BANK LOAN CASE STUDY

By Sudhansu

AGENDA

- ❑ Project Description
- ❑ Approach
- ❑ Tech-Stack Used
- ❑ Data Cleaning
- ❑ Insights
- ❑ Result

Python Jupyter Notebook hyperlink : click on this

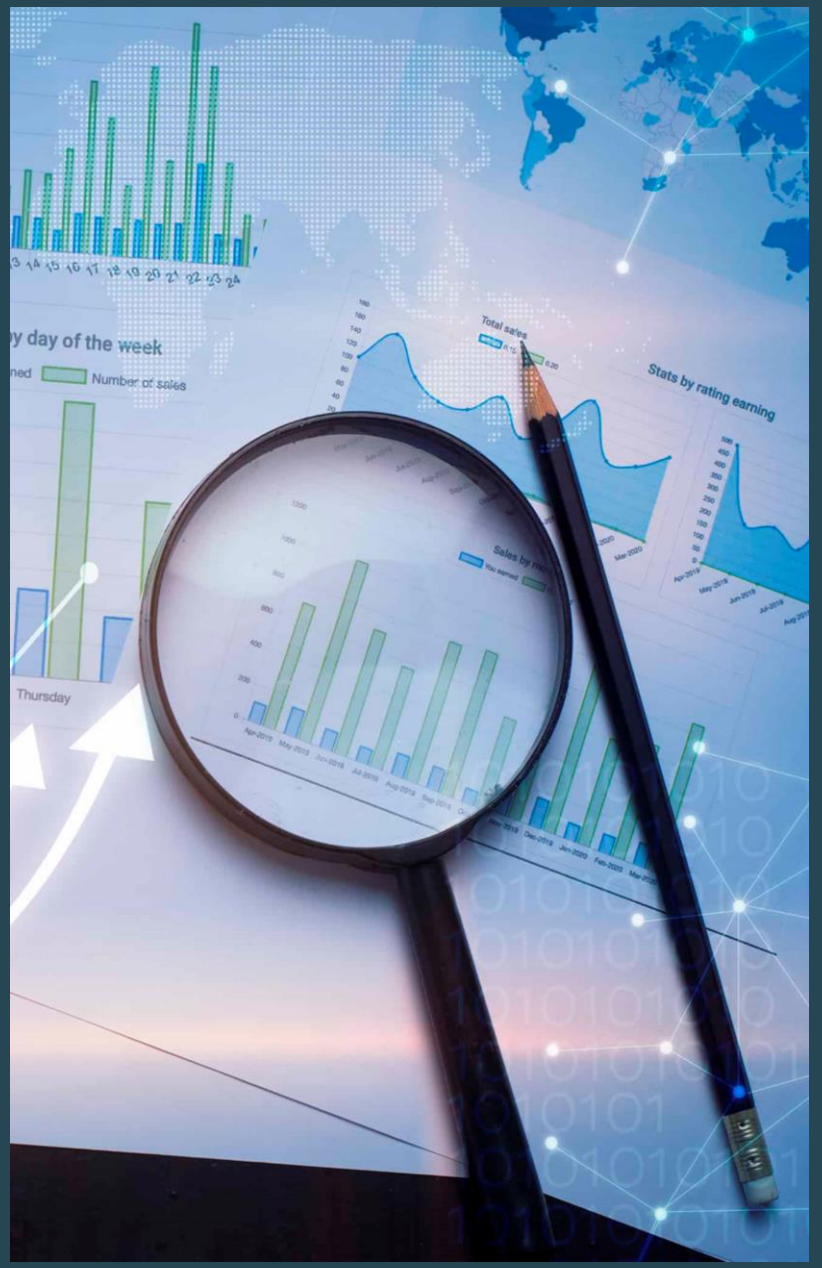
Jupyter Notebook as .ipynb

Link for: Juypyter notebook as pdf

Project Description

The bank is facing challenges, they are facing financial losses and losing business due to approval loans to the clients who can't repay and rejecting those who can repay.

Using EDA to understand how customer attributes & loan attributes influence the likelihood of default. Identifying key factors that indicate a customer credit worthiness. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants



Approach

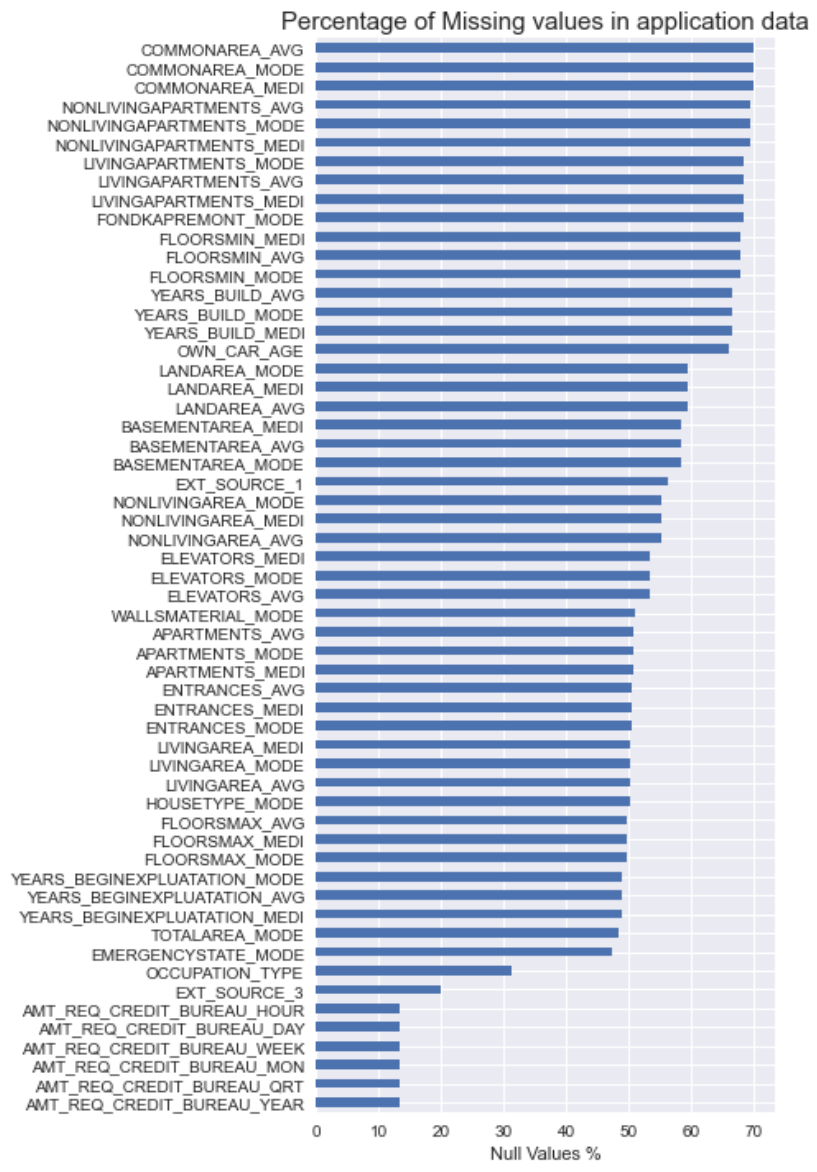
Downloaded and **imported** the dataset in Jupyter Notebook, performed **data cleaning** i.e. removing missing & invalid rows and handling outliers. Used **NumPy, pandas** and **matplotlib/seaborn chart** to draw graphs and finding insights. And finally drawn conclusions from insights and made recommendations

