Credit EDA Case Study

By:

Somake Mehrotra

(DS 43)

What is EDA?

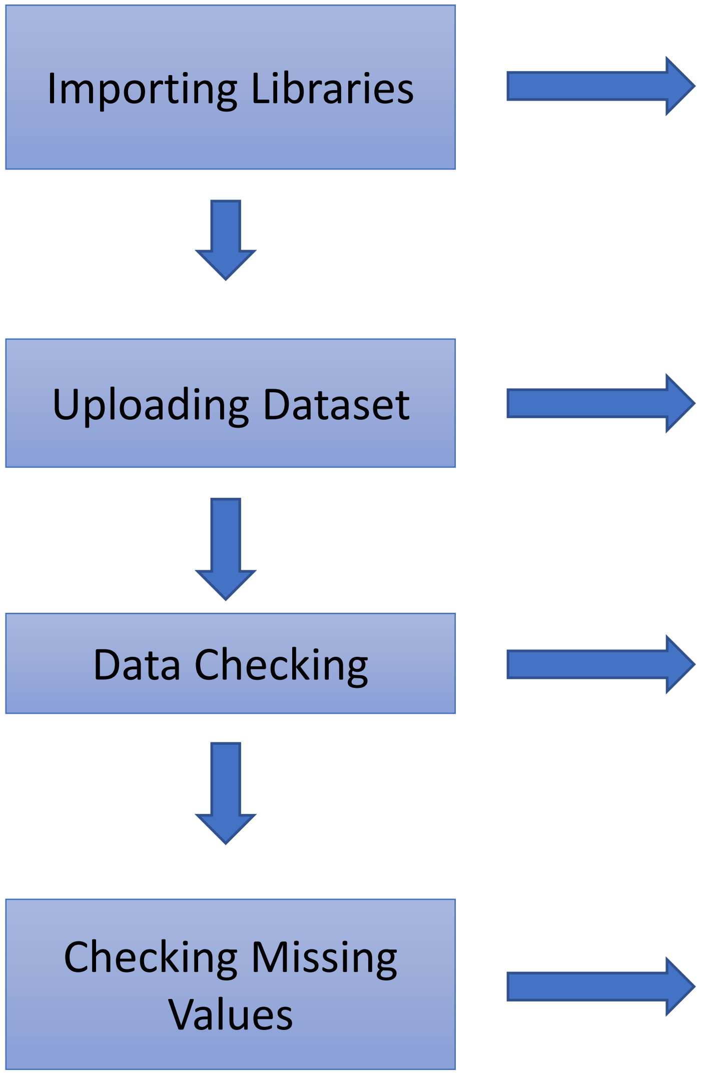
Exploratory Data Analysis (EDA) is a data visualization technique to draw inferences and obtain insights. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

1. Problem Statement

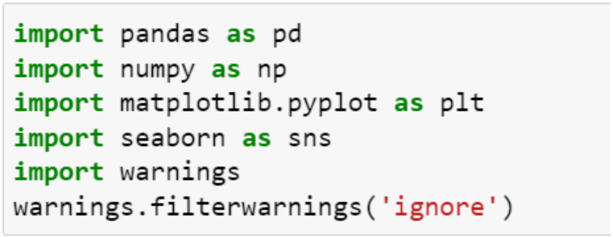
- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

2. Data Mining

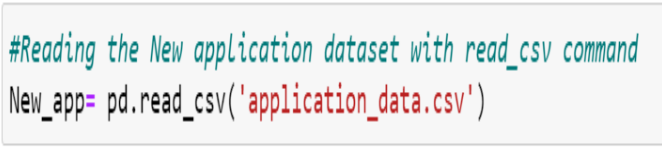


Importing Libraries



```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Uploading Dataset



```python
#Reading the New application dataset with read_csv command
New_app= pd.read_csv('application_data.csv')
```

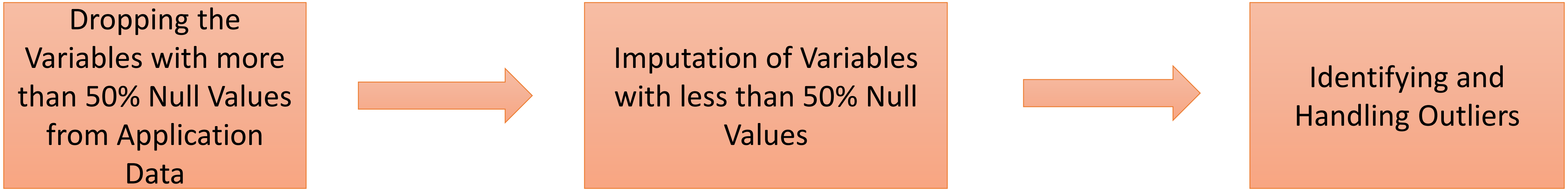
Data Checking

1. New_app.head()
2. New_app.shape
3. New_app.info()

Checking Missing Values

- New_app.isnull().sum()

3. Data Cleaning



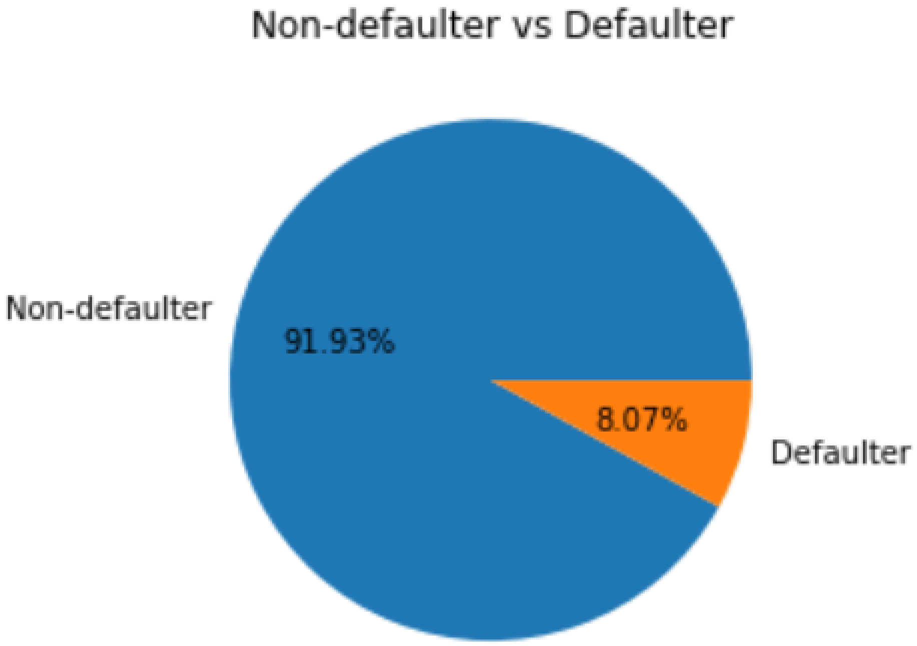
Dropping the Variables with more than 50% Null Values from Application Data

Imputation of Variables with less than 50% Null Values

Identifying and Handling Outliers

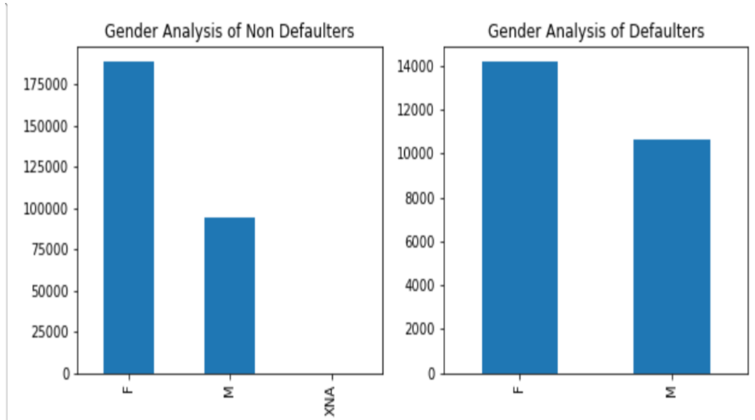
4. Data Exploration

Checking the Imbalance in Dataset

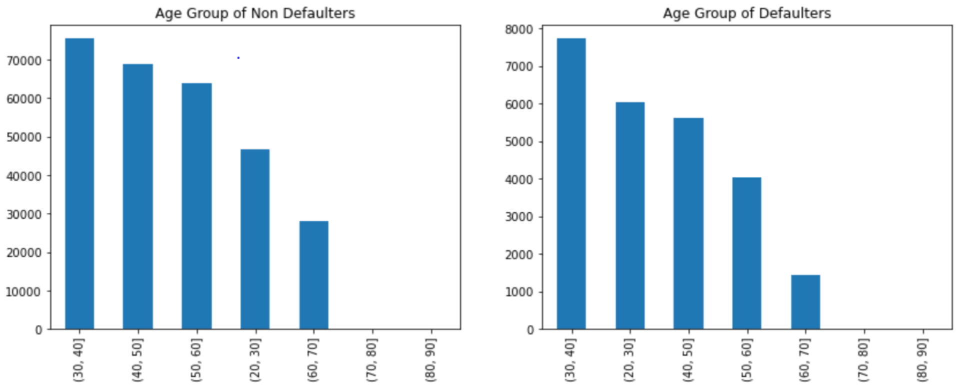


From the Pie chart we can see that the dataset is highly imbalanced between Non-Defaulters and Defaulters. So, we need to divide the dataset and then perform the analysis.

