Credit EDA Case Study

Credit EDA Case Study

This case study aims to identify patterns to understand the driving factors behind loan default.

By :

Akshay Joshi & Chinmay Gaikwad

Objectives:

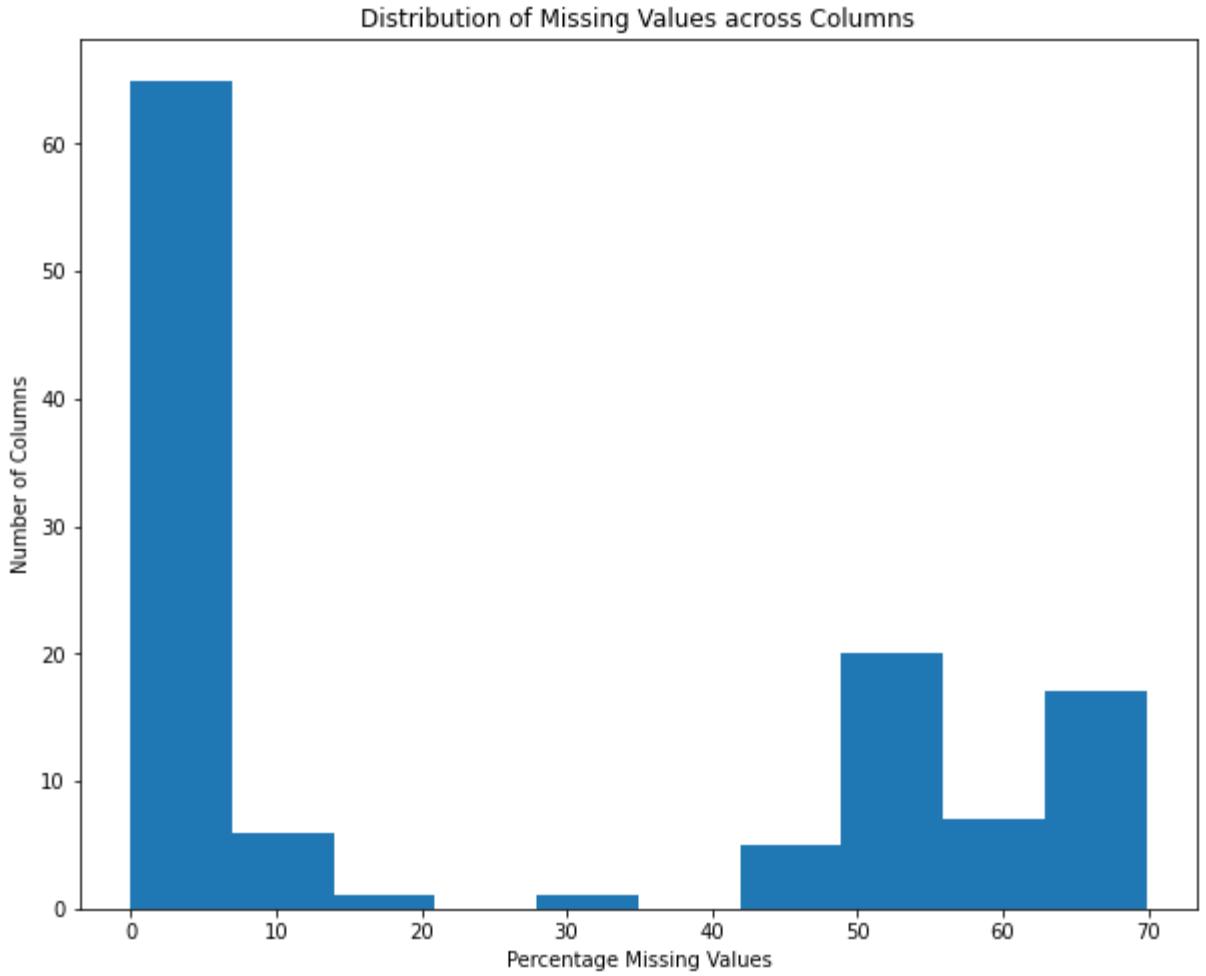
1.This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

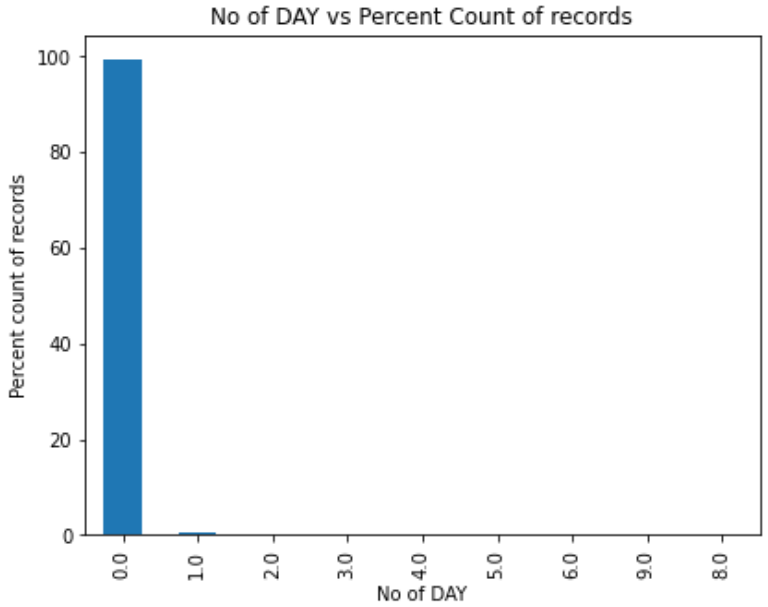
2. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

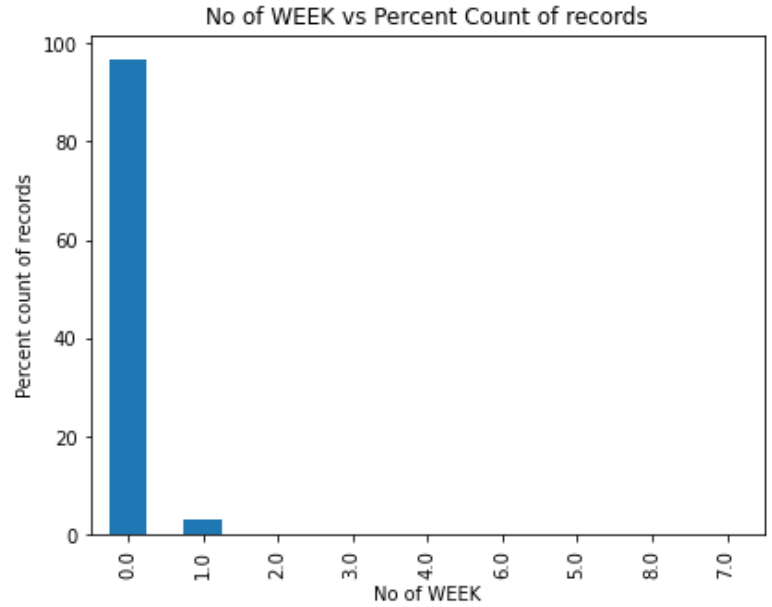
3.The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

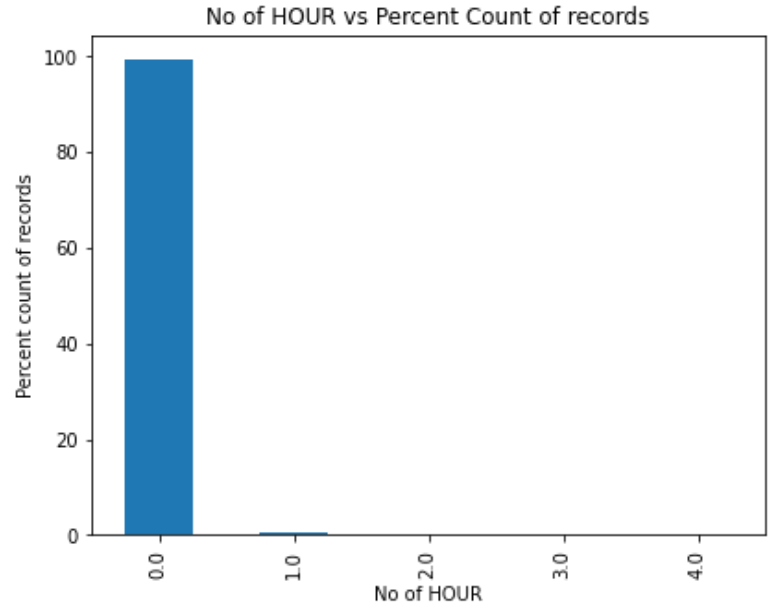
Handling Missing Values:

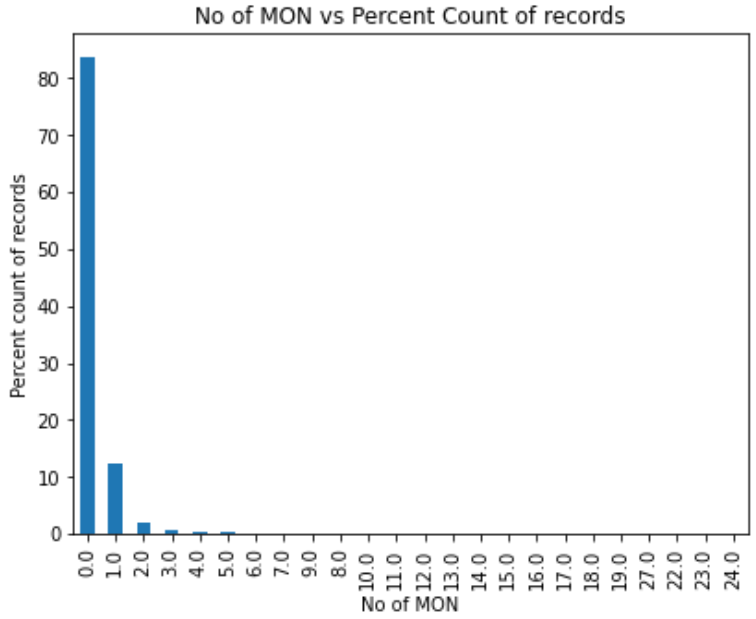
From the adjacent graph, it is evident that there are many columns with the missing values percentage more than 40%, since 40% is a significant number for missing values in the dataset, we can drop these columns.

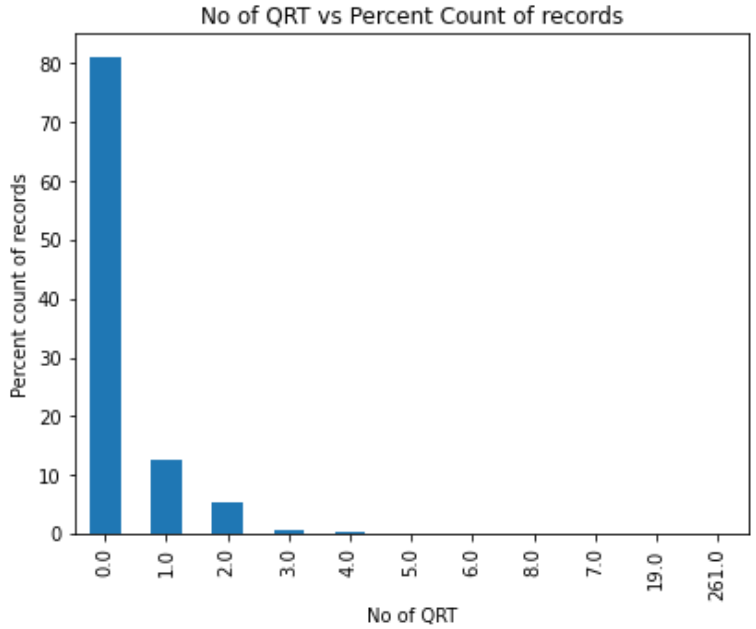


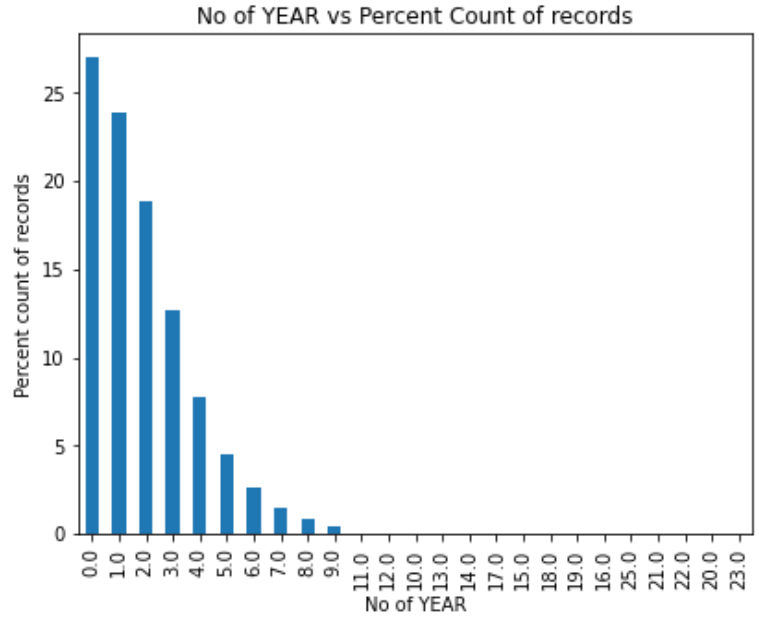

No of DAY vs Percent Count of records
Percent count of records
No of DAY
0.0
1.0
2.0
3.0
4.0
5.0
6.0
9.0
8.0


No of WEEK vs Percent Count of records
Percent count of records
No of WEEK
0.0
1.0
2.0
3.0
4.0
6.0
5.0
8.0
7.0


No of HOUR vs Percent Count of records
Percent count of records
No of HOUR
0.0
1.0
2.0
3.0
4.0


No of MON vs Percent Count of records
Percent count of records
No of MON
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
9.0
8.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
27.0
22.0
23.0
24.0


No of QRT vs Percent Count of records
Percent count of records
No of QRT
0.0
1.0
2.0
3.0
4.0
5.0
6.0
8.0
7.0
19.0
261.0

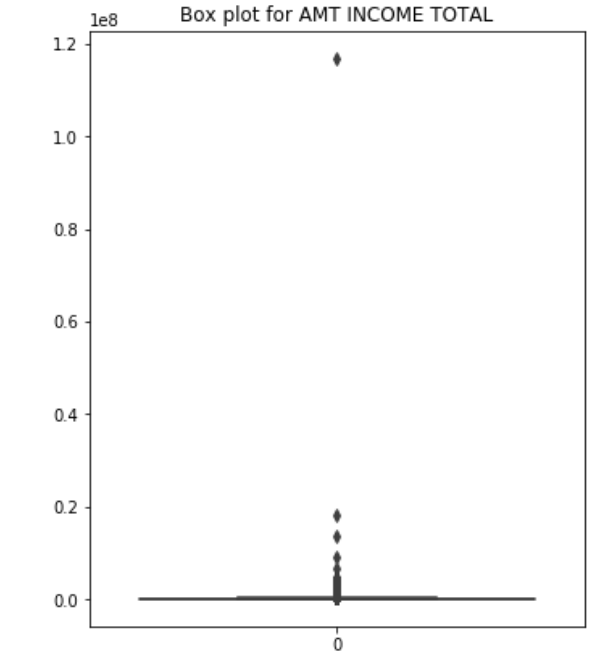

No of YEAR vs Percent Count of records
Percent count of records
No of YEAR
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
11.0
12.0
10.0
13.0
14.0
17.0
15.0
18.0
19.0
16.0
25.0
21.0
22.0
20.0
23.0

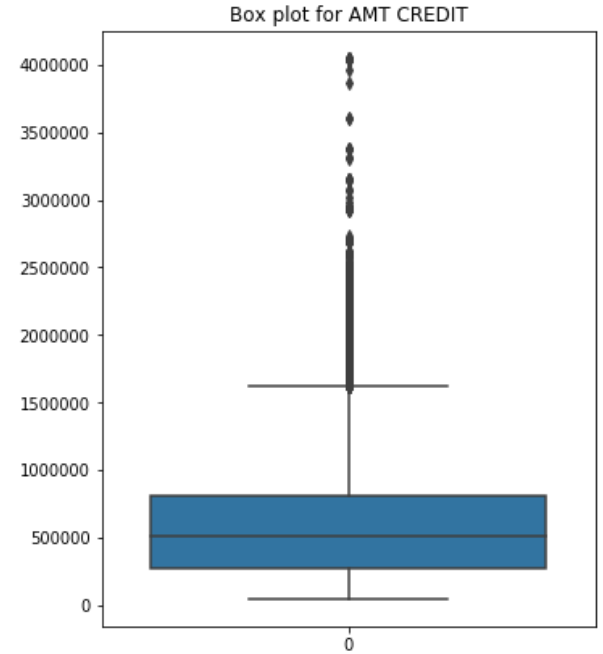
As 99% of the values in
AMT_REQ_CREDIT_BUREAU_HOUR,
AMT_REQ_CREDIT_BUREAU_DAY,
AMT_REQ_CREDIT_BUREAU_MON,
AMT_REQ_CREDIT_BUREAU_WEEK
and
AMT_REQ_CREDIT_BUREAU_QRT