Tech-Stack Used

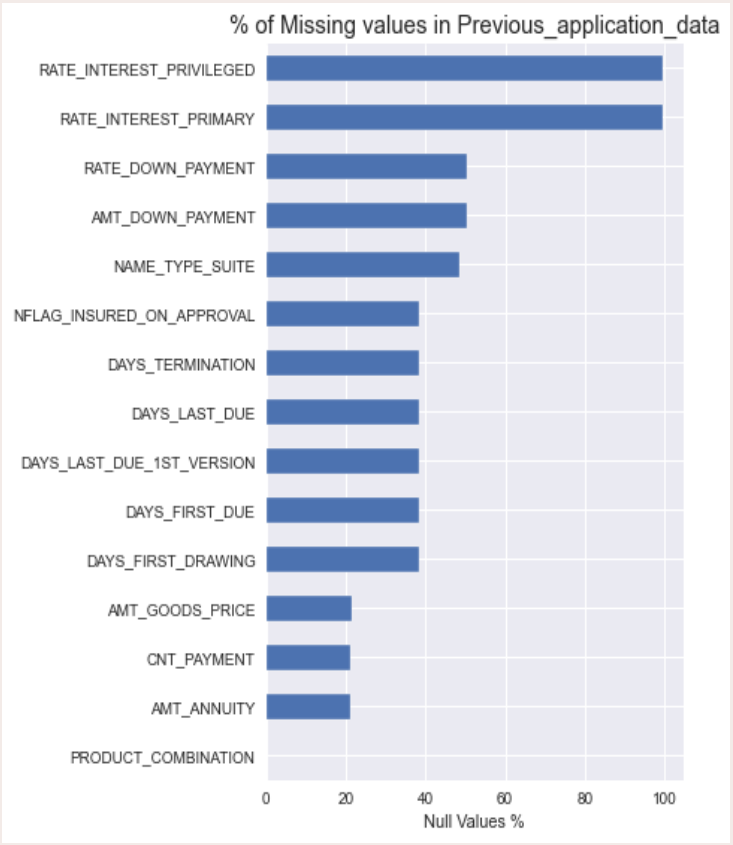
- **Software**: Python 3.10.12
- **Version**: Jupyter Notebook 6.4.6
- **Purpose**: Data cleaning and visualization

I used **Python** and **Jupyter Notebook** to clean and visualize data because these tools are **powerful** and **versatile**. Python is a general-purpose programming language that can be used for a variety of tasks, including data cleaning and visualization. Jupyter Notebook is an **interactive environment** that makes it easy to write and run Python code.

Data Cleaning



A. Handling Missing values of Application data & Previous Application data



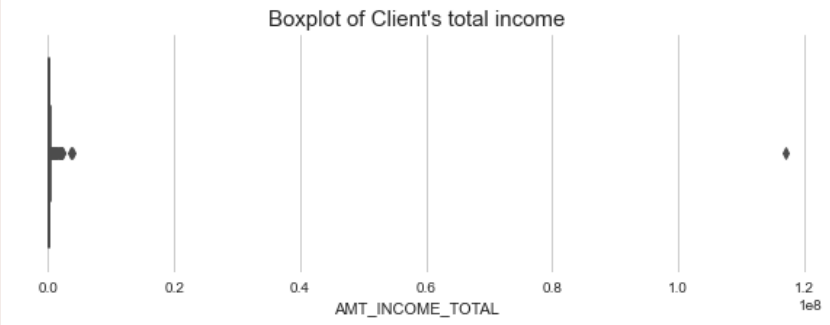
APPLICATION DATA

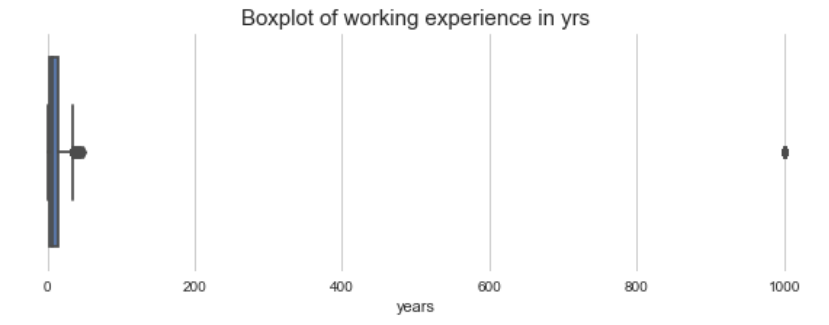
- **removed** all the **null values** from the application_data dropping **52 columns** and **194 rows**
- all variable with missing values **greater than 40%** are **dropped.**
- variables with missing values **below 13%** are **imputed** with **median** values.
- **insignificant null rows** of variables are **removed.**

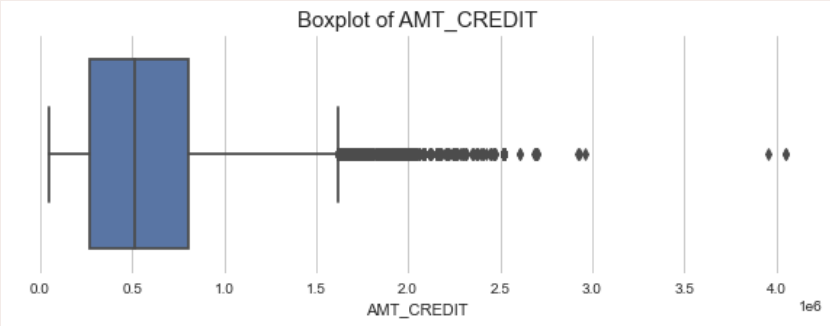
PREVIOUS APPLICAION DATA

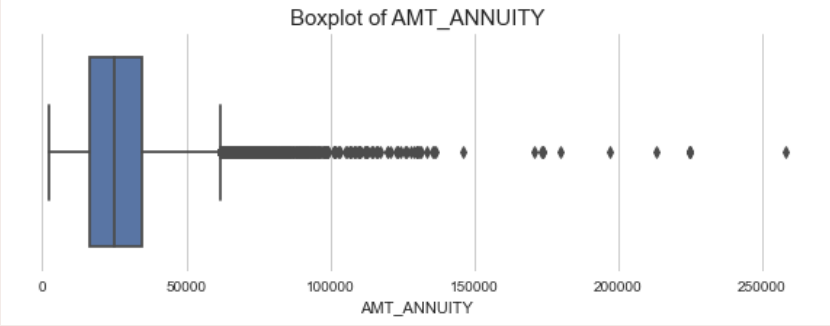
- **removed** all the **null values** from the previous_application_data dropping **11 columns** and **8 rows**
- all variable with missing values **greater than 38%** are **dropped.**
- and rest missing values are imputed with **median values** and **insignificant rows removed.**

