Project Report:

Bank Customer Segmentation

1. Introduction

In today's competitive banking landscape, understanding customers' diverse needs is crucial for success. Customer segmentation, the process of categorizing customers based on shared characteristics, behaviours and preferences is an invaluable tool for banks. By segmenting their customer base, banks can personalize services, optimize resource allocation and drive growth. This project explores the implementation of customer segmentation strategies to enhance customer experience and boost business performance.

Most banks have a large customer base - with different characteristics in terms of age, transaction_amount, transaction_time, location of customers, and more. Customer segmentation is the process of dividing a customer dataset into specific groups based on shared traits.

2. Research Objective

The Goal of this project is to create a model that will help bank to segment customers according to their Age, Recency value, Frequency value & Monetary value.

There are some questions I'd like to answer that will help us to learn more about customers and their behaviour.

(i) To find out the top 3 cities whose customers have the most money in their accounts.

(ii) To find out the top 3 cities whose customers have transacted the most money.

(iii) To find out whether male customers or female customers do more amount of transaction.

(iv) To find out whether male customers or female customers have more money in their accounts?

(v) To find out on which day of the week the highest and lowest transactions took place and how much.

(vi) To find out in which month the highest and lowest transactions occurred per month and for what amount.

(vii) To find out which age group of customers do maximum transaction Amount.

(viii) To find out which age group of customers has maximum account balance.

3. Data Collection

We took dataset from Kaggle. This dataset consists of 1 Million+ transaction by customers for a bank in India. The data contains information such as-TransactionID, CustomerID, CustomerDOB, CustGendere, CustLocation, CustAccountBalance, TransactionDate, TransactionTime, TransactionAmount (INR) etc.

4. Data Preprocessing

(i) Drop null values because we found very small number of null values as compare to our dataset.

(ii) Change datatypes of some columns, From float to int (for numerical columns) & object to datetime (for date columns).

(iii) Find value count of Gender column, we find there is only one count for Transgender, then we drop it.

(iv) Find Customer's Age (As a new column).

To find customer's age we have to need customer birth year and current year.

- We extract year from CustomerDOB and the extract year will act as customer birth year.
- Similarly, we extract transaction year from TransactionDate and the extract year will act as current year. Together with we will also extract transaction_day and transaction_month from transaction_year for further analysis.

(v) Drop unnecessary columns.

(vi) Treat outliers of the data.

(vii) We took/extract 50,000 rows data randomly for our further analysis.

(viii) Find/Extract RFM values (As new columns).

(ix) Finally, we make a new dataframe using RFM & Customer Age, which is used by us for further analysis.

(x) Then we treat duplicate values, outliers from the new dataframe that contains RFM.

5. Segmentation Methods

(a) Clustering Algorithm

K-Means

K-Means is a heuristic algorithm used for clustering data based on a measure of closeness. It groups data into K clusters, where K represents the number of desired clusters. The algorithm works iteratively by moving the centroids (cluster centers) to the mean position of their constituent points and reassigning data instances to their closest clusters. This process continues iteratively until no significant change in the cluster centers is possible, leading to the convergence of the algorithm.

(b) Dimensionality Reduction

PCA (Principal Component Analysis)

Principal Component Analysis (PCA) is a widely used technique in machine learning and statistics for reducing the dimensionality of data while preserving as much variance as possible. It achieves this by transforming the original features into a new set of orthogonal (uncorrelated) features called principal components.

6. Validation Technique/Evaluation Technique

Silhouette Score

The Silhouette Score is a way to measure how well data points are clustered. Imagine you have some data and you're trying to group similar points together. The Silhouette Score tells you how close each point is to the other points in its own group compared to the closest neighbouring group.

In simpler terms, it's like giving each data point a grade on how well it fits with its group. If a point is really close to other points in its group and pretty far from points in other groups, it gets a high grade (good silhouette score). But if a point is sort of in the middle, not too close to its group members and not too far from neighbouring groups, it gets a lower grade (lower silhouette score). So, the higher the Silhouette Score is for the whole dataset, the better the clustering is working.

