Bank Loan Case Study

Project Description:

We have a dataset of a finance company specializing in lending various loans to urban customers. I thoroughly analyzed loan application data to identify some of the most important variables affecting loan approval decisions. I looked for patterns and trends by analyzing variables including applicant income, credit history, loan amount, and loan term using statistical methods and data visualization.

Approach:

I used a systematic method to analyze the data for the bank loan case study. To guarantee data integrity, I first cleaned and preprocessed the data, addressing outliers and missing values. After that, in order to obtain preliminary insights, I carried out exploratory data analysis (EDA), utilizing visualizations to find connections between variables like earnings, credit history, and loan approval status.

Tech Stack Used:

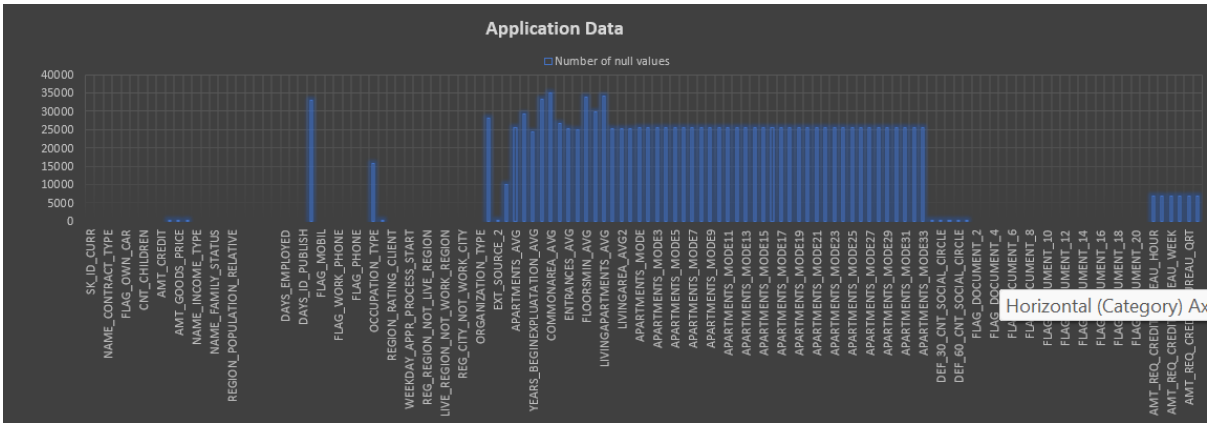
For this project, I used Microsoft Excel as the primary data analysis and visualization tool. Excel's powerful functions and pivot tables were employed for data cleaning, aggregation, and exploration, while charts and graphs were used to visualize key insights. Additionally, I used Microsoft Word for documenting the analysis process, creating comprehensive reports, and presenting findings in a structured format.

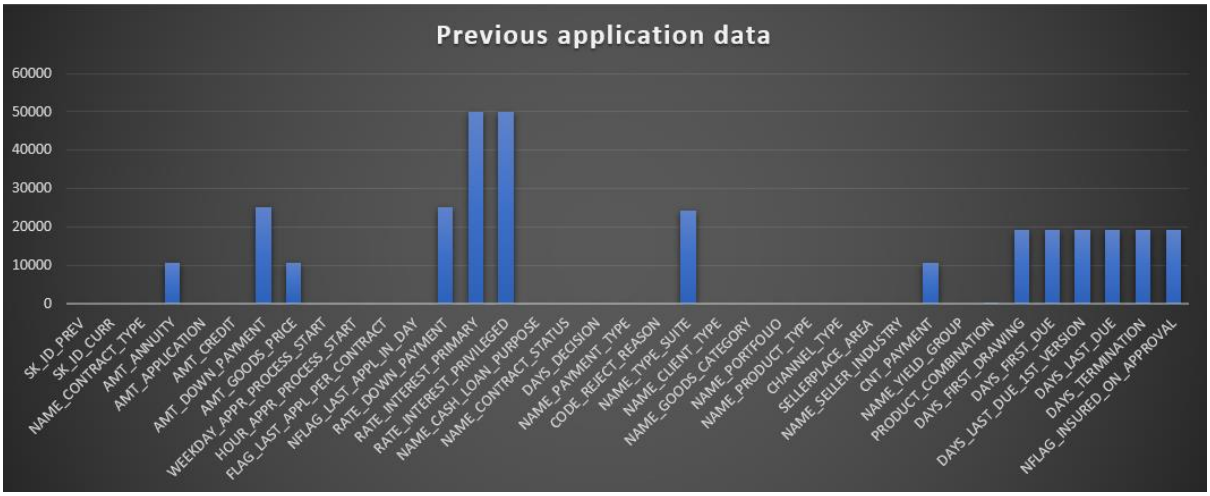
Insights:

1. **Identify Missing Data and Deal with it Appropriately:**

After going through the whole data, it was found that both tables, application data, and previous application data were full of empty cells, columns, and rows. It was thoroughly examined and dealt with the empty data. The columns with more than 40% of empty cells were marked and dropped so that it does not create errors and result in inaccurate insights.