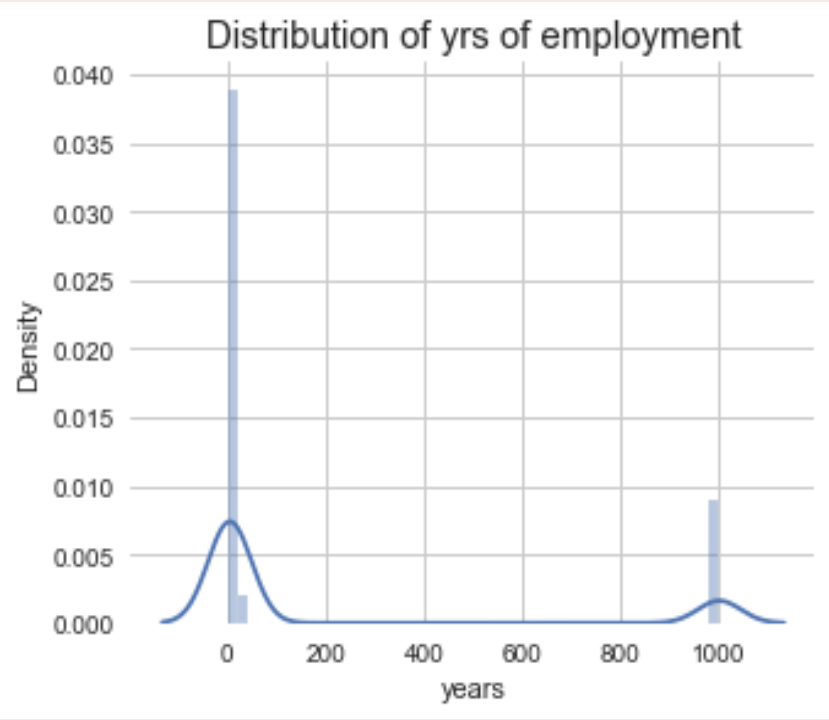
Data Cleaning







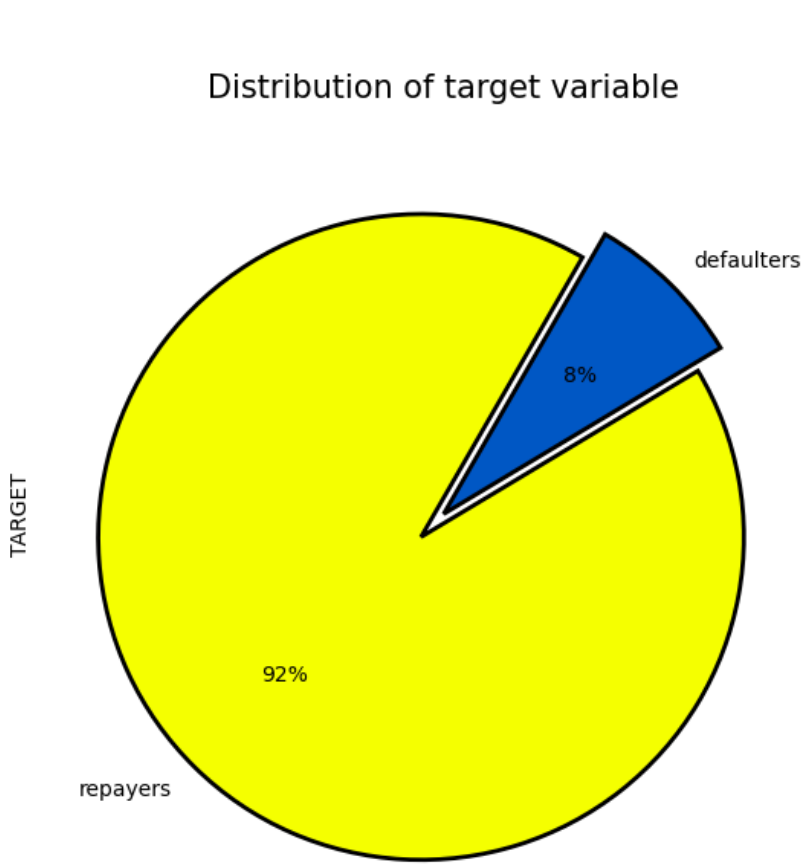


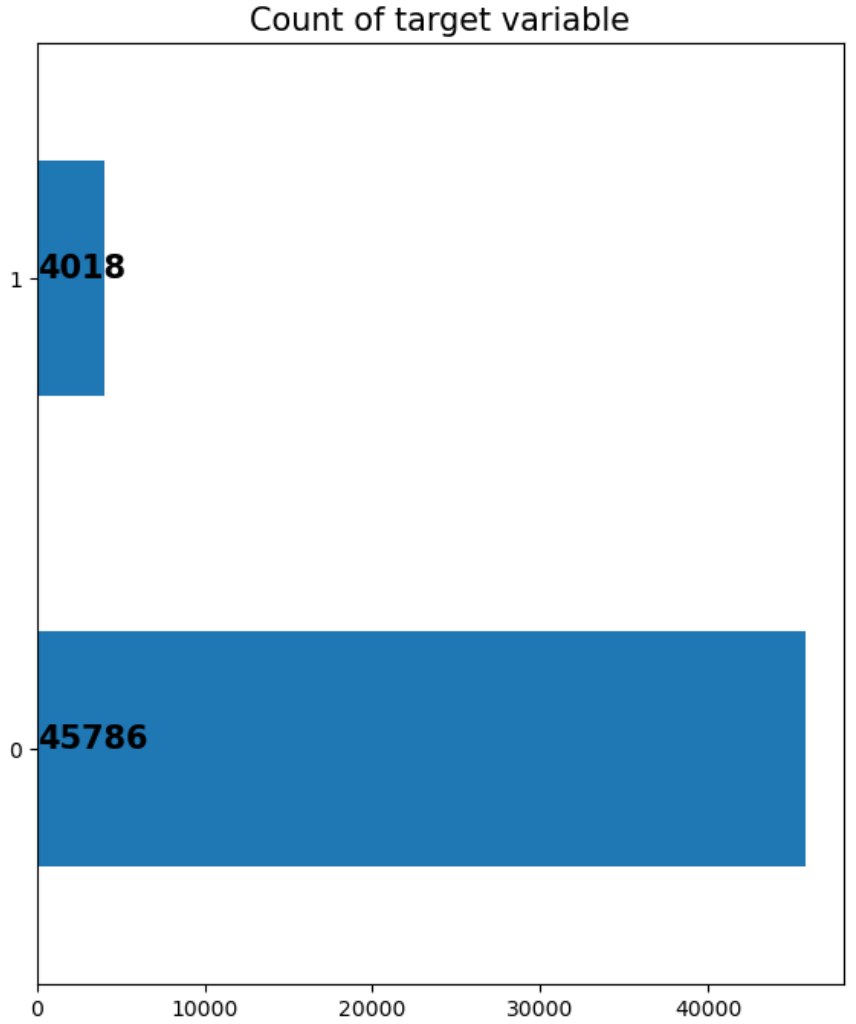


B. Handling Outliers of Application data & Previous Application data

- removed client with **exceptionally high income** as it can distort our analysis for general trend.
- **replaced invalid** values of years_employed
  ie. **"1000yrs"** by **median** value.
- created bucket for client's age, income and credit amount.
