

The background of the slide is a photograph of a classical bank building facade. It features grey stone walls, arched windows with decorative moldings, and large gold-colored letters spelling out 'BANK' on a dark horizontal band. The lighting is bright, suggesting daytime.

# **BANK LOAN CASE STUDY**

---

**By: VISHAL SONI**

---

---

# 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.

**Based on the scenarios a detailed analysis must be conducted and insights needs to be drawn to help 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®2016

**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.

---

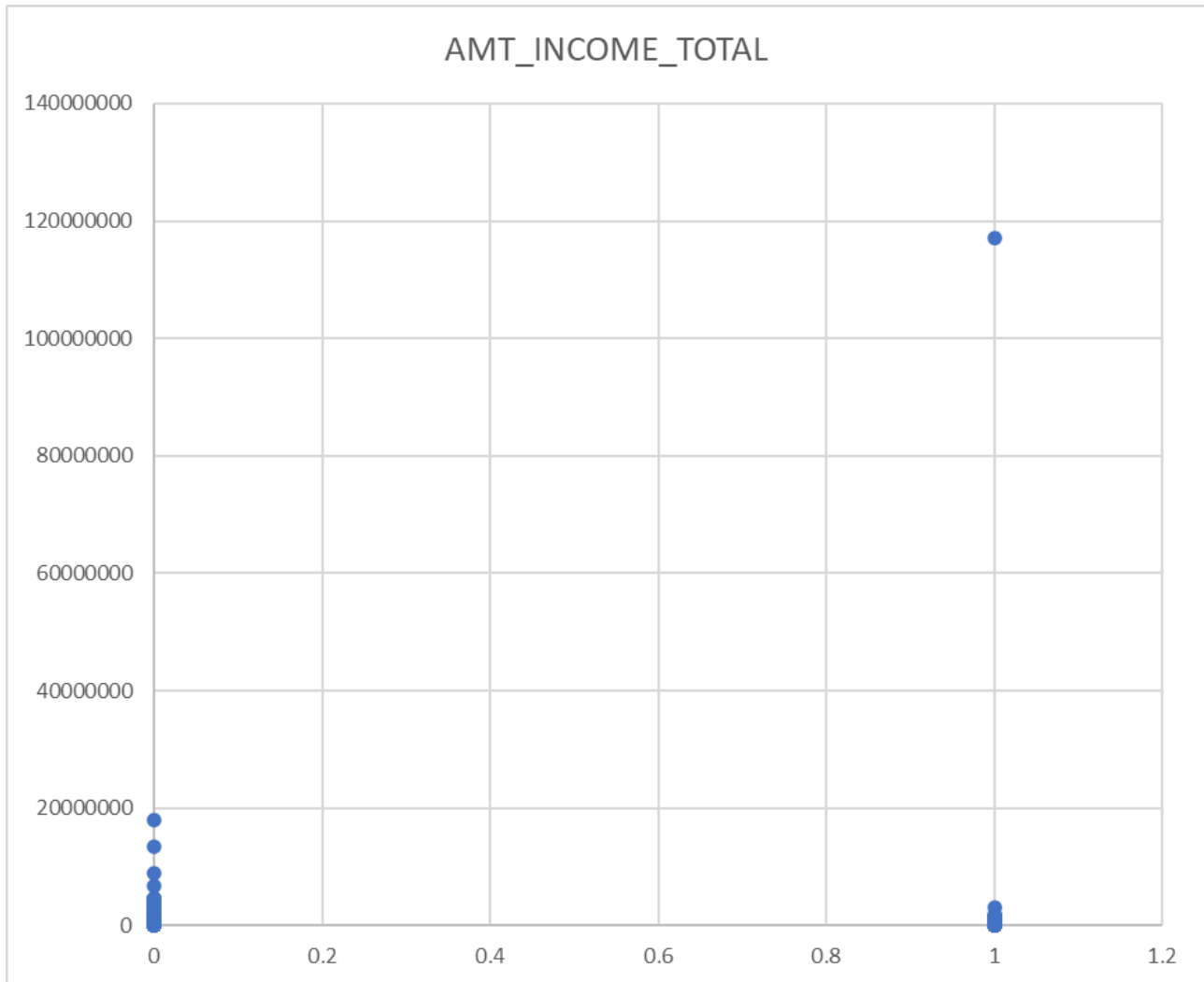
# APPROACH

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  $1 - (\text{Total Row Counts for each columns} / \text{Total Row Counts})$ . 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.