Following are the bar charts representing the number of null cells in each column for each table/data.





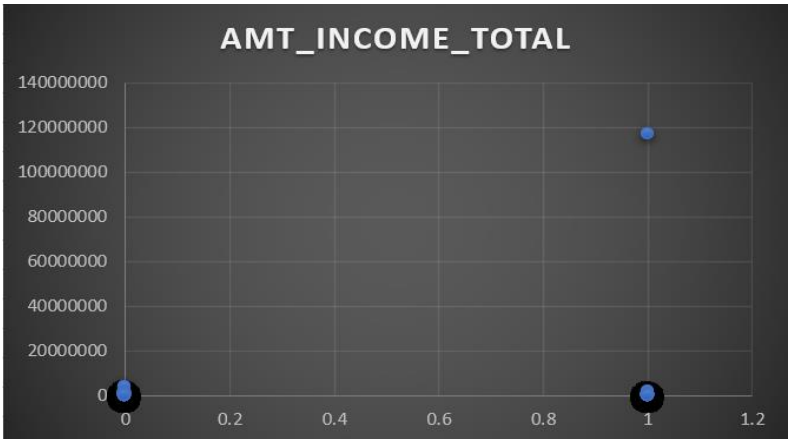
2. <u>Identify Outliers in the Dataset:</u>

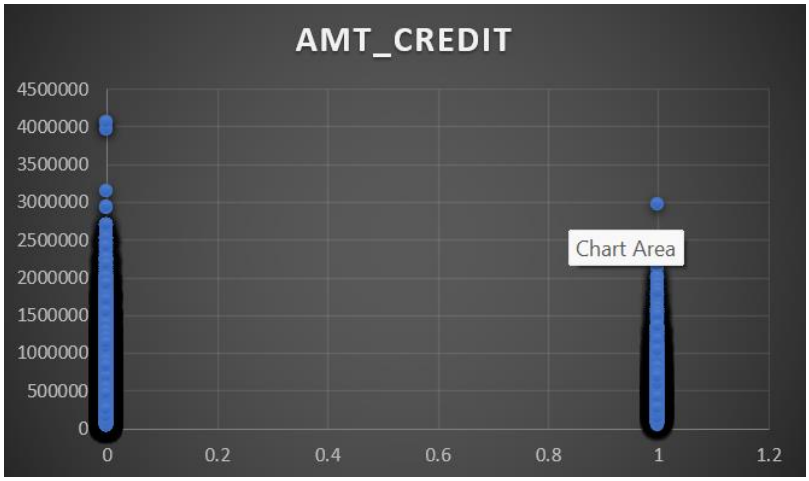
A statistical analysis was carried out to find quartiles, interquartile range, lower limits, and upper limits. Proper scatter plots were also formed in MS Excel to find the outliers in the dataset such as CNT_CHILDREN, DAYS_EMPLOYED, etc.

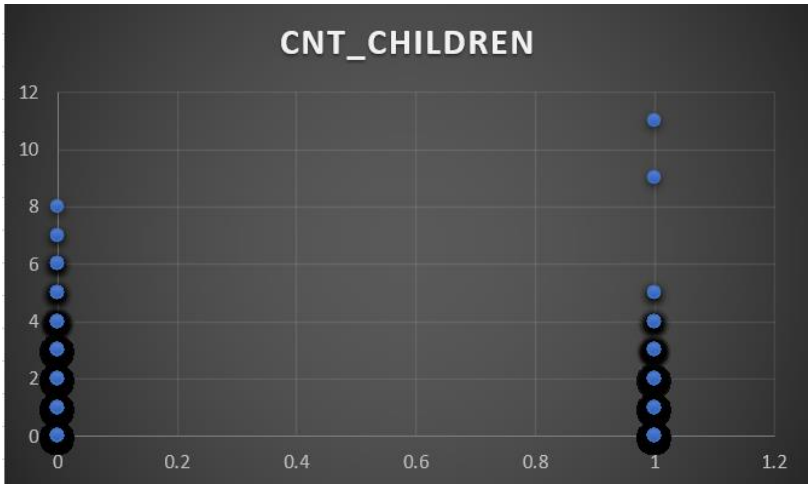
This analysis was done for both datasets, and the following are the useful insights that were found.

