

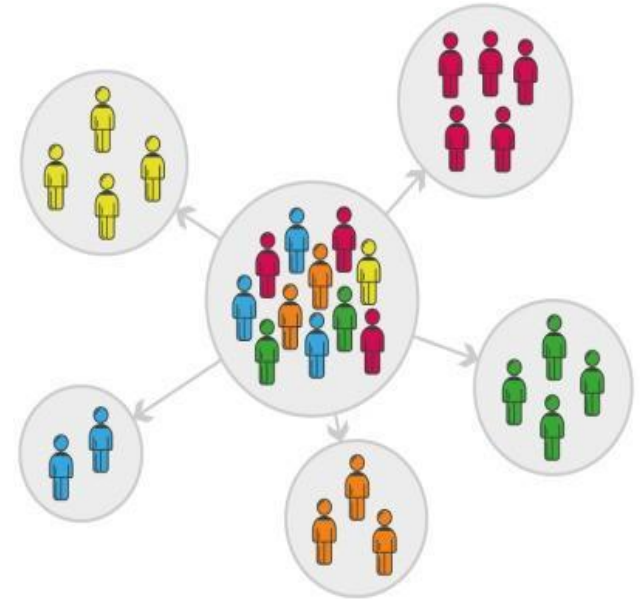
Capstone Project

Online Retail Customer Segmentation



Customer Segmentation Analysis follows:

- ❑ Problem Statements
- ❑ Data Information
- ❑ Data Analysis
- ❑ Data cleaning
- ❑ Data Preparation
- ❑ Model Training
- ❑ Summary
- ❑ Challenges
- ❑ Conclusion



Problem Statements:

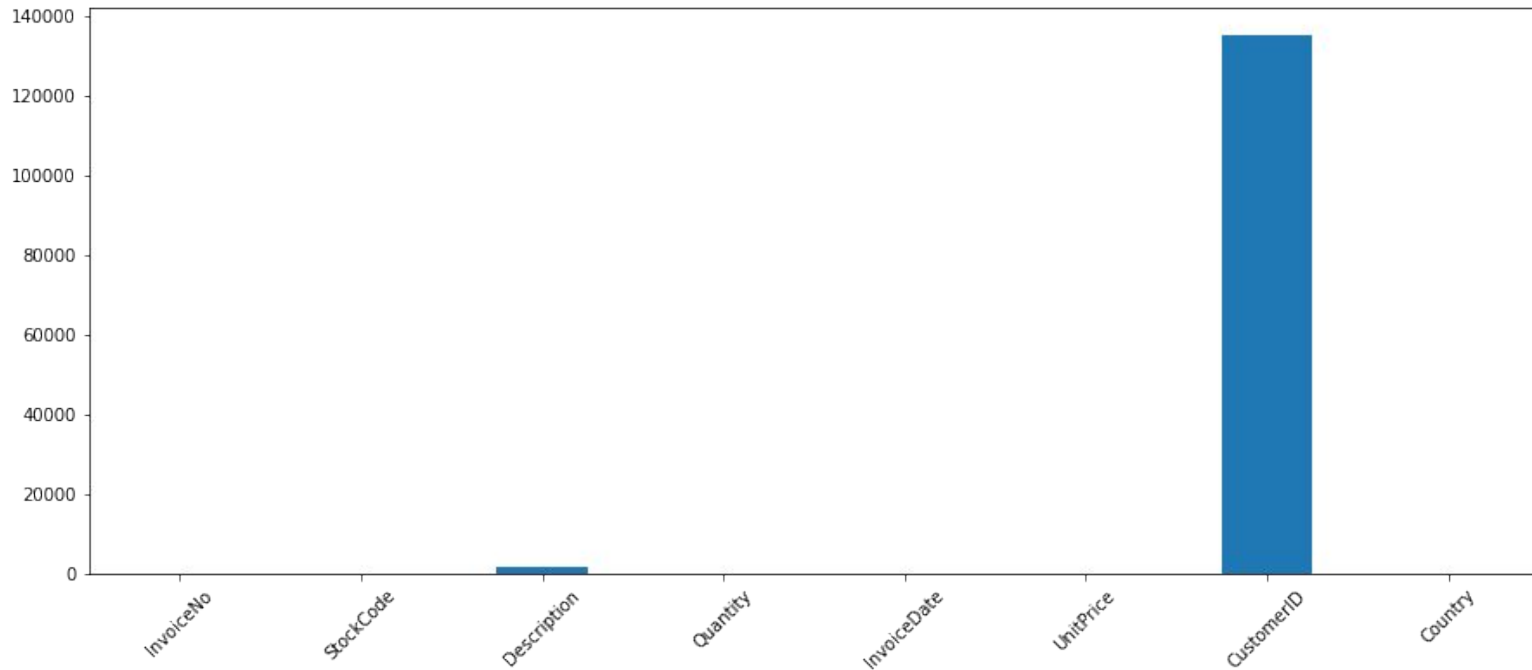
- ☐ How Null values in the data handled?
- ☐ Does data contains negative values?
- ☐ Does data contains outliers?
- ☐ Any country dominated in terms of data?
- ☐ Number of clusters based on silhouette score?
- ☐ Better visualization of clusters for different algorithms?

