



BANK LOAN CASE STUDY

By Abhishek Jagtap

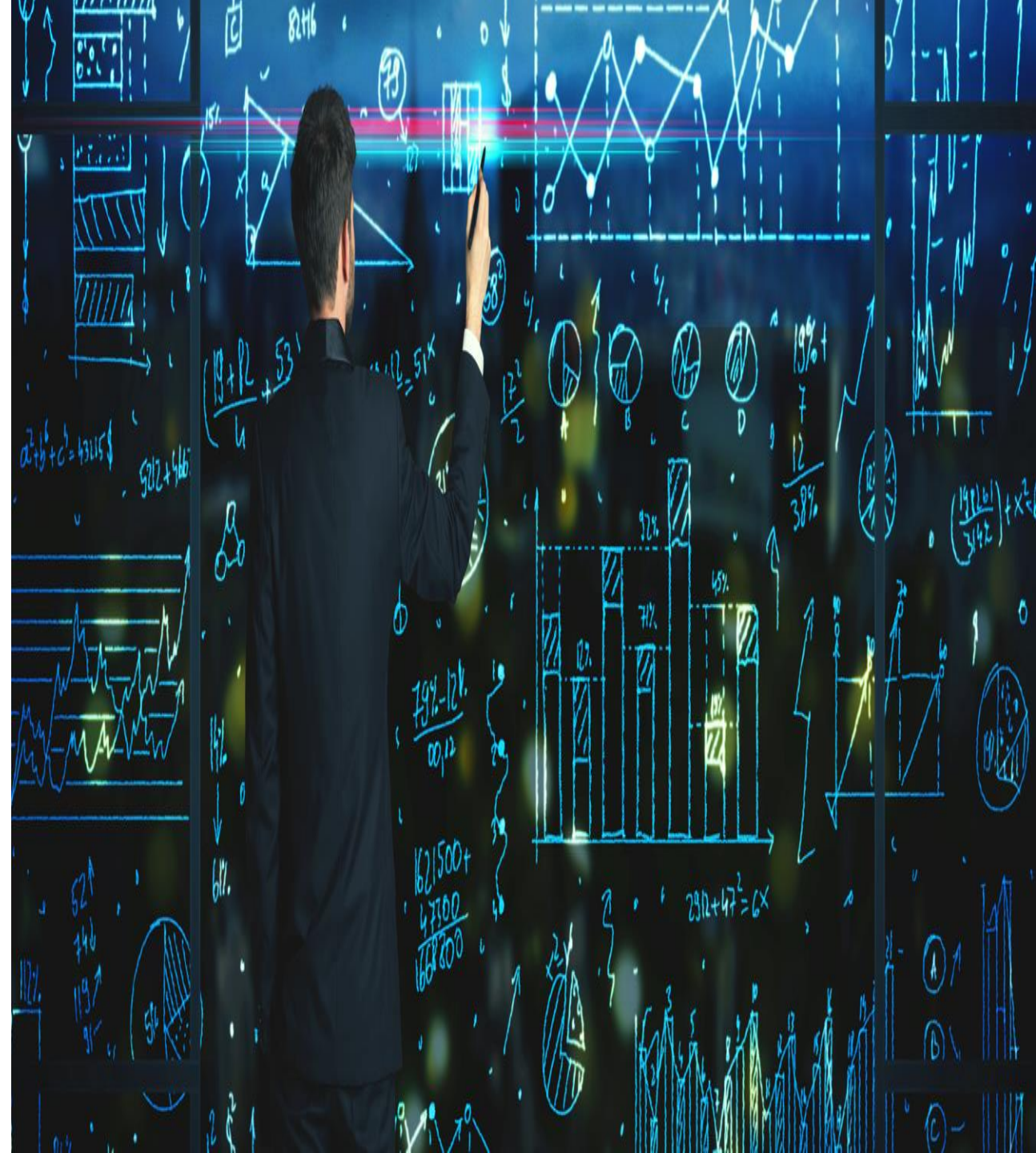
PROJECT DESCRIPTION

This case study aims to give you an idea of applying EDA in a real business scenario. You will 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 also by using EDA to analyze the patterns present in the data.



