EDA Case Study Assessment – Bank Loan

CREATED BY : MAYUR INGOLE

Exploratory Data Analysis

UNIVARIATE ANALYSIS :

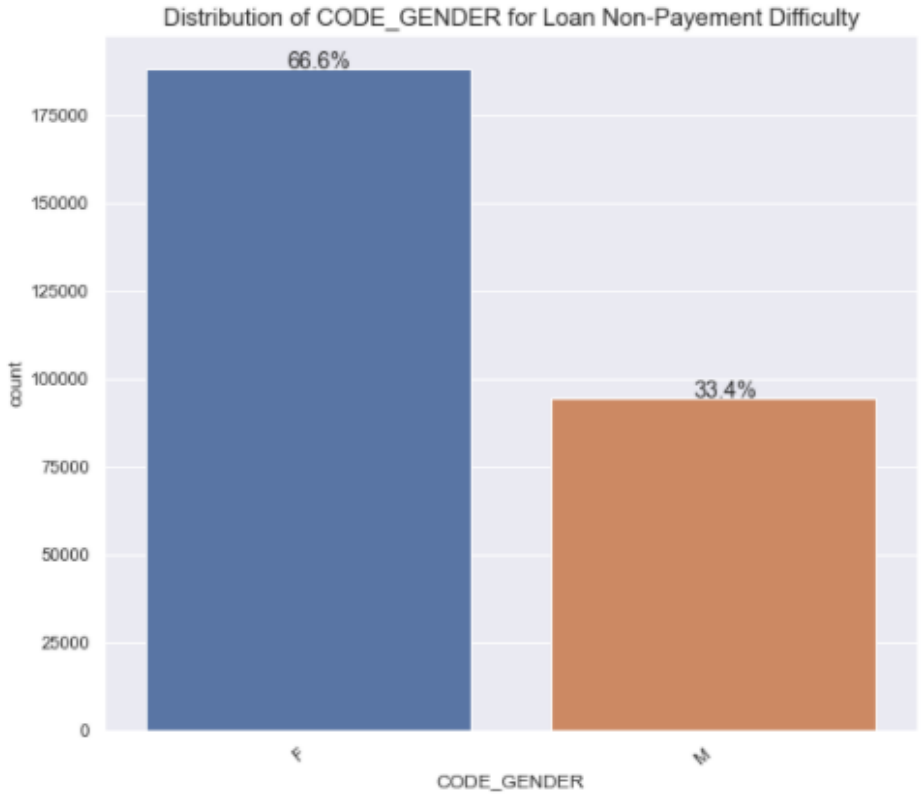
- Analysis On single Variable

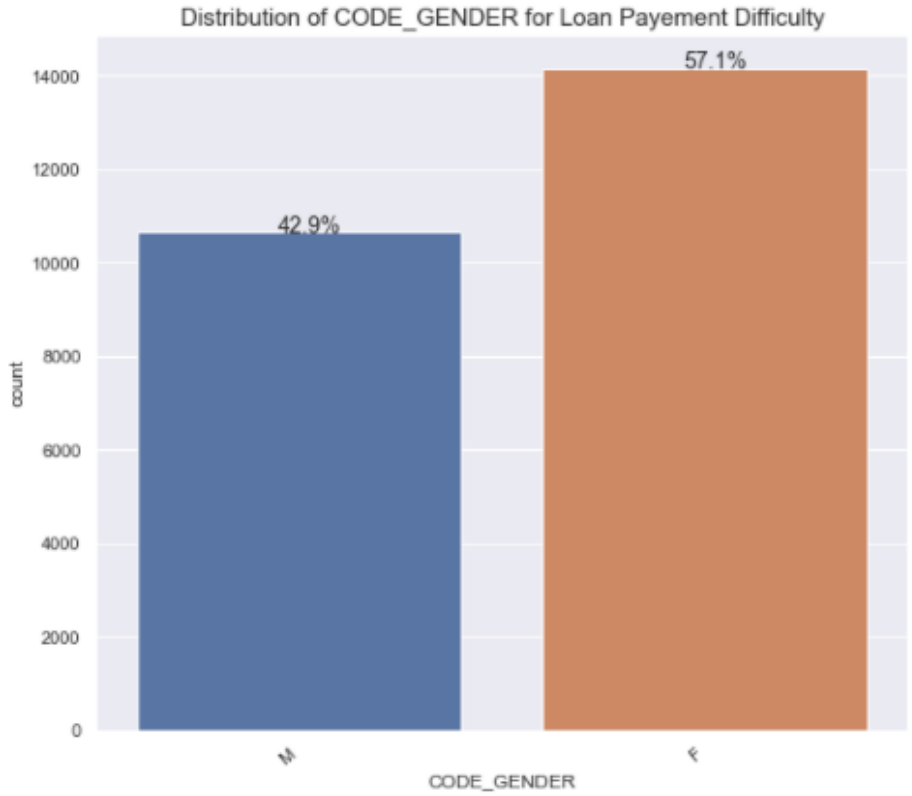
- Analysis On application_data.csv file.

- Getting Insights from the Data

Categorical Variable Analysis

1) CODE_GENDER Variable :

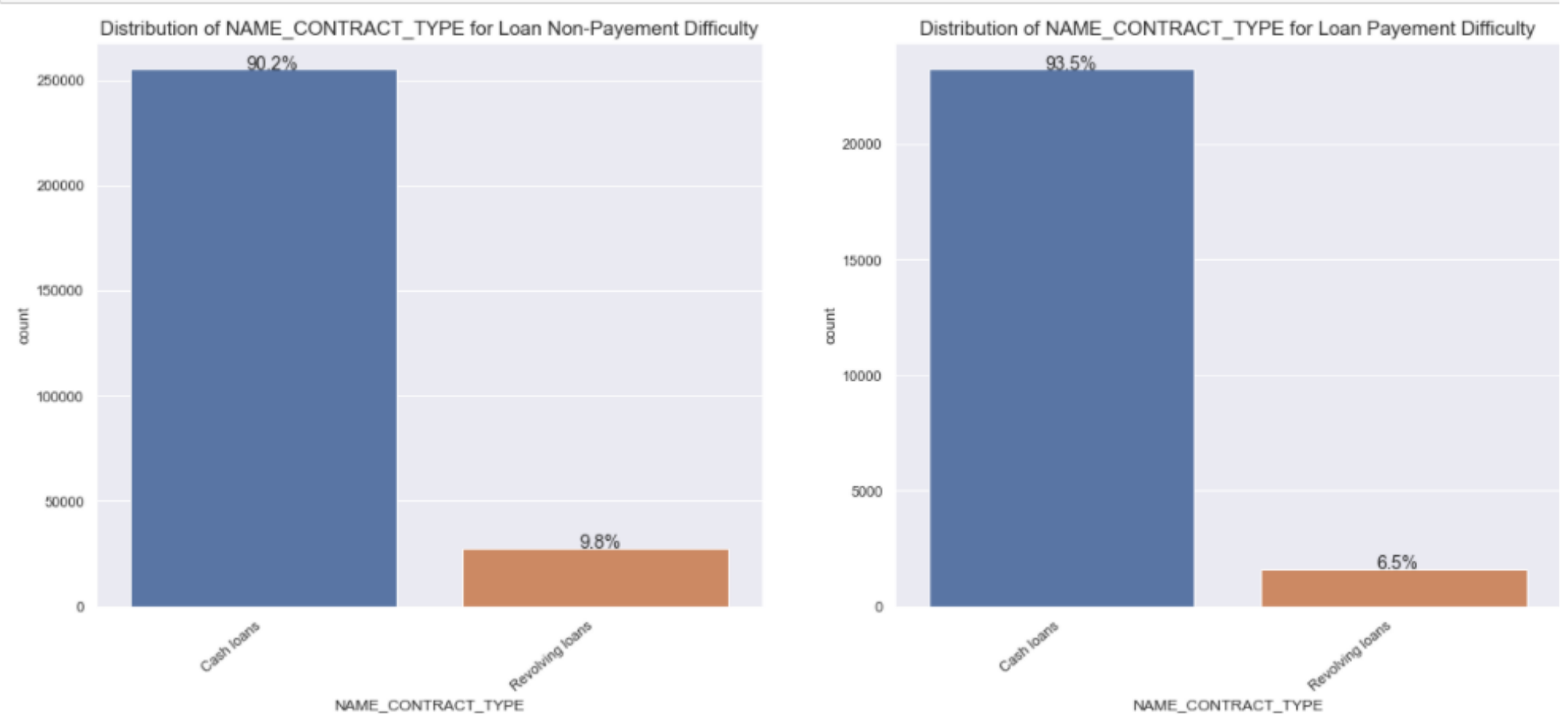




From the above graph we can conclude that:
- Women are more applying for loan as compared to men which is a unique point to note.
- We conclude that as female are more applying for loan as compared to men , because of this women are also high in defaulter.
- Men has 33.4% non defaulter while 42.9% are defaulter. It means men are likely to default as compared to women

Categorical Univariate Analysis

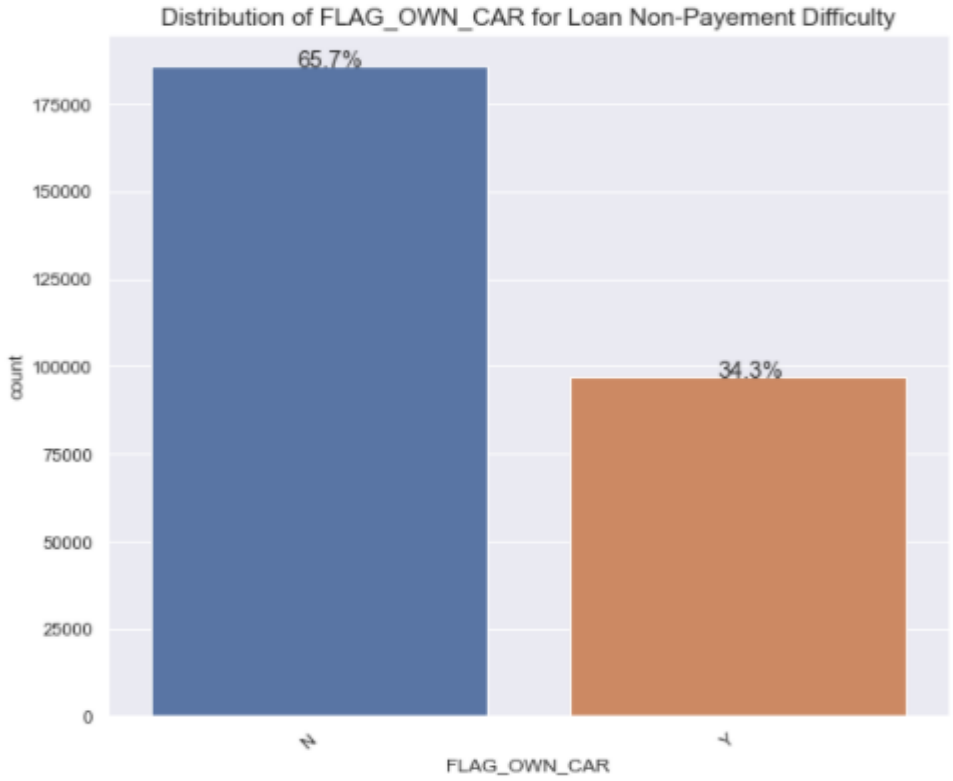
2) NAME_CONTRACT_TYPE:

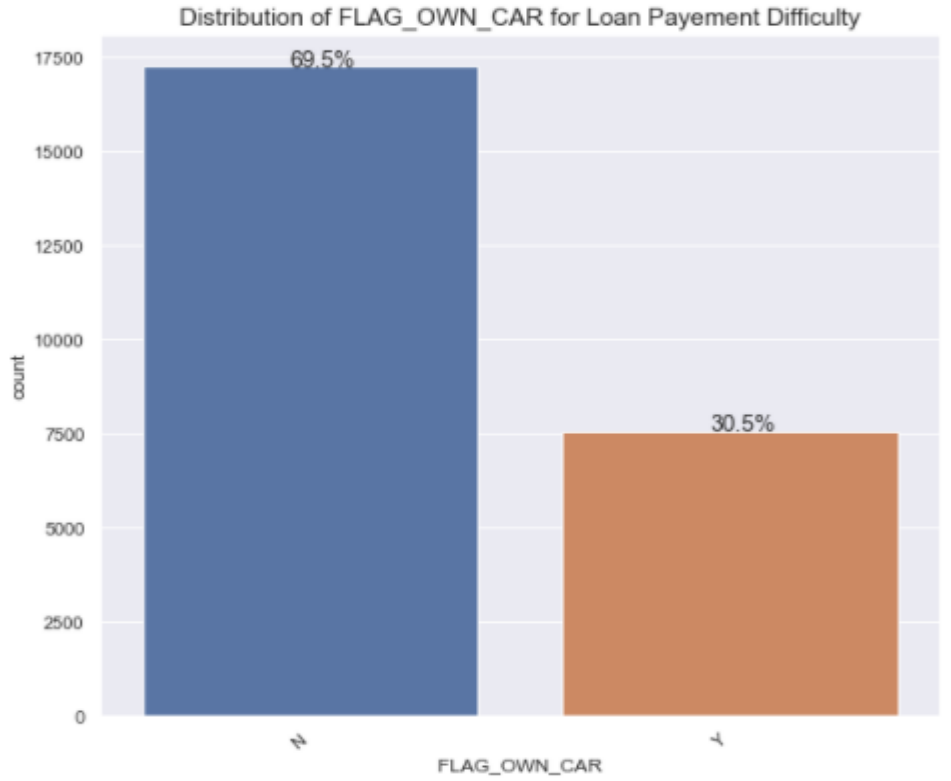


- From the above graph we concluded that people prefer to take cash loans as compared to revolving loan while the defaulter in revolving loans are less at about 6.5%

Categorical Variable Analysis

3) FLAG_OWN_CAR Variable :

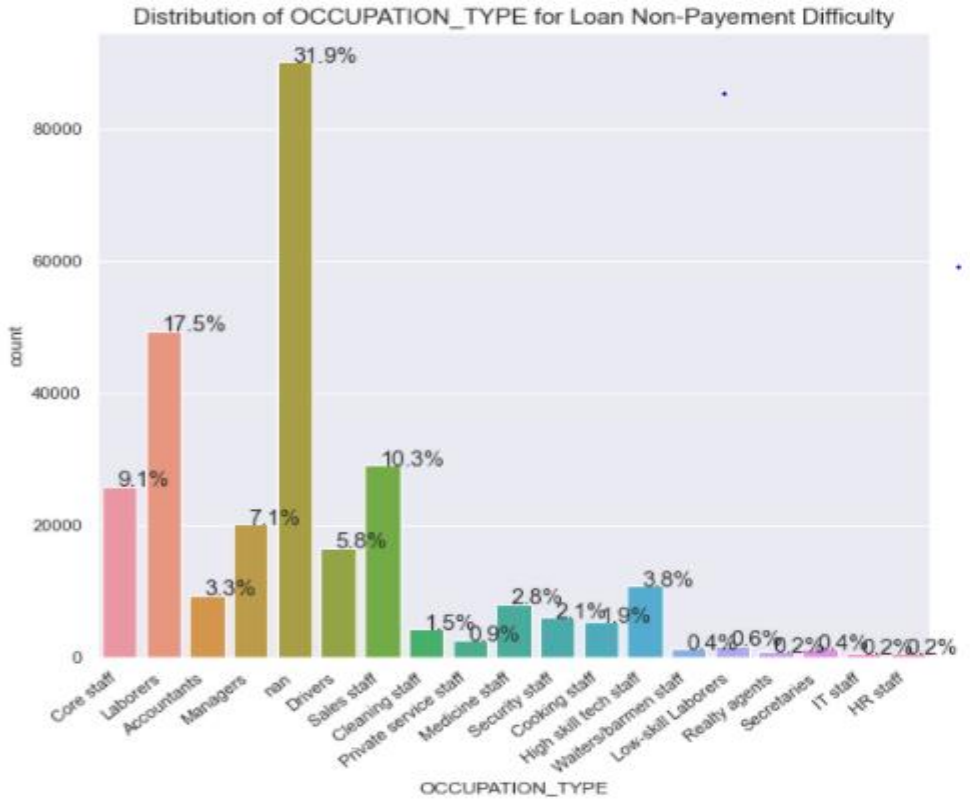


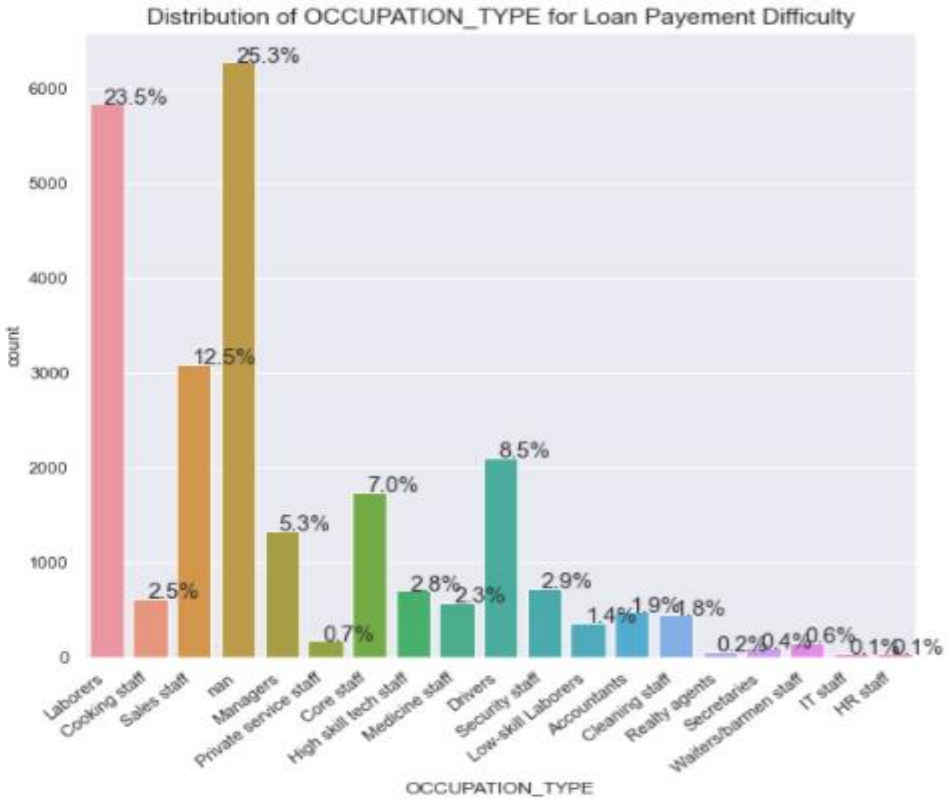


- From the graph ,we can conclude that the client without cars prefer more to take loan as compared to client who have car.
- While the defaulter is also high for the client without car

Categorical Variable Analysis

4) OCCUPATION_TYPE Variable :

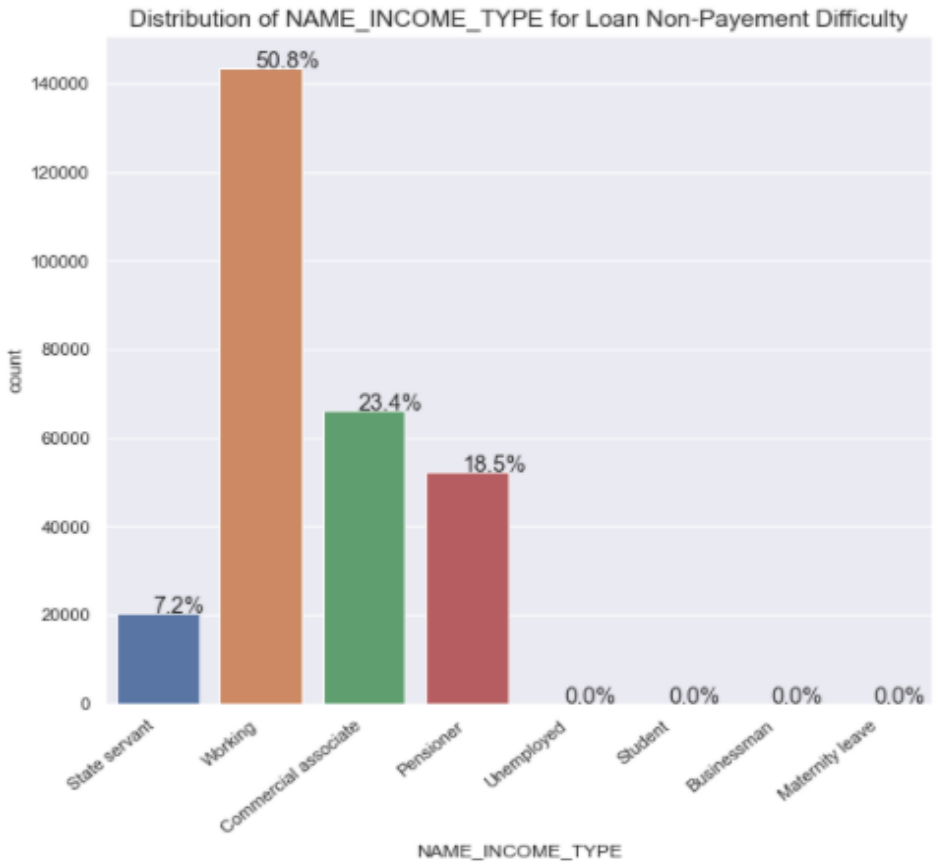


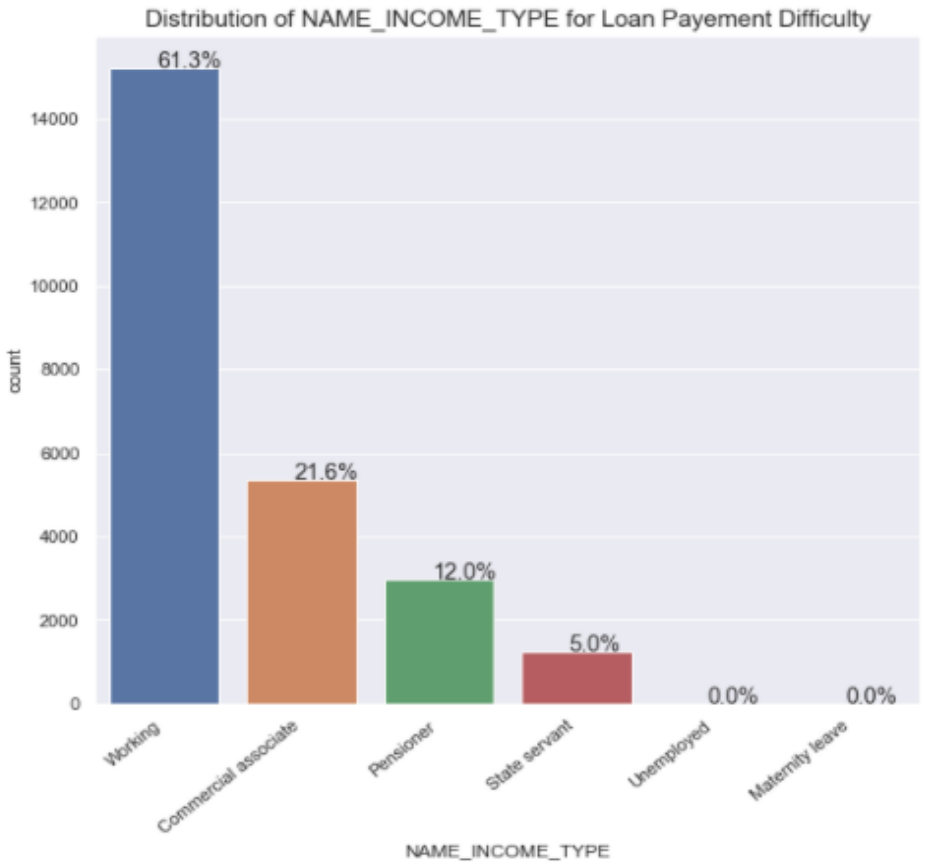


- From the graph ,we can conclude that Laborers, Drivers , Low Skilled labors are more chances of being an defaulter.

