A Micro Project report on

Banking system data analysis using R

Submitted to the CMR Institute of Technology, Hyderabad in partial fulfillment of the requirement for the award of the Laboratory of

Data Mining and Data Analytics

of

III-B.Tech. I-Semester

in

Computer Science and Engineering

Submitted by

A. Siri	**23R01A05CT**
A. Niranjan	**23R01A05CU**
S. Hasini Reddy	**23R01A05Y0**

Under the Guidance of

Mr. Amireddy Srinish Reddy

(Asst. Professor Dept of CSE)

CMR INSTITUTE OF TECHNOLOGY

(UGC AUTONOMUS)

(Approved by AICTE, Affiliated to JNTU, Kukatpally, Hyderabad)

Kandlakoya, Medchal Road, Hyderabad

2025-2026

CMR INSTITUTE OF TECHNOLOGY

(UGC AUTONOMUS)
(Approved by AICTE, Affiliated to JNTU, Kukatpally, Hyderabad)

Kandlakoya, Medchal Road, Hyderabad.

Department of CSE

CERTIFICATE

This is to certify that a Micro Project entitled with:

Banking system data analysis using R

is being submitted by

A. Siri	**23R01A05CT**
A. Niranjan	**23R01A05CU**
S. Hasini Reddy	**23R01A05Y0**

In partial fulfillment of the requirement for award of the **Data Mining and Data Analytics** of III-B. Tech I- Semester in CSE towards a record of a Bonafide work carried out under our guidance and supervision.

Signature of Guide

Mr. Amireddy Srinish Reddy

Signature of HOD

Prof. Dr. K. Pradeep Reddy

ACKNOWLEDGEMENT

We are extremely grateful to **Dr. M. Janga Reddy, Director, Dr. B.Satyanarayana, Principal and Prof. Dr. K. Pradeep Reddy, Head of Department,** Dept of COMPUTER SCIENCE AND ENGINEERING, CMR Institute of Technology for their inspiration and valuable guidance during entire duration.

We are extremely thankful to Data Mining and Data Analytics faculty in-charge **Mr.Amireddy Srinish Reddy**, Computer Science and Engineering department, CMR Institute of Technology for his constant guidance, encouragement and moral support throughout the project.

We express our thanks to all staff members and friends for all the help and coordination extended in bringing out this Project successfully in time. Finally, we are very much thankful to our parents and relatives who guided directly or indirectly for successful completion of the project.

A. Siri	**23R01A05CT**
A. Niranjan	**23R01A05CU**
S. Hasini Reddy	**23R01A05Y0**

CONTENTS

ABSTRACT

The Banking System Data Analysis project focuses on exploring and interpreting customer transaction behavior using a structured dataset of 30 customers, including details such as age, account number, transaction date, type, amount, and account balance. The primary objective is to uncover patterns in deposits and withdrawals, identify high-value customers, and understand trends across different age groups.

Through comprehensive analysis using R programming, various visualizations such as bar charts, histograms, pie charts, scatterplots, line charts, and boxplots were generated. These visualizations highlight key insights, including the distribution of transaction types, variations in transaction amounts, trends in account balances over time, and differences in behavior across age groups. High-value customers and outliers were identified, providing actionable insights for targeted marketing, personalized banking services, and monitoring unusual transactions for fraud prevention.

Overall, this project demonstrates how data analysis and visualization can support strategic decision-making in the banking sector, helping banks understand customer behavior, optimize transaction processes, and design customer-focused financial products.

INTRODUCTION

The banking sector is a cornerstone of any economy, playing a crucial role in managing financial transactions, savings, investments, and customer accounts. In today's digital era, banks collect vast amounts of transactional data daily. Analyzing this data effectively enables banks to understand customer behavior, optimize services, detect unusual patterns, and make informed decisions. Data analysis in banking not only improves operational efficiency but also supports strategies for customer retention, risk management, and fraud detection.

This project focuses on analyzing a dataset of 30 customers, including details such as age, account number, transaction type, transaction amount, transaction date, and account balance. The aim is to explore transaction patterns, identify high-value customers, and study how transaction behavior varies across different age groups. By leveraging **R programming**, the project applies data visualization techniques such as bar charts, line charts, pie charts, histograms, scatterplots, and boxplots. These visualizations provide a clear understanding of how deposits and withdrawals are distributed and highlight trends in customer transactions over time.

The insights gained from this analysis can guide banks in developing targeted financial products and personalized services. For instance, identifying customers with consistently high balances or frequent high-value transactions can help banks tailor premium offerings or investment plans. Additionally, understanding withdrawal patterns and outliers can enhance fraud detection systems. Overall, this project demonstrates the power of data analysis in transforming raw banking data into actionable insights that improve both customer satisfaction and operational performance.

PROCEDURE

The procedure for conducting the banking system data analysis involves several systematic steps to ensure accurate and actionable insights are derived from the dataset. Each step focuses on understanding customer behavior, transaction patterns, and trends in account balances. The process begins with the collection of data, followed by cleaning, analysis, and visualization to derive meaningful results.

The first step in the procedure is data collection. A structured dataset containing customer details, account numbers, transaction dates, transaction types, transaction amounts, and account balances was compiled for analysis. For this project, a dataset of 30 customers was used, providing a representative sample. Ensuring the accuracy and completeness of the data at this stage is crucial, as it forms the foundation for all subsequent analysis.

