



Credit EDA – Case Study

PRESENTED BY: LOVE AGARWAL &
KRISHNA AGRAWAL

Analysis On Application Data File

Data Understanding

1. File has 307511 rows with 122 columns
2. After observing the data we have come to the following conclusions :
 - a) Column SK_ID_CURR is of type int which does not make sense. Although it contains only integer values but we cannot perform any numerical operation on them thus we should convert its data type to string.
 - b) There are some columns like CODE_GENDER, NAME_CONTRACT_TYPE, NAME_EDUCATION_TYPE which are of type object. Such columns are best described as categorical columns rather than string.
 - c) The type of most of the Flag variables is int which is good because then we can use them for calculations more easily we can classify them as categorical but we think its best to leave them as int.
 - d) Columns like WEEKDAY_APPR_PROCESS_START need change in data type they must be classified as ordered categorical variables.

Handling Missing values in columns

1. There are 40 columns with high percentage of missing values(>50).
2. We can deal with them in the following ways:
 - a) Get more data, explore and try to fill those missing values with actual data.
 - b) We refrain from imputing values based on existing data because it may lead to data distortion. While imputing values to such high degree we would just be adding noise to the data which can significantly affect the results.
 - c) Drop the columns. Since the columns have very high missing value percentage they are less likely to give us an accurate results. So it is better to drop them
 - d) In our case since we do not have any alternative data source we have decided to drop all columns whose missing value percentage is greater than 50

Handling Missing values strategy

We have to deal with columns with low missing percentage values(<15). We can use the following methods to deal with it. We can deal with them in the following ways:

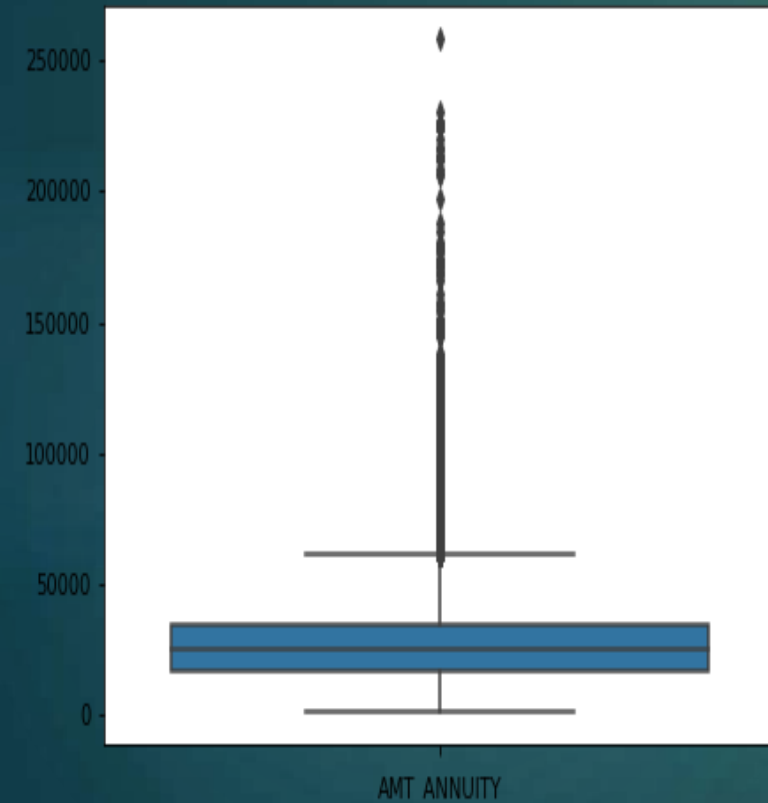
- A. Do not impute those values. Let them as it is and try to exclude them from the calculation. This way we can fill those values with actual data at a later time
- B. We can use mean or median to impute missing values. In some cases we may use specific number to fill up the missing values like 0,1 etc.
- C. There is no fixed approach to impute missing values. The approach may differ from person to person and column to column
- D. Now we will see some examples on how to handle missing values



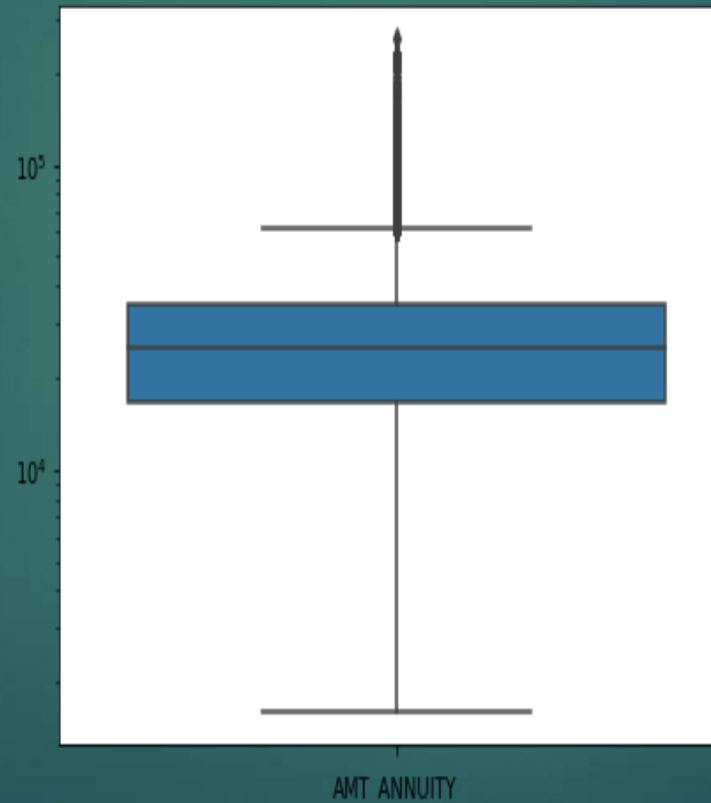
*Examples of how we dealt with
missing values in some columns*

Example 1: AMT_ANNUIITY Numerical Column

Boxplot to see outliers



Log Scale Representation



Inference

Boxplot show presence of outliers which means we cannot impute missing values with mean as mean is highly affected by outliers. We have the following options to choose from for imputing values.

1. Impute the missing values with median since it is not affected by outliers
2. We can remove the outliers and then impute the missing values with mean

Example 2: NAME_TYPE_SUITE Categorical Column

Value count of distinct values

```
curapp.loc[:, "NAME_TYPE_SUITE"].value_counts()
```

Unaccompanied	248526
Family	40149
Spouse, partner	11370
Children	3267
Other_B	1770
Other_A	866
Group of people	271

Name: NAME_TYPE_SUITE, dtype: int64

Inference

The column NAME_TYPE_SUITE is a categorical column so we have following options to choose for imputing missing values.

1. We can impute missing value with the most frequent value. In our case it is "Unaccompanied"
2. We can introduce a new value called "Unknown" to mark that this value is not known so that we can change it after getting additional data.

Example 3 :CNT_FAM_MEMBERS Numeric Column

Value count of distinct values

```
curapp.loc[:, "CNT_FAM_MEMBERS"].value_counts()
```

2.0	158357
1.0	67847
3.0	52601
4.0	24697
5.0	3478
6.0	408
7.0	81
8.0	20
9.0	6
10.0	3
14.0	2
16.0	2
12.0	2
20.0	2
11.0	1
13.0	1
15.0	1

```
Name: CNT_FAM_MEMBERS, dtype: int64
```

Inference

The column CNT_FAM_MEMBERS is a numerical column with discrete values. So we cannot impute missing values with mean. The most sensible choice would be to impute the missing values with 2 as it has maximum frequency

Remark

Based on our understanding we have selected some columns and we would be doing our analysis on those columns only.

List of columns which we deemed important are:

"SK_ID_CURR","TARGET","NAME_CONTRACT_TYPE","CODE_GENDER","FLAG_OWN_CAR","FLAG_OWN_REALTY","CNT_CHILDREN","AMT_INCOME_TOTAL","AMT_CREDIT","AMT_ANNUITY","AMT_GOODS_PRICE","NAME_TYPE_SUITE","NAME_INCOME_TYPE","NAME_EDUCATION_TYPE","NAME_FAMILY_STATUS","NAME_HOUSING_TYPE","DAYS_BIRTH","OCCUPATION_TYPE","CNT_FAM_MEMBERS","WEEKDAY_APPR_PROCESS_START","HOUR_APPR_PROCESS_START","ORGANIZATION_TYPE","OBS_60_CNT_SOCIAL_CIRCLE","DEF_60_CNT_SOCIAL_CIRCLE","AMT_REQ_CREDIT_BUREAU_QRT".

Important Note

The DAYS_BIRTH column has negative values. Although days cannot be negative but this is our understanding that the negative value is because the data source from where the data has been retrieved might have calculated it by subtracting it from the current date.

Why negative sign is important?

We think the negative sign in days column can work to an advantage. There are number of column which contain number of days like when the loan application expired etc. The sign indicates whether it is a coming date or the date has passed. Thus we think it is logical to follow the same sign convention and also signs + and - in the DAYS columns can prove useful in calculating dates thus preventing any confusion.

But since we as human beings cannot infer much data from number of days thus we have created a new column called AGE using DAYS_BIRTH column.

Treating Numerical columns

We will check data quality of the numerical columns in our data set. It will include the following things:

- Check data distribution of the column
- Identify the outliers.

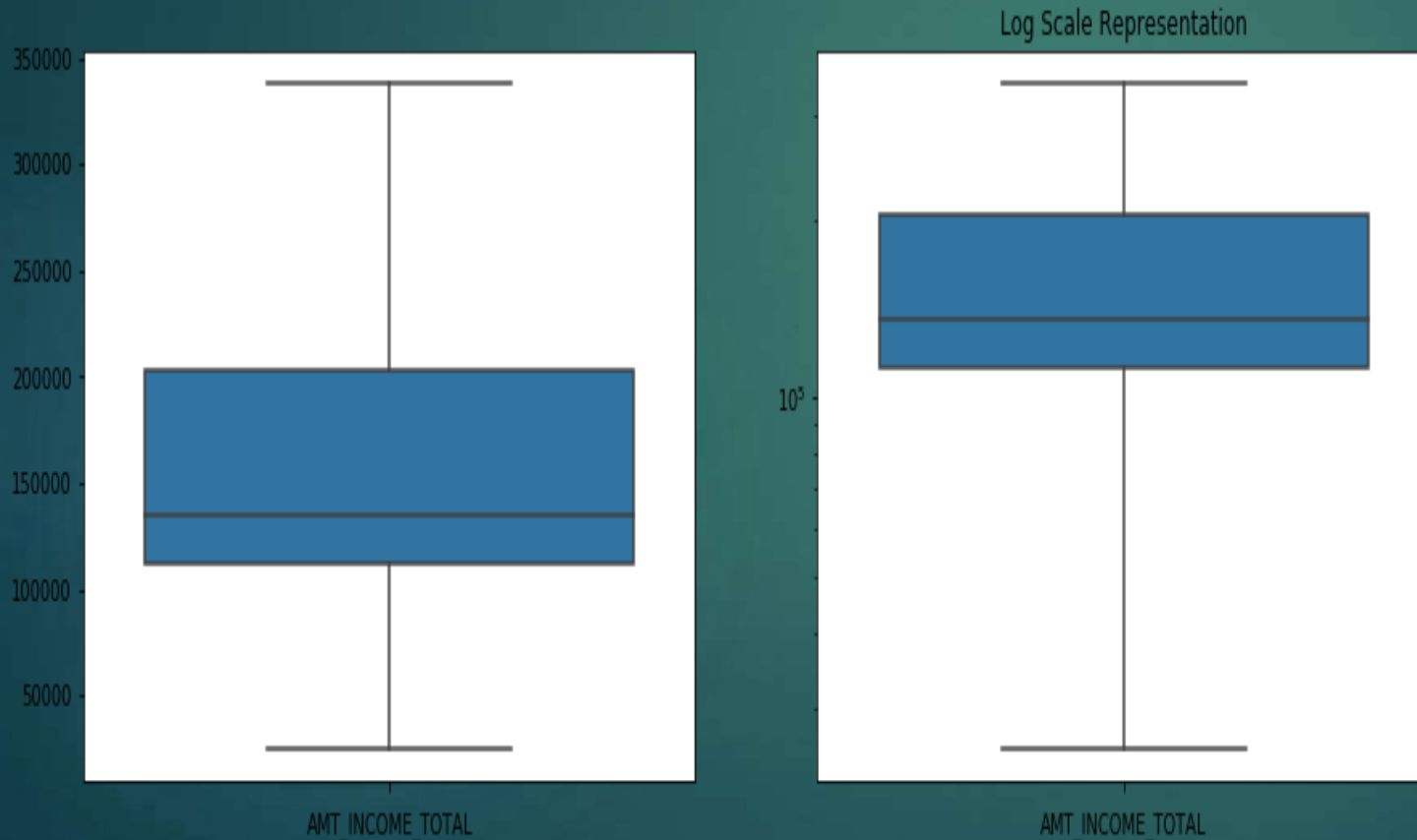
How we discovered outliers ?

1. We first used a box plot to check the presence of outliers in our data.
2. To find exact values we used interquartile distance method.
3. We first calculated interquartile distance which is given as difference of 75 and 25 percentile.
4. Once we calculated the IQD we used it to calculate a upper and lower bound.
 - $\text{Upperbound} = q3 + 1.5 * \text{IQD}$
 - $\text{Lowerbound} = q1 - 1.5 * \text{IQD}$
5. Any value that does not lie in range $[\text{Lowerbound}, \text{Upperbound}]$ is considered an outlier.

Treating Outlier

AMT_INCOME_TOTAL column

Boxplot After Outlier Treated

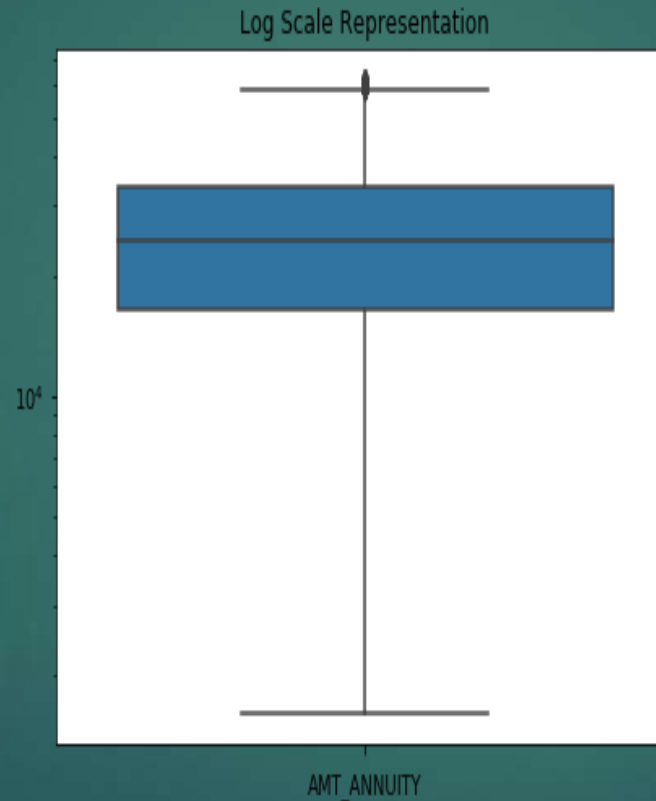
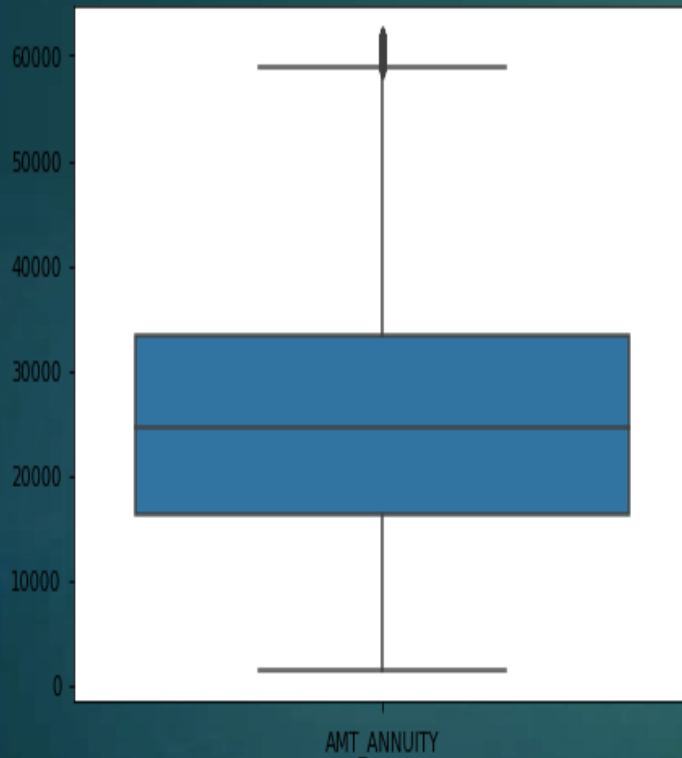


Inference

- ❑ The outliers have been treated successfully as it is evident from the boxplot
- ❑ The boxplot shows that most of the data is concentrated between 1,20,000-2,20,000
- ❑ Minimum value of AMT_INCOME_TOTAL is 25650 whereas maximum value is 337500
- ❑ The thickness of the boxplot suggest that data distribution is not very wide i.e data is not spread over a large range.

Treating Outlier in AMT_ANNUIITY

Boxplot After Outlier Treated



Inference

- ❑ The outliers have been treated successfully as it is evident from the boxplot. However there are still some values which can be considered as outliers this suggest we need to fine tune our outlier detection method.
- ❑ The boxplot shows that most of the data is concentrated between 16000-24000
- ❑ Minimum value of AMT_ANNUIITY is 1615 whereas maximum value is 61699
- ❑ The thickness of the boxplot suggest that data distribution is not very wide i.e data is not spread over a large range.
- ❑ The distance of the upper whisker from second quartile indicates that maximum value is differs largely from general distribution.

