Test Technique

Élaborée par :

Omar M’rad

Introduction:

Dans un monde de plus en plus connecté, le commerce de détail en ligne a connu une croissance exponentielle, offrant aux consommateurs un accès facile à une multitude de produits et aux entreprises une portée mondiale sans précédent. Dans ce contexte dynamique, l'analyse des données devient un outil essentiel pour comprendre les tendances du marché, optimiser les stratégies de vente et offrir une expérience client personnalisée.

L'ensemble de données fourni pour cette analyse représente un instantané des transactions effectuées par un détaillant en ligne basé au Royaume-Uni sur une période allant du 12/01/2010 au 12/09/2011. Comprenant plus de 500 000 instances, ces données offrent une opportunité précieuse d'explorer les dynamiques des ventes en ligne dans un environnement transnational.

Ce rapport vise à examiner en profondeur ces données pour identifier les principaux moteurs de vente, les tendances saisonnières, les segments de clients les plus rentables, et les opportunités d'amélioration des performances commerciales. À travers une combinaison d'analyse exploratoire des données et de visualisations interactives, nous chercherons à fournir des insights pertinents pour guider les décisions stratégiques de l'entreprise.

Au cours des sections suivantes, nous examinerons en détail la structure des données, procéderons à une exploration approfondie des tendances et des modèles, et présenterons un tableau de bord interactif basé sur **Power BI** pour synthétiser les résultats de l'analyse.

En mettant l'accent sur l'optimisation des performances commerciales et l'amélioration de l'expérience client, cette analyse aspire à fournir une base solide pour la prise de décision éclairée dans le domaine du commerce de détail en ligne.

1. Compréhension des données:

Les données fournies dans cet ensemble de données transnationales représentent toutes les transactions effectuées par un détaillant en ligne basé au Royaume-Uni entre le 12/01/2010 et le 12/09/2011.

Notre jeu de données est composé de 541 908 lignes et contient 8 colonnes, telles que:

- InvoiceNo (Catégorique) : Un identifiant attribué à chaque transaction, avec les annulations signalées par un préfixe "c".
- StockCode (Catégorique) : Un identifiant attribué à chaque produit distinct.
- Description (Catégorique) : Le nom du produit.
- Quantity (Entier) : Les quantités de chaque produit par transaction.
- InvoiceDate (Date) : La date et l'heure de chaque transaction.
- UnitPrice (Continu) : Le prix du produit par unité en livres sterling.
- CustomerID (Catégorique) : Un identifiant attribué à chaque client.
- Country (Catégorique) : Le pays où réside chaque client.

Ces données fournissent une vision détaillée des transactions, permettant une analyse approfondie des performances de vente en ligne dans diverses dimensions telles que la temporalité, la quantité, les produits et les clients.

	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
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France

541909 rows × 8 columns

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	541909.000000	541909	541909.000000	406829.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114	15287.690570
min	-80995.000000	2010-12-01 08:26:00	-11062.060000	12346.000000
25%	1.000000	2011-03-28 11:34:00	1.250000	13953.000000
50%	3.000000	2011-07-19 17:17:00	2.080000	15152.000000
75%	10.000000	2011-10-19 11:27:00	4.130000	16791.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	218.081158	NaN	96.759853	1713.600303

2. Préparation des données:

La préparation des données est une étape essentielle avant d'entreprendre toute analyse approfondie. Dans cette section, nous détaillerons les étapes de nettoyage et de prétraitement des données effectuées sur l'ensemble de données des ventes au détail en ligne.

2.1 Suppression les doublons:

```
count     541909
unique         2
top        False
freq      536641
dtype: object

True
```

Dans notre jeu de données Data_Retail, nous avons trouvé 5 268 doublons. Après les avoir supprimés, nous avons obtenu 536 641 lignes uniques.

Après cela, nous avons mis à jour les index à l'aide de la fonction reset_index.

2.2 Traitement de colonnes:

Dans cette section, nous décrivons les étapes spécifiques que nous avons entreprises pour traiter chaque colonne de l'ensemble de données des ventes au détail en ligne:

2.2.1 InvoiceNo:

On a vérifié s'il y avait des annulations de factures à supprimer, mais dans notre cas, nous n'en avons pas trouvé.

```
The 'InvoiceNo' column does not contain 'c'.
```

2.2.2 Quantity:

Nous avons trouvé des valeurs négatives dans la colonne Quantité, comme le montre la figure :

```
141       -1
154       -1
235      -12
236      -24
237      -24
           ..
535188   -11
536280    -1
536447    -5
536448    -1
536449    -5
Name: Quantity, Length: 10587, dtype: int64
```

Puisque la colonne Quantity est de type entier et ne devrait contenir que des valeurs positives, nous devons donc les supprimer.

2.2.3 Price:

Le même traitement s'applique à la colonne Price : elle ne doit pas contenir de valeurs négatives ni de zéros.

2.2.4 StockCode:

Tout d'abord, nous avons vérifié les valeurs de StockCode. Ensuite, nous avons constaté que certaines valeurs ne respectaient pas le format de données attendu, comme un numéro intégral à 5 chiffres. Nous avons donc effectué des analyses et des observations, puis appliqué des conditions en utilisant des expressions régulières pour nettoyer les données.

