EDA Case Study Presentation

Objective:

1. To find out different patterns in given dataset with right approach of cleaning and imputing missing irrelevant data.
2. To perform EDA on different understandable parameters and find out meaningful inferences.
3. Recognize pattern to avoid giving loans to applications, which can result to defaulter.
4. Recognize pattern to avoid not to giving loans to applications who are capable of repaying back.

Approach to clean the data:

1. Find out the percentage of missing values for all the columns in dataset. Given data set has 122 columns.
2. Remove columns, which has more than 50% of missing values. This reduced the column count to 81 columns.

Approach to impute the columns with very less missing data:

1. For numerical columns like, AMT_ANNUITY, AMT_GOOD_PRICE the missing values can be imputed with median so that it should not be impacted with the exceptionally higher values or outlines.
2. NAME_TYPE_SUITE is categorical value, and as per the understanding of the column, it can be imputed as 'Not Available' or 'Unaccompanied' as the most frequent one.
3. The observation variables (OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE) can be imputed with zero (0), as it can indicate that there was no such observation made for the given application.
4. Column OCCUPATION_TYPE has high missing values, and imputed with value like ‘Not Available’.
5. Narrowed down the columns to 24 understandable set and removed rows with more than 50% missing data.

Datatype check and deriving columns:

1. For the days related columns, removed the preceding hyphen/negative (-).
2. Derived the columns as AGE, EXP_IN_YEARS for DAYS_BIRTH, DAYS_EMPLOYED columns dividing by 365 from actual data.

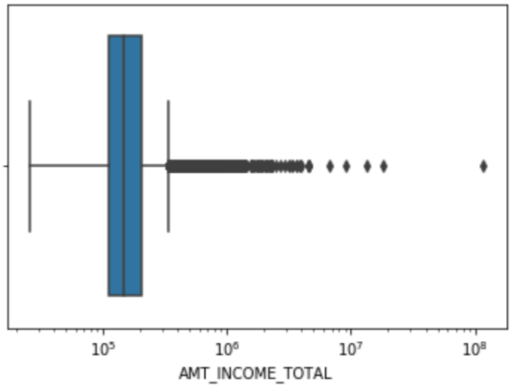
Handling outliers:

1. Identified a major chunk of data for numerical columns like AMT_INCOME_TOTAL, AMT_GOODS_PRICE, EXP_IN_YEARS.

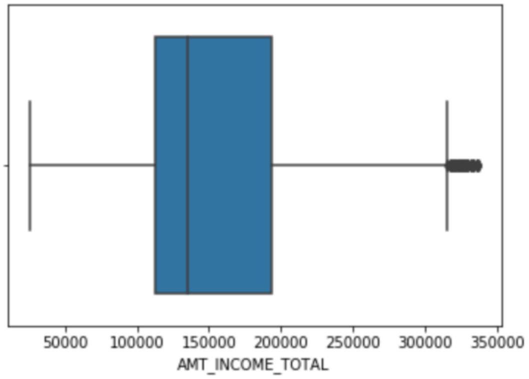
2. Removed the data, which are above 95% quantile for all the mentioned columns and resulted to sensible range.

For example, with the above approach the AMT_INCOME_TOTAL narrowed down to more meaningful range of around 120000 to 190000.

With Outliers (wider range in log scale)

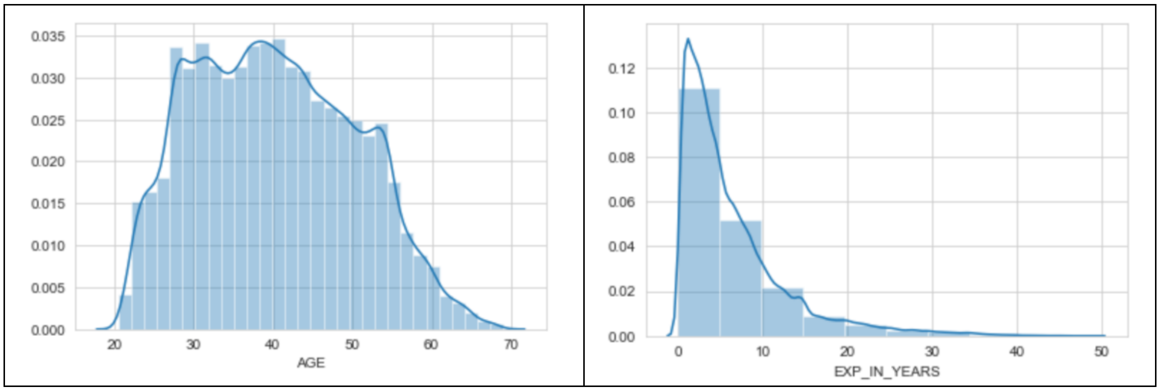


After removing 95% quantile data



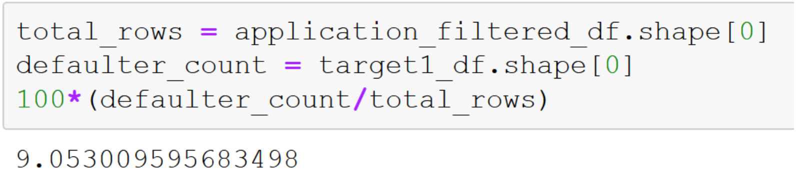
Plotting continuous variables:

1. Created bins and used histogram, density plot to plot meaningful continuous variables like AGE, EXP_IN_YEARS
2. The plot shows that the number of applicants mainly lies between age of 30 to 45, mostly termed as early mid-age and mid-age category. Majority is in between 38 to 40 years range.
3. The plot shows that the number of applicants are mainly in less than 10 years of experience, mostly early in career.



