EDA ASSIGNMENT

Contents

- Problem Statement / Business Objective
- Data Pre-Processing
- Analysis
  - Univariant Analysis
  - Bivariant Analysis
  - Segmented Univariant Analysis

Business Understanding

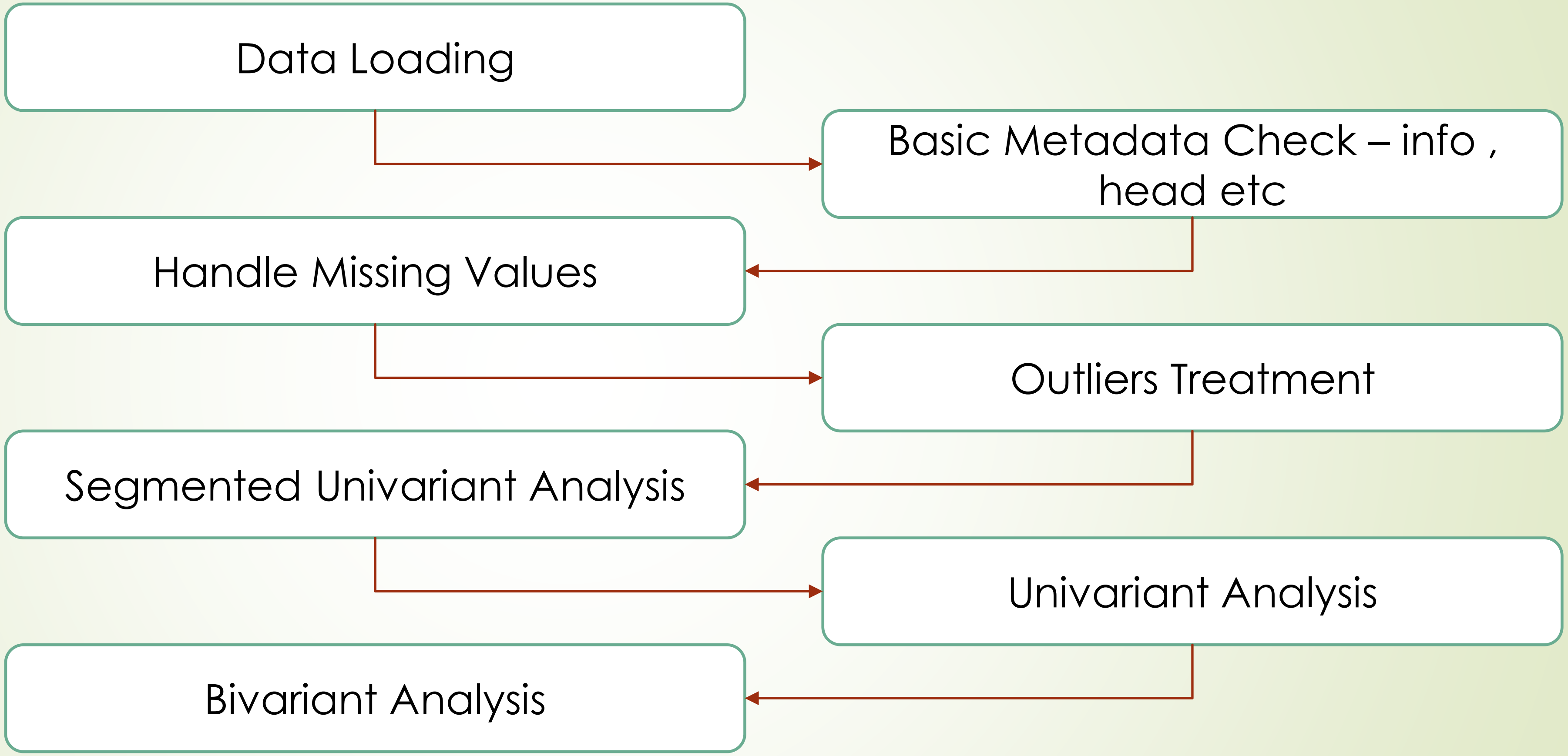
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile.
- Two types of risks are associated with the bank's decision:
  - ❑ If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - ❑ If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Problem Statement/Business Objective

- **Primary Aim:** Identifying various factors in Dataset that lead to default payments in loans, so that at the time of lending money to customers, the bank or the financial institution does not face any money loss. The company can utilize this knowledge for its portfolio and risk assessment.

- **Identifying Patterns:** Indicates if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Data Pre-Processing Flow Chart



Data Loading

Basic Metadata Check – info , head etc

Handle Missing Values

Outliers Treatment

Segmented Univariant Analysis

Univariant Analysis

Bivariant Analysis

Missing Values

- More than 40% of the values in 49 columns of the dataset are null. Drop those columns since the number of missing values is so high.
- Remove columns that are not relevant to the analysis.

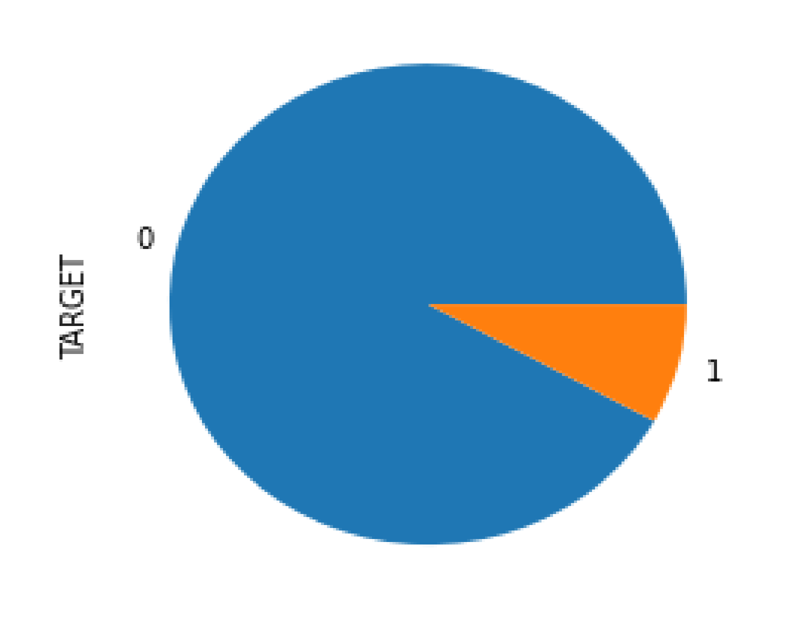
Outliers Treatment

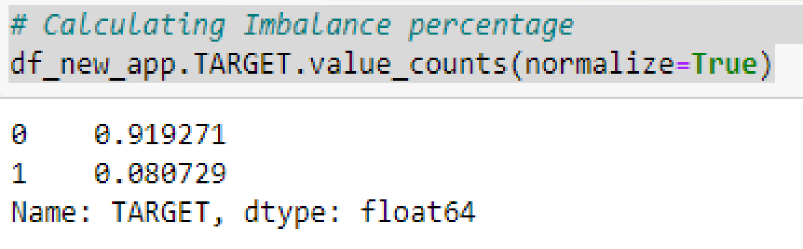
- AMT_GOODS_PRICE column have outliers

- Outliers in the DAYS_EMPLOYED column are at 365243. The important point to note in this situation is that the outlier figure is more than 1000 when we convert days to years. It is impossible for someone to work for a company for more than 1000 years.225375 rows in the DAYS EMPLOYED column out of 307511 records contain outliers, which significantly reduce the analyses' accuracy.

