Bank Loan Case Study

Final Project-2

Description: Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Approach:

1. **Understand the data**
   Understand the data to make a plan or analyze given data and make meaningful insights from it.

2. **Data Pre-Processing**
   Removing or Handling Null Values and as well as Outliers from the data.

3. **Data Analysis**
   Deriving conclusion by Analyze the data set

4. **Data Visualization**
   Plotting a data or insights from the data

5. **Results**
   Give a result from the Analysis

Tech-Stack Used:

Microsoft Excel 2019 - Easy to use and Very Powerful.

Insights / Data Analytics Task:

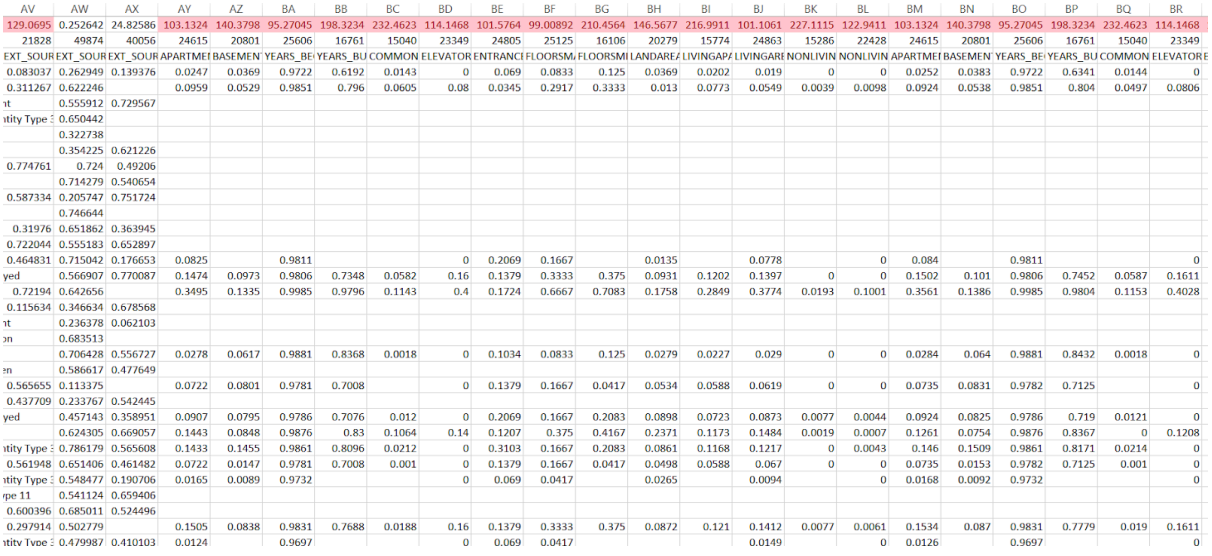
A. Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

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.

Total Rows = 50000

Total Columns = 122

- ➔ Find Percentage of Null values and delete having less than (<) 30% rows.
- ➔ Deleted Rows Highlighted with Light Red Color in Dataset.
- ➔ For Example…



AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR
129.0695	0.252642	24.82586	103.1324	140.3798	95.27045	198.3234	232.4623	114.1468	101.5764	99.00892	210.4564	146.5677	216.9911	101.1061	227.1115	122.9411	103.1324	140.3798	95.27045	198.3234	232.4623	114.1468
21828	49874	40056	24615	20801	25606	16761	15040	23349	24805	25125	16106	20279	15774	24863	15286	22428	24615	20801	25606	16761	15040	23349
EXT_SOUR	EXT_SOUR	EXT_SOUR	APARTMEI	BASEMEN	YEARS_BE	YEARS_BU	COMMON	ELEVATOR	ENTRANCI	FLOORSMA	FLOORSMI	LANDARE/	LIVINGAP/	LIVINGARE	NONLIVIN	NONLIVIN	APARTMEI	BASEMEN	YEARS_BE	YEARS_BU	COMMON	ELEVATOR
0.083037	0.262949	0.139376	0.0247	0.0369	0.9722	0.6192	0.0143	0	0.069	0.0833	0.125	0.0369	0.0202	0.019	0	0	0.0252	0.0383	0.9722	0.6341	0.0144	0
0.311267	0.622246		0.0959	0.0529	0.9851	0.796	0.0605	0.08	0.0345	0.2917	0.3333	0.013	0.0773	0.0549	0.0039	0.0098	0.0924	0.0538	0.9851	0.804	0.0497	0.0806
nt	0.555912	0.729567																				
ıtity Type	0.650442																					
	0.322738																					
	0.354225	0.621226																				
0.774761	0.724	0.49206																				
	0.714279	0.540654																				
0.587334	0.205747	0.751724																				
	0.746644																					
0.31976	0.651862	0.363945																				
0.722044	0.555183	0.652897																				
0.464831	0.715042	0.176653	0.0825		0.9811			0	0.2069	0.1667		0.0135		0.0778		0	0.084		0.9811			0
yed	0.566907	0.770087	0.1474	0.0973	0.9806	0.7348	0.0582	0.16	0.1379	0.3333	0.375	0.0931	0.1202	0.1397	0	0	0.1502	0.101	0.9806	0.7452	0.0587	0.1611
0.72194	0.642656		0.3495	0.1335	0.9985	0.9796	0.1143	0.4	0.1724	0.6667	0.7083	0.1758	0.2849	0.3774	0.0193	0.1001	0.3561	0.1386	0.9985	0.9804	0.1153	0.4028
0.115634	0.346634	0.678568																				
nt	0.236378	0.062103																				
on	0.683513																					
	0.706428	0.556727	0.0278	0.0617	0.9881	0.8368	0.0018	0	0.1034	0.0833	0.125	0.0279	0.0227	0.029	0	0	0.0284	0.064	0.9881	0.8432	0.0018	0
en	0.586617	0.477649																				
0.565655	0.113375		0.0722	0.0801	0.9781	0.7008		0	0.1379	0.1667	0.0417	0.0534	0.0588	0.0619	0	0	0.0735	0.0831	0.9782	0.7125		0
0.437709	0.233767	0.542445																				
yed	0.457143	0.358951	0.0907	0.0795	0.9786	0.7076	0.012	0	0.2069	0.1667	0.2083	0.0898	0.0723	0.0873	0.0077	0.0044	0.0924	0.0825	0.9786	0.719	0.0121	0
	0.624305	0.669057	0.1443	0.0848	0.9876	0.83	0.1064	0.14	0.1207	0.375	0.4167	0.2371	0.1173	0.1484	0.0019	0.0007	0.1261	0.0754	0.9876	0.8367	0	0.1208
ıtity Type	0.786179	0.565608	0.1433	0.1455	0.9861	0.8096	0.0212	0	0.3103	0.1667	0.2083	0.0861	0.1168	0.1217	0	0.0043	0.146	0.1509	0.9861	0.8171	0.0214	0
0.561948	0.651406	0.461482	0.0722	0.0147	0.9781	0.7008	0.001	0	0.1379	0.1667	0.0417	0.0498	0.0588	0.067	0	0	0.0735	0.0153	0.9782	0.7125	0.001	0
ıtity Type	0.548477	0.190706	0.0165	0.0089	0.9732			0	0.069	0.0417		0.0265		0.0094		0	0.0168	0.0092	0.9732			0
ype 11	0.541124	0.659406																				
0.600396	0.685011	0.524496																				
0.297914	0.502779		0.1505	0.0838	0.9831	0.7688	0.0188	0.16	0.1379	0.3333	0.375	0.0872	0.121	0.1412	0.0077	0.0061	0.1534	0.087	0.9831	0.7779	0.019	0.1611
ıtity Type	0.479987	0.410103	0.0124		0.9697			0	0.069	0.0417				0.0149		0	0.0126		0.9697			0

