Capstone Project-4

Customer Segmentation

Prepared by:

Suhail Shaikh

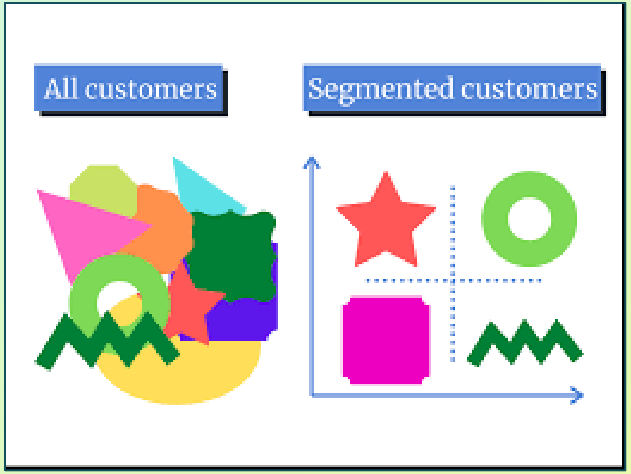
Contents

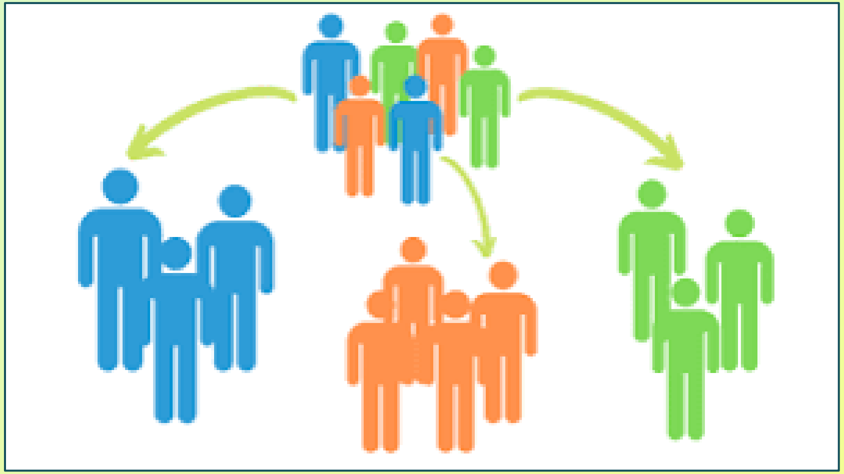
- ❑ Problem Statement
- ❑ Data Summary
- ❑ Data Cleaning
- ❑ Feature Engineering
- ❑ Exploratory Data Analysis
- ❑ Optimum Number of Clusters
- ❑ Model Implementation
- ❑ Model Validation
- ❑ Model Selection
- ❑ Conclusion
- ❑ Challenges
- ❑ References

Problem Statement

Problem Statement

- 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.





Data Summary

Data Summary

Data set has 541909 rows and 8 columns. The columns in data set have information as mentioned below :

- **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each 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.

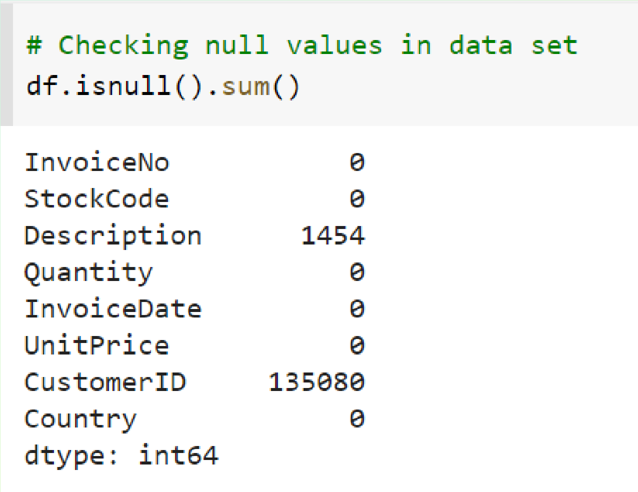
Cont….

- **UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
- **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- **Country**: Country name. Nominal, the name of the country where each customer resides.

Data Cleaning

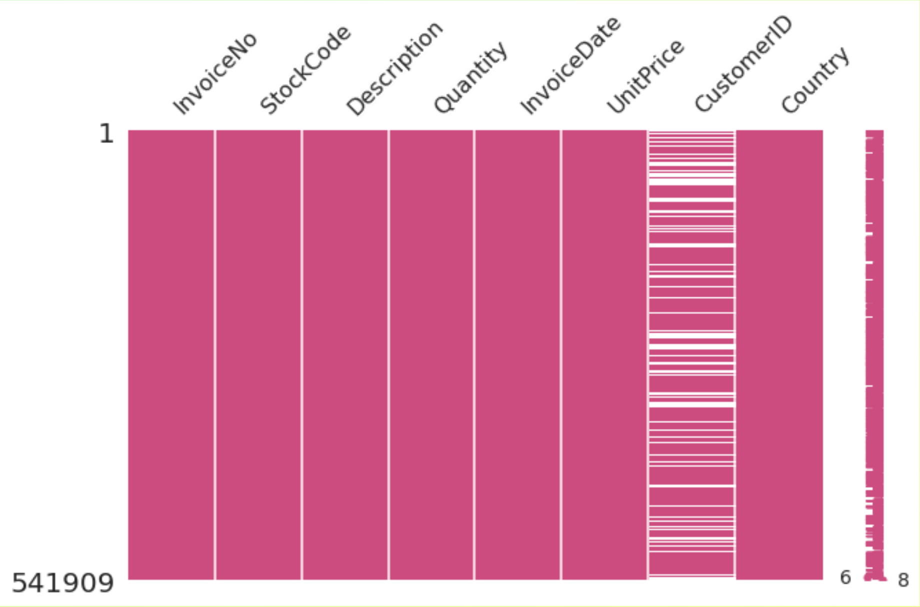
Null Values Detection

- **Checking the missing values in Data Set**



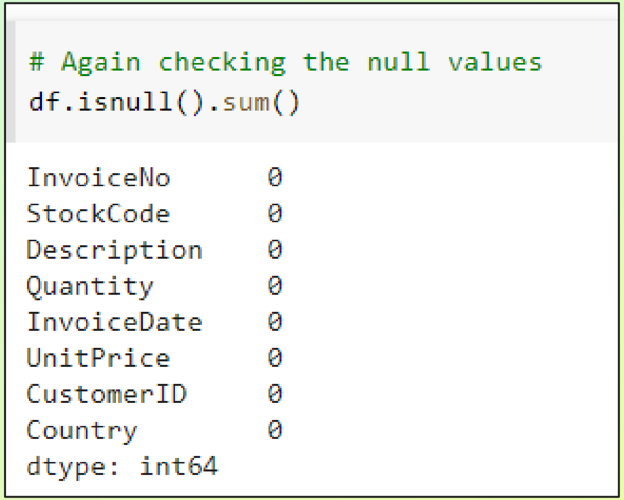
```
# Checking null values in data set
df.isnull().sum()
```

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```



Null Values Removal

- We have dropped the missing values from data set.