2.3 Détection valeurs Manquantes

Nous avons vérifié l'ensemble de données pour détecter la présence de valeurs manquantes dans chaque colonne, et nous avons constaté qu'il y en avait dans la colonne CustomerID:

```
Missing values in each column after cleaning customerID :
 InvoiceNo          0
StockCode           0
Description         0
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     131354
Country             0
dtype: int64
```

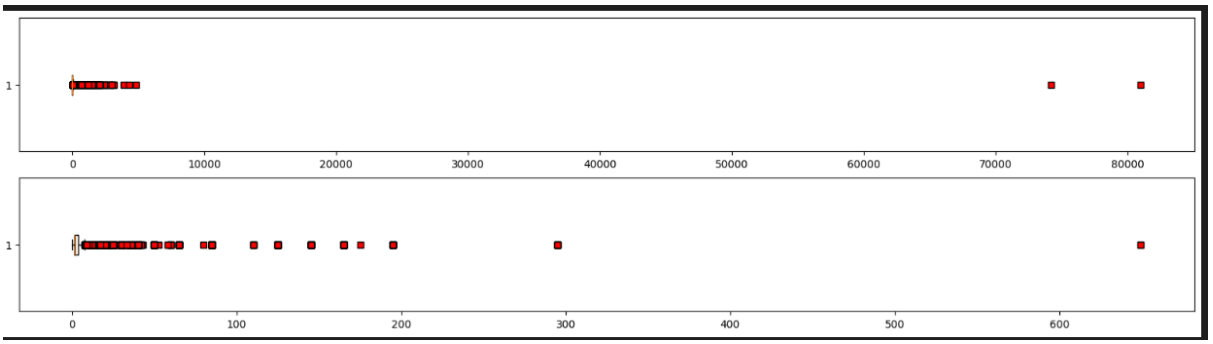
Donc, nous avons supprimé les lignes qui contiennent des valeurs NaN dans la colonne CustomerID.

2.4 Détection des valeurs aberrantes:

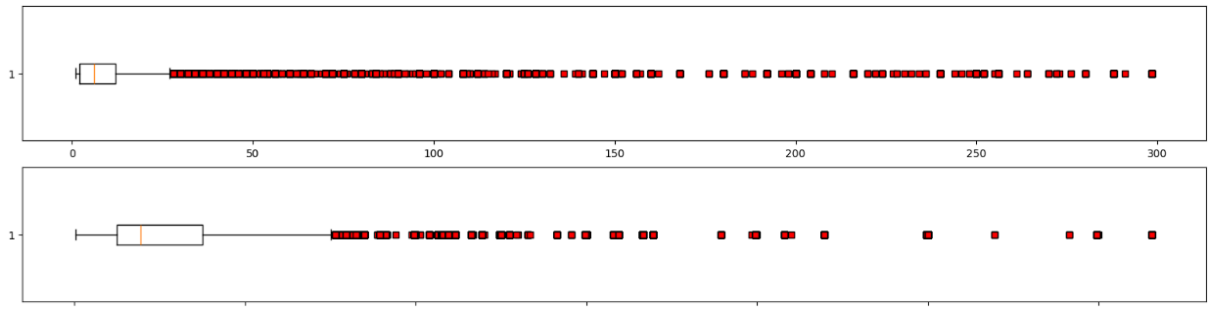
Nous avons également examiné les valeurs extrêmes ou aberrantes dans les variables quantitatives telles que la quantité et le prix unitaire.

Tout d'abord, nous avons effectué une visualisation à l'aide de subplots pour les détecter. Nous avons identifié quelques valeurs aberrantes, puis nous avons appliqué la méthode des quantiles (0.01 et 0.99) pour les éliminer.

- Avant l'élimination:



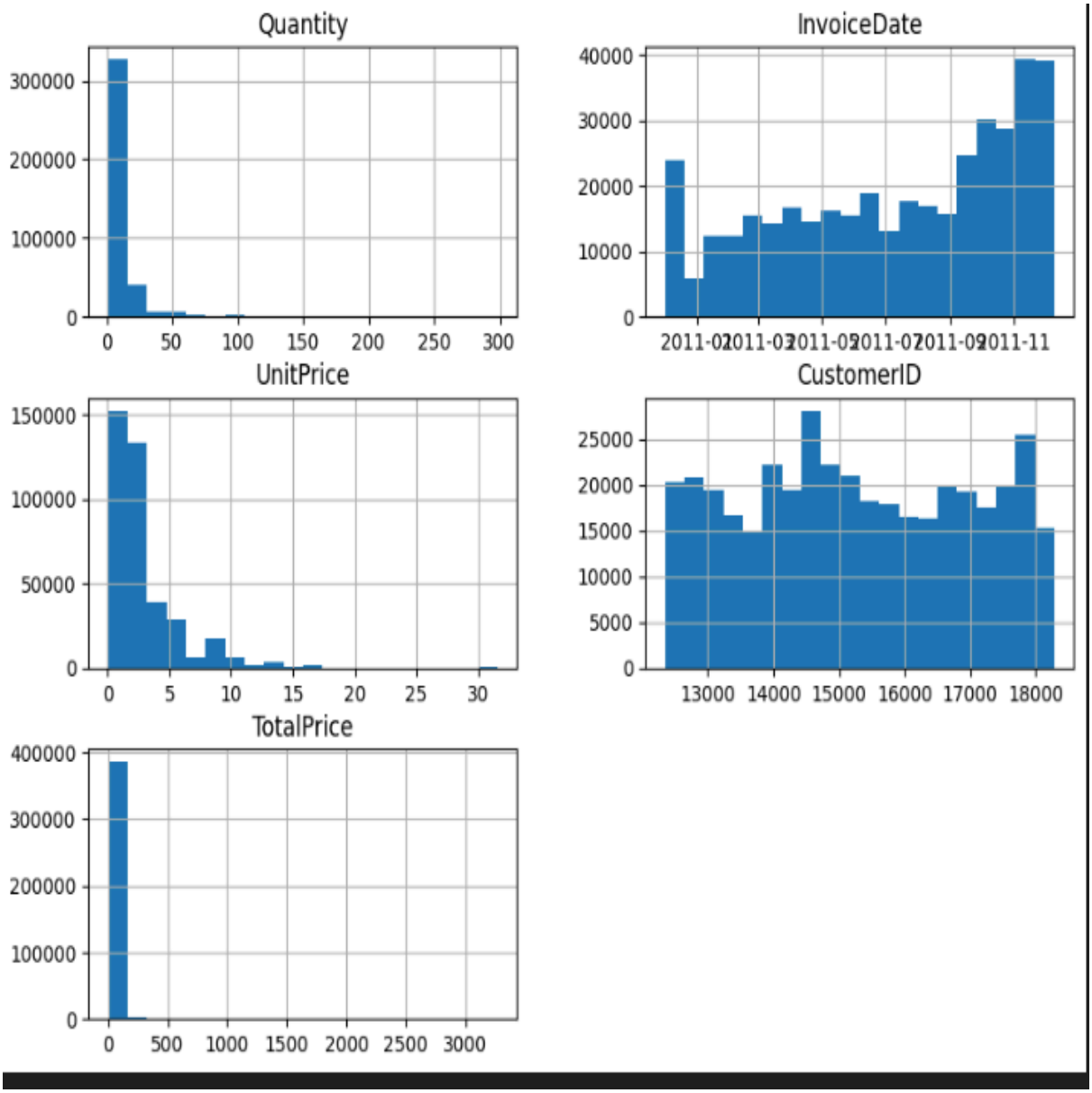
- Après l'élimination:



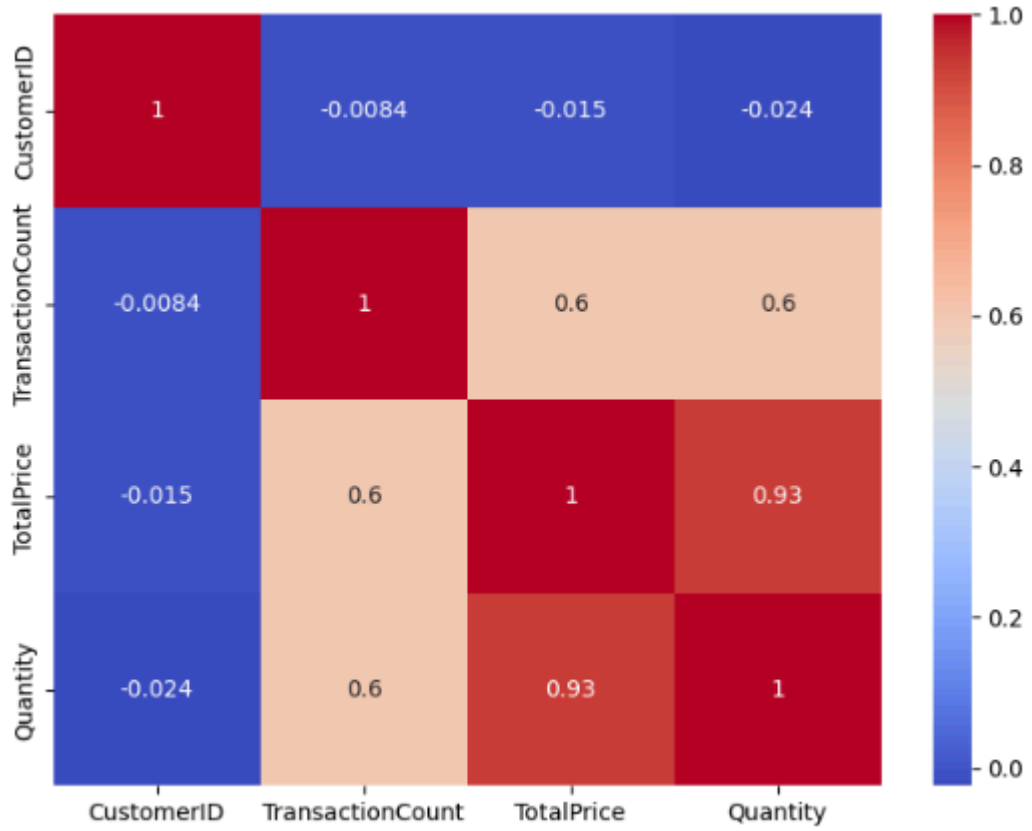
3. Exploration des données :

Dans cette section, nous explorerons les données pour identifier les tendances, les corrélations et les modèles significatifs. Cette exploration initiale nous permettra de mieux comprendre la nature des transactions de vente au détail en ligne et de générer des insights pertinents pour notre analyse.

3.1 Analyse des données:



3.1.1 Observations à partir du heatmap :



- CustomerId n'a aucune corrélation avec les autres caractéristiques, ce qui est attendu car CustomerId est un identifiant unique pour chaque client.

- TransactionCount et TotalPrice ont une corrélation positive modérée de 0.6, suggérant que les clients effectuant plus de transactions ont tendance à dépenser davantage au total.

