Credit EDA Case Study

Submitted By

Hariprasad Nomula

Ritij Srivastava

Batch-DSC-32May batch

Problem Statement

- Objective is to analysis the given data set and find out which customer is best suited to avail loan.
- There are two risk concerning while taking decision:-

    1)If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

    2)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.

Data Available

- Data about the client with payment difficulties.
- Data about the previous application.

Analysis approach

Following steps were carried out during analysis.

1. Loading the data-set using pandas.
2. Understanding the data.
3. Inspecting structure of the data.
4. Gathering information about various columns.
5. Quality inspection.

- Inspecting the data for any missing values.
- Finding the columns which have missing values more then 50%.

- Inspecting the columns which has missing values less then 13%.
- Check the data-type of columns and modify the column with appropriate data-type.
- We can impute the data for remaining missing values.
- For numerical column use mean median and others.
- For categorical we can use mode to imputing the value.
- For numerical column we have check for any outliers and decide appropriate method to deal with outliers.
- Etraction the data from certain column and creating a new column e.g with datetime column we extract date.
- Creating bins for continuous variables for further analysis of.
- Check the data imbalance based on variable target.

Analysis approach

Based on Target variable column in dataset it can be split into two.

1) Target 0 (non defaulters)
2) Target 1 (defaulters)

