CREDIT EDA ASSIGNMENT

Presented by:
RAHUL PRAKASH CHIKATE

TABLE OF CONTENTS

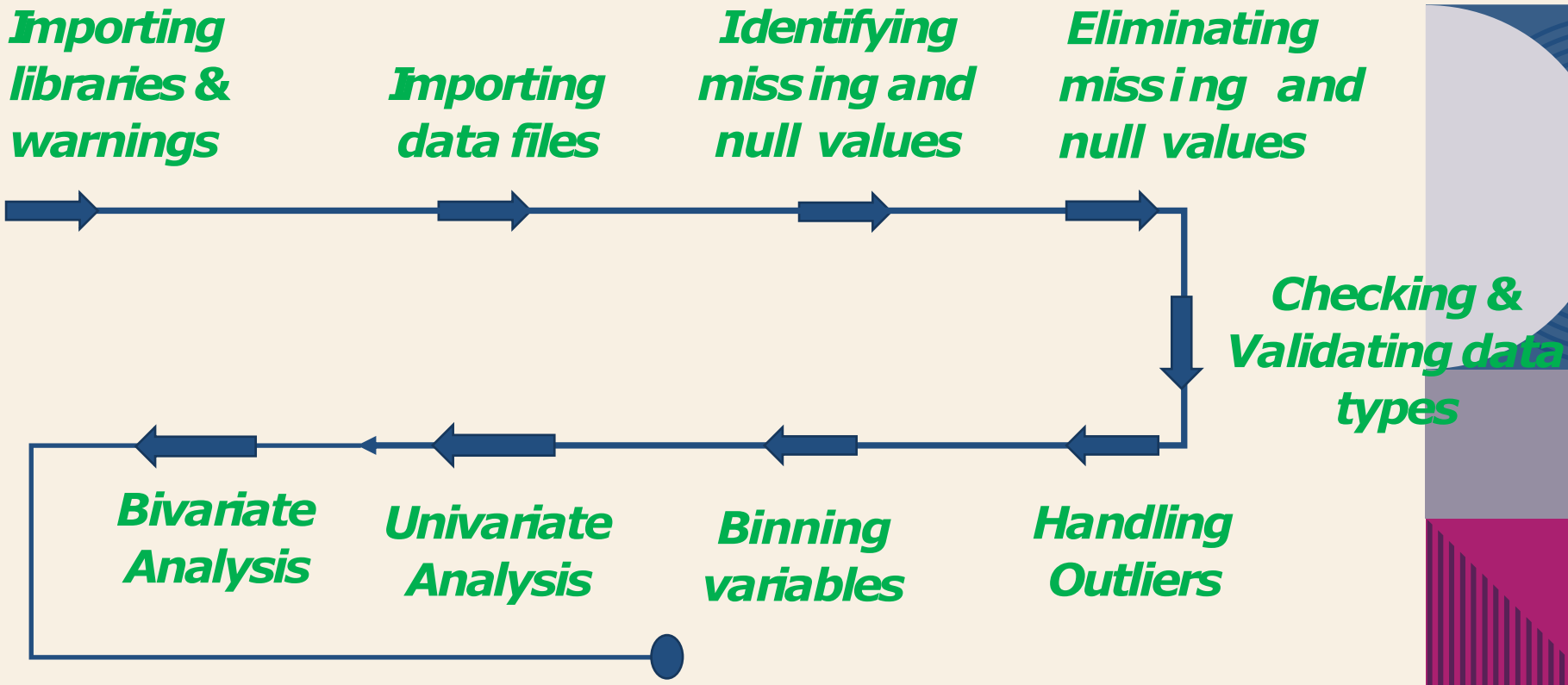
PROBLEM STATEMENT

AIM

The agenda is to identify patterns which indicate if a client has any difficulties in payment which will further help the bank to decide if

- The loan should be approved
- Plan new lending schemes
- Denying the loan

WORK FLOW


Importing libraries & warnings
Importing data files
Identifying missing and null values
Eliminating missing and null values
Checking & Validating data types
Handling Outliers
Binning variables
Univariate Analysis
Bivariate Analysis
Conclusion

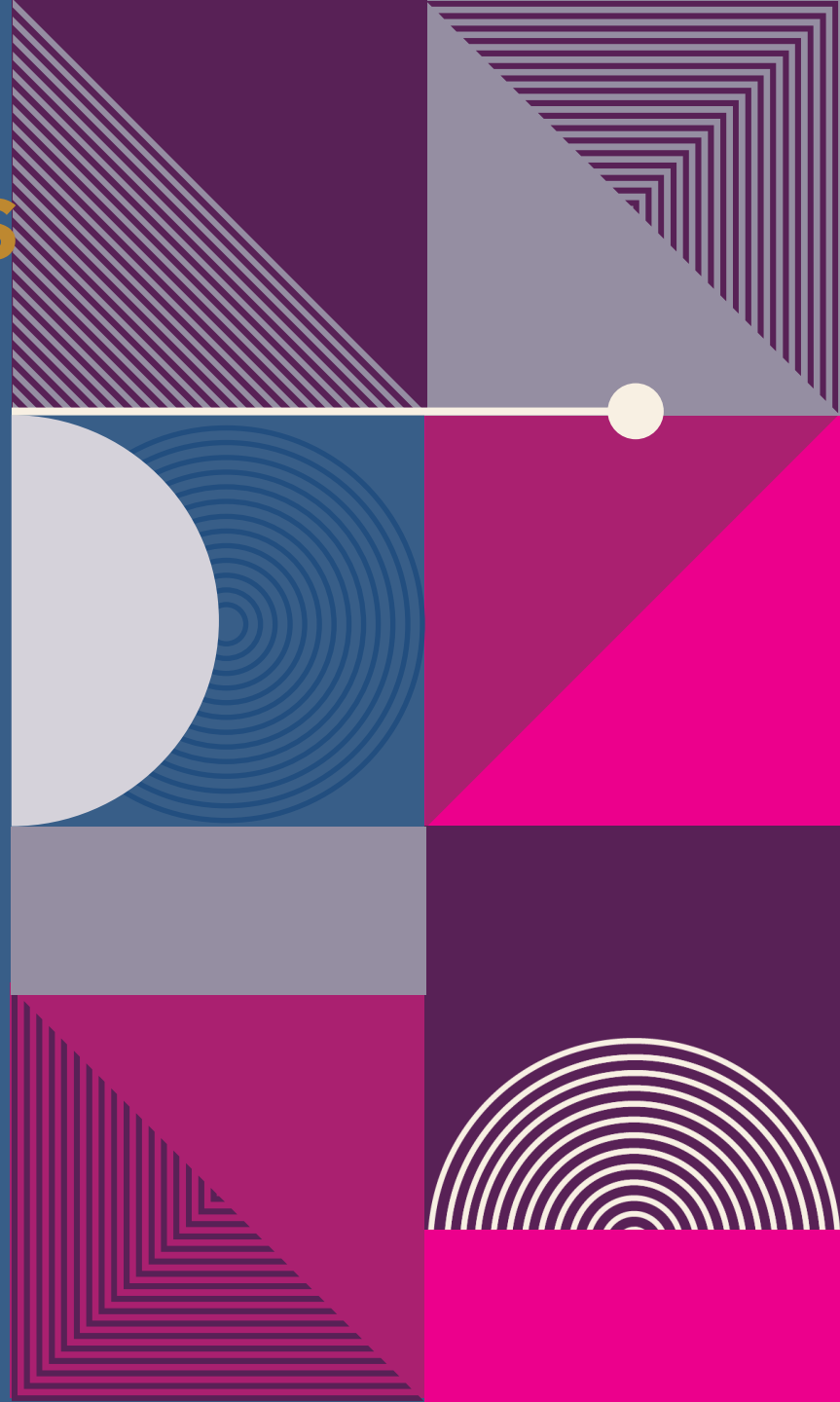
IMPORTING LIBRARIES AND WARNINGS

Importing libraries

Imported pandas, numpy, matplotlib & seaborn for data loading & visualization

Importing Warnings

Highlights warnings however the program runs.

