CCE – IIT Madras – Advanced Certification in Data Science and AI

Capstone Project – 2.2

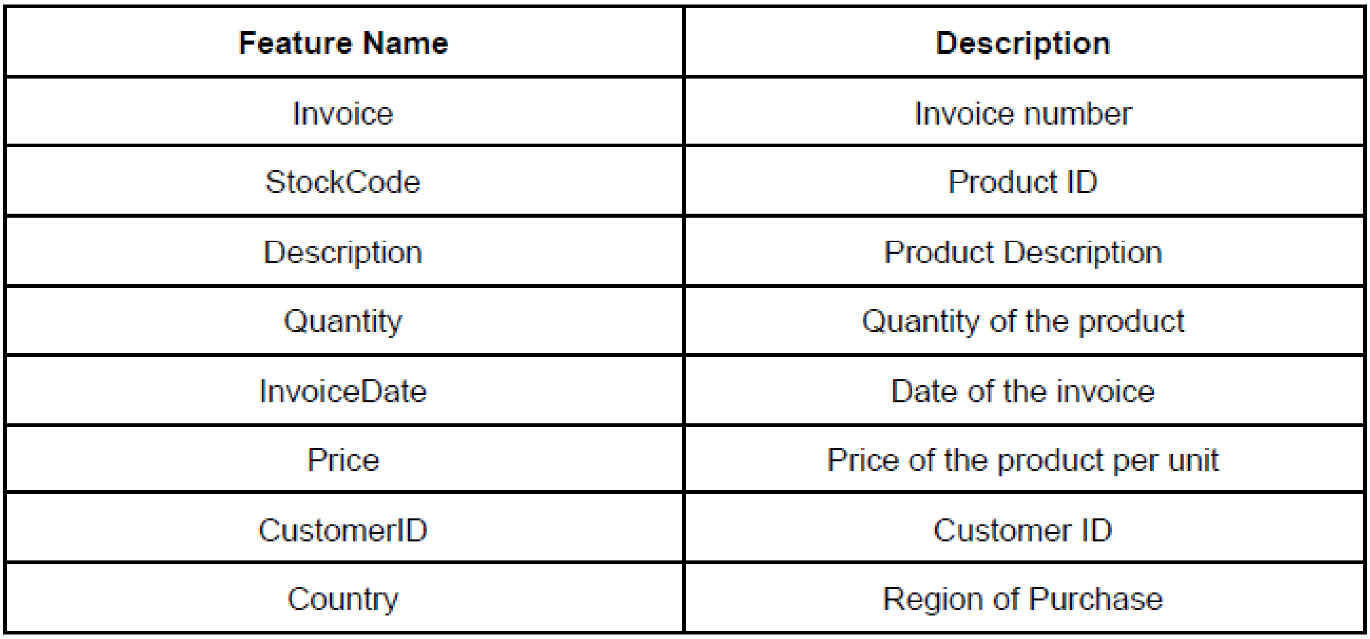
Customer Segmentation with Online Retail Dataset

Table of Contents

1. Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same.

Dataset Information:
The online_retail.csv contains 387961 rows and 8 columns.



Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	Customer ID
Country	Region of Purchase

---

2. Project Objective

1. Using the above data, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.
2. Segment the customers based on their purchasing behaviour.

However, it includes different steps to explore the trends and obtain something useful from the dataset, which are:

- Exploratory Data Analysis
- Data Pre-processing & cleaning of the data
- Dealing with Null entries and creating a dataset as per the problem statement
- Further, Feature Engineering to Extract and select features accordingly

---

3. Data Description

3.2. Detailed Overview of Dataset:

Number of rows in the dataset = 525461 ROWS

Number of columns in the dataset = 8 COLUMNS

1. InvoiceNo: A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'C', it indicates a cancellation (Nominal)
2. StockCode: A 5-digit integral number uniquely assigned to each distinct product (Nominal)
3. Description: Product (item) name. (Nominal)
4. Quantity: The quantities of each product (item) per transaction (Numeric)
5. InvoiceDate: The day and time when each transaction was generated (Numeric)
6. UnitPrice: Product price per unit in sterling (Numeric)
7. CustomerID: A 5-digit integral number uniquely assigned to each customer (Nominal)
8. Country: Name of the country where each customer resides (Nominal)

3.2. Features Information with unique entries:

- Total number of unique values for InvoiceNo: 25900
- Total number of unique values for StockCode: 4070
- Total number of unique values for Description: 4223
- Total number of unique values for Quantity: 722
- Total number of unique values for InvoiceDate: 23260
- Total number of unique values for UnitPrice: 1630
- Total number of unique values for CustomerID: 4372
- Total number of unique values for Country: 38

3.3. Missing value distribution:

Features	Missing Value Count	Percentage
CustomerID	135080	24.93
Description	1454	0.27
InvoiceNo	0	0.00
StockCode	0	0.00
Quantity	0	0.00
InvoiceDate	0	0.00
UnitPrice	0	0.00
Country	0	0.00

3.4. Summary:

- Number of invoices: 25900
- Number of products bought: 4070
- Number of customers: 4372
- Percentage of customers NA: 24.93 %
- The average quantity of the product purchased by a customer: 1122.0
- Average revenue generated per customer: 322.01
- Average product quantity sold per transaction: 10.0
- Average revenue generated per transaction: 4.61

4. Data Pre-processing Steps and Inspiration

4.1. Data Cleaning & removing outliers:

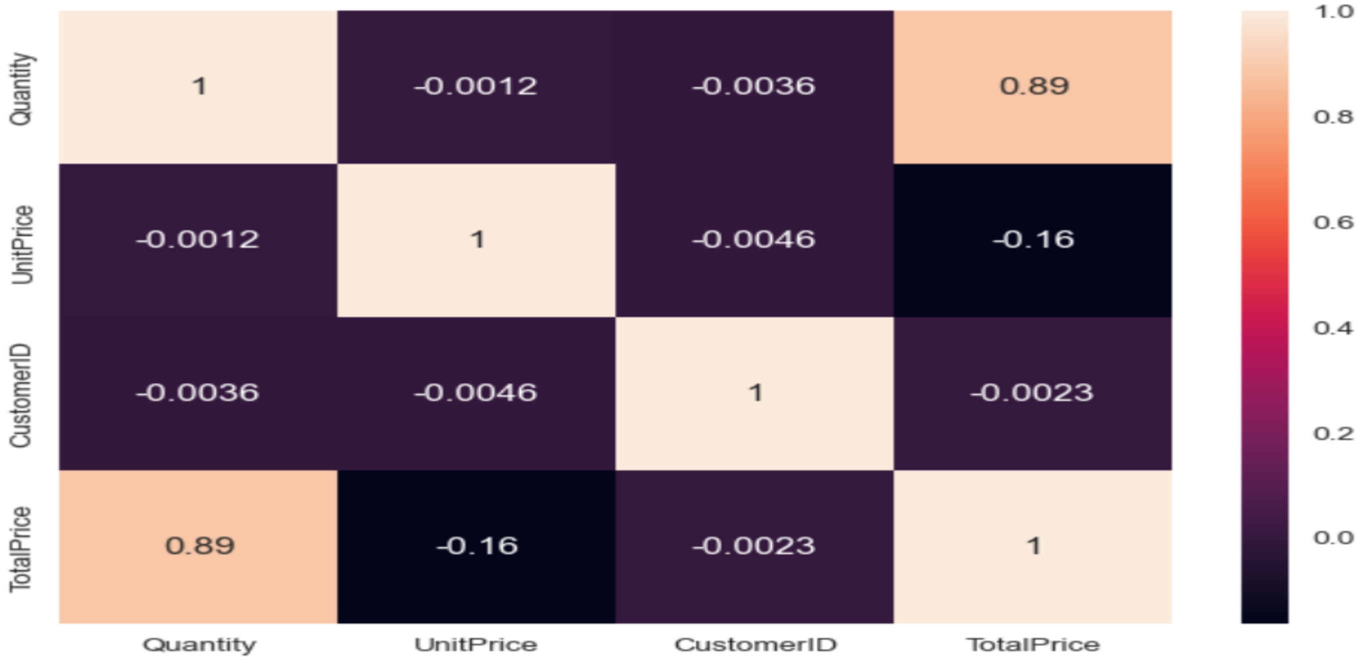
I have cleaned the data by handling & dropping the missing, duplicate values and fixing outliers using Robust statistical methods, such as the interquartile range (IQR). This step ensures the data is accurate and reliable.