Categorical Variable Analysis

5) NAME_INCOME_TYPE Variable :

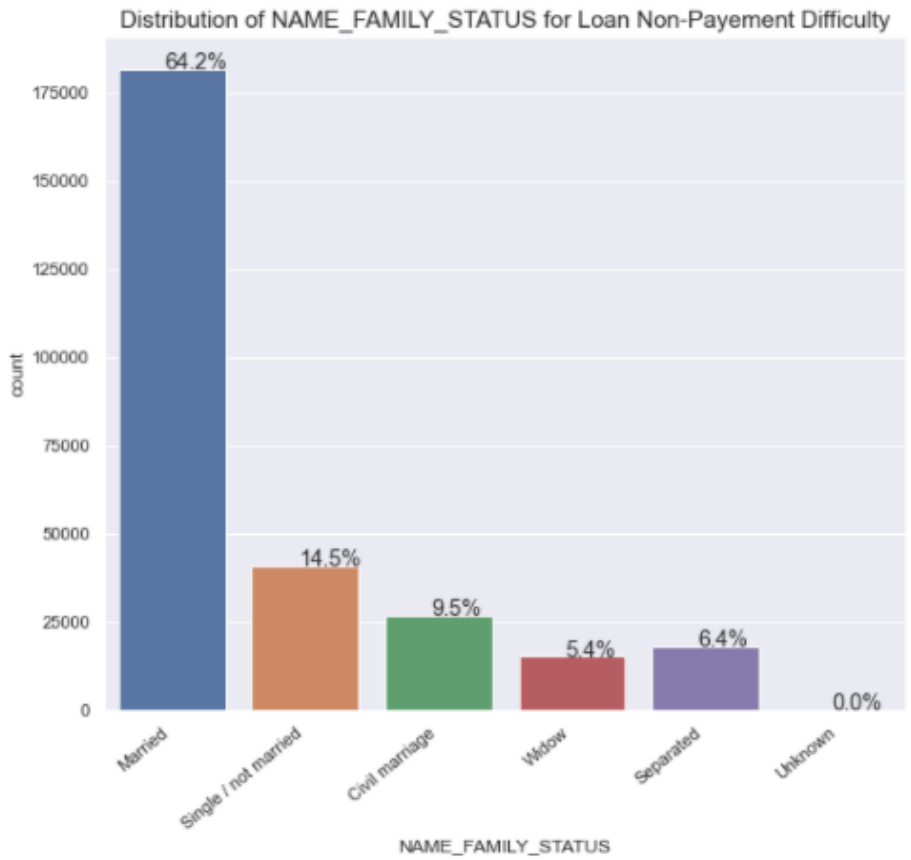


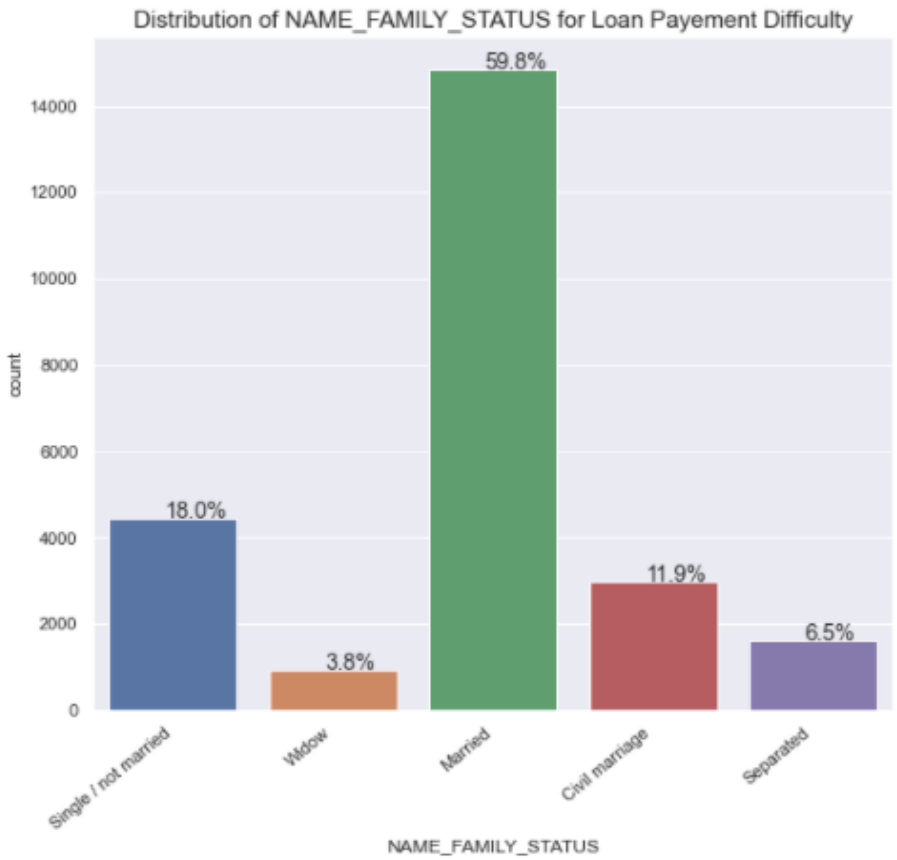


- From the above we conclude that the working class people contribute around 51 % as non deafulter and 61.3% as defaulter.
- Students, Businessman ,Unemployed person has 0% defaulter.

Categorical Variable Analysis

6) NAME_FAMILY_STATUS Variable :

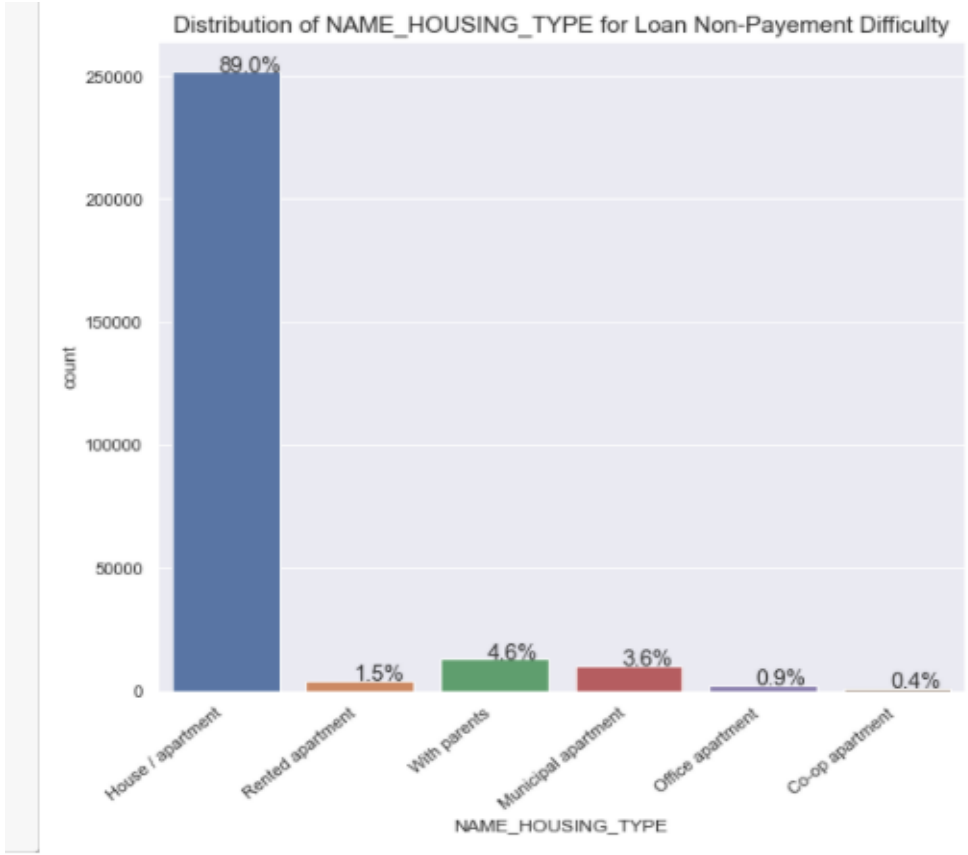


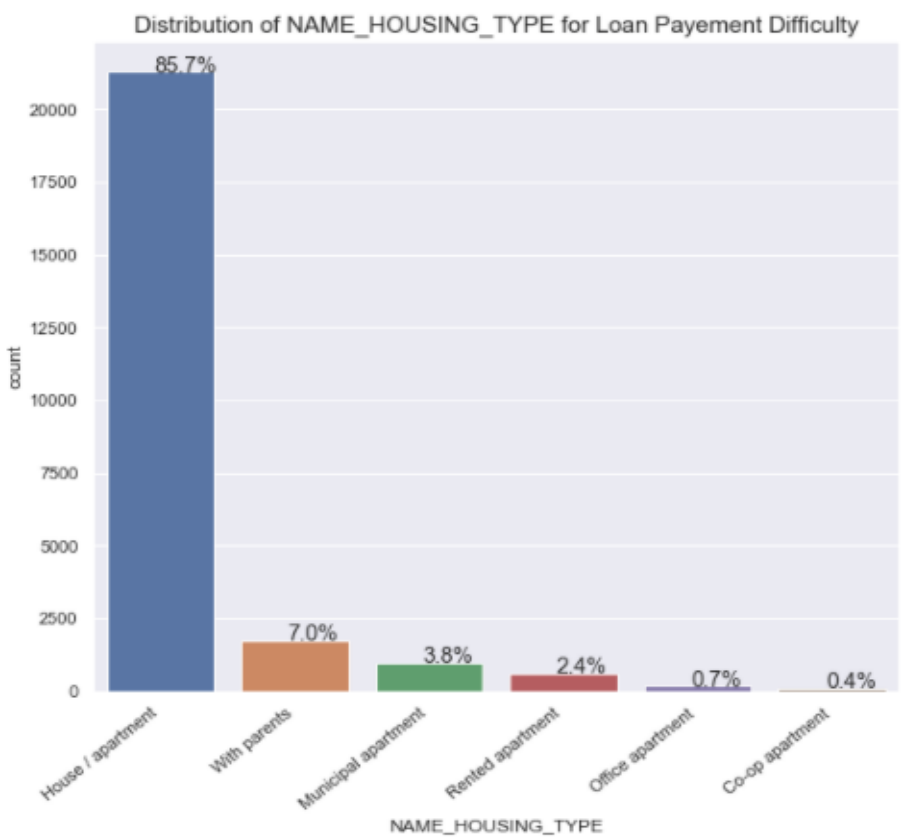


- We observed that the married people are high in taking loan on which 64.2% are non-defaulter while 59.8% are defaulter. It is because married people need to take a loan for various household purposes

Categorical Variable Analysis

7) NAME_HOUSING_TYPE Variable :

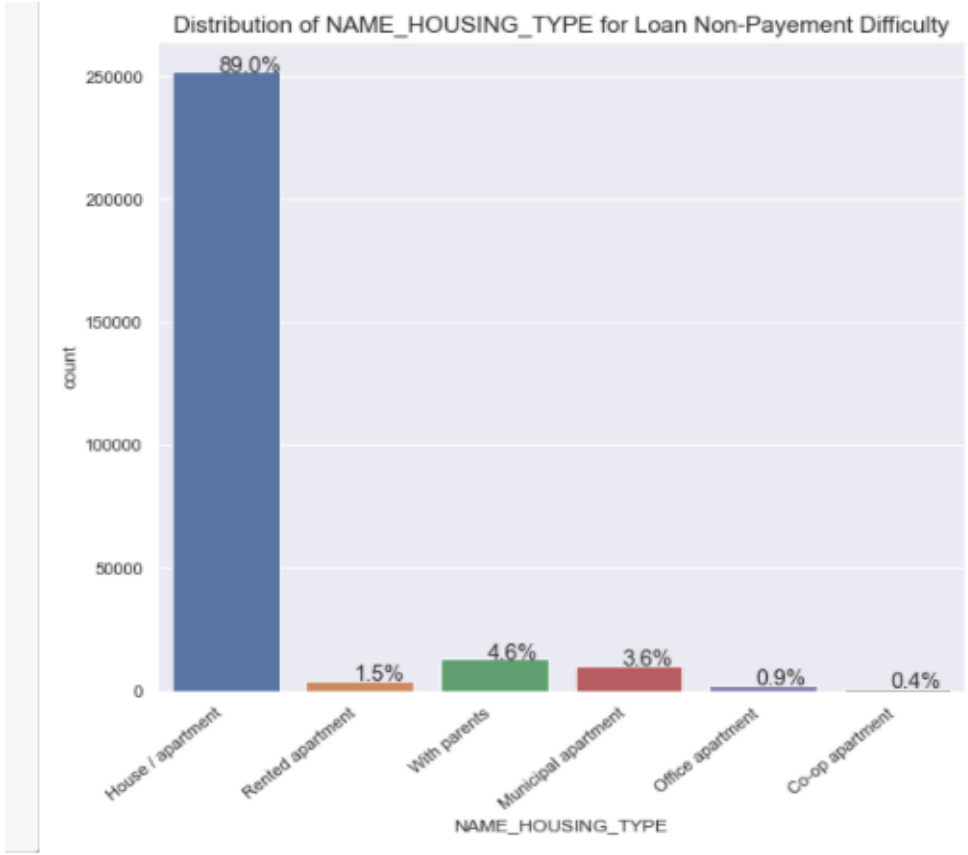


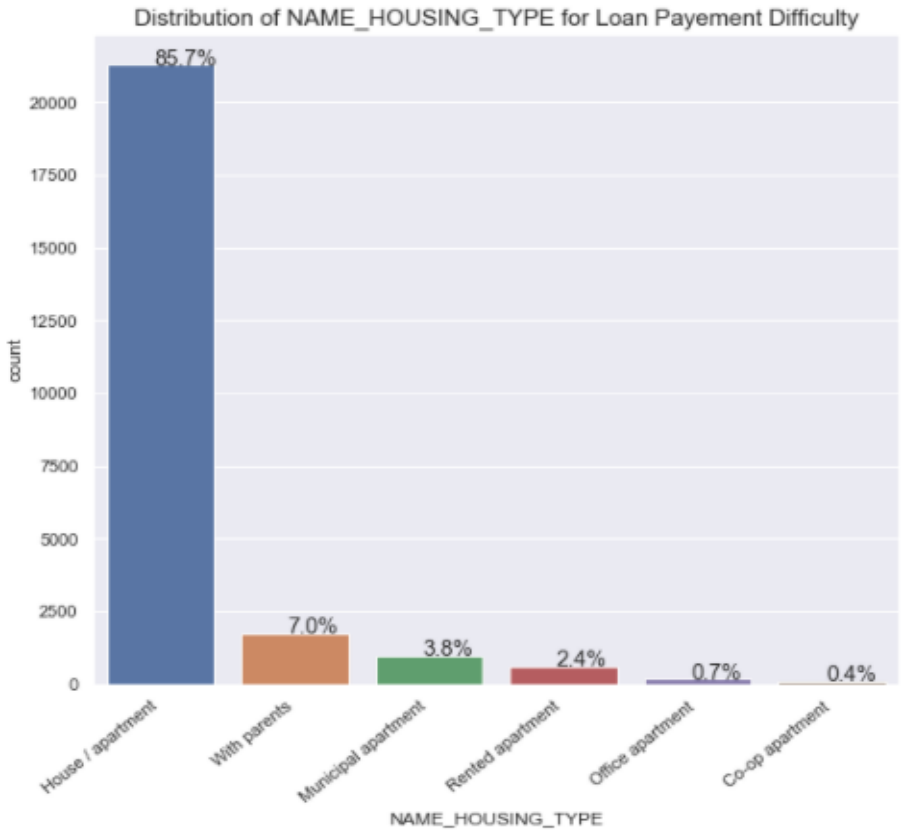


- From the above, we conculde that the client who have House/Apartment are more apply for taking the loans on which 89 % are non defaulter and 85.7% are defaulter. While people who live with their parents faces payment difficulty as compared to others.

Categorical Variable Analysis

8) AMT_INCOME_CATEGORY Variable :

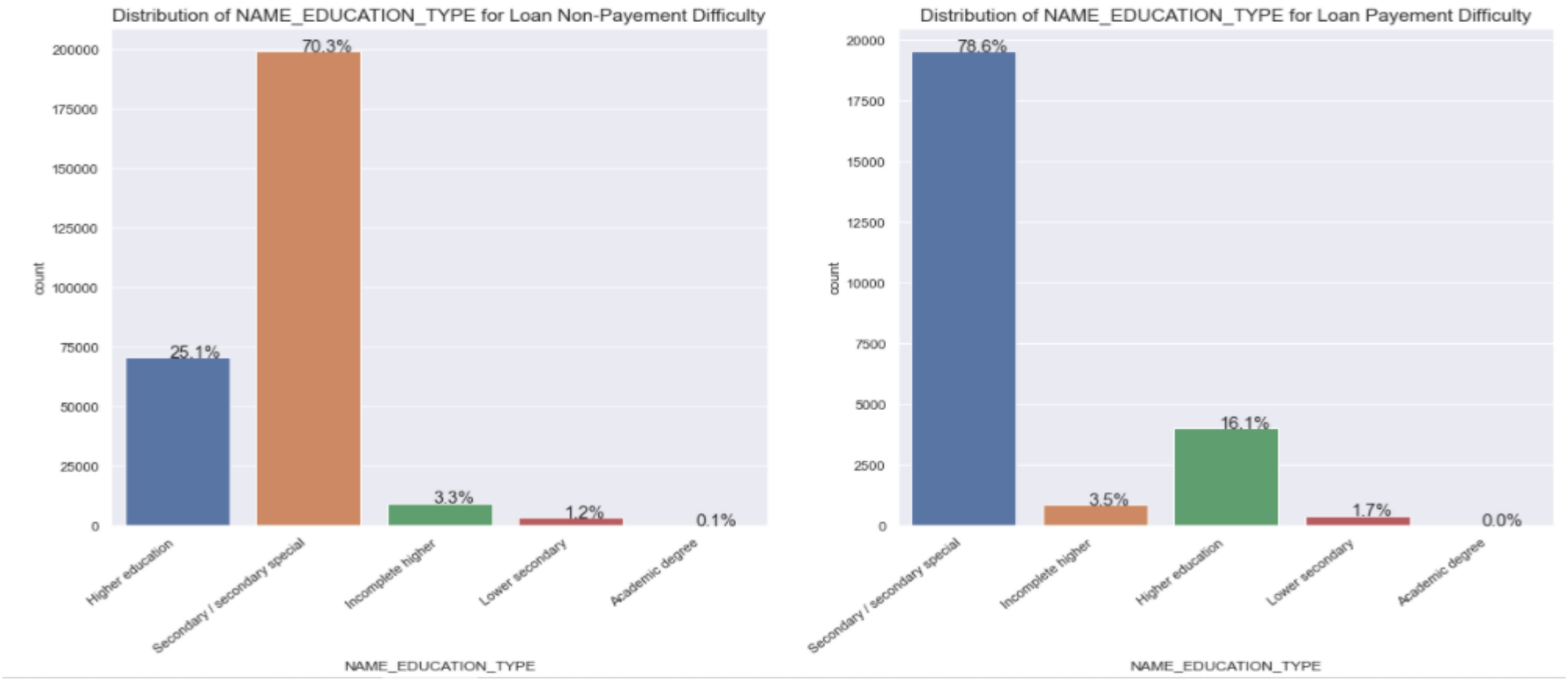




- The Low and High Income group are more applying for loan.
- The very high income group are less defaulter of about 12.1% and non defaulter as 15.3%.

