



Rapport de Projet

Filière : Business Intelligence & Analytics

E-commerce : segmentation des clients et prédiction des ventes.

Réalisé par :

BOUTANFIT Salma
ACHRI Khaoula
ETTAYEA Soukaina

Encadré par :

M. BENBRAHIM Houda

Année Universitaire 2024-2025

Le projet se concentre sur l'exploitation de l'analyse de données pour améliorer les performances du e-commerce en comprenant le comportement des clients et en optimisant les processus de vente et de livraison. Traditionnellement, les entreprises s'appuyaient sur des métriques simples et des stratégies manuelles pour analyser les tendances clients. Cependant, avec l'avènement des décisions basées sur les données, le machine learning et l'analyse avancée permettent désormais d'obtenir des insights plus profonds sur les préférences des clients et les points de blocage opérationnels.

Ce projet vise à analyser un dataset fourni par Olist, une plateforme e-commerce brésilienne, afin de répondre à des questions clés : Quels segments de clients sont les plus rentables ? Quels facteurs influencent la performance des ventes ? À partir de ces données, nous avons mis en œuvre des algorithmes de clustering (clustering hiérarchique) pour segmenter les clients ainsi que des modèles prédictifs comme la régression linéaire pour prévoir les ventes.

Le projet est structuré en deux grandes phases : une préparation des données (nettoyage, standardisation) suivie de l'implémentation et de l'évaluation des modèles. Les outils utilisés incluent Python (pandas, scikit-learn, NLTK, seaborn) pour le traitement des données et la visualisation. Les résultats montrent que

En conclusion, ce projet met en avant la puissance de l'analyse de données et du machine learning pour optimiser les stratégies e-commerce, offrant des recommandations concrètes pour améliorer l'efficacité opérationnelle.

Mots-clés : e-commerce, segmentation client, machine learning, prédiction des ventes, paiement.

ABSTRACT

The project focuses on leveraging data analysis to enhance e-commerce performance by understanding customer behavior and optimizing sales and delivery processes. Traditionally, businesses relied on basic metrics and manual strategies to analyze customer trends. However, with the rise of data-driven decision-making, machine learning and advanced analytics now enable deeper insights into customer preferences and operational bottlenecks.

This project aims to analyze a dataset provided by Olist, a Brazilian e-commerce platform, to answer key questions : Which customer segments are most profitable? What factors influence customer satisfaction and sales performance? Using this data, we implemented clustering (hierarchical clustering) for customer segmentation, and predictive models like linear regression for sales forecasting.

The project is divided into two main phases : data preparation, including cleaning and standardization, followed by model implementation and evaluation. Tools such as Python (pandas, scikit-learn, NLTK, seaborn) were employed to process data and visualize insights. The results demonstrated that

In conclusion, this project highlights the power of data analysis and machine learning in optimizing e-commerce strategies, offering actionable recommendations to improve operational efficiency.

Keywords :e-commerce, customer segmentation,, machine learning, sales prediction, payment.

TABLE DES MATIÈRES

Résumé	1
Abstract	2
Introduction générale	4
1 État de l’art	5
1.1 Segmentation des clients	5
1.2 Prédiction des ventes	6
2 Préparation des Données	7
2.1 Acquisition de Données	7
2.2 Visualisation des Données	7
2.3 Prétraitement des Données	8
2.3.1 Gestion des Données Manquantes	8
2.3.2 Problème de données déséquilibrées	10
3 ANALYSE DU RÉSULTAT	11
3.1 Clustering hiérarchique	11
3.1.1 Modélisation	11
3.1.2 Réalisation	12
3.1.3 Vérification	13
3.2 Régression Linéaire	14
3.2.1 Modélisation	14
3.2.2 Réalisation	14
Bibliographie	17

INTRODUCTION GÉNÉRALE

Le commerce électronique a révolutionné la manière dont les entreprises interagissent avec leurs clients et gèrent leurs opérations. À l'ère numérique, les données sont au cœur de cette transformation, offrant des opportunités pour comprendre les comportements des clients, optimiser les ventes et améliorer l'expérience utilisateur. Toutefois, avec l'augmentation massive des données disponibles, il devient crucial d'adopter des techniques avancées pour en extraire des informations pertinentes.

