CREDIT EDA ASSIGNMENT

PRESENTED BY -- NIPUN GARG

Contents

- Problem Statement
- Work Flow
- Reading Dataset
- Handling Null Values
- Outliers Analysis
- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis
- Top Correlation
- Conclusion

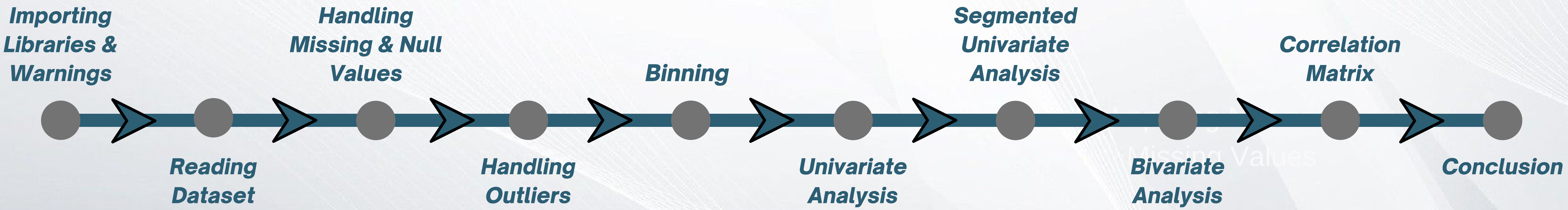
Problem Statement

Agenda

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as

- Denying the loan
- Reducing loan amount
- Lending (to risky applicants) at a higher interest rate

Work Flow



Importing Libraries & Warnings

Reading Dataset

Handling Missing & Null Values

Handling Outliers

Binning

Univariate Analysis

Segmented Univariate Analysis

Bivariate Analysis

Correlation Matrix

Conclusion

Reading Dataset

- The feature named as Target is our target variable which tells if client has made timely payments as 0 and if client has made default in payments as 1.

- There are 2 dataset available 'application_data.csv' and 'previous_data.csv'.

- The 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

- The 'previous_application.csv' contains information about the client’s previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

- Read dataset info, shape, statistical summary in Jupyter NB for analysis.

Handling Null Values

- There are 49 columns in 'application_data.csv' with more than 40% null values in them. These columns were dropped as they could create problem in analysis.

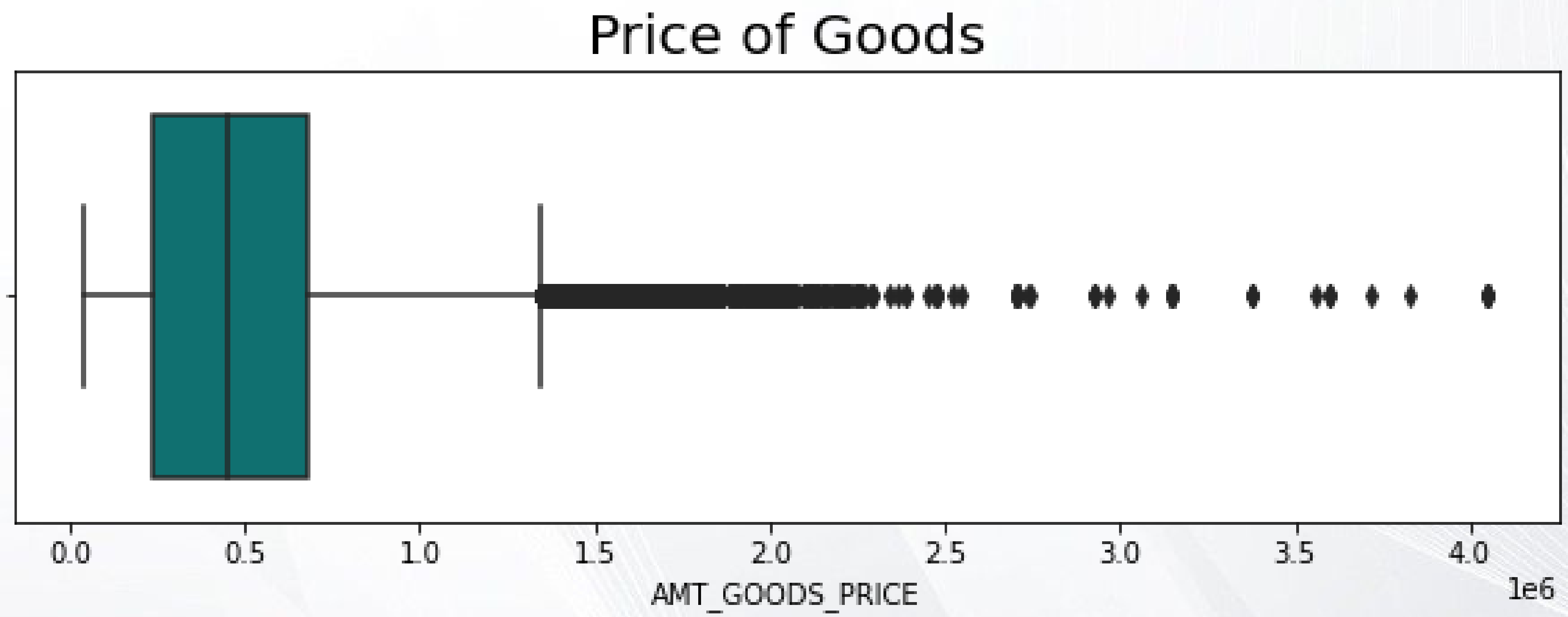
- There is one column OCCUPATION_TYPE with 31.3 % null values. It is a categorical column. Null percentage is high in this column so imputing with any mode will create imbalance in column. So 'Unknown' new category is created for these values.

- Numerical columns has been imputed using median. As there are outliers present in columns.

Outliers Analysis

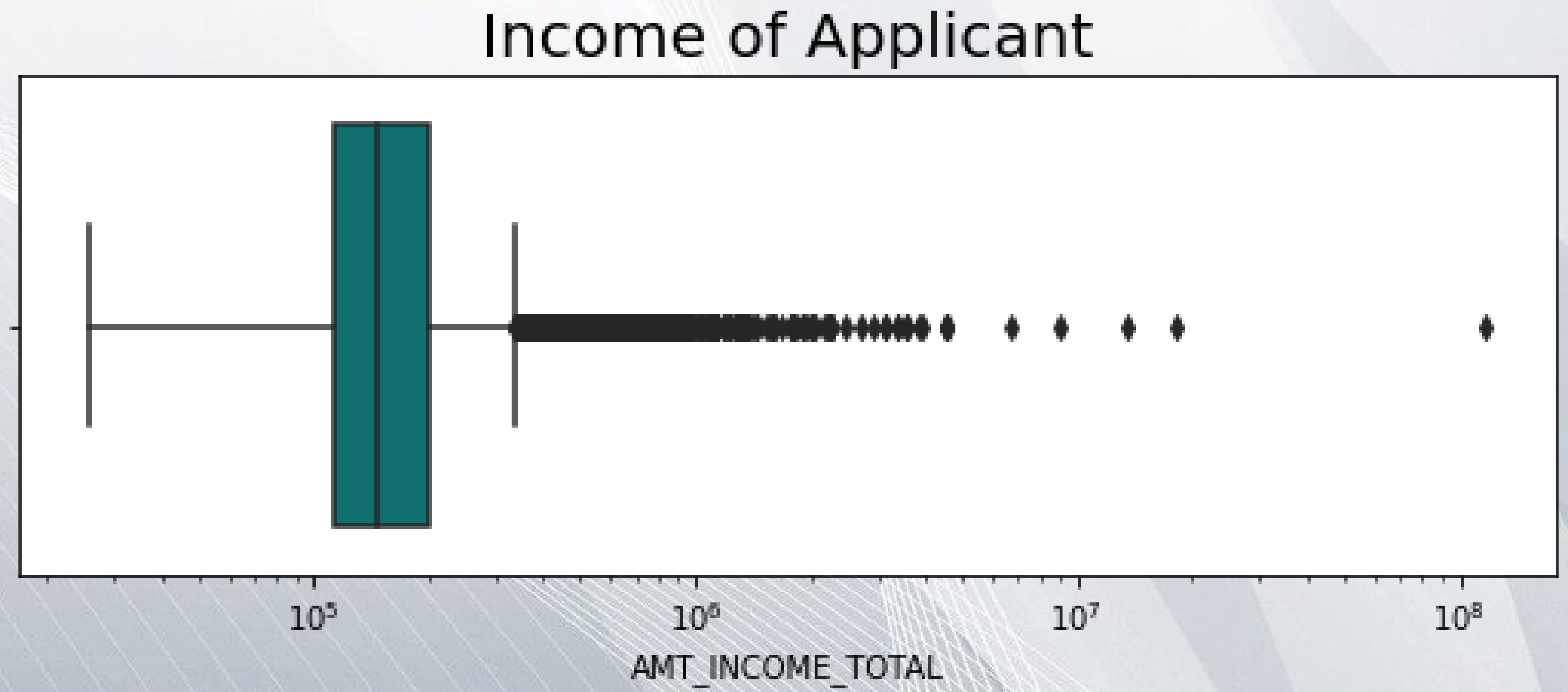
- **AMT_GOODS_PRICE**

Outliers are present in Goods price. Difference between 0.99 and 1.0 is very big. A new categorical variable created to gain insight from AMT_GOODS_PRICE



- **AMT_INCOME_TOTAL**

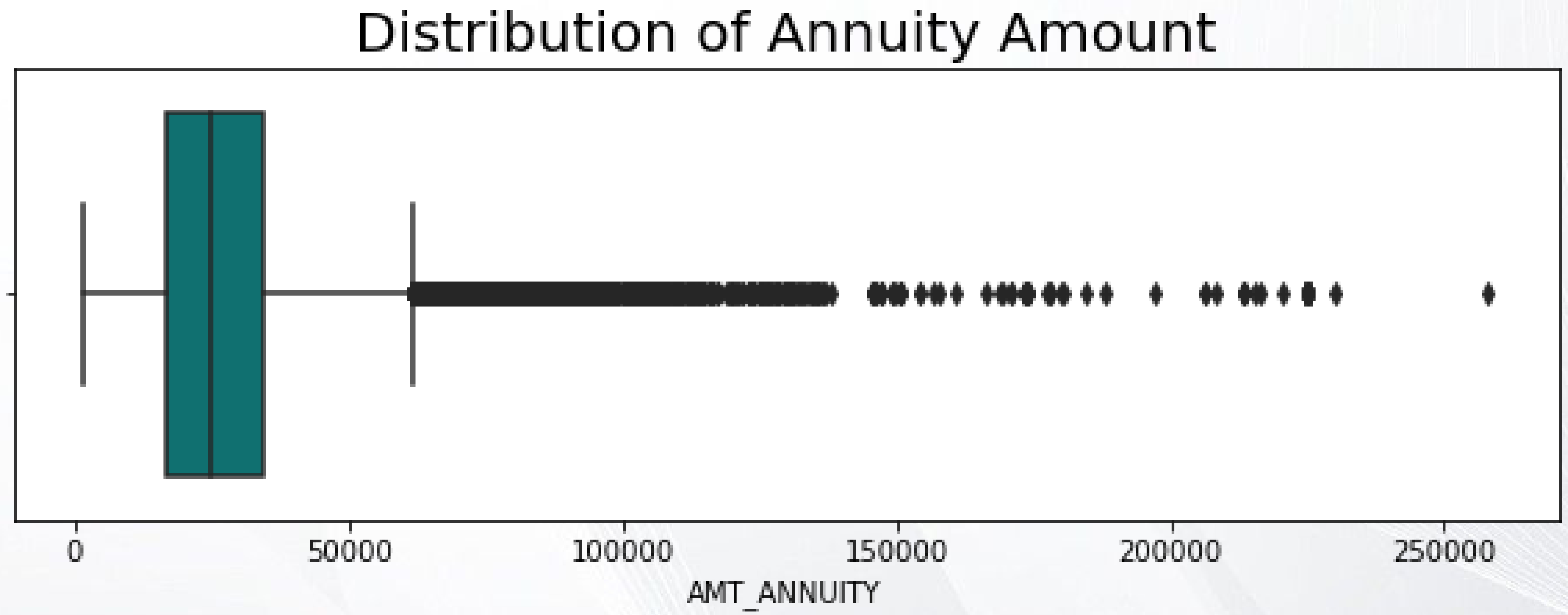
Outliers are present in column. Difference between 99th percentile and 100 percentile is 116527500. So we will create categorical column for analysis



