BANK LOAN CASE STUDY

FINAL PROJECT II

DONE BY:G.REVAN

PROJECT DESCRIPTION

This project aims to analyze the loan risk of bank. When the company receives a loan application, the company has to decide for loan approval on the based-on-application profile. Two types of risks associated with the banks decision:

- If the applicant is likely to reply the loan ,then not approving the loan result in a loss of the business to company.
- If the applicant is not likely to repay the loan, i.e; he/she is likely to defaulter, then approving the loan may lead to financial loss for the company.

The data given below contains the information about the loan application at the time of applying the loan.It contains two types of scenarios:

- ❑ The client with payment difficulties: He/She has done late payment of more than X days at least one of the first Y installments of the loan in our sample.
- ❑ All other cases: All other cases when the payment is paid on time.

PROJECT DESCRIPTION

A bank loan case study is given which involves 3 datasets.

- 1. `application_data.csv` contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- 2. `previous_application.csv` contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- 3. `columns_descrption.csv` is data dictionary which describes the meaning of the variables.

We are supposed to use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Approach Used:

- I have downloaded the application_data, previous_data and columns_descriptions csv files and load them into MSExcel .
- All the columns details are read from Columns_description.csv file and undersood them clearly for further analysis.
- Next ,I have done EDA on the data for giving the solutions for the tasks.

TECH_STACK USED: MSEXCEL,POWERPOINT

TASK 1 : Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

DATA CLEANING

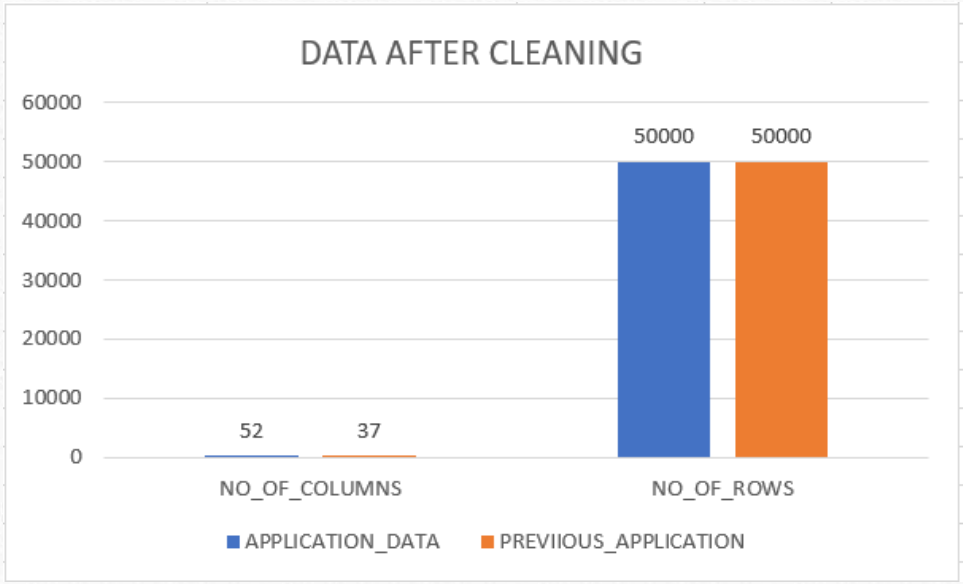
'previous_application.csv': There are 37 columns and 50000 rows.
'application_data.csv': There are 122 columns and 50000 rows.

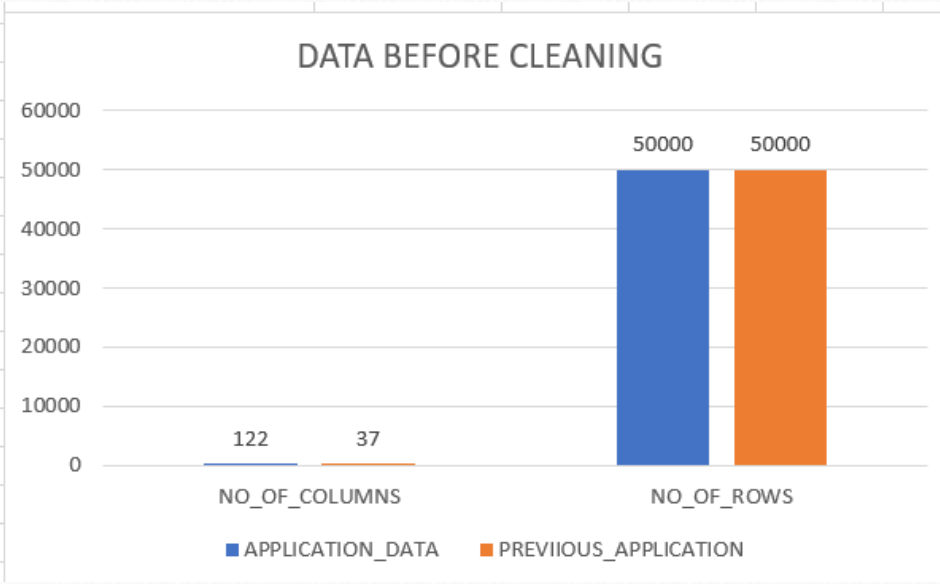
Handling Duplicates: There are no duplicates in the given datasets.
Handling Missing values: On both the datasets , I have used COUNT function to count the total rows and found the percentage of null values for each column using (blank rows in column/total rows)*100

Deleting columns:I have deleted the columns having more than 30% of null values in that column and the columns having less than 30% are used to perform the distribution statistics i.e;mean,median to find missing values for the data.

Replacing null values: I have replaced blank cells with the mean values of that columns.