4.2. Exploratory data analysis (EDA):

Identify relevant features that are most informative for customer segmentation. You can perform exploratory data analysis (EDA) to understand the distribution and correlations between variables. Additionally, techniques like feature importance ranking, and correlation analysis, can help in selecting the most relevant features.

1. To proceed further, I have first created a column that holds the total price by multiplying the unit price by the quantity of each product row-wise.

2. **Correlation heatmaps** are commonly used in feature selection, and identifying patterns or relationships in the data. They serve as a valuable tool to gain a visual understanding of the interdependencies between variables, aiding in data interpretation and decision-making.



Here, we can clearly see that the Quantity & TotalPrice are positively correlated. CustomerID is displayed as its data type is 'numerical', and I decided to keep it numerical as it will be helpful to create and extract features to segment the customers.

3. **Features importance,** we can apply the Random Forest, Gradient boosting, etc, to our dataset to find out their importance, however, we do not have enough numerical features to perform it. So, I decided to manually select the features,

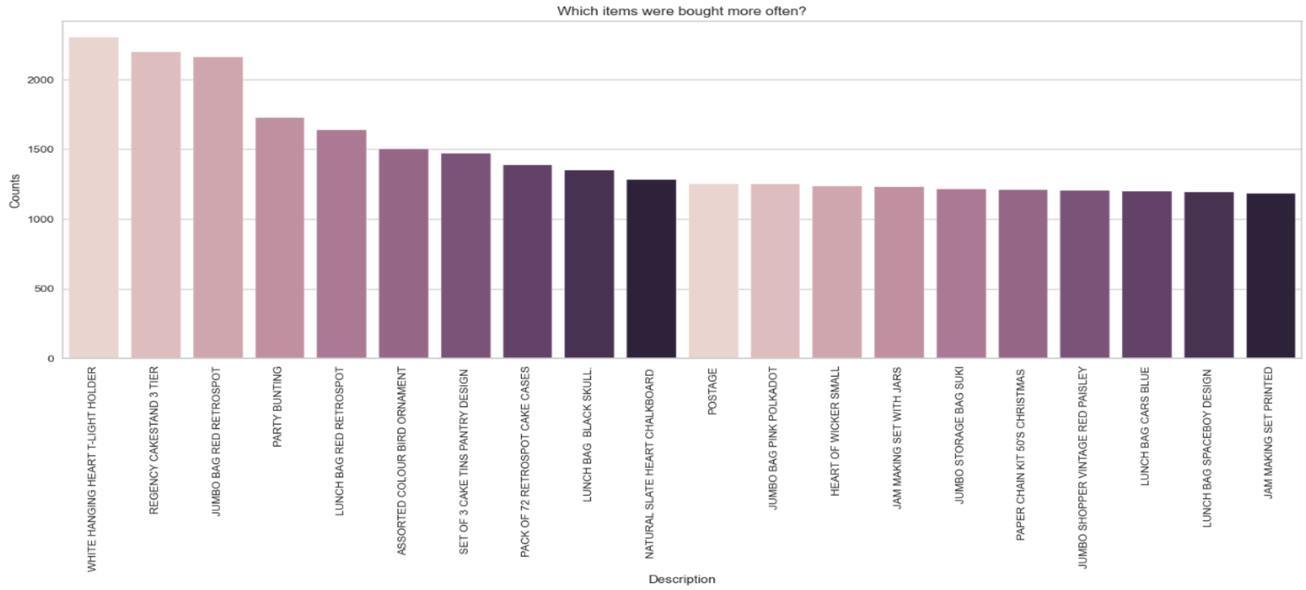
* **Invoice Number,** It will be an important feature along with the **Customer ID** to find out the customer's purchasing pattern.

***Country,** this would be less but not least important, because this column contains the customer's native country. However, most of the customers approx. 91.43% belongs to the United Kingdom.

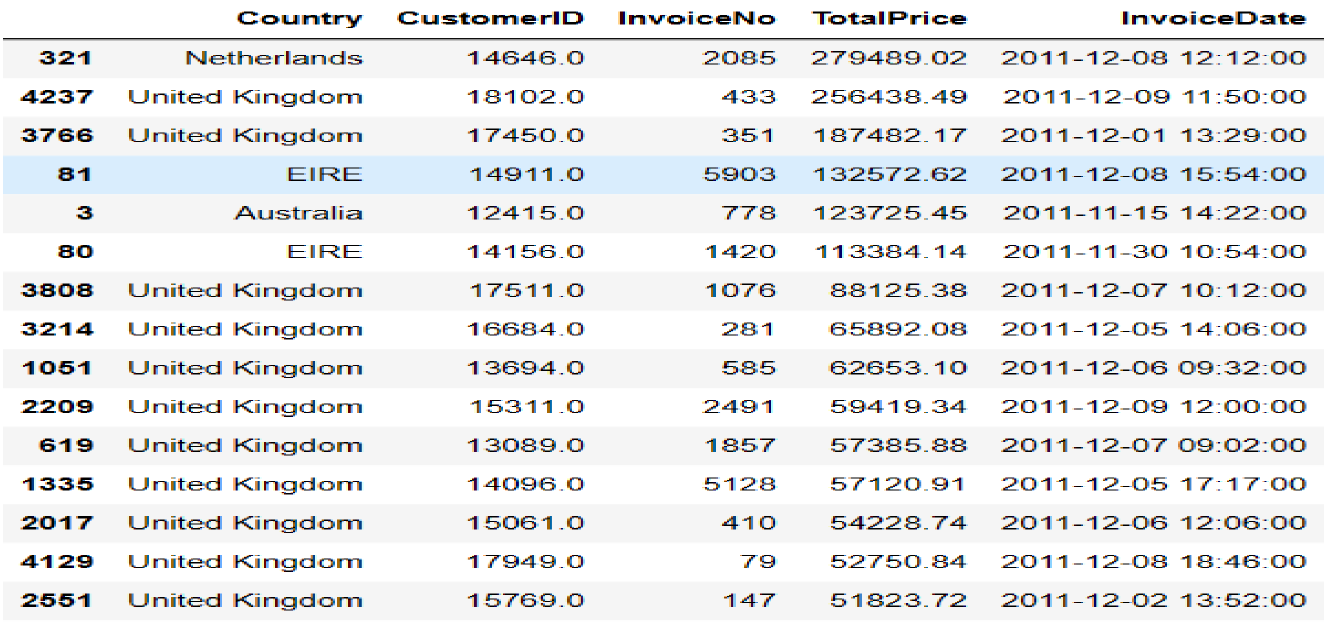
***Description & Stock code** contains the name of the products and their relevant stock number, which would be important for product-based analysis but not important for further model-building processes with the K means algorithm.

