Trainity Data Analytics Training

Final Project-2

Bank Loan Case Study

Name-Sudipta Samanta

PROJECT DESCRIPTION

This project focuses on analyzing a dataset of loan applications to identify patterns and factors that influence loan default. The primary aim is to support the finance company in making informed decisions about loan approvals, reducing financial risks, and ensuring capable applicants are not rejected. By applying Exploratory Data Analysis (EDA), we examine customer and loan attributes to uncover insights that can mitigate default risks while maximizing business opportunities.

APPROACH

The project was executed in the following steps using Python in Jupyter Notebook:

1. **Data Loading and Preprocessing:**

Imported the dataset and conducted an initial review of its structure.

Checked for missing values and handled them using appropriate imputation techniques, such as filling with the mean, median, or mode based on the variable type and distribution.

2. Handling Missing Data:

Used Python libraries like pandas to identify missing data.

Visualized missing values using a heatmap (seaborn) and bar plots (matplotlib).

3. Outlier Detection:

Used statistical methods like the Interquartile Range (IQR) to identify outliers.

Visualized distributions with box plots and histograms to understand the presence and impact of outliers.

4. Analyzing Data Imbalance:

Calculated the proportions of target classes using value_counts.

Visualized the class distribution using pie charts and bar graphs to understand data imbalance.

5. Univariate, Segmented Univariate, and Bivariate Analysis:

Conducted univariate analysis using descriptive statistics (mean, median, standard deviation) and visualized distributions with histograms and bar charts.

Performed segmented univariate analysis to compare variable distributions across different scenarios (e.g., payment difficulties vs. timely payments).

Conducted bivariate analysis using scatter plots, correlation matrices, and pair plots to explore relationships between customer attributes, loan attributes, and the likelihood of default.

6. Correlation Analysis:

Segmented the data into scenarios and calculated correlation coefficients (numpy.corrcoef) to identify the top correlated variables for each case.

Visualized correlations using heatmaps to highlight key indicators of loan default.

TECH-STACK USED

Jupyter Notebook: For interactive data analysis and visualization.

And **MS Excel**.

INSIGHTS

1. Missing Data Patterns:

Variables related to credit history and income had the highest percentage of missing values.

Imputation with median values proved effective for income-related variables, while mode imputation worked well for categorical variables.

2. Outlier Analysis:

High loan amounts and extended durations were identified as common outliers.

These were primarily linked to cases of loan default, suggesting they are risk indicators.

3. Class Imbalance:

The dataset exhibited significant class imbalance, with more non-default cases than default cases.

This highlights the need for strategies like oversampling or undersampling during model building.

4. Key Variables Influencing Loan Default:

Loan amount, loan duration, and applicant income were found to be the most correlated with loan default.

Customers with lower incomes and higher loan amounts exhibited a higher likelihood of default.

5. Segmented Analysis:

Customers with payment difficulties had distinct trends in income, credit history, and loan amount distributions compared to those without difficulties.

Applicants with insufficient credit history showed higher default rates.

6. RECOMMENDATIONS:

Introduce stricter approval criteria for high-risk applicants based on loan amount and credit history.

Offer loans with adjusted terms (e.g., higher interest rates or shorter durations) to mitigate risks for borderline cases.

Address data imbalance when building predictive models to ensure accurate default predictions.

