PROJECT 6

Project Description:

This project is about the Loan application Case study , which is to be solved using EDA(Exploratory Data Analysis).EDA(Exploratory Data Analysis)is an approach that is used to analyze the data and discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations.

Importance of using EDA for analyzing data sets is:

- Helps identify errors in data sets.
- Gives a better understanding of the data set.
- Helps detect outliers or anomalous events.
- Helps understand data set variables and the relationship among them.

Objective of Exploratory Data Analysis

The goal of EDA is to allow data scientists to get deep insight into a data set and at the same time provide specific outcomes that a data scientist would want to extract from the data set. It includes:

- List of outliers
- Estimates for parameters
- Uncertainties about those estimates
- List of all important factors
- Conclusions or assumptions as to whether certain individual factors are statistically essential
- Optimal settings
- A good predictive model

Approach:

The loan application consists of 2 datasets which was provides to us , to analyze the data and help in solving the below questions. The 2 datasets are :

- **application_data.csv (**we have 122 rows and 49999 columns in it **)**

- **previous_application.csv(**we have 37 rows and 50000 columns in it **)**

Each consists of number of columns and rows to perform EDA on it . The rows and column which are not required in further analysis and which carries most numbers of blanks I deleted it .

After that I categorized the data find out the outliers and then deleted . Then I started working on then began performing univariate and bivariate analysis using
pivot tables and charts.

Tech-Stack Used:

- Microsoft Excel 365.

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.

- **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
- **Hint:** 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.
-

ANSWER:

The 2 datasets were provided to us.

- **application_data.csv (**we have 122 rows and 49999 columns in it **)**

- **previous_application.csv(**we have 37 rows and 50000 columns in it

The given dataset has huge number of columns . I filtered the data .

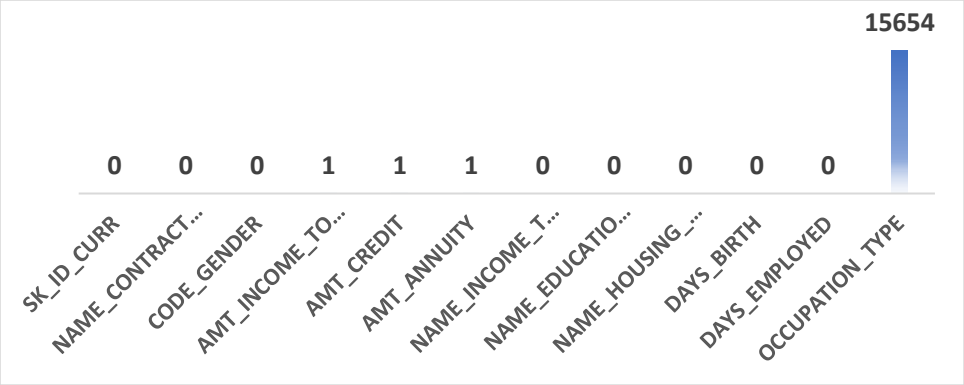
- I deleted the column which consists of most blank data around more than 5%.

-I also deleted some of the unused columns which were not necessary .

The Datasets are based on 2 categories of variables.

1. Categorical Variable
2. Numerical Variable

Categorical data	Numerical data
OCCUPATION_TYPE	SK_ID_CURR
NAME_HOUSING_TYPE	DAYS_EMPLOYED
NAME_EDUCATION_TYPE	DAYS_BIRTH
CODE_GENDER	AMT_ANNUITY
NAME_CONTRACT_TYPE	AMT_CREDIT
TARGET	AMT_INCOME_TOTAL