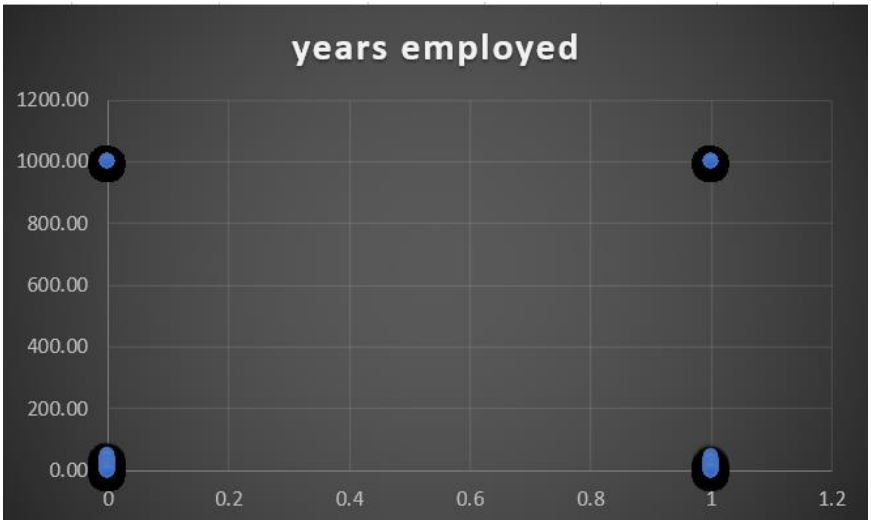
APPLICATION DATA:

Columns	Q1	Q3	interquartile range	Upper limit	Lower limit
AMT_INCOME_TOTAL	112500	202500	90000	337500	-22500
AMT_CREDIT	270000	808650	538650	1616625	-537975
AMT_ANNUITY	16456.5	34596	18139.5	61805.25	-10752.75
AMT_GOODS_PRICE	238500	679500	441000	1341000	-423000