Analysis

- An applicant with a value of 1 has some payment issues, while an applicant with a value of 0 has paid all of their EMIs on time, according to the "TARGET" variable in the application dataset.
- The TARGET column has 8.07% of 1s, meaning that 8% of clients are having payment issues and 91.92% are not



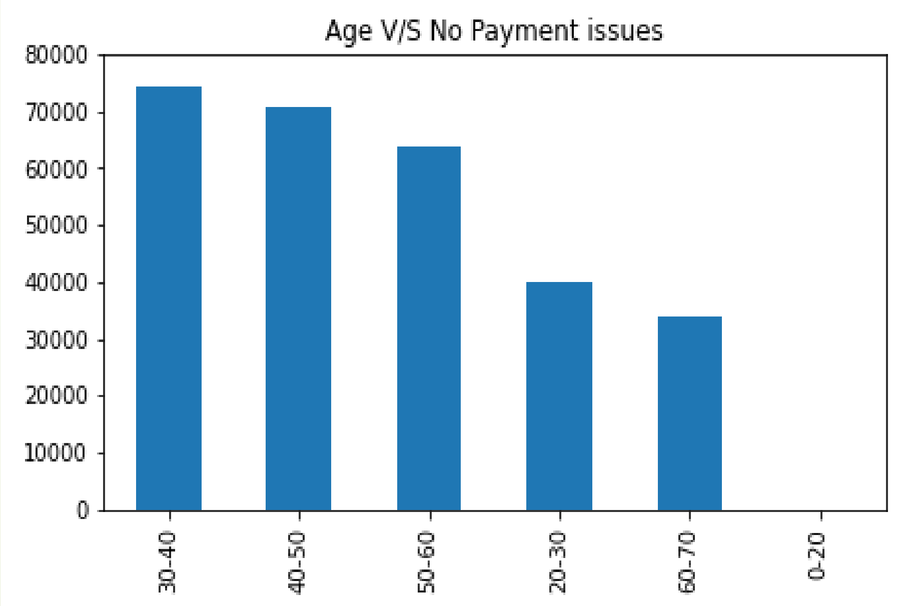


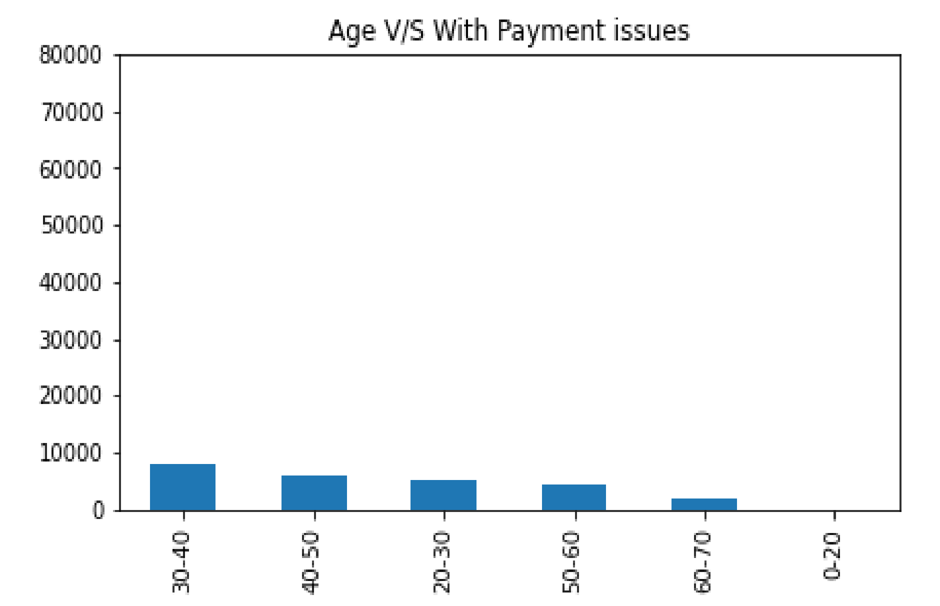
```
# Calculating Imbalance percentage
df_new_app.TARGET.value_counts(normalize=True)
```

```
0    0.919271
1    0.080729
Name: TARGET, dtype: float64
```

Segmented Univariate Analysis

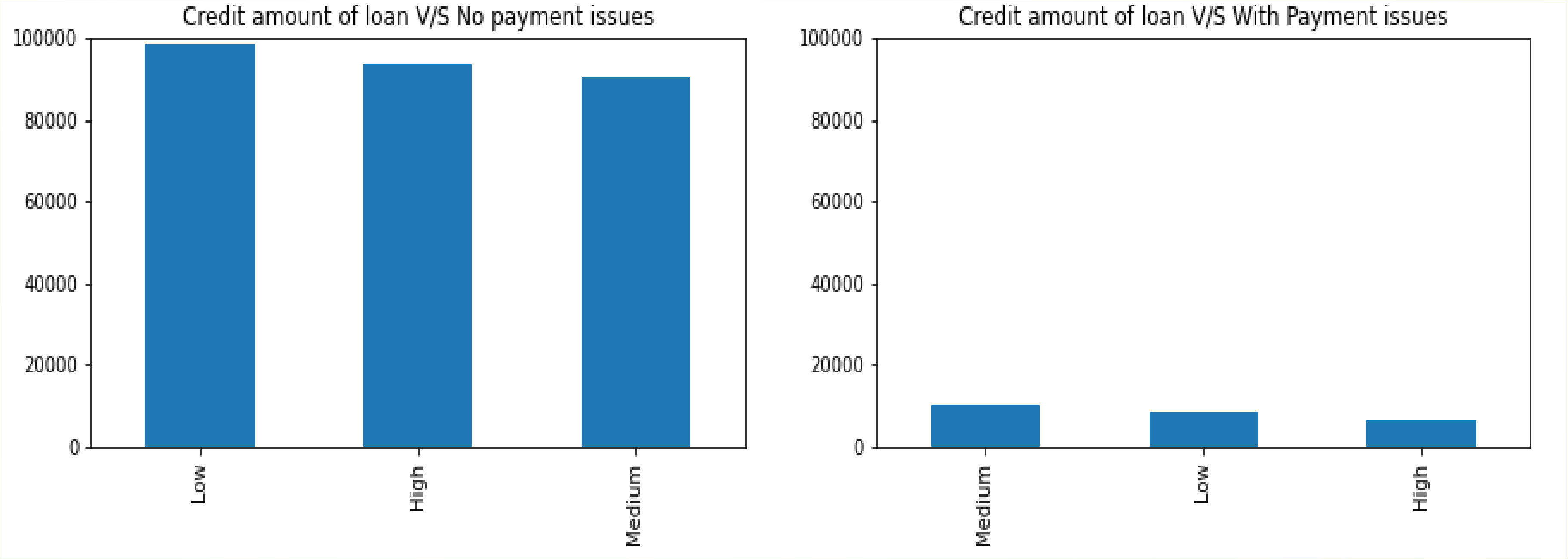
- We can observe that customers belonging to age group 30-40 are able to make payment on time and can be considered while lending loan! but on the other chart we can visualize that 30-40 age group customers are the one who has defaulted on Lone, Customer of Age group 60-70 are less likely to default on loans





