Bank Loan Case Study

Project Description:

This project aims to perform exploratory data analysis (EDA) on a loan application dataset in the context of a consumer finance company. The goal is to identify patterns and variables that influence the likelihood of loan default. By analysing the data, we can gain insights into the driving factors behind loan default, which can help the company make informed decisions regarding loan approvals, risk assessment, and portfolio management.

Approach:

The project will involve the following steps:

1.Data acquisition: Download the dataset containing client information and loan attributes.

2.Data preprocessing: Identify missing data and handle it appropriately by either removing columns or replacing missing values. Analyze outliers and determine if any action needs to be taken. Check for data imbalance and calculate the ratio of imbalance.

3.Perform EDA: Conduct univariate, segmented univariate, and bivariate analysis to explore the data. Interpret the results in business terms to understand the significance of variables in differentiating clients with payment difficulties from others.

4.Calculate correlations: Segment the data based on the target variable (payment difficulties) and calculate the top 10 correlations for each segment. Examine the correlations to identify any insights.

5.Visualization and summarization: Use visualizations such as histograms, bar plots, and heatmaps to present the most important results. Summarize the key findings and their implications for the business.

Tech-Stack Used:

MS Excel – To explore the dataset.

VS code – To use 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: For the data analysis, python is the best programming language.

Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly

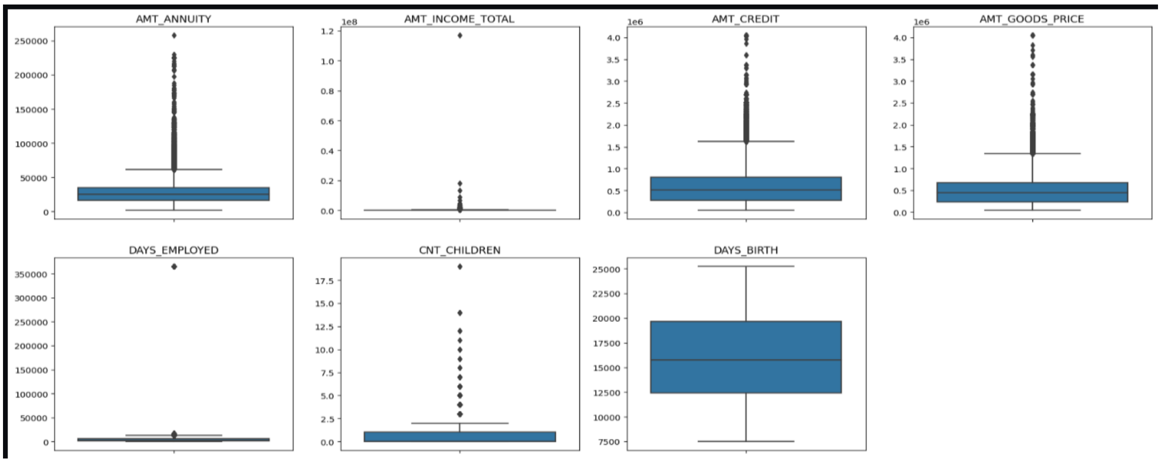
1. **Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)**
    1. There are total of 49 columns in Application_Data and 11 columns in Previous_Application which have missing values greater than 40%.
    2. On further analysis, I found that "EXT_SOURCE_2"," EXT_SOURCE_3" has no correlation with the "TARGET" column.
    3. On checking the relation of 'FLAG_DOCUMENT_X' with loan repayment status, we found that the clients applying for loans only submitted the 'FLAG_DOCUMENT_3'.
    4. 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.
    5. '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. 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 'NAME_TYPE_SUITE' using mode, 'OCCUPATION_TYPE' by adding an 'Unknown' category, numerical variables
    9. '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.
    10. 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 are not started.

2. **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.

