CREDIT EDA-CASE STUDY

PRESENTED BY- GANTAVYA BANGA

ROSHAN XAVIER

CASE STUDY OBJECTIVE

The objective behind the case study is to identify users on the basis of their data and segregate them into two different categories, one that includes the people who can pay back a loan with ease and others who have difficulties in paying back loans.

This segregation can help the bank decide on what customers need to be provided with the loans and what customers loan applications need to be rejected.

As a whole the analysis needs to be accurate to ensure that the customers who have been provided with the loan should not end up being defaulters and the customers who are not-defaulters or who can pay the loan amount should not be rejected.

Hence, the company is expecting to know the key metrics that lead to loan defaults, so that the bank can evaluate the data for assessing the risk and portfolio.

APPLICATION DATA DATASET

Contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**

TREATING MISSING VALUES

In the application_data file there are many columns with quite a lot of missing values. The processes that we have used to treat the missing values are defined below:

- Check for all the columns that have missing or null values.
- Drop the columns with more than 50% missing values.
- Recheck the dataset to confirm that the columns have been dropped.
- Check for the columns with missing values equal to 13%.
- Since, the columns with missing values equal to 13% have many outliers we will be imputing them with values.
- Here, we make use of the median values as the mean values are affected by the outliers in the data.
- Hence, not all the missing values can be imputed by using one method.

CHANGING THE DATA TYPES

```
SK_ID_CURR                        int64
TARGET                            int64
NAME_CONTRACT_TYPE               object
CODE_GENDER                      object
FLAG_OWN_CAR                     object
FLAG_OWN_REALTY                  object
CNT_CHILDREN                      int64
AMT_INCOME_TOTAL                float64
AMT_CREDIT                      float64
AMT_ANNUITY                     float64
AMT_GOODS_PRICE                 float64
NAME_TYPE_SUITE                  object
NAME_INCOME_TYPE                 object
NAME_EDUCATION_TYPE              object
NAME_FAMILY_STATUS               object
NAME_HOUSING_TYPE                object
DAYS_BIRTH                        int64
DAYS_EMPLOYED                     int64
OCCUPATION_TYPE                  object
CNT_FAM_MEMBERS                 float64
WEEKDAY_APPR_PROCESS_START       object
HOUR_APPR_PROCESS_START           int64
ORGANIZATION_TYPE                object
OBS_60_CNT_SOCIAL_CIRCLE        float64
DEF_60_CNT_SOCIAL_CIRCLE        float64
AMT_REQ_CREDIT_BUREAU_QRT       float64
dtype: object
```

- Here, in the application_data dataset can observe that "FLAG_OWN_CAR" & "FLAG_OWN_REALTY" are defined as objects, even though the data stored is 0 and 1. Hence, we need to rectify the data type to “int” for further operations.
- We can also see that "SK_ID_CURR" is stored as “int”. Since we won’t be using "SK_ID_CURR" in our numerical analysis , we can change it to “string”.
- Some categorical columns like "NAME_CONTRACT_TYPE', "CODE_GENDER", "NAME_EDUCATION_TYPE”,"NAME_HOUSING_TYPE" can be changed to category dtype for better insights.

HANDLING NEGATIVE VALUES IN THE DATAFRAME

```
0          -9461
1         -16765
2         -19046
3         -19005
4         -19932
           ...
307506     -9327
307507    -20775
307508    -14966
307509    -11961
307510    -16856
```

We can notice that the DAYS_BIRTH column defines the days past since a person was born, this data is not just confusing but can also lead to multiple errors during computations.

Hence, we will be converting the DAYS_BIRTH column to Age by dividing it by 365 and finding its absolute value to remove the negative sign.

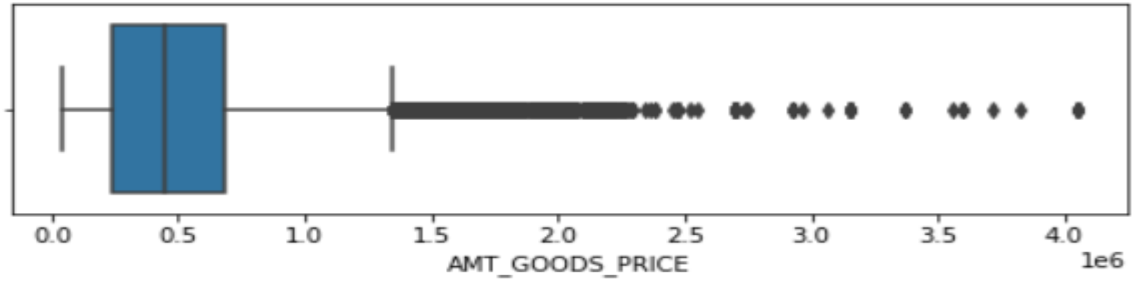
```
0            -637
1           -1188
2            -225
3           -3039
4           -3038
             ...
307506       -236
307507     365243
307508      -7921
307509      -4786
307510      -1262
```

We can notice that the DAYS_EMPLOYED column defines the number of days an employee has been employed for, but again the column does not define the data in a correct format.

Hence, we will be converting the DAYS_EMPLOYED column to Work_experience by dividing it by 365, in order to define the experience of any given user.

HANDLING OUTLIERS-EXAMPLE

Handling outliers in the AMT_GOODS_PRICE



Through the help of Boxplot, we can clearly identify the outliers present in the data. There are quite a lot of outliers in the AMT_GOODS_PRICE column with a missing percentage . Hence, we will use the quantile range method to remove the outliers.

count	3.036560e+05
mean	5.217324e+05
std	3.369093e+05
min	4.050000e+04
25%	2.385000e+05
50%	4.500000e+05
75%	6.795000e+05
max	1.795500e+06

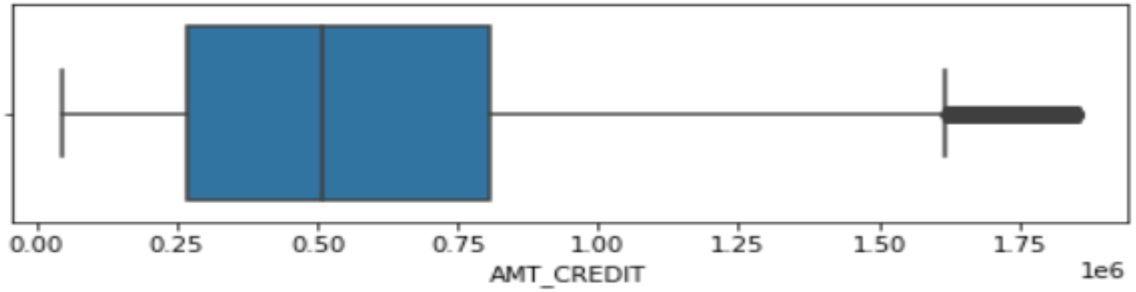
Outliers can similarly be removed by using different processes depending on the outlier.