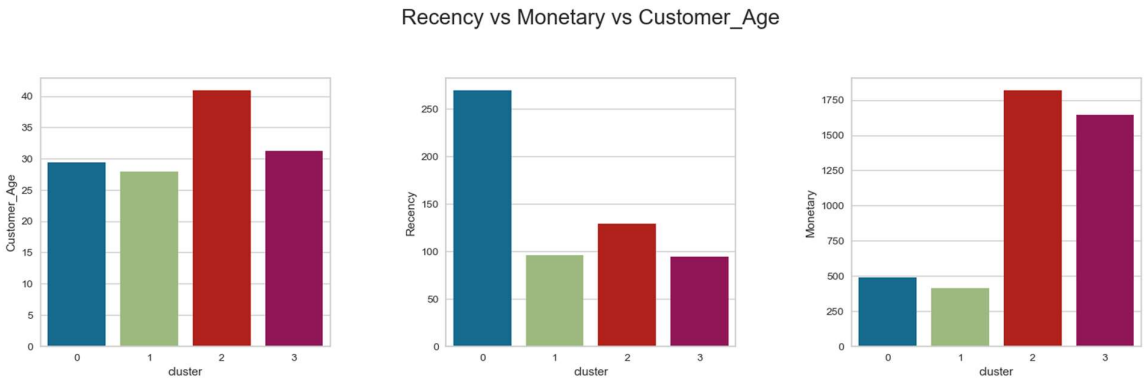
Silhouette score range is (-1 to 1). It means,

If Silhouette score is near to 1 or equal to 1: The clustering of datapoints is good and if silhouette score is near to -1 or equal to -1: The clustering of datapoints is not good. We should have to use another technique or algorithm for better clustering.

7. Finalize Model

We use PCA applies K-Means model. Because of their silhouette score i.e, [0.48] which is a good score or greater than simple K-Means (i.e, 0.40) that's why cluster separation provided by the model is more compelling. We can see group better.

8. Segmentation Analysis & Visualization



From the above graphs we can conclude the following: -

Cluster 0: Shows high average recency and low monetary with average customer age around 29.

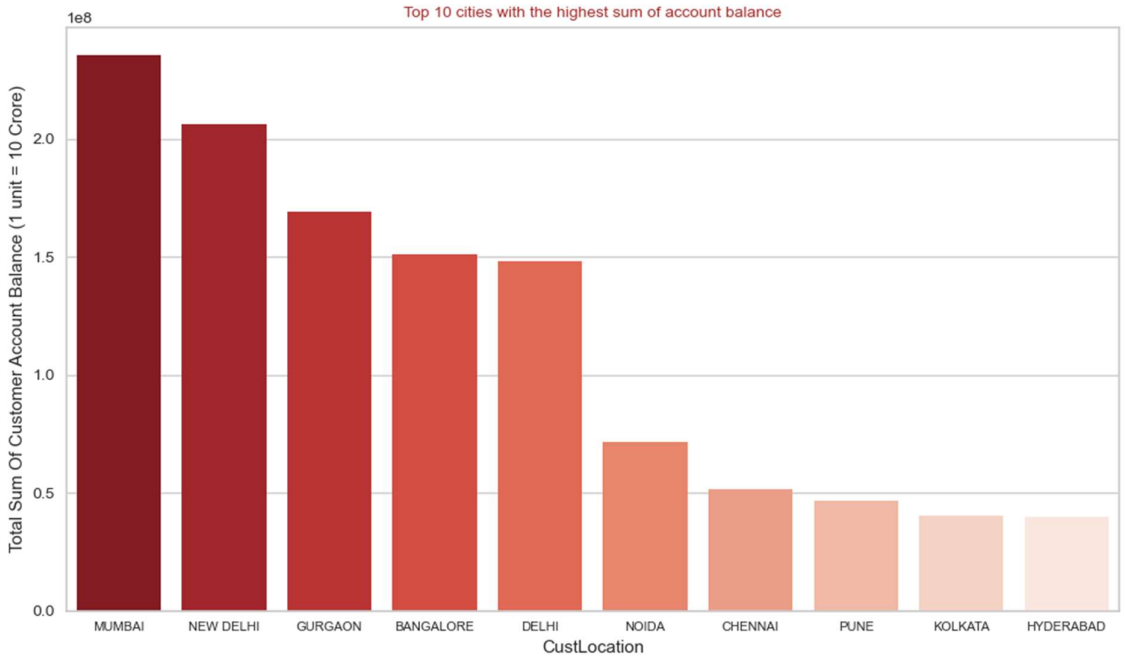
Cluster 1: Shows low average recency and low monetary with average customer age around 27.