Analysis is carried out on each of this data frame by: -

-Univariate Analysis

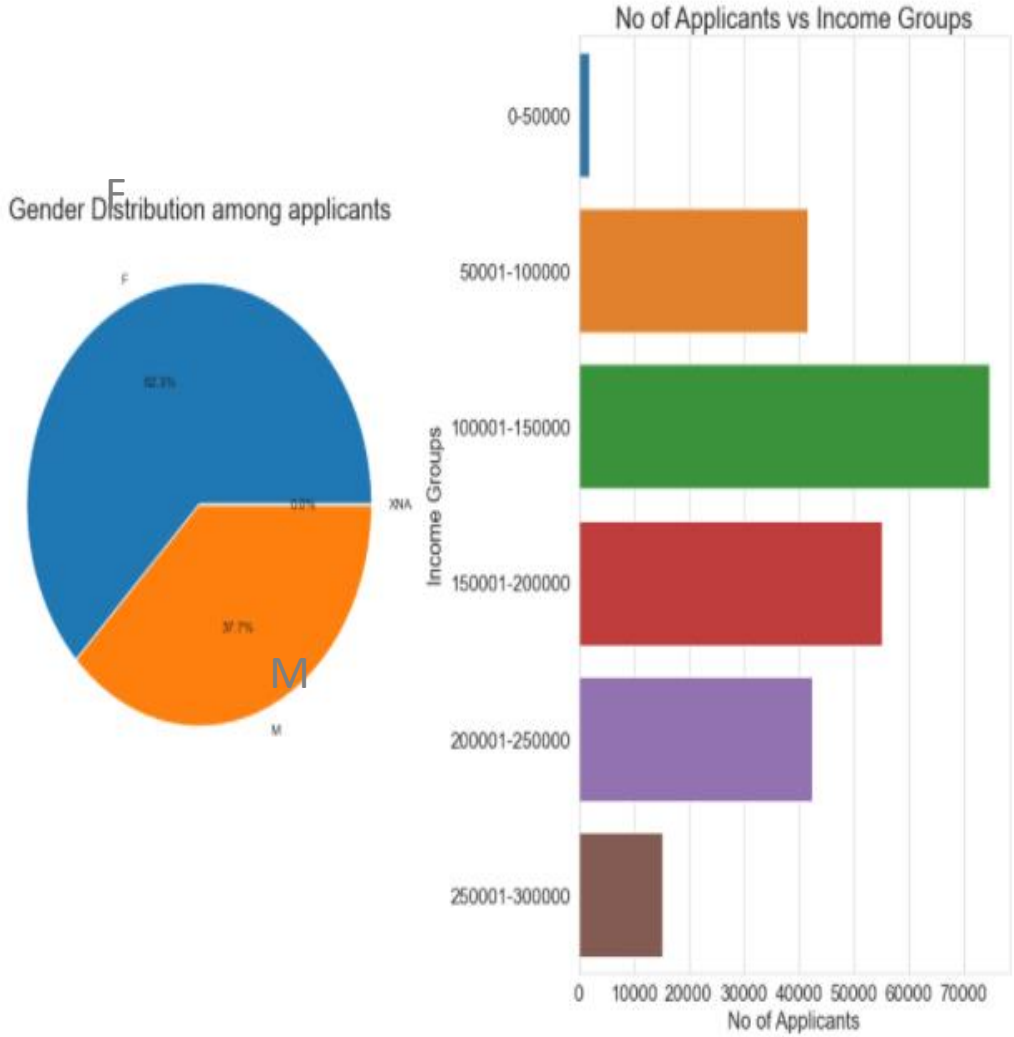
→Categorical variable

→Continuous Variable

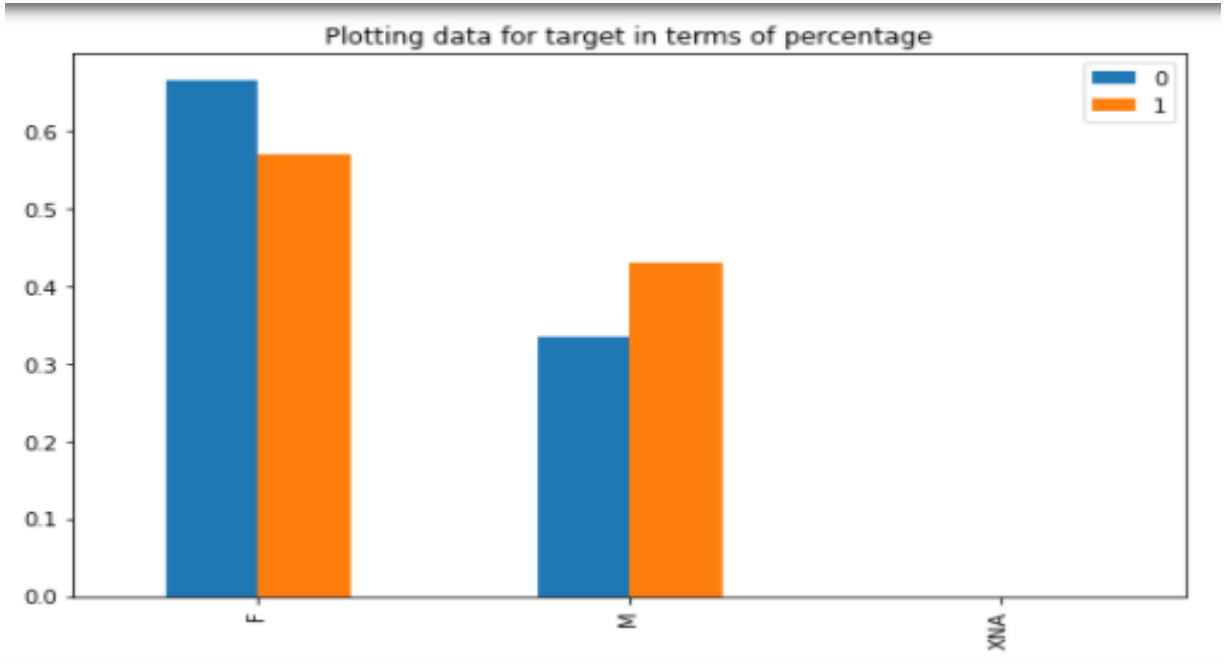
- Bi-variate Analysis:-
    - →categorical Vs Categorical
    - →Categorical Vs Continuous
    - → Continuous Vs Continuous
- Multi-variate Analysis

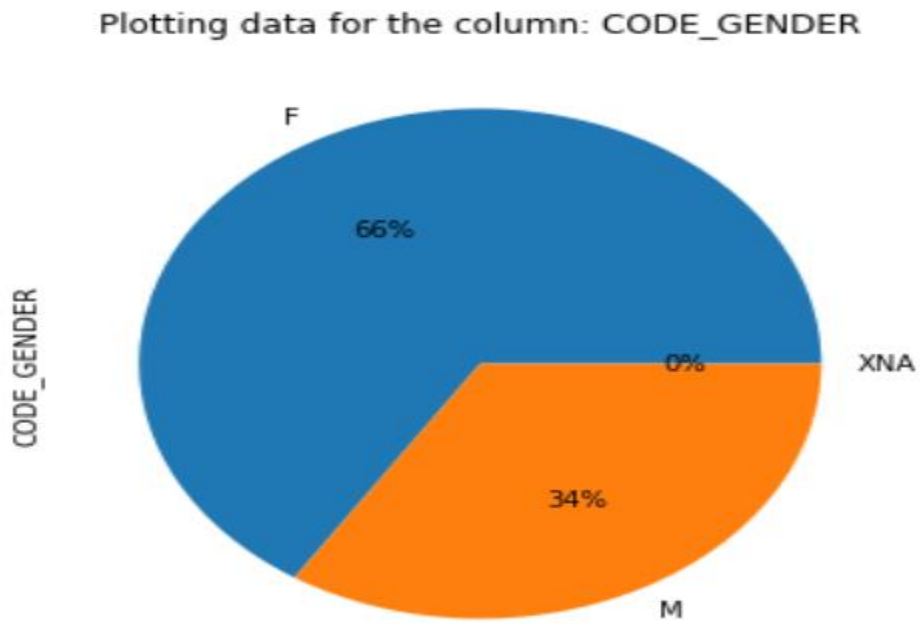
Gender count in data set and number of applicant with respect to income

- Here gender distribution in pie chart it is clearly visible that there are a greater number 63% of female applicant , men only 37%.
- Second bar chart shows the No of applicants Vs Income group. We can see that there many applicant in salary range of 10 thousand to 15 thousand



