CREDIT EDA- Case Study

By-Kavya Makhija

Introduction

This case study 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 minimize the risk of losing money while lending to customers.

Business Understanding-I

- 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 specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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-II

- 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.

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

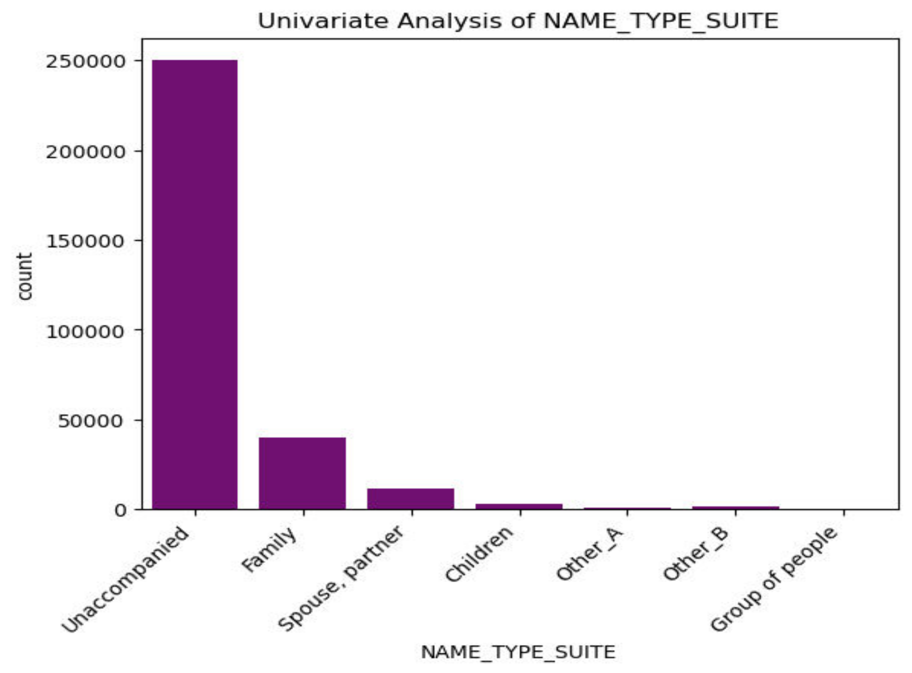
This dataset has 3 files as explained below:

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.**

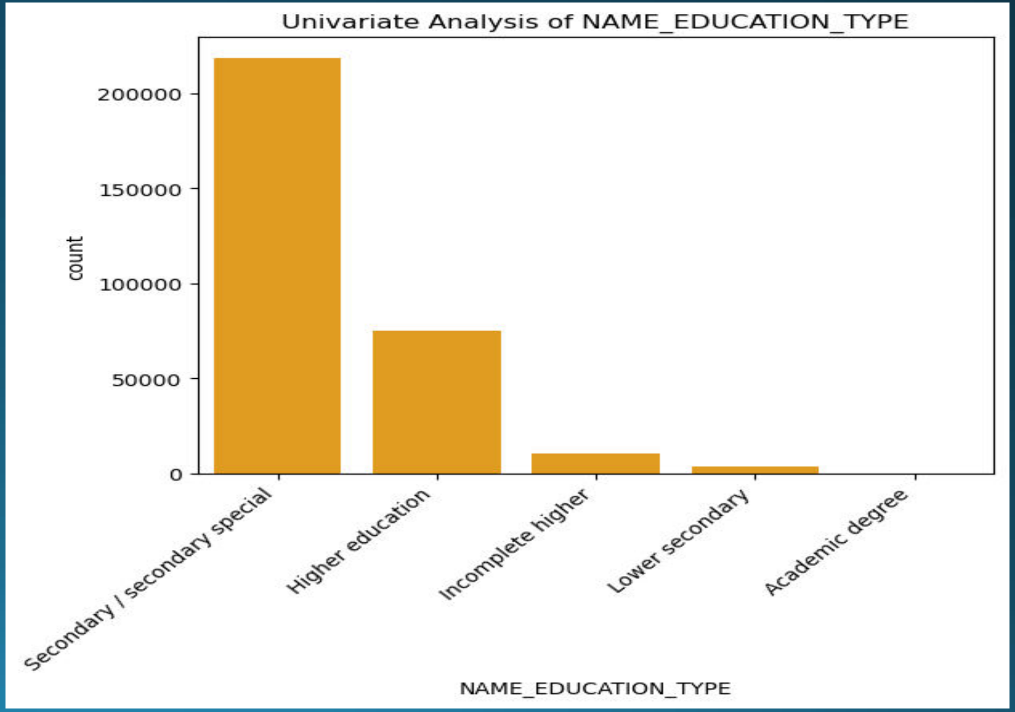
2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

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

UNIVARIATE ANALYSIS

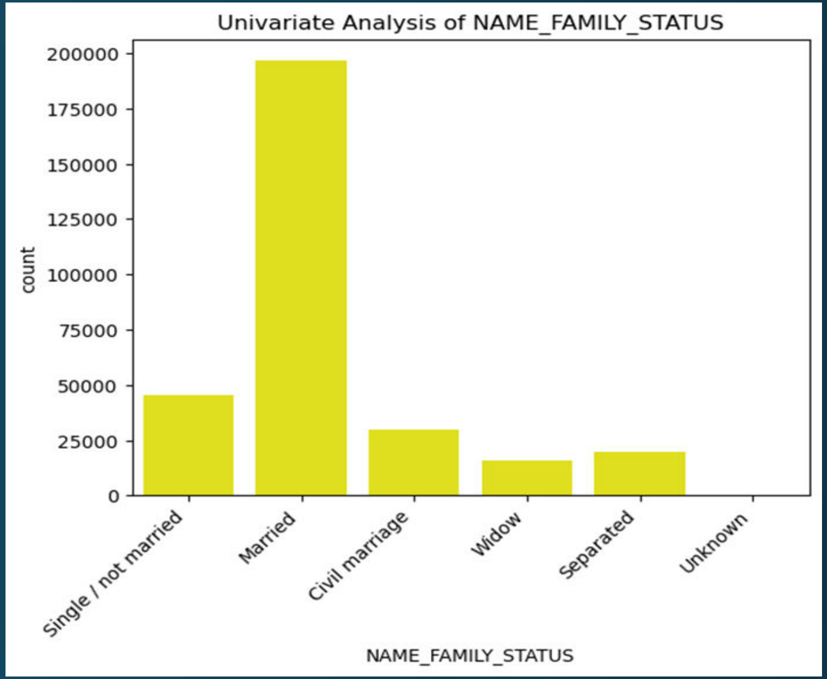


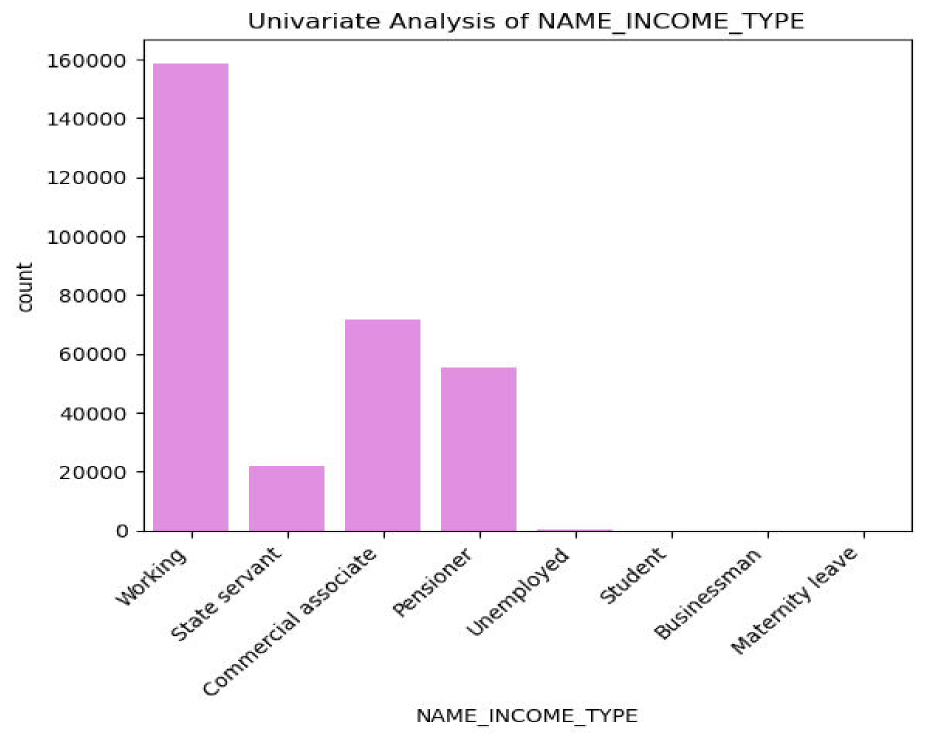
From the given chart on 'NAME_TYPE_SUITE' we can see most of the applicants are Unaccompanied.