Binning some continuous variables

Binning AGE column

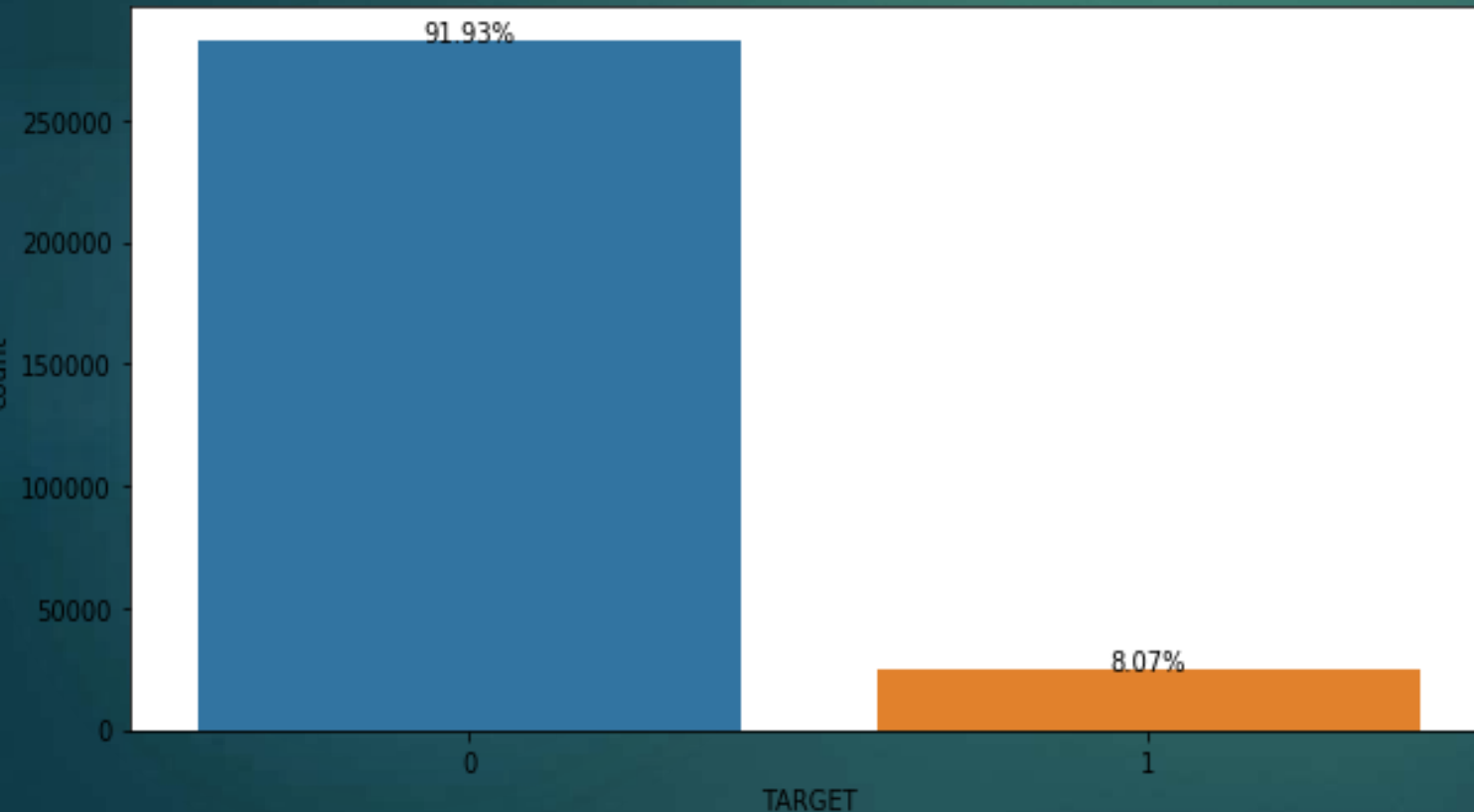
- ❑ We have divided age groups in three columns: Young, MIDDLE_AGED, Old
- ❑ People with age less than 30 are considered Young.
- ❑ People with age between 30-60 are classified as MIDDLE_AGE.
- ❑ People having age above 60 are categorized as OLD.

Binning AMT_INCOME_TOTAL Column

- ❑ We have divided AMT_INCOME in three columns: UPPER_CLASS, MIDDLE_CLASS, LOWER_MIDDLE_CLASS, LOWER_CLASS.
- ❑ People with income less than 30,000 are considered LOWER_CLASS.
- ❑ People with income between 30,000-60,000 are classified as LOWER_MIDDLE_CLASS.
- ❑ People having income between 60,000 and 2,00,000 are categorized as MIDDLE_CLASS
- ❑ People who earn more than 2,00,000 are classified as UPPER_CLASS

Checking Data Imbalance for Target variable

Frequency Distribution



Insights

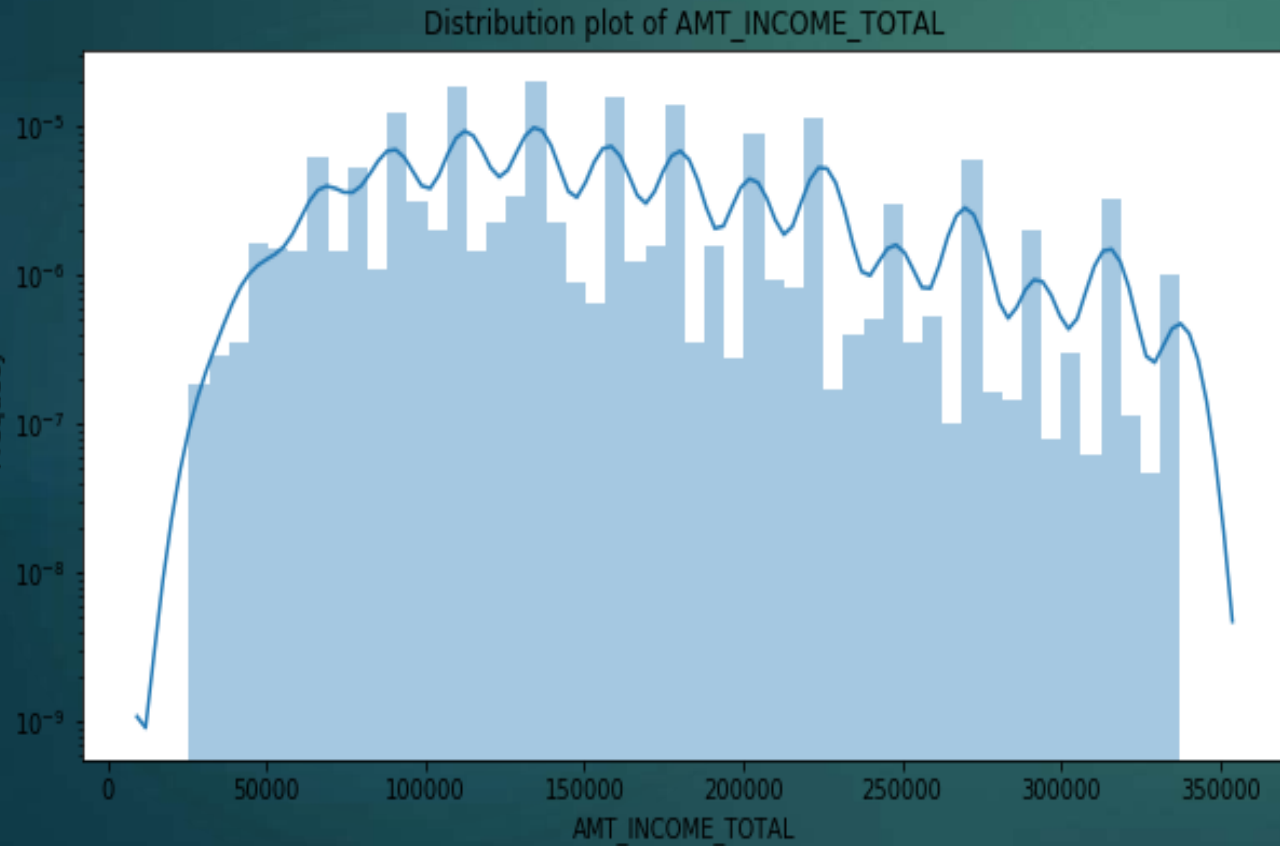
- ❑ We have found imbalance ratio as 11.3.
- ❑ The plot also shows that there is a significant imbalance in TARGET column

Approach

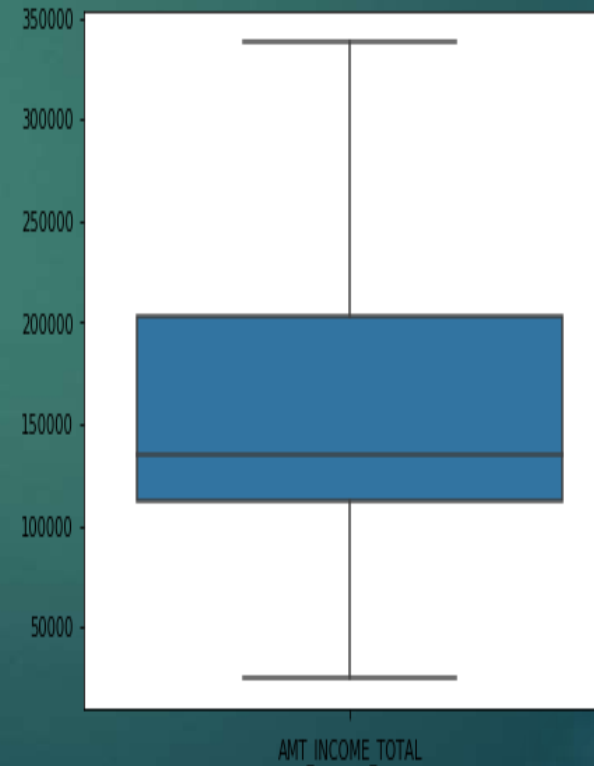
1. We will now divide the dataframe `df` into two data frames `dfdefault` and `dfnotdefault`.
2. These two data frames will contain data of defaulters and non defaulters respectively.
3. Using these dataframes we will now try to see if there are any patterns between different variables and their effect on people defaulting their loan.
4. We will see which group of people are safe to grant a loan and which characteristics can say that person is likely to default on his/her loan.

Univariate Analysis on Numerical Columns (AMT_INCOME_TOTAL)

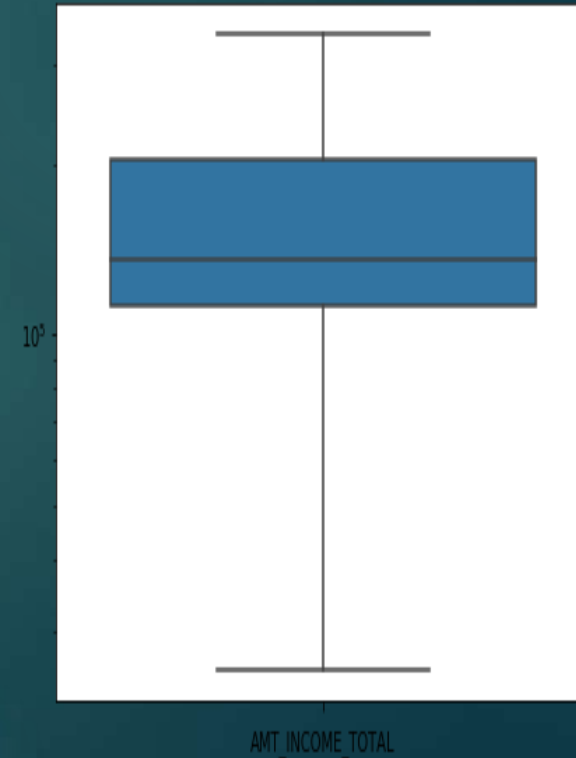
Distribution Plot



Boxplot



Log Scale Representation

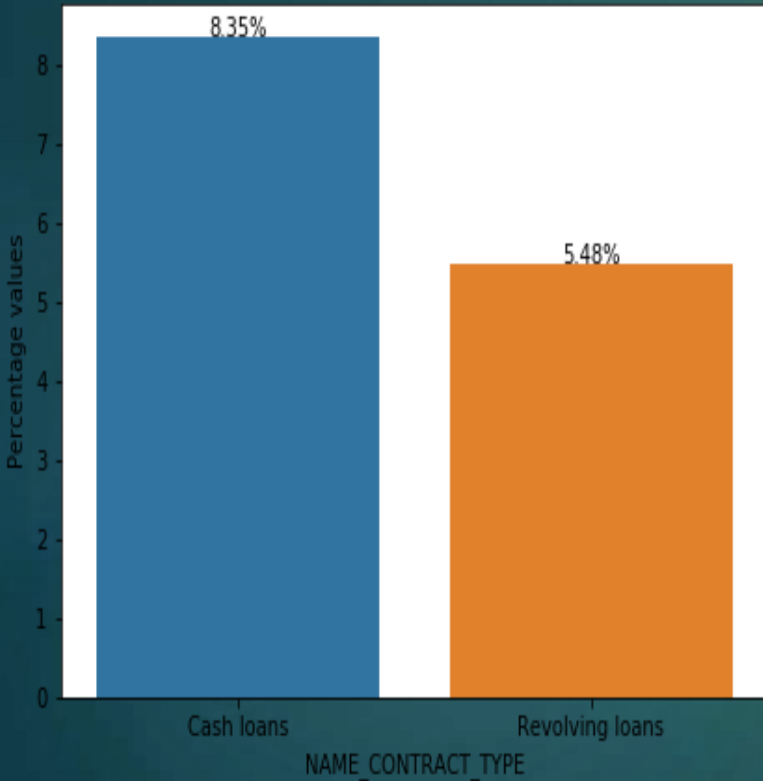


Insights

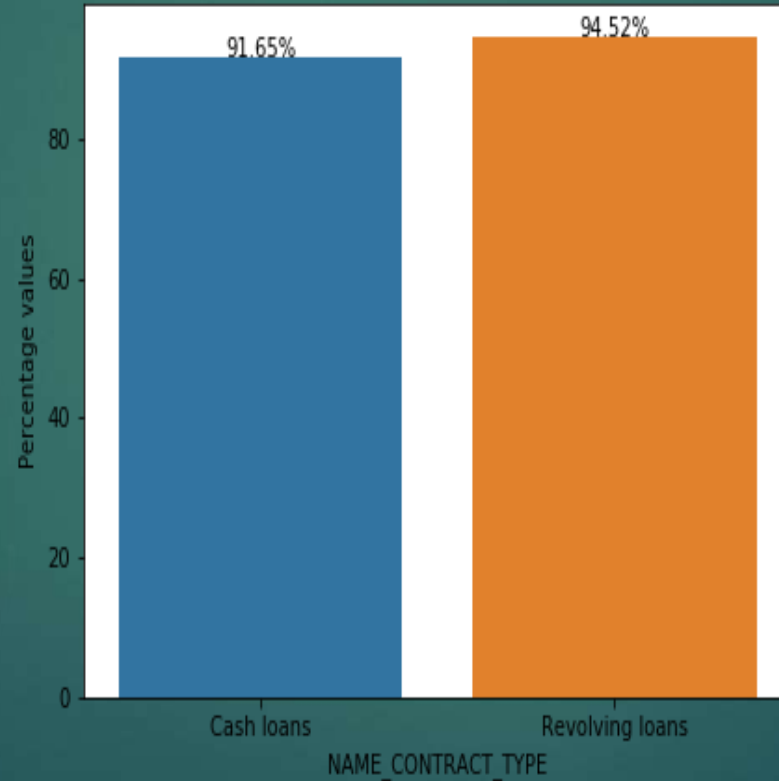
- ❑ *The Distribution plot did not provide much insight but we can say that for income less than 50,000 and greater than 3,40,000 the frequency is low(it drops).For all other values there is a continuous increase and decrease in frequency levels.*
- ❑ *From box plot we saw that majority of the population income lies between 1,00,000 and 2,00,000*
- ❑ *Minimum value is 25,650 and maximum value is 3,37.500. We can also estimate these values from the boxplot shown above.*

Univariate Analysis on Categorical Columns (NAME_CONTRACT_TYPE)

Percentage plot for Defaulters



Percentage plot for Non_Defaulters



Observation and Conclusion

- ❑ 5.48% of revolving loans are defaulted.
- ❑ 8.35% of people who take cash loans default on their loans.
- ❑ Default Percentage of cash loans is higher than revolving loans.
- ❑ The data suggests that granting cash loans to customers involves higher risk than granting them revolving loans.
- ❑ The data supports our understanding that revolving loans are less risky because of the flexibility they provide to the customer.

