Department of Computer & Information Science
University of Massachusetts, Dartmouth

DSC 550: Master's Project Final Report

BANKING RISK PROFILING SYSTEM

SRAVANI KAIRAM KONDA

skairamkonda@umassd.edu

02130417

Under the Guidance of Prof. Dr. Ashokkumar Patel

Submission Date: 04/26/2025

TABLE OF CONTENTS

1. ABSTRACT

In the evolving financial landscape, risk management plays a critical role in sustaining banking operations. Loan defaults and financial instability of customers can potentially significantly impact the profitability and reputation of a bank. This project aims to minimize lending risks through the development of a Banking Risk Profiling System that assesses customer demographics, financial behavior, and interaction patterns. With a customer profile constructed on several key metrics in combination, the bank may make lending or non-lending decisions on the basis of information, reducing potential losses. The project also points out how pre-analysis of customers raises a bank's portfolio quality as well as enhances loan recovery rates.

A simulated banking data set was created by combining features of over one freely available data set. Information was structured and uploaded into MySQL Workbench to enable relational modeling between tables such as Clients - Banking, Gender, Investment Advisor, and Banking Relationship. Jupyter Notebook was used for big data preprocessing, exploratory data analysis (EDA), and feature creation such as engagement length and income bands. Power BI was used to design real-time dashboards, which were directly linked to the MySQL database so that the updates would be reflected in real time. Key KPIs such as Total Clients, Total Loans, Deposits, Engagement Days, and Fees were tracked, and dynamic filters and drill-through functionality were introduced for deeper analysis.

The system not only demonstrates the application of SQL, Python, and BI tools but also highlights the effectiveness of data-driven risk assessment in banking. Key insights were generated, such as income-driven loan trends, risk segmentation by nationality, and advisor performance impacts on client portfolios. Through the use of this profiling system, banks can enhance their risk management systems, have better customer screening policies, develop more targeted marketing campaigns, and make better lending decisions. This project is an in-practice example of how data science can have a direct impact on financial sector stability and innovation.

2. INTRODUCTION

2.1 BACKGROUND

Credit risk management has been among the most critical issues for lenders and banks in the banking and lending business. Banks profit primarily through lending money, but every loan involves the risk of default, which can lead to immense financial losses. In the past, determining a client's creditworthiness was often determined by narrow factors, such as income or history with a bank, which at times could not accurately predict default behavior. Banks today have at their disposal massive customer information that can be leveraged for better decision-making since technology and data collection have broadened in scope.

Using customer demographics, behavior at transactions, loyalty statistics, and financial spending, banks can analyze the default or repayment probability better. The banking sector has been revolutionized through data science, as predictive analytics, profiling customers, and segmentation of risks are made possible. Banks ensure not only enhanced client targeting but also reduced non-performing assets (NPAs). Maximum utilization of data profiling can thus enhance operational efficiency, improve customer relationships, and ensure financial stability in competitive banking.

2.2 PROJECT OVERVIEW

This project involves the design of a Banking Risk Profiling System to help banks reduce the risk of default on loans. The system handles customer data, including demographic, financial, and behavioral attributes, to identify the risk level of each customer. A synthetic banking dataset was constructed by combining columns from different real-world datasets. The dataset was structured into a relational database based on MySQL Workbench. It is made up of many interlinked tables such as Clients - Banking, Gender, Investment Advisor, and Banking Relationship, replicating true banking transactions.

Jupyter Notebook was applied to perform broad data cleaning, feature engineering (e.g., creating engagement days and income bands), and exploratory data analysis (EDA). Statistical trends were observed, and principal KPIs were resolved to measure client financial behaviors. Later, Power BI was used to develop dynamic dashboards directly connected to the live MySQL database. Dashboards show KPIs such as Total Clients, Loan Amount, Deposits, Engagement Duration, and Processing Fees, along with interactive features such as drill-throughs and slicers. The project thus creates an end-to-end data analysis pipeline from raw data to actionable visual insights, giving banks a real-world tool to improve lending decisions.

3. RELATED WORK

Numerous research studies and practical uses have established the significance of data analytics in risk management in the banking sector. Numerous research studies have established the efficacy of credit scoring models driven by machine learning algorithms such as decision trees, logistic regression, and random forests over traditional methods of risk analysis. The significance of client segmentation, behavior analysis, and early warning systems for predicting loan defaults has been established through numerous studies. Organizations like the World Bank and large private banks have been increasingly investing in customer profiling predictive analytics.

Also, recent Kaggle competitions and open datasets like the Loan Default Prediction Dataset have provided researchers with an opportunity to experiment with real-world financial data. Projects in banking data science often focus on clustering customers, risk level scoring, and recommending appropriate financial products. This project is inspired by such comparable works but adapts the methodology to develop a more practical, dashboard-driven, mid-size solution combining SQL databases, Python analysis, and Power BI visualization. Unlike model-based studies alone, this project emphasizes integrated workflow development for banking decision support.

4. DATASET DESCRIPTION

The dataset created for this project contains structured information related to banking operations and customer financial behavior. It was designed to simulate a real-world banking environment by organizing client details, financial records, and relationship mappings into a relational database structure. To replicate real banking systems, multiple tables were interlinked using primary keys and foreign keys, ensuring logical relationships between clients, banking contacts, gender classifications, and investment advisors. This design allows for meaningful analysis across different dimensions of client activities and financial engagements.

Key Tables:

- **Clients** - Banking, Gender, Banking Relationship, Investment Advisor

These tables collectively capture vital aspects of client information.

Core data dimensions represented in the dataset are:

- **Demographics**: Age, Gender, Nationality, Occupation
- **Financials**: Deposits, Loans, Credit Card Balances, Processing Fees
- **Engagements**: Bank Join Date, Loyalty Classification, Advisor Assignment

Each client record is tied to their financial transactions, risk classifications, and engagement duration, enabling multi-dimensional analysis of risk patterns and financial behavior. The dataset thus provides a holistic view of client activities within a bank.

The dataset was manually constructed in Excel by combining attributes from multiple publicly available financial and banking data sources. This synthetic yet realistic structure was carefully curated to maintain diversity across demographics and financial patterns. Later, the dataset was uploaded into MySQL Workbench for relational database modeling, allowing for further data management, querying, and integration with analysis platforms. With its wide range of demographic and transactional data, the dataset laid a strong foundation for detailed exploratory data analysis (EDA), feature engineering, and risk profiling for the Banking Risk Profiling System project.

The dataset utilized for this project was synthetically created by compiling features from multiple public banking datasets and manually formatting them into an Excel workbook. The intent was to design a data structure that closely resembles real-world banking systems, capturing a wide range of client attributes, financial activities, and engagement relationships. To simulate an actual banking database, multiple tables were designed with primary keys and foreign keys, enabling logical associations between clients, advisors, banking contacts, and other operational factors.

Tables and Columns Description:

Clients - Banking: The main table containing customer details such as Client ID, Name, Age, Location ID, GenderId, Nationality, Occupation, Estimated Income, Fee Structure, Loyalty Classification, Bank Join Date, Bank Loans, Deposits, Checking and Saving Account Balances, Foreign Currency Accounts, Credit Card Balances, Business Lending, Properties Owned, Risk Weighting, and associated Banking Relationship IDs.

Gender: A simple mapping table linking GenderId to Gender (e.g., 1 - Male, 2 - Female).

Banking Relationship: It contains relationship details and identifiers to show which bank contact is associated with each client.

Investment Advisor: Maps clients to their designated financial advisors, facilitating advisory-related analyses.

Each of these tables was relationally linked using appropriate primary (e.g., Client ID, GenderId, BRId) and foreign keys to maintain the integrity of the database and allow flexible querying across different attributes.

Data Cleaning and Feature Engineering: Several critical data cleaning and feature enhancement steps were carried out after the dataset was structured:

Engagement Days: A new column, Engagement Days, was created by calculating the difference between the client's Joined Bank date and today's date. This provided insight into the length of the customer's relationship with the bank.

Income Bands:

Customers were categorized into income segments for easier risk analysis:

Low Income: Estimated Income < $100,000
Mid Income: Estimated Income between $100,000 and $300,000
High Income: Estimated Income > $300,000

Processing Fees:

Processing Fees were derived based on the client's Fee Structure (e.g., High = 5%, Mid = 3%, Low = 1%).

Handling Missing Values:

Missing or inconsistent values were cleaned by either imputing defaults where logical or removing erroneous entries during the preprocessing phase. By implementing these cleaning steps, the final dataset provided a robust and clean base for conducting exploratory data analysis (EDA) and developing dashboards that profile banking clients in a structured and insightful manner.