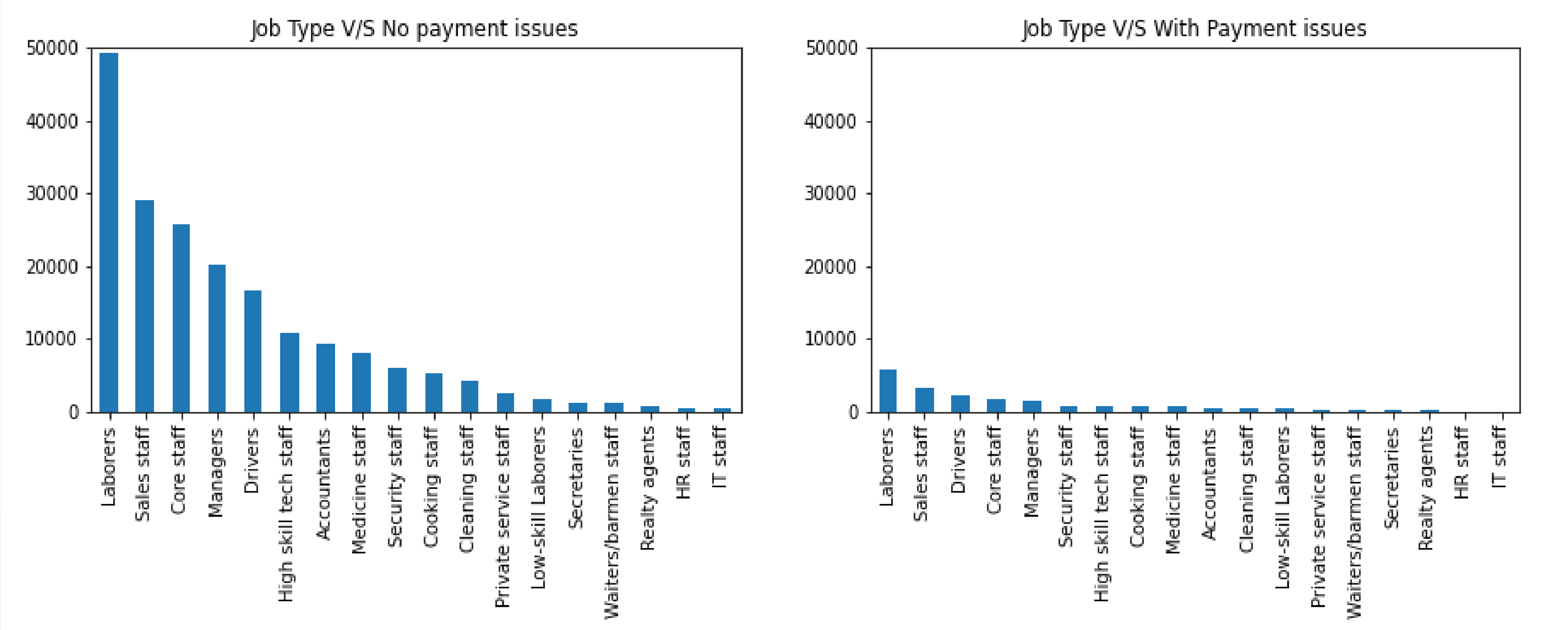
Segmented Univariate Analysis

- Customers with medium credit amounts are more likely to miss payments on lone payments, whereas customers with high credit amounts may be considered for leading lone payments.



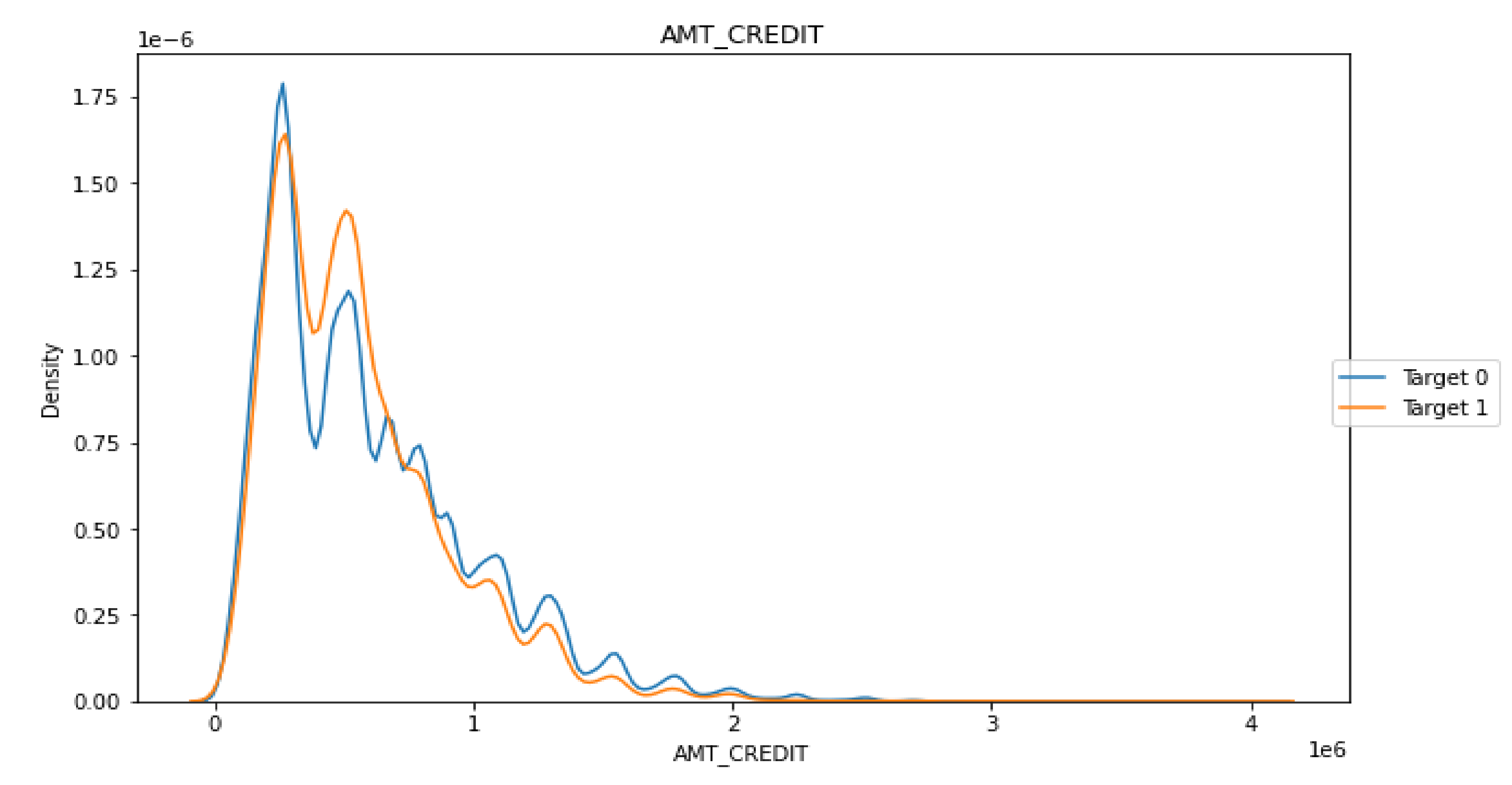
Segmented Univariate Analysis

- ➢ The plot unmistakably demonstrates that Laborers are more likely than IT Staff to pay their installments on time.



