CREDIT EDA ASSIGNMENT

PRESENTATION BY --
AISHWARYA KHATRI

CONTENTS

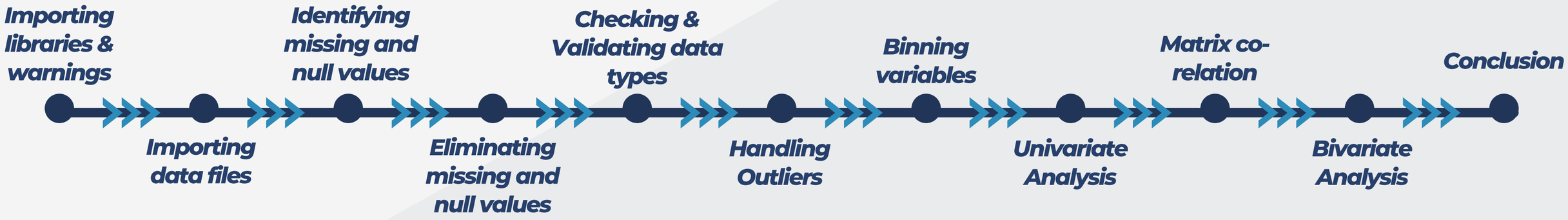
PROBLEM STATEMENT

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
Matrix co-relation
Bivariate Analysis
Conclusion

Importing libraries

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

Imported warnings

Highlights warnings however the program runs.

READING DATASET

- The flag variable is our target variable which highlights if the client will pay installment 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 null values in aapplication_data.csv and eliminated 49 columns which had null values more than 40%

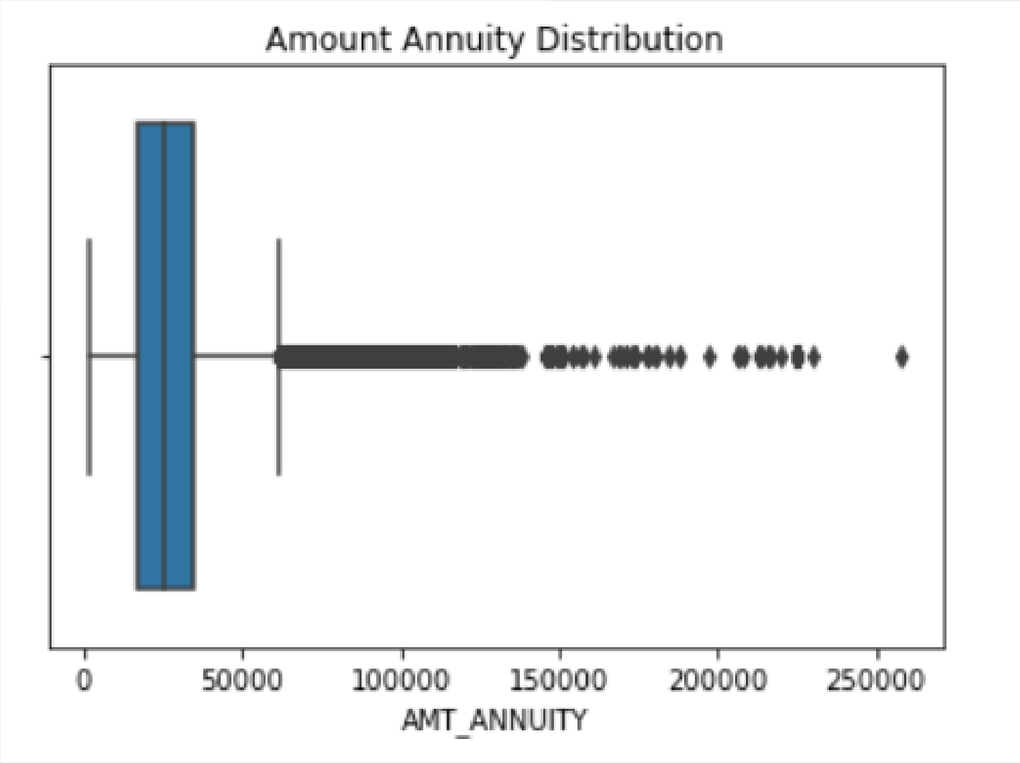
- Post that, AMT_ANNUITY, AMT_GOODS_PRICE, EXT_SOURCE_2, NAME_TYPE_SUITE, had less than 1% of null (& numeric) values. Hence, identified outliers and imputed using the best approach available.

- checked for unique values in columns by the following condition:
  1) If the count of unique values <=40, it's a categorical column
  2) If the count of unique values > 50, it's a continuous column

OUTLIERS HANDLING

- AMT_ANNUITY variable

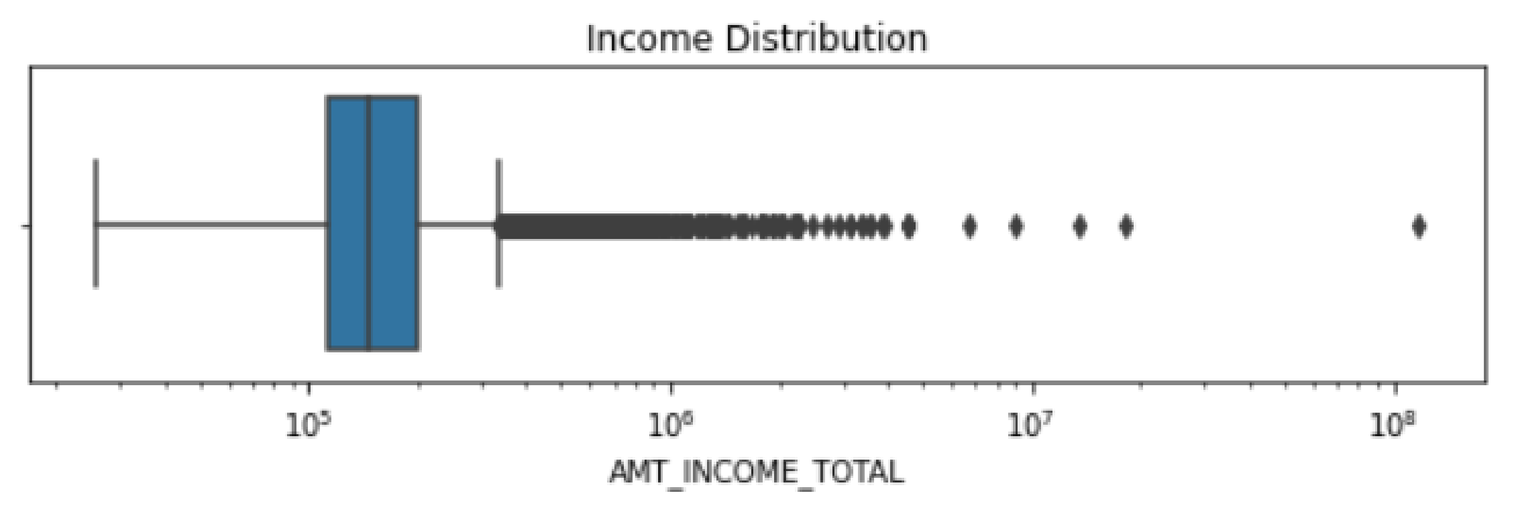
As seen here, outlier is present at 258025. Hence to impute the outlier values, we will use median here.



OUTLIERS HANDLING

- AMT_INCOME variable

  Here in 'AMT_INCOME_TOTAL' variable outlier values stands at 1.17x10^8. As the 95th and 99th quantile values differ largely, we can conclude the presence of outliers in the data set. Hence, we will cap the same here.



