NAME : ARIN KUMAR **DATE : 20/3/2023**

PROJECT-6

BANK LOAN CASE STUDY

Project Description:

In this project, we will be analyzing two datasets - application_data.csv and previous_application.csv to identify if a client has payment difficulties and if there are any factors affecting this. We will be performing exploratory data analysis to understand the data and draw insights from it. We will also be identifying missing values, outliers, and data imbalances in the data and taking appropriate steps to handle them.

Approach:

The approach for this analysis will involve the following steps:

1. Data Understanding: Understanding the data, its structure, and variables.
2. Data Cleaning: Identifying missing values and outliers and replacing them with appropriate methods.
3. Data Exploration: Exploring the data through univariate, segmented univariate, and bivariate analysis.
4. Correlation Analysis: Identifying the top 10 correlations for clients with payment difficulties and all other cases.
5. Visualization and Insights: Presenting the most important results through visualizations and summarizing the insights.

Tech-Stack Used:

For this project, we will be using Excel to perform exploratory data analysis and draw insights from the data.

Insights:

During our exploratory data analysis, we observed that the application_data.csv file contains information about 307,511 clients and 122 variables, while the previous_application.csv file contains information about 1,670,214 previous loan applications and 37 variables. We observed missing values in several columns in both datasets and used various methods to handle them. We also observed outliers in some columns, which we did not remove as they seemed valid and could provide valuable insights. We identified data imbalance in the target variable, with only 8.07% of clients having payment difficulties.

In our bivariate analysis, we identified several variables that were strongly correlated with payment difficulties, including the number of days before the application when the client changed his registration, the number of days before the application when the client's ID document was changed, and the number of days before the application when the client registered his phone number.

A. Identify if there are **outliers** in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

1	DAYS_EMPLOYED	OUTLIER	25%	75%	IQR	UPPER BOUND	LOWER BOUND
10	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
13	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
20	-7804	TRUE	-2760	-289	2471	3417.5	-6466.5
25	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
40	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
45	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
48	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
51	-9523	TRUE	-2760	-289	2471	3417.5	-6466.5
53	-6977	TRUE	-2760	-289	2471	3417.5	-6466.5
56	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
58	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
64	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
81	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
83	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
86	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
92	-8862	TRUE	-2760	-289	2471	3417.5	-6466.5
97	-7980	TRUE	-2760	-289	2471	3417.5	-6466.5
100	-6737	TRUE	-2760	-289	2471	3417.5	-6466.5
101	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
106	-8466	TRUE	-2760	-289	2471	3417.5	-6466.5
107	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
108	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
110	365243	TRUE	-2760	-289	2471	3417.5	-6466.5
119	365243	TRUE	-2760	-289	2471	3417.5	-6466.5

These are the outliers present in the DAYS_EMPLOYED column because this values has exceeded the upper bound or the values was less than the lower bound.

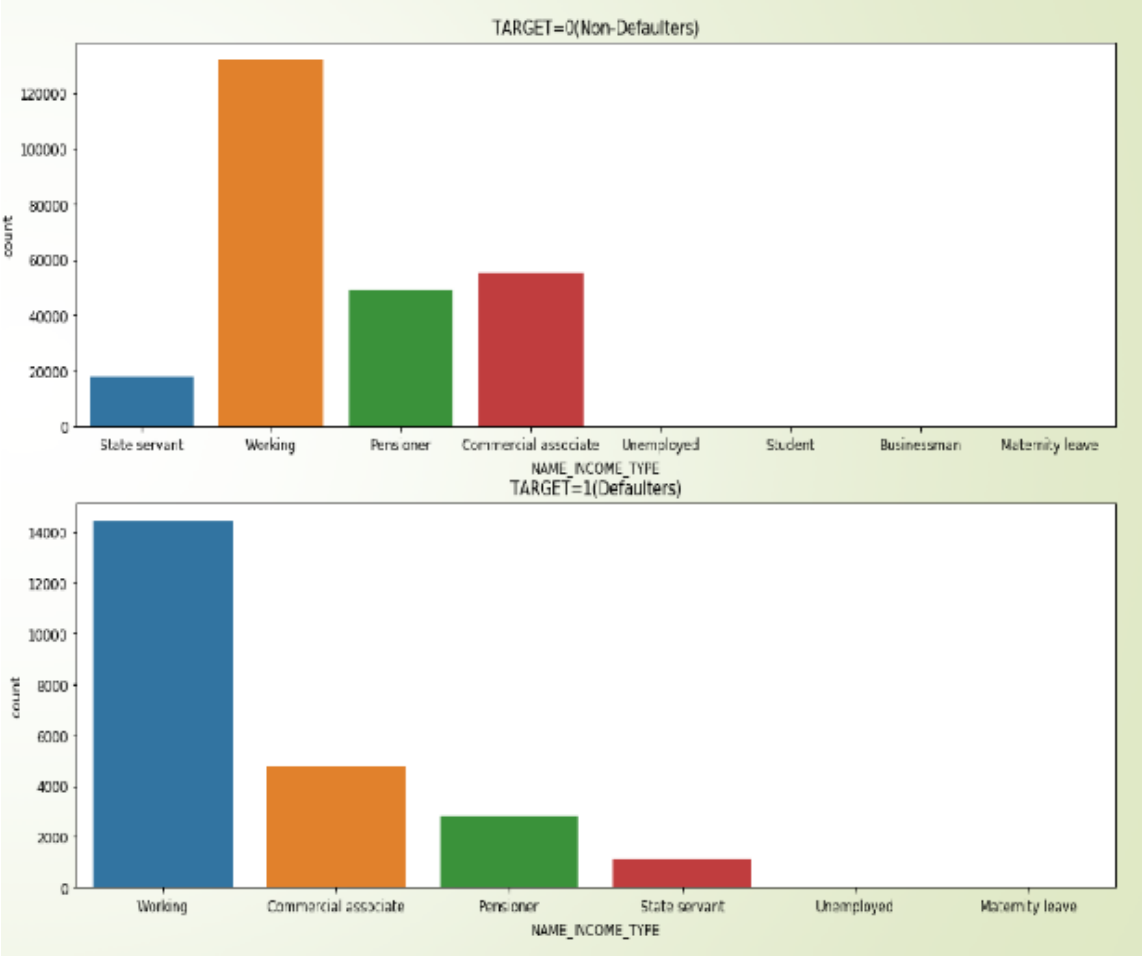
B. Identify if there is data imbalance in the data. Find the ratio of data imbalance.