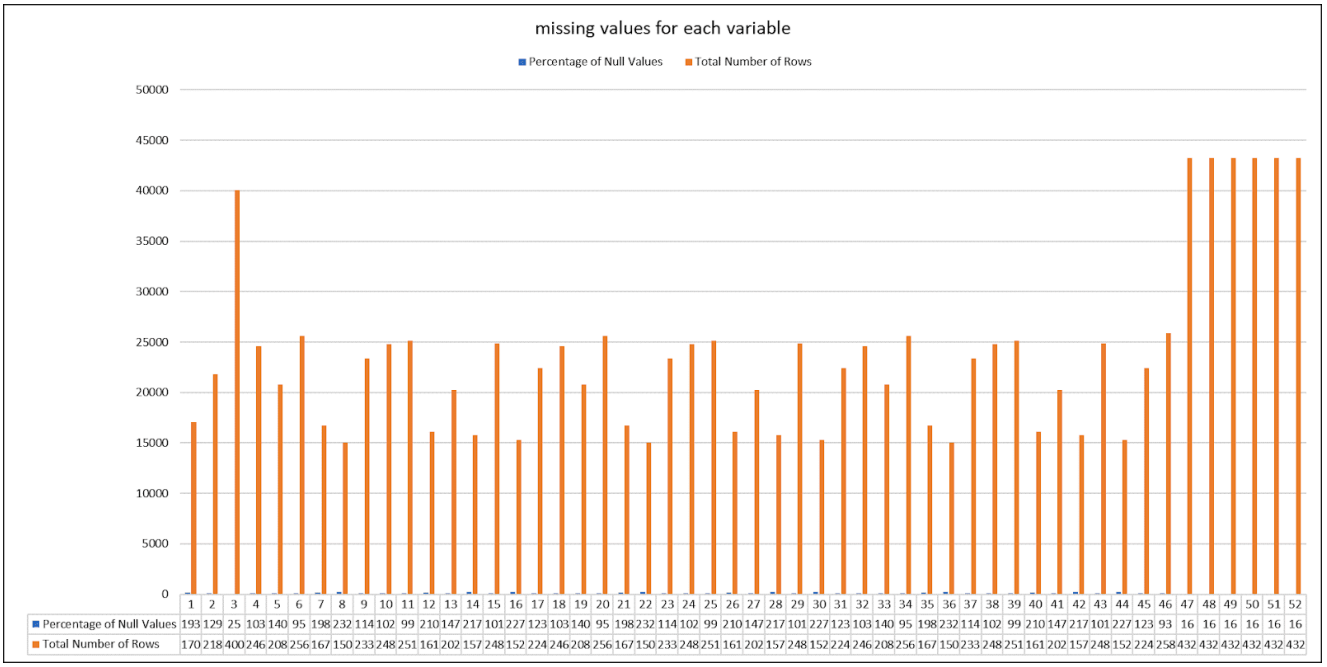
- ➔ After Deleting those Rows(Less than 30%)

Total Column=122-45 Remaining Column = 77

Formula:(To Find Null Values)

=IF(COUNT(B4:B50002)=0,0,
(COUNTBLANK(B4:B50002)/COUNT(B4:B50002))*100)

➔ chart to visualize the proportion of missing values for each variable.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
Percentage of Null Values	193	129	25	103	140	95	198	232	114	102	99	210	147	217	101	227	123	103	140	95	198	232	114	102	99	210	147	217	101	227	123	103	140	95	198	232	114	102	99	210	147	217	101	227	123	93	16	16	16	16	16	16
Total Number of Rows	170	218	400	246	208	256	167	150	233	248	251	161	202	157	248	152	224	246	208	256	167	150	233	248	251	161	202	157	248	152	224	246	208	256	167	150	233	248	251	161	202	157	248	152	224	258	432	432	432	432	432	432

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: Detects and identifies outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

➔ Finding Outliers in AMT_INCOME_TOTAL column using TARGET column.

Quartile 1 112500	Inter Quartile Range 90000
Quartile 2 145800	Upper Limit 337500
Quartile 3 202500	Lower Limit -22500

➔ Formula:

Quartile 1 =QUARTILE.INC(B:B,1)