Once the data is collected, it undergoes cleaning and preparation. This step involves checking for missing values, correcting inconsistent entries, and converting columns to appropriate data types, such as dates and categorical factors. Additional columns, such as age groups, are created to facilitate analysis. Proper preparation of the data ensures that the results of the analysis are reliable and valid.

After data preparation, exploratory data analysis (EDA) is conducted to gain an initial understanding of the dataset. Summary statistics such as the mean, maximum, minimum, and total transaction amounts are calculated. The structure and distribution of the data are examined to identify patterns, trends, and potential outliers. This step helps in forming a foundation for deeper analysis and ensures that important patterns are not overlooked.

The next step is data visualization, which provides a clear and intuitive way to interpret patterns, trends, and anomalies in the dataset. Various types of visualizations were created using R programming and the ggplot2 package. Bar charts were used to display transaction counts and total or average transaction amounts by customer, age group, or transaction type. Scatterplots were created to represent individual transactions for each customer, highlighting variations and outliers. Line charts were used to track trends in account balances and transaction amounts over time, while histograms illustrated the distribution of transaction amounts. Boxplots highlighted the spread and variability of transaction amounts, making it easier to identify unusual transactions, and pie charts were used to show the proportion of transaction types and to highlight top contributing customers.

After the visualizations were generated, the analysis and interpretation of the data were carried out. This involved identifying high-value customers, understanding age groups with larger account balances, examining transaction patterns, and detecting unusual behaviors. Correlations between transaction amounts and account balances were also explored to gain deeper insights into customer financial habits.

Finally, the findings were summarized and reported in a structured format. Each visualization was explained, key trends were highlighted, and recommendations were made for banking strategies such as targeted services, fraud monitoring, and improving transaction efficiency. Both quantitative analysis and qualitative interpretation were used to provide a comprehensive view of customer behavior, ensuring that the insights derived from the project are actionable and valuable for banking decision-making.

REQUIREMENTS AND SPECIFICATIONS

The successful execution of the Banking System Data Analysis project requires both **software and hardware resources**, as well as clearly defined specifications to ensure accurate and efficient analysis. The project is implemented using **R programming**, which is a powerful language for statistical computing and data visualization. The primary libraries used include **ggplot2** for creating visualizations and **dplyr** for data manipulation and preprocessing. RStudio is recommended as the integrated development environment (IDE) for executing scripts, managing datasets, and generating visual outputs.

The dataset forms a critical requirement for the project. It must include detailed information about customers, such as **Customer Name, Age, Account Number, Transaction Date, Transaction Type, Transaction Amount, and Account Balance**. The data should be clean, accurate, and in a structured format like CSV to facilitate easy loading and manipulation in R. Ensuring completeness and consistency in the dataset is essential to prevent errors during analysis and visualization.

From a hardware perspective, the project can be executed on a standard personal computer or laptop with at least 4 GB of RAM and sufficient storage to handle datasets and generated plots. An active internet connection is useful for installing required R packages and updating libraries when needed. For advanced visualization or handling larger datasets, higher processing power and memory may be beneficial but are not mandatory for the current scope.

The specifications of the analysis include several key objectives. First, the system should be able to read and process the dataset efficiently. It must handle missing values and correctly identify categorical and numerical data. The system should generate a variety of visualizations including **bar charts, scatterplots, line charts, histograms, boxplots, and pie charts**. These visualizations should be clearly labeled with appropriate titles, axis names, and legends to ensure interpretability. Finally, the system should allow for the **analysis of customer behavior**, identification of high-value customers, detection of outliers, and trends in transaction patterns to provide actionable insights for banking operations and strategy development.

IMPLEMENTATION

bank_dataset_30.csv

Customer_Name,Age,Account_Number,Transaction_Date,Transaction_Type,Transaction_Amount,Account_Balance