APPROACH

In this case study, there were two extensive datasets: application_data and previous_application. These datasets contained numerous unnecessary columns having high number of null values. The initial step involved cleaning the data, deleting the column with high percent of null values then eliminating outliers, and replacing missing values. Following the data cleaning process, univariate analysis, segmented univariate analysis, and bivariate analysis were conducted using pivot tables and charts. Additionally, the top 10 correlations were identified.



TECH STACK USED

MS excel was used for analysis as excel offers a grid layout where data can be organized in rows and columns, making it easy to input and manipulate information.

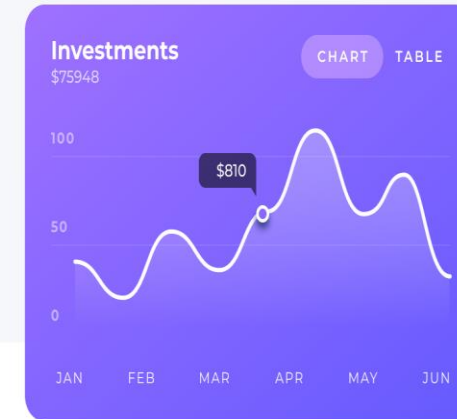
Link of the analyzed data – [Bank loan case study](#)



INSIGHTS

The objective of this case study is to conduct Exploratory Data Analysis (EDA) in order to identify patterns that can indicate if a client will have difficulty repaying their loan installments. The key insights was performing univariate, Segmented Univariate, and Bivariate Analysis which helps to finding different patterns.

Loan Case Study



AutoSave Off application_data Saved to this PC Search

File Home Insert Page Layout Formulas Data Review View Automate Help

Paste Font Alignment Number Styles

Calibri 11 Wrap Text General Conditional Formatting Format as Table

B I U Merge & Center

Clipboard Font Alignment Number Styles

0012 X ✓ fx

	A	B	C	D	E	F	G	H	I	J	K	L	M
88		157861	1	Cash loan M	Y	N		0	135000	270000	21330	270000	Unac
89		157862	0	Cash loan M	Y	Y		1	90000	900000	26446.5	900000	Unac
90		157863	0	Revolving F	Y	Y		0	162000	540000	27000	540000	Fami
91		157865	0	Revolving F	N	Y		0	135000	270000	13500	270000	Unac
92		157867	0	Revolving F	N	Y		0	112500	180000	9000	180000	Unac
93		157868	0	Cash loan F	Y	Y		0	135000	1078200	31653	900000	Unac
94		157869	0	Revolving F	N	Y		3	58500	157500	7875	157500	Unac
95		157870	0	Revolving F	N	N		0	202500	450000	22500	450000	Unac
96		157871	0	Cash loan F	N	N		0	180000	1206000	45936	1206000	
97		157872	0	Cash loan M	N	N		0	126000	1125000	47794.5	1125000	Unac
98		157873	0	Cash loan M	N	N		1	112500	900000	26316	900000	Unac
99		157874	0	Cash loan F	N	Y		0	270000	820638	34897.5	733500	Fami
00		157875	0	Cash loan F	N	Y		0	117000	254700	14751	225000	Unac
01									145800			450000	
02													
03													
04	null values	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
05	median values to replace	129076	0	#NUM!	#NUM!	#NUM!	#NUM!	0	145800	514777.5	24939	450000	
06													
07													
08													
09													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													