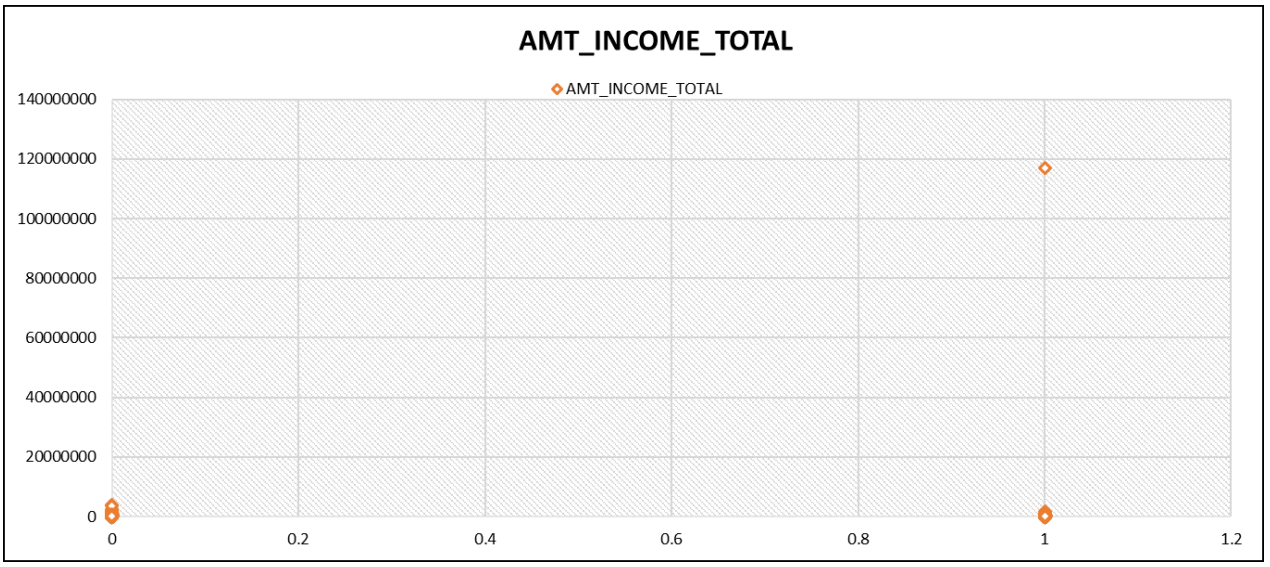
Quartile 2 =QUARTILE.INC(B:B,2)

Quartile 3 =QUARTILE.INC(B:B,3)

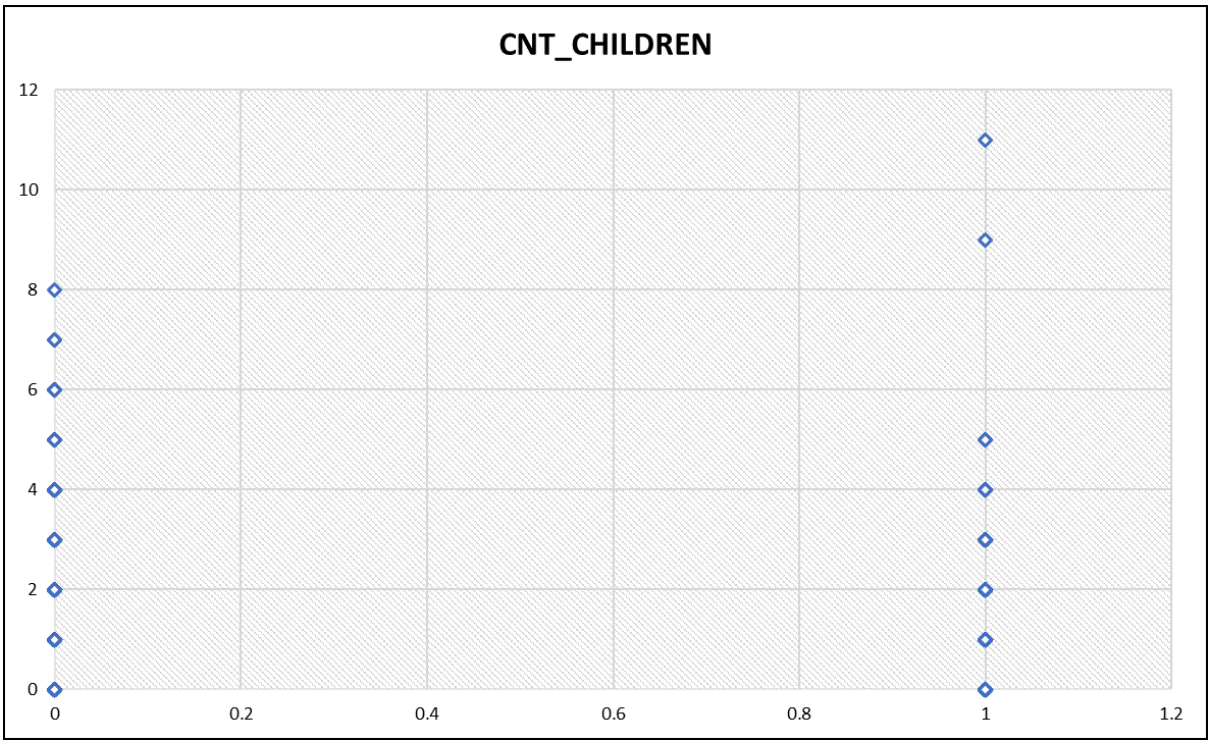
Inter Qua Range =D10-D4

Upper Limit =D10+(1.5*G4)

