Credit EDA Assignment

By: Ramakrushna Mohapatra

Batch: DS C47

August 2022 Batch

::Problem Statement::

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.

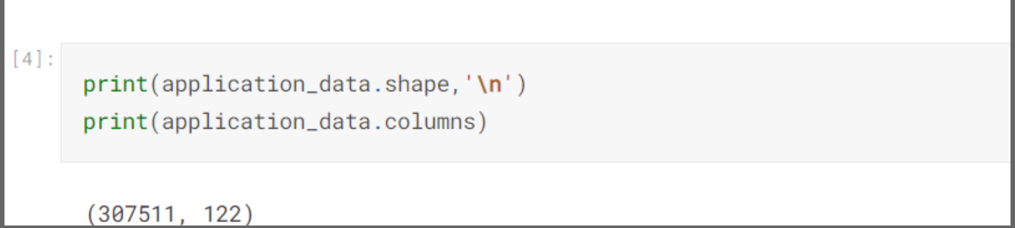
Major Problems:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. 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

Data Mining and Cleaning: (Application Data)

Shape:

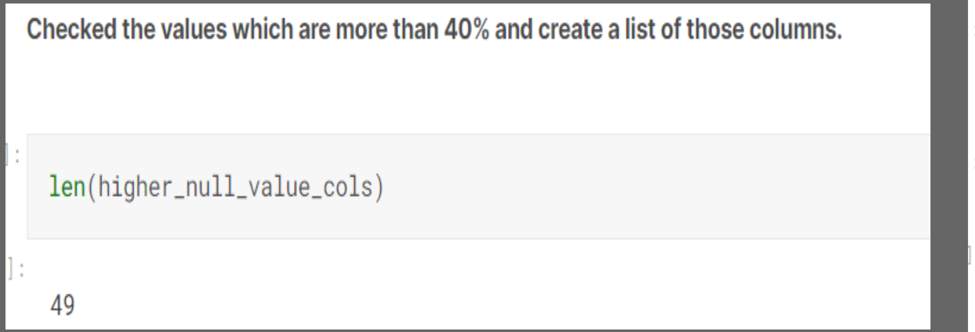
Total 307511 rows and 122 columns.



```
[4]:
print(application_data.shape,'\n')
print(application_data.columns)

(307511, 122)
```

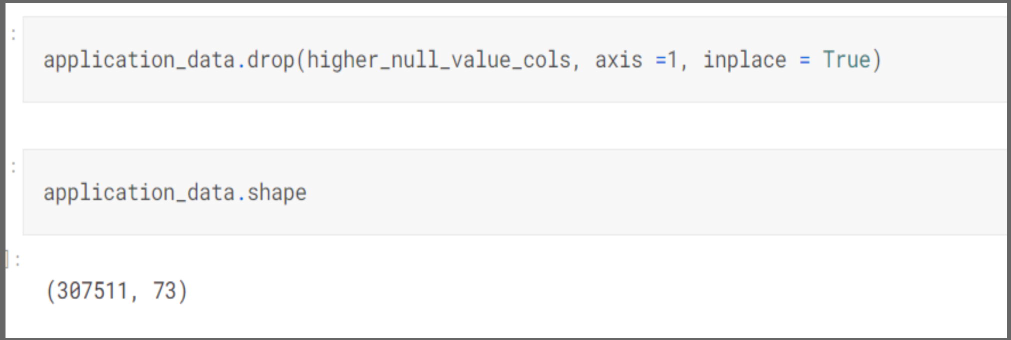
Missing Value Check and Imputation:

1. Dropped those columns which are having more than 40% of null values from the dataframe.
2. Columns which are having less than 30% of null values has been imputed with values mean, median and mode. Mostly Mode and mean.
3. 3. Total 49 no of columns has been dropped after checking as per the condition mentioned in first point. Total column reduced to 73.



Checked the values which are more than 40% and create a list of those columns.

