Credit EDA Assignment

Sushree Rutayanee

Batch DSC-59

Index:

- Problem statement
- Steps followed for performing EDA
- Plots and the respective inferences derived

Problem statement:

In this case study, we are using EDA to draw insights from the data (i.e. current loan application and previous loan application details of a client) to see on which combination of variables - a client will be able to repay the loan.

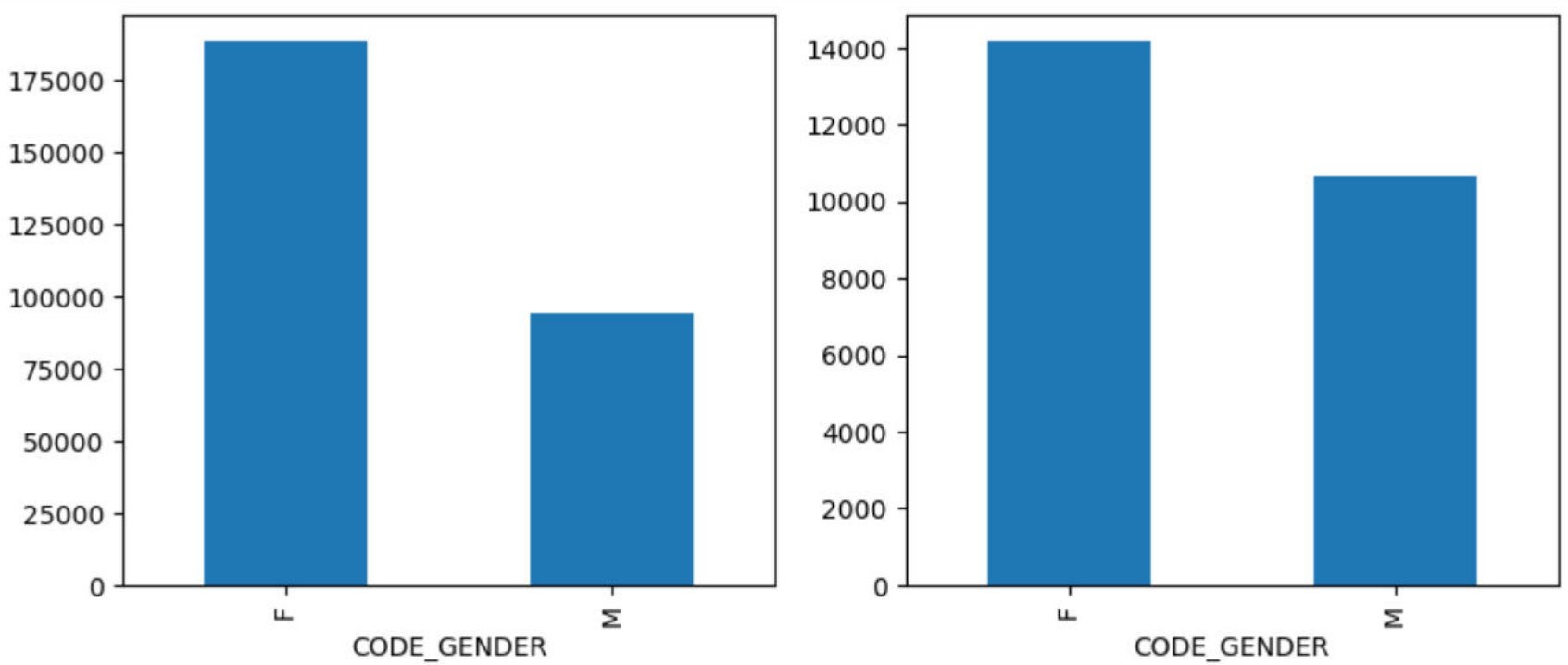
It will help us in understanding which clients are having payment difficulties and client who will be able to fulfil based on which the loan company will know which loan to approve and which not to.

Steps followed for performing EDA:

1. We imported all the necessary libraries like numpy (for numerical analysis), pandas (for data manipulation), matplotlib and seaborn (for visualization).
2. We loaded both the datasets provided (application_data and previous_application)
3. We checked the structures and performed data cleaning for both the datasets which included steps like fixing missing values (where >50% we dropped the columns and >13% we wrote suggestions on how they can be fixed), fixed the data types where needed, fixed negative values, then detected outliers and wrote suggestions for them in comments, binned few variables)
4. Next we performed the univariate, bivariate analysis on the application_data dataset by splitting it into 2 dataframes based on the target variable. We also plotted the correlation for the same.
5. Next we merged both the datasets and then performed univariate and bivariate analysis on both.

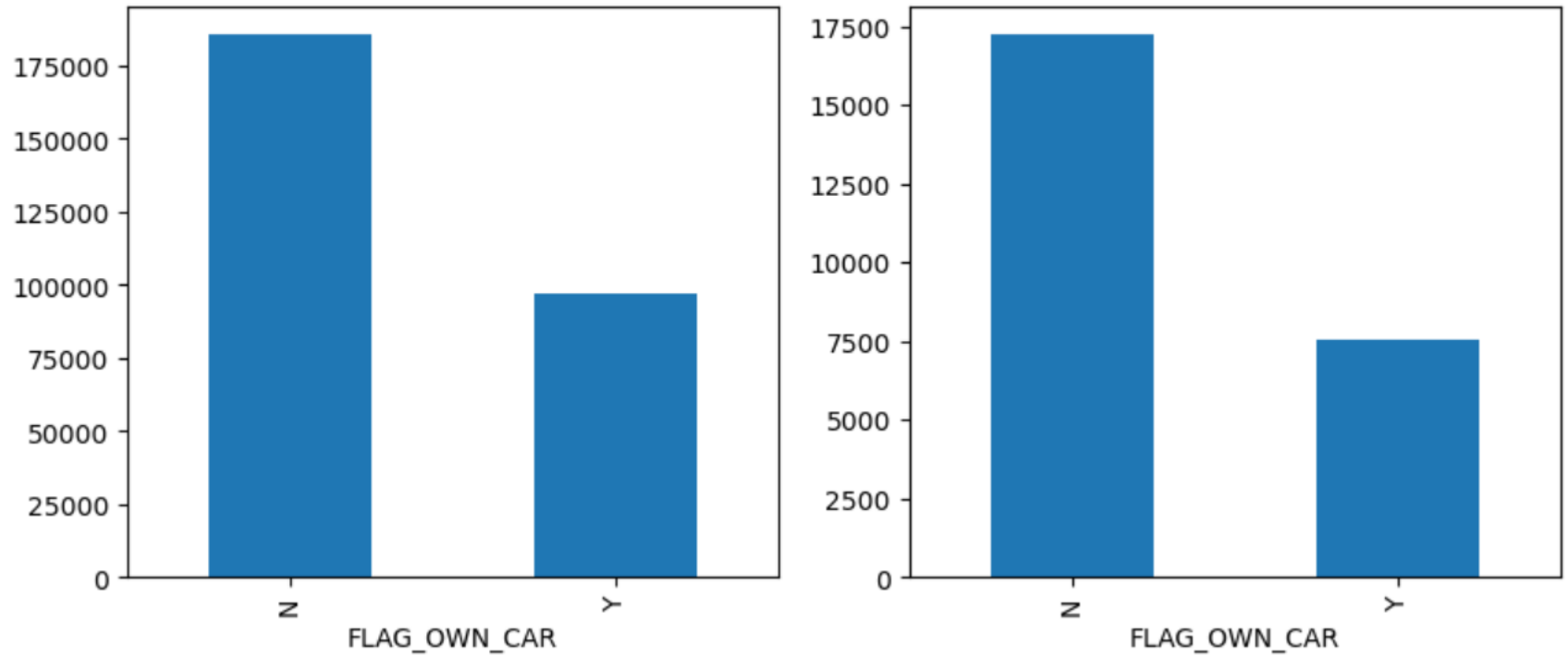
Plots and the respective inferences derived:

1. We splitted the application_data into 2 data frames based on the Target variable (i.e. 0 for clients with payment difficulties and 1 for other). We performed univariate analysis on gender column in both the dataframes. Left graph denotes people with payment difficulties and right denotes people who have done payments on time.



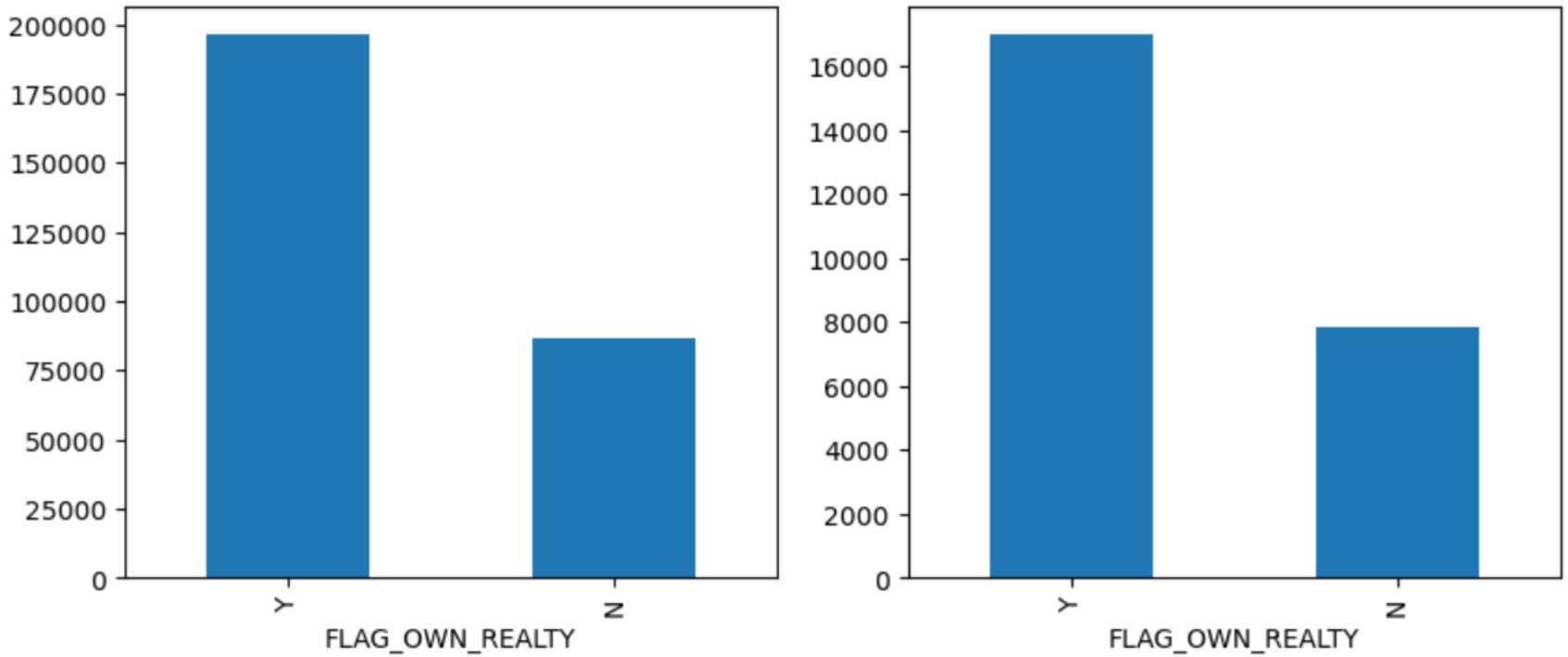
Inference: Defaulters are more in female category than male. Coming to non-defaulters, females are more here as well than males.