INSIGHTS

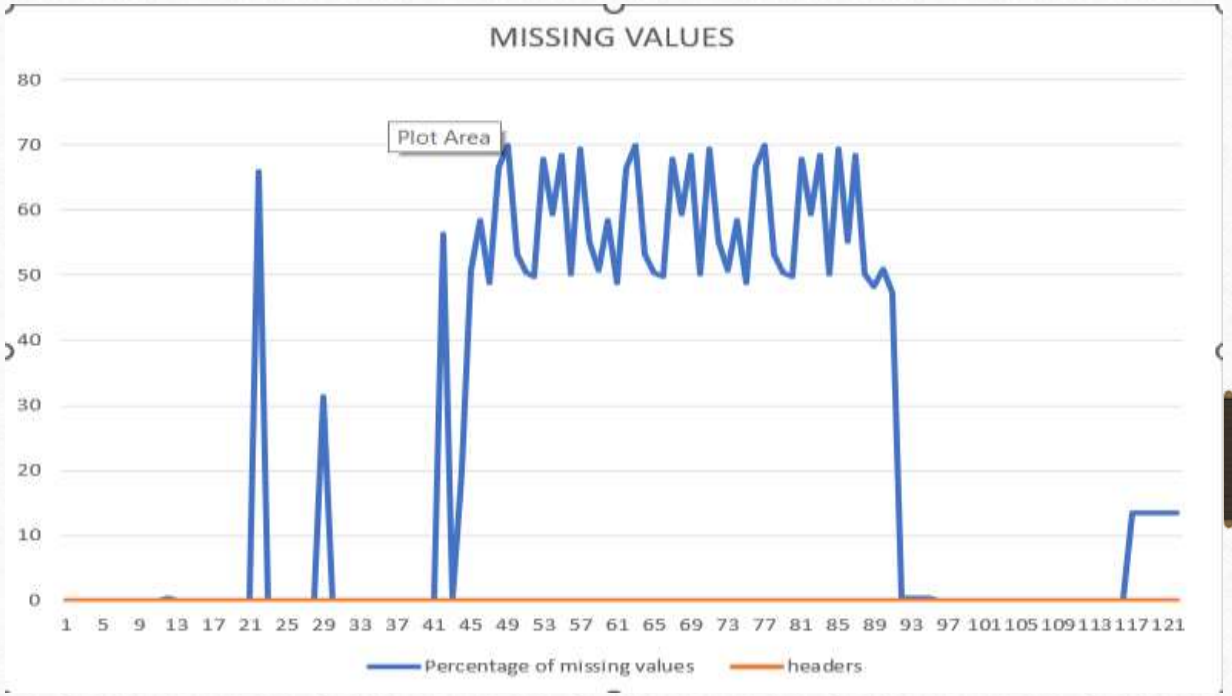




- The above figures describes the NO_OF-COLUMNS to NO_OF_ROWS ratio with respective to the data before and after cleaning.
- In PREVIOUS_APPLICATION the rows and columns are not changed whereas APPLICATION_DATA has changed for the columns.

INSIGHTS

Table	Row
application_data	SK_ID_CURR
application_data	NAME_CONTRACT_TYPE
application_data	AMT_CREDIT
application_data	AMT_ANNUITY
application_data	AMT_GOODS_PRICE
application_data	NAME_TYPE_SUITE
application_data	WEEKDAY_APPR_PROCESS_START
application_data	HOUR_APPR_PROCESS_START
previous_application.csv	SK_ID_CURR
previous_application.csv	NAME_CONTRACT_TYPE
previous_application.csv	AMT_ANNUITY
previous_application.csv	AMT_CREDIT
previous_application.csv	AMT_GOODS_PRICE
previous_application.csv	WEEKDAY_APPR_PROCESS_START
previous_application.csv	HOUR_APPR_PROCESS_START
previous_application.csv	NAME_TYPE_SUITE



- The first figure describes the same columns in both the datasets .
- The second figure describes the percentage of missing values in respective headers .

Task 2: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

- An outlier refers to a data point that significantly deviates from the rest of the data in a dataset.
- Outliers can arise due to various reasons such as measurement errors, data entry errors, natural variability in the data, or even genuine extreme observations.

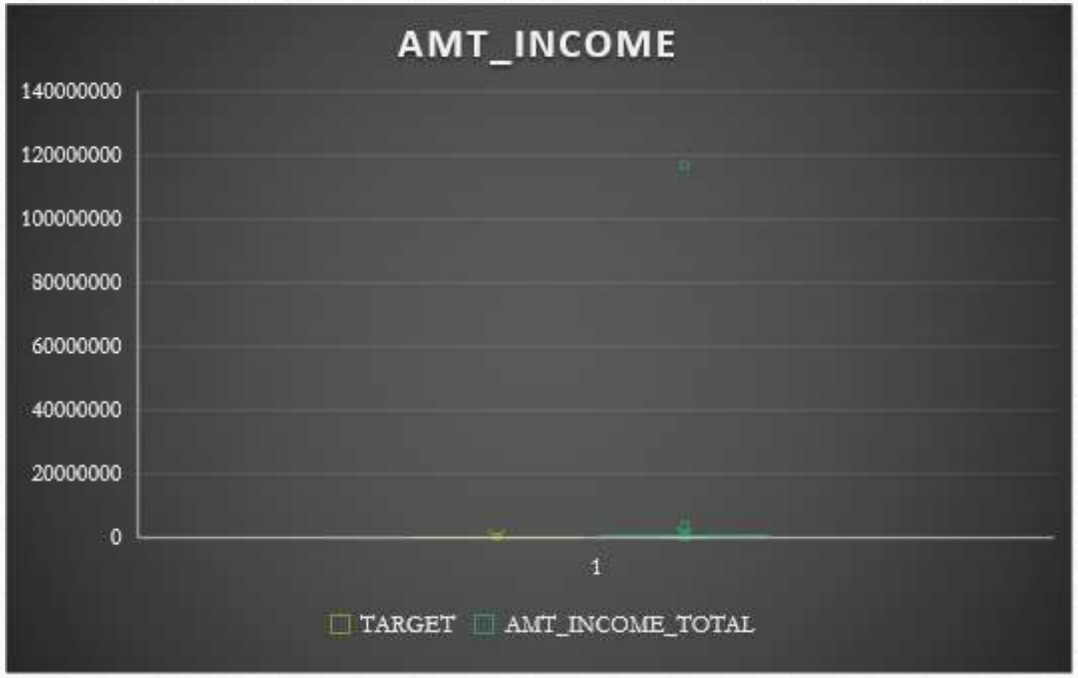
➢ Outliers can only be identified on Numeric variables. To identify outliers in Excel, follow these simple steps:

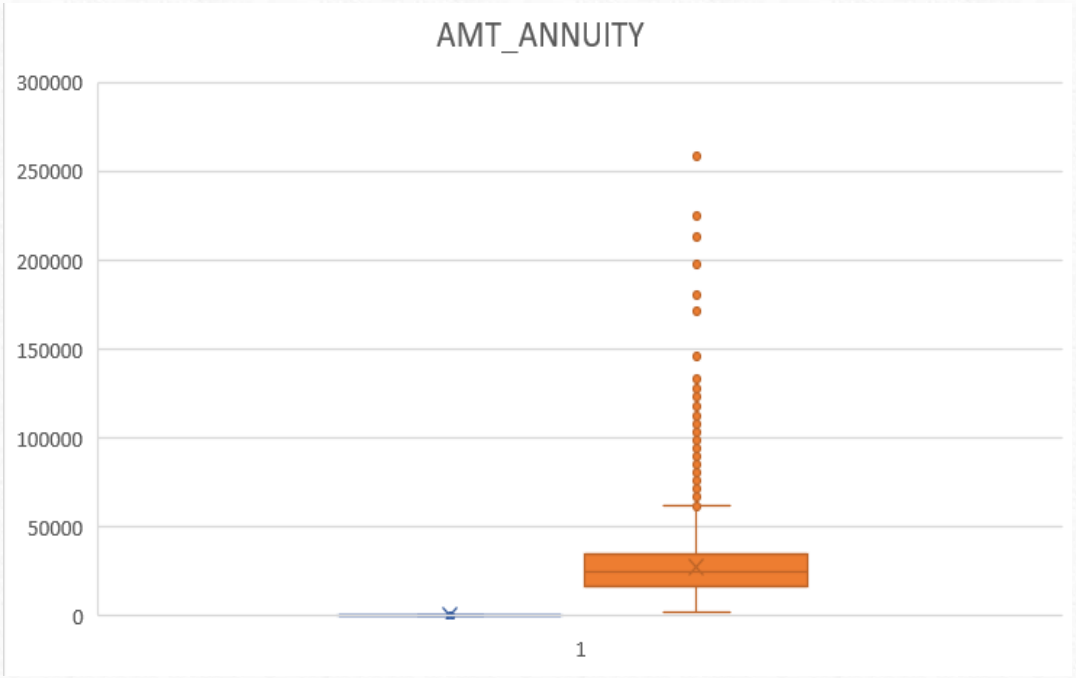
- Calculate the interquartile range (IQR) by subtracting the first quartile (25th percentile) from the third quartile (75th percentile) using the formula: "=QUARTILE.INC(range,3) – QUARTILE.INC(range,1)".
- Determine the lower bound for outliers by subtracting 1.5 times the IQR from the first quartile using the formula: "=QUARTILE.INC(range,1) – (1.5 * IQR)".
- Determine the upper bound for outliers by adding 1.5 times the IQR to the third quartile using the formula: "=QUARTILE.INC(range,3) + (1.5 * IQR)".