Reporting data imbalance:

We performed the data imbalance for Target=1 i.e. for defaulter applications and the calculation is captured below:

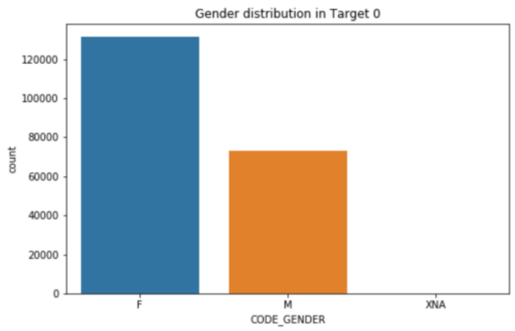
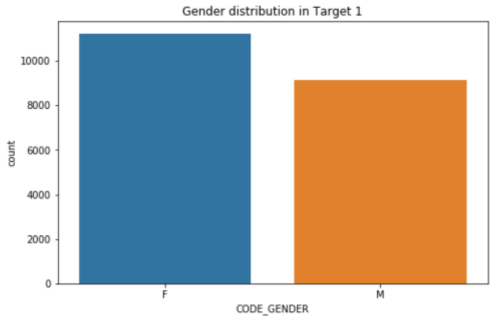
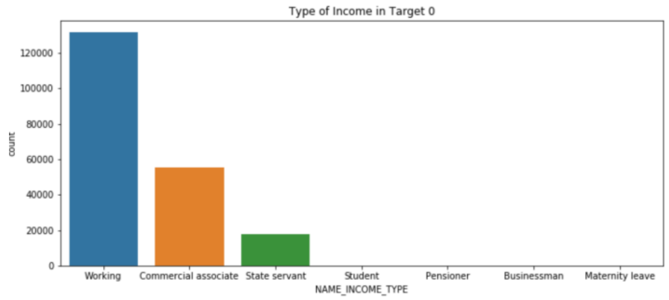
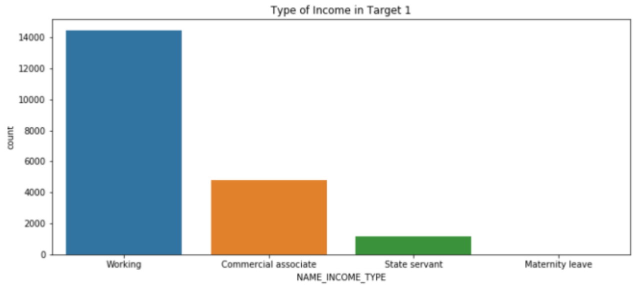


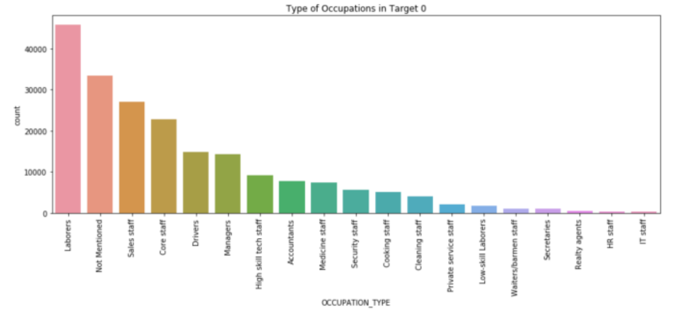
```
total_rows = application_filtered_df.shape[0]
defaulter_count = target1_df.shape[0]
100*(defaulter_count/total_rows)
```

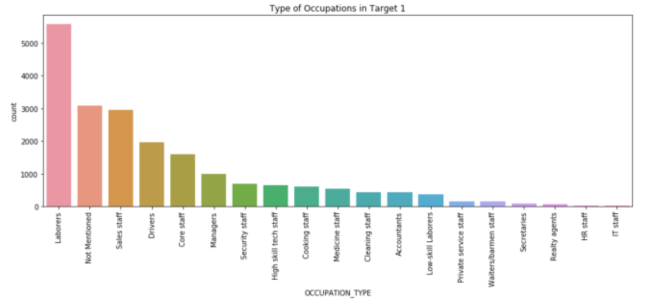
```
9.053009595683498
```

Performing univariate analysis:

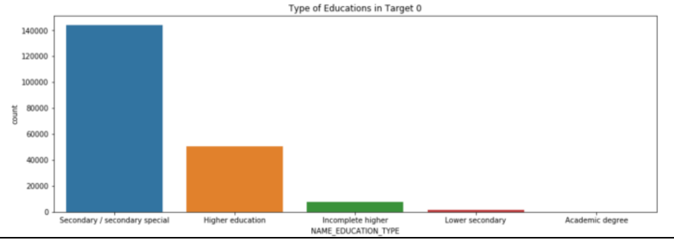
We segregated the master data to Target=0 and Target=1 data as asked and it was used to perform further analysis. The univariate analysis and the findings reported below:

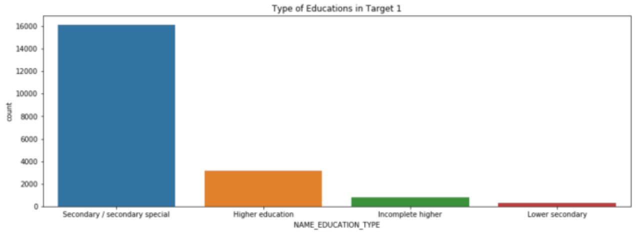
<table>
<tr><th colspan="2">Gender Distribution</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2">Distribution of Income Type</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2">Distribution of Occupations</td></tr>
</table>





Distribution of education levels





Key highlights:

1. The number of female applicants and defaulters is higher in comparison to male applicants.
2. Univariate Analysis for 'NAME_INCOME_TYPE' across Target1 and Target0 shows the 'Working' type have more number in both the cases, considering that the 'Working' applicants are more. It also shows that a major chunk is divided among 'Working', 'State Servant', 'Commercial associate' leaving the other NAME_INCOME_TYPE to negligible and also shows that 'State Servant' has less payment difficulties.
3. Laborers are the highest in loan application as well as defaulter category. Major applications do not have the occupation details mentioned. Drivers are more prone to be defaulters as well.
4. It clearly shows that secondary and secondary special category are more prone to apply for loans as well as being defaulters. The 'academic degree' has less number of applications, although with negligible defaulters.
5. The family status pattern is almost same for defaulters and non-defaulters. Married people are more prone to apply loans in compare to other categories, followed by single category.

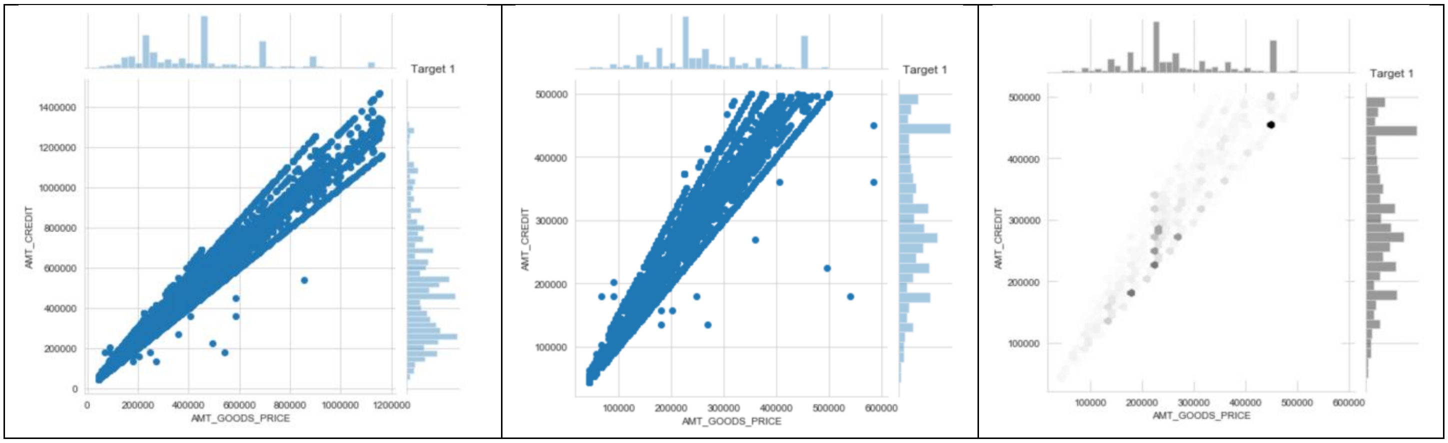
Performing bivariate analysis:

a. Credit Amount with Good Price:

Intention:

AMT_CREDIT, AMT_GOODS_PRICE should be aligned to avail the loan. The Loan approved and credited amount should not be beyond range of the good price. The target1 data will help to identify that range where it should not be a defaulter if these two aligned and needs to check deeply. For target0, we need to check the range where this two are beyond the range and should be treated as defaulter.

