Bank Loan Case Study

By Ashish Upadhyay

Project Description

We used the Bank Loan Dataset in this study to do Exploratory Data Analysis (EDA). This case study tries to give you a sense of how EDA is applied in a genuine business context. Due to their insufficient or absence of credit histories, loan providers find it hard to provide loans to individuals.
Because of this, some customers take advantage of it by defaulting.

Business Understanding

Imagine that we are employed by a consumer finance business specializing in providing different types of loans to urban clients. To assess the patterns found in the data, we must use EDA. By doing this, it will be ensured that only those applicants who can repay the loan will be accepted.
When a loan application is received, the business must evaluate whether to approve the loan based on the applicant's profile.
The bank's judgment is subject to two distinct types of risk:

- If the applicant is likely to repay the loan, denying the loan leads to a loss of business for the company.
- If the applicant is not likely to pay back the loan or is likely to default, then approving the loan may result in a loss of revenue for the business.

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 the 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 the client received worse pricing which he did not want.
3. **Rejected:** The company had rejected the loan (because the client does not meet their requirements etc.).

4. **Unused Offers:** The loan has been cancelled by the client but at different stages of the process.

In this case study, we'll utilize EDA to learn how borrower attributes and loan attributes affect the likelihood of default.

We must gain some insights from this project, such as:

- Which application types are more likely to repay the loan amount and which applicant types have taken advantage of the loan amount?
- identifying such defaulters and enforcing strict actions such as refusing the loan, lowering the loan amount, lending (too risky applicants) at a higher interest rate, etc. By doing this, it will be ensured that only borrowers who can repay the loan will be accepted.
- Presenting a method for removing useless columns, null values, and outliers from the data, describing the findings of univariate, segmented univariate, bivariate analysis, etc. in terms of business, and identifying the top 10 correlations for the client with payment issues and all other cases.
- Finding such outcomes by conducting EDA will assist the company in deciding whether to approve or deny the loan application.

Tech-Stack Used:

1. **MS Excel:-** I used MS Excel to understand the dataset and what exactly means, also valuable for data analysis for small datasets, and also useful in understanding the column description.
2. **Google Colab / Jupyter Notebook:-** Google Colaboratory ("Colab" for short) is a data analysis and machine learning tool that allows you to combine executable Python code and rich text along with charts, images, HTML, LaTeX, and more into a single document stored in Google Drive.
3. **Python** (**Programming Language**)**:-** Python is a programming language that has extensive support of libraries which makes data analysis easier.
4. **Google Doc:-** Used google docs to prepare the report.

Procedure:

1. In order to understand the dataset's objective and the significance of its columns, used MS Excel.
2. I imported the necessary Python libraries (Pandas, Numpy, Seaborn, etc.) into my colab notebook.
3. Datasets (Application_Data and Previous_Application) are imported after libraries.
4. Identified missing values and eliminated columns that weren't necessary for analysis.
5. Identified relationships between outliers across various columns.
6. Graphs were used to display the identified data imbalance and its ratio.
7. Conducted data analysis using univariate, segmented univariate, and bivariate methods.
8. List the top ten correlations between TARGET variables.
9. All of the meaningful data (analysis work) was finally presented in the form of various graphs and charts.

Understanding Data:

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 includes the data on whether the previous application had been Approved, Cancelled, Refused, or Unused offer.
3. **`columns_descrption.csv`** is a data dictionary that describes the meaning of the variables.

Identify the missing data and use appropriate methods to deal with it. (Remove columns/or replace them 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 -

1. 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.
2. Similarly, the missing values in the column 'AMT_ANNUITY' is imputed using the same median () operation.
3. In median imputation, the missing values are replaced with the median value of the entire feature column.
4. 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.

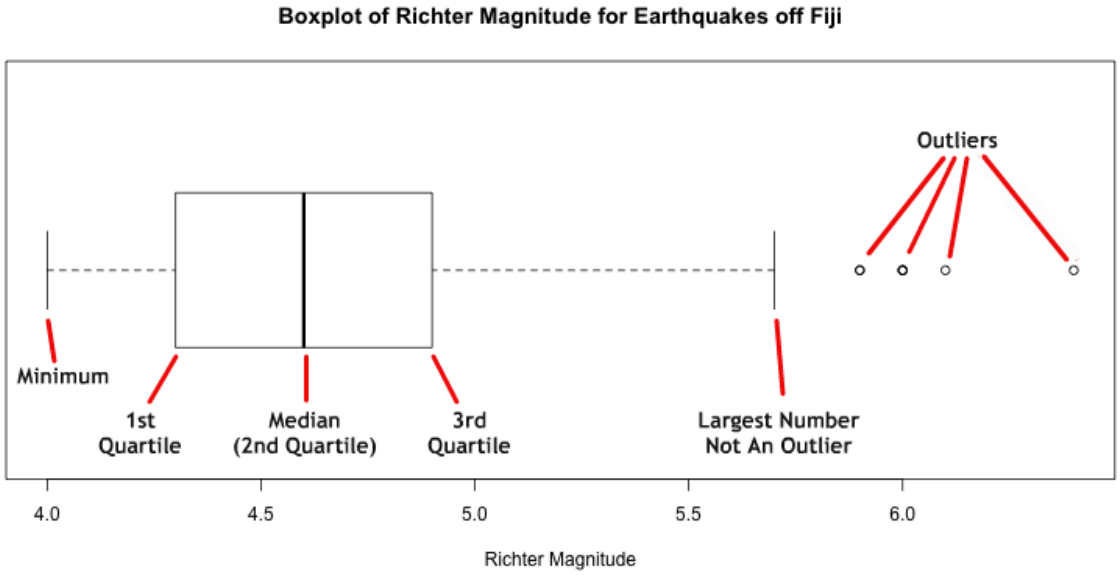
In Previous_Application -

1. Imputation in Previous application.
2. Converting negative days to positive as days can't be negative.
3. Days in Columns 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', and 'DAYS_ID_PUBLISH' are converted into positive values.
4. Imputing Null Values/Missing values in Previous Application
5. 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.
6. For imputing null values, in column 'AMT_ANNUITY' nulls are filled using the median Operation. In 'AMT_GOODS_PRICE' and 'AMT_CREDIT', we have used mode ().
7. And imputation in 'CNT_PAYMENT' is done by filling null by '0'.

Identify if there are outliers in the dataset. Also, mention why you think it is an outlier.

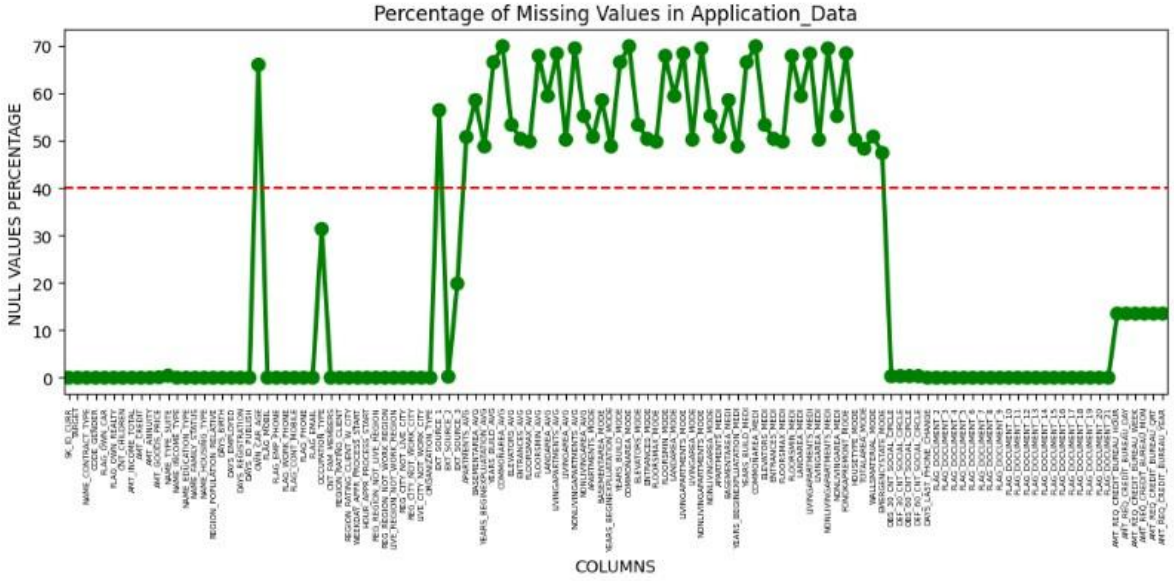
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. The outliers in this analysis work are identified using Box plots.
A box plot is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum. Also, it contains outliers that are 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.



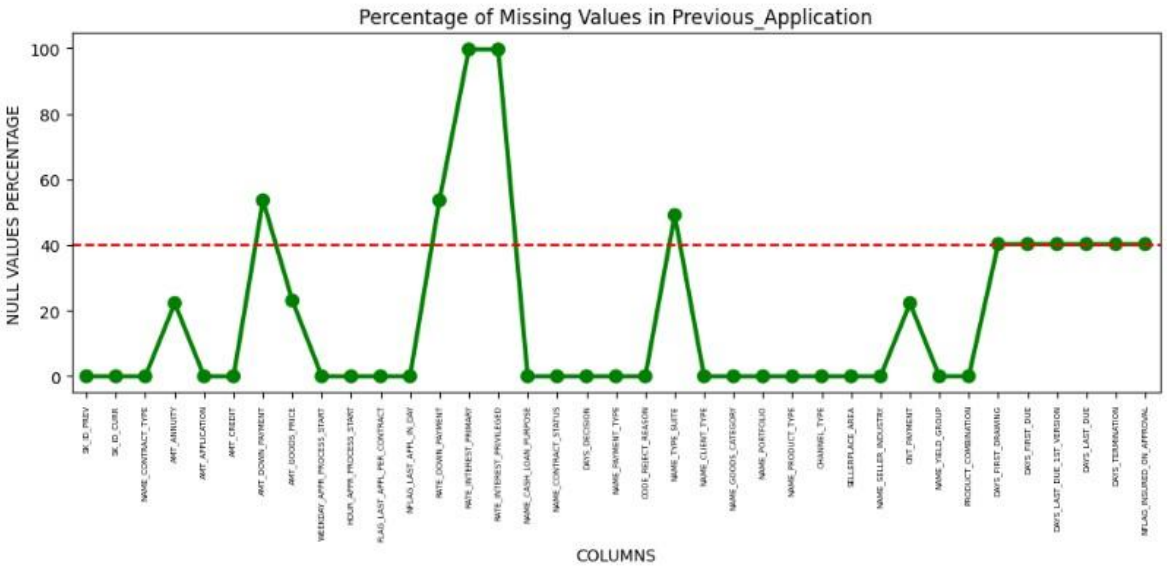
In Application Data:

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

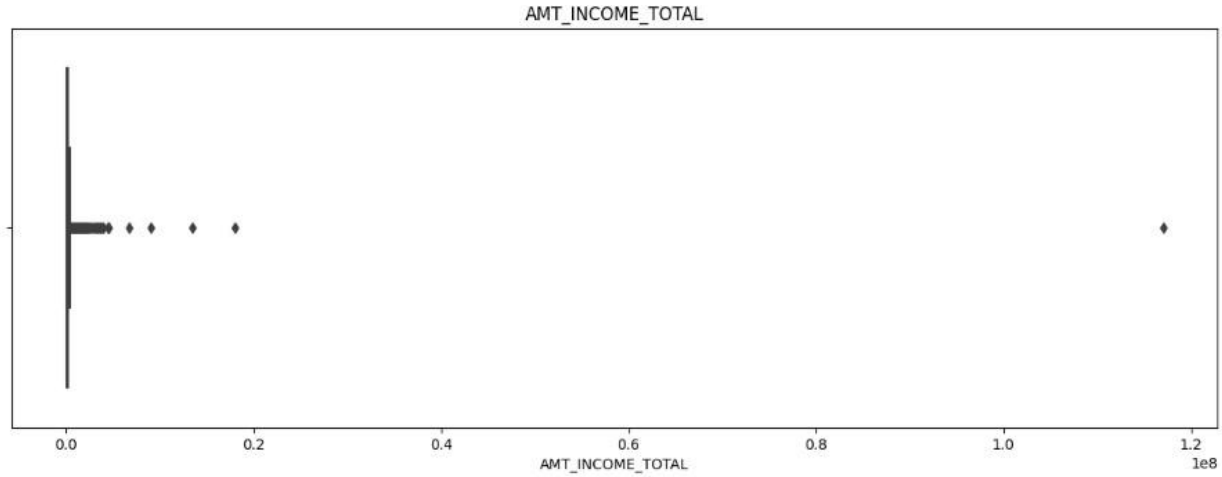


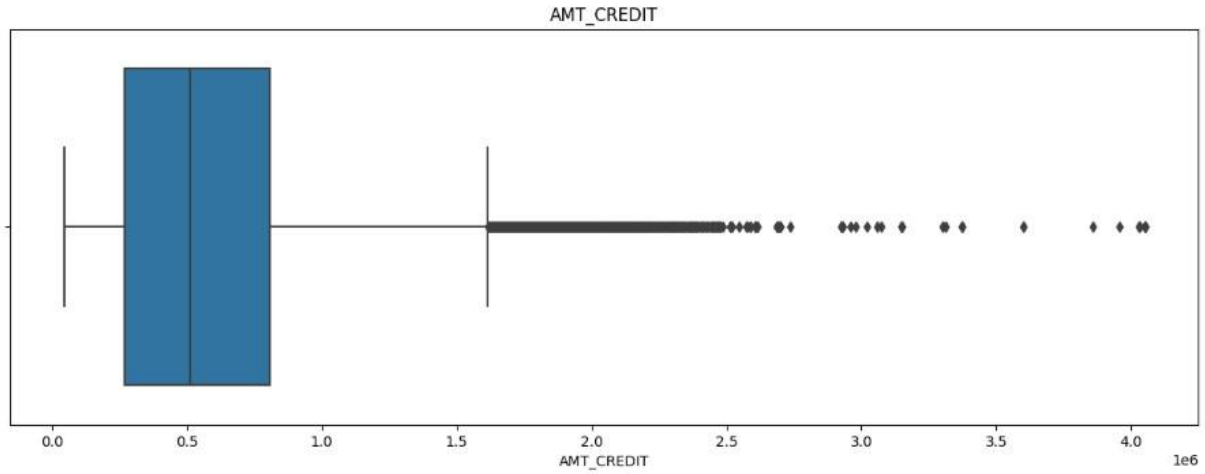
In Previous Application Data:

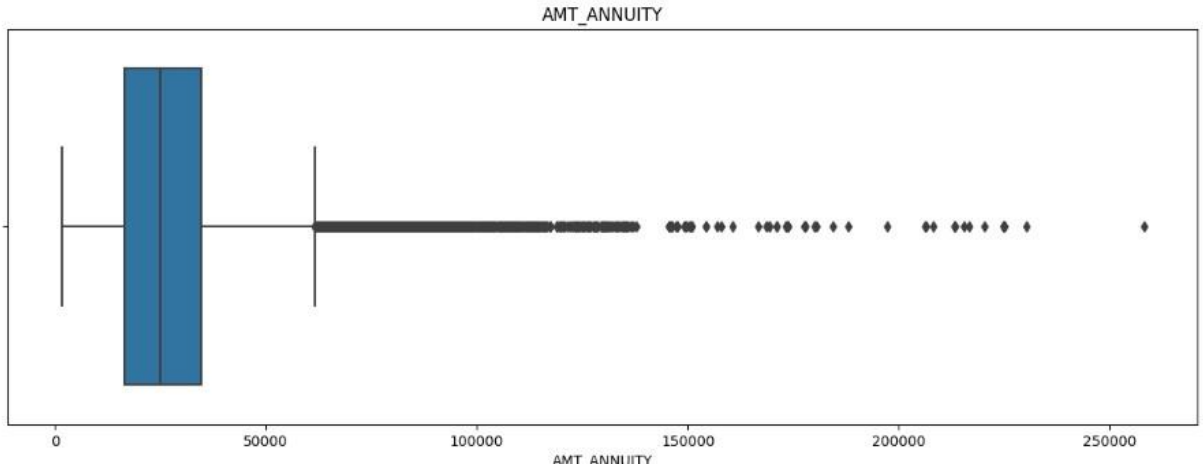
There are 11 columns where the outliers are more than 40%.

