Capstone Project – Online retail segmentation

Table of Contents

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.

Project Objective

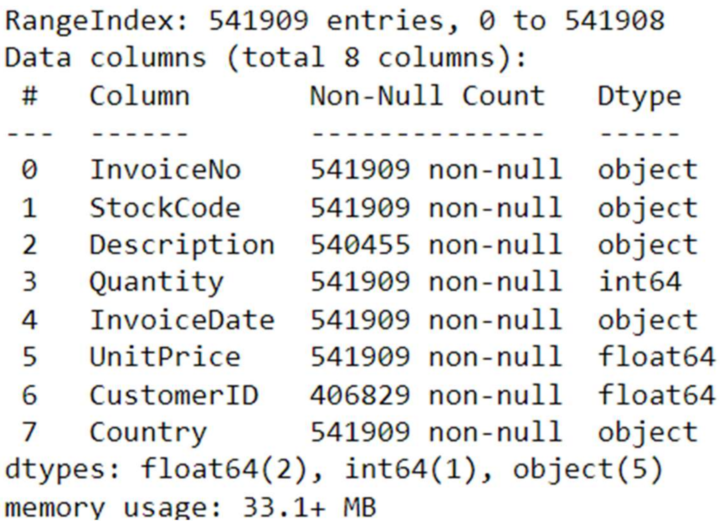
The objective of this project is to analyze customer purchase patterns for an online retail store and provide evidence-based insights to better understand customer behavior and preferences. This information can be used to inform marketing and sales strategies, improve customer engagement, and drive business growth.

Data Description

The online_retail.csv contains 387961 rows and 8 columns.

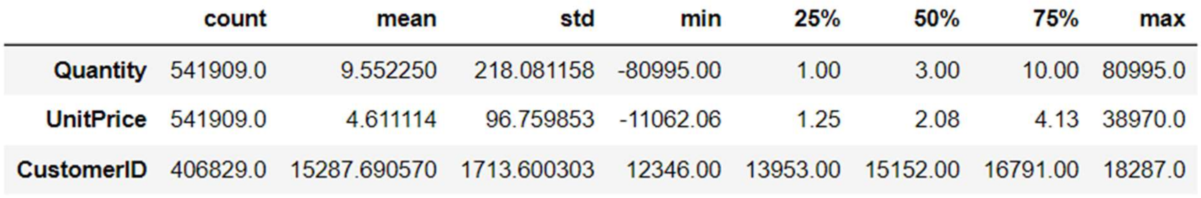
Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	Customer ID
Country	Region of Purchase

1.Data information: In the below table we can find the data type of each column and total non-null values.

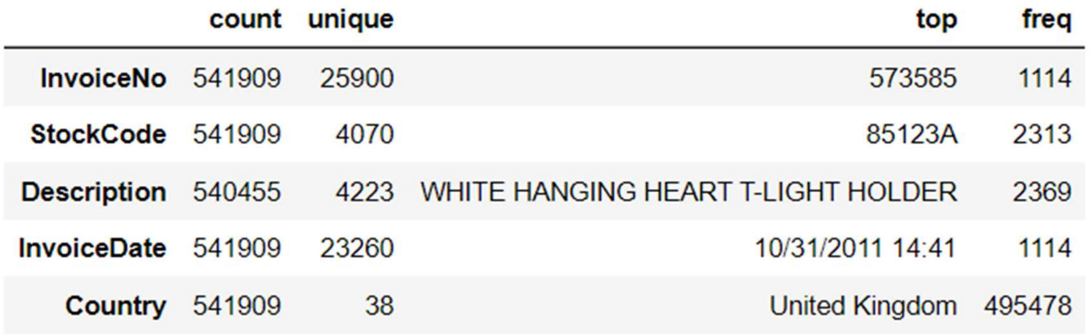


```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

2.Data Description: The data description showing the total count, mean, standard deviation(std), minimum, etc. for the numeric features is as shown in the Table.

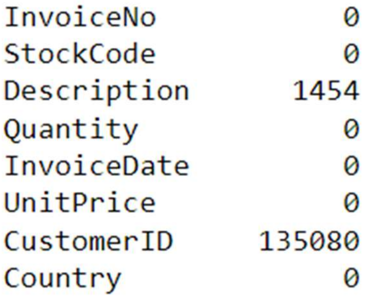


	count	mean	std	min	25%	50%	75%	max
Quantity	541909.0	9.552250	218.081158	-80995.00	1.00	3.00	10.00	80995.0
UnitPrice	541909.0	4.611114	96.759853	-11062.06	1.25	2.08	4.13	38970.0
CustomerID	406829.0	15287.690570	1713.600303	12346.00	13953.00	15152.00	16791.00	18287.0



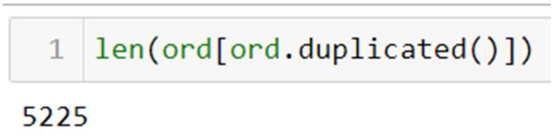
	count	unique	top	freq
InvoiceNo	541909	25900	573585	1114
StockCode	541909	4070	85123A	2313
Description	540455	4223	WHITE HANGING HEART T-LIGHT HOLDER	2369
InvoiceDate	541909	23260	10/31/2011 14:41	1114
Country	541909	38	United Kingdom	495478

3. check for Null values:



```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
```

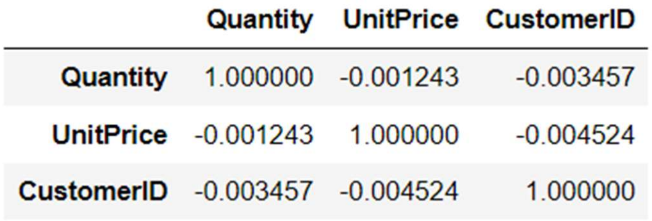
3. check for Duplicate values:



```
len(ord[ord.duplicated()])
```

```
5225
```

4. Correlation of the features:



	Quantity	UnitPrice	CustomerID
Quantity	1.000000	-0.001243	-0.003457
UnitPrice	-0.001243	1.000000	-0.004524
CustomerID	-0.003457	-0.004524	1.000000

Data Preprocessing Steps And Inspiration

The preprocessing of the data included the following steps:

1. Data Cleaning: Remove missing or inconsistent data, correct data types, and handle outliers.
2. Data Transformation: Normalize the data to account for differences in scale and unit of measurement.
3. Data Reduction: Reduce the number of variables in the dataset through techniques such as feature selection or feature extraction.
4. Data Normalization: Transform the data into a standard format to improve consistency and reduce complexity.

In terms of inspiration for analyzing an online retail dataset is for:

Customer Segmentation: Identify distinct groups of customers based on their purchasing behavior, such as frequency of purchases, types of products purchased, and spending patterns.

Choosing the Algorithm for the Project

The given data set consists of online sales dataset which have large amount of sales data.

1. K-Means Clustering: K-Means is a popular algorithm for customer segmentation because it is fast, simple, and easy to interpret. It works by grouping similar customers into clusters based on their purchasing behavior.
