EDA Assignment

DS C54 Batch

Problem Statement

- 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 to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to 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.

Problem Statement

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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 applicants using EDA is the aim of this case study.

Handling Outliers and Missing Values

Missing Value Treatment

- The first approach was to drop columns which have more than 40% missing values. This brought down the ‘application_data’ from 122 columns to 73 columns.
- Next, I explored the different columns and identified missing values. Note that missing values can come in various forms. Like for example: The gender column had a value as ‘XNA’. These were very less in number and hence I proceeded by dropping those values from the dataset.

- There could be different approaches while dealing with missing values. Instead of dropping the values, you can also choose to impute them with values. For example: You can replace missing values by the median value if it is a continuous variable or with mode if it is a categorical variable.
- The CNT_FAM_MEMBERS column had missing values which I replaced with the mode of that particular column.

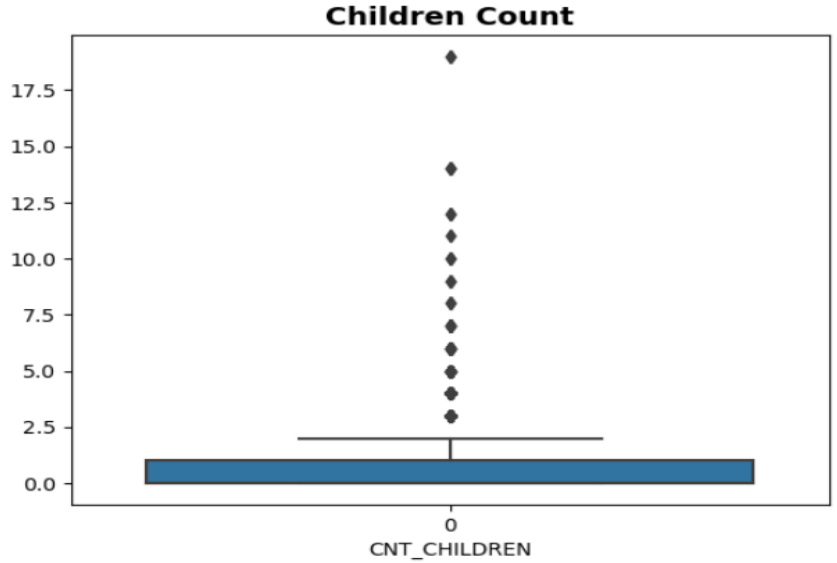
Outlier Treatment

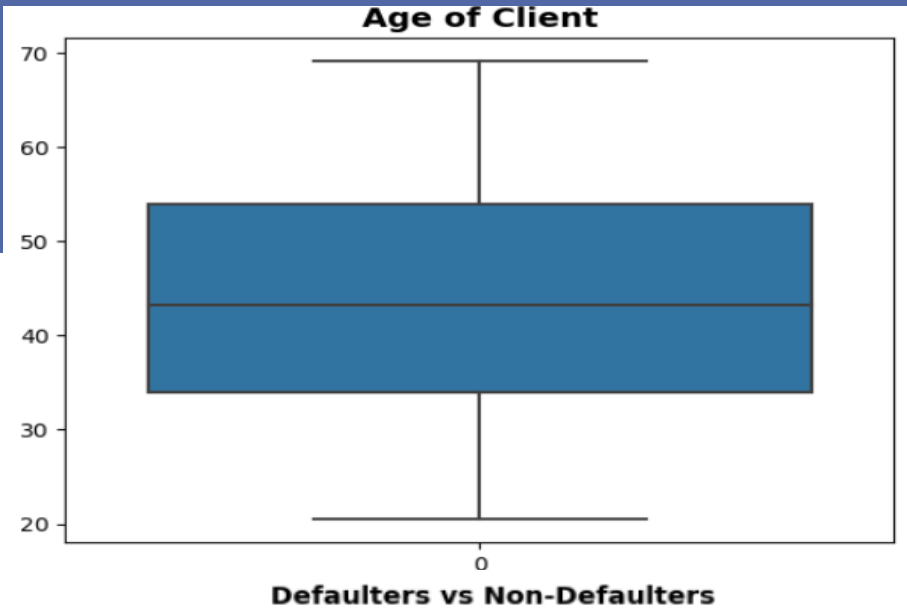
- Outliers are data points that significantly deviate from the majority of the data in a dataset. These data points are unusually distant or extreme compared to other values and can have a substantial impact on statistical analysis and modeling.
- You can identify outliers by plotting the data using scatter plots, box plots, or histograms which can reveal potential outliers visually.