HANDLING OUTLIERS-EXAMPLE

Handling outliers in AMT_CREDIT columns

Before Handling

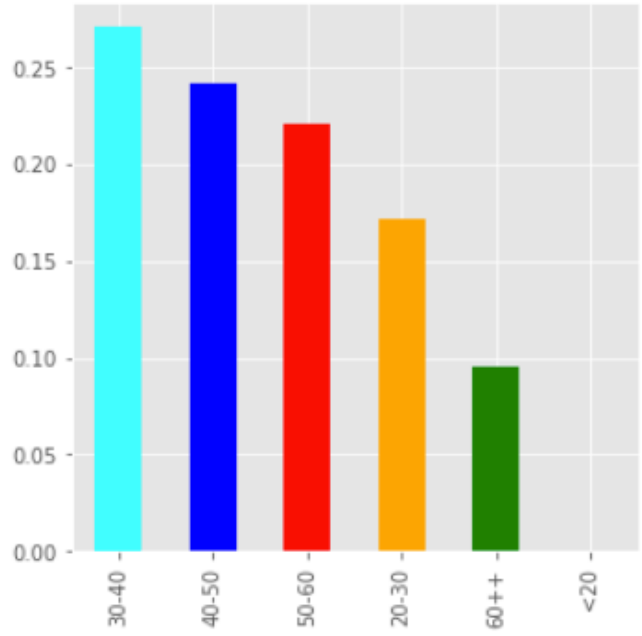
After Handling

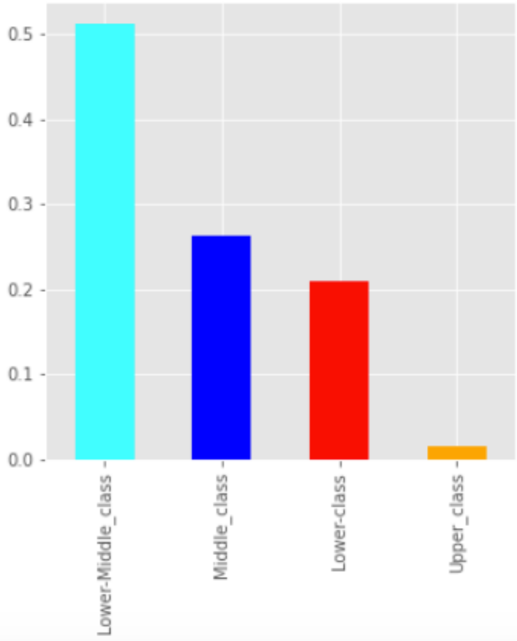


As we can see that the outliers have been treated. Their are still some outliers outside the quartile range. But hey are in the continuous straight line so its not required to remove them. From the boxplot we can see that most of the data lies within 270000 to 800000.

The min value is around 45000 while the max value is approximately 1852000

BINNING CATEGORICAL COLUMNS



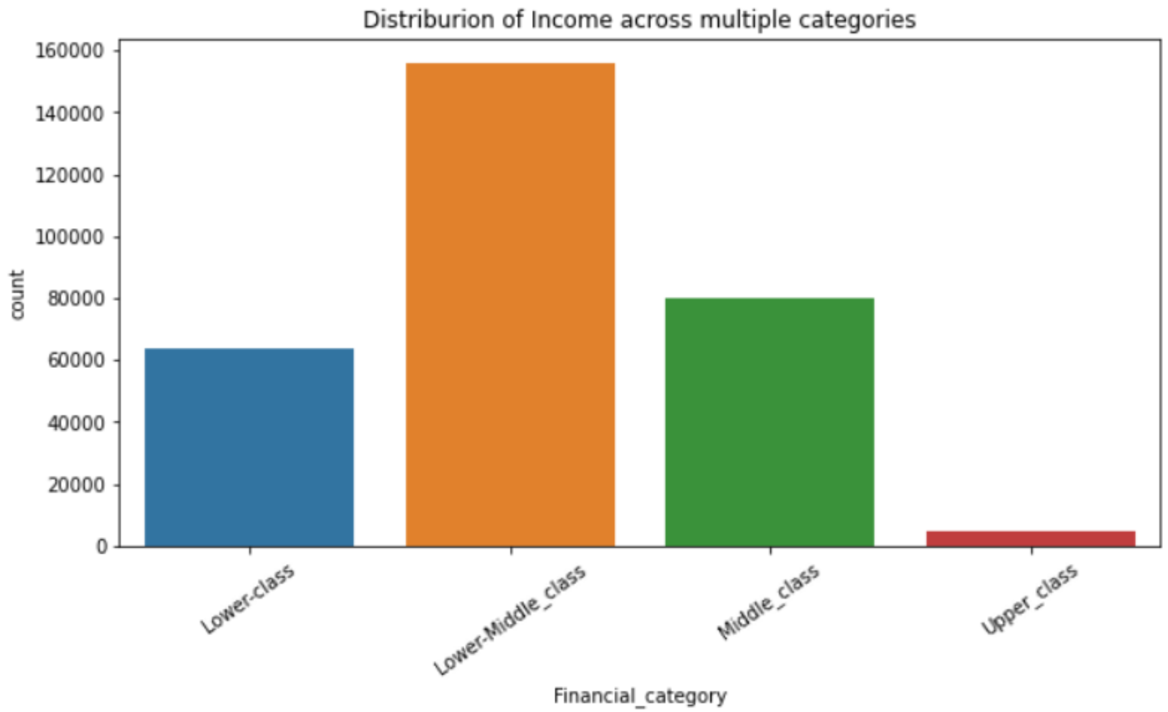


- We divided the Age column into 6 categories: >20, 20-30, 30-40, 40-50, 50-60, 60++

- We divided the Financial category column into 4 categories: Lower Class, Lower Middle Claps, Middle Class and Upper Class
- LC contains income less than 30000
- LMC contains value between 30k and 1 lakh
- Middle class contains Income between 1-2 Lakh
- Upper Class contain values between 2-4 Lakhs

