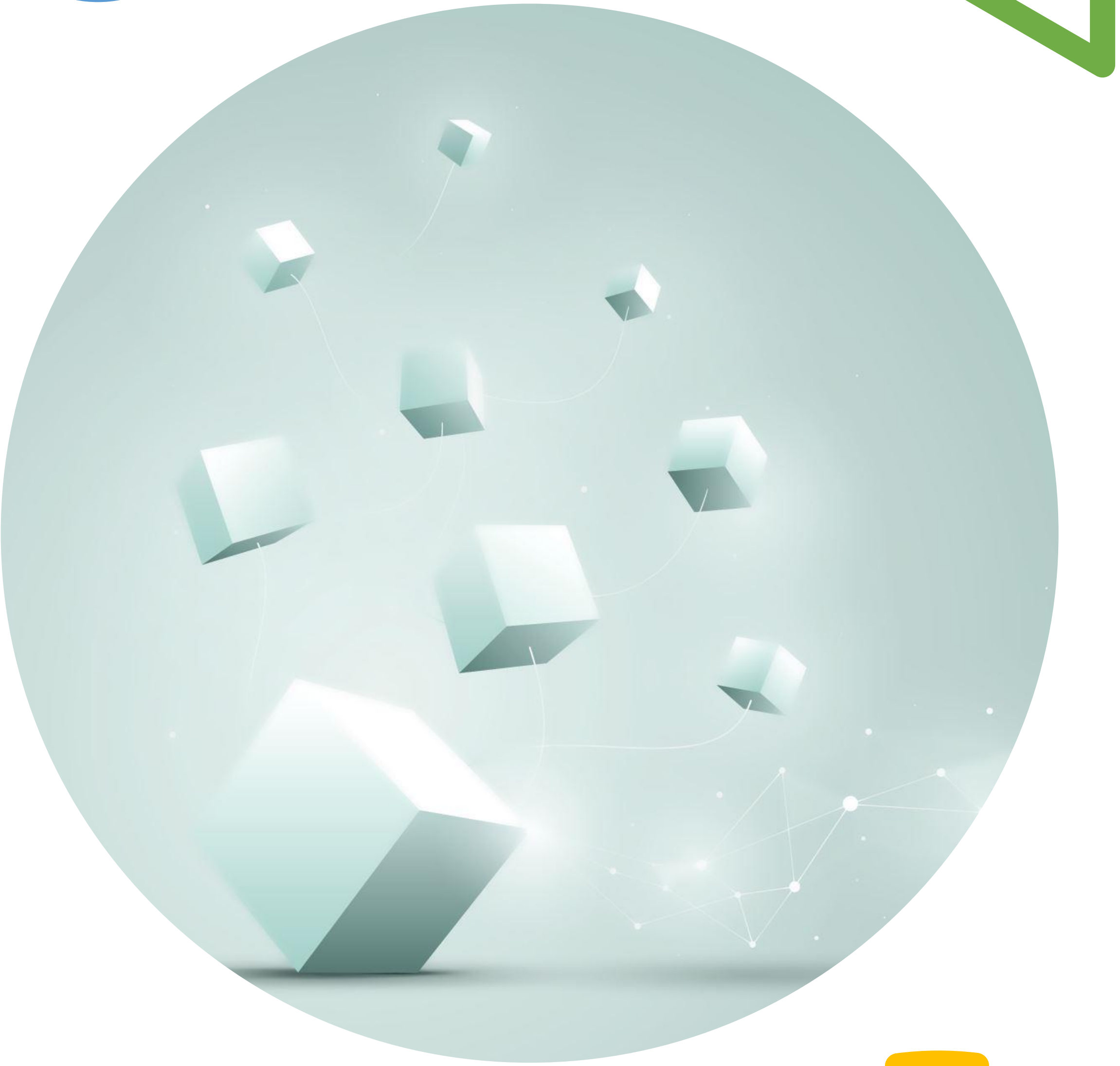


Credit EDA Case Study

By – Sakshi Agarwal

Problem Statement

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 utilize 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).

EDA Steps

- Understanding the Domain/Variables
- Read/Load the Dataset
- Inspect the Dataframe/Check the structure of the data
- Handling Missing Values
- Handling Incorrect Column Datatypes
- Handling Outliers
- Handling Incorrect Values in data
- Check the Imbalance Percentage/Ratio
- Perform Univariate Analysis
- Perform Bivariate Analysis

Handling Missing Values

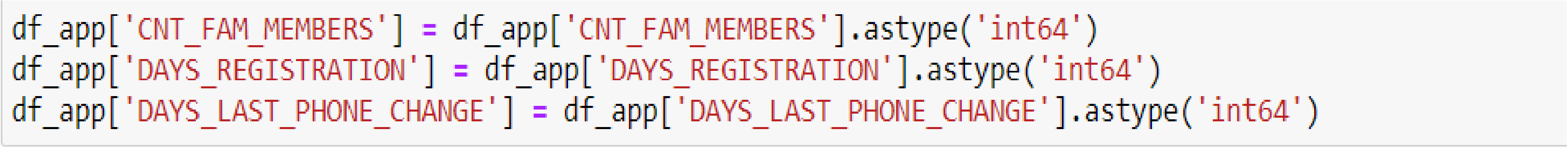
Since there are lot of missing values in our dataset so below techniques are used to handle missing values –

1. Drop all the columns where missing values > 50 percent.
2. Remove all the records where missing value lies between 10 and 20 percent.

Handling Incorrect Column Datatypes

There were multiple columns with incorrect datatypes which has been converted to its relevant datatype.

Float Columns converted to Integer Datatypes.

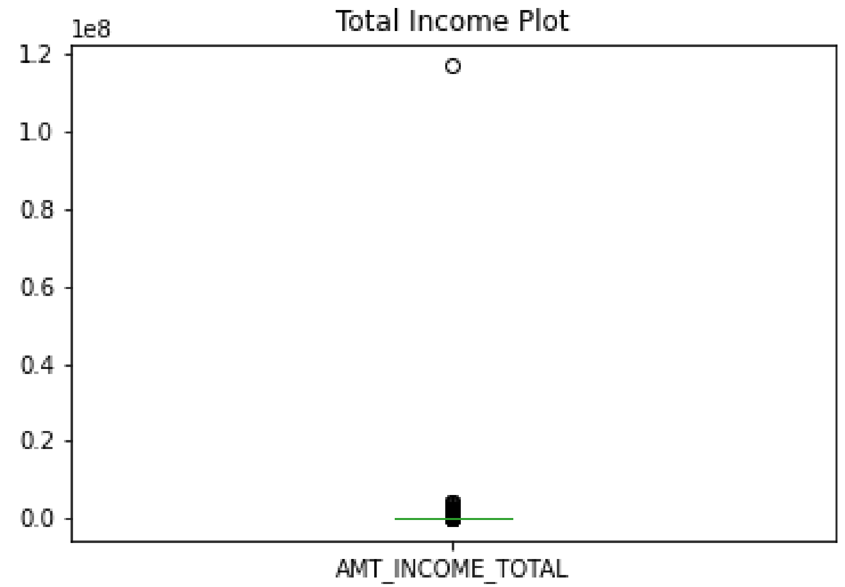


```
df_app['CNT_FAM_MEMBERS'] = df_app['CNT_FAM_MEMBERS'].astype('int64')
df_app['DAYS_REGISTRATION'] = df_app['DAYS_REGISTRATION'].astype('int64')
df_app['DAYS_LAST_PHONE_CHANGE'] = df_app['DAYS_LAST_PHONE_CHANGE'].astype('int64')
```

Handling Outliers

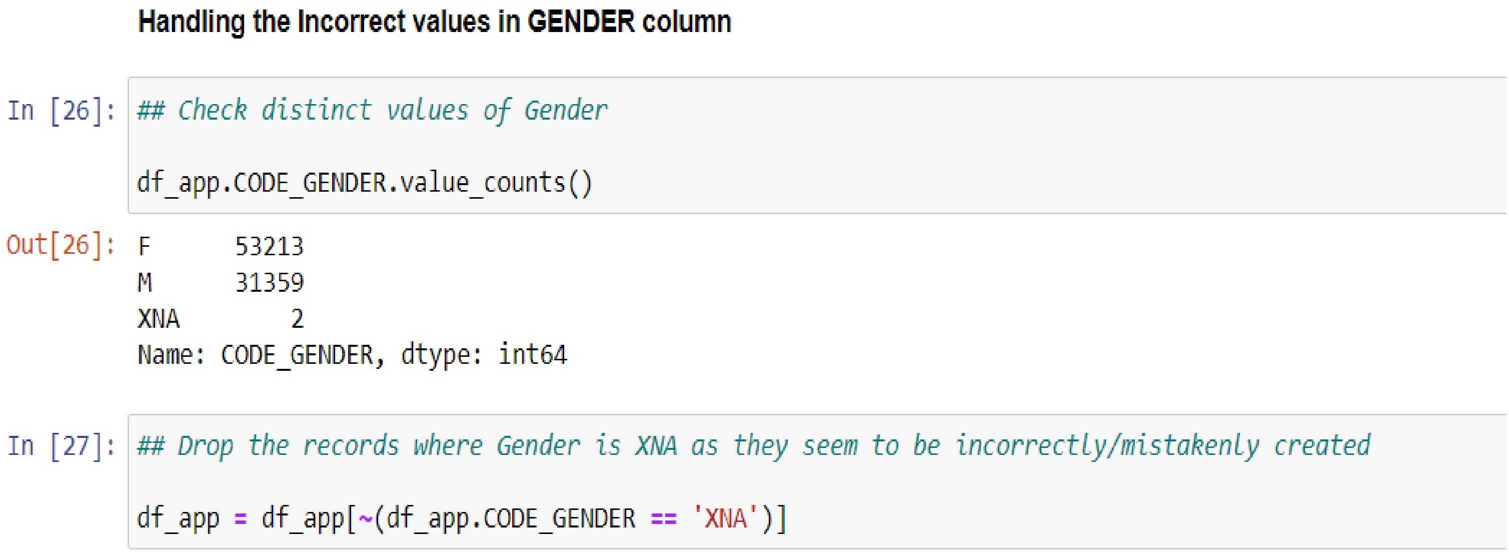
Outliers are the data points which are extremely high/low as compared to other data points in our dataset. They can be either removed/replaced with some other value.

