Northeastern University
College of Professional Studies

Northeastern University
College of Professional Studies

ALY6110 - Data Management and Big Data

Module 4 Final Project

Professor Olesya Agafontseva

11th June 2025

Group 6: Payal Sharma, Min Zeng, Hang Chen,Yuhong Tao

1. Introduction

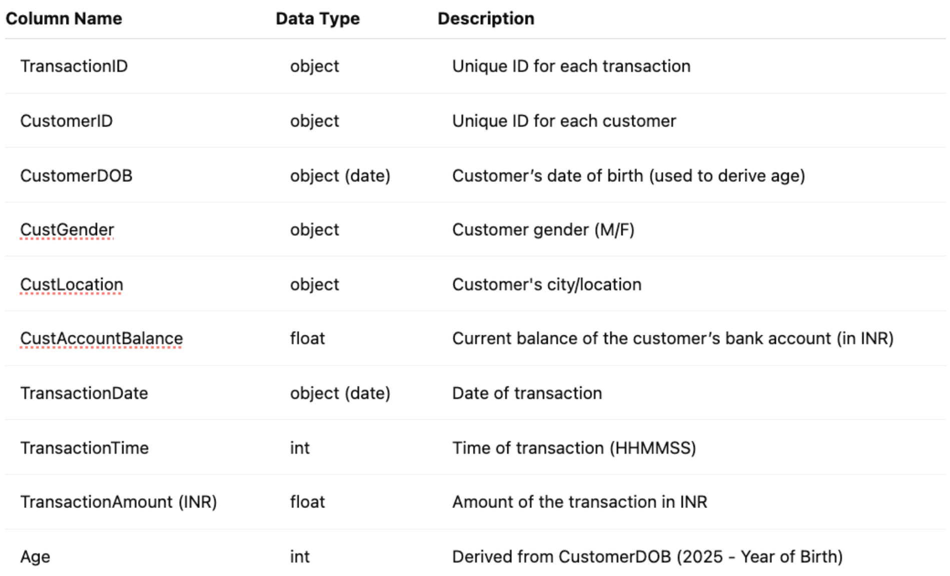
Understanding customer behavior is critical for effective decision-making in the banking industry. This report presents a basic analysis of a large bank transaction dataset using Python, focusing on identifying customer patterns through correlation analysis, clustering, and visual exploration. The insights generated aim to support strategic customer segmentation and engagement efforts.

2. Dataset Overview

2.1 Description

The dataset used is from Kaggle: Bank Customer Segmentation Dataset. It contains over 1 million transaction records including customer demographics, transaction details, and account balances.

2.2 Variables and Data Types



Column Name	Data Type	Description
TransactionID	object	Unique ID for each transaction
CustomerID	object	Unique ID for each customer
CustomerDOB	object (date)	Customer’s date of birth (used to derive age)
CustGender	object	Customer gender (M/F)
CustLocation	object	Customer's city/location
CustAccountBalance	float	Current balance of the customer’s bank account (in INR)
TransactionDate	object (date)	Date of transaction
TransactionTime	int	Time of transaction (HHMMSS)
TransactionAmount (INR)	float	Amount of the transaction in INR
Age	int	Derived from CustomerDOB (2025 - Year of Birth)

2.3 Notes:

- Age was derived from CustomerDOB using datetime conversion.
- Gender and location were used to identify customer segments.
- Transaction time was formatted for potential time-based pattern analysis.

