Credit EDA Case Study

By Srilathaa Vasu

Agenda

- Business Objectives
- Data Cleaning
- Imbalance in Data
- Univariate Analysis
- Bi/Multivariate Analysis
- Correlation Analysis
- Analyzing Previous Application Dataset
- Univariate Analysis
- Bi/Multivariate Analysis
- Merged Dataframe Analysis

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.

Data Cleaning

Steps Performed :

1. Import the data and analyse the basic information in the dataset.
2. Remove all the unwanted/unneccesary columns from the dataset.
3. Handle the missing values for some columns and transform those columns into 'int' datatype.
4. Calculate the percentage of null values in each columns. If the percentage of null values is greater than 50%, we remove those columns.
5. Handle missing values by imputation and dropping some values.
   Handle outliers for some columns.

```
#checking remaining null values columns
new_null = app1.isnull().sum()/len(app1)*100
new_null.sort_values(ascending=False).head(20)
```

```
FLOORSMAX_MEDI                  49.760822
FLOORSMAX_MODE                  49.760822
FLOORSMAX_AVG                   49.760822
YEARS_BEGINEXPLUATATION_MEDI    48.781019
YEARS_BEGINEXPLUATATION_MODE    48.781019
YEARS_BEGINEXPLUATATION_AVG     48.781019
TOTALAREA_MODE                  48.268517
EMERGENCYSTATE_MODE             47.398304
OCCUPATION_TYPE                 31.345545
EXT_SOURCE_3                    19.825307
NAME_TYPE_SUITE                  0.420148
EXT_SOURCE_2                     0.214626
AMT_GOODS_PRICE                  0.090403
AMT_ANNUITY                      0.003902
CNT_FAM_MEMBERS                  0.000650
ORGANIZATION_TYPE                0.000000
LIVE_CITY_NOT_WORK_CITY          0.000000
SK_ID_CURR                       0.000000
REG_CITY_NOT_LIVE_CITY           0.000000
OBS_30_CNT_SOCIAL_CIRCLE         0.000000
dtype: float64
```

Imbalance in Data

```
app1['TARGET'].value_counts().plot.bar()
plt.title('Target variable')
plt.show()
```

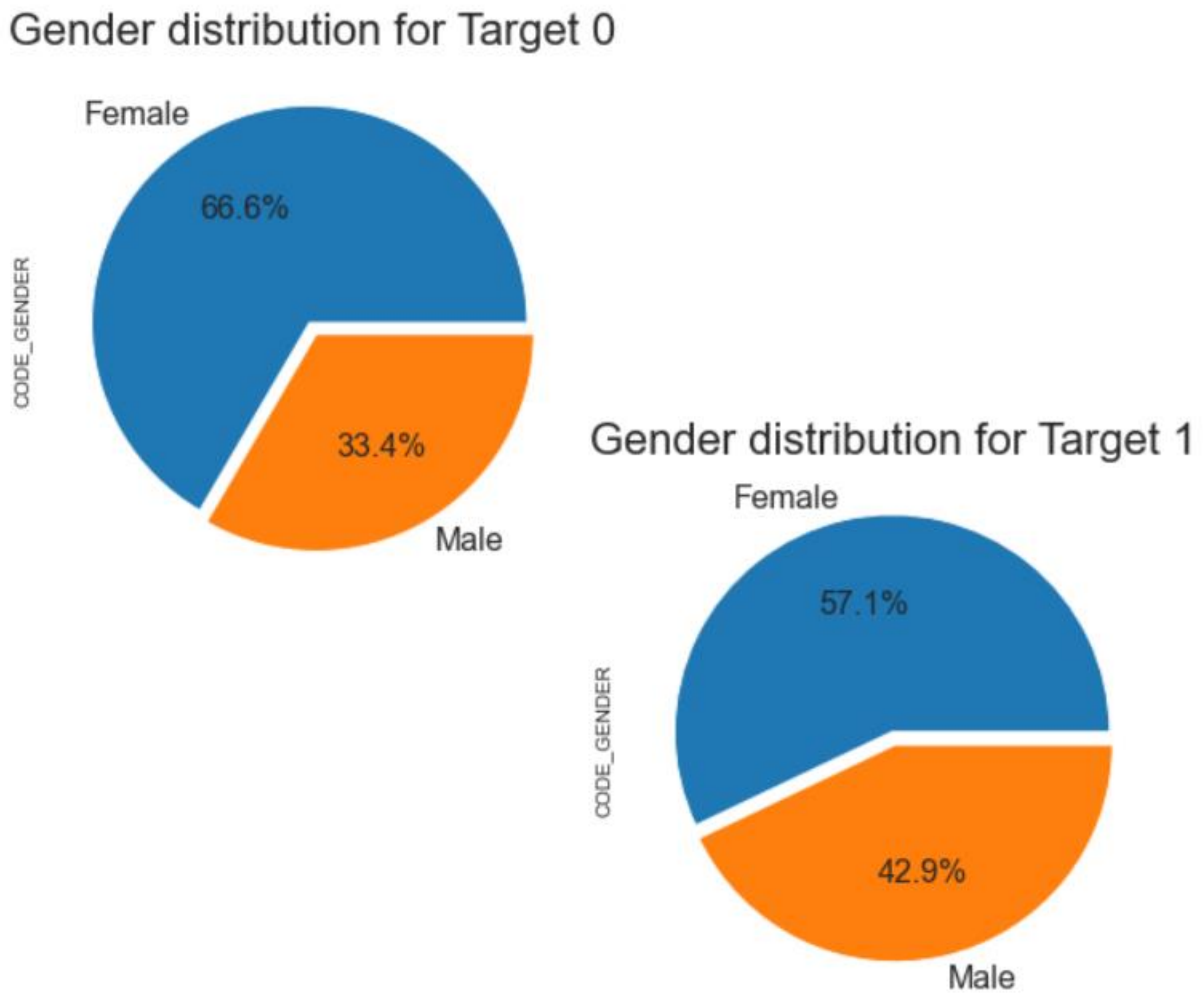
The imbalance is high in target variable.

- Client with no payment difficulties is 91.92%.
- Client with payment difficulties is 8.07%.
- Imbalance ratio is 11.39

Univariate Analysis

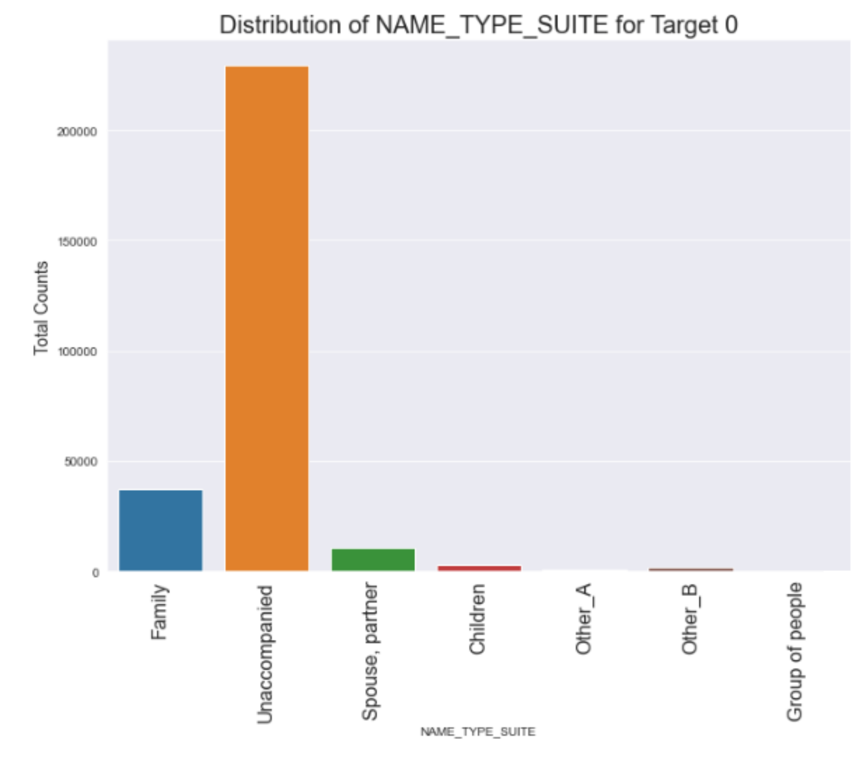
Gender variable :

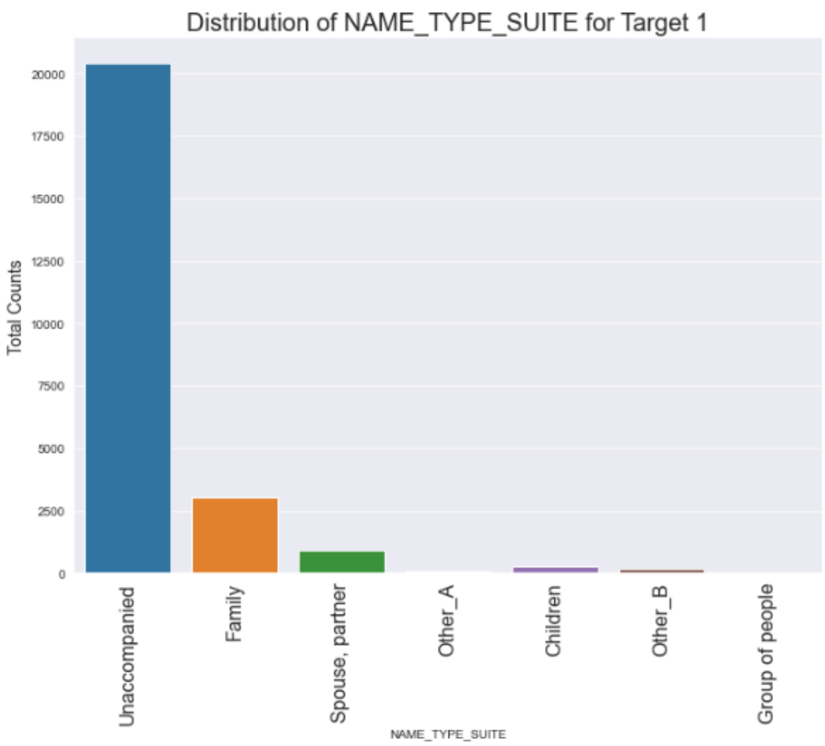
- From the pie charts, we can conclude that the number of female count is more than male.
- On the other hand, we can see that while repaying the loan males population is more likely to default than female customer.



NAME_TYPE_SUITE variable :

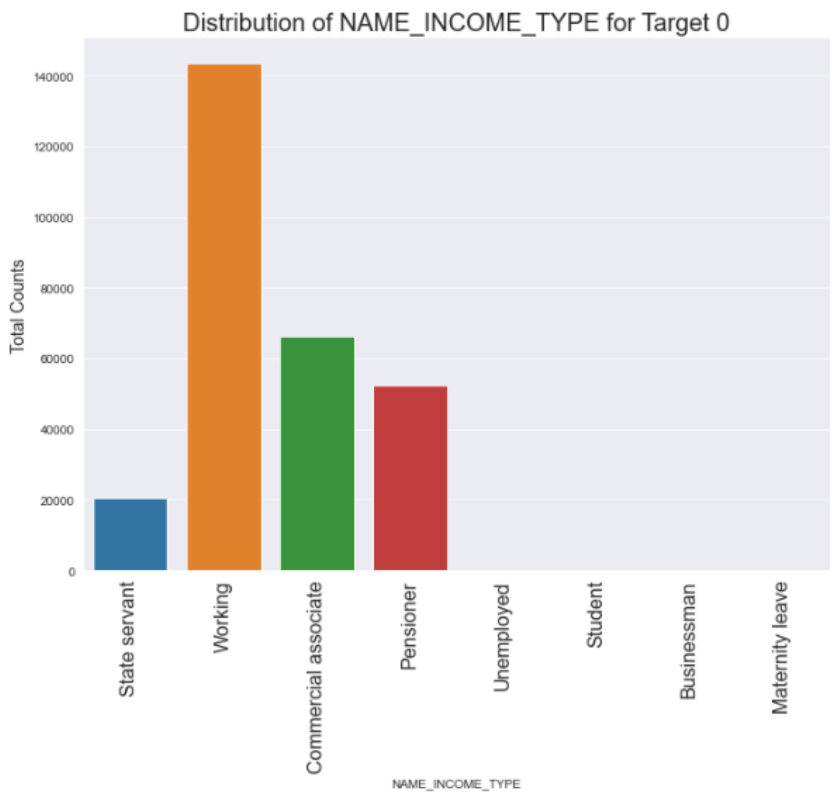
From the 'NAME_TYPE_SUITE' column, we can observe that mostly customers having no companion apply for the loan.

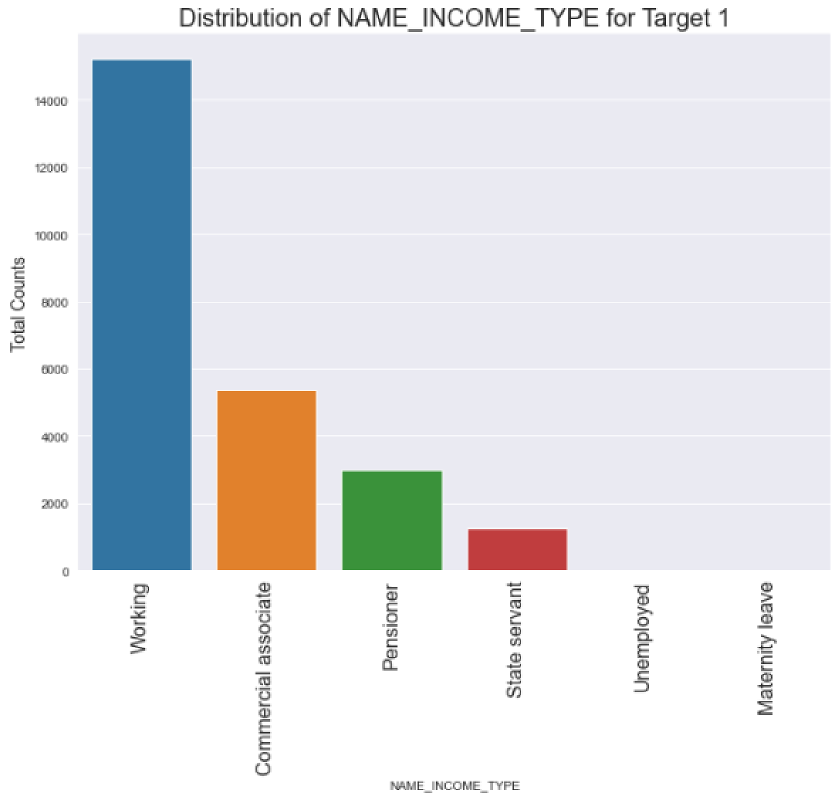




NAME_INCOME_TYPE variable :

- From the NAME_INCOME_TYPE column, we can observe that the Students and Businessman don't default.
- We can see that working customer apply for more loans followed by Commercial Associate.
- We can also observe that working class people are more likely to default the loan.

