Credit EDA Case-Study

The first project while starting with Data Science

Introduction

This assignment aims to give us an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that we have learnt in the EDA module, we will also develop a basic understanding of risk analytics in banking and financial services and underst and how data is used to minimise the risk of losing money while lending to customers.

Business Understanding 1:

- 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. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

- When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

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

★ 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.

Business Understanding 2:

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the

  loan in our sample,

- All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. Approved: The Company has approved loan Application
2. Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
4. Unused offer: Loan has been cancelled by the client but on different stages of the process.

In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Business Objectives:

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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. 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.

In other words, 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 utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Data Understanding:

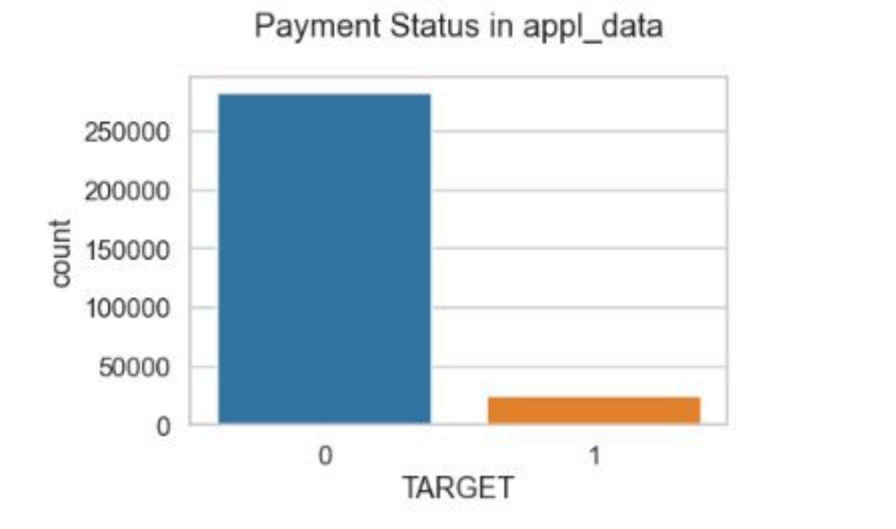
Given dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a client has payment difficulties.

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

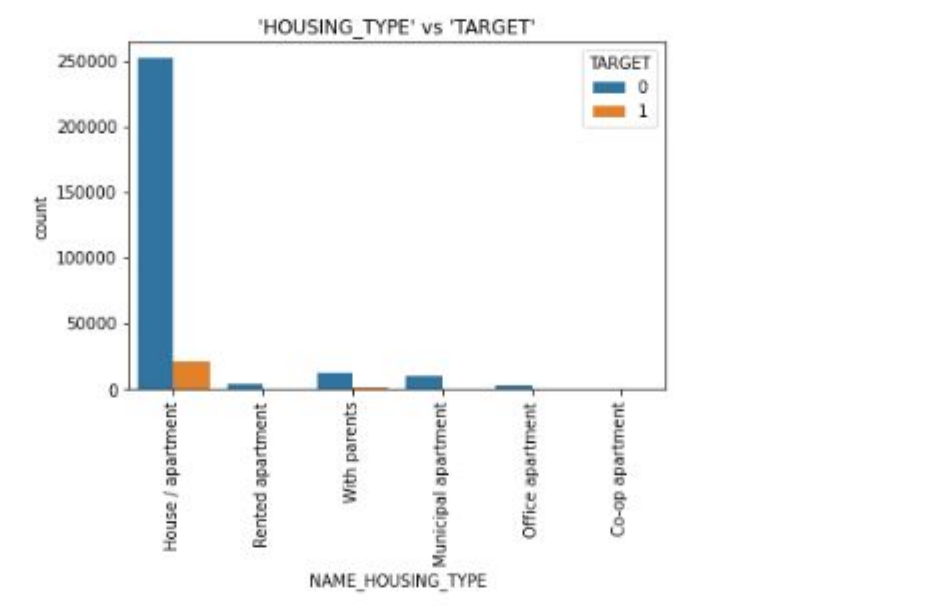
3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Firstly we will read our data sets with the help of pandas, so that we can observe those for analysis. We will check info() and describe() & shape to get some understanding on both the data-sets. We will try to Analyze the data in small segments as there are so many columns to look at altogether. We will also use some plot in the process of understanding/visualizing/analyzing data.

