Capstone Project : 4

Online Retail Customer Segmentation

By

Pankaj Kumar

Ankit Sharma

Data Science Trainee,

Alma Better

Understanding Business Problem

→ Topic – “Online Retail Customer Segmentation”

→ Problem Statement :

- “In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers”.

Aim and results expected through Project

- Through this project we will try to predict the buying behaviour of the customers .
- Also we will try to find how the profit of the retail company can be maximized.

Dataset Information

→This dataset contains 541909 observations and 8 features that contain the data of between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

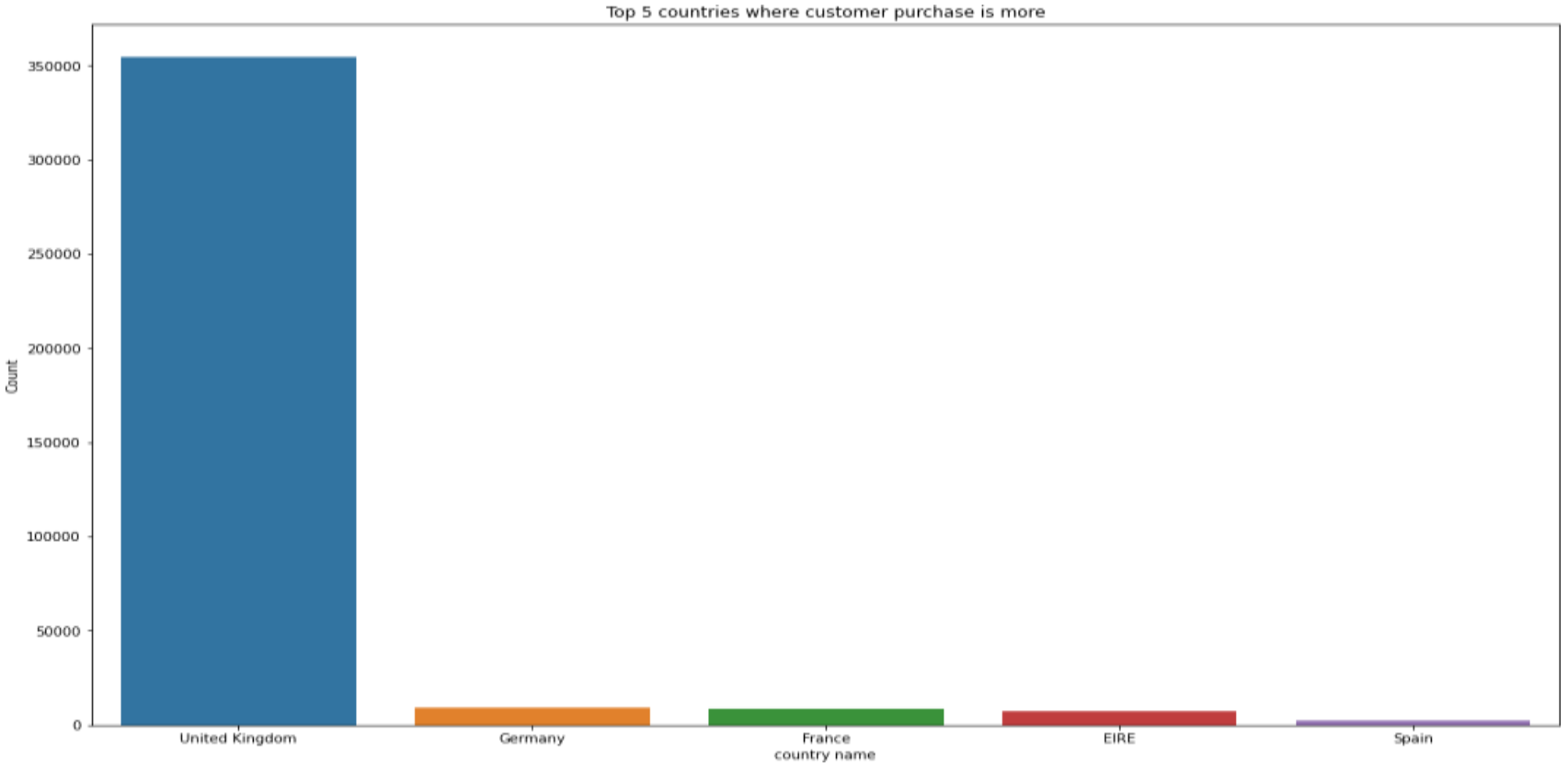
- There are 3 categorical features in our dataset.
- This dataset have null and duplicate values.