Segmented Univariate Analysis

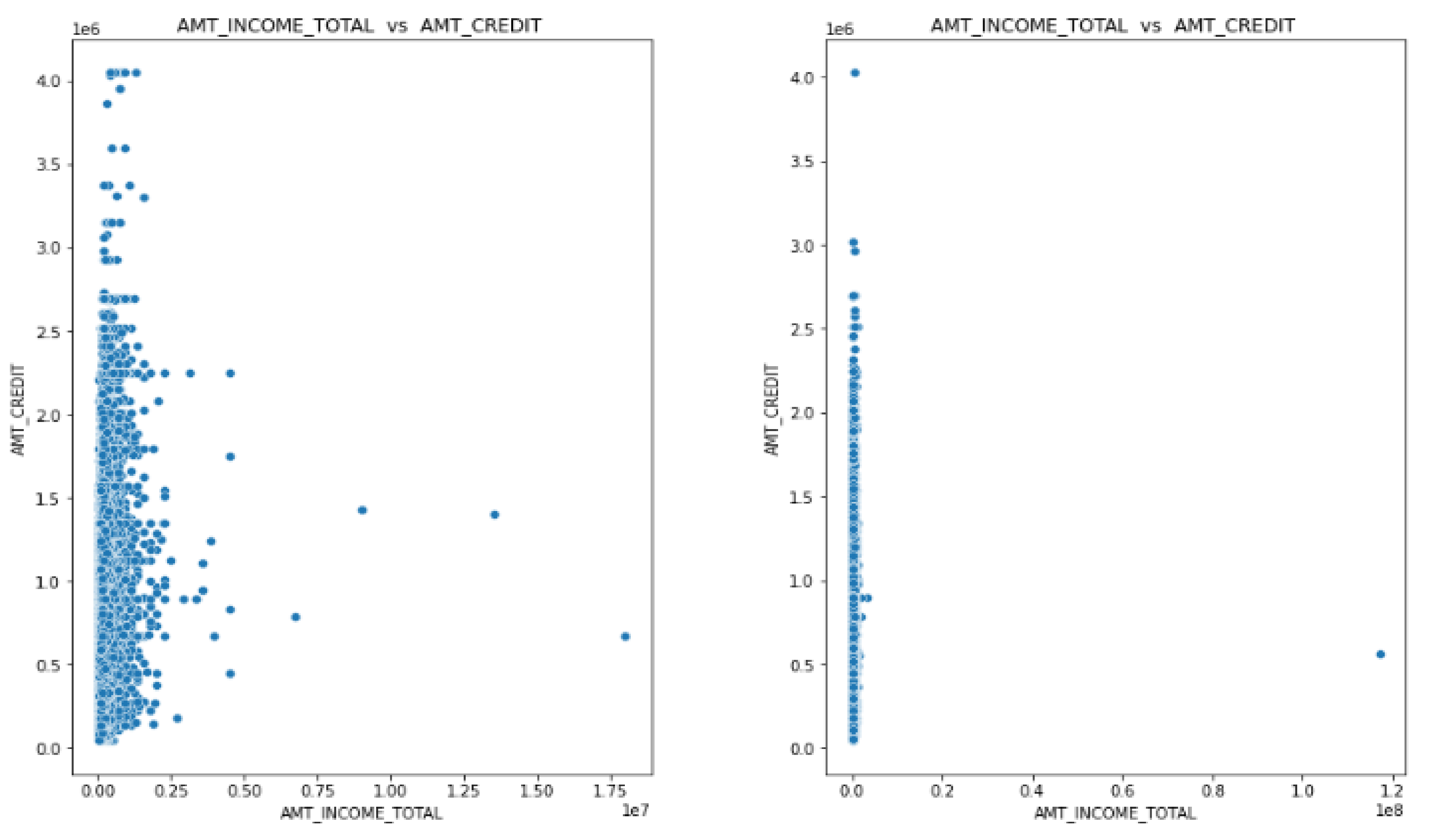
- The storyline demonstrates that an applicant who is given a large credit amount won't experience any problems paying their EMIs.



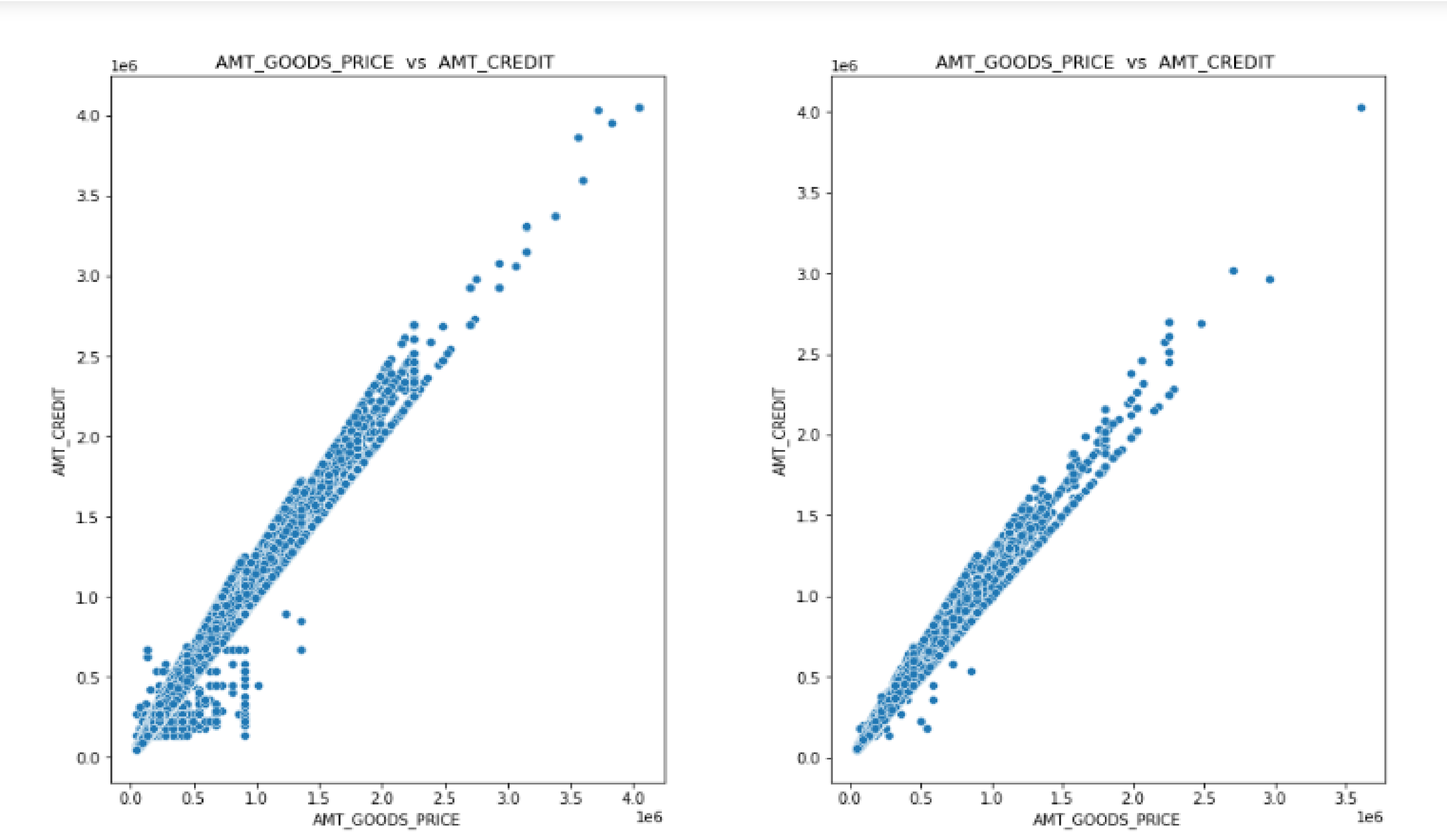
Bivariate Analysis

- In comparison to those who didn't pay or made late payments, individuals who had the loan amount paid on time or in whole were more likely to receive higher credits. In comparison to individuals with higher products prices but unpaid loans, those with higher goods prices and on-time payments have higher credit ratings.