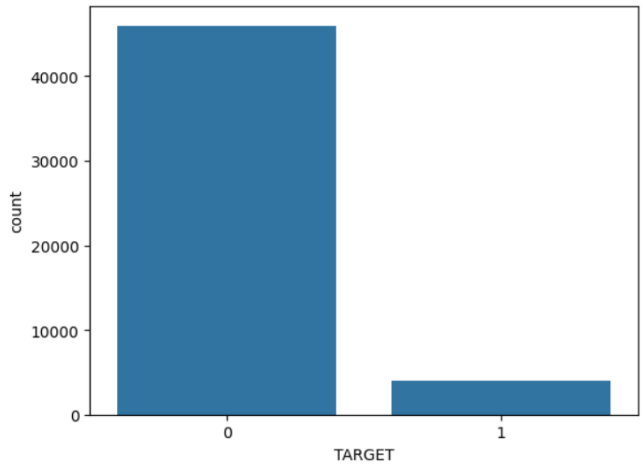
Results

A. Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.**

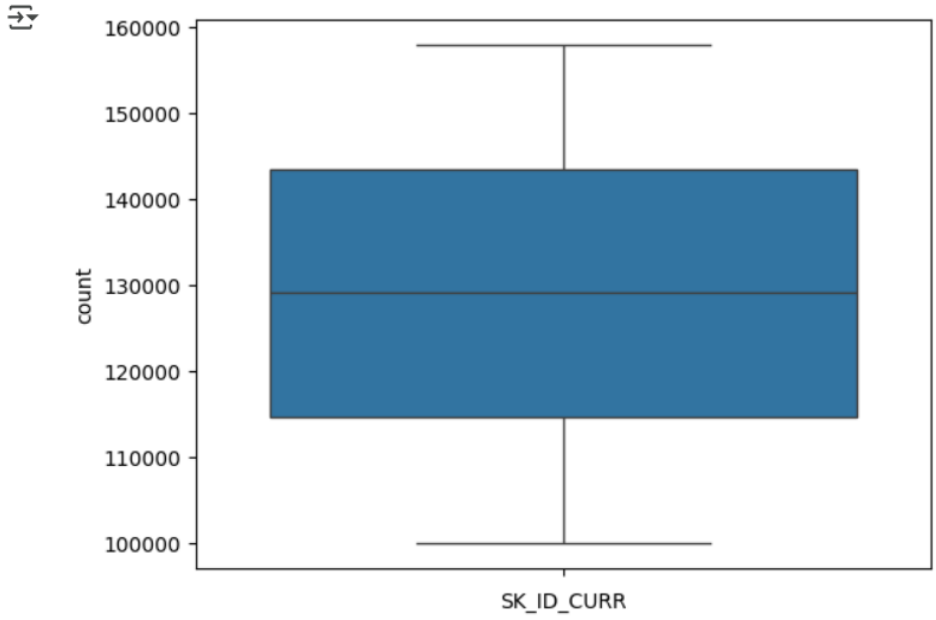
```
sns.countplot(data = n, x = 'TARGET') # no effect of defaulters
```

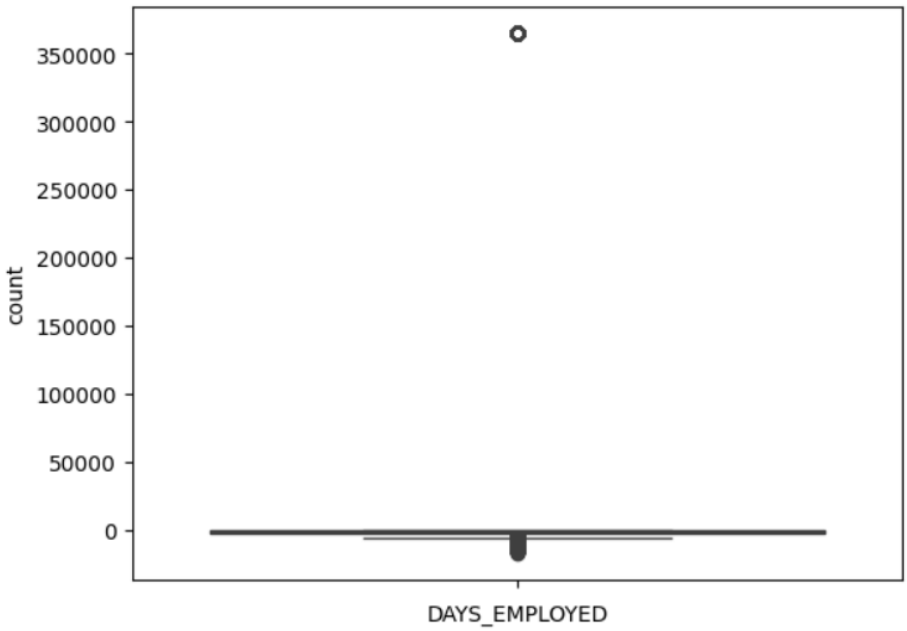
```
<Axes: xlabel='TARGET', ylabel='count'>
```

