Bank Loan Case Study

Exploratory Data Analysis (EDA) Report

1. Introduction

Objective of the Study:

As a data analyst at a finance company, our goal is to analyze loan application data to identify patterns that indicate whether an applicant is likely to default on their loan. By using **Exploratory Data Analysis (EDA)**, we can help the company make better loan approval decisions, minimizing financial losses while ensuring capable applicants are not rejected unnecessarily.

Challenges Faced by the Company:

- Some customers with insufficient credit history default on their loans.
- The company loses business if a capable applicant is rejected.
- If an incapable applicant is approved, the company faces financial risk.

By analyzing **customer attributes and loan attributes**, I aim to understand the factors influencing loan default and provide data-driven recommendations.

---

2. Understanding the Data:

We have three datasets:

1. **application_data.csv** – Contains details of loan applications, including applicant demographics, financial data, and loan details.
2. **previous_application.csv** – Includes information about applicants' past loan applications.
3. **columns_description.csv** – Describes the meaning of various columns in the datasets.

Key Variables to Analyze:

- **TARGET**: 1 = Client had payment difficulties, 0 = All other cases.
- **NAME_CONTRACT_TYPE**: Type of loan (Cash/Consumer loan, etc.).
- **AMT_INCOME_TOTAL**: Total income of the applicant.
- **AMT_CREDIT**: Amount of loan approved.
- **DAYS_BIRTH**: Applicant's age (in days).
- **NAME_EDUCATION_TYPE**: Education level of the applicant.
- **NAME_FAMILY_STATUS**: Marital status of the applicant.
- **OCCUPATION_TYPE**: Applicant's occupation category.
- **EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3**: External risk rating scores.

3.Approach:

1. **Data Collection**: The dataset consists of loan applications, including customer attributes and loan attributes.
2. **Data Cleaning**: Identified and handled missing values using appropriate imputation techniques.
3. **Outlier Detection**: Used quartile-based analysis to identify and handle outliers.
4. **Data Imbalance Analysis**: Evaluated the distribution of target variables to determine class imbalance.
5. **Univariate, Segmented Univariate, and Bivariate Analysis**: Analyzed distributions and relationships between variables.
6. **Correlation Analysis**: Identified key indicators of loan default.

4.Tech-Stack Used

- **Software**: Microsoft Excel 2019

- **Techniques**: Conditional Formatting, COUNT, COUNTIF, AVERAGE, MEDIAN, PIVOT Tables, QUARTILE, CORREL, and Data Visualization tools (Bar Charts, Histograms, Box Plots, Scatter Plots).

5.Data Preprocessing:

A. Handling Missing Data:

- Columns with more than **50% missing values** were dropped.
- Missing values in categorical columns were replaced with the most frequent value.
- Missing values in numerical columns were replaced with the **median** to avoid bias.

ALL THE COLUMN NAME WHICH ARE HIGHLIGHTED IN GREEN NEED TO BE DROPPED DOWN

AS THEY ARE IRRELEVANT COLUMNS FOR DOING OUR ANALYSIS

Column name	Total number of null values	Percentage of null v
FLAG_MOBIL	1	0.000325192
FLAG_EMPLOY_PHONE	55387	18.01138821
FLAG_WORK_PHONE	0	0
FLAG_CONT_MOBILE	0	0
FLAG_PHONE	0	0
FLAG_EMAIL	0	0
CNT_FAMILY_MEMBERS	2	0.000650383
REGION_RATING_CLENT	0	0
REGION_RATING_CLENT_W_CITY	0	0
EXT_SOURCE_3	60965	19.82530706
YEAR_BEGINEXPLUATATION_AVG	150008	48.78134441

YEAR_BEGINEXPLUATATION_MODE	150007	48.78101922
YEAR_BEGINEXPLUATATION_MEDIAN	150007	48.78101922
TOTAL_AREA_MODE	148431	48.26851722
EMERGENCYSTATE_MODE	145755	47.39830445
DAYS_LAST_PHONE_CHANGE	1	0.000325192
FLAG DOC 2	0	0
FLAG DOC 3	0	0
FLAG DOC 4	0	0
FLAG DOC 5	0	0
FLAG DOC 6	0	0
FLAG DOC 7	0	0
FLAG DOC 8	0	0
FLAG DOC 9	0	0
FLAG DOC 10	0	0
FLAG DOC 11	0	0
FLAG DOC 12	0	0
FLAG DOC 13	0	0
FLAG DOC 14	0	0
FLAG DOC 15	0	0
FLAG DOC 16	0	0
FLAG DOC 17	0	0
FLAG DOC 18	0	0
FLAG DOC 19	0	0
FLAG DOC 20	0	0
FLAG DOC 21	0	0

Insights:

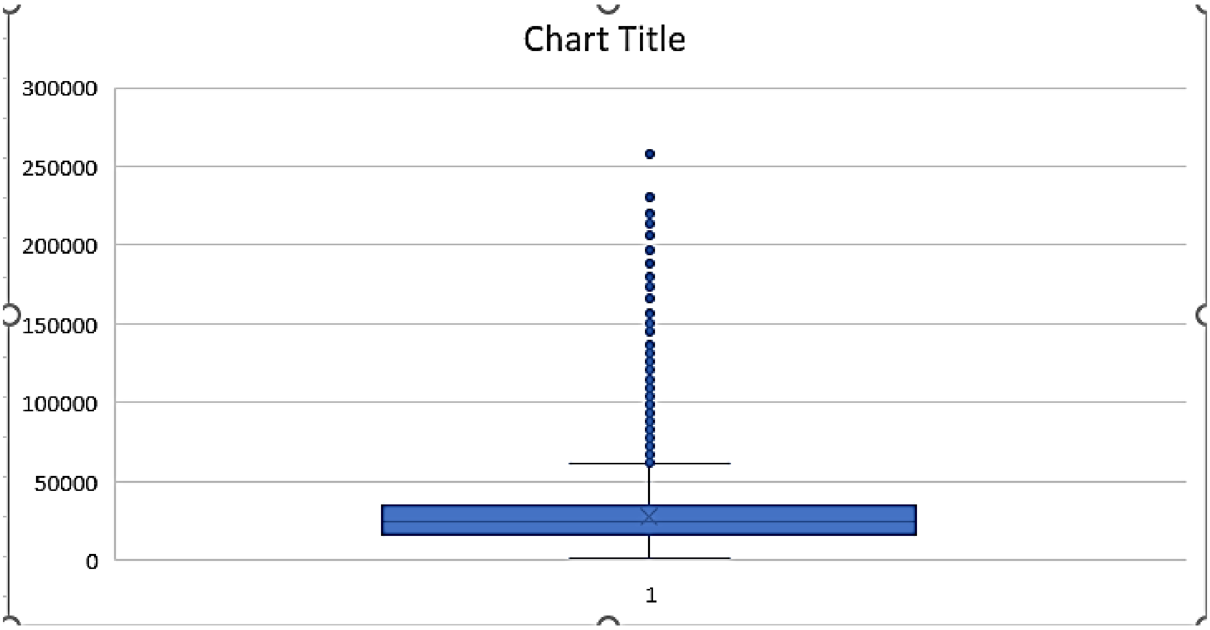
- Dropping columns with excessive missing values ensures that our dataset remains robust and avoids unreliable predictions.