***Quantity, Unit Price & Total Price** are correlated, as we can see in the above heatmap plot, so, they will be very important features in all.

4. **Grouping Columns to get Invoice count for their respective product:** I have grouped the Stock code, Description, and Invoice count to find the most & least-selling products. I have attached the screenshot with the top 20 products as per their invoice counts.

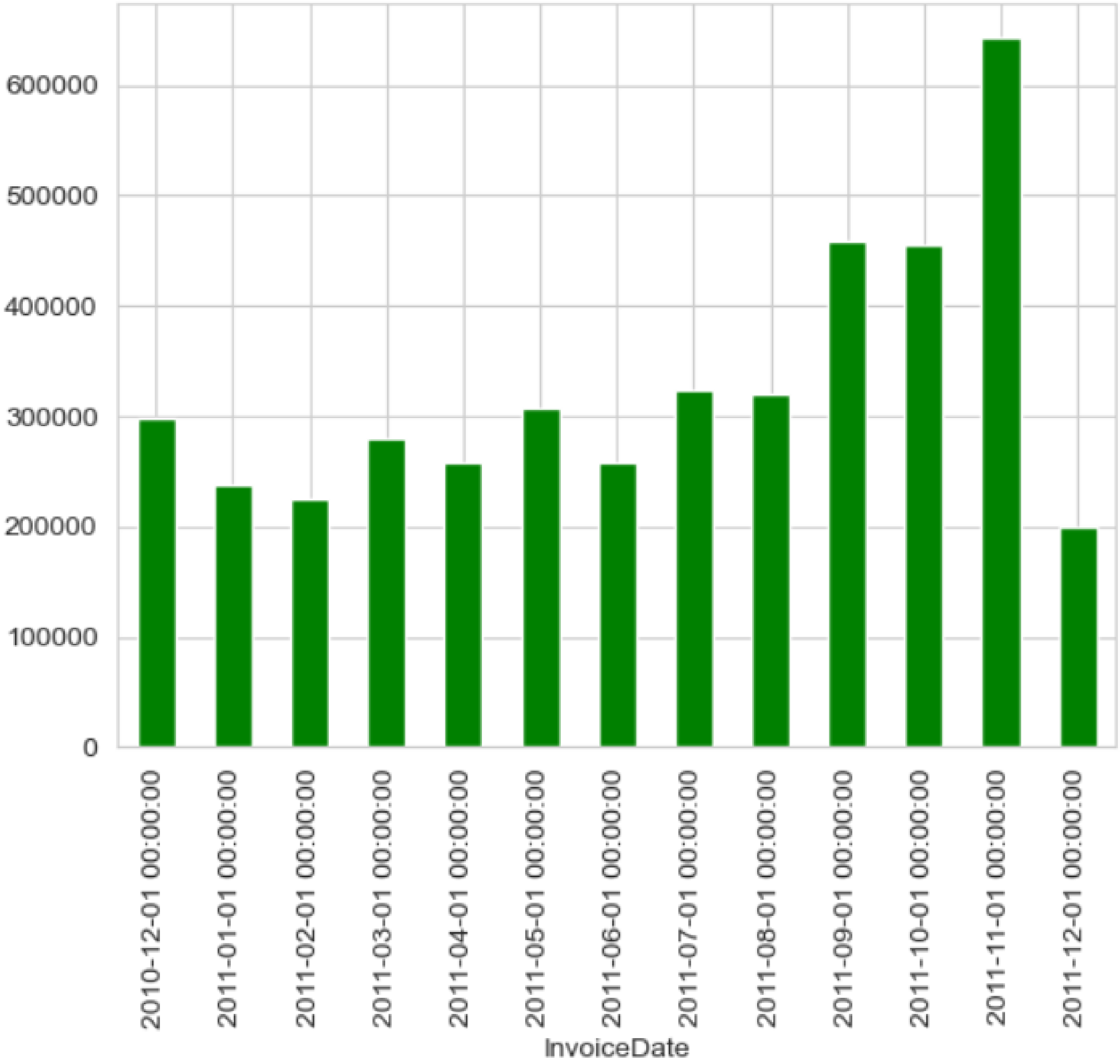


5. **Grouping to get multiple statistics for different columns:** Calculating recency, frequency, and customer lifetime value, in which top 15 are mentioned below:

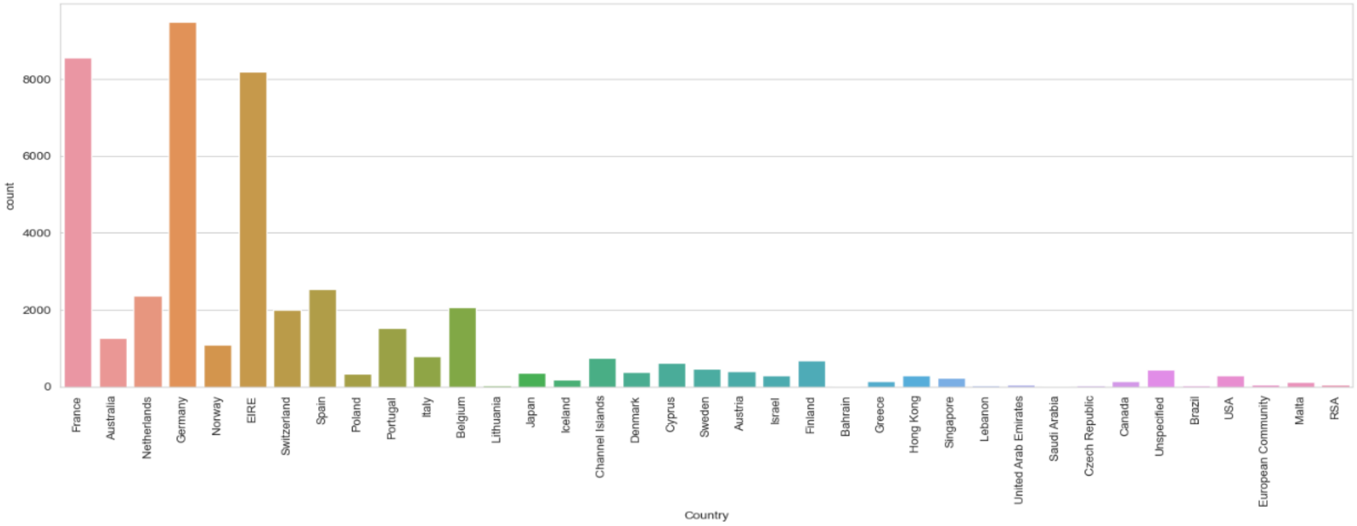


	Country	CustomerID	InvoiceNo	TotalPrice	InvoiceDate
321	Netherlands	14646.0	2085	279489.02	2011-12-08 12:12:00
4237	United Kingdom	18102.0	433	256438.49	2011-12-09 11:50:00
3766	United Kingdom	17450.0	351	187482.17	2011-12-01 13:29:00
81	EIRE	14911.0	5903	132572.62	2011-12-08 15:54:00
3	Australia	12415.0	778	123725.45	2011-11-15 14:22:00
80	EIRE	14156.0	1420	113384.14	2011-11-30 10:54:00
3808	United Kingdom	17511.0	1076	88125.38	2011-12-07 10:12:00
3214	United Kingdom	16684.0	281	65892.08	2011-12-05 14:06:00
1051	United Kingdom	13694.0	585	62653.10	2011-12-06 09:32:00
2209	United Kingdom	15311.0	2491	59419.34	2011-12-09 12:00:00
619	United Kingdom	13089.0	1857	57385.88	2011-12-07 09:02:00
1335	United Kingdom	14096.0	5128	57120.91	2011-12-05 17:17:00
2017	United Kingdom	15061.0	410	54228.74	2011-12-06 12:06:00
4129	United Kingdom	17949.0	79	52750.84	2011-12-08 18:46:00
2551	United Kingdom	15769.0	147	51823.72	2011-12-02 13:52:00

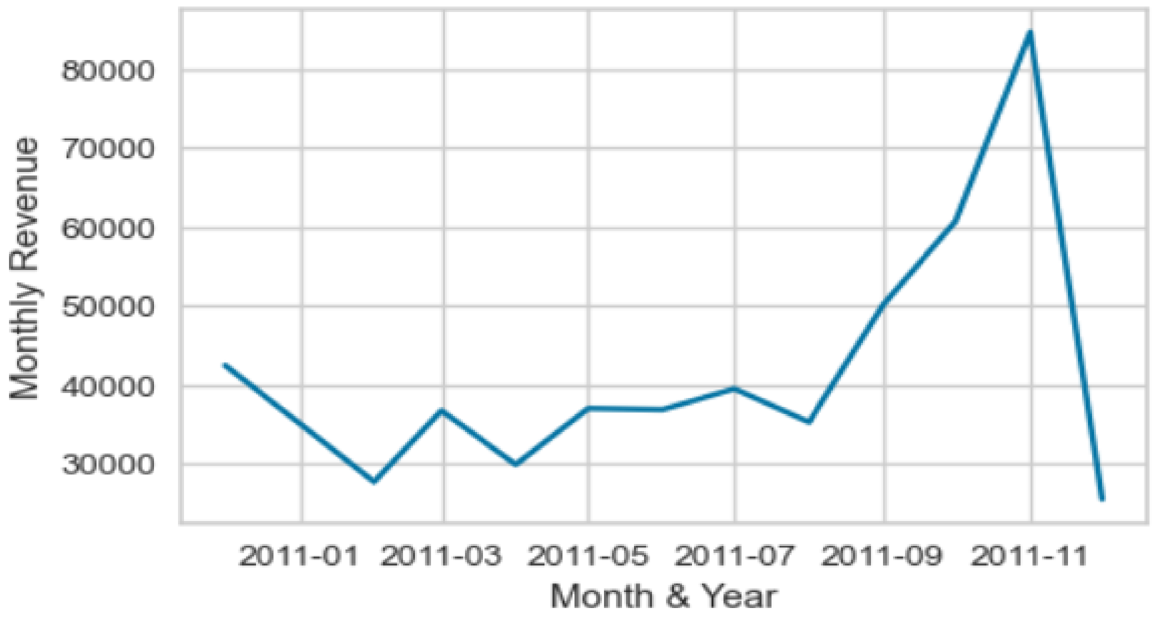
6. **Quantity-Based Analysis:** Naming grouped aggregate columns with multiple statistics, and resetting the index to include the date column, for which attached the plot:

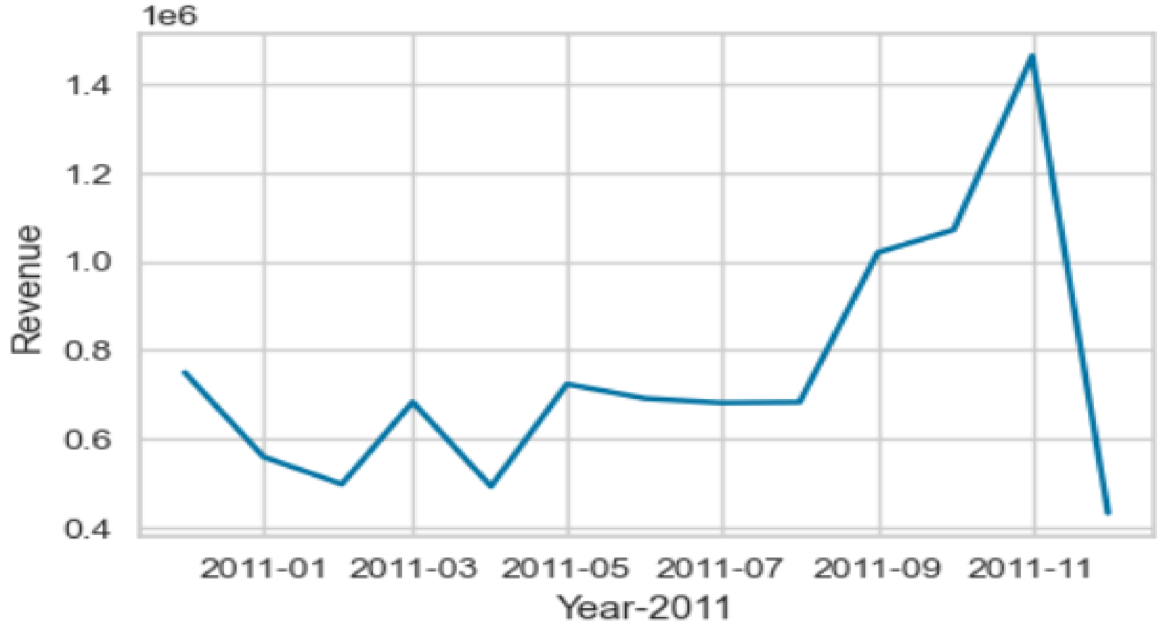


7. **Country-Based Analysis:** I have discovered after visualizing the overall sales that most of the customers are from the United Kingdom, i.e., almost 94% of all customers. So, I have visualized the country-based sales excluding the United Kingdom.