Outliers Analysis

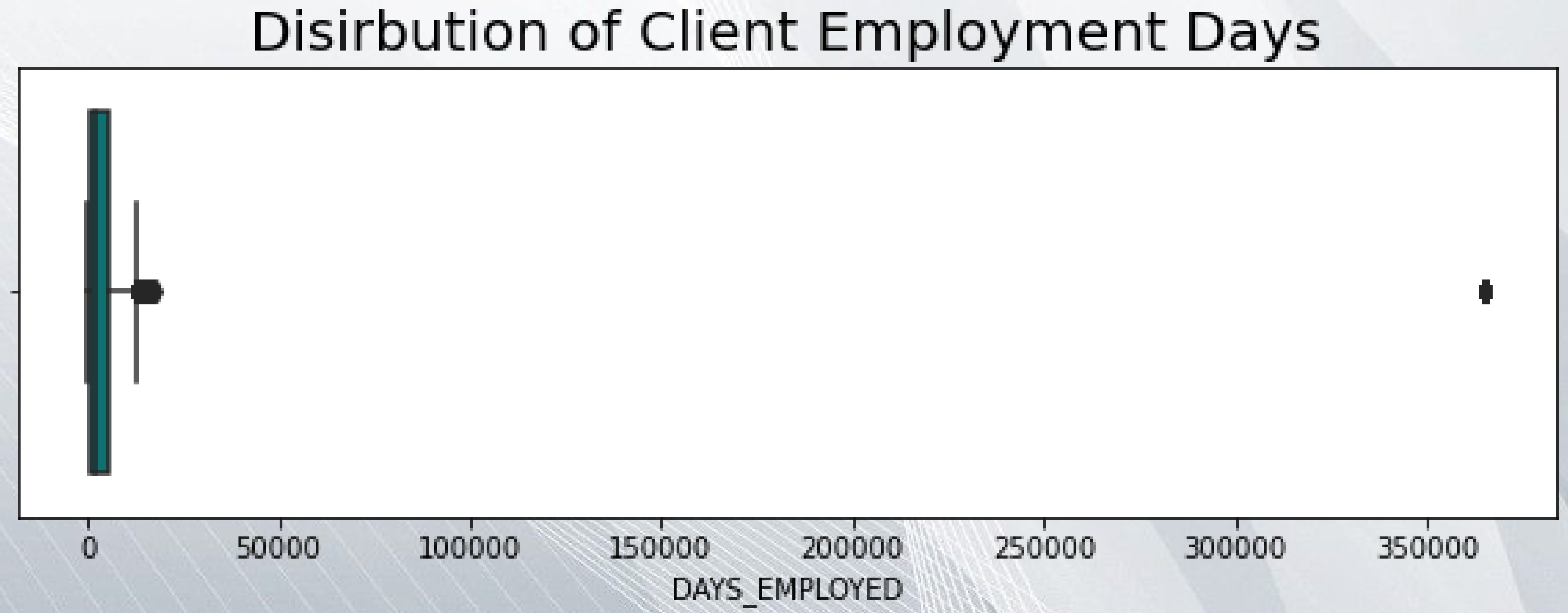
- **AMT_ANNUITY**

Categorical column created for further analysis.



- **DAYS_EMPLOYED**

Outliers are present in DAYS_EMPLOYED and it seems like a invalid value. For this binning into new category is best option.

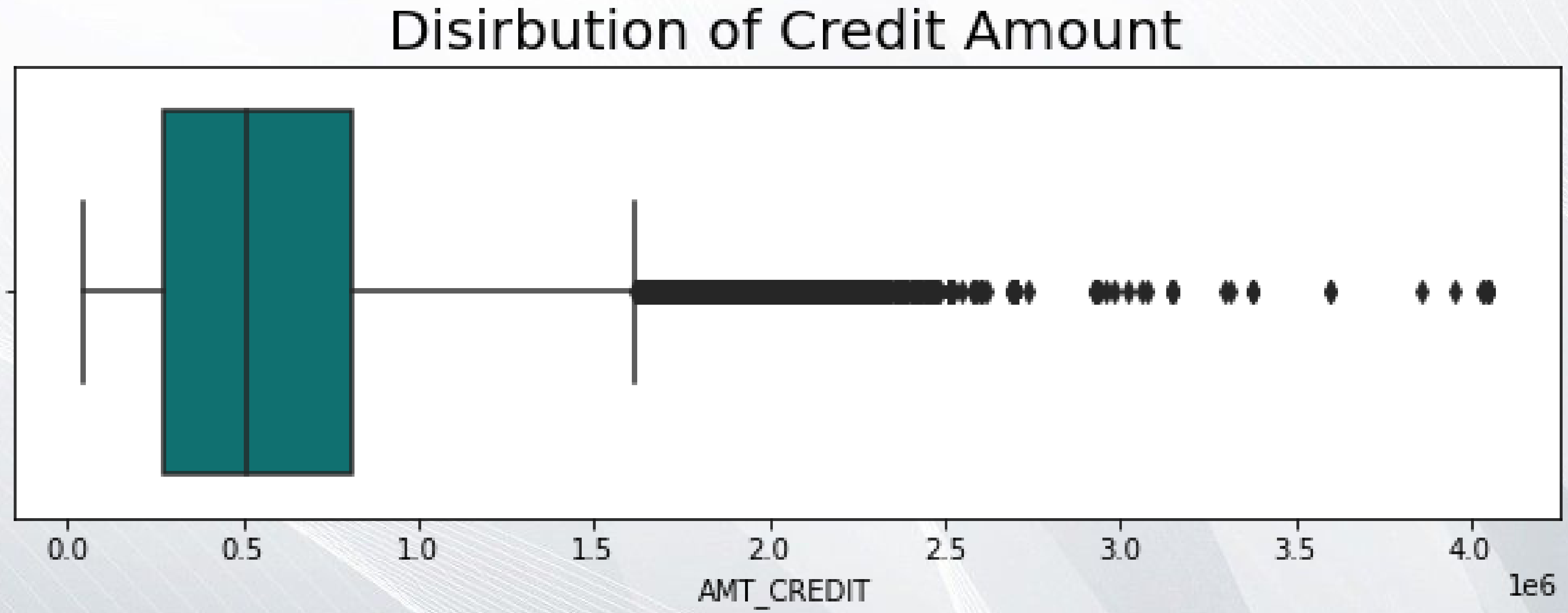


Outliers Analysis

- **AMT_CREDIT**

Outliers are available in Credit amount after 97th percentile.

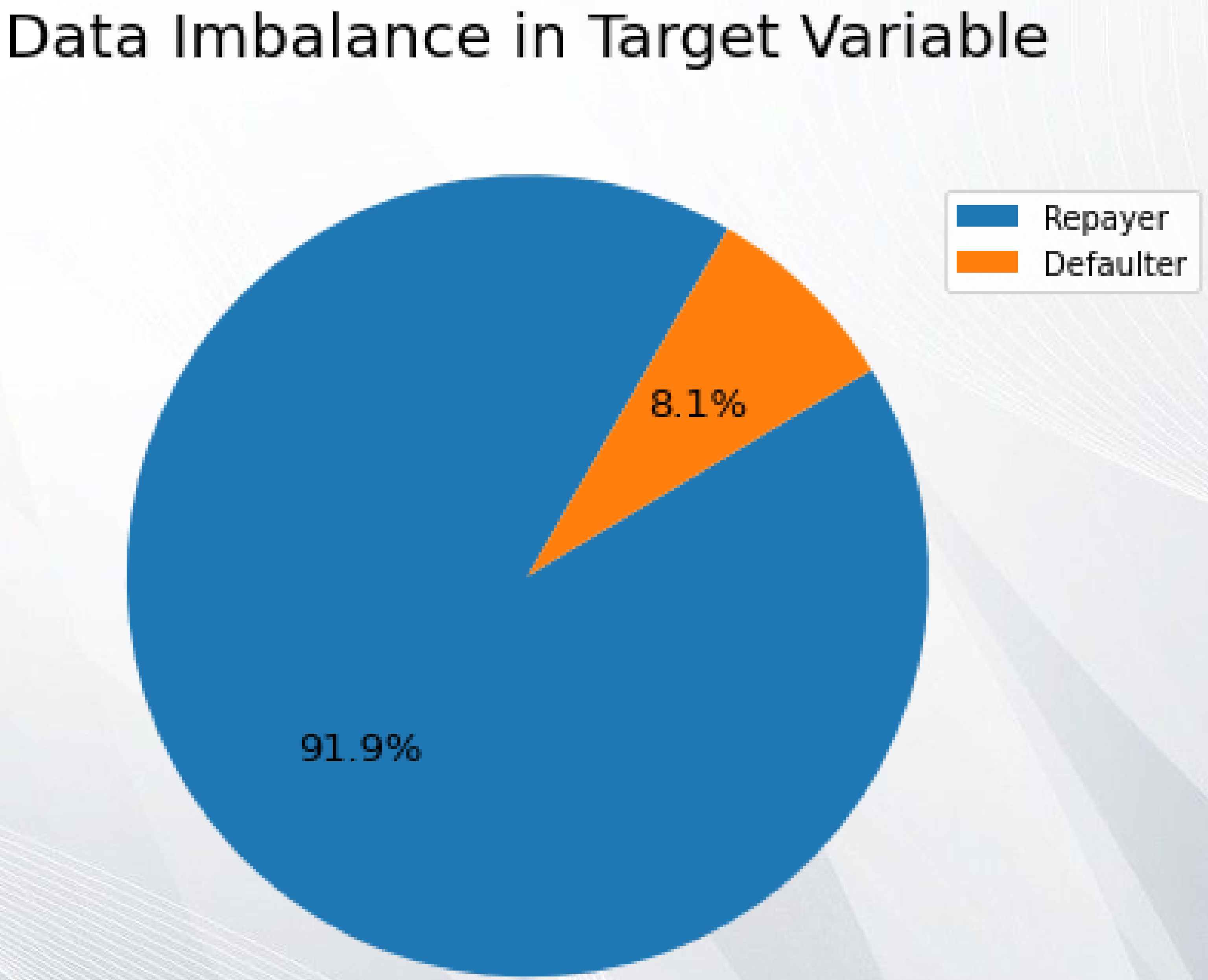
0.75	808650.0
0.85	1024740.0
0.90	1133748.0
0.97	1546020.0
0.99	1854000.0
1.00	4050000.0



Univariate Analysis

- **Data Imbalance**

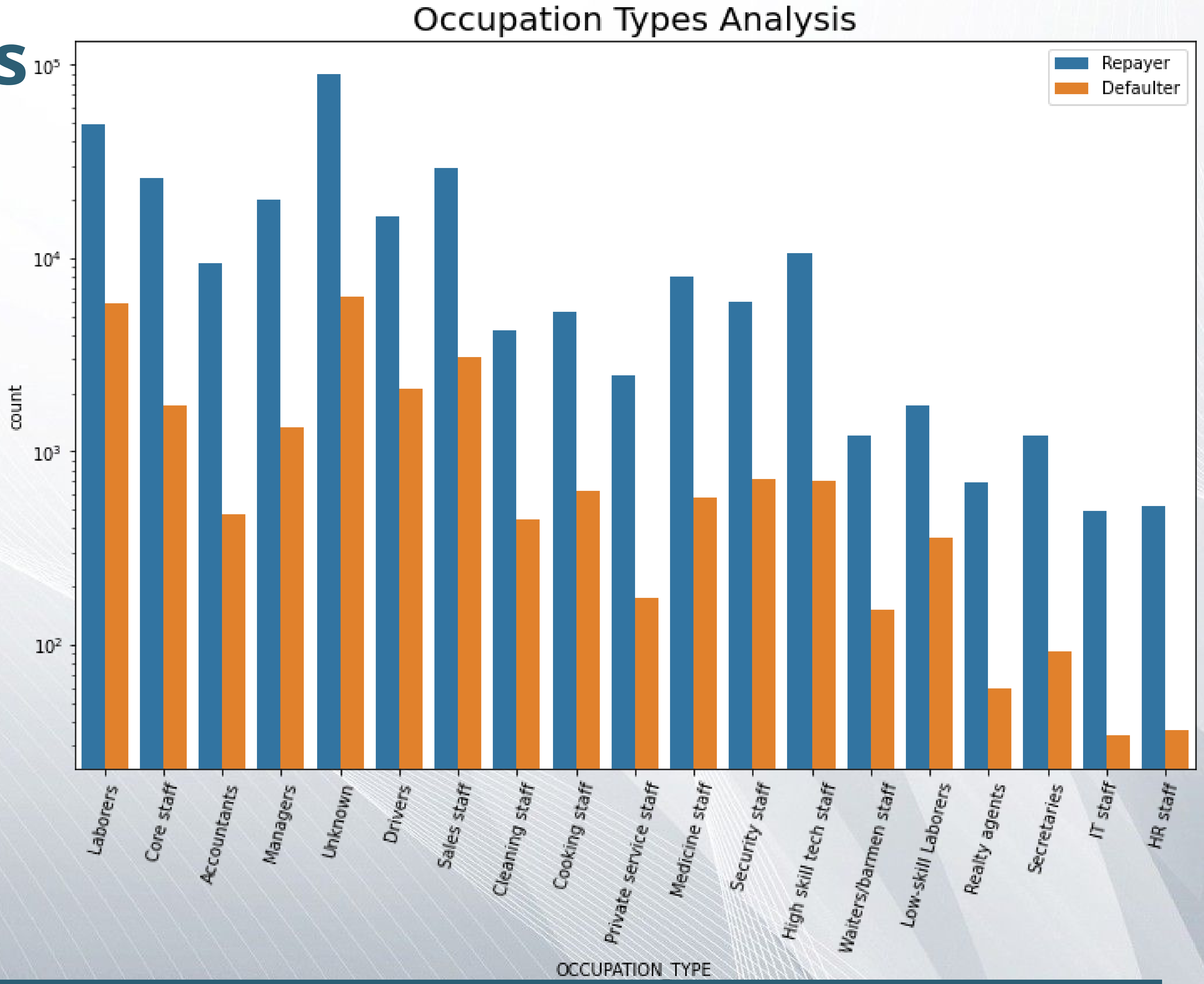
Here we can see repayer's percentage is higher than defaulter's percentage. 91.9% people pay there loan on time while 8.1% people face difficulties in making loan payments on time.