.

OUTLIERS

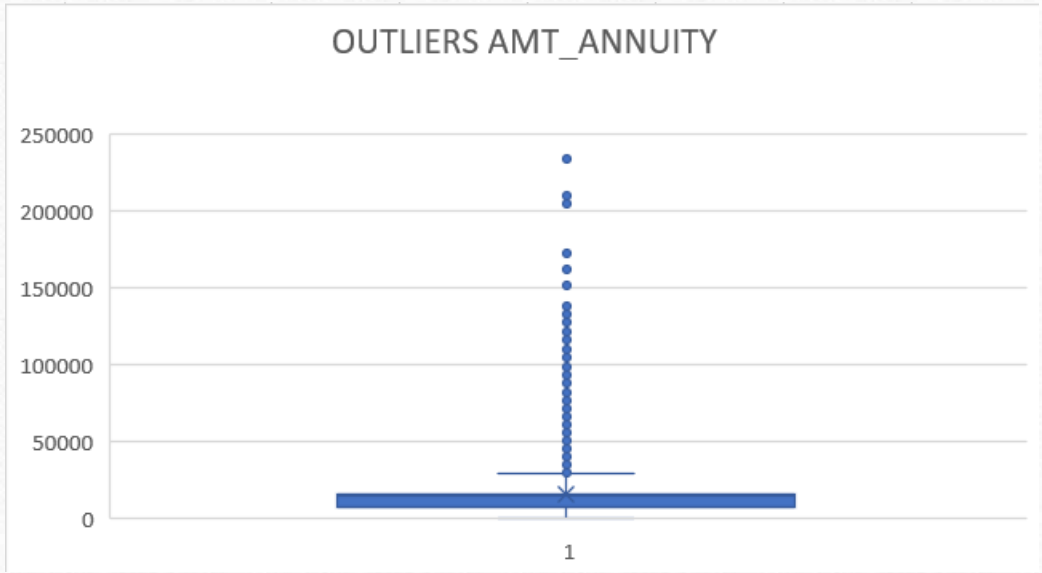
APPLICATION_DATA					
COLUMN NAME	Q1	Q3	IQR	UPPERBOUND	LOWERBOUND
AMT_ANNUITY	16456.5	34596	18139.5	61805.25	-10752.75
AMT_INCOME_TOTAL	112500	202500	90000	337500	-22500
AMT_CREDIT	270000	808650	538650	1616625	-537975
AMT_GOODS_PRICE	238500	679500	441000	1341000	-423000

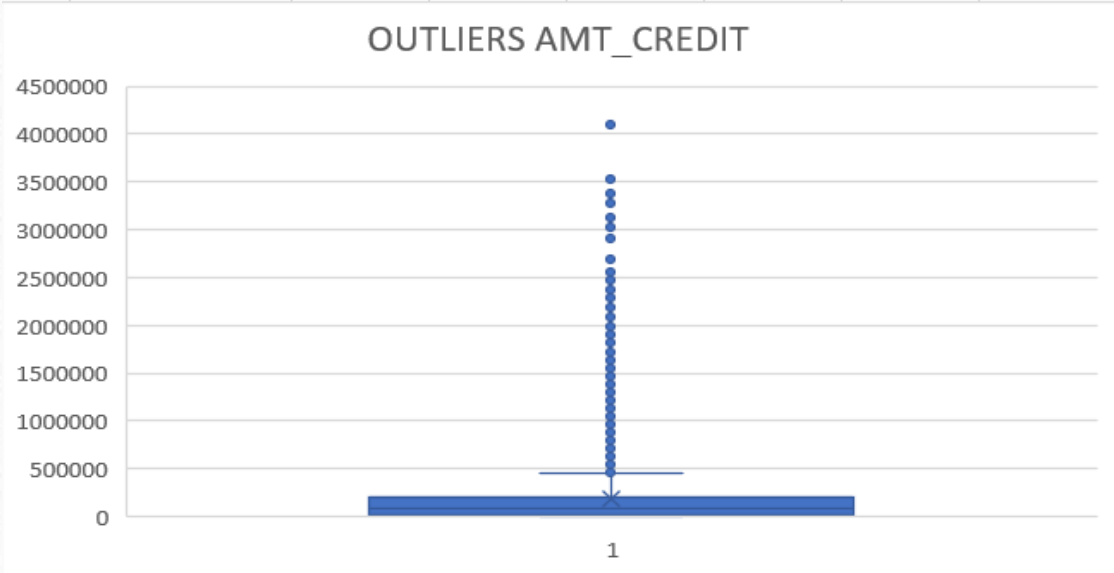




OUTLIERS

PREVIOUS_APPLICATION					
COLUMN NAME	Q1	Q3	IQR	UPPER BOUND	LOWER BOUND
AMT_ANNUITY	7189.74	16256.16	9066.42	29855.79	2656.53
APPLICATION_AMT	22045.5	180000	157954.5	416931.75	-56931.75
AMT_CREDIT	26055	198105.8	172050.8	456181.875	-59970.375
AMT_GODS_PRICE	63663.75	215141.4	151477.7	442357.9183	-12075.08365





INSIGHTS

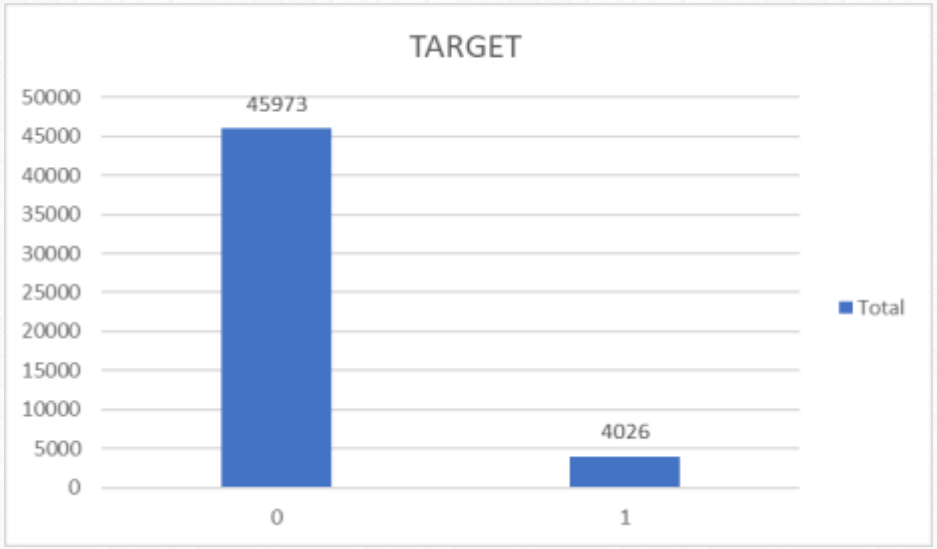
- In both datasets outliers are present.