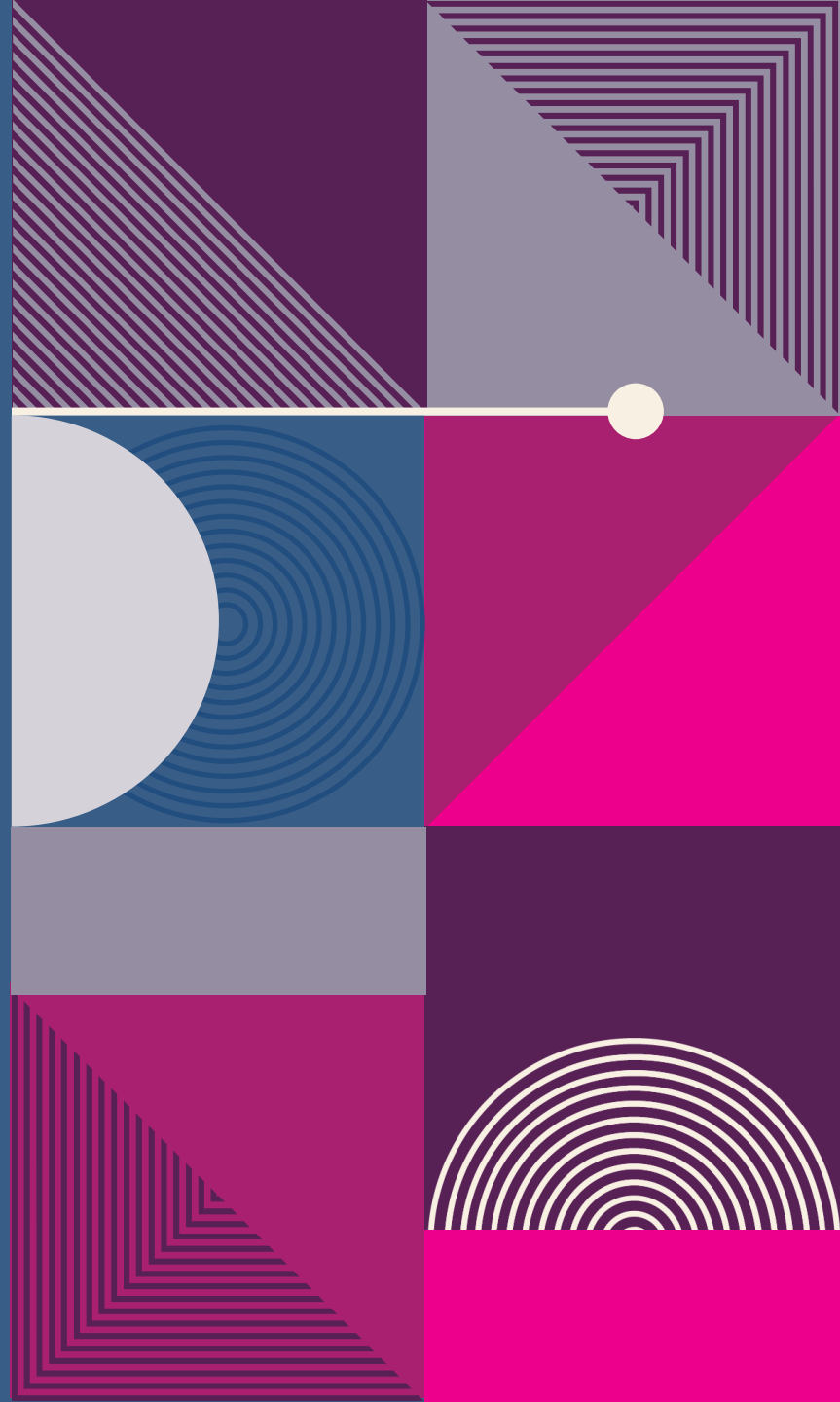


READING DATASET

The flag variable is our target variable which highlights if the client will pay instalment on time or not

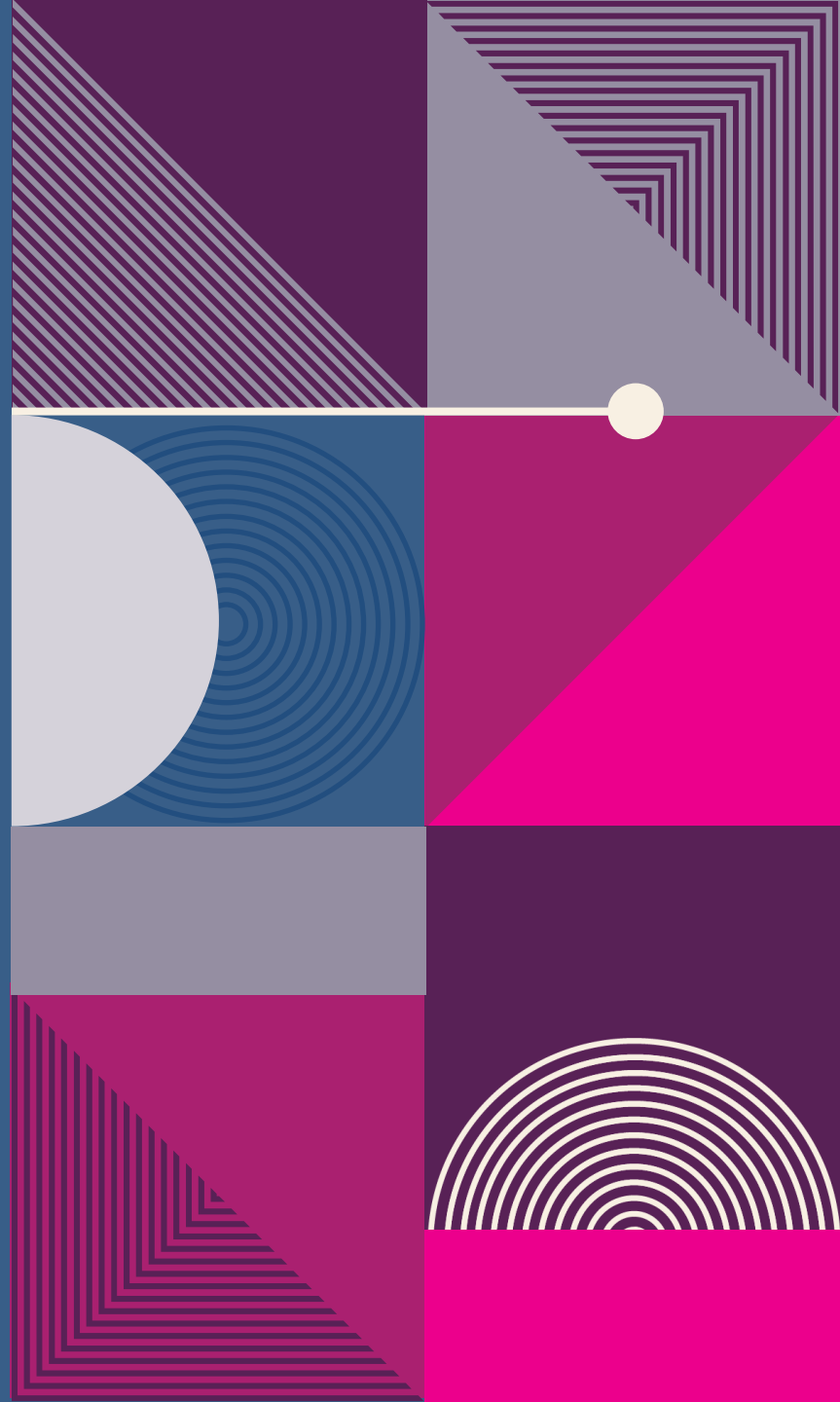
Two data files were extracted from the given dataset. namely - 'application_data.csv’ and 'previous_data.csv'

Highlighted datafile description, shape etc., in the notebook for elaborated experience in reading the data.



