BANK LOAN CASE STUDY

PROJECT SUBMITTED BY HARJAS

PROJECT DESCRIPTION

- The project aims at analyzing the risk appetite of banks. When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
2. If the applicant is not likely to repay the loan, that is, he or she is likely to default, then approving the loan may lead to a financial loss for the company.

PROJECT DESCRIPTION

- The data given contains 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 or she had a late payment of more than X days on at least one of the first Y installments of the loan in our sample.
2. All other cases: all other cases when the payment is paid on time.

- Based on the scenarios a detailed analysis must be conducted and insights need to be drawn to help the bank identify the pattern which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (too risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.

TECH STACK USED

- MICROSOFT EXCEL

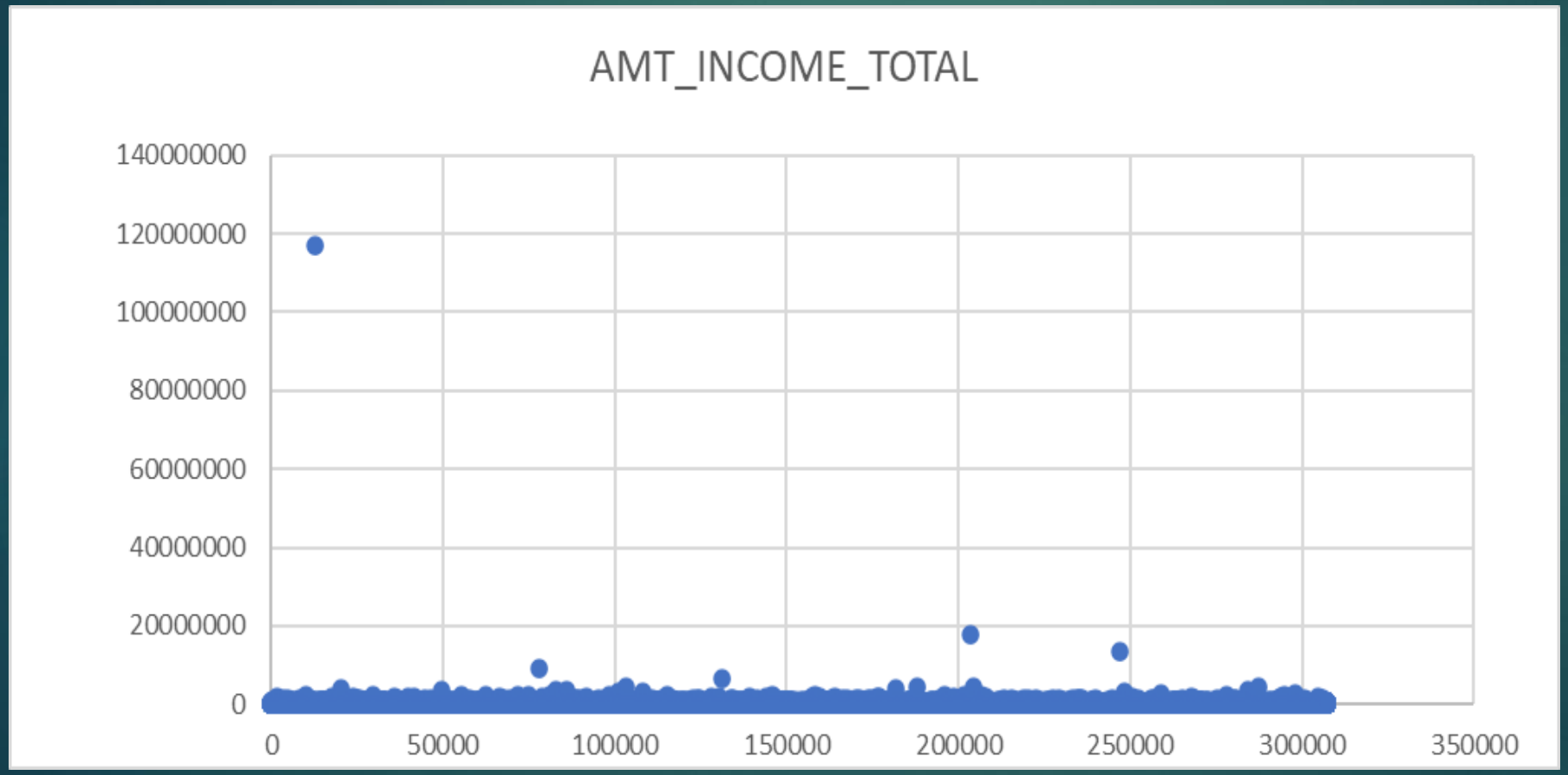
- Excel was used for all the analysis. Excel was also used to generate a graphical representation of the results and to better understand them.

