EDA – CREDIT ASSIGNMENT

NAME : PRATEEK JAIN
BATCH ID: 1975

Data Science Program - March 2022

INTRODUCTION

Given dataset of loan providing company	3 .csv files as dataset	1. Application 2. Previous Application 3. Column description

About Dataset

Application

- Contain information at the time of applying
  - **The client with payment difficulties:**
  - **All other cases:**

Previous Application

- Contain information about four types of decisions that could be taken
  - **Approved**
  - **Cancelled**
  - **Refused**
  - **Unused offer**

Column Description

- Contains information
  - Description/meaning of the columns so as to get better understanding of dataset

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile

Objective

Understanding

- How consumer attributes and loan attributes influence the tendency of default.

Identifying

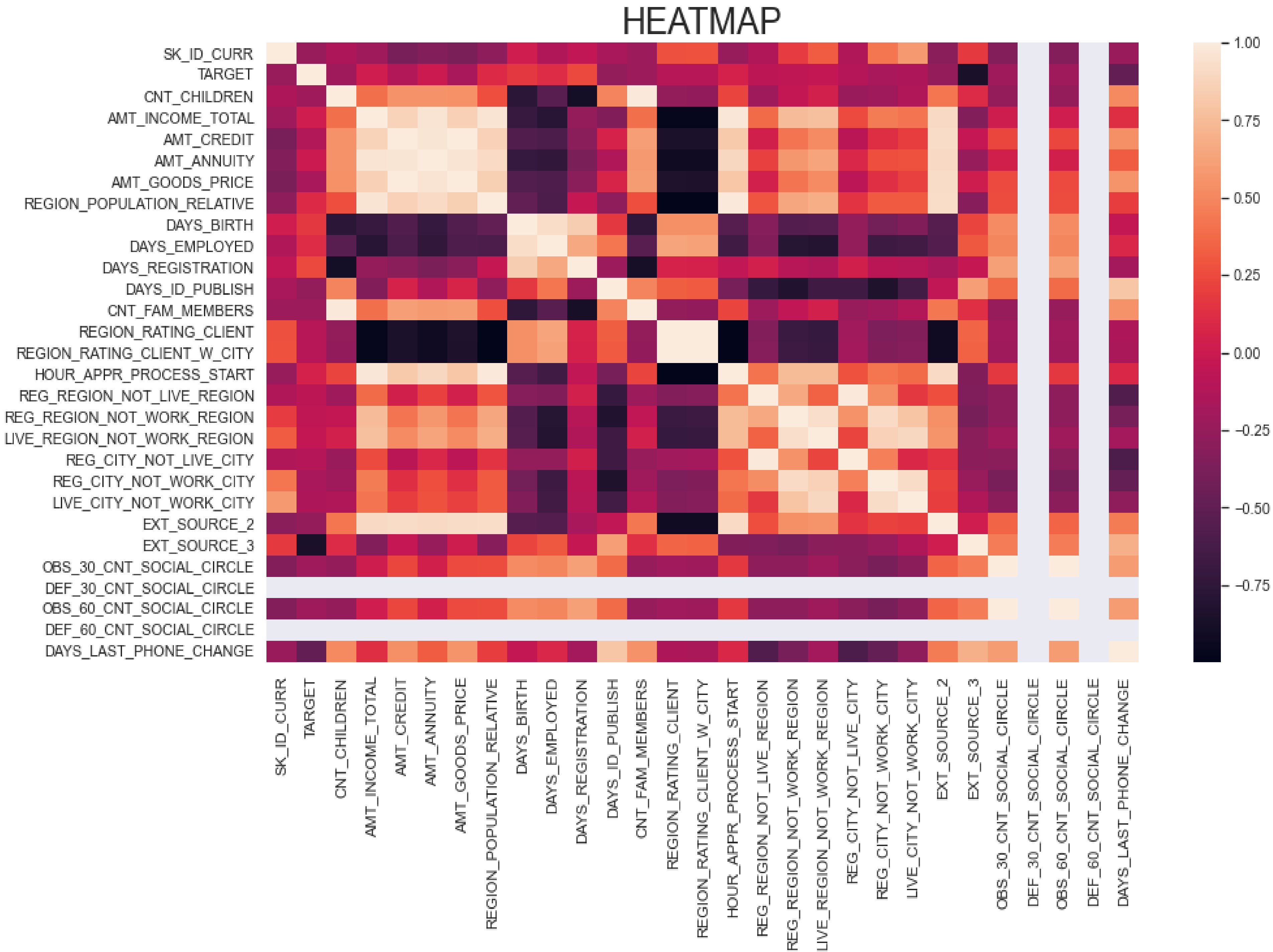
- Patterns indicating if a client has difficulty paying their installments
- Top 10 correlations

Predicting

- Consumers capable of repaying the loan are not rejected.
- Understand the driving factors (or driver variables) behind loan default

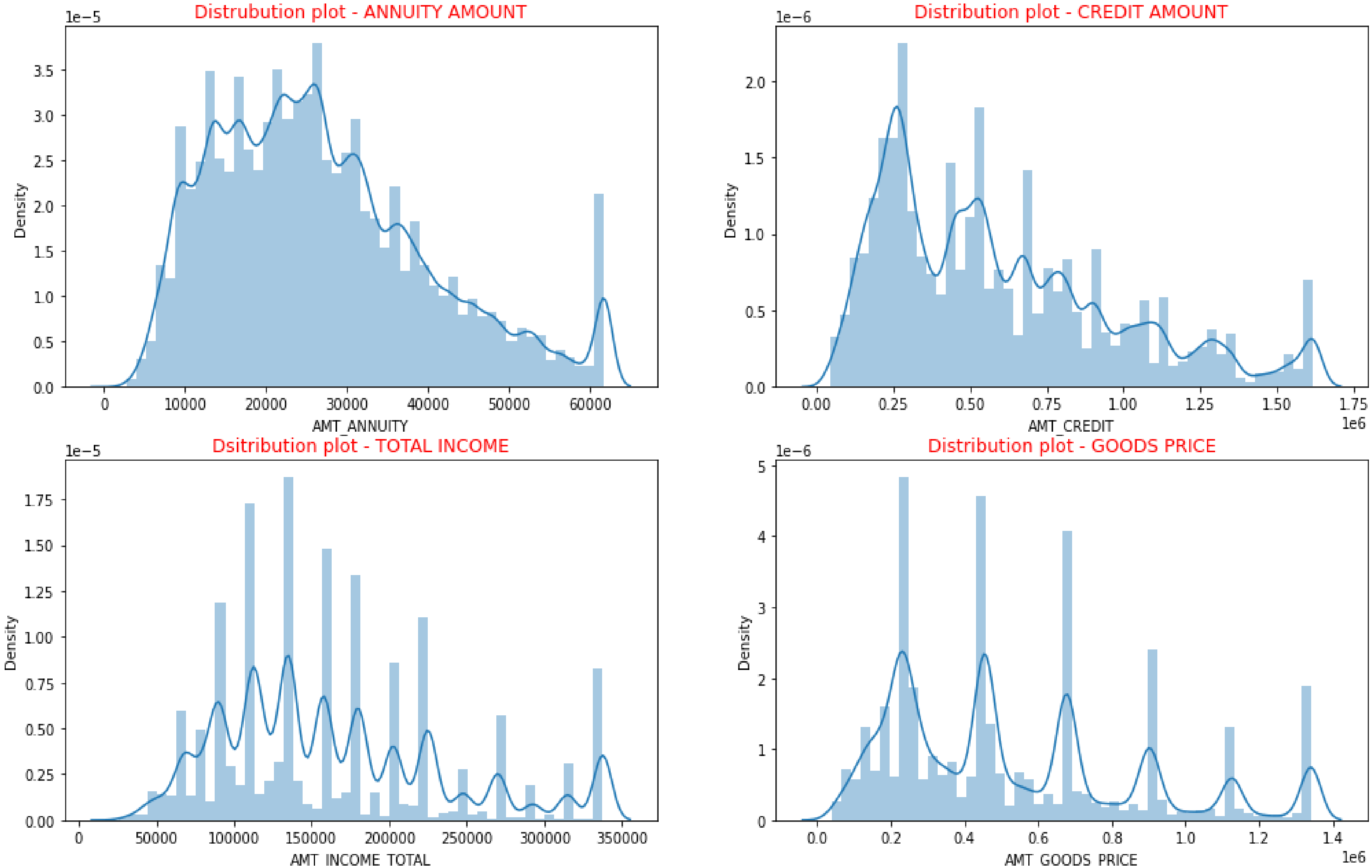
APPLICATION DATAFRAME

1. • Importing Libraries
2. • Reading the data set and finding percentage of null values
3. • Dropping columns with missing values >45%
4. • Identifying continuous and categorical columns/variable
5. • Continuous column - Columns containing unique values > 58
6. • Categorical column – Columns containing unique values < 58
7. • Imputation for missing value<45 (categorical – mode, continuous – median)
8. • Dropping unnecessary columns
9. • Detecting Outliers – Using Subplots
10. • Handling outliers by flooring and capping


HEATMAP
SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

UNIVARIATE ANALYSIS

SUBPLOTS



Most of the people taking loans have annuity between 20000 and 30000

DEFAULTER AND NON DEFAULTERS

APPLICATION LOAN DATASET – TWO MAIN VARIABLES

- DEFAULTER
- NON DEFAULTER

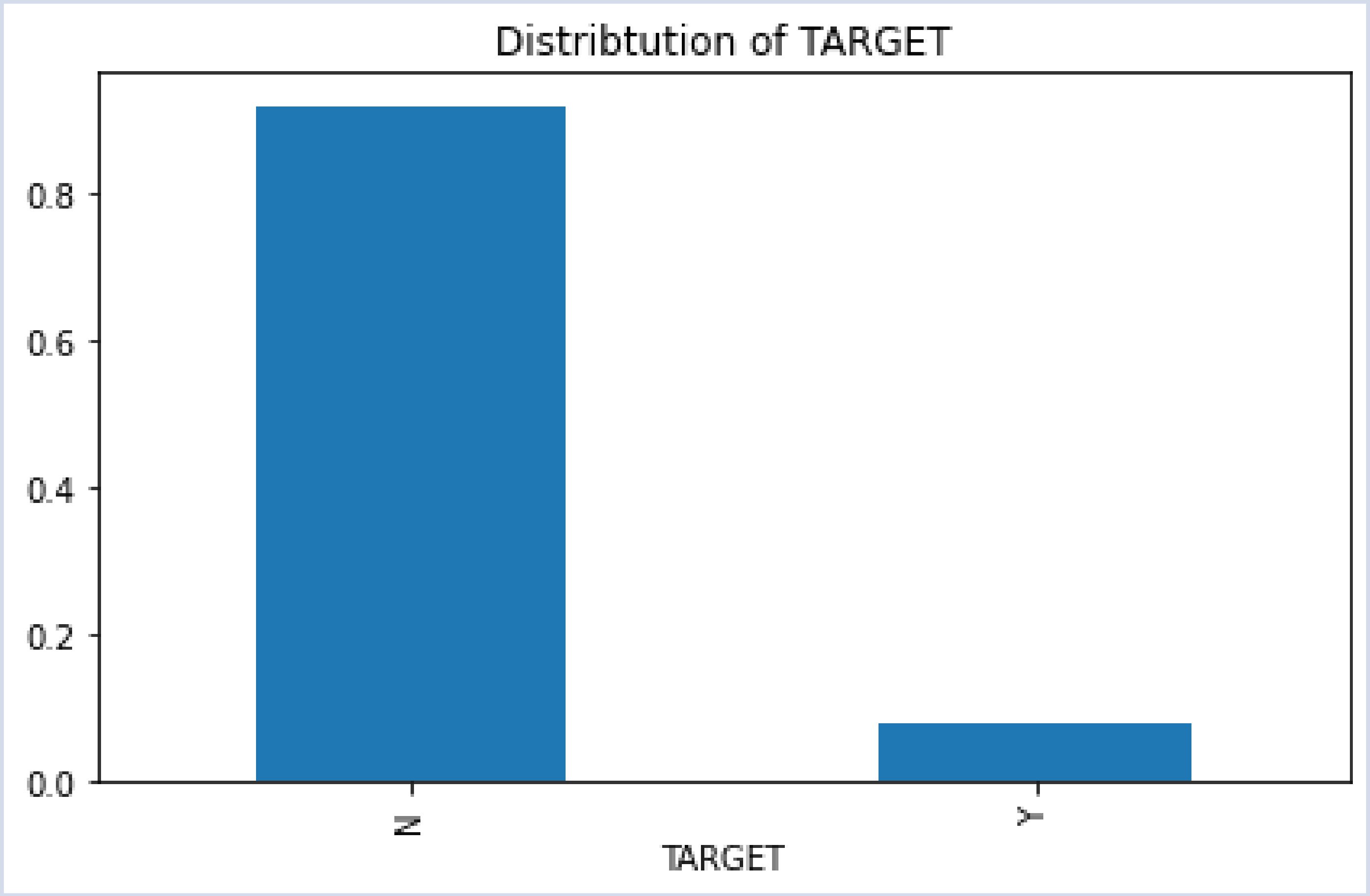
DEFAULTERS

- The client with payment difficulties
- Response is stored as 1 (‘Y’- Yes)

NON DEFAULTERS

- **All other cases**
- Response is stored as 0 (‘N’- No)

BARPLOT – DEFAULTER AND NON DEFAULTERS



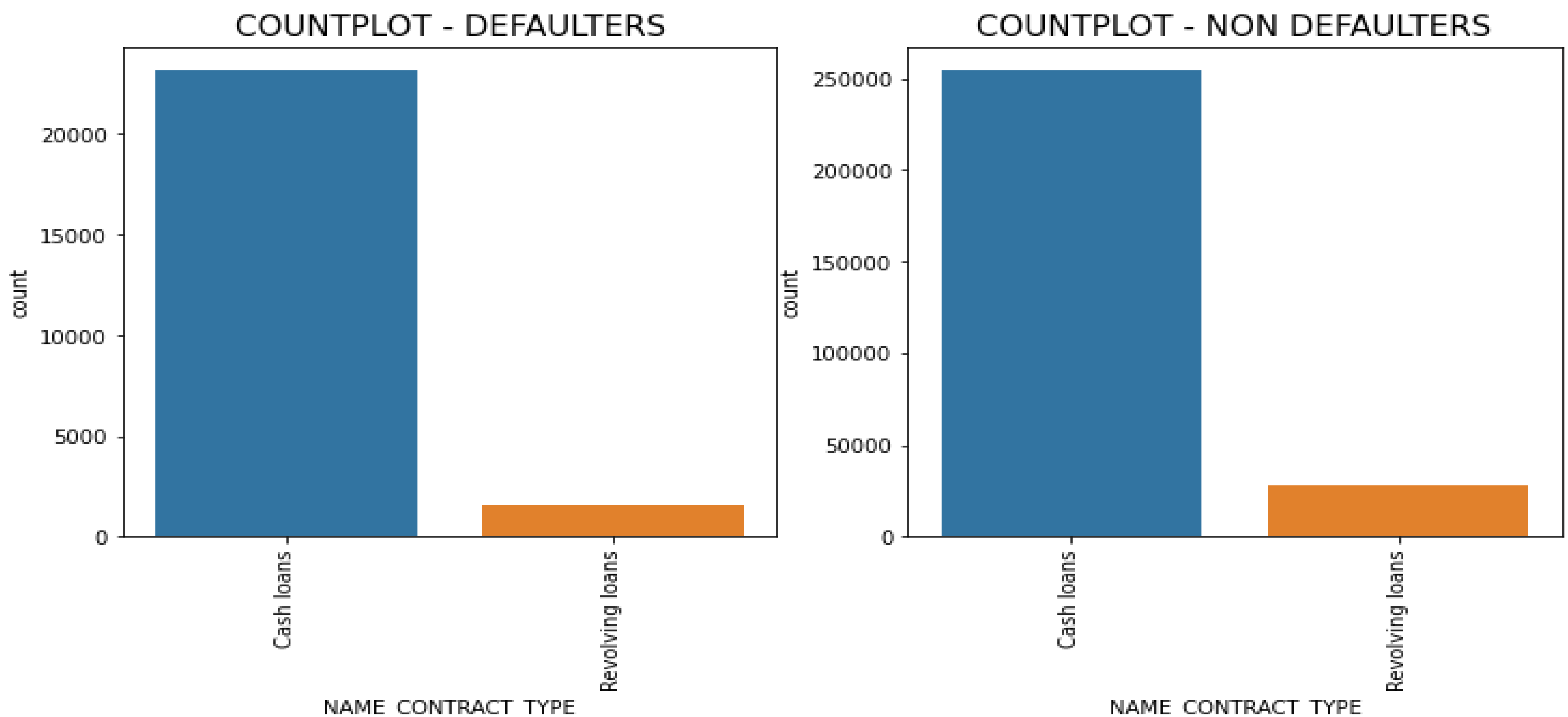
- **Large number of people applying for loans are non defaulters**