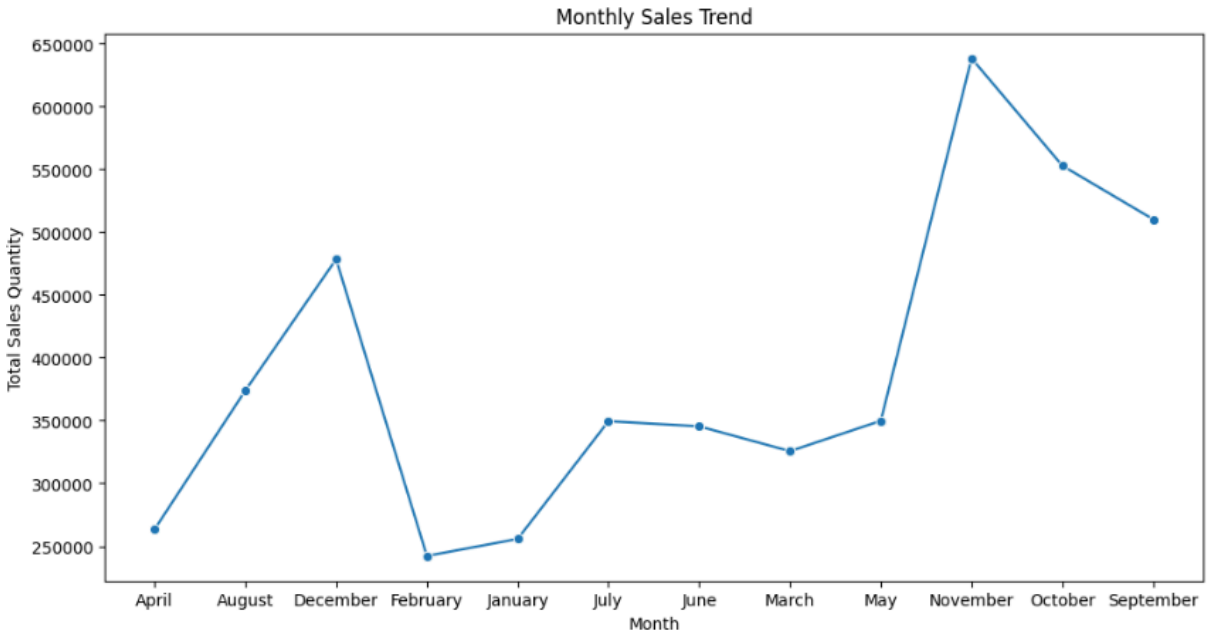
- TotalPrice et Quantity ont une corrélation très forte et positive de 0.93, ce qui indique que lorsque la quantité de produits achetés augmente, le prix total a tendance à augmenter également.

- TransactionCount et Quantity ont également une corrélation positive modérée de 0.6, ce qui implique que les clients qui achètent plus d'articles ont également tendance à effectuer davantage de transactions.

3.2 Analyse des tendances des ventes au fil du temps:

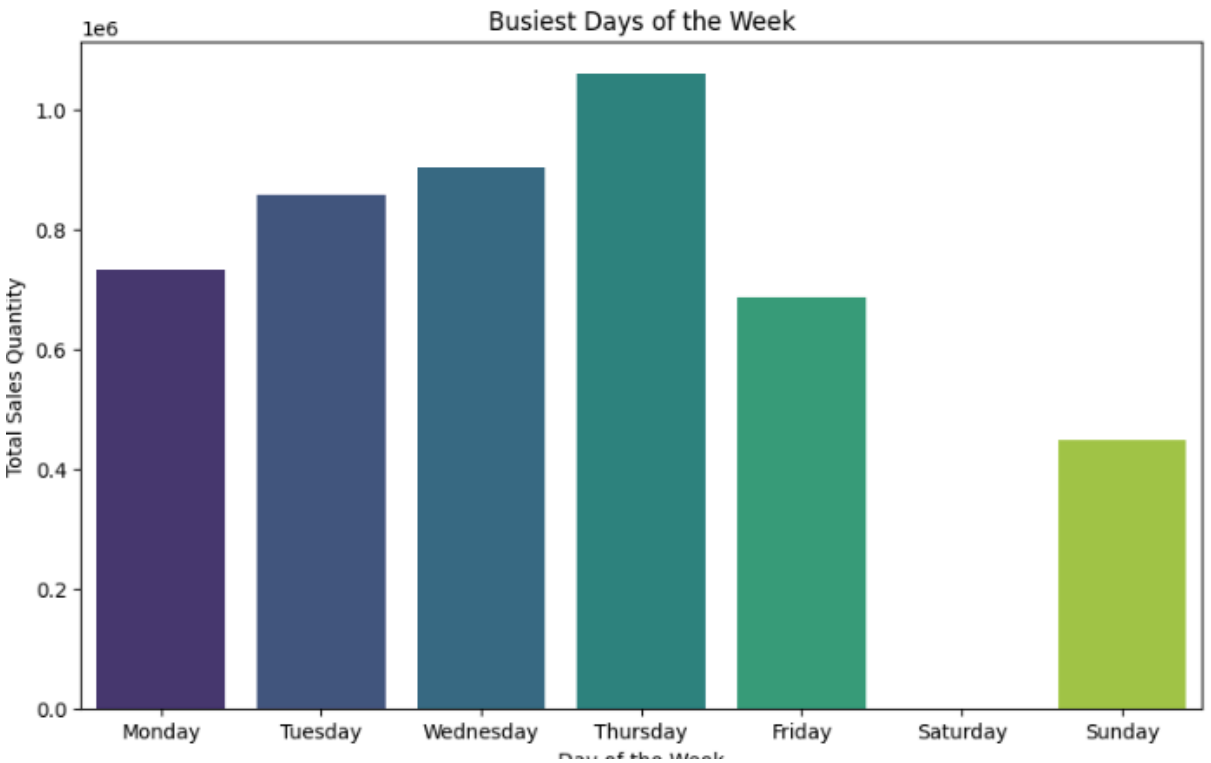
3.2.1 Analyse en fonction des mois:

Nous avons réalisé une analyse de la tendance de la quantité en fonction du mois. Le graphique montre des augmentations et des diminutions tout au long de l'année, mais le mois de novembre se distingue comme le point le plus important en termes de quantité de produits vendus.