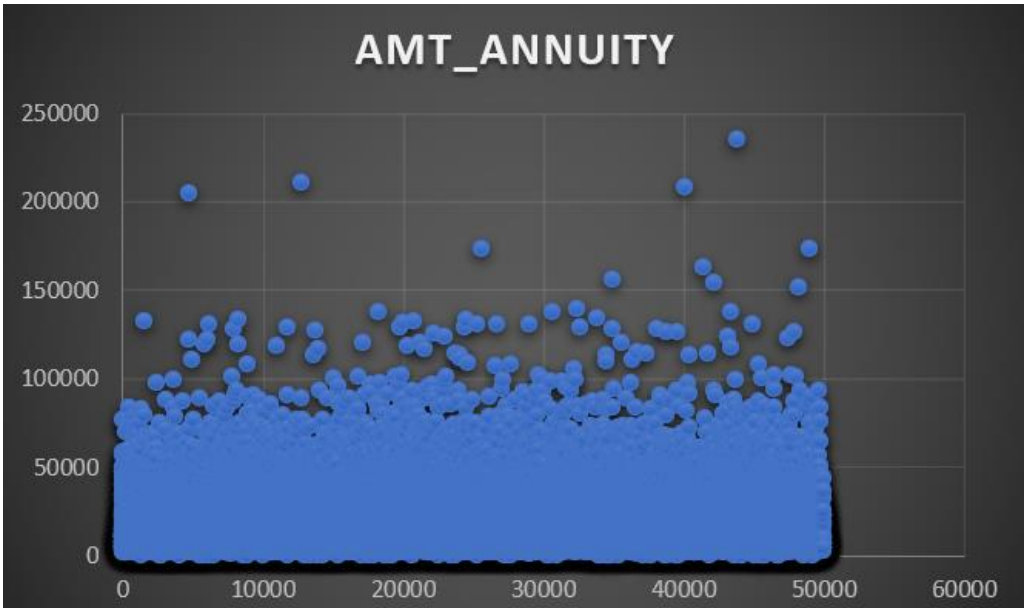


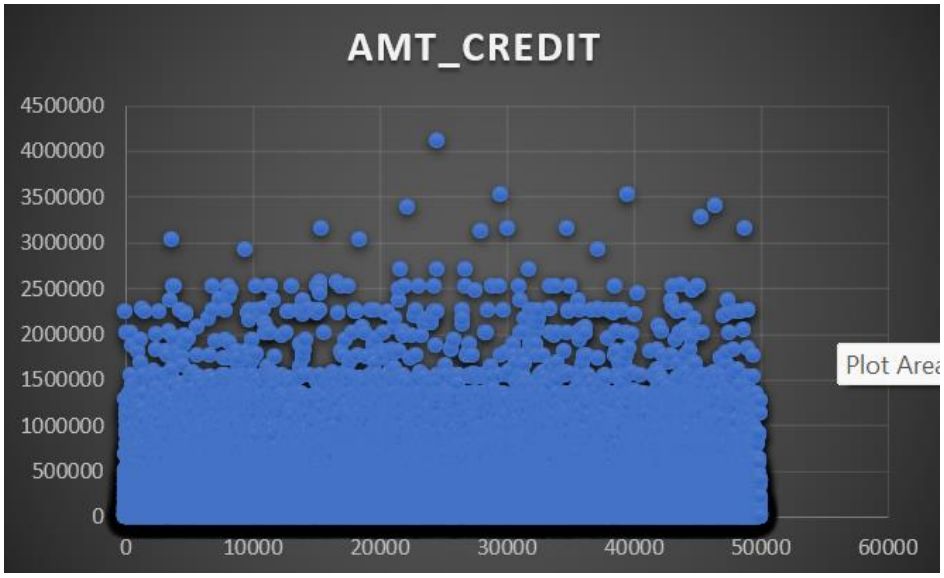


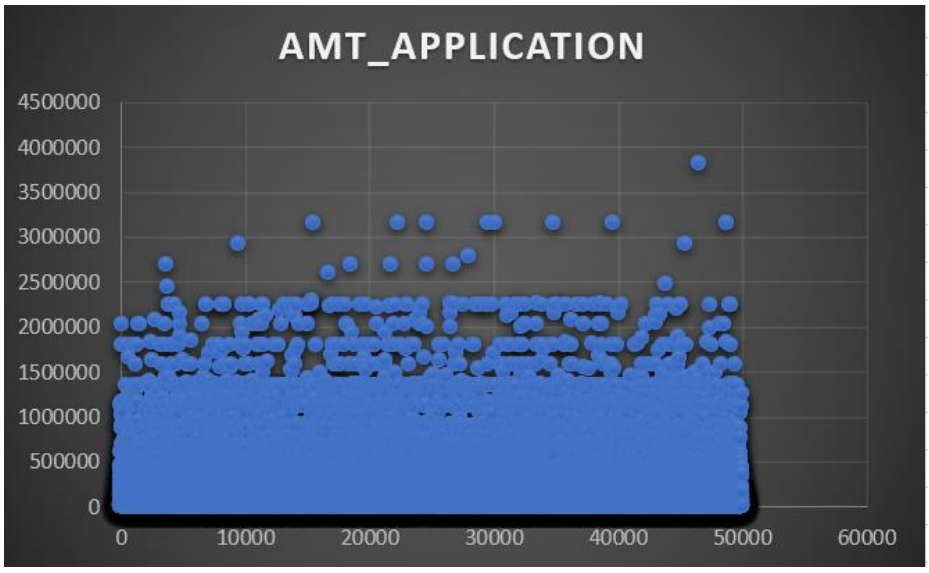


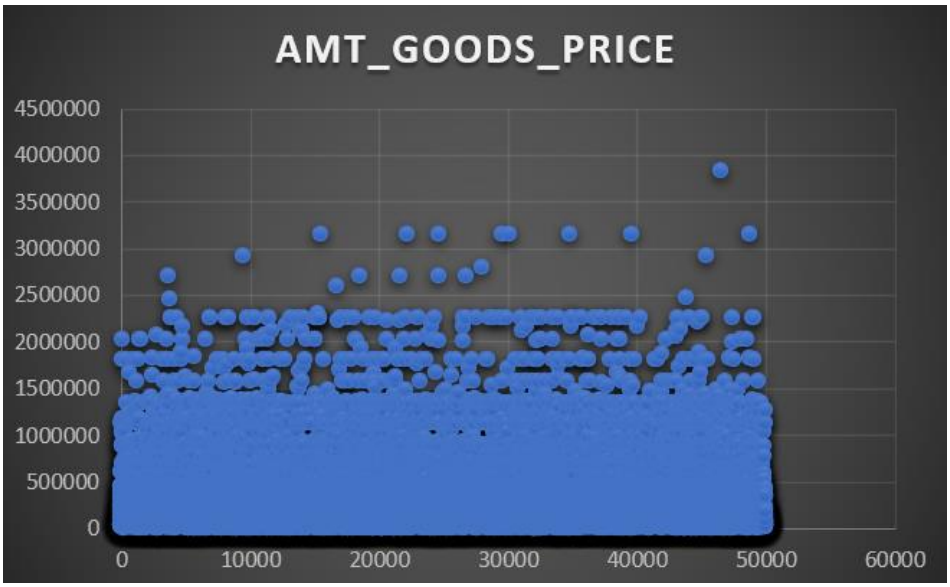
PREVIOUS APPLICATION DATA:

Columns	Q1	Q3	interquartile range	Upper limit	Lower limit
AMT_ANNUITY	238500	679500	441000	1341000	-423000
AMT_APPLICATION	22045.5	180000	157954.5	416931.75	-214886.25
AMT_CREDIT	26055	198106	172050.75	456181.875	-232021.13
AMT_GOODS_PRICE	71967.375	182989	111021.75	349521.75	-94565.25




AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000
60000
Plot Area

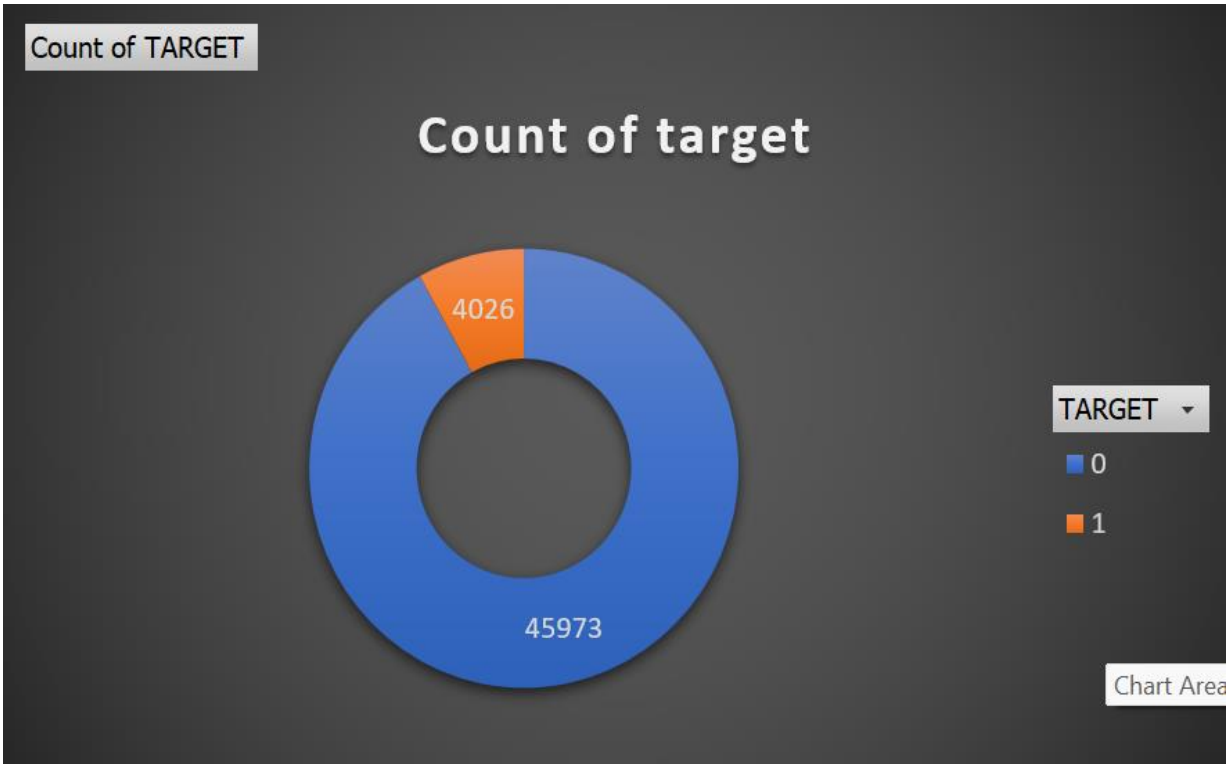

AMT_APPLICATION
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000
60000


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000
60000

3. Analyze Data Imbalance:

There were two segments people in the data represented as 0 and 1 under target column. After examining it was found that the data was highly imbalanced as people under category 0 were more than the people of category 1.