APPROACH

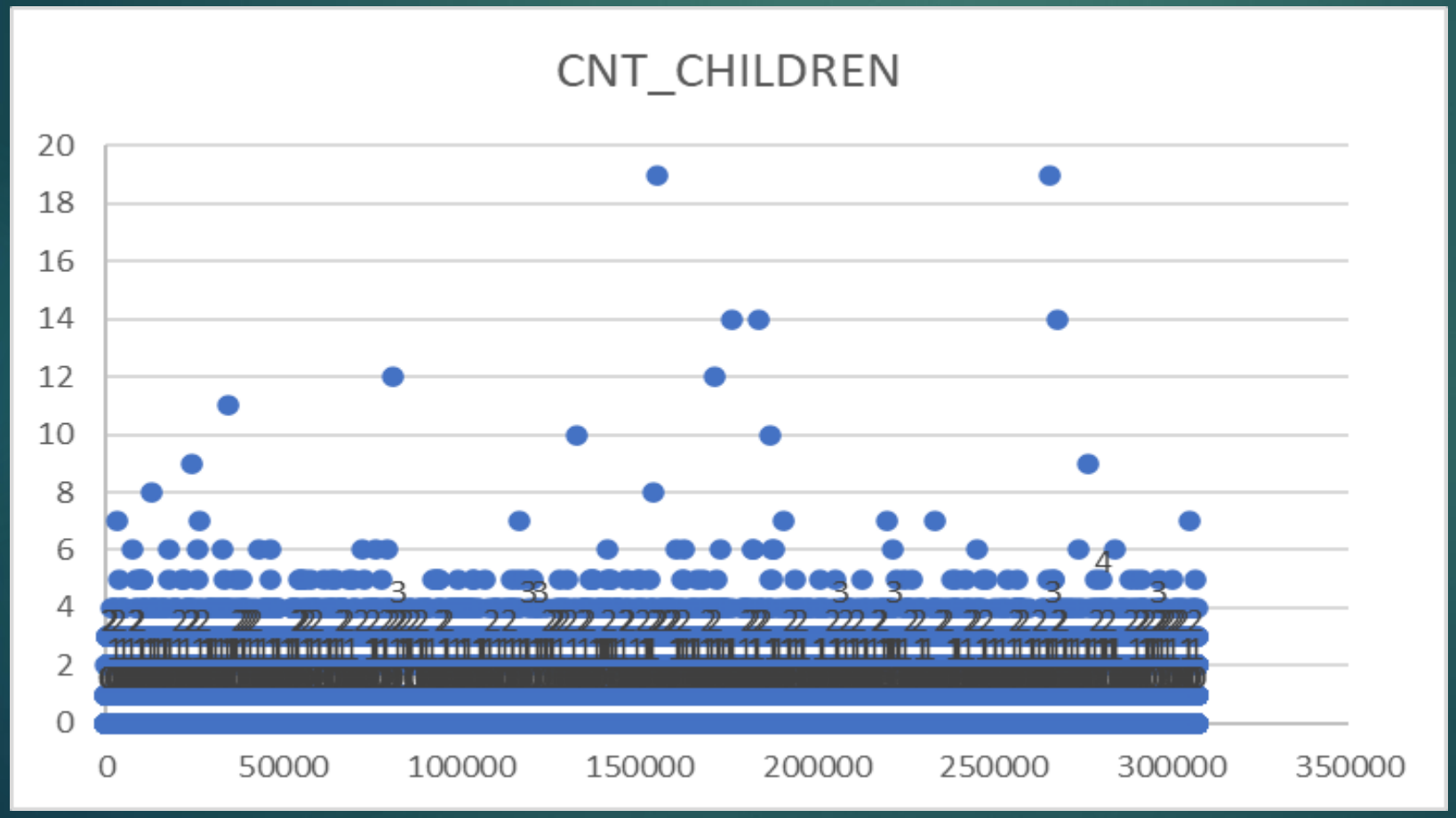
- Used the COUNTA function to count the total rows in each column.
- Secondly, found the percentage of null values in each column using the formula – total row counts for each column / total row counts.
- Further, removed all the columns having null value percentages of more than 30%. For columns having less than 30% null value percentages, have done mean, median, and mode imputations for the missing values for columns having null value percentages less than 30%.
- Also found the outliers using the interquartile range method considering relevant columns.
- After going through each column description, have kept only relevant columns to bring out the insights.
- The columns having days were converted into years by simpy dividing the days by 365.