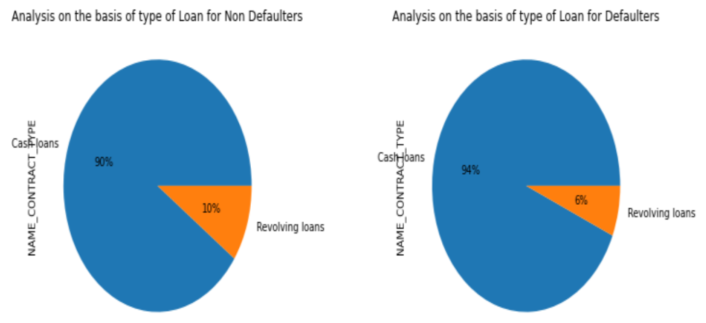
5. Univariate Analysis for Categorical Variables



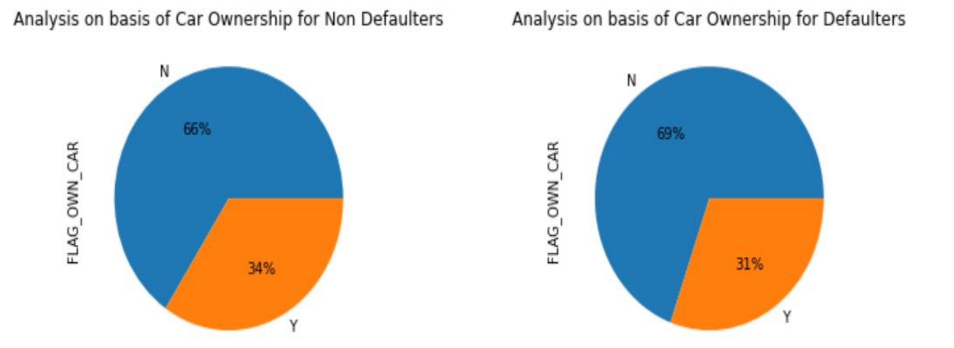
CODE_GENDER – Count of Female applicants is higher for both Defaulters and Non-Defaulters



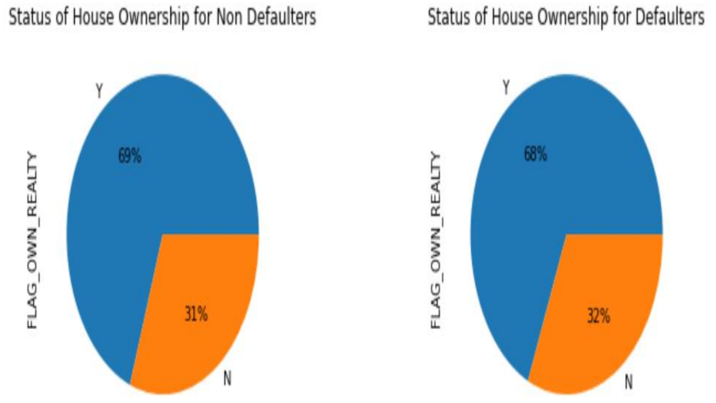
Age_Group - Major chunk of loan applicants belongs to age group of 30 to 40.



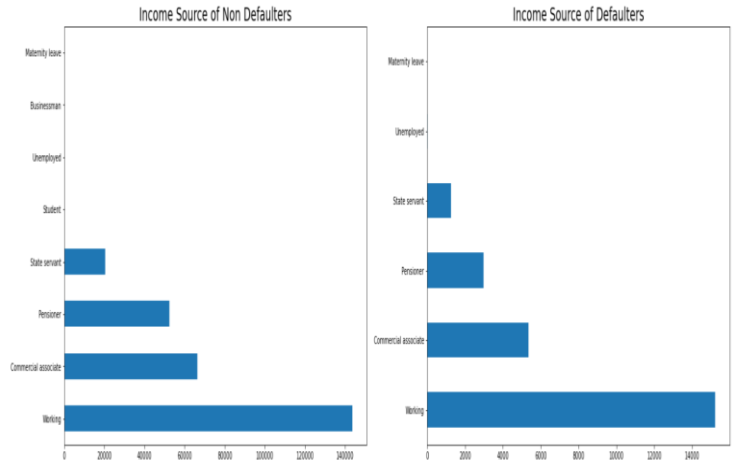
Name_Contract_Type – 90% of application is for cash loan



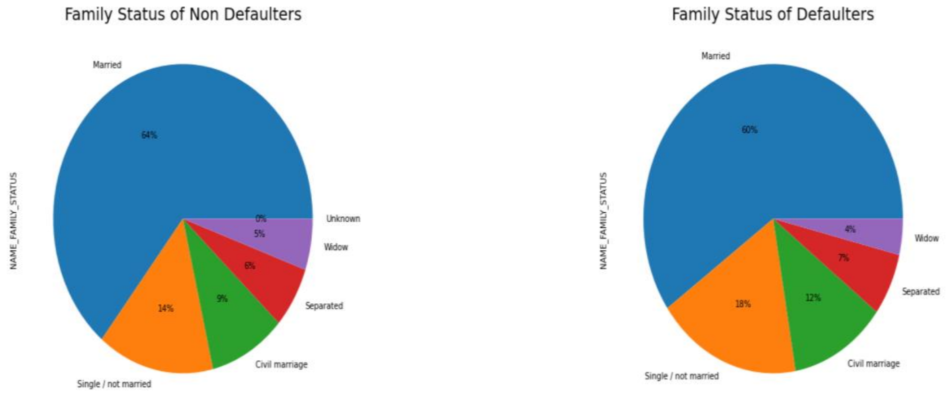
FLAG_OWN_CAR – More than 50% of loan applicants don't posses their own car.



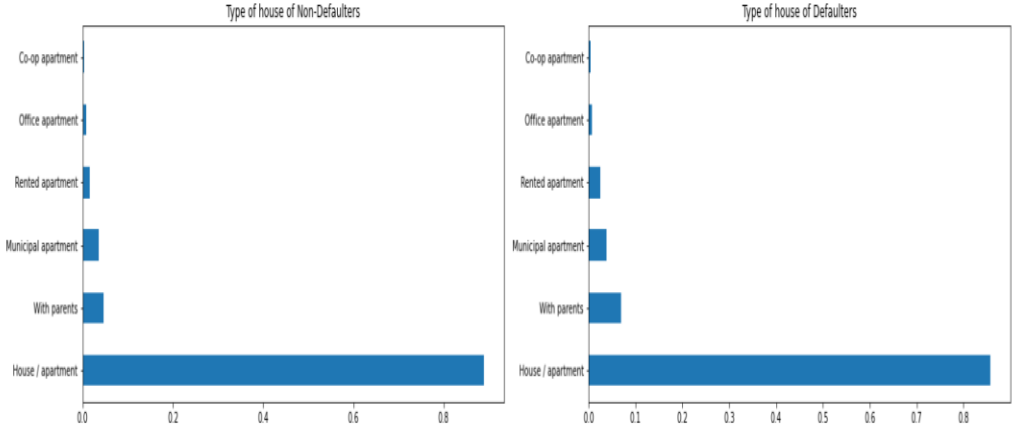
FLAG_OWN_REALTY – On Contrary, More than 50% applicants have their own house.