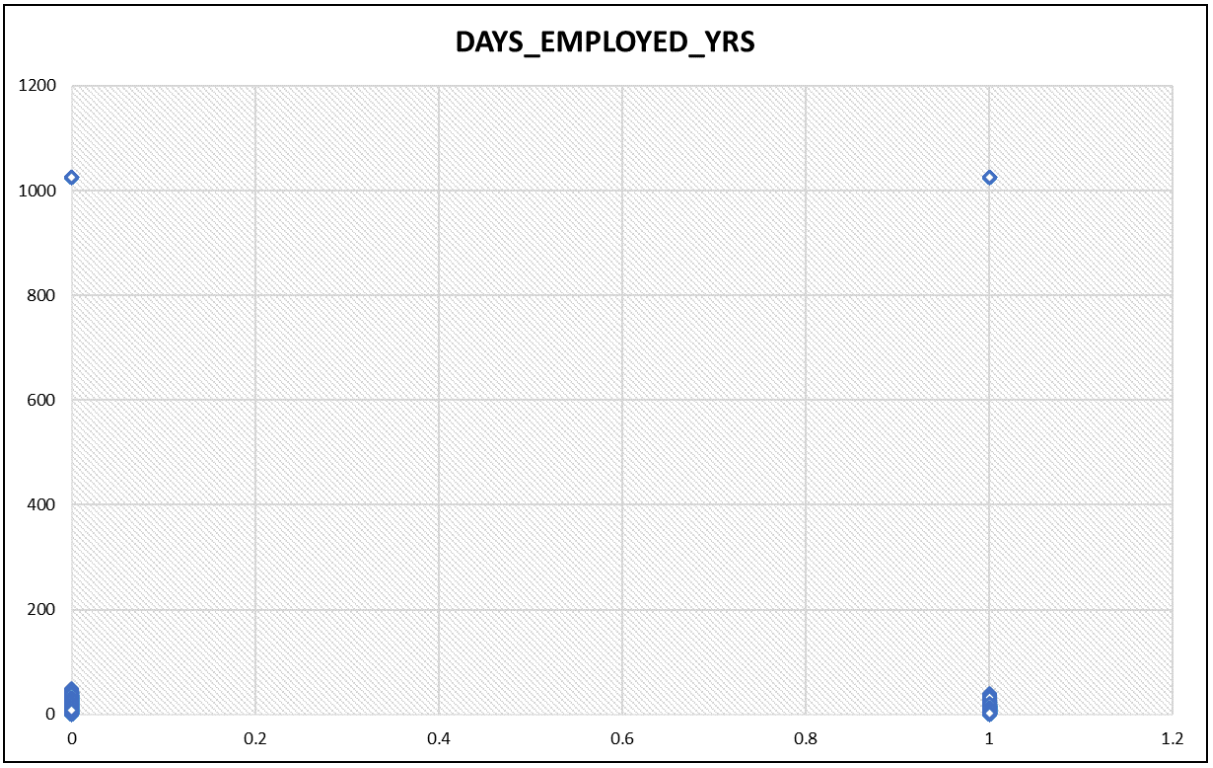
Lower limit =D4-(1.5*G4)



➔ Finding Outliers in the CNT_CHILDREN column using the TARGET column.



➔ Finding Outliers in the DAYS_EMPLOYED_YRS column using the TARGET column.