OUTLIERS

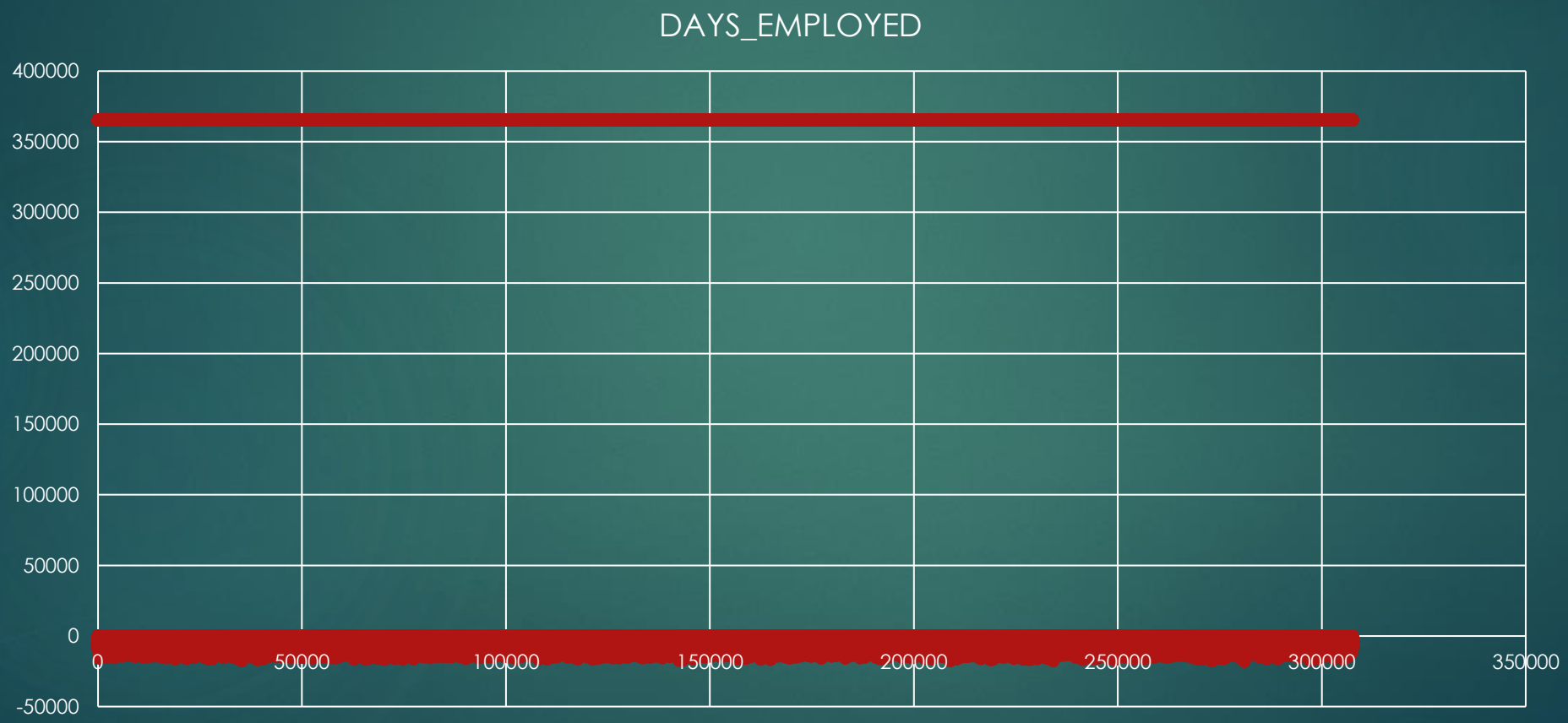
IN THE BELOW DISPLAYED XY PLOTTER ONE CAN SEE THAT THERE ARE INCOMES THAT ARE BEYOND THE LIMIT. THERE ARE APPLICANTS WHO ARE DRAWING AN INCOME OF AROUND 11 CRORES WHEREAS THE MAJORITY OF APPLICANTS ARE DRAWING INCOME IN LACS ONLY. REFER TO THE AMT_INCOME_TOTAL IN THE EXCEL SHEET FOR THE ANALYSIS.