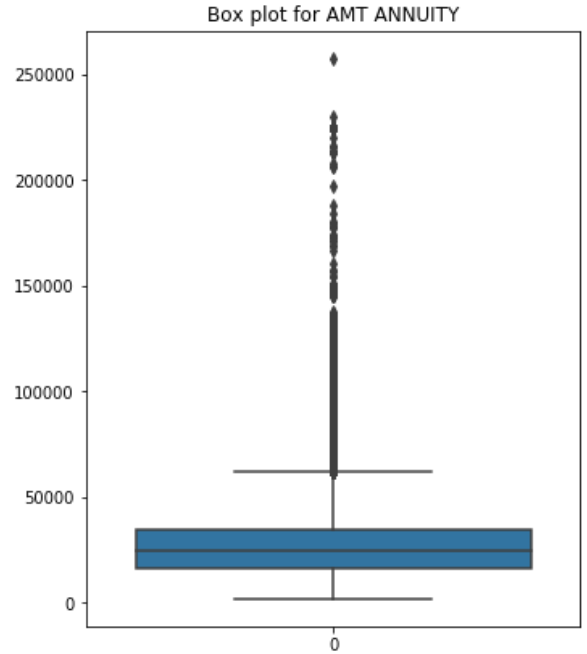
columns are 0, hence it is safe to impute missing values with 0

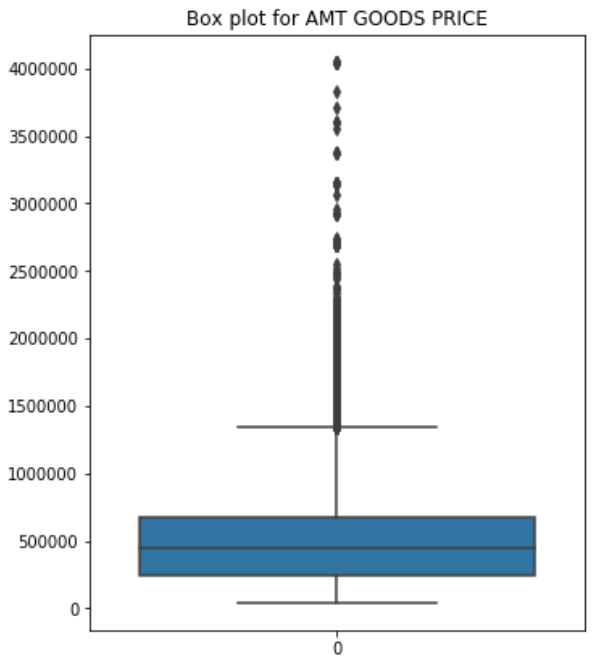
For AMT_REQ_CREDIT_BUREAU_YEAR column we can impute missing values with the median number of years