Univariate Analysis

- **Analysis of Occupation Types**

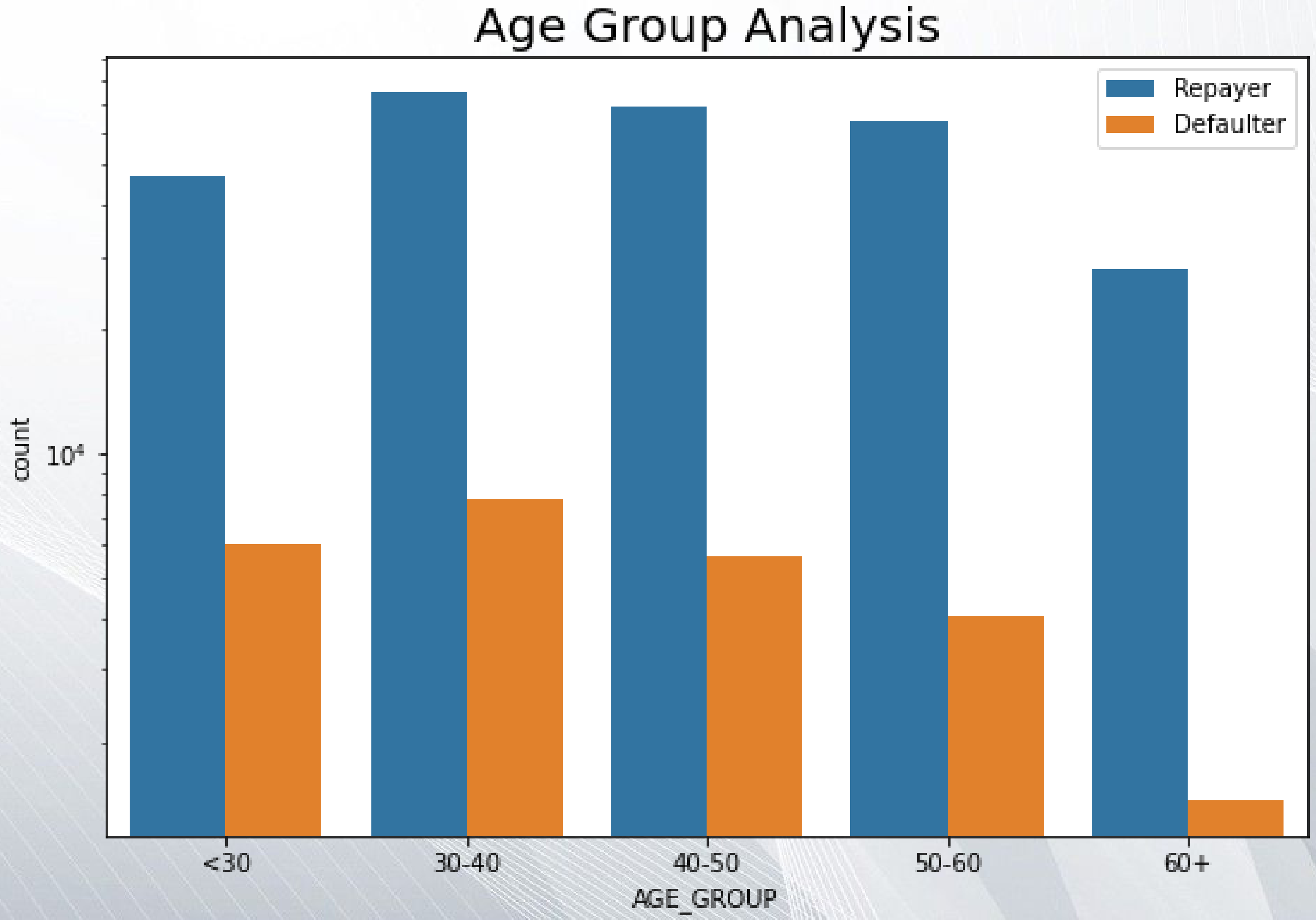
Low-skill laborers, Drivers and laborers categories have the high % of defaulters. IT staff, HR staff, Private Service staff and Accountant categories have low default percentage. Bank should focus more on these categories.



Univariate Analysis

- **Analysis of Age Group**

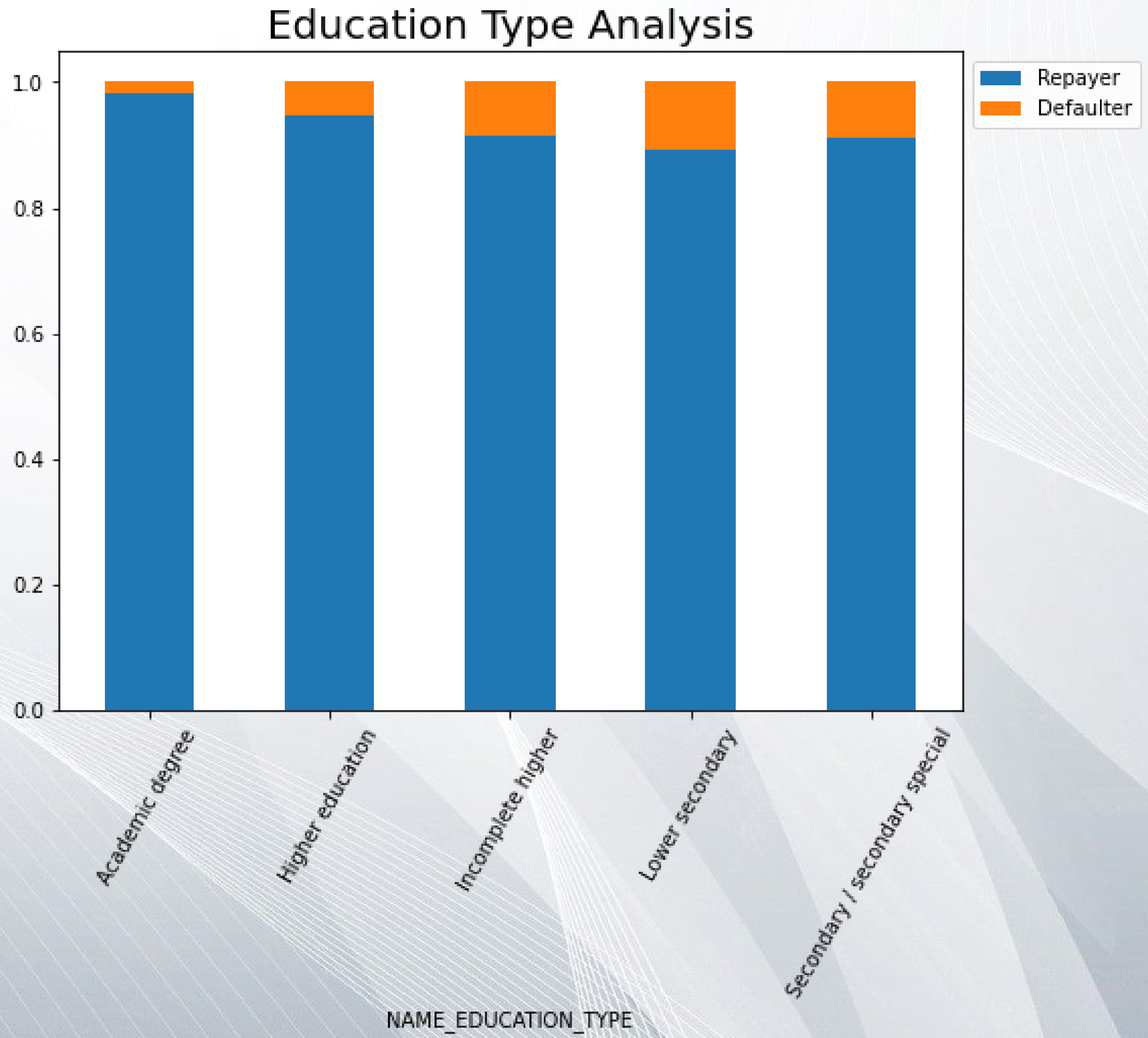
People with age 50 years and above are likely to repay loan than younger people.



Univariate Analysis

- **Education Background Analysis**

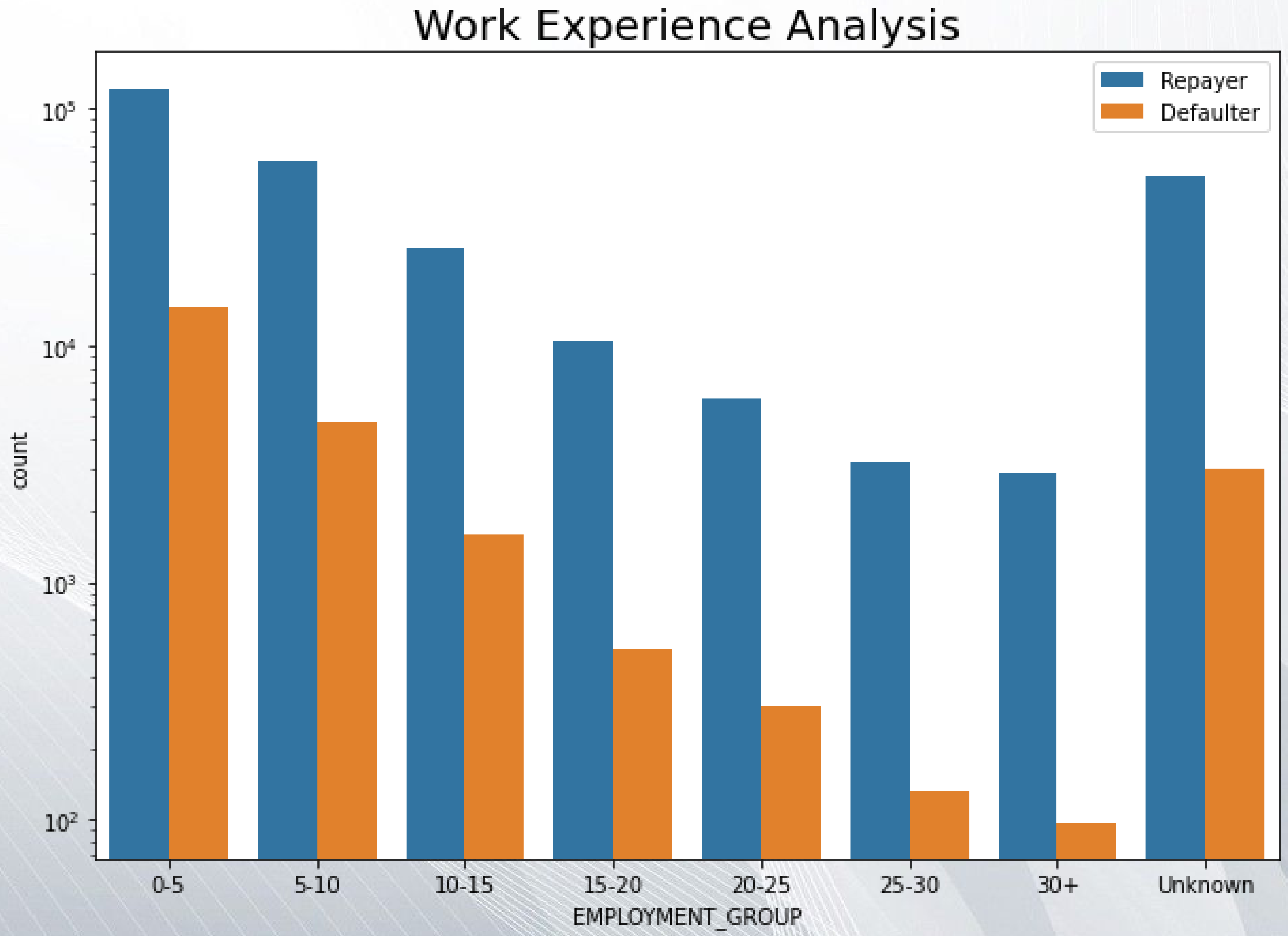
People with higher education repay their loan on time than people with lower education.



