Bank Customer Loyalty and Segmentation Analysis Using Hadoop MapReduce, Hive, and R

The project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering School of

Engineering and Sciences

Submitted by

Namratha Addagada - AP21110010048

Under the Guidance of

Dr. Rajiv Senapati

SRM University–AP

Neerukonda,Mangalagiri,Guntur

Andhra Pradesh – 522 240

[November, 2024]

Certificate

Date: 8/23/2024

This is to certify that the work present in this Project entitled "**Bank Customer Loyalty and Segmentation Analysis Using Hadoop MapReduce, Hive, and R**" has been carried out by **Sai Gopal Lanka, Namratha Addagada, and Haasitha Ambati** under my/our supervision. The Work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in the **School of Engineering and Sciences**.

Supervisor

(Signature)

Prof./Dr.[Name]

Designation,

Affiliation.

Co-supervisor

(Signature)

Prof./Dr.[Name]

Designation,

Affiliation.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof./Dr. Rajiv Senapati, for their continuous support, guidance, and encouragement throughout the project. I also thank my peers and family for their invaluable assistance.

Table of Contents

Abstract

This project comprehensively analyses customer behaviour, loyalty, and spending trends using big data analytics within the banking sector. Leveraging tools such as Hadoop, MapReduce, and Hive, we processed extensive datasets to uncover patterns in customer loyalty, spending behaviours, and cross-sell opportunities. Our analysis spans multiple segments, including Churn Prediction, Customer Segmentation, and Cross-Sell/Up-Sell Analysis, each aimed at understanding various dimensions of customer engagement. Through visualisations such as radial spending charts, sentiment heatmaps, and segmentation comparisons, we provide actionable insights that enable targeted recommendations and personalised offerings.

In addition to customer segmentation, this project examines monthly and quarterly spending trends, loyalty scores, and sentiment analysis to identify high-value customers and at-risk accounts. Using Hive for complex querying, we improved data accessibility, allowing deeper insights into customer behaviour and enhanced analytics capabilities. This project thus serves as a blueprint for improving customer experience and optimising engagement through big data analytics, demonstrating the transformative potential of Hadoop and Hive in the banking industry.

List of Figures

Introduction

In today's data-driven world, organizations in the banking and financial services sector are increasingly leveraging big data analytics to gain insights into customer behaviour and enhance decision-making processes. Customer loyalty, segmentation, and spending patterns hold critical value, enabling banks to offer targeted services, predict churn, and foster deeper engagement with high-value clients.

This project harnesses the power of big data technologies, specifically the Hadoop ecosystem, MapReduce, and Hive, to perform a detailed analysis of customer behaviour in the banking industry. By analysing extensive customer data, we identify patterns and trends that inform customer relationship strategies and drive more effective marketing and retention campaigns.

The scope of the project includes several key analyses: Churn Prediction, Customer Segmentation, Loyalty Scoring, and Cross-Sell/Up-Sell recommendations. Using MapReduce programming in Java and Hive for complex data queries, we processed and analysed datasets to understand customer loyalty levels, spending habits, and sentiment trends. Additionally, we developed various visualisations to present the findings in a clear and actionable format, helping stakeholders quickly grasp critical insights. By integrating Hive for streamlined data querying, this project exemplifies an advanced, integrated approach to customer analytics, demonstrating the significant impact of big data on customer engagement and satisfaction in the banking sector.

Methodology

The methodology for this project involves several key phases, each utilizing big data technologies and programming to extract, process, and analyze customer data within the banking sector. This systematic approach enabled the identification of meaningful patterns and insights regarding customer loyalty, spending behaviour, and cross-sell/up-sell opportunities.

- **Data Collection and Storage**:
  - Large datasets containing customer transaction details, spending patterns, and demographic information were collected and stored in a Hadoop Distributed File System (HDFS). Using HDFS provided a robust storage solution capable of handling the volume and variety of data in a distributed manner, ensuring both scalability and fault tolerance.
- **Data Processing with MapReduce**:
  - MapReduce programming in Java was used to process and analyze the raw data in a distributed environment. The MapReduce paradigm allowed us to filter, aggregate, and compute metrics efficiently, such as total spending, average spending, and transaction counts for each customer. This analysis enabled us to derive valuable insights for customer segmentation and loyalty scoring.
- **Data Querying with Hive**:
  - To facilitate complex data queries and improve data accessibility, we integrated Hive, which provided a SQL-like interface to query data stored in HDFS. Using Hive, we conducted advanced queries to segment customers by loyalty level, identify high-value customers, and evaluate spending trends across different time frames. Hive's structured query capabilities enhanced the depth of our analysis and streamlined data retrieval for further processing.
- **Data Analysis and Visualization**:
  - After obtaining the processed data, we applied various analytical methods to derive insights into customer behaviour. Key analyses included Churn Prediction, Customer Segmentation, Loyalty Scoring, and Cross-Sell/Up-Sell Analysis. To communicate findings effectively, we developed a series of visualizations, including radial bar charts, heatmaps, bubble charts, and network graphs. These visuals provided an accessible, intuitive way for stakeholders to understand customer trends and recommendations.
- **Insights and Recommendations**:
  - Based on the visual analysis and statistical outputs, we generated targeted insights and recommendations for customer engagement strategies. The findings identified high-value customers, at-risk accounts, and suitable product recommendations, equipping the bank with data-driven strategies for customer retention and personalized offerings.

Discussion

- Datasets:

1. **“account.txt”**

- The account.txt dataset contains essential banking information for each customer account. It includes the following fields:
    - **Account ID:** A unique identifier for each account.
    - **Customer ID:** A unique identifier linking accounts to individual customers, enabling analysis of multiple accounts under a single customer.
    - **Account Type:** Specifies whether the account is a "savings" or "current" account.
    - **Balance:** The current balance in the account, allows insights into the financial activity and liquidity of customers.
    - **Transactions:** The count of recent transactions, indicating account activity level.
    - **Overdraft Facility:** Indicates if the account has an overdraft facility, which can be analyzed for understanding customer risk levels and credit behaviour.

	0	1	2	3	4	5
0	OMOI8086920Z	1245015582	savings	3667	822	no
1	OMOI8086920Z	1245015582	savings	5806	1035	no
2	OMOI8086920Z	1245015582	savings	1601	635	no
3	OMOI8086920Z	1245015582	savings	4189	802	no
4	LSGS200789GG	1099638085	current	1590	1439	yes

2. **“Person.txt”**

- The person.txt dataset provides demographic information for each account holder, which includes:
    - **Account ID**: Links to the corresponding account in the account.txt dataset, enabling cross-referencing between customer demographics and account activity.
    - **First Name**: The customer’s first name.
    - **Last Name**: The customer’s last name.
    - **Age**: The age of the customer, allowing segmentation based on age groups.
    - **Gender**: The gender of the customer, can support gender-based analysis of account behaviour and financial patterns.
    - **City**: The customer's city of residence, is useful for geographical analysis of spending patterns and customer distribution.

	0	1	2	3	4	5
0	OMOI8086920Z	Allison	Abbott	21	female	Chicago
1	LSGS200789GG	Arthur	Acevedo	20	female	Boston
2	KNBL191477BE	Ana	Acosta	38	male	Los Angeles
3	POOH2832110T	Alex	Adams	43	male	New York
4	KGNZ630334NA	Arlene	Adkins	23	male	Chicago

- R Visualisations:

1. **Customer Spending Analysis:**
   - Visualise total and average spending across customers for Churn Prediction and Loyal Segmentation.
   - Shows recent and past spending for customers in the **Churn Prediction** dataset.

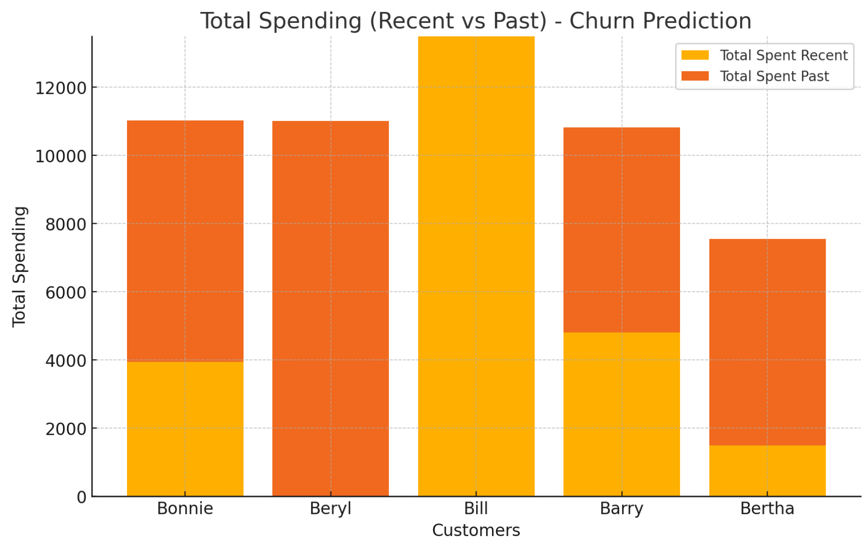

Fig 1: Present and past expenditure of customers in Churn prediction

2. **Customer Segmentation**:
   - Show distribution by **Segment** (e.g., High Spender) in **Loyal Segmentation**.
   - Show **Loyalty** levels in **Loyalty Score**.

