RISK ANALYTICS IN BANKING AND FINANCIAL SERVICES

SUBMITTED BY

SUJATA JHA

PROBLEM STATEMENT

- Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. . While processing any loan one may need to think what criteria to follow so that least number of default cases arises.
- More default cases means more loss to the bank. Sometimes bank do not have sufficient credit history for many applicants. Because of that, some consumers use it to their advantage by becoming a defaulter.
- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. So that company can utilize this knowledge for its portfolio and risk assessment.

DATA PROVIDED

- *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.
- Target variable (1 - client with payment difficulties, 0 - all other cases)
- *previous_application.csv:* contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- *Columns_description.csv* : It contains description of all column in above two files.

APPROACH AND METHODOLOGY(1/3)

- Data Understanding and Loading
  - We understand the data from column_description.csv .
  - We load the data as dataset by importing important libraries. Dataset for both the files are created.
  - We check the structure /metadata of the two datasets. This step includes analysing shapes, columns, datatype , description like min, max, median or deviation, quantiles etc.
- Data Cleaning for both datasets
  - Missing value check for datasets: If the percentage of missing values is equal or more than the 40% in a column we drop the columns as imputing such big number of values may change the nature of the data provided for analysis.

APPROACH AND METHODOLOGY(2/3)

We drop the columns which are of less importance related to Target variable.

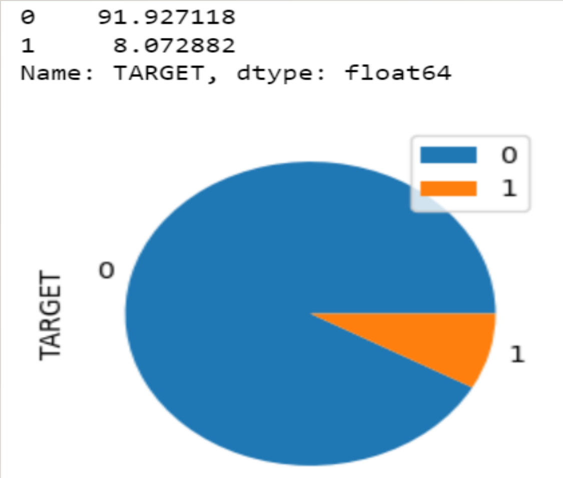
- Imputing Values:
  - For numerical columns impute missing value with mean when the distribution of data is normal for the variable. We impute median value for missing values when the distribution of data is not normal(skewed graph/outliers presence).
  - For categorical columns impute mode value for the columns.
- Standardize values like all days related variables are changed to positive values ex- Birth days are converted to positive values.
- Outliers are just identified.

APPROACH AND METHODOLOGY(3/3)

- Univariate , Bivariate and Multivariate Analysis
  - In both datasets three list of variables of type object, int and float are created. So that plotting is easier. Insights are drawn from the graph.

GRAPHS AND INSIGHTS

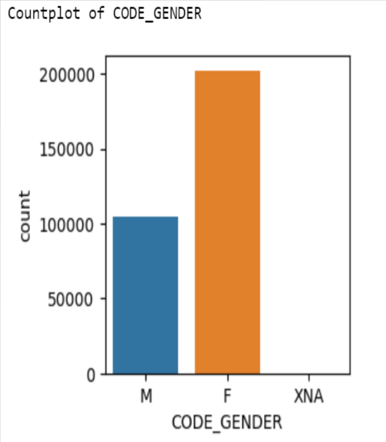
- Data Imbalance for Target variable mentioned in Problem statement: 8% of applicants have payment difficulty. Banks will have less defaulter then only banks can survive. Imbalance is expected.