HANDLING DATA, NULL & MISSING VALUES

- Checked for unique Value in application_data.csv and replaced XNA values with np.NAN. Also checked for null values in application_data.csv and eliminated 49 columns which had null values more than 40%
- Post that, AMT_ANNUITY , AMT_GOODS_PRICE, NAME_TYPE_SUITE , EXT_SOURCE_2, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, DAYS_LAST_PHONE_CHANGE had less than 1% of null (& numeric) values. Hence drop the rows which had null values.
- After that replaced null values with mode value for categorical numeric column .
- Changed Negative values in dataset to positive one using ‘abs’ Function.
- Changed the data type of CNT_FAM_MEMBERS from float to int

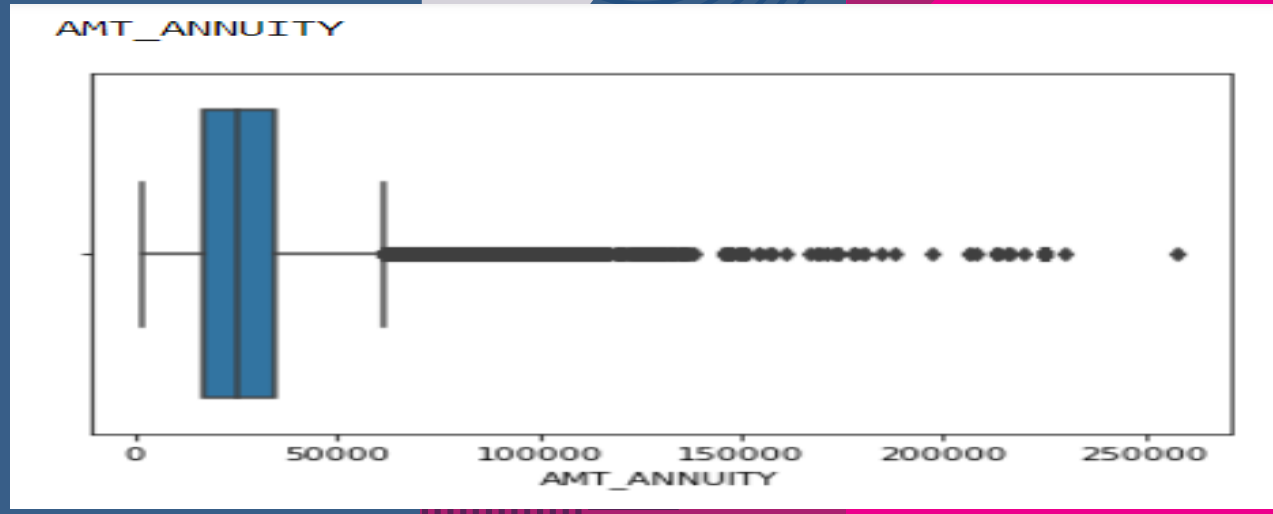


OUTLIERS HANDLING

- Checked for outliers theoretically and by using boxplot

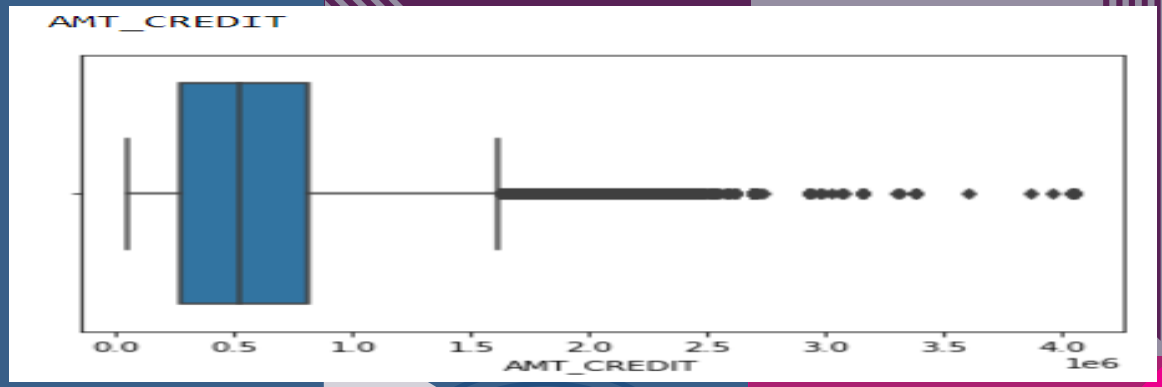
  - AMT_ANNUITY variable

As seen here, outlier is present in this columns and to remove the outlier we will use IQR method.



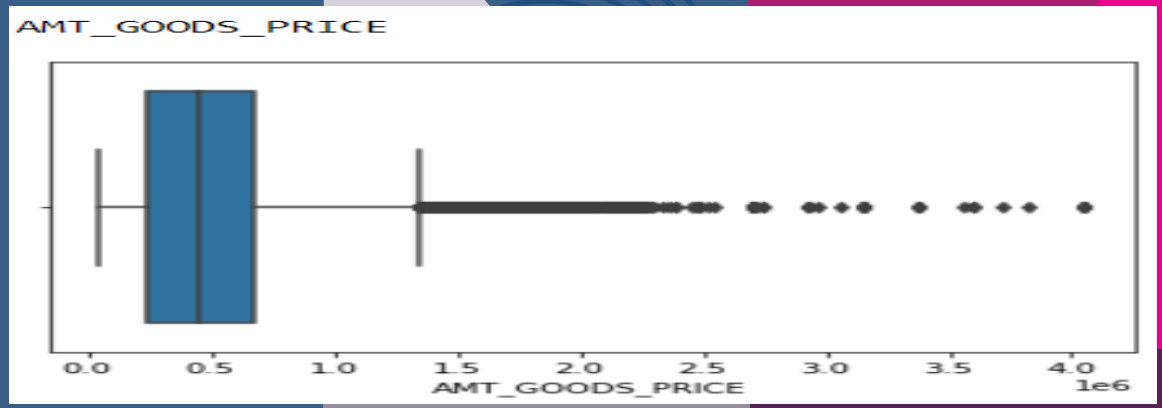
- AMT_CREDIT variable

As seen here, outlier is present in this columns and to remove the outlier we will use IQR method.



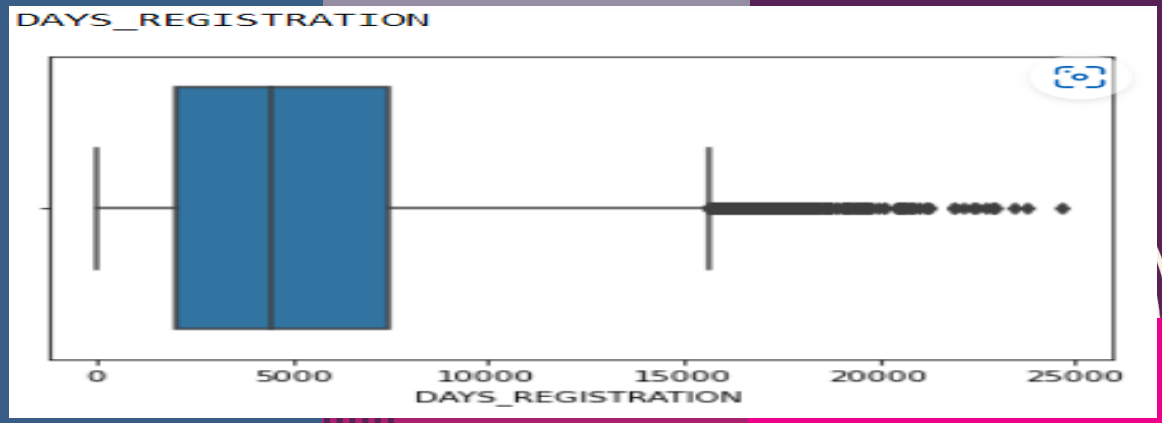
- AMT_GOODS_PRICE variable

As seen here, outlier is present in this columns and to remove the outlier we will use IQR method.



