EDA Credit Case Study

By Vivek Yadav

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

- 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 specializes 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.

Risks are associated with the bank’s

- 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.

Problem Of Customers

- **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.

Decision Taken by Client/Company

- 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.

Business Objectives

- 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. 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.

Detasets

- 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 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.

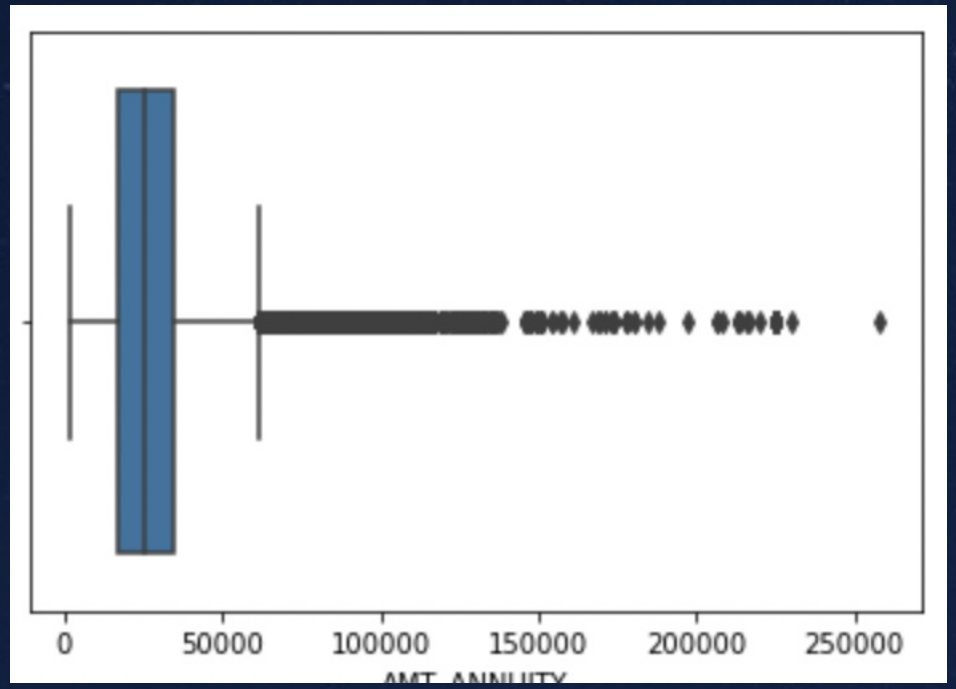
Exploratory Data Analysis

Missing Value Analysis

- ☆ Dropping the Column which have missing percentage more than 50 because we didn't get any insight from that Column because of data insufficient.
- ☆ Now impute the remaining column through statistical approach.
- ☆ For Numerical impute the Null Values with the mean/median.
- ☆ For Categorical Column impute the missing value by mode or name separate category like 'missing', 'Others'.

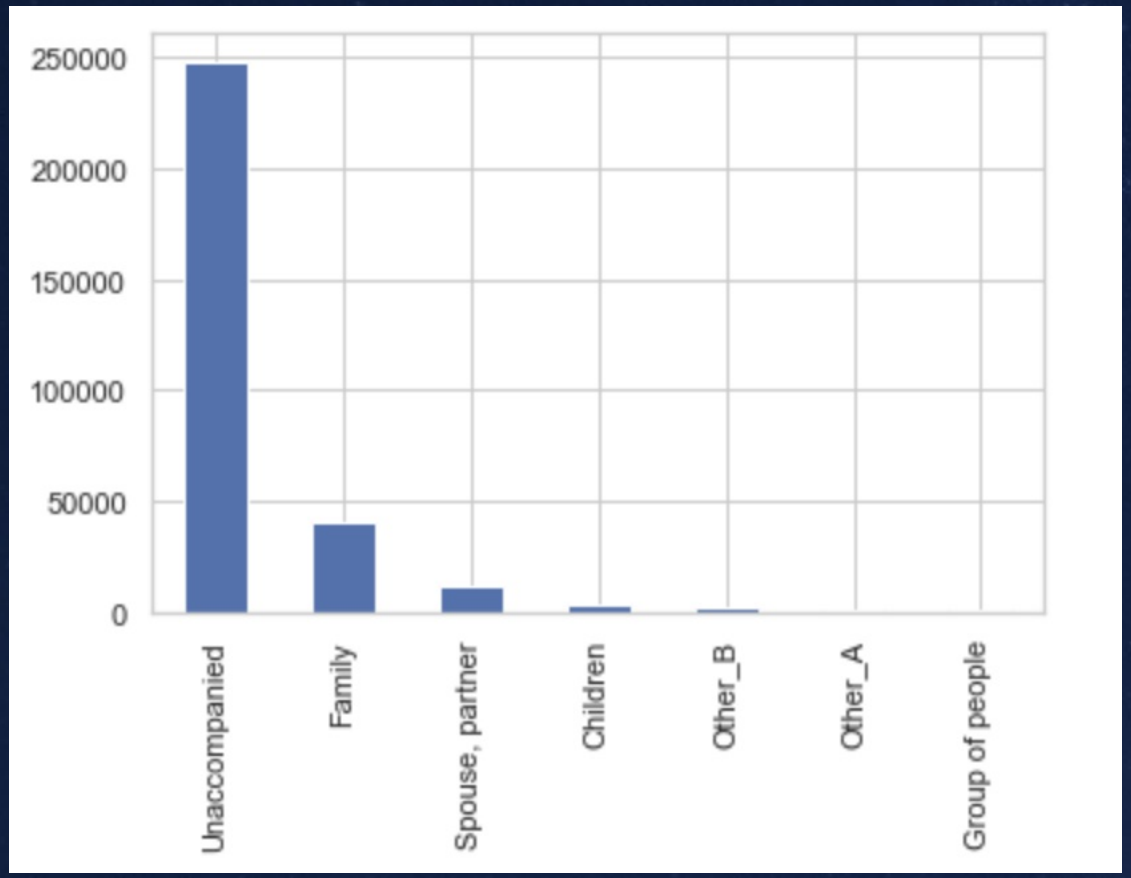

Detecting
Outliers

AMT_ANNUITY



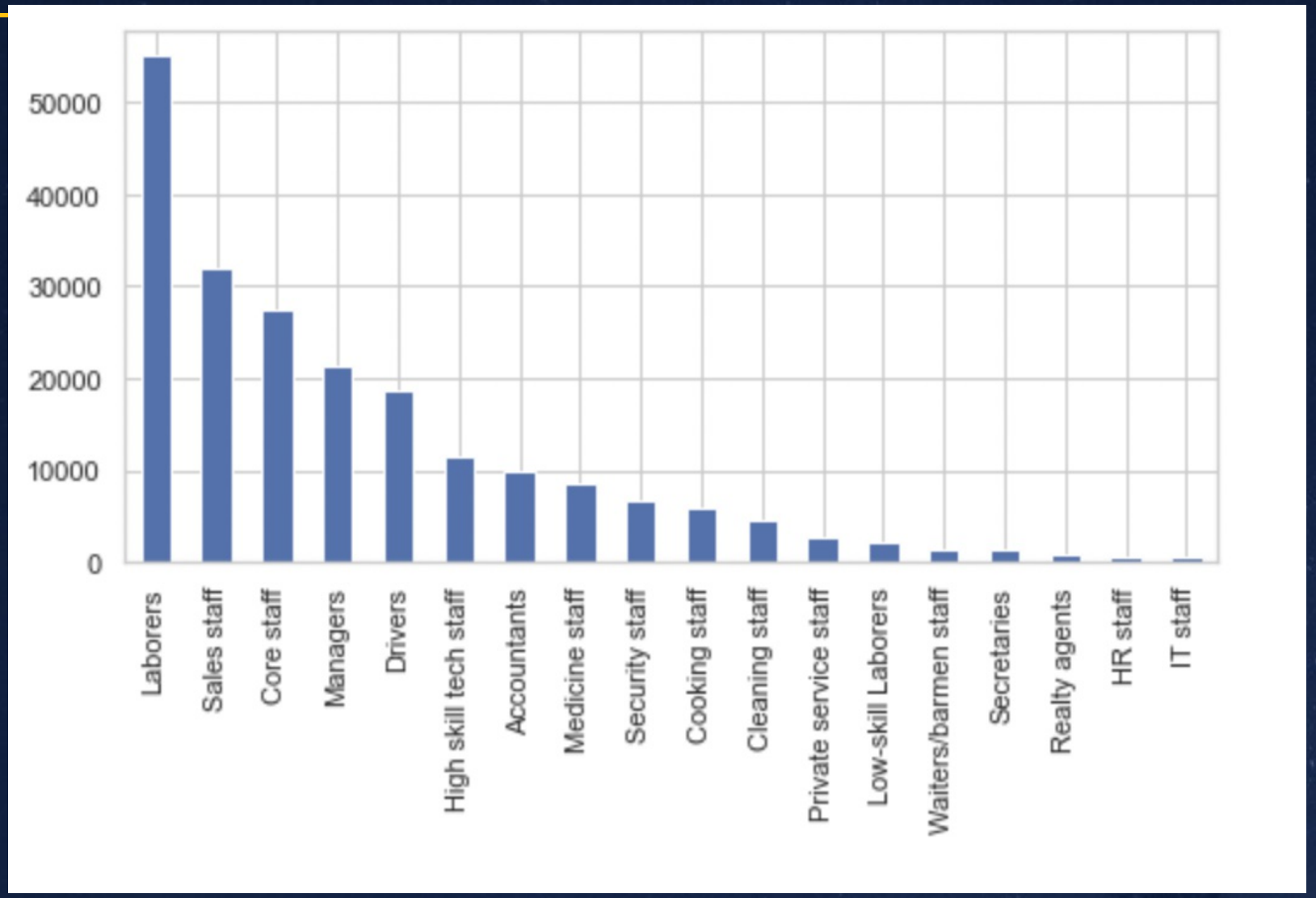
As we can easily understand if we imputing with mean then its affected by outlier so median is good option

Imputing Missing For Categorical variable



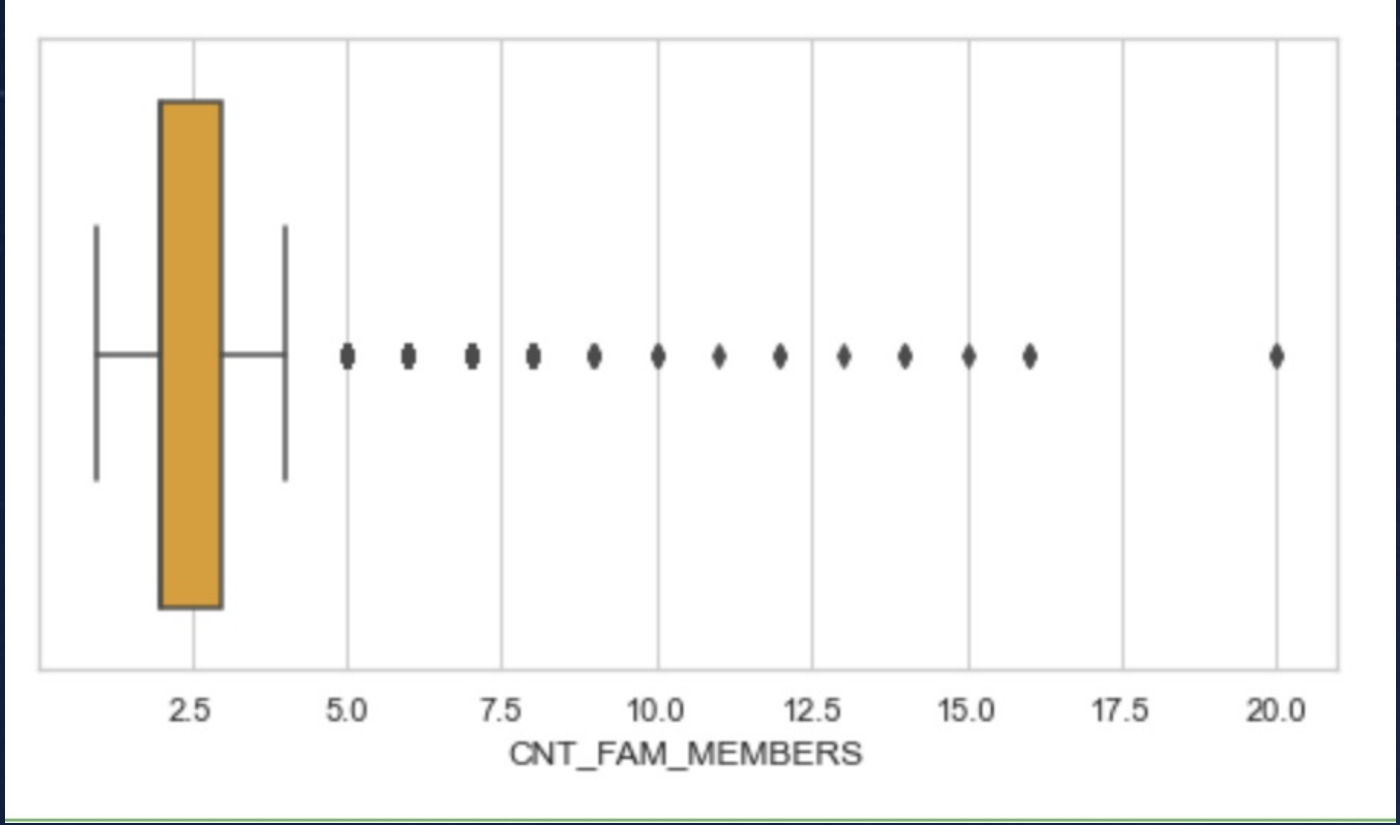
Since as compare to no of rows of dataframe the no of rows having null values is less so for this we can easily fill the null values with mode it didn't affect our analysis