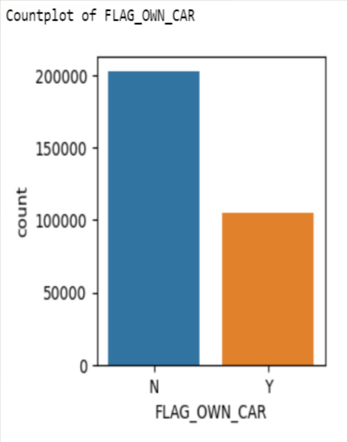


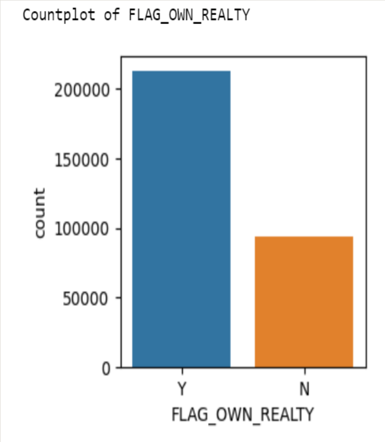
```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

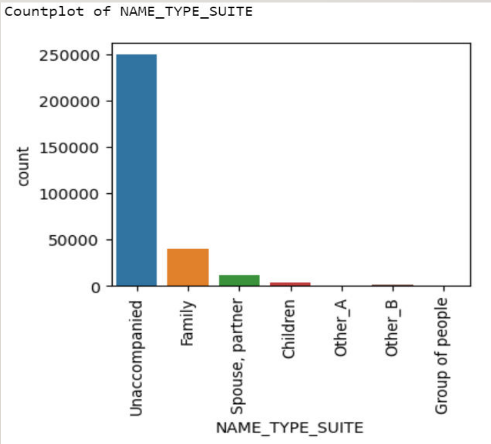
INSIGHTS FROM UNIVARIATE ANALYSIS

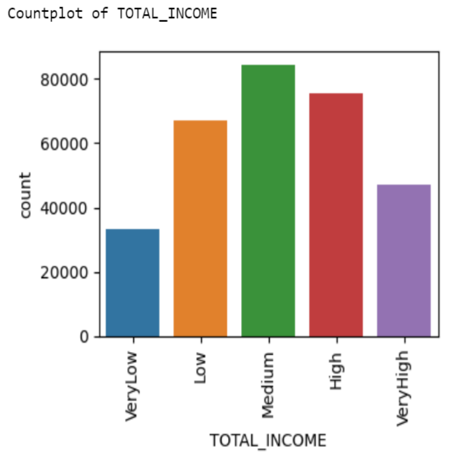
- 65% of applicants are females while males are only 34%. Assuming that this might be due to some off in interest rate given to the Females.
- There are 66% of applicants who do not have car.
- There are 69% of applicants who own a house or flat.
- There are 81% of people who came unaccompanied while applying for loan
- Medium and High salary people are more likely to take loan.





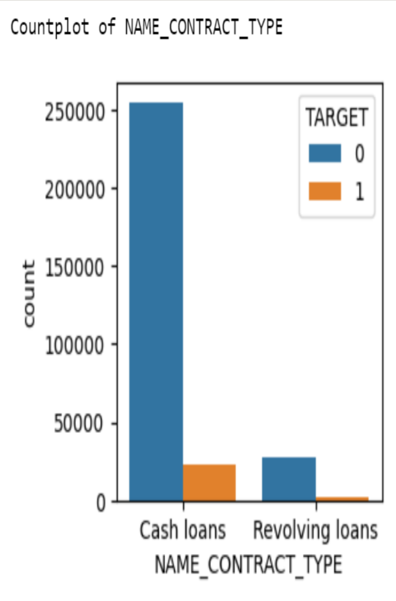


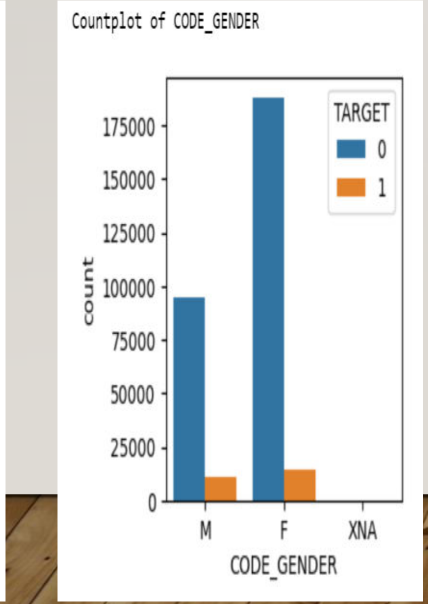


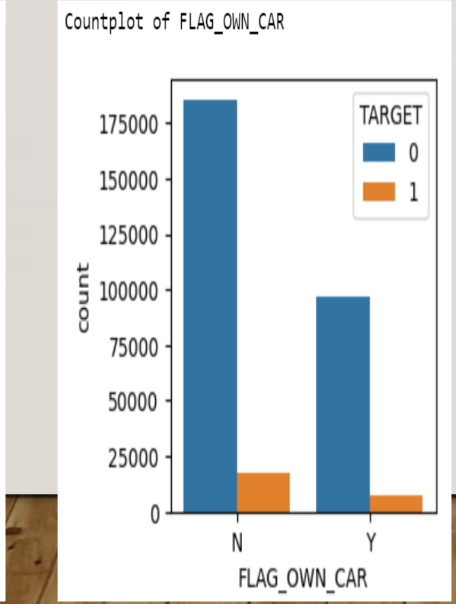


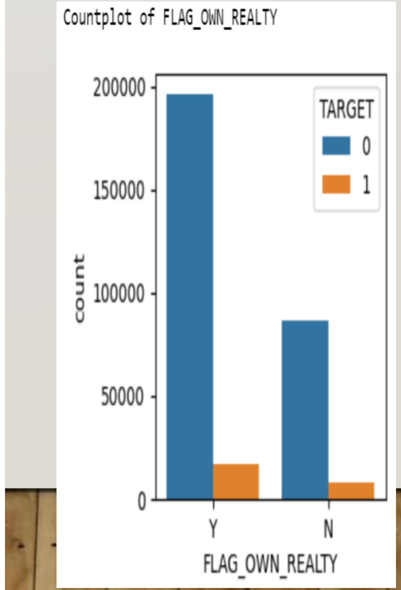