Amit,25,2001,2025-09-01,Deposit,15000,45000

Riya,32,2002,2025-09-02,Withdrawal,8000,22000

Karan,45,2003,2025-09-03,Deposit,30000,120000

Priya,28,2004,2025-09-04,Deposit,12000,48000

Neha,39,2005,2025-09-05,Withdrawal,5000,65000

Rohit,50,2006,2025-09-06,Deposit,40000,150000

Sonal,42,2007,2025-09-07,Withdrawal,12000,58000

Arjun,31,2008,2025-09-08,Deposit,28000,87000

Deepak,29,2009,2025-09-09,Withdrawal,7000,33000

Seema,36,2010,2025-09-10,Deposit,18000,76000

Vijay,55,2011,2025-09-11,Deposit,50000,200000

Asha,48,2012,2025-09-12,Withdrawal,15000,90000

Kavya,23,2013,2025-09-13,Deposit,9000,25000

Manish,60,2014,2025-09-14,Deposit,35000,140000

Renu,34,2015,2025-09-15,Withdrawal,4000,41000

Suresh,27,2016,2025-09-16,Deposit,11000,37000

Isha,41,2017,2025-09-17,Withdrawal,6000,72000

Harish,46,2018,2025-09-18,Deposit,22000,98000

Meera,30,2019,2025-09-19,Deposit,8000,44000

Gaurav,52,2020,2025-09-20,Withdrawal,20000,120000

Priyanka,38,2021,2025-09-21,Deposit,16000,68000

Sunil,61,2022,2025-09-22,Deposit,45000,230000

Alpa,26,2023,2025-09-23,Withdrawal,3000,27000

Rajat,33,2024,2025-09-24,Deposit,14000,59000

Nisha,44,2025,2025-09-25,Withdrawal,9000,81000

Tarun,49,2026,2025-09-26,Deposit,26000,115000

Bhavna,21,2027,2025-09-27,Deposit,6000,20000

Kiran,56,2028,2025-09-28,Withdrawal,25000,90000

Veda,37,2029,2025-09-29,Deposit,20000,72000

Omar,64,2030,2025-09-30,Withdrawal,18000,150000

CODE:

```
# Set working directory and confirm

setwd("C:/Users/cyber/Desktop/Dmda")

getwd()


# Install required libraries (only the first time)

install.packages("ggplot2")

install.packages("dplyr")


# Load libraries

library(ggplot2)

library(dplyr)


cat("Libraries loaded successfully ■ \n")


# Load the dataset

bank_data <- read.csv("bank_dataset_30.csv")


# Display first few rows to check data

head(bank_data)


# Check the structure of the dataset

str(bank_data)
```

 # Check for missing values

```
colSums(is.na(bank_data))
```

 # Summary statistics of dataset

```
summary(bank_data)
```

 # Quick metrics

```
cat("Total Transactions:", nrow(bank_data), "\n")
cat("Average Transaction Amount:", mean(bank_data$Transaction_Amount), "\n")
cat("Max Account Balance:", max(bank_data$Account_Balance), "\n")
cat("Min Account Balance:", min(bank_data$Account_Balance), "\n")
```

DATA VISUALIZATION

 # Bar chart: Transaction Amount by Customer and Type

```
ggplot(bank_data, aes(x=Customer_Name, y=Transaction_Amount, fill=Transaction_Type)) +
geom_bar(stat="identity") +
theme_minimal() +labs(title="Transaction Amount by Customer and Type", x="Customer", y="Transaction Amount")
```

 # Scatter plot: Transaction Amount by Customer

```
ggplot(bank_data,aes(x=Customer_Name, y=Transaction_Amount, color=Transaction_Type)) +geom_point(size=3) +theme_minimal() +labs(title="Transaction Amount by Customer", x="Customer", y="Transaction Amount") +theme(axis.text.x=element_text(angle=45, hjust=1))
```

 # Boxplot: Transaction Amount Distribution by Type

```
ggplot(bank_data, aes(x=Transaction_Type, y=Transaction_Amount, fill=Transaction_Type)) +geom_boxplot() +theme_minimal() +labs(title="Transaction Amount Distribution by Type", x="Transaction Type", y="Transaction Amount")
```

 # Line chart: Account Balance Trend Over Time

```
ggplot(bank_data, aes(x=Transaction_Date, y=Account_Balance, group=1)) +
geom_line(color="blue", size=1) +geom_point(color="darkblue", size=2) +
theme_minimal() +labs(title="Account Balance Trend Over Time", x="Transaction Date", y="Account Balance")
```

 # Line chart: Transaction Amount Trend Over Time by Type

```
ggplot(bank_data,aes(x=Transaction_Date, y=Transaction_Amount, color=Transaction_Type, group=Transaction_Type)) +geom_line(size=1) +geom_point(size=2) +theme_minimal() +
labs(title="Transaction Amount Trend Over Time", x="Transaction Date", y="Transaction Amount")
```

 # Histogram: Distribution of Transaction Amounts

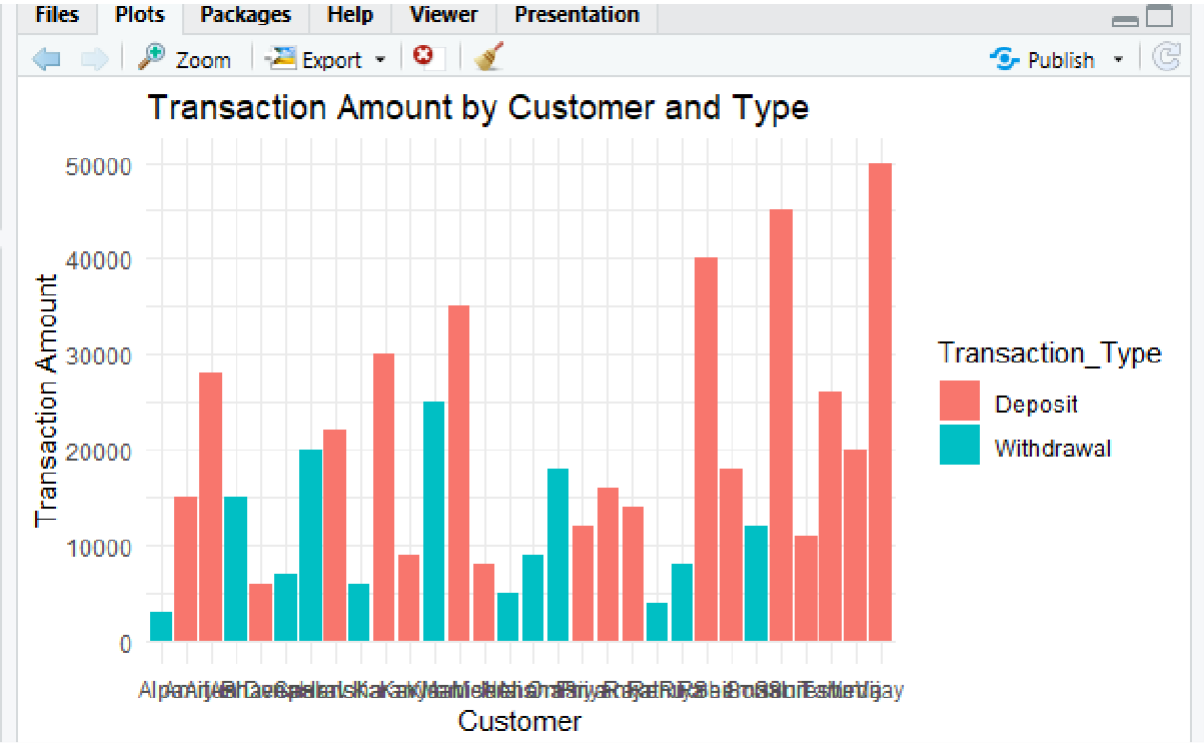
```
ggplot(bank_data, aes(x=Transaction_Amount, fill=Transaction_Type)) +
geom_histogram(binwidth=5000, color="black", alpha=0.7) +theme_minimal() +
labs(title="Distribution of Transaction Amounts", x="Transaction Amount", y="Frequency") +
scale_fill_manual(values=c("Deposit"="steelblue", "Withdrawal"="salmon"))
```