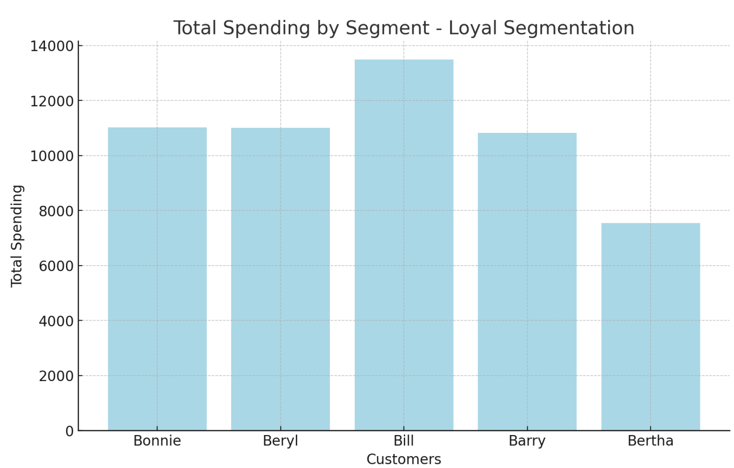

Fig 2: Total Expenditure by Segment

3. **Recommendation Analysis by Account Type**:
    - Display recommendations per account type from **Cross Sell Upsell Analysis**.

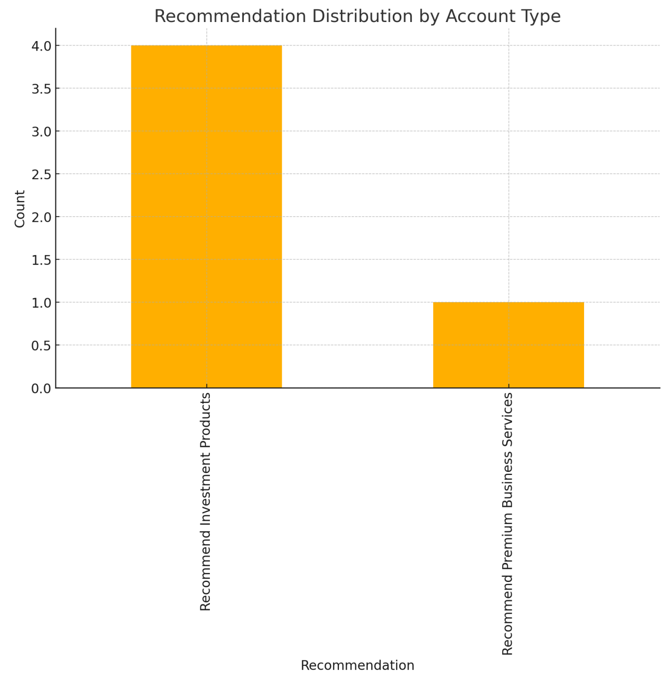

Fig 3: Recommendation Distribution by Account Type

4. **Quarterly Spending Trend:**
    - Shows spending trends over quarters for selected customers in **Spending Trend Analysis**.

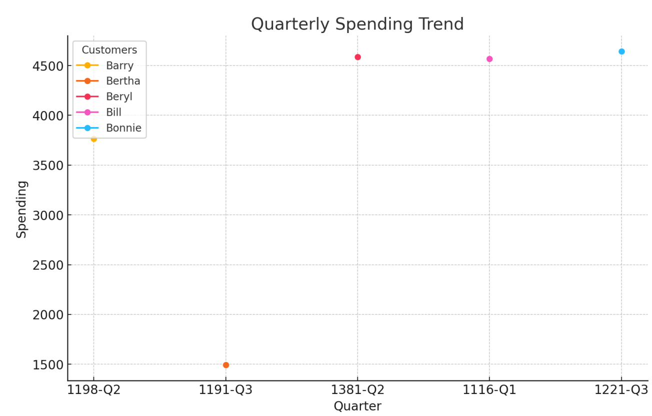

Fig 4: Quarterly Expenditure Trend

5. **Loyalty Distribution**:
   - Visualizes the distribution of loyalty levels in **Loyalty Score**.

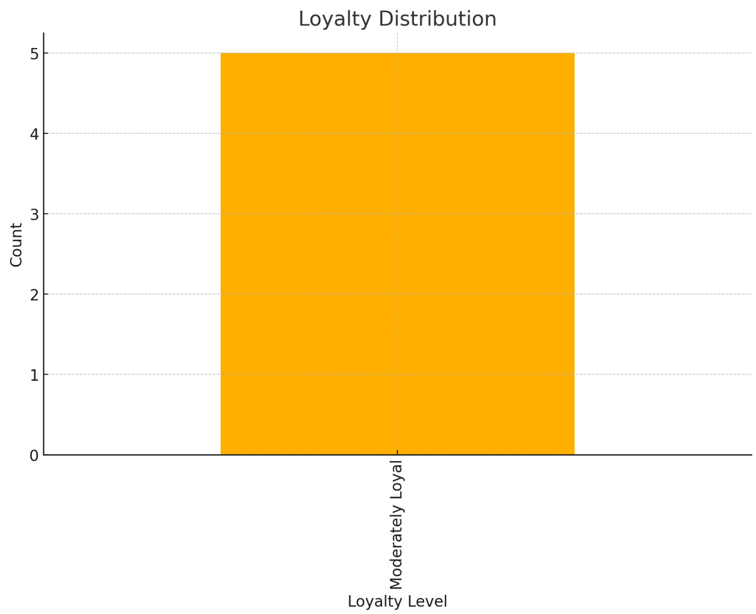

Fig 5: Distribution of Loyalty

6. **Average Spending Comparison**:
   - A comparison of average recent and past spending from **Churn Prediction**.

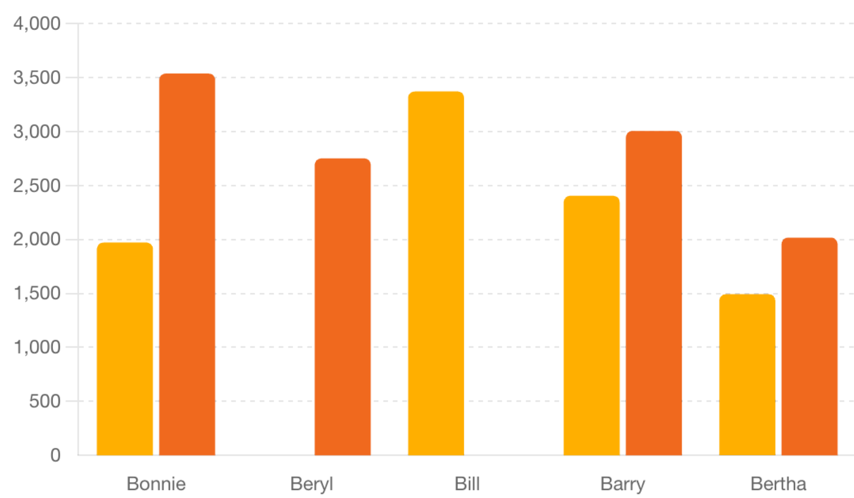

Fig 6: Comparison of Present and past average expenditure

7. **Transactions vs. Total Spending**:
    - A scatter plot of transactions vs. total spending in **Cross Sell Upsell Analysis**.

