CREDIT EDA ASSIGNMENT

HARSHIT KAMANI

PROBLEM STATEMENT

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 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.

PROBLEM STATEMENT CONT.

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, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

PROBLEM STATEMENT CONT.

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.

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).

Steps Involved:

- Importing Libraries
- Reading Data Set – application_data.csv
- Checking basic data about DS like shape, info, describe, etc. to understang DS in a better way,
- Checking for Null Values & XNA Values – Imputing/Dropping as per scenario.
- Dropping unnecessary columns that are not useful in analysis.
- Converting negative data to positive data and changing Data Types where it’s necessary.
- Checking for outliers and binning the data/creating new column where it’s necessary.
- Same steps to be followed for previous_application.csv
- Merging both Data Sets
- Checking for imbalance and dividing them into TARGET_0, TARGET_1, Rejected_Loans, Approved_Loans Datasets.
- Performing Univariate Analysis for Categorical & Numerical columns.
- Performing Bivariate Analysis for Categorical & Numerical columns.
- Performing Univariate Analysis for Categorical vs Numerical columns.
- Correlation – Checking for top 10 & making heatmap.

Importing Libraries, Reading Data Set & Checking Basics

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings("ignore")

pd.set_option('display.max_columns', 125)
pd.set_option('display.max_rows',200)
```

```
### Step 2 - Reading application_data DS
application_data=pd.read_csv("application_data.csv")
```

```
application_data.head()
```

STRATION	DAYS_ID_PUBLISH	OWN_CAR_AGE	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE	FLAG_PHONE	FLAG_EMAIL	OC
-3648.0	-2120	NaN	1	1	0				
-1186.0	-291	NaN	1	1	0				
-4260.0	-2531	26.0	1	1	1				
-9833.0	-2437	NaN	1	1	0				
-4311.0	-3458	NaN	1	1	0				

```
### Step 3 - Checking Shape/info etc of DS
```

```
application_data.shape
```

```
(307511, 122)
```

```
application_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

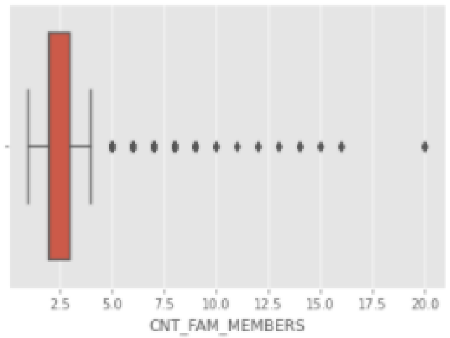
```
application_data.dtypes
```

```
SK_ID_CURR                    int64
TARGET                        int64
```

Null Value Treatment

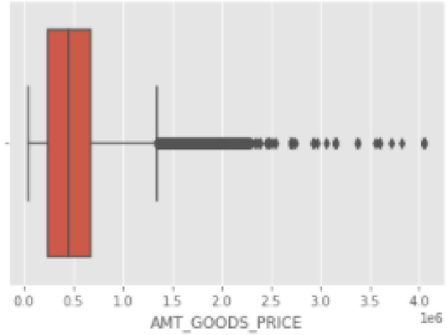
- Deleting all those columns having more than 50% values as NULL as they won't be helpful in analysis (insufficient data).
- Using **mode** to impute null values of categorical column – **Occupation Type**
- Using **mean** to impute null values of numerical columns no having outliers - **CNT_FAM_MEMBERS**
- Using **median** to impute null values of numerical columns having outliers - **AMT_GOODS_PRICE, AMT_ANNUITY**

```
sns.boxplot(application_data.CNT_FAM_MEMBERS)
plt.show()
```

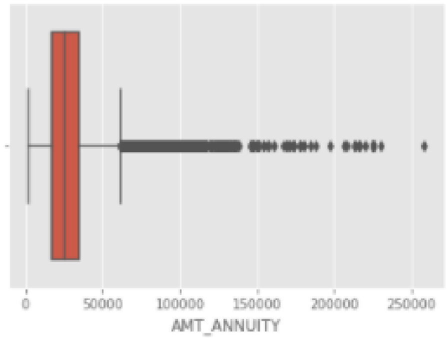



```
#Since these are numerical columns and have no outliers, so using mean to impute missing values
application_data['CNT_FAM_MEMBERS'].fillna((application_data['CNT_FAM_MEMBERS'].mean()), inplace=T
```

```
sns.boxplot(application_data.AMT_GOODS_PRICE)
plt.show()
```

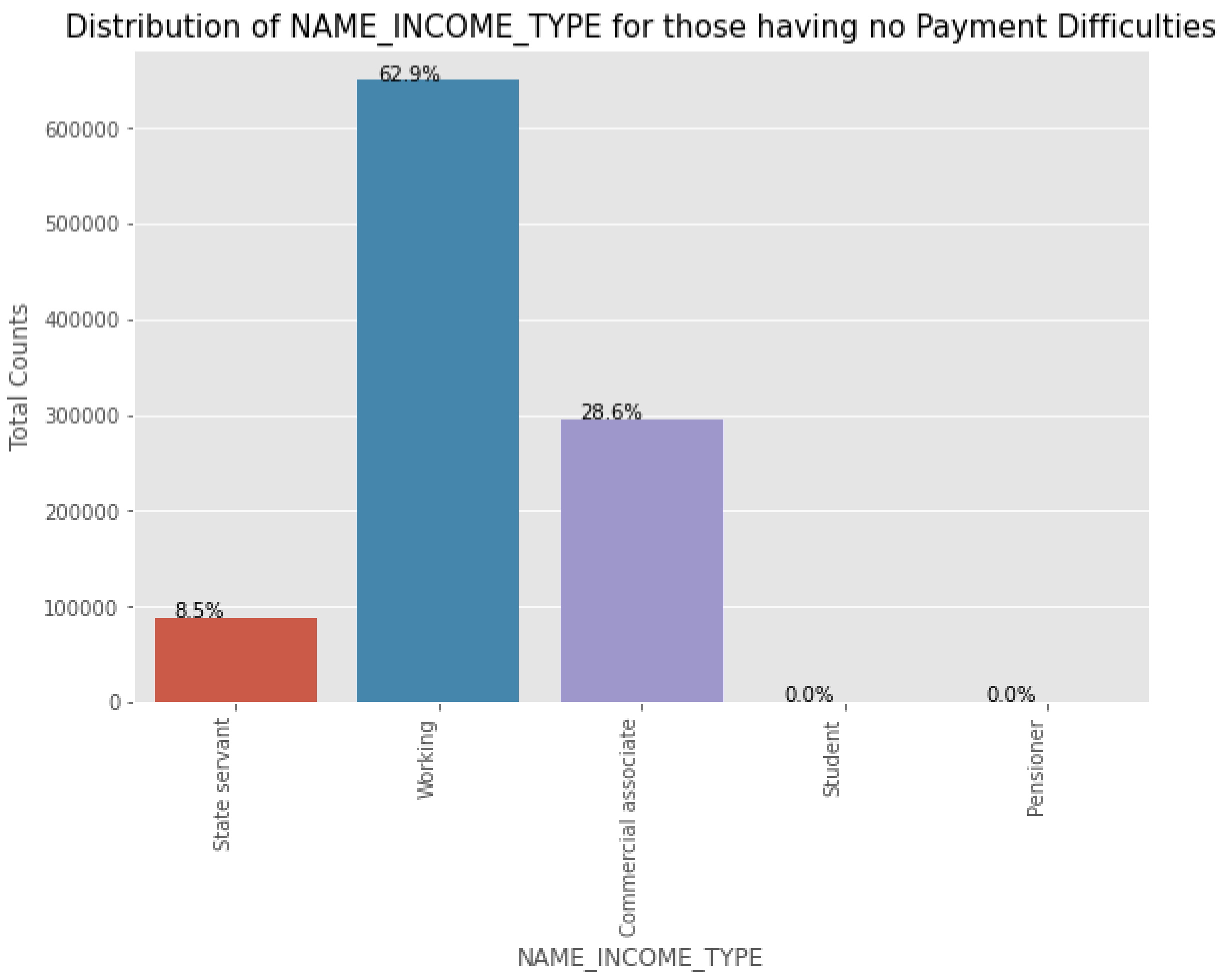


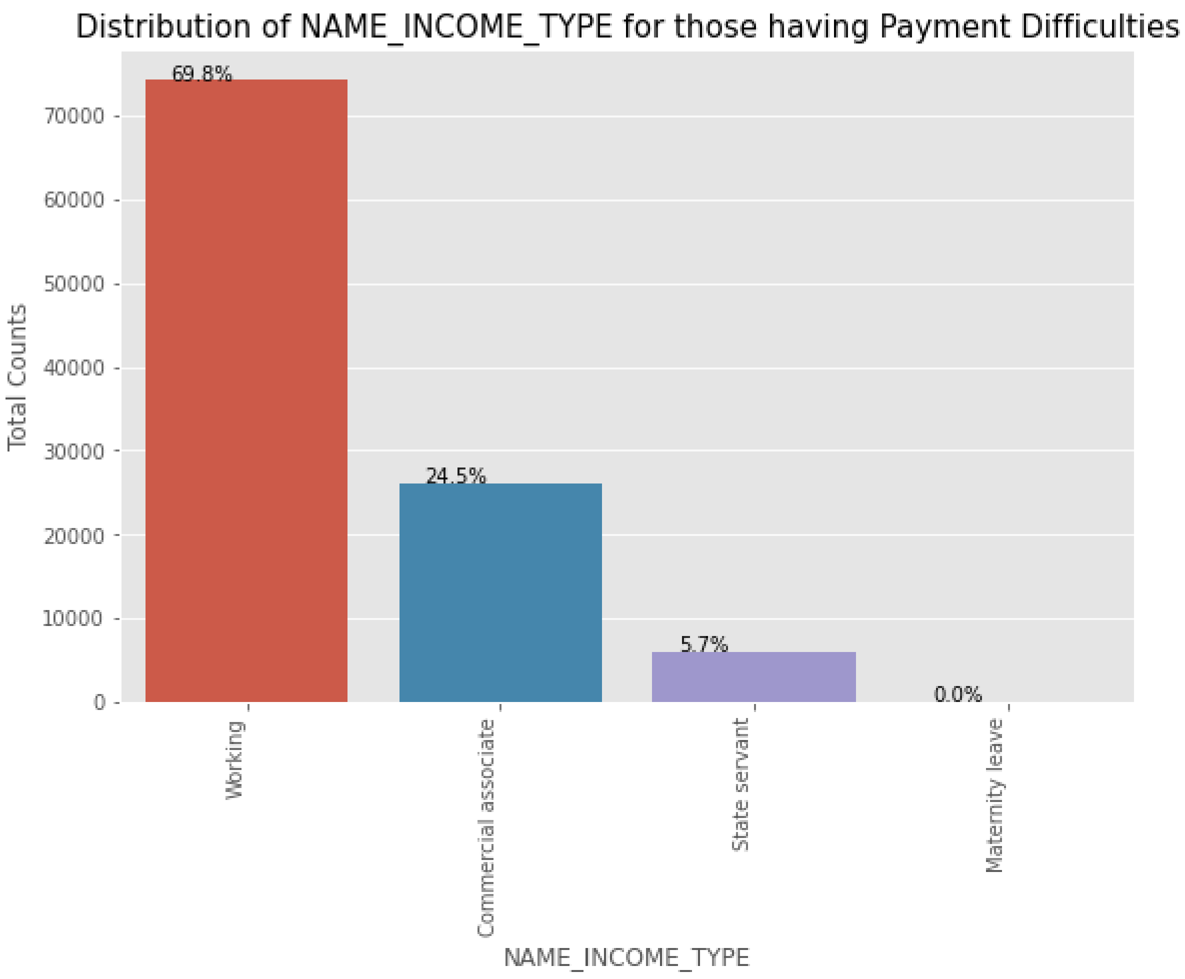

```
sns.boxplot(application_data.AMT_ANNUITY)
plt.show()
```




```
#Since these are numerical columns and have outliers, so using median to impute missing values
```