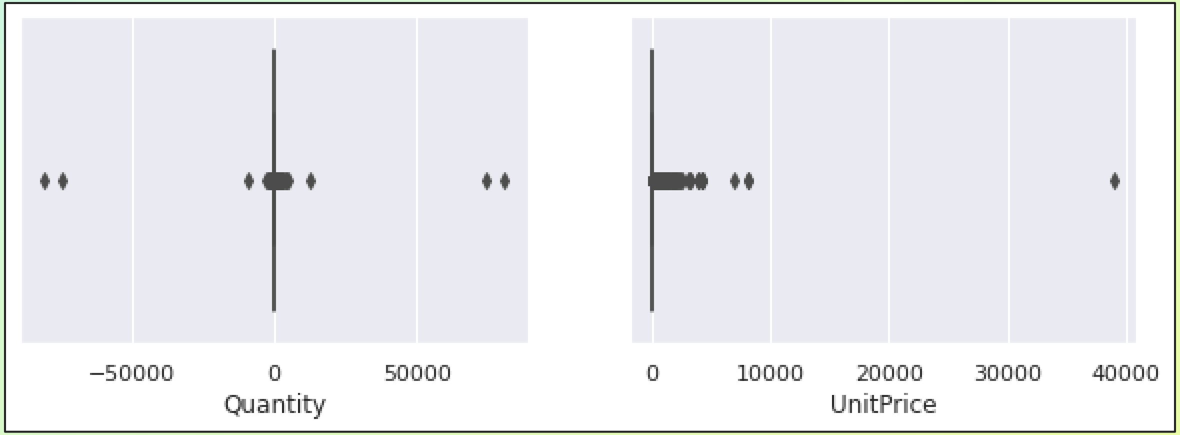


```
# Again checking the null values
df.isnull().sum()
```

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

Outlier Detection

- Box Plot Before Outlier Removal



- We have plotted boxplot for numerical features to visualize the outliers.

IQR (Interquartile Range) Method

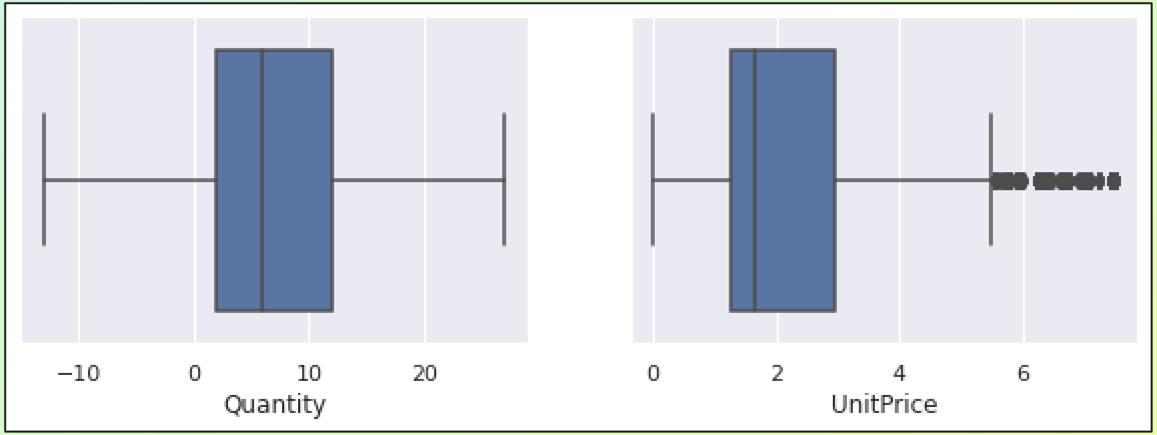
- We have used IQR Method to remove the outliers from data set. The important parameters of IQR are as follows:

  Q1 = 25% quantile
  Q3 = 75% quantile
  IQR = Q3-Q1
  Lower limit = Q1 - 1.5 * IQR
  Upper limit = Q3 + 1.5 * IQR

- We identified lower and upper limit for our numerical features and set the limits to remove the outliers.

Outliers Removal

- **Box Plot After Outliers Removal**



Feature Engineering

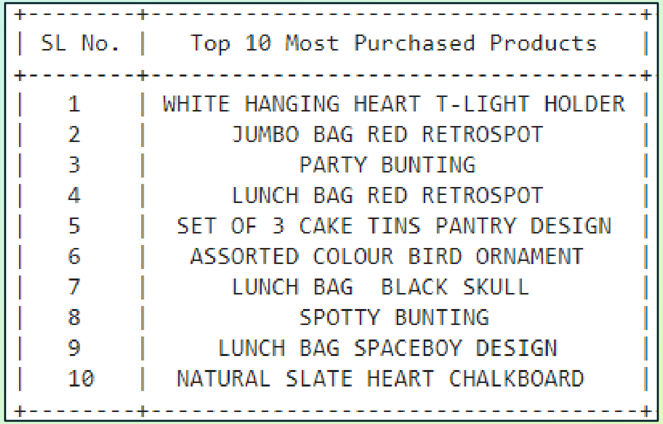
Feature Engineering

We have added the below mentioned features in our data frame

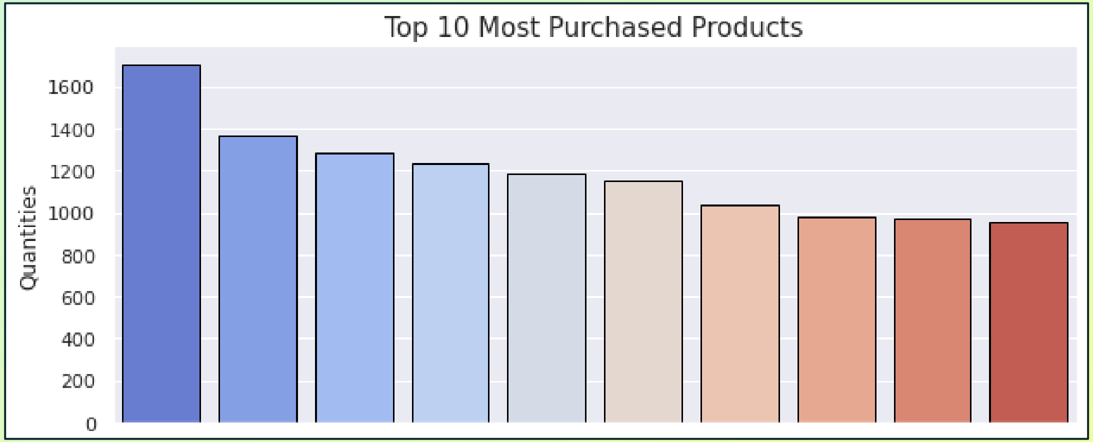
- **'Day’** - represents the name of day
- **'Month’** - represents the name of month
- **"year“** - represents the year
- **"month_num“** - shows the month number
- **"day_num“** -shows the day number
- **"hour"** - shows the time in hour
- **"minute“** - shows the time in minutes
- **"sales“** - total amount of sales

