A PROJECT REPORT ON

BANK LOAN DATA ANALYSIS

Submitted By: -

Pritesh Kumar Bag

Trainity

C-97, C-97, Ahinsa Cir, Panch Batti, C Scheme,

Ashok Nagar, Jaipur, Rajasthan 302001

AUGUST 29,2023

PROJECT DESCRIPTION

Bank loan data analysis offers crucial insights into customer behaviour, credit risk, and portfolio performance. By scrutinizing trends, default probabilities, and customer segments, it enables informed decision-making. It aids in optimizing lending strategies, mitigating risks, and enhancing profitability by identifying potential defaults, improving credit scoring models, and tailoring offerings to diverse customer needs. This analysis ensures regulatory compliance, facilitates proactive measures against defaults, and supports the development of effective loan products, fostering a more robust and resilient banking system while improving overall customer satisfaction and financial outcomes. Analysing bank loan data involves various steps aimed at extracting insights, identifying patterns, and making informed decisions.

APPROACH

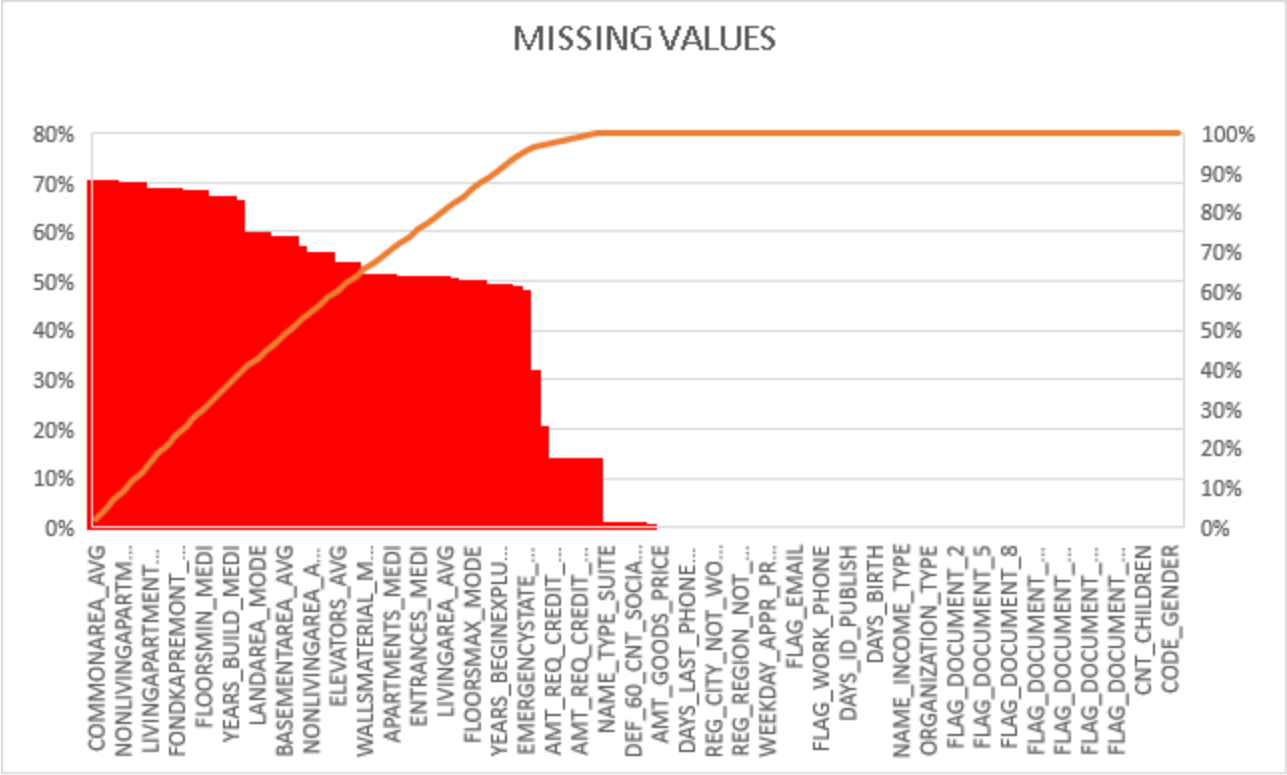
- Downloaded the given datafile
- Understanding the datafile
- Checking the blanks and outliers
- Removing unwanted data
- Drawing summary from the data
- Used formulas, pivot table and charts