- Using median imputation prevents skewing of numerical data due to extreme values.

Task B: Outlier Detection:

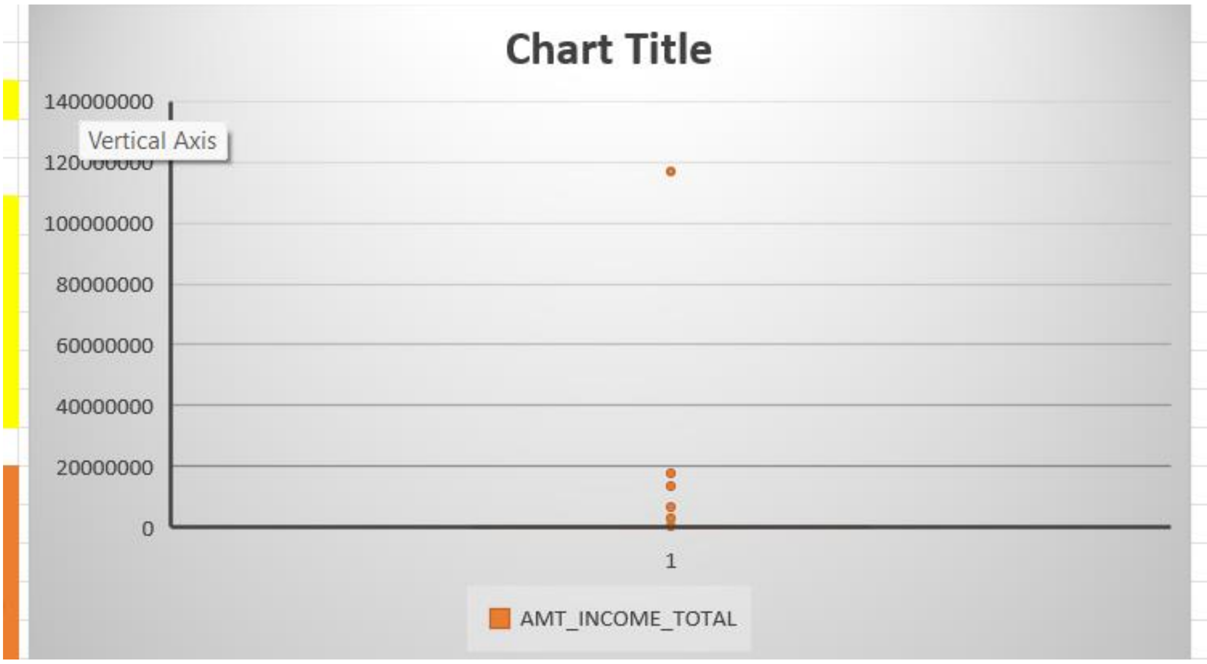
- Outliers were identified using **Interquartile Range (IQR)** and box plots.
- Key outliers in **AMT_ANNUITY, AMT_INCOME_TOTAL, and DAYS_EMPLOYED** were analyzed.
- Certain extreme outliers were replaced with the **median**.
- **Visualization**: Box plots were used to illustrate the outlier distribution.

Box plot for **AMT_ANNUITY**



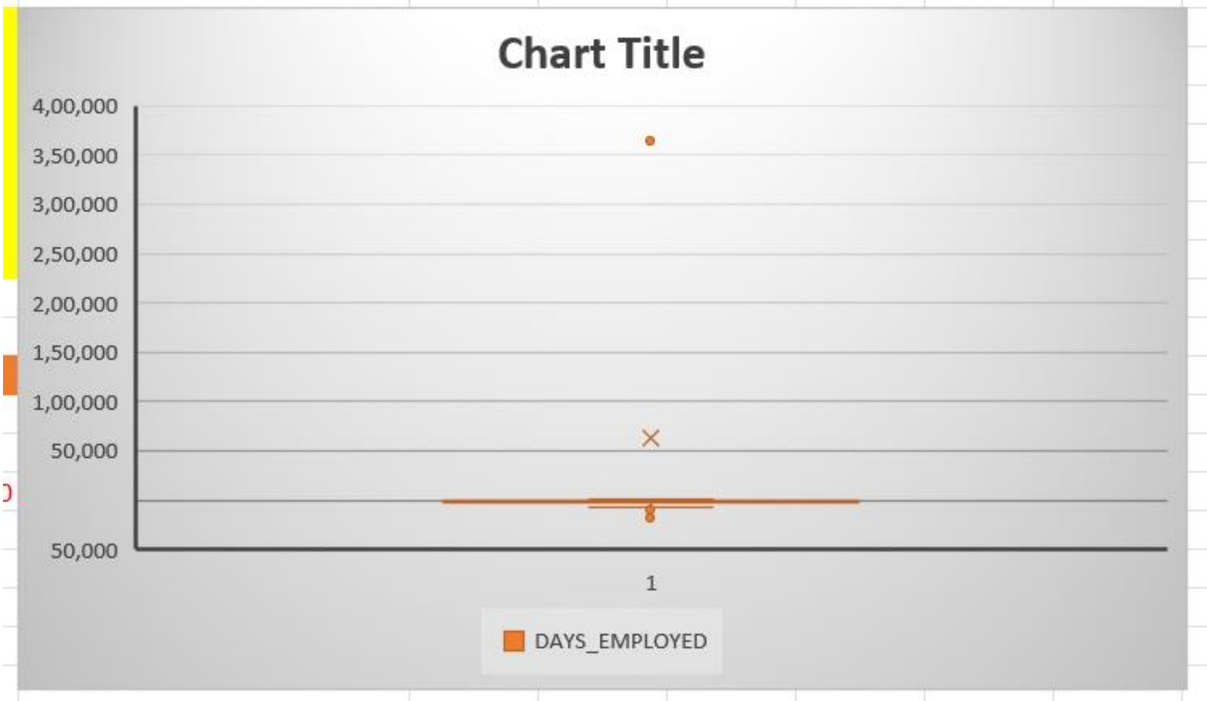
Insight: the outlier is above 2,50,000 which can be replaced by median 24903