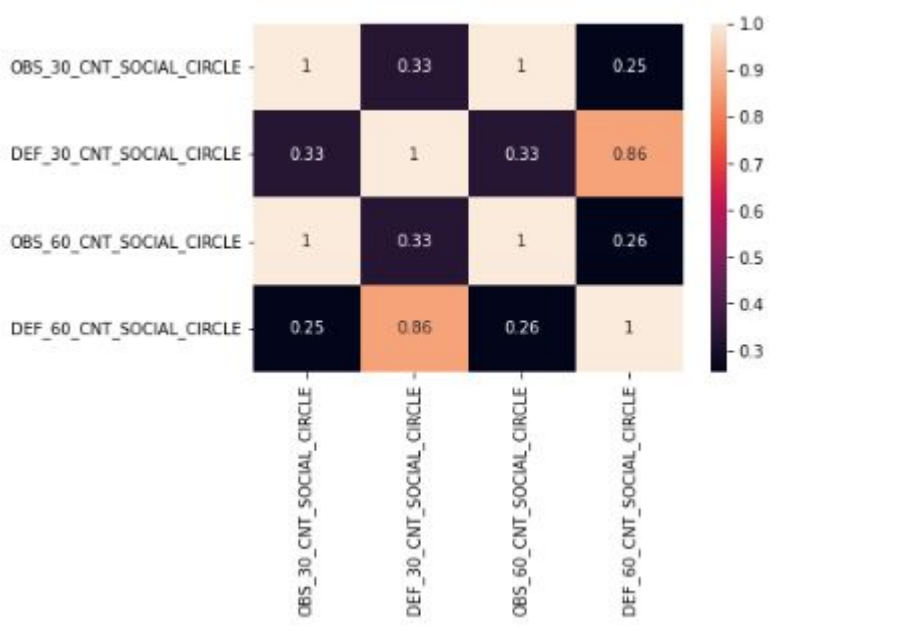


By looking at the above graph we see the Data Imbalance is present, now we will use column TARGET to understand more about Data Imbalance in appl_data.

By looking at this plot we can say most of the applicants have House/Apartments and after that there are people who live with their parents or in municipal apartment.



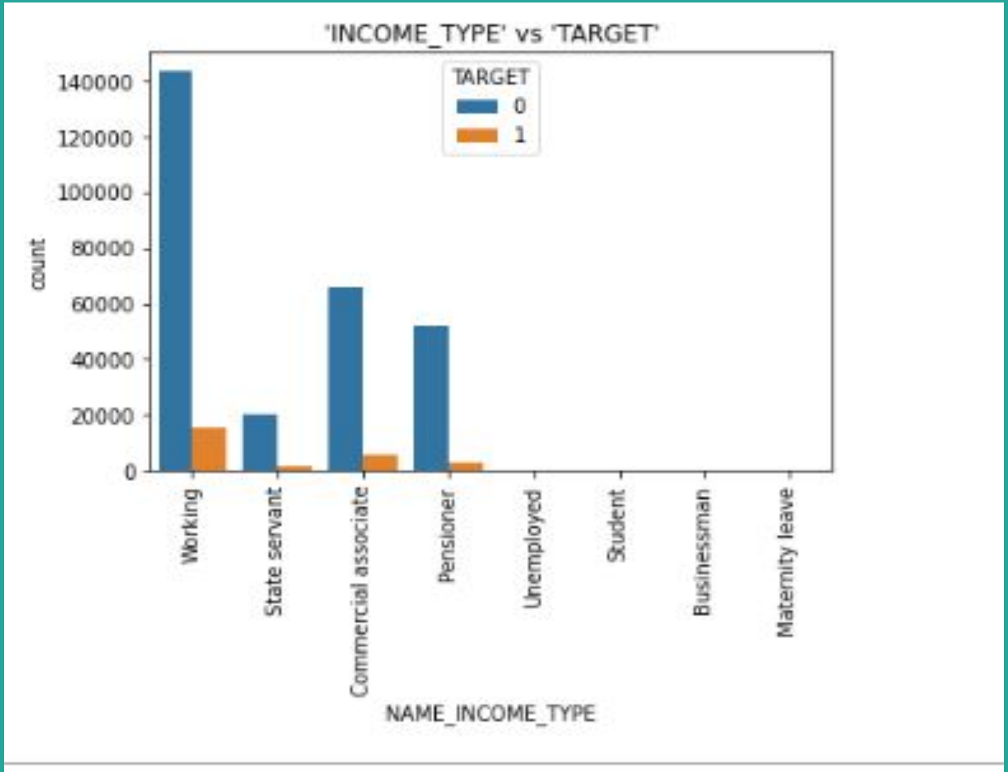
#DEF_30_CNT_SOCIAL_CIRCLE & DEF_60_CNT_SOCIAL_CIRCLE are highly correlated.
#OBS_30_CNT_SOCIAL_CIRCLE & OBS_60_CNT_SOCIAL_CIRCLE are identical.