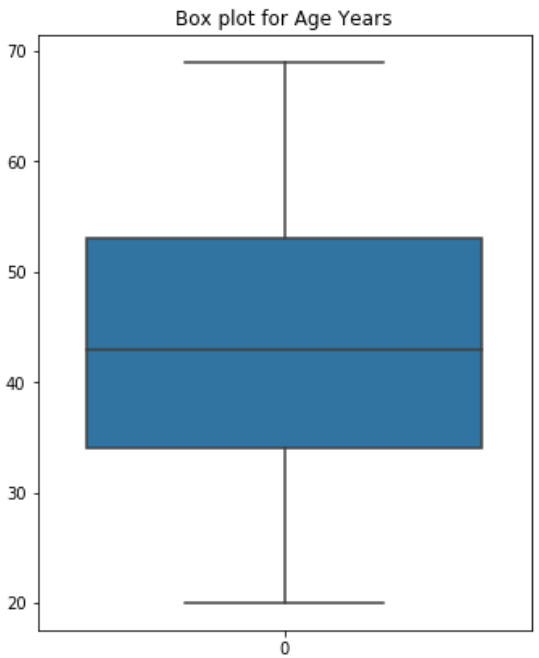
Before removing the outliers

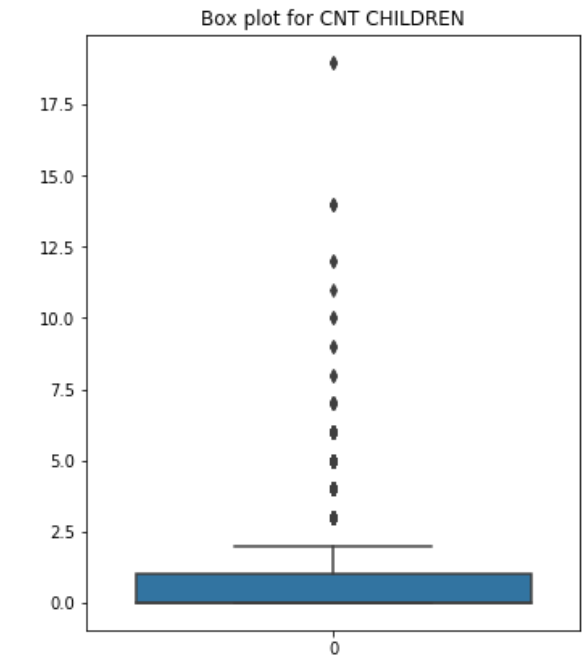




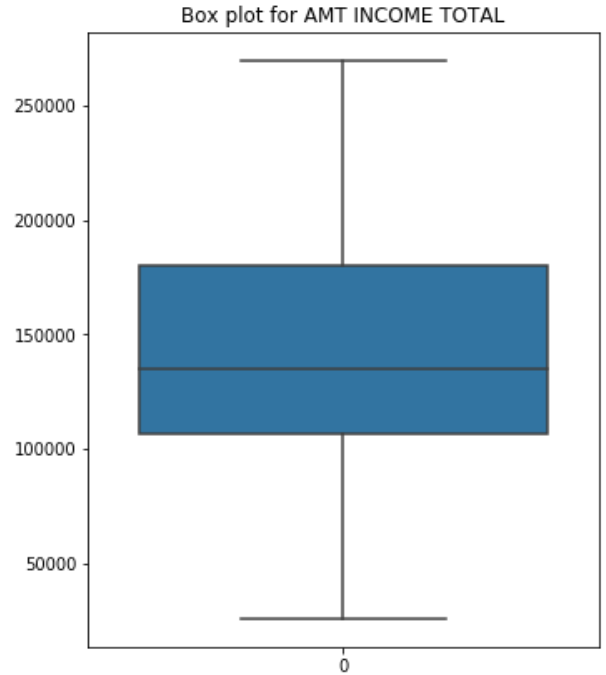


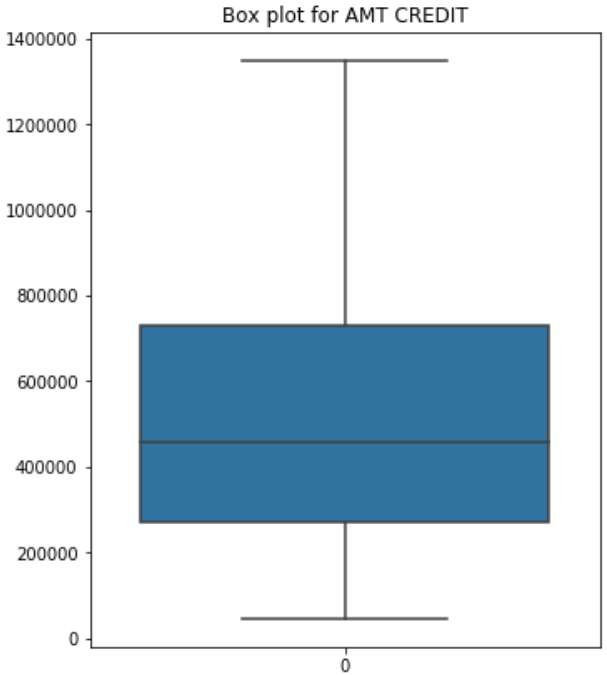


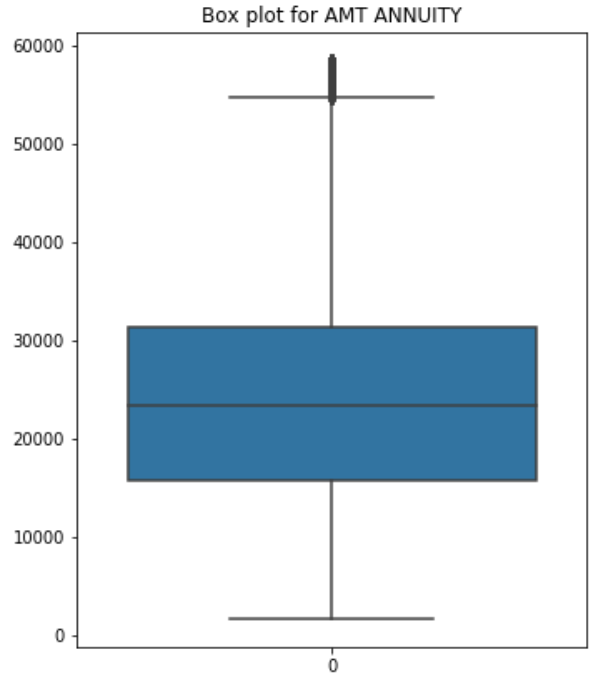


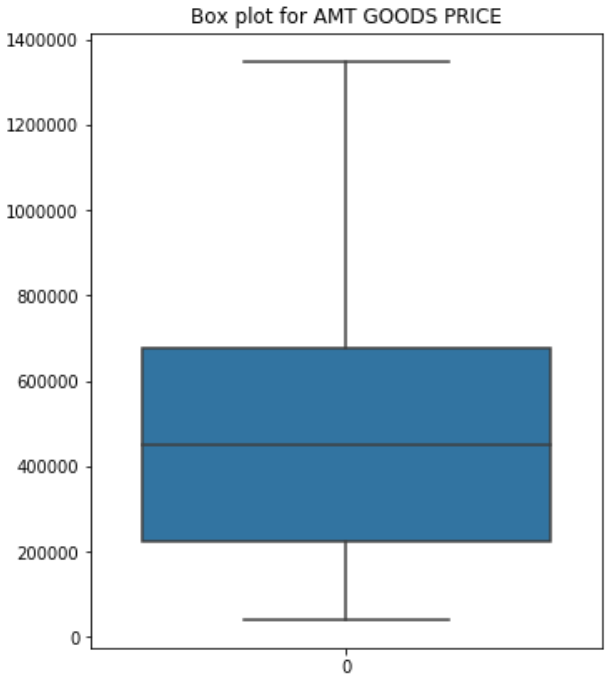


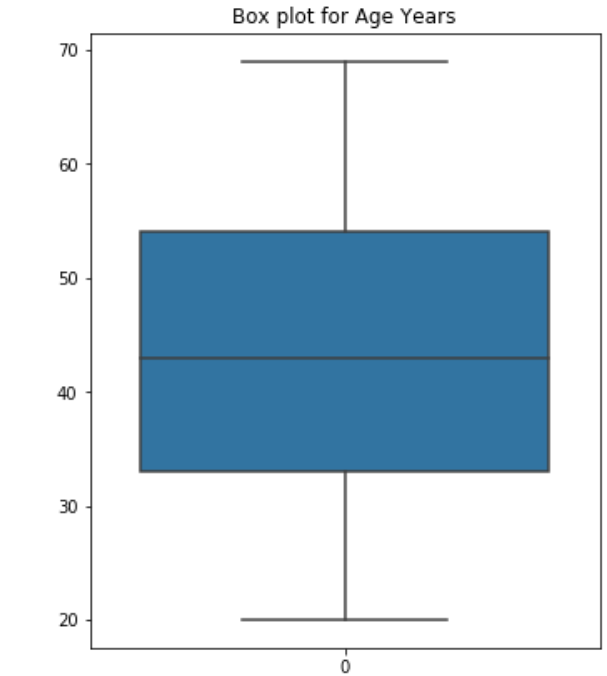
After removing the outliers

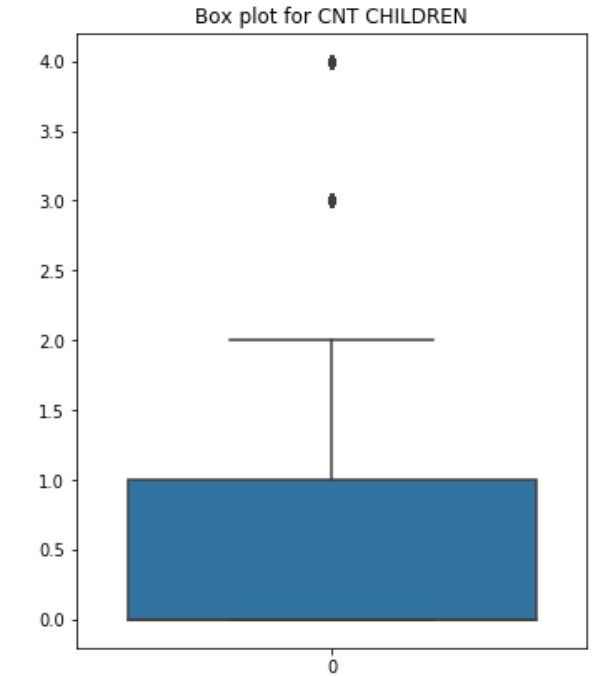


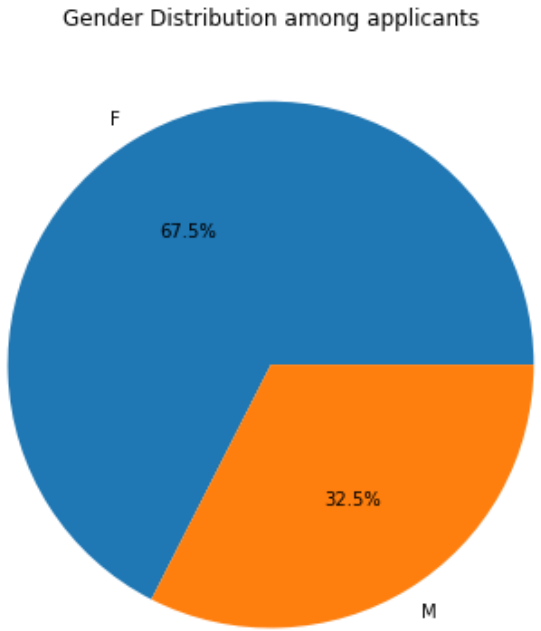


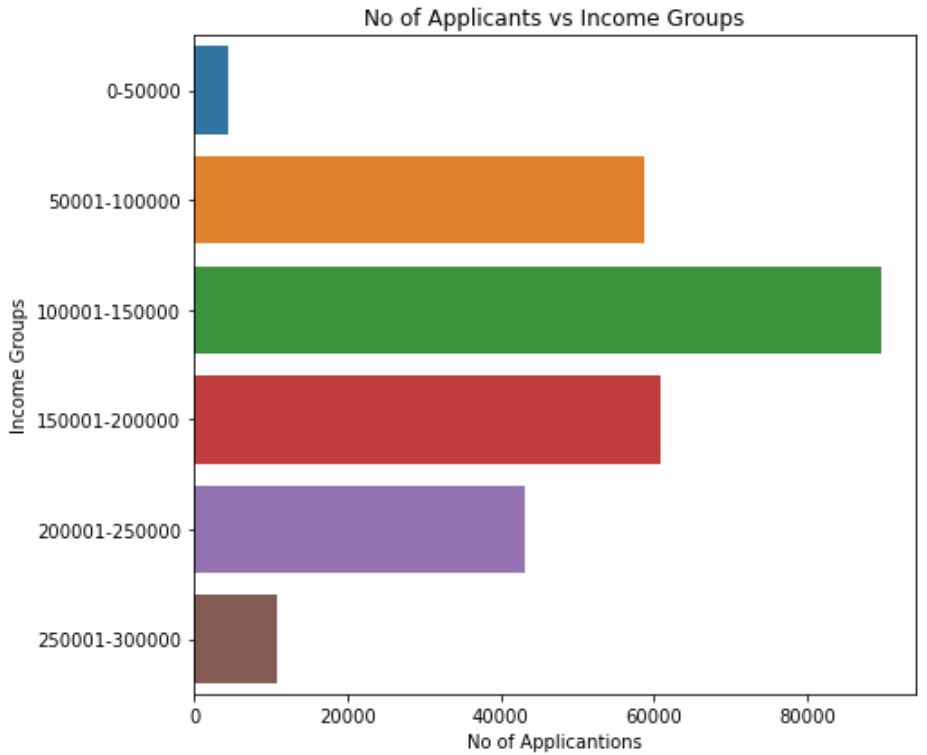


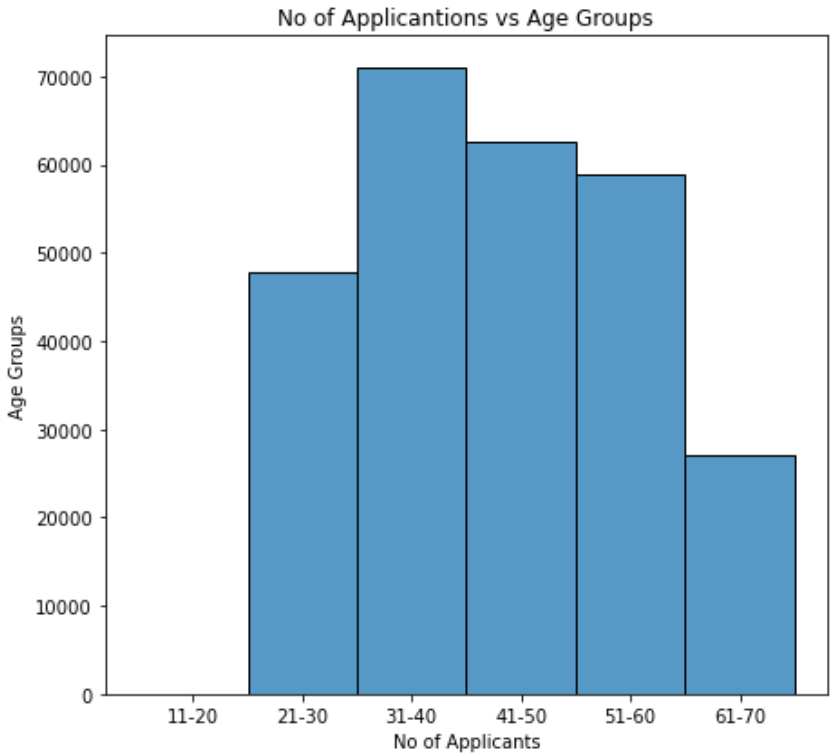




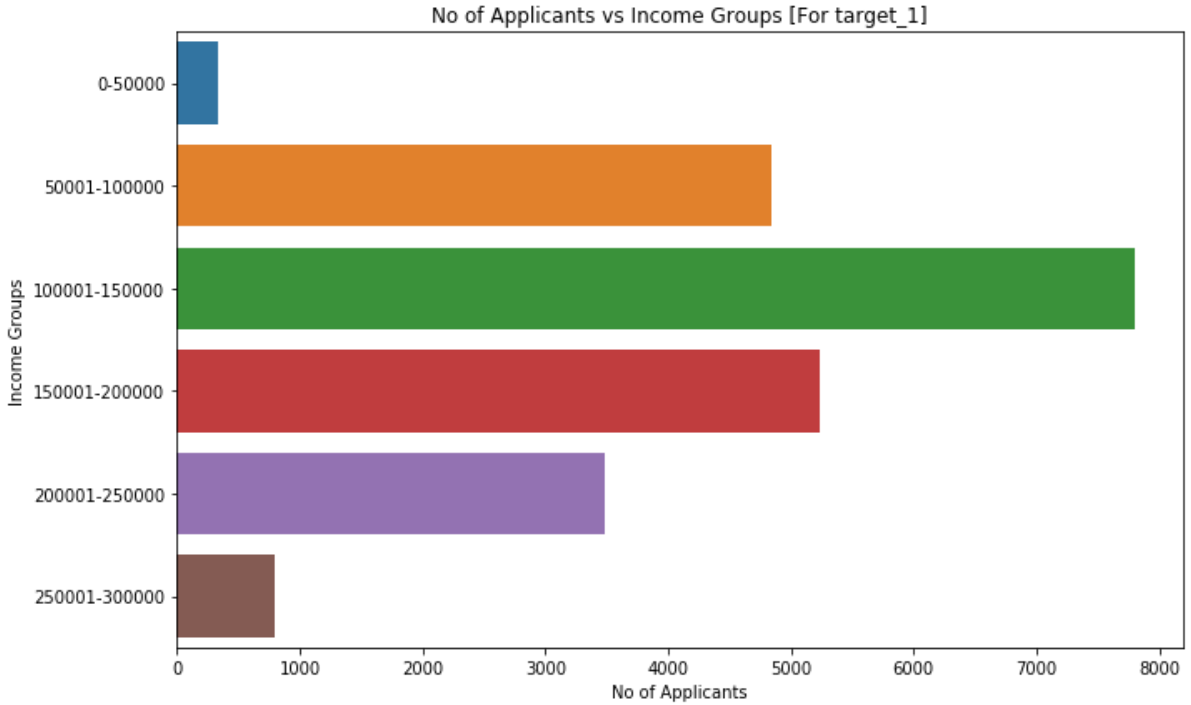


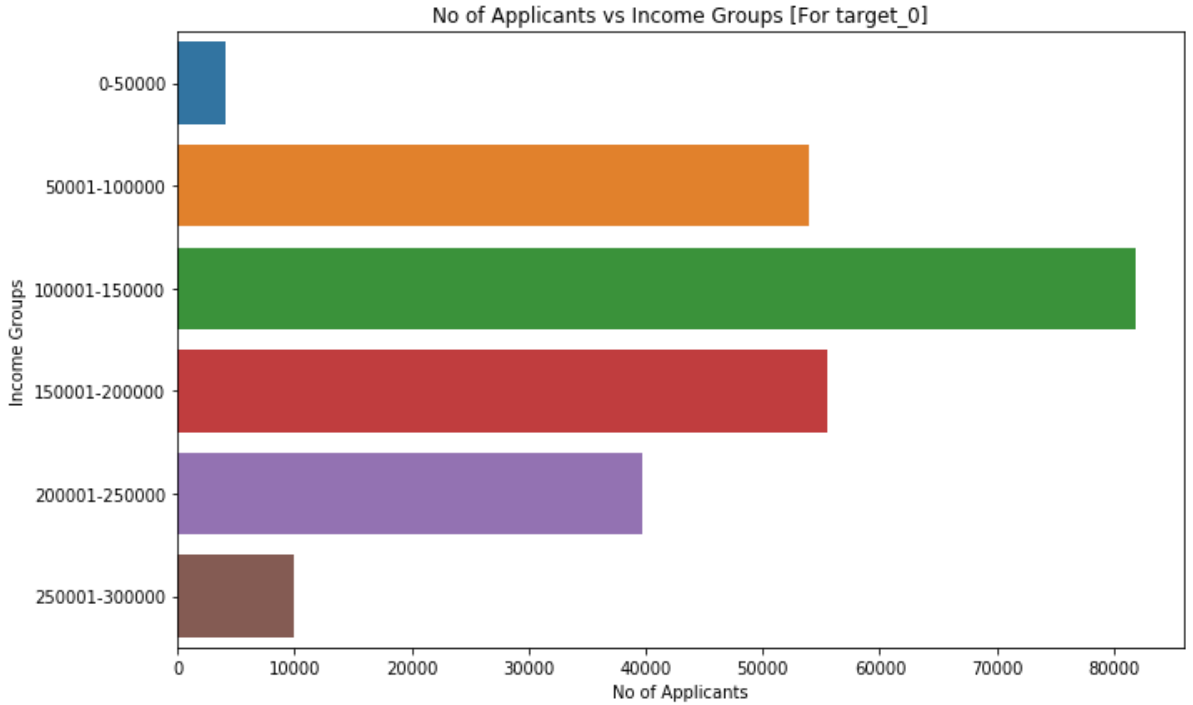






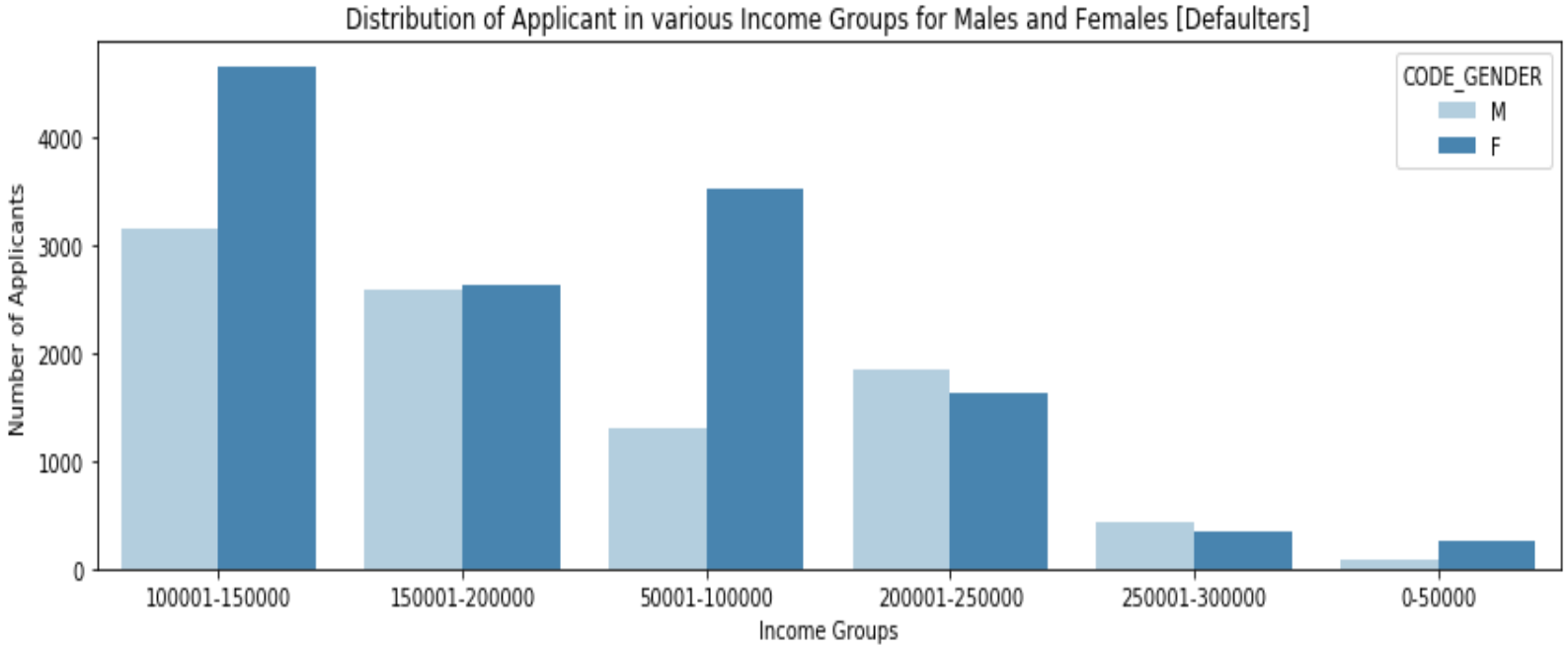
- The percentage of females is higher compared to males in loan applications
- Higher loan applications have come from applicants with income group 100001-150000
- People in age group 31-40 have applied the highest number of loans