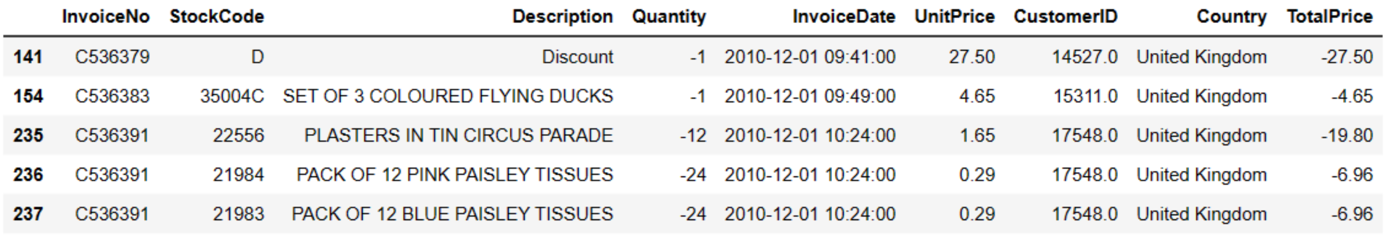


8. **Time-Based revenue analysis:** I have grouped the feature to find the monthly & overall revenue trend for the year 2011.





9. **Cancelled Invoice/Orders:** As mentioned in the description of the dataset and we also observed above that some InvoiceNo starts with the letter "c" = cancelled. Let's see if our hypothesis is correct about the negative quantity: -80995. We will look for the list of cancelled invoices and check if there is an invoice with that quantity.



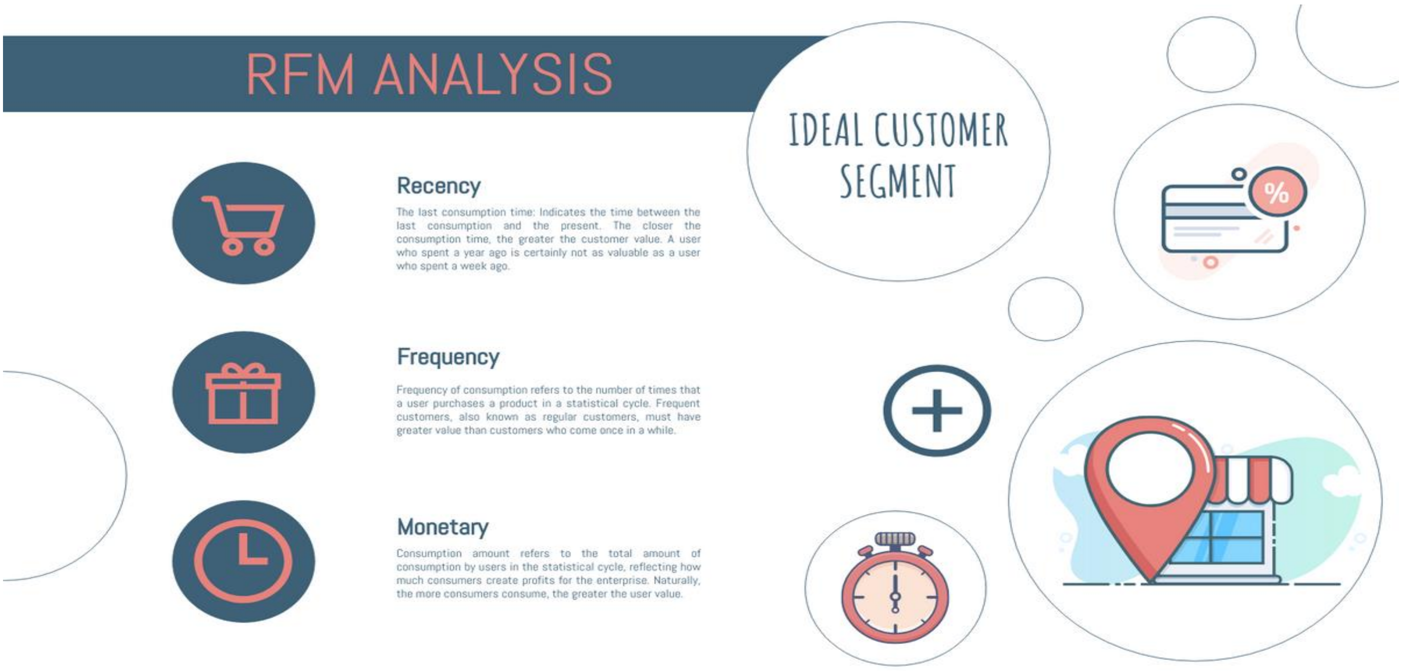
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom	-27.50
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom	-4.65
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom	-19.80
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom	-6.96
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom	-6.96

Total number of cancelled invoices/orders: 9288, and Percentage of orders canceled: 9288/25900 (35.86%)

4.3. Feature Extraction with the Help of RFM Segmentation:

Overview of RFM Segmentation

RFM segmentation is a technique used in marketing and customer relationship management (CRM) to categorize customers based on their recent purchase behavior. RFM stands for Recency, Frequency, and Monetary Value, which are three key factors used to analyze customer behavior and identify different segments.



Recency

The last consumption time: Indicates the time between the last consumption and the present. The closer the consumption time, the greater the customer value. A user who spent a year ago is certainly not as valuable as a user who spent a week ago.

Frequency

Frequency of consumption refers to the number of times that a user purchases a product in a statistical cycle. Frequent customers, also known as regular customers, must have greater value than customers who come once in a while.

Monetary

Consumption amount refers to the total amount of consumption by users in the statistical cycle, reflecting how much consumers create profits for the enterprise. Naturally, the more consumers consume, the greater the user value.

Here's a breakdown of each component in RFM segmentation:

1. **Recency (R):** Recency refers to the time elapsed since a customer's last purchase. It captures how recently a customer has interacted with a business. Customers who have made a purchase more recently are considered more engaged and potentially more valuable.

2. **Frequency (F):** Frequency measures how often a customer makes purchases within a specific period. It quantifies the customer's level of engagement or loyalty. Customers who make frequent purchases are generally more loyal and likely to generate consistent revenue.

3. **Monetary Value (M):** Monetary Value represents the amount of money a customer has spent on purchases. It reflects the customer's purchasing power and their potential value to the business. Customers with higher monetary value are typically more profitable.

To perform RFM segmentation, the following steps are typically followed:

1. **Calculation of RFM Scores:** Calculate the RFM scores for each customer based on their individual recency, frequency, and monetary value. This involves assigning numerical values or rankings to each customer based on their behavior.

