UPGRAD DS IITB – C62

EDA CREDIT ASSIGNMENT

- Nivethini Senthilselvan
nivethini2530@gmail.com

CONTENTS OF THE ASSIGNMENT

- What is EDA ?
- Business objective of case study.
- Assumptions of the case study.
- Approaches and Methodlogies Used.
- Data visualation of case study
- Conclusion

WHAT IS EDA ?

- In data analytics, exploratory data analysis is how we describe the practice of investigating a dataset and summarizing its main features. It's a form of **descriptive analytics**.
- EDA aims to spot patterns and trends, to identify anomalies, and to test early hypotheses. Although exploratory data analysis can be carried out at various stages of the **data analytics process**, it is usually conducted before a firm hypothesis or end goal is defined.
- Exploratory data analytics often uses visual techniques, such as graphs, plots, and other visualizations. This is because our natural pattern-detecting abilities make it much easier to spot trends and anomalies when they're represented visually.
- As a simple example, outliers (or data points that skew a trend) stand out much more immediately on a scatter graph than they do in columns on a spreadsheet.

BUSINESS OBJECTIVE OF CASE STUDY

- 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.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

ASSUMPTIONS OF THE CASE STUDY

- In order to find the defaulters from application_data.csv, we need to focus on TARGET variable which has only two values either 1 or 0.
- 1 – represents clients with payment difficulties and 0 – represents all other cases when the payment is paid on time.
- First, we need to understand both the type of clients by comparing and analyzing critical variables like age, total income, credit amount, annuity amount, count of children, occupation type etc.
- After completely analyzing the variables for both the type of clients using univariate analysis, we need to understand the dataset by comparing and contrasting two critical variables using graphs and plots.
- For any person, who wants to repay the installments on time, their annuity amount should be less than 40% of their total income.

ASSUMPTIONS OF THE CASE STUDY

- Later, we need to analyze the family status of the client, whether his family is small or big. So that he/she can afford the installments.
- The next thing, we need to analyze the social circle of the client, whether his/her circle has any defaulters already. By doing so, we can predict whether that client may or may not become a defaulter in the near future.
- If the clients social circle has more than 1 defaulter and his last phone change was within 5days, then we need to analyze if those client are worthy of approving loans.
- By merging the two datasets ie application_dataset and previous_application_dataset by matching the current application ID, we need to analyze the previous application status if any. If the previous application is rejected, then we need to understand the reason for its rejection.

STEPS AND METHODOLOGIES TO PERFORM EDA

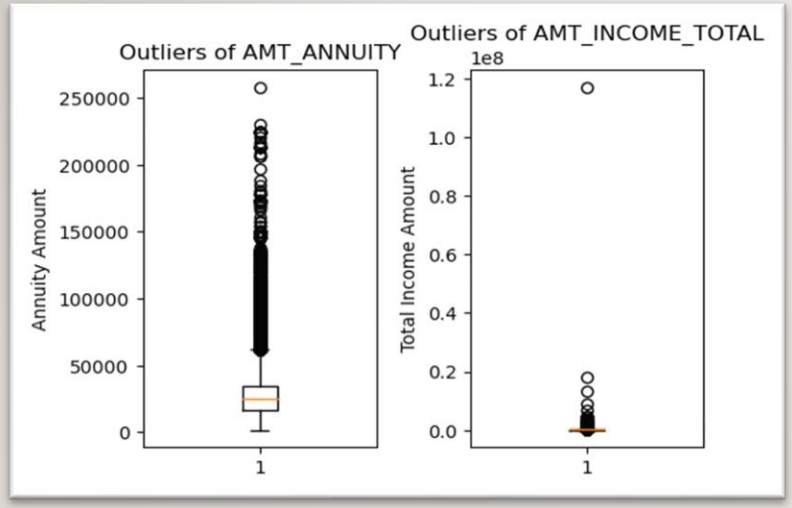
- **DATA CLEANING :**
  - Drop the unnecessary columns in the dataset which are irrelevant to achieve our goal.
  - Check the percentage of null values in each column
  - Drop the columns with more than 40% of null values.
  - For the columns less than 40% of null values, simply impute the appropriate mean, median or mode value of that column
  - Fix the rows and columns by checking whether they have the correct data type and practical values.
  - Identify any outliers in the dataset by analysing each critical column and take necessary actions.
  - Check for data imbalance in the dataset. Especially for the TARGET column. Ignore if the imbalance is practical.

STEPS AND METHODOLOGIES TO PERFORM EDA

- **PERFORM UNIVARIATE, BIVARIATE OR MULTIVARIATE ANALYSIS:**
  - Univariate Analysis – Analyzing and understanding about a single variable.
  - Bivariate Analysis - Analyzing and understanding about two variables.
    - Categorical – Categorical variables
    - Categorical – Numerical variables
    - Numerical – Numerical variables
  - Multivariate Analysis – Analyzing and understanding more than two variables.
  - For Univariate analysis, we can use pie chart, line chart, histogram, etc.
  - For Bivariate analysis, we can use bar plot, scatter plot, etc.
  - For Multivariate analysis, we can use pair plots, heatmaps, etc.