Univariate Analysis

- **Work Experience Analysis**

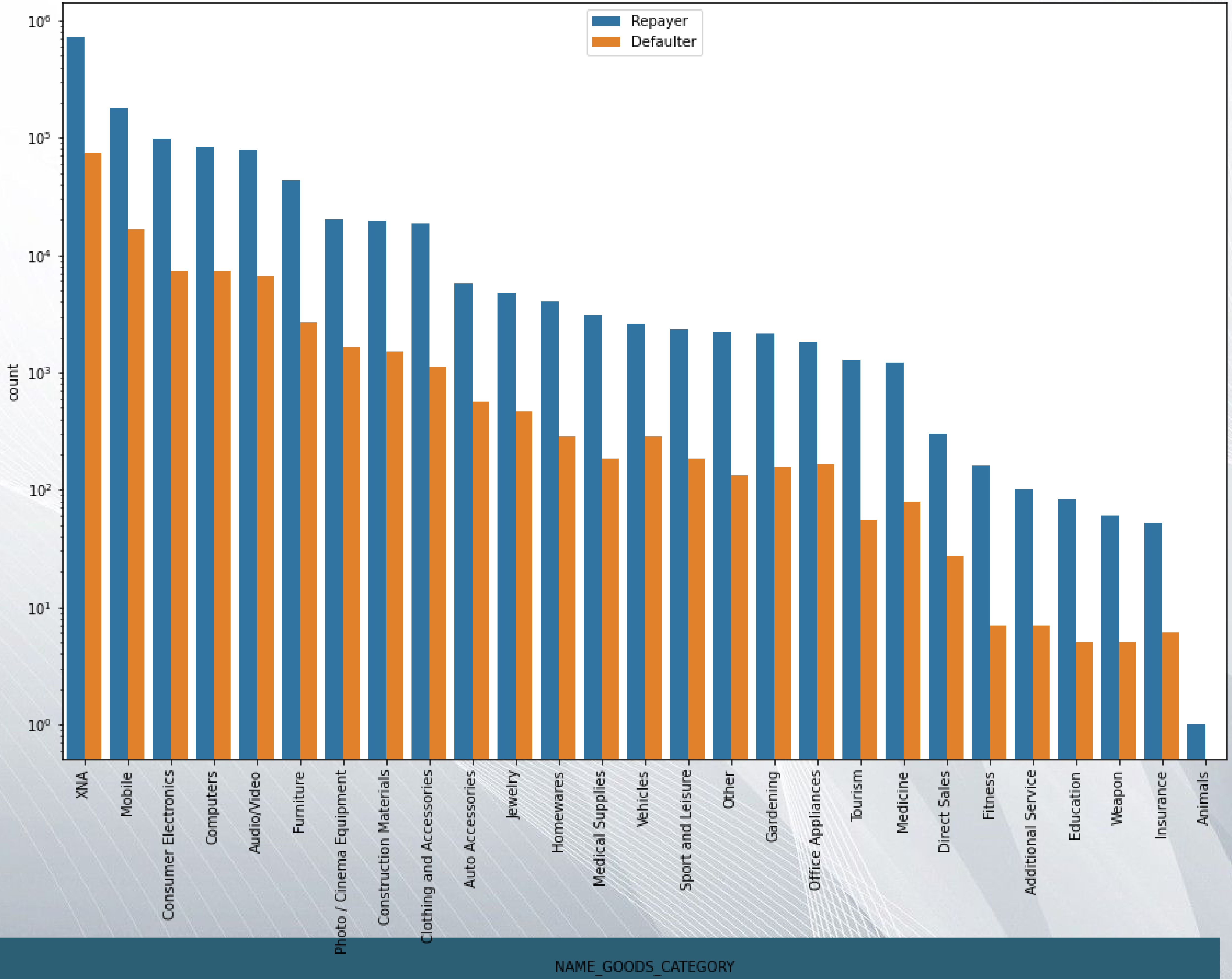
People with more professional experience are paying loan on time. People with work experience 0-5 years have high chances of making payment defaults.



Univariate Analysis on Merged data

- **NAME_GOODS_CATEGORY**

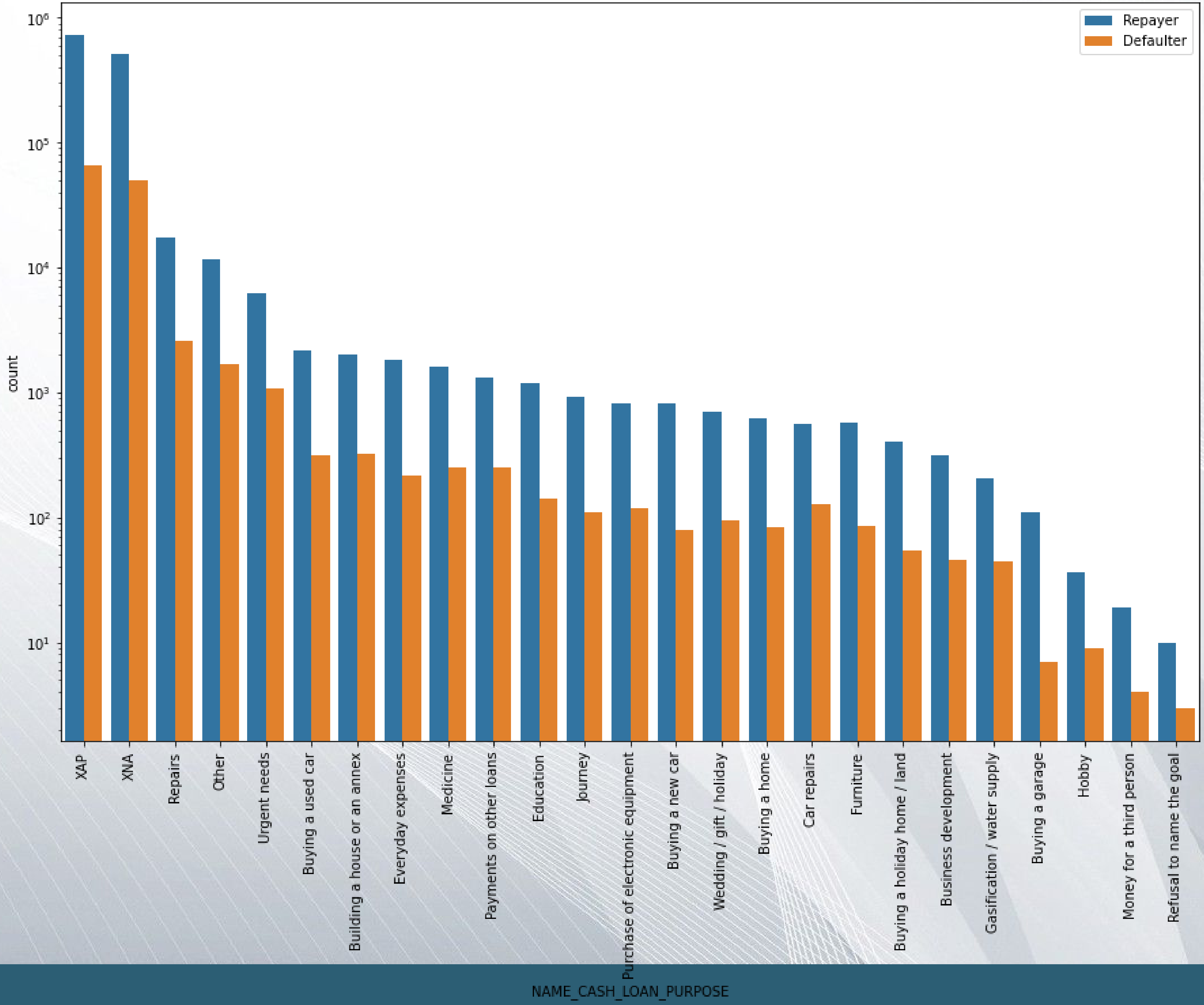
People are taking consumer loans more for buying electronics products.