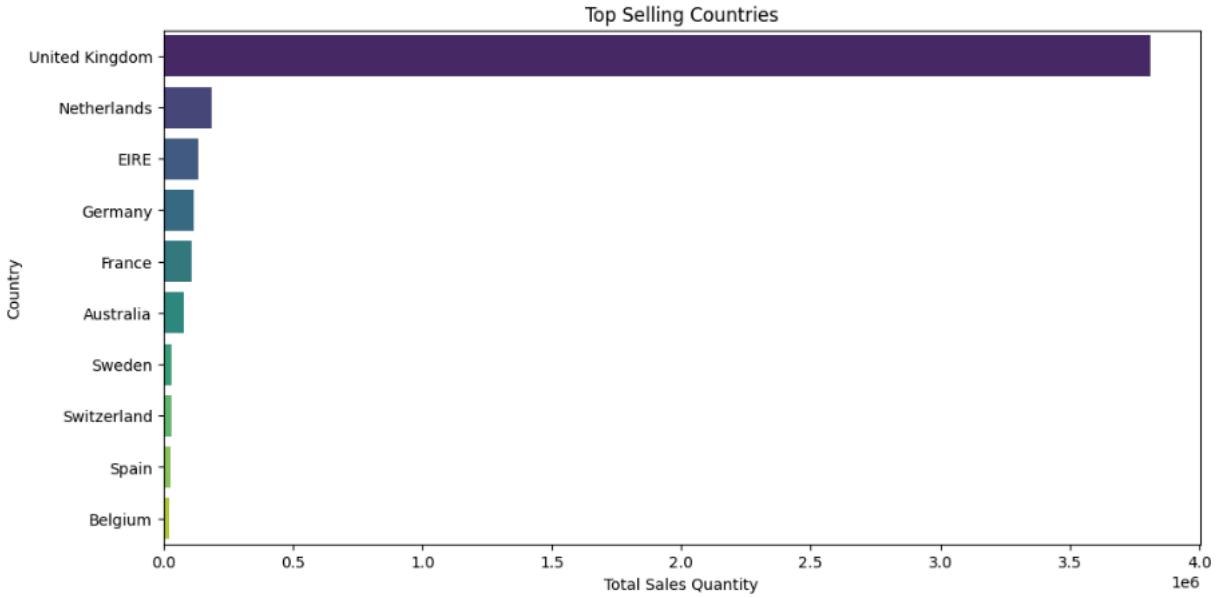


3.2.2 Analyse en fonction du jour:

D'après la figure, nous avons constaté que le jour le plus chargé est le jeudi, suivi du mercredi et du mardi.



3.3 Exploration des produits les plus vendus et des pays:

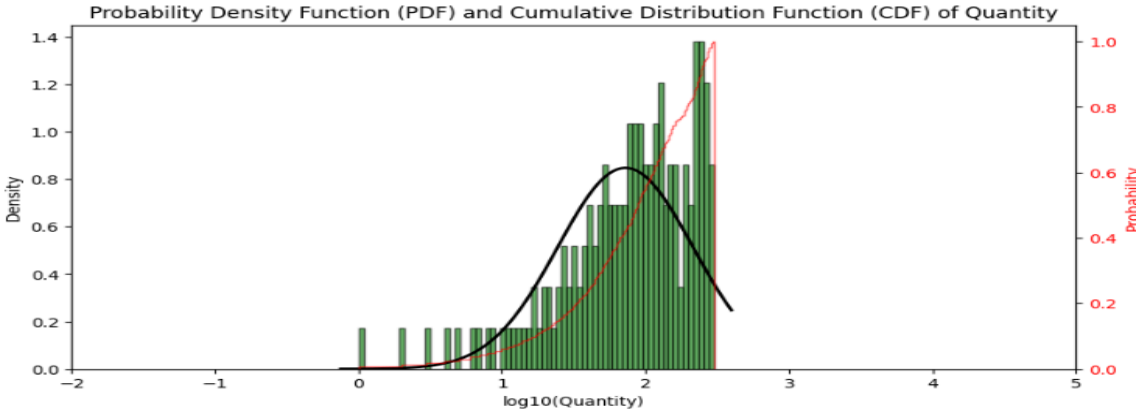


D'après cette figure, nous avons constaté que le United Kingdom est le pays où les ventes sont les plus élevées par rapport aux autres pays, suivi par Netherlands en deuxième position, puis EIRE et l'Allemagne.

3.4 Analyse de la distribution des valeurs dans les colonnes:

3.4.1 Quantity:

La fonction de densité de probabilité (PDF) montre la probabilité qu'une valeur donnée de la quantité se produise. Dans ce graphique, la PDF est la plus élevée autour de 0 sur l'axe des x, ce qui signifie que les valeurs les plus courantes de la quantité sont autour de 1. La fonction de distribution cumulative (CDF) montre la probabilité qu'une valeur de la quantité soit inférieure ou égale à une certaine valeur. Par exemple, la CDF est de 0,8 à x = 1, ce qui signifie qu'il y a 80 % de chances que la quantité soit inférieure ou égale à 1.