Box plot for **AMT_INCOME_TOTAL**



Insight : the outlier is near 12000000

Box plot for **DAYS_EMPLOYED**



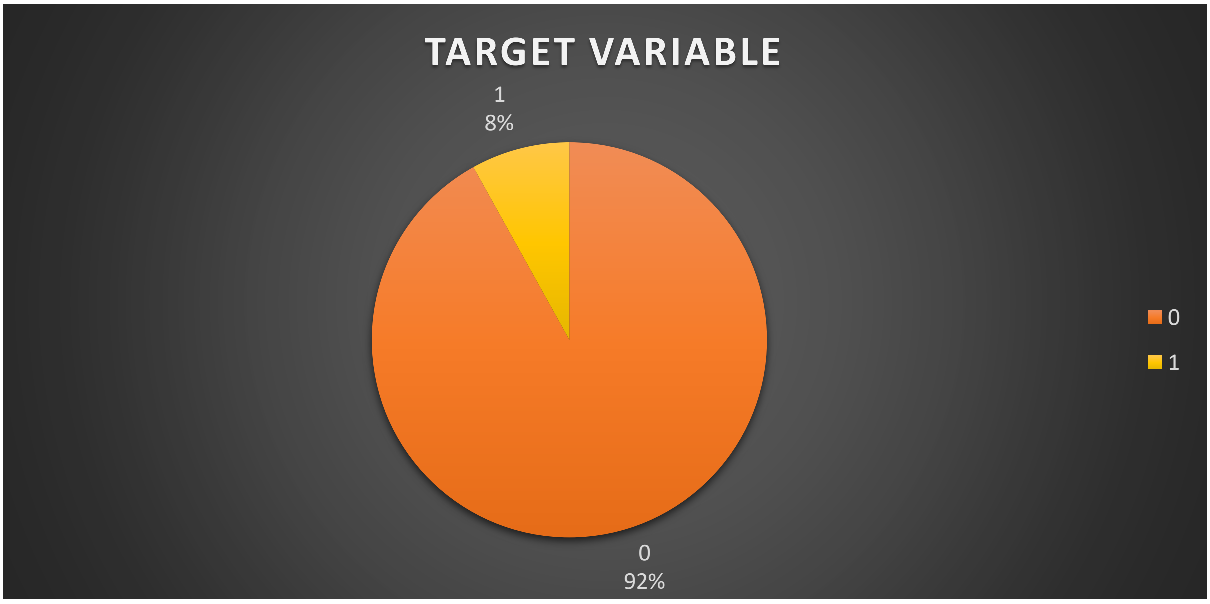
insight: the outlier is 365243 which can be replaced by median 1213

Insights:

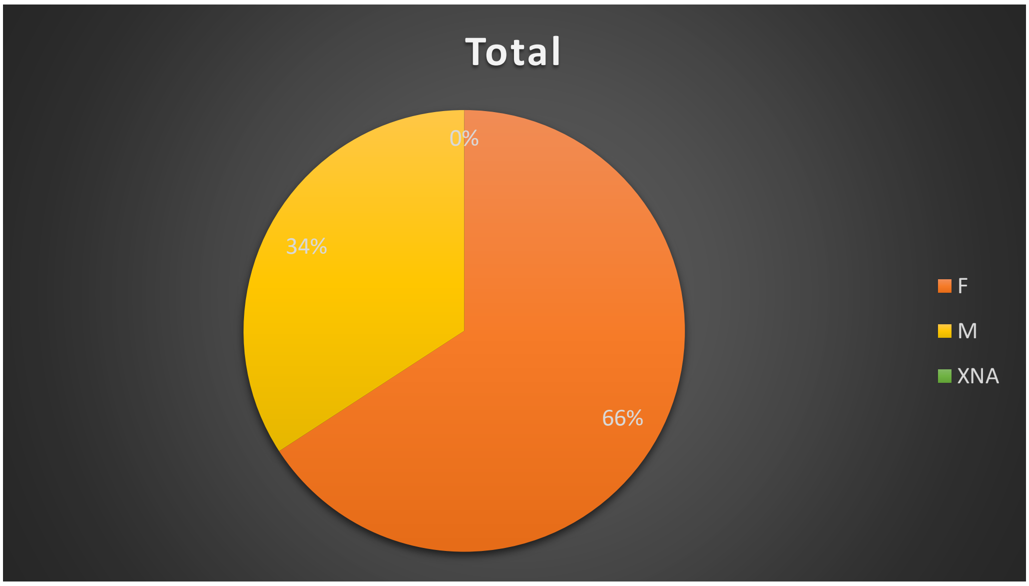
- Outliers can significantly impact the analysis and model predictions; replacing extreme outliers helps maintain data accuracy.
- Some outliers, like variations in income, were retained since they represent real-world scenarios.

