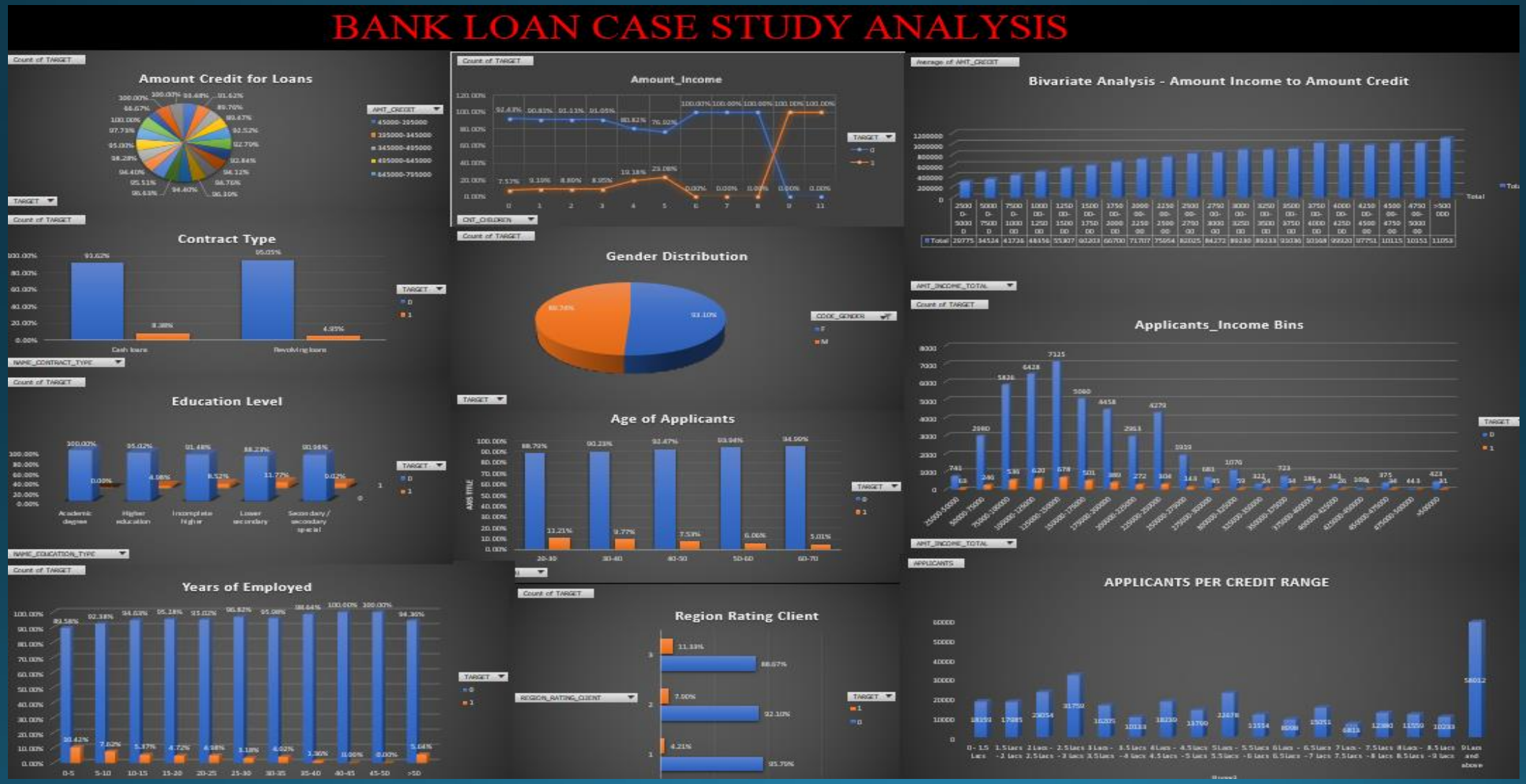
Bank Loan Case Study

By :- Mr. Raj Rathod



Business Overview

- The primary objective of this project is to discern patterns indicative of customers encountering challenges in paying their installments.
- This insight will inform strategic decisions, including loan denial, adjustment of loan amounts, or offering loans at higher interest rates to high-risk applicants.
- The company seeks to unravel the critical factors contributing to loan default to enhance decision-making regarding loan approvals.
- Employing Exploratory Data Analysis (EDA) to comprehend how customer characteristics and loan features influence default likelihood.
- EDA aims to enhance our understanding of the relationship between customer attributes and loan characteristics, ultimately aiding in risk assessment.

Project Description

1. Description:
   - I am a data analyst at a finance company specializing in lending to urban customers.
   - Challenge: Some customers with insufficient credit history default on loans.
   - Task: Use Exploratory Data Analysis (EDA) to analyze data patterns and prevent capable applicants from being rejected.

2. Risks Faced:
   - Rejection of capable applicants leads to lost business.
   - Approval of incapable applicants results in financial losses.

3. Dataset Overview:
   - Contains loan application information.
   - Scenarios:- Customers with payment difficulties: Late payment of more than X days on first Y installments.- Other cases: Payments made on time.

4. Possible Outcomes of Loan Application:
   - Approved: Loan application accepted.
   - Cancelled: Customer cancels application during approval process.
   - Refused: Loan application rejected by the company.
   - Unused Offer: Loan approved but not utilized by the customer.

Project Introduction

- Illustrates practical implementation of Exploratory Data Analysis (EDA) in a real-world business setting which is a Bank in My Case.
- Application of EDA techniques acquired during the module.
- Development of a fundamental understanding of risk analytics in banking and financial services.
- Demonstrates leveraging data to mitigate risks associated with consumer lending.
- Highlights challenges in loan granting at banks.
- Shows how data and risk analytics can minimize lending risks.
- Utilizes data relationships and visualizations to identify potential loan defaulters.

Tech-Stack Used

Microsoft Excel 2021

Microsoft PowerPoint

- Microsoft Excel 2021: Utilized for data cleaning, outlier detection, and conducting univariate and bivariate analysis using pivot tables and charts.
- Microsoft PowerPoint: Employed for presentation purposes, summarizing key findings and insights derived from the bank loan case study analysis project.

Approach

- This case study involves two extensive datasets: the ‘application_data’ and the ‘previous_application’.
- These datasets contained numerous unnecessary columns and blank entries, which were deemed irrelevant for risk assessments, prompting an initial cleaning process.
- To assess this vast amount of data effectively, the first step involved cleaning the datasets, identifying and removing outliers.
- Subsequently, univariate and bivariate analysis was conducted using pivot tables and charts to delve deeper into the data and derive meaningful insights.

Data Landscape

- application_data.csv: Contains comprehensive information about the client at the time of application, focusing on their payment difficulties.
- previous_application.csv: Provides insights into the client's prior loan data, including approval, cancellation, refusal, or unused offers.
- columns_description.csv: Serves as a data dictionary, detailing the meaning of variables present in the datasets.

A. To Identify Missing Data & Deal with it Appropriately

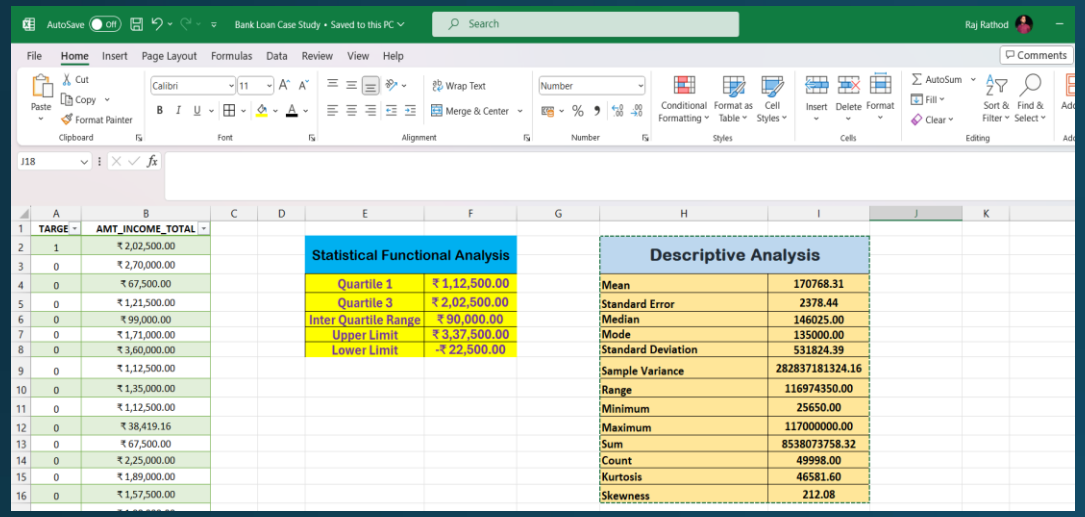
- Initially, I utilized the COUNTA function to determine the total number of rows in each column/variable. Then, I calculated the percentage of null values in each column using the formula: (Total rows count for each column / Total rows count) * 100.
- Subsequently, columns with a null value percentage exceeding 30% were eliminated, while those with less than 30% underwent distribution statistics analysis, including mean and mode, to address missing values.
- I retained only pertinent variables to extract meaningful insights and standardized the time units by converting days to years, dividing by 365.
- Furthermore, outliers were identified utilizing the interquartile range method, focusing on relevant variables.