Categorical Variable Analysis

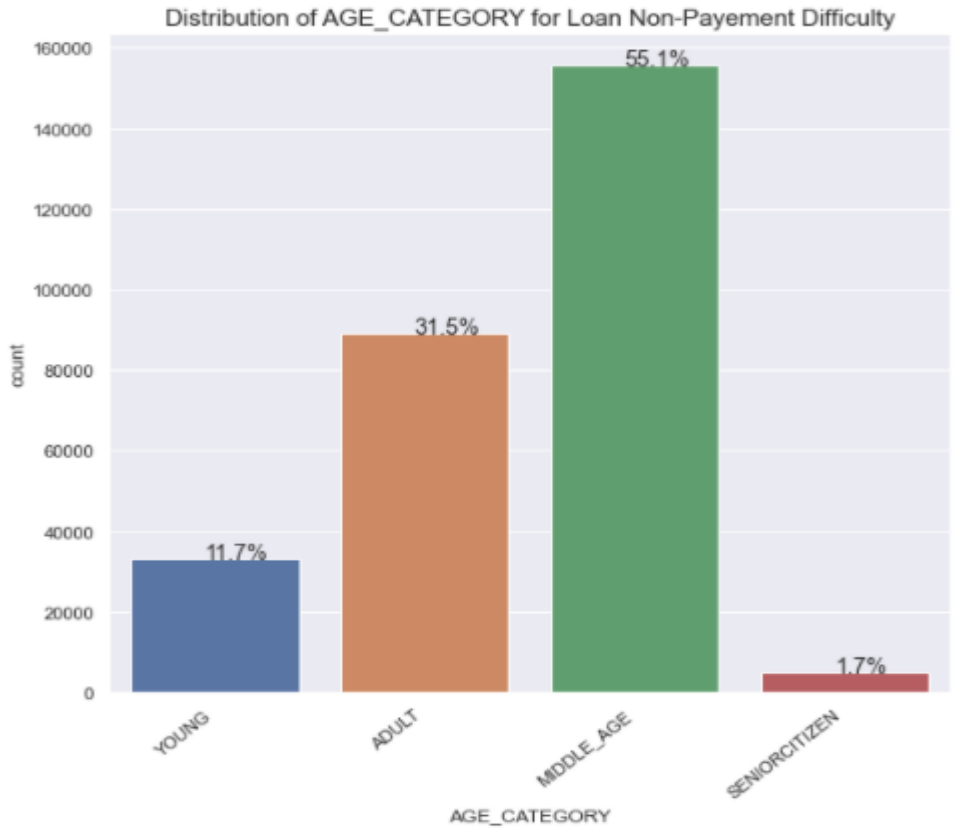
9) NAME_EDUCATION_TYPE Variable :

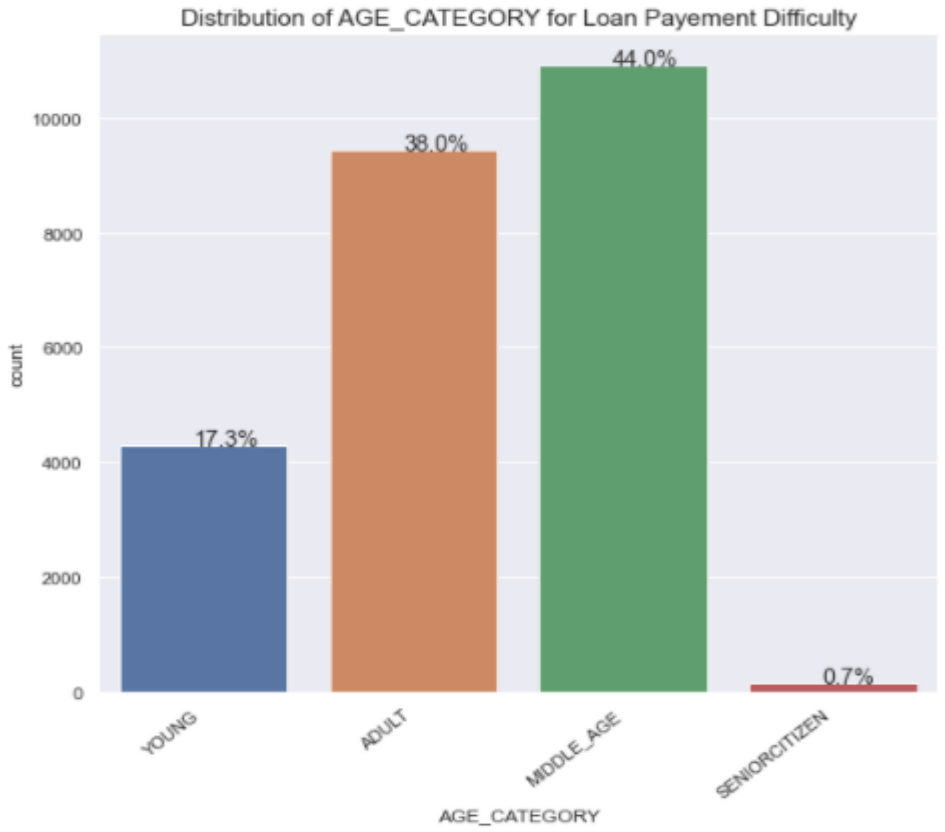


- We observe that the people who have secondary degree are high for applying loans while 70.3 % are non defaulter and 78.6% are defaulter. People who has Academic degree have 0% defaulter and almost 0.1% non defaulter.

Categorical Variable Analysis

10) AGE_CATEGORY Variable :



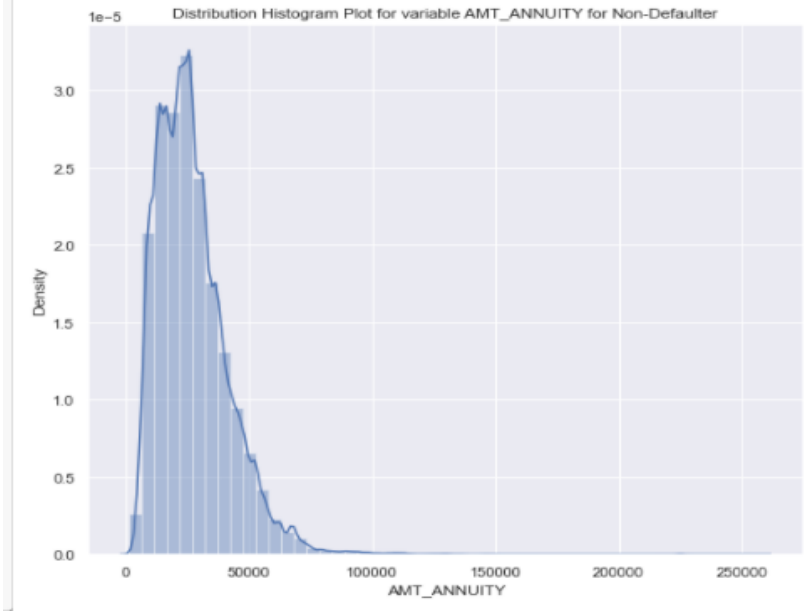


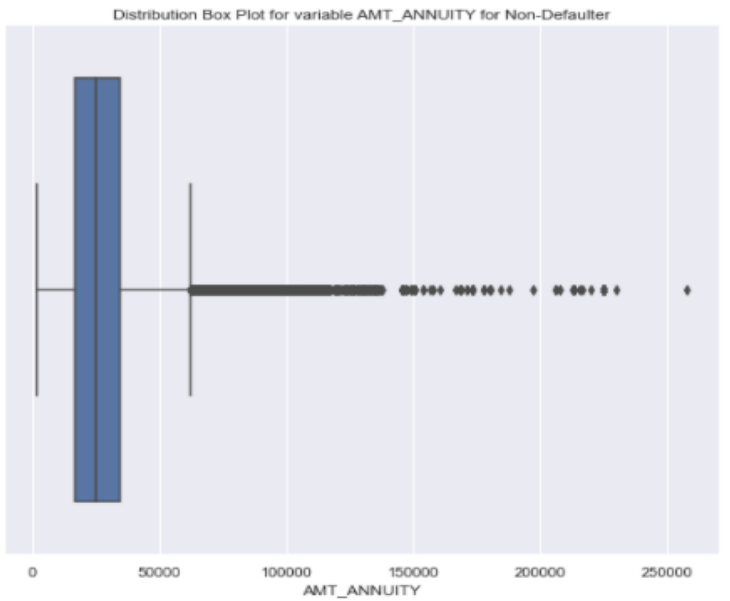
- The senior citizen are less defaulter as compared to other age category. Middle_age are the highest defaulter of about 44.0% and adult are also likely to defaulter at around 38%.

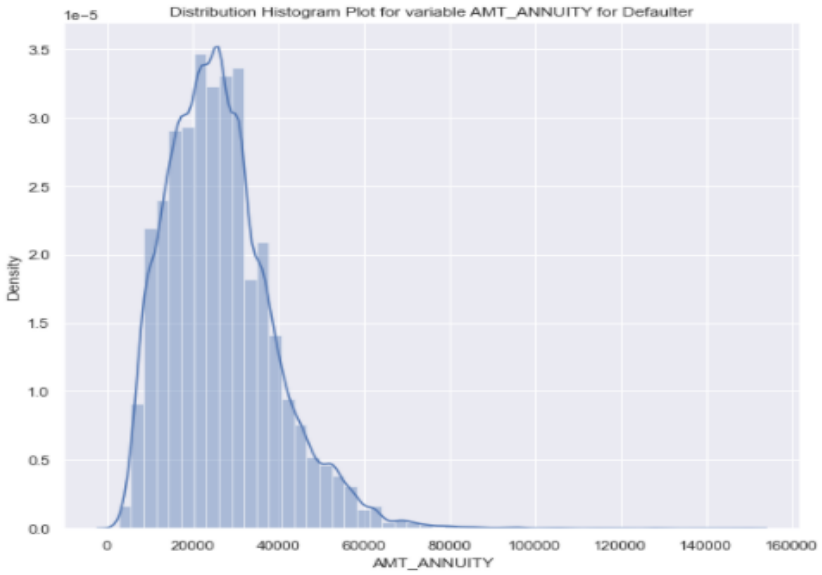
Continuous Variable Analysis

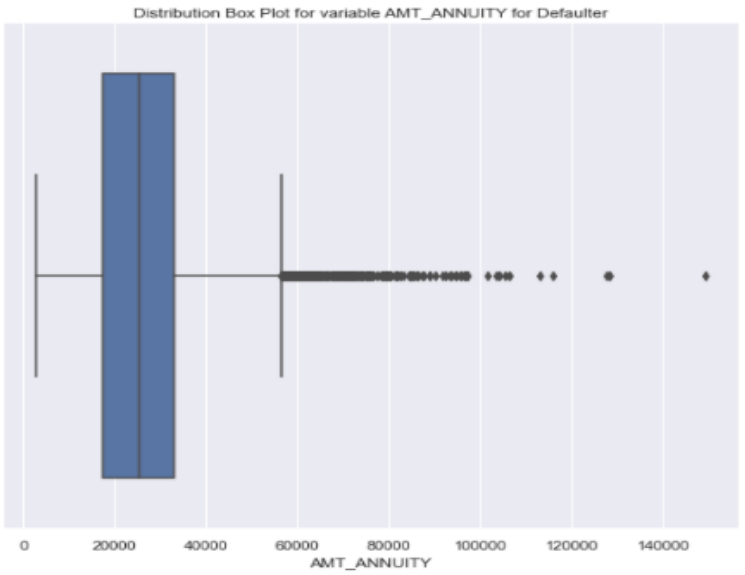
1) AGE_CATEGORY Variable for Target =0 & Target = 1 :

- From the above graph we observe that the distribution of AMT Annuity in case of non defaulter is maximum approximately in between 10000 to 25000 while in case of defaulter the distribution is maximum approximately between 25000 to 40000. It is possible that the maximum Annual Annuity will lead to the defaulter.
- We also observer that the there are outliners in both the cases.





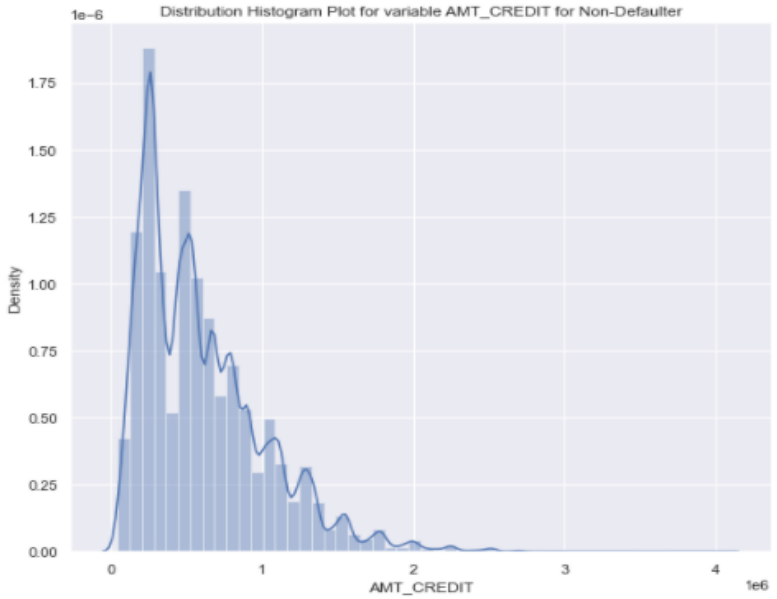


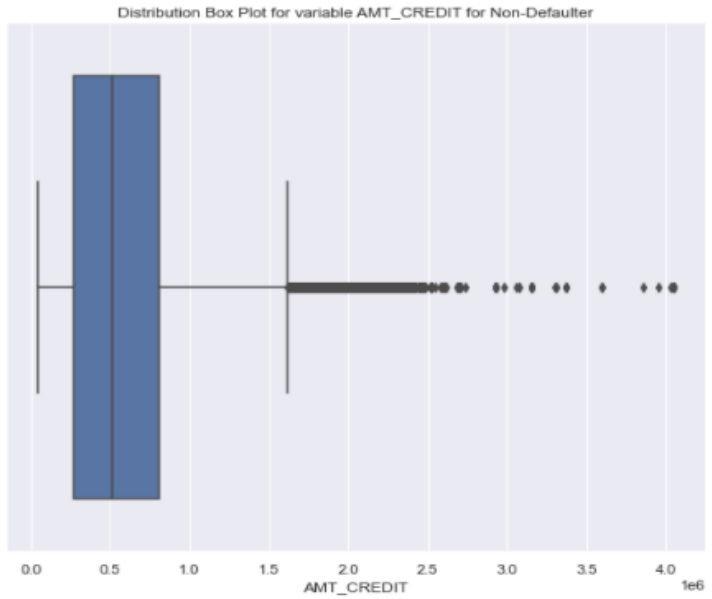


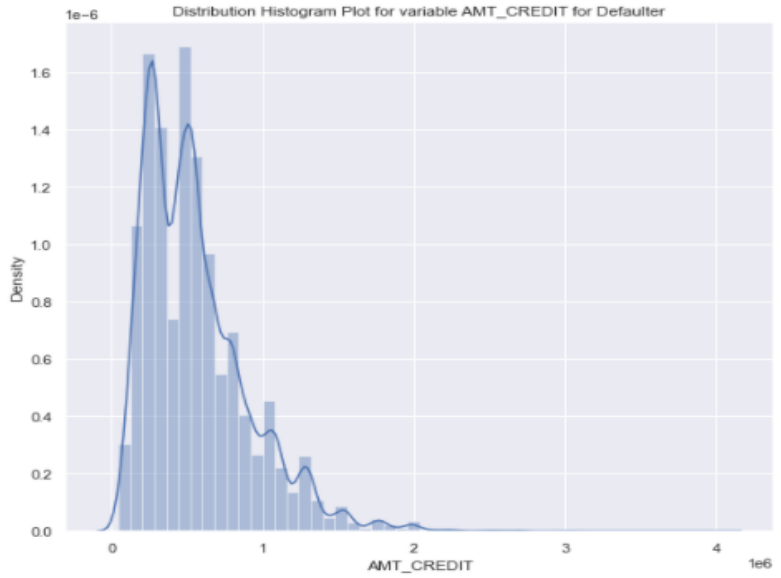
Continuous Variable Analysis

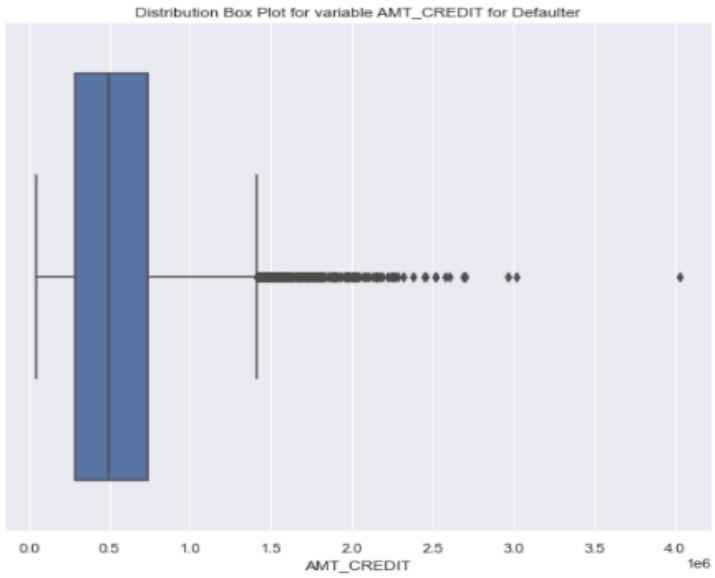
2) AMT_CREDIT Variable for Target =0 & Target = 1 :

- We can observe that some outliners in the data. There is the maximum distribution of data in the first quantile as compared to third quantile it means that most of the client are from first quantile having credit between 0 to 0.5 from the graph





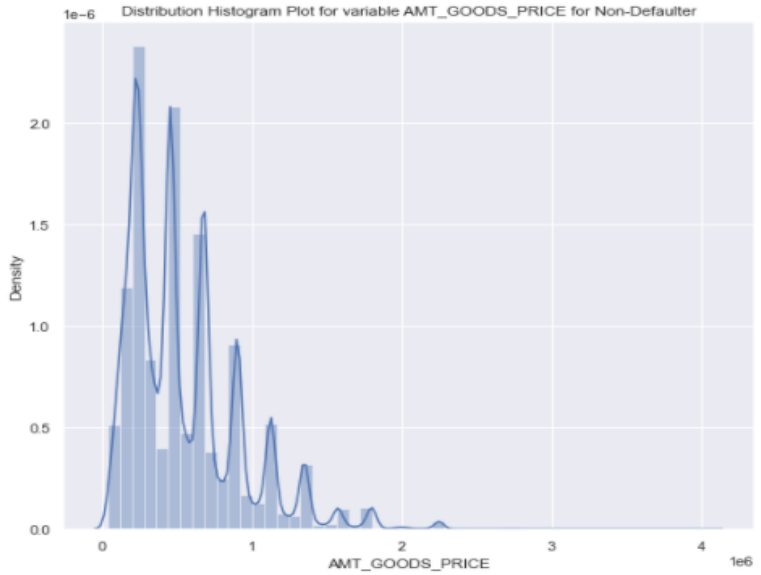


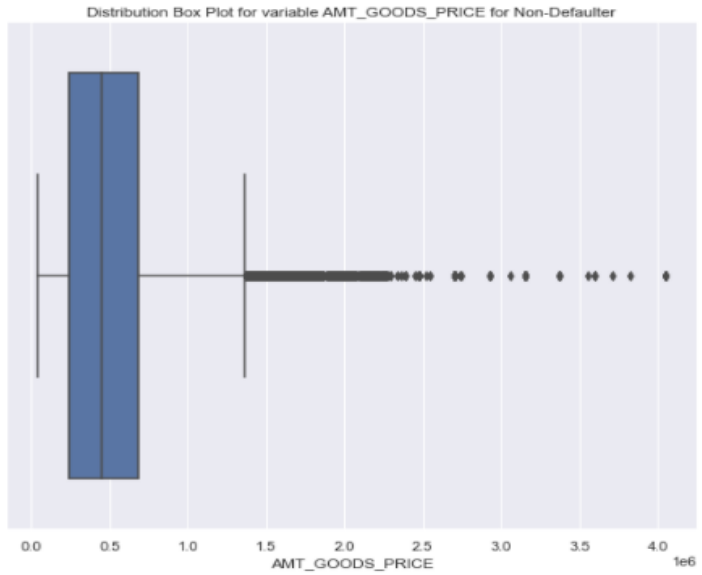


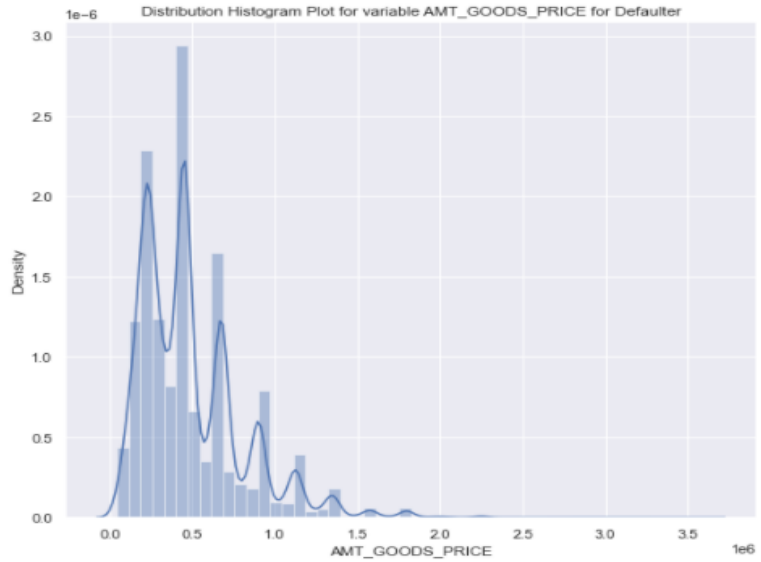
Continuous Variable Analysis

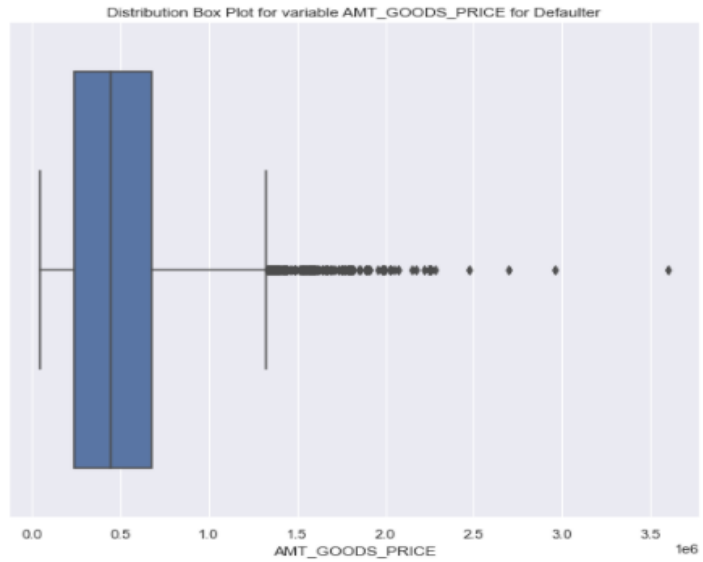
3) AMT_GOODS_PRICE Variable for Target =0 & Target = 1 :

- We can observe that some outliners in the data. There is the maximum distribution of data in the first quantile as compared to third quantile it means that most of the client are from first quantile having credit between 0 to 0.5. There is a similar distribution for both defaulter and non defaulter.