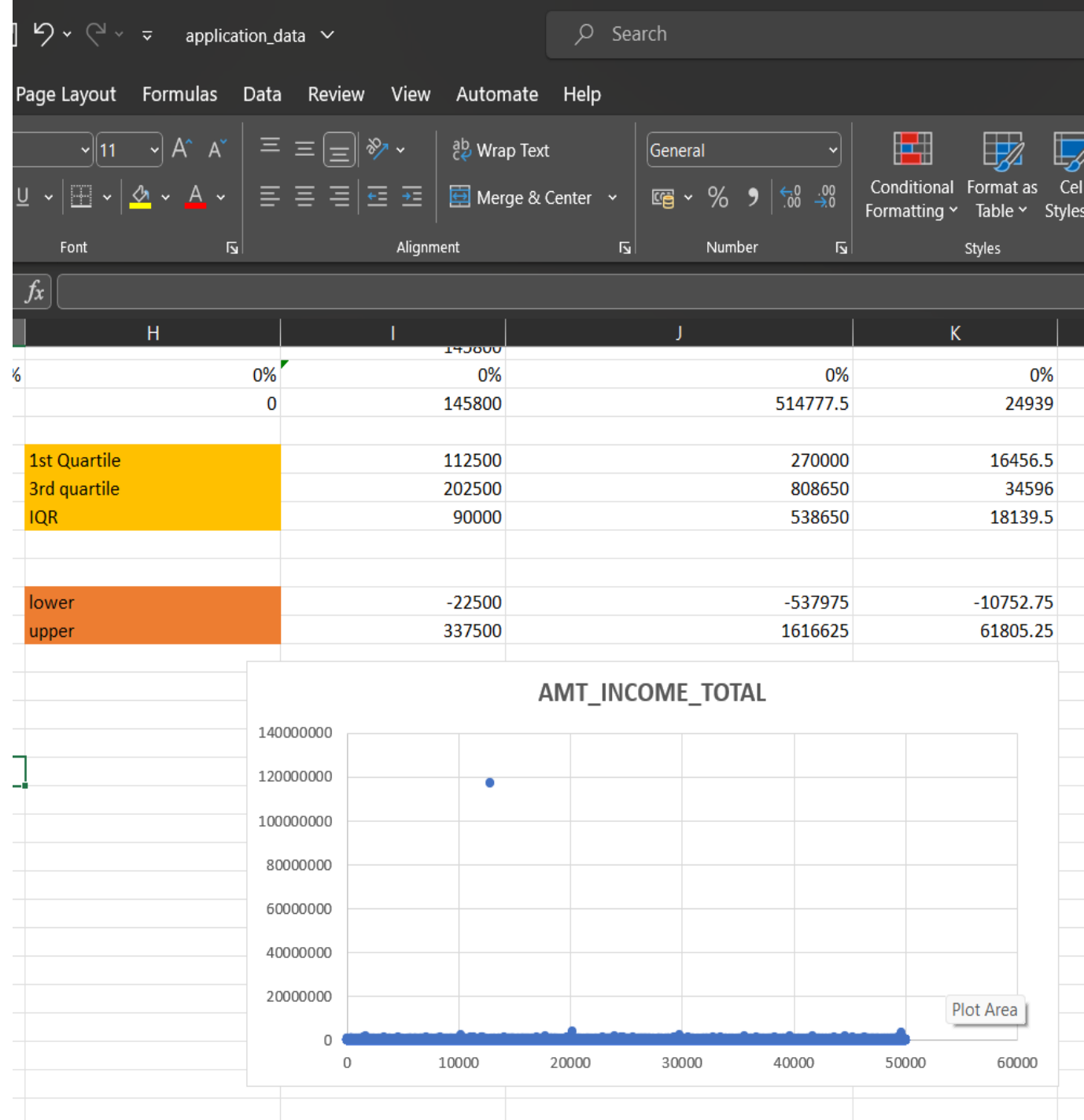
application_data outliers data imbalance univariate analysis segmented univariate bivariate analy ... + :

B. Identify Outliers in the Dataset

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

Ans:- Outliers being identified by Quartile function and marked with conditional formatting also the scatter chart was created to visualize outlier value