Making two Dataframes

1. Defaulters (TARGET = 0)
2. Non- Defaulters (TARGET = 1)

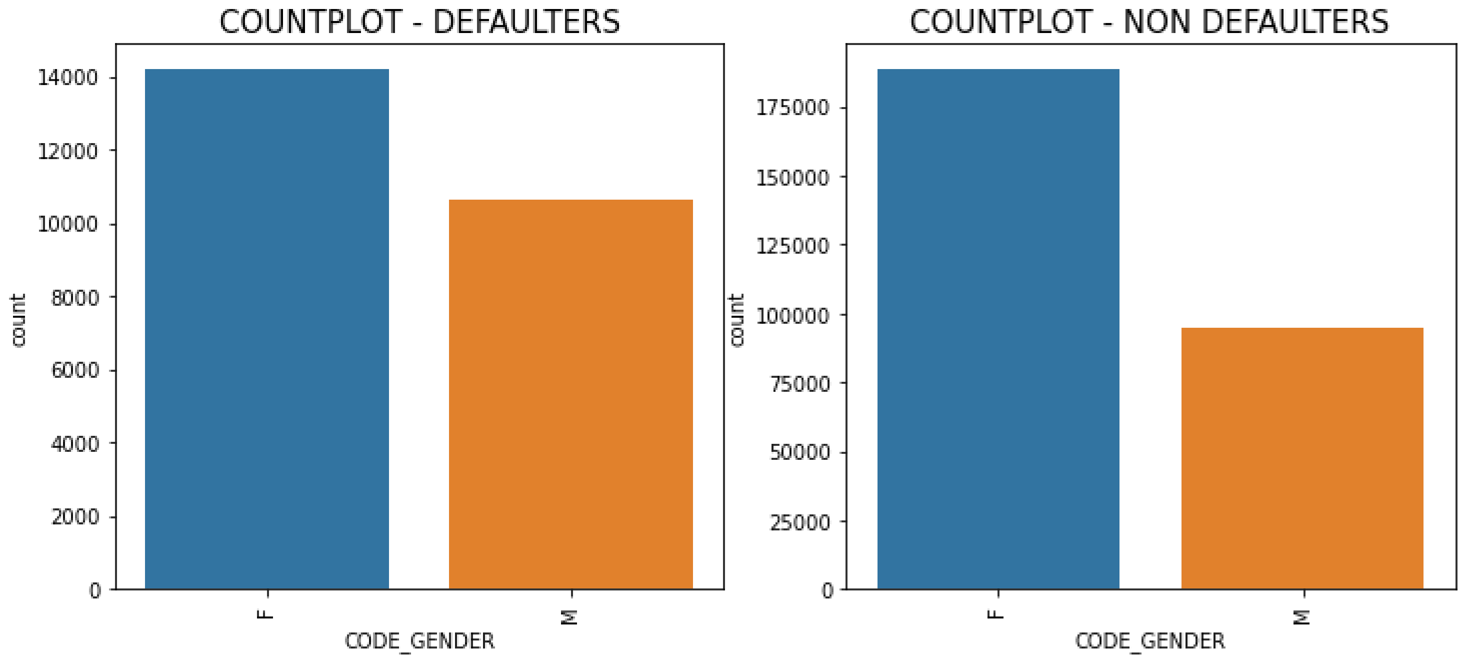
Univariate analysis for each Dataframes using subplots

COUNTPLOT – LOAN TYPE



Number of people whether defaulters or non defaulters prefer to take 'CASH LOAN' as compared to 'REVOLVING LOAN'

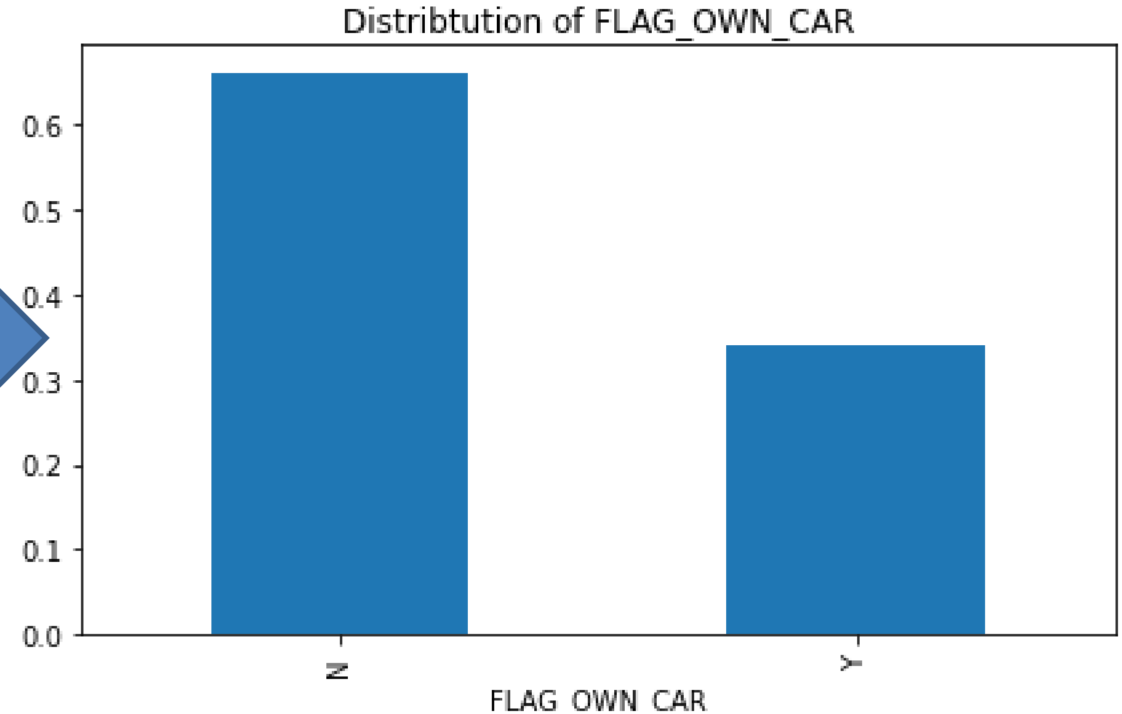
COUNTPLOT - GENDER



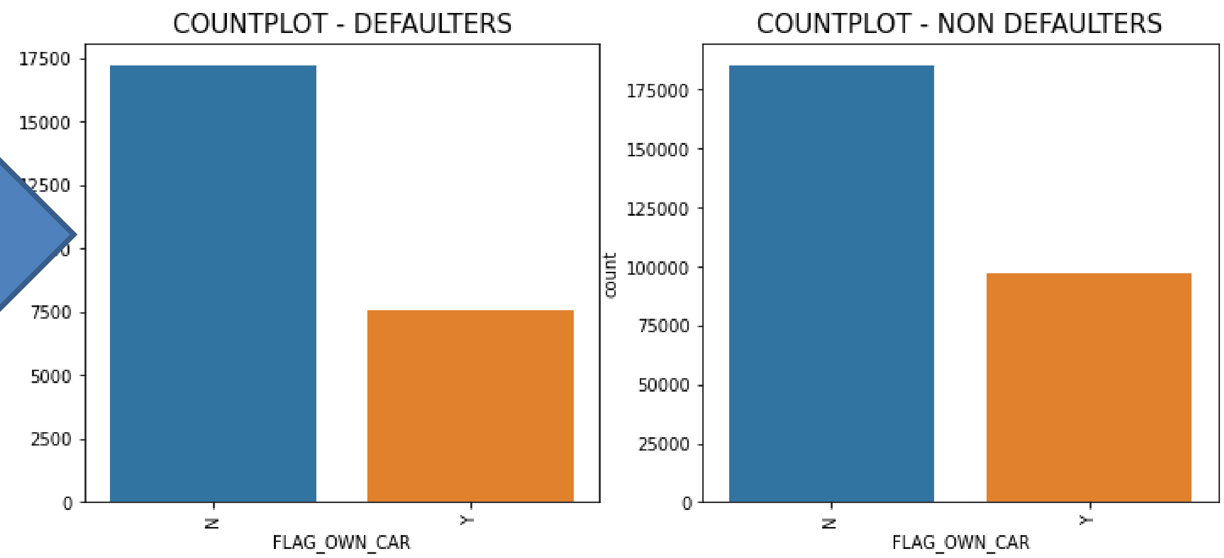
Number of "FEMALE" taking loans is much higher than the number of "MALE" for both the target variables

COUNTPLOT - CAR

People who “DON’T OWN CAR” are applying more for loan

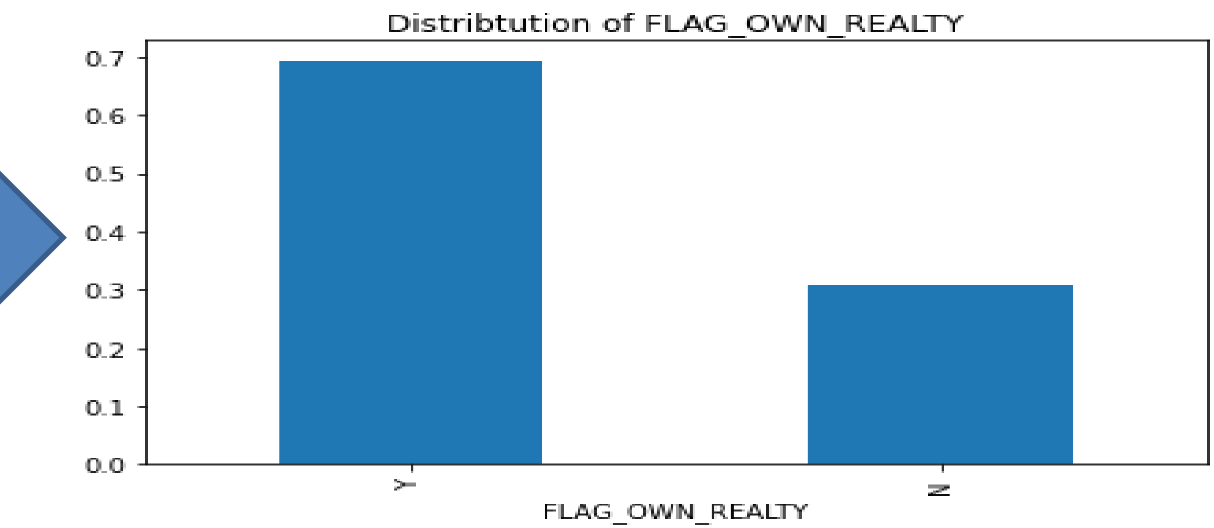


Most people applying for the loan and have car are "NON DEFUALTER"

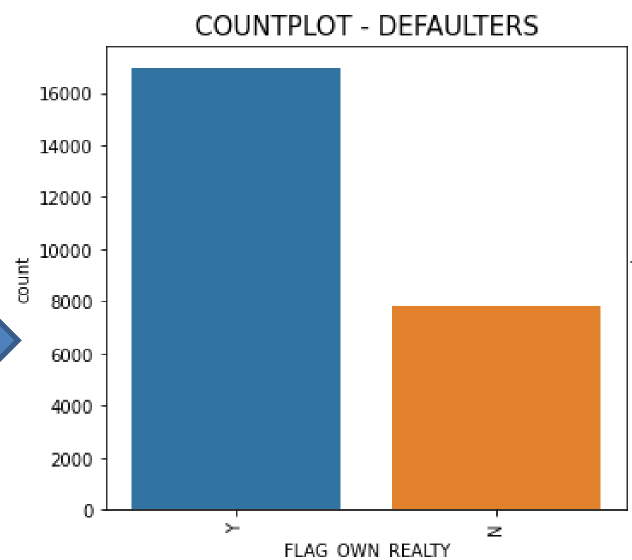


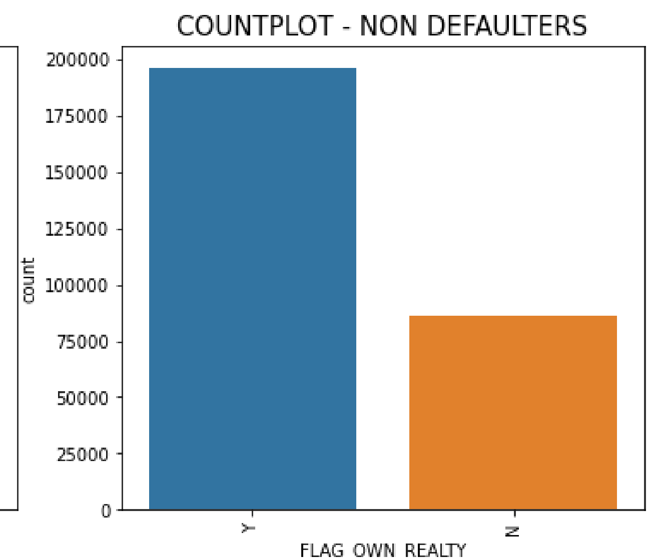
COUNTPLOT - HOUSE

Most people applying for the loan have own house

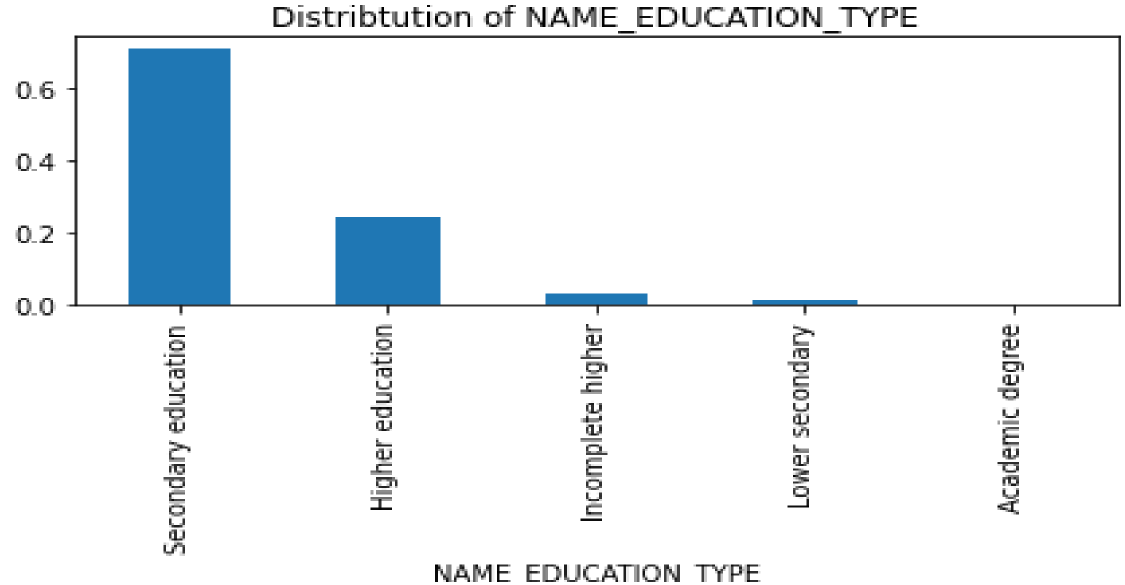


People having 'OWN HOUSE' and non defaulters are more



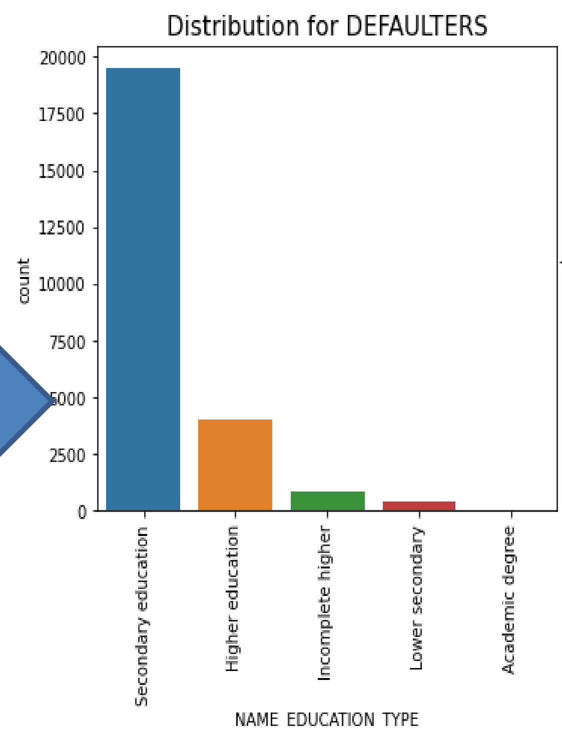


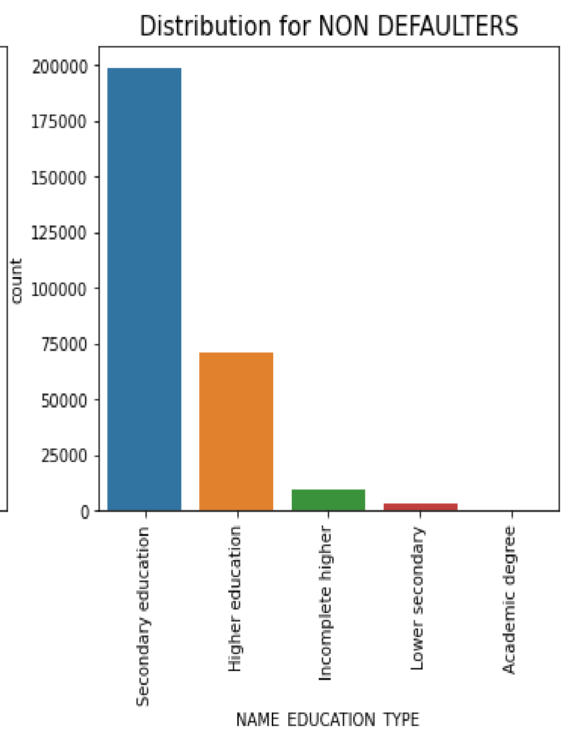
People applying for the loan, mostly have secondary education



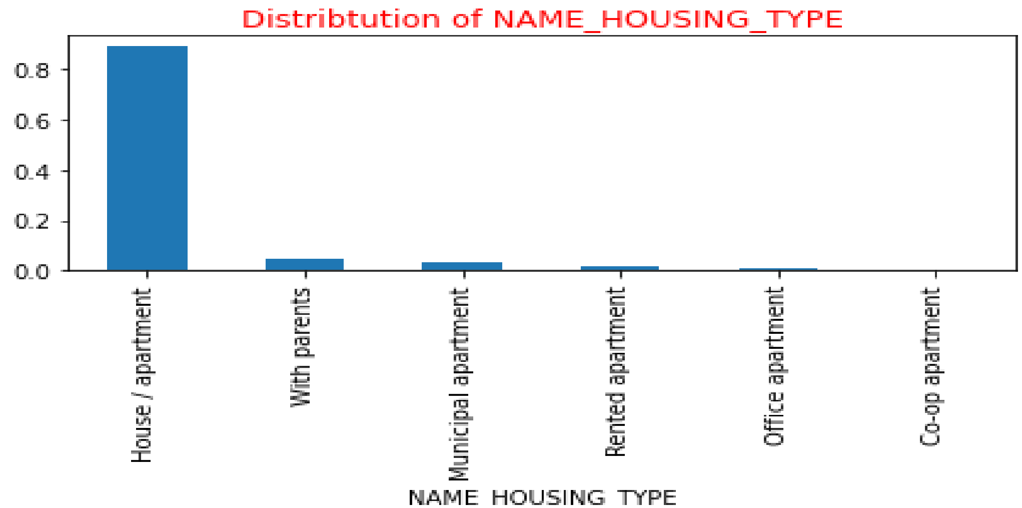
•People having secondary & higher education have less payment difficulties

•People have academic degrees can also be targeted as they apply less for the loans