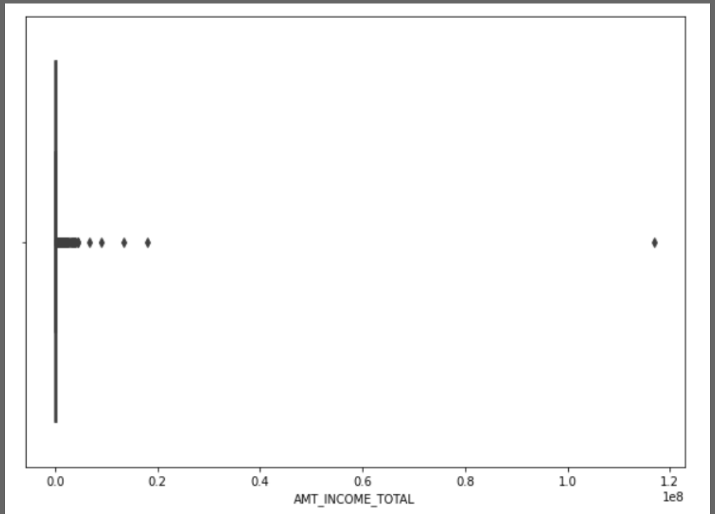
```
len(higher_null_value_cols)

49
```

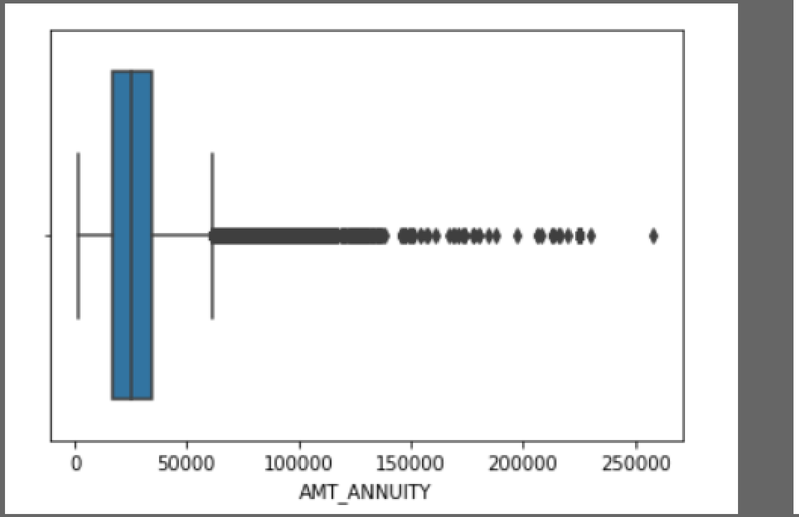


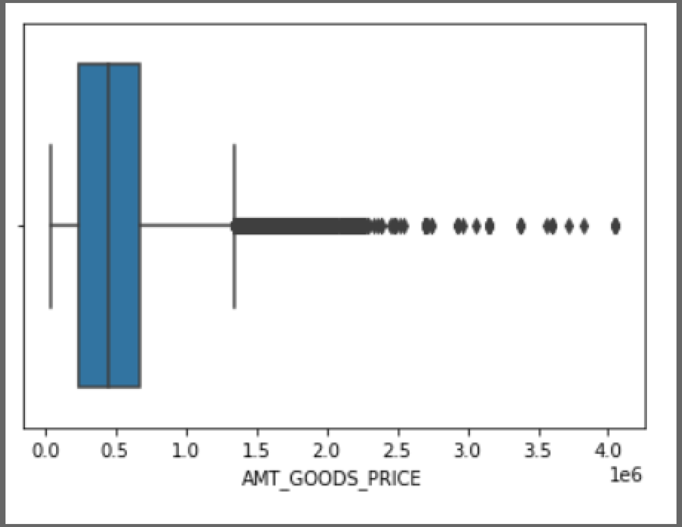
```
application_data.drop(higher_null_value_cols, axis =1, inplace = True)

application_data.shape

(307511, 73)
```

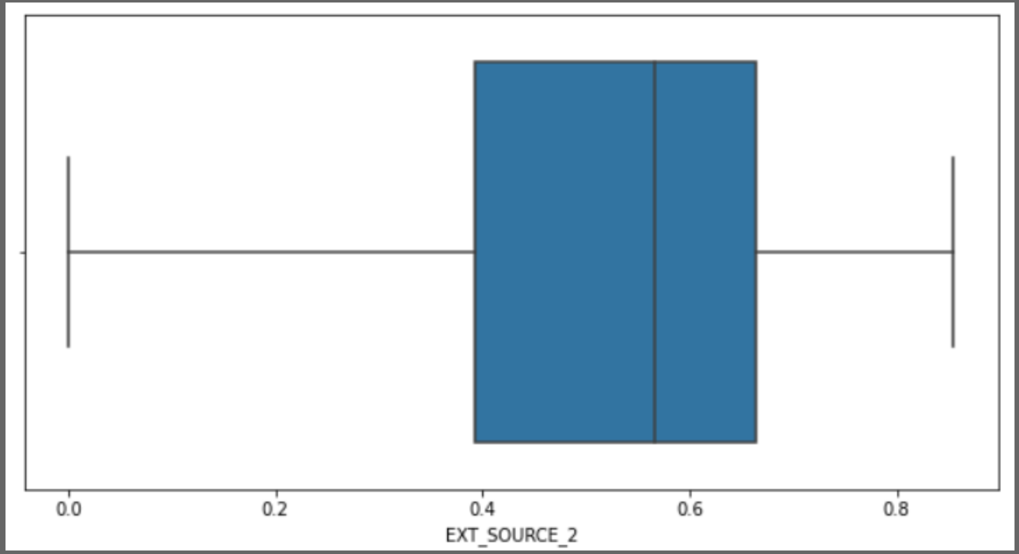
Imputation Of Columns:

* EXT_SOURCE_3, EXT_SOURCE_2, AMT_GOOD_PRICE, AMT_ANNUITY, AMT_REQ_CREDIT_BUREAU_QRT, AMT_INCOME_TOTAL and AGE.