2. Hierarchical Clustering: Hierarchical clustering is another commonly used algorithm for customer segmentation. It starts by grouping individual customers into clusters and then aggregates these clusters into larger groups as needed.

Motivation and Reasons for Choosing the Algorithm:

The reasons for choosing the KMeans and Hierarchical Clustering algorithms include:

1. Simplicity: Both KMeans and Hierarchical Clustering are relatively simple algorithms that are easy to implement and understand. They are also well suited for large datasets, making them a popular choice for customer segmentation projects.
2. Speed: Both algorithms are computationally fast, allowing for quick results and the ability to run multiple iterations and compare results.
3. Interpretability: KMeans provides clear and interpretable results, as it assigns each customer to a specific cluster based on their attributes. Hierarchical Clustering also provides a clear visual representation of the customer segments in the form of a dendrogram.
4. Scalability: Both algorithms are scalable, meaning that they can handle large datasets and many clusters without significant performance degradation.

Assumptions:

The following assumptions were made to create the segmentation model for Online retail project.

1. Symmetrical Distribution of Data: K-Means and Hierarchical Clustering assume that the data is symmetrically distributed, which means that the data points are evenly spread out in the feature space.
2. Independence of Variables: K-Means and Hierarchical Clustering assume that the variables in the dataset are independent, which means that the variables do not affect each other.
3. The Number of Clusters: Both K-Means and Hierarchical Clustering require the number of clusters to be specified ahead of time, which can be challenging in some cases. The number of clusters can be determined using methods such as the elbow method or silhouette score.

Model Evaluation and Technique

There are several methods for evaluating the performance of K-Means and Hierarchical Clustering models:

1. Silhouette Score: The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, with a higher score indicating a better fit.

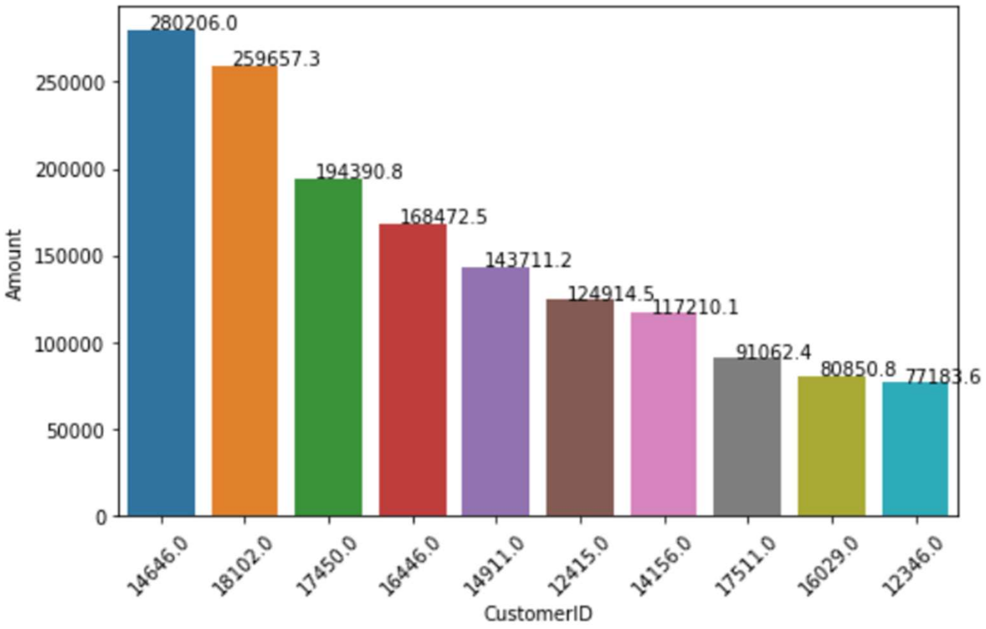
2. Calinski-Harabasz Index: The Calinski-Harabasz index is a measure of the ratio of the between-cluster variance to the within-cluster variance. A higher score indicates a better fit.

3. Davies-Bouldin Index: The Davies-Bouldin index is a measure of the average similarity between each cluster and its most similar cluster. A lower score indicates a better fit.

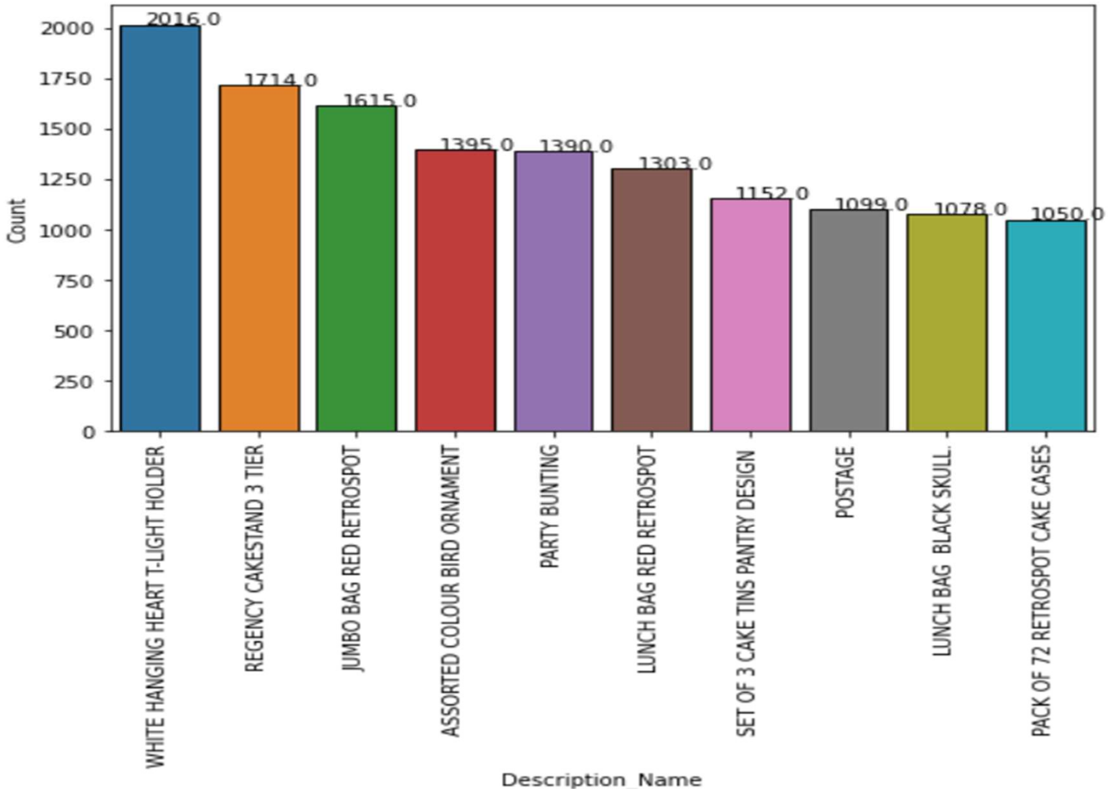
4. Elbow Method: The elbow method is a visual method for determining the optimal number of clusters by plotting the inertia against the number of clusters and selecting the number of clusters where the decrease in inertia begins to level off.