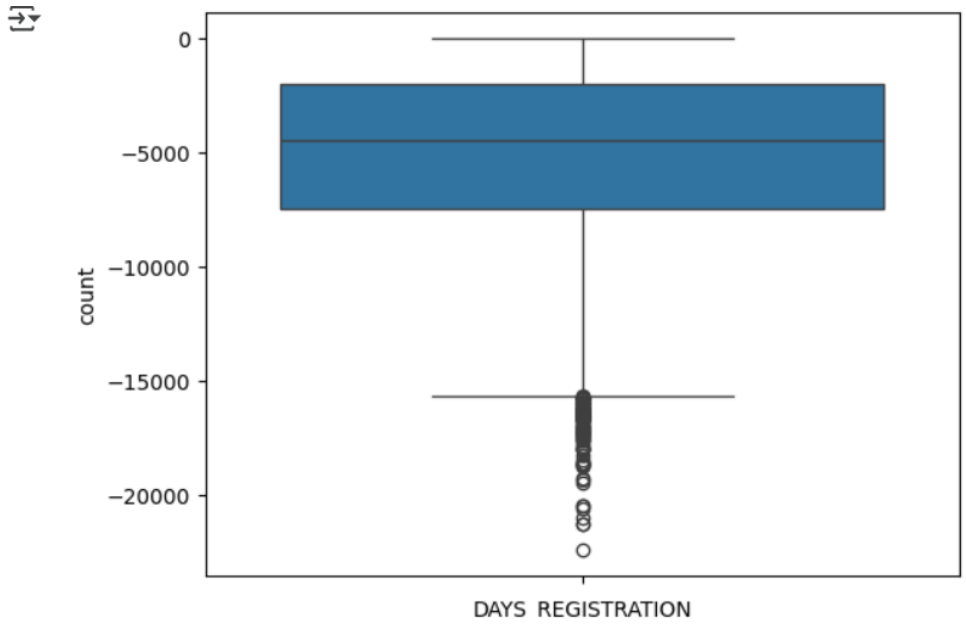


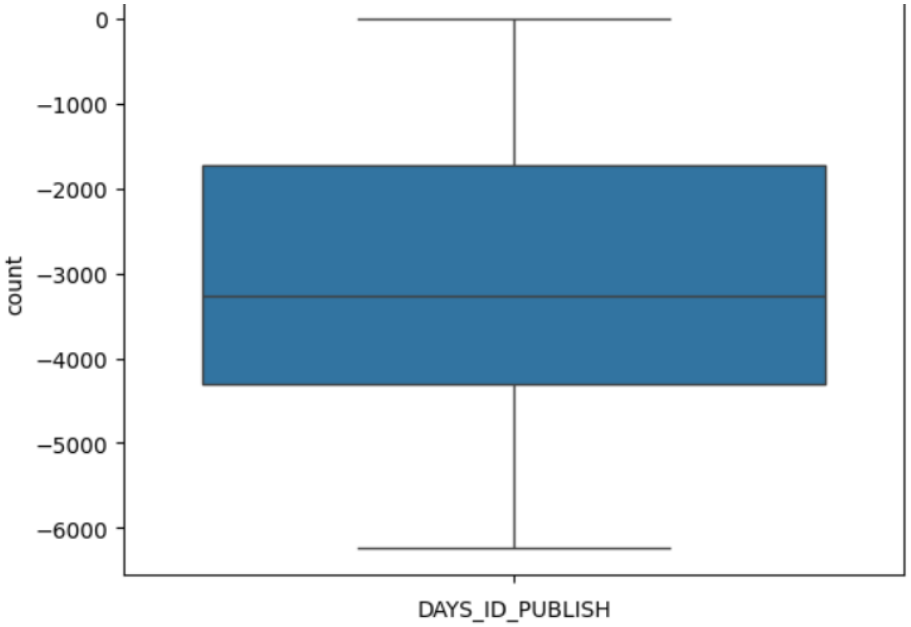
B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

- **Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.**









C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

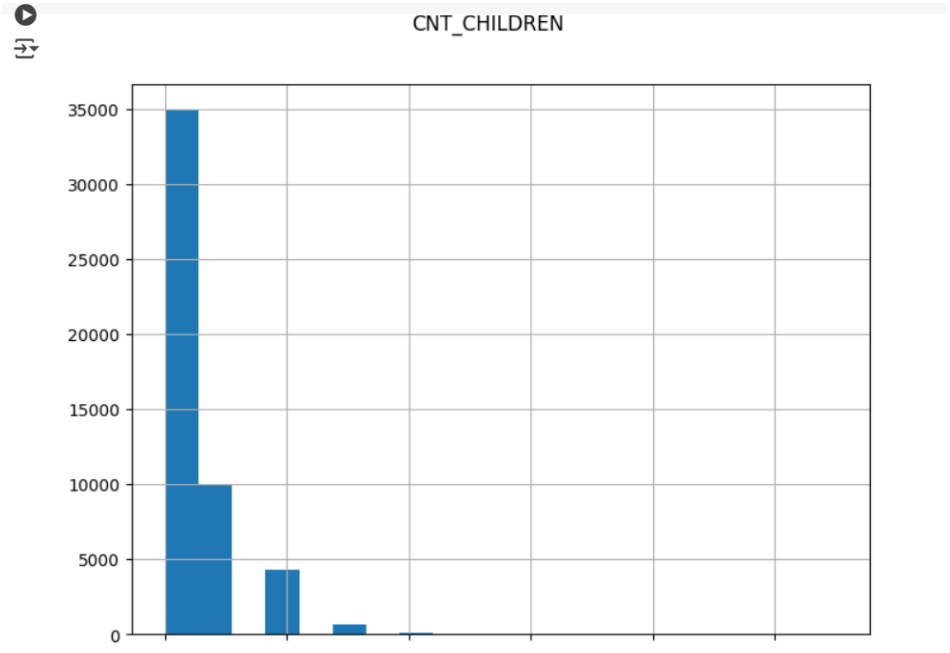
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