Correlation on Numeric Columns

List of columns which we deemed important for finding correlation are :

1. AMT_INCOME_TOTAL
2. AMT_CREDIT
3. AMT_ANNUITY
4. AMT_GOODS_PRICE
5. DAYS_BIRTH
6. OBS_60_CNT_SOCIAL_CIRCLE
7. DEF_60_CNT_SOCIAL_CIRCLE
8. AMT_REQ_CREDIT_BUREAU_QRT

Correlation for Defaulter



Observation

- ❑ *AMT_CREDIT and AMT_GOODS_PRICE are highly positive correlated with correlation 0.98*
- ❑ *AMT_CREDIT and DAYS_BIRTH are negatively correlated with correlation -0.14*
- ❑ *Top 5 positively correlated columns in decreasing order:*
 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*
- ❑ *Top 5 negatively correlated columns in decreasing order:*
 1. *AMT_CREDIT and DAYS_BIRTH*
 2. *AMT_GOODS_PRICE and DAYS_BIRTH*
 3. *DEF_60_CNT_SOCIAL_CIRCLE and AMT_CREDIT*
 4. *DEF_60_CNT_SOCIAL_CIRCLE and AMT_GOODS_PRICE*
 5. *DEF_60_CNT_SOCIAL_CIRCLE and AMT_ANNUITY*

Correlation for Non Defaulter



Observation

- ▶ *AMT_CREDIT and AMT_GOODS_PRICE are highly positive correlated with correlation 0.98*
- ▶ *AMT_CREDIT and DAYS_BIRTH are negatively correlated with correlation -0.049*
- ▶ *Top 5 positively correlated columns in decreasing order:*
 - ▶ *AMT_CREDIT and AMT_GOODS_PRICE*
 - ▶ *AMT_CREDIT and AMT_ANNUITY*
 - ▶ *AMT_ANNUITY and AMT_GOODS_PRICE*
 - ▶ *AMT_ANNUITY and AMT_INCOME_TOTAL*
 - ▶ *AMT_CREDIT and AMT_INCOME_TOTAL*
- ▶ *Top 5 negatively correlated columns in decreasing order:*
 - ▶ *AMT_CREDIT and DAYS_BIRTH*
 - ▶ *AMT_GOODS_PRICE and DAYS_BIRTH*
 - ▶ *DEF_60_CNT_SOCIAL_CIRCLE and AMT_INCOME_TOTAL*
 - ▶ *DEF_60_CNT_SOCIAL_CIRCLE and AMT_CREDIT*
 - ▶ *DEF_60_CNT_SOCIAL_CIRCLE and AMT_ANNUITY*



Conclusions

1. *Top 5 positivity correlated column pairs are same in defaulter and non defaulter case but the the order of correlation is not same.*
2. *Top 5 negatively correlated column pairs are not same in defaulter and non defaulter case*

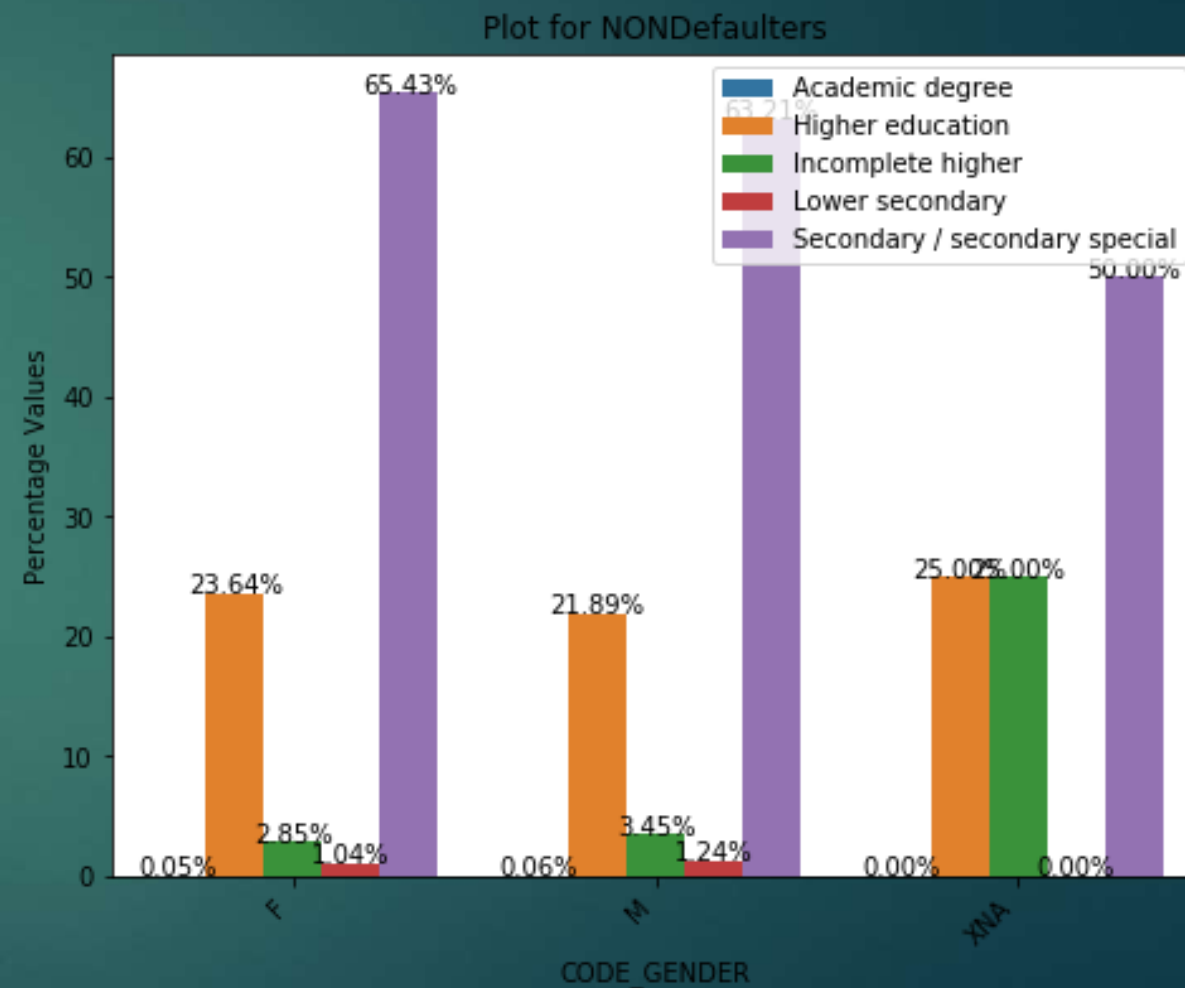
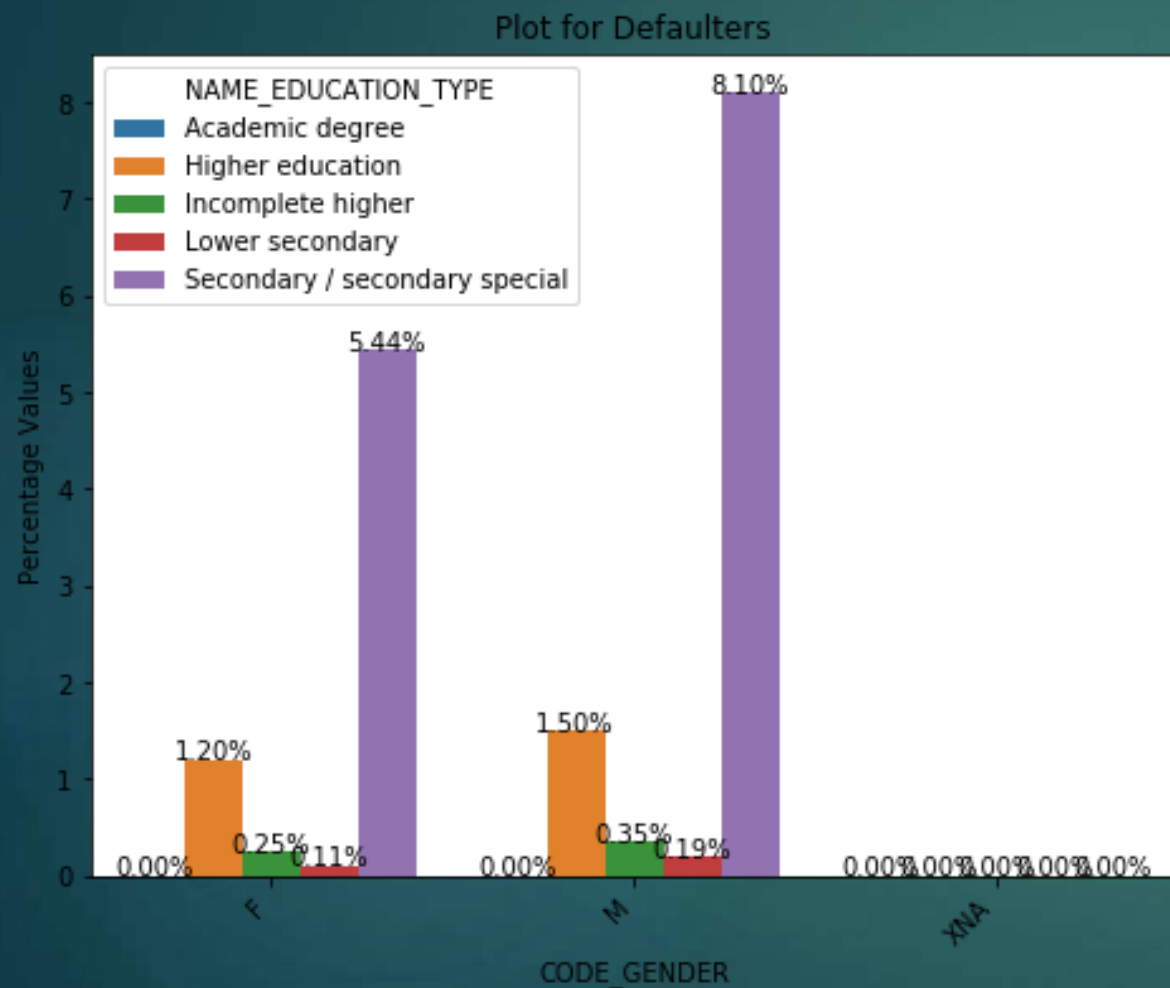
Bivariate Analysis (Categorical Columns)

We have made some conclusions considering a single categorical variable and comparing it with respect to Target variable (Defaulter, Non Defaulter). Now we will take it a step further and see how combination of two categorical variables affect the target variables. We will draw inferences taking two cases.

Example 1: Columns involved CODE_GENDER, NAME_EDUCATION_STATUS

- ❑ For CODE_GENDER we observed that male customers are more likely to default than females.
- ❑ For NAME_EDUCATION_STATUS we observed that education level has impact on target variable. More educated people are less prone to defaulting on their loans.
- ❑ We will now see if we observe the similar trend when we divide people based on gender.

Example 1: Columns involved CODE_GENDER, NAME_EDUCATION_STATUS



Observations:

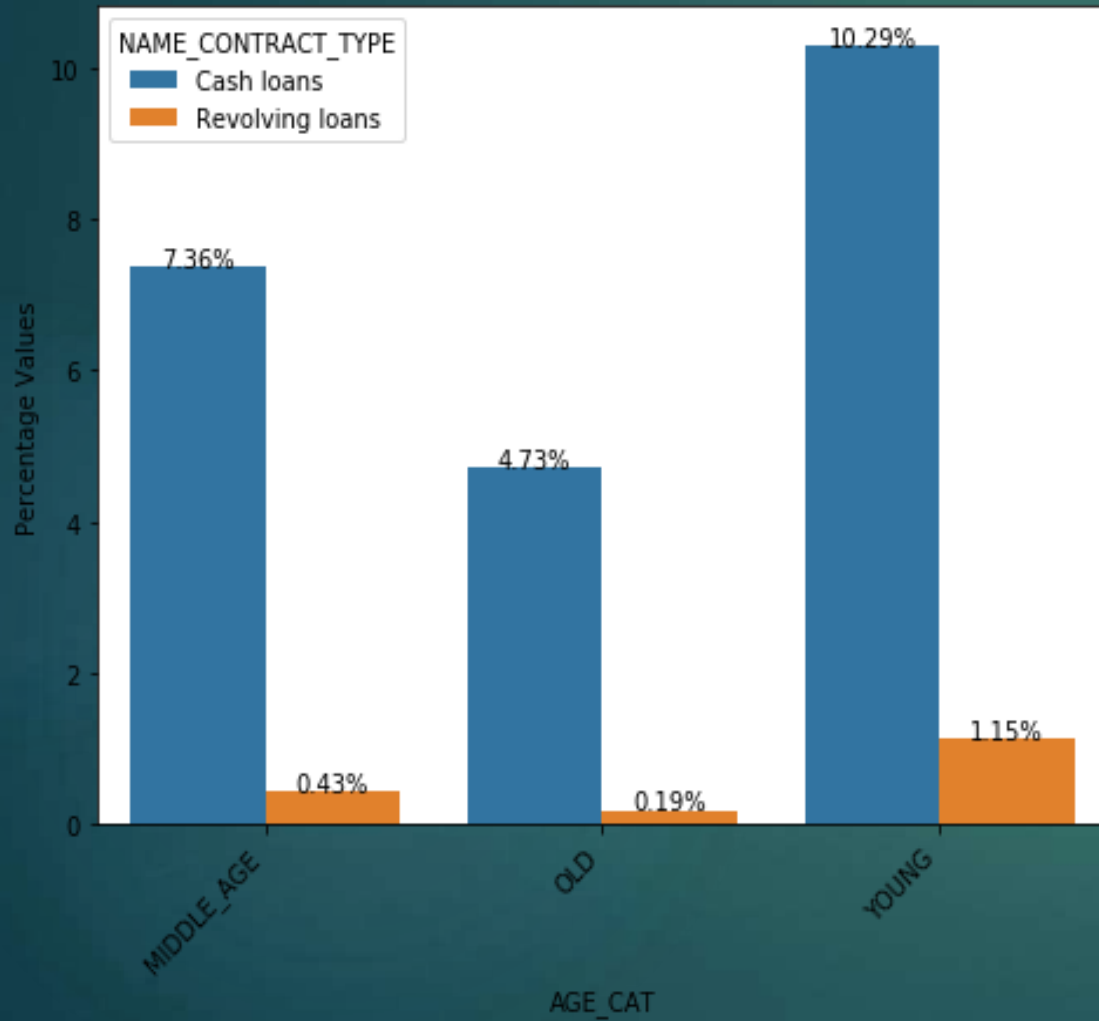
- ❑ Earlier we saw that people with less education are more prone to defaulting their loans and males are more likely to default than females.
- ❑ Among Females who defaulted on their loans 82% (5/7)have secondary education.
- ❑ Among males who defaulted 80% (8/10) of them have secondary education

Conclusion:

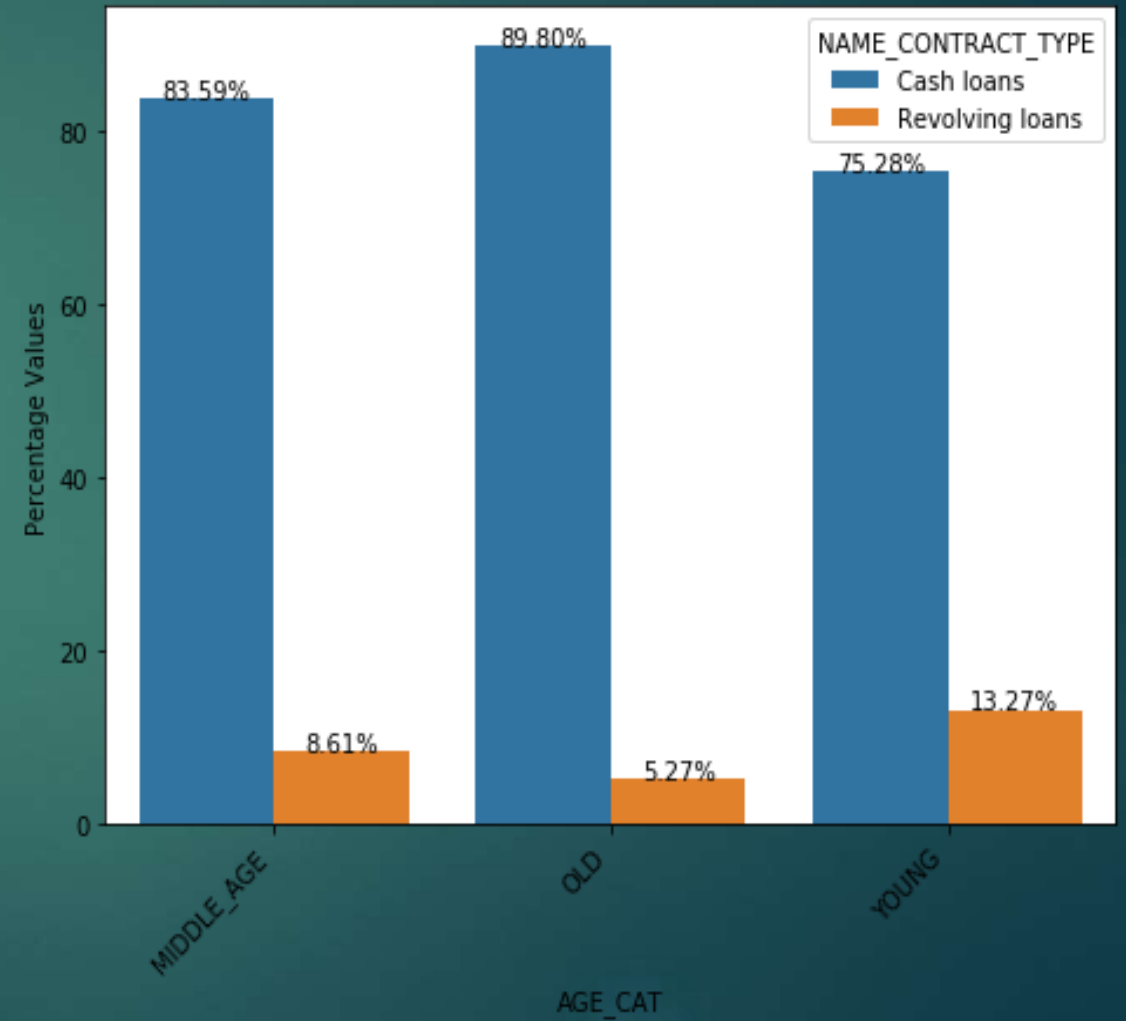
- ❑ The chart still supports our understanding that people who are highly educated are less prone to defaulting on their loans. This is evident from the fact that no person who has an academic degree defaulted.
- ❑ If we see overall than people with low education level default more and people with high education level default less. However among dividing people in males and females this trend is not always true.
- ❑ People with secondary education defies this trend .Both males and females who have secondary education show unusually high percentage in both defaulting and non defaulting.
- ❑ For all other education levels our observation still holds true.
- ❑ If we observe the Non_defaulters chart it can be seen that for both males and females categories that people with higher education do not default(Except for the unusual trend seen in secondary education)

Example 2: Columns involved AGE_CAT, NAME_CONTRACT_TYPE

Plot for Defaulters



Plot for NONDefaulters



Observations:

- 8% of middle aged people were defaulters. 7% of total middle age people defaulted on cash loans
- 5% of old people were defaulters. 4% of the total old people defaulted on cash loans.
- 11% of young people defaulted on their loans, among those 10% were cash loan defaulters.