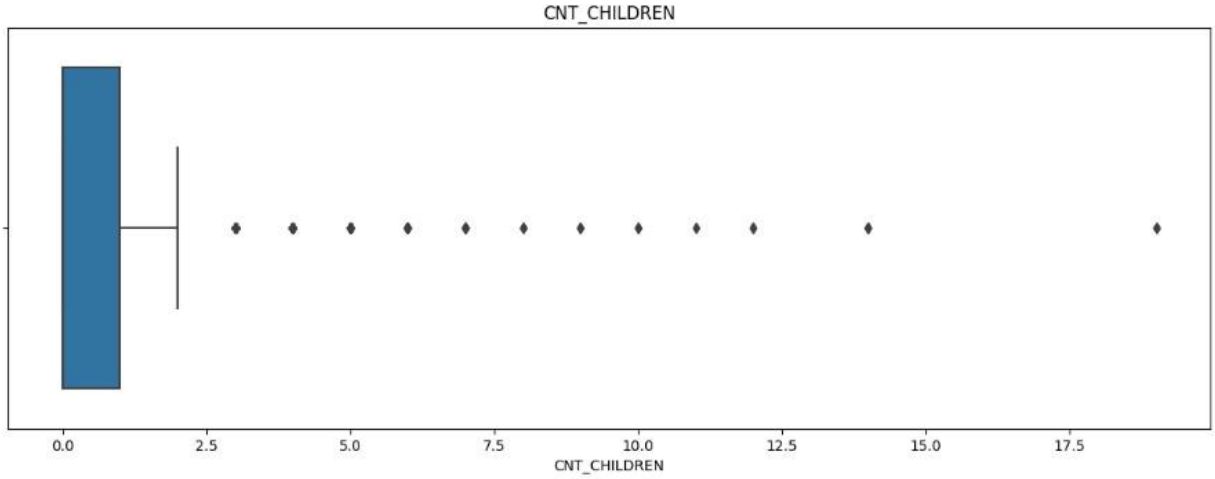


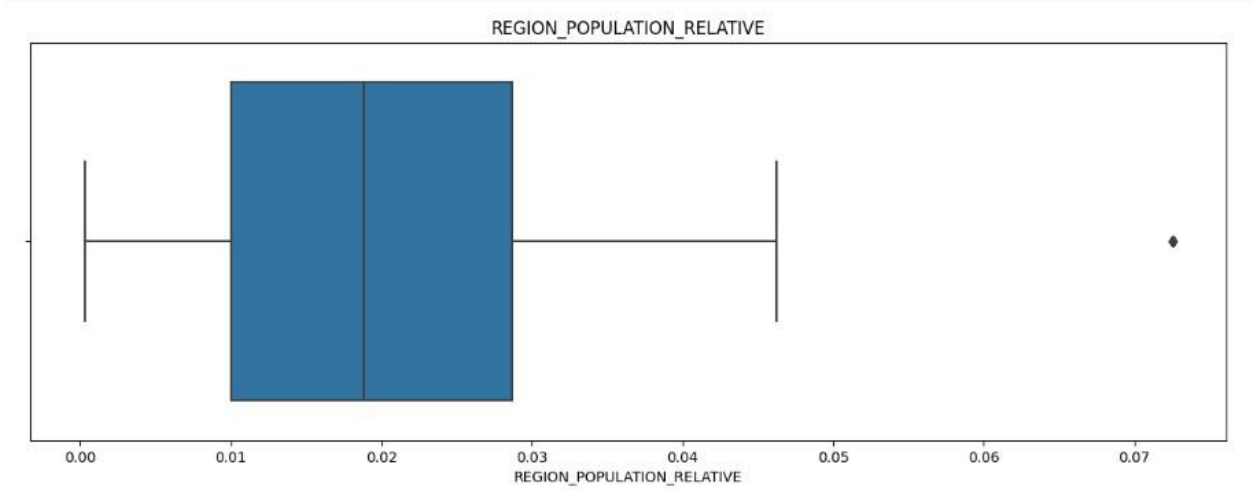
Outliers in Application Data

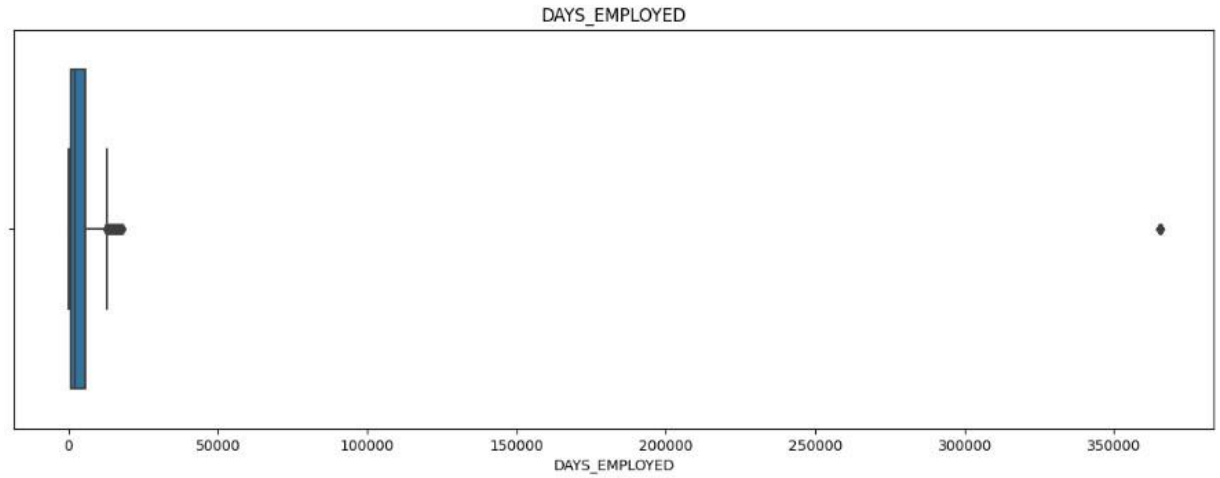


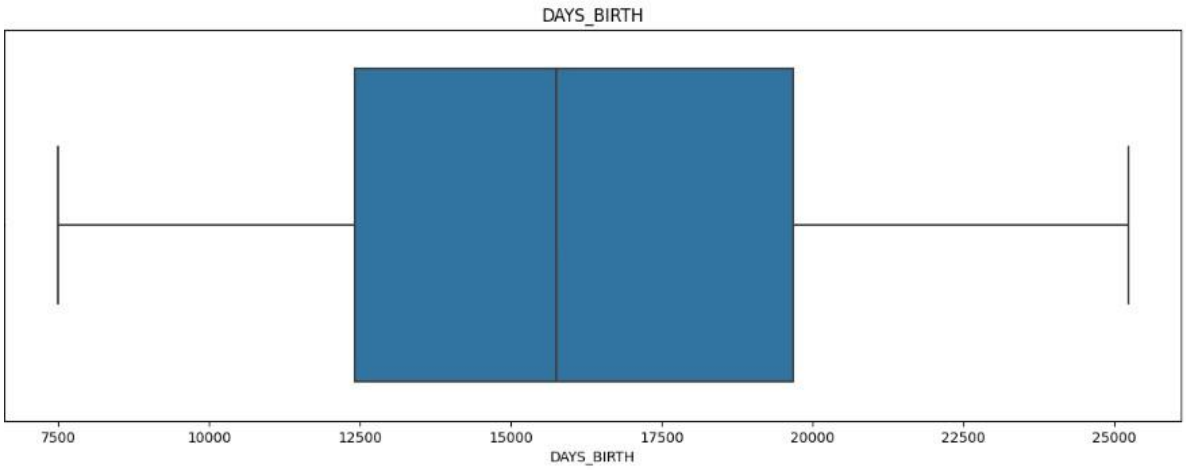

AMT_CREDIT
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
AMT_CREDIT
1e6


AMT_ANNUITY
0
50000
100000
150000
200000
250000
AMT_ANNUITY


CNT_CHILDREN
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
CNT_CHILDREN





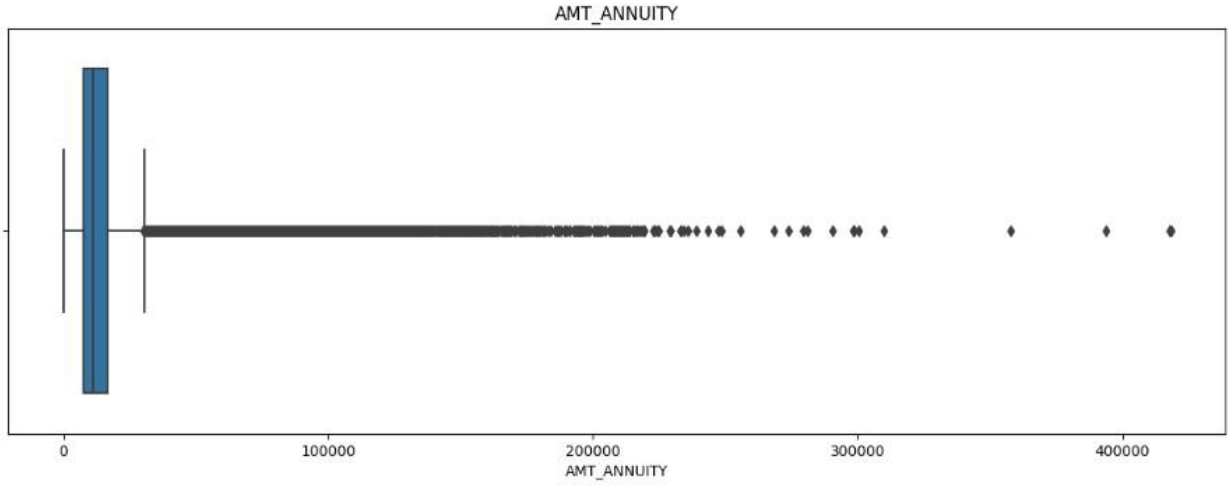


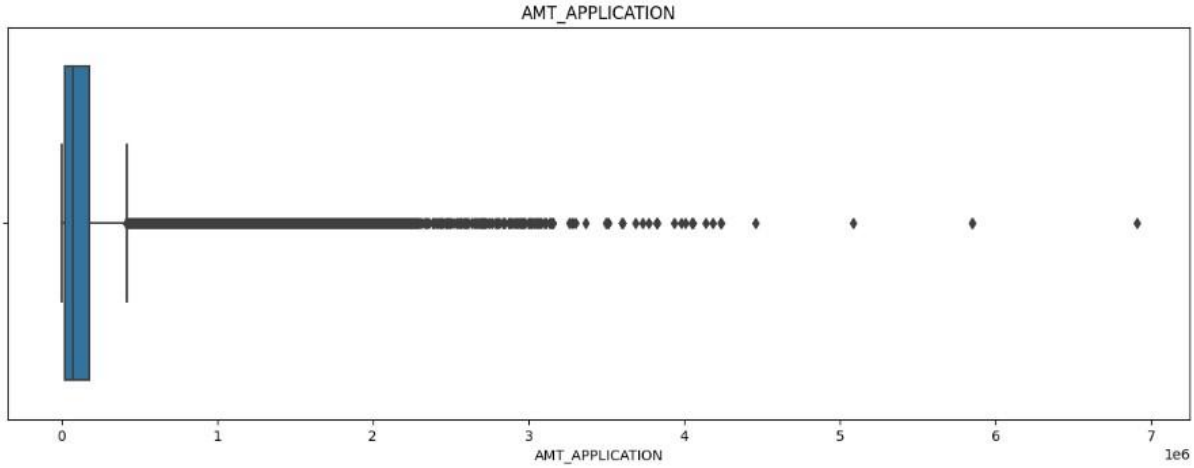
Insight:

1. AMT_ANNUITY, AMT_CREDIT, and CNT_CHILDREN have some number of outliers.

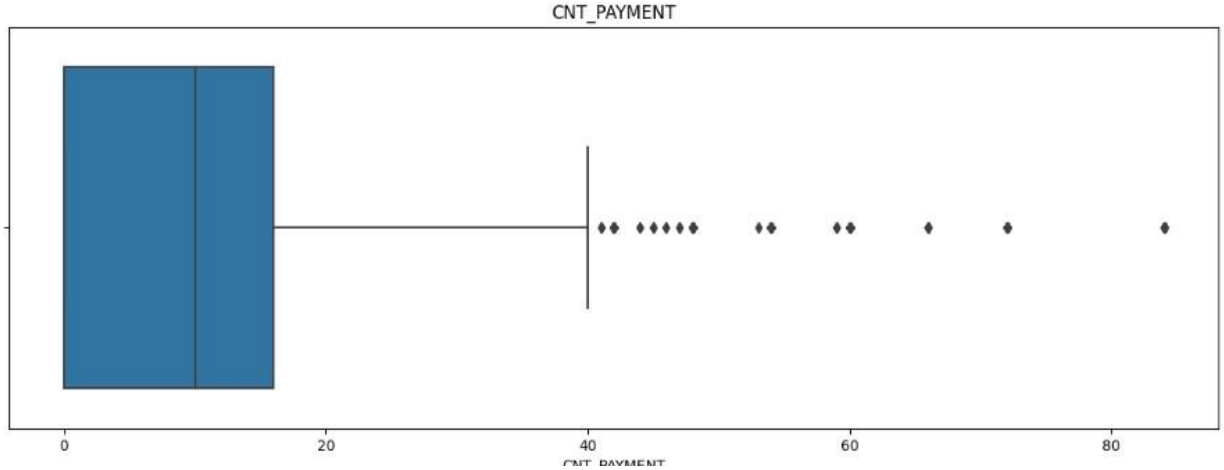
2. AMT_INCOME_TOTAL has a huge number of outliers which indicates that few of the loan applicants have high incomes compared to the others.
3. DAYS_BIRTH has no outliers which means the data available is reliable.
4. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which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

