Bank loan Analysis

Project: This project is based on Bank loan, there are client who are interested in taking loan from the bank. Some may repay the amount but some may not. Before giving loan to customer bank need to assure that the customer may repay the debt or not

Objective: As a Data Analyst first we need to clean the Data, handling missing values, imputation of values. We need to find the primary key to use as foreign key to link two data sets to perform joins so that data can be extracted. Our goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default. The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

Approach: We have the dataset that contains the data of application which contains huge amount of data, so first we need to clean the data, we need to check which columns are not used so we can remove that, also we need to check the missing values, outliers to impute or remove from the dataset.

Tech Stack Used: I have used MS excel, as it is most efficient tool to clean analyse huge amount of data.

Tasks:

A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

Ans: In the dataset provided we can clearly see that we have hugged amount of data to process. We have numerous numbers of columns around 172, so after considering the data and the tasks we have removed few columns that were no longer needed.
So, we have checked for number of null values in each column and if any column has more that 30% of num values then we have dropped the column. And if it has lesser value then we have replaced it with relevant values.
We have used count if, is null function to calculate that.

B. **Identify Outliers in the Dataset:** Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

Ans: We have calculated the outliers for the application data set we can see that few datapoints are way higher than the range of the data.

Let's take example of **Amount_Income columns:**

AMT_INCOME_TOTAL	
MEAN	170767.5905
MEDIAN	145800
MODE	0
STANDARD DEVIATION	531819.0951
VARIANCE	282831549942
MAX	117000000
MIN	25650
SUM	8538208758
COUNT	49999