OUTLIERS HANDLING

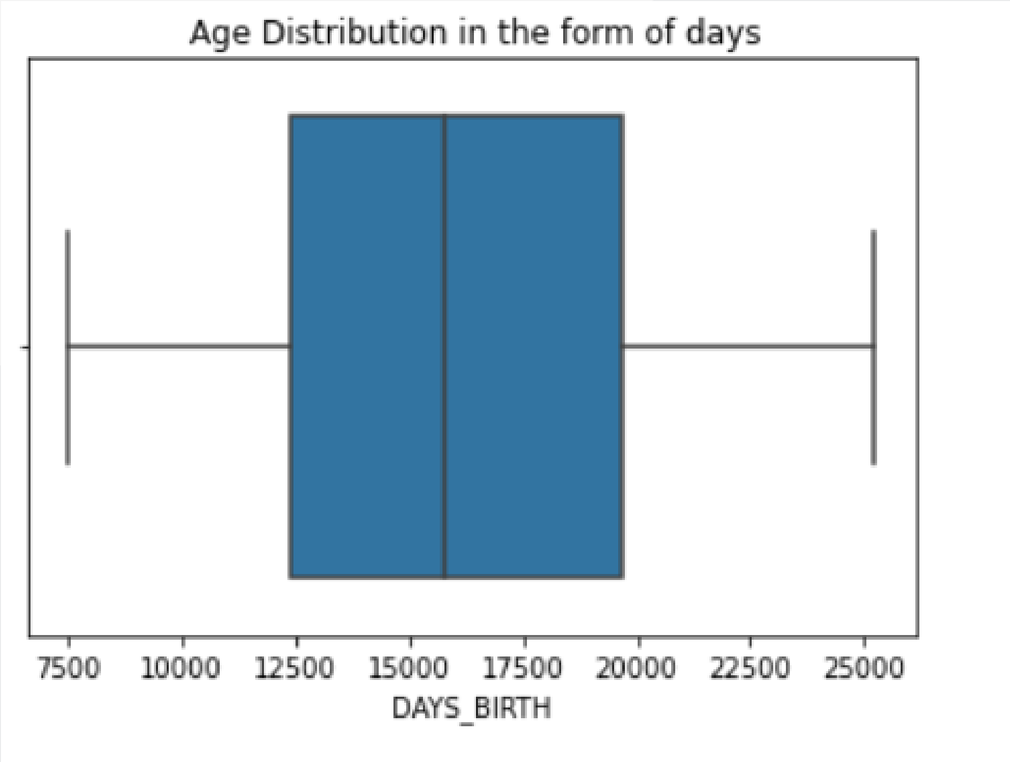
- AMT_CREDIT variable

The outliers here are present after the 99th quantile.

OUTLIERS HANDLING

- DAYS_BIRTH variable

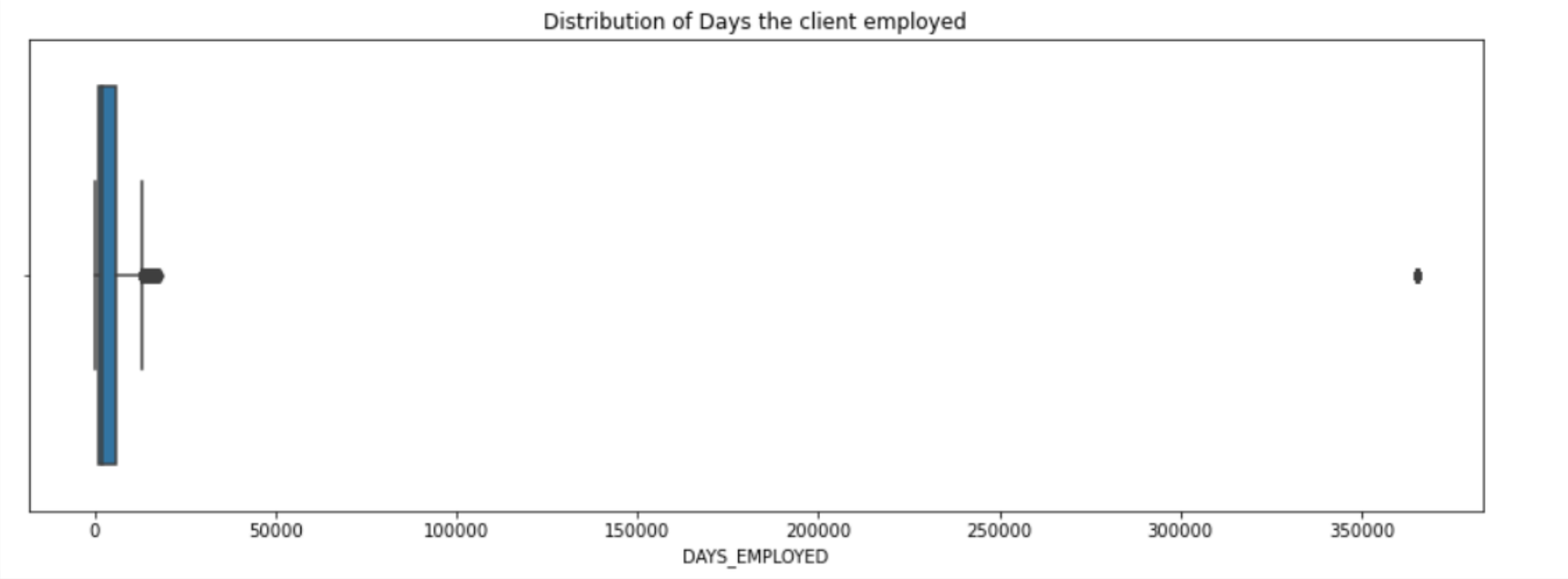
No outliers are present here.



OUTLIERS HANDLING

- DAYS_EMPLOYED variable

We have an outlier at 365243



ANALYSIS

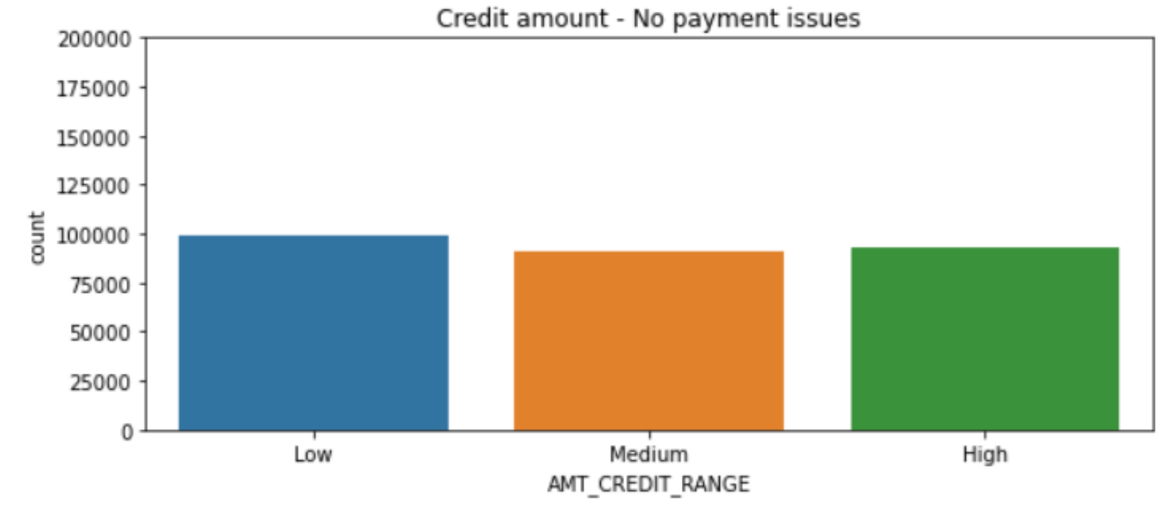
When we find the imbalance percentage, We can conclude that the 'TARGET VARIABLE' has 91.92% of 0s and 08.07% of 1s. Hence, we can conclude here that 91.92% of people here make timely payments and only 8.02% face challenges.