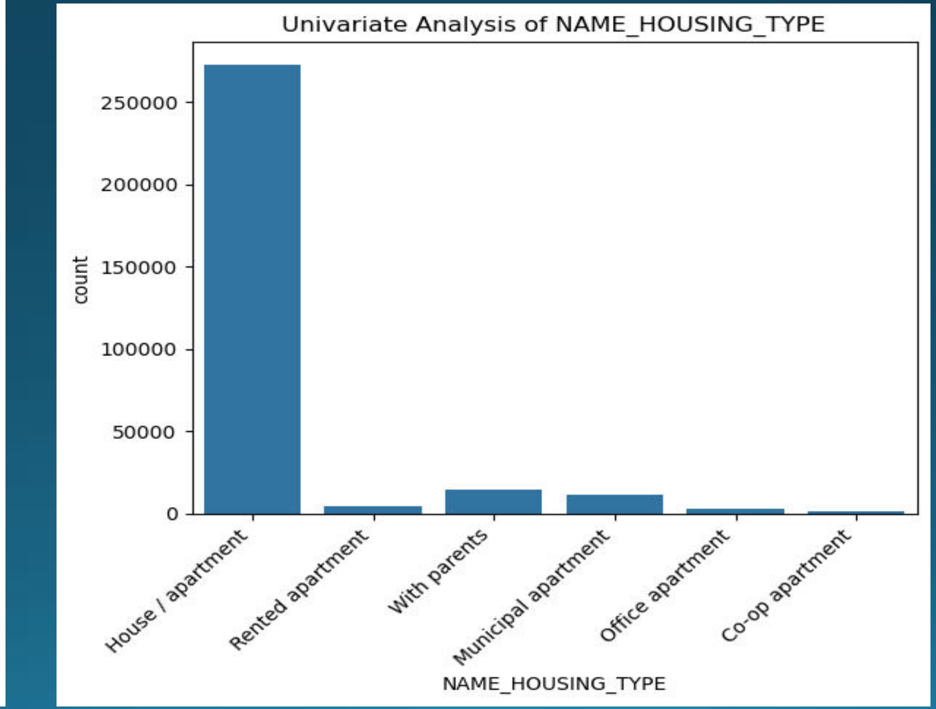


From the given chart on 'NAME_EDUCATION_TYPE' we can see most of the applicants have Secondary/secondary special education.

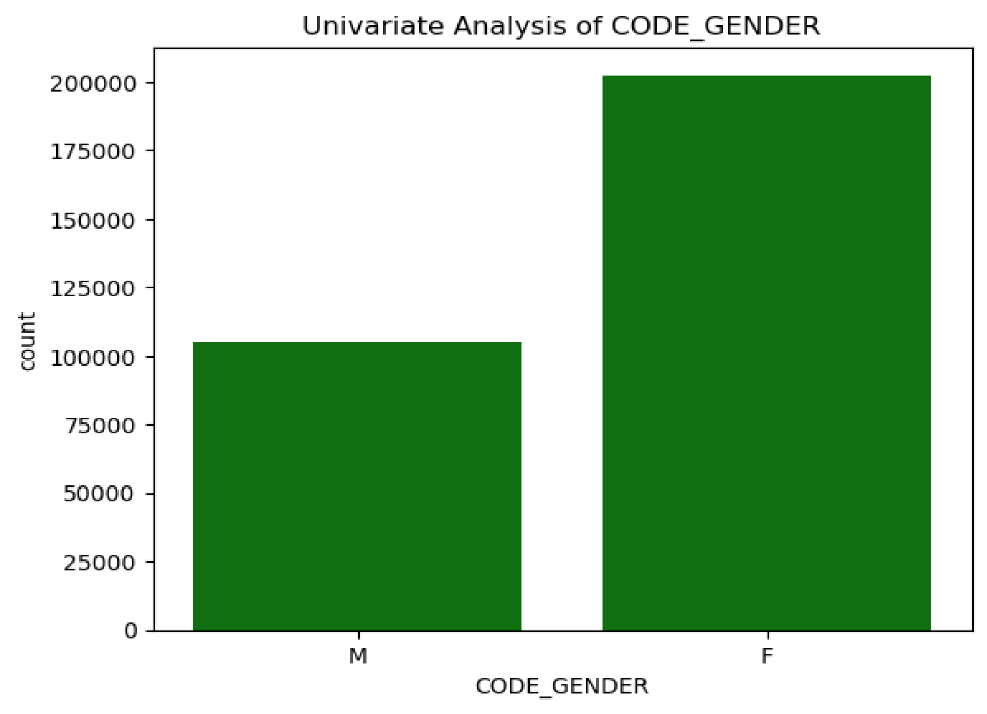
- From the charts we can depict that most of the applicants are ‘Working’
- Most of the applicants are Married.
- Moreover maximum number of applicants own their own house/apartment.

