CIE 417 project task II

Hazem Muhammad Tarek 201800283

Ibrahim Hamada Ibrahim 201800739

Sohaila Islam Zaki 201800998

Abu Dhabi Commercial Bank(ADCB) loans' dataset

Data set

- **Total O/S:** Total loan amount
- **TENOR_@Booking:** Duration in months for the loan
- **Loan Term:** Duration in years
- **Booking Date:** The date on which the loan data is entered
- **Maturity_Date:** the date on which a borrower's final loan payment is due.
- **DPD:** Days past dues, a metric that indicates whether a customer has been consistent in his/her repayments and if he/she has missed any, how many installments did the customer miss, and by how many days.
- **DOB:** Date of Birth of the client
- **Age:** The age of the client.
- **Gender:** The gender of the client.
- **Customer Segment:** classification of the customer based on the type of employment.

Data pre-processing

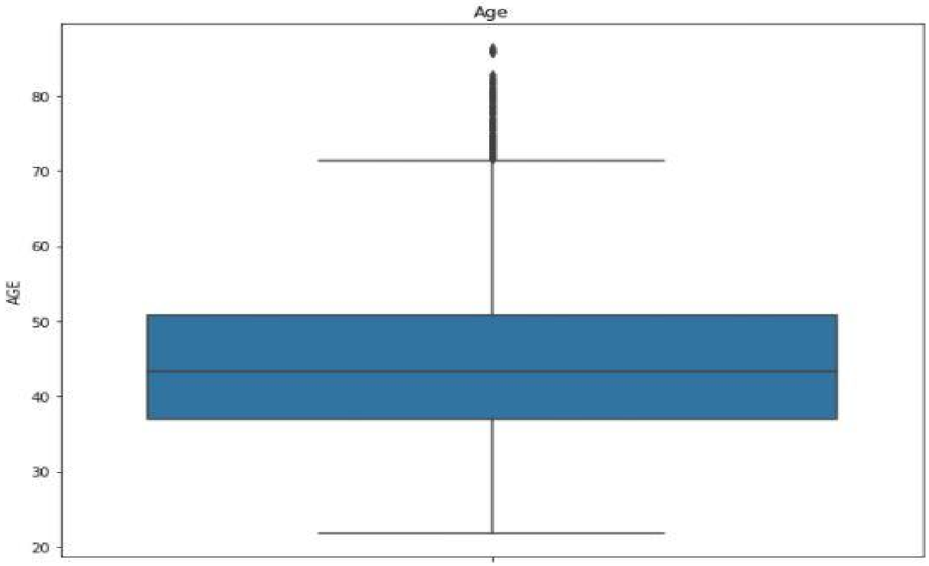
- 279 nulls were found and dropped.
- 798 duplicates were found and dropped.

- Loan Term (Duration in years) was dropped because it doesn't add new information since TENOR_@Booking is the Duration in months for the loan.
- Drop DOB Feature since we have Age feature, so DOB is not needed anymore.
- Age at maturity column was dropped and recalculated(age+loan term in years) as a number instead of date(DD/MM/YYYY).
- Naming issues in the gender column were fixed, possible values were [MALE, Male, FEMALE, Female] and were changed to [Male, Female].
- Customer segment and gender were one-hot encoded.

Data Visualization

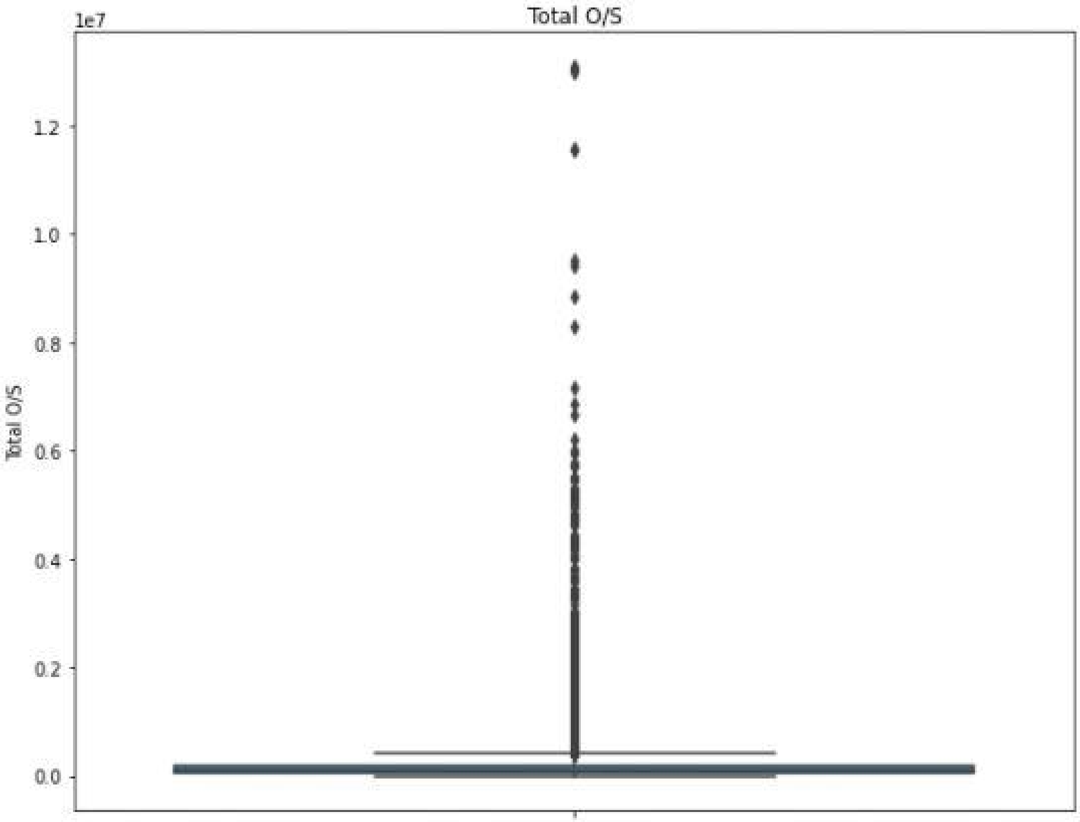
Outliers in age:

- There are some negative ages which need to be dropped.
- The legal age for loans is 21 years.
- So, all ages less than 21 years will be dropped.