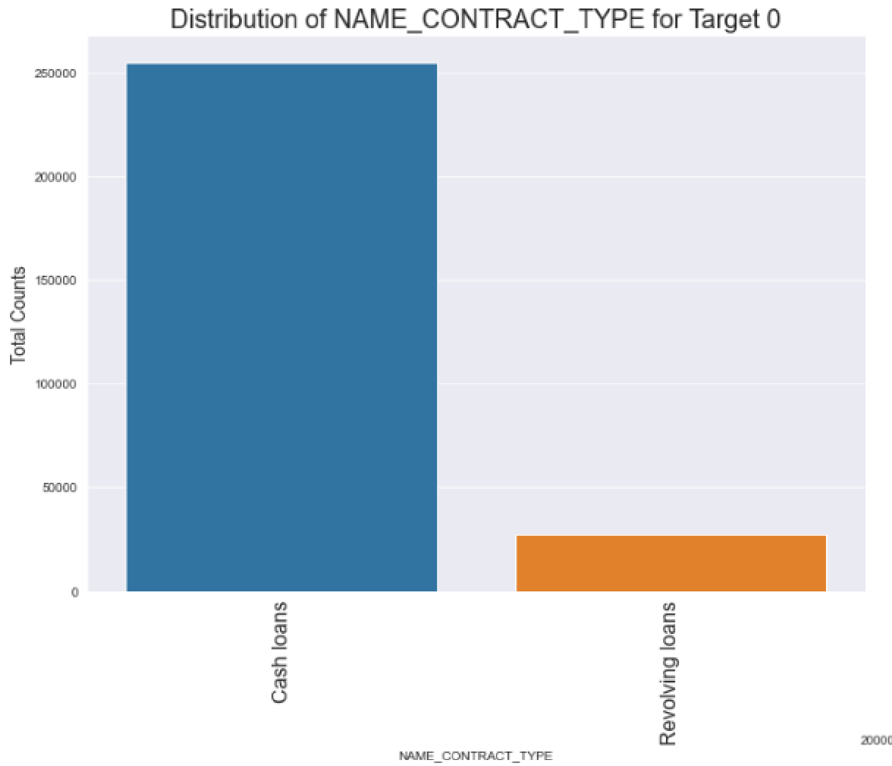


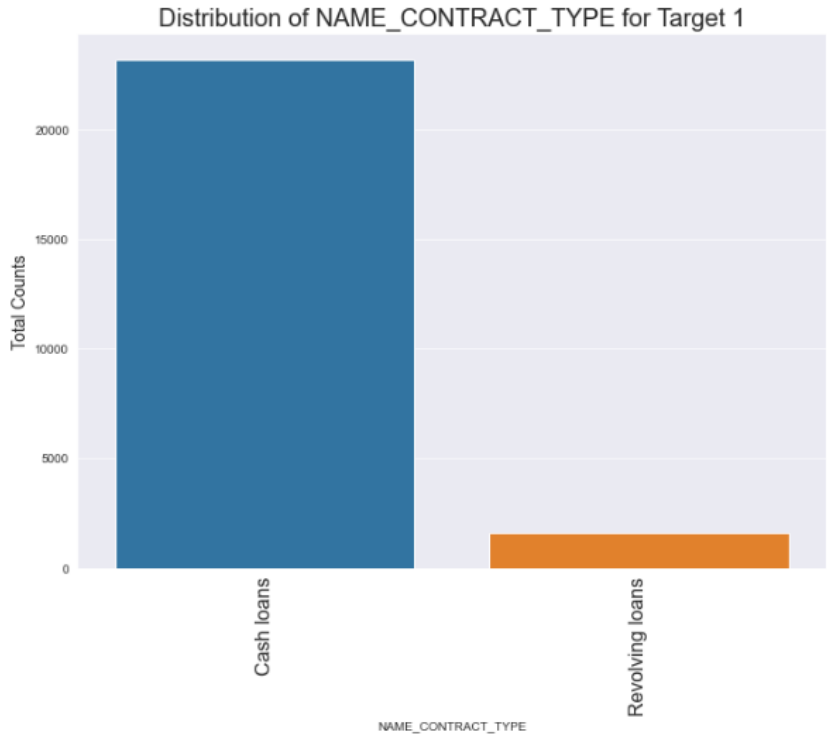


NAME_CONTRACT_TYPE variable :

- From 'NAME_CONTRACT_TYPE' column, we can observe that cash loans have higher percentage when compared with revolving loans.

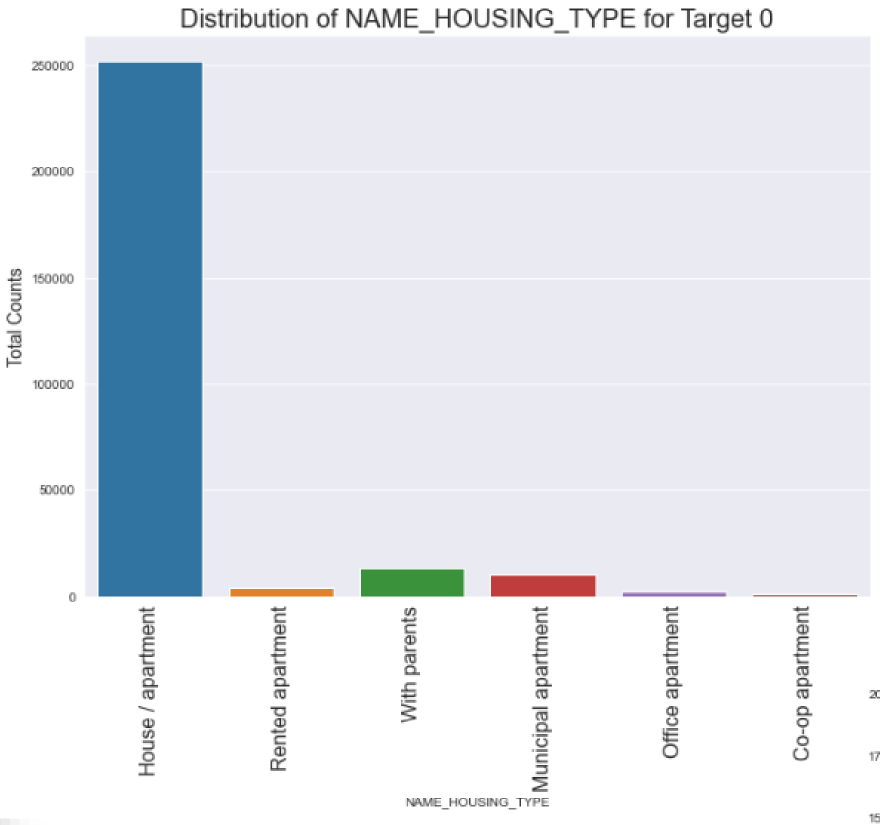
- Customers who took cash loans are more likely to have difficulty in paying the loans and there is a decrease in the percentage of Payment Difficulties who opt for revolving loans.

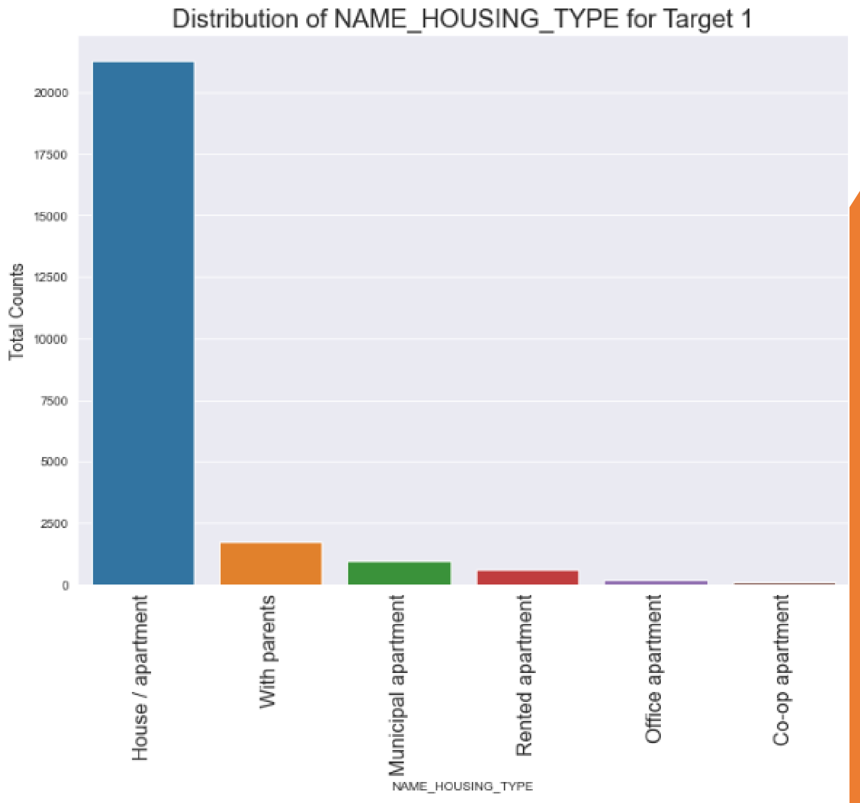




NAME_HOUSING_TYPE variable :

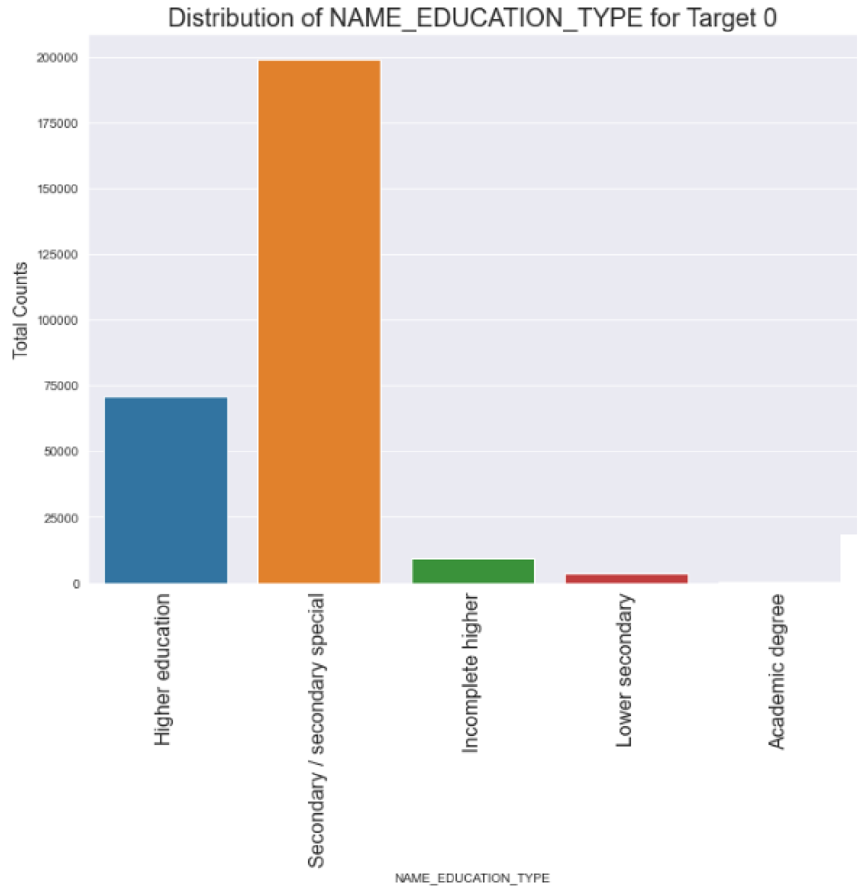
- From the NAME_HOUSING_TYPE column, we can observe that customers having house/apartments apply for more loans.
- We can also notice that customers with house/apartments are having more payment problems and customers with co-op apartments have least problems.

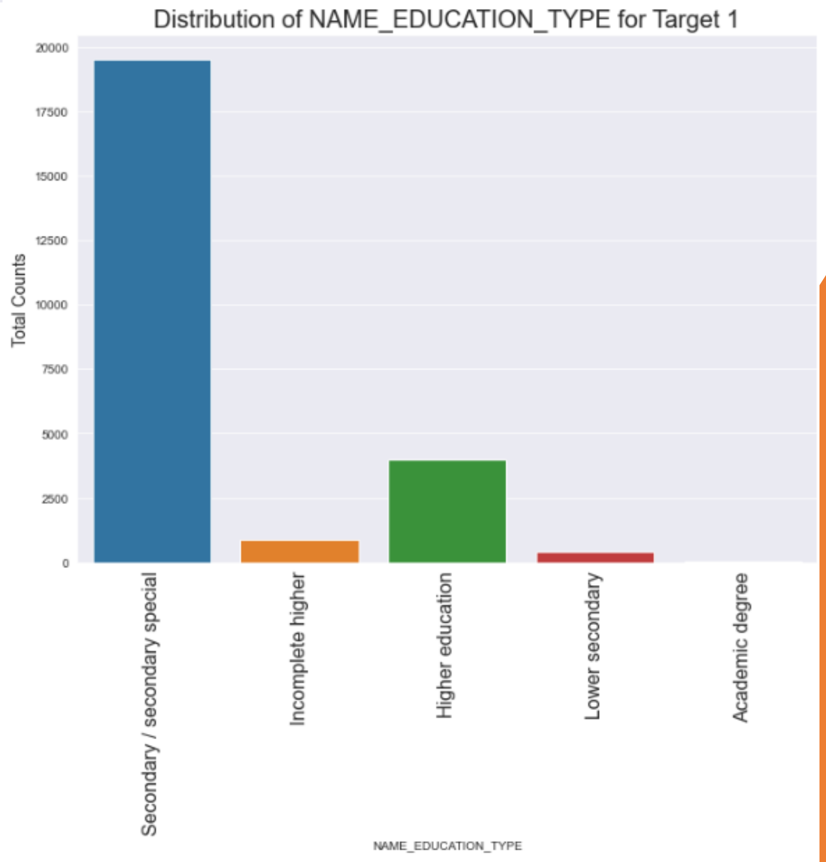




NAME_EDUCATION_TYPE variable :

- From NAME_EDUCATION_TYPE column, we can observe that majority of people who have taken loan have completed secondary special.