UNIVARIATE ANALYSIS

Univariate Analysis for Distribution of Income across multiple categories

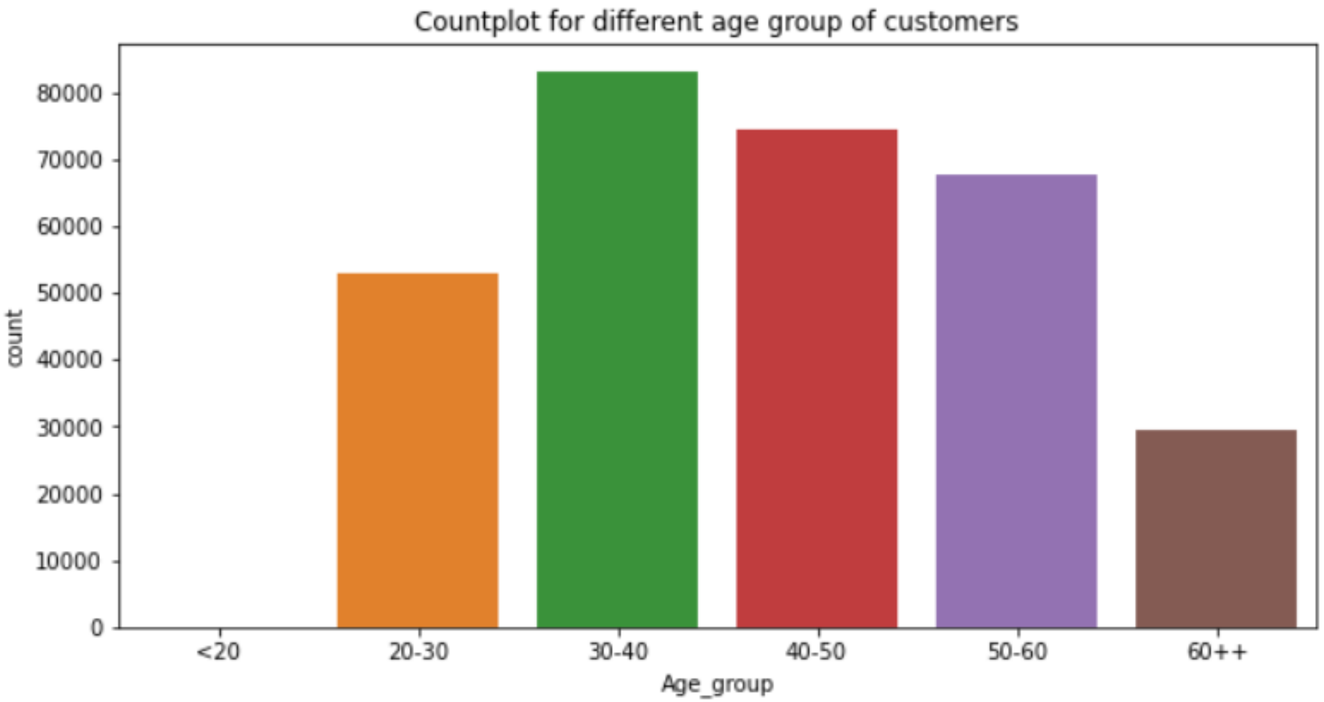


There are a huge majority of people who belong to the Lower-Middle_class category.

- There are very few, almost 15 times lesser number of people who belong to the Upper_class category.

UNIVARIATE ANALYSIS

Univariate Analysis – Countplot for different age group of customers



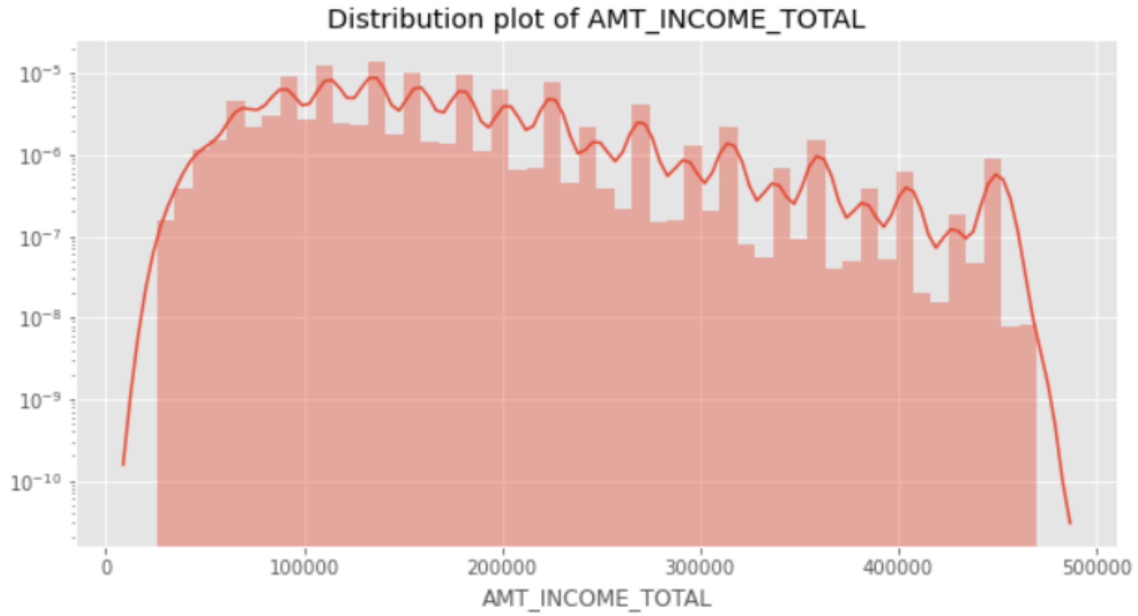
A wide majority of the customers belong to the age group of 30-40

Most people in the age group of 40-50 and 50-60 are active customers.

There are just 30% people above the age group of 60 in comparison to the people in the age group of 30-40