Task C: Data Imbalance Analysis

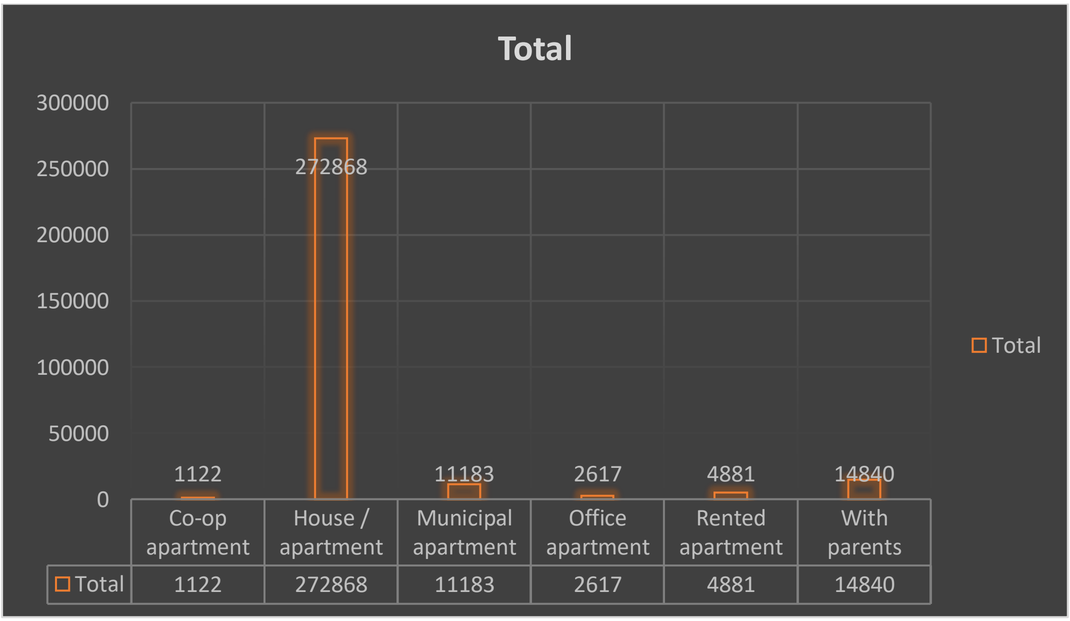
- The target variable (loan default status) was **highly imbalanced**.
    - **92% of customers had no payment issues**.
    - **8% of customers had payment difficulties**.
- **Visualization**: A pie chart highlighted this imbalance.



- Gender Imbalance:
    - **66% of applicants were female**, 34% were male.
    - **Female applicants had lower default rates compared to male applicants**.



- Name Housing Type Imbalance:
    - **88.73% of applicants lived in a house/apartment**, while others lived in rented apartments or with parents.
    - **Applicants in rented apartments had a higher likelihood of defaulting**



	Co-op apartment	House / apartment	Municipal apartment	Office apartment	Rented apartment	With parents
Total	1122	272868	11183	2617	4881	14840

Insights:

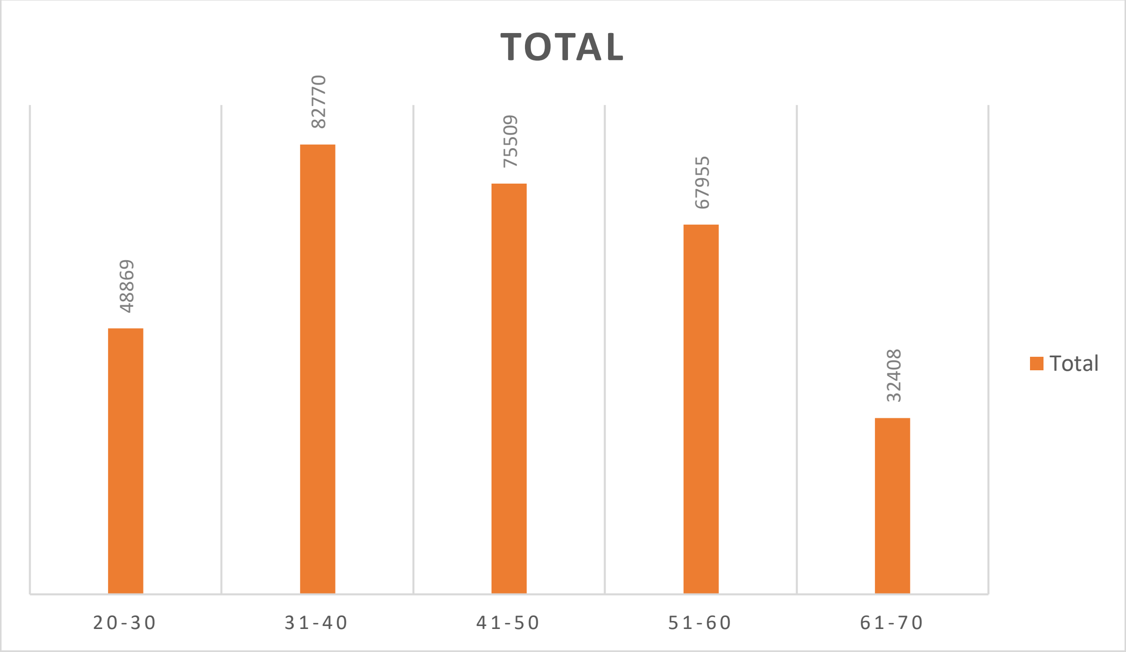
- The imbalance suggests that a simple predictive model might be biased towards non-defaulters.

- The bank lends more to female applicants, who generally have lower default rates, making them safer borrowers.
- Clients in rented apartments or with unstable housing conditions show **higher risk of default**.
- Additional techniques like **sampling strategies** or **weighted classification models** might be required for future predictive modeling.

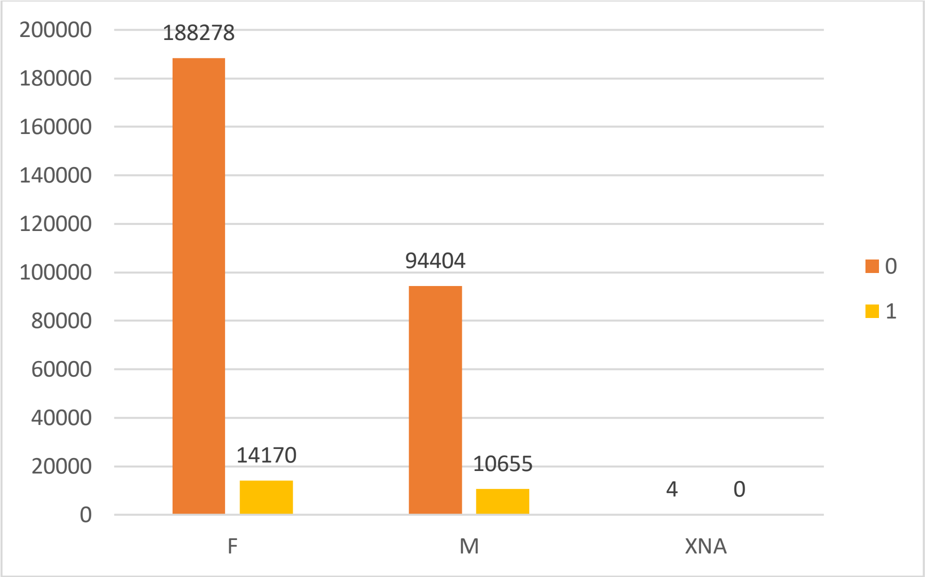
Task D: Univariate, Segmented Univariate, and Bivariate Analysis