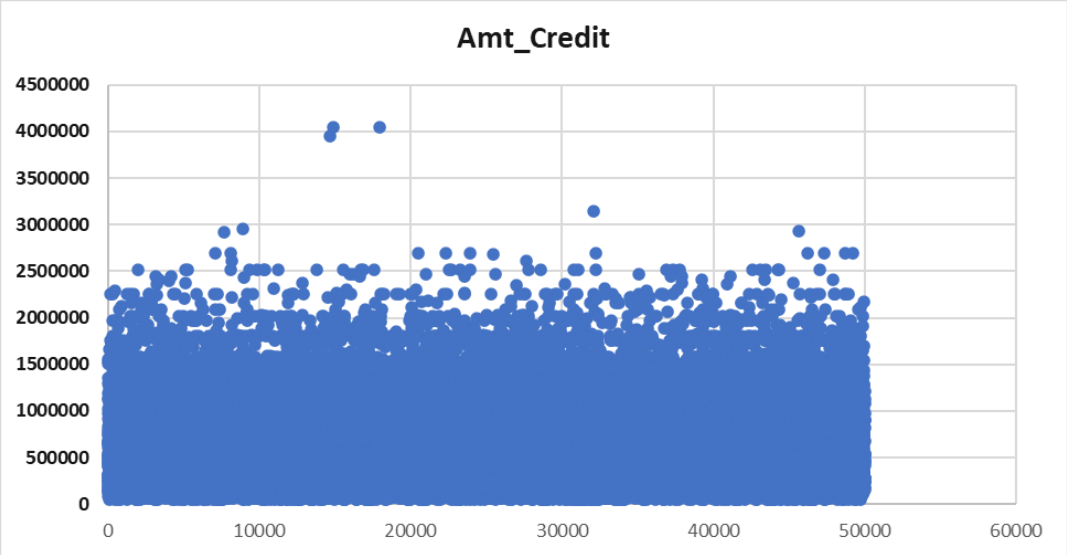
- This chart is to visualize the proportion of missing values for each variable.

.

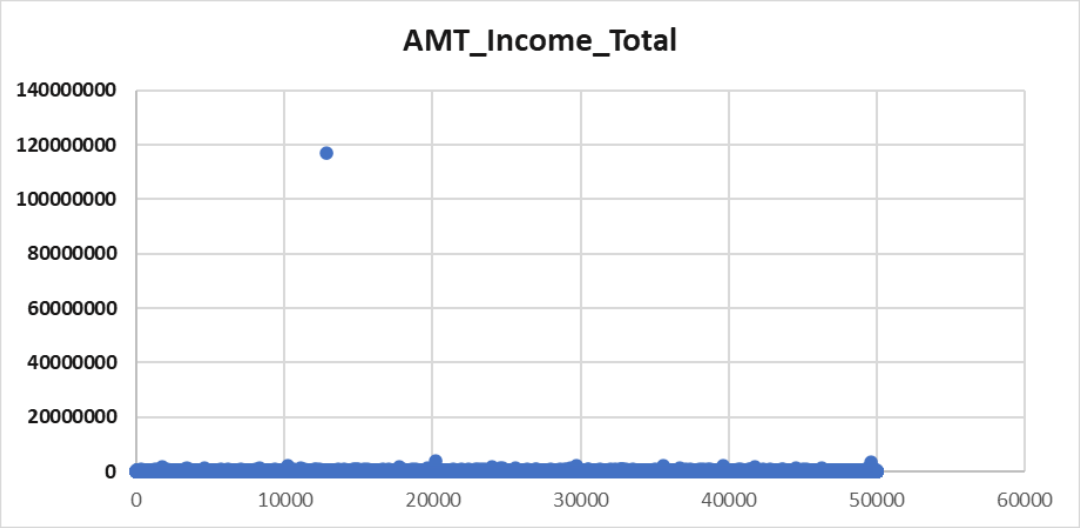
B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- **Hint:** 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.