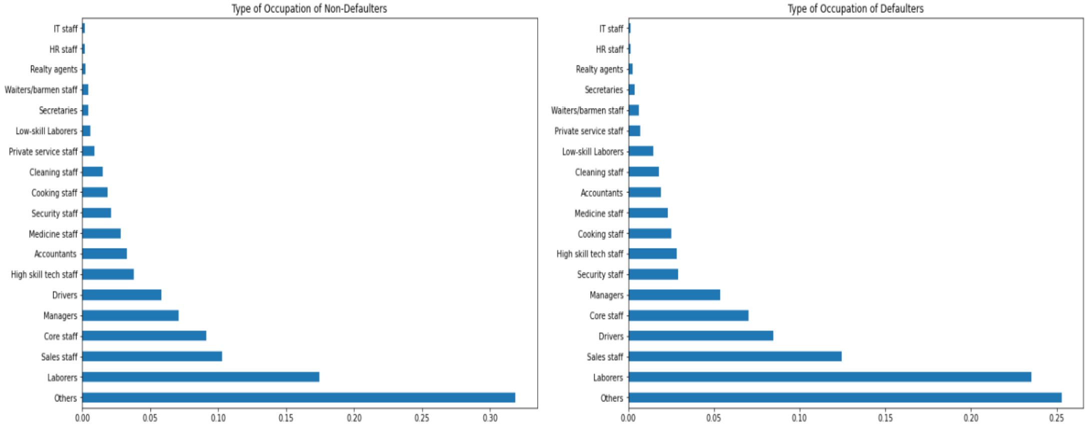
NAME_INCOME_TYPE – The Highest number of loan applicants are working class



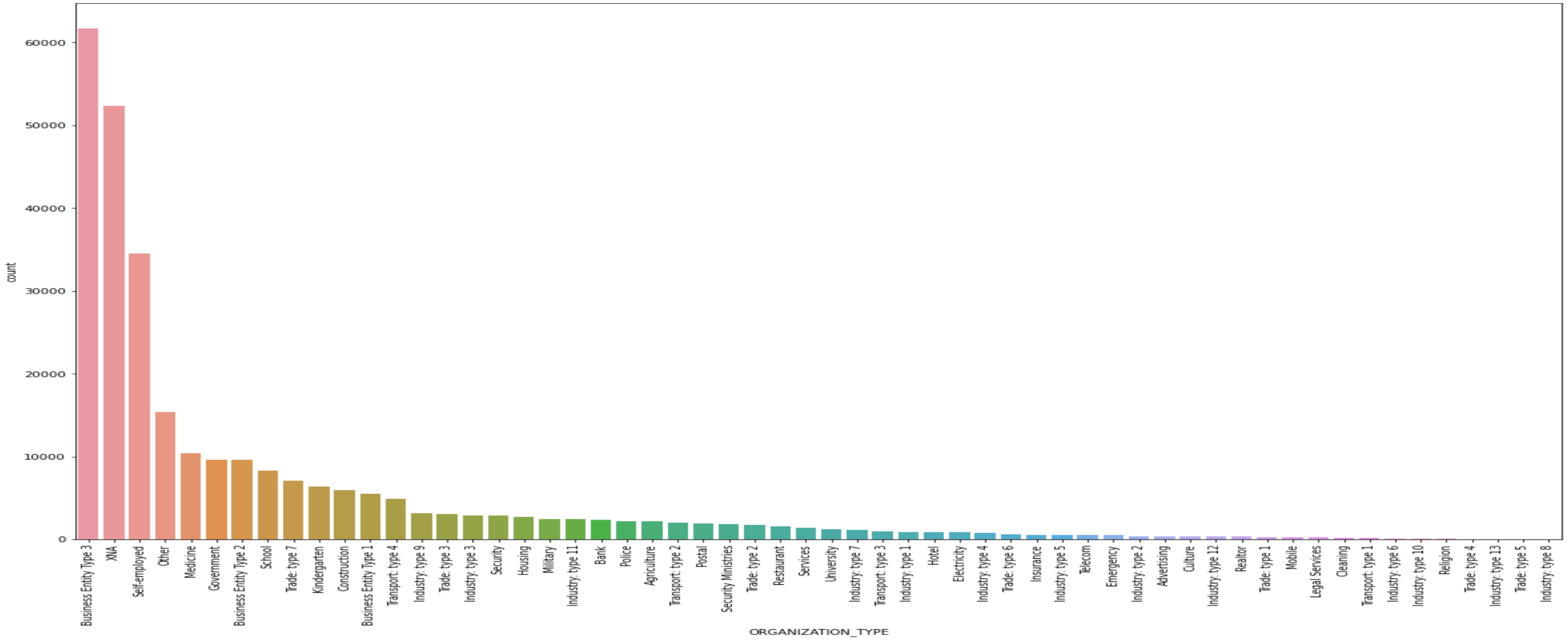
NAME_FAMILY_STATUS – The percentage of Married loan applicants is highest for both Defaulters and Non-Defaulters.

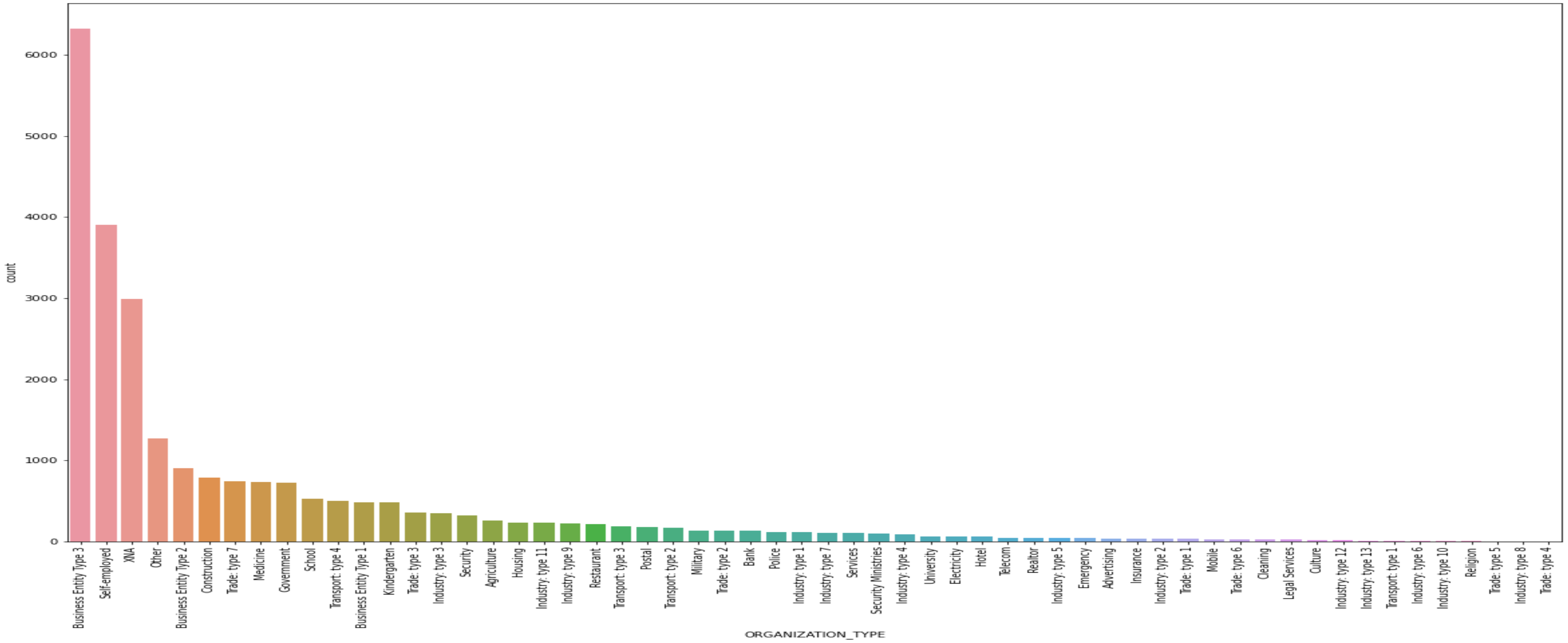


NAME_HOUSING_TYPE – Majority of loan applicants resides in their own house or apartment.