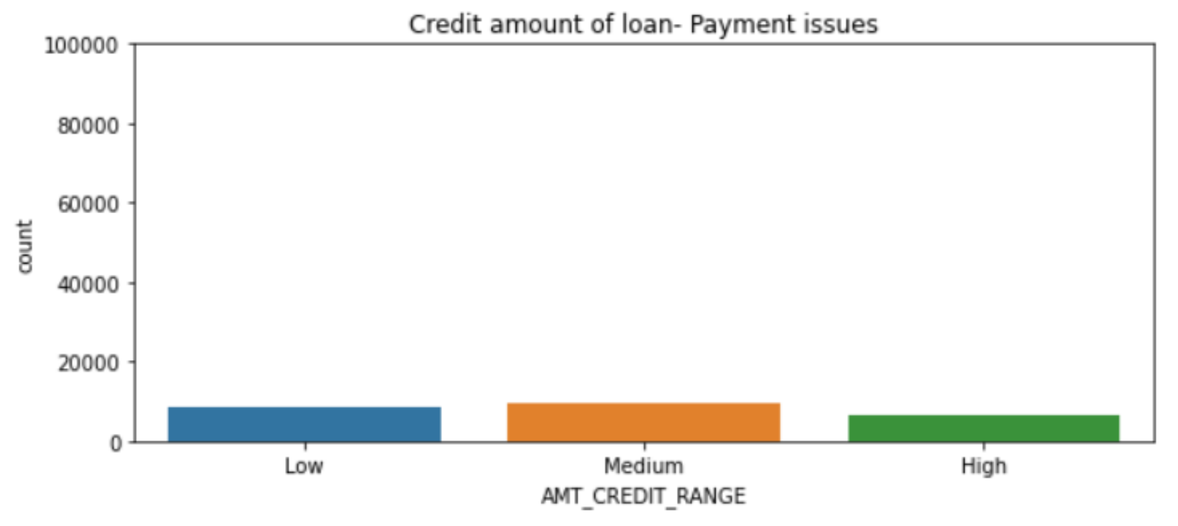
```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

UNIVARIATE ANALYSIS - 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.

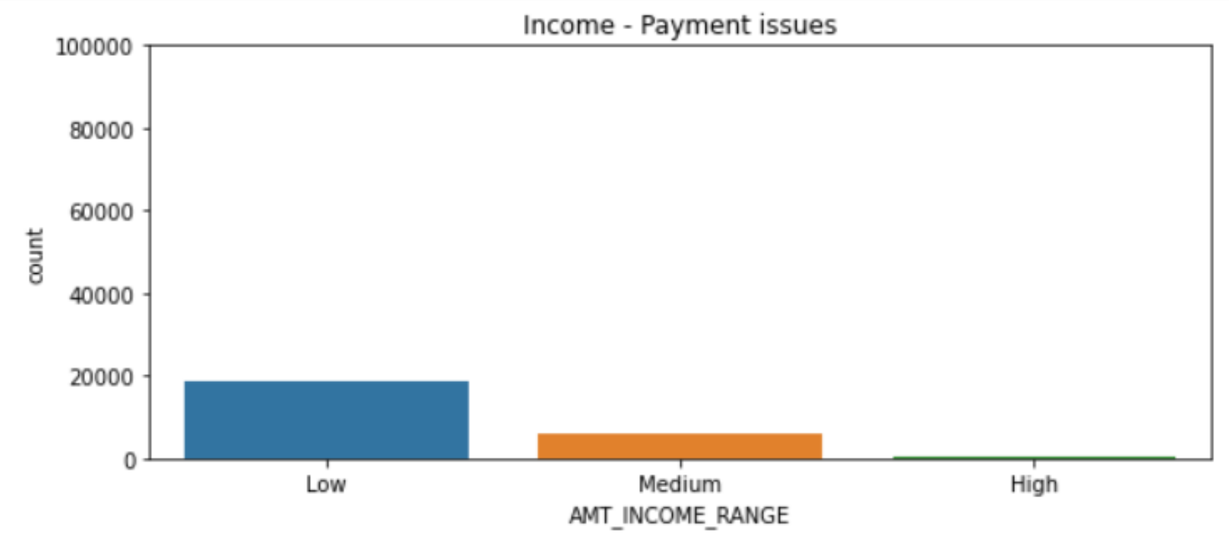


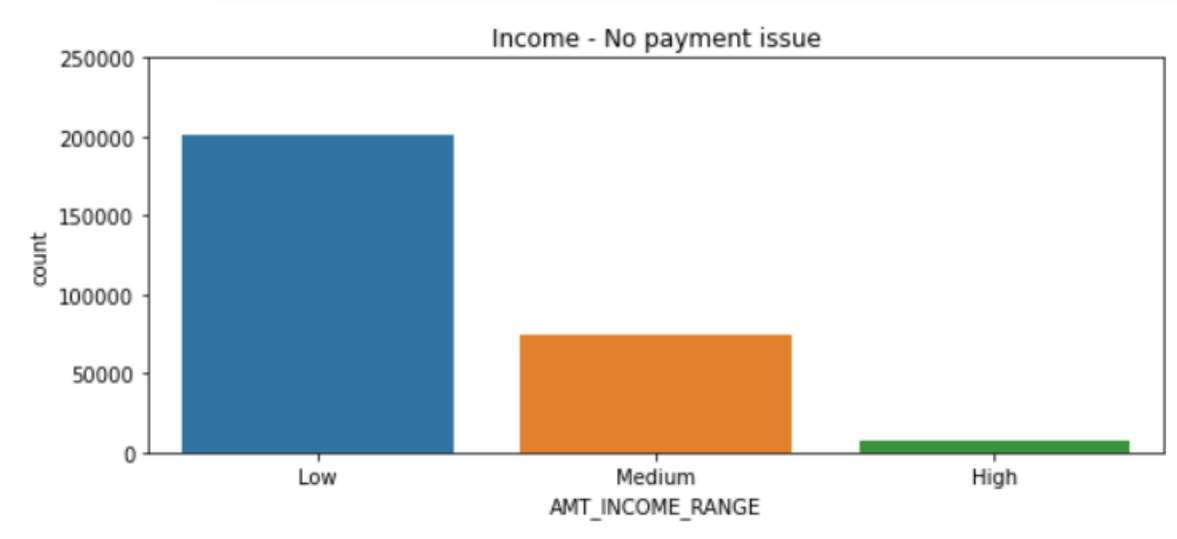


UNIVARIATE ANALYSIS - NUMERIC VARIABLES

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.

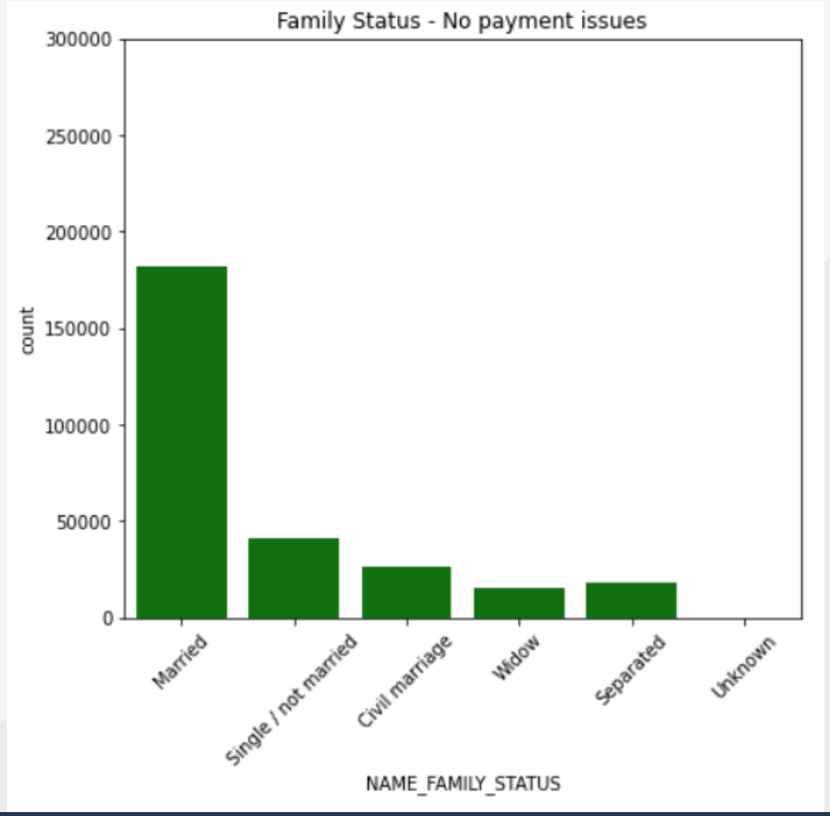