	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

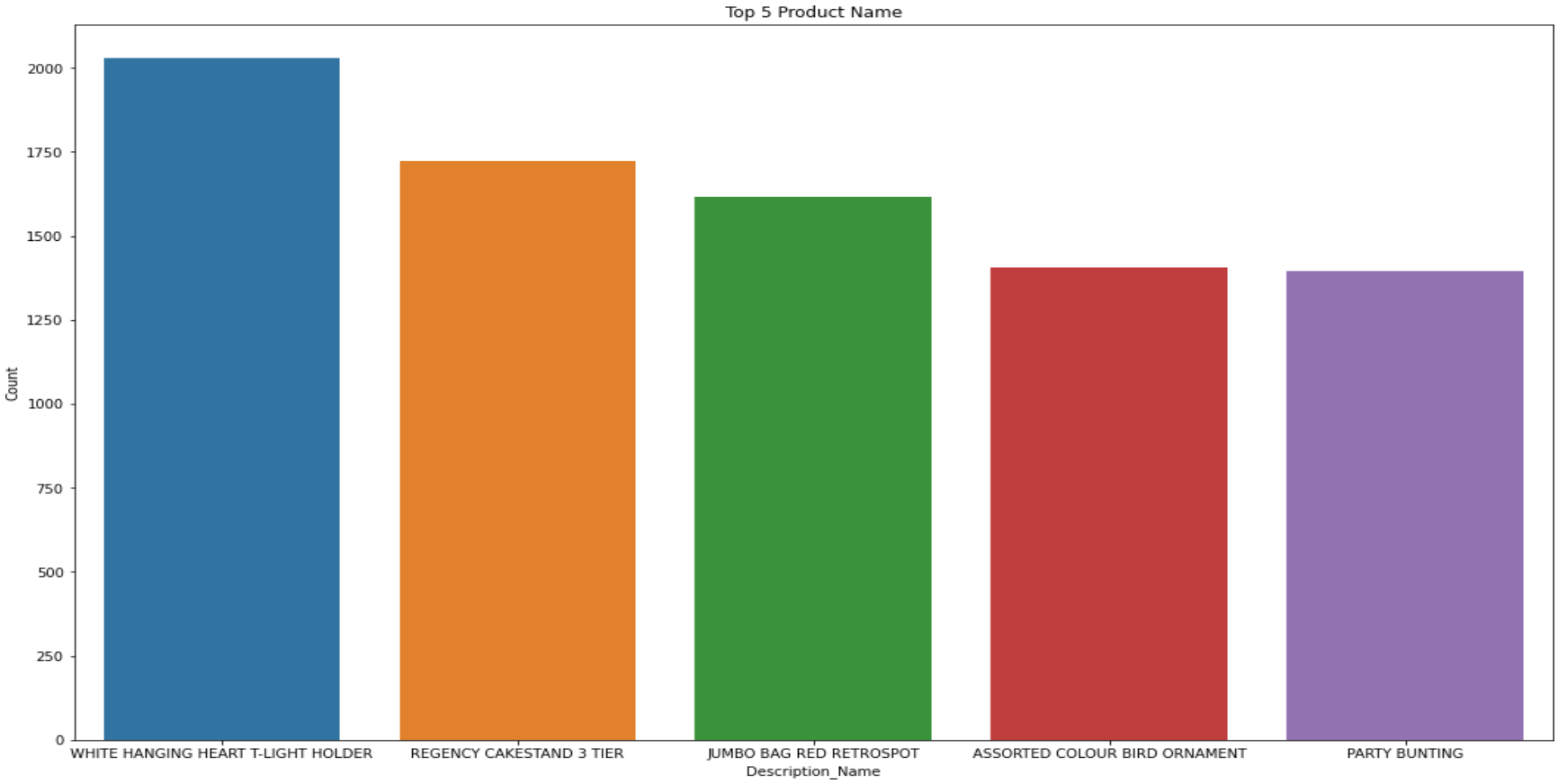
Feature Summary

- **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates acancellation.
- **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned toeach distinct product.
- **Description**: Product (item) name. Nominal.
- **Quantity**: The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate**: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
- **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned toeach customer.
- **Country**: Country name. Nominal, the name of the country where each customer resides.