	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE	1	0.33	1	0.25
DEF_30_CNT_SOCIAL_CIRCLE	0.33	1	0.33	0.86
OBS_60_CNT_SOCIAL_CIRCLE	1	0.33	1	0.26
DEF_60_CNT_SOCIAL_CIRCLE	0.25	0.86	0.26	1

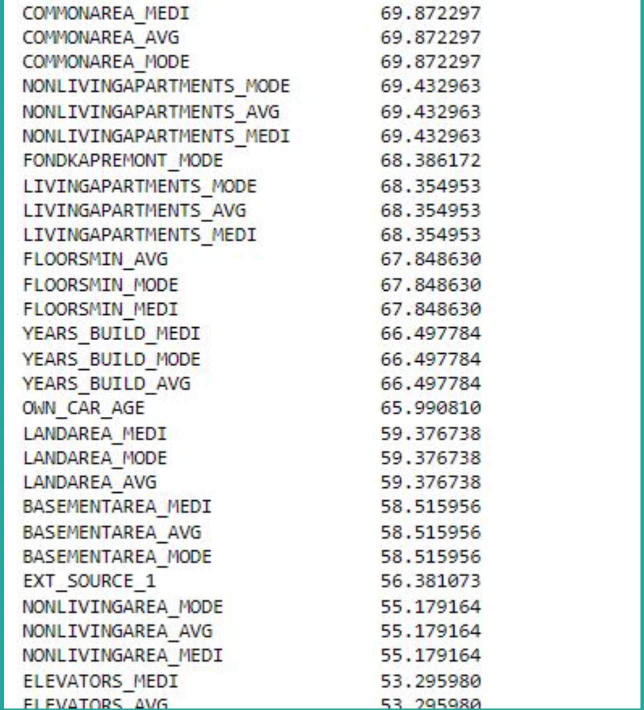
Education & Occupation Info:

Plot below says, there are more working people who applied for loan.