OCCUPATION_TYPE



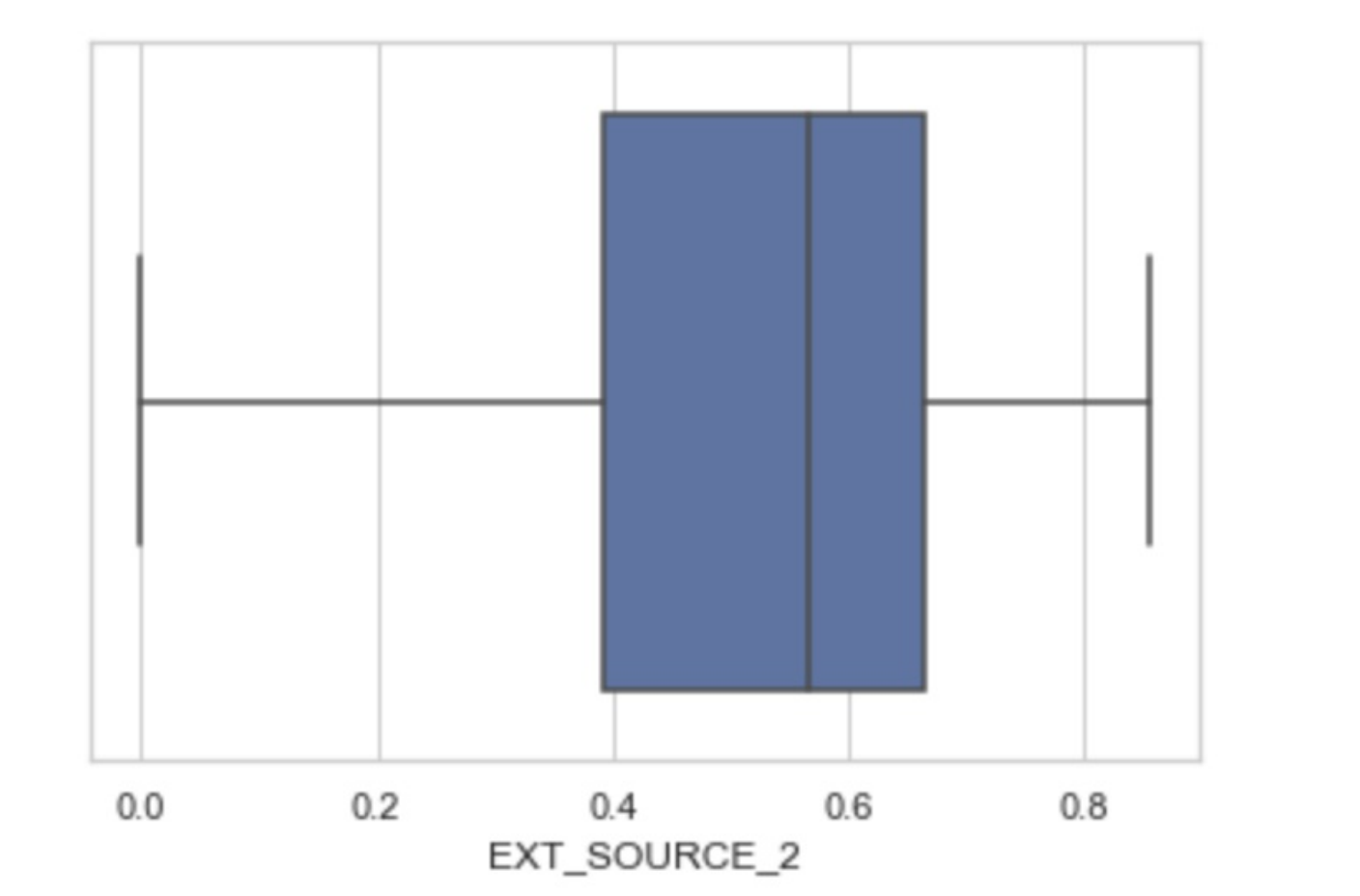
From Above observation what i understand is that lets assign new variable for the null values with others or missing for the Null value

CNT_FAM_MEMBERS



1.Here 20 is an Outlier even more than 10 is very rare. So on the basis of my observation median is good statistical way impute the null values because median is not affected with an oulier.Here we can also impute with mean beacuse here mean and median both are equal.

EXT_SOURCE_2

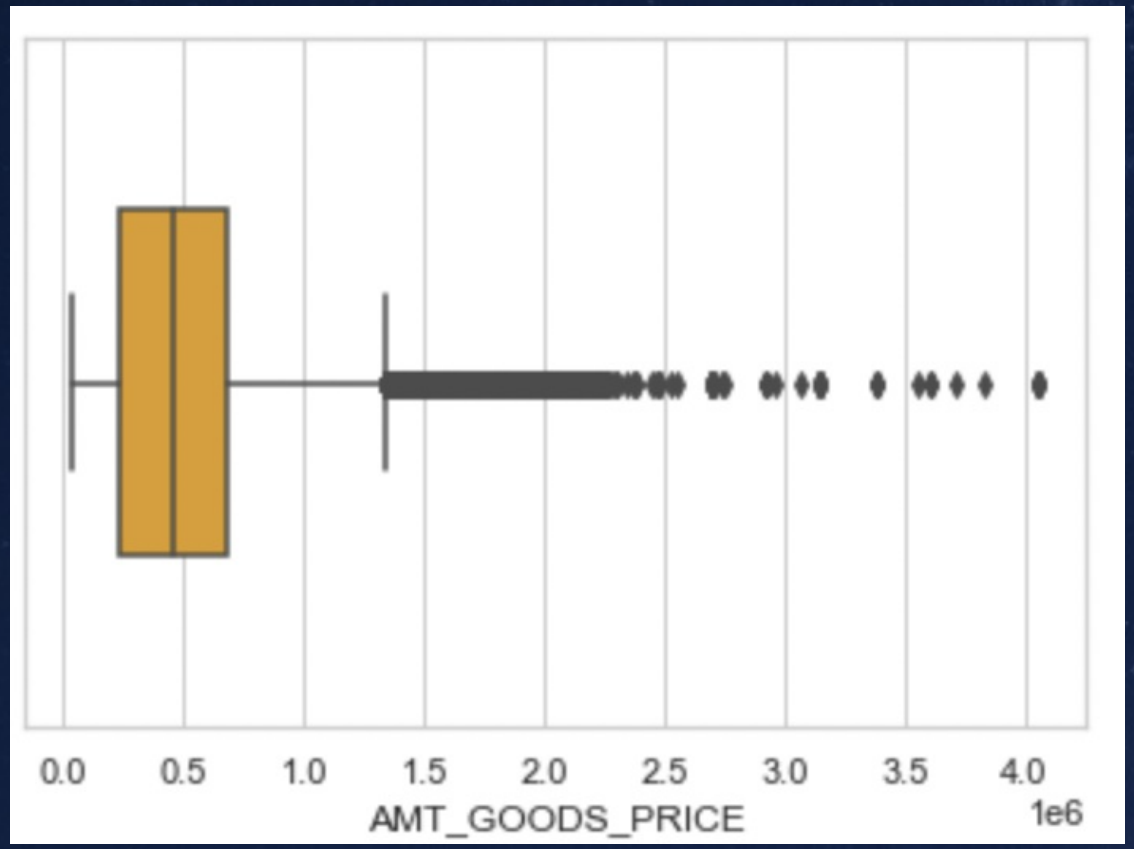


As we seen from boxplot there is no outlier so we can impute both mean and median but from spread point of view it is right Skiwed means that most of data points is toward right side so median is good option here

Some Correctness is Needed in Some Column

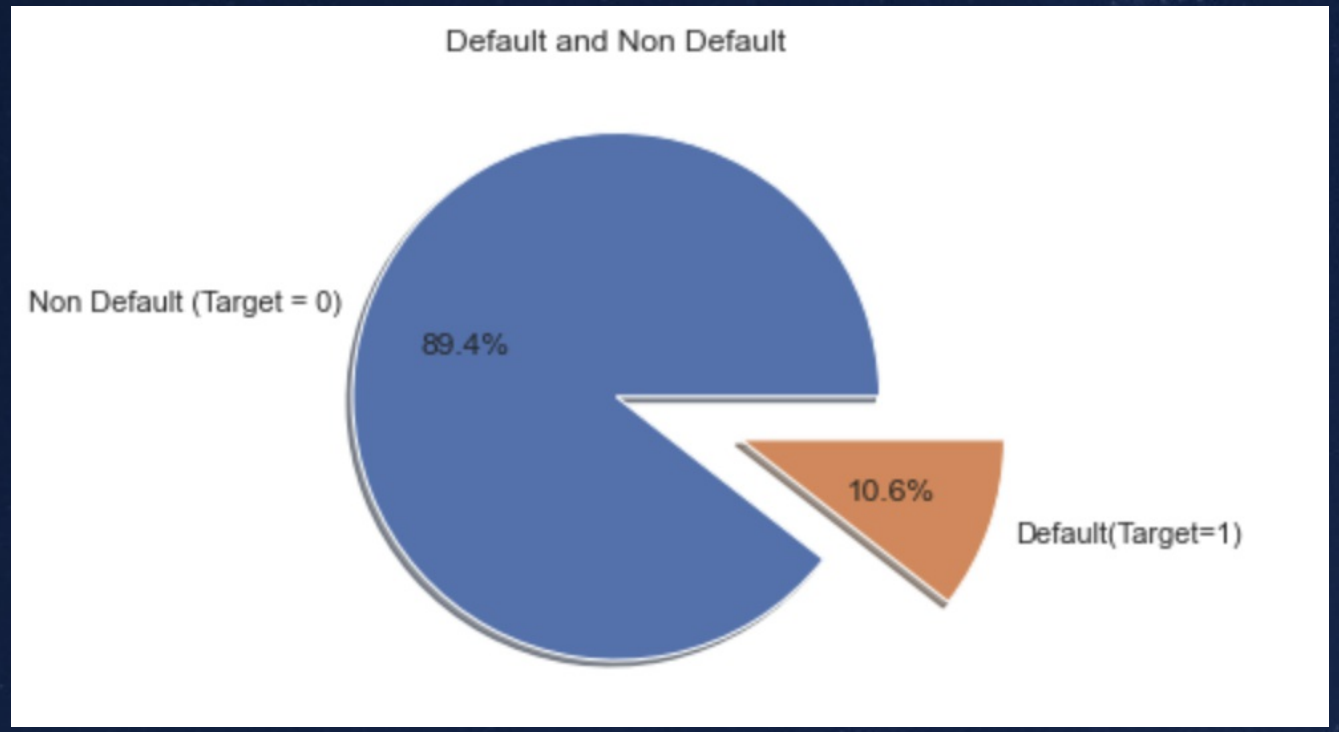
- ☆ In Days Column Having Values which are negative so change into positive
- ☆ Group data into category i.e Binning the Days_birth Column because we can easily Visualize each category easily like which group of people belong to adult ,Teenent or senior Citizen also get better insight from there
- ☆ Changing data types

AMT_GOOD_PRICE



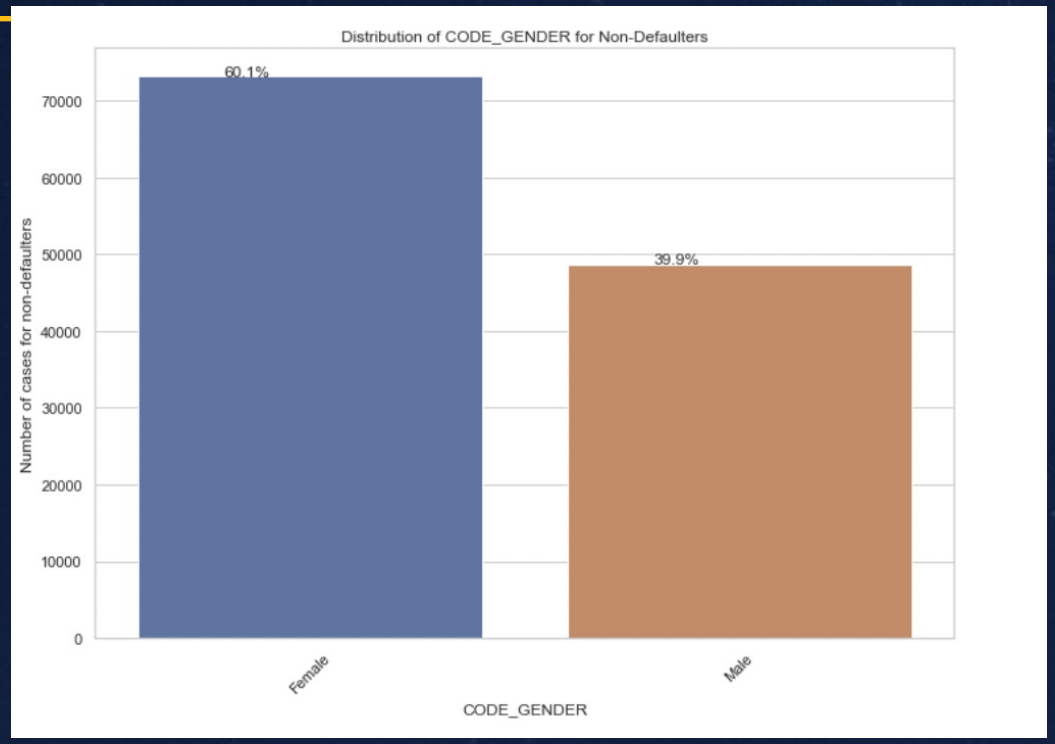
From the Boxplot Distribution we see so many outlier so for this we impute the null values with median value

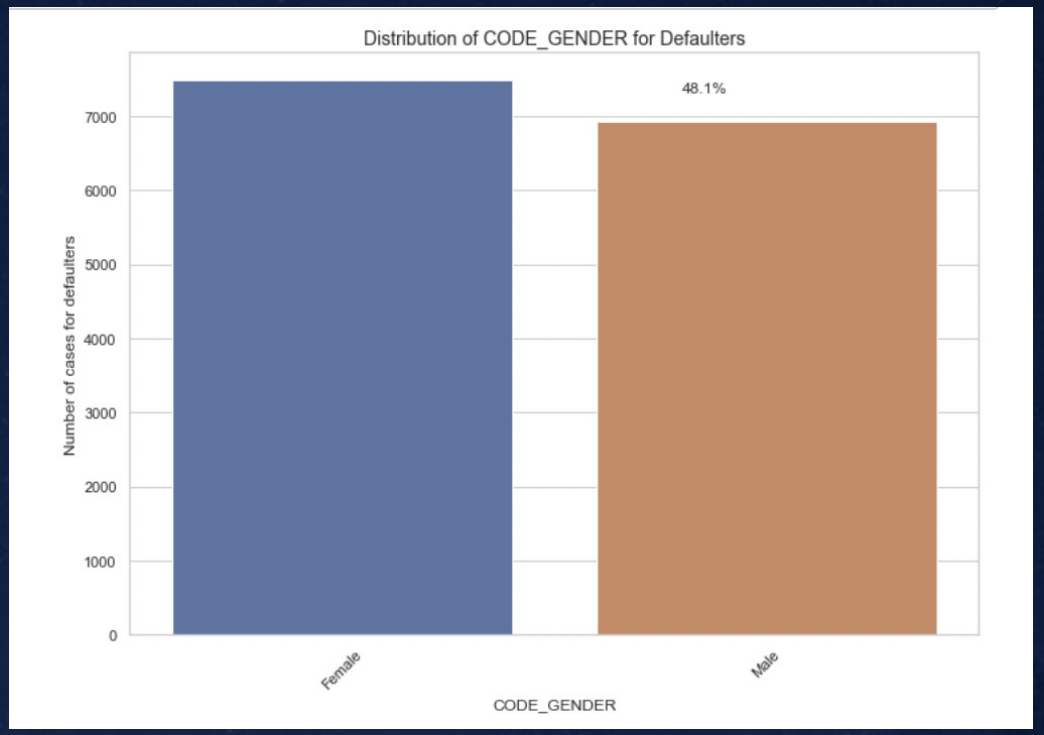
Checking the Imbalance in Target



From the Pie chart we can easily identify that approx 90% of Customer are non Defaulter while 10% of Customer Got Defaulter
Means that out of 10 ,one got defaulted so it is very important to analyse that what actually reason behind this for any company