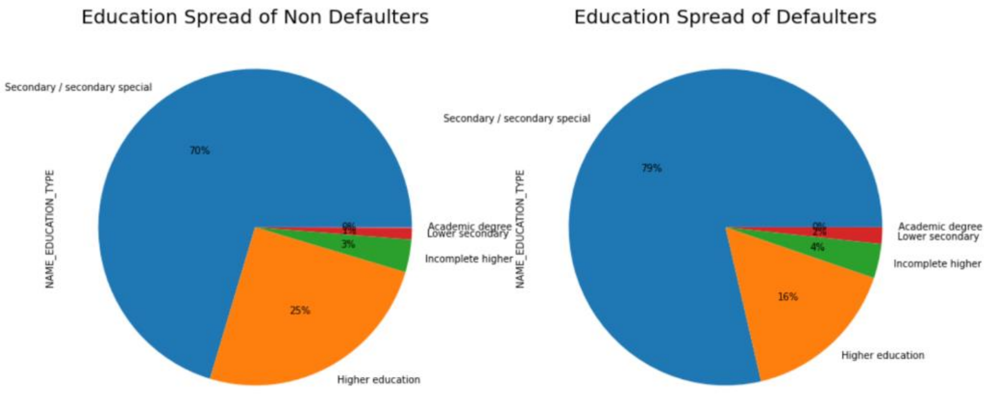
OCCUPATION_TYPE – The top 5 Occupation categories are:- Others, Laborers, Sales staff, Core staff, Managers.



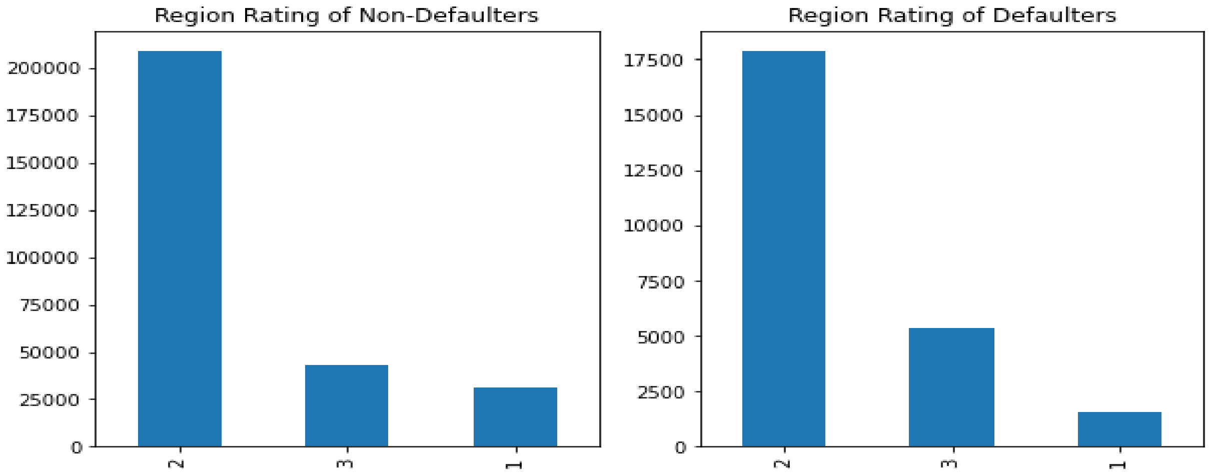


Organization Type – The Highest number of loan applicants belongs to Business Entity.

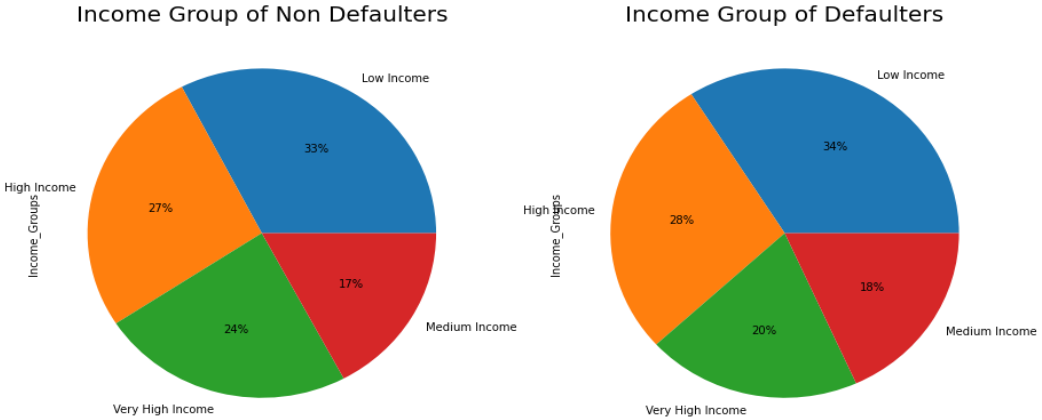
6. Univariate Categorical Ordered Analysis



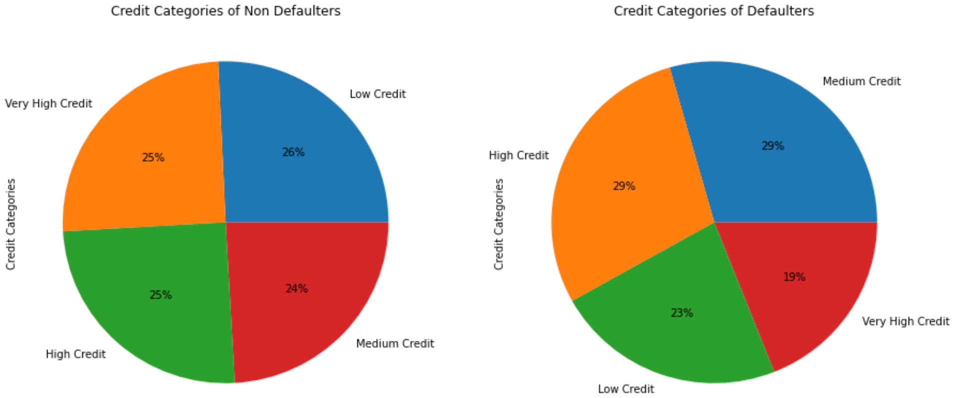
NAME_EDUCATION_TYPE – 70% and above loan applicants have attained secondary level of education.



REGION_RATING – The highest count of applicants belongs to region rating 2.

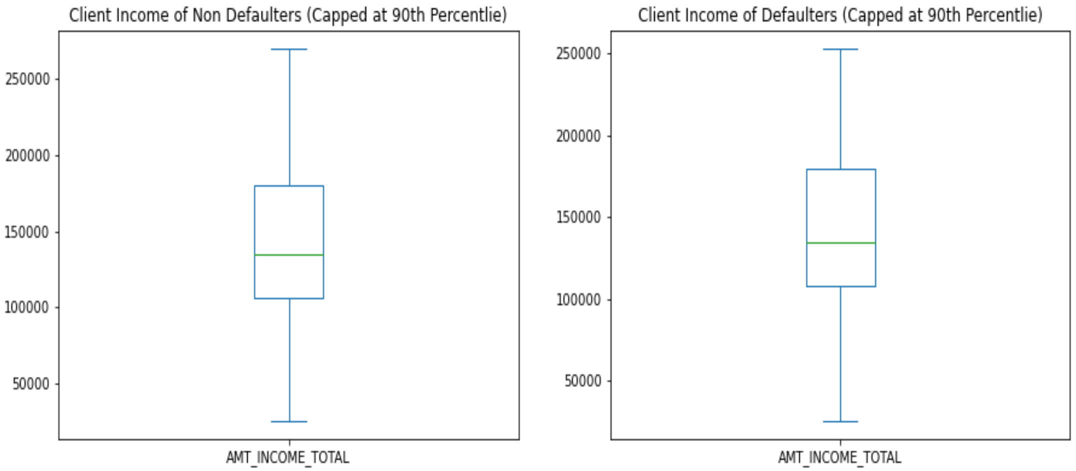


INCOME_GROUPS **-** The maximum percentage of loan applicants pertains to low income level followed by high income level.

