

## TRAINITY PROJECT 6

### BANK LOAN CASE STUDY

HYPERLINK OF EXCEL FILE:

<https://docs.google.com/spreadsheets/d/1zs7h9MI52Rltt-UMvt0wk6oYZnnTNNf8/edit?usp=sharing&ouid=106990321423670318865&rtpof=true&sd=true>

**MAYANK KUMAR GANDHARV**

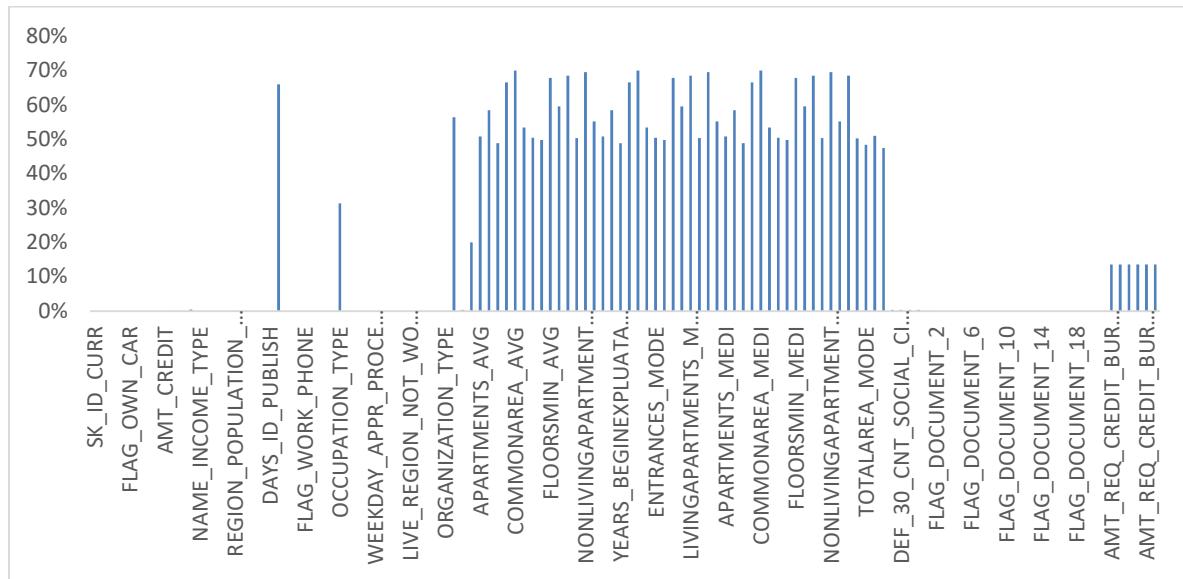
**TECH STACK USED:** Microsoft Excel 2021.

#### DATA ANALYTICS TASKS:

- 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.

Steps involved to do this task:

1. First I checked for any duplicate data.
2. Then I inserted 2 columns and did further analysis to find missing data.
3. First I used “**COUNTA**” Function to count each row for a particular column.
4. After that I calculated the missing value in form of percentage (Syntax: **=1-A2/\$A\$2**).
5. At last I got missing values for each column in form of percentage and marked red for every value which is above 30%.



