Q6. Bank Customer Clustering

Reference

https://www.kaggle.com/code/shawkyelgendy/customer-segmentation-eda-k-means-pca
https://www.kaggle.com/code/abduallahhussien/customer-segmentation-using-four-clustering-types

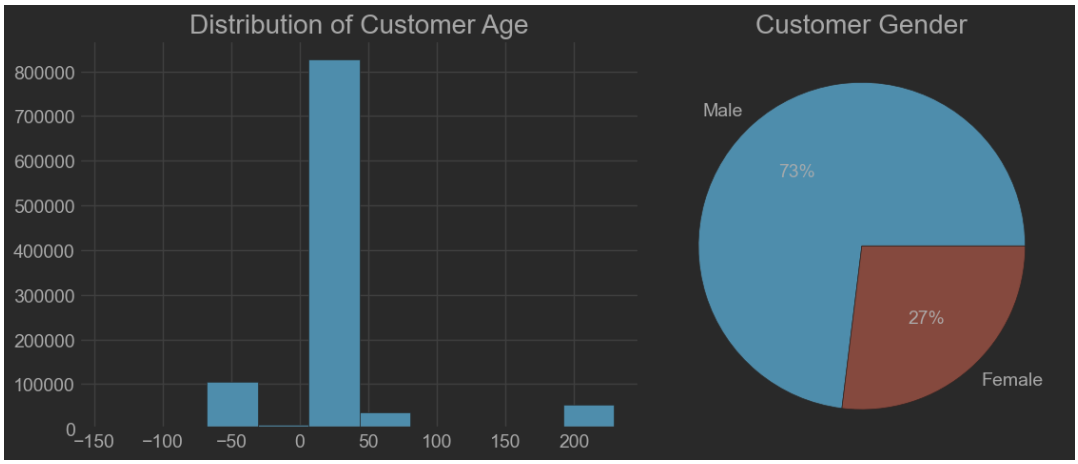
Cleaning data

1. Delete missing data.
2. Convert type of columns 'TransactionDate', 'CustomerDateOfBirth' from string to datetime.
3. Calculate customer age.
4. Drop outliers in genders

Explore data

1. Distribution of Customer Age

Most customers are between 0-50 years old.

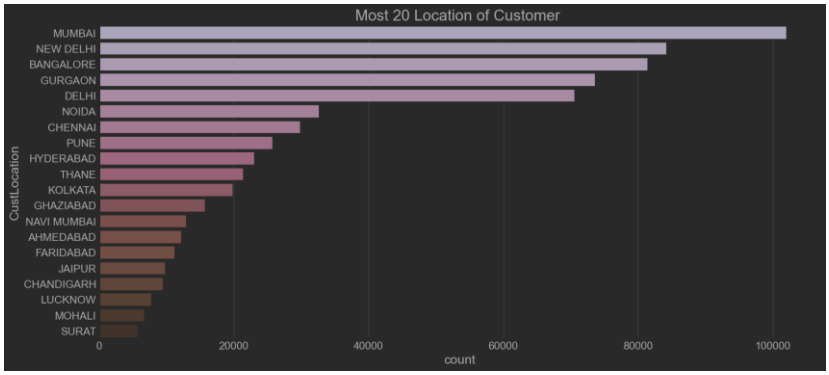


2. Customer Gender

Male customers account for 73%, while female customers only account for 27%, which is less than half of male customers.