Univariate Analysis

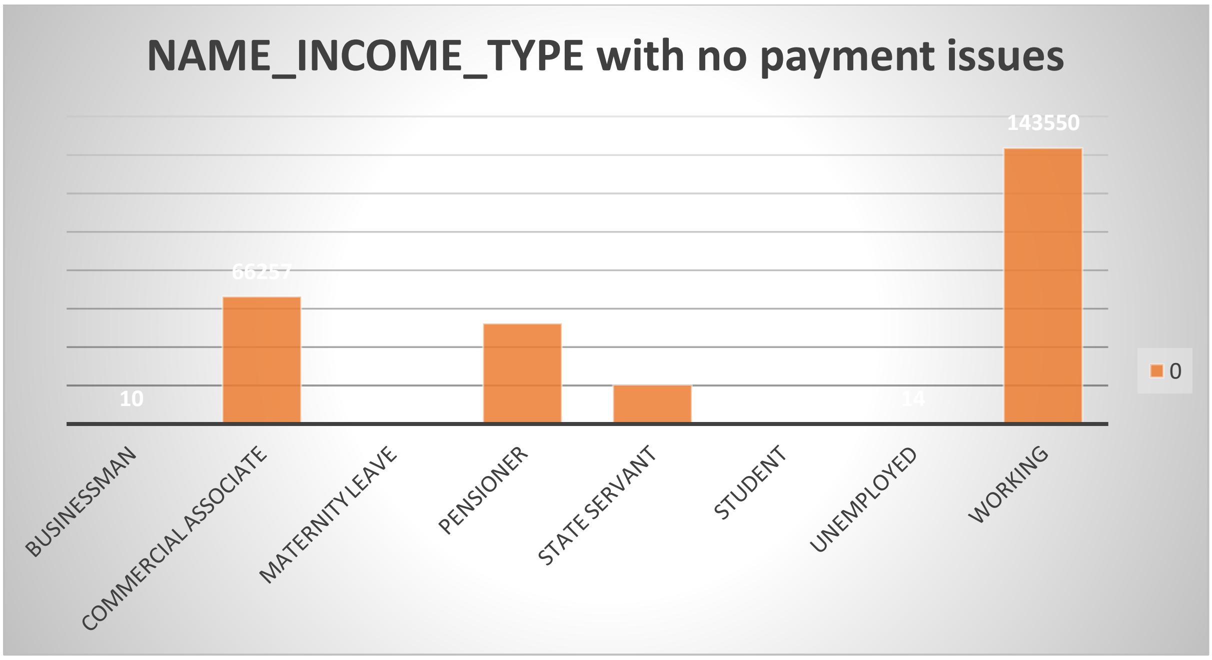
- **Age Group Analysis**: Most applicants were between **31-40 years old**.



- **Gender Analysis**: **66% of applicants were female**, 34% were male.

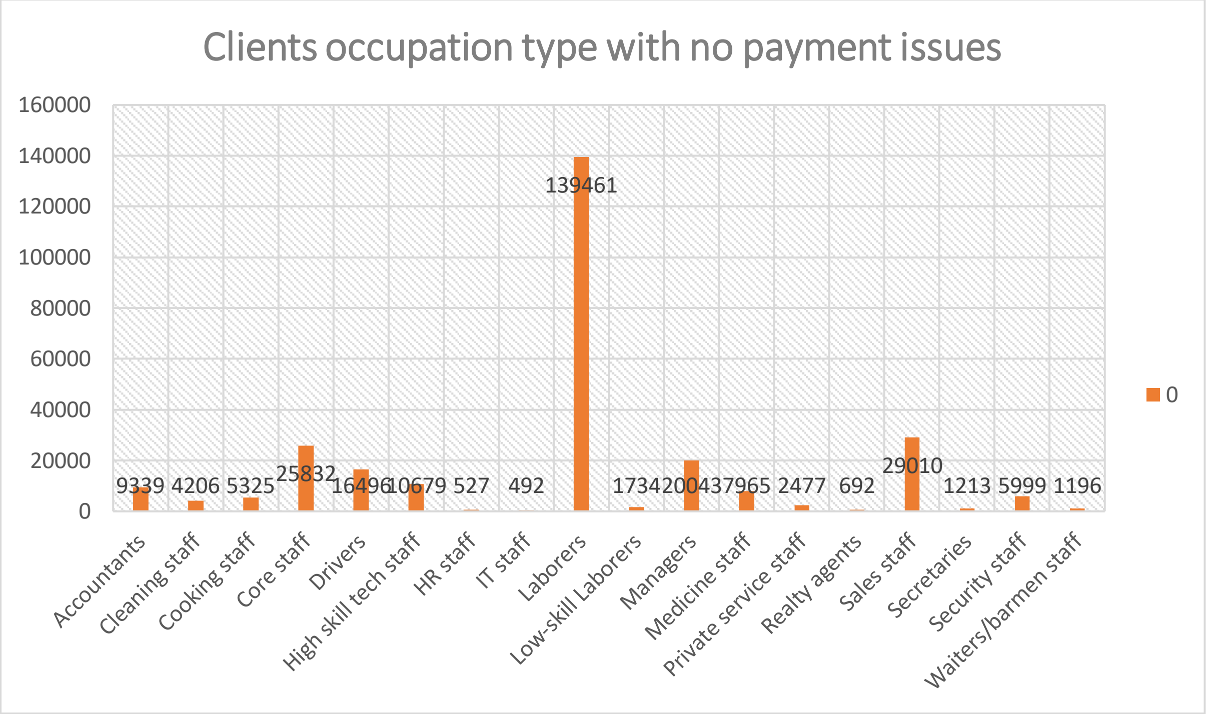


- **Income Type Analysis**: **Working professionals** had the highest number of applications.