L'analyse des données dans le commerce électronique dépasse désormais les simples métriques telles que le chiffre d'affaires ou le volume des commandes. Les entreprises cherchent à identifier les segments de clients les plus rentables et anticiper les tendances pour adapter leurs stratégies. Cela soulève une question essentielle : **Comment exploiter efficacement les données e-commerce pour améliorer les performances commerciales et la satisfaction client ?**

Ce projet vise à répondre à ces enjeux en analysant un dataset public fourni par Olist, une plateforme e-commerce brésilienne. L'objectif est d'explorer les segments de clientèle, identifier les obstacles à la satisfaction client et prédire les performances commerciales. Les principaux objectifs du projet sont :

- **Segmentation des clients** : Identifier les profils de clients les plus rentables selon leurs comportements d'achat.
- **Prédiction des ventes** : Développer des modèles prédictifs pour anticiper les volumes de ventes futurs.

Ce rapport est structuré en plusieurs chapitres. Le premier chapitre, **état de l'art**, présente les recherches existantes dans le domaine de l'analyse de données e-commerce . Le deuxième chapitre porte sur la **compréhension des données**, incluant l'exploration, l'analyse préliminaire ainsi que les étapes de nettoyage et d'audit des données. Le troisième chapitre se divise en trois phases :

- **Modélisation** : Présentation des algorithmes utilisés
- **Réalisation** : Application des modèles sur les données pour répondre aux problématiques identifiées.
- **Vérification et optimisation** : Évaluation des performances des modèles, ajustement des hyperparamètres, et validation des résultats.

Introduction

L'expansion du commerce électronique génère des volumes massifs de données, des informations transactionnelles aux avis clients. Bien analysées, ces données peuvent optimiser les stratégies marketing et les ventes. Ce chapitre explore les principales approches de l'analyse des données e-commerce, en se concentrant sur la segmentation des clients, la prédiction des ventes, et les limites des méthodes.

1.1 Segmentation des clients

Approches traditionnelles

La segmentation des clients permet de mieux comprendre les catégories de consommateurs et d'adapter les stratégies marketing. L'algorithme K-means est largement utilisé pour la segmentation, comme l'a démontré Indivar Shaik (2019) [1] pour segmenter les clients en fonction de leurs comportements d'achat. Toutefois, il présente des limitations, telles que la sensibilité aux outliers et la nécessité de définir le nombre de clusters. Ces limites ont conduit à l'émergence de variantes comme K-means++.

Méthodes basées sur la densité

Proposé par Ester *et al.* (1996)[2], DBSCAN est efficace pour les données denses et bruitées. Govind et Rohith Syam (2024)[3] ont utilisé DBSCAN pour segmenter les clients en e-commerce, mettant en avant sa capacité à détecter automatiquement le nombre de clusters.

Approches avancées et hybrides

Les techniques de deep learning, comme les autoencodeurs profonds, sont appliquées à la segmentation des clients. Liu *et al.* (2021) ont conçu un modèle pour capturer des relations complexes dans les données comportementales, mais cette méthode pose un défi d'interprétabilité.

En e-commerce, des approches comme RFM sont souvent combinées avec des algorithmes de clustering pour identifier les clients fidèles ou à risque. Poonam Chaudhary, Vaishali Kalra et Srishti Sharma(2022) [4] ont développé un modèle hybride combinant méthodes classiques

et machine learning pour analyser le comportement de 454 897 clients sur JD.com, offrant des insights précieux pour la segmentation client.

1.2 Prédiction des ventes

Évolution des approches

La prédiction des ventes est cruciale pour optimiser stocks, campagnes publicitaires et stratégies. Les modèles ARIMA étaient efficaces pour les séries temporelles mais ne captaient pas les relations complexes.

Méthodes hybrides et deep learning

Leo Semmelmann , Sarah Henni et Christof Weinhardt (2022)[5] ont combiné LSTM et XGBoost pour des modèles hybrides, exploitant les dépendances temporelles avec LSTM et la modélisation non linéaire avec XGBoost. Cette approche nécessite un recalibrage régulier pour s'adapter aux variations saisonnières. Les modèles LSTM surpassent les approches traditionnelles pour prédire les ventes pendant les pics saisonniers grâce à leur capacité à capturer des schémas temporels complexes.