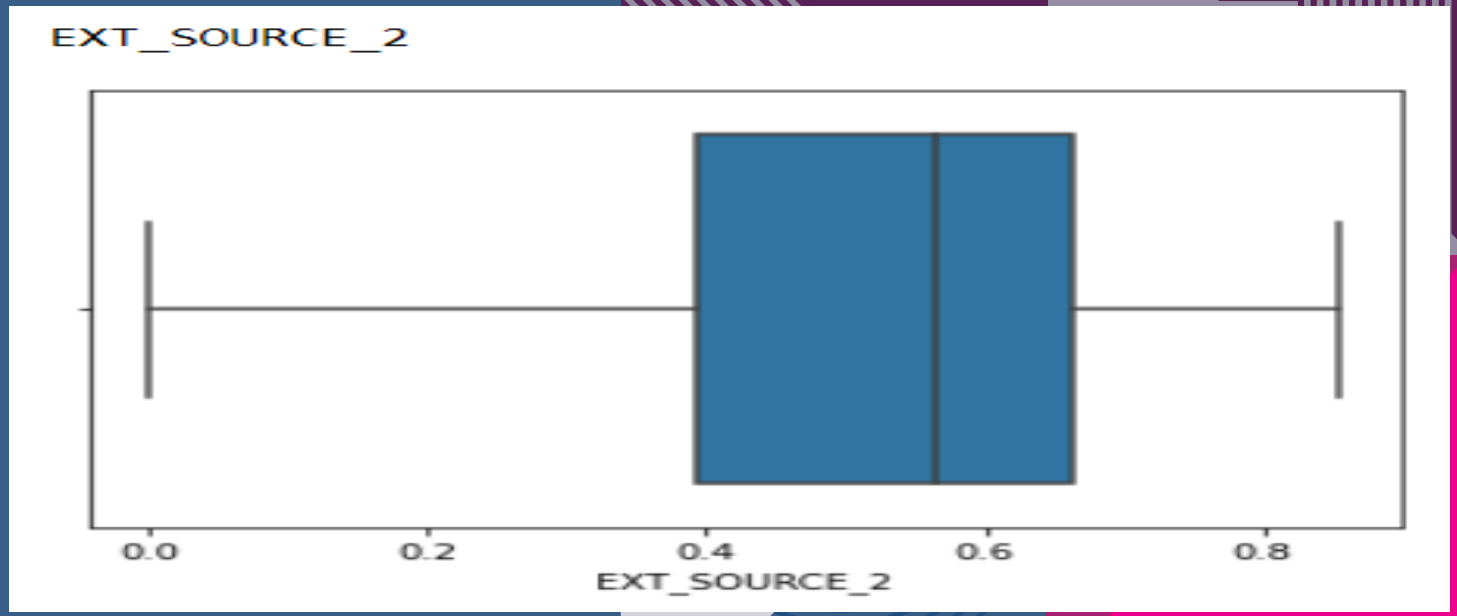
- DAYS_REGESTRATION variable

As seen here, outlier is present in this columns and to remove the outlier we will use IQR method.



- EXT_COURCE_2 variable

No outliers are present here.



ANALYSIS

When we find the imbalance percentage, We can conclude that the 'TARGET VARIABLE' has 91.90% of 0s and 08.10% of 1s. Hence, we can conclude here that 91.90% of people here make timely payments and only 8.10% face challenges.

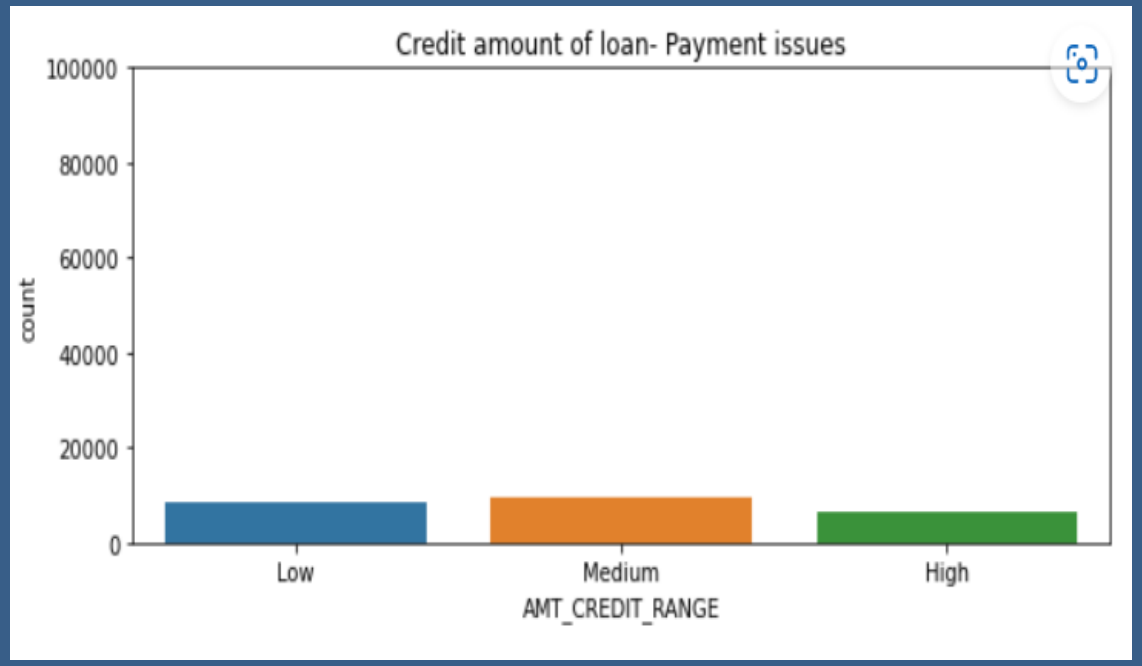
```
1 # check data inbalance in target column
2 (app_data.TARGET.value_counts())/ (len(app_data))*100
```

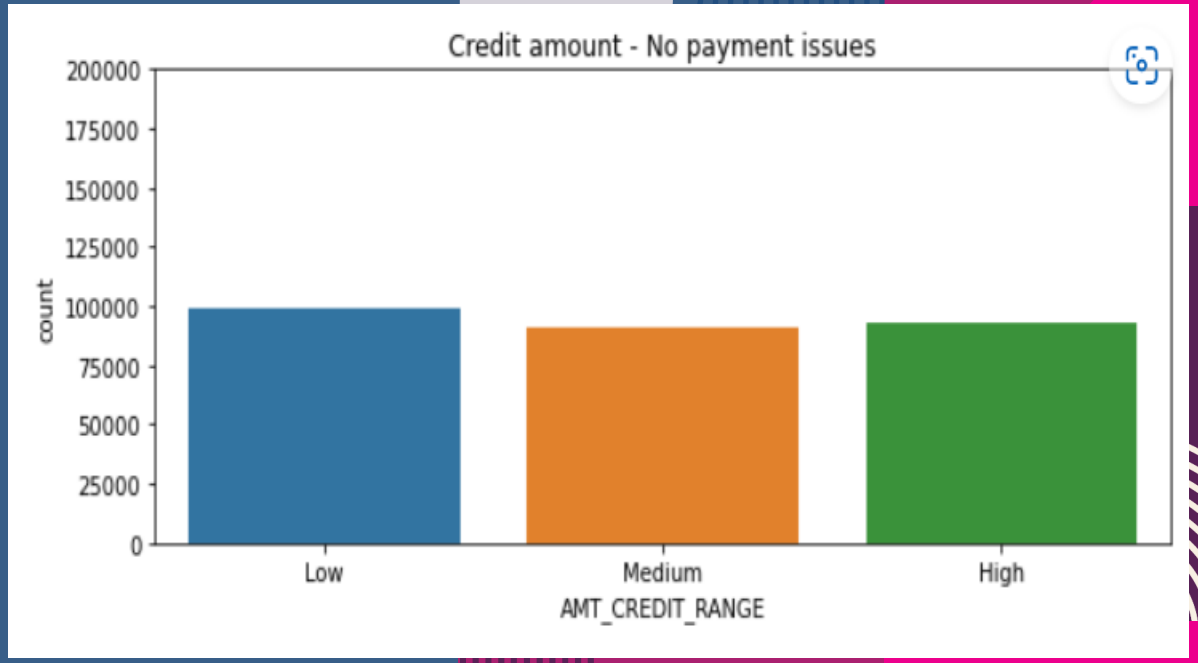
```
0    91.899897
1     8.100103
Name: TARGET, dtype: float64
```

UNIVARIATE ANALYSIS FOR NUMERIC VARIABLES

Credit Amount Analysis

When verified, it has been observed that, Customers who have low credit amount are more likely to pay back the loan.

