Bank Loan Case Study

Project description:

This project is about the bank giving loan to the consumer and it aims to give an idea of applying EDA in a real business scenario and also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customer.

Approach:

In order to execute this project, I have used the EDA to analysis the pattern present in the data and also independently researched about the risk analysis. Find out the mean, median and mode in excel and find the missing value and outliers.

Tech stack used:

Microsoft excel is used to create the pivot table and to represent the graphs and, Microsoft word to present the report and google drive to submit the project.

Insights:

- **Problem statement:**

  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 company wants to understand the driving factors (or driver variables) behind the loan default. i.e., the variable which are strong in loan default.

- **Identify the missing data and use appropriate method to deal with it.**

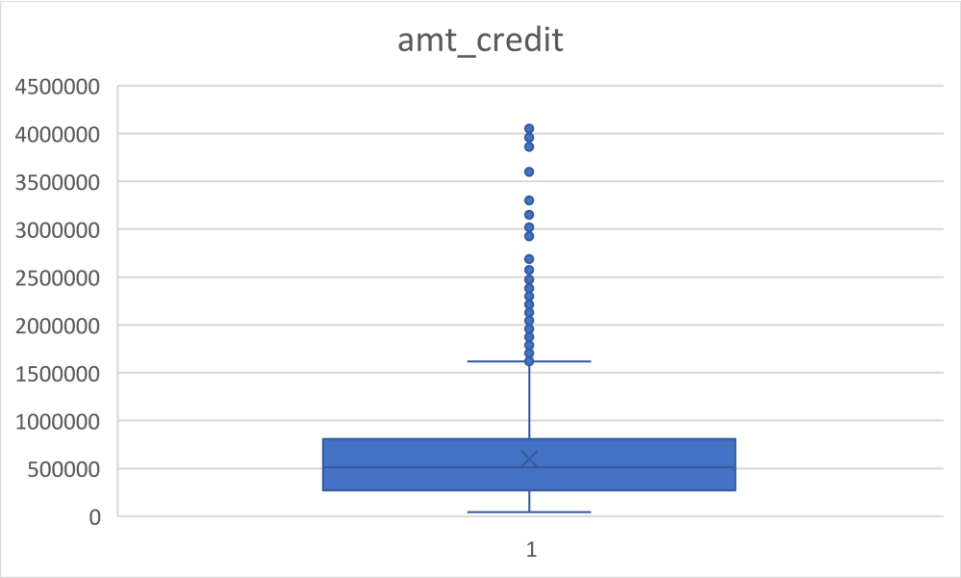
  In order to deal with the missing data, we have to find the value which is missing has any impact on the dataset, if not then we should remove it. If it does impact the we have to find the type of variable. If it is categorical data then mode, if not then we can use mean or median.

- **Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier.**

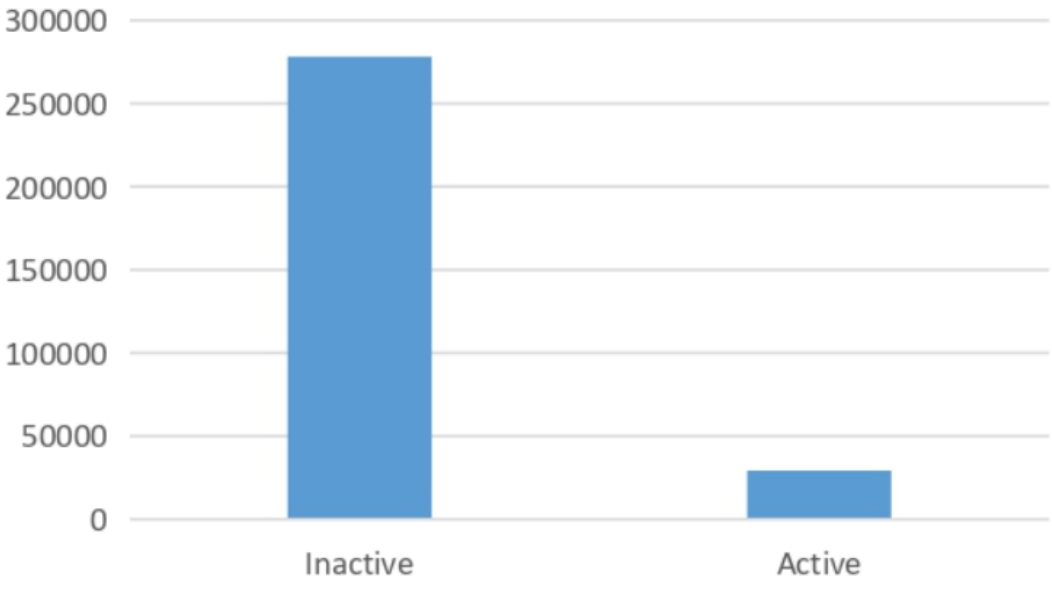
  Outliers: Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests.