Bivariate Analysis

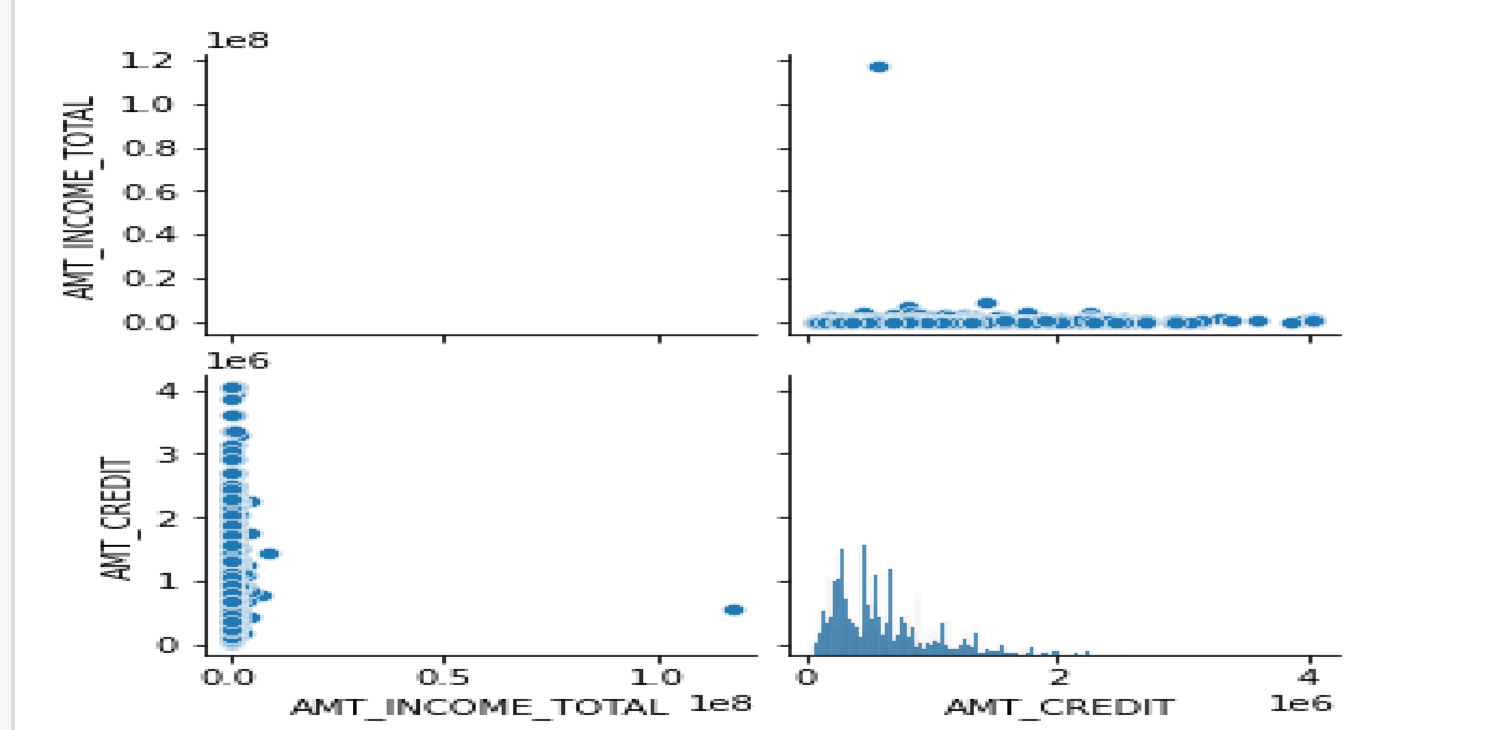


Bivariate Analysis



Bivariate Analysis

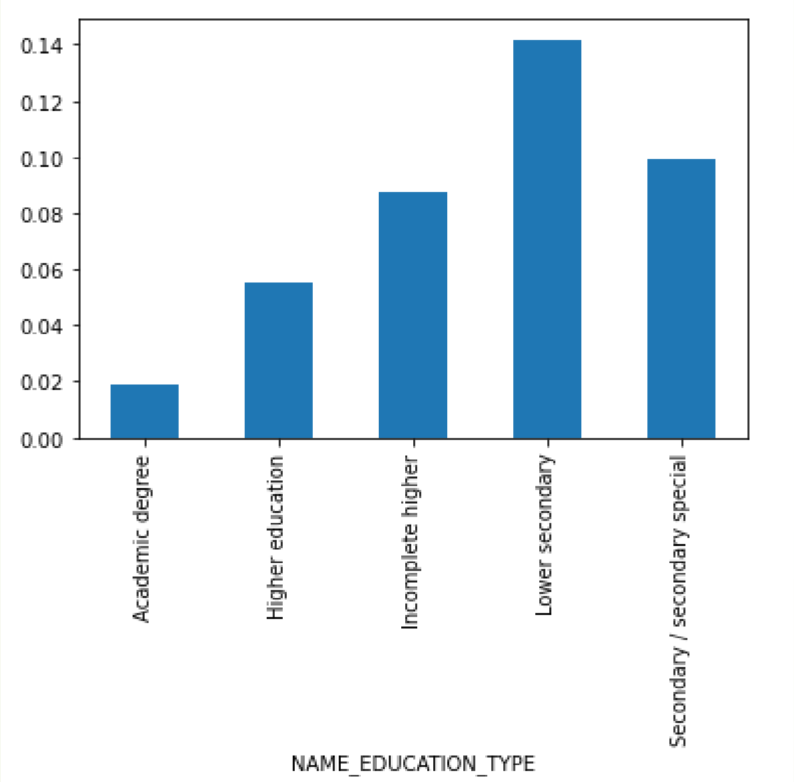
- We can infer from the aforementioned bivariate analysis that AMT CREDIT to applicants who pay loan instalments on time is higher than AMT CREDIT to applicants who do not. Applicants with low Good Price default on loans.