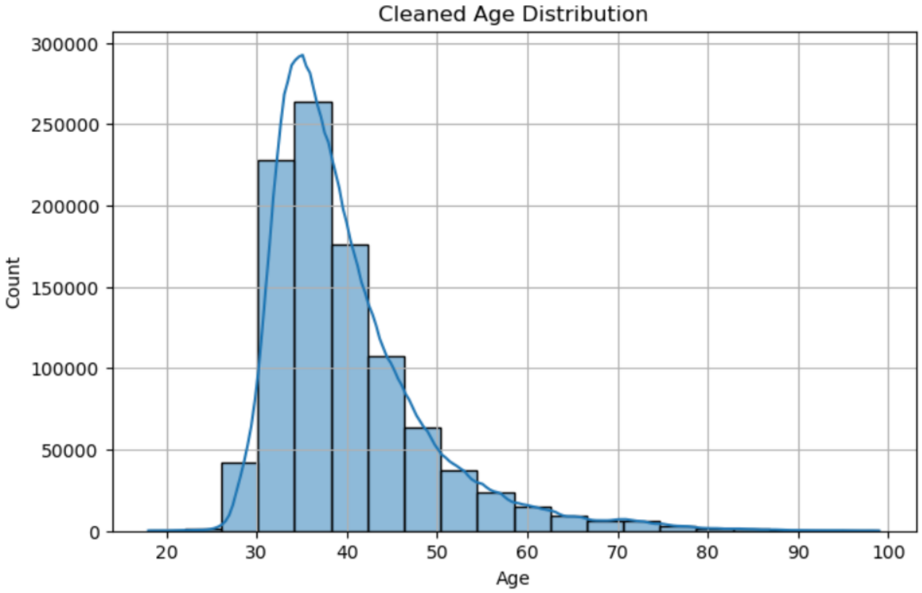
2.4 Data Collection and Cleaning

The dataset was imported in CSV format and cleaned using Python. Missing values in CustomerDOB, CustGender, CustLocation, and CustAccountBalance were identified. The CustomerDOB column was converted to a datetime format and used to calculate customer age.

Categorical columns were checked for consistency, and numeric columns were reviewed for outliers.

2.5 Exploratory Data Analysis (EDA)

2.5.1 Age Distribution



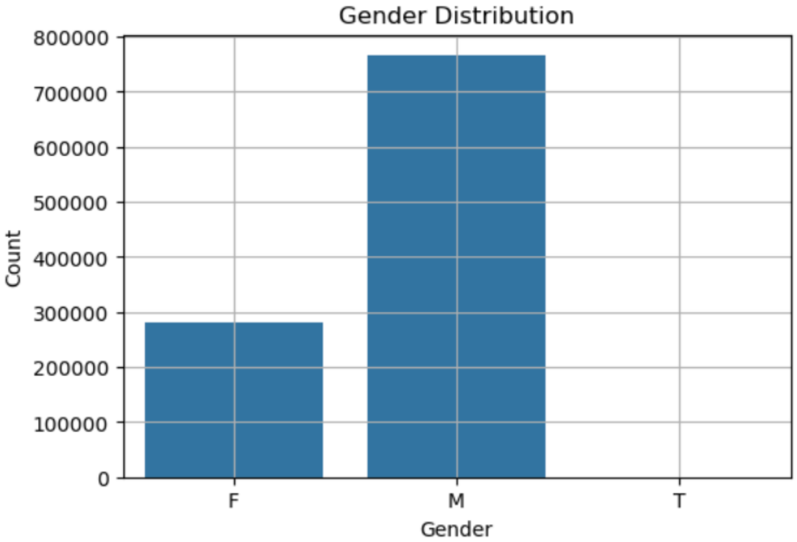
The histogram clearly shows a realistic distribution of customer ages:

- Most customers are between 30 and 40 years old, making this the predominant age segment.
- There's a notable peak around the mid-30s, suggesting this age group is highly active and engaged.
- Customers above 60 years old become less frequent, gradually decreasing towards the higher ages.

Insights:

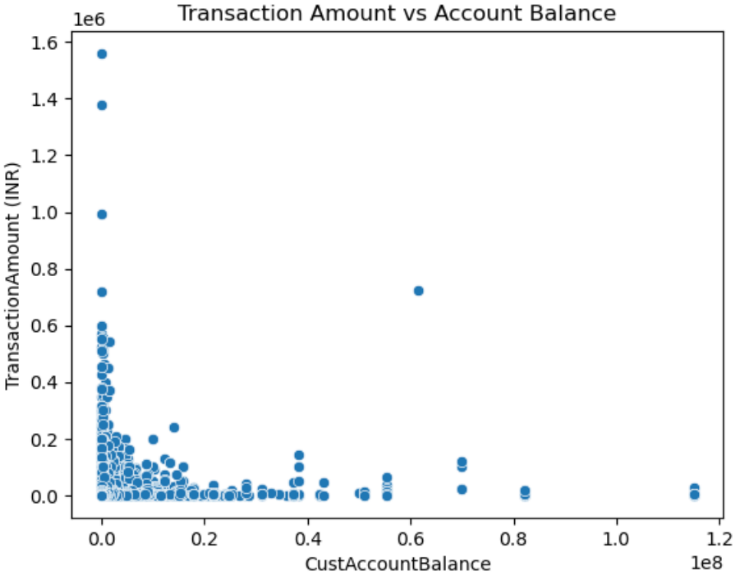
- The 30–40 age range could be an ideal demographic for targeted financial products, loans, or investments.
- Marketing and customer retention efforts should prioritize this dominant group.
- The corrected distribution demonstrates effective data cleaning, ensuring reliable analytics moving forward.

