BANK LOAN CASE STUDY

FINAL PROJECT-2

Project Description

This 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:

- 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.

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

- 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.
- All other cases: All other cases when the payment is paid on time.

Approach

I have used COUNTA function to count the total rows in each column. After that I have found the percentage of null values in each column using the formula 1- (Total Row Counts for each columns / Total Row Counts). After that I have removed all the columns having null value percentages more than 30%. For column having less than 30% null value percentages I have done mean, median and mode imputations for the missing values for columns having null value percentages less than 30%. I have also found the outliers using interquartile range method considering relevant columns. After going through each column description, I have kept only relevant columns to bring out the insights. The columns having days are converted in to years by simply dividing the days by 365.Click on the below link to open the excel file. The excel file contain all the analysis.

Excel Worksheet Link→ https://1drv.ms/x/s!As8HBhJc-A2NgTCpwJVNuEcr_-Yn?e=kgXILT

Tech-Stack Used

Microsoft Excel 2019

Purpose – All the analysis has been performed in excel. This tool is also used to create graphical representation of the results and to understand the result set better.

Project Analysis

IDENTIFICATION

We have identified how we will approach the data , finding missing dataset and working on it accordingly to gain the required results I.

OUTLIERS

Identify Outliers and show how they play any role in our data.

IMBALANCE

Understanding the ratio of imbalance in our data.

Results Of Univariate, Segmented Univariate, Bivariate Analysis.