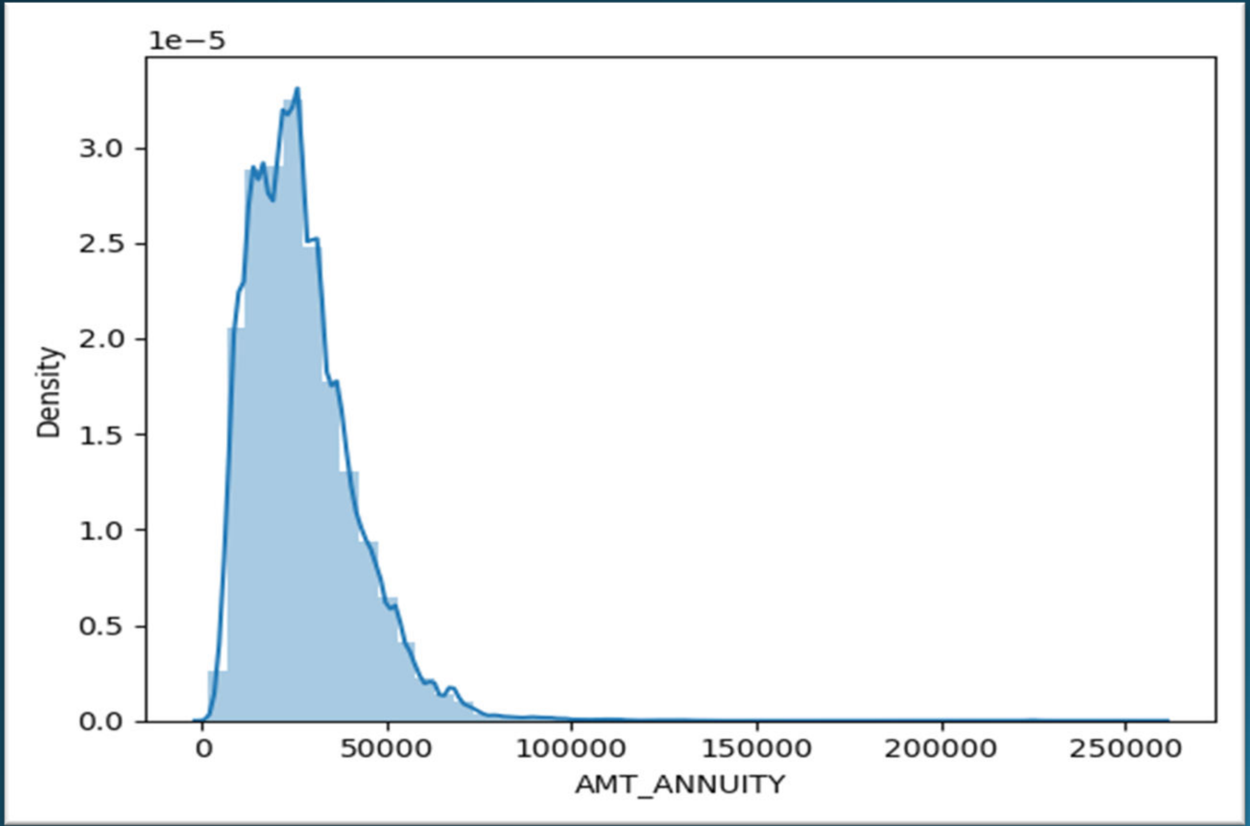


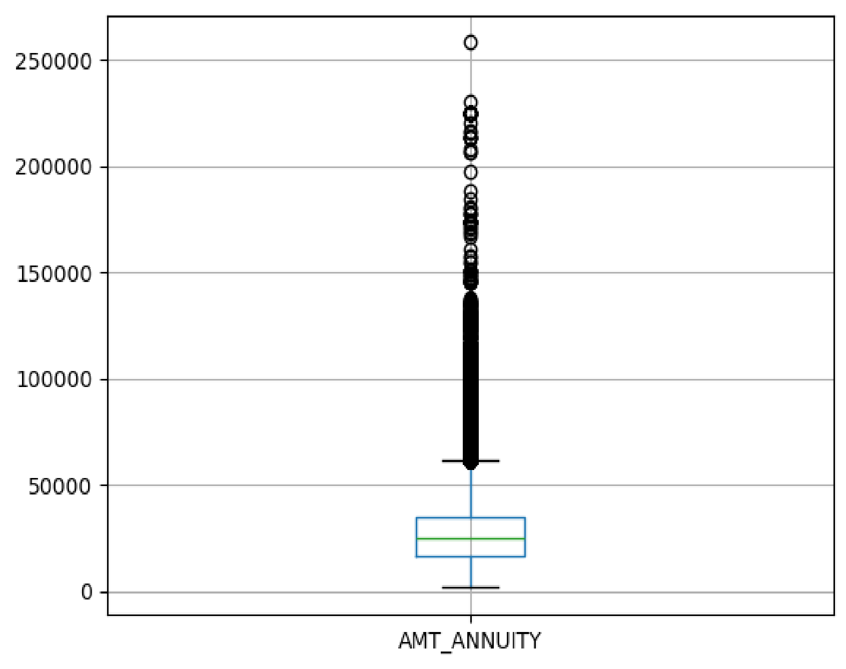




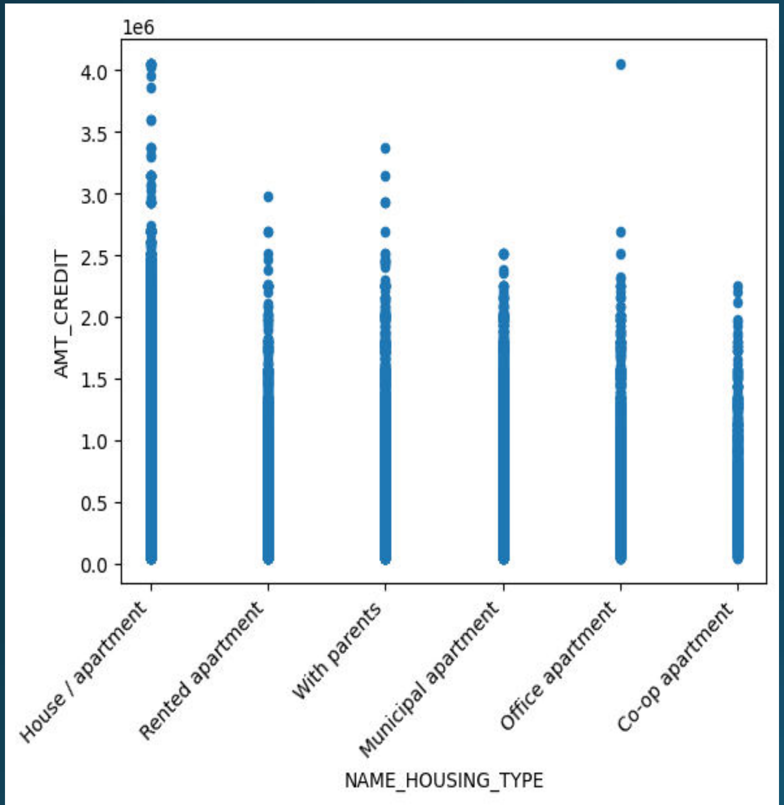
- From the chart we can depict that maximum number of applicants are females.
- The graph on 'AMT_ANNUITY' depicts that maximum number of loan annuity is between 20000-40000.



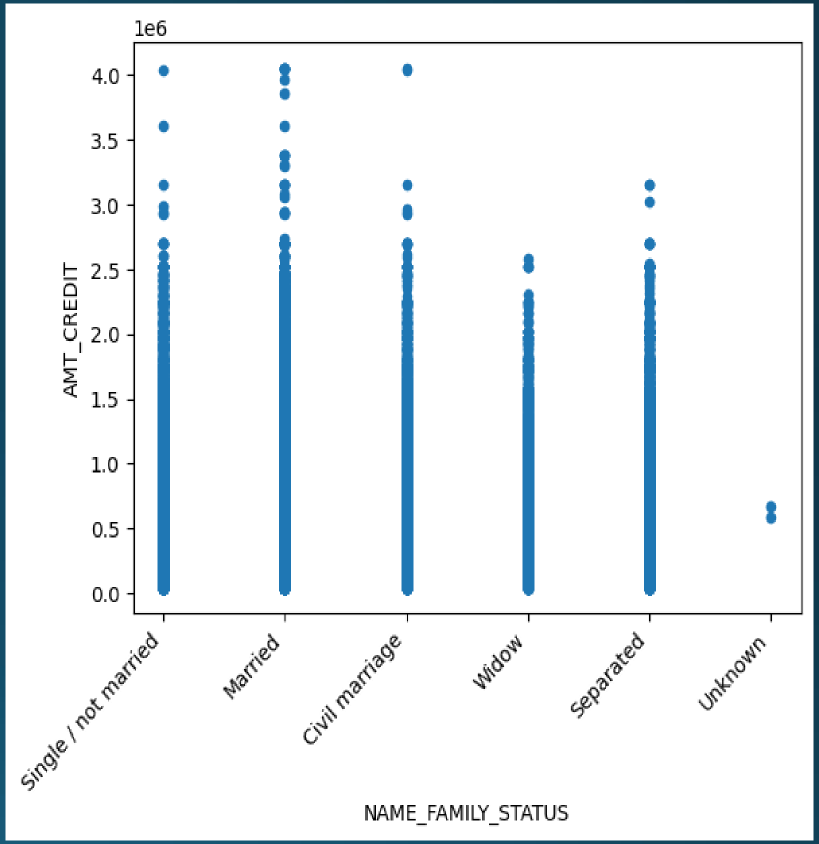




BIVARIATE ANALYSIS

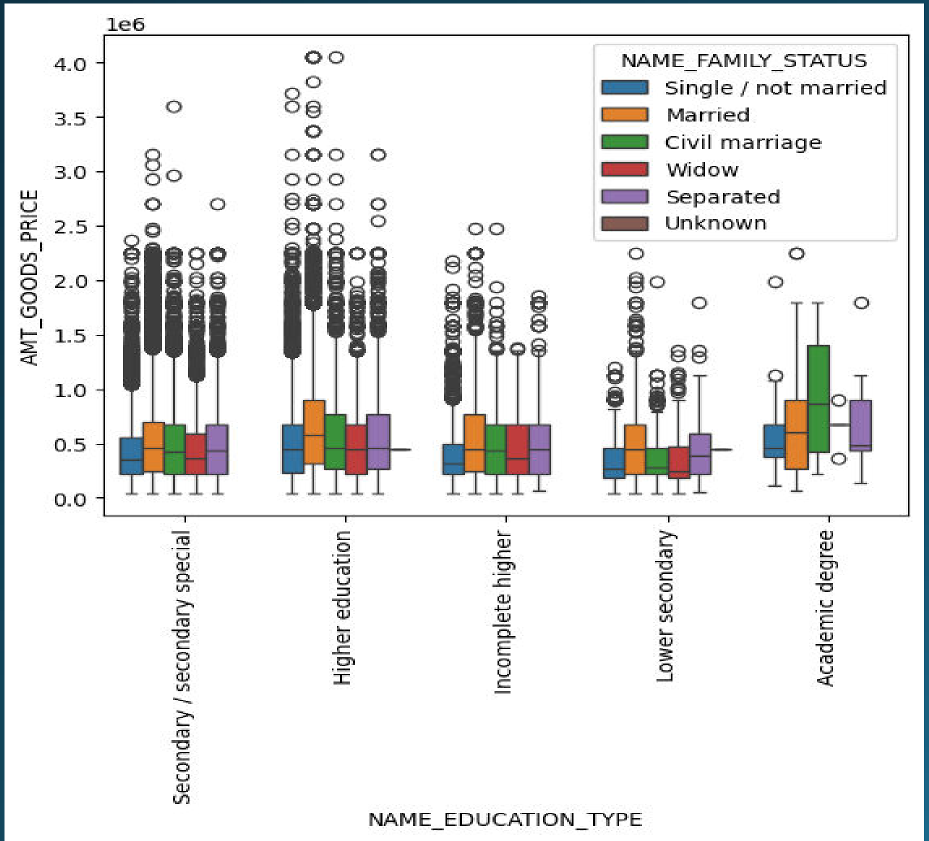


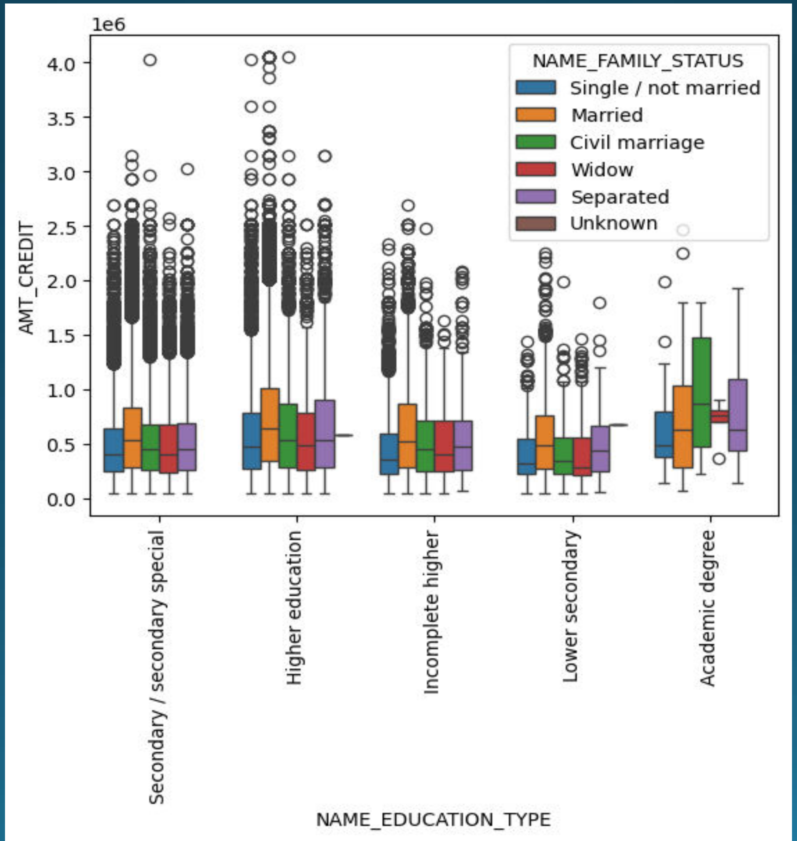
Applicants with high credit score own House/apartment.