5. TOOLS AND TECHNOLOGIES USED

Successful execution of the Banking Risk Profiling System project relied on a combination of database management systems, data analysis platforms, and data visualization tools. The chosen technologies ensured an integrated workflow from raw data handling to final business intelligence dashboard development.

MySQL Workbench

MySQL Workbench was used for database creation, data uploading, relational modeling, and structured querying. Tables representing clients, gender classifications, banking relationships, and advisory assignments were created and normalized. Primary and foreign key relationships were enforced to maintain data integrity, and queries were written to retrieve, join, and manipulate datasets efficiently for further analysis.

Key Functions Used:

Database creation and relational schema modeling

Data importation from Excel

SQL querying for data validation and preparation

Jupyter Notebook

Jupyter Notebook provided a flexible and powerful environment for data cleaning, feature engineering, and exploratory data analysis (EDA). Using Python and several scientific libraries, the raw extracted tables were transformed into ready-to-analyze formats.

Python Libraries Utilized:

Pandas: For data manipulation, cleaning, and tabular operations.

Seaborn: For creating statistical visualizations like bar plots, histograms, and box plots.

Matplotlib: For detailed plotting of distributions, relationships, and client behaviors.

NumPy: For numerical operations and handling missing values.

Major Tasks:

Handling missing or incorrect data

Feature creation (Engagement Days, Income Bands, Processing Fees)

Generating exploratory visualizations to identify patterns and insights

Power BI

Power BI was used to build dynamic, interactive dashboards directly connected to the live MySQL database. It enabled the development of key performance indicators (KPIs) and detailed visual representations of client risk profiling metrics.

Features Implemented:

KPI Cards: Total Clients, Total Loans, Deposits, Engagement Days, Total Fees

Drill-through Capabilities: Allowing deep dives from summary statistics into detailed client-level data

Dynamic Slicers: Enabling filtering by Income Band, Gender, Nationality, etc.

Real-Time Data Connection: Dashboards automatically updated when the MySQL database was modified

Power BI enhanced the interpretability of the data and allowed stakeholders to quickly and intuitively assess client risk profiles without the need for technical querying.

6. PROJECT METHODOLOGY

The Banking Risk Profiling System was developed from a highly structured five-step methodology. The project moved sequentially through a rational pipeline from taking in structured banking data into a relational database, data cleaning and exploratory analysis, creation of key performance indicators (KPIs), presentation of insights from dynamic dashboards, and final interpretation of results to support actual banking decisions.

Step 1: Data Uploading to MySQL Workbench

Step one involved loading the processed dataset onto MySQL Workbench. The data had been previously prepared in Excel through manual piecing together of different publicly available dataset attributes to mimic realistic banking transactions. Tables titled Clients - Banking, Gender, Banking Relationship, and Investment Advisor were created to hold client economic and demographic information.

To maintain relational integrity:

Primary keys were assigned to important fields like Client ID, Gender ID, and Banking Relationship ID.

Foreign key relationships were established between tables in order to ensure data consistency.

Data were imported into MySQL through the Import Wizard. Upon uploading, simple SELECT, JOIN, and GROUP BY queries were used to validate data by ensuring relationships among clients, genders, banking contacts, and advisors were accurately established. This provided a clean-structured backend database that provided a solid foundation for data analysis and visualization.

Step 2: Data Cleaning and EDA in Jupyter Notebook

After the data was structured in MySQL, the next step was to link the database to Jupyter Notebook using Python's mysql-connector library. SQL queries were employed to import tables into Pandas DataFrames for cleaning and exploration.

Data Cleaning Activities:

Missing Values Handling: Scanned through.isnull() and.sum(), and filled or dropped sensibly where needed.

Standardizing Data Types: Ensured that fields like dates, income, loans, and deposits were of proper data types.

Fixing Inconsistencies: Validated and corrected any inconsistencies between related fields (e.g., Gender IDs, Banking Relationships).

Exploratory Data Analysis (EDA):

Visualizing Distributions: Seaborn and Matplotlib were employed to plot income levels, deposit amounts, loan amounts, and credit card balances. Financial Behavior Trends: Trends at the beginning across nationalities, genders, and loyalty categories were examined.

Outlier Detection: Looked for unusually high loan values or deposit values that would distort analysis.

EDA enabled the determination of major trends and the understanding of client financial patterns, which paved the way for KPI design and dashboard configuration.

Step 3: Building Key Performance Indicators (KPIs)

Following data cleaning and an understanding of basic trends, the next step was to develop a list of Key Performance Indicators (KPIs) in order to quantify client profiles and banking operations.

Key KPIs Developed:

Total Clients: Count of unique client IDs.

Total Loans: Aggregate of Bank Loans, Business Lending, and Credit Card Balances.

Total Deposits: Sum of Checking Accounts, Saving Accounts, Bank Deposits, and Foreign Currency Accounts.

Processing Fees: Calculated as a percentage of Total Loan Amounts based on Fee Structure type (High, Mid, Low).

Total Credit Card Amount: Amount corresponding to credit card balances.

The KPIs were initially verified using Python calculations and re-written in Power BI using DAX measures to facilitate real-time dynamic updation of the dashboard.

Step 4: Power BI Dashboard Development

After KPIs were developed and tested, the next step involved creating interactive Power BI dashboards. The Power BI dashboards were linked with the MySQL database for real-time updation and visualization of data directly.

KPI Cards: Provided key headline statistics like Total Clients, Total Loan, Total Deposit, and Total Fees.

Bar and Pie Charts:

Shown loan distribution by income band and nationality.

Showed deposits breakdown by account types.

Drill-Through Pages: Enabled users to drill down from summary stats to client detail.

Dynamic Slicers: Enabled filtering by Gender, Nationality, Loyalty Classification, Fee Structure, and Advisor assignment.

The visualizations were constructed so that both general summaries and fine-grained deconstructions would be available, allowing stakeholders to easily analyze client portfolios and money behaviors.

Step 5: Deriving Insights and Business Interpretation

The final step was to go through the dashboards and KPIs to obtain meaningful insights about client financial behavior and banking risk profiles.

Key Observations:

High-Income Clients: Larger loan sizes were taken by higher estimated income clients but with higher deposits, indicating lower risk.

Loan Exposure by Nationality: A few nationalities demonstrated higher exposure of loans, which had to be tracked with more attention for risk management.

Advisor Influence: Investment advisors with more successful client portfolios can be identified, offering strategic advantages.

Account Type Trends: Savings and checking accounts remained the top deposit vehicles for clients across most segments.

These findings confirmed the project's fundamental purpose — to enable smarter, data-driven loan approval decisions and enable banks to better understand potential client risk.

The project's end product — a set of interactive dashboards backed by a clean, relational database and detailed financial KPIs — created an efficient decision-support system that banking institutions may utilize for loan default risk reduction and maximizing client engagement.

7. DATA CLEANING (MYSQL WORKBENCH)

Creating a new column Engagement Timeframe in client-banking column which tells about the timeline of the clients in banks

```
Engagement Timeframe =
SWITCH(TRUE(),
'Clients - Banking'[Engagment Days] < 365 , "< 1 Years",
'Clients - Banking'[Engagment Days] < 1825 , "< 5 Years",
'Clients - Banking'[Engagment Days] < 3650 , "< 10 Years",
'Clients - Banking'[Engagment Days] < 7300 , "< 20 Years",
"> 20 Years")
```

Processing Fees	Engagment Days	Engagement Timeframe	Income Band	Client ID	Name	Age	Location ID	Joined Bank	Banking Contact	Nationality	Occupation
0.05	858	< 5 Years	Mid	IND41067	Louis Ramirez	45	2764	15 April 2020	Shawn Cook	European	Office Assistant I
0.05	725	< 5 Years	Low	IND33193	Scott Ward	27	41166	26 August 2020	Todd Roberts	European	Accountant IV
0.05	8220	> 20 Years	Mid	IND17984	Gary Bell	18	31814	18 February 2000	Nicholas Cunningham	European	Office Assistant IV
0.05	1161	< 5 Years	Mid	IND34859	David Fernandez	38	10258	17 June 2019	Nicholas Cunningham	European	Recruiter
0.05	611	< 5 Years	Low	IND55021	Brandon Morrison	75	27514	18 December 2020	Dennis Ruiz	European	Tax Accountant
0.05	844	< 5 Years	Mid	IND87992	Walter Matthews	36	31048	29 April 2020	Anthony Torres	European	Geologist I
0.05	3351	< 10 Years	Low	IND31796	Heather Welch	26	7151	18 June 2013	Joshua Bennett	European	Account Executive
0.05	1592	< 5 Years	Mid	IND56699	Paul Bell	45	31169	12 April 2018	Donald Reynolds	European	Statistician IV
0.05	2639	< 10 Years	Mid	IND35302	Jessica Black	78	35348	31 May 2015	Joe Hanson	European	Product Engineer
0.05	660	< 5 Years	Mid	IND34318	Nancy Black	40	18785	30 October 2020	Bruce Butler	European	Engineer IV
0.05	7214	< 20 Years	Mid	IND79633	Samuel Gilbert	54	32020	20 November 2002	Jesse Evans	European	Geologist I
0.05	829	< 5 Years	Mid	IND25477	Anthony Gardner	69	39052	14 May 2020	Joshua Ryan	European	Health Coach IV