INSIGHTS FROM BIVARIATE ANALYSIS

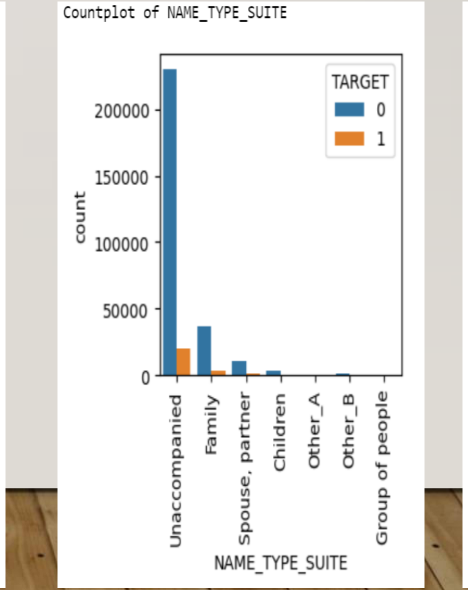
- Cash loans are more smoothly paid than Revolving type. Revolving loans are risky.
- Compared to males, females are paying loans without difficulty. So giving loan to males is more risky.
- Applicants who don't have car/own house are paying loans without difficulty.
- Applicants who came unaccompanied during loan registration are more reliable.
- High salary applicant is more risky than medium salaried applicant.

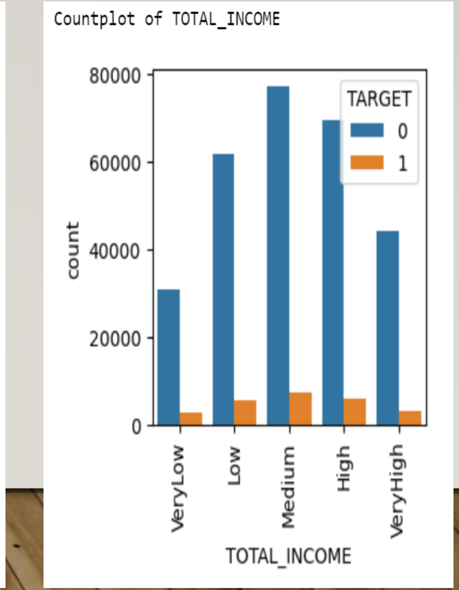

Countplot of NAME_CONTRACT_TYPE


Countplot of CODE_GENDER


Countplot of FLAG_OWN_CAR

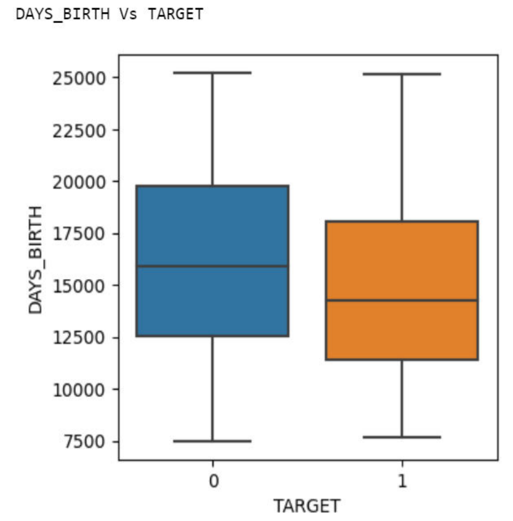

Countplot of FLAG_OWN_REALTY

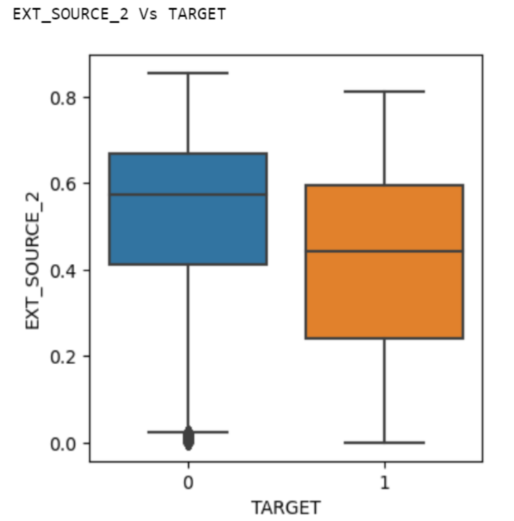

Countplot of NAME_TYPE_SUITE

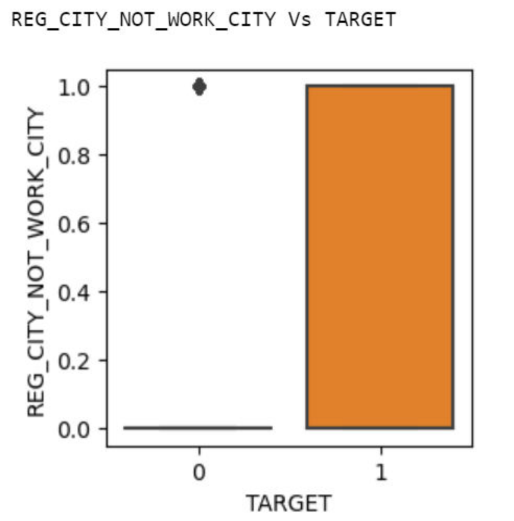