Exploratory Data Analysis

BIVARIATE ANALYSIS :

- Analysis On Two Variable

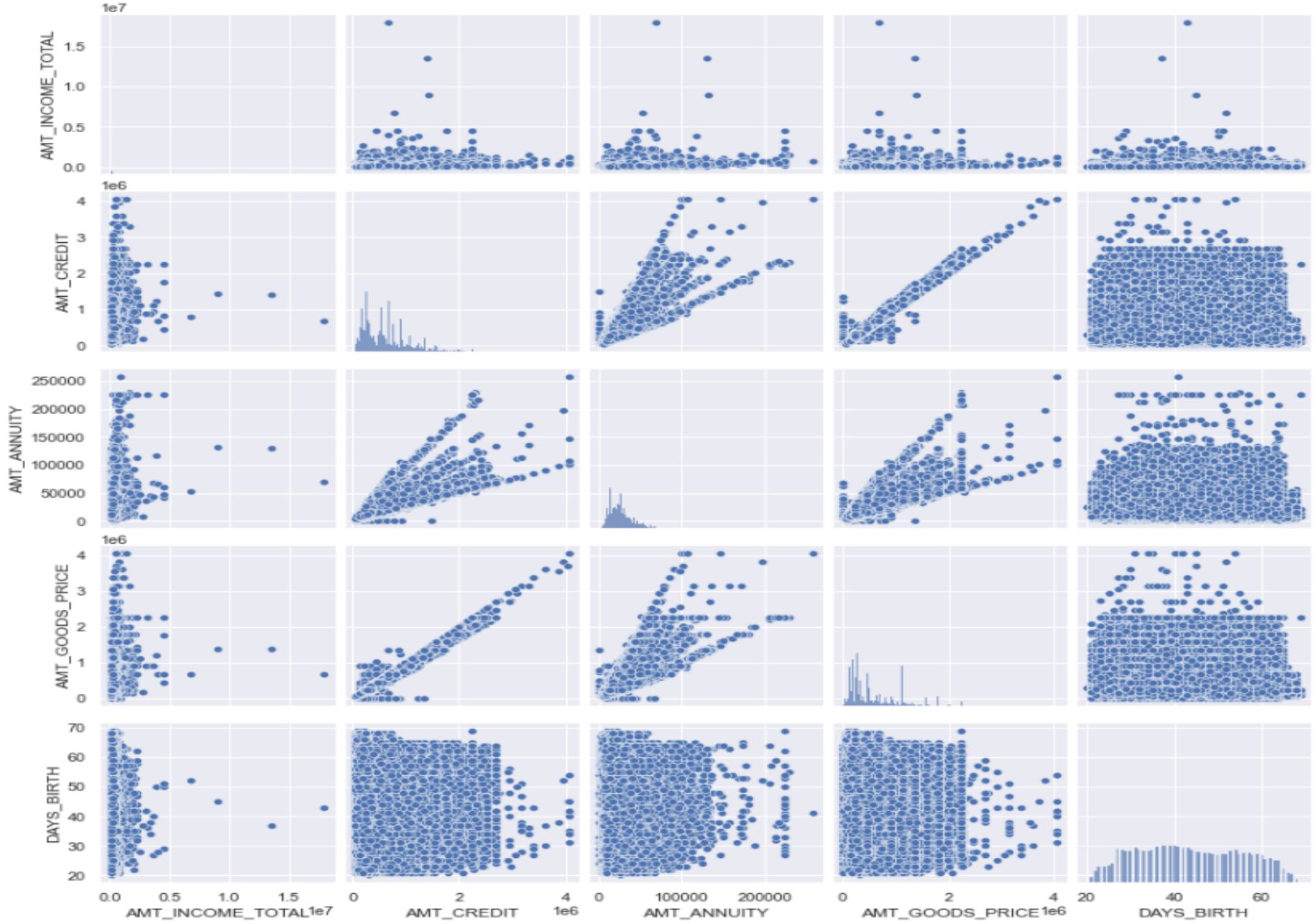
- Analysis On application_data.csv File.

- Getting Insights from the Data

Bivariate Analysis on numerical column

Pair Plot for Target =0 for following variable :

- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- DAYS_BIRTH



Bivariate Analysis on numerical column

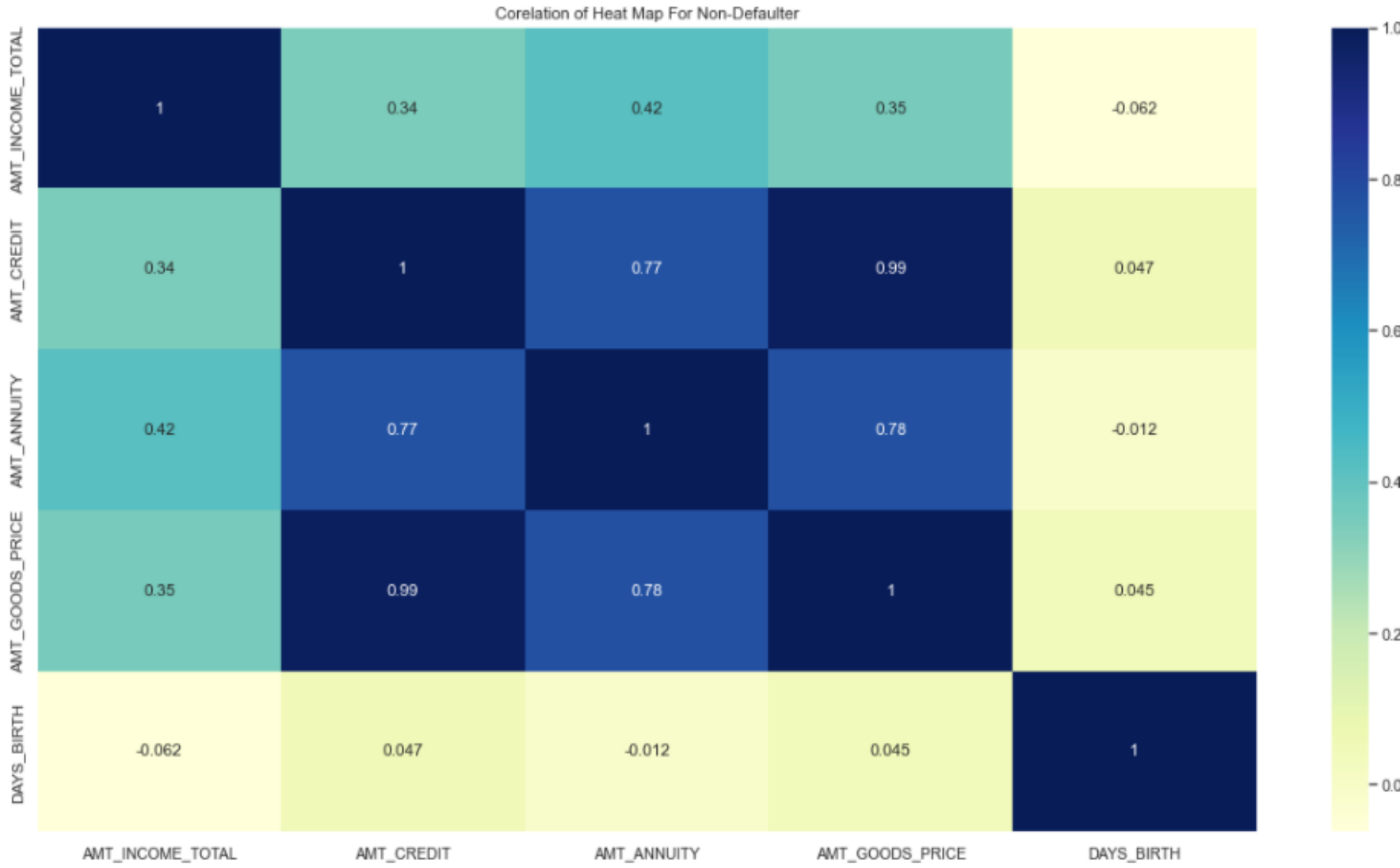
Pair Plot for Target =1 for following variable :

- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- DAYS_BIRTH

Bivariate Analysis on numerical column

Heat Map for Target =0 for following variable :

- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- DAYS_BIRTH



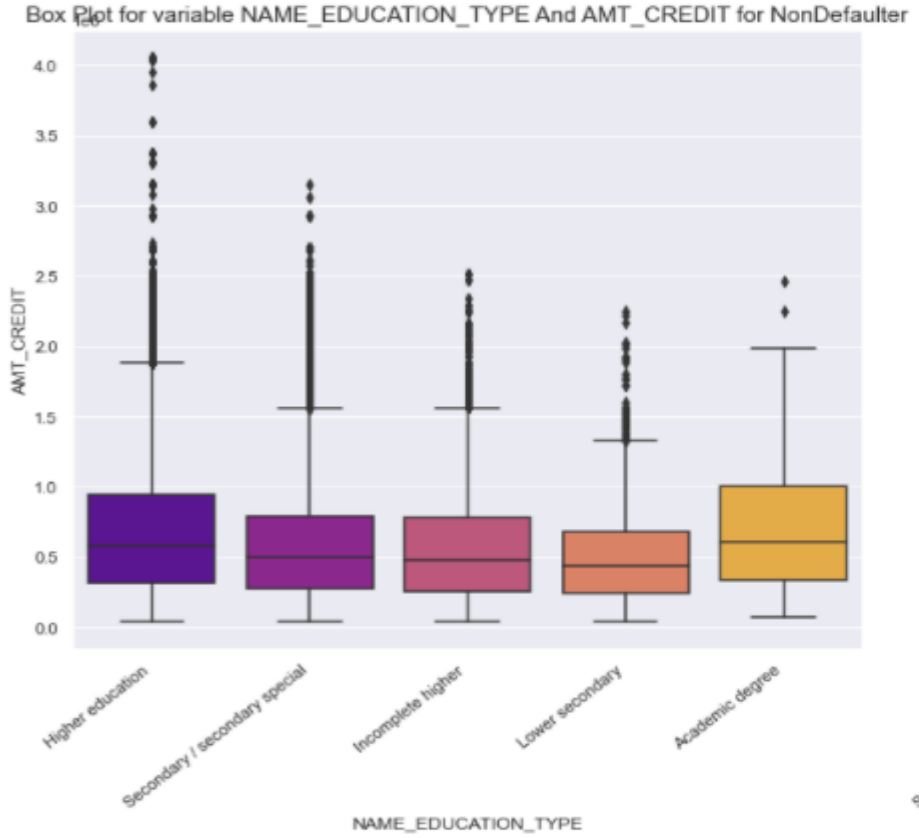
Heat Map for Target =1 for following variable :

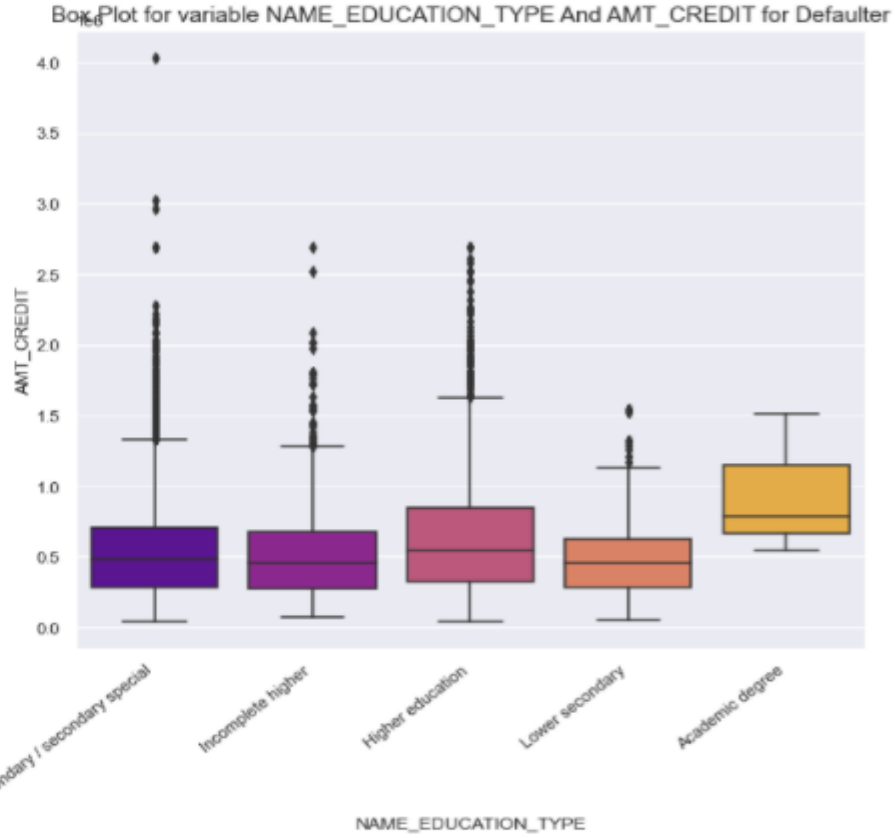
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- DAYS_BIRTH

-We observed from the above chart that there is a strong and linear correlation between AMT_CREDIT and AMT_GOODS_PRICE.
- Also there is some linear correlation between AMT_CREDIT and AMT_ANNUITY to the some extent and after that the point are scatter as the AMT_CREDIT increases.

Bivariate Analysis For Category And Continuous Variable

1) NAME_EDUCATION_TYPE VS AMT_CREDIT Variable :

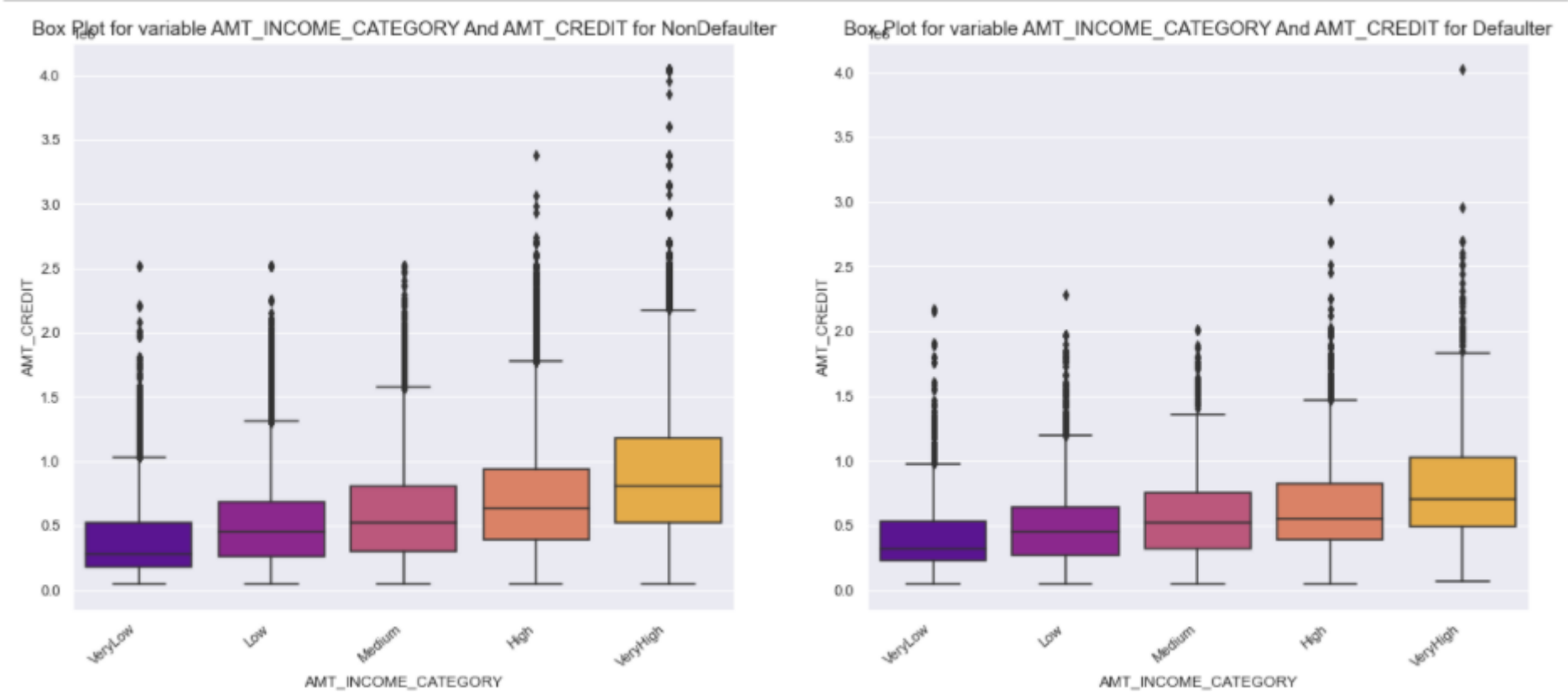




- From the above box plot we observe that the there is less chance of default if the AMT_Credit of the client is more than 3.0 for almost all the Education type except Secondary education, who has less number of client defaulter.
- Client who has credit 2.5 are less likely to default. So Company has to provide them loan by considering other factor also because some people are defaulter while some are not defaulter.

Bivariate Analysis For Category And Continuous Variable

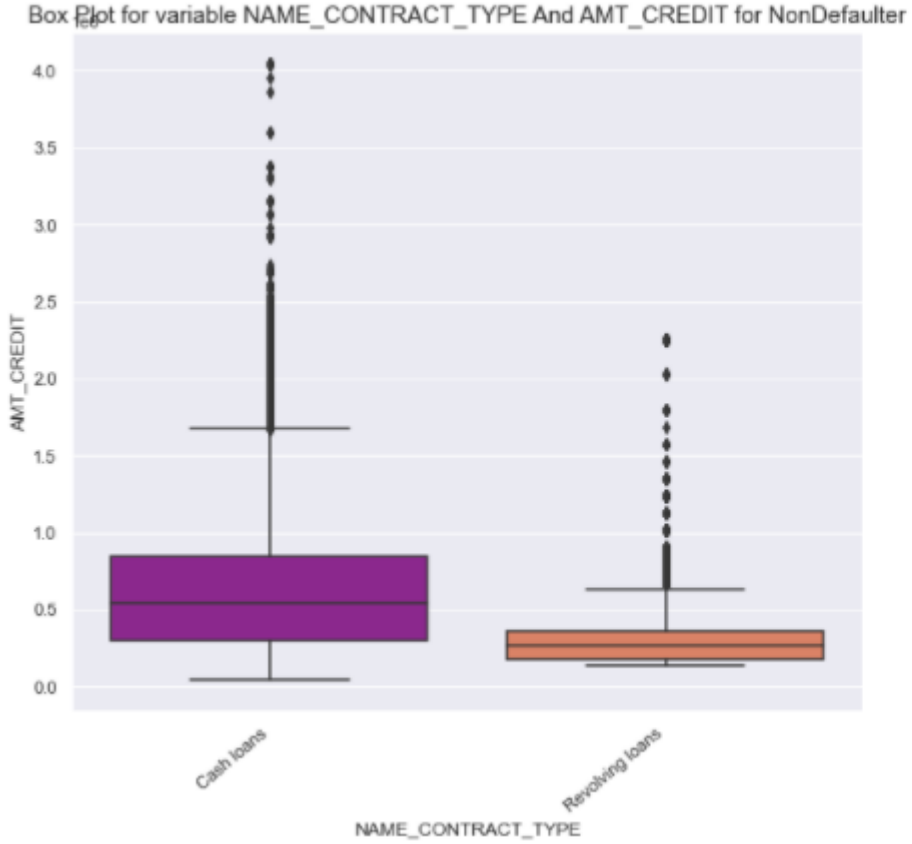
2) AMT_INCOME_CATEGORY VS AMT_CREDIT Variable :

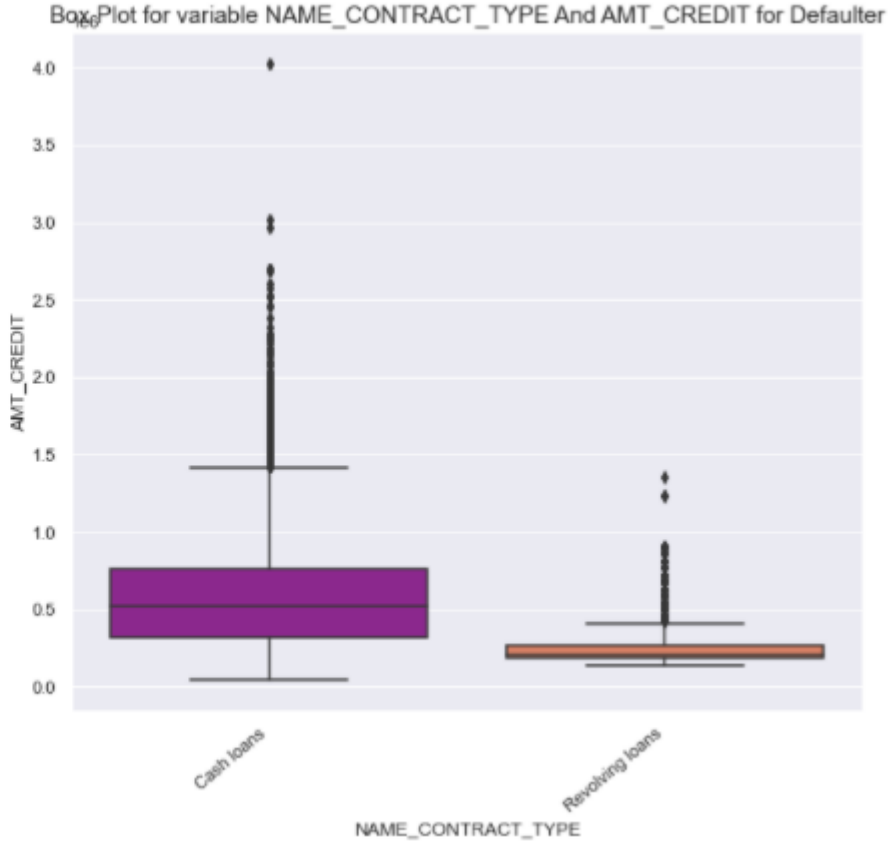


- From the above graph we conclude that for all the income category who has credit more than 3.0 as per graph have no defaulter. So they face no difficulty to pay. And there are some client who has credit less than 3.0 some are defaulter and some are not. So to give the loan company have to consider other factor also.