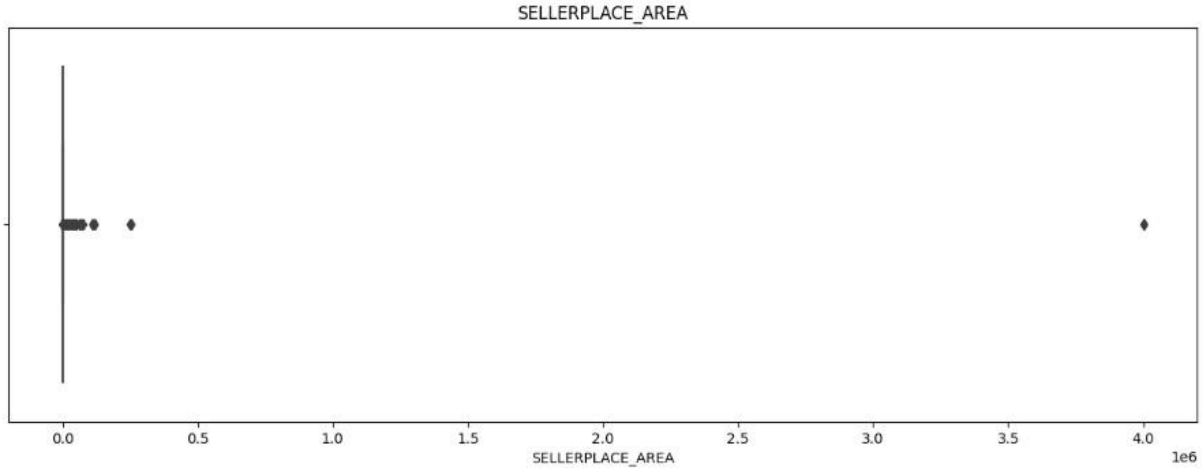
Outliers in Previous Data

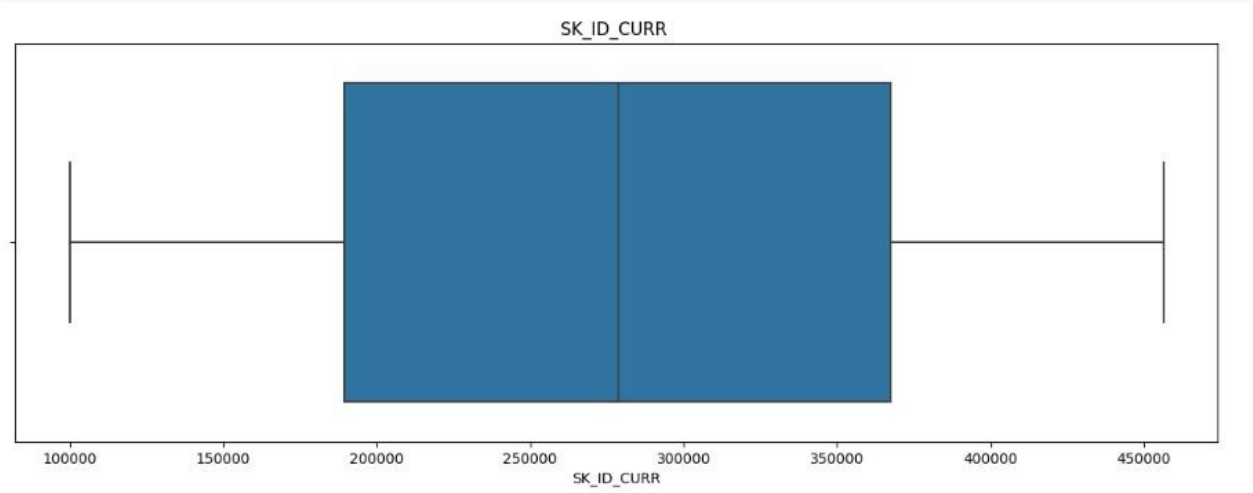


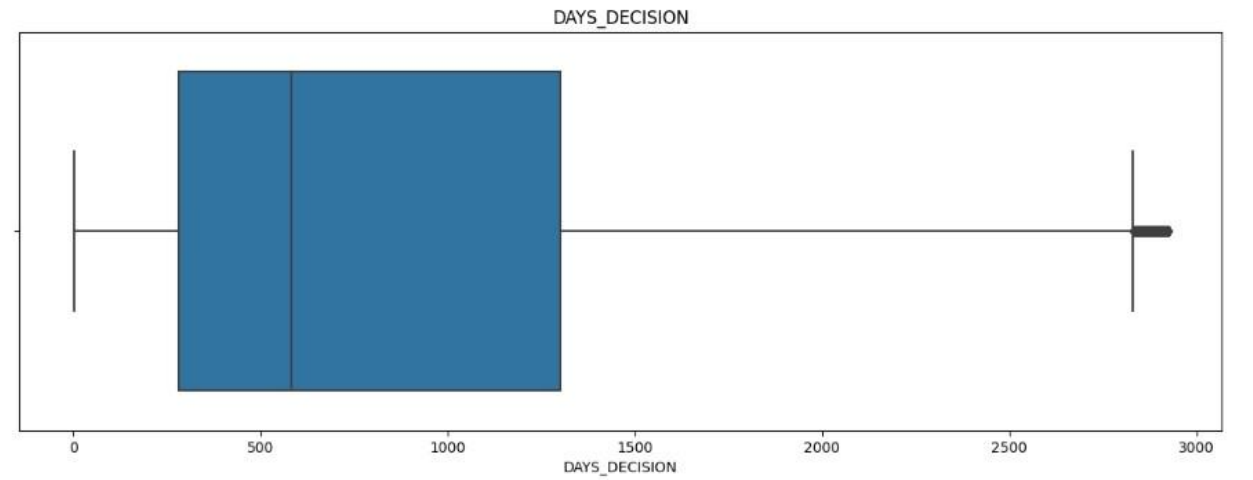



AMT_GOODS_PRICE
0
1
2
3
4
5
6
7
AMT_GOODS_PRICE
1e6


CNT_PAYMENT
0
20
40
60
80


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





	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	292779.762386	318574.557319	287141.316090	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:

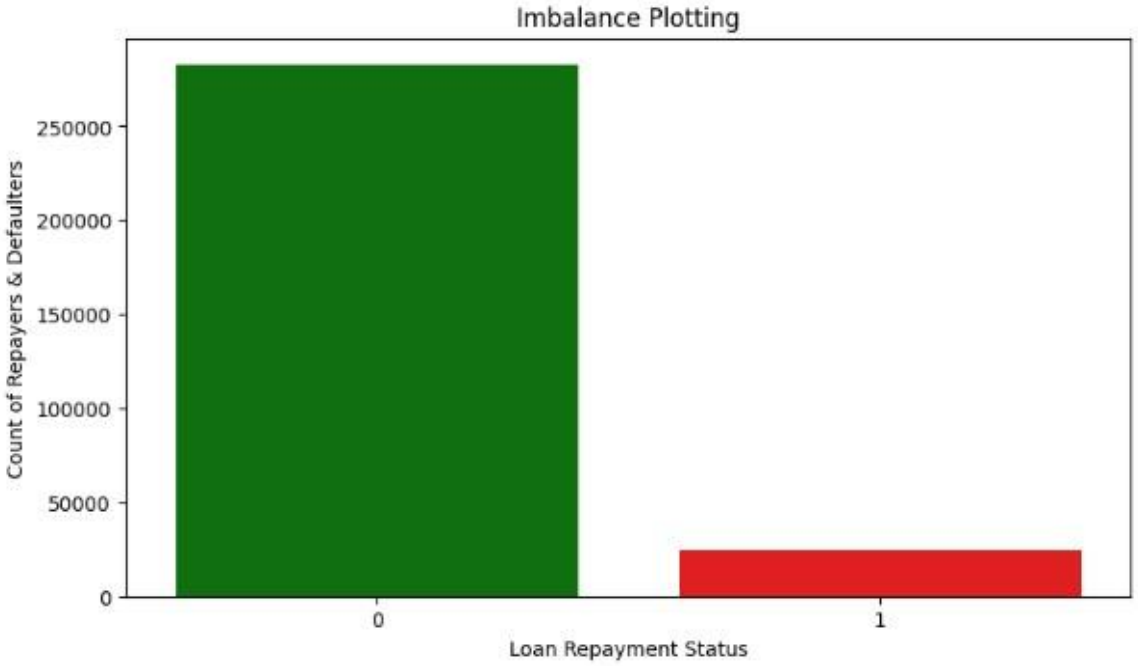
1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, and SELLERPLACE_AREA have a huge number of outliers.
2. CNT_PAYMENT has few outlier values.
3. SK_ID_CURR is an ID column and hence no outliers.

4. DAYS_DECISION has few numbers of outliers indicating that these previous application decisions were taken long back.

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

The result of data imbalance shows-

The ratio of data imbalance relative with respect to Repayor and Defaulter data is 11.39: 1.



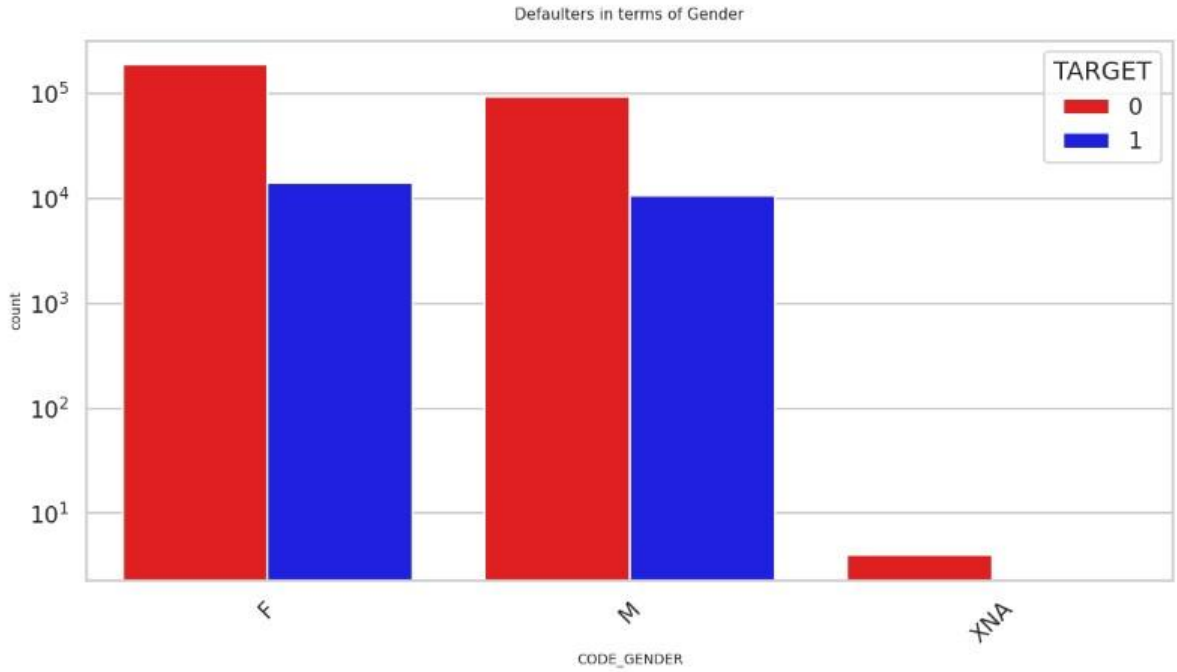
Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