Data Summary

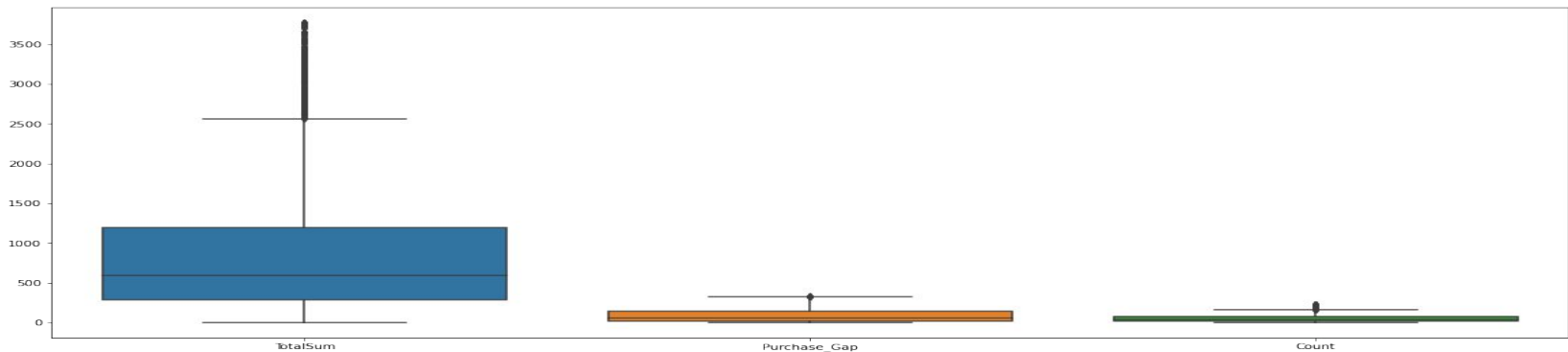
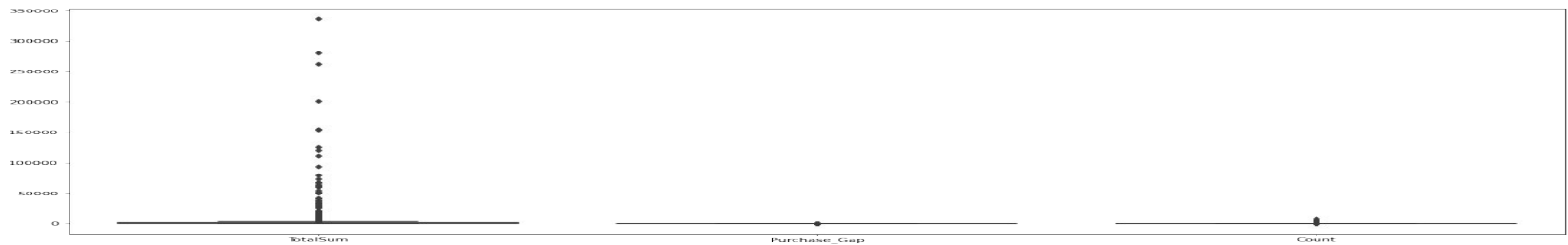
- ❑ **File type:** Excel
- ❑ **File size:** 23MB
- ❑ **Dataset shape:**
 - ❖ Rows:541909
 - ❖ Columns:8
- ❑ **Important Columns:**
 - ❖ CustomerID
 - ❖ Quantity
 - ❖ UnitPrice
 - ❖ InvoiceDate