Exploratory Data Analysis

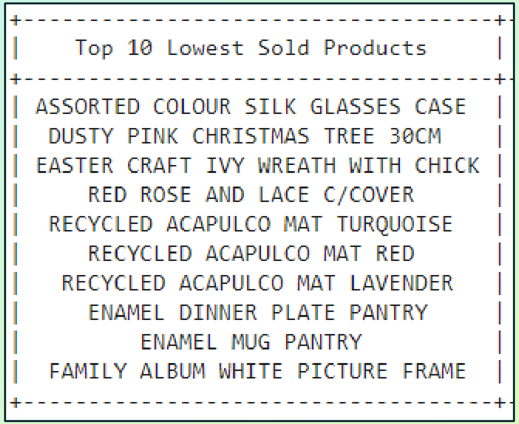
Top Ten Most Purchased Products



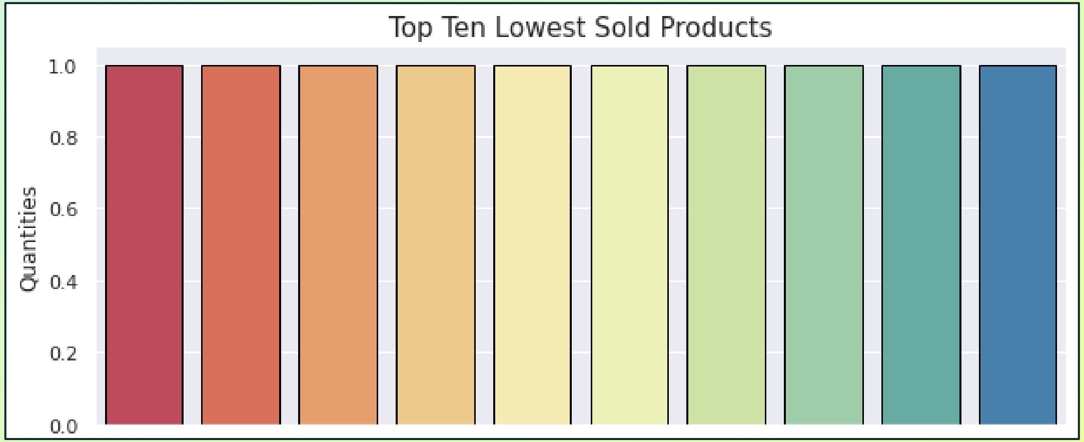
```
+---------+--------------------------------------+
| SL No.  |   Top 10 Most Purchased Products     |
+---------+--------------------------------------+
|    1    | WHITE HANGING HEART T-LIGHT HOLDER   |
|    2    |      JUMBO BAG RED RETROSPOT         |
|    3    |           PARTY BUNTING              |
|    4    |      LUNCH BAG RED RETROSPOT         |
|    5    |  SET OF 3 CAKE TINS PANTRY DESIGN    |
|    6    |   ASSORTED COLOUR BIRD ORNAMENT      |
|    7    |      LUNCH BAG  BLACK SKULL          |
|    8    |           SPOTTY BUNTING             |
|    9    |     LUNCH BAG SPACEBOY DESIGN        |
|   10    |   NATURAL SLATE HEART CHALKBOARD     |
+---------+--------------------------------------+
```



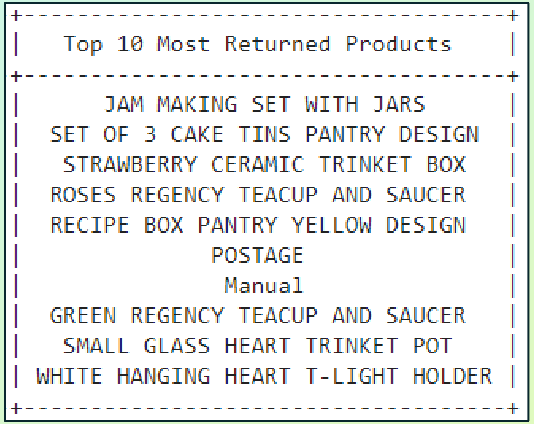
Top Ten Lowest Sold Products



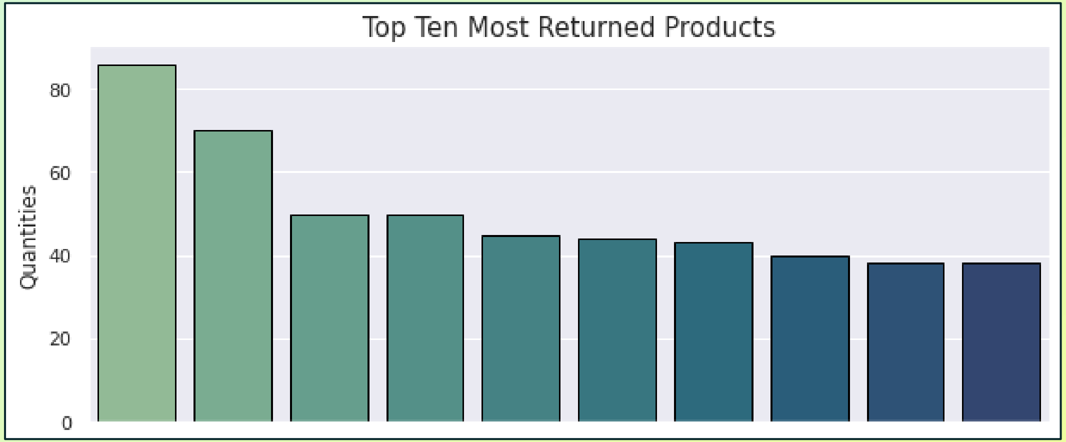
```
+-------------------------------------+
|     Top 10 Lowest Sold Products     |
+-------------------------------------+
| ASSORTED COLOUR SILK GLASSES CASE   |
|  DUSTY PINK CHRISTMAS TREE 30CM     |
| EASTER CRAFT IVY WREATH WITH CHICK  |
|     RED ROSE AND LACE C/COVER       |
|  RECYCLED ACAPULCO MAT TURQUOISE    |
|     RECYCLED ACAPULCO MAT RED       |
|   RECYCLED ACAPULCO MAT LAVENDER    |
|     ENAMEL DINNER PLATE PANTRY      |
|         ENAMEL MUG PANTRY           |
|  FAMILY ALBUM WHITE PICTURE FRAME   |
+-------------------------------------+
```



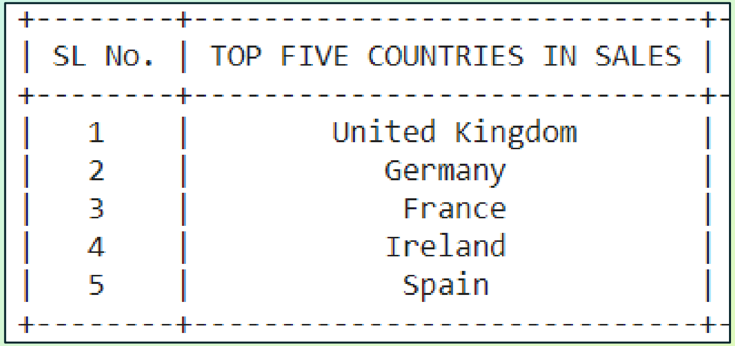
Top Ten Most Returned Products



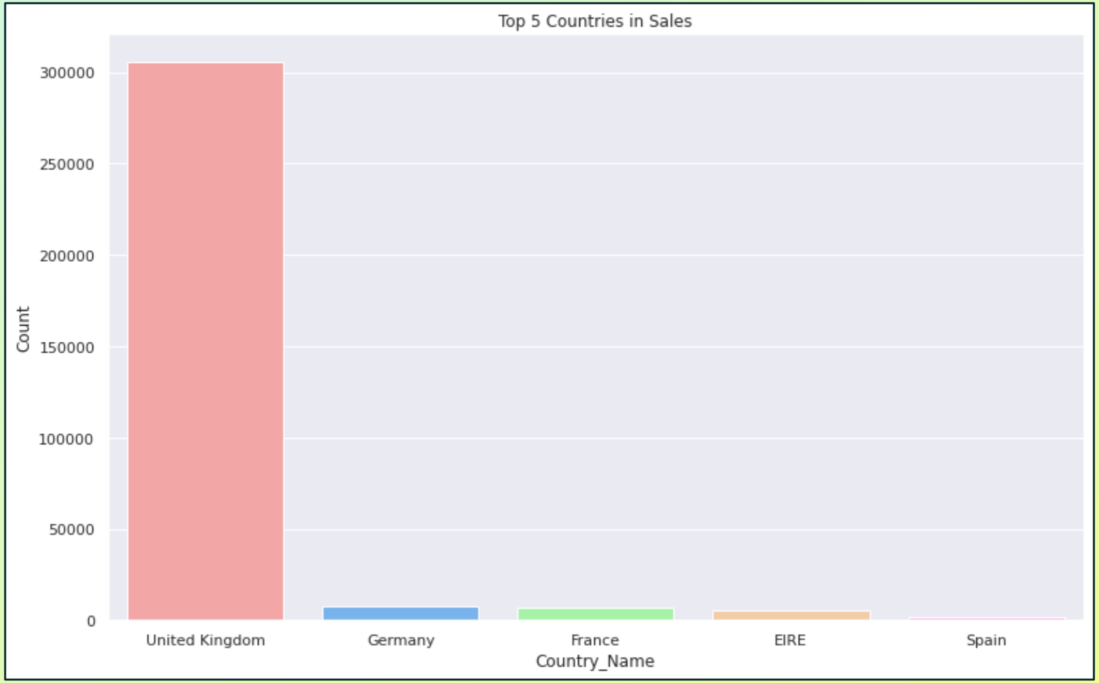
```
+-----------------------------------+
|    Top 10 Most Returned Products  |
+-----------------------------------+
|       JAM MAKING SET WITH JARS    |
|   SET OF 3 CAKE TINS PANTRY DESIGN |
|    STRAWBERRY CERAMIC TRINKET BOX |
|   ROSES REGENCY TEACUP AND SAUCER |
|   RECIPE BOX PANTRY YELLOW DESIGN |
|              POSTAGE              |
|              Manual               |
|   GREEN REGENCY TEACUP AND SAUCER |
|    SMALL GLASS HEART TRINKET POT  |
|  WHITE HANGING HEART T-LIGHT HOLDER |
+-----------------------------------+
```