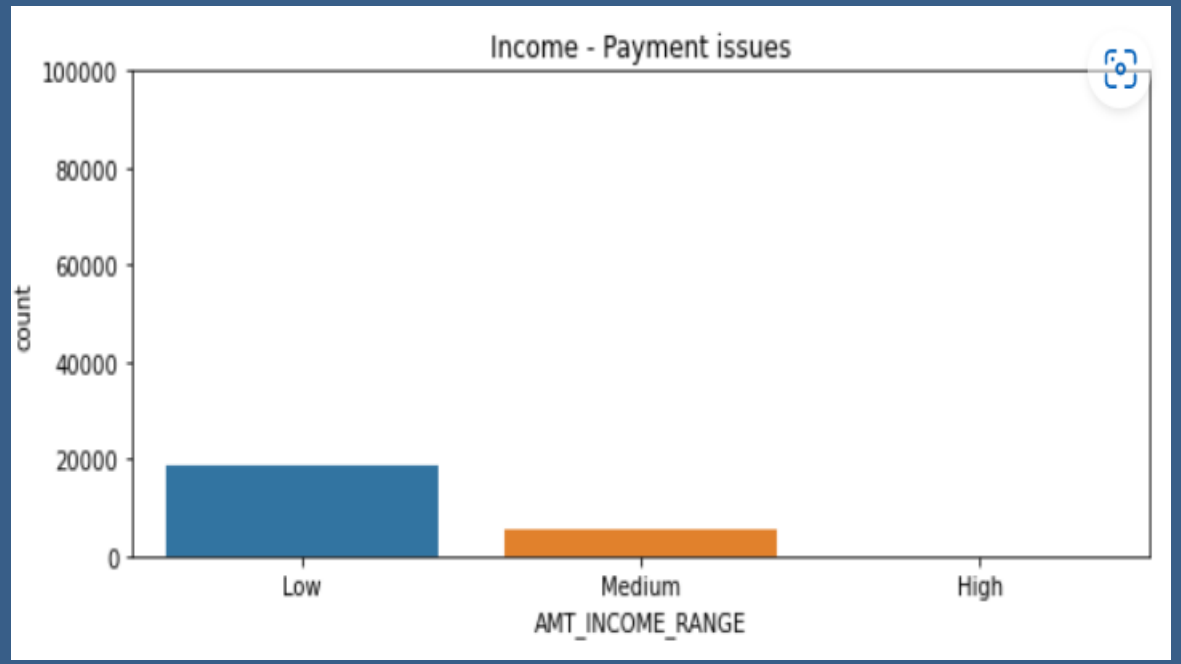


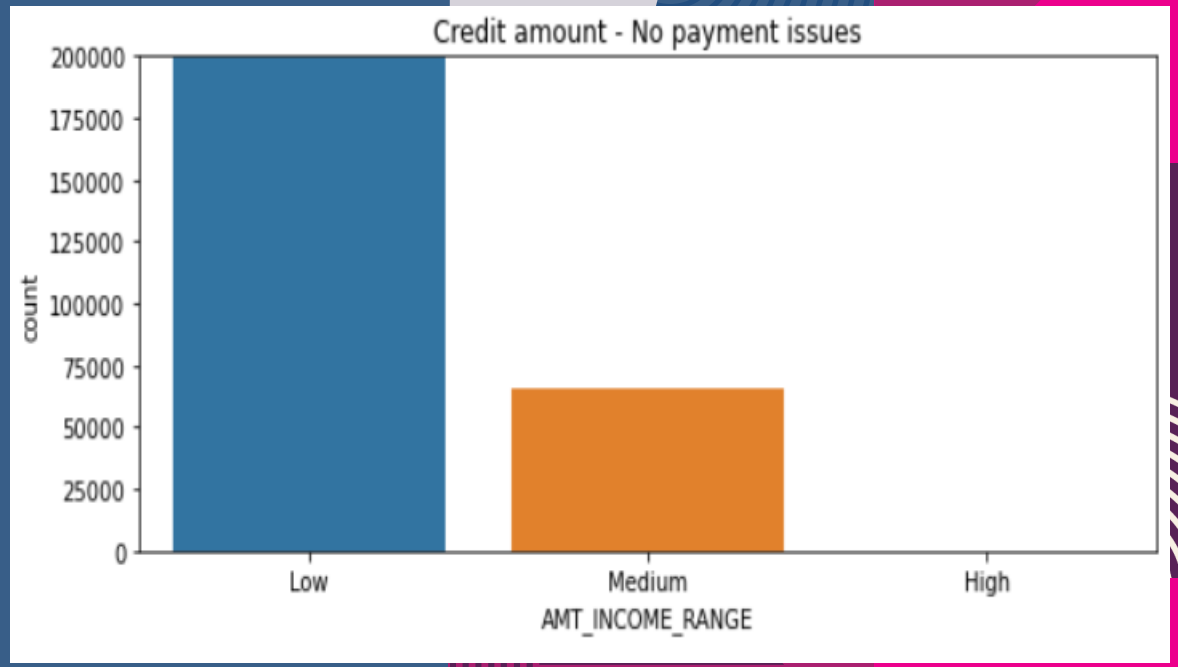


UNIVARIATE ANALYSIS FOR NUMERIC VARIABLES

Income Analysis

After completing the analysis, it has been observed that as compared to the other categories, clients having low income are more likely to repay the loan.