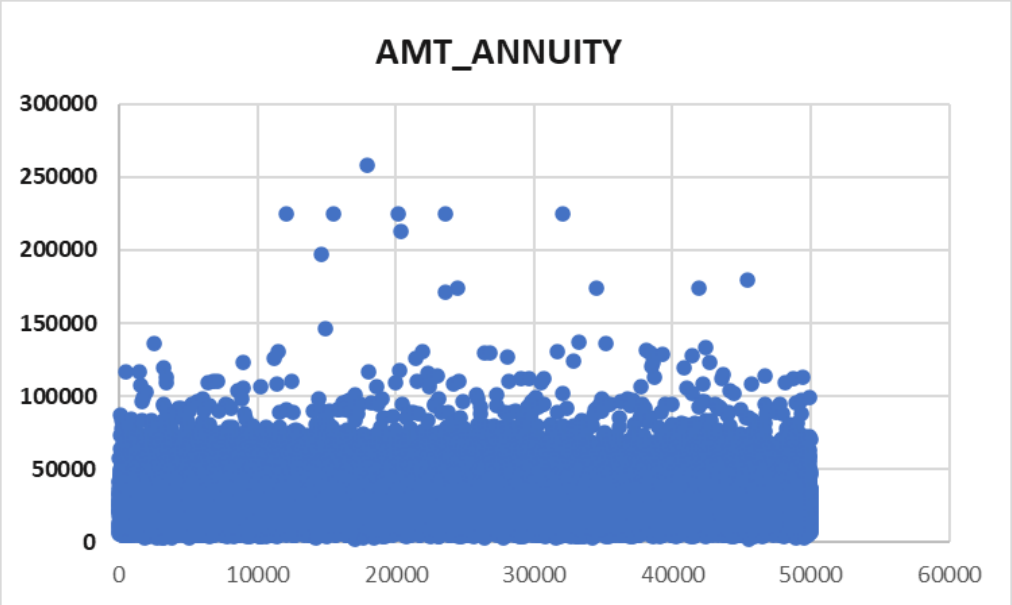
An Outlier is a data point that differs significantly from other observations.



Q1	Q3	IQR	Upper Bound	Lower Bound	Min	Max
270000	808650	538650	1616625	-537975	45000	4050000



Q1	Q3	IQR	Upper Bound	Lower Bound	Min	Max
112500	202500	90000	506250	-22500	25650	117000000

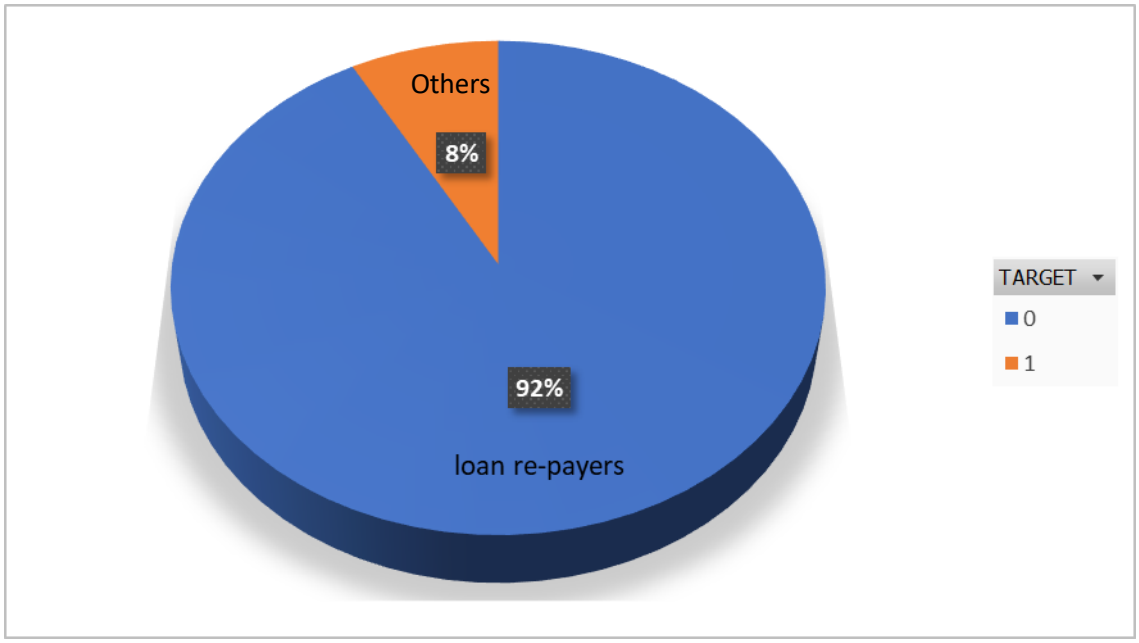


Q1	Q3	IQR	Upper Bound	Lower Bound	Min	Max
16456.5	34596	18139.5	61805.25	-10752.75	2052	258025.5

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.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- **Hint:** Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.

- **Graph suggestion:** Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.



