BANK LOAN CASE STUDY

Davidraju Lakkamthoti

AGENDA

Project description

Approach

Tech-stack used

Insights

Result

Excel file hyperlink

PROJECT DESCRIPTION

This project is about Risk analytics of a bank who is lending loans.

some customers who don't have a sufficient credit history take advantage of this and default on their loans.

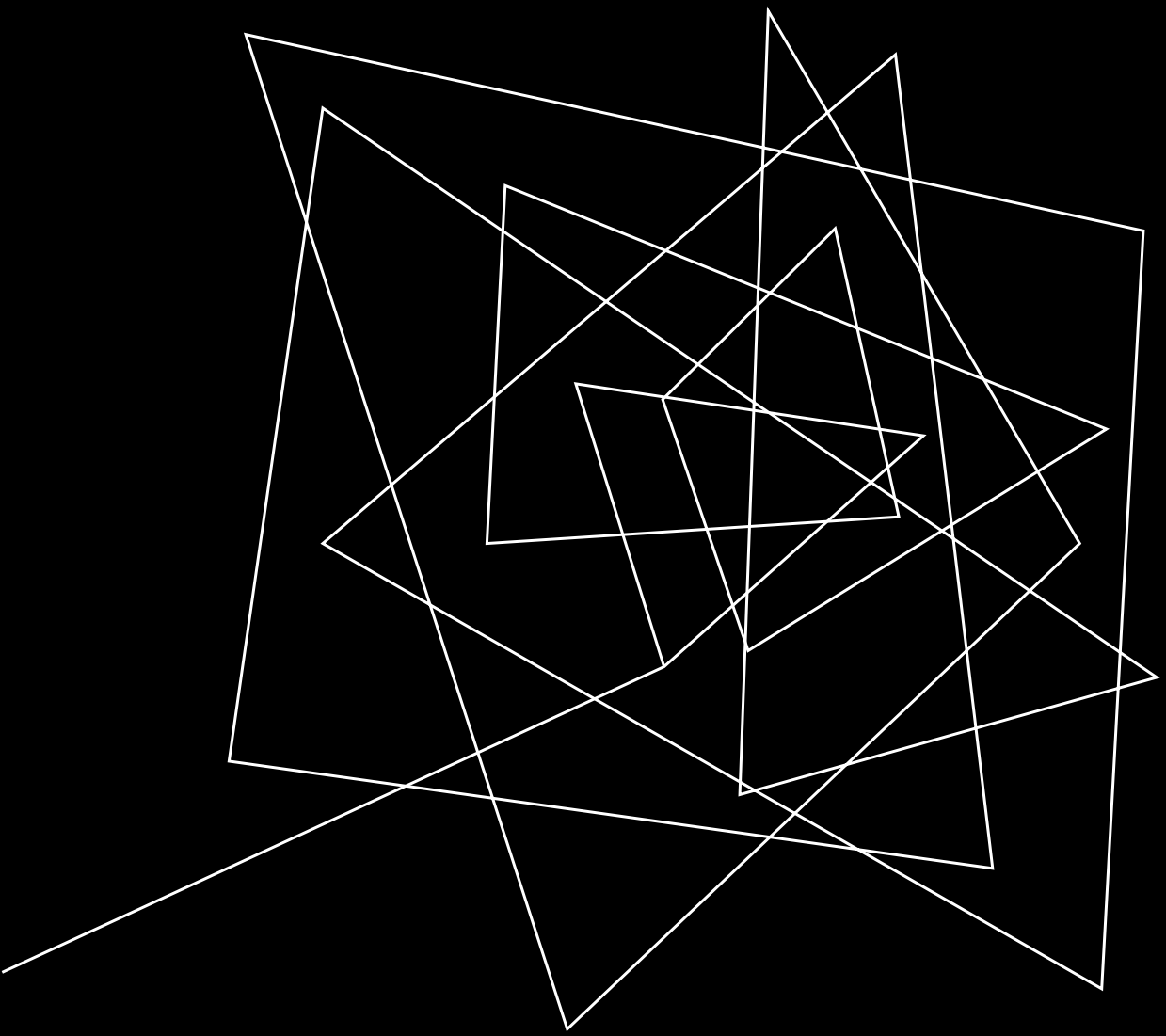
Some customers can repay the loan but is not approved, the bank loses business.