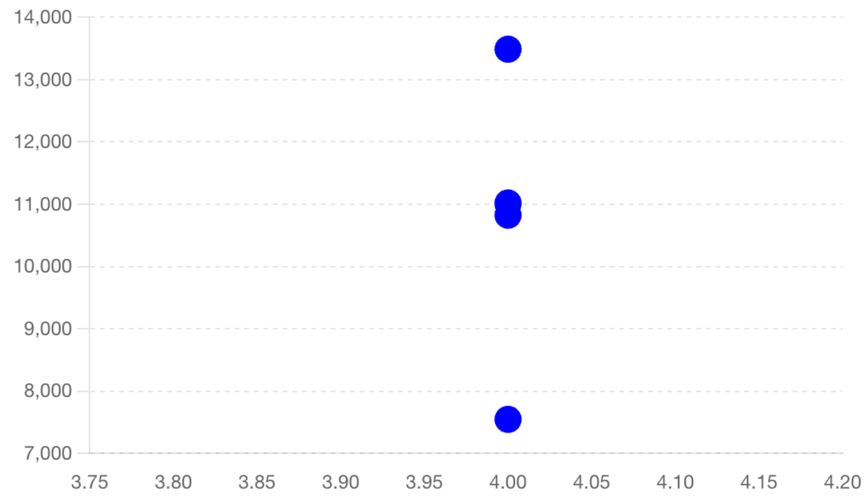

Fig 7: Scatter plot of Transactions and Total Spending

8. **Monthly Spending Trend**:
    - Displays monthly spending trends for an example customer in **Spending Trend Analysis**.

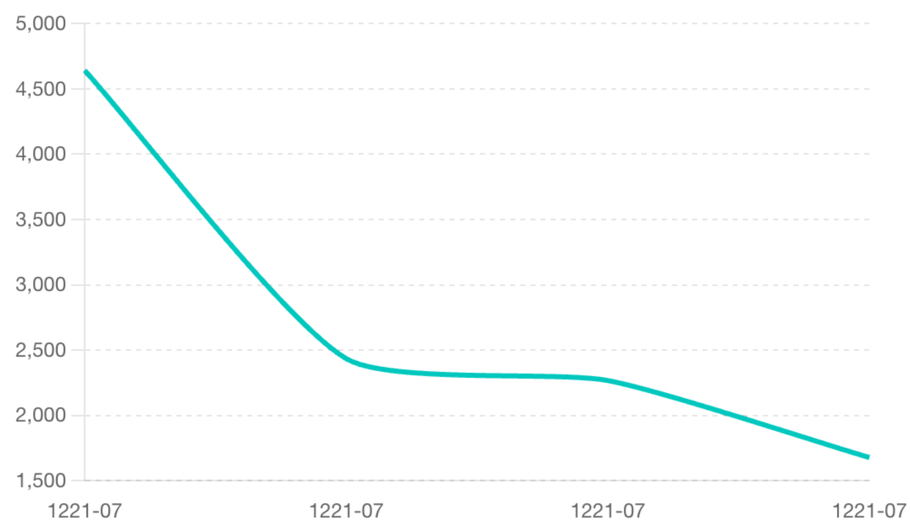

Fig 8: Monthly Expenditure Trend of a customer

9. Comparison of Loyalty Levels and Spending:

- Compare spending in **Loyalty Score** and **Loyal Segmentation**.

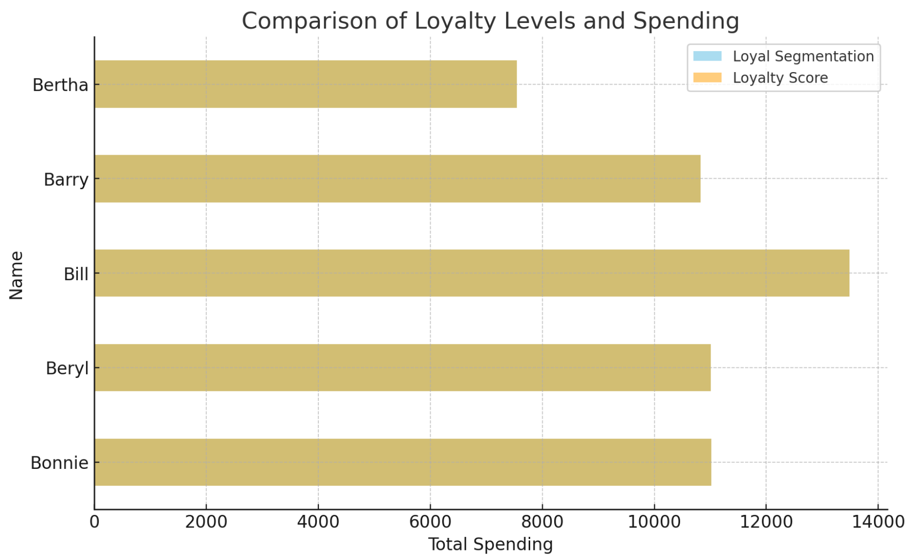

Fig 9: Comparison between Loyalty levels and spending

10. Heatmap with Annotations:

- Show monthly vs. quarterly spending with added detail from **Spending Trend Analysis**

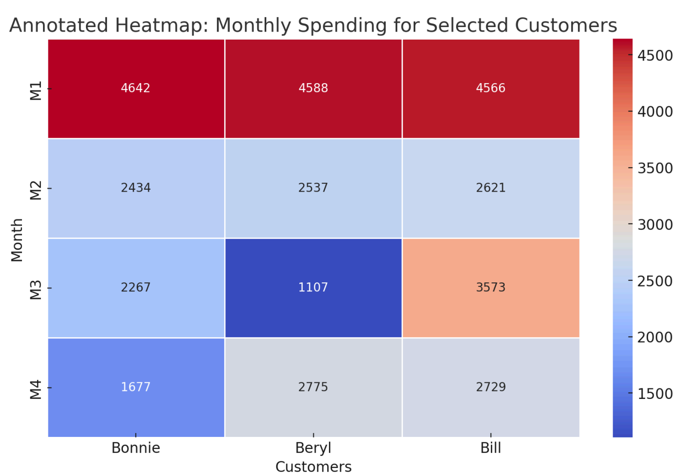

Fig 10: Heatmap of Monthly Spending of Selected Customers

11. Bubble Chart:

- Shows the relationship between total spending, transactions, and loyalty level from **Loyalty Score**.

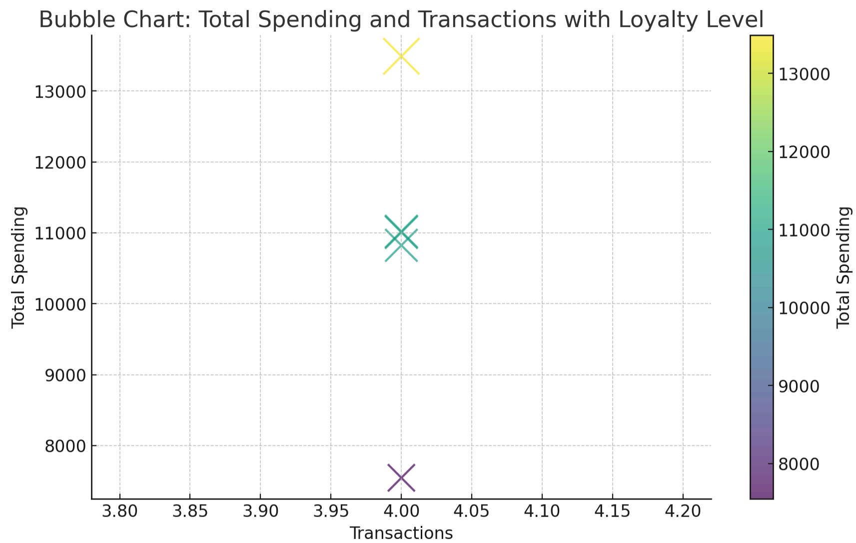

Fig 11: Relationship between Total spending and Transactions

12. Radar Chart:

- Compares spending, loyalty score, transactions, and frequency across customers.

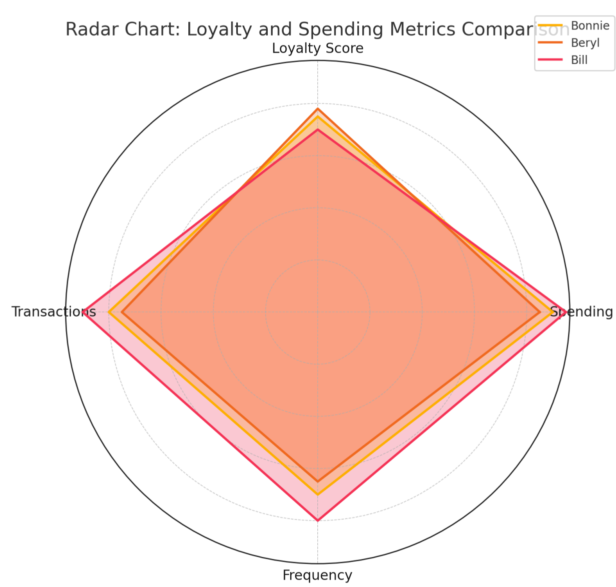

Fig 12: Radar Chart of Loyalty and Spending Metrics

13. Network Graph:

- Illustrates customer connections to recommended products from **Cross Sell Upsell Analysis**.