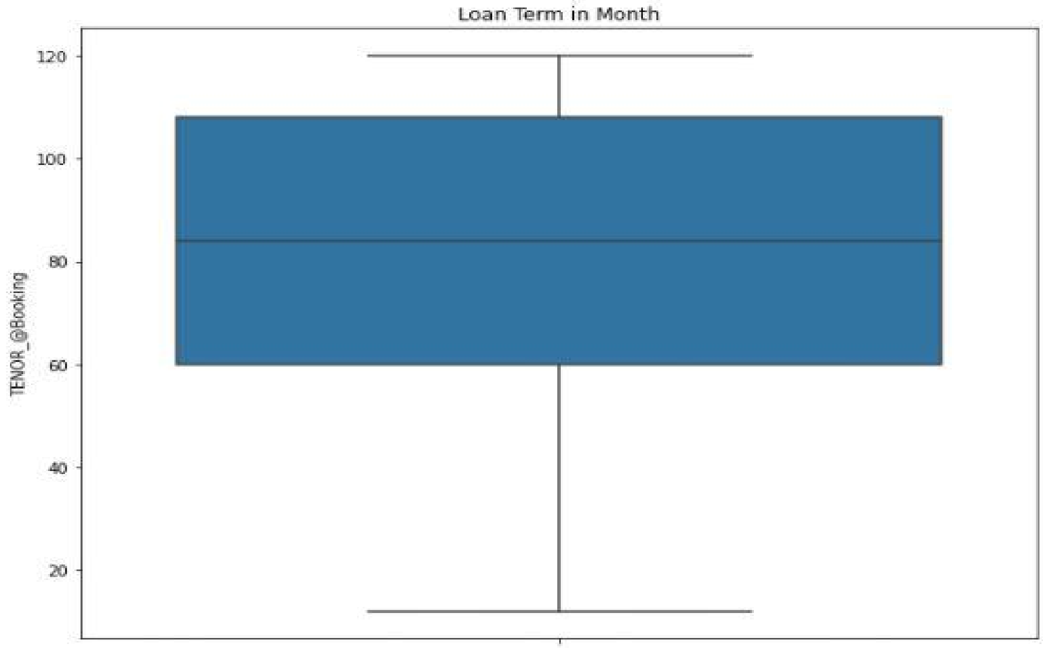


Outliers in total loan amount:

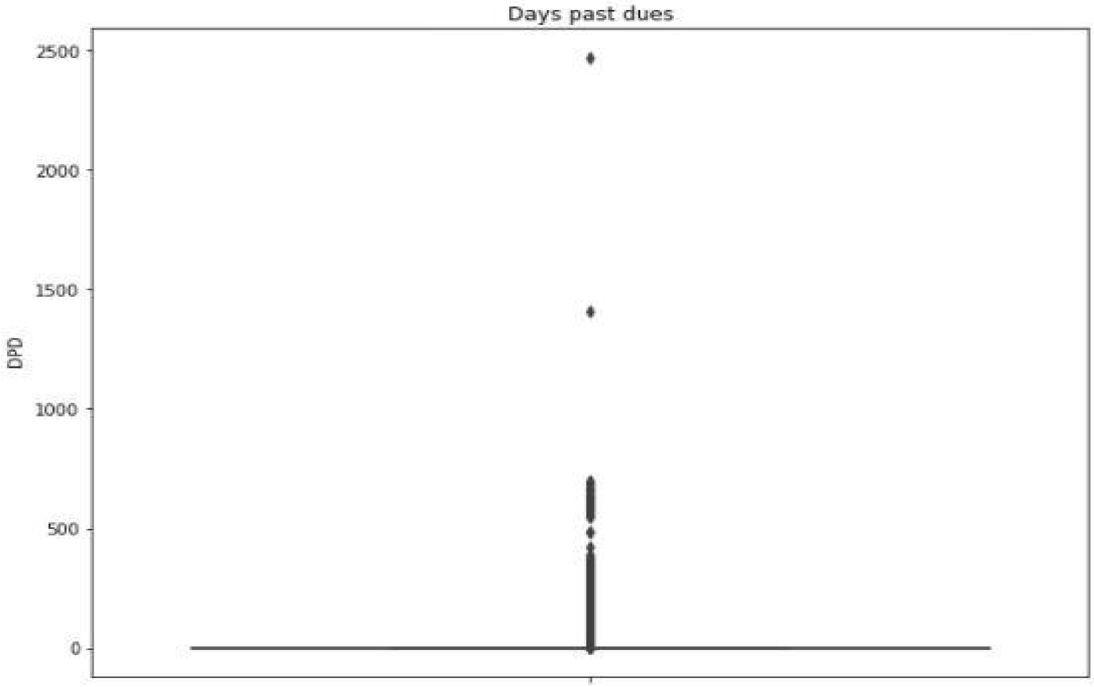
- Abu Dhabi Commercial Bank (ADCB) Website says that the minimum value for a loan is 10K EGP.
- So, all amounts less than 10K EGP will be dropped.
- Also, it can be noted that there is an outlier with Loan Amount equals 25 Million EGP, so Loans greater than 15 Million will be dropped.



Boxplot of loan term in month.

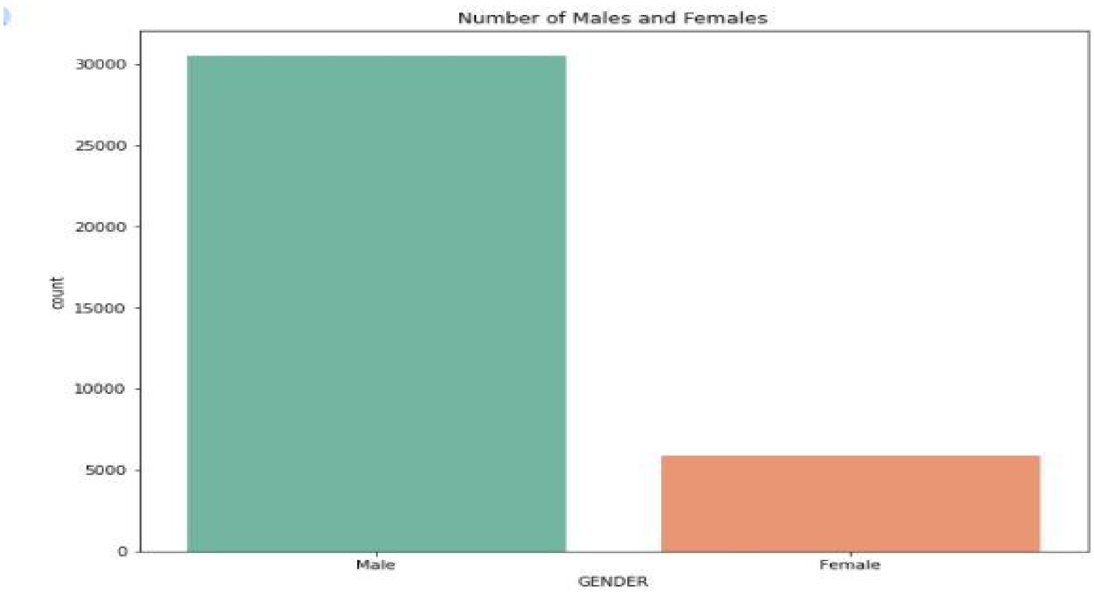


Boxplot of past dues.