The dataset we'll be working with contains information about loan applications. It includes two types of scenarios:

1.Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.

2.All other cases: These are cases where the payment was made on time.

Excel file hyperlink



APPROACH & TECH STACK USED

In every data analytics project , we first basically clean our data.

That included: 1) Finding the percentage of null values in each column and eliminating the columns with more than 30% of null values.

2) Calculated Inter Quartile Ranges to find outliers and vomit.

3) Imputation of mean values for the missing values in relevant columns.

Then did the tasks asked.

Ms Excel used for whole project.

INSIGHTS

Outliers in AMT_TOTAL_INCOME

Quartile 1
112500

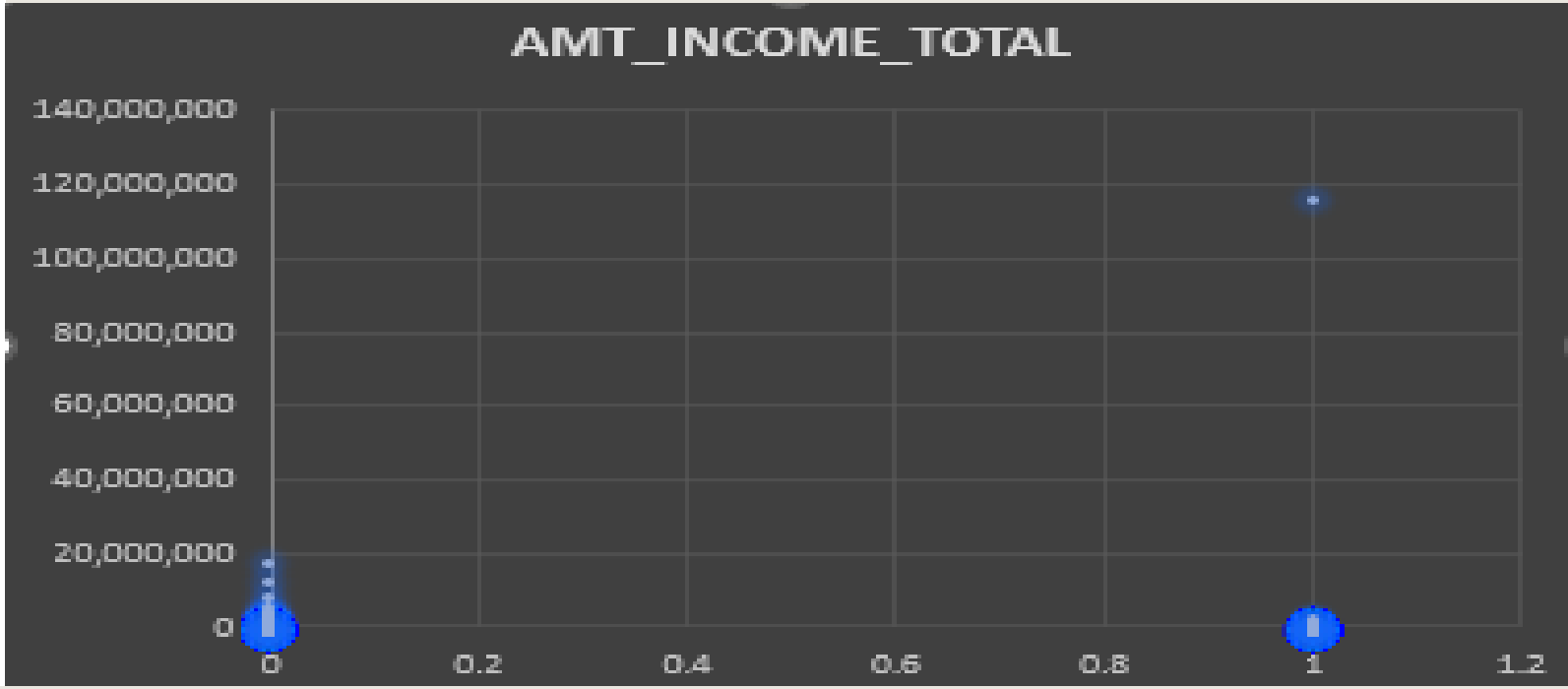
Quartile 3
202500

Inter Quartile Range
90000

Upper Limit
337500

Lower Limit
-22500

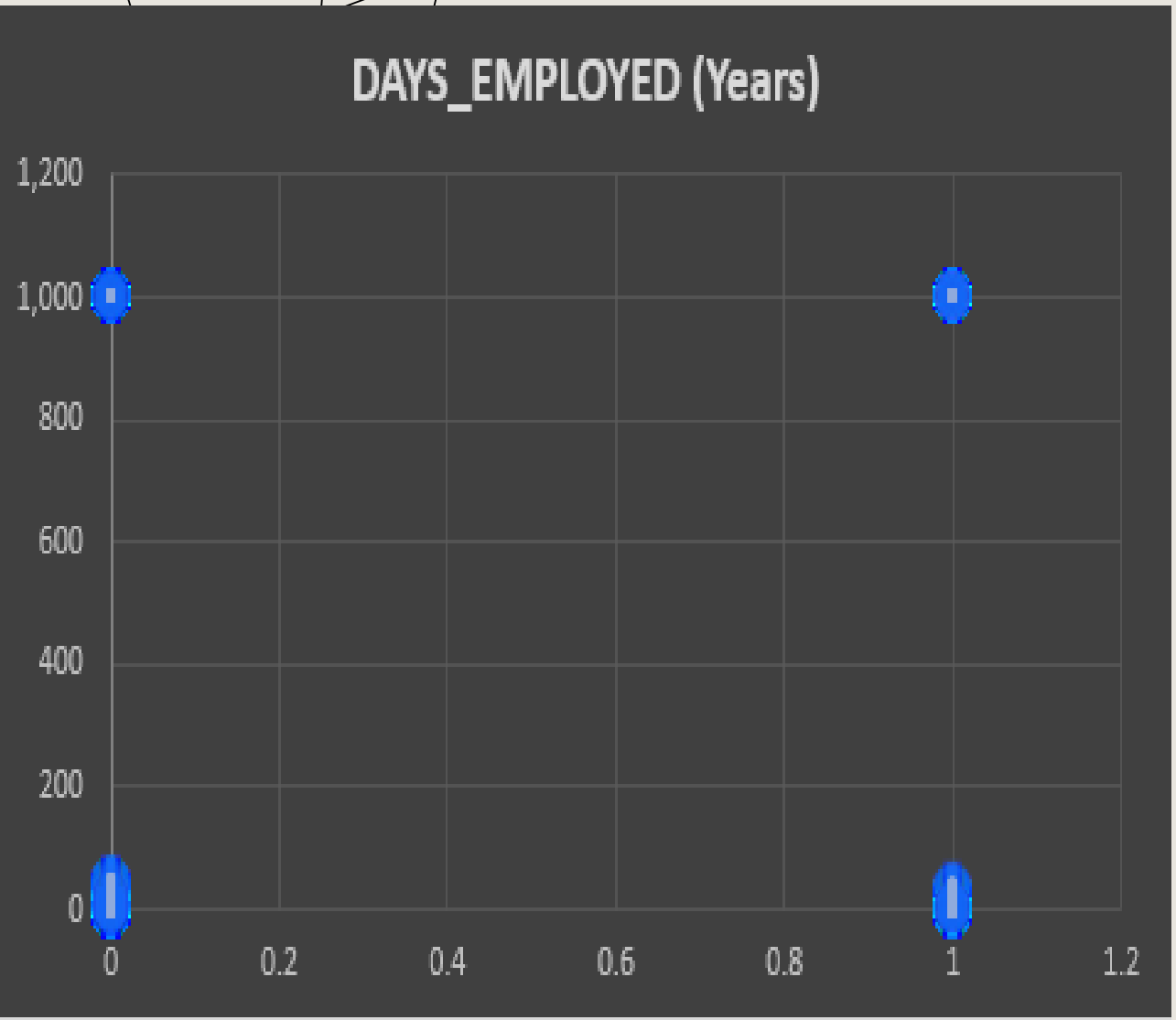
AMT_INCOME_TOTAL	
Mean	168797.9193
Standard Error	427.6058332
Median	147150
Mode	135000
Standard Deviati	237123.1463
Sample Variance	56227386501
Range	116974350
Minimum	25650
Maximum	117000000
Sum	51907216961
Count	307511

