TRAINITY PROJECT 6

BANK LOAN CASE STUDY

HYPERLINK OF EXCEL FILE:

https://docs.google.com/spreadsheets/d/1zs7h9Ml52Rltt-UMvt0wk6oYZnnTNNf8/edit?usp=sharing&ouid=106990321423670318865&rtpof=true&sd=true

MAYANK KUMAR GANDHARV

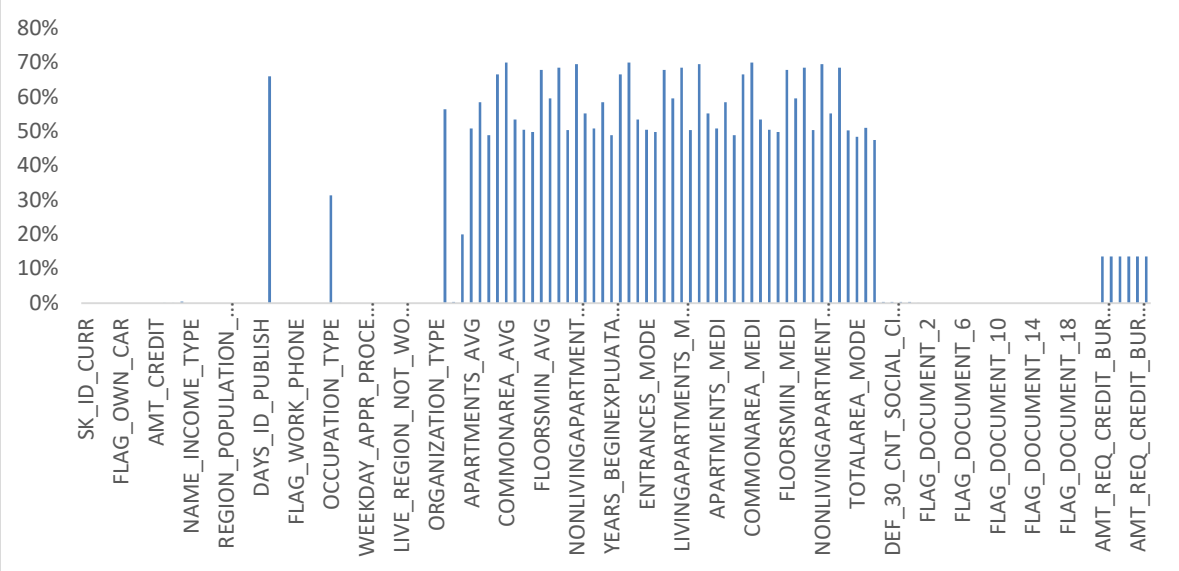
TECH STACK USED: Microsoft Excel 2021.

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/A2**).
5. At last I got missing values for each column in form of percentage and marked red for every value which is above 30%.