Cluster 2: Shows moderate average recency and high monetary with average customer age around 41.

Cluster 3: Shows low average recency and high monetary with average customer age around 31.

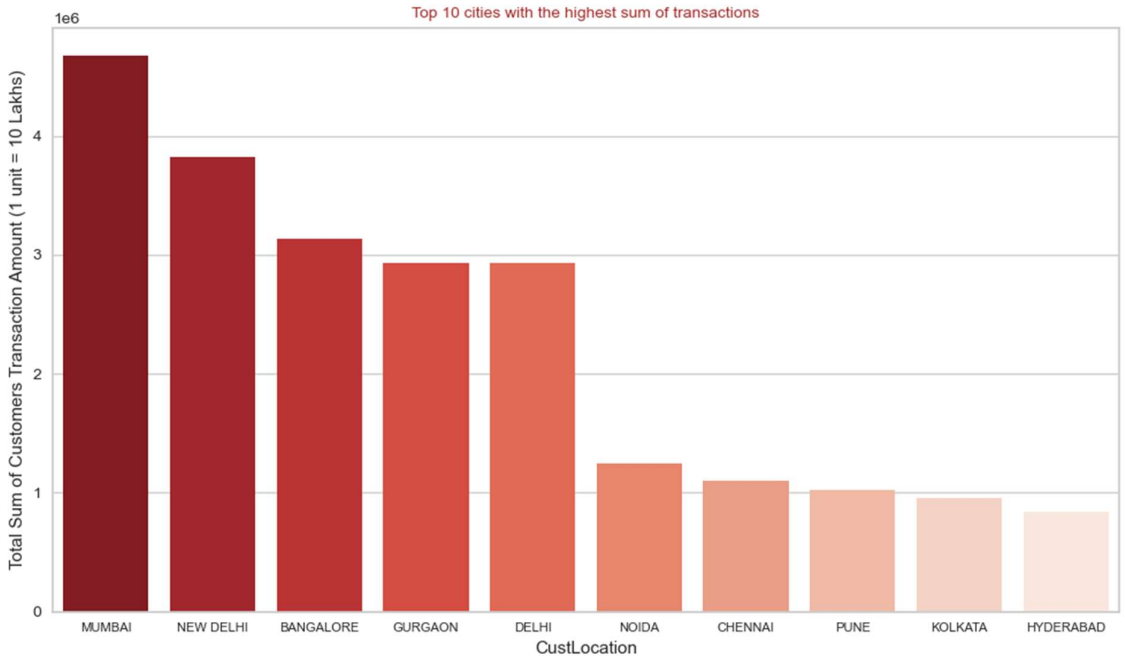
9. Answers to the Questions I wanted to find are as below serial wise: -

(i)