2. Next we performed univariate analysis on FLAG_OWN_CAR column in both the data frames. Left graph denotes people with payment difficulties and right denotes people who have done payments on time.



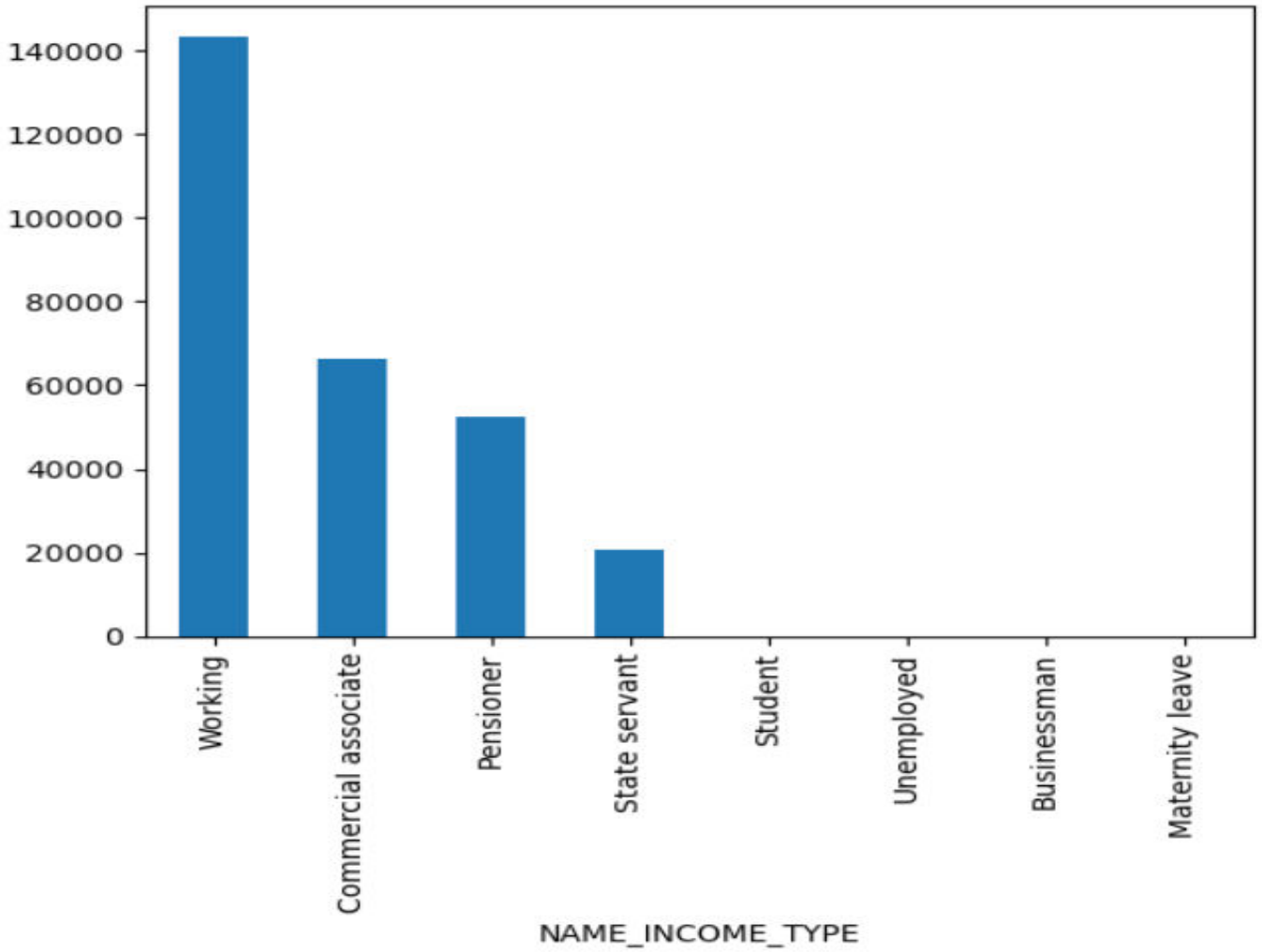
Inference: People who own a car have less payment difficulties than people who don't own a car.

3. Next we performed univariate analysis on FLAG_OWN_REALTY column in both the data frames. Left graph denotes people with payment difficulties and right denotes people who have done payments on time.



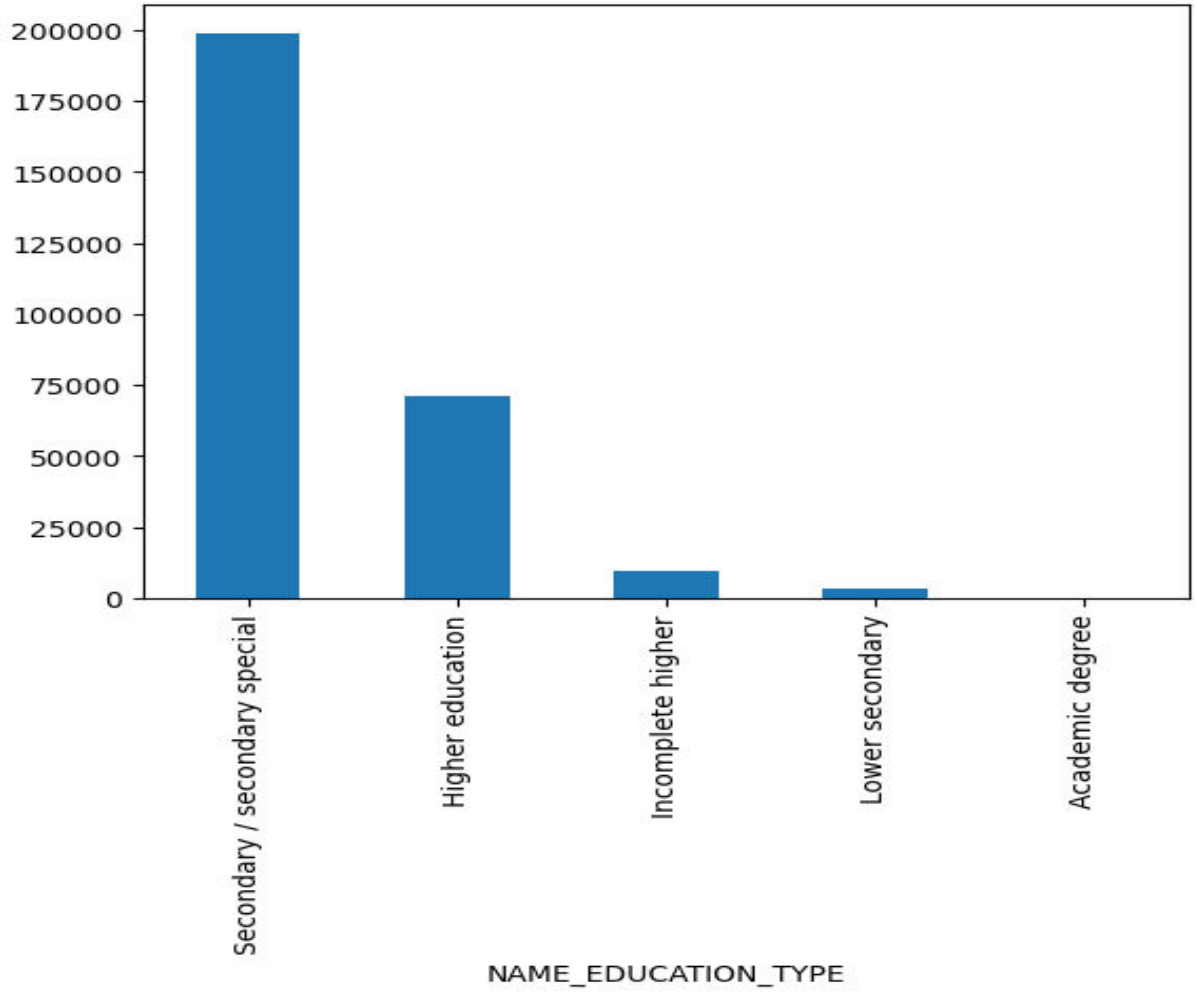
Inference: People who own a house or flat have more payment difficulties than people who don't own.

4. Next we performed univariate analysis on NAME_INCOME_TYPE column in the data frame where Target is 0 i.e. the data frame which contains people with payment difficulties.