Conclusion:

- The chart supports our initial understanding that old people are less likely to default. If we observe the chart on the right we will observe that 90% of old people paid back their cash loans and 5.27% paid back their revolving loans. So if a person is old and is asking for cash loan he is very likely to pay it back so bank should approve it.
- Similar trend is observed in middle aged people. 83% of total middle aged people paid back their cash loan and 9% of middle age people paid back their revolving loan thus only 8% of middle aged people defaulted on their loans.
- Upon further observation we conclude that if a person is middle aged and is asking for revolving loan he is likely to pay it back. But if he is asking for cash loan there is some risk associated with it.
- For young people we saw that 85% (75+10) of them have cash loans. 10% of total young people defaulted on cash loan whereas 75% paid them.
- For young people we saw that there is 1 in 14 approx. 7% chance that they will default on revolving loans whereas there is 10 in 85 approx. 11.76% chance that they will default on cash loans so we conclude that providing revolving loans to young people is a less risky endeavour



Analysis On Previous Application File

Data Understanding

1. File has 1670214 rows with 37 columns
2. After observing the data we have come to the following conclusions :
 - ❑ Column `SK_ID_PREV`, `SK_ID_CURR` are of type `int` which does not make sense. Although it contains only integer values but we cannot perform any numerical operation on them thus we should convert its data type to string.
 - ❑ There are some columns like `NAME_CONTRACT_TYPE`, `NAME_CONTRACT_STATUS`, `CODE_REJECT_REASON`, `NAME_CLIENT_TYPE` which are of type `object`. Such columns are best described as categorical columns rather than string.
 - ❑ The type of few of the Flag variables is `int` which is good because then we can use them for calculations more easily we can classify them as categorical but we think its best to leave them as `int`.
 - ❑ Columns like `WEEKDAY_APPR_PROCESS_START` need change in data type they must be classified as ordered categorical variables.

Checking Missing values in columns

1. There are 4 columns with high percentage of missing values(>50).
2. We can deal with them in the following ways :
 - ❑ Get more data, explore and try to fill those missing values with actual data.
 - ❑ We refrain from imputing values based on existing data because it may lead to data distortion. While imputing values to such high degree we would just be adding noise to the data which can significantly affect the results.
 - ❑ Drop the columns. Since the columns have very high missing value percentage they are less likely to give us an accurate results. So it is better to drop them.
 - ❑ In our case since we do not have any alternative data source we have decided to drop all columns whose missing value percentage is greater than 50

Handling Missing values strategy

There are 4 columns with high percentage of missing values(>50). We can deal with them in the following ways.

- ❑ Get more data, explore and try to fill those missing values with actual data.
- ❑ We refrain from imputing values based on existing data because it may lead to data distortion. While imputing values to such high degree we would just be adding noise to the data which can significantly affect the results
- ❑ Drop the columns. Since the columns have very high missing value percentage they are less likely to give us an accurate results. So it is better to drop them
- ❑ In our case since we don't have any alternative data source we have decided to drop all columns whose missing value percentage is greater than 50

Handling Missing values strategy

- 1) Now we have to deal with columns with low missing percentage values(<15).We can use the following methods to deal with it.
- 2) Do not impute those values.Let them as it is and try to exclude them from the calculation.This way we can fill those values with actual data at a later time
- 3) We can use mean or median to impute missing values.In some cases we may use specific number to fill up the missing values like 0,1 etc.
- 4) There is no fixed approach to impute missing values.The approach may differ from person to person and column to column



*Examples of how we dealt with
missing values in some columns*

Example 1: PRODUCT_COMBINATION

Categorical Column

Count of values

Cash	285990
POS household with interest	263622
POS mobile with interest	220670
Cash X-Sell: middle	143883
Cash X-Sell: low	130248
Card Street	112582
POS industry with interest	98833
POS household without interest	82908
Card X-Sell	80582
Cash Street: high	59639
Cash X-Sell: high	59301
Cash Street: middle	34658
Cash Street: low	33834
POS mobile without interest	24082
POS other with interest	23879
POS industry without interest	12602
POS others without interest	2555

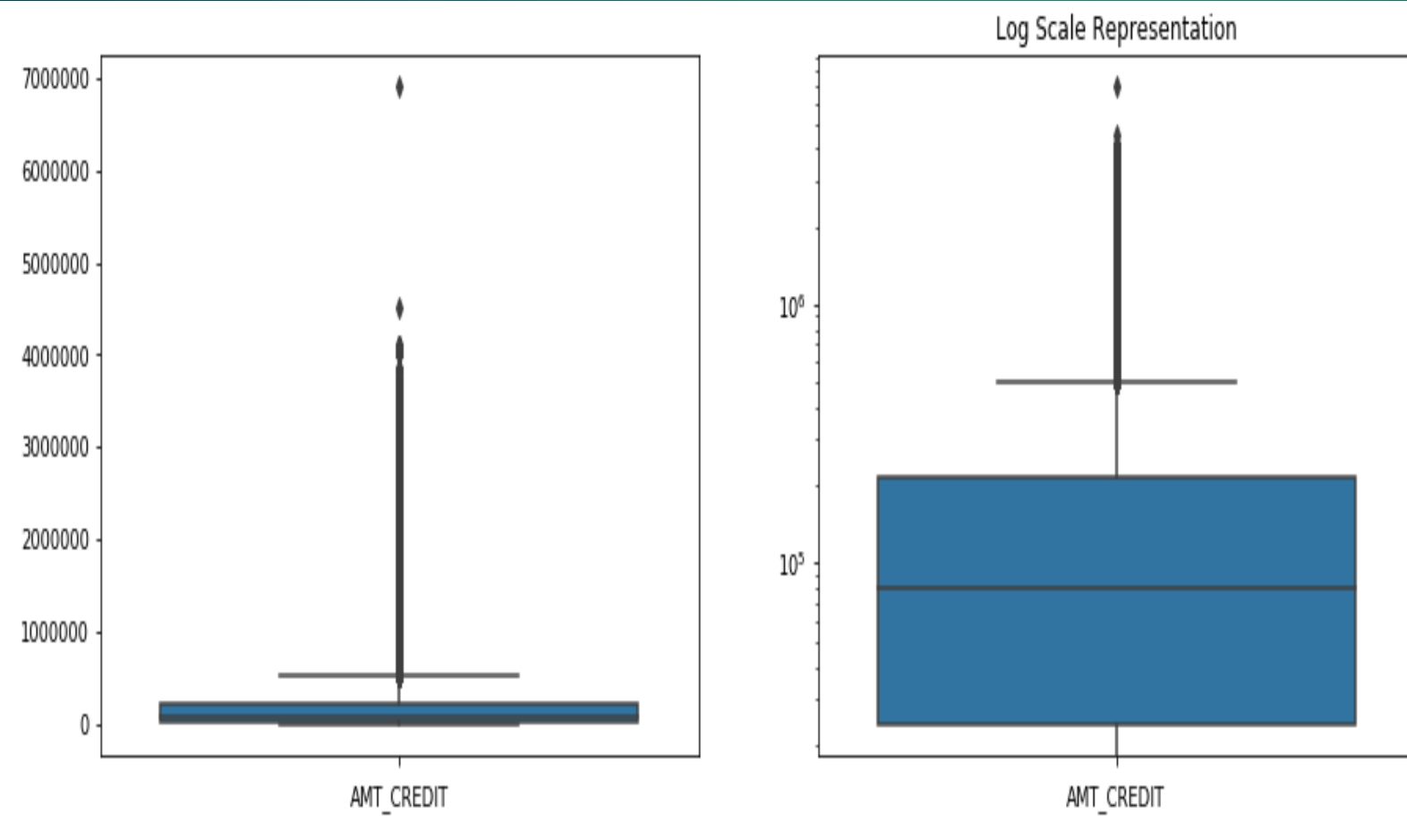
Name: PRODUCT_COMBINATION, dtype: int64

The column `PRODUCT_COMBINATION` is a categorical column so we have following options to choose for imputing missing values.

- ▶ We can impute missing value with the most frequent value. In our case it is "Cash"
- ▶ We can introduce a new value called "Unknown" to mark that this value is not known so that we can change it after getting additional data


Example 2 AMT_CREDIT Numerical Column

Boxplot to see outliers



Inference

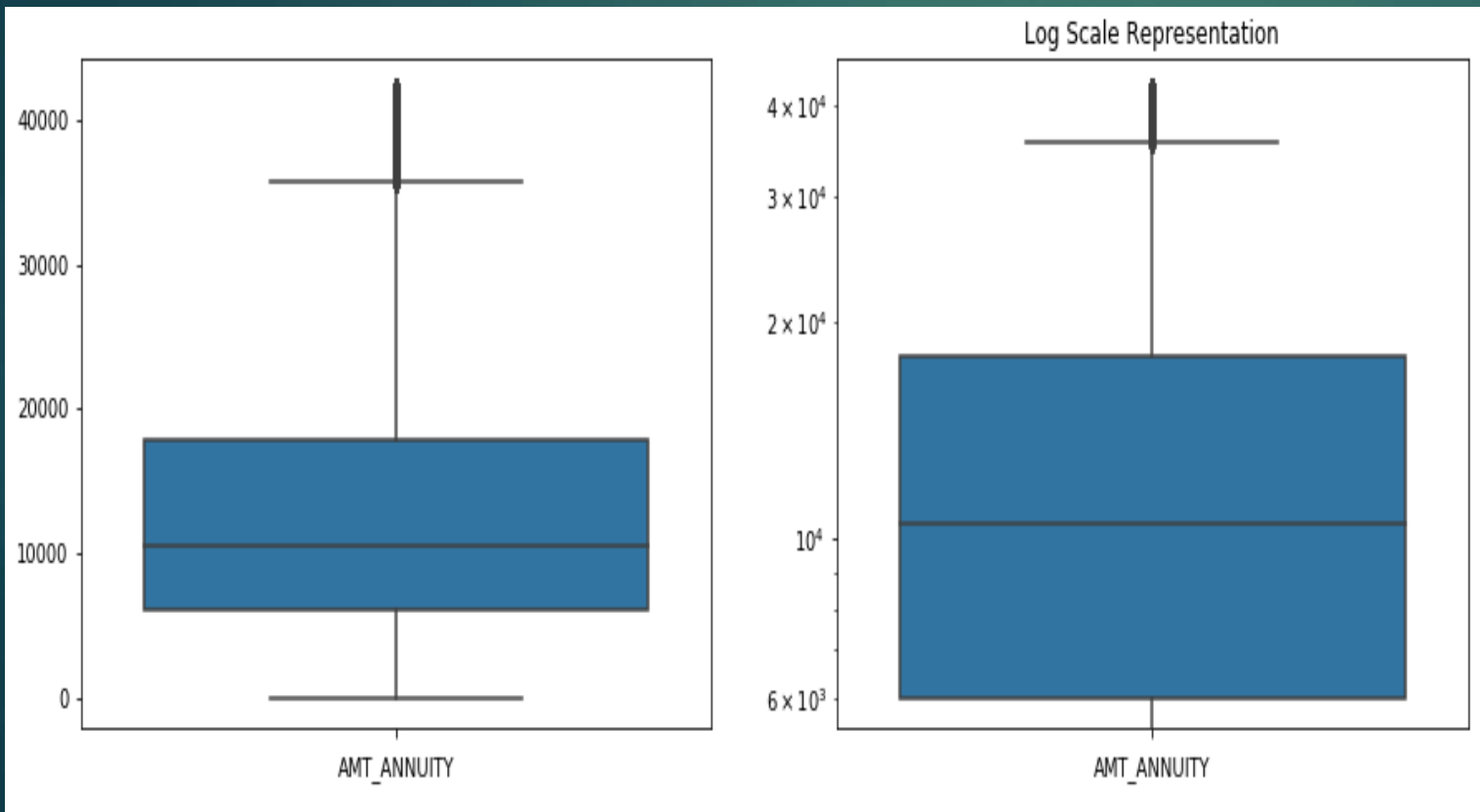
- ❑ Box plot show presence of outliers which means we cannot impute missing values with mean as mean is highly affected by outliers. We have the following options to choose from for imputing values.
- ❑ Impute the missing values with median since it is not affected by outliers
- ❑ We can remove the outliers and then impute the missing values with mean



Treating Outliers *(Some Examples)*

Treating Outliers:AMT_ANNUIITY

After outlier treatment

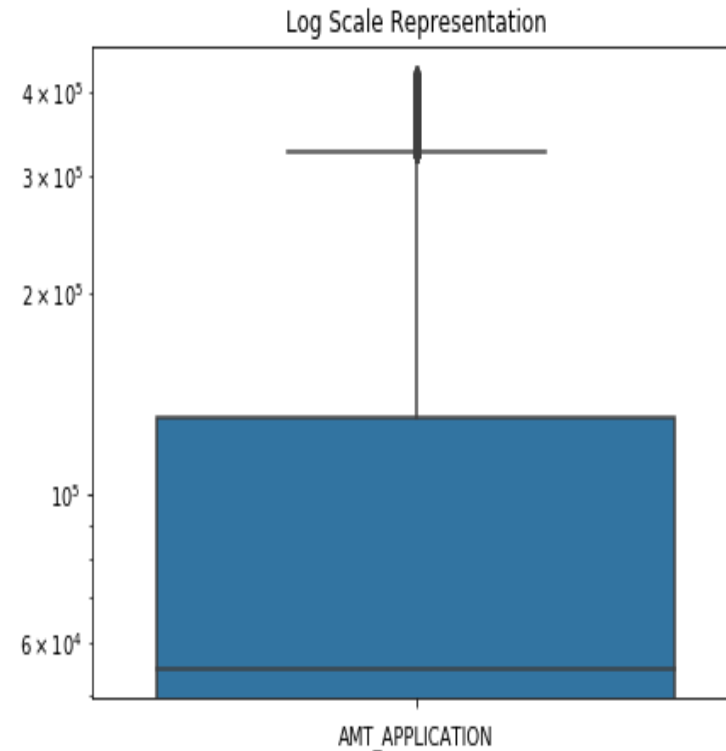
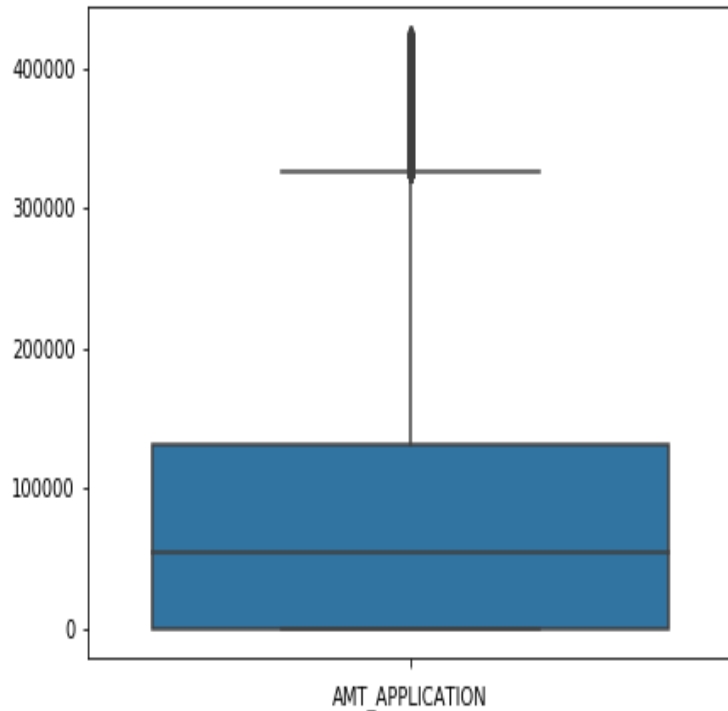


Inference

- ❑ The outliers have been treated but still there are some outliers present. This suggests we need to fine tune our outlier detection method.
- ❑ The boxplot shows that most of the data is concentrated between 6,000-18,000
- ❑ Minimum value of AMT_ANNUIITY is 0 whereas maximum value is 42161
- ❑ The thickness of the boxplot suggests that data distribution is not very wide i.e. data is not spread over a large range

Treating Outliers:AMT_APPLICATION

After outlier treatment

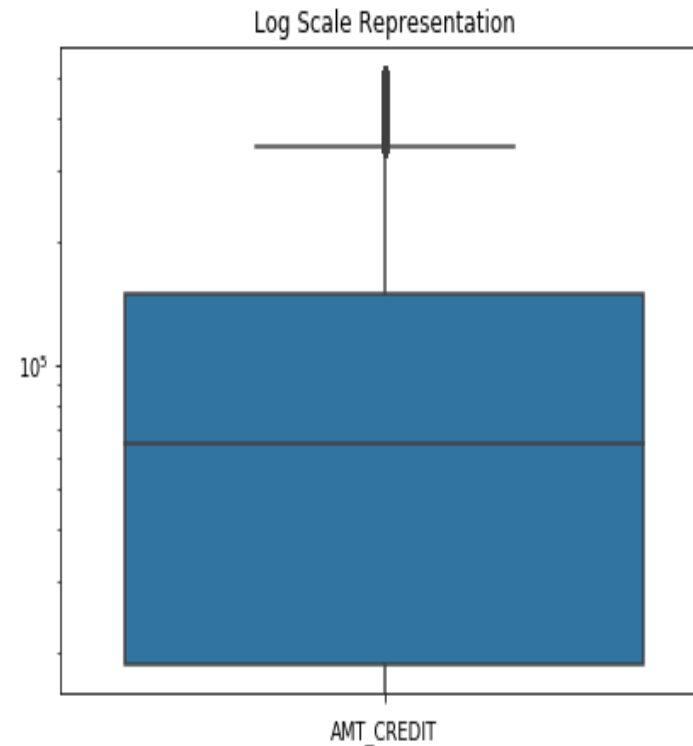
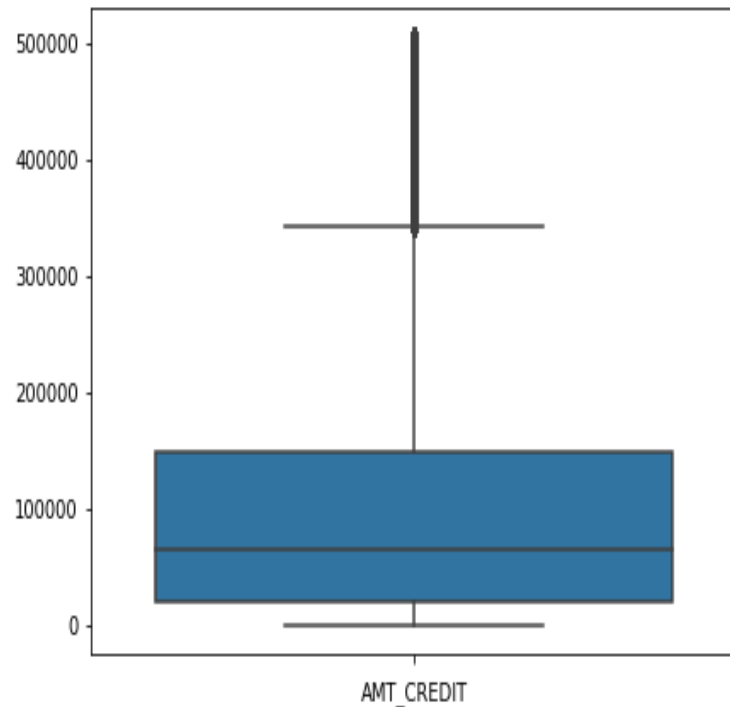


Inference

- ❑ The outliers have been treated but still there are some outliers present. This suggests we need to fine tune our outlier detection method.
- ❑ The boxplot shows that most of the data is concentrated between 0-1,20,000
- ❑ Minimum value of AMT_APPLICATION is 0 whereas maximum value is 422820
- ❑ The thickness of the boxplot suggests that data distribution is not very wide i.e data is not spread over a large range.