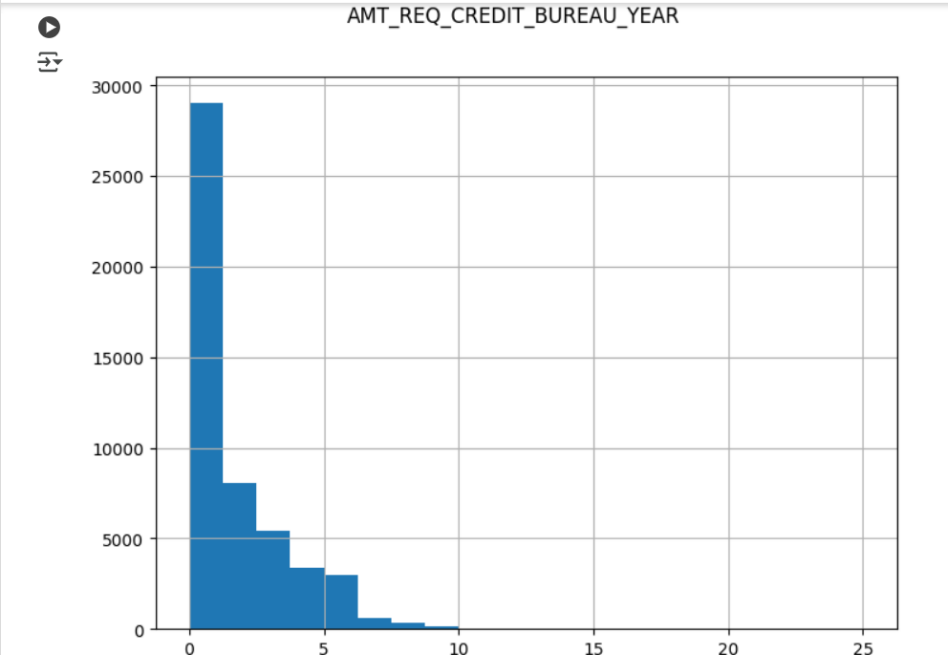
Imbalance Ratio=11:53

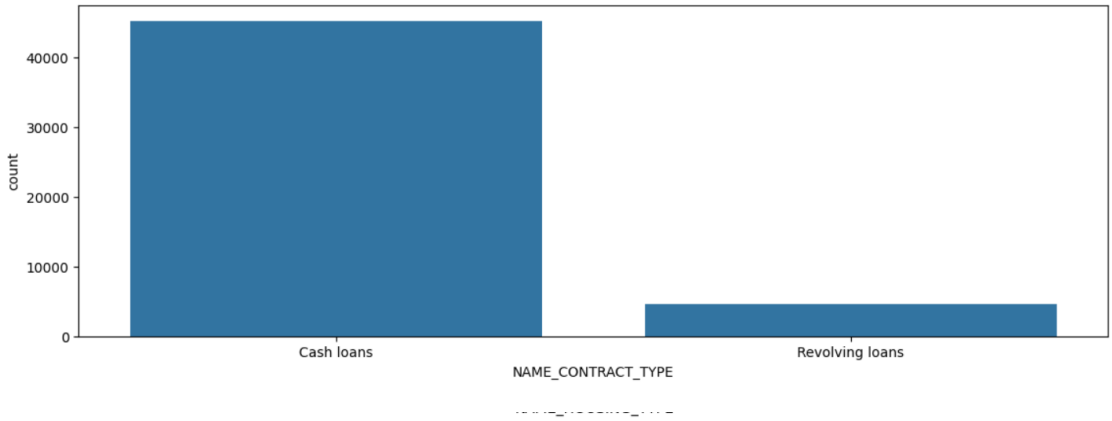
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