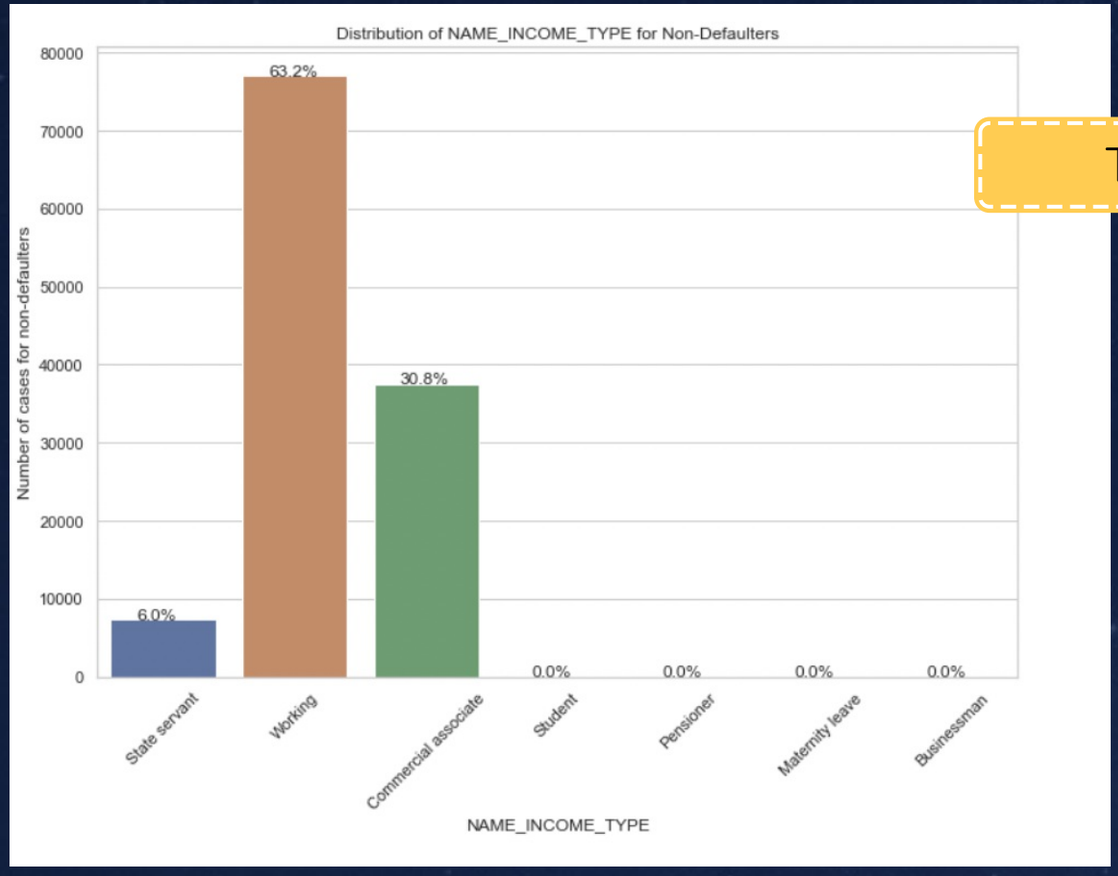
Univariate Analysis
Categorical:CODE_GENDER



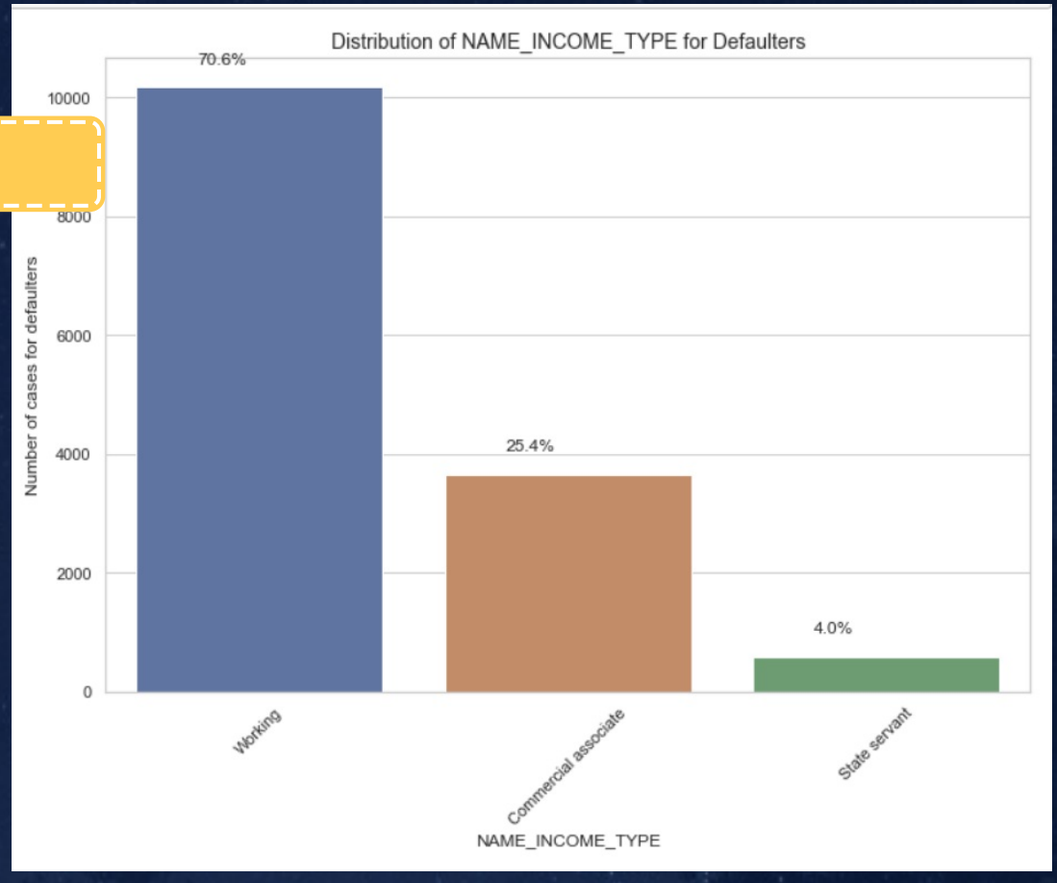


We see Female Percentage in Non-Defaulter is 60% and but in case of Defaulters its decrease by 8% i.e 52%
Its means More no of Female apply for Loan in compare to male and also default more than the male
But the rate of Defaulting the loan for female is less than man

Categorical:NAME_INCOME_TYPE



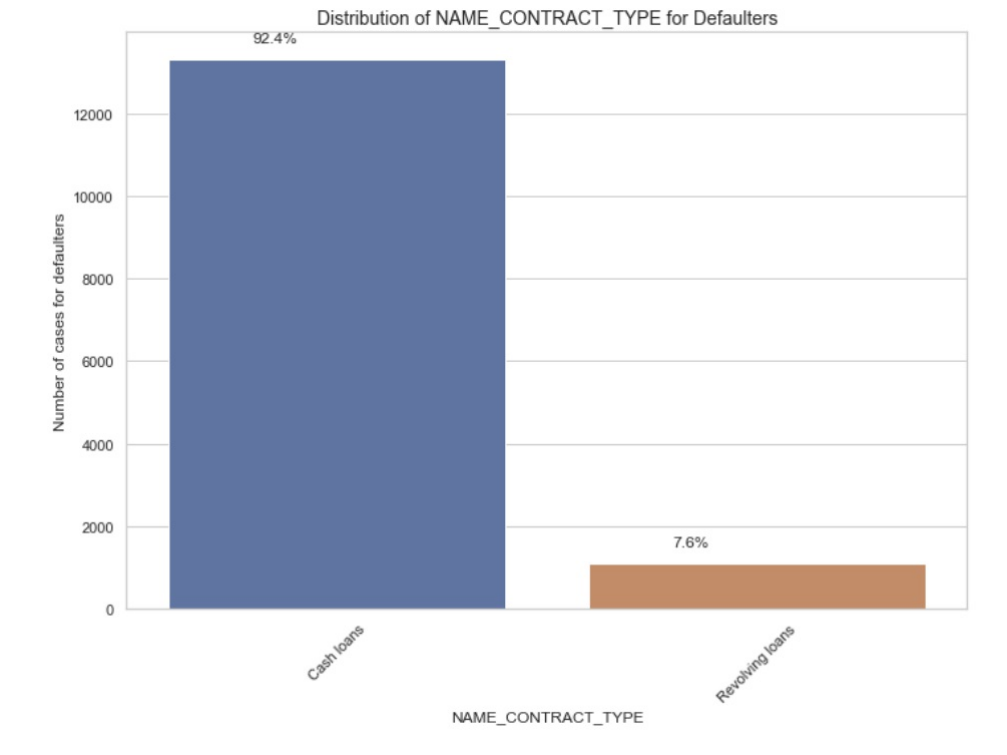
Text

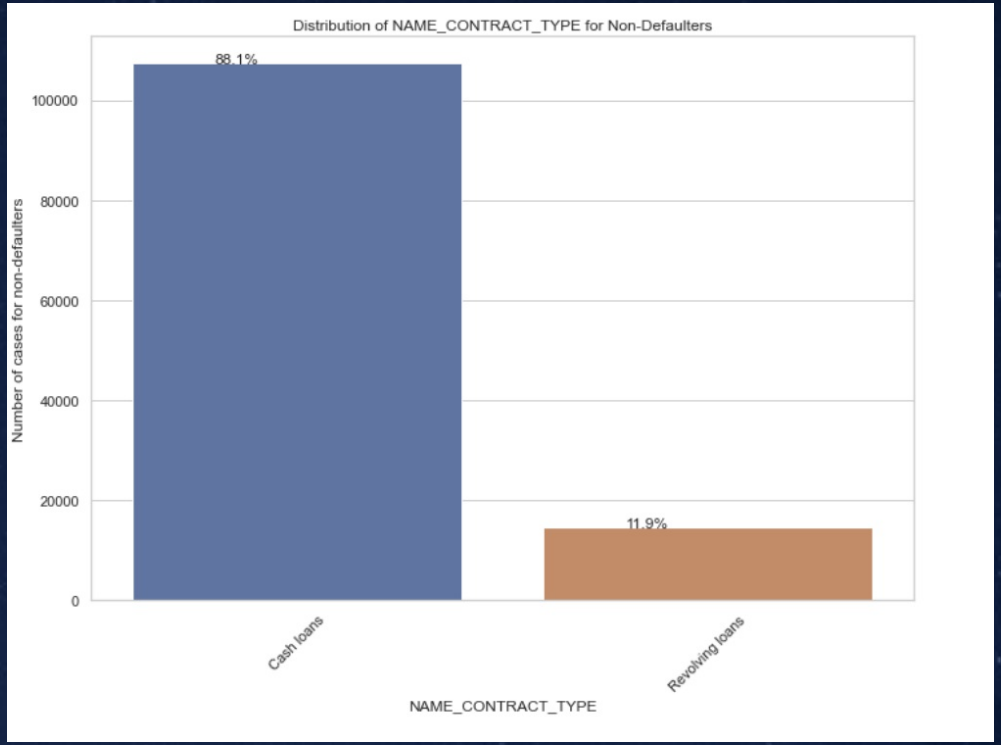


Observation:

- ☆ 63 % of Working Employees cames into Non-Default and 70% comes under the Defaulter so what we get from here is that no of the taking the loan for working employees is more and also at same time no of case of Defaulting is more.Usually Working employees taking the loan in huge number ..
- ☆ If we talk about Student ofcoarse they have no income source so usually they are not taking the loan and also probability of getting Default is less
- ☆ State servent taking the loan in rare case and also the loan amount is very less so chance of getting default is less in compare to others
- ☆ Commertial associate percentage for the Non_default case is more than the default cases