TECH STACK USED

- Microsoft Excel 2016
- SOLVED DATASET- CLICK HERE

INSIGHTS

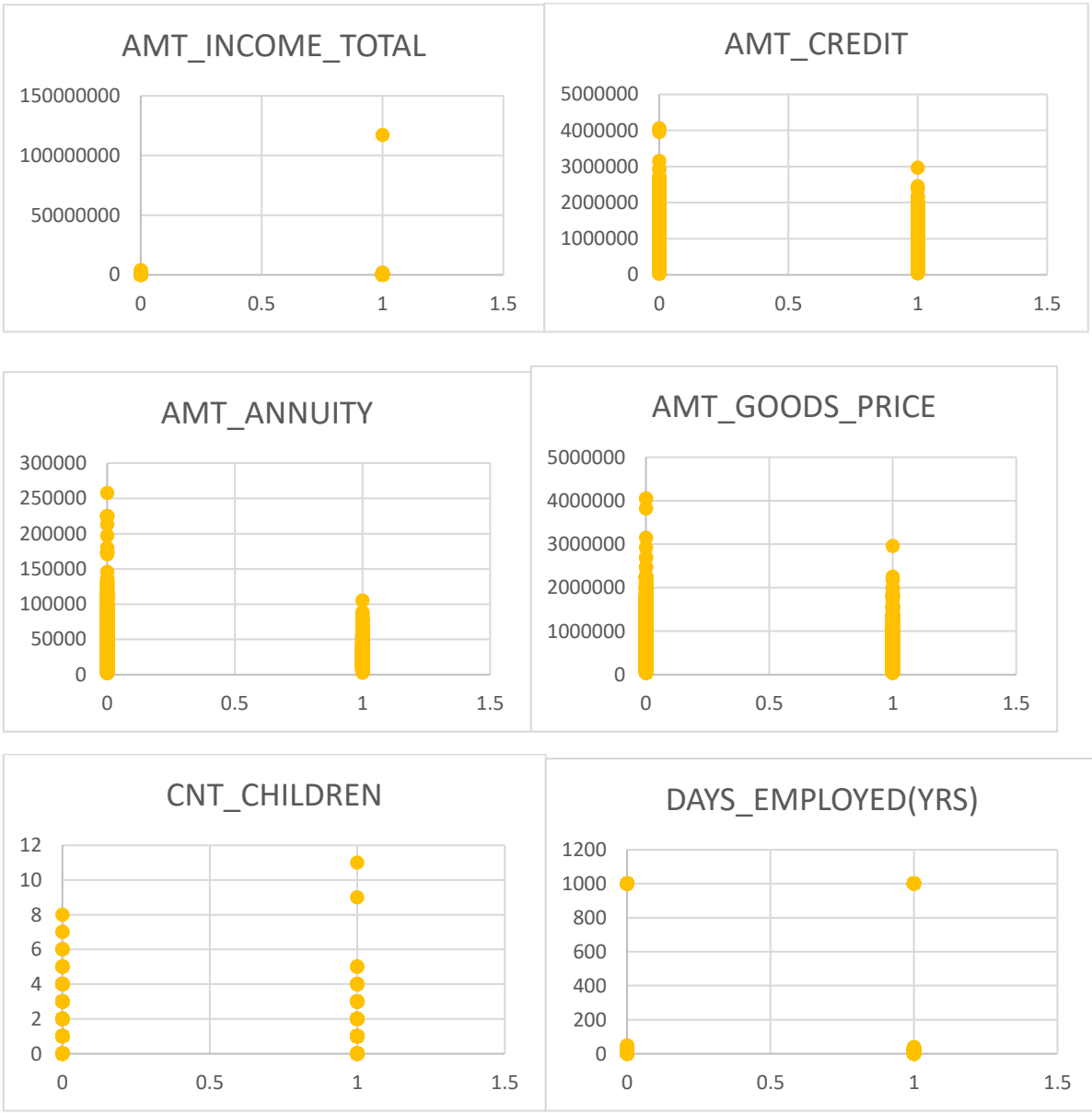
A. Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis. Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features. Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel. Graph suggestion: Create a bar chart or column chart to visualize the proportion of missing values for each variable.