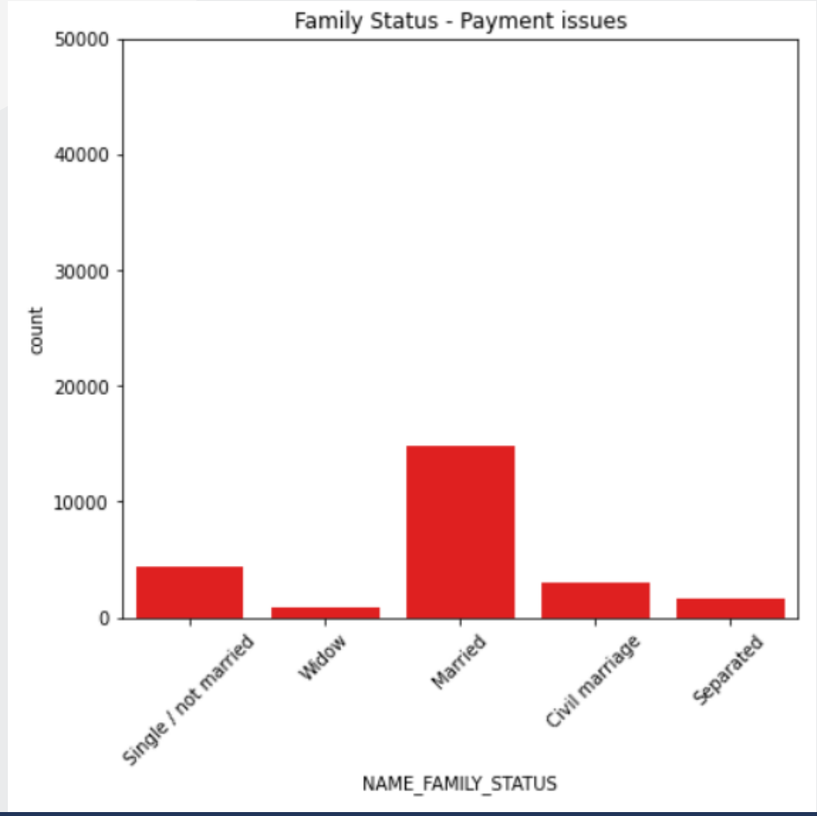


UNIVARIATE ANALYSIS - CATEGORICAL VARIABLES

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.

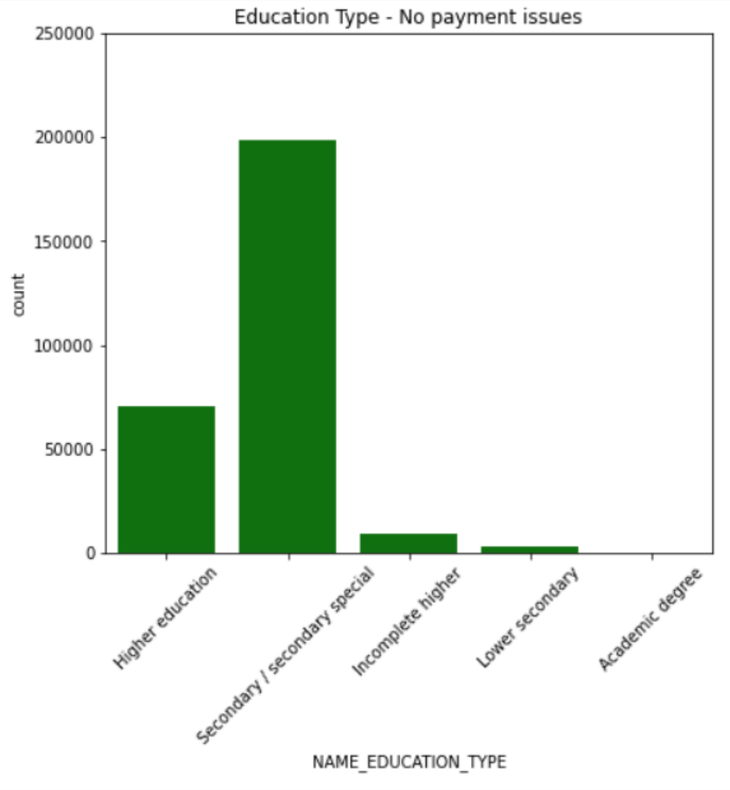


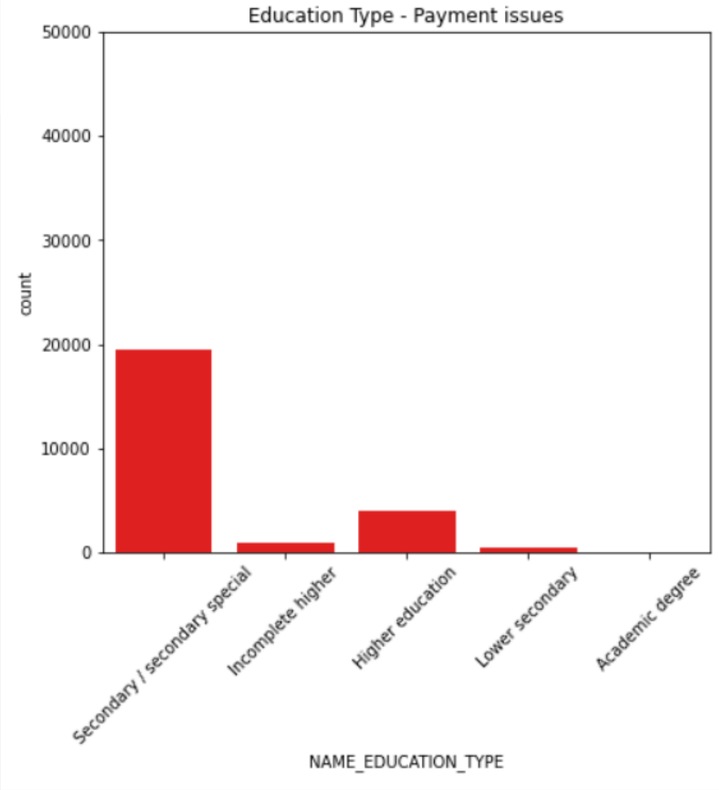


UNIVARIATE ANALYSIS - CATEGORICAL VARIABLES

Education type analysis

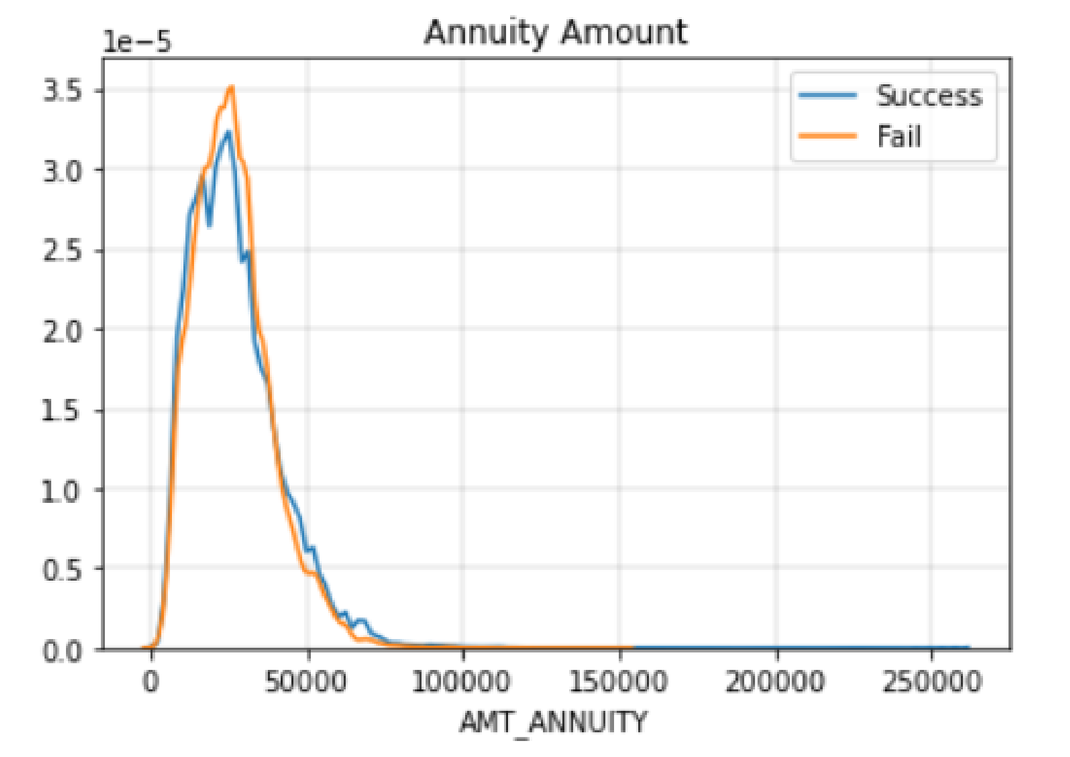