DATA VISUALIZATION OF THE CASE STUDY

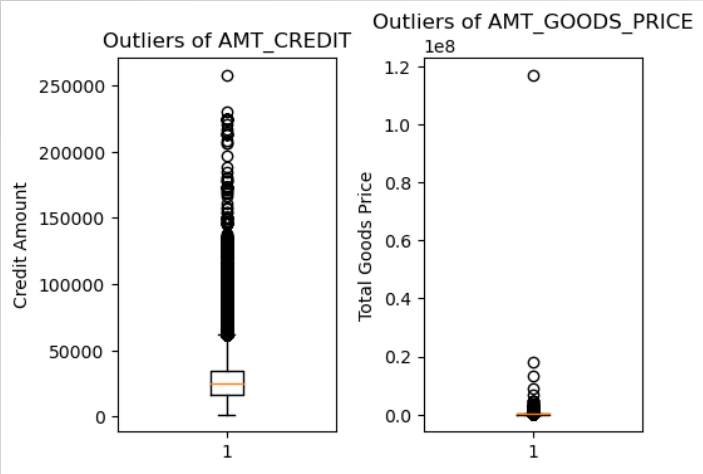
- **UNIVARIATE ANALYSIS OF APPLICATION_DATASET:**

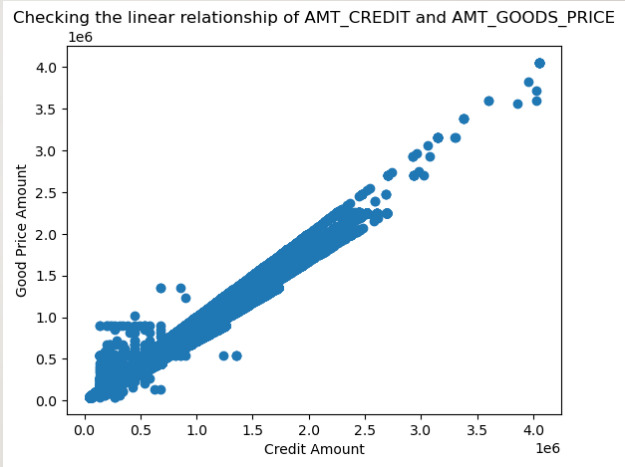


Observation:

- It is observed that, there are outliers present in the AMT_ANNUITY and AMT_INCOME_TOTAL column, where in AMT_ANNUITY the 100th percentile falls between 50,000 and 1,00,000 and IQR is below 50,000.
- The higher values present in the AMT_ANNUITY and AMT_INCOME_TOTAL may represent some exceptional business tycoons and financially successful clients.

DATA VISUALIZATION OF THE CASE STUDY

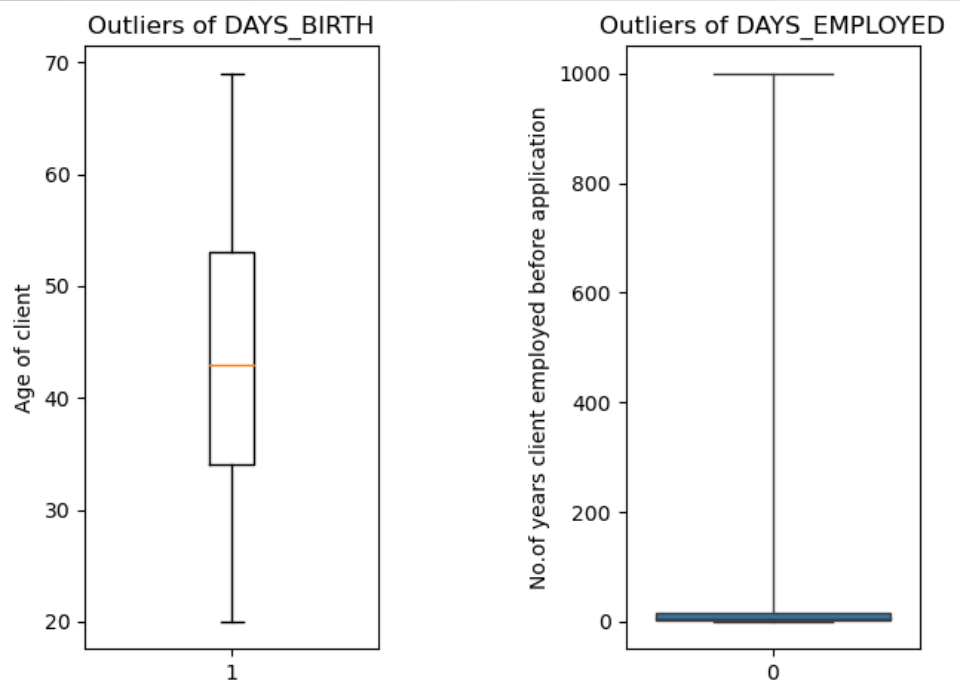




Observation:

1.Though AMT_CREDIT and AMT_GOODS_PRICE has few outliers, it is cut and clear that, there is a high positive correlation between two variables, ie higher the credit amount, higher the goods price.

DATA VISUALIZATION OF THE CASE STUDY

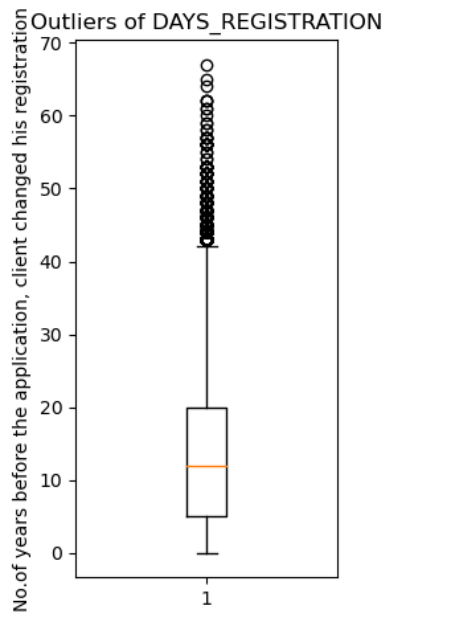


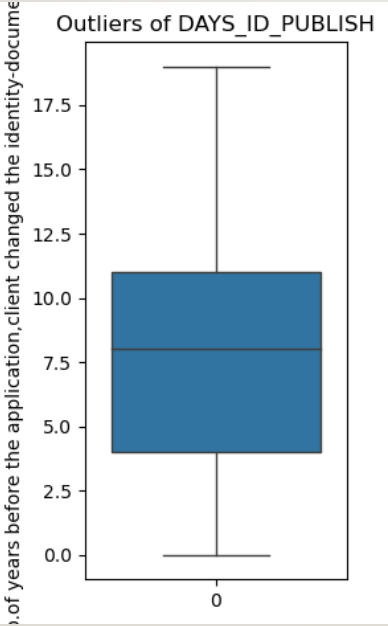
Observation:

1.The DAYS_BIRTH column is absolutely fine having the age group between 20 to 70.

