République Française
MINISTERE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE
LA RECHERCHE SCIENTIFIQUE

Université Claude Bernard Lyon 1
Master Data Science

Projet Data Mining

Thème:

Analyse et exploration des données d'une entreprise de e-commerce système de recommandation

Réalisé par:

- SIMOUD Achour
- El KAOUT Mohamed Amine

Année universitaire 2023- 2024

Table des matières :

Projet Data Mining

1. Introduction :

Le secteur du commerce électronique a connu une croissance exponentielle au cours des dernières décennies, transformant la manière dont les consommateurs interagissent avec les produits et services. Dans ce contexte dynamique, l'importance de comprendre les comportements d'achat des clients et d'améliorer l'expérience utilisateur est cruciale pour la réussite d'une entreprise de commerce électronique.

C'est dans cette optique que ce projet s'inscrit, se concentrant sur l'analyse et l'exploration des données d'une entreprise de commerce électronique, avec un accent particulier sur le développement d'un système de recommandation.

L'idée derrière ce projet est d'anticiper les préférences des clients, afin d'améliorer significativement leur expérience d'achat. Nous appliquerons la méthode de détection de patterns fréquents que nous avons abordé en cours. Cette étude permettrait de personnaliser l'expérience des clients en leur proposant des produits plus pertinents selon leur profil, ainsi augmenter le taux de conversion des visiteurs du store.

2. Description des données :

Le dataset s'agit des données de transactions survenues entre le 01/12/2010 et le 09/12/2011, pour une entreprise de e-commerce basée au Royaume unis mais sans magasin physique. La société vend principalement des cadeaux pour différentes occasions, une majorité des clients de cette entreprise sont des grossistes.

2.1 Format des données :

Les données se présentent sous forme d'un tableau contenant les numéros de facture, le code des produits, leurs descriptions, la quantité achetée, la date de la transaction, le prix unitaire, l'identifiant du client ainsi que son pays.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Figure 1 : Echantillon des données

3. Nettoyage des Données :

Le processus de nettoyage des données est crucial pour garantir la qualité et la fiabilité des analyses ultérieures. Dans notre ensemble de données, plusieurs problèmes nécessitent une attention particulière, tels que les identifiants clients manquants, les descriptions divergentes pour le même produit, et la répétition de codes de produit. Voici comment nous avons résolu ces problèmes :

3.1 Traitement des Identifiants Clients Manquants :

Les transactions ou le l'identifiant client est manquant ont tout simplement été supprimées du jeu de données.

3.2 Consolidation des Descriptions de Produits :

Plusieurs code produits sont associés avec des descriptions différentes, pour résoudre ce problème nous avons d'abord passer toutes les descriptions en majuscule nous avons filtré les caractères spéciaux. Ensuite pour chaque code produit nous avons remplacé sa description par celle qui était la plus fréquente, cela permet de réduire les divergence causée par des erreurs de saisie humaine.

3.3 Élimination des Codes de Produit Répétés :

Une fois les descriptions corrigées, on remarque qu'il existe des code produit différents qui désignent le même produit, nous avons utilisé le même raisonnement que pour les description, nous avons attribué à chaque produit le code le plus fréquemment utilisé.

3.4 Élimination des transactions qui sont pas des ventes :

Il existe des transactions enregistrées qui sont en réalité des frais bancaires, des frais de transport etc, ces transactions n'étant pas pertinentes dans le cadre de notre projet nous avons décidé de les éliminer.

4- Analyse géographique :

Pour comprendre la distribution des ventes de l'entreprise nous avons réalisé une analyse géographique sur les données.
Nous avons identifié les 3 pays qui font le plus de ventes, ainsi que les 3 produits les plus vendus par pays. La figure suivante illustre les résultats obtenus :