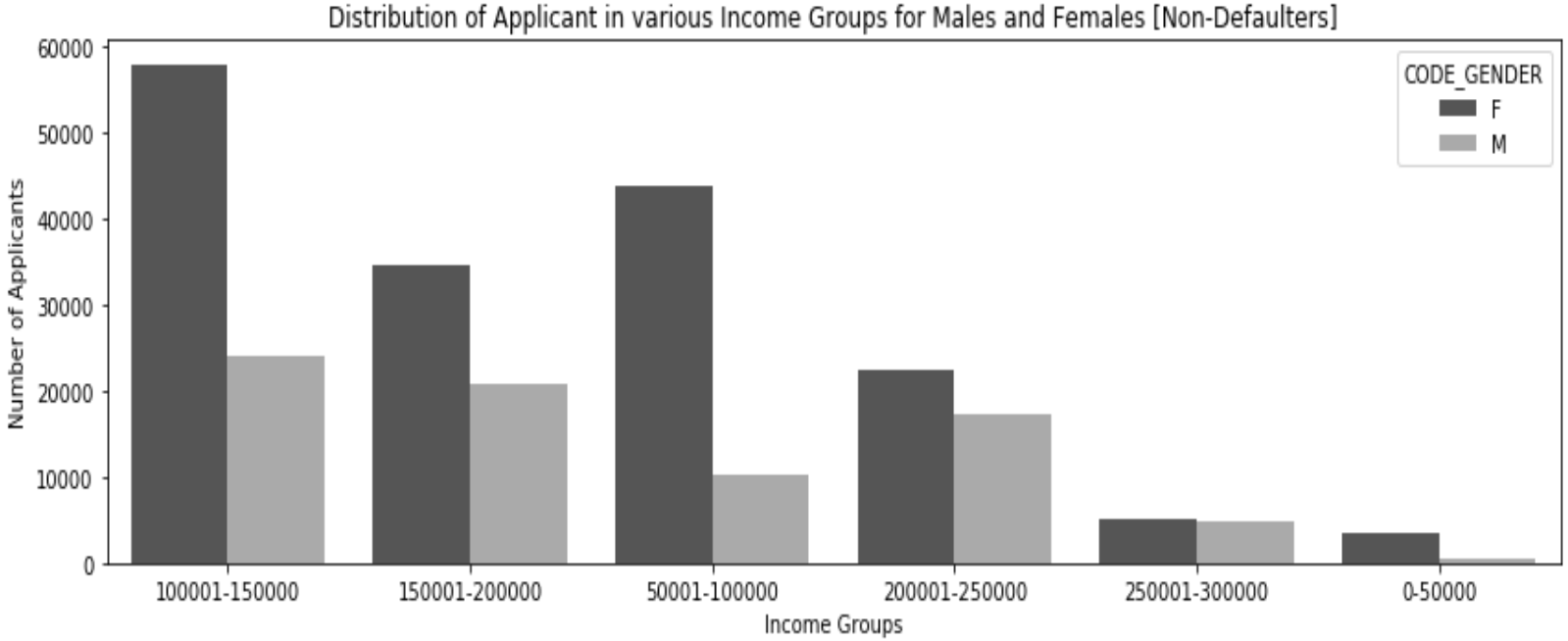


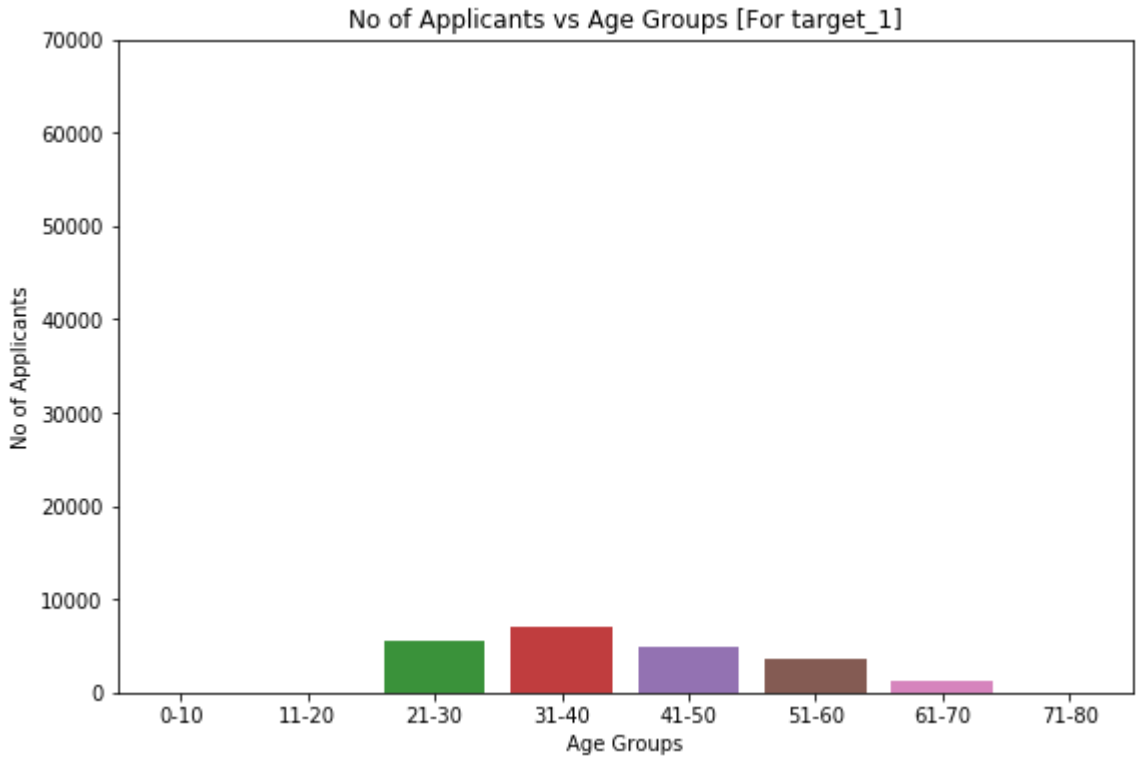


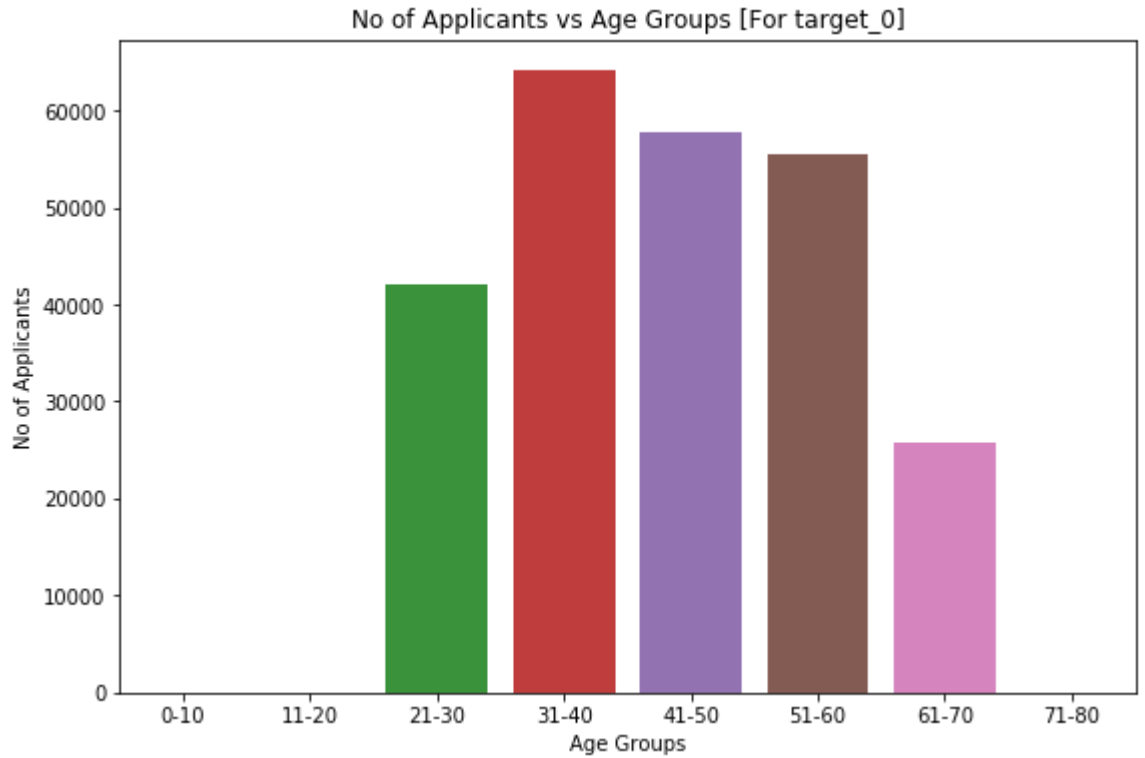
- The number of applications in income group 100001-150000 have most defaults as well as success
- Whereas, income group 0-50000 has lowest defaults as well as non-defaults

Number of loan applications from females are greater in all the income groups compared to males

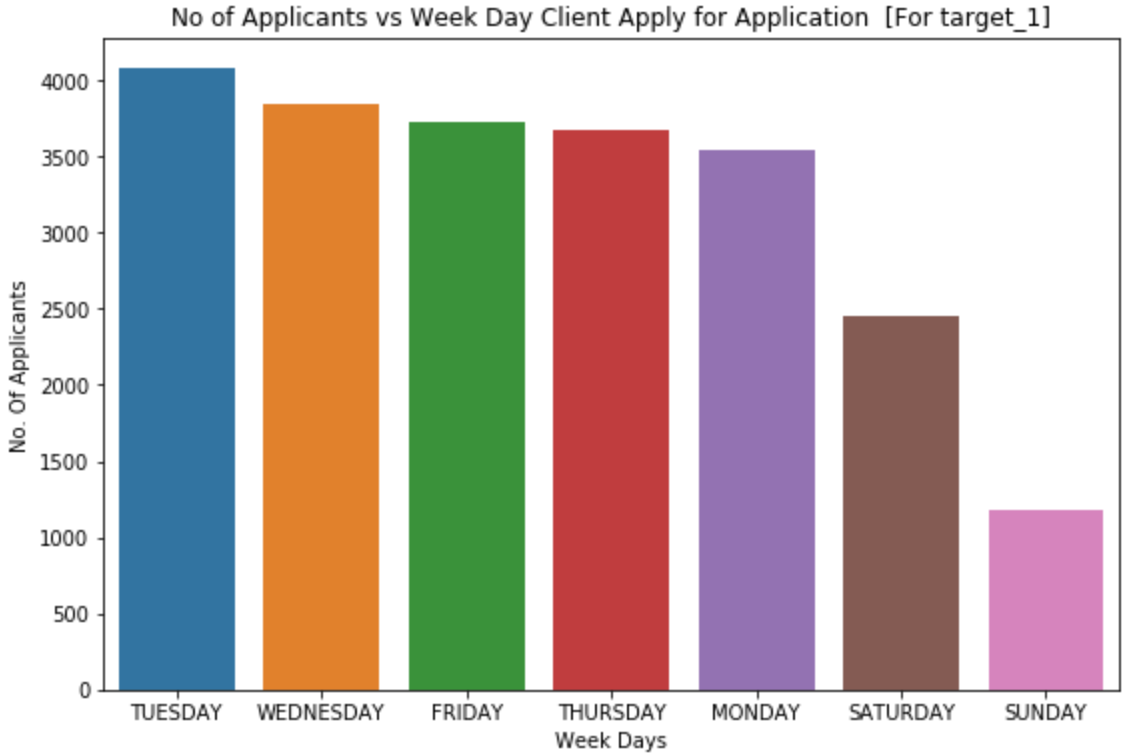


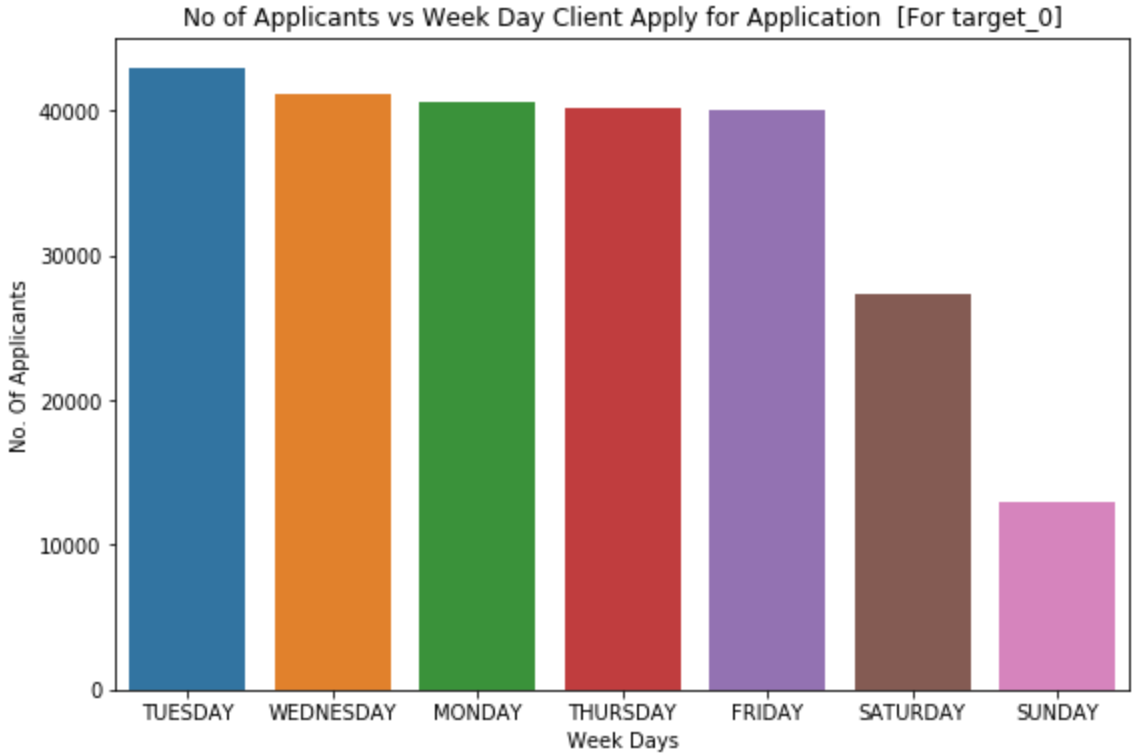






Number of loans that were defaulted and not defaulted are highest in age group 31 to 40 years and it decreases as age increases or decreases

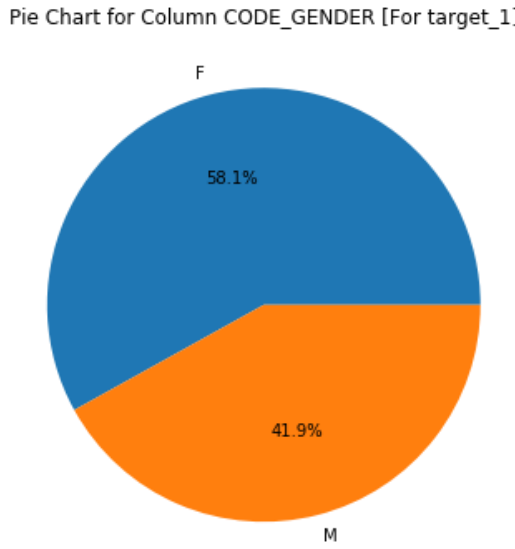


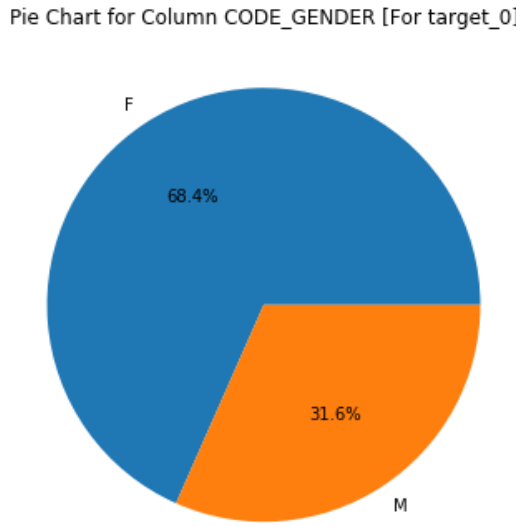


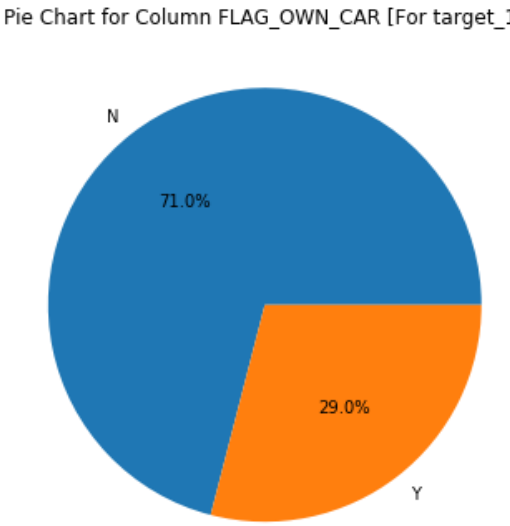
From above graphs we can see that on Tuesday there are highest number of loan applications registered whereas on Sunday least number of application were registered

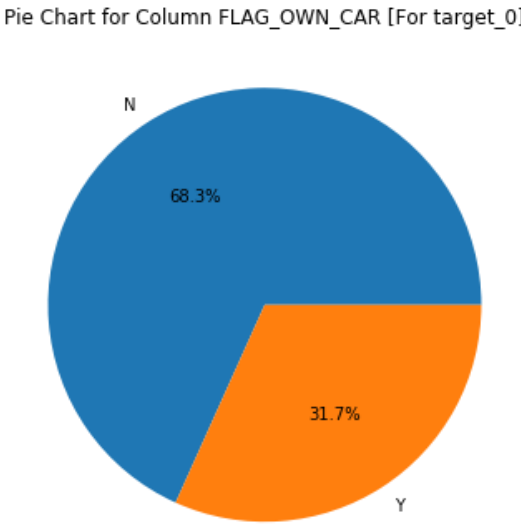
Proportion of females is higher than males in both defaulters and non-defaulters

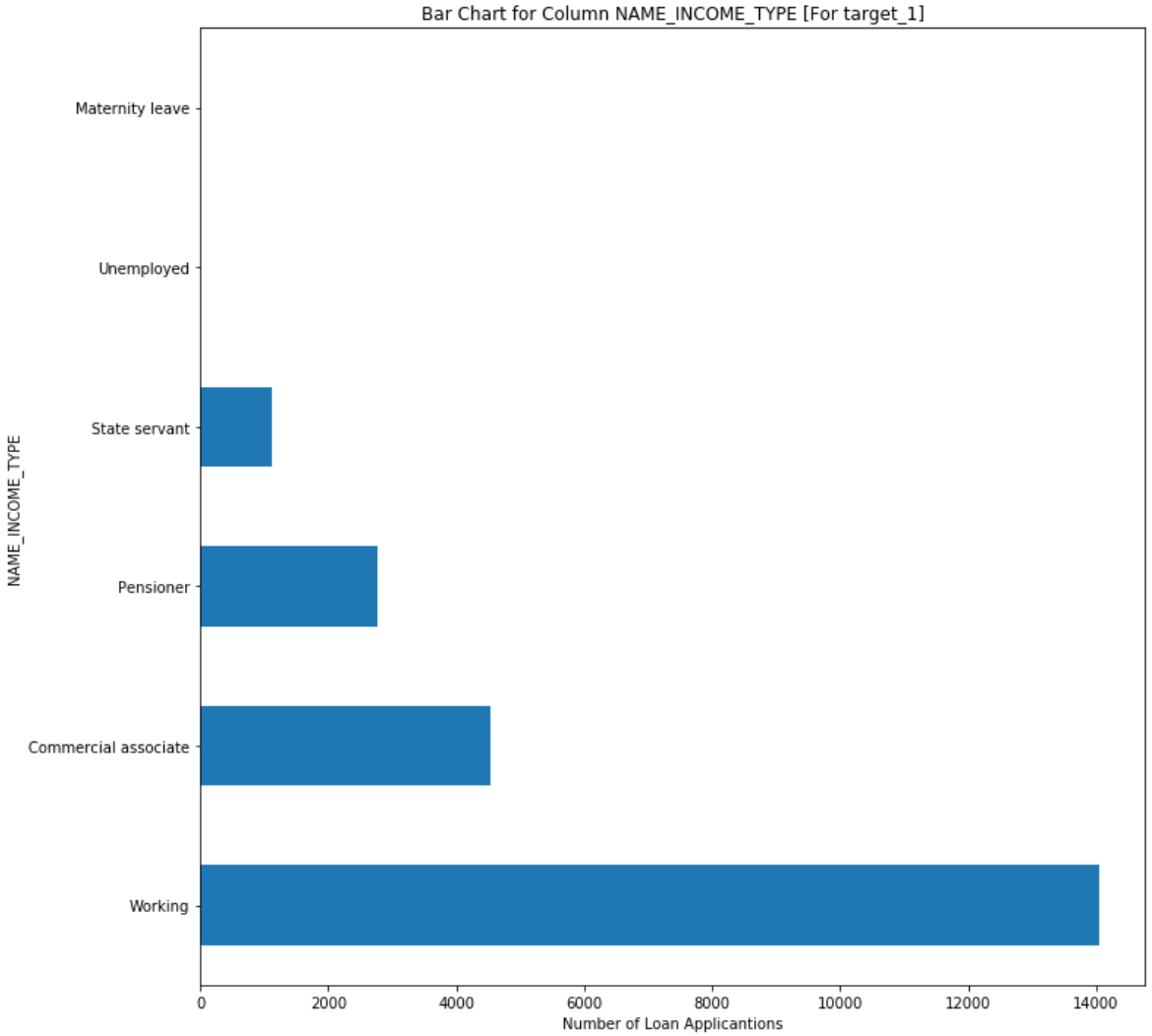
Proportion of both defaulters and non-defaulters not having a car is higher than those who have it