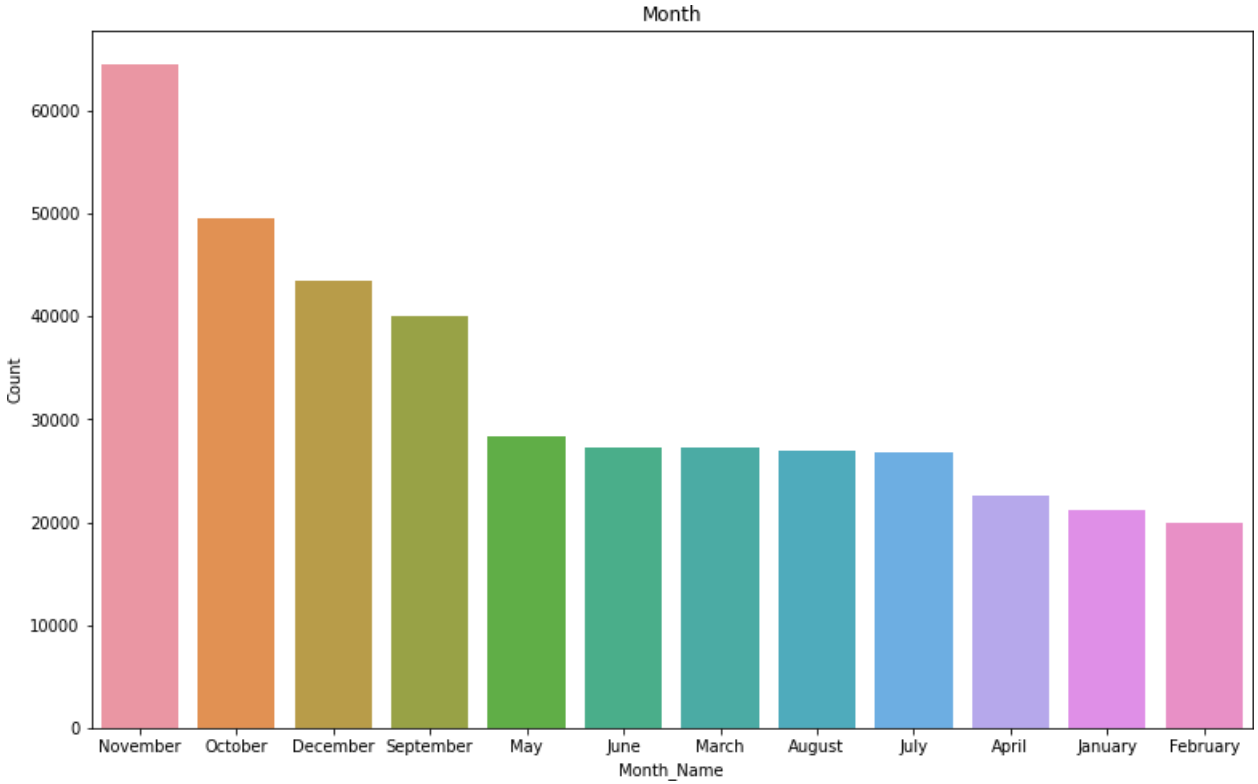
Categorical Feature Analysis Of Country column