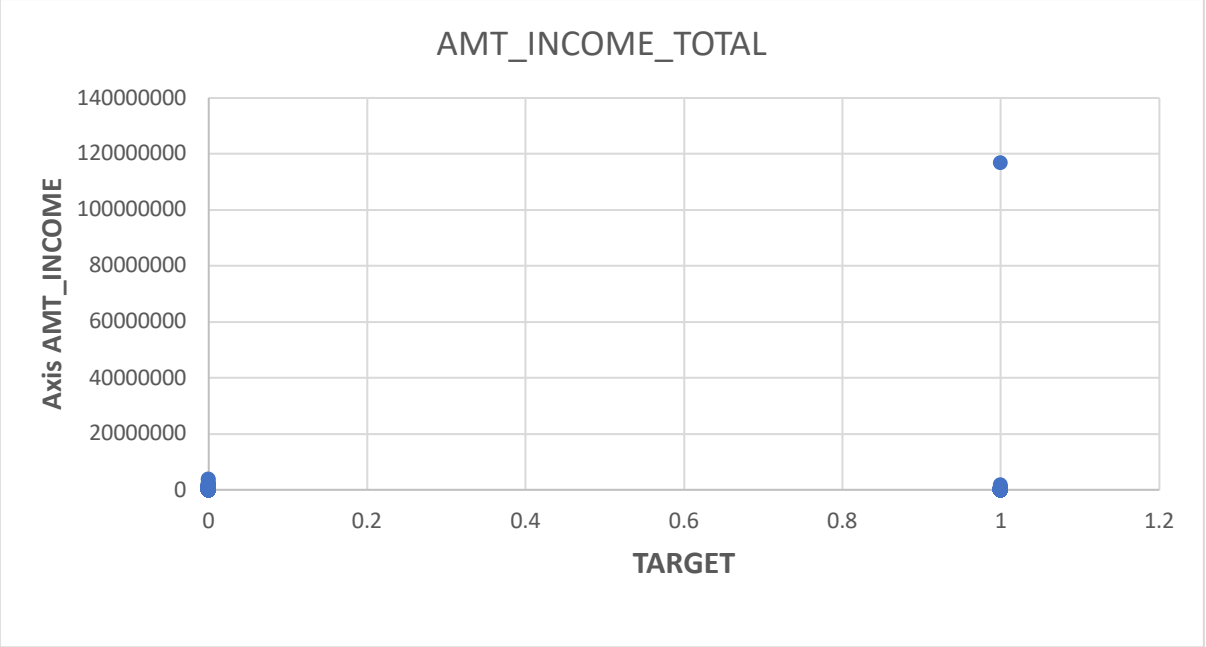
QUARTILE 1
112500

QUARTILE 3
202500

INNER QUARTILE RANGE
90000

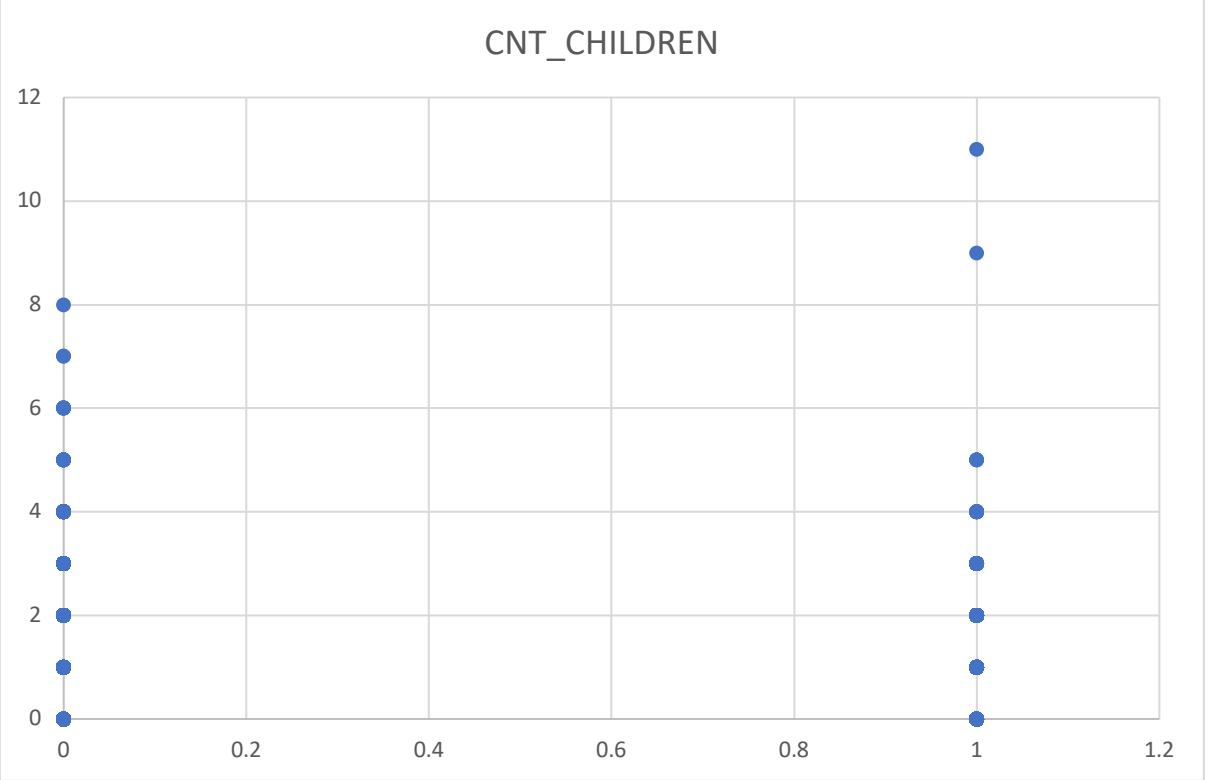
UPPER LIMIT
135000

LOWE LIMIT
-22500



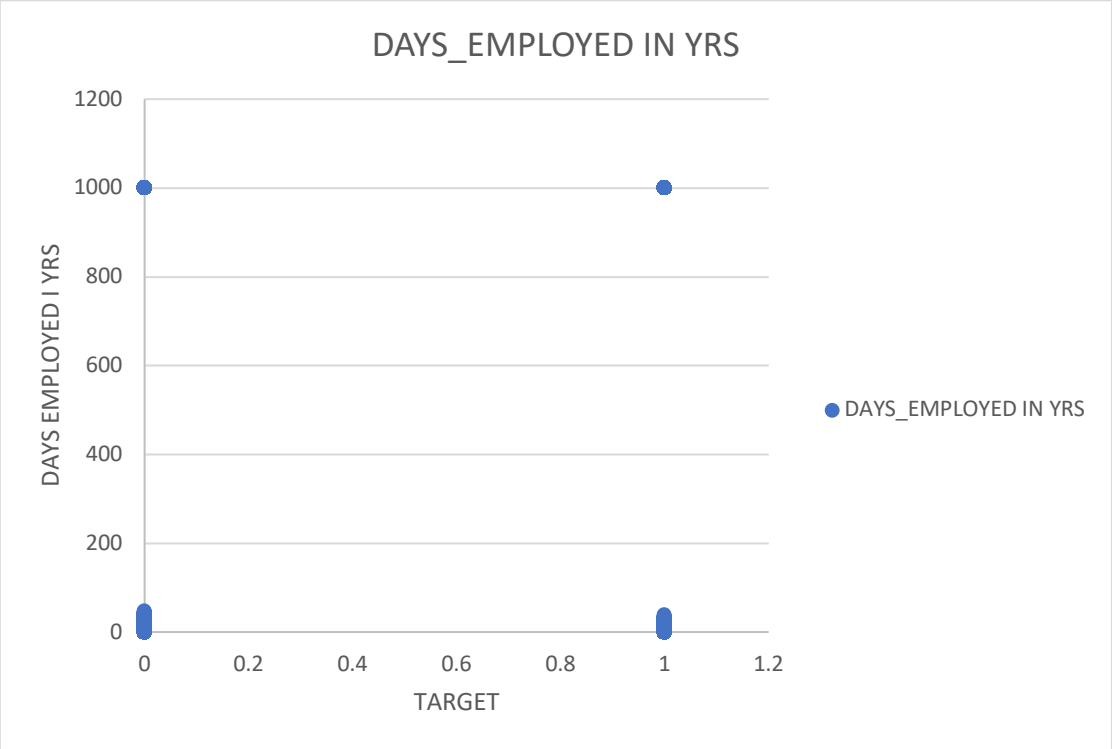
Here one data point which is near to 120000000 is outliear for this columns, it may produce ambiguity to the dataset and may also produce inappropriate result when building machine learning models.

Let's take another example of **count of children column**:



We can see that there are family which has near to 12 children that is biologically possible but it's very rare to have in these days.

Let's take example of **DAYS_EMPLOYEED_IN_YEARS:**



Here from the above table, we can see that there are applicants who have been working in the last 100 years which is not possible in a normal human life cycle, hence their outliers can be amputate or removed.

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

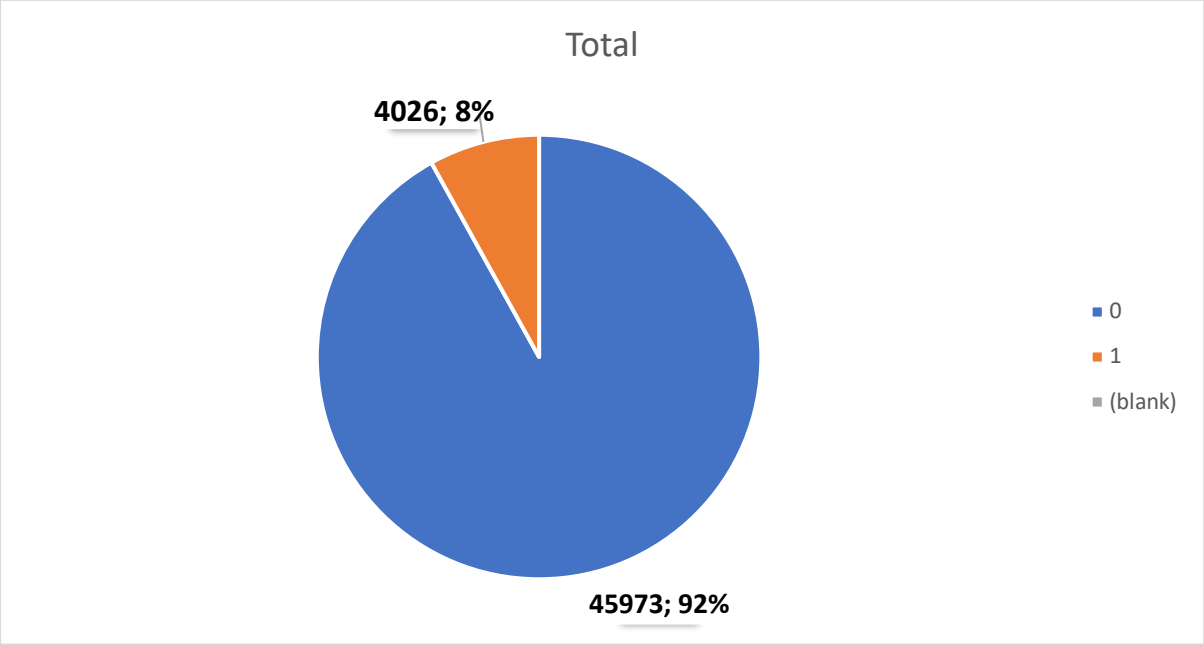
Ans: To calculate the data imbalance we need to calculate the Target count of each i.e. 0's and 1's.