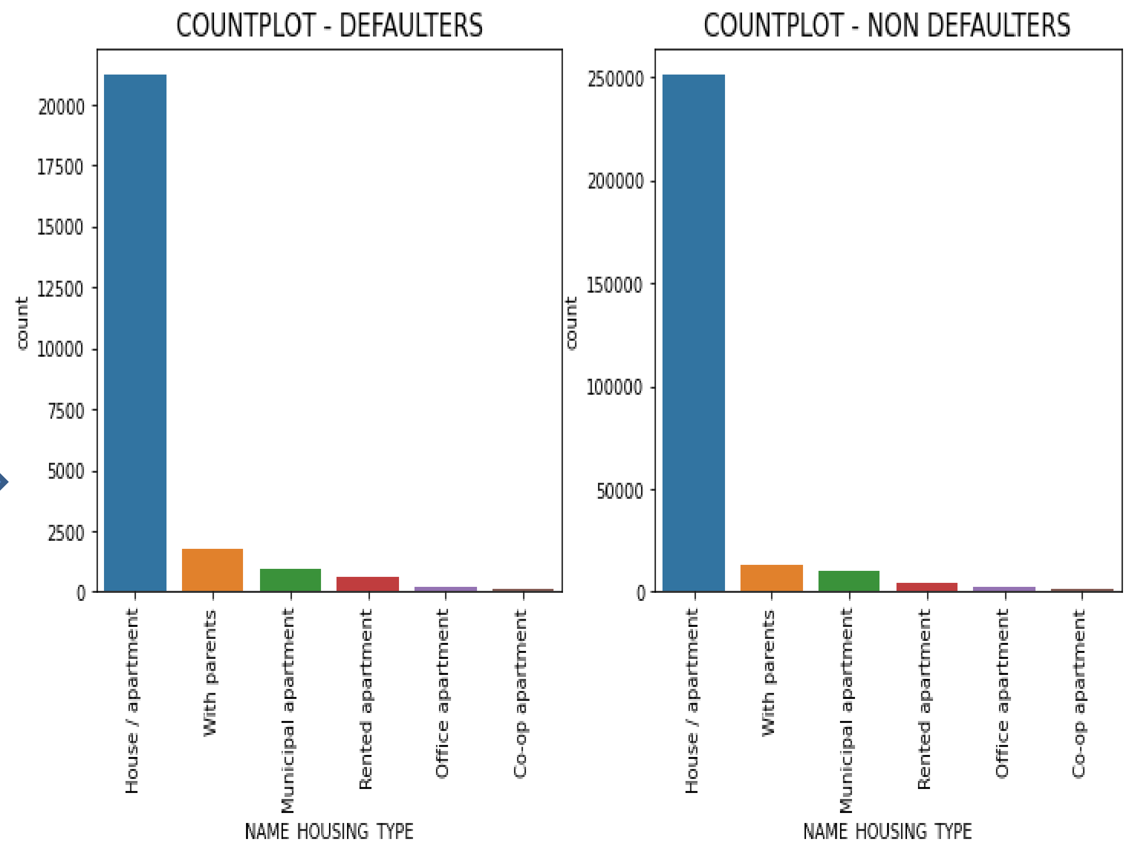


People applying for the loan are mostly living in a house/apartment

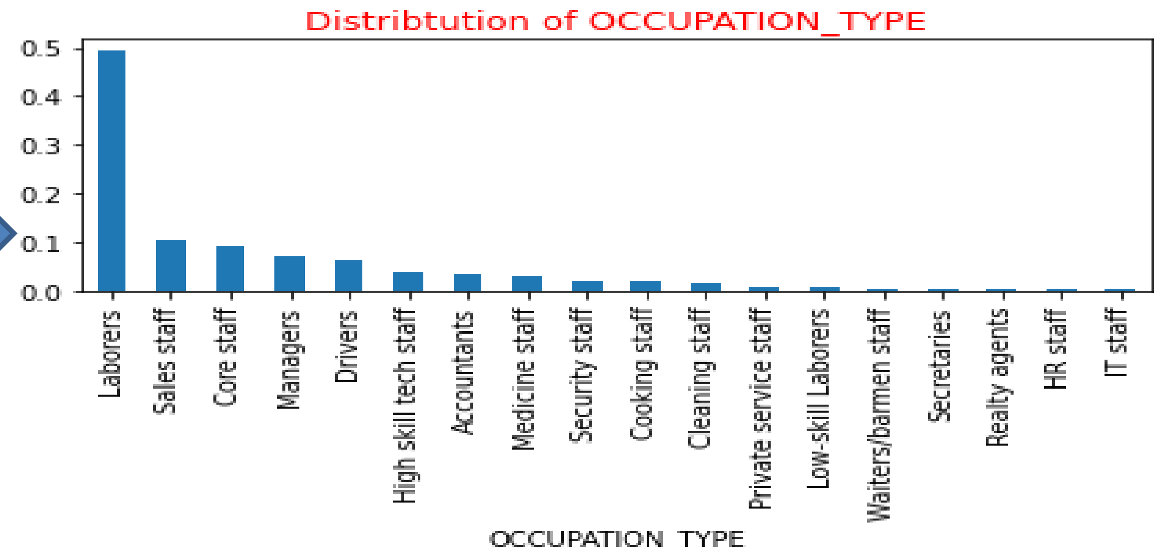


- Large number of people applying for loans have there own 'HOUSE/APARTMENT'

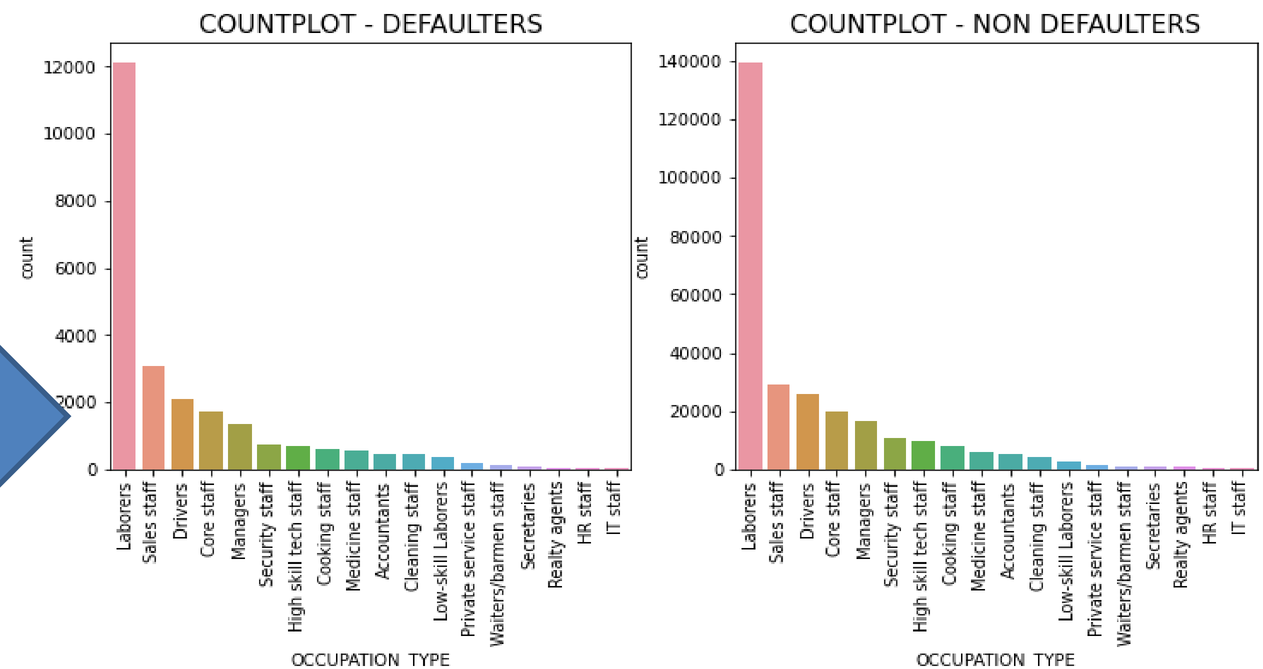
- People living 'WITH PARENTS' are more likely to have payment difficulties.



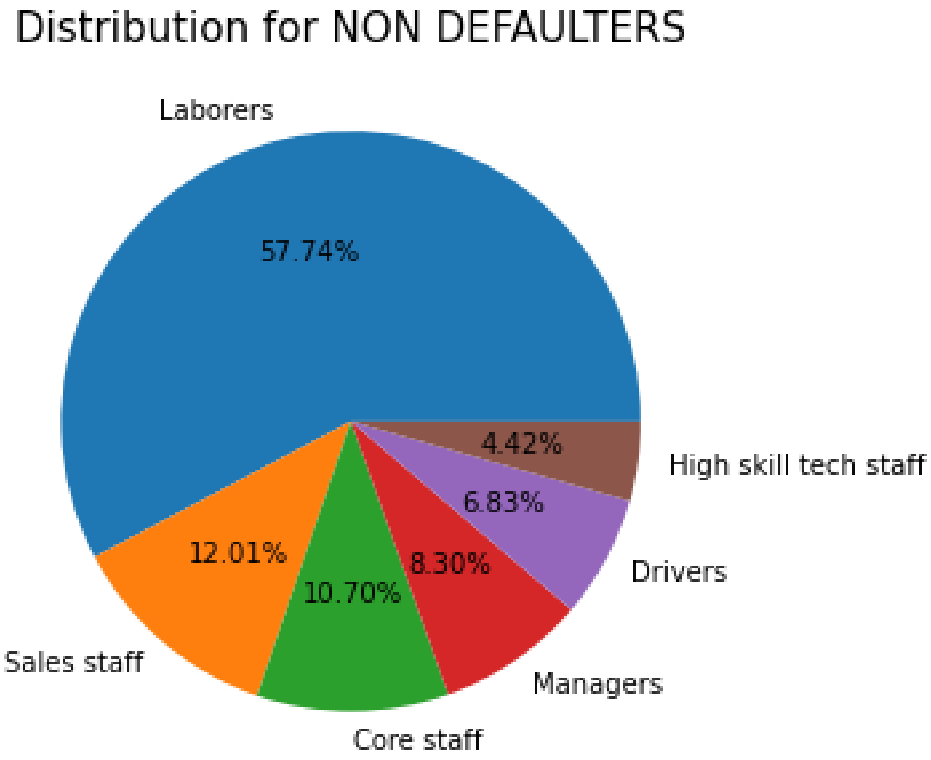
Labourers are applying more for loan as compared to other category

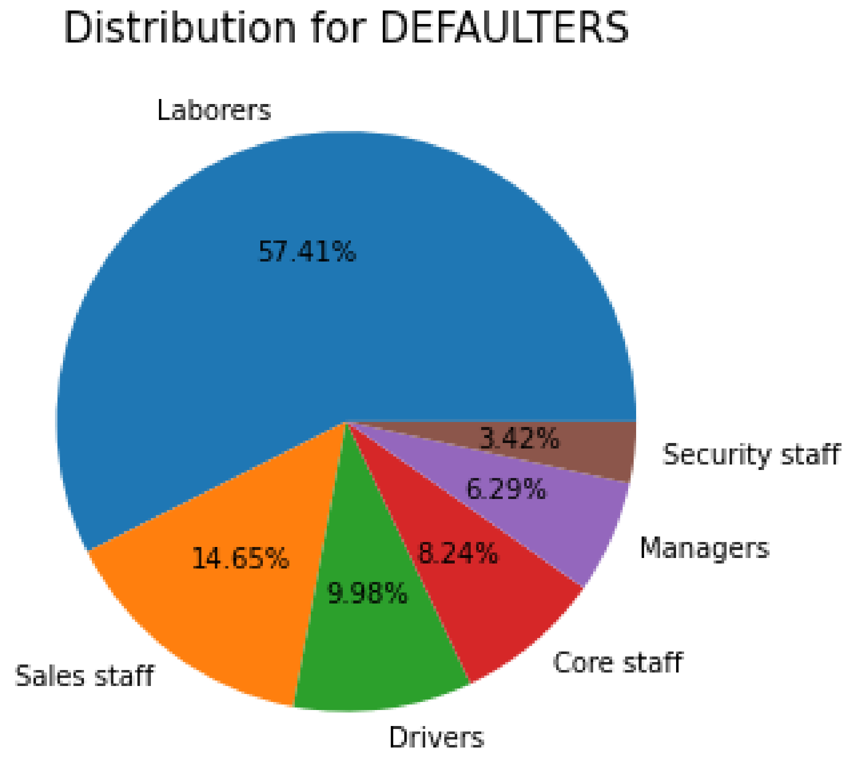


Labourers are the maximum number of defaulters and non defaulters



From the previous graph we took only 5-6 categories so as to more insight





INFERENCES FROM PREIVOIUS SLIDE

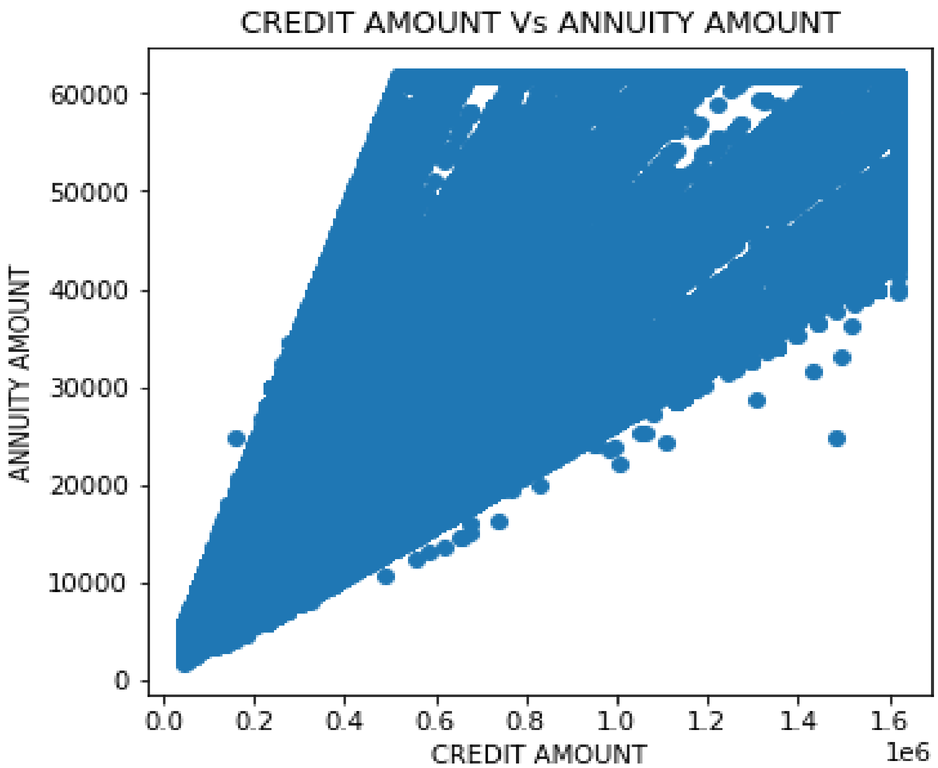
accept the applications of Non Defaulter where % is higher.

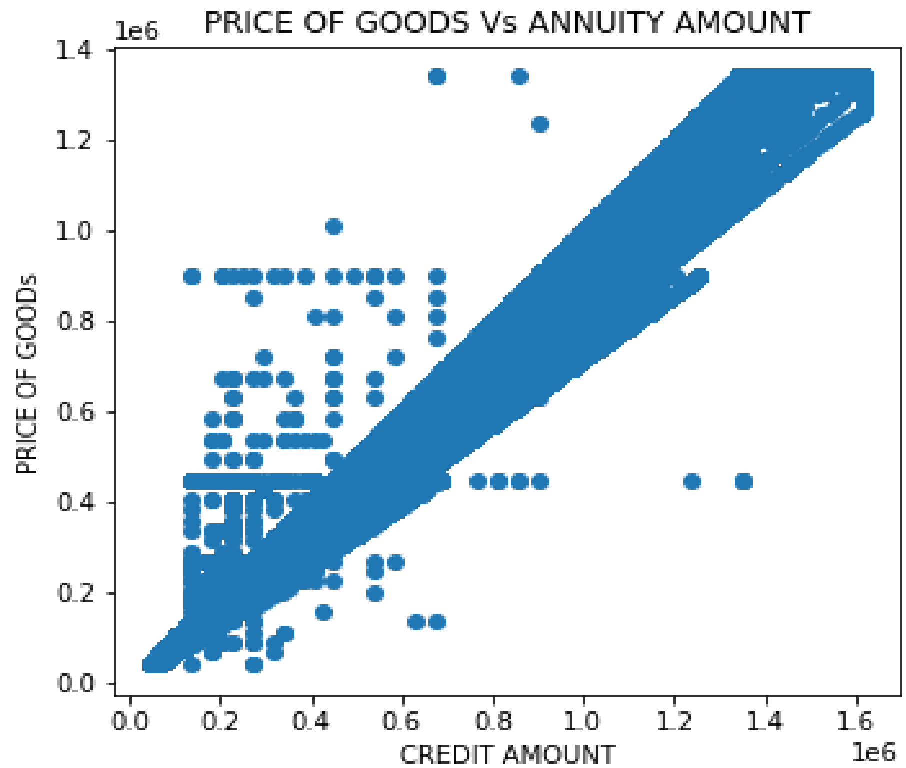
Core Staff - 5.34% higher than defaulters	***Managers - 1.94% higher than defaulters***	***High skill tech staff - Not listed in Defaulters list as we have taken only 6 maximum values***