DATA CLEANING - COLUMNS REMOVED

FLAG_WORK_PHONE FLAG_COUNT_MOBILE FLAG_PHONE FLAG_DOCUMENT_2-21 FLAG_DOCUMENT_3
FLAG_DOCUMENT_ 4 FLAG_DOCUMENT_ 5 FLAG_DOCUMENT_6 FLAG_DOCUMENT_7
FLAG_DOCUMENT_ 8 FLAG_DOCUMENT_9 FLAG_DOCUMENT_10 FLAG_DOCUMENT_11
FLAG_DOCUMENT_ 12 FLAG_DOCUMENT_13 FLAG_DOCUMENT_14 FLAG_DOCUMENT_15
FLAG_DOCUMENT_16 FLAG_DOCUMENT_17 FLAG_DOCUMENT_ 18 FLAG_DOCUMENT_ 19
FLAG_DOCUMENT_20 FLAG_DOCUMENT_ 21 EXT_SOURCE_2 EXT_SOURCE_3

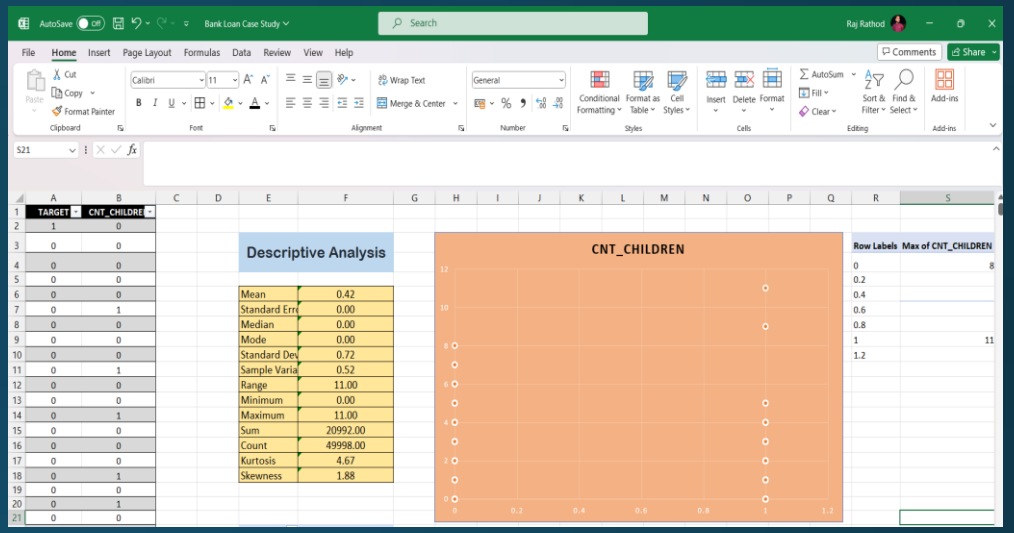
Column1	Column2	Column3	Column4
FLAG_WORK_PHONE	FLAG_DOCUMENT_6	FLAG_DOCUMENT_13	FLAG_DOCUMENT_20
FLAG_COUNT_MOBILE	FLAG_DOCUMENT_7	FLAG_DOCUMENT_14	FLAG_DOCUMENT_ 21
FLAG_PHONE	FLAG_DOCUMENT_ 8	FLAG_DOCUMENT_15	EXT_SOURCE_2
FLAG_DOCUMENT_2-21	FLAG_DOCUMENT_9	FLAG_DOCUMENT_16	EXT_SOURCE_3
FLAG_DOCUMENT_3	FLAG_DOCUMENT_10	FLAG_DOCUMENT_17	
FLAG_DOCUMENT_ 4	FLAG_DOCUMENT_11	FLAG_DOCUMENT_ 18	
FLAG_DOCUMENT_ 5	FLAG_DOCUMENT_ 12	FLAG_DOCUMENT_ 19	

B. Identify Outliers in the Dataset:

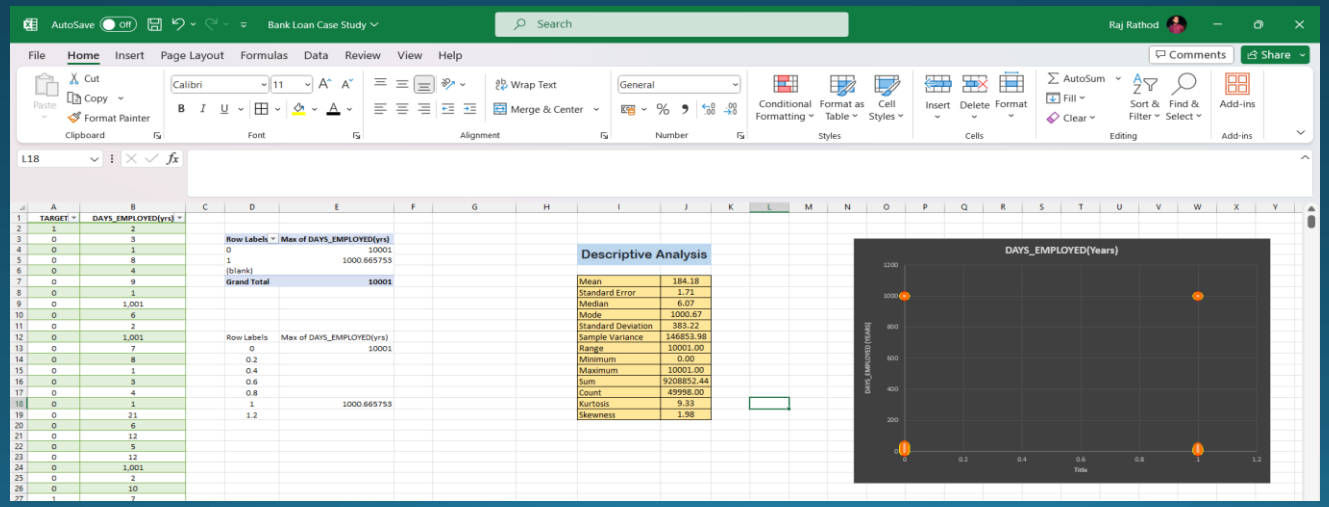


Statistical Functional Analysis	
Quartile 1	₹ 1,12,500.00
Quartile 3	₹ 2,02,500.00
Inter Quartile Range	₹ 90,000.00
Upper Limit	₹ 3,37,500.00
Lower Limit	-₹ 22,500.00

Descriptive Analysis	
Mean	170768.31
Standard Error	2378.44
Median	146025.00
Mode	135000.00
Standard Deviation	531824.39
Sample Variance	282837181324.16
Range	116974350.00
Minimum	25650.00
Maximum	117000000.00
Sum	8538073758.32
Count	49998.00
Kurtosis	46581.60
Skewness	212.08