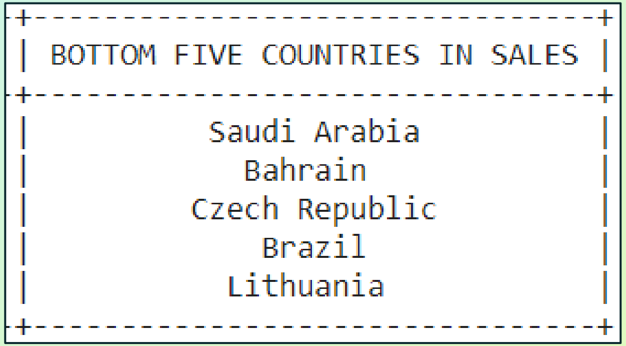
Top Five Countries in Sales



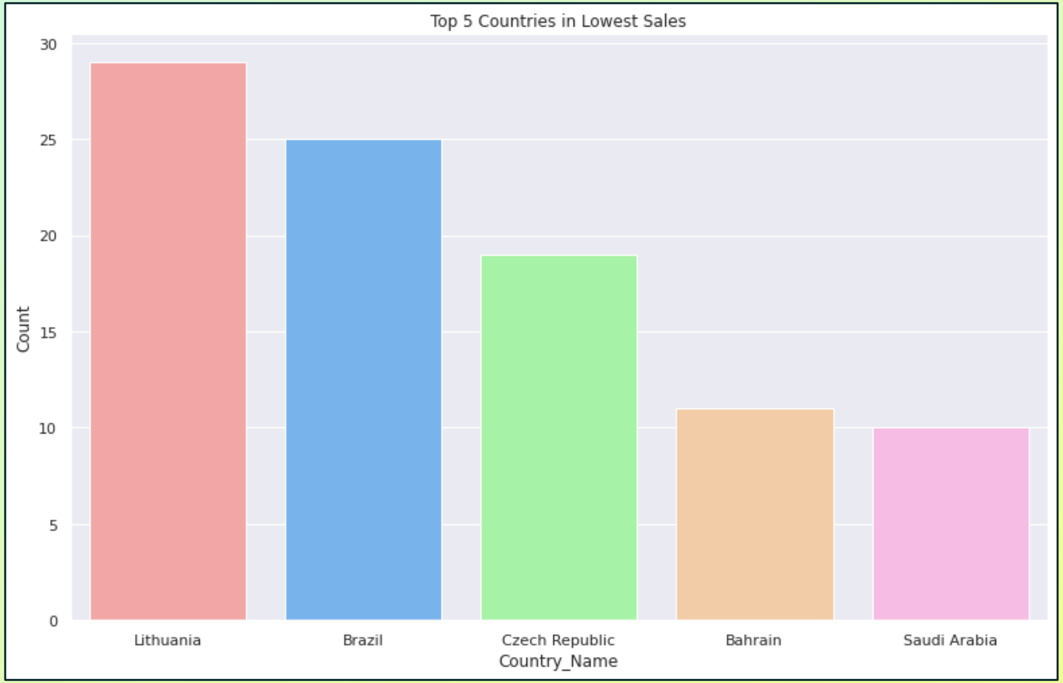
```
+--------+-----------------------------+-
| SL No. | TOP FIVE COUNTRIES IN SALES |
+--------+-----------------------------+-
|   1    |        United Kingdom       |
|   2    |           Germany           |
|   3    |            France           |
|   4    |           Ireland           |
|   5    |            Spain            |
+--------+-----------------------------+-
```



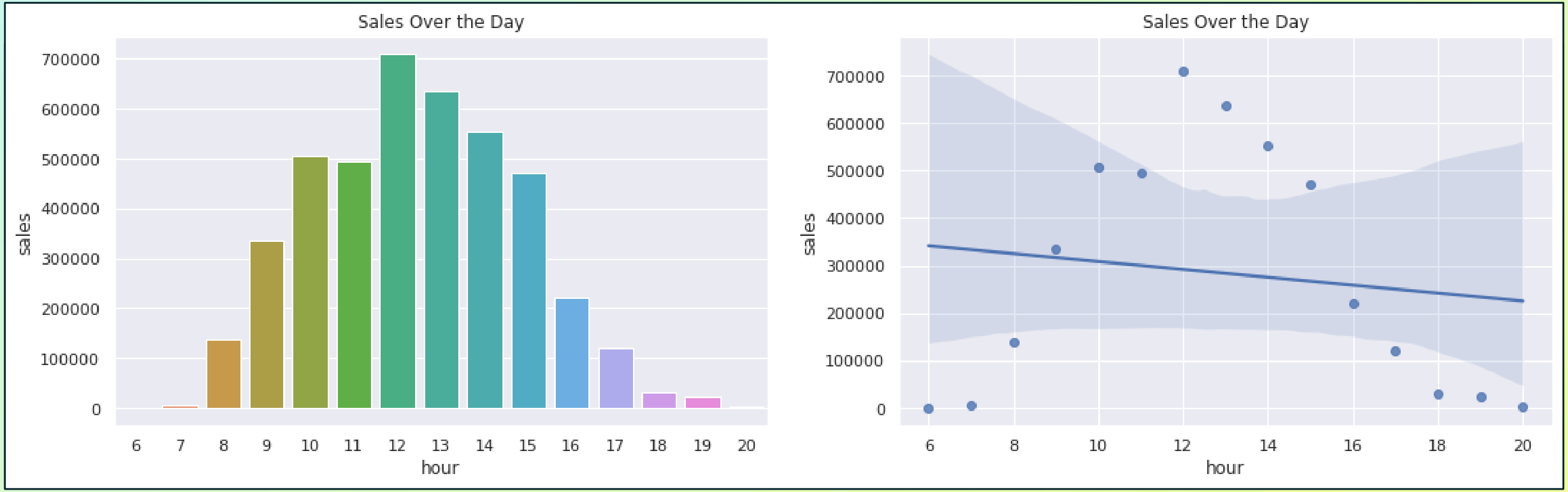
Top Five Countries in Lowest Sales



```
+--------------------------------+
| BOTTOM FIVE COUNTRIES IN SALES |
+--------------------------------+
|          Saudi Arabia          |
|            Bahrain             |
|         Czech Republic         |
|             Brazil             |
|           Lithuania            |
+--------------------------------+
```