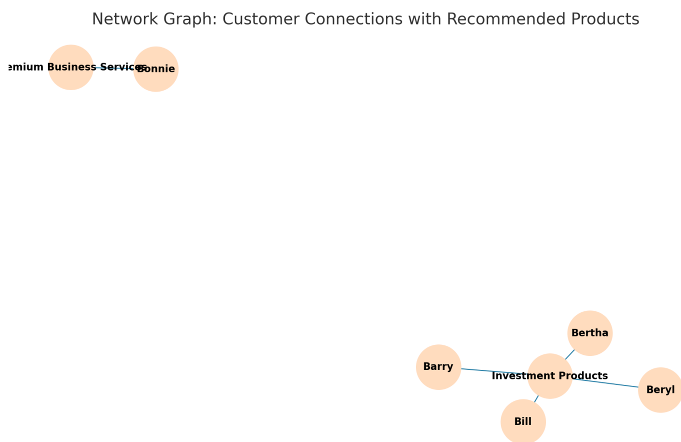

Fig 13: Customer connections with recommended products

14. Horizontal Bar Chart:

- Visualizes customer spending from **Churn Prediction**.

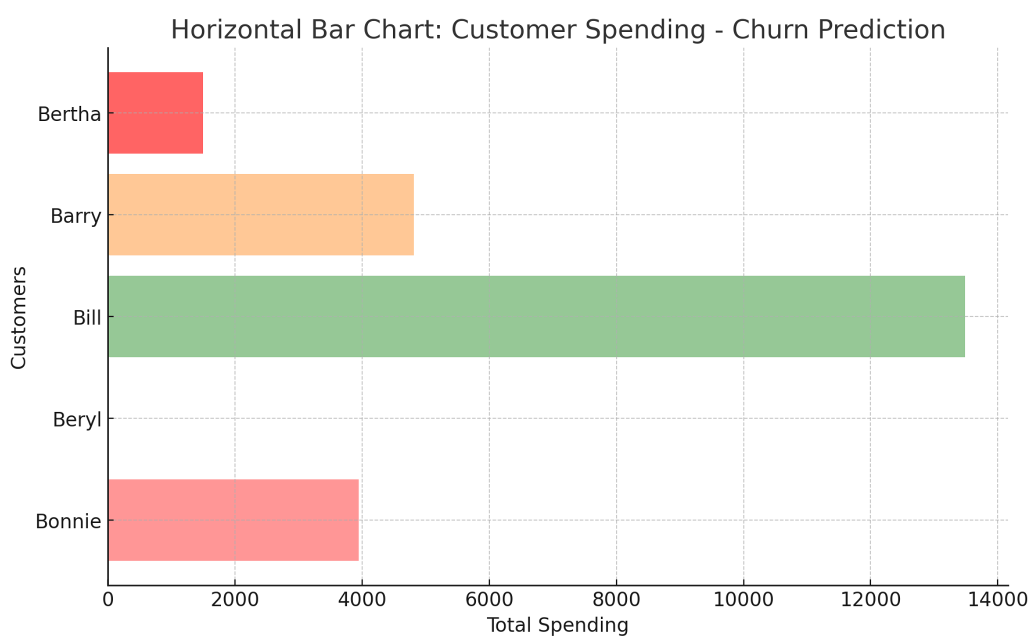

Fig 14: Customer Spending from Churn prediction

15. Chord Diagram Simulation:

- Represents connections between account types and recommendations from **Cross Sell Upsell Analysis**.