Univariate Categorical Analysis

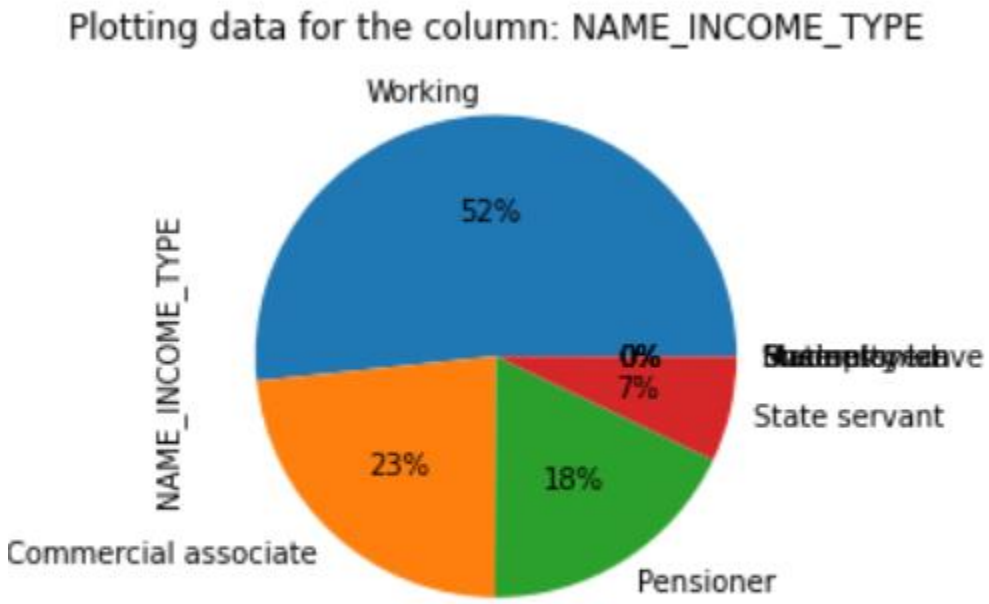


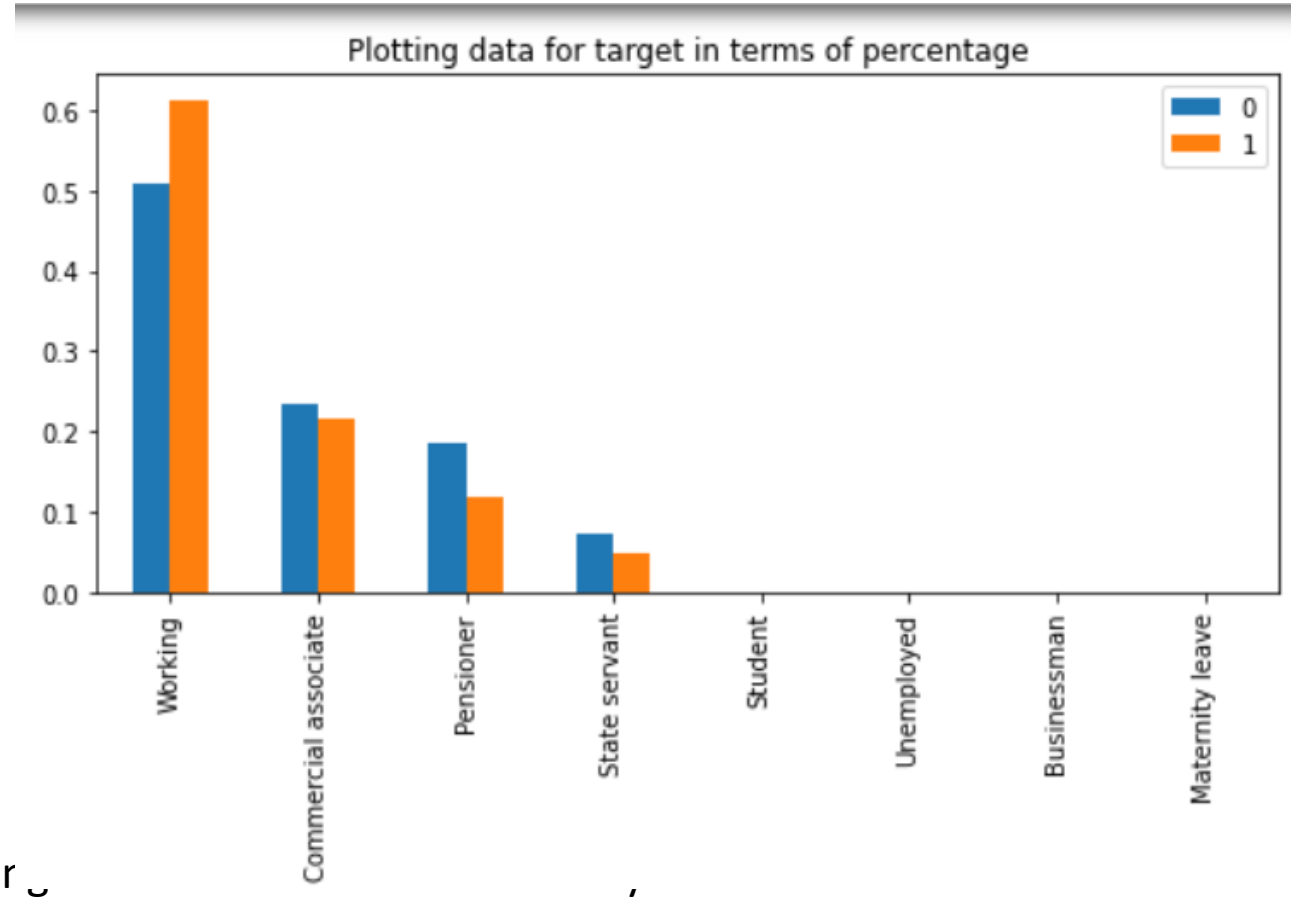


As per graph there are most applicant of the female

From CODE_GENDER column we can see that out of total customer 66% are female and only 34% are male who opted for loan. Where as while repalying the loan males population is more likely to defalut than female customer.