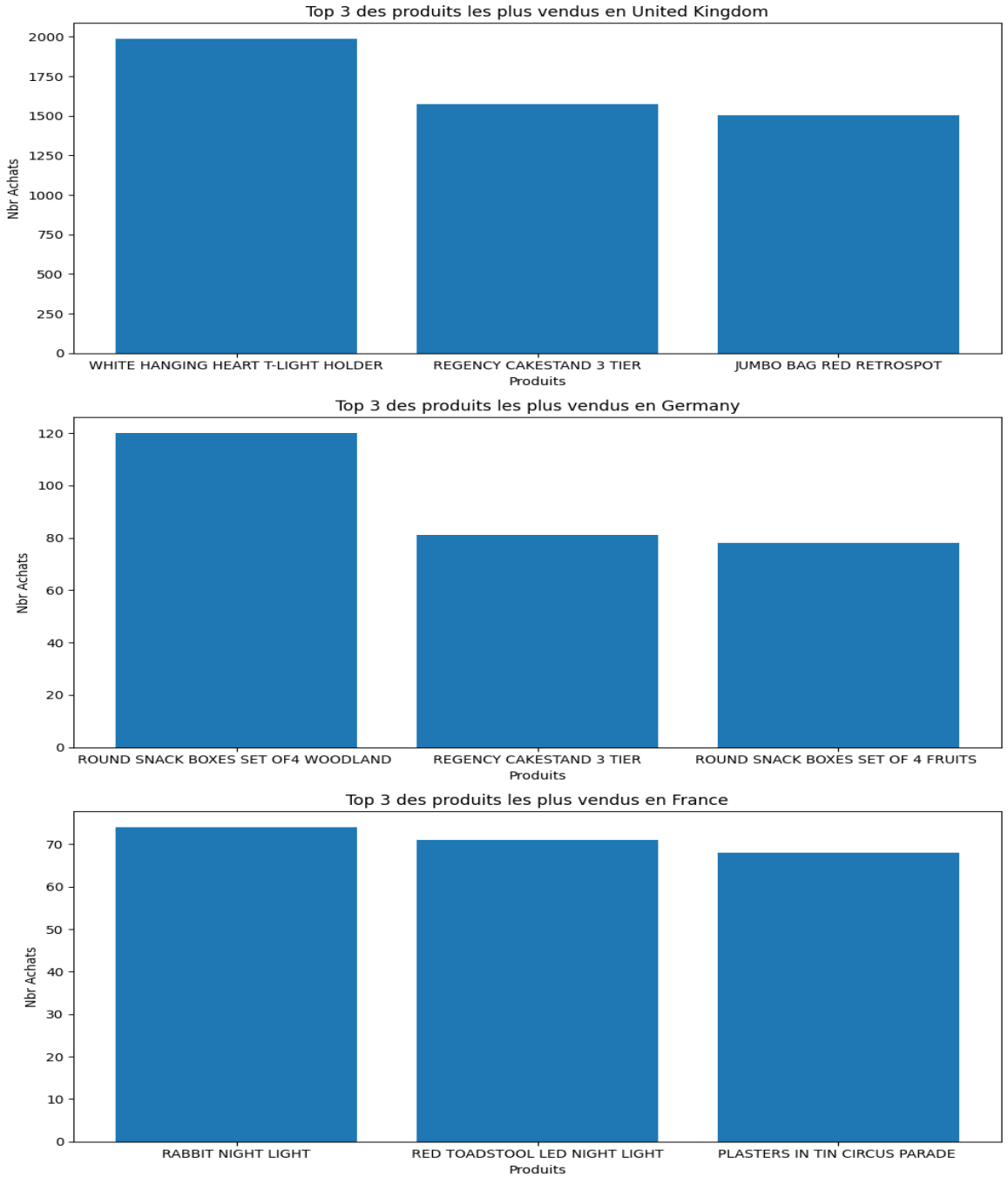


Figure 2 : Les produits les plus achetés par pays.

D'après la figure on s'aperçoit que le Royaume unis est le leader en vente ce qui logique vu que l'entreprise est implantée là bas, ensuite les deux pays qui achètent le plus sont des pays voisins au Royaume unis. Cette analyse permet d'améliorer la stratégie de marketing soit en concentrant les efforts sur les pays qui sont déjà de très bons acheteurs soit en décidant de scaler en essayant d'atteindre d'autres pays.

