



Bank Loan Case Study

By Darshan Hedgire

Project Description:

This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, we will develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

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.

➤ How are we going to handle the things ?

Suppose we work for a consumer finance company which specialises in lending various types of loans to urban customers. We will 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 has to 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, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

➤ **Following are the things that we are going to find out through this case study:**

- Our aim is to identify the patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- The **driving factors** (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- Presenting the overall approach of the **data analysis, cleaning the dataset, finding outliers, data imbalance, univariate, segmented univariate, bivariate analysis**, etc.
- The **top 10 correlation for the Client with payment difficulties** and all **other cases** (Target variable).

Tech-Stack Used:

- **Microsoft Excel 2016:** It enables users to format, organize and calculate data in a spreadsheet. It organizes data in an easy-to-navigate way. It has been used to have an overall look of the data and for understanding the different column descriptions
- **Jupyter Notebook:** It is used for the data cleaning and imputing the data. As the dataset was very large, so it is used for the whole data analysis purpose, visualizing the data and summarizing it to get the necessary insights for the client.
- **Python Programming (Version 3.8):** For the data analysis, python is the best and the easiest to use programming language.
- **Microsoft Word 2016:** It is used to make a report (PDF) to be presented to the leadership team.

Approach:

Problem Statement:

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicant's using EDA is the aim of this case study.

The dataset given by the client contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

1. **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample
2. **All other cases:** All other cases when the payment is paid on time.

We will use **EDA** to understand how consumer attributes and loan attributes influence the tendency of default.

Analysis Approach:

Following steps done for the analysis purpose:

1. Imported the NumPy, pandas, matplotlib and seaborn python libraries.
2. Imported the datasets (Application_Data & Previous_Application)
3. Identification: We have identified how we will approach the data, finding missing dataset and working on it accordingly to gain the required results.
4. Outliers: Identified outliers and showed how they play if any role in our dataset.
5. Imbalance: Understanding the ratio of imbalance in our data.
6. Correlation Analysis: Finding the correlation between the variables with respect to the target variables and find the top three correlation.
7. Visualisation: Visualized the data with the help of charts and graphs.

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

1. There are total of 49 columns in Application_Data and 11 columns in Previous_Application which have missing values greater than 40%.
2. On further analysis, we found that "EXT_SOURCE_2","EXT_SOURCE_3" has no correlation with the "TARGET" column.
3. There is almost no correlation of 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL' with the "TARGET" column.
4. 'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START', 'FLAG_LAST_APPL_PER_CONTRACT', 'NFLAG_LAST_APPL_IN_DAY' are the column in the Previous_Application which are not needed for the analysis.
5. Dropping all the above mention columns which will total 76 in Application_Data and 15 in Previous_Application.

6. Converting the negative days column into positive days.

7. Imputing the remaining null values columns needed for data analysis with mean, median (numerical data) and mode (categorical data).

8. Imputed categorical variable 'OCCUPATION_TYPE' by adding an 'Unknown' category, numerical variables

'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' with median.

9. Imputed AMT_ANNUITY with median, AMT_GOODS_PRICE with mode, CNT_PAYMENT with 0 as the NAME_CONTRACT_STATUS for these indicate that most of these loans were not started.

• **Identify if there are outliers in the dataset. Also, mention why do 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. An outlier can be identified from a box-plot graph. If the value lies above maximum and below minimum, they are considered as outliers.

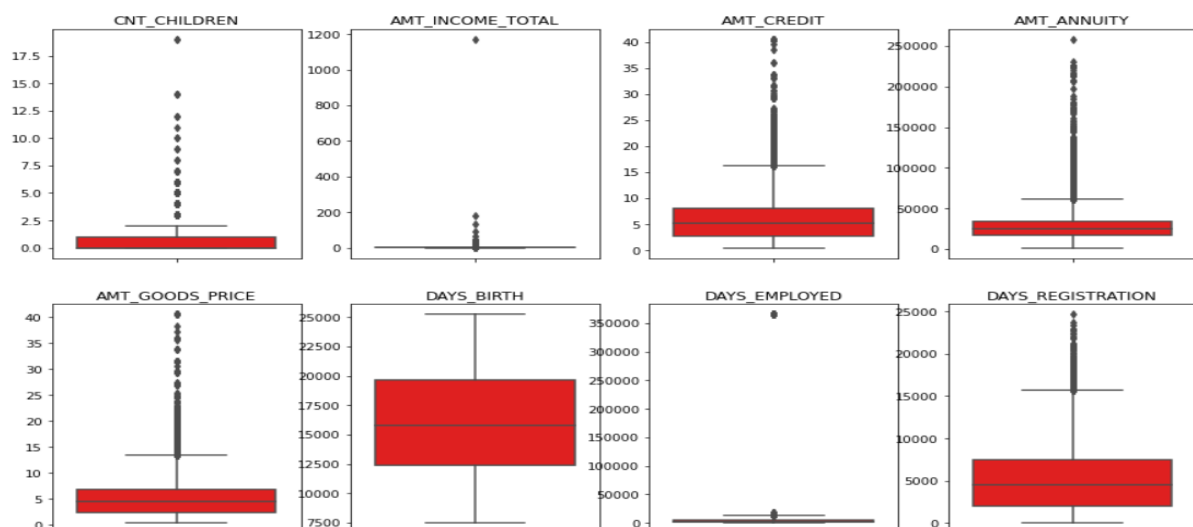
Application_Data:

1. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, 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.

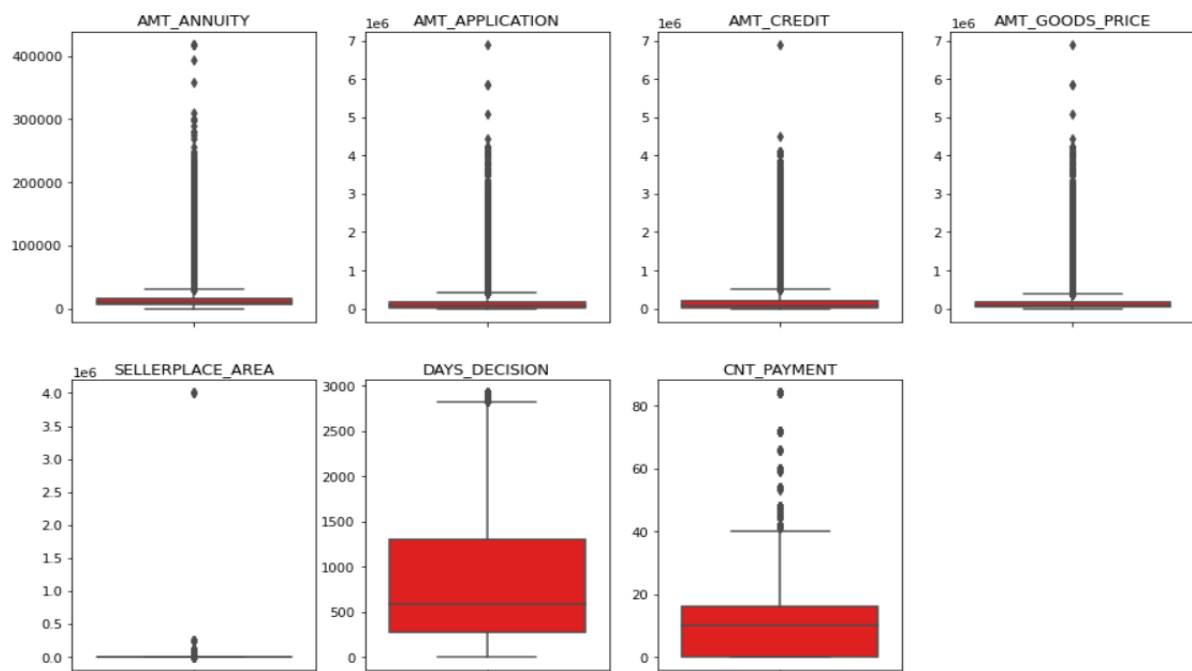
3. DAYS_BIRTH has no outliers which means the data available is reliable.

4. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.