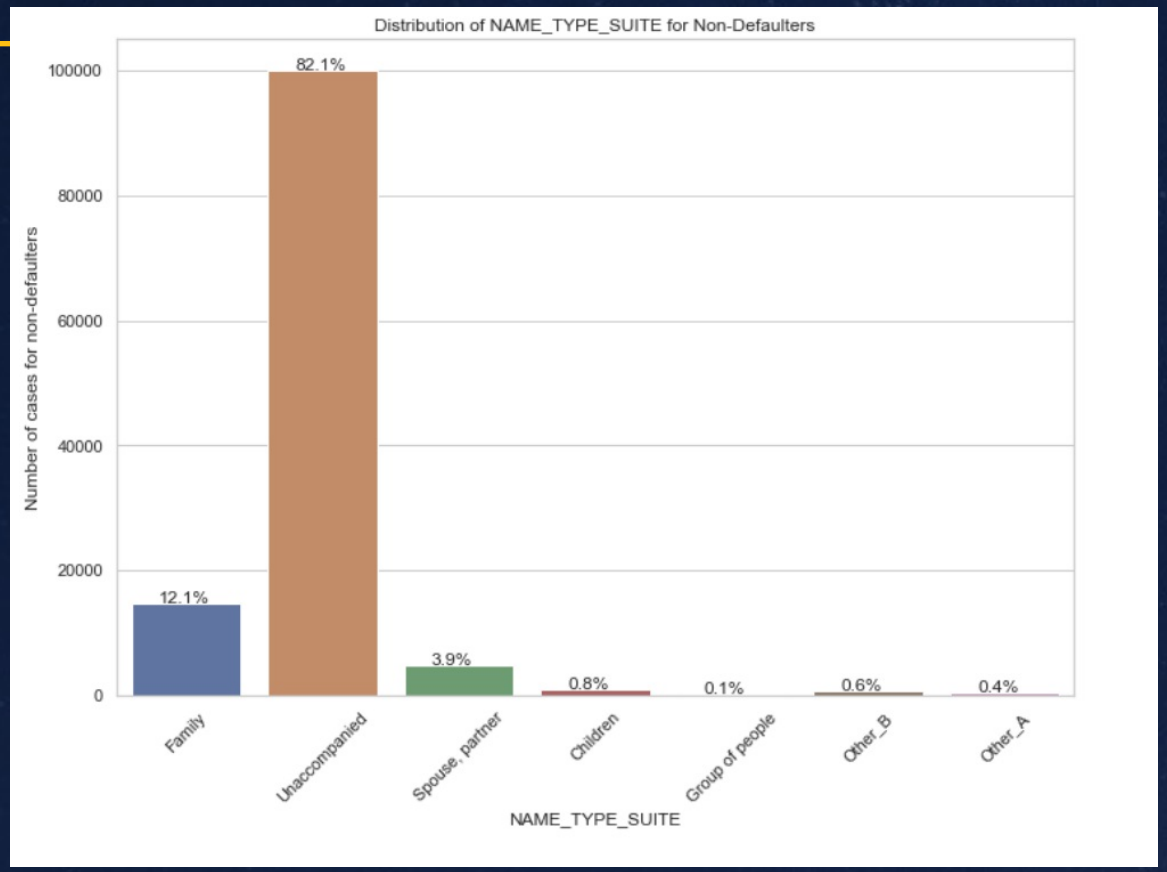
Categorical:NAME_CONTRACT_TYPE

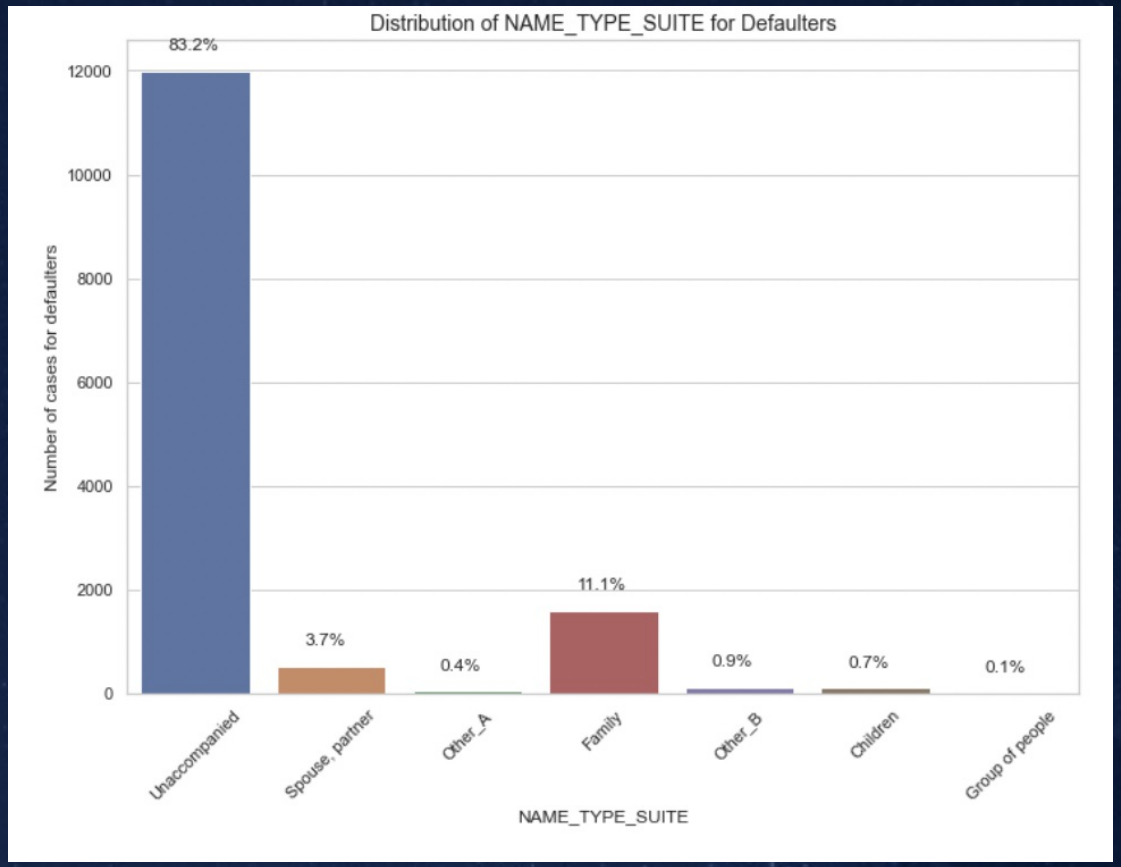




Most Of loan Taken via Cash Loan Type in both the Case Non-Defaulter and Defaulter
Revolving Loan are less Frquent because its is basically for short Term period
Most of Defaulter Occur in Cash loans

Categorical:NAME_TYPE_SUITE

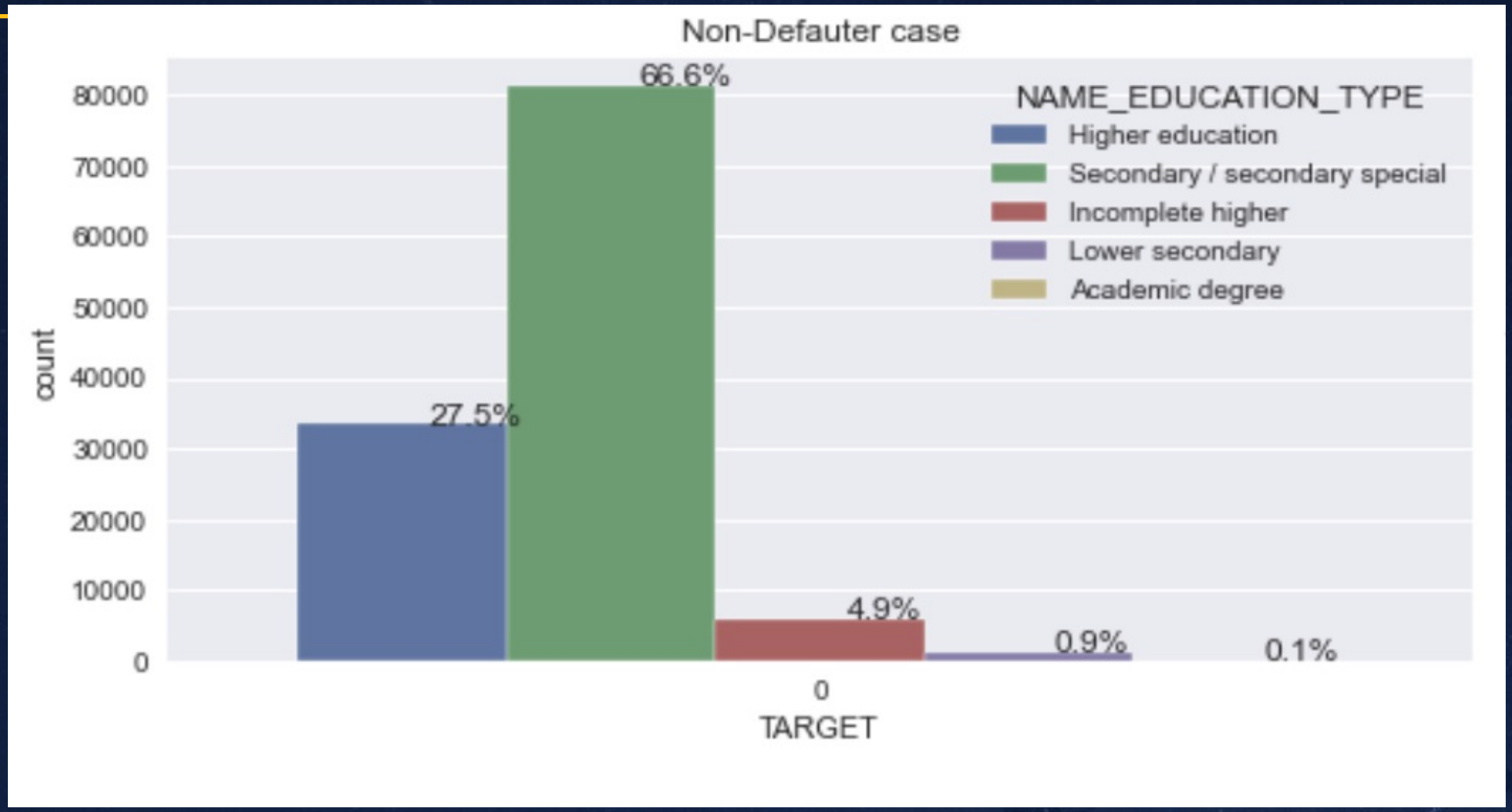


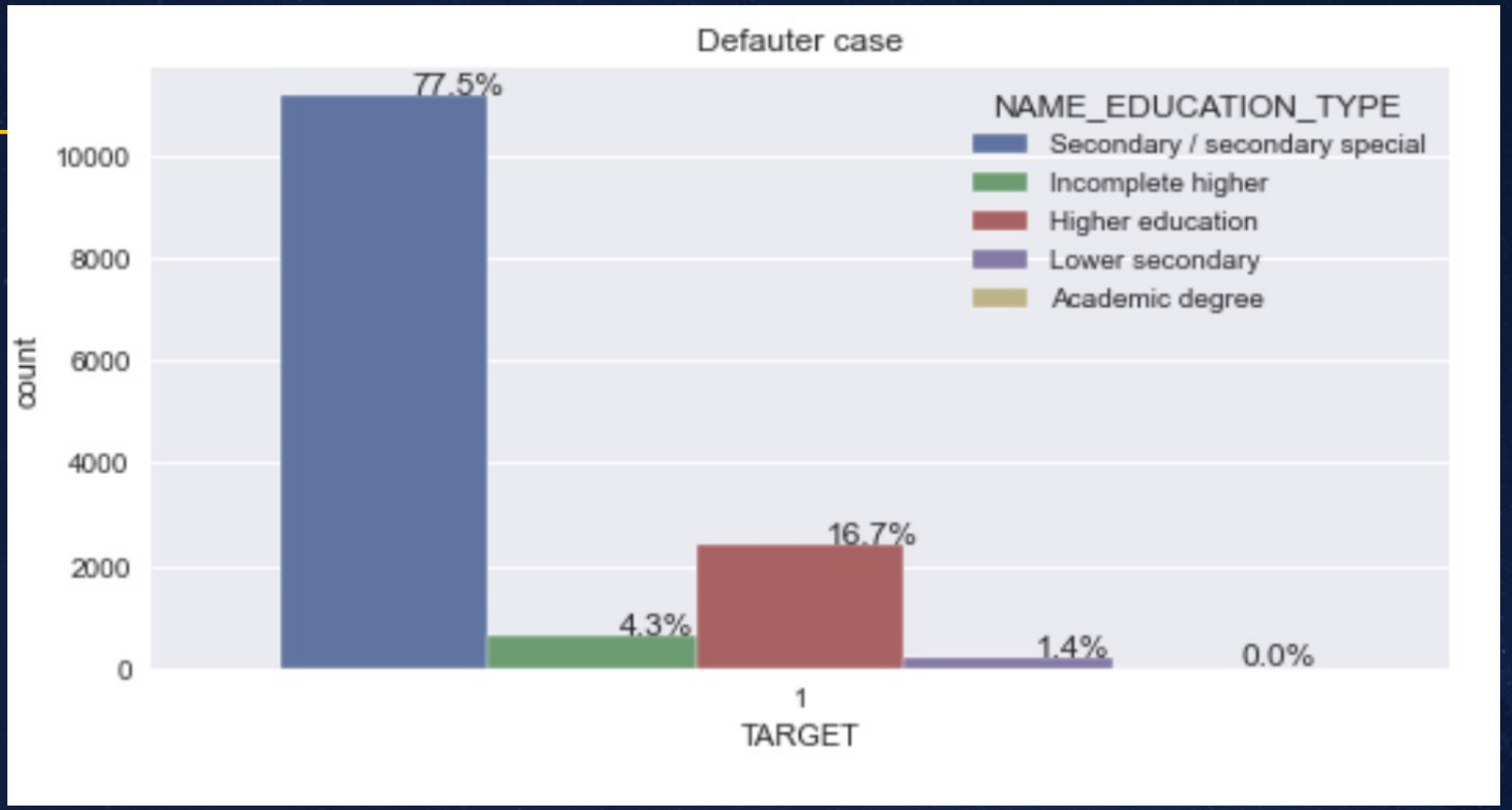


Observation:

- ☆ Means 82% of customer taking loan from Unaccompanied so that most of case of Default occur in Unaccompanied
- ☆ Some Cases comes under the Family which are basically less Compare to Unaccompanied so that Default case also less,
- ☆ In case of Children Taking loan is few but at same time ratio between Non-Default and Default is high because mostly student unable pay their loan so they comes under Defaulter Case.

Categorical:NAME_EDUCATION_TYPE

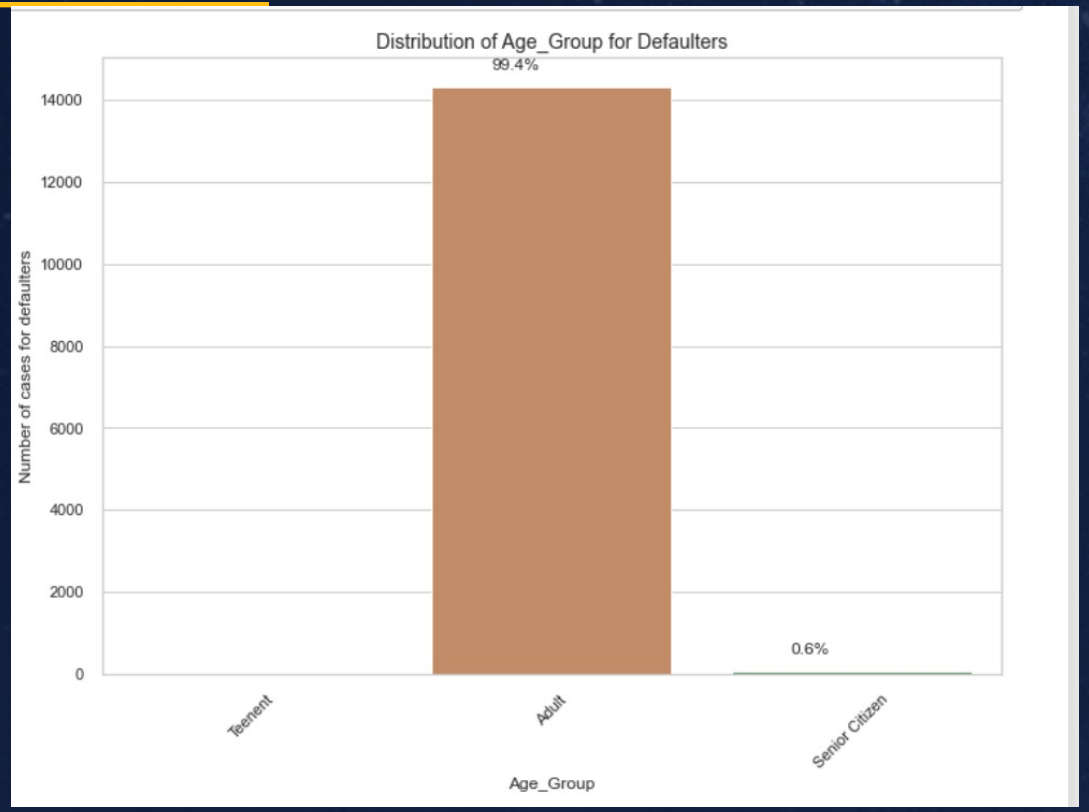


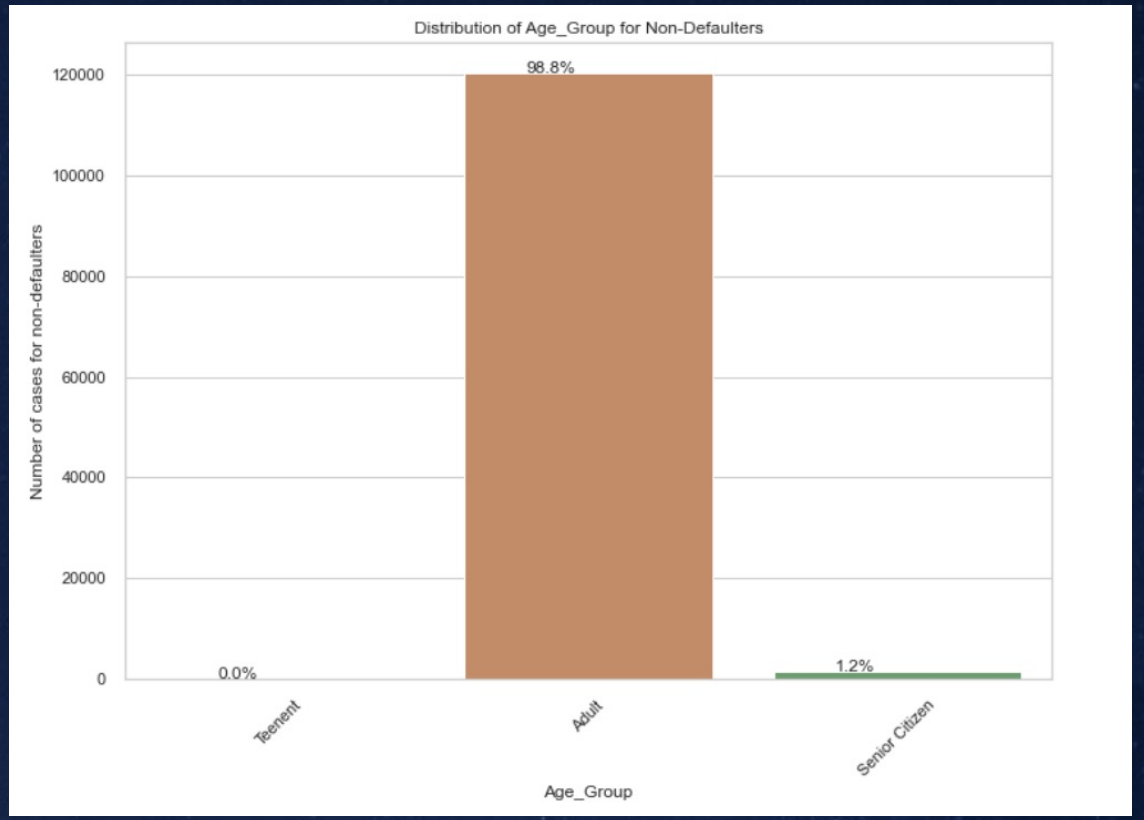

Defauter case
77.5%
4.3%
16.7%
1.4%
0.0%
NAME_EDUCATION_TYPE
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
10000
8000
6000
4000
2000
0
count
1
TARGET