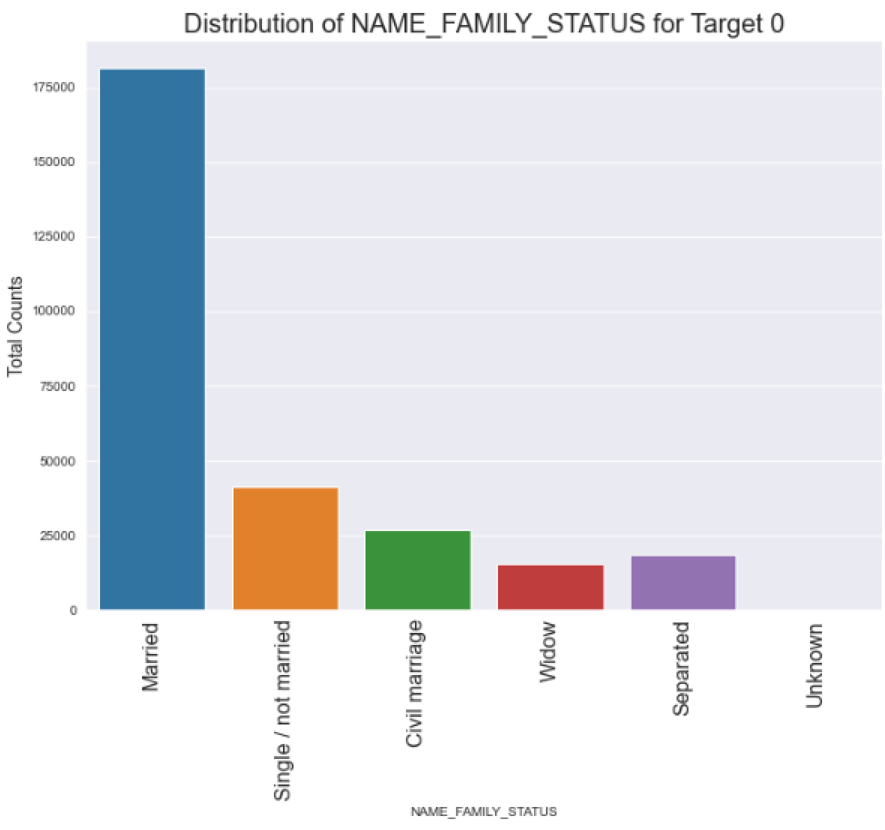


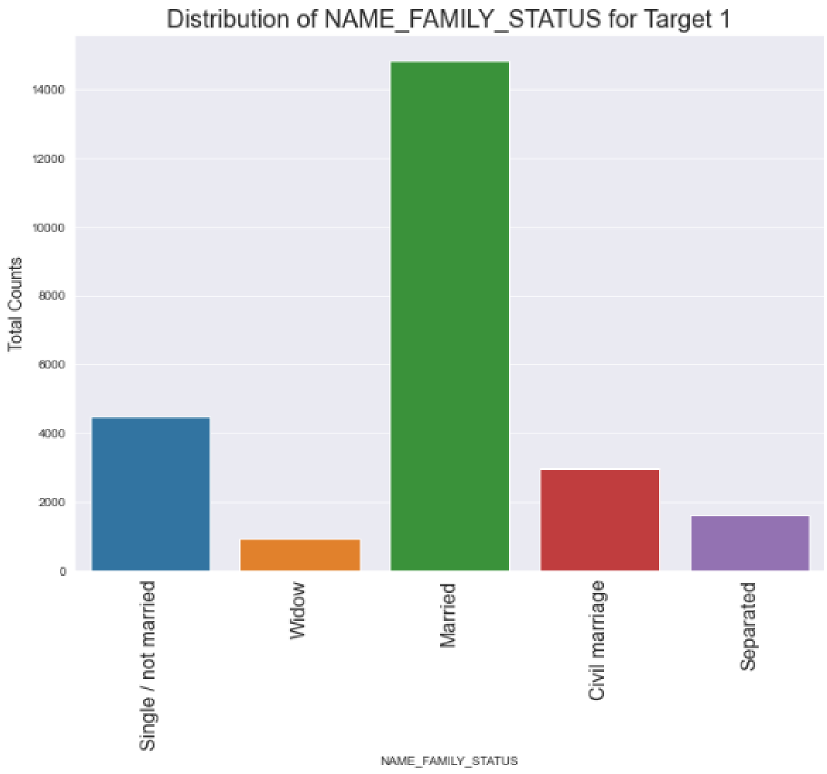


NAME_FAMILY_STATUS VARIABLE :

- From the NAME_FAMILY_STATUS column, we can observe that most of the customers who took loan are married followed by single/not married.

- Customers who are married also have highest percentage of payment problems followed by single/not married.

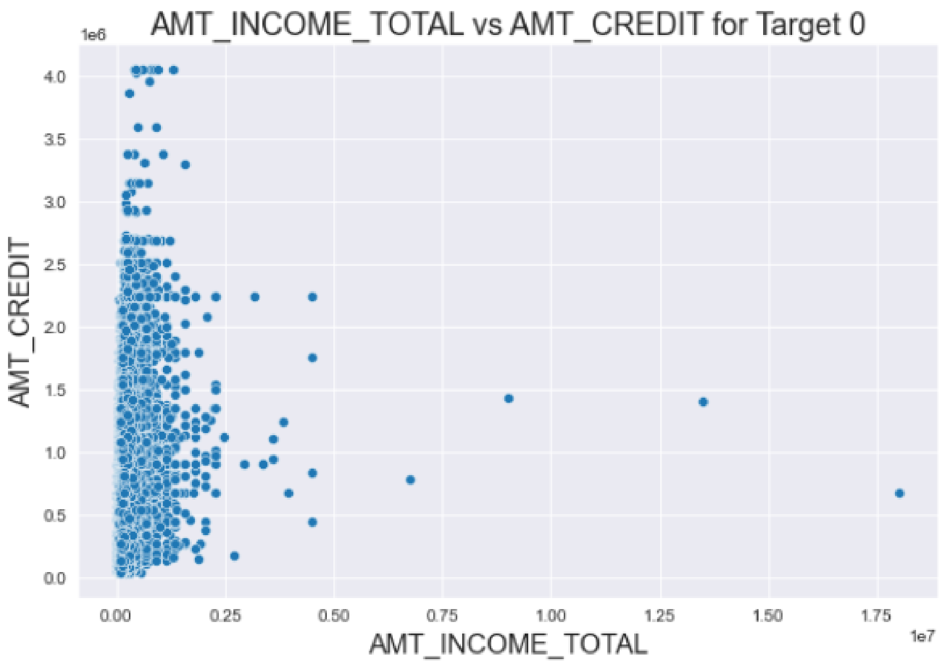


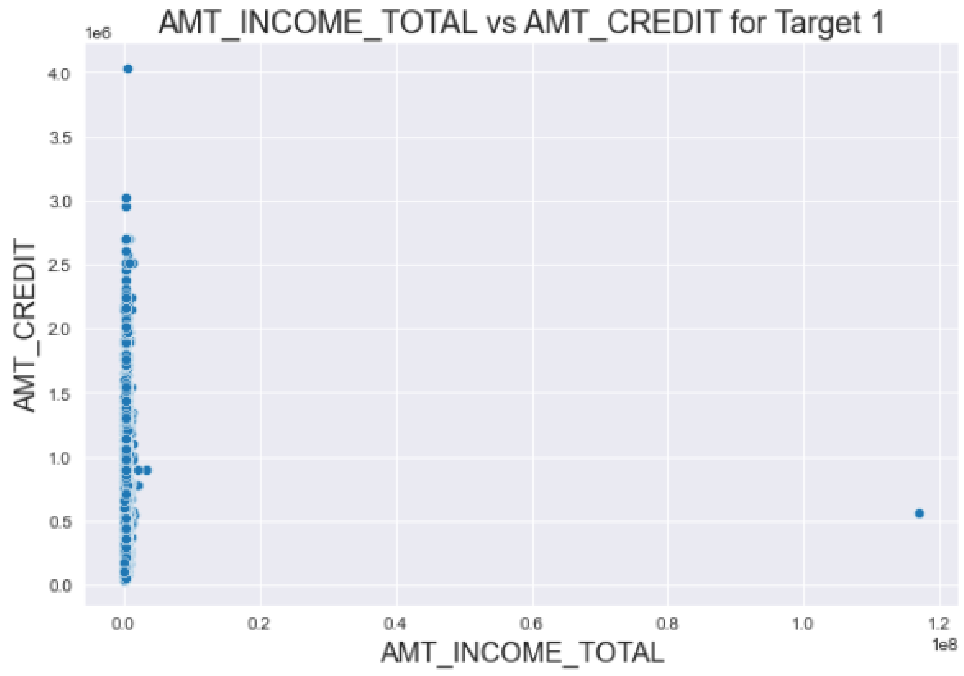


Bivariate Analysis

AMT_INCOME_TOTAL vs AMT_CREDIT

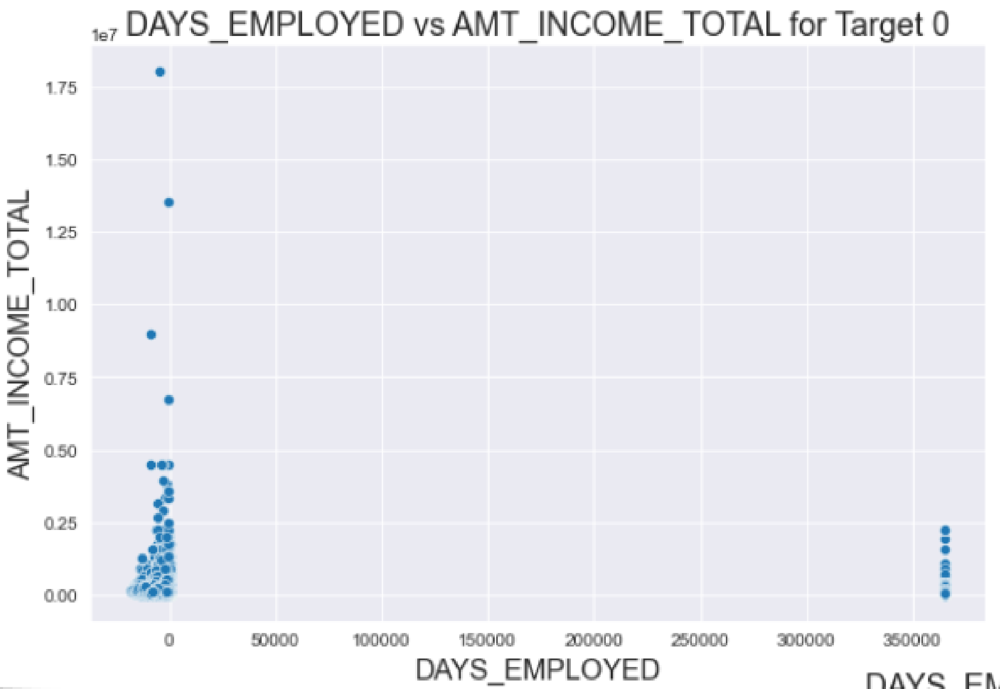
- From the above scatter plot, we can observe that customers with low income take more loans compared to others.
- Also, customers with low income have more payment problems.