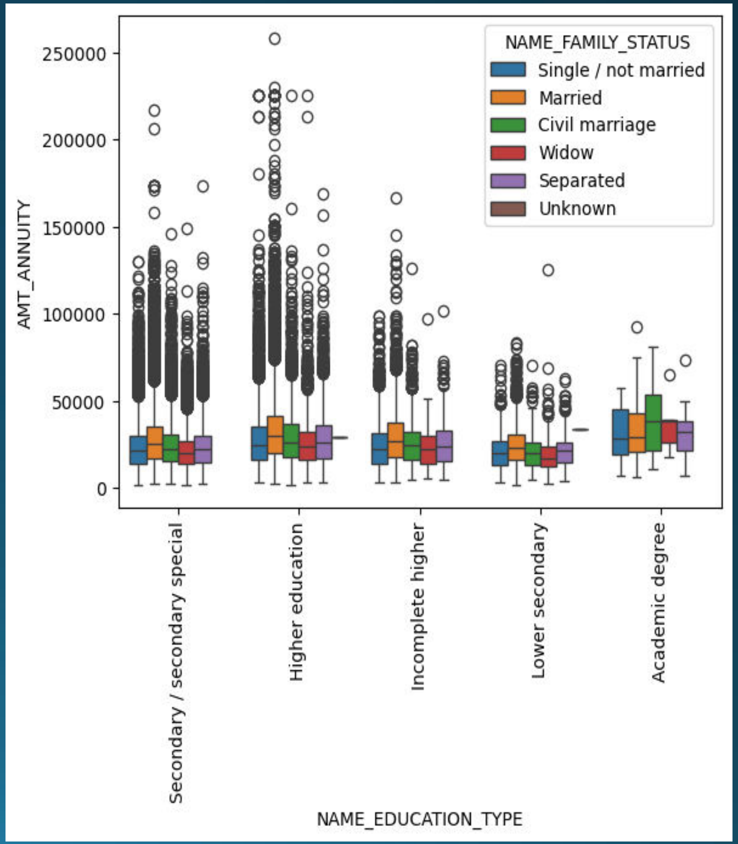


Applicants with high credit scores are single.

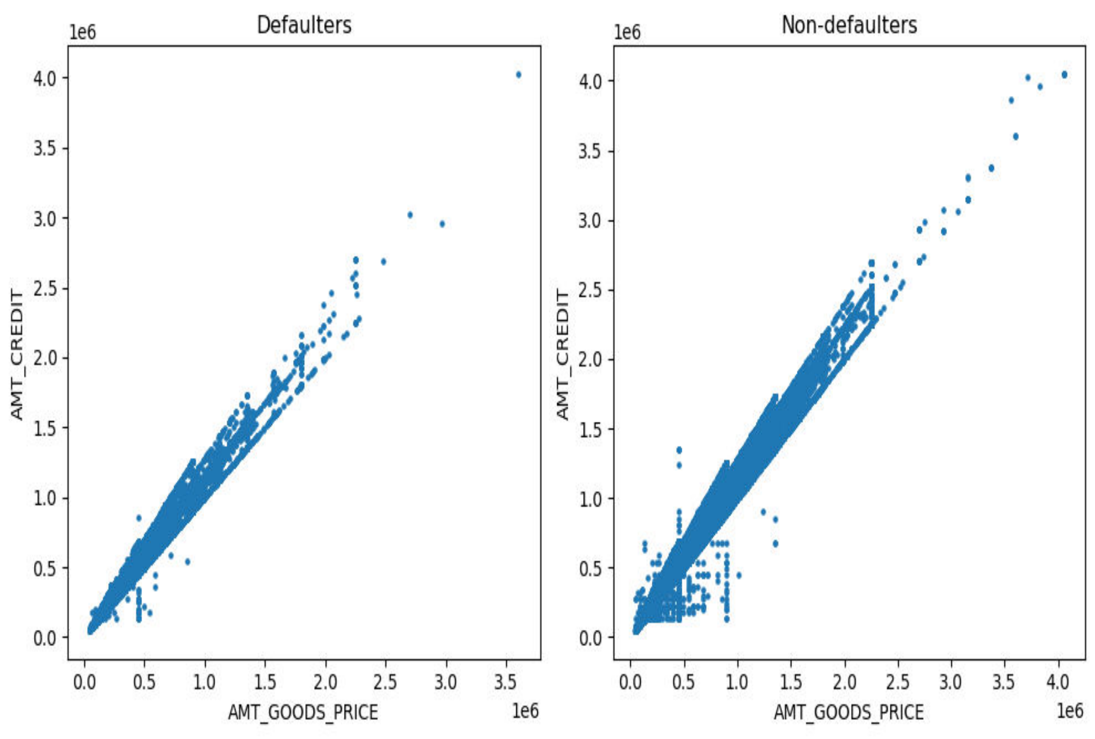
- From the graph of ‘AMT_CREDIT’ and ‘NAME_EDUCATION_TYPE’ we can see that Family status of Civil Marriage, Marriage and Separated of Academic degree education have higher credit than others.
- From the graph of ‘AMT_GOODS_PRICE’ and ‘NAME_EDUCATION_TYPE’ we can see that Family status of Civil Marriage,Married,Single and Separated of Academic degree education have higher credit than others.
- From the graph of ‘AMT_ANNUITY’ and ‘NAME_EDUCATION_TYPE’ we can see that Family status of Civil Marriage, Widow and Separated of Academic degree education have higher credit than others. A high amount of outliers are present especially in Higher Education. Moreover, ‘Widows' have lower annuity in education categories

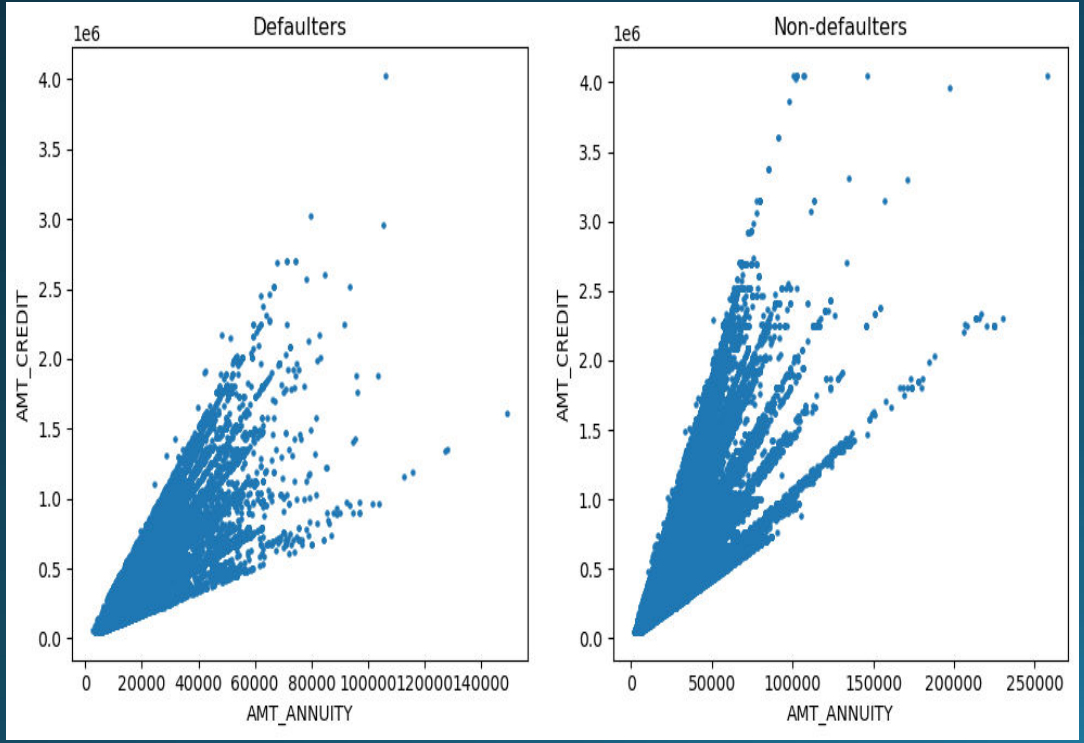






'AMT_GOODS_PRICE' and 'AMT_CREDIT' have strong positive correlation. This means as Goods price increases, so does Credit Amount.