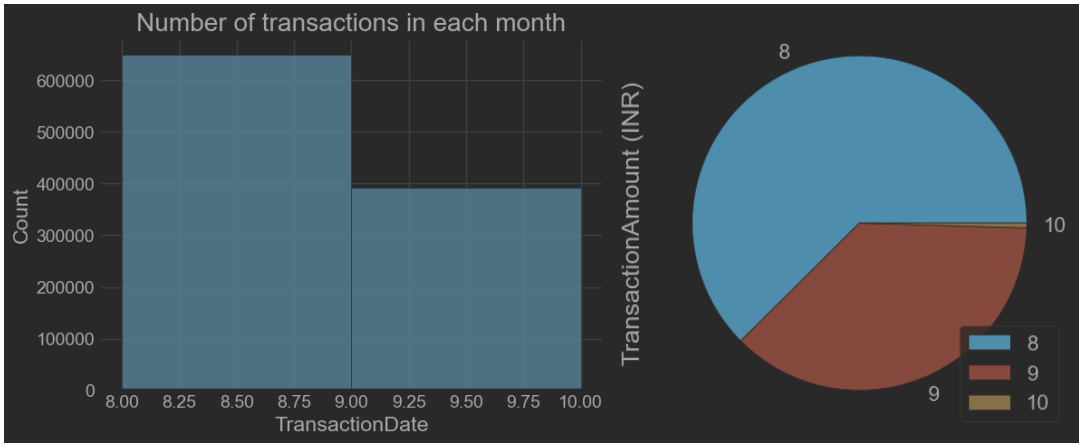
3. Most 20 Location of Customer

This chart shows the top 20 cities with the highest number of customers, followed by Mumbai and New Delhi.



4. Number of transactions in each month

In this dataset, there are only transactions in August, September, and October. The trading volume in August is greater than that in September, and there is only a small portion of transactions in October.

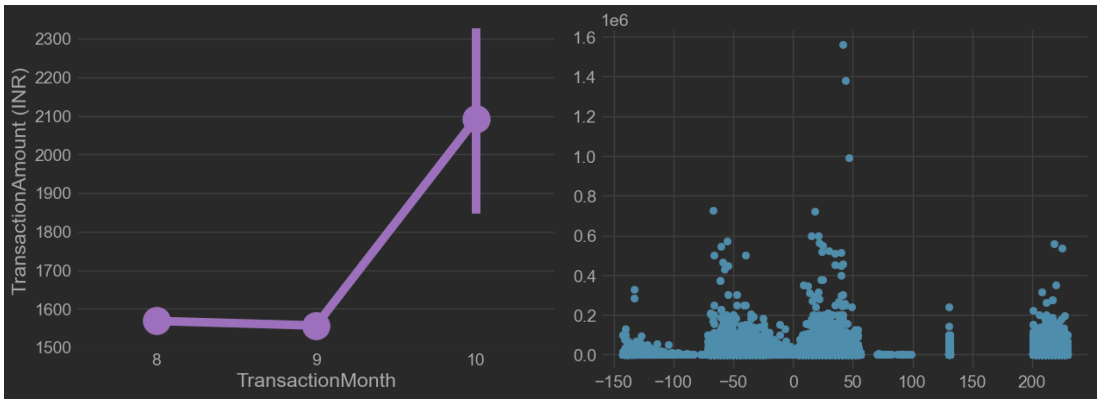


5. Pie chart for TransactionAmount in each month

This is the pie chart version of the image above, which more intuitively reflects the transaction amount of each month.

6. Point plot for TransactionAmount in each month

This plot give some additional information including the mean of each category of categorical variables, confidence interval, trend change.

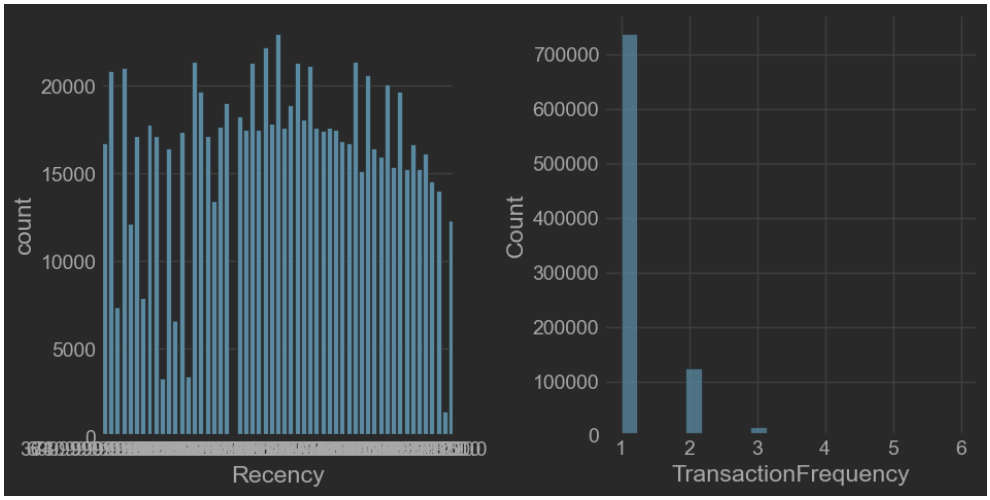


7. Scatter plot of age and transaction amount

This graph shows the correlation and distribution trend between age and transaction amount. We can see that most of the points are concentrated in four areas. The areas on both sides have fewer discrete points, while the middle area has more discrete points.