For this case study – we have a client whose Total Income is close to 120M has been dropped from the dataset.



Handling Incorrect Values in Data

Incorrect Values namely XNA existed in GENDER column, so we removed them from the dataset.



Handling the Incorrect values in GENDER column

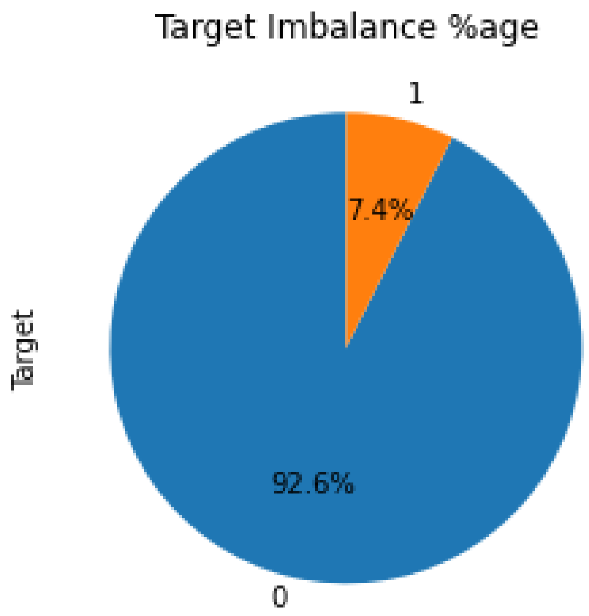
```
In [26]: ## Check distinct values of Gender

         df_app.CODE_GENDER.value_counts()

Out[26]: F      53213
         M      31359
         XNA        2
         Name: CODE_GENDER, dtype: int64
```

```
In [27]: ## Drop the records where Gender is XNA as they seem to be incorrectly/mistakenly created

         df_app = df_app[~(df_app.CODE_GENDER == 'XNA')]
```

Imbalance Data Ratio

Imbalance Data is when one variable tends to have higher count/percent compared to the other.

For our case study – we have similar kind of behavior where no of clients with Payment difficulties have less datapoints as compared to other cases.



Strategy for Univariate/Bivariate Analysis

We would divide the whole dataset on the two types of TARGET variable and continue our analysis based on whether the client has faced any payment difficulties/not.

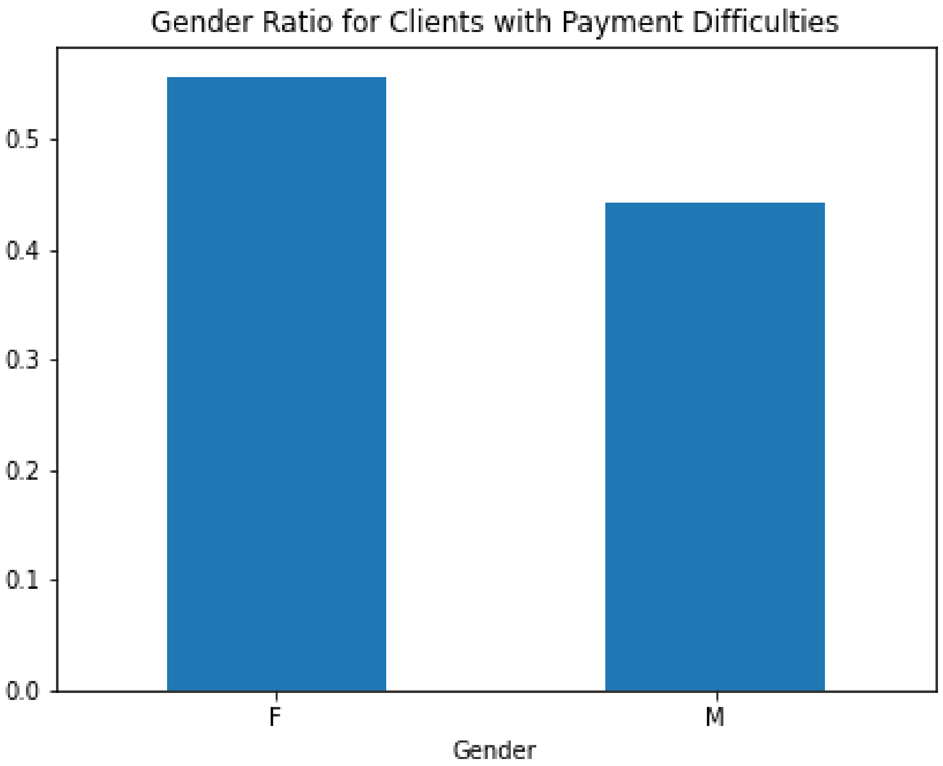
Creating Buckets for Analysis

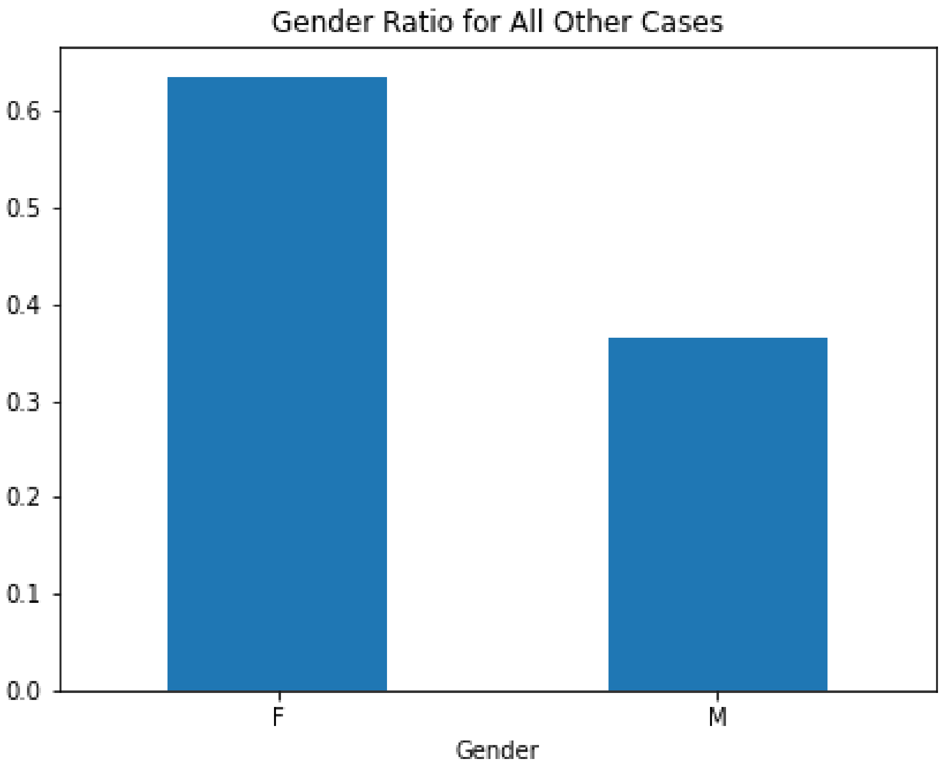
Total Income column has been divided into 5 buckets of below range for the analysis –

1. **VERY LOW** : 0-100000
2. **LOW** : 100000-200000
3. **MEDIUM** : 200000-300000
4. **HIGH** : 300000-400000
5. **VERY HIGH** : 4000000-1000000