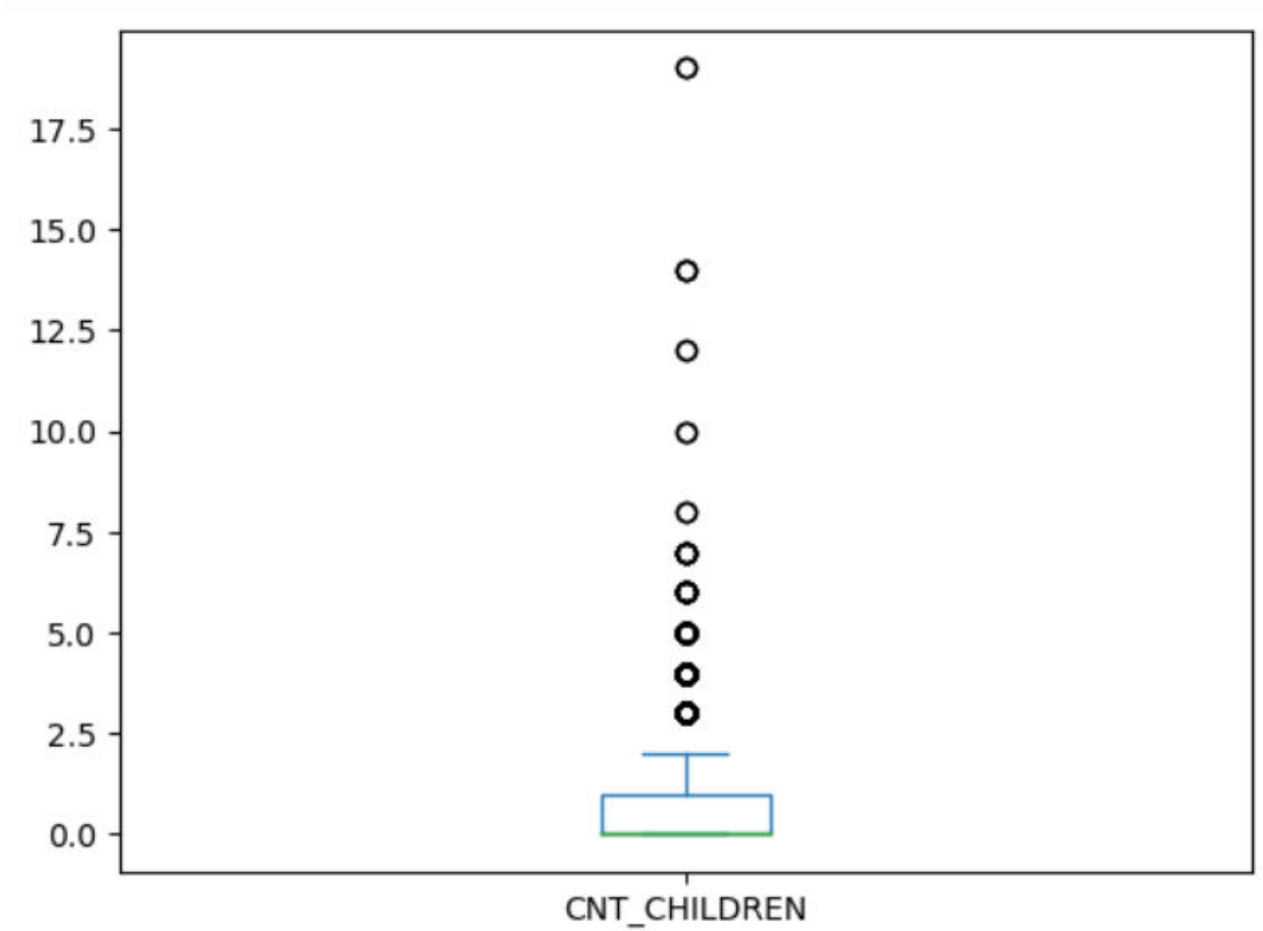
Inference: People who are working have more payment difficulties than other categories like student, businessman, commericial associate, pensioner, etc.

5. Next we performed univariate analysis on NAME_EDUCATION_TYPE column in the data frame where Target is 0 i.e. the data frame which contains people with payment difficulties.



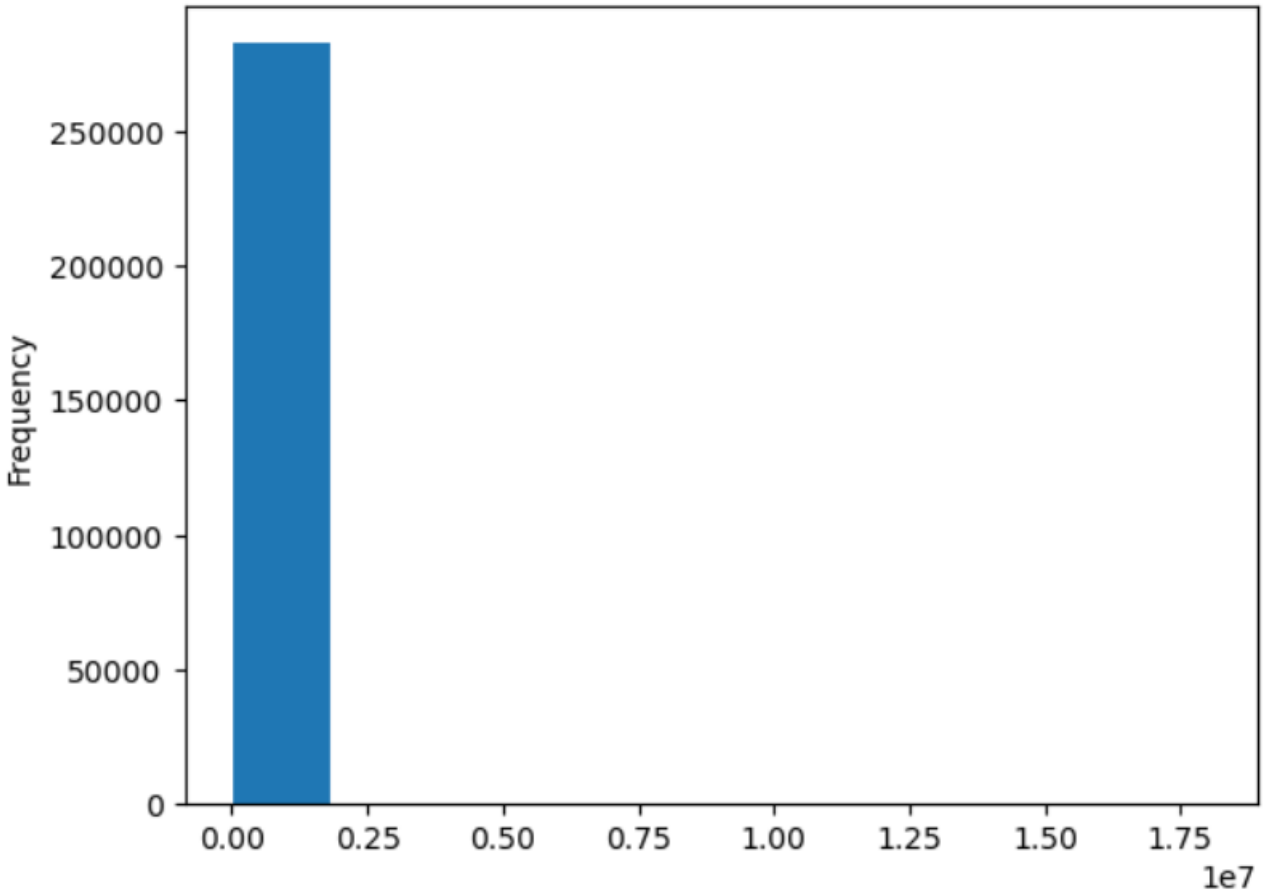
Inference: Secondary education people have more payment difficulties than people with higher education.

6. Next we performed univariate analysis on CNT_CHILDREN column in the data frame where Target is 0 i.e. the data frame which contains people with payment difficulties.



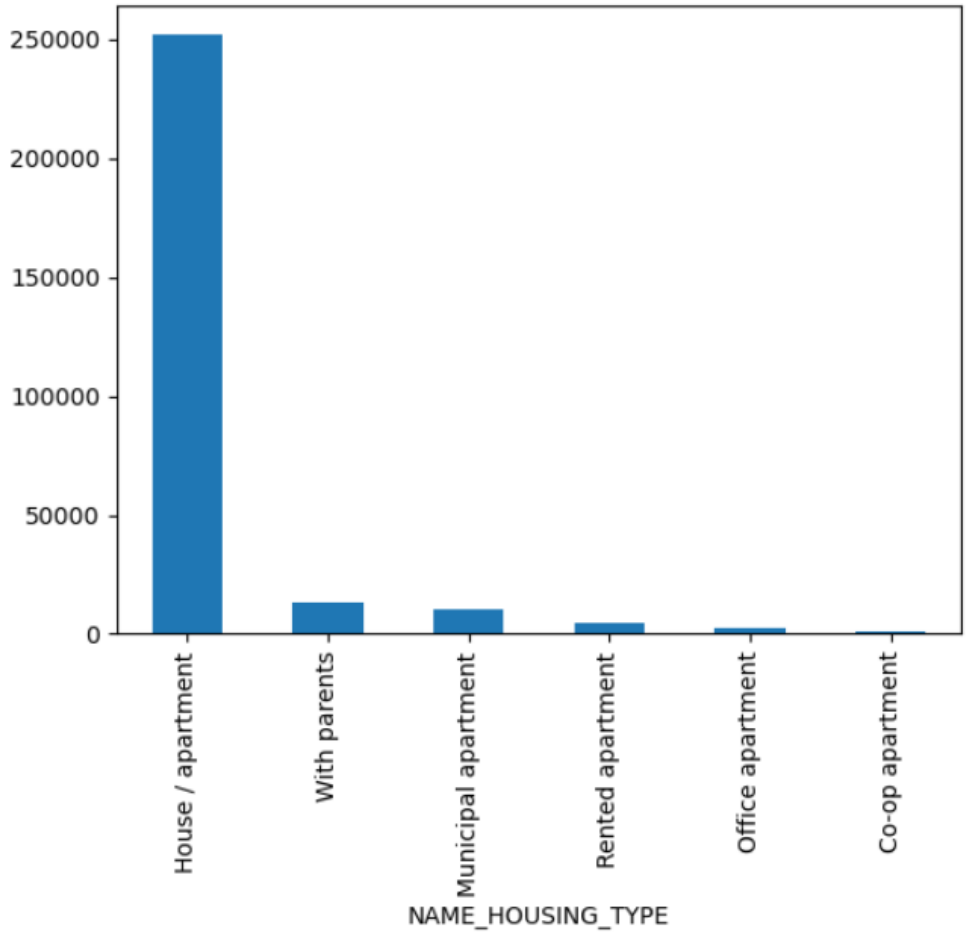
Inference: People with more children have more payment difficulties.

7. Next we performed univariate analysis on AMT_INCOME_TOTAL column in the data frame where Target is 0 i.e. the data frame which contains people with payment difficulties.