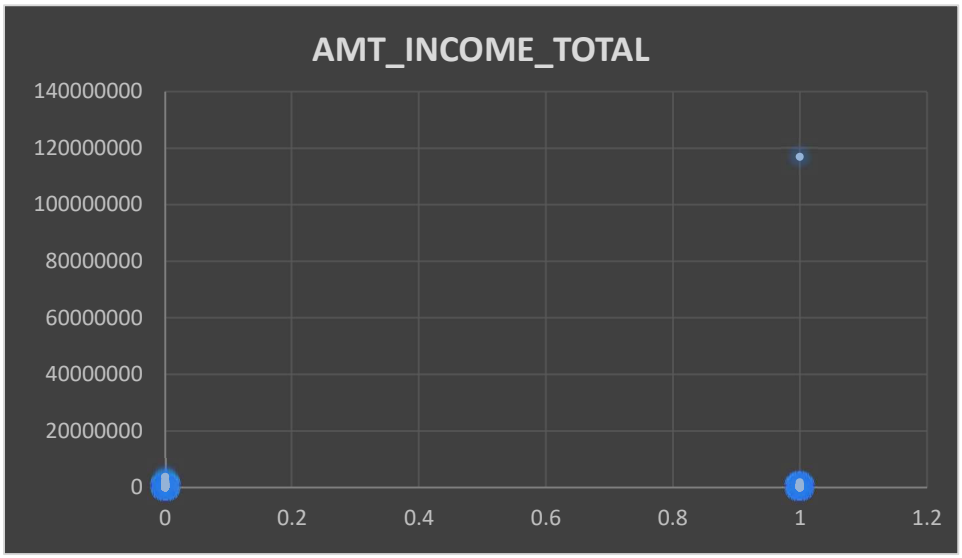


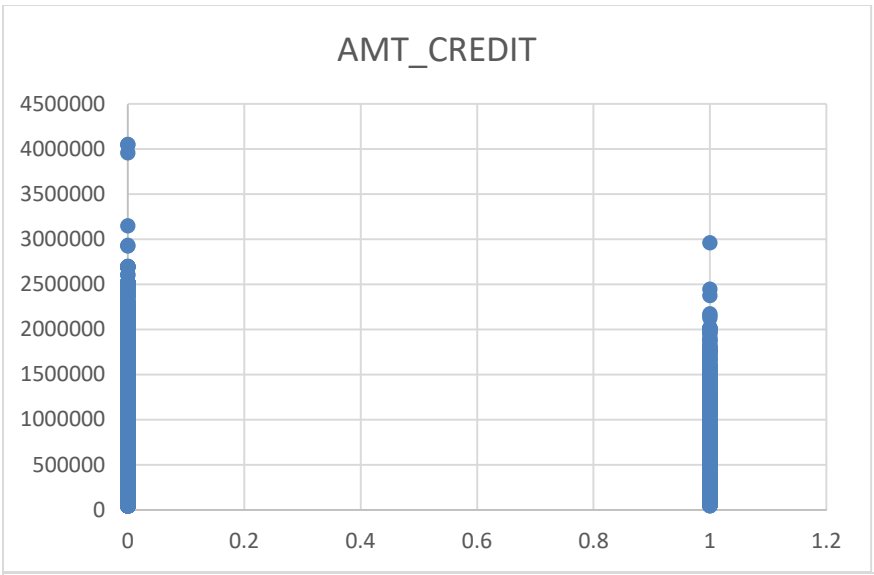
For Imputation of data in the dataset (for data whose missing value is less than 30%), I used MEDIAN AND AVERAGE.

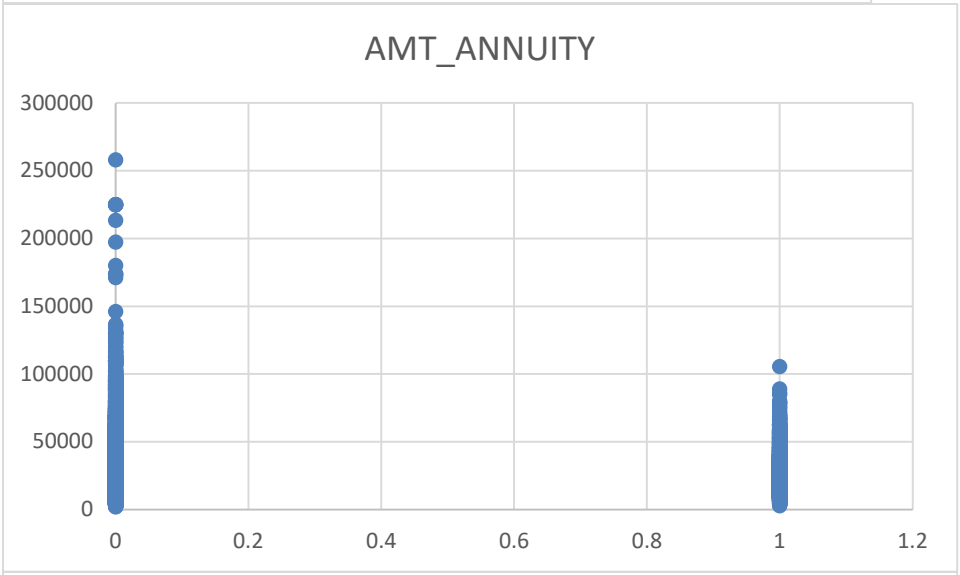
Inference: There are many variables with a lot of missing values but can play an important role in analysis. In order to make the analysis more accurate data should be imputed. Now, for imputation there can be 2 ways, first being taking the average of the whole column and impute all the blanks with the 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:

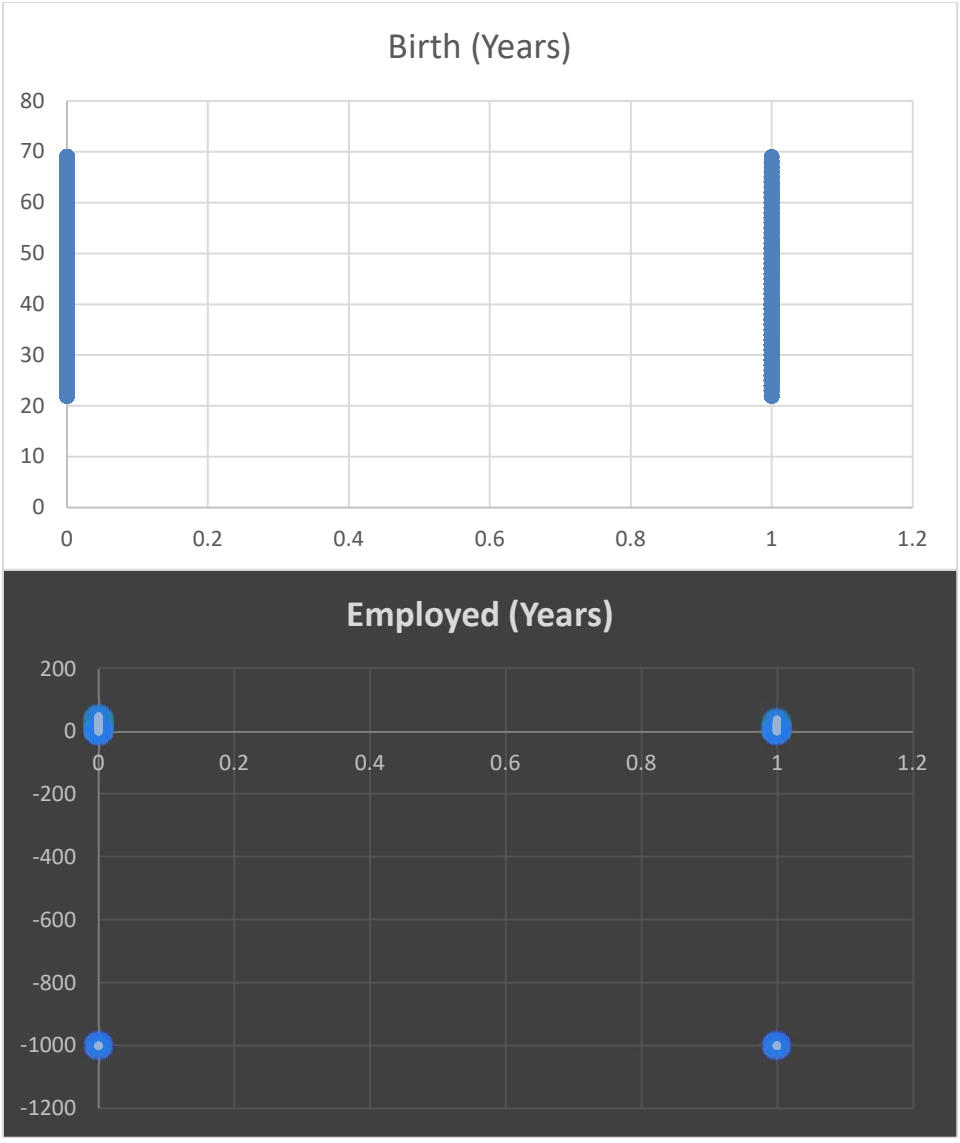
1. I calculated quartile 1 and 3 by using syntax **=QUARTILE.EXC([AMT_INCOME_TOTAL],1) and =QUARTILE.EXC([AMT_INCOME_TOTAL],3)** respectively.
2. After calculating Q1 and Q3 I calculated Inner Quartile Range [Q3-Q1].
3. Lastly I calculated the upper bound and lower bound by using SYNTAX **=[@Q3]+(1.5*[@IQR])** 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.