Univariate Analysis

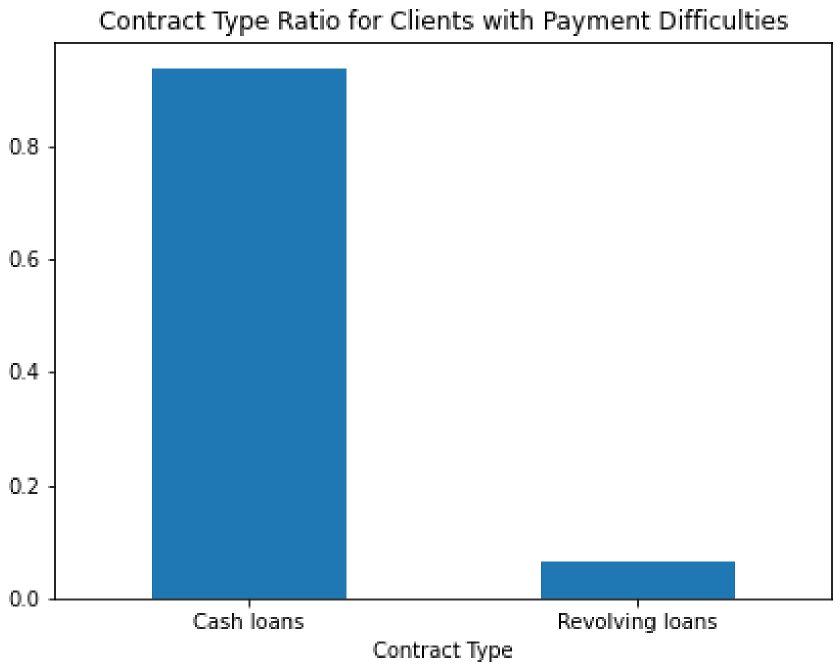
Gender Ratio Plot

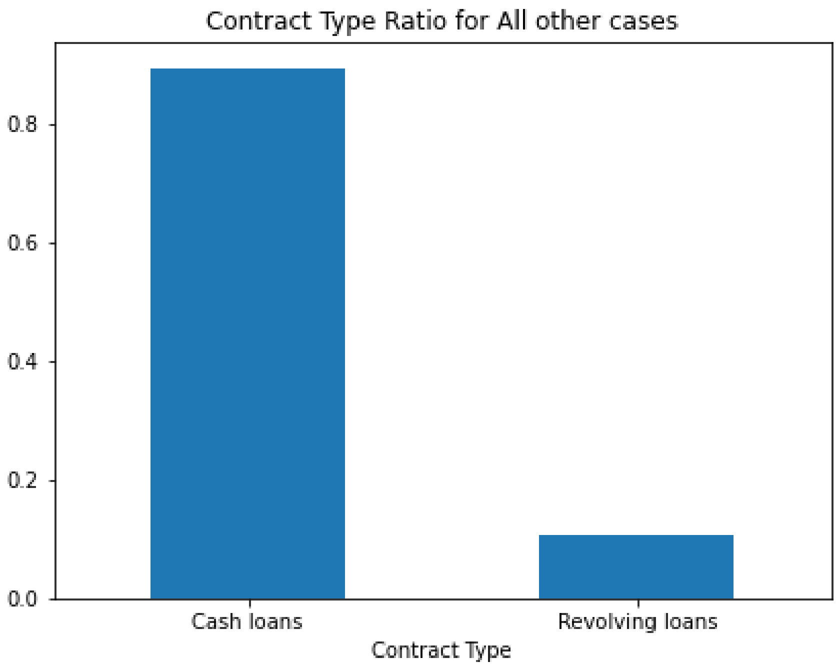




Inference – Ratio of Female to Male client is lesser for loan payment difficulties as compared to other cases.

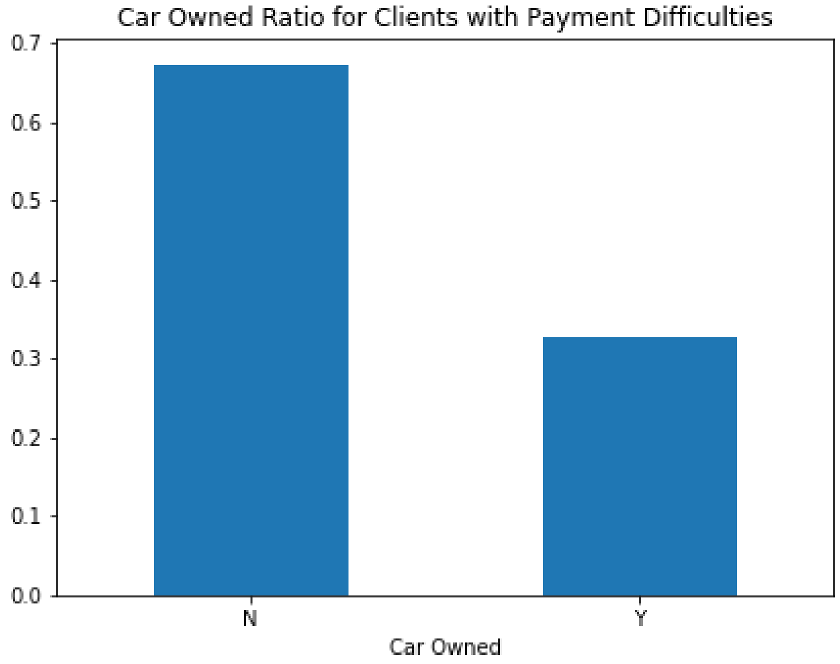
Contract Type Ratio Plot

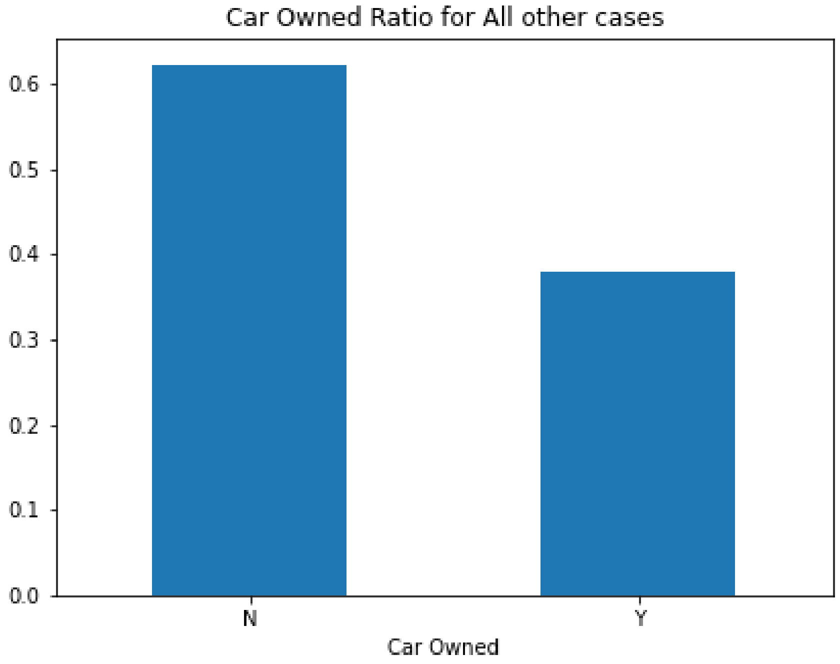




Inference – Revolving Loans Contract Types are lesser in case of loan payment difficulties when compared with other cases.

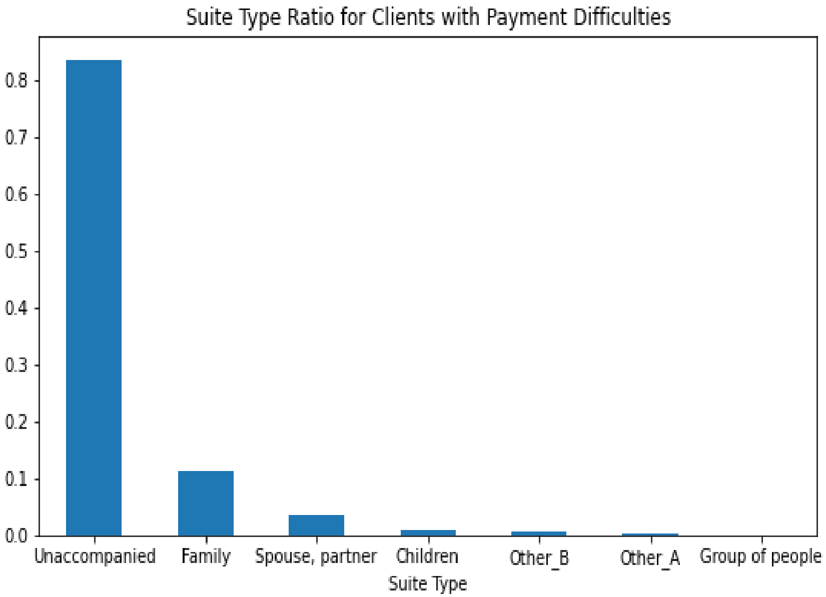
Car Owned Ratio Plot



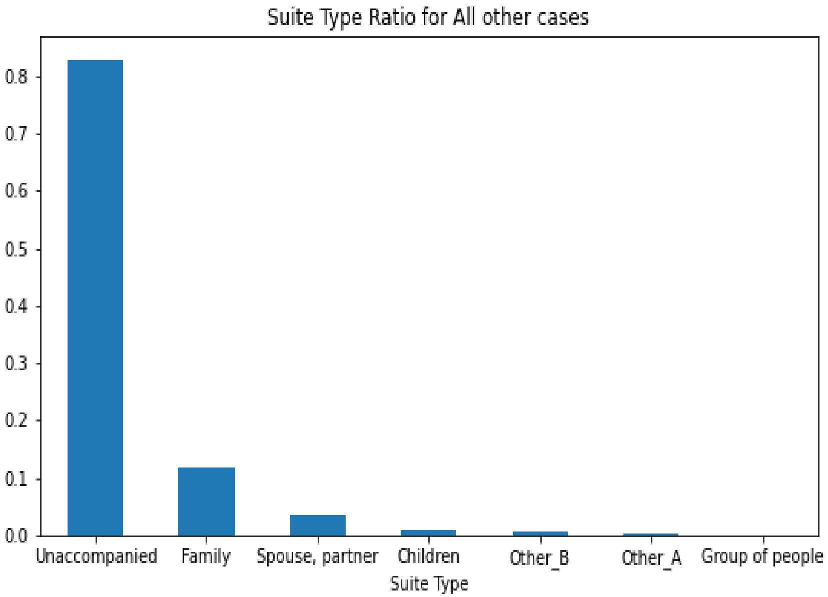


Inference – People who don’t own a car are more in case of Payment difficulties as compared to other cases. Also, car owners are lesser in number when they have payment difficulties.