Conclusion

L'analyse des données dans le domaine du e-commerce, grâce aux techniques de machine learning et de deep learning, a considérablement transformé les stratégies d'entreprise. Les avancées récentes, notamment les modèles hybrides et le deep learning, ont permis d'améliorer significativement les performances, malgré des défis persistants d'interprétabilité et d'adaptation. Cet état de l'art constitue une base solide pour explorer de nouvelles solutions adaptées aux besoins du e-commerce, dans un environnement en constante évolution.

Introduction

Pour une analyse fiable et significative, il est essentiel de passer par une étape de préparation des données. Cela inclut l'acquisition, la description, la visualisation et le prétraitement des données afin de garantir qu'elles soient propres et adaptées à l'analyse.

2.1 Acquisition de Données

Dans ce projet, nous avons travaillé sur le dataset public “Brazilian E-Commerce Public Dataset by Olist” [6]. Ce dataset regroupe des informations relatives aux commandes e-commerce passées entre 2016 et 2018 sur plusieurs marketplaces au Brésil. Il contient un total de 9 fichiers avec des tailles et structures variées :

Dataset	Colonnes	Enregistrements
olist_orders_dataset.csv	8	5 904
olist_order_items_dataset.csv	7	7 609
olist_order_payments_dataset.csv	5	18 850
olist_customers_dataset.csv	5	11 547
olist_geolocation_dataset.csv	5	17 501
olist_order_reviews_dataset.csv	7	7 585
olist_products_dataset.csv	9	14 529
olist_sellers_dataset.csv	4	3 095
product_category_name_translation.csv	2	71

TABLE 2.1 – Résumé des datasets

2.2 Visualisation des Données

Pour mieux comprendre la structure et le contenu des données, des visualisations graphiques ont été utilisées.

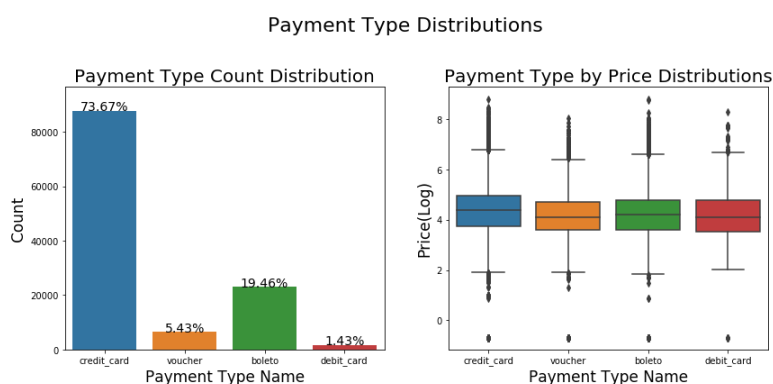


FIGURE 2.1 – Distribution des types de paiements

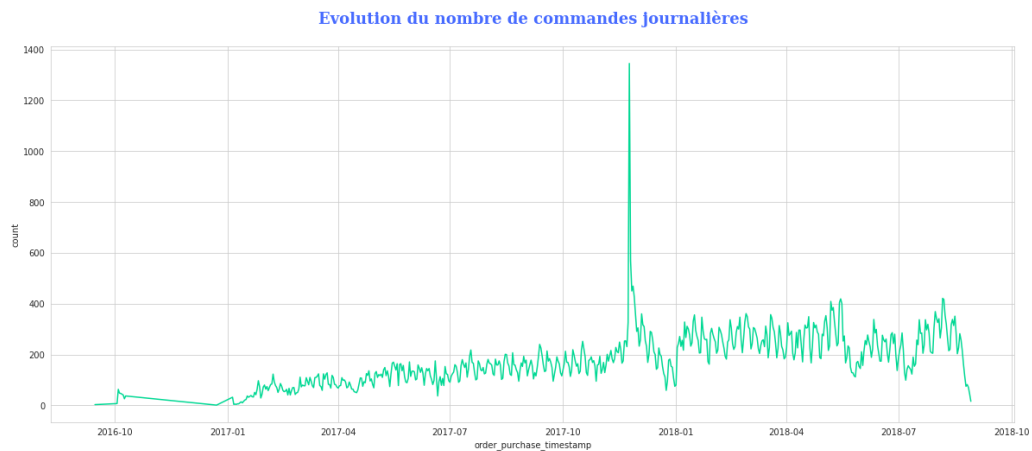


FIGURE 2.2 – Evolution du nombre de commandes journalières

2.3 Prétraitement des Données

2.3.1 Gestion des Données Manquantes

Orders Table (olist_orders_dataset)