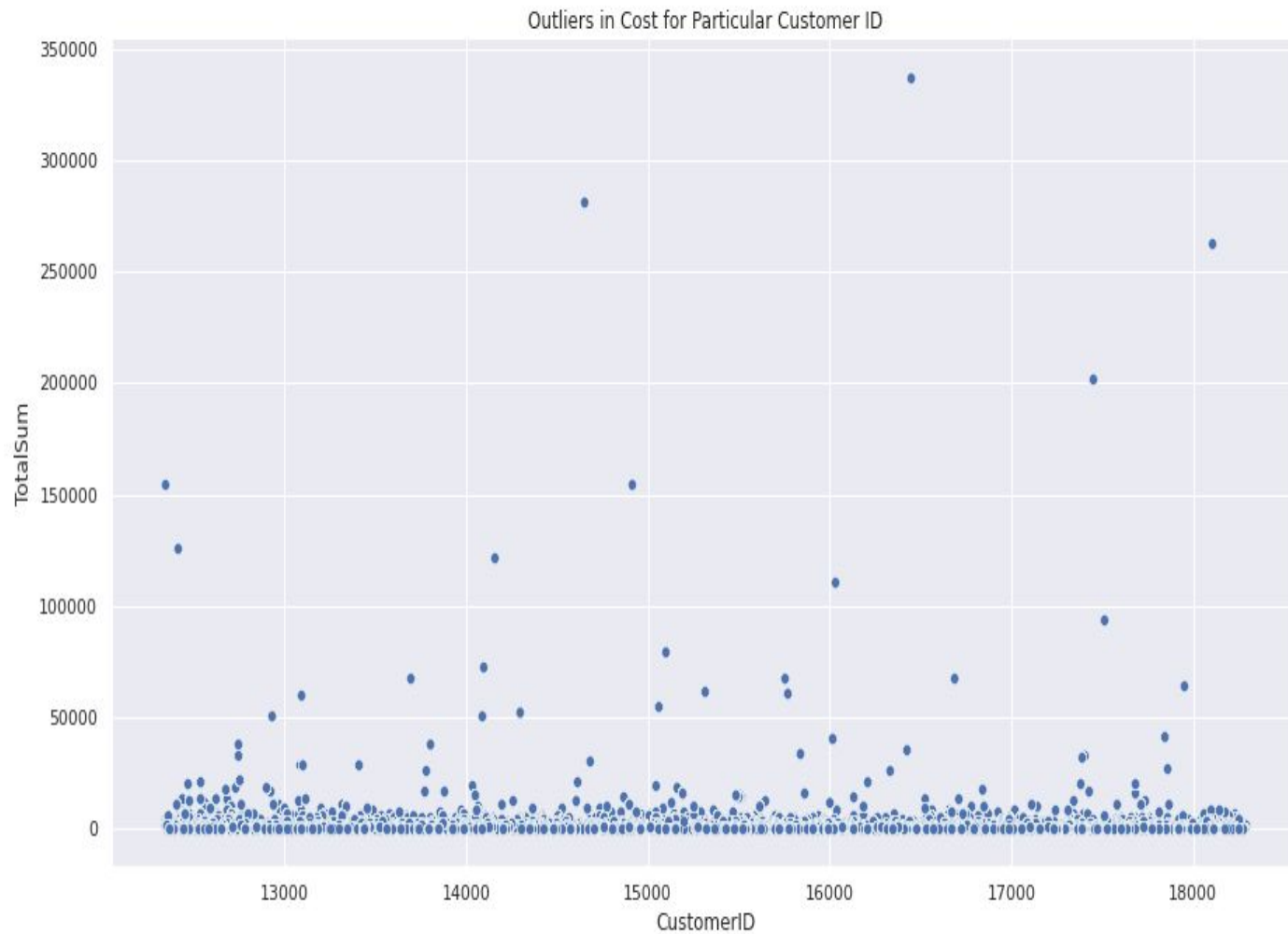


Does data contains Null values?

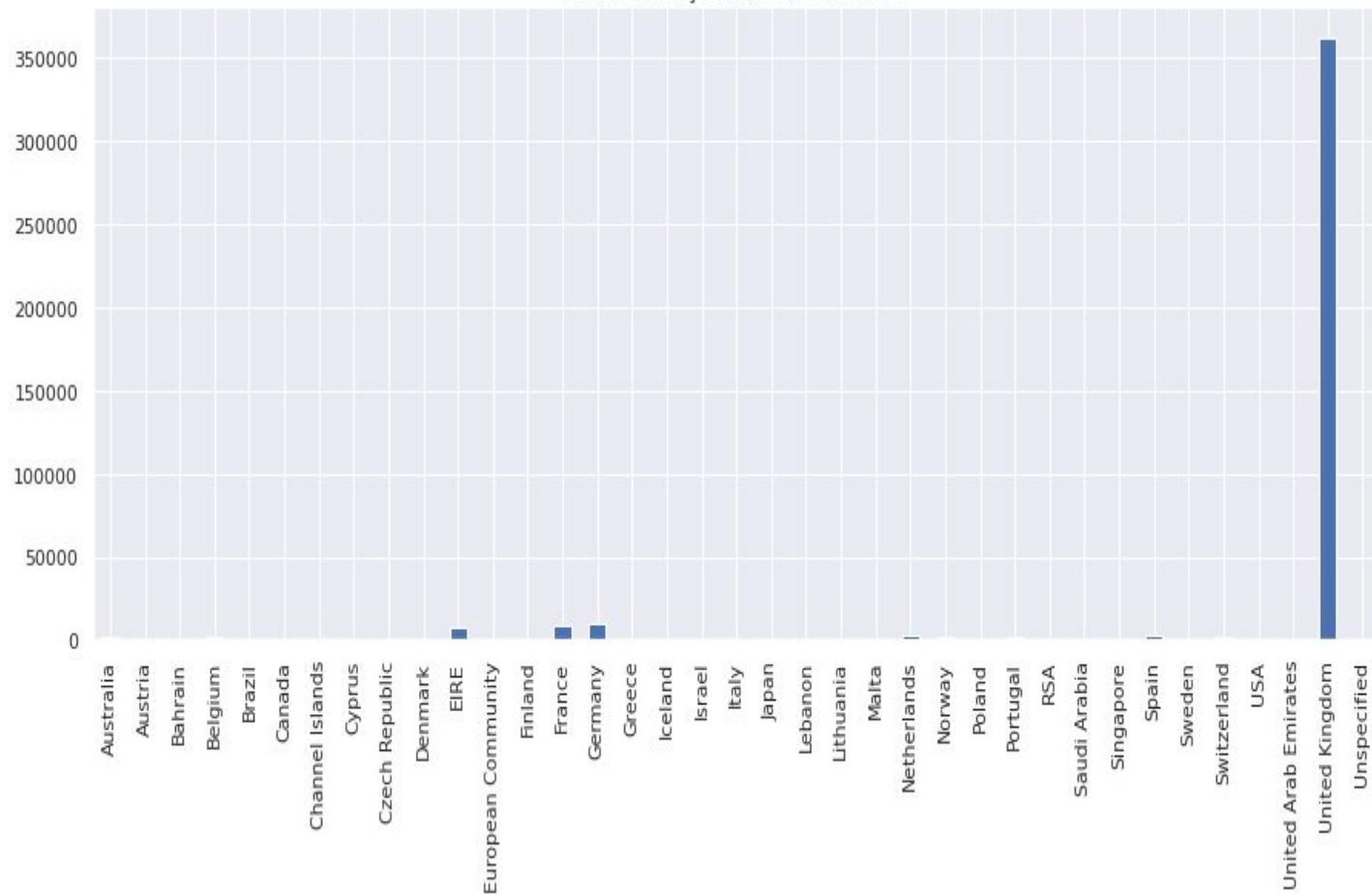


Does data contains outliers?