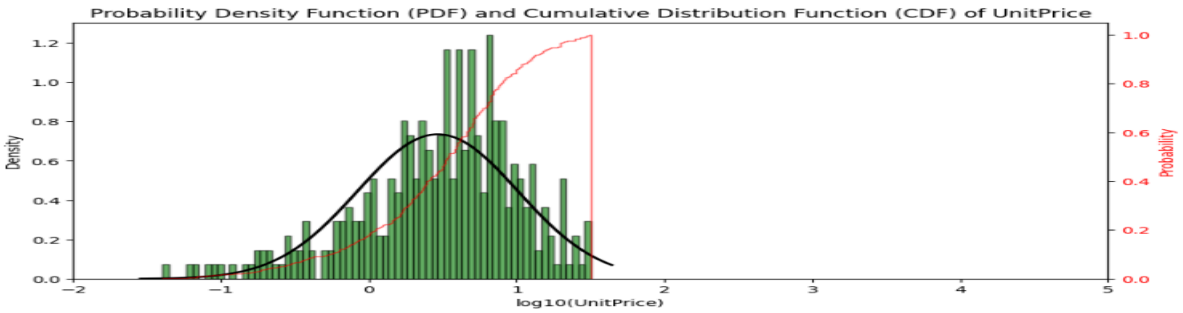


3.4.2 UnitPrice:

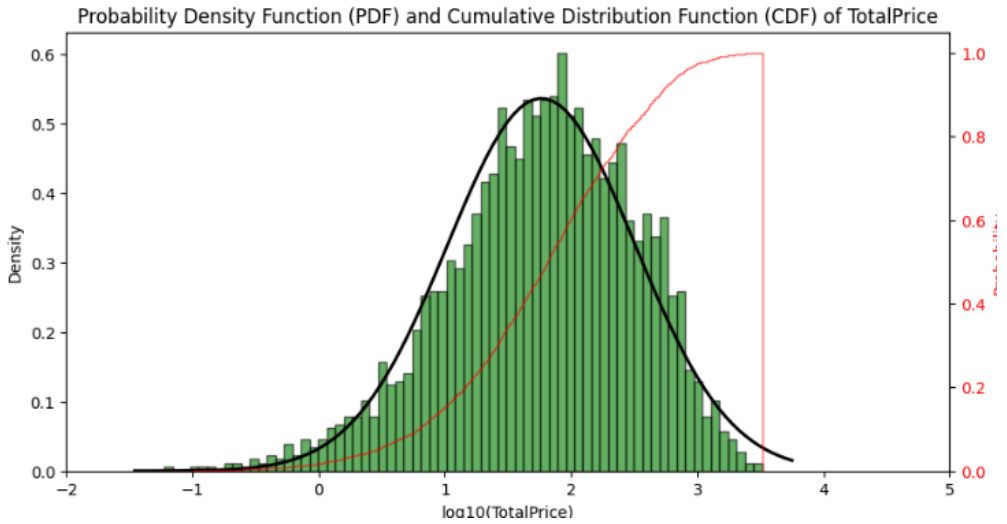
La PDF vous indique la probabilité de trouver un prix unitaire à une certaine valeur.

La CDF vous indique la probabilité que le prix unitaire soit inférieur ou égal à une certaine valeur.

Par exemple, la PDF est la plus élevée autour d'une valeur de 0 sur l'axe des x. Cela signifie que les prix unitaires sont plus susceptibles d'être autour de 1 (en échelle logarithmique). La CDF est de 0,8 à environ une valeur de 0,5 sur l'axe des x. Cela signifie qu'il y a 80% de chances que le prix unitaire soit inférieur ou égal à 1 (en échelle logarithmique), ou 10 fois plus cher que le prix unitaire le plus courant.



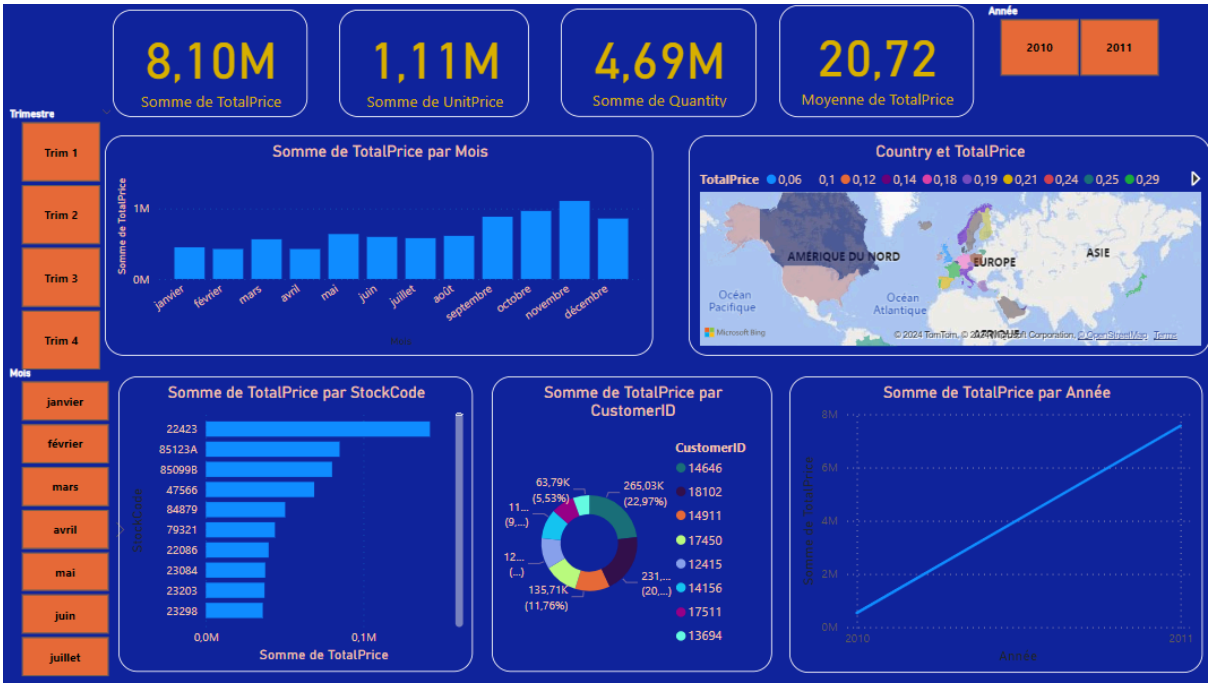
3.4.3 TotalPrice:



Fonction de densité de probabilité (PDF) : Elle montre la probabilité d'observer un certain prix total. Plus la courbe est haute pour une valeur donnée sur l'axe des x, plus ce prix total est fréquent. Ici, la PDF est la plus élevée autour de 0 sur l'axe des x (équivalent à 1 en échelle logarithmique). Cela signifie que les prix totaux les plus courants se situent autour de 1.