For the column CNT_CHILDREN, there are outliers for target column 0 as well as 1.



For days employed, there are outliers for both target columns 0 and 1.

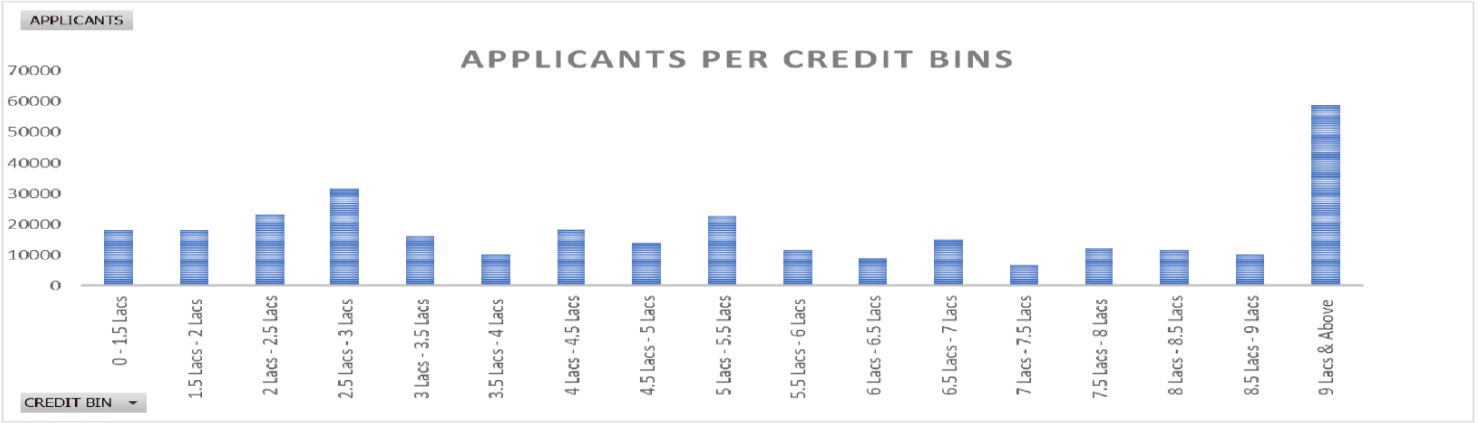


DATA IMBALANCES

- Data imbalance shows the ratio of total applicants with payment difficulties (1) to the total applicants with installments being paid on time (0) to be 11.39.
- That is out of total applications of 3075011, 92% of applicants paid installments on time thus making the majority class, and the rest of the 8% of the applicants had payment difficulties thus making the minority class.