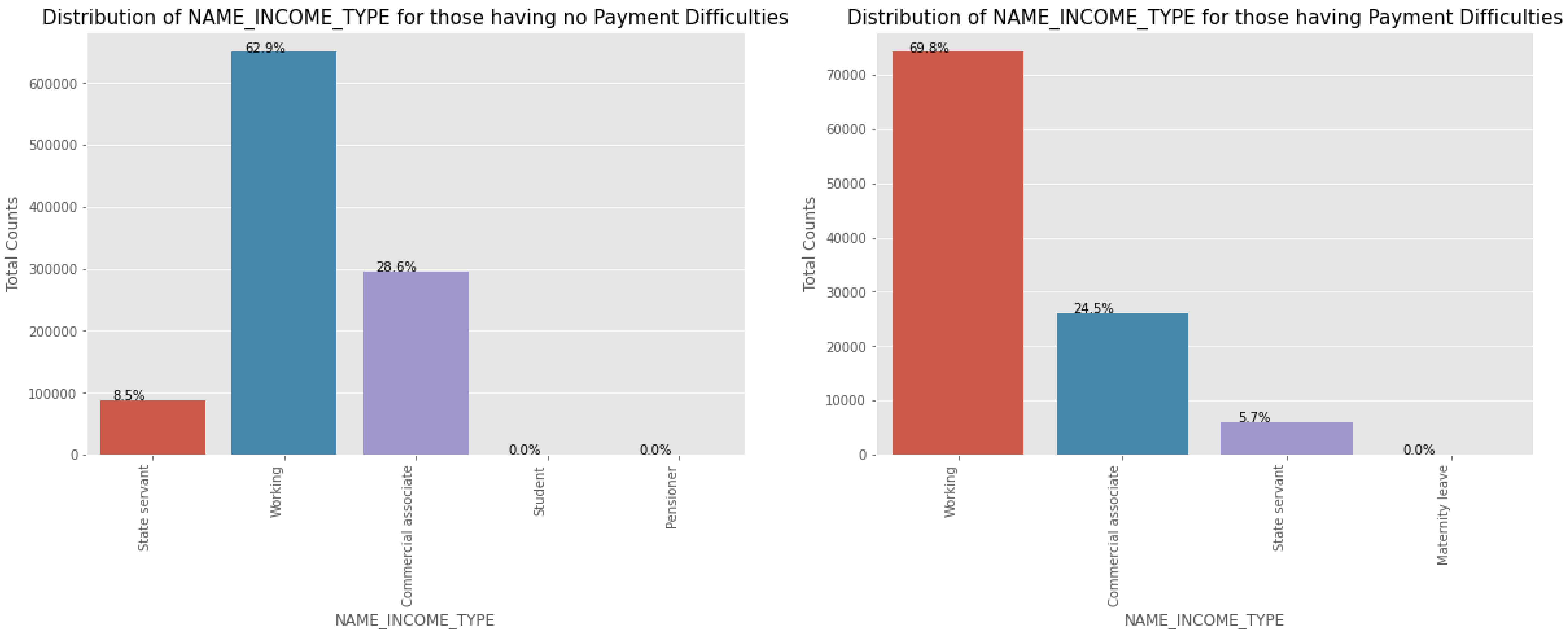
Analysis – Univariate Categorical





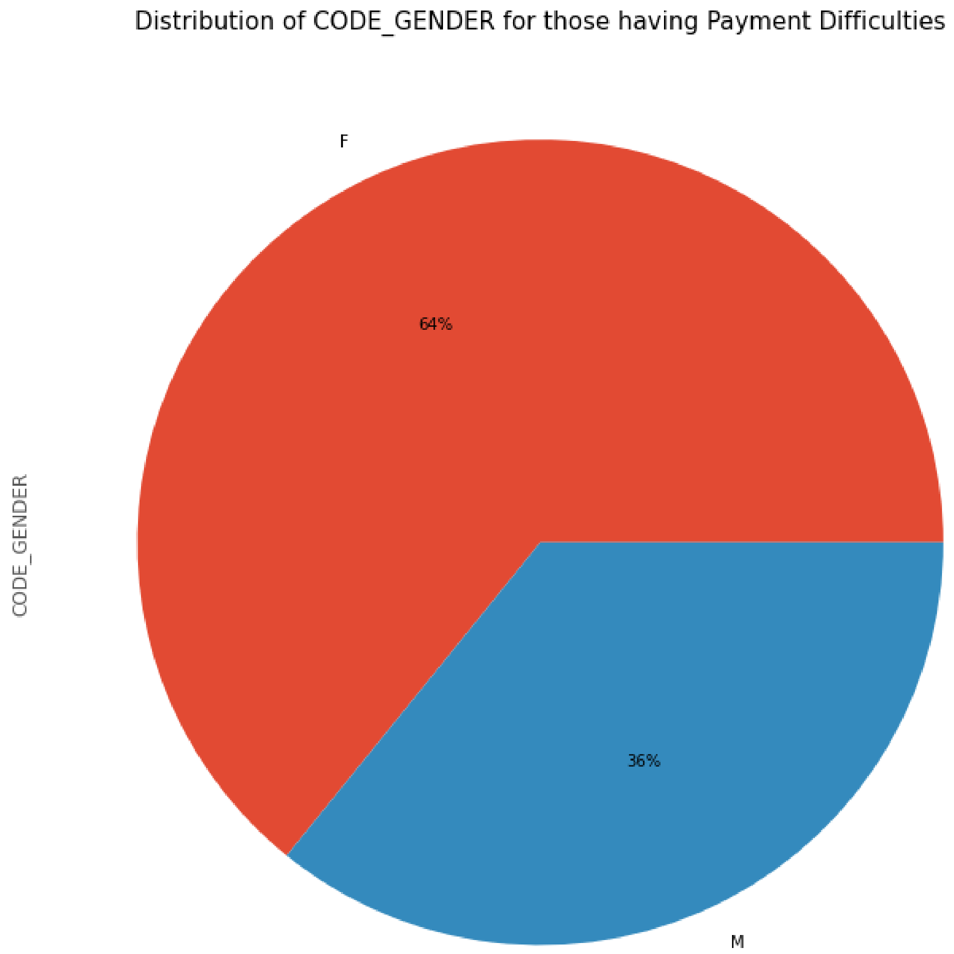
#Insights - State Servant and Commercial Associates have less Payment Difficulties and overall they are tapped less so they should be targetted more compared to working ones where we find more payment difficulties.

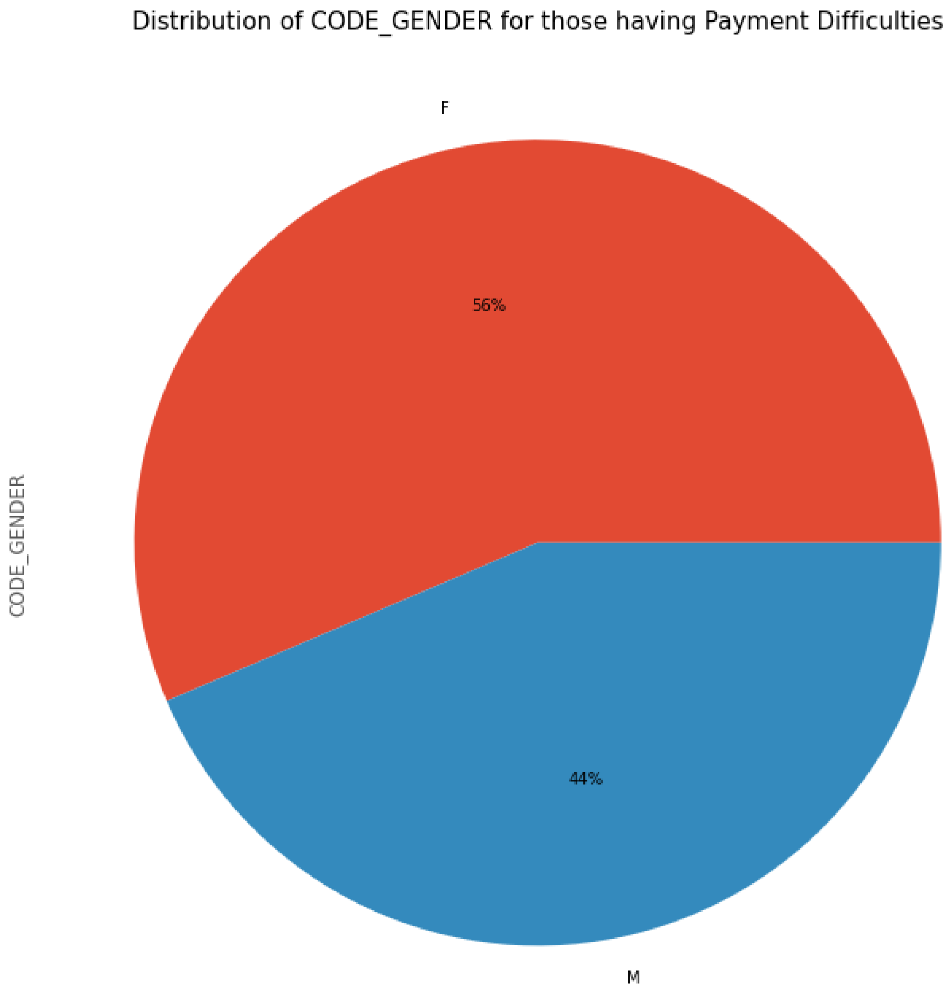
Analysis – Univariate Categorical



#Insights - percentage of loans approved to clients having payment difficulties is much higher, needs to be reduced.