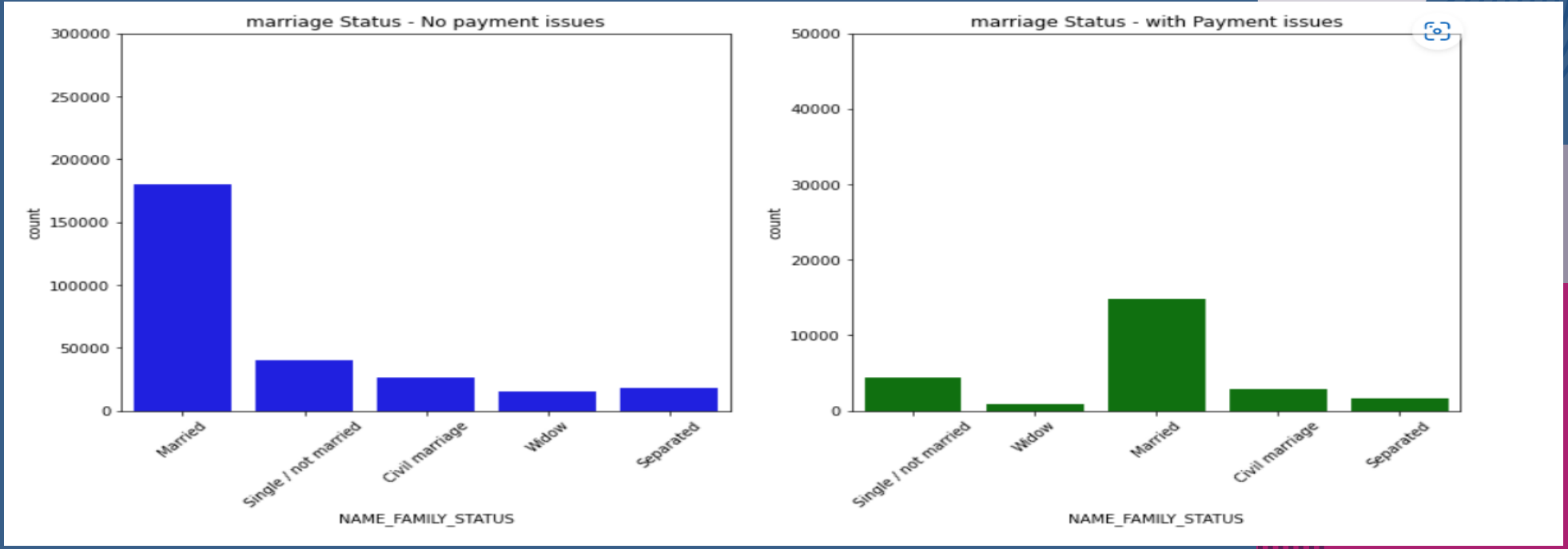




UNIVARIATE ANALYSIS FOR NUMERIC VARIABLES

Family status type analysis

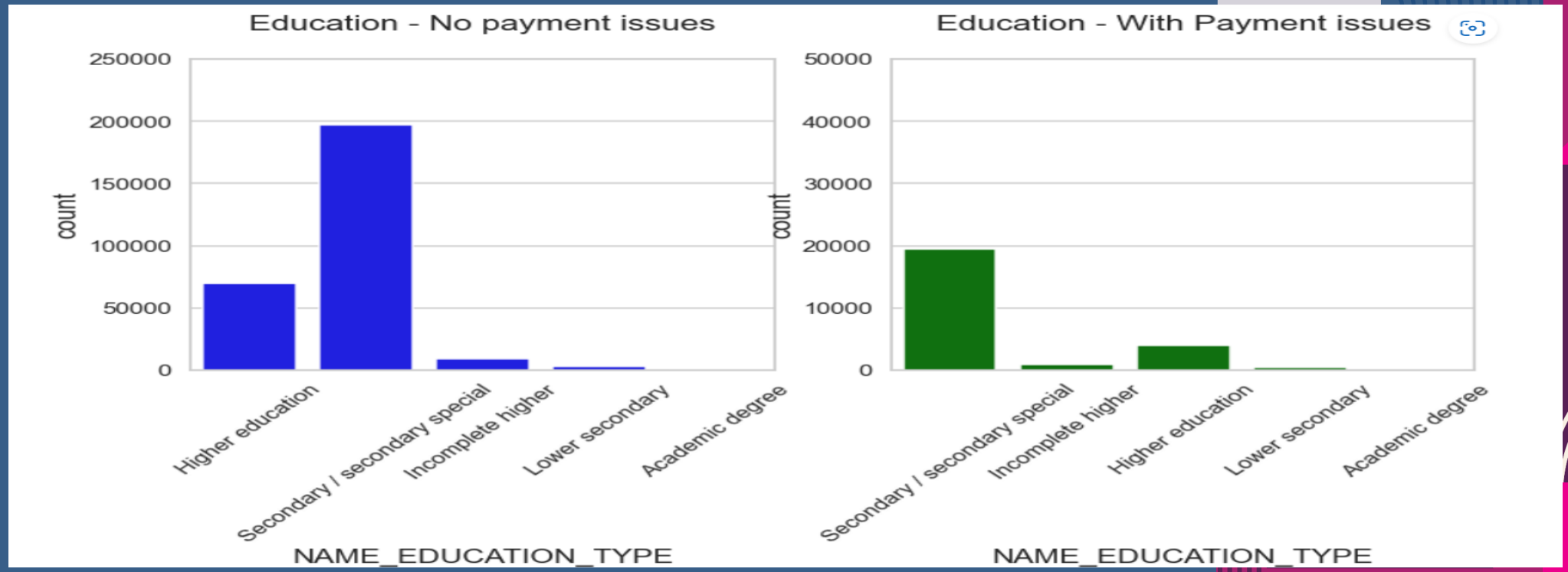
As seen here, Widows are the least likely category to repay whereas as seen here, Married category is most likely to repay the loan amount.



UNIVARIATE ANALYSIS FOR NUMERIC VARIABLES

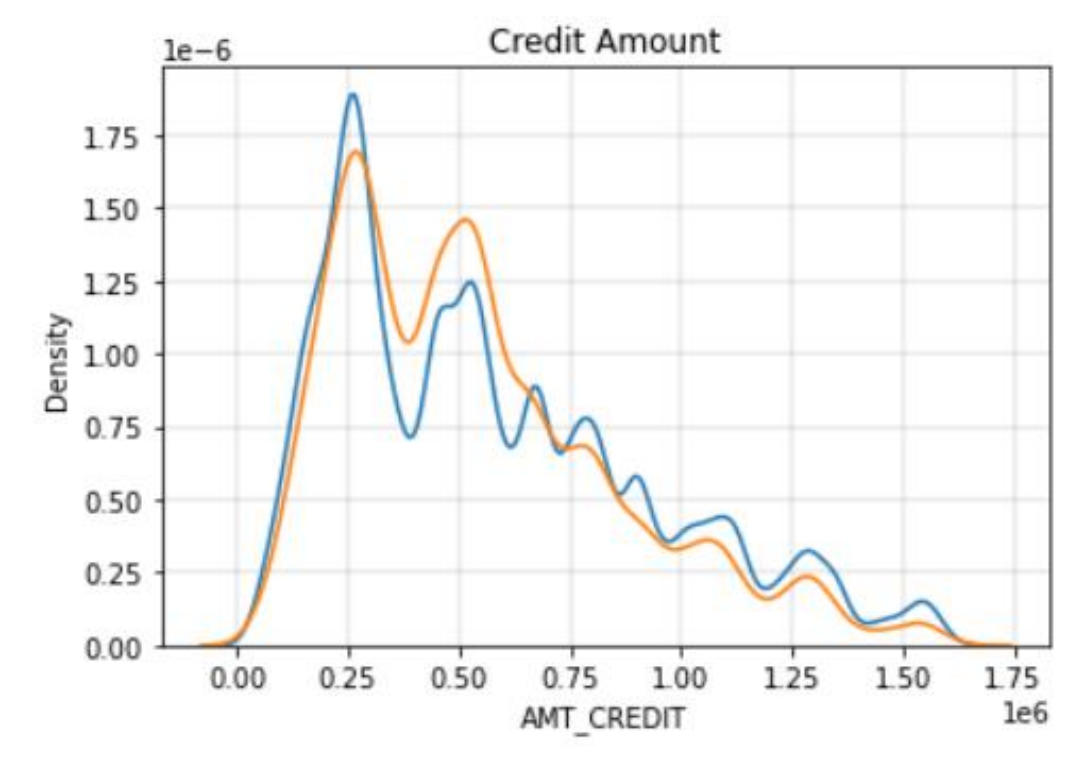
Education type analysis

People most likely to repay the loan have secondary/ Secondary special education status whereas clients having an academic degree are the most defaulters.