People most likely to repay te loan have secondary/Secondary special education status whereas clients having an academic degree are the most defaulters.



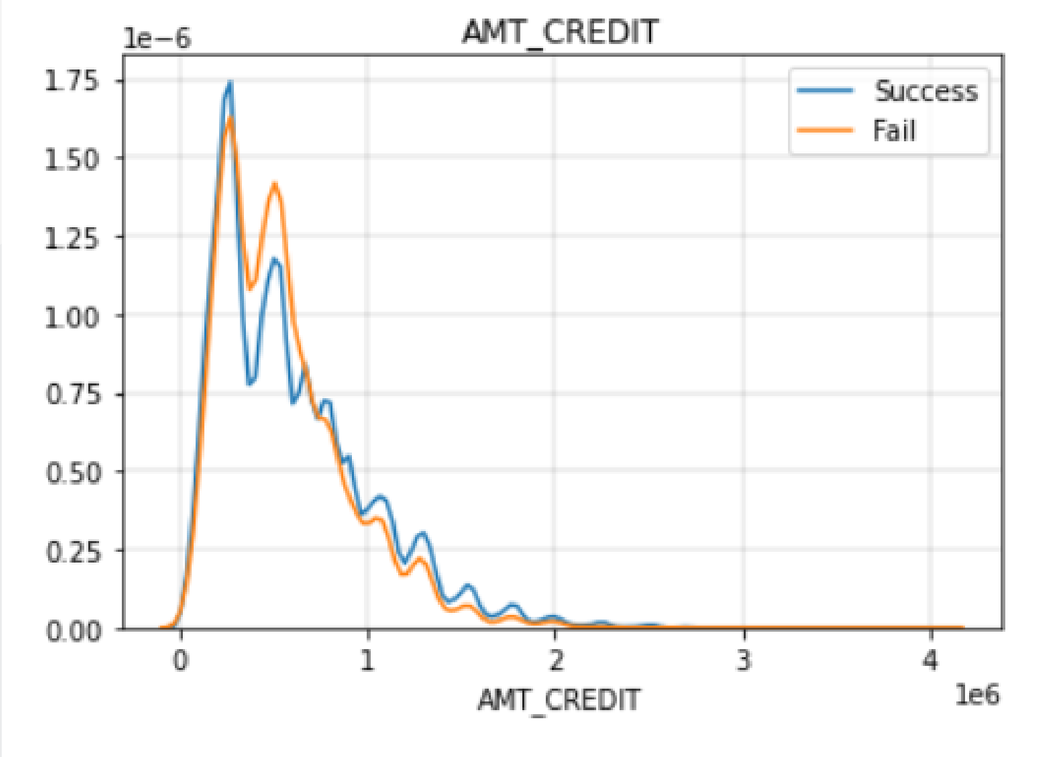


UNIVARIATE ANALYSIS - CONTINUOUS VARIABLES

Annuity Amount

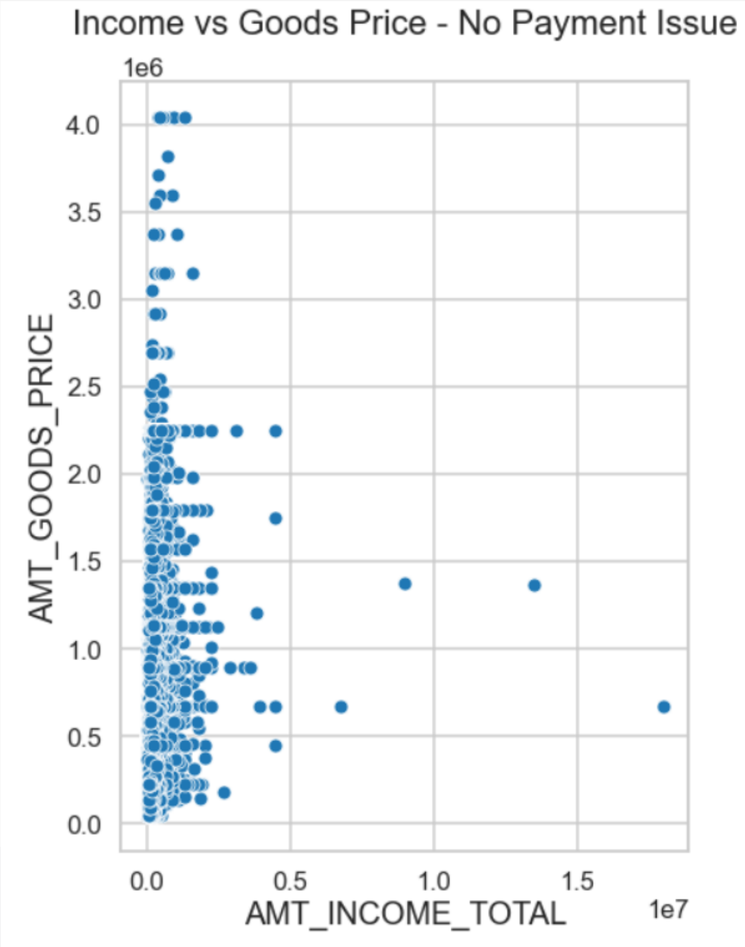


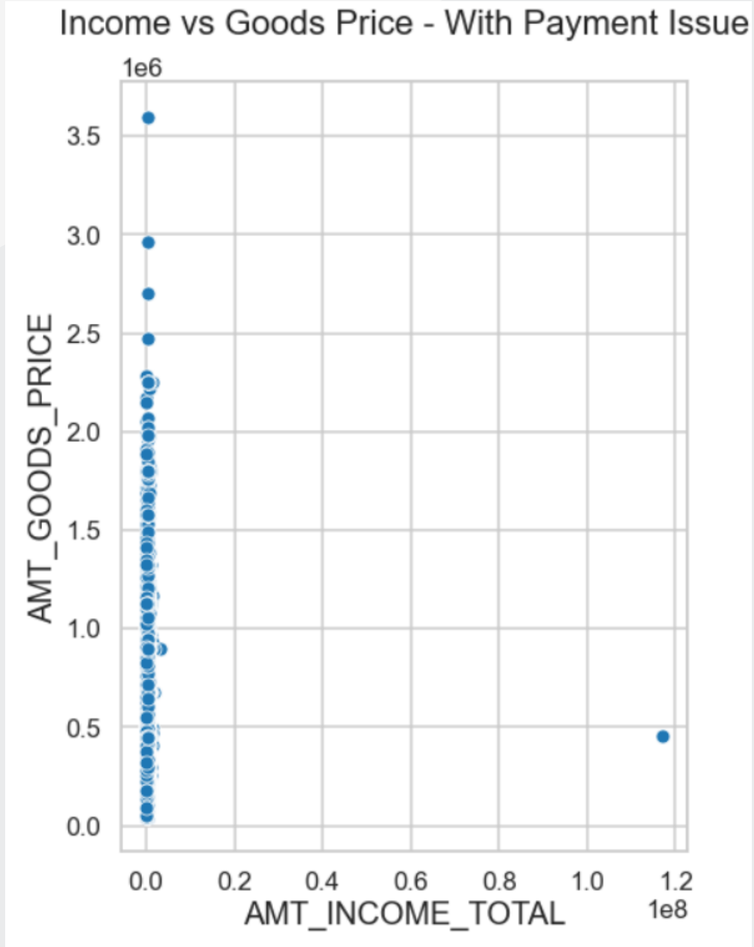
Credit Amount



BIVARIATE ANALYSIS - NUMERIC - NUMERIC