- **rest** all outlier values **seems valid** as they are variables with just very **high skewed data**, thus leaving them as it is & **using quantiles** for analysis instead means.

Data Imbalance



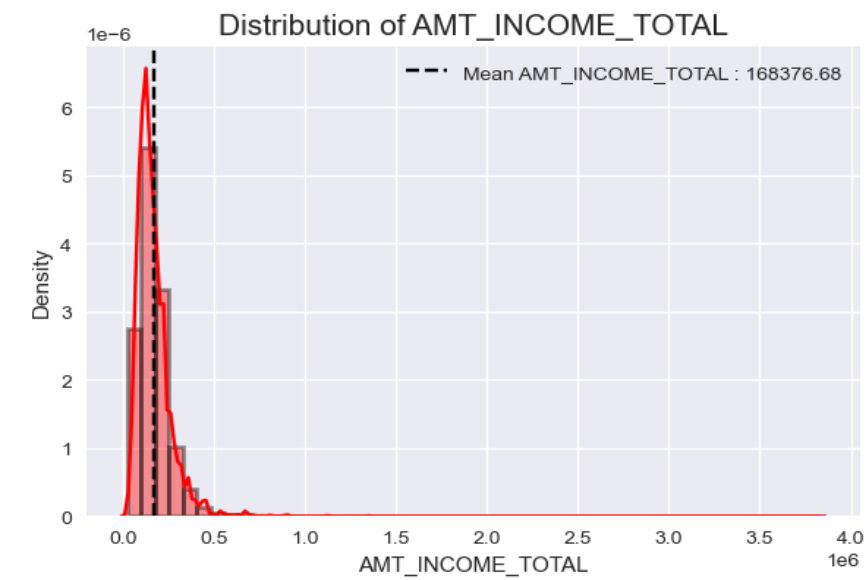


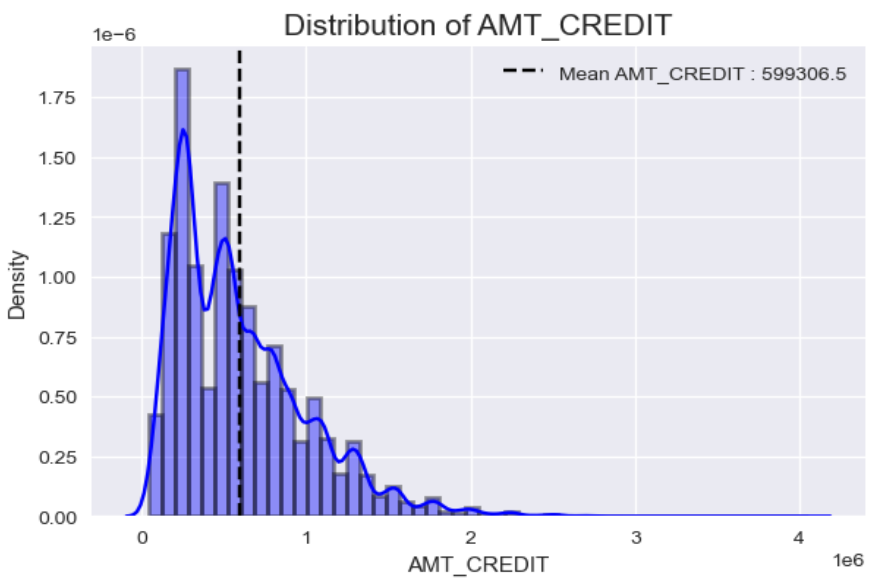
C. Data Imbalance analysis

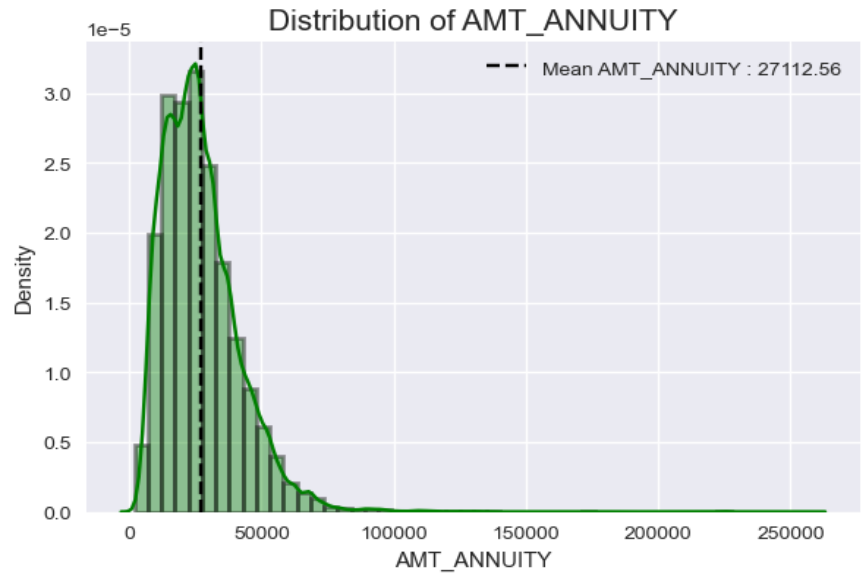
- **Default rate** is **0.08** in the application_data