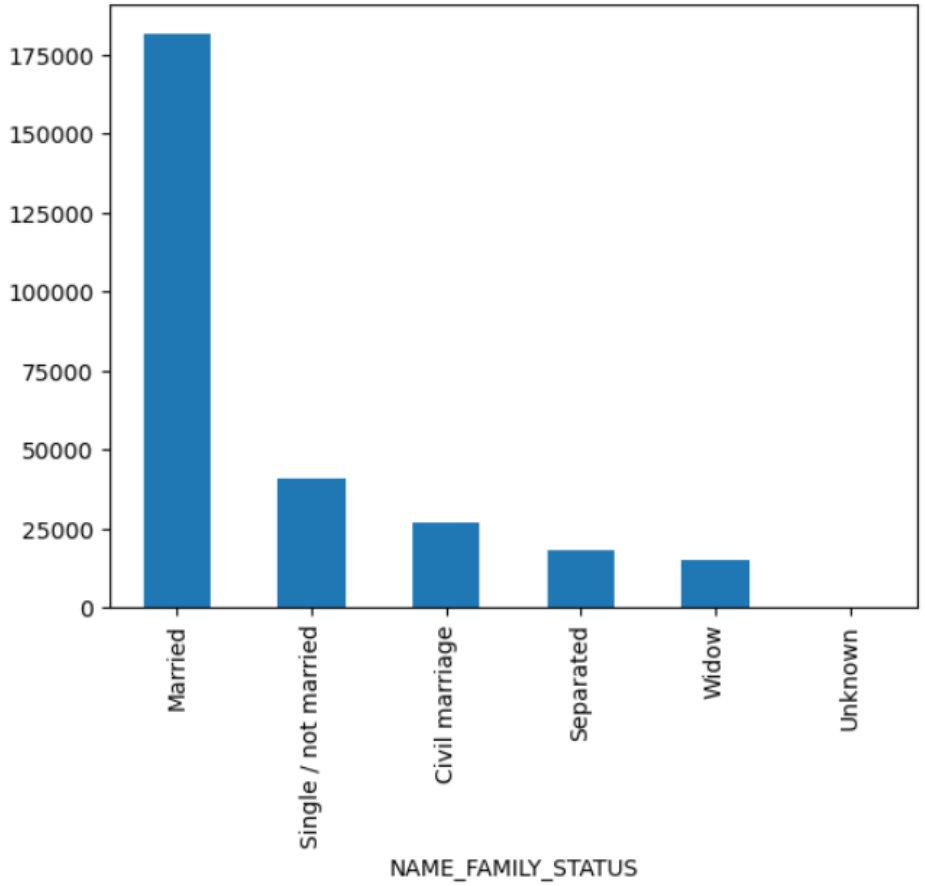
Inference: People with more income have less payment difficulties

8. Next we performed univariate analysis on NAME_HOUSING_TYPE column in the data frame where Target is 0 i.e. the data frame which contains people with payment difficulties.



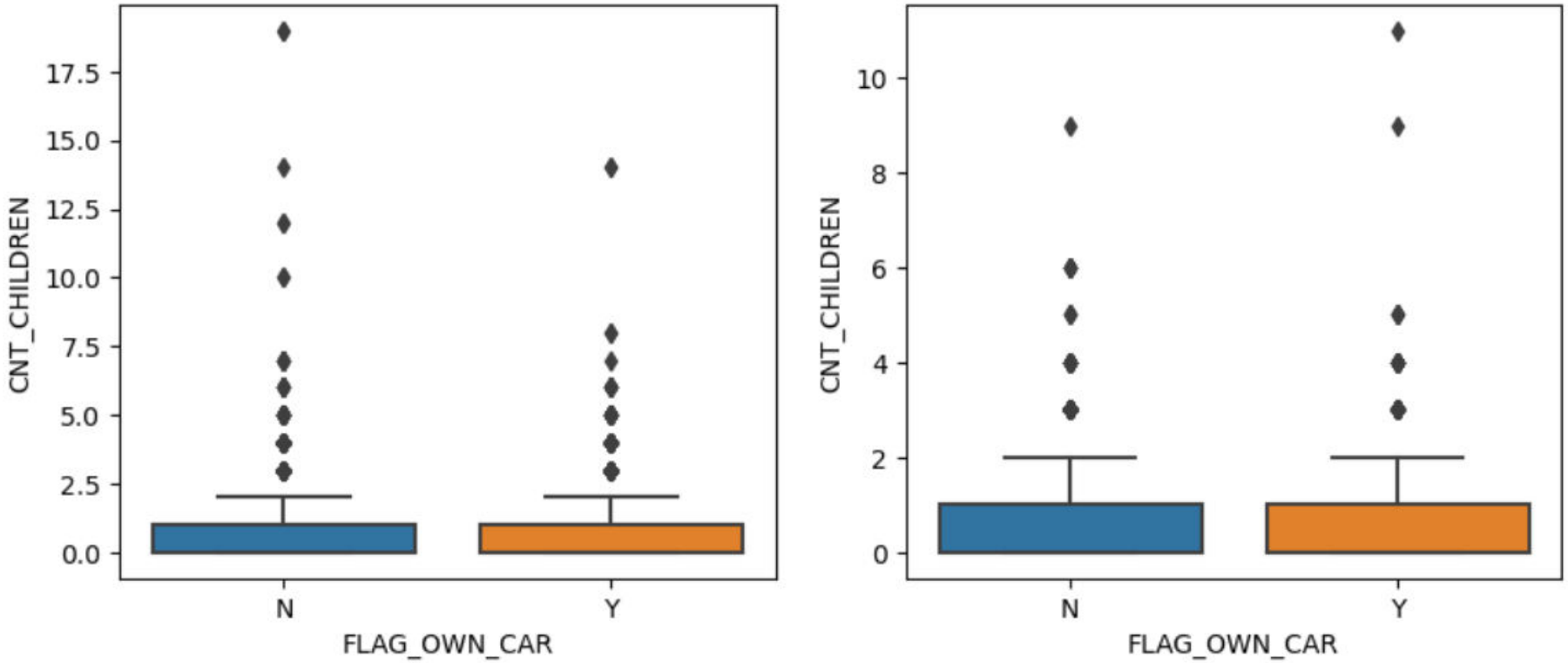
Inference: People living in house/apartment have more payment difficulties than people living in rented apartments, office apartments and co-op apartments.

9. Next we performed univariate analysis on NAME_FAMILY_STATUS column in the data frame where Target is 0 i.e. the data frame which contains people with payment difficulties.



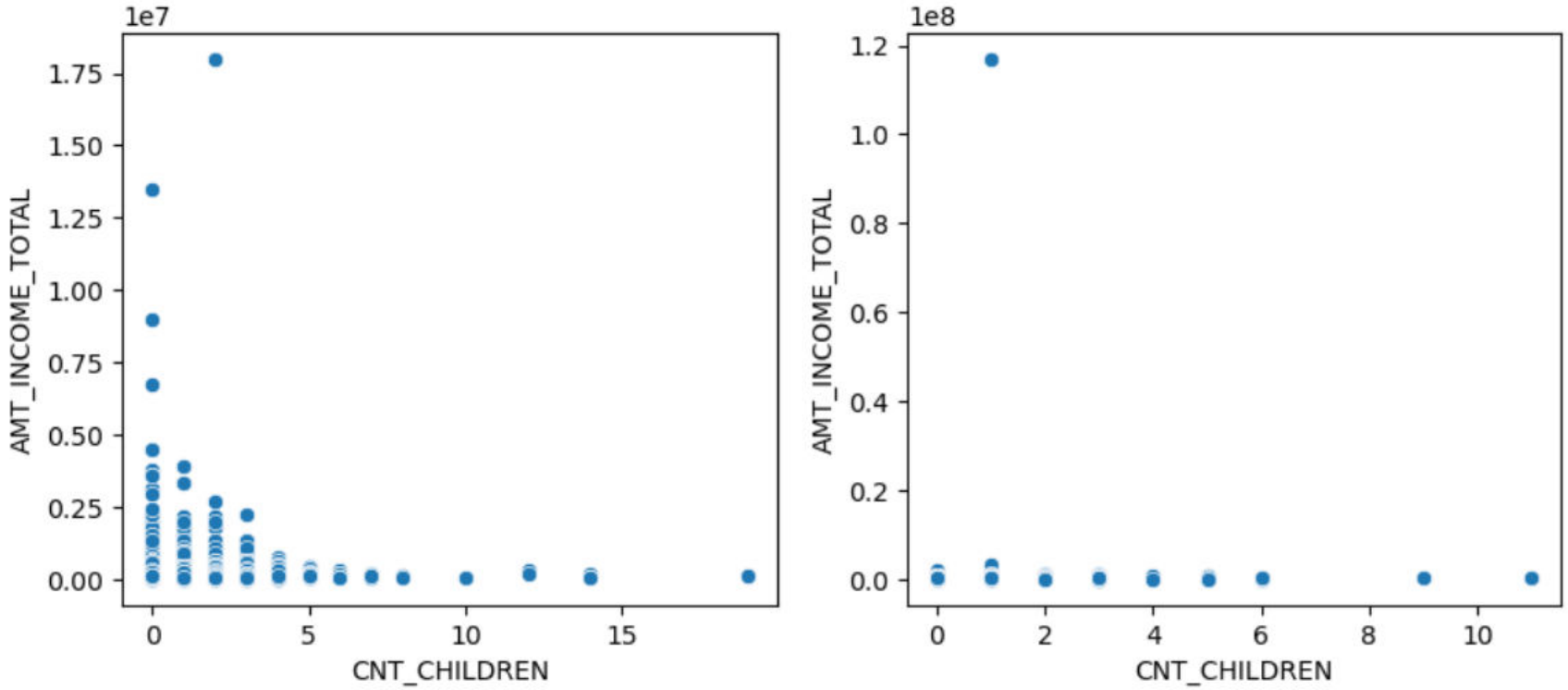
Inference: People who are married have more payment difficulties than people who are seperated or widowed.

10. Next we performed bivariate analysis between FLAG_OWN_CAR and CNT_CHILDREN columns in both the data frames (one with Target = 0 and second with Target = 1). Left graph denotes people with payment difficulties and right denotes people who have done payments on time.