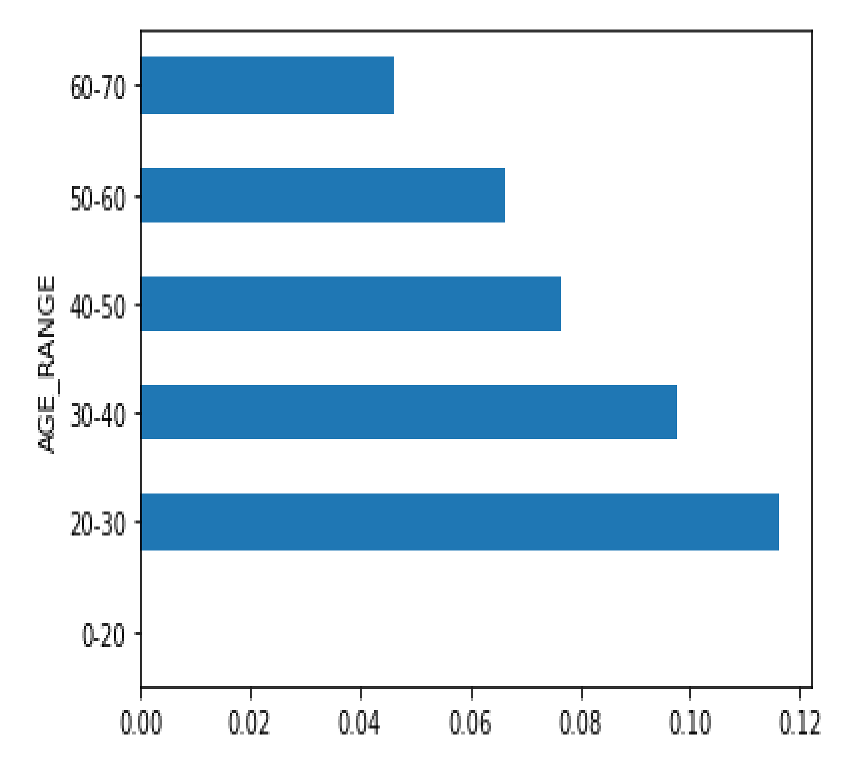


Bivariate Analysis

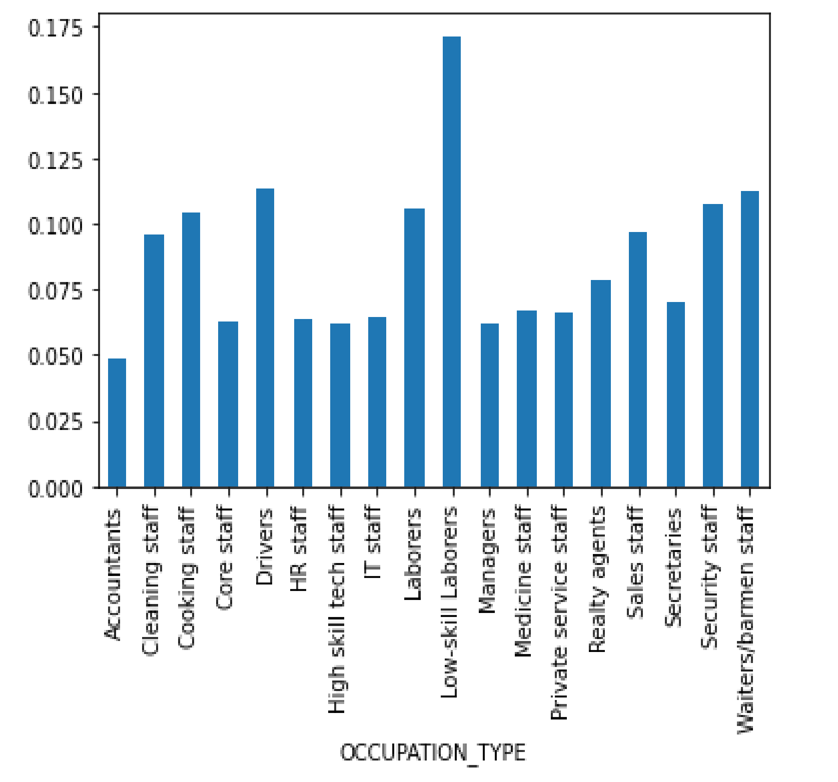
- The accompanying graphic shows that candidates with less formal secondary education are more likely to default on loans than candidates with formal education. applicants between the ages of 20 and 30 and those working in low-skilled labour occupations are more likely to default on loans. The likelihood of defaulting on a loan is higher for applicants with the family status of Single/Not Married & Civil Marriage and those who reside in rented apartments.