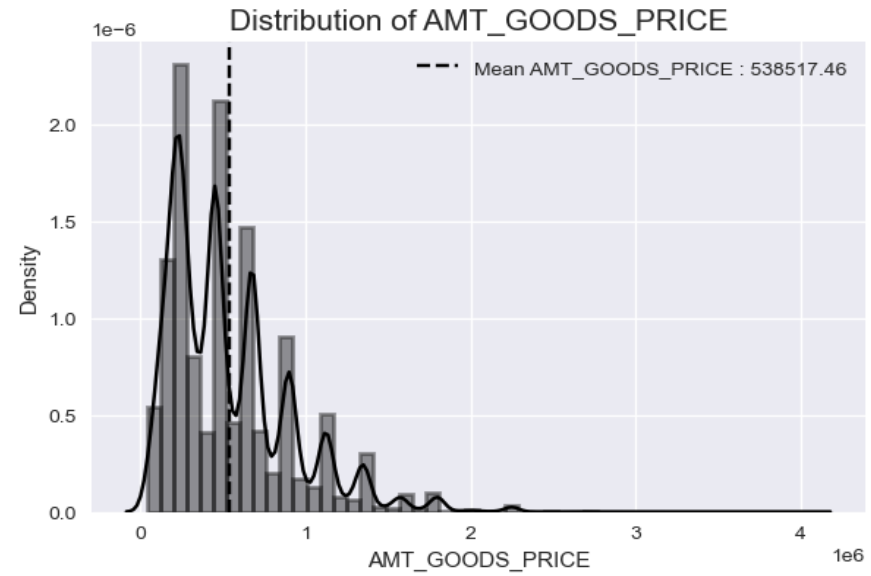
- application_data is **highly imbalanced** with **only 8%** data belongs to **dafault(1)** category and **92%** data are in **repayers(0)** category.

Insights





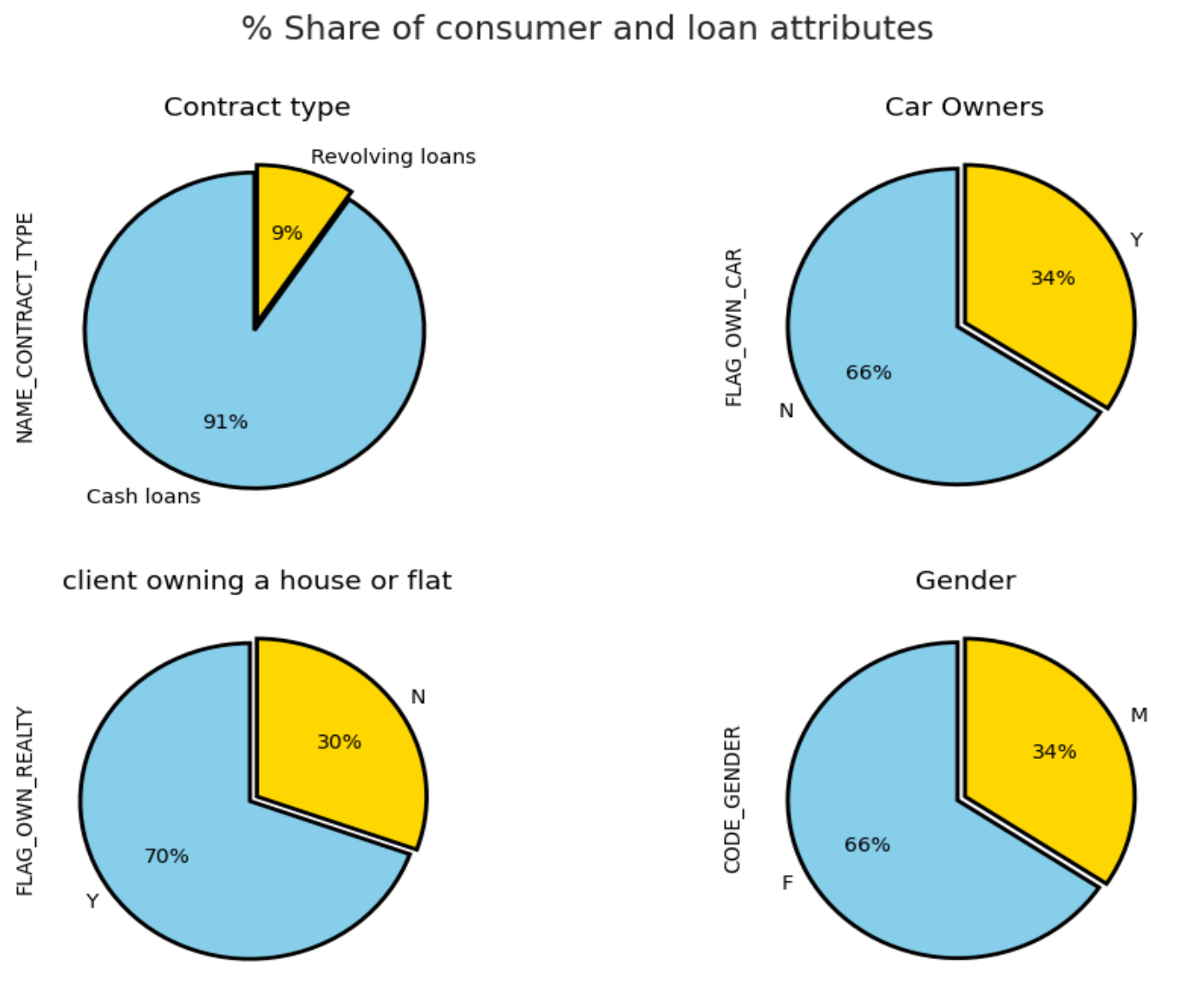




D. Univariate Analysis of numerical variables

- **income** of the client's are **highly skewed**, about **99%** of the borrowers have income **less than 5lakhs**.
- **Credit** amount of the loan is mostly **less** then **10 lakhs**
- Most **people** have **annuity below 50,000** for the loan
- Most no. of loans are given for **goods** price **below 10 lakhs**