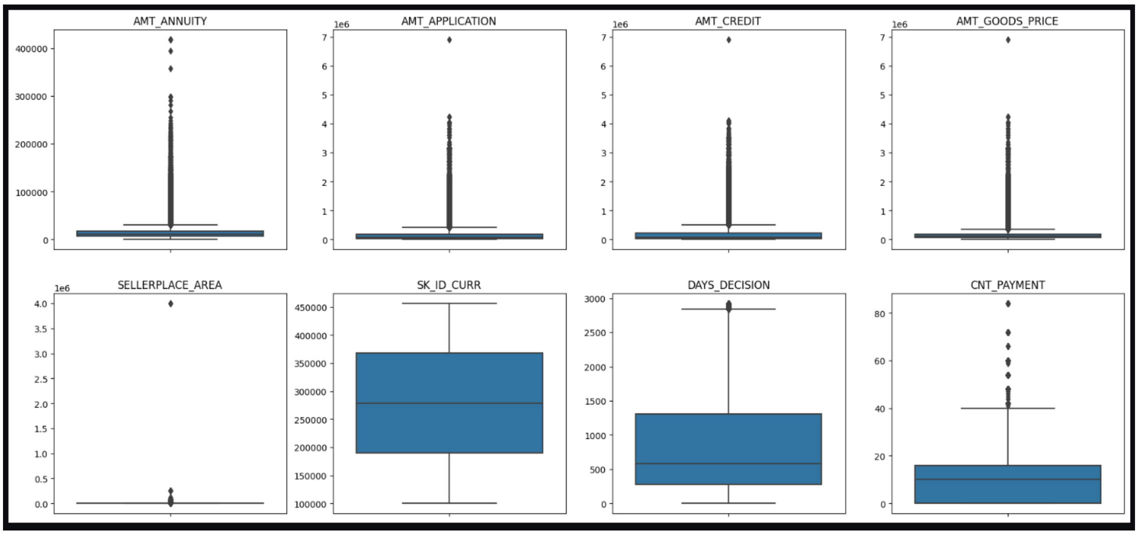
Application_Data:

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.



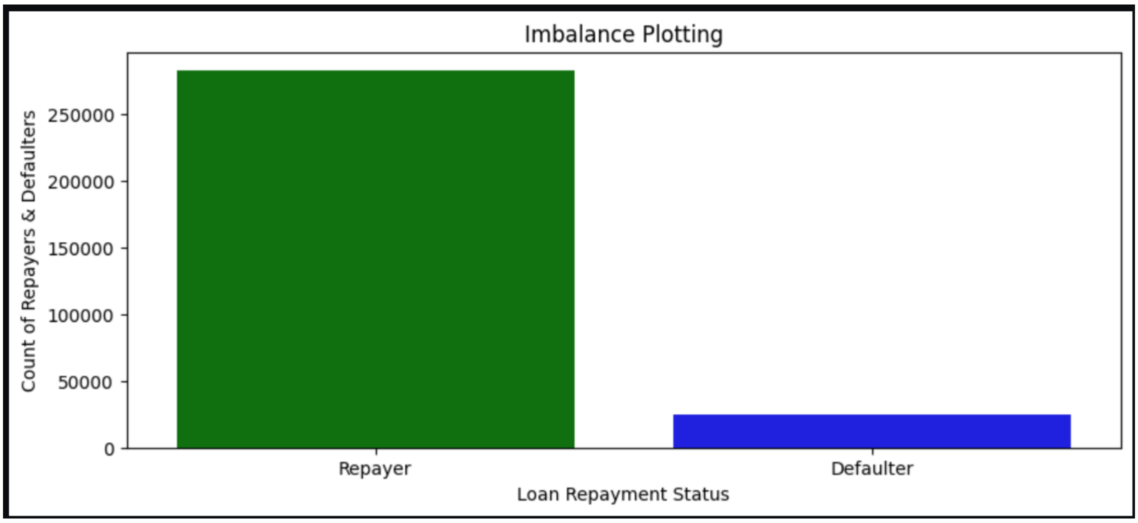
Previous_Application:

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



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

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



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

1. 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%).
2. Clients who own a car are half in number of the clients who don't own a car. But based

on the percentage of default, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost same.

3. 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.

4. Majority of people live in House/apartment

5. People living in office apartments have lowest default rate

6. People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting

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

8. 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).

9. Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree

10. 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.

11. Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.

12. 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.

13. Student and Businessmen, though less in numbers do not have any default record. Thus these two category are safest for providing loan.

14. Most of the applicants are living in Region_Rating 2 place.

15. Region Rating 3 has the highest default rate (11%).

16. Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans.

17. Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans.

18. 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.

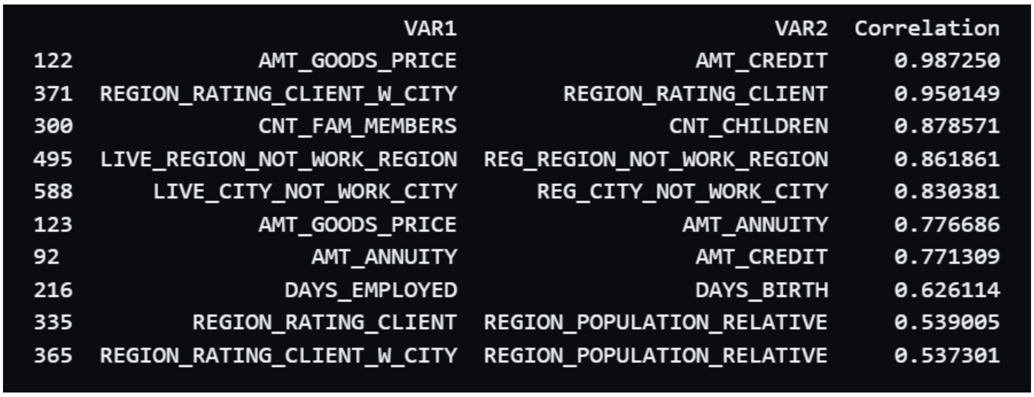
19. There is no significant correlation between non-defaulters and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (~9%) than who have not submitted the document (6%).

20. 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.

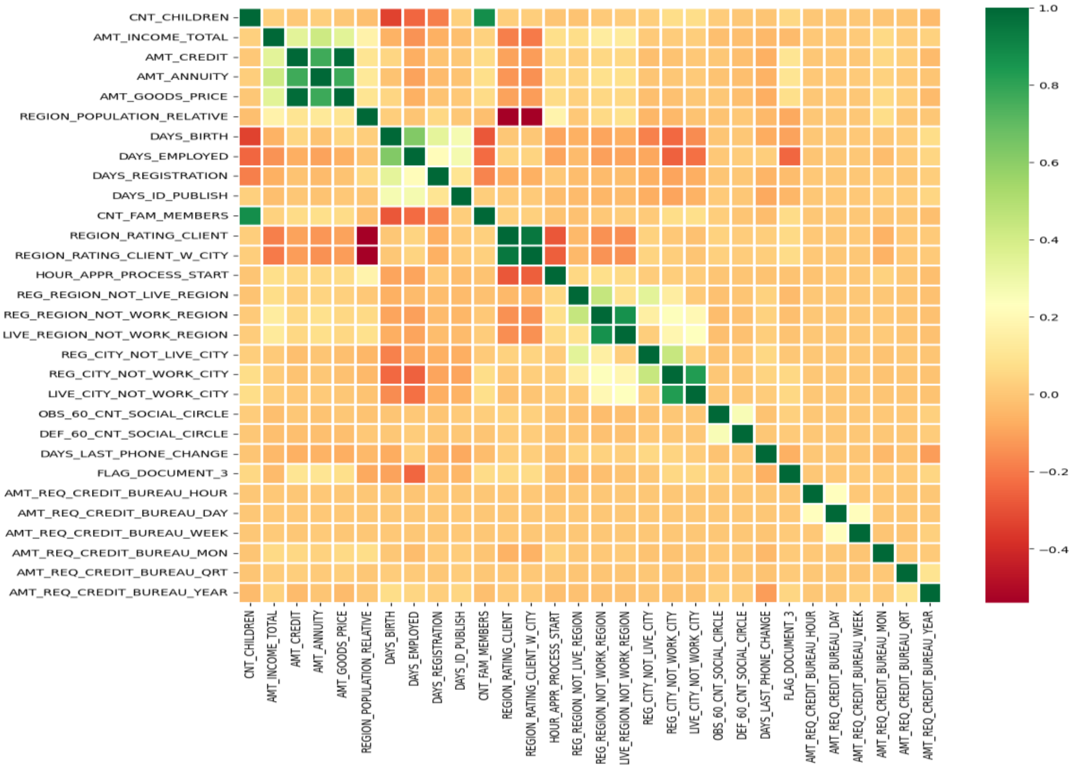
21. Family members follow the same trend as Children, where, having more family members increases the risk of defaulting.