2.The DAYS_EMPLOYED column looks fishy, a max of 50 years service is acceptable and anything above 50 years are incorrect data.

DATA VISUALIZATION OF THE CASE STUDY



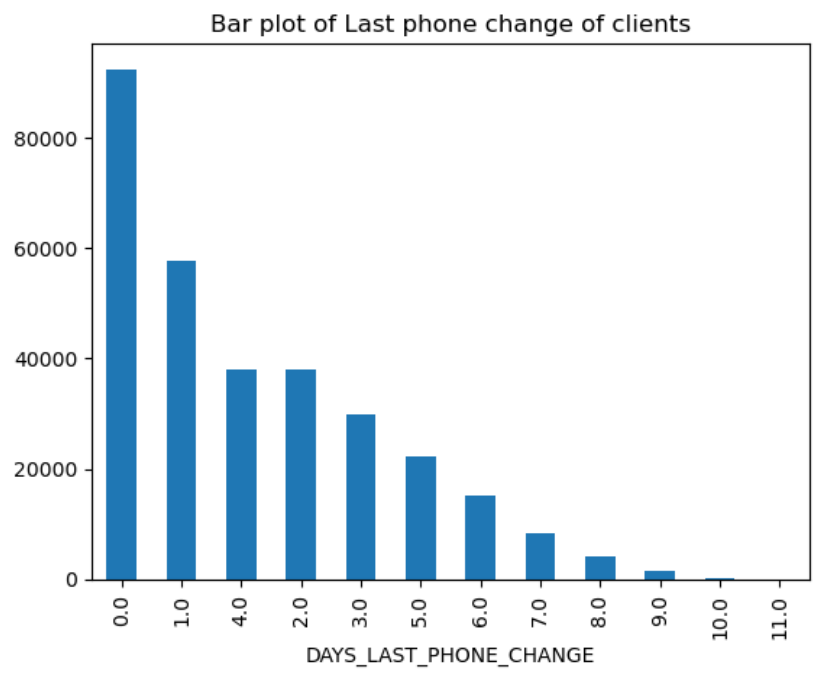


Observation:

1.From the above boxplot, it is obvious that there are some outliers present in DAYS_REGISTRATION column. But these outliers are acceptable because some may live in their ancestral property which may have registered some 50/60 years back.

2.The DAYS_ID_PUBLISH has no outliers and it is perfectly fine.

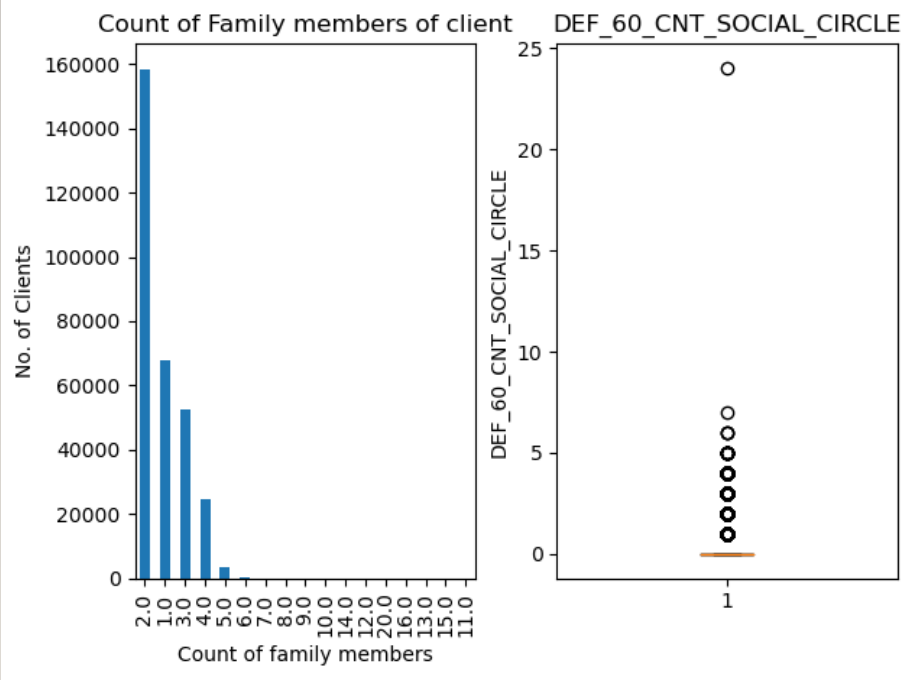
DATA VISUALIZATION OF THE CASE STUDY



Observation:

•It is observed that, more than 80,000 clients have changed their phone number very recently before the application