Fonction de distribution cumulative (CDF) : Elle indique la probabilité d'avoir un prix total inférieur ou égal à une certaine valeur. Par exemple, un CDF de 0.8 à x = 1 signifie qu'il y a 80% de chances d'avoir un prix total inférieur ou égal à 10 (10 étant la valeur équivalente à 1 en échelle logarithmique sur l'axe des x).

4. Analyse des données avec PowerBI :



À travers l'analyse des visualisations de notre tableau de bord Power BI, plusieurs insights clés émergent quant aux performances de vente au détail en ligne. Tout d'abord, en examinant la visualisation des produits les plus vendus, il est clair que certains articles se démarquent nettement en termes de ventes. Cette information nous permettra de mieux comprendre les préférences de nos clients et d'adapter notre inventaire et notre stratégie marketing en conséquence.

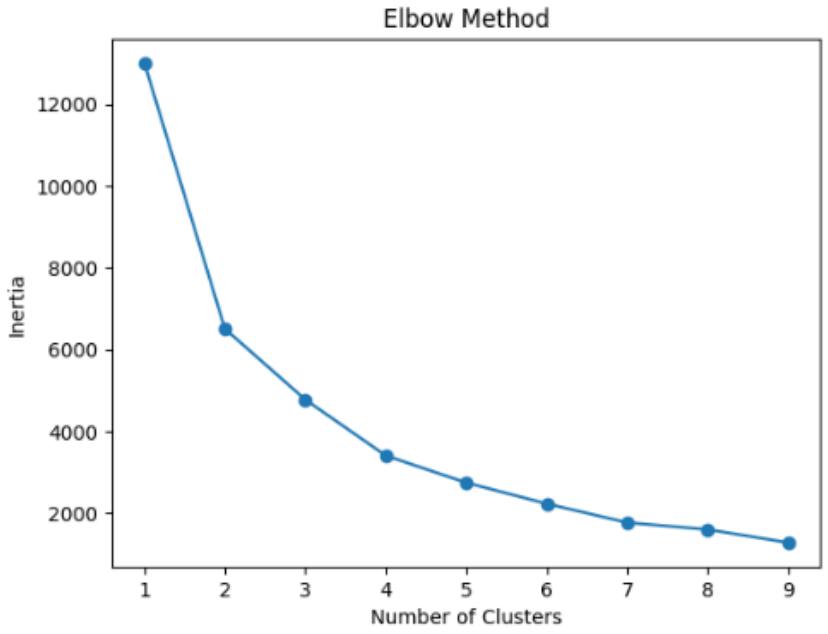
De plus, la répartition géographique des ventes par pays révèle des tendances intéressantes. Nous constatons une forte concentration des ventes dans certains pays, ce qui suggère des opportunités de croissance supplémentaires dans ces marchés. Cette analyse nous incitera à explorer davantage ces régions et à investir dans des efforts de marketing ciblés pour maximiser notre présence sur ces marchés.

En examinant les tendances temporelles des ventes, nous observons des variations saisonnières qui correspondent aux périodes de fêtes et de promotions spéciales. Cette compréhension des fluctuations de la demande nous permettra de mieux planifier nos campagnes de marketing et nos initiatives de vente pour capitaliser sur ces périodes de pic d'activité.

Enfin, l'analyse des ventes par client met en lumière les clients les plus rentables et fidèles à notre entreprise. Cette information nous permettra de personnaliser notre service client et nos offres pour fidéliser ces clients précieux et encourager leur engagement continu avec notre marque.

En combinant ces insights, nous sommes bien positionnés pour prendre des décisions stratégiques éclairées qui stimuleront la croissance de notre entreprise et renforceront notre position sur le marché de la vente au détail en ligne.