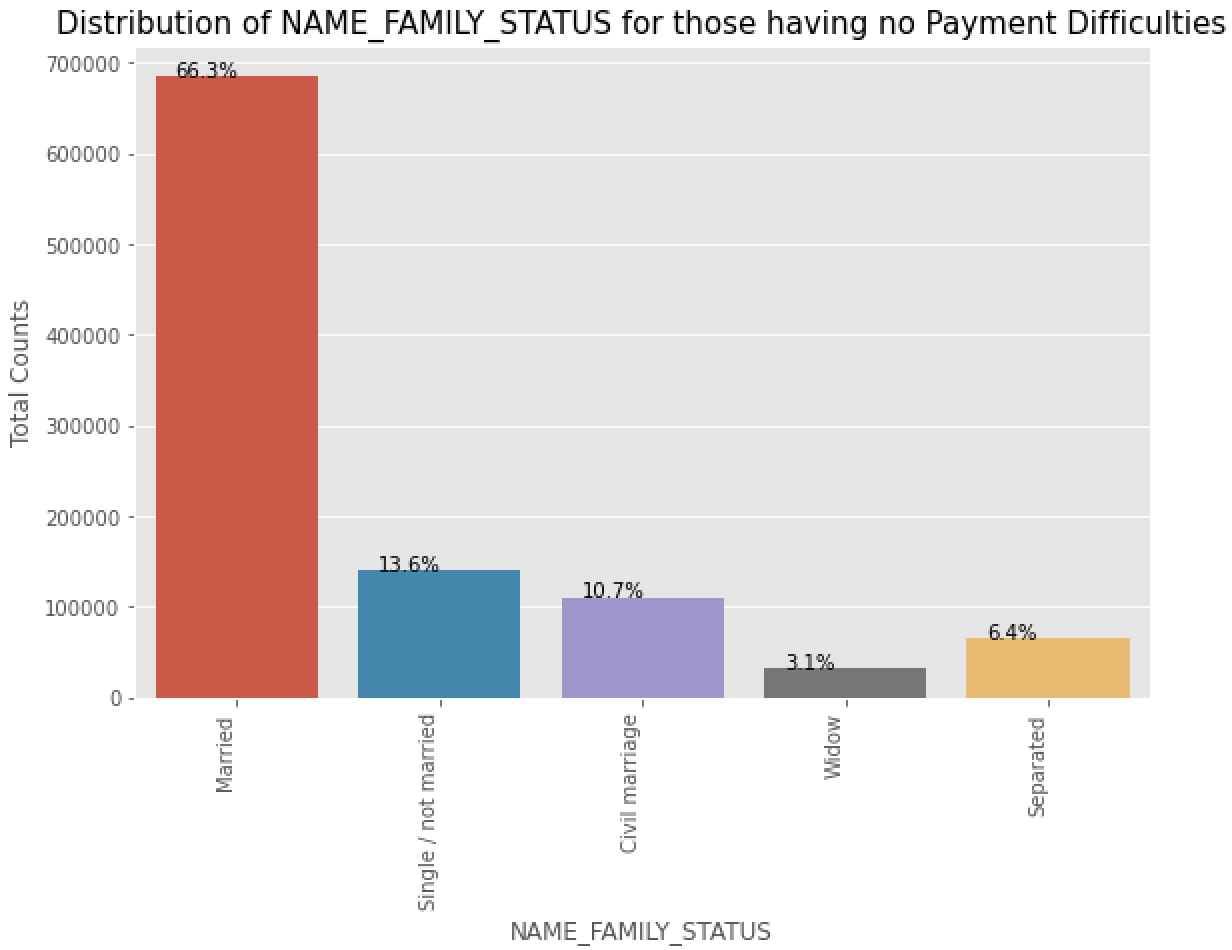
Analysis – Univariate Categorical

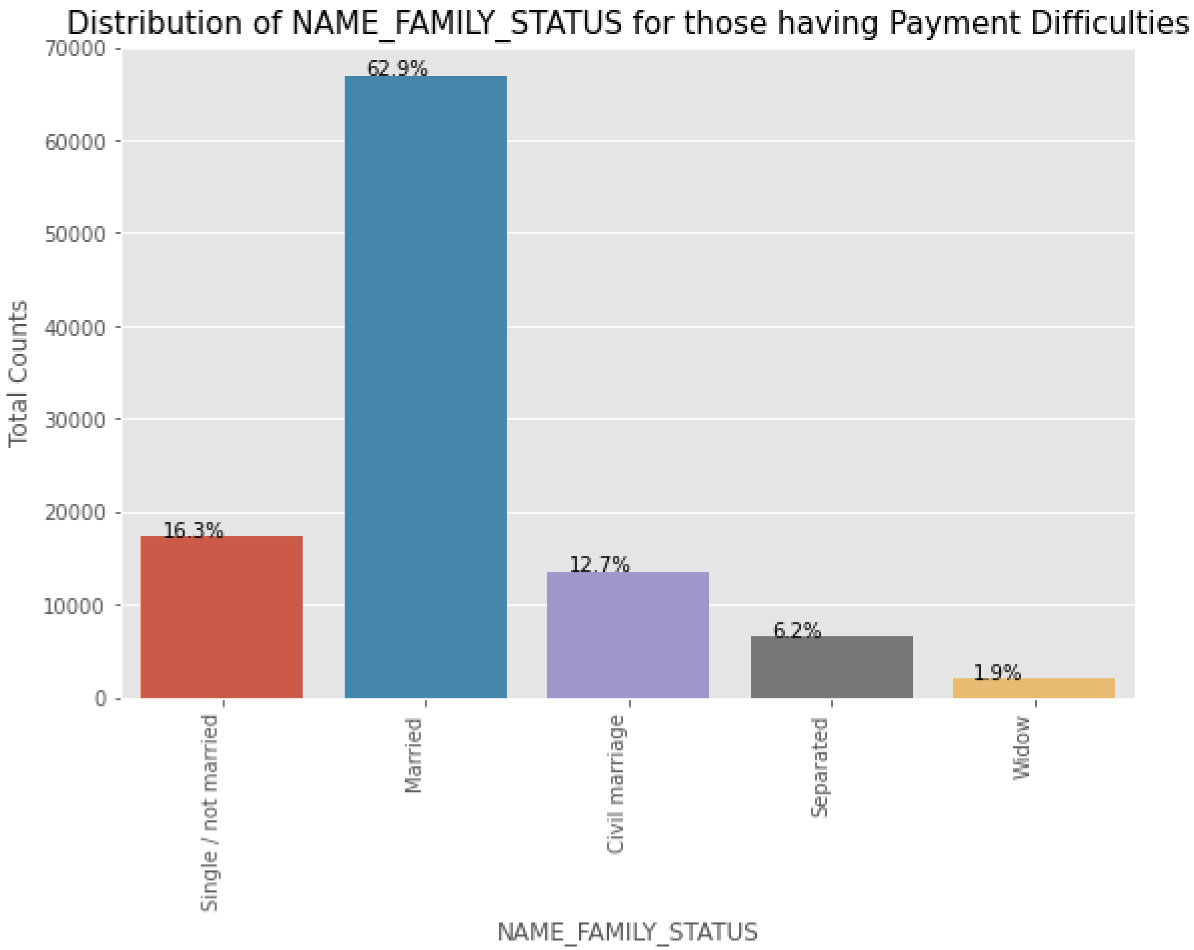




#Insights - Male members have less payment difficulties compared to females, as females are given more loans compared to male.