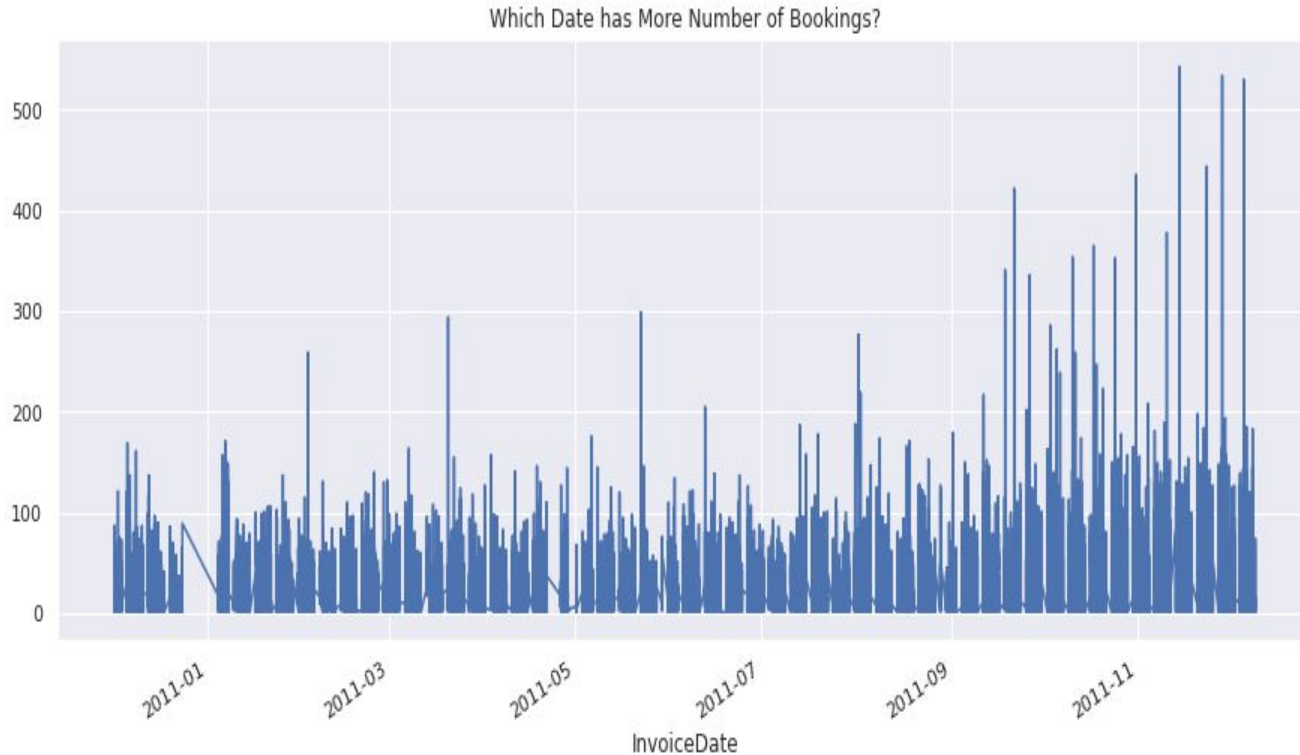




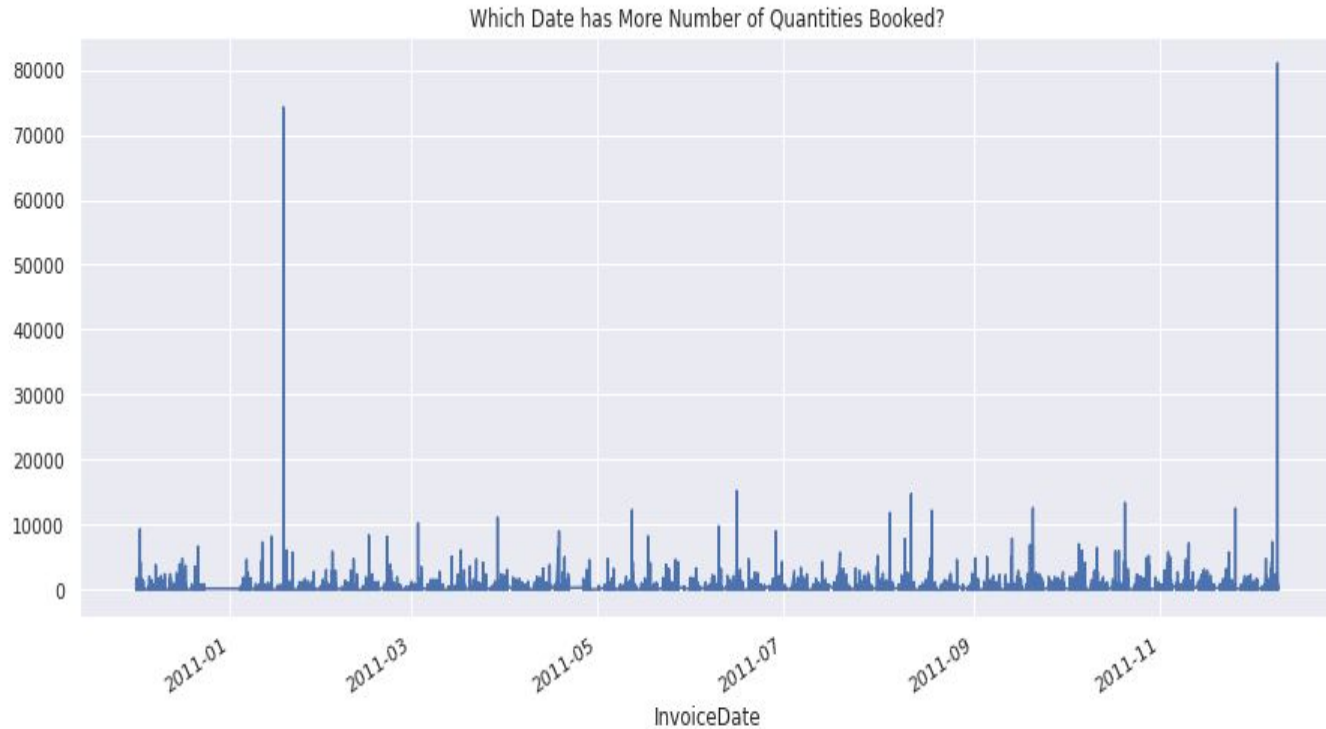
Which Country is dominated in data?




```
eda_df.groupby('InvoiceDate').count()['InvoiceNo'].plot.line(figsize=(15,6))  
plt.title('Which Date has More Number of Bookings?')  
plt.show()
```



```
eda_df.groupby('InvoiceDate')['Quantity'].sum().plot.line(figsize=(15,6))  
plt.title('Which Date has More Number of Quantities Booked?')  
plt.show()
```