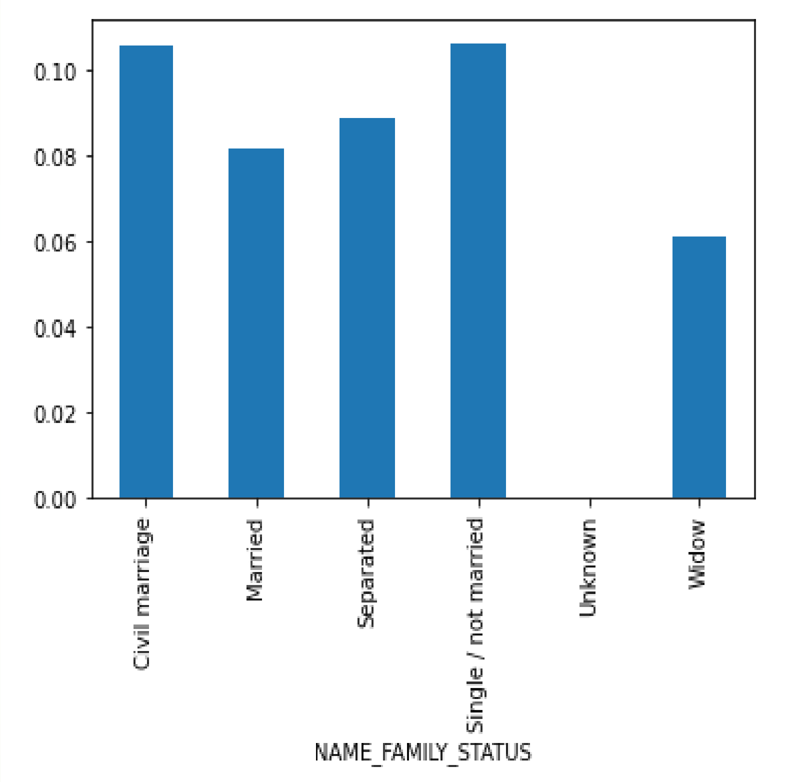
Bivariate Analysis



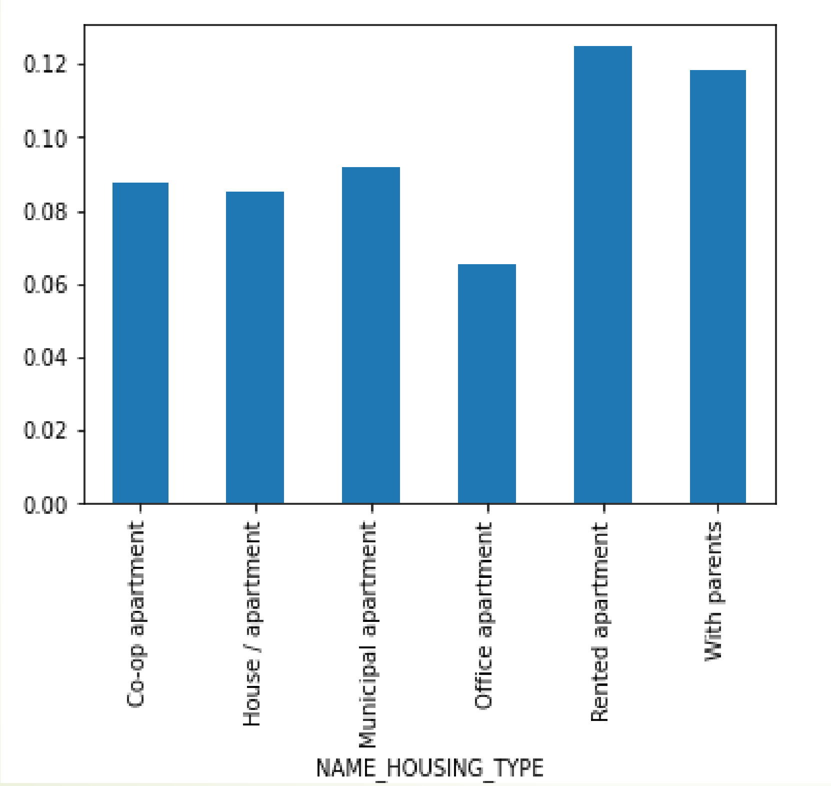


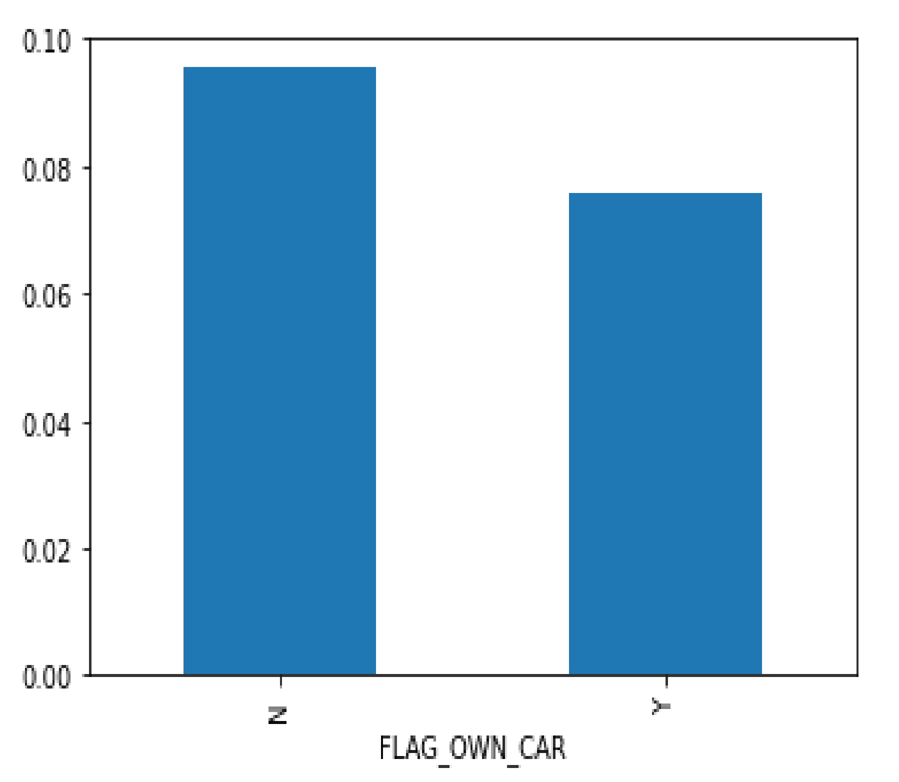
Bivariate Analysis





Bivariate Analysis



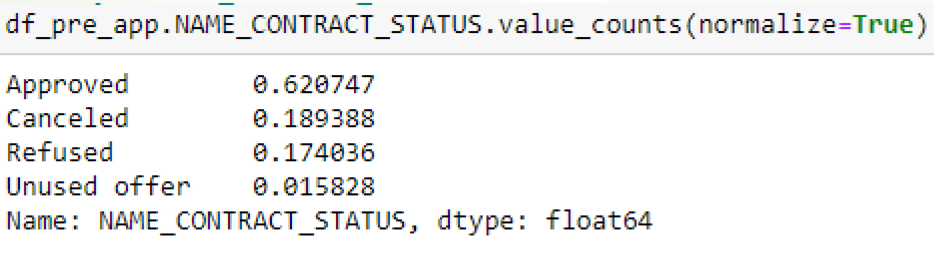


Previous Application Data

- There are missing values in 11 out of the 37 columns, which is more than 40%.
- This dataset's NAME_CONTRACT_STATUS column, which has 4 distinct values, is the target column.

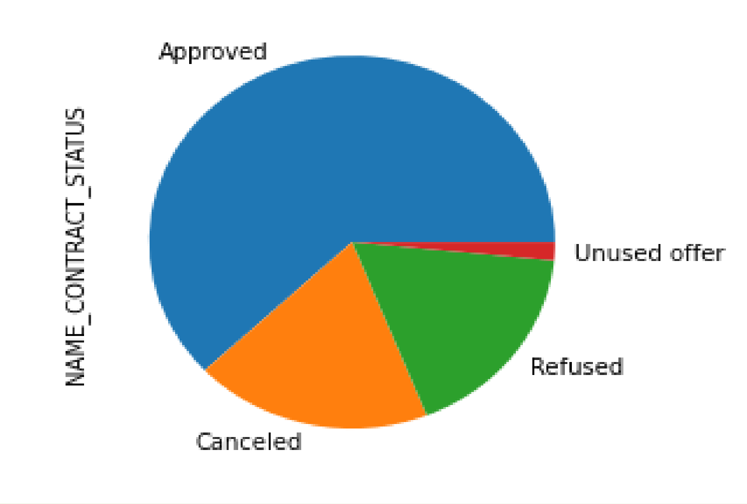
Previous Application Data (Univariant Analysis)

- Distribution of the Target Column



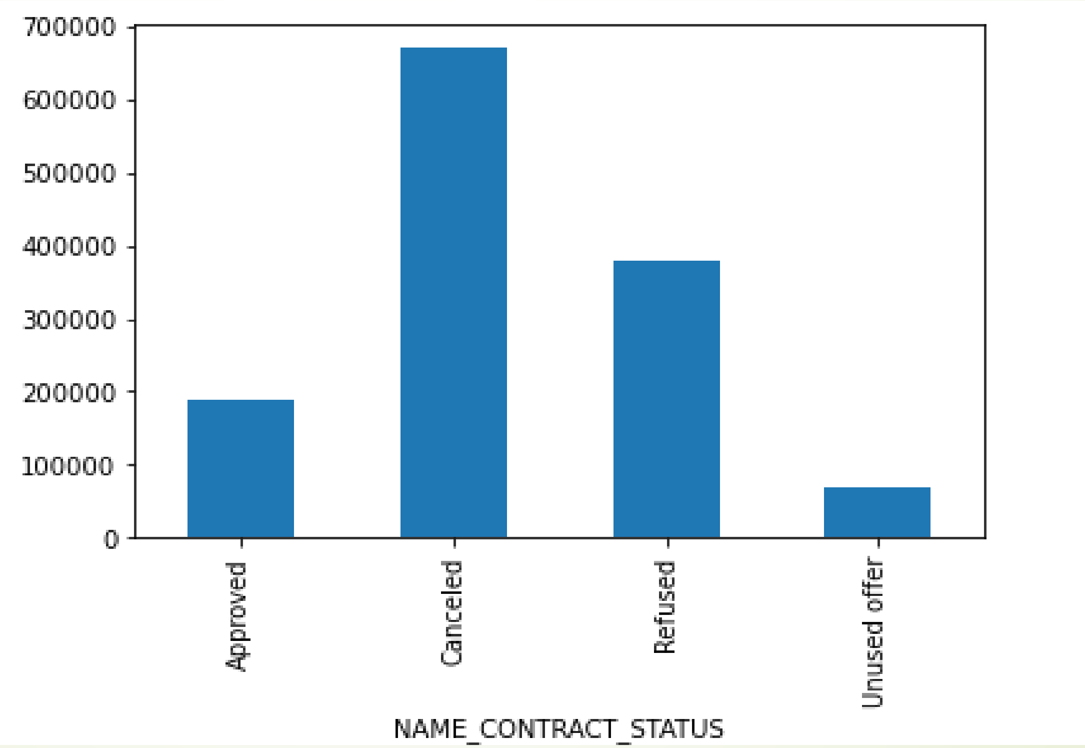
```
df_pre_app.NAME_CONTRACT_STATUS.value_counts(normalize=True)

Approved        0.620747
Canceled        0.189388
Refused         0.174036
Unused offer    0.015828
Name: NAME_CONTRACT_STATUS, dtype: float64
```



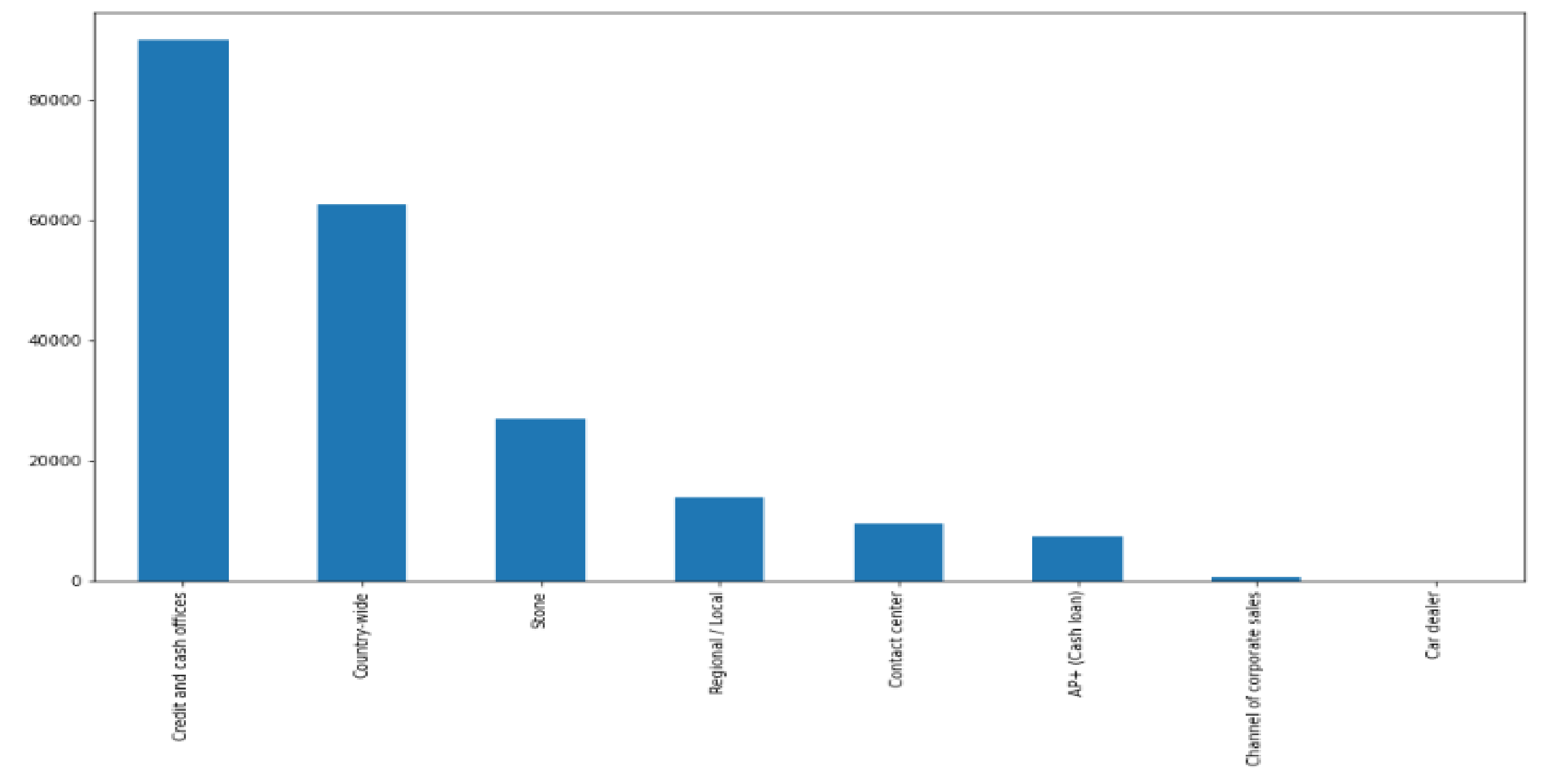
Previous Application Data (Univariant Analysis)