Creating a new column Engagement Days in Client-Banking table how many days the client spent from the date of joining in banks

```
Engagment Days =
DATEDIFF('Clients - Banking'[Joined Bank],TODAY(), DAY )
```

Processing Fees	Engagment Days	Engagement Timeframe	Income Band	Client ID
0.05	858	< 5 Years	Mid	IND41067
0.05	725	< 5 Years	Low	IND33193
0.05	8220	> 20 Years	Mid	IND17984
0.05	1161	< 5 Years	Mid	IND34859
0.05	611	< 5 Years	Low	IND55021
0.05	844	< 5 Years	Mid	IND87992
0.05	3351	< 10 Years	Low	IND31796
0.05	1592	< 5 Years	Mid	IND56699
0.05	2639	< 10 Years	Mid	IND35302
0.05	660	< 5 Years	Mid	IND34318
0.05	7214	< 20 Years	Mid	IND79633
0.05	829	< 5 Years	Mid	IND25477

Creating bins for the Estimated Income < 100000 as low and <300000 as Mid with the column named as Income Band in Clients-Banking table.

```
1 Income Band =
2 SWITCH(TRUE(),
3 'Clients - Banking'[Estimated Income] < 100000,"Low",
4 'Clients - Banking'[Estimated Income] < 300000,"Mid",
5 "High" )
```

Processing Fees	Engagment Days	Engagement Timeframe	Income Band	Client ID
0.05	2425	< 10 Years	Mid	IND16101
0.05	5966	< 20 Years	Low	IND26283
0.05	1522	< 5 Years	Mid	IND97689
0.05	1669	< 5 Years	High	IND88778
0.05	4100	< 20 Years	Low	IND92423
0.05	7884	> 20 Years	Mid	IND38441
0.05	1610	< 5 Years	Mid	IND79955
0.05	858	< 5 Years	Mid	IND41067
0.05	725	< 5 Years	Low	IND33193
0.05	8220	> 20 Years	Mid	IND17984
0.05	1161	< 5 Years	Mid	IND34859
0.05	611	< 5 Years	Low	IND55021
0.05	844	< 5 Years	Mid	IND87992
0.05	3351	< 10 Years	Low	IND31796
0.05	1592	< 5 Years	Mid	IND56699
0.05	2639	< 10 Years	Mid	IND35302
0.05	660	< 5 Years	Mid	IND34318

Creating a new column named Processing Fees for the column Fee Structure like if fee structure is high then processing fee would be 0.05

```
1 Processing Fees =
2 SWITCH('Clients - Banking'[Fee Structure],
3 "High",0.05,
4 "Mid" ,0.03,
5 "Low" , 0.01 , 0
6 )
```

Processing Fees	Engagment Days	Engagement Timeframe	Income Band
0.05	2425	< 10 Years	Mid
0.05	5966	< 20 Years	Low
0.05	1522	< 5 Years	Mid
0.05	1669	< 5 Years	High
0.05	4100	< 20 Years	Low
0.05	7884	> 20 Years	Mid
0.05	1610	< 5 Years	Mid
0.05	858	< 5 Years	Mid
0.05	725	< 5 Years	Low
0.05	8220	> 20 Years	Mid
0.05	1161	< 5 Years	Mid
0.05	611	< 5 Years	Low

KPI’S:

In which followings KPIS are present:

Total Clients:

Total Clients KPI represents total number of clients in banking.

Total Clients = DISTINCTCOUNT('Clients - Banking'[Client ID])

Total Loan:

Total Loan gives you information about the bank loan + Business lending + credit cards balance of particular investor , gender.

Total Loan = [Bank Loan] + [Business Lending] + [Credit Cards Balance]

Bank Loan:

Bank Loan gives you information about what is the loan amount of loan to be repaid by the client to bank.

Bank Loan = SUM('Clients - Banking'[Bank Loans])

Business Lending:

Business lending gives you information about the loan amount given to small business.

Business Lending = SUM('Clients - Banking'[Business Lending])

Total Deposit:

Total Deposit gives you information about the amount deposited by particular investors in bank

Total Deposit = [Bank Deposit] + [Savings Account] + [Foreign Currency Account] + [Checking Accounts]

Total Fees:

Total Fees is nothing but the amount charged by the bank for account set-up , maintenance charges etc.

Total Fees = SUMX('Clients - Banking' , [Total Loan] * 'Clients - Banking'[Processing Fees])

Bank Deposit:

Bank deposit is the money put in the bank.

Bank Deposit = SUM('Clients - Banking'[Bank Deposits])

Checking Account Amount:

Checking account amount is nothing but which offers easy access to your money for daily transactional needs.

Checking Accounts = SUM('Clients - Banking'[Checking Accounts])

Total CC Amount:

Total CC Amount is a short-term source of financing for a company by a bank.

Total CC Amount = SUM('Clients - Banking'[Amount of Credit Cards])

Saving Account Amount:

A savings account is an interest-bearing deposit account held at a bank.

Savings Account = SUM('Clients - Banking'[Saving Accounts])

Foreign Currency Amount:

Foreign Currency Account means an account held in a currency that is not the currency of India or Bhutan or Nepal.

Foreign Currency Account = SUM('Clients - Banking'[Foreign Currency Account])

Engagement Account:

Engagement Banking is nothing but puts the customer at the center and aims to deliver the digital experiences they expect.

Engagment Length = SUM('Clients - Banking'[Engagment Days])

Credit Cards Balance:

It is the total amount of money currently owned by a cardholder to their credit card bank.

Credit Cards Balance = SUM('Clients - Banking'[Credit Card Balance])

8. EXPLORATORY DATA ANALYSIS (EDA)

The Exploratory Data Analysis (EDA) process was utilized to understand the distribution, organization, and connections in the banking data set. Python libraries such as Pandas, Seaborn, Matplotlib, and NumPy were employed to perform statistical summaries and visualizations with insight. The EDA process was utilized to uncover patterns and to prepare the data for KPI building and developing dashboards.

1. Distribution Analysis

Income Band Distribution

A bar plot was employed to observe the distribution of clients across different Income Bands (Low, Mid, High). It was observed that:

The majority of clients belonged to the Mid Income Band.

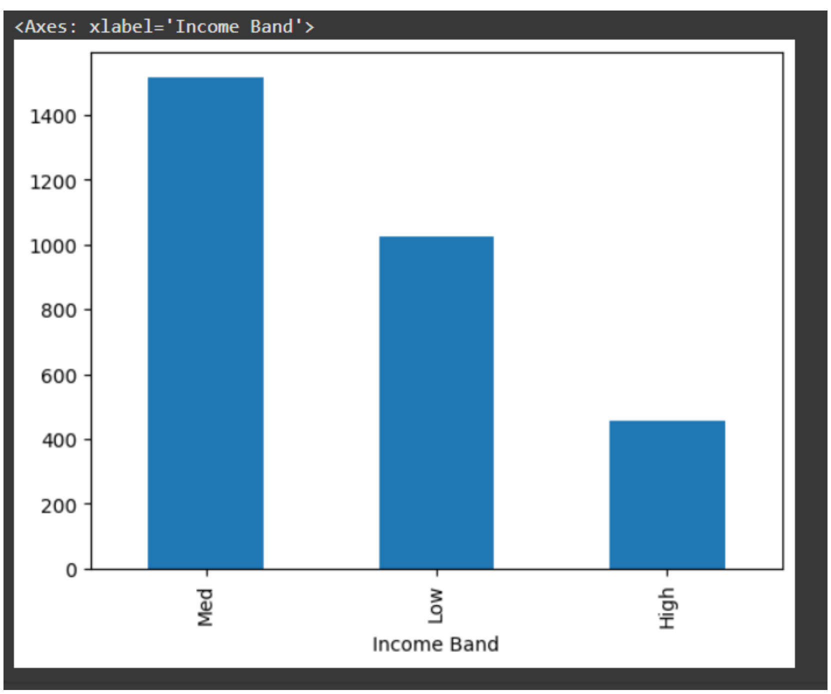
Fewer clients belonged to the High Income Band, which were high-net-worth clients.