2.5.2 Gender Distribution



This bar chart shows that the majority of customers are **male (M)**, with over **750,000 records**, while **female (F)** customers make up a significantly smaller portion (around 280,000). There are also a few entries marked as **'T'** (possibly for transgender or invalid/missing values), but their count is almost negligible. This imbalance in gender distribution could affect modeling and should be considered when interpreting behavior patterns.

2.5.3 Transaction Amount vs. Account Balance

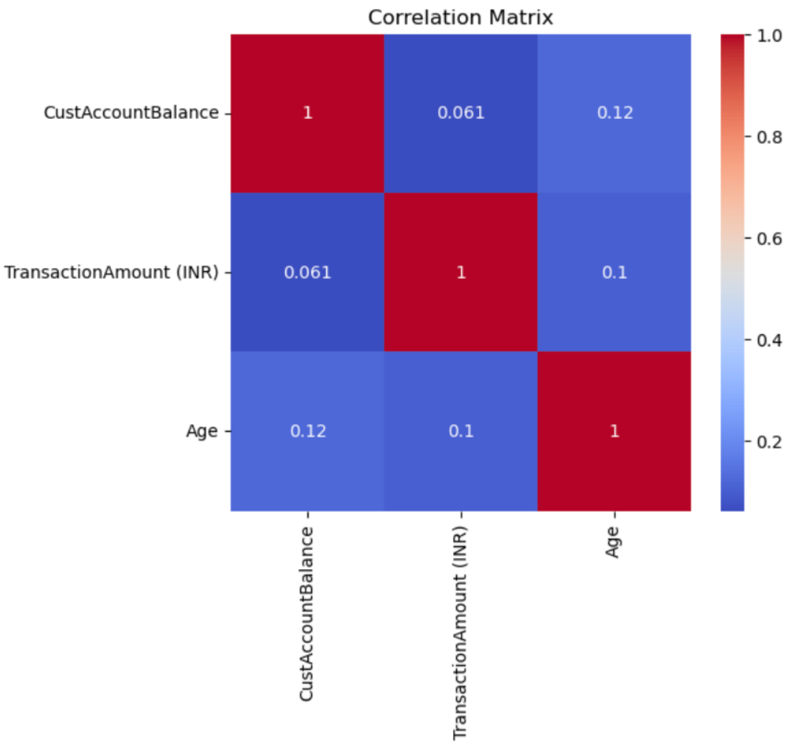


This scatter plot shows that **most transactions are small and occur from accounts with lower balances**. Surprisingly, several high-value transactions (above ₹1,000,000) are also made from accounts with **very low balances**, which may indicate overdrafts, business activity, or anomalies.

There is **no strong visible linear relationship** between account balance and transaction amount—customers with large balances don't necessarily spend more.

3. Analysis

3.1 Correlation Analysis



The correlation matrix shows relationships between **Customer Account Balance**, **Transaction Amount**, and **Age**:

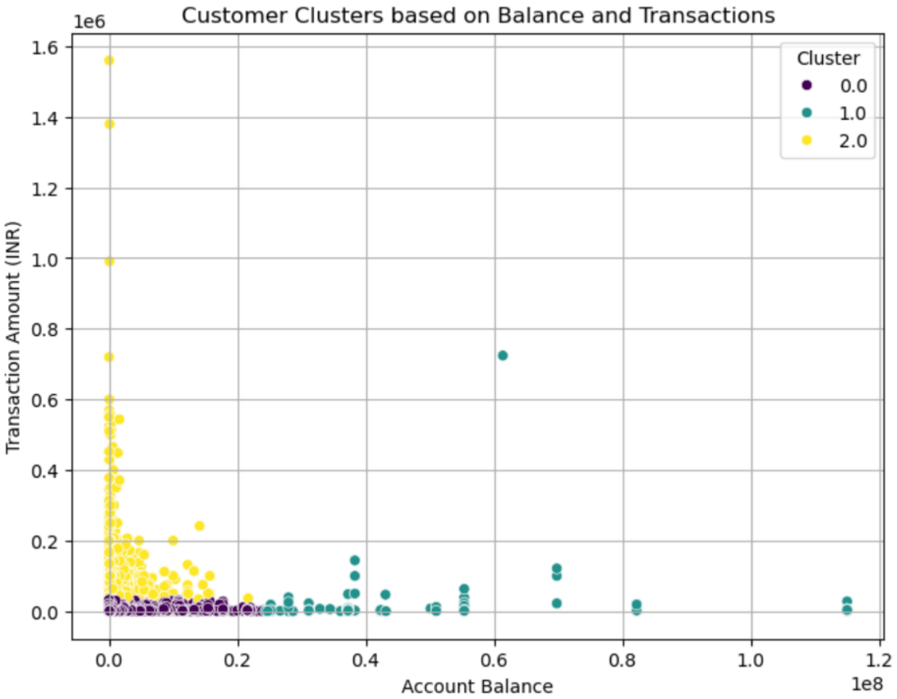
- **Customer Account Balance and Transaction Amount** have a **weak positive correlation (0.061)**, indicating customers with higher account balances slightly tend to make larger transactions, but the relationship isn't strong enough to be highly predictive.
- **Age and Customer Account Balance** show a slightly stronger but still modest positive correlation (**0.12**), suggesting older customers might maintain slightly higher account balances.
- **Age and Transaction Amount** exhibit a minimal positive correlation (**0.1**), indicating age has little influence on the transaction amount customers typically engage with.