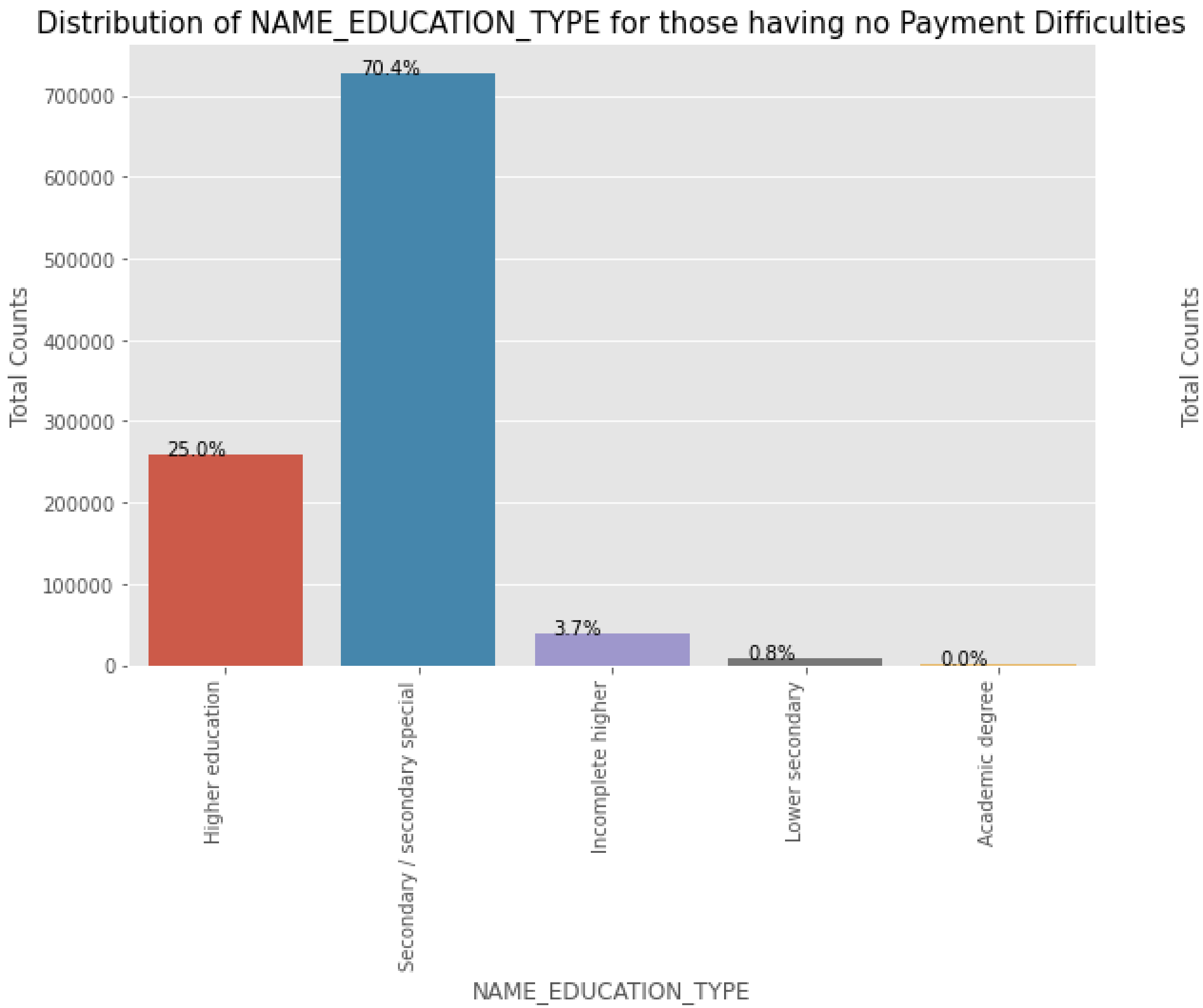
Analysis – Univariate Categorical

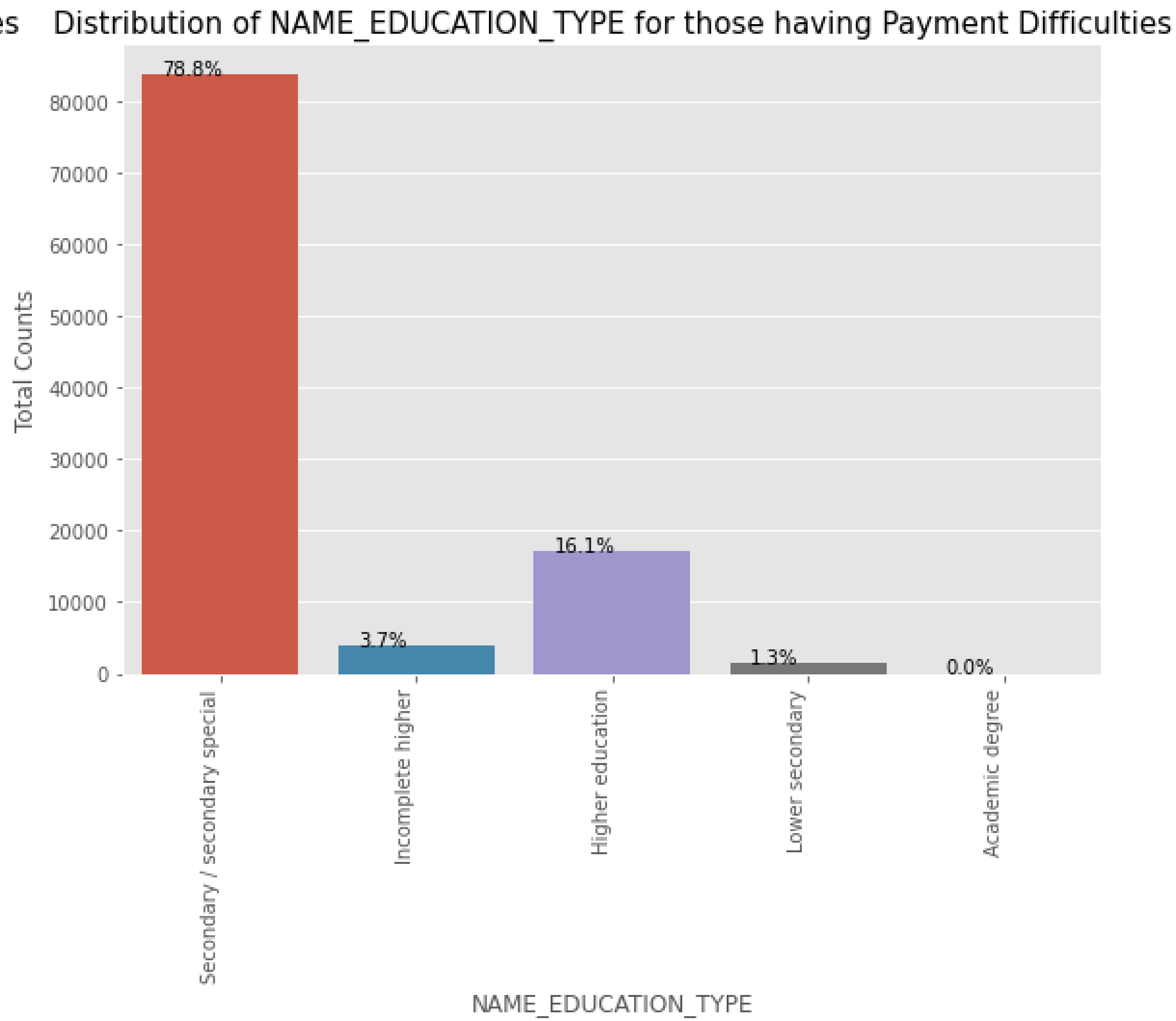




#Insights - though percentage of loan given to single persons & civil marriage is less - rejected applications are more,
#but comparatively they have high percentage of payment difficulties so they should be tapped less.
#loans given to widow have low payment difficulty percentage, so that segment can be tapped more.

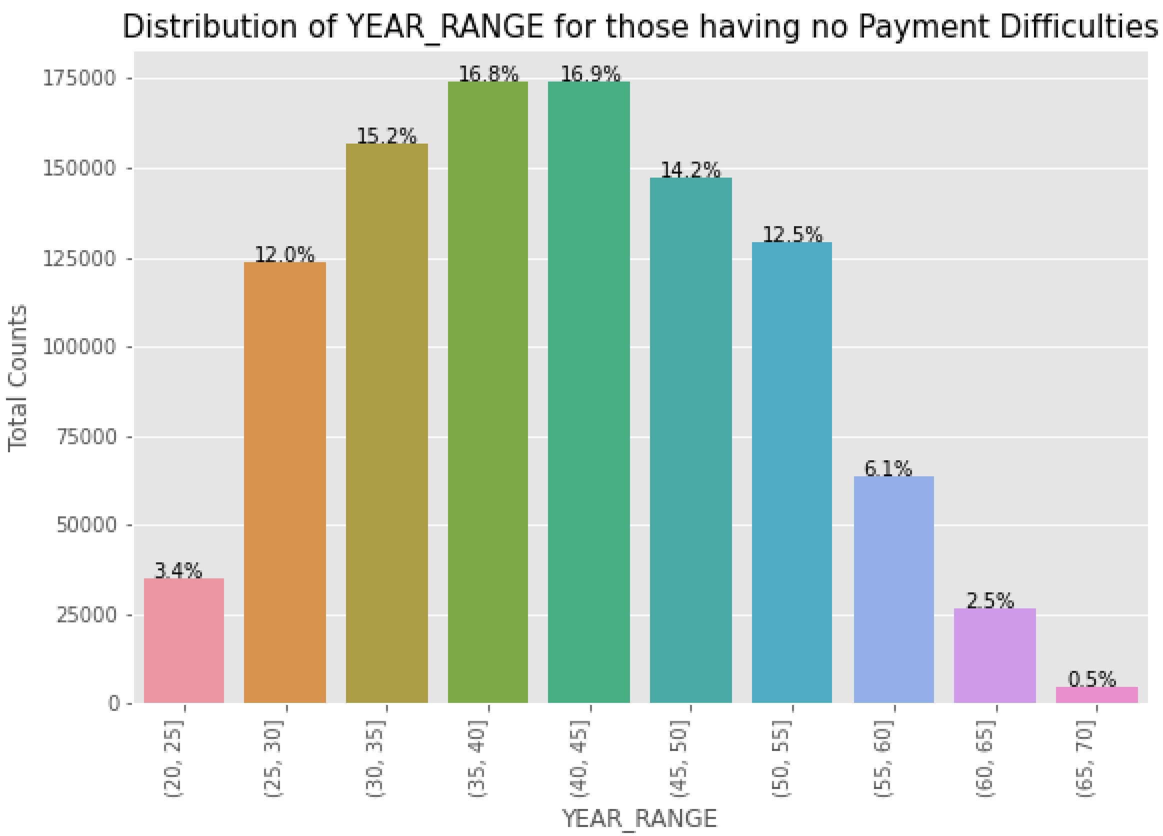
Analysis – Univariate Categorical

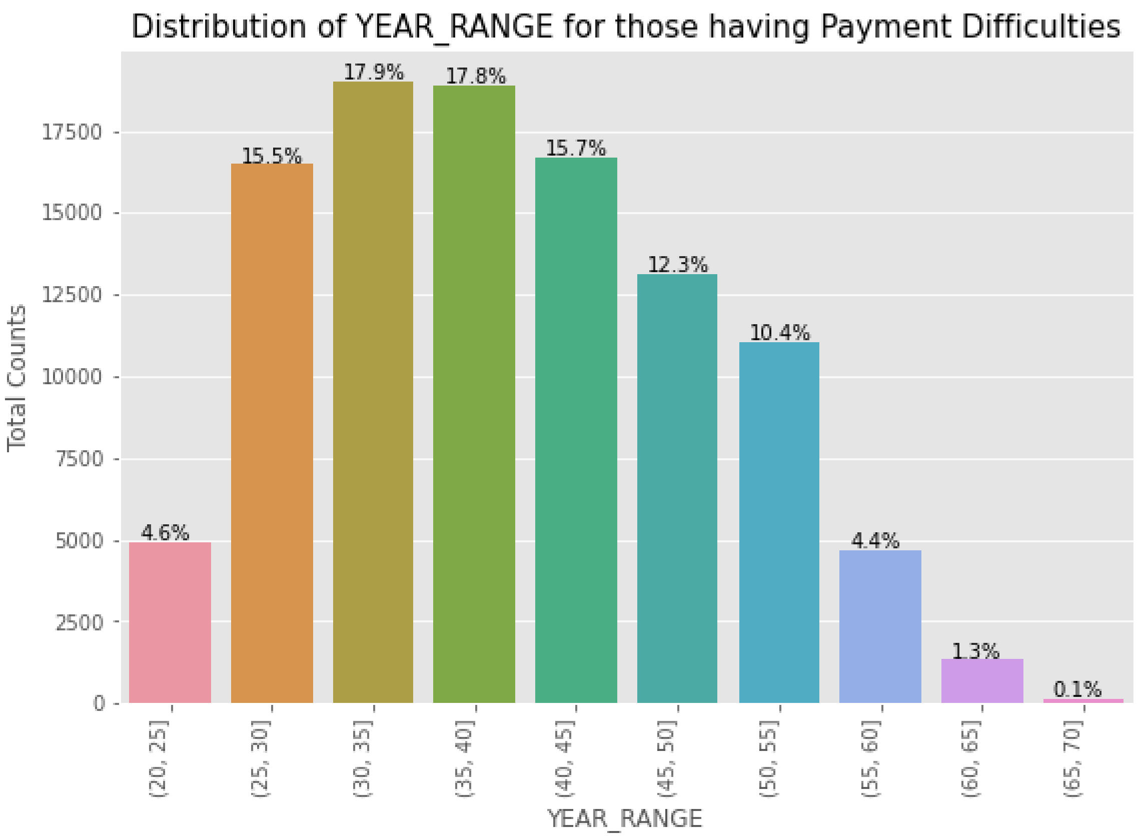




#Insights - loans given to those having higher education are low in number and
#percentage of payment difficulties is less for them so they should be tapped more.
#Secondary Education people are having more payment difficulties, so they should be tapped less.