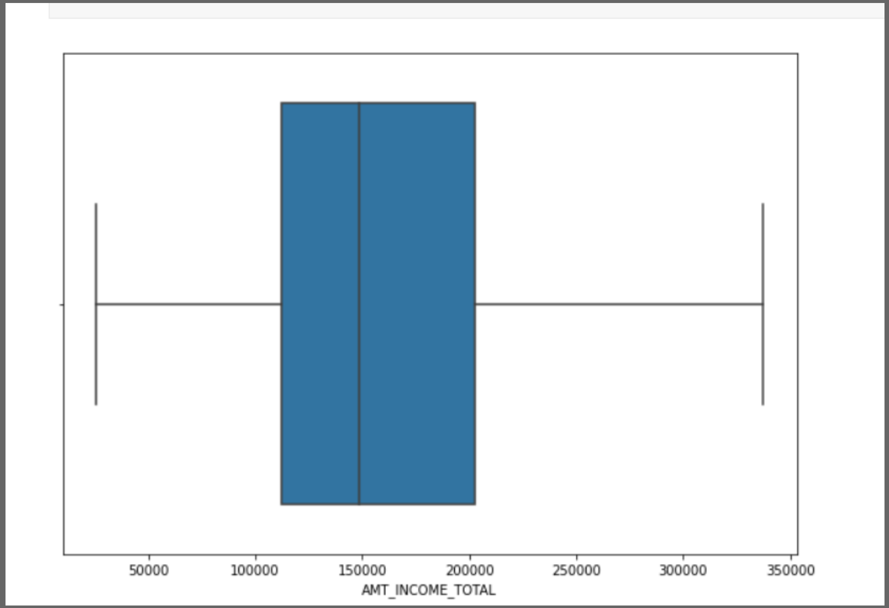


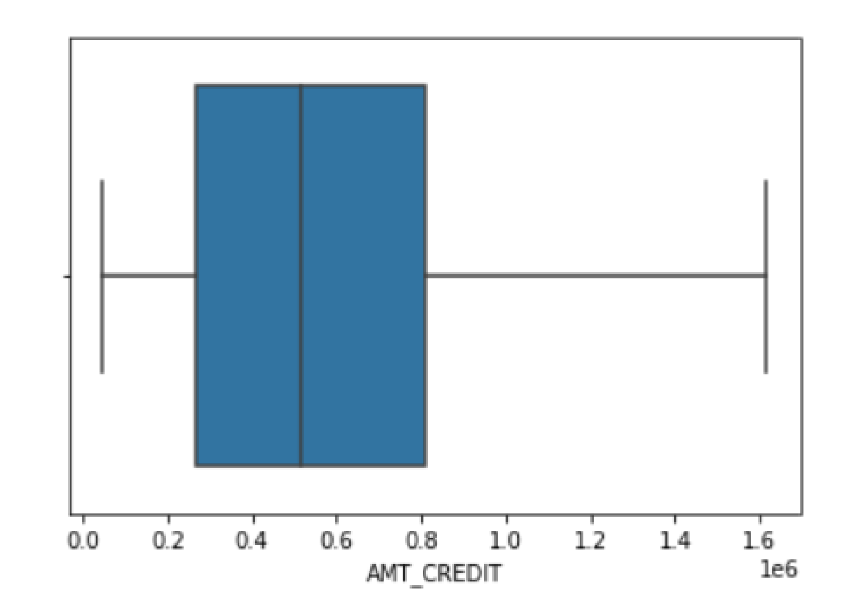




- EXT_SOURCE_3: There is not much more difference between mean and median value while showing the describe() method. So, I went with mean value imputation.
- EXT_SOURCE_2: There is no outliers and very less missing value. I went with mean() as well.
- OCCUPATION_TYPE: This is object data type column where most of the columns values are filled with Labour but when I checked the no of missing value which is really higher than the labour value present in that column. We can't take mode of it. So, I left the missing value as it is.
- AMT_GOOD_PRICE: Here the std is very high, we simply can't put the mean or median value here. So, I just dropped those rows which are having NaN value. Also, the no of missing value rows are really less, nearly 0.09% of total column.
- AMT_REQ_CREDIT_BUREAU_QRT: Here I took the mode as this is also a object type datatype and most widely used value was 0.0. So, I just replaced 0.0 in place of missing value in this column.
- Same goes with AMT_REQ_CREDIT_BUREAU_WEEK, YEAR and MON.
- HANDLING ERRORS: DAYS_BIRTH, DAYS_EMPLOYED, DAYS_ID_PUBLISH column used to having negative values. I have converted them to absolute values.