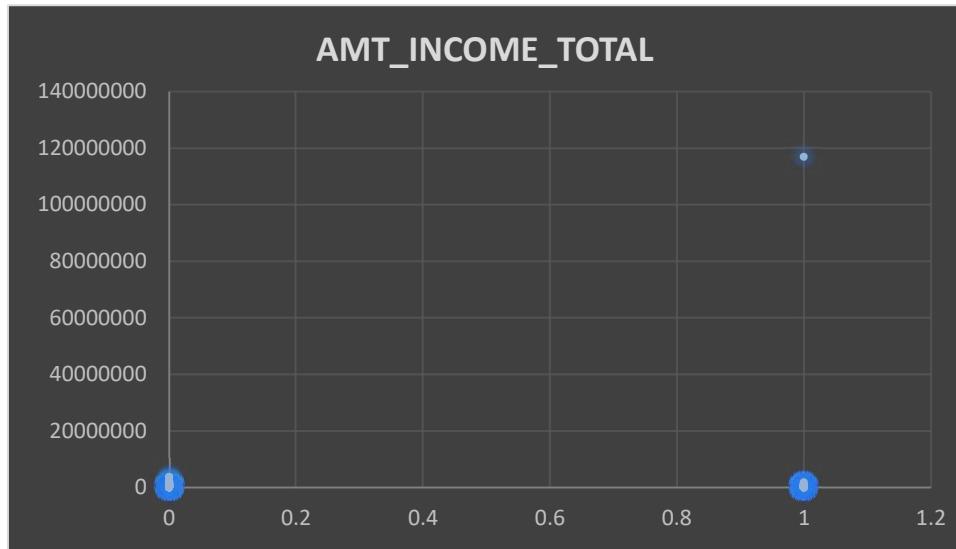
For Imputation of data in the dataset (for data whose missing value is less than 30%), I used MEDIAN AND AVERAGE.

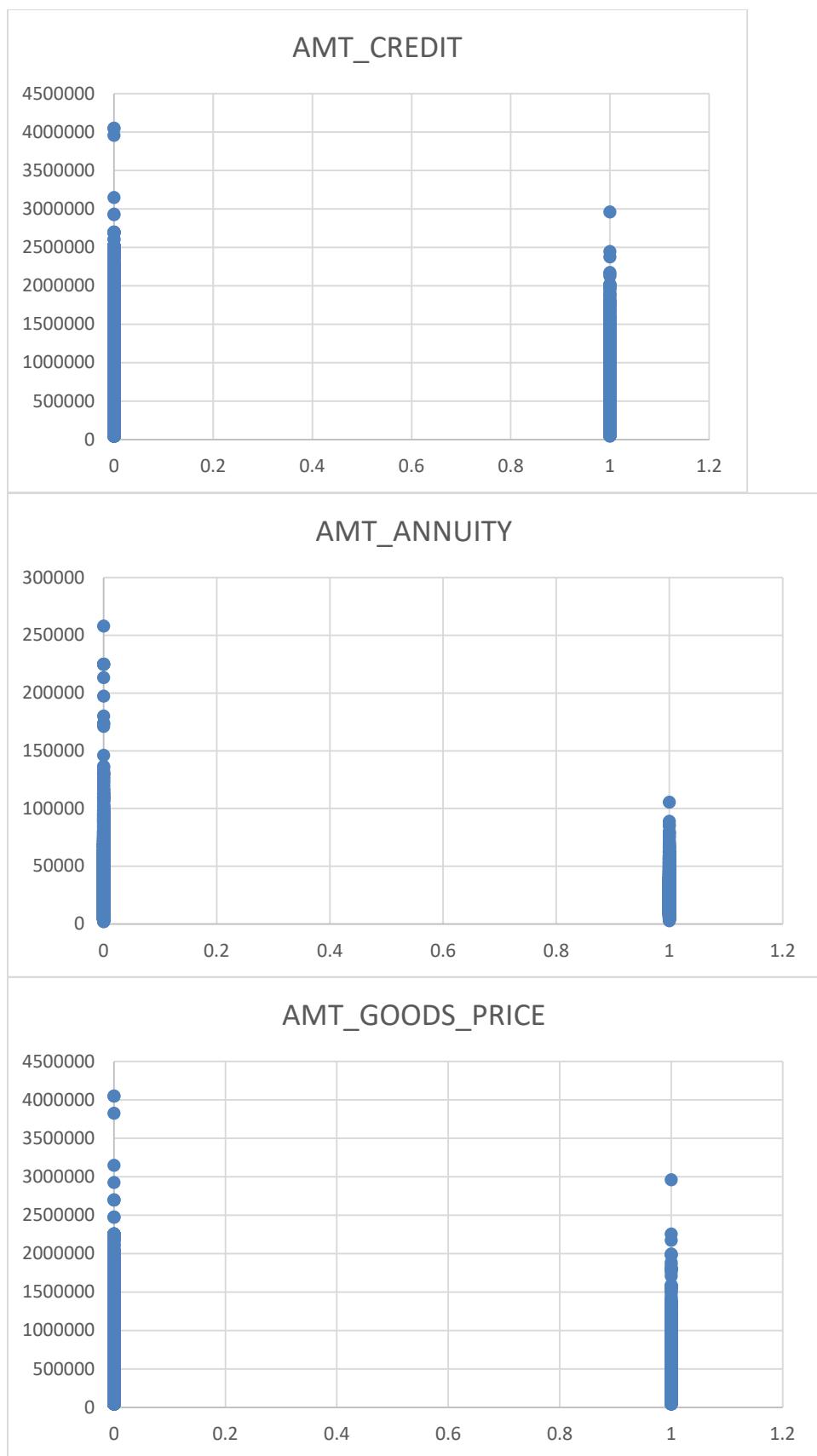
**Inference:** There are many variables with a lot of missing values but can play an important role in analysis. In order to make the analysis more accurate data should be imputed. Now, for imputation there can be 2 ways, first being taking the average of the whole column and impute all the blanks with the average value. Second being the median way, this data imputation way is only viable if the data column has fixed variable values such as (1,2,3,4....), in this case the most occurring value should be imputed in the blanks.