<Figure size 1440x360 with 0 Axes>



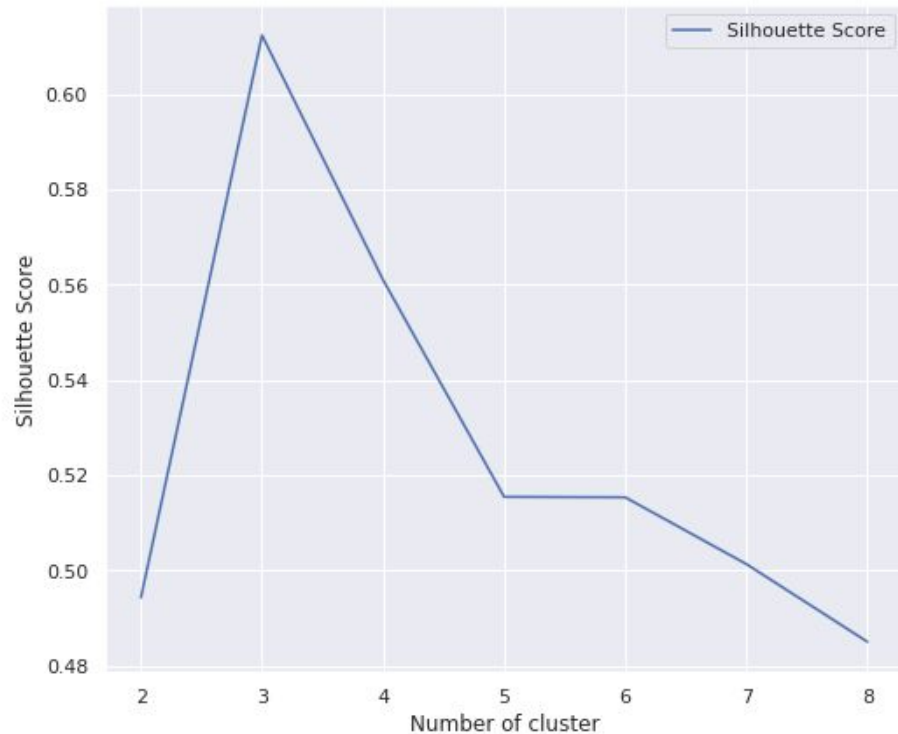
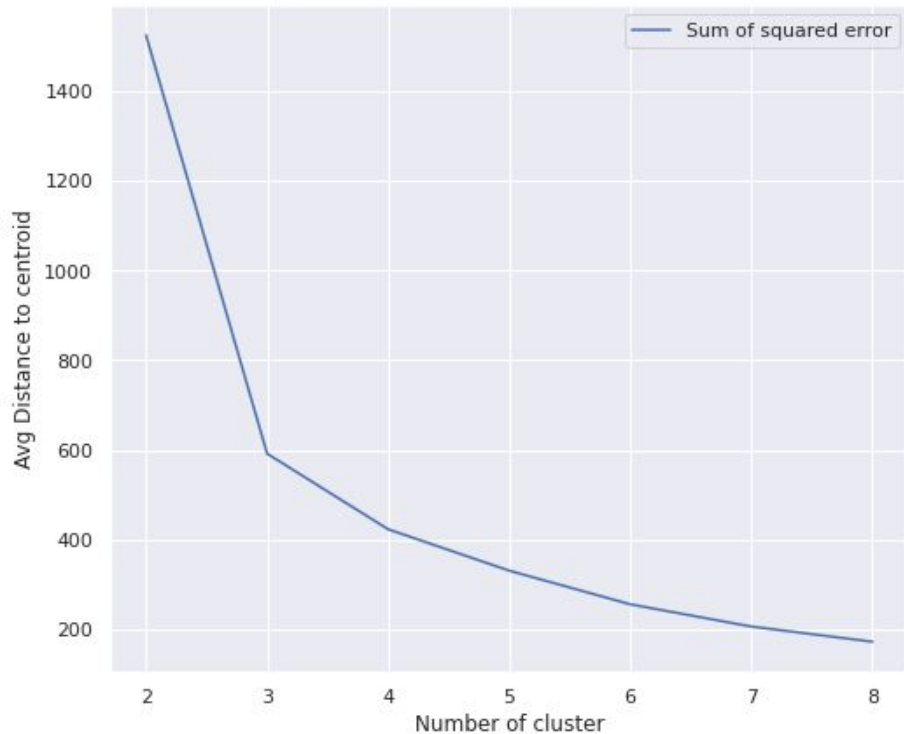
How to find optimal cluster size?

Elbow method: it runs on k-means clustering, it measures the sum of square distances from each point to its assigned center.