INFERENCE

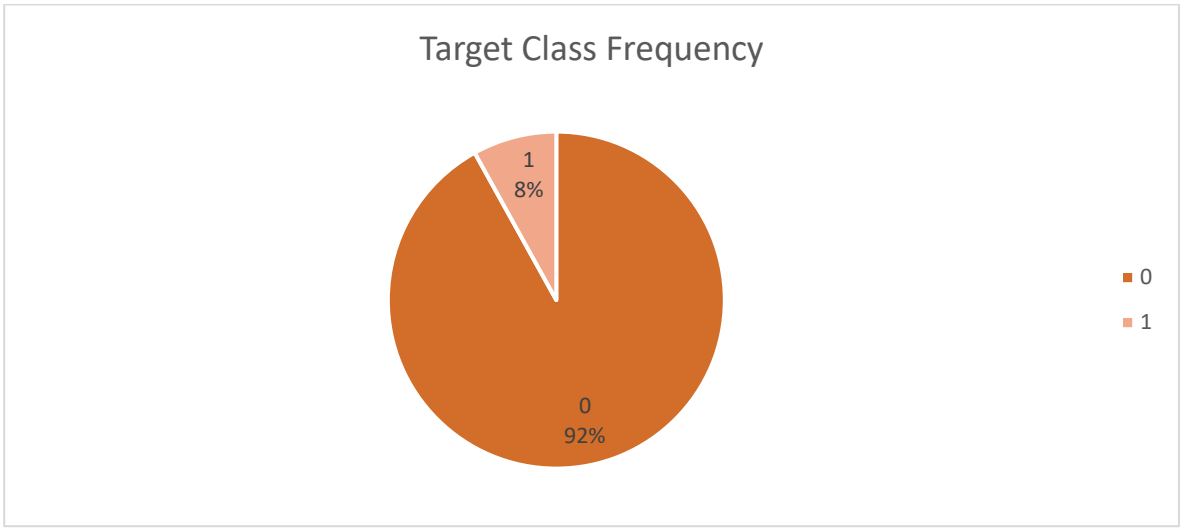
APPLICATION DATA
COLUMNS COUNT=122 MISSING VALUES COLUMNS ABOVE 30%=50 COLUMNS COUNT AFTER COLUMNS REMOVED=72

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. Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables. Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation. Graph suggestion: Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.