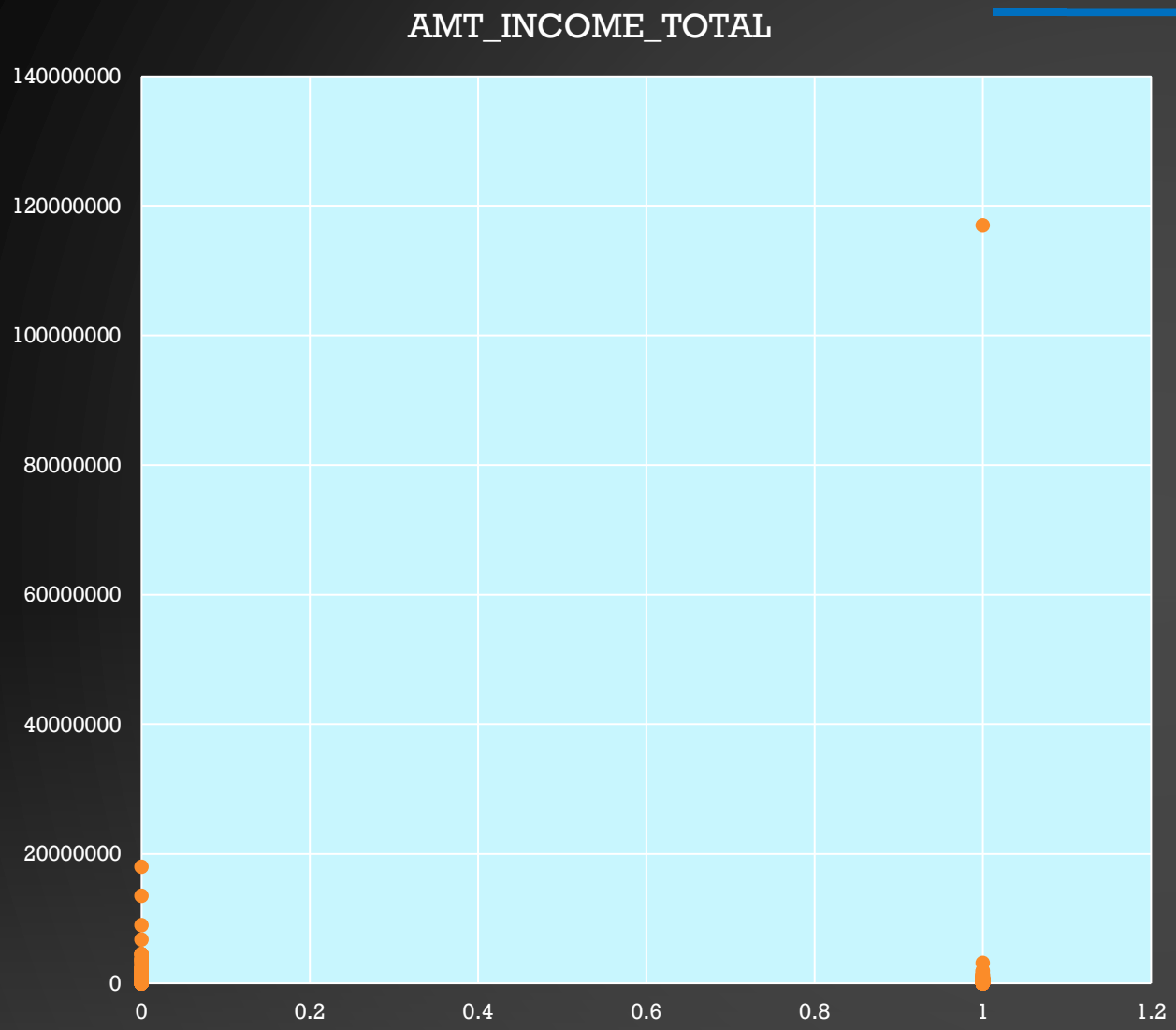
Correlation Analysis

Finding the correlation between the 5 variables with respect to the target variables and find the top three correlation.

VISUALISATION

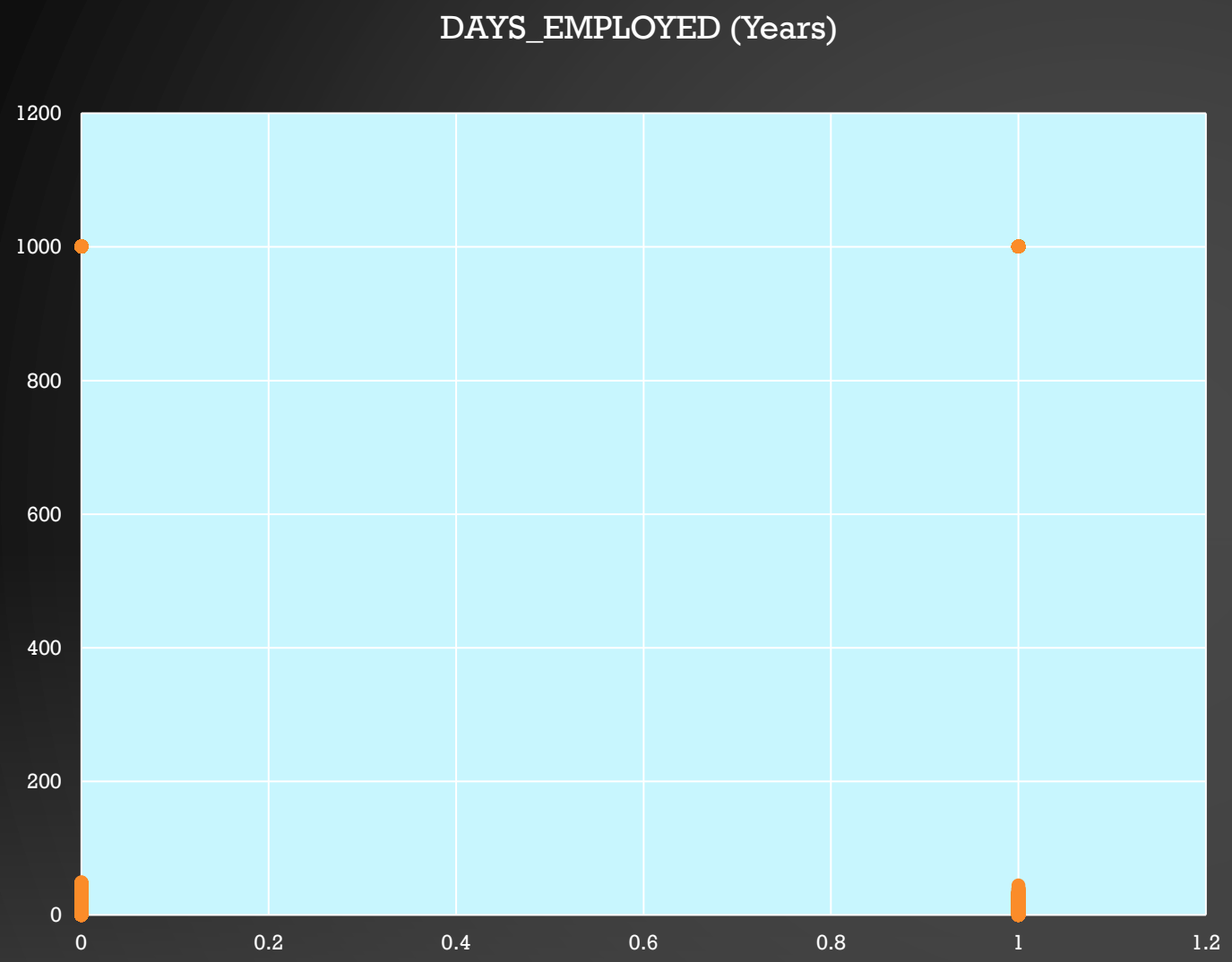
Visualize data with the help of charts and graphs.