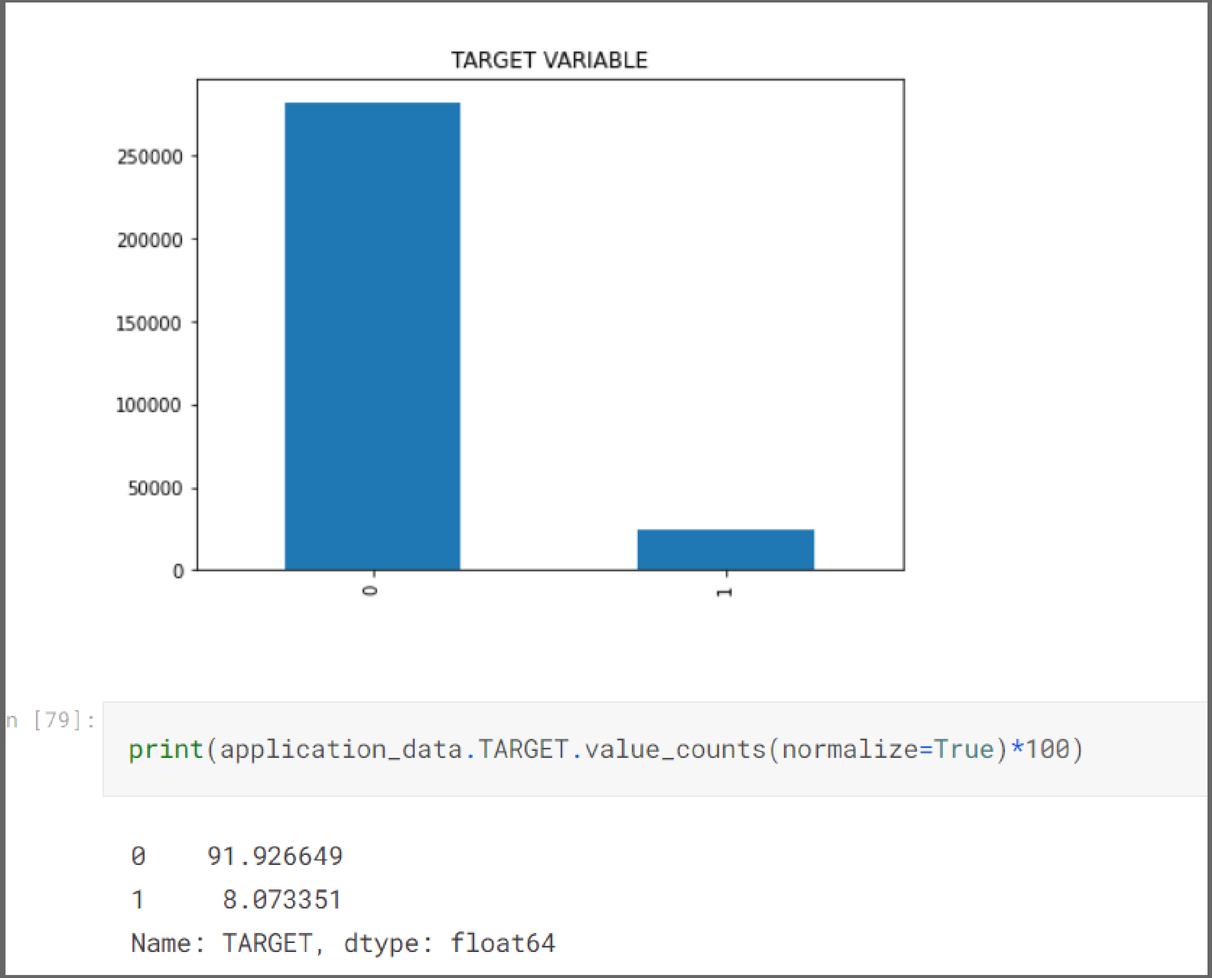
- In EXT_SOURCE_2 there is no outliers.
- For other columns like AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE outliers has been handle with the help of calculation of IQR.
- Formula Used for this are :
- IQR = Q3-Q1
- Q1 = Datafrme[‘column’].describe()[’25%’]
- Q1 = Datafrme[‘column’].describe()[’75%’]
- After handling Outliers, boxplots looks like this:





Data Imbalance Checking:

* More than 90% of people are coming under Non-Defaulter category where as only 8% of people are coming under Defaulter category.



