Bank Loan Case Study

By Rahul Inchal

Project Description

In this project, we have performed Exploratory Data Analysis (EDA) on Bank Loan Dataset. This case study aims to give you an idea of applying EDA in a real business scenario. The loan providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming defaulters.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

Suppose we work for a consumer finance company which specialises in lending various types of loans to urban customers. You must use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company:

1. **Approved**: The company has approved loan application
2. **Cancelled**: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused**: The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused Offer**: Loan has been cancelled by the client but on different stages of the process. In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

From this project, we must discover some insights such as-

- Which type of applicants are likely to pay the loan amount and which type of applicants are taking advantage of the loan amount?
- Identifying such defaulters and taking strict actions like denying the loan, reducing the amount of loan, lending (too risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.
- Presenting approach of removing non-useful columns, and null values, identifying outliers and data imbalance in the data, explaining the results of univariate, segmented univariate, bivariate analysis, etc. in business terms, drawing out top 10 correlation for the client with payment difficulties and all other cases.
- Discovering such results by performing EDA to help the company to decide whether to approve or refuse the loan application.

Tech-Stack used:

1. **MS Excel:** It is useful for understanding what the dataset exactly means, also useful in analysis work when the data is small, and for understanding the columns' description of data.
2. **Jupyter notebook:** Jupyter notebook is an interactive environment for arranging notebooks, code, and data using many programming languages. I have used it for whole analysis work using the python language.
3. **Python:** Python is a simple and easy programming language used in web development and machine learning applications, etc.
4. **MS word:** It is used for making the report.

Approach:

1. Used MS Excel to understand what the dataset is about and its columns' meaning, which is important.
2. Then imported the required python libraries (Pandas, Numpy, Seaborn, etc.) into Jupyter notebook.
3. After libraries, datasets (Application_Data and Previous_Application) are imported.
4. Identified missing values and removed columns which are not useful for analysis work.
5. Identified outliers of different columns and their relations.
6. Identified data imbalance and its ratio and presented using the graphs.
7. Performed univariate, segmented univariate, and bivariate analysis of data.
8. Found out the top 10 correlations with respect to TARGET variables.
9. Lastly presented all the senseful data (analysis work) in the form of different graphs and charts.

Data Understanding:

1. **`application_data.csv`** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. **`previous_application.csv`** 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.
3. **`columns_descrption.csv`** is data dictionary which describes the meaning of the variables.

Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

1. There are two datasets namely 'Applicaton_Data' and 'Previous_Application'.
2. Initially calculated the null value percentage of each column and found out the columns containing null values above 40% in Application data are 64 columns and in Previous Application are 11 columns.

In Application_Data -

3. In application_data, we found 'AMT_GOODS_PRICE' as a useful column in data analysis work that's why using the median () operation we imputed the missing values in the 'AMT_GOODS_PRICE' columns.
4. Similarly, the missing values in the column 'AMT_ANNUITY' is imputed using the same median () operation.
5. In median imputation, the missing values are replaced with the median value of the entire feature column.
6. The columns like 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_DOCUMENT_1' and others which are not useful for analysis work are dropped off from data frame.

In Previous_Application -

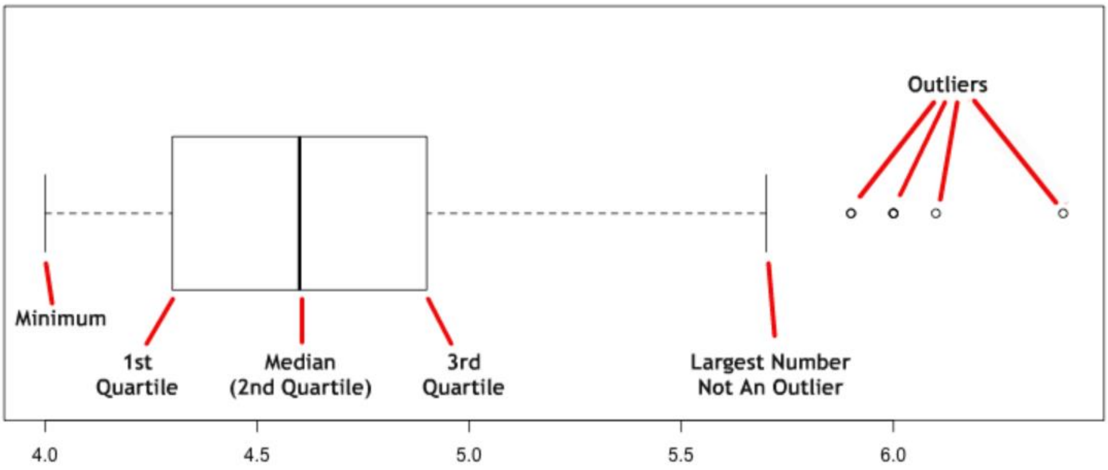
7. Imputation in Previous application.
8. Converting negative days to positive as days can't be negative.
9. Days in Columns 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH' are converted into positive values.
10. Imputing Null Values/Missing values in Previous Application
    1. After checking the null values percentage, we discovered columns like AMT_ANNUITY, AMT_GOODS_PRICE, AMT_DOWN_PAYMENT, and CNT_PAYMENT, have some null values/missing values.
    2. For imputing null values, in column 'AMT_ANNUITY' nulls are filled using the median operation.
    3. In 'AMT_GOODS_PRICE' and 'AMT_CREDIT', we have used mode ().
    4. And imputation in 'CNT_PAYMENT' are done by filling null by '0'.

- **<u>Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier.</u>**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. The outliers in this analysis work are identified using Box plots.

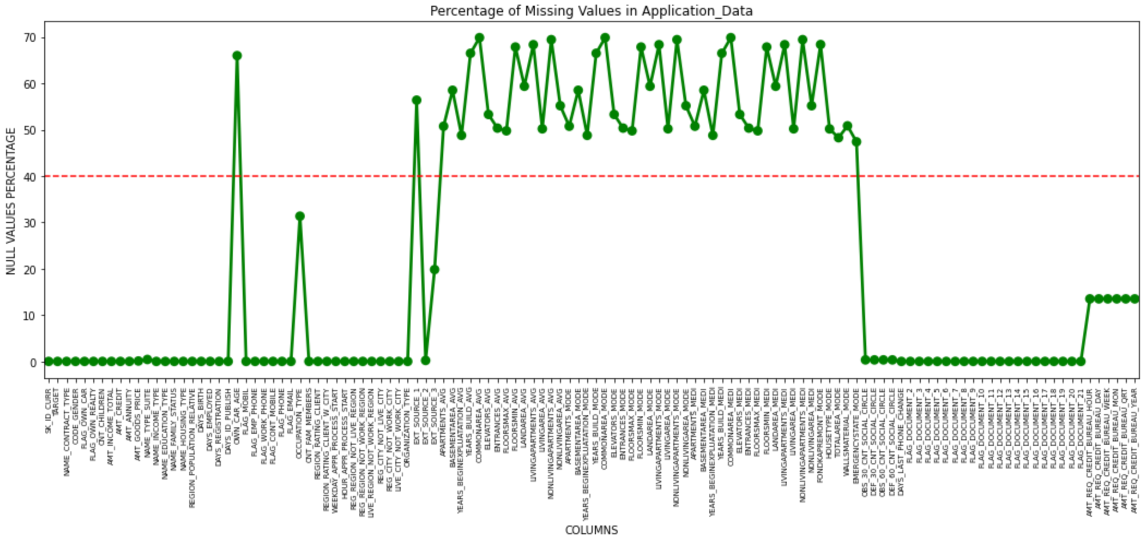
A box plot is a graphical rendition of a statistical data based on the minimum, first quartile, median, third quartile and maximum. Also, it contains outliers which is detected outside the plot.

The term boxplot comes from the fact that it looks like a rectangle with lines extending from the top to bottom.



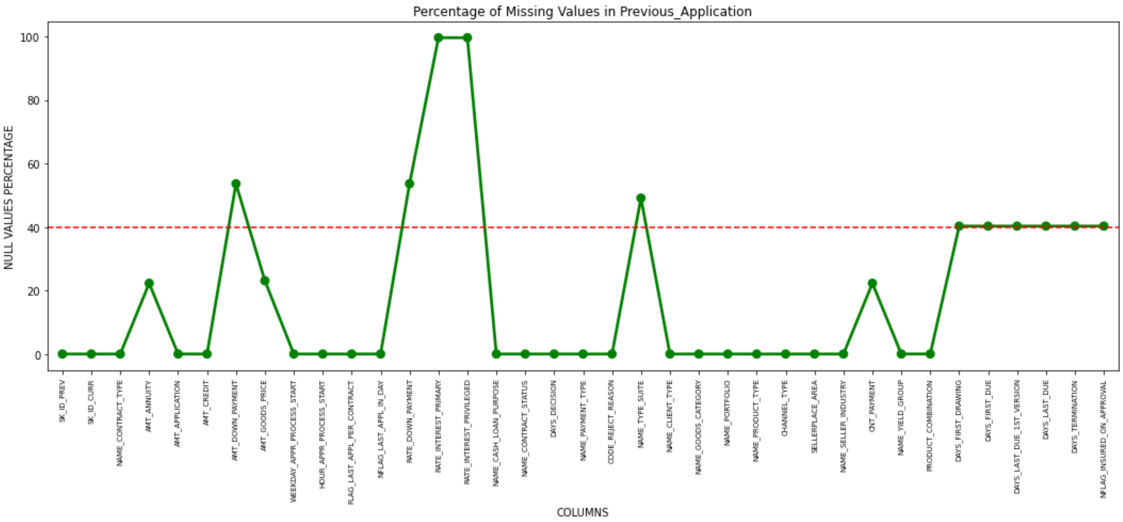
<u>In Application Data.</u>

There are 64 columns where the outliers are more than 40%.

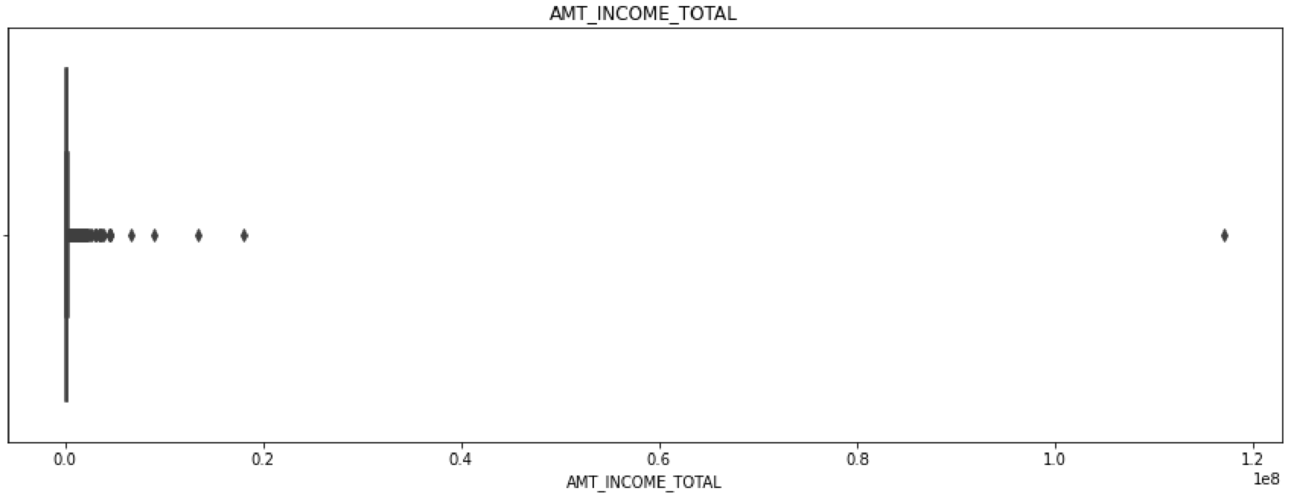


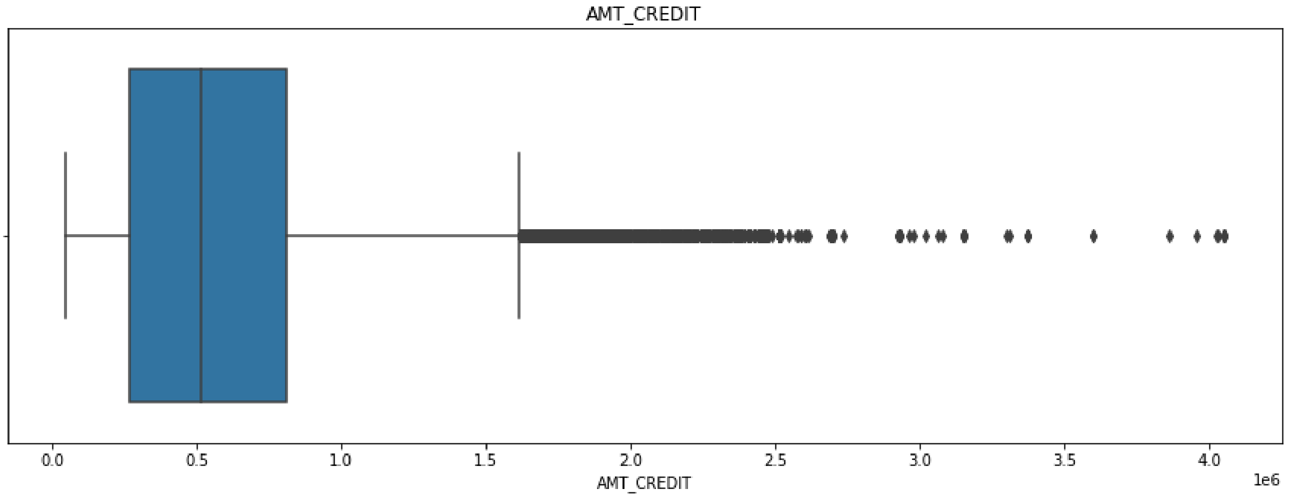
In Previous Data.