Row Labels	Count of TARGET
0	45973
1	4026
(blank)	
Grand Total	**49999**

Here we can see that there is huge difference between the 0's target and 1's Target.

	COUNT OF 0'S & 1'S	RATIO	TARGET	CONTRIBUTION
0	45973	11.41902633	0	91.95
1	4026		1	8.05

From the chart we can see that there is unbalanced distribution of data as we can see that approx. 92% of the data to belongs to Target 0 and only 8% belongs to Target 1, so this creates imbalanced data.



Only 4026 data is of Target 1 out of 49999 rows which is very less as compared to Target 0 which has 45973 rows.

The ratio between both the data is approx.. 11.4%

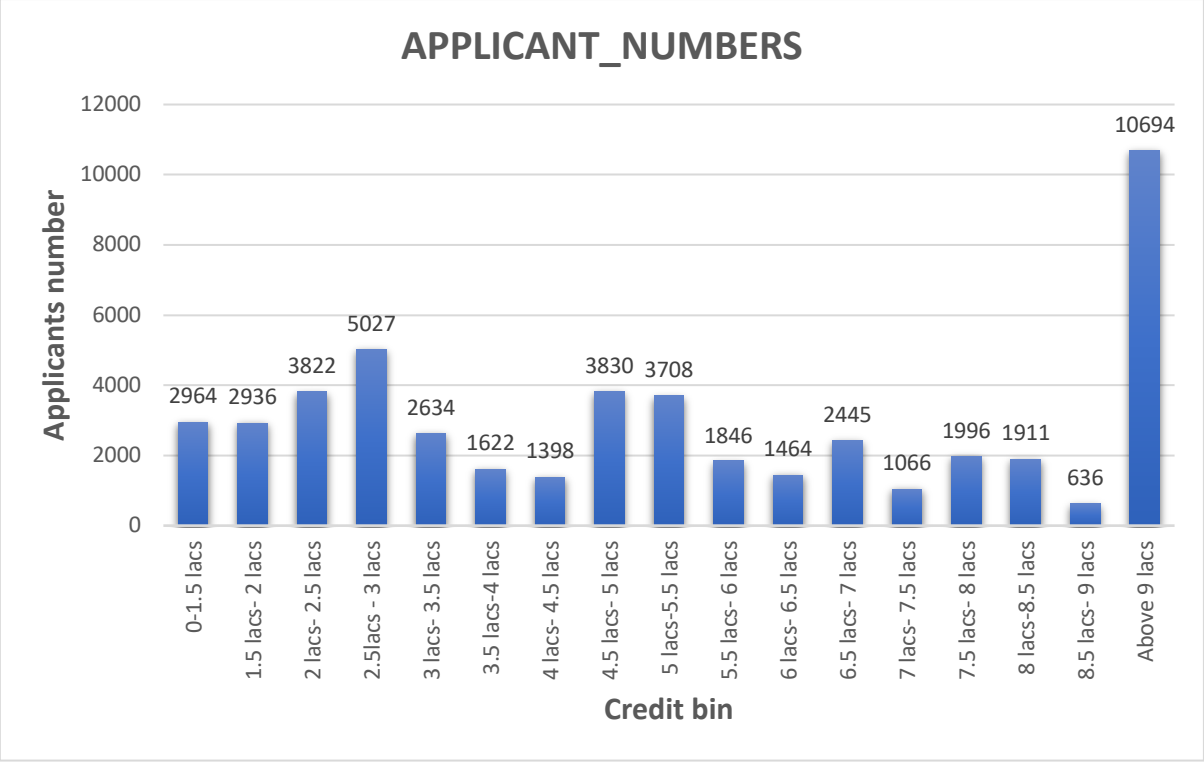
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: **To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes**.

Ans: Univariate analysis : Univariate analysis is a statistical method focused on examining one variable at a time to understand its characteristics, distribution, and behaviour within a dataset. It involves descriptive statistics such as mean, median, mode, and measures of variability like standard deviation and range. Through graphical representations like histograms, box plots, and bar charts,

univariate analysis helps identify patterns, outliers, and trends within the data. This analysis serves as a foundational step in exploring and understanding individual variables' properties before delving into more complex multivariate analyses, providing crucial insights for decision-making in various fields such as research, business, and academia.