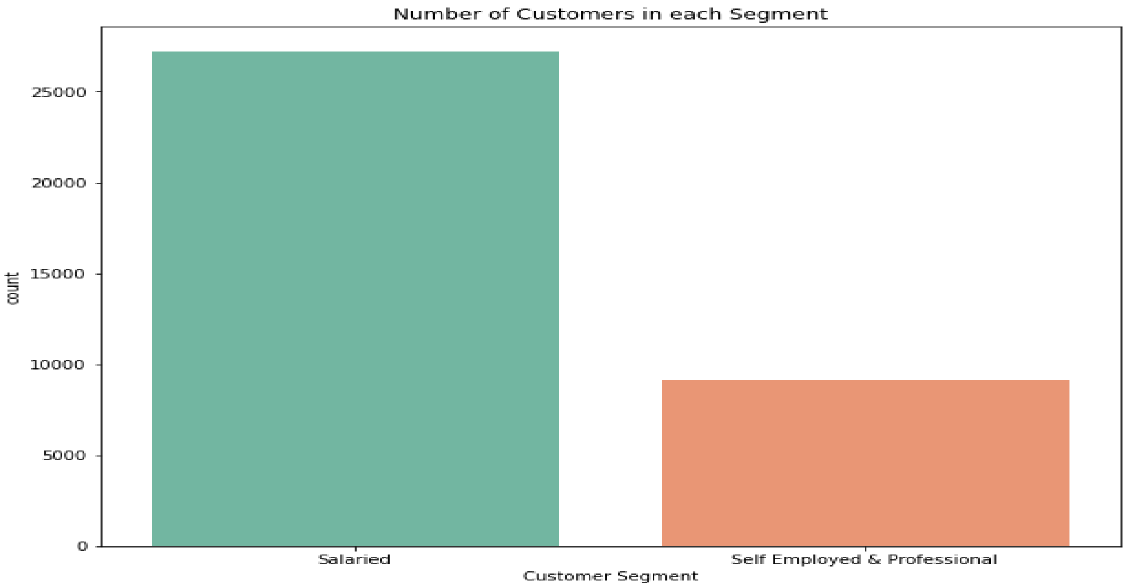
The number of males and females:

- The number of males is much higher than that of females.

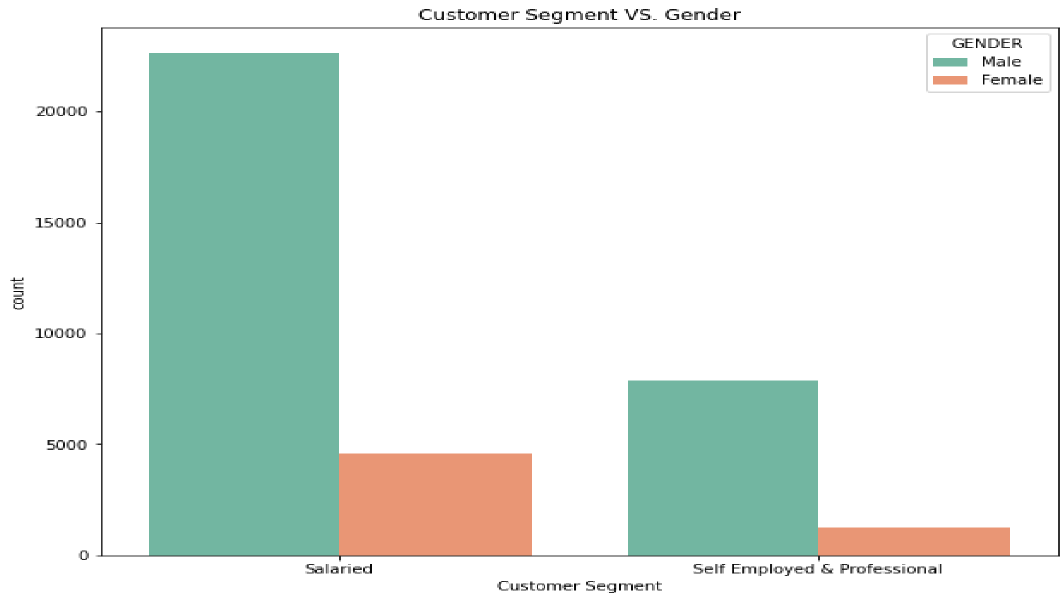


The Number of customers in each segment:

- The number of People in the Salaried Customer Segment is much more than people in the Self Employed & Professional Segment.

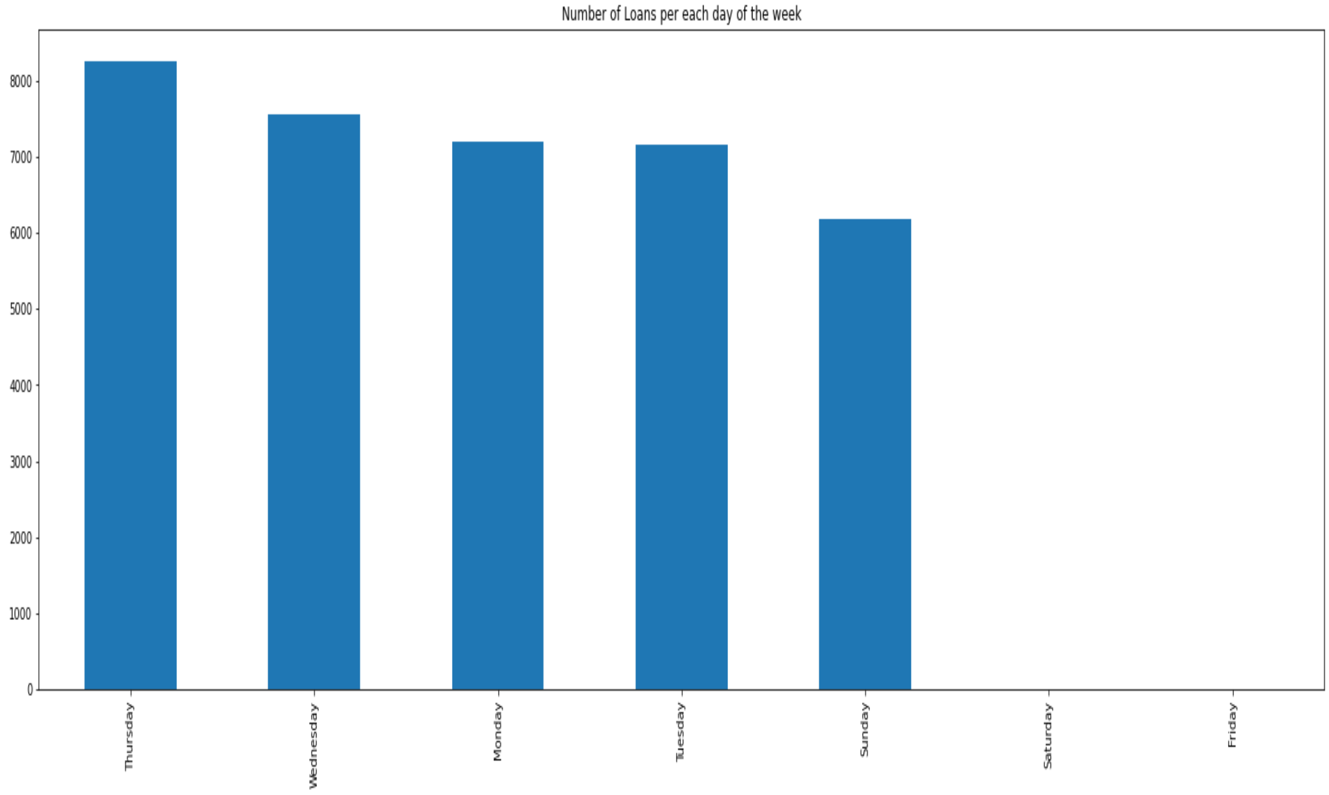


Customer segments and gender.