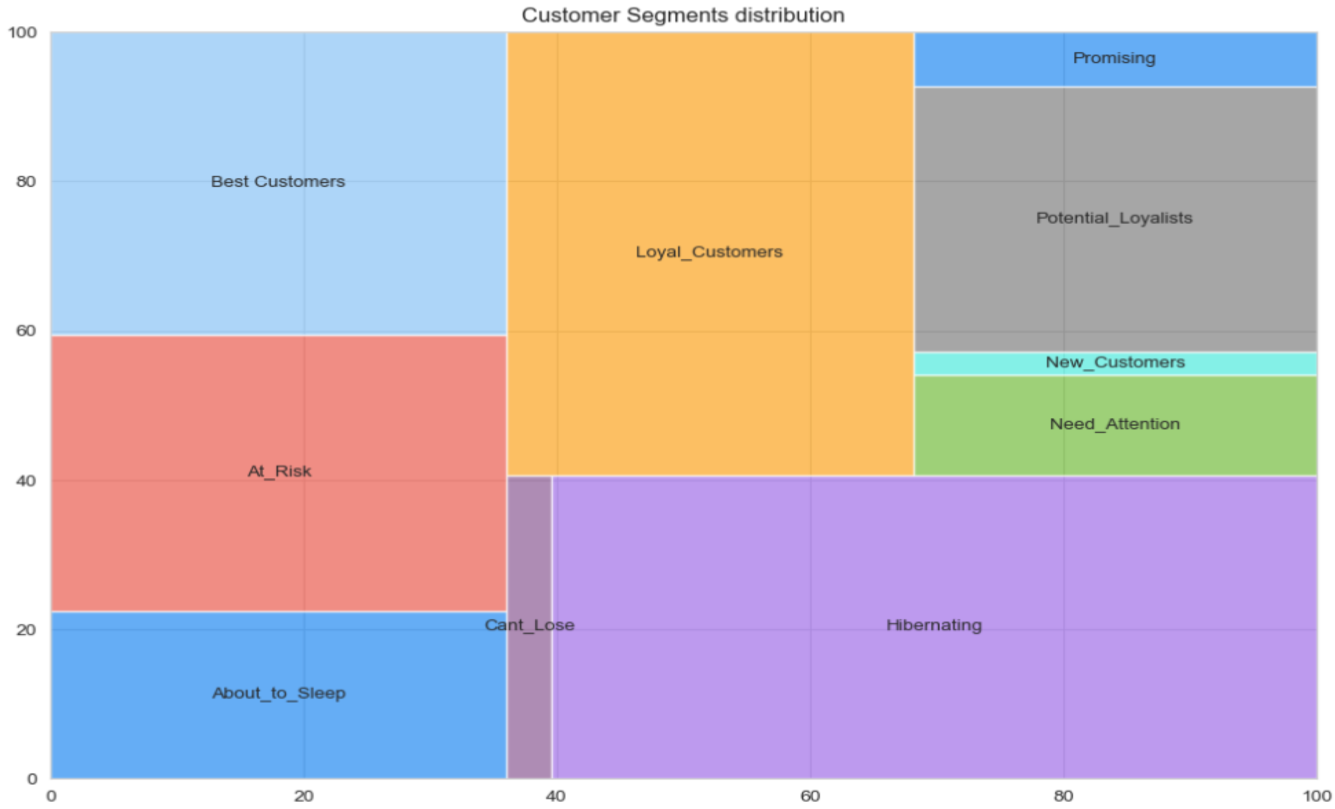
- **Recency score:** Determine the recency score based on the time elapsed since the customer's last purchase. For example, a score of 5 can be assigned to customers who made a purchase within the last 30 days, while a score of 1 can be assigned to customers who made a purchase more than 365 days ago.

- **Frequency score:** Assign a frequency score based on the number of purchases made by each customer within a specific timeframe. Customers with a higher frequency of purchases receive a higher score.

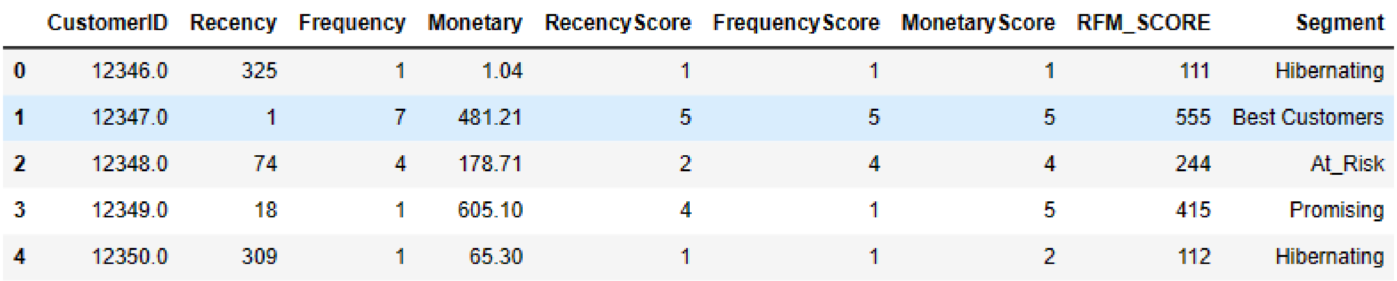
- **Monetary value score:** Assign a monetary value score based on the total amount spent by each customer. Customers who have spent more receive a higher score.

2. **Segment Formation:** Combine the RFM scores to create distinct customer segments. This can be done by dividing the scores into quartiles or by using clustering algorithms (such as K-means or hierarchical clustering) to group customers with similar RFM characteristics.
3. **Analysis and Action:** Analyze each segment to gain insights into customer behavior and preferences. Tailor marketing strategies, promotions, and customer communications to effectively engage and retain customers within each segment. Different segments may require different marketing approaches based on their RFM profiles.
4. **Implementing RFM Scores to segment the customers:**

- The min number of Recency metric means that this customer has just purchased, so the highest score (5) should be given to the lower number of Recency.
- The max number of Frequency and Monetary metrics mean that the customer is purchasing frequently and spending more money, so the highest score (5) should be given to the highest Frequency and Monetary values.
- By using the RFM scores I have segmented the customer as per their purchasing pattern into a few classes, which is clearly visible below,



Also, I have extracted the following features after performing the RFM segmentation and analysis and our new dataset has 4339 records with 9 features and from which the top 5 records are;



	CustomerID	Recency	Frequency	Monetary	RecencyScore	FrequencyScore	MonetaryScore	RFM_SCORE	Segment
0	12346.0	325	1	1.04	1	1	1	111	Hibernating
1	12347.0	1	7	481.21	5	5	5	555	Best Customers
2	12348.0	74	4	178.71	2	4	4	244	At_Risk
3	12349.0	18	1	605.10	4	1	5	415	Promising
4	12350.0	309	1	65.30	1	1	2	112	Hibernating

5. Choosing the Algorithm for the Project

We have studied the RFM segmentation and for a further procedure, we required a clustering algorithm (K means, Hierarchical clustering, etc.), because these algorithms help us to achieve their relevant classes and K means are the most used & preferred for the clustering purpose.

Additionally, it requires specifying the number of clusters (k) in advance and K-means clustering remains a popular and effective choice for customer segmentation in the online retail industry.