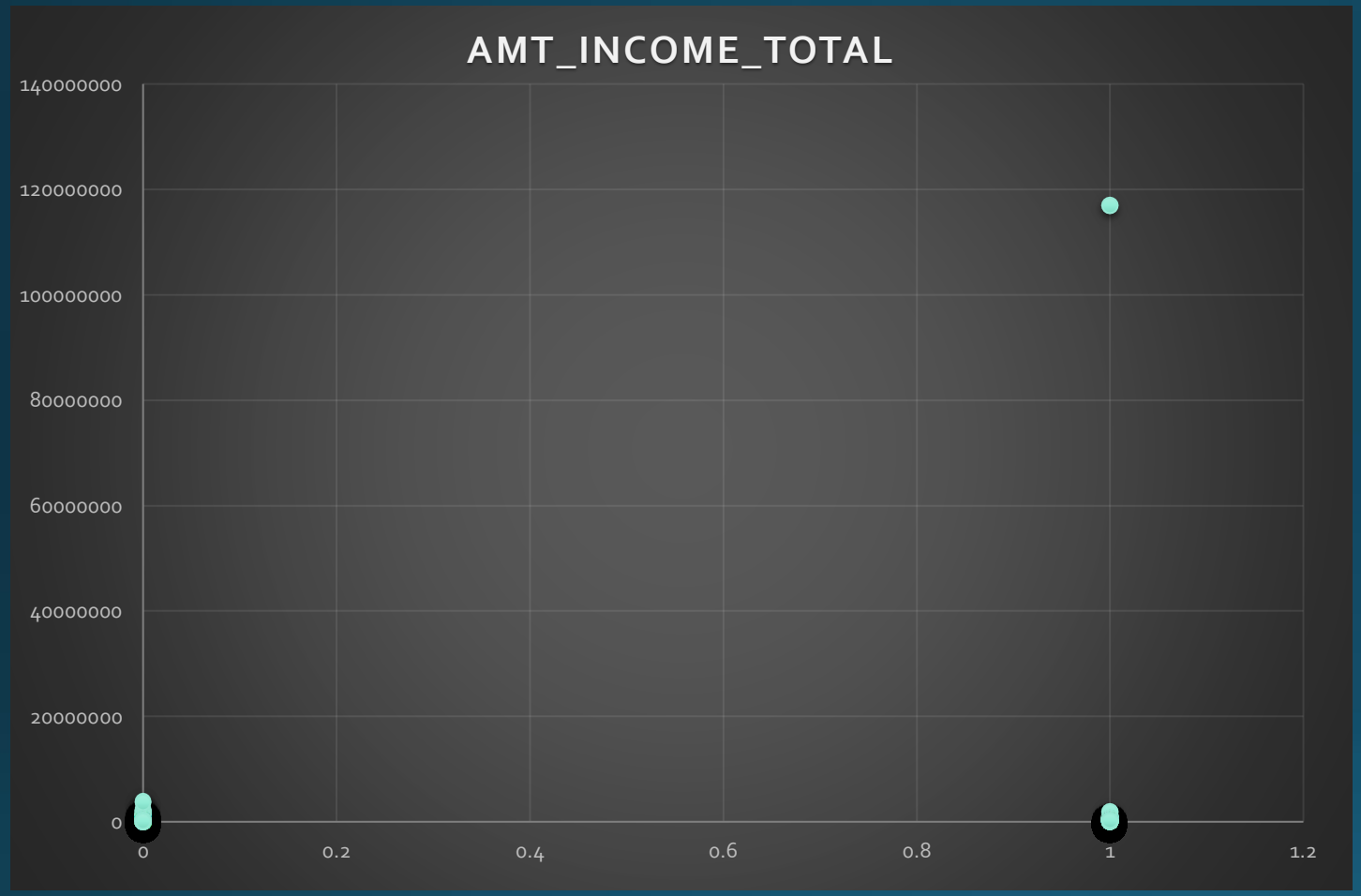


Descriptive Analysis	
Mean	0.42
Standard Err	0.00
Median	0.00
Mode	0.00
Standard Dev	0.72
Sample Varia	0.52
Range	11.00
Minimum	0.00
Maximum	11.00
Sum	20992.00
Count	49998.00
Kurtosis	4.67
Skewness	1.88



Descriptive Analysis	
Mean	184.18
Standard Error	1.71
Median	6.07
Mode	1000.67
Standard Deviation	383.22
Sample Variance	146853.98
Range	10001.00
Minimum	0.00
Maximum	10001.00
Sum	9208852.44
Count	49998.00
Kurtosis	9.33
Skewness	1.98

1. Outliers for AMT_INCOME_TOTAL -



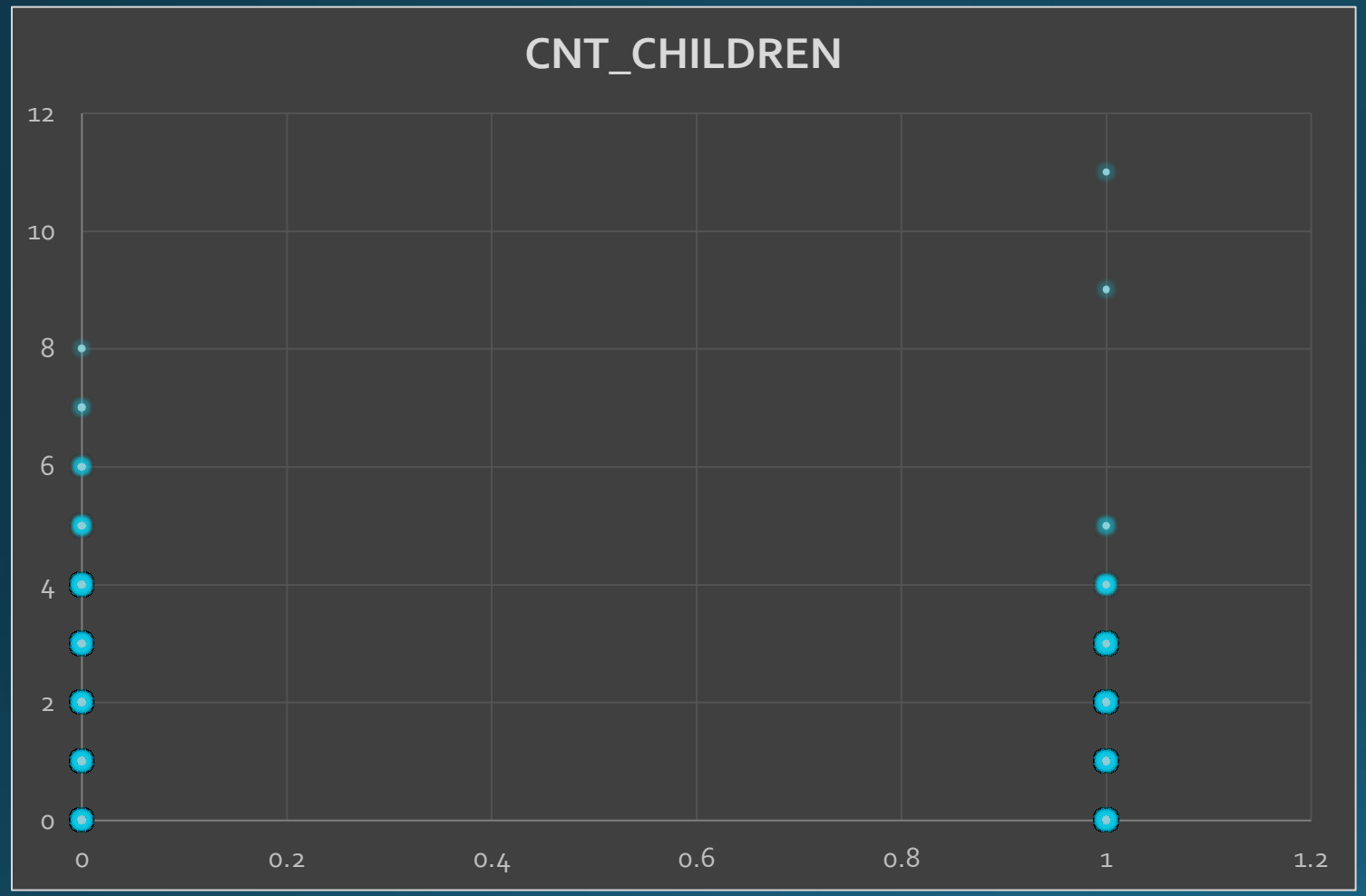
Descriptive Analysis	
Mean	170768.31
Standard Error	2378.44
Median	146025.00
Mode	135000.00
Standard Deviation	531824.39
Sample Variance	282837181324.16
Range	116974350.00
Minimum	25650.00
Maximum	117000000.00
Sum	8538073758.32
Count	49998.00
Kurtosis	46581.60
Skewness	212.08

Row Labels	AMT_INCOME_TOTAL
0	3825000
0.2	
0.4	
0.6	
0.8	
1	117000000
1.2	

Statistical Functional Analysis	
Quartile 1	₹ 1,12,500.00
Quartile 3	₹ 2,02,500.00
Inter Quartile Range	₹ 90,000.00
Upper Limit	₹ 3,37,500.00
Lower Limit	-₹ 22,500.00

- In this XY plotter, it's evident that for the target variable 1, there's a small number of applicants with an income of 11 crores, while the majority have incomes in the lakhs range.