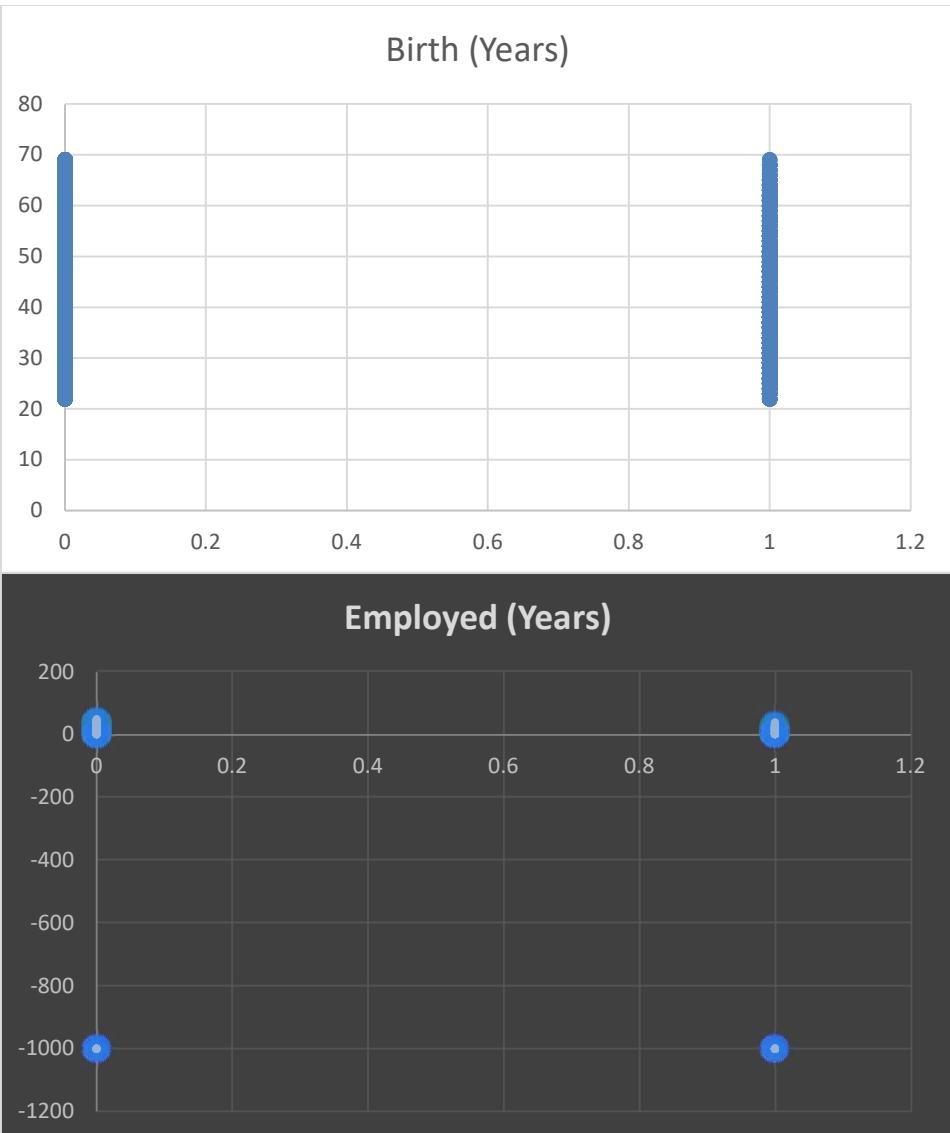
- B. **Identify Outliers in the Dataset:** Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