Here we can see there are:

- 45973 people who are falling in a category of repaying the loan on a time.
- And almost 8% falls in other category

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:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

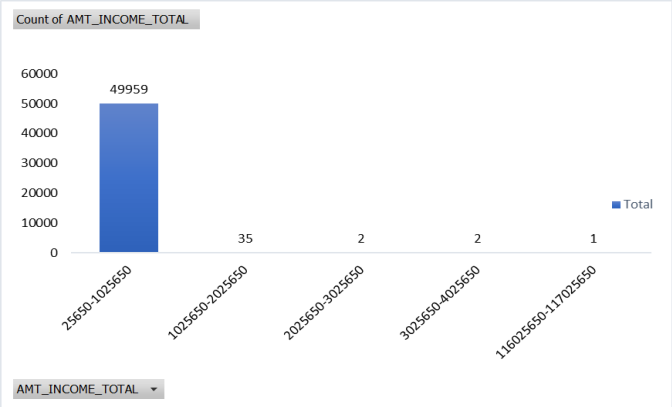
- **Hint:** Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.

- **Graph suggestion:** Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Univariate Analysis:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

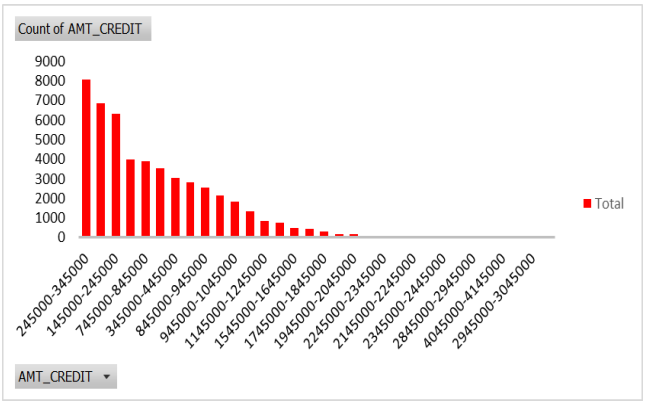
Univariate analysis on Amount_income:



Univariate Analysis on AMT_INCOME	Value
Mean	170767.6
Median	145800
Mode	135000
Standard Deviation	531819.10
Interquartile Range	90000
Range	116974350

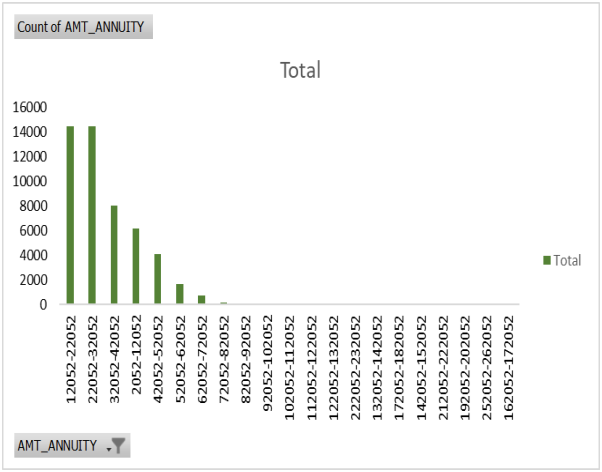
We can say that person having income of 25650 -1025650 has higher rate of taking the loan.

Univariate analysis on Amount_Credit:



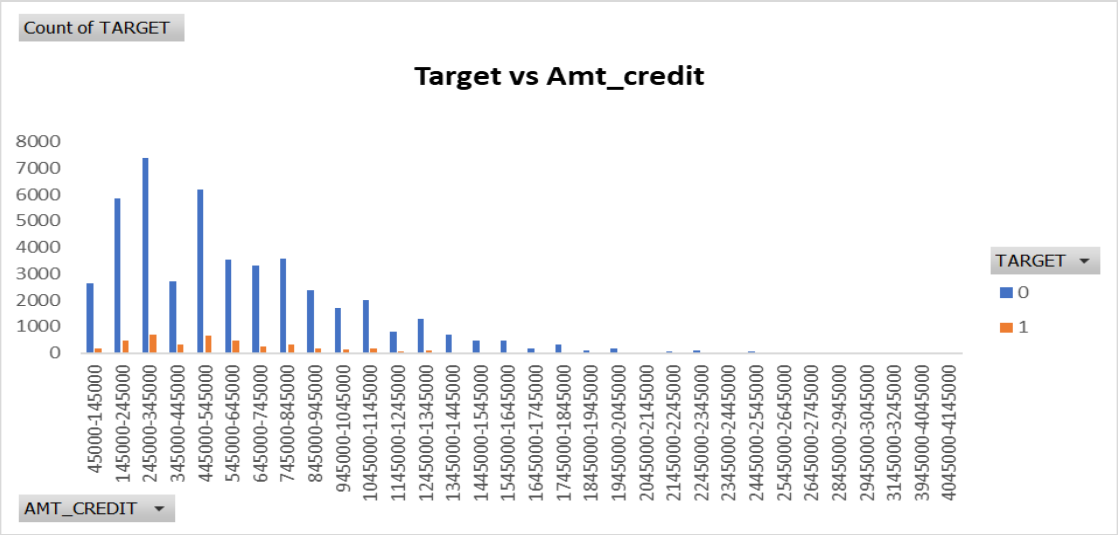
Univariate Analysis on AMT_CREDIT	Value
Mean	599700.58
Median	514777.5
Mode	450000
Standard Deviation	402415.43
Interquartile Range	538650
Range	4005000