Low-income clients formed a moderate percentage and were crucial for loan risk assessment.

Code:

```
df['Income Band'].value_counts().plot(kind='bar')
```

Figure 1: Distribution of Clients Across Income Bands



Categorical Feature Distributions

Count plots were generated for significant categorical variables like Gender, Banking Relationship, Fee Structure, Loyalty Classification, and Occupation.

Key observations are:

Gender distribution was evenly balanced, though slightly in favor of male clients.

A very large number of clients were assigned a Mid Fee Structure.

Clients belonging to the category "High Loyalty" were less than those belonging to "Medium" or "Low Loyalty".

Numerical Feature Distributions

Histograms were plotted for key financial variables including Estimated Income, Bank Deposits, Bank Loans, Checking Accounts, and Credit Card Balances.

Observations:

Estimated Income was right skewed, with most of the clients having incomes in the range $100,000 to $300,000.

Bank Deposits and Saving Accounts had fairly normal distributions with minimal right skew.

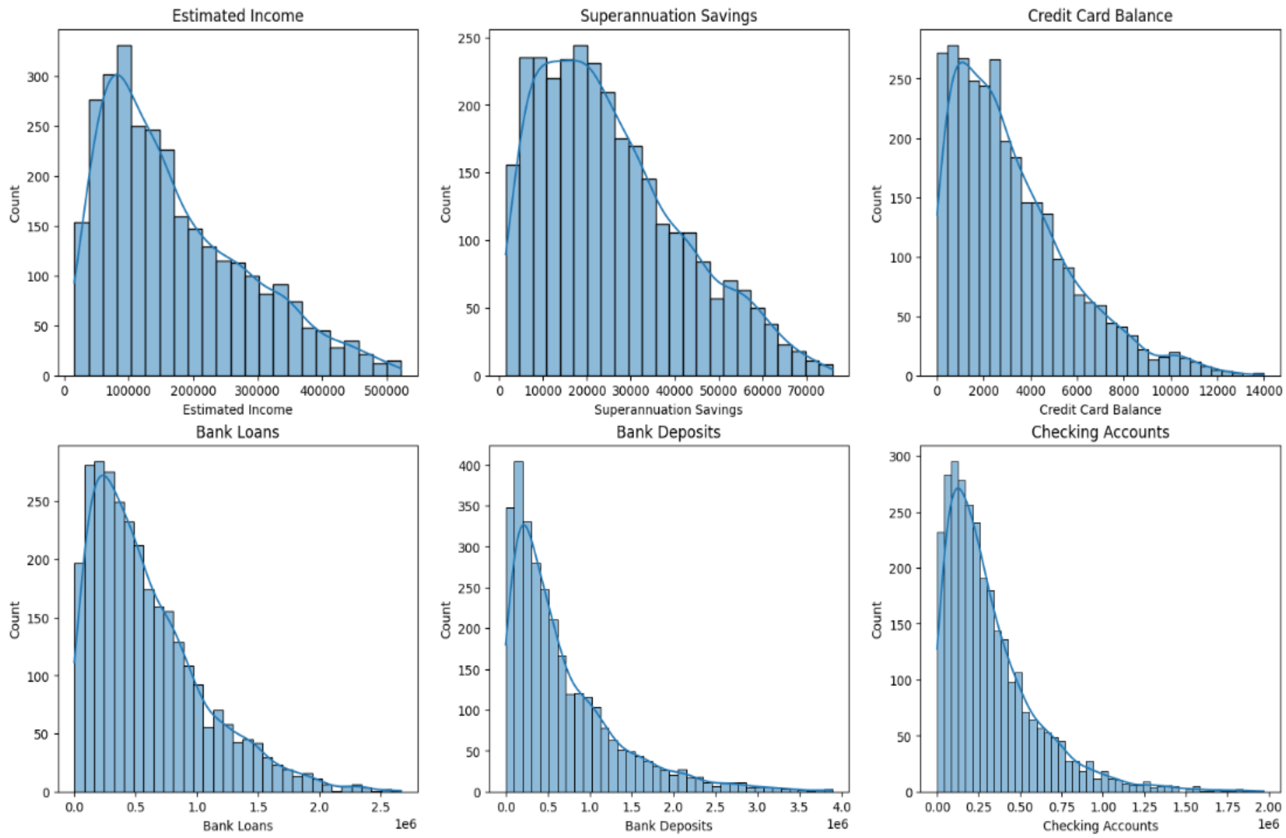
Credit Card Balances had an enormous spread with some of the clients having substantial balances.

Code:

```
numerical_cols = ['Estimated Income', 'Superannuation Savings', 'Credit Card Balance', 'Bank Loans', 'Bank Deposits', 'Checking Accounts', 'Saving Accounts', 'Foreign Currency Account', 'Business Lending']

# Univariate analysis and visualization
plt.figure(figsize=(18,22))
for i,col in enumerate(numerical_cols):
  plt.subplot(4,3,i+1)
  sns.histplot(df[col],kde=True)
  plt.title(col)
plt.show()
```

Figure 2: Distribution of Numerical Financial Features



2. Relationship and Trend Analysis

Feature Relationships (Scatter Plots)

Notable financial features were plotted with relationships between them (Scatter Plots)

Regression plots (regplots) were utilized in plotting key financial features in a bid to observe linear relations:

Bank Deposits vs. Saving Accounts depicted a positive linear trend, confirming that clients who have more saving also tend to have higher deposit balance.

Business Lending vs. Bank Loans similarly depicted a positive relationship, portraying that clients with more bank loans also tended to borrow more businesses.

These two pairs are also important when doing customer financial health profiling.

Code:

```
pairs_to_plot = [
    ('Bank Deposits', 'Saving Accounts'),
```

```
    ('Checking Accounts', 'Saving Accounts'),
    ('Checking Accounts', 'Foreign Currency Account'),
    ('Age', 'Superannuation Savings'),
    ('Estimated Income', 'Checking Accounts'),
    ('Bank Loans', 'Credit Card Balance'),
    ('Business Lending', 'Bank Loans'),
]

for x_col, y_col in pairs_to_plot:
    plt.figure(figsize=(8, 6))
    sns.regplot(
        data=df,
        x=x_col,
        y=y_col,
        scatter_kws={'alpha': 0.4},     # semi-transparent points
        line_kws={'color': 'red'}       # best-fit line color
    )
    plt.title(f'Relationship between {x_col} and {y_col}', fontsize=14)
    plt.xlabel(x_col, fontsize=12)
    plt.ylabel(y_col, fontsize=12)
    plt.tight_layout()
    plt.show()
```