5- Analyse temporelle :

Dans la même optique, nous avons aussi réalisé une analyse temporelle de l'activité de l'entreprise en nous intéressant à l'évolution des ventes sur une période d'une année pour essayer de repérer des périodes clés.
La figure suivante représente l'évolution des ventes en fonction du temps:

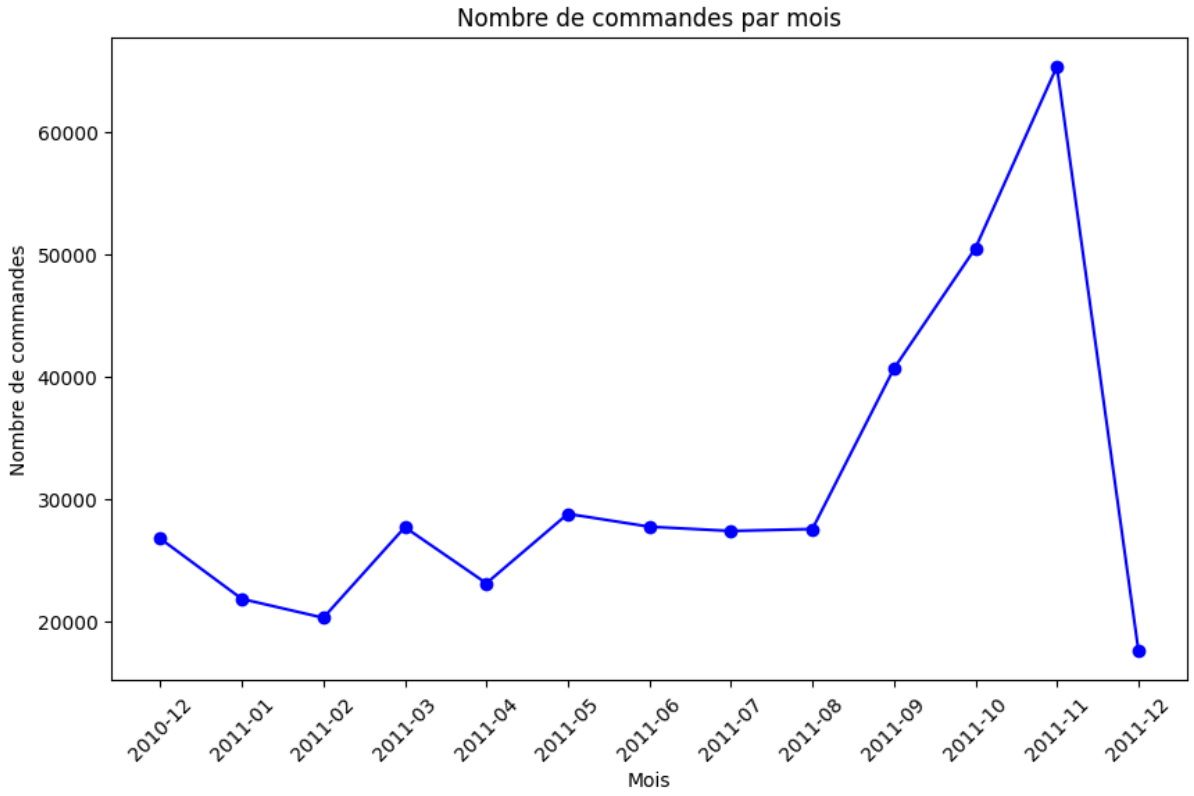


Figure 3 : Evolution du nombre de commandes au cours du temps

La figure montre clairement une hausse importante de ventes entre le mois de septembre et décembre, cette hausse peut être expliquée par la nature des produits vendus par l'entreprise qui sont principalement des cadeaux, donc la période correspond au moment où les acheteurs remplissent leurs stocks pour les fêtes de fin d'année.
Ces résultats peuvent être pertinents pour améliorer la gestion des stocks et anticiper les périodes de forte demande.

6- Système de recommandation de produits :

Le système de recommandation de produits se base sur une détection de patterns fréquents dans les commandes qui ont été faites par les clients.
Ces méthode nous permet d'obtenir des couples triplets de produits qui sont souvent achetés ensemble, elle offre la possibilité d'optimiser le système de recommandation selon deux leviers :

a- Recommander les produits les plus adaptés à un client donné :

Cela peut se faire en recherchant les conséquents engendrés par les produits déjà achetés par le client et lui proposer ceux avec le score le plus élevé.
Cette figure représente un exemple de cette utilisation pour un produit donné.

