Banking Project

Problem Statement:

To study and apply risk analytics concepts in the banking domain by analyzing customer financial data and building analytical and predictive models that help identify and classify customer risk levels, thereby supporting effective risk management and lending decisions.

Solution:

The solution combines Power BI dashboards for descriptive analysis with machine learning models for predictive risk assessment. While dashboards provide insights into customer behaviour, the predictive model classifies applicants into risk categories, supporting informed and data-driven loan approval decisions.

About Dataset:

This dataset basically contains information about bank details ,various client details which consists of multiple tables which are interlinked with each other through keys like primary key and foreign key.

The various tables are Banking Relationship, Client-Banking, Gender and Investment Advisor

Database Connectivity:

The banking dataset was stored in a MySQL database and accessed using MySQL Workbench. A secure connection was established between MySQL and Jupyter Notebook using Python's database connectivity libraries. This enabled direct querying and data retrieval from the database into the analysis environment. The extracted data was then loaded into Pandas DataFrames for further analysis and processing.

```
!pip install pymysql
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pymysql in c:\users\soham\appdata\roaming\python\python312\site-packages (1.1.2)

[notice] A new release of pip is available: 25.2 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
import pymysql

conn = pymysql.connect(
    host="127.0.0.1",
    user="root",
    password="",
    database="banking_Case"
)

print("Connected successfully 🎉")
```

```
Connected successfully 🎉
```

EDA (Exploratory Data Analysis):

Exploratory Data Analysis was conducted to understand the structure, quality, and underlying patterns within the banking customer dataset. This step was essential to identify data characteristics, detect anomalies, and guide subsequent feature engineering and modelling decisions.

Initially, the dataset structure was examined using functions such as info(), describe(), and shape to understand the number of records, data types, presence of missing values, and overall data distribution. This helped in identifying numerical and categorical variables and assessing data completeness.

Univariate analysis was performed to analyse the distribution of individual variables such as age, income, account balances, loans, and credit card usage. Visualizations including histograms and bar charts were used to identify skewness, outliers, and common value ranges.

Bivariate analysis was carried out to study relationships between key variables and the target-related features. Comparisons between financial attributes and risk-related indicators helped uncover patterns such as higher credit utilization or loan amounts being associated with increased risk levels.

Correlation analysis was conducted using heatmaps to identify relationships among numerical variables and detect multicollinearity. This analysis provided insights into how different financial features interact with each other and informed feature selection and engineering decisions.

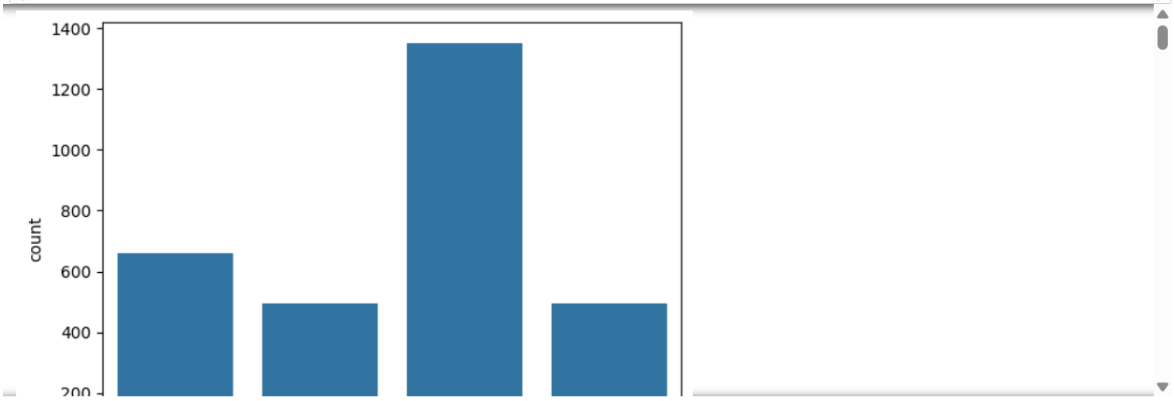
Overall, the EDA phase provided critical insights into customer financial behaviour and laid the foundation for dashboard creation and machine learning model development.

```
# generating descriptive statistics:
df.describe()
```