Outliers



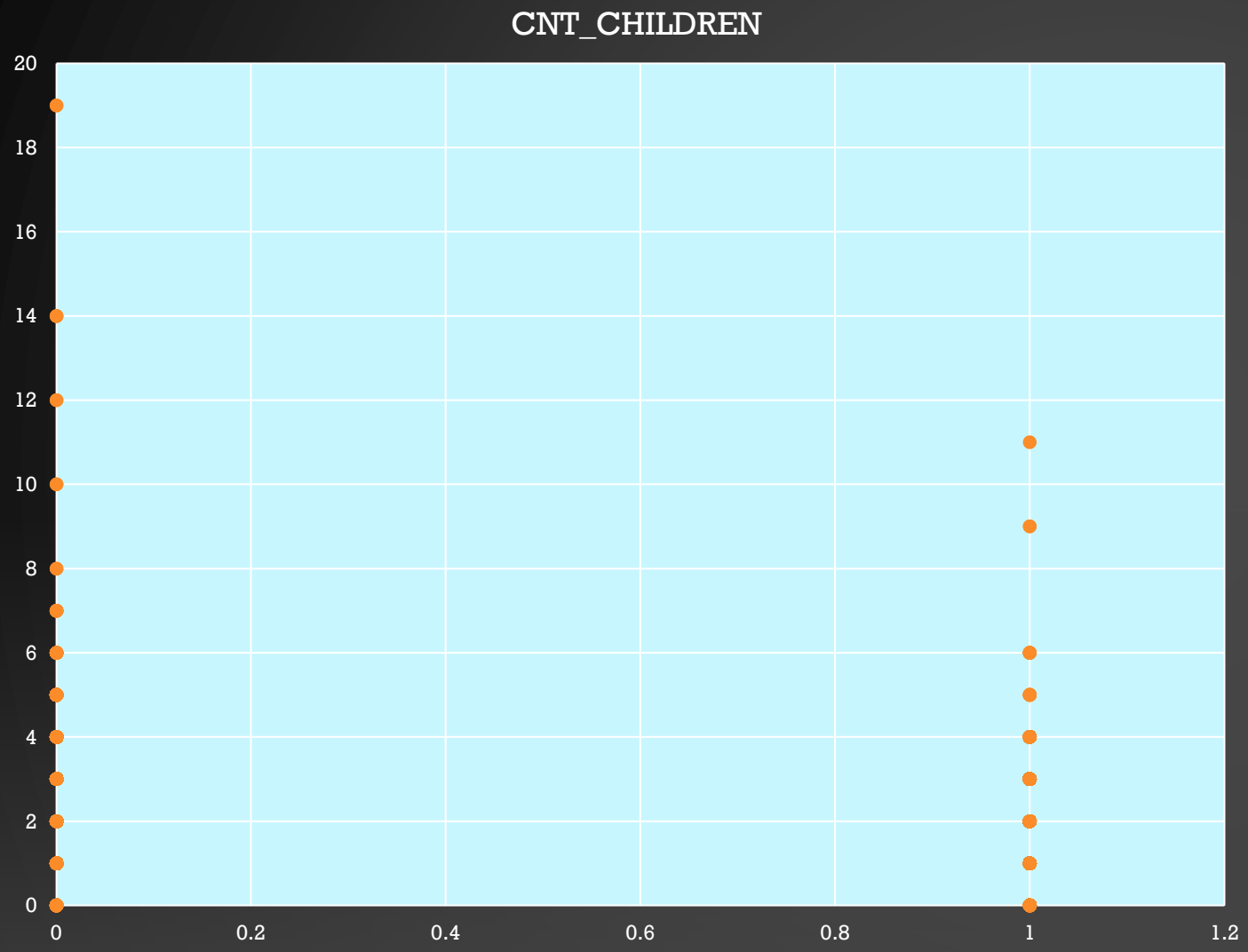
In the above XY plotter we can see that for the target variable 1 there are income which are beyond the limit. There are applicants who are drawing an income of around 11 crores whereas majority of applicants are drawing income in lacs only. For analysis refer the sheet outliers for AMT_TOTAL_INCOME in the above link.