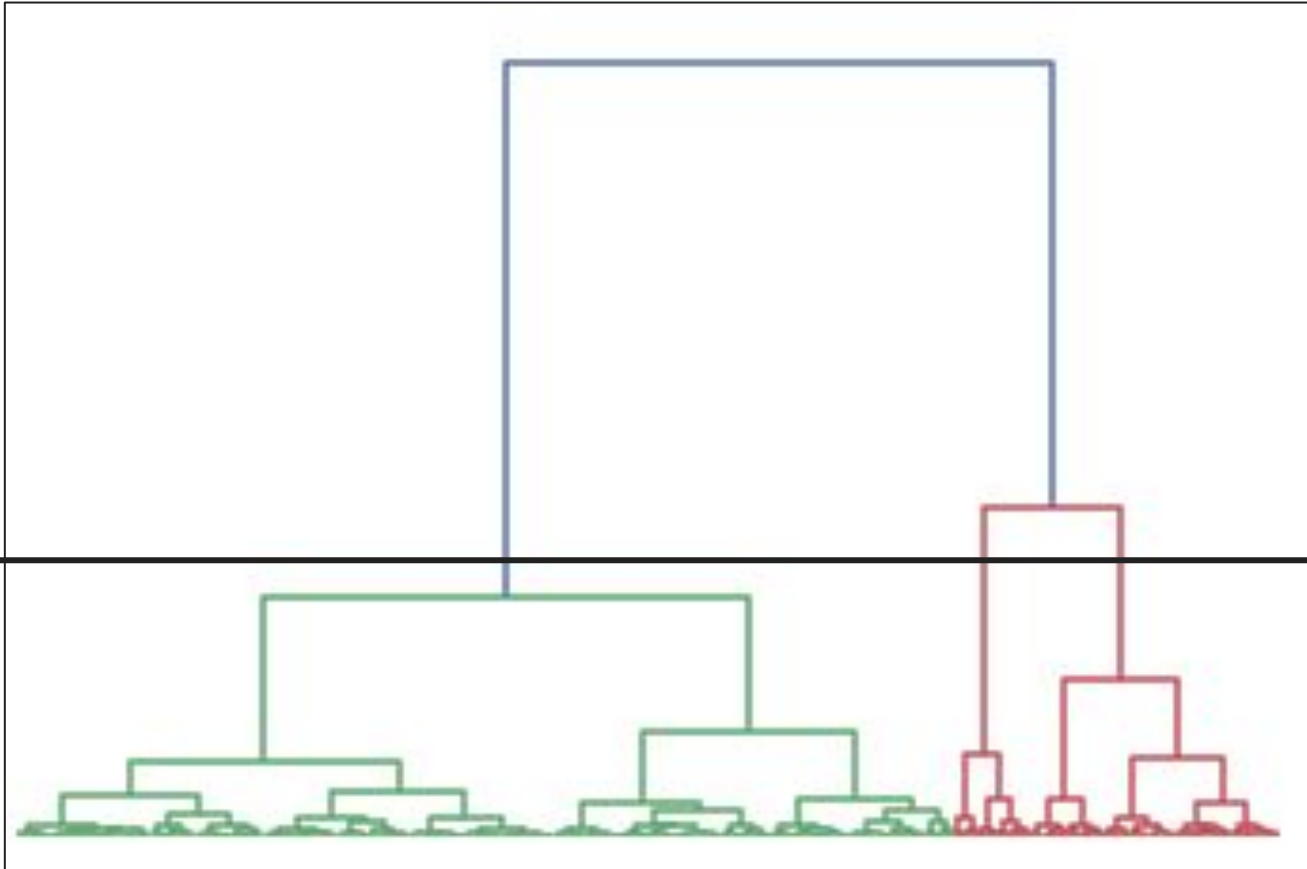
Silhouette score: The silhouette ranges from -1 to $+1$, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

```
for n_clusters=2, The Silhoutte Coefficient is=0.49428728628278296
for n_clusters=3, The Silhoutte Coefficient is=0.6123894802353901
for n_clusters=4, The Silhoutte Coefficient is=0.561019099786192
for n_clusters=5, The Silhoutte Coefficient is=0.5154420877925352
for n_clusters=6, The Silhoutte Coefficient is=0.5153482718550463
for n_clusters=7, The Silhoutte Coefficient is=0.5013357480604806
for n_clusters=8, The Silhoutte Coefficient is=0.48501437515080126
```

How does Elbow and silhouette graph look?



Visualization of Dendrogram Linkage



What are the Clustering algorithms used?