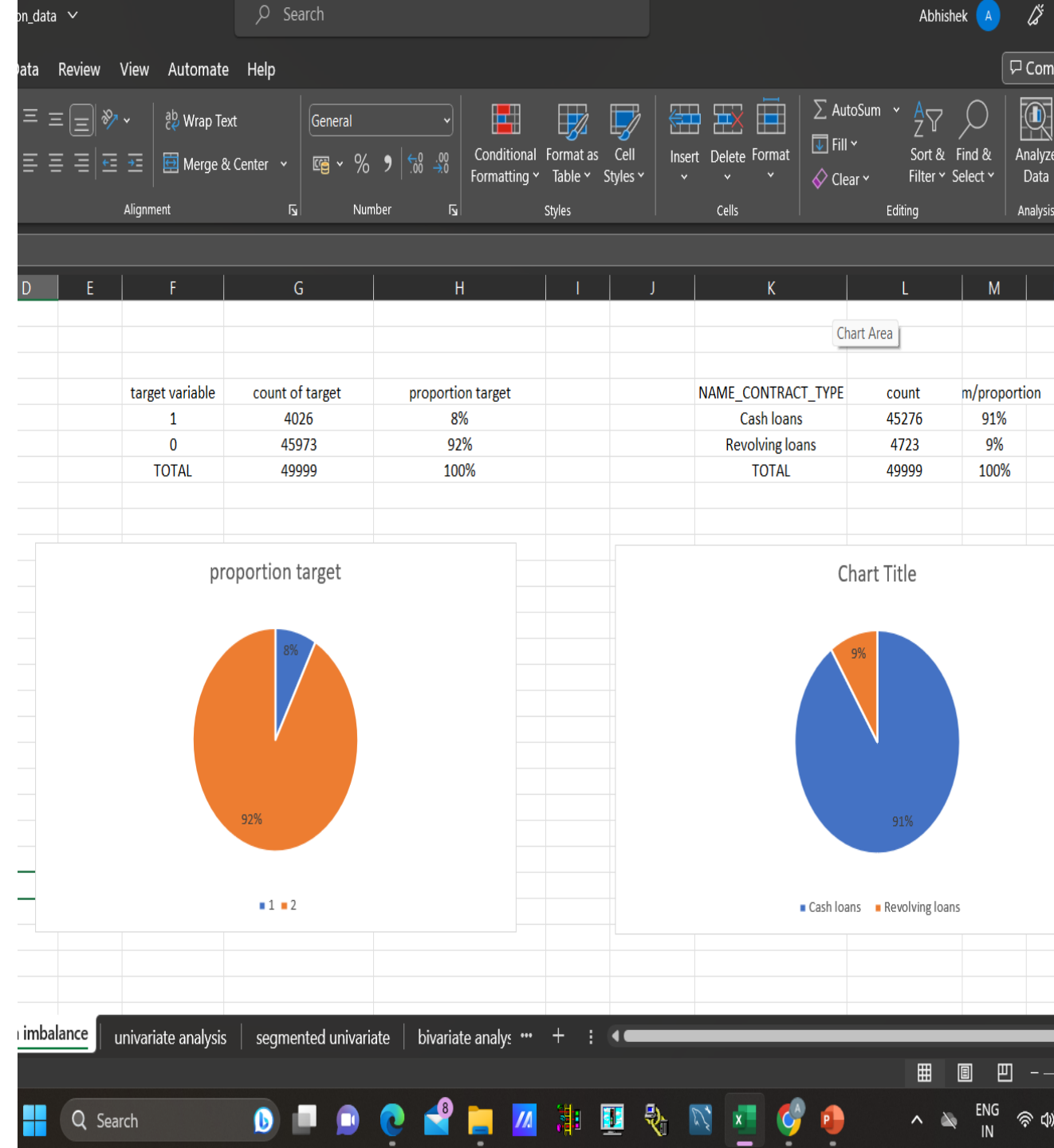
The outliers value for income can be valid as income can be high but for days employed, days birth value can't be high.



C. Analyse Data Imbalance

Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Ans:- Data imbalance for target and contract type is calculated with count formula and proportion for the same is shown.