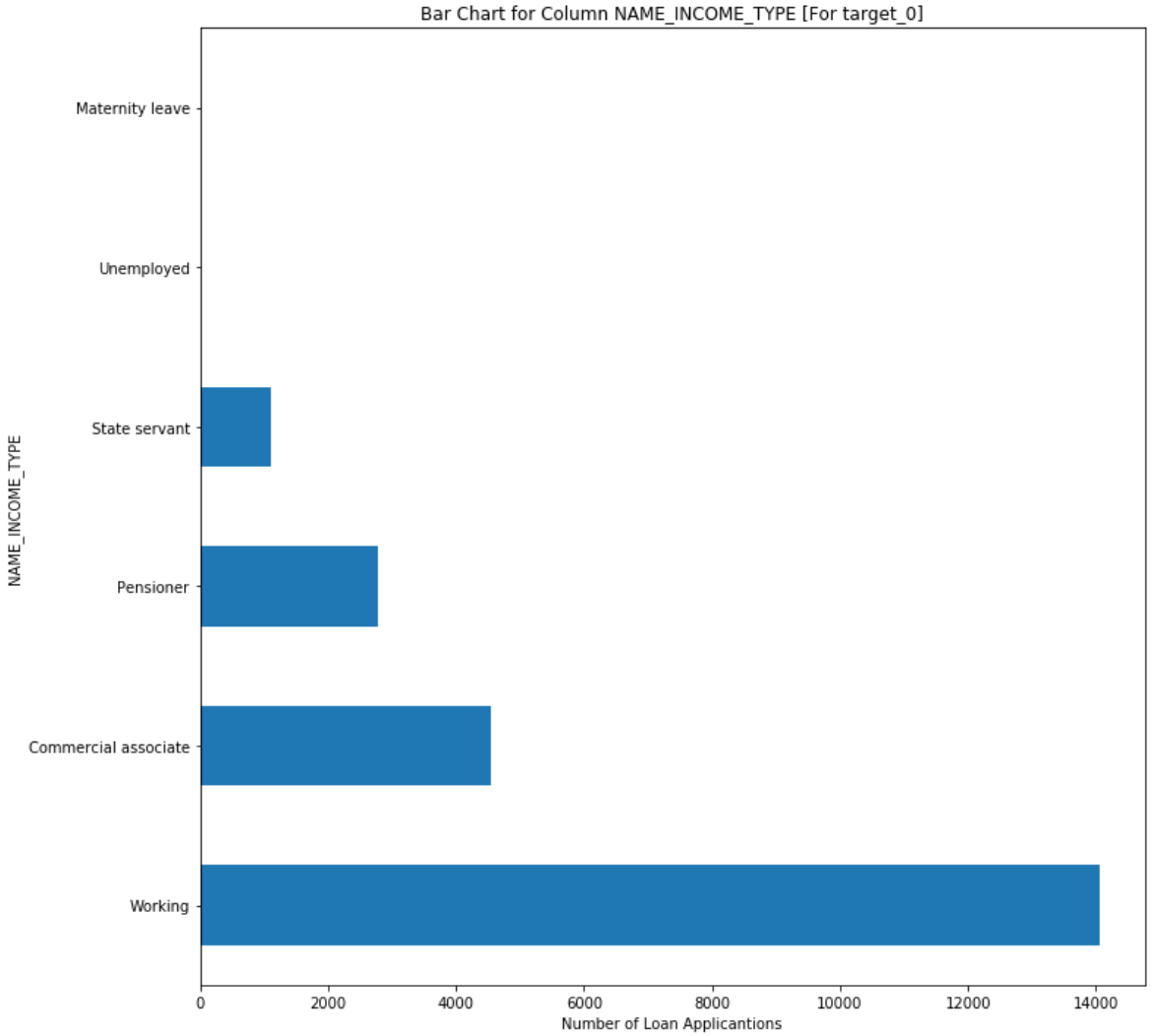




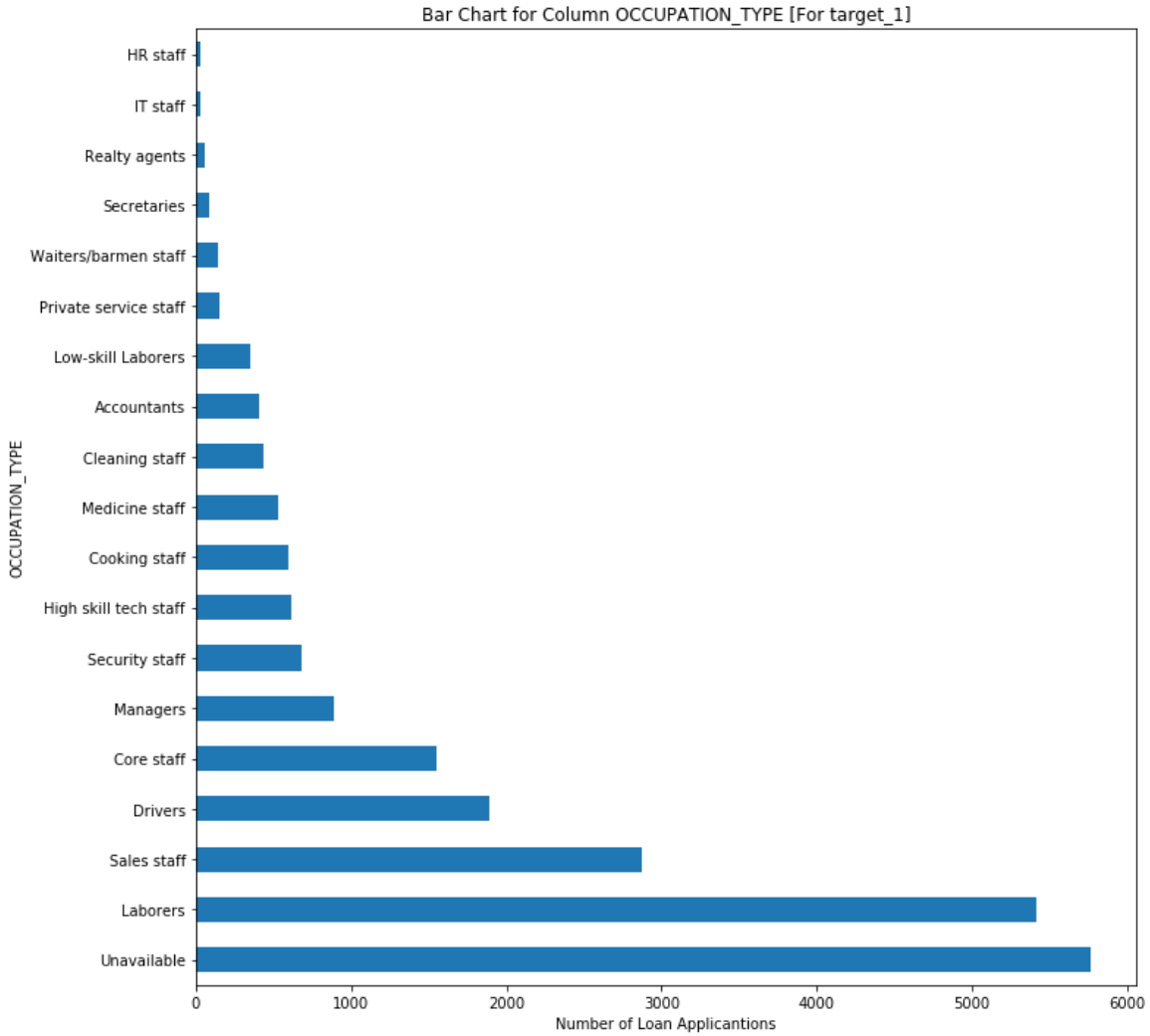


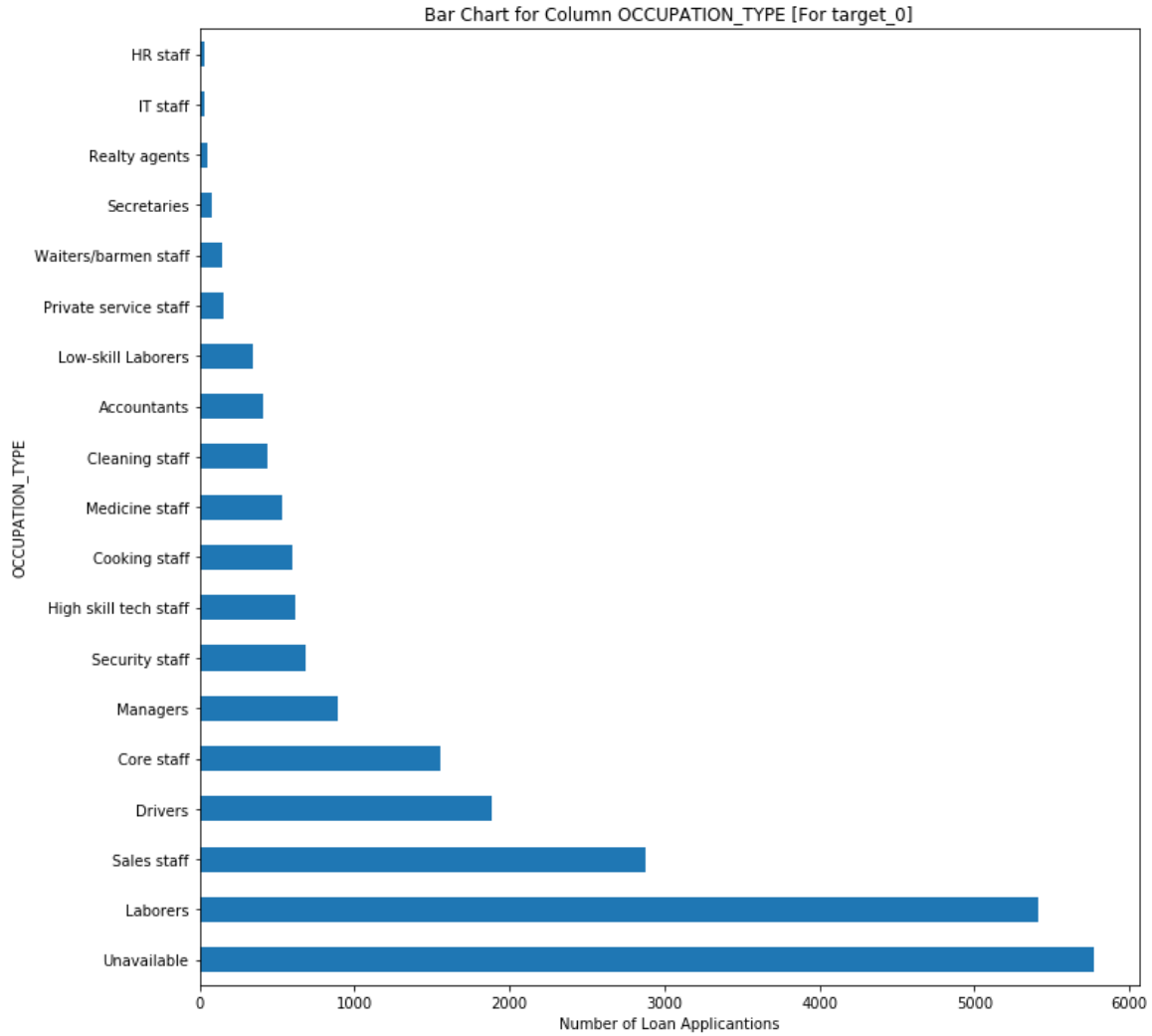




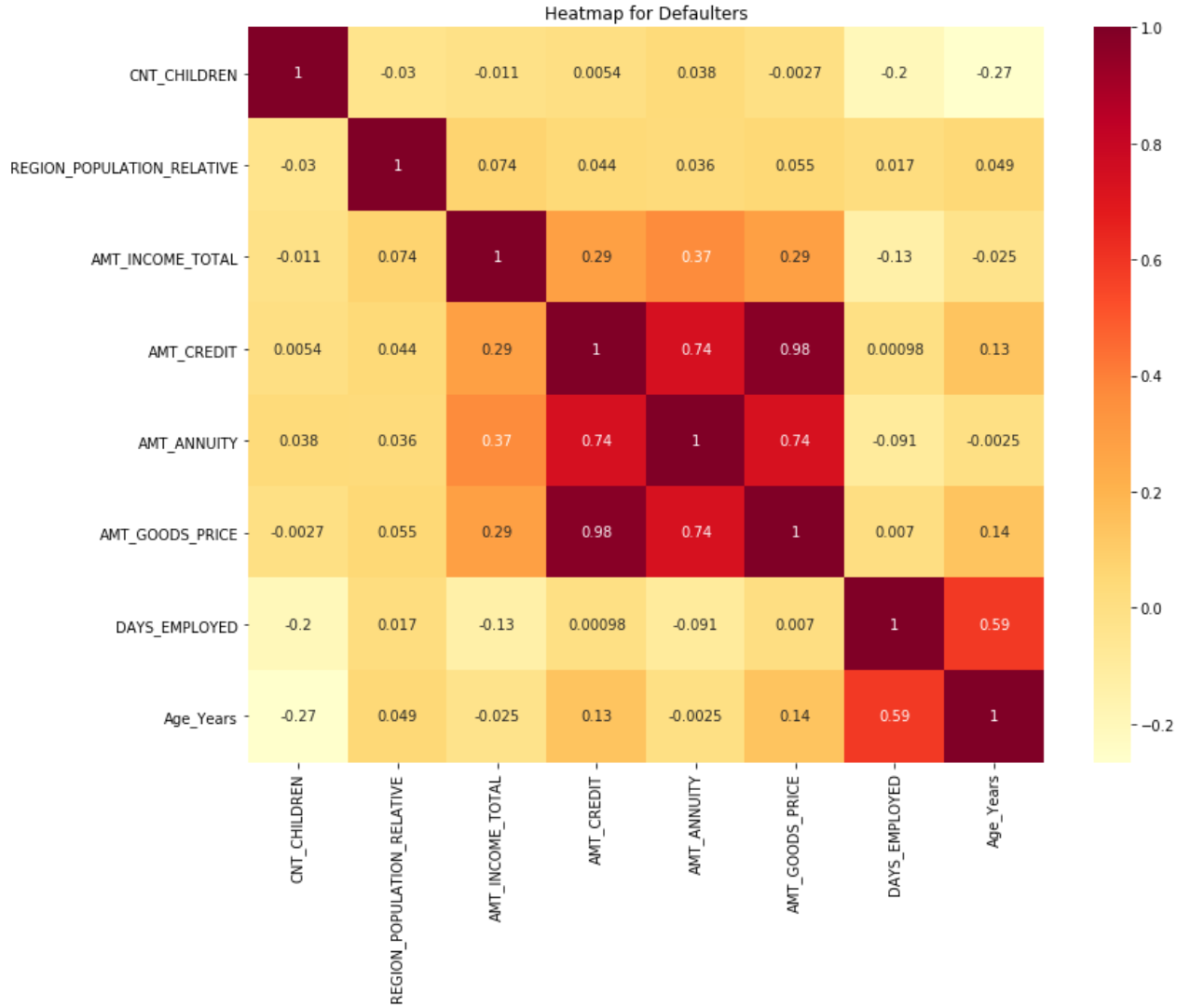


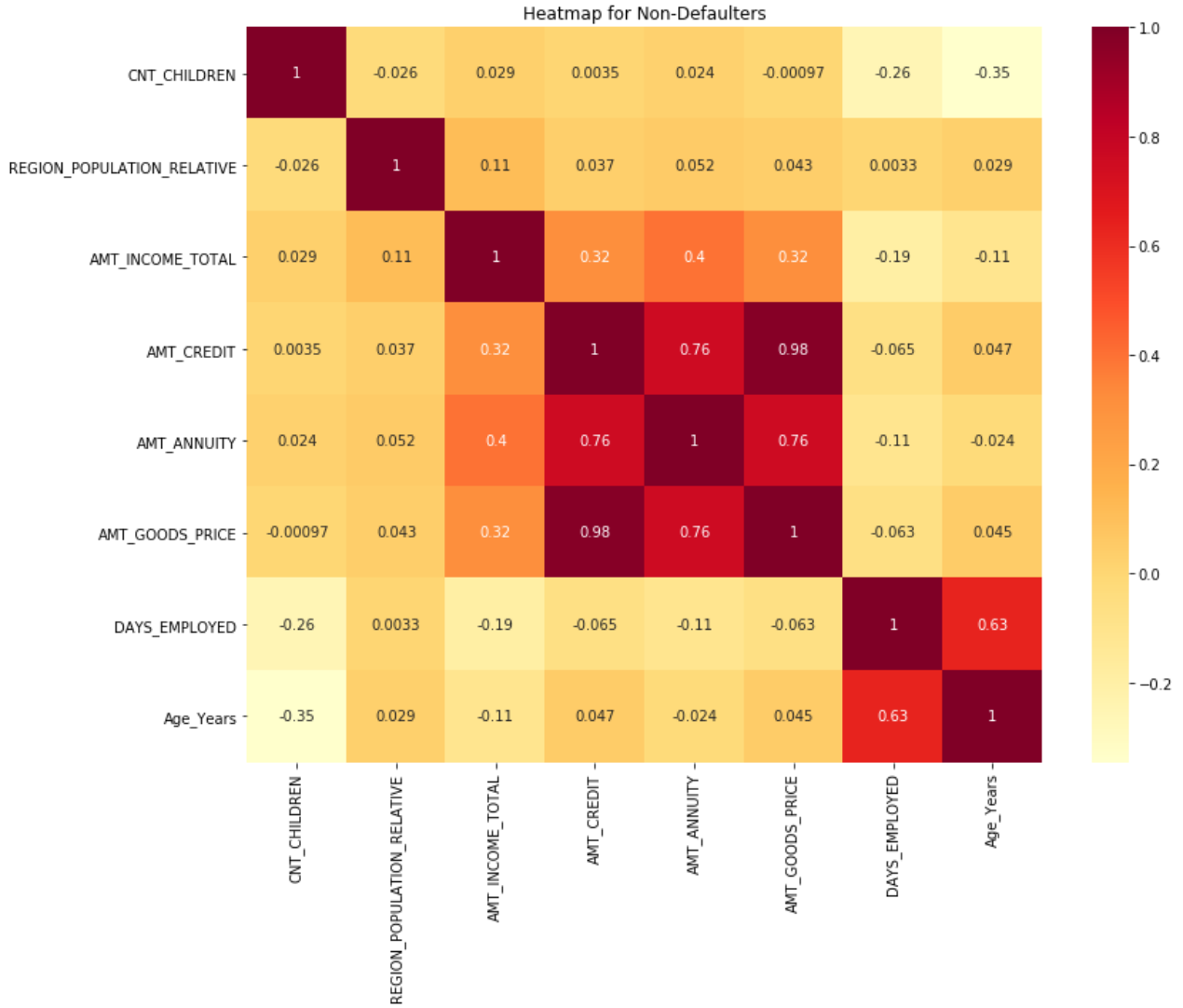
People from working income background have applied higher number of loan applications than other income type and this category has highest defaulted loans