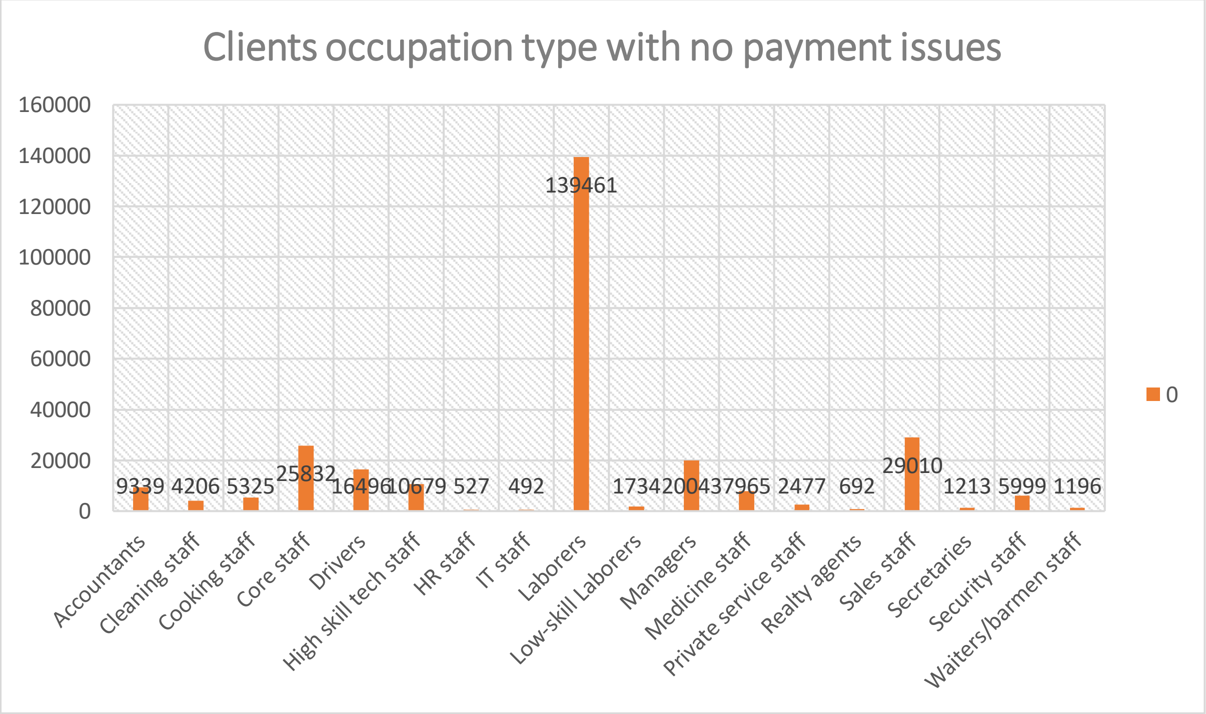


- **Occupation Type Analysis**: Majority of applicants were **Laborers**.