CONTINOUS UNIVARIATE ANALYSIS

Distribution plot of AMT_INCOME_TOTAL

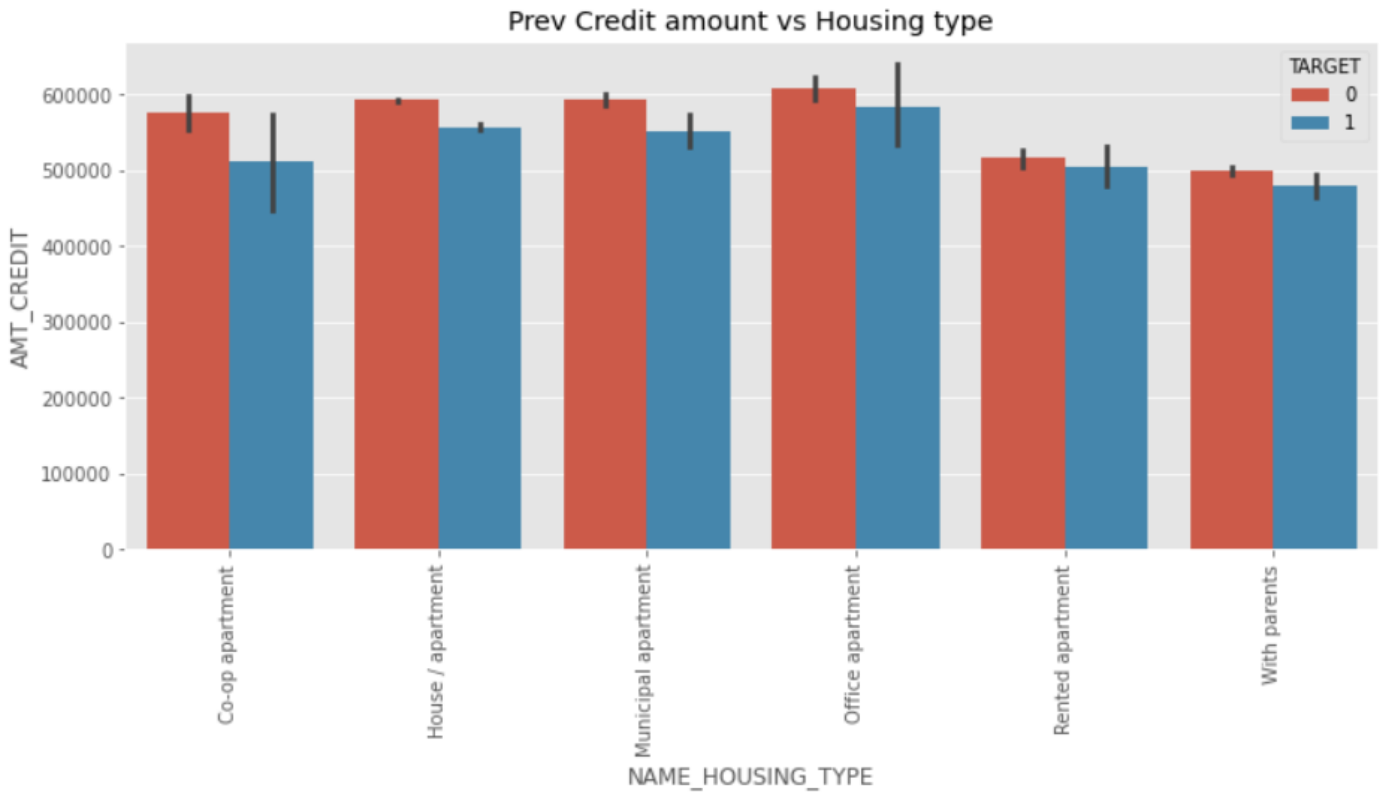


Majority of the population lied between 112500 and 202500

The mean of total population is around 160000 while median is 144000

Min value is 25650 and max value is 469800

BIVARIATE ANALYSIS

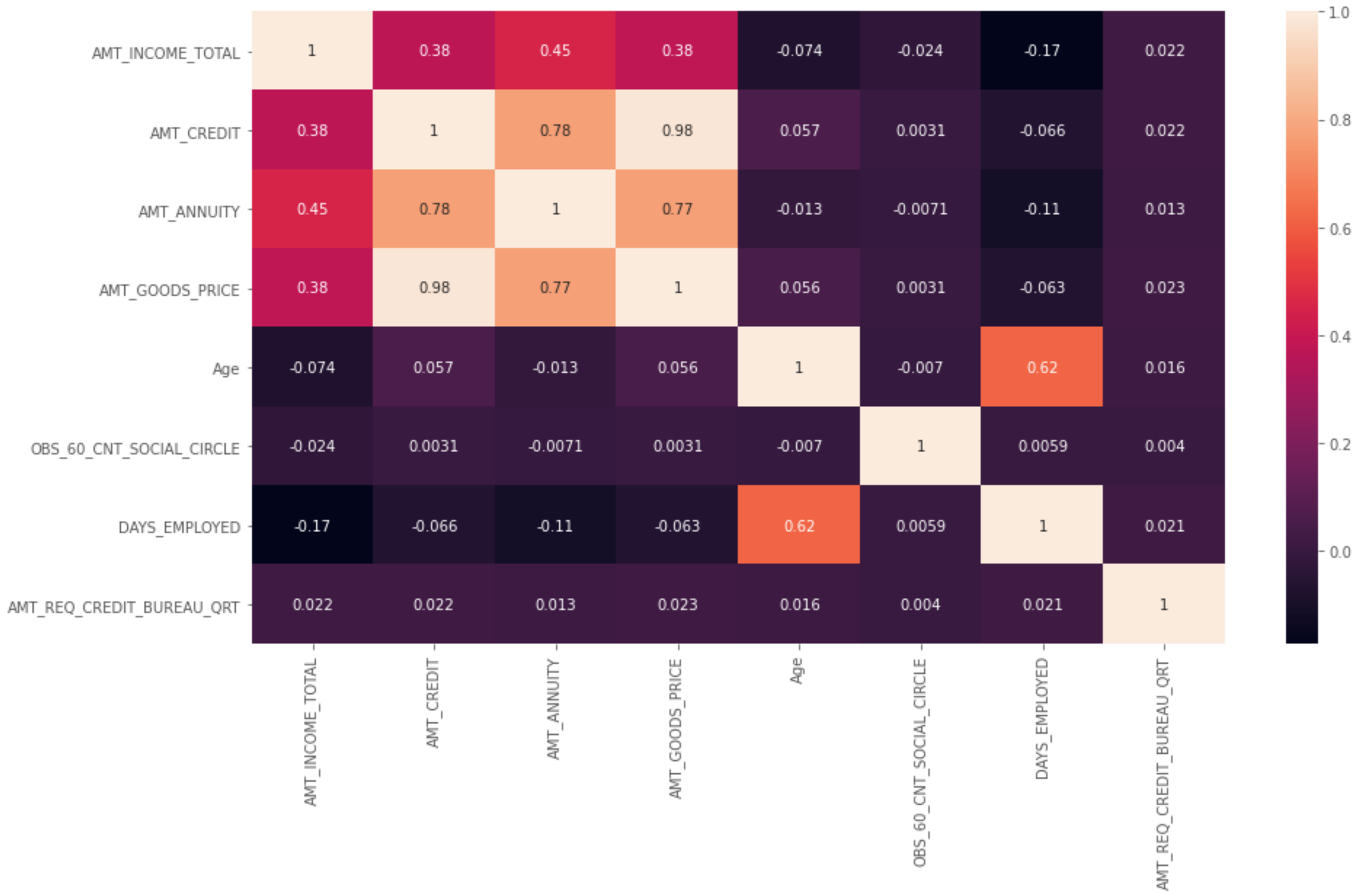


The people living in office apartments tend to have the highest credit amount in both Target0 and Target1, i.e. for people with payment difficulties and for other people.

Amongst all of the applicants, regardless of the housing type, there is always a higher number of people who have difficulties in making the repayments.

The customers who are living with their parents tend to have the least credit amount.