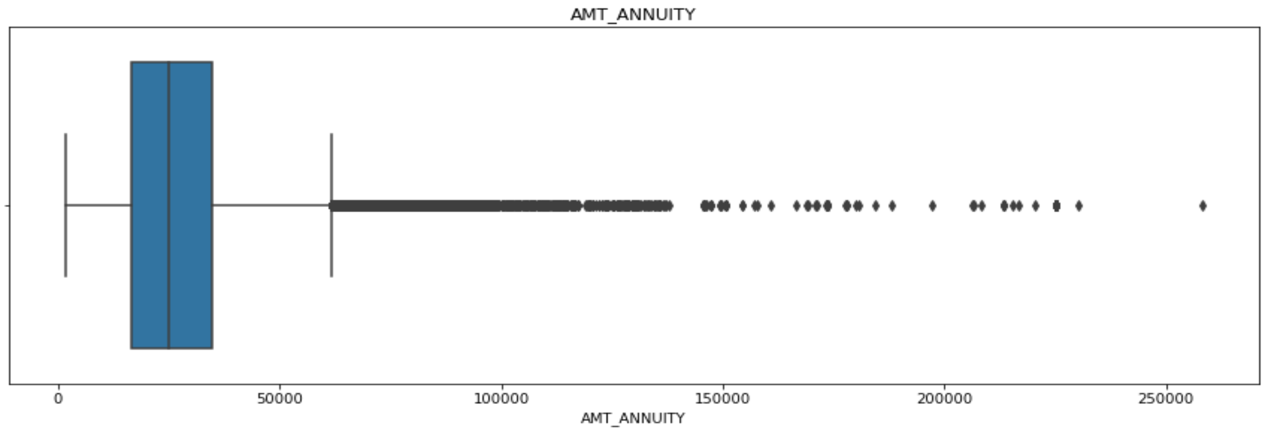
There are 11 columns where the outliers are more than 40%.