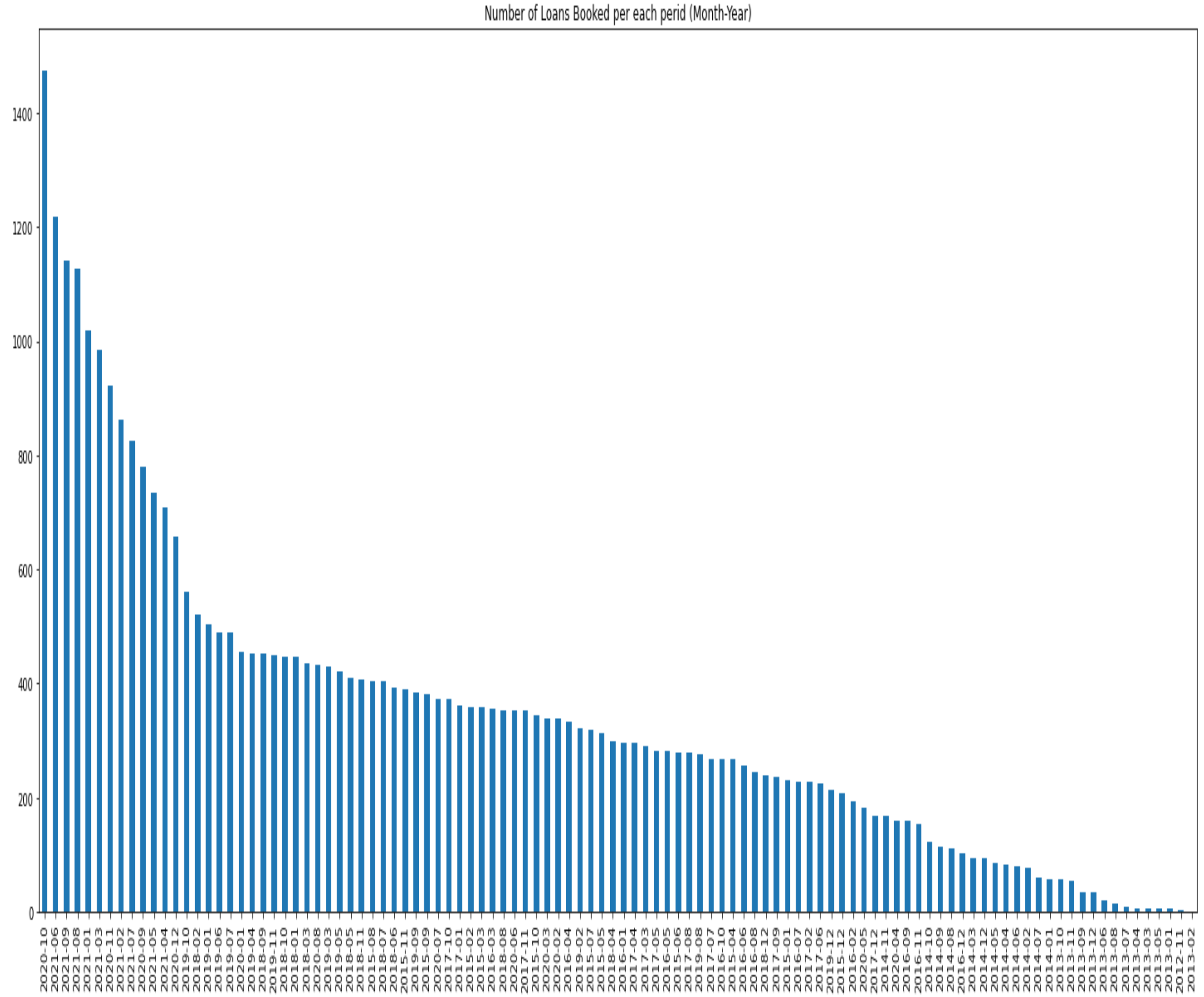
The number of loans booked on each day of the week:

- Most loans were booked on Thursday.



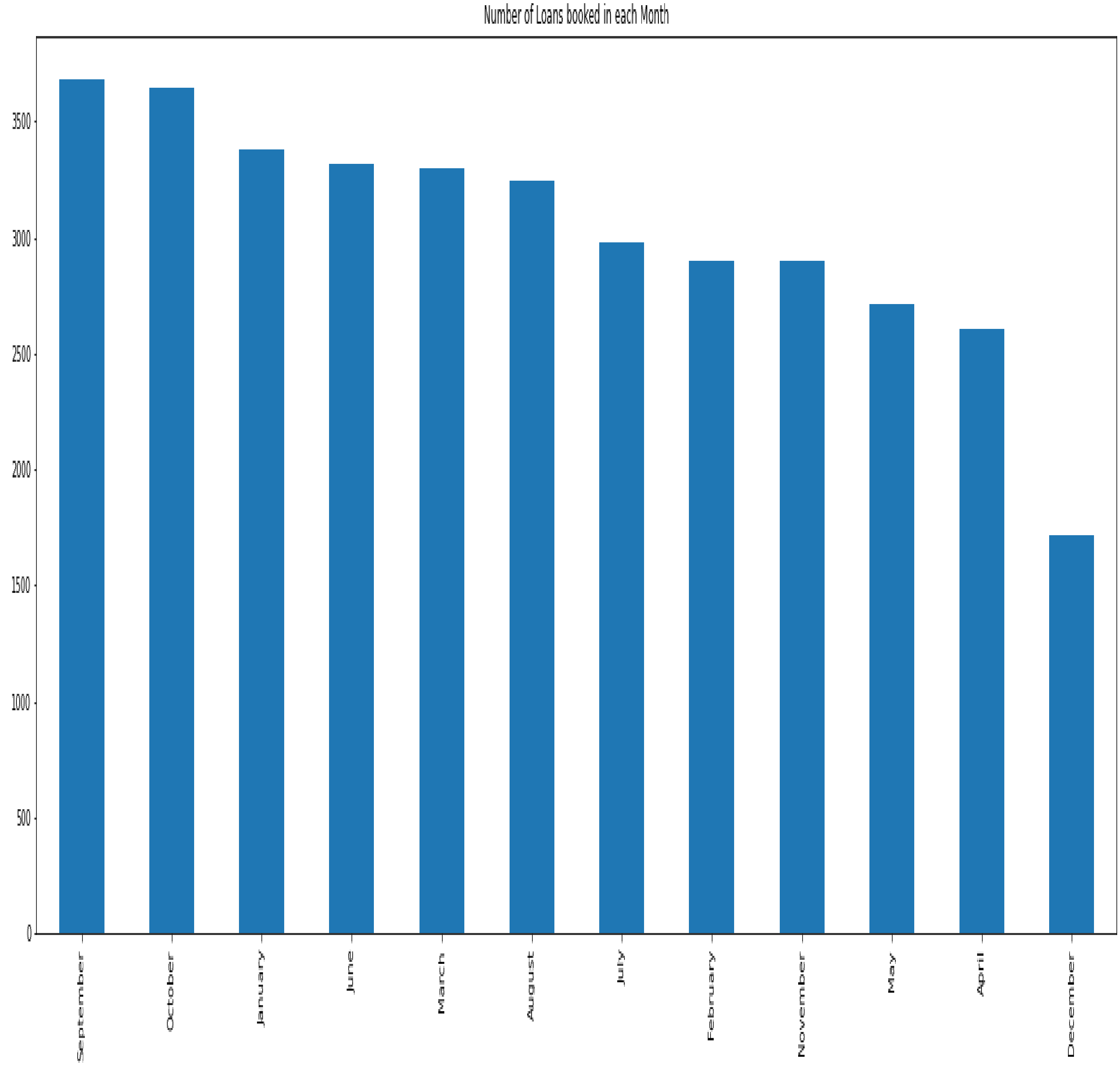
The number of loans booked per each period(month-year):

- Most loans were booked in 10-2020.