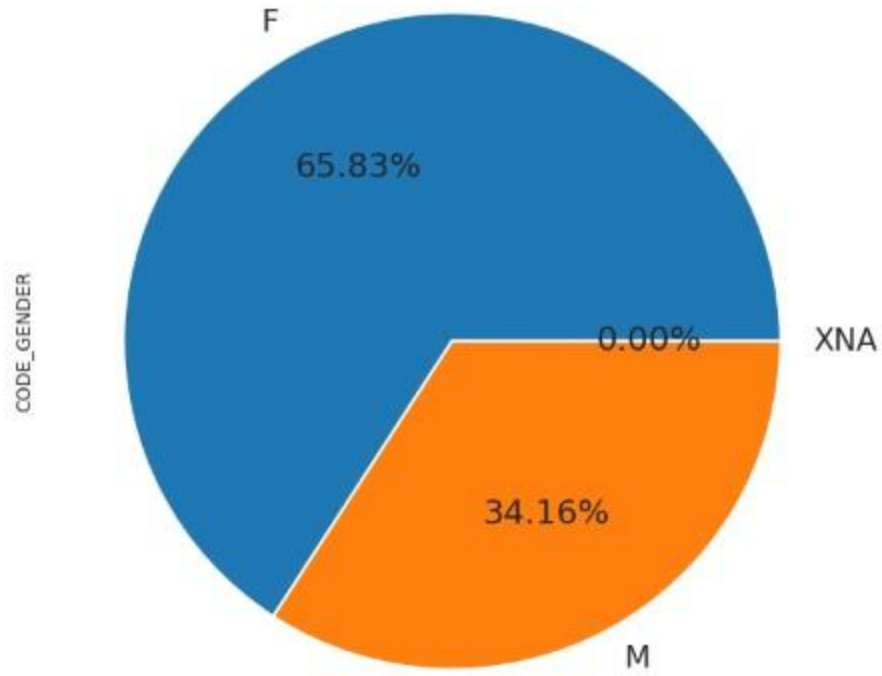
Univariate Analysis for Target 0

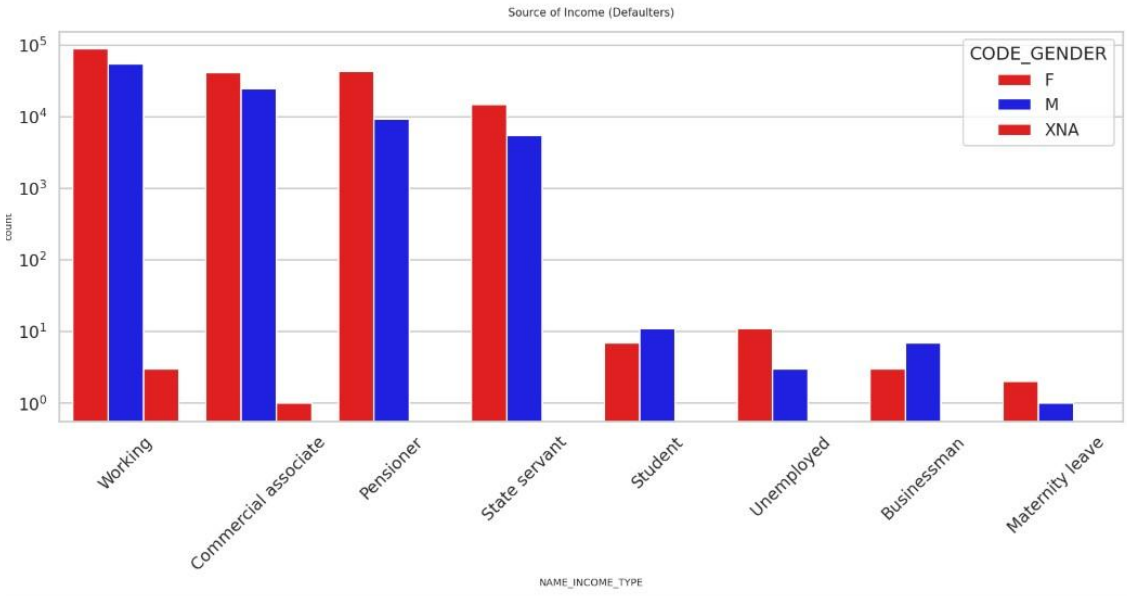
For target Variable 0(defaulters)

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 use loan money above 10 lakhs to 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.



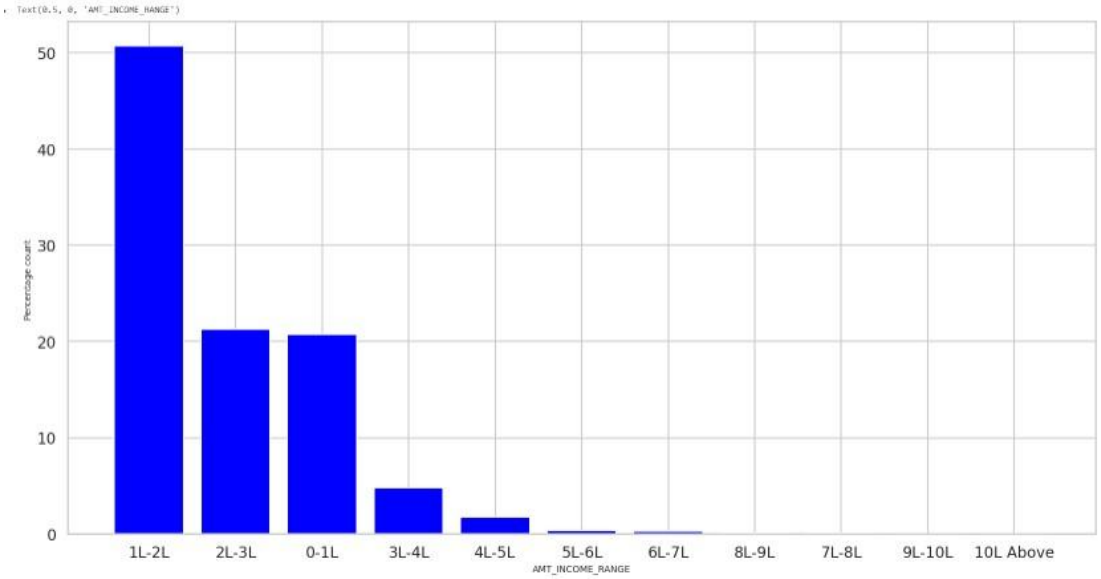




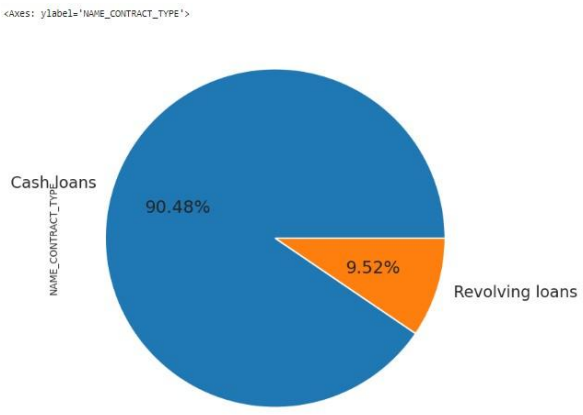
Insight:

1. Most Females have 'Working' as a source of income.
2. As compared to males, females are in larger numbers in terms of not repaying the loan.

3. Meanwhile, Commercial Associates (both female and male) have a great contribution to defaulters.
4. Maternity leave people have the least defaulter values.

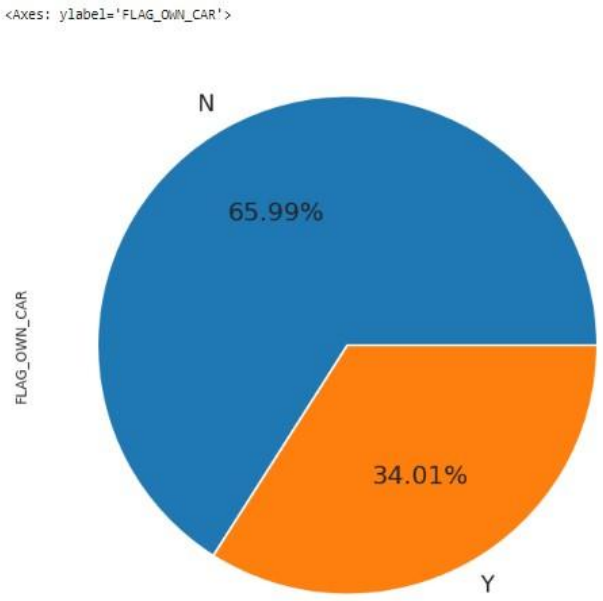


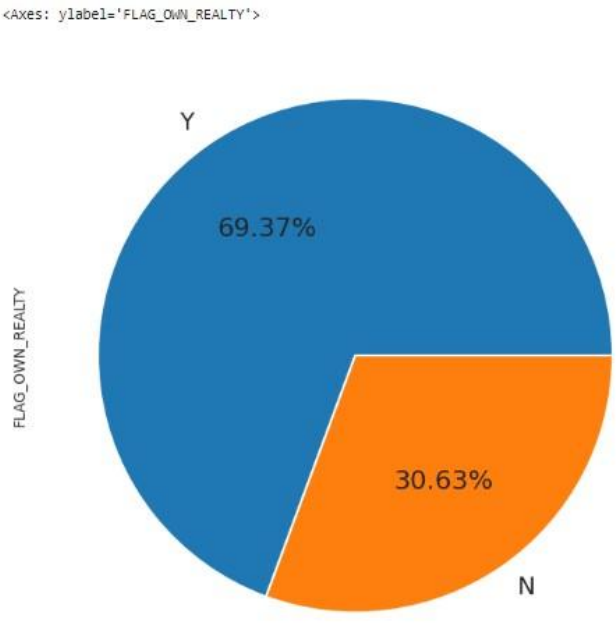

```
1L-2L        50.73
2L-3L        21.21
0-1L         20.73
3L-4L         4.78
4L-5L         1.74
5L-6L         0.36
6L-7L         0.28
8L-9L         0.10
7L-8L         0.05
9L-10L        0.01
10L Above     0.01
Name: AMT_INCOME_RANGE, dtype: float64
```



Insight

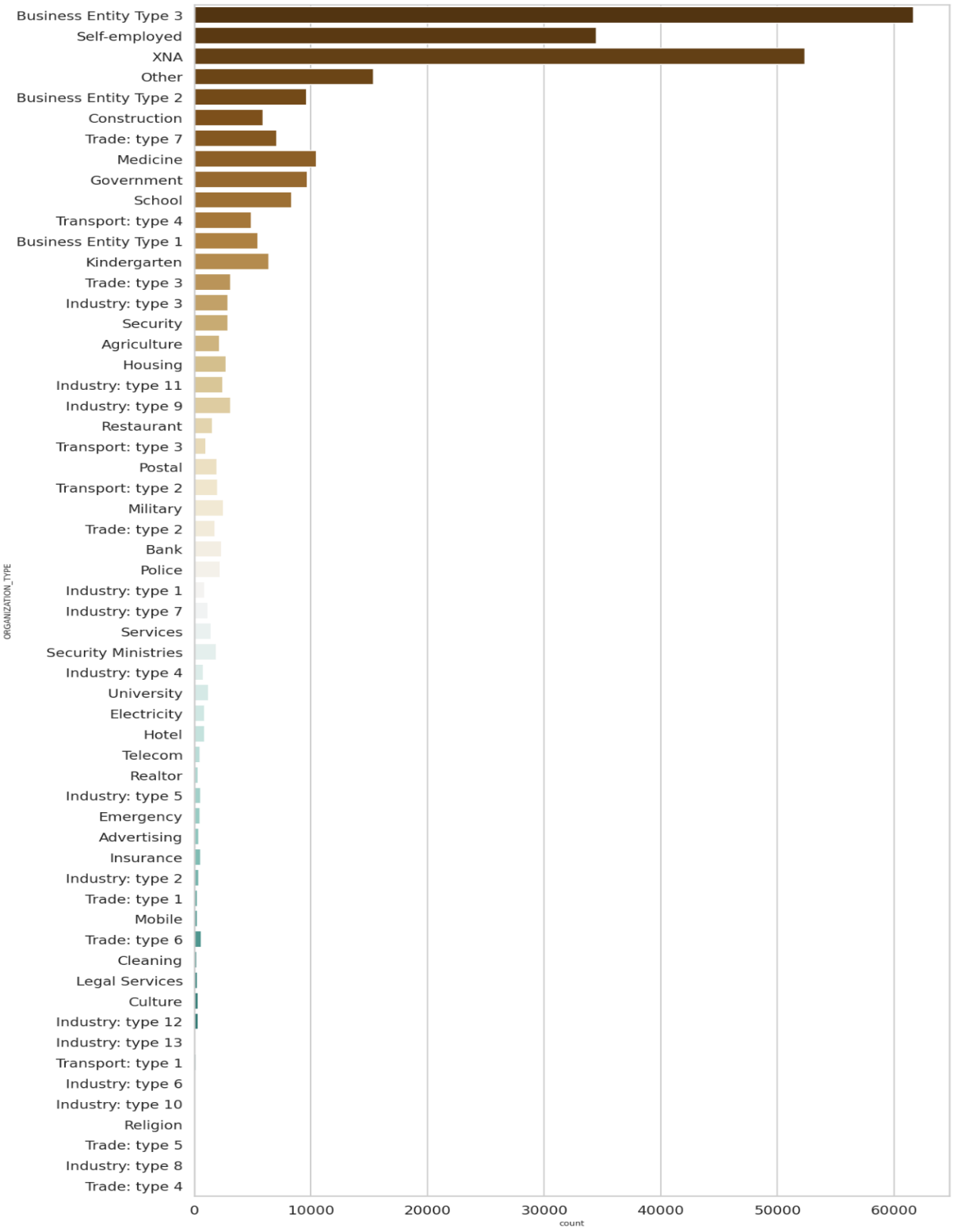
1. Hence as compared to revolving loans, cash loans are taken in bulk
2. Almost clients applied for cash loans later they turned into defaulters





Insights

1. Clients having no car have more defaulter rate
2. In case of not returning loans, clients have a house or flat in large numbers.