Inferences from the Project

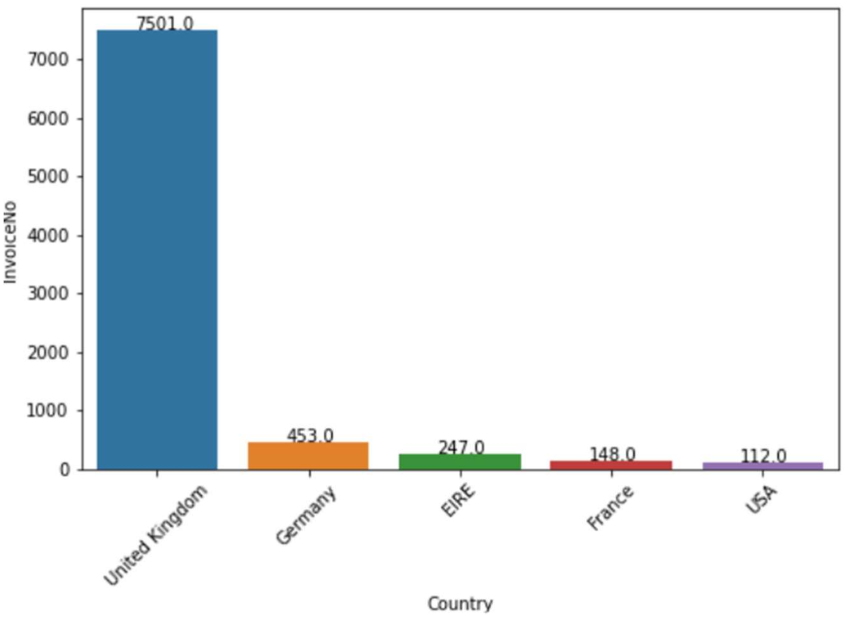
Top 10 customers who spent more amount than others.



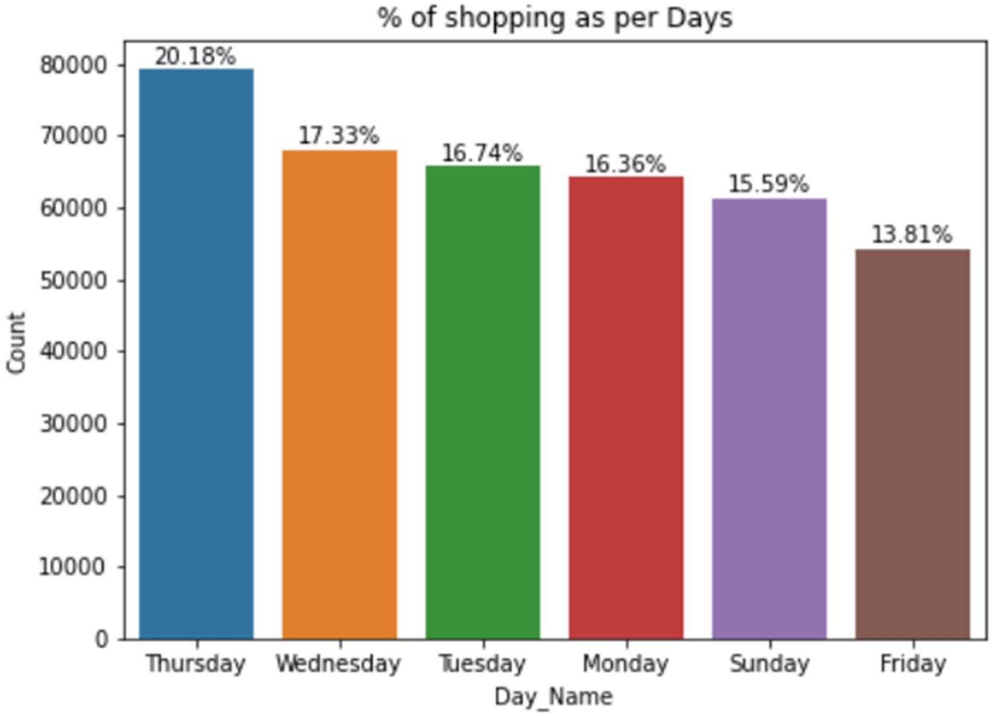
Top 10 most bought products:



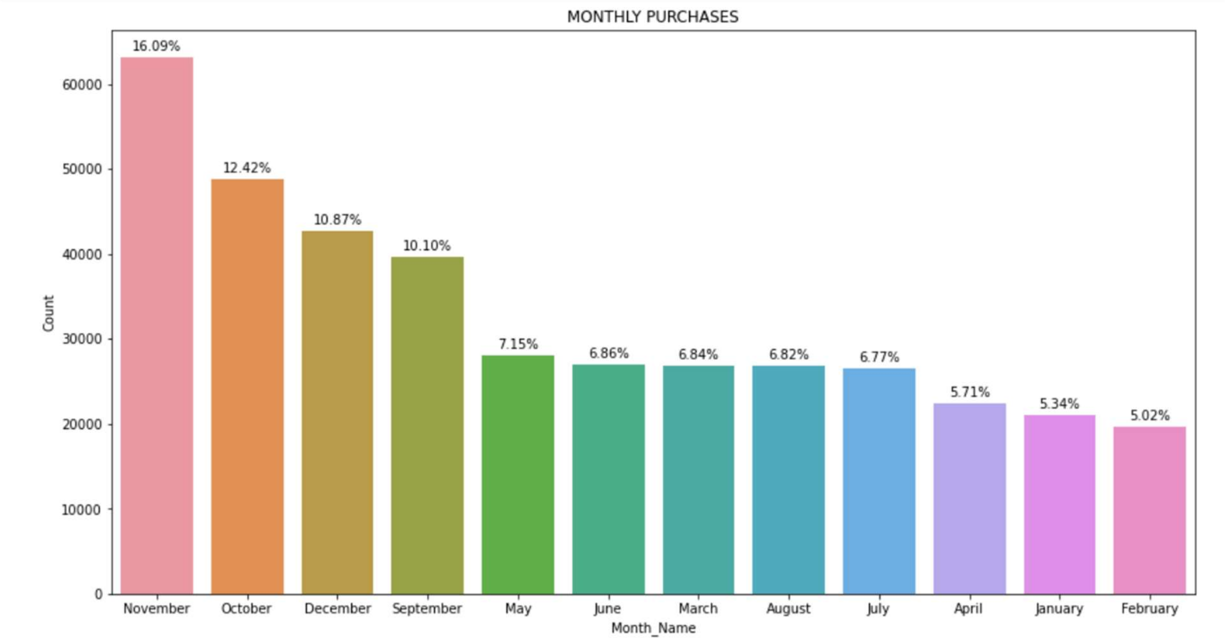
Top 5 countries orders are cancelled.



Percentage of shopping as per days



Percentage of monthly purchases.



Future scope:

The future scope of this project is summarized as follows:

1. Personalized Marketing: By better understanding the customer purchase patterns, the firm can create targeted marketing campaigns and personalized promotions to increase customer engagement and sales.

2. Predictive Analysis: The data collected from this project can be used for predictive analysis to identify future trends, forecast sales, and make informed decisions.

3. Customer Segmentation: Further refinement of the customer segments can be done to get a deeper understanding of the behavior and needs of each segment, leading to improved customer experience and satisfaction.

Conclusion

After evaluating the performance of both algorithms using various evaluation techniques, KMeans Clustering provides a better silhouette score, Calinski-Harabasz index, and Davies-Bouldin index and a lower inertia, it can be concluded that KMeans Clustering is a better fit for the data and should be chosen for the segmentation.