Univariate Analysis on Merged data

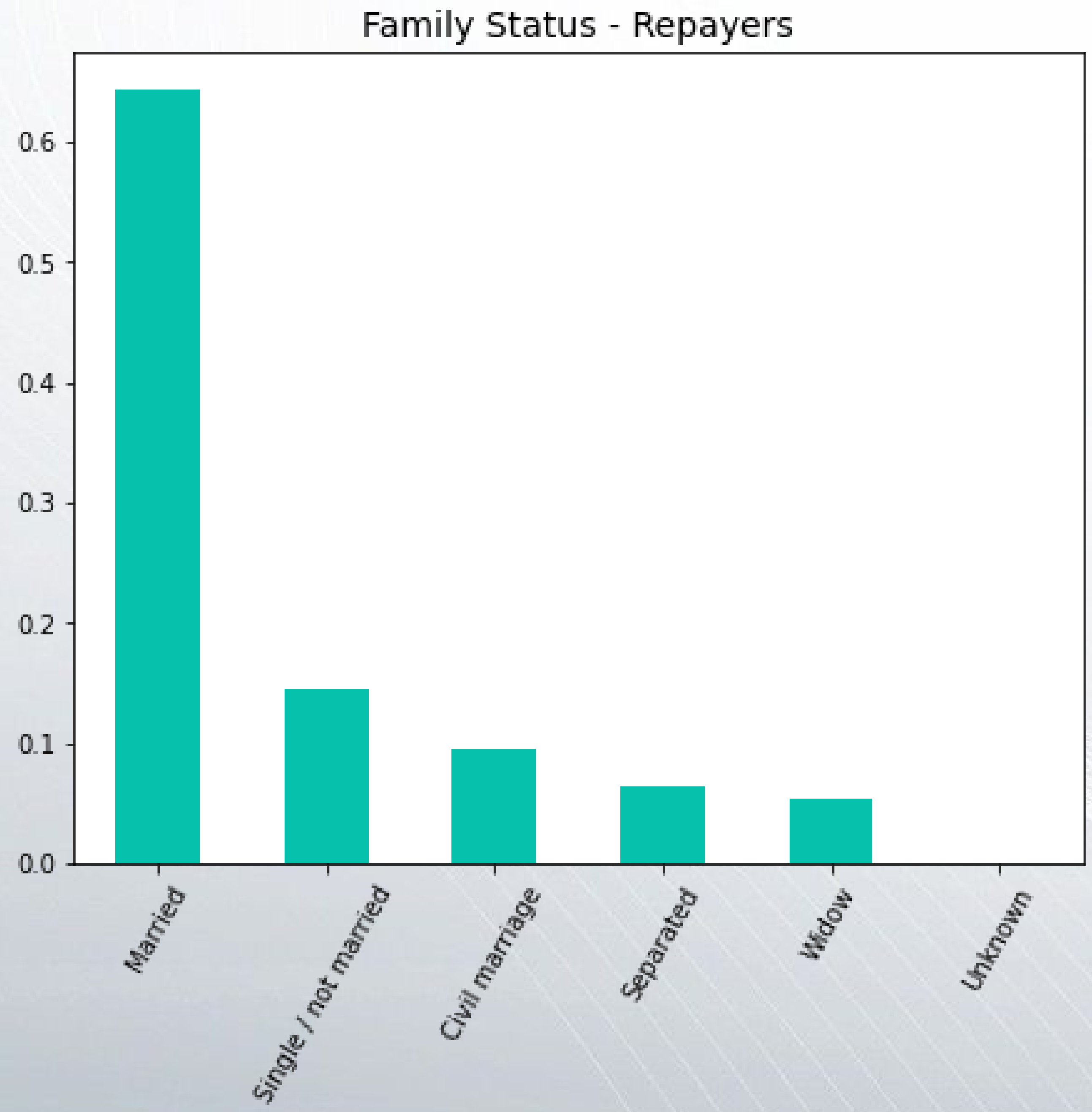
- **NAME_CASH_LOAN_PURPOSE**

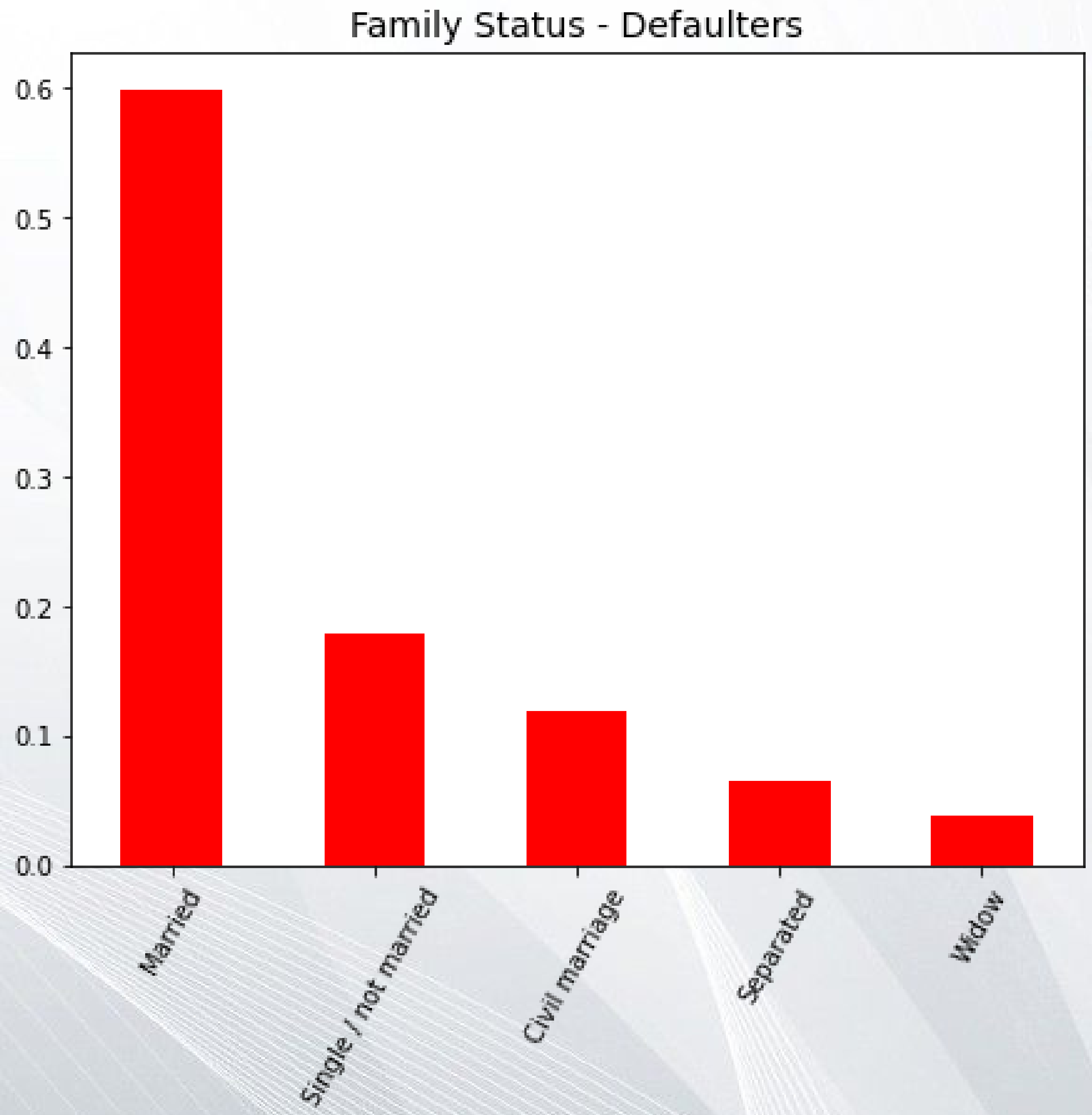
People are taking cash loan for buying property, buying or repairing car, education and for medical purpose.



Segmented Univariate Analysis

- **NAME_FAMILY_STATUS**

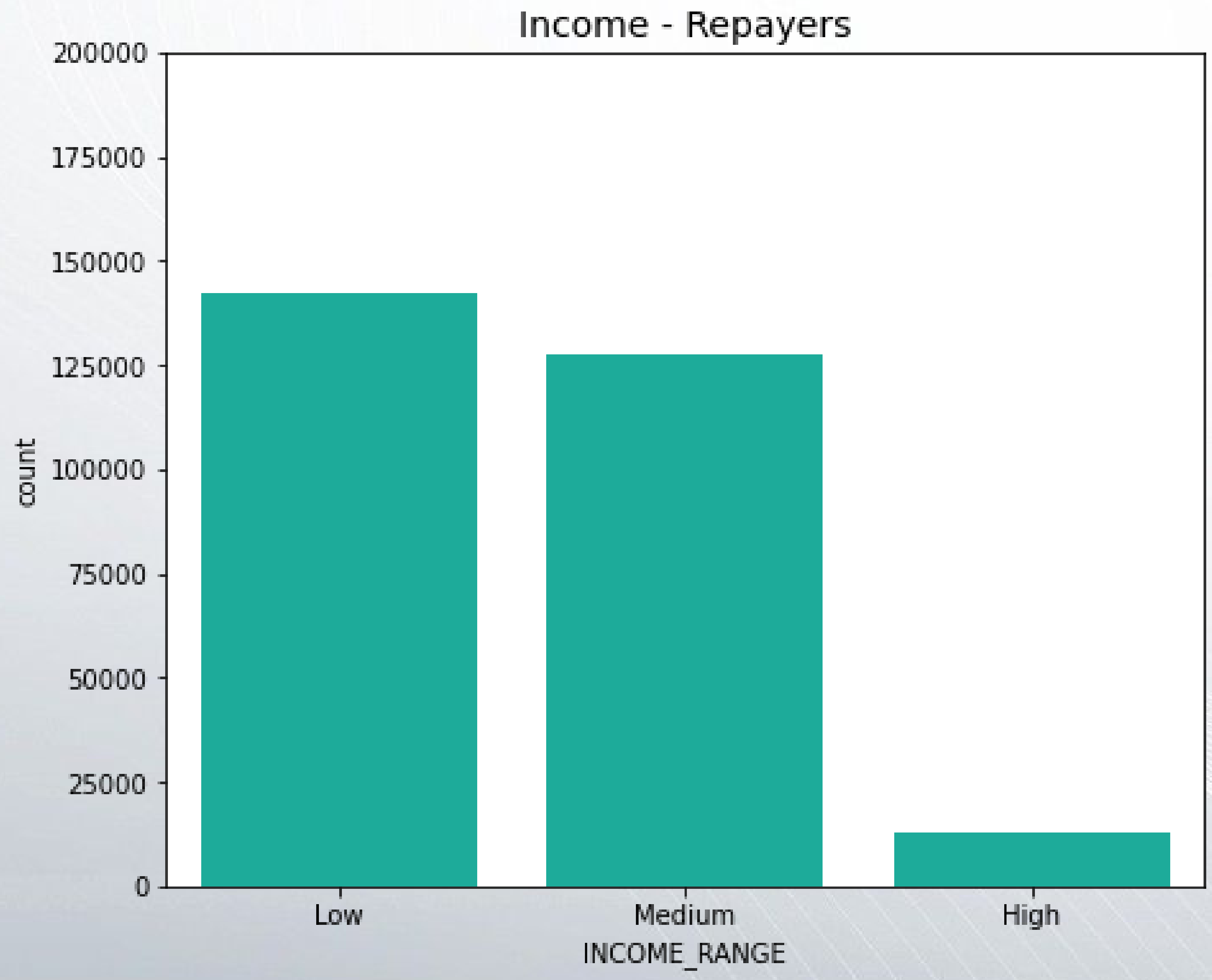


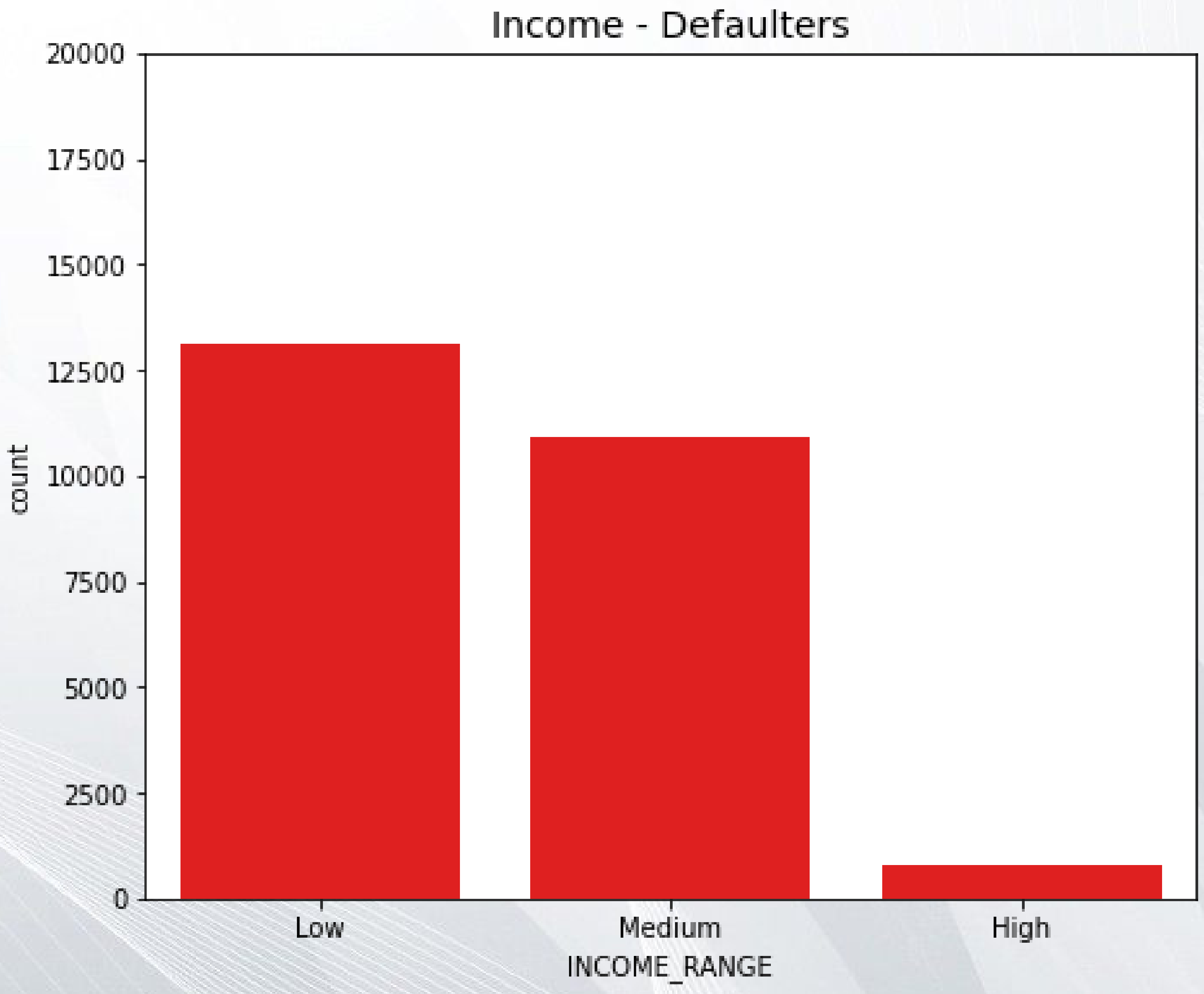


We can see here Married people have the highest percentage in both repayers and defaulters. But Single or not married person have the more default percentage than repaying percentage. Bank should be careful while giving loans to this category.