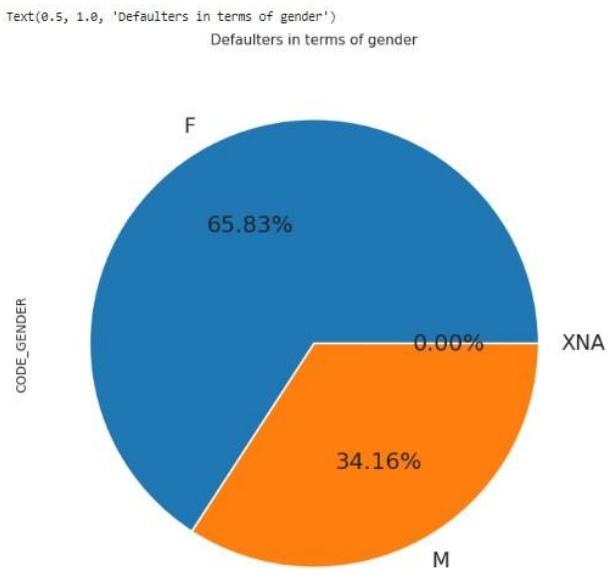


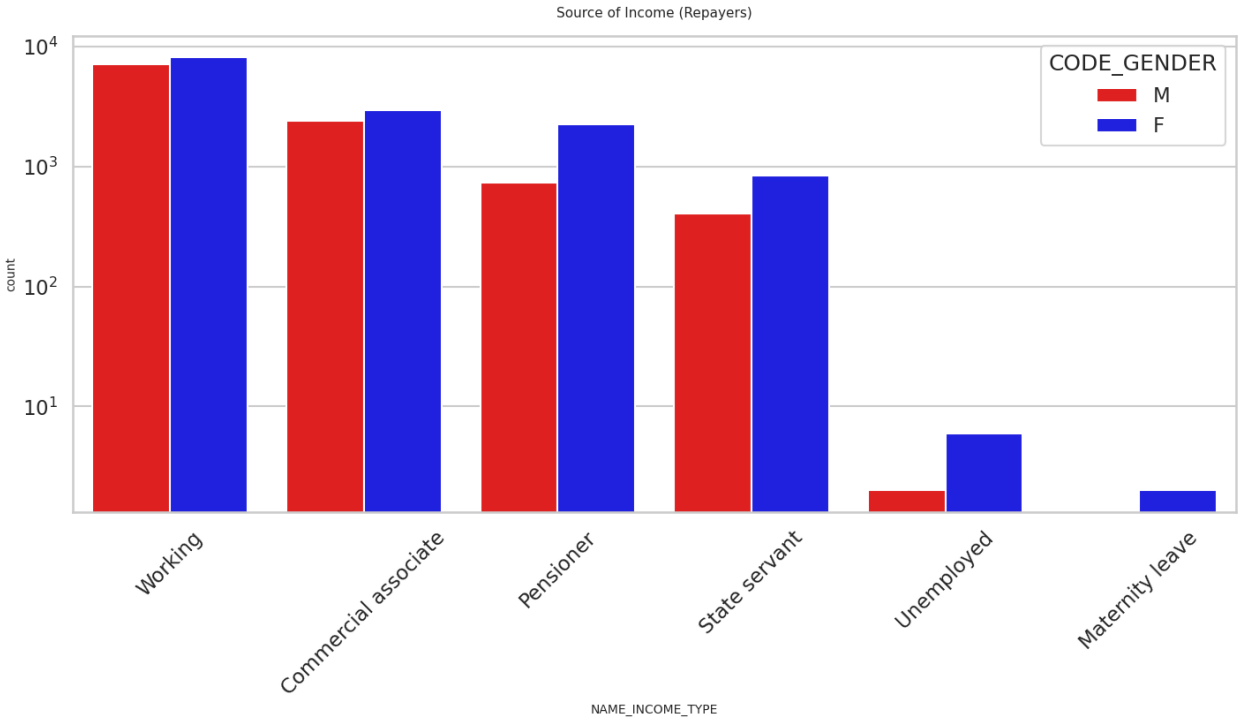
Insight

• Business Entity type 3 people have a big bar in defaulter cases followed by Self-Employed people and others.

For target Variable 1(Repayors)

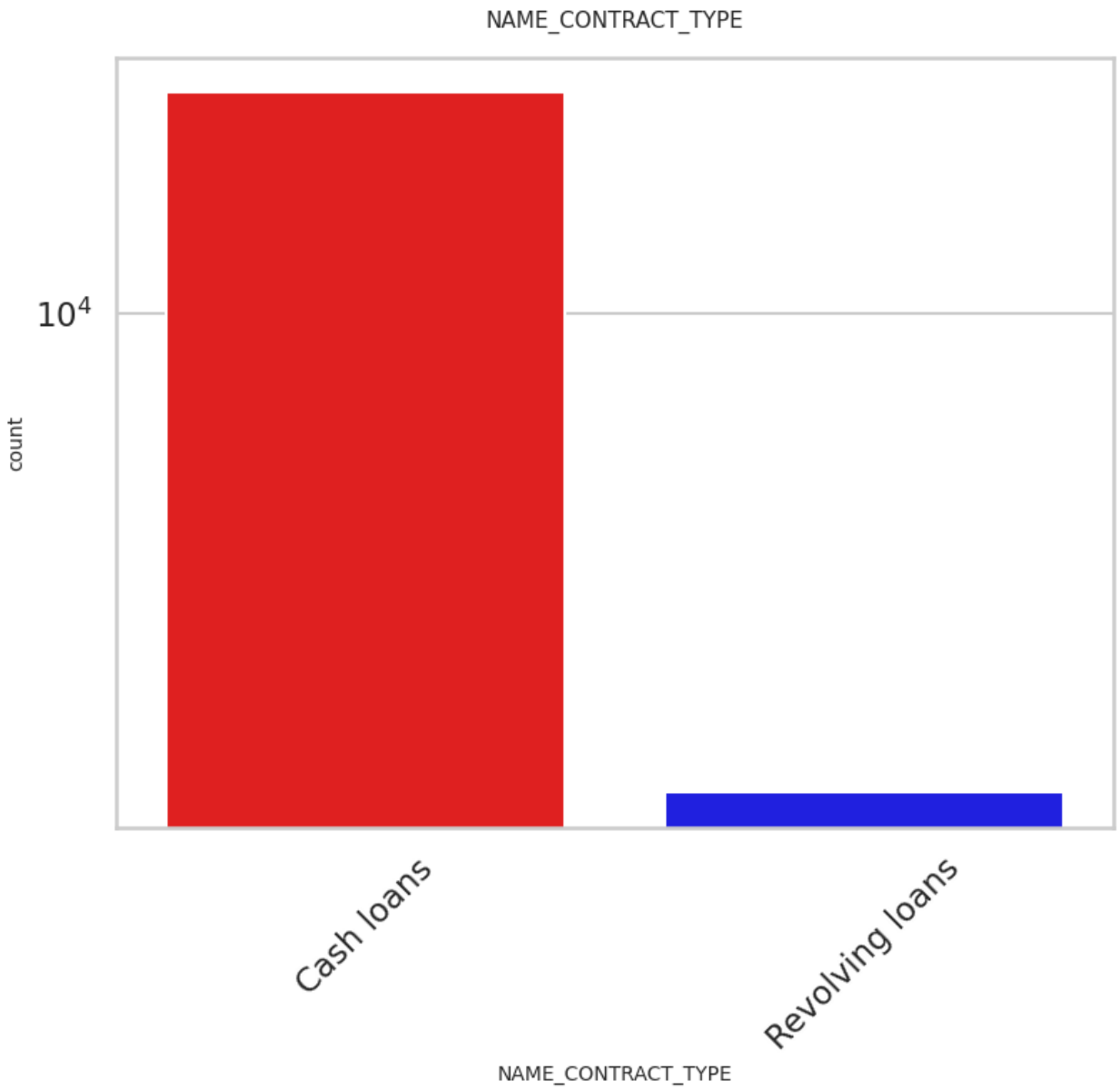
1. 57% of females are in the repayors counts which is less than females count in defaulters.
2. Here students and businessmen are missing whereas people with 'working' as the source of income are leading.
3. In terms of repaying loans, the percentage of people got increased from 50% in defaulter rate 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%).