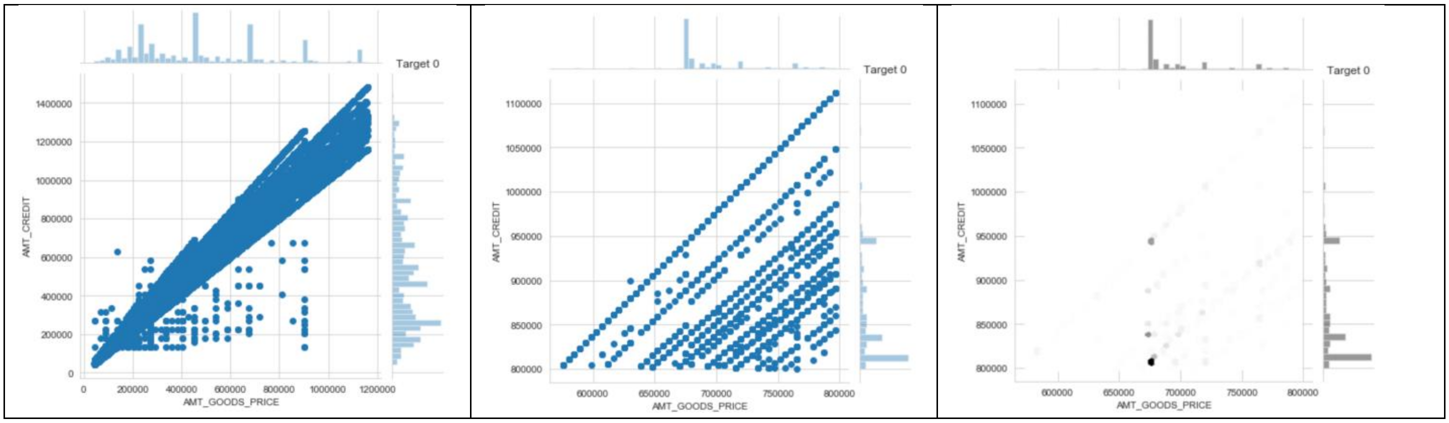
Target = 1



Findings and Highlights:

1. The first plot is for the entire Target=1 data which show exponential behavior of AMT_CREDIT with respect to AMT_GOODS_PRICE. We further narrowed down to credit less than 500000 to find out some useful pattern here. The plots are shown here from left to right.
2. For Target=1, both the parameters are increasing exponentially, but we can see some of the outlined scatters where the good price is low, but the credit amount is high, and on the other hand, some case good price is high, but credit amount is low.
3. In the rightmost sample above for Target=1, there is a dark dot between 400000 and 500000, it shows both the parameters are aligned, but still it is a defaulter, so we need to deep dive on other parameters to find out the actual reason, or it should be marked as non-defaulter or not.

Target = 0



Findings and Highlights:

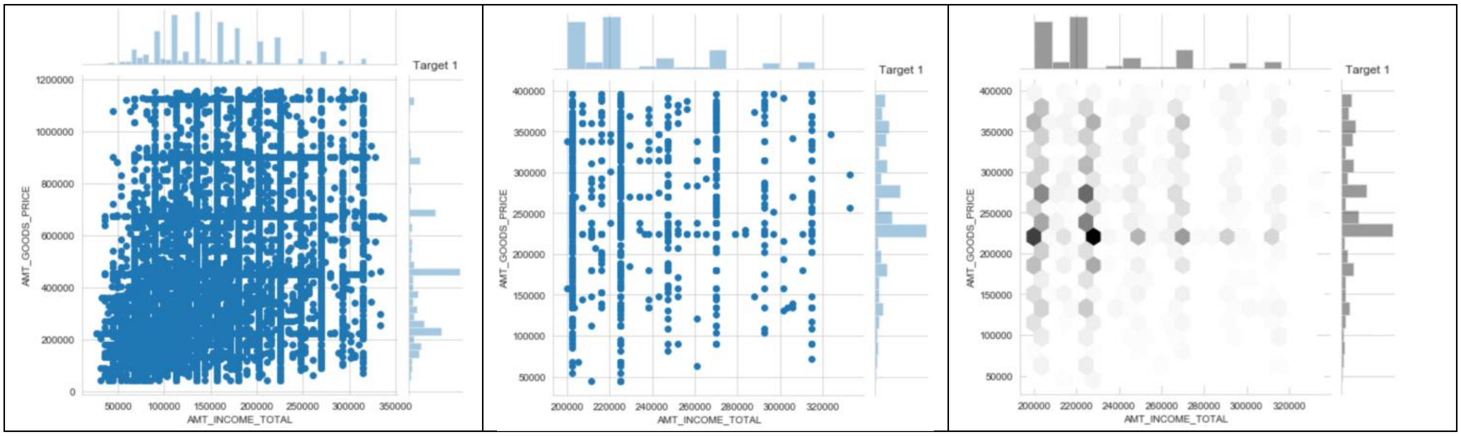
1. Initially the entire Target=0 dataset was plotted. Then for a better analysis, we narrowed down to a region where the AMT_GOOD_PRICE is less than 800000, but the AMT_CREDIT is more than 800000. The plots are shown starting from left to right.
2. In the first left plot, there are some data points where credited amount is more than the actual good price. This should be avoided and rechecked.
3. For Target=0, here the point to note is, in the dark region the good price and the credit amount are not aligned and it is not under defaulter as well. Therefore, the credit amount is way more high than the actual required loan amount.

b. Good Price and Income Total:

Intention:

For any loan application, there should be a defined range ratio for the applicant's income to the price of the good being availed for. If the price of the good's price is beyond range, it should be checked thoroughly. Similarly, it is within defined range, it is more likely to approve to get the loan.