Outliers



In the sheet outliers for Days Employed there are outliers for both target column 0 and 1. The XY plotter shows there are applicants being employed for 1000 years from the day of application which is clearly an anomaly.

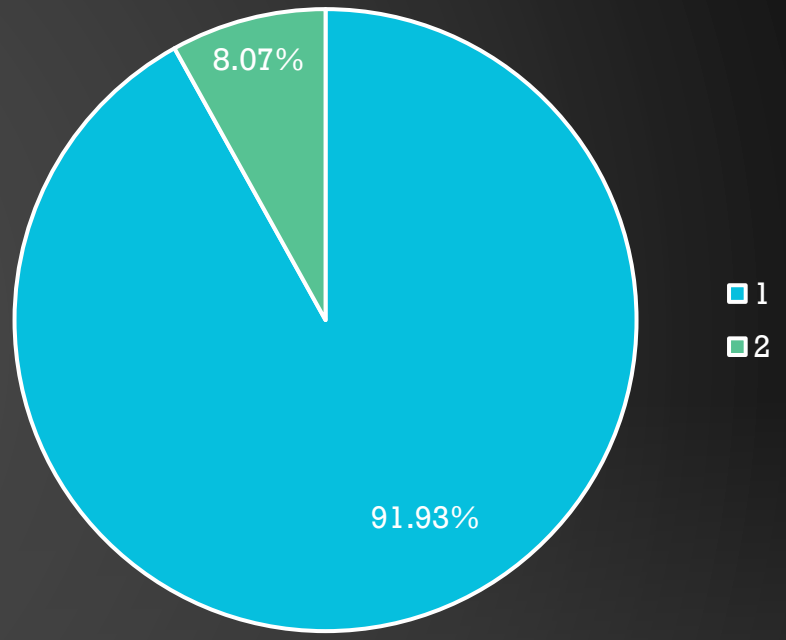
Outliers



In the sheet outliers for CNT_CHILDREN there are outliers for the target column 0 and as well as 1. The XY Plotter for 0 shows 19 children which is highly unusual these days. The XY plotter for 1 shows more than 7 children.