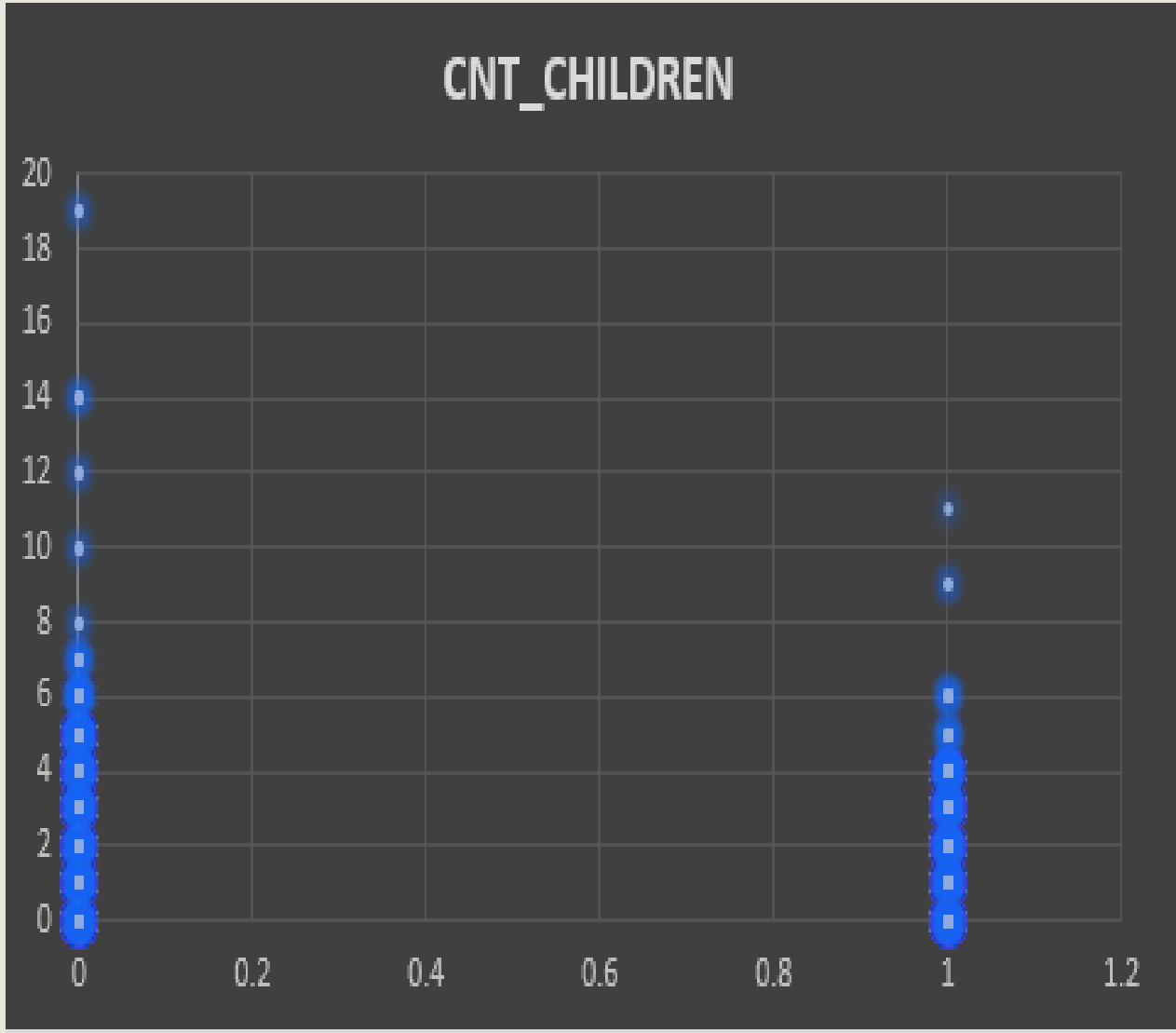


Here, we can see the IQR is calculated to find the outliers. In the AMT_INCOME_TOTAL column we have outliers for target variable 1 with an income more than 10 crores .As there are majority of the people earning in lakhs only.