DATA VISUALIZATION OF THE CASE STUDY



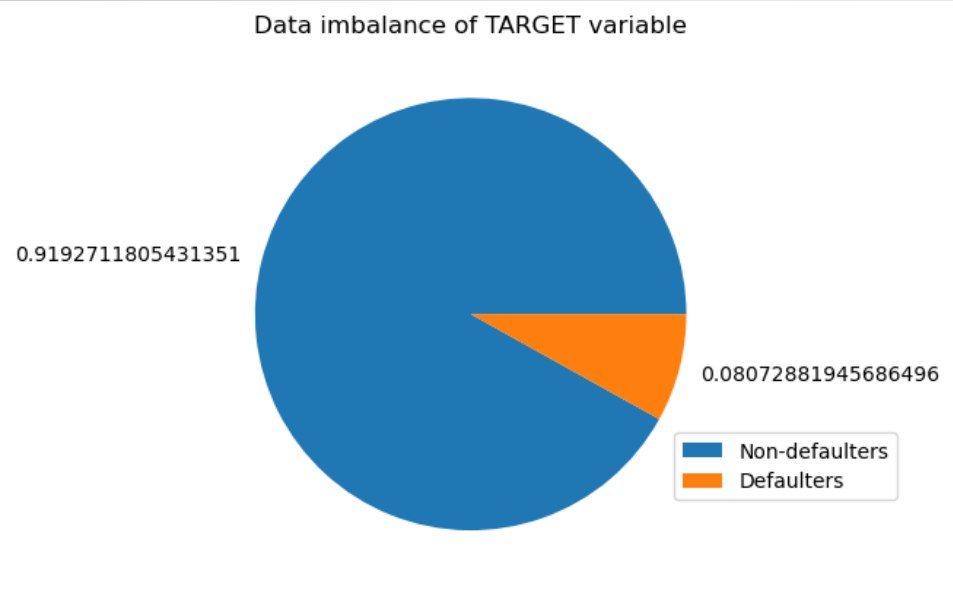
Observation:

1.From the bar plot of CNT_FAM_MEMBERS, it is obvious that most of the clients have two members in the family and a max of 12, which appears to be quite common. Hence no major outlier is seen

2.For the DEF_60_CNT_SOCIAL_CIRCLE column, there are outliers present. Those outliers may be the suspected defaulters we are looking for.

DATA VISUALIZATION OF THE CASE STUDY

- **DATA IMBALANCE:**

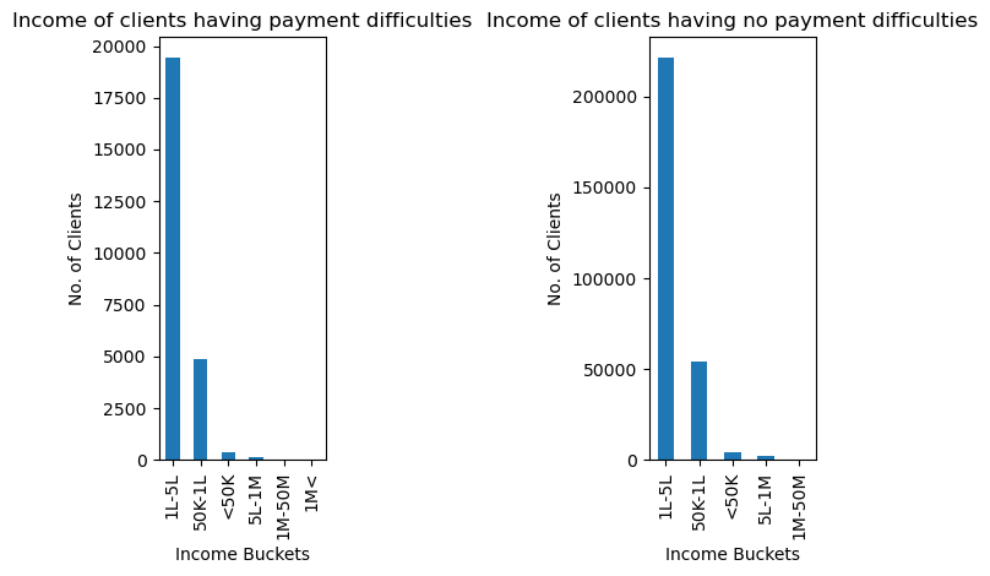


Notes:

•It is obvious that, the no. of non_defaulters greater than no.of defaulters, which is practical in banks. If not, the banks would have corrupted.

DATA VISUALIZATION OF THE CASE STUDY

- **Univariate and Bivariate Analysis:**

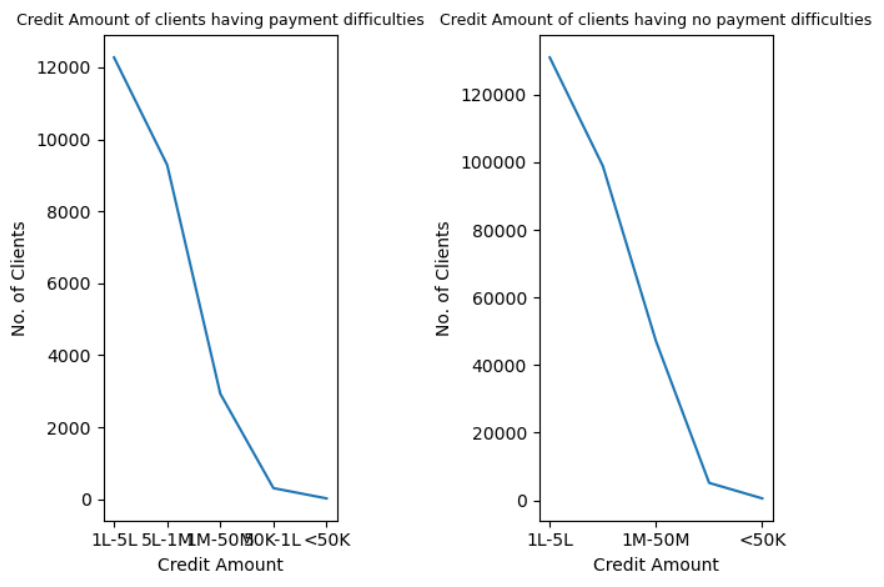


Observation:

1.It is quite surprising that, more client's income falls in the group of 1L-5L in both the bar graphs. But clients with no payment difficulties have more number of people who are receiving an income of 5L-1M than people in clients with payment difficulties.