	Age	Location ID	Estimated Income	Superannuation Savings	Amount of Credit Cards	Credit Card Balance	Bank Loans	Bank Deposits	Checking Accounts	Saving Accounts	
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3.000000e+03	3.000000e+03	3.000000e+03	3.000000e+03	3000.
mean	51.039667	21563.323000	171305.034263	25531.599673	1.463667	3176.206943	5.913862e+05	6.715602e+05	3.210929e+05	2.329084e+05	29883.
std	19.854760	12462.273017	111935.808209	16259.950770	0.676387	2497.094709	4.575570e+05	6.457169e+05	2.820796e+05	2.300078e+05	23109.
min	17.000000	12.000000	15919.480000	1482.030000	1.000000	1.170000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	45.
25%	34.000000	10803.500000	82906.595000	12513.775000	1.000000	1236.630000	2.396281e+05	2.044004e+05	1.199475e+05	7.479440e+04	11916.
50%	51.000000	21129.500000	142313.480000	22357.355000	1.000000	2560.805000	4.797934e+05	4.633165e+05	2.428157e+05	1.640866e+05	24341.
75%	69.000000	32054.500000	242290.305000	35464.740000	2.000000	4522.632500	8.258130e+05	9.427546e+05	4.348749e+05	3.155750e+05	41966.
max	85.000000	43369.000000	522330.260000	75963.900000	3.000000	13991.990000	2.667557e+06	3.890598e+06	1.969923e+06	1.724118e+06	124704.