2. Outliers for CNT_CHILDREN



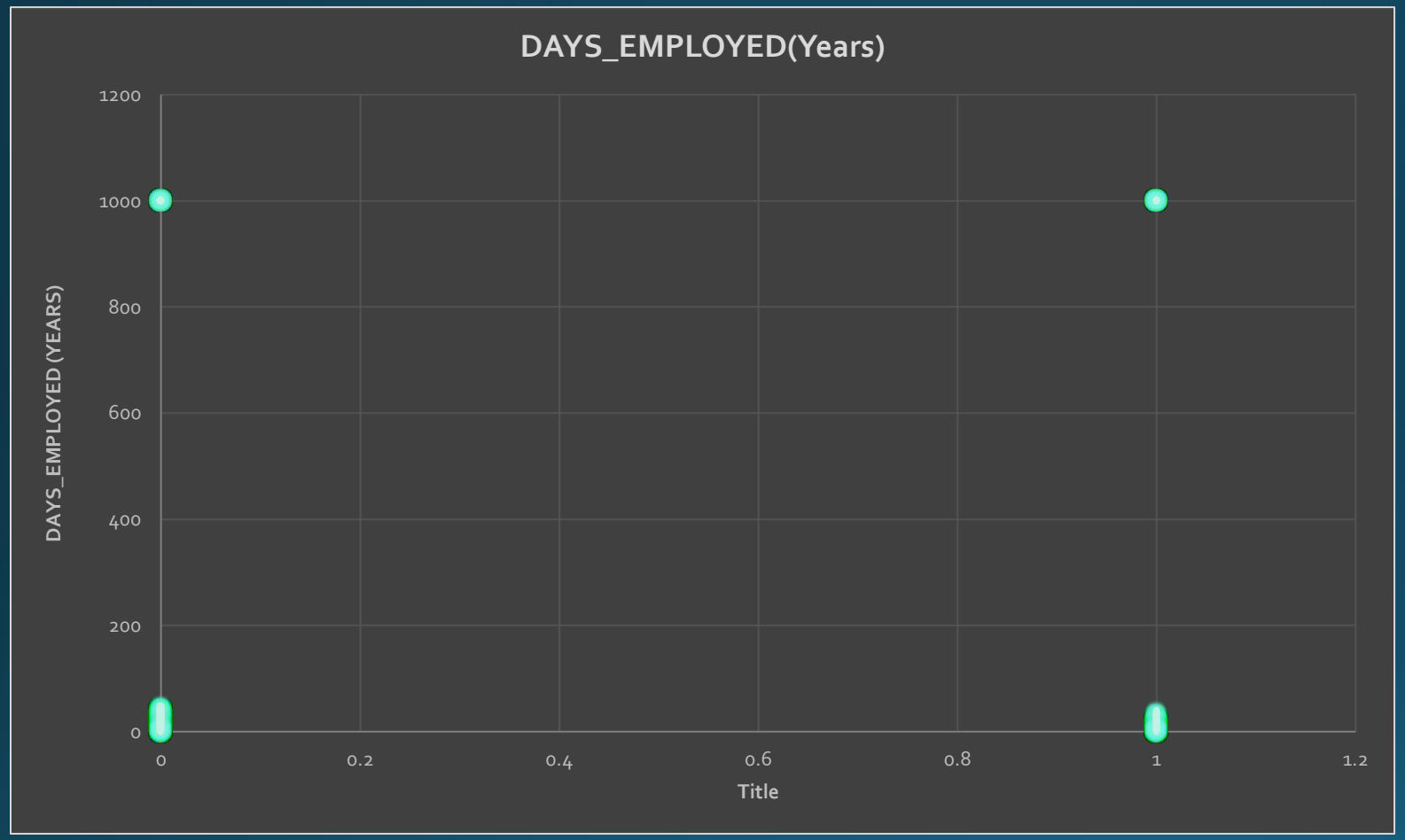
Descriptive Analysis

Mean	0.42
Standard Error	0.00
Median	0.00
Mode	0.00
Standard Deviation	0.72
Sample Variance	0.52
Range	11.00
Minimum	0.00
Maximum	11.00
Sum	20992.00
Count	49998.00
Kurtosis	4.67
Skewness	1.88

Row Labels	Max of CNT_CHILDREN
0	8
0.2	
0.4	
0.6	
0.8	
1	11
1.2	

In this XY plotter, it's evident that among the target variable O applicants, the maximum number of children observed is 8, a rarity in contemporary contexts. Conversely, for target 1 applicants, the maximum number of children observed is 11.

3. Outliers for DAYS_EMPLOYED(Years)



Descriptive Analysis	
Mean	184.18
Standard Error	1.71
Median	6.07
Mode	1000.67
Standard Deviation	383.22
Sample Variance	146853.98
Range	10001.00
Minimum	0.00
Maximum	10001.00
Sum	9208852.44
Count	49998.00
Kurtosis	9.33
Skewness	1.98

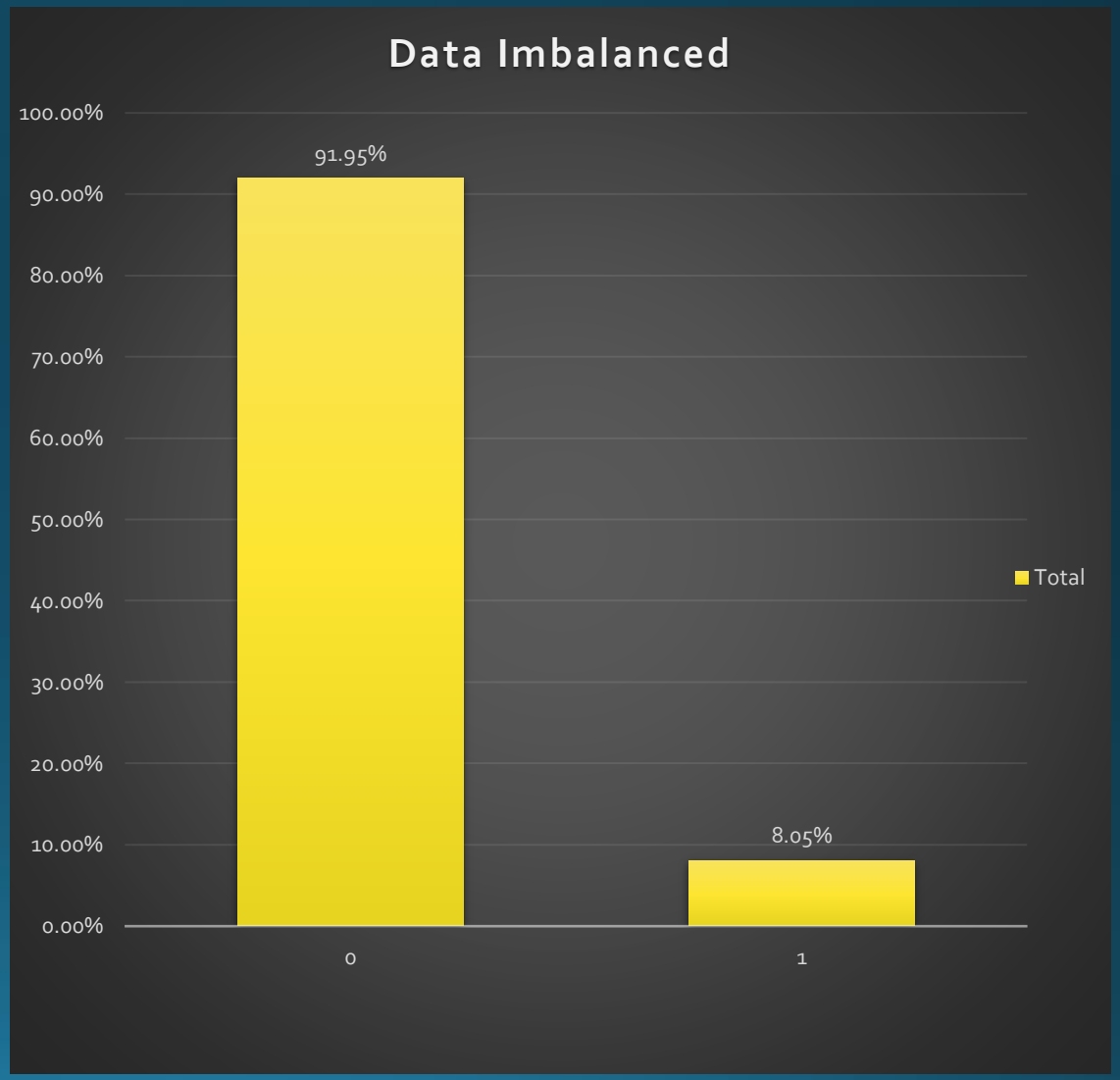
Row Labels	Max of DAYS_EMPLOYED(yrs)
0	10001
0.2	
0.4	
0.6	
0.8	
1	1000.665753
1.2	