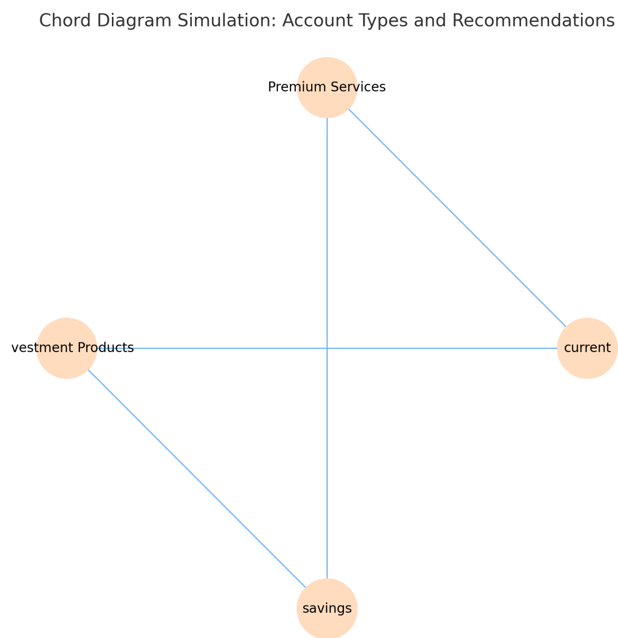

Fig 15: Chord diagram of Account types and Recommendations

16. Sankey Diagram:

- Segment to Recommendation Flow

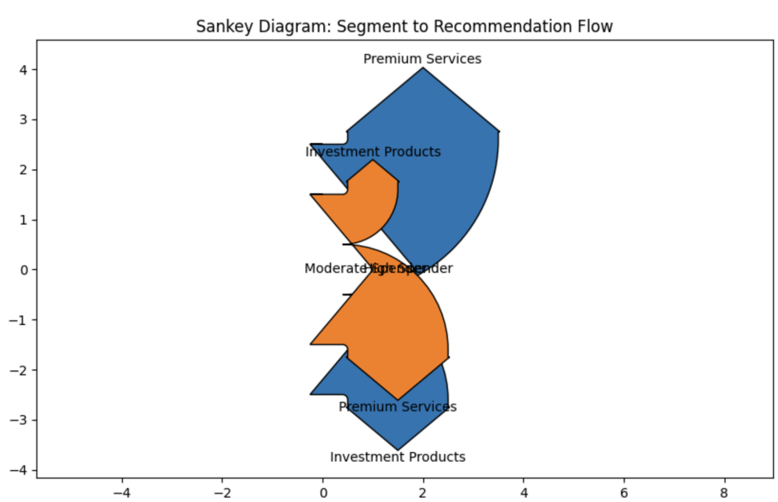

Fig 16: Segment to Recommendation flow

17. Customer Hierarchy (Segment, Loyalty Level, Recommendation):

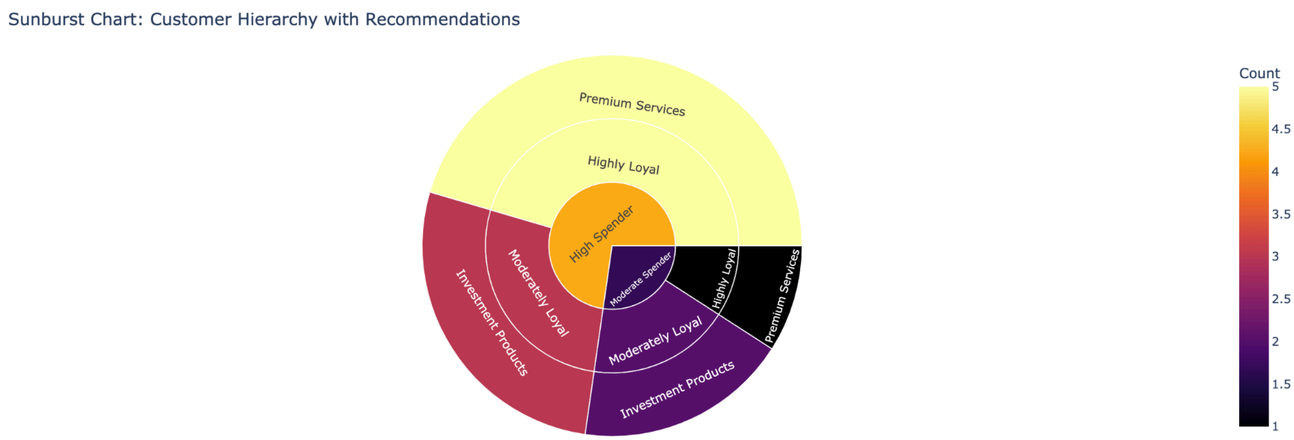

Fig 17: Customer Hierarchy with recommendations

18. Proportional Spending and Transactions by Segment:

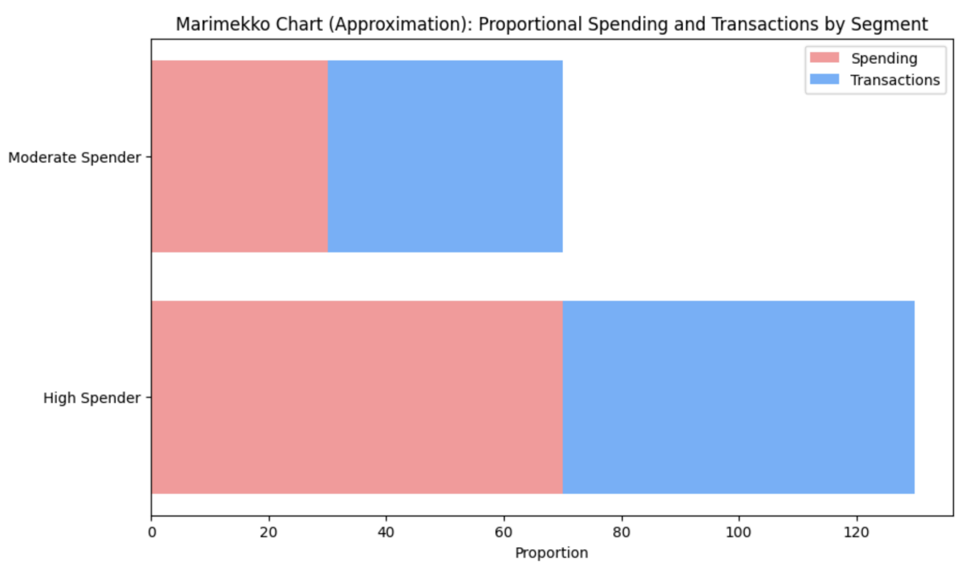

Fig 18: Proportional spending and Transactions by Segment

19. Distribution of spending across customer segments:

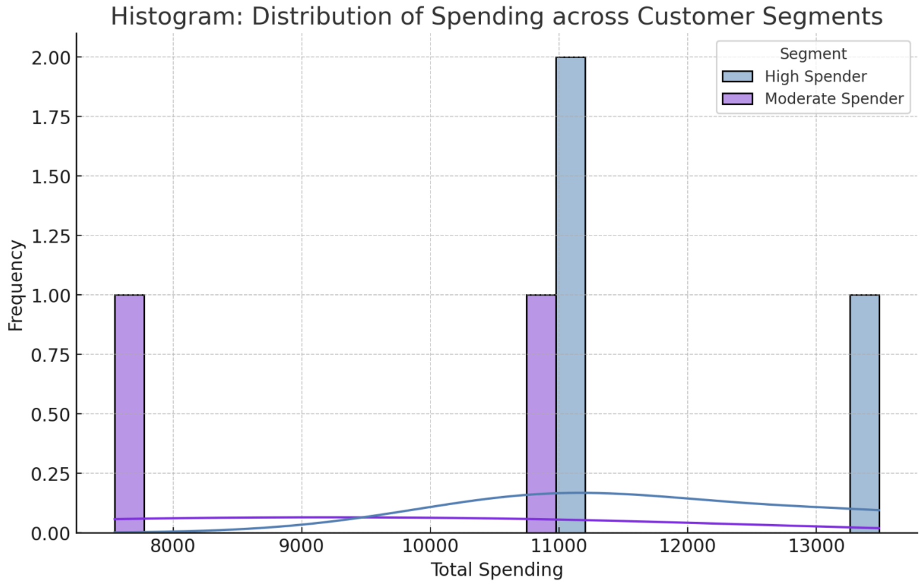

Fig 19: Distribution of Spending across Customer Segments

20. Performance Across Customer Segments:

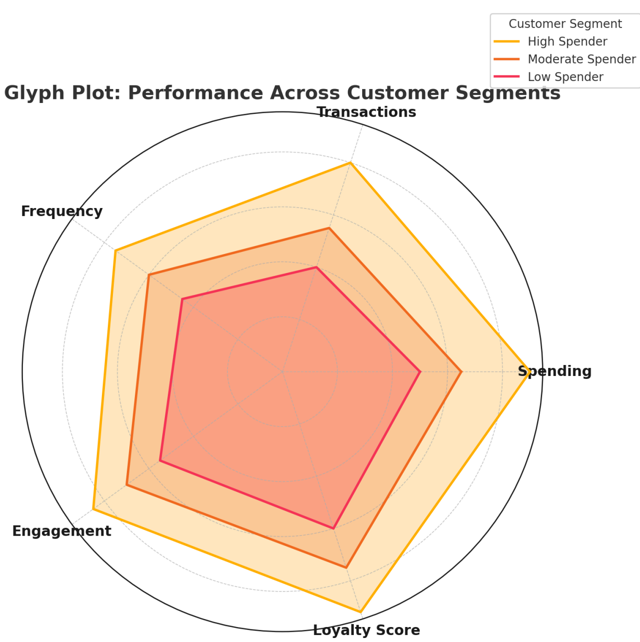

Fig 20: Performance across customer segments

21. Ternary Plot illustrating the distribution of Spending, Transactions, and Frequency across different customers:

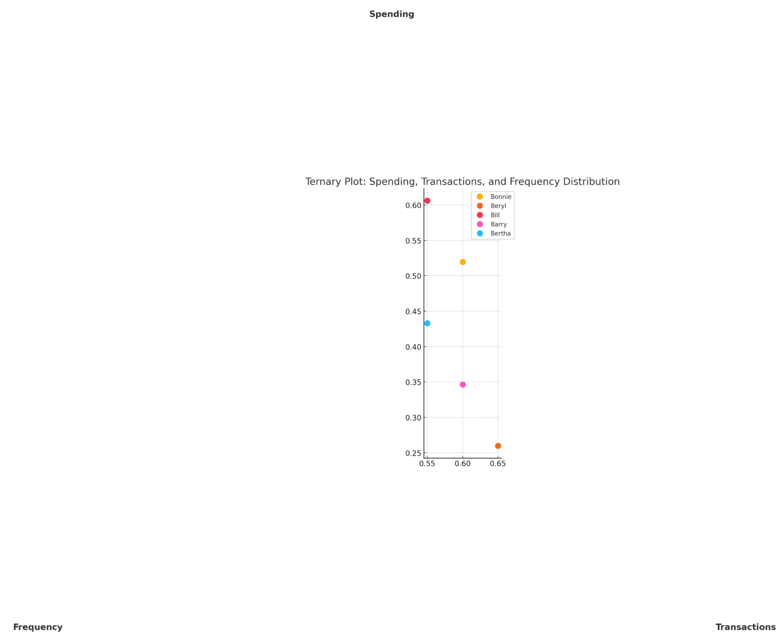

Fig 21: Distribution of Spending, Transactions and Frequency

22. Customer spending patterns and groups similar patterns together:

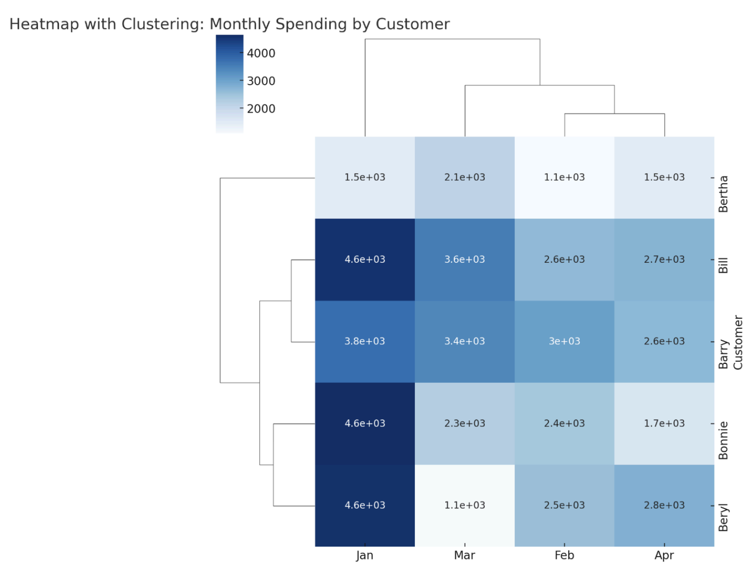

Fig 22: Monthly Expenditure by Customer

23. Network Diagram:

- Illustrates connections between customer segments and recommendations, with edge thickness representing the strength of each cross-sell/upsell recommendation.