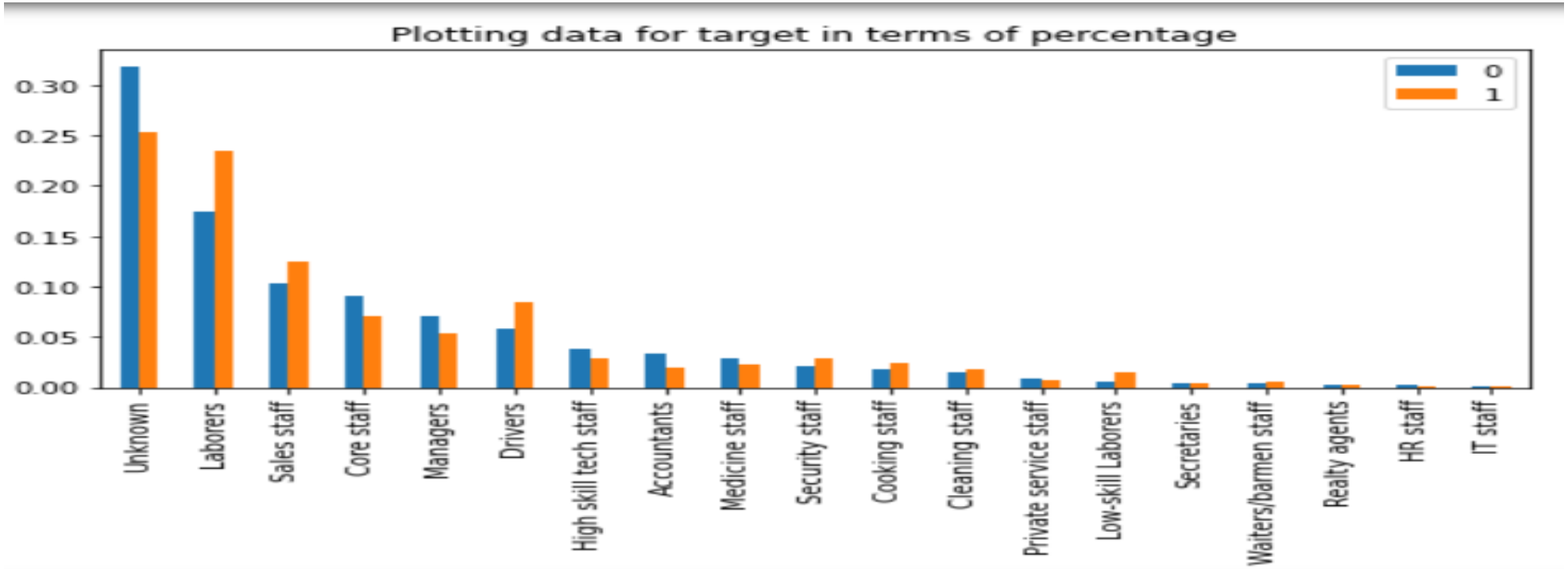
Univariate Categorical Analysis - Distribution of Income Type





From NAME_INCOME_TYPE we can see that workir g , , by Commercial Associate where as pensioner customer are less likely to default the loan. And also working customer are more likely to apply for loan follwed by Commercial Associate

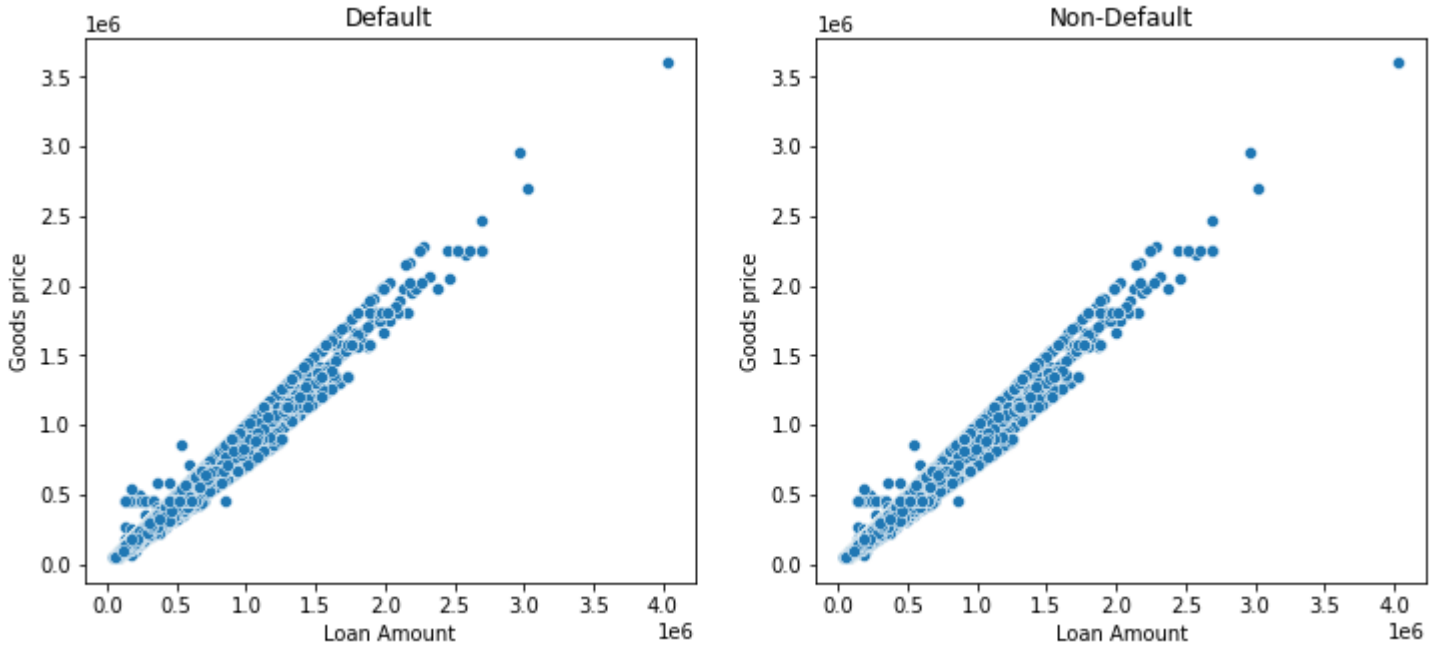
Univariate Categorical Analysis -By occupation type



1) Laboure's, sales, driver are more likely to default.