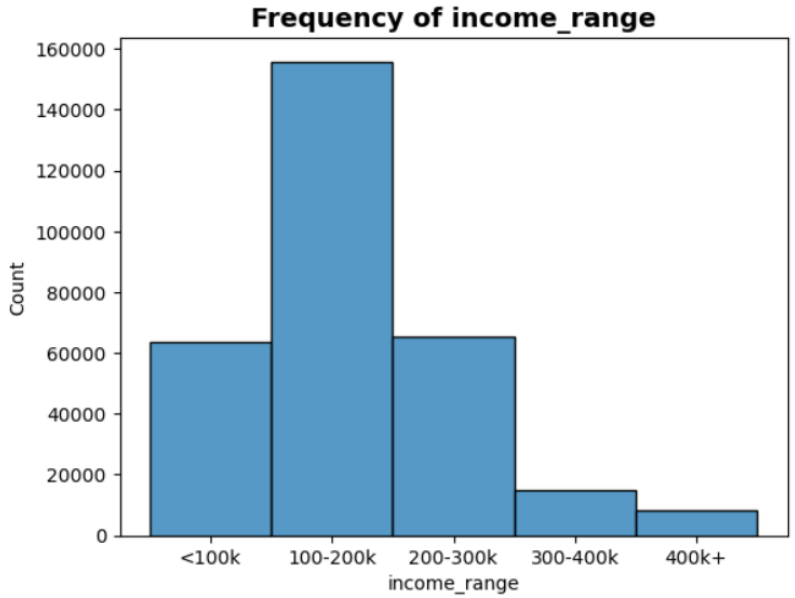
- I took a similar technique while working with the dataset. I plot a boxplot for the ‘CNT_CHILDREN’ and ‘CNT_FAM_MEMBERS’ column to identify outliers.
- Once outliers are identified you can use various techniques to handle them like using median or mode depending on if it is a continuous or a categorical variable. You can also choose to drop them but only if you are confident that it is not going to affect your analysis.