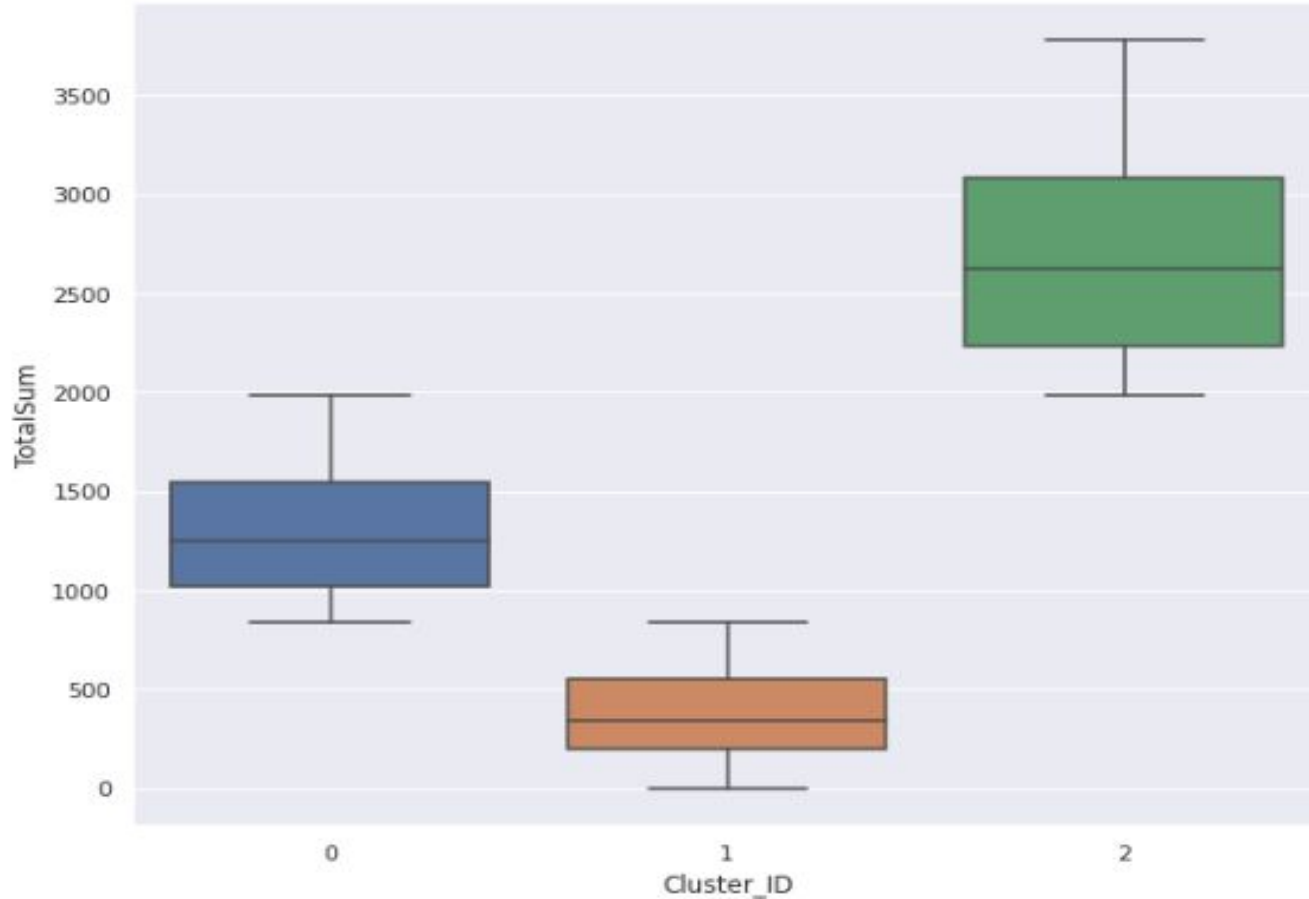
```
#Cluster=3, Parameter Tuning
kmeans=KMeans(n_clusters=3,max_iter=50)
ckmeans=kmeans.fit(X)
y_kmeans=kmeans.predict(X)
centers = kmeans.cluster_centers_
new_df['Cluster_ID']=kmeans.labels_
```

```
from sklearn.cluster import AgglomerativeClustering
ac=AgglomerativeClustering(n_clusters=3 ,affinity='euclidean',linkage='ward')
y_ac=ac.fit_predict(assign)
```

