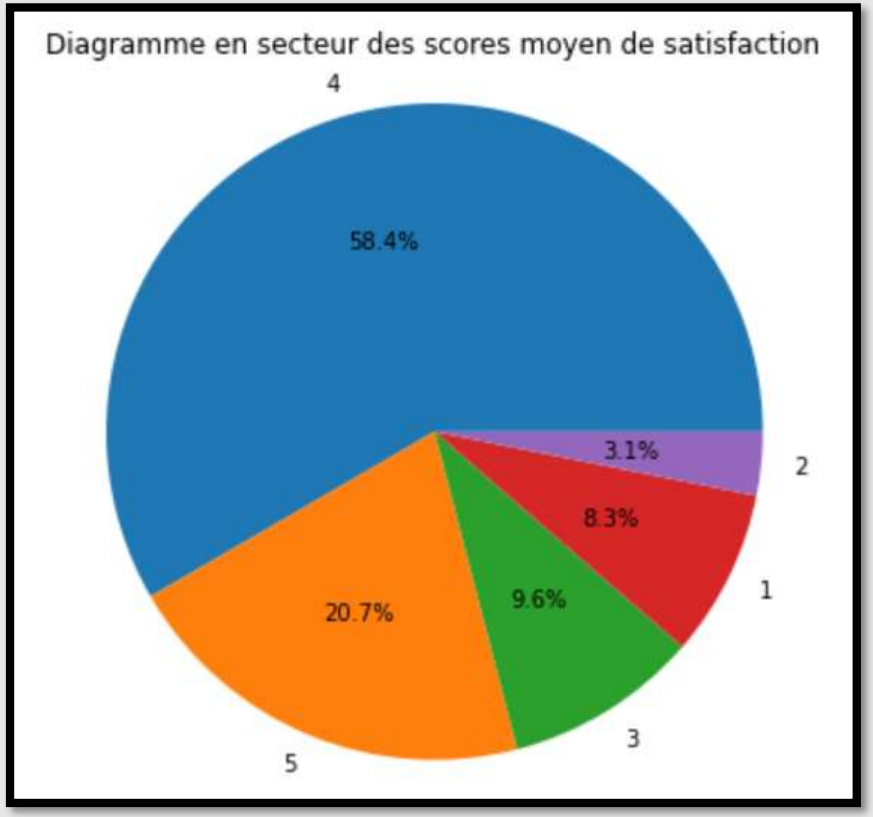
III. Feature Engineering

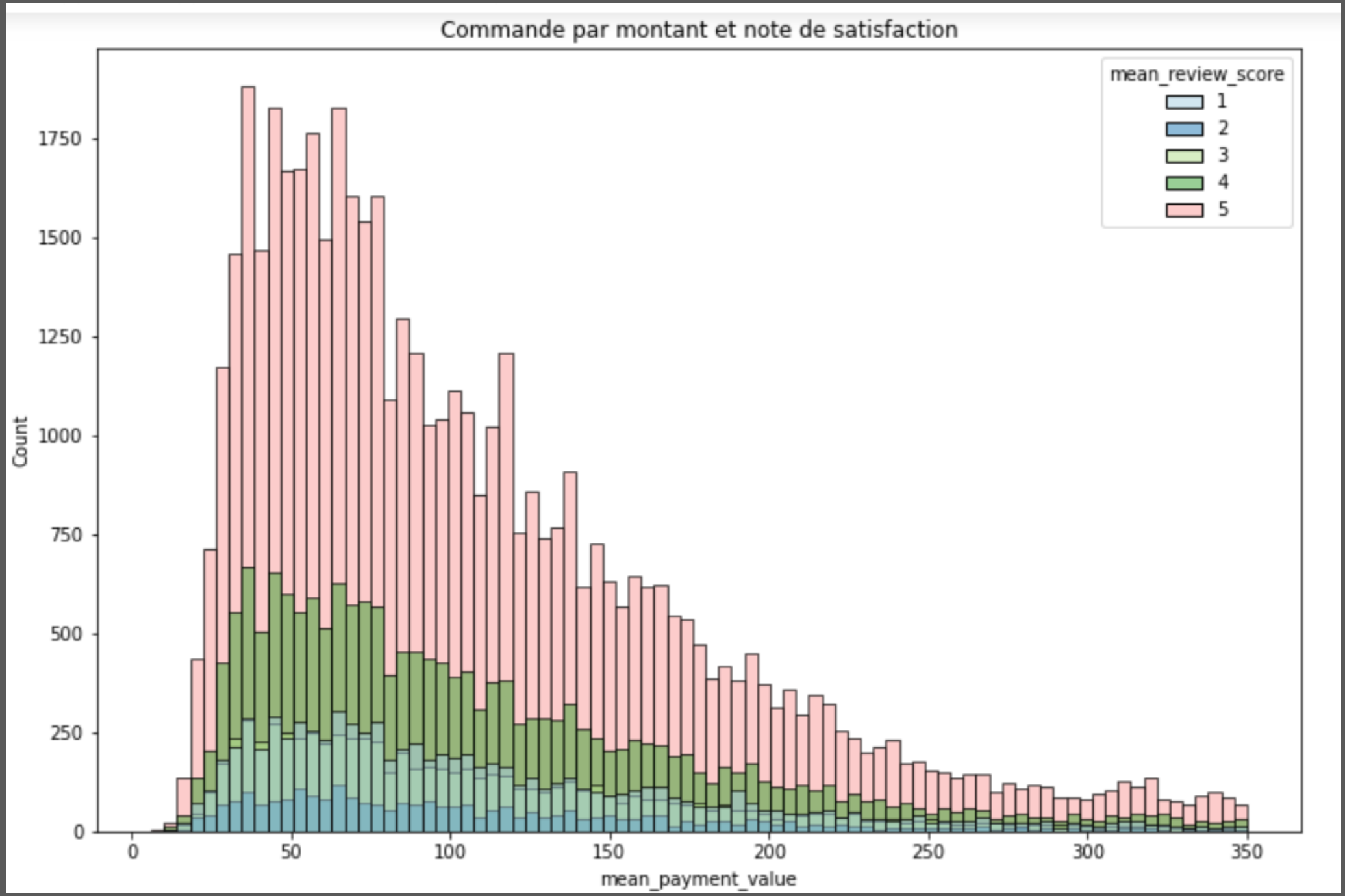
delivery_time	estimated_delivery_time
8 days 19:30:00	19 days 08:54:25
16 days 15:52:55	24 days 03:11:36
26 days 01:51:06	24 days 07:52:15
14 days 23:57:47	27 days 07:53:22
11 days 11:04:18	16 days 14:08:30