2.This univariate analysis of AMT_INCOME_TOTAL column of both the types of client does not provide us great insights about the defaulters. Since both the bar graphs looks alike more or less.

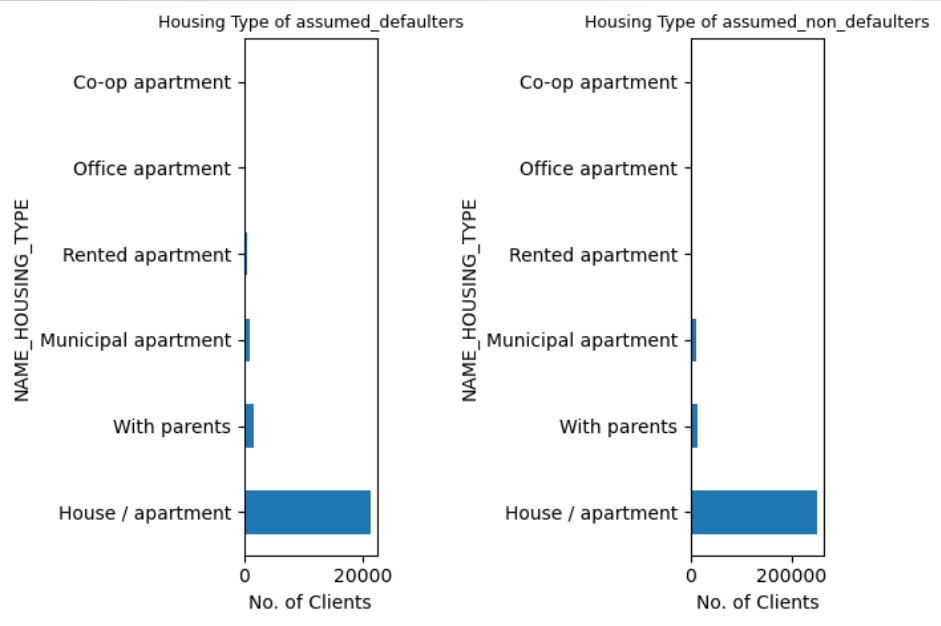
DATA VISUALIZATION OF THE CASE STUDY



Observation:

•The two-line graph of both types of clients looks almost similar. And most of the client from both the types have credit amount between 1L-5L.

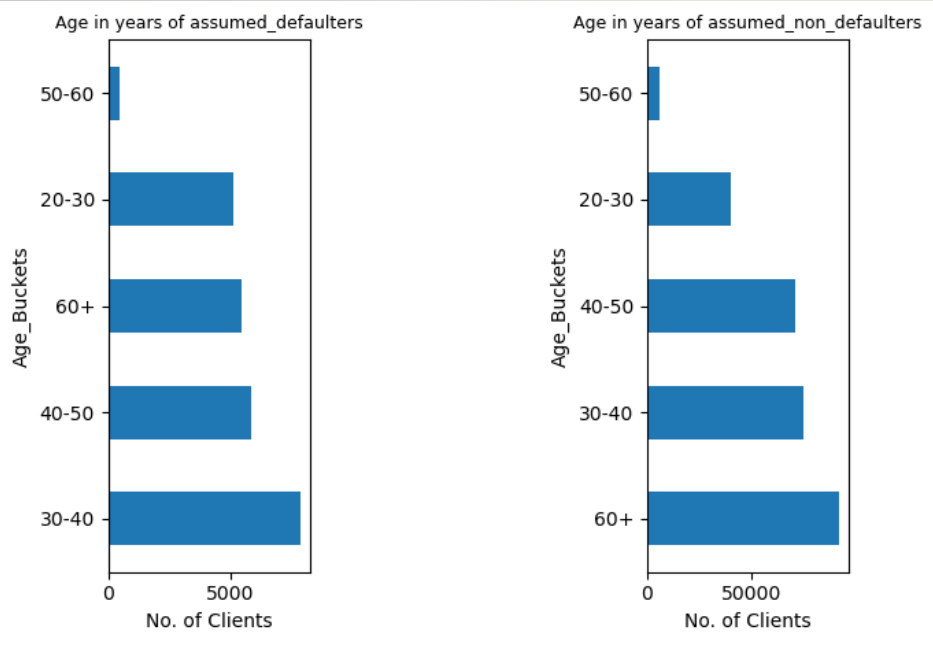
DATA VISUALIZATION OF THE CASE STUDY



Observation:

•Most of the clients of both types, have own house/apartment and with parents. Rented apartment are quite a few clients only. Therefore, 50% of the clients have their own house. This is good fact to be known, because it makes us curious to do our analysis by questioning us, what stops the assumed defaulters to make their payments.