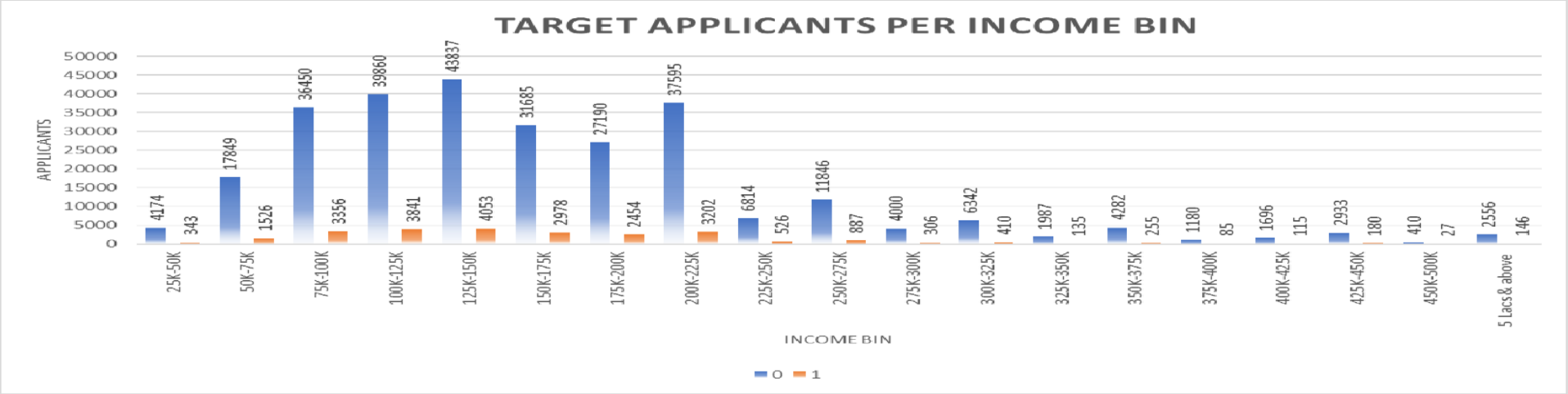
UNIVARIATE ANALYSIS

UNIVARIAITE ANALYSIS



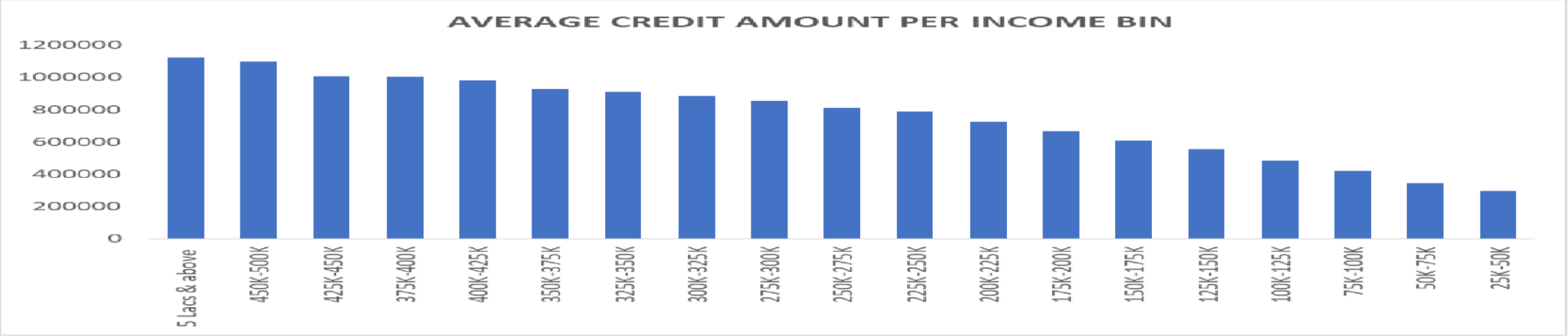
Univariate Analysis refers to the analysis of data that contains only one variable. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The above graph is an example of univariate analysis which depicts simply the count of applicants for the variable AMT_CREDIT grouped in different credit bins. Majority of the applicants were offered loans in the credit range of 9 Lacs and above.

UNIVARIAITE SEGMENTED ANALYSIS



Univariate Analysis refers to the analysis of data that contains only one variable. Segmented analysis here means that the data variable is analyzed in subsets. The above graph is an example of univariate segmented analysis which depicts simply the count of segmented applicants (0 & 1) for the variable AMT_TOTAL_INCOME grouped in different income bins. As evident from the graph there are very few targets 1 applicant who draw an income of more than 50 Lacs and above which can be the reason for the difficulties in the payments. Also, maximum applicants (0,1) draw an income between 1.25 Lacs to 1.5 Lacs but there are applicants which are having payment difficulties despite belonging to the same income range.