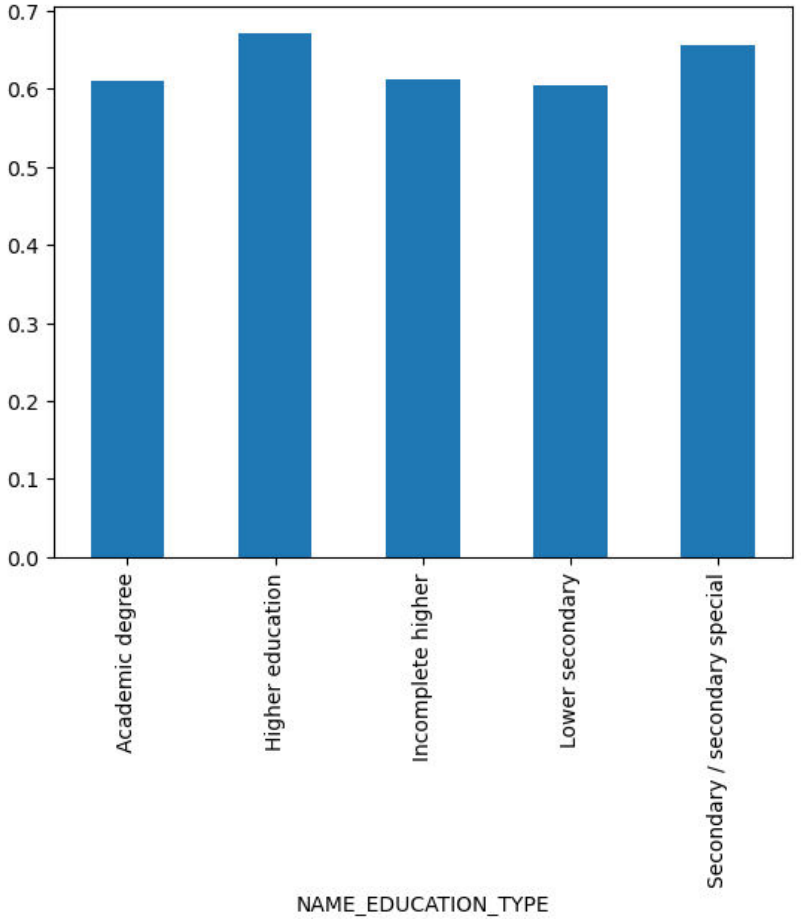
Inference: People with payment difficulties who have more children don't own a car. Few people with payment difficulties who having less children can at least own a car. People who don’t have much payment difficulties own a car even though having more children.

11. Next we performed bivariate analysis between CNT_CHILDREN and AMT_INCOME_TOTAL columns in both the data frames (one with Target = 0 and second with Target = 1). Left graph denotes people with payment difficulties and right denotes people who have done payments on time.



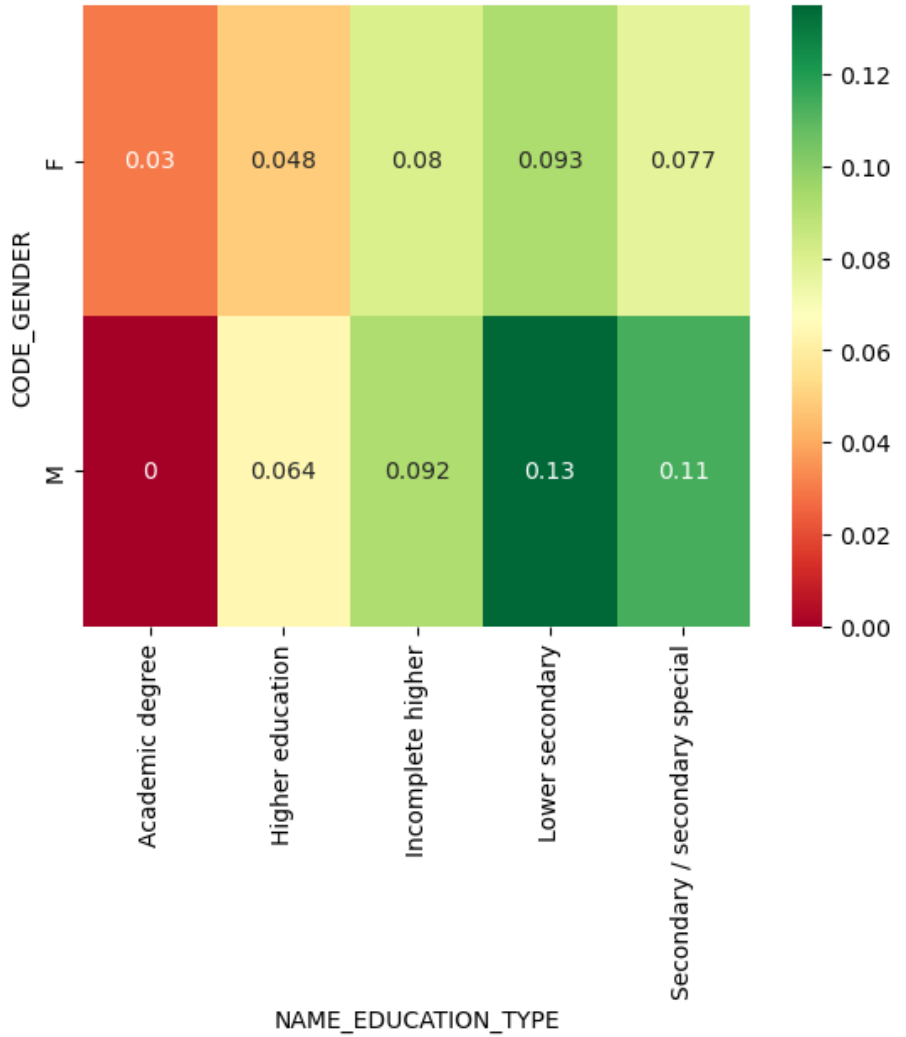
Inference: People who have more children and less income have payment difficulties. People who have less children and good income have less payment difficulties.

12. Next we performed bivariate analysis between NAME_EDUCATION_TYPE and Gender_Num columns in the application_data dataframe.



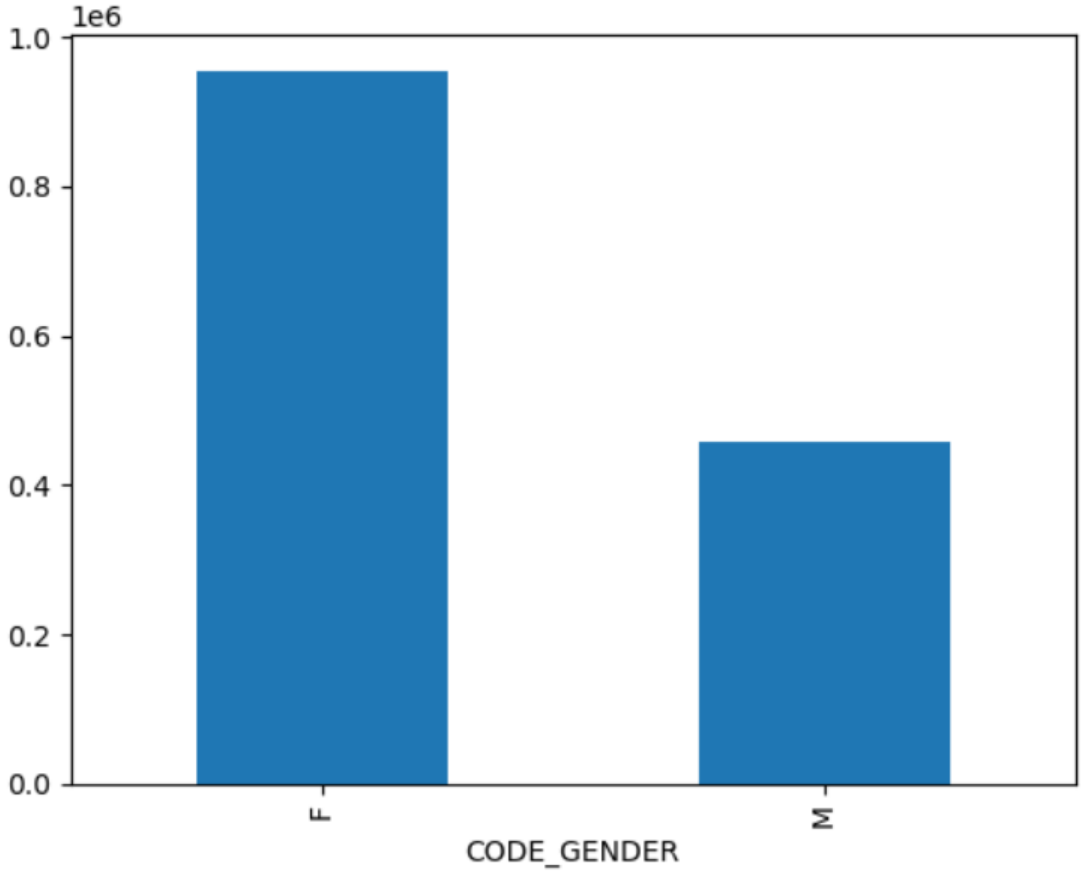
Inference: Females seem to be more educated than males.

13. Next we plotted a correlation matrix (heatmap) on a pivot table consisting of index (CODE_GENDER), columns (NAME_EDUCATION_TYPE) and values (TARGET). Target 0 indicates people with payment difficulties and 1 indicates people who give payments on time.



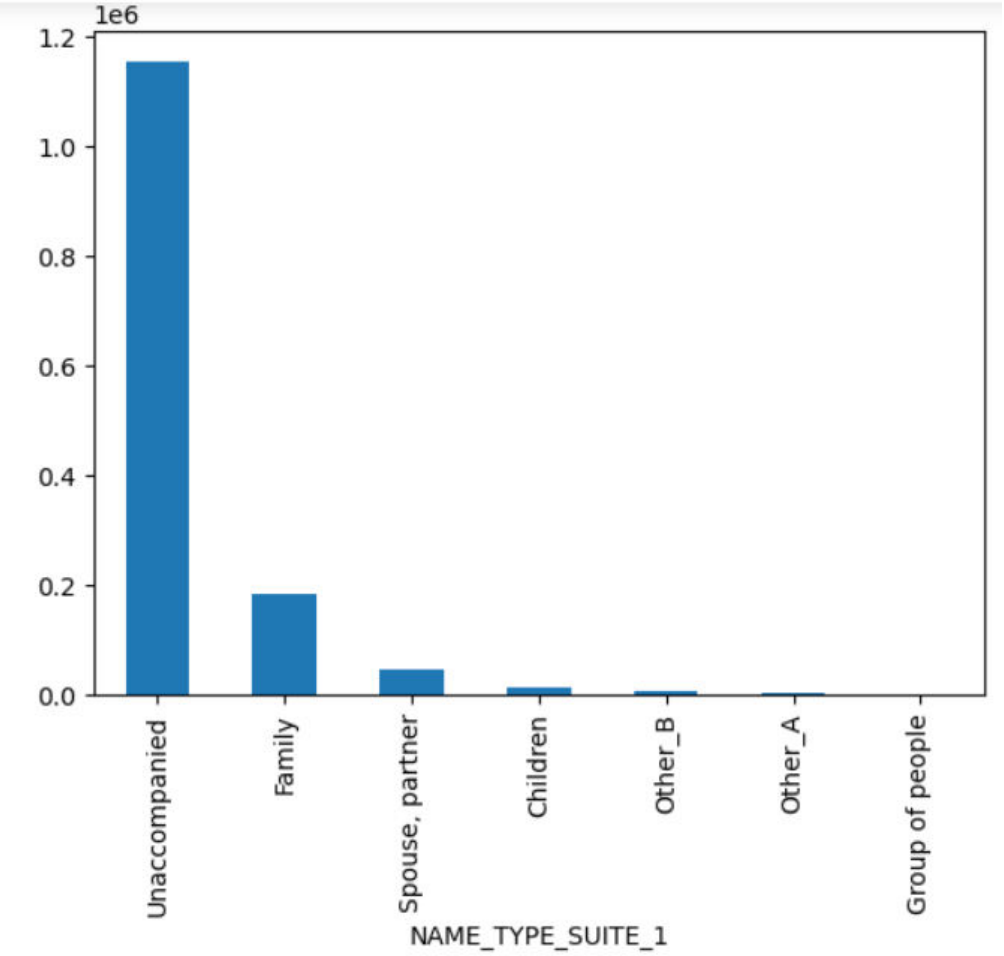
Inference: Men with secondary education have less payment difficulties whereas males and females with academic degree have more payment difficulties

14. Next we merged both the dataframes “application_data” and “previous_application”. We performed univariate analysis on CODE_GENDER column.