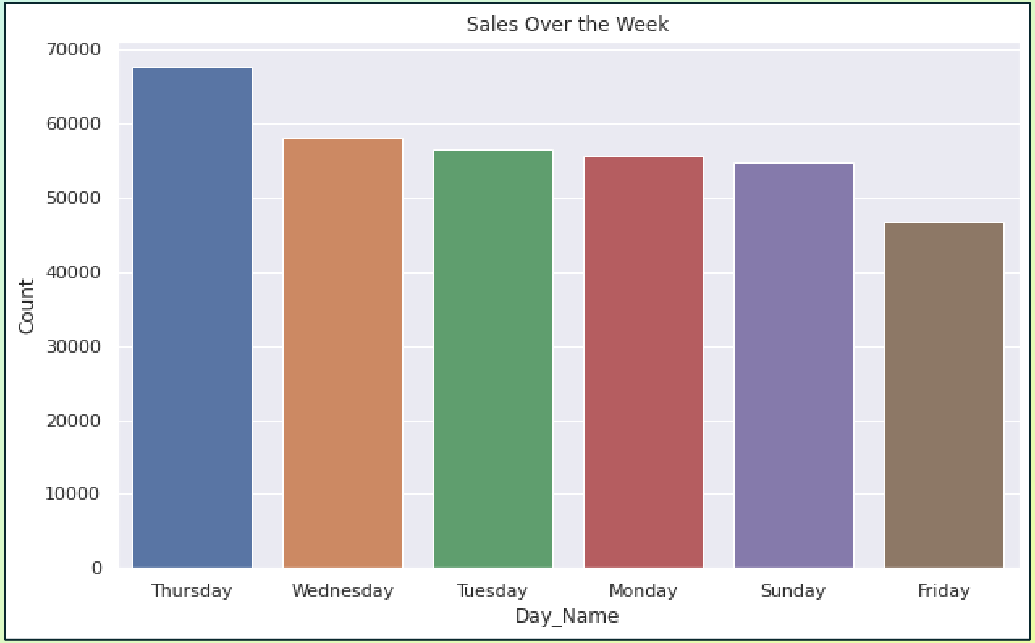


Sales Over the Day



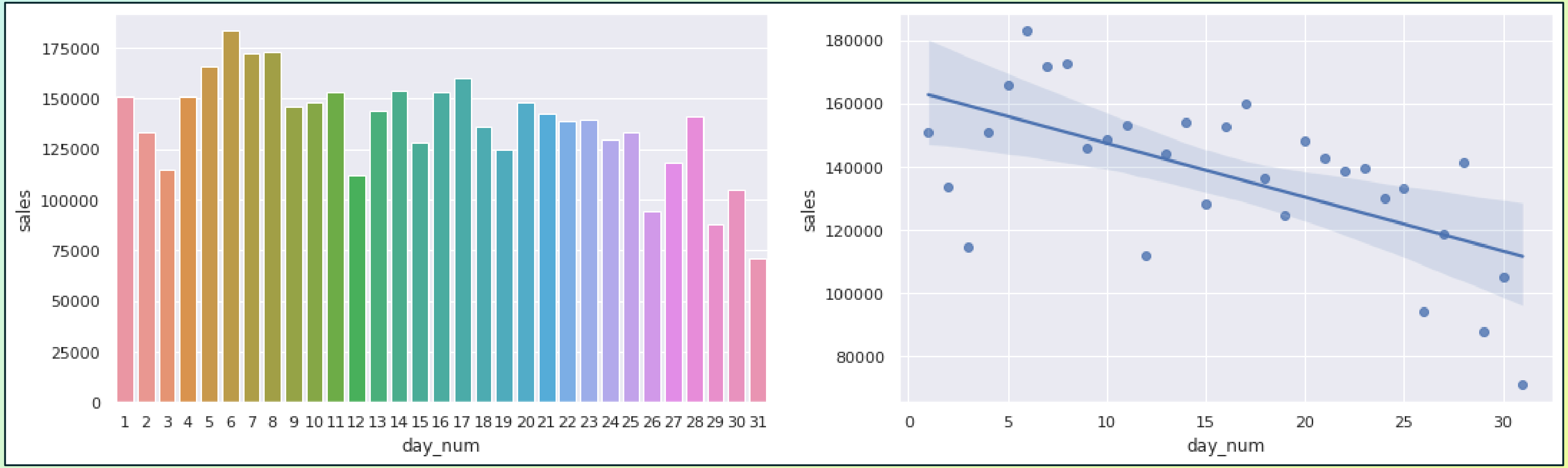
- Most transactions are done between 10 am to 4 pm.

Sales Over the Week



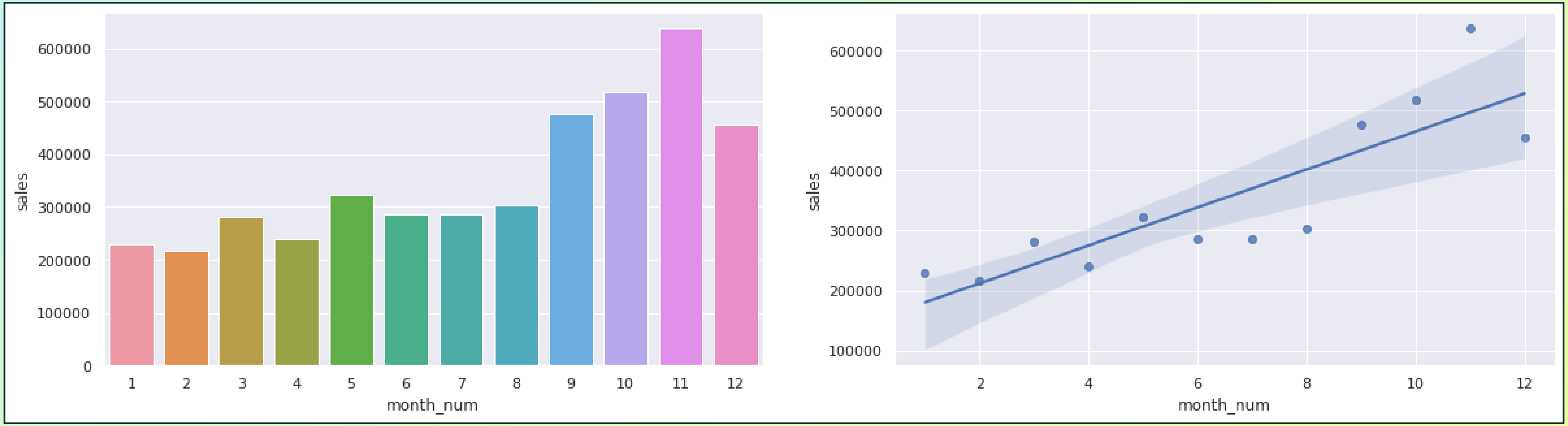
- ***From the graph we can see that sales is high on Thursday***

Sales Over the Month



- ***Sales of store is good in first half of the month and it is dropping on second half of the month.***

Sales Over the Year



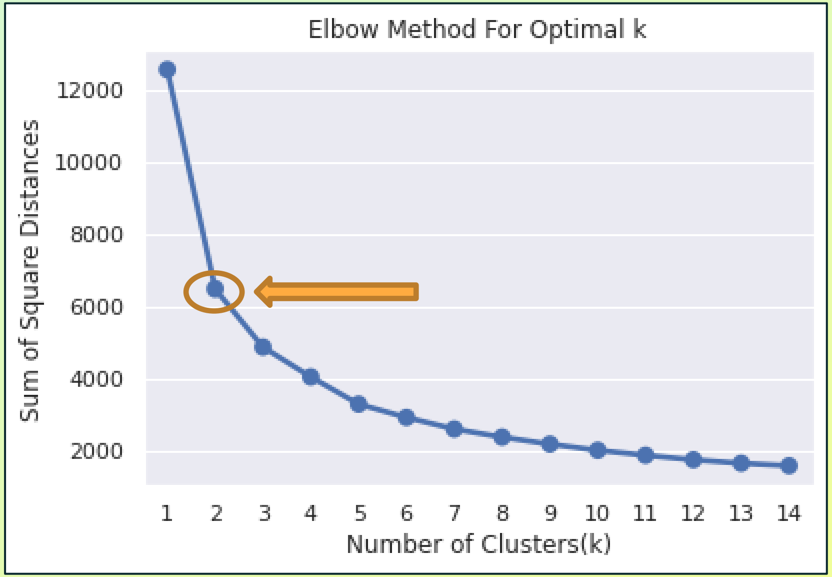
- ***Sales performance is good in second half of each year but it is less in first half of the year.***

Optimum Number of Clusters

Elbow Method

- In order to find the optimum number of clusters for ***KMeans Clustering***, we have used Elbow Method. We plotted the elbow graph for 15 clusters.