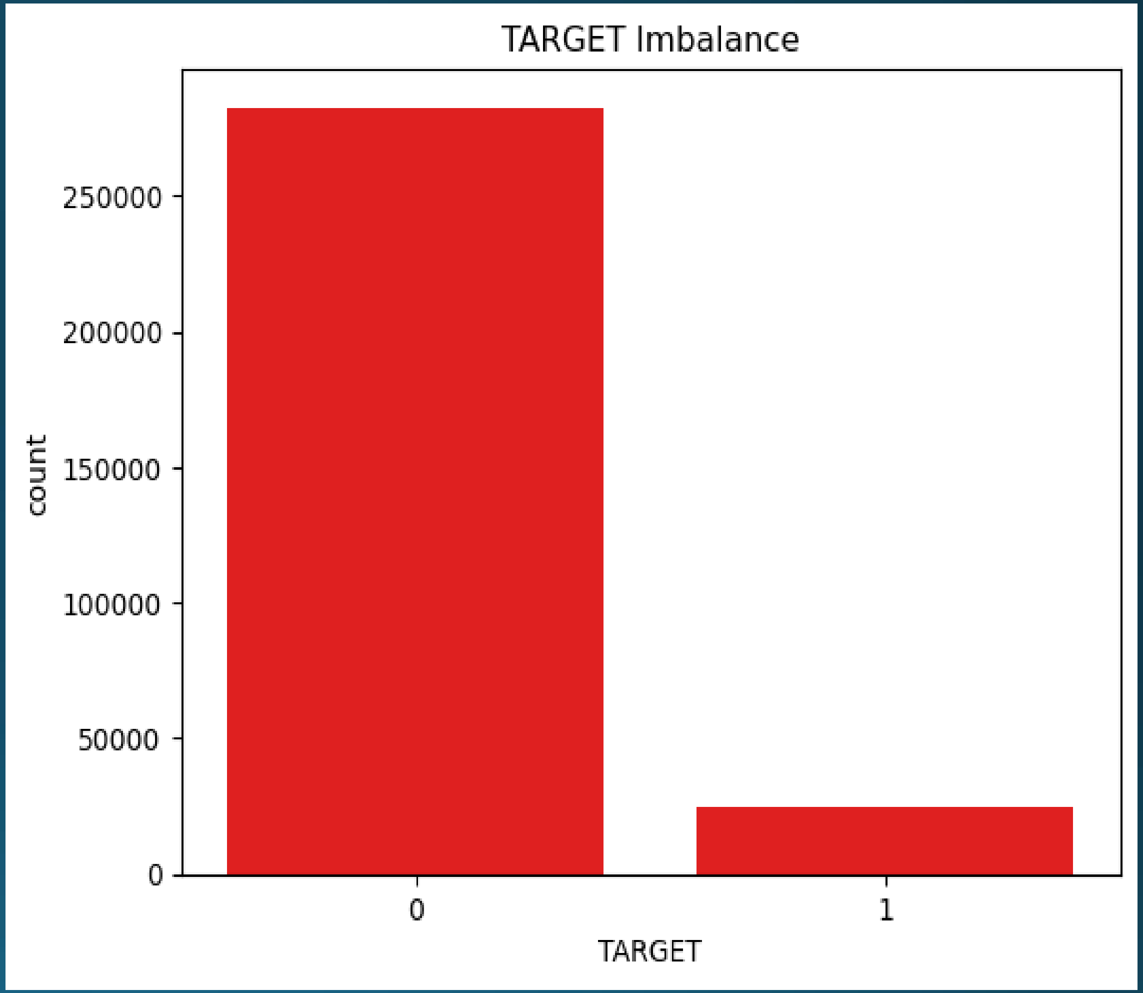




'AMT_ANNUITY' and 'AMT_CREDIT' have strong positive correlation. This means as Annuity Amount increases, so does Credit Amount.

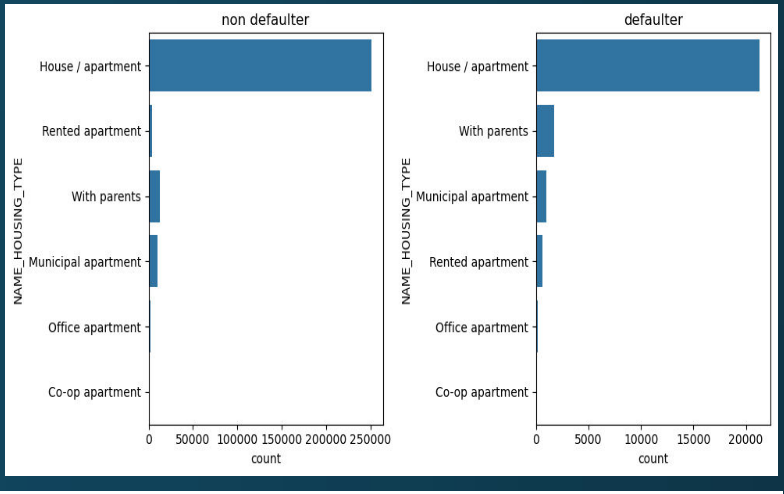
SEGMENTED- ANALYSIS

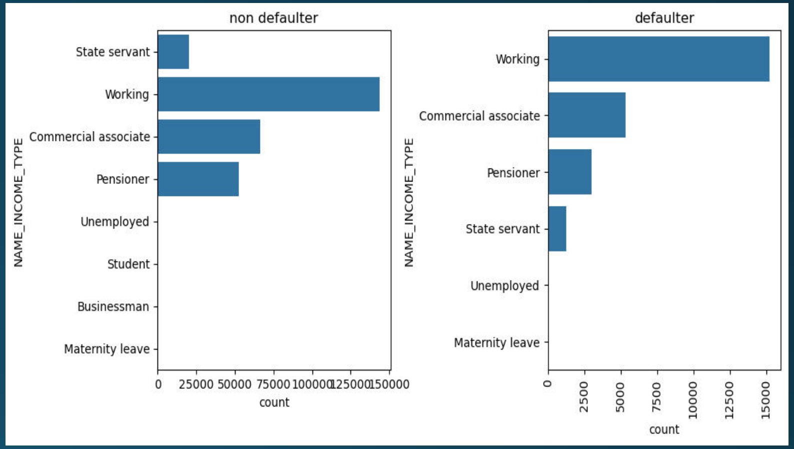
- From the chart we can depict the TARGET Imbalance. We can see 92% of client pay on time and 8% of clients have payment difficulties. The TARGET Imbalance ratio is 11.4(approximately)
- Target variable(1- client with payment difficulties: he/she had late payment more than X days or at least one of the first Y installments of the loan in our sample, 0- all other cases)

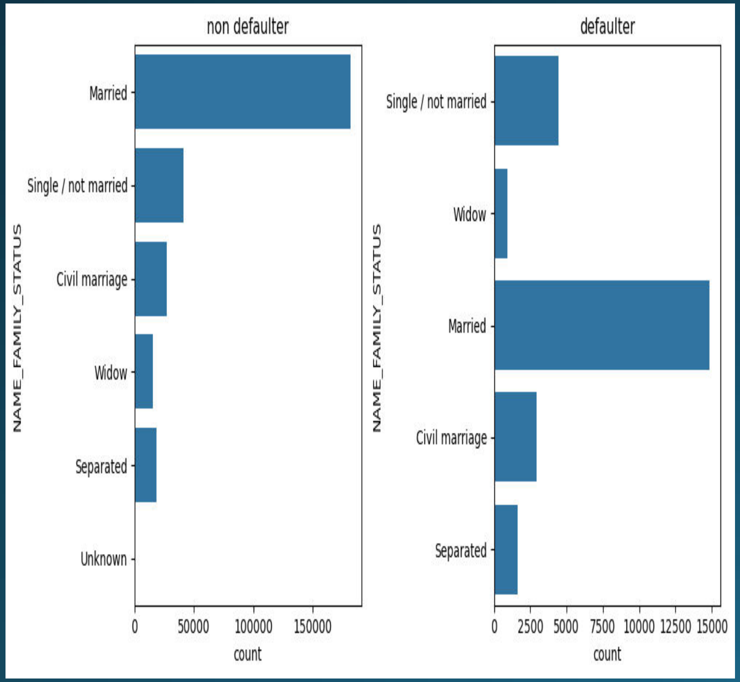


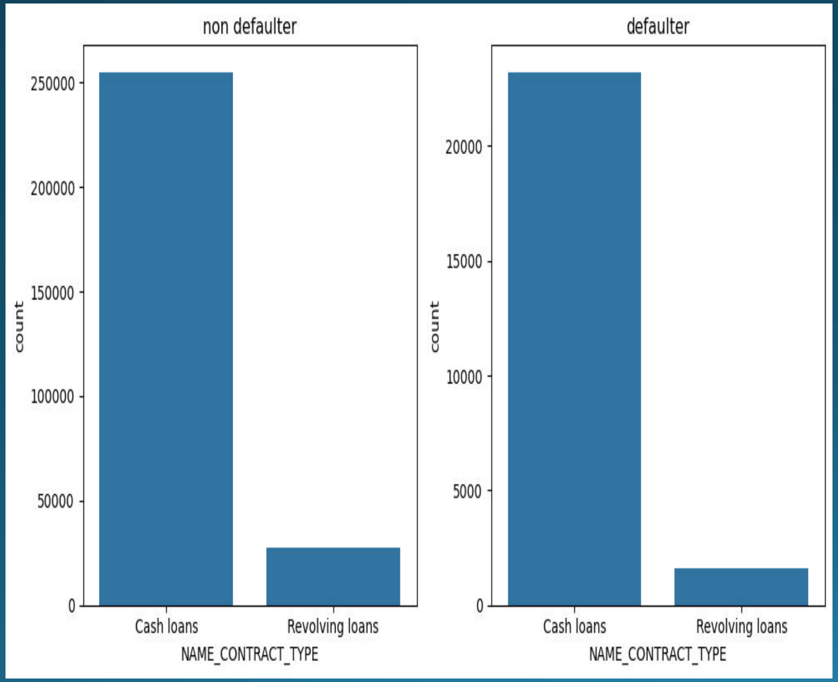
SEGMENTED-UNIVARIATE ANALYSIS

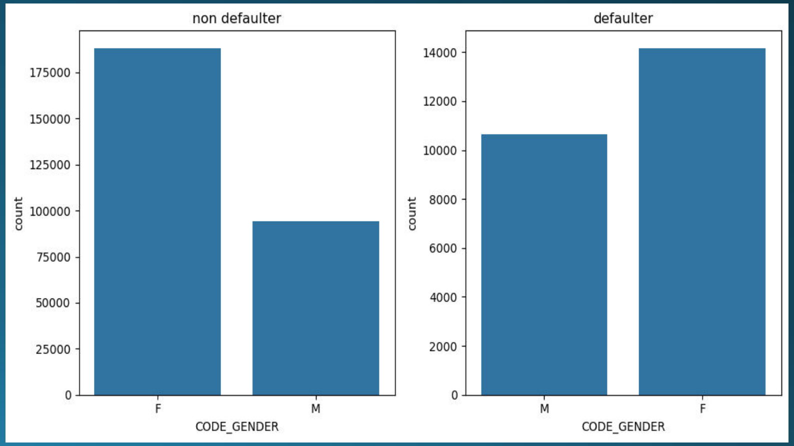
- From the graph of 'NAME_CONTRACT_TYPE' we can see more defaulters can be expected for applicants for cash loans vs revolving loans.
- The graph of 'NAME_HOUSING_TYPE' we can depict that people who own their own house/apartment are highest in both the cases.
- From the graph of 'NAME_INCOME_TYPE' we can depict that Working applicants are highest in both the cases followed by Commercial associate.
- Married applicants are maximum in number in both the cases in the graph of 'NAME_FAMILY_STATUS'.
- From the graph of 'CODE_GENDER' it is depicted that Females are maximum in both the categories.