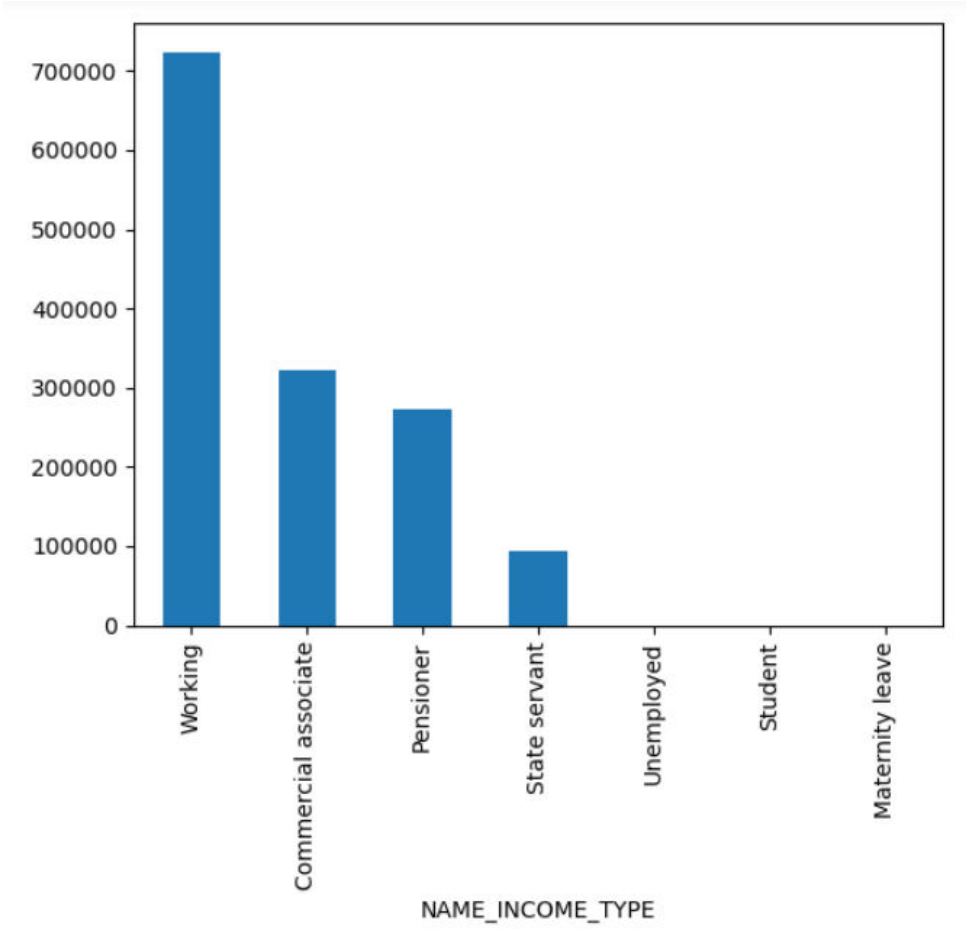
Inference: As per this dataset merge, females are more in number who came for applying loan than male.

15. Next we merged both the dataframes “application_data” and “previous_application”. We performed univariate analysis on NAME_TYPE_SUITE_1 column.



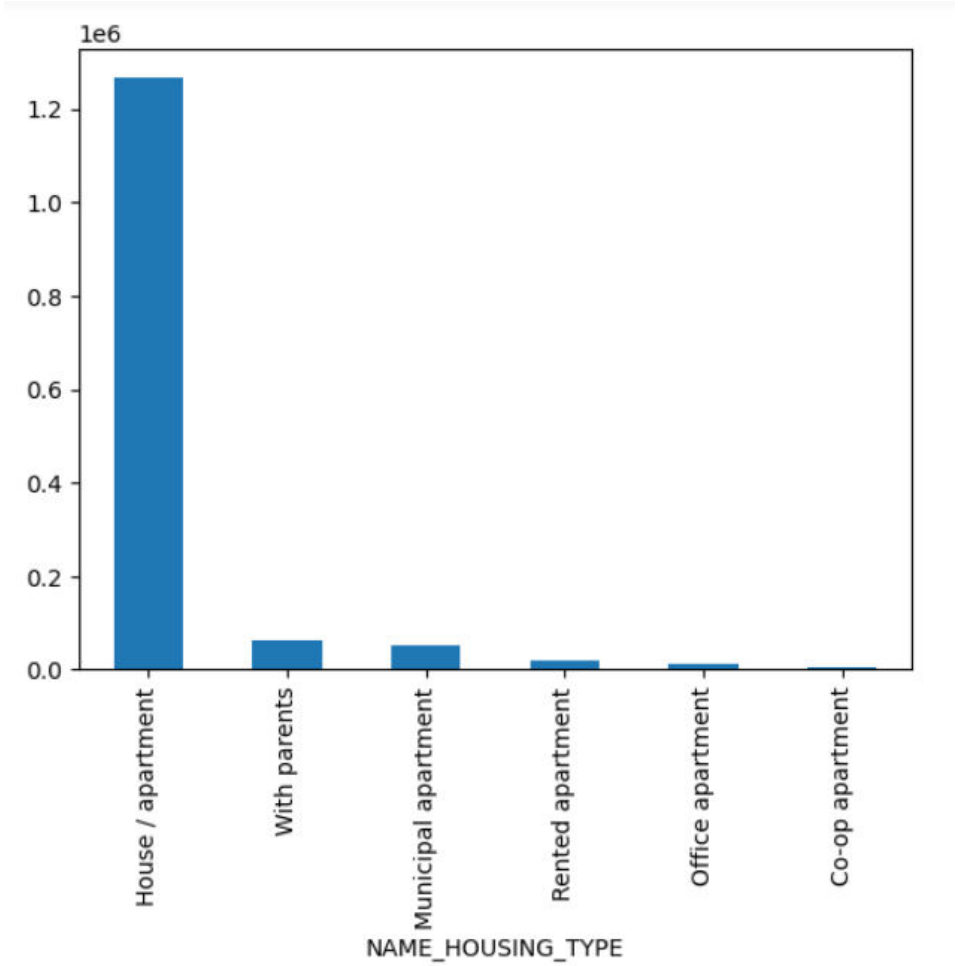
Inference: Most of the people who came for applying loan came alone and very less people came with their family or spouse or children.

16. Next we merged both the dataframes “application_data” and “previous_application”. We performed univariate analysis on NAME_INCOME_TYPE column.



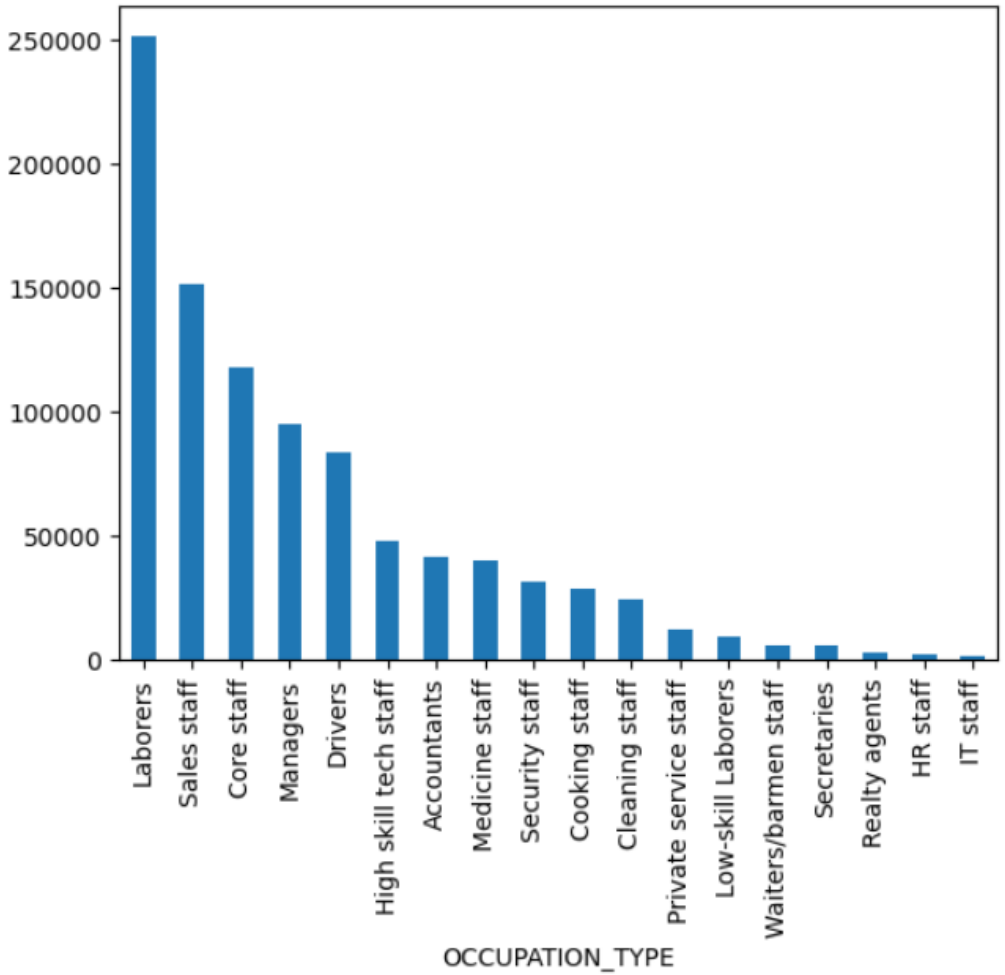
Inference: Working people are the ones who are more in need of applying for loan than other categories.