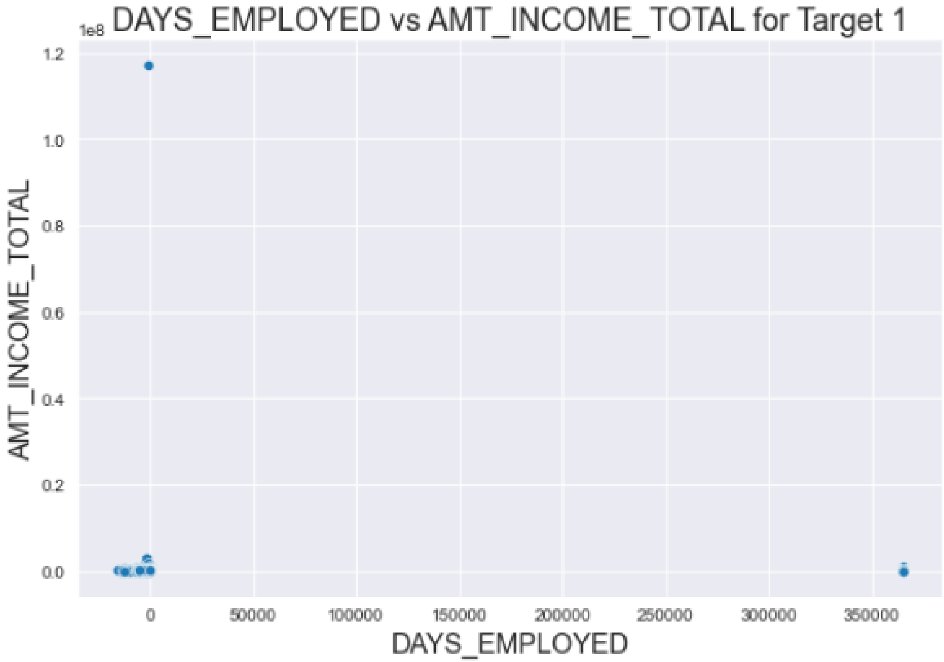




DAYS_EMPLOYED vs AMT_INCOME_TOTAL

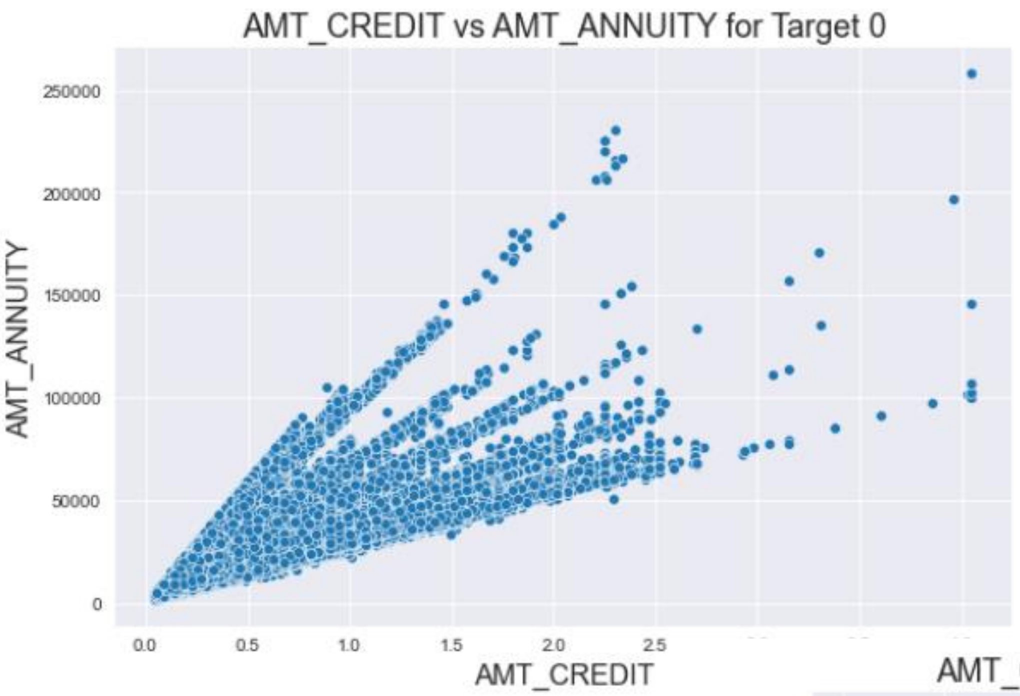
- From the scatter plot, we can see that people with low incomes have more payment problems.

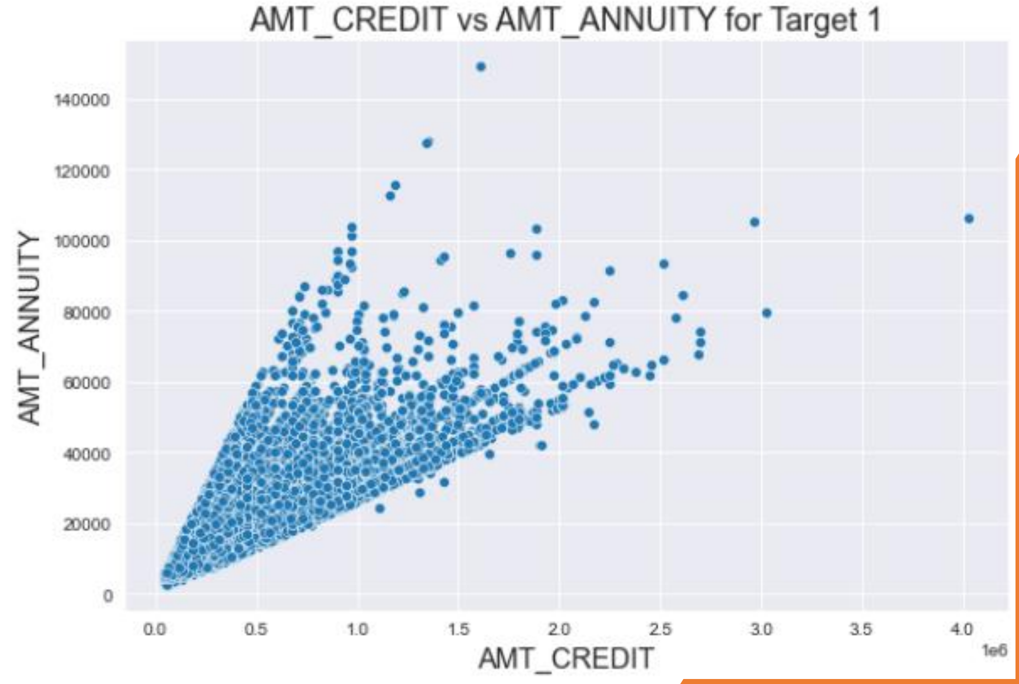




AMT_CREDIT vs AMT_ANNUITY

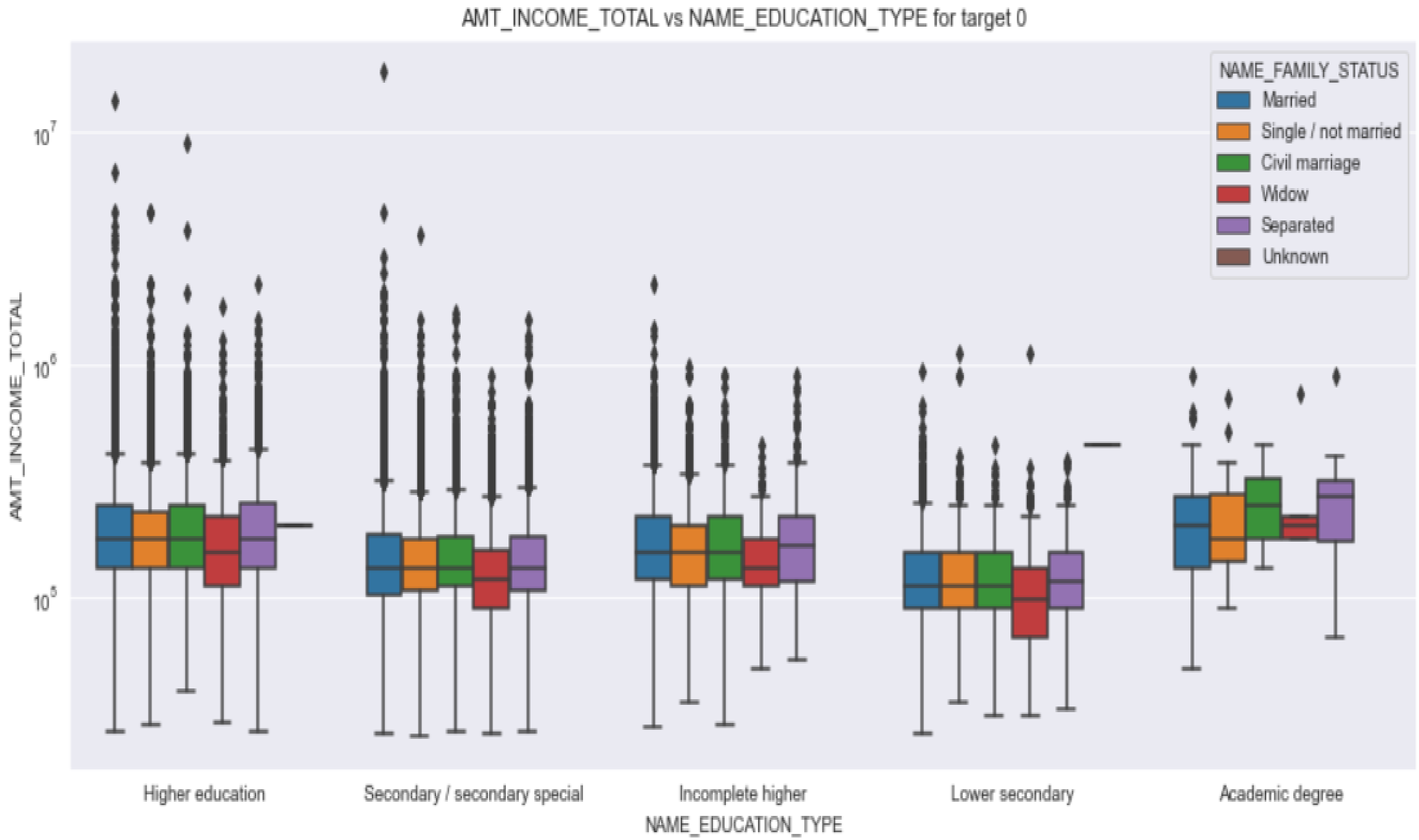
From the plot, we can notice that as the loan amount increases, the repayment term also increases.





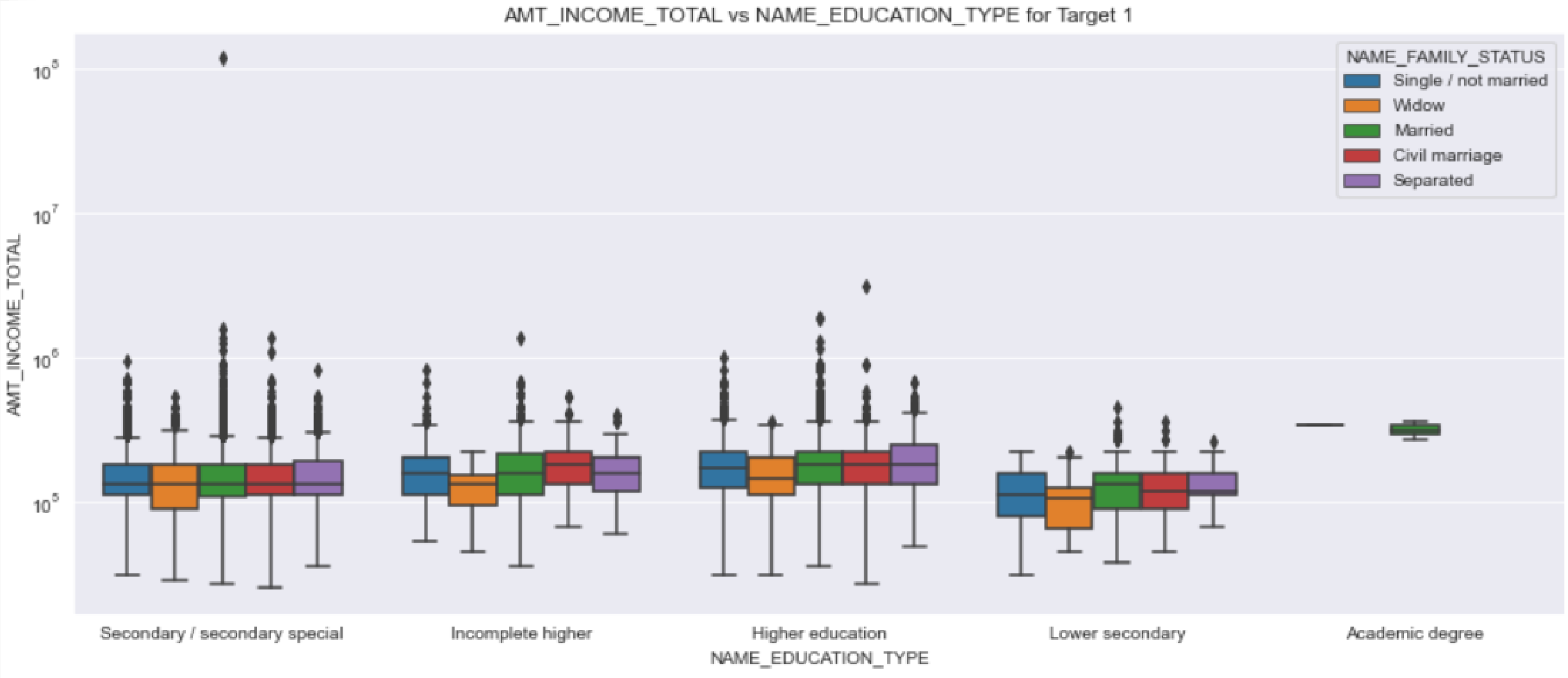
AMT_INCOME_TOTAL vs NAME_EDUCATION_TYPE

- From the boxplots, we can notice that for Education type 'Higher education' the income amount is mostly equal with family status.



- Less outlier are present for Academic degree but their income amount is little higher than Higher education.

- Lower secondary have less income amount than others.



Correlation Analysis

- From corr_0 & corr_1, we can observe that the correlations are almost same for both cases.

```
corr_0.head(15)
```

```
OBS_30_CNT_SOCIAL_CIRCLE      OBS_60_CNT_SOCIAL_CIRCLE      0.998510
OBS_60_CNT_SOCIAL_CIRCLE      OBS_30_CNT_SOCIAL_CIRCLE      0.998510
FLOORSMAX_MEDI                FLOORSMAX_AVG                 0.997018
FLOORSMAX_AVG                 FLOORSMAX_MEDI                0.997018
YEARS_BEGINEXPLUATATION_MEDI  YEARS_BEGINEXPLUATATION_AVG   0.993582
YEARS_BEGINEXPLUATATION_AVG   YEARS_BEGINEXPLUATATION_MEDI  0.993582
FLOORSMAX_MEDI                FLOORSMAX_MODE                0.988153
FLOORSMAX_MODE                FLOORSMAX_MEDI                0.988153
AMT_GOODS_PRICE               AMT_CREDIT                    0.987250
AMT_CREDIT                    AMT_GOODS_PRICE               0.987250
FLOORSMAX_AVG                 FLOORSMAX_MODE                0.985603
FLOORSMAX_MODE                FLOORSMAX_AVG                 0.985603
YEARS_BEGINEXPLUATATION_MODE  YEARS_BEGINEXPLUATATION_AVG   0.971032
YEARS_BEGINEXPLUATATION_AVG   YEARS_BEGINEXPLUATATION_MODE  0.971032
YEARS_BEGINEXPLUATATION_MEDI  YEARS_BEGINEXPLUATATION_MODE  0.962064
dtype: float64
```

- OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE shows the highest correlation.