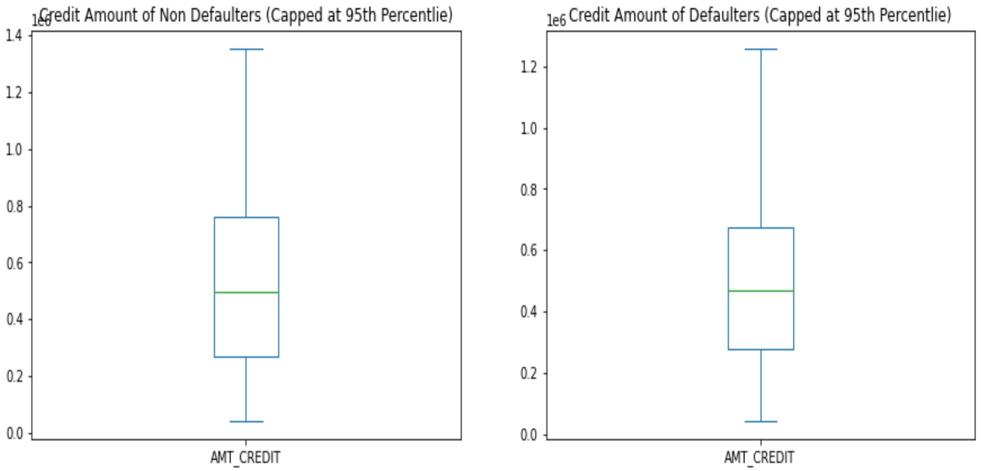


Credit Categories- Non-Defaulters has applied for low credit whereas for Defaulters the percentage of loan with Medium credit is highest.

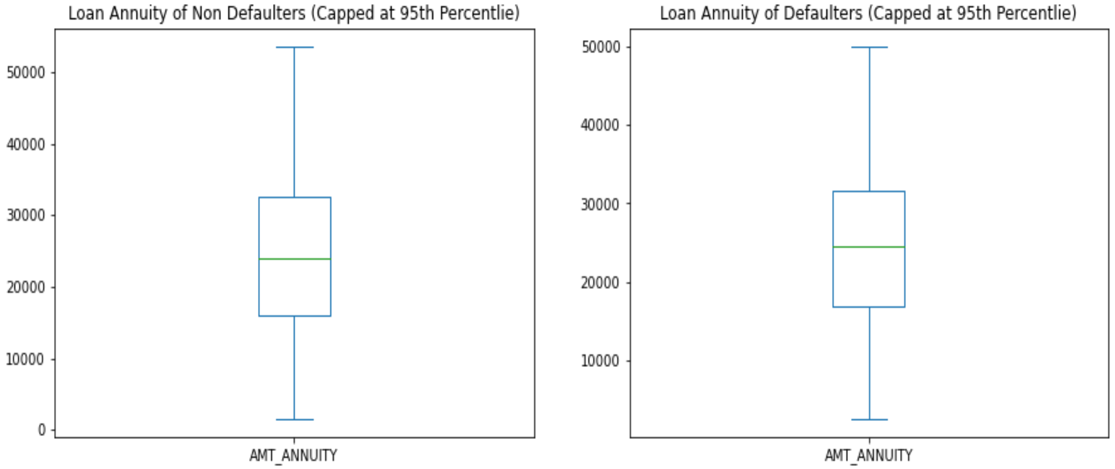
7. Univariate Analysis of Continuous Variables



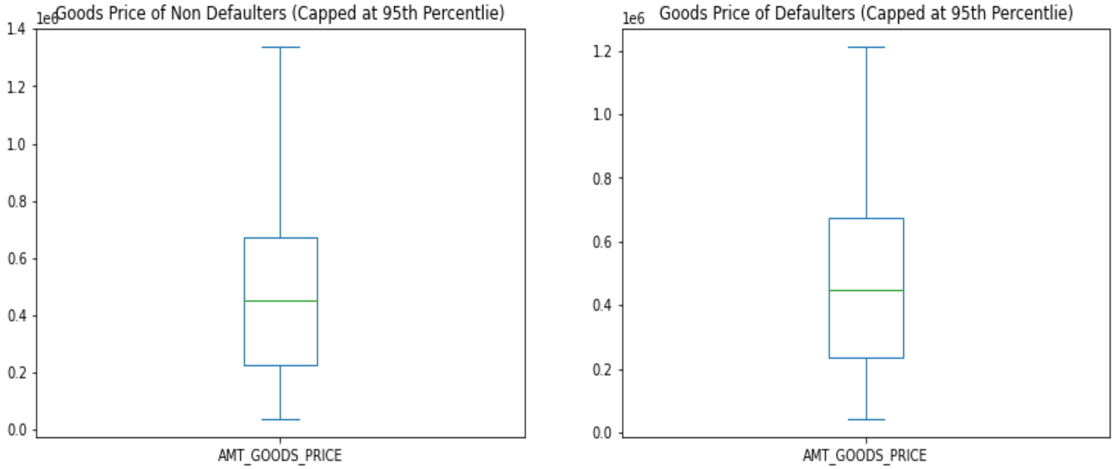
AMT_INCOME_TOTAL – The income level has been capped at 90th percentile to remove outliers and from box plot we can see that the IQR and median is almost same for both categories.



AMT_CREDIT- Credit level limited to 95th percentile and box plot shows that Median is almost same for both categories.

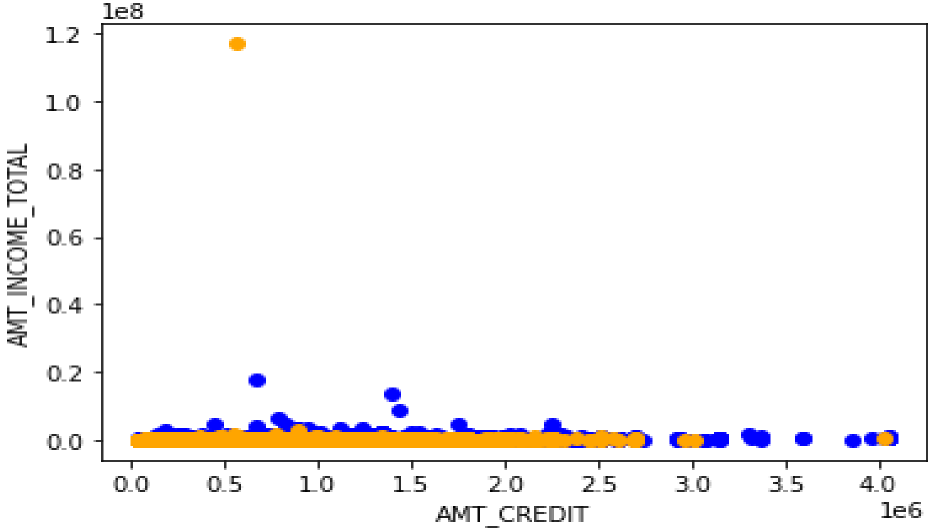


AMT_ANNUITY – Loan annuity level limited to 95th percentile and IQR and median is almost same for both.

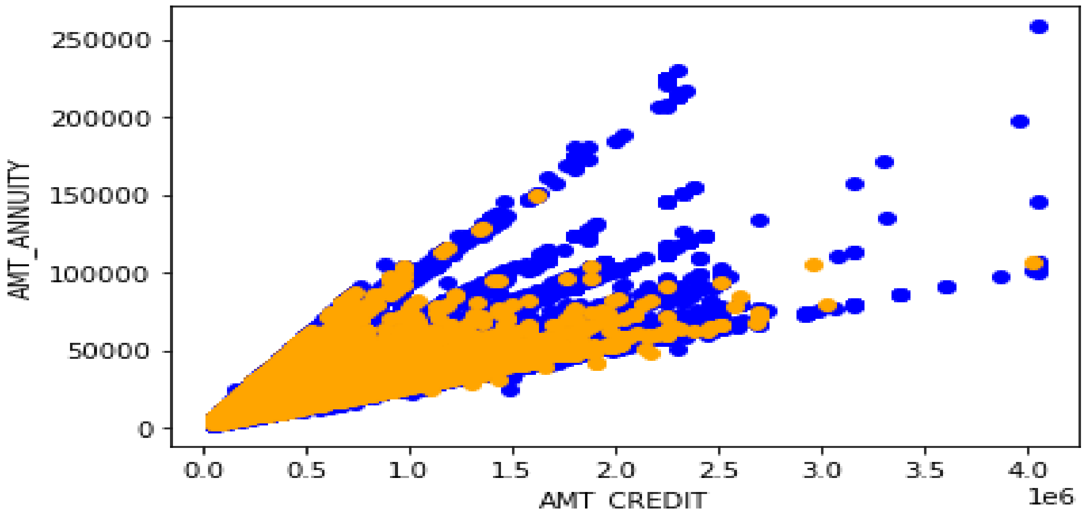


AMT_GOODS_PRICE – The Median for Goods price is exactly same for both categories however the goods price of Non-Defaulters is slightly higher than Defaulters.

