Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand 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 to 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:

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.

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):

Approved: The Company has approved loan Application

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.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but at different stages of the process.

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

Data Understanding

This 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.

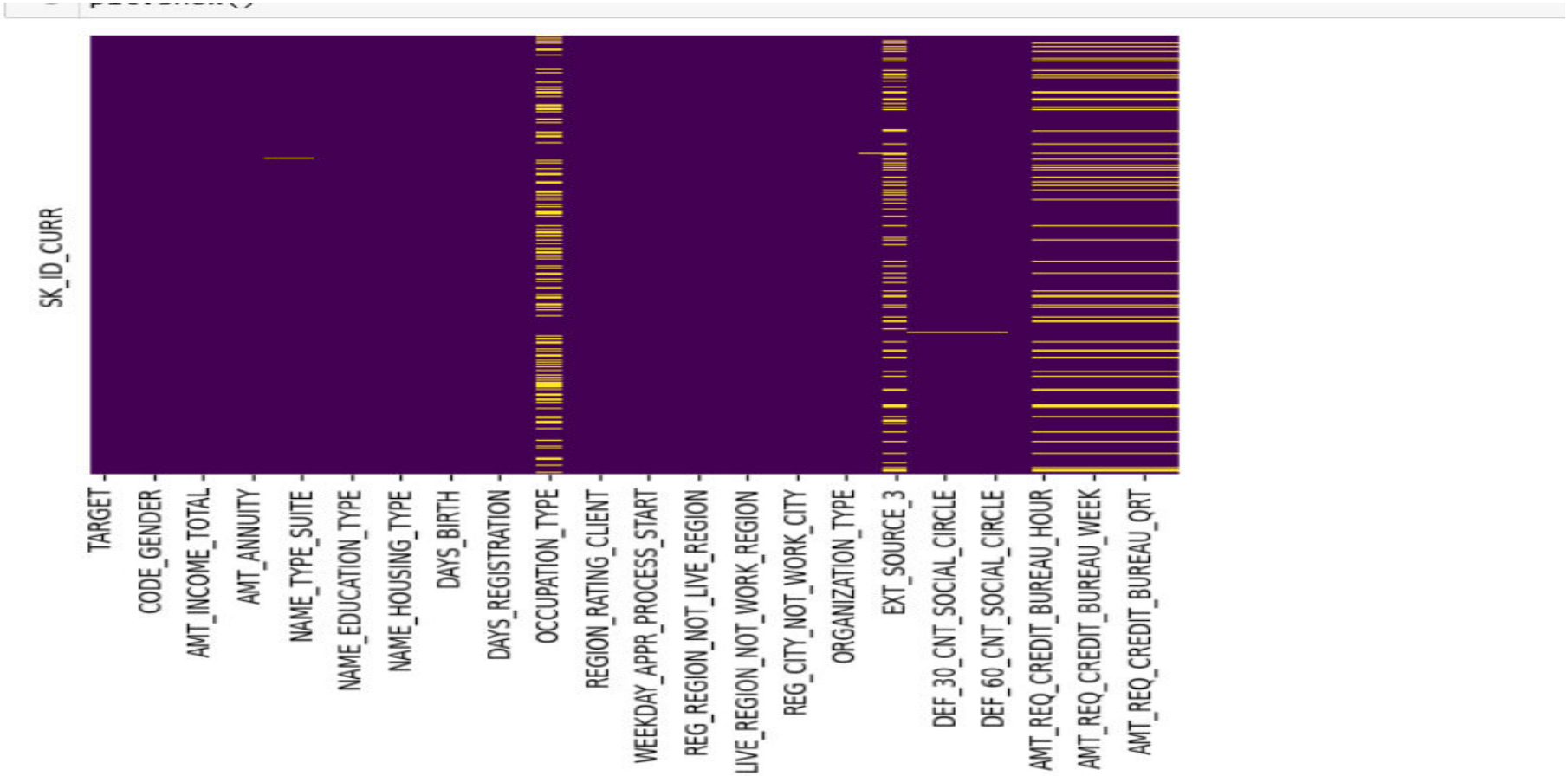
2. 'previous_application.csv' contains information about the client’s previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

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

Cleaning and imputing the dataset

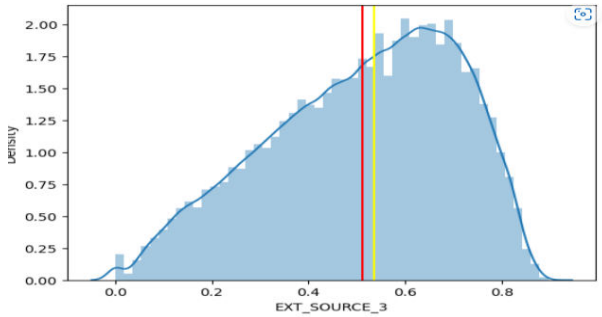
Unwanted data is deleted which is not required in further calculation or assumptions (most of names start with prefix "FLAG") in the application dataset. Data related to clients documents, whether client own a car or not.