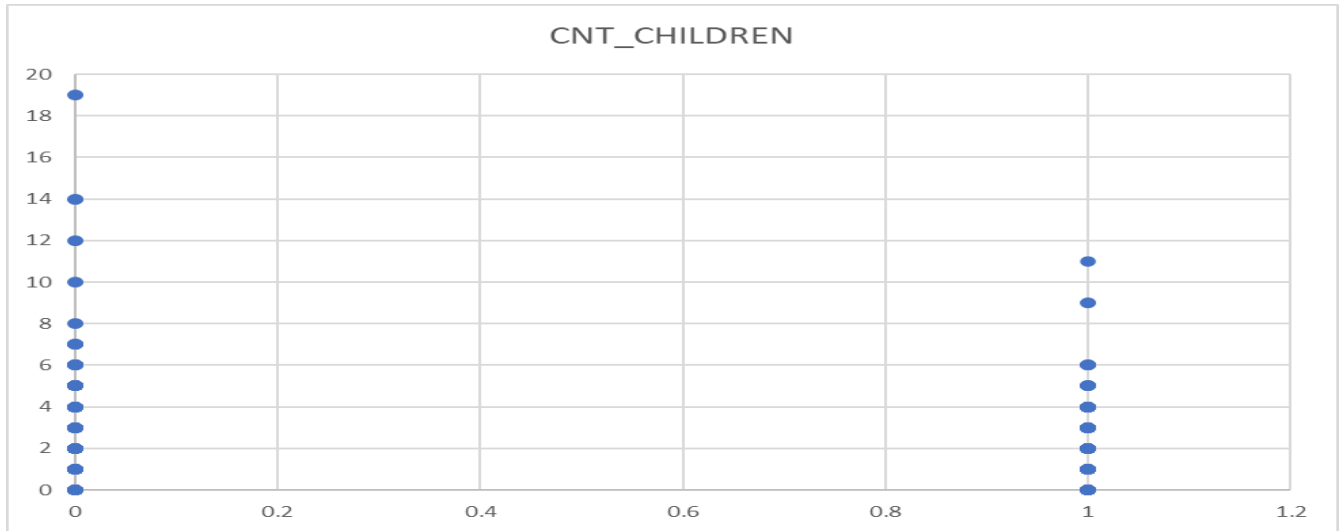
[https://docs.google.com/spreadsheets/d/1wAmpXp5r76j-fIqqpf6z6A1erqhQqjjN/edit?usp=share\\_link&oid=114888847224303662552&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1wAmpXp5r76j-fIqqpf6z6A1erqhQqjjN/edit?usp=share_link&oid=114888847224303662552&rtpof=true&sd=true)

---

# 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.



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.