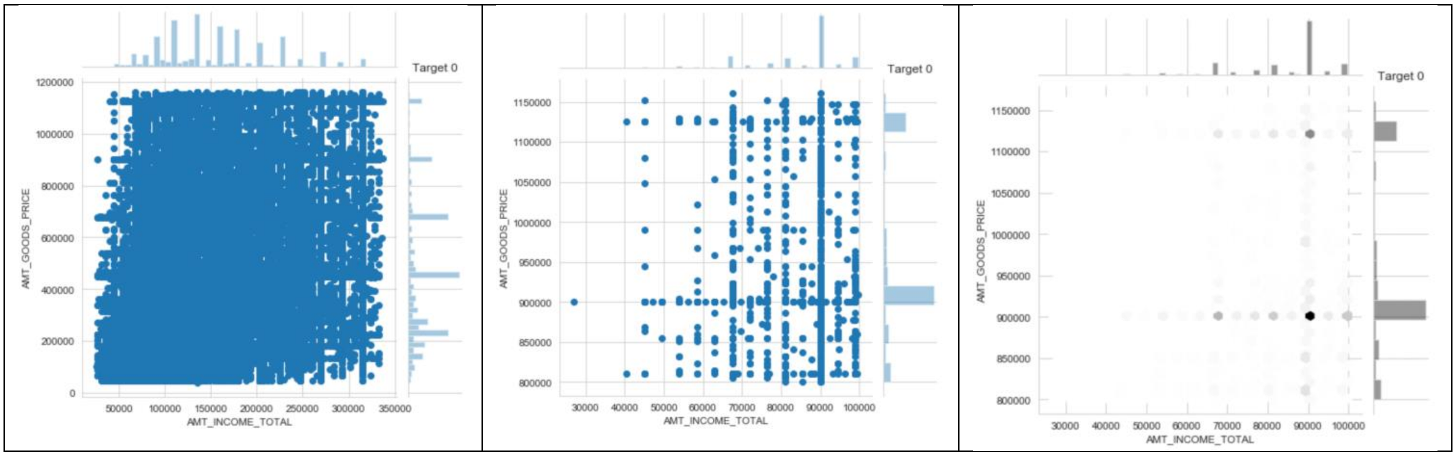
Target = 1



Findings and Highlights:

1. Initially entire dataset for Target=0 was plotted. It shows that the more density is mainly where the good price is near to double of the total income.
2. We further narrowed down to a region where income is high but good price is low, still it comes under defaulter. In this sample, we took a region where AMT_INCOME_TOTAL is higher than 200000 and AMT_GOODS_PRICE is lesser than 400000.
3. For the above sample on Target=1, there is an intensity where the good's price is almost near to the income value and resulted to defaulter. We need to do more check on this why it has become the defaulter in this pattern.

Target = 0



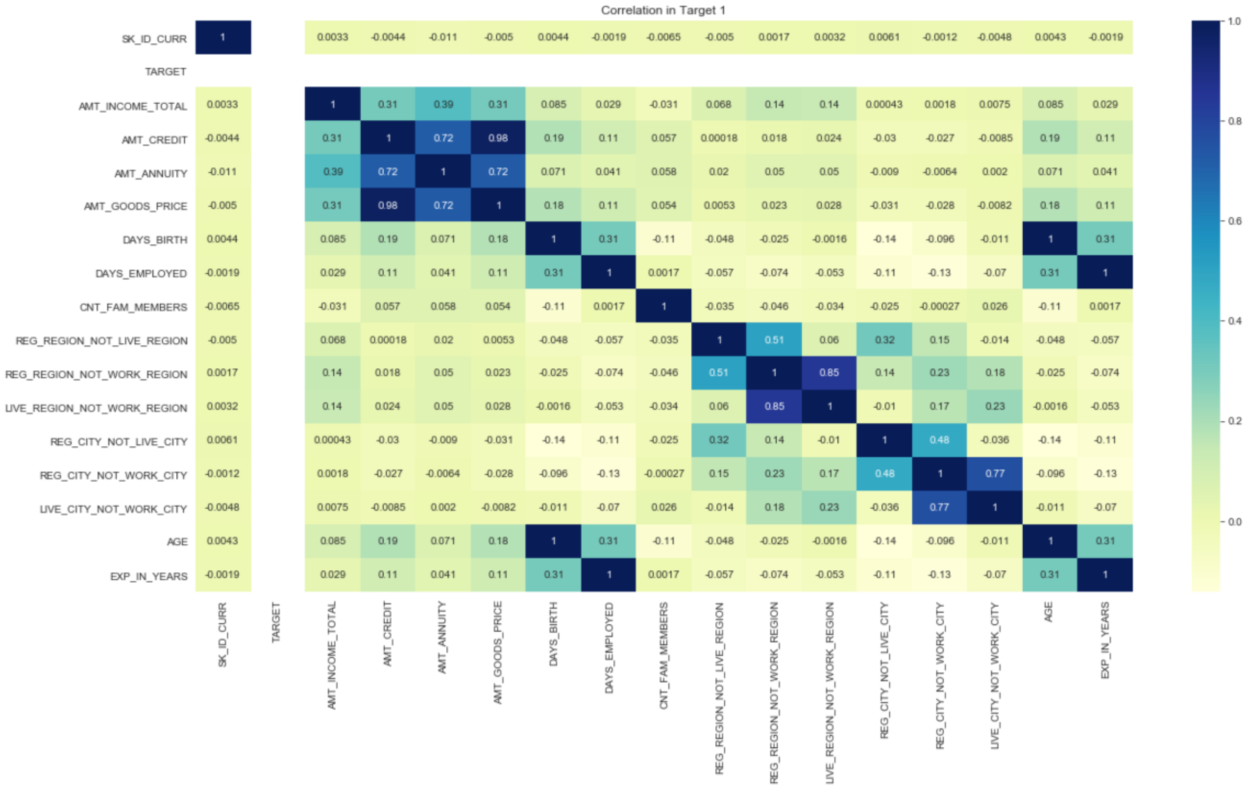
Findings and Highlights:

1. Initially entire dataset for Target=0 was plotted. It shows that for the income of 100000 and 150000 range, there is high density in the plot.
2. We further narrowed down to a range where the AMT_INCOME_TOTAL is less than 100000 but the AMT_GOOD_PRICE is more than 800000. It will show the distribution where the good price can reach to beyond affordable limit of the income.
3. From the above sample of rightmost plot on Target=0, the dark point is where the good's price is almost 10 times of the income total and it is not under defaulter as well. We need to check more on these kind of transactions.