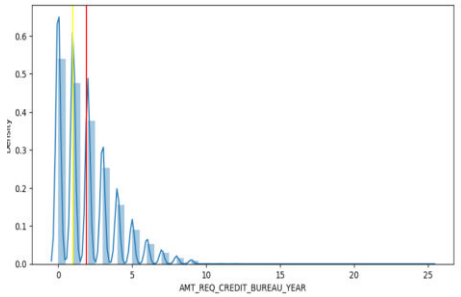
2. Checking the null value through heatmap. yellow lines show the missing values :-



Imputing the value based on mean, median and skew

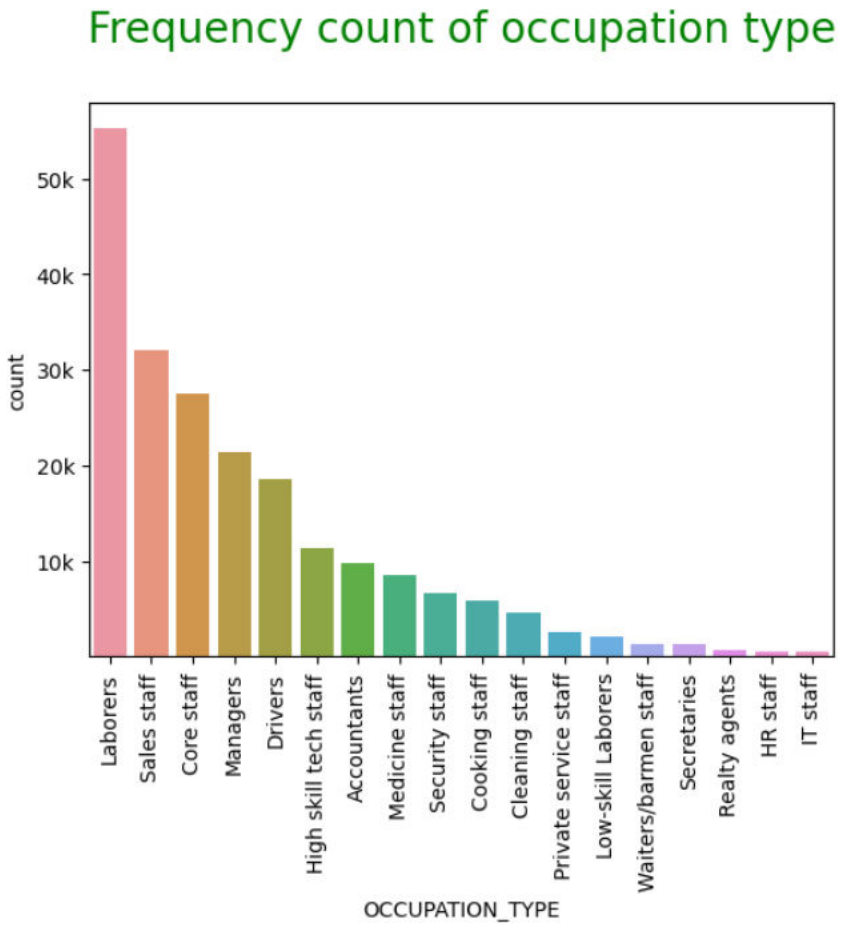
1. EXT_SOURCE_3 column is negatively Skewed so we will use median to fill the NaN value
2. AMT_REQ_CREDIT_BUREAU_YEAR column is positively Skewed so we will use median to fill the NaN value. Hence replacing the the NaN with median values.
3.AMT_REQ_CREDIT_BUREAU_QRT column is postive skew so we will use median to fill the NaN value. Hence replacing the the NaN with median values.
4.AMT_REQ_CREDIT_BUREAU_MON column is postive skew so we will use median to fill the NaN value. Hence replacing the the NaN with median values.
4.AMT_REQ_CREDIT_BUREAU_WEEK column is postive skew so we will use median to fill the NaN value. Hence replacing the the NaN with median values
5.AMT_REQ_CREDIT_BUREAU_DAY column is postive skew so we will use median to fill the NaN value. Hence replacing the the NaN with median values. The mean is greater than the median in positively distributed data, and most people fall on the lower side.





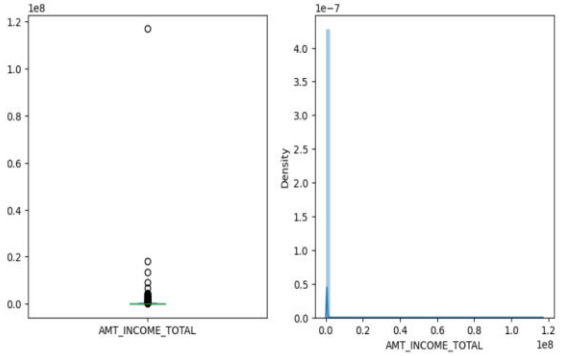
Univariate Analysis

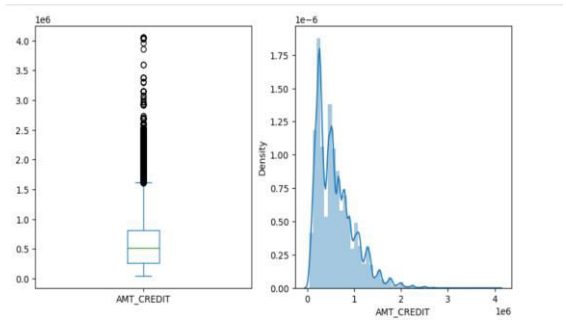
Laborer is highest as compared to other occupation type and Sales staff is the second highest on application loan