Suite Type Ratio Plot



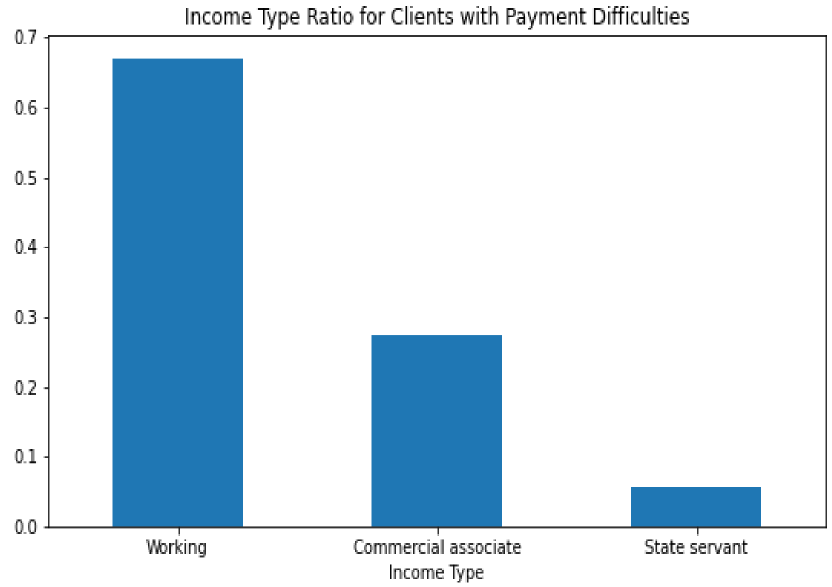
Suite Type Ratio for Clients with Payment Difficulties

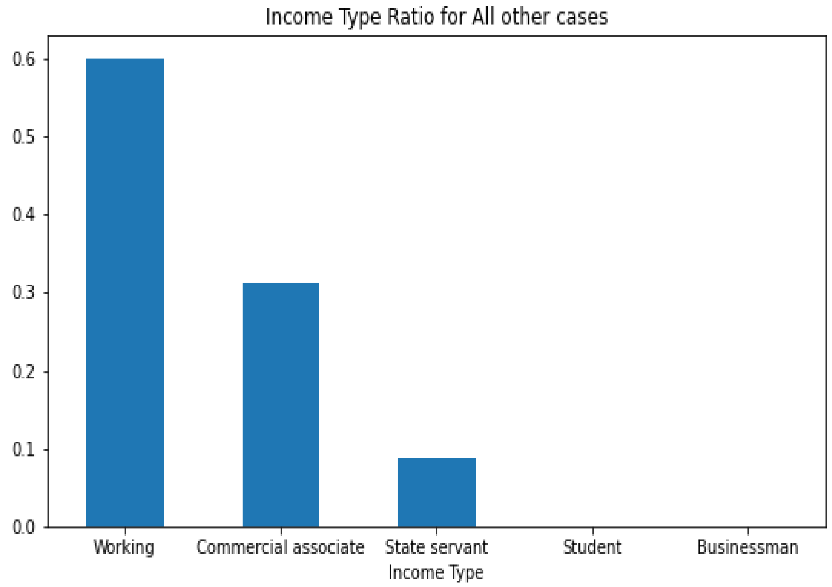


Suite Type Ratio for All other cases

Inference – Suite Type Ratio is almost similar for both the categories.

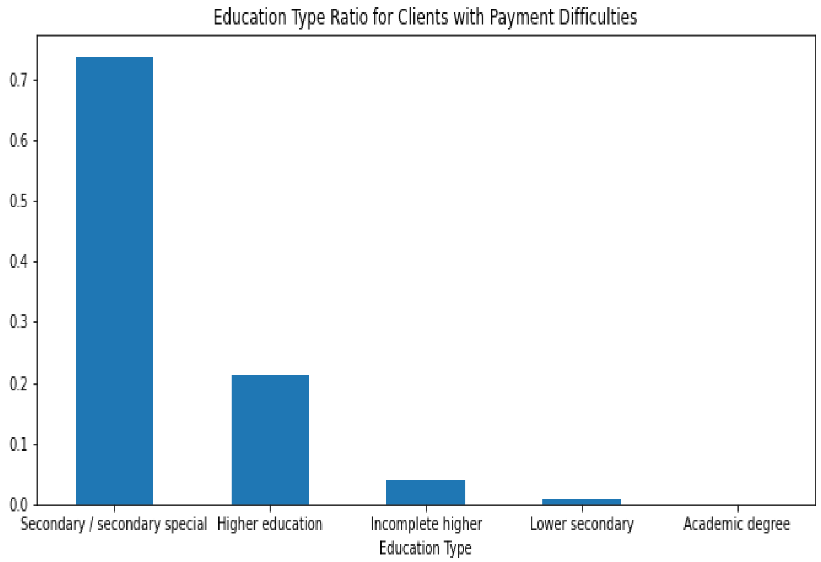
Income Type Ratio Plot

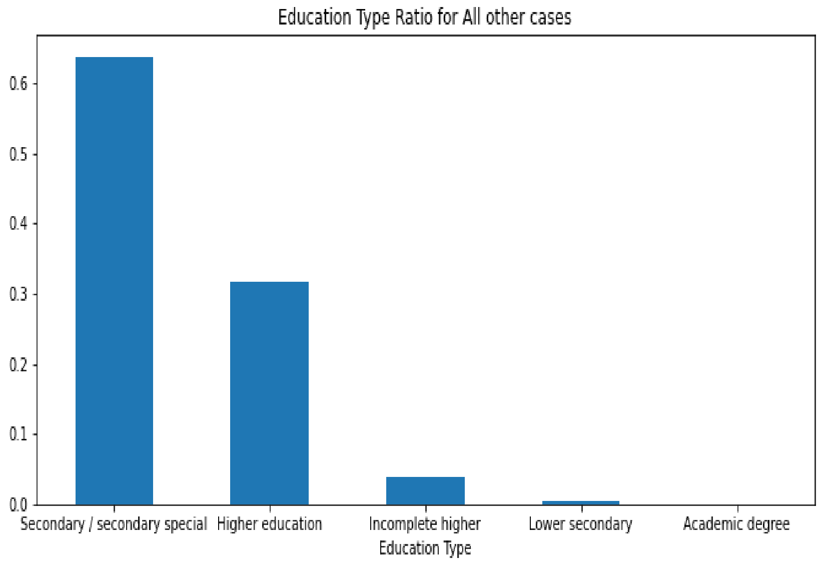




Inference – Working People tend to have Payment difficulties, Commercial Associates have less payment difficulties while Student and businessman do not have any payment difficulties.
Although total number of Student and businessman are very less in the dataset itself.

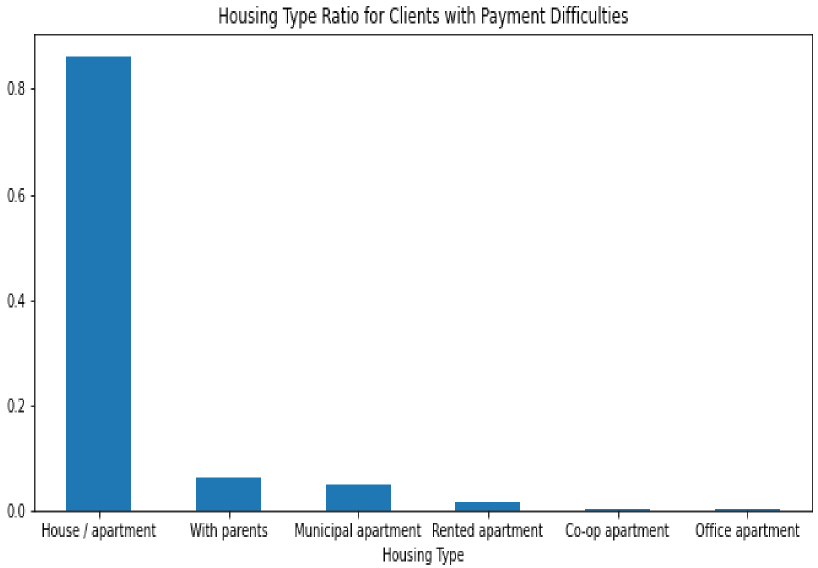
Education Type Ratio Plot

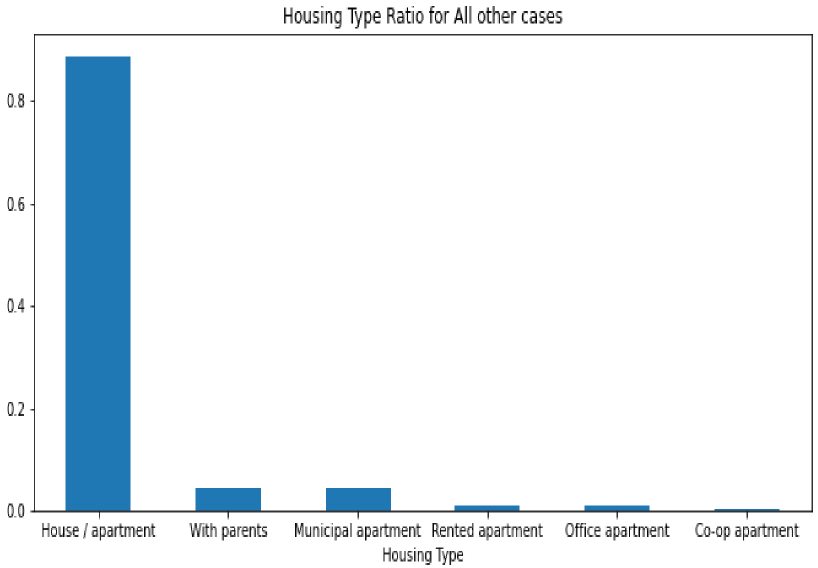




Inference – Secondary/Secondary Special Education Type people tend to have more payment difficulties when compared with others while Higher Education people have less payment difficulties.