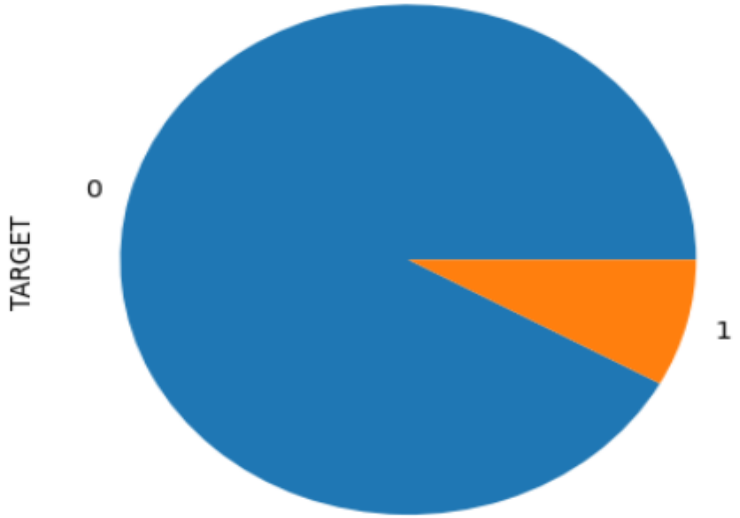
Univariate and Bivariate Analysis

Univariate Analysis

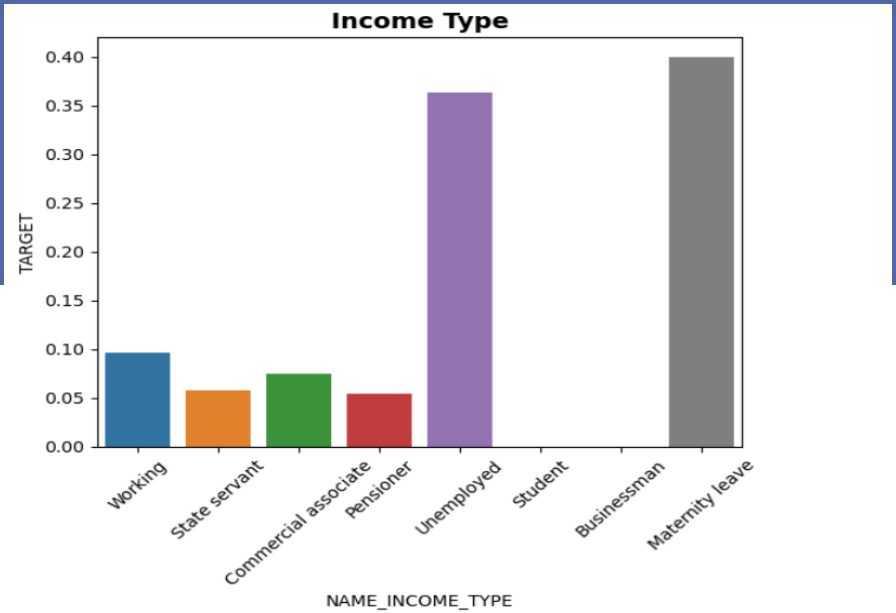


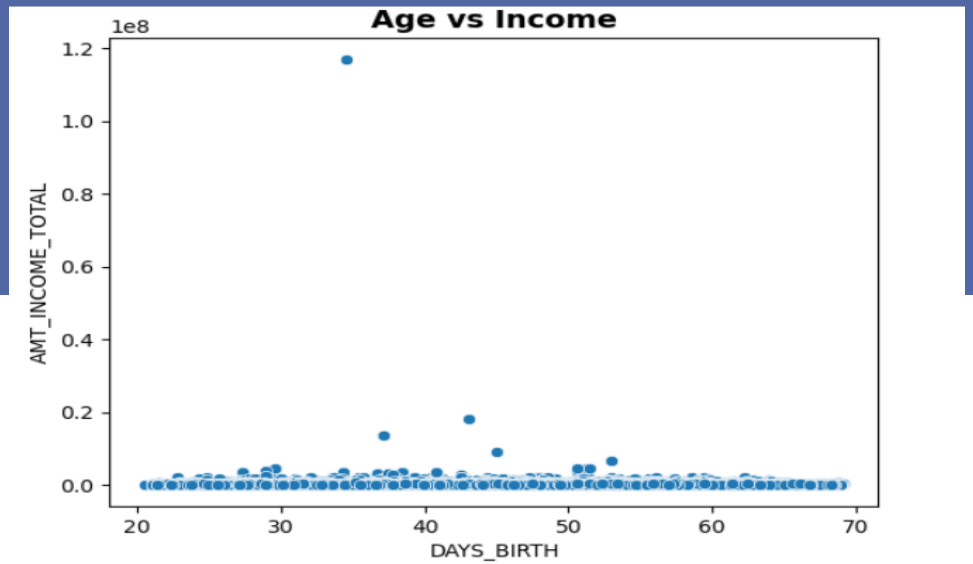


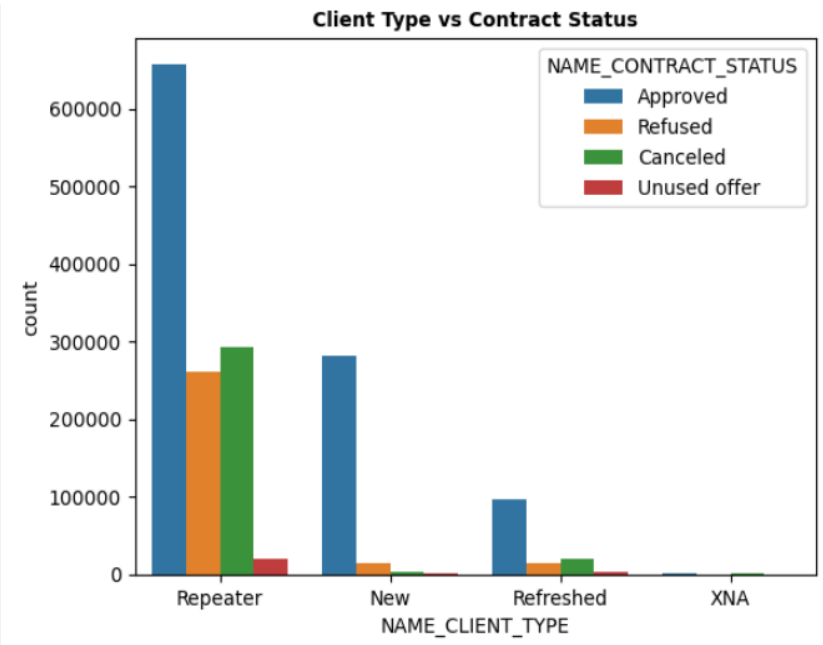




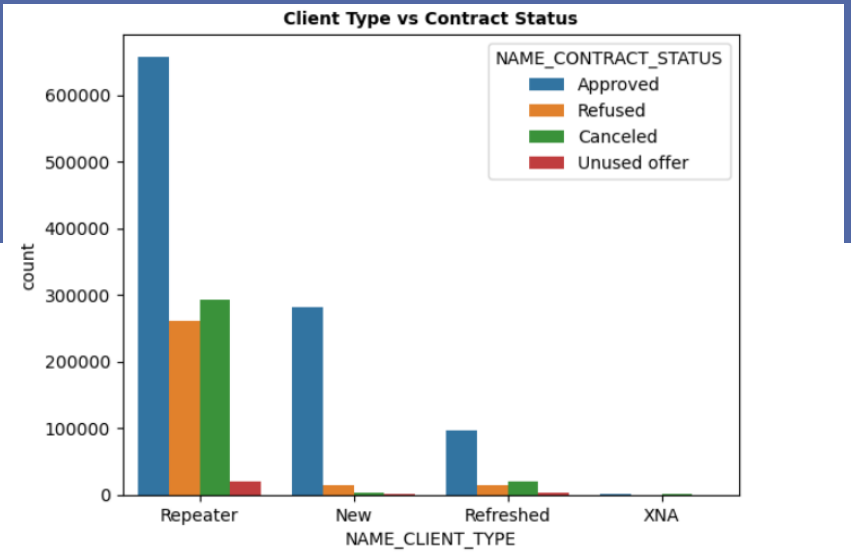
Bivariate Analysis

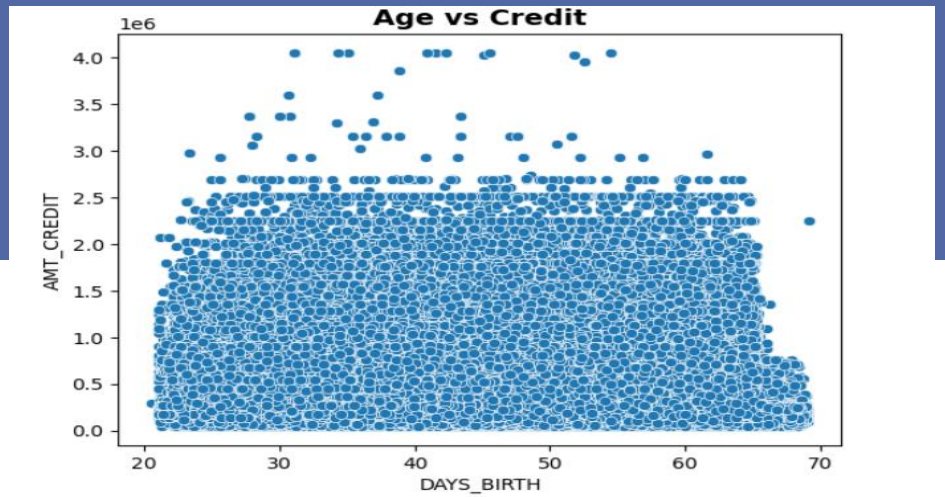


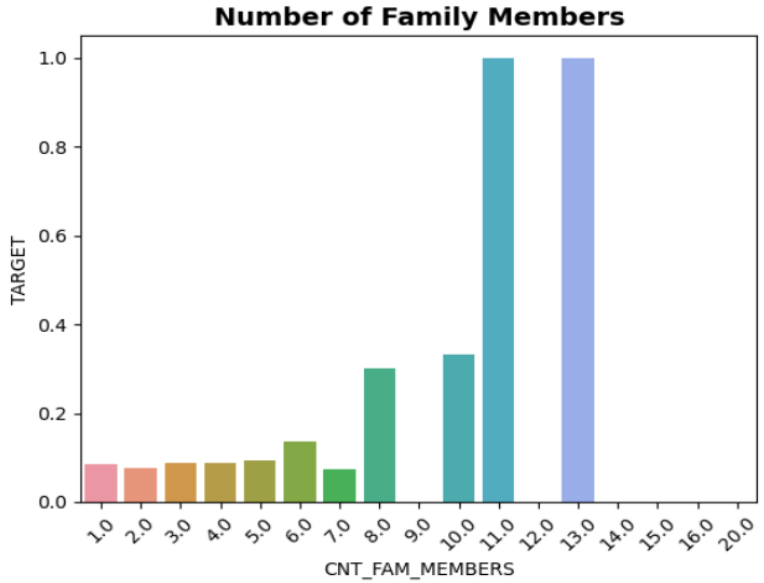


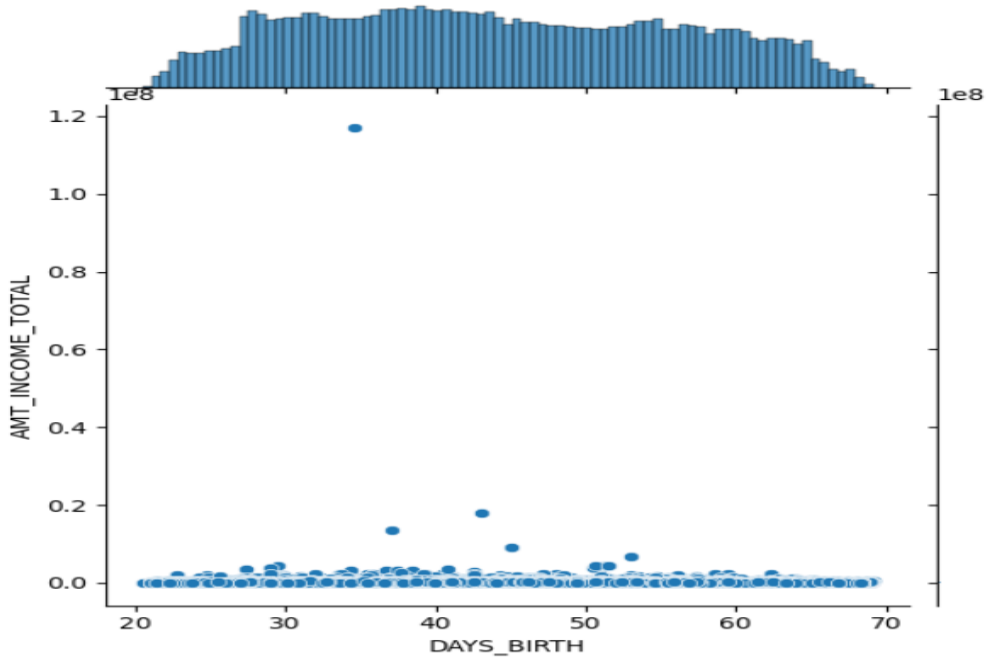


Bivariate Analysis









Top 10 Correlations

Target = 1

FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999705
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
AMT_GOODS_PRICE	AMT_CREDIT	0.983103
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
AMT_ANNUITY	AMT_CREDIT	0.752195

dtype: float64

Target = 0

FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999756
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
AMT_GOODS_PRICE	AMT_CREDIT	0.987253
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950148
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878569
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859289
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
AMT_ANNUITY	AMT_CREDIT	0.771308

dtype: float64

Insights

- While performing the univariate and bivariate analysis there were certain patterns which I have noticed. Clients with more salaries tend to default less than compared to clients with less salaries.