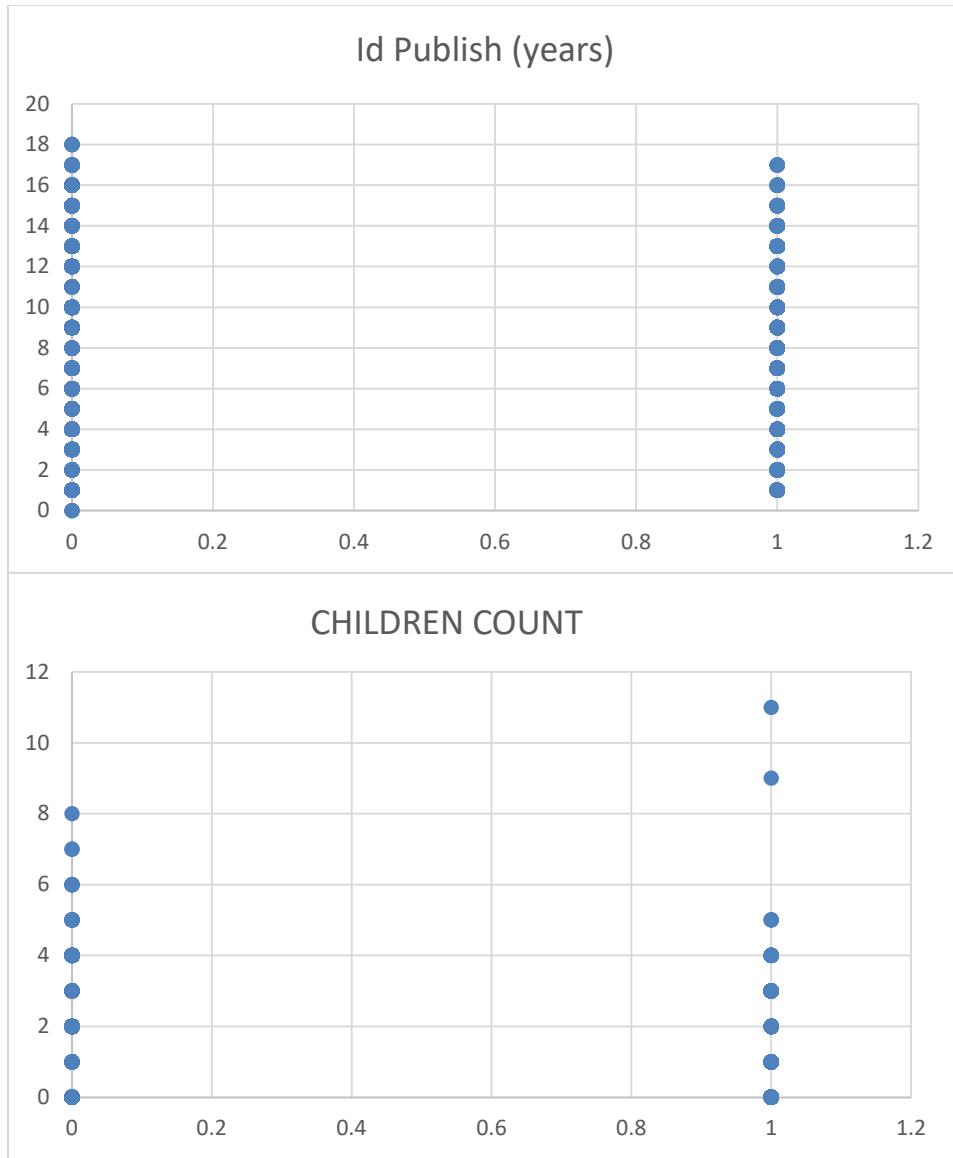
To find the outliers:

1. I calculated quartile 1 and 3 by using syntax `=QUARTILE.EXC([AMT_INCOME_TOTAL],1)` and `=QUARTILE.EXC([AMT_INCOME_TOTAL],3)` respectively.
2. After calculating Q1 and Q3 I calculated Inner Quartile Range [Q3-Q1].
3. Lastly I calculated the upper bound and lower bound by using SYNTAX `=[@Q3]+(1.5*[@IQR])` and `=[@Q1]-(1.5*[@IQR])` Respectively.
4. Some date columns were provided with negative values, so I changed the whole column into the positive value and also divided each row with 365 so that I could get the data in year form.
5. Also, I formed a scatter plot of some numeric data with respect to Target (0,1), to make the outliers visual.









#### INSIGHTS:

For this dataset, the potential outliers can be in the column “**Employed (Years)**” and “**AMT\_INCOME\_TOTAL**”. The value “1001” in Employed (years) is not practically possible and the value “117000000” in AMT\_INCOME\_TOTAL is also an outlier for this data. Rest outliers based on IQR is considerable for further analysis.

- 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.

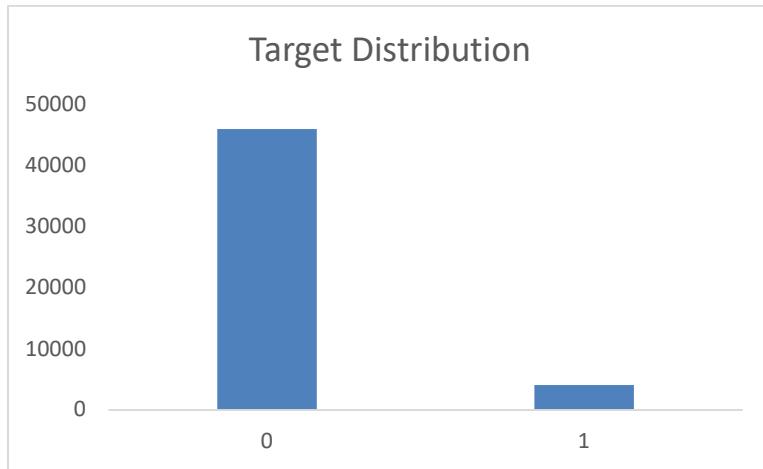
For this task following steps were taken:

1. Since, there are only 2 variables through which the data is revolving around which is Target variable (1 - client with payment difficulties: he/she had late payment more than X days on

at least one of the first Y installments of the loan in our sample, 0 - all other cases), the imbalance is found for these 2 variables.

2. In order to find the imbalance, first I pivoted the data of Target and SK\_ID\_CURR and took count of target based on 1 and 0.
3. After that I calculated the percentage distribution of 1 and 0 in the dataset.