 # Pie chart: Proportion of Transaction Types

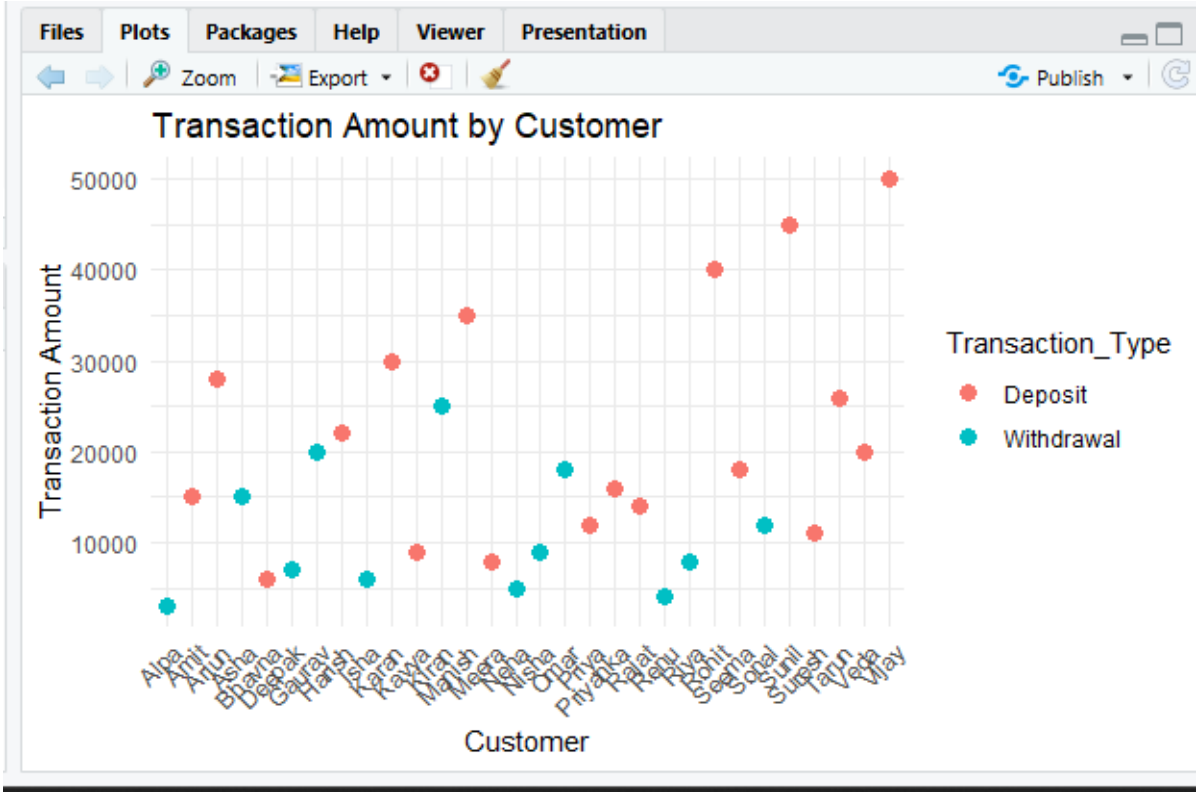
```
transaction_counts <- table(bank_data$Transaction_Type)
pie(transaction_counts, labels=paste(names(transaction_counts), "\n", transaction_counts),
main="Proportion of Transaction Types", col=c("skyblue","salmon"))
```