Treating Outliers:AMT_CREDIT

After outlier treatment



Inference

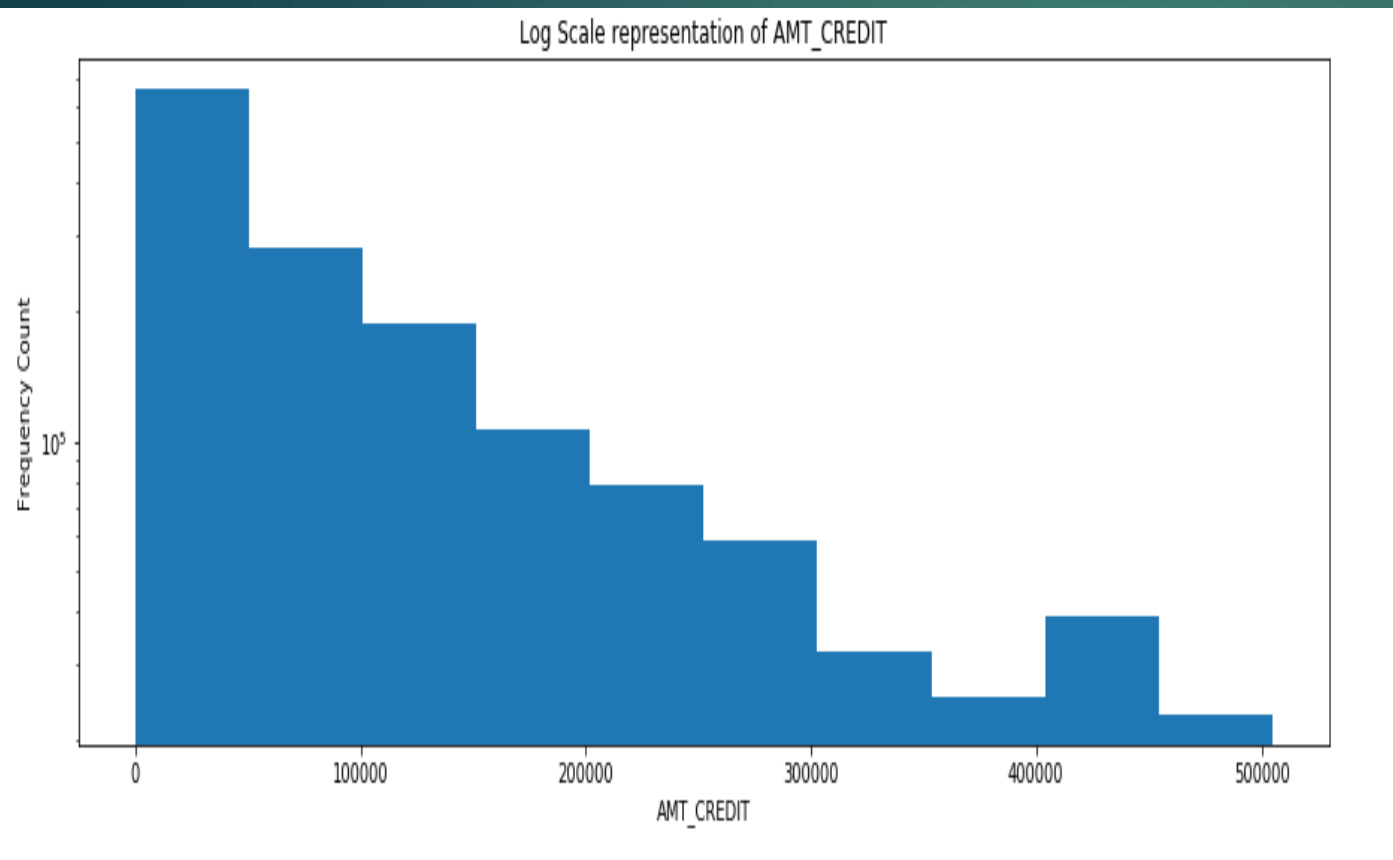
- ❑ The outliers have been treated but still there are some outliers present. This suggests we need to fine tune our outlier detection method.
- ❑ The boxplot shows that most of the data is concentrated between 0-1,20,000
- ❑ Minimum value of AMT_CREDIT is 0 whereas maximum value is 504805.
- ❑ The thickness of the boxplot suggests that data distribution is not very wide i.e data is not spread over a large range.

APPROACH

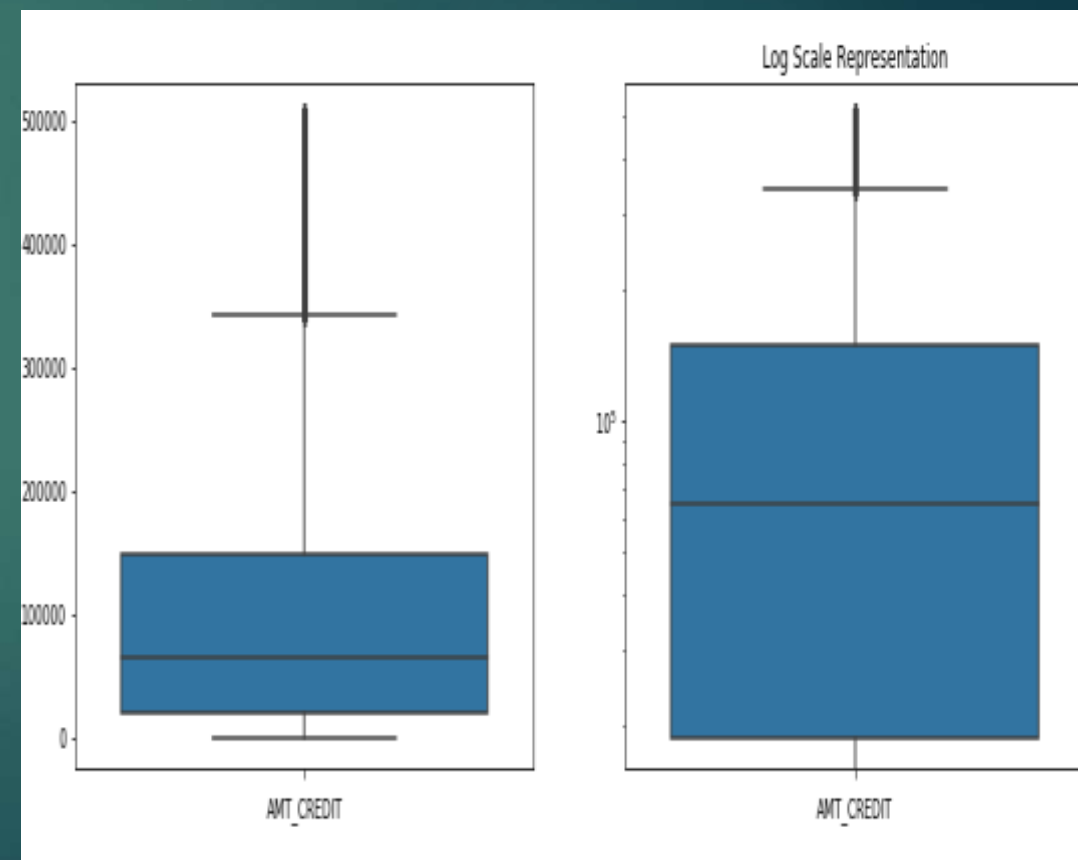
- ❑ We will now divide the dataframe df into four data frames Approved,Canceled,Refused and Unused Offer.
- ❑ These four data frames will contain data of Approved,Canceled,Refused and Unused Offer respectively
- ❑ Using these dataframes we will now try to see if there are any patterns between different variables and their effect on loan getting approved.
- ❑ We will see which group of people are getting loans approved and which characteristics can say that person is likely to get loan approved

Univariate Analysis on Numerical Columns (AMT_CREDIT)

Histogram Plot



Boxplot

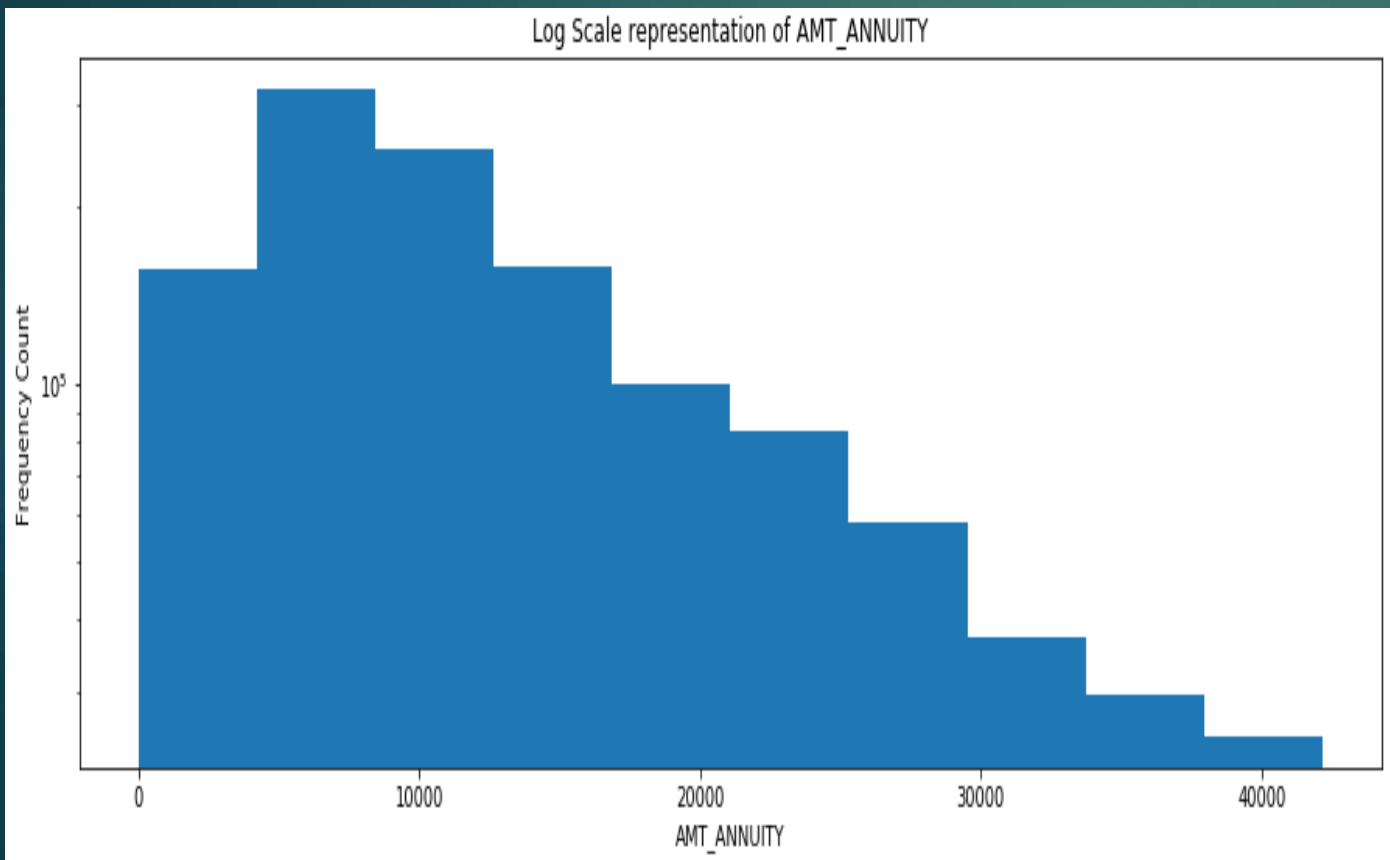


Insights

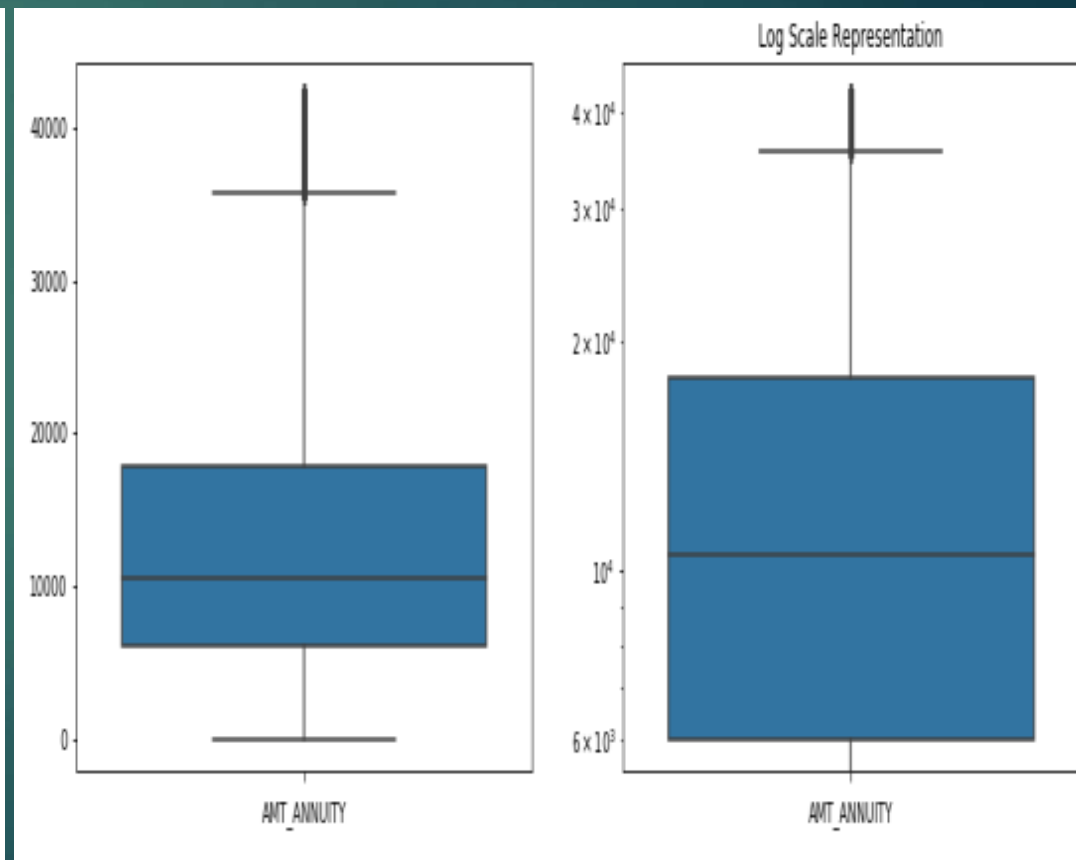
- ❑ The bar plot suggest that AMT_CREDIT follow power law
- ❑ Minimum value is 0 and maximum value is 5048055. We can also estimate these values from the boxplot shown above.
- ❑ Box plot suggest that most of the values of AMT_CREDIT are concentrated between 20,000 and 1,60,000
- ❑ Frequency of AMT_CREDIT between 0-50,000 is highes

Univariate Analysis on Numerical Columns (AMT_ANNUITY)