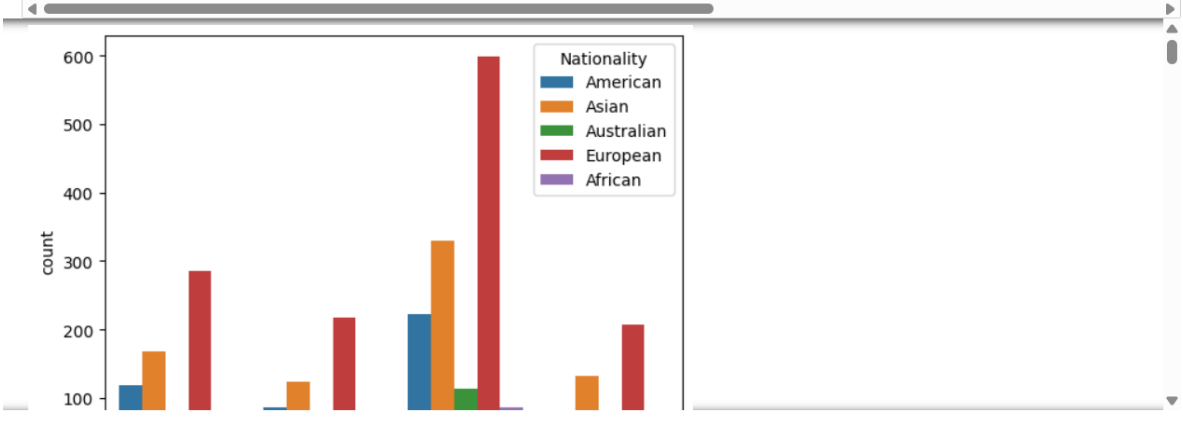
Univarite analysis

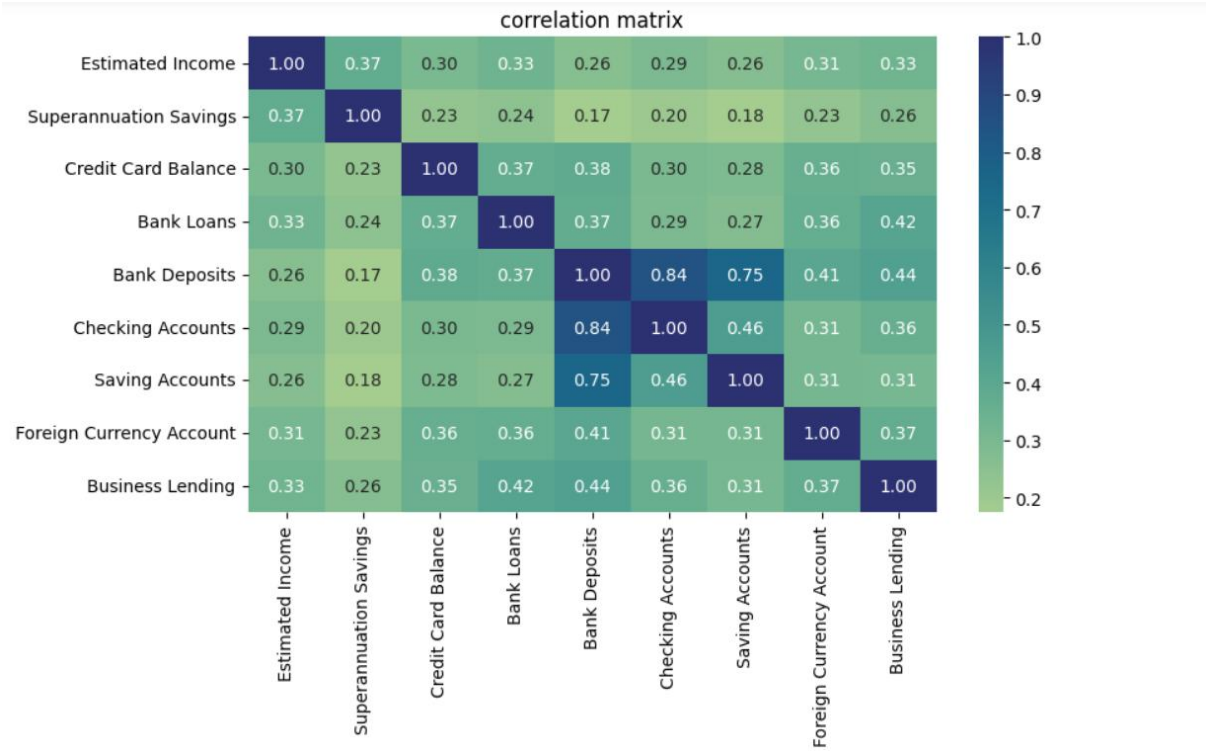
```
for i , predictor in enumerate(df[['BRId','GenderId','IAId','Amount of Credit Cards','Nationality','Occupation','Fee Structure','
    plt.figure(i)
    sns.countplot(data=df,x=predictor)
```



Bivariate analysis

```
for i , predictor in enumerate(df[['BRId','GenderId','IAId','Amount of Credit Cards','Nationality','Occupation','Fee Structure','
    plt.figure(i)
    sns.countplot(data=df,x=predictor,hue='Nationality')
```





Power BI Dashboard:

After completing exploratory data analysis, the cleaned dataset was imported into Power BI for interactive visualization and business intelligence reporting. Power Query was used to perform additional data transformations such as data type corrections, column renaming, and minor preprocessing to ensure consistency and accuracy across visuals.

Custom measures and calculated fields were created using DAX to derive meaningful business metrics required for analysis. These measures enabled dynamic filtering, aggregation, and comparative analysis across different customer segments.

A comprehensive four-page interactive dashboard was developed to present insights in a structured and user-friendly manner. The dashboard consists of the following pages:

- Home: Provides a high-level overview of key customer and financial metrics, serving as a navigation and summary page.
- Loan Analysis: Focuses on customer loan behaviour, credit exposure, and lending-related trends to support loan risk assessment.
- Deposit Analysis: Analyzes customer deposit patterns, account balances, and savings behaviour to understand liquidity and customer value.
- Summary, Q&A: Presents consolidated insights and allows users to interactively query the data for quick answers and ad-hoc analysis.

The Power BI dashboard enables stakeholders to explore customer profiles, identify financial patterns, and support data-driven decision-making through intuitive visuals and interactive filters.

Calculated Functions:

Some of the new measures/columns created:

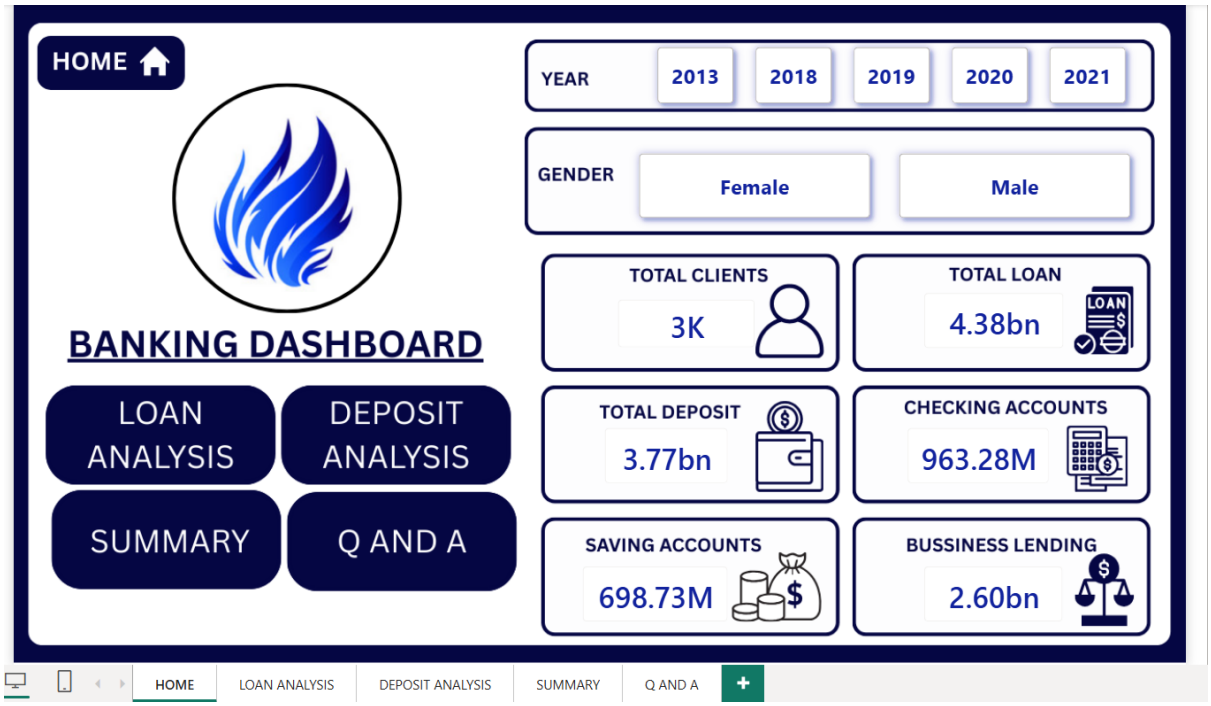
```
total deposit = SUM('banking_case customer'[Bank Deposits]) + SUM('banking_case customer'[Saving Accounts])+ SUM('banking_case customer'[Checking Accounts]) + SUM('banking_case customer'[Foreign Currency Account])
```

```
total loan = SUM('banking_case customer'[Bank Loans])+SUM('banking_case customer'[Business Lending]) + SUM('banking_case customer'[Credit Card Balance])
```

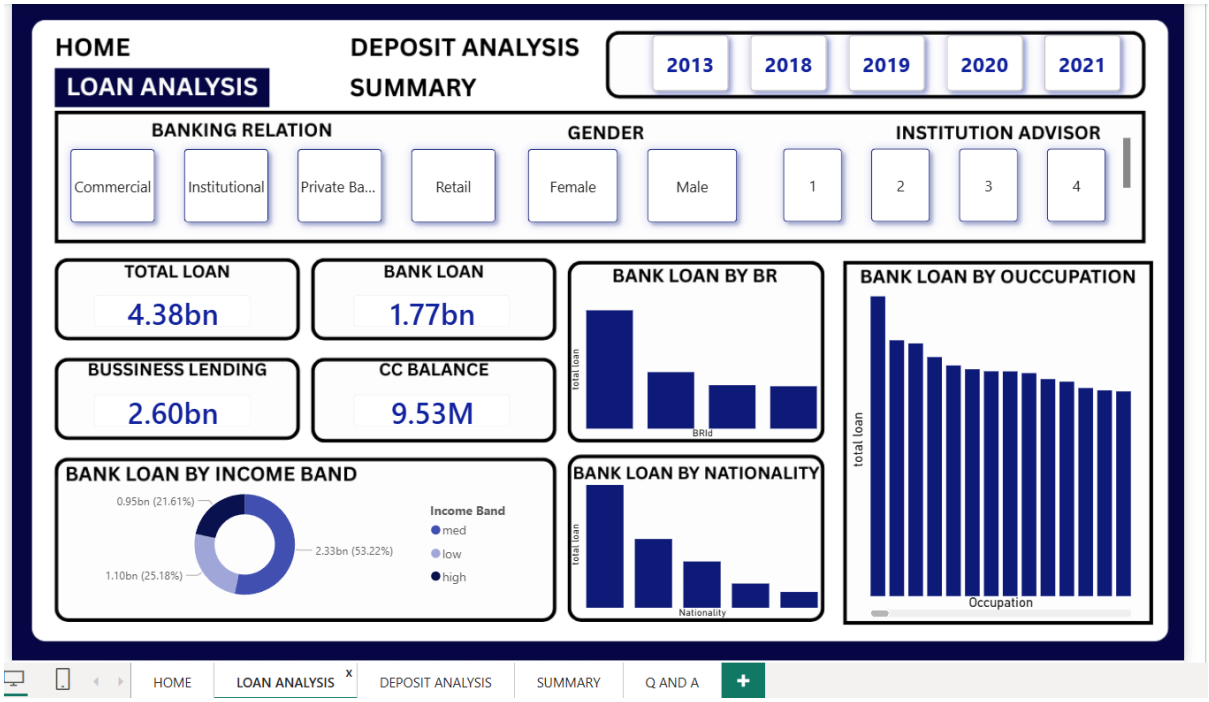
```
year of joining = YEAR('banking_case customer'[Joined Bank])
```