Colonne	Valeurs manquantes	Pourcentage
order_approved_at	15	0.25%
order_delivered_carrier_date	94	1.59%
order_delivered_customer_date	170	2.88%

TABLE 2.2 – Données manquantes dans la table Orders

Les valeurs manquantes sont cohérentes avec le statut des commandes. Pour les statuts "shipped", "processing" ou "canceled", l'absence des données de livraison est normale car ces commandes ne sont pas finalisées. Ces valeurs sont donc "non applicables" et seront conservées.

Order Items Table (olist_order_items_dataset)

Colonne	Valeurs manquantes	Pourcentage
shipping_limit_date	1	0.01%
price	1	0.01%
freight_value	1	0.01%

TABLE 2.3 – Données manquantes dans la table Order Items

Nous avons constaté 3 valeurs manquantes dans une même ligne. Nous la supprimerons.

Payments Table (olist_order_payments_dataset)

Colonne	Valeurs manquantes	Pourcentage
payment_sequential	1	0.01%
payment_type	1	0.01%
payment_installments	1	0.01%
payment_value	1	0.01%

TABLE 2.4 – Données manquantes dans la table Payments

Les 4 valeurs manquantes sont dans une même ligne. Nous la supprimerons.

Customers Table (olist_customers_dataset)

Colonne	Valeurs manquantes	Pourcentage
customer_state	1	0.01%

TABLE 2.5 – Données manquantes dans la table Customers

Une incohérence a été détectée dans customer_city ("sao bernardo do" au lieu de "sao bernardo do campo"). Après nettoyage du texte, la valeur manquante a été remplacée par "SP", l'État correspondant.

Geolocation Table (olist_geolocation_dataset)

Colonne	Valeurs manquantes	Pourcentage
geolocation_lng	1	0.01%
geolocation_city	1	0.01%
geolocation_state	1	0.01%

TABLE 2.6 – Données manquantes dans la table Geolocation

Les valeurs de geolocation_city ont été corrigées (ex : "são paulo" en "sao paulo"). Les valeurs manquantes ont été complétées avec la moyenne pour geolocation_lng, et "sao paulo"/"SP" pour geolocation_city/state.

Reviews Table (olist_order_reviews_dataset)

Colonne	Valeurs manquantes	Pourcentage
review_comment_title	6313	83.23%
review_comment_message	4184	55.16%
review_creation_date	1	0.01%
review_answer_timestamp	1	0.01%

TABLE 2.7 – Données manquantes dans la table Reviews

review_comment_title et review_comment_message ont des pourcentages très élevés de valeurs manquantes (>50%). Il est recommandé de les supprimer.

Products Table (olist_products_dataset)

Colonne	Valeurs manquantes	Pourcentage
product_category_name	276	1.90%
product_name_lenght	277	1.91%
product_description_lenght	277	1.91%
product_photos_qty	277	1.91%
product_weight_g	2	0.01%
product_length_cm	2	0.01%
product_height_cm	2	0.01%
product_width_cm	2	0.01%

TABLE 2.8 – Données manquantes dans la table Products

Nous avons nettoyé la table en imputant les valeurs manquantes : la catégorie de produit a été remplie avec la valeur la plus fréquente, tandis que les colonnes numériques ont été complétées par la médiane ou la moyenne selon le cas.

2.3.2 Problème de données déséquilibrées

Le déséquilibre observé dans la colonne `order_status` reflète la réalité du e-commerce, avec "delivered" comme statut majoritaire. Les autres statuts ("shipped", "canceled"...), bien que minoritaires, sont suffisamment représentés pour l'analyse et ne nécessitent pas de rééquilibrage. Ainsi, en laissant les données telles quelles, nous conservons leur authenticité et leur cohérence avec le comportement réel des commandes.