Univariate analysis on Amount_Annuity:



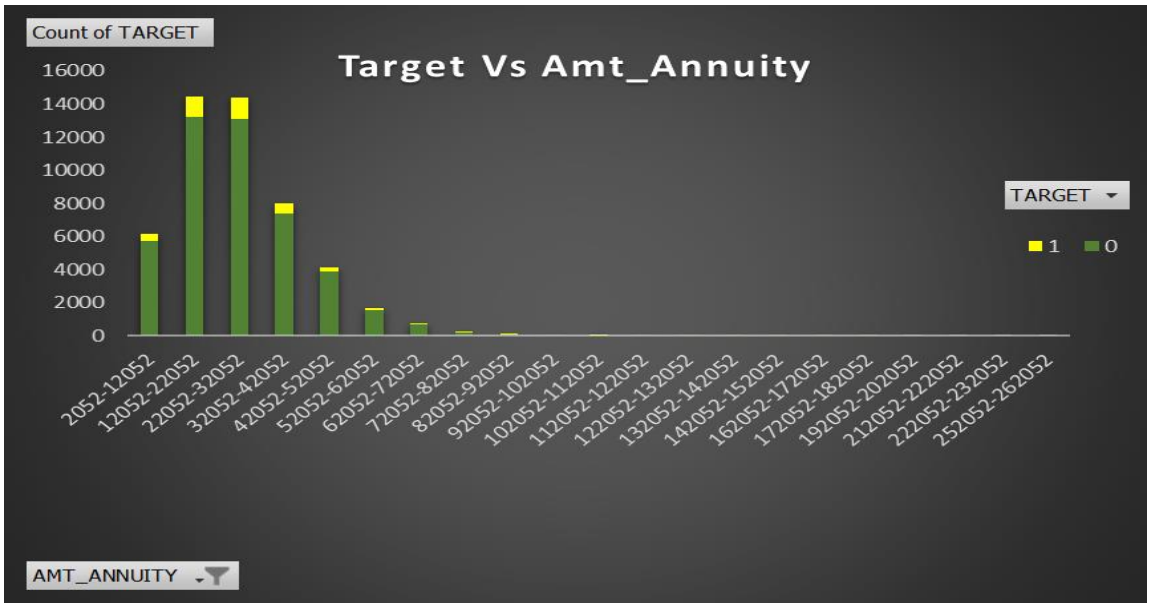
Univariate Analysis on AMT_ANNUITY	Value
Mean	27107.38
Median	24939
Mode	9000
Standard Deviation	14562.94
Interquartile Range	18139.5
Range	255973.5

Segmented Univariate analysis on Amount_CREDIT:



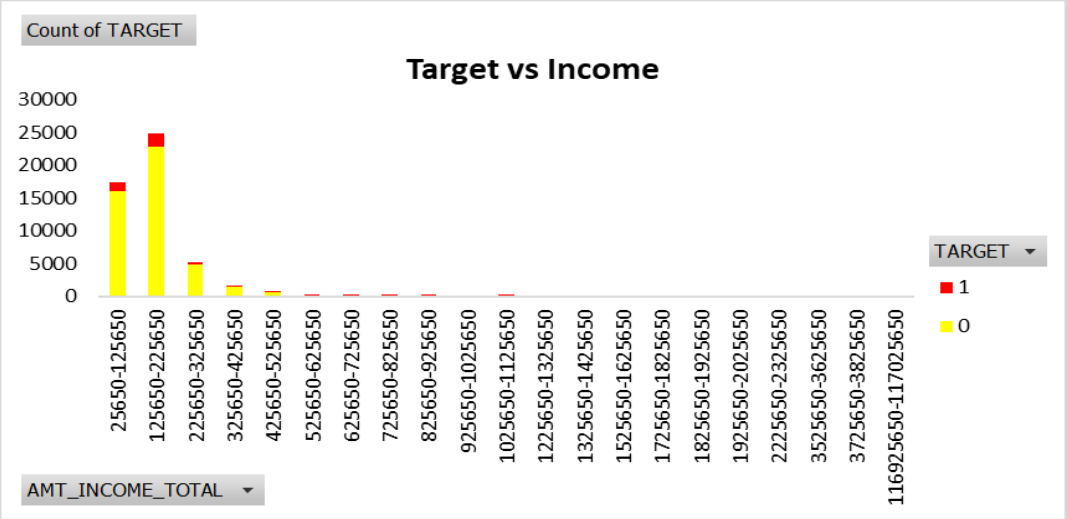
Here we can see that the person having Amt_credit in between 245000-345000 , tends to pay the loan on time.

Segmented Univariate analysis on Amount_Annuity:



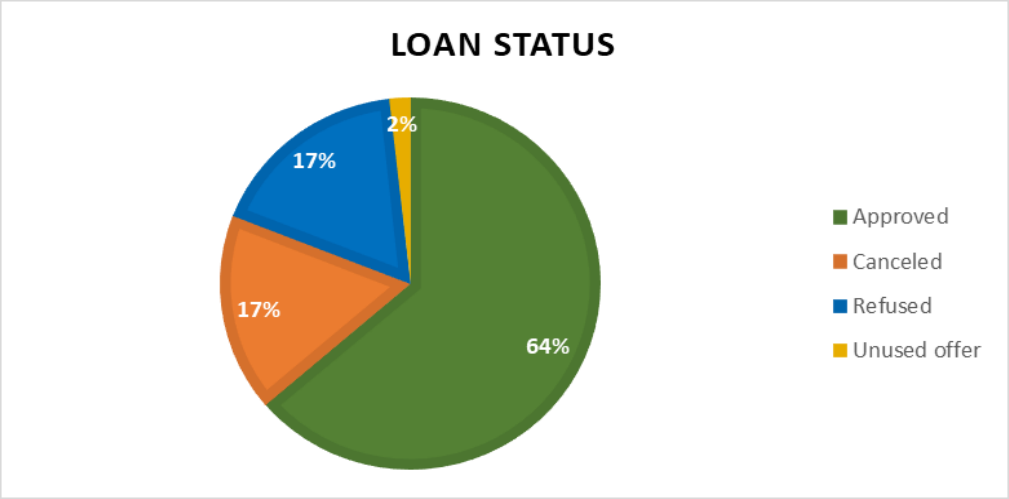
Here we can see the the person having Amt_Annuity in between 12052-22052 have paid the loan on time.

Segmented Univariate analysis on Amount_Income:



The person having income in between 125650-225650 have higher chances of repaying the loan on time.