In this XY plotter, it's evident that a small number of applicants in both target groups 0 and 1 are shown as being employed for 1000 years, which is clearly impossible. Conversely, the majority of applicants appear to have employment durations ranging from around 80 to 90 years.

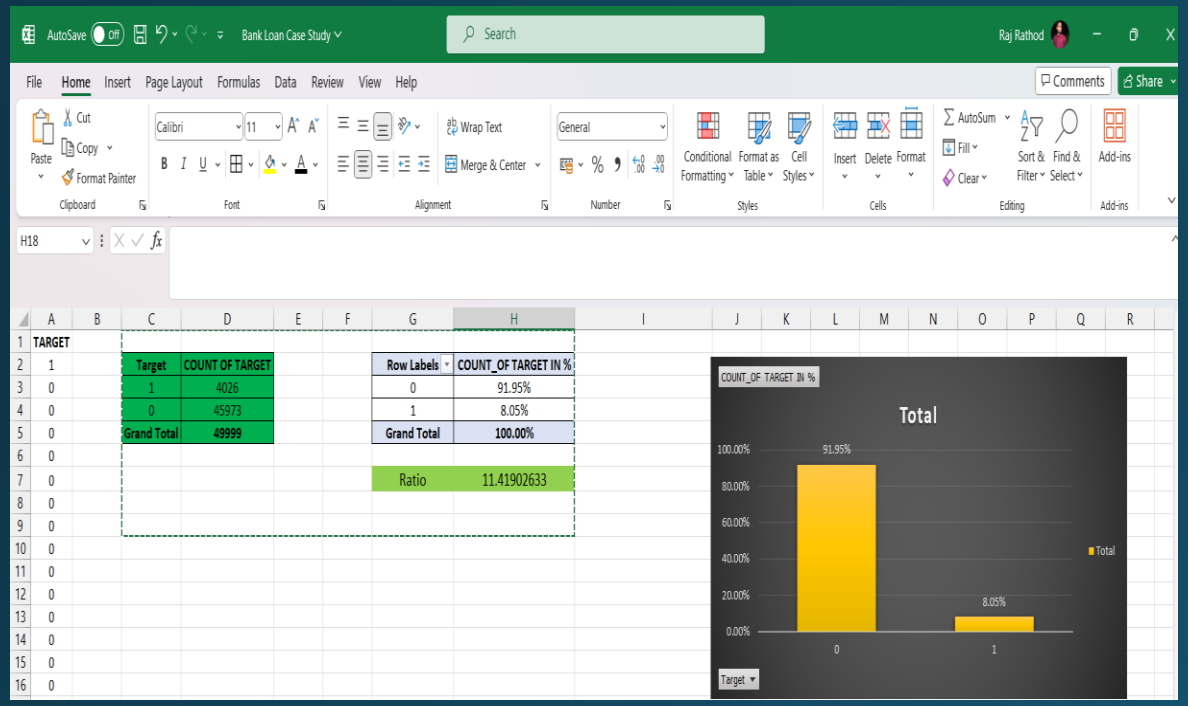
C. Analyse Data Imbalanced

- In the Excel file provided, the "Data Imbalanced" sheet illustrates the ratio between two categories: applicants facing payment difficulties (Target 1) and those making payments on time (Target 0), with a ratio of 11:41.

- Out of the total applicant pool of 49999, 91.95% successfully make payments on time, constituting the majority class. Conversely, the remaining 8.05% experience payment difficulties, forming the minority class.



Data Imbalanced

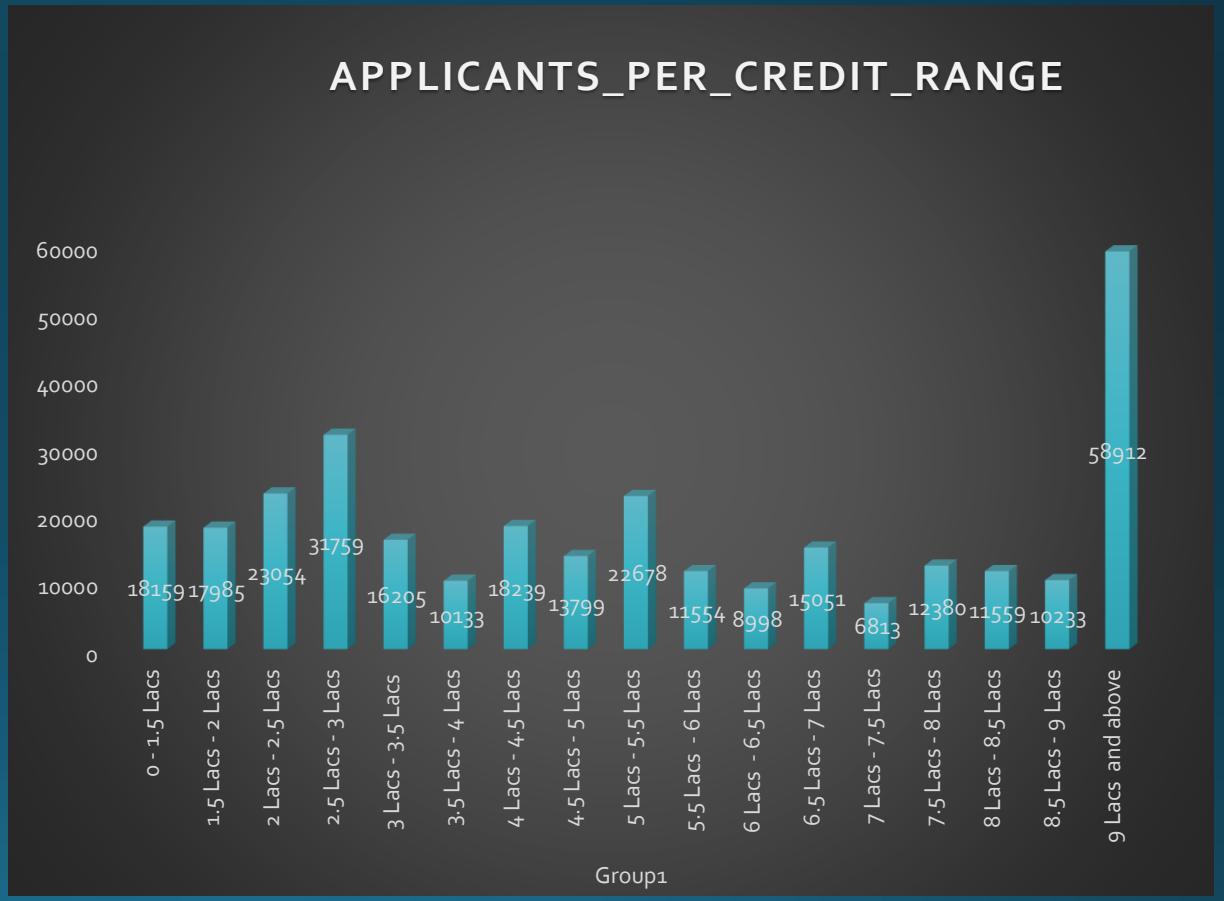


Target	COUNT OF TARGET			Row Labels	COUNT_OF TARGET IN %
1	4026			0	91.95%
0	45973			1	8.05%
Grand Total	49999			Grand Total	100.00%
				Ratio	11.41902633