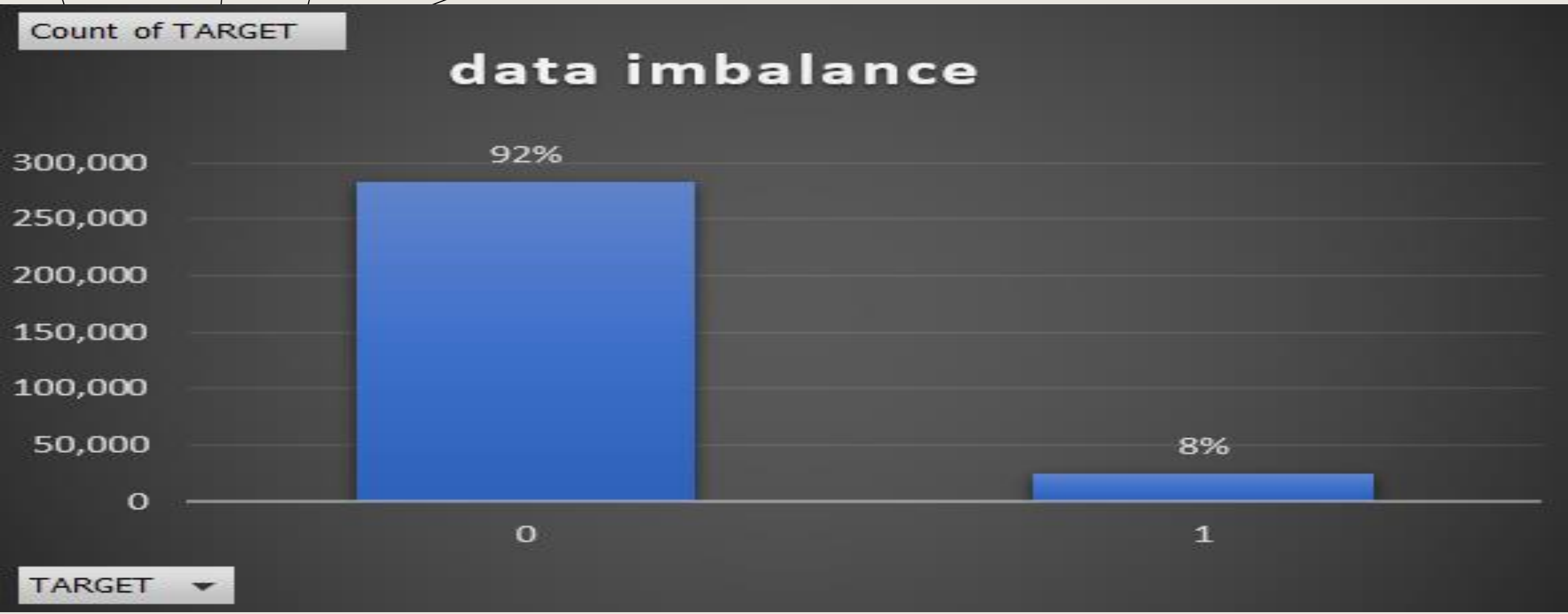
Outlier plots for other relevant columns are shown in next slides