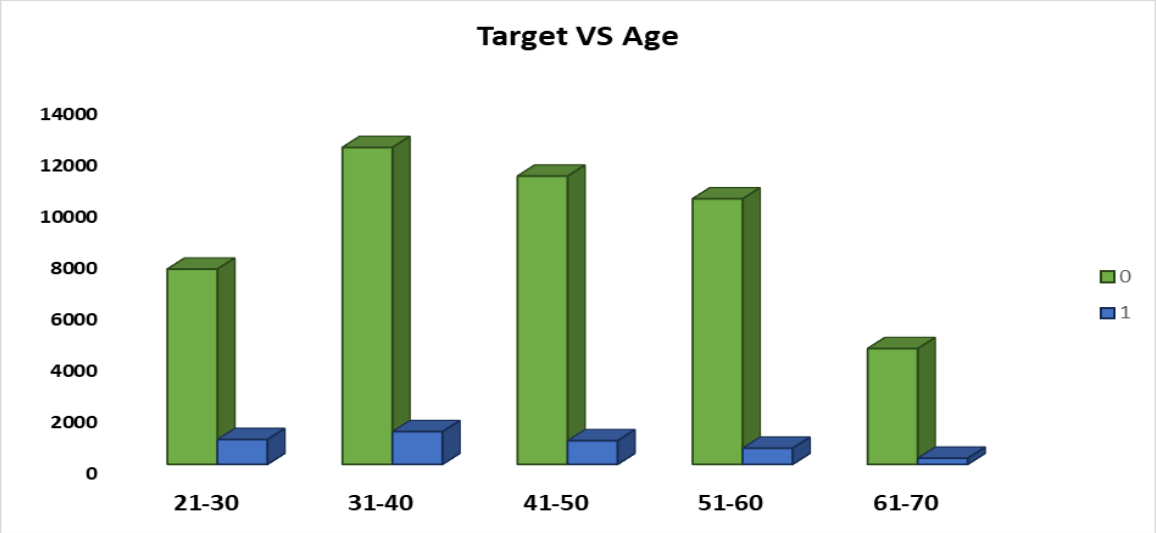
Segmented Univariate analysis on Loan Status:



We have only 64% of approved loan status with us and rest are as seen above.

File:[previous application.xlxs]

Segmented Univariate Analysis Target VS Age

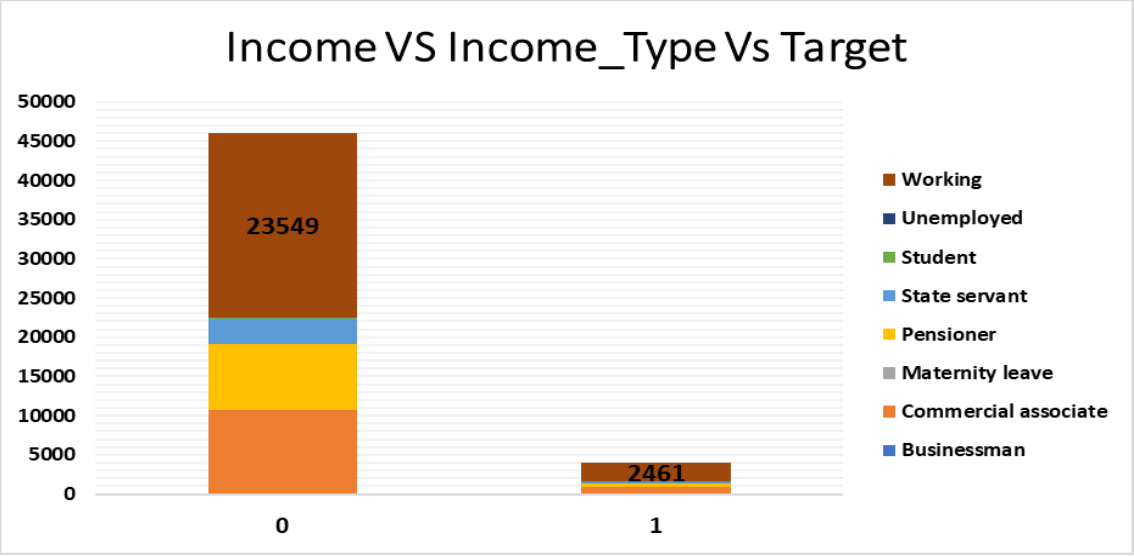


We can assume that the person aged between 31-40 has high tendency to pay the loan on time.

Bivariate Analysis:

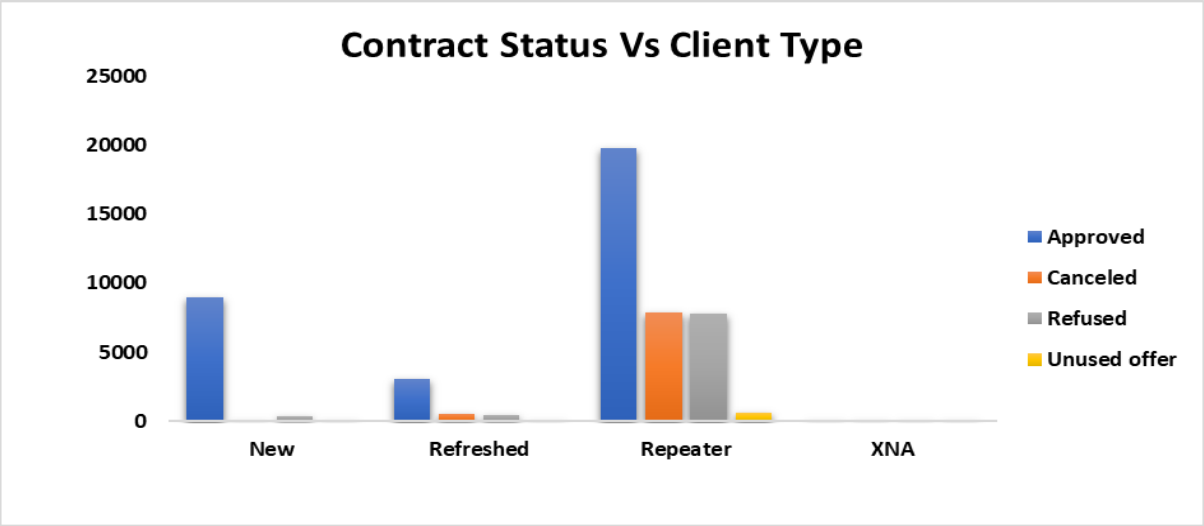
Bivariate analysis is an analysis of two variables to determine the relationships between them. They are often reported in quality of life research. It is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (it is often denoted as X, Y), for the purpose of determining the empirical relationship between them.

Bivariate Analysis Income Vs Target Vs Income_Type



People who are working seems to pay loan on time more than others.

Bivariate Analysis Contract Status Vs Target Vs Client_type



The Repeater has higher chances of getting the loan approve faster.

File:[previous application.xlxs]

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 colors or shading.

CNT_CHILDREN	1	0.036	0.006	0.026	0.002	-0.336	-0.119	-0.183	0.033
AMT_INCOME_TC	0.036	1	0.378	0.451	0.385	-0.074	-0.006	-0.069	-0.032
AMT_CREDIT	0.006	0.378	1	0.771	0.987	0.051	0.065	-0.008	0.009
AMT_ANNUITY	0.026	0.451	0.771	1	0.776	-0.010	0.027	-0.035	-0.009
AMT_GOODS_PRI	0.002	0.385	0.987	0.776	1	0.049	0.067	-0.011	0.010
DAYS_BIRTH	-0.336	-0.074	0.051	-0.010	0.049	1	0.408	0.335	0.269
DAYS_EMPLOYED	-0.119	-0.006	0.065	0.027	0.067	0.408	1	0.191	0.138
DAYS_REGISTRAT	-0.183	-0.069	-0.008	-0.035	-0.011	0.335	0.191	1	0.103
DAYS_ID_PUBLISH	0.033	-0.032	0.009	-0.009	0.010	0.269	0.138	0.103	1
	CNT_CHILDREN	AMT_INCOME	AMT_CRED	AMT_ANN	AMT_GO	DAYS_BI	DAYS_EM	DAYS_REGIS	DAYS_ID_PUBLI

Result:

- The person earning and having a income in between 25650 -1025650 has higher rate of taking the loan.
- The person having income in between 125650-225650 have higher rate of repaying the loan on time.
- The Repeater has higher chances of getting the loan approve faster.
- People who are working seems to pay loan on time more than others.
- The person aged between 31-40 has high tendency to pay the loan on time.
- The Repeater has higher chances of getting the loan approve faster.

Video : https://www.loom.com/share/07e19660ceb1497b9421768ed2b3dd8a?sid=dff577e5-f9e2-42f3-8ffd-ac75be22d09f

Application Dataset: https://docs.google.com/spreadsheets/d/1bw5X-hh7q-1fp8zDrjUP10RxZwODyE9c/edit?usp=sharing&ouid=113818516476537685883&rtpof=true&sd=true

Previous Application Dataset:
https://docs.google.com/spreadsheets/d/1_2Ydz1pqP1q54e3boHv3rAU8tUtmNLZs/edit?usp=drive_link&ouid=113818516476537685883&rtpof=true&sd=true

Thank you !!