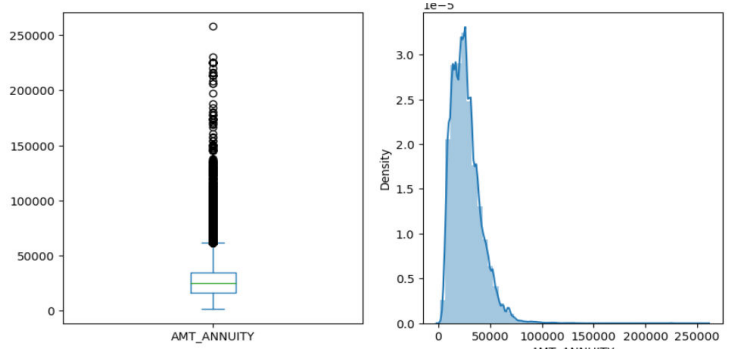


Outlier Analysis

1 . 'AMT_INCOME_TOTAL' lower to upper fence value range : [-22500.0, 337500.0]
2. 'AMT_ANNUITY' lower to upper fence value range : [-10584.0, 61704.0]
3. As observed from distplot and boxplot, the outliers tend to exist after 1616625. Applicants with AMT_CREDIT above 1616625 are outliers.

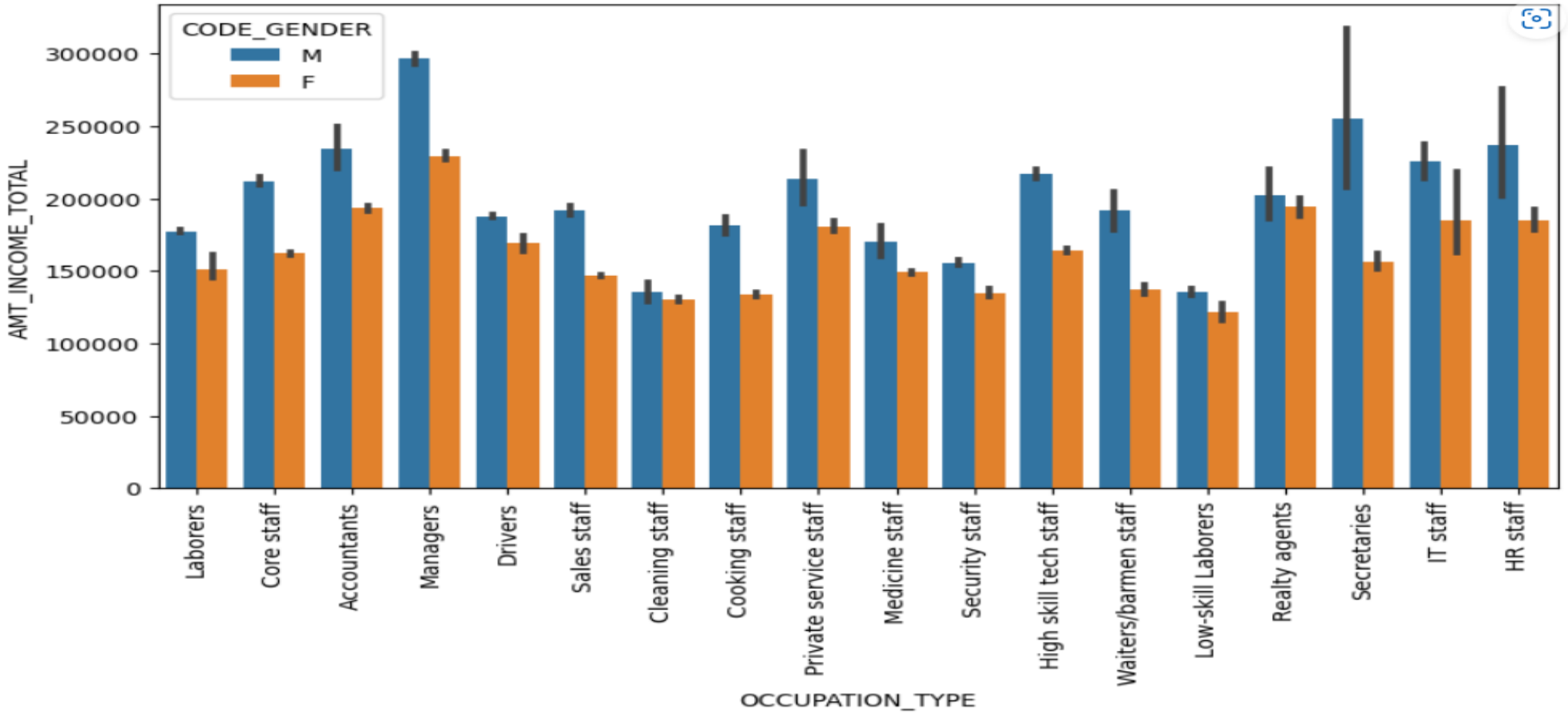