Power BI: Visualization And Result:

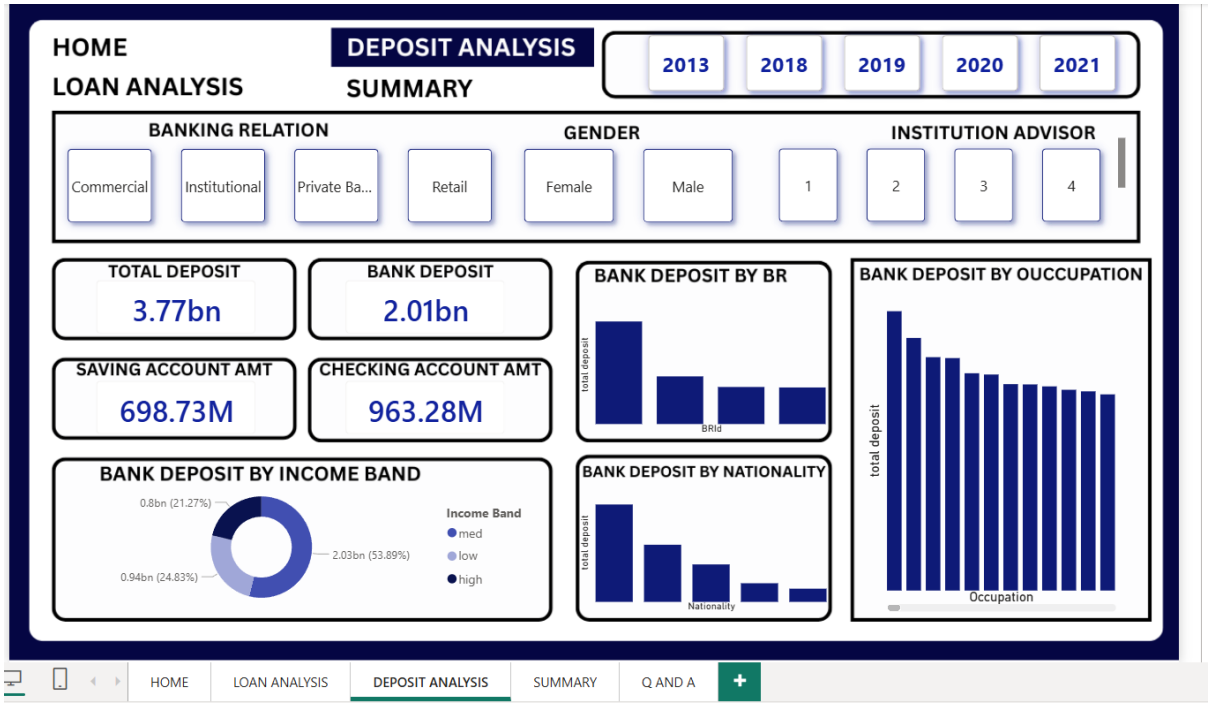
Home :