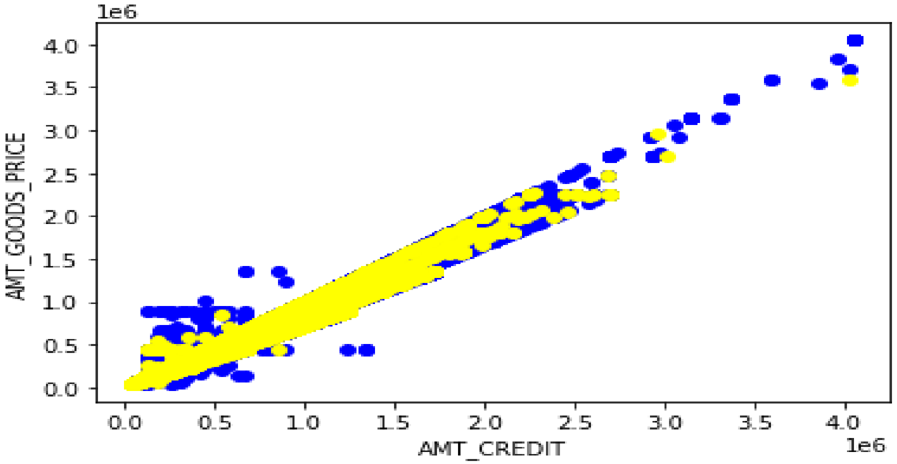
8. Bivariate Analysis of Continuous Variables.



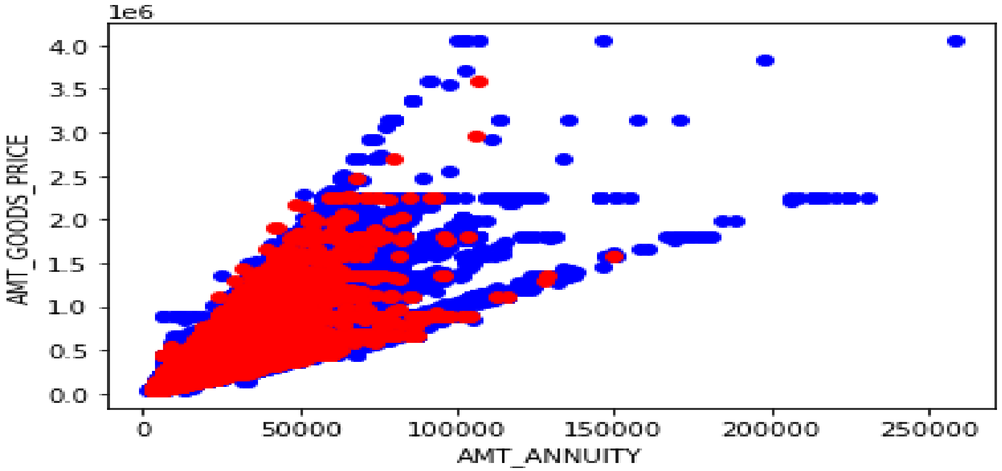
AMT_CREDIT & AMT_INCOME_TOTAL – Very low correlation is observed.



AMT_CREDIT & AMT_ANNUITY – There is a High positive correlation between credit amount and loan annuity.

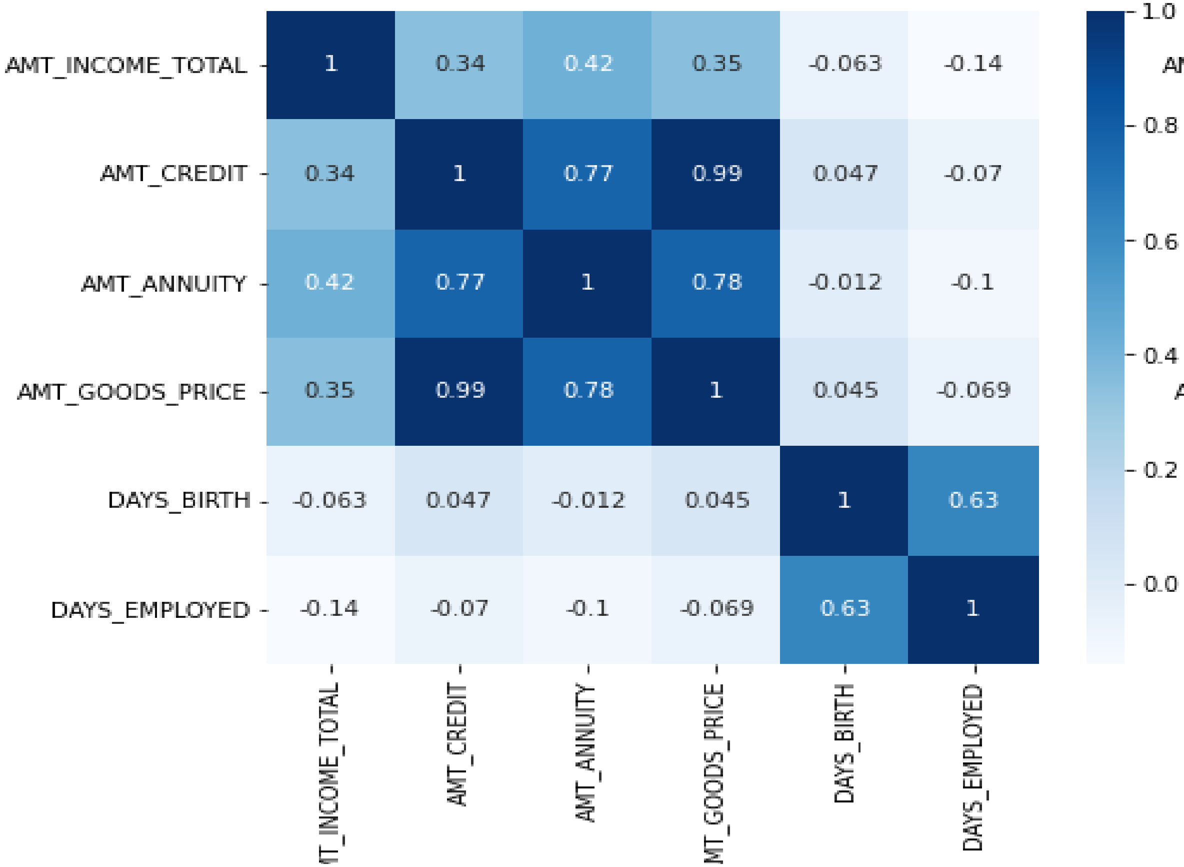


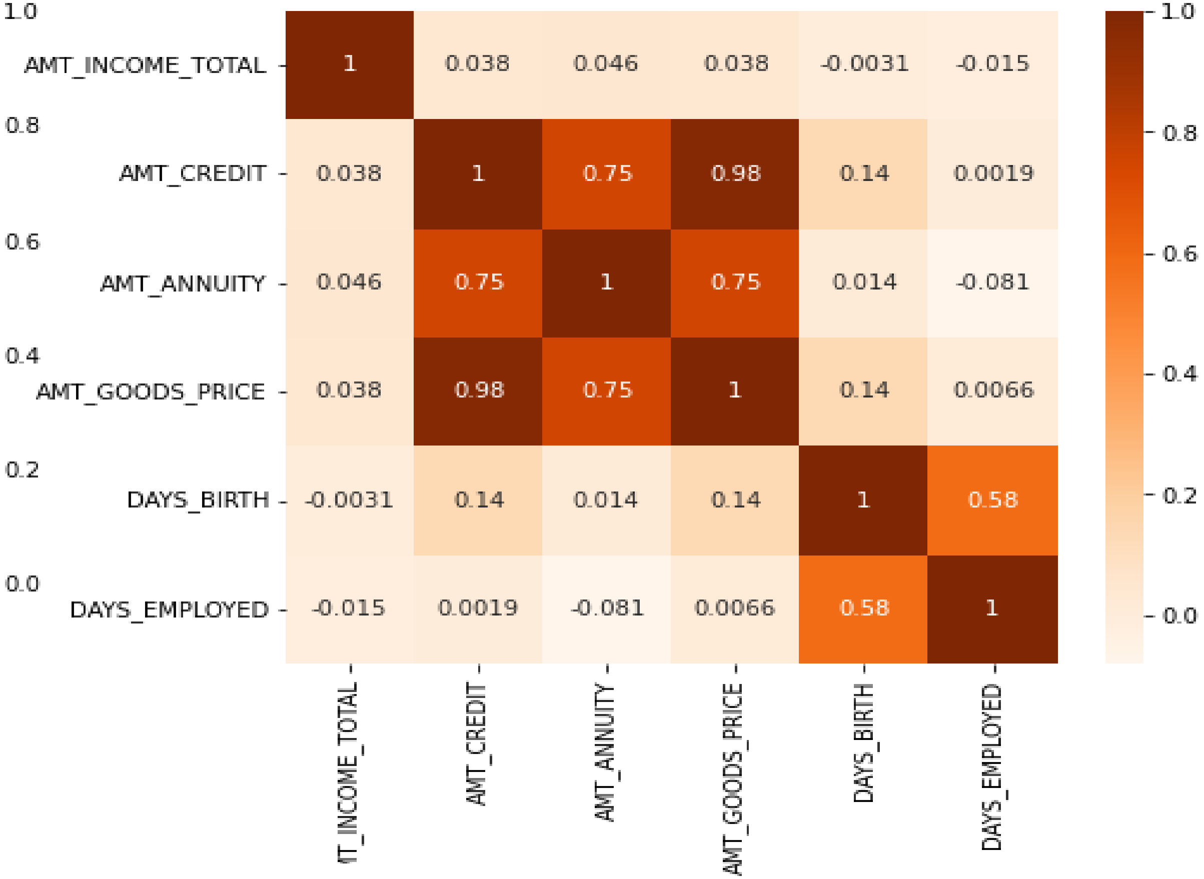
AMT_CREDIT & AMT_GOODS_PRICE- Very High Correlation is observed between Credit amount and amount of Goods price for which loan is availed.