BIVARIATE & MULTIVARIATE ANALYSIS

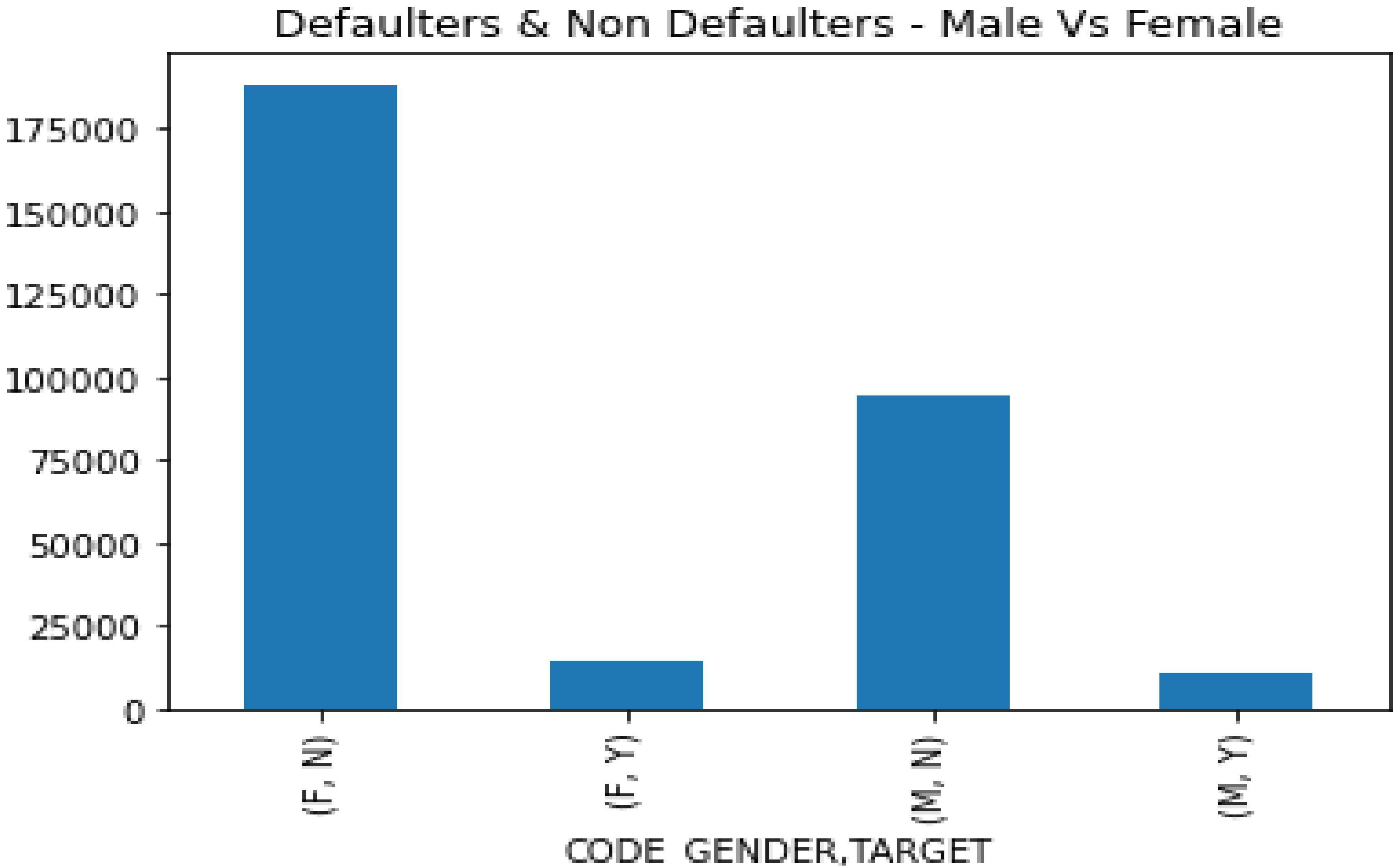
- Now we will find out the defaulters
- We will work upon the separate dataset for deaulters (TARGET = 1)
- We will try to plot relationships between different variables toget more insight

SCATTER SUBPLOTPLOT – CREDIT Vs ANNUITY AMOUNT, CREDIT Vs PRICE OF GOODS

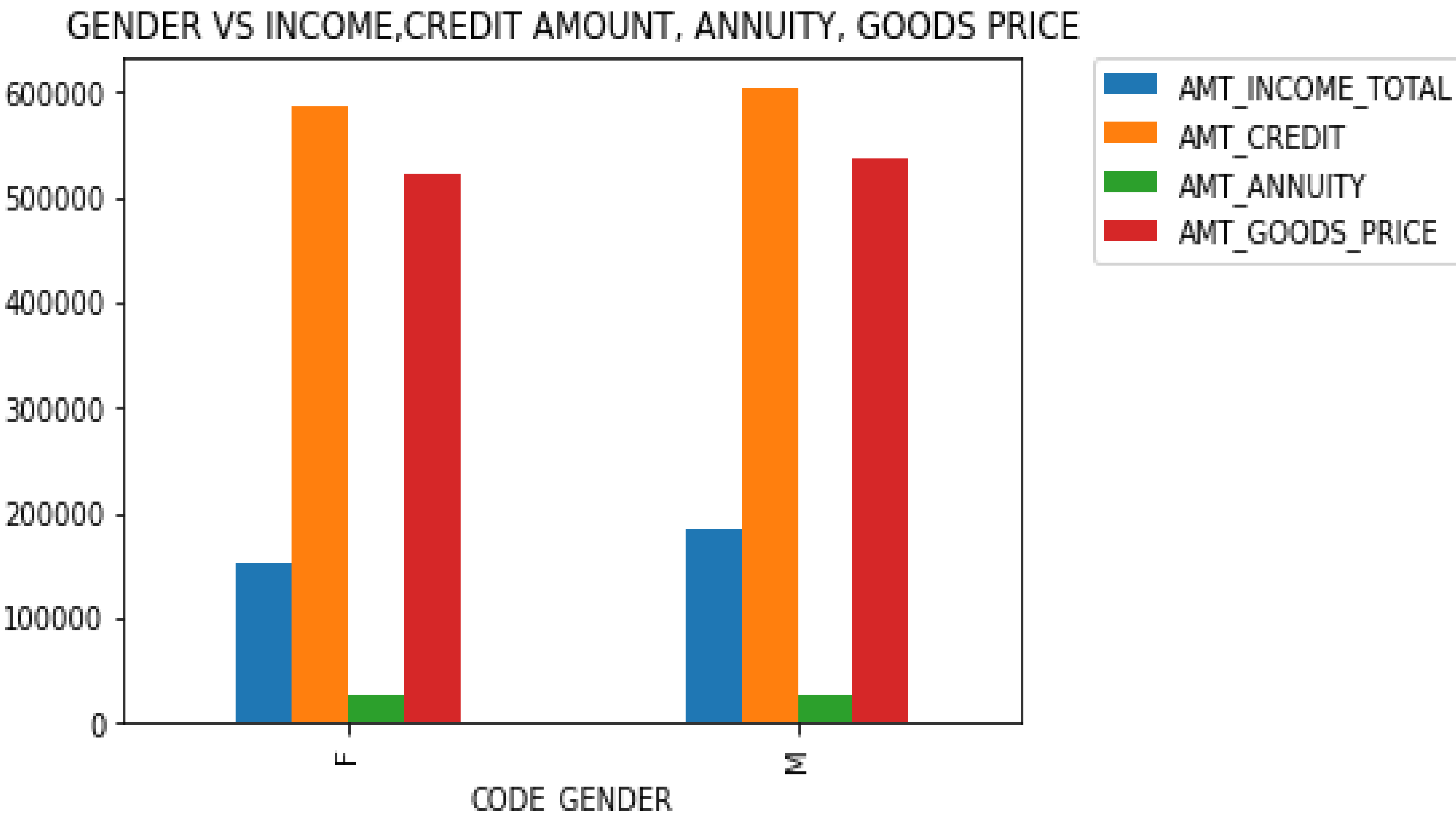




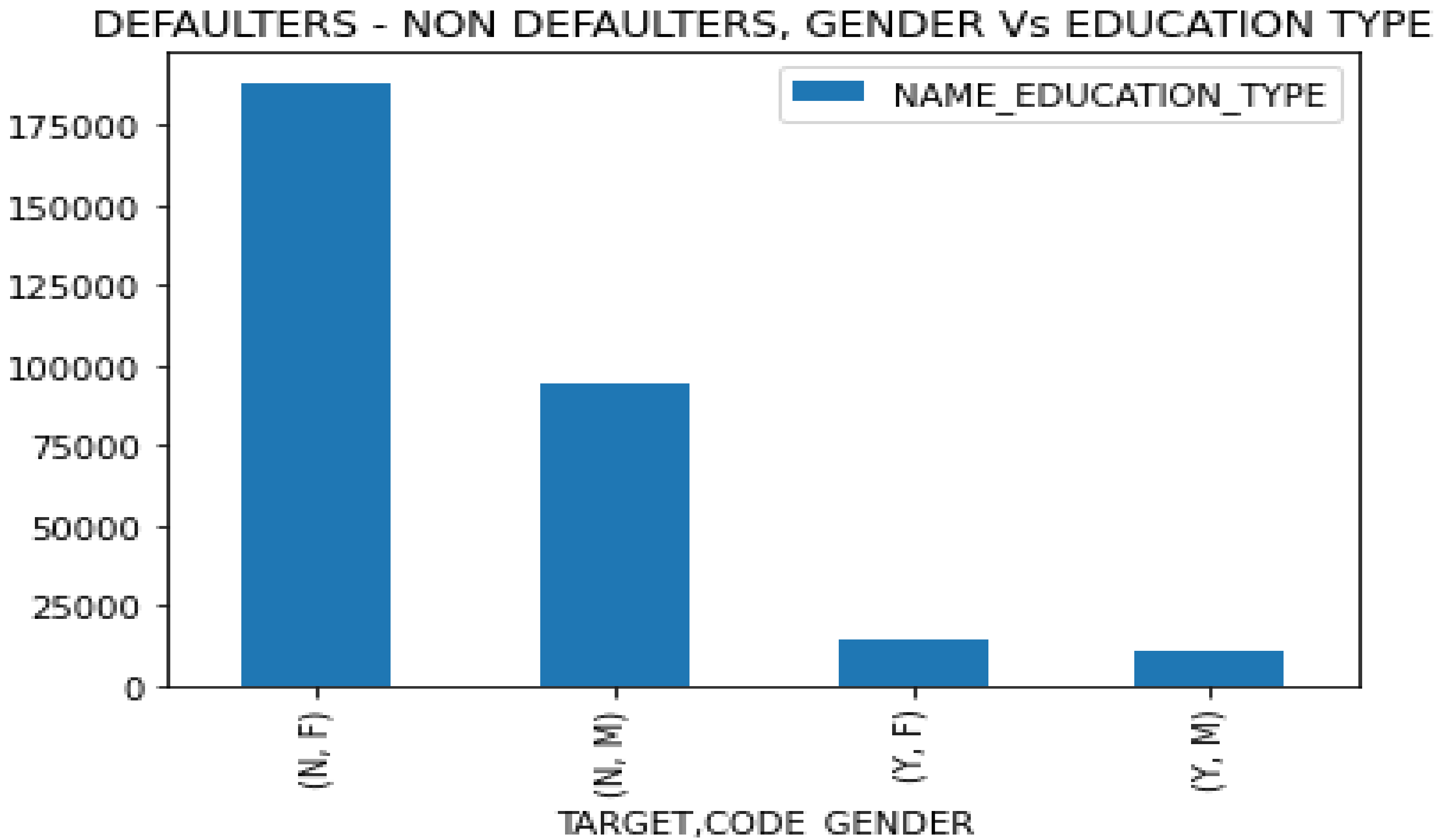
- **Positive correlation between ANNUITY & CREDIT AMOUNT**
- **Positive correlation between PRICE OF GOODS & CREDIT AMOUNT**



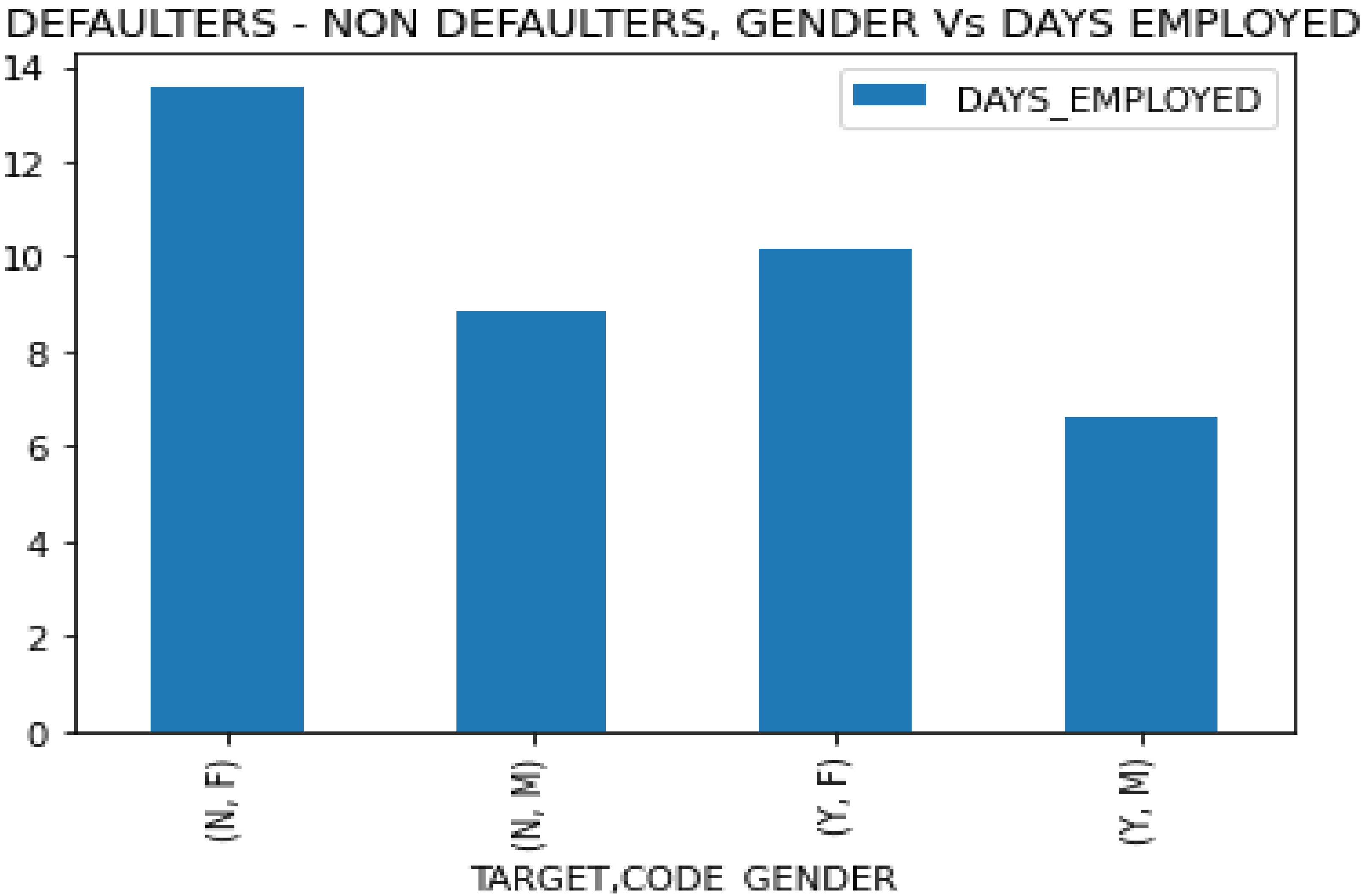
- **Number of female defaulter and non defaulter are more than males**



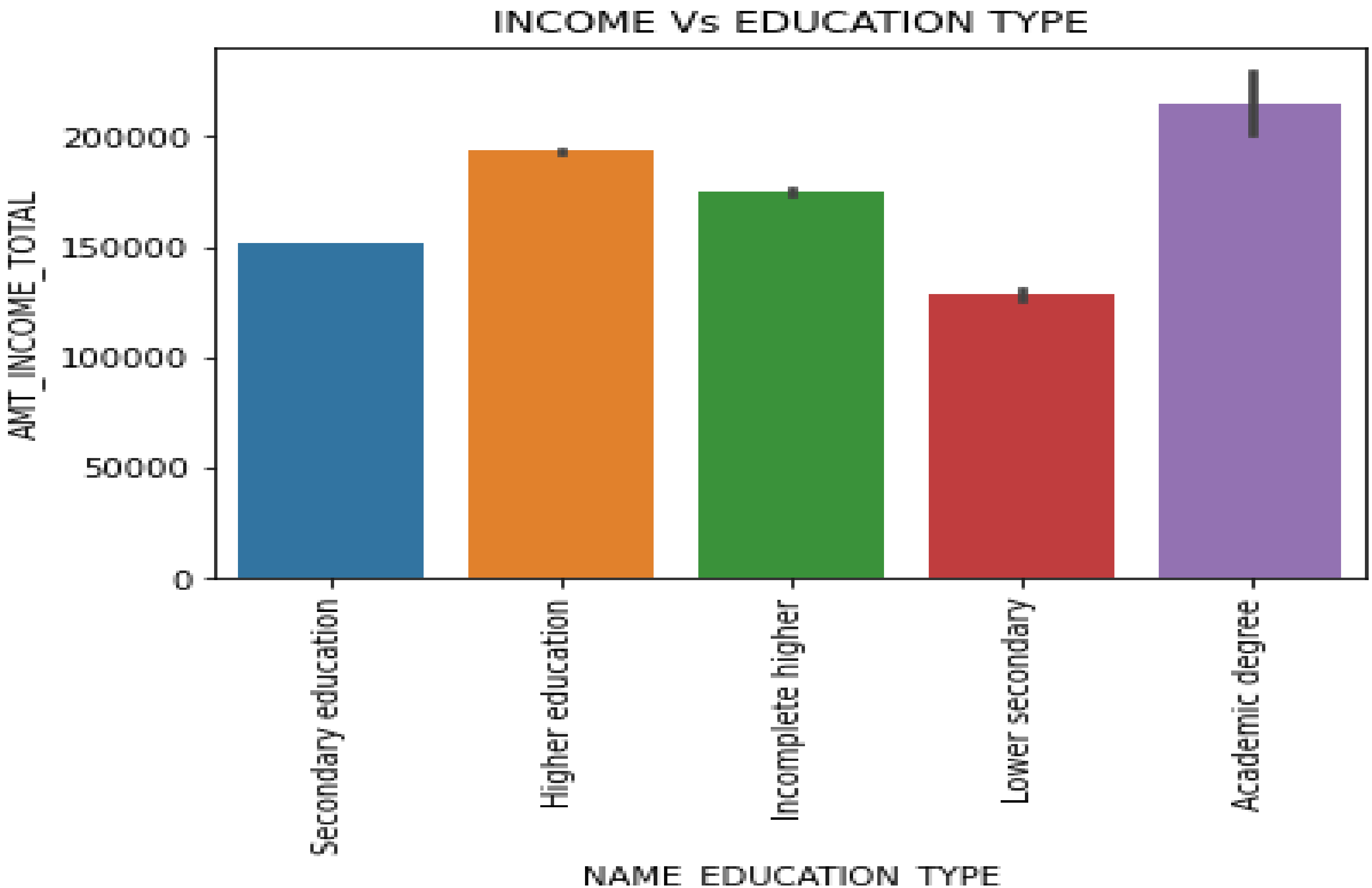
- **Income of male is pretty higher than of female so we can say that males have less loan payment difficulty.**
- **Males can get loan easily as compared to female.**