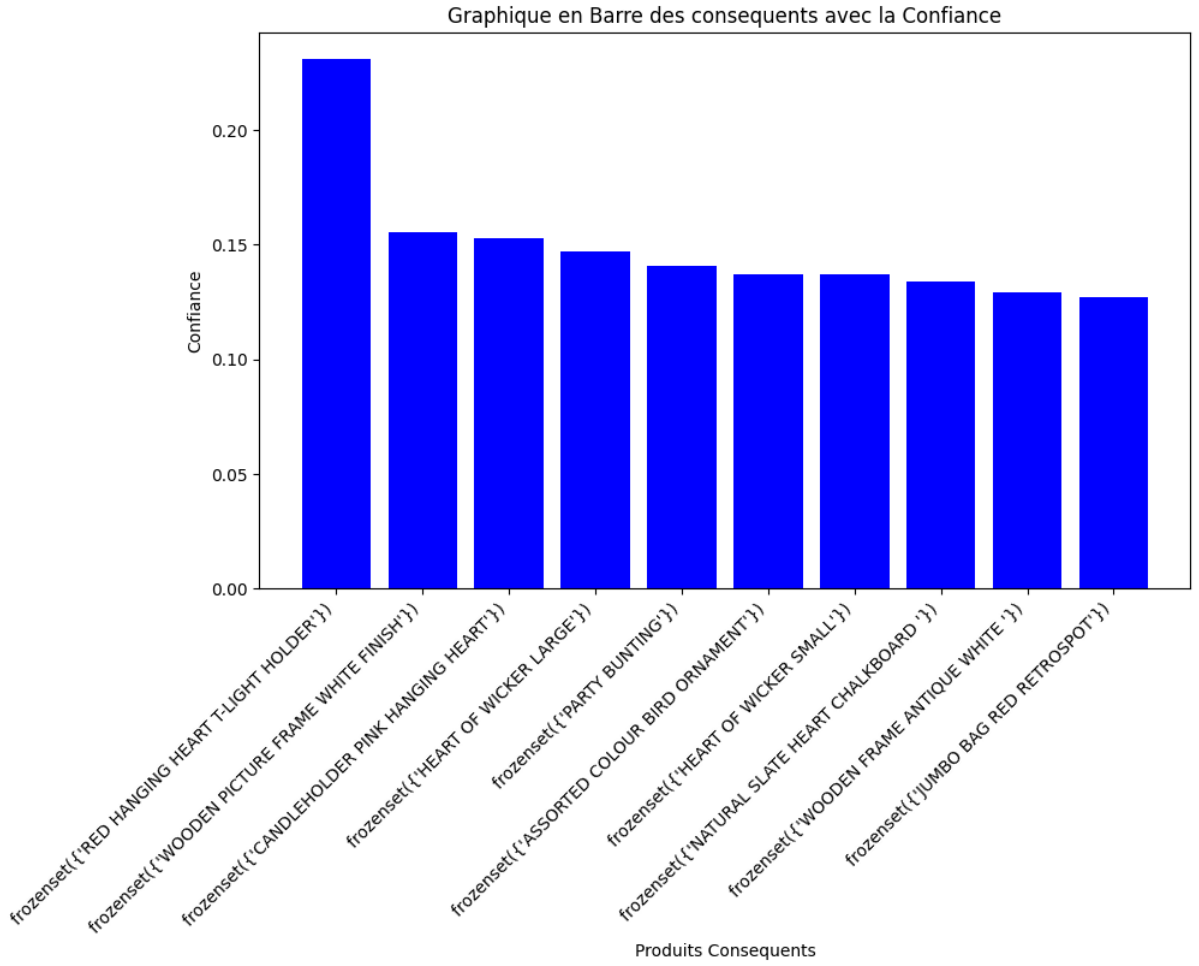
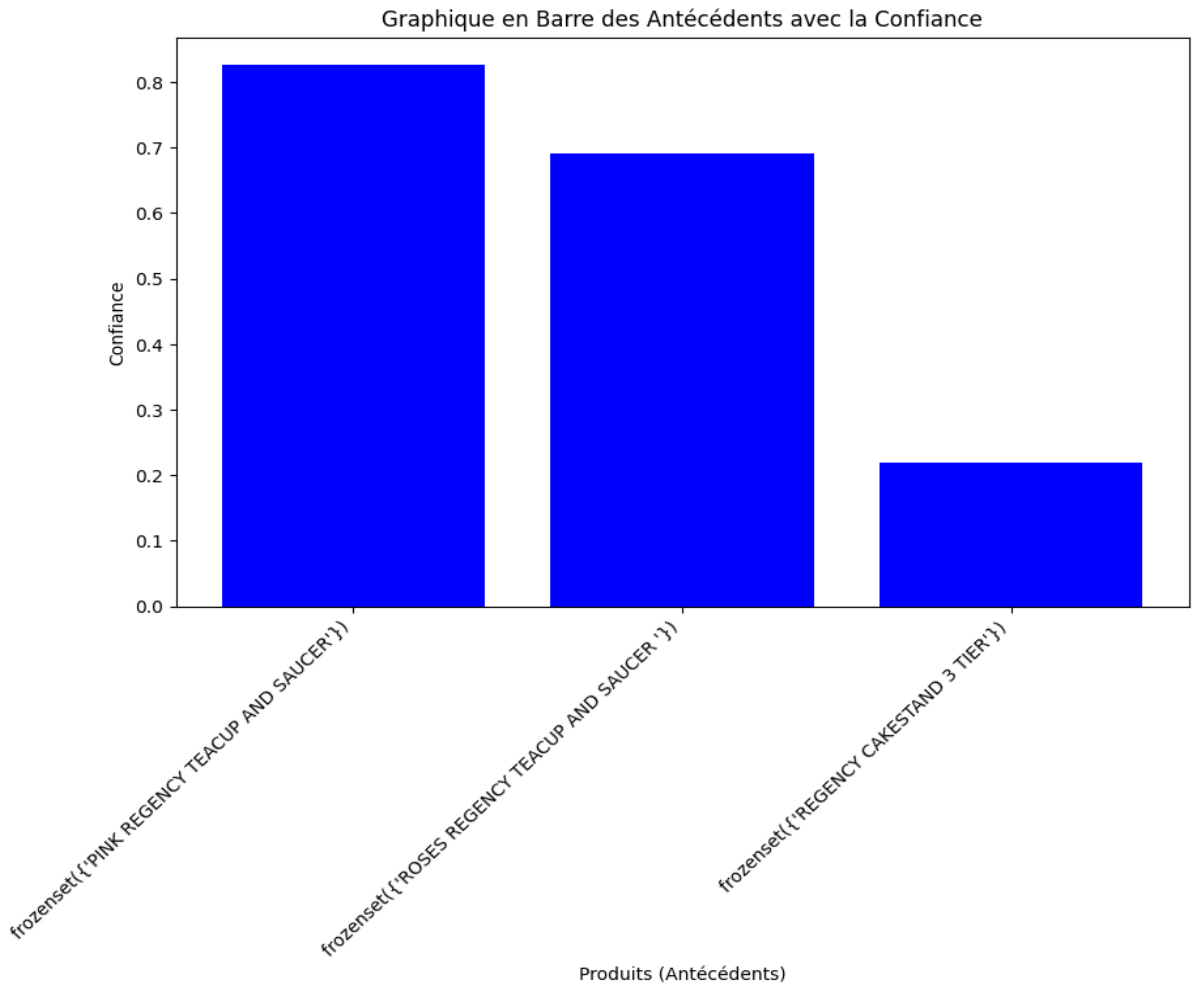