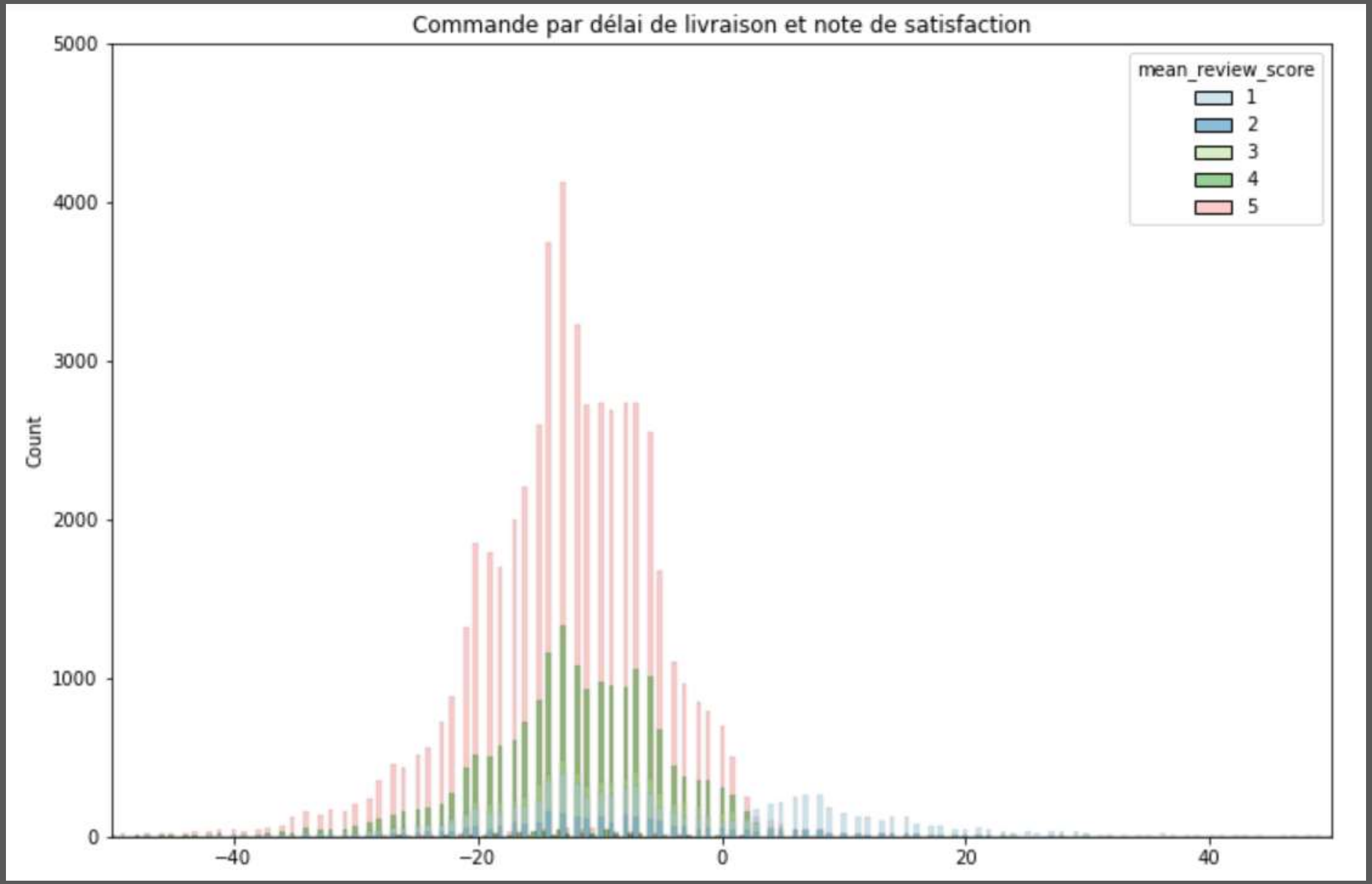
- Création de nouvelles variables
  - Temps de livraison moyen
  - Nombre totale de commandes
  - Somme totale dépensée
  - Somme moyenne d'une commande
  - Note moyenne
  - Nombre moyen de paiements
  - Dernière date de commande
  - Catégorie de produit la plus fréquente
  - Quantité moyenne de photos
- Traduction des catégories