Bivariate Analysis For Category And Continuous Variable

3) AMT_INCOME_CATEGORY VS AMT_CREDIT Variable :

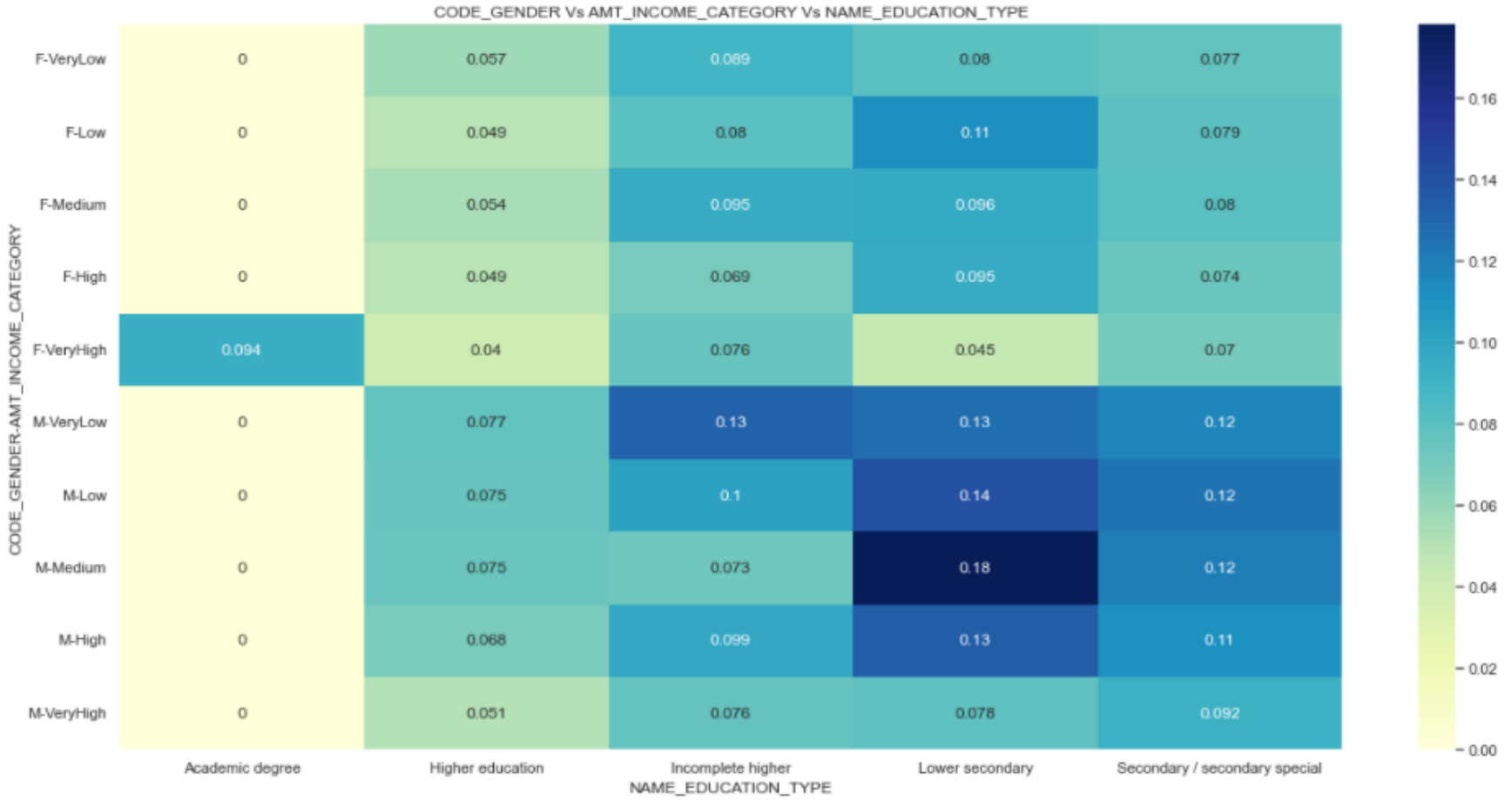




- From the above analysis we conclude that there are high chance that people who prefer cash loan and has credit less than 1.5 have more chances for Defaulter

Multivariate Analysis For Category And Category Variable

3) CODE_GENDER Vs AMT_INCOME_CATEGORY Vs NAME_EDUCATION_TYPE:



- From the table, we conclude that female who has Low income category and LowerEducation has high loan payment difficulty.
- Male who has medium income and lower secondary education has high loan paying difficulty.

TOP CORRELATION

For Target 0 dataset

	Column1	Column2	Correlation	Abs_Correlation
412	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999756	0.999756
734	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508	0.998508
190	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
639	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149	0.950149
560	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
766	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332	0.859332
191	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
159	AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309
287	DAYS_EMPLOYED	DAYS_BIRTH	0.626028	0.626028
411	FLAG_EMP_PHONE	DAYS_BIRTH	-0.621989	0.621989

For Target 1 dataset

	Column1	Column2	Correlation	Abs_Correlation
412	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999705	0.999705
734	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269	0.998269
190	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
639	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637	0.956637
560	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
766	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994	0.868994
191	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
159	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
287	DAYS_EMPLOYED	DAYS_BIRTH	0.582441	0.582441
411	FLAG_EMP_PHONE	DAYS_BIRTH	-0.578783	0.578783

- From above table we observe that the TOP 10 correlation columns are same for Target 0 and Target 1 dataframes.

Exploratory Data Analysis

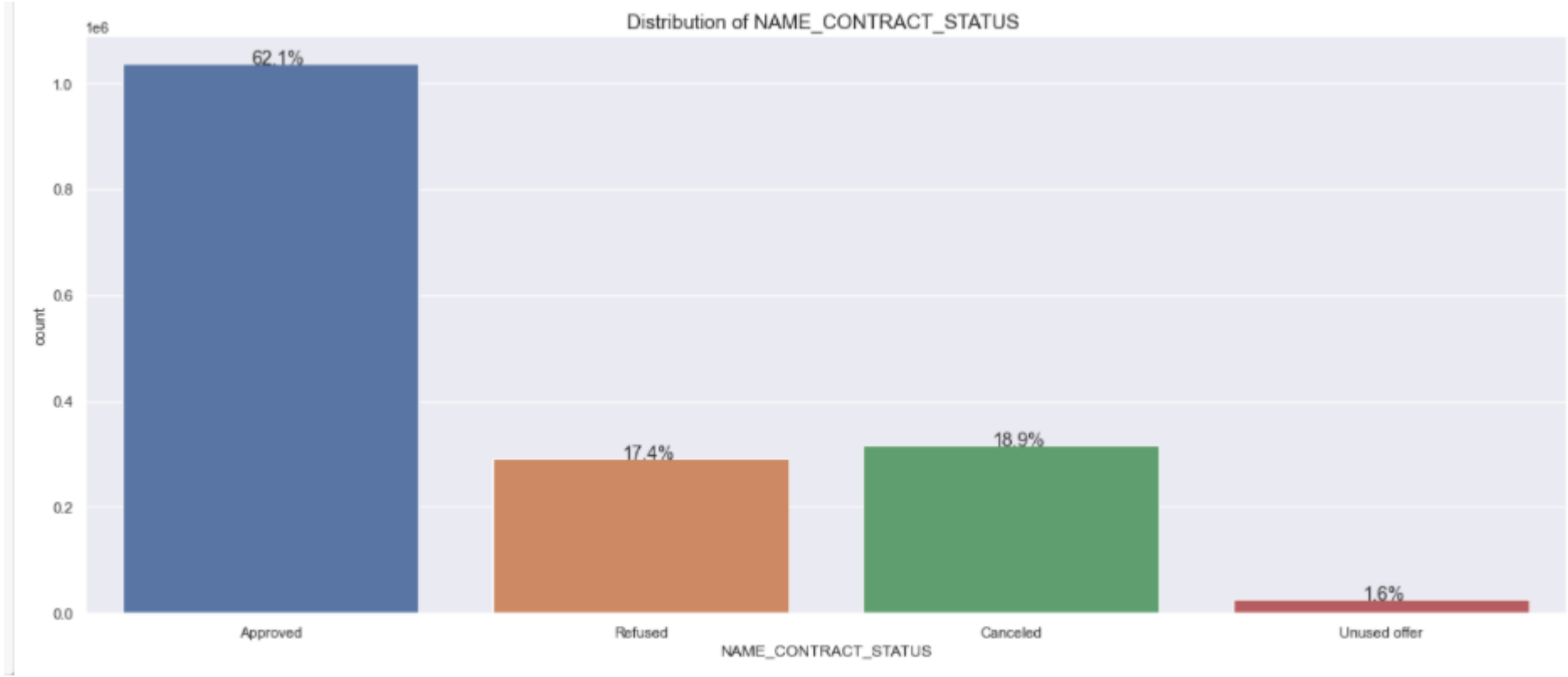
Data ANALYSIS : previous_application.csv

- Univariate Analysis, Bivariate Analysis, Multi variate Analysis

- Analysis On previous_application.csv file.

Categorical Variable Analysis

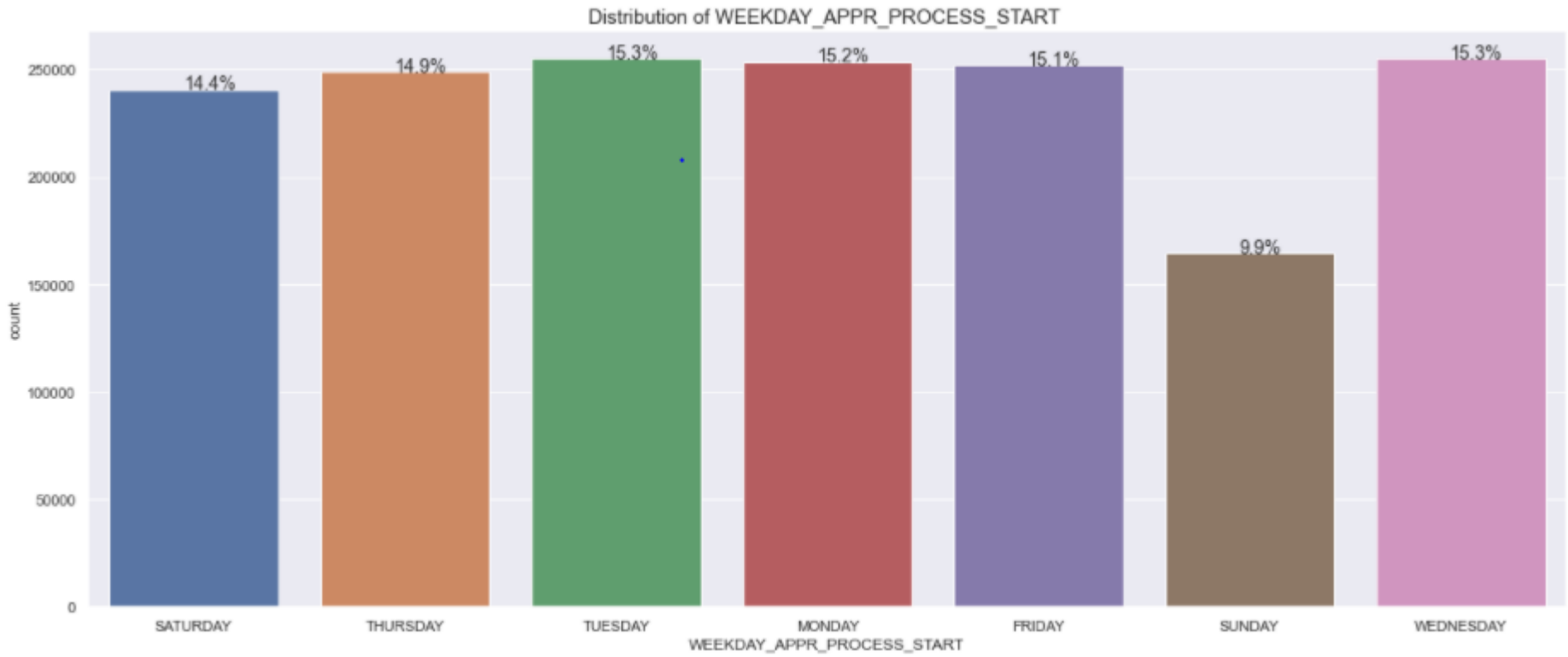
1) NAME_CONTRACT_STATUS Variable :



From the above graph we can conclude that:
- We can easily observed that the majority of loans are approved and very less percentage of loan which is 1.6% are unused offer and 17.4% of loan application are refused and 18.9% loan are canceled

Categorical Variable Analysis

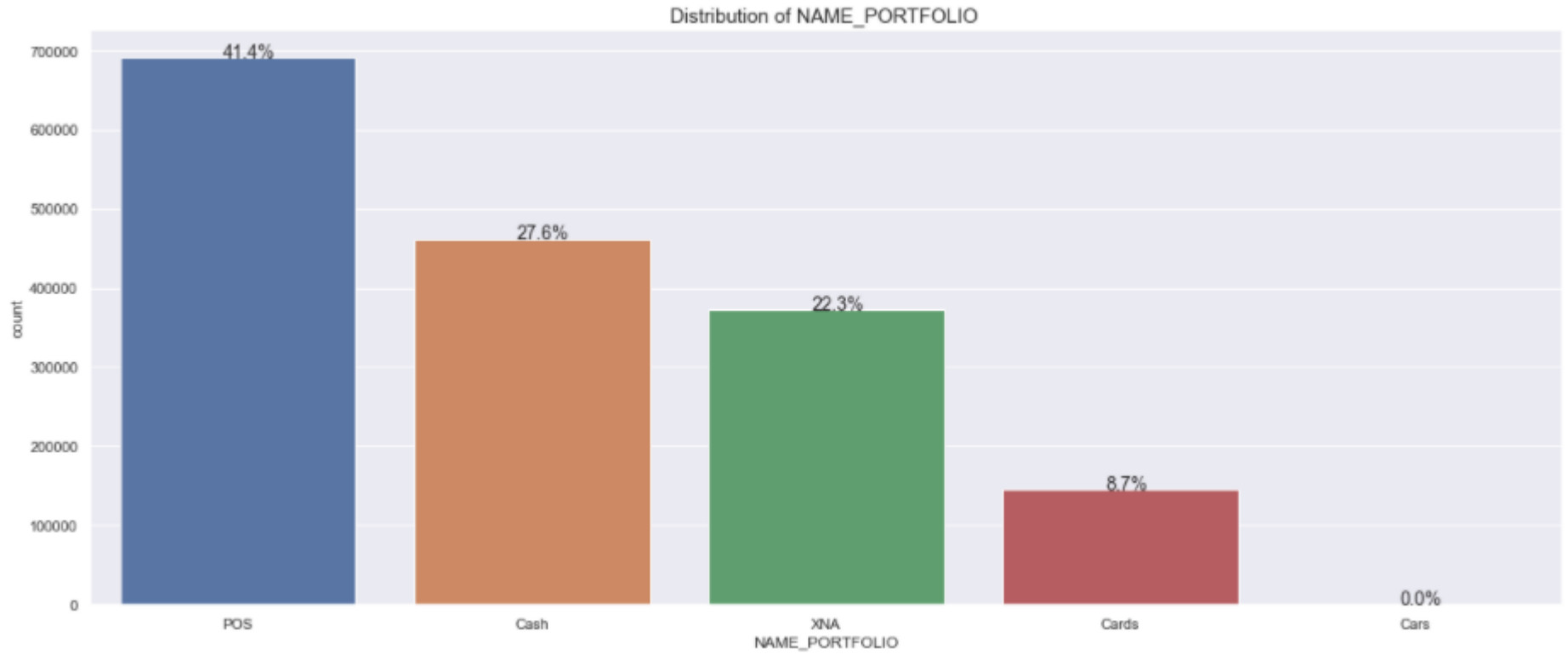
2) WEEKDAY_APPR_PROCESS_START Variable :



From the above graph:
- We can observe that there are less number of people apply on Sunday as compared to other days,which is similar.

Categorical Variable Analysis

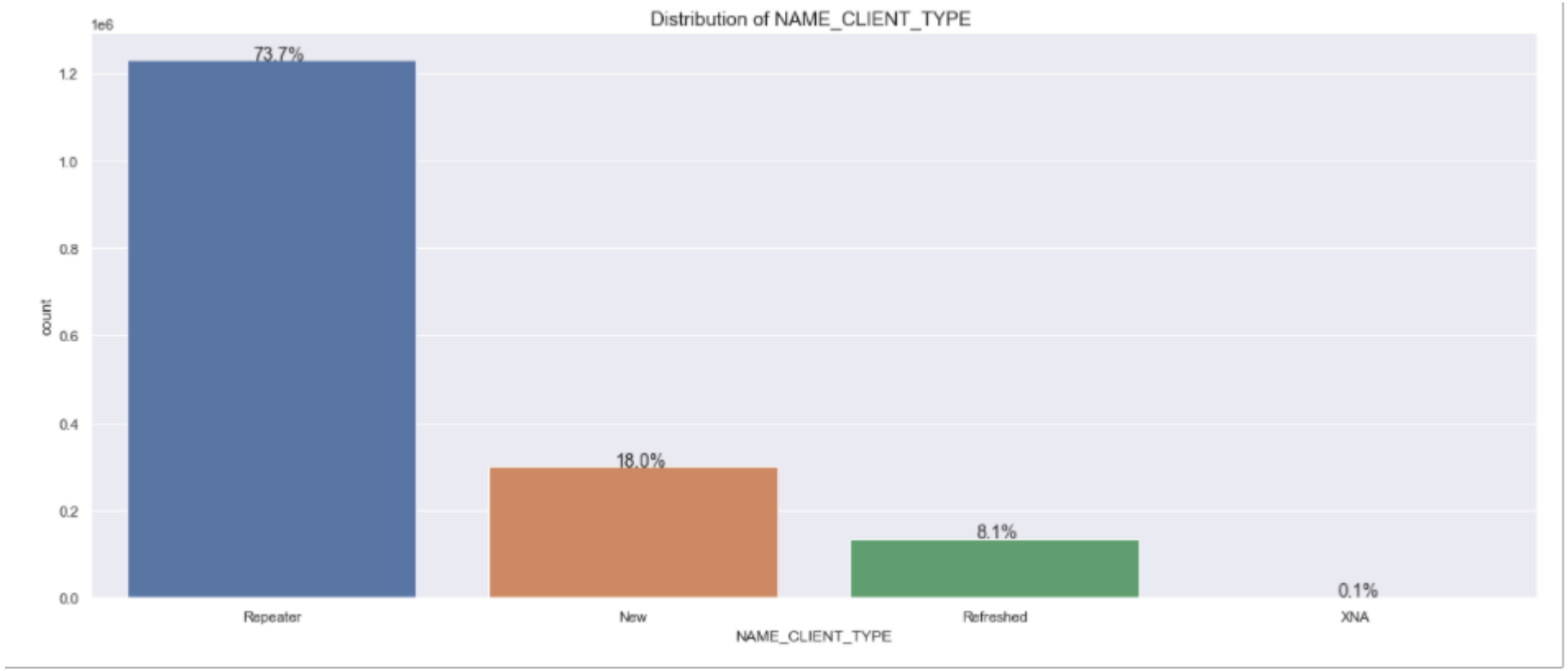
3) NAME_PORTFOLIO Variable :



From the above graph:
- We observe that people prefer to apply for POS which is around 41% and also a fair amount of people apply for cash as well.