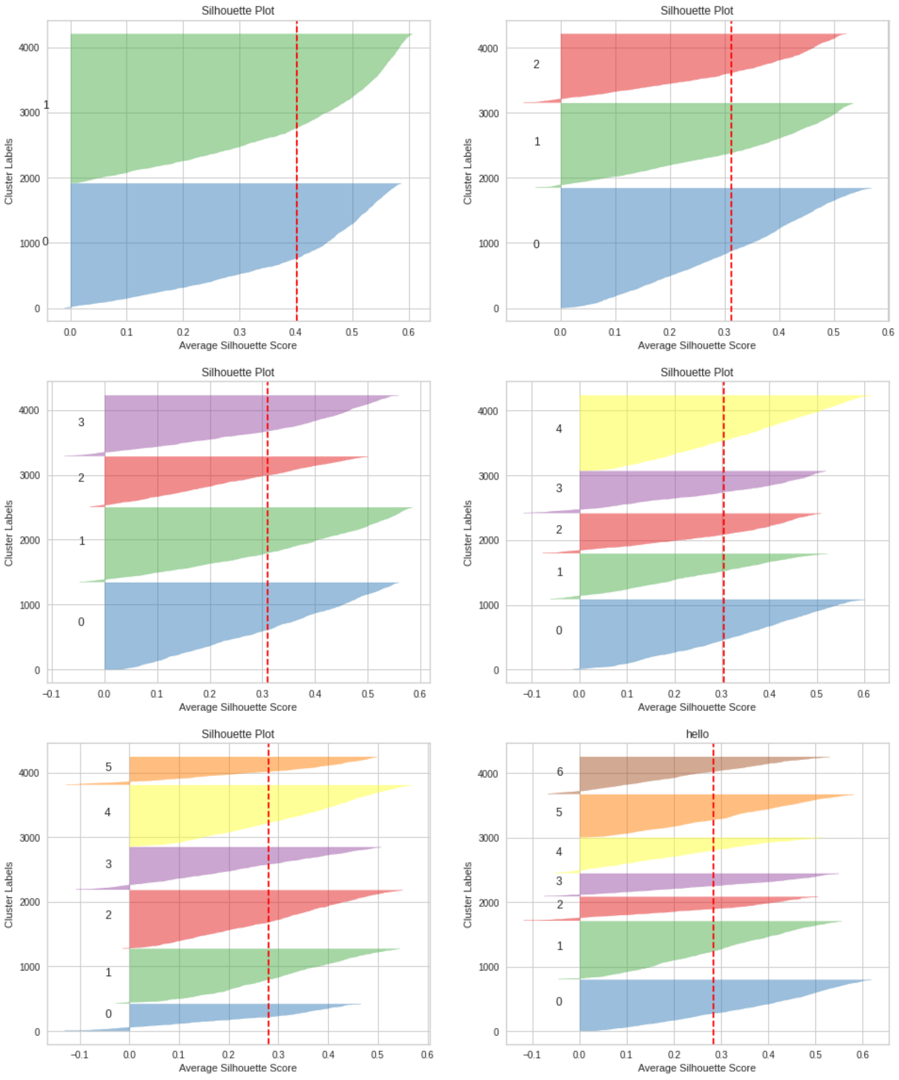
- The point before which the distortion or inertia is decreasing in a Linear fashion is nothing but the optimal number of cluster. Hence ***optimal number of clusters is 2.***



Silhouette Analysis

- In this method, we plotted the Silhouette Plot for range of clusters between 2 to 7. The plot represents the average silhouette score against the number of clusters. From this plot we can easily get the optimal number of clusters for ***KMeans Clustering***.

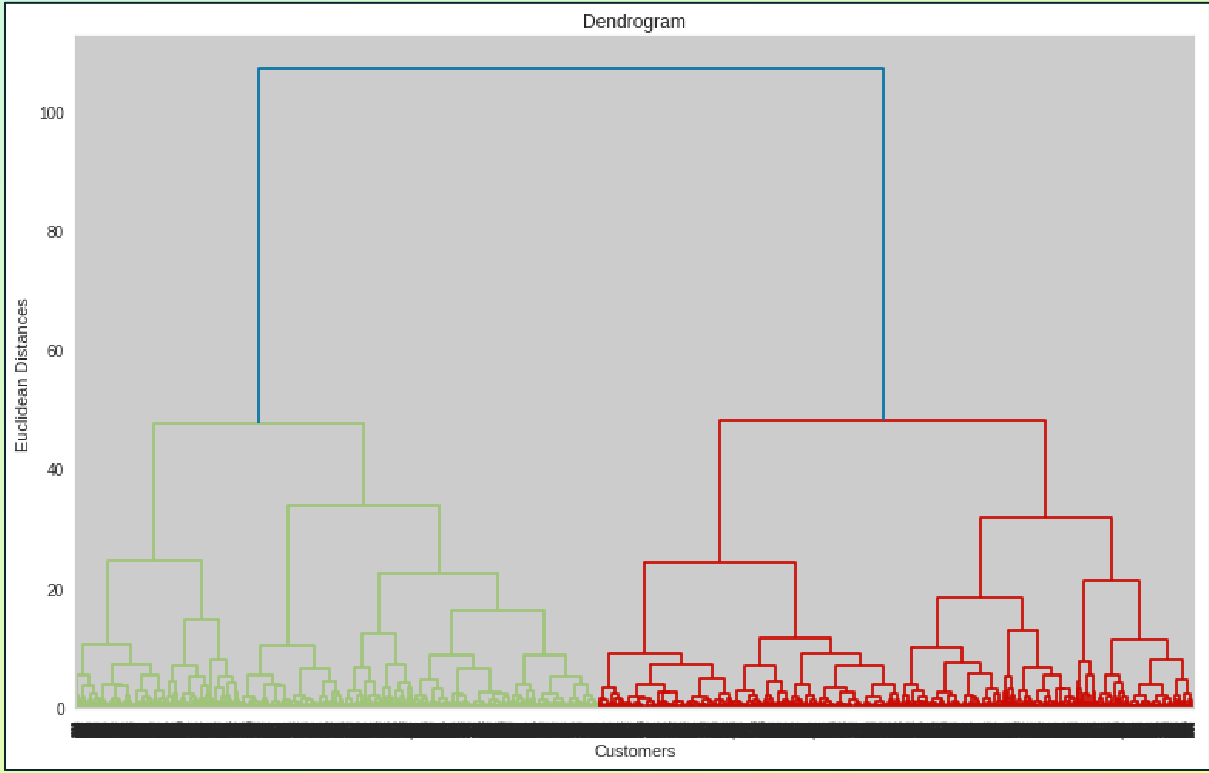
- From these Silhouette Plots we can observe that average silhouette score is highest against the number of clusters 2.***Hence the optimal number of clusters is 2.***



Dendogram

- To find the optimal number of clusters for ***Hierarchical Clustering***, we will plot Dendogram.

- The number of vertical lines which are being intersected by the line drawn using the threshold=90 represent the optimal number of clusters. ***Hence optimal number of Clusters = 2.***

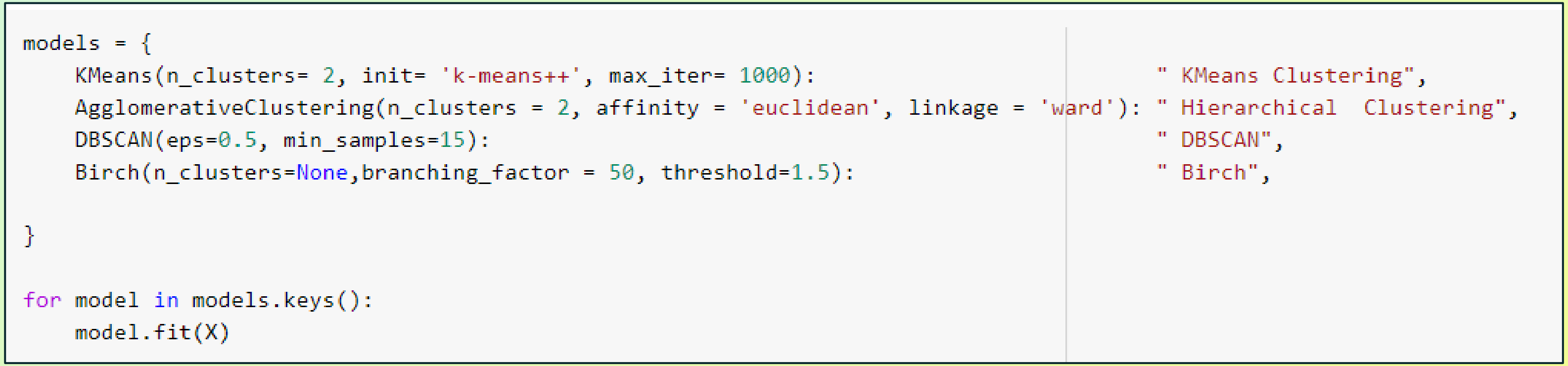


Model Implementation

Model Implementation

We have implemented below mentioned clustering models to our data set.

