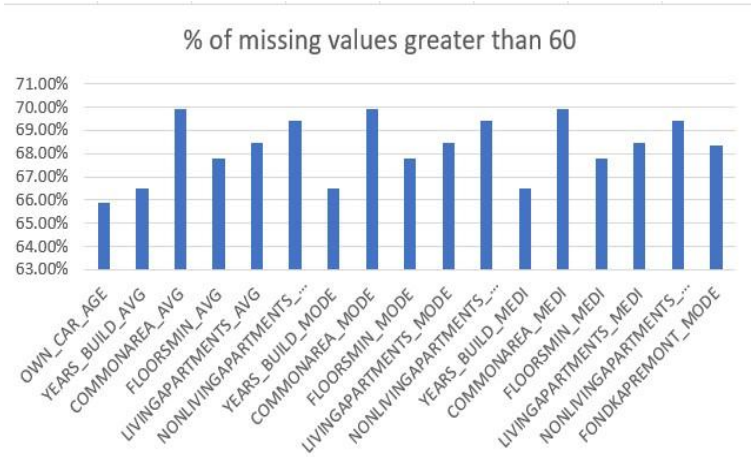
BANK LOAN CASE STUDY

Project Description: This project analyses loan application data to identify patterns that influence the likelihood of loan default. It also focuses on understanding the key factors and customer behaviour behind loan default so it can make better decisions about loan approval and rejection, reduce financial risks and ensure capable applicants are not rejected.

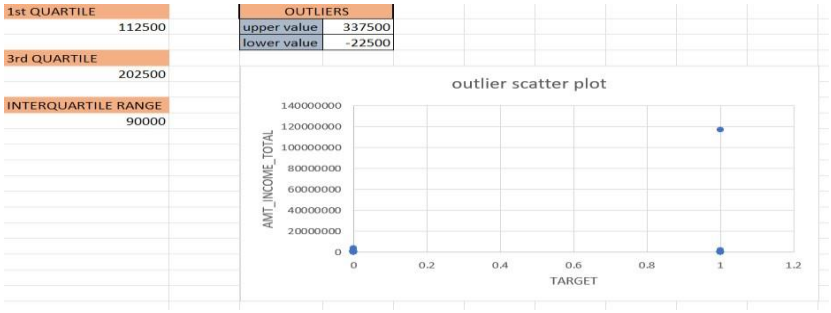
Approach: To perform the analysis of the Bank Loan Case Study datasets, I performed Exploratory Data Analysis focusing on handling missing data, filtering and filling missing values, deleting rows and deleting unnecessary columns especially the ones with more than 50% missing values. After that further Data Analysis tasks like identifying outliers, addressing data imbalance, performing univariate, segmented univariate, bivariate analysis and top correlations are carried out to uncover key factors driving loan default.

Data Analytics Tasks:

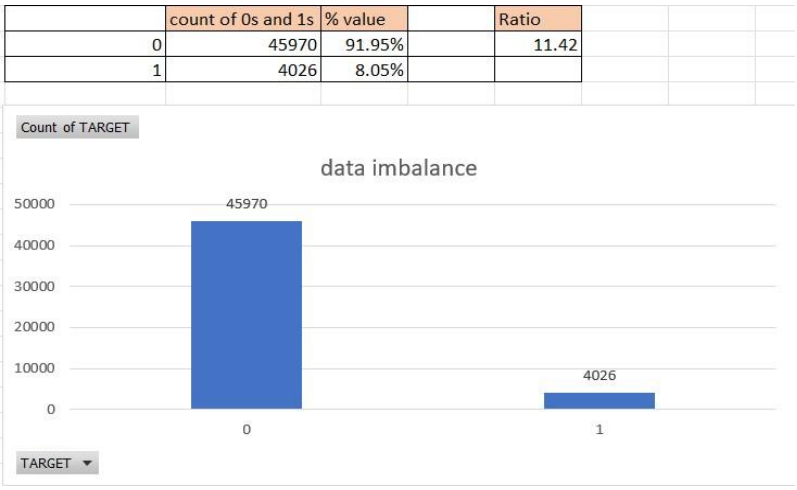
A. **Identify Missing Data:** The task involves identifying missing data in the dataset using countblank function and highlighting missing data for more than 60% using conditional formatting. For columns having missing values less than 40%, I imputed the numerical values using average formula and for categorical columns I simply put unknown.