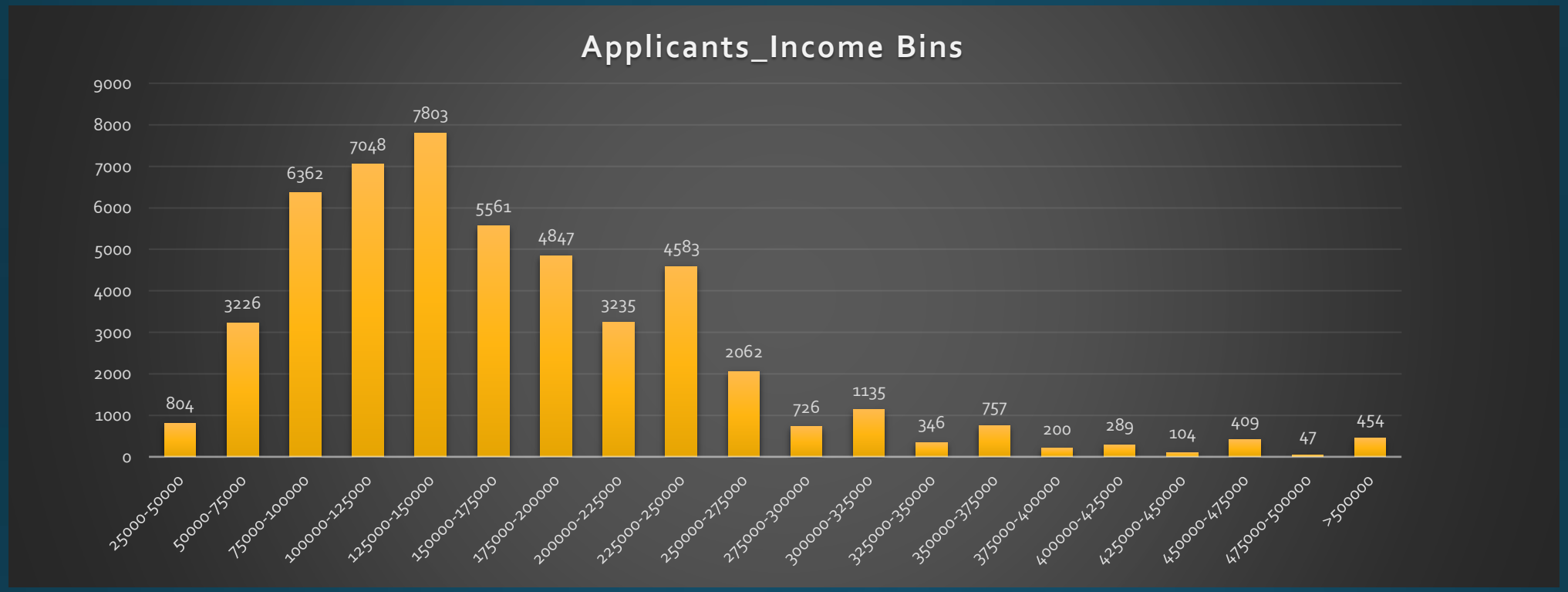
D. To Perform: Univariate, Segmented Univariate and Bivariate Analysis:

- Univariate Analysis involves examining data that consists of a single variable. It focuses on describing the data and identifying existing patterns, rather than exploring causes or relationships.

- The graph presented here illustrates univariate analysis by displaying the count of loan applicants (0 & 1) for various income brackets within the "AMT_CREDIT" column. The majority of applicants received loan approvals within the credit range of 9 lakhs and above.



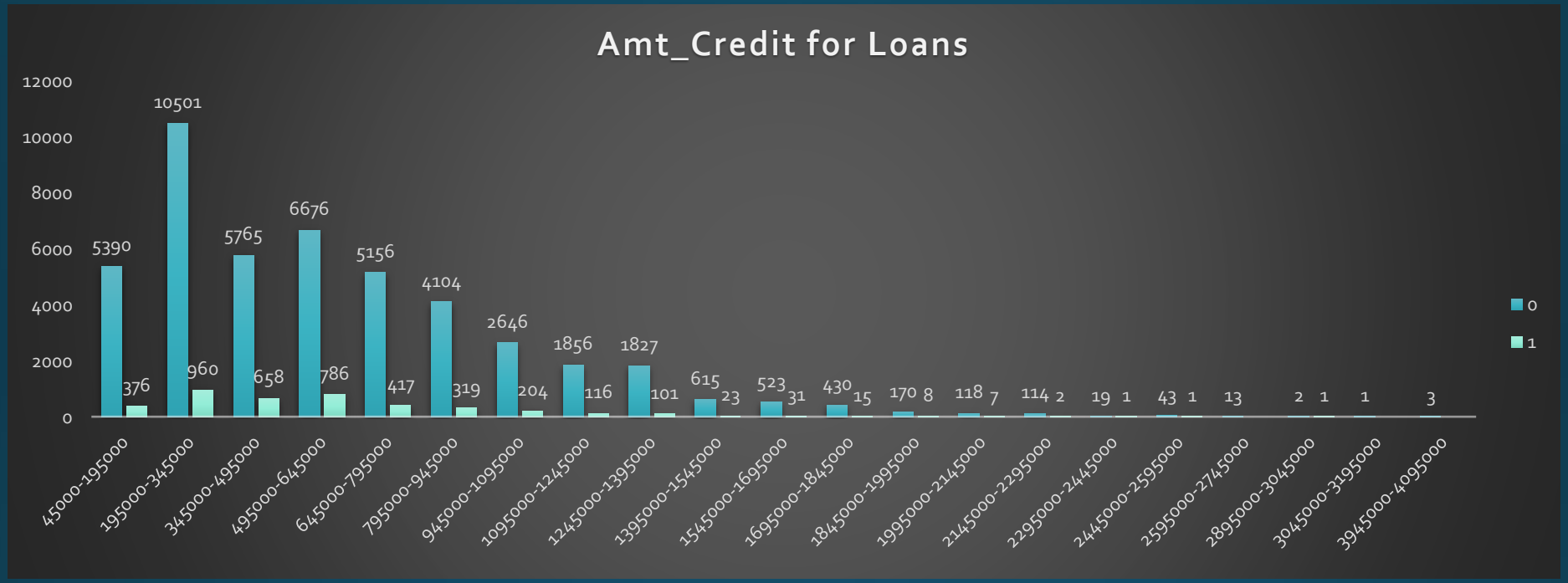
2.Segmented Variate Analysis



- Segmented Univariate analysis involves examining data containing only one variable. In this context, segmented analysis refers to dissecting a variable into subsets for analysis.