Segmented Univariate Analysis

- **AMT_INCOME(INCOME_RANGE)**

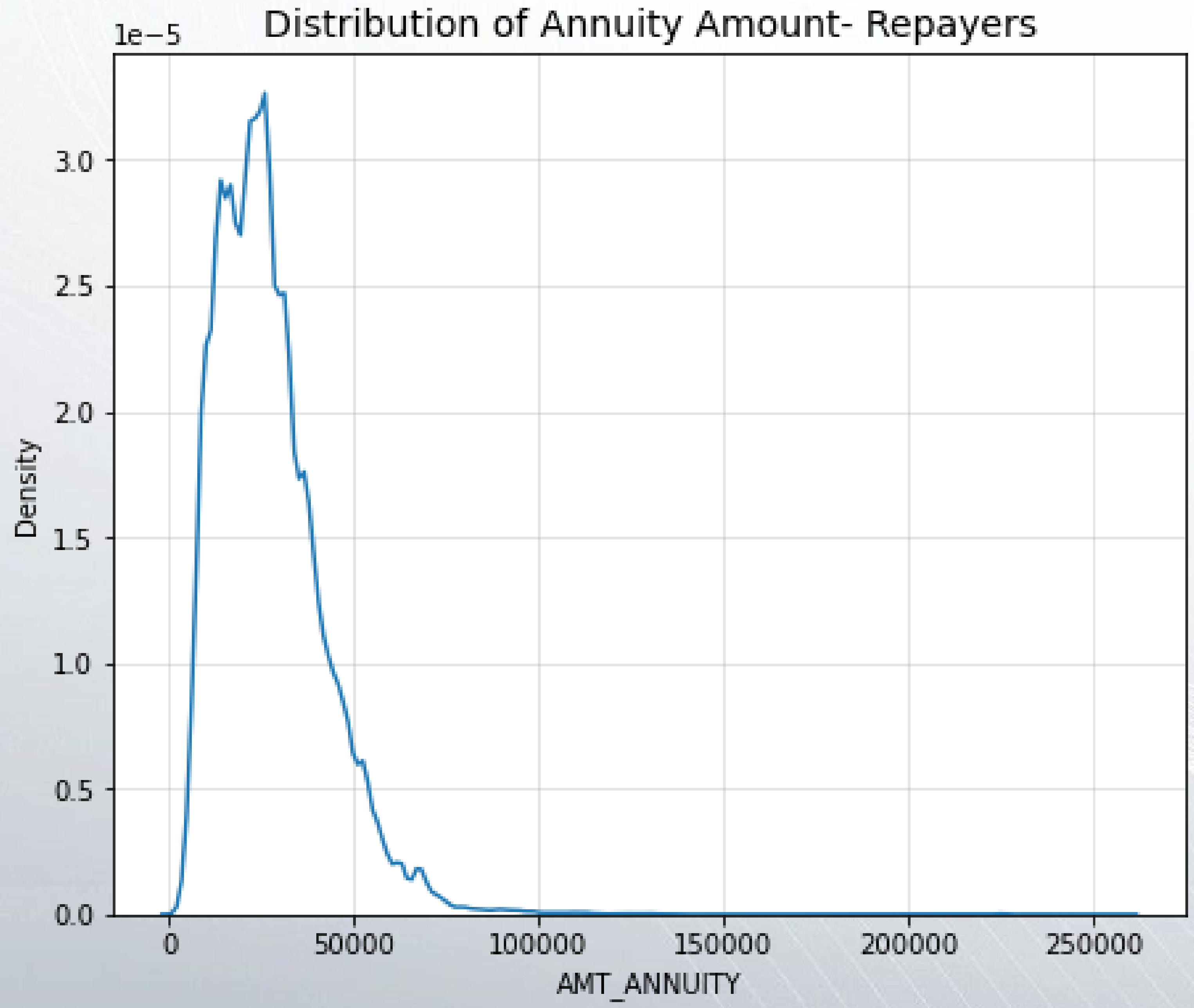


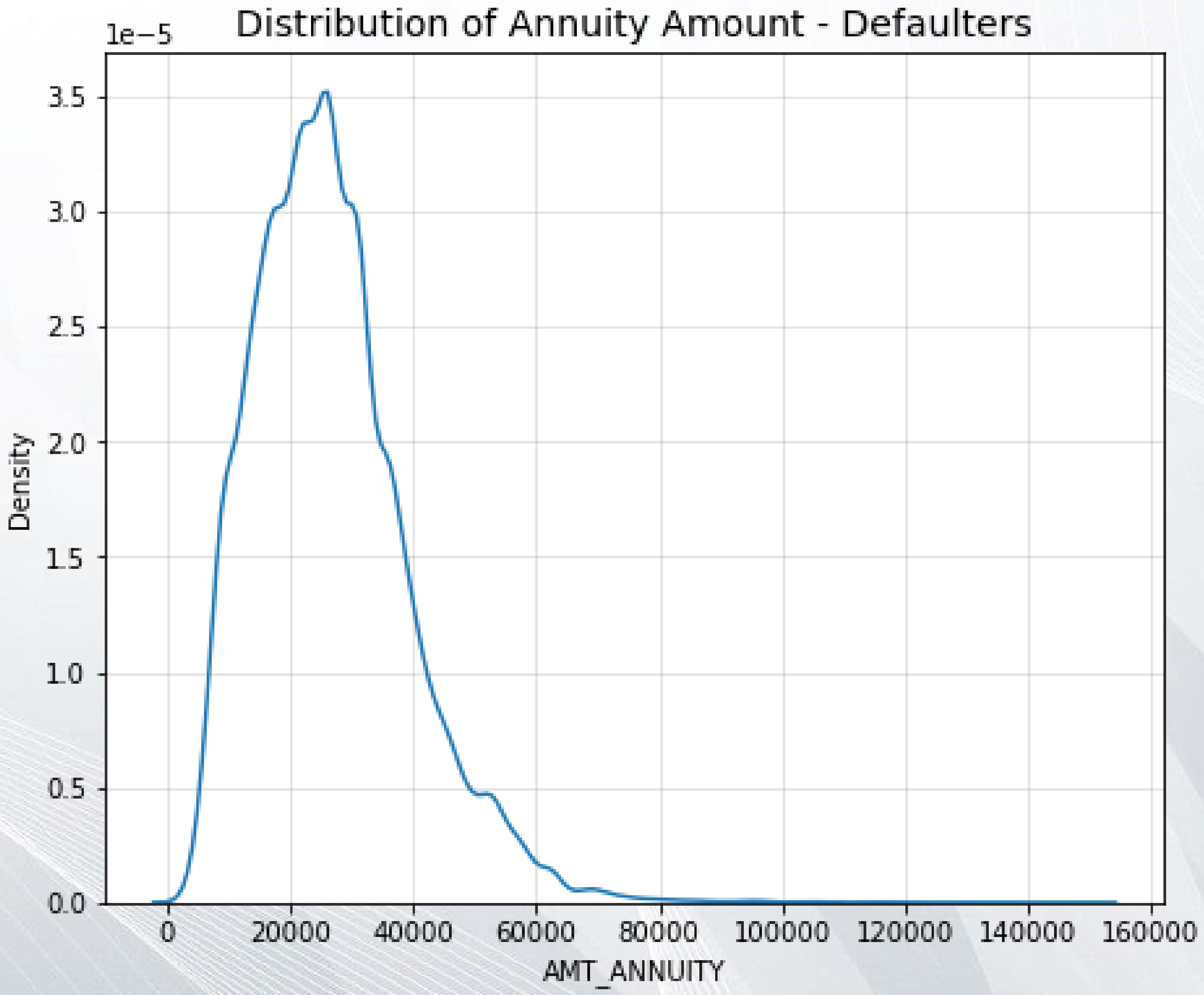


Majority of the people have income between 1 lacs to 3 lacs. People with low income has high chances of defaulting the loan.

Segmented Univariate Analysis

- **AMT_ANNUITY**

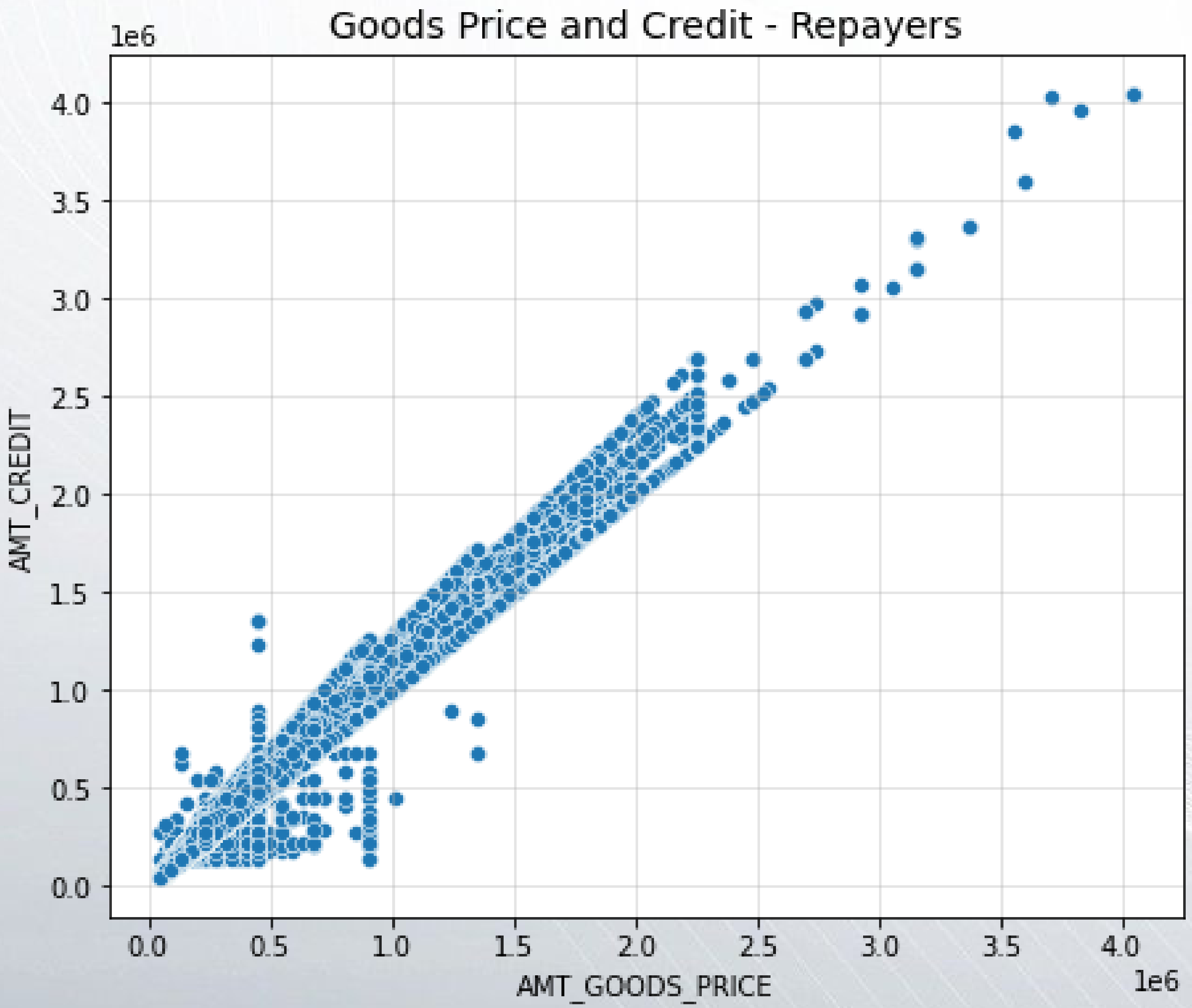


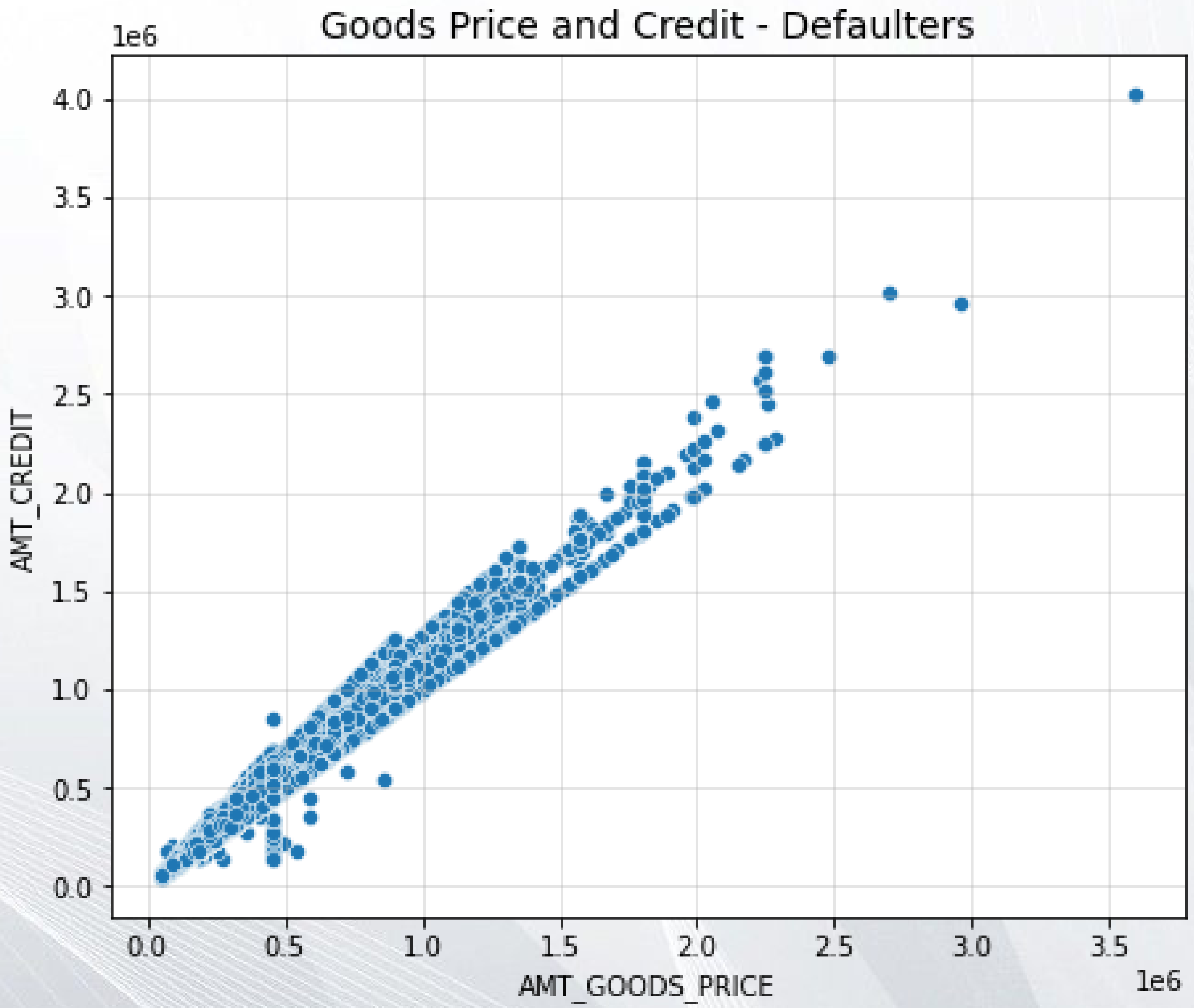


Mostly repayers paying annuity amount upto Rs 50000. Mostly defaulters paying annuity amount approx Rs 30000. There are more defaulters in low credit category. Bank needs to be more careful while giving loans in credit range.