Previous_Application:

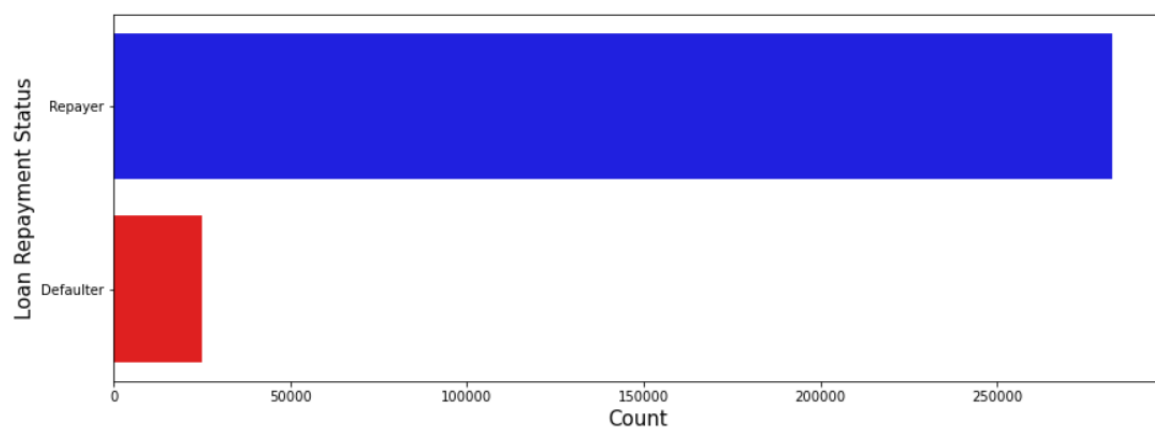
1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
2. CNT_PAYMENT has few outlier values.
3. SK_ID_CURR is an ID column and hence no outliers.
4. DAYS_DECISION has little number of outliers indicating that these previous applications decisions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



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

- This data is highly imbalanced as number of defaulter is very less in total population.
Data Imbalance Ratio with respect to Repayment and Default: 11.39 : 1 (approx.)

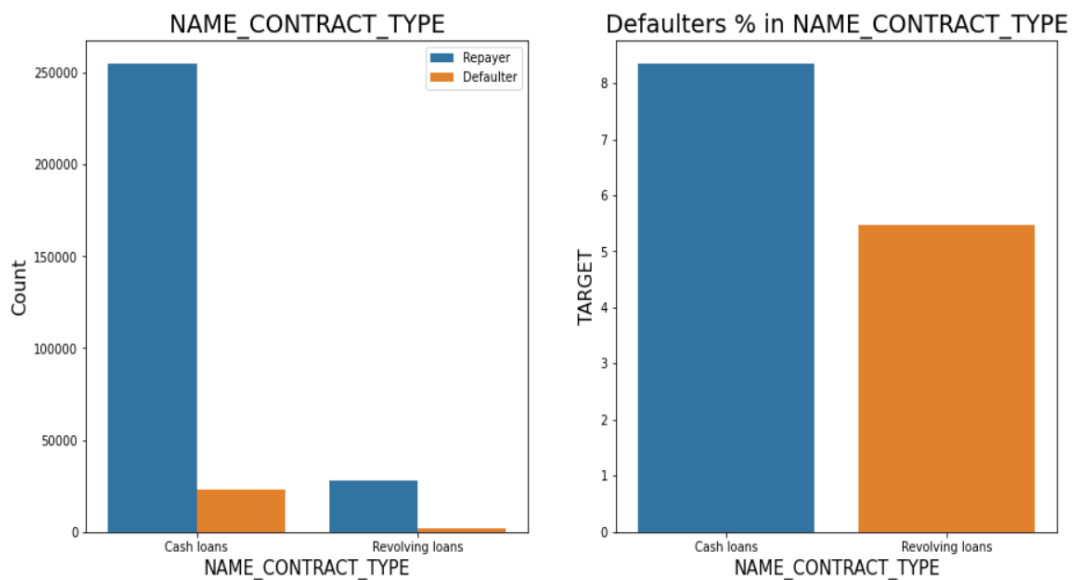
Imbalance Plotting (Repayer Vs Defaulter)



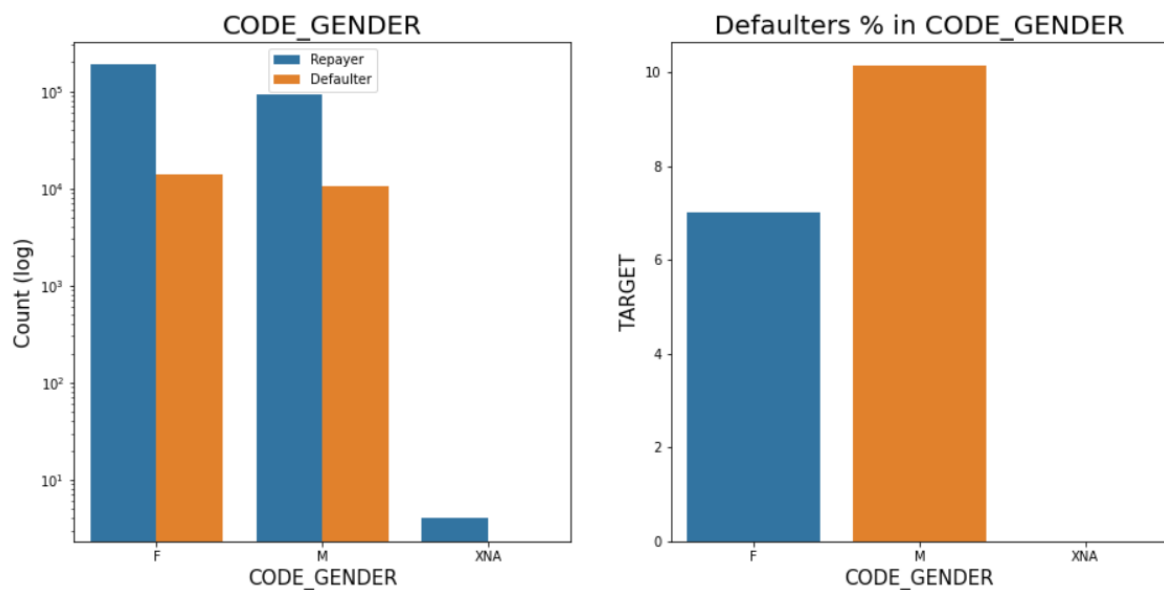
- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

Categorical variable analysis (Segmented Univariate Analysis) :

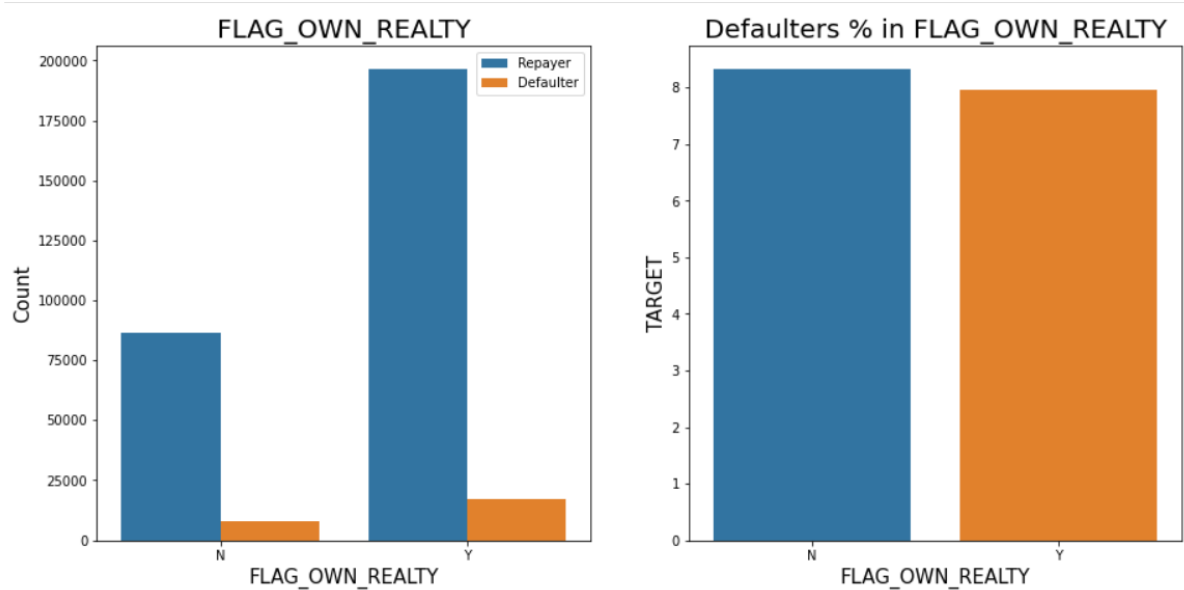
- Revolving loans are just a small fraction (10%) from the total number of loans. Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters



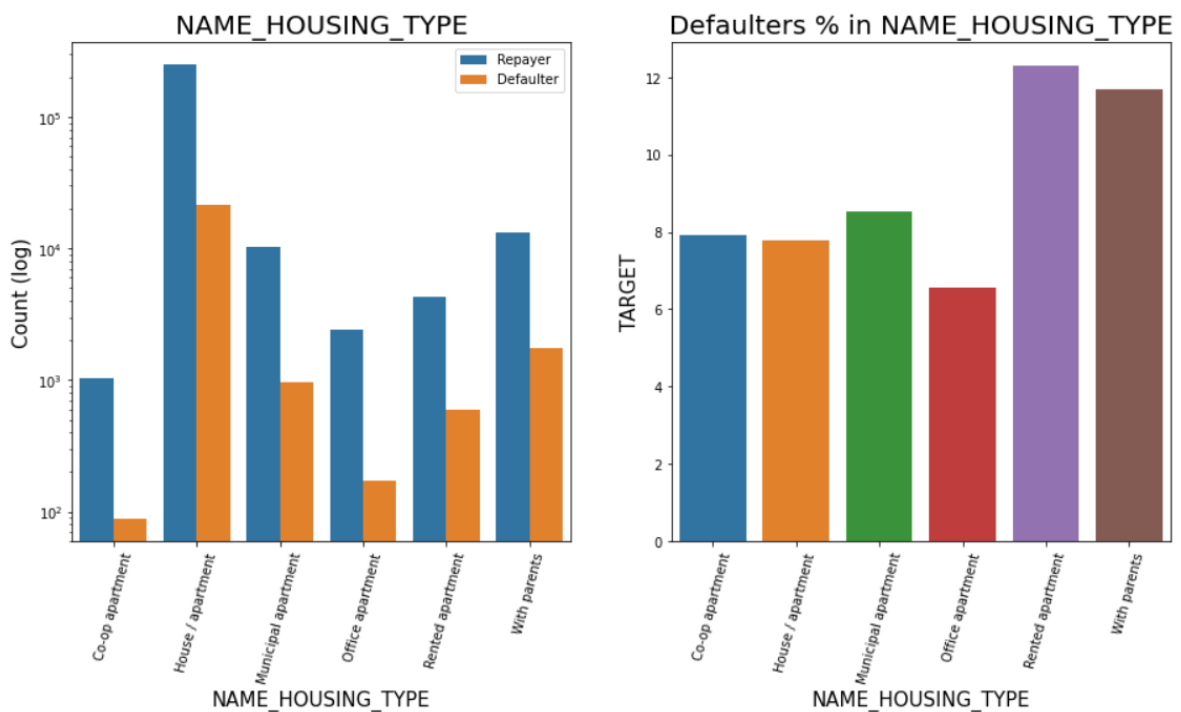
- The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (7%).



- The clients who own real estate are more than double of the ones that don't own. But the defaulting rate of both categories are around the same (~8%). Thus there is no correlation between owning a reality and defaulting the loan.

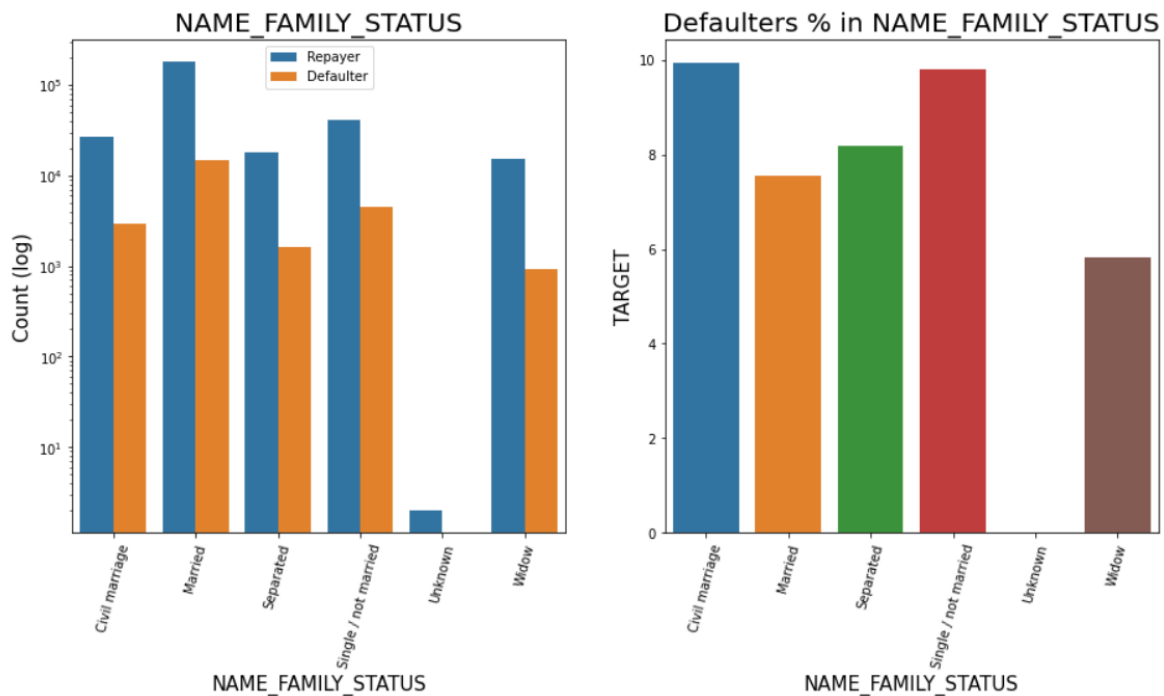


- Majority of people live in House/apartment.
- People living in office apartments have lowest default rate.
- People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting

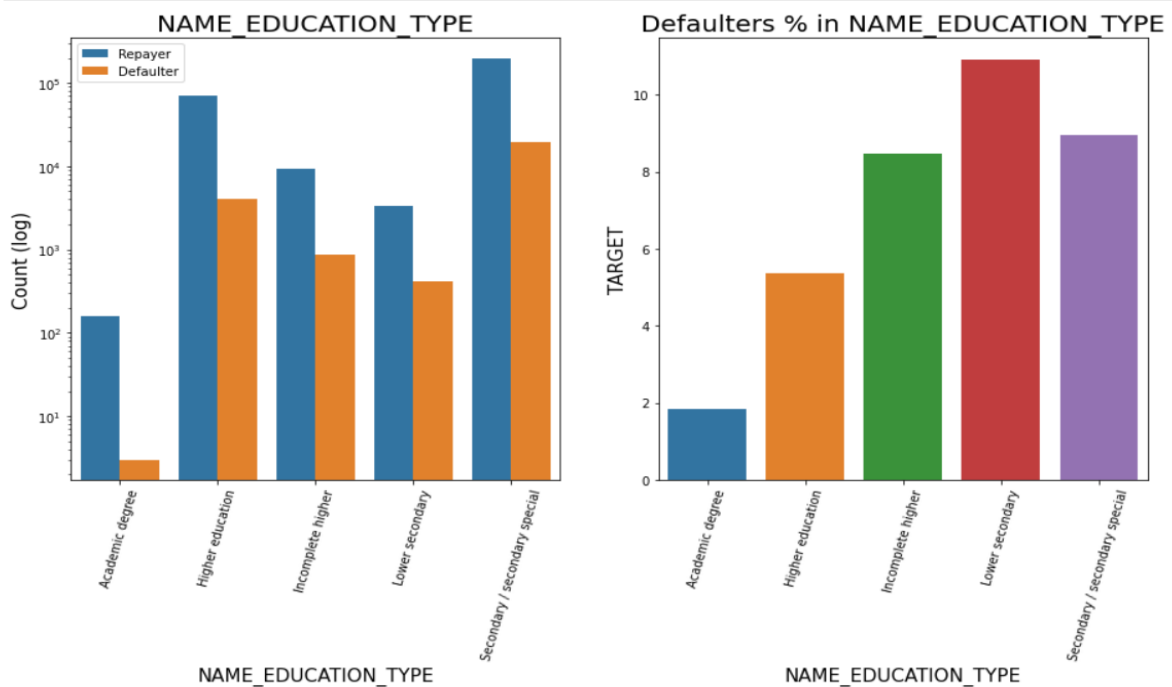


- Most of the people who have taken loan are married, followed by Single/not married and civil marriage.