- **Most of the Non defaulter women are educated as compared to males.**

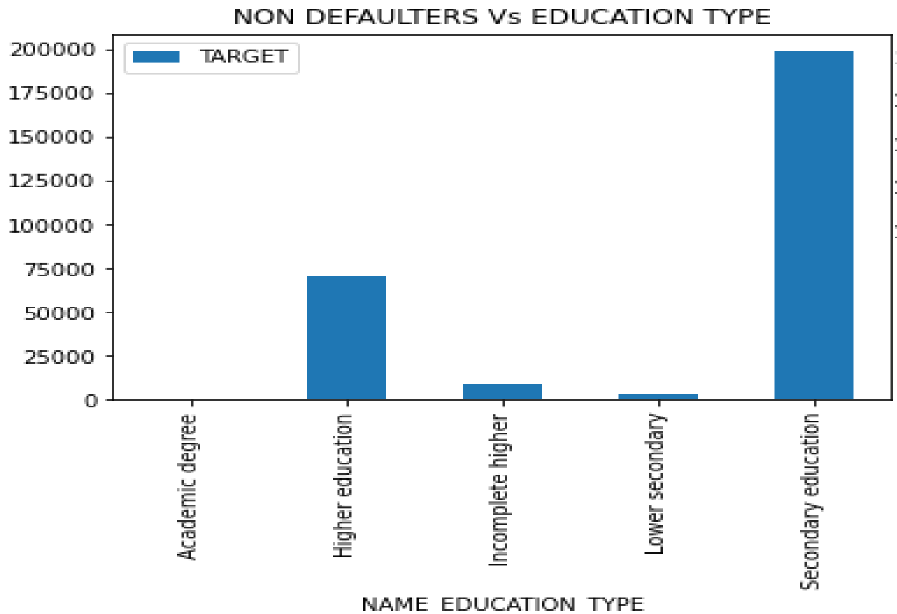


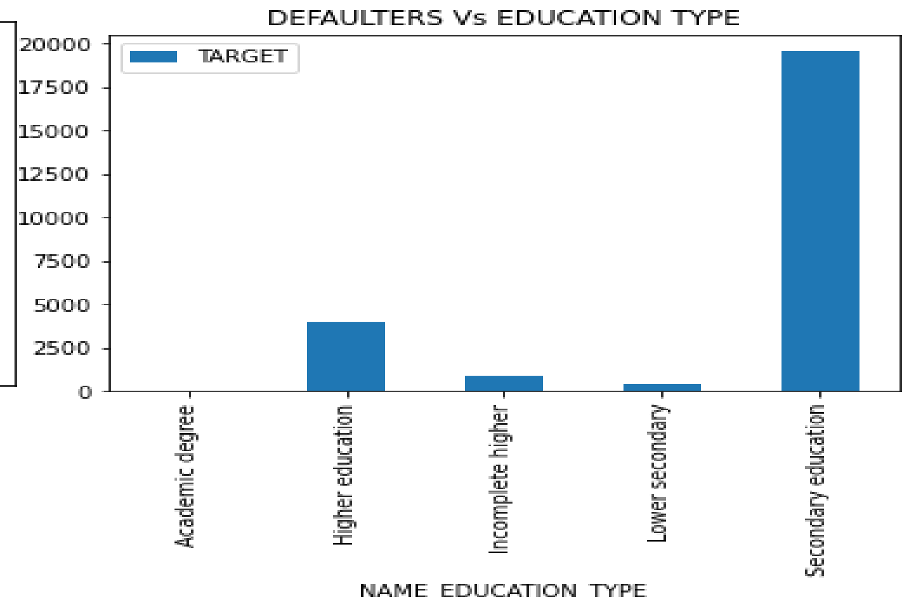
- Females employed days are higher than males



➢ **People having academic degree have high median salary. So people with Academic Degree can be targeted**

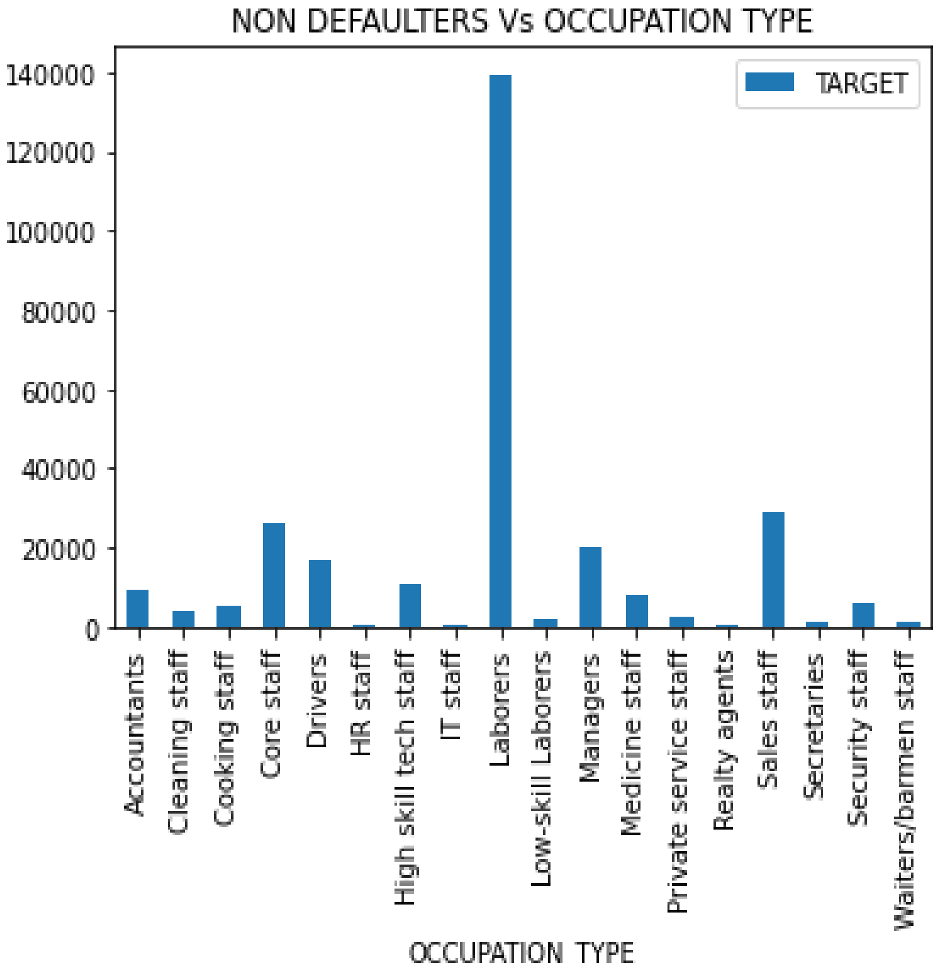
DEFAULTER – NON DEFAULTERS VS EDUCATION TYPE

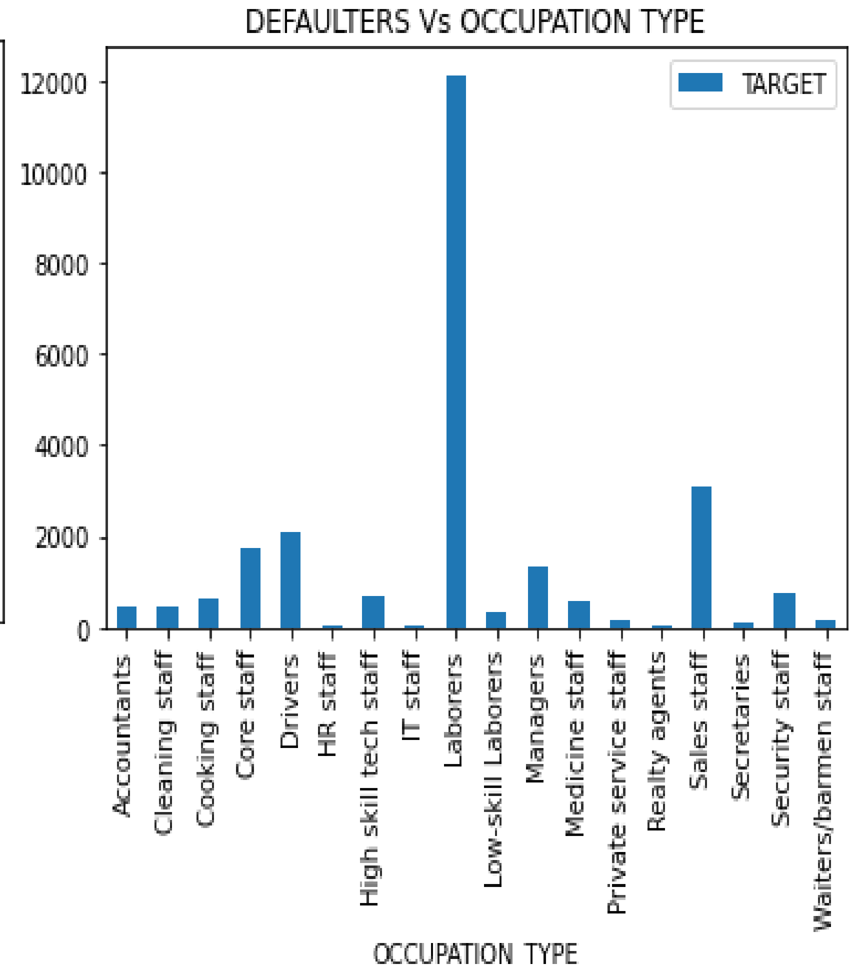




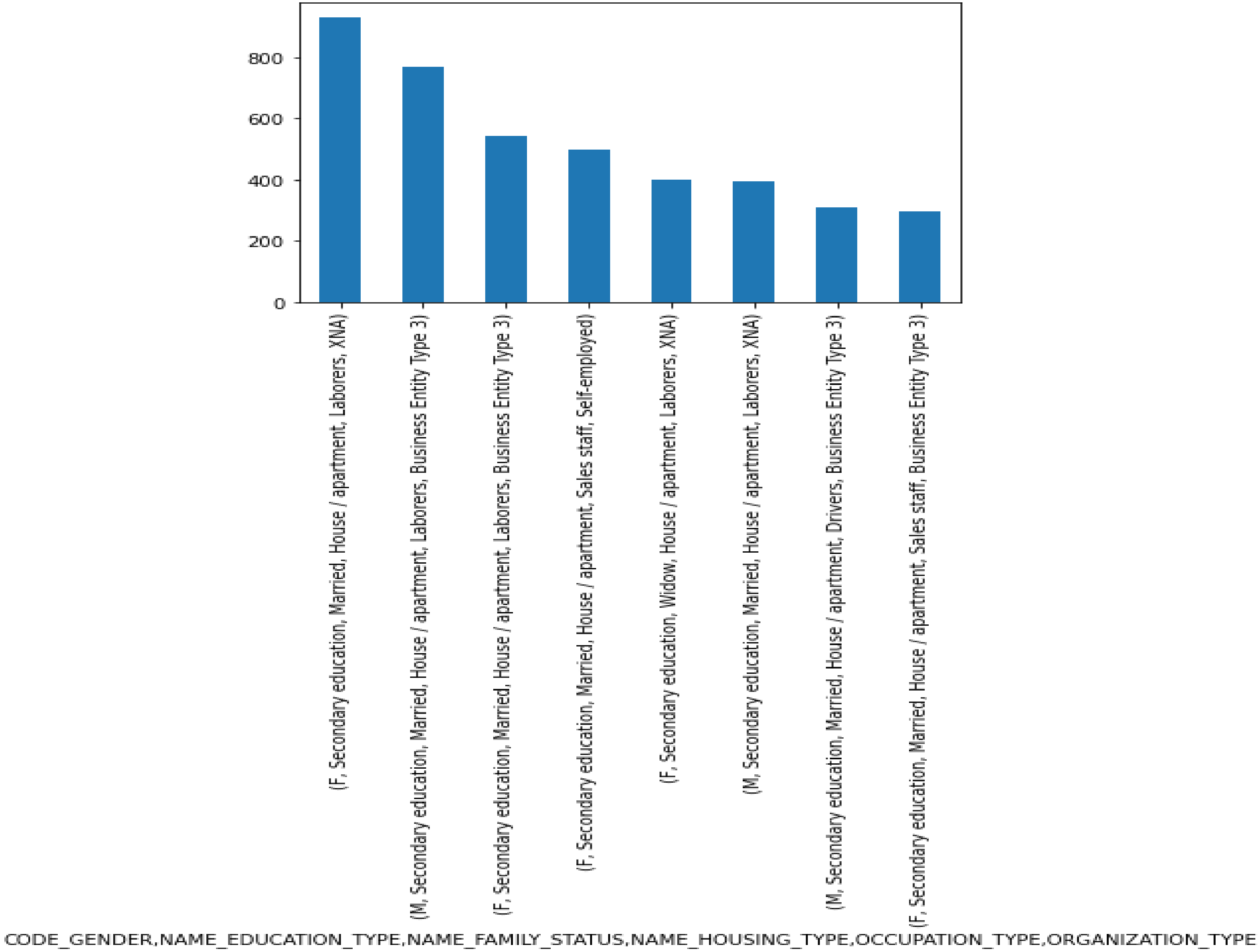
- **People having secondary education are highest number of defaulter and non defaulters**

- **People having Higher education are more in non defaulter bucket so they can be targeted**





- **Maximum number of Labourers are the defaulters and non defaulters both**



800
600
400
200
0
(F, Secondary education, Married, House / apartment, Laborers, XNA)
(M, Secondary education, Married, House / apartment, Laborers, Business Entity Type 3)
(F, Secondary education, Married, House / apartment, Laborers, Business Entity Type 3)
(F, Secondary education, Married, House / apartment, Sales staff, Self-employed)
(F, Secondary education, Widow, House / apartment, Laborers, XNA)
(M, Secondary education, Married, House / apartment, Laborers, XNA)
(M, Secondary education, Married, House / apartment, Drivers, Business Entity Type 3)
(F, Secondary education, Married, House / apartment, Sales staff, Business Entity Type 3)
CODE_GENDER,NAME_EDUCATION_TYPE,NAME_FAMILY_STATUS,NAME_HOUSING_TYPE,OCCUPATION_TYPE,ORGANIZATION_TYPE