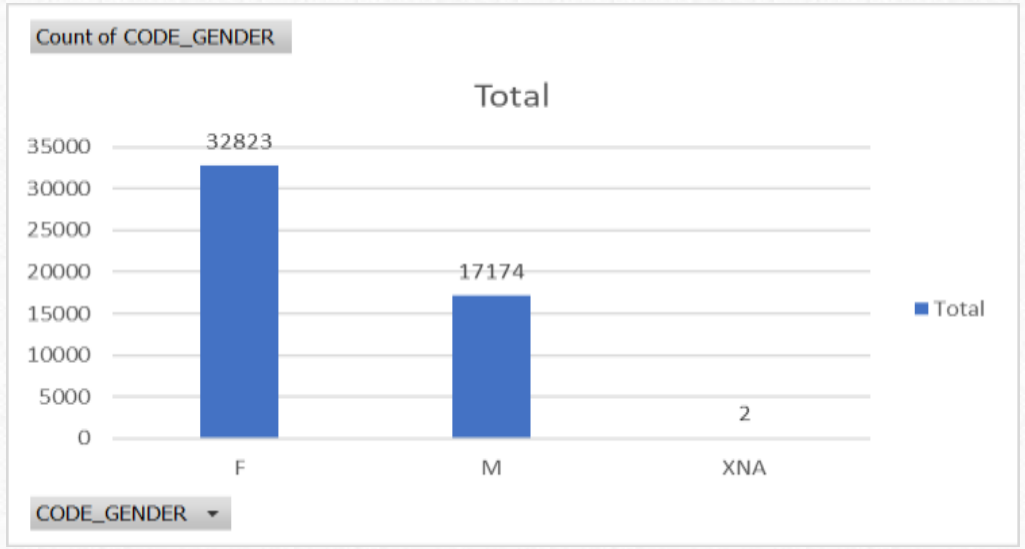
Task 3: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

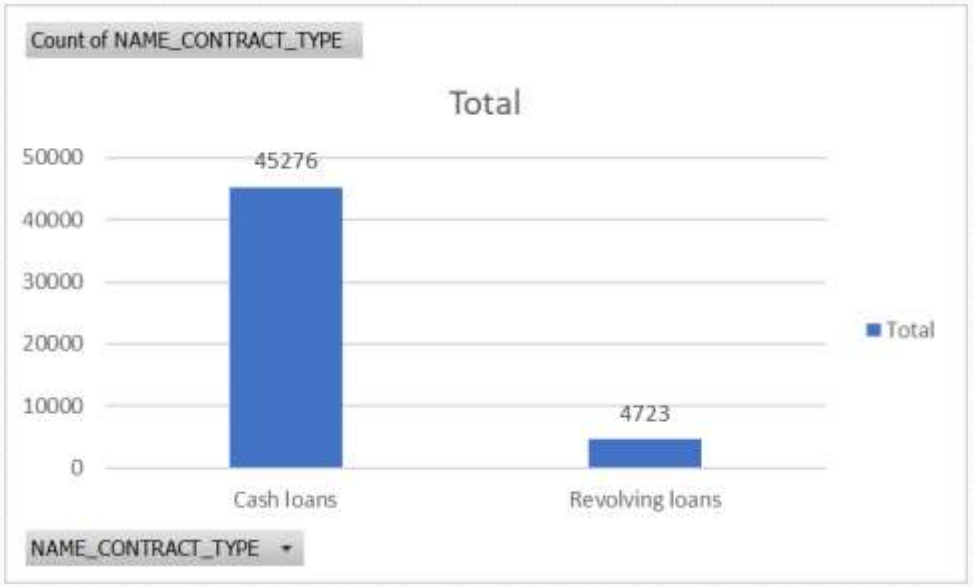
Data imbalance occurs when data is disseminated in an unequal manner. Data imbalance in Excel refers to an uneven distribution of data across different categories or classes. It can cause issues when analyzing or modeling the data, especially if one class is significantly underrepresented compared to others.

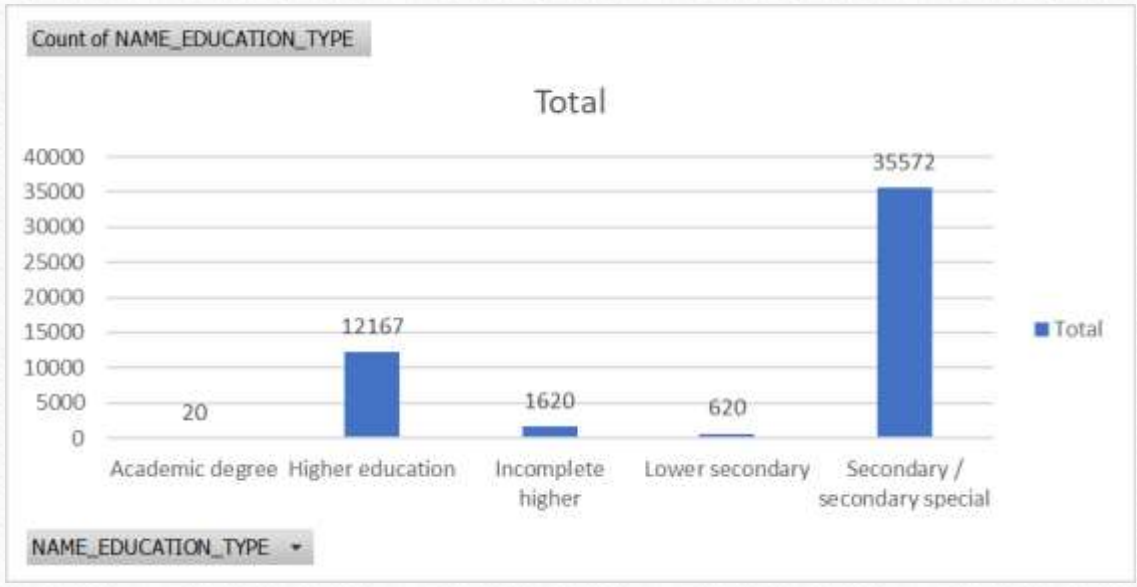
Data imbalance can be effectively visualized and analyzed using Pivot charts in Excel. Here's a step-by-step:

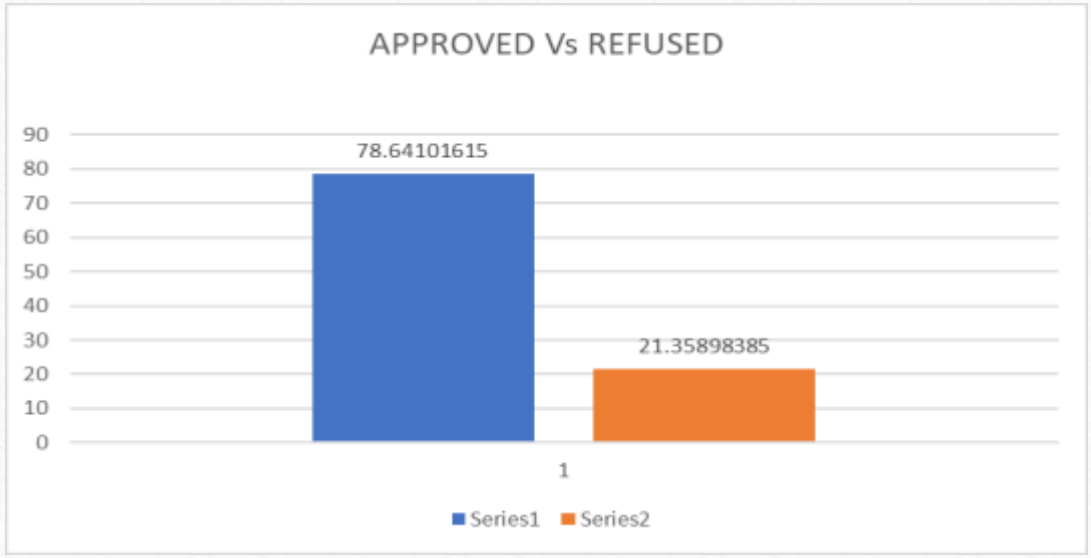
- Select the dataset in Excel that contains the categorical variable you want to assess for data imbalance.
- Navigate to the "Insert" tab in the Excel ribbon and click on "PivotTable." A dialog box will appear.
- Specify the range of your dataset and choose the destination for the Pivot Table, such as a new worksheet.
- In the PivotTable Field List, drag the categorical variable representing the class or category you want to analyze into the "Rows" area.
- Drag the same categorical variable into the "Values" area. Excel will automatically calculate the count of occurrences for each category.
- Create a Pivot chart based on the PivotTable. With the PivotTable selected, go to the "Insert" tab and choose the desired chart type, such as a bar chart, column chart, or pie chart.

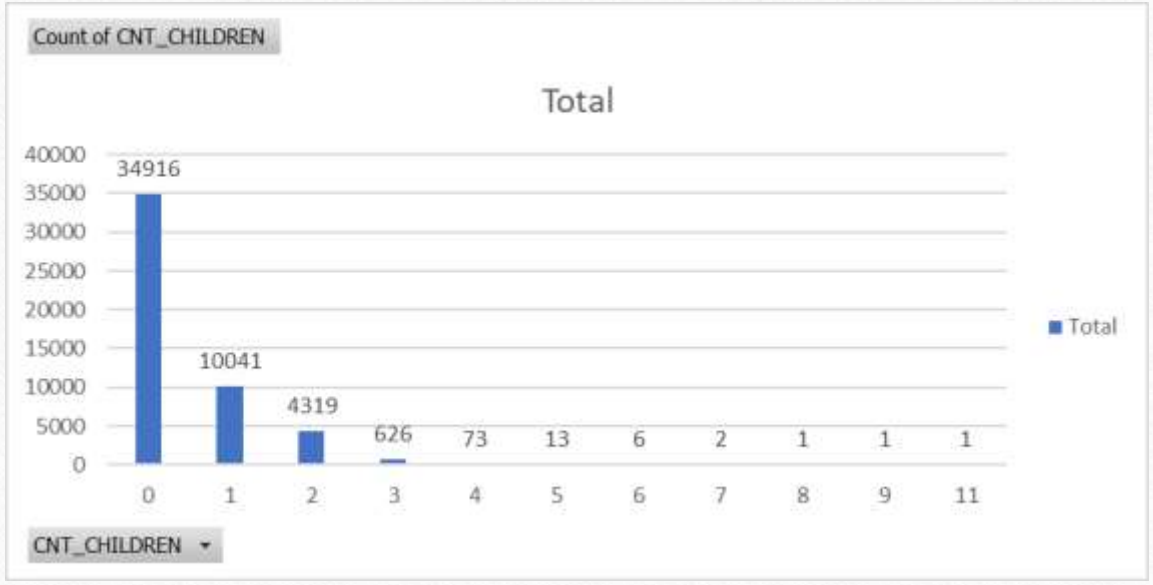

TARGET
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
45973
4026
0
1
Total


Count of CODE_GENDER
Total
35000
30000
25000
20000
15000
10000
5000
0
32823
17174
2
F
M
XNA
Total
CODE_GENDER


Count of NAME_CONTRACT_TYPE
Total
50000
40000
30000
20000
10000
0
45276
4723
Cash loans
Revolving loans
Total
NAME_CONTRACT_TYPE


Count of NAME_EDUCATION_TYPE
Total
40000
35000
30000
25000
20000
15000
10000
5000
0
20
12167
1620
620
35572
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
Total
NAME_EDUCATION_TYPE





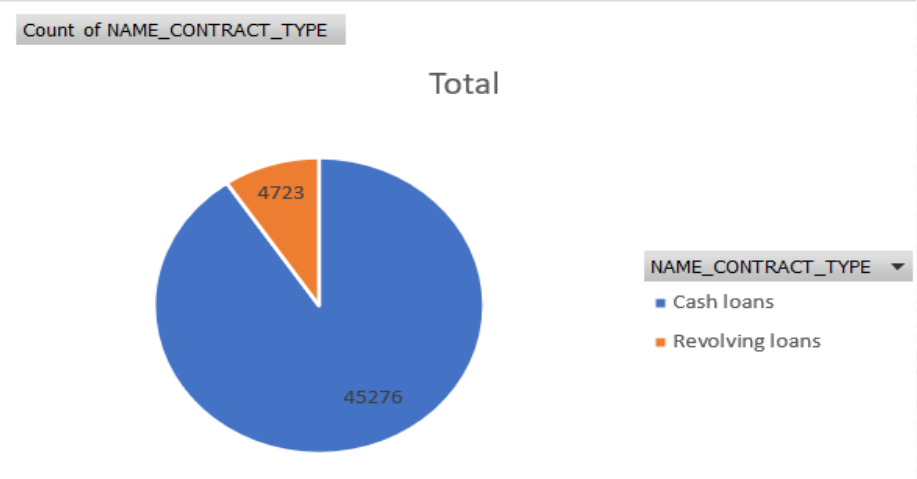
INSIGHTS:

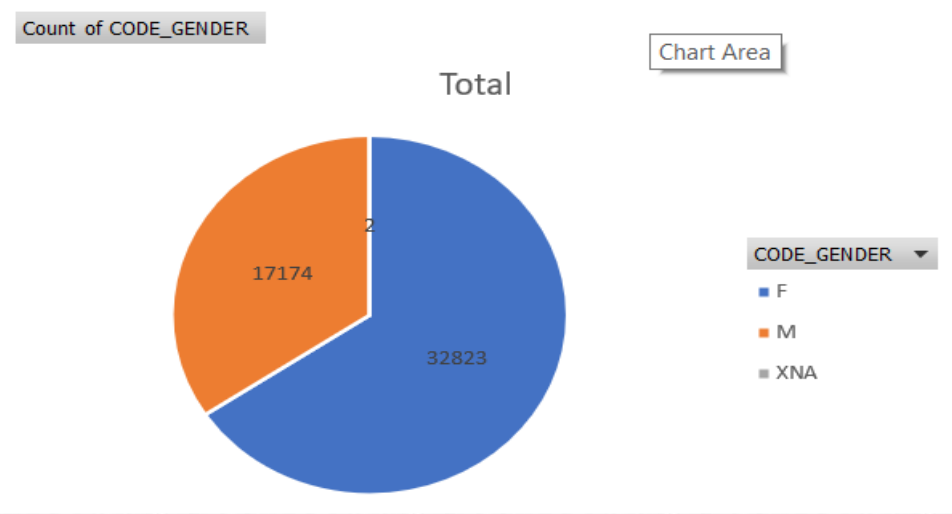
The charts shows that there is a high imbalance data in the applications.

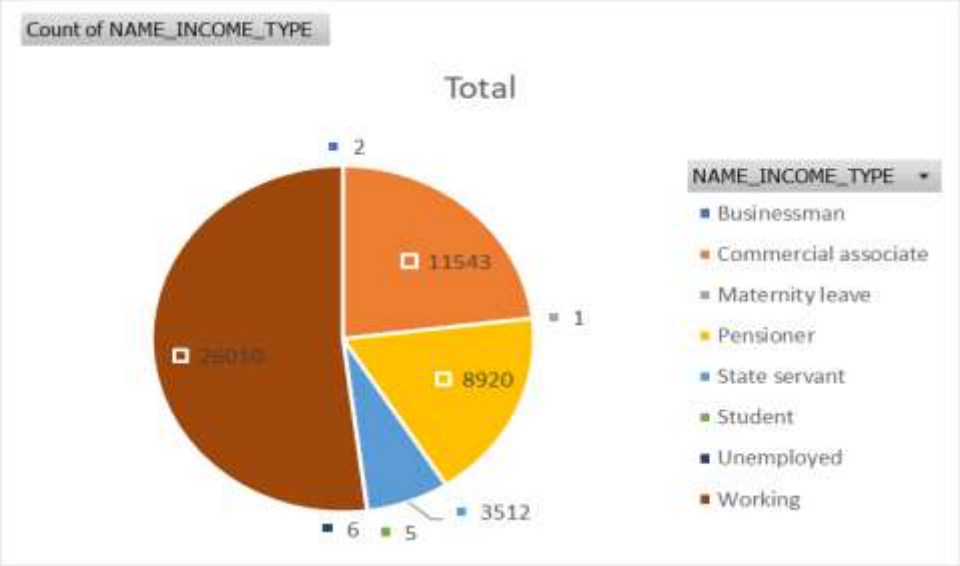
TASK 4:Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

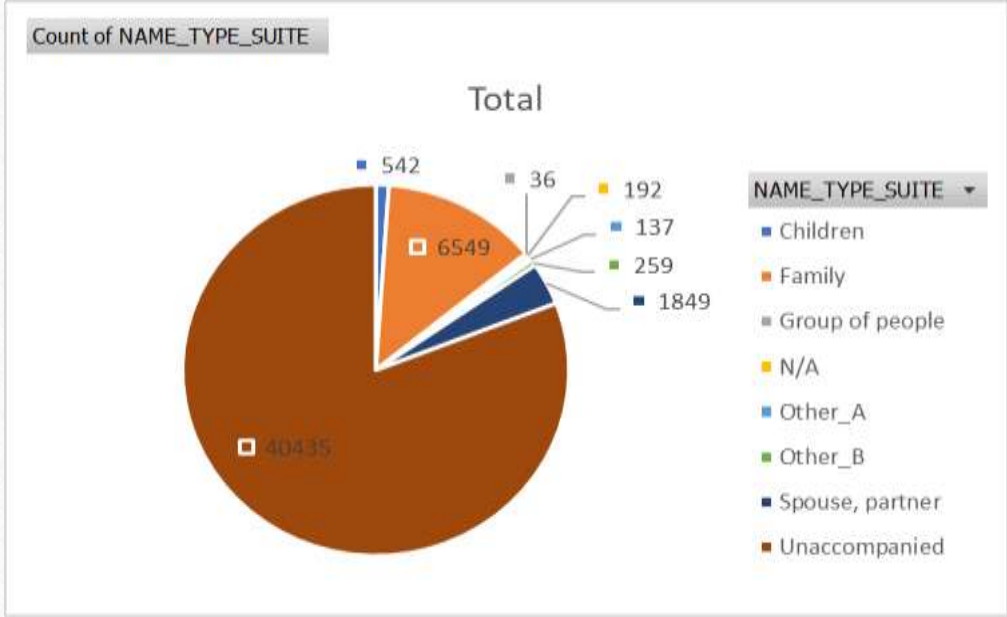
- Univariate analysis is a statistical approach that focuses on analyzing and interpreting data for a single variable without considering its relationship with other variables. It involves examining the distribution, measures of central tendency, variability, and other characteristics of the variable to gain insights and understand its behavior.
- Descriptive Statistics is found for numerical data which includes mean, median, standard deviation, mode, sample variance, kurtosis, skewness, range, max, min, sum, count, largest (2), smallest (2).
- Pie chart is used for representing categorical data.
- Here we have considered mainly the common variables among both datasets and some other ones of relevance, which are shown in the coming slides.
- Bivariate analysis is a statistical technique that examines the relationship between two variables. It explores how changes in one variable correspond to changes in another variable.
- The purpose is to understand the association, patterns, and dependencies between the two variables.
- Bivariate analysis aids in making informed decisions and predicting outcomes based on the observed relationship between variables.