6. Motivation and Reasons for Choosing the Algorithm

There are several motivations and reasons for choosing the K-means clustering algorithm for customer segmentation using an online retail dataset. Here are some specific points:

Scalability: K-means clustering is scalable and can handle large datasets efficiently. Online retail datasets often consist of a large number of customers, transactions, and variables such as purchase history, demographics, and browsing behavior. The scalability of K-means makes it suitable for analyzing and clustering such large-scale datasets.

Quick Insights: K-means provides a fast way to gain initial insights into customer behavior. By clustering customers based on their purchasing patterns, businesses can quickly identify distinct groups of customers with similar preferences. These insights can help in making data-driven decisions, such as creating targeted promotions or designing personalized marketing campaigns.

Time Complexity: The time complexity of the K-means clustering algorithm is generally considered to be relatively efficient compared to some other clustering algorithms.

- O (I * K * n * d), where I is the number of iterations, K is the number of clusters, n is the number of data points, and d is the number of dimensions (features) in the dataset.
- The K-means algorithm iteratively assigns data points to clusters and updates the cluster centroids until convergence. The number of iterations required for convergence can vary, but in practice, it often converges quickly.

Interpretability: K-means provides clusters that are relatively easy to interpret. The algorithm aims to minimize the within-cluster sum of squares, which means it groups data points that are closer to each other. This property makes it easier to interpret and analyze the resulting customer segments.

Robustness: K-means is generally robust to noise and outliers in the data. Outliers are less likely to significantly affect the resulting clusters since each data point is assigned to the nearest centroid. This robustness is beneficial when dealing with real-world customer data that may contain noise or outliers.

---

7. Assumptions

I have decided to go with 3 and 5 '**n_clusters**' as we have discovered earlier on the basis of RFM scores that customers can be classified under 10 different segments. To proceed further we will evaluate the model with **n_clusters** 3 and 5 before training the final model.

8. Model Evaluation and Techniques

K-means is an unsupervised machine-learning algorithm that is used to group data into clusters based on similarity. Unlike supervised algorithms, there is no clear metric for evaluating the performance of a K-means model. However, there are some techniques that can be used to evaluate the quality of the clustering results. Here are some common evaluation techniques for K-means in Python:

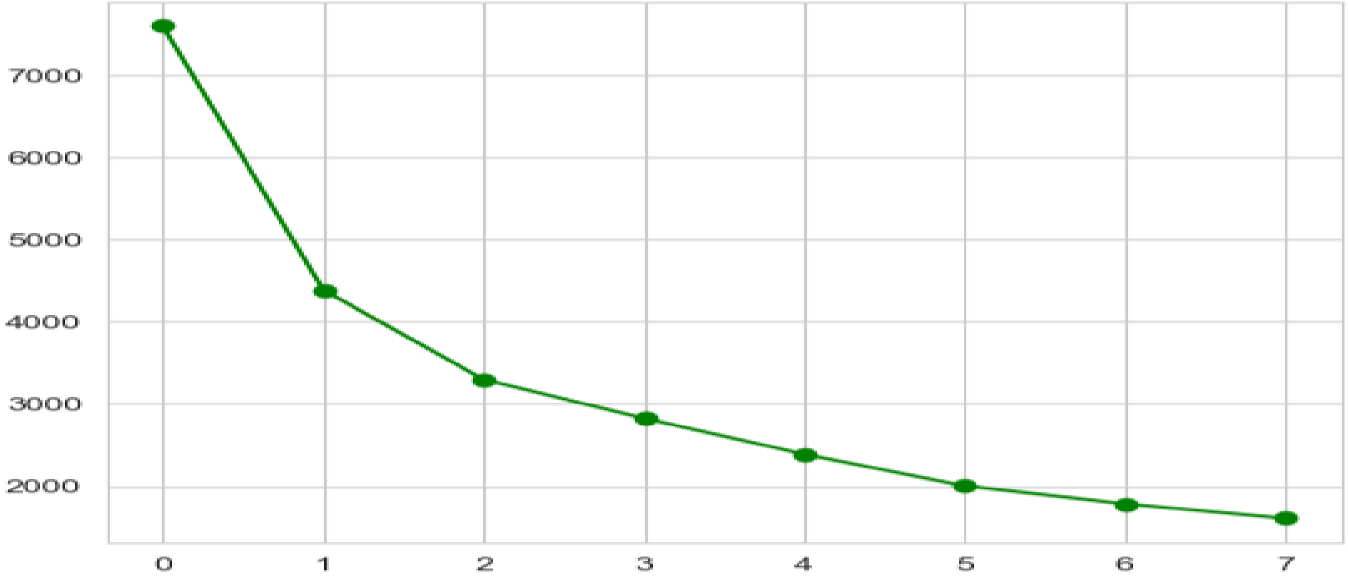
1. Inertia: Inertia measures the sum of squared distances of samples to their closest cluster center. A lower inertia value indicates better clustering performance. We can use the **inertia_** attribute of the K Means object in Scikit-Learn to calculate the inertia of the model.

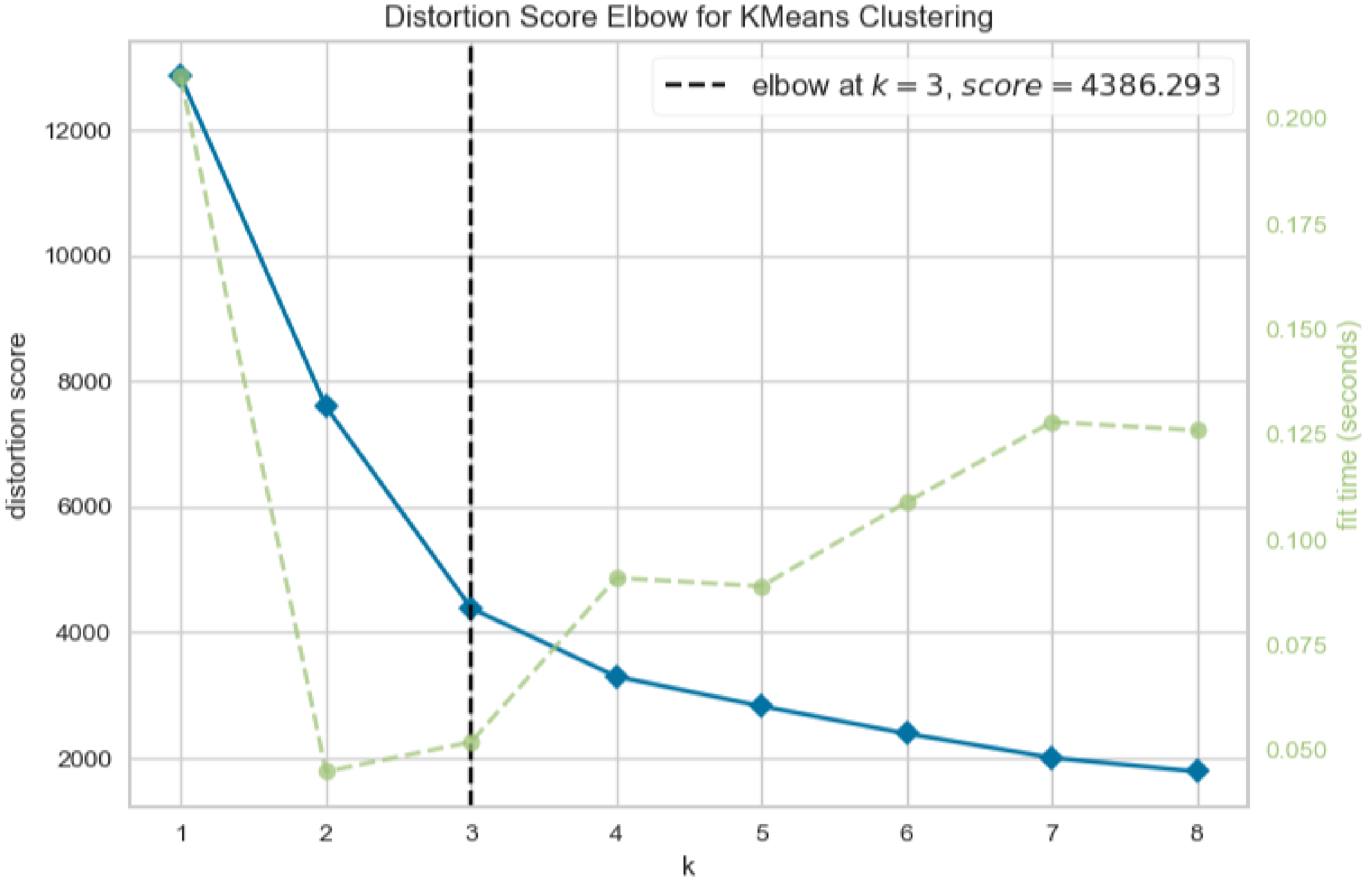
Inertia for K = 3, 4386.29, and K = 5 is 2826.45. Here, Inertia the K value is less as compared to other K values which can be observed better in the relevant notebook file of this project.