- We can also see that some columns are directly related to each other like 'AMT_GOODS_PRICE, AMT_CREDIT'

```
corr_1.head(15)
```

```
OBS_60_CNT_SOCIAL_CIRCLE        OBS_30_CNT_SOCIAL_CIRCLE        0.998270
OBS_30_CNT_SOCIAL_CIRCLE        OBS_60_CNT_SOCIAL_CIRCLE        0.998270
FLOORSMAX_MEDI                  FLOORSMAX_AVG                   0.997187
FLOORSMAX_AVG                   FLOORSMAX_MEDI                  0.997187
YEARS_BEGINEXPLUATATION_AVG     YEARS_BEGINEXPLUATATION_MEDI    0.996124
YEARS_BEGINEXPLUATATION_MEDI    YEARS_BEGINEXPLUATATION_AVG     0.996124
FLOORSMAX_MEDI                  FLOORSMAX_MODE                  0.989195
FLOORSMAX_MODE                  FLOORSMAX_MEDI                  0.989195
                                FLOORSMAX_AVG                   0.986594
FLOORSMAX_AVG                   FLOORSMAX_MODE                  0.986594
AMT_GOODS_PRICE                 AMT_CREDIT                      0.983103
AMT_CREDIT                      AMT_GOODS_PRICE                 0.983103
YEARS_BEGINEXPLUATATION_MODE    YEARS_BEGINEXPLUATATION_AVG     0.980466
YEARS_BEGINEXPLUATATION_AVG     YEARS_BEGINEXPLUATATION_MODE    0.980466
YEARS_BEGINEXPLUATATION_MODE    YEARS_BEGINEXPLUATATION_MEDI    0.978073
dtype: float64
```

Analyzing Previous Application Datset

```
app_pre.shape
```

```
(1670214, 37)
```

```
app_pre.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
 20  NAME_TYPE_SUITE              849809 non-null   object
 21  NAME_CLIENT_TYPE             1670214 non-null  object
 22  NAME_GOODS_CATEGORY          1670214 non-null  object
 23  NAME_PORTFOLIO               1670214 non-null  object
```

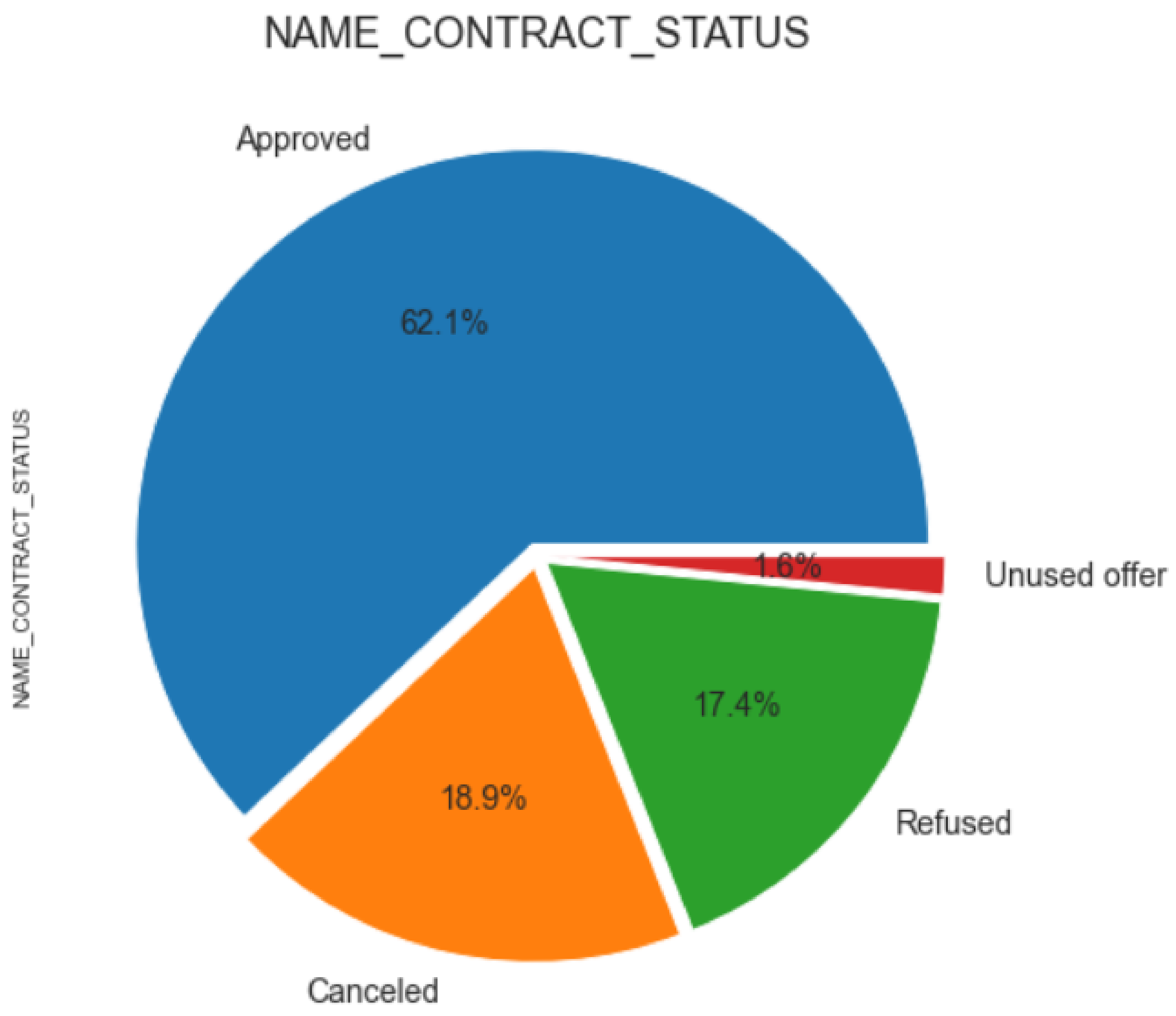
```
# Removing all the columns with more than 50% of null values
app_pre1 = app_pre.loc[:,app_pre.isnull().mean()<=0.5]
app_pre1.shape
```

```
(1670214, 33)
```

Univariate Analysis

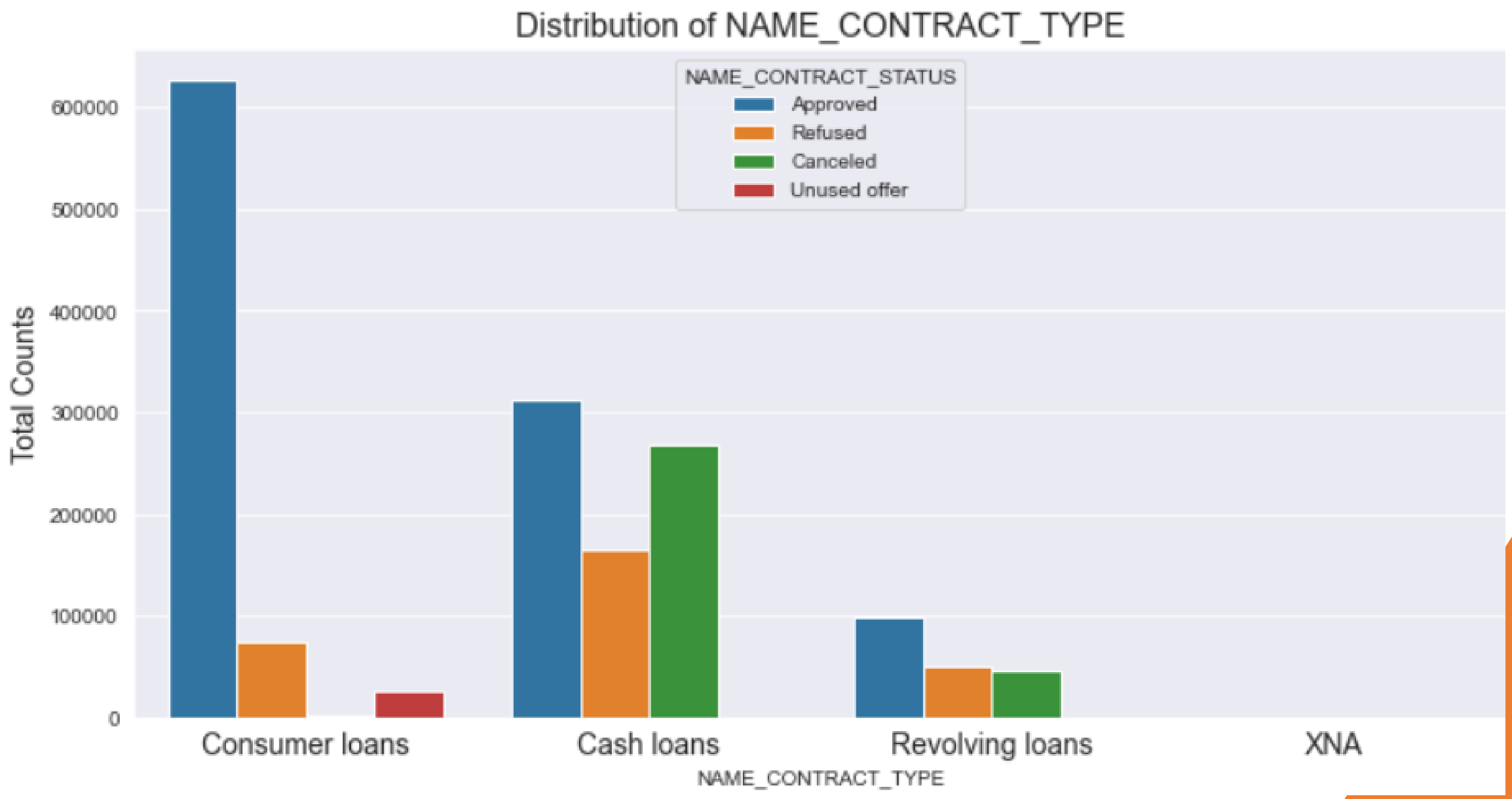
NAME_CONTRACT_STATUS :

- From the pie chart, we can observe that approved loans has the highest percentage.
- We can also see that only few loans have been refused.



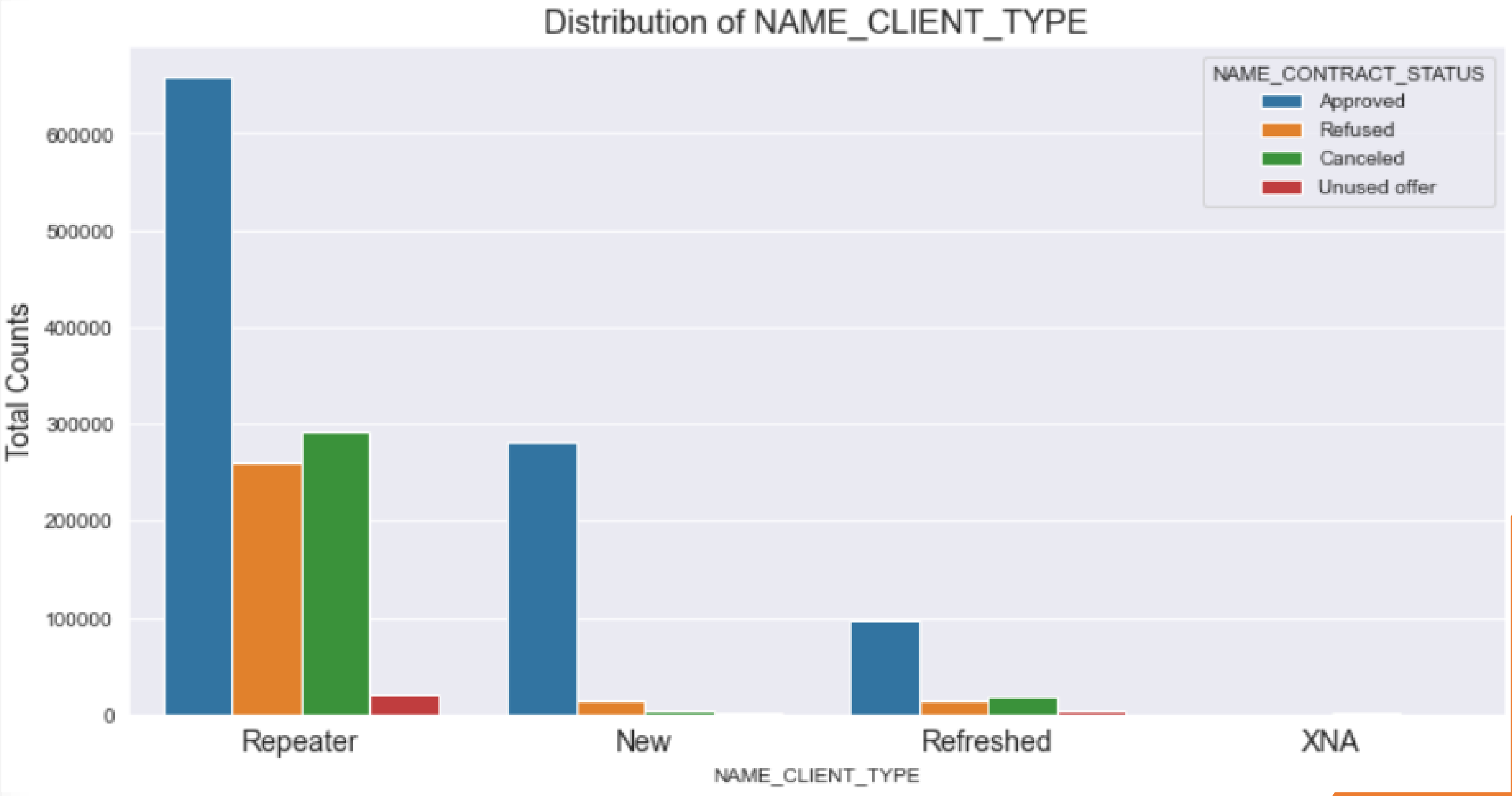
NAME_CONTRACT_TYPE :

From the graph, we can notice that most of the previous applications are consumer loans and cash loans.