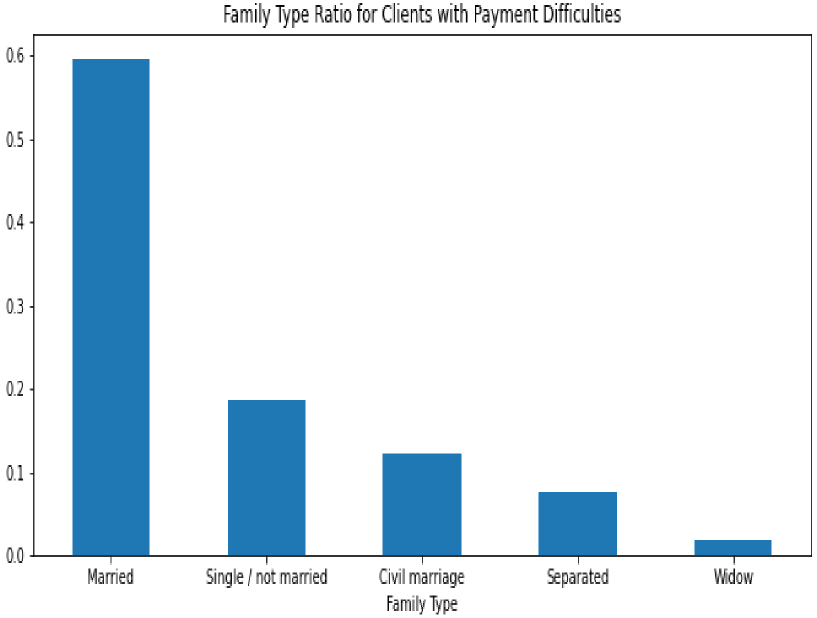
Housing Type Ratio Plot

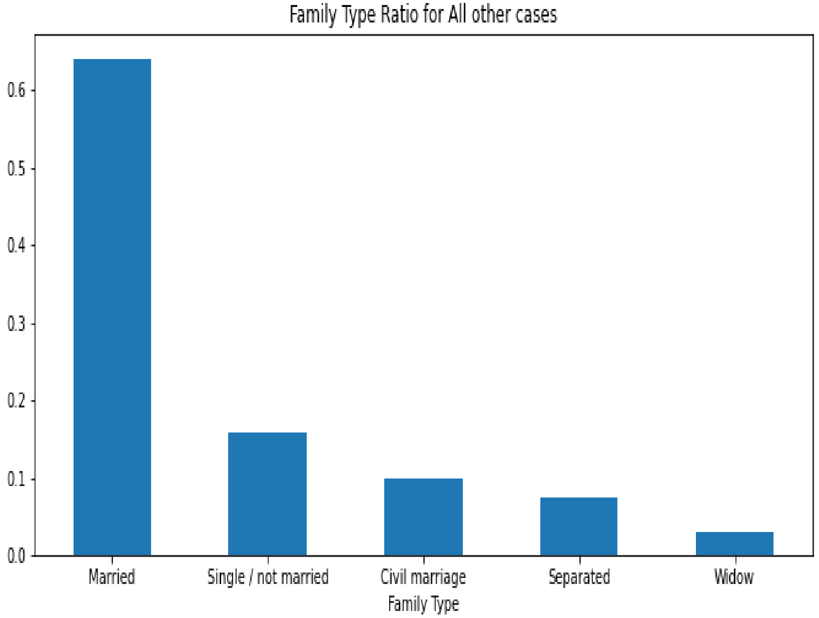




Inference – People living in House/Apartment tend to have more payment difficulties when compared with people living with Parents/Municipal/Rented/Office Apartments.

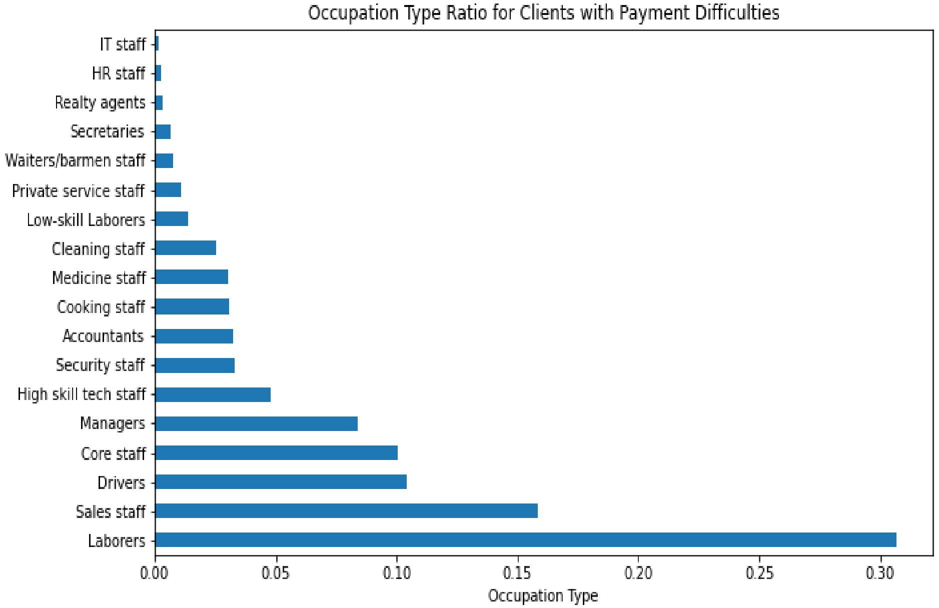
Family Type Ratio Plot

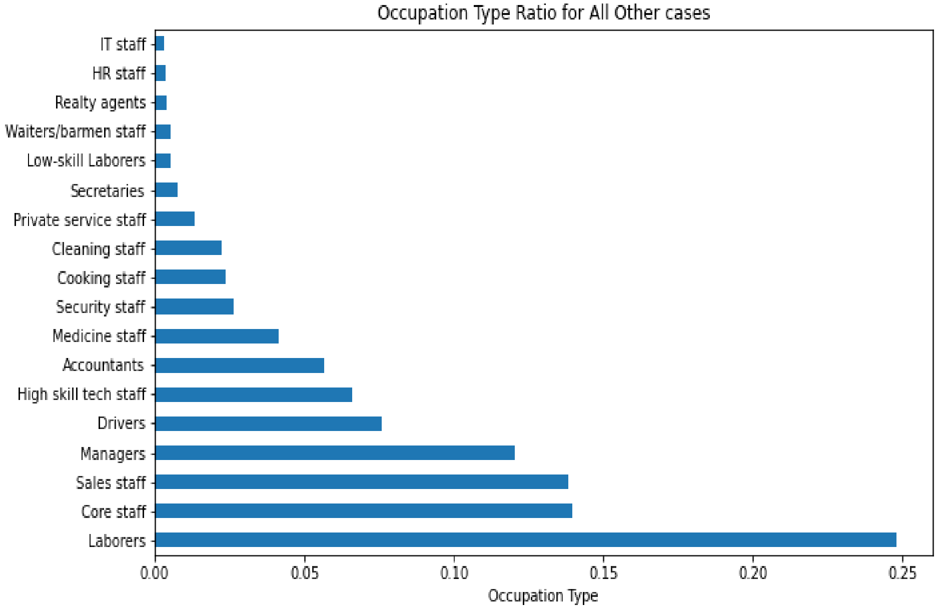




Inference – In the dataset provided, number of Married people are more. Though, Married people tend to have more payment difficulties when compared with Unmarried/Widow/Separated etc.

Occupation Type Ratio Plot

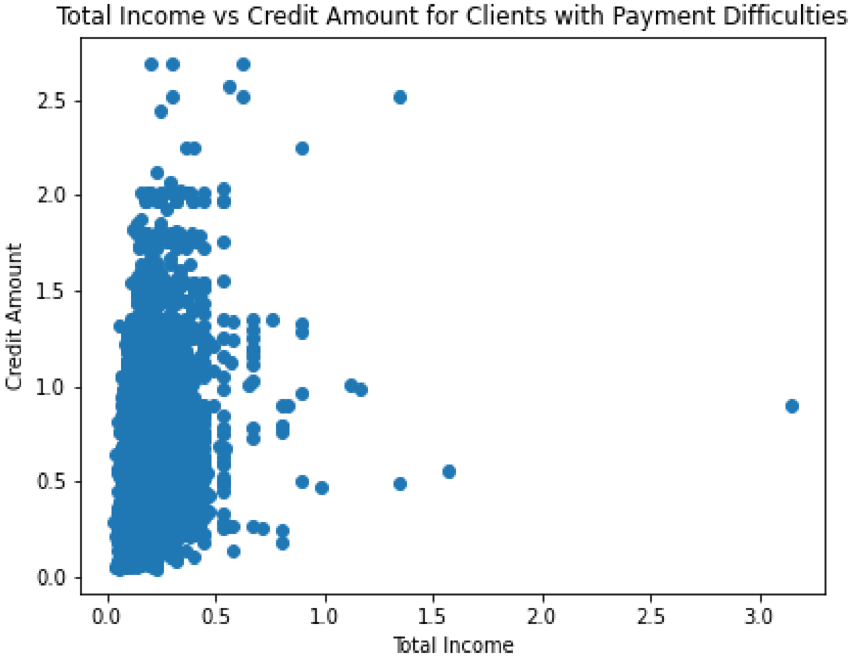


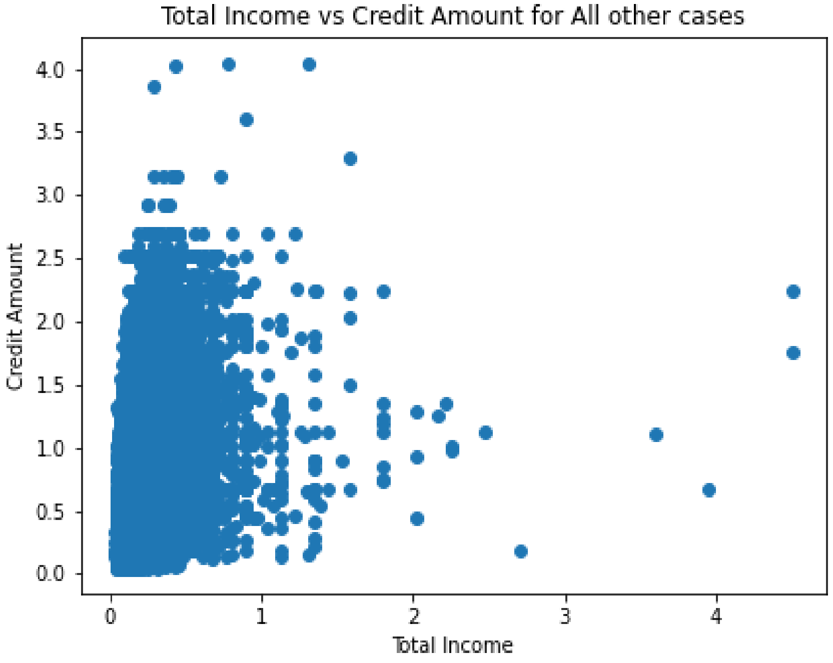


Inference – Laborers have more payment difficulties when compared with other Occupation Types.