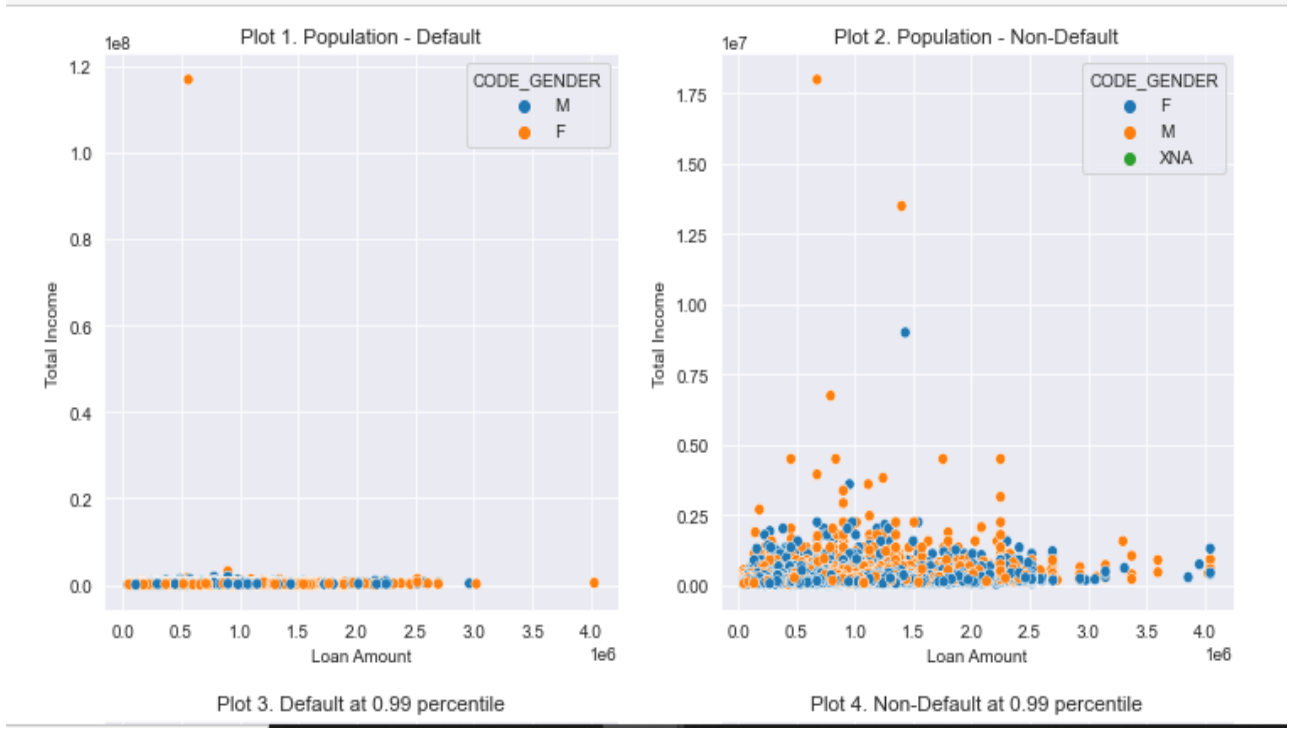
2We can inference that customer who are employed as Labourers are more likely to default the loan followed by Driver and Sales Staff. However, Unknown category (which we replaced for Null values in our data set earlier) has maximum capability of replaying the loan followed by Managers and Accountants. We can consider calling customer whose occupation type is missing again for collecting information.

Bivariate Analysis goodprice vs loan amount:



Most of the customers are availing loan almost equal to the goods price.