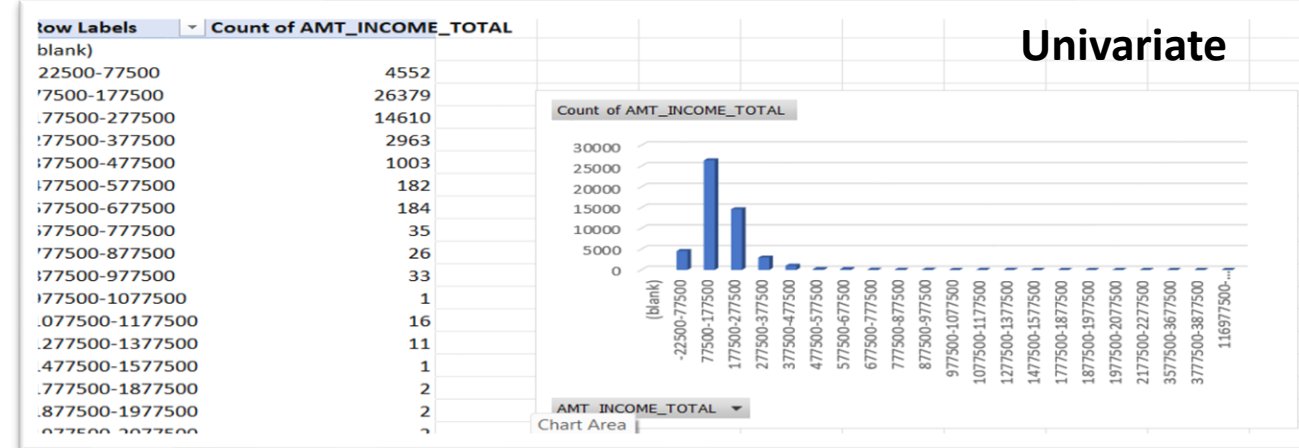


D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

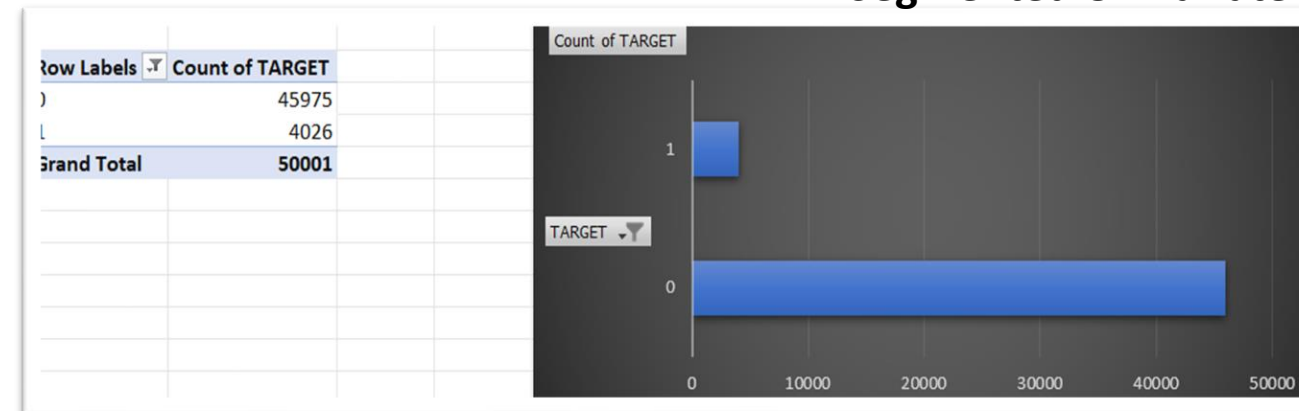
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Ans:- Univariate, Segmented Univariate and Bivariate Analysis was performed on different variables for understanding the pattern