Bivariate Analysis

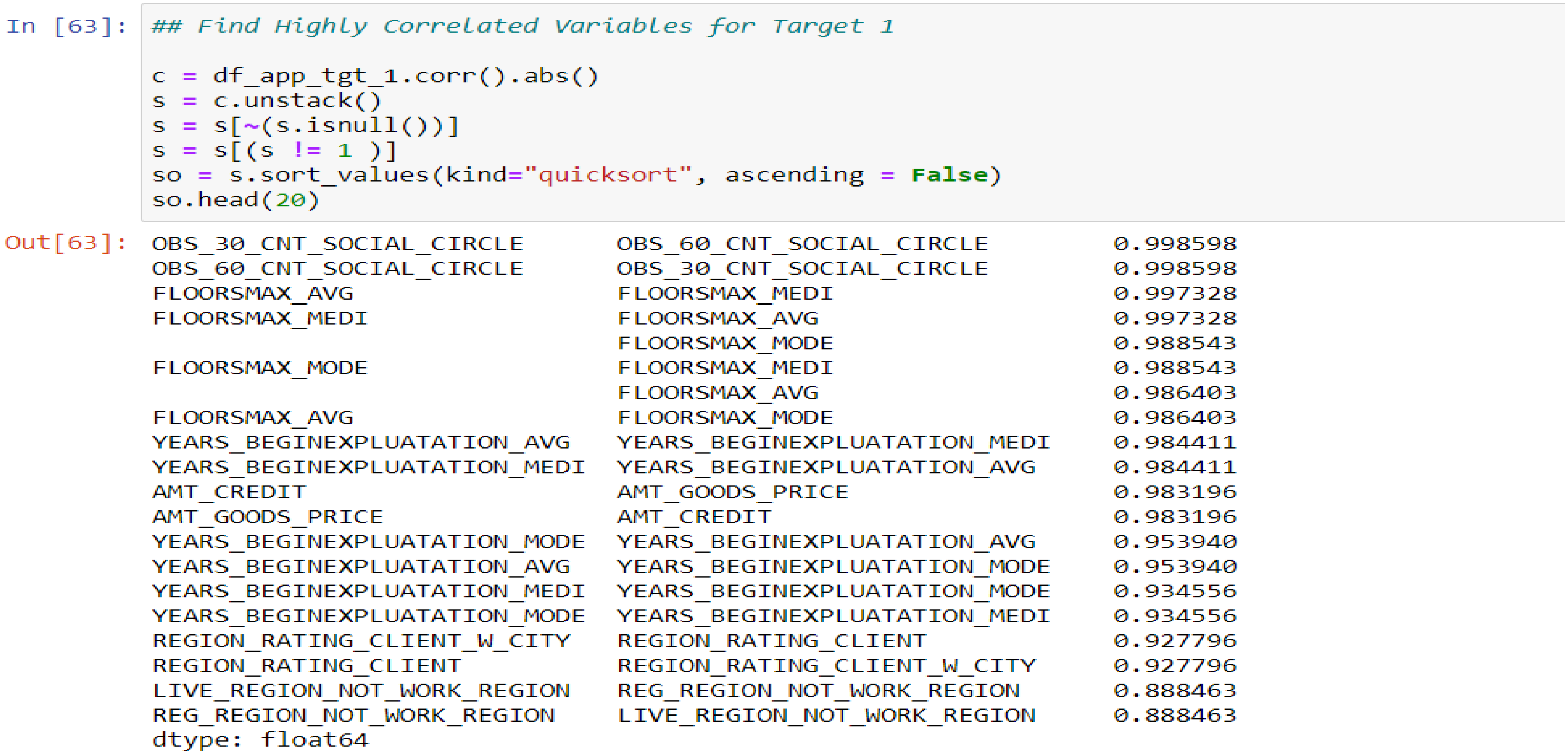
Total Income vs Credit Amount


Total Income vs Credit Amount for Clients with Payment Difficulties


Total Income vs Credit Amount for All other cases

Inference – Credit Amount did not show high correlation with total income in either of the categories.

Correlated Variables for TARGET - 1



In [63]:

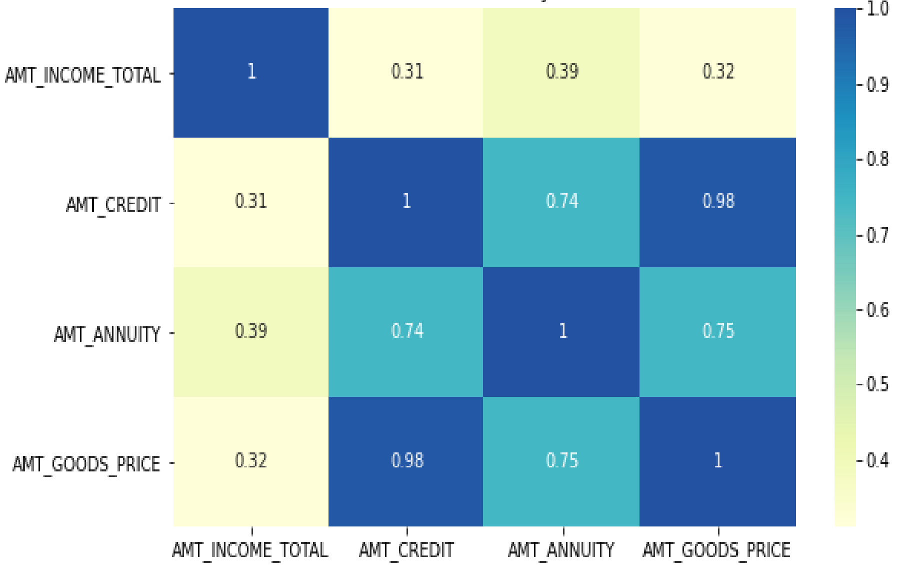
```
## Find Highly Correlated Variables for Target 1

c = df_app_tgt_1.corr().abs()
s = c.unstack()
s = s[~(s.isnull())]
s = s[(s != 1 )]
so = s.sort_values(kind="quicksort", ascending = False)
so.head(20)
```

Out[63]:

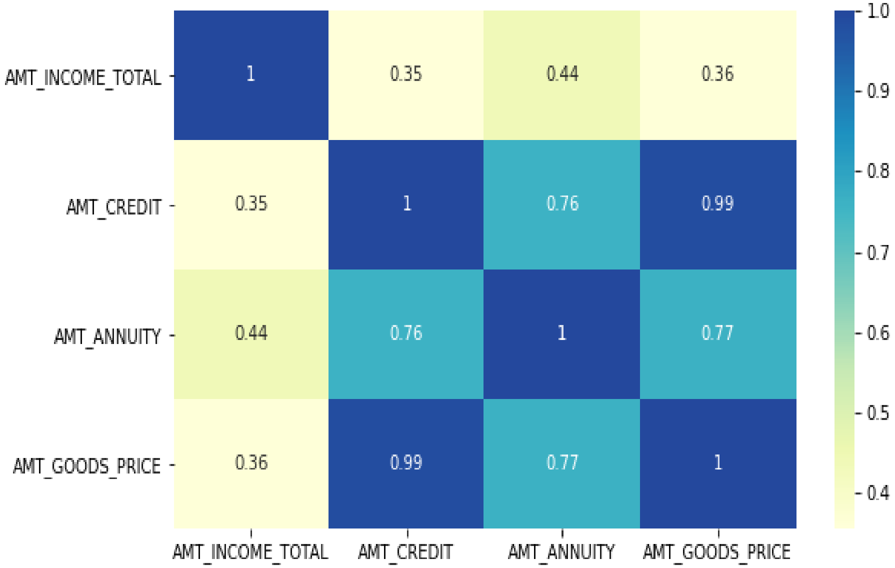
```
OBS_30_CNT_SOCIAL_CIRCLE        OBS_60_CNT_SOCIAL_CIRCLE        0.998598
OBS_60_CNT_SOCIAL_CIRCLE        OBS_30_CNT_SOCIAL_CIRCLE        0.998598
FLOORSMAX_AVG                   FLOORSMAX_MEDI                  0.997328
FLOORSMAX_MEDI                  FLOORSMAX_AVG                   0.997328
                                FLOORSMAX_MODE                  0.988543
FLOORSMAX_MODE                  FLOORSMAX_MEDI                  0.988543
                                FLOORSMAX_AVG                   0.986403
FLOORSMAX_AVG                   FLOORSMAX_MODE                  0.986403
YEARS_BEGINEXPLUATATION_AVG     YEARS_BEGINEXPLUATATION_MEDI    0.984411
YEARS_BEGINEXPLUATATION_MEDI    YEARS_BEGINEXPLUATATION_AVG     0.984411
AMT_CREDIT                      AMT_GOODS_PRICE                 0.983196
AMT_GOODS_PRICE                 AMT_CREDIT                      0.983196
YEARS_BEGINEXPLUATATION_MODE    YEARS_BEGINEXPLUATATION_AVG     0.953940
YEARS_BEGINEXPLUATATION_AVG     YEARS_BEGINEXPLUATATION_MODE    0.953940
YEARS_BEGINEXPLUATATION_MEDI    YEARS_BEGINEXPLUATATION_MODE    0.934556
YEARS_BEGINEXPLUATATION_MODE    YEARS_BEGINEXPLUATATION_MEDI    0.934556
REGION_RATING_CLIENT_W_CITY     REGION_RATING_CLIENT            0.927796
REGION_RATING_CLIENT            REGION_RATING_CLIENT_W_CITY     0.927796
LIVE_REGION_NOT_WORK_REGION     REG_REGION_NOT_WORK_REGION      0.888463
REG_REGION_NOT_WORK_REGION      LIVE_REGION_NOT_WORK_REGION     0.888463
dtype: float64
```