Correlation in the defaulter dataset:

Top correlated attributes in Target = 1

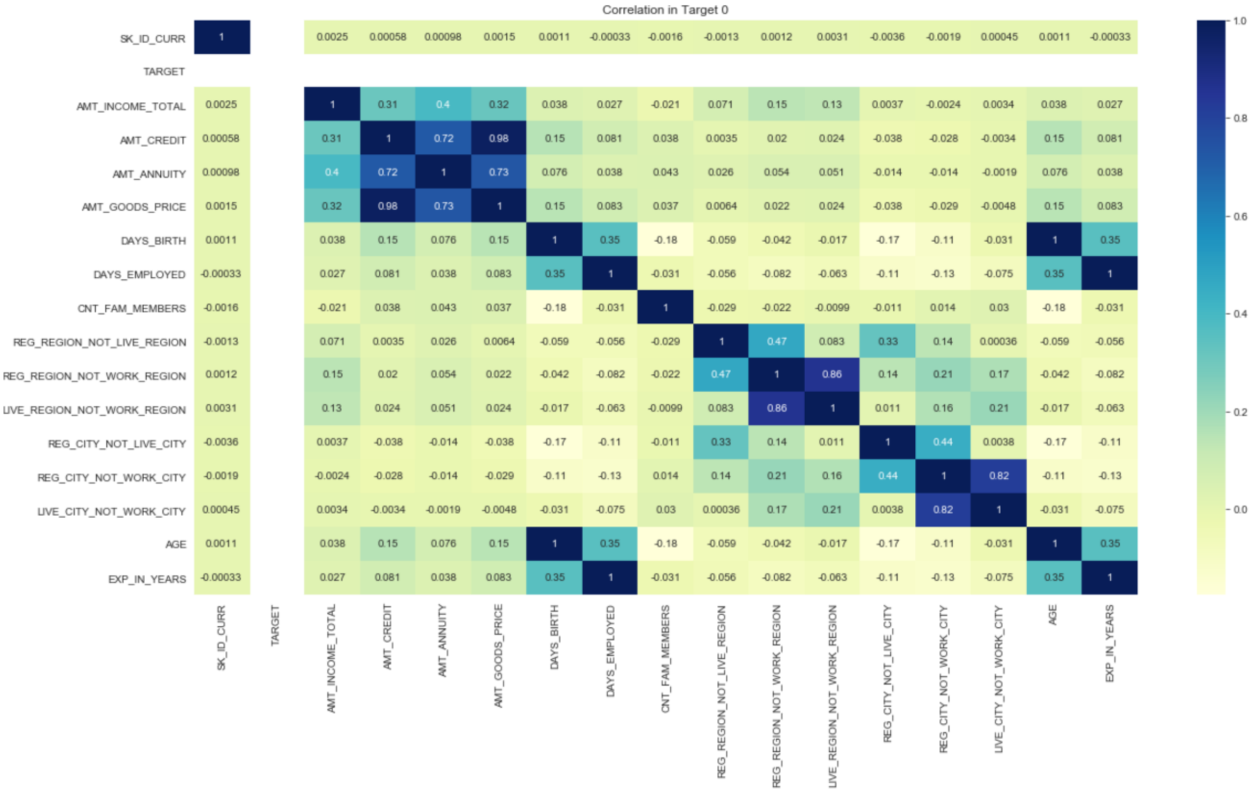
1. AMT_GOOD_PRICE and AMT_CREDIT - 98%
2. LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION - 85%
3. LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY - 77%
4. AMT_ANNUITY and AMT_CREDIT, AMT_ANNUITY and AMT_GOOD_PRICE - 72%



Correlation in the non-defaulter dataset:

Top correlated attributes in Target = 0

1. AMT_GOOD_PRICE and AMT_CREDIT - 98%
2. LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION - 86%
3. LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY - 82%
4. AMT_ANNUITY and AMT_GOOD_PRICE - 73%