5. **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:

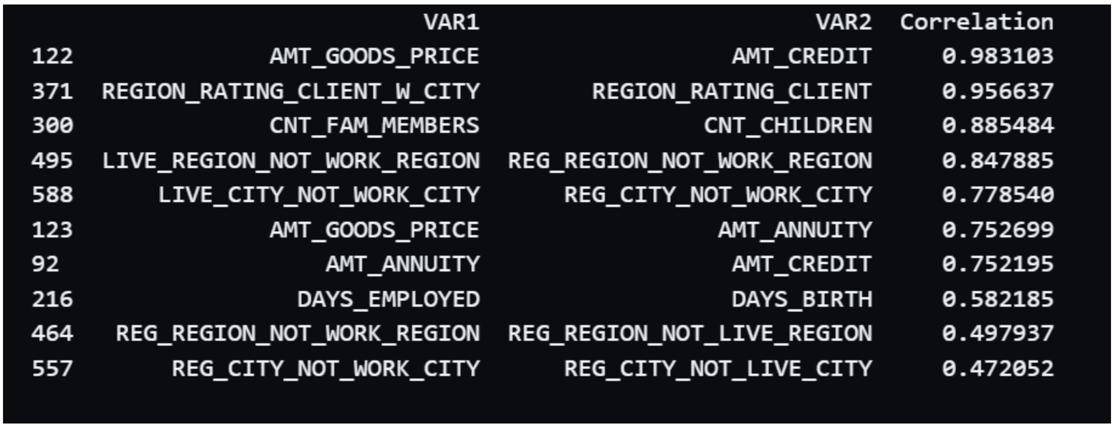


```
                              VAR1                         VAR2  Correlation
122                AMT_GOODS_PRICE                   AMT_CREDIT     0.987250
371    REGION_RATING_CLIENT_W_CITY         REGION_RATING_CLIENT     0.950149
300                CNT_FAM_MEMBERS                 CNT_CHILDREN     0.878571
495    LIVE_REGION_NOT_WORK_REGION   REG_REGION_NOT_WORK_REGION     0.861861
588        LIVE_CITY_NOT_WORK_CITY       REG_CITY_NOT_WORK_CITY     0.830381
123                AMT_GOODS_PRICE                  AMT_ANNUITY     0.776686
92                     AMT_ANNUITY                   AMT_CREDIT     0.771309
216                  DAYS_EMPLOYED                   DAYS_BIRTH     0.626114
335           REGION_RATING_CLIENT   REGION_POPULATION_RELATIVE     0.539005
365    REGION_RATING_CLIENT_W_CITY   REGION_POPULATION_RELATIVE     0.537301
```