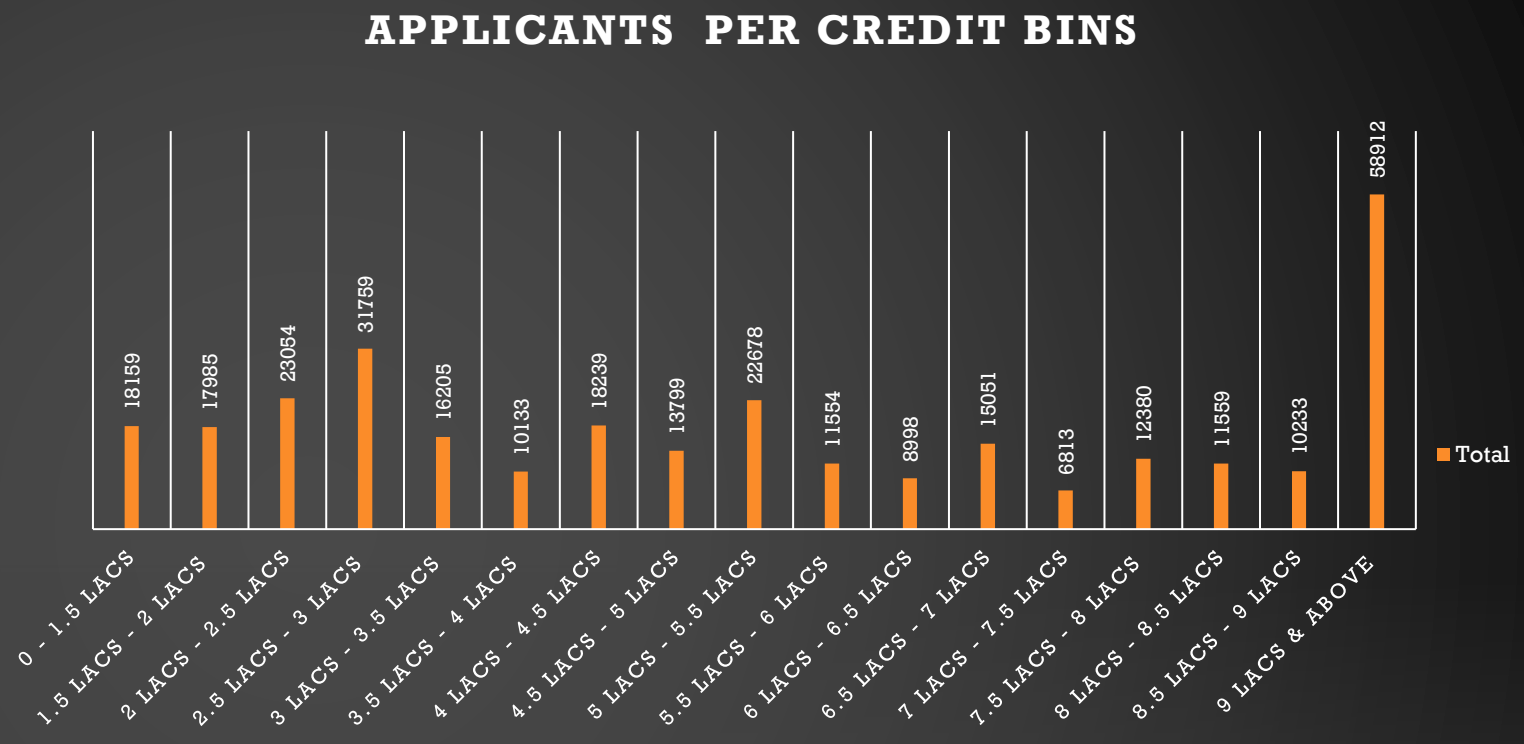
Data Imbalance

In the excel file attached above the sheet 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, 91.93% applicants paid installments on time thus makes the majority class and the rest of the 8.07% of applicants had payment difficulties thus makes the minority class.