- The pattern we saw was that people with less than 200K salary tend to default a bit more and the trend is decreasing as the salary bucket increases.

- The next pattern that I noticed was in the education variable. People with academic degrees or higher tend to default less compared to people with lower secondary education or secondary education.
- The next pattern that I noticed was in the Income type column where there was a distinct difference between unemployed and people on maternity leave tend to default way more than compared to people from other categories.

Insights Contd

- While analysing the CNT_FAM_MEMBERS column, I noticed a pattern that people with a bigger family have a chance to default more than compared with people with a small family. The same trend with people having more children who tend to default more compared to people with less children. This pattern is also aligned with the high correlations between the two columns which means people with more children have more family members.
- I noticed a pattern in the Region Rating column where people living in Region 1 tend to default less compared to people living in other regions. This is also the same trend with Region Rating Client with city and the same was proved by the correlation coefficient which shows .956 between the two columns

- There was also a pattern in the 'Occupation_Type' column where Low skill laborers category tend to default way more compared to the other categories.
- After segmenting the data, I noticed people towards the late 20s tend to default the most compared to others.
- After merging the two datasets I found out clients who were 'Refused' earlier tend to default more compared to other clients.

Conclusion

- After analysing the dataset, there were some clear patterns in the Income Type Column, income amount column, education type column where people of a certain categories tend to default more than compared to the other categories of people.
- People who have more children also tend to default more which is aligned with people who have more family members.
- There is clear pattern in the Occupation Type column where low skill laborers have a chance to default more compared to the others.
- Keeping these points in mind, the bank can make decisions to give out loans or reject them. They can maybe increase the rate of interest to a certain category of people like people in the age group 25-30 and give out loans at a lower interest rate to people who are in the age group from 60-70.

Thank You