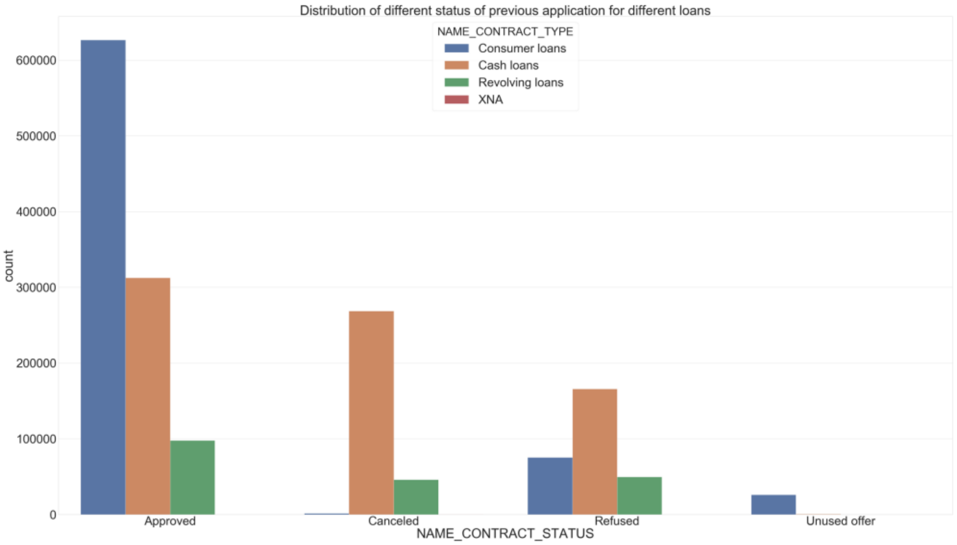
Analysis on previous application data:

Basic Procedure:

1. Import the CSV file in data and have the general checks performed.
2. Found of the percentage of missing values in all the columns.
3. Remove columns, which all have more than 30% data. Actual dataset has 37 columns and it reduced 26 columns
4. Narrowed down to almost 15 meaningful columns by dropping some unwanted/non-understandable columns.

Univariate and Bivariate Analysis on different attributes:

a. Contract Status and Contract Type:

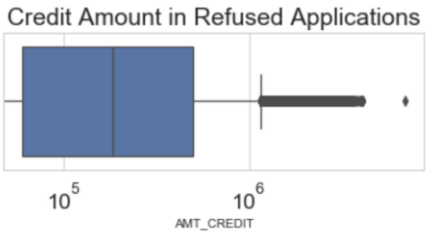


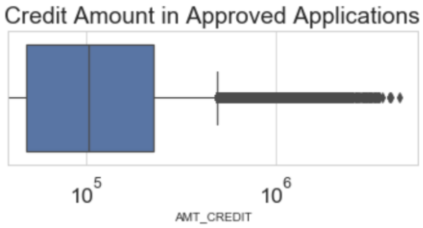
Highlights:

1. Consumer Loans are the most approved loans.
2. Cash loans are the most refused, but more of them are canceled.

b. Credit Amount:

We further narrowed down to two dataset of approved and refused data. Then found out below pattern for credit amount (AMT_CREDIT).



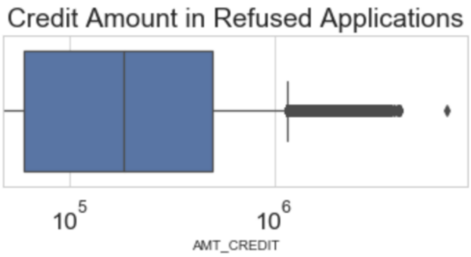


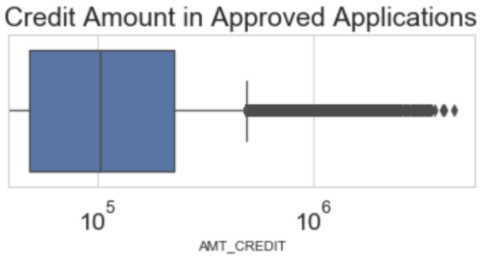
Highlights:

The range of credit amount is higher in refused applications compared to the approved one and the median is high for the same.

c. Application Amount:

We have tried to find out similar pattern for two dataset for application amount (AMT_APPLICATION).