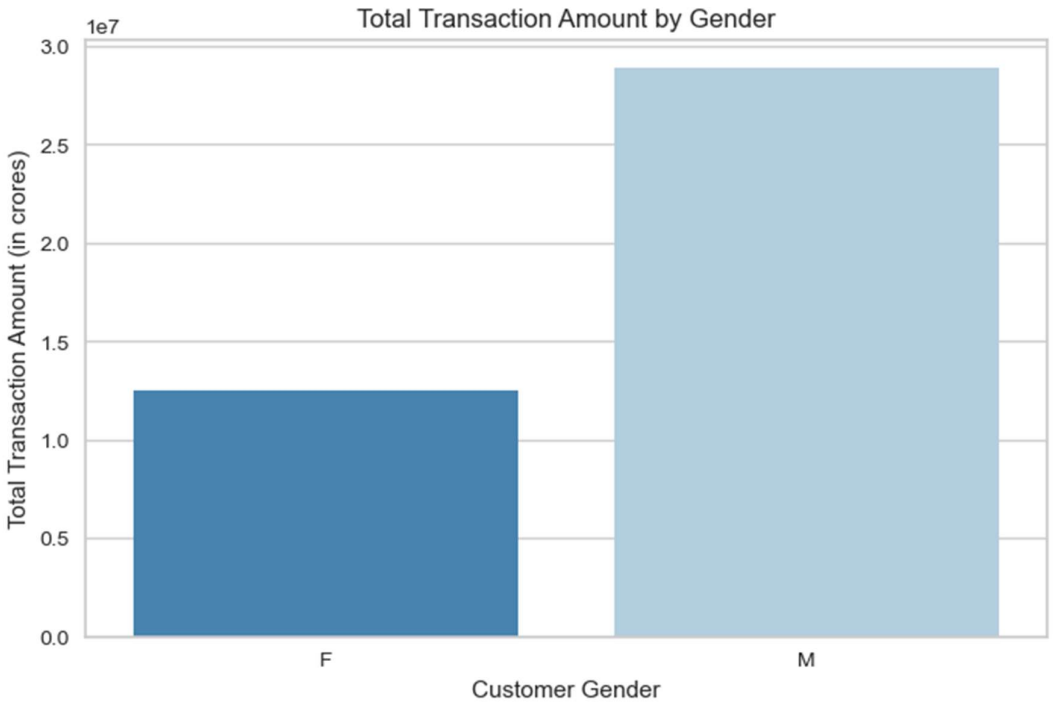
Above graph shows top 10 cities with highest amount of money in the account of their customers.

(ii)



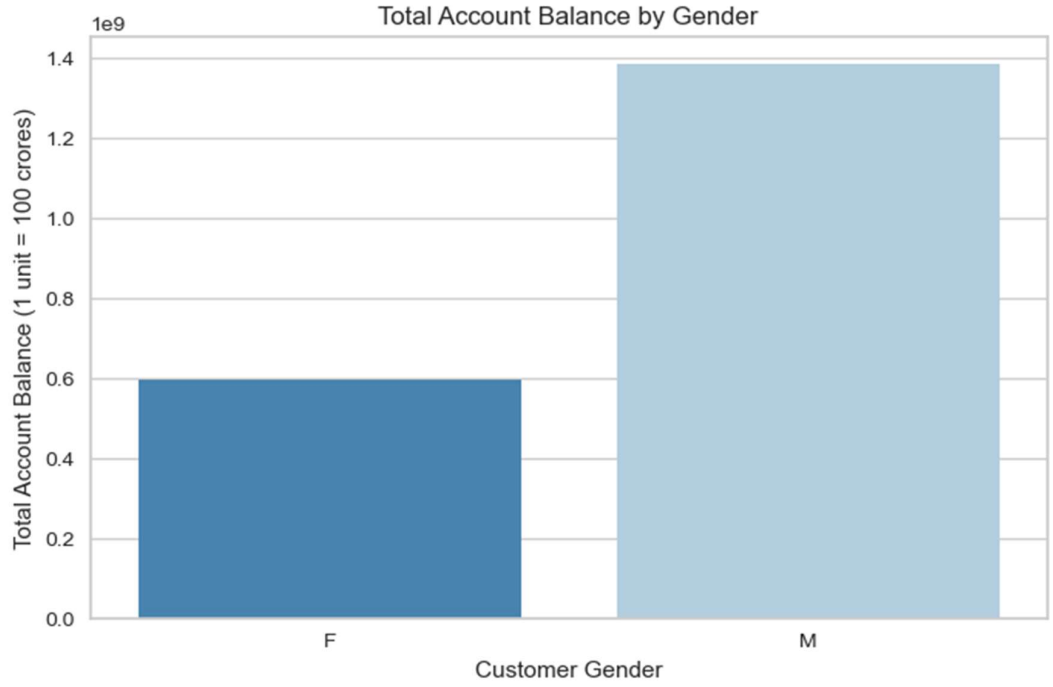
Above graph shows top 10 cities customers have transacted the highest amount.