Target	Count of TARGET	Percentage
0	45973	92%
1	4026	8%
<b>Grand Total</b>	<b>49999</b>	<b>100%</b>

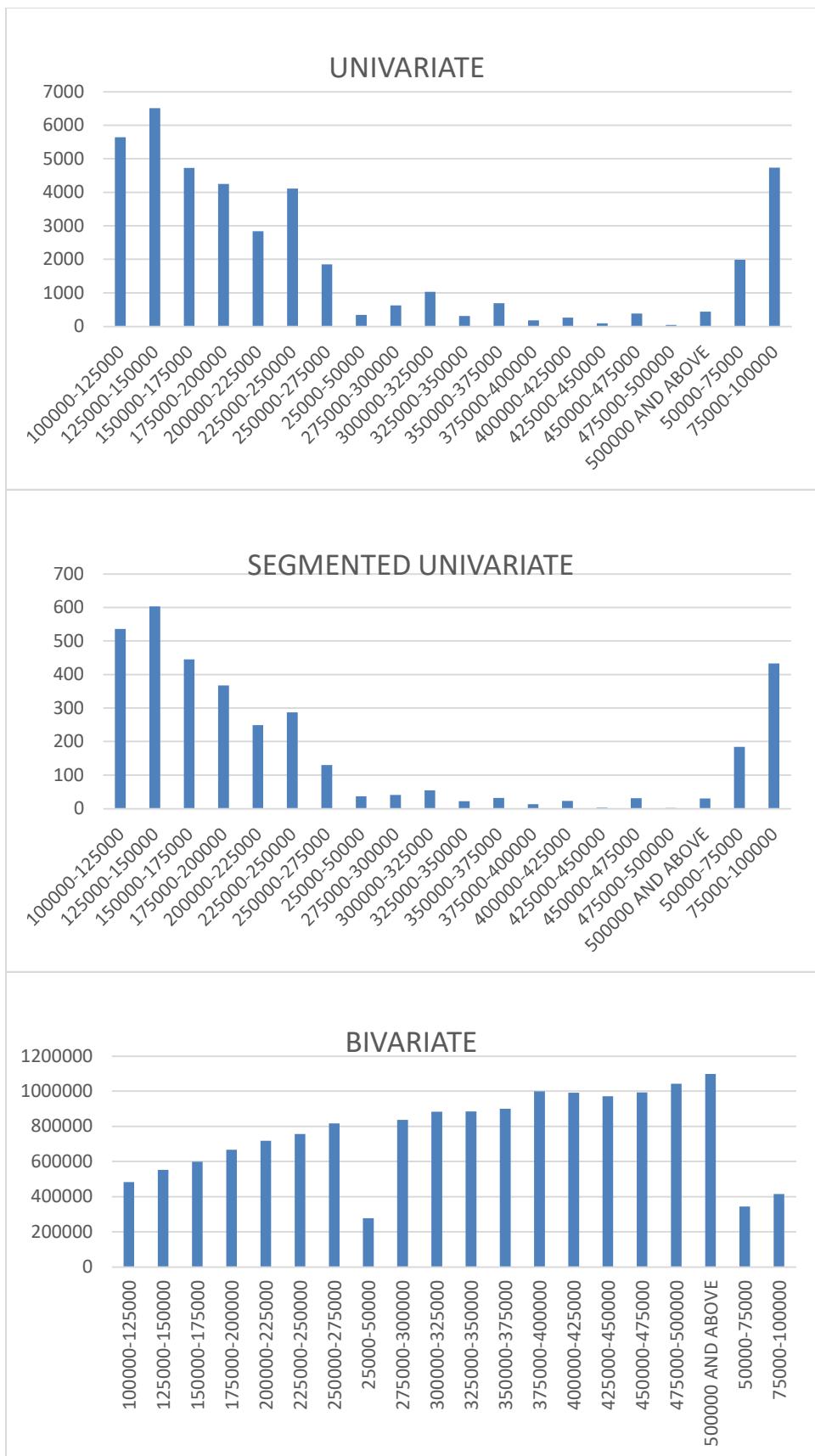


**Inference:** the distribution of target values has huge difference.

- 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.