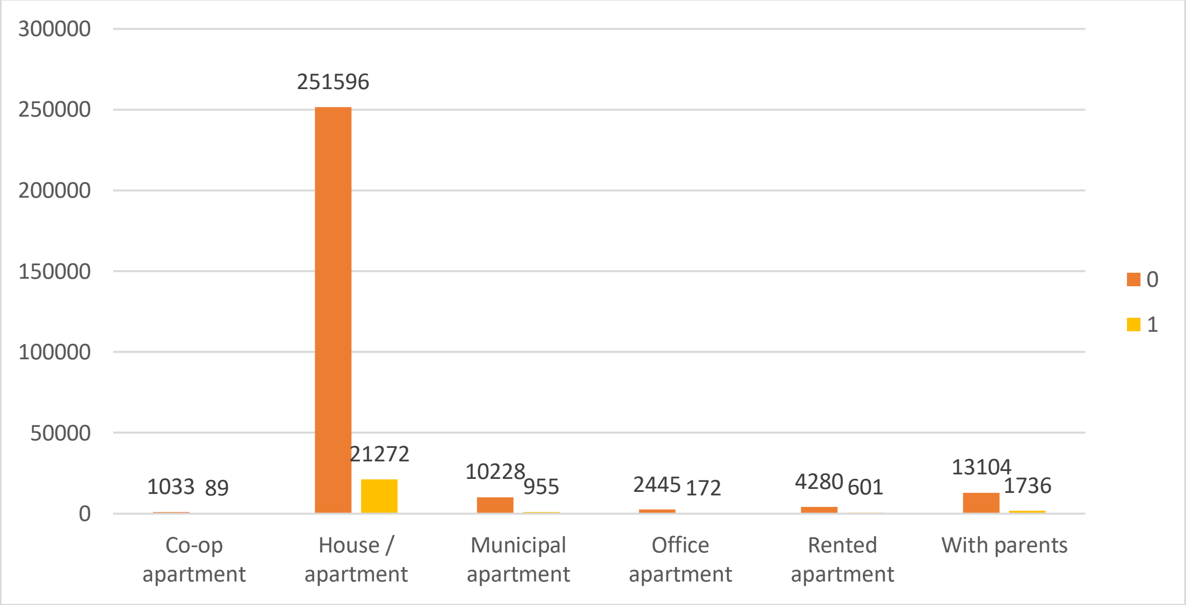
- With no payment issue



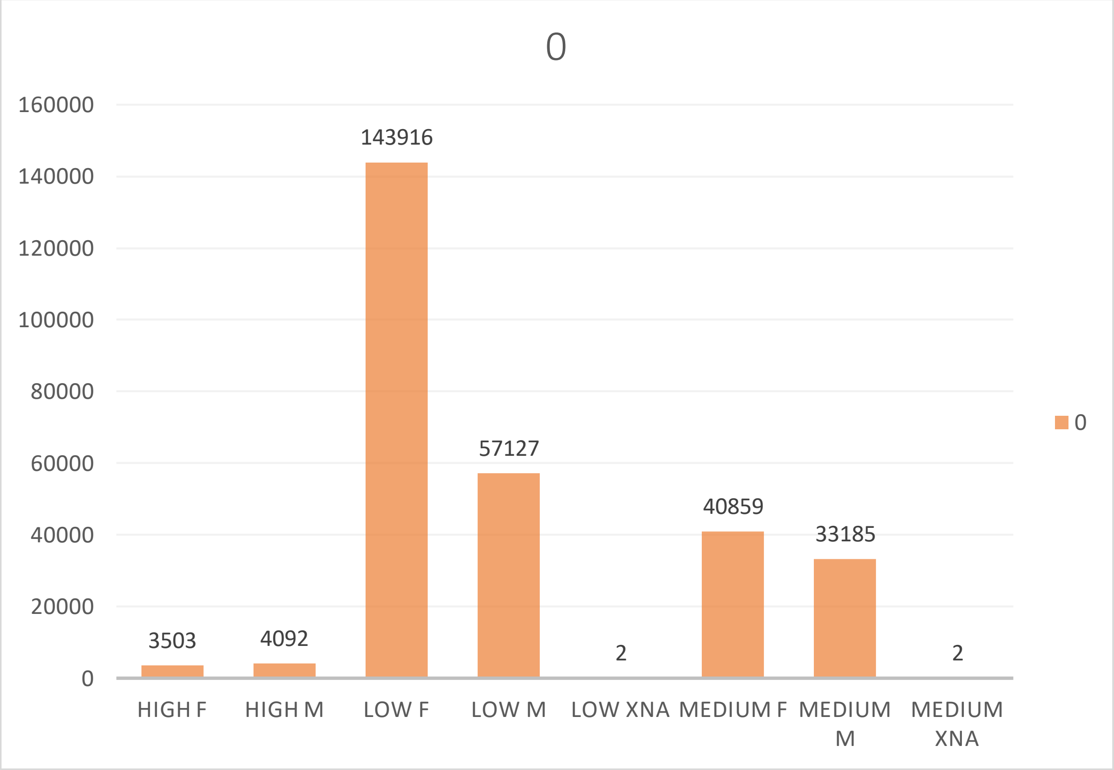
With payment issue



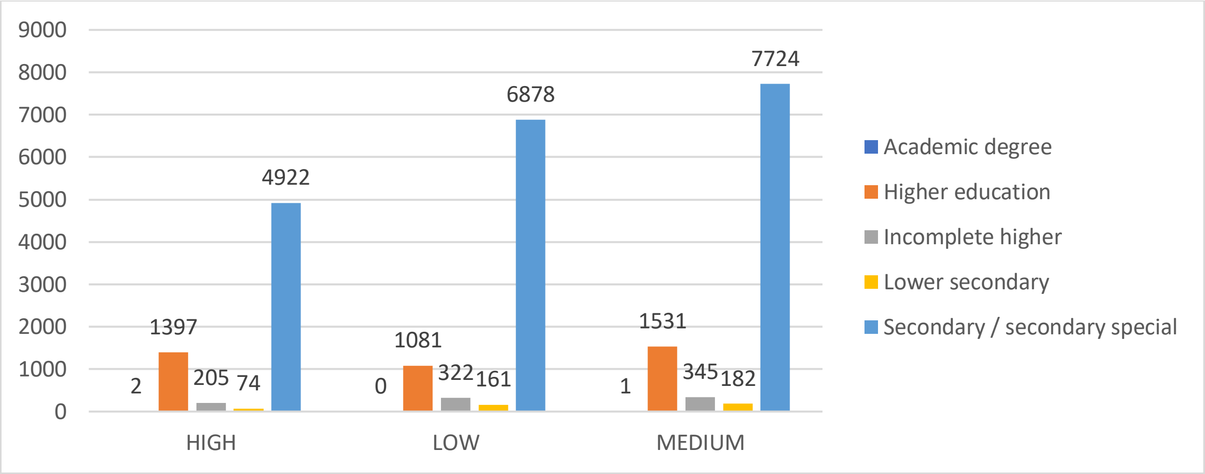
- **Housing Type Analysis**: Most applicants lived in **houses/apartments**.



- **Segmented Univariate Analysis**
  - Compared distributions of various attributes across different target variable classes.
- **Gender vs Loan Default**: Female applicants had a lower default rate than males.

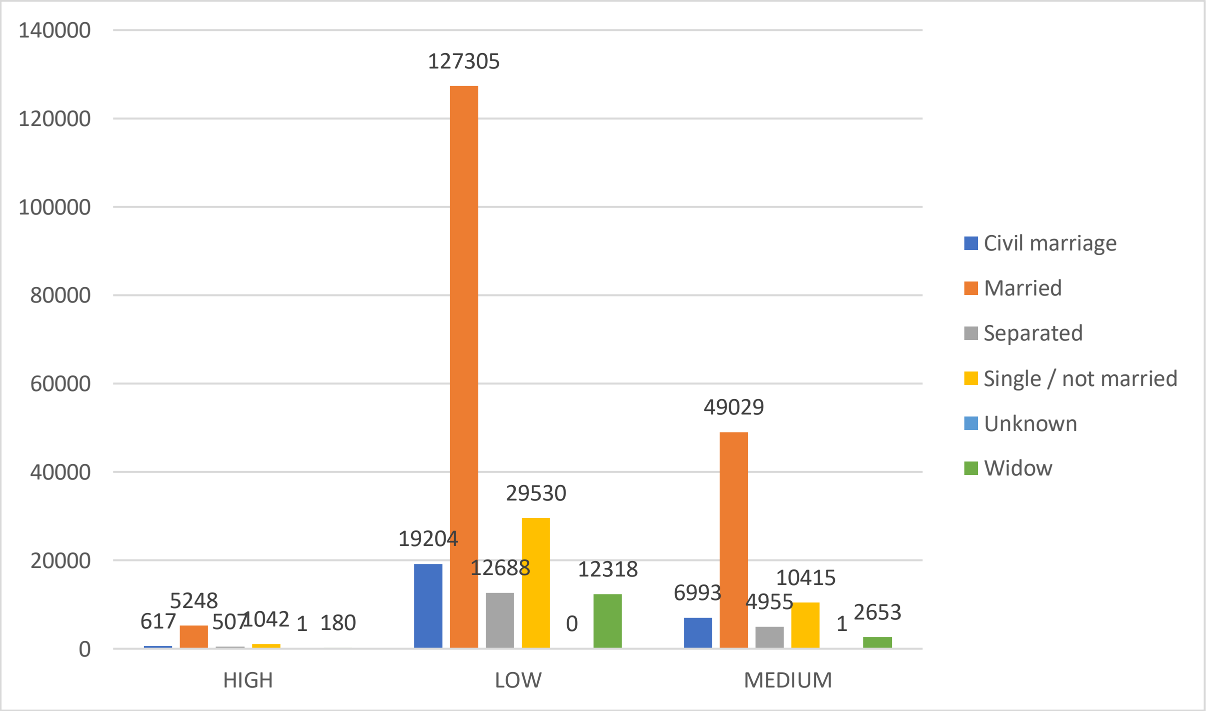


- **Credit amount vs Educational status**: highest credit amount is for Academic degree holders.