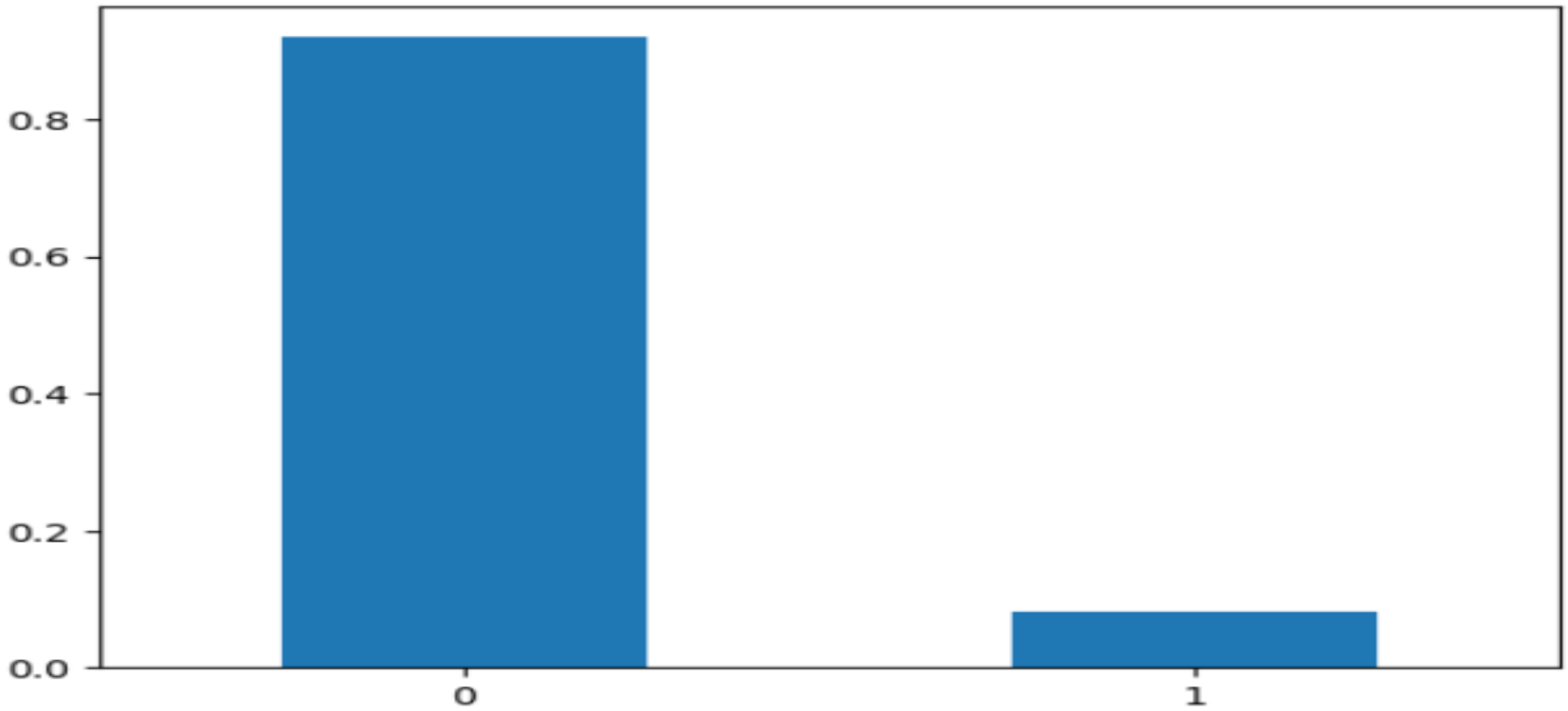
Bivariate Analysis

In Gender, Male have the highest income total compare to the female in all the occupation type. Male and Female managers have the highest income compare to other occupation type. Male with occupation as Secretaries and female with occupation as Realty agents are the second highest respectively.