UNIVARIATE ANALYSIS(APPLICATION_DATA)

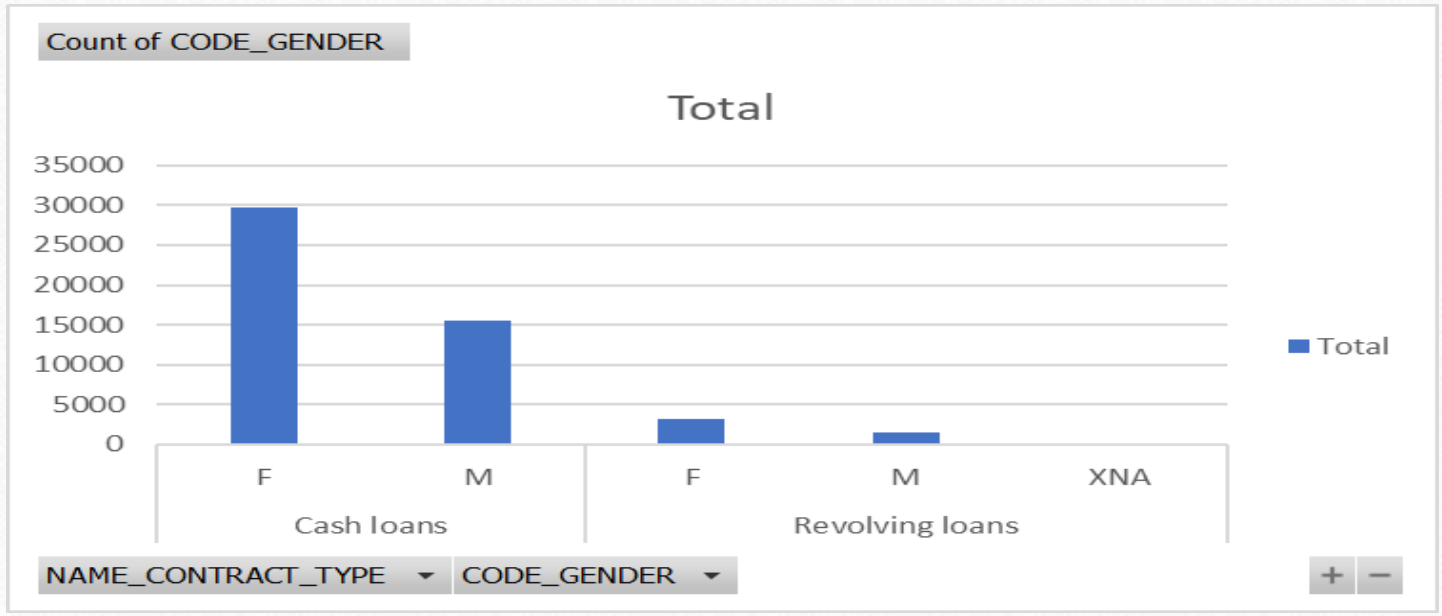






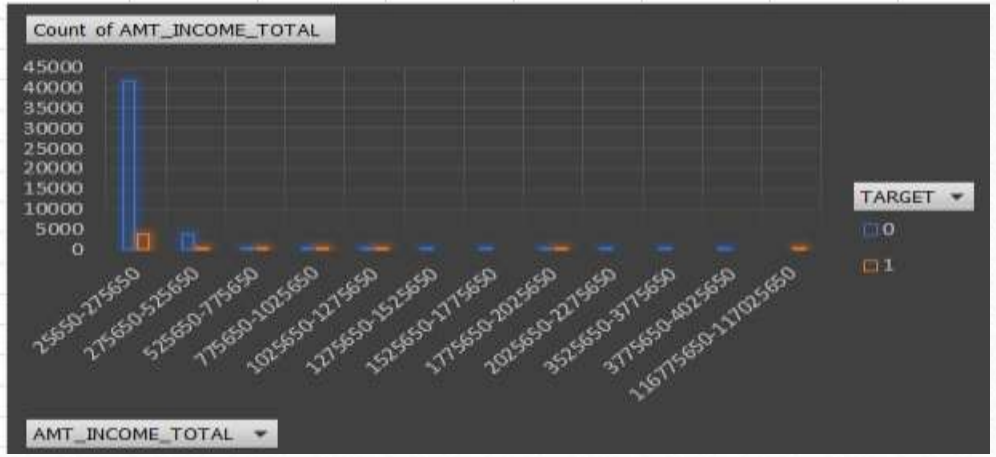


SEGMENTED UNIVARIATE

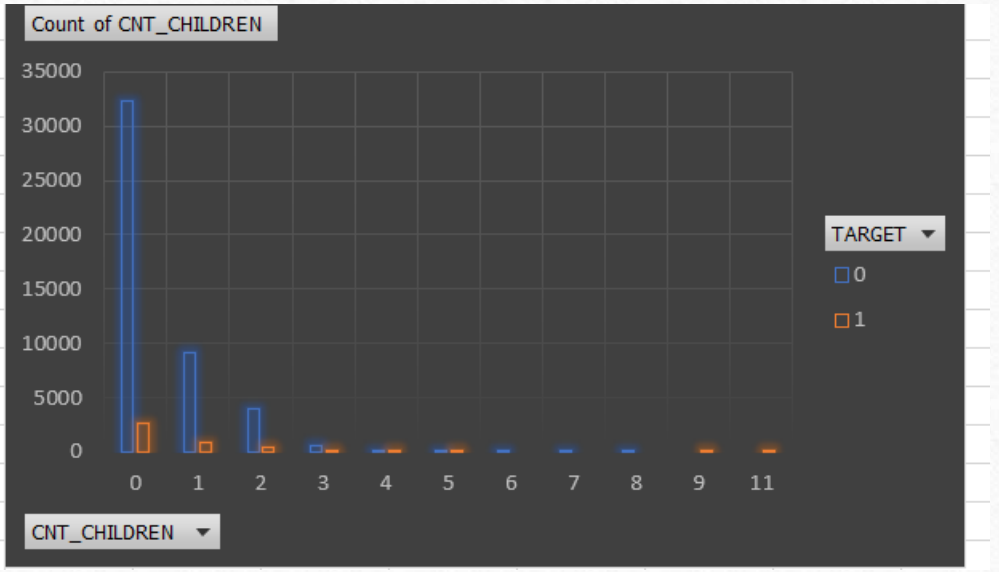


BIVARIATE ANALYSIS

Count of AMT_INCOME_TOTAL	Column Labels		
AMT_INCOME_TOTAL	0	1	Grand Total
25650-275650	41780	3752	45532
275650-525650	3779	243	4022
525650-775650	322	24	346
775650-1025650	55	4	59
1025650-1275650	16	1	17
1275650-1525650	11		11
1525650-1775650	1		1
1775650-2025650	5	1	6
2025650-2275650	2		2
3525650-3775650	1		1
3775650-4025650	1		1
116775650-117025650		1	1
Grand Total	45973	4026	49999



Count of CNT_CHILDREN	Column Labels		
Row Labels	0	1	Grand Total
0	32272	2644	34916
1	9118	923	10041
2	3935	384	4319
3	570	56	626
4	59	14	73
5	10	3	13
6	6		6
7	2		2
8	1		1
9		1	1
11		1	1
Grand Total	45973	4026	49999



INSIGHTS

- More number of females are there in gender for Application_data
- Working class people are higher.
- Unaccompanied NAME_TYPE of people are higher.
- CASH_LOANS are higher than resolving loans.