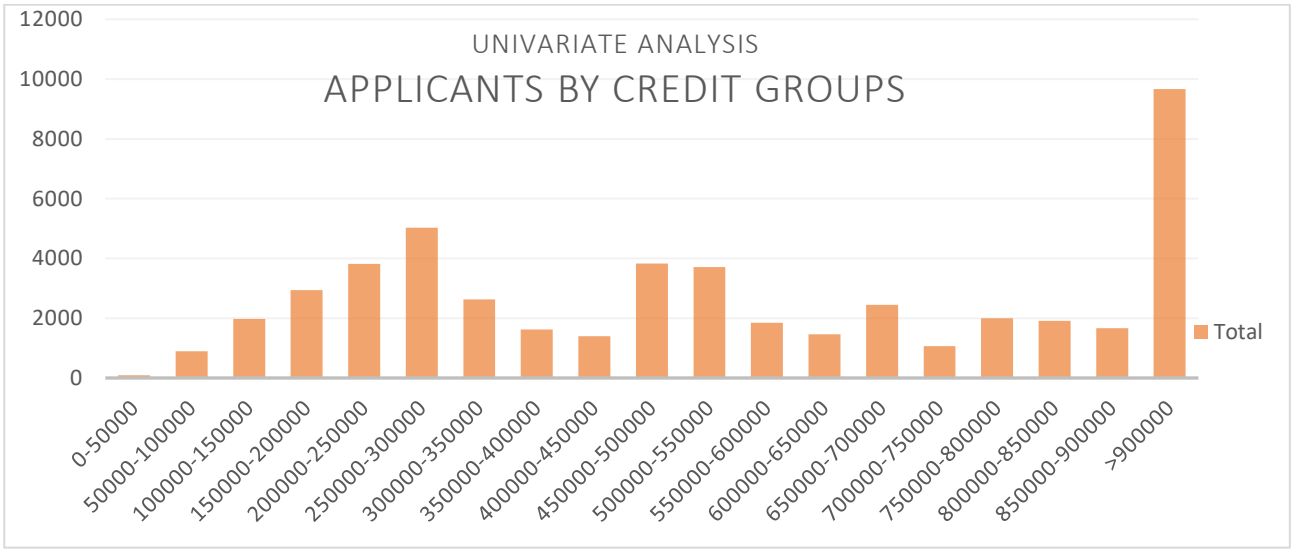
INFERENCE- Outliers were found comparison the values of mean and median and of all columns. Further Maximum value and Quartile 3 difference was compared with Quartile 1 and Minimum value to get the outliers of all columns and then columns with outliers have been plotted on graph to get a clear picture.

C. Analyse 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. Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions. Graph suggestion: Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

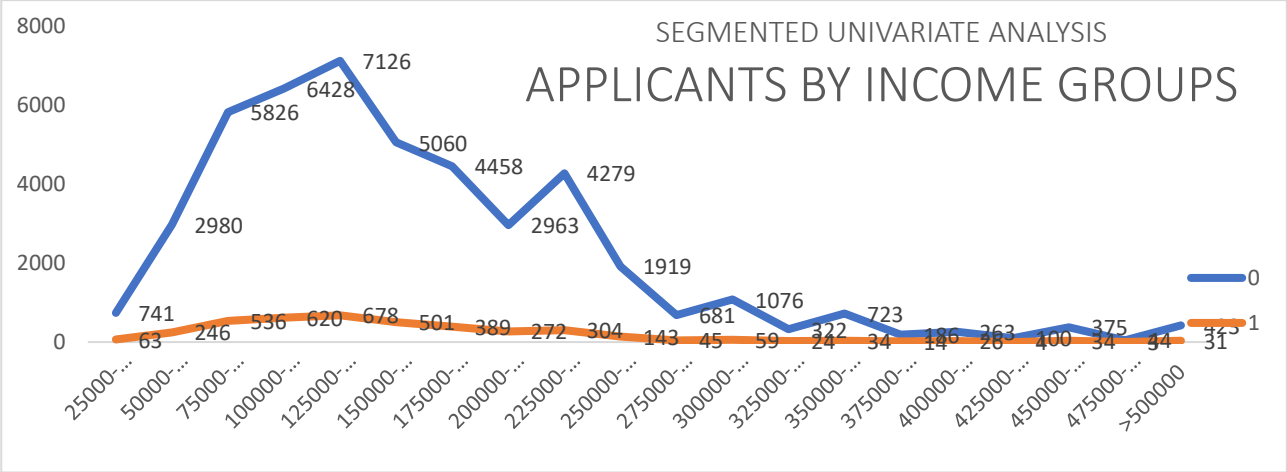


INFERENCE- There is a huge data imbalance as the target 1 applicants represent only 1 percent of the total applicants, whereas applicants with other problem that is target 0 is 92 percent