Figure 3: Bank Deposits vs Saving Accounts

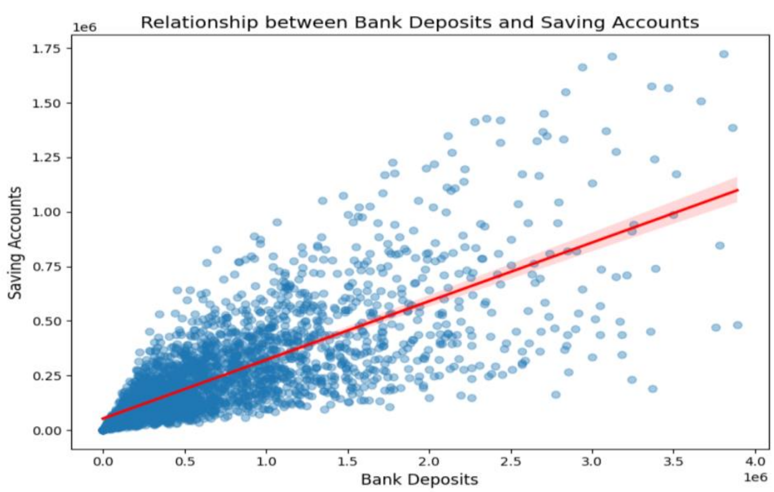
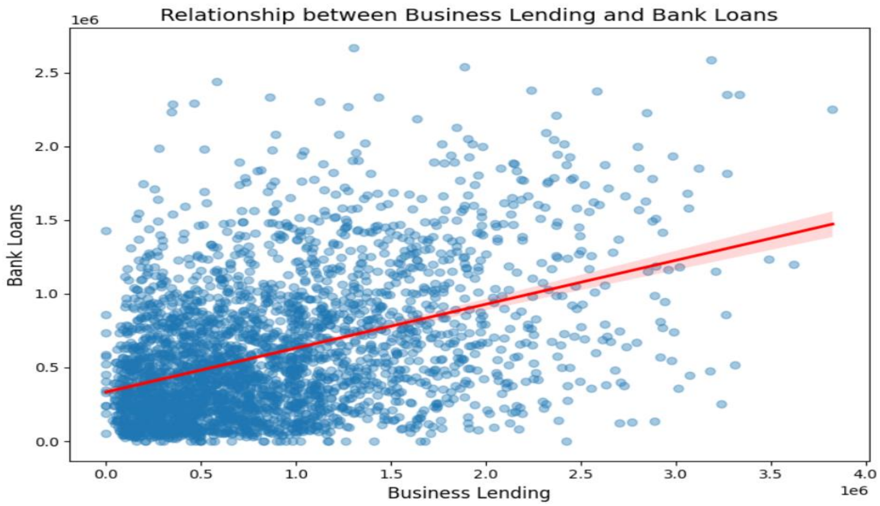


Figure 4: Bank Loans vs Business Lending



3. Correlation Analysis

We created a correlation heatmap in a bid to find out how the different numeric finance variables were associated with one another. Some noteworthy observations from the correlation matrix are:

Bank Deposits and Saving Accounts were considerably positively correlated.

Checking Accounts was moderately positively correlated with Bank Deposits.

Bank Loans and Credit Card Balances were weakly positively correlated, indicating customers who took loans also had higher card balances.

The heatmap played a key role in determining influential variable pairs that were driving financial behavior and risk classification.

Code:

```
numerical_cols = ['Estimated Income', 'Superannuation Savings', 'Credit Card Balance', 'Bank Loans', 'Bank Deposits', 'Checking Accounts', 'Saving Accounts', 'Foreign Currency Account', 'Business Lending']

correlation_matrix = df[numerical_cols].corr()

plt.figure(figsize=(12,12))
sns.heatmap(correlation_matrix, annot=True, cmap='crest', fmt=".2f")
plt.title("Correlation Matrix")
plt.show()
```

Figure 5: Major Financial Variables Correlation Matrix

Summary of EDA Findings

EDA effectively uncovered key financial trends and customer behaviors

Mid-Income customers are the largest customer segment and are most critical for loan targeting.

High Income customers have high loans but are typically followed by big deposits.

There are close associations between saving behavior and having deposits.

Certain financial segments, like Bank Loans and Business Lending, are interdependent and need to be assessed together under risk profiling.

These insights guided the development of KPIs and helped to create dashboards that focus on the principal risk drivers and client opportunities of the bank.

9. DASHBOARD AND KPI DESIGN

The final phase of this project was to design an interactive business intelligence solution in Power BI. The dashboard enables real-time visual understanding of customer banking behavior by displaying financial KPIs such as loans, deposits, credit activity, and account balances. It gives users the ability to apply filters based on gender, banking relationship, investment advisor, and year to immediately change the insights displayed.

The dashboard is divided into four principal pages:

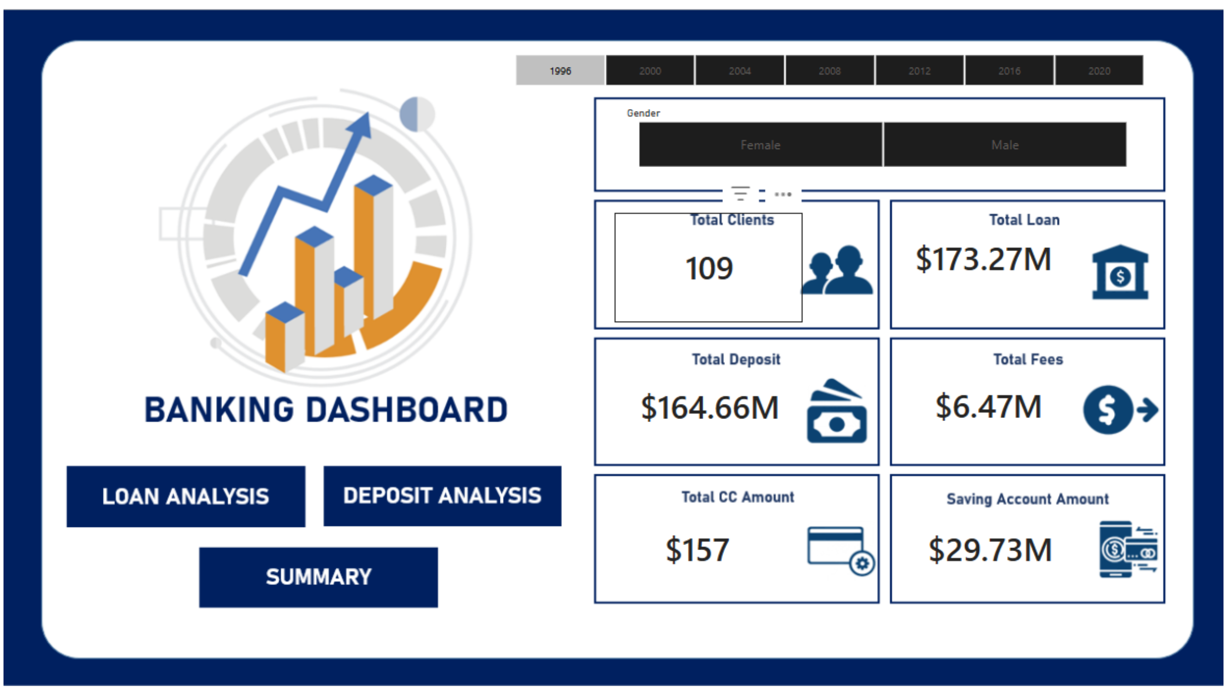
Home Page
Loan Analysis
Deposit Analysis
KPI Summary
Each of them is focused on a specific client banking behavior facet, providing actionable information to bank decision-makers and analysts.

1. Home Page Overview

Home Page is the principal navigation page of the dashboard. It includes interactive buttons for quick navigation to Loan Analysis, Deposit Analysis, and Summary pages. Additionally, slicers allow users to filter by: Gender, Year of Bank Joining

Looking at, higher-level parameters like Total Clients, Total Loan, Total Deposit, Total Credit Card Amount, Total Fees, and Saving Account Amount are showcased on the homepage.

Figure: Power BI Home Page



The interface works to facilitate efficiency in navigation by users and performance viewing without the necessary technical background.

2. Loan Analysis Dashboard

Loan Analysis Page is concerned with credit and its concentration along various aspects of clients.

Top features are:

Loan Categories: Bank Loan, Business Lending, Credit Cards, Visual Charts:

Bar chart: Bank Loan by Relationship Type

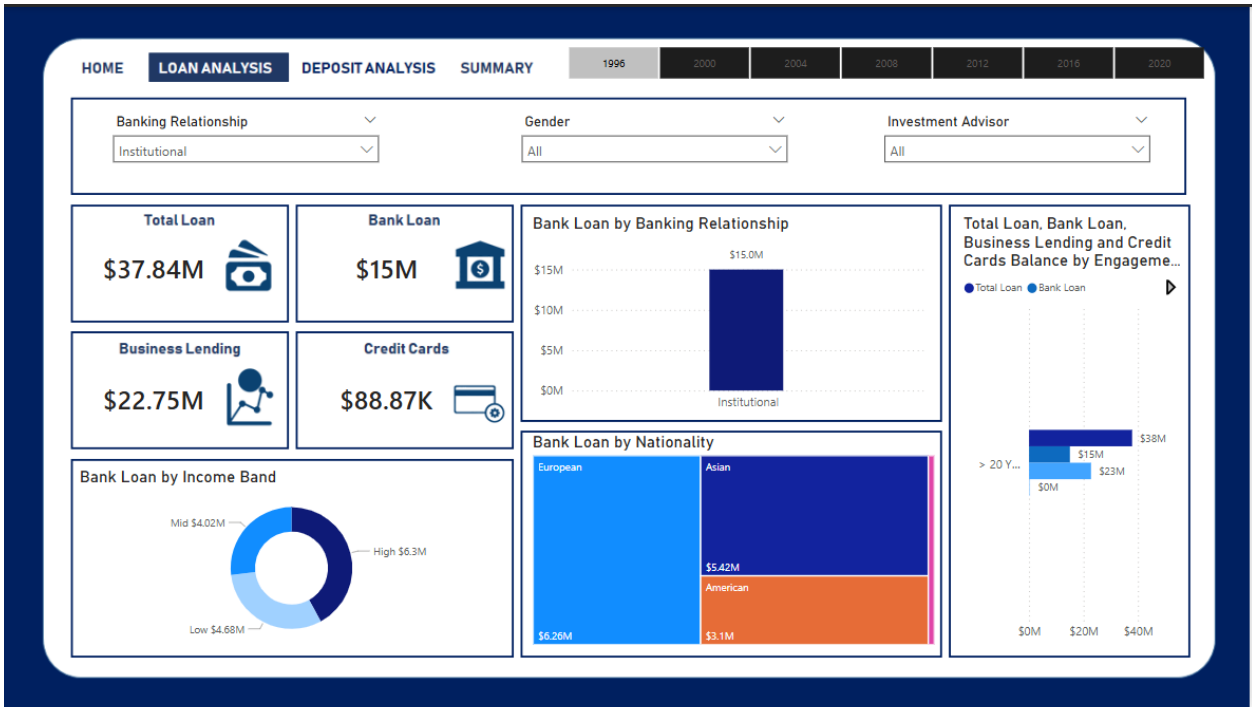
Donut chart: Bank Loan by Income Band

Tree map: Bank Loan by Nationality

Comparative bar: Loan amount vs. Bank Loan by engagement timeframe

The above page helps finance analysts to compute which parts of clients possess the majority of loans and also identify concentration risk.

Figure: Loan Analysis Dashboard



3. Deposit Analysis Dashboard

The Deposit Analysis Page presents analysis of how funds are spread across deposit types and client profiles. It incorporates the following elements:

Deposit Metrics: Total Deposit, Bank Deposit, Saving Account Amount, Checking Account Amount, Foreign Currency Balances

Visual Charts:

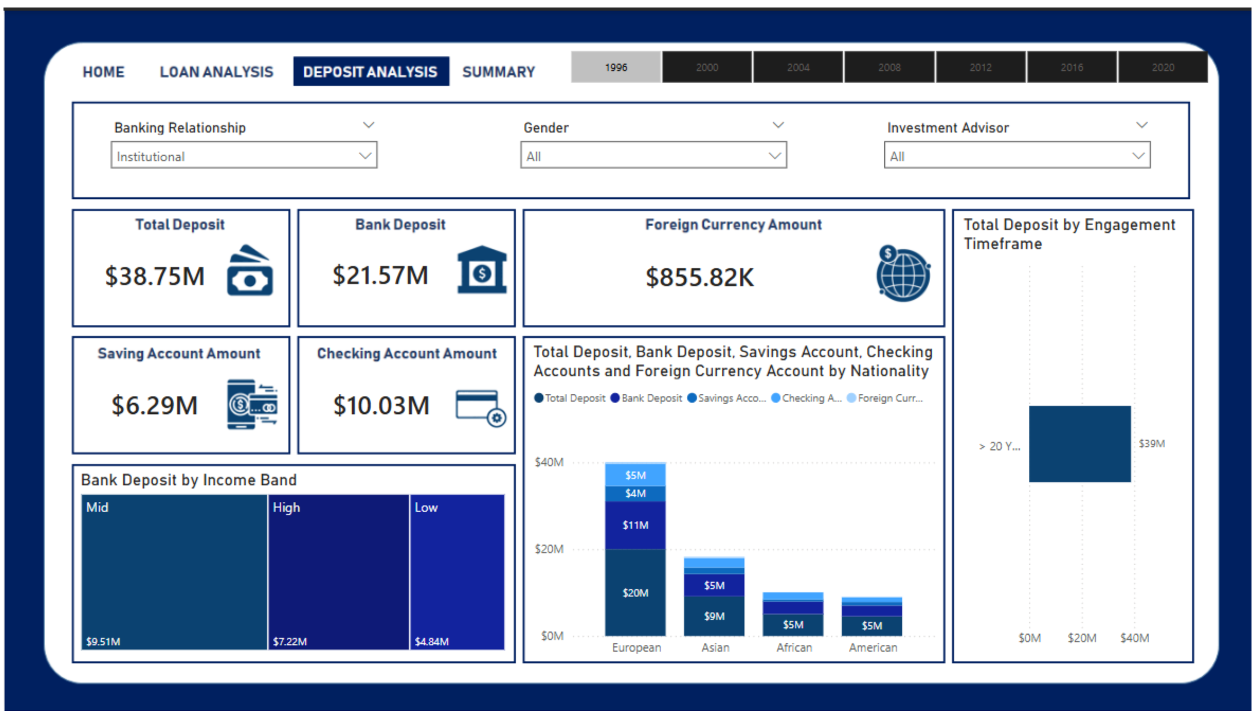
Tree map: Bank Deposit by Income Band

Bar chart: Nationality-wise distribution of deposits

Deposit value by Engagement Timeframe

This dashboard assists in determining where deposits are concentrated and what kind of customer segments is creating the most deposit growth.

Figure: Deposit Analysis Dashboard

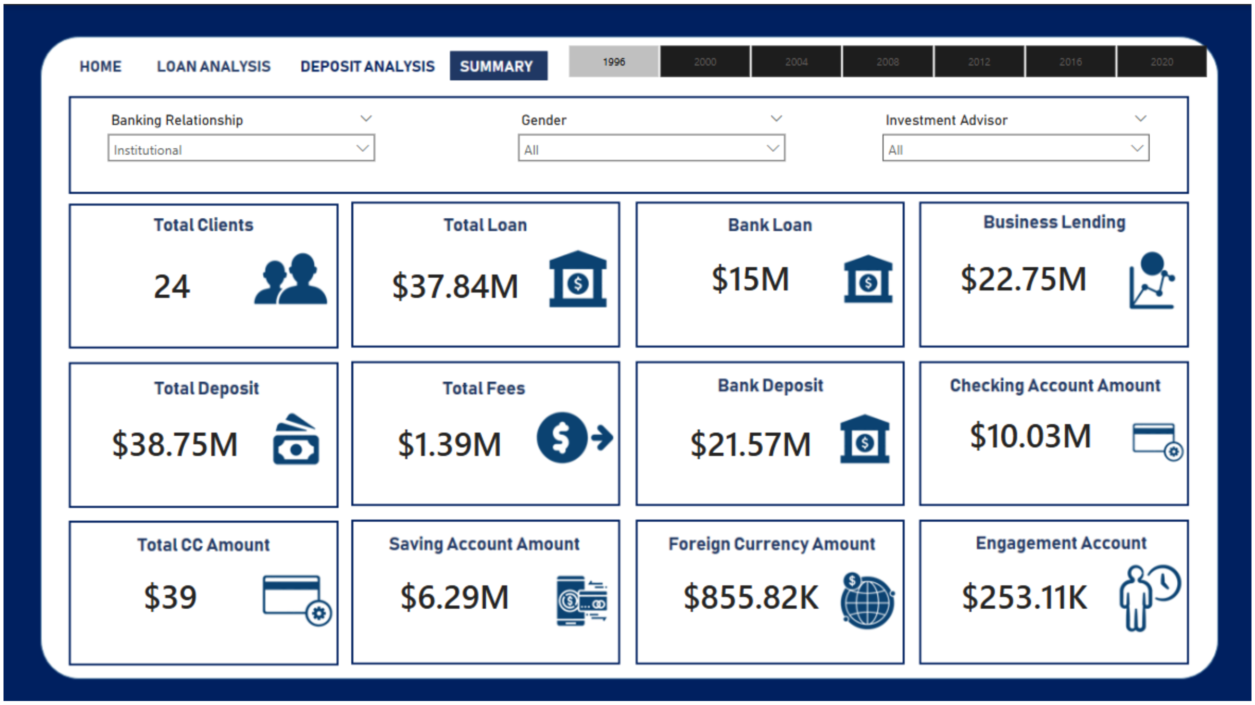


4. KPI Summary Dashboard

The Summary Page presents a one-glance summary of key banking performance metrics for selected filters (i.e., banking relationship or sex). The visually organized layout includes:

Total Clients, Total Loan, Bank Loan, Business Lending, Total Deposit, Total Fees, Credit Card Amount, Saving and Checking Account Amounts, Foreign Currency Amount, Engagement Account Value

Figure: KPI Summary Dashboard



This dashboard is designed to present analysts with an overall impression of performance based on both client engagement and finance distribution.