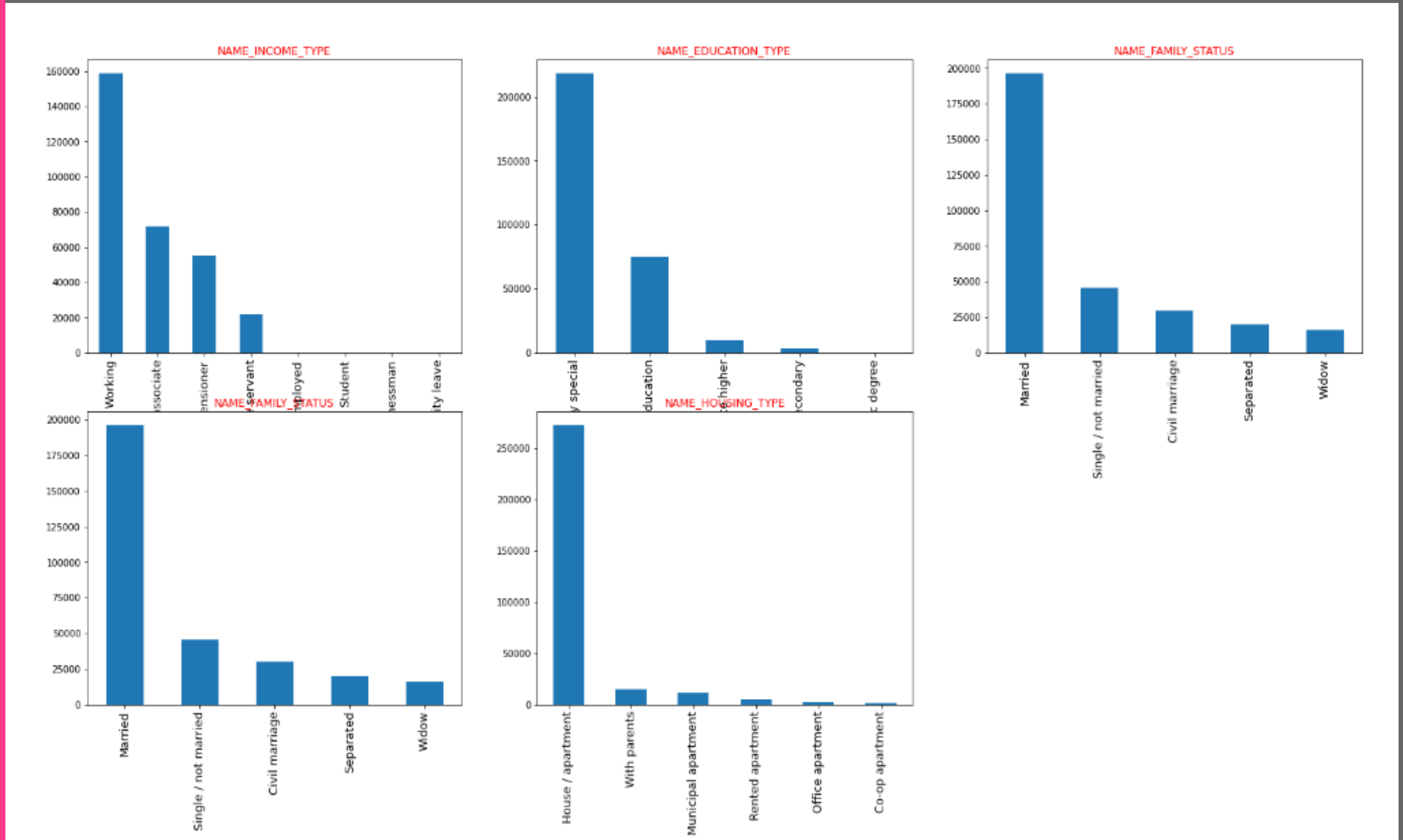
```
n [79]:  print(application_data.TARGET.value_counts(normalize=True)*100)
```

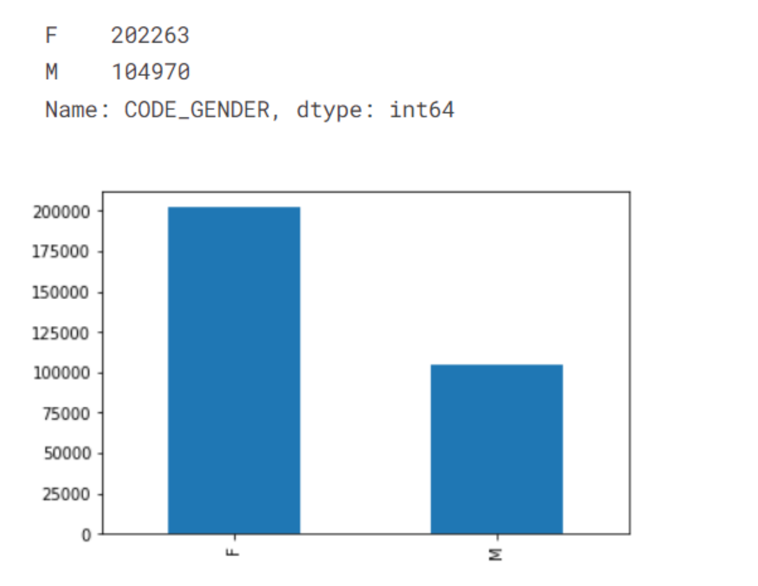
```
0    91.926649
1     8.073351
Name: TARGET, dtype: float64
```

* With the help of TARGET Variable/Column, I have checked some of the relationship using Bivariate and Univariate Analysis.