AMT_ANNUITY & AMT_GOODS_PRICE- Positive and high correlation is observed between Annuity amount and Goods Price.

9. Correlation of All Numeric Variables Using Heatmap

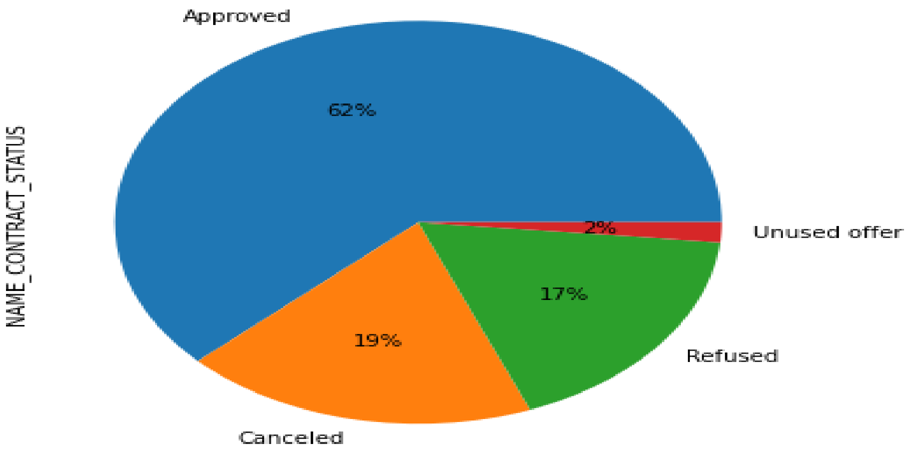

Correlation for Non-Defaulters


Correlation for Defaulters

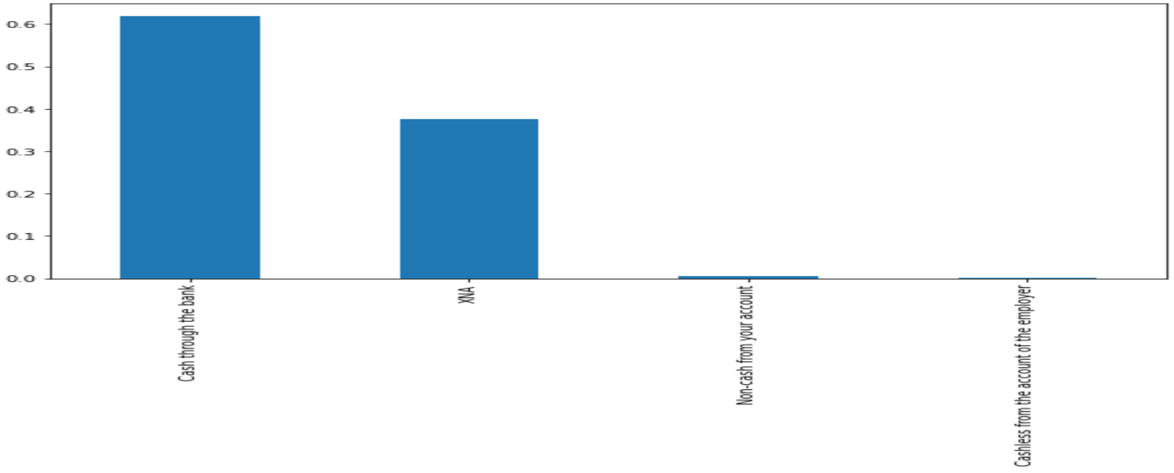
The Highest Correlating Variables for both the dataset are same as:-

- AMT_CREDIT & AMT_GOODS_PRICE
- AMT_ANNUITY & AMT_GOODS_PRICE
- AMT_ CREDIT & AMT_ANNUITY

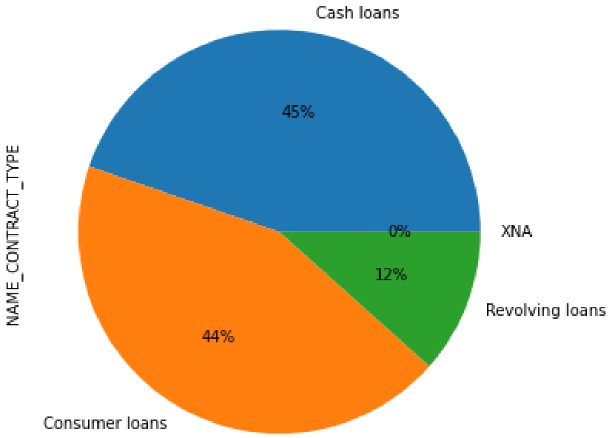
10. Univariate Analysis of Previous Application Variables



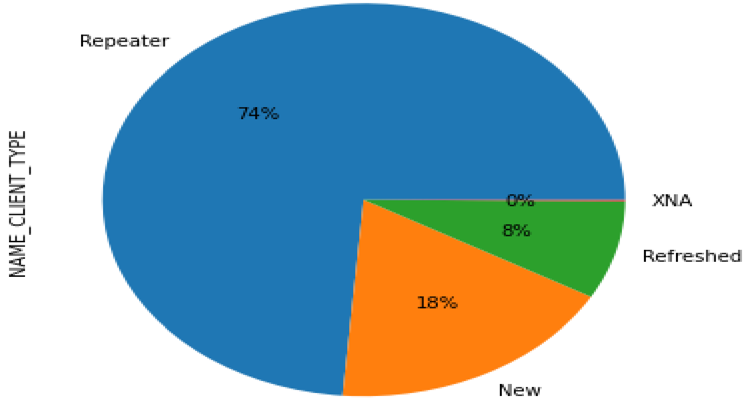
Contract_Status – More than 60% of loan applications were approved previously.



PAYMENT_TYPE – Most preferred payment mode is Cash through Bank.

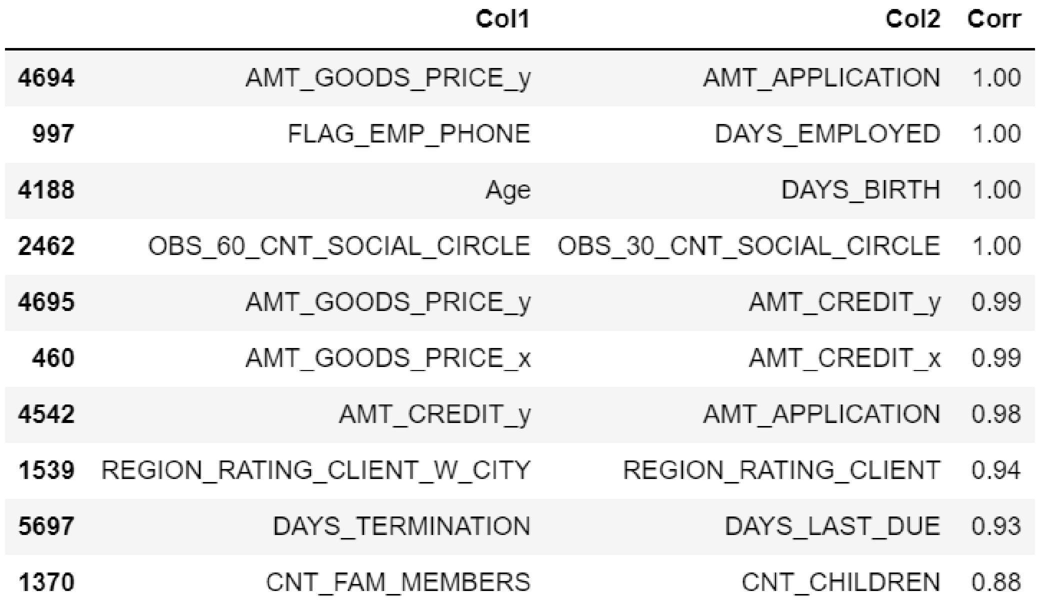


CONTRACT_TYPE - As similar to New application dataset , the major category of loan type is Cash Loan. But we can see that previously Consumer loans were also highly applied.