  In applications.csv and previous_applications.csv we found the outliers and in columns_description.csv we did not get the outliers. Because the values above the 1.5IQR are considered as the outliers.

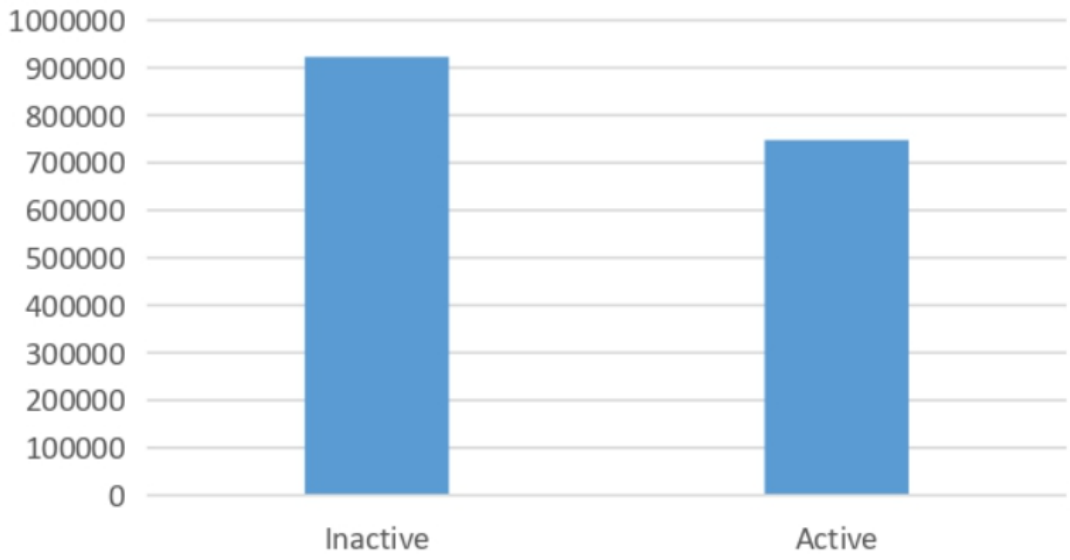
Q1	270000
Q2	513531
Q3	808650
MAX	4050000
MIN	45000
IQR	538650
Lower bound	-537975
upper bound	1616625



- **Identify if there is data imbalance in the data. Find the ratio of data imbalance.**
  In application.csv data, the percentage of data imbalance is 10.5 %, where the number of active and inactive variable are 278232 and 29279.



In previous_application.csv data, the percentage of data imbalance is 81.02%, where the number of active and unactive variables are 922661 and 747553.



- **Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms**

Univariate analysis:

It is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a univariate analysis technique is used.

Segmented univariate analysis:

It is one of the simplest forms of visualization to analyse data. In its name 'uni' means one which itself describes that it considers only a single data variable for analysis. The data variable is analysed in subsets and is very useful as it can show the change metric in pattern across the different segments of the same variable.

Bivariate analysis:

It is slightly more analytical than univariate analysis, when the data set contains two variables and researchers aim to undertake comparisons between the two data set then bivariate analysis is the right type of analysis technique.

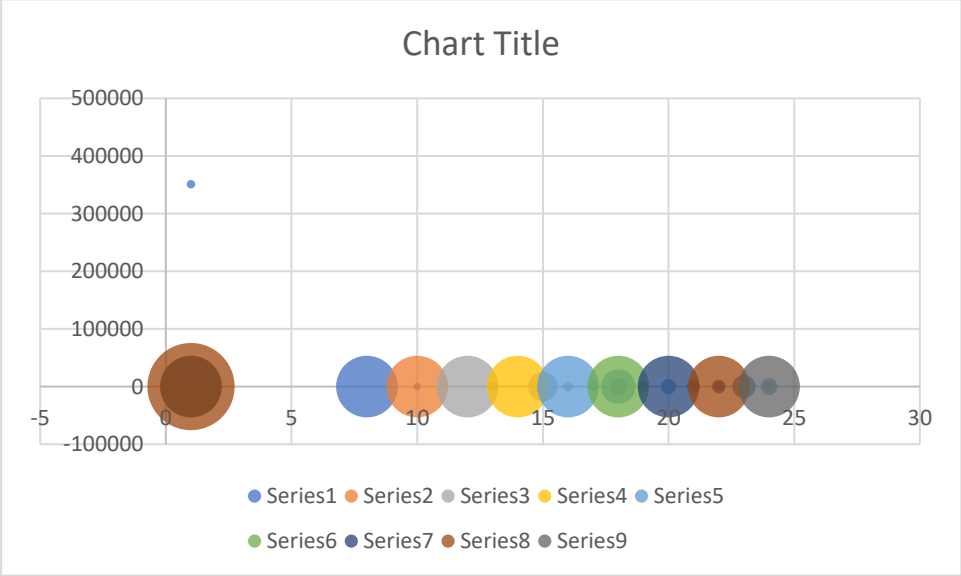
This study explores the relationship of two variables as well as the depth of this relationship is to figure out if there are any discrepancies between two variables and any causes of this difference, we have conducted correlation analysis on the basis of our dataset.

- Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing
- Include visualizations and summarize **t**he most important results in the presentation.

Result: When the target variable is 0:

351000	*0.0188*	*-9461*	*-637*	*-3648*	*-2120*		*1*	*1*	*0*

1							
0.076049	1						
-0.13579	-0.04819	1					
0.003581	0.015102	-0.5751	1				
-0.02567	-0.05622	0.289111	-0.18893	1			
-0.05608	-0.01554	0.252851	-0.22647	0.096829	1		
-0.07806	-0.05163	-0.00174	0.023443	-0.02401	0.017019	1	
#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1

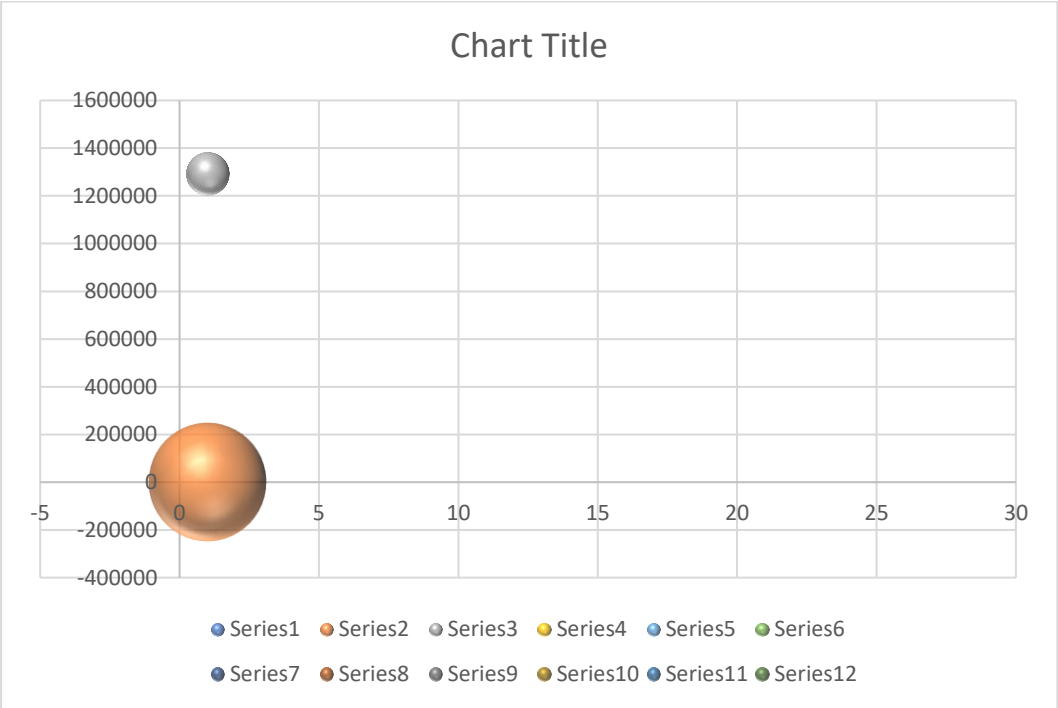


When the target variable is 1:

270000	*1293503*	*35698.5*	*1129500*	*0.00354*	*-16765*	*-1188*	*-1186*	*-291*

1							
0.38152	1						
0.454034	0.773828	1					
0.388084	0.987078	0.779252	1				
0.182193	0.094012	0.118997	0.097709	1			
0.072285	-0.05359	0.010164	-0.05078	-0.03115	1		
-0.15795	-0.07415	-0.1122	-0.07228	-0.01162	-0.61693	1	
0.059985	0.009999	0.035059	0.014097	-0.06822	0.337731	-0.20723	1

0.028533	-0.01024	0.006582	-0.01088	-0.00311	0.26632	-0.27054	0.104556		1
-0.14995	-0.09469	-0.09756	-0.1024	-0.08953	0.012863	0.025538	-0.02328	0.007935	1



Result:

From doing this project I have the gained the knowledge of EDA, risk analysis, and came to know how to work with the data and outliers, univariate and bivariate analysis, correlation data analysis.