INSIGHT

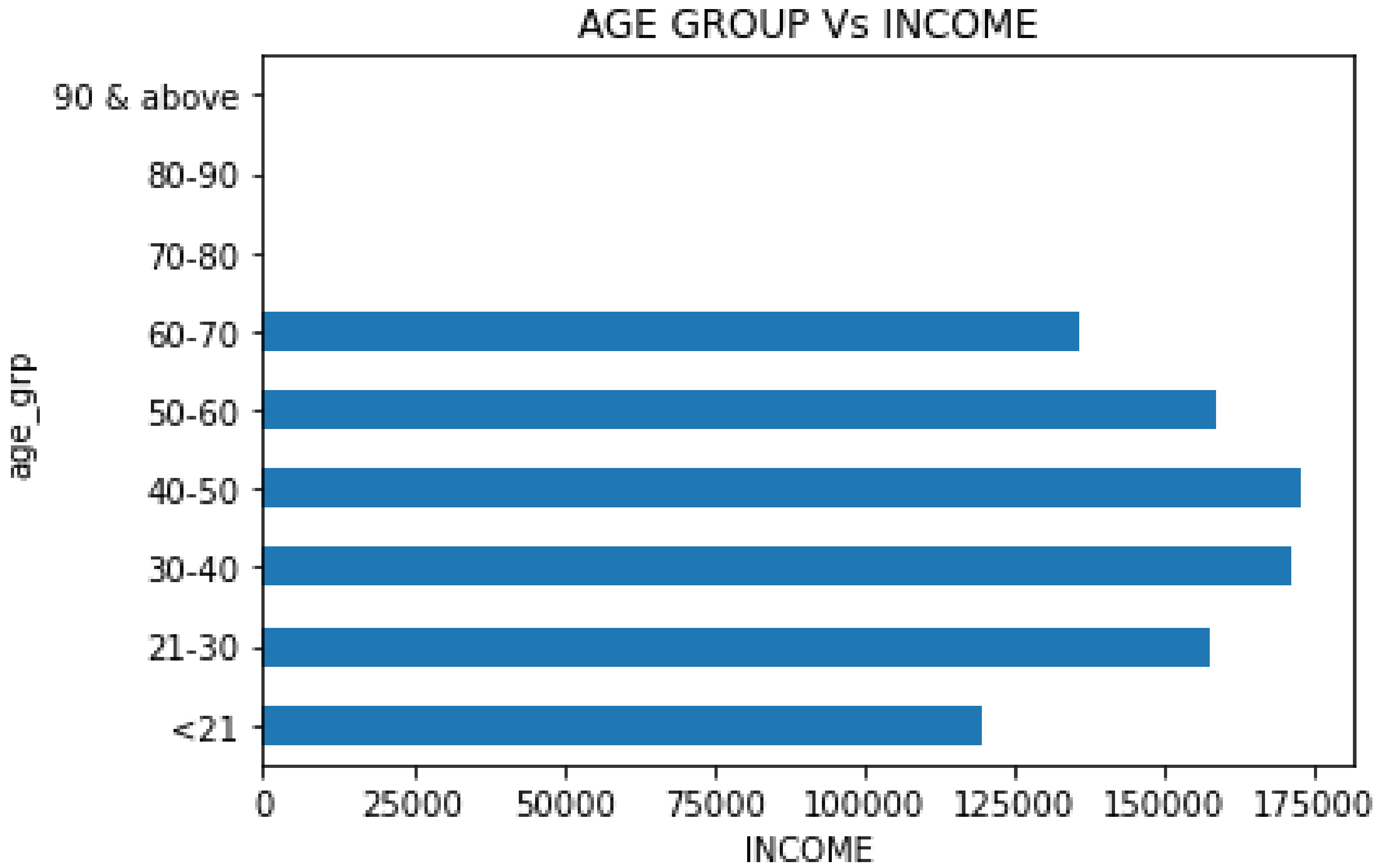
MAJOR DEFAULTERS

MALE & FEMALE

MARRIED

LIVING IN APARTMENT

WORKING IN A BUSINESS ENTITY TYPE 3

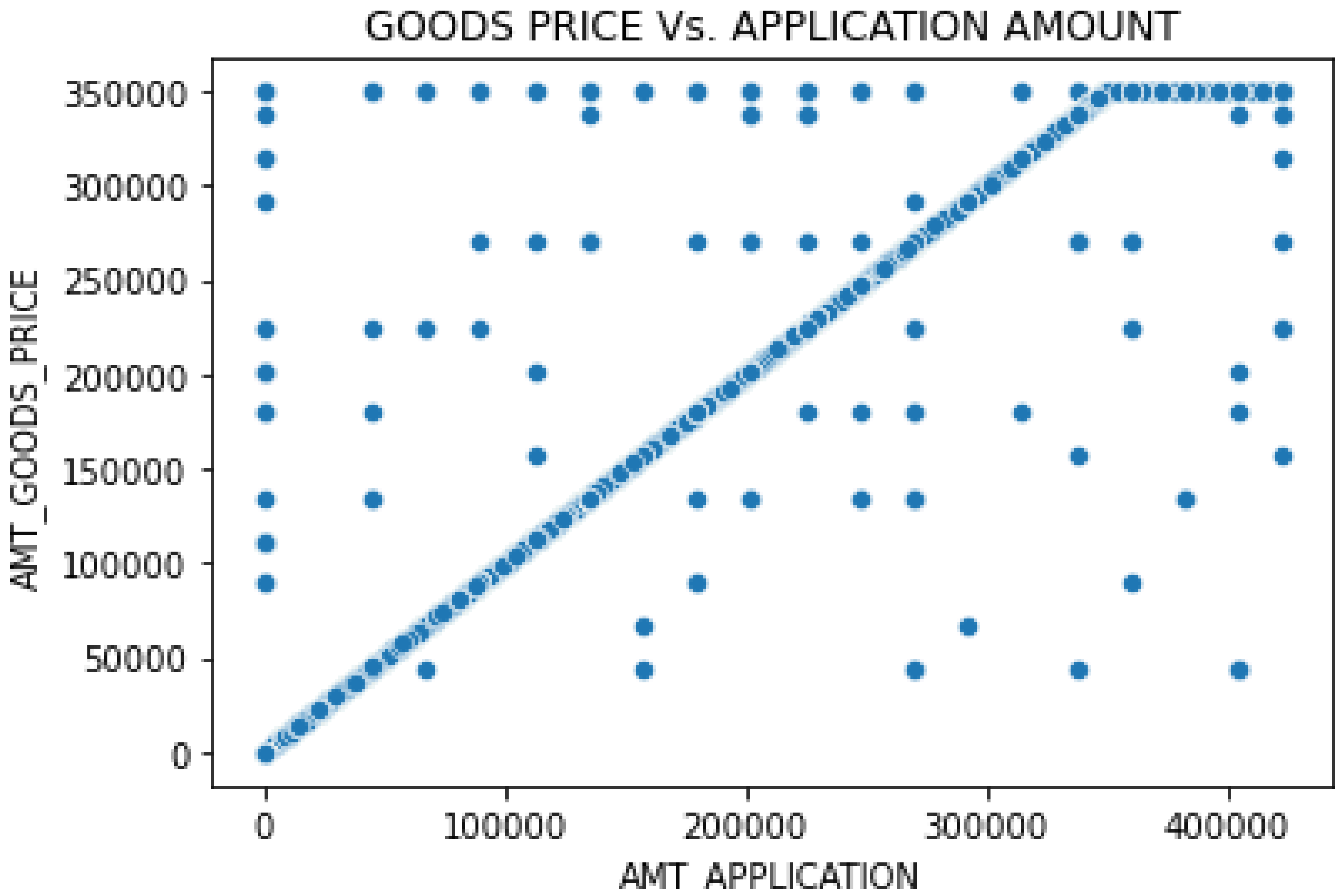


Maximum number of the people applying for the loan are between 40-50 age and have high mean salary

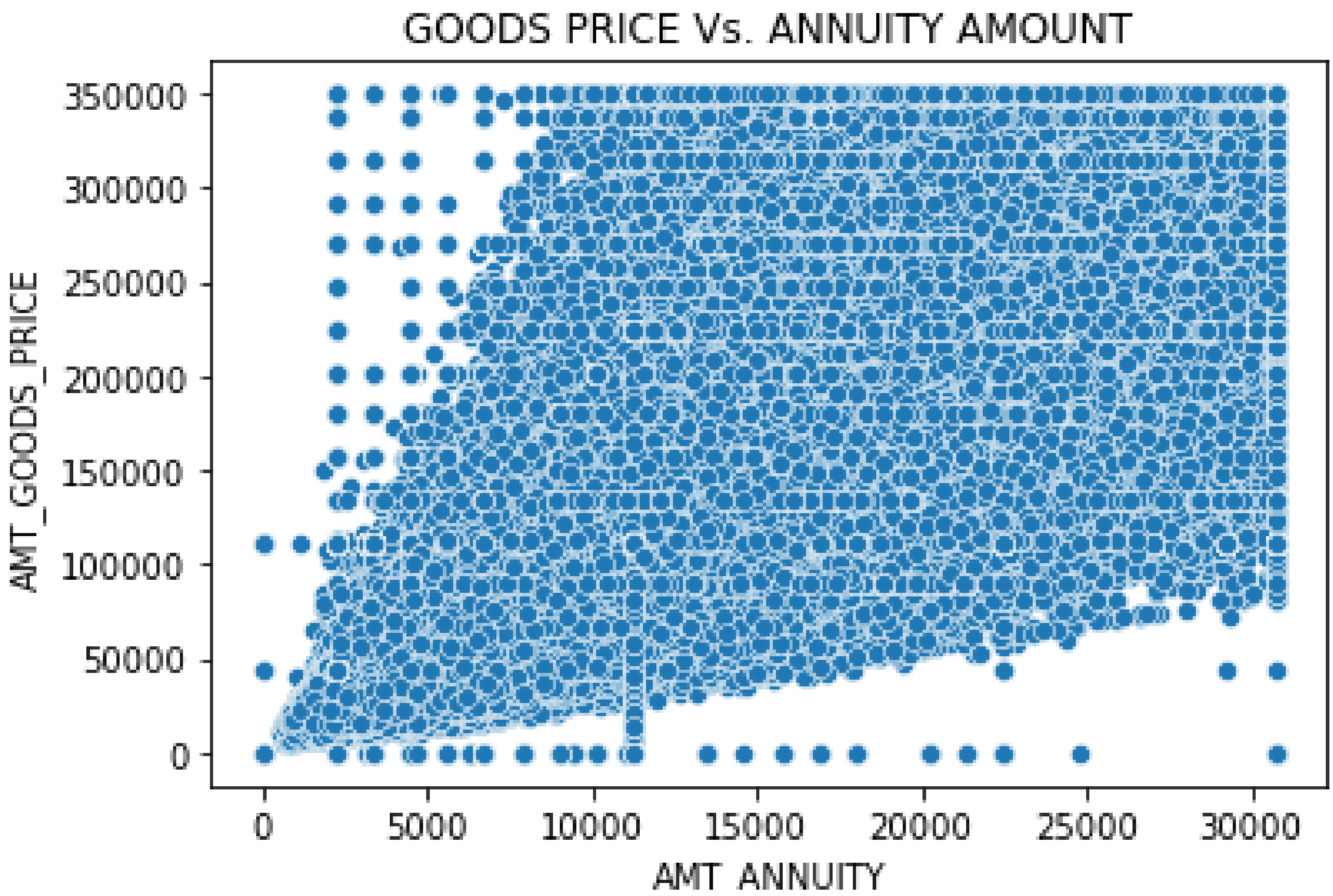
PREVIOUS APPLICATION DATASET

Data cleaning approach is same as in application dataset

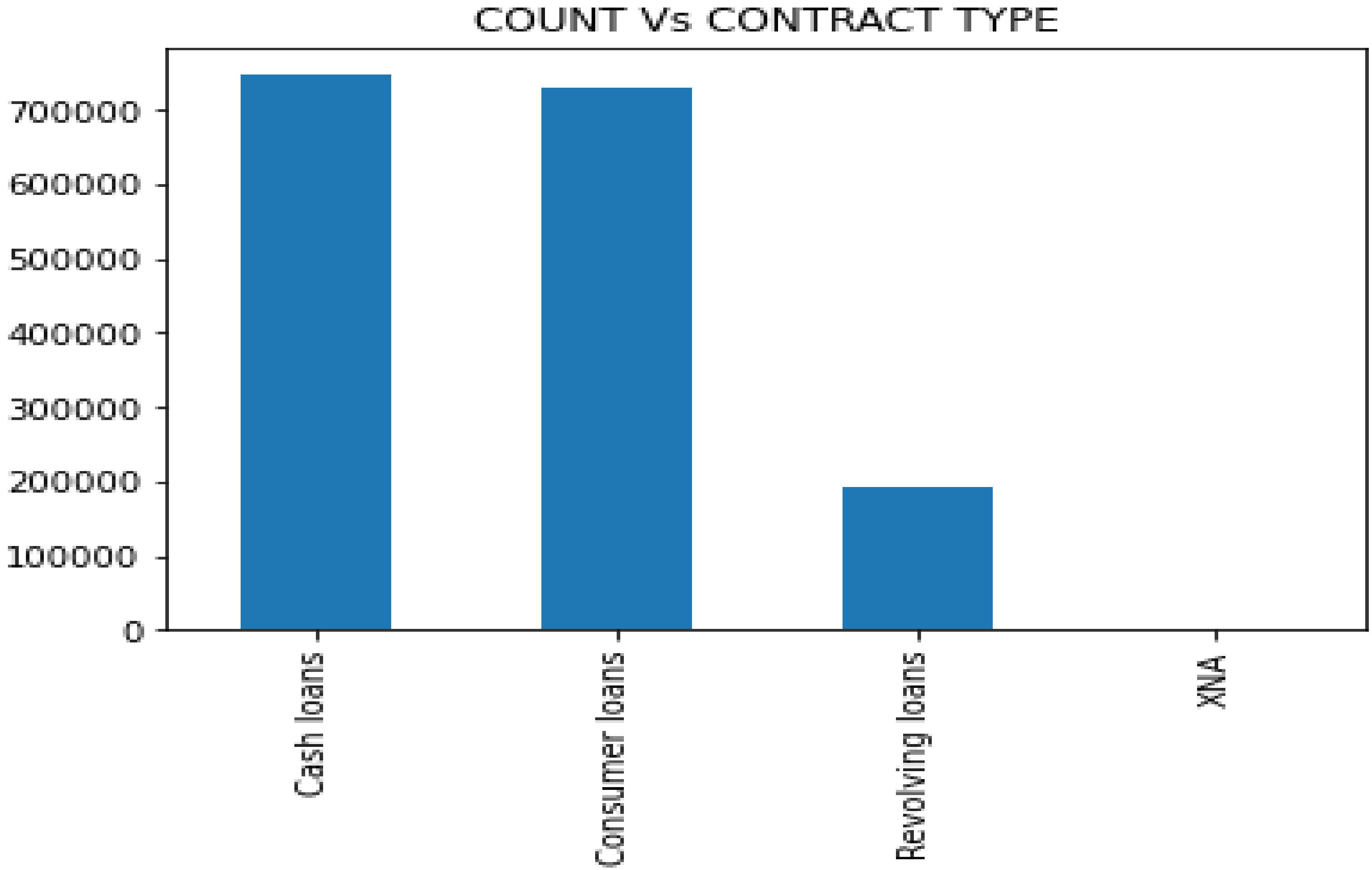
1. • Importing Libraries
2. • Reading the data set and finding percentage of null values
3. • Dropping columns with missing values >45%
4. • Identifying continuous and categorical columns/variable
5. • Continuous column - Columns containing unique values > 58
6. • Categorical column – Columns containing unique values < 58
7. • Imputation for missing value<45 (categorical – mode, continuous – median)
8. • Dropping unnecessary columns
9. • Detecting Outliers – Using Subplots
10. • Handling outliers by flooring and capping



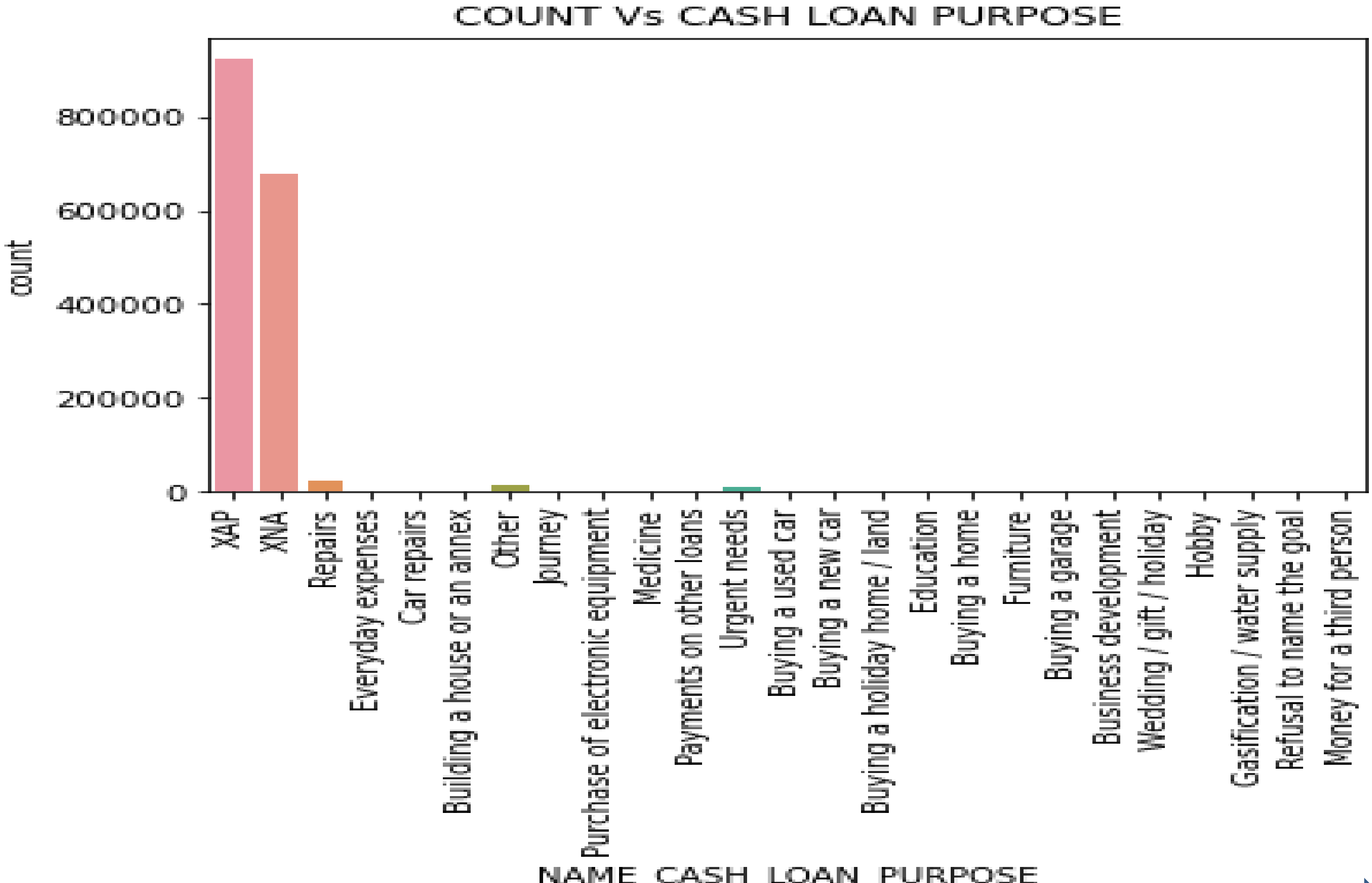
➢**As the goods price increases amount of loan application also increases**



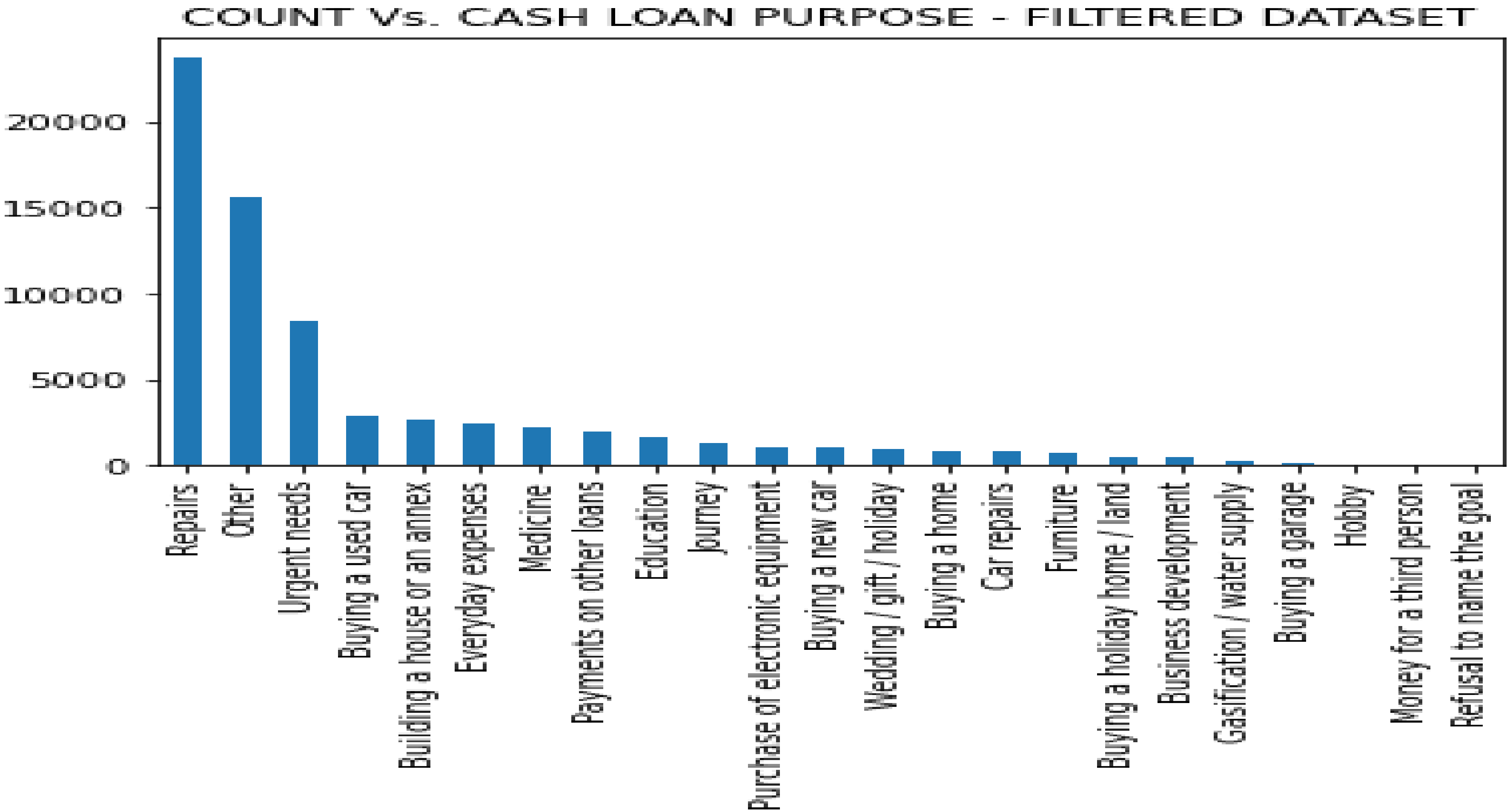
➢**As the goods price increases annuity amount also increases**