Countplot of TOTAL_INCOME

INSIGHTS FROM BIVARIATE ANALYSIS

- Mostly loan is taken by applicants age from 12000 to 17000 days...approx. 30 to 45 years
- More the age more is the density in the loan payment. Middle age around 50% of quartile shows the payment difficulty as well.
- Applicant with EXT_SOURCE_2 more score are more likely to pay loan without difficulty.
- REG_CITY_NOT_WORK_CITY Vs TARGET indicates that applicant with different permanent address and work address have high density for default.

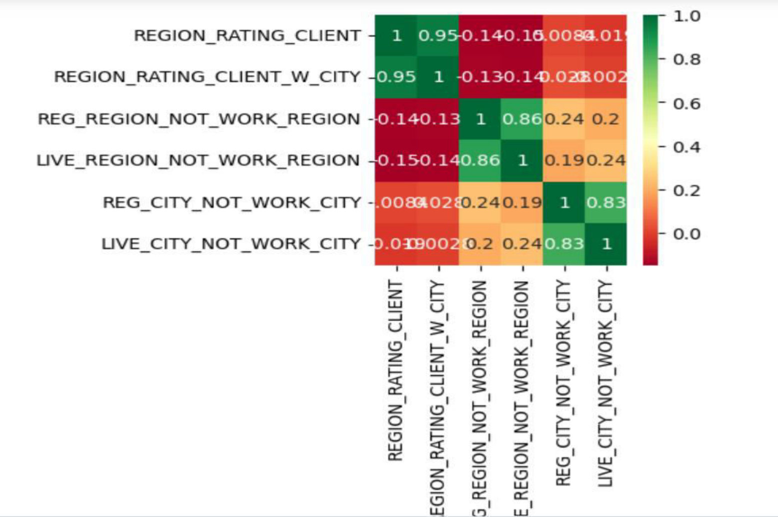


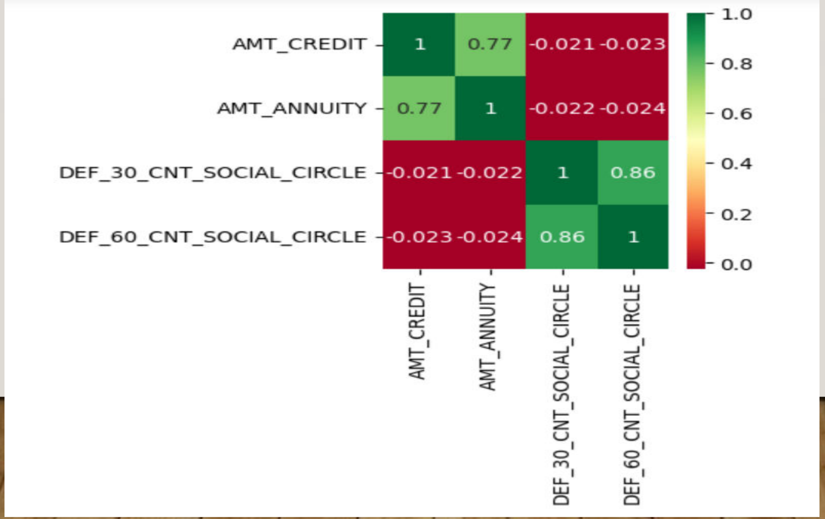


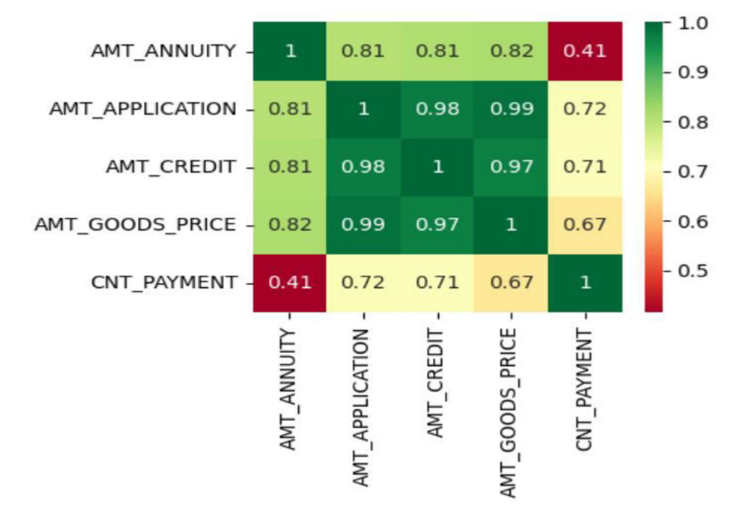


INSIGHT FROM MULTIVARIATE ANALYSIS

- Positive linear correlation is observed among below variables
  - REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY
  - REG_REGION_NOT_WORK_REGION and LIVE_REGION_NOT_WORK_REGION
  - REG_CITY_NOT_WORK_CITY and LIVE_CITY_NOT_WORK_CITY
  - AMT_CREDIT and AMT_ANNUITY
  - AMT_ANNUITY and AMT_GOODS_PRICE
  - DEF_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE
  - AMT_ANNUITY, AMT_APPLICATION,AMT_CREDIT,AMT_GOODS_PRICE have high positive correlation.

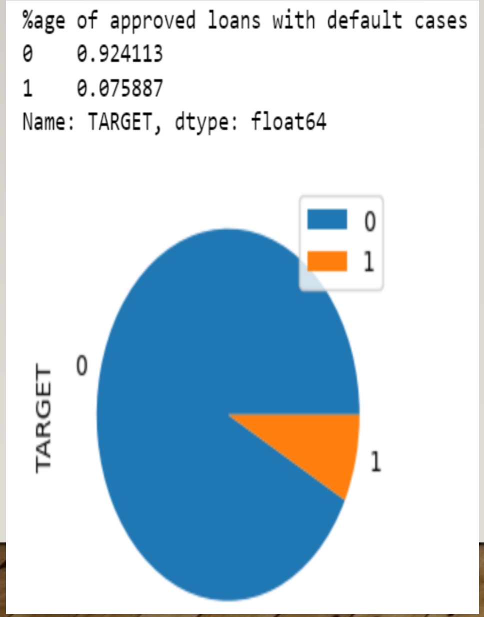




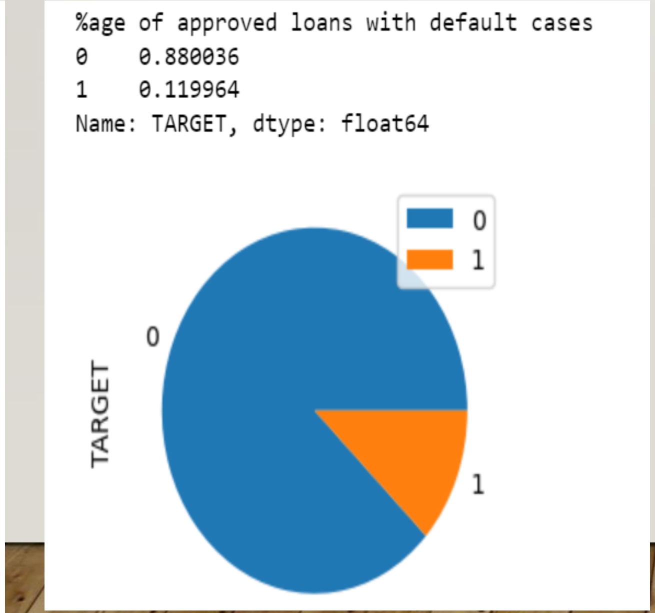


INSIGHTS FROM MERGED DATA(BOTH DATASETS)

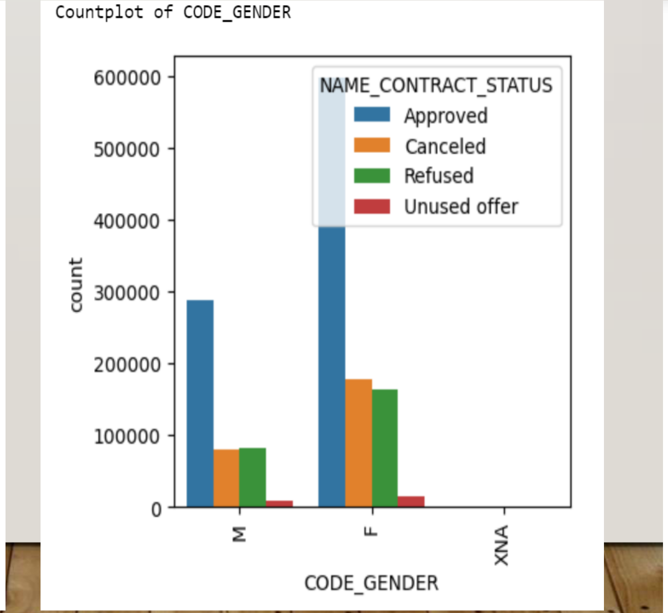
- There are around 8% of default cases in approved loans. This is a loss to banks.
- 88% of the refused applicant were able to pay the loans. Only 22% could be the default case. This is also considered as loss for banks.
- More females got approved loans as well.