UNIVARIAITE ANALYSIS

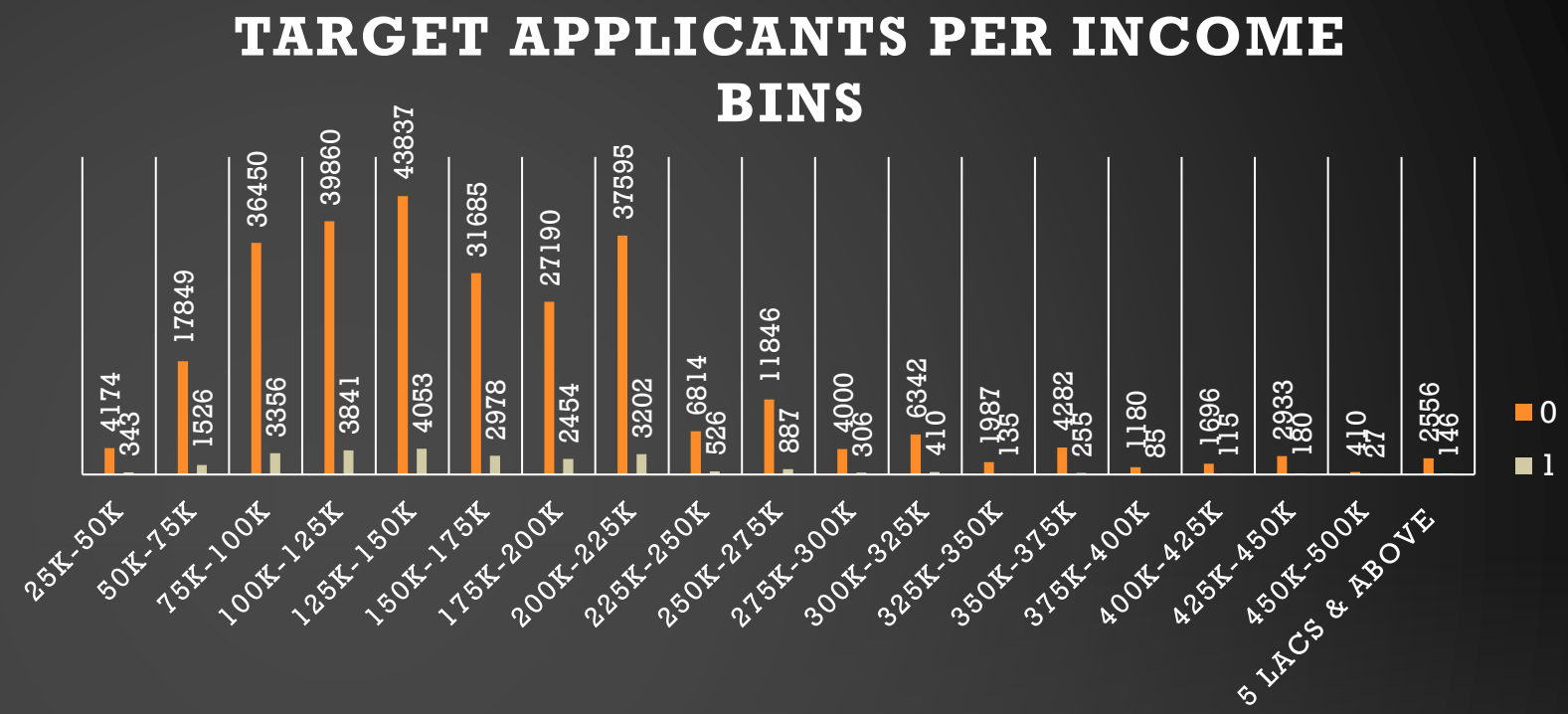
Row Labels	Count of SK_ID_CURR
0 - 1.5 Lacs	18159
1.5 Lacs - 2 Lacs	17985
2 Lacs - 2.5 Lacs	23054
2.5 Lacs - 3 Lacs	31759
3 Lacs - 3.5 Lacs	16205
3.5 Lacs - 4 Lacs	10133
4 Lacs - 4.5 Lacs	18239
4.5 Lacs - 5 Lacs	13799
5 Lacs - 5.5 Lacs	22678
5.5 Lacs - 6 Lacs	11554
6 Lacs - 6.5 Lacs	8998
6.5 Lacs - 7 Lacs	15051
7 Lacs - 7.5 Lacs	6813
7.5 Lacs - 8 Lacs	12380
8 Lacs - 8.5 Lacs	11559
8.5 Lacs - 9 Lacs	10233
9 Lacs & Above	58912
Grand Total	**307511**