Correlation of Amount Variables

Correlation for Clients with Payment Difficulties

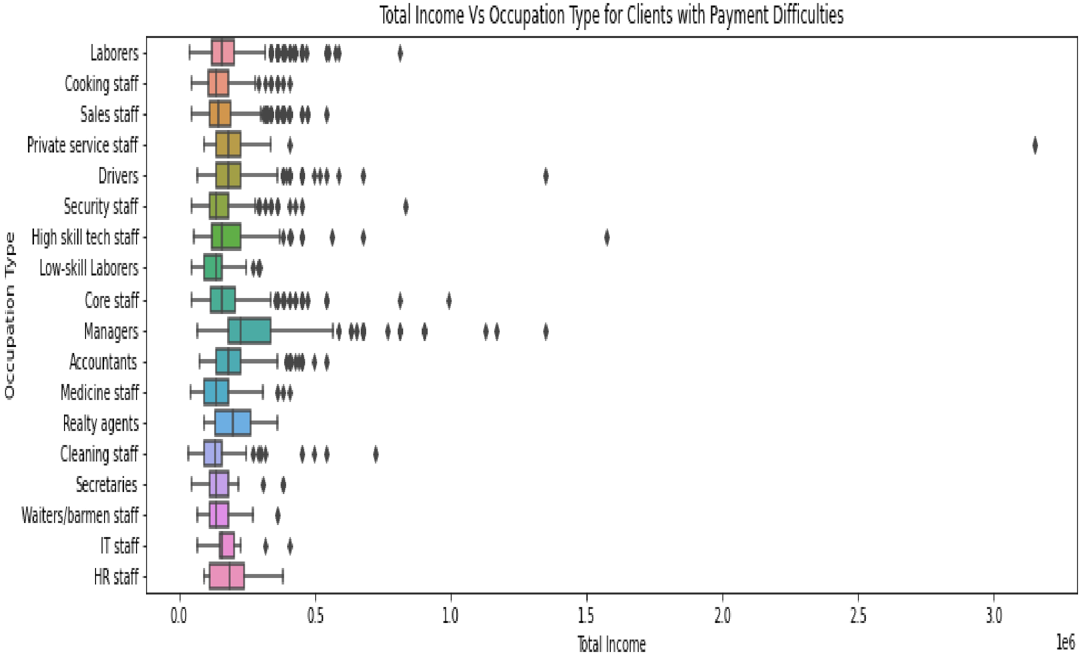


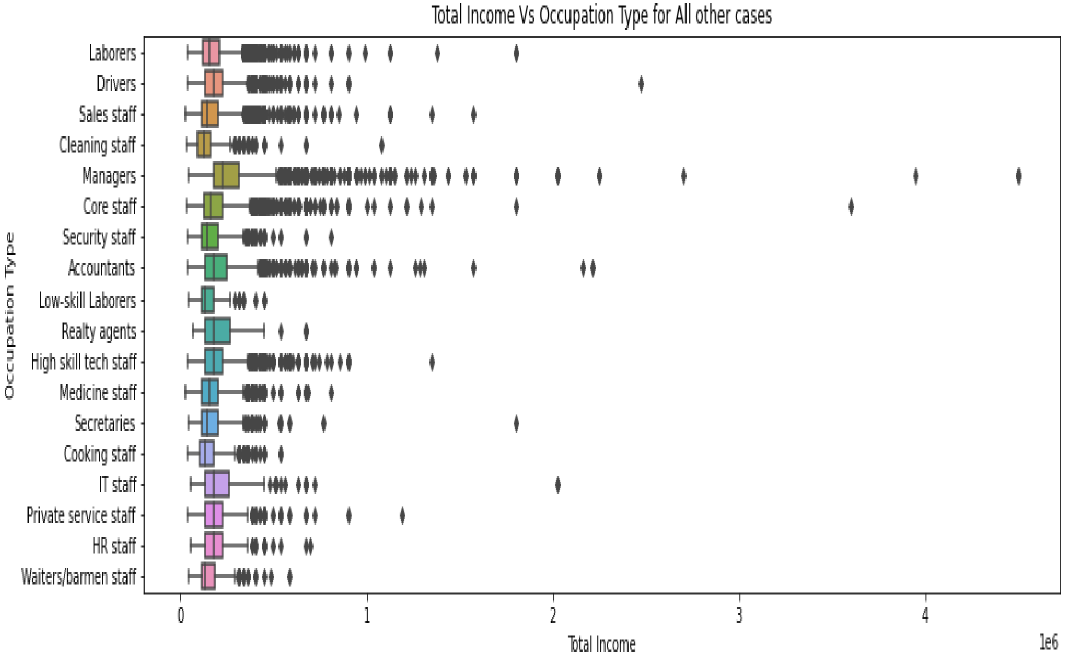
Correlation for All other cases



Inference – Credit Amount and Amt Goods Price are highly correlated while Total Income and Credit Amount have less correlation for both the categories.

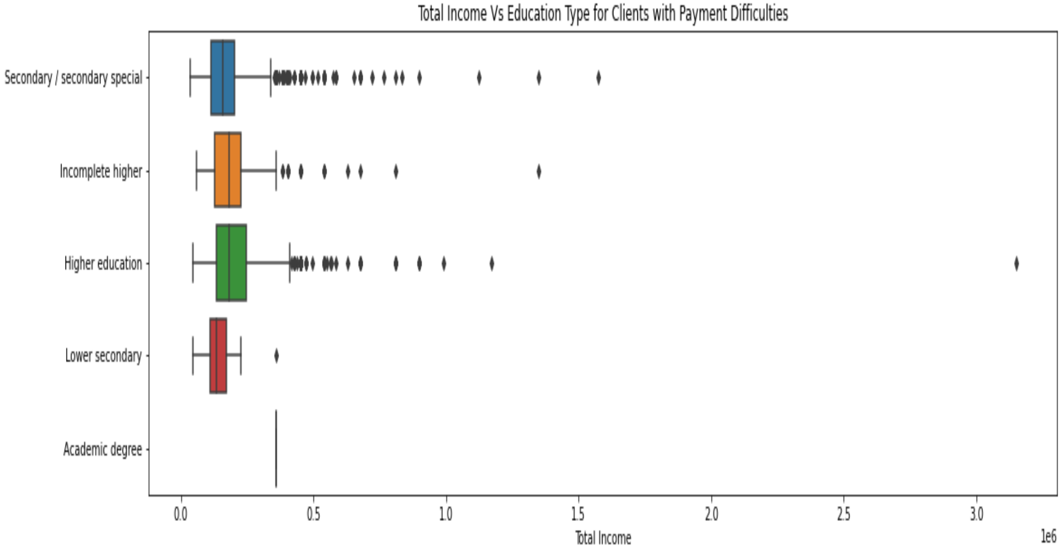
Total Income vs Occupation Type

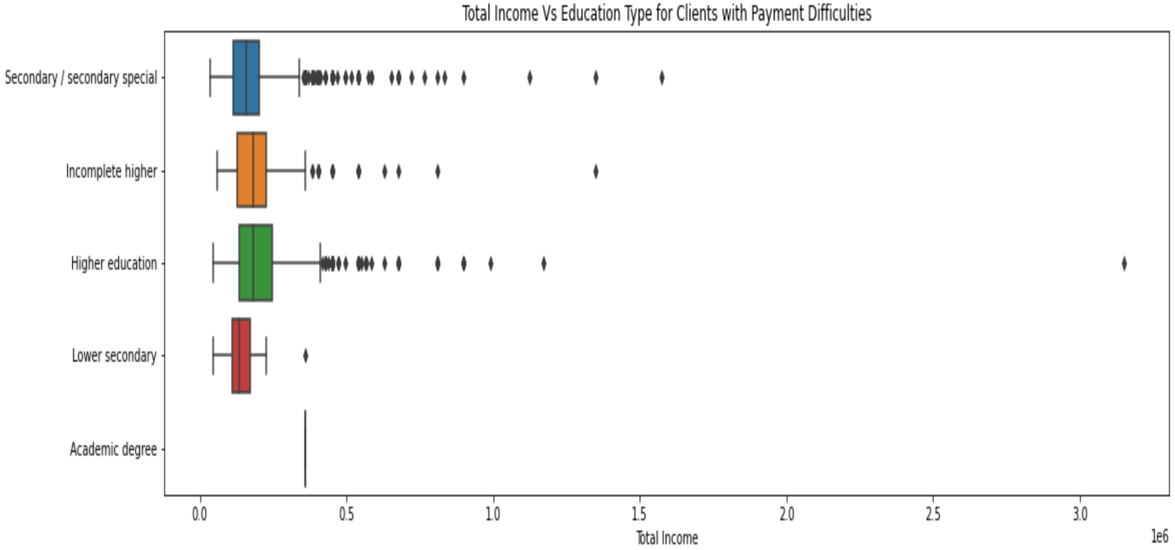




Inference – Managers have higher Income as compared to other Occupations for both the categories.