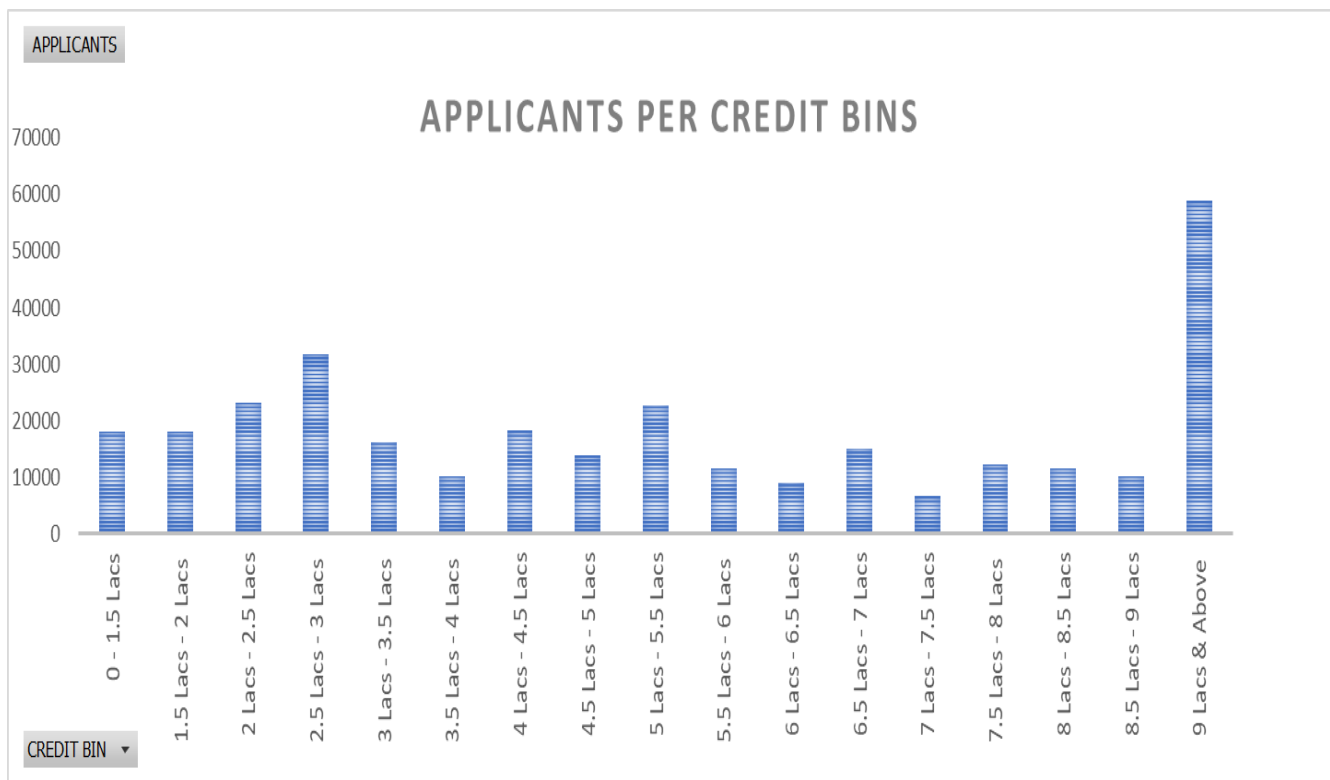
In the sheet outliers for Days Employed there are outliers for both target column 0 and 1. The XY plotter shows there are applicants being employed for 1000 years from the day of application which is clearly an anomaly.

---

# 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, 92% applicants paid installments on time thus makes the majority class and the rest of the 8% of applicants had payment difficulties thus makes the minority class.