NAME_CLIENT_ TYPE :

- From the graph, we can observe that most of the loan applications are from repeaters.

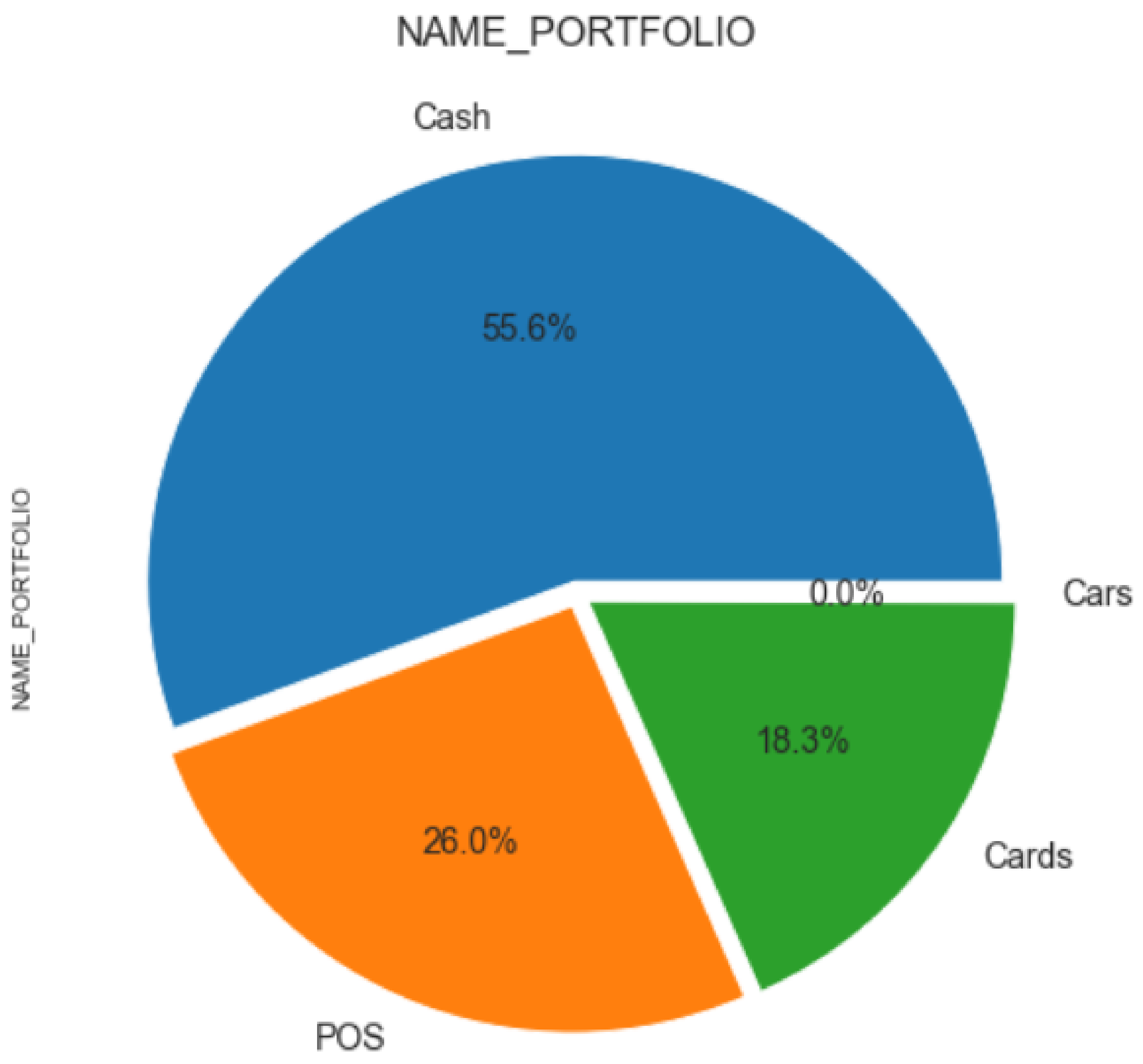


CODE_REJECT_ REASON :

From the CODE_REJECT_ REASON column, we can see that 'HC' is the reason for majority of the loans to be rejected.

NAME_ PORTFOLIO :

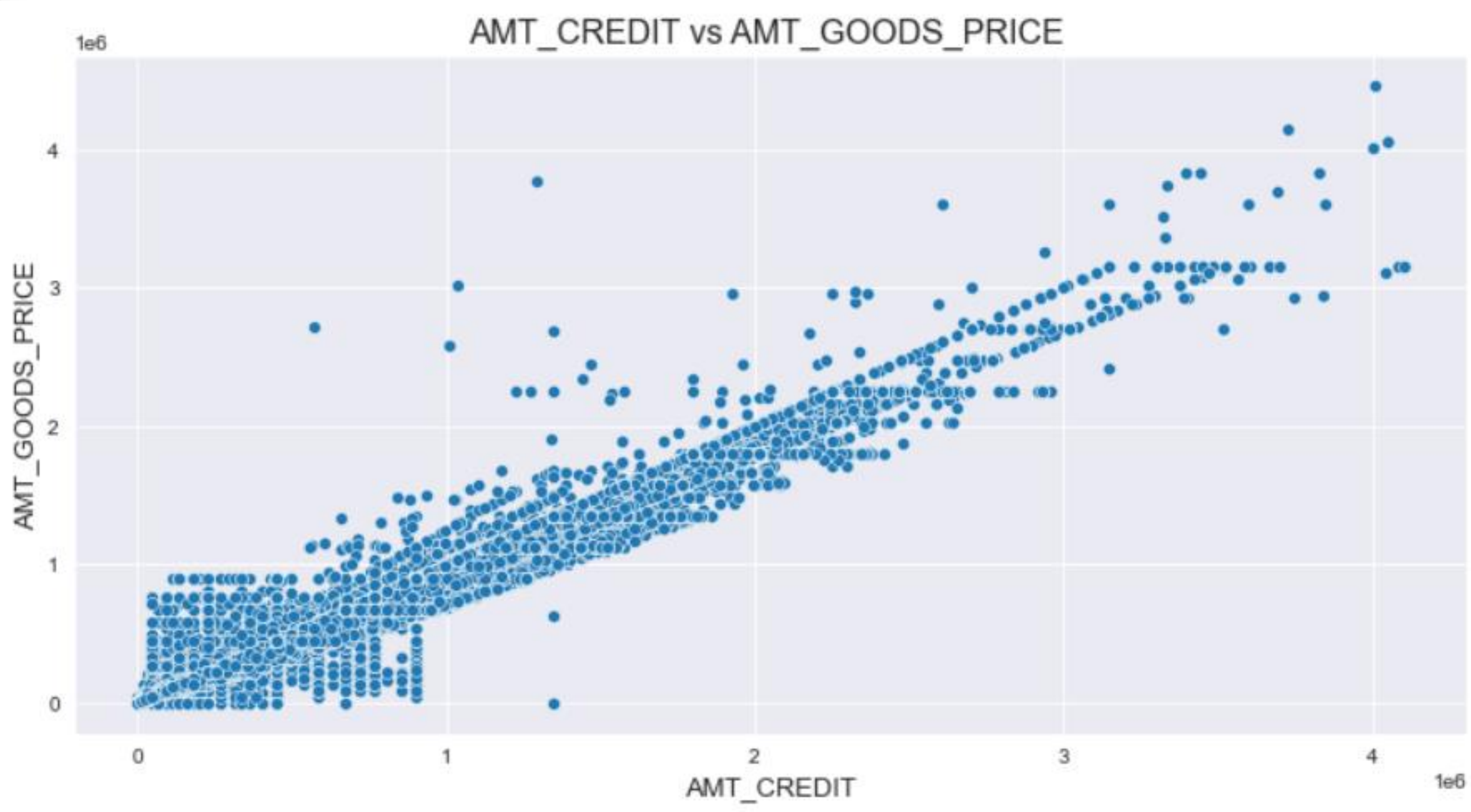
From the pie chart, we can infer that most of previous applications have been applied for 'POS' followed by 'Cash'.



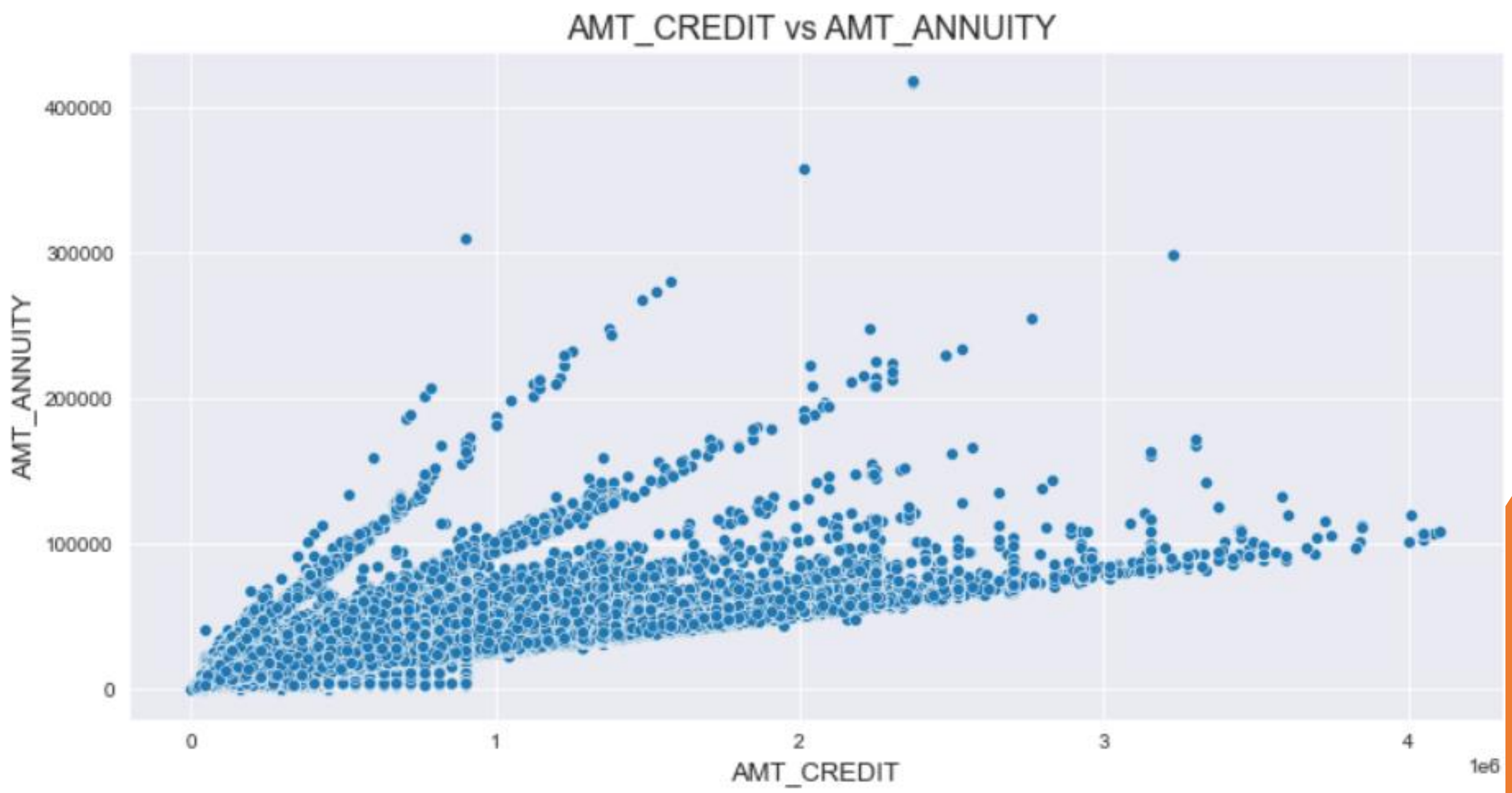
Bivariate Analysis

AMT_CREDIT vs AMT_GOODS_PRICE

In the previous applications, we can observe that the amount credited is highly influenced by the price of the goods.