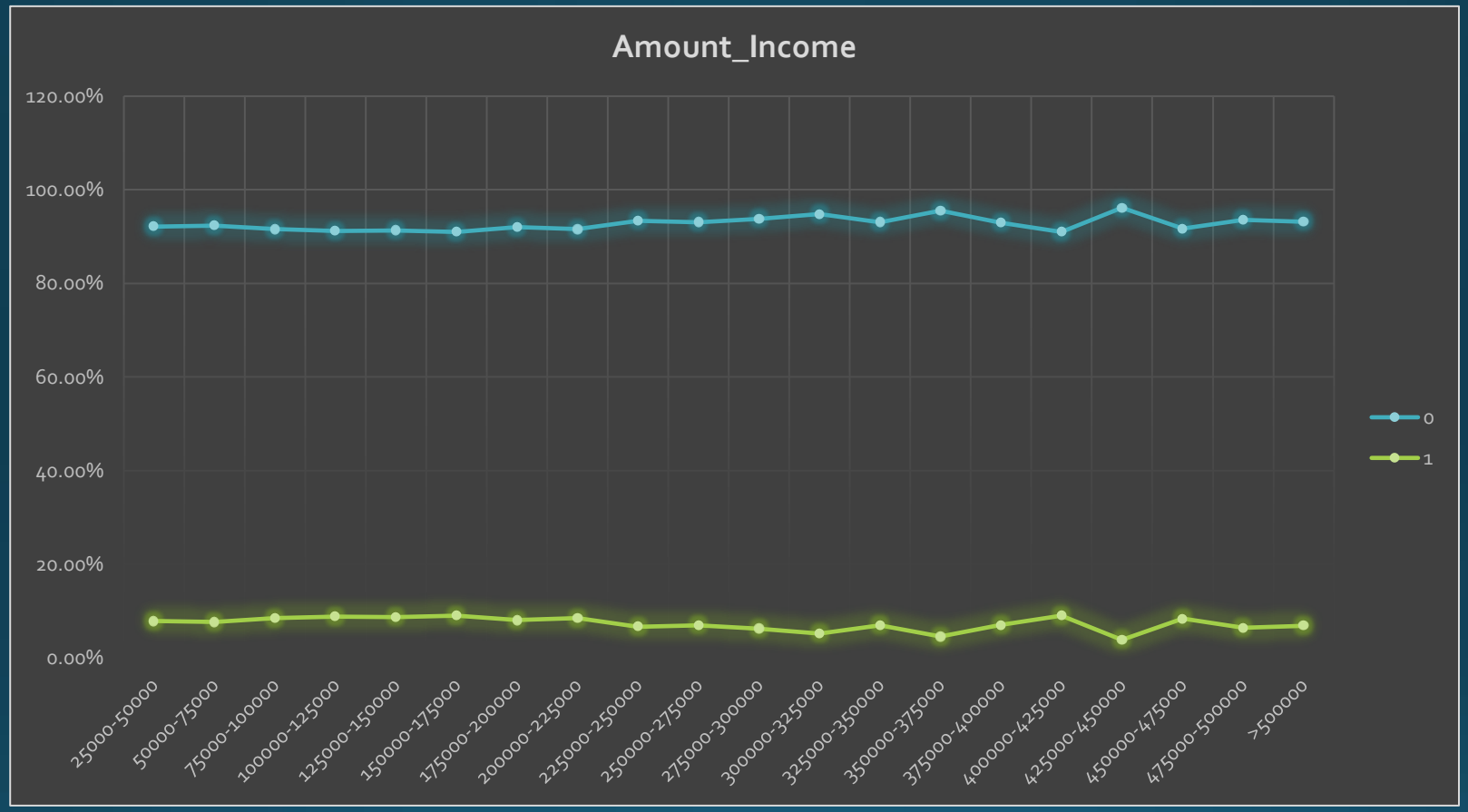
- The graph above illustrates segmented analysis, revealing that the majority of applicants (0 & 1) earn between 1 lakh and 2.25 lakhs. Notably, very few target 1 applicants earn 5 lakhs or more, which could contribute to payment difficulties.

3.Bivariate Analysis



- Bivariate analysis involves examining data that includes two distinct variables. It explores the connections and causations between these variables, aiming to uncover their relationship.
- The graph above illustrates the correlation between applicants and various income brackets, demonstrating a direct proportionality between the two.
- Consequently, as income rises, the amount of credit also increases.

3.Bivariate Analysis



- The line above illustrates the correlation between applicants and various income brackets, demonstrating a direct proportionality between the two.
- Consequently, as income rises, the amount of credit also increases.

E .To Identify Top Correlations for Different Scenarios:

- Understanding correlations between variables and the target variable can unveil significant indicators of loan default.
- Task: Utilize Excel functions to segment the dataset based on various scenarios, such as clients experiencing payment difficulties and all other cases.
- Identify the top correlations within each segmented dataset to discern key relationships contributing to loan default.

Correlation for Applicants with Payment Made On Time

Correlation For Applicants With Payments Made On Time										
CNT_CHILDREN	1.000	0.036315621	0.006	0.026	0.002	-0.025	-0.336	-0.246	0.879	0.021
AMT_INCOME_TOTAL	0.036	1.000	0.378	0.451	0.385	0.182	-0.074	-0.162	0.042	-0.205
AMT_CREDIT	0.006	0.377963032	1.000	0.771	0.987	0.096	0.051	-0.075	0.065	-0.103
AMT_ANNUITY	0.026	0.451137151	0.771	1.000	0.776	0.117	-0.010	-0.111	0.078	-0.130
AMT_GOODS_PRICE	0.002	0.384573329	0.987	0.776	1.000	0.099	0.049	-0.072	0.063	-0.105
REGION_POPULATION_RELATIVE	-0.025	0.181936304	0.096	0.117	0.099	1.000	Chart Area	-0.007	-0.023	-0.539
DAYS_BIRTH(yrs)	-0.336	-0.073764968	0.051	-0.010	0.049	0.030	1.000	0.623	-0.284	-0.009
DAYS_EMPLOYED(yrs)	-0.246	-0.161685009	-0.075	-0.111	-0.072	-0.007	0.623	1.000	-0.235	0.041
CNT_FAM_MEMBERS	0.879	0.041598095	0.065	0.078	0.063	-0.023	-0.284	-0.235	1.000	0.022
REGION_RATING_CLIENT	0.021	-0.205032782	-0.103	-0.130	-0.105	-0.539	-0.009	0.041	0.022	1.000
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	_POPULATION_R	DAYS_BIRTH(yrs)	DAYS_EMPLOYED(yrs)	CNT_FAM_MEMBERS	REGION_RATING_CLIENT

- The heatmap above illustrates the correlation among various variables for target 0, representing applicants who made payments on time.
- In the heatmap, red indicates the strongest correlation, while green indicates the weakest correlation between variables.
- Key correlations observed include:- AMT_TOTAL_INCOME to AMT_CREDIT- DAYS_BIRTH to DAYS_EMPLOYED- DAYS_EMPLOYED to DAYS_ID_PUBLISH

Correlation for Applicants with Payment Difficulties

	Correlation For Applicants With Payment Difficulties									
CNT_CHILDREN	1	0.010110177	0.007601905	0.029172977	-0.001079665	-0.020359154	-0.2496732	-0.189773227	0.892521875	0.055515557
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	0.018004594	0.013269502	-0.006180303	-0.009033662	-0.011758681	0.013121678	-0.012846697
AMT_CREDIT	0.007601905	0.015271444	1	0.749665201	0.982267963	0.067775624	0.142506035	0.018782223	0.06124869	-0.045024534
AMT_ANNUITY	0.029172977	0.018004594	0.749665201	1	0.74950403	0.073123998	0.008751713	-0.078113894	0.075838463	-0.061578289
AMT_GOODS_PRICE	-0.001079665	0.013269502	0.982267963	0.74950403	1	0.076635488	0.141005898	0.023181572	0.055135807	-0.051296281
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.067775624	0.073123998	0.076635488	1	0.016468731	0.007710059	-0.017257146	-0.430032303
DAYS_BIRTH(yrs)	-0.2496732	-0.009033662	0.142506035	0.008751713	0.141005898	0.016468731	1	0.588242824	-0.199141397	-0.045027112
DAYS_EMPLOYED(yrs)	-0.189773227	-0.011758681	0.018782223	-0.078113894	0.023181572	0.007710059	0.588242824	1	-0.183362962	-0.009237108
CNT_FAM_MEMBERS	0.892521875	0.013121678	0.06124869	0.075838463	0.055135807	-0.017257146	-0.199141397	-0.183362962	1	0.057279521
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.045024534	-0.061578289	-0.051296281	-0.430032303	-0.045027112	-0.009237108	0.057279521	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH(yrs)	DAYS_EMPLOYED(yrs)	CNT_FAM_MEMBERS	REGION_RATING_CLIE

Chart Area

The heatmap above illustrates the correlation among various variables for target o, representing applicants facing payment difficulties.

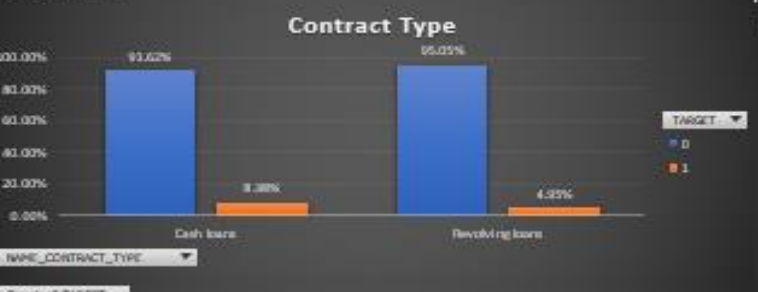
The heatmap colour scheme ranges from red, indicating the strongest correlation, to green, indicating the weakest correlation between variables.