Correlation Matrix



Positive-Correlations

1)AMT_CREDIT and AMT_GOODS_PRICE

2)AMT_CREDIT and AMT_ANNUITY

3)AMT_CREDIT and AMT_INCOME_TOTAL

4)AMT_ANNUITY and AMT_GOODS_PRICE

5)AMT_ANNUITY and AMT_INCOME_TOTAL

Negative-Correlations

1) AMT_CREDIT and DAYS_EMPLOYES

2) AMT_GOODS_PRICE and DAYS_EMPLOYES

3) AMT_ AMMUNITY and Age

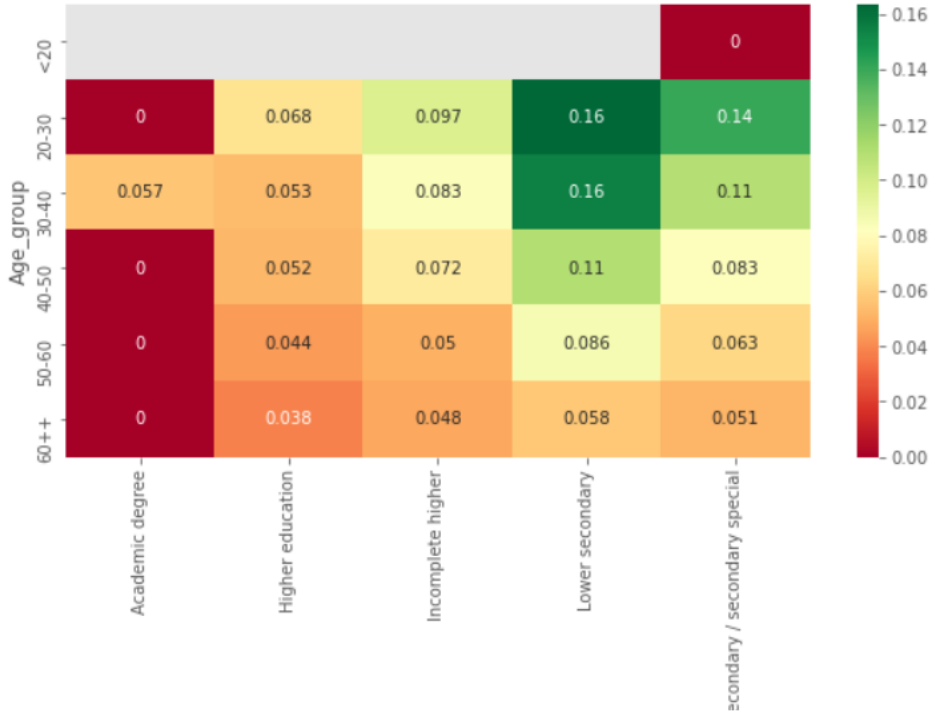
4) AMT_INCOME TOTAL and DAYS EMPLOYES

5) AMT_AMMUNITY and DAYS_EMPLOYES

The goods price and the credit amount have a very even correlation of 0.99.

The AMT_ANNUITY is highly correlated with AMT_CREDIT and AMT_GOODS_PRICE.

Correlation between Age Group and Education qualifications



The correlation of people in the age group of 20-30 and 30-40 with an education of lower secondary is much higher than the rest which shows that its very likely to get defaulted

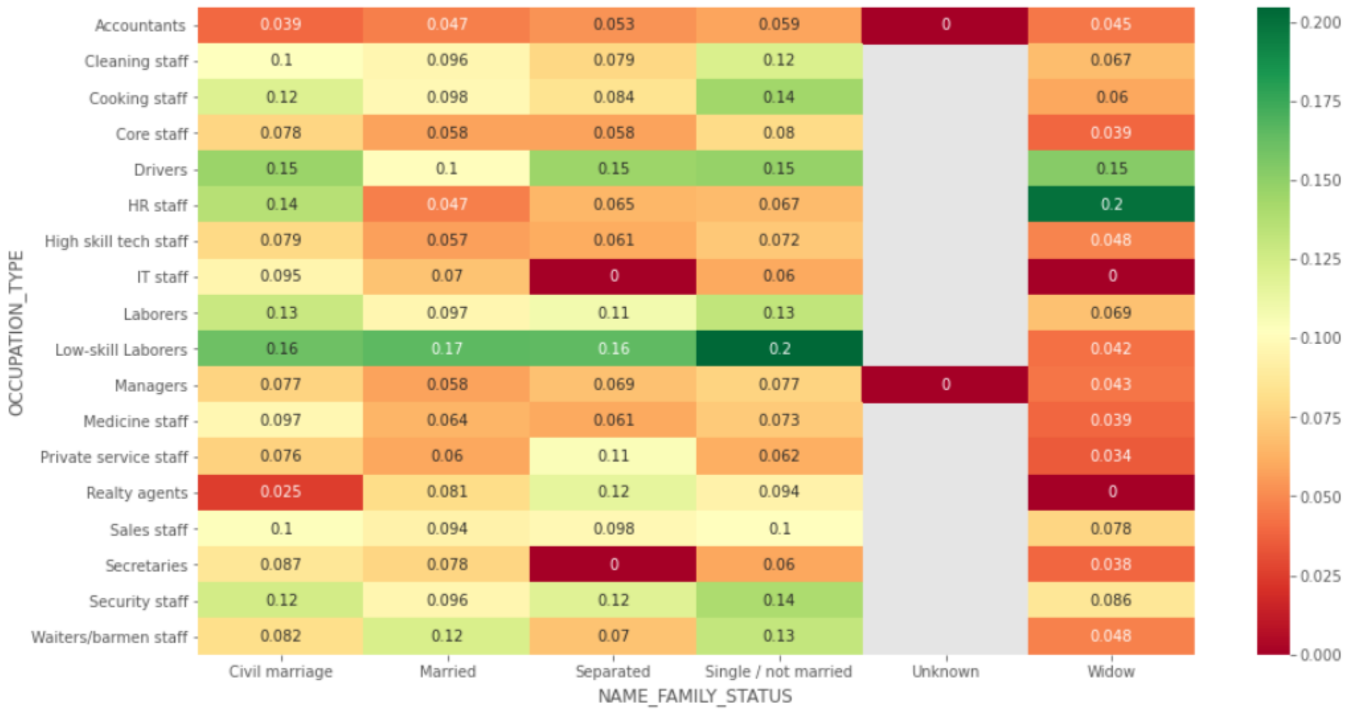
Except for the age group of 30-40, there are few people who have an academic degree.

There is an even distribution of people who have completed their secondary level education.

Lower secondary education with age group 20-30 & 30-40 have correlation of 0.16 are most likely to get defaulted also 40-50 age group with lower secondary education have correlation of 0.11

- we can also observe that population with Academic degree are least likely to get defaulted with all the age groups have a correlation of 0 except 30-40 age group which have correlation of 30-40

Correlation between Family Status and Occupation Type



Most of the customers who are either single or widow are highly correlated to low-skilled labour and HR respectively.

Most of the single customers have the highest correlation with different occupation types.

PREVIOUS DATA DATASET

Contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

CHANGING THE DATA TYPES

```
<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                  1285287 non-null  float64
 4   AMT_APPLICATION              1645828 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  int64  
 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 
 24  NAME_PRODUCT_TYPE            1670214 non-null  object 
 25  CHANNEL_TYPE                 1670214 non-null  object 
 26  SELLERPLACE_AREA             1670214 non-null  int64  
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object 
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object 
 30  PRODUCT_COMBINATION          1669868 non-null  object 
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(7), object(15)
memory usage: 471.5+ MB
```

- We are analyzing the previous data so here we can change the 'SK_ID_CURR' from integer to string
- FLAG_LAST_APPL_PER_CONTRACT needs to changed from object on int value for analysis
- Also, we can change the 'WEEKDAY_APPR_PROCESS_START', 'NAME_CONTRACT_STATUS', 'NAME_PAYMENT_TYPE','CODE_REJECT_REASON, 'NAME_CLIENT_TYPE‘ from object to categorical data.

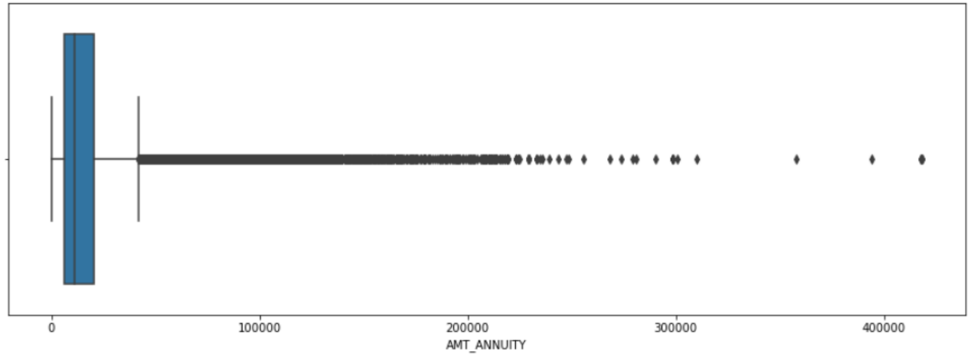
HOW DO WE FIND OUTLIERS HERE?