To do the Univariate Analysis , I have created the BIN size of Amount, top identify the count in each interval.

CREDITE BIN	APPLICANT_NUMBERS
0-1.5 lacs	2964
1.5 lacs- 2 lacs	2936
2 lacs- 2.5 lacs	3822
2.5lacs - 3 lacs	5027
3 lacs- 3.5 lacs	2634
3.5 lacs-4 lacs	1622
4 lacs- 4.5 lacs	1398
4.5 lacs- 5 lacs	3830
5 lacs-5.5 lacs	3708
5.5 lacs- 6 lacs	1846
6 lacs- 6.5 lacs	1464
6.5 lacs- 7 lacs	2445
7 lacs- 7.5 lacs	1066
7.5 lacs- 8 lacs	1996
8 lacs-8.5 lacs	1911
8.5 lacs- 9 lacs	636
Above 9 lacs	10694



Univariate Analysis

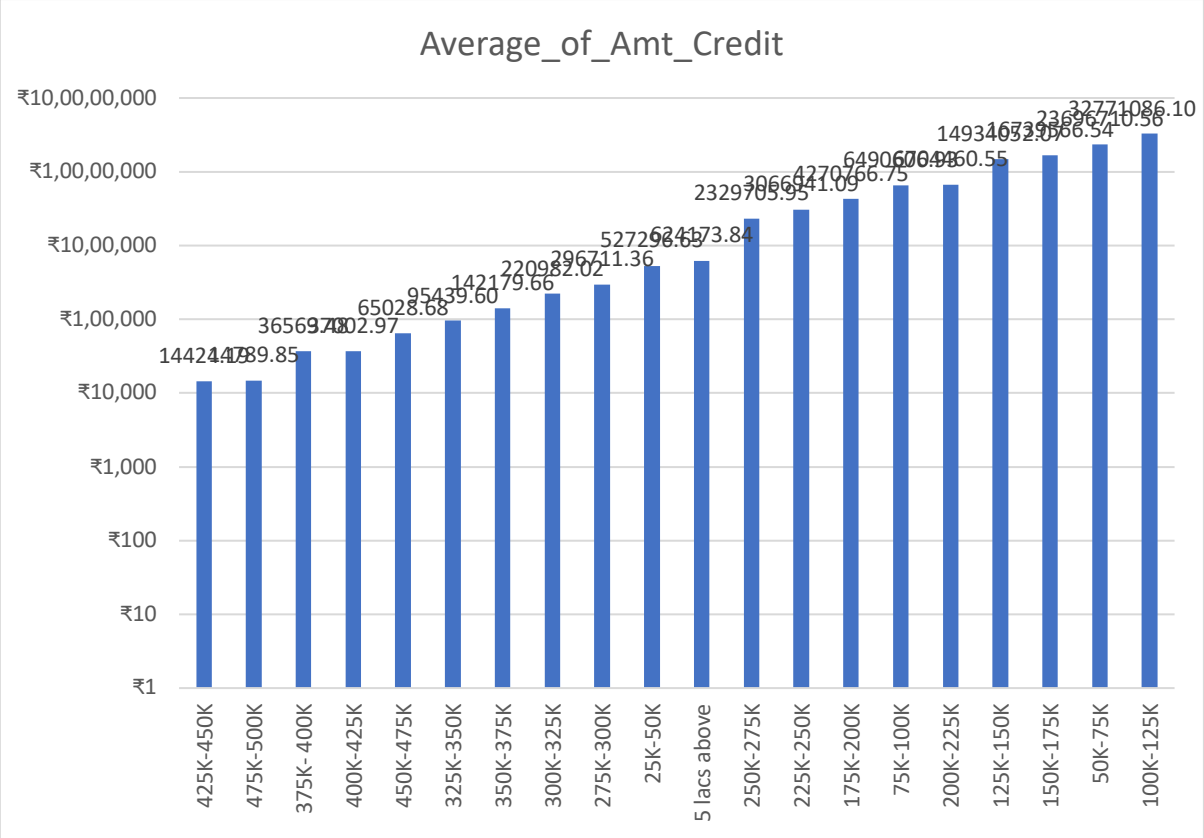
Here there is only to dimension involved to analyse the data.