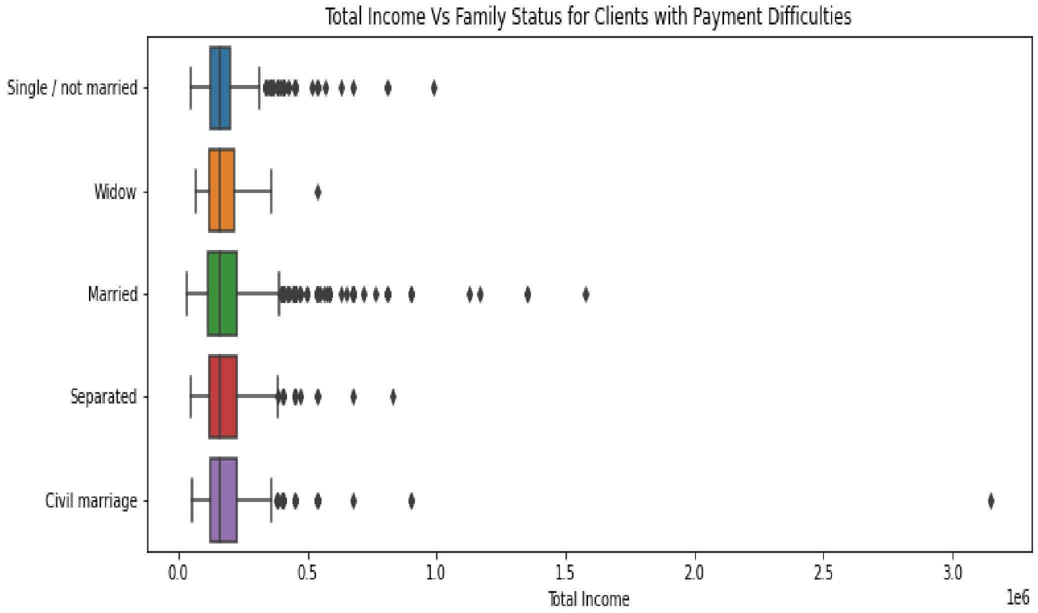
Total Income vs Education Type

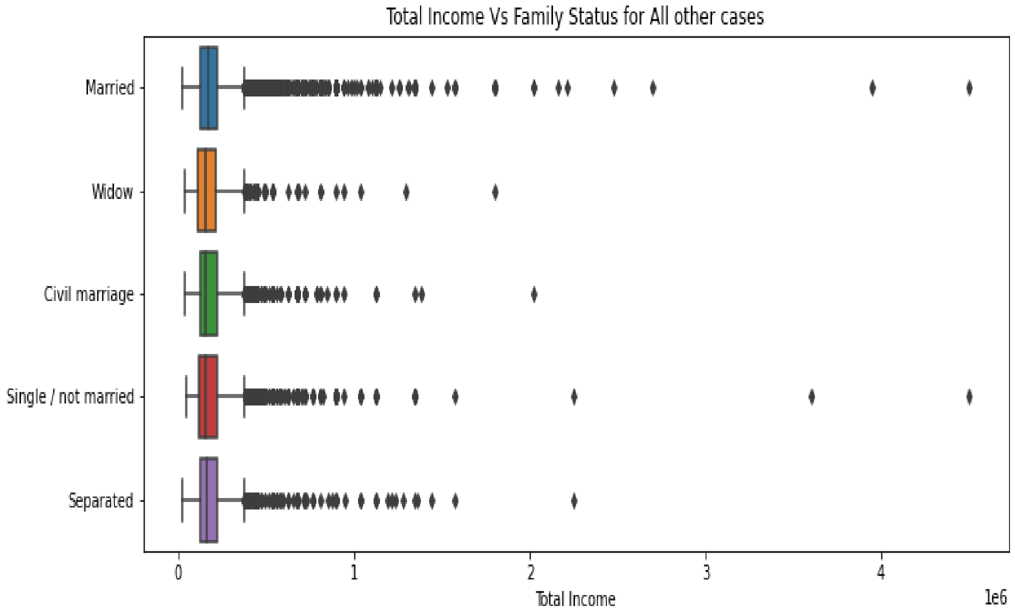




Inference – Higher Education People have higher Income as compared to others for both the categories.

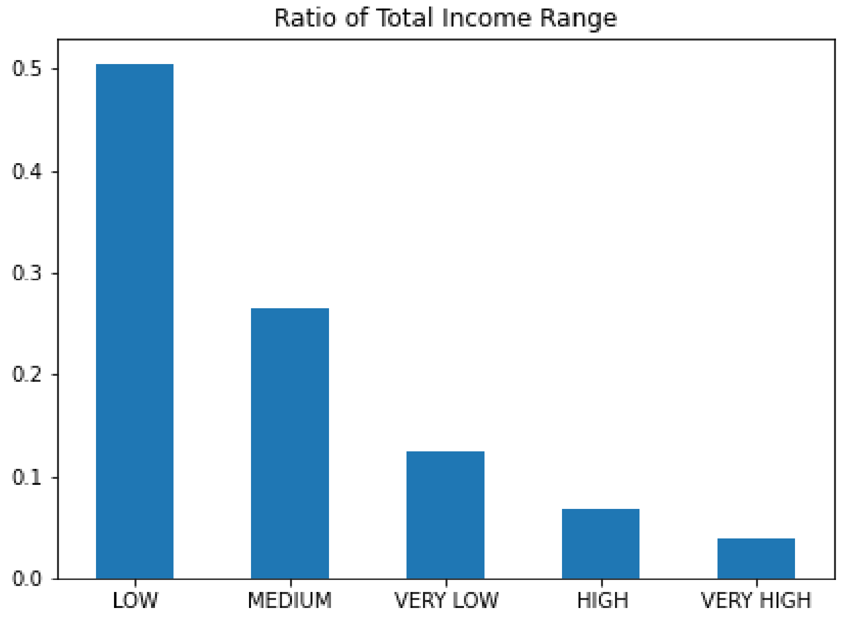
Total Income vs Family Status



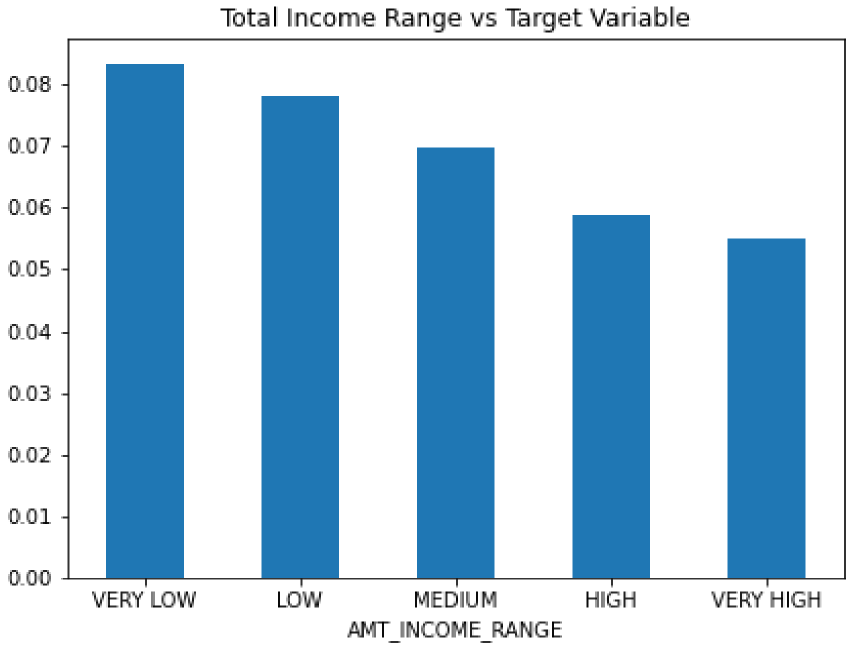


Inference – Married People have higher Income as compared to others for both the categories.

Total Income Range vs Target Variable



Ratio of Total Income Range



Total Income Range vs Target Variable

Inference – People falling in Very Low Income Range Category tend to have payment difficulties as compared to people in Very High Income Range Category. This graph clearly indicates the decrease in trend as the income rises/increases.

Conclusion

- Almost 50% of the columns were dropped from application dataset due to missing values/not relevant for analysis.
- Few columns were incorrectly defined in the dataset which were converted to its relevant type.
- Incorrect values removed from the dataset s/a XNA present in Gender Column.
- Total number of people with payment difficulties were very less as compared to other cases which signifies imbalanced data.
- Focus should be more on Student/Businessman Income Type as they have successful loan re-payments.
- Focus should be less on Secondary Educated People while more on Higher Educated People for successful loan re-payments.
- Focus should be more on People living with parents/Rented/Office Apartments for successful loan re-payments.
- Banks should focus less on Married People and more on Single/Separated etc for successful loan re-payments.
- Banks should focus less on Laborers while more on others like High skilled staff/managers etc for successful loan re-payments.
- People having Higher Income have less payment difficulties hence banks should target these people.