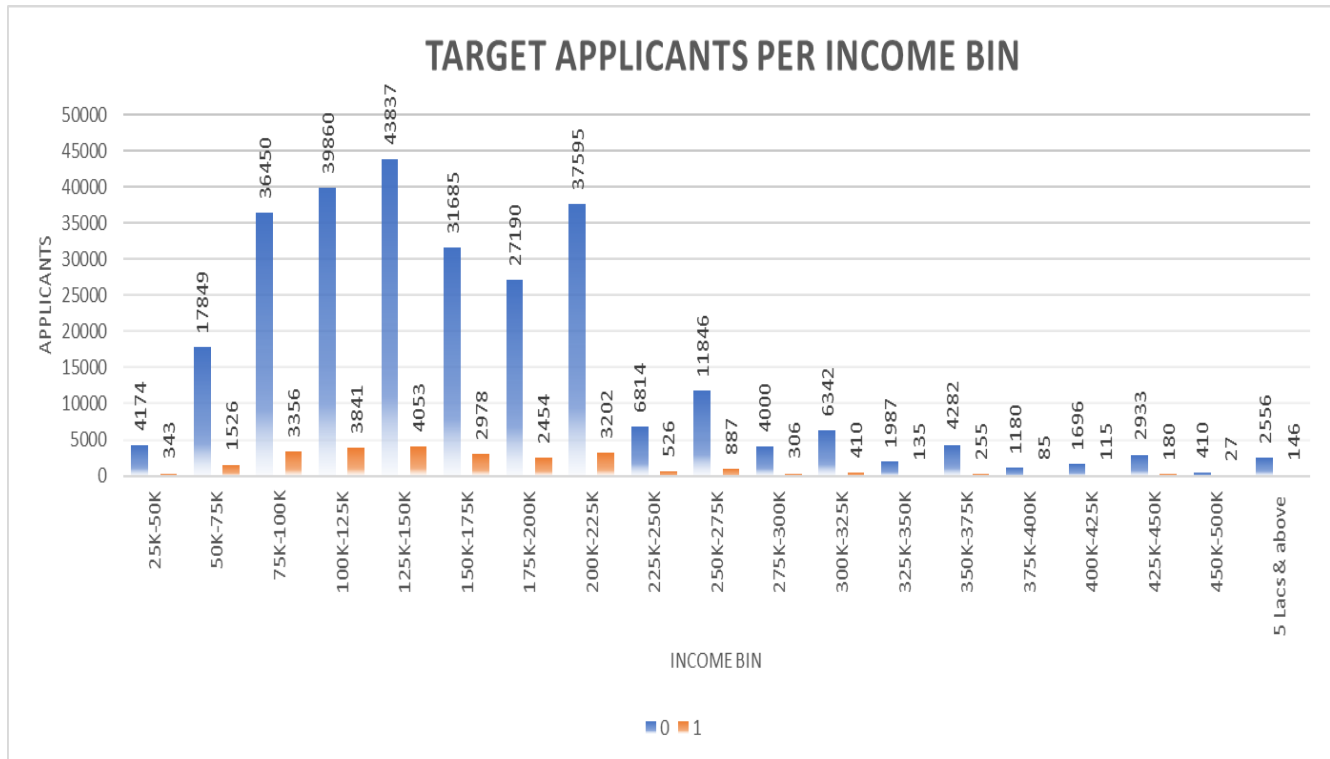
# UNIVARIANTE ANALYSIS



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.

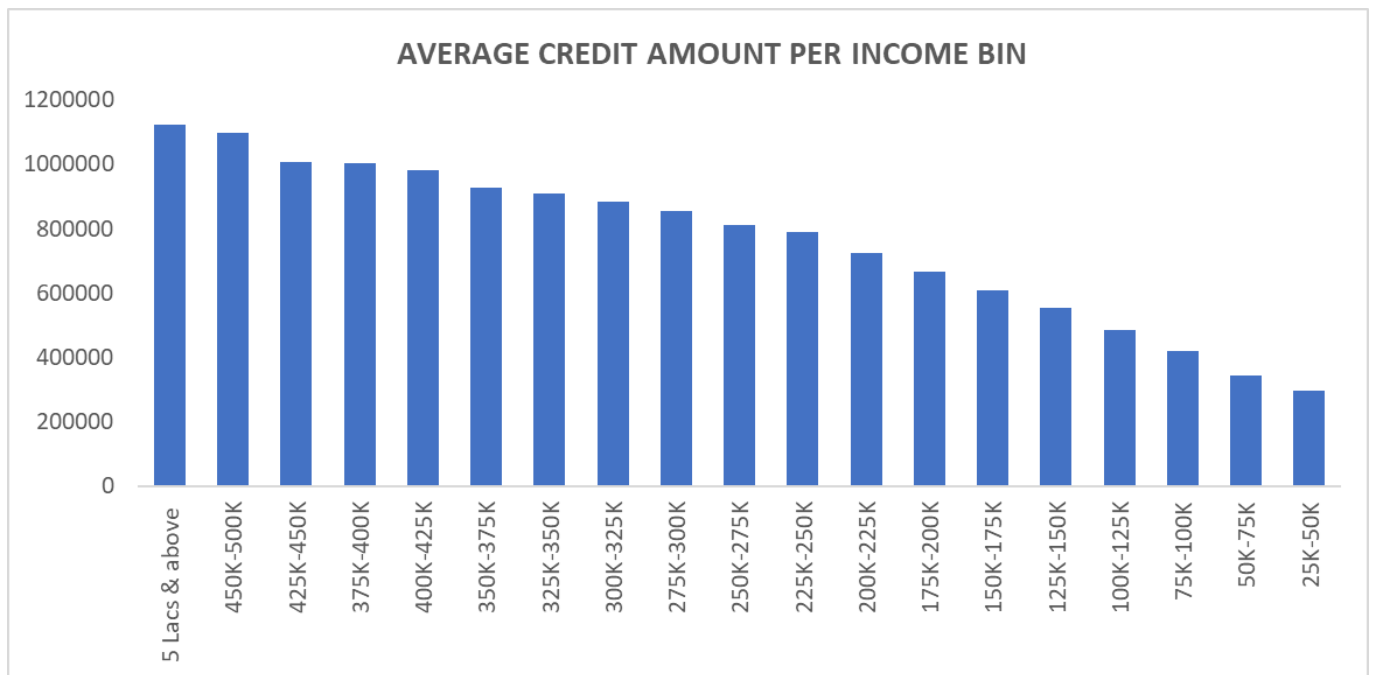


# UNIVARIATE SEGMENTED ANALYSIS



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	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 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	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
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH (Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH (Years)	REGION_RATING_CLIENT

**\*\* 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

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.