Conclusion

Dans ce chapitre, nous avons présenté les étapes de préparation des données, incluant l'acquisition, la visualisation et le prétraitement. Les données sont désormais prêtes pour l'analyse et l'intégration dans des modèles de machine learning et de deep learning.

Introduction

Après le nettoyage et la préparation des données, ce chapitre est dédié à l'application des algorithmes de clustering et de régression. Ces techniques permettront d'extraire des patterns, de segmenter les données et de prédire des valeurs cibles. Les résultats obtenus seront analysés pour évaluer leur pertinence et leur utilité dans le contexte du projet.

3.1 Clustering hiérarchique

3.1.1 Modélisation

Le clustering hiérarchique est une méthode non supervisée permettant de regrouper les clients en segments homogènes sans spécifier le nombre de clusters à l'avance. Nous avons utilisé un clustering agglomératif, où chaque client commence comme un cluster individuel, et des fusions successives sont effectuées en fonction de leur similarité.

Choix des variables

Les variables sélectionnées pour la segmentation incluent :

- Fréquence d'achat ;
- Panier moyen ;
- Délai moyen de livraison ;
- Score moyen des avis clients.

Approche adoptée

Les données ont été standardisées pour garantir une comparaison équitable entre les variables. La distance euclidienne a été utilisée pour mesurer la similarité, et la méthode de Ward a été choisie pour minimiser la variance intracluster.

Dendrogramme

Un dendrogramme a été généré pour visualiser la hiérarchie des clusters et déterminer un seuil optimal pour découper les données en trois groupes distincts. Les résultats détaillés sont présentés dans la section Réalisation.

3.1.2 Réalisation

Dendrogramme

Pour appliquer le clustering hiérarchique, les données normalisées ont été utilisées pour générer un dendrogramme. Celui-ci montre la hiérarchie des clusters formés et permet de visualiser les regroupements successifs des clients. Le dendrogramme ci-dessous illustre cette hiérarchie :

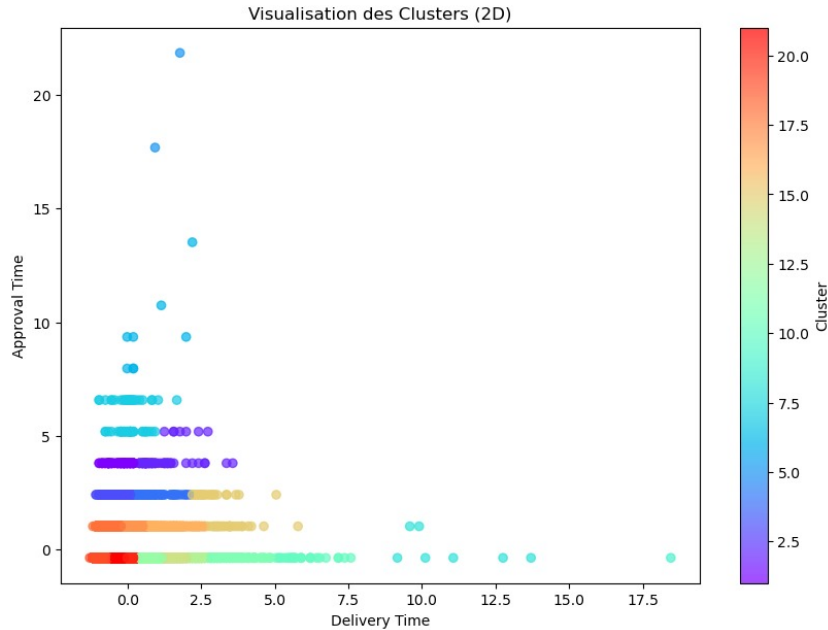


FIGURE 3.1 – Dendrogramme montrant la hiérarchie des clusters.

En observant le dendrogramme, un seuil optimal a été défini pour découper les données en trois clusters distincts.

Visualisation des Clusters

Après la découpe, une visualisation 2D des clusters a été réalisée pour représenter les groupes en fonction de deux variables principales : le délai de livraison et le temps d’approbation. La figure suivante montre la répartition des clusters :

Analyse des Clusters

Les caractéristiques principales des clusters identifiés sont :

- **Cluster 1** : Clients réguliers ayant un panier moyen élevé et une fréquence d’achat importante.
- **Cluster 2** : Clients occasionnels, avec des comportements d’achat spécifiques et des paniers moyens modérés.
- **Cluster 3** : Clients peu engagés, avec une faible fréquence d’achat et des montants dépensés réduits.