Bivariate Analysis

- **Goods Price vs Credit Amount**

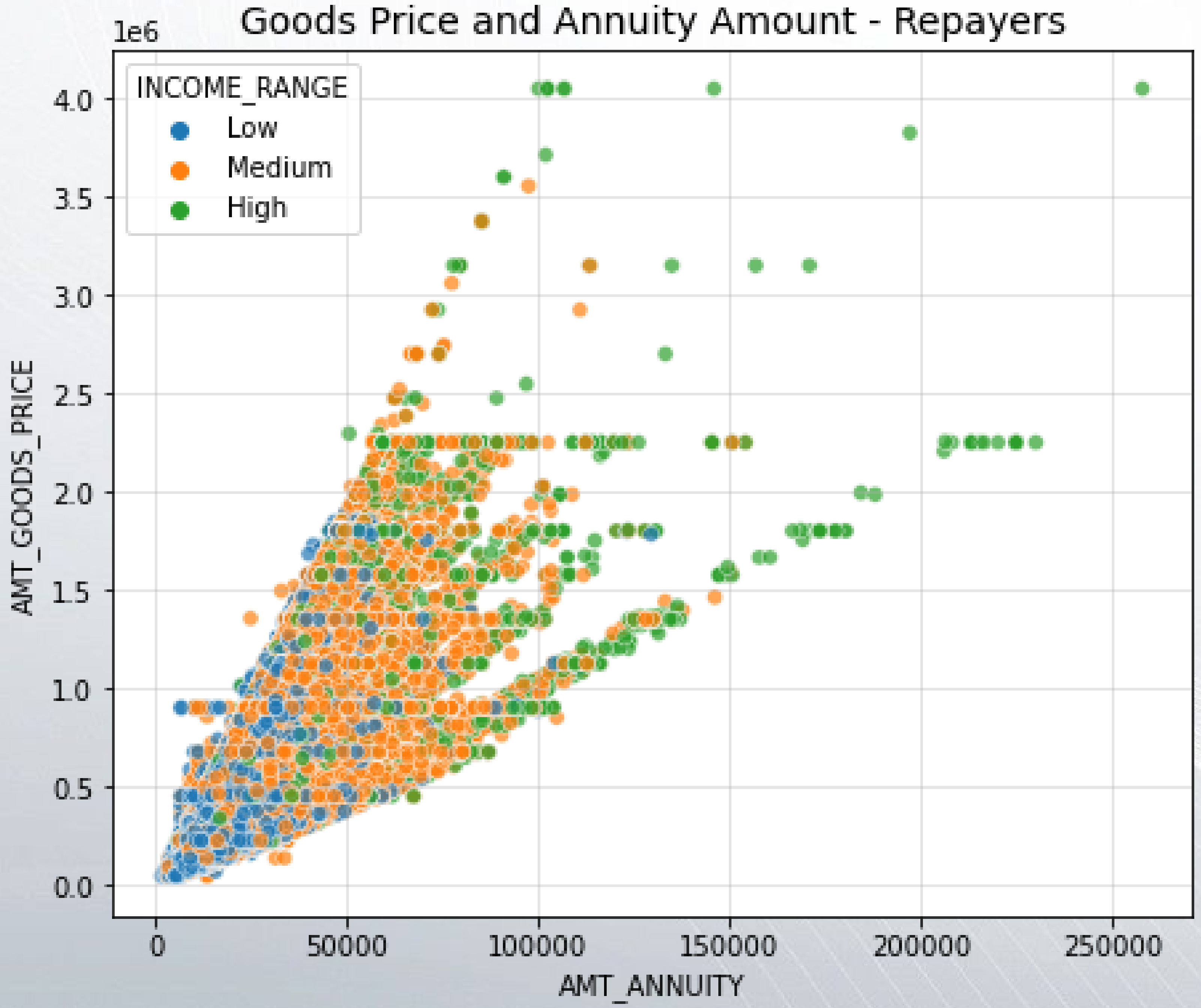


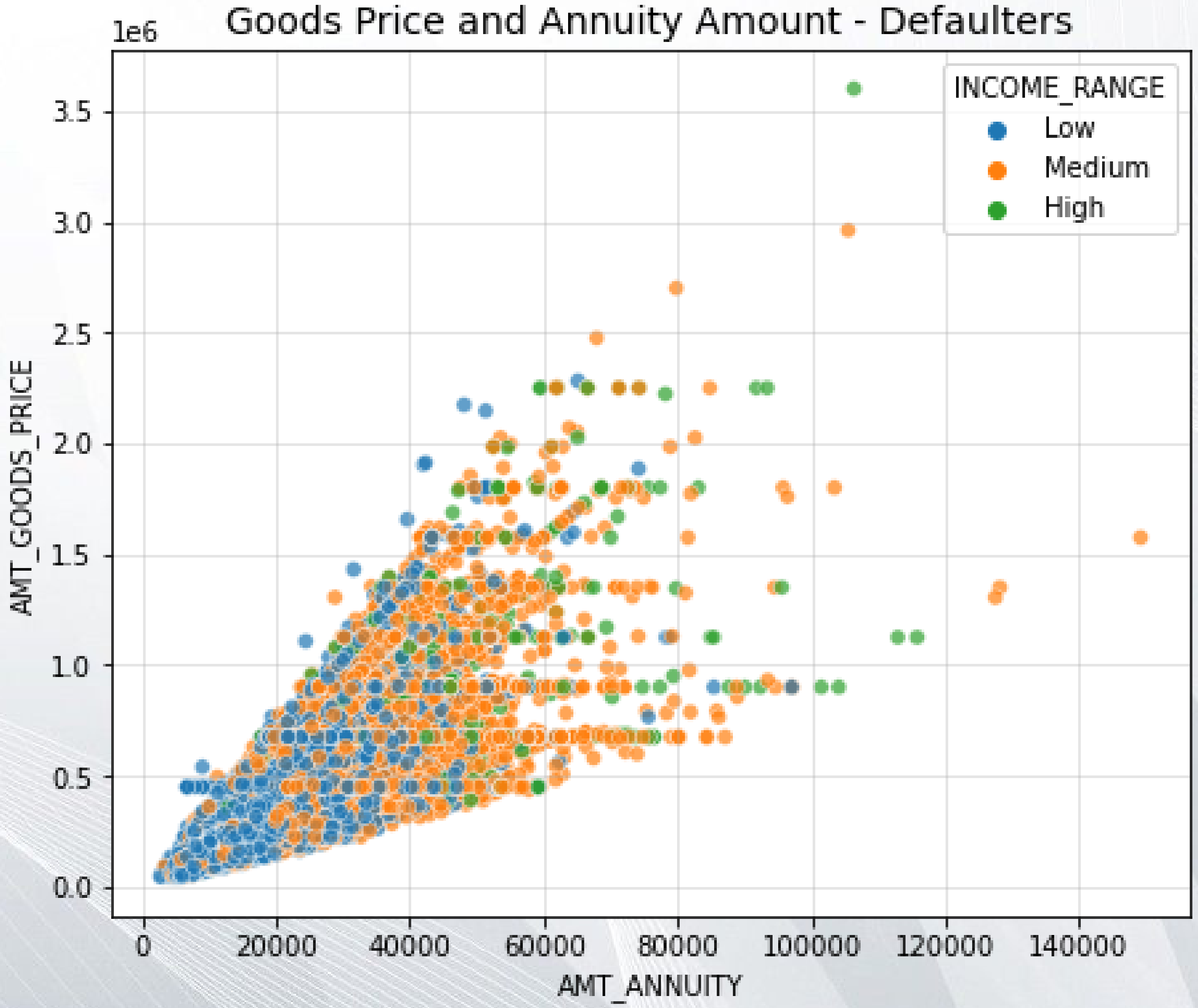


Repayers have high chances of getting high credit for expensive goods.