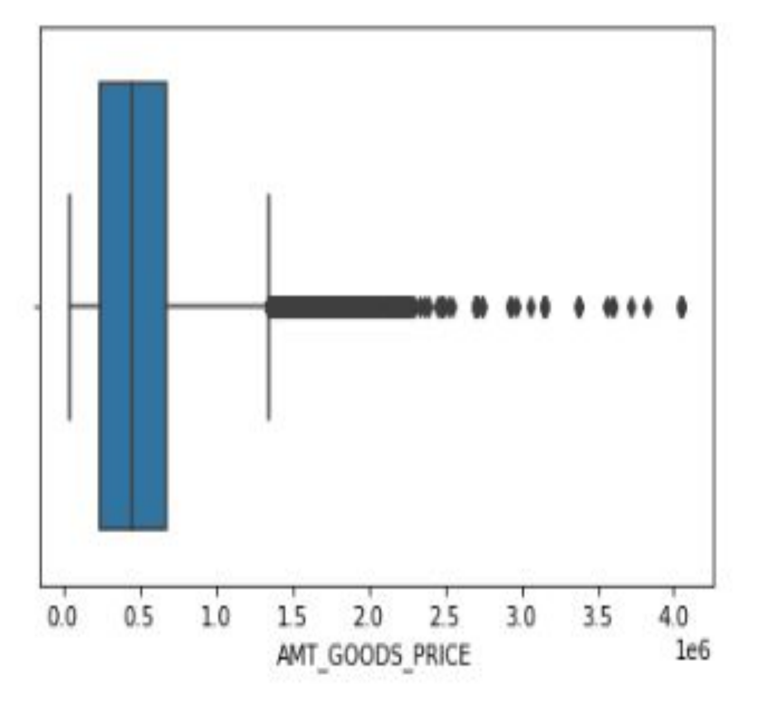


Data Cleaning:

We will check for missing values and their percentage, we will drop columns which have more than 45% missing rows and then in the remaining column, we will check for outliers and we will handle them by imputing mean/median or other necessary values.

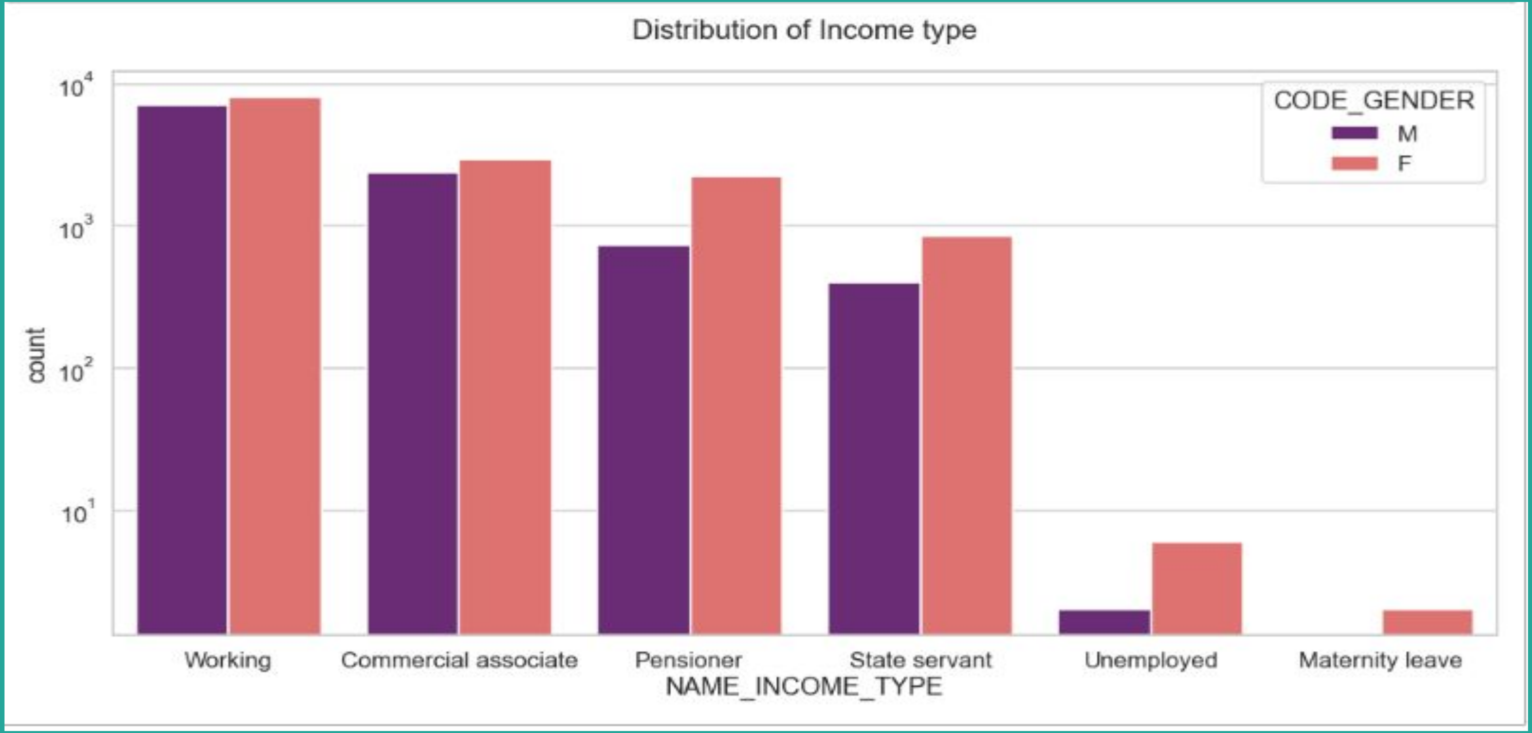


```
COMMONAREA_MEDI                 69.872297
COMMONAREA_AVG                  69.872297
COMMONAREA_MODE                 69.872297
NONLIVINGAPARTMENTS_MODE        69.432963
NONLIVINGAPARTMENTS_AVG         69.432963
NONLIVINGAPARTMENTS_MEDI        69.432963
FONDKAPREMONT_MODE              68.386172
LIVINGAPARTMENTS_MODE           68.354953
LIVINGAPARTMENTS_AVG            68.354953
LIVINGAPARTMENTS_MEDI           68.354953
FLOORSMIN_AVG                   67.848630
FLOORSMIN_MODE                  67.848630
FLOORSMIN_MEDI                  67.848630
YEARS_BUILD_MEDI                66.497784
YEARS_BUILD_MODE                66.497784
YEARS_BUILD_AVG                 66.497784
OWN_CAR_AGE                     65.990810
LANDAREA_MEDI                   59.376738
LANDAREA_MODE                   59.376738
LANDAREA_AVG                    59.376738
BASEMENTAREA_MEDI               58.515956
BASEMENTAREA_AVG                58.515956
BASEMENTAREA_MODE               58.515956
EXT_SOURCE_1                    56.381073
NONLIVINGAREA_MODE              55.179164
NONLIVINGAREA_AVG               55.179164
NONLIVINGAREA_MEDI              55.179164
ELEVATORS_MEDI                  53.295980
ELEVATORS_AVG                   53.295980
```