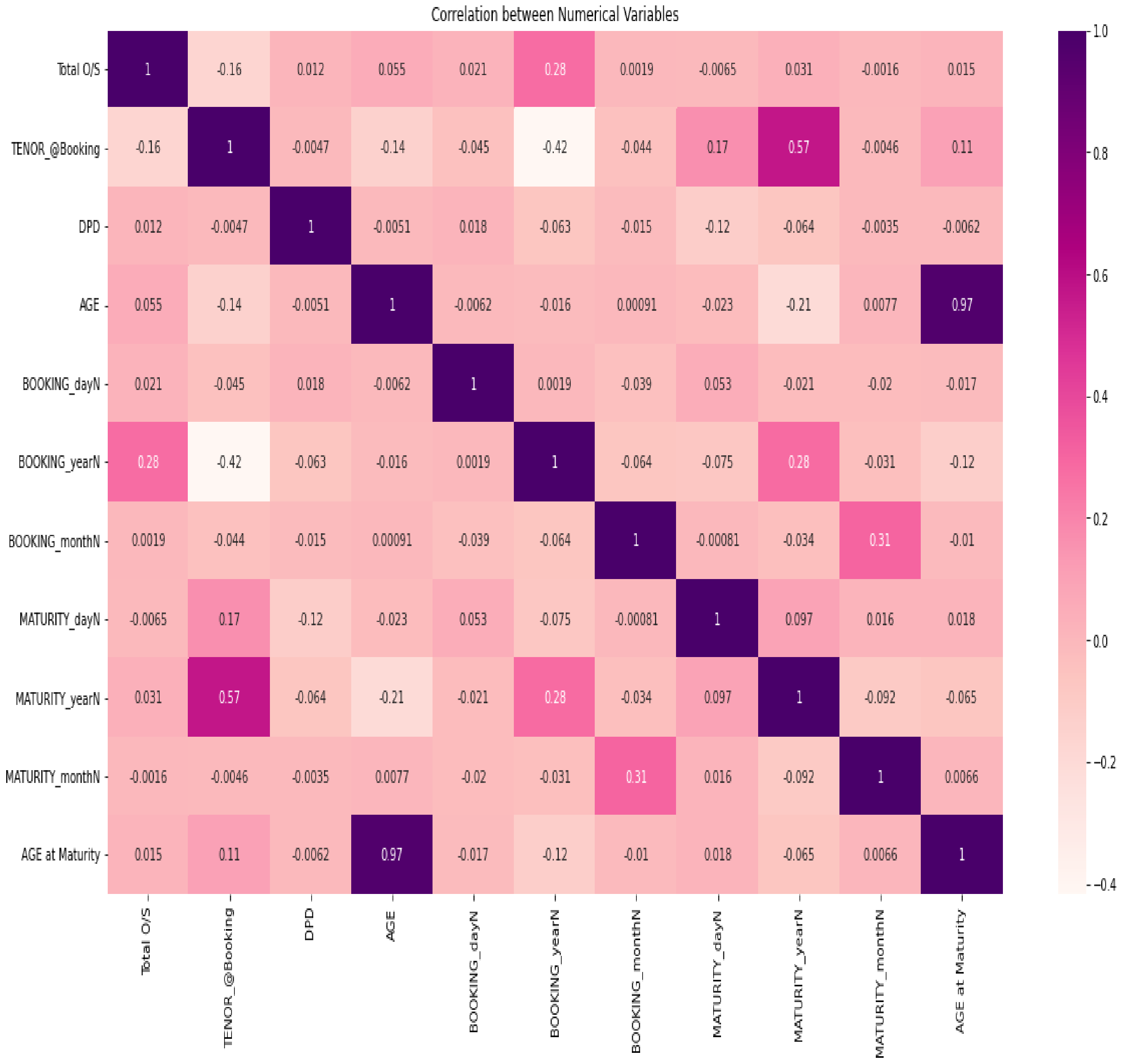


The number of loans booked in each month:

- Most loans were booked in September and October.

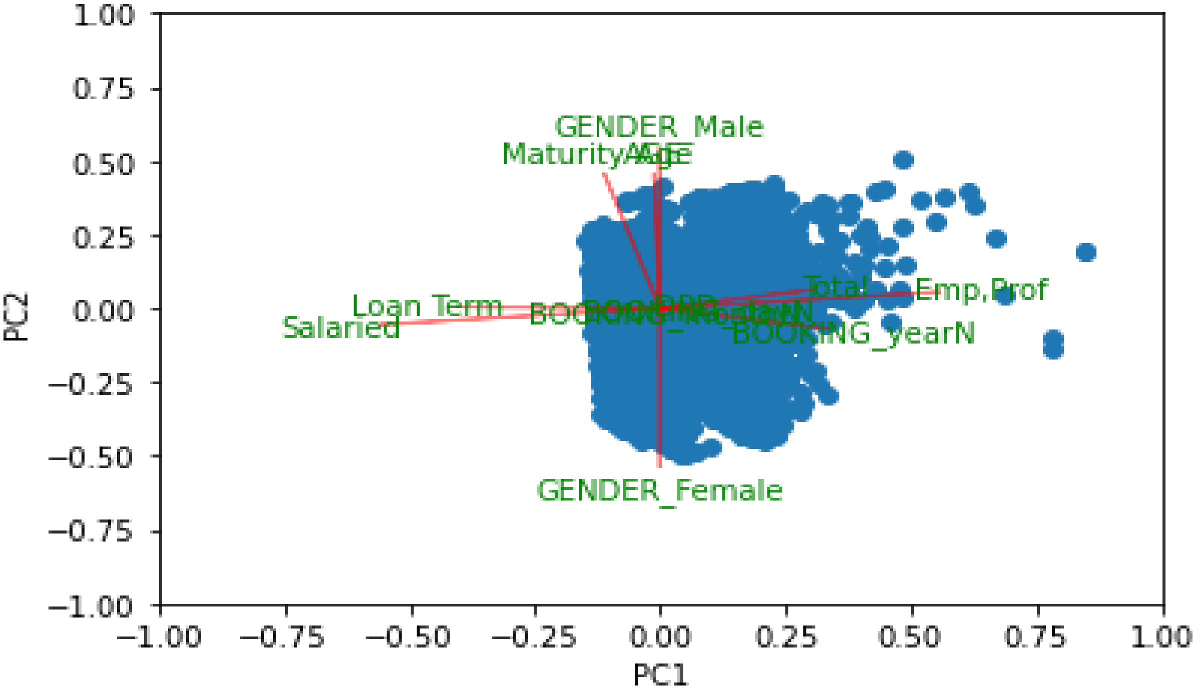


Correlation matrix.