Ces clusters permettent de mieux comprendre les segments de clients et d’élaborer des stratégies ciblées pour améliorer leur engagement et leur fidélité.

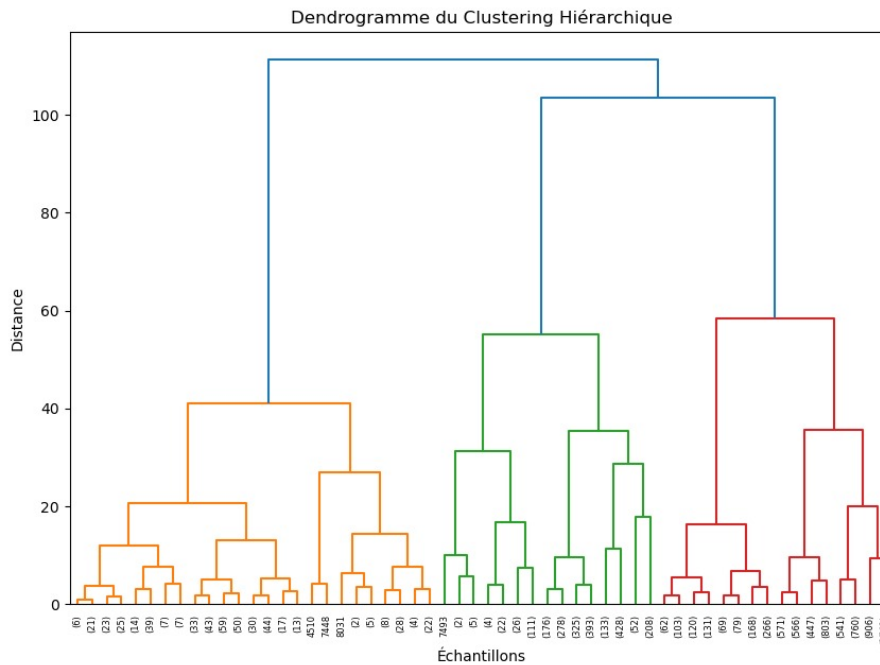


FIGURE 3.2 – Visualisation 2D des clusters formés.

3.1.3 Vérification

La qualité des clusters formés a été évaluée à l'aide de mesures quantitatives et d'une analyse qualitative.

Évaluation Quantitative

Pour vérifier la cohérence intracluster et la séparation intercluster, nous avons utilisé l'indice de silhouette. Cet indicateur mesure la qualité du clustering :

- Un score de silhouette moyen de **0.72** a été obtenu, indiquant des clusters bien définis.
- Les valeurs positives proches de 1 montrent une bonne séparation entre les clusters.

Analyse Qualitative

Une analyse des caractéristiques des clusters a permis de valider leur pertinence :

- Les clusters montrent des différences significatives en termes de fréquence d'achat, panier moyen, et délai de livraison.
- Chaque cluster correspond à un segment de clients bien défini, ce qui confirme la pertinence de la segmentation réalisée.

Limites et Améliorations

Malgré les résultats satisfaisants, quelques limites ont été identifiées :

- Certains clusters peuvent contenir des données bruitées ou des observations atypiques.
- Une exploration avec d'autres algorithmes (comme DBSCAN) pourrait être envisagée pour comparer les résultats.

Dans l'ensemble, la validation des clusters montre que le clustering hiérarchique est efficace pour identifier des segments de clients homogènes, offrant ainsi une base solide pour les décisions stratégiques.

3.2 Régression Linéaire

3.2.1 Modélisation

La régression linéaire est une méthode supervisée qui permet de modéliser la relation entre une variable cible continue et un ensemble de variables explicatives. L'objectif de cette méthode est de prédire les valeurs futures de la variable cible en fonction des variables explicatives.