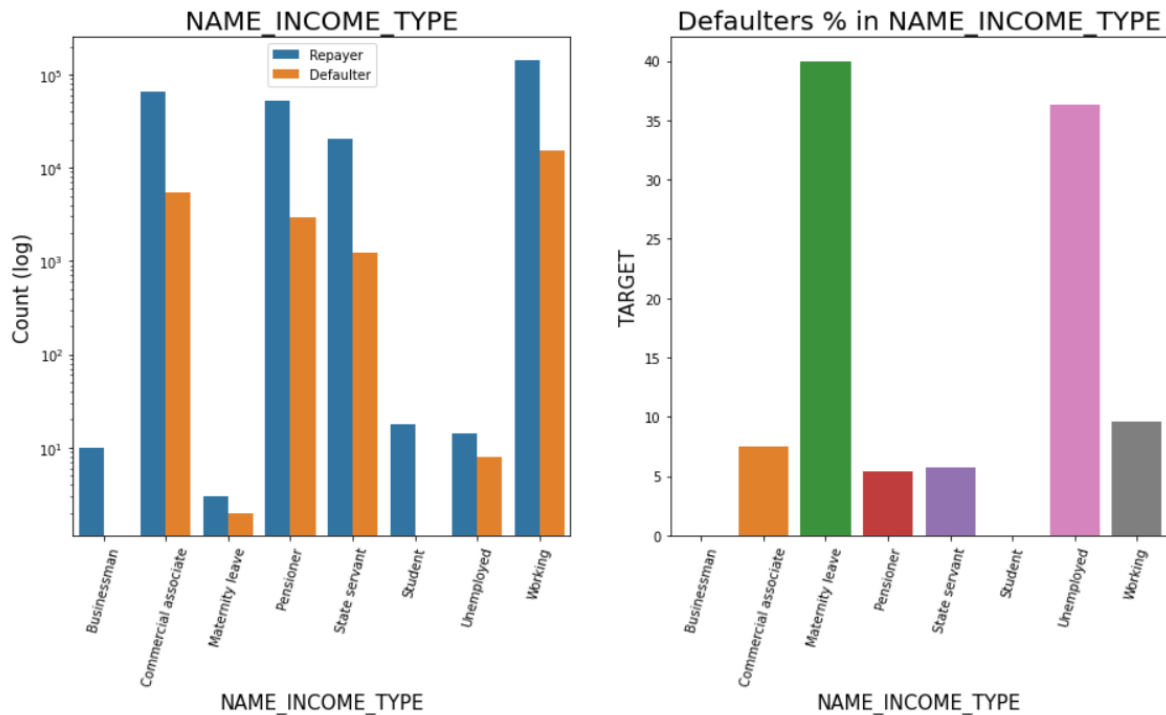
- In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest (exception being Unknown).



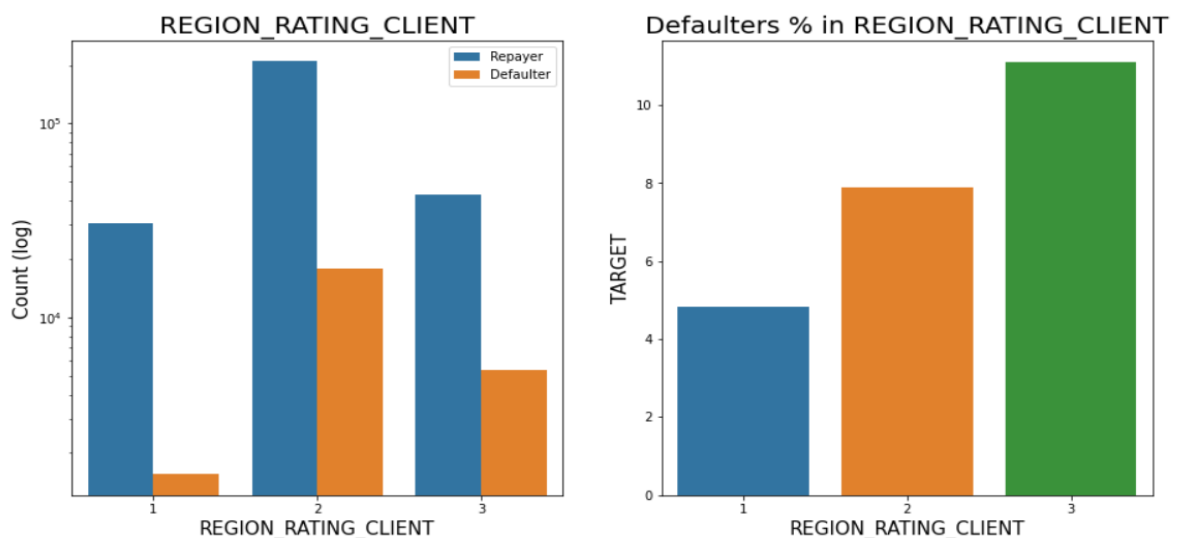
- Majority of the clients have Secondary / secondary special education, followed by clients with Higher education.
- Only a very small number having an academic degree.
- The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% defaulting rate.



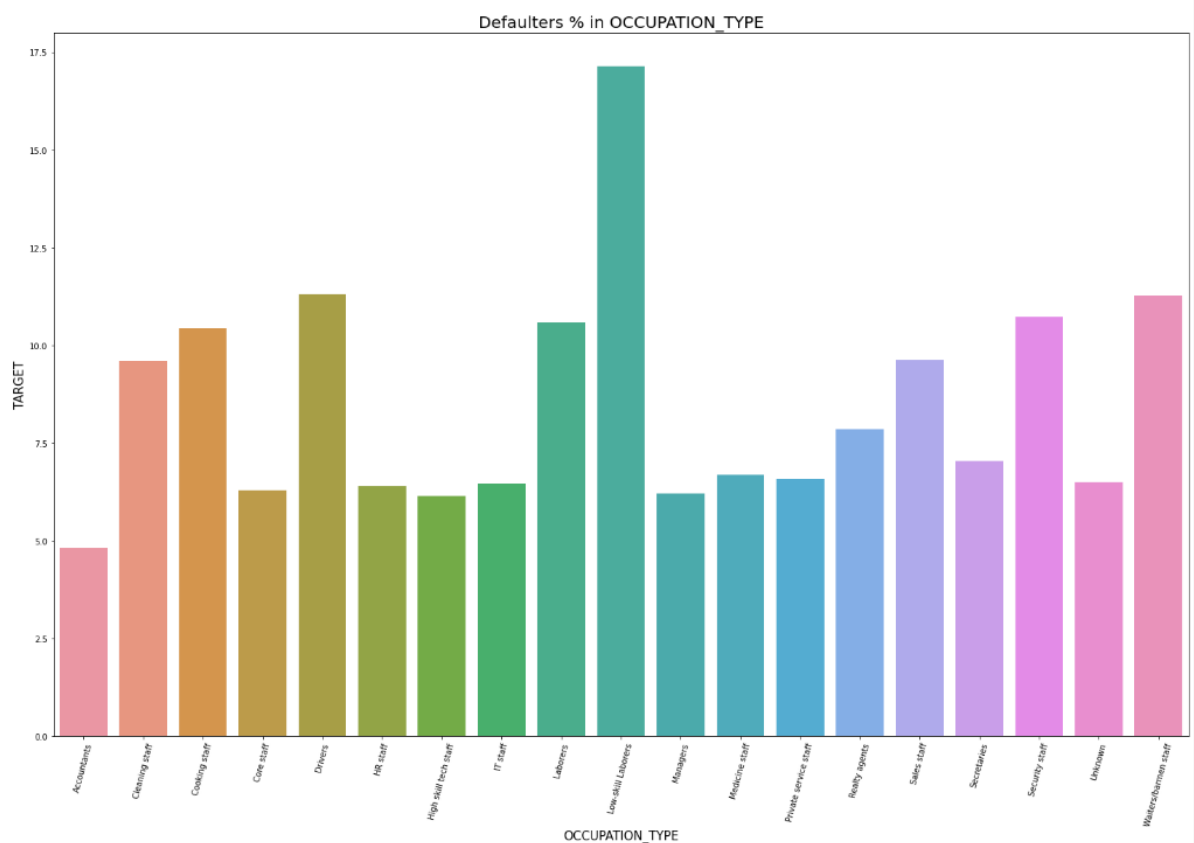
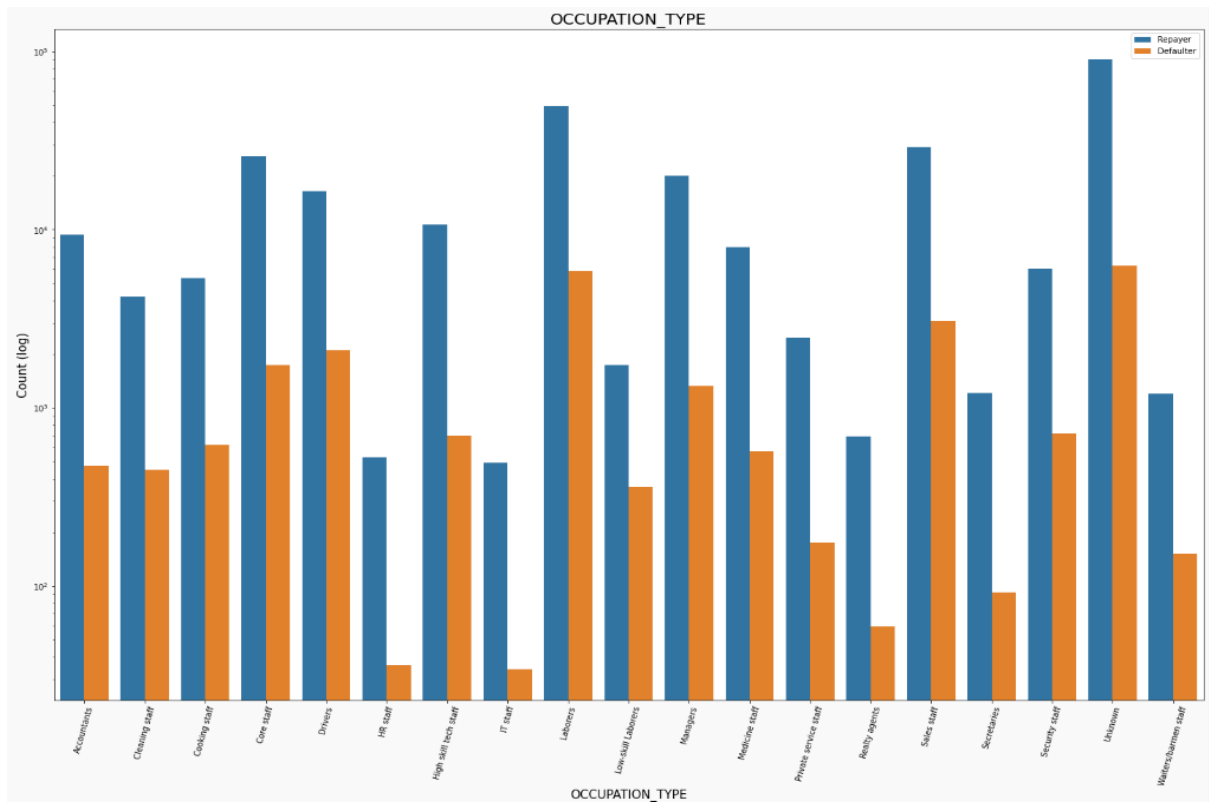
- Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.
- The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.
- Student and Businessmen, though less in numbers do not have any default record. Thus these two category are safest for providing loan.