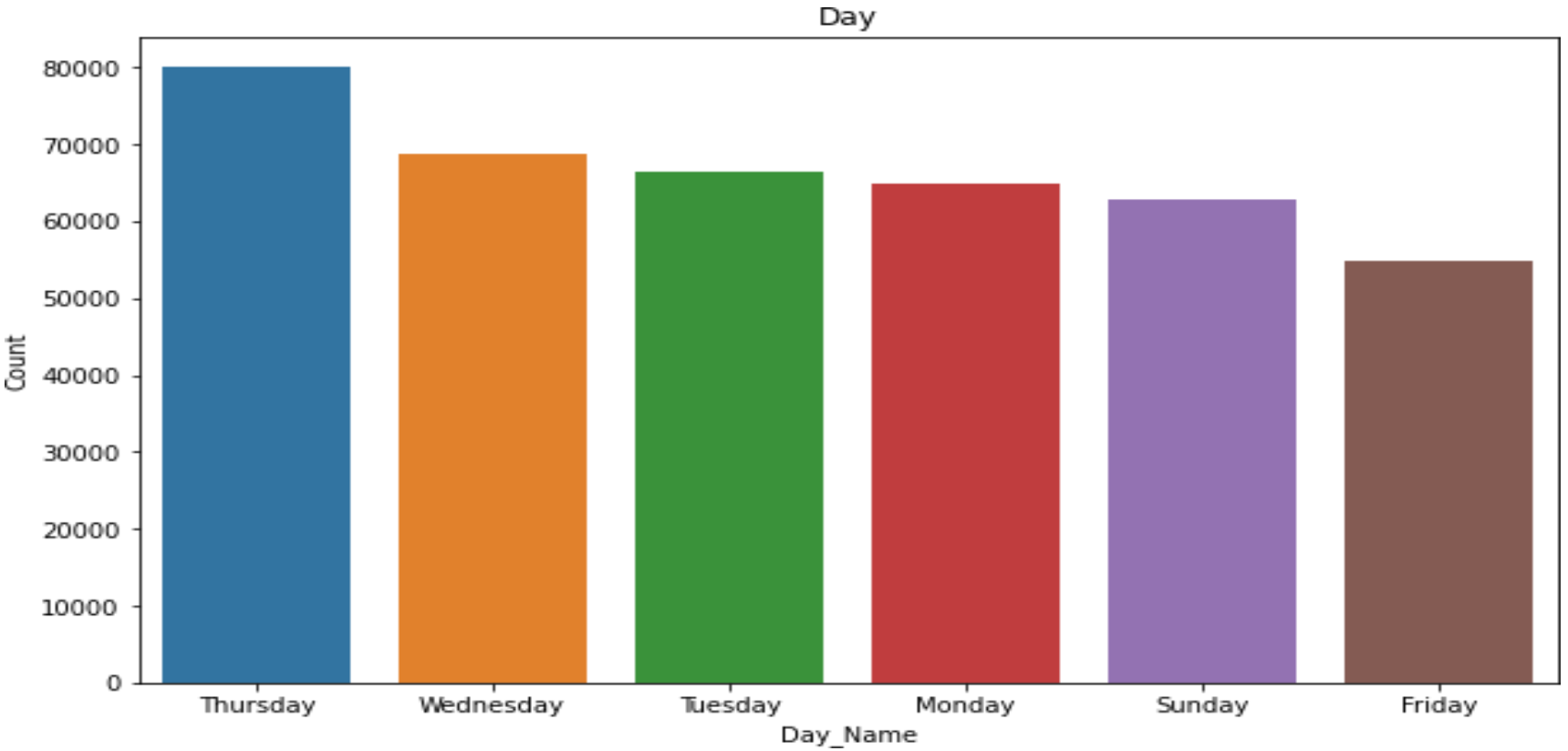
Analysis Of Product column



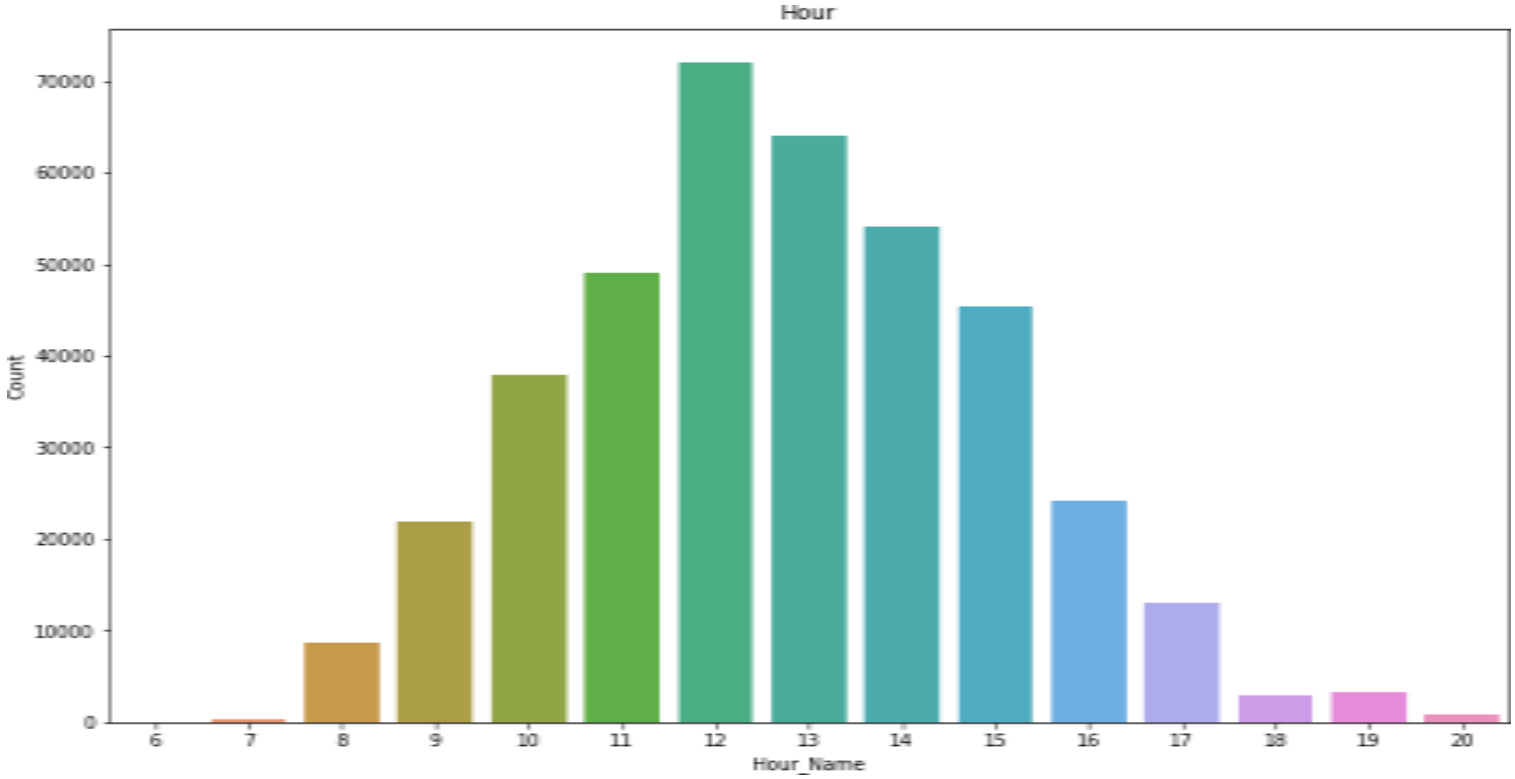
Analysis Of Month Column



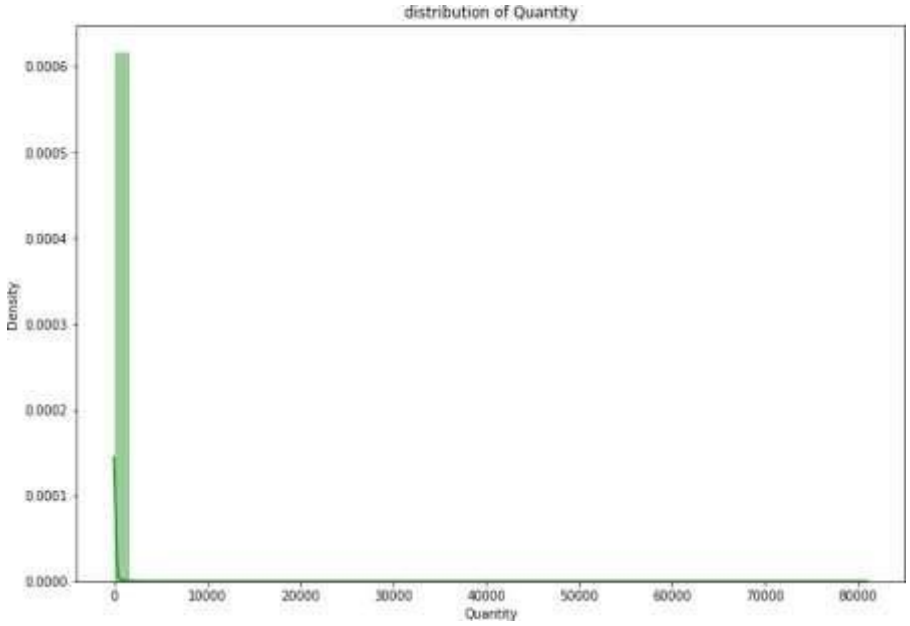
Analysis Of Day Column

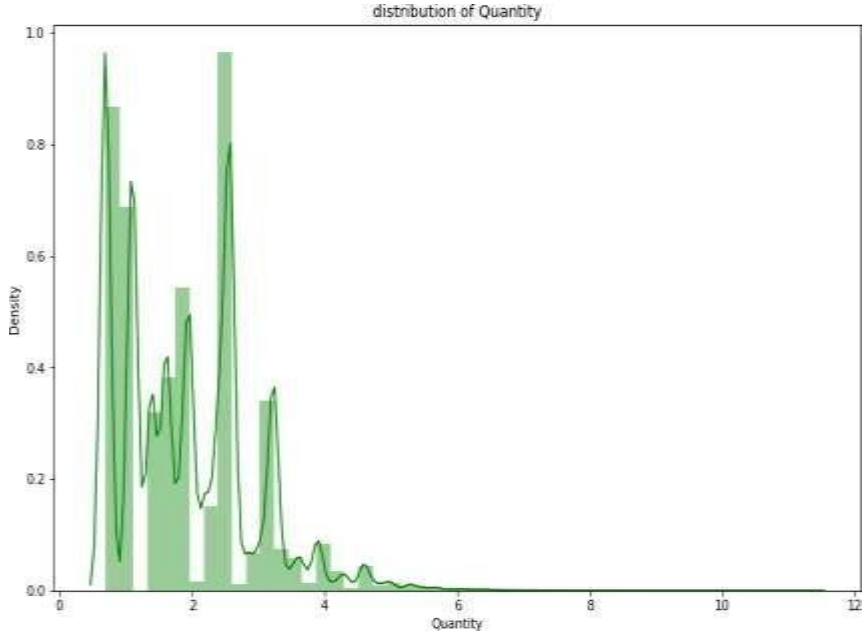


Analysis of Day Column Continued



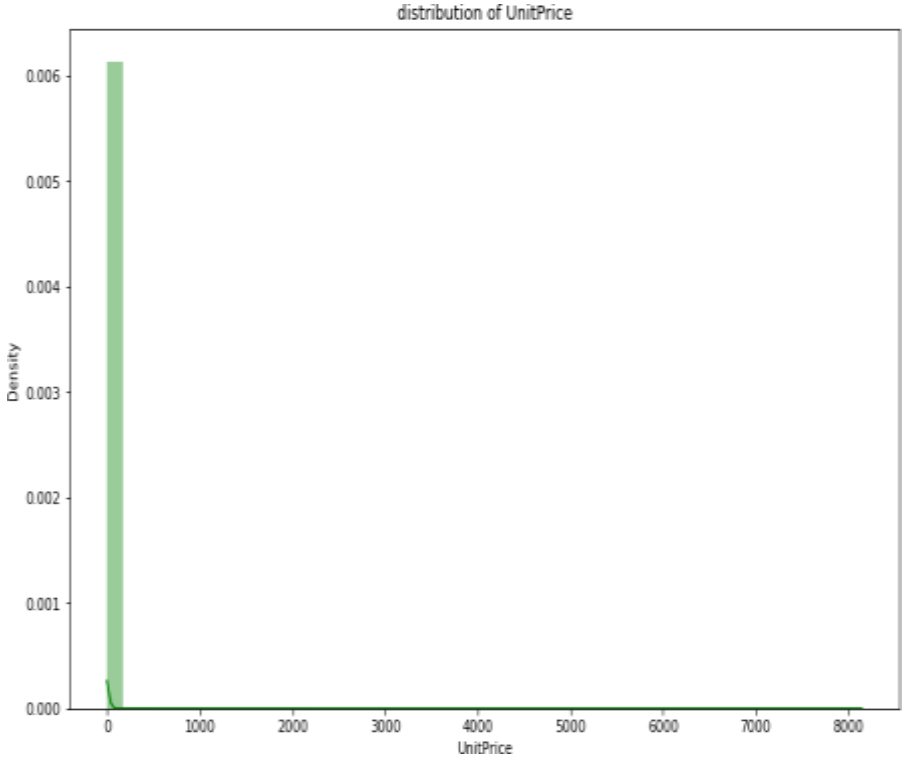
Analysis Of Numerical column

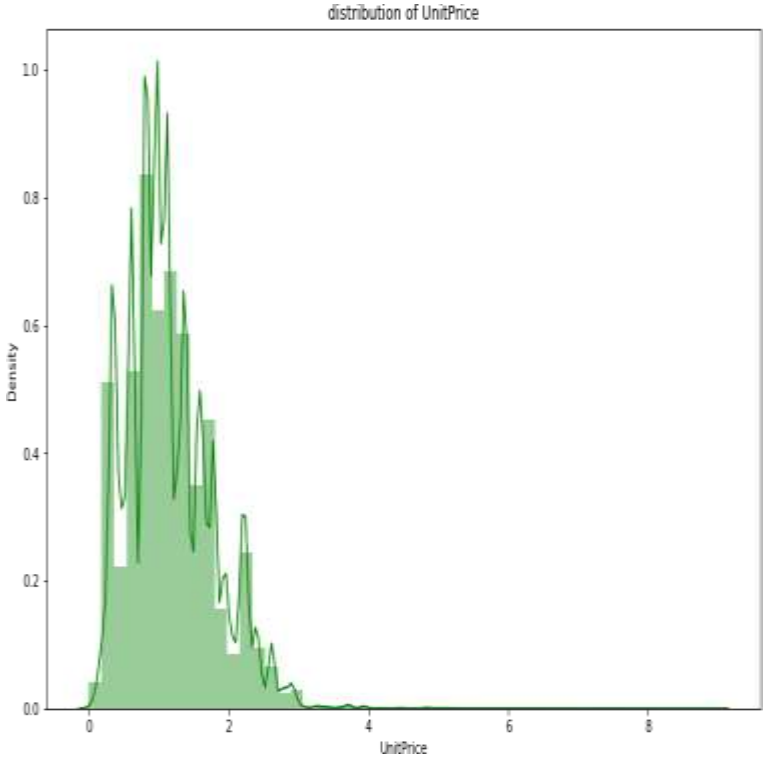




Data transformation