2. Silhouette Score: The silhouette score measures how similar an object is to its own cluster compared to other clusters. It takes values between -1 and 1, with values closer to 1 indicating better clustering performance. We can use the silhouette_score() function from Scikit-Learn to calculate the silhouette score of the model.

For n_clusters=2, the silhouette score is 0.5322166790563784
For n_clusters=3, the silhouette score is 0.5018866215764767
For n_clusters=4, the silhouette score is 0.48340729083864337
For n_clusters=5, the silhouette score is 0.40138887489017216
For n_clusters=6, the silhouette score is 0.40492416091010375
For n_clusters=7, the silhouette score is 0.394889950795094
For n_clusters=8, the silhouette score is 0.3784855169421591
For n_clusters=9, the silhouette score is 0.37172758026211017

3. Elbow Method: The elbow method is a graphical technique for evaluating the performance of a K-means model by plotting the inertia as a function of the number of clusters. The idea is to choose the number of clusters at the "elbow" of the curve, where the inertia starts to decrease at a slower rate. We can use the WCSS score to visualize it.





4. Visual Inspection: We can also visually inspect the clustering results to evaluate the quality of the model. You can use scatter plots or other visualization techniques to visualize the data points with their assigned cluster labels. If the clusters are well separated and there is minimal overlap, it indicates good clustering performance.

9. Inferences from the Same

1. Three Clusters (Customer Segments): Carefully examining the three cluster classification, we observe the following groups of customers:

a. High-value customer: Cluster 1 is the high-value customer segment for the online retail store as the customers in this group place the highest value orders with a very high relative frequency than other members. They are also the ones who have transacted the most recently.

b. Medium-value customer: Cluster 0 appears to be the medium-valued customer segment. These customers place an order of a considerable amount, though not as much as high-valued customers, but still quite higher than low-valued customers. Also, their orders are relatively more frequent than the lowest-value segment.

c. Low-value customer: It is quite evident that Cluster 2 has customers who rarely shop and when they order, their orders are pretty low value.

Apart from the numbers, the visualization of clusters in Silhouette Analysis show that all three customer segments are quite distinct with very less overlap between them. The general trend resonated in these 3 clusters is that high monetary value is correlated with a high frequency of orders and more recent ones.

2. Five Clusters (Customer Segments): In five clusters, we find the following customer segments:

a. Overall high-valued customers: Cluster 0 is the typical high-value customer who has shopped recently and shops regularly for high-value orders.

b. High monetary value but less frequent: Cluster 4 represents a peculiar customer segment who place quite a high-valued order but do not do so frequently or have not done much recently. But, these customers do hold a lot of promise if targeted to improve sales.

c. Medium value - low frequency - recent customers: The customers from Cluster 3 have recently placed medium-valued orders but do not do so frequently.

d. Medium value - low frequency - older customers: The customers from Cluster 2 and 0 happen to place medium valued orders quite a long time ago and they do not do so frequently.

e. Low-valued customers: Cluster 1 is the segment of customers who have not shopped in the longest time, nor do they shop frequently and their orders are of the lowest values.

The visualization of clusters in Silhouette Analysis shows some overlap between the customer segments.

However, the dataset does not distinguish between wholesale and retail customers, it is quite likely that high-value frequent clients are the wholesale dealers and medium low-valued ones are individual retail purchasers.

10. Future Possibilities of the Project

At this juncture, it makes sense to show interested stakeholders the cluster solutions and get their input. The decision should be based upon how the business plans to use the results, and the level of granularity they want to see in the clusters. What range of customer behavior from high-to-low-value customers are the business stakeholders interested in exploring? And from the answer to that question, various methods of clustering can be further exploited whether applied to RFM variables or directly to the transaction dataset available.

In the near future, we can do Classification for future customers in appropriately defined clusters by trying different models like SVC, and Logistic Regression on the training set and compare their performance on the test set in order to choose the best one to use for our predictions.

11. Conclusion

We can check the median of each variable (Frequency, Monetary, Recency) in each cluster in order to understand what customers each cluster represents.

The customer segments thus deduced can be very useful in targeted marketing, scouting for new customers, and ultimately revenue growth. After knowing the types of customers, it depends upon the retailer's policy whether to chase the high-value customers and offer them better service and discounts or try and encourage low & medium value customers to shop more frequently or of higher monetary value.

These disadvantages of k-means mean that for many datasets (especially low-dimensional datasets) it may not perform as well as you might hope. Here comes Gaussian Mixture Model (GMM) can help by providing greater flexibility due to clusters having unconstrained covariances and allowing probabilistic cluster assignment.

Reference: Python Data Science Handbook by Jake VanderPlas.

12. References:

1. https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/#:~:text=The%20two%20most%20popular%20metrics,which%20you%20will%20explore%20next.
2. https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/
3. https://machinelearningmastery.com/clustering-algorithms-with-python/
4. https://www.kaggle.com/code/thiagopanini/customer-segmentation-eda-and-kmeans
5. https://www.kaggle.com/code/thiagopanini/customer-segmentation-eda-and-kmeans
6. https://www.putler.com/rfm-analysis/
7. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a