Graphical Analysis Of Variables:

- We can see that Female loanees are higher as compared to male loanees.
- Most Of the loanees are working professional, Secondary education, Married and having their own house/apartment.
- Most of the Flag_Documents were submitted correctly escepts some like Document_6,3,and 8. (Image not attached here)

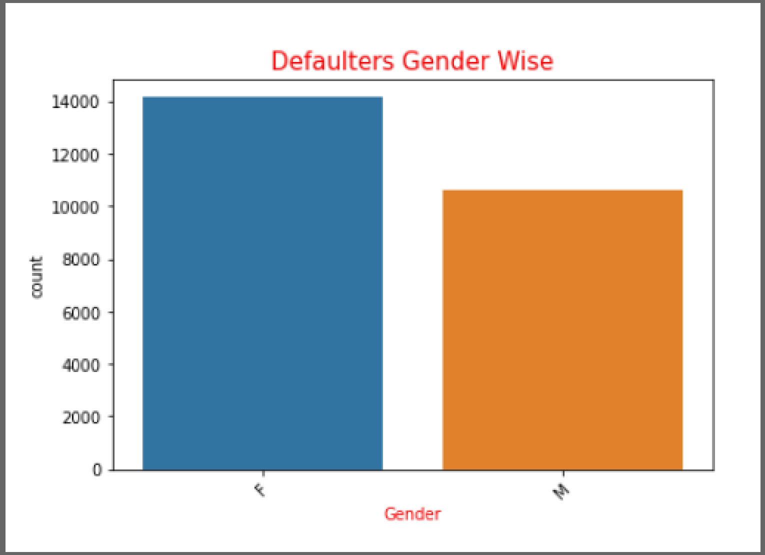


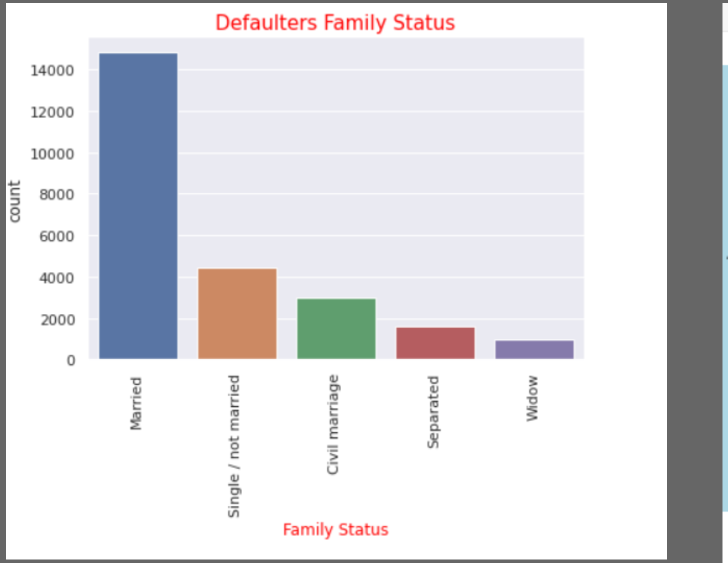


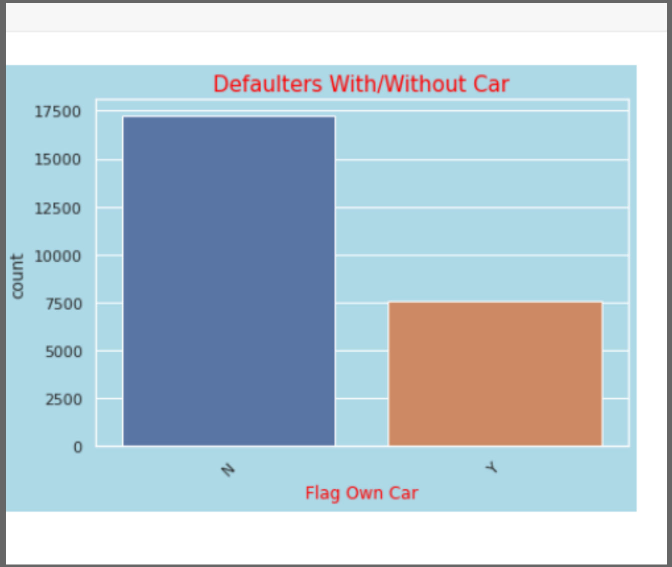
```
F    202263
M    104970
Name: CODE_GENDER, dtype: int64
```

Univariate Analysis Using Target Variable:

- Females are more in defaulters list compared Male.