(iii)



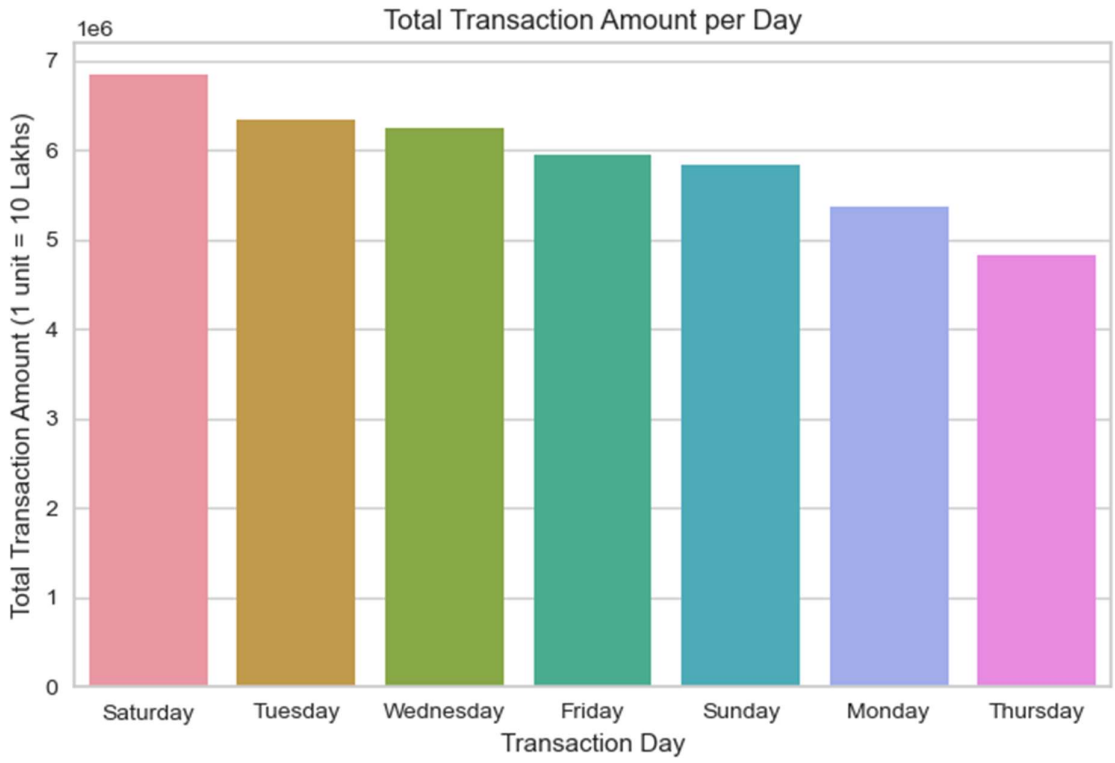
Above graph shows that male customers do more amount of transaction.

(iv)