3	Row Labels	Count of TARGET
4	0	282686
5	1	24825
6	(blank)	
7	**Grand Total**	**307511**
8		
9	DATA IMBALANCE	0.087818286

C. Explain the results of univariate, bivariate analysis, etc. in business terms.

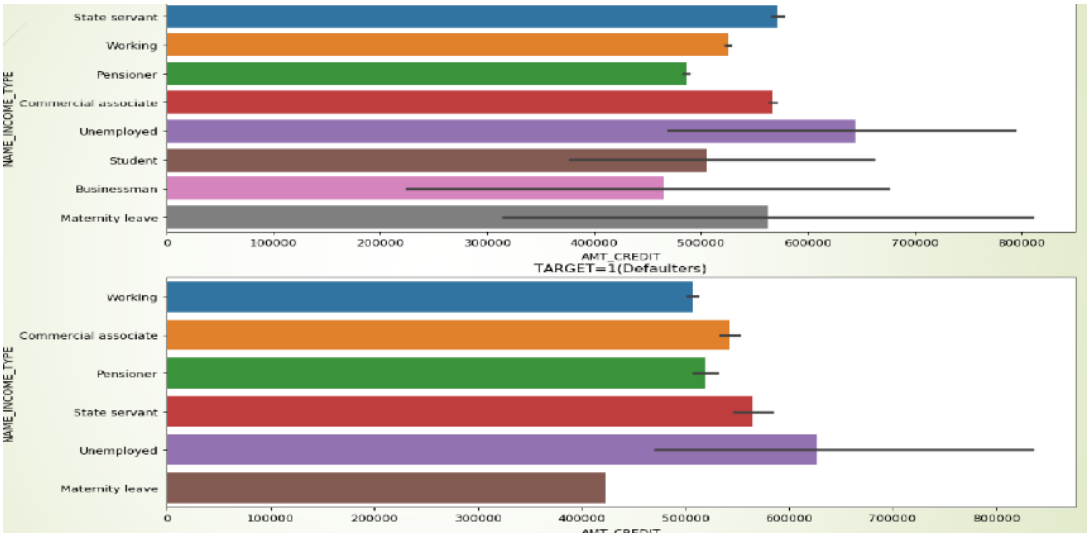
UNIVARIATE ANALYSIS

Target variable for defaulters and non defaulters

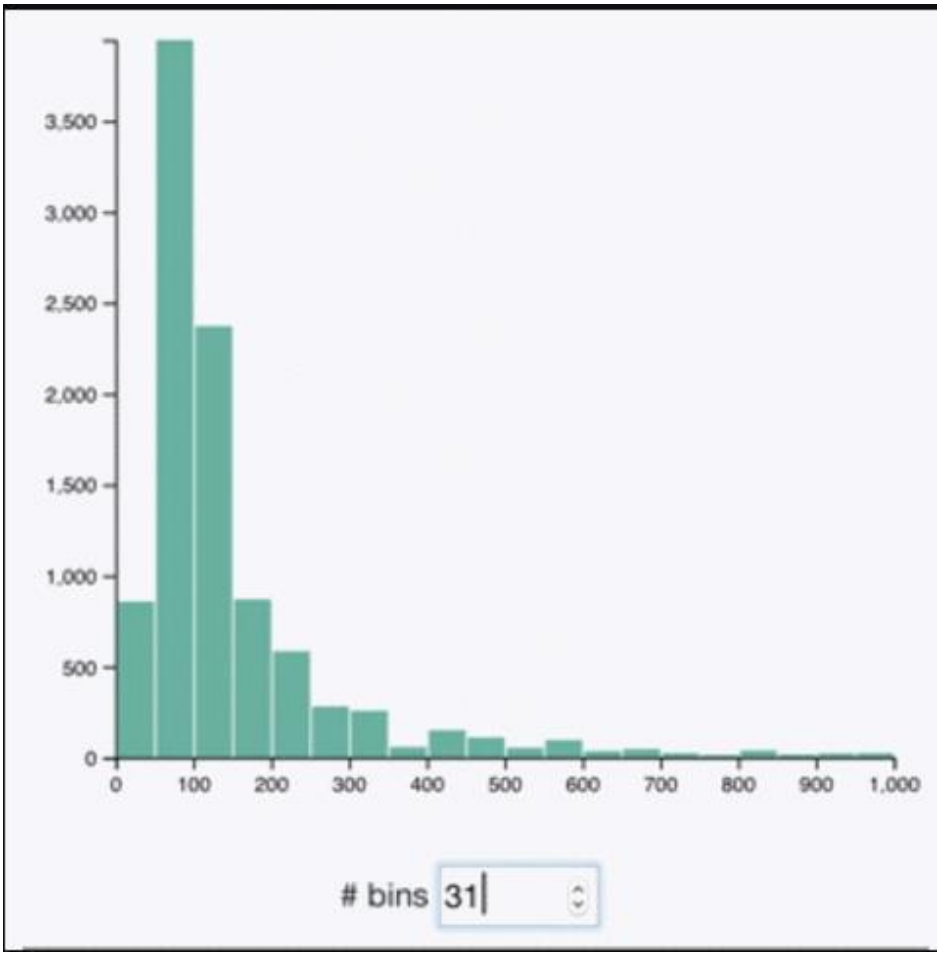


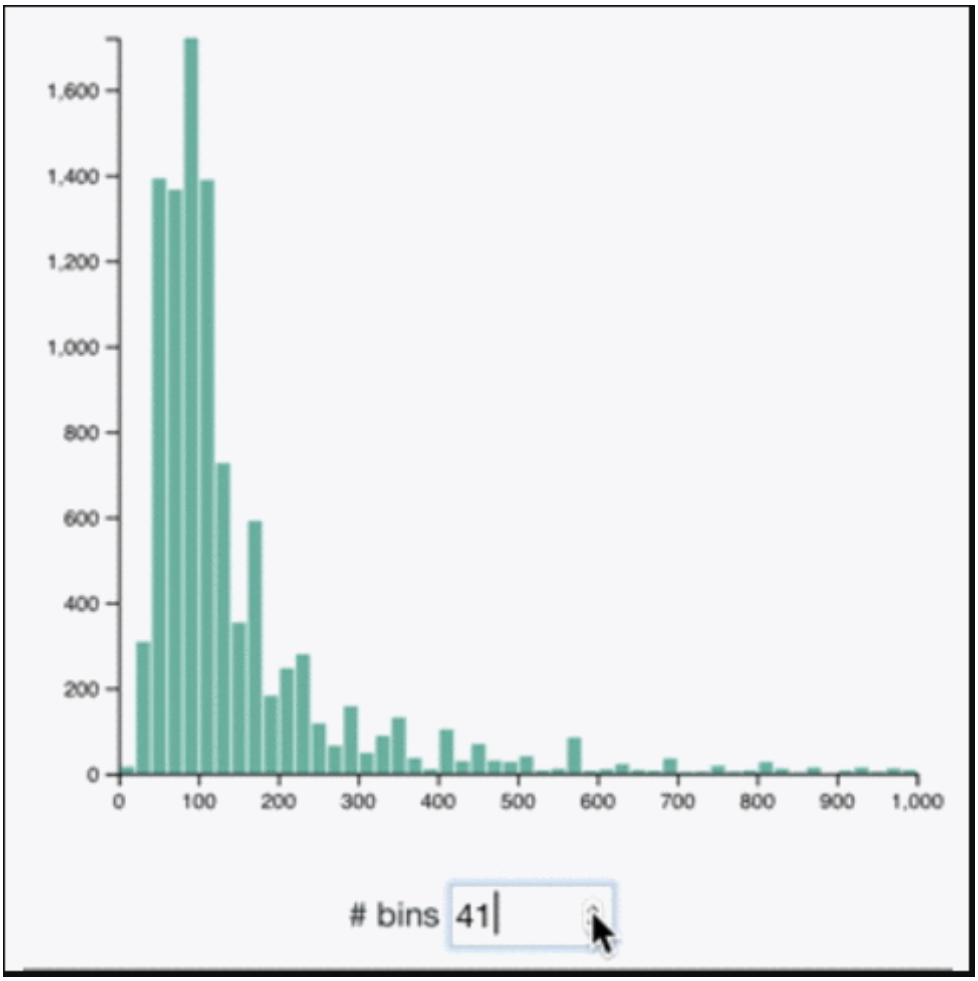
BIVARIATE ANALYSIS

Name_income_type vs amt_credit



D. 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.





Result:

Through this project, we were able to gain valuable insights into the factors affecting payment difficulties for clients. We identified several variables that were strongly correlated with payment difficulties and could be used to predict if a client is likely to have payment difficulties. This information can be used by the company to identify high-risk clients and take appropriate steps to mitigate risk.