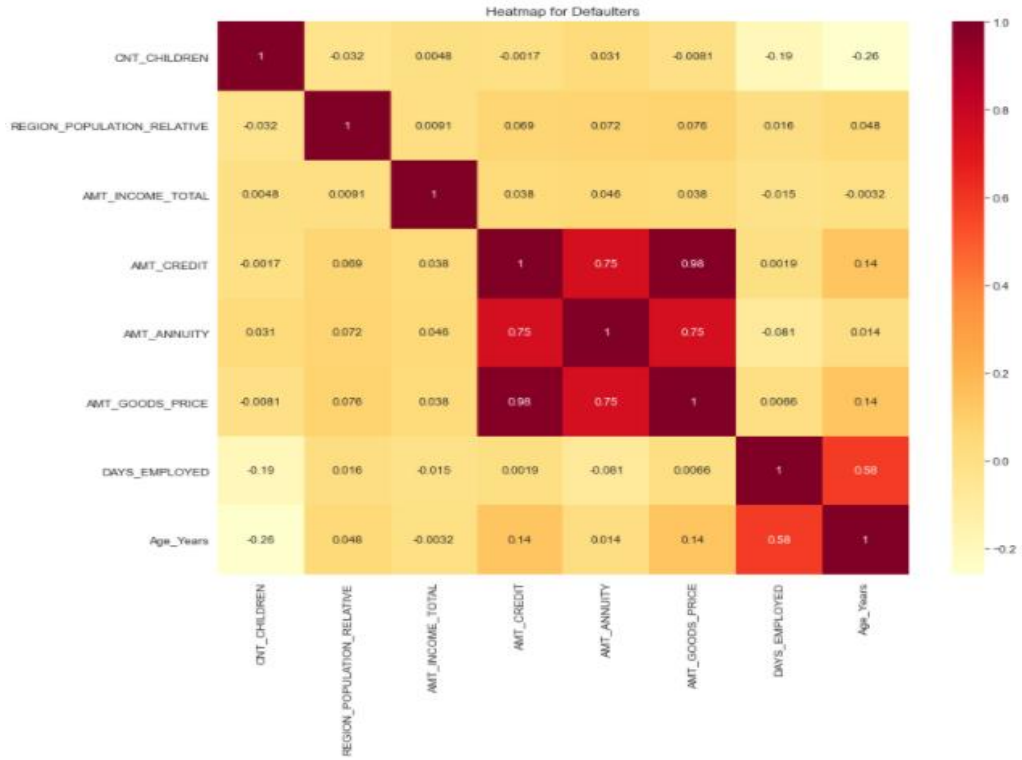
Bivariate Analysis

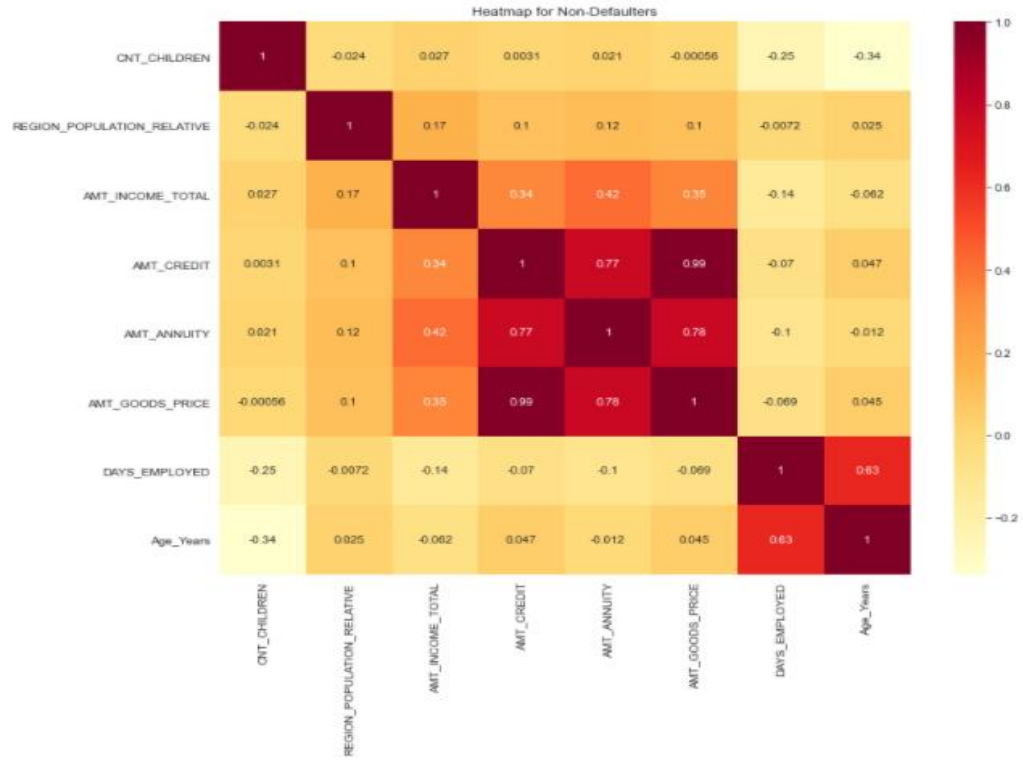


- Plot 1 & 2 for entire population is tightly packed and difficult to draw inferences

--most of the customer are applying for loan ranging below 2.5 and anthything above 2.5 is an outlier.

Multivariate Analysis





From the above heatmap, it evident that there is strong correlation between the Goods Price and Loan Amount that was credited. Also, higher the age the number of days people were employed is also high. On the other hand, There is weak correlation between the age and count of children and between the loan annuity and number of days applicant was employed

top 10 correlation

FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999756
DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999756
FLAG_EMP_PHONE	Years_Employment	0.999756
Years_Employment	FLAG_EMP_PHONE	0.999756
DAYS_BIRTH	Age_Years	0.999711
Age_Years	DAYS_BIRTH	0.999711
DAYS_REGISTRATION	Years_REGISTRATION	0.999554
Years_REGISTRATION	DAYS_REGISTRATION	0.999554
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998510
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510
DAYS_ID_PUBLISH	Years_publish	0.997518
Years_publish	DAYS_ID_PUBLISH	0.997518
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997018
FLOORSMAX_AVG	FLOORSMAX_MEDI	0.997018
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993582
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	0.993582
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.988153
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988153