- First of all we make use of boxplots to see if we can spot the outliers.
- Secondly, we have used the quantile method to identify the values of outliers
- In most cases anything above 99% quantile is an outliers and we remove them for the better analysis of our data

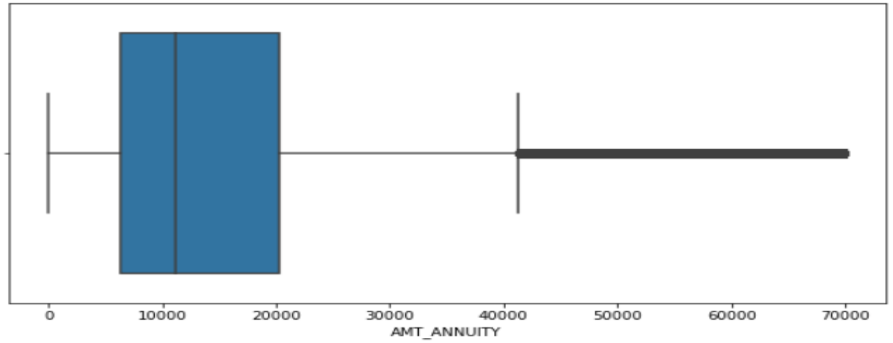
HANDLING OUTLIERS

Handling outliers in AMT_ANNUITY

Before Handling



After Handling



As we can see that the outliers have been treated but their are still some continuous outliers making a line outside the quartile range.

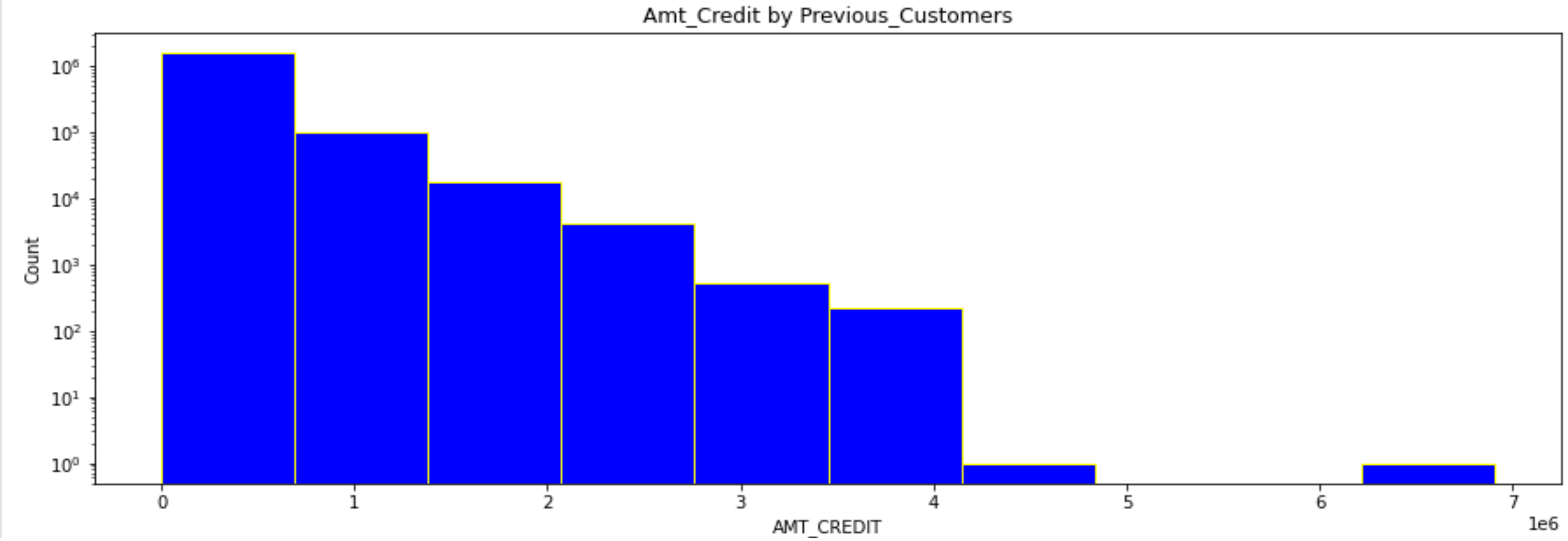
Its not required to removed them as its useful for data analysis .