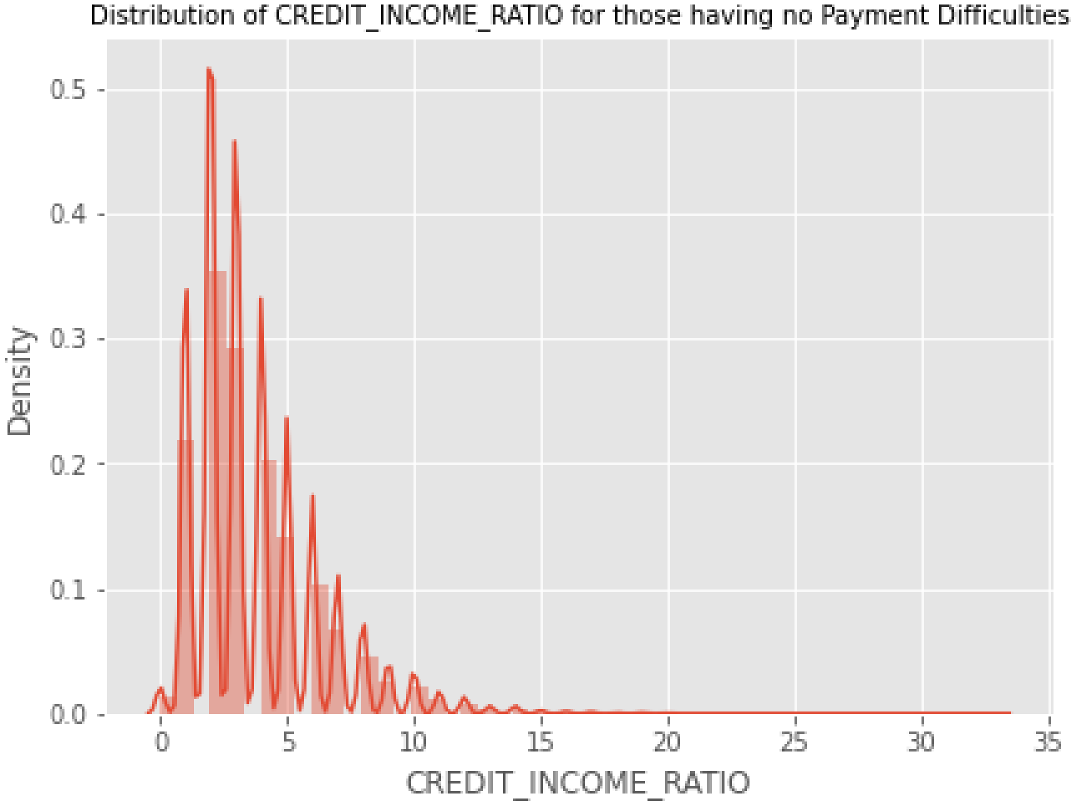
Analysis – Univariate Categorical

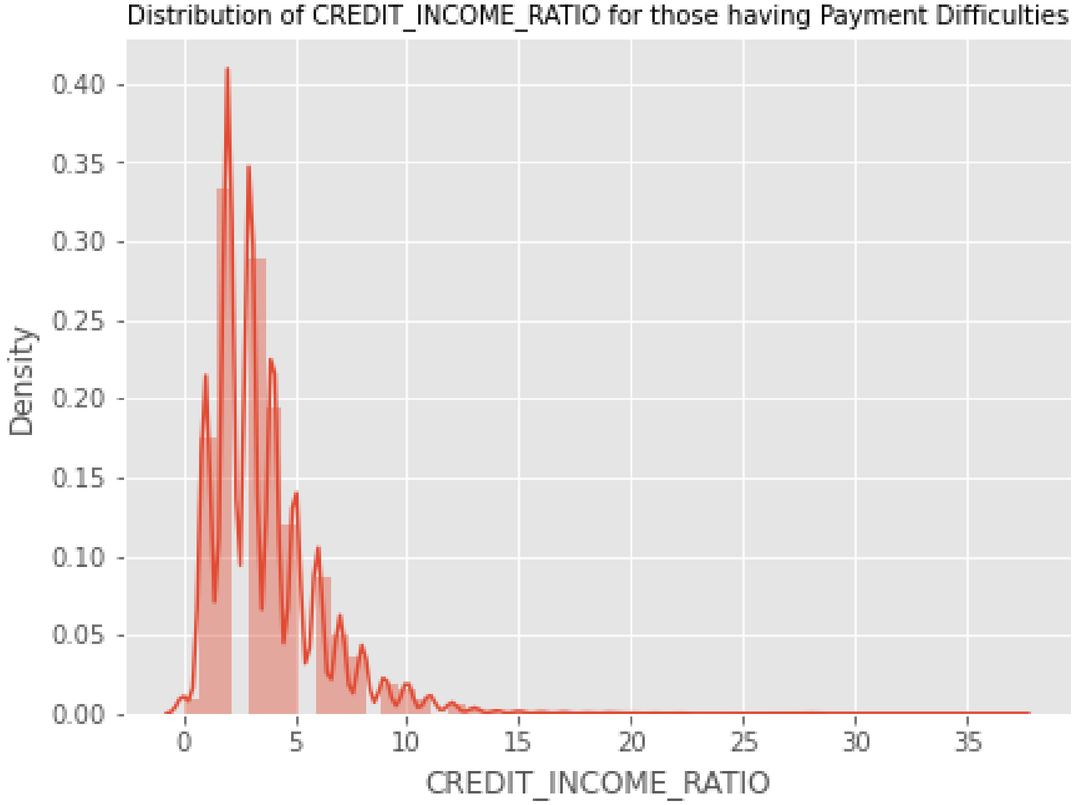




#Insights - Age Group (30-40) have more difficuties in payment compared to those in Age Group (45-55)

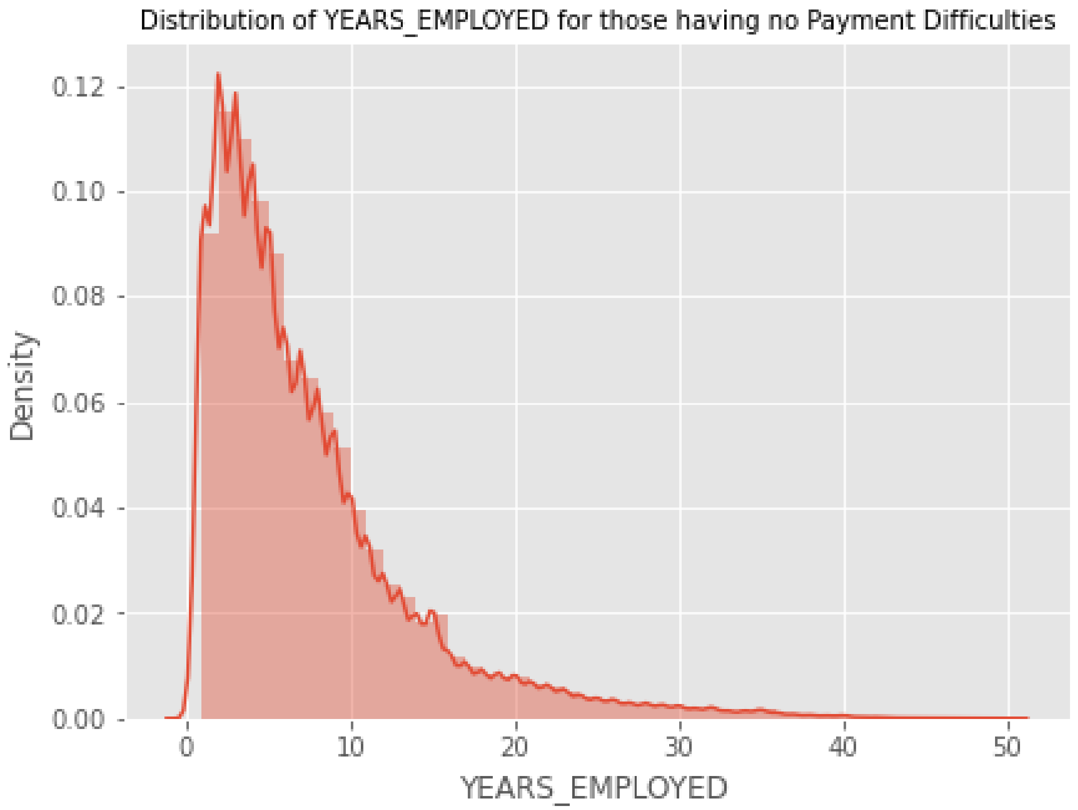
Analysis – Univariate Numerical

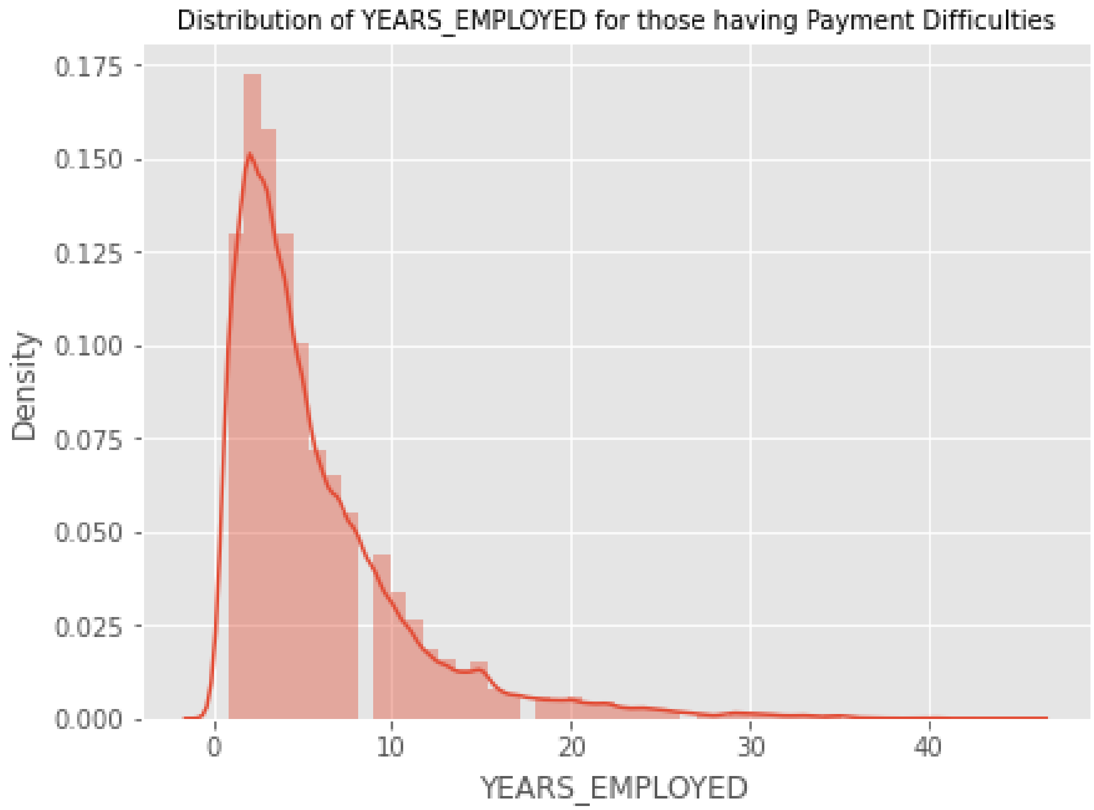




#Insights - When CREDIT_INCOME_RATIO is approx 2.5, more poeple have payment difficulties.