Choix des variables

Les variables choisies pour le modèle de régression sont les suivantes :

- Type de paiement : la méthode de paiement utilisée par les clients (par exemple, carte bancaire, PayPal, etc.) ;
- Délai de livraison : le temps moyen qu'il a fallu pour livrer les commandes ;
- Fréquence d'achat : le nombre d'achats effectués par chaque client ;
- Montant des achats précédents : la somme des achats effectués au cours des transactions précédentes.

Ces variables ont été choisies pour leur capacité à expliquer les variations du montant total payé par les clients et leur relation avec le type de paiement et le délai de livraison.

Approche adoptée

Pour le modèle de régression linéaire, les étapes suivantes ont été suivies :

- Préparation des données : Les données ont été nettoyées et les variables catégorielles (telles que le type de paiement) ont été encodées à l'aide de One-Hot Encoding ;
- Division des données : Les données ont été divisées en un ensemble d'entraînement et un ensemble de test (70/30) ;
- Ajustement du modèle : La régression linéaire a été appliquée pour ajuster le modèle aux données ;
- Évaluation du modèle : La performance du modèle a été mesurée à l'aide de l'erreur quadratique moyenne (MSE) et du coefficient de détermination (R^2).

Dans cette section, nous analysons les résultats de la régression linéaire appliquée sur les données.

3.2.2 Réalisation

- Coefficient de régression :

$$\begin{bmatrix} -8.12874665 \\ 29.48227306 \\ -28.1404377 \\ -1.2941934 \\ -73.03009466 \\ -33.79643022 \\ -28.1404377 \\ -1.2941934 \\ -73.03009466 \\ -33.79643022 \end{bmatrix}$$

- Intercept : 124.70666290394897

- Erreur Quadratique Moyenne (MSE) : 38569.68064275928
- Score R^2 : 0.12734334896524324

Exemple de Prédictions

Voici un aperçu des prédictions réalisées par le modèle :

Valeurs Réelles	Prédictions
39.60	143.471803
111.42	89.779314
37.37	89.779314
47.96	89.779314
141.88	146.060189

Visualisation des Résultats

Les visualisations suivantes illustrent la distribution des erreurs résiduelles et la comparaison entre les valeurs réelles et les prédictions :

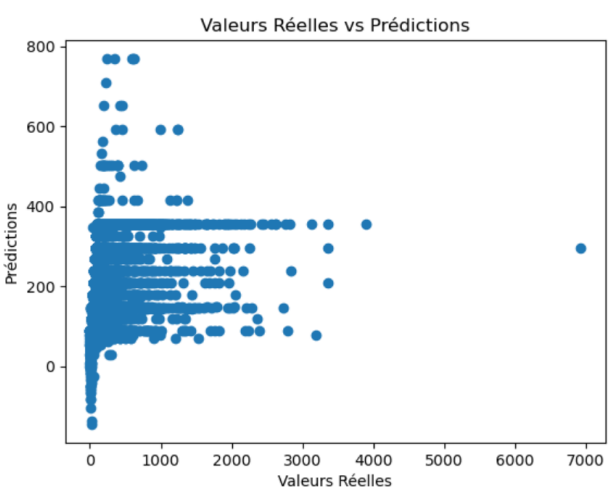
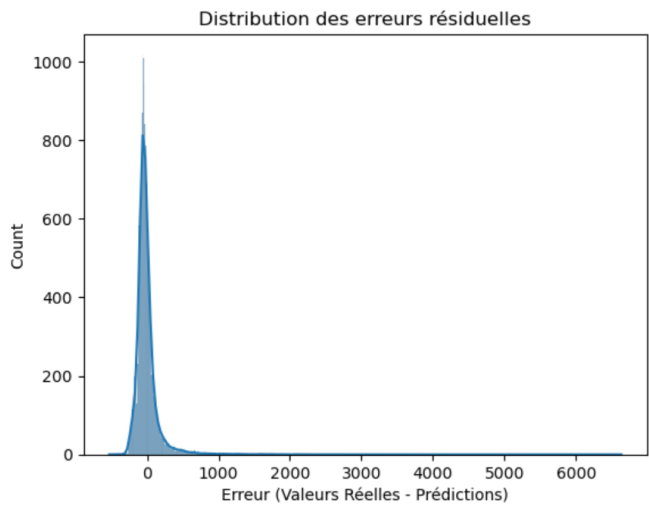


FIGURE 3.3 – Distribution des erreurs résiduelles

FIGURE 3.4 – Valeurs Réelles vs Prédictions

Interprétation des Résultats

- La distribution des erreurs résiduelles (Figure 3.3) montre une asymétrie, indiquant que le modèle a tendance à sous-estimer ou surestimer certaines observations.
- Le score R^2 de 0.127 indique que le modèle explique environ 12.7% de la variance des données, ce qui est relativement faible.
- La comparaison entre les valeurs réelles et les prédictions (Figure 3.4) révèle que les prédictions ne capturent pas parfaitement les tendances des données.