Row Labels	Count of TARGET	Count	Percentage
0	45973	45973	91.95
1	4026	4026	8.05
Grand Total	**49999**	49999	

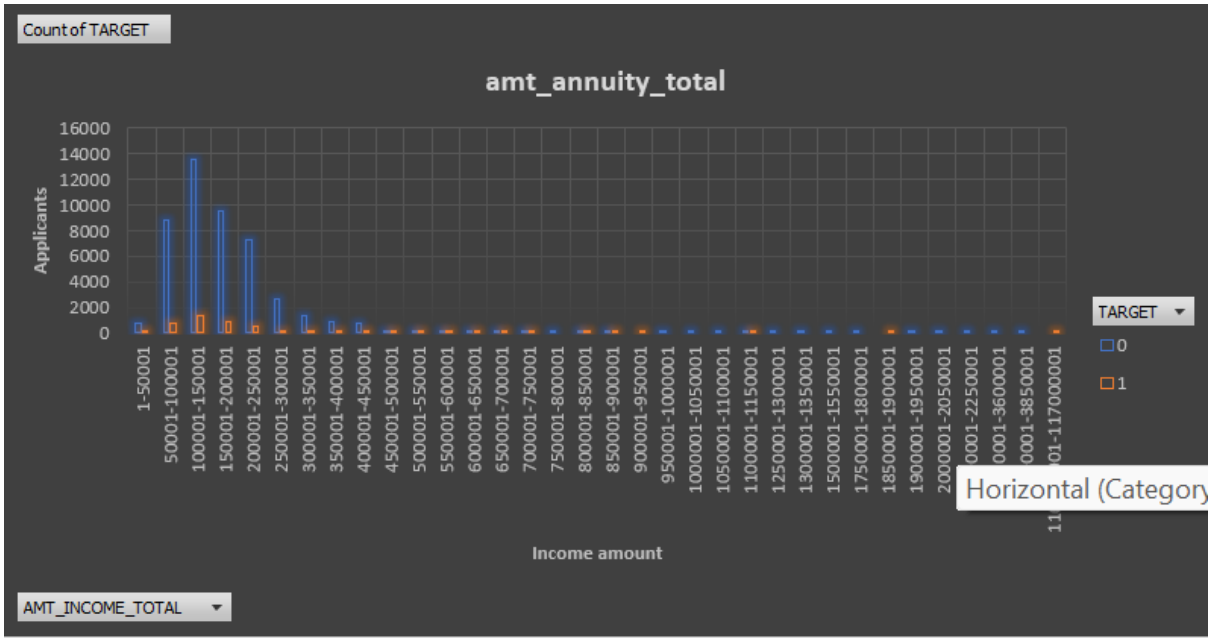


4. Univariate, Segmented Univariate, and Bivariate Analysis:

Segmented Univariate:

A segmented univariate analysis was done on the income of segmented people i.e. 0 and 1. First, a group of ranges was made for the income of people with a gap of 50000, and the number of people from each segment that belongs to that range was arranged.