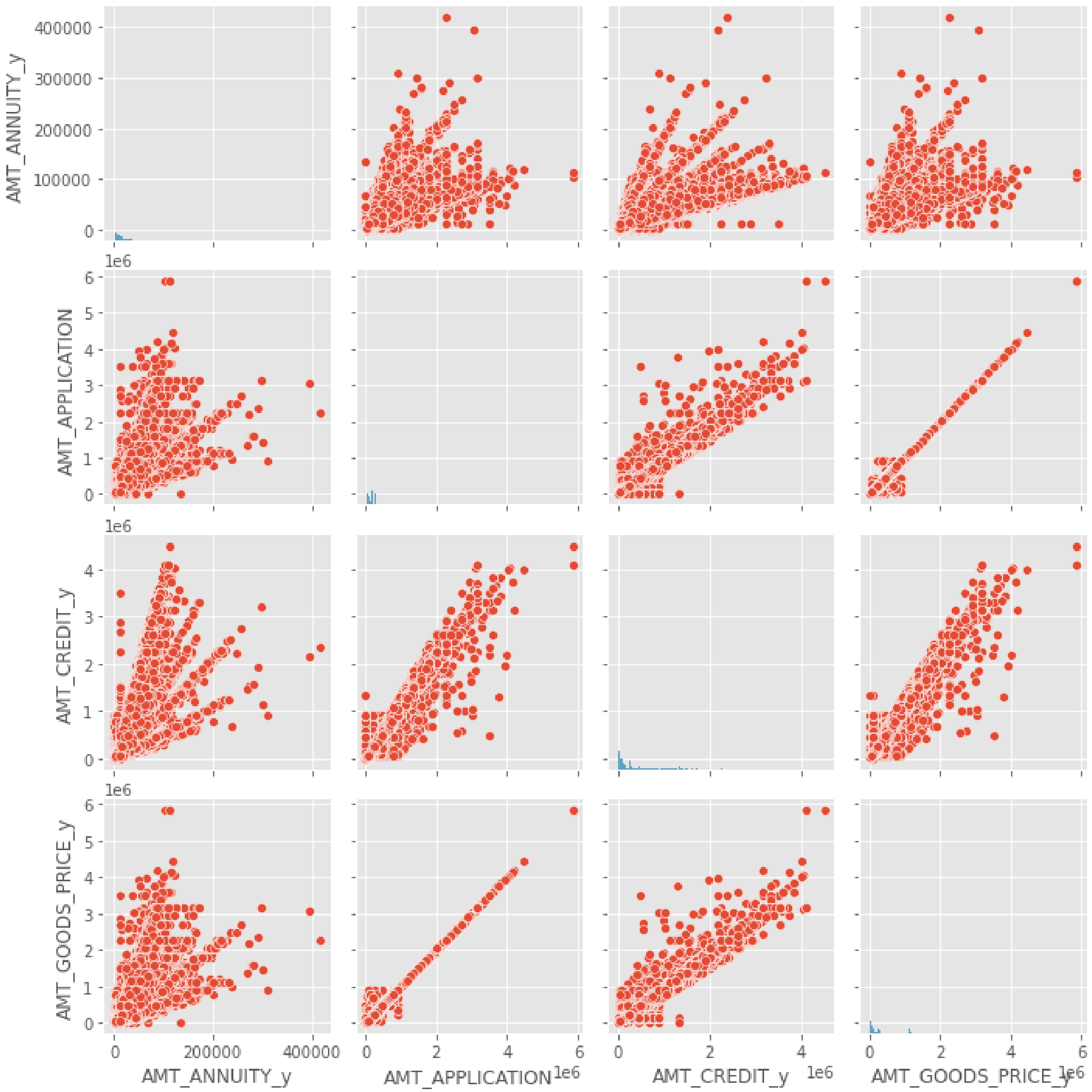
Analysis – Univariate Numerical



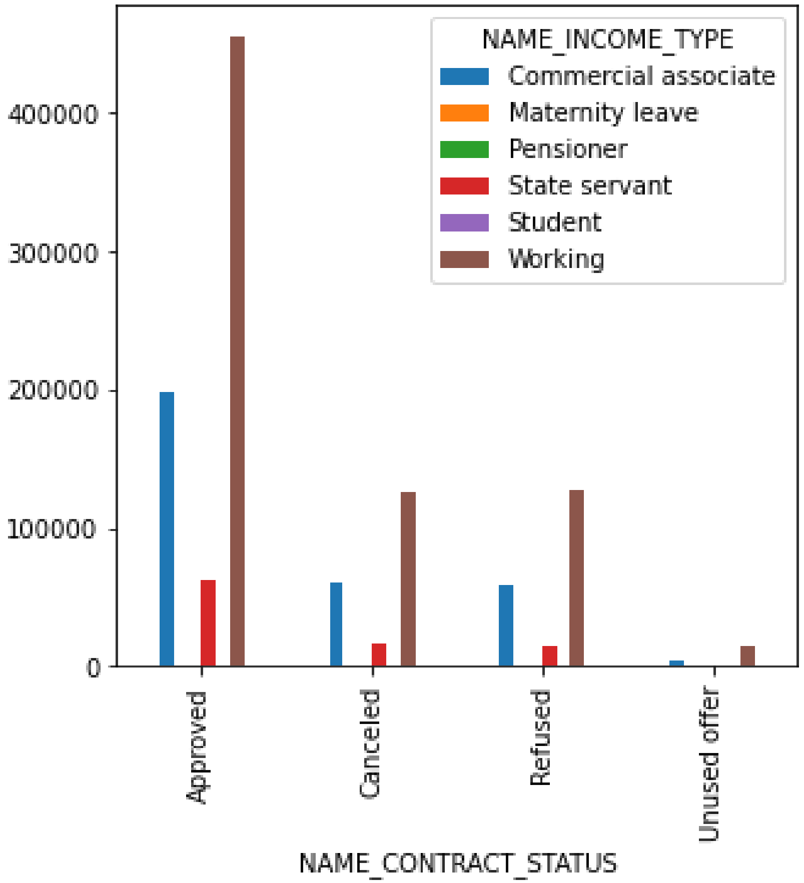


Analysis – Bivariate Numerical

#Insights - Annuity having high positive (directly proportion) impact on :
credit asked by client, credit approved,
price of goods for which loan was taken
credit asked by client is directly proportional to price of good for which loan is being taken

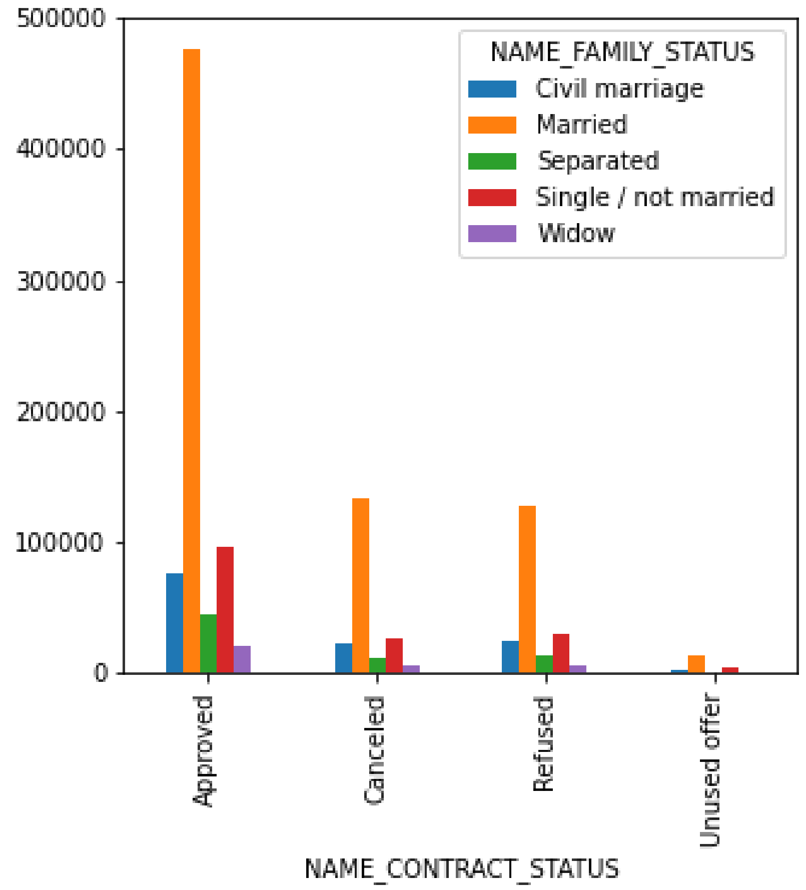


Analysis – Bivariate Categorical



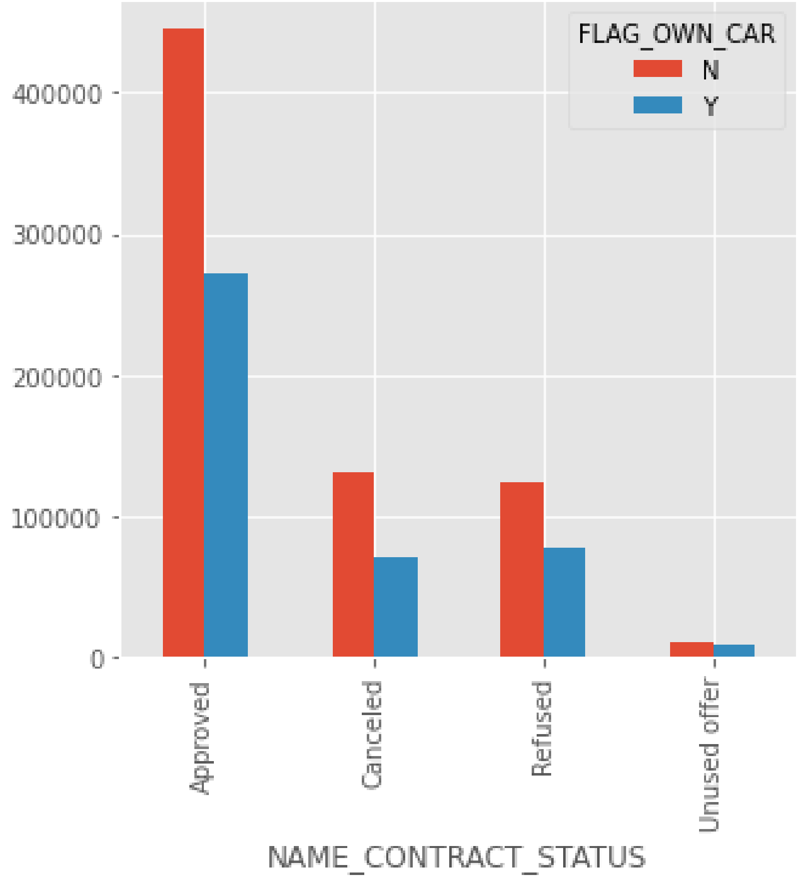
NAME_INCOME_TYPE NAME_CONTRACT_STATUS	Commercial associate	Maternity leave	Pensioner	State servant	Student	Working
Approved	198507	10	44	61630	20	455720
Canceled	59785	2	14	15679	3	126282
Refused	58117	3	26	15597	1	127832
Unused offer	5072	1	0	1518	0	14255

Analysis – Bivariate Categorical



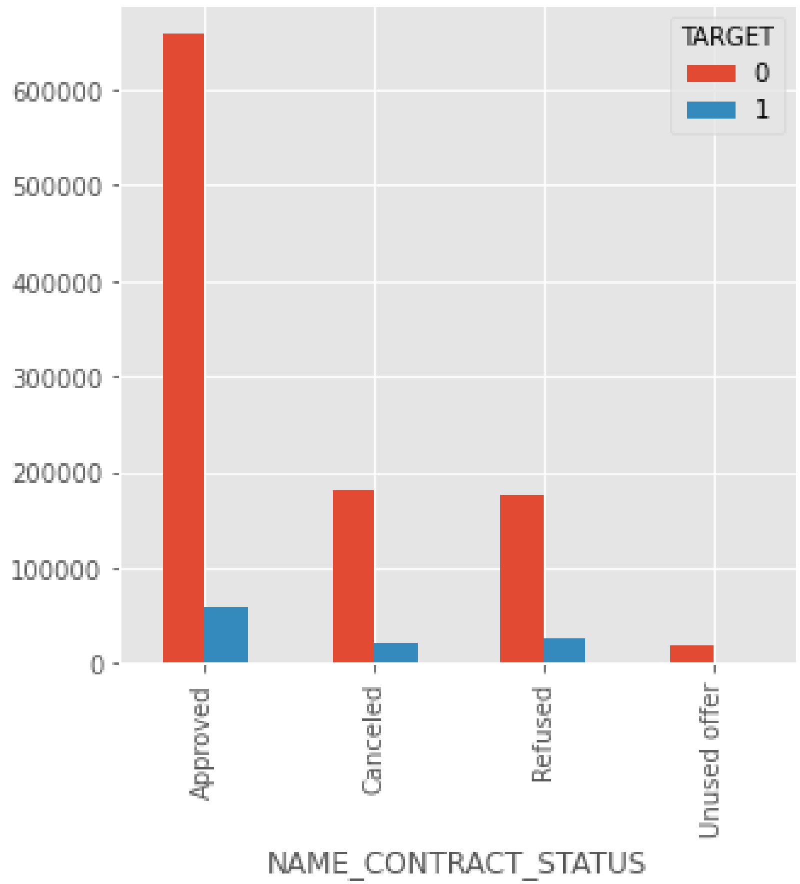
NAME_FAMILY_STATUS NAME_CONTRACT_STATUS	Civil marriage	Married	Separated	Single / not married	Widow
Approved	75455	477237	45155	97113	20971
Canceled	22098	133590	12605	27209	6263
Refused	24500	128003	13177	29453	6443
Unused offer	1862	13313	1448	3816	407