To perform this task, following steps were taken:

1. Before performing any analysis, first I divided the annual income into intervals using VLOOKUP and formed a column "BUCKET".
2. After creating the bucket I then deleted all the outliers.
3. Now, all the data is ready for further analysis.
4. For univariate analysis I created a pivot table grouping the data by BUCKET and count of Target.
5. For Segmented Univariate, I grouped "BUCKET" and count of Target also, filtered the target in order to segment the data into two segments (1 and 0).
6. For Bivariate, I considered 2 data column. First being the AMT\_INCOME\_TOTAL and second being the AMT\_CREDIT. Then pivoted the data and grouped the Bucket and average of AMT\_CREDIT.



application\_data new - Excel

AMT_CREDIT	BUCKET	Employed(Y)	Column1	NUMBER	RANGE	UNIVARIATE	TARGET	COUNT OF TARGET
406597.5	20000-22500K	2 N		25000	25000-50000	Row Labels	SEGMENTED UNIVARIATE	1
1293501.5	25000-27500K	4 N		50000	50000-75000	5637	100000-125000	536
135000	50000-75000	1 N		75000	75000-100000	125000-150000	125000-150000	603
312628.5	125000-150000	9 N		100000	100000-125000	150000-175000	150000-175000	445
513000	100000-125000	9 N		125000	125000-150000	175000-200000	175000-200000	367
490495.5	75000-100000	5 N		150000	150000-175000	200000-225000	200000-225000	249
1560726	150000-175000	9 N		175000	175000-200000	225000-250000	225000-250000	287
1530000	35000-37500K	2 N		200000	200000-225000	250000-275000	250000-275000	130
405000	125000-150000	6 N		225000	225000-250000	250000-50000	250000-50000	37
652500	100000-125000	2 N		250000	250000-275000	275000-300000	275000-300000	41
				275000	275000-300000	300000-325000	300000-325000	55
80865	50000-75000	8 N		300000	300000-325000	325000-350000	325000-350000	22
918468	225000-250000	9 N		325000	325000-350000	350000-375000	350000-375000	32
773860.5	175000-200000	1 N		350000	350000-375000	375000-400000	375000-400000	13
299772	150000-175000	4 N		375000	375000-400000	400000-425000	400000-425000	23
509602.5	100000-125000	4 N		400000	400000-425000	425000-450000	425000-450000	3
270000	100000-125000	1 N		425000	425000-450000	450000-475000	450000-475000	21
157500	100000-125000	22 N		450000	450000-475000	475000-500000	475000-500000	2
644491	75000-100000	6 N		475000	475000-500000	500000 AND ABOVE	500000 AND ABOVE	30
427500	125000-150000	12 N		500000	500000 AND ABOVE	50000-75000	50000-75000	184
1132573.5	20000-22500K	5 N		75000	75000-100000	4737	75000-100000	433
497520	450000-475000	12 N		Grand Total	41075	Grand Total	3523	Grand Total
24								
25								
26								
27								
28								

Ready Circular References

Task 3 TARGET IMBALANCE    Task 4 UNI AND BIVARIATE    Task 4 DASHBOARD    Sheet4    Sheet ...

Type here to search

**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.

To perform this task I followed these steps:

1. First, I extracted data in form of "0" and "1" so that I could differentiate the correlation for both the variables.
2. Then, I made a table where I wrote the column names both Horizontally and Vertically, to find the correlation of subsequent columns.
3. Using the formula =CORREL('task 5 "0"'!\$B:\$B,'task 5 "0"'!B:B), I then calculated the correlation for every variable corresponding to the other variable.