BIVARIAITE ANALYSIS



Bivariate Analysis refers to the analysis of data that contains only two variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables. The above graph is an example of bivariate analysis which depicts the relation between AMT_CREDIT and AMT_TOTAL_INCOME. As evident from the graph applicants drawing higher income were offered higher loan amount. Thus, these two variables follow a directionally proportional relation.

CORRELATIONS FOR APPLICANTS WITH PAYMENT MADE ON TIME

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH (Years)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.027	0.003	-0.024	-0.337	-0.245	0.029	0.023
AMT_INCOME_TOTAL	0.027	1	0.343	0.168	-0.063	-0.140	-0.023	-0.187
AMT_CREDIT	0.003	0.343	1	0.101	0.047	-0.070	0.001	-0.103
REGION_POPULATION_RELATIVE	-0.024	0.168	0.101	1	0.025	-0.007	0.001	-0.539
DAYS_BIRTH (Years)	-0.337	-0.063	0.047	0.025	1	0.626	0.271	-0.002
DAYS_EMPLOYED (Years)	-0.245	-0.140	-0.070	-0.007	0.626	1	0.277	0.038
DAYS_ID_PUBLISH (Years)	0.029	-0.023	0.001	0.001	0.271	0.277	1	0.009
REGION_RATING_CLIENT	0.023	-0.187	-0.103	-0.539	-0.002	0.038	0.009	1

** The Analysis can be found on the above attached link on page 4 on sheet "Correlation for Target 0" in excel file Bank Loan Case Study.

The heat map in the above slide shows the correlations between the different variables for the target (0) that is applicants with no payment difficulties.

The color scheme used for the heat map in the above slide is green to white which indicates the strongest correlations are in green and the weakest correlations being in whites.

The most relevant correlations can be seen between the variables are:

- AMT_TOTAL_INCOME to AMT_CREDIT
- DAYS_EMPLOYED to DAYS_BIRTH
- REGION_POPULATION_RELATIVE to AMT_INCOME_TOTAL

CORRELATIONS FOR APPLICANTS WITH PAYMENT DIFFICULTIES

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH (Years)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.005	-0.002	-0.032	-0.259	-0.193	0.032	0.041
AMT_INCOME_TOTAL	0.005	1	0.038	0.009	-0.003	-0.015	0.004	-0.021
AMT_CREDIT	-0.002	0.038	1	0.069	0.135	0.002	0.052	-0.059
REGION_POPULATION_RELATIVE	-0.032	0.009	0.069	1	0.048	0.016	0.016	-0.443
DAYS_BIRTH (Years)	-0.259	-0.003	0.135	0.048	1	0.582	0.253	-0.034
DAYS_EMPLOYED (Years)	-0.193	-0.015	0.002	0.016	0.582	1	0.229	0.003
DAYS_ID_PUBLISH (Years)	0.032	0.004	0.052	0.016	0.253	0.229	1	-0.001
REGION_RATING_CLIENT	0.041	-0.021	-0.059	-0.443	-0.034	0.003	-0.001	1

** The Analysis can be found on the above attached link on page 4 on sheet "Correlation for Target 1" in excel file Bank Loan Case Study.

The heat map in the above slide shows the correlations between the different variables for the target (0) that is applicants with no payment difficulties.

The color scheme used for the heat map in the above slide is green to white which indicates the strongest correlations are in green and the weakest correlations being in whites.

The most relevant correlations can be seen between the variables are:

- AMT_TOTAL_INCOME to AMT_CREDIT
- DAYS_EMPLOYED to DAYS_BIRTH
- REGION_POPULATION_RELATIVE to AMT_INCOME_TOTAL

CONCLUSION

- The project helps in handling large datasets. How EDA can be applied to large datasets. When dealing with large datasets it is also important to select only those columns which are extremely useful to our analysis. Finding correlation columns can become very convenient while dealing with large datasets as it saves time selecting which columns should be considered for analysis. The project helped in understanding the various terminologies used in the banking domain. The insights drawn from the project are included in the next page.

- Applicants drawing higher incomes were offered higher loan amounts by the bank.

  The majority of applicants drew an income range between 1.25 Lacs- 1.5 Lacs, also the defaults drew income between the same range.

  The majority of applicants were offered loans in the credit range of 9 Lacs and above.

THANK YOU