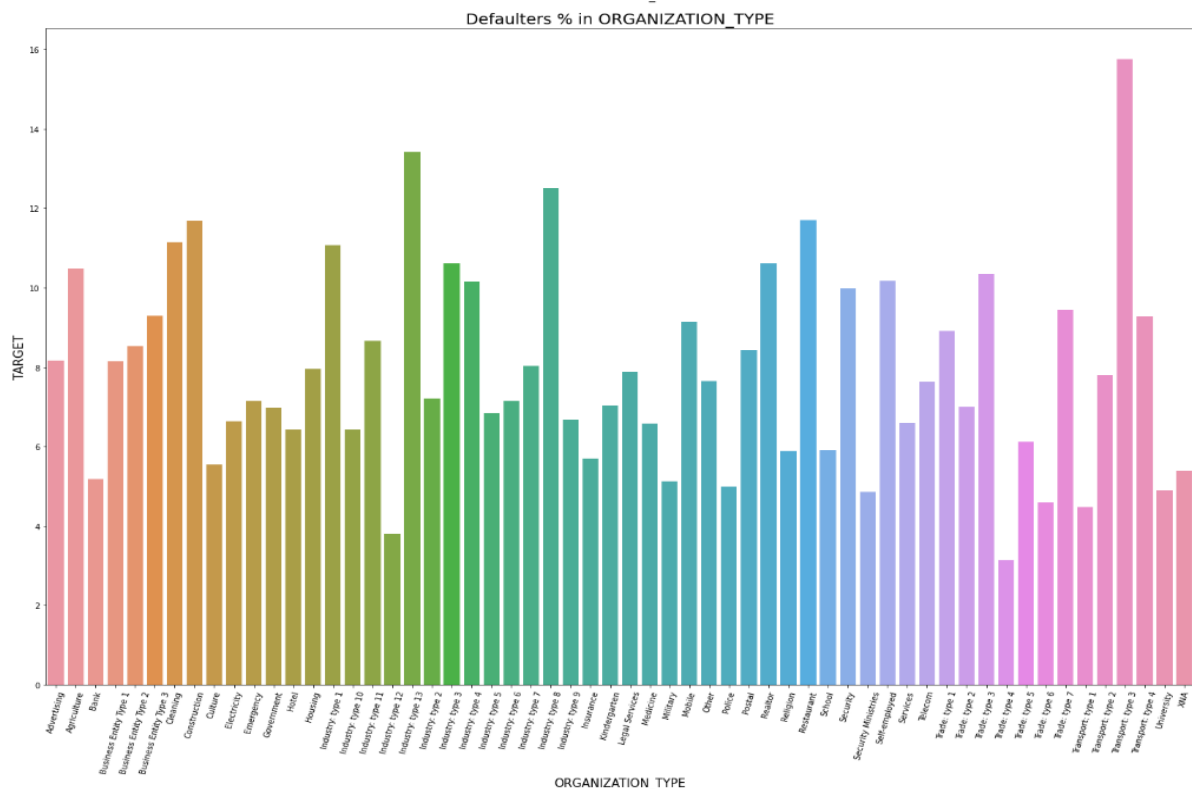
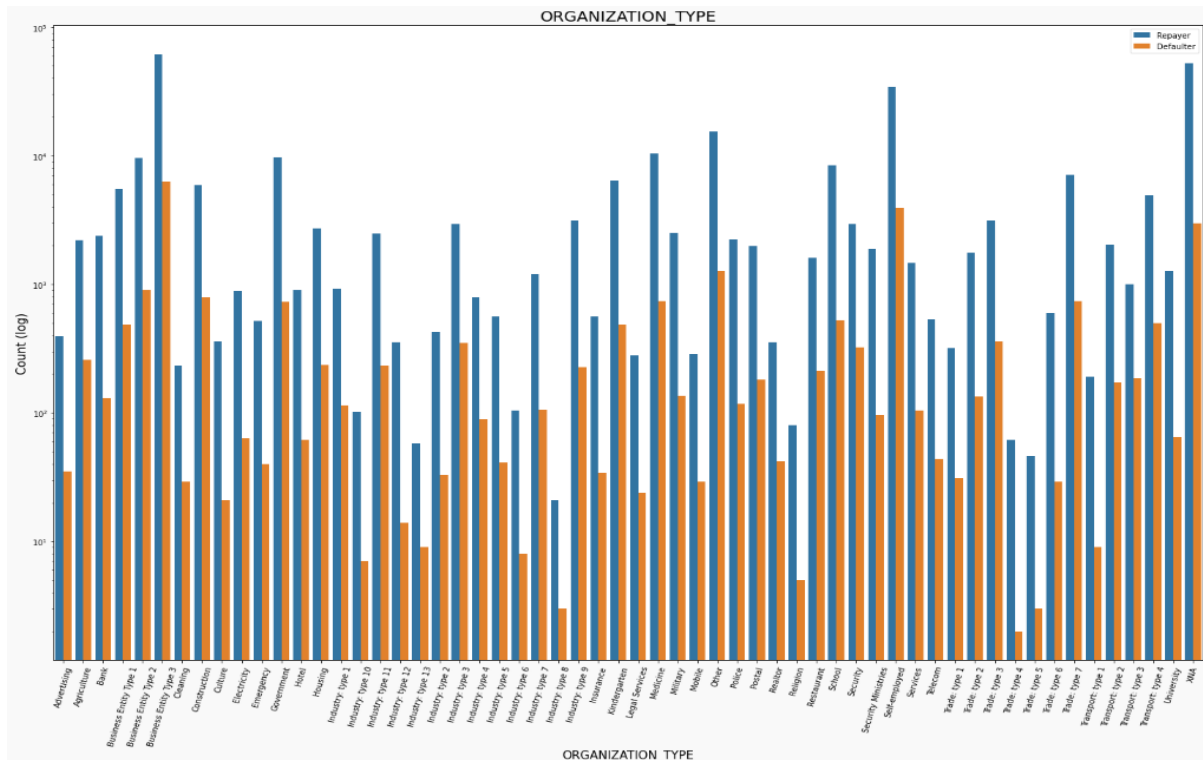
- Most of the applicants are living in Region_Rating 2 place.
- Region Rating 3 has the highest default rate (11%).
- Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans.



- Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans. The category with highest percent of not repaid loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.

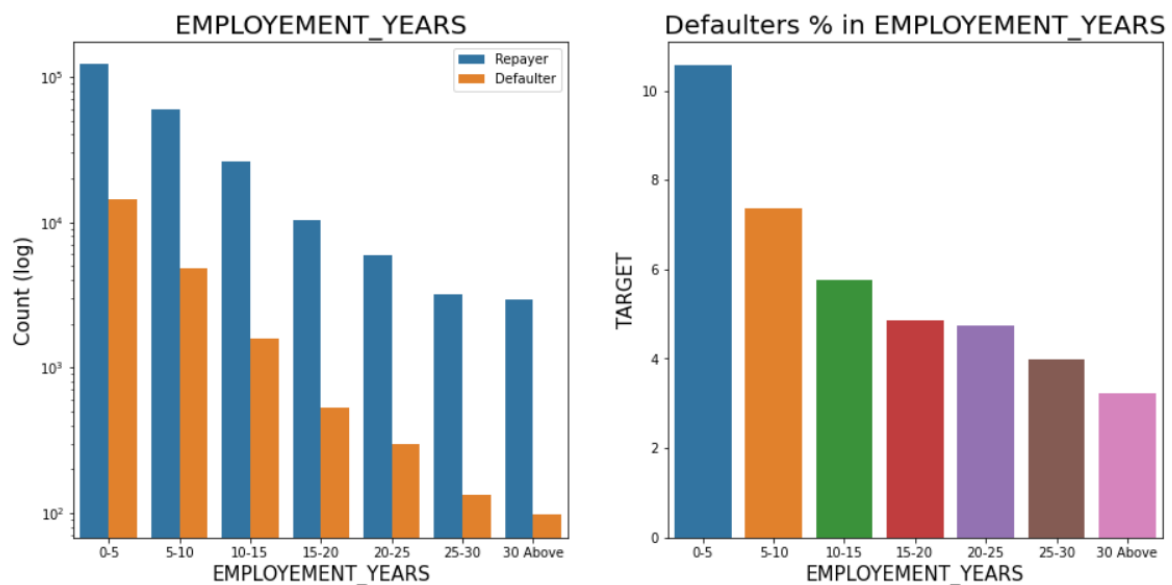


- Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting. Most of the people application for loan are from Business Entity Type 3. For a very high number of applications, Organization type information is unavailable(XNA).

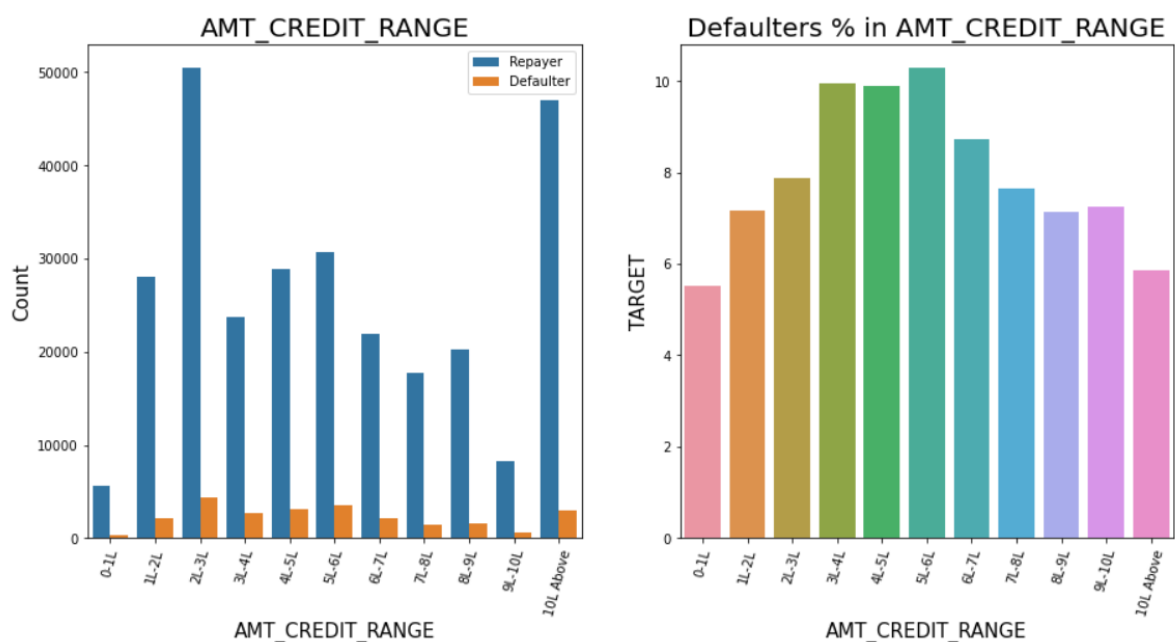


- Category of organization type has lesser defaulters thus safer for providing loans:
Trade Type 4 and 5 & Industry type 8

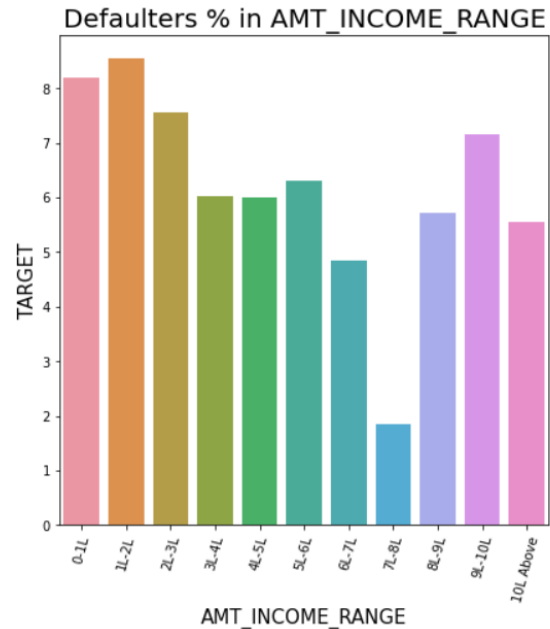
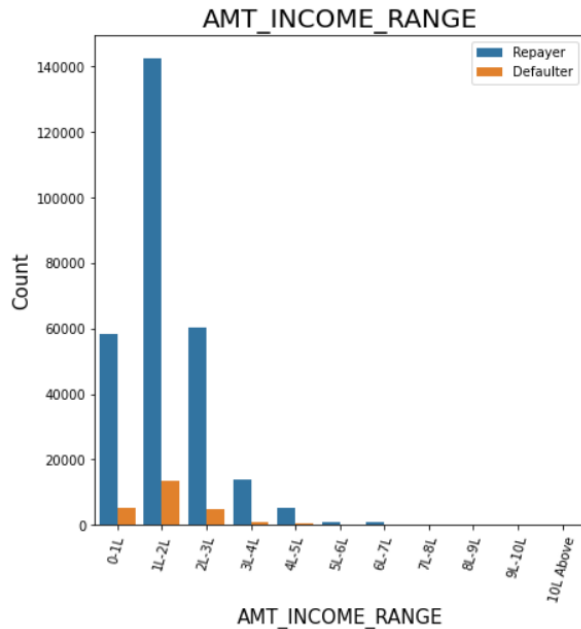
- Majority of the applicants have been employed in between 0-5 years. The defaulting rating of this group is also the highest which is 10%
- With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experiences having less than 1% default rate