RESULTS

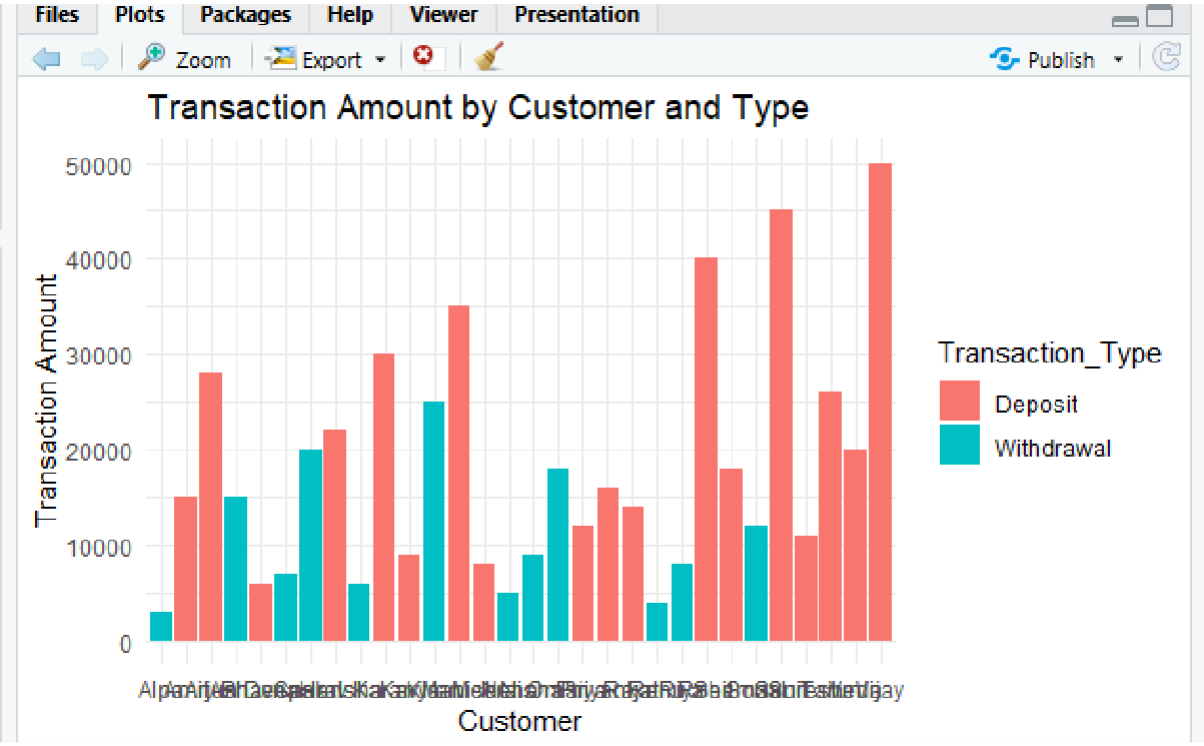
1. BAR CHART:



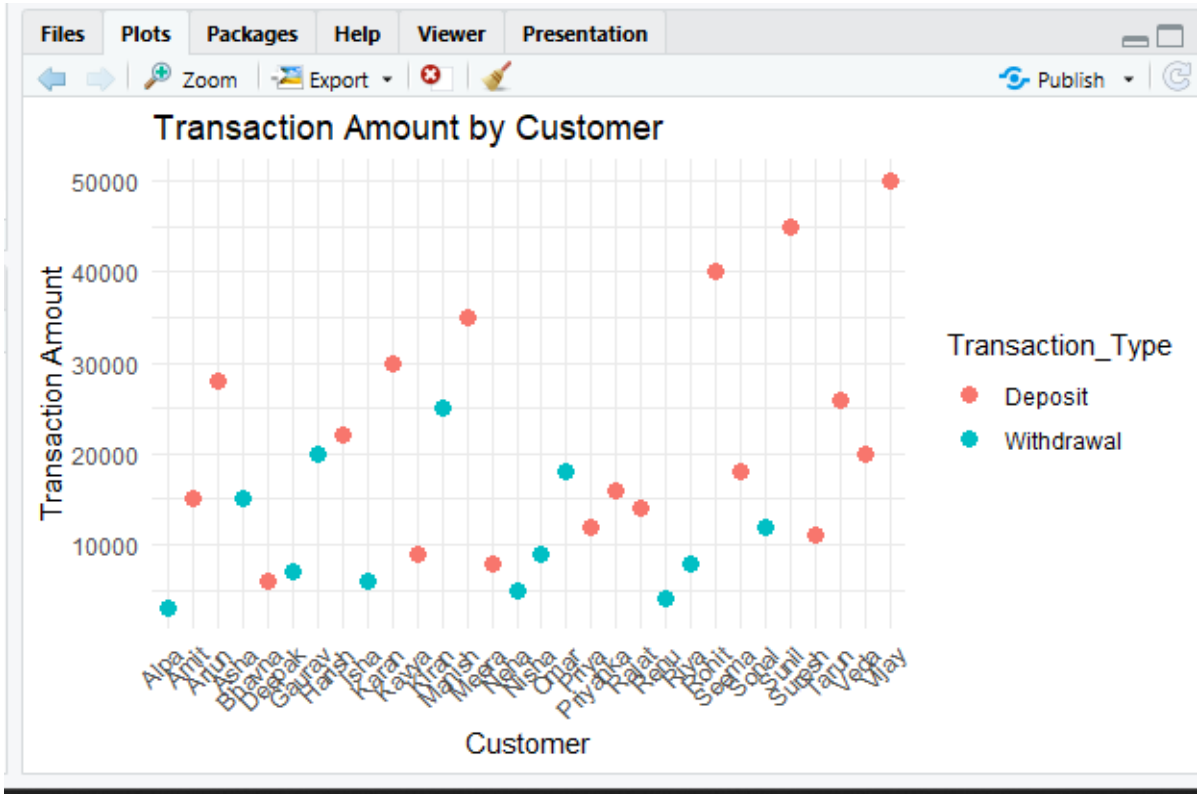
2. SCATTERPLOT:



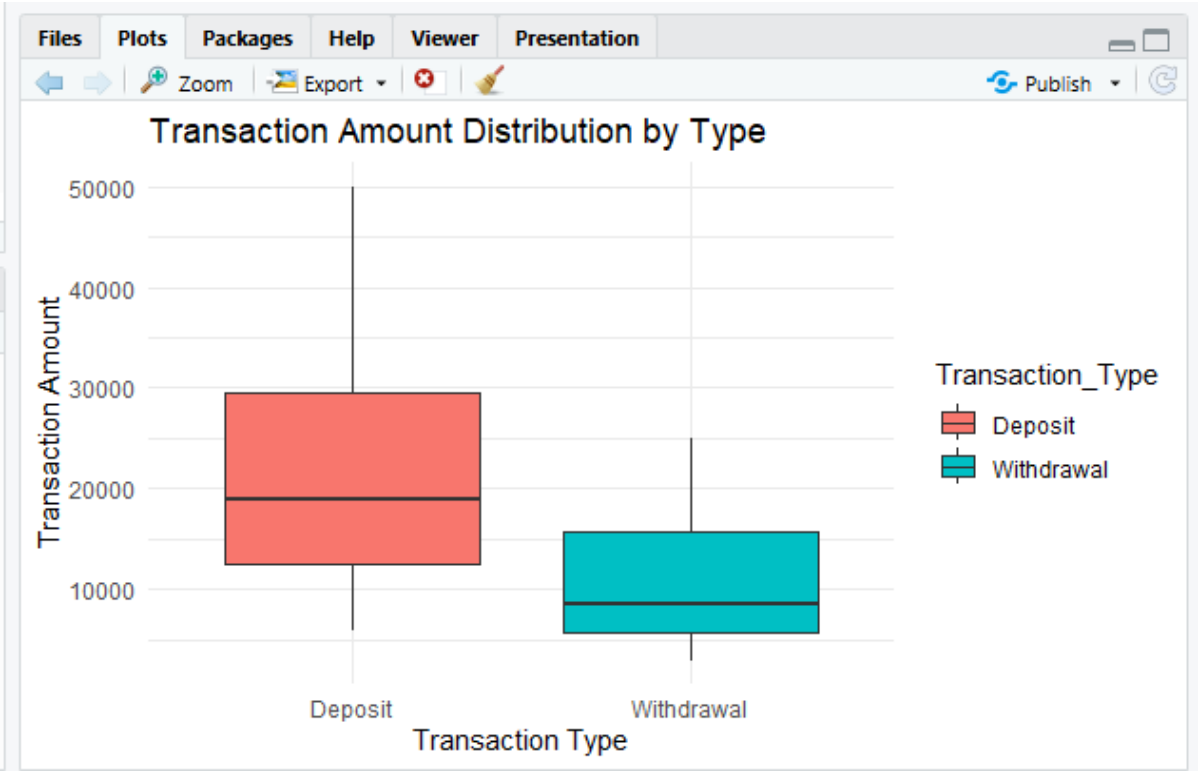
3. BAR CHART:



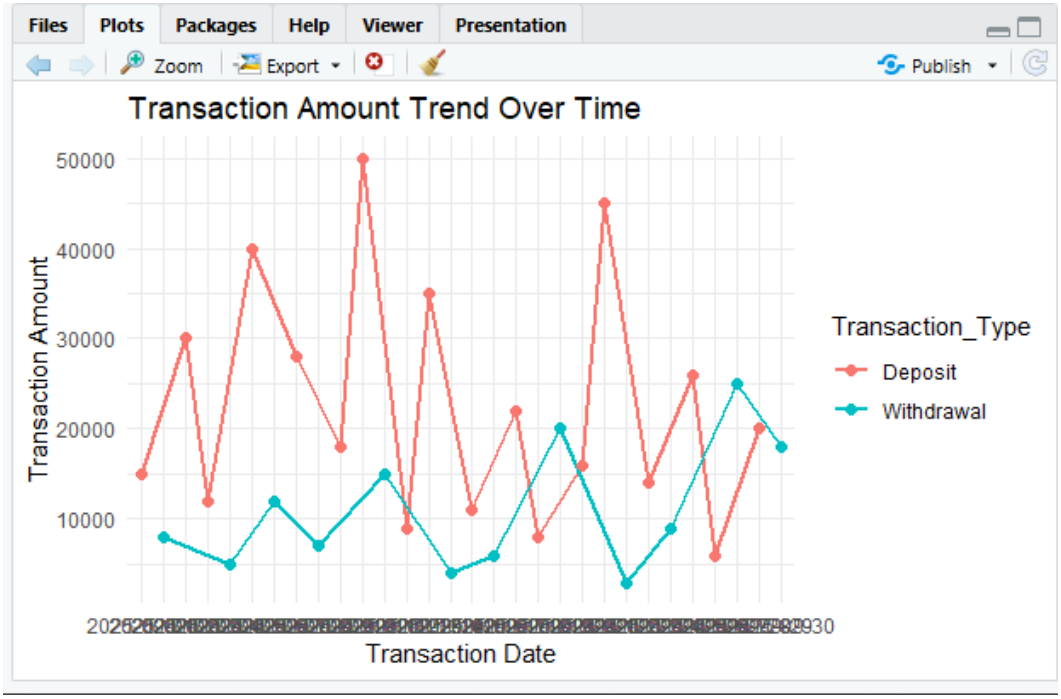
4. SCATTERPLOT:



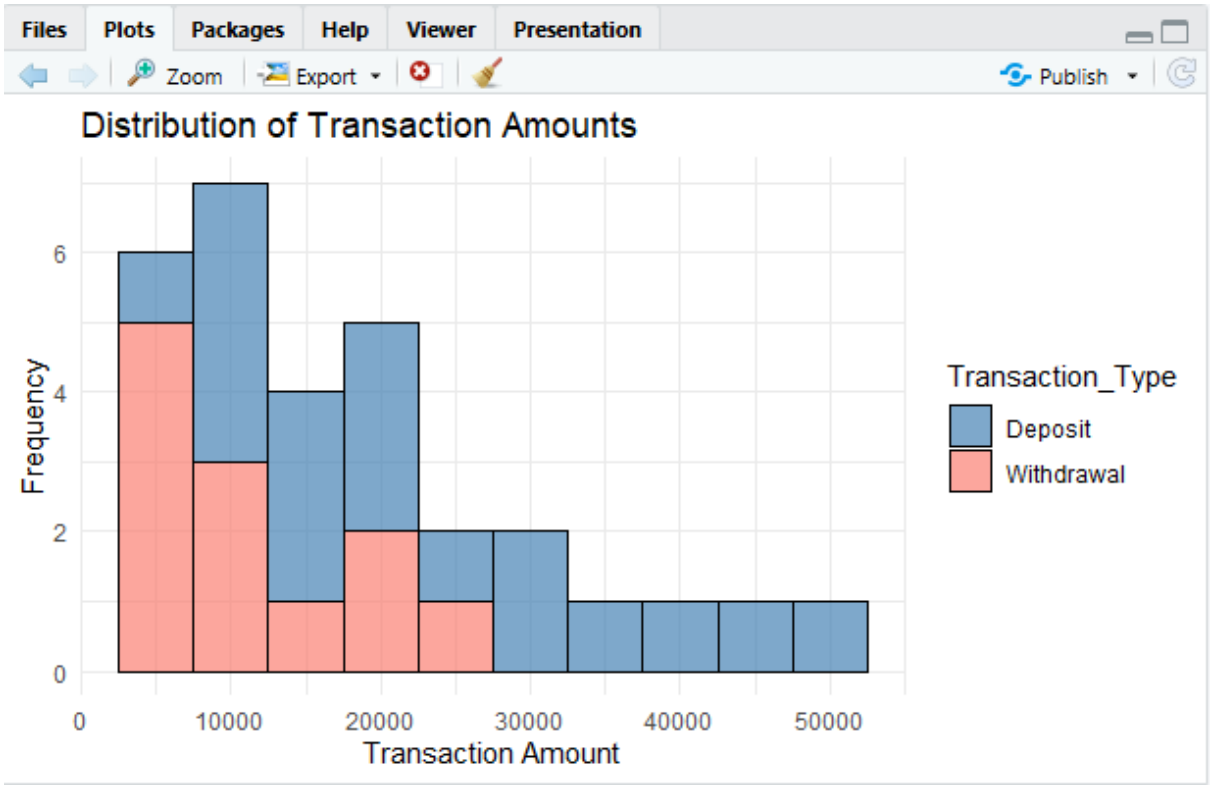
5. BOX PLOT:



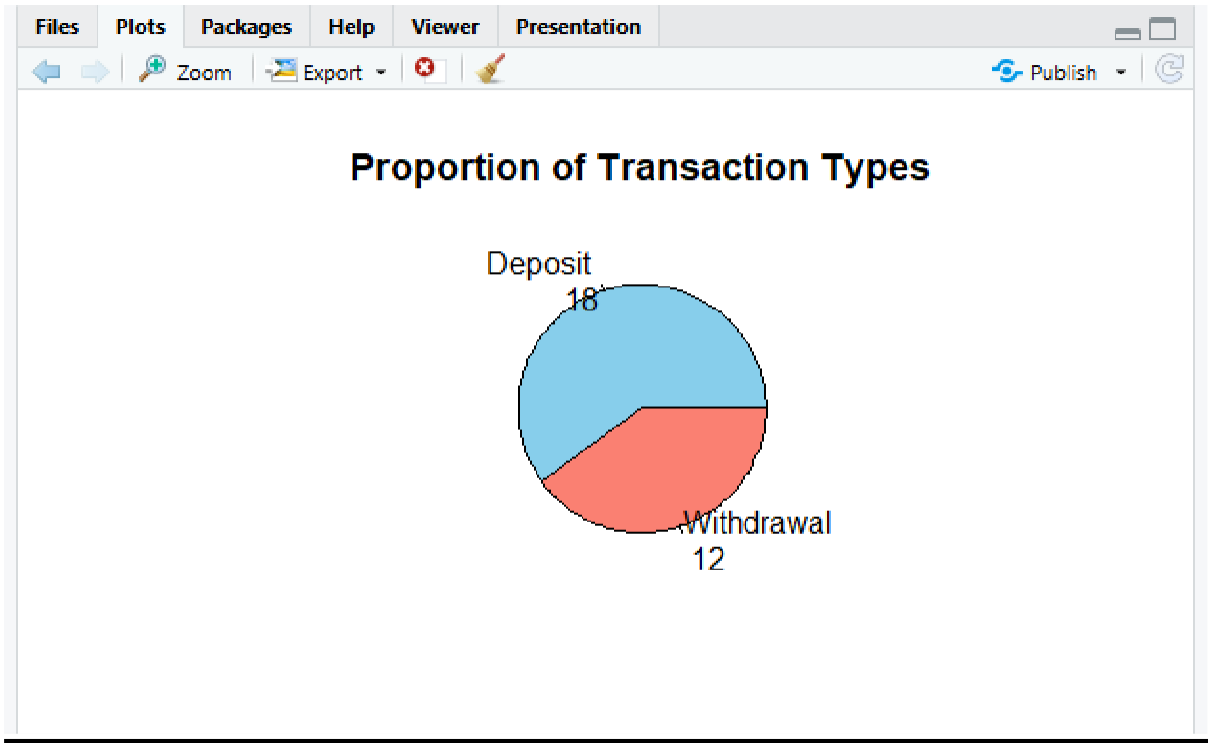
6. LINE CHART:



7. HISTOGRAM



8. PIE CHART



REPORT

The analysis of the banking dataset reveals clear patterns in customer transactions and account balances. High-value customers, such as Vijay, Sunil, and Rohit, consistently perform large deposits, indicating strong financial engagement and potential for premium banking services. In contrast, younger or low-balance customers generally conduct smaller, frequent transactions, representing typical day-to-day banking activity. Visualizations such as bar charts and scatter plots helped identify these variations, highlighting both transaction amounts and individual customer behavior.

Transaction type analysis shows that deposits slightly outnumber withdrawals, which is consistent with normal account maintenance behavior. Boxplots and histograms reveal that deposits typically have higher median amounts and a wider range compared to withdrawals, while certain large withdrawals act as outliers that could warrant further attention for fraud monitoring or unusual financial activity. Line charts tracking transaction amounts and account balances over time further show trends, including peaks corresponding to salary or bill cycles, as well as consistent high balances for financially stable customers.

Overall, the insights gained from this analysis can inform banking strategies such as targeting high-value customers with personalized services, designing age-specific financial products, monitoring unusual transactions for fraud detection, and optimizing operational efficiency. By combining quantitative data with visual interpretation, the project demonstrates how structured banking data can be transformed into actionable insights, helping banks enhance customer engagement, improve transaction management, and make data-driven decisions.

CONCLUSION

The Banking System Data Analysis project provides a comprehensive understanding of customer behavior, transaction patterns, and account balance trends using a structured dataset of 30 customers. By employing R programming and visualization tools such as bar charts, scatterplots, line charts, histograms, boxplots, and pie charts, the analysis successfully identified high-value customers, frequent transaction patterns, and unusual financial activities. These insights enable banks to segment customers effectively, target specific groups with personalized services, and monitor transactions for potential fraud.

The visualizations also revealed trends over time, such as fluctuations in account balances and spikes in transaction amounts, which can be linked to salary cycles or significant financial events. This temporal analysis helps banks anticipate customer needs and plan operations accordingly, ensuring efficient resource allocation and better customer satisfaction. The project demonstrates that combining quantitative metrics with visual interpretation provides a clearer and more actionable understanding of financial data.

Overall, this analysis highlights the importance of data-driven decision-making in the banking sector. By leveraging transaction data, banks can enhance operational efficiency, improve customer engagement, design targeted financial products, and maintain robust monitoring mechanisms. The project underscores how structured data analysis transforms raw banking data into meaningful insights that support strategic planning, risk management, and improved customer experience.

FUTURE SCOPE

The Banking System Data Analysis project lays the foundation for more advanced and large-scale analysis in the future. With a larger dataset covering multiple branches, longer time periods, and additional customer demographics, the analysis can be extended to identify more comprehensive trends and patterns. Incorporating machine learning techniques could allow predictive modeling, such as forecasting customer account balances, predicting high-value transactions, or detecting potential fraudulent activity before it occurs.

In addition, integrating real-time banking data and automating analysis using dashboards can provide dynamic insights to bank managers and decision-makers. This would allow instant tracking of transactions, customer activity, and financial trends, enabling timely interventions and personalized recommendations. Advanced analytics could also explore correlations between customer behavior, transaction frequency, and financial products usage to design more effective marketing and service strategies.

Furthermore, the project can be expanded to include sentiment analysis and customer feedback, linking transactional behavior with satisfaction levels. This holistic approach would provide banks with a deeper understanding of customer needs, improving engagement and loyalty. Overall, the future scope of this project includes predictive analytics, real-time monitoring, personalized service strategies, and integration with broader financial systems, making banking operations more data-driven and customer-centric.

REFERENCES

1. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

2. Grolemund, G., & Wickham, H. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.

3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

4. R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available at: https://www.r-project.org

5. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd Edition). Morgan Kaufmann.

6. Indian Banks Association. (2020). *Banking Trends and Customer Behavior Analysis in India*. Available at: https://www.iba.org.in

7. TutorialsPoint. (2025). *R Programming Basics and Data Visualization*. Available at: https://www.tutorialspoint.com/r/index.htm

8. Analytics India Magazine. (2021). *Data Analytics in Indian Banking Sector*. Available at: https://analyticsindiamag.com