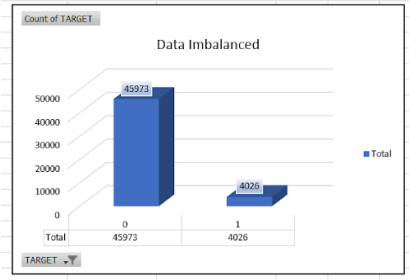


C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

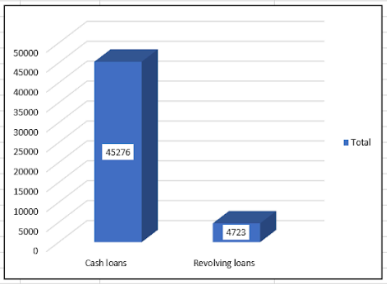
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

- ➔ Here,I'm finding data imbalance between two different columns with each variable.
- ➔ For Example...

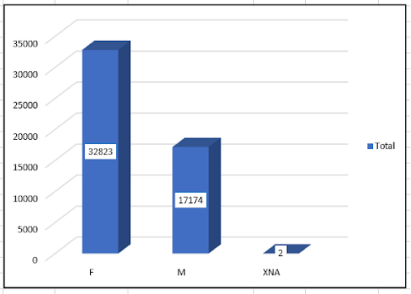
Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	49999



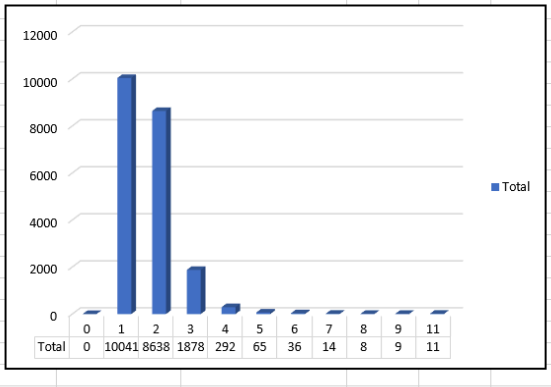
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45276
Revolving loans	4723
Grand Total	49999



Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	49999



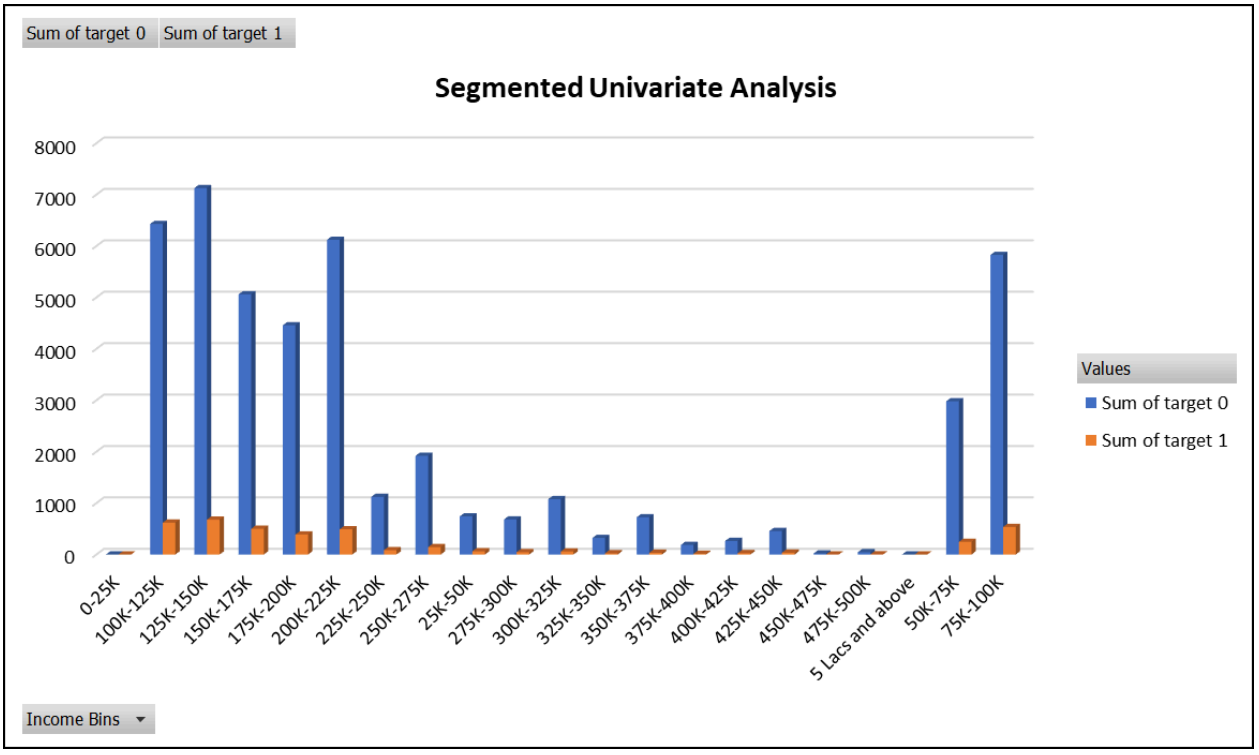
Row Labels	Sum of CNT_CHILDREN
0	0
1	10041
2	8638
3	1878
4	292
5	65
6	36
7	14
8	8
9	9
11	11
Grand Total	20992