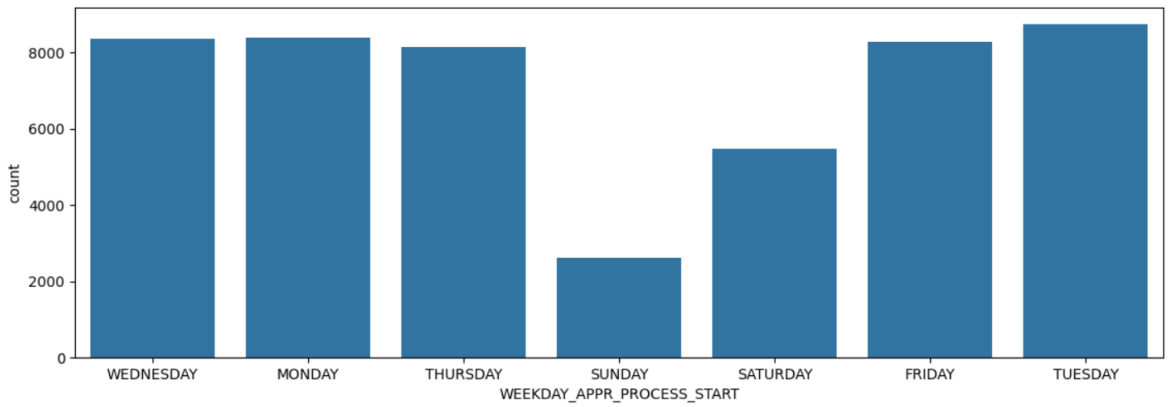
- **Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.**