B2 : =CORREL('task 5 "0"'!\$B:\$B,'task 5 "0"'!B:B)

A	B	C	D	E	F	G	H	I	J	K	L	M	
FOR TARGET "0"													
CNT_CHILDREN	1	-0.04603882	-0.01664573	-0.002563237	-0.020519895	-0.03141276	-0.24402217	-0.06564223	-0.16142737	0.12259874			
AMT_INCOME_TOTAL		-0.004603882	1	0.36237622	0.438291513	0.370267175	0.189034572	0.04950801	0.035684917	-0.03377063	0.028603295		
AMT_CREDIT			-0.01664573	0.83632762	0.1	0.76203902	0.986867101	0.09704598	0.160473618	0.094185238	0.023771303	0.045050564	
AMT_ANNUITY				-0.02551327		0.76203902	1	0.76701503	0.116075265	0.154591524	0.09503774	0.019716902	0.04521235
AMT_GOODS_PRICE					-0.020519895		0.70705173	0.986687101	0.76701503	0.100152038	0.154591524	0.09503774	0.019716902
REGION_POPULATION_RELATIVE						-0.031413276		0.189034572	0.09704598	0.100152038	0.149062827	-0.00522065	0.06455793
Birth (Years)							-0.241402217		0.049062827	1	0.351853547	0.304921246	0.106398424
Employed (Years)								-0.035684917	0.094185238	1	0.17559655	0.08209305	
Registration (years)									-0.03377063		1	0.036776855	
id Publish (years)										-0.16142737		1	
CNT_CHILDREN	1	-0.03009158	0.009855795	0.011434752	0.001984197	-0.01880466	-0.16142737	-0.016142524	-0.16142737	0.102811863			
AMT_INCOME_TOTAL		-0.030091588	1	0.312610193	0.376504151	0.316021702	0.09625773	0.08815784	0.021961217	-0.02840764	0.036047479		
AMT_CREDIT			-0.009855796	0.312610193	1	0.745212312	0.98210942	0.055227555	0.194608161	0.105859073	0.04277175	0.050000849	
AMT_ANNUITY				0.011434752		0.745212312	1	0.746710504	0.065554391	0.086150322	0.054912442	-0.012227862	0.050579826
AMT_GOODS_PRICE					-0.019841937		0.312610193	0.98210942	1	0.060820794	0.188436095	0.113346396	0.041569018
REGION_POPULATION_RELATIVE						-0.0180466		0.056554391	0.060820794	1	0.01633897	-0.000570196	0.047928204
Birth (Years)							-0.056525773		0.08815784	1	0.303074844	0.239659081	0.123915511
Employed (Years)								-0.016142524		0.036776855	1	0.15054485	0.098240742
Registration (years)									-0.000570196		1	0.04265344	
id Publish (years)										-0.16142737		1	
CNT_CHILDREN	1	0.036045749	0.050000849	0.050579826	0.05977521	0.047928204	0.239659081	0.15054485	0.04265344	0.04265344	1		

Ready Circular References

Task 4 DASHBOARD    Task 5 dataset    Task 5 "0"    Task 5 "1"    Task 5 CORRELATION

Type here to search

**The most correlated variables are:**

FOR TARGET “0”:

1. **AMT\_ANNUITY AND AMT\_CREDIT (VICE-VERSA):** 0.762033902
2. **AMT\_CREDIT AND AMT\_GOODS\_PRICE (VICE-VERSA):** 0.986687101
3. **AMT\_ANNUITY AND AMT\_GOODS\_PRICE (VICE-VERSA):** 0.767011503

FOR TARGET “1”:

1. **AMT\_CREDIT AND AMT\_GOODS\_PRICE (VICE-VERSA):** 0.982109142

**The least correlated variables are:**

FOR TARGET “0”:

1. **BIRTH (YEARS) AND CNT\_CHILDREN (VICE-VERSA):** -0.241402217
2. **REGISTRATION (YEARS) AND CNT\_CHILDREN (VICE-VERSA):** -0.161412737

FOR TARGET “1”:

1. **BIRTH (YEARS) AND CNT\_CHILDREN (VICE-VERSA):** -0.162383055
2. **REGISTRATION (YEARS) AND CNT\_CHILDREN (VICE-VERSA):** -0.130736906