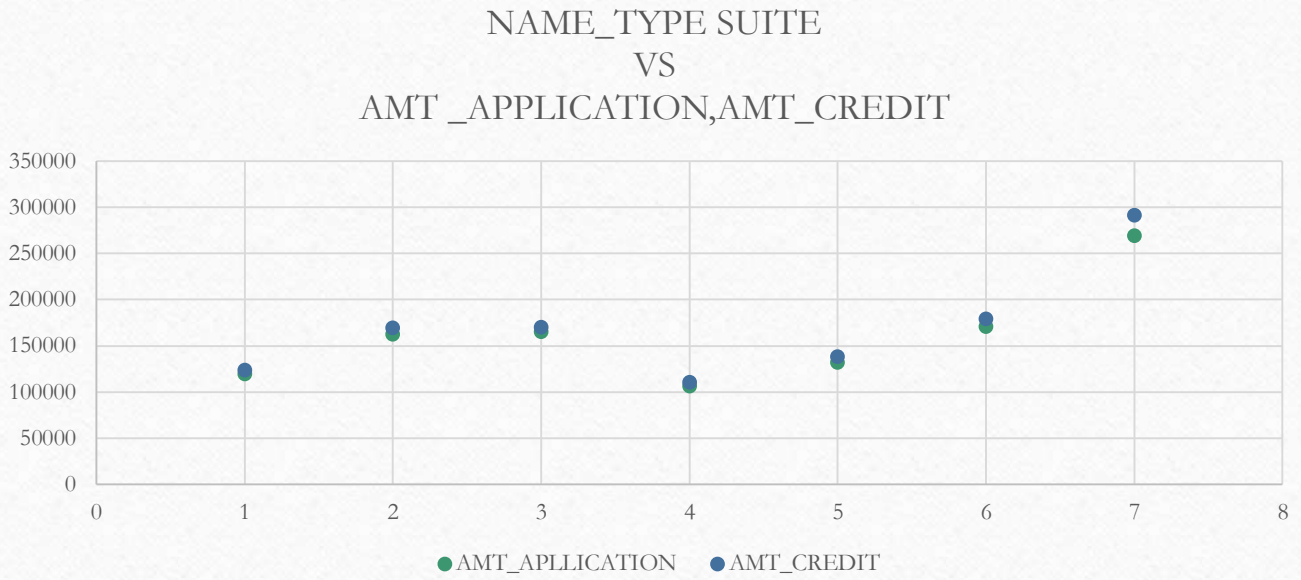
Task 5: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

- Correlation measures the relationship between two variables. It shows how they change together: positively, negatively, or not at all. The correlation coefficient, denoted as "r," ranges from -1 to 1.
- A coefficient of 1 means a perfect positive correlation, -1 means a perfect negative correlation, and 0 means no correlation. Correlation helps us understand the strength and direction of the relationship, but it doesn't imply causation. It's used to analyze data and predict outcomes in various fields.

CORREALTIONAL ANALYSIS:

An example of mean AMT_APPLICATION and AMT_CREDIT for different NAME_TYPE_SUITE.
An illustration using Scatter plot is also given.

NAME_TYPE	X-AXIS	AMT_APLLICATION	AMT_CREDIT
Children	1	119942.8998	124062.3603
Family	2	162632.5004	169499.4187
Group of people	3	165884.06	170155.125
Other_A	4	106861.4035	110788.0941
Other_B	5	132323.2253	138436.0328
Spouse, partner	6	171455.2867	179376.2682
Unaccompanied	7	269389.7817	292069.9744



	TARGET	T_CHILDRI	INCOME_T	MT_CREDI	MT_ANNUI	_GOODS_P	PULATIO	AYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	EXT_SOURCE_2	EXT_SOURCE_3	DAYS_LAST_PHONE_CHANGE
TARGET	1														
CNT_CHILDREN	0.02636	1													
AMT_INCOME_TOTAL	0.01089	0.00959	1												
AMT_CREDIT	-0.0324	0.00497	0.06932	1											
AMT_ANNUITY	-0.0124	0.02618	0.08301	0.7695	1										
AMT_GOODS_PRICE	-0.0413	0.00025	0.06988	0.98661	0.77402	1									
REGION_POPULATION_RELATIVE	-0.0408	-0.0256	0.02984	0.09511	0.11511	0.09916	1								
DAYS_BIRTH	0.07679	0.32926	0.016	-0.0593	0.00771	-0.0576	-0.0325	1							
DAYS_EMPLOYED	-0.0403	-0.2397	-0.0316	-0.0705	-0.1104	-0.0678	-0.0041	-0.6136	1						
DAYS_REGISTRATION	0.04234	0.18122	0.00995	0.00345	0.03322	0.0061	-0.0593	0.33363	-0.204680611	1					
DAYS_ID_PUBLISH	0.04693	-0.0321	0.00351	-0.0122	0.00672	-0.014	-0.0043	0.27083	-0.270382022	0.104298561	1				
CNT_FAM_MEMBERS	0.01299	0.88045	0.01123	0.064	0.07738	0.0616	-0.023	0.27724	-0.22981846	0.170110711	-0.026077905	1			
EXT_SOURCE_2	-0.1583	-0.0176	0.01952	0.13802	0.12879	0.14677	0.20093	-0.0938	-0.026097761	-0.060920774	-0.04748406	0.002664617	1		
EXT_SOURCE_3	-0.1596	-0.0395	-0.0215	0.03754	0.02093	0.041	-0.0085	-0.1872	0.10185073	-0.101681509	-0.11880614	-0.022910399	0.092231966	1	
DAYS_LAST_PHONE_CHANGE	0.05614	-0.002	-0.0048	-0.0762	-0.0673	-0.0797	-0.0478	0.08019	0.027517977	0.052146121	0.091373703	-0.022704641	-0.192414193	-0.070537374	1

THE TOP 10 VARIABLES WITH HIGHEST CORRELATION:

1.AMT_GOODS_PRICE,AMT_CREDIT-0.98661
2.CNT_FAMILY_MEMBERS,CNT_CHILDREN-0.880451437
3.AMT_GOODS_PRICE,AMT_ANNUITY-0.77402
4.AMT_ANNUITY,AMT_CREDIT-0.7695
5.DAYS_EMPLOYED,DAYS_BIRTH-0.6136
6.DAYS_REGISTRATION,DAYS_BIRTH-0.33363
7.DAYS_BIRTH,CNT_CHILDREN-0.32926
8.DAYS_EMPLOYED,CNT_CHILDREN-0.2397
9.DAYS_ID_PUBLISH,DAYS_EMPLOYED-0.270382022
10.CNT_FAM_MENBERS,DAYS_EMPLOYED -0.22981846

RESULTS:

- In this task,I have developed and improved the skills in performing data cleaning and analyzing different columns with different functionalities.
- I have found outliers for different columns and the imbalance data columns in the given dataset.
- The Univariate,segmented univariate and Bivariate analysis for same column with different categories.
- The correlation for a set of columns to check their correlation level.

DRIVE LINK FOR PROJECT:
https://drive.google.com/drive/folders/14BG_PbbxVlKgSmjHzaXuyrxAj87gkuGE?usp=drive_link