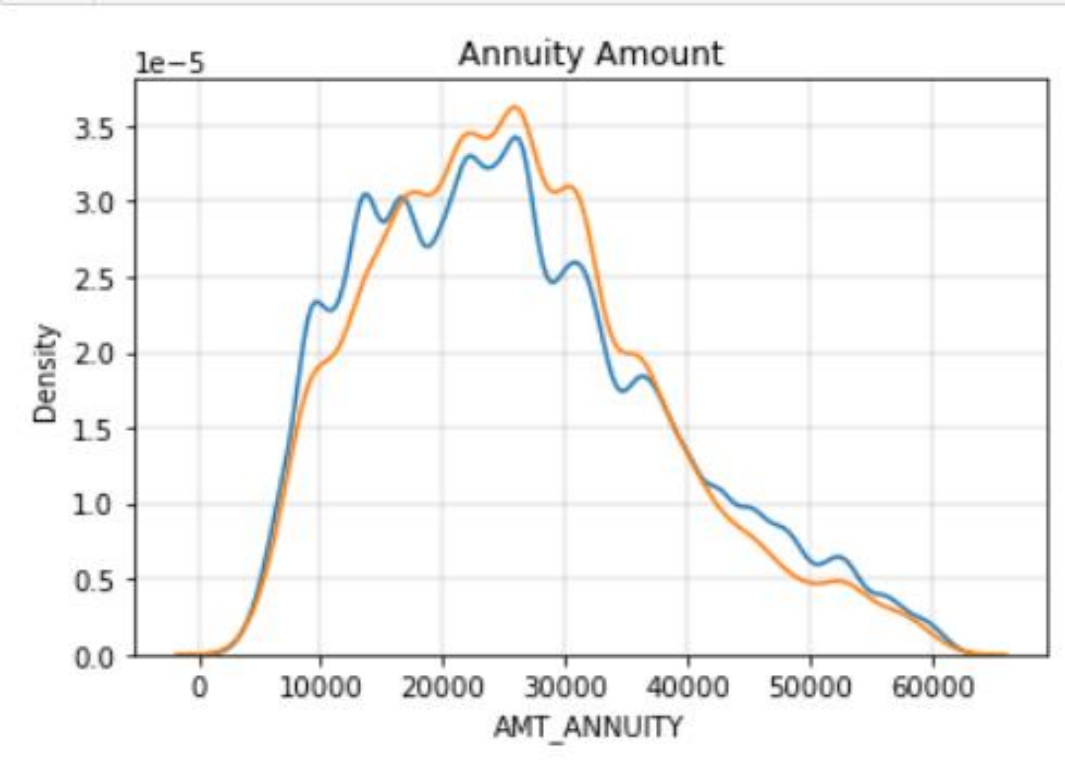


UNIVARIATE ANALYSIS - CONTINUOUS VARIABLES

CREDIT AMOUNT

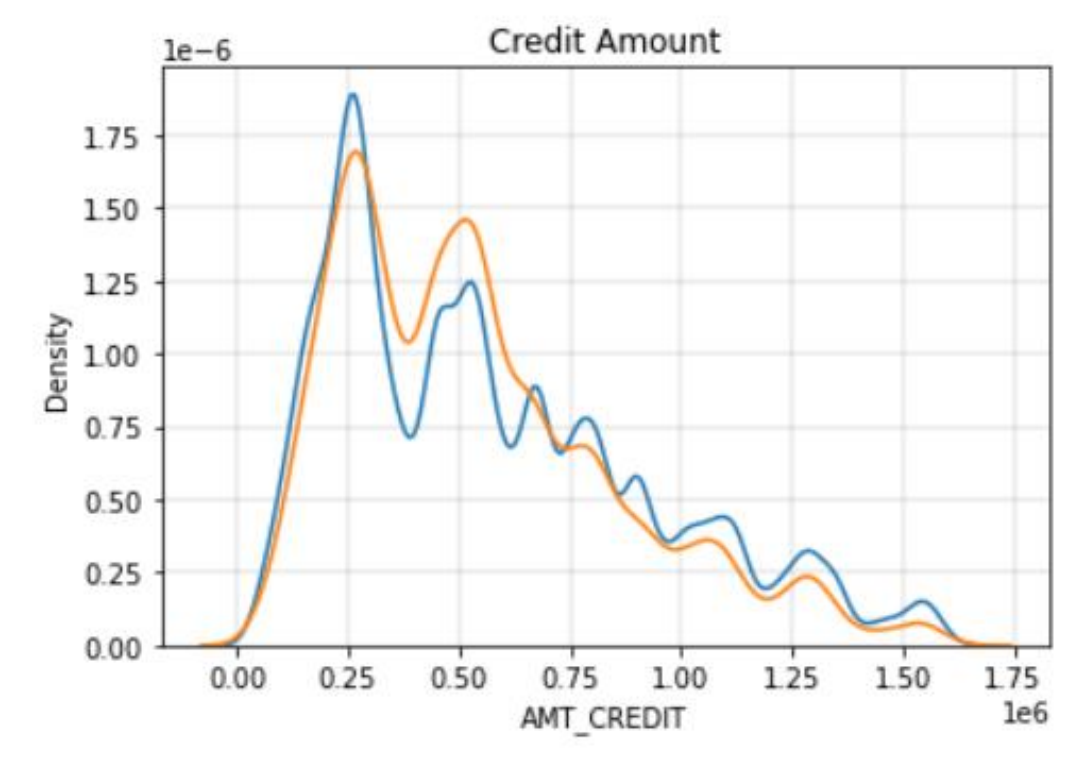


ANNUITY AMOUNT

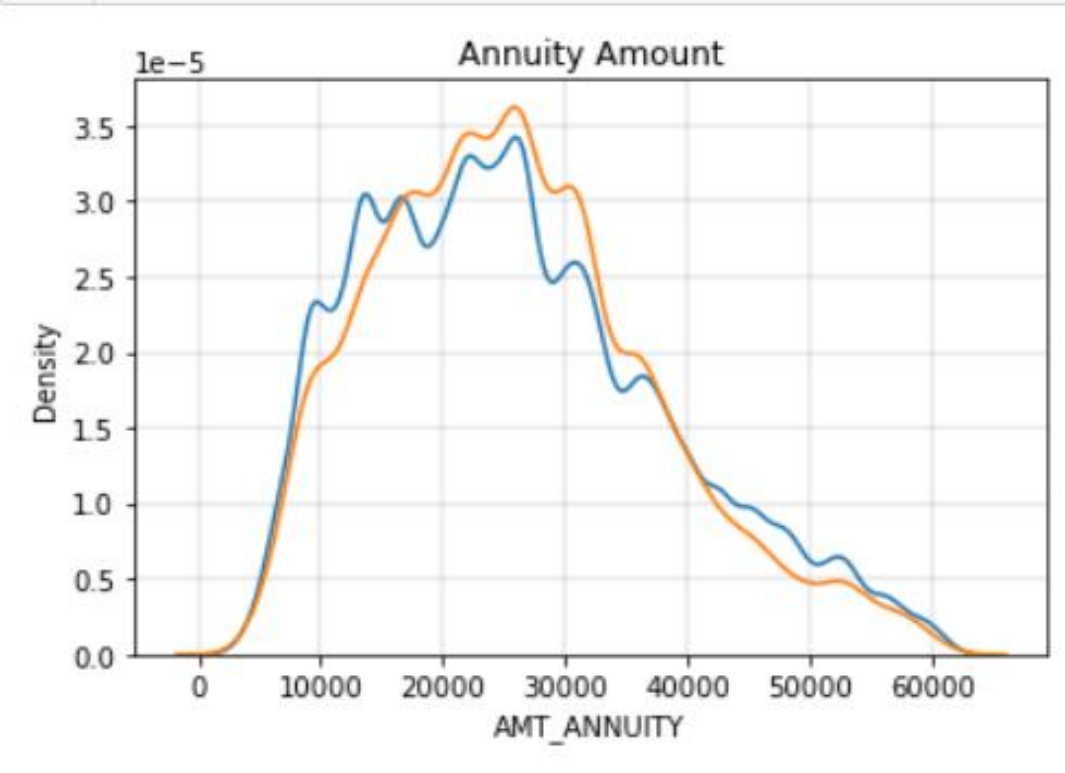


UNIVARIATE ANALYSIS - CONTINUOUS VARIABLES

CREDIT AMOUNT



ANNUITY AMOUNT



BIVARIATE ANALYSIS - NUMERIC - NUMERIC

INCOME VS GOOD PRICE

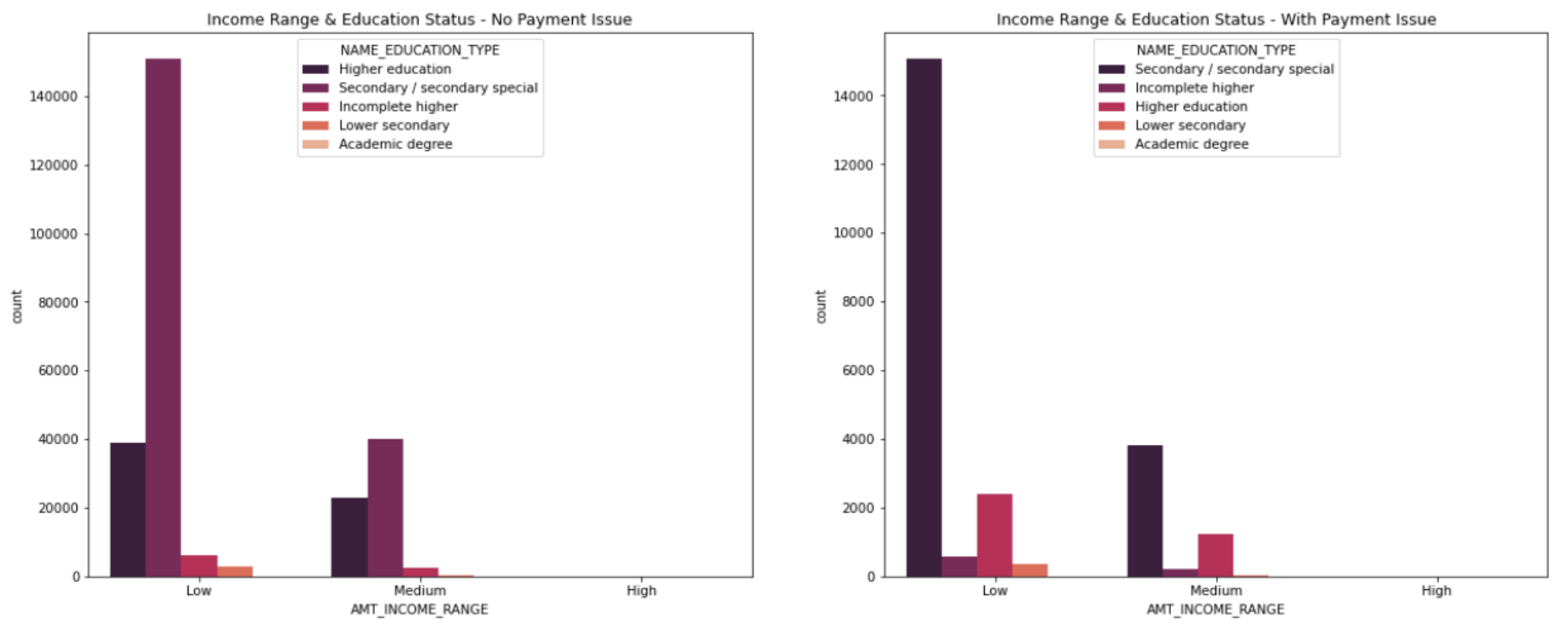
BIVARIATE ANALYSIS - NUMERIC - NUMERIC

GOODS PRICE VS CREDIT

BIVARIATE ANALYSIS - NUMERIC - CATEGORICAL

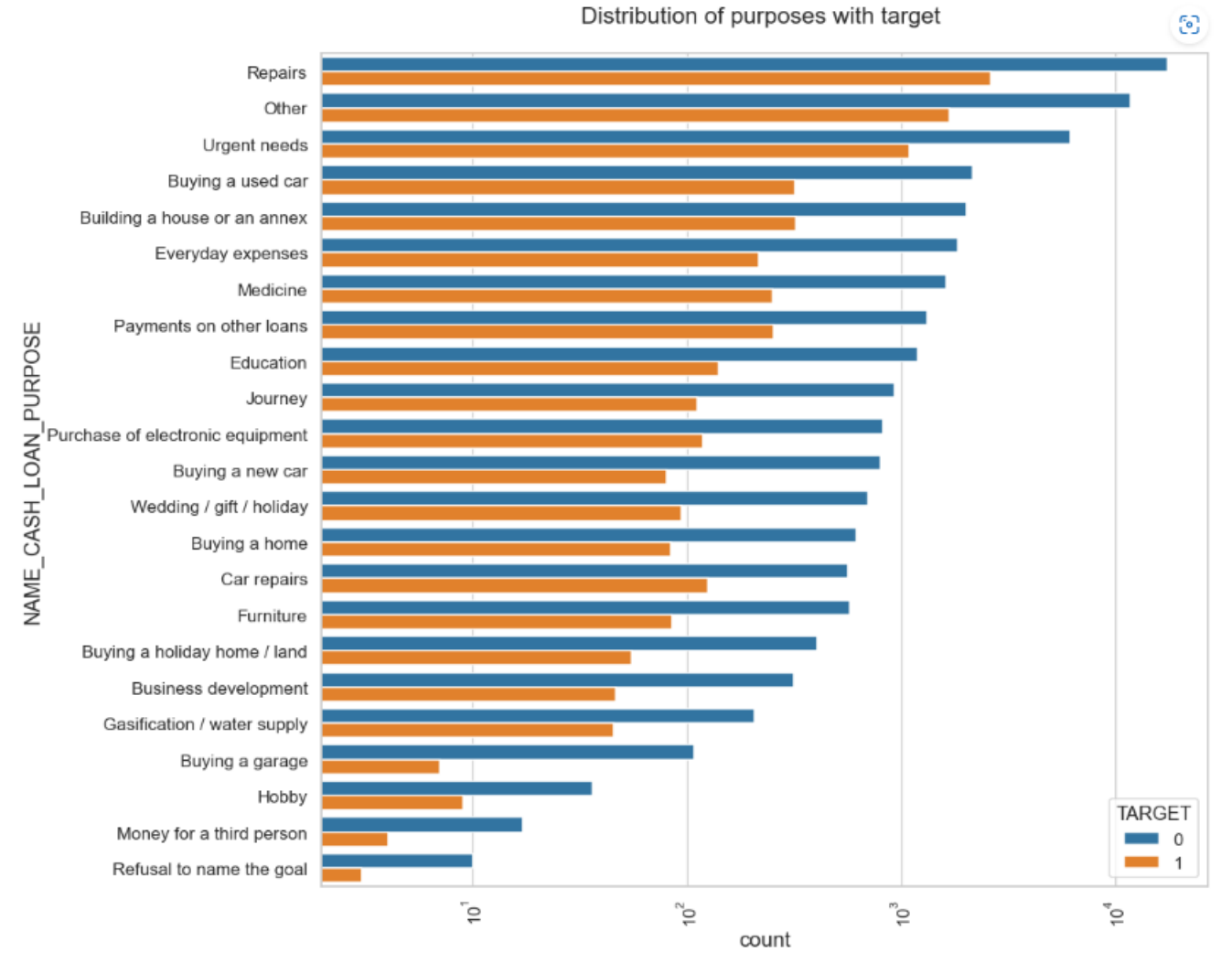
INCOME VS EDUCATION STATUS

We can see here, that people with secondary/secondary special education status and a low income range are most likely to repay the loan.



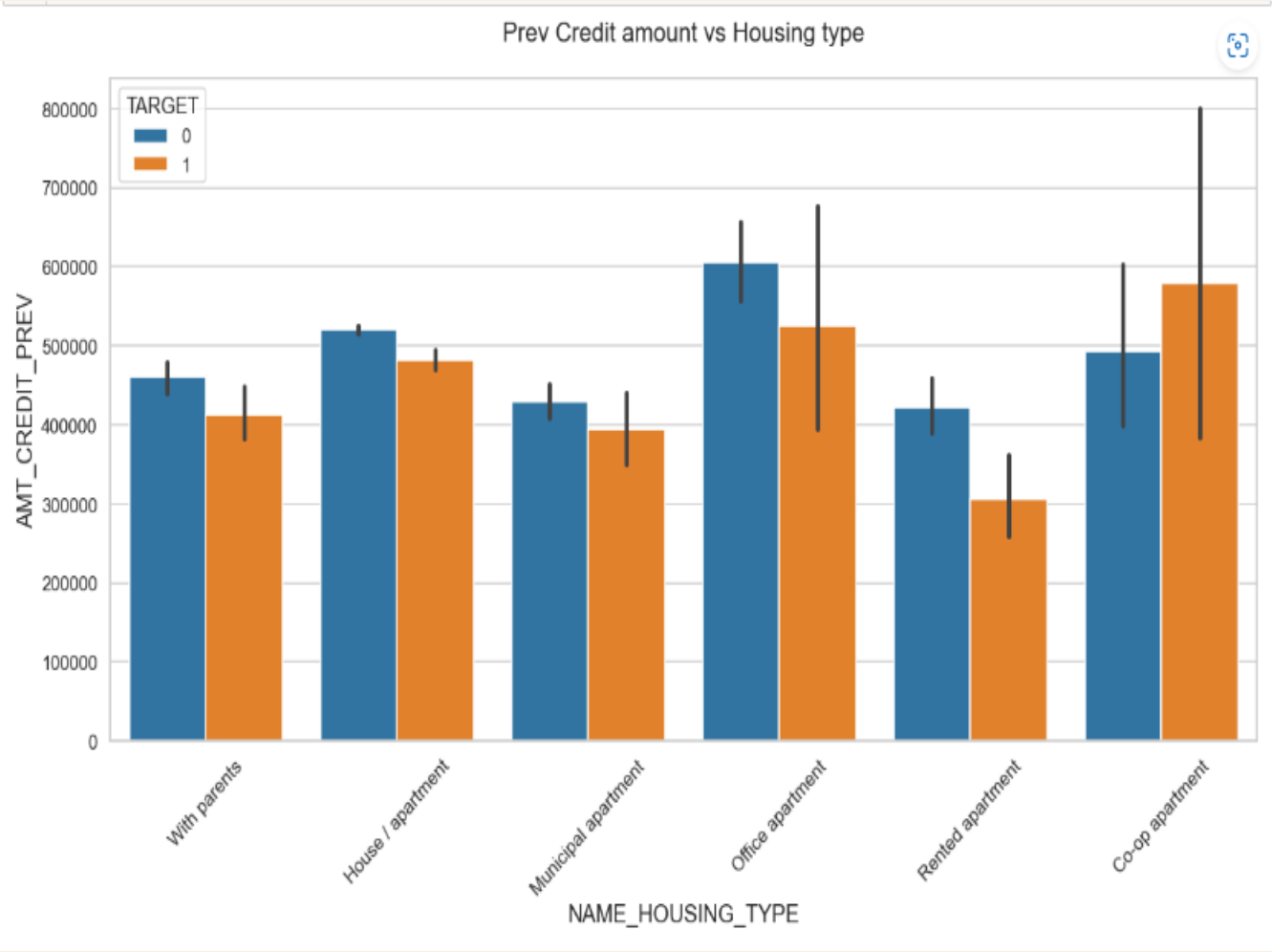
UNIVARIATE ANALYSIS - PREVIOUS MERGED DATA

Here, we can see that Categories like 'Buying a garage', 'Money for a third person', etc are the categories who have a high chance of successful loan repayment and should be prioritized.



BIVARIATE ANALYSIS - PREVIOUS MERGED DATA

When it comes to the ease of payment, "office apartment" category has the higher credit as compared to others. Also, bank should be careful while approving loans for 'co-op apartment' categories.



CONCLUSION

1) *People taking credit for commodities like new car or new garage are beneficial as they are loyal customers and pay back religiously.*
2) *Bank needs to be careful while giving away loans to widows in comparison of married people as they have a high chance of paying back.*
3) *Customers who have low credit amount are more likely to payback, Hence bank should focus more on small loan amounts.*
4) *People with high education status are more likely to pay back as compared to people with lower education.*

THANK YOU