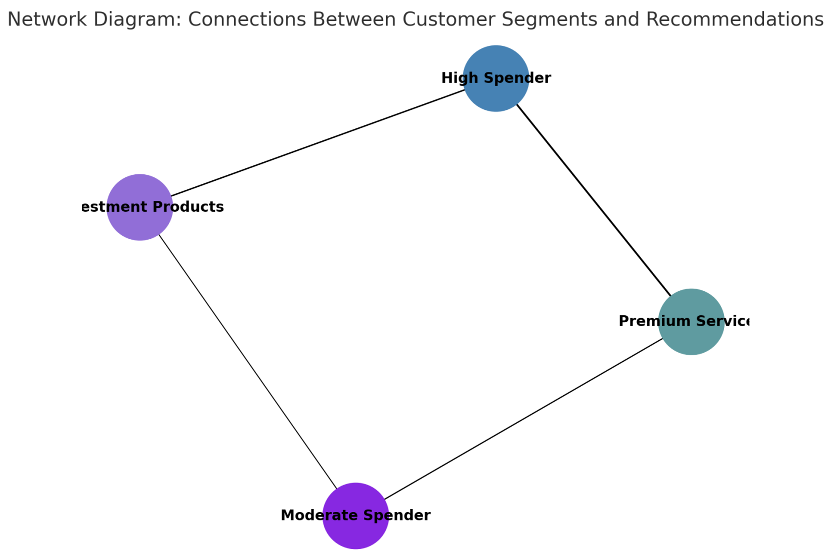

Fig 23: Network Diagram of Customer Segments and Recommendations

24. Dumbbell Plot:

- comparing **Previous Spending** to **New Spending** for each customer after recommendations.

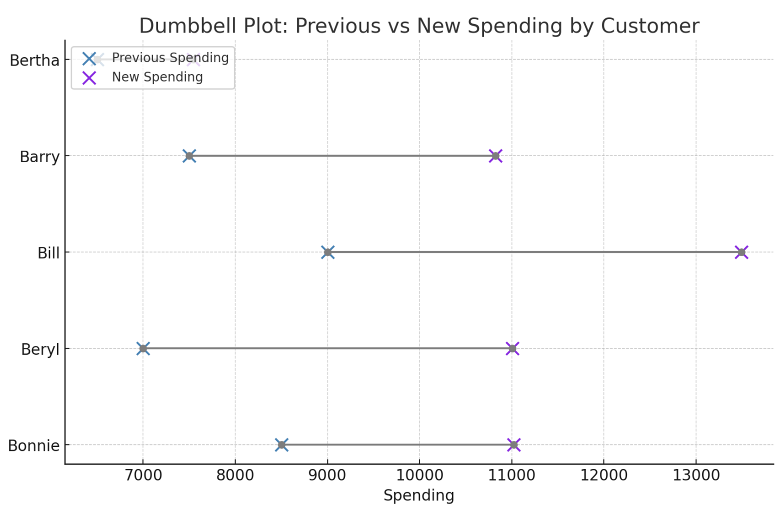

Fig 24: Comparison of Past and Present expenditure

25. Recommendation distribution by account type:

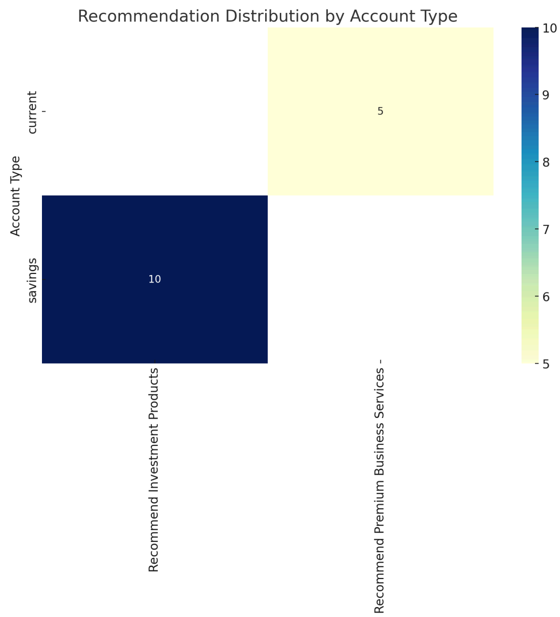

Fig 25: Recommendation System by Account Type

26. Stacked Bar Chart showcasing the monthly spending distribution across customers:

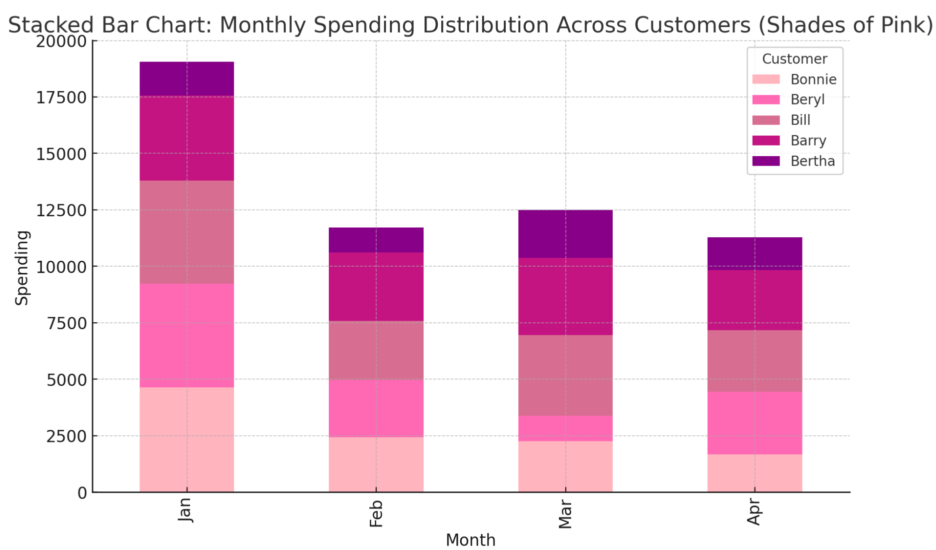

Fig 26: Monthly Expenditure Distribution

- Data Visualizations and Insights Derived from Hive Queries:

1. Total Balance by Account Type:

```
hive> INSERT OVERWRITE DIRECTORY '/user/namratha/hive_outputs/total_balance_by_account_type'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT account_type, SUM(balance) AS total_balance
    > FROM account_table
    > GROUP BY account_type;
```

Fig 27: Command for Total Balance by Account Type

- **Output:**

total_balance_by_account_type.txt

```
current,72270
savings,95668
```

Fig 28: Output Displaying the Total Balance by Account type

- **Visualisation:**

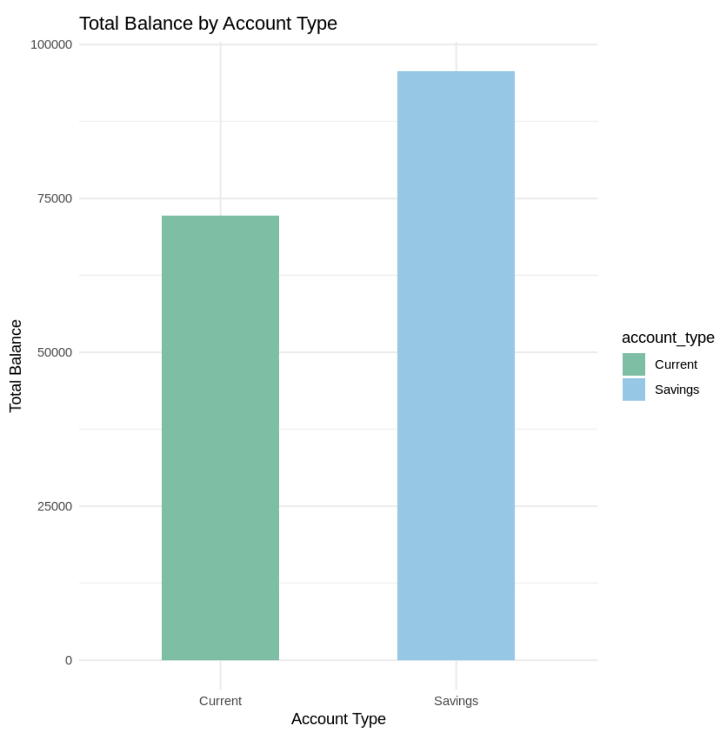

Fig 29: Bar Graph representing Total Balance by Account type

2. High Balance Accounts:

```
hive> INSERT OVERWRITE DIRECTORY '/user/namratha/hive_outputs/high_balance_accounts'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT *
    > FROM account_table
    > WHERE balance > 3000;
```