- **Bivariate Analysis**

  - **Total income range vs family status**: From the Bar plot we can infer that clients with total_income_range as 'Low' and family_status as 'Married' have the highest count for clients having payment issues

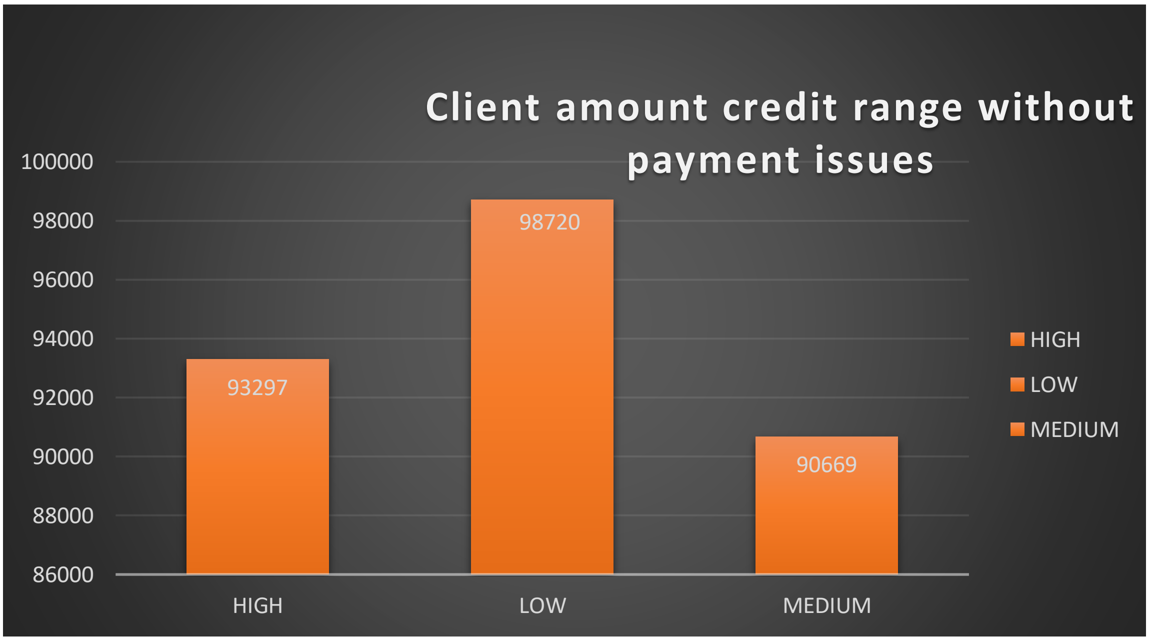


Task E: Identifying Top Correlations for Different Scenarios

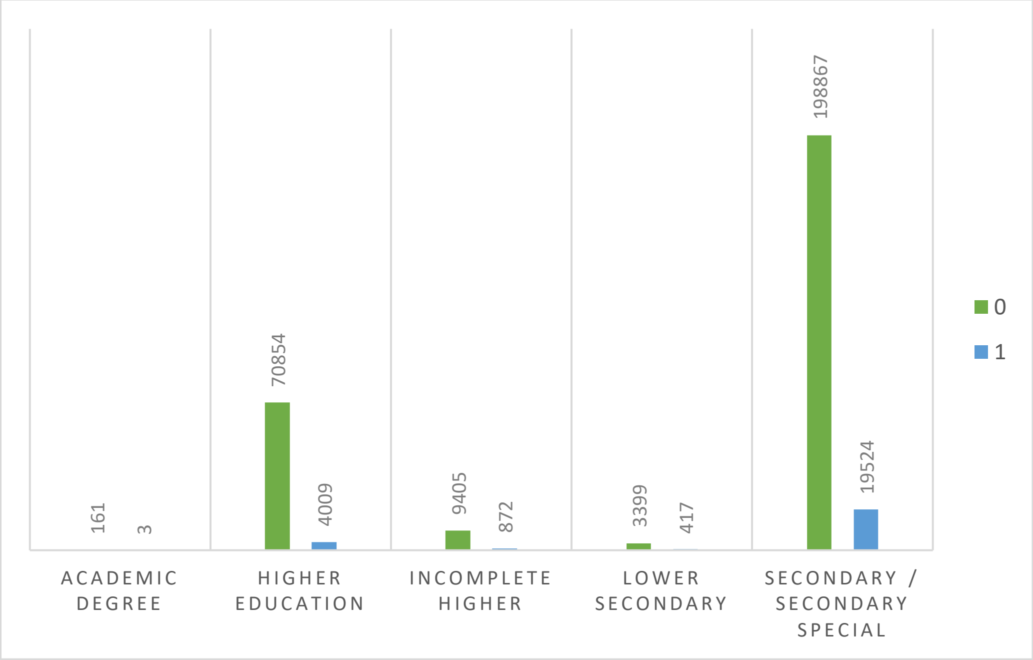
- **Segmented the dataset based on different scenarios** (e.g., clients with payment difficulties and others).
- Used **CORREL function** to calculate correlation coefficients between variables and the target variable.
- Ranked the correlations to identify the **top indicators of loan default**.
- **Visualization**: Heatmaps and correlation matrices were used.

Insights:

- **AMT_INCOME_TOTAL** had a **negative correlation** with default probability.



- EDUCATION has a direct correlation with default probabilty



6.Key Insights and Recommendations

- **Target the right customers**: Clients with higher education and stable employment history should be prioritized.
- **Adjust lending policies**: Higher interest rates should be applied to **risky applicants** identified through data patterns.
- **Refine approval criteria**: Customers with low-income and unstable job history should undergo a **more thorough credit check**.
- **Focus on Loan Purpose**: Loans for **home purchases** had a lower default rate, while those for **repairs had higher defaults**.
- **Gender-Based Lending**: Female applicants had **lower default rates**, suggesting they might be a safer lending group.

7.Conclusion

This project provided key insights into **customer behavior and risk factors affecting loan defaults**. The findings can be used to **optimize loan approval**

processes and minimize financial losses while ensuring that capable applicants receive loans.

Supporting Documents

- **Hyperlinked Excel File**: [clich here for Google Drive Link]