- Goods Price of amount under 200000 are more likely to be approved by the bank.

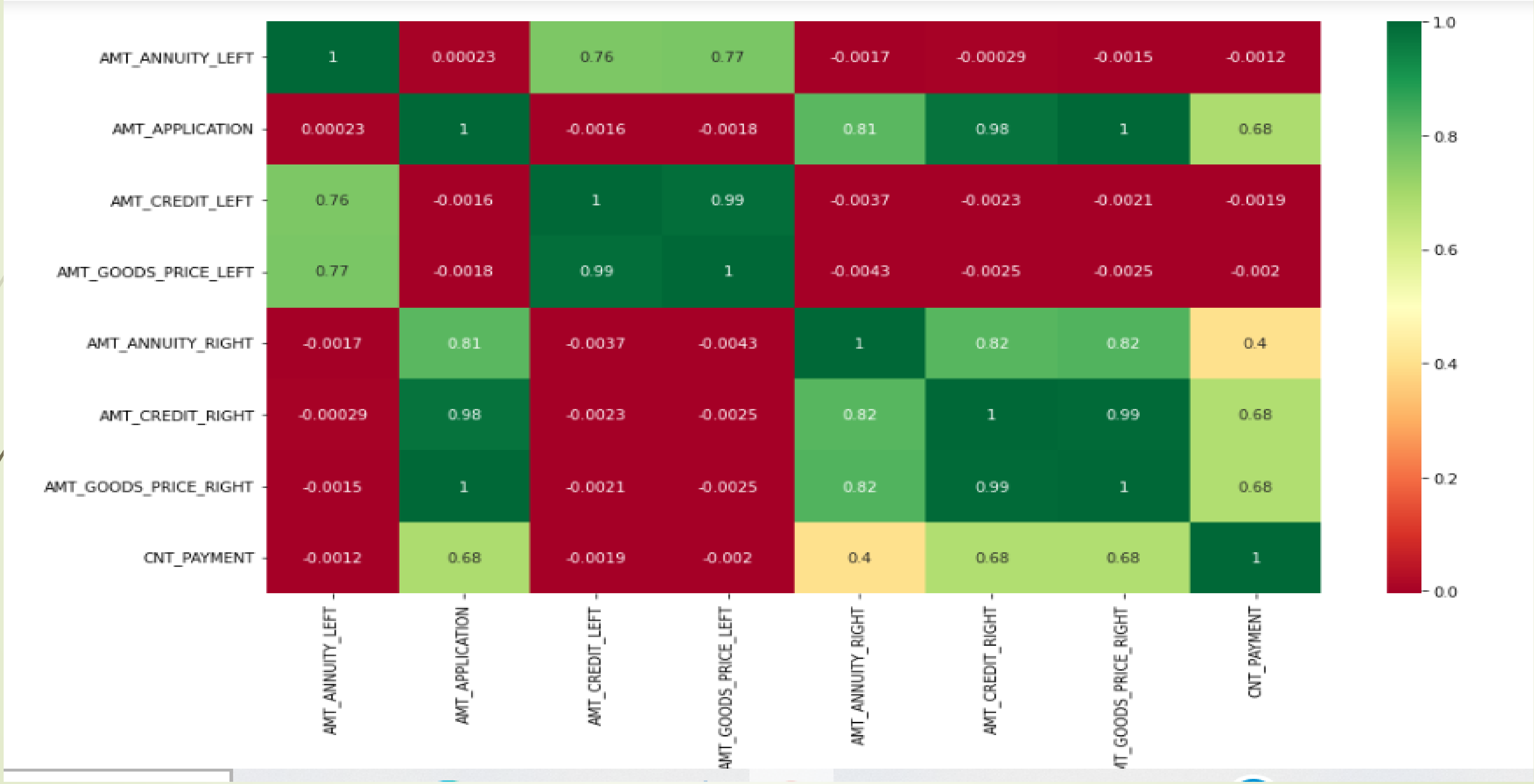


Merge Data (Univariant Analysis)

- Credit and cash offices is the highest Channel Type among all loan applications



Correlation Matrix



	AMT_ANNUITY_LEFT	AMT_APPLICATION	AMT_CREDIT_LEFT	AMT_GOODS_PRICE_LEFT	AMT_ANNUITY_RIGHT	AMT_CREDIT_RIGHT	MT_GOODS_PRICE_RIGHT	CNT_PAYMENT
AMT_ANNUITY_LEFT	1	0.00023	0.76	0.77	-0.0017	-0.00029	-0.0015	-0.0012
AMT_APPLICATION	0.00023	1	-0.0016	-0.0018	0.81	0.98	1	0.68
AMT_CREDIT_LEFT	0.76	-0.0016	1	0.99	-0.0037	-0.0023	-0.0021	-0.0019
AMT_GOODS_PRICE_LEFT	0.77	-0.0018	0.99	1	-0.0043	-0.0025	-0.0025	-0.002
AMT_ANNUITY_RIGHT	-0.0017	0.81	-0.0037	-0.0043	1	0.82	0.82	0.4
AMT_CREDIT_RIGHT	-0.00029	0.98	-0.0023	-0.0025	0.82	1	0.99	0.68
AMT_GOODS_PRICE_RIGHT	-0.0015	1	-0.0021	-0.0025	0.82	0.99	1	0.68
CNT_PAYMENT	-0.0012	0.68	-0.0019	-0.002	0.4	0.68	0.68	1

Correlation Matrix

- AMT_APPLICATION has a high correlation with AMT_ANNUITY_RIGHT,AMT_CREDIT_RIGHT,AMT_GOODS_PRICE_RIGHT and decent correlation with CNT_PAYMENT

- AMT_GOODS_PRICE_RIGHT has a high correlation with AMT_ANNUITY_RIGHT,AMT_CREDIT_RIGHT,AMT_APPLICATION and decent correlation with CNT_PAYMENT

- AMT_CREDIT_RIGHT has a high correlation with AMT_GOODS_PRICE_RIGHT and decent correlation with CNT_PAYMENT

- AMT_ANNUITY_LEFT has a high correlation with AMT_GOODS_PRICE_RIGHT,AMT_CREDIT_RIGHT

- AMT_ANNUITY_LEFT has a high correlation with AMT_GOODS_PRICE_LEFT,AMT_CREDIT_LEFT

- AMT_CREDIT_LEFT has a high correlation with AMT_GOODS_PRICE_LEFT