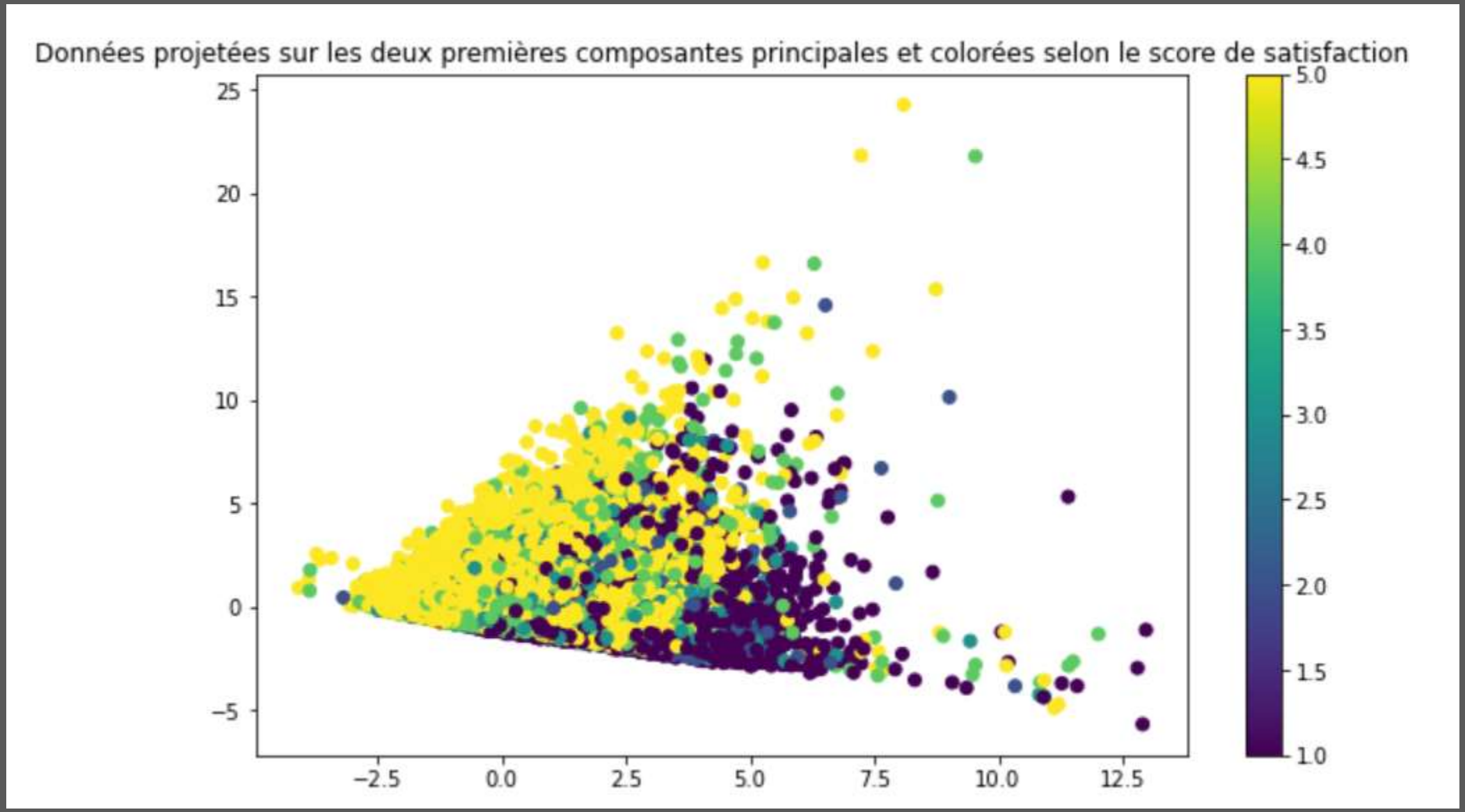
IV. Exploration des données

- Analyse Univariée
- Analyse Bivariée
  - Corrélation de variables
- Réduction dimensionnelle
  - ACP
  - T-SNE
  - Visualisation des clusters




Commande par montant et note de satisfaction
mean_review_score
1
2
3
4
5
Count
0
250
500
750
1000
1250
1500
1750
0
50
100
150
200
250
300
350
mean_payment_value


Commande par délai de livraison et note de satisfaction
mean_review_score
1
2
3
4
5
Count
5000
4000
3000
2000
1000
0
−40
−20
0
20
40


Données projetées sur les deux premières composantes principales et colorées selon le score de satisfaction
25
20
15
10
5
0
−5
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0