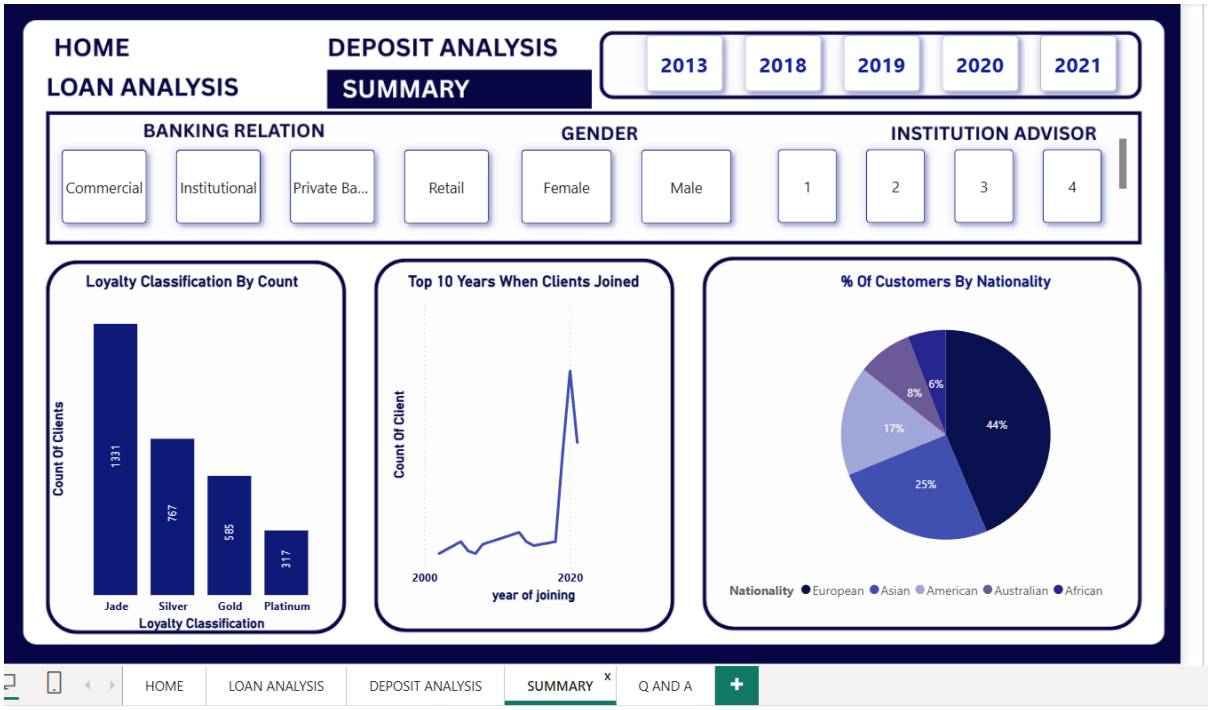
Loan Analysis:



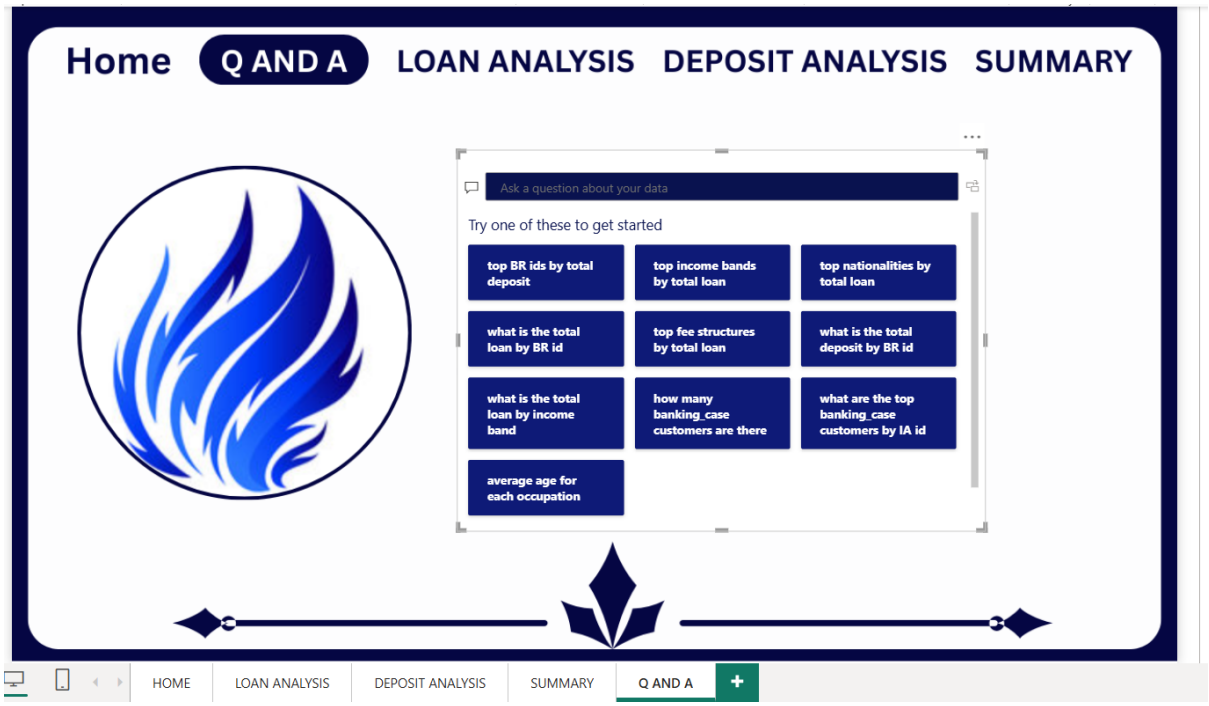
Deposit Analysis:



Summary Dashboard:


HOME
LOAN ANALYSIS
DEPOSIT ANALYSIS
SUMMARY
2013
2018
2019
2020
2021
BANKING RELATION
Commercial
Institutional
Private Ba...
Retail
GENDER
Female
Male
INSTITUTION ADVISOR
1
2
3
4
Loyalty Classification By Count
Count Of Clients
1331
767
585
317
Jade
Silver
Gold
Platinum
Loyalty Classification
Top 10 Years When Clients Joined
Count Of Client
2000
2020
year of joining
% Of Customers By Nationality
44%
25%
17%
8%
6%
Nationality
European
Asian
American
Australian
African
HOME
LOAN ANALYSIS
DEPOSIT ANALYSIS
SUMMARY
Q AND A

Q and A:


Home
Q AND A
LOAN ANALYSIS
DEPOSIT ANALYSIS
SUMMARY
Ask a question about your data
Try one of these to get started
top BR ids by total deposit
top income bands by total loan
top nationalities by total loan
what is the total loan by BR id
top fee structures by total loan
what is the total deposit by BR id
what is the total loan by income band
how many banking_case customers are there
what are the top banking_case customers by IA id
average age for each occupation
HOME
LOAN ANALYSIS
DEPOSIT ANALYSIS
SUMMARY
Q AND A

Machine Learning Implementation:

Machine Learning Methodology:

Following data analysis and visualization, a machine learning pipeline was developed to predict customer risk levels based on financial and demographic attributes. The objective of this phase was to build a reliable predictive model that could classify customers into predefined risk categories.

Feature Engineering:

Feature engineering was performed to enhance the predictive power of the dataset by creating meaningful financial indicators. New features such as **Total Balance**, **Credit Utilization**, and **Loan-to-Income Ratio** were derived from existing financial variables. These engineered features capture customer financial behaviour more effectively than raw variables.

Irrelevant and identifier-based columns were removed to prevent noise and data leakage during model training. The final feature set consisted of a balanced combination of demographic, account-related, and risk-related attributes.

Data Preparation & Encoding:

The dataset was divided into numerical and categorical features. Categorical variables were encoded using label encoding to convert them into numerical form suitable for machine learning algorithms. Numerical variables were standardized using feature scaling techniques to ensure uniform contribution across features.

Train-Test Split:

The prepared dataset was split into training and testing sets to evaluate model performance on unseen data. This ensured that the models generalized well and did not overfit to the training data.

Model Development:

The target variable selected for prediction was **Risk Weighting**, representing different levels of customer financial risk. Multiple machine learning algorithms were implemented and compared, including:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier

To address class imbalance in the target variable, oversampling techniques such as **SMOTE** were applied, improving the model's ability to learn minority risk classes.

Model Evaluation:

Models were evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC (One-vs-Rest for multi-class classification). Comparative analysis across models was conducted to identify the best-performing algorithm.

Among the evaluated models, **Gradient Boosting** demonstrated superior performance with balanced classification across risk categories and higher predictive reliability.

Prediction & Risk Scoring:

The final model was used to generate predictions on test data, providing both predicted risk classes and probability scores. These probability estimates enable more nuanced risk assessment, supporting practical applications such as customer risk profiling and lending decision support.

Outcome:

The machine learning component successfully transformed descriptive insights into actionable predictions, enabling proactive risk management and supporting data-driven financial decisions.