8. Count of Recency

Recency: number of days since the last purchase or order so I will create a new column of TransactionDate to subtract the last transaction from the first transaction



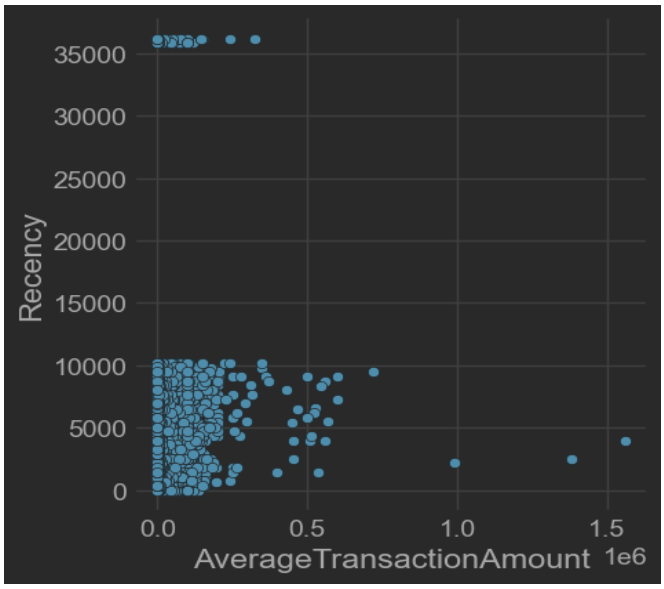
9. Histogram of TransactionFrequency

It can be seen that the higher the frequency, the fewer people there are, and the curvature decreases. The frequency of 1 is the highest, several times that of 2, and 2 is also several times that of 3. There are almost no more frequencies of 4, 5, or 6.

10. Scatter plot of AverageTransactionAmount and Recency

It can be seen that most of the points are clustered between 0 to 10000 in currency and 0 to 0.2 in

amount. A small portion of them gather at a revenue of over 35000.

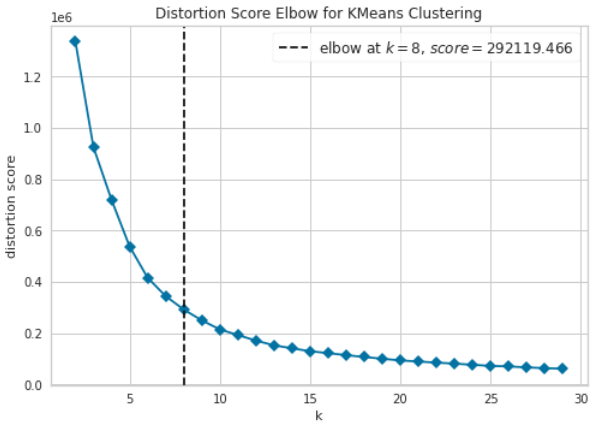


Clustering customers

1. Centroid-based clustering: k-means

K-Means algorithm:

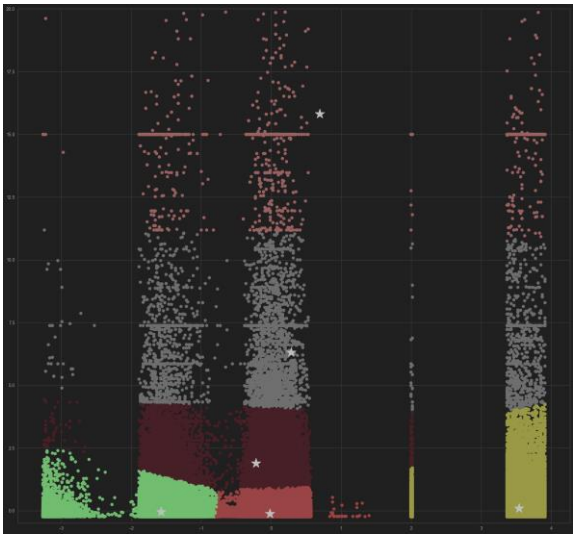
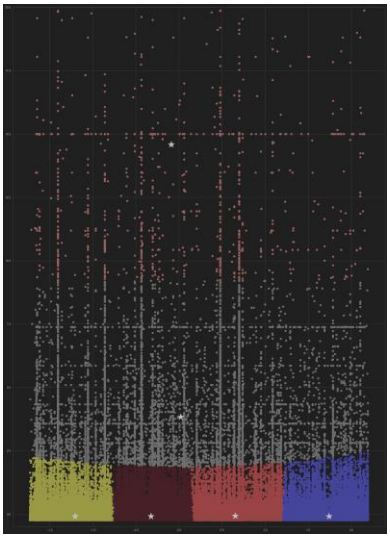
1. Select the number of clusters K
2. Initialize the K centroids
3. Assign each data point to their closest centroid
4. For each cluster calculate the average of its assigned examples and let it the new position for that centroid
5. Reassign each data example to the new closest centroid of each cluster
6. Update the centroid position



Result:

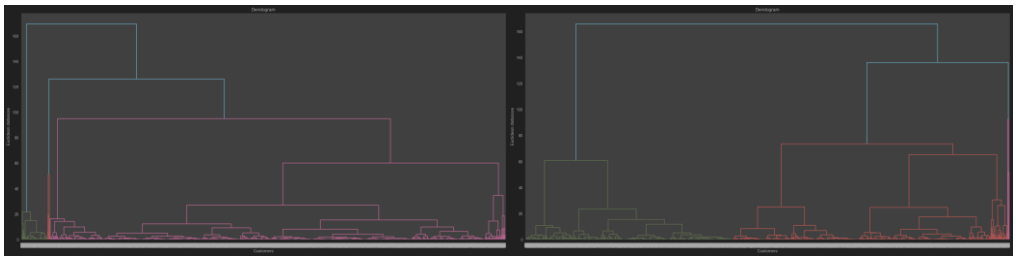
Clustering into 8 categories based on age and location with INR,

It can be seen that as the INR increases, age and location no longer affect clustering, while those with lower INR will be clustered together due to age and location.

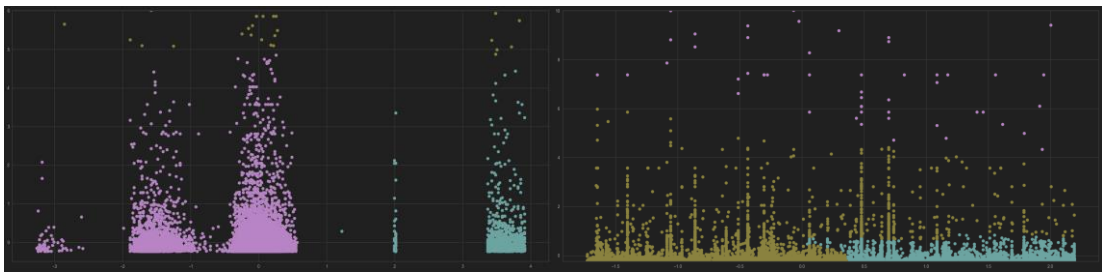
On age	On location
	

2. Hierarchical clustering: Agglomerative

Hierarchical clustering is where you build a dendrogram to represent data, where each group links to two or more successor groups. The groups are nested and organized as a tree, which ideally ends up as a meaningful classification scheme.



Each node in the cluster tree contains a group of similar data; Nodes group on the graph next to other, similar nodes. Clusters at one level join with clusters in the next level up, using a degree of similarity. The process carries on until all nodes are in the tree, which gives a visual snapshot of the data contained in the whole set. The total number of clusters is not predetermined before you start the tree creation.



Agglomerative is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

3. Density-based clustering: DBSCAN

Key Characteristics of DBSCAN Algorithm

1. It does not require the number of clusters as input.
2. It is can detect outliers while finding clusters.
3. DBSCAN algorithm can detect clusters that are complex or randomly shaped and sized.

Finding the Optimal value of Epsilon

The average distance between each point and its k nearest neighbors is calculated where k = the MinPts selected by us. We then plot the average k-distances in ascending order on a k-distance graph. The optimal value for epsilon is the point with maximum curvature or bend, i.e. at the greatest slope.