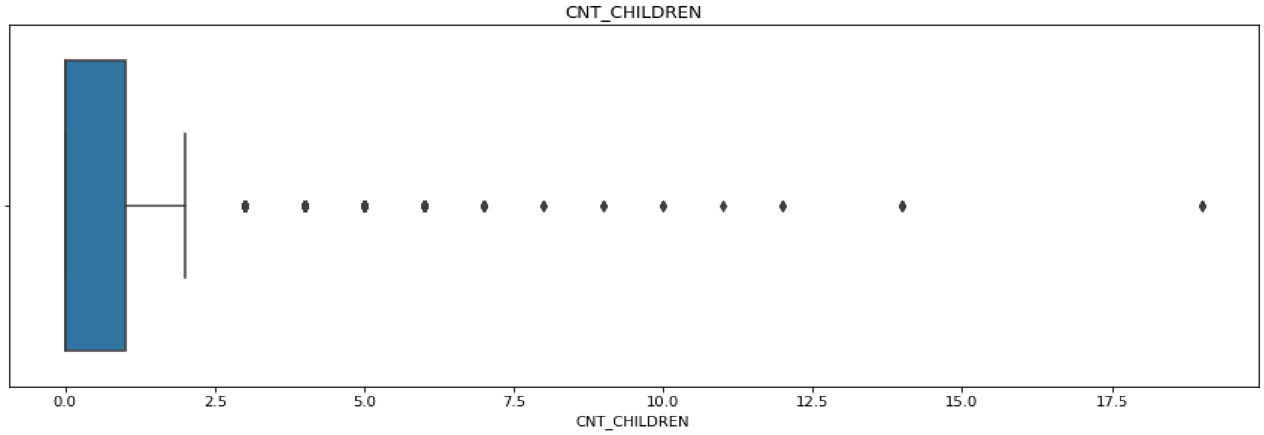


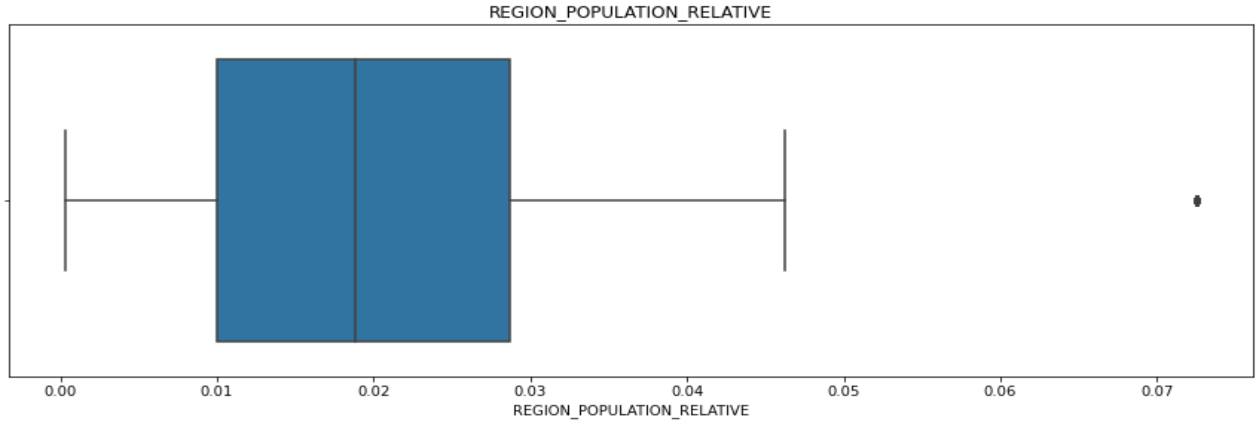
3A. Outliers in Application Data

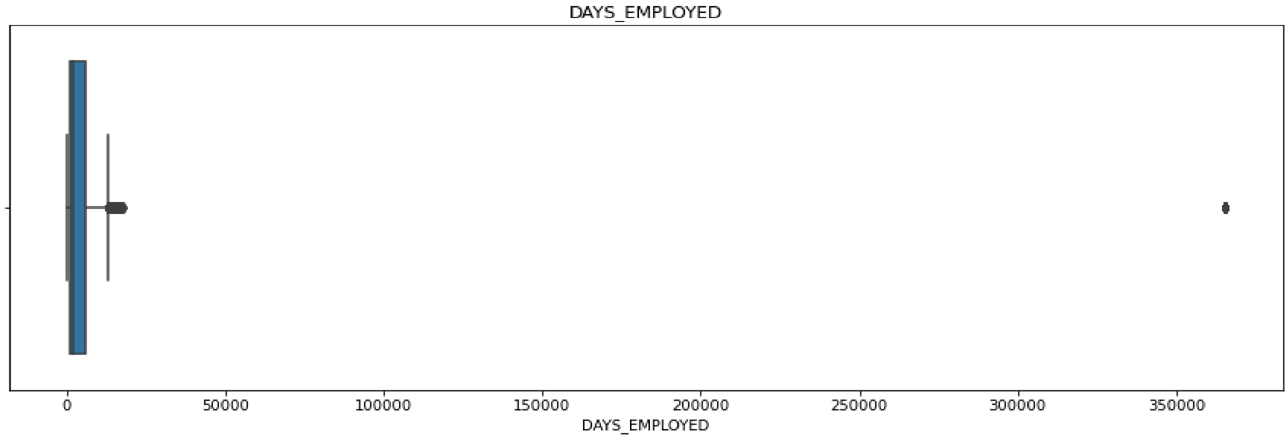


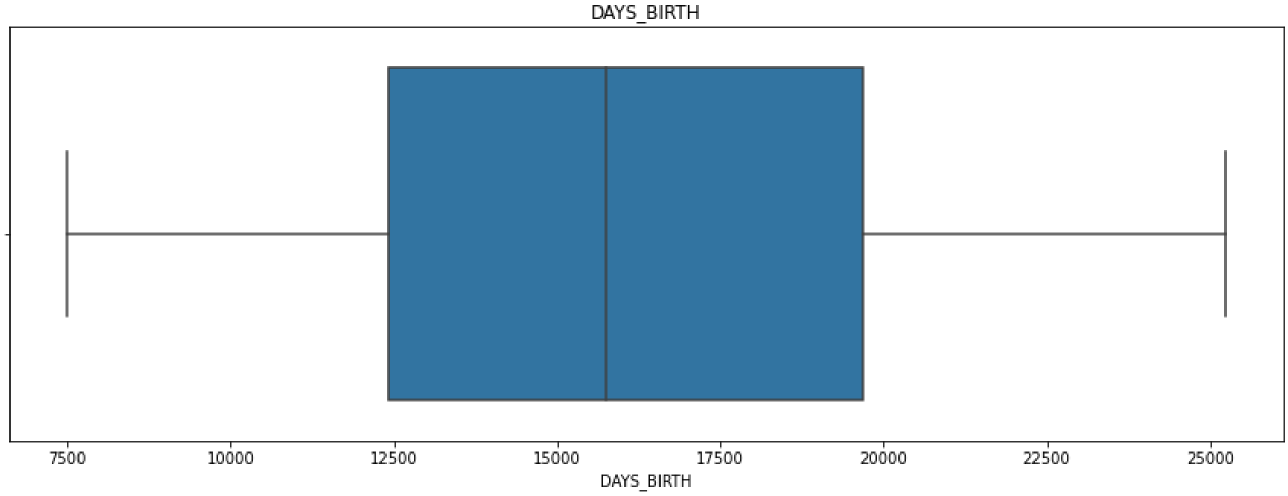



AMT_ANNUITY


CNT_CHILDREN


REGION_POPULATION_RELATIVE

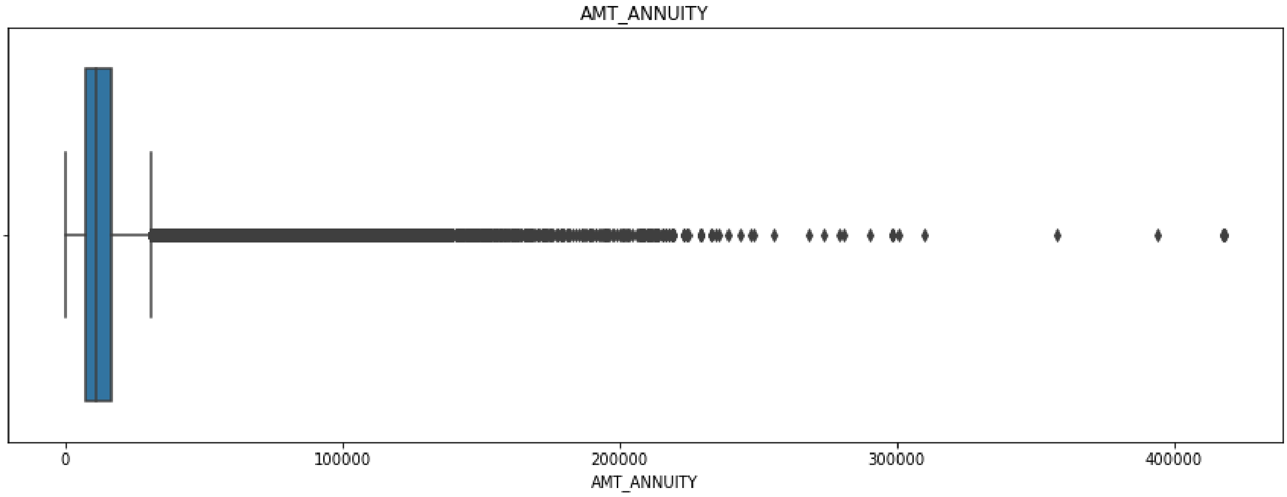

DAYS_EMPLOYED

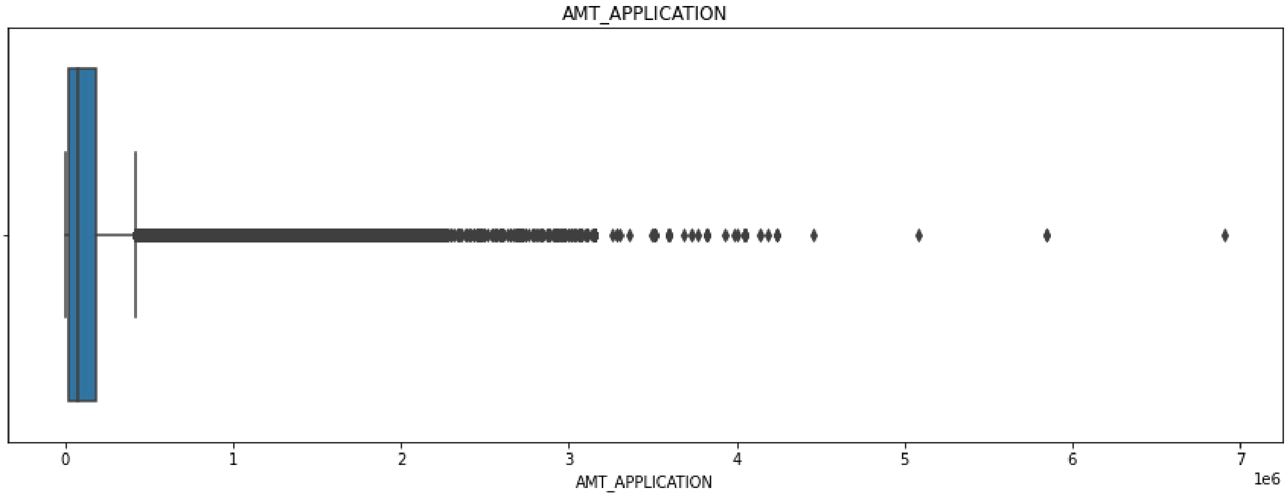


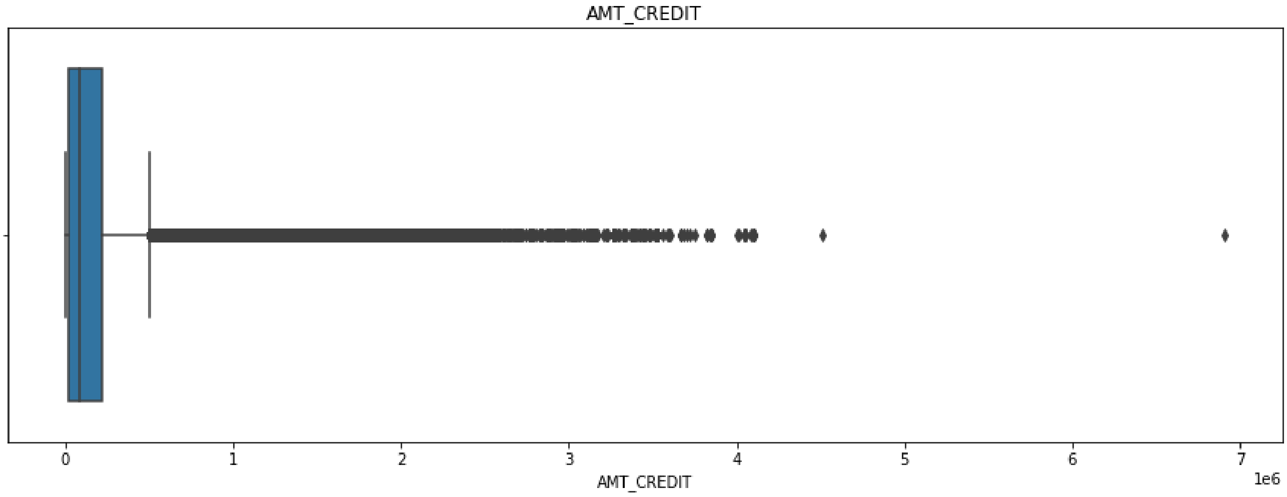
Insight:

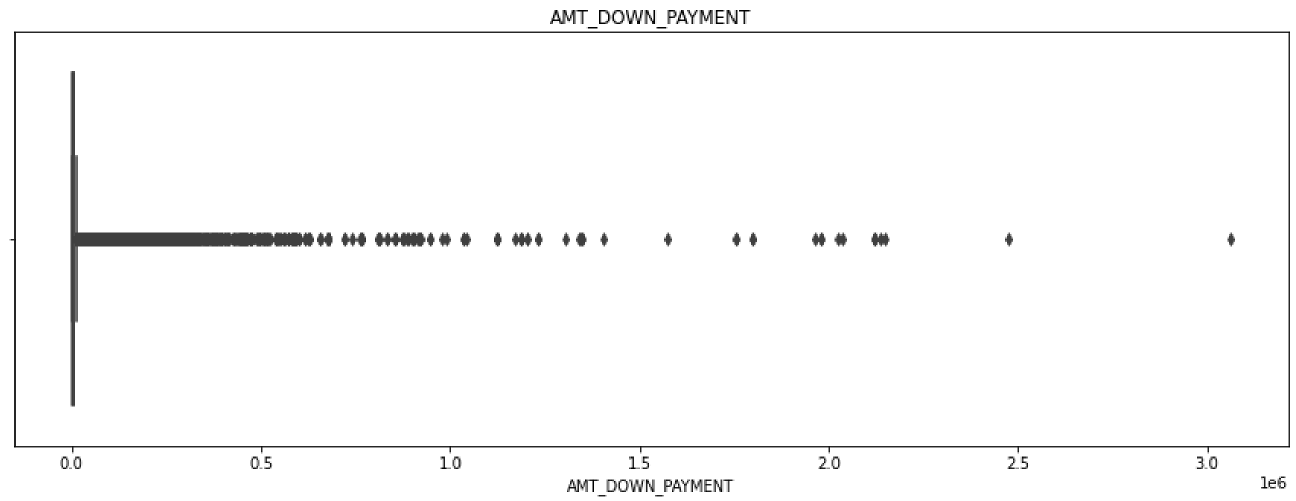
1. AMT_ANNUITY, AMT_CREDIT,CNT_CHILDREN have some number of outliers.
2. AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income compared to the others.
3. DAYS_BIRTH has no outliers which means the data available is reliable.
4. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this is an incorrect entry.

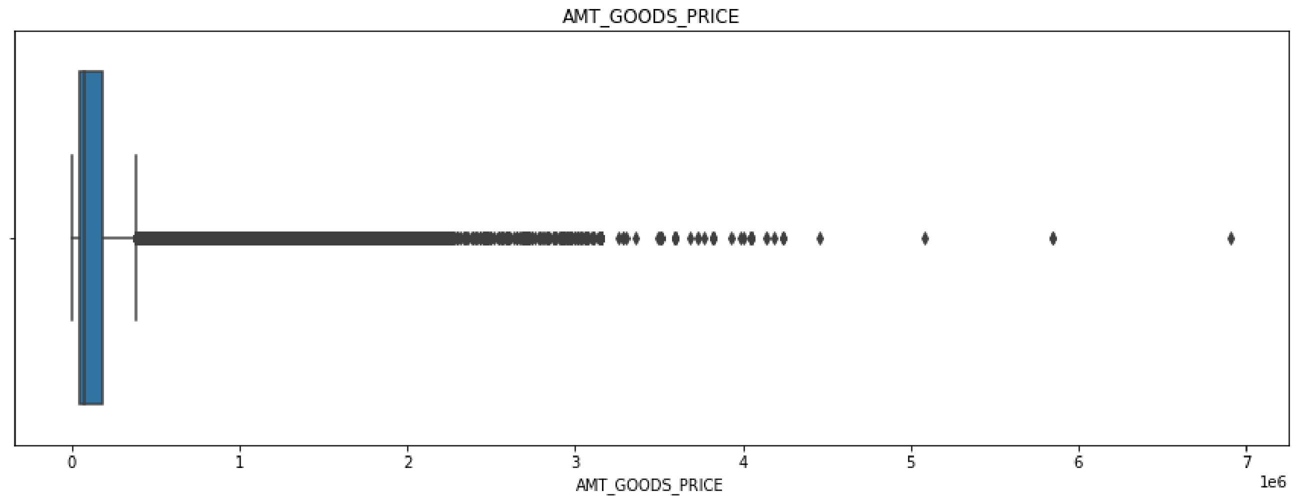
3B. Outliers in Previous Data

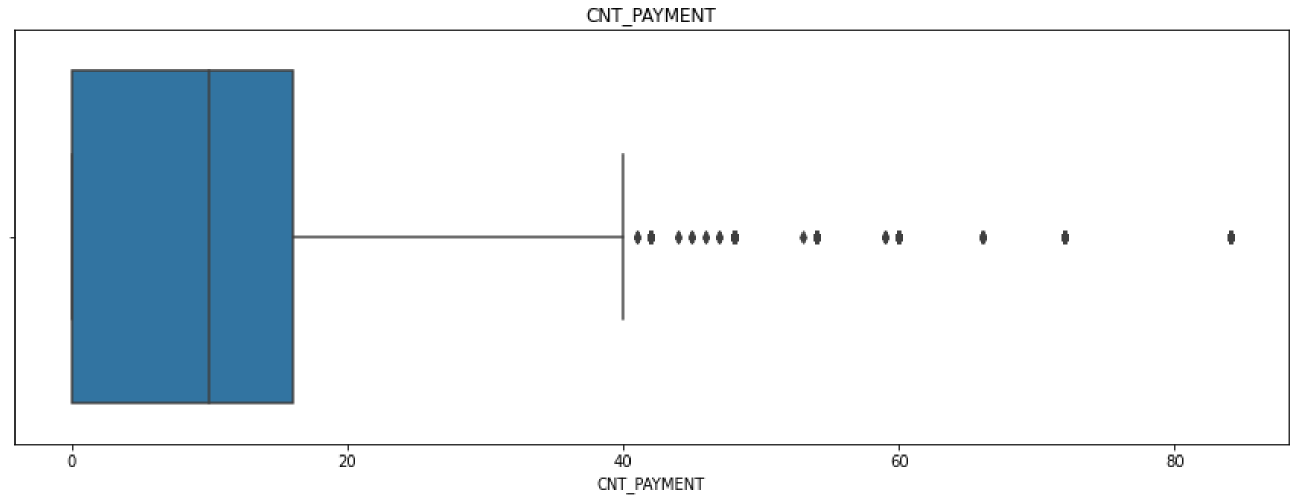


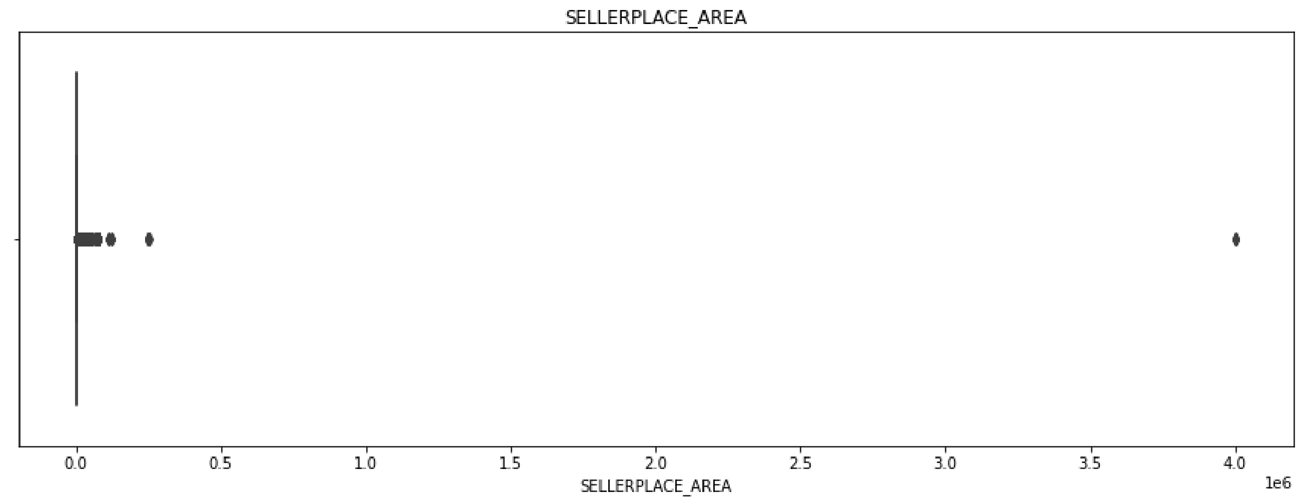


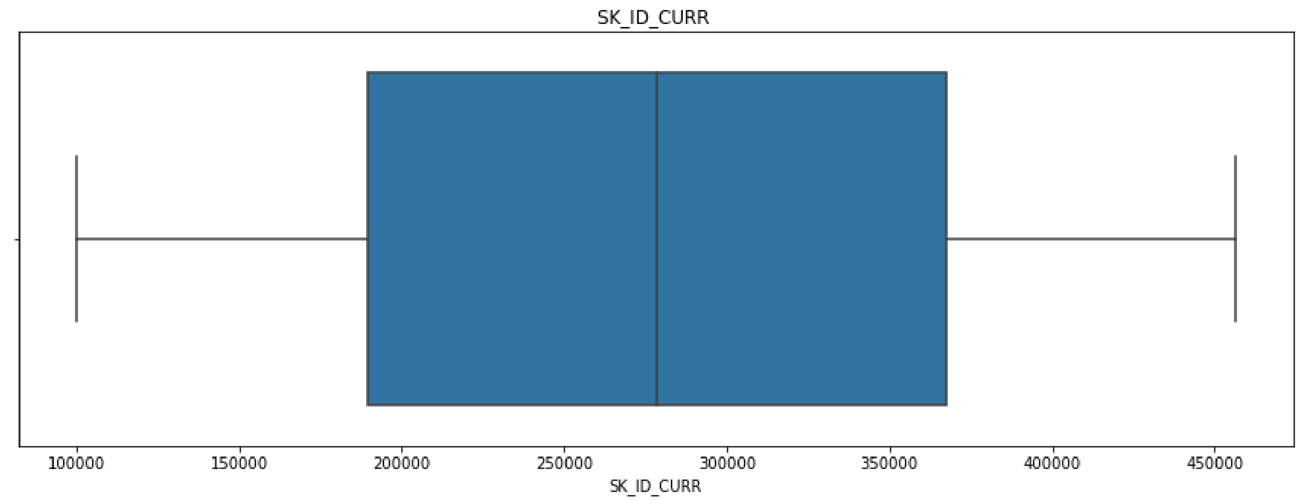


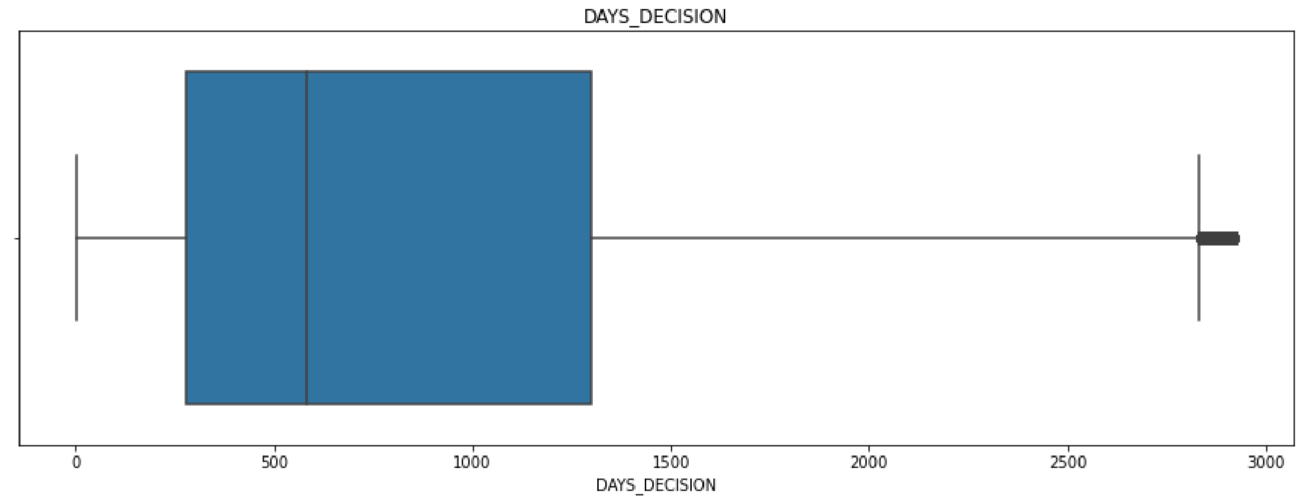

AMT_DOWN_PAYMENT
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
AMT_DOWN_PAYMENT