Analysis Of Numerical column





Data transformation

Conclusion

- ✓ We can conclude that most of the customers have purchase the items in Thursday Wednesday and Tuesday
- ✓ The most numbers of customers have purchase the gifts in the month of November ,October and December September
- ✓ Most of the customers have purchase the items in Afternoon ,moderate numbers of customers have purchase the items in Morning and least numbers of customers have purchase the items in Evening
- ✓ Top countries with respect to purchase are UK Germany France
- ✓ Top Items are WHITE HANGING HEART T-LIGHT HOLDER,REGENCY CAKESTAND 3 TIER JUMBO BAG RED RETROSPOT ASSORTED COLOUR BIRD ORNAMENT

RFM Segmentation

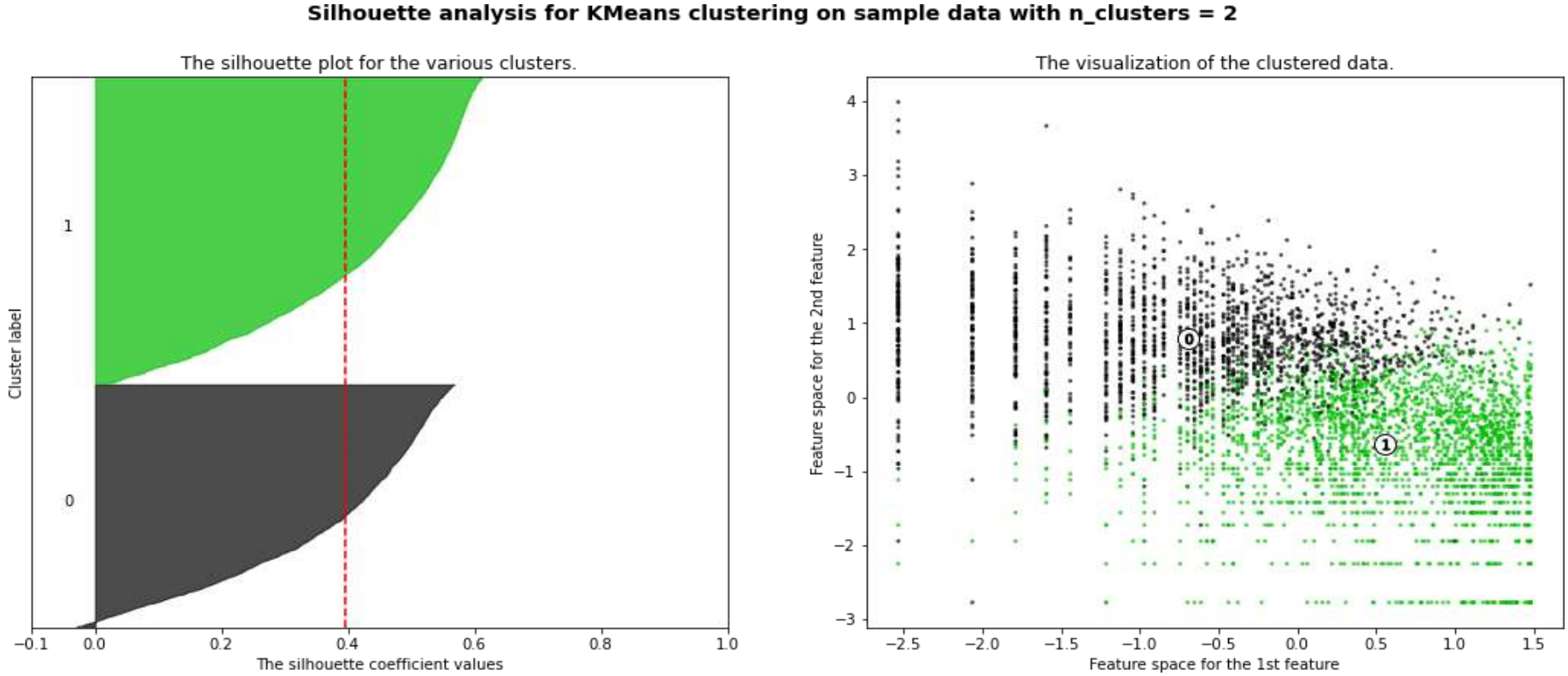
- RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their transaction was (Monetary).
- RFM analysis helps to answer the following questions: Who was our most recent customer? How many times has he purchased items from our shop? And what is the total value of his trade? All this information can be critical to understanding how good or bad a customer is to the company.
- After getting the RFM values, a common practice is to create ‘quartiles’ on each of the metrics and assigning the required order.