Analysis – Bivariate Categorical



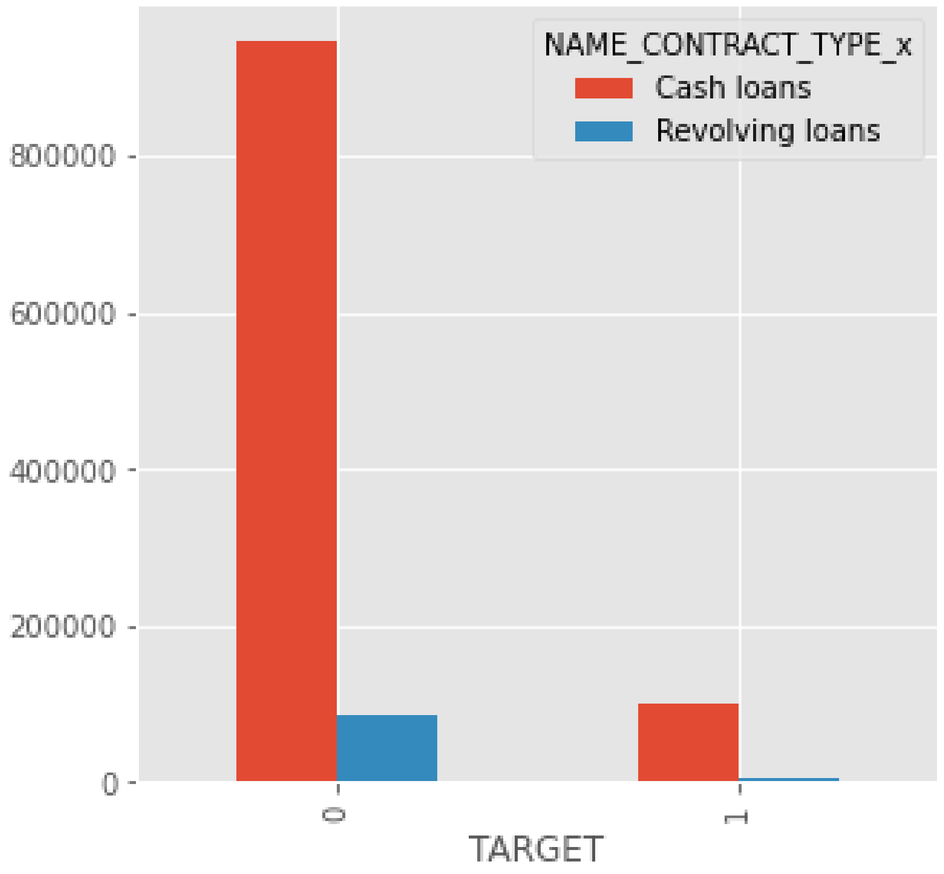
FLAG_OWN_CAR	N	Y
NAME_CONTRACT_STATUS		
Approved	444448	271483
Canceled	130771	70994
Refused	124214	77362
Unused offer	11019	9827

Analysis – Bivariate Categorical



TARGET	0	1
NAME_CONTRACT_STATUS		
Approved	657634	58297
Canceled	181483	20282
Refused	175597	25979
Unused offer	19069	1777

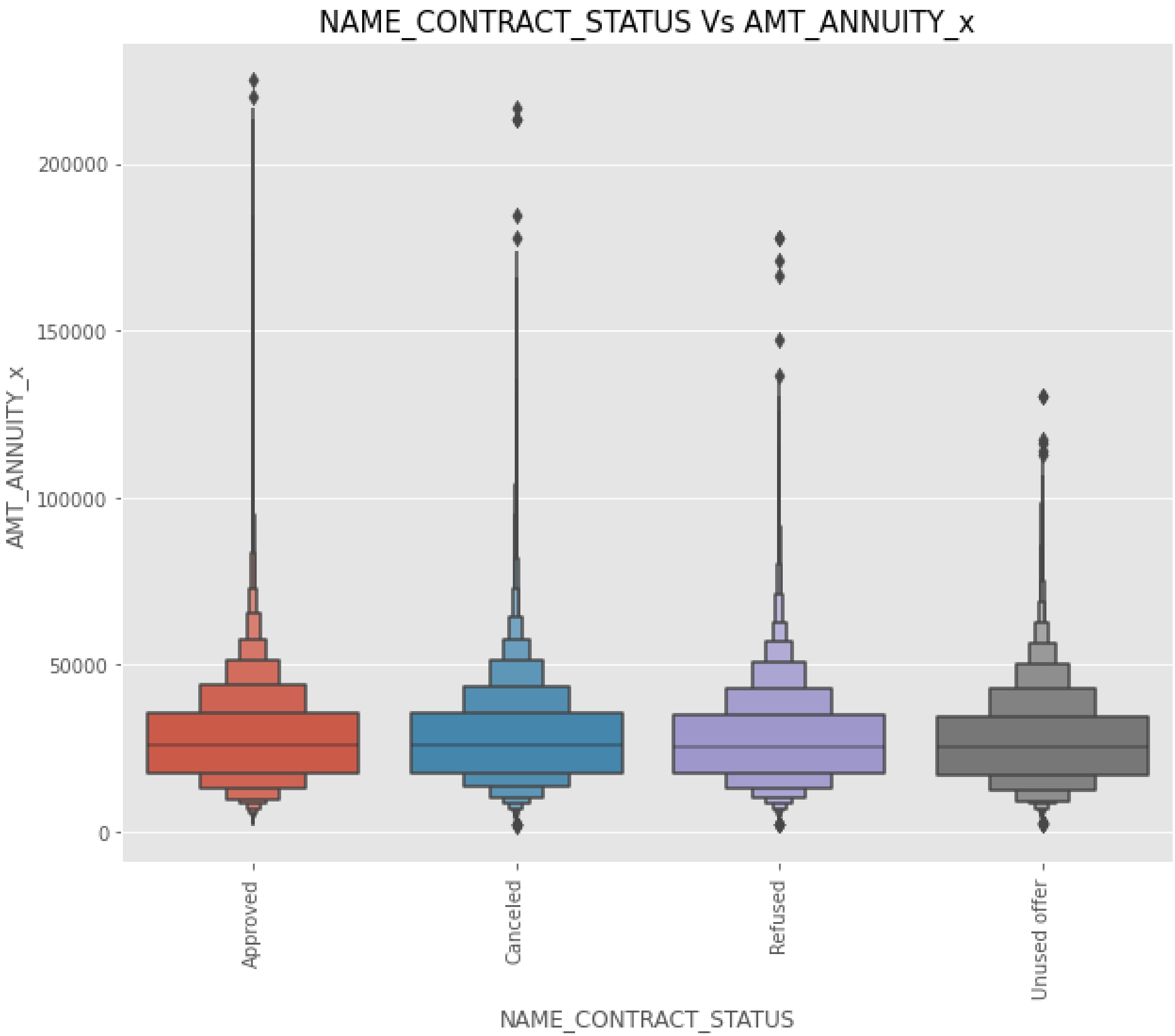
Analysis – Bivariate Categorical



NAME_CONTRACT_TYPE_x	Cash loans	Revolving loans
TARGET		
0	947657	86126
1	101231	5104

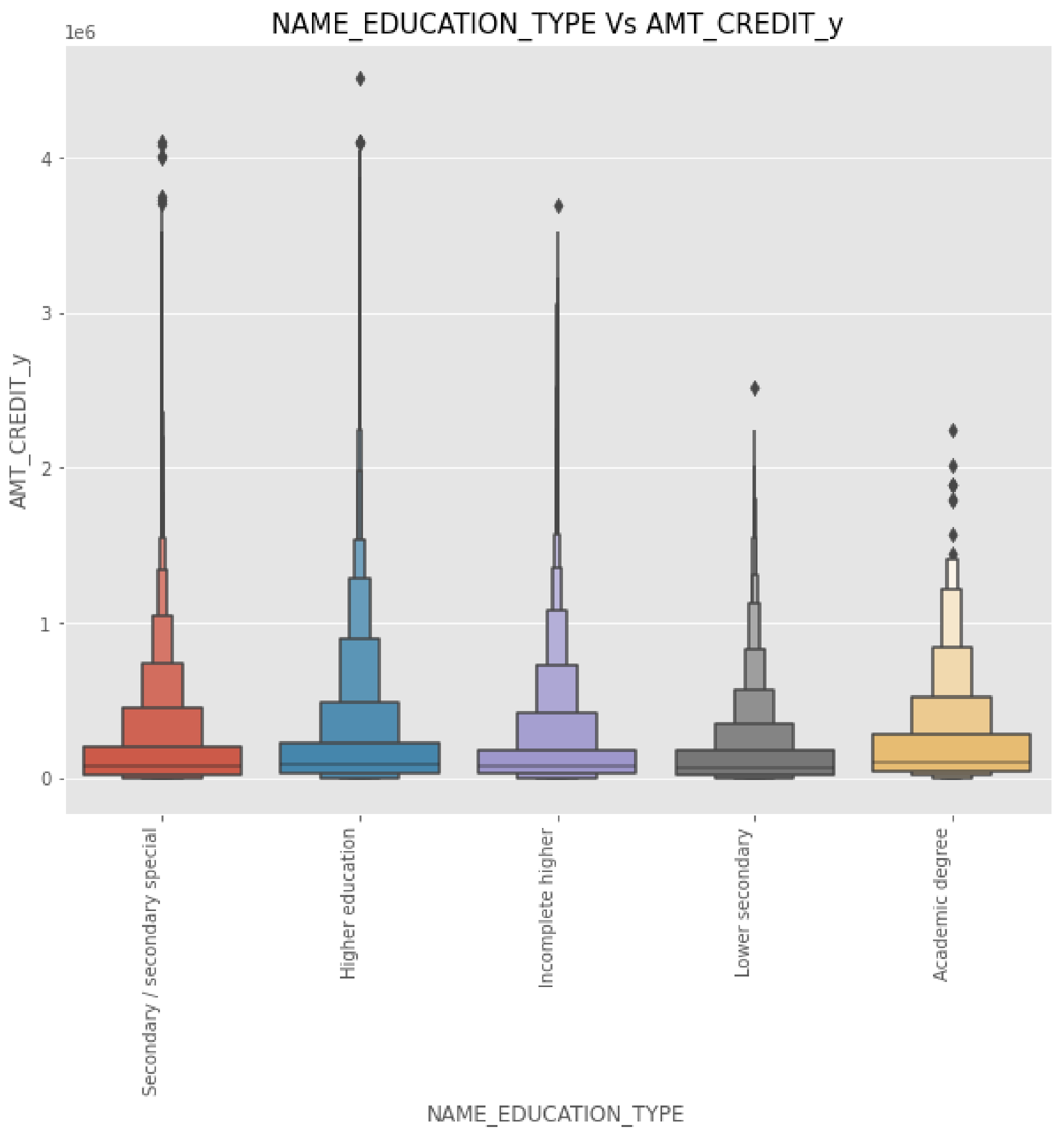
Analysis – Categorical vs Numerical

#Insights - Loan Application of people having too high or low annuity gets rejected more often.