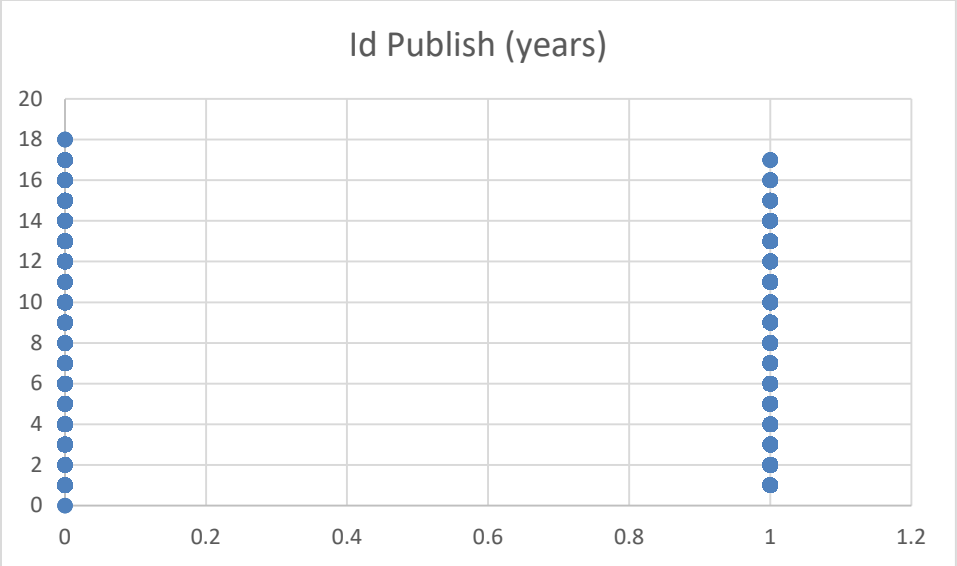


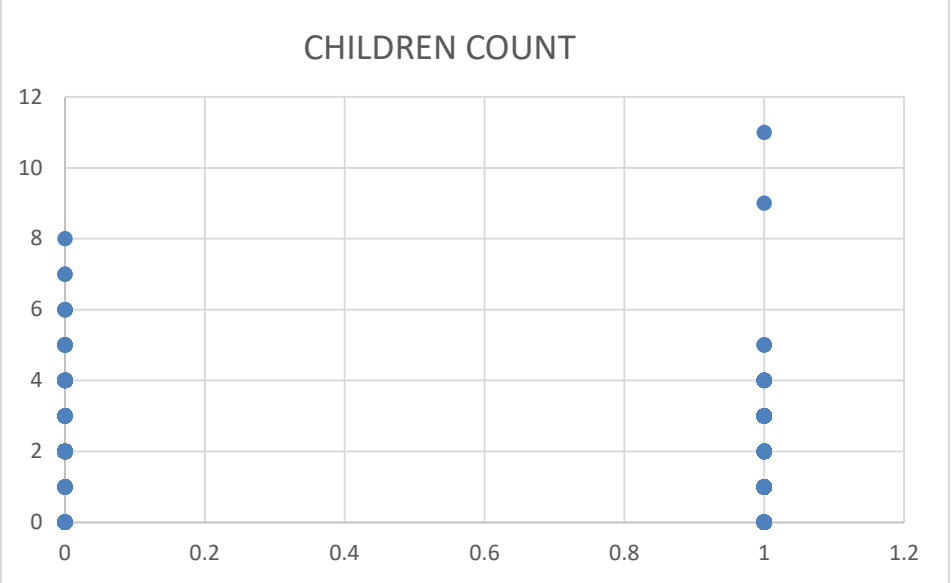

AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
0.2
0.4
0.6
0.8
1
1.2


AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
0
0.2
0.4
0.6
0.8
1
1.2


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
0.2
0.4
0.6
0.8
1
1.2


Birth (Years)
80
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
1.2
Employed (Years)
200
0
-200
-400
-600
-800
-1000
-1200
0
0.2
0.4
0.6
0.8
1
1.2





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.