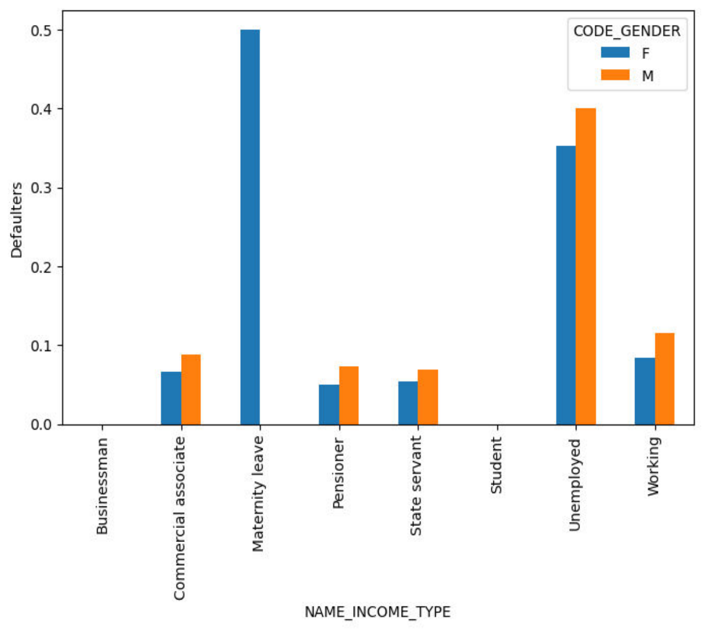


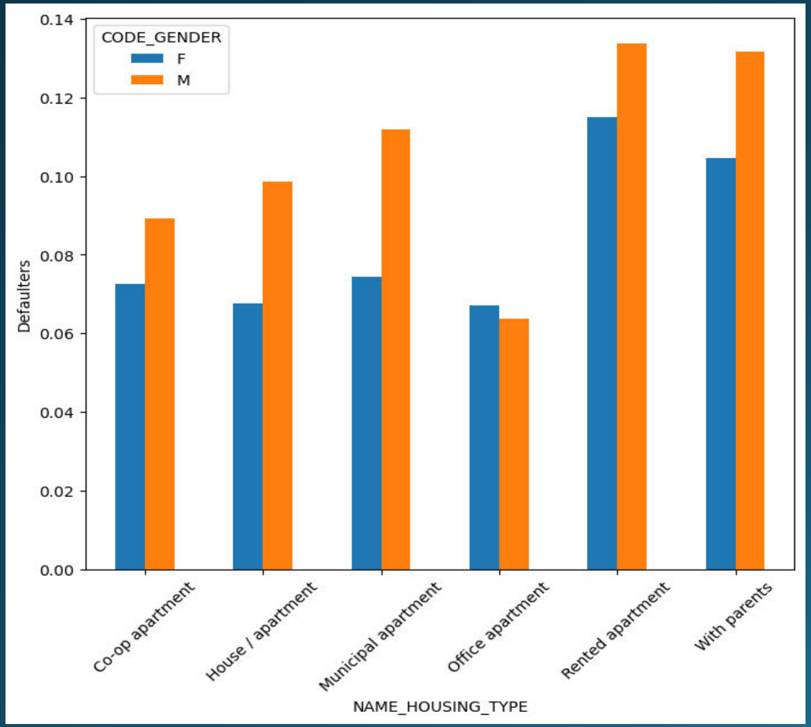


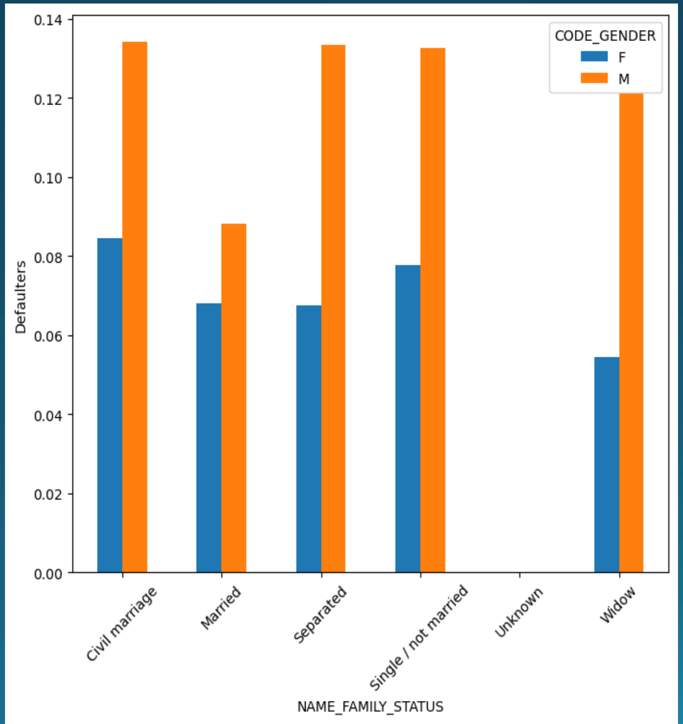


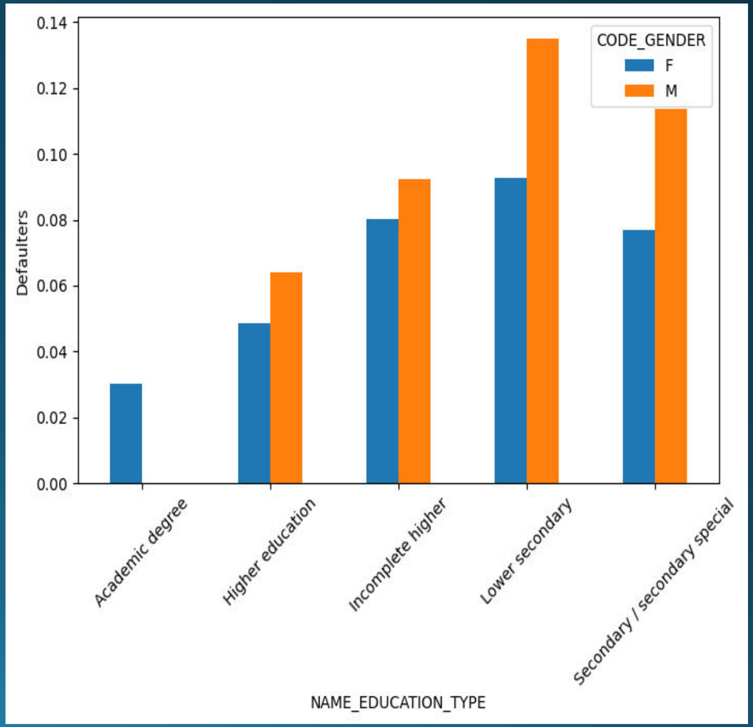
SEGMENTED-BIVARIATE ANALYSIS

- From the graph ‘NAME_EDUCATION_TYPE’ we can see that**.** lower secondary educated clients are more defaulted followed by Secondary and Incomplete higher educated clients.Moreover,the Higher educated people are less defaulted.
- The ‘NAME_INCOME_TYPE’ graph depicts that the unemployed' clients are more defaulted and clients with maternity leave are defaulted more. Males are more defaulted with their respective professions compared to females.as well**.**
- Across all Family status the Male clients are more defaulted than Female in ‘NAME_FAMILY_STATUS’ graph.
- The graph of ‘NAME_HOUSING_TYPE’ depicts that Males with 'Rented apartment' are more at fault.