17. Next we merged both the dataframes “application_data” and “previous_application”. We performed univariate analysis on NAME_HOUSING_TYPE column.



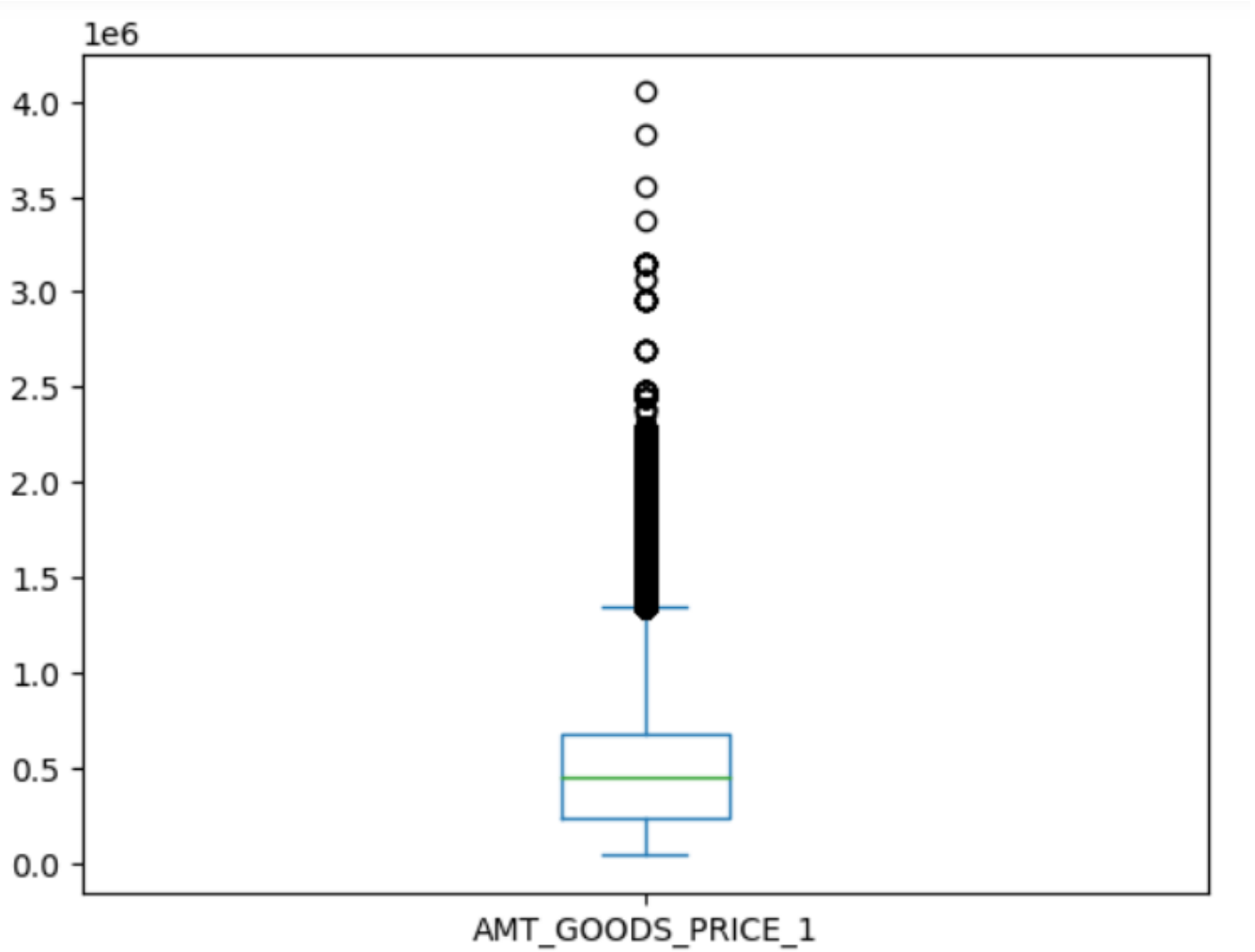
Inference: People living in apartments are coming in more to apply for loans than people who live with their parents or in rented apartments or office apartment

18. Next we merged both the dataframes “application_data” and “previous_application”. We performed univariate analysis on OCCUPATION_TYPE column.



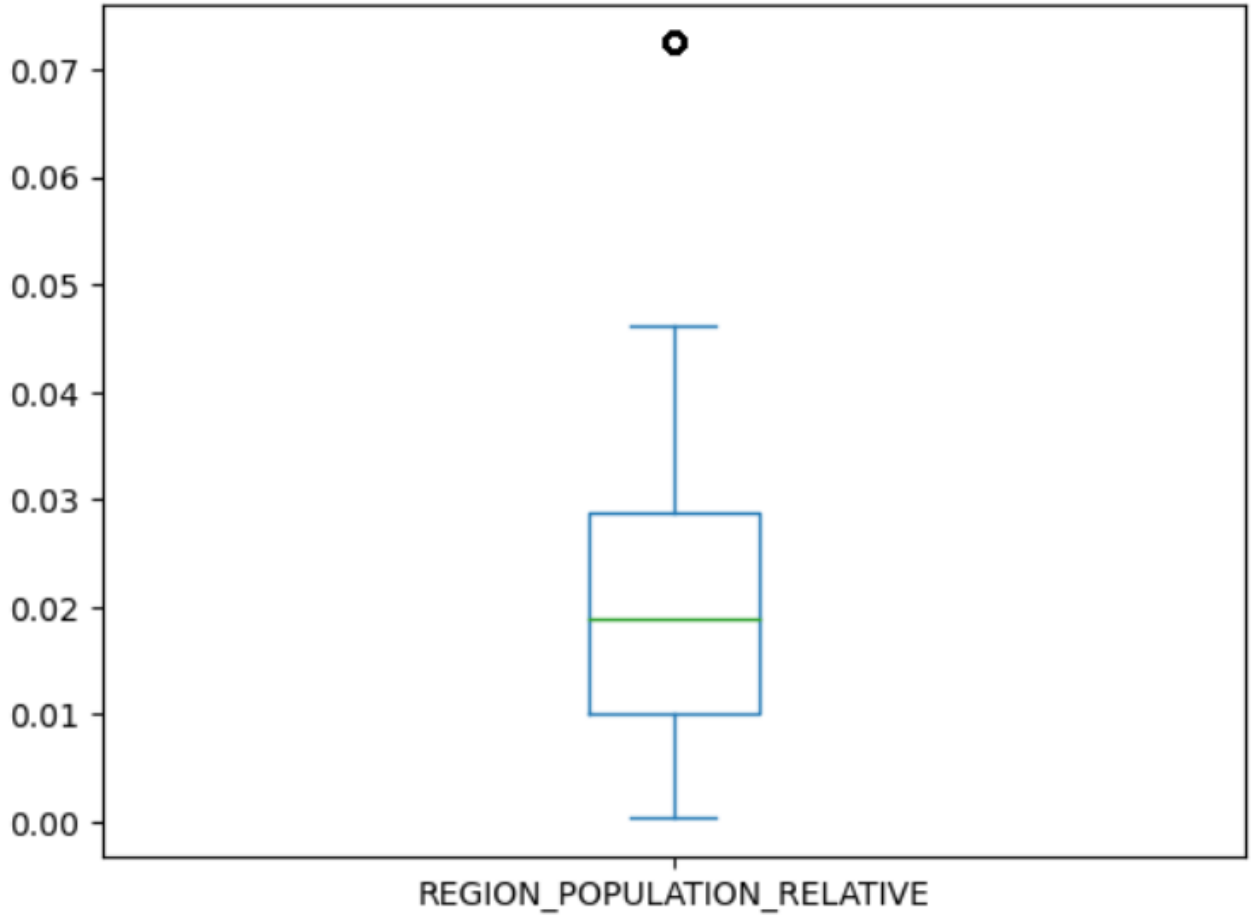
Inference: People who are laborers are coming in more to apply for loans than people of other occupation.

19. Next we merged both the dataframes “application_data” and “previous_application”. We performed univariate analysis on AMT_GOODS_PRICE_1 column.