OUTLIER PLOTS





DATA IMBALANCE ANALYSIS



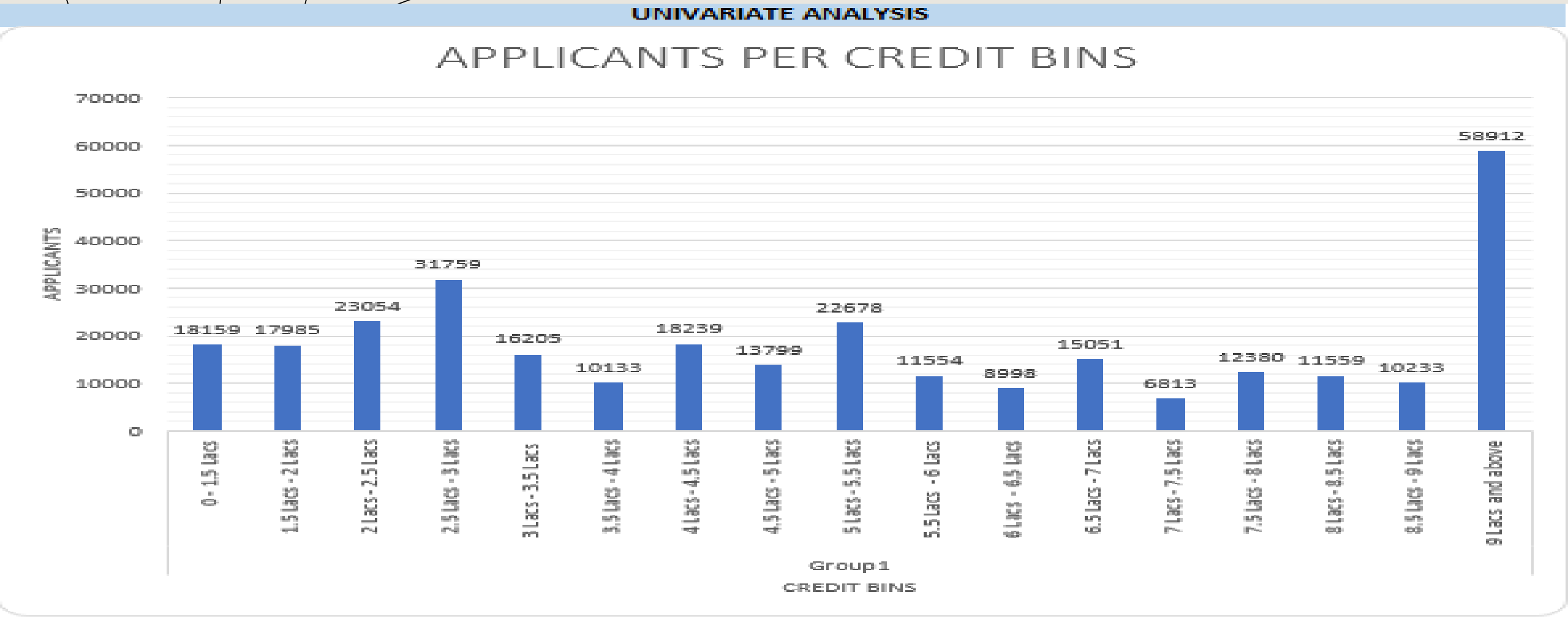
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

Here, Target 0 is the percentage of people making on time payments, Target 1 is people missing the bill date.

We have a ratio of 11:39 for Target 0 and Target 1.

So, 92% of people are making on time payments only 8% people are failing to pay on time.