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.

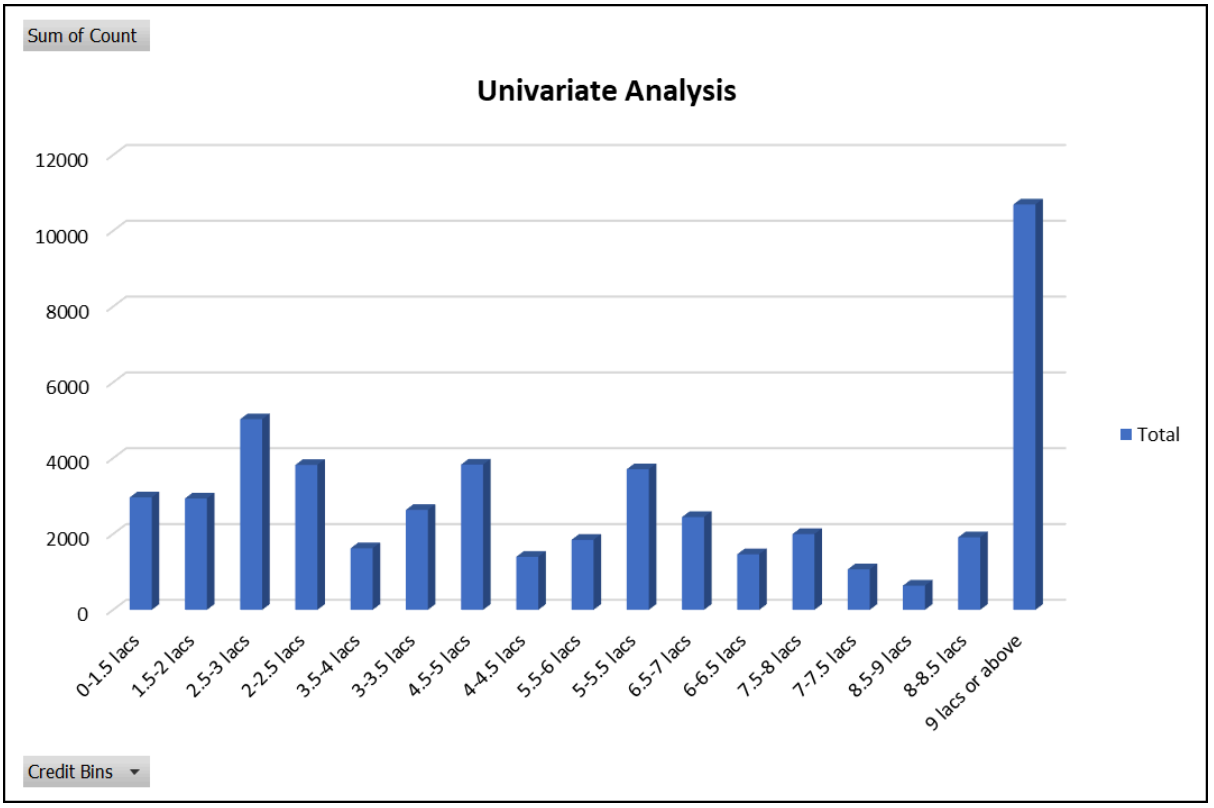
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

➔ **Segmented Univariate Analysis**
Creating an Income Bins based on AMT_INCOME_TOTAL column and divide them to different target values like 1 or 0.