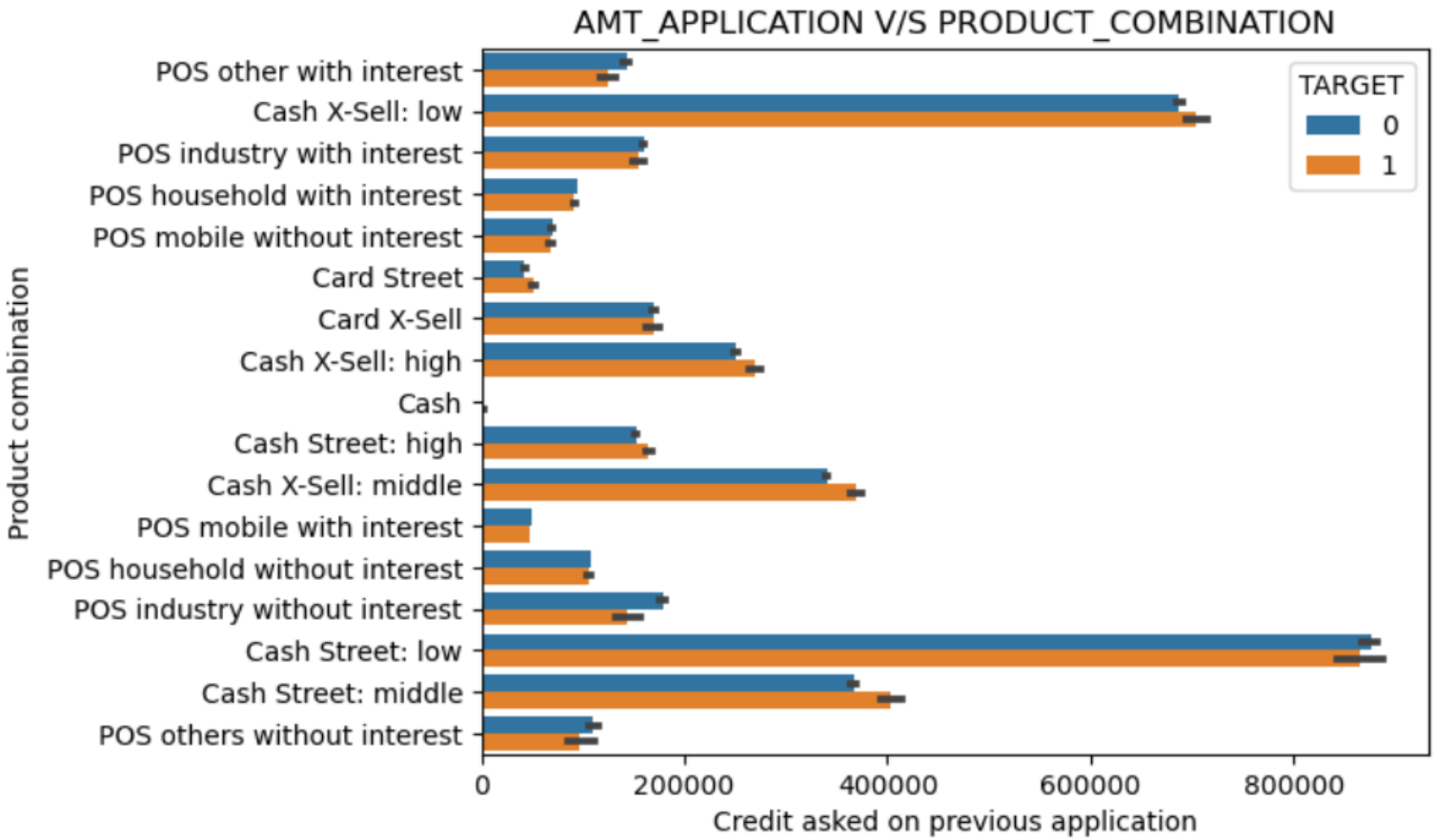
Inference: Loan applications are taken more on costly goods.

20. Next we merged both the dataframes “application_data” and “previous_application”. We performed univariate analysis on REGION_POPULATION_RELATIVE column.



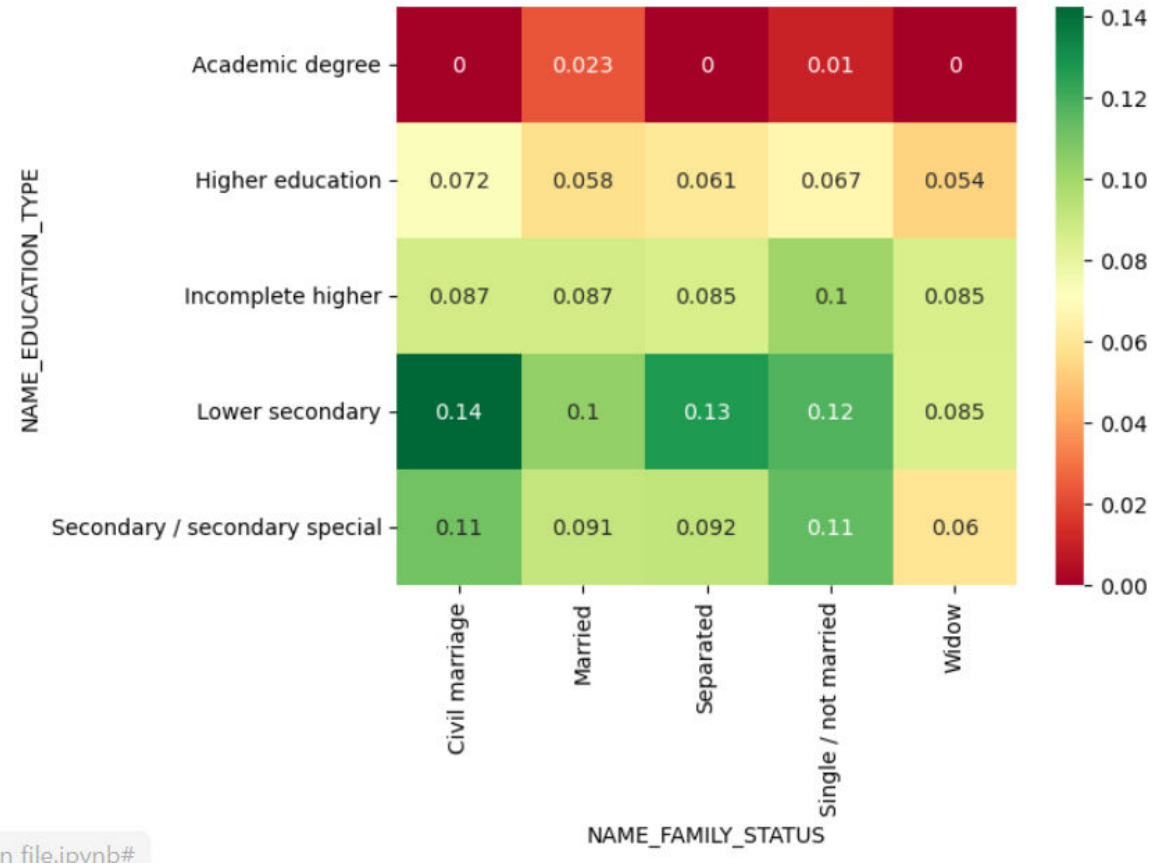
Inference: People living in more populated area are coming in to apply for Loan.

21. Next we merged both the dataframes “application_data” and “previous_application”. We performed bivariate analysis on AMT_APPLICATION and PRODUCT_COMBINATION column.



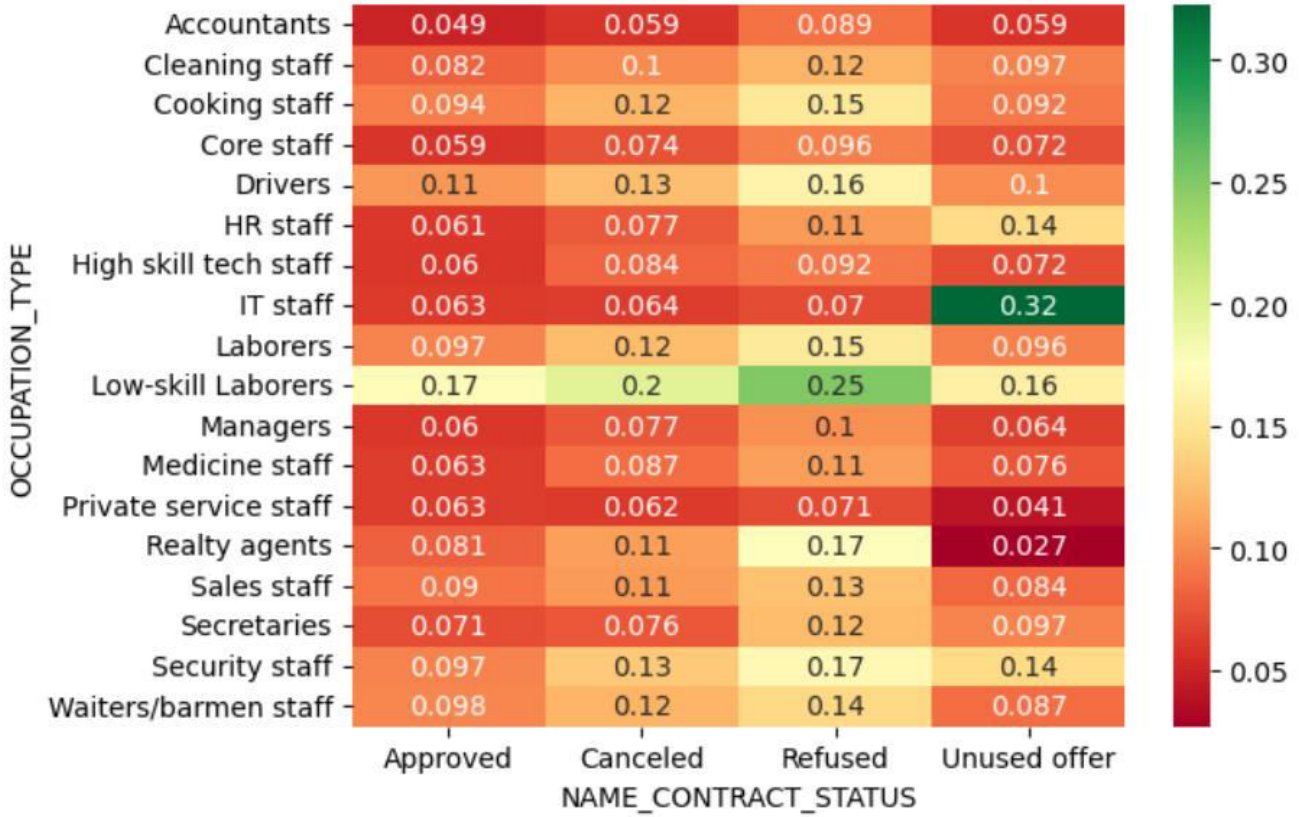
Inference: Many People with payment difficulties have asked more credit on low cash street. Few people with payment difficulties have asked less credit on card street. People who don’t have much payment difficulties have asked less credit on POS mobile with interest

22. Next we merged both the dataframes “application_data” and “previous_application”. We performed correlation by using a heatmap by using index as NAME_EDUCATION_TYPE, columns as NAME_FAMILY_STATUS and values as TARGET. Target 0 indicates people with payment difficulties and 1 indicates people who give payments on time.



Inference: Many People with payment difficulties have asked more credit on low cash street. Few people with payment difficulties have asked less credit on card street. People who don’t have much payment difficulties have asked less credit on POS mobile with interest

23. Next we merged both the dataframes “application_data” and “previous_application”. We performed correlation by using a heatmap by using index as OCCUPATION_TYPE, columns as NAME_CONTRACT_STATUS and values as TARGET. Target 0 indicates people with payment difficulties and 1 indicates people who give payments on time.



Inference: IT Staff who have less payment difficulties compared to others seem to have cancelled the loan during various stages of the process. Low skill laborers got their loan approved better than other occupation types.

Conclusions from the above study:

1. People with higher education can be given a loan approval as they have less payment difficulties.
2. People from high income group can be given a loan approval.
3. Men with secondary education have less payment difficulties so they can be given a loan approval.
4. People living in rented apartments, office apartments or co-op apartments should be given a loan approval as they have less payment difficulties.
5. People who are separated or widowed should be given a loan approval as they have less payment difficulties.

Thank You