Principal Component analysis

- PCA is used to decompose a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance. In scikit-learn, **PCA** is implemented as a *transformer* object that learns components in its fit method, and can be used on new data to project it on these components.
- Firstly, we should scale the data, so we scaled it using standard scalar.
- Applying PCA on the scaled data results in these principal components.

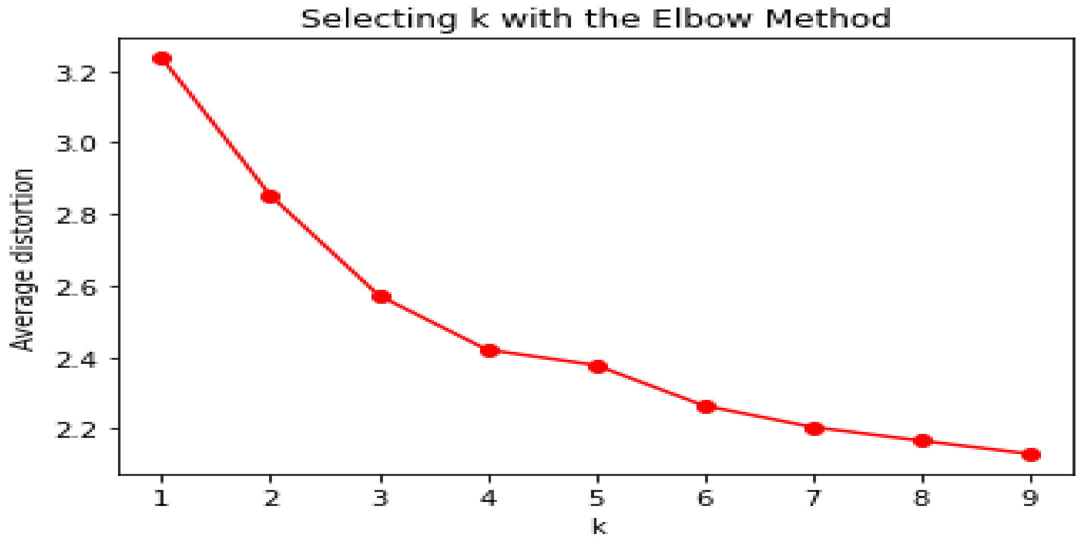


The feature with the most ability to distinguish between other features in the dataset is **GENDER**.

Clustering

K-Means Clustering:

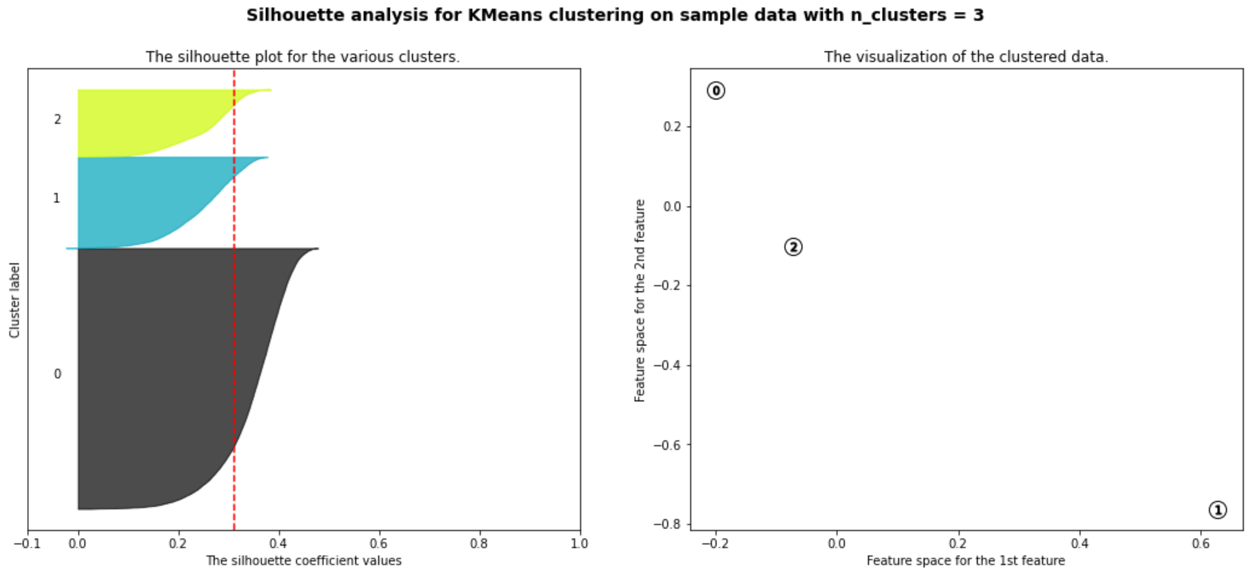
- The K-means algorithm aims to choose centroids that minimize the **inertia**, or **within-cluster sum-of-squares criterion**:
- Using the Elbow method to choose the best K value which is the number of clusters.



The best K value is 4 but let's compare between the average silhouette score of 3, 4, 5, and 6 clusters.

- Using The silhouette samples and score to visualize the scaled data:

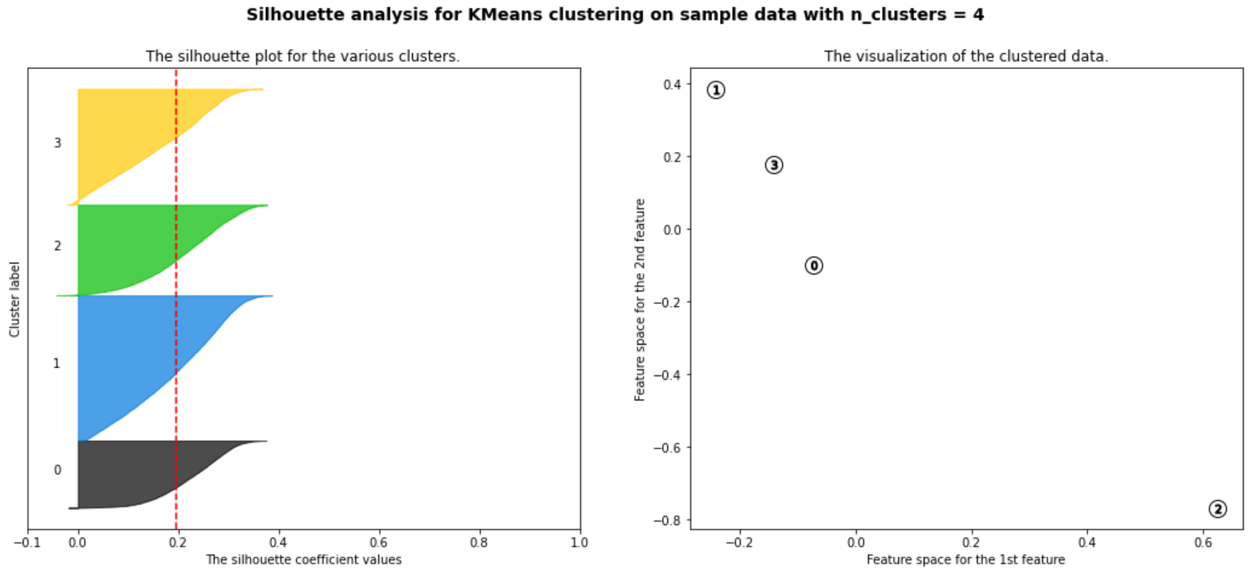
```
For n_clusters = 3 The average silhouette_score is :
0.31220752106783983

For n_clusters = 4 The average silhouette_score is :
0.1949636970359901

For n_clusters = 5 The average silhouette_score is :
0.18870629751290185

For n_clusters = 6 The average silhouette_score is :

0.19566370179644024
```



Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

The silhouette plot for the various clusters.