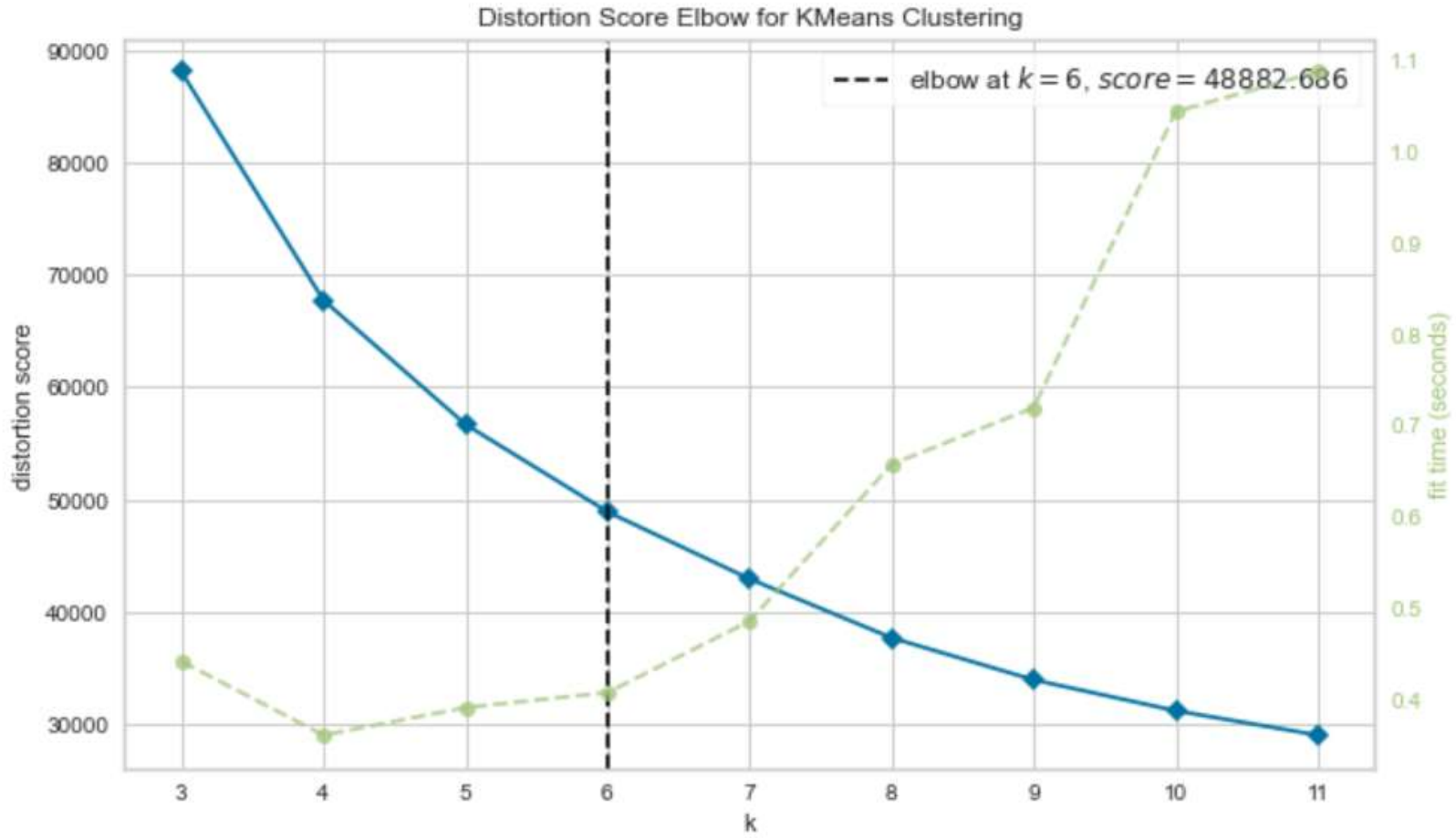
V. Modélisations du problème

Modèles utilisés

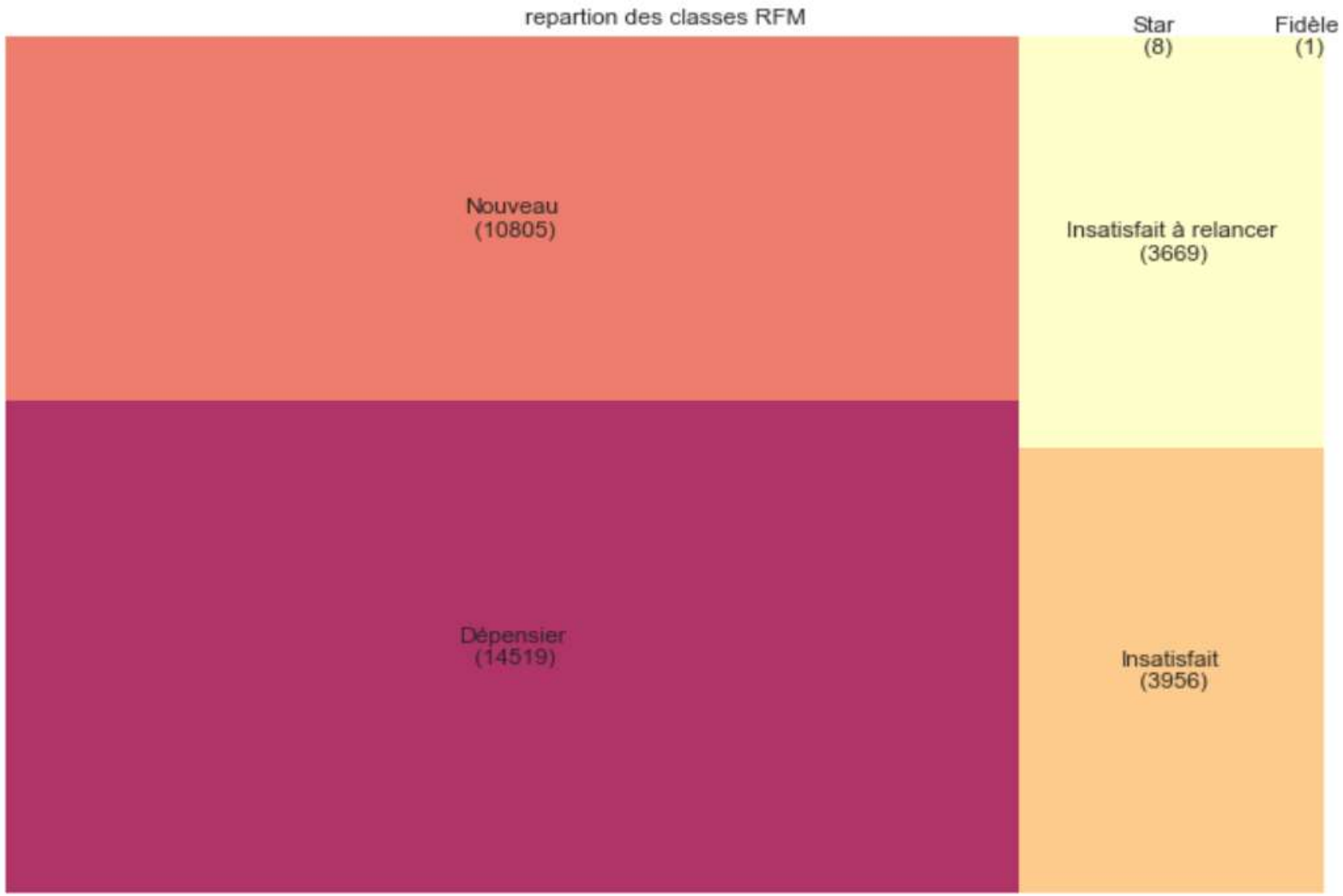
- Méthodes traditionnelles
  - RFM
  - Loi de Pareto
- Méthodes de clustering
  - K-Means
  - DBSCAN
  - Agglomerative Clustering