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

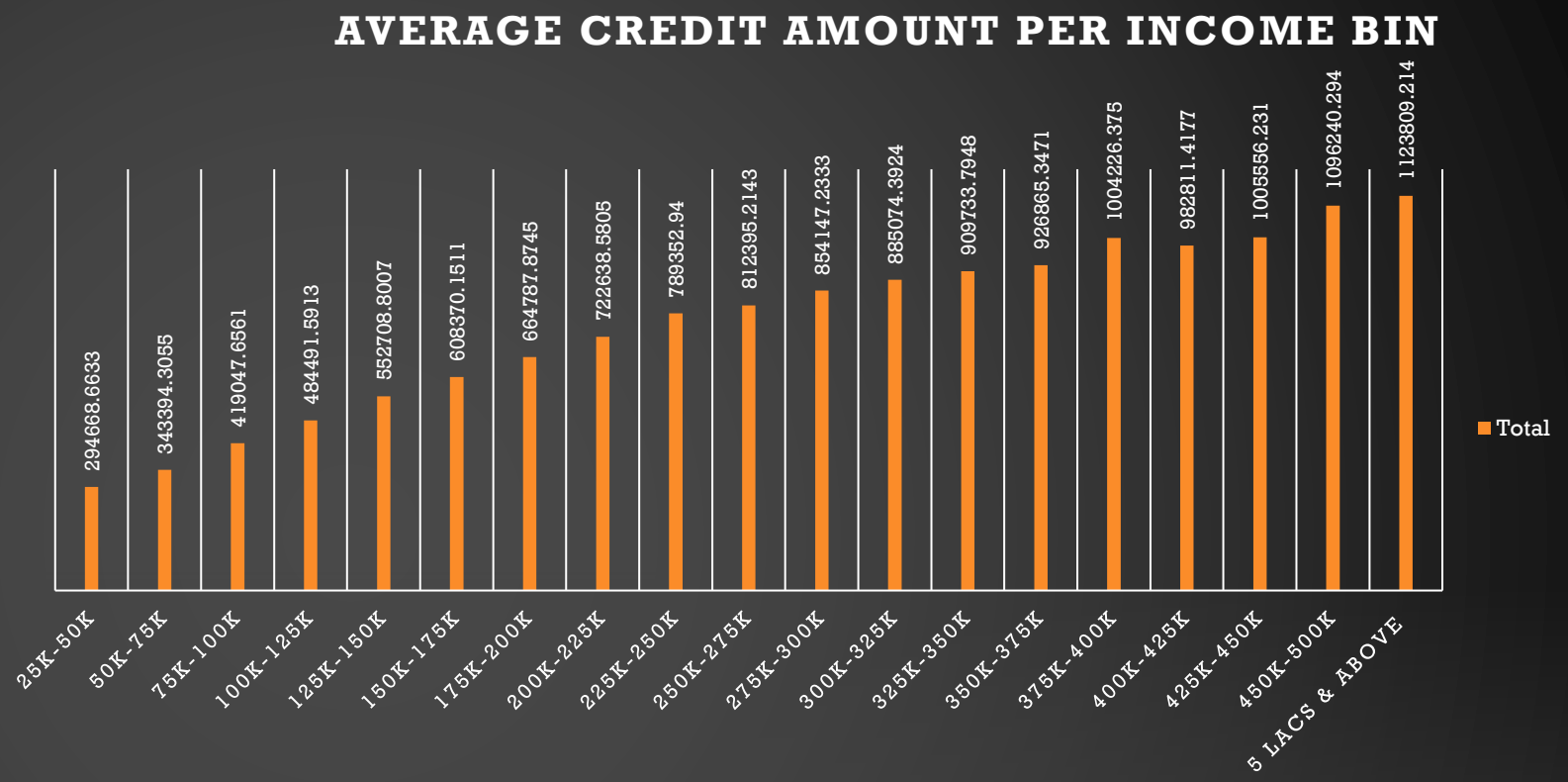
Count of SK_ID_CURR	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
25K-50K	4174	343	4517
50K-75K	17849	1526	19375
75K-100K	36450	3356	39806
100K-125K	39860	3841	43701
125K-150K	43837	4053	47890
150K-175K	31685	2978	34663
175K-200K	27190	2454	29644
200K-225K	37595	3202	40797
225K-250K	6814	526	7340
250K-275K	11846	887	12733
275K-300K	4000	306	4306
300K-325K	6342	410	6752
325K-350K	1987	135	2122
350K-375K	4282	255	4537
375K-400K	1180	85	1265
400K-425K	1696	115	1811
425K-450K	2933	180	3113
450K-500K	410	27	437
5 Lacs & above	2556	146	2702
Grand Total	**282686**	**24825**	**307511**