5. Interactive Features and Filters

Interactivity was incorporated through slicers on Gender, Banking Relationship, Investment Advisor, and Year.

Visuals update dynamically based on selected filters, and tooltips were added for key KPI cards.

These features allow users to segment customers in real time, examine loan and deposit performance, and identify trends across different demographics and relationships.

Note: Drill-through pages were intentionally excluded from this version in order to prioritize clarity, simplicity, and user access of main visuals.

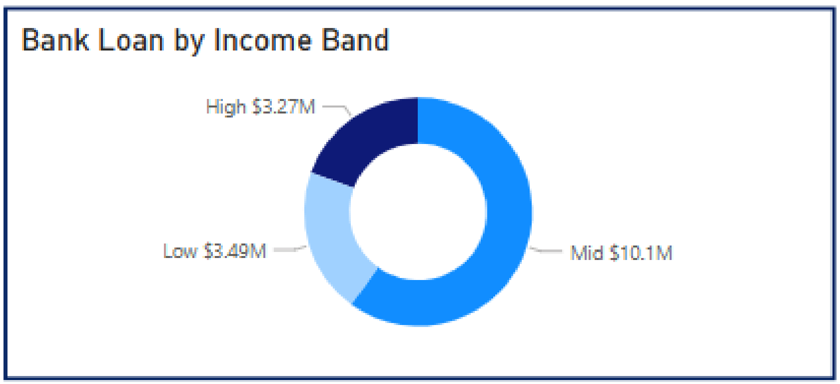
10. RESULTS AND INSIGHTS

This section will highlight the most significant findings of the Power BI dashboard. These findings were established based on a number of various visual components, KPIs, and filters used in the report. They encompass the trends between income and loan behavior, national-level risk distribution, deposit patterns, and customer engagement patterns-all extremely vital for formulating targeted strategies from banking organizations.

1. Income Band and Loan Behavior

The first major observation is related to the distribution of the loan amount between the income bands. The following chart illustrates that, for example, clients in the Mid Income Band hold the most loan volume-which is above $10M-followed by Low and High Income Bands, respectively.

Figure: Bank Loan Distribution by Income Band



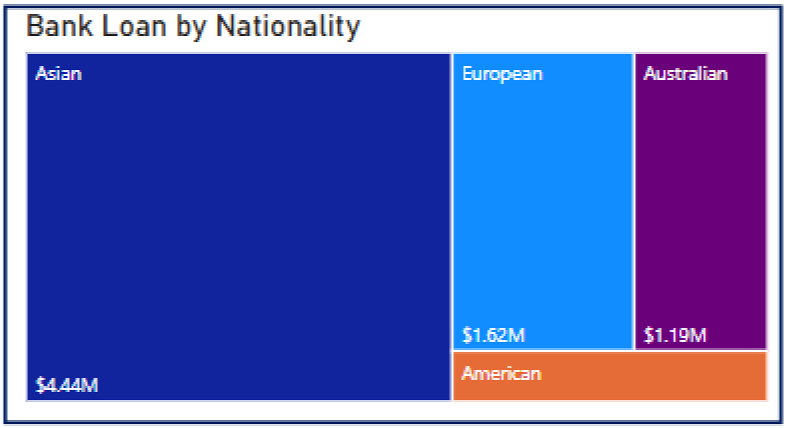
Interpretation:

While High Income customers have the potential for large loans, Mid Income customers are apparently more dynamic in availing loans. This could be reflective of both approval rates that are higher and active financial needs. Banks must monitor this segment for opportunity in loans as well as risk exposure.

2. Loan Risk by Nationality

Another key observation is from the analysis of loan data nationality-wise. The tree map shows that Asian customers have the largest loan amount ($4.44M), much larger compared to other categories.

Figure: Bank Loan by Nationality



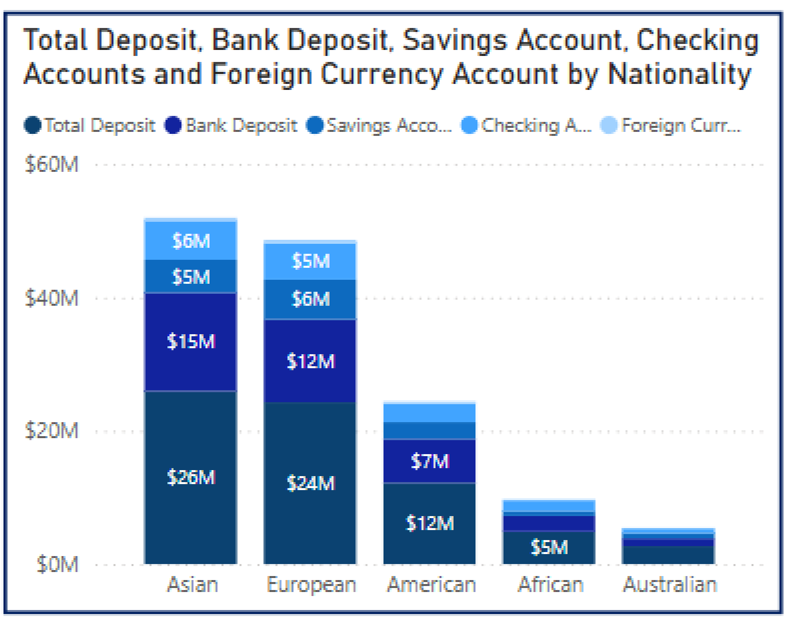
Interpretation:

This indicates that loan exposure is disproportionately concentrated on specific nationalities. Although it is not inherently negative, this pattern merits closer monitoring for ensuring that risks are well distributed. It also informs the credit policy to target regions or national groups that experience higher loan demand.

3. Deposit Patterns by Nationality

The following stacked bar chart shows a closer look at deposits among checking, savings, and foreign currency accounts of various nationalities. Asian and European customers constitute the highest deposit amount, comprising over $45M in total.

Figure: Total Deposits by Nationality and Account Type



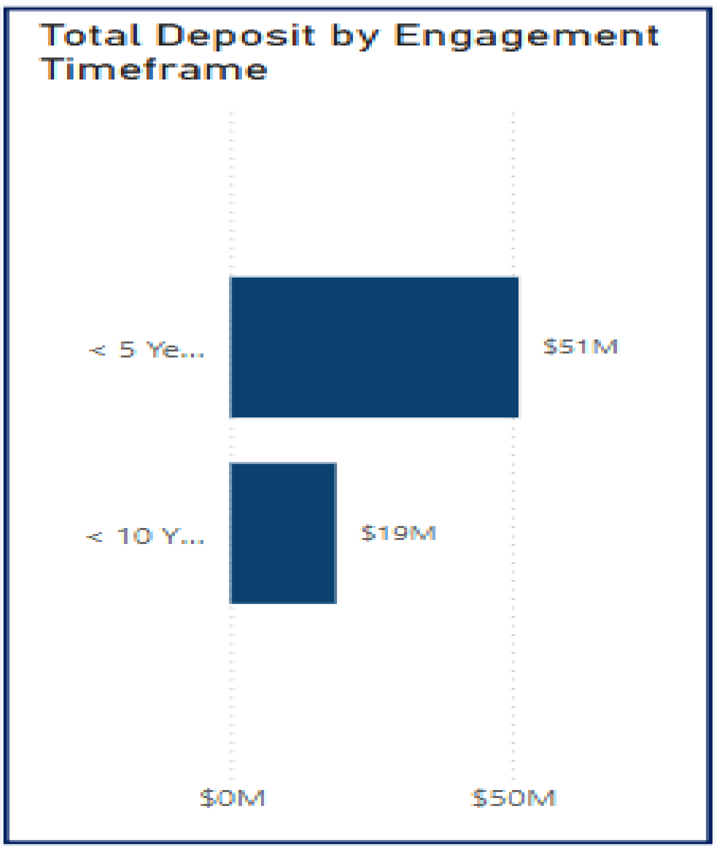
Interpretation:

These client groups exhibit saving and deposit behavior, which is low-risk and high-value for banks. This client segment needs to be retained and expanded by offering customized saving and investment plans.

4. Engagement Timeframe vs. Deposits

Engagement time also has an impact on deposit activity. As the bar chart illustrates, customers with <5 years of engagement accumulate over $51M in deposits, significantly more than the longer tenured customers.