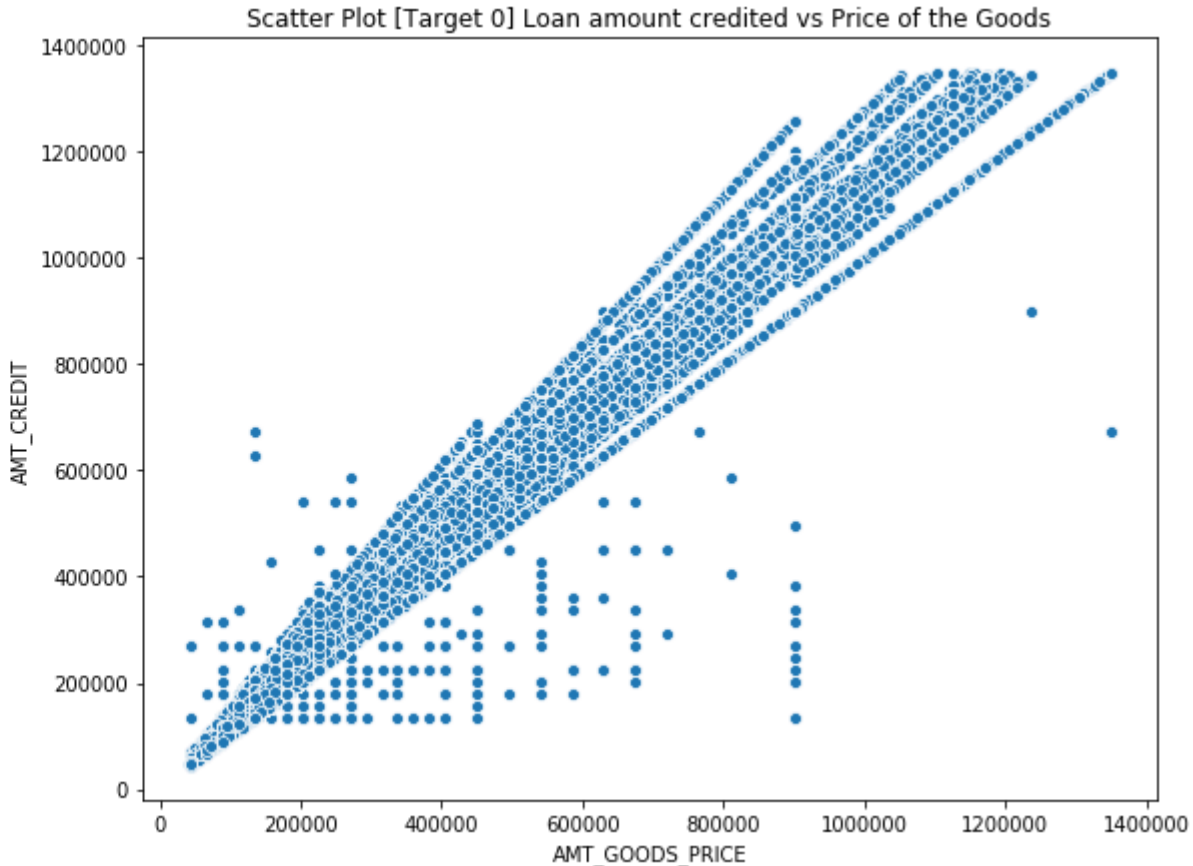


Occupation type which has highest number of defaulted and non-defaulted loans is unknown



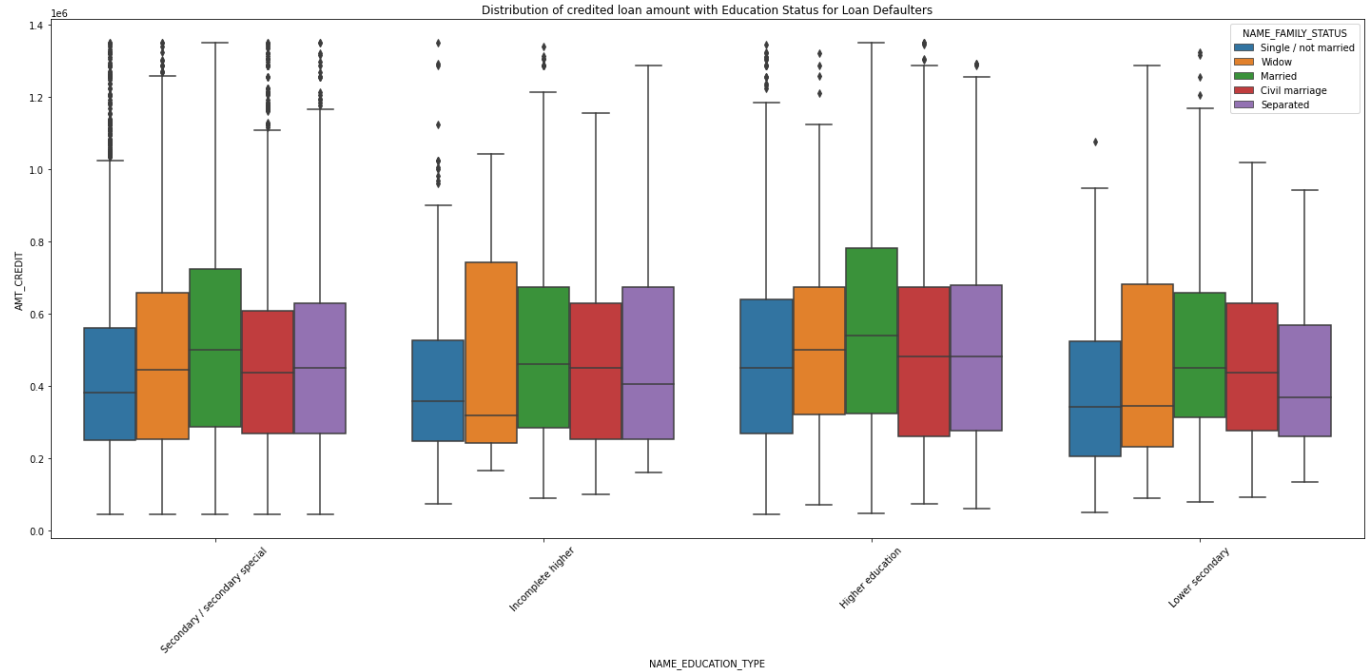


From the above heatmap, it evident that there is strong correlation between the Goods Price and Loan Amount that was credited. Also, it can be seen that higher the age the number of days people were employed is also high.
On the other hand, There is weak correlation between the age and count of children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

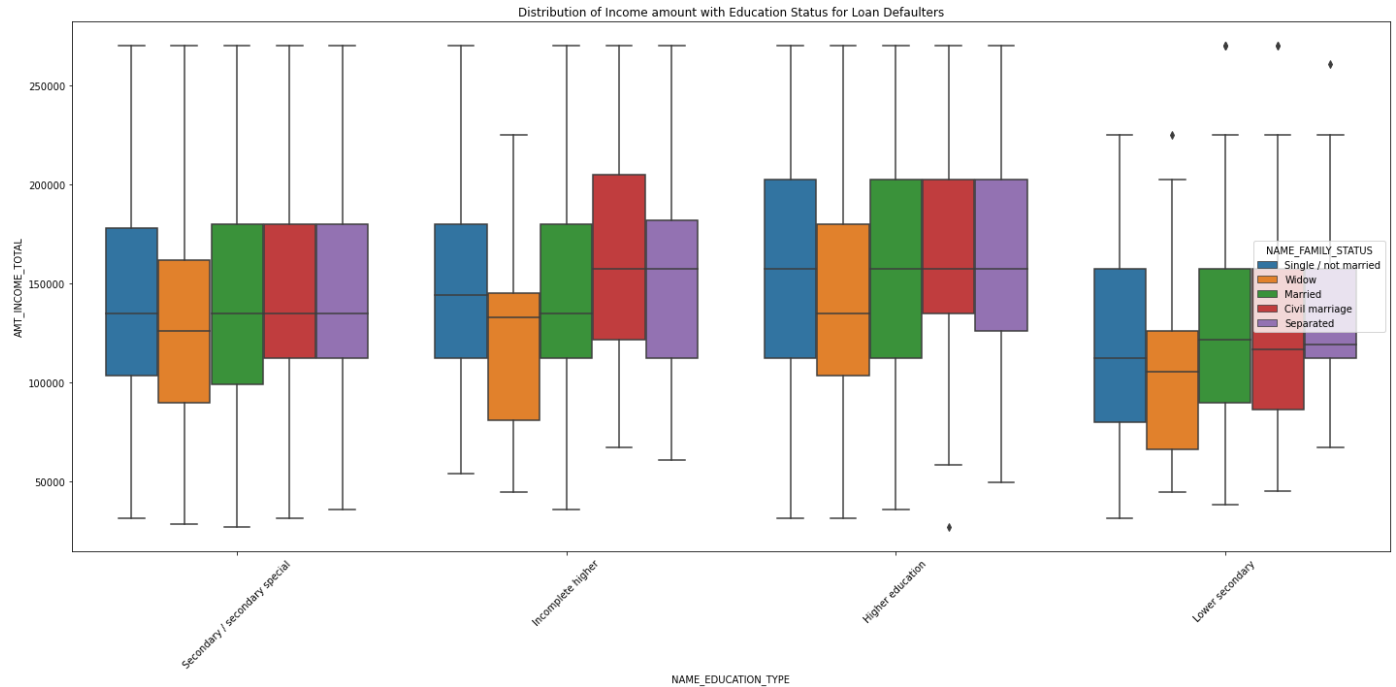


From the above scatterplot, we can confirm that there is strong correlation between the Goods Price and Loan Amount that was credited

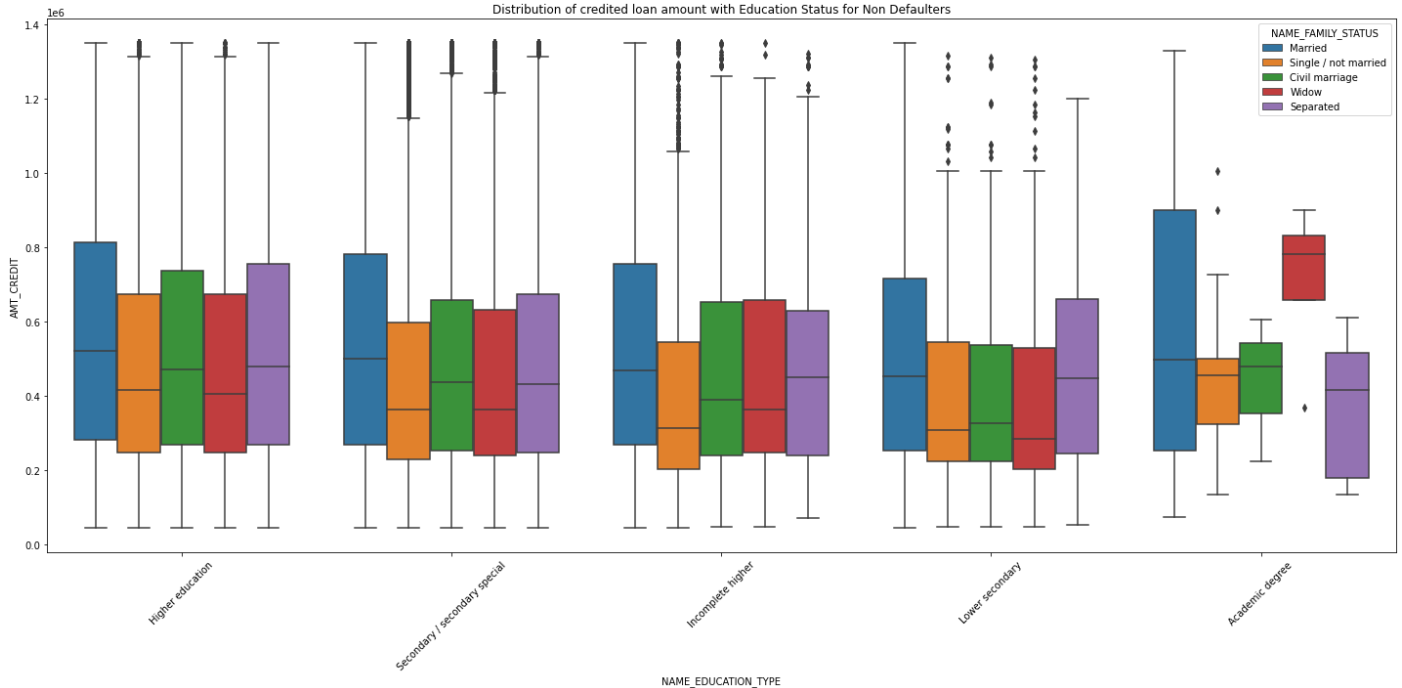
Married people with higher education background have defaulted loans for higher distribution of credited loan amounts



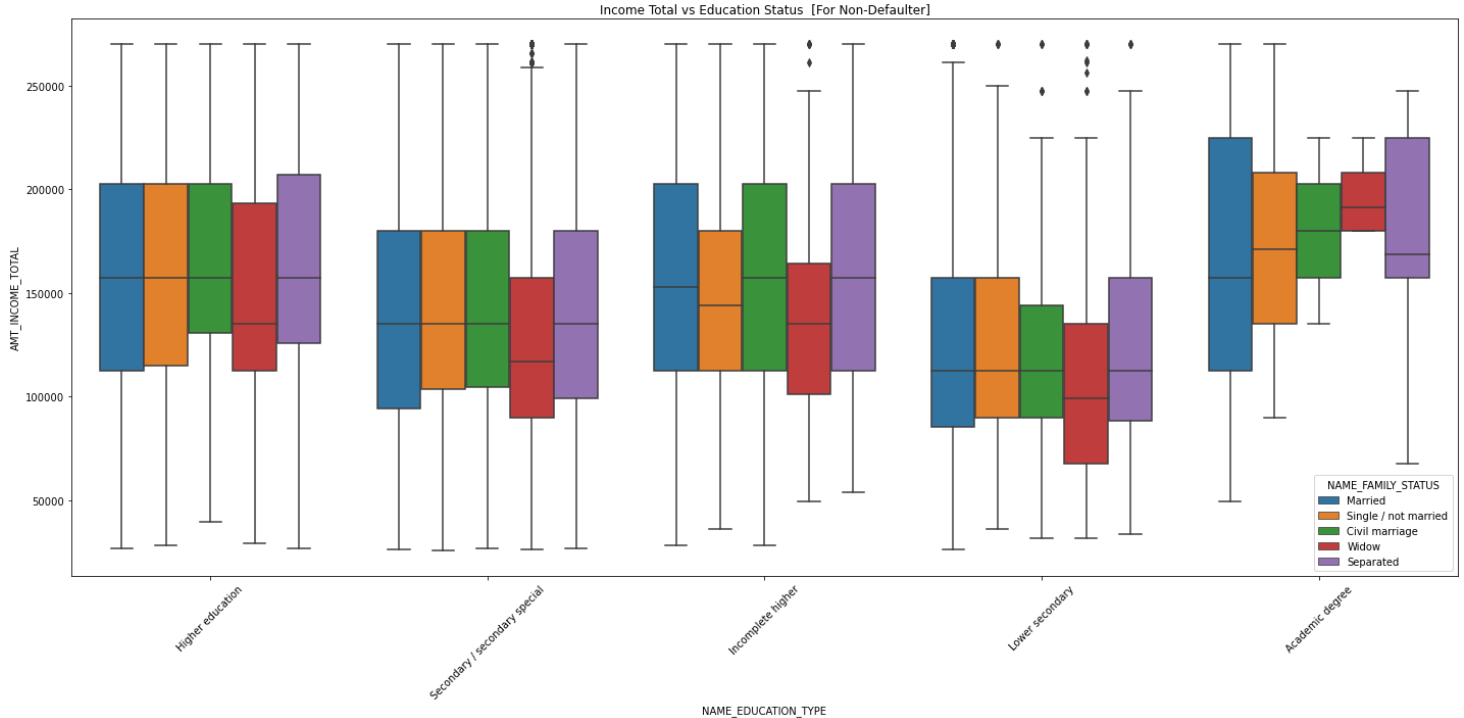
Married people with higher education background have defaulted loans for higher distribution of income amounts