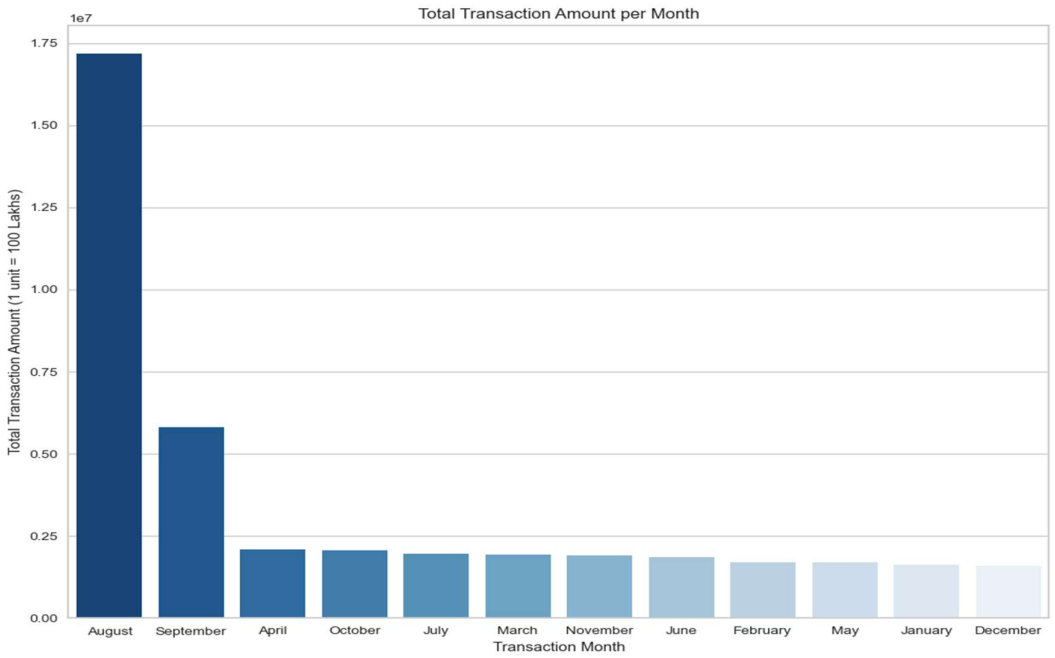
Above graph shows male customers have more money in their accounts.

(v)



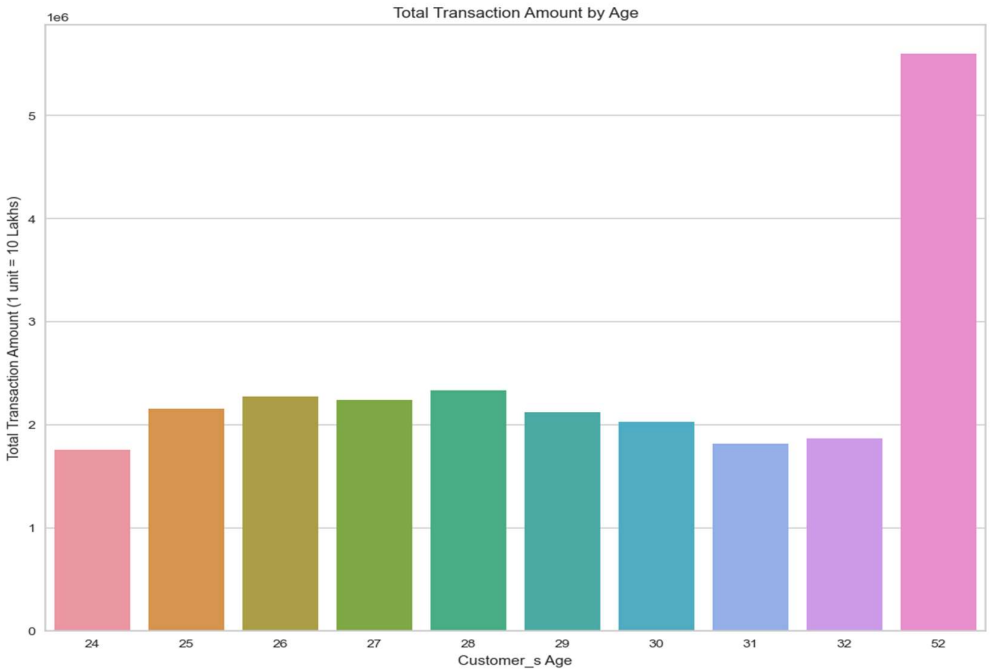
Above graph shows that Customers do maximum amount of transaction on Saturday and minimum amount of transaction on Thursday according to the days in a week.