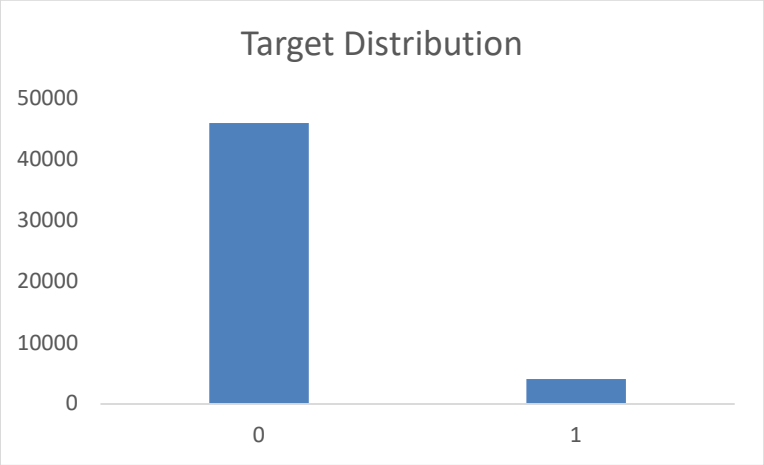
For this task following steps were taken:

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%
Grand Total	**49999**	**100%**

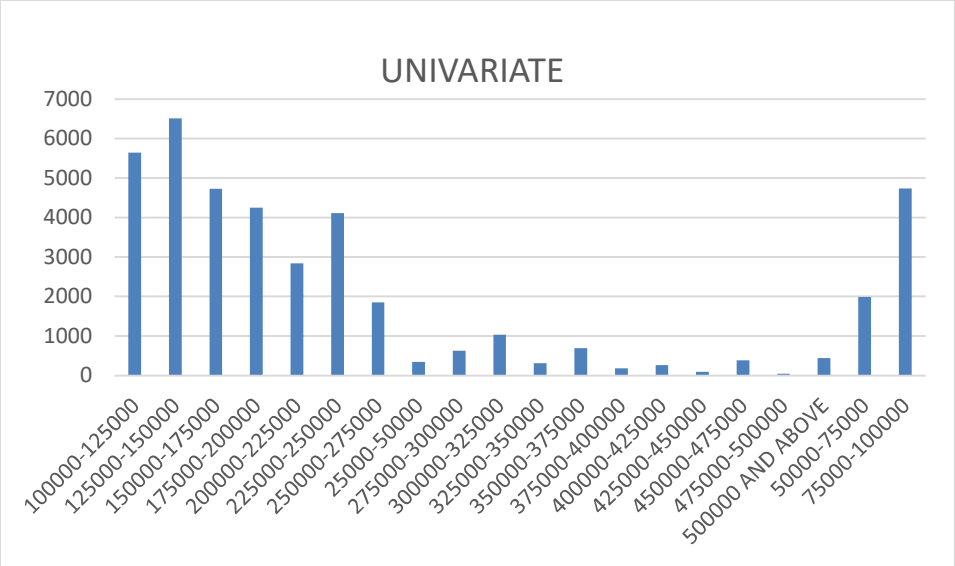


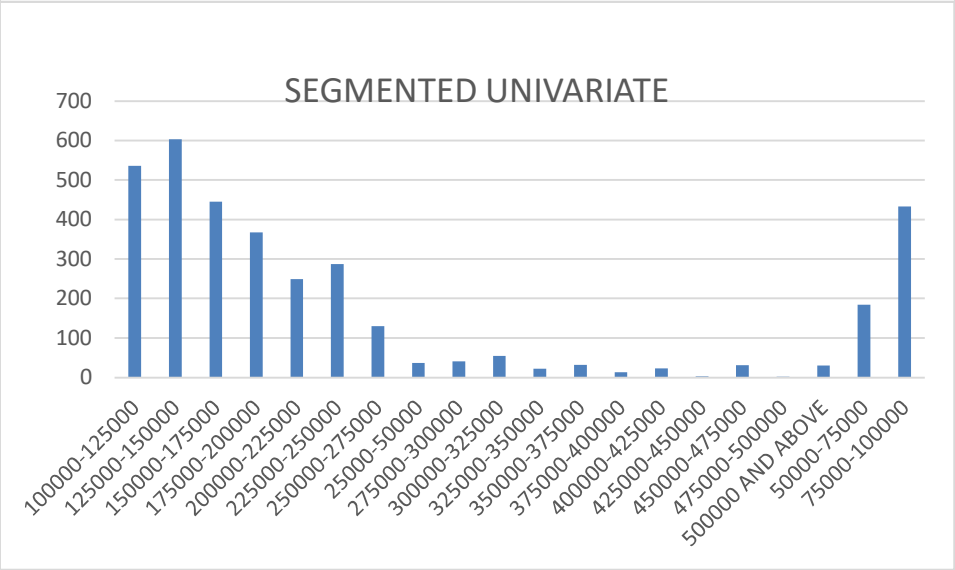
Inference: the distribution of target values has huge difference.