Observation:

- ☆ 66% of the customer Having Secondary/Secondary Special taking the loan Comes under the Non-Defaulter as well as defaulter while percentage of case in defualter is 77.5%.Because Usually More no of customer taking loan more the chance of Default
- ☆ But the Other Hand People having Lower Education mostly likely Default

Categorical:Age_Group

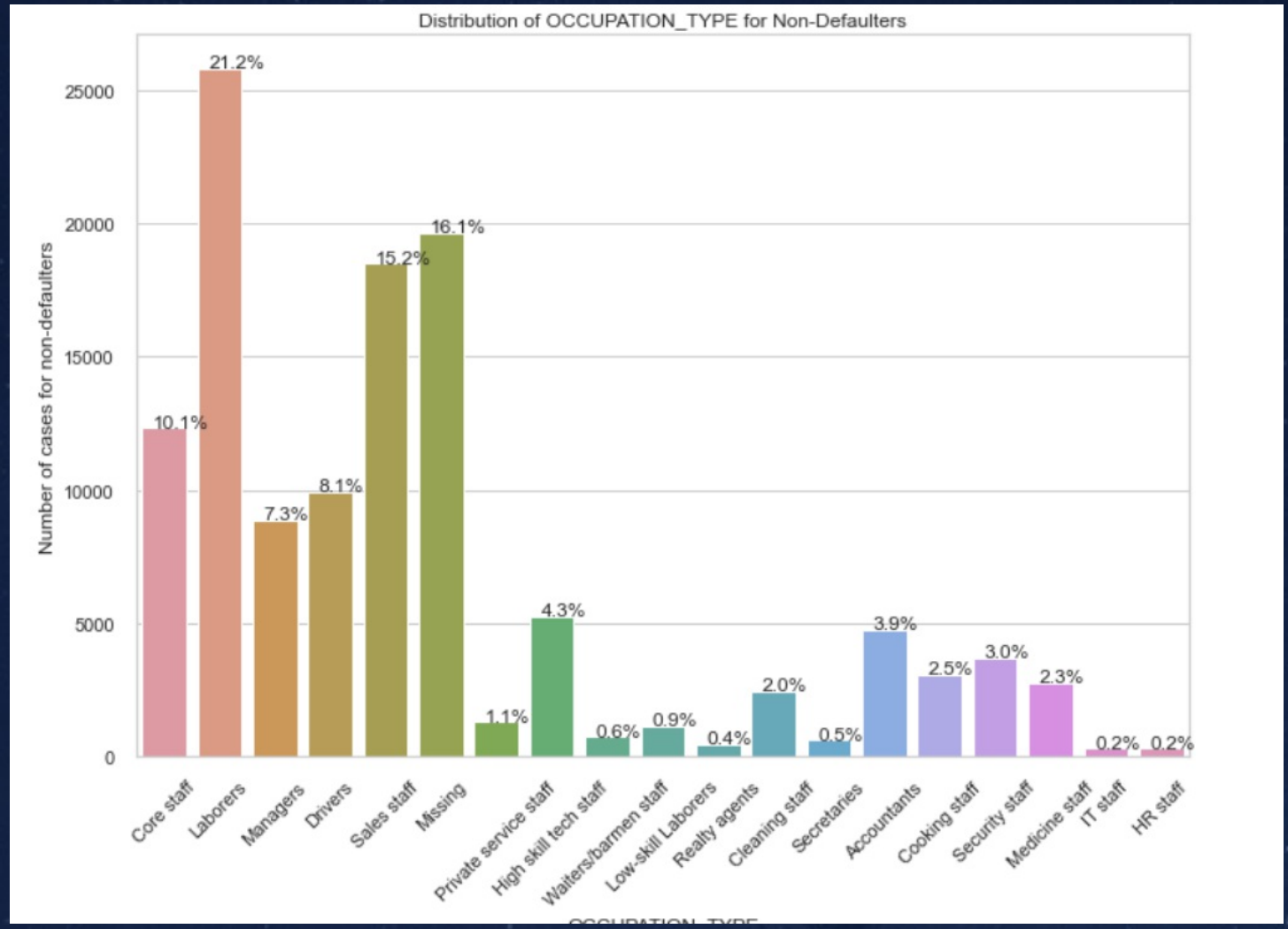


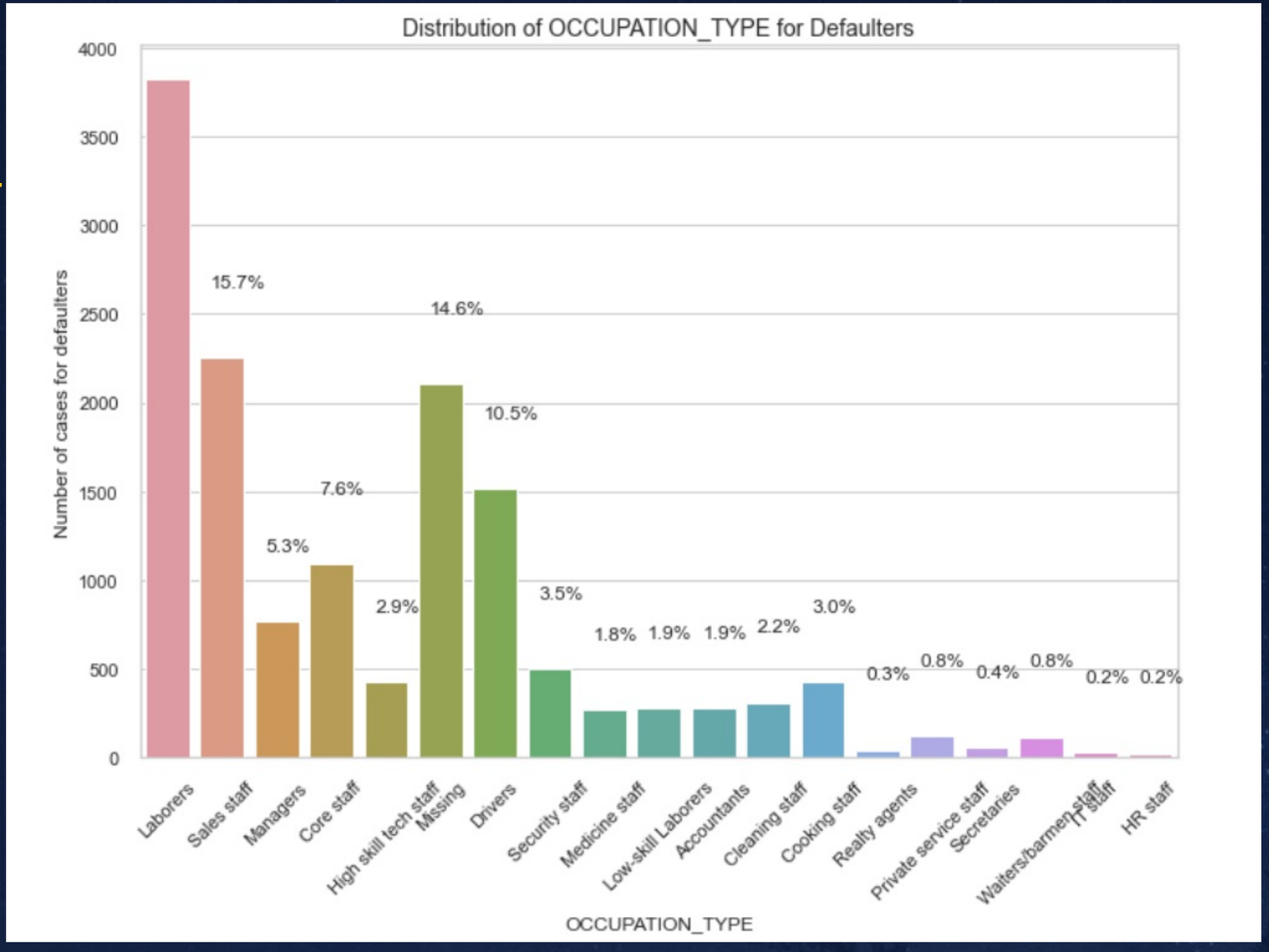


Observation:

- ☆ Most of customer taking the loan in Adult section at same time Most Of Defaulter cases in Adult
- ☆ Bank do not entertain the Senior citizen and Teenet Customer Becuase senior Citizen are pensionar and teenent are non_service

Categorical:OCCUPATION_TYPE

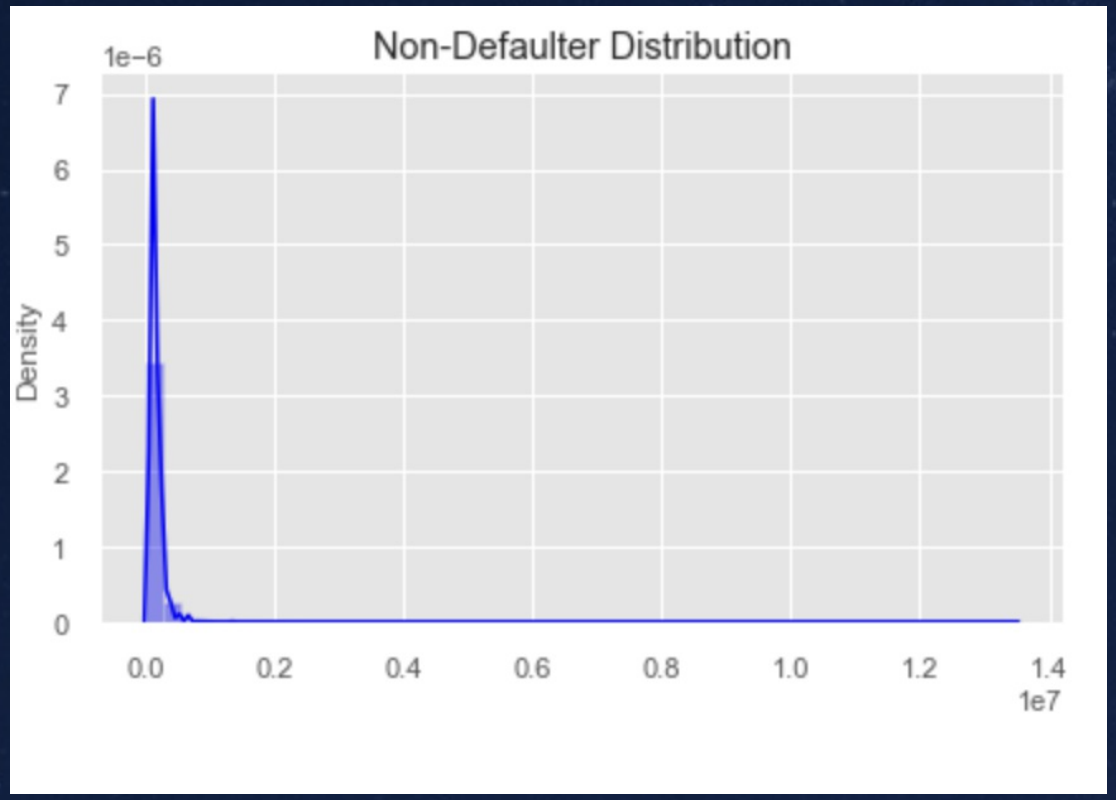


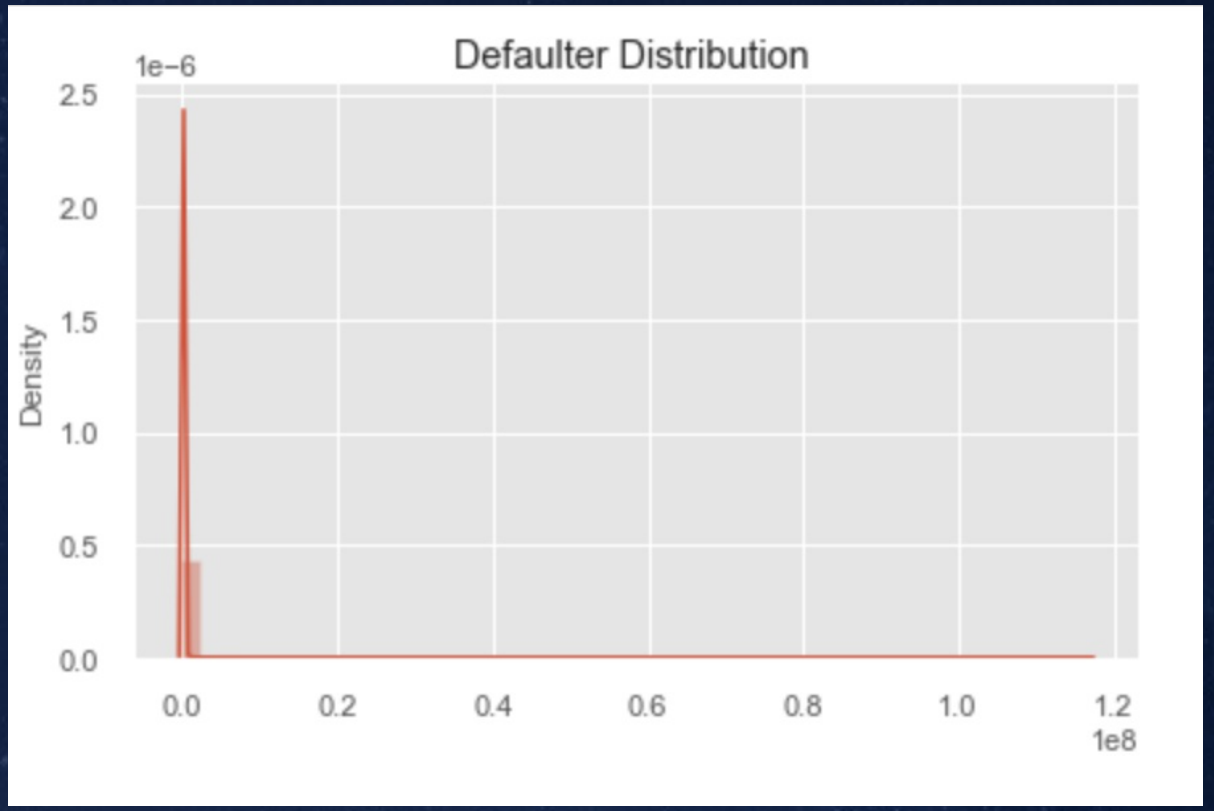

Distribution of OCCUPATION_TYPE for Defaulters
Number of cases for defaulters
4000
3500
3000
2500
2000
1500
1000
500
0
15.7%
14.6%
10.5%
7.6%
5.3%
3.5%
2.9%
3.0%
2.2%
1.8%
1.9%
1.9%
0.3%
0.8%
0.4%
0.8%
0.2%
0.2%
Laborers
Sales staff
Managers
Core staff
High skill tech staff
Missing
Drivers
Security staff
Medicine staff
Low-skill Laborers
Accountants
Cleaning staff
Cooking staff
Realty agents
Private service staff
Secretaries
Waiters/barmen staff
IT staff
HR staff
OCCUPATION_TYPE