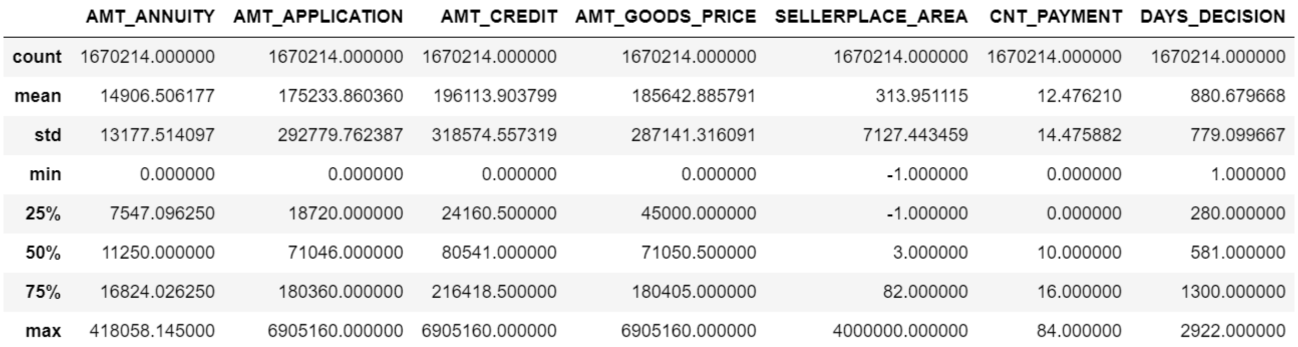

AMT_GOODS_PRICE
0
1
2
3
4
5
6
7
1e6
AMT_GOODS_PRICE


CNT_PAYMENT
0
20
40
60
80
CNT_PAYMENT


SELLERPLACE_AREA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
SELLERPLACE_AREA


SK_ID_CURR
100000
150000
200000
250000
300000
350000
400000
450000
SK_ID_CURR


DAYS_DECISION
0
500
1000
1500
2000
2500
3000
DAYS_DECISION



	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	SELLERPLACE_AREA	CNT_PAYMENT	DAYS_DECISION
count	1670214.000000	1670214.000000	1670214.000000	1670214.000000	1670214.000000	1670214.000000	1670214.000000
mean	14906.506177	175233.860360	196113.903799	185642.885791	313.951115	12.476210	880.679668
std	13177.514097	292779.762387	318574.557319	287141.316091	7127.443459	14.475882	779.099667
min	0.000000	0.000000	0.000000	0.000000	-1.000000	0.000000	1.000000
25%	7547.096250	18720.000000	24160.500000	45000.000000	-1.000000	0.000000	280.000000
50%	11250.000000	71046.000000	80541.000000	71050.500000	3.000000	10.000000	581.000000
75%	16824.026250	180360.000000	216418.500000	180405.000000	82.000000	16.000000	1300.000000
max	418058.145000	6905160.000000	6905160.000000	6905160.000000	4000000.000000	84.000000	2922.000000

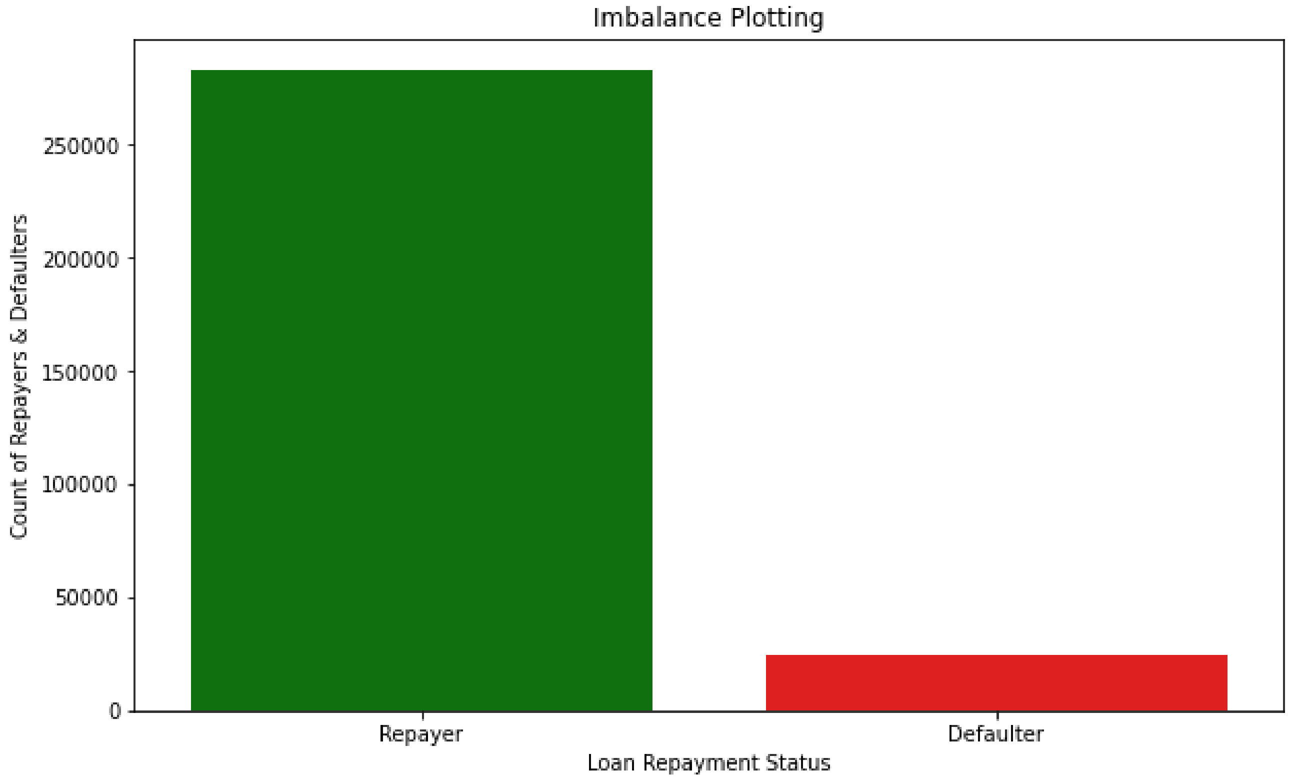
Insight:

1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
2. CNT_PAYMENT has few outlier values.
3. SK_ID_CURR is an ID column and hence no outliers.
4. DAYS_DECISION has few number of outliers indicating that these previous applications decisions were taken long back.

4. Identify if there is data imbalance in the data. Find the ratio of data imbalance.

The result of data imbalance shows-

Ratio of data imbalance in relative with respect to Repayor and Defaulter data is **11.39: 1.**

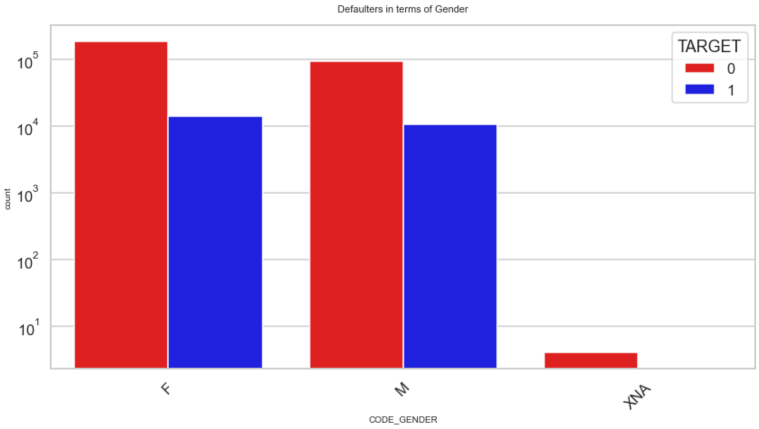


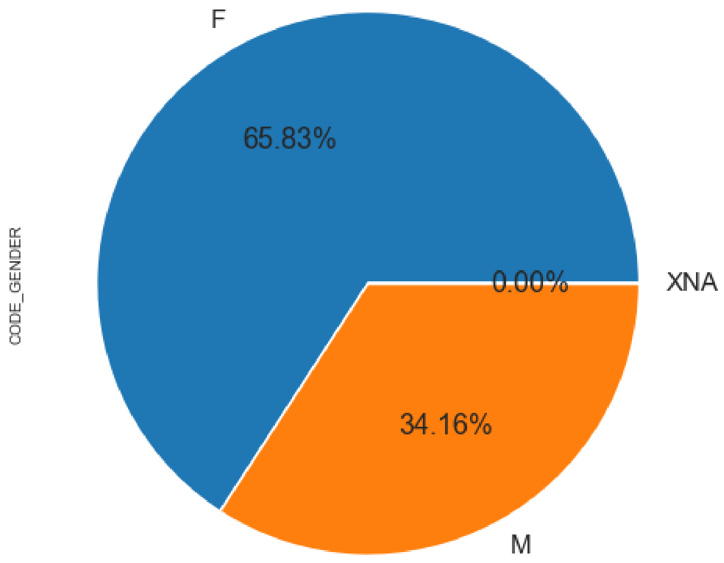
5. Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

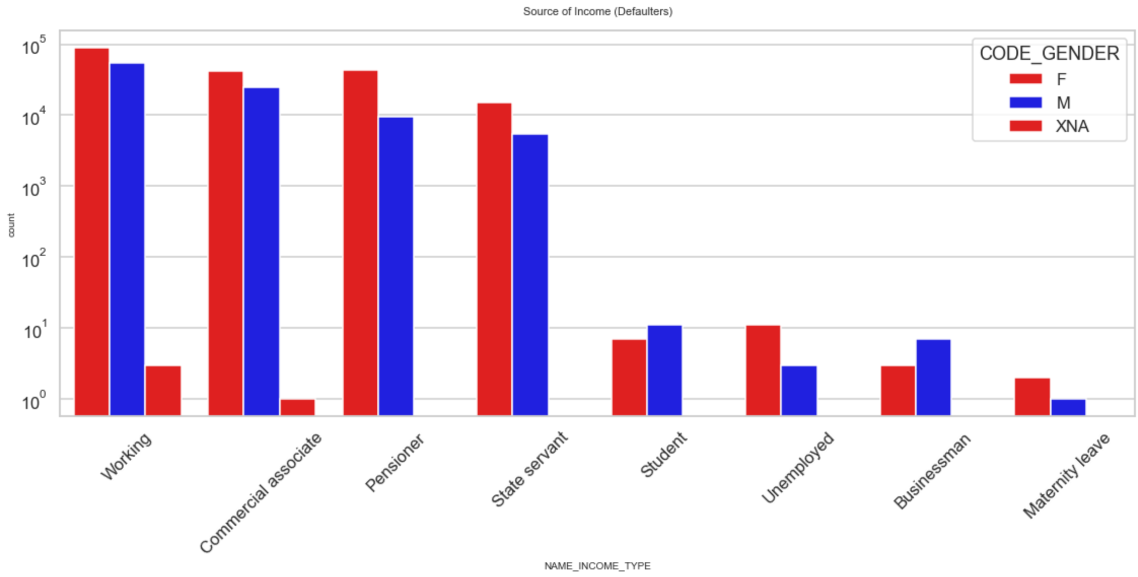
5A. Univariate Analysis for target 0

For target Variable 0(defaulters)

1. Most of the applicants are female but the twist is a smaller number of loans are taken by males but in case of not repaying the loan (defaulters), the number of males is in large numbers.
2. About 188278 females were found to be defaulters whereas males were 94404 in number.
3. Females have overtaken males in terms of not repaying the loan.
4. Most Females have 'Working' as their source of income, as compared to males, females are in larger numbers in terms of not repaying the loan.
5. Meanwhile, Commercial Associates (both female and male) have a great contribution to defaulters.
6. Maternity leave people have the least defaulter count.
7. Here, we can see most of the defaulters have income between 1Lakh-2Lakh whereas approx.
8. Rich clients (income above 1M) are less in the defaulters' line.
9. 20% of defaulters have income less than a Lakh.
10. More than 16% of people uses loan money above 10 lakhs as for their advantage.
11. Meanwhile, loans ranging from 2Lakh-3Lakh are taken with the intention of not repaying where females are in large numbers.
12. Hence as compared to revolving loans (9.21%), cash loans (90.79%) are taken in bulk.
13. Almost clients applied for cash loans later they turned into defaulters.
14. Clients having no car have more defaulter rate (69%).
15. In case of not returning loans, clients have a house or flat in large numbers (65%).
16. Business Entity type 3 people have a big bar in defaulter cases followed by Self Employed people and others.