Histogram Plot



Boxplot



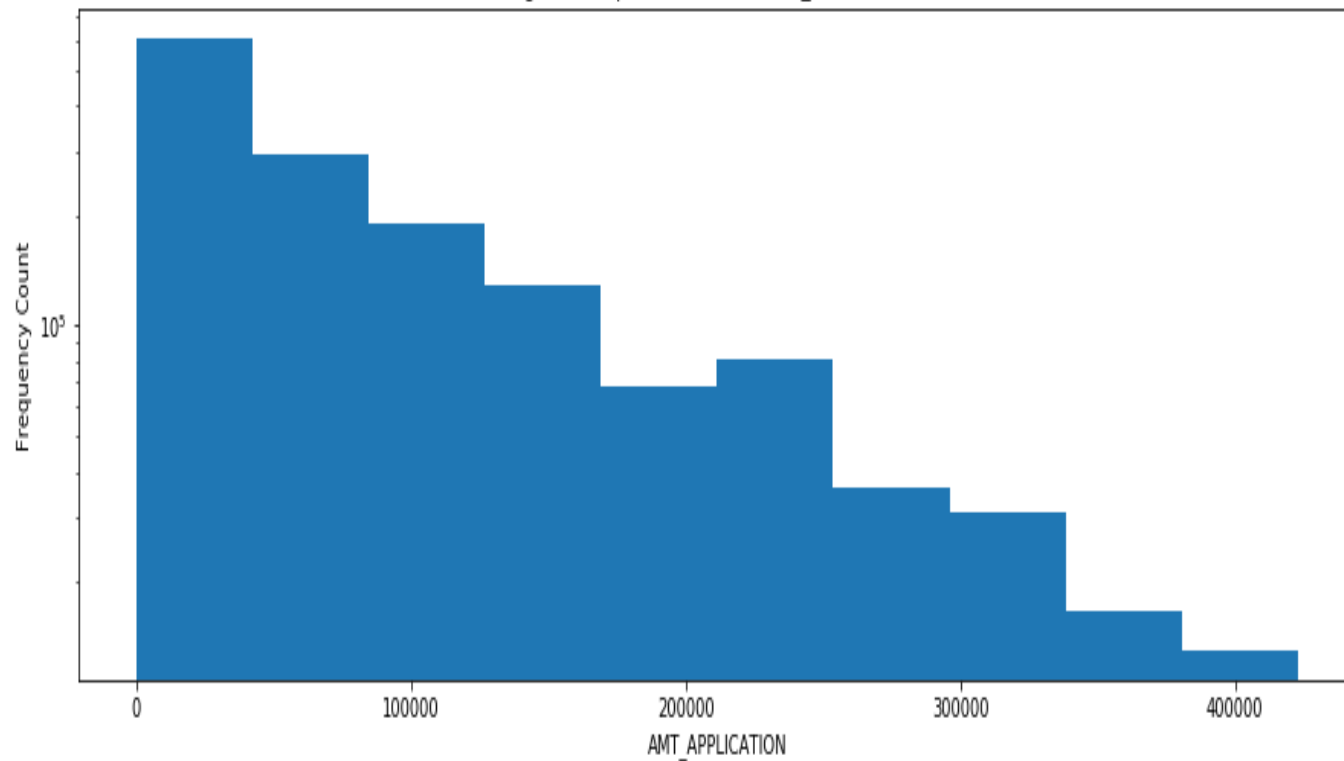
Insights

- ❑ The bar plot suggest that *AMT_ANNUIITY* follow power law after 5,000. This suggest frequency distribution after 5,000 follows a geometric progression.
- ❑ Minimum value is 0 and maximum value is 42,161. We can also estimate these values from the boxplot shown above.
- ❑ Box plot suggest that most of the values of *AMT_ANNUIITY* are concentrated between 5,000 and 18,000

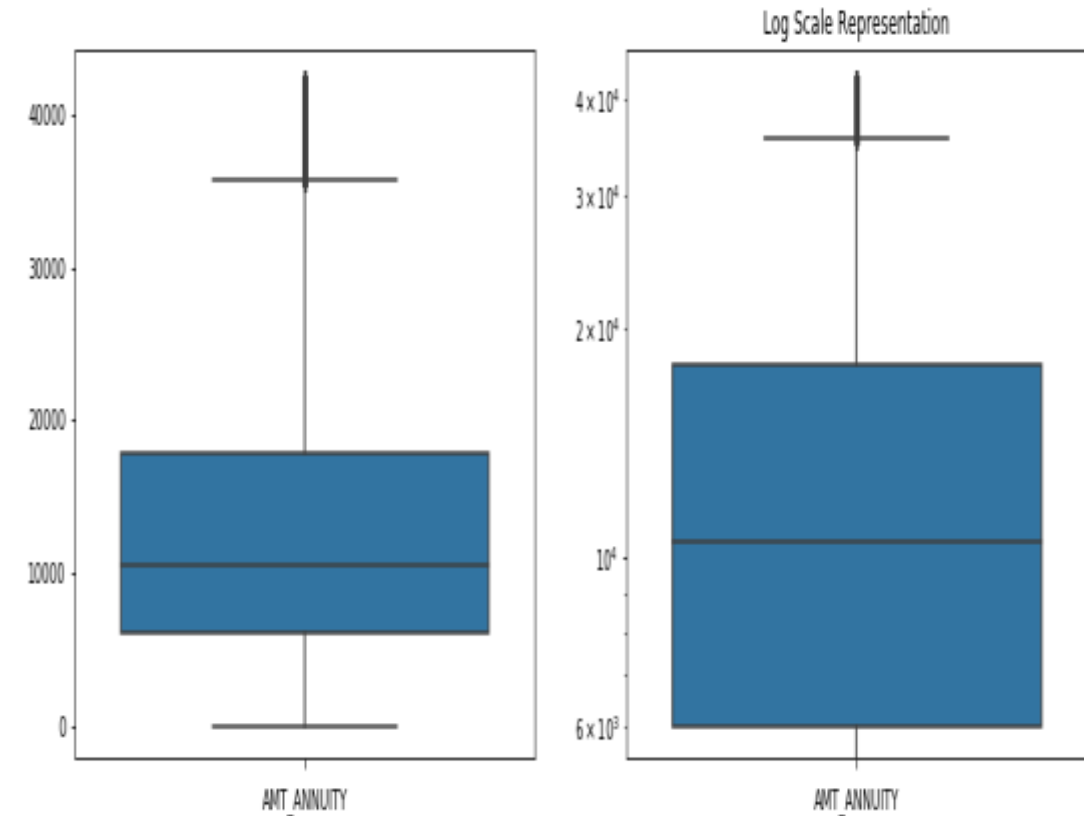
Univariate Analysis on Numerical Columns (AMT_APPLICATION)

Histogram Plot

Log Scale representation of AMT_APPLICATION



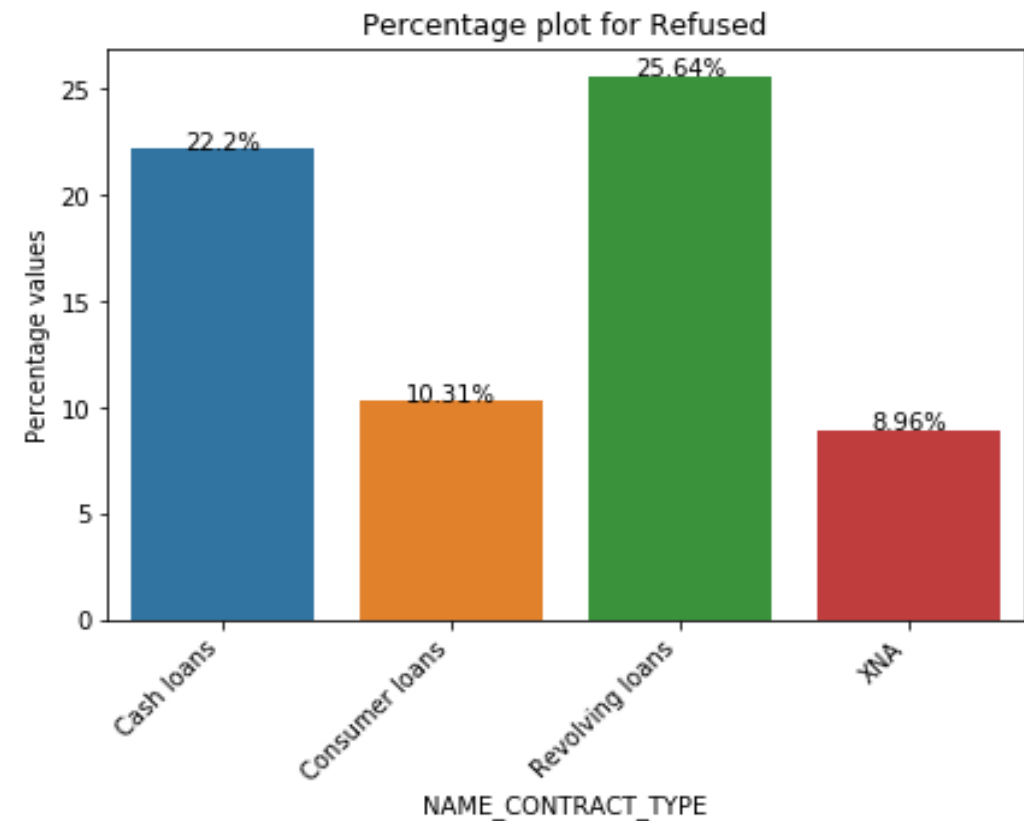
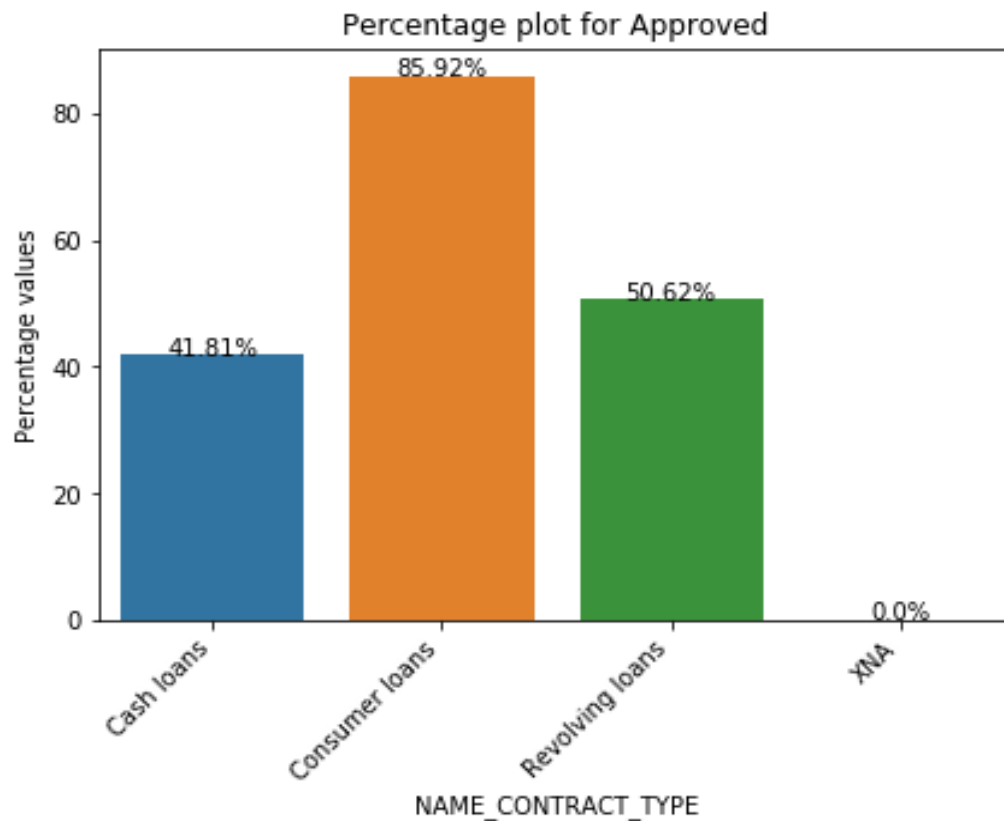
Boxplot



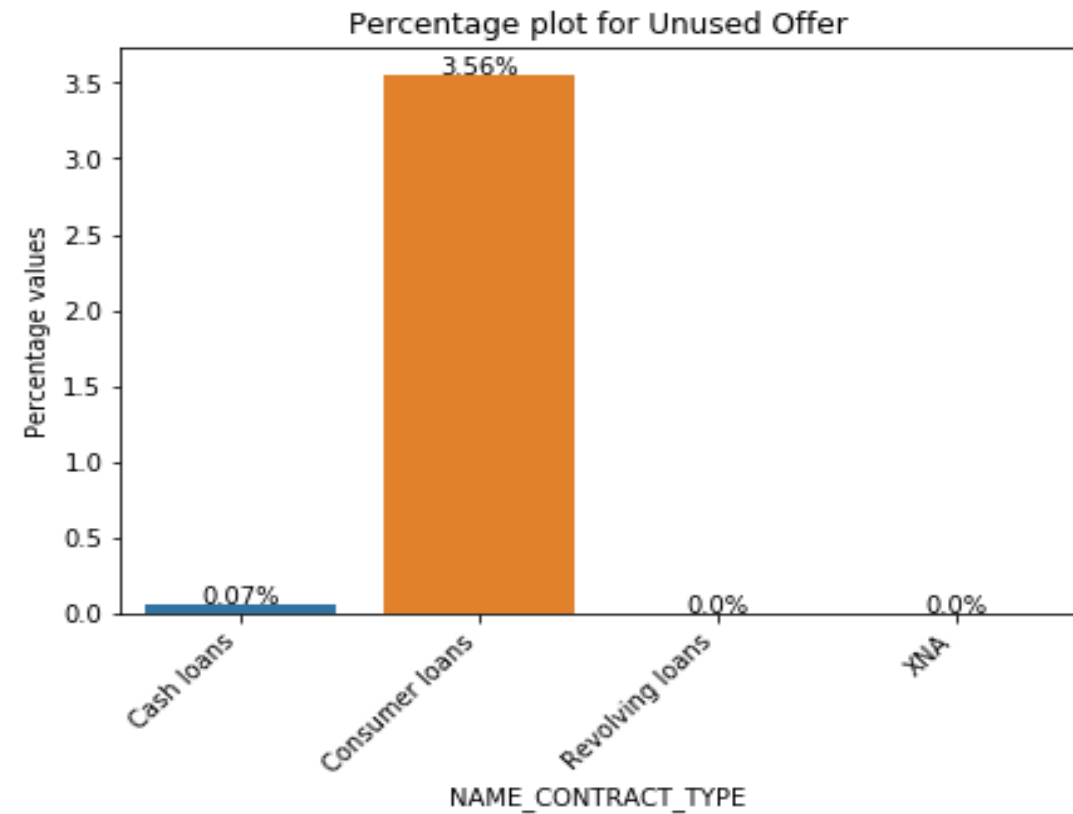
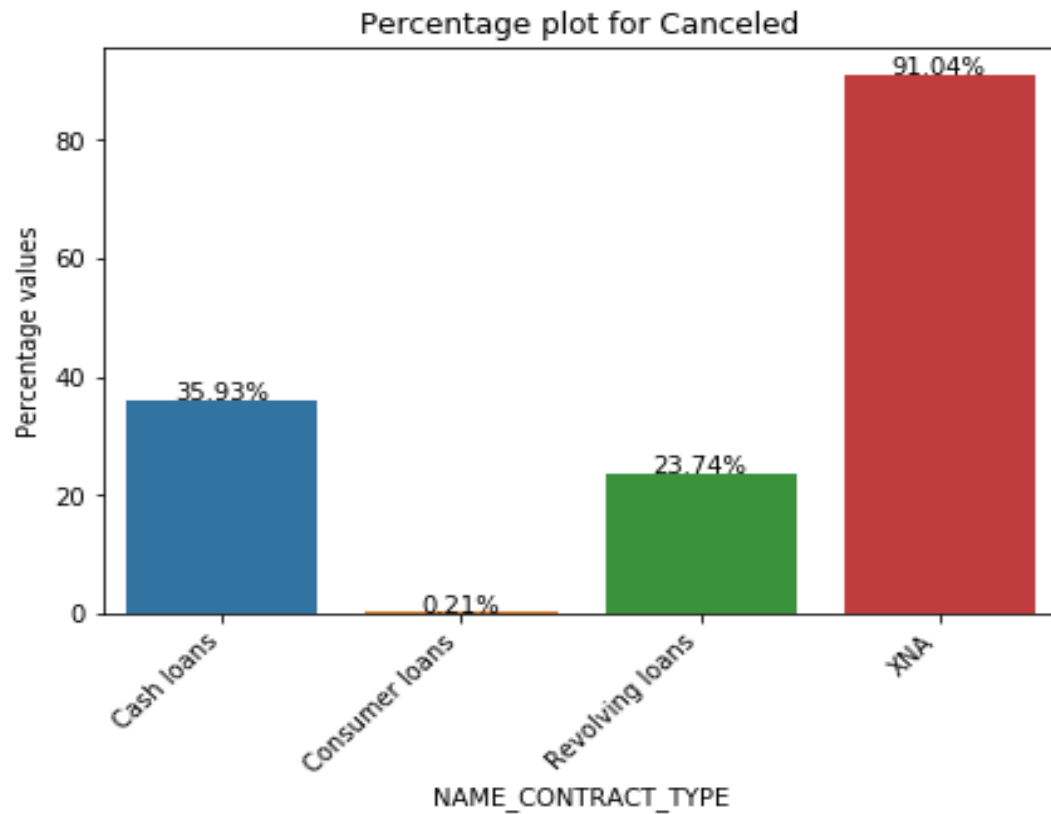
Insights

- ❑ *The AMT_APPLICATION follows poer law as is evident from bar chart.*
- ❑ *Minimum value is 0 and maximum value is 4,22,820. We can also estimate these values from the boxplot shown above.*
- ❑ *Box plot suggest that most of the values of AMT_APPLICATION are concentrated between 0 and 1,20,000*

Univariate Analysis on Categorical Columns (*NAME_CONTRACT_TYPE*)



Univariate Analysis on Categorical Columns (NAME_CONTRACT_TYPE)