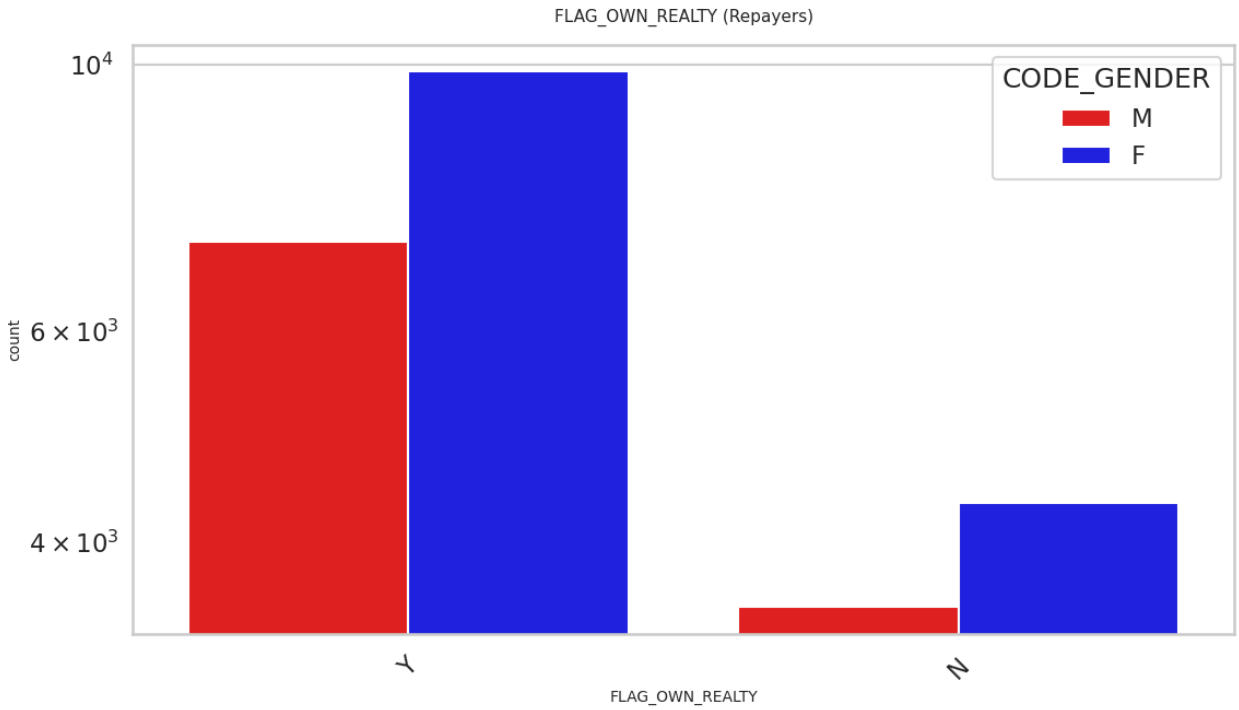




Insights

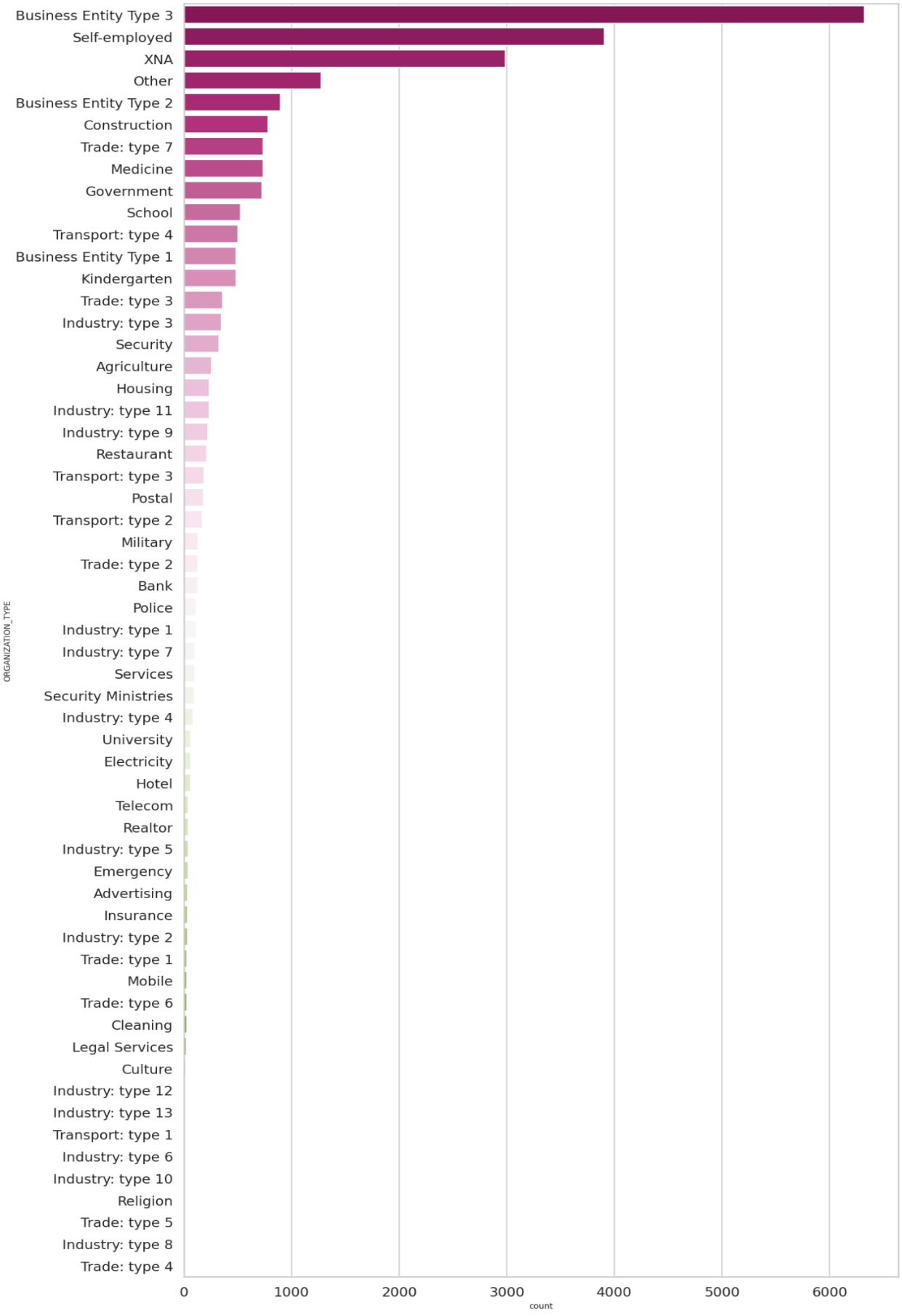
1. Here students and businessmen are missing whereas people with 'working' as a source of income are leading.
2. In terms of repaying loans, the percentage of people got increased from 50% in defaulter rate to 61% in repayment.


NAME_CONTRACT_TYPE
10^4
count
Cash loans
Revolving loans
NAME_CONTRACT_TYPE



Insights

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

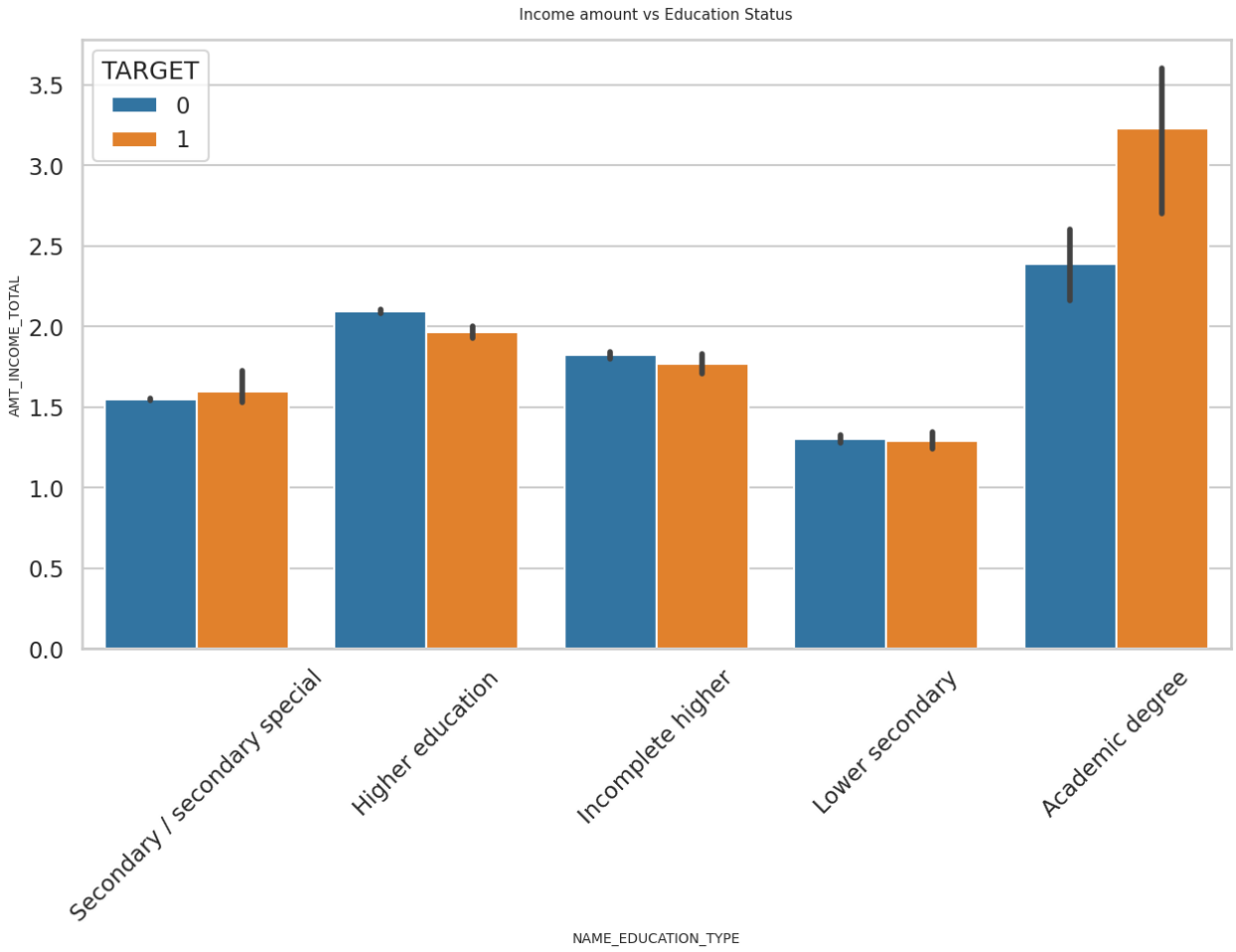

Business Entity Type 3
Self-employed
XNA
Other
Business Entity Type 2
Construction
Trade: type 7
Medicine
Government
School
Transport: type 4
Business Entity Type 1
Kindergarten
Trade: type 3
Industry: type 3
Security
Agriculture
Housing
Industry: type 11
Industry: type 9
Restaurant
Transport: type 3
Postal
Transport: type 2
Military
Trade: type 2
Bank
Police
Industry: type 1
Industry: type 7
Services
Security Ministries
Industry: type 4
University
Electricity
Hotel
Telecom
Realtor
Industry: type 5
Emergency
Advertising
Insurance
Industry: type 2
Trade: type 1
Mobile
Trade: type 6
Cleaning
Legal Services
Culture
Industry: type 12
Industry: type 13
Transport: type 1
Industry: type 6
Industry: type 10
Religion
Trade: type 5
Industry: type 8
Trade: type 4
ORGANIZATION_TYPE
0
1000
2000
3000
4000
5000
6000
count

Insights

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 numbers (68%).