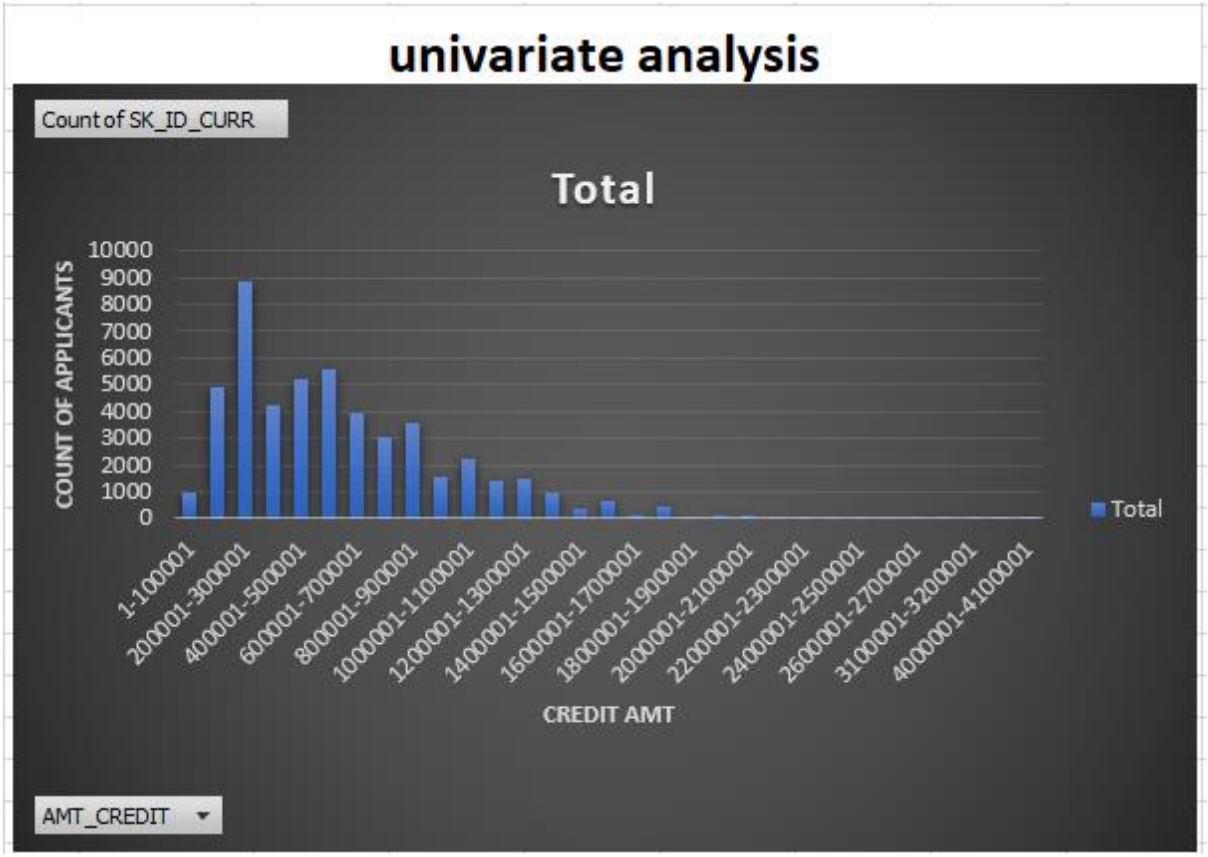
Count of TARGE Column La			
Row Labels	0	1	Grand T
1-50001	741	63	804
50001-100001	8806	782	9588
100001-150001	13554	###	14852
150001-200001	9518	890	10408
200001-250001	7242	576	7818
250001-300001	2600	188	2788
300001-350001	1398	83	1481
350001-400001	909	48	957
400001-450001	719	62	781
450001-500001	63	5	68
500001-550001	115	9	124
550001-600001	38	5	43
600001-650001	39	1	40
650001-700001	109	8	117
700001-750001	21	1	22
750001-800001	11		11
800001-850001	19	2	21
850001-900001	33	1	34
900001-950001		1	1
950001-1000001	1		1
1000001-1050001	1		1
1050001-1100001	3		3
1100001-1150001	12	1	13
1250001-1300001	1		1
1300001-1350001	10		10
1500001-1550001	1		1
1750001-1800001	2		2
1850001-1900001		1	1
1900001-1950001	1		1
2000001-2050001	2		2
2200001-2250001	2		2
3550001-3600001	1		1
3800001-3850001	1		1
116950001-117000001		1	1
Grand Total	45973	##	49999



Univariate:

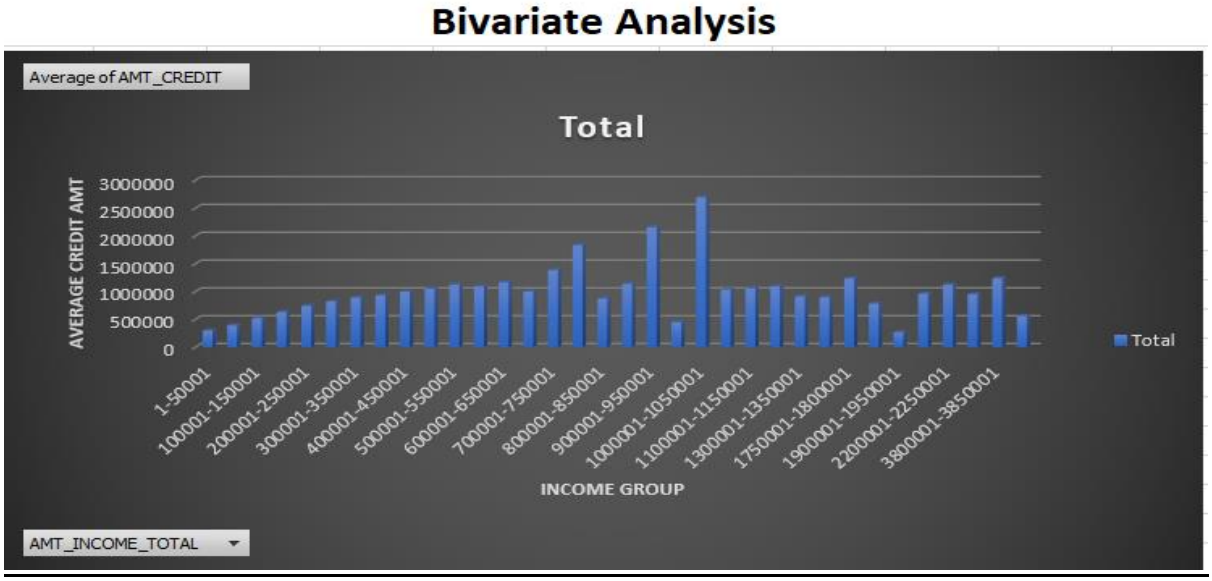
Univariate analysis was carried out by analyzing the credit/loan amount of each applicant. Again, a group of ranges was made for credit amount with a gap of 100000, and the number of applicants belonging to that range was arranged.