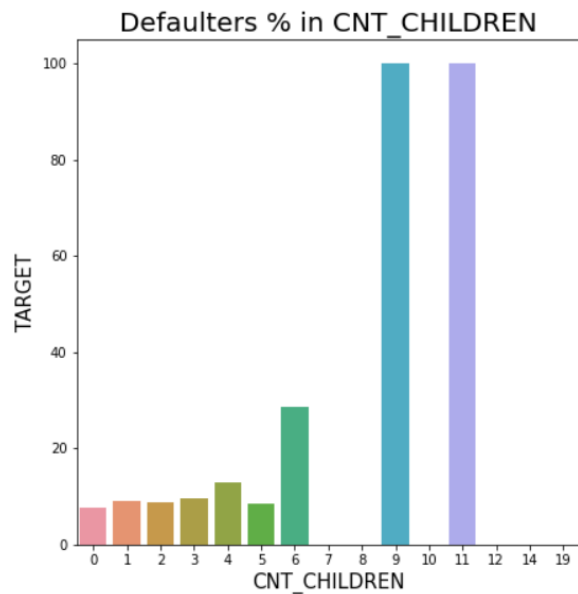
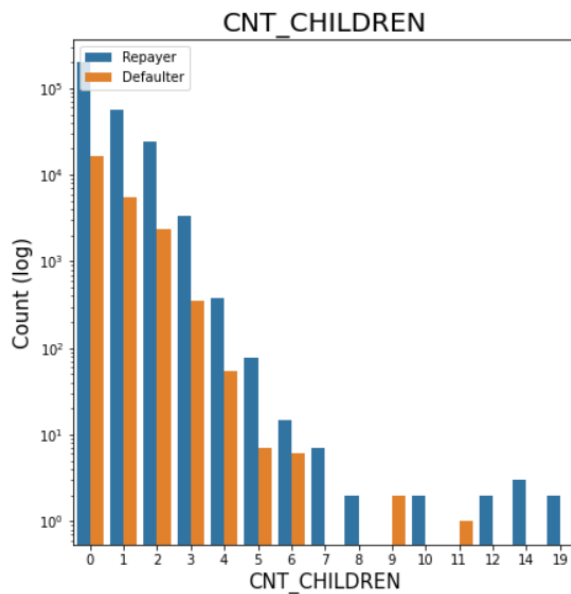
- There are high number of applicants have loan in range of 2-3 Lakhs followed by 10 Lakh above range
- More than 80% of the loan provided are for amount less than 900,000. People who get loan for 300-600k tend to default more than others.



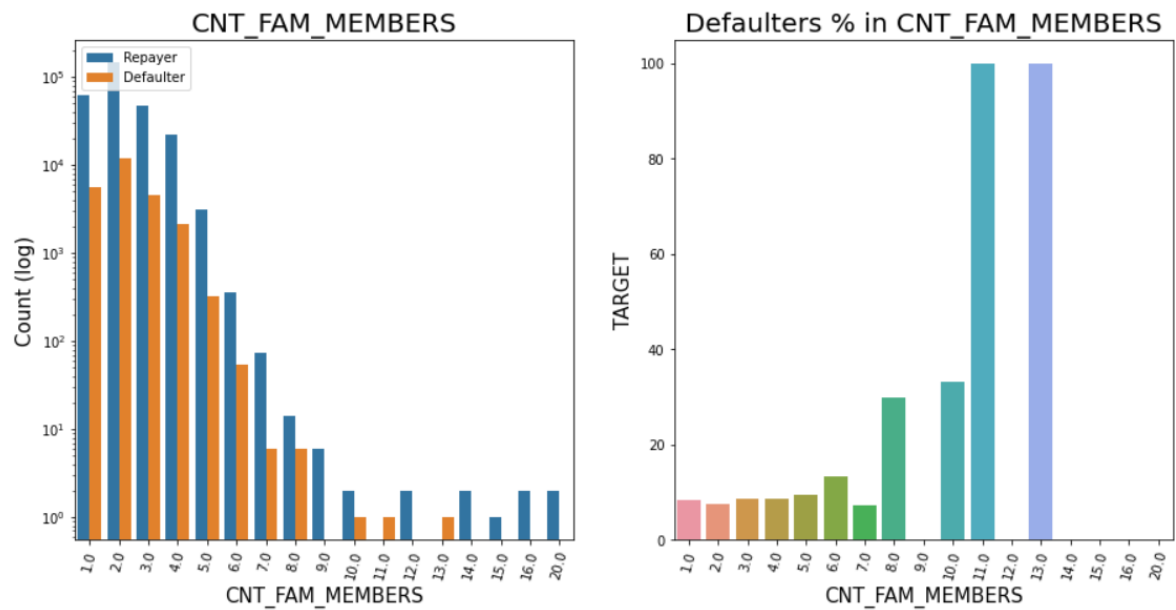
- 90% of the applications have Annual Income less than 300,000.
- Application with Income less than 300,000 has high probability of defaulting.
- Applicant with Income more than 700,000 are less likely to default.



- Most of the applicants do not have children.
- Very few clients have more than 3 children.
- Clients who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate.

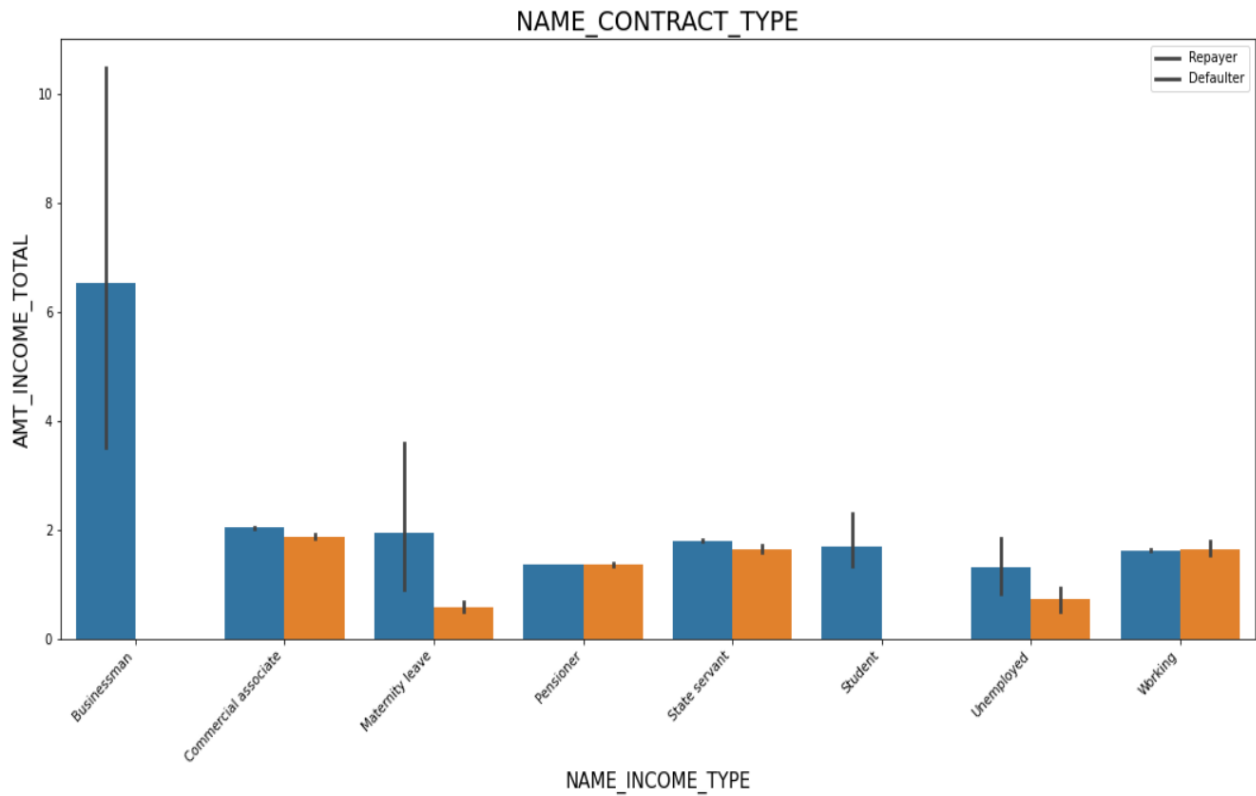


- Family members follow the same trend as Children, where, having more family members increases the risk of defaulting.
- Having more than 10 family members tends to have 100% default rate.



Categorical Bivariate or Multivariate Analysis :

- It can be seen that Business man's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a business man could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs.