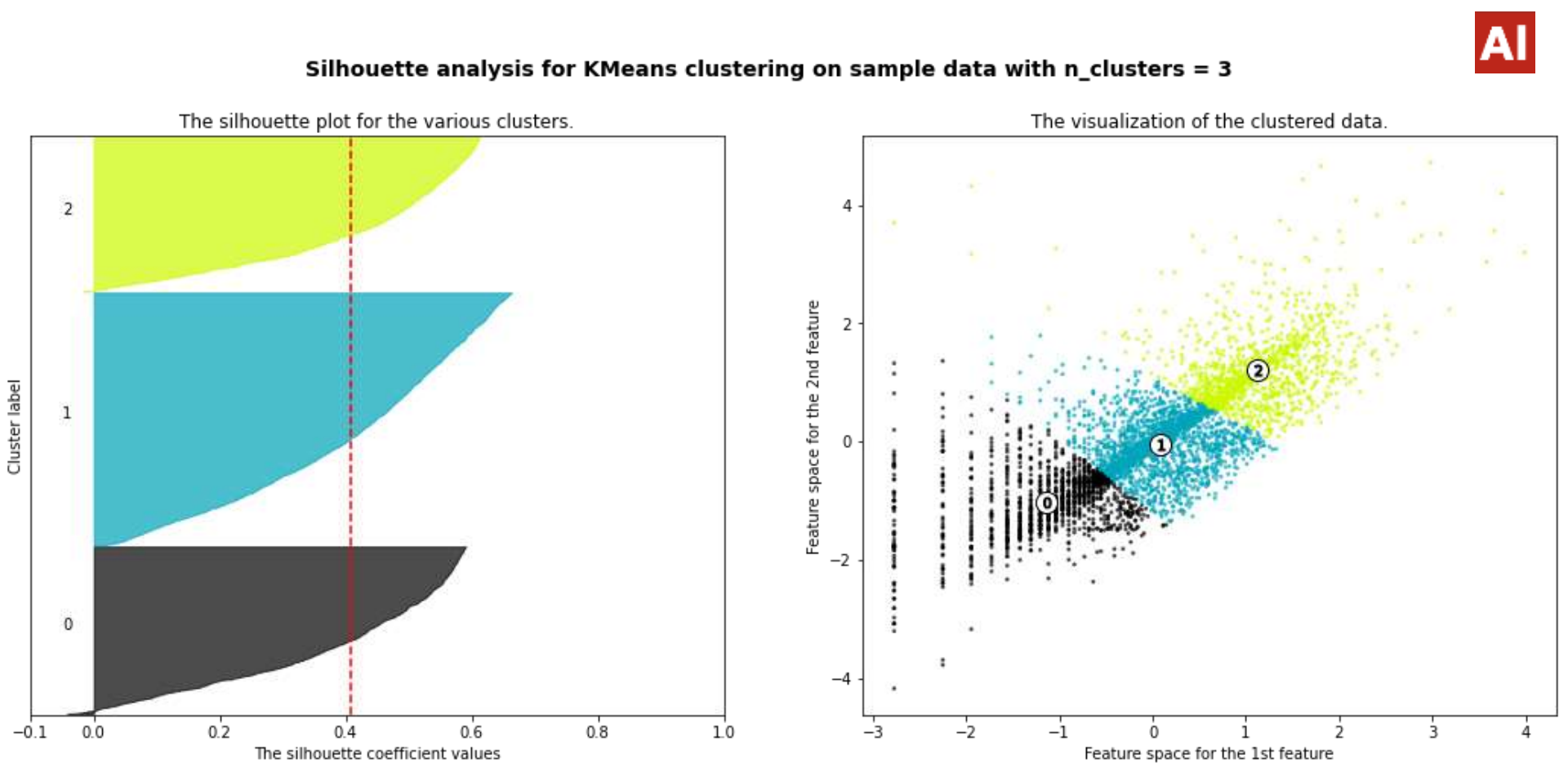
RFM Implementation

- **RFM** simply mean Recency, Frequency, Monetary

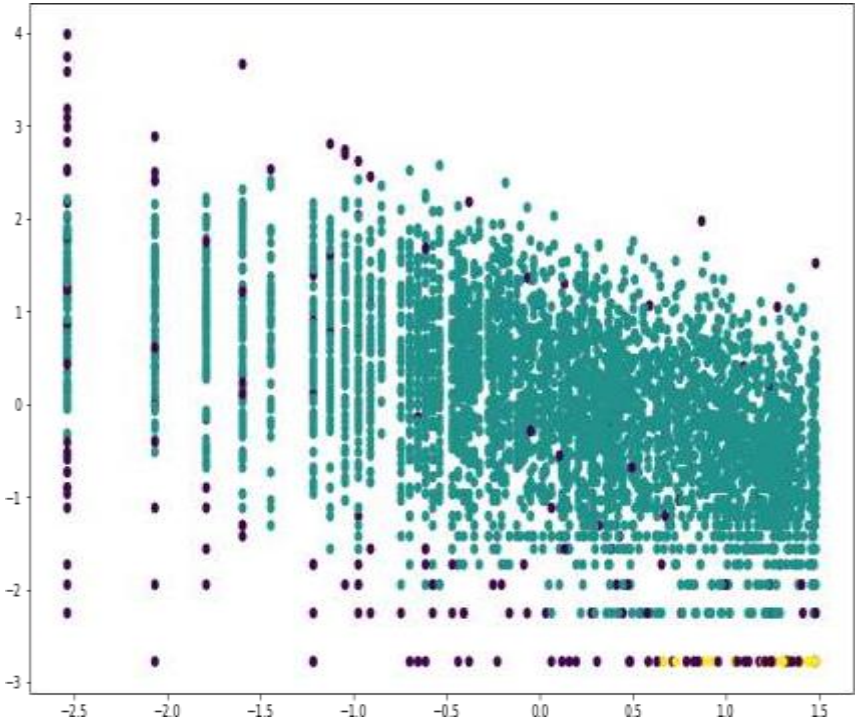
- The RFM model is based on three quantitative factors
- Frequency: How often a customer makes a purchase.
- Monetary Value: How much money a customer spends on. (Sum of Total)
- Recency = Latest Date - Last Invoice Data
- Split into four segments using quantiles

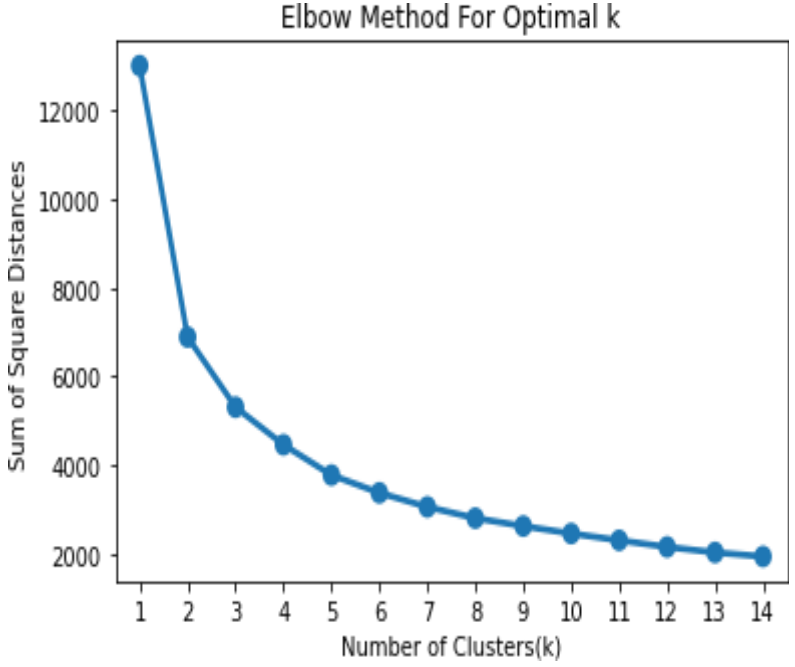
K mean clustering and Silhouette Method on Recency ,Frequency and Monetary




Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4
Feature space for the 1st feature

DBSCAN and Elbow method to Recency Frequency and Monetary





Conclusion

Although we didn't obtain two clearly separated clusters, we were able to build a model that can classify new customers into "low value" and "high value" groups. Generally, if a customer only transacted with us a few times, they needed to be at least in the top 50th percentile in monetary spending to be considered a "high value customer".

1) K-Means with silhouette_score of RFM Optimal_Number_of_cluster are 2
2) K-Means with Elbow methos of RFM Optimal_Number_of_cluster are 2

Challenges

- Large Dataset to handle
- Need to analyze lot of variable
- Null value handling
- Feature engineering
- Selecting Optimum number of cluster
- Deciding the flow of the presentation

Thank you