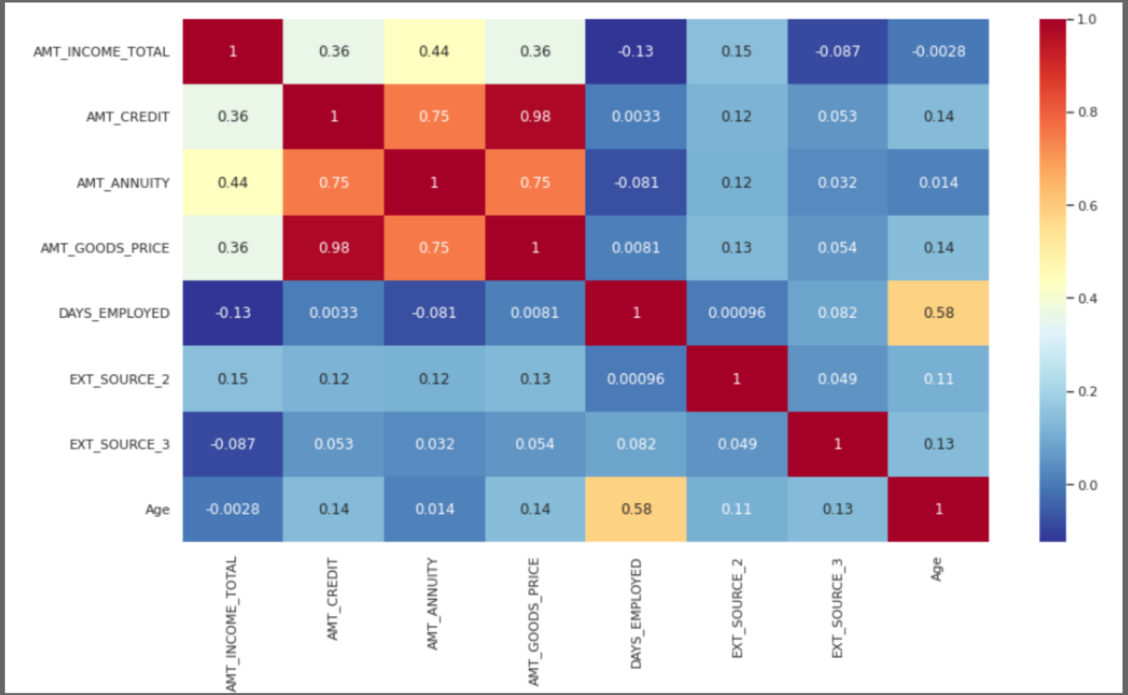


- Most of the defaulters are not having any car.
- Most of the defaulters are having their own House/apartments.
- Most of the defaulters are married and than single people are struggling.
- Most of the defaulters income type is working and then commercial associate.
- Most of the defaulters education type is secondary/ secondary special education.
- Cash loans contract type defaulters are high compared to revolving loans.

- Most of the defaulters are earning between 1lakhs to 1.7lakhs annually.
- The credit amount of the loan, the more chances of being a defaulter. We can see that the spike is high till 0.50 which is 500000.

CORRELAIONS:

- For Defaulters, there is a strong relation between credit amount and goods price amount. Which is nearly 98%.
- Good price and annuity amount is having strong correlation of 75% which is same as credit amount and annuity amount.
- Age and Days Employed also not strong not weakly correlated but somewhere in the medium. They are 58% correlated.