Figure 4 :

b- Trouver les acheteurs potentiels d'un produit :

Cette approche consiste à déterminer les clients potentiels pour un certain produit, pour se faire il suffit de chercher les antécédents avec les scores les plus élevés pour un consequent donné et on choisit les clients qui ont acheté ces produits antécédents.
Les résultats obtenus sont illustrés dans la figure suivante.



7- Conclusion :

Au cours de ce projet nous avons eu l'opportunité de travailler avec un jeu de données réelles et nous confronter aux problématiques que ce genre de données peuvent présenter telles que les erreurs de saisie, les informations manquantes etc. Il nous a aussi permis de mettre en pratique la notion de la détection de patterns fréquents que nous avons vu en cours et d'apprendre à interpréter les résultats auxquels cette méthode aboutit.
Au terme de ce travail nous avons retenu quelques perspectives qu'il serait intéressant d'explorer, comme la mise en place d'un système de recommandation en se basant sur les profils des clients, la mise en oeuvre d'un algorithme pour catégoriser les produits et qui pourrait ainsi permettre de détecter les tendance de consommation des clients ou même des pays.
Au terme de ce projet, nous avons retenu quelques perspectives qu'il serait intéressant d'explorer, comme la fusion des données géographique au système de recommandation qui prendrait en compte les tendances fréquentes dans chaque pays, la catégorisation des produits qui permettrait de détecter les préférences des clients. Ces perspectives permettent d'affiner plus le système de recommandation ainsi proposer une meilleur expérience pour les clients de l'entreprise.