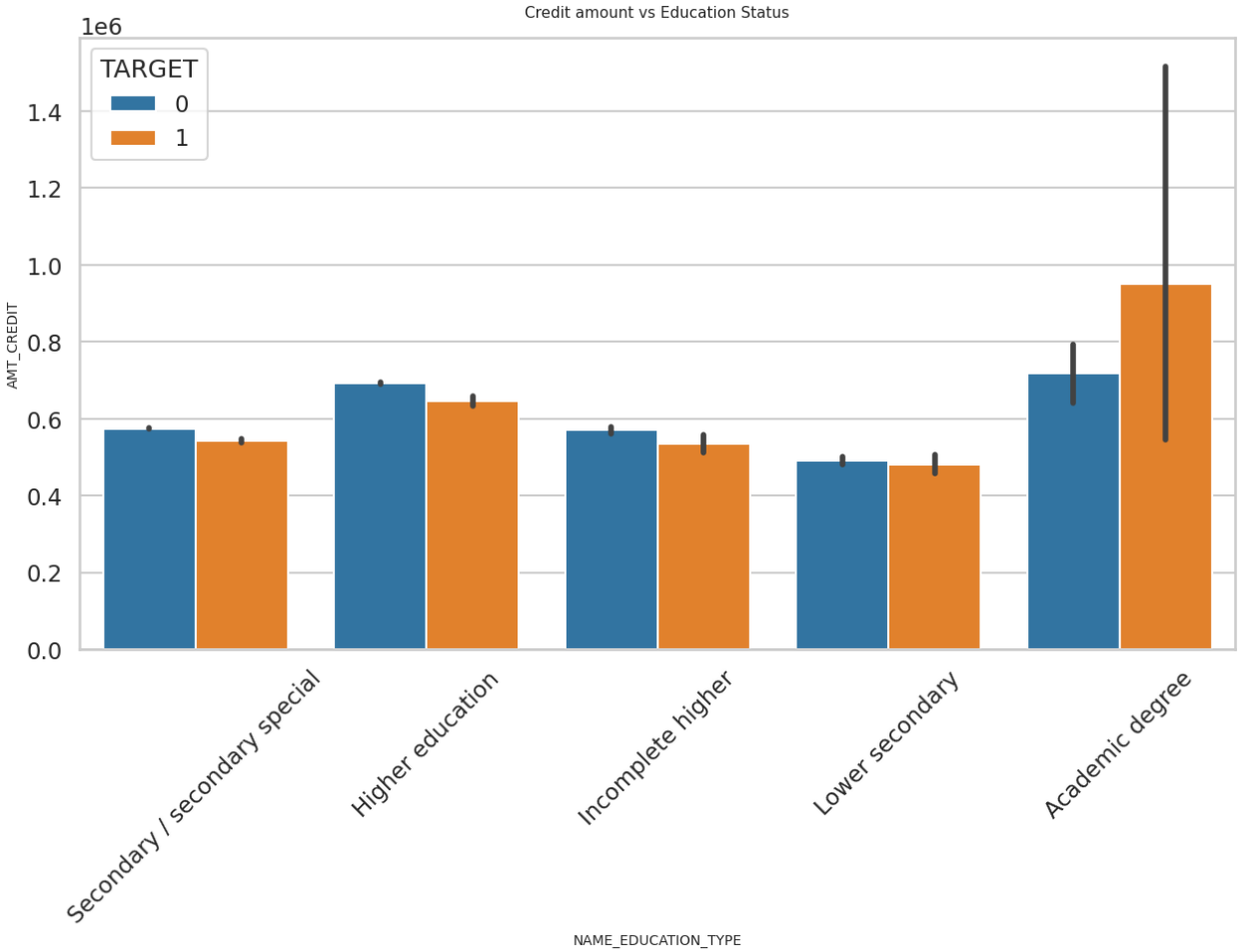
Result of bivariate analysis:

1. Clients with academic degrees have the most income and 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.

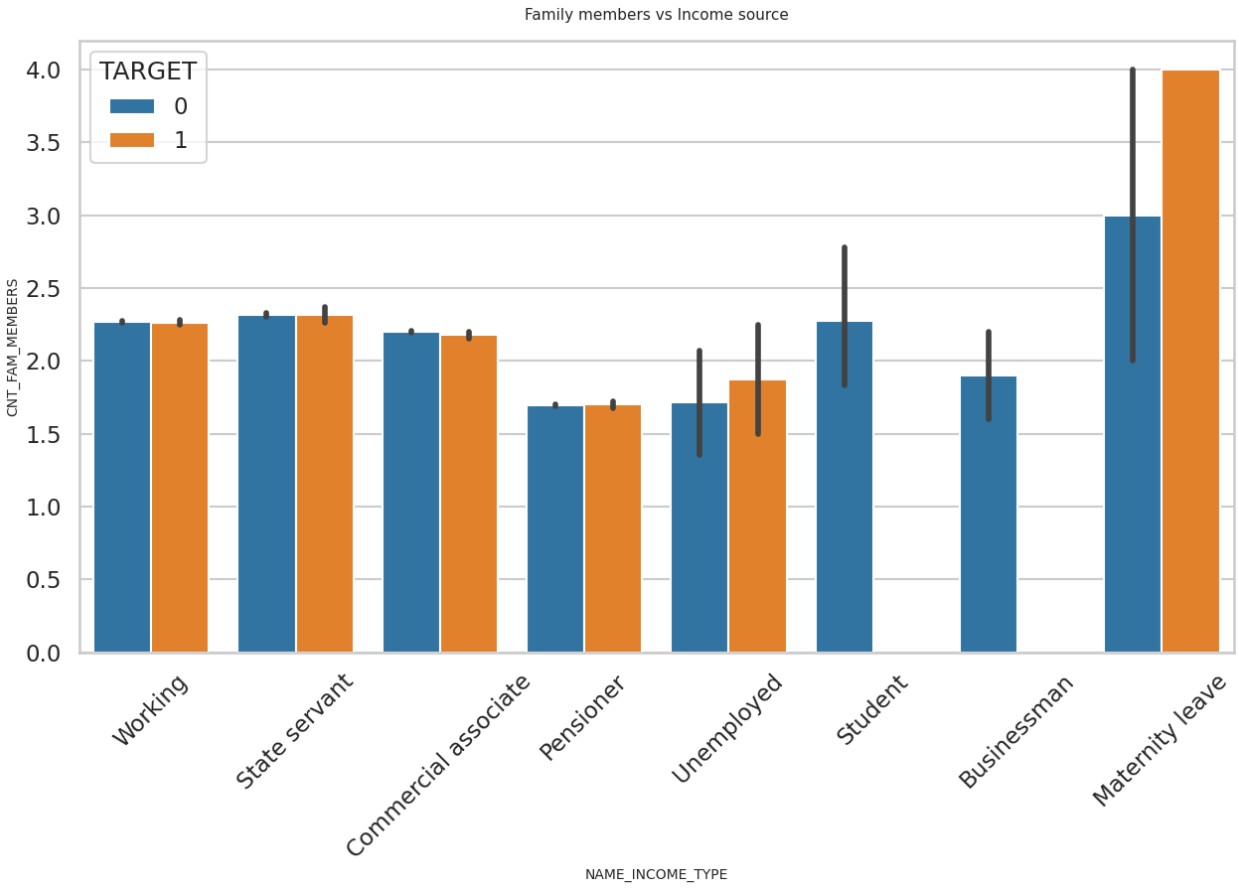


Insights

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

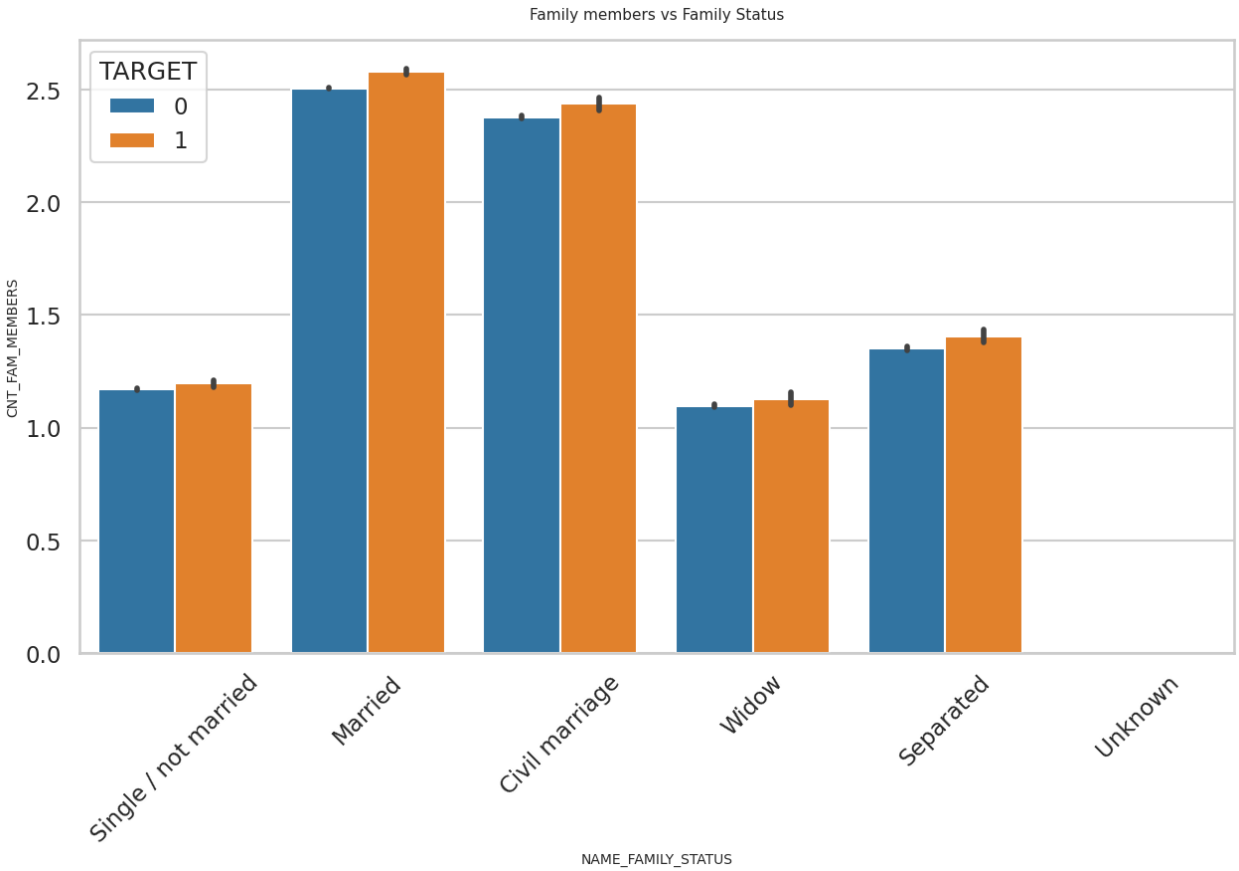


Academic degree people are on the safe side of giving loans as they have credited a huge amount and they have returned less than the credited amount to the company.



Insights

1. Working people having 2 family members are an equal number of repayers and defaulters.
2. People who get income through Maternity Leave tend to be more Defaulter when they have more Family Members.



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

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)

SK_ID_CURR	SK_ID_CURR	1.000000
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_CREDIT	0.752195
DAYS_EMPLOYED	DAYS_BIRTH	0.582185
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937
REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.472052
REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.322628
DAYS_REGISTRATION	DAYS_BIRTH	0.289114
DAYS_BIRTH	DAYS_ID_PUBLISH	0.252863

dtype: float64

Top 10 correlations for non-defaulters

SK_ID_CURR	SK_ID_CURR	1.000000
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878571
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
AMT_CREDIT	AMT_ANNUITY	0.771297
DAYS_EMPLOYED	DAYS_BIRTH	0.626114
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.446101
REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.435514
AMT_INCOME_TOTAL	AMT_ANNUITY	0.418948
AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.341571

dtype: float64

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

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. Companies 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 the most counts in terms of not paying loan status.
15. The company should pay attention to clients whose previous application was for POS, Cash, etc because they are also contributing to the defaulters counts.

Learning

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 extracting useful 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.