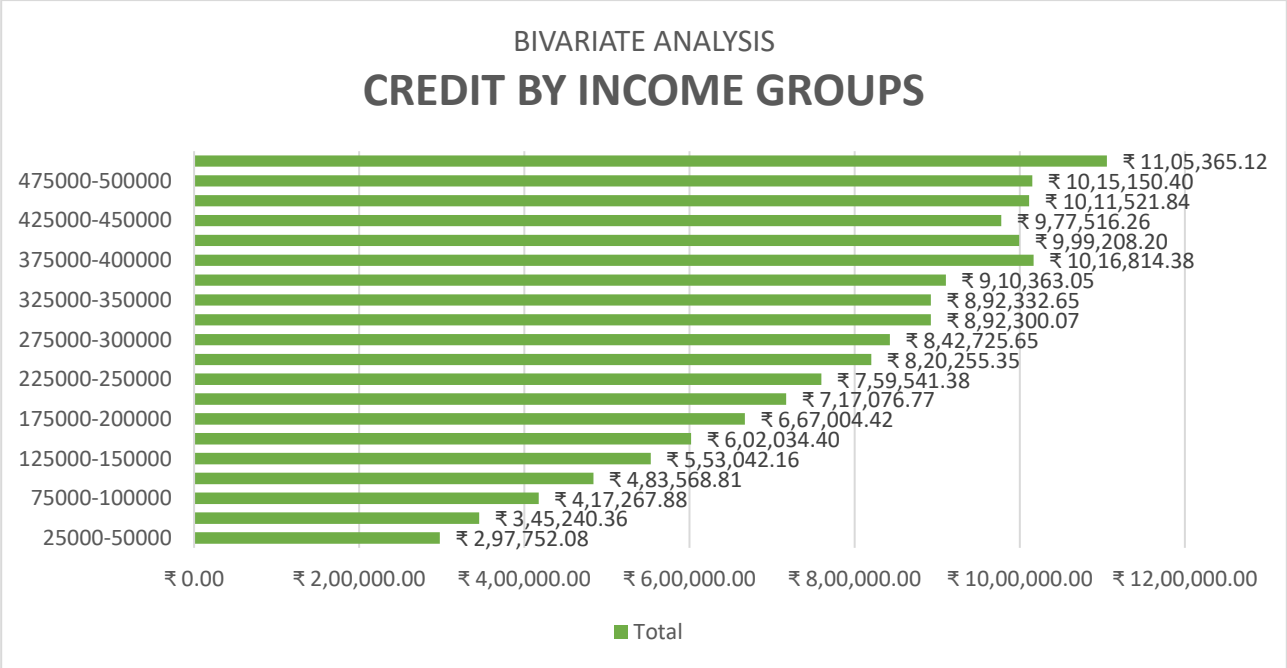
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes. Task: Graph suggestion: Create histograms, bar charts, or box plots to visualize the distributions of variables.



INTERPRETATION- This univariate analysis shows the total number of applicants from different credit bins based upon the target. There are more than 9000 applicants who wants a loan above 9 lakhs.



INTERPRETATION- Segmented univariate analysis shows the total number of applicants from different income groups based upon the target 1 or target 0 type of problems faced by them. Income bin 125000-150000 have huge applicants both on target 1 and target 0 kind of problems



INTERPRETATION- This bivariate analysis shows the amount of credit taken by people of different income bins. There is a positive correlation among the people for opting higher credit on increasing of incomes.

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. Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions. Hint: Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario. Graph suggestion: Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colours or shading.