DATA VISUALIZATION OF THE CASE STUDY



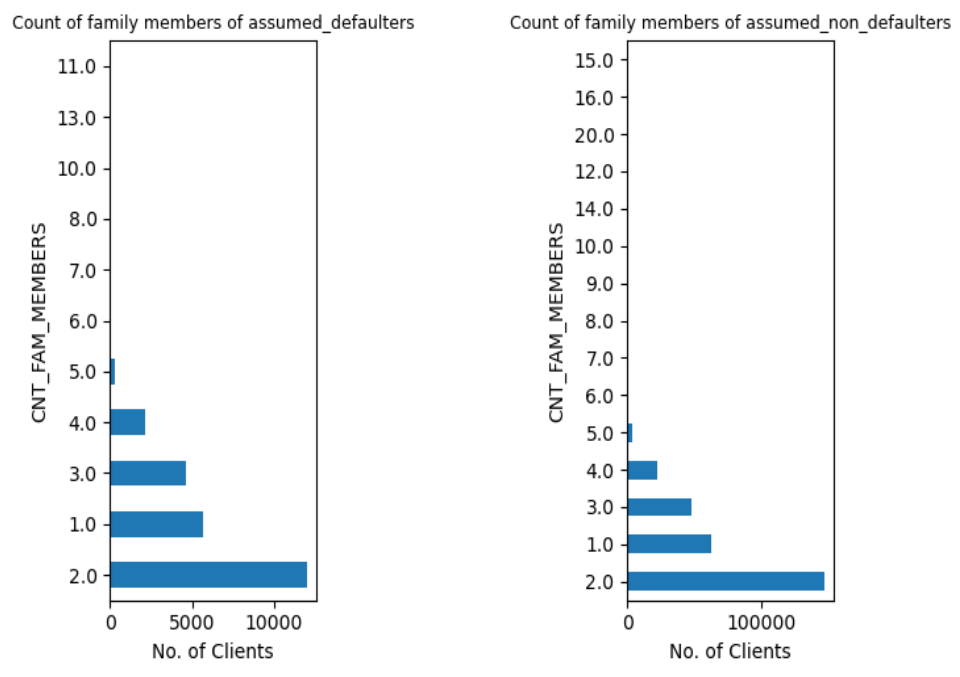
Observation:

•Finally, here comes the difference between two graphs, where most of assumed defaulters are coming under the age group of 30-40, whereas in the assumed_non_defaulters most of the clients were 60+.

•From this difference, it is known that clients who are 60 and above have no difficulties in repaying the loan, since they may receive any pension amount or through other resources.

•But clients under the age buckets of 30-40, who are basically the family man having more responsibilities regarding family growth. So, they are having difficulties in replaying the loan amount. This is practical.

DATA VISUALIZATION OF THE CASE STUDY

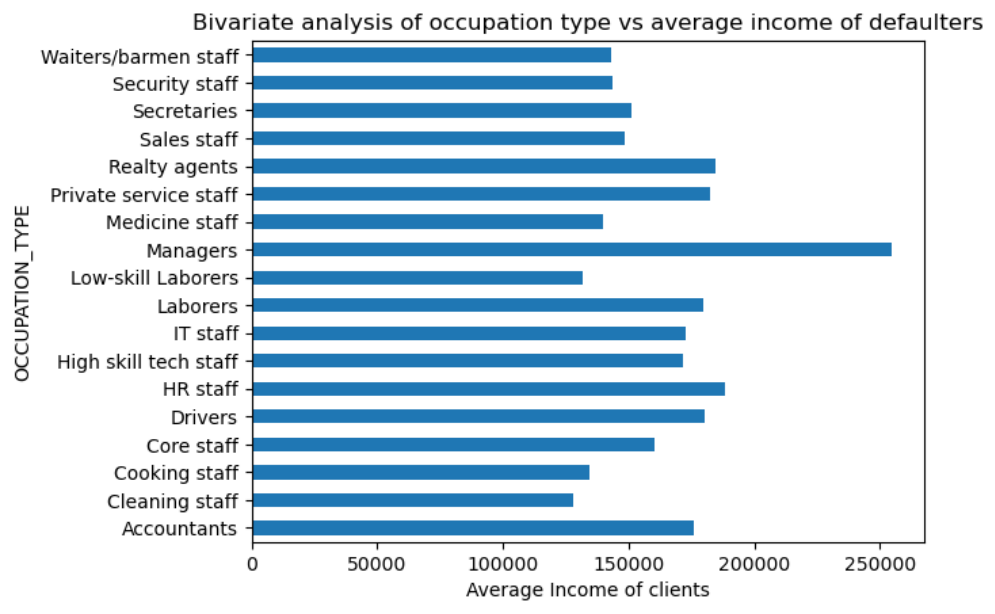


Observation:

•It is quite surprising that, most of assumed defaulters also have only 2 in the family, but what stopping them from replaying the loan is still unknown. Let's analyze much deeper to get the answer.