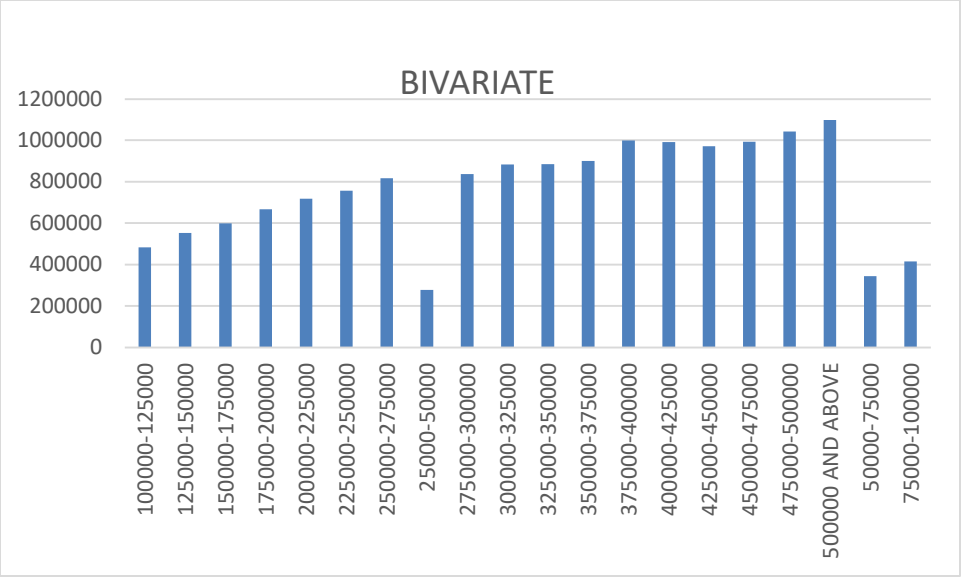
D. **Perform Univariate, Segmented Univariate, and Bivariate Analysis:** To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

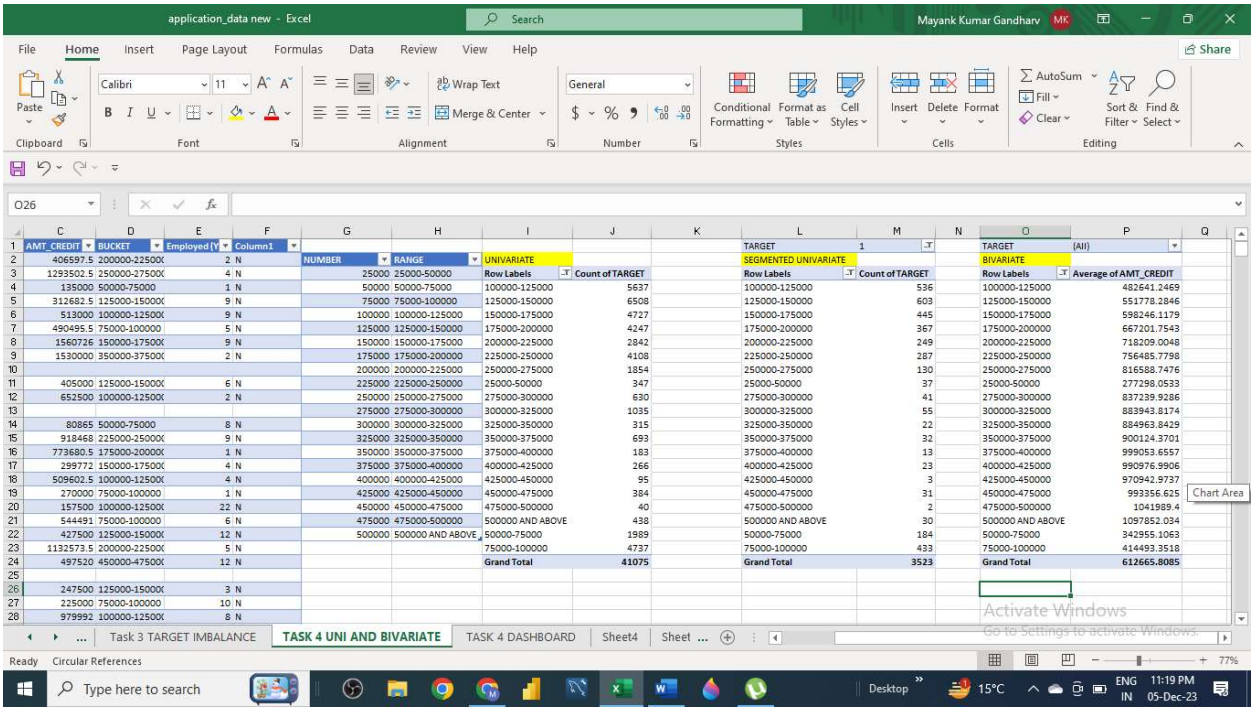
To perform this task, following steps were taken:

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.


UNIVARIATE
7000
6000
5000
4000
3000
2000
1000
0
100000-125000
125000-150000
150000-175000
175000-200000
200000-225000
225000-250000
250000-275000
25000-50000
275000-300000
300000-325000
325000-350000
350000-375000
375000-400000
400000-425000
425000-450000
450000-475000
475000-500000
500000 AND ABOVE
50000-75000
75000-100000

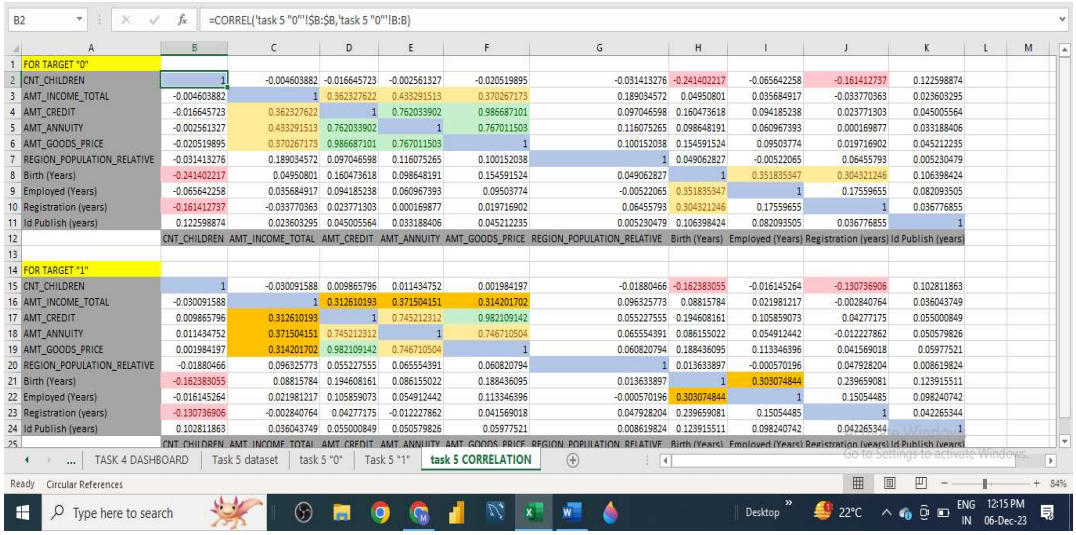

SEGMENTED UNIVARIATE
700
600
500
400
300
200
100
0
100000-125000
125000-150000
150000-175000
175000-200000
200000-225000
225000-250000
250000-275000
25000-50000
275000-300000
300000-325000
325000-350000
350000-375000
375000-400000
400000-425000
425000-450000
450000-475000
475000-500000
500000 AND ABOVE
50000-75000
75000-100000


BIVARIATE
1200000
1000000
800000
600000
400000
200000
0
100000-125000
125000-150000
150000-175000
175000-200000
200000-225000
225000-250000
250000-275000
25000-50000
275000-300000
300000-325000
325000-350000
350000-375000
375000-400000
400000-425000
425000-450000
450000-475000
475000-500000
500000 AND ABOVE
50000-75000
75000-100000



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.



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