TARGET-0

	CNT_CHIL	AMT_INC	AMT_CRE	REGION_P	DAYS_BIR	DAYS_EM	DAYS_ID_	REGION_RATING_	CNT_FAM_MEMBERS
CNT_CHILDREN	1	0.03632	0.005705	-0.02491	-0.33588	-0.24552	0.032537	0.021288992	0.87923936
AMT_INCOME_TOTAL	0.03632	1	0.377966	0.181941	-0.07377	-0.16168	-0.03229	-0.205031899	0.041613404
AMT_CREDIT	0.005705	0.377966	1	0.095539	0.051084	-0.07473	0.00829	-0.102556478	0.064877635
REGION_POPULATION_RELA	-0.02491	0.181941	0.095539	1	0.030435	-0.00677	0.002236	-0.539333113	-0.023006667
DAYS_BIRTH(YRS)	-0.33588	-0.07377	0.051084	0.030435	1	0.623475	0.270073	-0.00902485	-0.284384945
DAYS_EMPLOYED(YRS)	-0.24552	-0.16168	-0.07473	-0.00677	0.623475	1	0.274516	0.040937165	-0.234767657
DAYS_ID_PUBLISH(YRS)	0.032537	-0.03229	0.00829	0.002236	0.270073	0.274516	1	0.008097427	0.025058177
REGION_RATING_CLIENT	0.021289	-0.20503	-0.10256	-0.53933	-0.00902	0.040937	0.008097	1	0.022204177
CNT_FAM_MEMBERS	0.879239	0.041613	0.064878	-0.02301	-0.28438	-0.23477	0.025058	0.022204177	1

INFERENCE-We can see that there is a close relation between the amount of income, amount of credit, region. Count of family member is also dependent upon count of children.

TARGET-1

	CNT_CHILDREN	AMT_INC	AMT_CRE	REGION_P	DAYS_BIR	DAYS_EM	DAYS_ID_	REGION_RATING_CL	CNT_FAM_MEMBERS
CNT_CHIL	1	0.01011	0.007602	-0.02036	-0.24967	-0.18977	0.042361	0.055515557	0.892521875
AMT_INC	0.010110177	1	0.015271	-0.00618	-0.00903	-0.01176	0.009122	-0.012846697	0.013121678
AMT_CRE	0.007601905	0.015271	1	0.067776	0.142506	0.018782	0.043772	-0.045024534	0.06124869
REGION_P	-0.020359154	-0.00618	0.067776	1	0.016469	0.00771	0.005119	-0.430032303	-0.017257146
DAYS_BIR	-0.2496732	-0.00903	0.142506	0.016469	1	0.588243	0.247897	-0.045027112	-0.199141397
DAYS_EMI	-0.189773227	-0.01176	0.018782	0.00771	0.588243	1	0.232662	-0.009237108	-0.183362962
DAYS_ID_	0.042360717	0.009122	0.043772	0.005119	0.247897	0.232662	1	-0.025335227	0.044037815
REGION_R	0.055515557	-0.01285	-0.04502	-0.43003	-0.04503	-0.00924	-0.02534	1	0.057279521
CNT_FAM_	0.892521875	0.013122	0.061249	-0.01726	-0.19914	-0.18336	0.044038	0.057279521	1

INFERENCE-We can see that there is a close relation between the amount of income, amount of credit, region. Count of family member is also dependent upon count of children. There is close relation between days birth and days employed

RESULTS

- Outliers were found comparison the values of mean and median and of all columns. Further Maximum value and Quartile 3 difference was compared with Quartile 1 and Minimum value to get the outliers of all columns and then columns with outliers have been plotted on graph to get a clear picture.

- There is a huge data imbalance as the target 1 applicants represent only 1 percent of the total applicants, whereas applicants with other problem that is target 0 is 92 percent

- Segmented univariate analysis shows the total number of applicants from different income groups based upon the target 1 or target 0 type of problems faced by them. Income bin 125000-150000 have huge applicants both on target 1 and target 0 kind of problems

- This univariate analysis shows the total number of applicants from different credit bins based upon the target. There are more than 9000 applicants who wants a loan above 9 lakhs.

- This bivariate analysis shows the amount of credit taken by people of different income bins. There is a positive correlation among the people for opting higher credit on increasing of incomes.

- We can see that there is a close relation between the amount of income, amount of credit, region. Count of family member is also dependent upon count of children. There is close relation between days birth and days employed