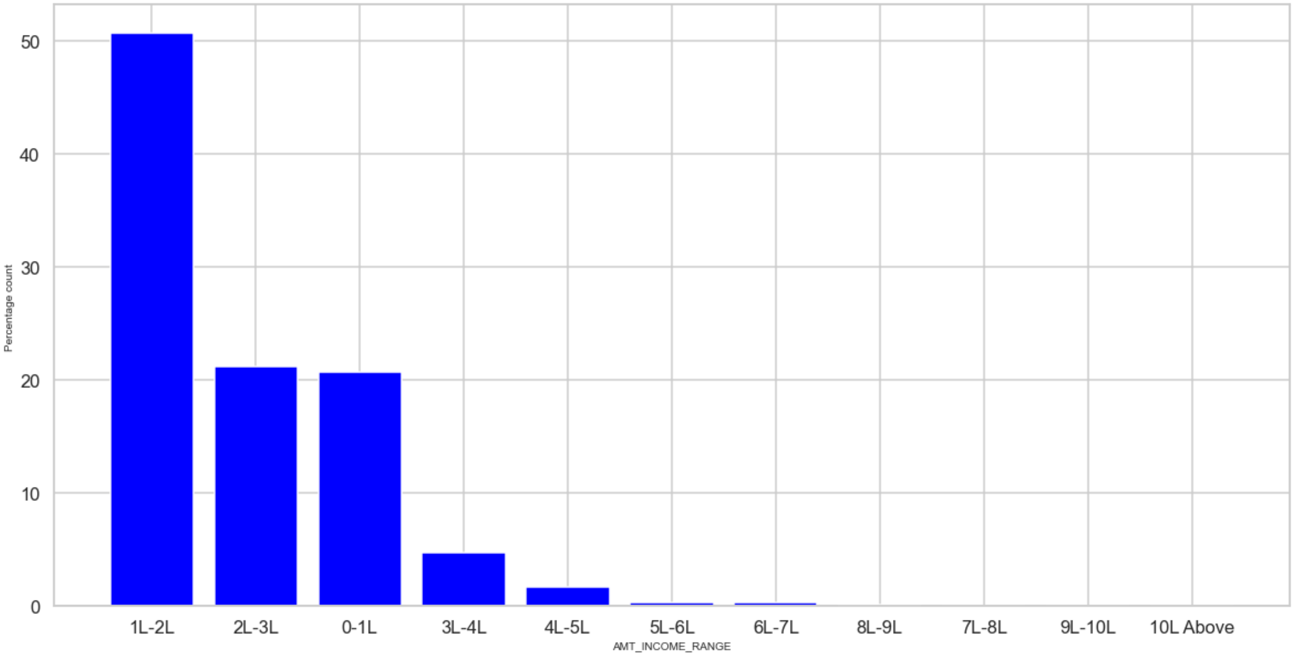


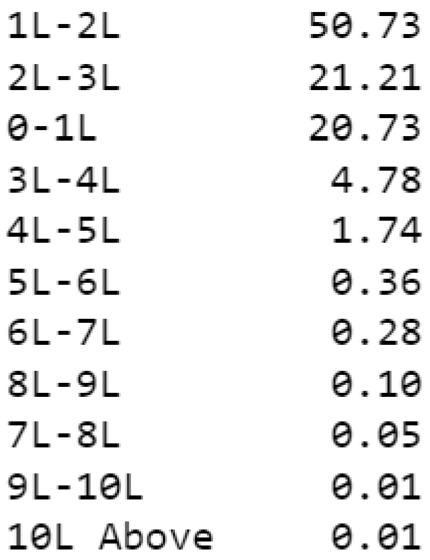




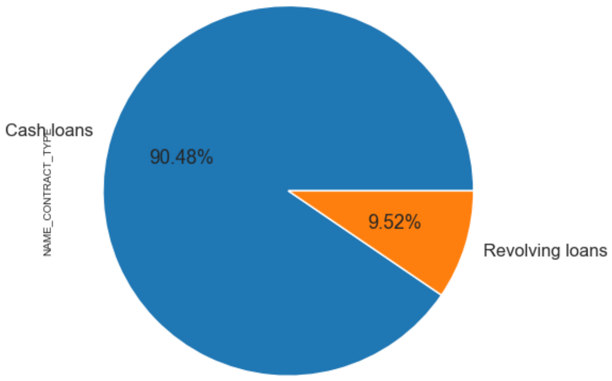
Insight:

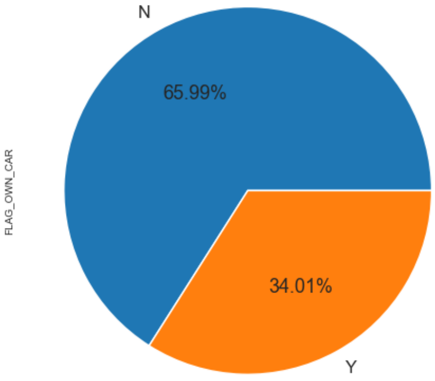
1. Most of Females have 'Working' as source of income.
2. As compared to male, females are in larger number in terms of not repaying the loan.
3. Meanwhile, Commercial Associates (both female and males) have great contribution in defaulters.
4. Maternity leave people have least defaulter values.

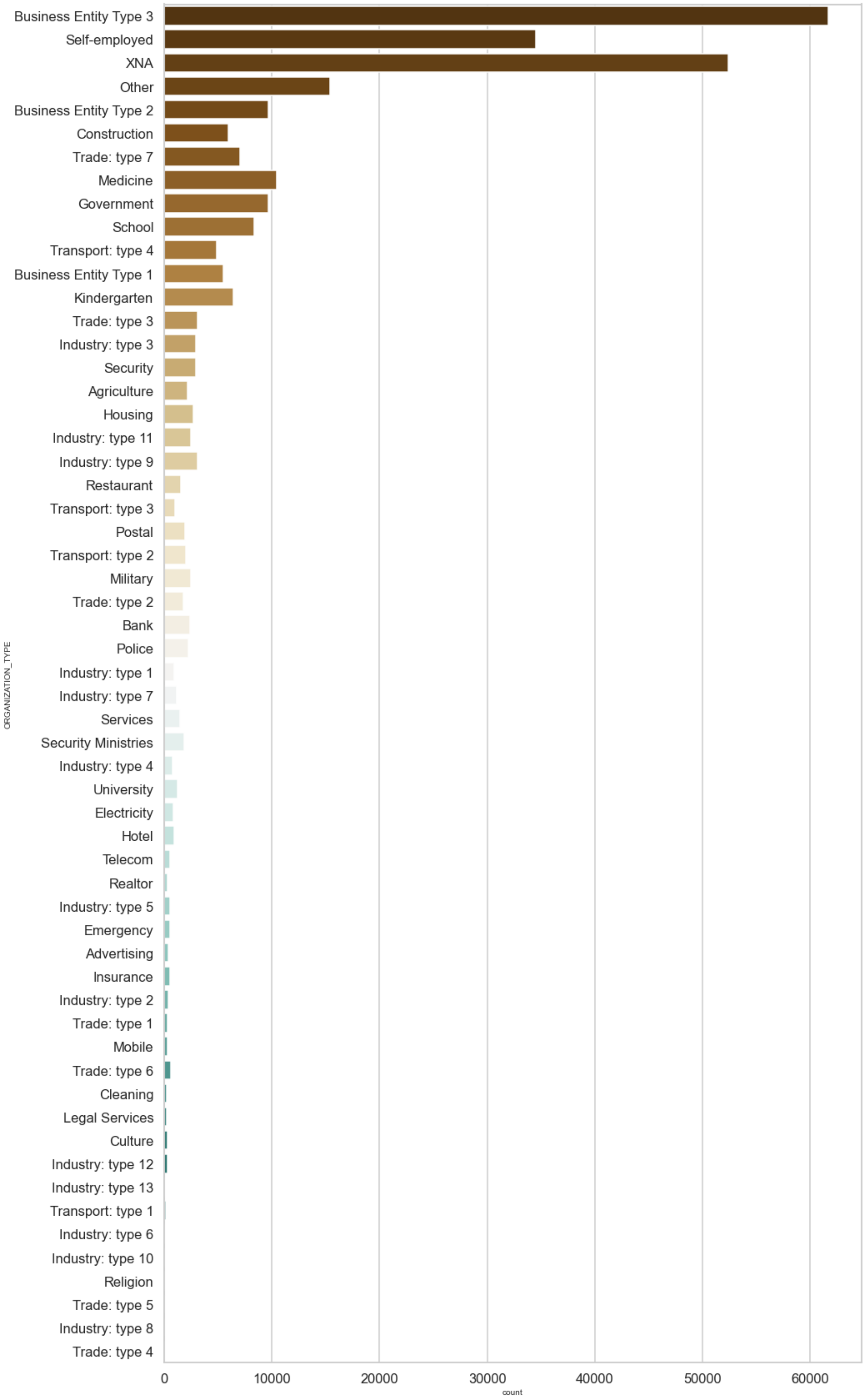




```
1L-2L        50.73
2L-3L        21.21
0-1L         20.73
3L-4L         4.78
4L-5L         1.74
5L-6L         0.36
6L-7L         0.28
8L-9L         0.10
7L-8L         0.05
9L-10L        0.01
10L Above     0.01
```





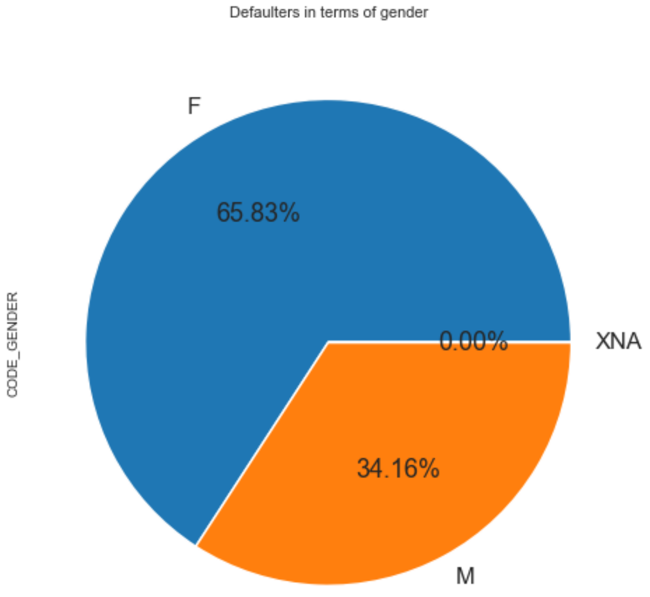


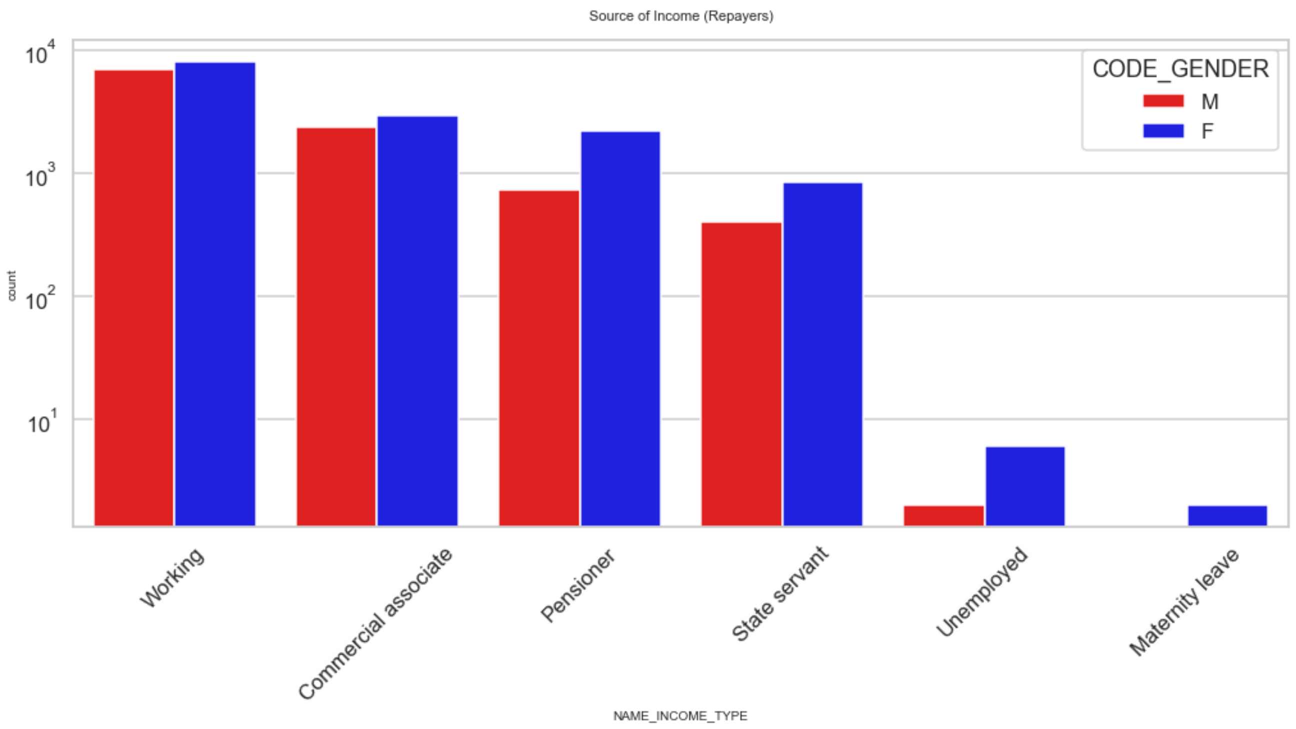
Insight

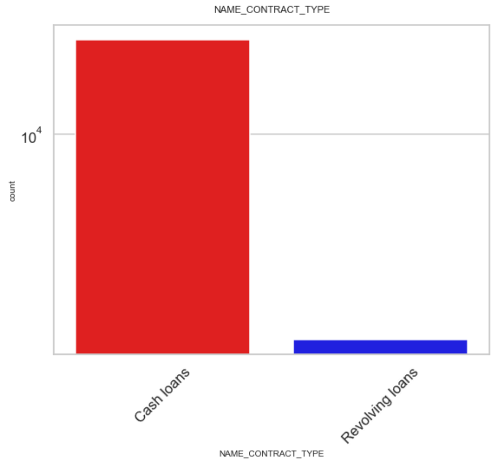
- Business Entity type 3 people have a big bar in defaulter case followed by Self-Employed people and other