dtype: float64

From application dataset following inference can be made

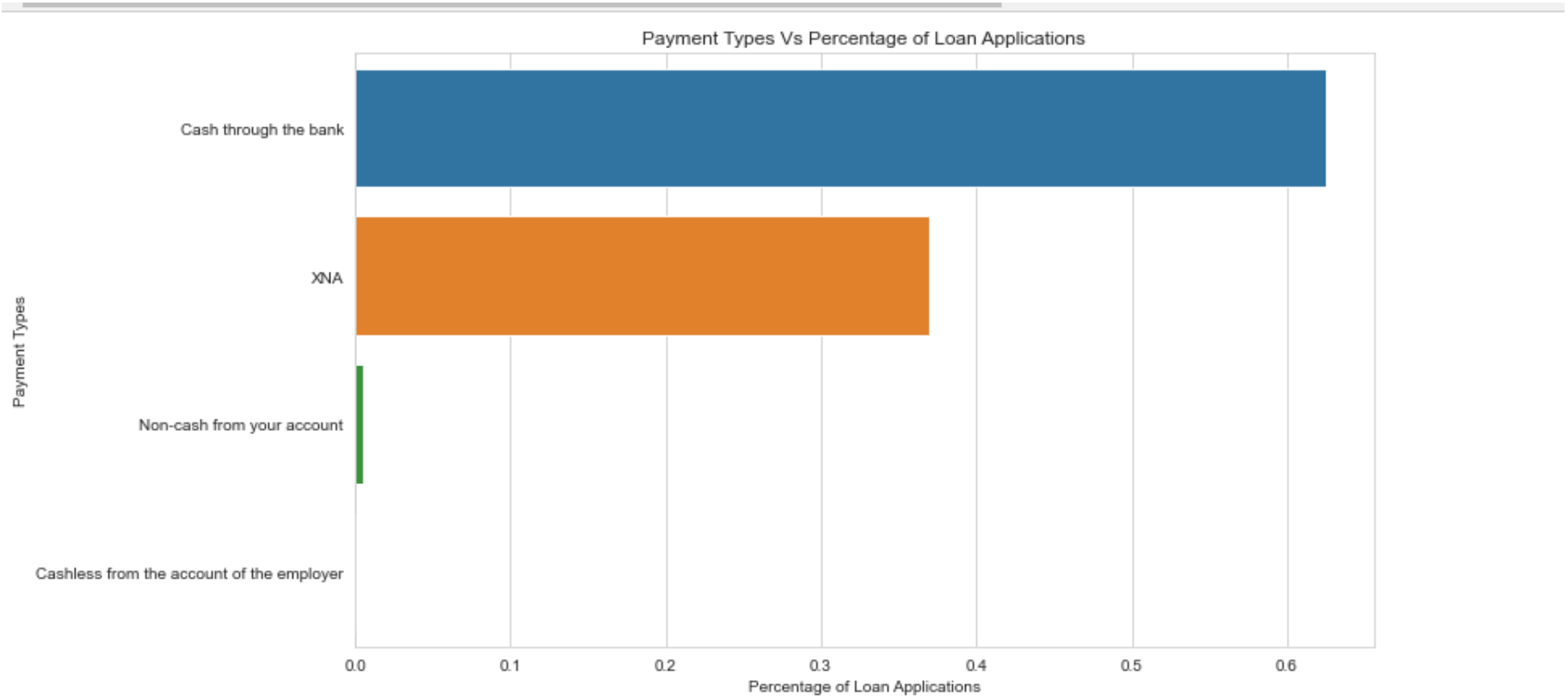
After analysing the datasets, we can see that there are quite a few variables through which the bank can see what are the driving factors as to who can repay the loan.
Factors which indicate that the person will be a **Non-Defaulter** are:

1) LOANS_EDUCATION_TYPE: People who have Academic Degrees have less defaults as compared to other people.
2) AMT_INCOME_TOTAL: Customers who have income in the range of 700K and 800K are least likely to default.
3) ORGANIZATION_TYPE: Clients with Trade Type:4 & 6, Industry Type:12 Transport Type: 1 are least likely to defaul

Factors which indicate that the person will be a **Defaulter** are:

1) CODE_GENDER: Male Customers are more likely to default than females.
2) NAME_FAMILY_STATUS: People who are single or have done Civil Marriage are more likely to default.
3) NAME_INCOME_TYPE: Clients who are on maternity leave or are Unemployed are most likely to default on their payments.
4) NAME_HOUSING_TYPE: People who live with in rented apartments or with their parents are more likely to default on loan.

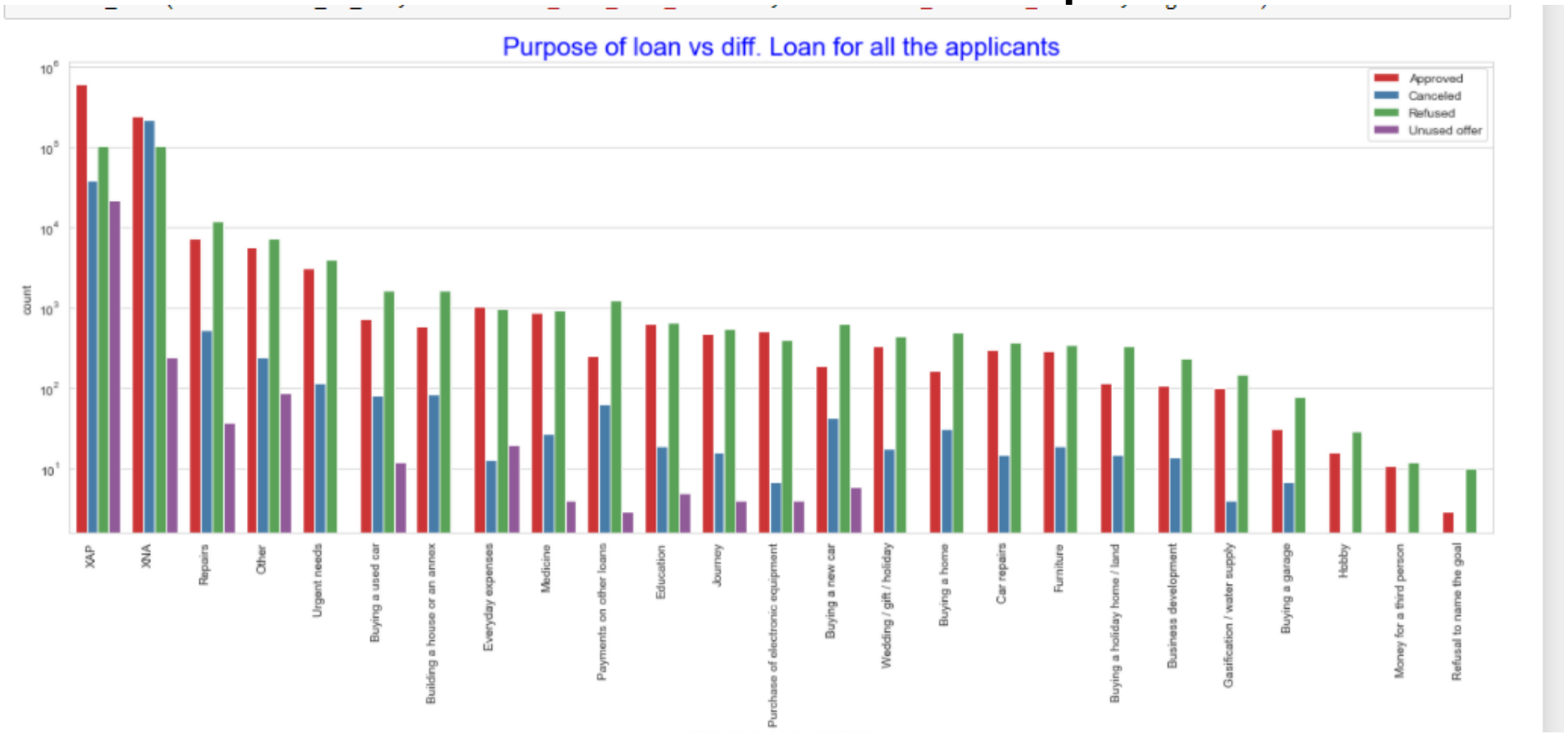
Bivariate Payment Types Vs Percentage of Loan Application