Married people with Academic degree have successful loans for higher distribution of credited loan amounts



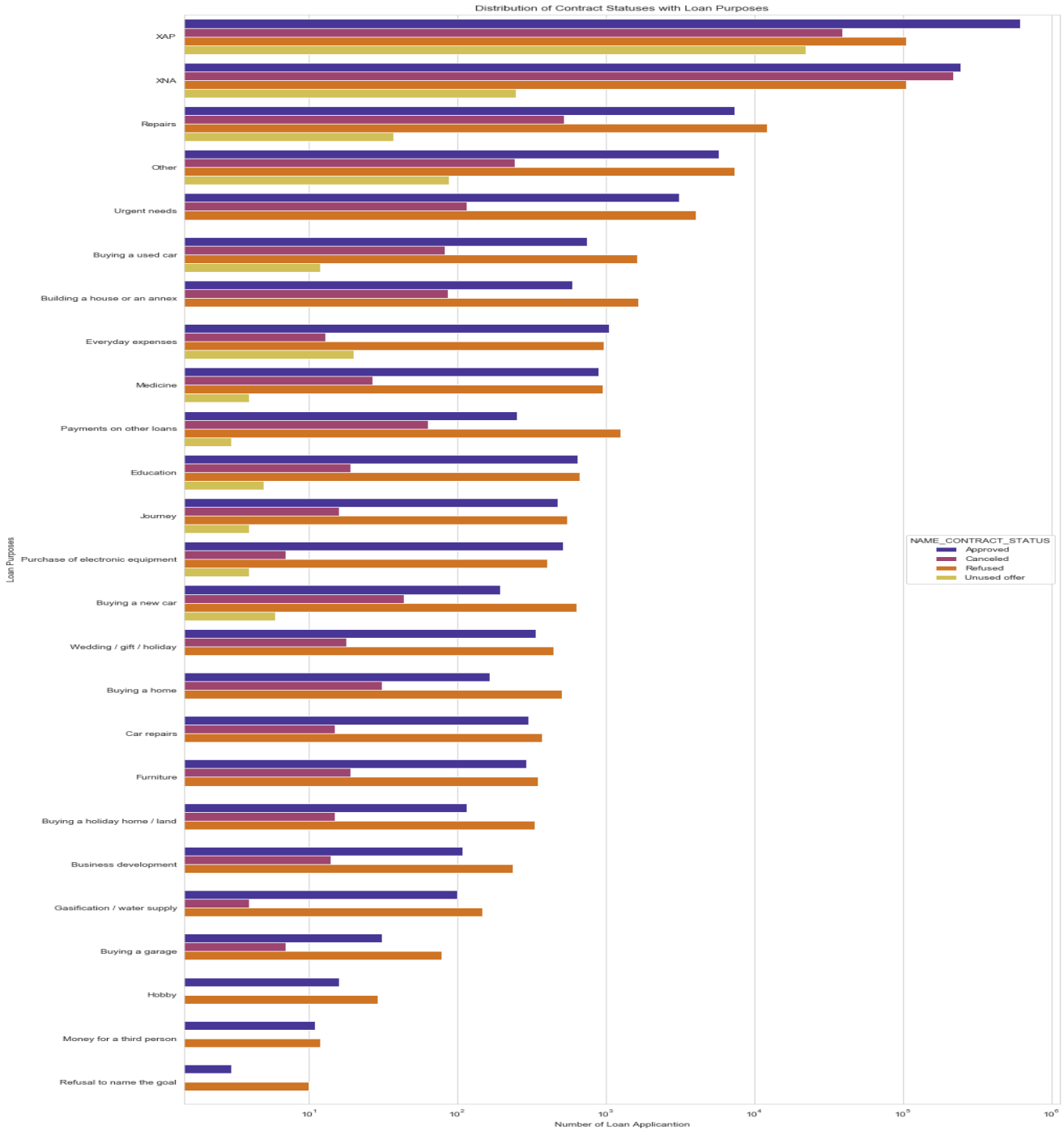
Married people with Academic degree have successful loans for higher distribution of Income amounts.

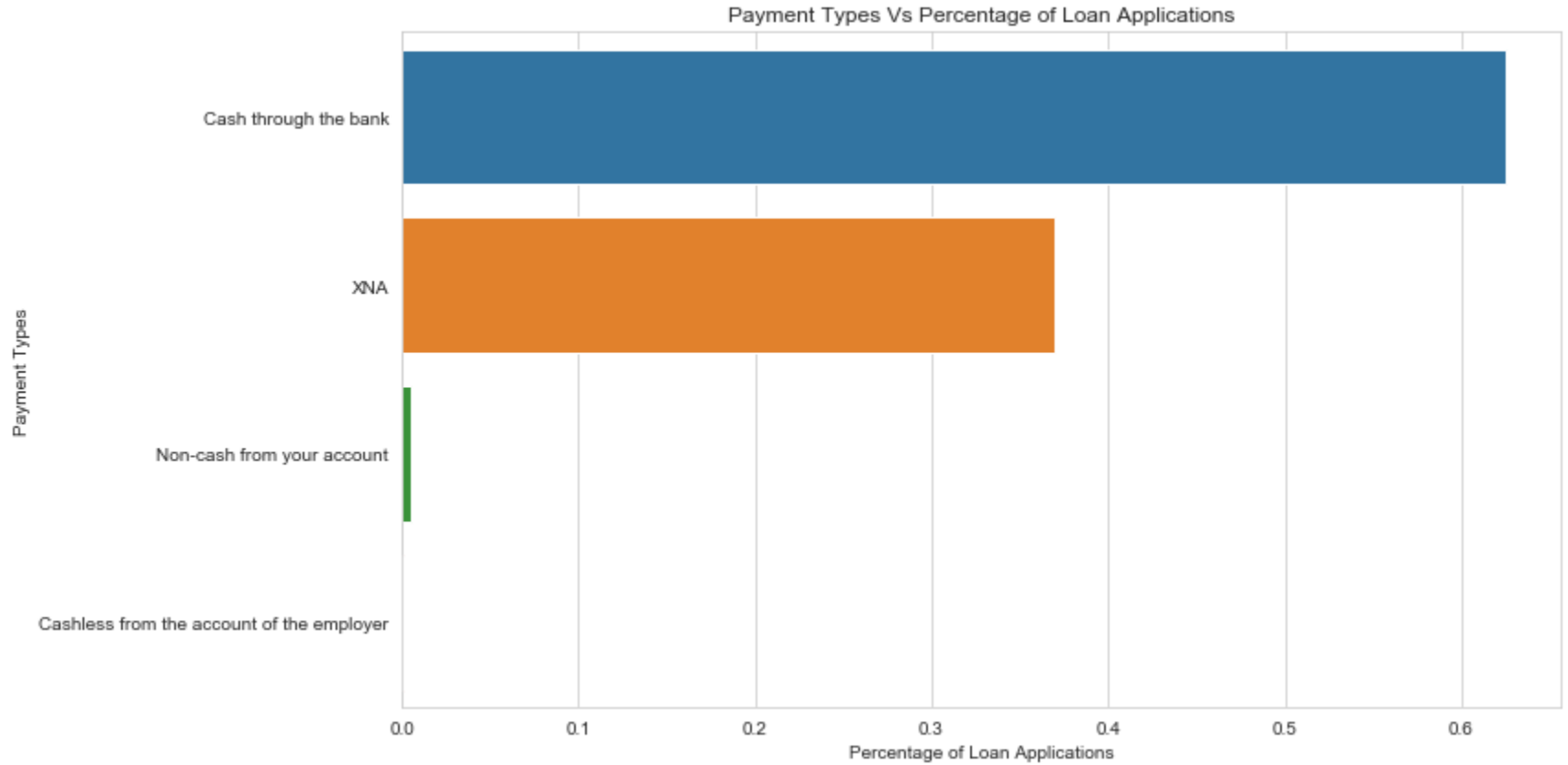


When customers do not communicate their purpose for the loan bank has approved low number of applications.

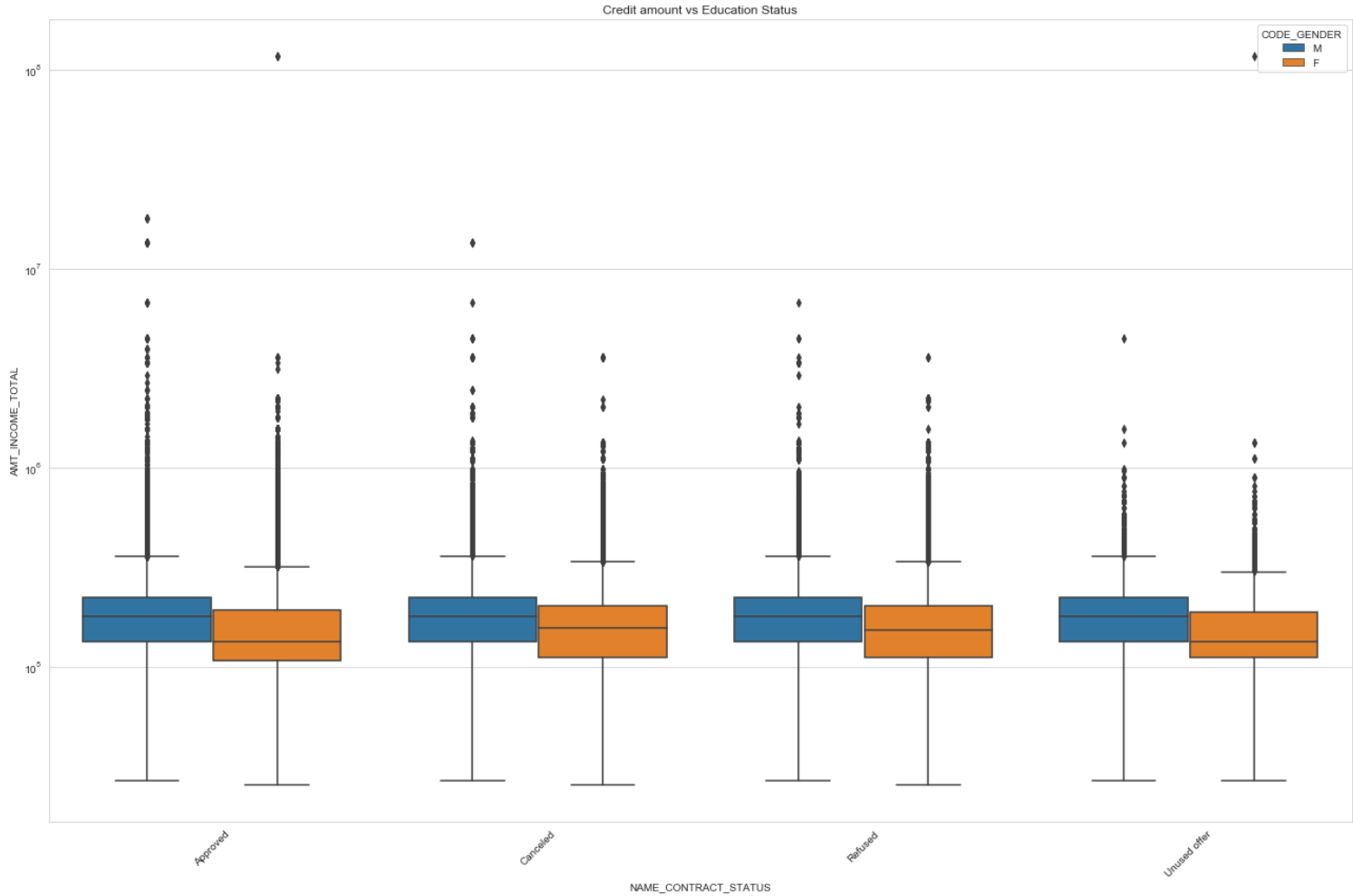
For all the purposes of taking loans, bank has refused more applications that it approved.

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.

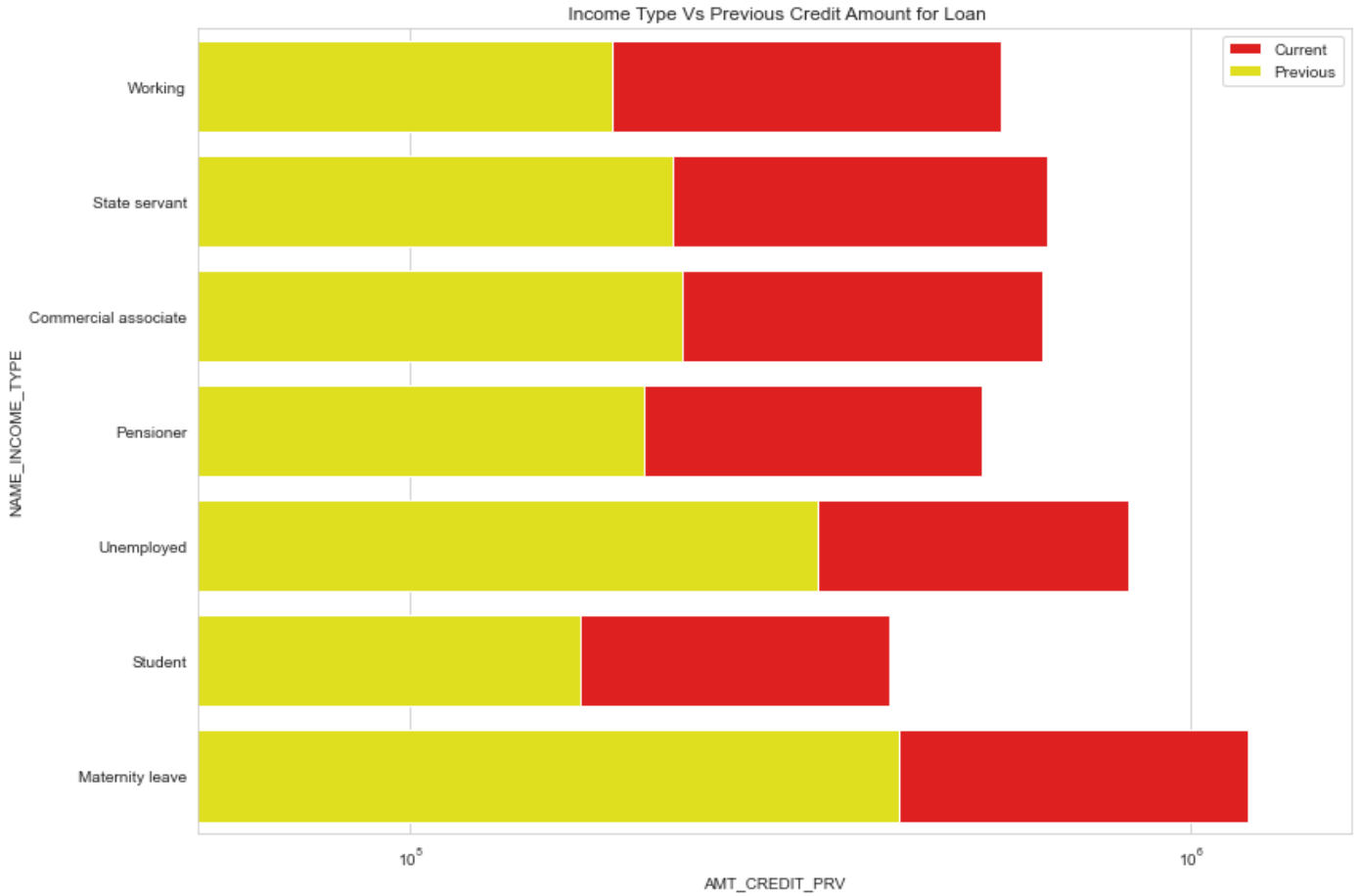




Most of the previous loans were repaid by Cash deposits.
The second highest mode of previous loan repayment has undisclosed methods (XNAs)
Non-Cash and Cashless contribute to extreme low percent of loan repayment methods