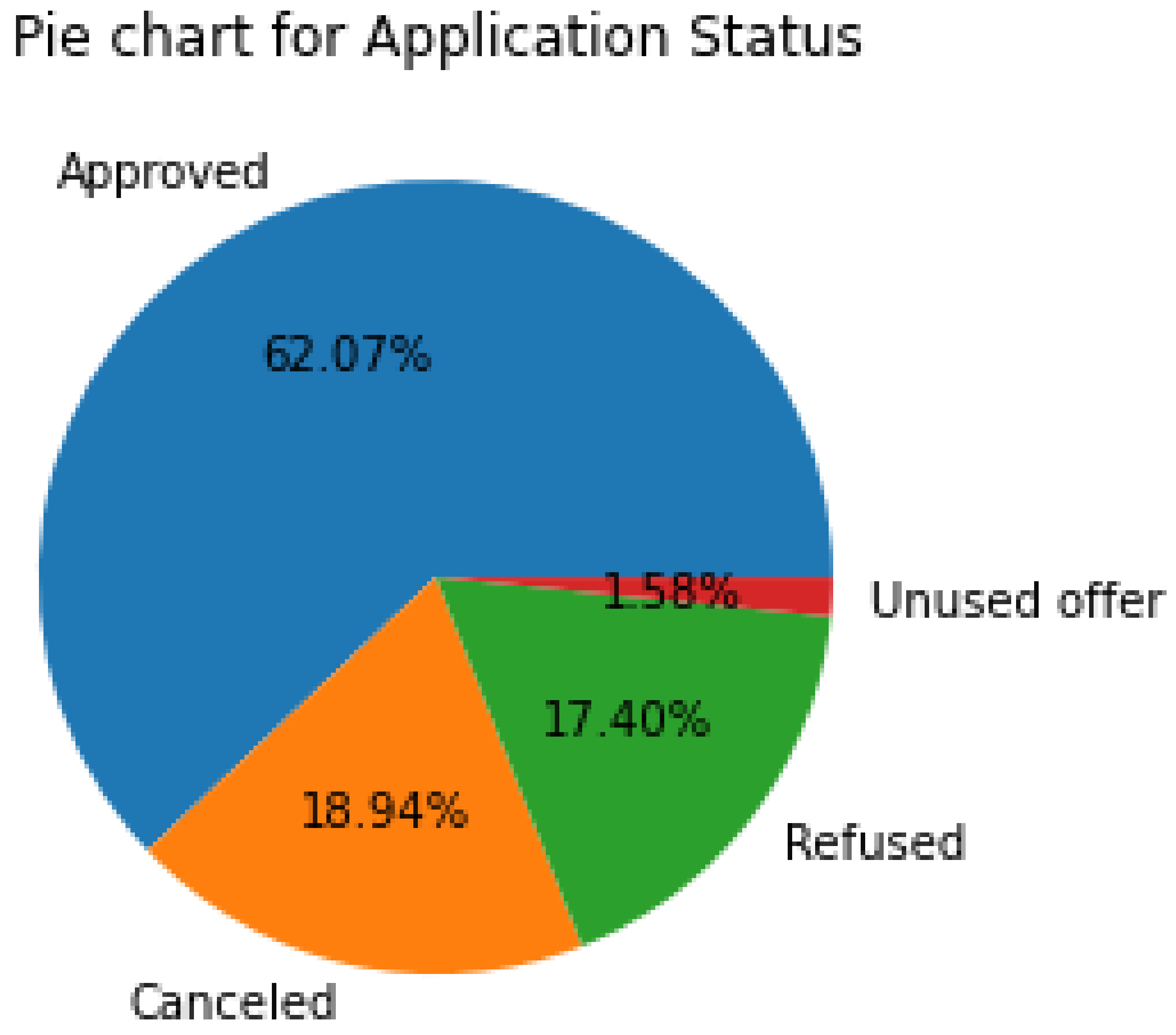
- **People previously applied for cash loans more**

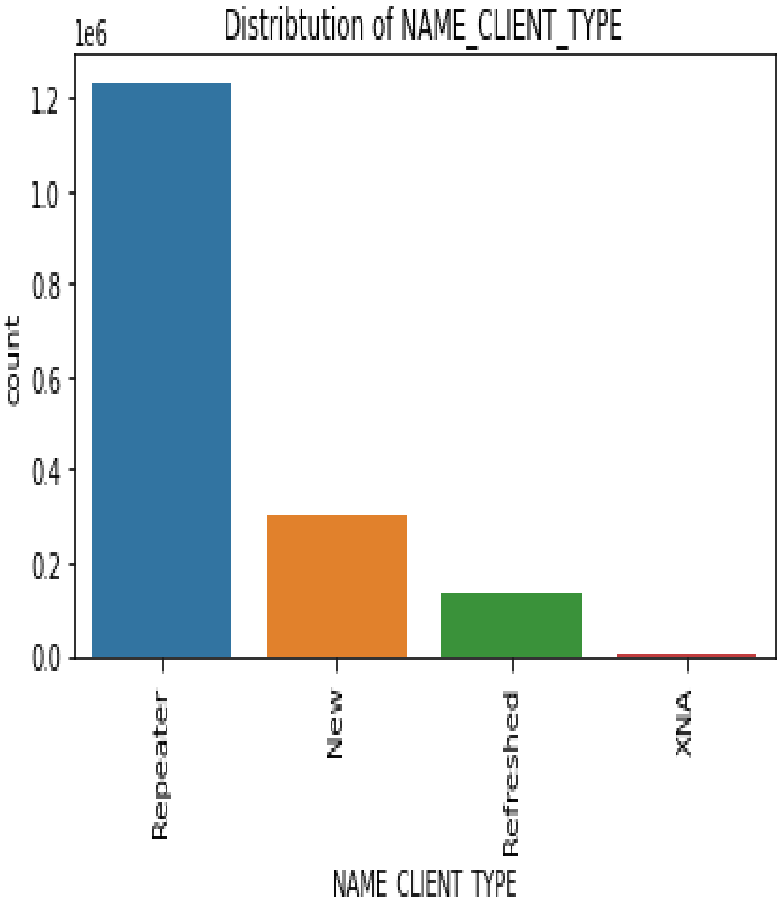


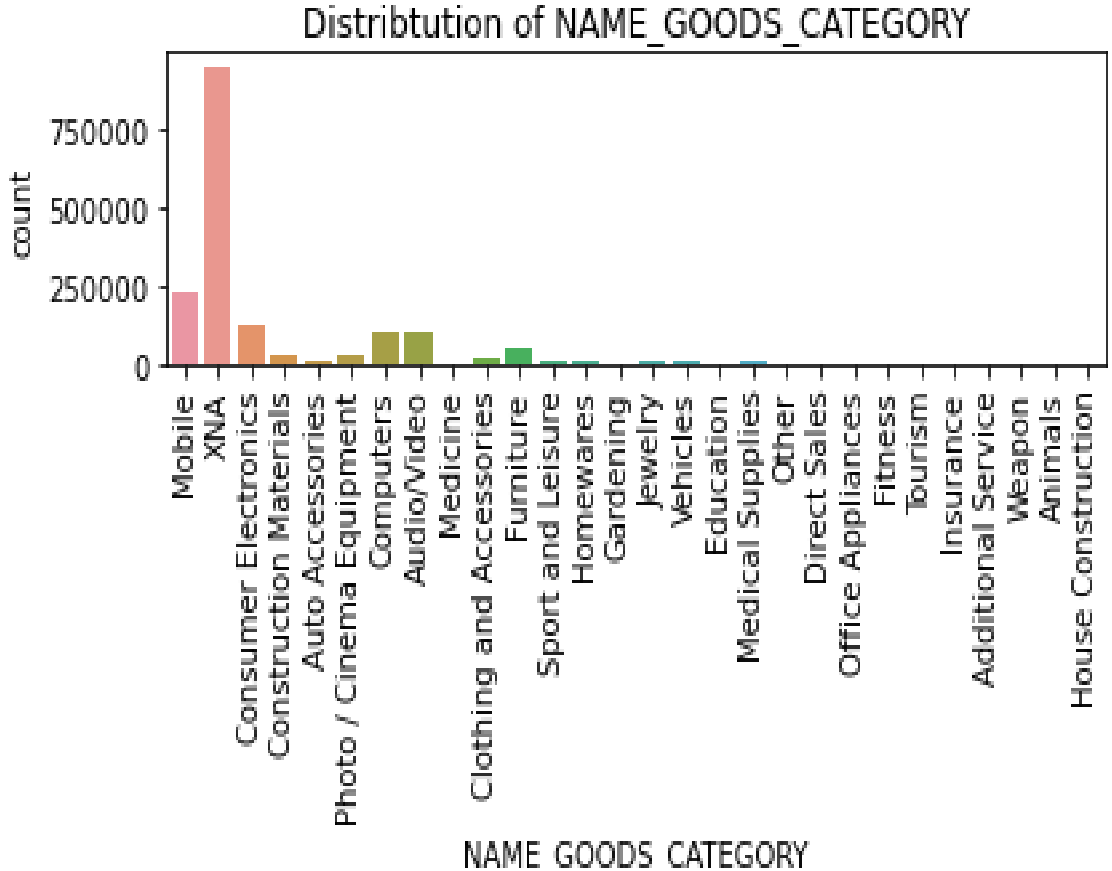
IGNORING XAP, XNA. GOING DEEPER FOR OTHER CASH LOAN PURPOSE



- **Maximum number of people are applied for loan for repair purpose**
- **Secondly people applies, for other reasons**
- **People also require loan for urgent needs which are not specified**
- **People applies loan for buying a used car, building a house.**


Pie chart for Application Status
Approved
62.07%
1.58%
Unused offer
17.40%
18.94%
Refused
Canceled

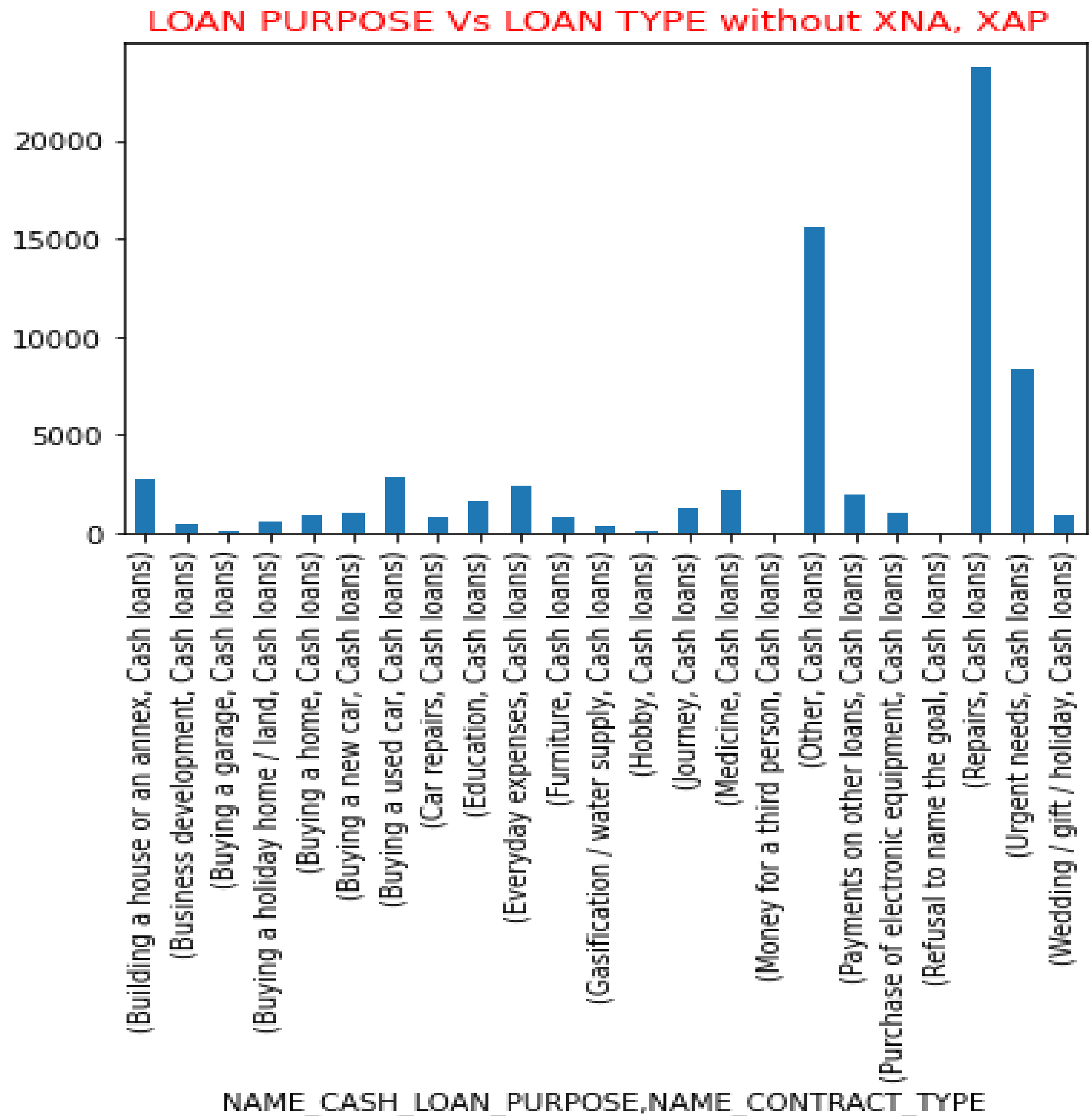




- **Maximum number of old clients are applying for the loan again**

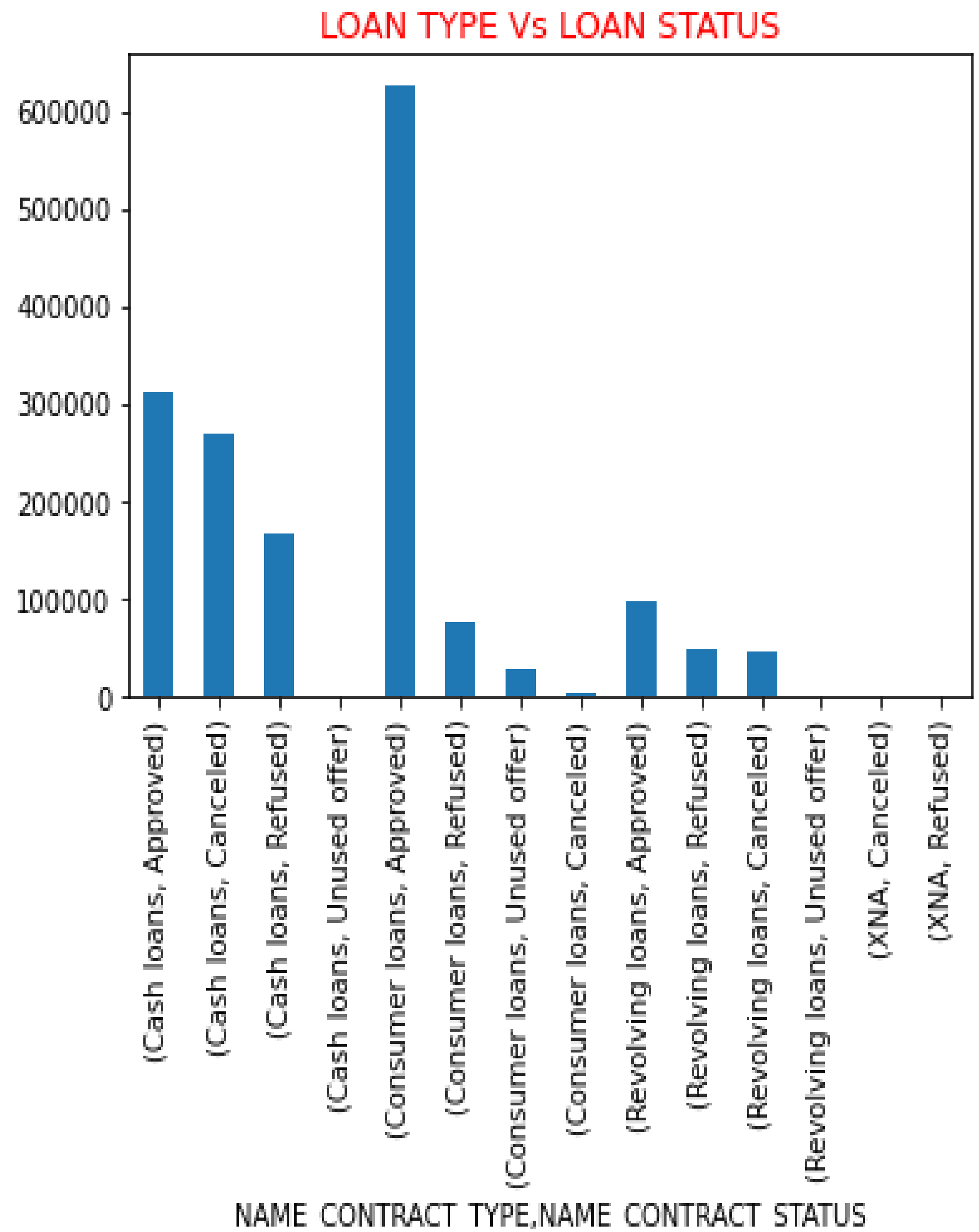
- **Other than XNA, people are applying loan for mostly mobile phones.**