From the boxplot we can see that most of the data lies within 6000 to 20000

The mix value is 0 while the max value is approximately 70000

UNIVARIATE ANALYSIS

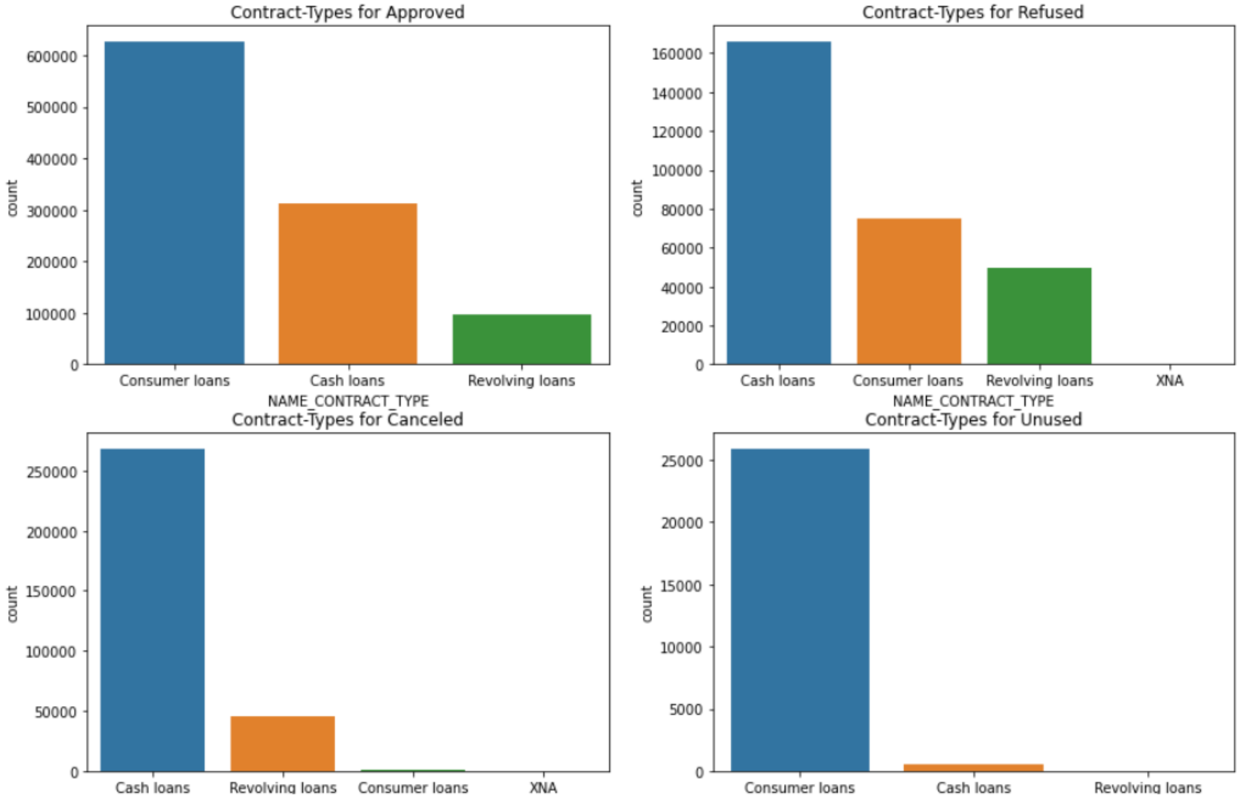
AMT_CREDIT by Previous_Customers



As we can see from the bar plot that as the AMT_CREDIT is low, the frequency count is at the highest. the higher the AMT_CREDIT, the lower the Frequency count .

From the bar plot we can also observe that the max value is around 1500000 and the min value is 0.

UNIVARIATE ANALYSIS



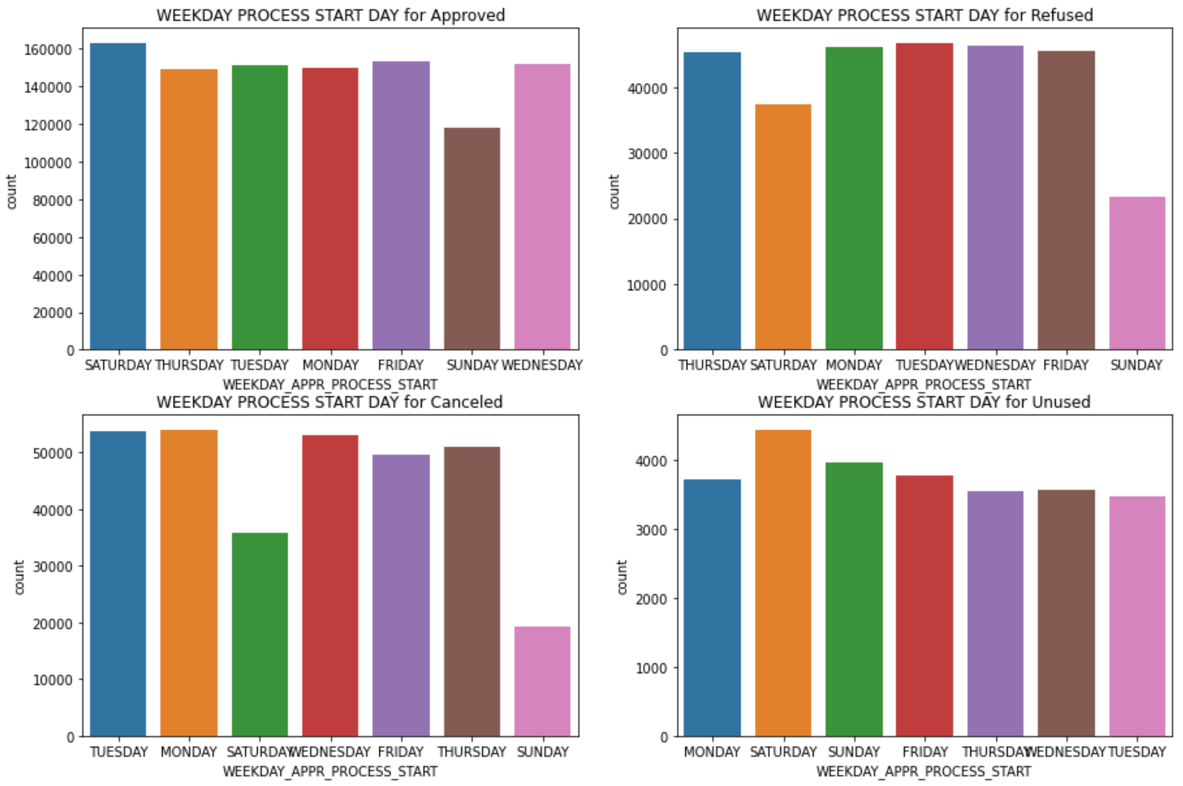
As we can see from the above plots that 60% of consumers loans and 30% of cash loans gets approved.

Also, 57% of cash loans and 25% of consumer loans and 17% of revolving loans gets refused.

Cash Loans gets canceled the most at around 85%.

And finally consumers loan are the most unused by the customers

UNIVARIATE ANALYSIS



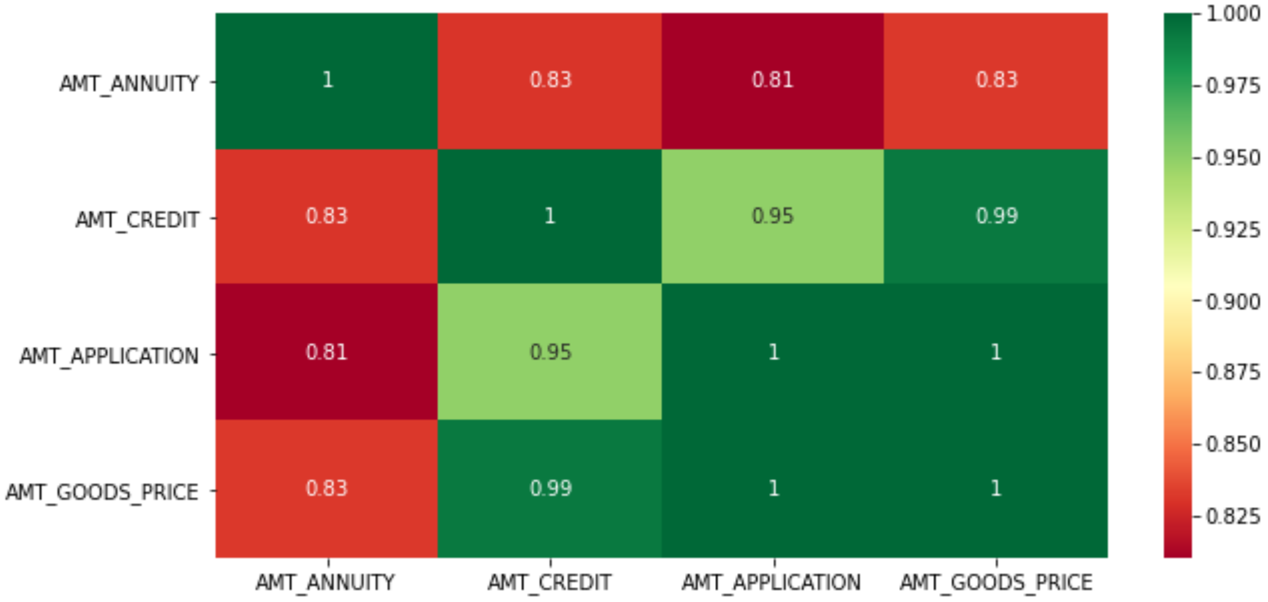
As we can see from the above plots that most of the loans applied on Saturday gets approved.