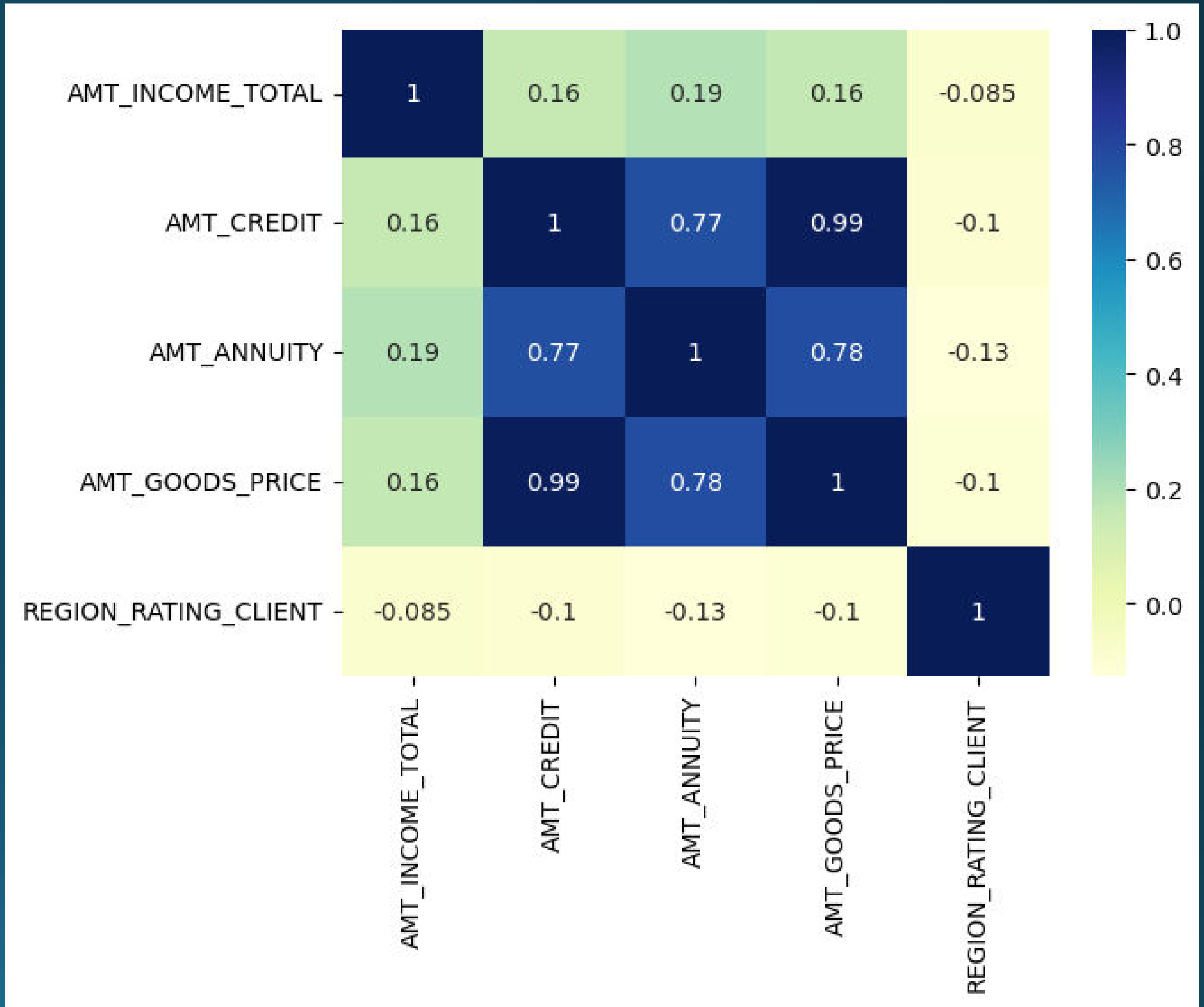


CORRELATION & HEATMAP

- The correlation coefficient ranges from -1 to 1.
- If the correlation coefficient is close to 1, it means that there is a strong positive correlation between the two variables.
- If the correlation coefficient is close to -1, it means that there is a strong negative correlation between the two variables.
- If the correlation coefficient is close to 0, it means that there is no correlation between the two variables.
- From the graph we can depict that

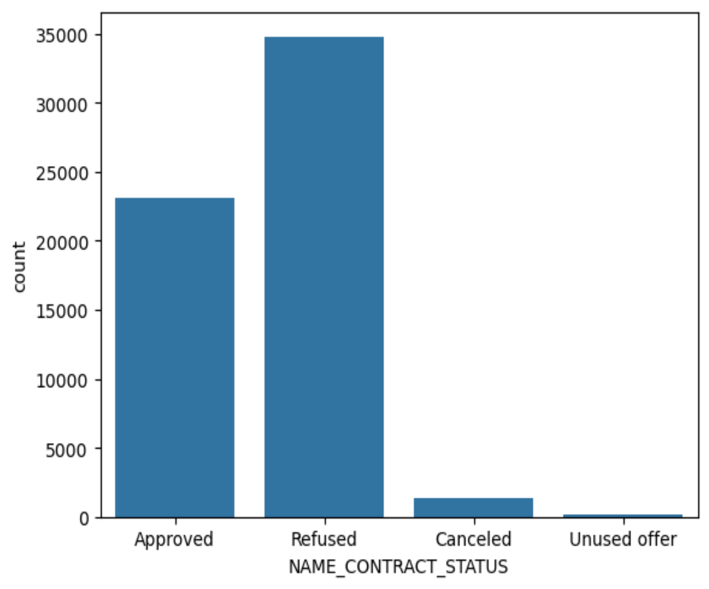
➢ 'AMT_CREDIT' and 'AMT_GOODS_PRICE' have a strong positive correlation(0.99). This means as the credit amount increases, the price of goods also increases.

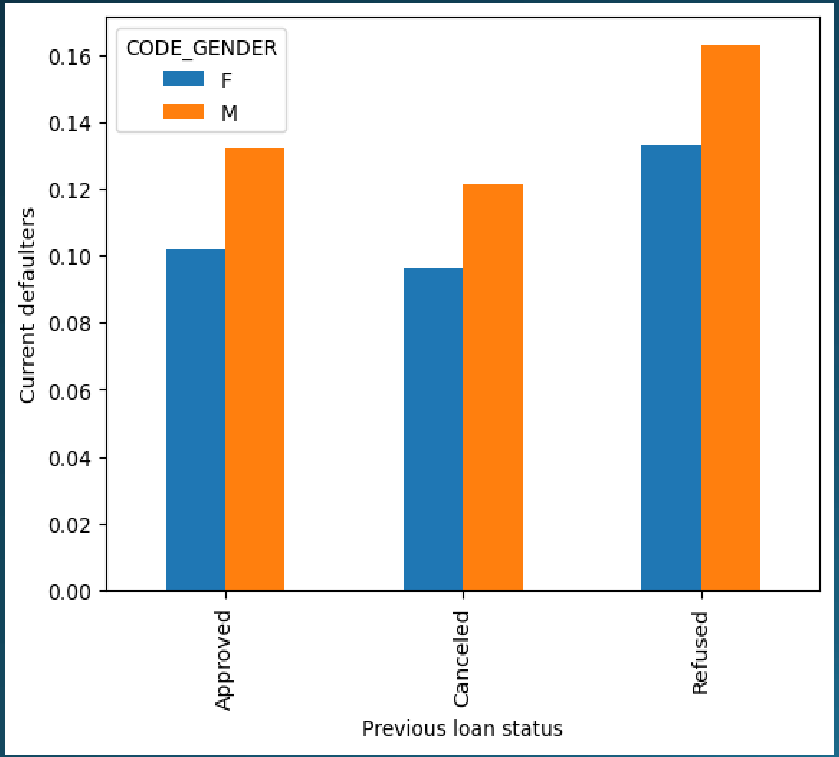
➢ AMT_ANNUITY' and 'AMT_GOODS_PRICE' have moderate positive correlation(0.78). Whereas,AMT_INCOME_TOTAL' shows weak positive correlation with 'AMT_CREDIT','AMT_ANNUITY' and 'AMT_GOODS_PRICE'

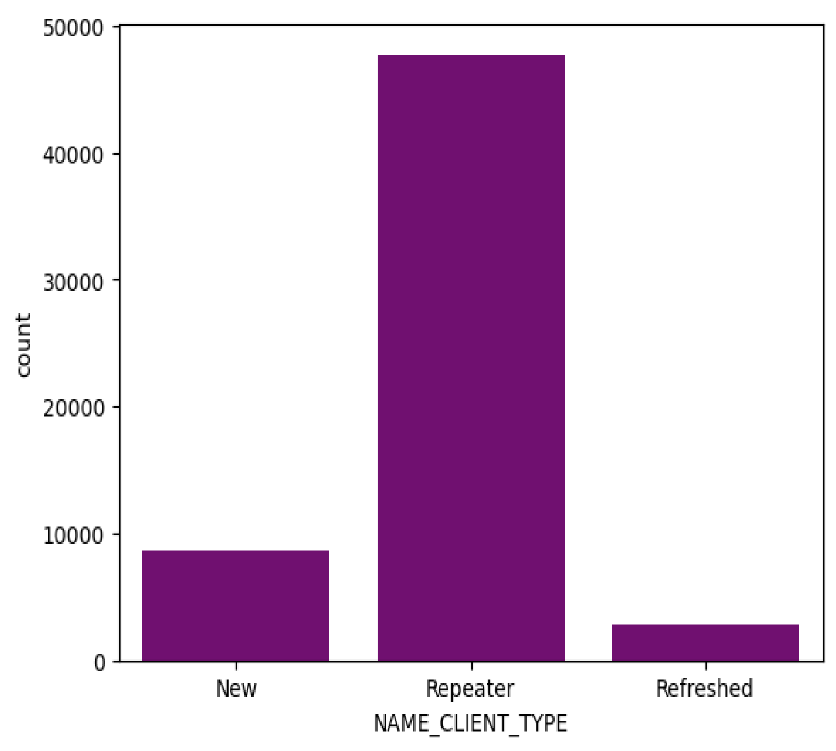


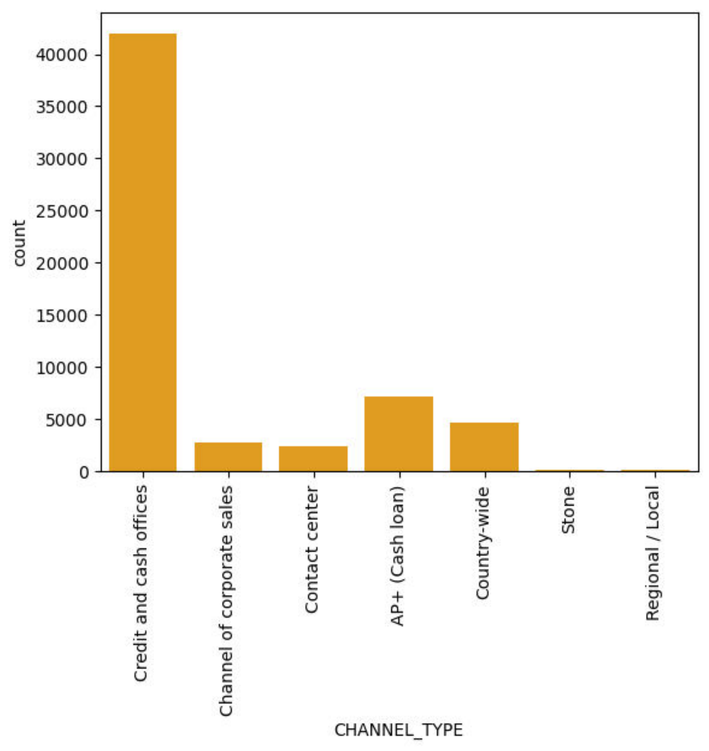
MERGED DATA ANALYSIS

- In ‘NAME_CONTRACT_STATUS’ graph we can see that most of the applicants have 'Refused' loans.
- Most of the applicants are repeater as depicted by ‘NAME_CLIENT_TYPE’.
- The ‘CHANNEL_TYPE’ graph depicts that clients of 'Credit and cash offices' are more in number.
- The graph between ‘NAME_CONTRACT_TYPE’ and ‘CODE_GENDER’ shows Previously 'Refused' clients are more defaulted than previously 'Approved' clients. Moreover Males are more defaulted than Females.