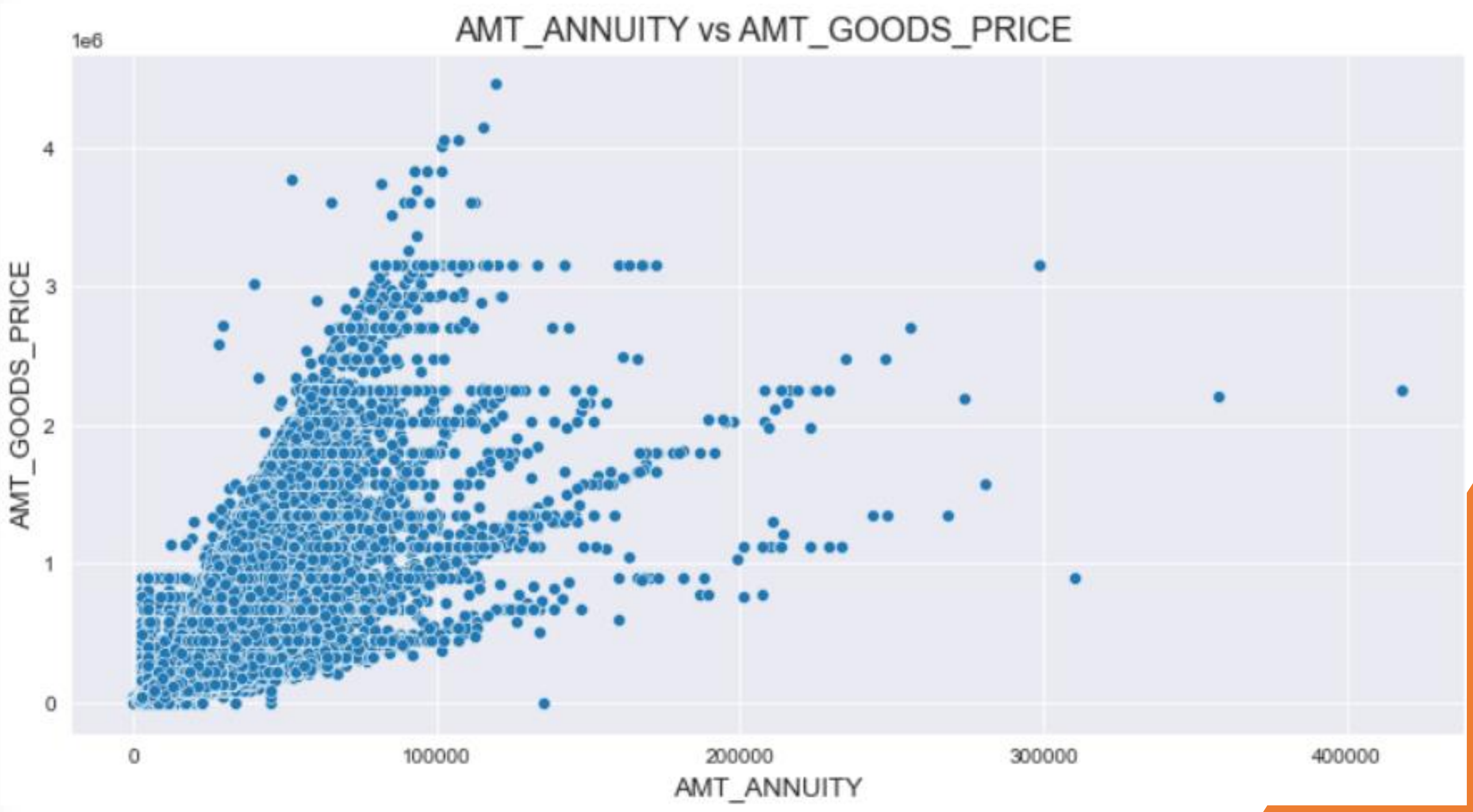


AMT_CREDIT vs AMT_ANNUITY



AMT_ANNUITY vs AMT_GOODS_ PRICE

From the two scatter plots, we can notice that AMT_ANNUITY has a high influence over the goods price and credit.



Merged Dataframe Analysis

```python
# Lets merge both the files and analyse patterns in the data
new_df = pd.merge(left = app1, right = app_pre, how='inner', on='SK_ID_CURR',suffixes='_x')
```

```python
new_df.shape
```

(1413701, 91)

```python
# Renaming the column names after merging

new_df = new_df.rename({'NAME_CONTRACT_TYPE_' : 'NAME_CONTRACT_TYPE','AMT_CREDIT_':'AMT_CREDIT','AMT_ANNUITY_':'AMT_ANNUITY',
                        'WEEKDAY_APPR_PROCESS_START_' : 'WEEKDAY_APPR_PROCESS_START',
                        'HOUR_APPR_PROCESS_START_':'HOUR_APPR_PROCESS_START','NAME_CONTRACT_TYPEx':'NAME_CONTRACT_TYPE_PREV',
                        'AMT_CREDITx':'AMT_CREDIT_PREV','AMT_ANNUITYx':'AMT_ANNUITY_PREV',
                        'WEEKDAY_APPR_PROCESS_STARTx':'WEEKDAY_APPR_PROCESS_START_PREV',
                        'HOUR_APPR_PROCESS_STARTx':'HOUR_APPR_PROCESS_START_PREV'}, axis=1)
```

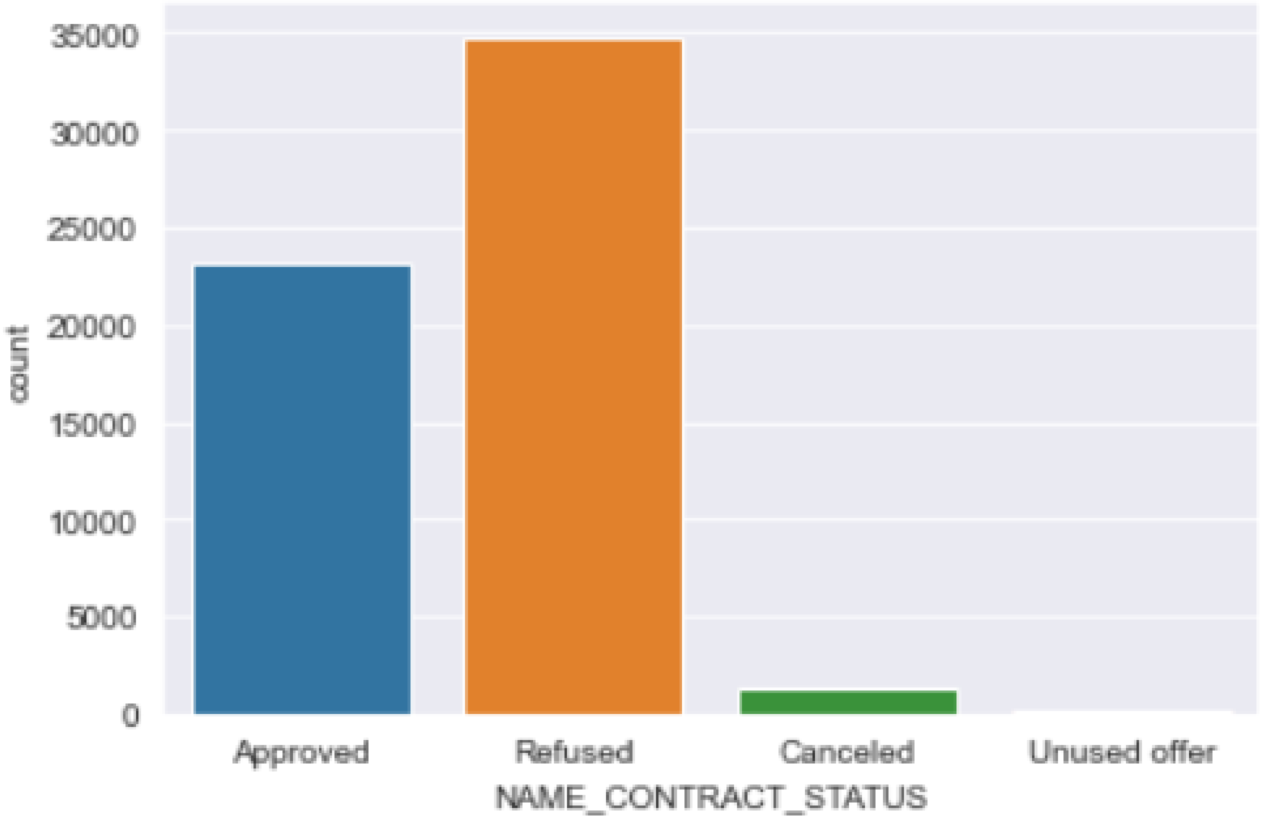
```python
# Removing the column values of 'XNA' and 'XAP'

new_df=new_df.drop(new_df[new_df['NAME_CASH_LOAN_PURPOSE']=='XNA'].index)
new_df=new_df.drop(new_df[new_df['NAME_CASH_LOAN_PURPOSE']=='XAP'].index)

new_df.shape
```

(59413, 91)

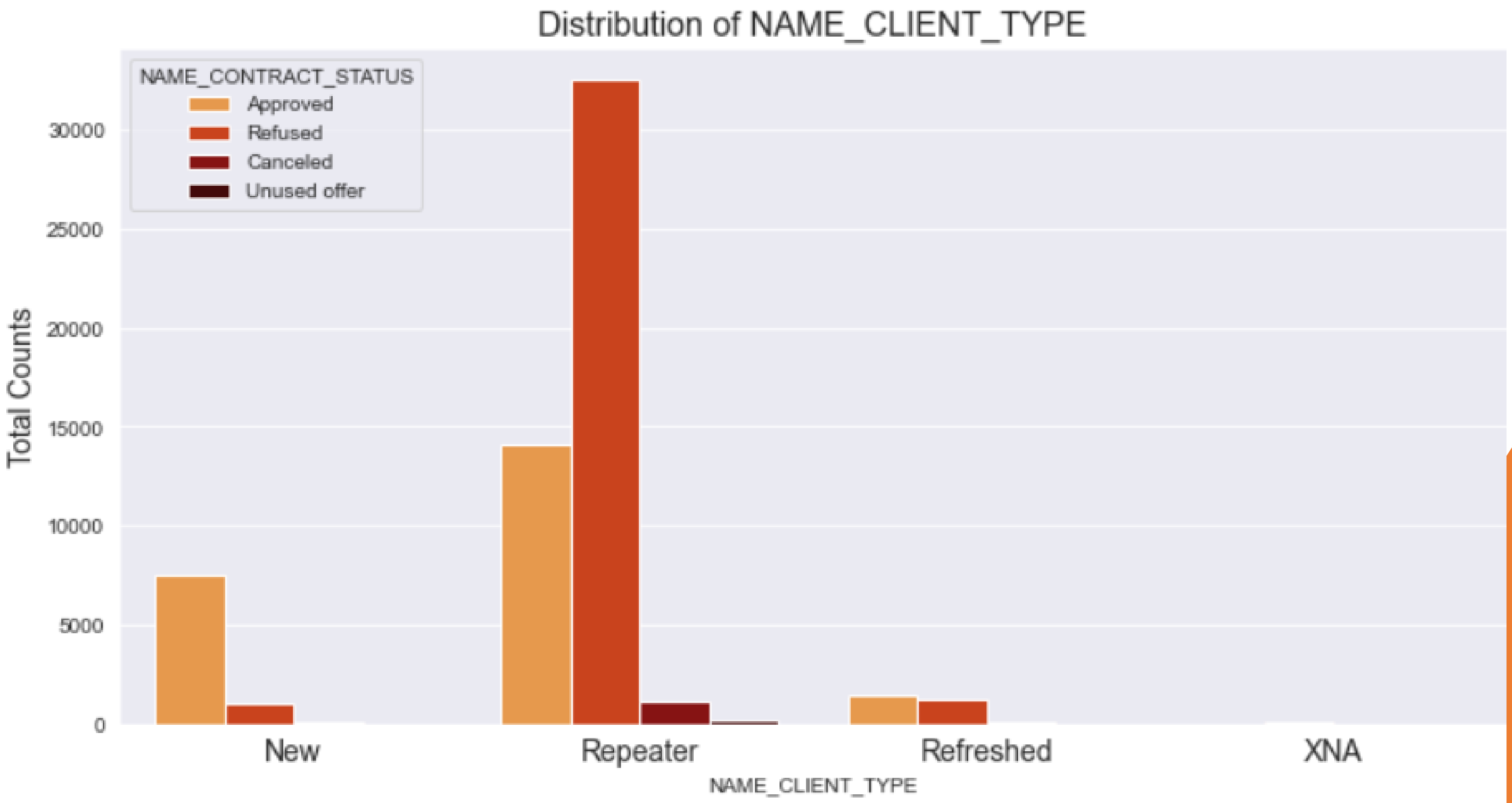
From the graph, we can notice that refused applications has the highest percentage.



Univariate Analysis

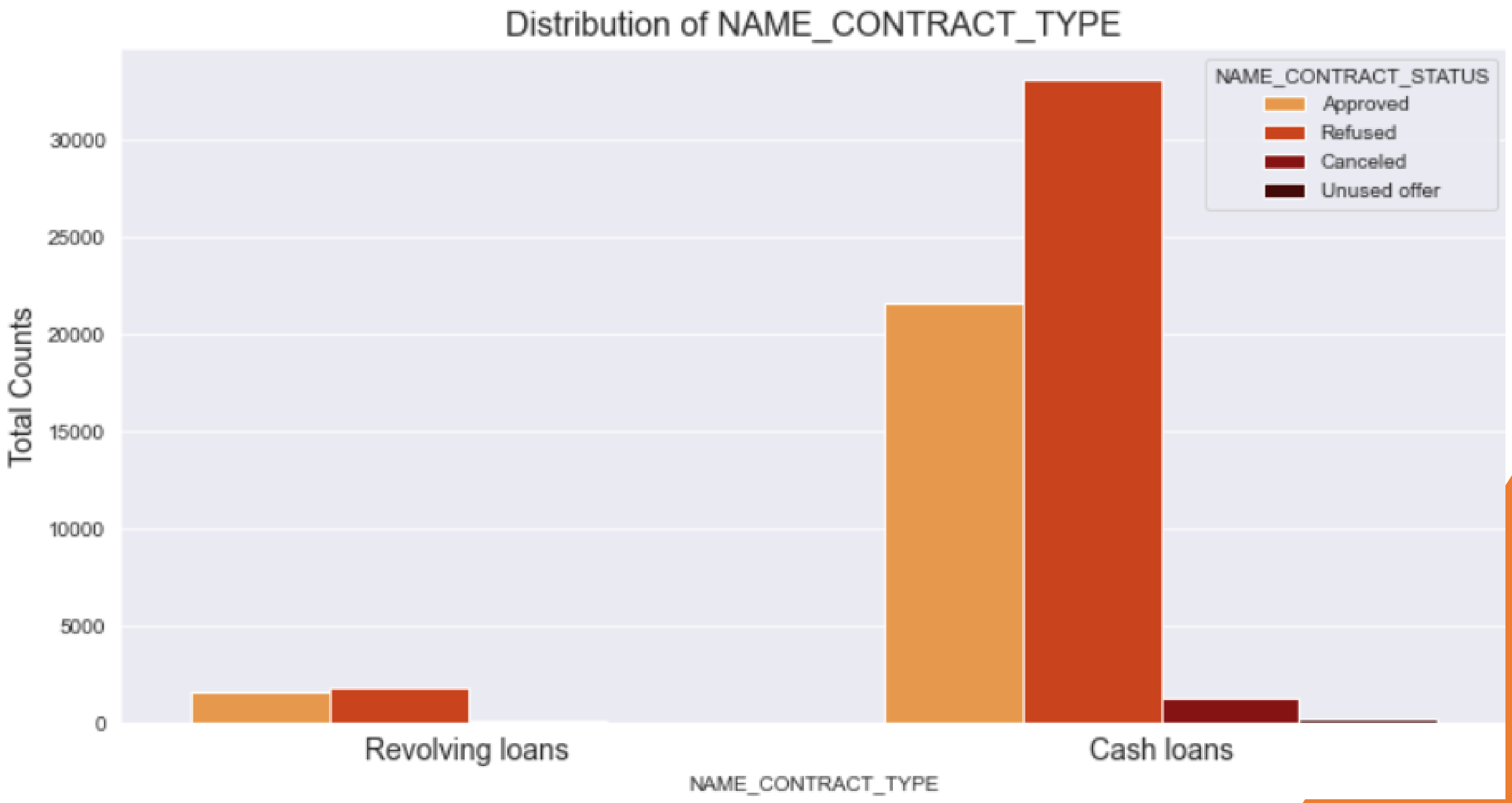
NAME_CLIENT_TYPE :

- From the plot, we can notice that most of the repeater applications have been refused.
- But we can even notice that the repeater applications have the highest approval.