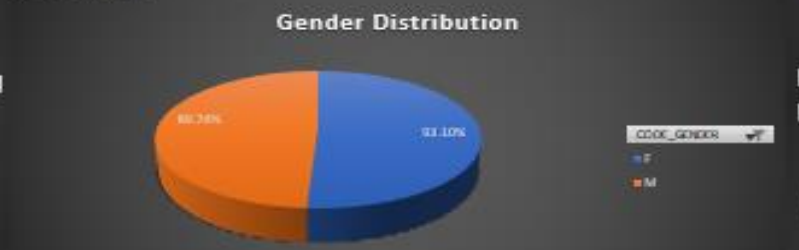
Consequently, the most significant correlations are observed between DAYS_BIRTH and DAYS_EMPLOYED, as well as between DAYS_ID_PUBLISH and DAYS_BIRTH.

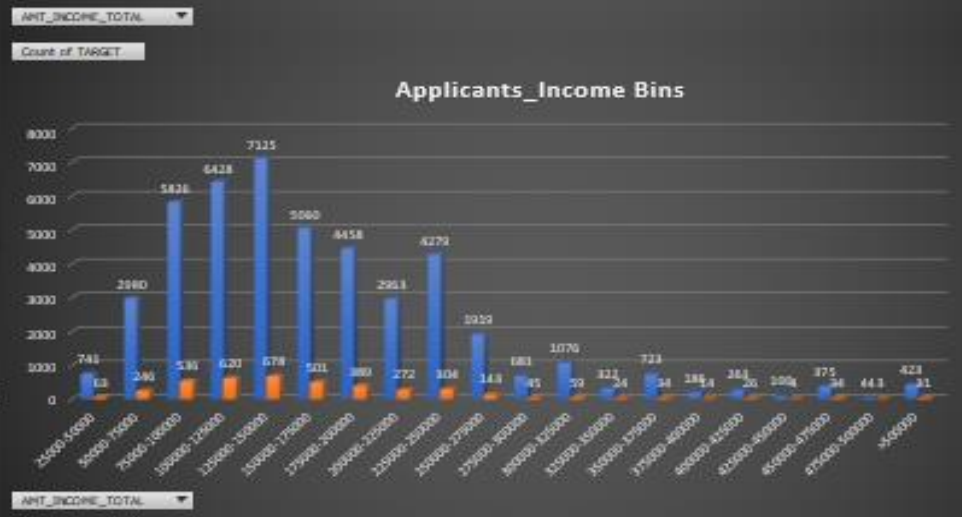
Insights

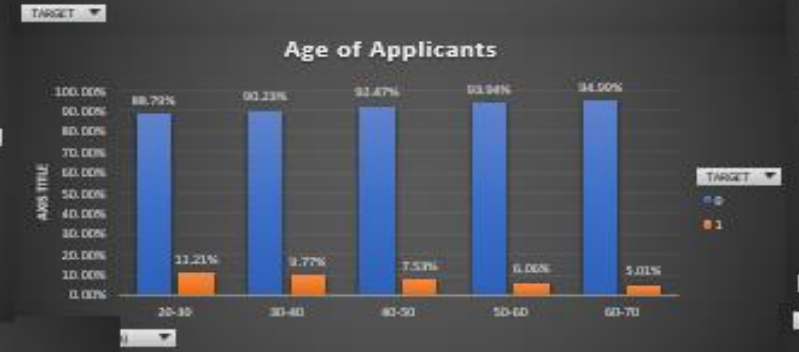
- Applicants drawing the higher income were offered the higher credit amount.
- Majority of the applicants got loan approval of credit range 9 lacs and above.
- Most of the applicants(0&1) drawing an income between 1lacs - 2.25 lacs.
- Through our analysis, we addressed several factors influencing the likelihood of client default:
- Age and experience correlate inversely with default probability. Priority may be given to older, more experienced clients by the bank.
- Educated clients, particularly those with higher levels of education, demonstrate lower default rates compared to those with lower educational attainment.
- Male clients exhibit a higher likelihood of default compared to female clients.
- Region Rating 3 comprises the highest percentage of defaulters, suggesting the need for stricter loan conditions. Conversely, clients from Region 1 present the lowest default risk.
- An observable trend indicates that as clients age, loan amounts tend to increase. Despite this, older clients with higher ages may pose lower default risks and potentially yield higher profitability.
- Furthermore, clients with more than two children exhibit higher default frequencies compared to others.

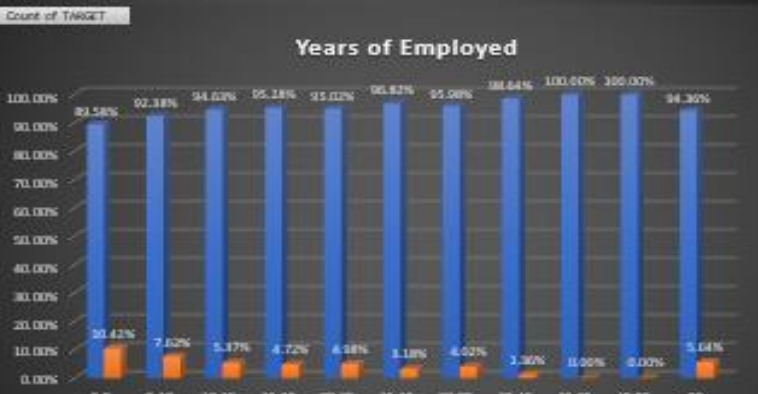
BANK LOAN CASE STUDY ANALYSIS

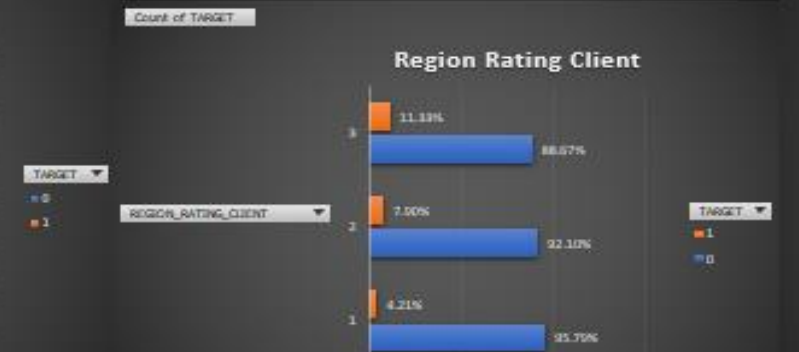


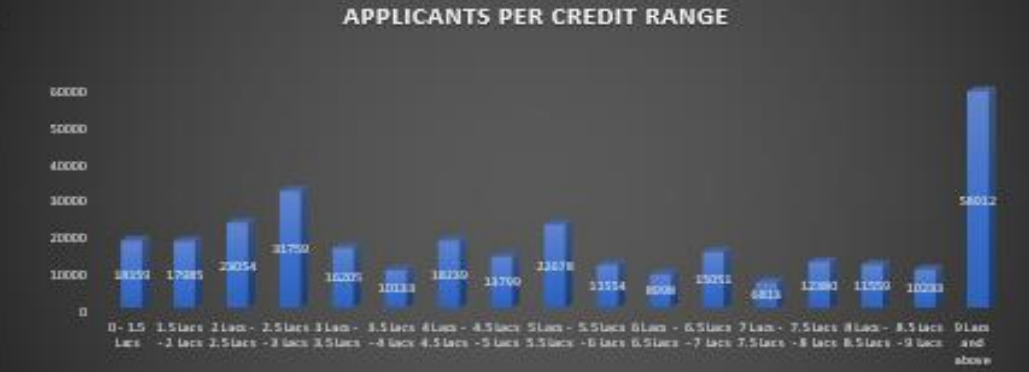












Result

- This case study exemplifies the practical application of Exploratory Data Analysis (EDA) within a real business context, specifically Bank Loan Analysis.
- Through this project, I acquired a fundamental understanding of risk analytics in banking and financial services, grasping how data is instrumental in mitigating lending risks.
- The project presented formidable challenges, particularly in analyzing correlations among variables to derive meaningful insights for clients.
- I gained insights into addressing data imbalances, identifying outliers, and understanding the key drivers within datasets.
- Visualization played a pivotal role in synthesizing complex datasets, facilitating the communication of crucial findings beneficial to the client.

THANK YOU!

Bank Loan Case Study Analysis By EDA Method by Mr. Raj Rathod

LinkedIn – www.linkedin.com/in/rajrathod54321

Excel Sheet Link

Video Presentation Link