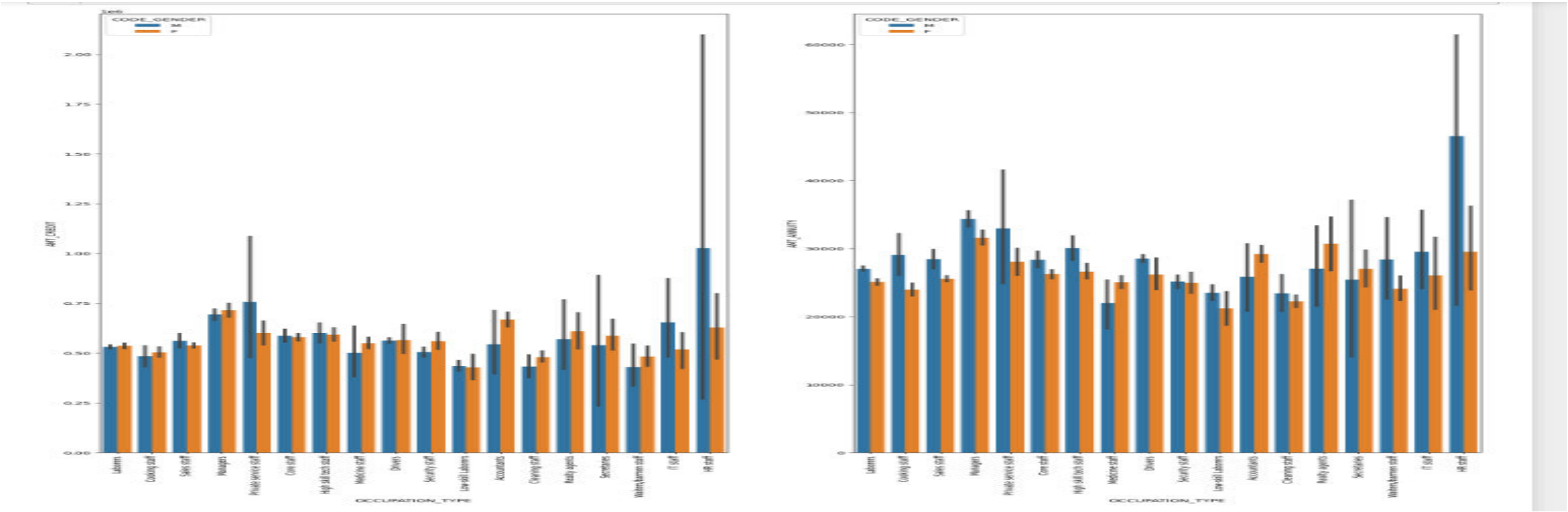
Checking the imbalance of the Target variable

1. Target value 1 represents client with payment difficulties (he/she had late payment more than X days on at least one of the first Y installments of the loan). This is only 8.07% of the data
2. Target value 0 represents all other cases than 1. This is 91.93% of the data



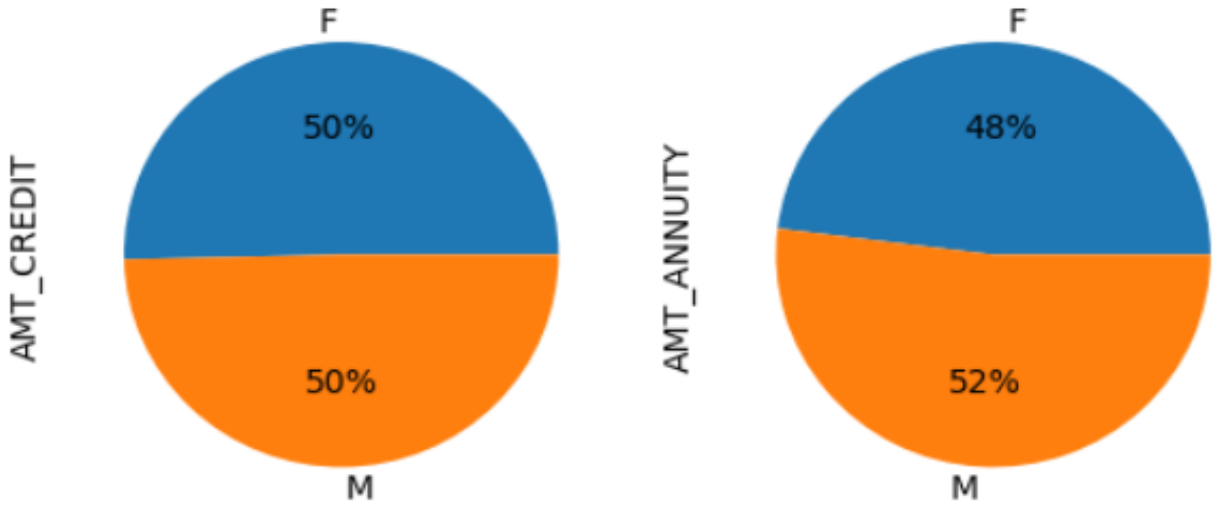
Based on ORGANIZATION_TYPE

1. Male and female with organization type 'Realtor' have the highest annuity and credit compared to all the organization type.
2. Male with organization type 'Cleaning' have the lowest credits compare to all the male in different organization type and have high annuity amount
3. Female with organization type 'Postal' havs lowest credits compare to all the female in different organization type and have high annuity amount.