Observation:

- ☆ Most of Customer in Labourers segment approx 21.2% comes in Non_defaulter and 22% in Defaulter.Its is true because they have less source of income
- ☆ Most of Customers Taking the Loan which have Less Income like Core Staff,driver etc

Numerical: AMT_INCOME_TOTAL

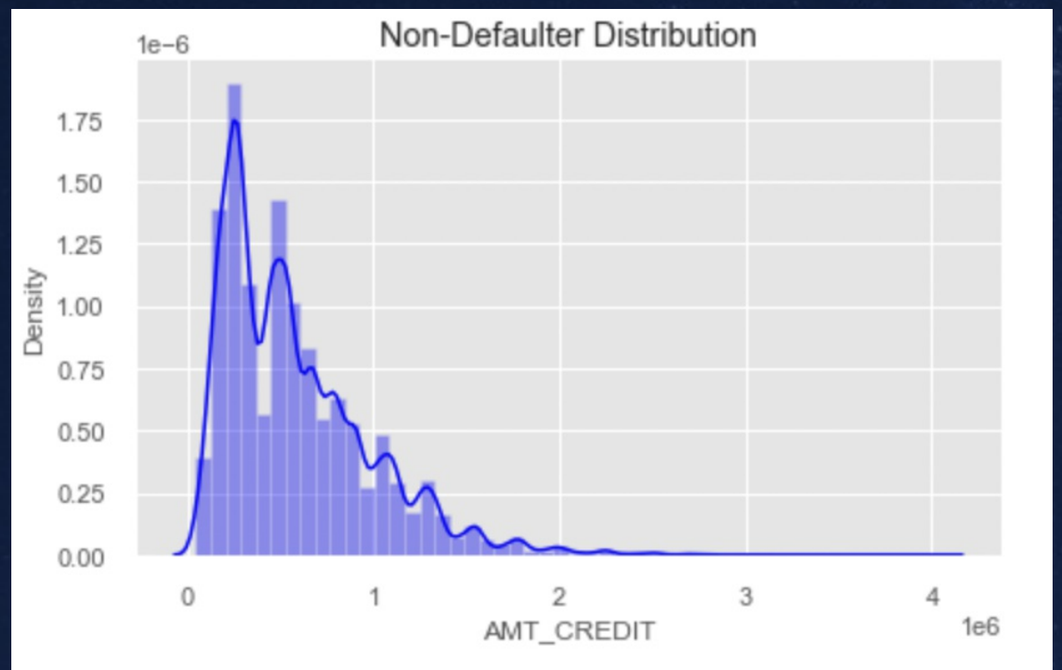


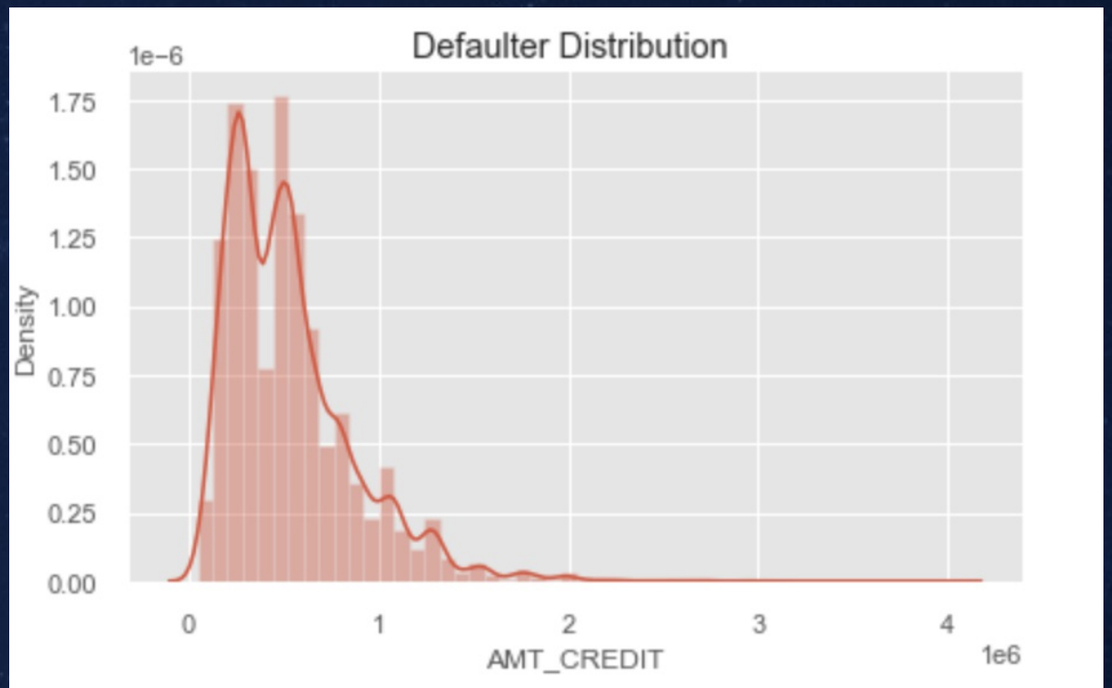


Observation:

- ☆ Mostly Customer Having Less Income Taking The Loan in Both the Cases Non _defaulter and Defaulter
- ☆ From Graph we can easily identify most of Distribution is left Skewed

Numerical: AMT_CREDIT





Observation:

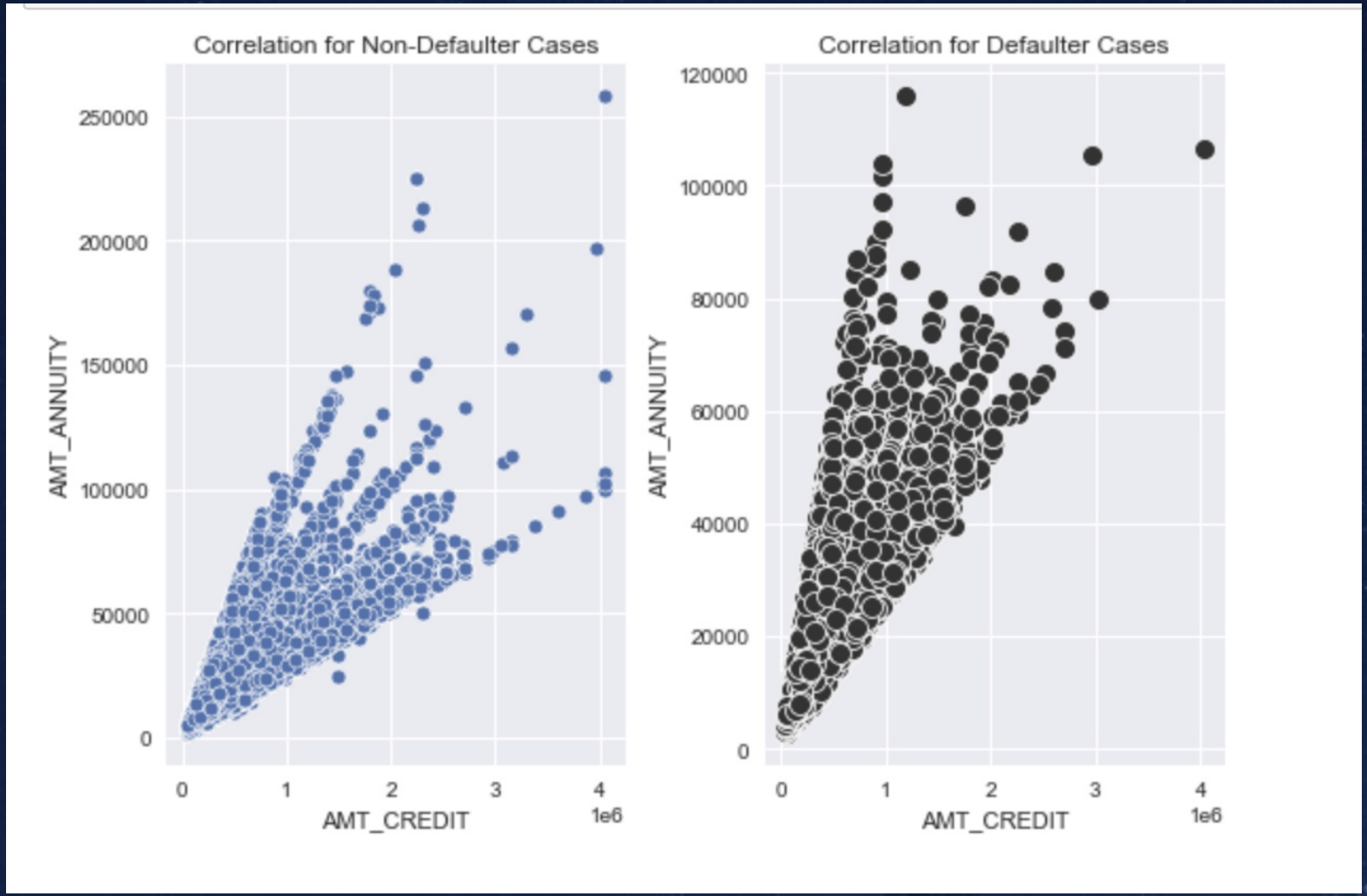
- ☆ Most of Customer taking the loan which are less amount as the credit Amount increases the no of the Customer decreases
- ☆ Some in Case of Defaulter there is up and down in graph but after sometimes its decreases Gradually..

Correlation Between Numerical Variabble

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	EXT_SOURCE_2
AMT_INCOME_TOTAL	1.000000	0.030726	0.035532	0.030075	0.001238
AMT_CREDIT	0.030726	1.000000	0.746683	0.982227	0.125183
AMT_ANNUITY	0.035532	0.746683	1.000000	0.746994	0.115786
AMT_GOODS_PRICE	0.030075	0.982227	0.746994	1.000000	0.134946
EXT_SOURCE_2	0.001238	0.125183	0.115786	0.134946	1.000000

there is high correlation between AMT_GOODS_PRICE and AMT_CREDIT also in AMT_GOODS_PRICE and AMT_ANNUITY As the AMT_GOODS_PRICE increases at same time AMT_CREDIT increases

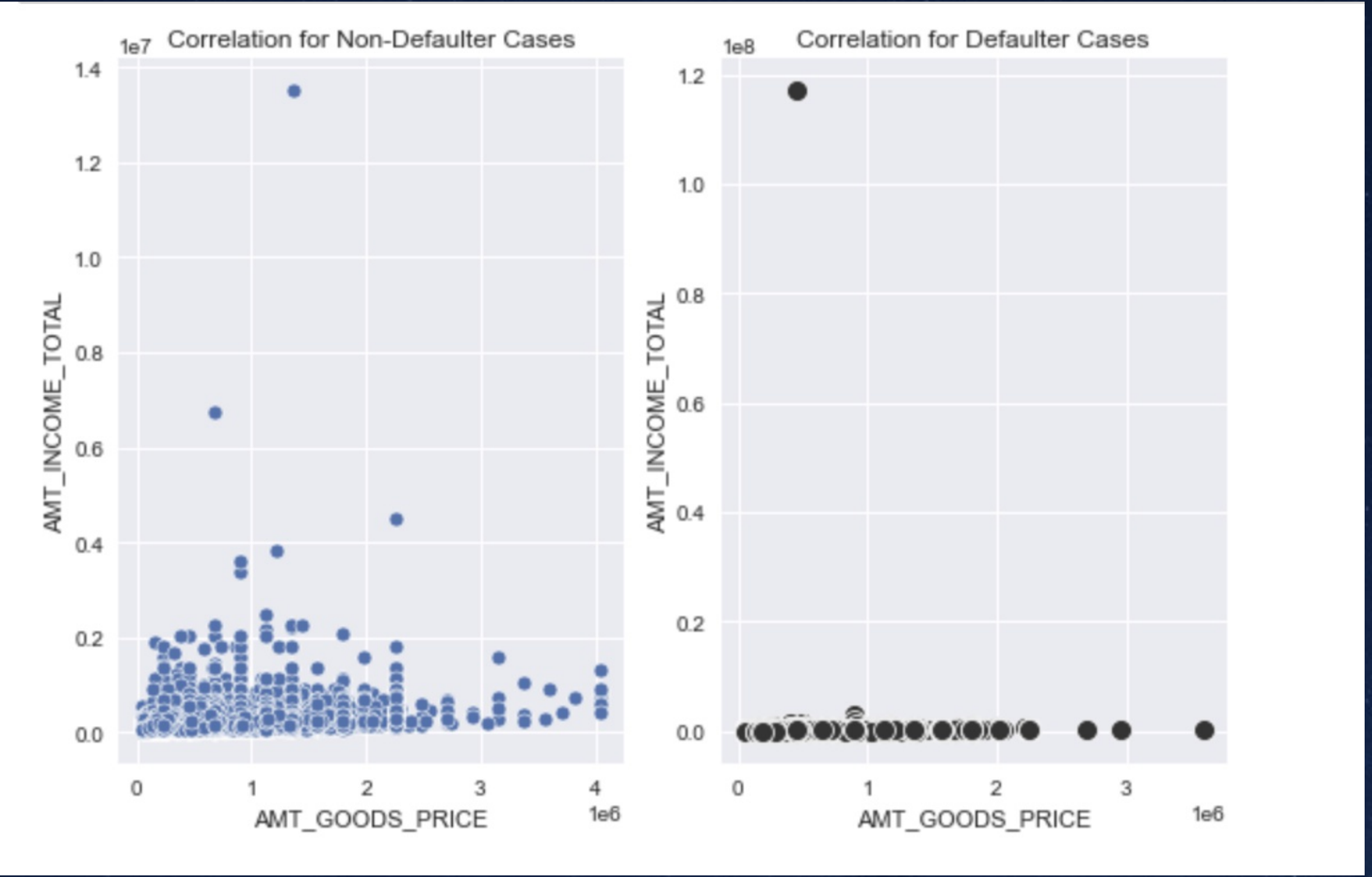
Correlation:AMT_CREDIT and AMT_ANNUITY



Observation:

- From the graph we can say high correlation between AMT_CREDIT and AMT_ANNUITY particular lower amount of loan
- As the AMT_CREDIT increase the AMT_ANNUITY increases in both the Cases Non-defaulter as well as defaulter

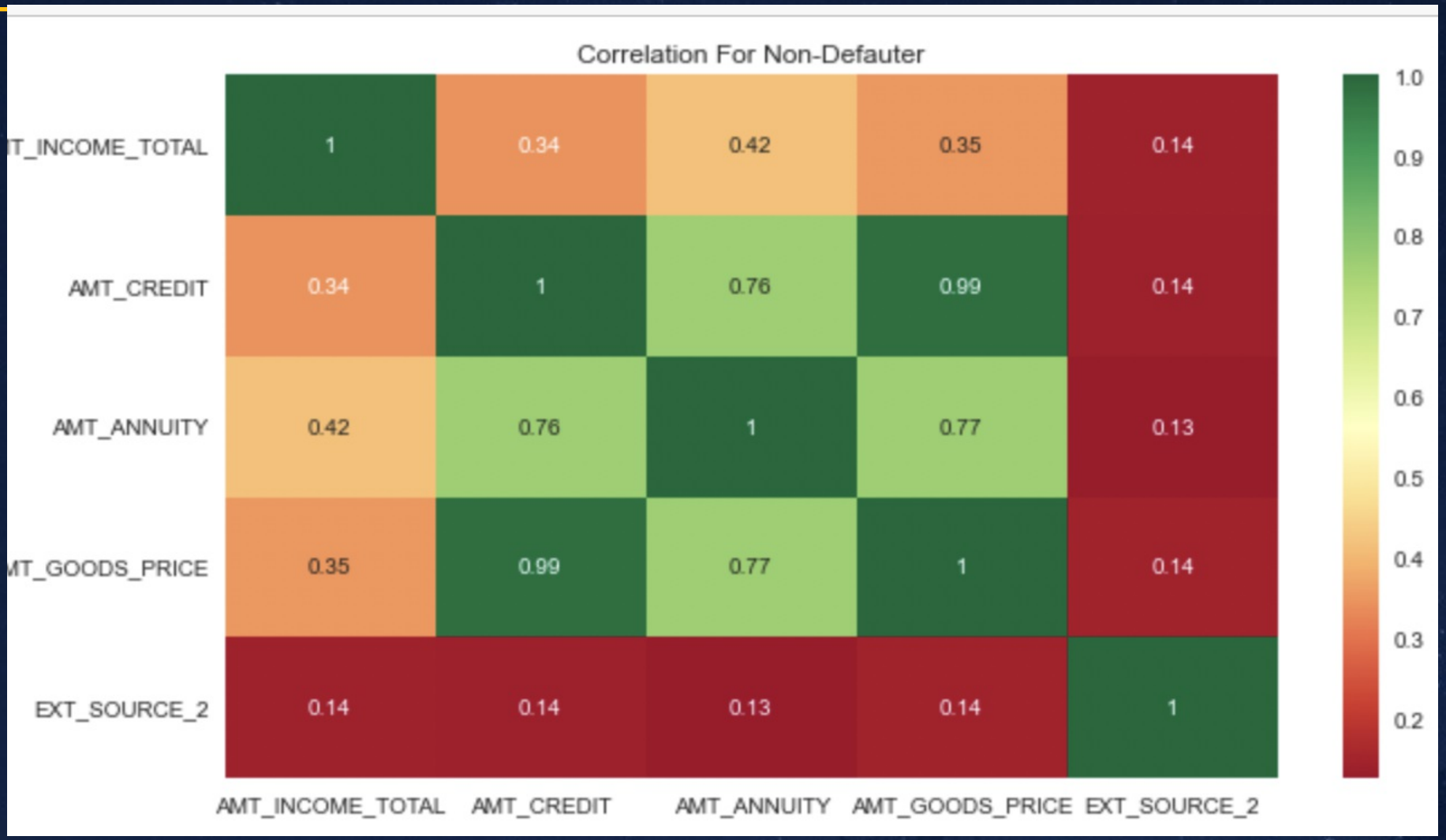
Correlation:AMT_GOODS_PRICE and AMT_INCOME_TOTAL



Observation:

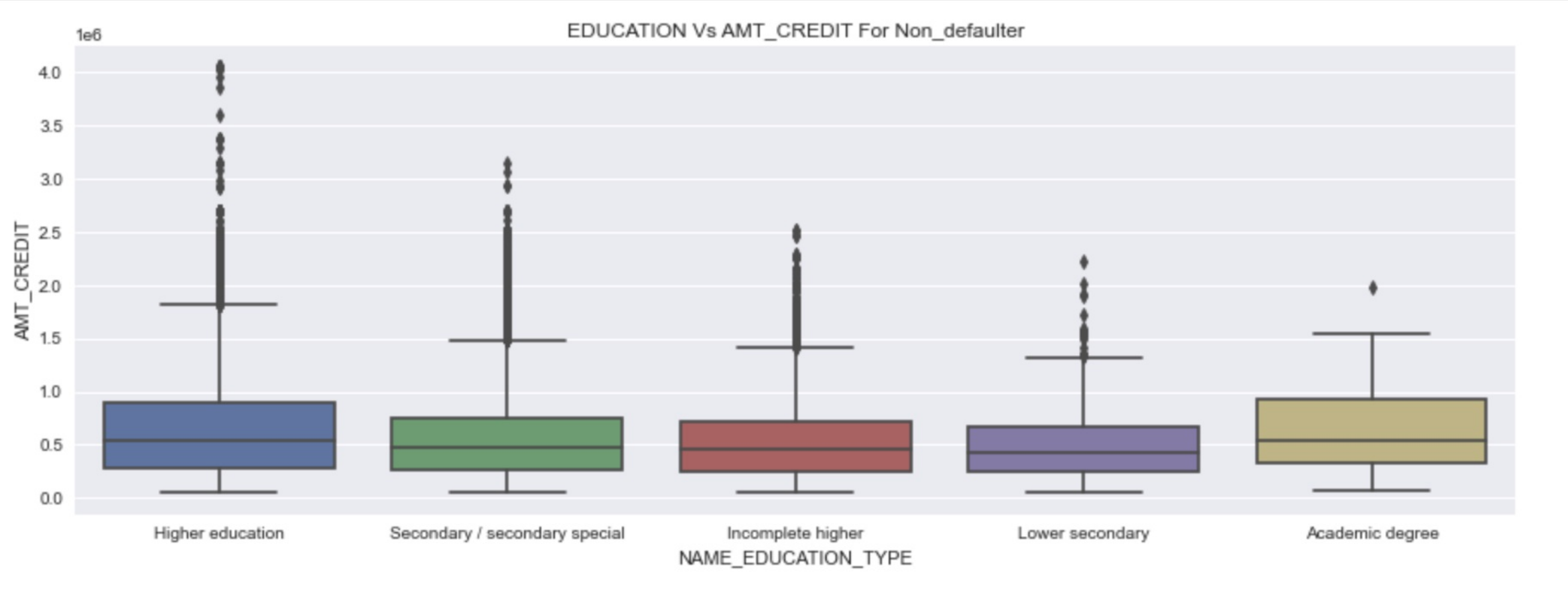
- ☆ There is no such linear relation in AMT_GOODS_PRICE and AMT_INCOME_Totat in both the Cases

Graphical View for Correlation



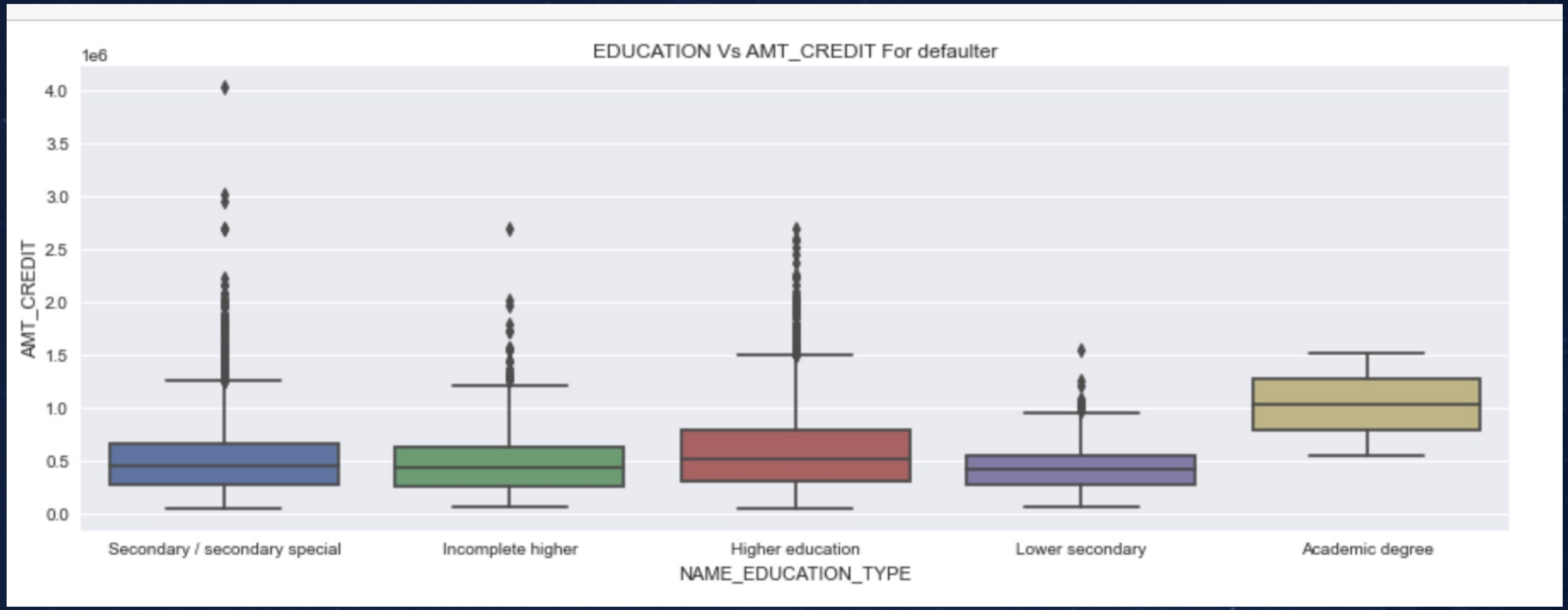

Correlation For Defauter
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
EXT_SOURCE_2
1
0.031
0.036
0.03
0.0012
0.031
1
0.75
0.98
0.13
0.036
0.75
1
0.75
0.12
0.03
0.98
0.75
1
0.13
0.0012
0.13
0.12
0.13
1
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
EXT_SOURCE_2
1.0
0.8
0.6
0.4
0.2

Bivariate /Multivariate Analysis
NAME_EDUCATION_TYPE vs AMT_CREDIT



Max value for Higher Education is high and median value almost Same for all Category

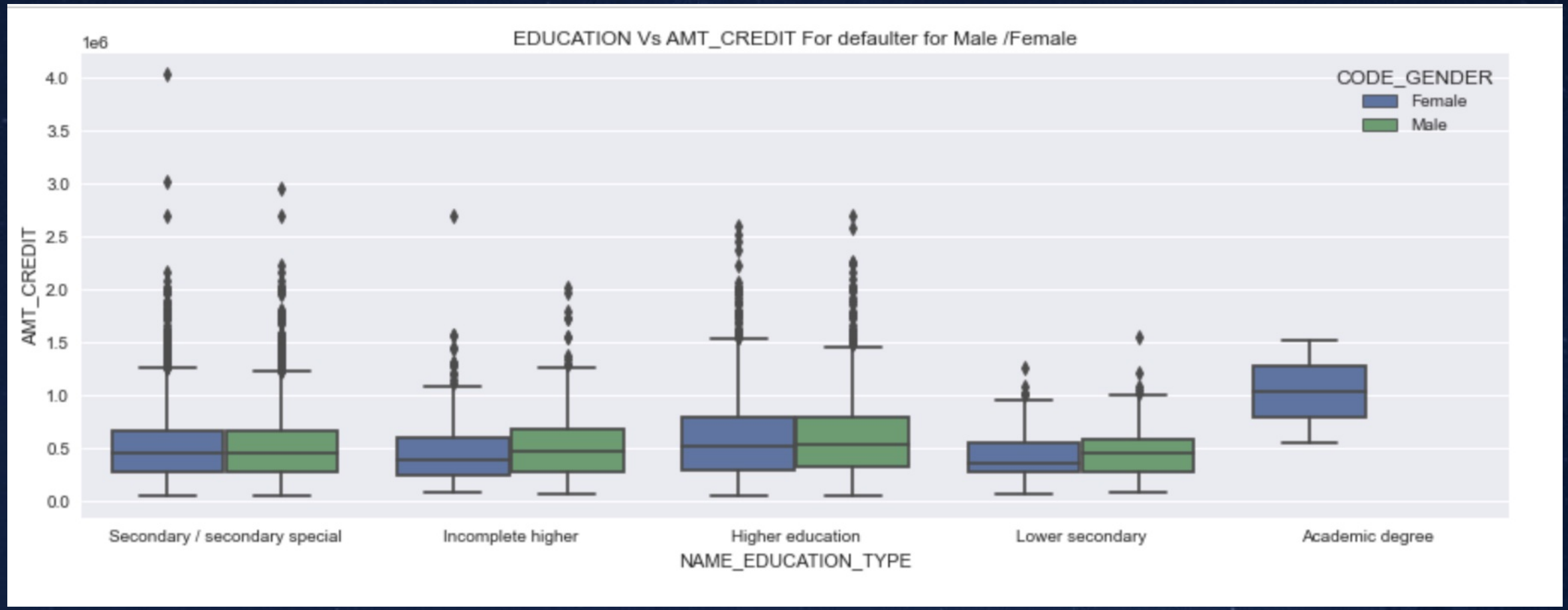
NAME_EDUCATION_TYPE vs AMT_CREDIT



Observation:

- ☆ Here what we see maximum value of AMT_CREDIT for Academy Degree
- ☆ Median value for Academy degree is high
- ☆ Mostly customer having lower Degree have more Probabilty to do Default