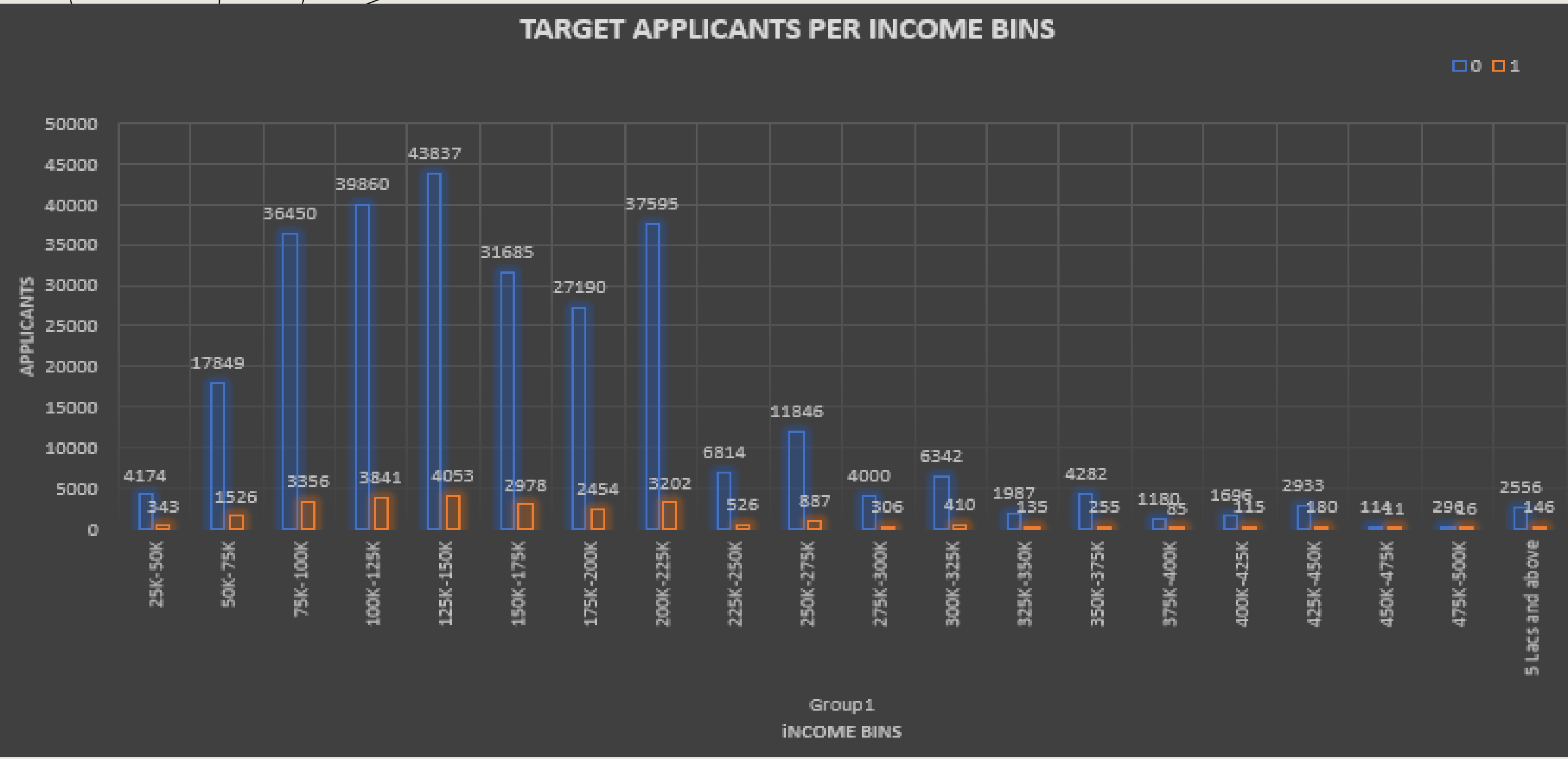
UNIVARIATE ANALYSIS



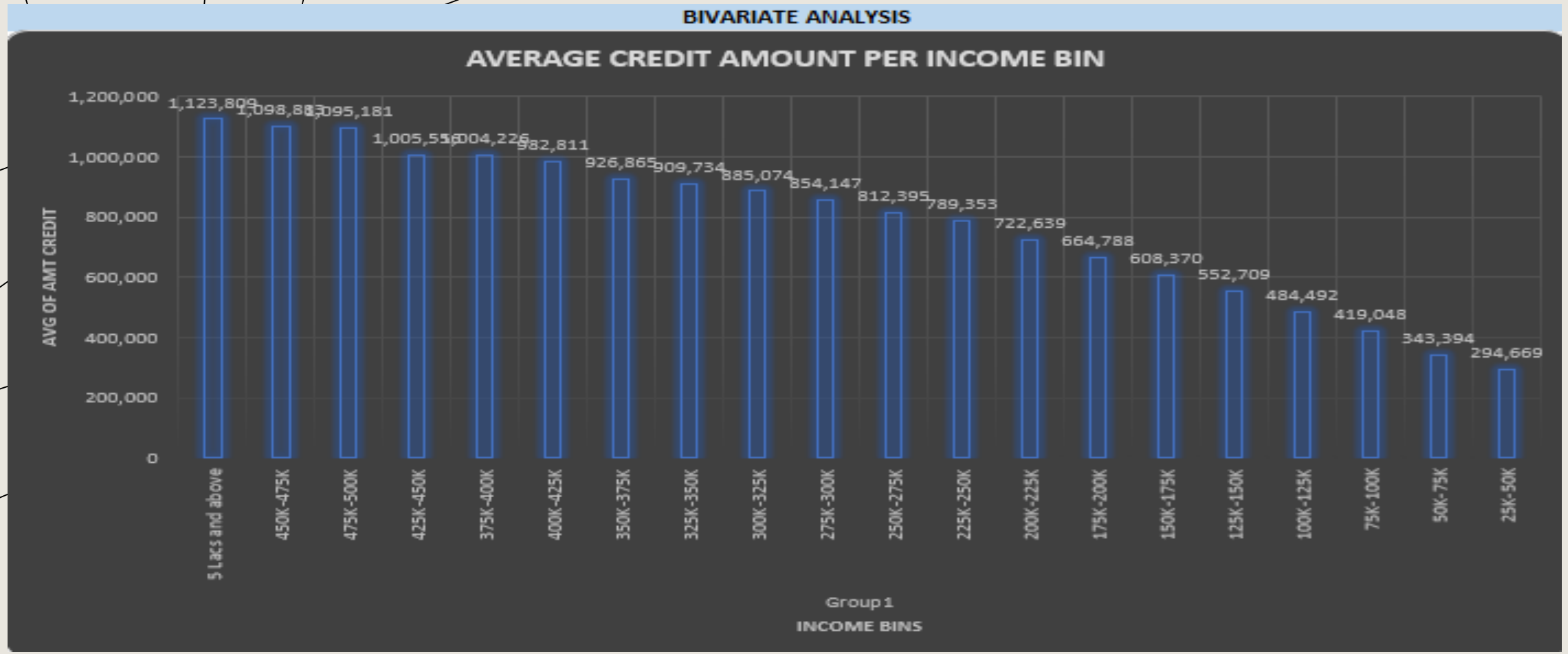
Univariate analysis refers to analysis of each varible, we will see how is it pattern and central tendencies.

The above plot is an univariate analysis of count of every applicants of columns AMT_CREDIT grouped in different income bins . So there is an observation that most of the applicants loan got approved with a range of 9 lakhs and above.

SEGMENTED UNIVARIATE ANALYSIS



BIVARIATE ANALYSIS



Bivariate analysis refers to observation of two variables how they are related to each other.

The above plot shows the relationship between applicants and different income bins which are directly proportional to each oyher. It clearly shows increase in income results in credit increase.

CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME

CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME								
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
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH(Years)	REGION_RATING_CLIENT

The Above correlation heat map shows correlation of different variables with applicants who made payment on time.

Green color intensity shows the amount of high correlation.

So. AMT_TOTAL_INCOME TO AMT_CREDIT, DAYS_BIRTH TO DAYS_EMPLOYED AND DAYS_EMPLOYED TI DAYS_ID_PUBLISH.

CORRELATION FOR APPLICANTS WITH PAYEEMENT DIFFICULTIES

CORRELATION FOR APPLICANTS WITH PAYMENT DIFFICULTIES								
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	0.023	-0.021	-0.059	-0.443	-0.034	0.003	-0.001	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH(Years)	REGION_RATING_CLIENT

The above heat map shows the correlation between the different variables for applicants with payment difficulties.

Same , green color intensity in heat map shows the level of correlation.

So, most relevant correlation among DAYS_BIRTH to DAYS_EMPLOYED and DAYS_ID_PUBLISH to DAYS_BIRTH.

RESULT

- Higher number of applicants got loan approved with a credit range of 9 lakhs and above.
- Applicants with higher income were given higher credit amount.
- Mean income is distributed between 1-2.5 lakhs

- Excel file hyperlink

THANK YOU

Davidraju Lakkamthoti

ge19lakkamthoti@mse.ac.in