Categorical Variable Analysis

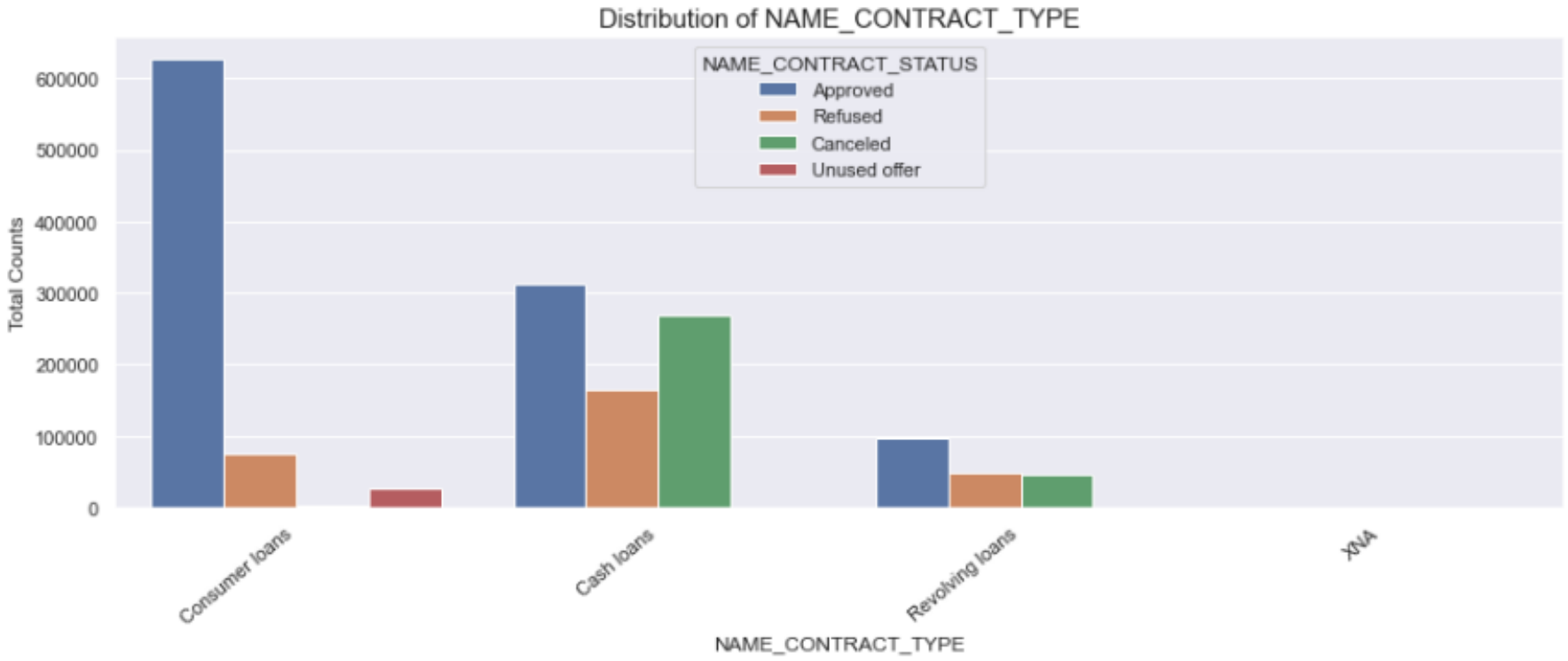
4) NAME_CLIENT_TYPE Variable :



From the above graph:
- We conclude that almost 74% client are repeater and 18% are New client.

Bivariate Categorical Variable Analysis

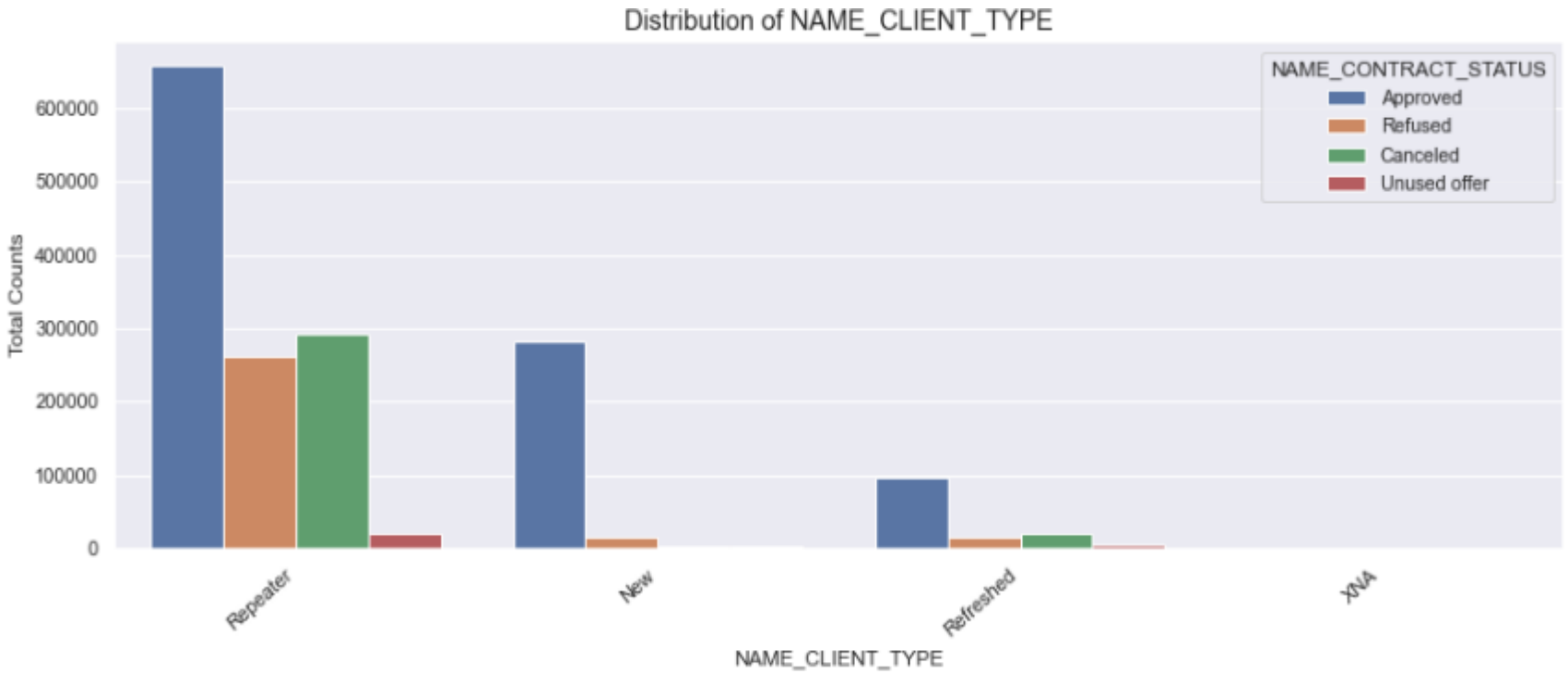
5) NAME_CONTRACT_TYPE Variable :



From the above graph:
- From the above chart, we analyze that the consumer loan and cash loans application are high. Although, the cash loans are refused more than others.

Bivariate Categorical Variable Analysis

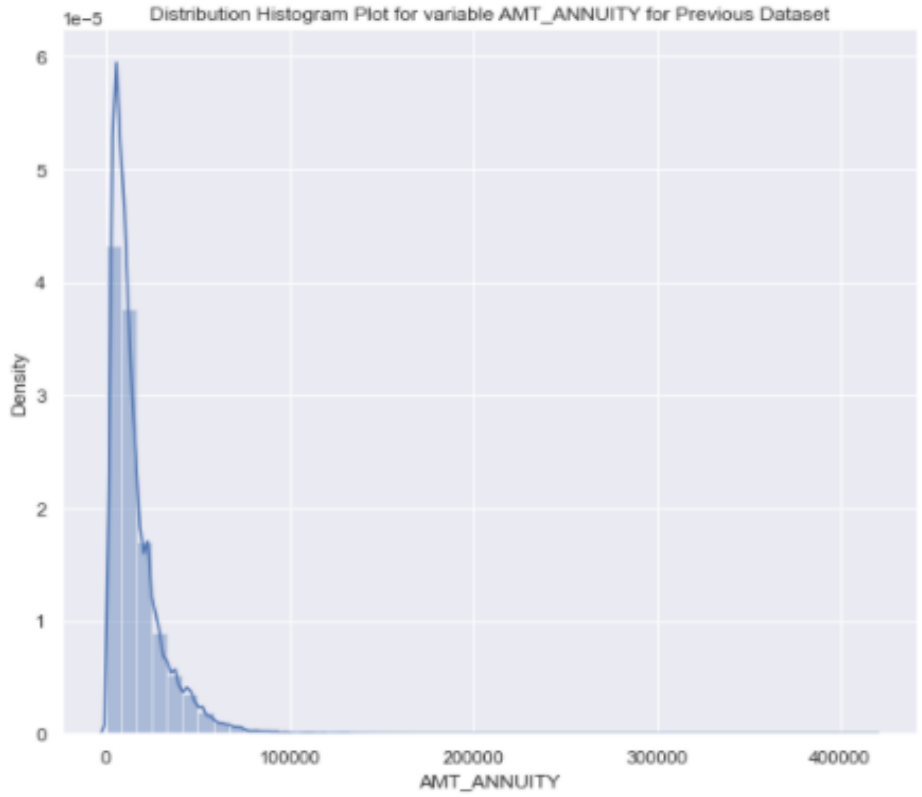
6) NAME_CONTRACT_TYPE Variable :

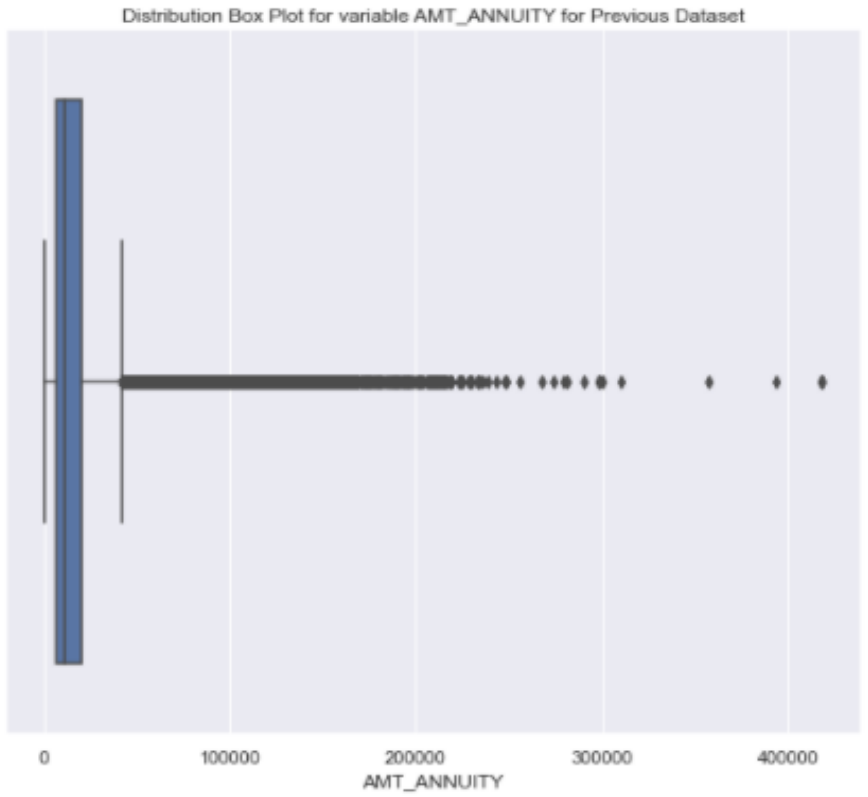


From the above graph:
- Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters. They also get refused most often.

Univariate analysis of numerical columns

1) AMT_ANNUITY Variable

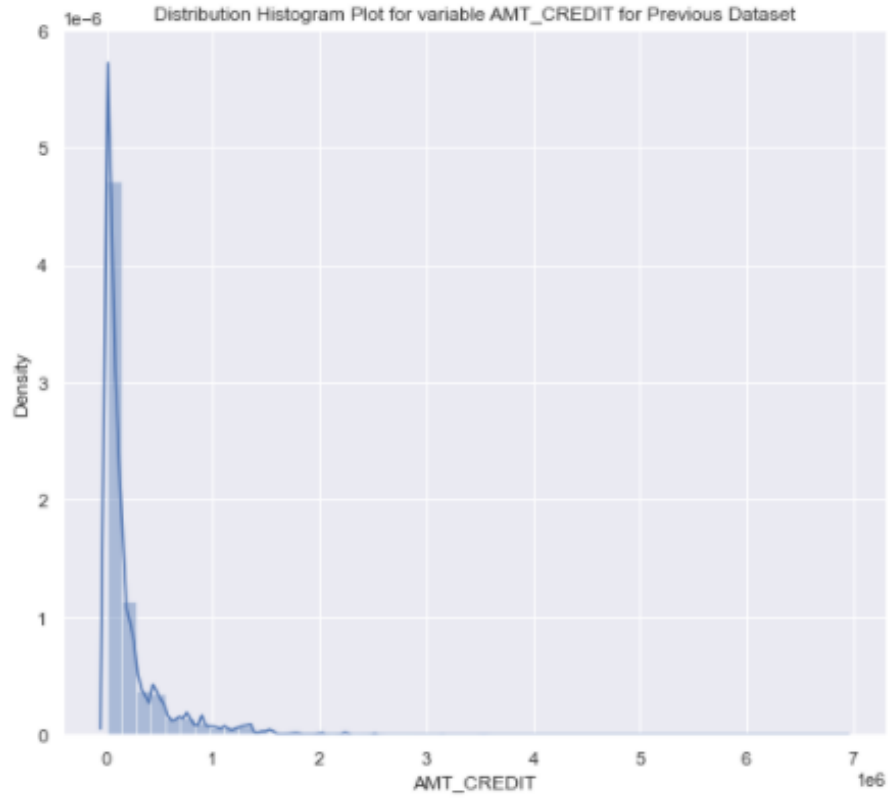


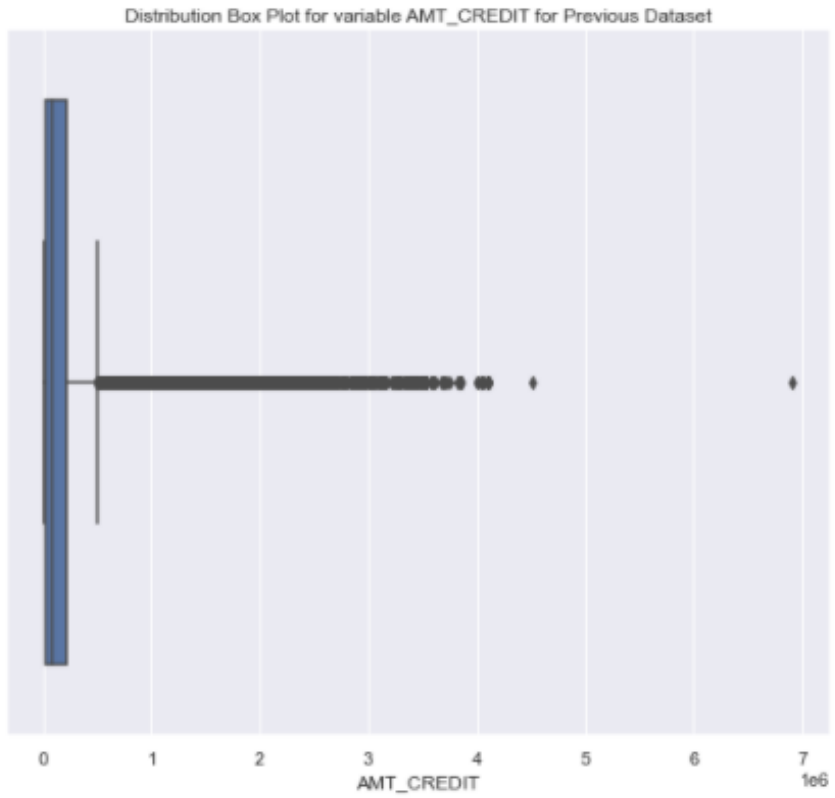


- We observe that the major distribution of AMT_ANNUITY is less than 50000. There are some outliners in the boxplot and the curve is not normal.

Univariate analysis of numerical columns

2) AMT_CREDIT Variable

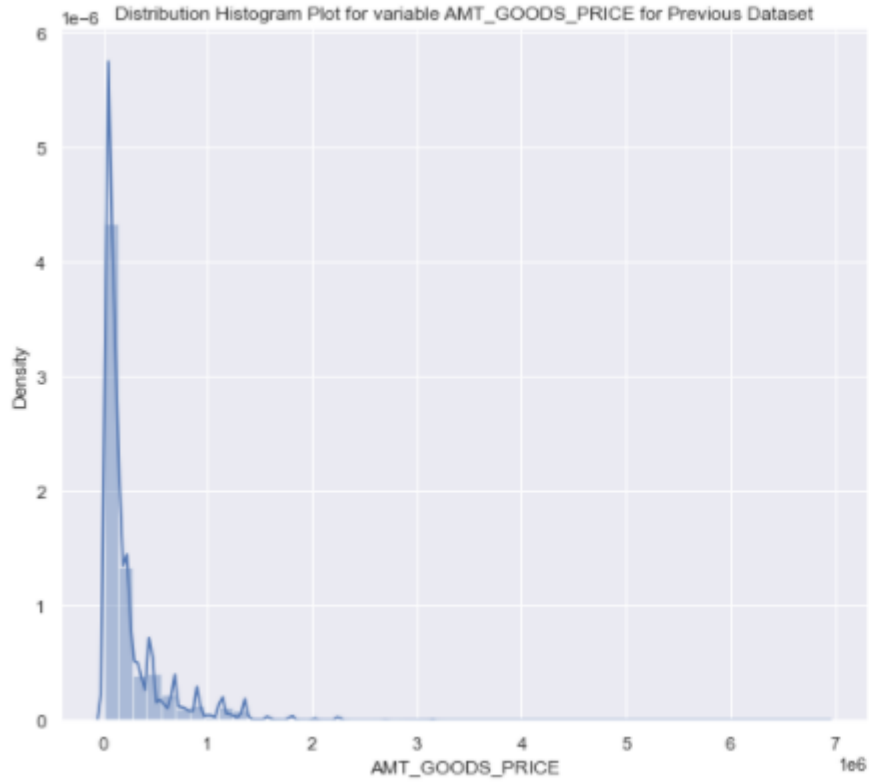


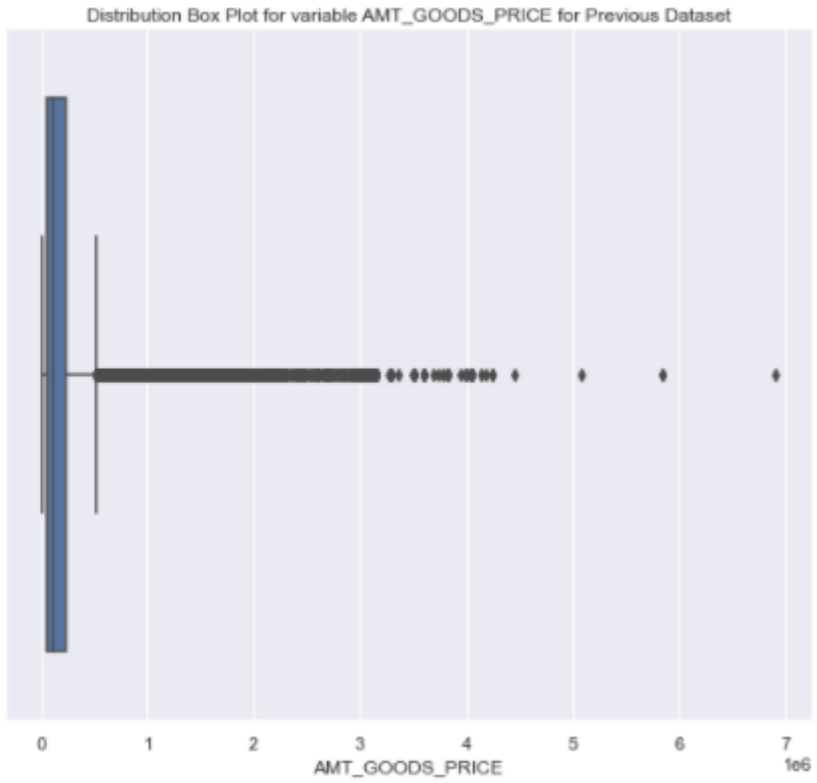


- We observe that the major distribution of AMT_CREDIT is less than 0.5 as per graph values. There are some outliners in the boxplot and the curve is not normal.