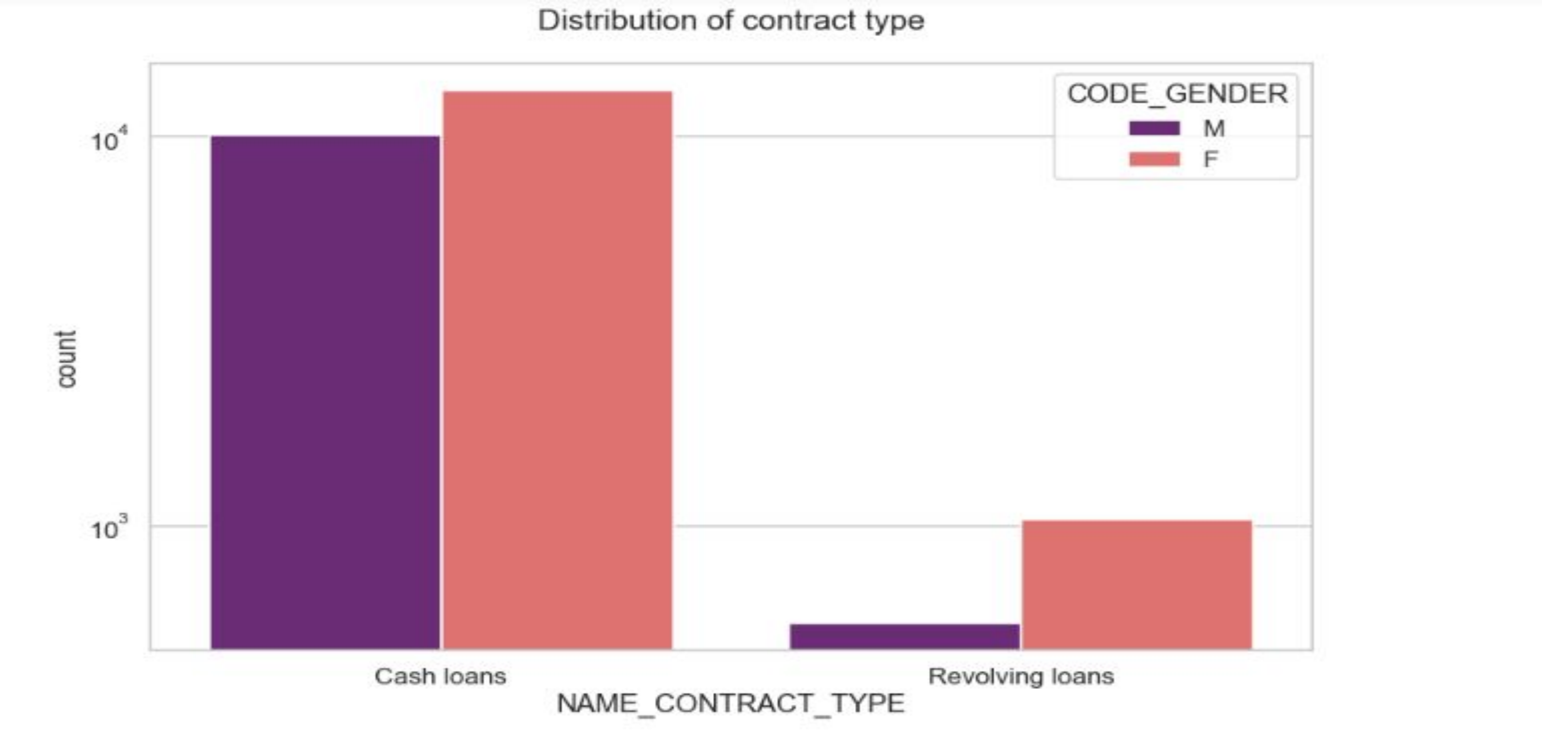
Conclusion from Variate Analysis for Categories:

For income type 'Working', 'Commercial Associate', and 'State Servant' the number of credits are higher than others, for this Females are having more number of credits than male.
Less number of credits for income type 'student' ,'pensioner', 'Businessman' and 'Maternity leave'.

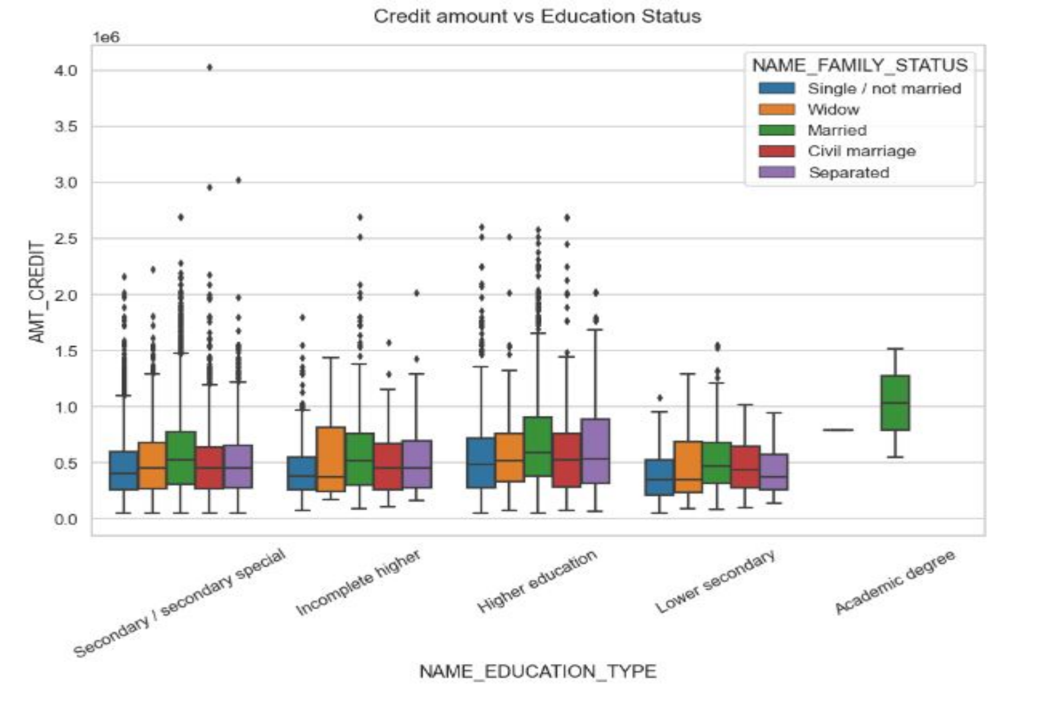


Conclusion 2:

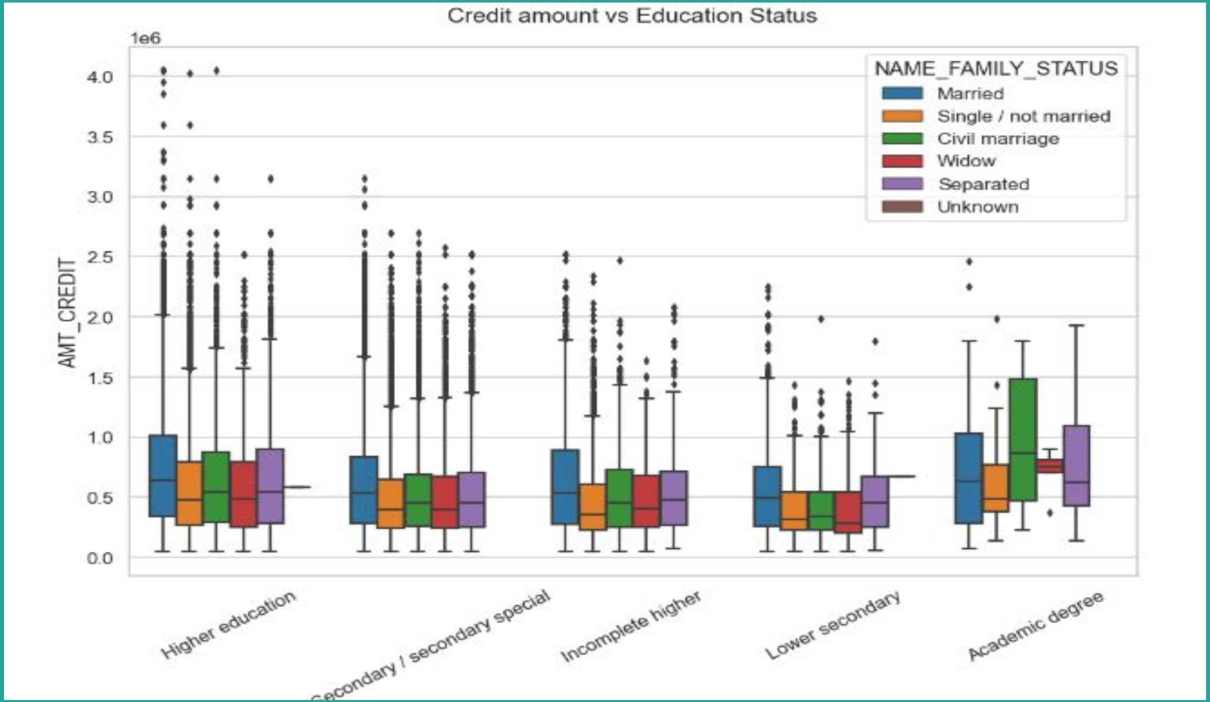
For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type. For this also Female is leading for applying credits.



Bivariate analysis for numerical variables:



For Defaulters, From the above box plot we can conclude that Family status of 'civil marriage', 'married' and 'separated' of Academic degree education are having higher number of credits than others. Also, higher education of family status of 'married', 'single' and 'civil marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.



For Non-Defaulters quite similar with defaulters From the above box plot we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Most of the outliers are from Education type 'Higher education' and 'Secondary'. Civil marriage for Academic degree is having most of the credits in the third quartile.

Thank You!