DATA VISUALIZATION OF THE CASE STUDY

- **Bivariate Analysis:**

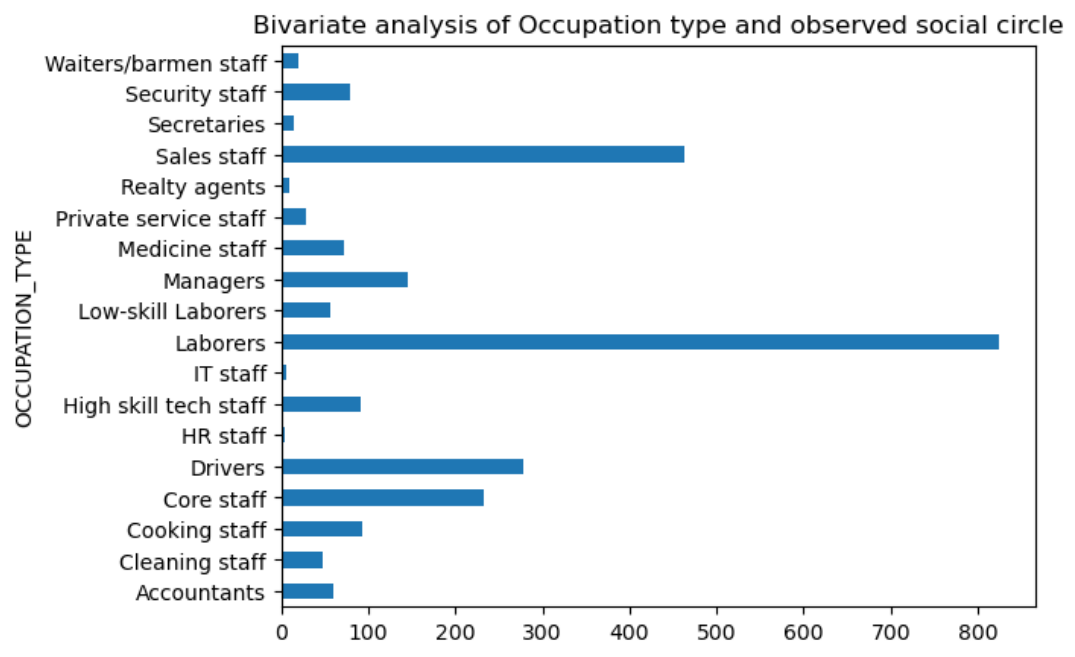


Observation:

•Among all of the occupation type, managers are getting higher income, which is obvious.

•On an average, There is no occupation type receiving income below 1Lakhs.

DATA VISUALIZATION OF THE CASE STUDY



Observation:

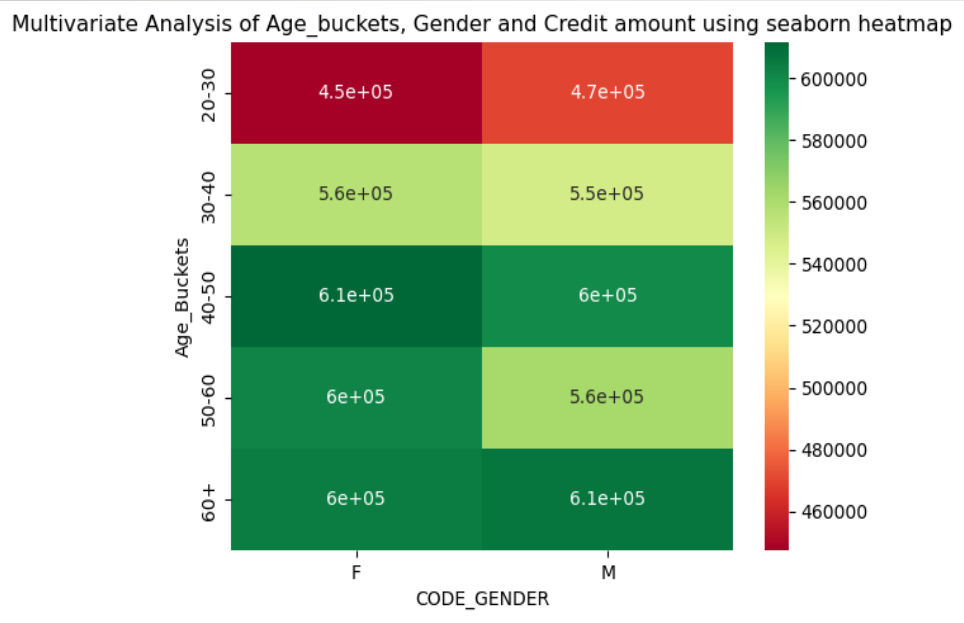
•It gives us great insight above the occupation of the suspected defaulters.

•Laborers have the highest observation of social surroundings defaulted on past 60 days followed by Sales staffs and drivers.

•From the information above, we can suspect most of the Laborers and sales staffs to be defaulters in this new application

DATA VISUALIZATION OF THE CASE STUDY

- **Multivariate analysis**



Observation:

•From the above heatmap, it is obviously seen that, assumed defaulters above 60+ have got the maximum credit amount regardless of gender.

•Whereas assumed defaulters between age group of 30-40 have got comparatively less credit amount regardless of gender.

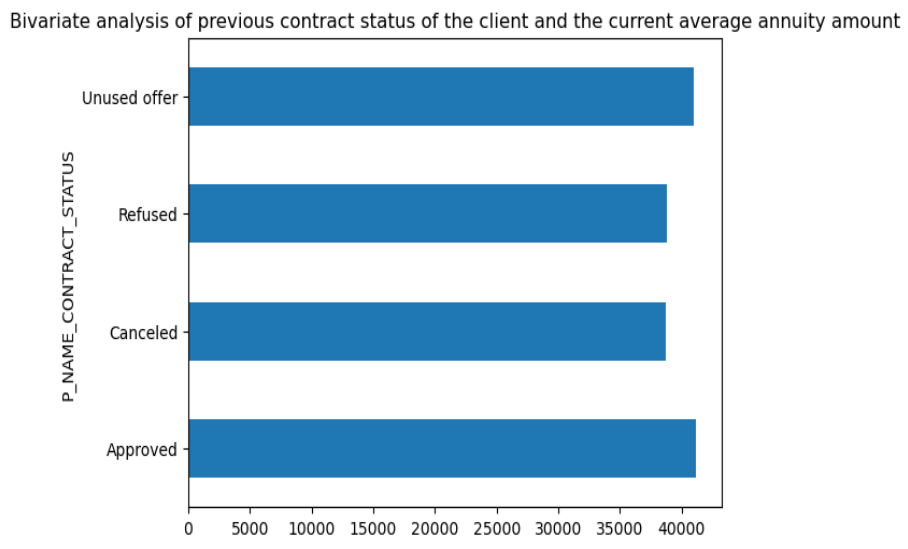
•Female assumed defaulters of age group 50-60 relatively have high credit amount than Male assumed defaulters of same age group.