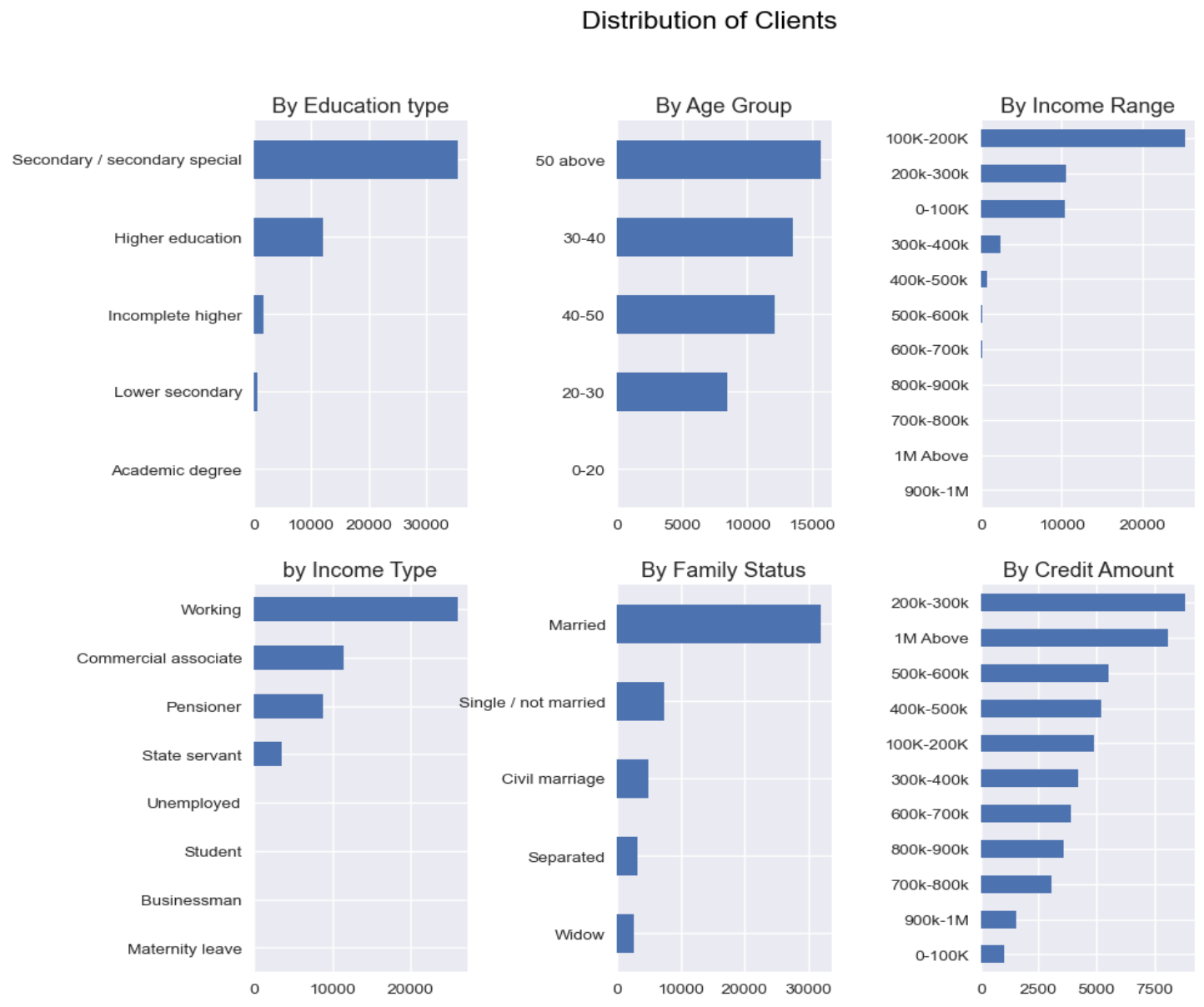
66% Clients of the bank are Females, 70% have house or flat



D. Univariate Analysis of categorical variables

- most of the contract type is **cash loans(91%)**
- only **34%** clients are **car owners**
- majority clients(**70%**) have **house or flat.**
- and majority clients(**66%**) are **females**

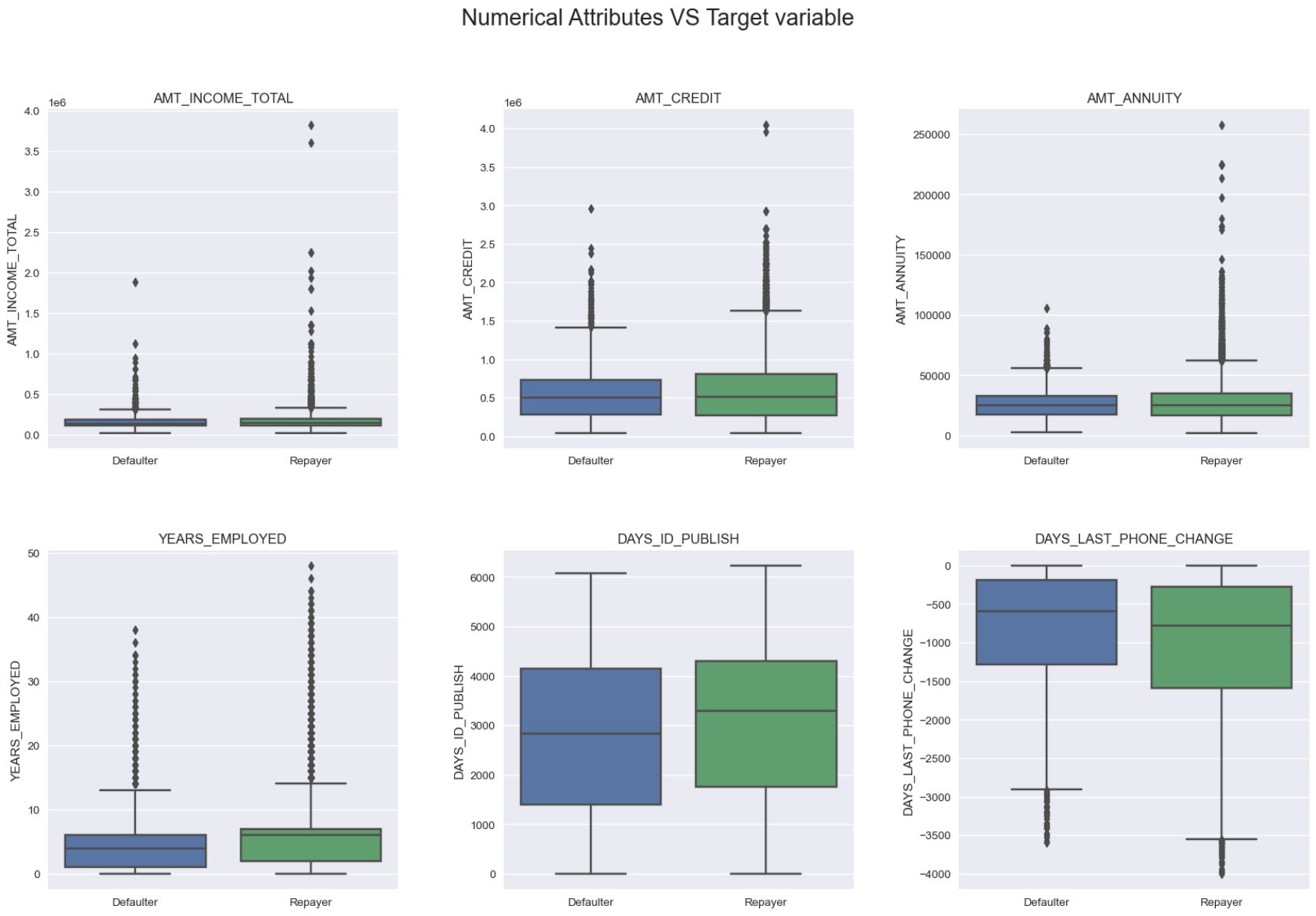
Maximum clients of the bank are above 40 yrs. old, working type and married



D. Univariate Analysis of categorical variables

- **secondary** & **Higher** education are **top education type** of clients
- most clients are in the **age** group of **30-40 and 50 above**
- most clients are in the range of **100-200k income**
- most clients are **credited amount** in the range of **200k-300k**
- **working type** and **married type** are **top** clients