Insights:

- The generally weak correlations suggest that customer spending and balance behaviors aren't heavily influenced by age alone, pointing towards the necessity of considering additional factors such as income, location, or lifestyle.

- Although these correlations are low, understanding even small relationships can guide more targeted marketing or personalized financial services for specific customer segments.

3.2 Clustering Analysis

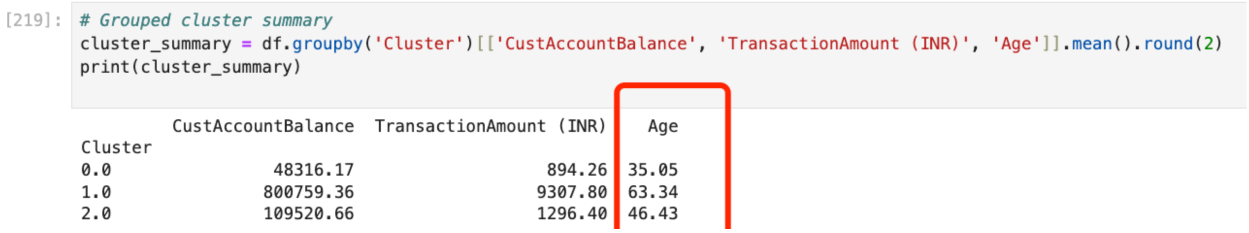


The clustering plot reveals three distinct groups of customers based on their account balance, transaction amount, and age:

- **Cluster 0 (Purple):** Represents the majority of customers who have moderate account balances and low transaction amounts.
- **Cluster 1 (Green):** Includes customers with significantly high account balances but relatively low transaction activity.
- **Cluster 2 (Yellow):** Contains customers who make very large transactions—even though many of them have low account balances.

These clusters suggest different customer behaviors and risk profiles. For example, customers in Cluster 2 might be flagged for potential overdraft or high-value transaction monitoring, while Cluster 1 customers may be prime targets for premium financial services.

3.3 Pattern Identification



```
[219]: # Grouped cluster summary
       cluster_summary = df.groupby('Cluster')[['CustAccountBalance', 'TransactionAmount (INR)', 'Age']].mean().round(2)
       print(cluster_summary)
```

```
         CustAccountBalance  TransactionAmount (INR)    Age
Cluster                                                    
0.0                48316.17                   894.26  35.05
1.0               800759.36                  9307.80  63.34
2.0               109520.66                  1296.40  46.43
```