Univariate Analysis

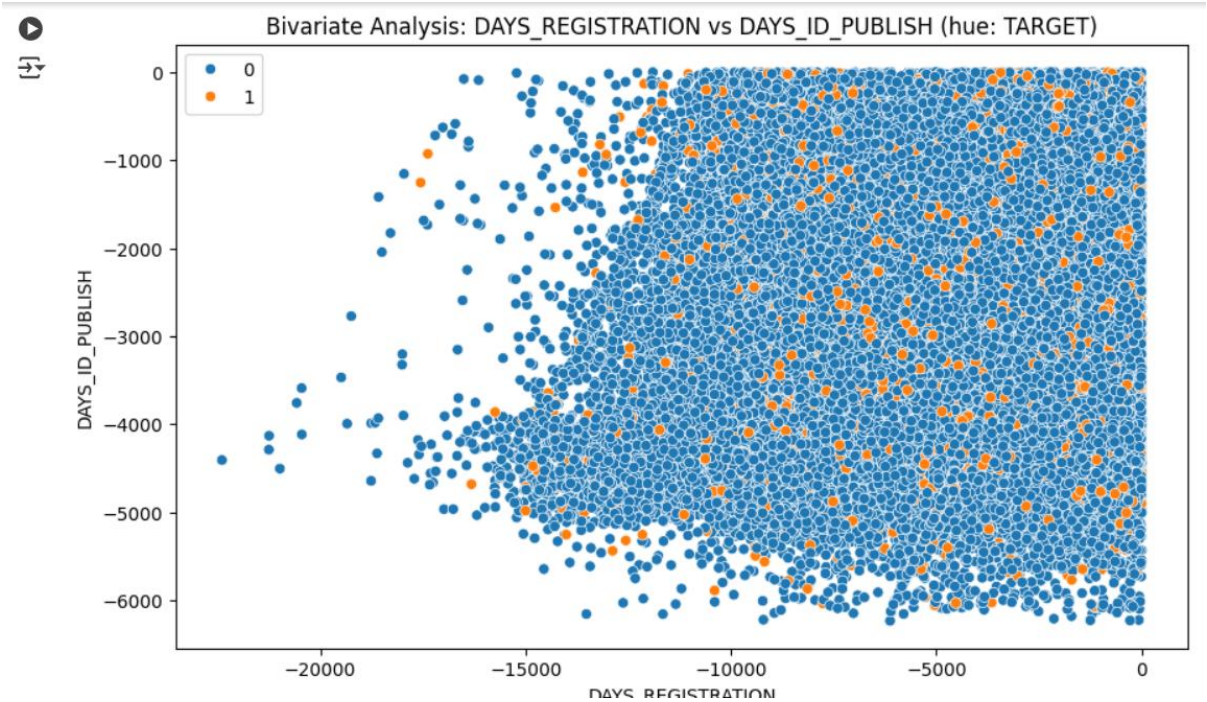






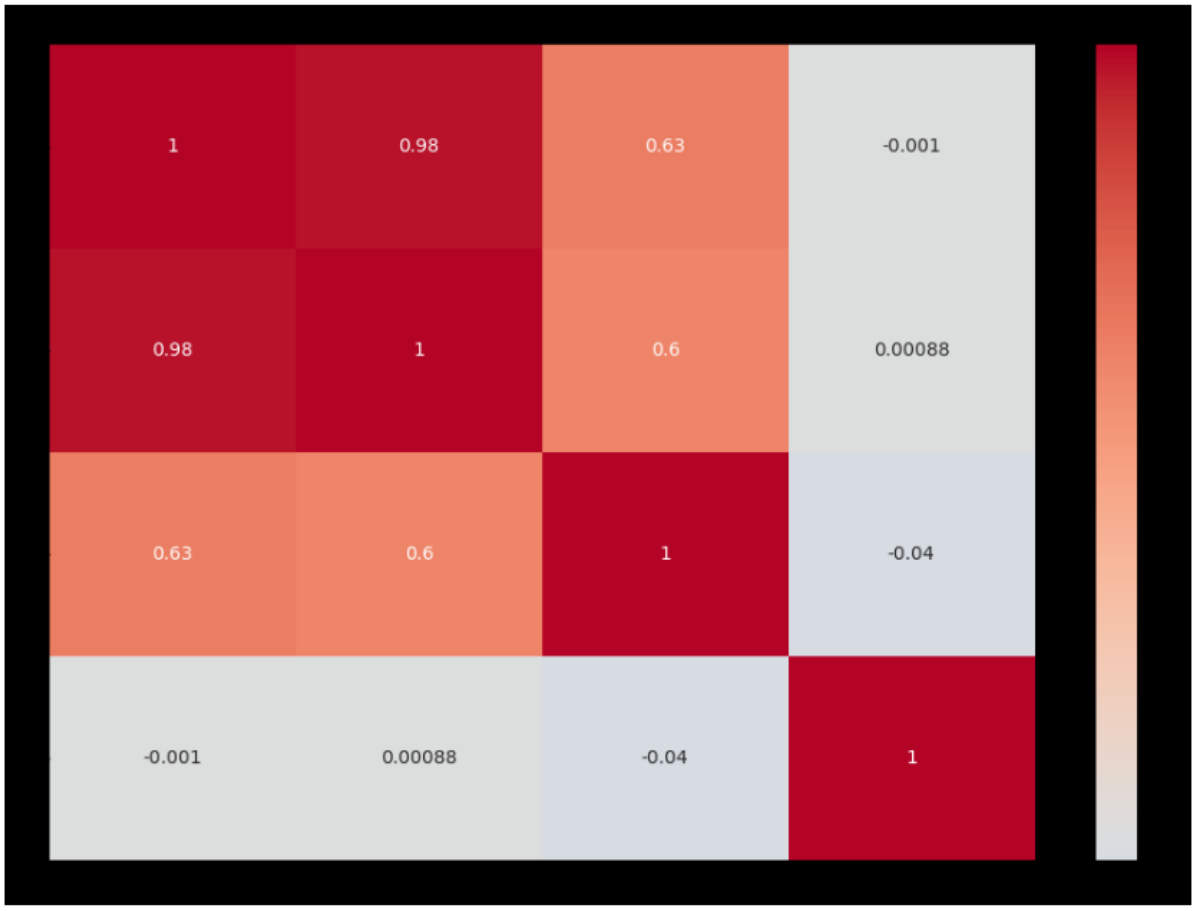


Bivariate Analysis



E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- **Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.**



The project successfully identified patterns and insights that can guide loan approval decisions and mitigate financial risks. The use of EDA helped uncover key indicators of default and provided actionable recommendations to enhance the company's decision-making process. These findings were presented in a comprehensive PDF report and visualized using clear and interpretable graphs.

Click on the link: (Access to Jupyter Notebook)

https://colab.research.google.com/drive/152CVqsqUNB027f9M1qx-Pd-GjP_Jdk-3?usp=sharing

Click on the link:

..\Documents\application_data.csv