Fig 30: Command for High Balance Accounts

- Output:

high_balance_accounts.txt

```
OMOI808692OZ,1245015582,savings,3667,822,no
OMOI808692OZ,1245015582,savings,5806,1035,no
OMOI808692OZ,1245015582,savings,4189,802,no
LSGS200789GG,1099638085,current,4383,1242,no
KNBL191477BE,1143512881,current,3304,917,no
KNBL191477BE,1143512881,current,4031,1291,no
KNBL191477BE,1143512881,current,3945,1698,yes
KNBL191477BE,1143512881,current,3374,548,no
POOH283211OT,1113028282,savings,4453,1359,no
POOH283211OT,1113028282,savings,4177,977,no
POOH283211OT,1113028282,savings,3055,1050,no
KGNZ630334NA,1366578552,savings,3751,823,no
KGNZ630334NA,1366578552,savings,6113,1608,no
JRWG458623WI,1399565991,savings,4403,1386,no
HHBH449945BH,1164949711,savings,4686,829,no
HHBH449945BH,1164949711,savings,3764,1567,no
JQOX824533OJ,1183721016,savings,3250,1534,no
BOFX584607FM,1104343425,savings,3573,588,no
BOFX584607FM,1104343425,savings,4566,1902,no
BFBR875138BZ,1188939438,current,4642,1809,no
OBIC304336IX,1400967582,current,4554,1349,no
OBIC304336IX,1400967582,current,4293,1077,no
OBIC304336IX,1400967582,current,4627,684,no
OBIC304336IX,1400967582,current,3759,900,no
BMHX832394HI,1354777658,savings,4588,1169,no
UNMZ436655MA,1108235342,current,3224,1695,no
UNMZ436655MA,1108235342,current,3605,1137,no
NPIQ652416IZ,1152923562,current,3711,1794,no
```

Fig 31: Output Displaying the High Balance Accounts

- Visualisation:

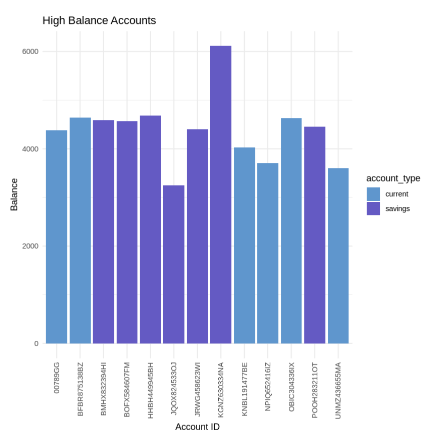

Fig 32: Bar Graph representing High Balance Accounts

3. **Average Balance by Account Type:**

```
hive> INSERT OVERWRITE DIRECTORY '/user/namratha/hive_outputs/average_balance_by_account_type'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT account_type, AVG(balance) AS average_balance
    > FROM account_table
    > GROUP BY account_type;
```

Fig 33: Command for Average Balance by Account type

- **Output:**

average_balance_by_account_type.txt

```
current,3011.25
savings,2989.625
```

Fig 34: Output Displaying the Average Balance by Account type

- **Visualisation:**

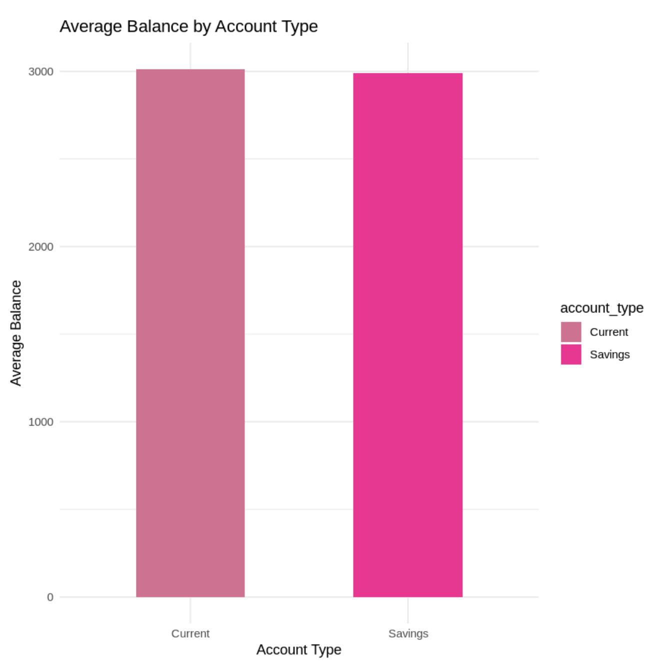

Fig 35: Bar Graph representing Average Balance by Account type

4. Accounts with Overdraft Facility:

```
hive> INSERT OVERWRITE DIRECTORY '/user/namratha/hive_outputs/accounts_with_overdraft'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT *
    > FROM account_table
    > WHERE is_overdraft_allowed = 'yes';
```

Fig 36: Command for Accounts with overdraft facility

- **Output:**

accounts_with_overdraft.txt

```
LSGS200789GG,1099638085,current,1590,1439,yes
LSGS200789GG,1099638085,current,2240,892,yes
LSGS200789GG,1099638085,current,2577,1557,yes
KNBL191477BE,1143512881,current,3945,1698,yes
JRWG458623WI,1399565991,savings,2583,1009,yes
JRWG458623WI,1399565991,savings,1192,1410,yes
JQOX8245330J,1183721016,savings,1493,712,yes
JQOX8245330J,1183721016,savings,1404,546,yes
JQOX8245330J,1183721016,savings,1399,1528,yes
UNMZ436655MA,1108235342,current,1547,1014,yes
UNMZ436655MA,1108235342,current,1088,1378,yes
```

Fig 37: Output displaying accounts with overdraft

- **Visualisation:**

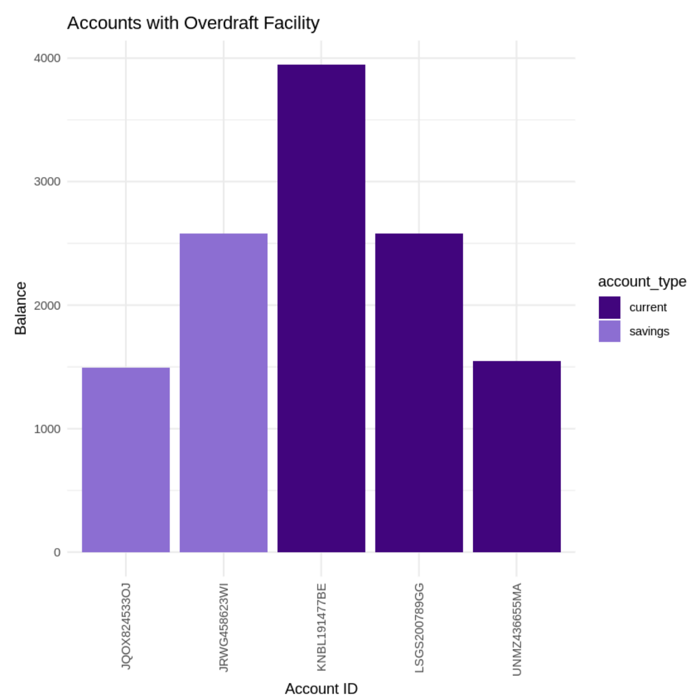

Fig 38: Bar graph showing Accounts with overdraft facility

5. **Account Holder by Gender:**

```
hive> INSERT OVERWRITE DIRECTORY '/user/namratha/hive_outputs/account_holders_by_gender'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT gender, COUNT(*) AS count_of_account_holders
    > FROM person_table
    > GROUP BY gender;
```