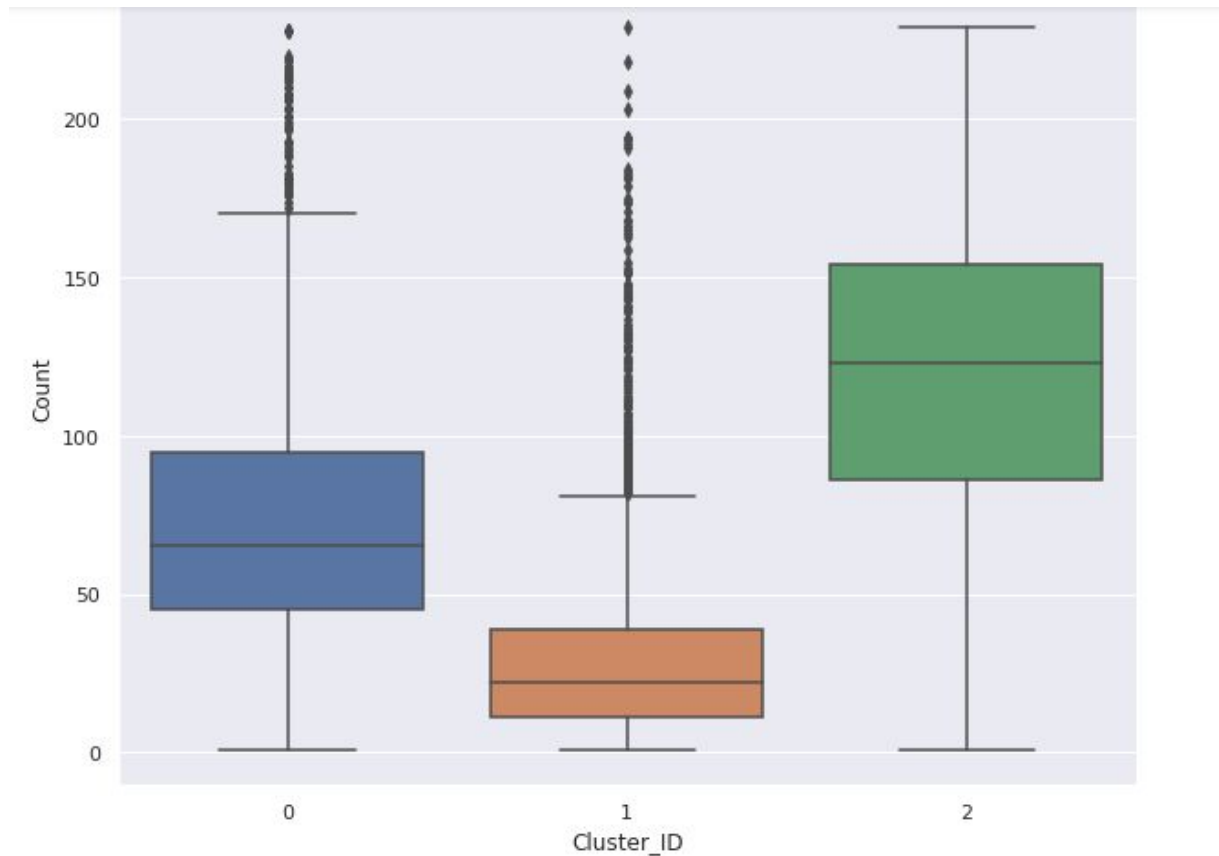

For Cluster size=3, Centers representation



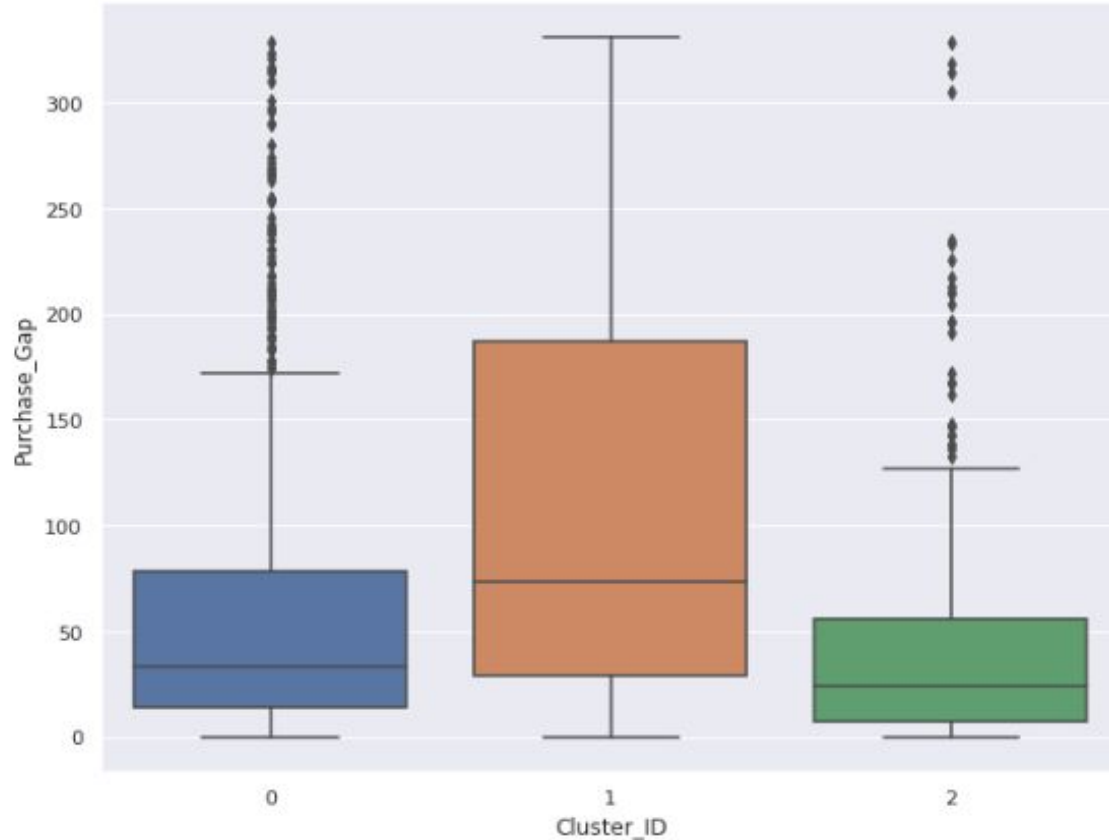
Which Cluster is having high spendings?



Which Cluster ID has made Highest Order Count?



Which Cluster takes more time gap to Purchase?



Summary

- First imported the libraries and dataset which was in excel file and This dataset contains 541909 rows and 8 columns, then checked for duplication of data and null values.
- There were more than 120000 null values present in CustomerID Column it main column as other column was filled with zero and drop all values.
- We Converted data into High Spending, Late Purchases ID, Number of Order till date.
- Various plots are visualized to see Outliers and Applied Interquartile Range method.
- Data was used different units so its scaled using Standard Scaler and normalise data.
- Applied Principle Component Analysis and reduced to 2-dim
- To find Number Clusters we applied Elow Method and silhouette score the Selected Cluster Size=3 with Visualize Graph.
- K-Means Clustering was applied with cluster size=3.
- Dendrogram Linkage and Hierarchical Agglomerative Clustering Models are applied
- Detailed visualization of each Cluster Center and their Spendings, Delay, Frequency of Number of Orders.

Conclusion

- Given Data for Customer Segmentation most of them are irrelevant like StockCode, Description.etc and there is no relation.
- After Applying Elbow and Silhouette score are more at cluster size =3, score was than 0.6,
- Same results applied with Dendrogram results of Kmeans Clusters Centers in plots appears better than Hierarchical Agglomerative Clustering
- Cluster_0 CustomerID's take more time gap between Each Oder they Placed(Rarely).
- Cluster_1 CustomerID's Makes always a Bulk Purchases which leads high Spendings(Retailers).
- Cluster_2 CustomerID's Has Highest Orders Placed(Small Shopes with less inventory).

Challenges and Future Analysis

- Whole data Consists of duplicated data initially there was more than 5-lakh data after grouped Based Customer ID it was left with 3.5 thousand Customers left.
 - Grouping them based on certain assumptions and there was many negative values.
-
- Expand our Analysis to multiple cities and compare patterns and trends amongst these Customer_ID.
 - Expectation-Maximization algorithm using Gaussian mixture model where not only mean as well as variance of data is considered.

Q&A

<https://slides.app.goo.gl/ojJmj>

THANK YOU