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

Row Labels	Average of AMT_CREDIT
25K-50K	294668.6633
50K-75K	343394.3055
75K-100K	419047.6561
100K-125K	484491.5913
125K-150K	552708.8007
150K-175K	608370.1511
175K-200K	664787.8745
200K-225K	722638.5805
225K-250K	789352.94
250K-275K	812395.2143
275K-300K	854147.2333
300K-325K	885074.3924
325K-350K	909733.7948
350K-375K	926865.3471
375K-400K	1004226.375
400K-425K	982811.4177
425K-450K	1005556.231
450K-500K	1096240.294
5 Lacs & above	1123809.214
Grand Total	599025.9997



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							
AMT_INCOME_TOTAL	0.027397188	1						
AMT_CREDIT	0.003081225	0.34279945	1					
REGION_POPULATION_RELATIVE	-0.024362658	0.167850636	0.100603799	1				
DAYS_BIRTH (Years)	-0.336966484	-0.062609158	0.047377831	0.025244113	1			
DAYS_EMPLOYED (Years)	-0.245174065	-0.140392466	-0.070104314	-0.007197856	0.626113878	1		
DAYS_ID_PUBLISH (Years)	0.028750653	-0.022896393	0.00146417	0.001070788	0.271314395	0.27666316	1	
REGION_RATING_CLIENT	0.022842107	-0.186573418	-0.103336744	-0.539004783	-0.002332327	0.038327694	0.00899835	1

** The Analysis can be found on the above attached link on page 3 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 red to green which indicates the strongest correlations are in red and the weakest correlations being in green.
- The most relevant correlations can be seen between the variables are:

→AMT_TOTAL_INCOME to AMT_CREDIT

→REGION_POPULATION_RELATIVE to AMT_INCOME_TOTAL

→DAYS_EMPLOYED to DAYS_BIRTH

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							
AMT_INCOME_TOTAL	0.004795787	1						
AMT_CREDIT	-0.001674961	0.038131435	1					
REGION_POPULATION_RELATIVE	-0.0319749	0.009134586	0.069161087	1				
DAYS_BIRTH (Years)	-0.259108666	-0.003096245	0.135316369	0.048190366	1			
DAYS_EMPLOYED (Years)	-0.192863828	-0.014977396	0.001930183	0.015531849	0.582185148	1		
DAYS_ID_PUBLISH (Years)	0.032298597	0.004214856	0.05232898	0.015536882	0.252862836	0.229090254	1	
REGION_RATING_CLIENT	0.040680482	-0.021486257	-0.059192754	-0.443235509	-0.033927932	0.003489989	-0.001397237	1

** The Analysis can be found on the above attached link on page 3 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 (1) that is applicants with payment difficulties.
- The color scheme used for the heat map in the above slide is red to green which indicates the strongest correlations are in red and the weakest green being in yellow.
- 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

This project helps in handling the large datasets. How exploratory data analysis can be applied to large datasets. When dealing with the large datasets it is also important to select only those columns which are extremely useful to our analysis. Finding correlations columns can become very convenient while dealing with large datasets as it saves time selecting which columns should be considered for analysis. The project also helps in understanding the various terminologies used in the banking domain. The insight drawn from the project are as follows:

→Applicants drawing higher income were offered higher loan amount by the bank.

→Majority of applicants drawn an income range between 1.25 Lacs – 1.5 Lacs, also the defaults drawn income between the same range.

→Majority of applicants were offered loans in the credit range of 9 Lacs and above.