The machine learning phase transformed descriptive insights into predictive intelligence, enabling proactive and data-driven risk assessment.

```
dfc['Total_Balance'] = (
    dfc['Checking Accounts'] +
    dfc['Saving Accounts'] +
    dfc['Bank Deposits']
)
```

```
dfc['Credit_Utilization'] = (
    dfc['Credit Card Balance'] /
    (dfc['Amount of Credit Cards'] + 1)
)
```

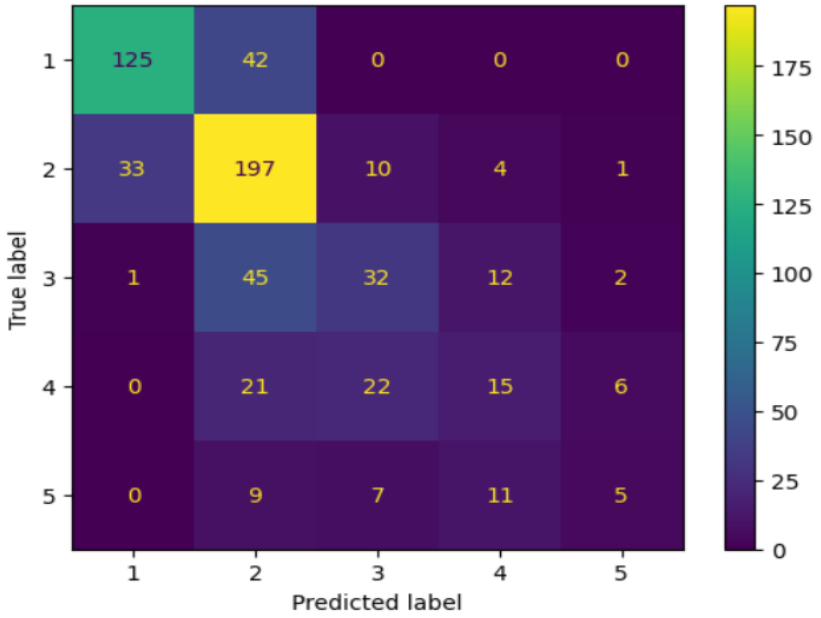
```
dfc['Loan_to_Income'] = dfc['Bank Loans'] / (dfc['Estimated Income'] + 1)
```

```
evaluate(gb_pipeline)
```

```
              precision    recall  f1-score   support

           1       0.79      0.75      0.77       167
           2       0.63      0.80      0.70       245
           3       0.45      0.35      0.39        92
           4       0.36      0.23      0.28        64
           5       0.36      0.16      0.22        32

    accuracy                           0.62       600
   macro avg       0.52      0.46      0.47       600
weighted avg       0.60      0.62      0.60       600

ROC-AUC: 0.8546724217082096
```

```
from sklearn.metrics import ConfusionMatrixDisplay
ConfusionMatrixDisplay.from_estimator(gb_pipeline, X_test, y_test)
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x20418cd2c10>
```



Conclusion:

This project successfully demonstrated an end-to-end data analytics and machine learning workflow applied to a banking customer dataset. Beginning with database connectivity using MySQL Workbench and transitioning through exploratory data analysis, business intelligence dashboarding, and predictive modelling, the project showcased how data can be transformed into actionable insights.

Exploratory Data Analysis provided a deep understanding of customer demographics, financial behaviour, and risk-related patterns. The Power BI dashboard enabled interactive and intuitive visualization of these insights, allowing stakeholders to analyse loan behaviour, deposits, and overall customer profiles. Finally, machine learning models were developed to predict customer risk levels, with Gradient Boosting emerging as the most effective model.

Overall, the project highlights the importance of combining descriptive analytics, visualization, and predictive modelling to support data-driven decision-making in the banking domain.

Improvements:

While the project achieved its objectives, certain improvements could further enhance its effectiveness and robustness:

- Advanced feature selection techniques could be applied to reduce redundancy and improve model interpretability.
- More sophisticated encoding methods, such as target encoding or one-hot encoding, could be explored for categorical variables.
- Hyperparameter tuning techniques could be used to further optimize model performance.
- Incorporation of additional evaluation metrics such as precision-recall curves could improve assessment for imbalanced risk classes.
- Enhanced dashboard interactivity with drill-throughs and tooltips could provide deeper analytical insights.

Future Scope:

The project offers several opportunities for future enhancement and real-world deployment:

- Deployment of the predictive model as a web application using Flask or Streamlit for real-time risk assessment.
- Integration of explainable AI techniques (e.g., SHAP) to improve transparency and trust in model predictions.

- Expansion of the dataset with real-time or historical data to improve model generalization.
- Automation of data pipelines for continuous data ingestion and model retraining.
- Extension of the solution to support additional banking use cases such as credit scoring, customer segmentation, and fraud detection.