Bivariate Analysis

- **Annuity amount, Goods Price and Income Range**

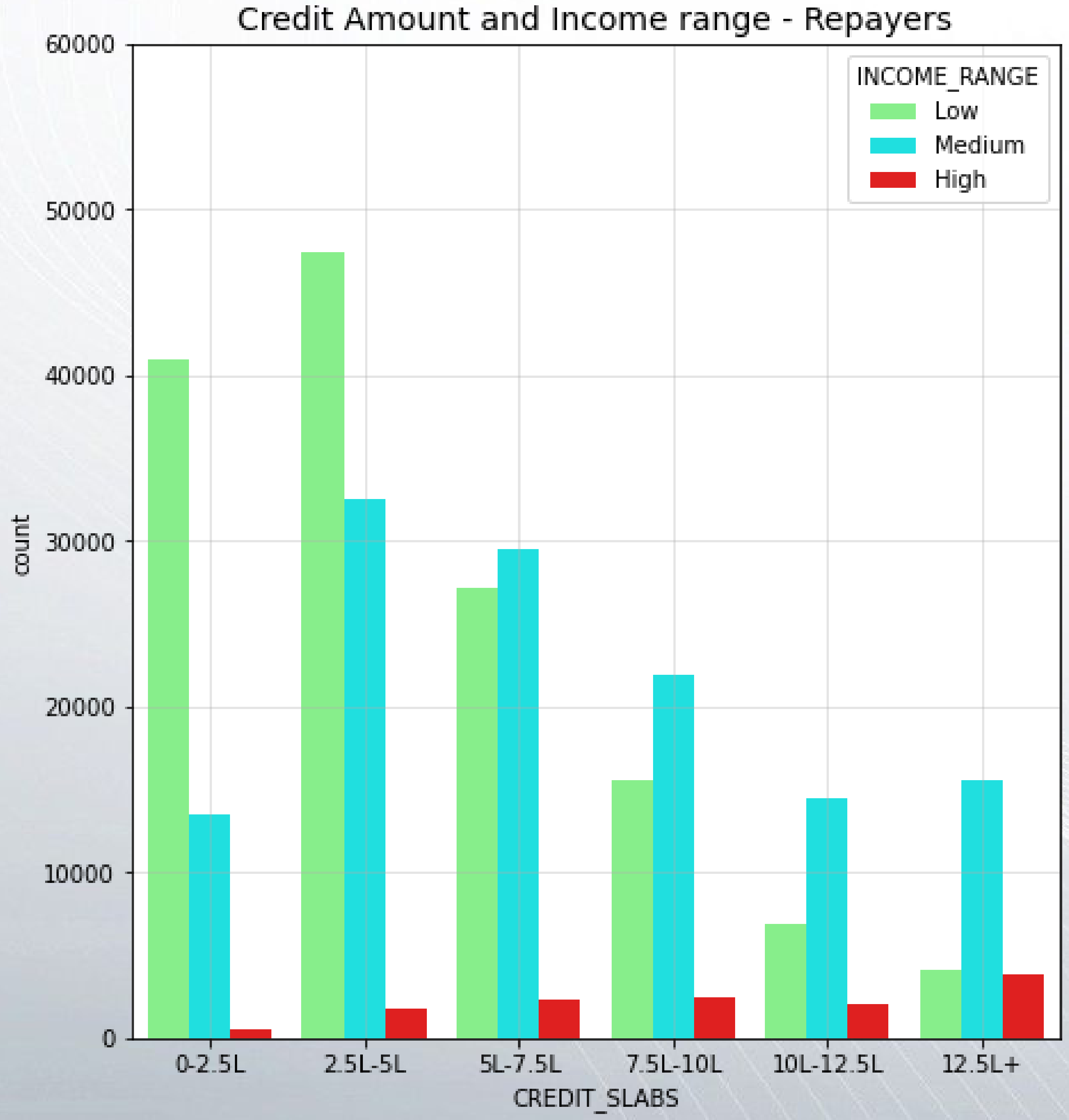

Goods Price and Annuity Amount - Repayers

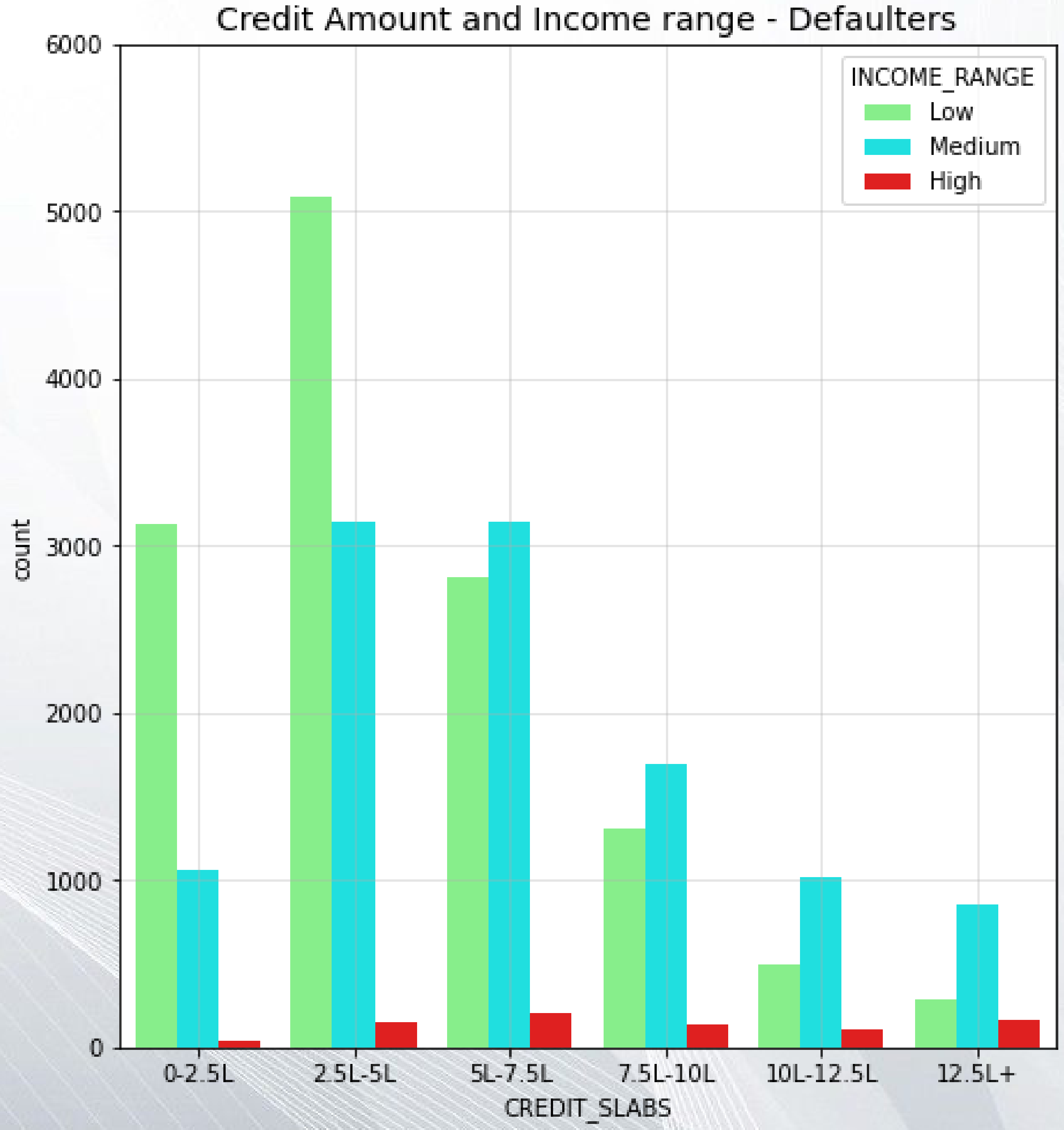

Goods Price and Annuity Amount - Defaulters

Goods price and annuity amount is more in all income category for repayers and for defaulters annuity amount and goods price is low.

Bivariate Analysis

- **Credit Amount vs Income Range**

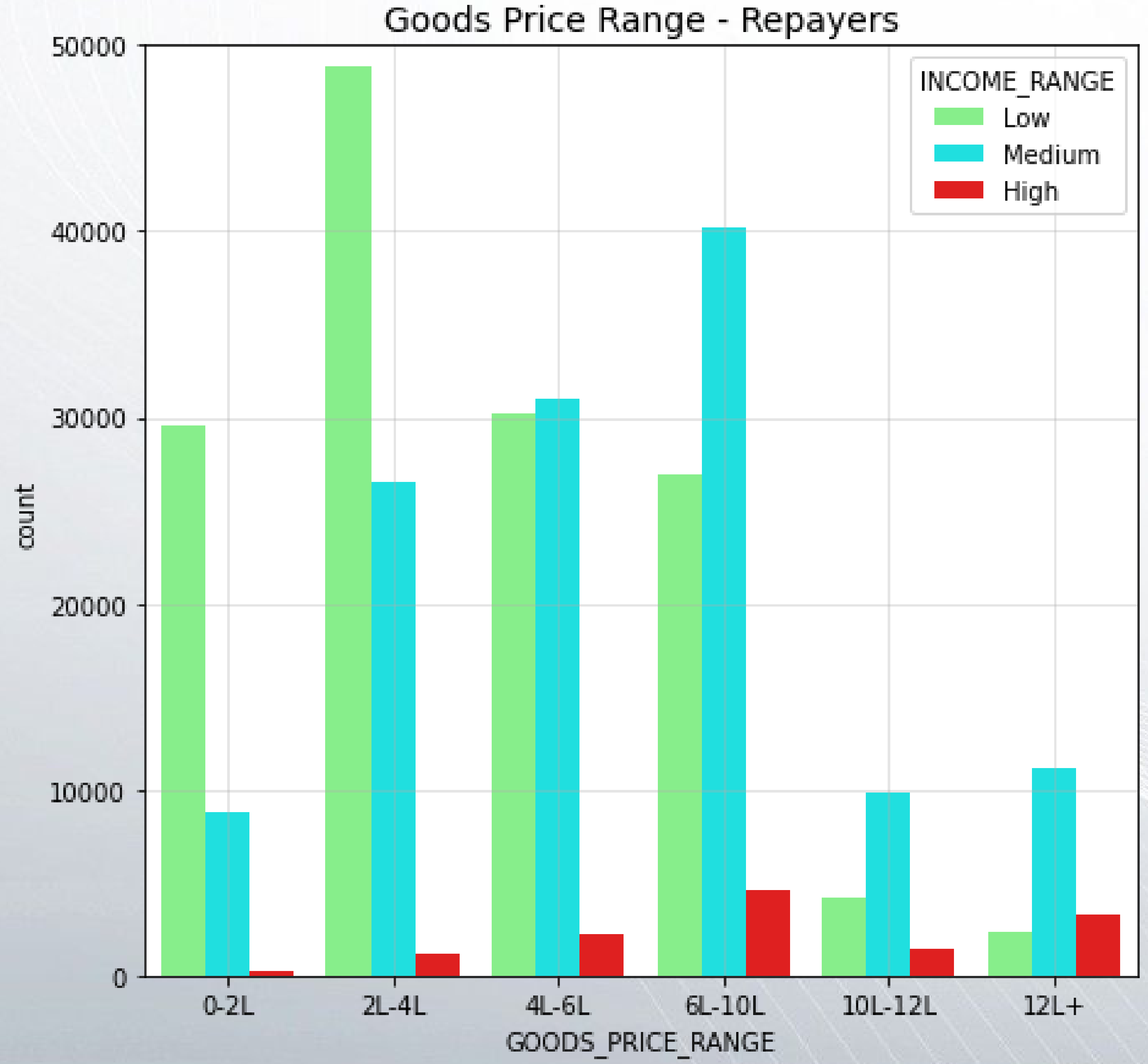




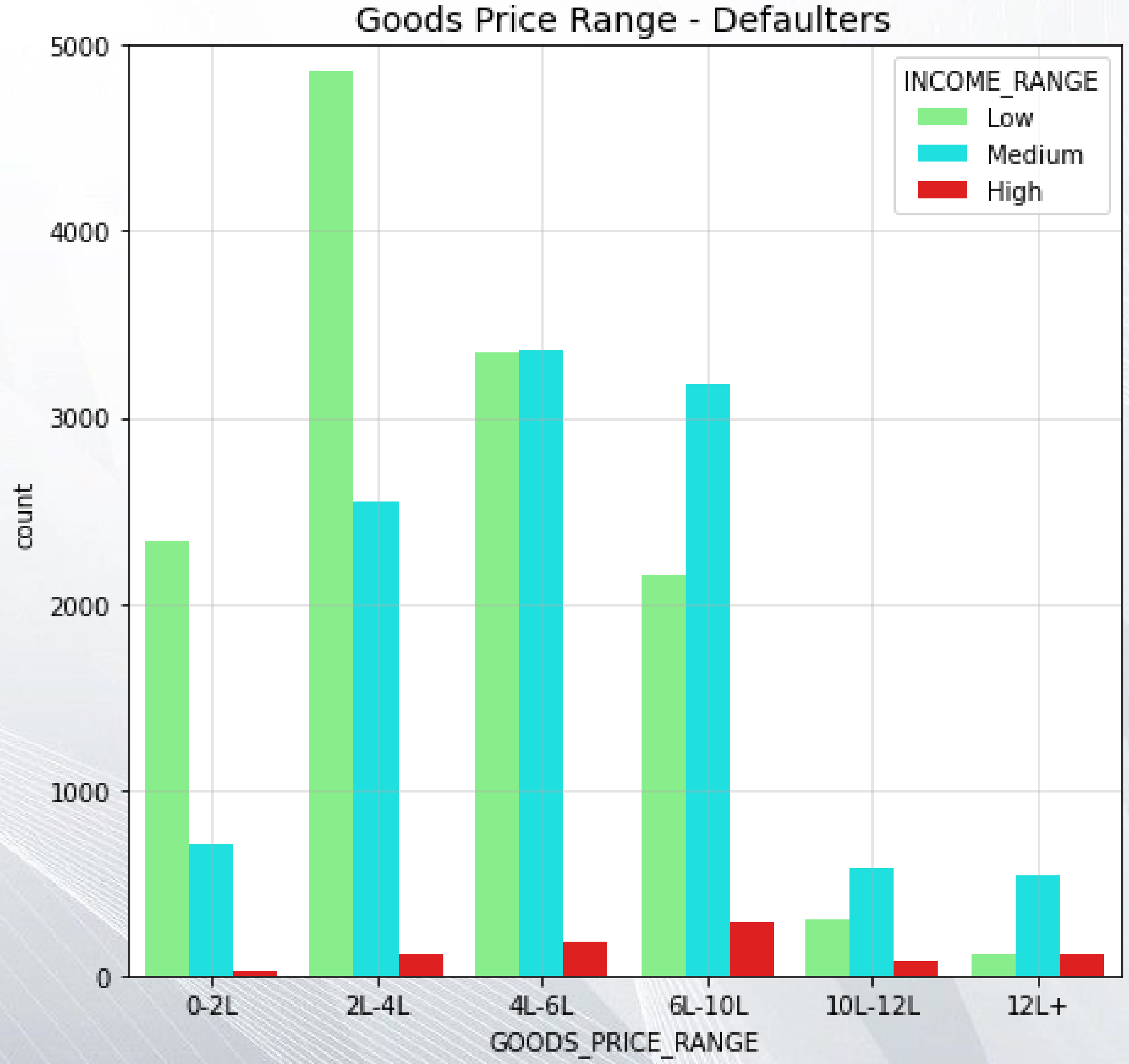
Low-income people are more likely to default on loans. While in higher credit maount people with high and medium income are paying loan on time.

Bivariate Analysis

- **Goods Price and Income Range**



Goods Price Range - Repayers



Goods Price Range - Defaulters

People with high and medium income buying expensive goods paying loan on time. Goods price 4 lacs to 6 lacs have high % of defaulters in all income category.

Correlation Matrix

- **Top 10 Correlation of Defaulters**

Variable-1	Variable-2	Correlation
DAYS_EMPLOYED	FLAG_EMP_PHONE	1
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1
AMT_GOODS_PRICE	AMT_CREDIT	0.98
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.96
CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
AMT_GOODS_PRICE	AMT_ANNUITY	0.75
AMT_ANNUITY	AMT_CREDIT	0.75
DAYS_EMPLOYED	DAYS_BIRTH	0.58
DAYS_BIRTH	FLAG_EMP_PHONE	0.58

Correlation Matrix

- **Top 10 Correlation of Repayers**

Variable-1	Variable-2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1
DAYS_EMPLOYED	FLAG_EMP_PHONE	1
AMT_GOODS_PRICE	AMT_CREDIT	0.99
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.95
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
AMT_GOODS_PRICE	AMT_ANNUITY	0.78
AMT_ANNUITY	AMT_CREDIT	0.77
DAYS_EMPLOYED	DAYS_BIRTH	0.63
DAYS_BIRTH	FLAG_EMP_PHONE	0.62

Conclusion

- People are taking cash loan for buying property, buying or repairing car, education and for medical purpose.
- People are taking consumer loans more for buying electronics products.
- People with higher education and good income should be classified as priority customer. These people are paying loan on time.
- People with more professional experience are paying loan on time. People with work experience 0-5 years have high chances of making payment defaults.
- People with lower education and low income are under high risk category.
- Bank should be careful while giving loan to single/unmarried people.

THANK YOU