Bi Variates: Bivariate analysis is a statistical technique aimed at understanding the relationship between two variables within a dataset. It explores how changes in one variable correlate or influence changes in another. Through methods like correlation analysis, scatter plots, and regression analysis, bivariate analysis uncovers patterns, dependencies, and associations between the variables. By examining the strength and direction of these relationships, it offers valuable insights into cause-and-effect dynamics, predictive modelling, and decision-making processes. Bivariate analysis is fundamental in fields like economics, social sciences, and data science, providing a basis for deeper exploration into multivariate relationships and complex phenomena.

INCOME_BIN	*Sum_of_Amt_Credit*	*Average_of_Amt_Credit*	Applicant_count
5 lacs above	501835765.5	624173.8377	804
475K-500K	47712069	14789.854	3226
450K-475K	413712432	65028.67526	6362
425K-450K	101661691.5	14424.19005	7048
400K-425K	288771169.5	37002.96893	7804
375K- 400K	203362875	36569.47941	5561
350K-375K	689144827.5	142179.6632	4847
325K-350K	308747097	95439.59722	3235
300K-325K	1012760582	220982.0165	4583
275K-300K	611818821	296711.3584	2062
250K-275K	1691366522	2329705.952	726

225K-250K	3480978137	3066941.089	1135
200K-225K	2319743349	6704460.546	346
175K-200K	3232970429	4270766.748	757
150K-175K	3347913308	16739566.54	200
125K-150K	4315941050	14934052.07	289
100K-125K	3408192954	32771086.1	104
75K-100K	2654658234	6490606.929	409
50K-75K	1113745397	23696710.56	47
25K-50K	239392669.5	527296.6289	454

INCOME_BIN	*Average_of_Amt_Credit*
425K-450K	14424.19
475K-500K	14789.85
375K- 400K	36569.48
400K-425K	37002.97
450K-475K	65028.68
325K-350K	95439.60
350K-375K	142179.66
300K-325K	220982.02
275K-300K	296711.36
25K-50K	527296.63
5 lacs above	624173.84
250K-275K	2329705.95
225K-250K	3066941.09
175K-200K	4270766.75
75K-100K	6490606.93
200K-225K	6704460.55
125K-150K	14934052.07
150K-175K	16739566.54
50K-75K	23696710.56
100K-125K	32771086.10