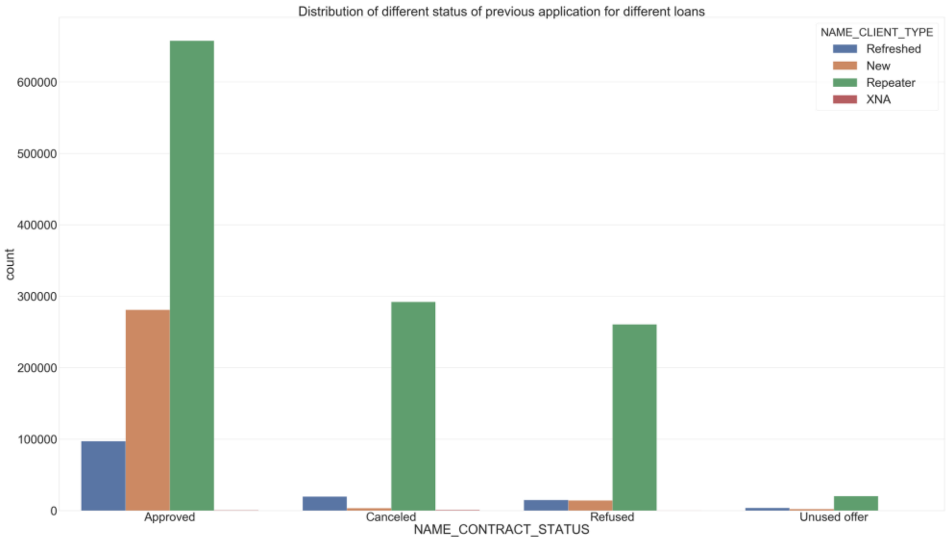


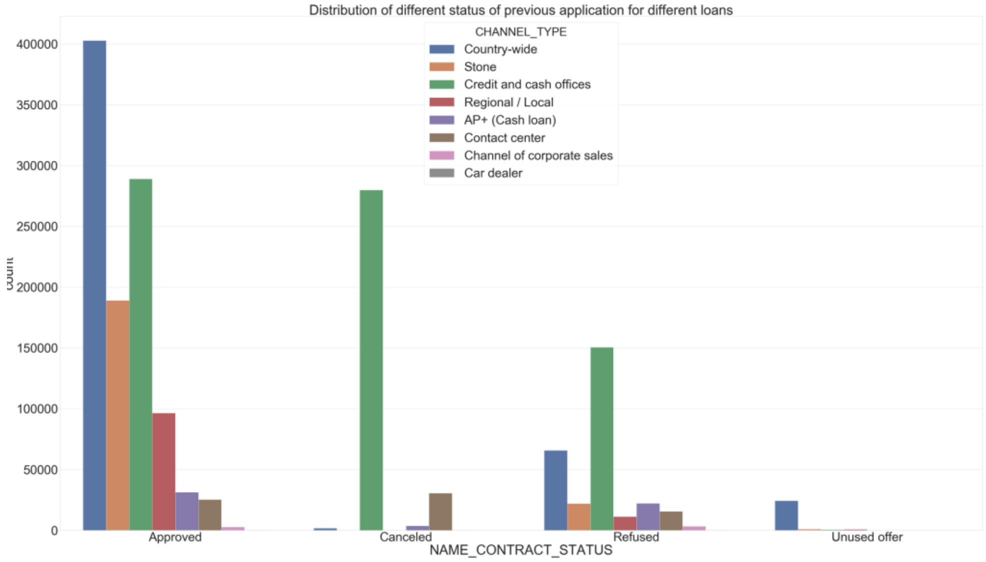
Highlights:

The application amount range is higher in refused applications and the median is high in comparison to approved applications.

d. Client Type and Channel Type

We have plotted different type of client and channels over different status available in the master dataset. The patterns shows as follows,





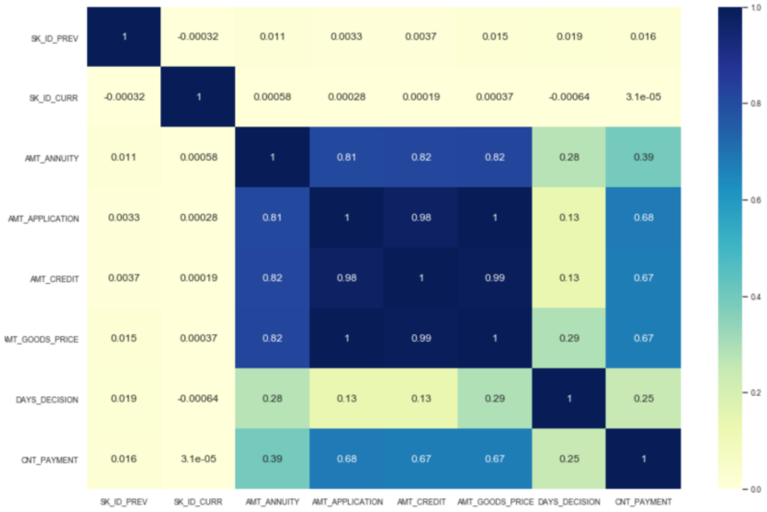
Highlights:

1. There are more number of repeater in trends, compared to 'New'. But New applications has less rejections in compare to repeated applications.
2. 'Country-wide' has more number of approved applications, but 'Credit-Cash offices' are more prone to be refused or canceled.

Correlation between different attributes in previous application:

Top correlation in previous application data:

1. AMT_GOOD_PRICE and AMT_CREDIT - 99%
2. AMT_APPLICATION and AMT_CREDIT - 98%
3. AMT_CREDIT and AMT_ANNUITY - 82%
4. AMT_ANNUITY and AMT_GOOD_PRICE - 82%



Final Words:

1. The given dataset has good number of null values, which are neglected or imputed as necessary.
2. The major number of applications are from people in salary range of 120000 to 190000.
3. The major applicants are female in gender, married in family status and working sector. Most of the working category belongs to laborers.
4. The major applicants are in age 38 to 42 among the applicants and work experience of major applicants are 0-5 years.
5. There is data imbalance of around 9.5% for defaulters over the total number of applicants.
6. There are applications where the person's income is in line with the credit amount application, but still in defaulter. Similarly, there are applications where the asked credit amount is almost 10 times of income and not in defaulter.
7. There are data where the asked credit amount is much higher than the good's price.
8. The correlation metrics is almost similar to defaulter and non-defaulters dataset. The credit amount, good's price and annuity are highly correlated.
9. Previous application shows mainly three kind of loans- consumer loan, cash loan and revolving loans.
10. Higher credit amount and higher application amount are more prone to refuse.
11. Repeated applications have higher rate of approval as well as refusal.
12. New applications has lower rejection rate compared to repetitive applications.
13. Credit and cash offices channel type has highest rate of refusal in previous data.
14. Credit amount, application amount and annuity amount has higher correlation in previous data as well.

Future Scope:

We can perform a further merge of these dataset based on the current application id which are present in previous dataset. We can figure out those who all were approved and refused. We can find the payment pattern of the previously approved applications and deep dive on the refusal cases to find more informative and meaningful patterns.