Figure: Total Deposits by Engagement Timeframe



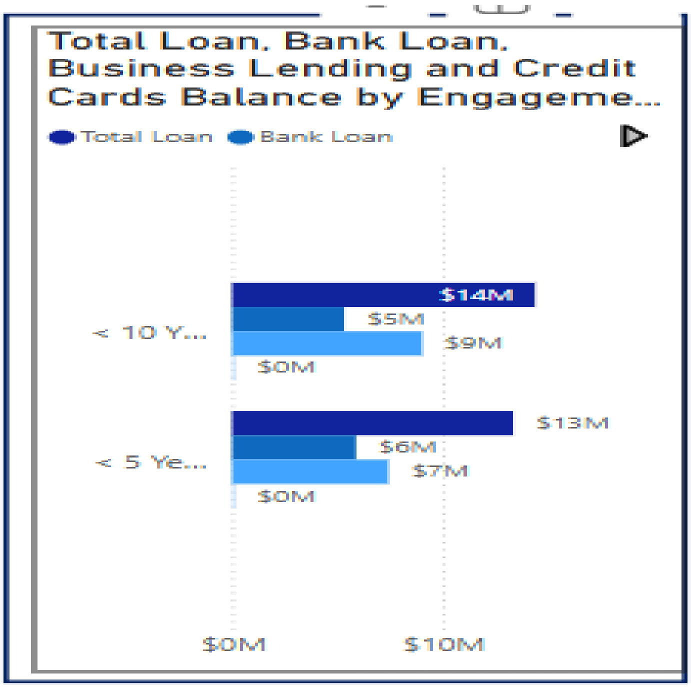
Interpretation:

This trend is counterintuitive but significant. It implies that newer clients are very active in depositing funds, due either to onboarding offers or short-term goals. Banks need to try and retain new clients by keeping this deposit momentum.

5. Loan Behavior by Tenure in Engagement

The last visual presents loan distribution by engagement duration. Customers with less than 5 years of tenure have greater overall loan balances ($13M) than those with a longer engagement.

Figure: Total Loan and Bank Loan by Engagement Timeframe



Interpretation:

Newer clients have more credit-seeking behavior. Though this shows good engagement and service usage, it also brings early-stage credit risk. A tight watch at the onboarding and approval process is required for the same.

11. BUSINESS VALUE DELIVERED

The Banking Profiling System developed in this project delivers tremendous value to banks by transforming raw customer data into business intelligence. Through a series of interactive dashboards, KPIs, and filtering capabilities, banks are now in a position to make informed decisions in a host of business functions — including credit risk management, customer segmentation, deposit forecasting, and strategic client engagement.

1. Improved Loan Risk Management

The dashboard enables banks to determine loan segments with high risk by examining loan volumes over income bands, nationalities, and engagement durations. By emphasizing trends such as increased borrowing by new or middle-income customers, decision-makers can take early steps to revise lending criteria, establish targeted loan caps, or heighten surveillance of high-exposure cohorts.

2. Improved Client Segmentation

By Gender, Banking Relationship, Advisor, and Income Band slicers and filters, the dashboard enables banks to segment the client base in real time.

This helps in: Developing customized products, Identifying low-engagement clients, Offering loyalty rewards or premium services for high-value customers

3. Strategic Deposit Planning

The deposit analytical tools help to gain an understanding of which clients make the largest contributions to checking and saving accounts.

This helps banks:

Tailor deposit programs or investment accounts. Initiate marketing campaigns to grow balances in under-performing segments. Prioritize onboarding new customers with positive initial deposit activity

4. Faster, Data-Driven Decision Making

With automated calculation of key metrics like Total Loan, Deposits, Fees, and Credit Balances, the dashboard significantly reduces financial report preparation time. Executives and branch managers have key KPIs at their fingertips in real time, allowing for rapid operational adjustments and earlier response to market shifts.

5. Simple to Use and Scalable

The Power BI platform works seamlessly with MySQL, and the dashboard becomes scalable from the departmental to the branch levels. With the use of intuitive visuals and reduced manual inputs, it becomes simple for even non-technical users to derive insights easily, thus enhancing organizational productivity and collaboration among teams.

Overall, this system allows banks to move from static reporting to interactive, insight-generating analysis. It supports both tactical decision-making—e.g., individual loan approval—and strategic planning—e.g., informing which segments to defend or grow. As a result, the dashboard delivers strong business value by reducing financial risk, improving client retention, and facilitating growth through data insights.

12. CHALLENGES FACED AND LIMITATIONS

Several challenges were faced in the creation of this project, particularly at the data preparation and dashboard design stages. While these problems were overcome successfully, they raise important considerations for future real-time banking analytics implementations.

1.Data Cleaning and Preparation Problems

The data set required critical restructuring to be imported into MySQL. Columns such as "Engagement Days," "Income Band," and "Banking Relationship" had to be normalized manually through consistency.

Unreliable or missing information in columns such as processing fees, account levels, or advisor allocations required logical imputation or filtering.

Having relations between numerous tables (e.g., Clients, Bank Deposits, Loans) involved creating appropriate primary and foreign keys as well as relational integrity.

In Python transformation, some numeric columns required to be converted from string types and removed outliers without losing valuable business insights.

2. Dashboard Development Constraints

Splitting large datasets into interactive visualizations within Power BI required performance optimization. Excessive filtering on dimensions at times made dashboard sluggish.

Some KPIs such as "Engagement Account" were computed using assumptions and hardcoded formulas, which may vary in real banking settings.

The current system relies on manual Excel imports to MySQL, which is not feasible for real-time.

3. Dataset Limitations

The data set for the project was artificially generated and combined from a wide range of banking-related sources, rather than live institutional data.

Realistic in form and distribution, but potentially short on depicting real customer behavior, regulatory constraints, or production system surprises in anomalies, artificial data can nevertheless be employed in most instances.

Certain variables such as credit score, default history, and transaction history were not included in it, which would have added complexity to risk analytics.

13. FUTURE WORK

While the system presently constitutes a solid foundation for descriptive banking analytics, there are numerous areas where the solution could be strengthened and augmented as discussed below:

1. Predictive Modeling

Integration of machine learning algorithms would allow the system to:

Make loan default forecasts based on demographic and financial behavior

Forecast future customer churn or deposit patterns

Offer credit scoring recommendations directly within the dashboard

Python libraries like Scikit-learn, XGBoost, and LightGBM could be used to train and deploy these models on historical data.

2. Real-Time Dashboard Integration

The Power BI dashboard is presently built on static data fetched from MySQL. Future releases could:

Use real-time data streaming from production databases or APIs

Have automatic refresh schedules enforced

Support push notifications or alerts on threshold breaches (e.g., loan levels)

This would add significantly to the usefulness of the system in real-time decision-making.

3. Enhanced Client Segmentation

Further granularity is achievable by dividing clients based on:

Spending patterns, frequency of transactions, or loyalty factor

Product use (loans + deposits + investments)

Behavioral clusters from unsupervised learning (e.g., K-means)

This would make banks able to deliver more individualized services and streamline client lifecycle management.

14. CONCLUSION

The project was successful in developing an efficient Banking Risk Profiling System by combining organized client and financial data, cleaning and formatting it with the help of Python, and designing interactive dashboards in Power BI. The solution provides banks with a real-time insight into client activity, financial interaction, and loan risk, enabling them to take data-driven decisions based on facts and not assumptions.

Analytical process derived key insights such as loan behavior income-wise, risk exposure on a nationality basis, deposit trends, and performance on an engagement basis. These insights, being well-designed KPIs and charts, enable stakeholders to understand and respond to dynamic customer needs at a faster pace.

Such data-driven systems support banks in:

Minimizing risk through the detection of at-risk client segments

Segmenting clients and tailoring services

Optimizing profitability through target-specific marketing and credit planning

Improving compliance and reporting through real-time dashboards

Through the adoption of advanced data analytics, banks can shift from traditional, reactive decision-making to proactive and predictive decision-making that improves operating efficiency and customer delight. This project provides a foundation for more advanced solutions such as predictive credit scoring, automated notifications, and intelligent advisory systems.

15. REFERENCES

https://www.kaggle.com/datasets/teajay/global-bank-customer-churn-prediction

https://www.kaggle.com/datasets/ruchi798/bank-loan-dataset

https://www.kaggle.com/datasets/sanjanabasu/banking-dataset-marketing-targets

Pandas Official Documentation

https://pandas.pydata.org/docs/

Seaborn Visualization Library

https://seaborn.pydata.org/