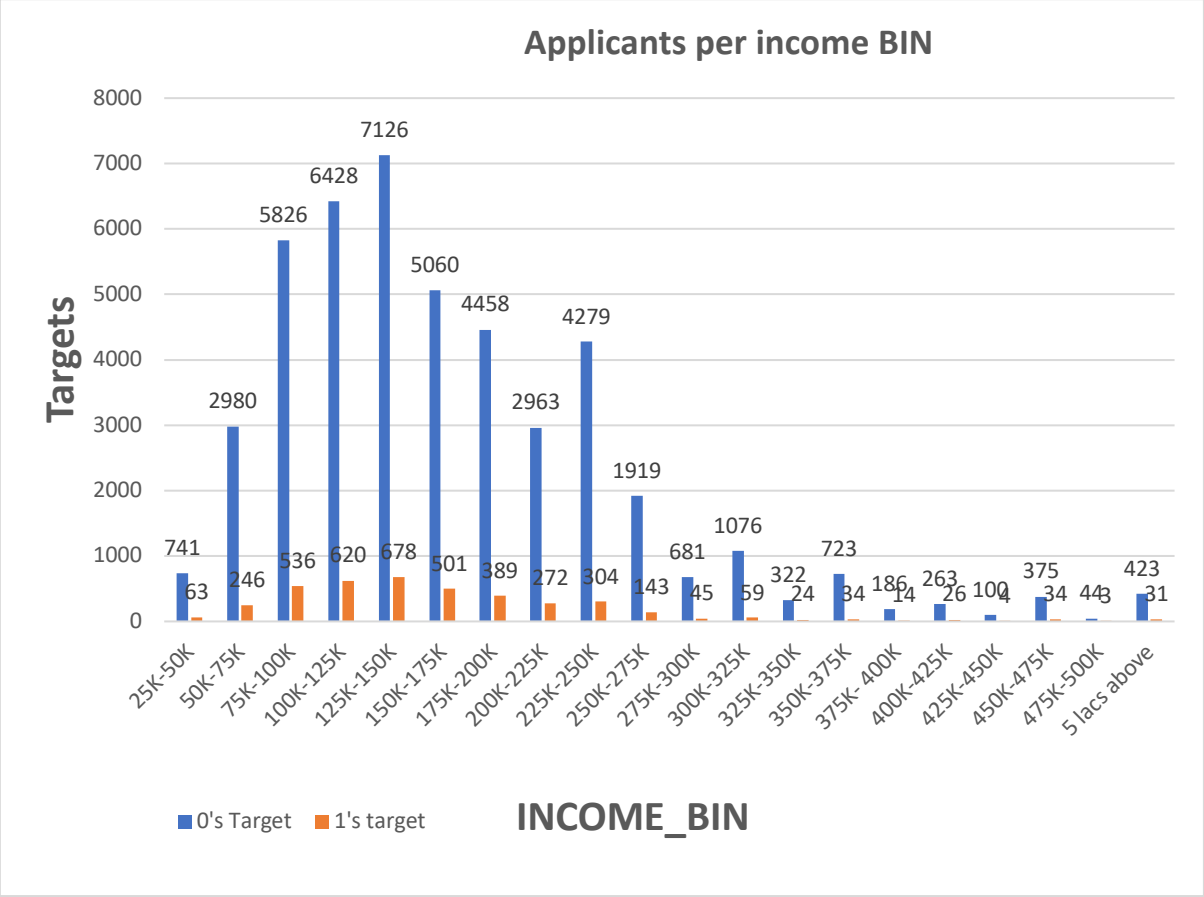
Bivariate Analysis

Segmented variate : Segmented variate analysis is a statistical method used to examine the relationship between two or more variables while considering the effect of additional factors, or segments, on this relationship. By dividing the data into distinct segments based on a categorical variable, such as age groups or geographic regions, segmented variate analysis allows for a more nuanced understanding of how different factors interact and influence the relationship between variables. This approach enables researchers to identify variations in the relationship across segments, providing valuable insights for targeted interventions, tailored strategies, and more precise decision-making in various fields, including marketing, healthcare, and social sciences.

INCOME_ BIN	0's Target	1's target	Total
25K-50K	741	63	804
50K-75K	2980	246	3226
75K-100K	5826	536	6362
100K-125K	6428	620	7048
125K-150K	7126	678	7804
150K-175K	5060	501	5561
175K-200K	4458	389	4847
200K-225K	2963	272	3235
225K-250K	4279	304	4583
250K-275K	1919	143	2062

275K-300K	681	45	726
300K-325K	1076	59	1135
325K-350K	322	24	346
350K-375K	723	34	757
375K- 400K	186	14	200
400K-425K	263	26	289
425K-450K	100	4	104
450K-475K	375	34	409
475K-500K	44	3	47
5 lacs above	423	31	454
Total	**45973**	**4026**	**49999**

This table illustrates the distribution of targets (0's and 1's) across various income bins. It appears to be structured with counts of targets for each income range, facilitating analysis and understanding of how income levels relate to the target outcomes.



The provided charts represents a segmented variate analysis, examining the relationship between income levels and target outcomes (0's and 1's). The data is segmented into income bins ranging from 25K to 5 lacs above. Within each income bin, the counts of 0's and 1's targets are recorded, along with the total count. This segmented analysis enables

a detailed understanding of how different income levels correlate with target outcomes, providing insights into potential patterns or associations between income and the likelihood of achieving the target.

E. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Ans: To find the understanding of correlation between different attributes we will divide the data into two parts i.e. Target 0 which means the payment is done successfully and Target 1 means payment difficulties.

After separating the table we will find the correlation of applicant with payment success.