NAME_EDUCATION_TYPE vs AMT_CREDIT



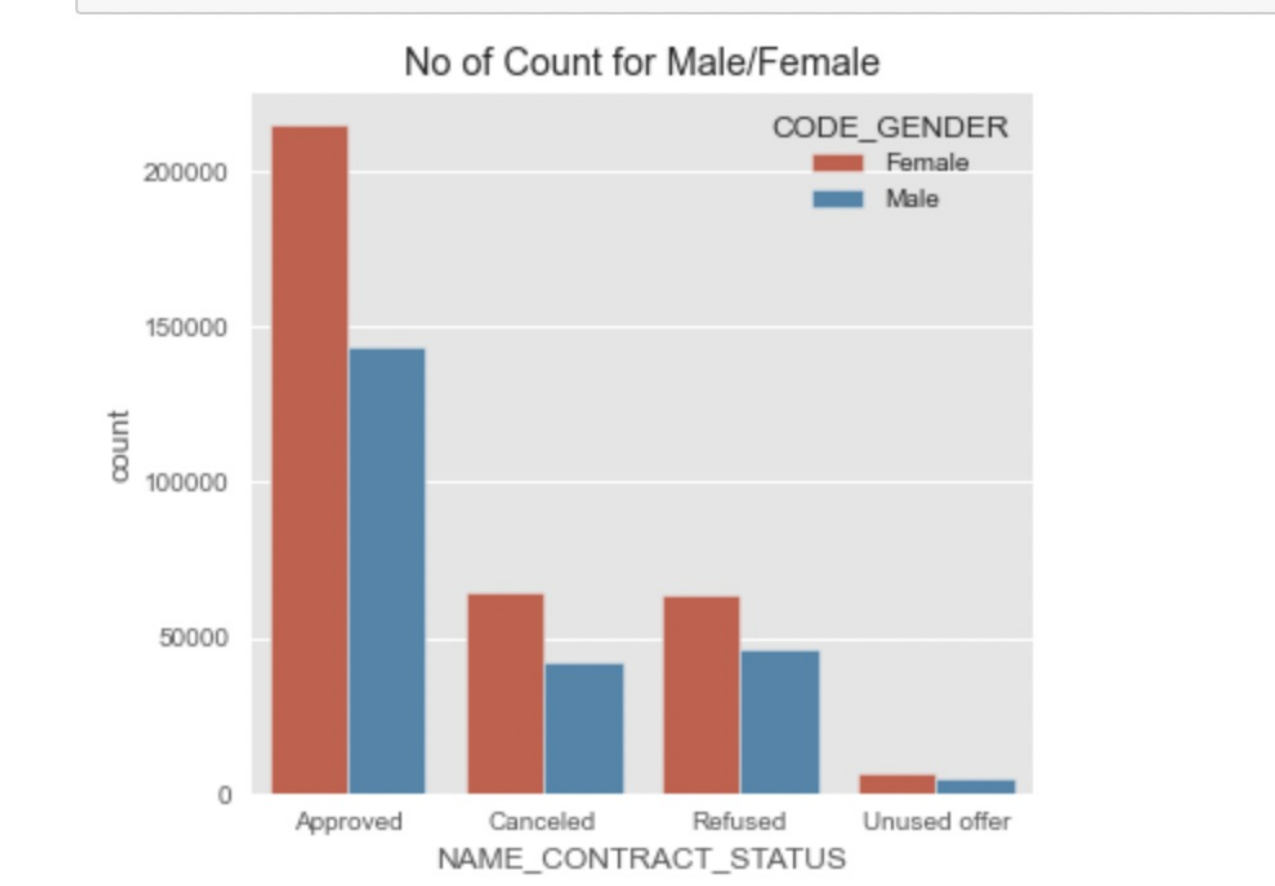
Observation:

- ☆ Median value of the customer having Academy Degree is very high and belong to female for those customer they taking the loan amount is high.
- ☆ Median amt_credit for male and female for the category of secondry/secondary special is almost same

Merging the Previous Application with New application

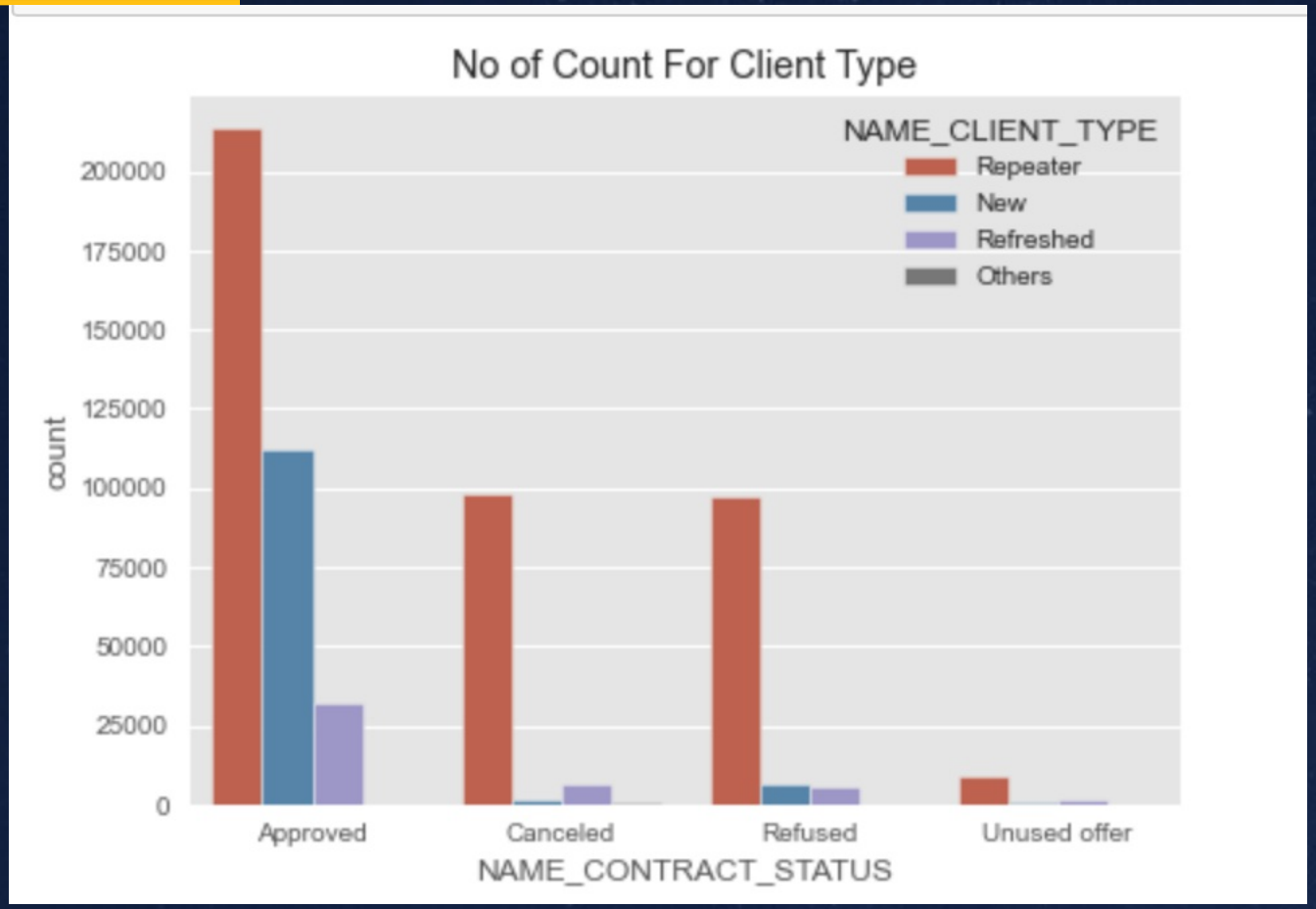
- **When a client applies for a loan, there are four types of decisions that could be taken by the client/company):**
- Approved: The Company has approved loan Application
- Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan 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 on different stages of the process.

Analysis:



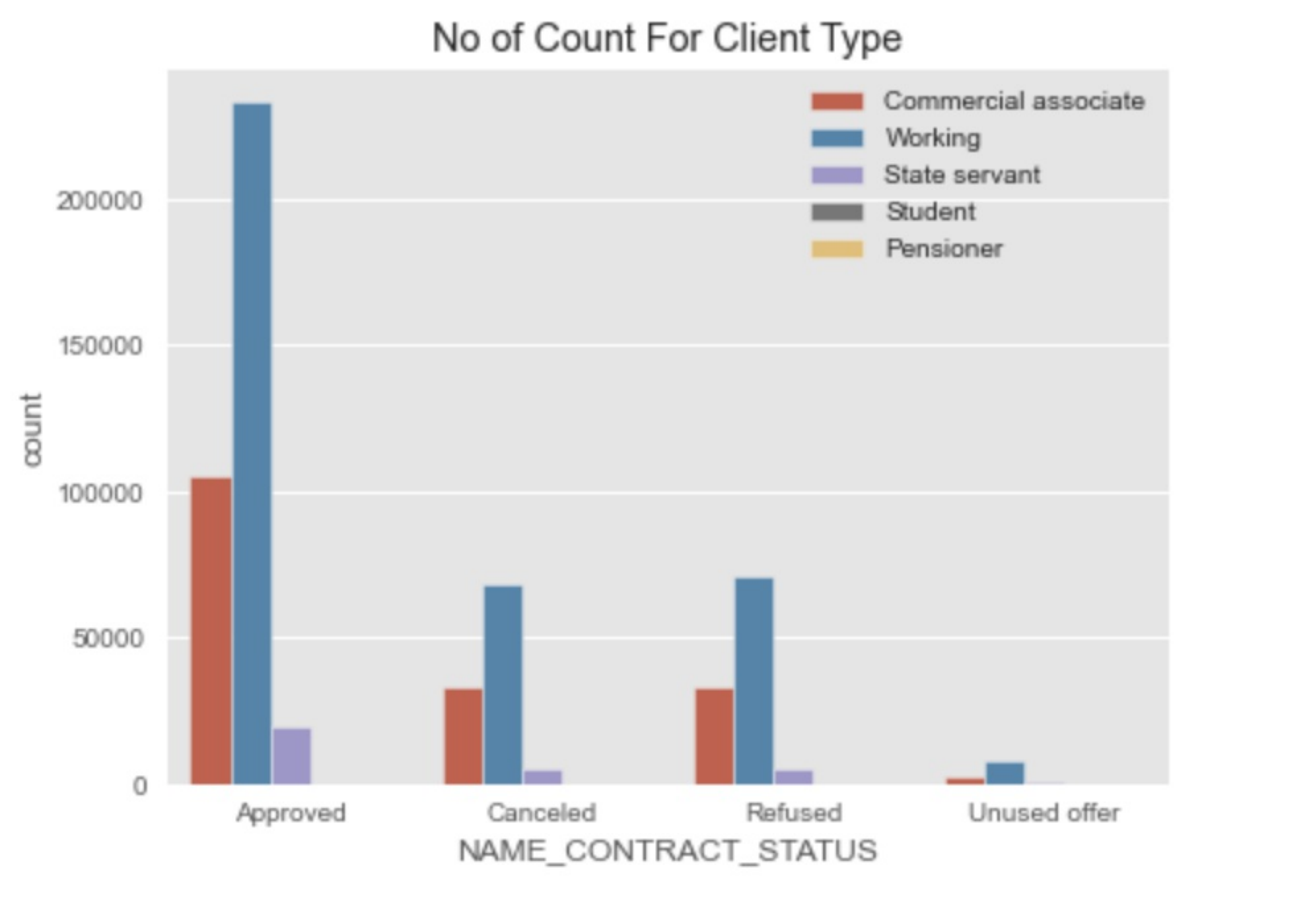
Observation:
No of Approval/Cancelled/Refused/ Unused offer for Female is greater than male

NAME_CLIENT_TYPE



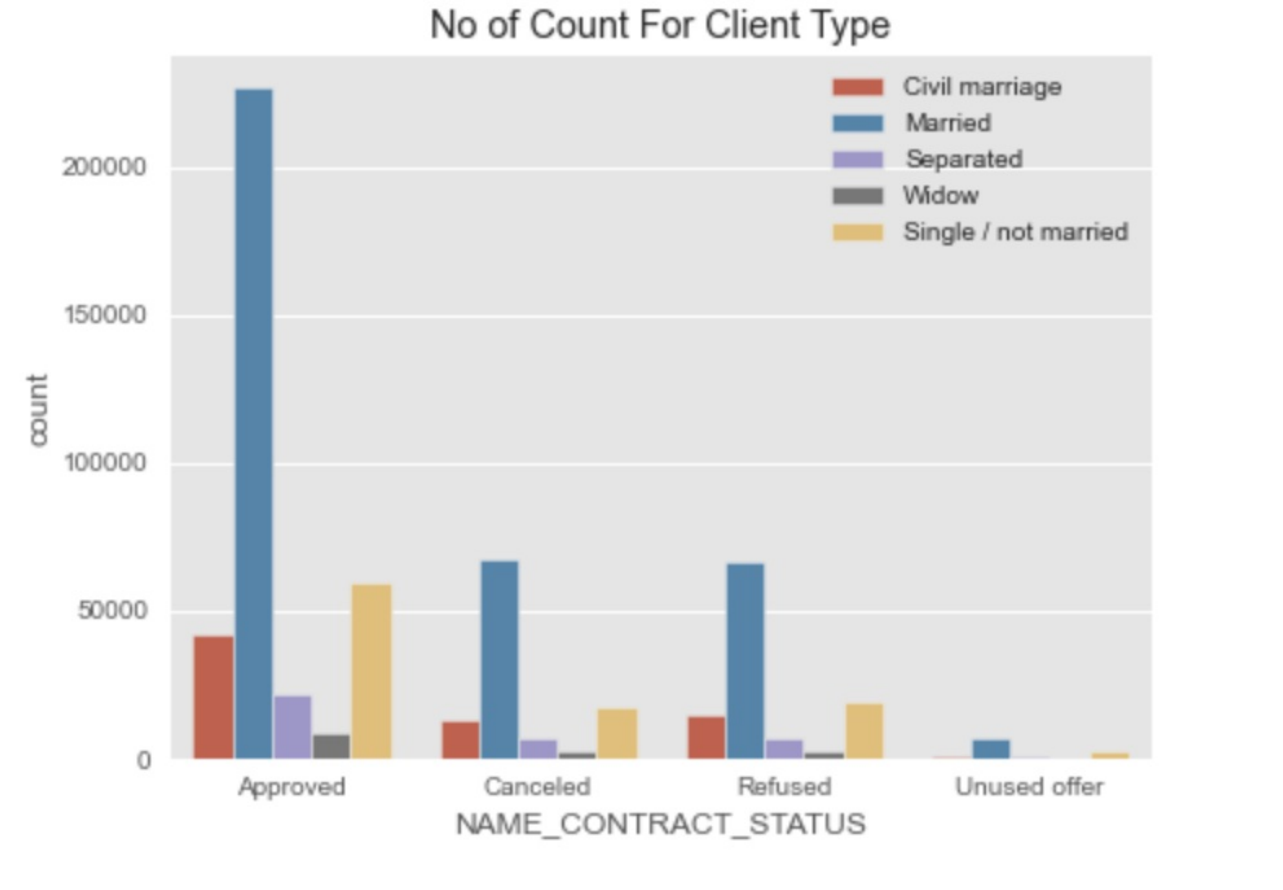
Observation:
Most of Client who are old approved the most and new client cancelled count are less

NAME_INCOME_TYPE



Observation:
Working Customer having approved number is greater than other profession
ALso the concelled count for working is greater

NAME_FAMILY_STATUS



Obseravtion:
For Married the Approved Count is Greater and Widow approved count is less

INSIGHTS OF THIS CASE STUDY

Insight of Case Study:

Conclusion: The people Who less likely to default

- Client Who working as state Govt
- Old People of any income Group
- Any Client who previous Loan was approved
- Customer having high income Group
- Female Customer Who are Older