- Mostly medium and high income applicants got loan approved.



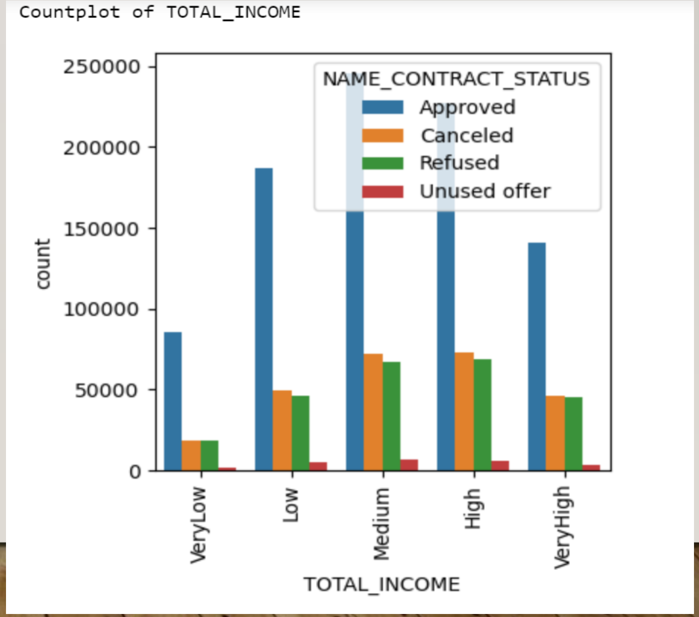
```
%age of approved loans with default cases
0    0.924113
1    0.075887
Name: TARGET, dtype: float64
```



```
%age of approved loans with default cases
0    0.880036
1    0.119964
Name: TARGET, dtype: float64
```



```
Countplot of CODE_GENDER
```



```
Countplot of TOTAL_INCOME
```

CONCLUSION

- Deriving variables for default cases
  - Males are more likely to be default cases.
  - EXT_SOURCE_2 less score shows more default cases.
  - Credit amount has strong relation with annuity amount, goods price. Higher amount may lead to higher loss as well.
  - High and Medium salaried people get the most of the approved loans. High salary applicant is more risky than medium salaried applicant.
  - Applicants who came unaccompanied during loan registration are more reliable.
  - Applicants who have car/own house are paying loans with difficulty
  - Applicant with different permanent address and work address have high density for default.
  - Revolving loans are risky

SUGGESTION

- There are 31% of Missing data for occupation type. Well this is important criteria for financial firms to consider.
- Purpose of the loan has a lot values as not known. Might be these values are not shared by the clients.
- 80% of Loan rejection reason is not specified. Banks need to work on it(might be hidden)
- 61% of payment is done via cash. There is a huge percentage of XNA which means data is missing it seems or not shared by applicants. If we achieve these data we can work on for reducing defaulters.

THANK YOU!!