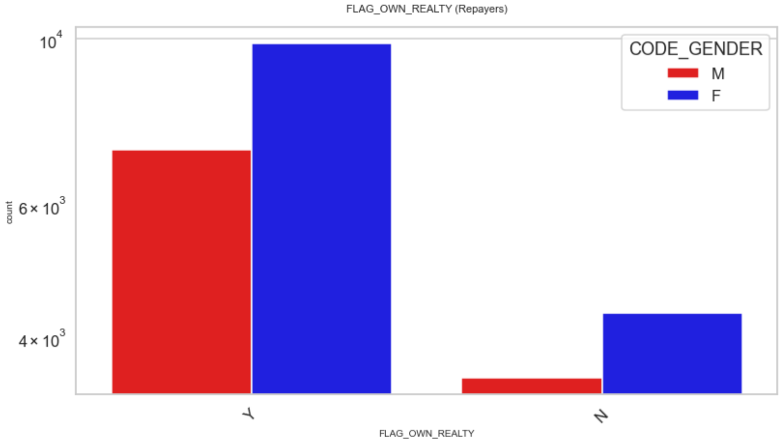
For target Variable 1(Repayors)

1. 57% of females are in the repayors counts which is less than females count in defaulters.
2. Here students and businessmen are missing whereas people with 'working' as the source of income is leading.
3. In terms of repaying loans, the percentage of people got increased from 50% in defaulter rate to 61% in repayment.
4. People having income 2lakh-3lakh are in more number in case of not repaying loan where here in terms of returning loan they are in fewer numbers as compared to defaulters' chart.
5. Vice versa for the people having income less than or equal to 1Lakh they are in more number for returning loans as compared to defaulter status.
6. Revolving loans returned by males are very less in numbers. Case loans in repayment status and defaulting status in quite the same.
7. Females who don't have a car have a big chance of repaying the loan and vice versa.
8. Males owning no flat or house have less chance of returning the loan.
9. Loan applicants not having a car are in better numbers in repayment status (69%).
10. Loan applicants owning a house or flat are also in good numbers (68%).



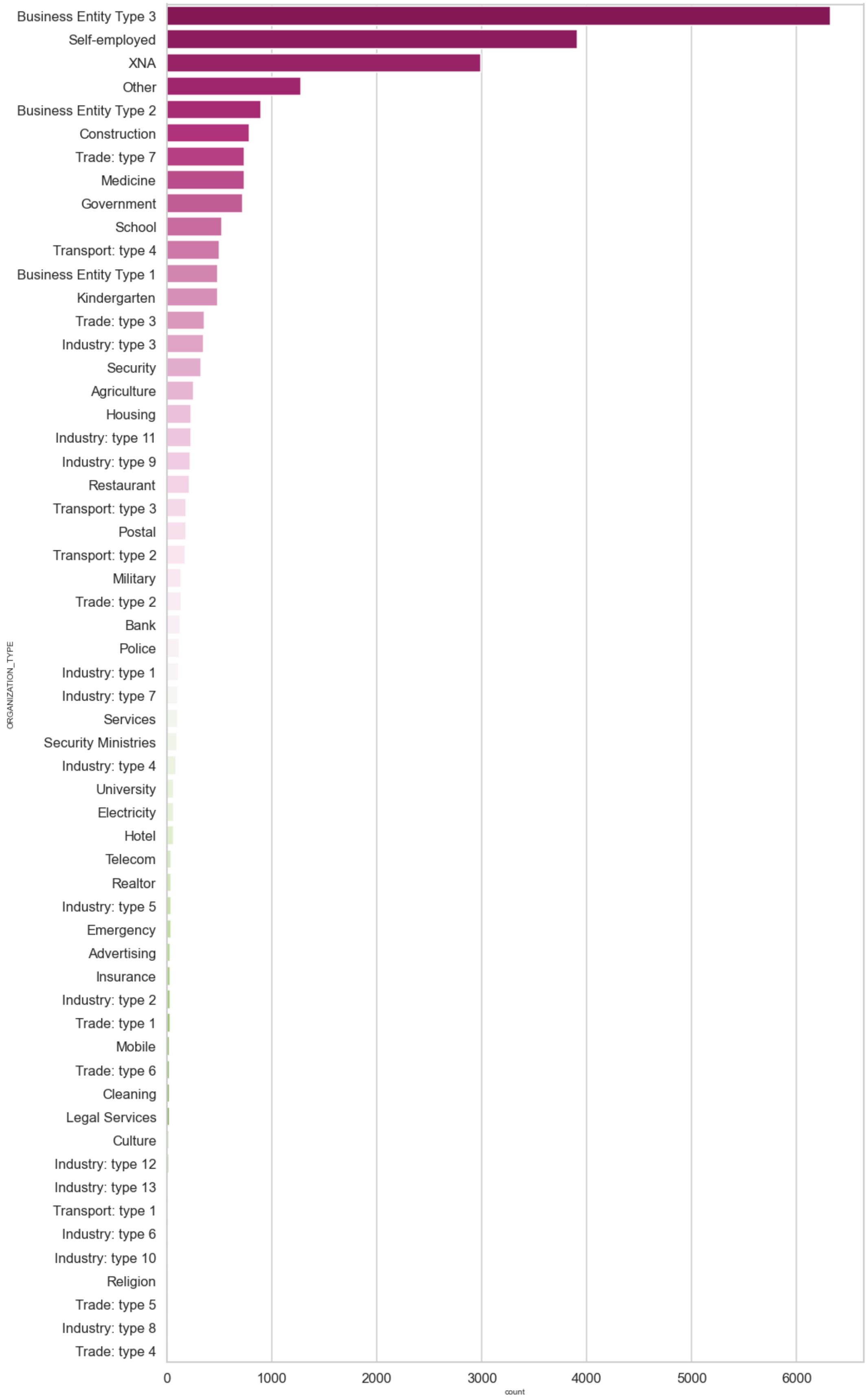






Insights:

1. Females who don't have a car have big chance of repaying the loan and vice versa.
2. Males owning no flat or house have less chance of returning the loan.



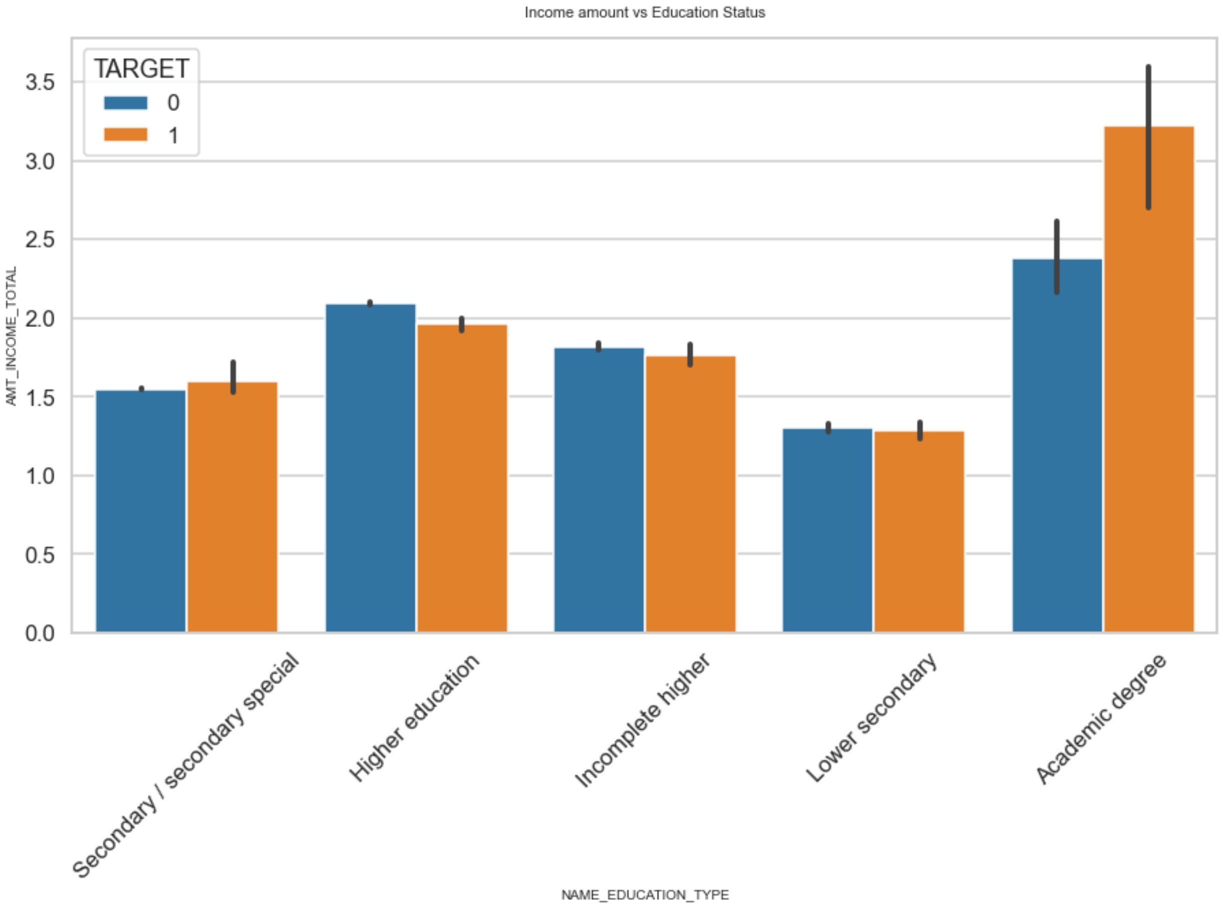
<u>Insights:</u>

1. loan applicants not having a car are in better numbers in repayment status (69%).

2. Loan applicants owning a house or flat are also in good number (68%

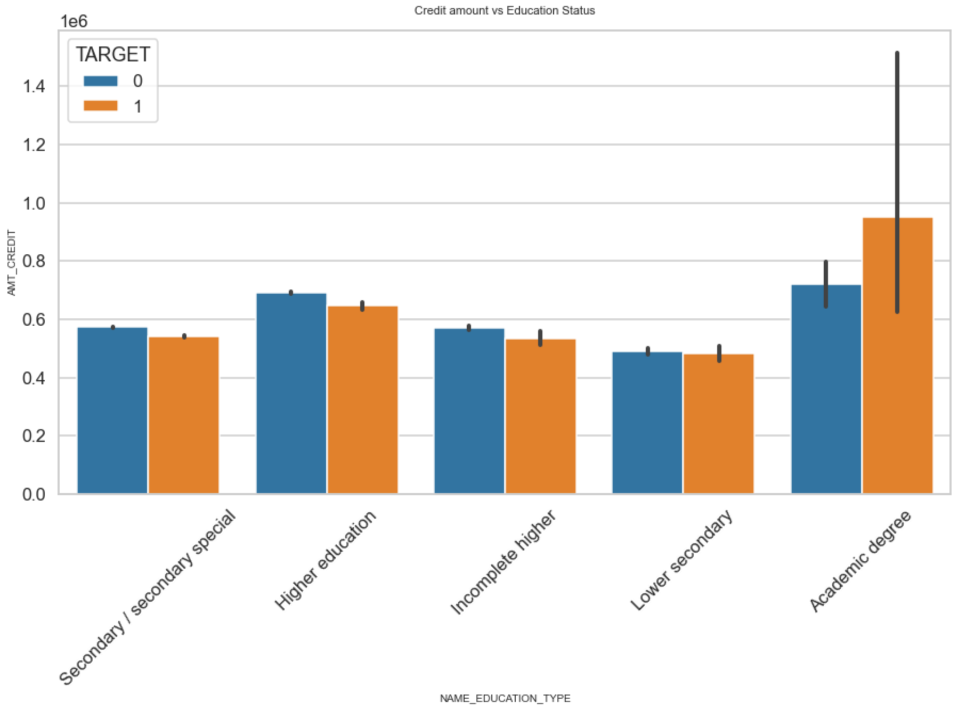
Result of bivariate analysis:

1. Clients having academic degrees who have the most income and have more counts in defaulters than repayors.
2. Clients educated from lower secondary are an equal number of counts in defaulters and repayors.
3. Working people having 2 family members are an equal number of repayors and defaulters.
4. People who get income through Maternity Leave tend to be more Defaulter when they have more Family Members.
5. Married people are high in defaulter counts as they have more children this may be a reason for default.