Based on ORGANIZATION_TYPE calculation

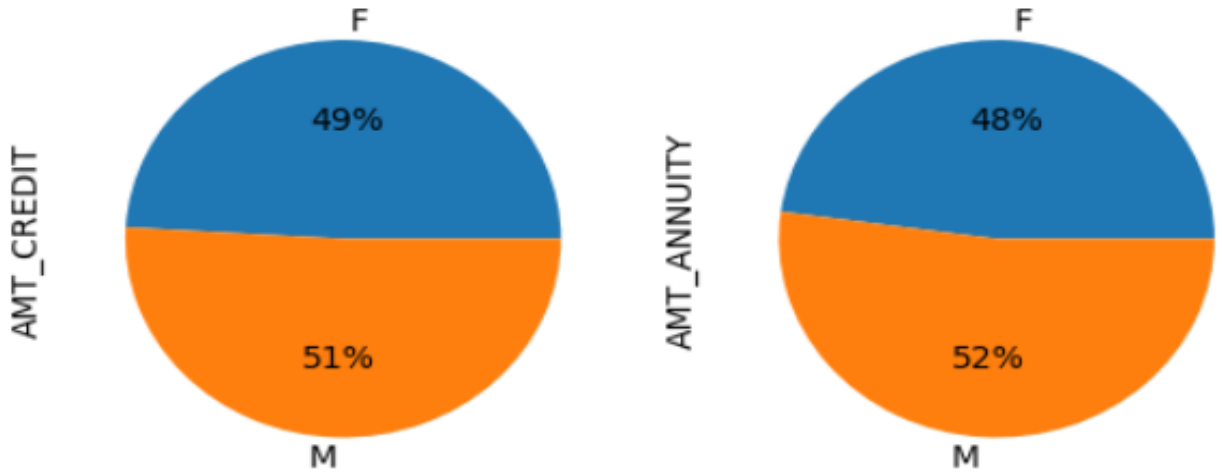
```
Average value of loan taken by male and female :
 CODE_GENDER
F    561140.131757
M    553307.957062
Name: AMT_CREDIT, dtype: float64

----------------------------------------------------------------

Average value of installments clients with payments difficulty categories by male and female :  CODE_GENDER
F    25584.223147
M    27675.350540
Name: AMT_ANNUITY, dtype: float64
```



Male (had late payment more than X days on at least one of the first Y installments of the loan) have 50% of the credit amount and same with the female. But 48% of female and 52% of men had late payment respectively.

Based on ORGANIZATION_TYPE calculation

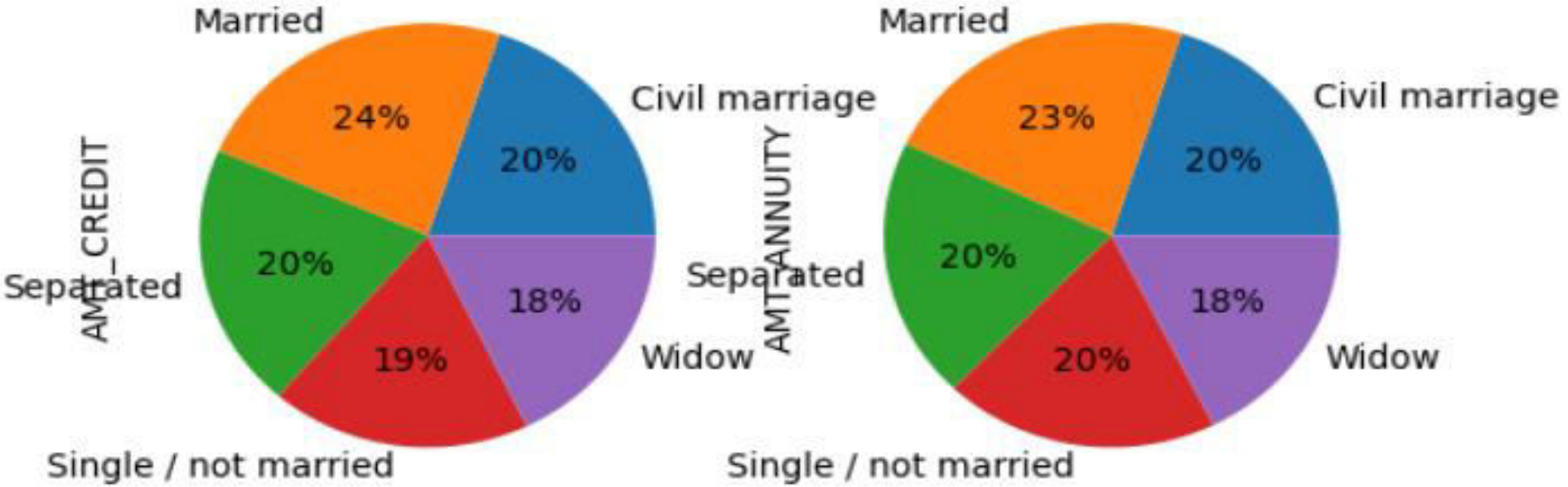
```
Average value of loan taken by male and female :
 CODE_GENDER
F    595142.808224
M    617617.410576
Name: AMT_CREDIT, dtype: float64

----------------------------------------------------------------

Average value of installments clients without any payments issues segregated male and female :
 CODE_GENDER
F    26358.852437
M    28768.393797
Name: AMT_ANNUITY, dtype: float64
```



51% of the male (Non-defaulter) have credit amount and female have 49%. But 48% of female and 52% of men had paid without any problem respectively.