Around 16% of the loans applied on Tuesdays gets refused.

17% of loans applied on Mondays gets canceled.

And finally 16% of loans applied on Saturday went unused

BIVARIATE ANALYSIS

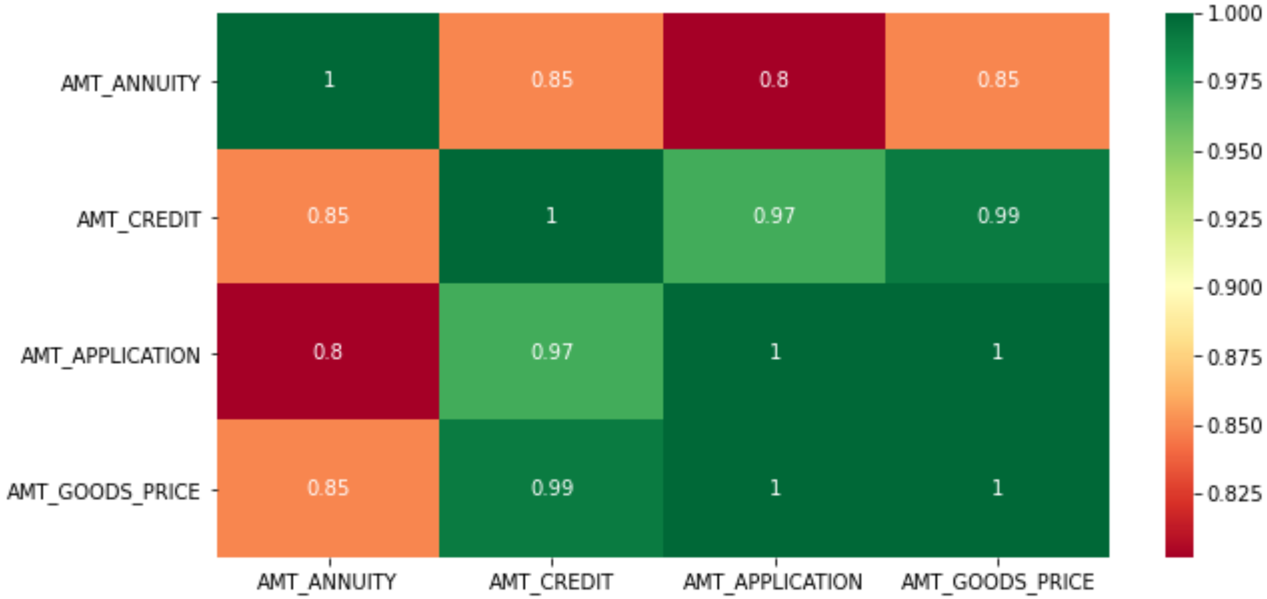


We can observe that all the columns are correlating pretty well.

AMT_APPLICATION & AMT_GOODS_PRICE got a perfect correlation of 1

AMT_CREDIT & AMT_Gppds_price are correlating at 0.99

BIVARIATE ANALYSIS



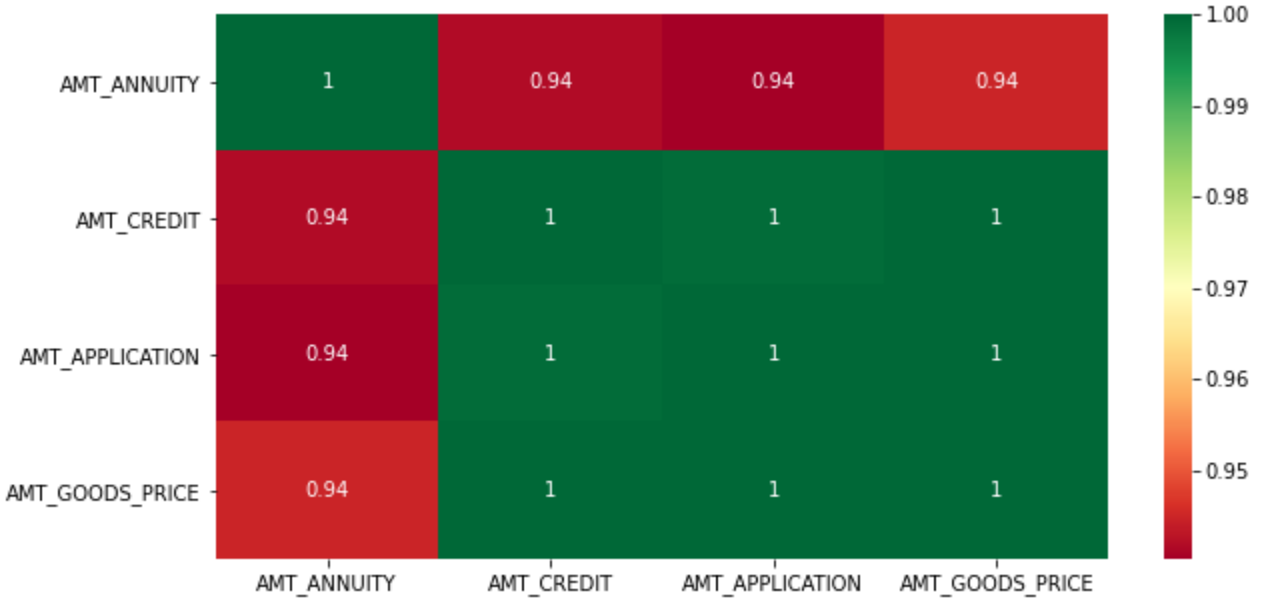
We can observe that all the columns are highly correlating like we seen before.

AMT_APPLICATION & AMT_GOODS_PRICE got a perfect correlation of 1

 AMT_CREDIT & AMT_GOODS_PRICE are correlating at 0.99

 AMT_APPLICATION & AMT_CREDIT are correlating at 0.96

BIVARIATE ANALYSIS



We can observe that all the columns are highly correlating

AMT_APPLICATION & AMT_GOODS_PRICE got a perfect correlation of 1

AMT_CREDIT & AMT_GOODS_PRICE is correlating at 1

AMT_APPLICATION & AMT_CREDIT is correlating at 0.97

AMT_ANUUITY & AMT_GOODS_RICE is correlating at 0.94

FINAL INSIGHTS

- 60% of the consumer loans and 30% of the cash loans tend to get approved while 57% of cash loans and 25% of consumer loans get refused.
- Cash Loans are more likely to get Default than Revolving loans.
- The analysis also suggests that Males are more likely to Default loans than Females
- Customers with Age group 30_40 are most likely to default loans and customers over 60++ are least likely to default loans
- Education qualification is highly correlated to the default percentage. We found that customers with secondary education default the most while compared to customers with Academic Education.
- Customers who are married are most likely to default than customers who are widow who are just 5% likely to default their loan
- People living with their parents have the most difficulty in paying back loans.
- People living in office apartments have the least difficulty in paying back loans.

Thank You