5. Clustering avec K-Means:

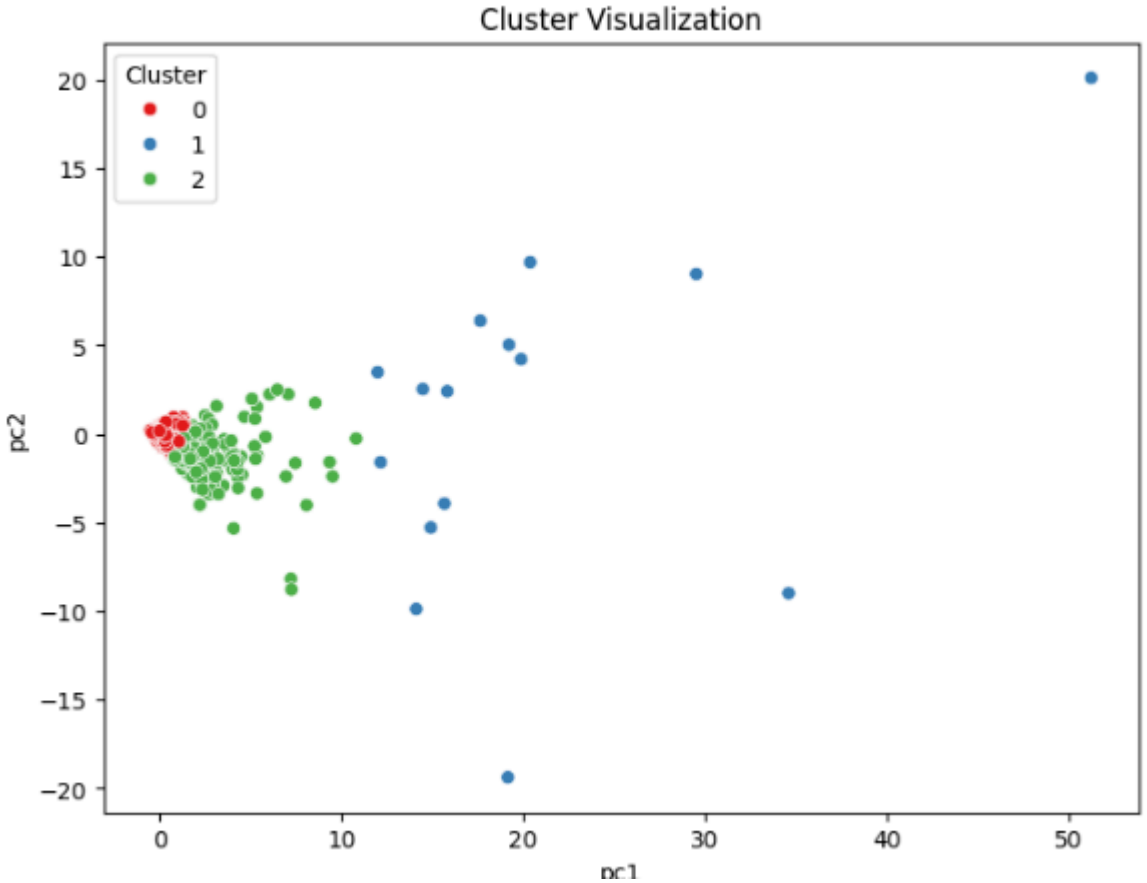


Dans le graphique, l'inertie diminue significativement lorsque le nombre de clusters passe de 1 à 2 et continue de diminuer lorsque nous passons à 3 clusters. Après 3 clusters, le taux de diminution de l'inertie ralentit, suggérant que des clusters supplémentaires n'expliquent pas beaucoup plus de la variance dans les données. Par conséquent, nous créerons 3 clusters pour ce cas d'utilisation.

- Analyse cluster centers:

```
   TransactionCount  TotalPrice   Quantity
0         -0.159392   -0.108256  -0.106089
1          9.774646   12.927967  12.924162
2          2.136349    1.051235   1.013977
Cluster
0    4083
2     236
1      15
Name: count, dtype: int64
         TransactionCount  TotalPrice   Quantity
Cluster
0               -0.159392   -0.108256  -0.106089
1                9.774646   12.927967  12.924162
2                2.136349    1.051235   1.013977
```

- Visualisation des clusters:



- Les clients du Cluster 2 constituent le groupe de clients principaux avec des habitudes de dépenses élevées et pourraient être ciblés avec des stratégies marketing générales.

- Les clients du Cluster 0 sont précieux avec des dépenses moyennes et pourraient être réceptifs à des activités promotionnelles encourageant une fréquence ou un volume d'achat accru.

- Les clients du Cluster 1 sont probablement les plus précieux et pourraient être au centre de services premium, d'offres exclusives ou d'autres efforts marketing haut de gamme.

N.B: On peut même utiliser K-Means avec RFM : R (Récence), F (Fréquence), M (Monétaire), qui est une méthode d'analyse de segmentation des clients largement utilisée dans le domaine du marketing et du commerce. Ensuite, on peut évaluer le

modèle en utilisant le coefficient de silhouette, qui est une métrique utilisée pour évaluer la qualité d'une technique de regroupement.

6. Pistes D'amélioration:

Suite à notre analyse détaillée, plusieurs opportunités d'amélioration pour notre entreprise de vente en ligne sont apparues. Nous pouvons ajuster notre inventaire pour mieux répondre à la demande saisonnière et augmenter la disponibilité des produits populaires. En parallèle, nous devrions concentrer nos efforts marketing sur les pays et segments de clients les plus rentables, en lançant des campagnes ciblées pour maximiser notre retour sur investissement publicitaire. Explorer de nouveaux marchés émergents identifiés dans notre analyse peut diversifier notre clientèle et stimuler nos revenus. Il est essentiel d'améliorer continuellement l'expérience client en offrant des services personnalisés et en fidélisant nos clients les plus précieux. Ajuster nos stratégies de tarification et nos offres promotionnelles en fonction des insights obtenus permettra d'optimiser nos revenus tout en maintenant notre compétitivité sur le marché. Enfin, un investissement dans des solutions technologiques avancées renforcera notre efficacité opérationnelle et nous permettra de prendre des décisions commerciales plus éclairées à l'avenir. En mettant en œuvre ces recommandations, nous sommes confiants dans notre capacité à consolider notre position sur le marché et à assurer une croissance durable de notre entreprise.

7. Conclusion:

En conclusion, l'analyse approfondie des données du détaillant en ligne a permis de découvrir des insights précieux pour guider les décisions commerciales. En comprenant les tendances de vente, les préférences des clients et les opportunités de croissance, l'entreprise est mieux positionnée pour réussir dans le marché compétitif du commerce de détail en ligne. En utilisant les données comme guide, elle peut ajuster ses stratégies marketing, optimiser son inventaire et offrir une expérience client personnalisée, ce qui contribuera à stimuler sa croissance et à renforcer sa position sur le marché.