Income vs Good Price

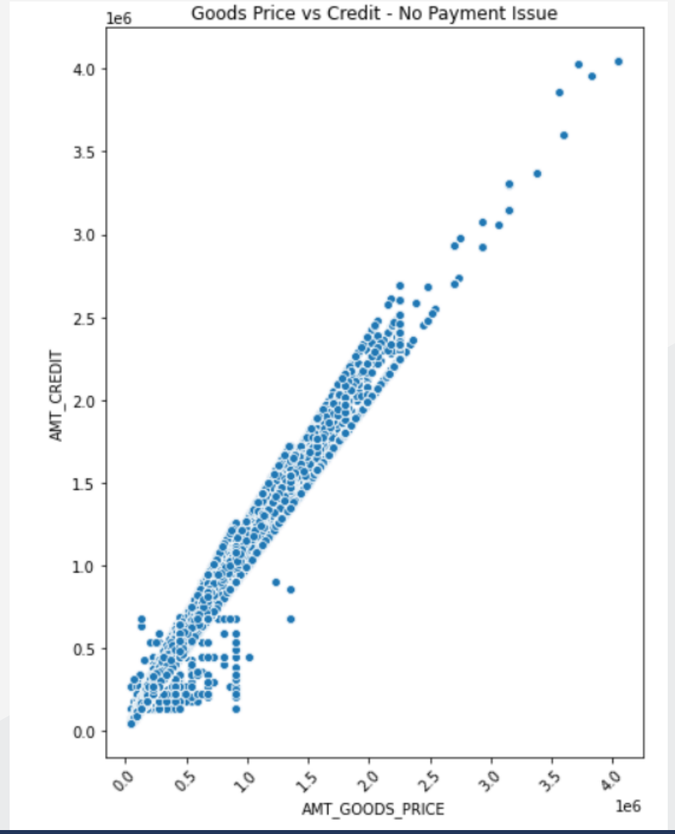


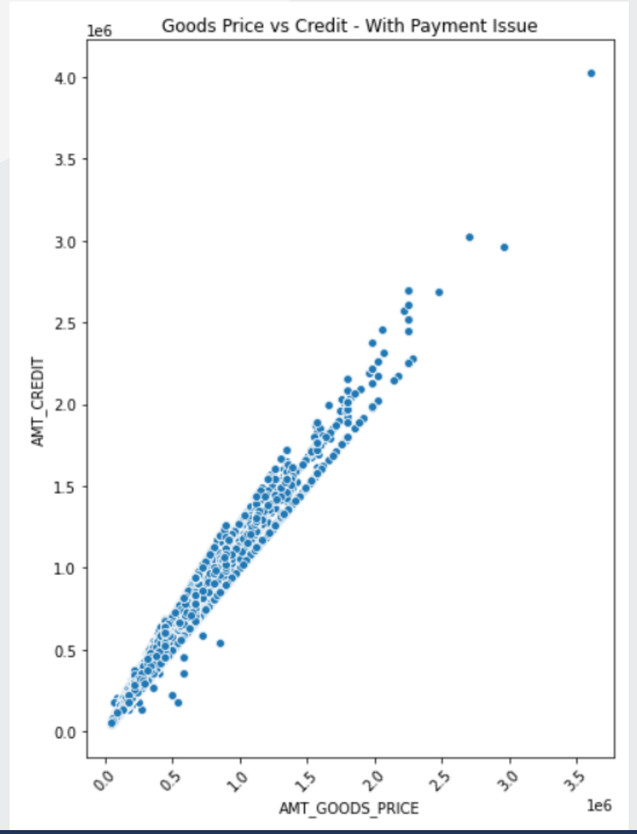


BIVARIATE ANALYSIS - NUMERIC - NUMERIC

Goods Price vs Credit

People who repay the loan on time have a chance of getting a loan for more expensive goods. Also, people who repay loan on time, have a higher chance of getting credit for a particular goods value.