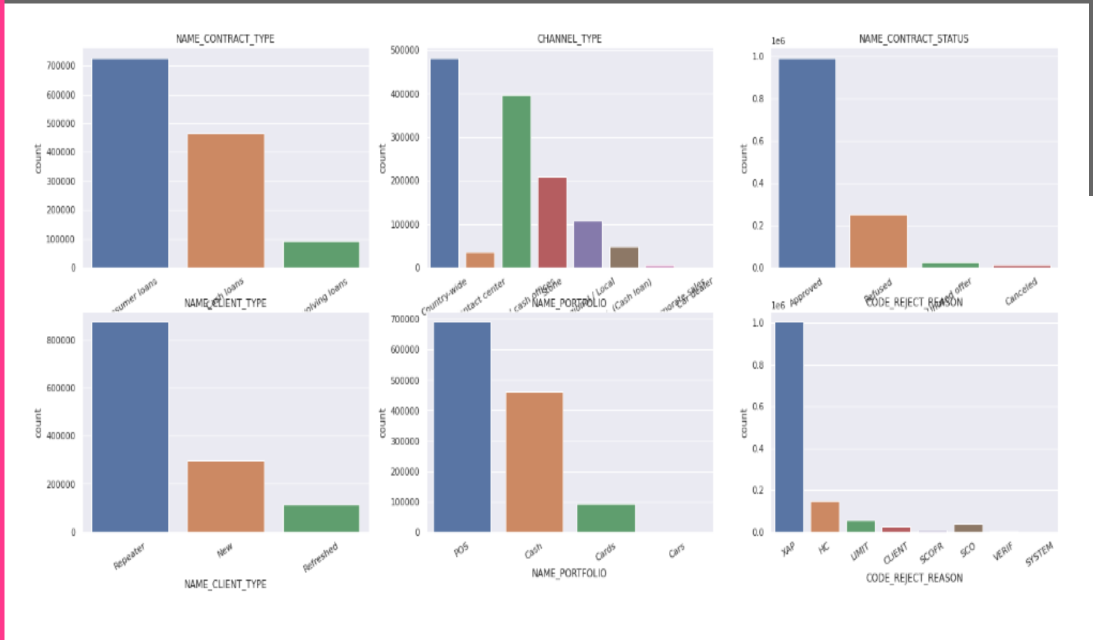


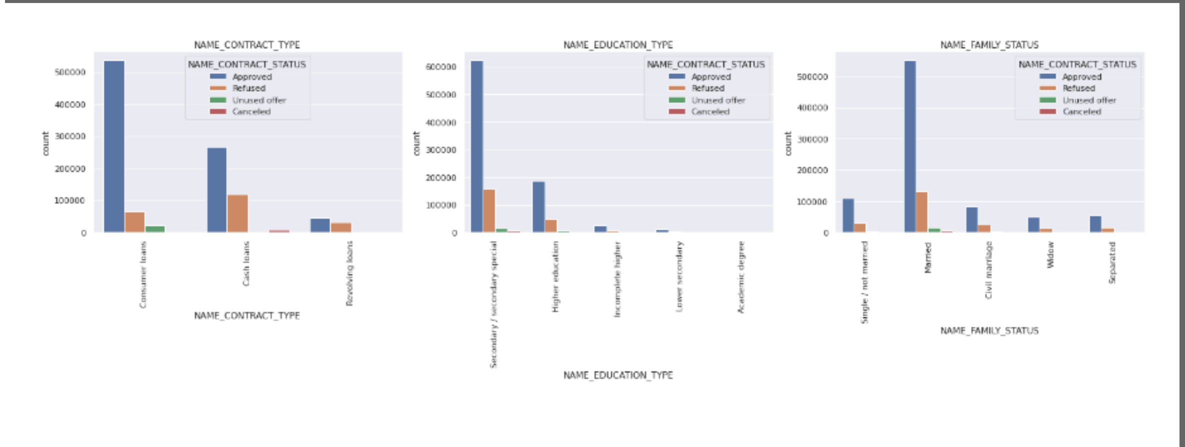
Bivariate Analysis:

- Defaulters: Female Academic Degree holders are earning more but still they are in defaulter list more. Where as not s single Academic male holders are in defaulter list.
- Defaulters: There are more congested values in initial or lower areas of both AMT_CREDIT and AMT_INCOME_TOTAL. With the Income the LOAN value also increases.
- But for rest of the education type we can see that males are earning more as well being the more defaulters.
- Cash loans by males are more compared to female and they are more defaulters as well.
- We can see that Unemployed people are more defaulters in both male and female case. Maternity leave females are also in higher no in defaulters list.
- The occupation type core staff who owns of House/apartment are more in number of non-Defaulters whereas the occupation type Laborers who owns of House/apartment are high in number to likely to default.

Previous Data Analysis:

- Here also more no of female loanees compare to male. So, they the one with most defaulters.
- Secondary/ Secondary special education people are struggling to pay their loan bills and are defaulters.
- Repeater clients are getting approved more compare new clients.
- Consumer contract types are approved more compared to other contract type.
- Highest no of loans got approved to those who are married.





Merge Data Analysis:

- In Contract type, female with consumer loans type are in more no of defaulters.
- Married people are struggling to pay the loan bill compared to single and separated people. More no of approved people are married. From next time onwards we need to give loan to those who are either single or divorced.
- Secondary/ secondary special people with education are struggling to pay their loan bill and becoming defaulters.
- Lower credit amount people else really higher credit amount people are having high chance to becoming a defaulters.
- Female Gender are more likely to not face payment difficulties then the male and hence it is recommended to approve more loans of Female Gender than the male gender at the same Female are High in number than who face difficulties than males.
- The Repeater applicant has High chance of non-Defaulting and also has high chance of defaulting when compared to new applicants.

THANK YOU

Regards

Ramakrushna Mohapatra