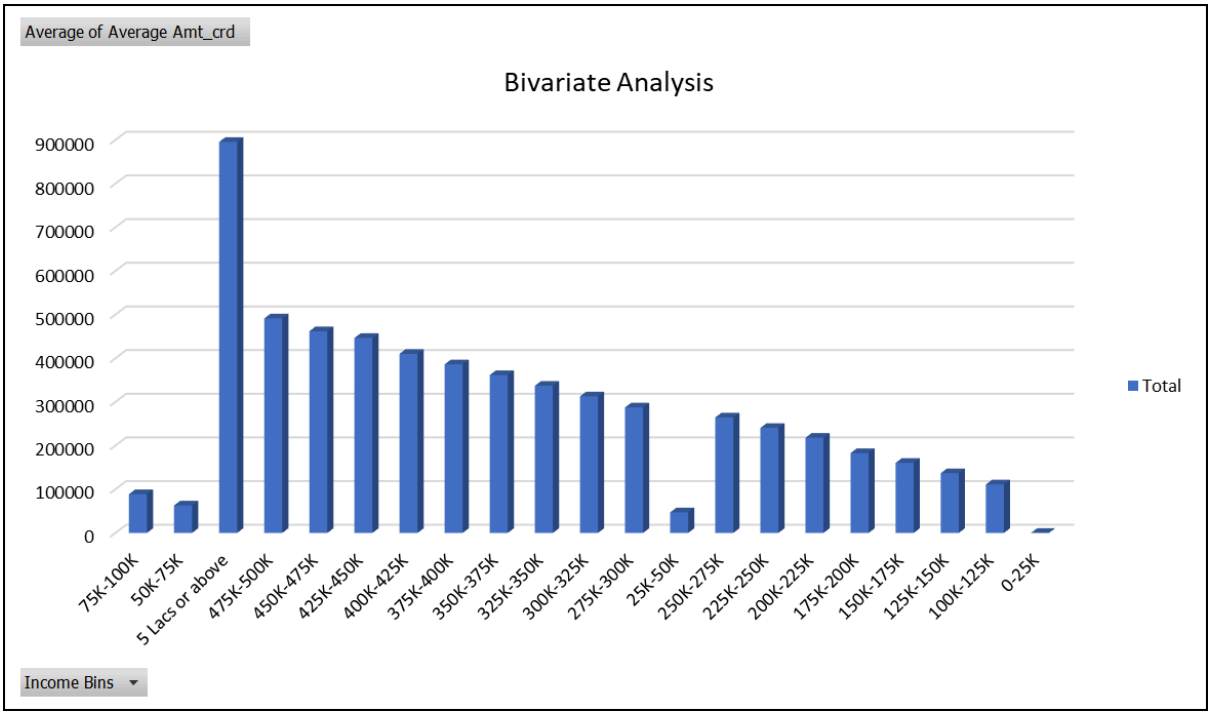
➔ Univariate Analysis

Creating Credit bins based on AMT_CREDIT column and count number of users within given bins.



➔ Bivariate Analysis

Finding Average amount of credit based on income bins.



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.

➔ Finding Correlation between variable and target variable.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH_YRS	DAYS_EMPLOYED_YRS	DAYS_ID_PUBLISH_YRS	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.009588558	0.00497156	-0.025555665	-0.329213627	-0.241534635	0.032330629	0.025913889
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	0.029841469	-0.015988003	-0.031512146	-0.004014667	-0.038188511
AMT_CREDIT	0.00497156	0.069315897	1	0.095111221	0.059192263	-0.067744072	0.012696596	-0.100507425
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	1	0.032436913	-0.004159946	0.004858946	-0.532667302
DAYS_BIRTH_YRS	-0.329213627	-0.015988003	0.059192263	0.032436913	1	0.621515057	0.269801734	-0.016698305
DAYS_EMPLOYED_YRS	-0.241534635	-0.031512146	0.012696596	-0.532667302	0.621515057	1	0.272331917	0.034559774
DAYS_ID_PUBLISH_YRS	0.032330629	-0.004014667	0.012696596	0.004858946	0.269801734	0.272331917	1	0.002138674
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.10050742	-0.532667302	-0.016698305	0.034559774	0.002138674	1

Result: This project involved extensive use of excel with this the major challenge was working with huge data.This project helped me to understand huge data and how to work with them and also helped me to learn new things in journey of data analyst.

Drive Link: Click Here to See the Excel File.

THANK YOU😀😀