Démarche de modélisation

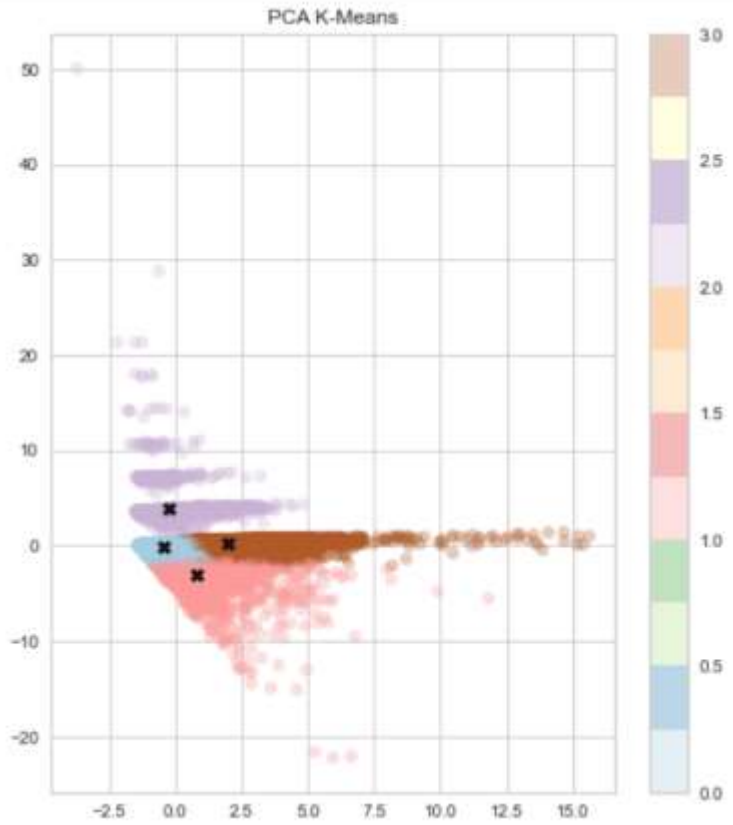
- Optimisation des hyperparamètres
- Proportions des clusters
- Visualisation des clusters
- Boxplot des features par cluster
- Caractérisation des clusters