Univariate analysis of numerical columns

3) AMT_GOODS_PRICE Variable



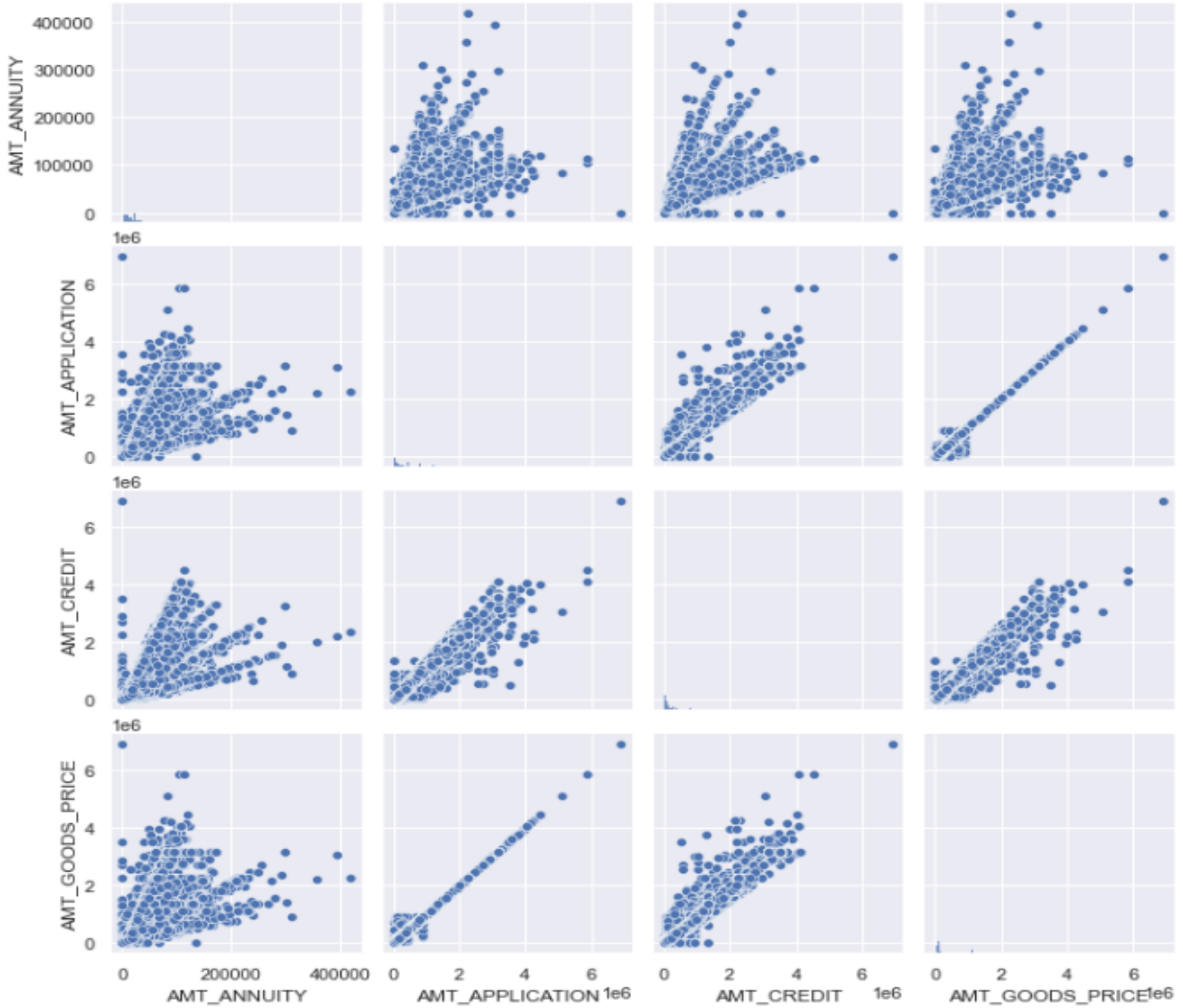


- There are some outliners in the boxplot and the curve is not normal..

Bivariate Analysis On Numerical Column

Pair Plot for following variable :

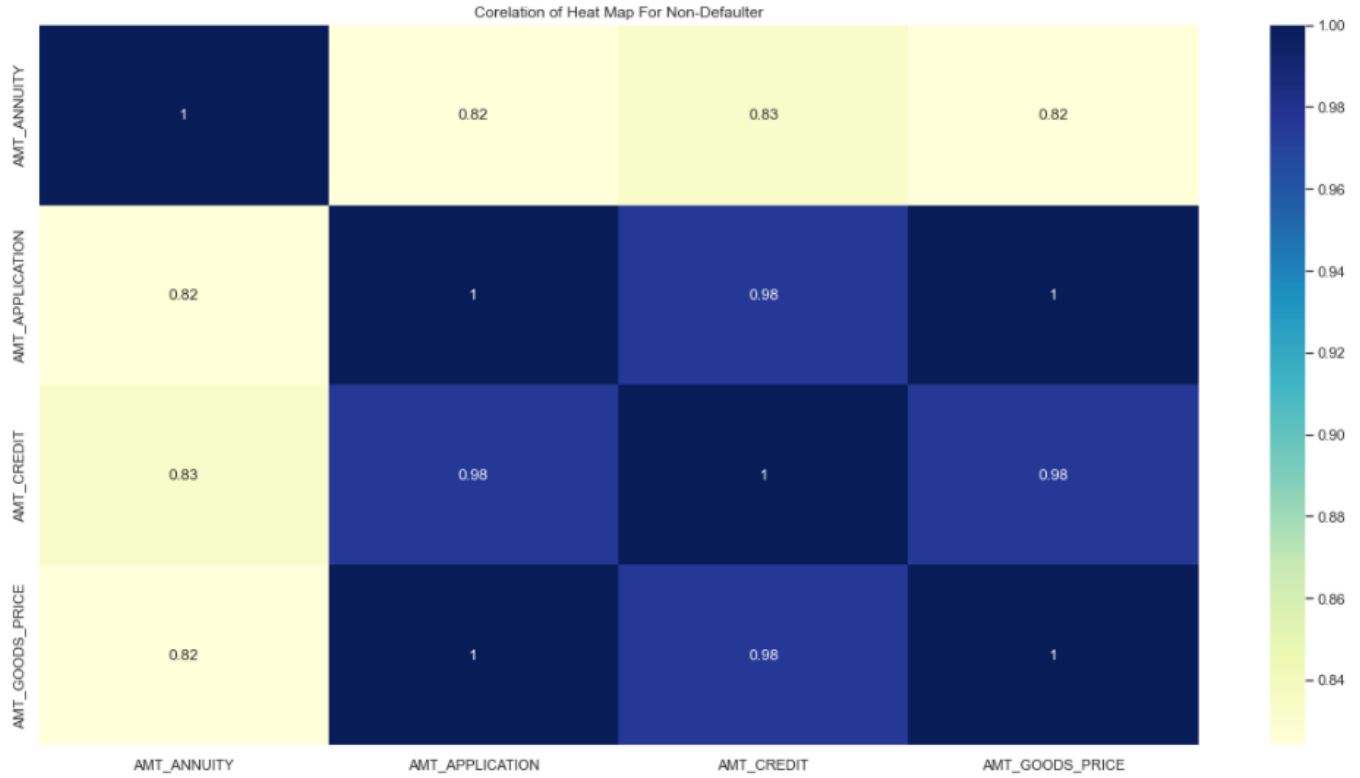
- AMT_APPLICATION
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE



Bivariate Analysis On Numerical Column using Heatmap

Heatmap for following variable :

- AMT_APPLICATION
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE



From the Above charts, we observe that there is a strong correlation between AMT_APPLICATION and AMT_GOOD_PRICE. Also , there is a strong correlation between AMT_CREDIT and AMT_APPLICATION.

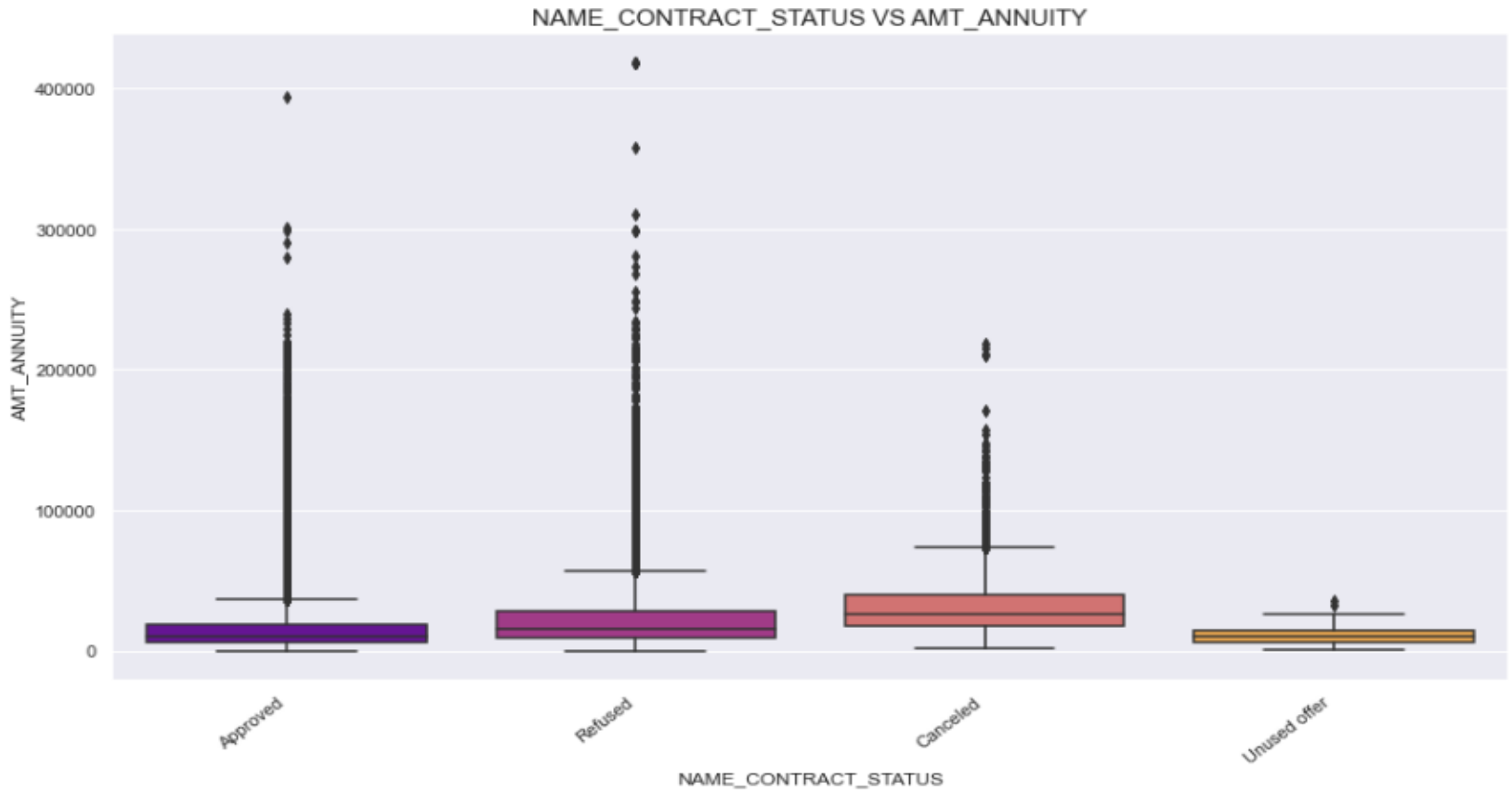
- Increase of annuity increases by the below factors
(1) How much credit did client asked on the previous application (AMT_APPLICATION)
(2)Final credit amount on the previous application that was approved by the bank (AMT_CREDIT)
(3) Goods price of good (AMT_GOODS_PRICE) that client asked for on the previous application.

- For how much credit did client ask on the previous application(AMT_APPLICATION) is highly influenced by the Goods price of good(AMT_GOODS_PRICE) that client has asked for on the previous application

- Final credit amount disbursed to the customer previously(AMT_CREDIT), after approval is highly influence by the application amount(AMT_APPLICATION) and also the goods price of good (AMT_GOODS_PRICE) that client asked for on the previous application.

Bivariate Analysis For Category And Continuous Variable

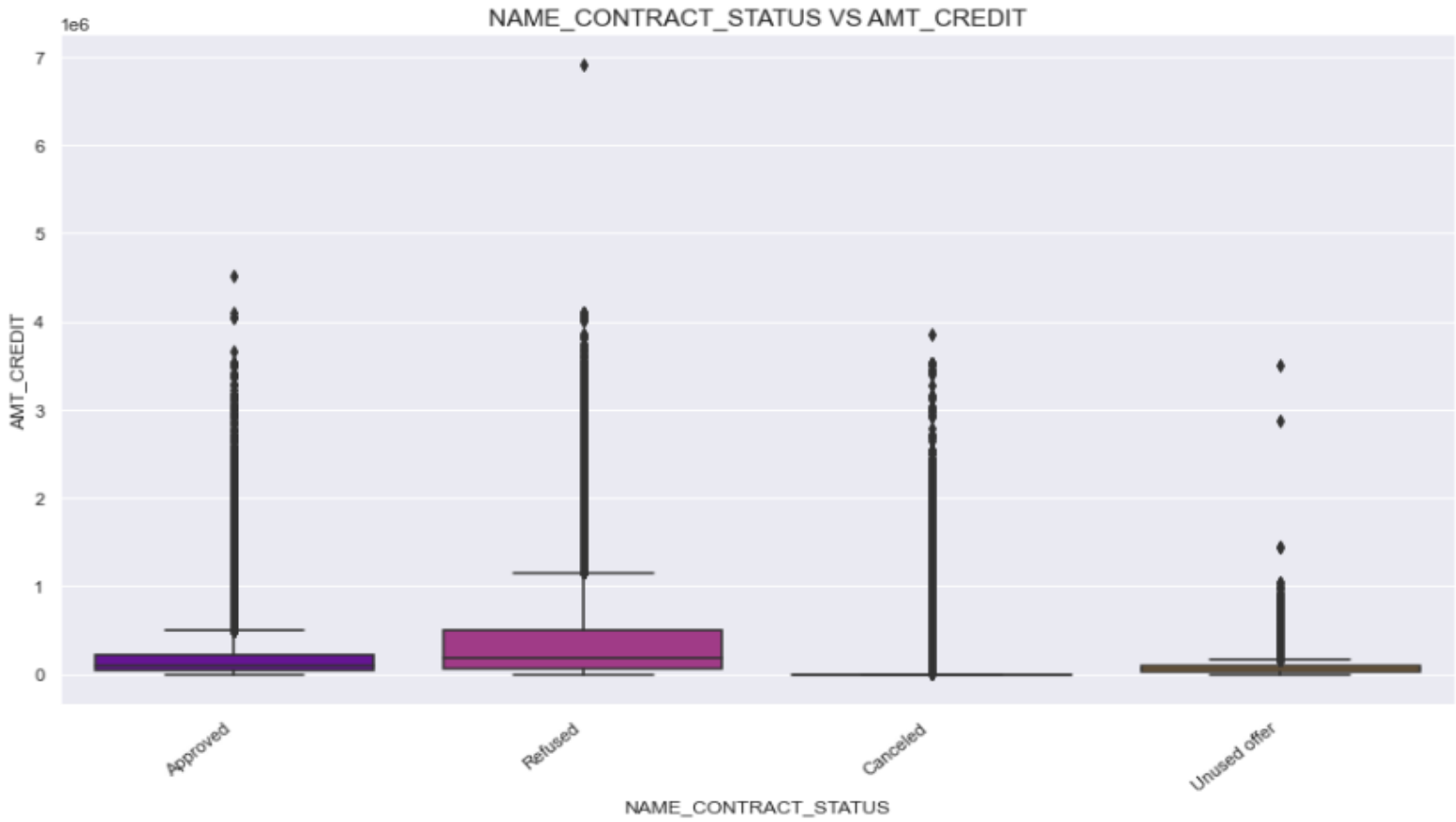
1) NAME_CONTRACT_STATUS vs AMT_ANNUITY



- From the above plot we can see that loan application for people with lower AMT_ANNUITY gets canceled or Unused most of the time.
- We also see that applications with too high AMT ANNUITY also got refused more often than others.

Bivariate Analysis For Category And Continuous Variable

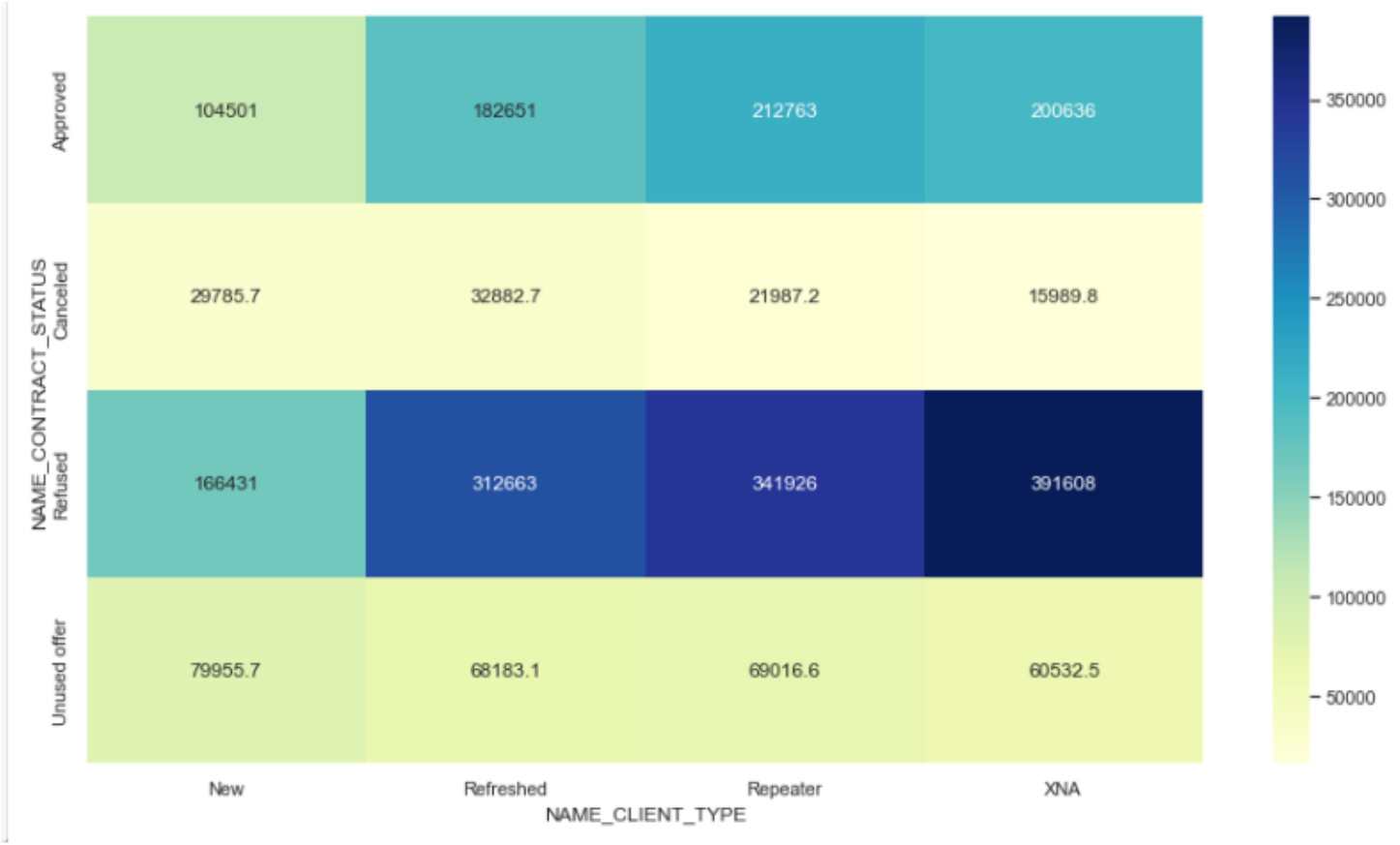
2) NAME_CONTRACT_STATUS vs AMT_CREDIT



- From the above graph we can infer that when the AMT_CREDIT is too low, it get's cancelled/unused most of the time.

Multivariate Analysis

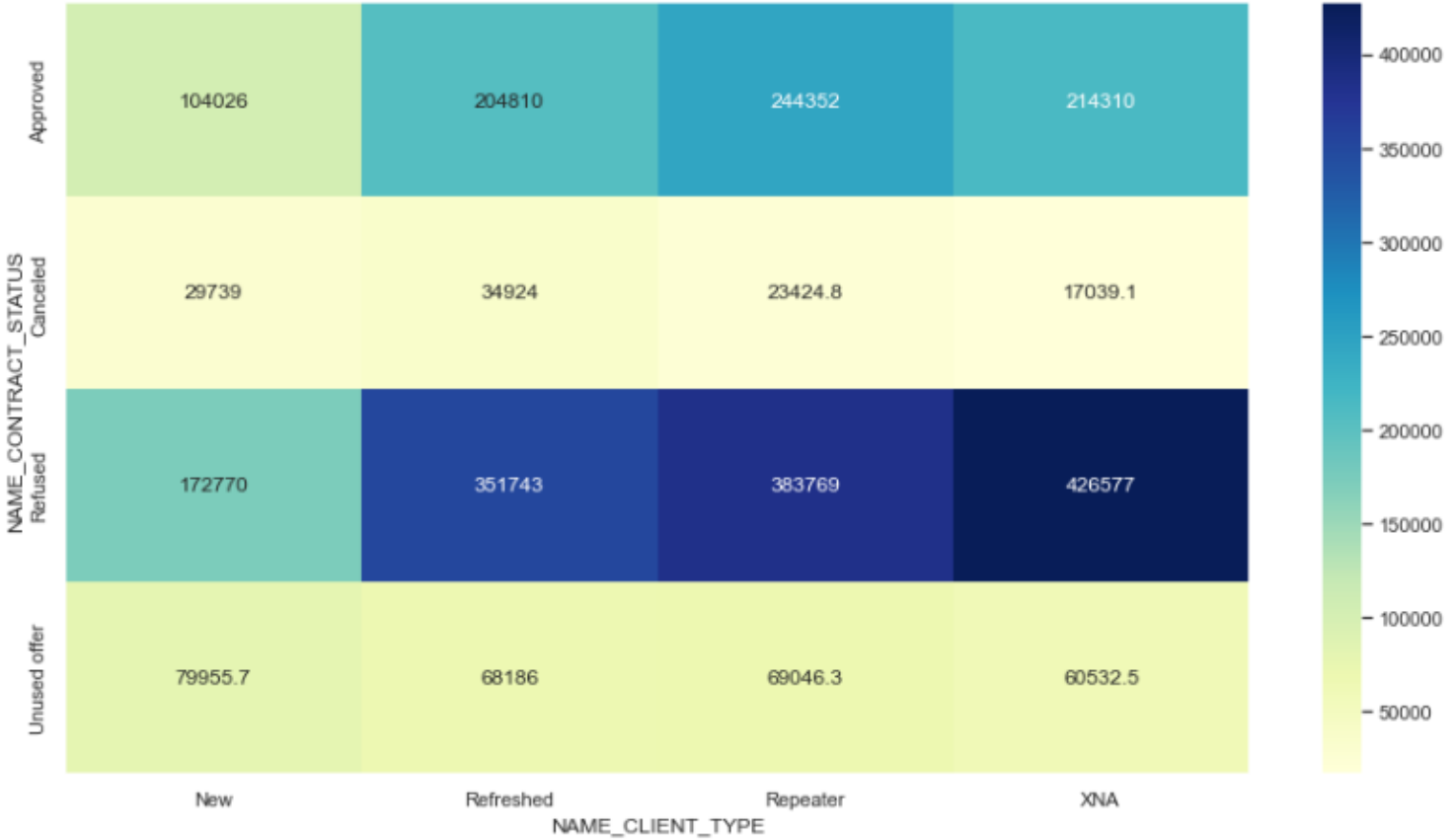
1) NAME_CONTRACT_STATUS vs NAME_CLIENT_TYPE vs AMT_APPLICATION:



- Unused offer as very low application amount for almost all the client type.
- Cancelled and Refused application amount is high. The bank may be refusing these application because possible client has very low credit amount.
- Repeater's application amount is higher than the New client. This indicate that may be bank has more conducted attractive policies/rate of interest for repeat applicants.