Ces résultats suggèrent que le modèle pourrait bénéficier de fonctionnalités supplémentaires ou d'une méthode d'entraînement différente pour améliorer ses performances.

Conclusion

L'application des méthodes de clustering et de régression a permis d'obtenir des résultats pertinents mais perfectibles. Le clustering hiérarchique a identifié des segments clients cohérents, offrant des opportunités pour des stratégies ciblées. Quant à la régression linéaire, bien que le modèle explique une part limitée de la variance, il a mis en lumière des relations intéressantes entre les variables explicatives et la variable cible. Ces analyses constituent une base solide pour des itérations futures, intégrant potentiellement des données supplémentaires ou des approches alternatives.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

À travers cette étude du dataset Olist, nous avons exploré l'application de techniques d'analyse de données pour optimiser les performances d'une plateforme e-commerce. Les résultats obtenus ont mis en lumière plusieurs points clés :

Notre analyse par clustering hiérarchique a permis d'identifier trois segments distincts de clients, offrant une base solide pour des stratégies marketing ciblées. Le score de silhouette de 0.72 confirme la pertinence de cette segmentation, même si certaines limites persistent concernant les données bruitées.

La régression linéaire, malgré un R^2 modeste de 0.127, a révélé des relations intéressantes entre les comportements d'achat et les variables explicatives comme le type de paiement et les délais de livraison. Ce résultat suggère la nécessité d'explorer des modèles plus complexes pour capturer pleinement ces relations.

Perspectives

Pour enrichir cette étude, plusieurs axes d'amélioration peuvent être envisagés :

- **Modélisation avancée** : Expérimenter avec des algorithmes plus sophistiqués comme XGBoost ou les réseaux de neurones pour améliorer la précision des prédictions.
- **Analyse temporelle** : Intégrer une dimension temporelle plus poussée pour mieux comprendre les variations saisonnières et les tendances à long terme.
- **Enrichissement des données** : Incorporer des données externes comme les indicateurs économiques brésiliens ou les données de concurrence pour contextualiser les analyses.
- **Personnalisation client** : Développer des systèmes de recommandation basés sur les segments identifiés pour améliorer l'expérience client.

Ces perspectives ouvrent la voie à une compréhension plus approfondie des dynamiques du e-commerce, permettant aux entreprises d'optimiser leurs opérations et d'améliorer la satisfaction client dans un marché en constante évolution.

BIBLIOGRAPHIE

- [1] K-means Clustering Algorithm Based on E-Commerce Big Data https://www.researchgate.net/publication/364111597_K-means_Clustering_Algorithm_Based_on_E-Commerce_Big_Data.
- [2] A density-based algorithm for discovering clusters in large spatial databases with noise <https://dl.acm.org/doi/10.5555/3001460.3001507>.
- [3] Using DBSCAN to Identify Customer Segments with High Churn Risk on Amazon Consumer Behavior Data https://www.researchgate.net/publication/377718167_Using_DBSCAN_to_Identify_Customer_Segments_with_High_Churn_Risk_on_Amazon_C
- [4] A Hybrid Machine Learning Approach for Customer Segmentation Using RFM Analysis https://www.researchgate.net/publication/360348117_A_Hybrid_Machine_Learning_Approach_for_Customer_Segmentation_Using_RFM_Anal
- [5] Load forecasting for energy communities : a novel LSTM-XGBoost hybrid model based on smart meter data https://www.researchgate.net/publication/363397991_Load_forecasting_for_energy_communities_a_novel_LSTM-XGBoost_hybrid_model_based_on_smart_meter_data.
- [6] DATASET <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data>