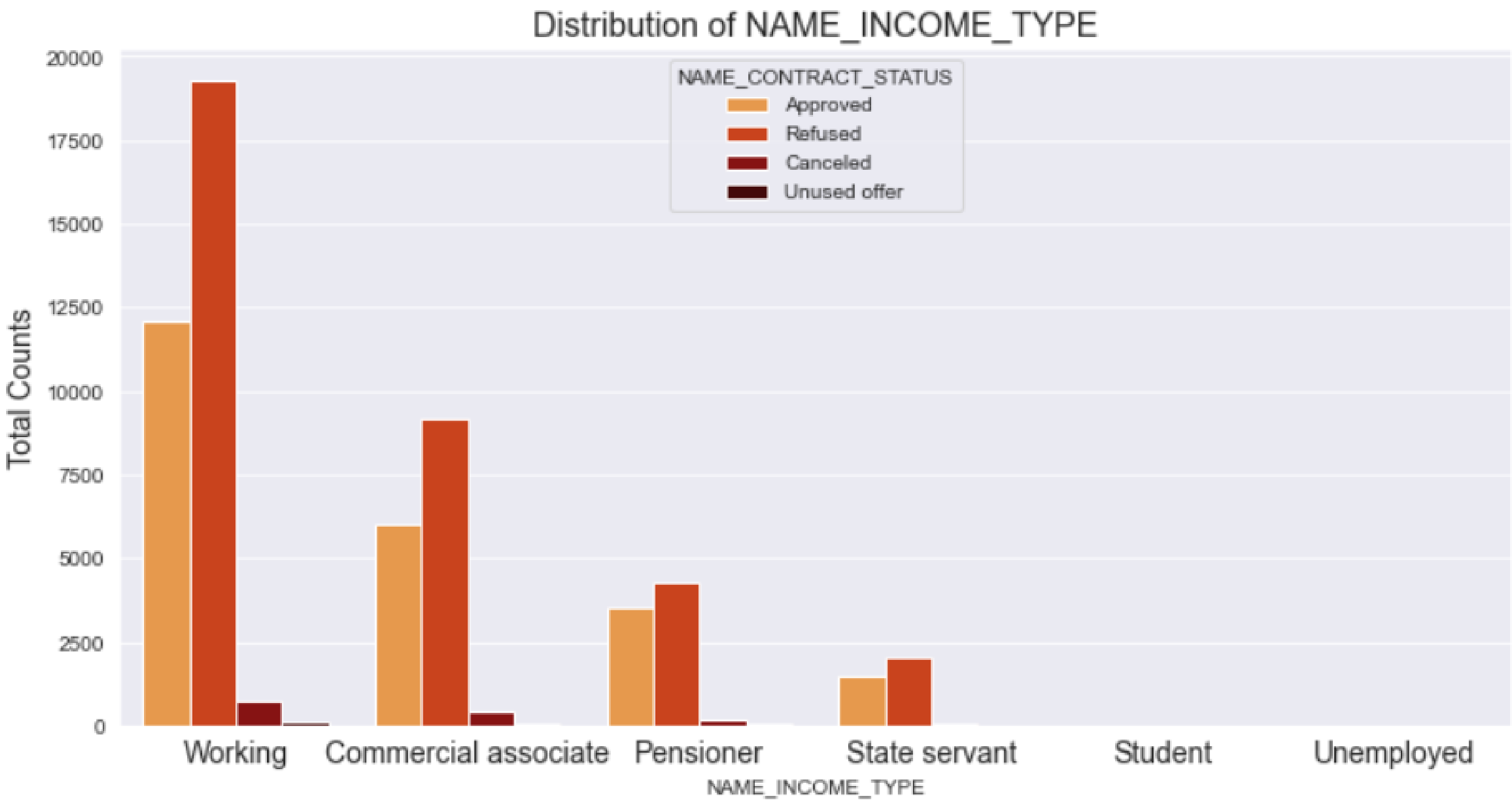
NAME_CONTRACT_TYPE :

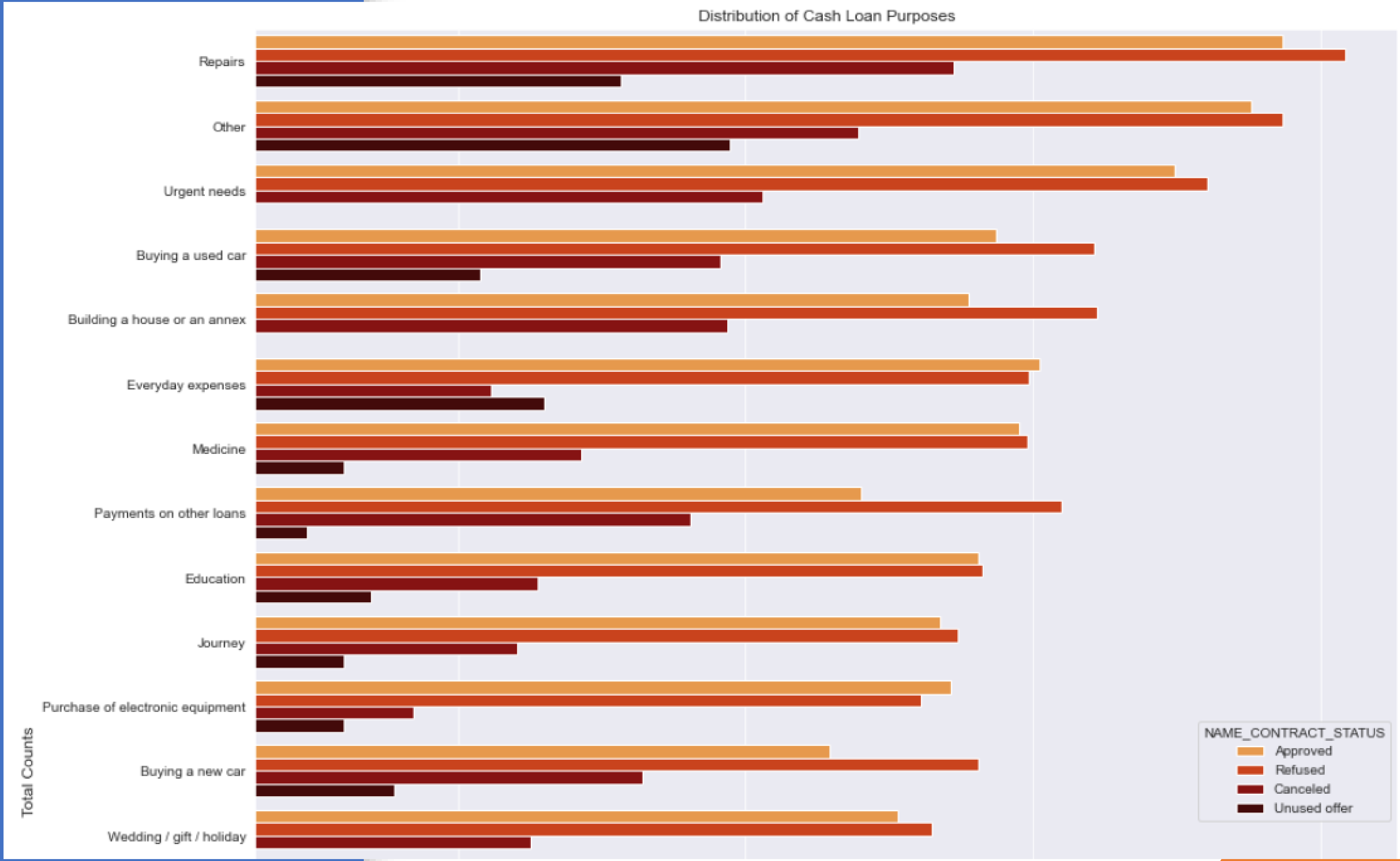
- From the plot, we can notice that most of the applications for cash loans have been rejected.
- On the other hand, highest percentage of approved loans is also cash loans.



NAME_INCOME_TYPE :

- From the above plot, we can notice that working people apply for more loans followed by commercial associate.
- But we can also observe that the maximum application rejection is also from working customers.

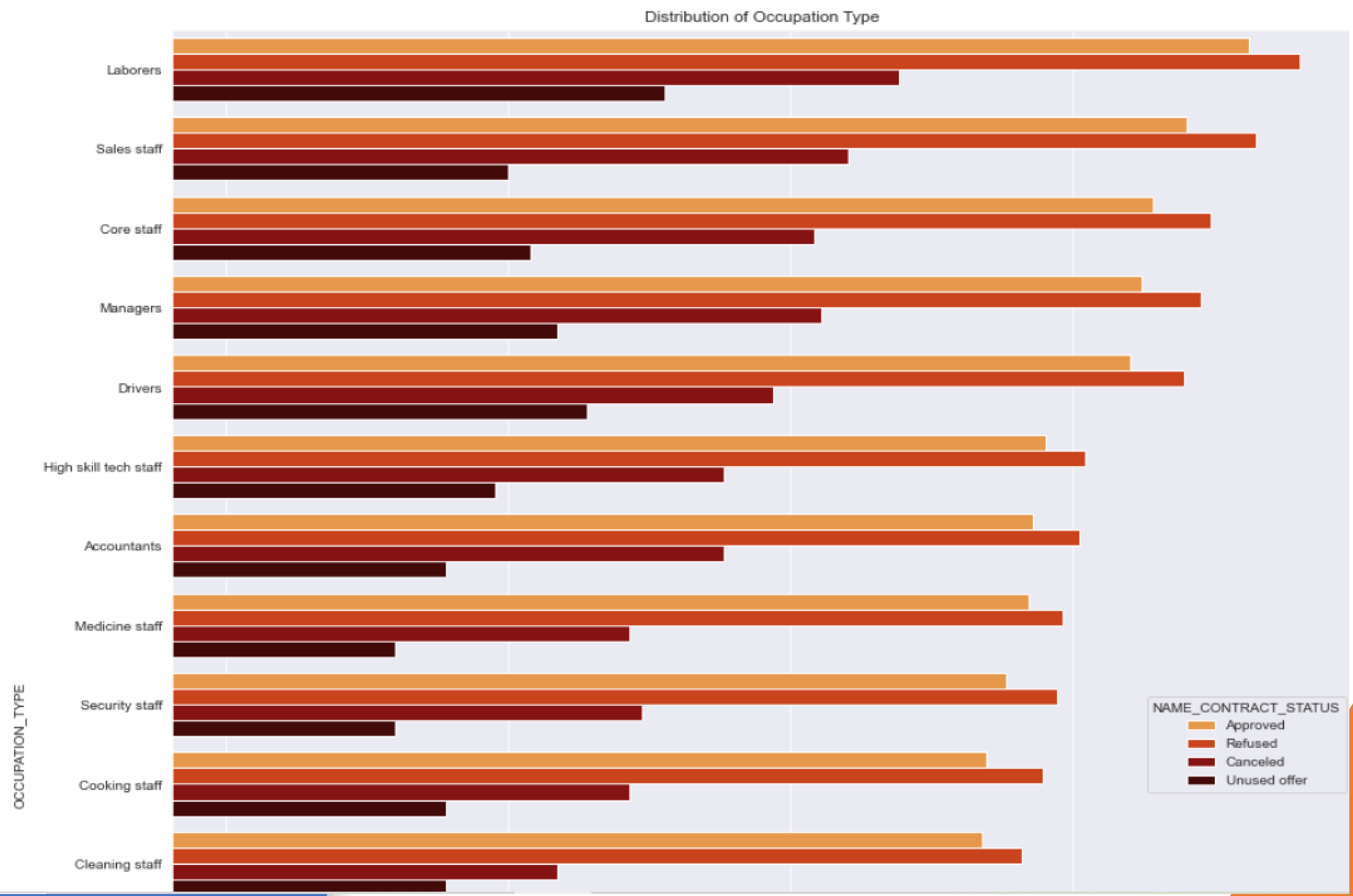




Distribution of Cash Loan Purposes
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Purchase of electronic equipment
Buying a new car
Wedding / gift / holiday
Total Counts
NAME_CONTRACT_STATUS
Approved
Refused
Canceled
Unused offer

NAME_CASH_ LOAN_PURPOSE

- From the previous plot, we can observe that most of the rejected loans are for 'repairs' purpose.
- We can also notice that 'education' and 'medicine' have equal number of approval and rejection.


Distribution of Occupation Type
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
OCCUPATION_TYPE
NAME_CONTRACT_STATUS
Approved
Refused
Canceled
Unused offer

OCCUPATION_ TYPE

- From the previous plot, we can observe that most of the 'laborers' loan applications are rejected followed by 'sales staff'.
- We can also notice that 'Secretaries' have equal number of approvals and rejections.

Conclusion

- In the gender category, we can conclude that the loan applications from female customers are more when compared with male. On the other hand, we can see that while repaying the loan males population is more likely to default than female customer.

- 'Working' customers are having most number of unsuccessful payments. So bank should focus less on this group of people.

- AMT_ANNUITY has a high influence over the goods price and credit.

- Bank should avoid customers with low income as they have more payment problems.

- Bank should focus more on students and businessman as they don't default.

THE END