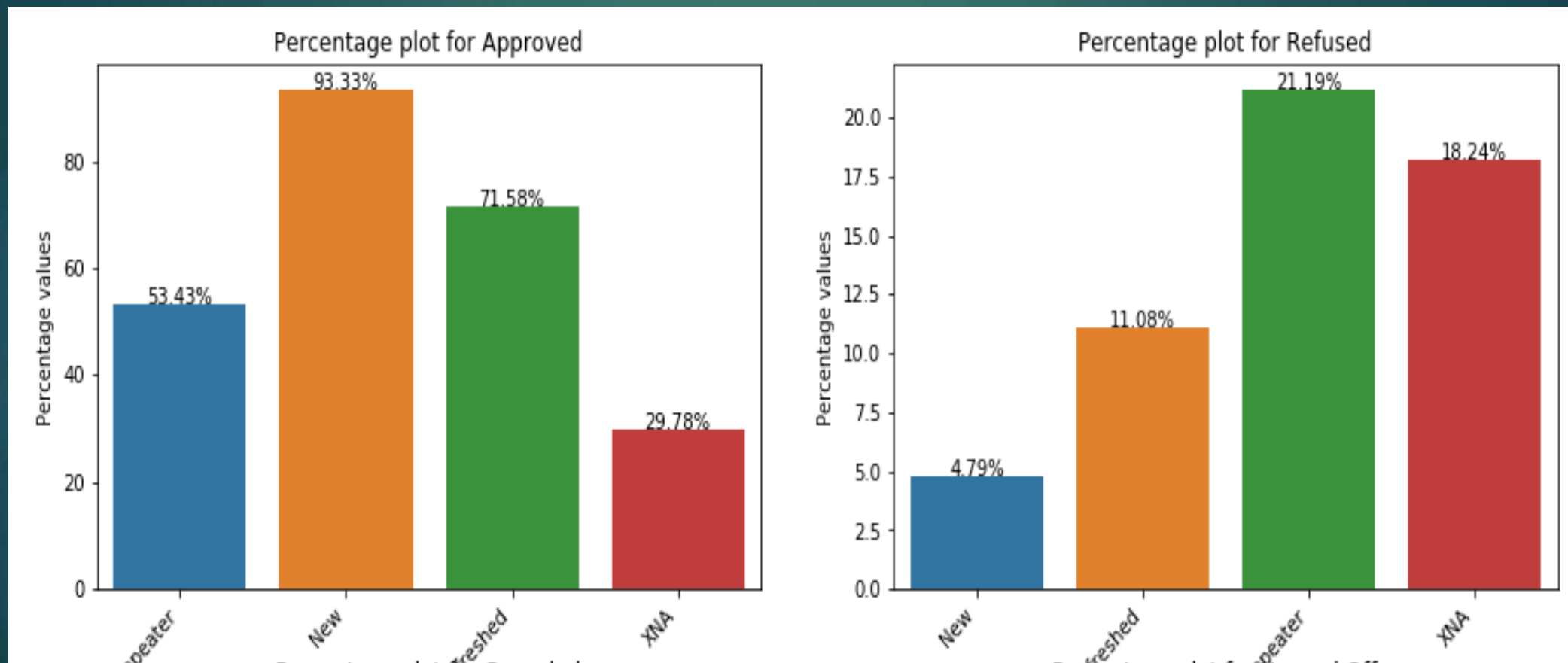
Observations

- ❑ 85.92% of Consumer loans are approved.
- ❑ 25.64% of Revolving loans are refused.
- ❑ 91.04% of XNA are canceled.
- ❑ 3.56% of consumer loans are unused offer

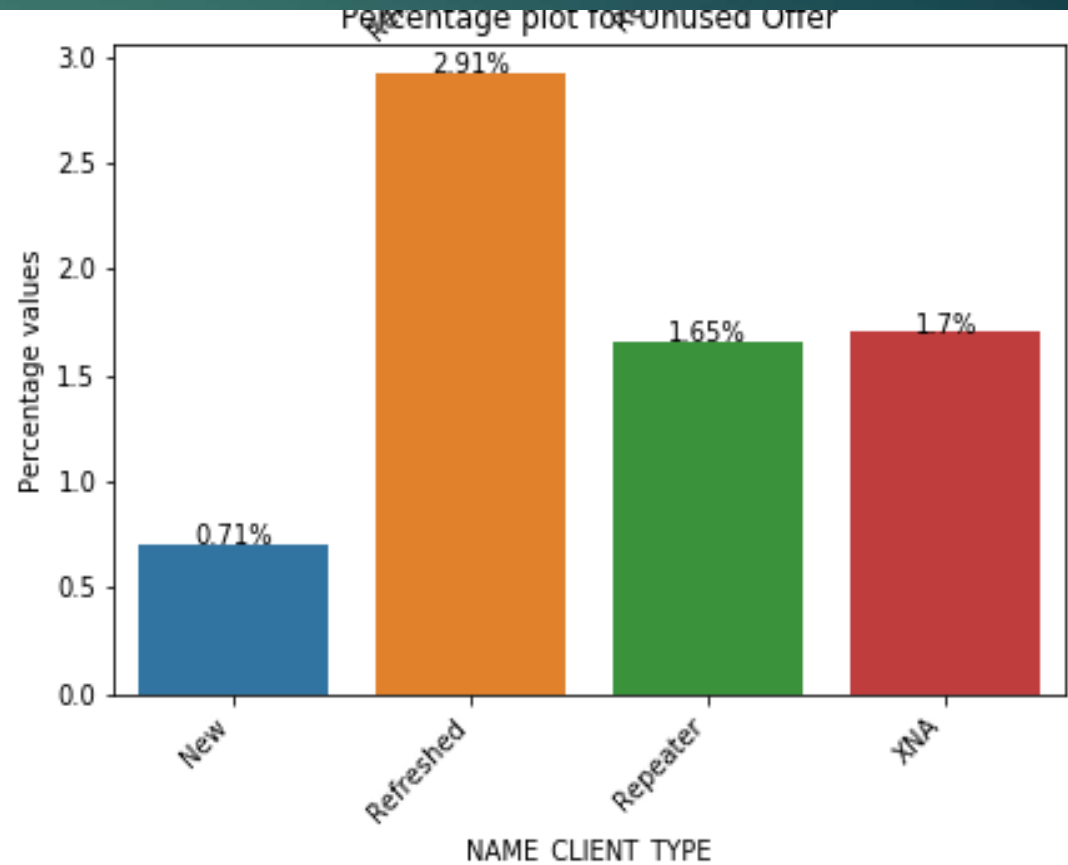
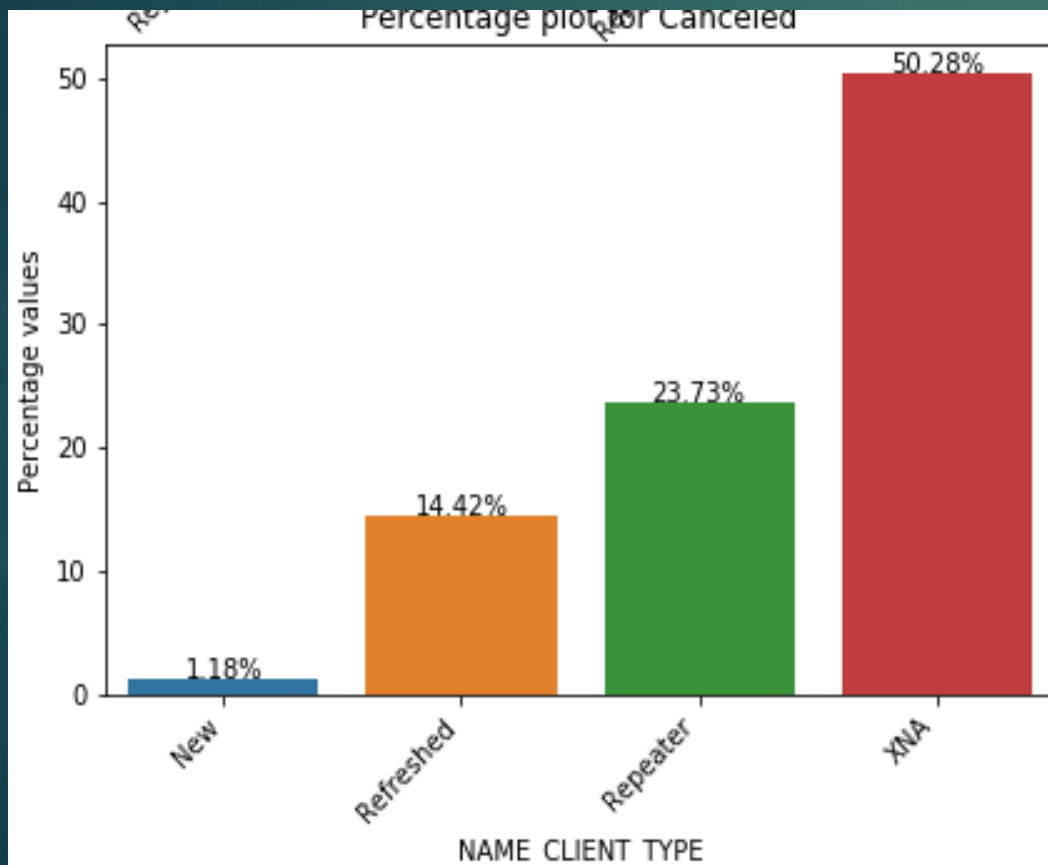
Conclusions

- ❑ 85% of the consumer loans get approved. This means that if a customer apply for a consumer loan then it has very high chance of getting approved.
- ❑ Almost 25% Revolving loans and 22% of cash loans are refused by the bank
- ❑ Cash loans are more canceled by client than revolving loans.
- ❑ Percentage of unused offers is very low for each category but Consumer loan offers are loans which are unused the most.

Univariate Analysis on Categorical Columns (NAME_CLIENT_TYPE)



Univariate Analysis on Categorical Columns (NAME_CLIENT_TYPE)



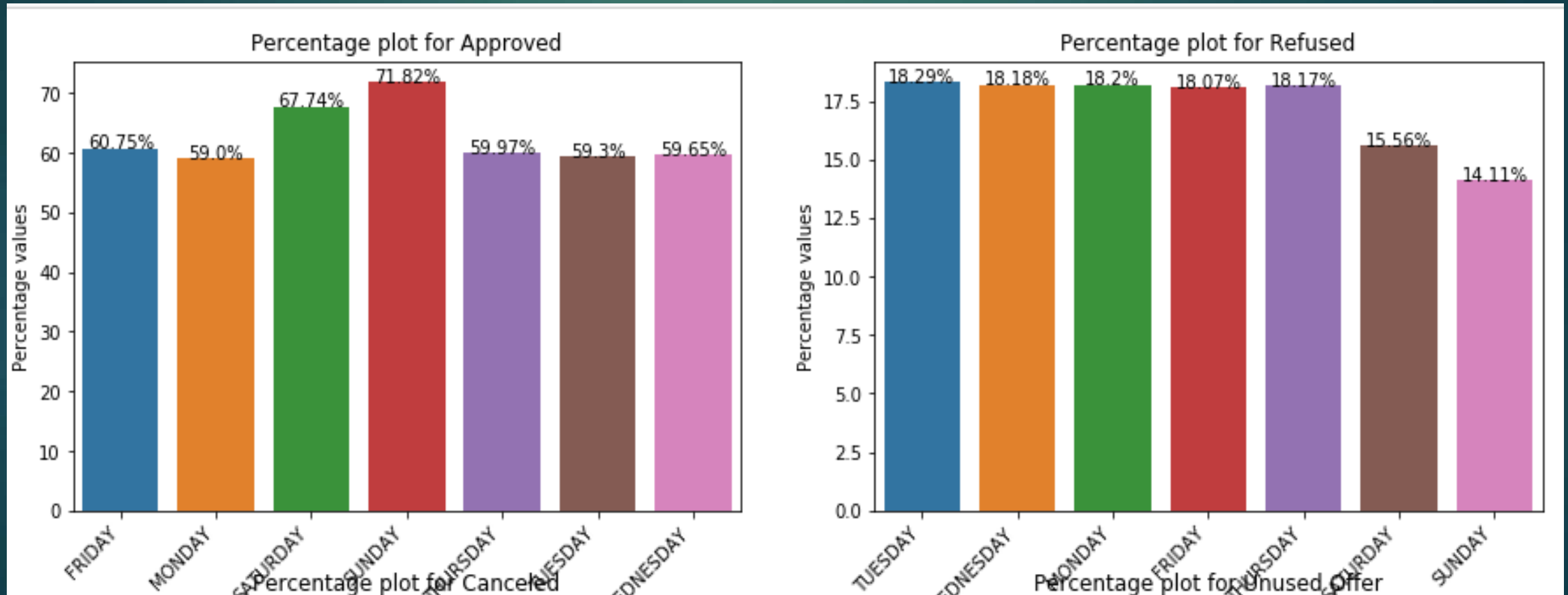
Observations

- ❑ 93.33% of loans applied by New client are approved whereas only 72% of refreshed client loans are approved. The number is worse for repeater clients.
- ❑ Repeater clients loans are refused the most. 21.19% loans applied by Repeater clients loans are refused. New clients loans are refused the least only 5%
- ❑ Repeater clients often cancel their loan at some stage. Almost 24% of them canceled the procedure. New clients rarely cancel the loan. Only 1% for new clients canceled their loans.
- ❑ Very rarely client loans are unused. The percentage is highest for refreshed clients but still the number is not very significant (only 3%)

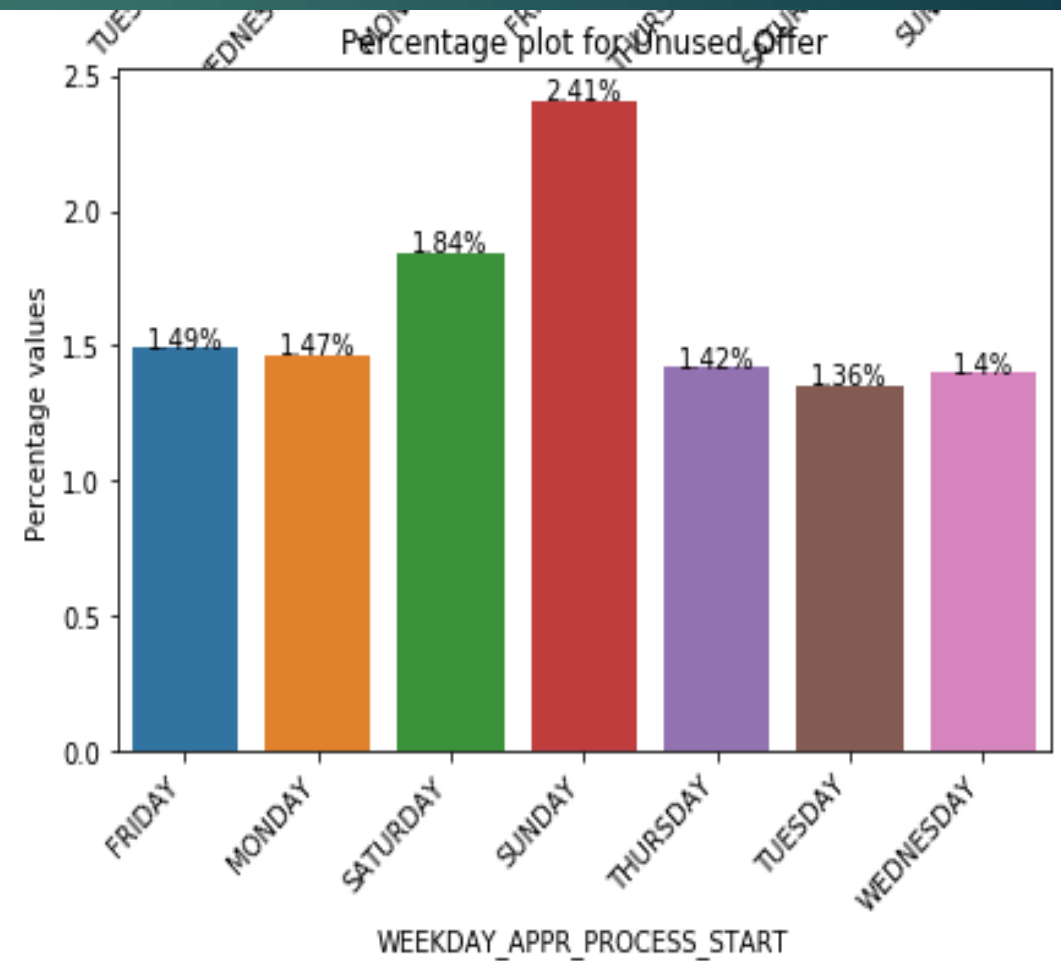
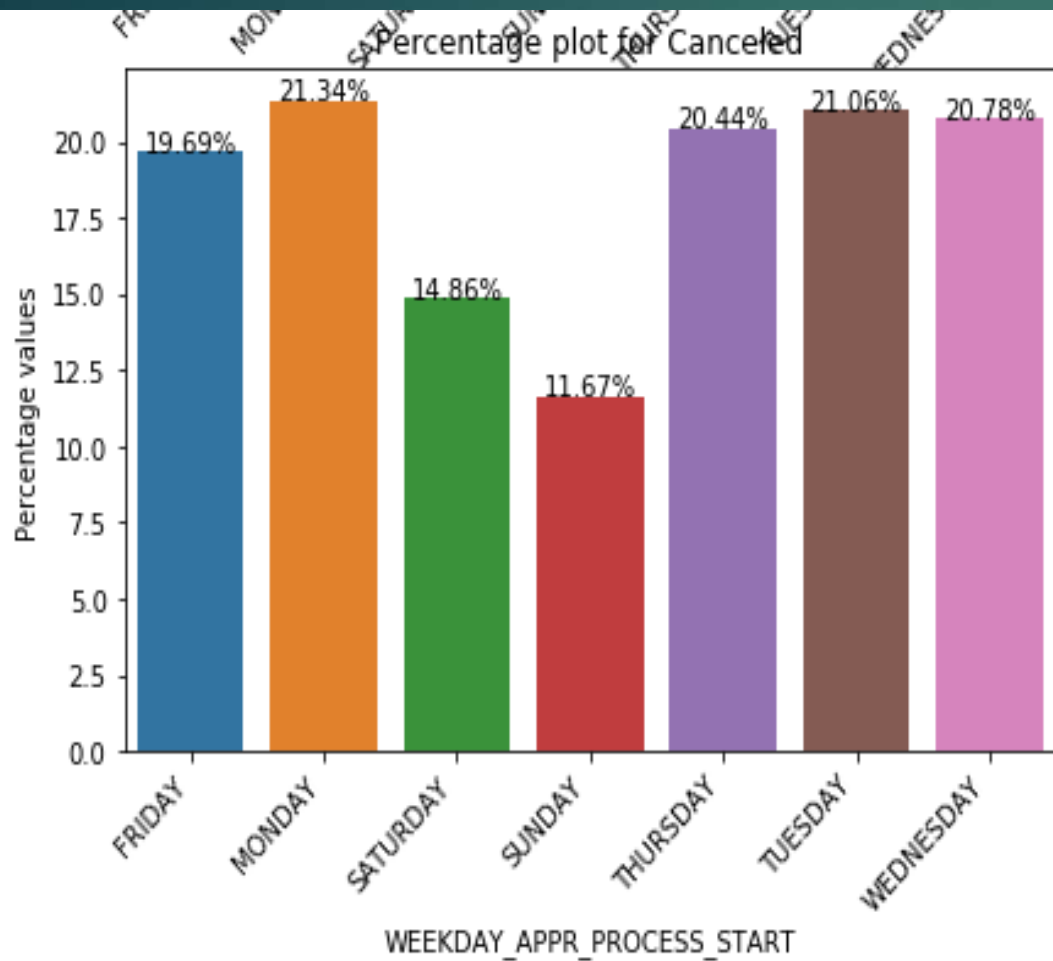
Conclusions

- ❑ New clients have very high chance around 93% of their loans getting approved.
- ❑ Repeater client has less chances around 50% of loan getting approved.
- ❑ Only 5% of the new clients loan request are refused whereas the number is around 22% for repeater client.
- ❑ Although repeater clients loans are less approved still around 24% of them canceled their loan at some point.
- ❑ Refreshed clients loans are more likely to be unused than any other type client

Univariate Analysis on Categorical Columns (WEEKDAY_APPR_PROCESS_START)



Univariate Analysis on Categorical Columns (WEEKDAY_APPR_PROCESS_START)



Observations

- ❑ 71.82% of loans applied on Sunday are approved.
- ❑ 18.29% of loans applied on Tuesday are refused.
- ❑ 21.34% of loans applied on Monday are canceled.
- ❑ 2.41% of loans applied on Sunday are unused offer.