B. **Identify Outliers in the Dataset:** The task involves detecting outliers in the dataset using excel functions like quartile and Interquartile Range on numerical columns.

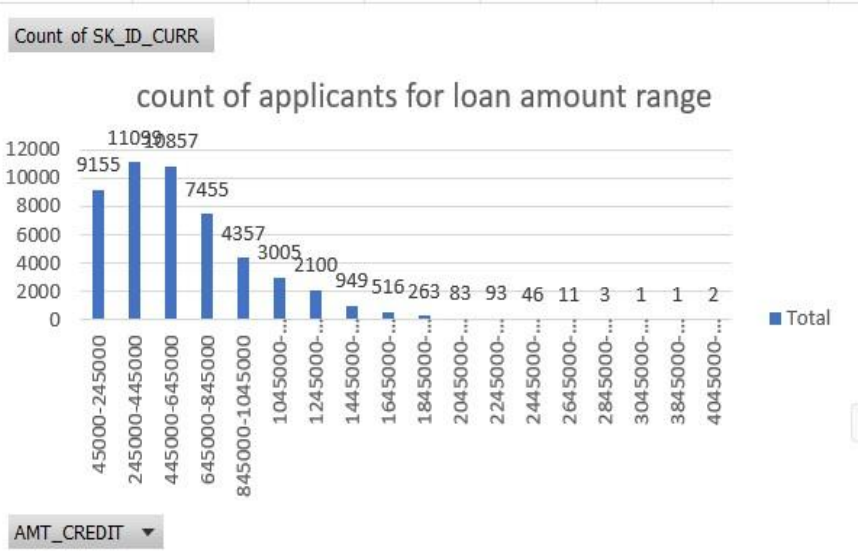


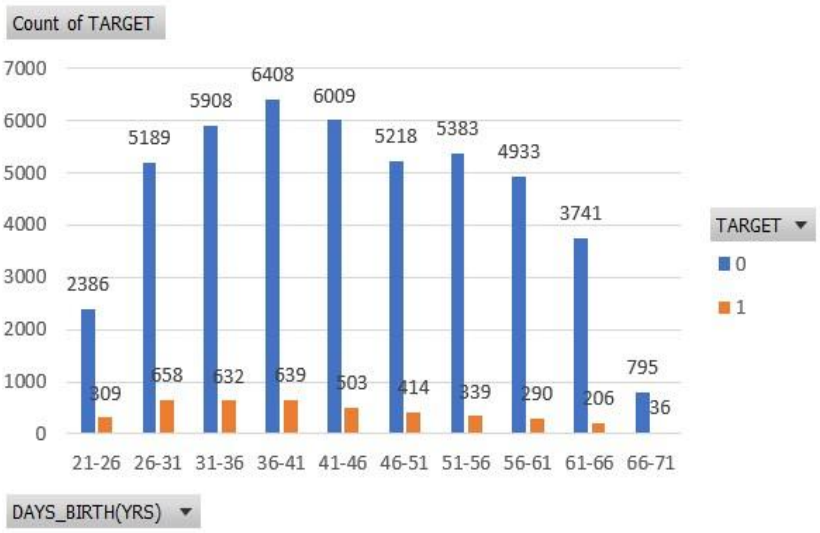
C. **Analyze Data Imbalance:** The task involves analysing data imbalance in the dataset and calculating the ratio of data imbalance using pivot tables and excel formulas.