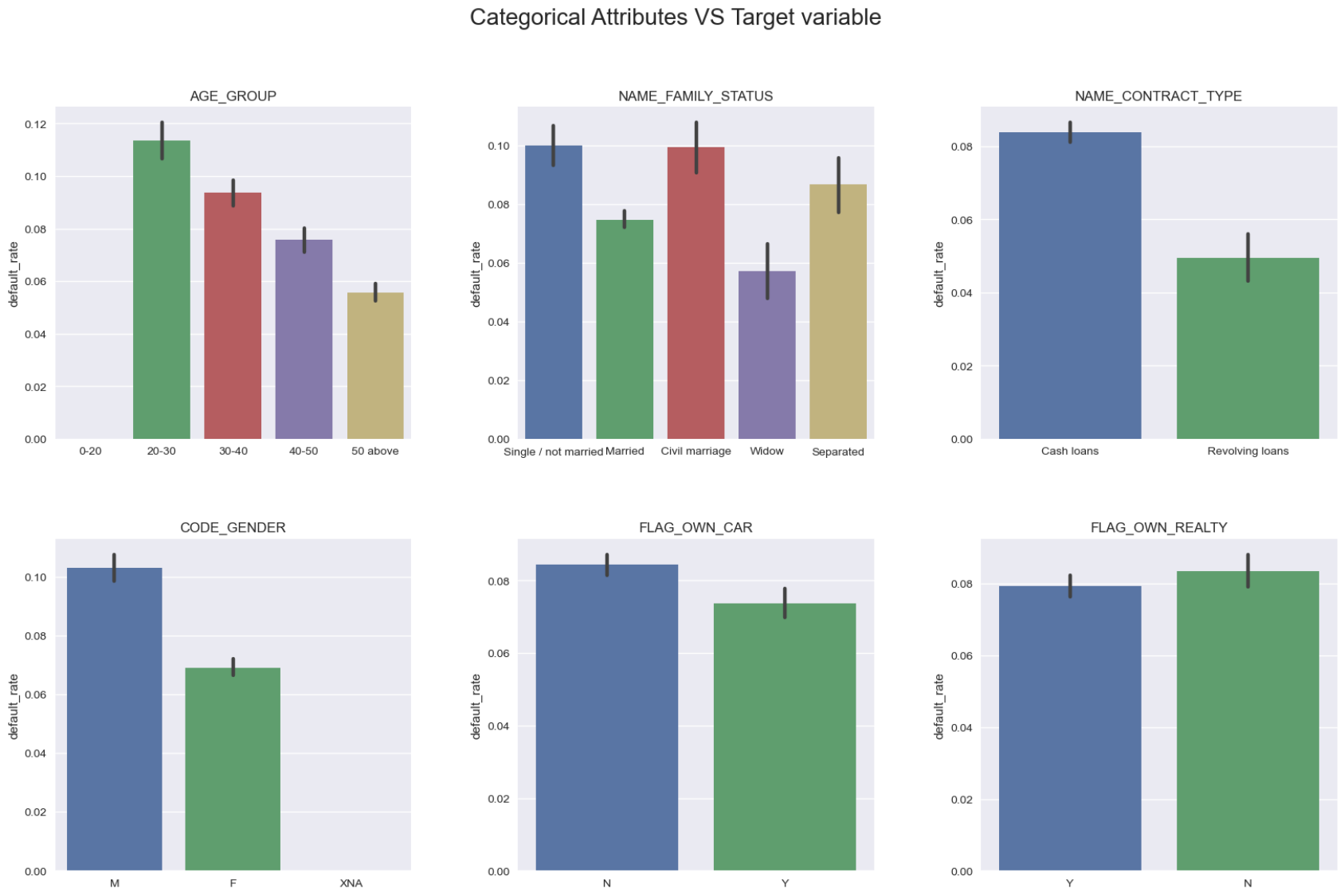
Median income of defaulters are same as repayors, but annuity, years of employment has lower values



D. Bivariate analysis to explore relationships between variables and the target variable

- **median income** of **defaulters** are **same** as repayors, **very high income** clients are **not defaulters**.
- **median credit** amount of defaulters are **same** as repayors, very **high** credit amount clients are **not defaulters**.
- **defaulters** have **median annuity** amount, **employment yrs**, **days last id updated** values are **slightly lower** than repayors
- defaulters have **median days** last **phone change** value is **higher** than repayors

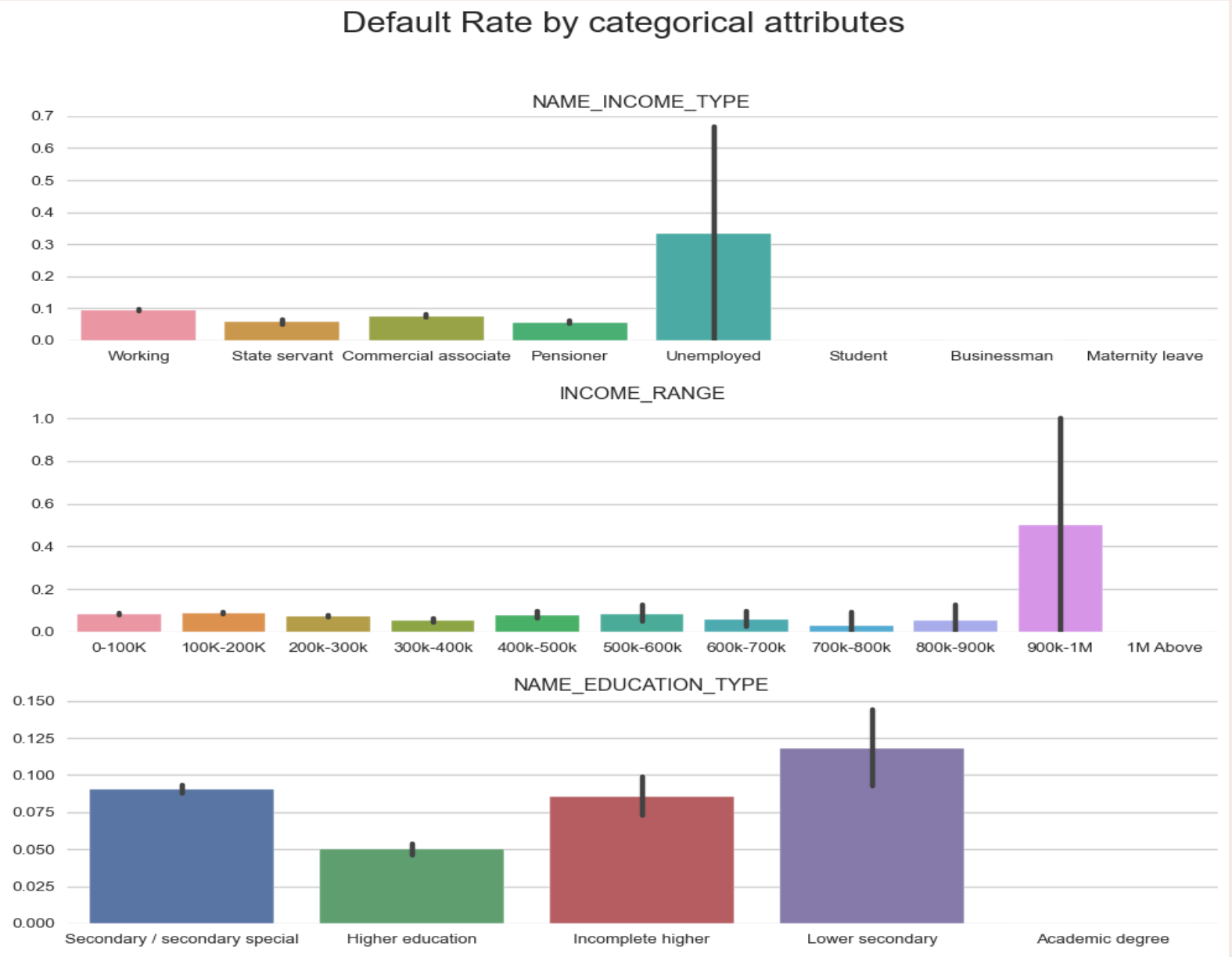
20-30 age group, civil married, singles have high default rate