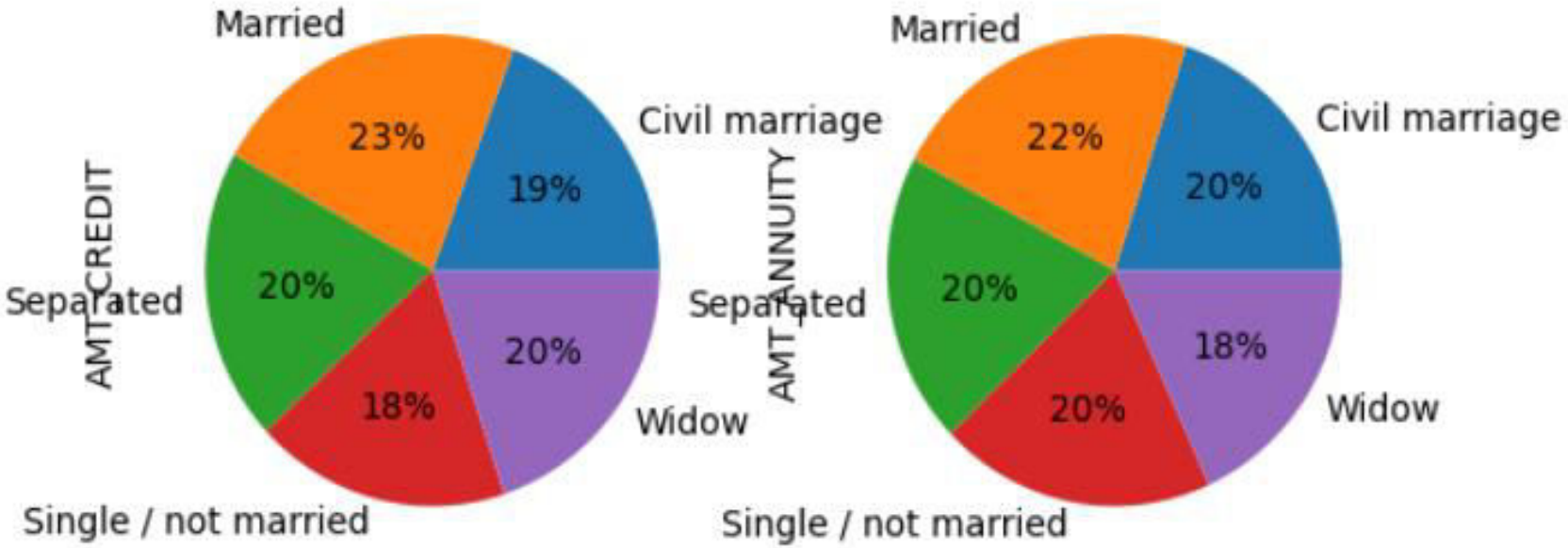
NAME_FAMILY_STATUS

People having difficulties with payment :-
1. Married people have 24% credit amount and 23% of amount annuity
2. Separated people have 20% credit amount and 20% of amount annuity
3. Civil marriage have 20% credit amount and 20% of amount annuity
4. Window have 18% credit amount and 18% of amount annuity
5. single/not married have 19% credit amount and 20% of amount annuity



NAME_FAMILY_STATUS

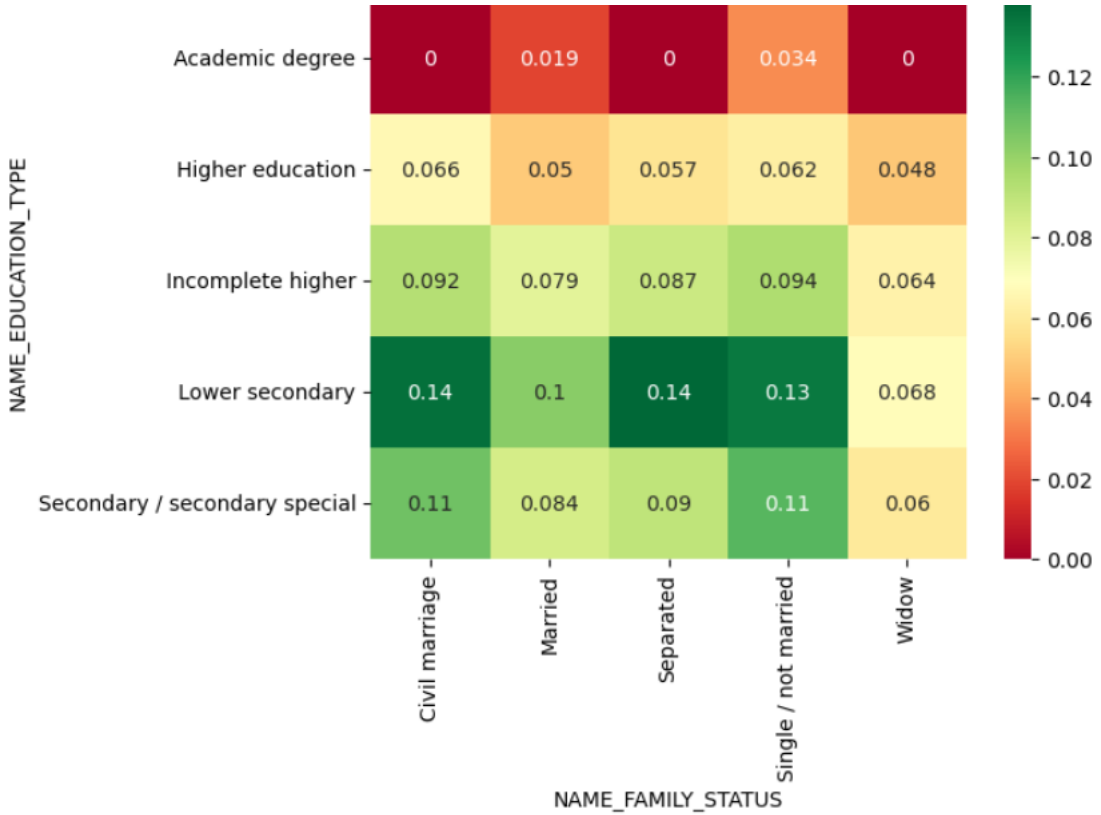
People with no issues in payment
1. Married people have 23% credit amount and 22% of amount annuity
2. Separated people have 20% credit amount and 20% of amount annuity
3. Civil marriage have 19% credit amount and 20% of amount annuity
4. Window have 20% credit amount and 18% of amount annuity
5. single/not married have 18% credit amount and 20% of amount annuity