- KMeans Clustering
- Hierarchical Clustering
- DBSCAN
- Birch



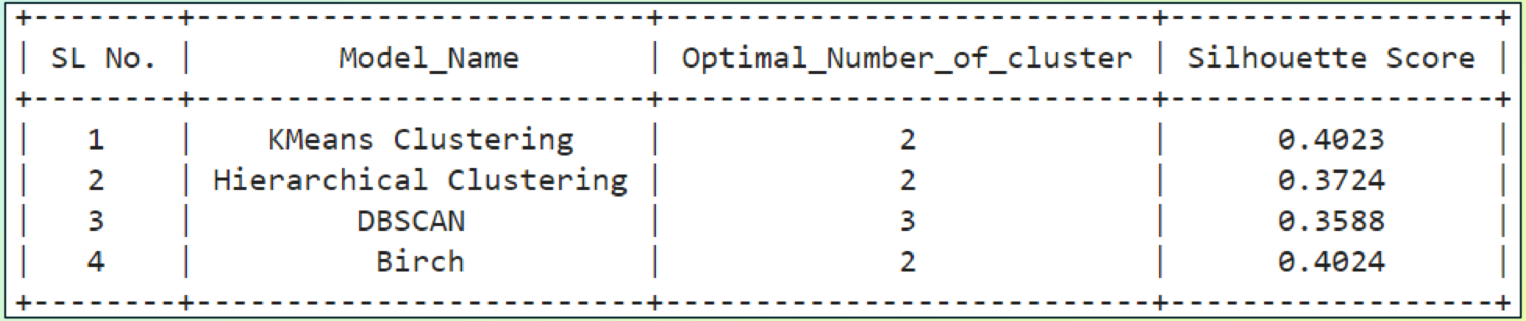
```
models = {
    KMeans(n_clusters= 2, init= 'k-means++', max_iter= 1000):                          " KMeans Clustering",
    AgglomerativeClustering(n_clusters = 2, affinity = 'euclidean', linkage = 'ward'): " Hierarchical  Clustering",
    DBSCAN(eps=0.5, min_samples=15):                                                   " DBSCAN",
    Birch(n_clusters=None,branching_factor = 50, threshold=1.5):                       " Birch",

}

for model in models.keys():
    model.fit(X)
```

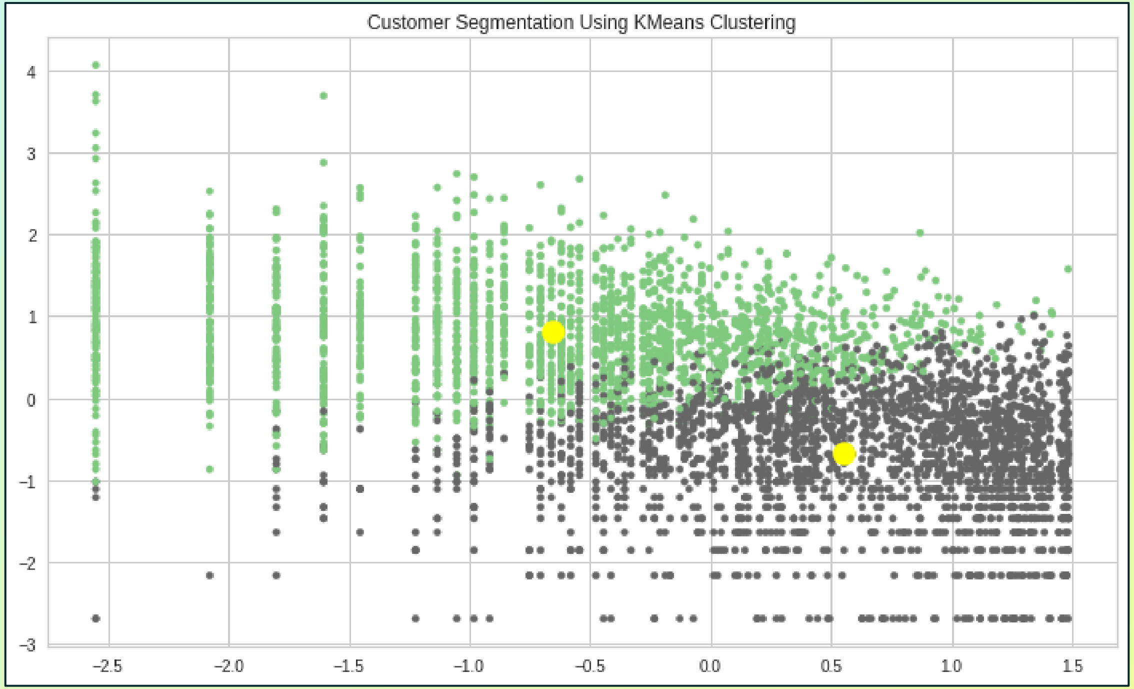
Model Evaluation

Model Accuracy



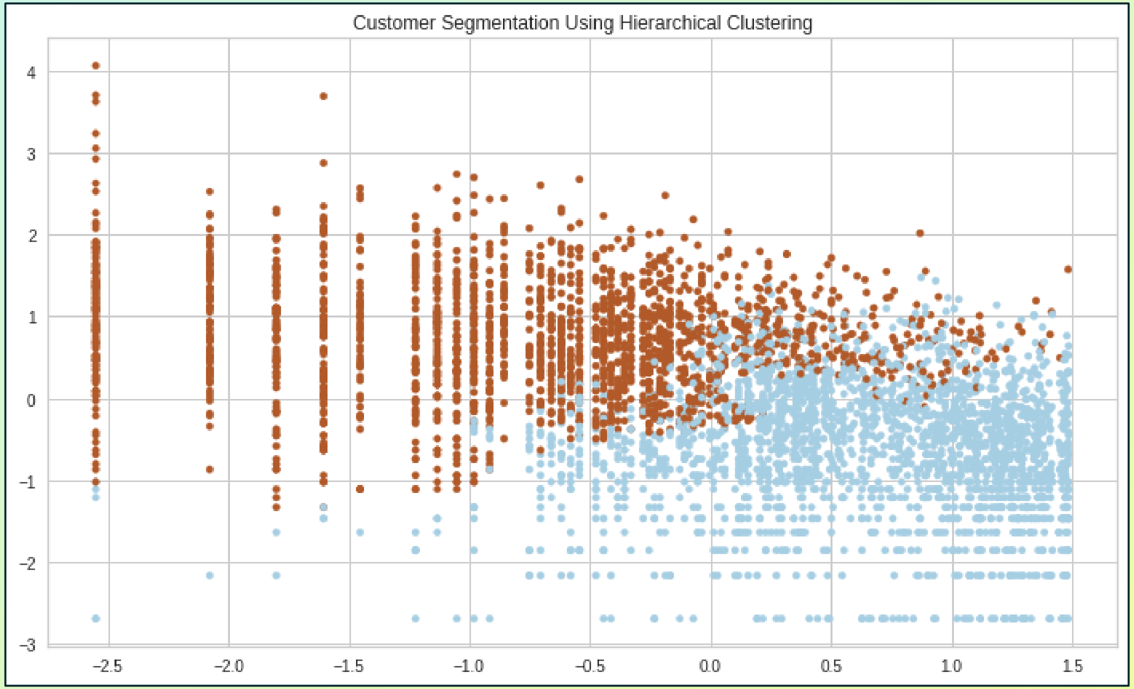
```
+--------+-------------------------+---------------------------+------------------+
| SL No. |        Model_Name       | Optimal_Number_of_cluster | Silhouette Score |
+--------+-------------------------+---------------------------+------------------+
|   1    |    KMeans Clustering    |             2             |      0.4023      |
|   2    | Hierarchical Clustering |             2             |      0.3724      |
|   3    |          DBSCAN         |             3             |      0.3588      |
|   4    |          Birch          |             2             |      0.4024      |
+--------+-------------------------+---------------------------+------------------+
```