Multivariate Analysis

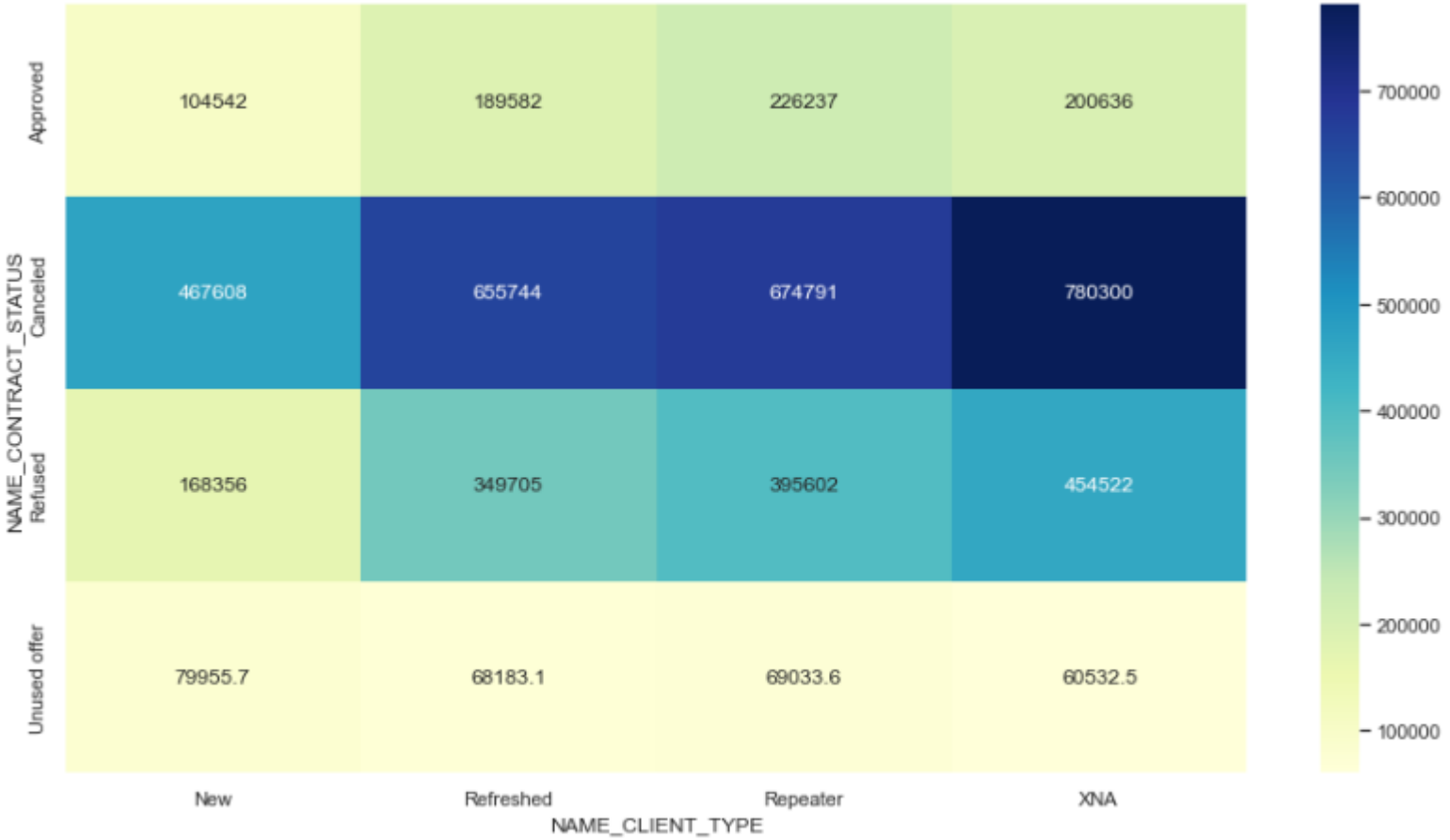
2) NAME_CONTRACT_STATUS vs NAME_CLIENT_TYPE vs AMT_CREDIT:



- Not able to understand why cancelled and refused application has credit amount?

Multivariate Analysis

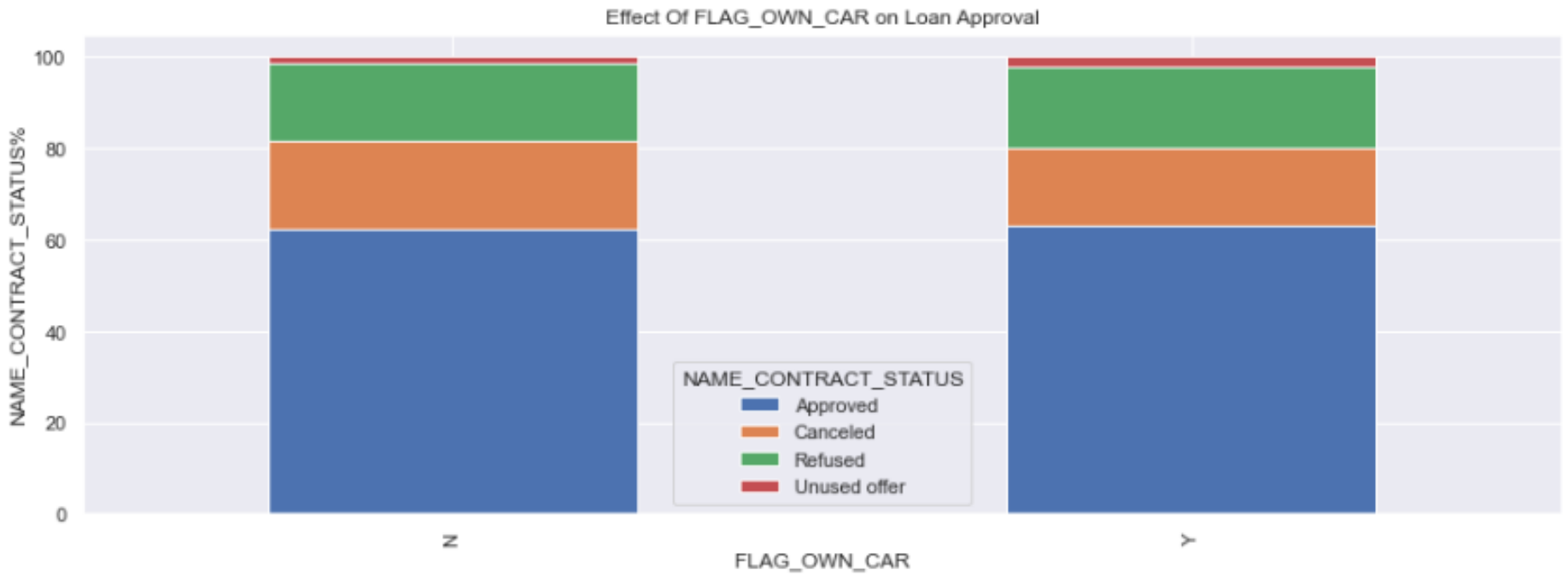
3) NAME_CONTRACT_STATUS vs NAME_CLIENT_TYPE vs AMT_GOODS_PRICE:



- From the above we concluded that all cancelled and refused cases have higher value of goods than other categories. This may be because they have low credit score.

Merge DataFrame Analysis

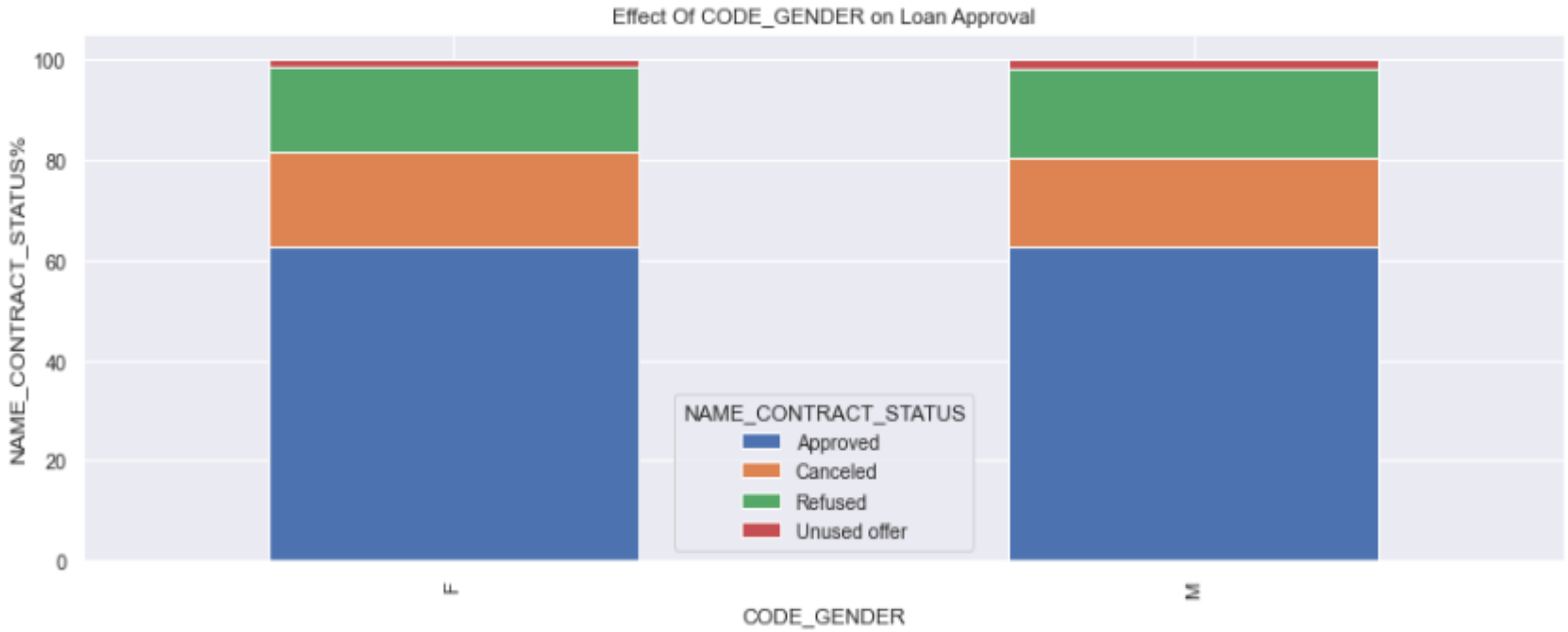
1) Effect Of NAME_CLIENT_TYPE on Loan Approval:



- From the above stacked bar chart we see that car ownership does not have any effect on application approval or rejection. But previous we observe that client who has a car has lesser chances of defaulter. For the higher loan amount , the bank can consider car ownership to accept the loan application.

Merge DataFrame Analysis

2) Effect Of CODE_GENDER on Loan Approval:



- From the above graph, we observe that code gender doesn't have any effect on application approval or rejection. But we observed earlier that female have lesser chances of default compared to males. So the bank can consider more weightage to female while approving a loan amount.

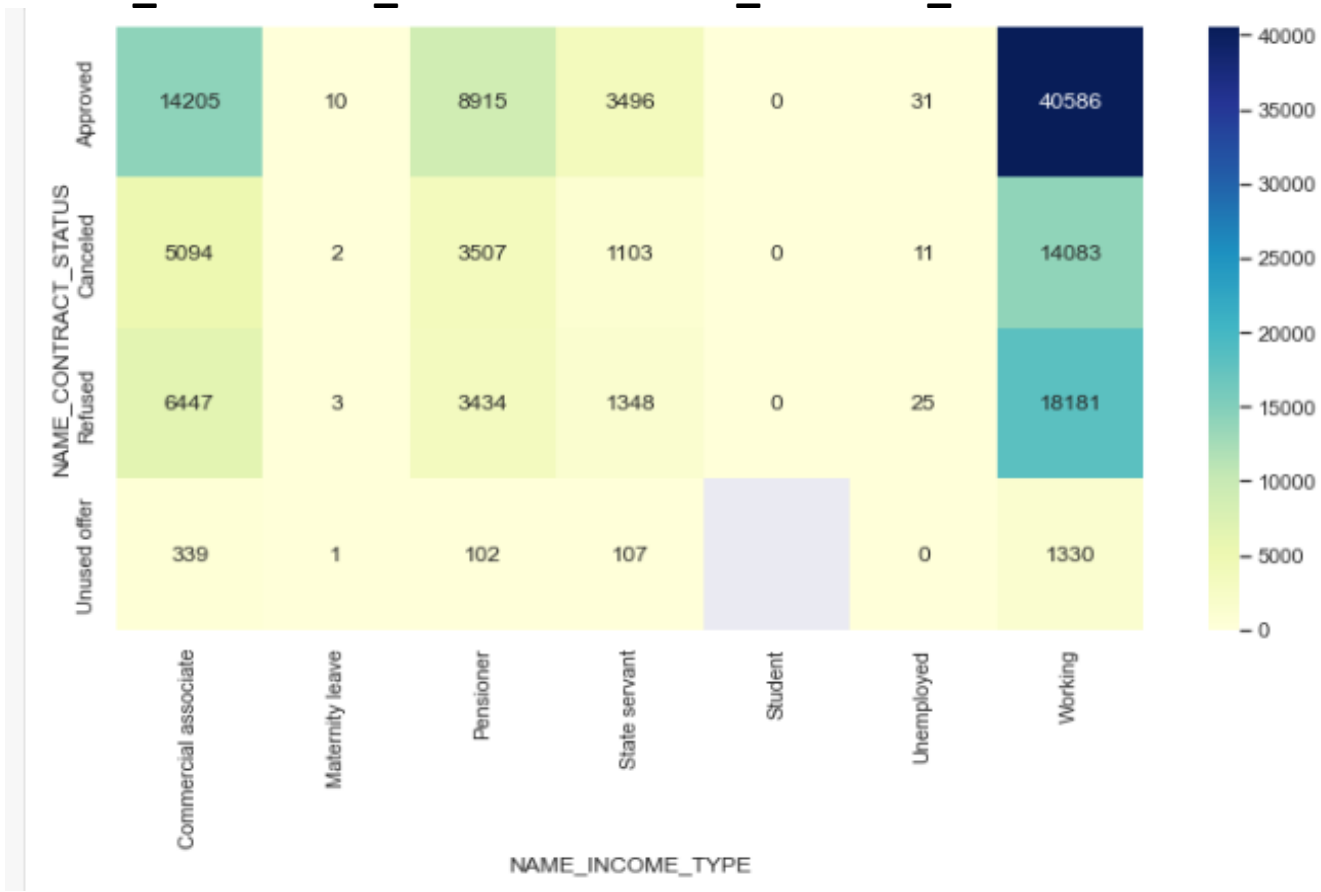
Merge DataFrame Analysis

3) Effect Of TARGET on Loan Approval:

- From the above We observe that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.

Merge DataFrame Analysis

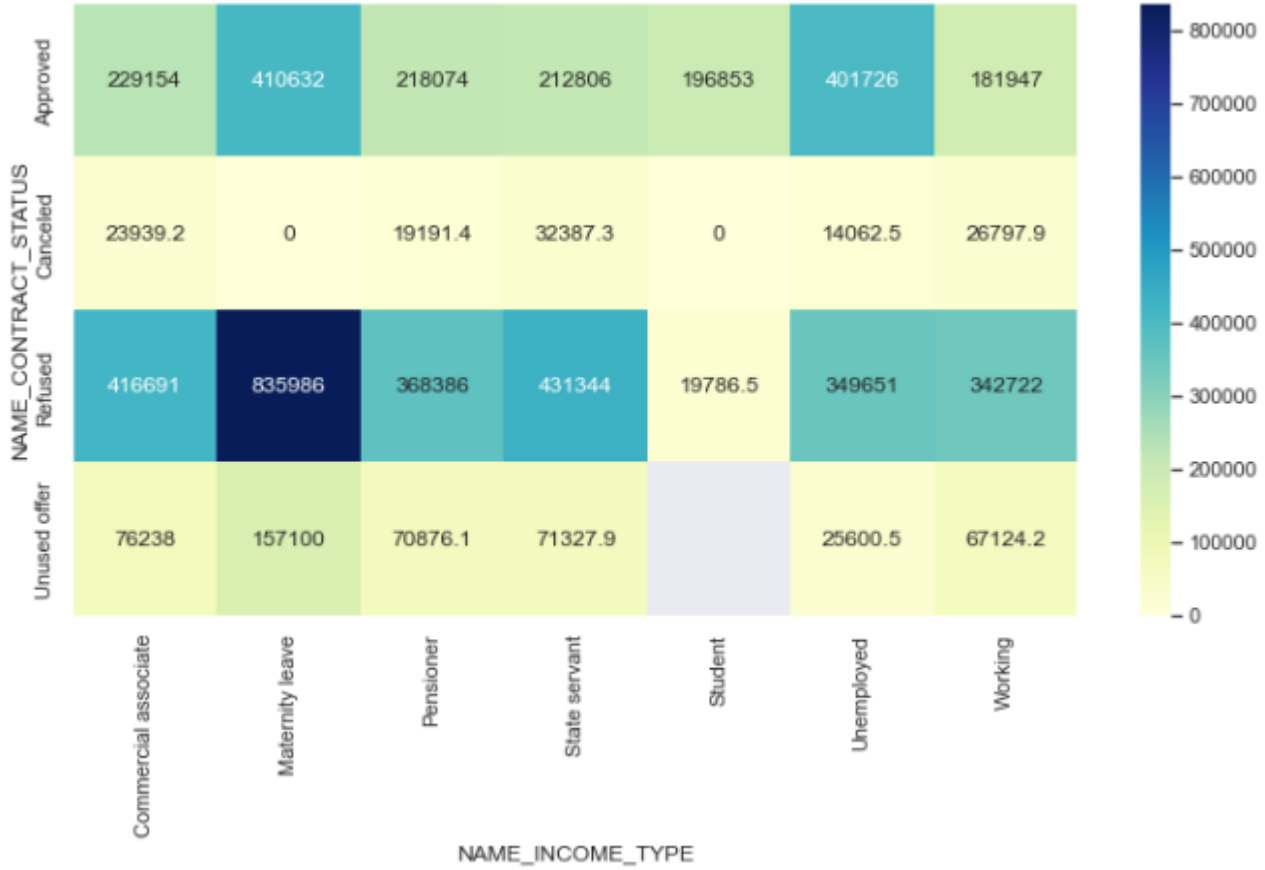
4) NAME_CONTRACT_STATUS Vs NAME_INCOME_TYPE Vs TARGET:



- Since Target 1 is defaulter, higher on the above matrix shows correlation to defaulter.
- Previous applications with Cancelled , Unused,Refused loans also have default. This indicates that the financial company had Refused/cancelled previous application, but has approved the current loan application and is facing default on these loans.
- Working client with Approved status have highest number in defaulted as compared to others.

Merge DataFrame Analysis

5) NAME_CONTRACT_STATUS Vs NAME_INCOME_TYPE Vs AMT_CREDIT_y :



- It is noticeable point that higher credit is offered to unemployment client and maternity leave.
- Client whose previous application is refused also has a fairly high credit except student.

SUMMARY

Insights For Defaulter:

- Female who has Low Income category and Lower Education has a high chances of Defaulter while for Male who has medium income and lower secondary education has high chances of Defaulter.
- The Age group in between 28 to 40 has higher chances to be a defaulter. So the bank should have to scritinize the other factor for giving the loan.
- Some married people pay on time while some are facing difficulty in paying. So the bank should have to take a other factor also while approving the loan.
- Working class people are more defaulter but some working class people pay on time. So bank have to consider the other factor also while approving the loan.
- Previous applications with Cancelled, Refused, Unused loans also have default which is a matter of concern. This indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing defaulter.
- Client whose previous application is cancelled, Refused are higher chances of a defaulter.
- Laborers, Drivers , Low Skilled labors are more chances of being an defaulter.
- Client who has without cars are fair chances of being an defaulter as compared to those who has cars but most of them are paying on time. So bank should have to consider other factor also.

Insights For Non-Defaulter:

- Bank should give more weightage to Female as they are regular pay on time.
- The senior citizen person who has more than 45 years of age are higher chances of non dafaulter.
- Business Man, Students are very less number of defaulter.
- Client whose previous application is approved are regular paying on time.