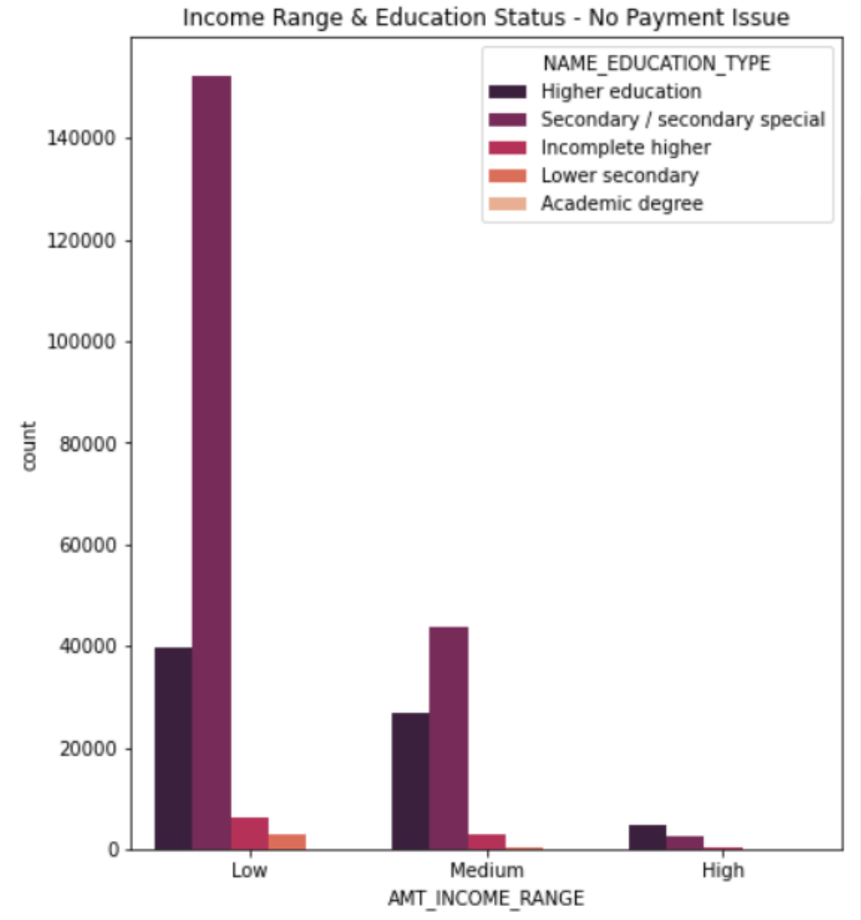


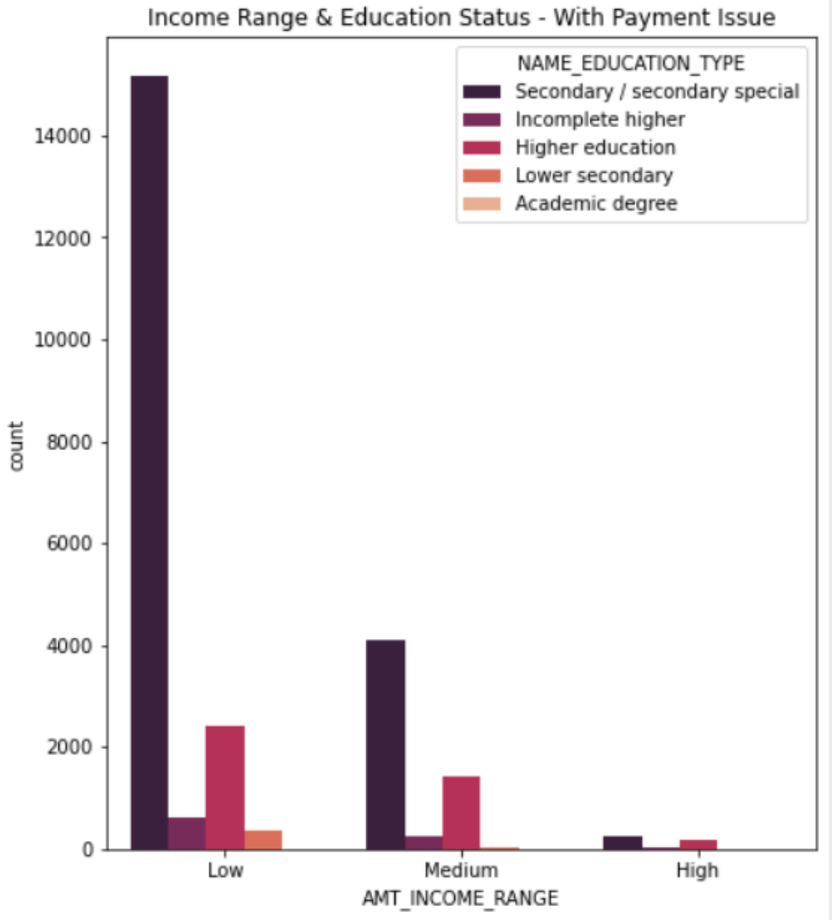


BIVARIATE ANALYSIS - NUMERIC - CATEGOTICAL

Income Range - 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.

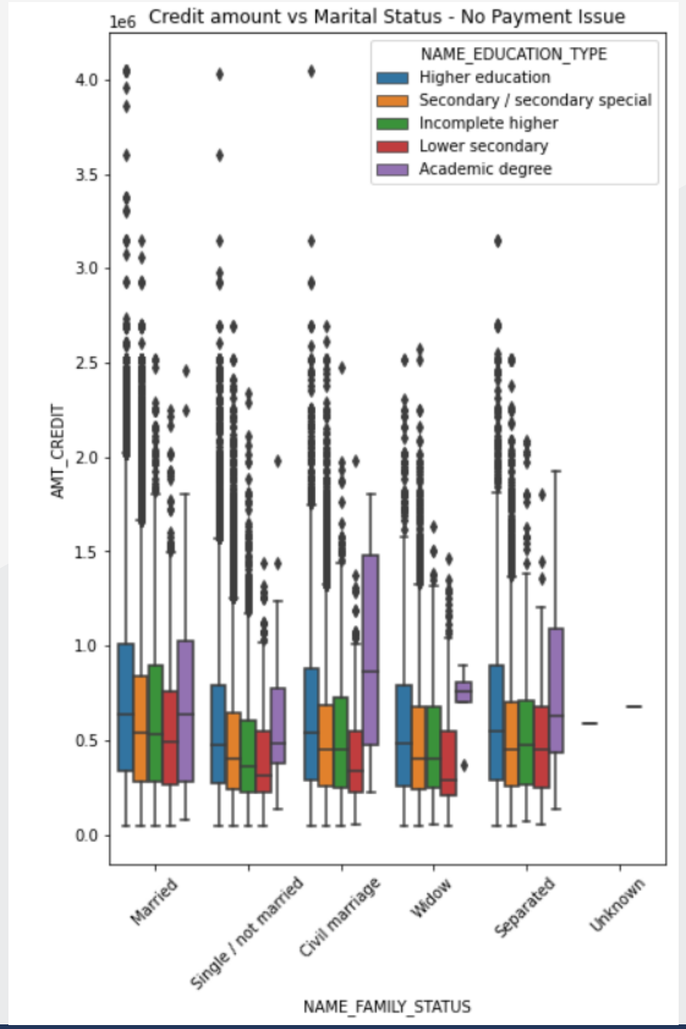


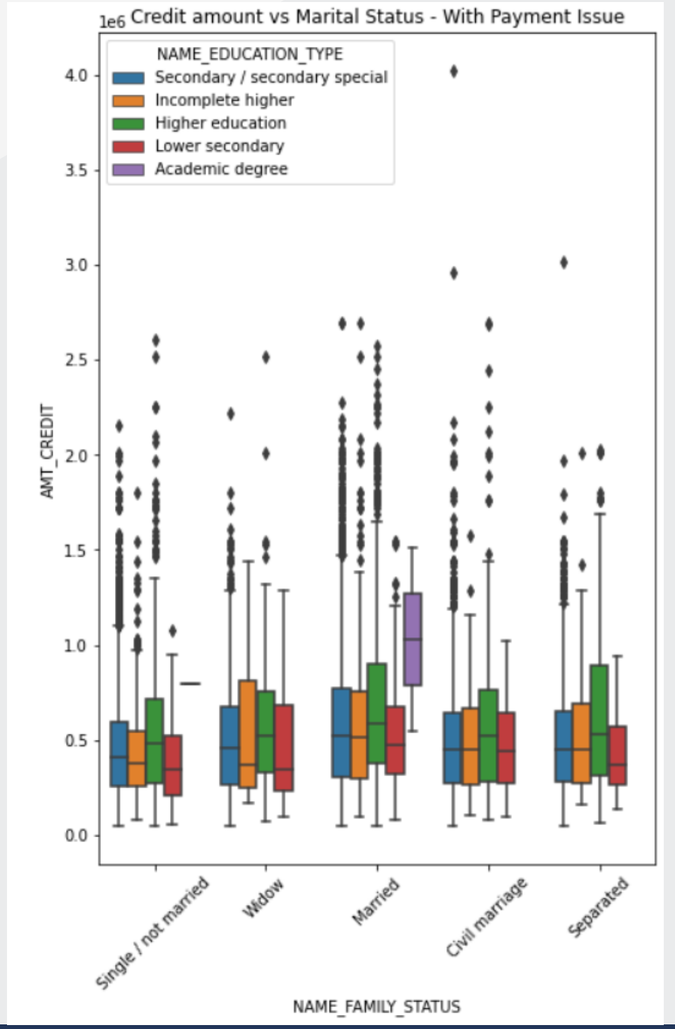


BIVARIATE ANALYSIS - NUMERIC - CATEGOTICAL

Credit Amount - Marital Status

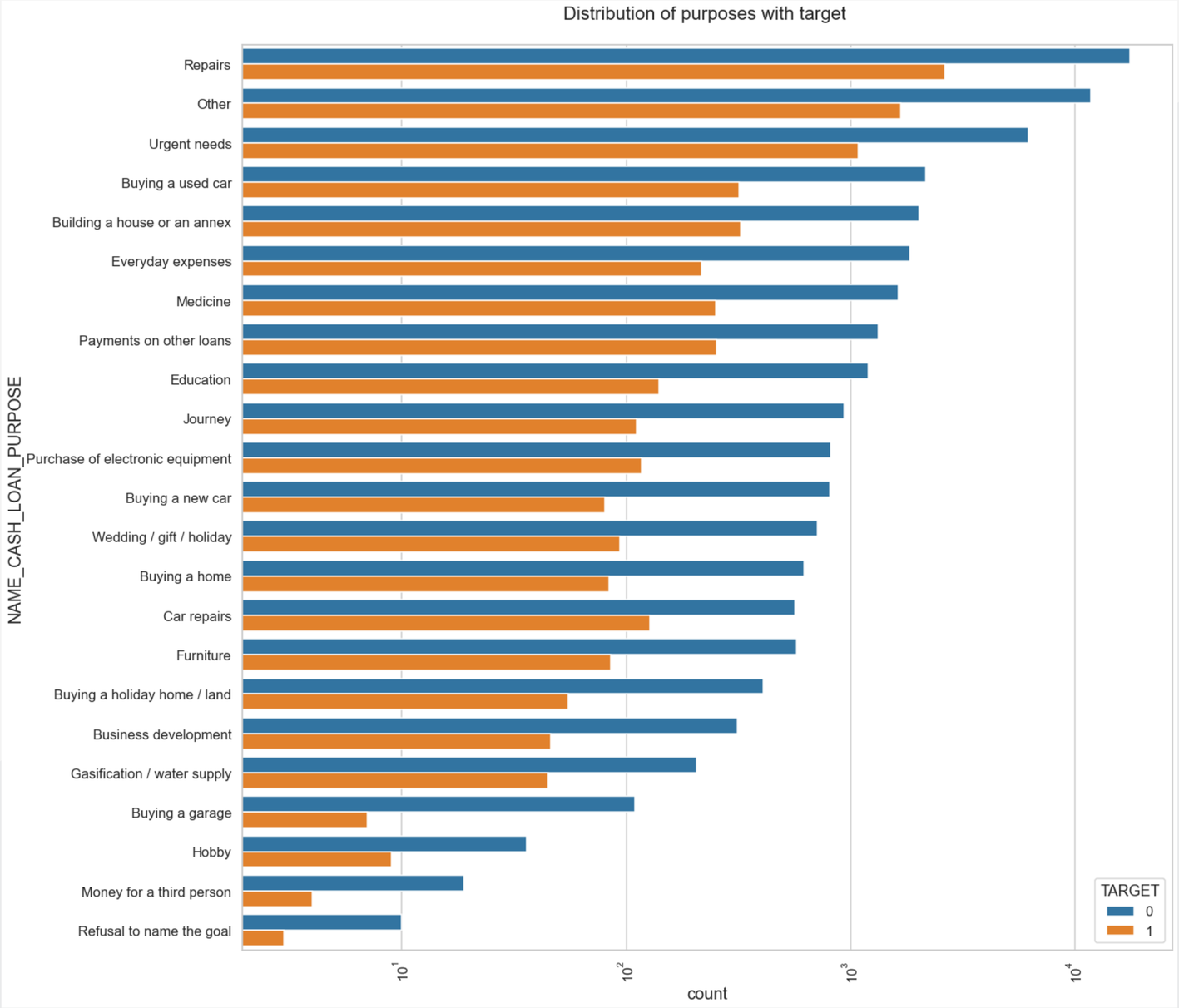