KMeans Clustering



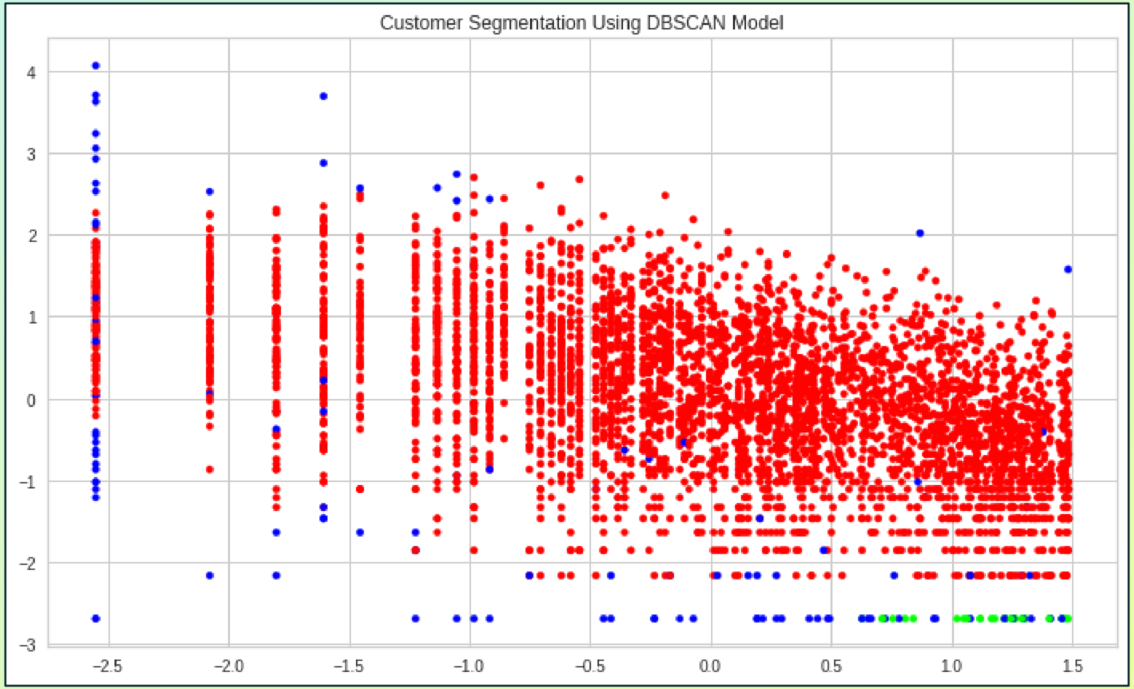
Number of clusters : 2 Silhouette Score : 0.4023

Hierarchical Clustering



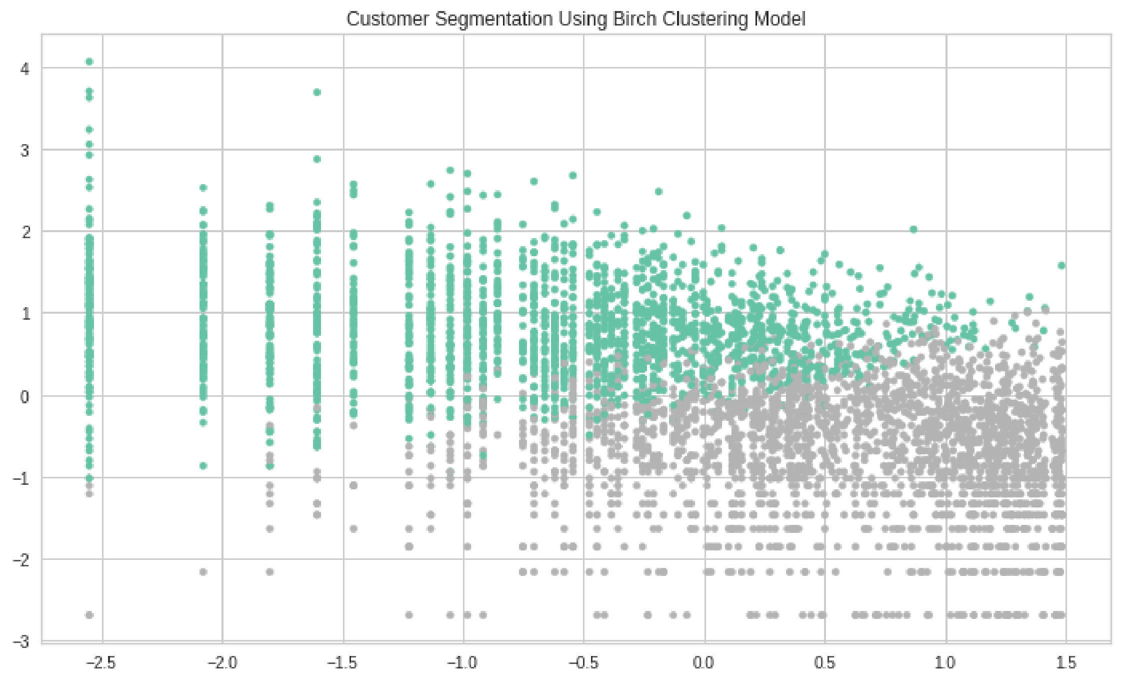
Number of clusters : 2 Silhouette Score : 0.3724

DBSCAN Clustering



Number of clusters : 3   Silhouette Score : 0.3588

DBSCAN Clustering



Number of clusters : 2　　　Silhouette Score : 0.4024

Model Selection

Model Selection

- We have calculated Silhouette score for all the clustering algorithms with optimum number of clusters.
- The Silhouette score for hierarchical clustering is 0.3724 and for DBSCAN is 0.3588.Both the algorithms performed well but score is comparatively less.
- KMeans and Birch Clustering have almost same Silhouette score i.e. 0.4023 and 0.4024 .
- KMeans clustering do not perform clustering very efficiently and it is difficult to process the large datasets with limited amount of resources. Hence for this dataset we selected Birch model for Clustering.

Conclusion

Conclusion

- In order to do the customer segmentation, we created RFM model and calculated RFM score. Higher RFM Score represents the most valuable customers of store.
- The optimal number of clusters for KMeans Clustering using Elbow Method is 2.
- We performed Silhouette Analysis and got the optimal number of clusters as 2.
- The optimal number of clusters for Hierarchical Clustering using Dendograph is 2.
- The Silhouette Score is highest for birch and KMeans Clustering (0.4024 & 0.4023).
- For DBSCAN clustering, the Silhouette Score is 0.3588 with optimum number of clusters 3.The hierarchical clustering model performed well but Silhoutte score was comparatively less 0.3724.
- KMeans clustering do not perform clustering very efficiently and it is difficult to process the large datasets with limited amount of resources. Hence for this dataset we selected Birch model for Clustering.

Suggestions to Improve Sales

- ❑ Customer should be segmented based on their recency, frequency, and monetary. Customer with high RFM will be more valuable customer for store.
- ❑ Store should offer credit limit to most valuable customers to maintain the connectivity.
- ❑ To attract new customers, we can offer a discount on first three orders to new customers.

Challenges

Challenges

- The data set was huge, so computational time involved was high.
- The Silhouette analysis was bit lengthy and time consuming process. As the number of clusters k increases, computational time also increases.
- Due to huge data set, time required for figures plotting was high.

References

References

- □ Kaggle
- □ Youtube
- □ Github
- □ Towards data science
- □ Analytics Vidya
- □ Code basics
- □ Stack over flow

Thank You