Multivariate Analysis of application data

Correlation value should be range between [-1 to +1] : -

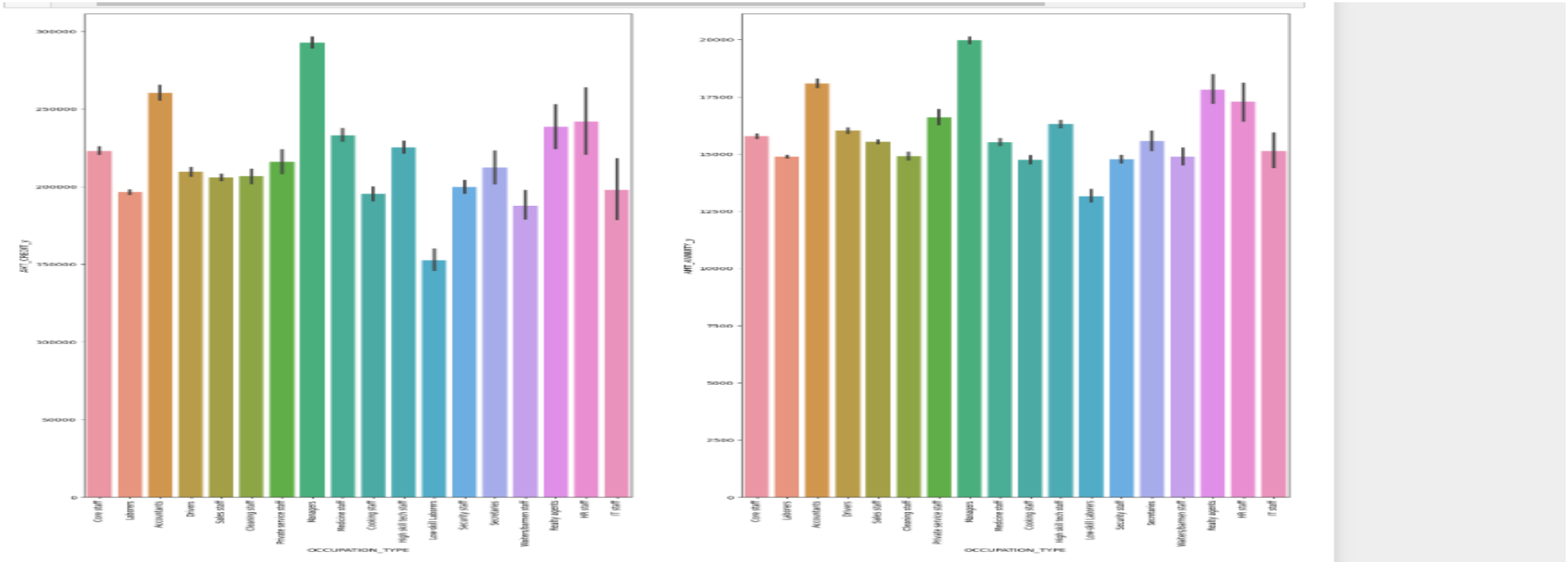
1. Lower secondary have high correlation with civil marriage and with separated
2. Academic degree hav zero correlation with civil marraige, separated and window
3. Incomplete higher have 0.094 correlation with single/not married
4. Secondary/secondary special have 0.11 correlation with cicil marriage and single/not married



Repeater client without payment difficulties

Clients without payment difficulties with repeaters:-
1. Manager credit amount is higher compare to all the occupation type in the account credit and in amount annuity of previous loan data.
2. Accounts have second highest amount credit and amount annuity respectively.

Clients with payment difficulties with repeaters:-
3. Low-skilled laborers have lowest amount credit and lowest amount annuity.
4. Waiter/barmen staff have second lowest amount credit and lowest amount annuity.



Conclusion

1. We should more focus on Manager and Realtor and HR-staff ocuupation type
2. we should more target to married people more
3. we should target male applicant more compare to female