Row Labels	Count of SK_ID_CURI
1-100001	989
100001-200001	4911
200001-300001	8849
300001-400001	4256
400001-500001	5228
500001-600001	5554
600001-700001	3909
700001-800001	3062
800001-900001	3571
900001-1000001	1524
1000001-1100001	2219
1100001-1200001	1396
1200001-1300001	1463
1300001-1400001	945
1400001-1500001	389
1500001-1600001	649
1600001-1700001	144
1700001-1800001	410
1800001-1900001	86
1900001-2000001	122
2000001-2100001	115
2100001-2200001	36
2200001-2300001	89
2300001-2400001	14
2400001-2500001	17
2500001-2600001	32
2600001-2700001	13
2900001-3000001	3
3100001-3200001	1
3900001-4000001	1
4000001-4100001	2
Grand Total	49999



Bivariate:

A bivariate analysis was also executed to examine the average credit amount and the range of income of applicants. It was done to find what is the average credit amount that applicants want from each range of income.



total income	Average of AMT_CREDIT
1-50001	297752.076
50001-100001	393033.336
100001-150001	520073.660
150001-200001	632290.904
200001-250001	741970.003
250001-300001	826106.650
300001-350001	892307.68
350001-400001	932609.929
400001-450001	999173.829
450001-500001	1051508.84
500001-550001	1124198.81
550001-600001	1091708.16
600001-650001	1165433.73
650001-700001	1001836.26
700001-750001	1386633.68
750001-800001	1836769.09
800001-850001	876760.071
850001-900001	1138641.08
900001-950001	215640(
950001-1000001	450000
1000001-1050001	269550(
1050001-1100001	103500(
1100001-1150001	1062698.88
1250001-1300001	109511
1300001-1350001	914911.2
1500001-1550001	900000
1750001-1800001	123750(
1850001-1900001	78192(
1900001-1950001	26955(
2000001-2050001	961827.75
2200001-2250001	112500(
3550001-3600001	95346(
3800001-3850001	1241023.5
116950001-117000001	56249
Grand Total	599700.581

5. **<u>Identify Top Correlations for Different Scenarios:</u>**

<u>To find the correlation between the different variables such as CNT_CHILDREN, AMT_INCOME, AMT_CREDIT, etc, it was needed to use a correlation formula in MS Excel every time for different variables. I made a table to properly visualize it and also did conditional formatting to differentiate between positive correlation and negative correlation.</u>

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	years employed	years registration	years birth
CNT_CHILDREN	1	0.009588558	0.00497156	-0.025555665	-0.241539565	-0.181217183	-0.329264
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	0.029841469	-0.03151033	-0.009952379	-0.016003
AMT_CREDIT	0.00497156	0.069315897	1	0.095111221	-0.06773941	-0.003448569	0.0593427
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	1	-0.004158337	0.059322344	0.0325137
years employed	-0.241539565	-0.03151033	-0.06773941	-0.004158337	1	0.209172133	0.6217283
years registration	-0.181217183	-0.009952379	-0.00344857	0.059322344	0.209172133	1	0.3336325
years birth	-0.329263754	-0.016002774	0.059342658	0.032513748	0.62172831	0.333632509	1

By:- SOURABH GUPTA
EXCEL FILE LINK: LINK