Numeric Variables Analysis

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

The top 10 correlation for the Client with repayment:

	VAR1	VAR2	Correlation
64	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
65	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
43	AMT_ANNUITY	AMT_CREDIT	0.771309
131	DAYS_EMPLOYED	DAYS_BIRTH	0.626114
42	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
63	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
21	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
152	DAYS_REGISTRATION	DAYS_BIRTH	0.333151
174	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.276663
173	DAYS_ID_PUBLISH	DAYS_BIRTH	0.271314

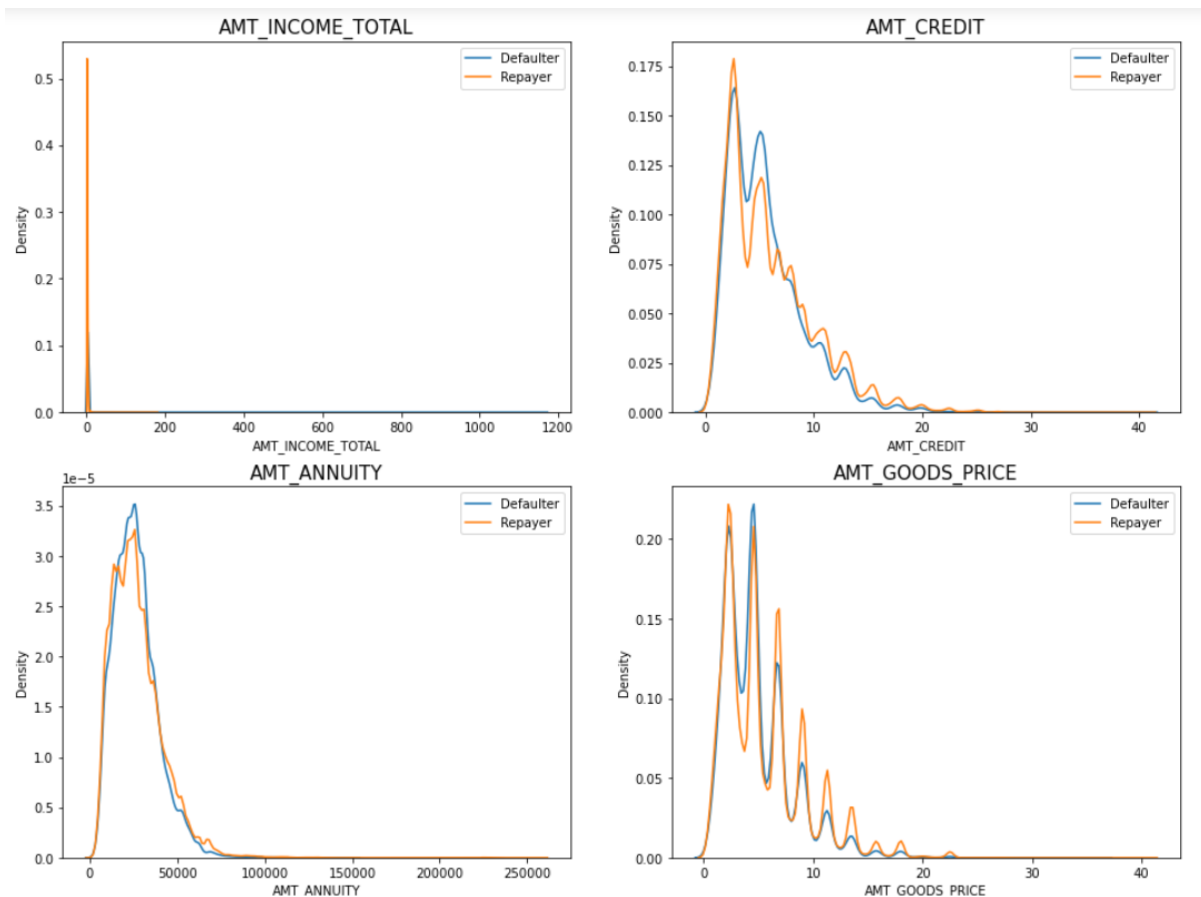
- Credit amount is highly correlated with amount of goods price, loan annuity, total income
- We can also see that repayment have high correlation in number of days employed.

The top 10 correlation for the Client with default:

	VAR1	VAR2	Correlation
64	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
65	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
43	AMT_ANNUITY	AMT_CREDIT	0.752195
131	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
152	DAYS_REGISTRATION	DAYS_BIRTH	0.289114
300	FLAG_DOCUMENT_3	DAYS_EMPLOYED	0.272169
263	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159
173	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863
351	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_HOUR	0.247511
174	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090

- Credit amount is highly correlated with amount of goods price which is same as repayments.
- But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayment(0.77).
- We can also see that repayment have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayment.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayment.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayment (0.254).

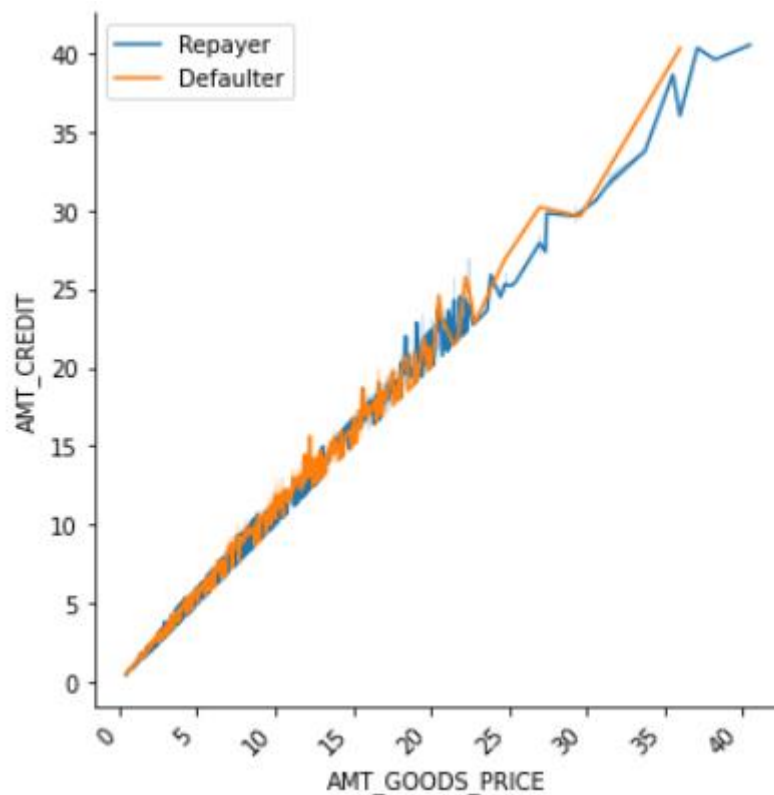
Numerical Univariate Analysis :



- Most no of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs
- The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

Numerical Bivariate Analysis :

Checking the relationship between Goods price and credit and comparing with loan repayment status.



- When the credit amount goes beyond 30 Lakhs, there is an increase in defaulters.

Insights:

Decisive Factors whether an applicant will Repay:

1. **NAME_EDUCATION_TYPE:** Academic degree has less defaults.
2. **NAME_INCOME_TYPE:** Student and Businessmen have no defaults.
3. **REGION_RATING_CLIENT:** RATING 1 is safer.
4. **ORGANIZATION_TYPE:** Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
5. **DAYS_BIRTH:** People above age of 50 have low probability of defaulting
6. **DAYS_EMPLOYED:** Clients with 40+ year experience having less than 1% default rate
7. **AMT_INCOME_TOTAL:** Applicant with Income more than 700,000 are less likely to default

8. NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, buying garage are being repaid mostly.

9. CNT_CHILDREN: People with zero to two children tend to repay the loans.

Decisive Factors whether an applicant will Default:

1. ODE_GENDER: Men are at relatively higher default rate

2. NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.

3. NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education

4. NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.

5. REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.

6. OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.

7. ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.

8. DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting

9. DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.

10. CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.

11. AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.

Results:

- In this case study, I applied the EDA in the real business case scenario.
- I learned basic of risk analytics in banking and financial services and understood how data is used to minimize the risk of losing money while lending to customers.
- This project was very challenging. I implemented the study of correlation between different variables to extract the necessary insights for the clients.
- I learned about data imbalance, outliers, driving factors for the datasets.
- It helped me in visualizing the huge dataset and summarizing the most important results helpful to the client.

Python programming Code (Data Analysis) Drive Link:

https://drive.google.com/drive/folders/1RHeOV8CVX5NY6fv6VEfFOu4KAPx7_hU?usp=sharing