(vi)



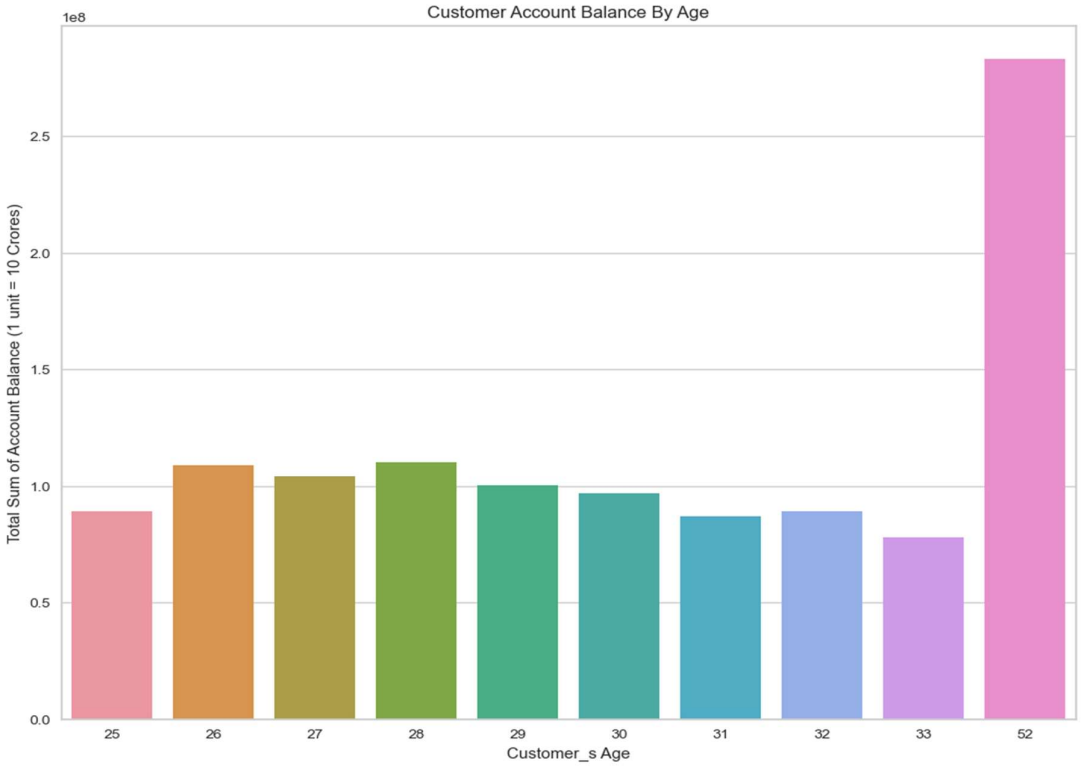
Above graph shows that Customers do maximum amount of transaction in August and minimum amount of transaction in December according to the months in a year.

(vii)



Above graph shows Customers of 52 years age do maximum amount of transaction.

(viii)



Above graph shows that Customers of 52 years age have maximum balance in their accounts.

10. Business Insights & Recommendations

From the above insights, we get –

Number 1

(a) Customer's whose average age around 29 are lies in top 5 in list of customers age group which have maximum account balance.

(b) Average Monetary value of these customers is low & their average recency value is high.

Therefore, we will offer them to make FD of their amount (with different tenures i.e, as long as possible) which will be profitable for both, customers & bank also.

- **Customer benefit: -**

  They will get high interest as compare to normal.

- **Bank benefit: -**

  Bank will earn double interest by renting the FD money.

Number 2

(a) Customer's whose average age around 27 has high amount in their accounts.

(b) Average Recency value of these customers is low.

Due to their low recency, it becomes easier for the bank employees to persuade them to invest their money in schemes that will be beneficial to both the parties. For example – Motivate them to invest money in government schemes (like Sukanya Samriddhi Yojana).

- **Customer benefit: -**

  They will get good interest rates and will also get tax exemption.

- **Bank benefit: -**

  Bank will get money for long term which is profitable for it.

Number 3

(a) Customer's whose average age around 41 has low amount in their accounts.

(b) Average Recency value of these customers is moderate and average monetary value is high.

Therefore, we will give some offers to them to decrease their recency value.

- **Bank benefit: -**

  Bank will get more money after decreasing their average recency value because their average monetary value is already high.

Number 4

(a) Customer's whose average age around 31 has moderate amount in their accounts.

(b) Average Recency value of these customers is low and average monetary value is high.

Due to their low recency, it becomes easier for the bank employees to persuade them to take loan.

- **Customer benefit: -**

  They will get money for their future plan or to start some startup.

- **Bank benefit: -**

  Bank gets interest.