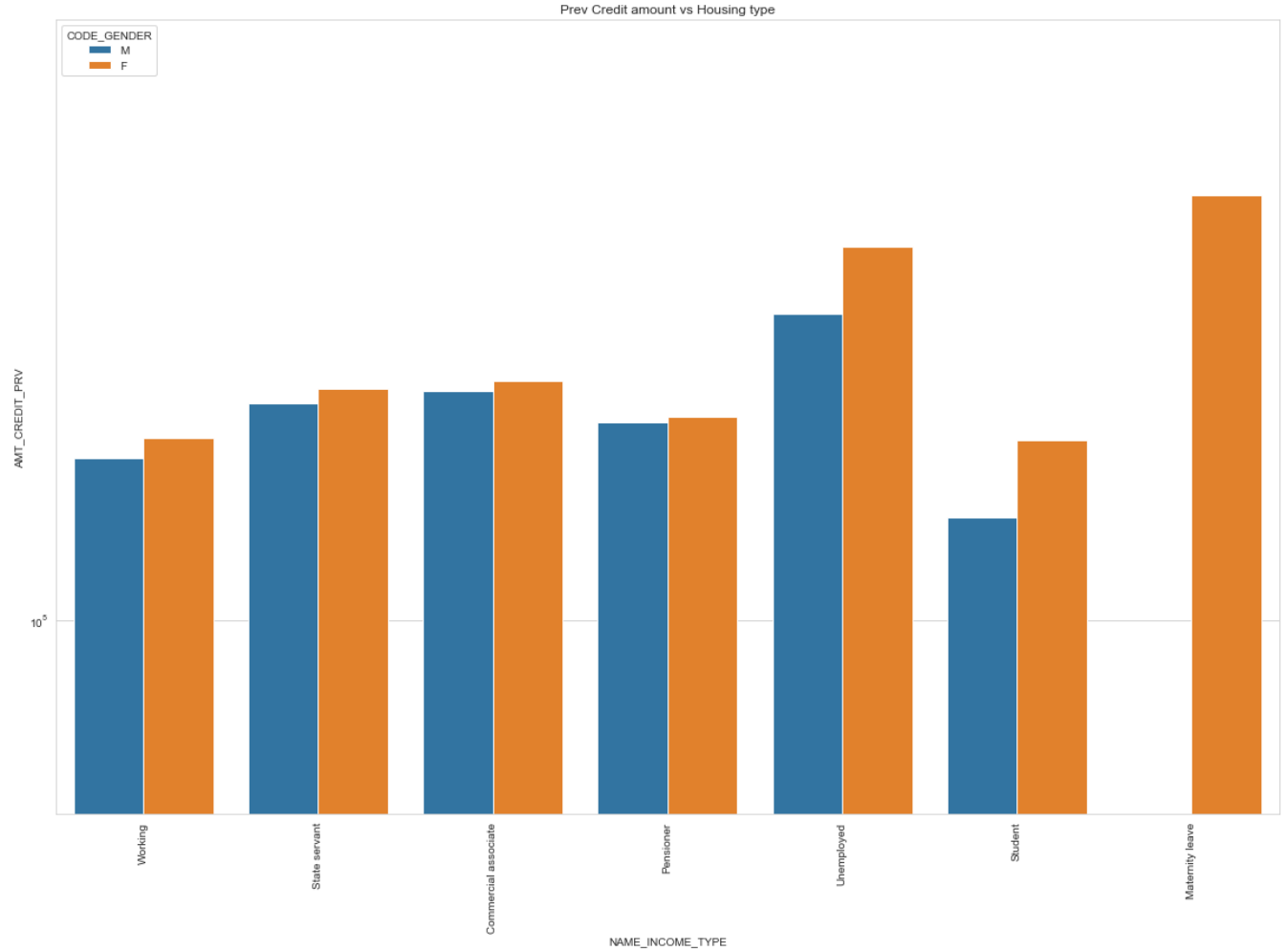


The loans that have been approved, cancelled, refused and unused have same median income amount for male customers.
The income of male customers is slightly higher than females in all the loan contract statuses.
The distribution of income amount is almost similar in all the loan contract statuses



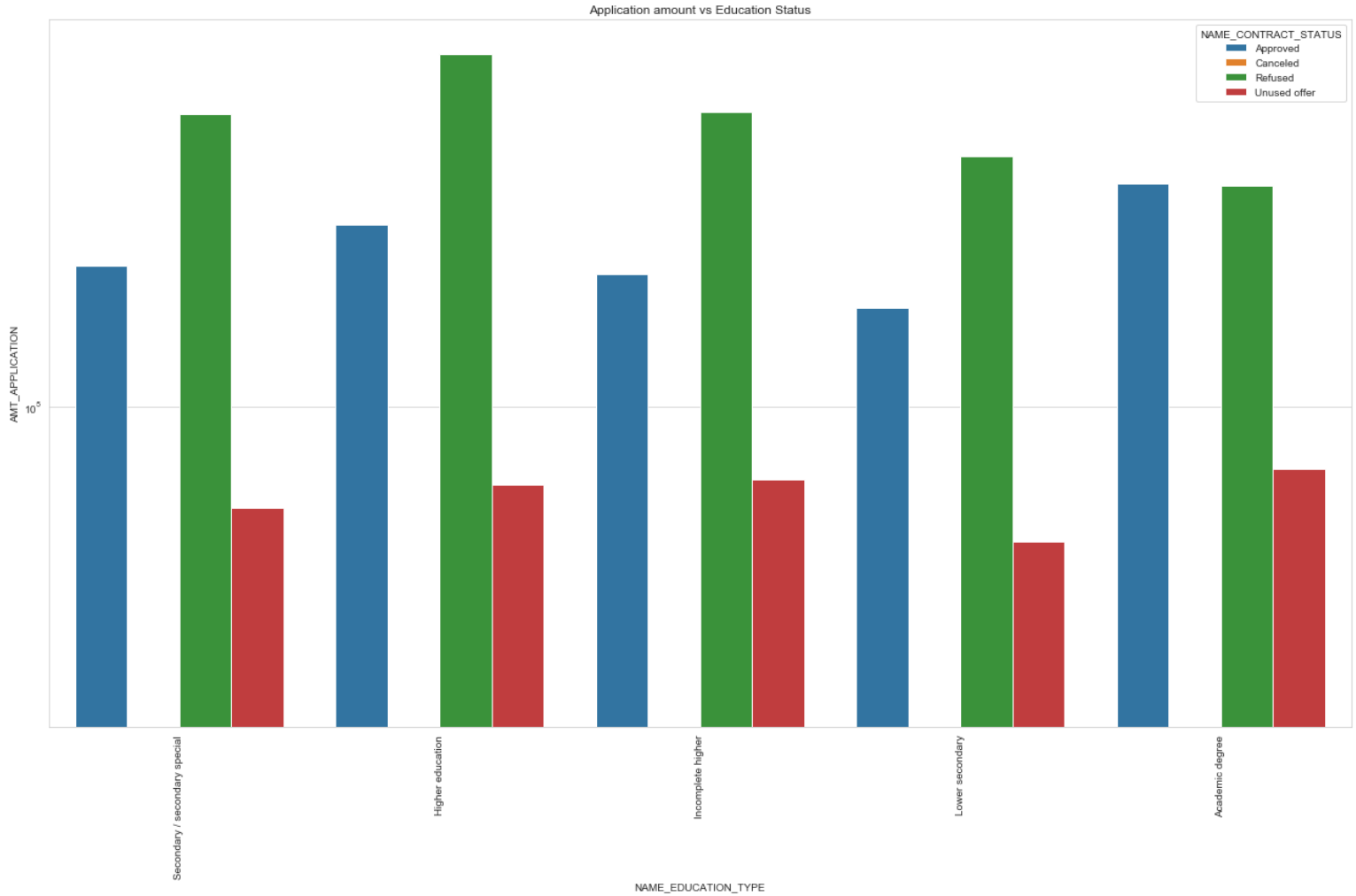
The loan amount that is credited in the current application is higher than that in previous application for all the income type groups

Highest loan amount is credited for people on maternity leave in both current and previous application.



Females across all the occupation type were given higher loan amount compared to males

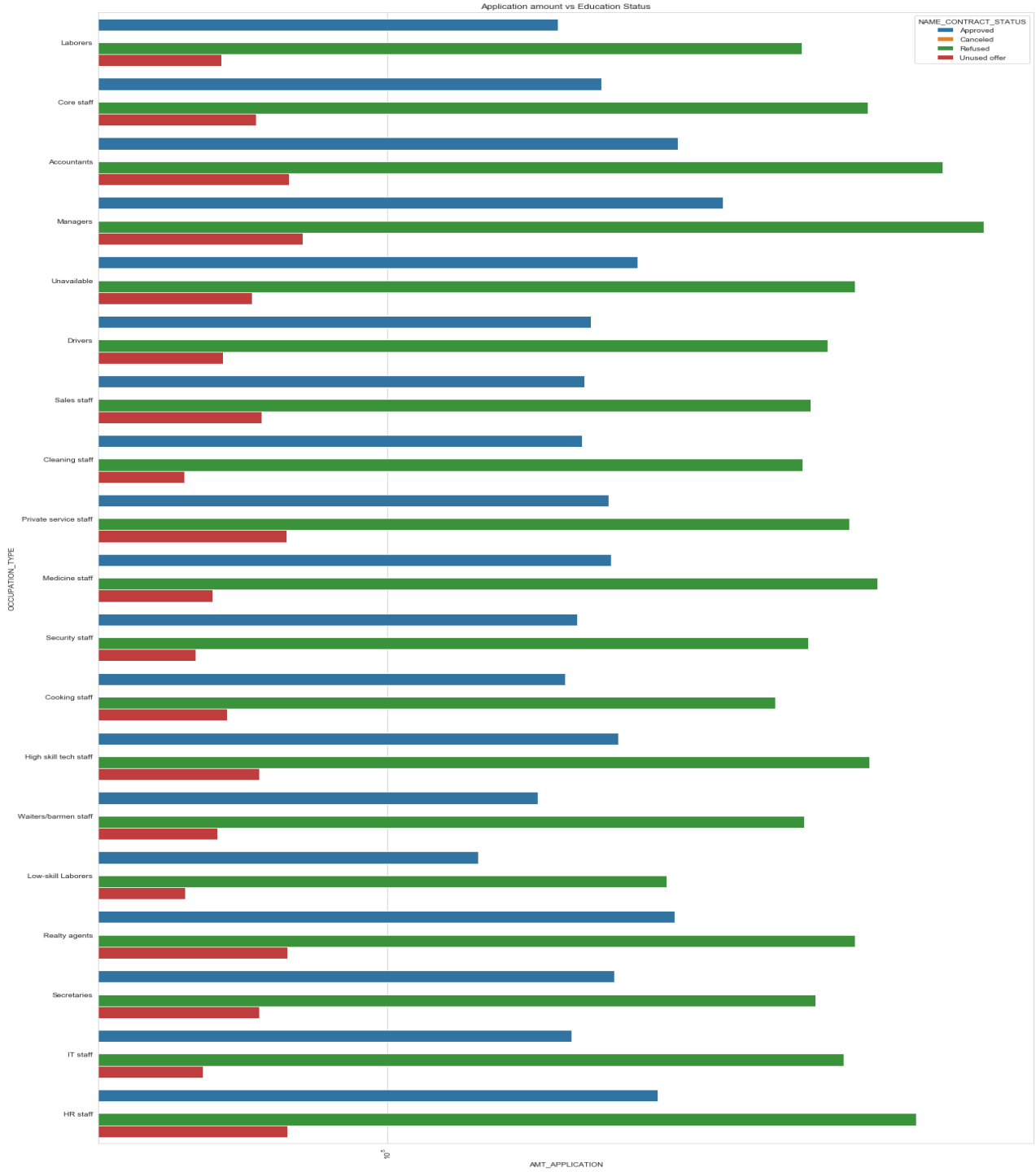
Unemployed females had highest loan amount in their previous loan application

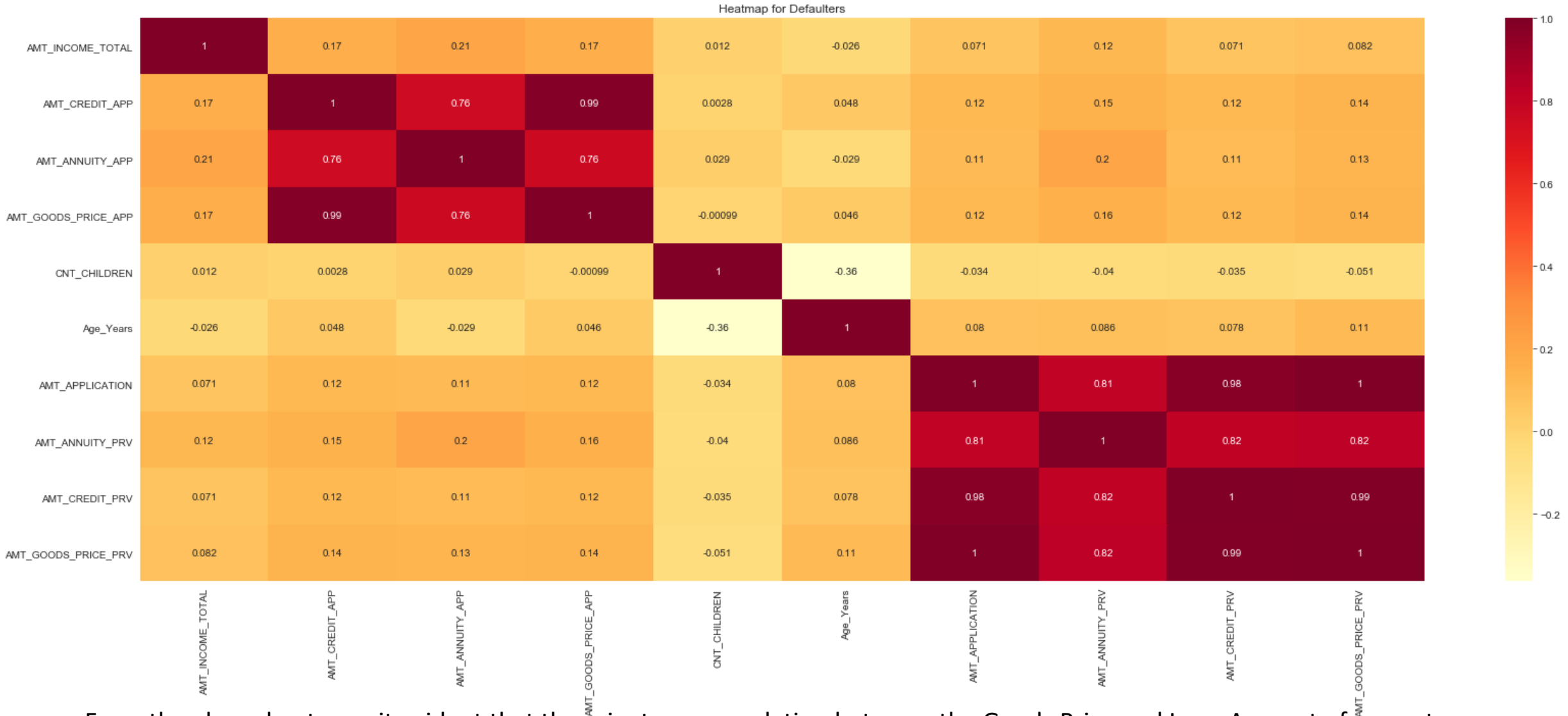


For the customers with academic degree as educational qualification, the amount of loan asked by the customer that was accepted and rejected is nearly same. Also, people in this category have cancelled loan with highest loan amounts. Amount of loan offered by the bank and which were unused by customers are almost same across all education types.

For the higher amount of loan asked by customers there is high chance of getting loan application refused, this is highest with customers on Managerial posts.

For the people with occupation type Accountants and Real estate agents the number of application that were accepted are high and highest for the customer on Managerial post.





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.

On the other hand, There is weak correlation between the age and count of children and also between the Goods price for the loan, loan amount credited and loan anuity of current and previous application

Conclusion

- Applicants having an academic degree have defaulted less number of loans compared to applicants from other educational backgrounds.
- People with higher secondary and married status have defaulted loans with higher loan amounts.
- Number of defaulters are higher in females than males.
- There is strong relation in loan amount credited by bank and loan amount asked by customer against goods.
- Also with loan purpose ‘Repair’ is having higher number of approved and rejected loans.
- Bank is expected to receive higher number of loan applications on Tuesdays than other days

Thank You !