We can see from this graph, that married people with a higher education status are likely to get much higher credit as compared to a widow with a lower education status.





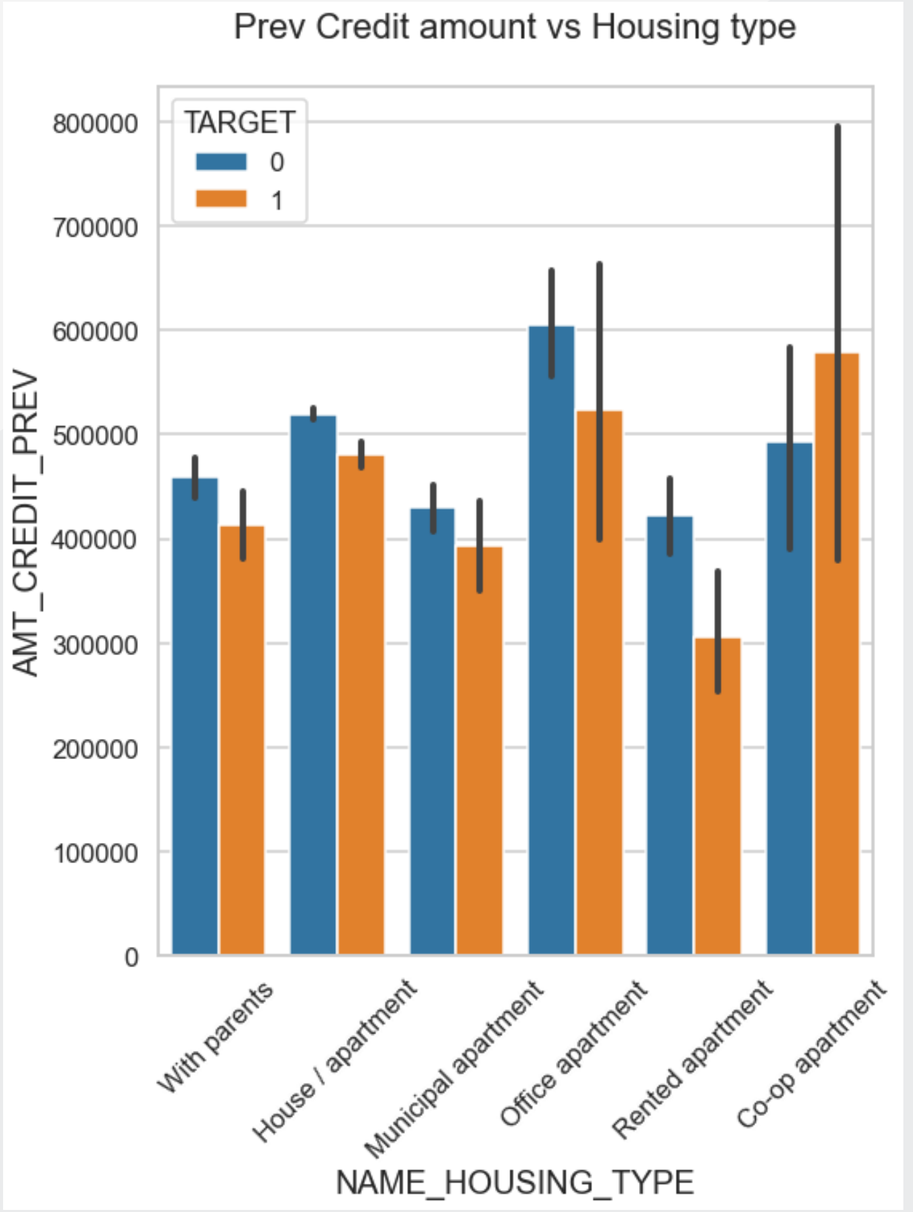
UNIVARIATE ANALYSIS - PREVIOUS APPLICATION

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 APPLICATION

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