Fig 39: Command for Account Holder by Gender

- **Output:**

account_holders_by_gender.txt

```
female,7
male,7
```

Fig 40: Output displaying account holders by gender

- **Visualisation:**

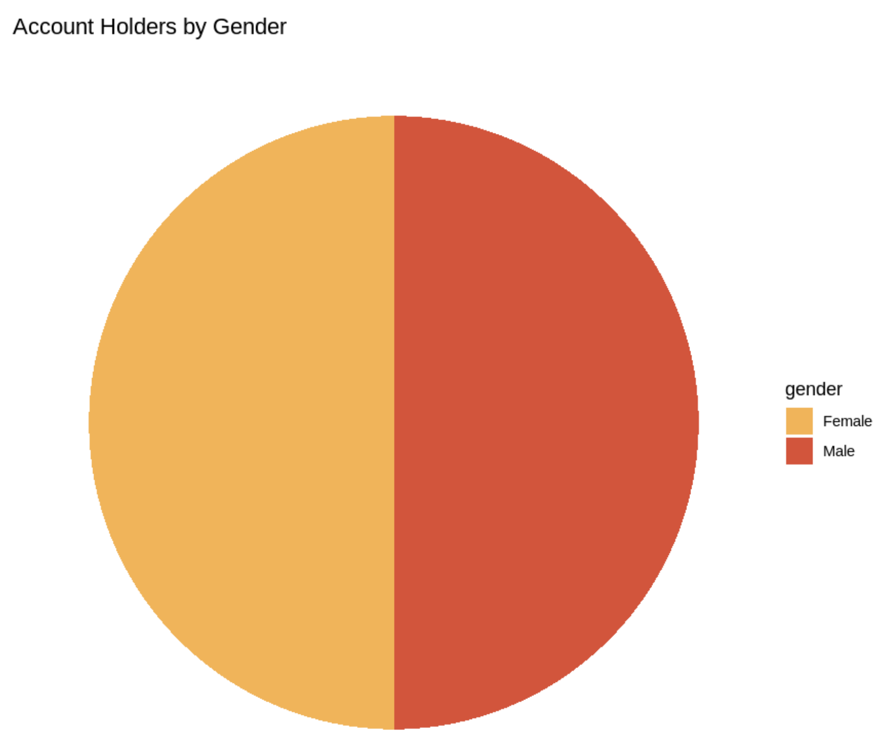

Fig 41: Pie chart showing the account holders by gender

6. **Account Holders Above Age 30:**

```
hive> INSERT OVERWRITE DIRECTORY '/user/namratha/hive_outputs/account_holders_above_30'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT person_id, first_name, last_name, age, city
    > FROM person_table
    > WHERE age > 30;
```

Fig 42: Command for Account holders above age 30

- **Output:**

account_holders_above_30.txt

```
KNBL191477BE,Ana,Acosta,38,Los Angeles
POOH2832110T,Alex,Adams,43,New York
JRWG458623WI,Alberto,Aguilar,33,Los Angeles
BMHX832394HI,Beryl,Allison,40,Los Angeles
NPIQ652416IZ,Cristobal,Alvarado,31,Washington
```

Fig 43: Output showing the account holders above age 30

- **Visualisation:**

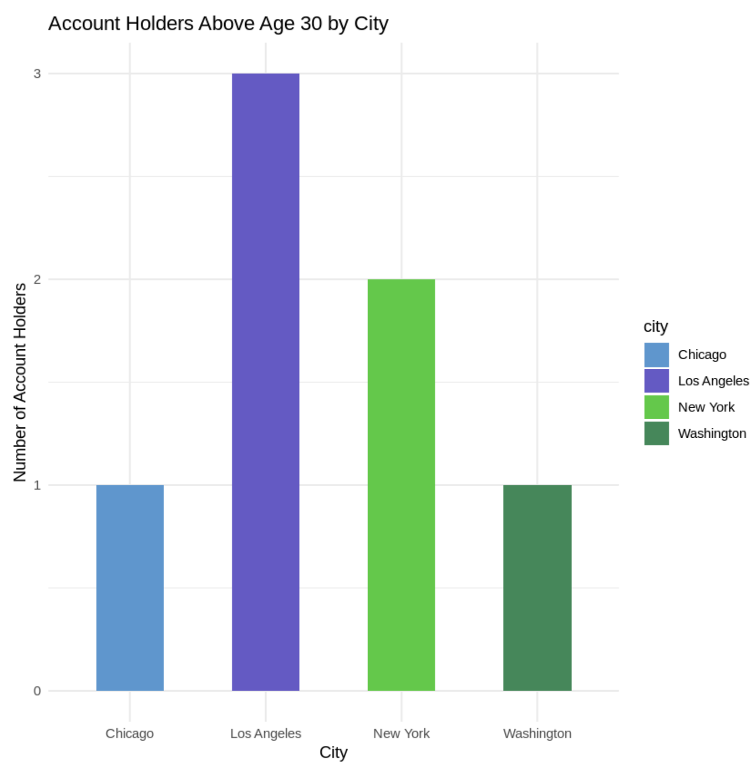

Fig 44: Bar graph showing account holders above age 30 by city

Concluding Remarks

This project effectively explored customer spending patterns, transaction frequencies, and the impact of targeted cross-sell and upsell recommendations. Through detailed analysis and visualisations—including radar charts, stacked bar charts, network diagrams, and other advanced plotting techniques—the project provided insights into customer segmentation, spending distribution, and the effectiveness of various recommendations.

Key findings highlighted the significant role of customer segmentation in tailoring cross-sell and upsell strategies. High spenders, for instance, demonstrated a higher likelihood of engaging with premium services, while moderate spenders showed interest in investment products. These insights are crucial for designing data-driven marketing strategies that align with specific customer needs, ultimately fostering stronger customer loyalty and increased revenue potential.

The various visualisations not only simplified complex data but also allowed for an in-depth comparison of customer behaviours across different metrics. By leveraging these insights, businesses can enhance their understanding of customer preferences, optimize recommendation strategies, and make informed decisions that drive customer engagement and profitability.

In conclusion, this project underscores the importance of data analytics in identifying valuable cross-sell and upsell opportunities. Future work could further refine these strategies by incorporating real-time data, predictive modelling, or machine learning to dynamically adapt recommendations based on evolving customer behaviours.

Future Works

While this project provides comprehensive insights into customer behaviour, loyalty, and segmentation using Hadoop, MapReduce, Hive, and R, there are several areas for further exploration and enhancement. Future work could focus on refining the analysis, improving the predictive accuracy, and expanding the scope of the research. Some potential directions for future research and development include:

1. Real-Time Analytics and Dynamic Modeling: Integrating real-time data processing frameworks like Apache Kafka and Spark Streaming could enable continuous monitoring of customer behavior and churn prediction. Real-time analytics would allow banks to adjust strategies dynamically, enhancing responsiveness to changing customer needs.

2. Advanced Machine Learning and Deep Learning Models: Future work could incorporate more sophisticated machine learning algorithms, such as gradient boosting and deep neural networks, for improved churn prediction and customer segmentation. These models can uncover complex patterns that traditional approaches might miss.

3. Personalized Recommendations with Hybrid Systems: A hybrid recommendation engine combining collaborative filtering and content-based filtering would offer more accurate and contextually relevant product suggestions. By better understanding customer preferences, banks can provide highly personalized cross-sell and upsell opportunities.

4. Customer Lifetime Value (CLV) Modeling: Developing predictive models for Customer Lifetime Value (CLV) could help banks prioritize high-value customers and tailor retention strategies. CLV prediction would complement churn analysis and allow for long-term customer engagement optimization.

References

1. White, T., 2015. Hadoop: The Definitive Guide. O'Reilly Media. Ref
2. Dean, J. & Ghemawat, S., 2008. MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM. 51(1), pp. 107-113. Ref
3. Bhat, S., 2013. Mastering Apache Hive. Packt Publishing. Ref
4. Kumar, V. & Shah, D., 2004. Building and Sustaining Profitable Customer Loyalty for the 21st Century. Journal of Retailing. 80(4), pp. 317-330. Ref
5. Lemmens, A. & Croux, C., 2006. Bagging and Boosting Classification Trees to Predict Churn in the Telecommunications Industry. Expert Systems with Applications. 31(3), pp. 513-522. Ref
6. He, H. & Ma, Y., 2017. Machine Learning for Financial Data Analysis. Springer. Ref
7. Nair, R. & Subramanian, R., 2011. Cross-selling Financial Products to Bank Customers: A Data Mining Approach. Expert Systems with Applications. 38(9), pp. 11589-11603. Ref
8. Zeng, Y. & Liu, Y., 2019. Ethics and Data Privacy in Big Data Analytics. Springer. Ref
9. Kreps, J., Narkhede, N. & Rao, J., 2011. Kafka: A Distributed Messaging System for Log Processing. Proceedings of the NetDB, pp. 1-7. Ref
10. Ricci, F., Rokach, L. & Shapira, B., 2015. Recommender Systems Handbook. Springer. Ref