Most of the previous loans were repaid by Cash deposits.
The second highest mode of previous loan repayment has undisclosed methods (XNAs)
Non-Cash and Cashless contribute to extreme low percent of loan repayment methods

Previous data set and application dataset combination analysis

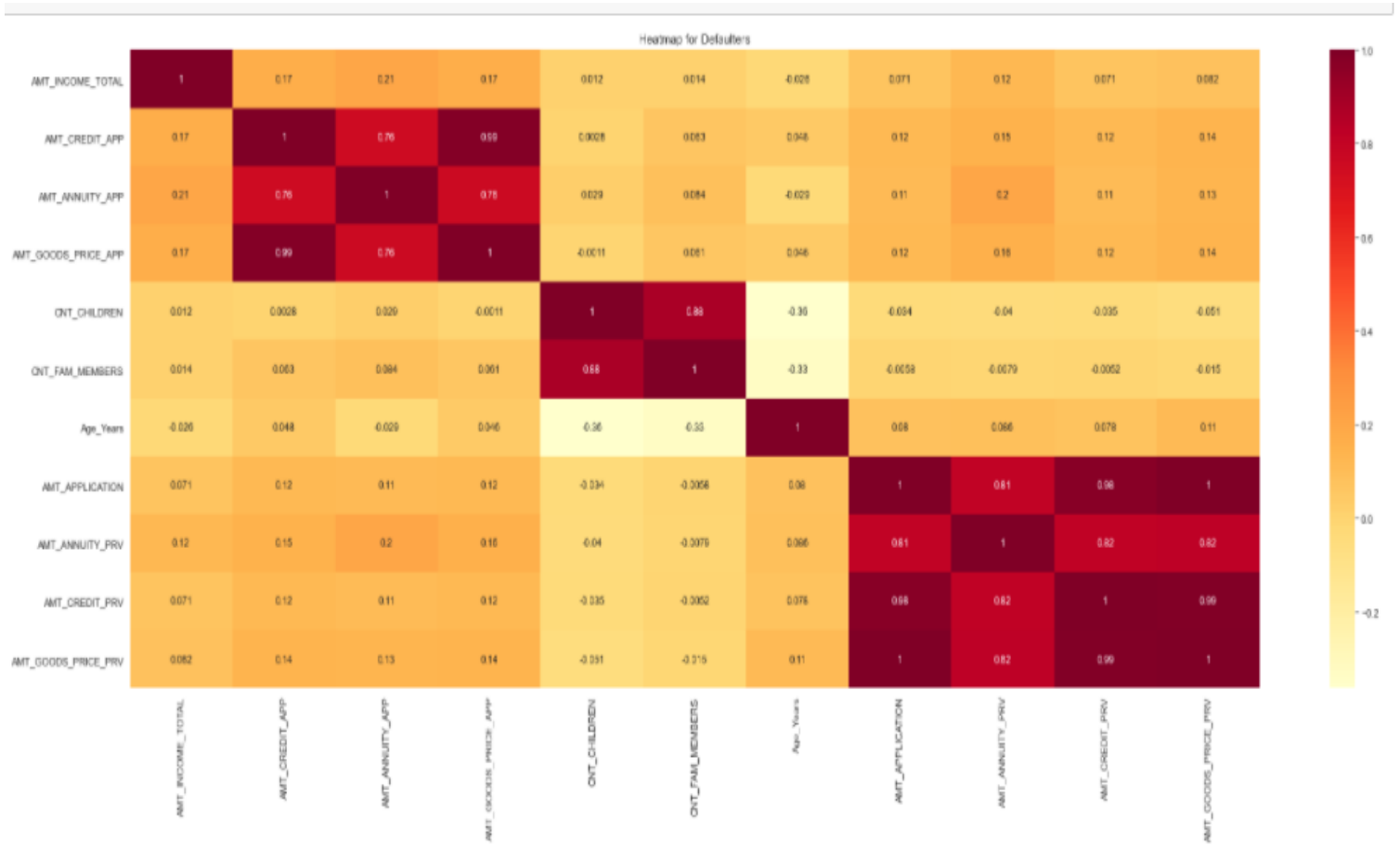
Distribution of Contract Statuses with Loan Purposes



For all the purposes of taking loans, bank has refused more applications that it approved.

The data related to purpose of taking loan is not clearly available, since there are highest number of application that were both approved and rejected for these purposes.

Multivariate Analysis



From the above heatmap, it evident that there is strong correlation between the Goods Price and Loan Amount of current application that is credited. Also, there is strong correlation between the Loan Amount asked by applicant and Loan Amount credited in previous application.

The following variables have a key role in deciding whether customers will default or not.

1. **Income Type**
2. **Income Amount**
3. **Annuity**
4. **Educational Type**
5. **Credit Amount**
6. **Contract Status**
7. **Contract Type**
8. **Gender**
9. **Age**

These can be called as driver variables and the insight was inferred from the graphs represented.

THANK YOU