DATA VISUALIZATION OF THE CASE STUDY

- Merging of two dataframe ie. application_dataset and previous_application_dataset to derive more valuable insights.
- Analyzing the merged dataset and approach towards the goal of finding the defaulters by
- Creating a dataframe called "final_defaulters" by conditioning (merged_dataset.TARGET == 1) & (merged_dataset.DEF_60_CNT_SOCIAL_CIRCLE > 0) & (merged_dataset.DAYS_LAST_PHONE_CHANGE <= 5)

DATA VISUALIZATION OF THE CASE STUDY

- Bivariate Analysis



Observation

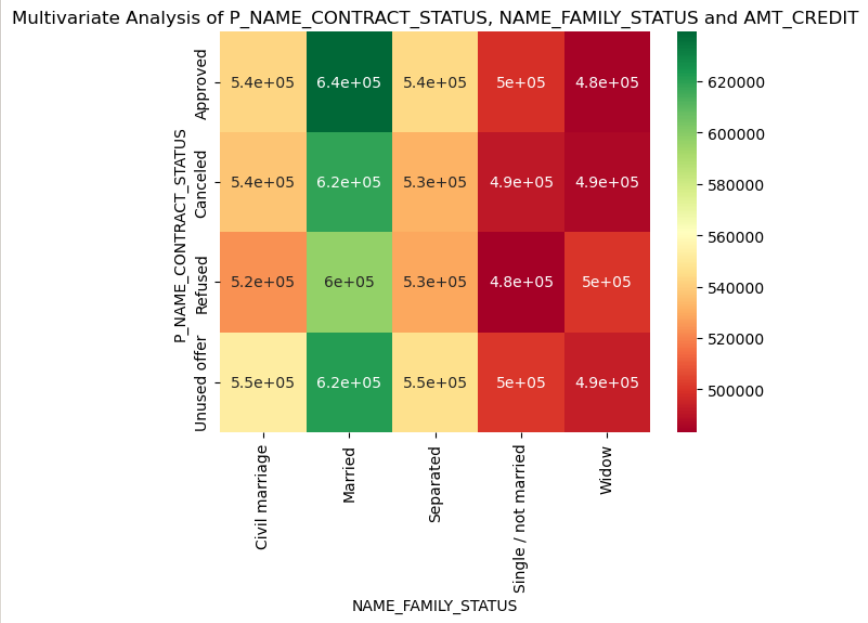
1.There is high current annuity amount for clients who have previous unused offer. This insight may give us prediction like, even in new application, these clients will unuse the offer even after approved.

2.The previously approved application despite of how high the annuity amount is, there are chances where these clients can repay the installments perfectly.

3.The clients who canceled their previous application may have 30% chances of repaying the loan in the new application.

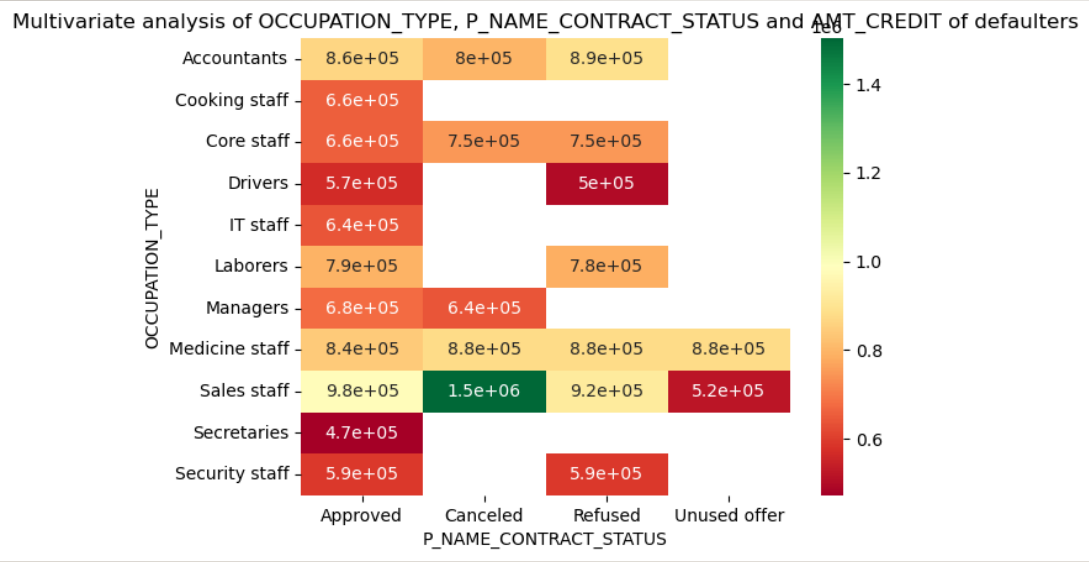
DATA VISUALIZATION OF THE CASE STUDY

- Multivariate Analysis



•From the above heatmap, it is evident that, married clients whose previous application was approved have the highest average credit amount in this new application.
•Single/not married clients who have canceled the previous application have the lowest credit amount in this new application. If their excepted credit amount is not sanctioned, they are chances that these clients would again cancel the application.
•Most of Window client's previous application has been refused. But there are chances for widow to repay the loan only after considering other criterions.

DATA VISUALIZATION OF THE CASE STUDY



Observations:

•Sales staff who cancelled their previous application have high average credit amount in this new application. But the chances of repaying the loan in installments is quite miserable again.

•Security staff whose previous application has been approved have 40% of chances to repay the loan in installments for this new application.

CONCLUSION

- From the analysis made through the case study by EDA process, So far 273 defaulters have been suspected from the entire applications of 3,07,511.
- These 273 predicted defaulters needs to undergo further detailed analysis to predict more accurate clients having payment difficulties for whom the bank may take the necessary actions like denying the loan, reducing the amount of loan, lending at a higher interest rate, etc.