Maximum number of people are taking cash loans for Repair



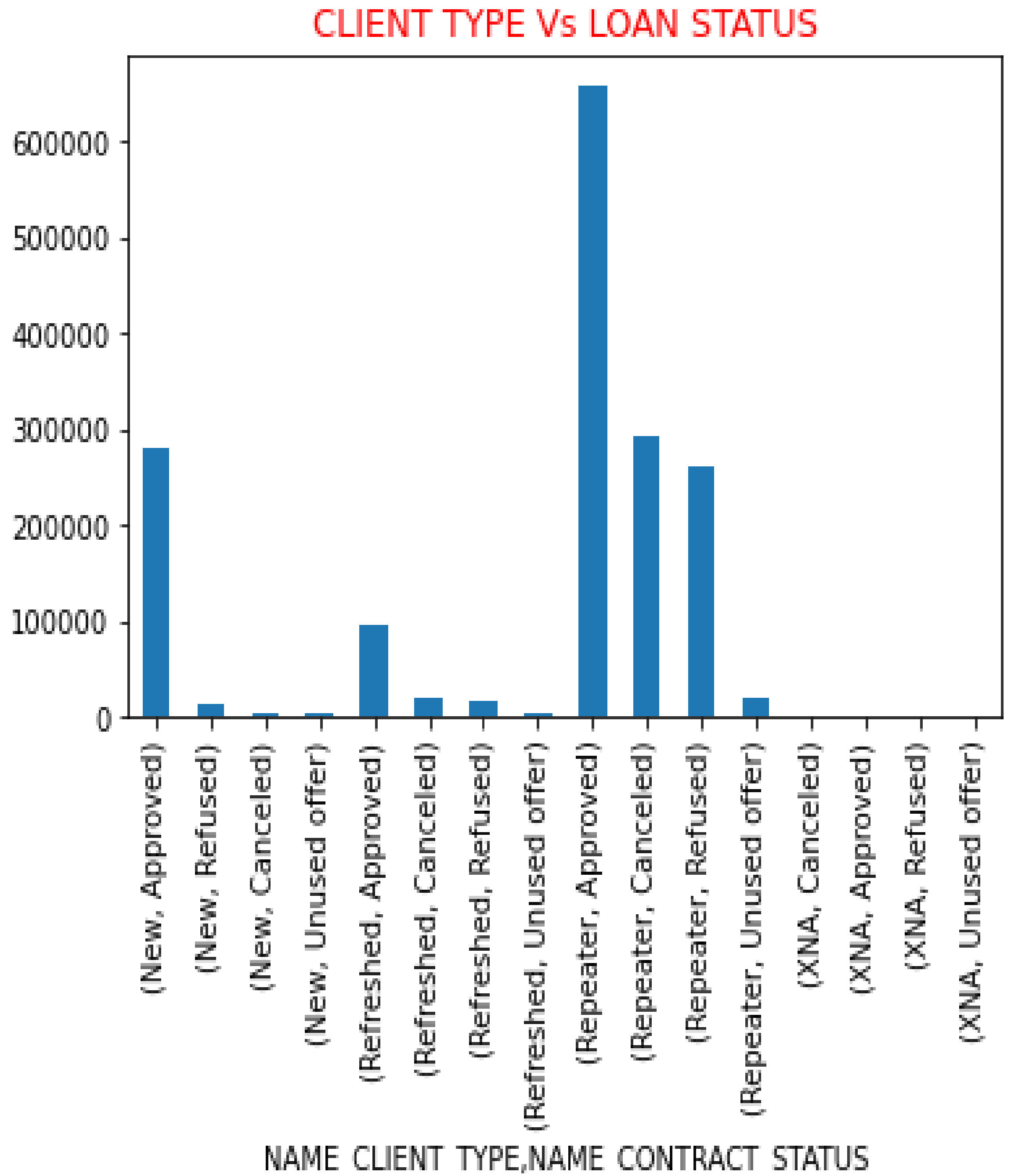
- ➢**Maximum number of loans that are approved is Consumer loans**

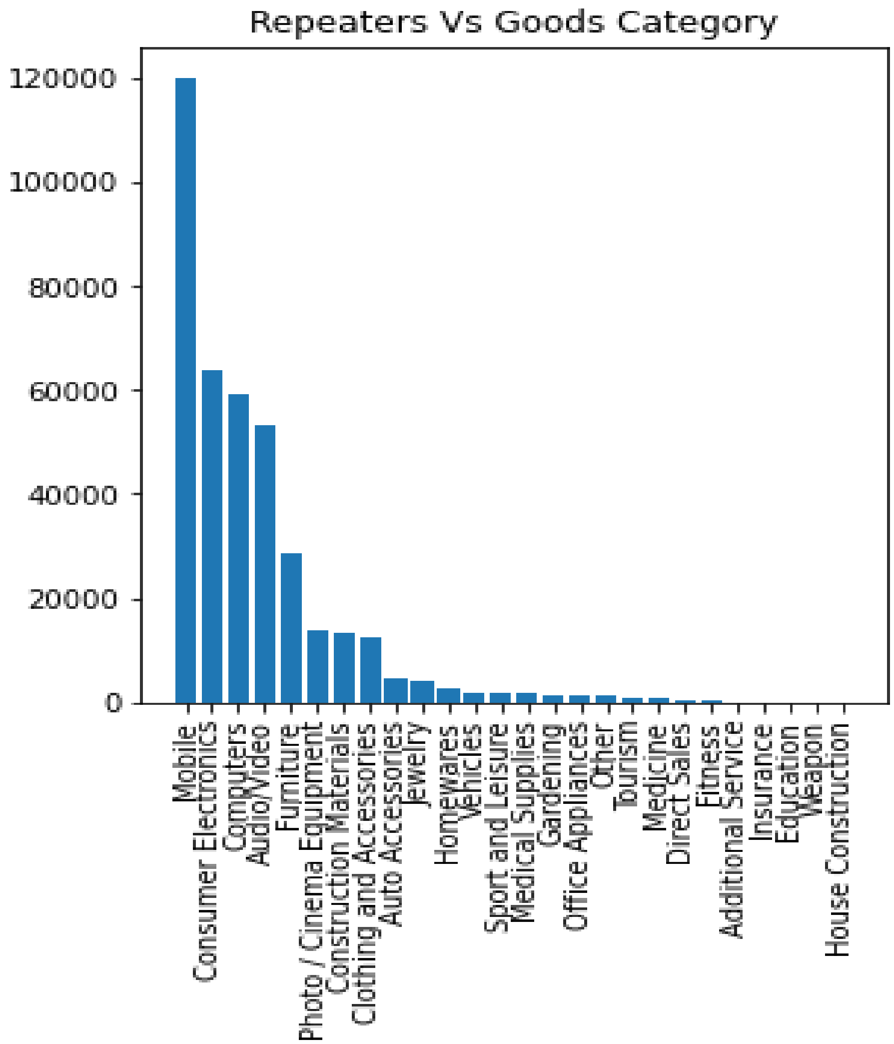
- ➢**Cancellation of Consumer loans is very less as compared to cash and revolving loans**

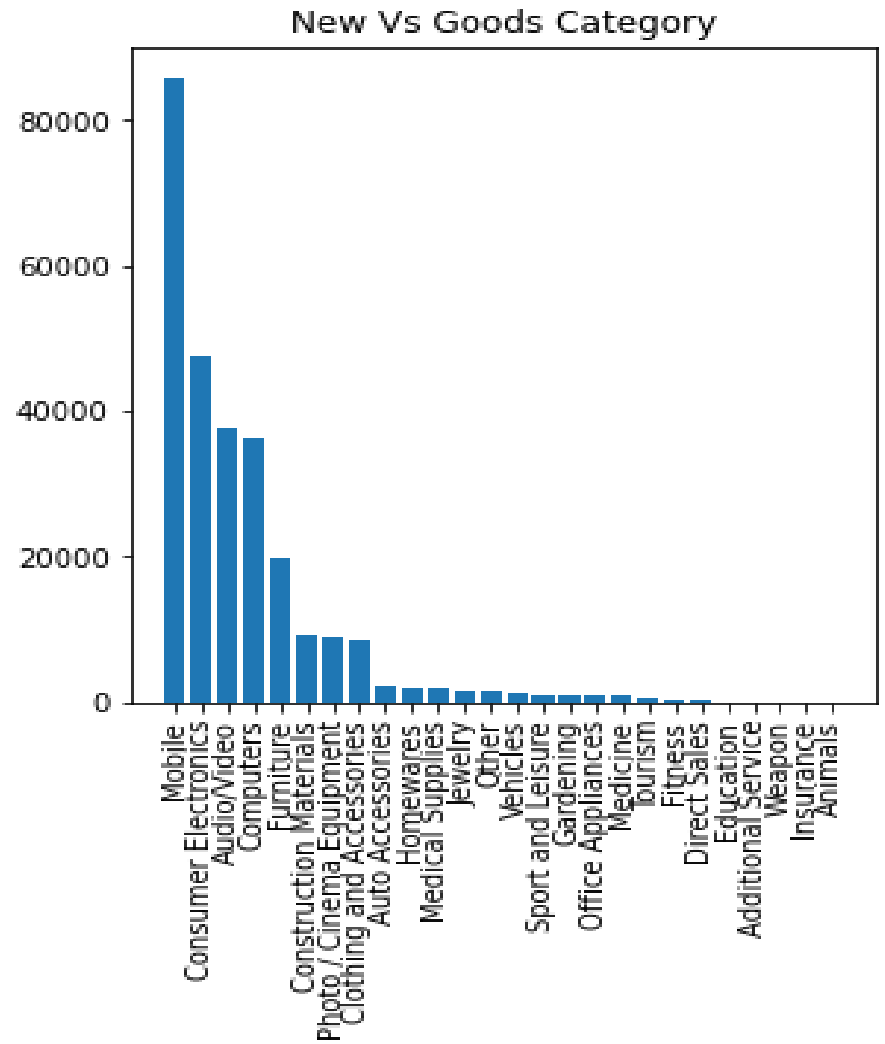


- ➢**Company approves the loan for maximum number of people who are applying for the new loan again**

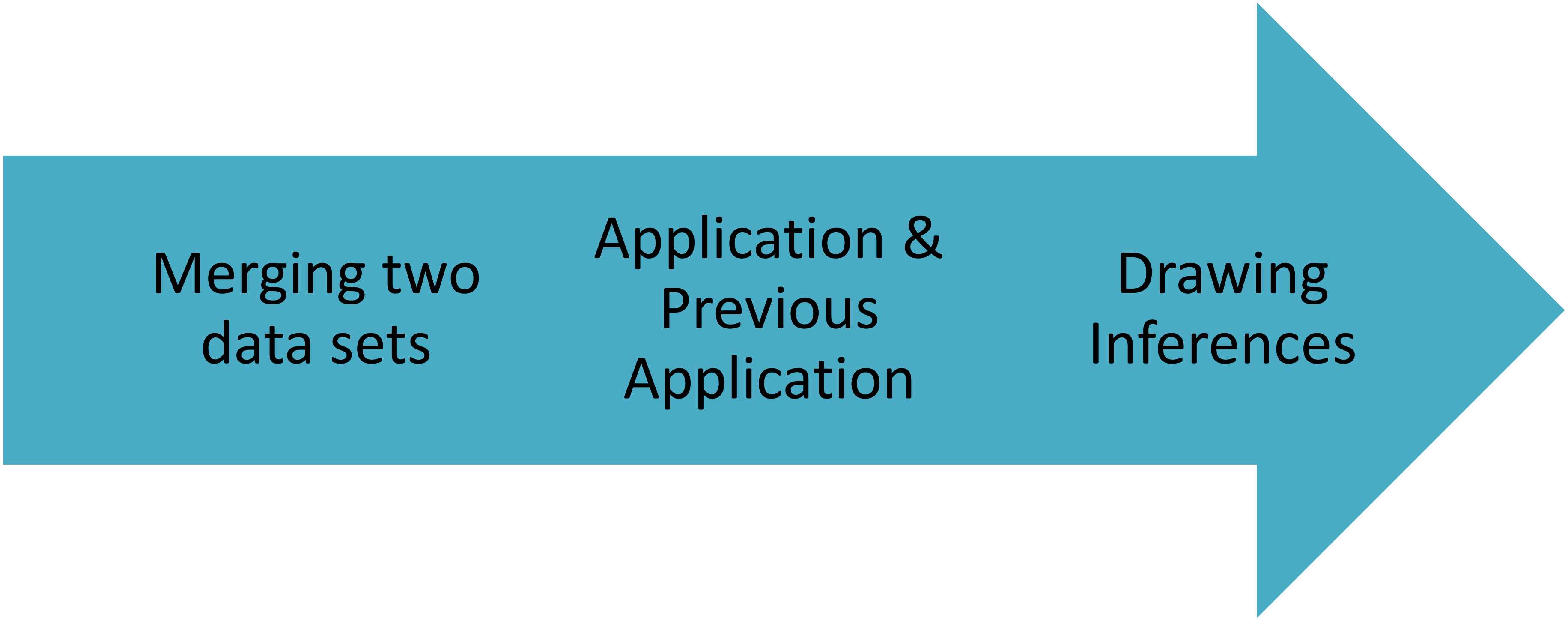
- ➢**For People applying for the first time, company tries not to cancel the loan application that is why cancellation for new people is very less as compared to refreshed and repeaters**

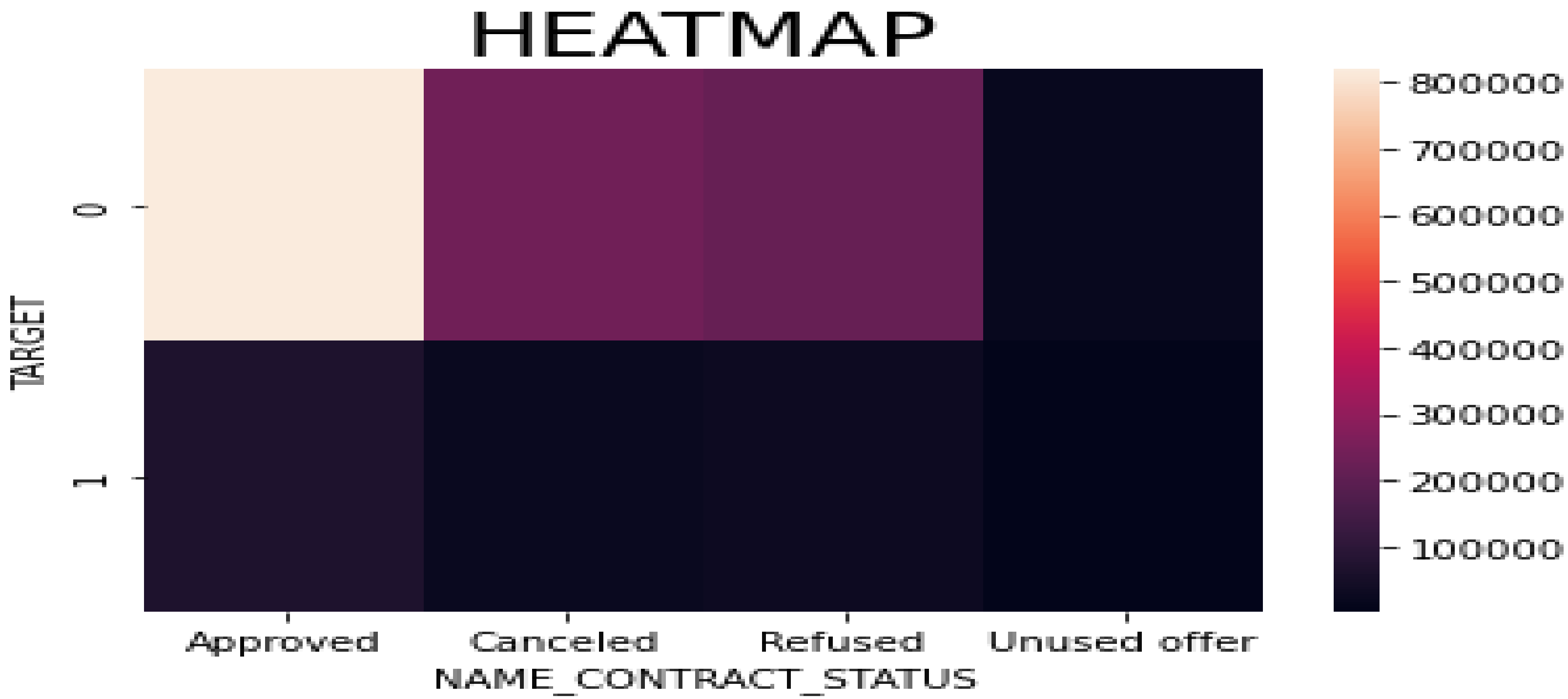






➢**Maximum number of people (Repeaters and New) are applying loan for mobile phones.**


Merging two data sets
Application & Previous Application
Drawing Inferences



- **Maximum number of loan is approved for non defaulters**
- **Loans are approved for defaulters also.**
- **Loans are refused for non defaulters also.**

MAJOR – INSIGHT

By taking mode of categorical columns with **Target = 0**, **CONTRACT_STATUS = Refused**
By taking mode of categorical columns with **Target = 1**, **CONTRACT_STATUS = Approved**

LOAN REFUSED PREVOUSLY FOR DEFAULTERS

LOAN APPROVED PREVIOUSLY FOR NON DEFAULTERS

APPLYING FOR CASH LOANS

FEMALE, MARRIED, HAVE SECONDARY EDUCATION

DOESNOT OWNS CAR, BUT HAVE HOUSE

LABOUR BY OCCUPATION AND WORKING IN BUSINESS ENTITY 3

THANK YOU