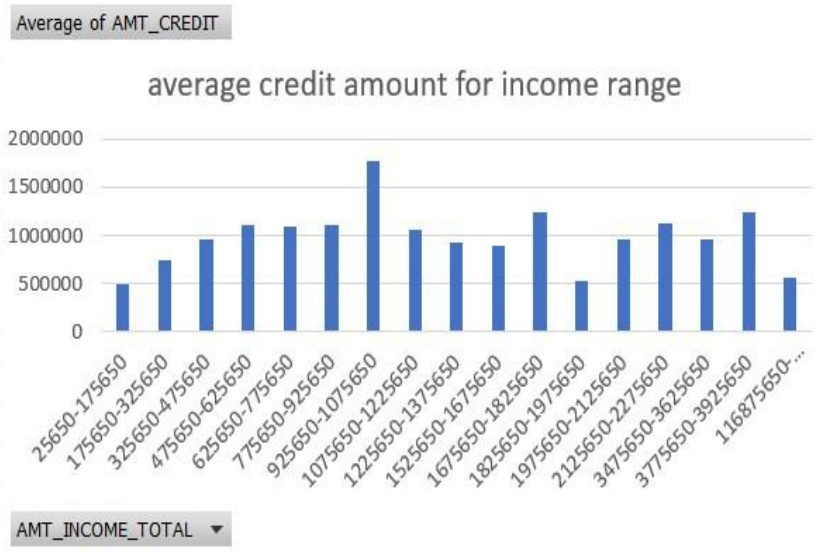


	count of 0s and 1s	% value		Ratio
0	45970	91.95%		11.42
1	4026	8.05%		

D. **Perform Univariate, Segmented Univariate and Bivariate Analysis:** The task involves analysis on consumer and loan attributes. For univariate analysis, distribution of individual variables like income range, loan amount range and age range are conducted. For segmented univariate analysis, variables like income range, loan amount range and age are compared with the target variable. For bivariate analysis, the income range is compared with the average credit amount using pivot tables and pivot charts.


Count of SK_ID_CURR
count of applicants for loan amount range
12000
10000
8000
6000
4000
2000
0
9155
11099
10857
7455
4357
3005
2100
949
516
263
83
93
46
11
3
1
1
2
45000-245000
245000-445000
445000-645000
645000-845000
845000-1045000
1045000-...
1245000-...
1445000-...
1645000-...
1845000-...
2045000-...
2245000-...
2445000-...
2645000-...
2845000-...
3045000-...
3845000-...
4045000-...
Total
AMT_CREDIT


Count of TARGET
7000
6000
5000
4000
3000
2000
1000
0
2386
309
5189
658
5908
632
6408
639
6009
503
5218
414
5383
339
4933
290
3741
206
795
36
21-26
26-31
31-36
36-41
41-46
46-51
51-56
56-61
61-66
66-71
TARGET
0
1
DAYS_BIRTH(YRS)


Average of AMT_CREDIT
average credit amount for income range
2000000
1500000
1000000
500000
0
25650-175650
175650-325650
325650-475650
475650-625650
625650-775650
775650-925650
925650-1075650
1075650-1225650
1225650-1375650
1525650-1675650
1675650-1825650
1825650-1975650
1975650-2125650
2125650-2275650
3475650-3625650
3775650-3925650
AMT_INCOME_TOTAL

E. **Identify Top Correlations for Different Scenarios:** The task involves identifying the top correlations for numerical variables in the dataset using the Excel correl function and visualizing the correlations using conditional formatting.

TOP CORRELATIONS	
CNT_CHILDREN	0.026354
AMT_INCOME_TOTAL	0.010897
AMT_CREDIT	-0.03243
REGION_POPULATION_RELATIVE	-0.0408
AGE	-0.07681

Tech-Stack Used: For this project, the software used is Microsoft Excel 2021 as it is a very popular tool for both beginners and advanced professionals. It is widely used for data analysis because of its accessibility, powerful functionality, versatility and ease of use.

Insights:

- Certain less significant variables having missing values were also removed to ensure data quality.
- Outliers were detected in variables like amount of income and number of children.
- The dataset was imbalanced with a higher proportion of non-default cases compared to default cases.
- As the age of applicants increases, the chances of defaulting decreases.
- Lowest income range has highest number of defaults.

Result: This project provided comprehensive insights into the Bank Loan Case Study dataset and helped me understand how key factors like customer attributes and loan attributes help us understand loan defaults and how we can make better decisions about loan approvals and loan rejections so that capable applicants are not rejected as to not lose business and incapable applicants do not get approved mistakenly to avoid financial loss.