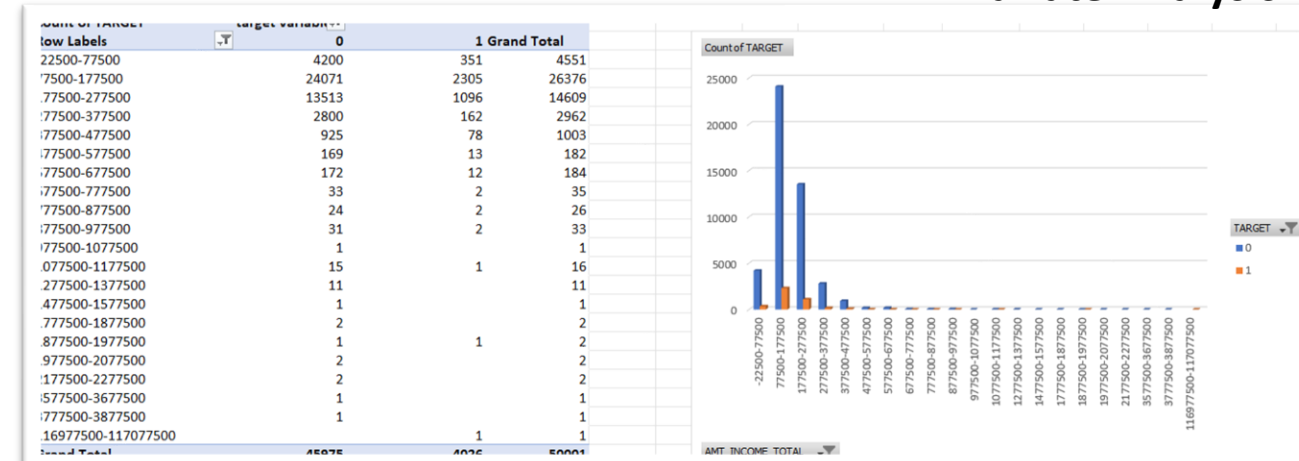
Univariate



Segmented Univariate



Bivariate Analysis



E. Identify Top Correlations for Different Scenarios

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.

Ans:- Correlation between variables and the target variable within each segment was calculated and correlation matrix was implemented. The top rank for correlations for different segments is shown.

	SK_ID_CURR	TARGET	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	CNT_FAM_MEMBERS	CNT_CHILDREN	AMT_CREDIT	AMT_GOODS_PRICE	AMT_ANNUITY	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH
SK_ID_CURR	1												
TARGET	0.003295	1											
REGION_RATING_CLIENT	0.004509	0.066130148	1										
REGION_RATING_CLIENT_W_CITY	0.003209	0.067079294	0.950710179	1									
CNT_FAM_MEMBERS	0.001954	0.012992443	0.025985049	0.025164908	1								
CNT_CHILDREN	0.005538	0.026363931	0.025913889	0.022777663	0.880454	1							
AMT_CREDIT	-0.00073	-0.032428347	-0.100507425	-0.109486833	0.063998	0.004972	1						
AMT_GOODS_PRICE	-0.00074	-0.04127611	-0.103635237	-0.111707308	0.061572	0.000233	0.986704	1					
AMT_ANNUITY	-0.00208	-0.012399094	-0.125802654	-0.139321767	0.077379	0.026179	0.769499	0.774134	1				
DAYS_BIRTH	0.001324	0.076787685	0.016779196	0.014551531	0.277246	0.329264	-0.05934	-0.05767	0.007712	1			
DAYS_EMPLOYED	-0.00437	-0.040294905	0.034321673	0.036829676	-0.22982	-0.23969	-0.07047	-0.06779	-0.11045	-0.61355	1		
DAYS_REGISTRATION	-0.00371	0.042342679	0.087517643	0.079791927	0.170111	0.181217	0.003449	0.006084	0.033219	0.333633	-0.20468	1	
DAYS_ID_PUBLISH	-0.00874	0.046926745	-0.002307011	-0.007312572	-0.02608	-0.03212	-0.01223	-0.01403	0.006716	0.270825	-0.27038	0.104299	1

TOP Correlations		
var1	var2	correl
AMT_GOODS_PRICE	AMT_CREDIT	0.986704386
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950710179
CNT_CHILDREN	CNT_FAM_MEMBERS	0.880454498
AMT_ANNUITY	AMT_GOODS_PRICE	0.774133945
AMT_ANNUITY	AMT_CREDIT	0.769498914

The background is a black field filled with a complex pattern of orange and blue dots, lines, and rectangles. The dots are arranged in various clusters and streams. The lines are thin and wavy, some in orange and some in blue. The rectangles are thin and outlined, some in orange and some in blue, arranged in a somewhat regular pattern. The overall effect is a dense, abstract, and colorful composition.

THANK YOU !