D. Bivariate analysis to explore relationships between variables and the target variable

- clients in **20-30 age group** have **highest default rate**
- **civil married** and **single/not married** clients have **high default rate**
- **cash loans** contract type, people **not owning car** and **house** have **higher default rate**
- **males** have **higher default** rate than females.

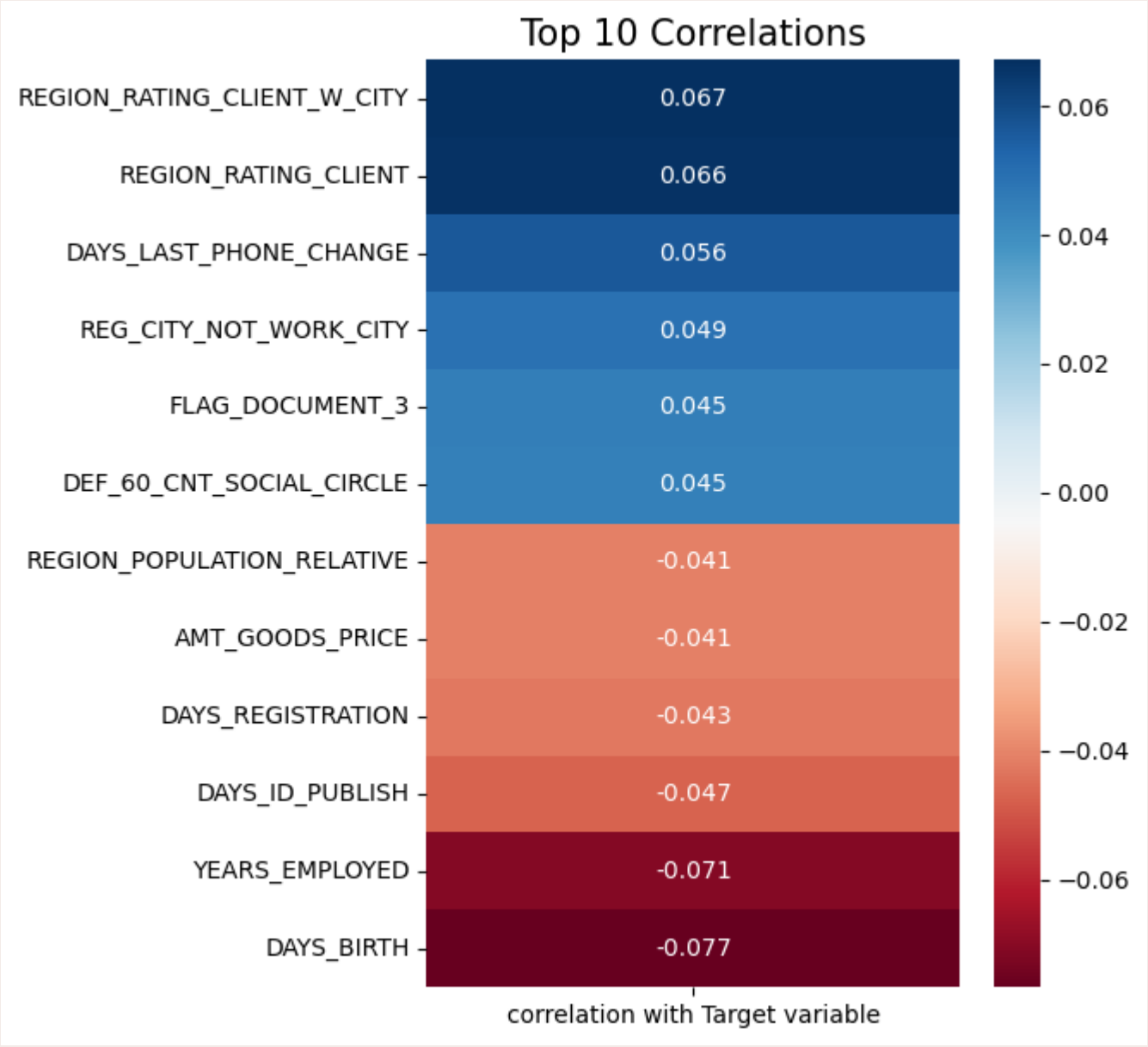
Unemployed, lower secondary and incomplete higher education type clients have high default rate



D. Bivariate analysis to explore relationships between variables and the target variable

- **unemployed** clients have the **highest default rate**
- people having **income** in the **category 900k-1m** have the **highest** default rate which is unexpected.
- **lower secondary** and **incomplete higher education** type clients have **higher default rate**.

Client's age, years of employment, Region rating, days last phone change are highly correlated with target variable



E. correlation between variables and the target variable

- target variable is **positively correlated** with REGION_RATING_CLIENT, DAYS_LAST_PHONE_CHANGE, REG_CITY_NOT_WORK_CITY, FLAG_DOCUMENT_3 etc.

- target variable is **negatively correlated** with DAYS_BIRTH, YEARS_EMPLOYED, DAYS_ID_PUBLISH, DAYS_REGISTRATION etc.

Results

- Conclusions and recommendations :

  - Clients with **High Region rating**, **not provided document 3**, **low days employment**, **days ID published**, **days registration** should be **denied** loans as they are at high risk of default.

  - Clients with **low annuity amount**, **unemployed**, **lower secondary or incomplete education type**, **single** or **civil married**, belongs to **age group 20-30** should be **lend** at **high interest** rate or **lower amount** as they are also at moderate risk of default.

  - Clients with **very high income**, **annuity**, **credit amount**, having **higher education**, owns **car**, **house** or **flat** are **good customers** for the bank. The bank should make strategies to **attract them** and make **loan process easier** for them.

Thank you