From the graph:

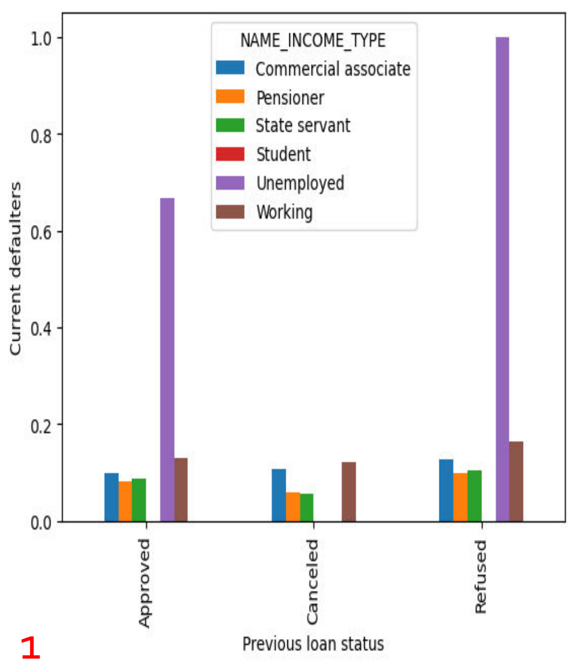
1 we can depict that for previously Approved status Unemployed clients were more defaulted, for previously Refused applicants the defaulters are more Unemployed clients and for previously Canceled applicants Working clients are maximum defaulters. No unemployed clients are defaulters.

2 we can depict that previous loan clients that are unemployed and have Secondary/Secondary special education are more number of defaulters.

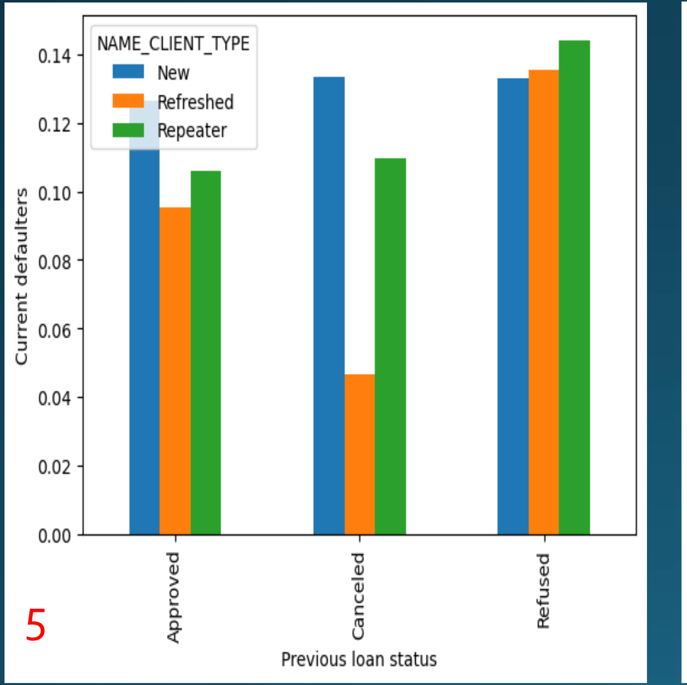
3 we can depict that for previous clients who are 'Widow' thatlive with their parents are maximum defaulters.

4 we can depict that for previously clients having Municipal Apartment having Lower secondary education are maximum defaulters.

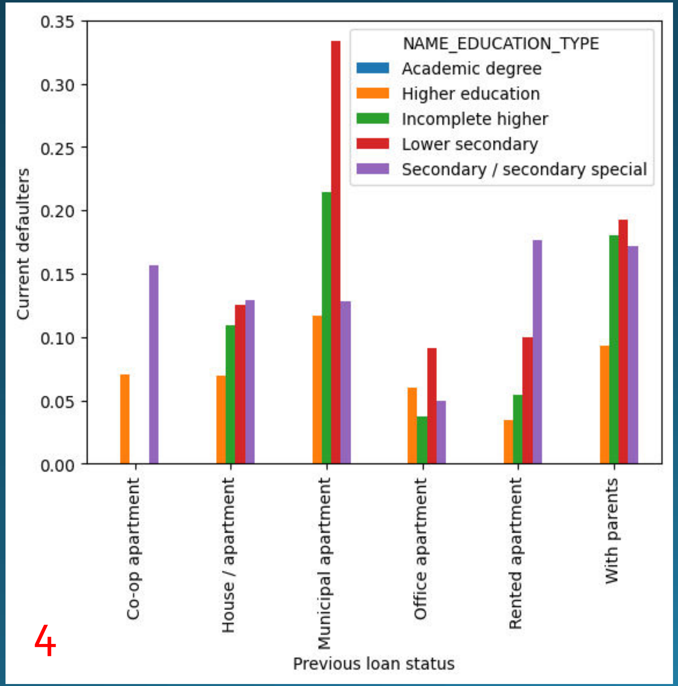
5 we can depict that For previously Approved status the New clients were more defaulted followed by Repeater and for previously Refused applicants the Defaulters are more Refused clients.Moreover,for previously Canceled applicants the Defaulters are more New clients.



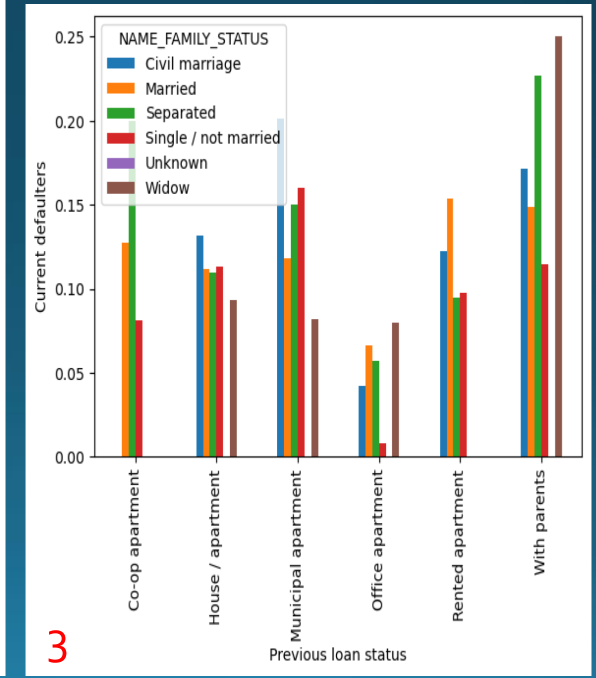
1



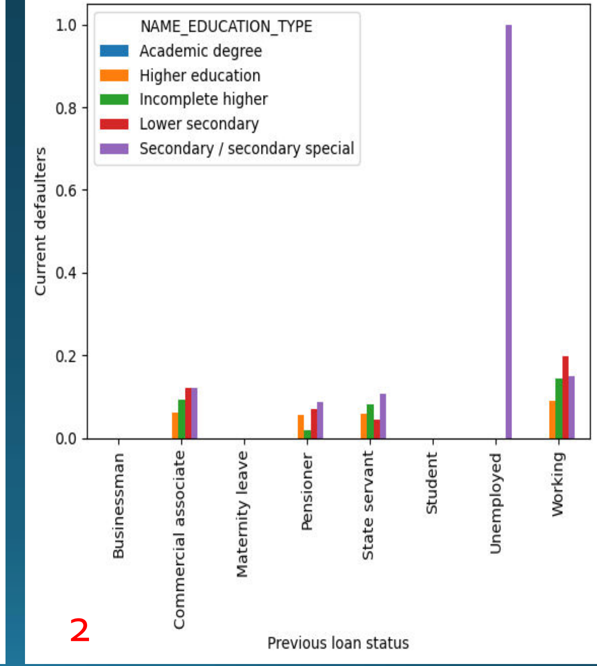
5



4



3



2

CONCLUSION & RECOMMENDATIONS

- Data is highly imbalanced. 92% of clients with no payment difficulties whereas 8% of clients have payment difficulties.
- The imbalance ratio is 11.4%(approximately)
- Most of the applicants who applied for loan are Unaccompanied have Clients with secondary/secondary special education are likely to apply for loan.
- The applicants that applied for loan are working professionals.
- Maximum number of clients own their own house/apartment.
- Maximum number of applicants have loan annuity between 20000-40000.
- Maximum number of applicants are females.
- Applicants with high credit score own House/apartment and are single.
- 'AMT_GOODS_PRICE' and 'AMT_CREDIT' have strong positive correlation. This means as Goods price increases, so does Credit Amount.
- 'AMT_ANNUITY' and 'AMT_CREDIT' have strong positive correlation. This means as Annuity Amount increases, so does Credit Amount.
- Bank should be careful while granting loan to individuals who have loan annuity between 20000-40000.
- Bank should target clients with higher education as they have less number of defaulters.
- Bank should be careful while granting loan to individuals that have rented apartment as they have maximum number of defaulters.

THANK YOU