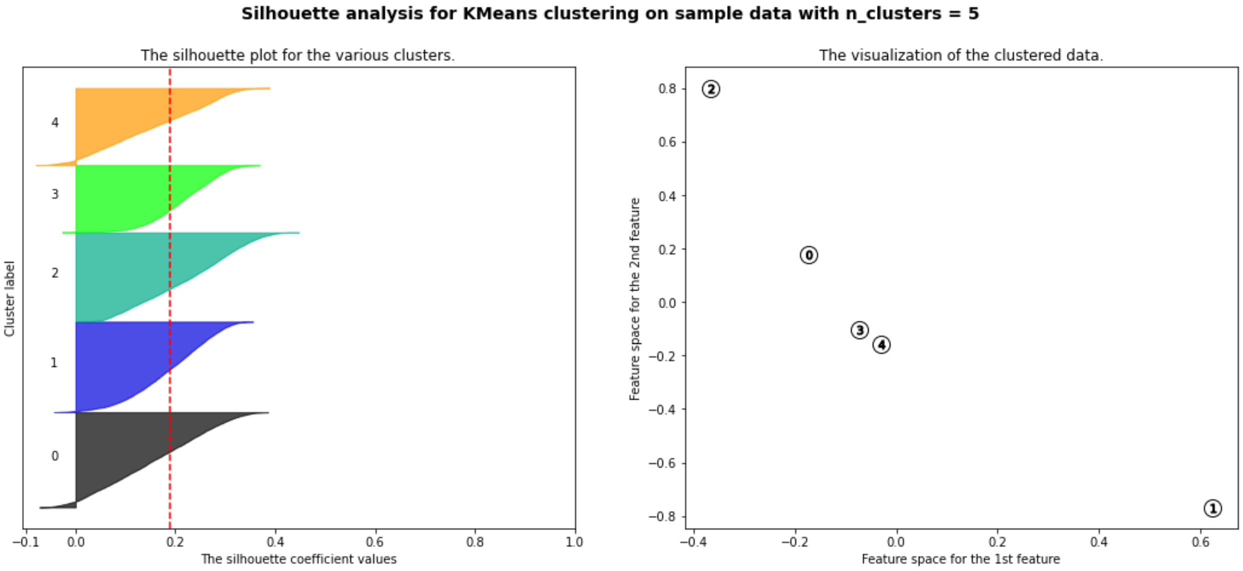
The visualization of the clustered data.

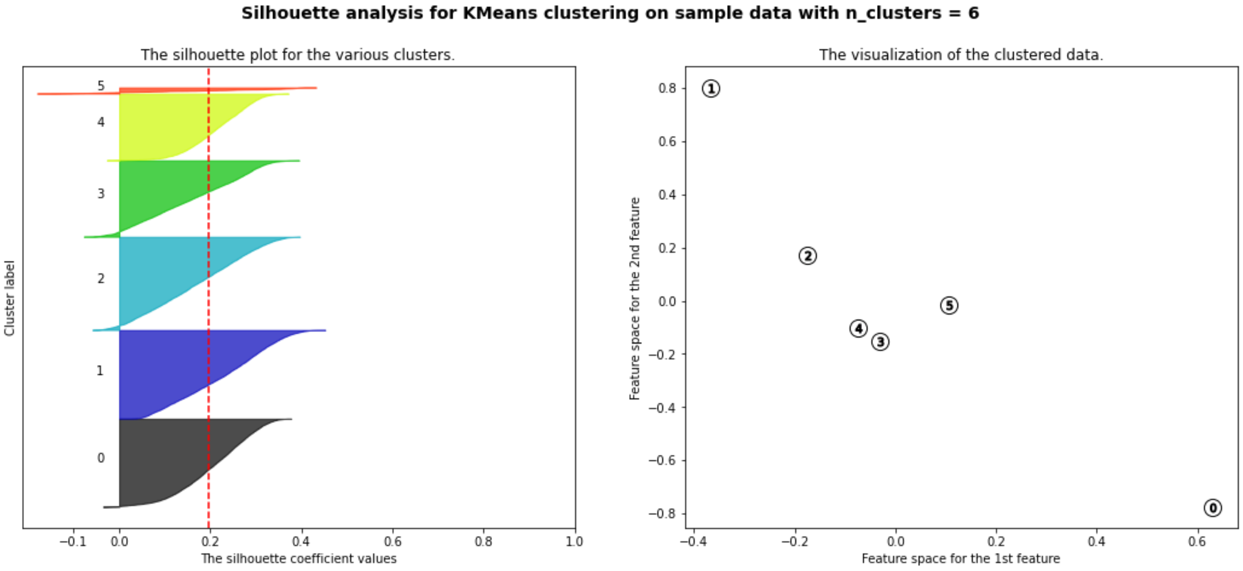


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4

The silhouette plot for the various clusters.

The visualization of the clustered data.





In the left figures, the thickness of each cluster label represents the number of examples in this group. While in the right figure, the numbers represent the cluster label and the coordinates are the centroid of each group to illustrate the distances between the centroids of each group.

Conclusion:

Therefore according to the average silhouette coefficient, the best number of clusters is 3 clusters. This means that the bank clients should be classified into three categories which are: accepted, rejected, and under_investigation.

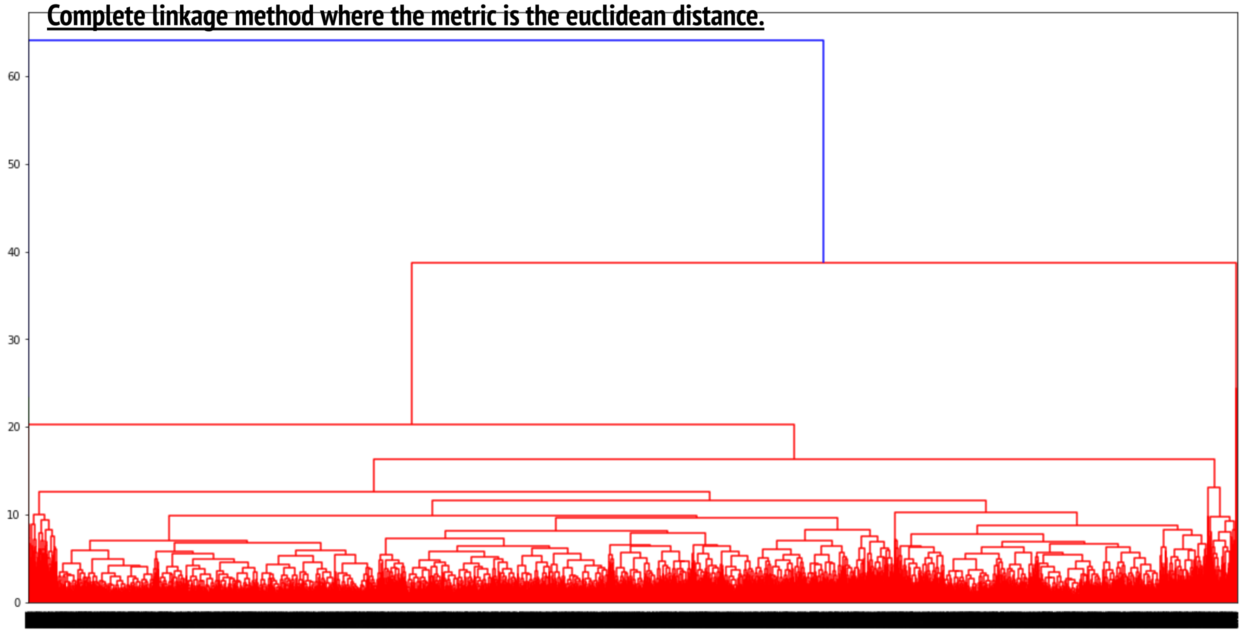
Hierarchical Clustering:

Hierarchical clustering is a clustering algorithm that builds nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a dendrogram. The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

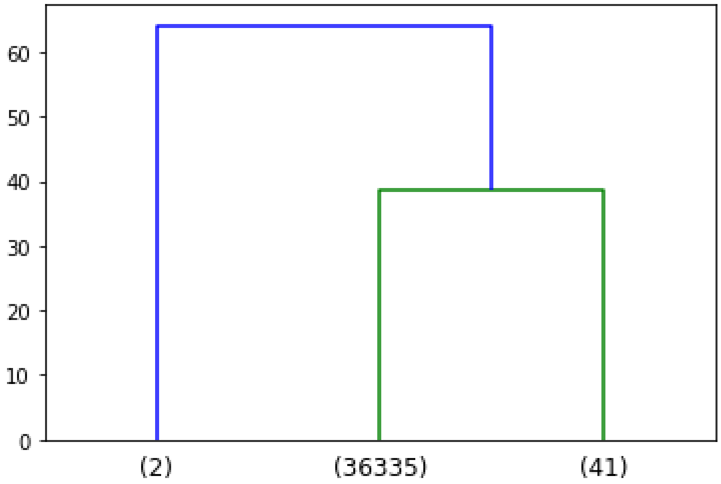
The **Agglomerative Clustering** object performs a hierarchical clustering using a bottom-up approach: each observation starts in its own cluster, and clusters are successively merged together. The linkage criteria determine the metric used for the merge strategy:

- **Ward** minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.
- **Maximum** or **complete linkage** minimizes the maximum distance between observations of pairs of clusters.
- **Average linkage** minimizes the average of the distances between all observations of pairs of clusters.
- **Single linkage** minimizes the distance between the closest observations of pairs of clusters.

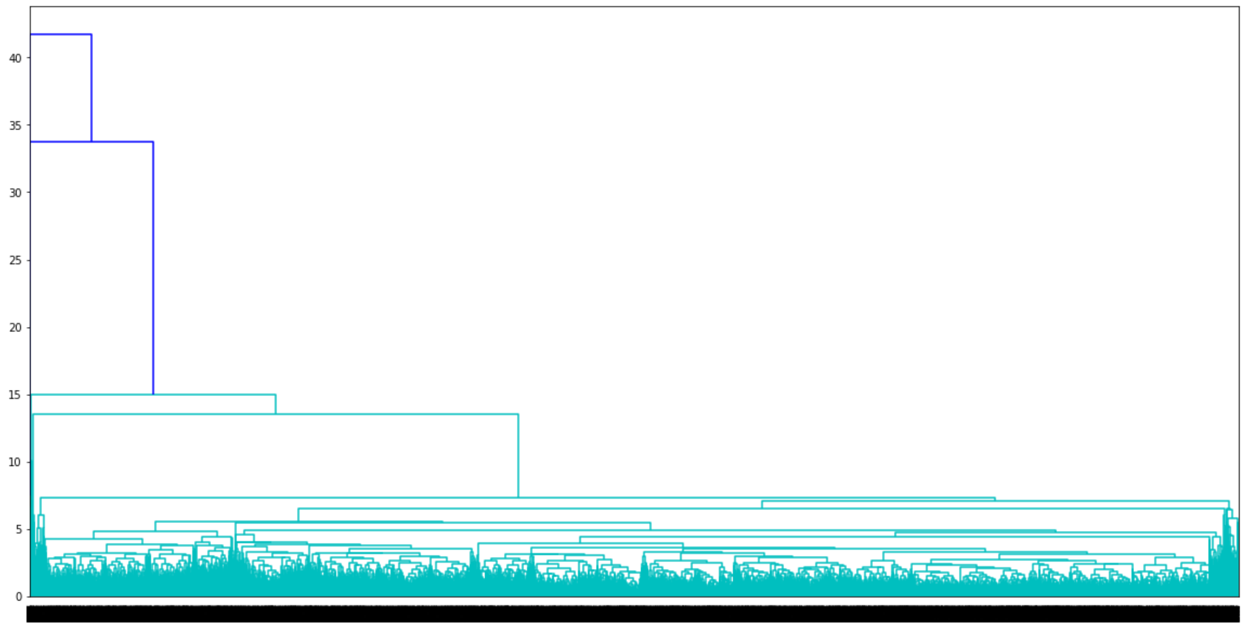
Complete linkage method where the metric is the euclidean distance.



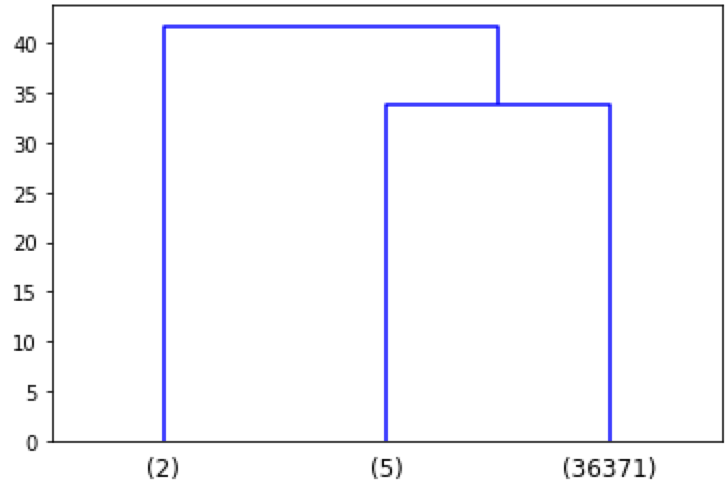
A truncated version using the last three clusters.



Average linkage method where the metric is the euclidean distance.



A truncated version using the last three clusters.



Conclusion:

- Based on the silhouette_scores, Average Linkage with k = 7 can be chosen as the final model for Hierarchical clustering.
- This means that the bank should classify the clients into 7 categories varying in the degree of acceptance.
- The models that can be developed using the presented data are the **unsupervised models**.
- The dataset doesn’t contain any target columns/features, hence **unsupervised models** such as K-means can be used with this data.
- We see that it might be useful if we have additional features/predictors like: Monthly Income, to check if there is a relationship between the Loan amount and the Monthly Income.