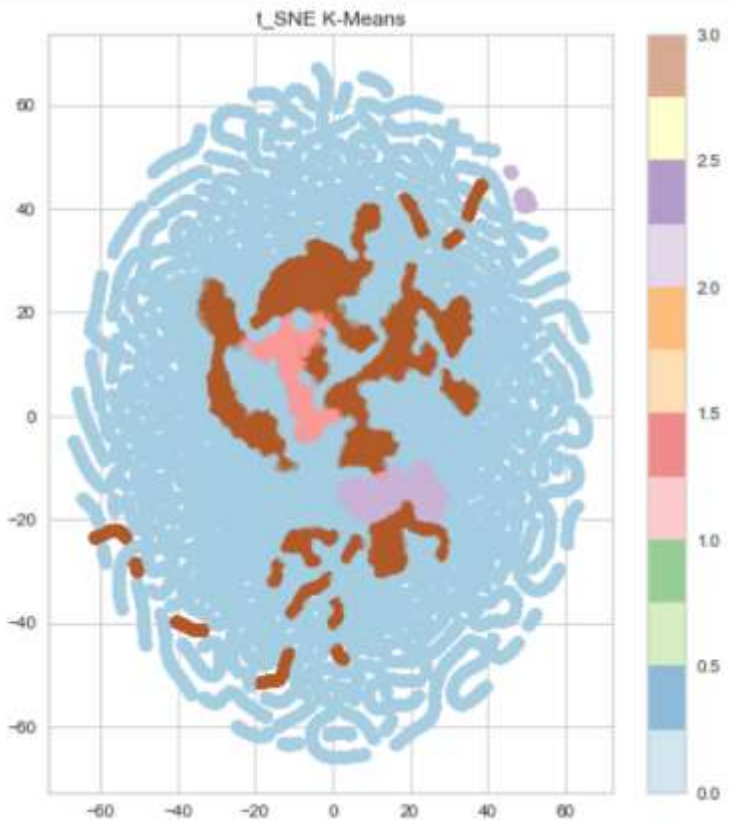


Optimisation hyperparamètre

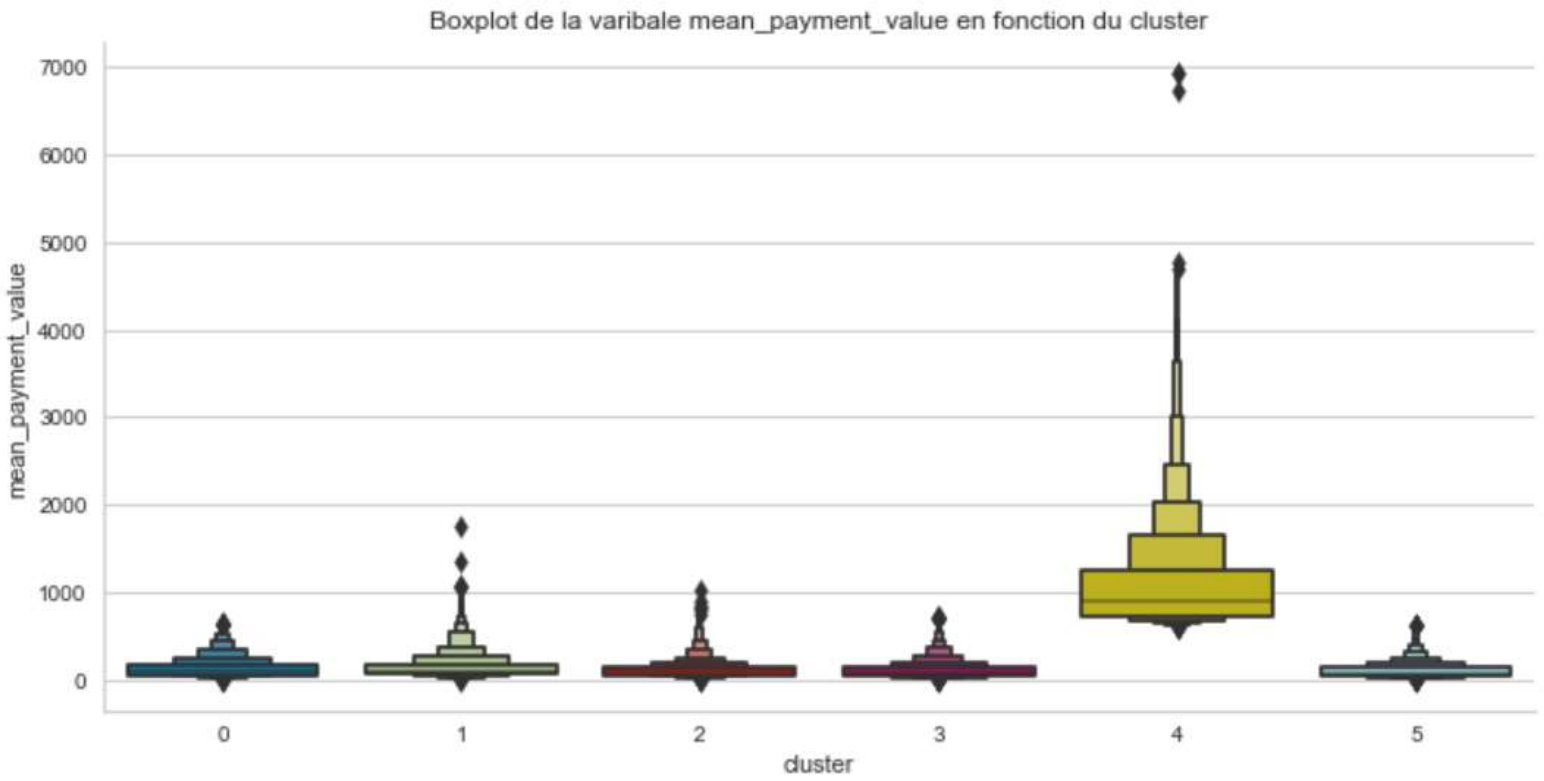


Proportion clusters

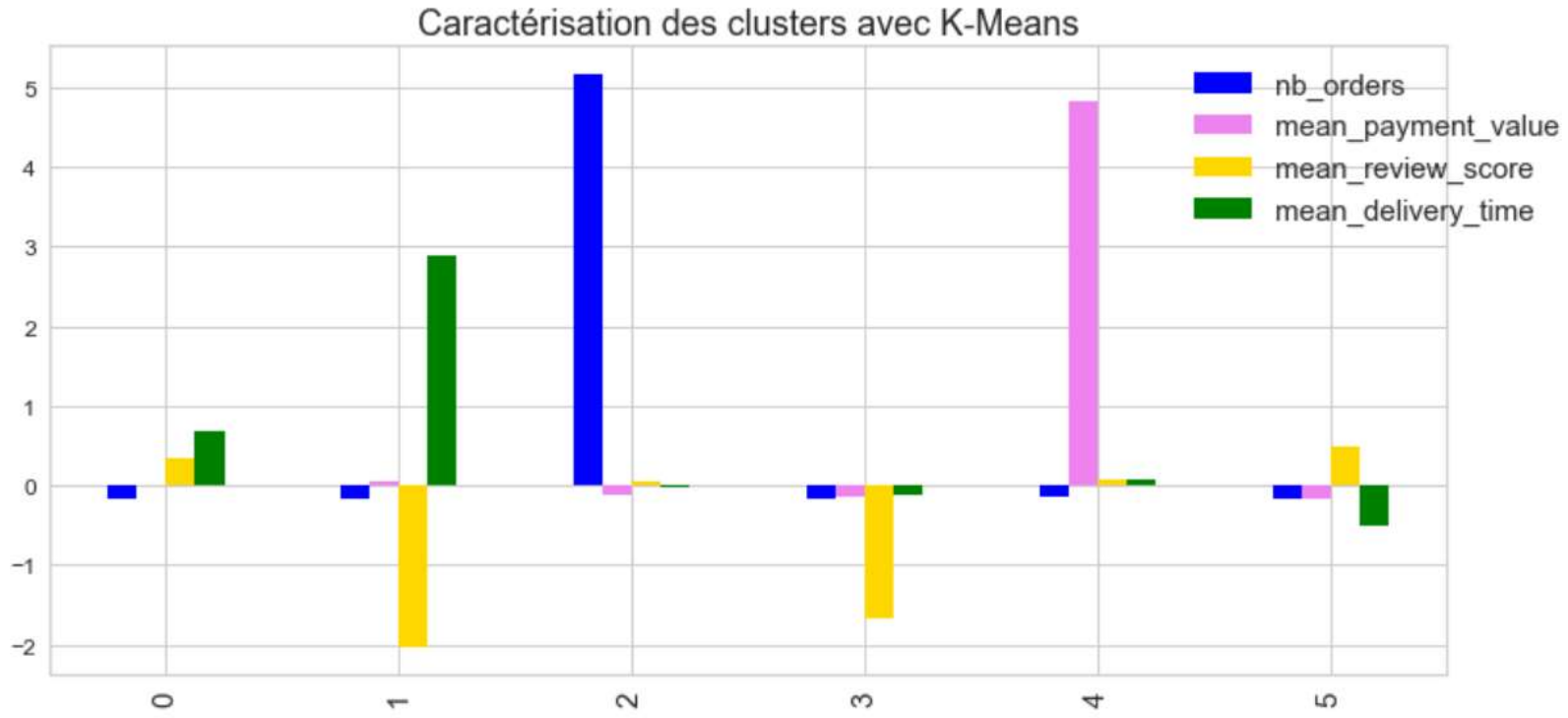




Visualisation clusters



Boxplot features



Caractérisation clusters

VI. Choix du modèle

- Importance du temps de calcul
- Clusters utilisables pour le problème métier
- Choix du K-Means avec 6 clusters

	K-Means 4 clusters	K-Means 6 clusters	K-Means 8 clusters	DBSCAN eps=0.3	DBSCAN eps=0.2	Agglomerative Clustering 6 clusters	Agglomerative Clustering 5 clusters
Score de Silhouette	0.547565	0.386366	0.390223	0.162327	0.090754	0.342169	0.333536
Score Davies Bouldin	0.804298	0.852013	0.820280	1.665363	1.775293	0.888254	0.974400
Temps de Calcul	0.939960	1.363340	1.889955	29.487150	15.057859	35.943649	39.912353

Clusters	Description	Nom	Proportion
Cluster 0	- Très contents - Dépenses dans la moyenne - Délais de livraison plus longs que la moyenne	**Prometteurs**	22%
Cluster 1	- Pas contents - Dépenses dans la moyenne - Délais de livraison très longs	**Mécontents**	5%
Cluster 2	- Contents - Beaucoup de commandes - Dépenses plus faibles que la moyenne - Délais de livraison dans la moyenne	**Fidèles**	3%
Cluster 3	- Pas contents - Dépenses plus faibles que la moyenne - Délais de livraison courts	**Râleurs**	15%
Cluster 4	- Contents - Dépenses très élevées - Délais de livraison dans la moyenne	**Dépensiers**	2%
Cluster 5	- Très contents - Dépenses plus faibles que la moyenne - Délais de livraison très rapides	**Lambdas**	53%

VII. Simulation de maintenance du modèle

- Evolution mensuelle des proportions des clusters
- Calcul du score ARI après 1,3,6,9 et 12mois
- Maintenance souhaitable après 6 mois

	1 mois	3 mois	6 mois	9 mois	12 mois
ARI	0.992213	0.990967	0.957353	0.640886	0.786978