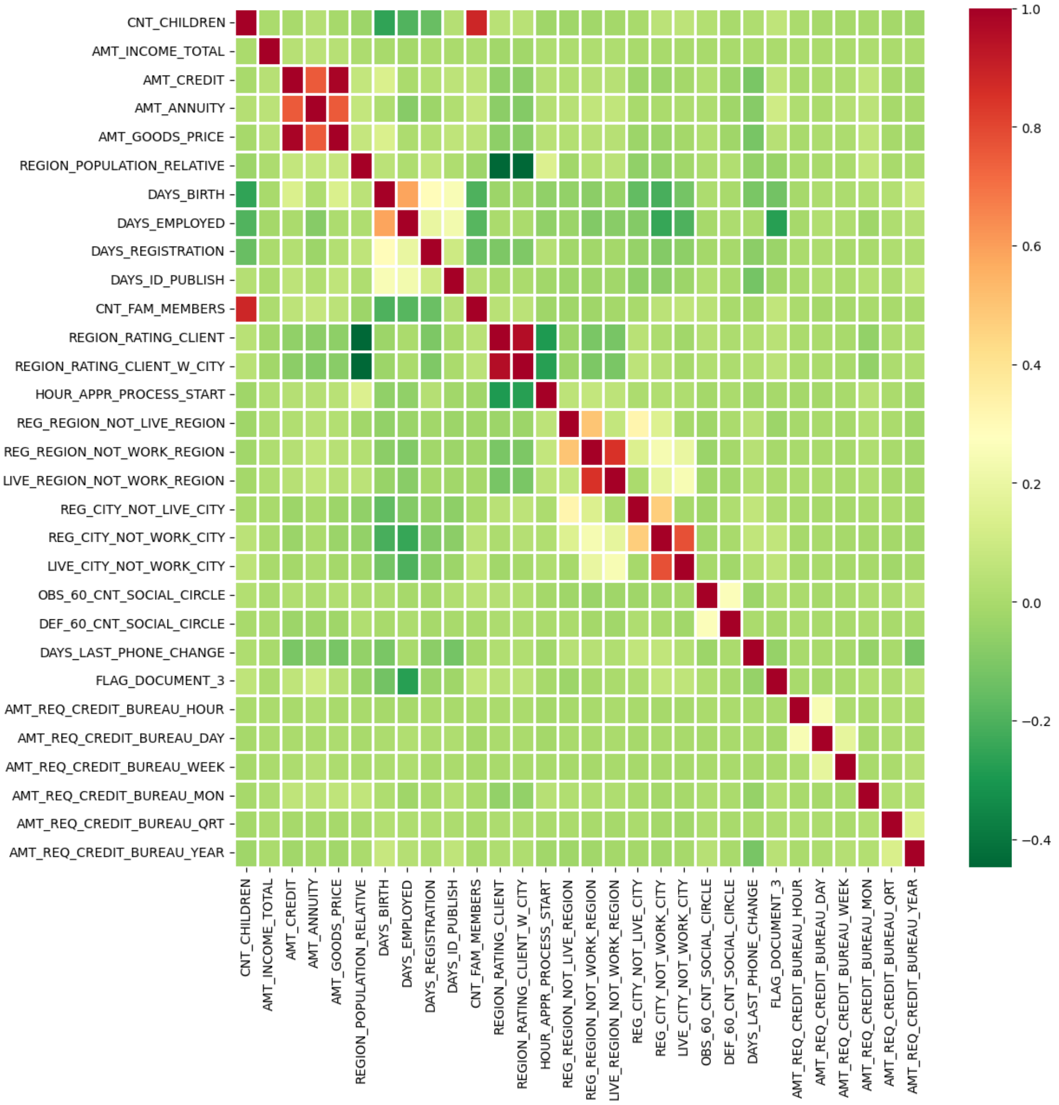


Credit amount is highly correlated with amount of goods price, loan annuity, total income

The top 10 correlation for the Client with default:



```
                             VAR1                        VAR2  Correlation
122                AMT_GOODS_PRICE                  AMT_CREDIT     0.983103
371   REGION_RATING_CLIENT_W_CITY        REGION_RATING_CLIENT     0.956637
300                CNT_FAM_MEMBERS                CNT_CHILDREN     0.885484
495   LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION     0.847885
588        LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY     0.778540
123                AMT_GOODS_PRICE                 AMT_ANNUITY     0.752699
92                     AMT_ANNUITY                  AMT_CREDIT     0.752195
216                  DAYS_EMPLOYED                  DAYS_BIRTH     0.582185
464     REG_REGION_NOT_WORK_REGION  REG_REGION_NOT_LIVE_REGION     0.497937
557         REG_CITY_NOT_WORK_CITY      REG_CITY_NOT_LIVE_CITY     0.472052
```



1.Credit amount is highly correlated with amount of goods price which is same as repayments.

2. But the loan annuity correlation with credit amount has slightly reduced in Defaulters (0.75) when compared to repayment (0.77).

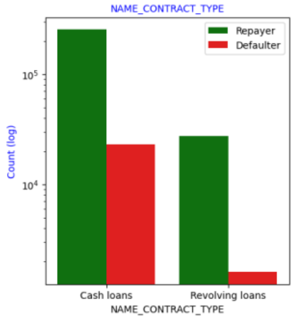
3. We can also see that repayment have high correlation in number of days Employed (0.62) when compared to defaulters (0.58).