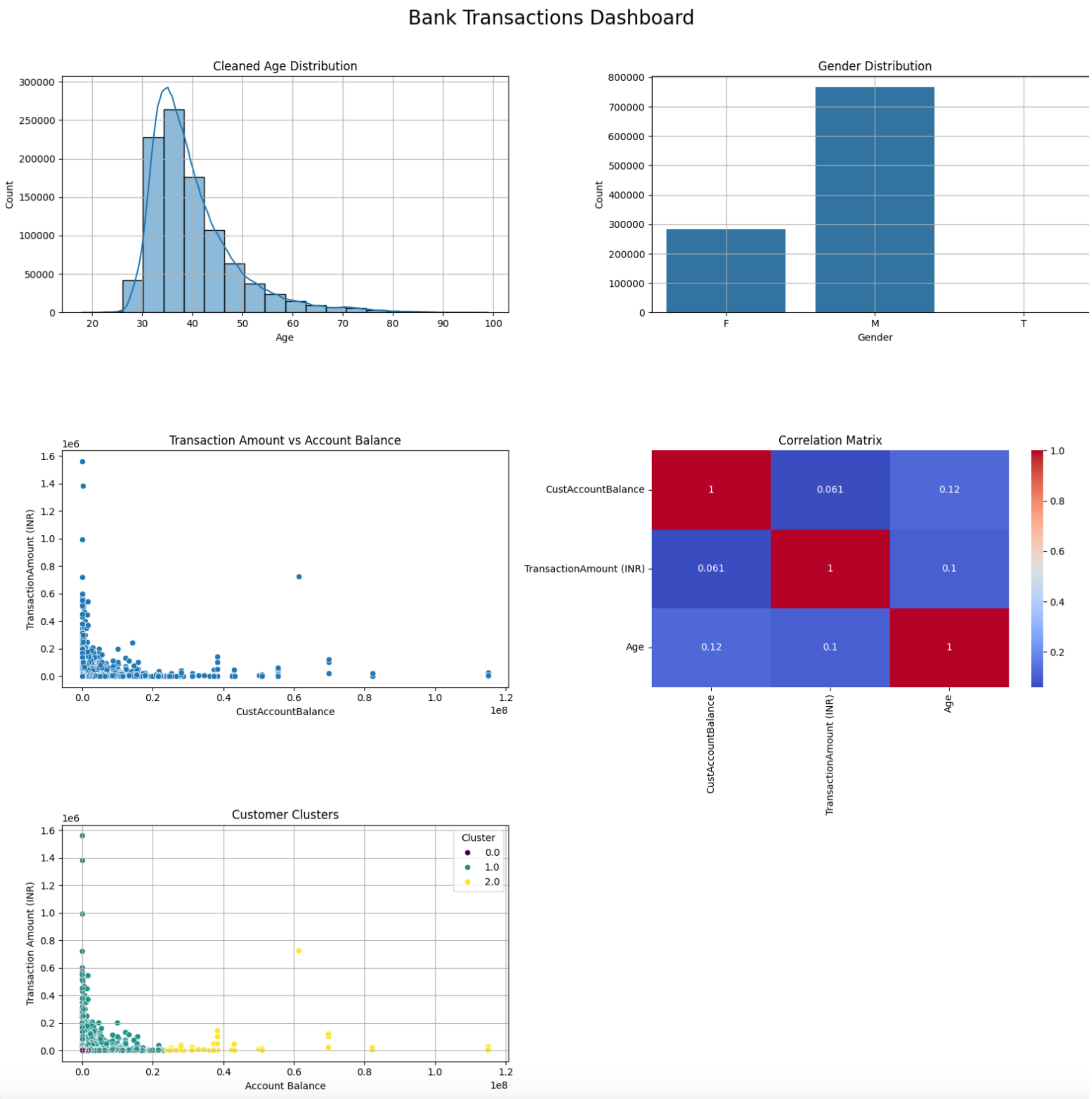
- **Cluster 0 (Young and Low Balance Group)**:
  - Customers here are typically younger (average age: 35).
  - They maintain lower account balances and carry out smaller transactions.
  - **Interpretation**: This segment is ideal for basic banking services, introductory financial products, or digital banking solutions tailored to younger users.
- **Cluster 1 (Senior High-Net-Worth Group)**:
  - These customers are older (average age: 63) with significantly higher account balances and larger transactions.
  - **Interpretation**: This group represents wealthier senior customers who may prefer premium or specialized financial services, investment management, or personalized financial advice.
- **Cluster 2 (Middle-aged Moderate Group)**:
  - Average age of customers is about 46, with moderate balances and transaction sizes.
  - **Interpretation**: This segment likely includes customers in their prime earning years who might benefit from tailored financial products like loans, mortgages, or retirement planning services.

This clustering analysis clearly segments customers by their financial behaviors and age demographics, enabling targeted, customized marketing strategies and more effective resource allocation.

4. Key Insights

1. **Customer Age Clusters:** Age is a strong segmentation factor. Different age groups exhibit different balance and transaction behaviors, which can inform personalized marketing or product offers.
2. **High Transactions with Low Balances:** Many large transactions occur from accounts with low balances, highlighting a potential risk area or a unique behavior worth targeting for credit product offerings.
3. **Urban Concentration:** Most active customers are concentrated in metro areas like Mumbai and Navi Mumbai, suggesting that regional targeting in these cities may yield high returns.

5. Dashboard Design



6. Conclusion

This analysis explored patterns and relationships within a large-scale bank customer transaction dataset. Using Python, we performed data cleaning, visualization, correlation assessment, and clustering. The insights generated can support customer segmentation strategies and product targeting. The dashboard will further visualize these findings for stakeholders.