Conclusions

- ❑ Loans which are applied on weekends(Saturday and Sunday) are approved more.Around 72% of them get approved.
- ❑ If a client has applied for a loan on weekdays there is a 18% chance that it is refused by the bank.
- ❑ Similar trend is observed for cancelled.There is around 20% chance that clients who applied for loans on weekdays will cancel it at some point.
- ❑ Around 2% of loans which are approved by bank on weekends are unused by client.The percentage is 1.5% for loans which were applied on weekdays

Bivariate Analysis (Categorical Columns)

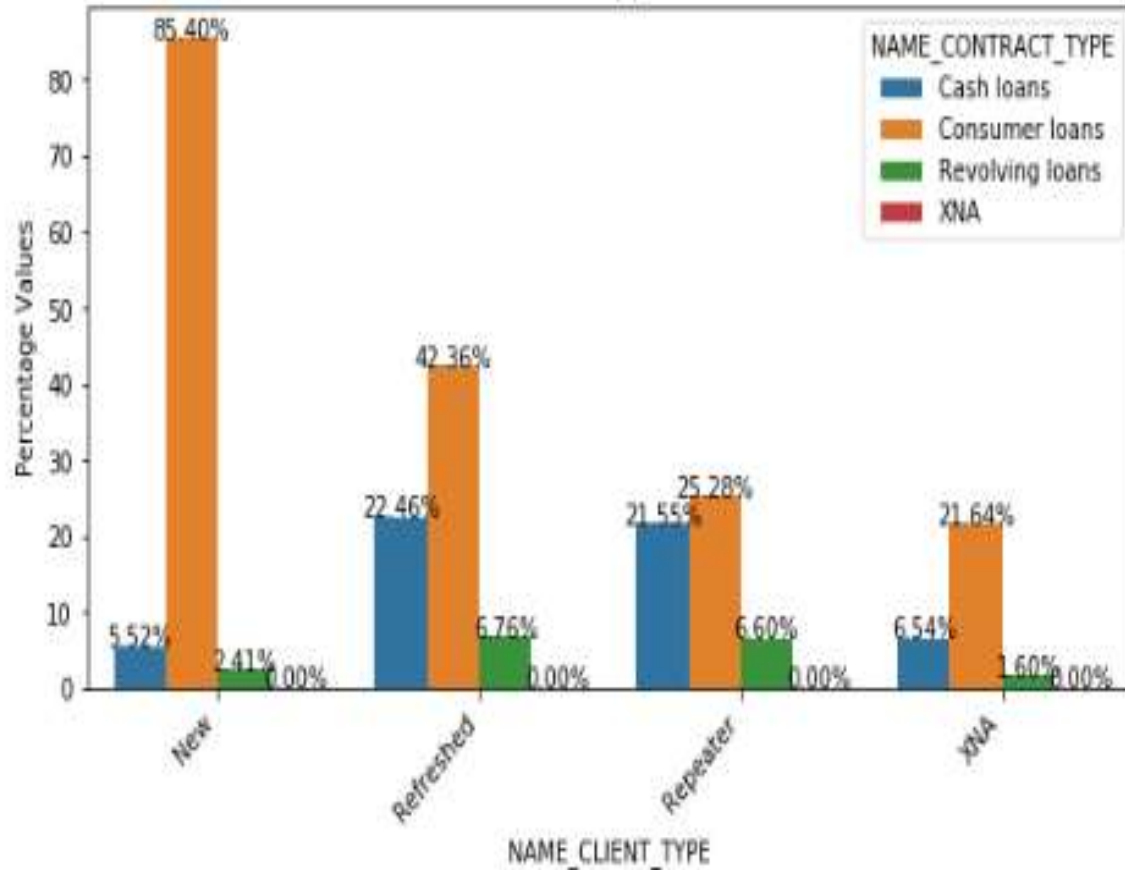
We have made some conclusions considering a single categorical variable. Now we will take it a step further and see how combination of two categorical variables affect the `CONTRACT_STATUS`. We will use the following example.

Example 1: Columns involved `NAME_CONTRACT_TYPE`, `NAME_CLIENT_TYPE`

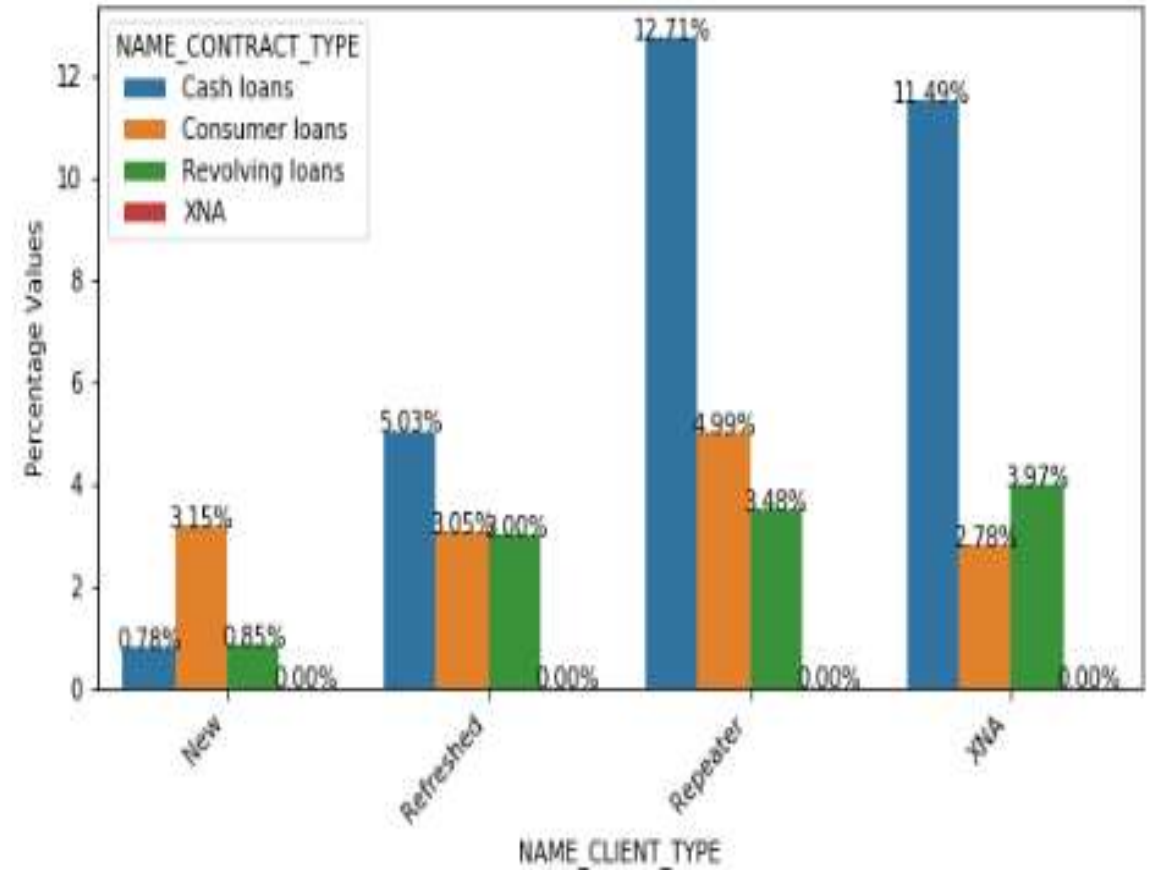
- ❑ For `NAME_CLIENT_TYPE` we observed that there is high chance of new client loan getting approved.
- ❑ For `NAME_CONTRACT_TYPE` we observed that there is high chance of consumer loan getting approved.
- ❑ We will now see if we observe the similar trend when we divide data based on client type.

Example 1: Columns involved NAME_CONTRACT_TYPE, NAME_CLIENT_TYPE

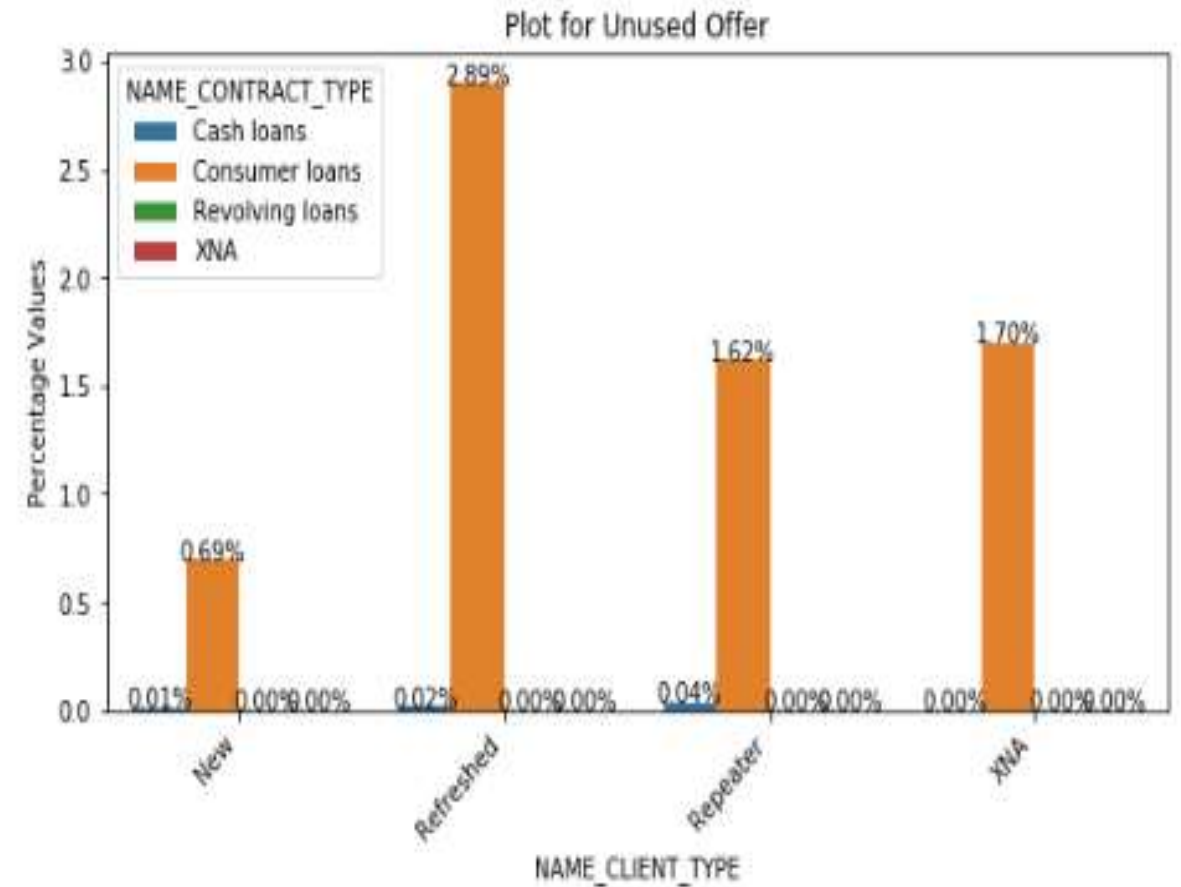
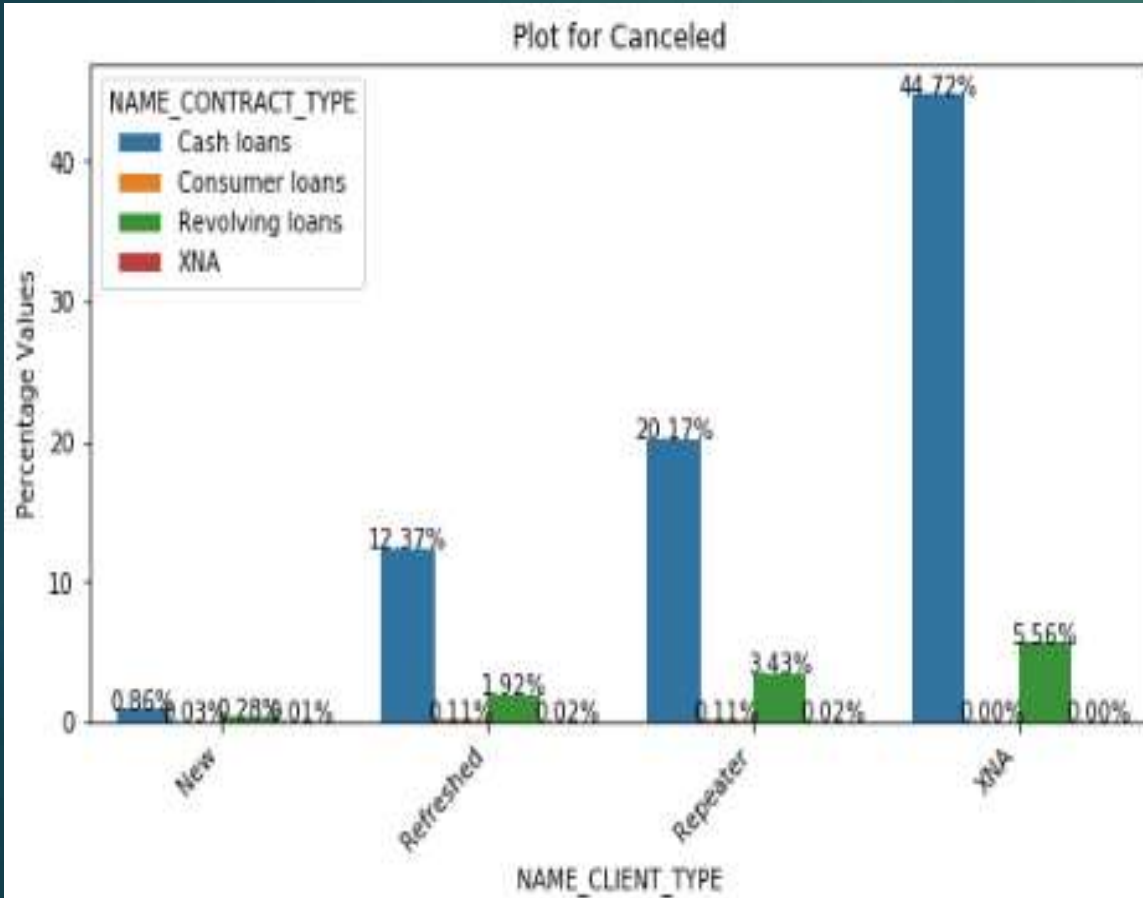
Plot for Approved



Plot for Refused



Example 1: Columns involved NAME_CONTRACT_TYPE, NAME_CLIENT_TYPE



Observations AND CONCLUSIONS

- ❑ Earlier we saw that new client have more chance of getting loans approved and Consumer loans have higher chance of getting approved
- ❑ 92% of new clients have approved loans.85% of total new clients got approved consumer loans.
- ❑ The chart still supports our understanding that client who apply for consumer loans have higher chance of getting loans approved.This is evident from the fact for different client type whose loans got approved we can observe they applied for consumer loans.
- ❑ In each client category,loans which are approved majority of them are consumer loans.
- ❑ Refreshed and repeater clients have considerable number of approved cash loans.The number is considerably higher than new clients.
- ❑ CASH loans by repeater clients are rejected the most.For new clients consumer loans are rejected most.
- ❑ Almost 20% of repeater clients have canceled their application of cash loans.The reason cannot be understood at this point but the number suggest some investigation is in order.

Correlation on Numeric Columns

List of columns which we deemed important for finding correlation are :

1. AMT_APPLICATION
2. AMT_CREDIT
3. AMT_ANNUITY

Correlation for Approved



Inference

- ❑ All the columns are highly correlated.
- ❑ Correlation of AMT_ANNUIITY with AMT_CREDIT is .77 whereas correlation with AMT_APPLICATION is .71
- ❑ Correlation of AMT_APPLICATION with AMT_CREDIT is highest and is equal to .83

Correlation REFUSED



Inference

- ▶ All the columns are highly correlated.
- ▶ Corelation of AMT_ANNUIITY with AMT_Credit is .73 whereas corelation with AMT_APPLICATION is .65
- ▶ Corelation of AMT_APPLICATION with AMT_CREDIT is highest and is equal to .88

Correlation for CANCELED



Inference

- ❑ All the columns are not as highly correlated as we saw in previous cases.
- ❑ Correlation of AMT_ANNUIITY with AMT_Credit is .69
- ❑ Correlation of AMT_ANNUIITY with AMT_APPLICATION is .6
- ❑ Correlation of AMT_APPLICATION with AMT_CREDIT is highest and is equal to .99

Correlation for *UNUSED OFFER*



Inference

- ▶ All the columns are highly correlated.
- ▶ Corelation of AMT_ANNUIITY with AMT_Credit is .94
- ▶ Corelation of AMT_ANNUIITY with AMT_APPLICATION is .94
- ▶ Corelation of AMT_APPLICATION with AMT_CREDIT is highest and is equal to 1. This is somewhat unexpected result.

Some Insights

- ❑ All columns are Highly correlated for all categories but Unused offer showed an unusual amount of correlation. We think this is because very less loan offers remain unused.
- ❑ AMT_CREDIT and AMT_APPLICATION have the highest correlation in all the four categories(ACCEPTED,REFUSED,UNUSED,CANCELLED)