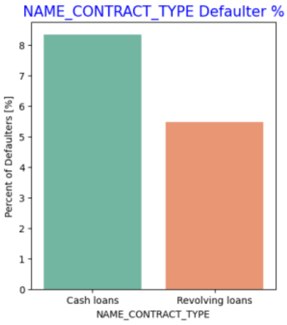
4. There is a severe drop in the correlation between total income of the client and the credit amount (0.038) amongst defaulters whereas it is 0.342 among repayment.

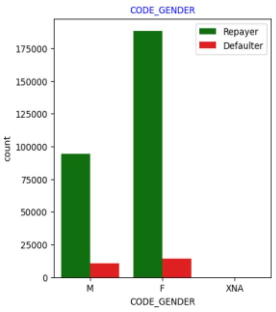
5. Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayment.

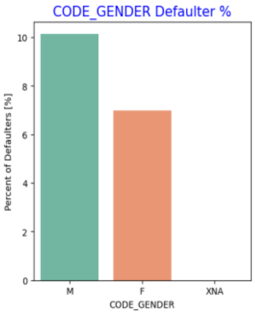
6. There is a slight increase in defaulted to observed count in social circle among Defaulters (0.264) when compared to repayment (0.254).

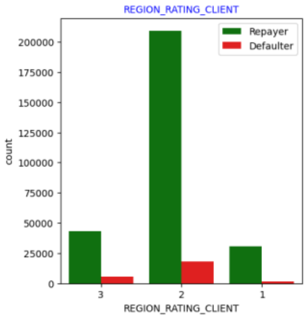
6. Include visualizations and summarize the most important results in the presentation.

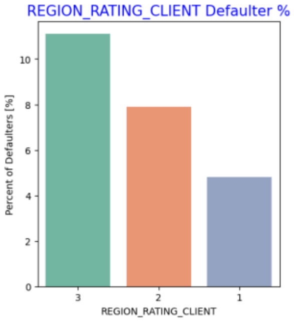


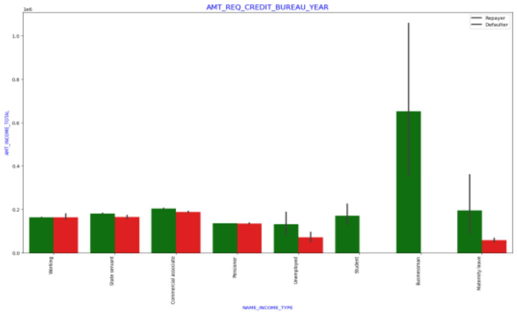


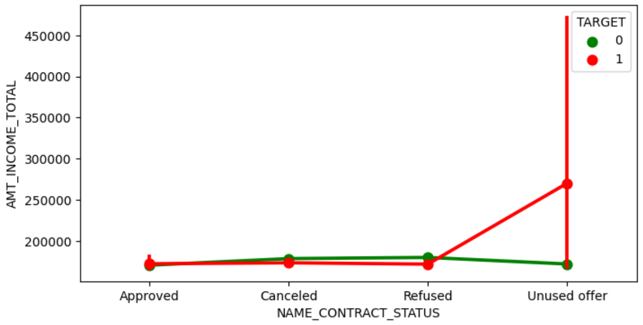


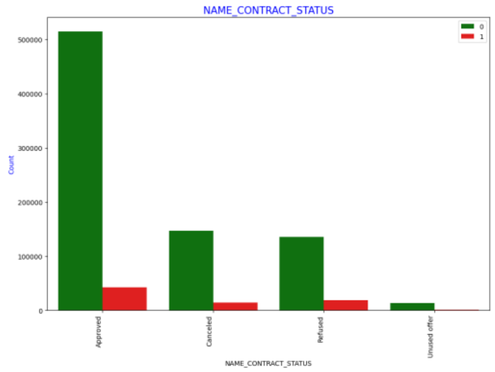












Insights:

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. CNT_CHILDREN: people who have zero or two children tend to repay the loans.
5. NAME_FAMILY_STATUS: Widows are more likely to repay the loans

Decisive Factors whether an applicant will Default:

1. CODE_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. 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.

Result:

- 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 case study helped me in learning how to summarize a huge dataset to gain the valuable insights.
- 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 result helpful to the client