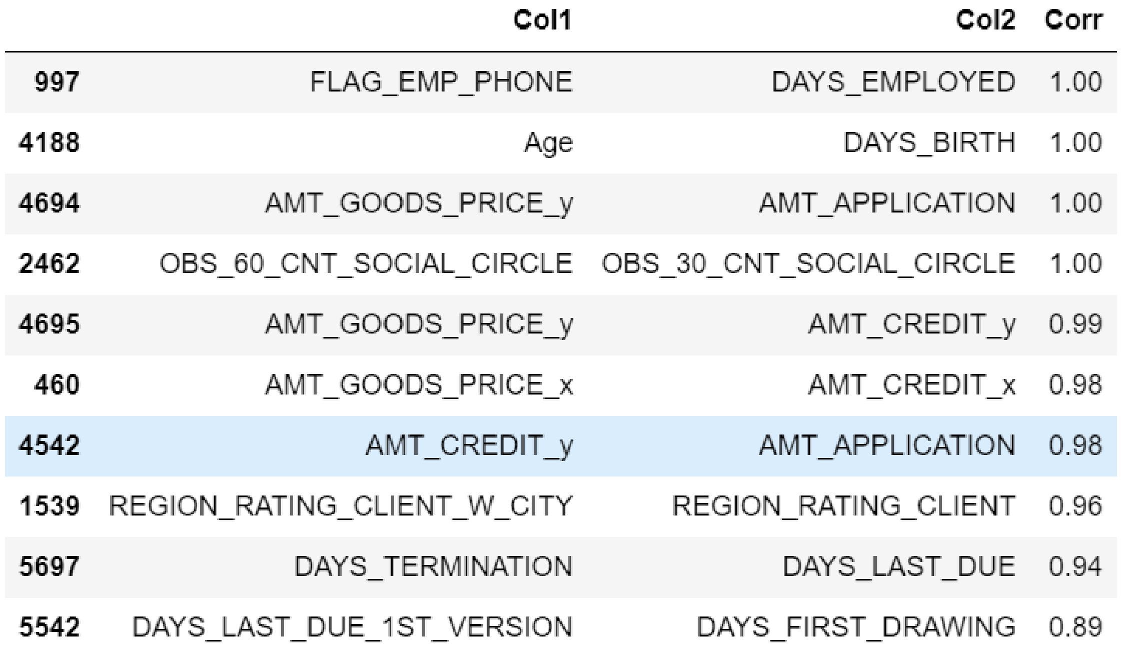


CLIENT_TYPE – More than 70% of clients were Repeaters.

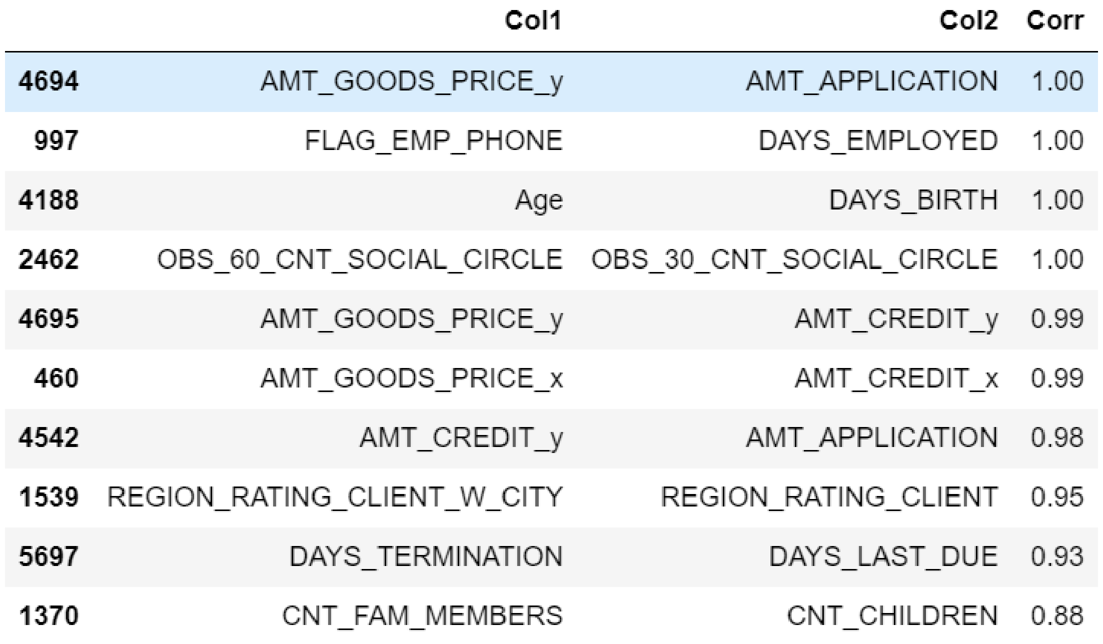
11. Correlation of Numeric Variables for Combined Dataset



	Col1	Col2	Corr
4694	AMT_GOODS_PRICE_y	AMT_APPLICATION	1.00
997	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
4188	Age	DAYS_BIRTH	1.00
2462	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
4695	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.99
460	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.99
4542	AMT_CREDIT_y	AMT_APPLICATION	0.98
1539	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.94
5697	DAYS_TERMINATION	DAYS_LAST_DUE	0.93
1370	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88



	Col1	Col2	Corr
997	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
4188	Age	DAYS_BIRTH	1.00
4694	AMT_GOODS_PRICE_y	AMT_APPLICATION	1.00
2462	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
4695	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.99
460	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.98
4542	AMT_CREDIT_y	AMT_APPLICATION	0.98
1539	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96
5697	DAYS_TERMINATION	DAYS_LAST_DUE	0.94
5542	DAYS_LAST_DUE_1ST_VERSION	DAYS_FIRST_DRAWING	0.89



	Col1	Col2	Corr
4694	AMT_GOODS_PRICE_y	AMT_APPLICATION	1.00
997	FLAG_EMP_PHONE	DAYS_EMPLOYED	1.00
4188	Age	DAYS_BIRTH	1.00
2462	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
4695	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.99
460	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.99
4542	AMT_CREDIT_y	AMT_APPLICATION	0.98
1539	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
5697	DAYS_TERMINATION	DAYS_LAST_DUE	0.93
1370	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88

Recommendations/Conclusion

- In Univariate Categorical Variables Analysis, we can notice that the pattern for both Defaulters and Non-Defaulters is same as there is no major difference in the characteristics of both.
- More focus should be on Cash loan as we can clearly see from both data set the percentage of Cash loan applicants are much higher in comparison to other loan types.
- The data reflects that number of female applicants are more than the male applicants. So, strategies should be designed in such manner that it becomes easier and more attractive for women to avail loans.
- People belonging to Business class tend to apply for more loans and give better yield of interest.
- High percentage of loan applicants have attained secondary level of education.
- Married couple are more willing to apply for loans whereas widows and separated are very less interest which shows that there is a high chance they can default.
- Individuals under 40 are more interested in loan than higher age backet. So, major chunk of loan should be provided to younger age category as people with higher age are more likely to face difficulty in repayments and may default.
- More than 50% of loan applicants don’t posses their own car. On Contrary, more than half have their own house. So, there is no significant pattern basis which we can decide the approval or rejection of loan.
- Loan availability should become easier and more attractive for people with Low-income group as from our data we can notice that percentage of Low-income individuals is highest. However company should be more cautious and vigil while providing loans to such group.