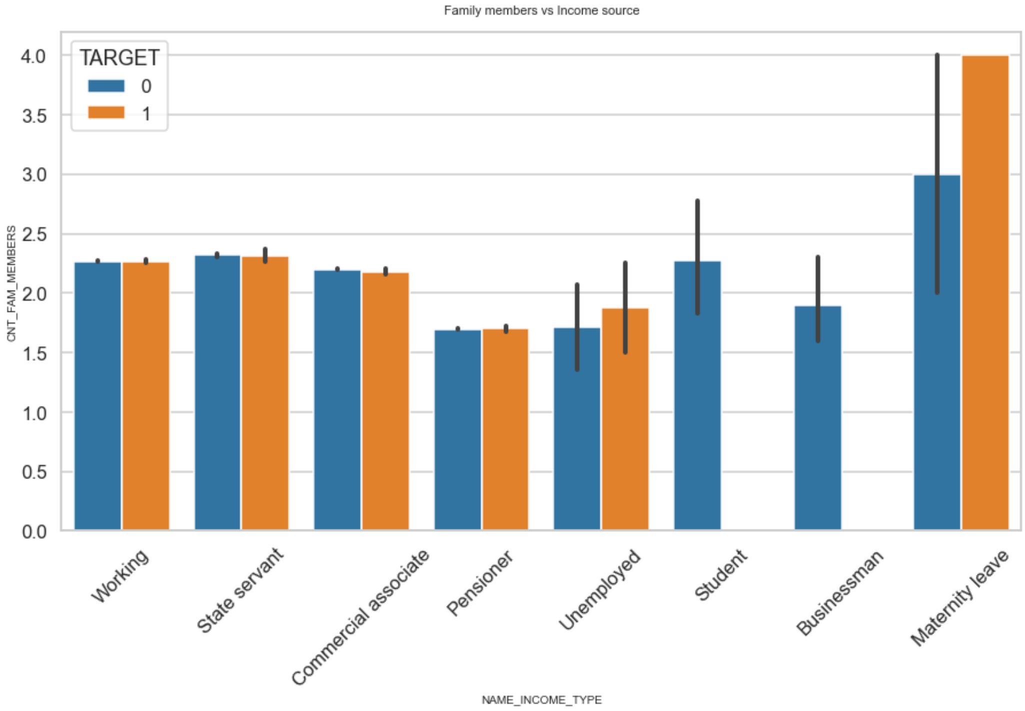


Insights:

1. Clients having academic degree have the most income have more counts in defaulters than repayors.
2. Clients educated from lower secondary are equal number of counts in defaulters and repayors.

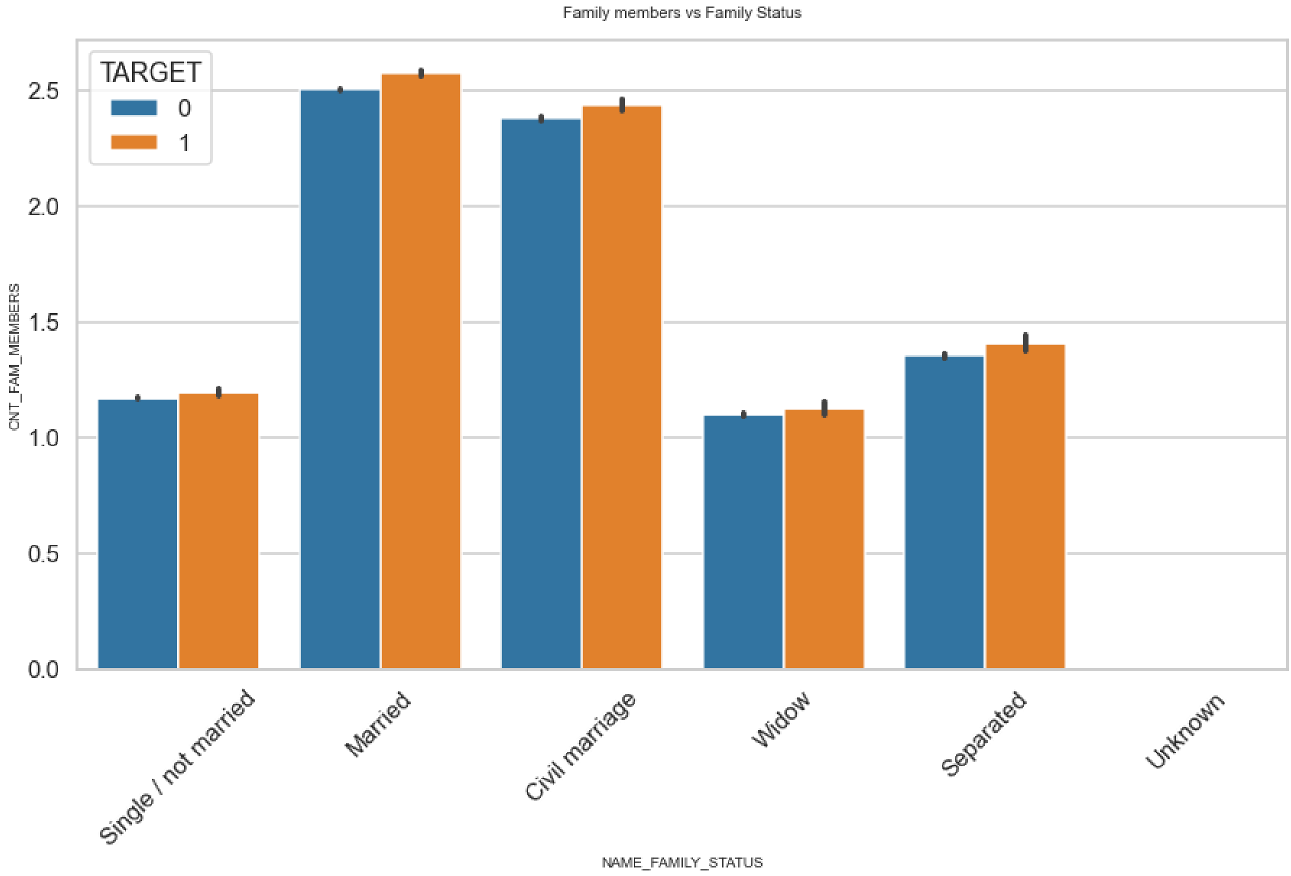


Academic degree people are safe side of giving loans as they have credited huge amount and they have returned less than credited amount to the company.



Insights:

1. Working people having 2 family members are equal number of repayors and defaulters,
2. People who getting income through Maternity Leave tends to be more Defaulter when they have more Family Members

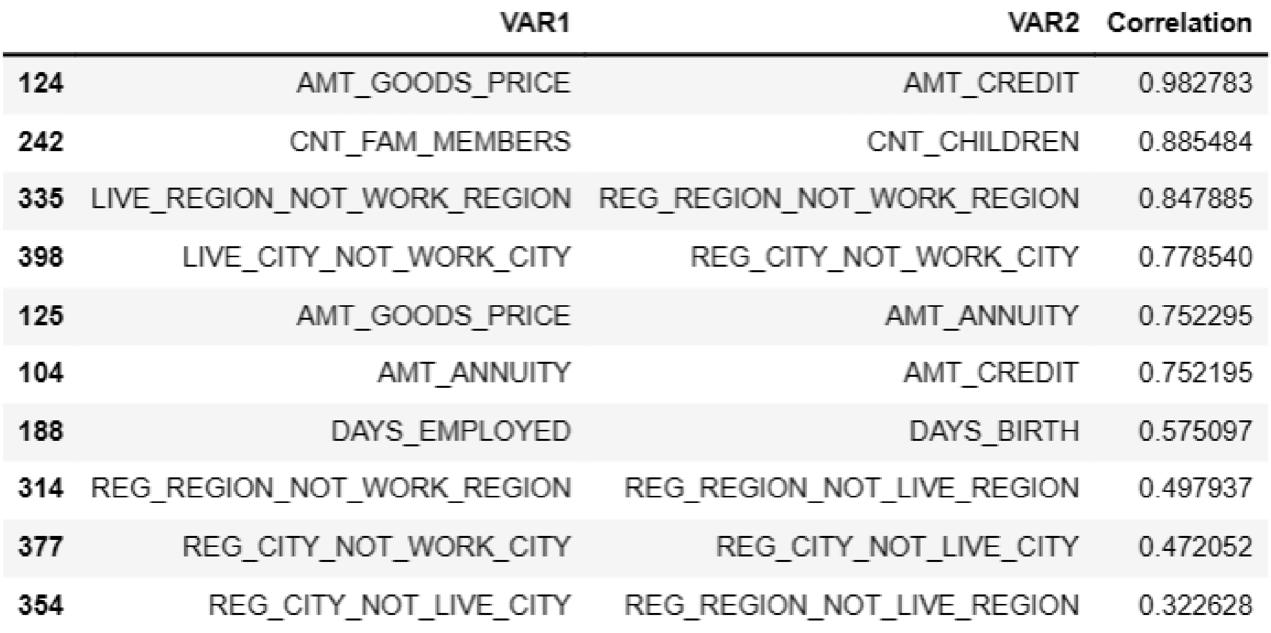


Married people are high in defaulter counts as they have more children’s this may be a reason of defaulting.

6. Find the top 10 correlations for the Client with payment difficulties and all other cases (Target variable).

Top 10 correlations for the client with payment difficulties (Repayor)

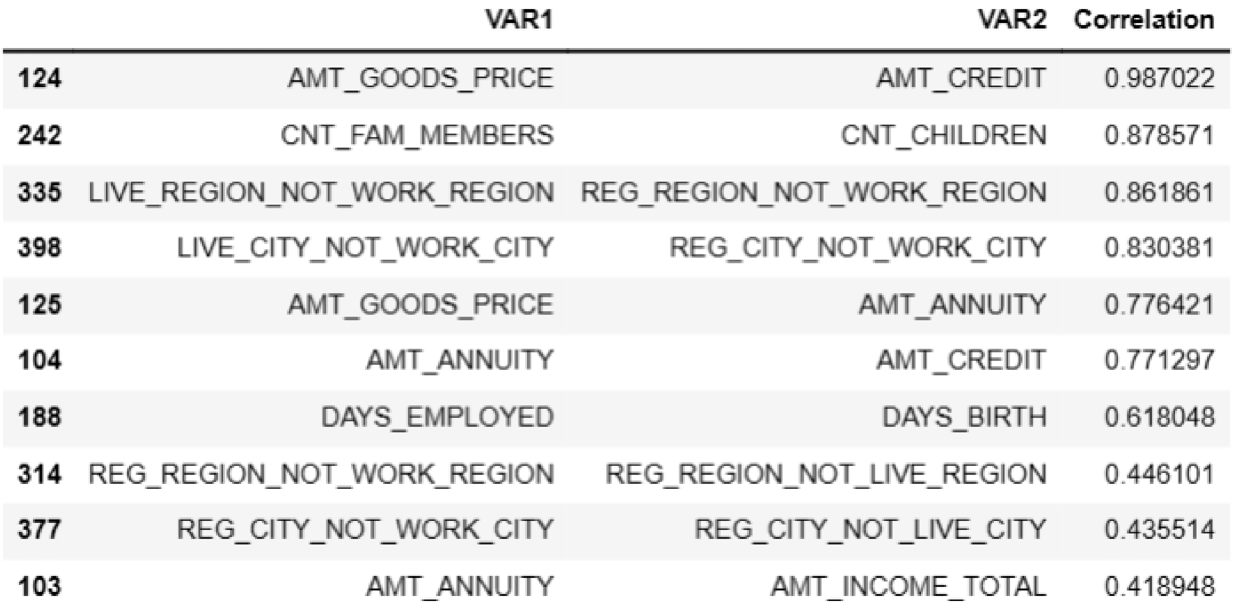
1. AMT_GOODS_PRICE and AMT_CREDIT is highly correlated (0.98).
2. Also, AMT_GOODS_PRICE is correlated with AMT_ANNUITY at 0.75.
3. CNT_FAM_MEMBERS and CNT_CHILDREN are also correlated at 0.88.
4. LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION comes in pictures in term of correlation (0.84).