CORRELATION OF APLLICANT WITH PAYMENT SUCCESSFUL								
CNT_CHILDREN	1							
AMT_INCOME_T	0.036	1						
AMT_CREDIT	0.006	0.378	1					
REGION_POPULA	-0.025	0.182	0.096	1				
DAYS_BIRTH IN Y	-0.004	-0.004	-0.004	-0.005	1			
DAYS_EMPLOYEL	-0.246	-0.162	-0.075	-0.007	0.005	1		
DAYS_ID_PUBLIS	0.033	-0.032	0.008	0.002	0.000	0.275	1	
REGION_RATING	0.018	-0.220	-0.112	-0.537	0.006	0.043	0.013	1
	CNT_CHILD	AMT_IN	AMT_CI	REGION_	DAYS_B	DAYS_E	DAYS_ID_	REGION_

The provided correlation matrix reveals the correlation coefficients between various applicant attributes and the success rate of payment. Here's a summary of the correlations:

- **CNT_CHILDREN**: There is a weak positive correlation (0.036) between the number of children and payment success.
- **AMT_INCOME_TOTAL**: Income shows a very weak positive correlation (0.006) with payment success.
- **AMT_CREDIT**: A moderate positive correlation (0.378) exists between credit amount and payment success.
- **REGION_POPULATION_RELATIVE**: The population relative to region demonstrates a weak negative correlation (-0.025) with payment success.
- **DAYS_BIRTH IN YRS**: Age (in years) doesn't exhibit a significant correlation with payment success.

- **DAYS_EMPLOYED IN YRS**: Employment duration in years displays a moderate negative correlation (-0.246) with payment success, suggesting longer employment tenure relates to higher success rates.

- **DAYS_ID_PUBLISH IN YRS**: The age of ID publication indicates a weak positive correlation (0.033) with payment success.

- **REGION_RATING_CLIENT_W_CITY**: The rating of the region where the client lives in the city shows a weak positive correlation (0.018) with payment success.
- These correlations provide insight into which applicant attributes may have stronger associations with payment success, aiding in risk assessment and decision-making processes.

Now Let's check the correlation of attributes in Target 1 which shows payment with difficulties.

CORRELATION OF APLLICANT WITH PAYMENT DIFICULTIES

CNT_CHILDREN	1							
AMT_INCOME_TOTAL	0.010	1						
AMT_CREDIT	0.008	0.015	1					
REGION_POPULATION_	-0.020	-0.006	0.068	1				
DAYS_BIRTH IN YRS	0.250	0.009	-0.143	-0.016	1			
DAYS_EMPLOYED IN YI	-0.190	-0.012	0.019	0.008	-0.588	1		
DAYS_ID_PUBLISH IN Y	0.042	0.009	0.044	0.005	-0.248	0.233	1	
REGION_RATING_CLIEI	0.055	-0.013	-0.053	-0.432	0.038	-0.004	-0.014	1
	CNT_CHILDREI	AMT_INCOME_	AMT_CREDIT	REGION_POPU	DAYS_BIRTH I	DAYS_EMPLOY	DAYS_ID_PUB	REGION_RATIN

The correlation matrix provided displays the correlation coefficients between various applicant attributes and the likelihood of payment difficulties. Here's a summary of the correlations:

- **CNT_CHILDREN**: There's a very weak positive correlation (0.010) between the number of children and payment difficulties.

- **AMT_INCOME_TOTAL**: Income shows a very weak positive correlation (0.008) with payment difficulties.

- **AMT_CREDIT**: Credit amount exhibits a very weak positive correlation (0.015) with payment difficulties.

- **REGION_POPULATION_RELATIVE**: The population relative to region displays a very weak negative correlation (-0.020) with payment difficulties.

- **DAYS_BIRTH IN YRS**: Age (in years) has a moderate positive correlation (0.250) with payment difficulties, indicating that older applicants may face more difficulties.

- **DAYS_EMPLOYED IN YRS**: Employment duration in years demonstrates a moderate negative correlation (-0.190) with payment difficulties, implying longer employment tenure may reduce the likelihood of difficulties.

- **DAYS_ID_PUBLISH IN YRS**: The age of ID publication shows a weak positive correlation (0.042) with payment difficulties.

- **REGION_RATING_CLIENT_W_CITY**: The rating of the region where the client lives in the city displays a weak positive correlation (0.055) with payment difficulties.
- These correlations offer insights into which applicant attributes may be more closely associated with the occurrence of payment difficulties, aiding in risk assessment and decision-making processes in lending or financial contexts.

Results: We have successfully cleaned the data and derived the charts of various attributes which show the distribution of income and the amount credited to various applicants. There applicants who can pay the loan easily and there are applicants who faces difficulties in repaying the loan. And on various conditions bank may reject and approve the loan of the applicants.

Drive Link:
https://drive.google.com/drive/folders/1keuBSn4sYD642ZqeU95hZSBzLEOMbtZk?usp=sharing

**********THE END**********