Analysis – Categorical vs Numerical

#Insights - Median Amt Credit seems
to be equal for all but 75 percentile onwards
AMT_CREDIT varies a lot
with each NAME_EDUCATION_TYPE



Correlation

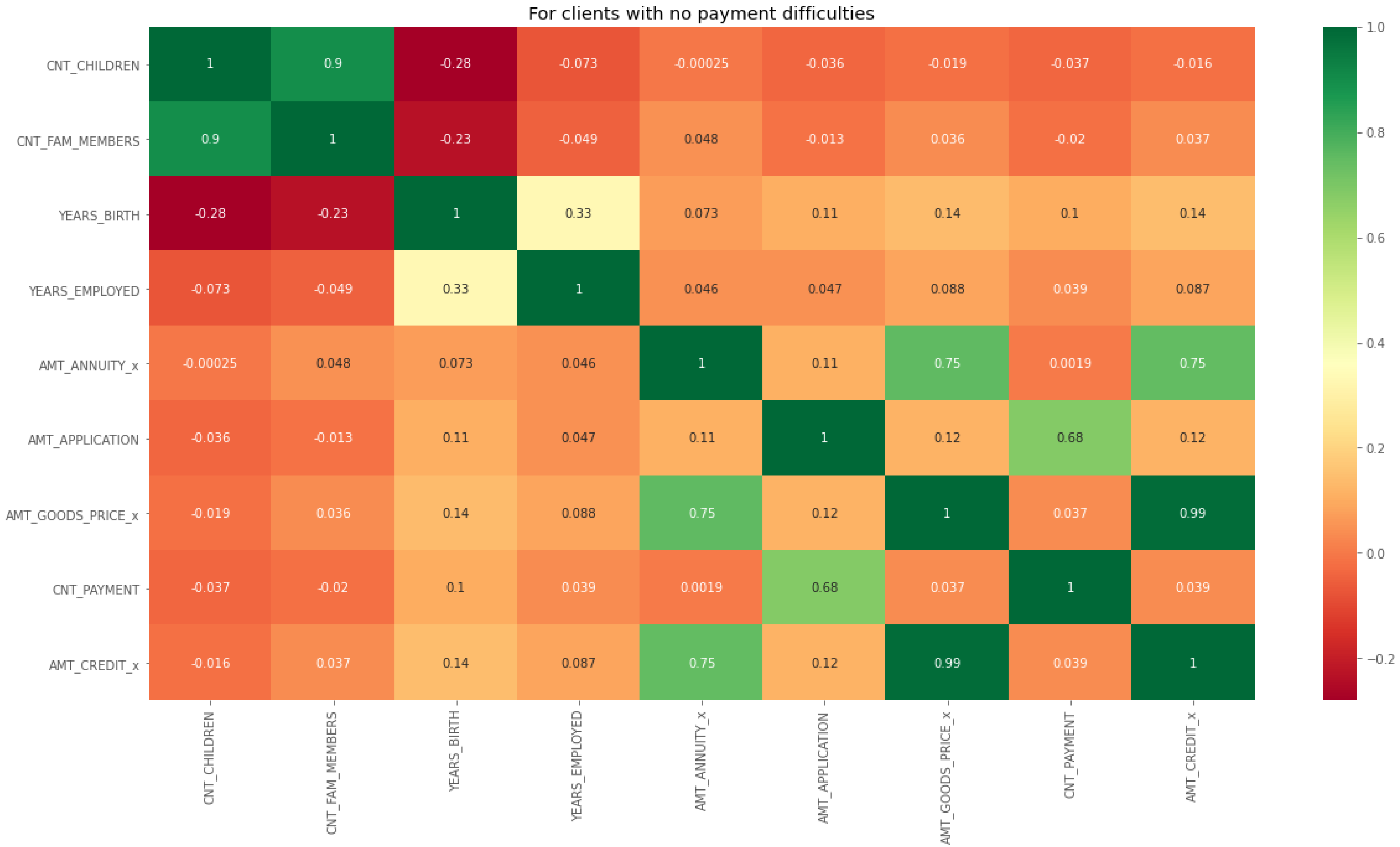
```
#Getting the top 10 correlation in TARGET_0
Top10Corr=TARGET_0.corr()
Top10Corr_df = Top10Corr.where(np.triu(np.ones(Top10Corr.shape),k=1).astype(np.bool)).unstack().reset_index()
Top10Corr_df.columns=['Column1','Column2','Correlation']
Top10Corr_df.dropna(subset=['Correlation'],inplace=True)
Top10Corr_df['Abs_Correlation']=Top10Corr_df['Correlation'].abs()
Top10Corr_df = Top10Corr_df.sort_values(by=['Abs_Correlation'], ascending=False)
Top10Corr_df.head(10)
```

	Column1	Column2	Correlation	Abs_Correlation
916	AMT_GOODS_PRICE_y	AMT_APPLICATION	0.987969	0.987969
202	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.985643	0.985643
883	AMT_CREDIT_y	AMT_APPLICATION	0.973433	0.973433
917	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.970464	0.970464
339	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.943188	0.943188
266	CNT_FAM_MEMBERS	CNT_CHILDREN	0.895207	0.895207
475	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.874522	0.874522
577	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.825895	0.825895
915	AMT_GOODS_PRICE_y	AMT_ANNUITY_y	0.814021	0.814021
882	AMT_CREDIT_y	AMT_ANNUITY_y	0.812540	0.812540

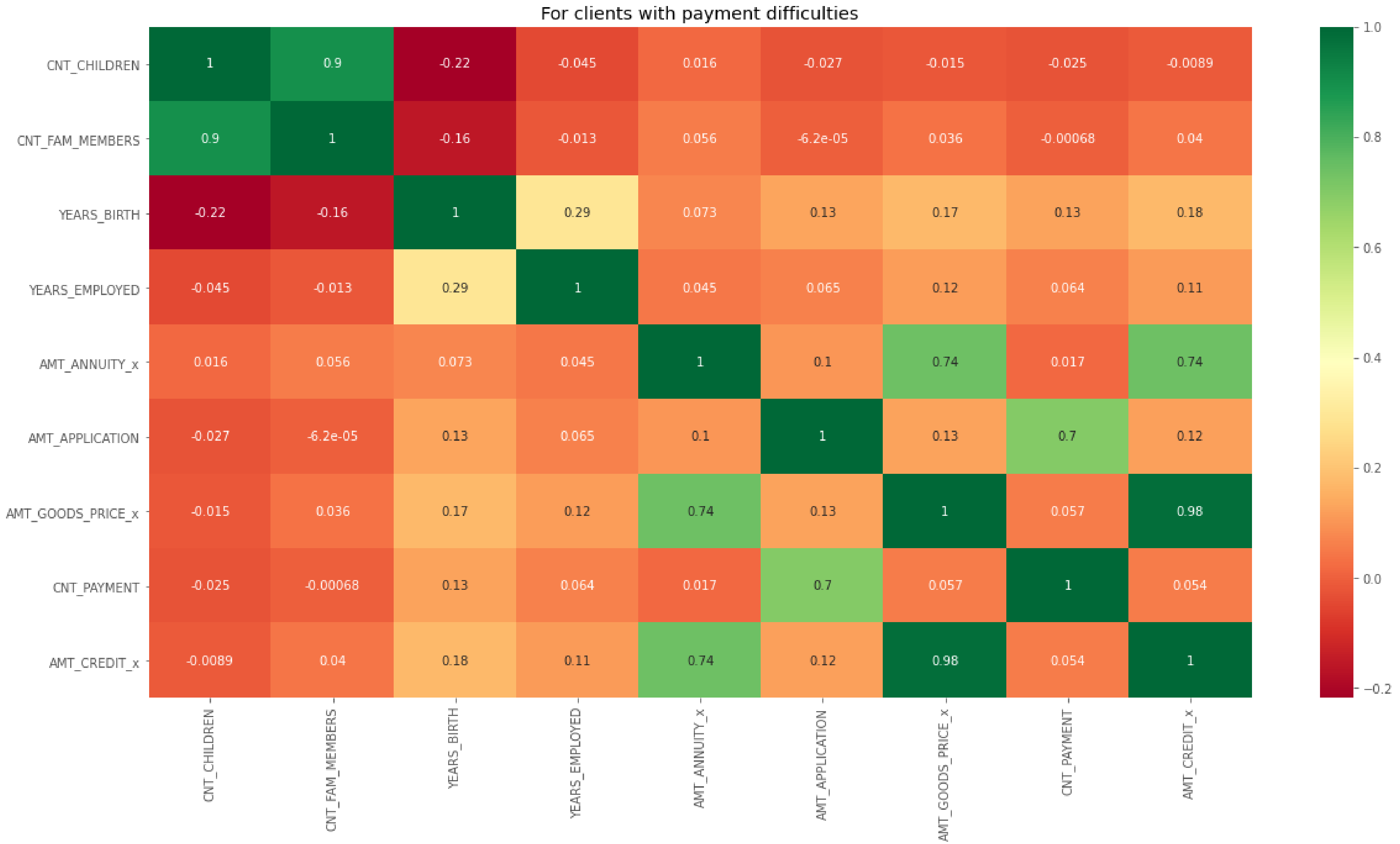
```
#Getting the top 10 correlation in TARGET_1
Top10Corr=TARGET_1.corr()
Top10Corr_df = Top10Corr.where(np.triu(np.ones(Top10Corr.shape),k=1).astype(np.bool)).unstack().reset_index()
Top10Corr_df.columns=['Column1','Column2','Correlation']
Top10Corr_df.dropna(subset=['Correlation'],inplace=True)
Top10Corr_df['Abs_Correlation']=Top10Corr_df['Correlation'].abs()
Top10Corr_df = Top10Corr_df.sort_values(by=['Abs_Correlation'], ascending=False)
Top10Corr_df.head(10)
```

	Column1	Column2	Correlation	Abs_Correlation
916	AMT_GOODS_PRICE_y	AMT_APPLICATION	0.985620	0.985620
202	AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.981802	0.981802
883	AMT_CREDIT_y	AMT_APPLICATION	0.973595	0.973595
917	AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.967739	0.967739
339	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956500	0.956500
266	CNT_FAM_MEMBERS	CNT_CHILDREN	0.895919	0.895919
475	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.872302	0.872302
915	AMT_GOODS_PRICE_y	AMT_ANNUITY_y	0.830791	0.830791
882	AMT_CREDIT_y	AMT_ANNUITY_y	0.829964	0.829964
849	AMT_APPLICATION	AMT_ANNUITY_y	0.815271	0.815271

Correlation



Correlation



Correlation

#Insights - Heatmap for both TARGET_0 & TARGET_1 are almost same!

#Highest Correlation is between AMT_GOODS_PRICE & AMT_CREDIT

#Lowest Correlation is between CNT_CHILDREN & YEARS_BIRTH

#AMT_CREDIT is inversly proportional to CNT_CHILDREN AND YEARS_BIRTH


THANK YOU