	VAR1	VAR2	Correlation
124	AMT_GOODS_PRICE	AMT_CREDIT	0.982783
242	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
335	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
398	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
125	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295
104	AMT_ANNUITY	AMT_CREDIT	0.752195
188	DAYS_EMPLOYED	DAYS_BIRTH	0.575097
314	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937
377	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.472052
354	REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.322628

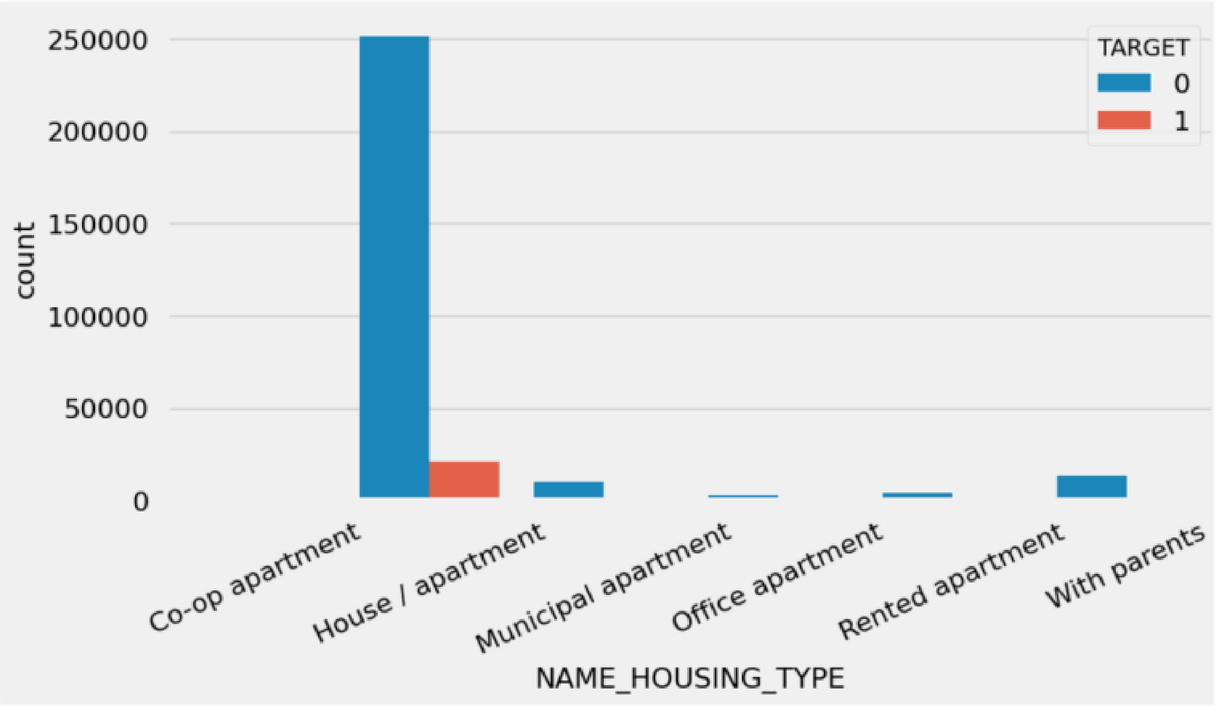
Top 10 correlations for the client with other cases (Defaulter)

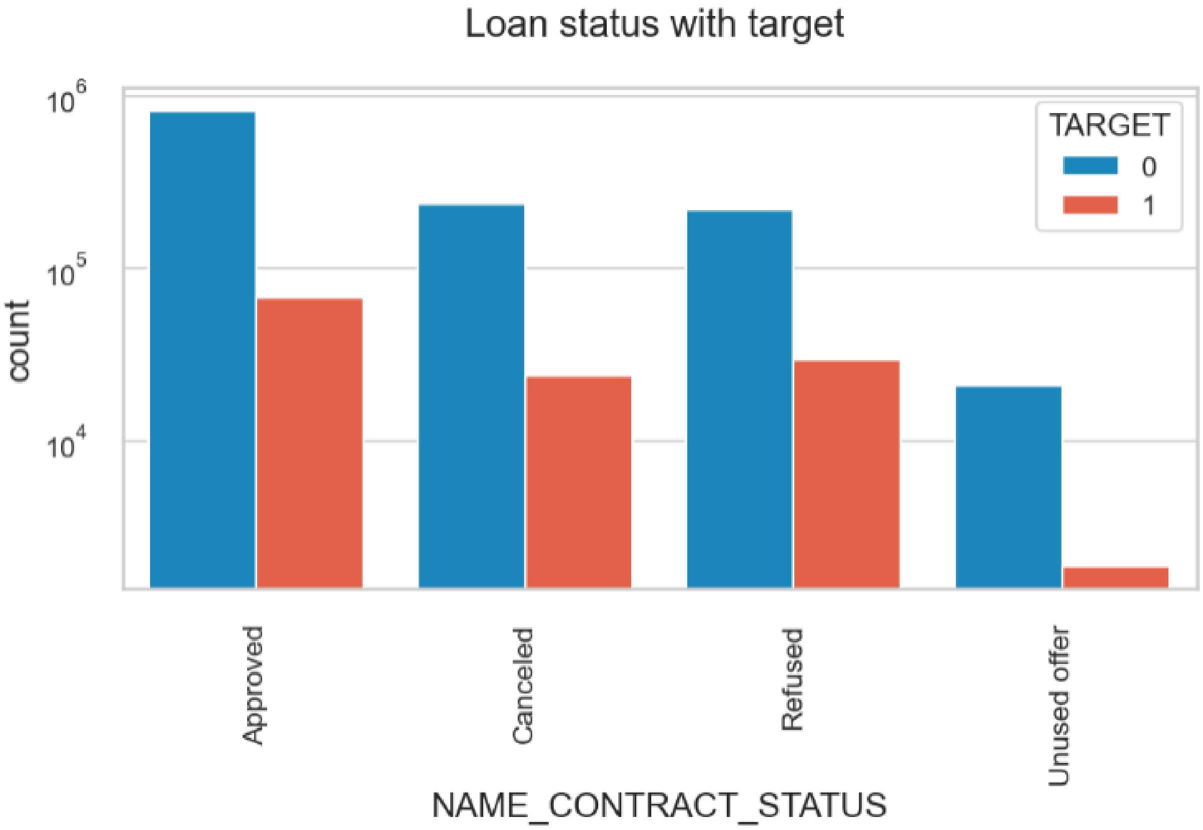
1. The correlation of AMT_GOODS_PRICE with AMT_CREDIT AND AMT_ANNUITY is quite same as in Repayment(above).
2. Whereas it's a bit increase in correlation of CNT_FAM_MEMBERS and CNT_CHILDREN (0.88) in defaulters as compared to repayment status (0.87).
3. Slightly decrease in correlation between AMT_GOODS_PRICE and AMT_ANNUITY (0.75) in the default case in comparison with repayment (0.77).
4. Same as in the case of AMT_ANNUITY and AMT_CREDIT, the correlation reduced from 0.77 to 0.75 in the defaulter case.
5. The correlation of DAYS_EMPLOYED and DAYS_BIRTH seems to be increased from 0.57 in repayment to 0.61 in defaulter.
6. Almost top 6 correlation are same but with slightly different values.
7. In defaulter case, AMT_ANNUITY and AMT_INCOME_TOTAL at 0.41 come in top_10 correlation.

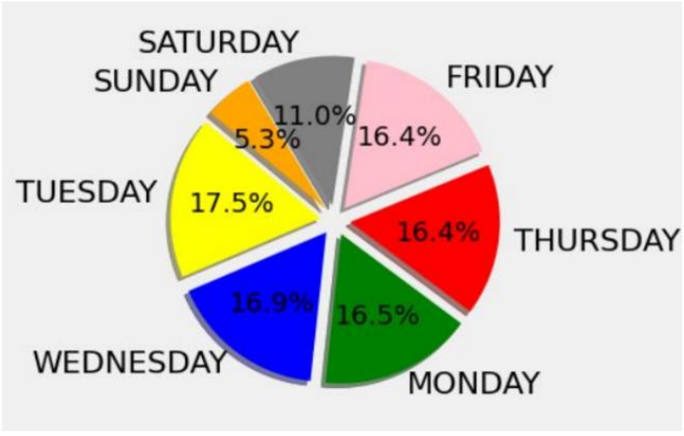


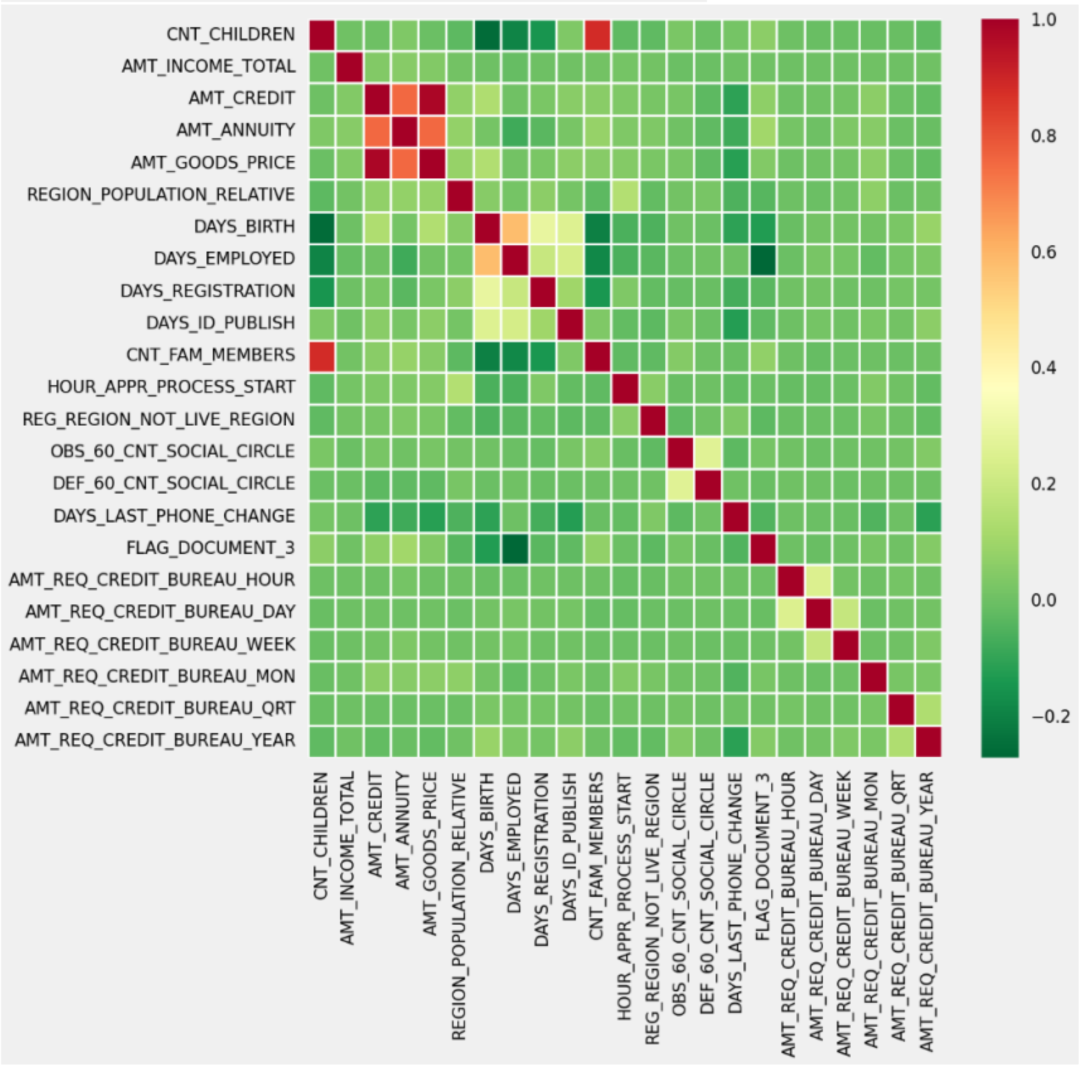
	VAR1	VAR2	Correlation
124	AMT_GOODS_PRICE	AMT_CREDIT	0.987022
242	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
335	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
398	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
125	AMT_GOODS_PRICE	AMT_ANNUITY	0.776421
104	AMT_ANNUITY	AMT_CREDIT	0.771297
188	DAYS_EMPLOYED	DAYS_BIRTH	0.618048
314	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.446101
377	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.435514
103	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418948

7. <u>Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables.</u>






SATURDAY
11.0%
SUNDAY
5.3%
FRIDAY
16.4%
TUESDAY
17.5%
THURSDAY
16.4%
WEDNESDAY
16.9%
MONDAY
16.5%


CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
CNT_FAM_MEMBERS
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
FLAG_DOCUMENT_3
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Insights:

1. Females were found to be more defaulters as compared to males.
2. Out of 100 only 16% of clients have applied on weekend days (Saturday & Sunday) for current loans meanwhile on TUESDAY most of the clients about 17% have shown interest in applying.
3. Clients living in apartments and houses have the most defaulters and it's not safe side to approve such clients' loan applications.
4. 18% of clients have applied for the very first time whereas almost 73% of clients have applied for loans again.
5. After XAP, the common reason behind the rejection of a loan is HC.
6. In Previous applications, about 41% of clients have applied for POS followed by Cash with 22%.
7. Other than XAP and XNA, for repairing purposes most of the defaulters have taken advantage of the loan amount.
8. Buying a used car is also a major reason for applying for a loan after urgent needs.
9. Company should think before approving loan applications for such reasons.
10. People living in office apartments is having higher credit for defaulting and people in co-op apartments have repaid the loan despite taking huge amounts as a loan.
11. Municipal apartments also have huge bars in not repaying the loan.
12. Defaulters use the unused offer for their benefit whereas only more than the half approved loans are repaid.
13. New clients have returned their loan payments but there are still defaulters more than repayors.
14. Whereas clients applying for loans again has most counts in terms of not paying loan status.
15. Company should pay attention on client whose previous application was for POS, Cash, etc because they are also contributing to the defaulters counts.

Result:

1. From the project we get to know how a company should manage risk during giving loans to clients.
2. Understood visualizing techniques using python libraries (pandas, matplotlib, seaborn, etc.) and the senseful data from the graphs and charts.
3. Learned how to present valuable insights and driving factors from the huge dataset.
4. Learned about correlations of important variables and the idea of presenting them.
5. Removal of null values, imputation of null values, data imbalance, outliers, univariate, bivariate analysis, etc. have also been studied.
6. Helped me to use EDA (Exploratory Data Analysis) in real business case scenes.