BANK LOAN CASE STUDY

Final Project-2

By - Anindya Das

PROJECT DESCRIPTION

This project aims at analyzing the risk appetite of banks. When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

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

1. The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample.
2. • All other cases: All other cases when the payment is paid on time.

Based on the scenarios a detailed analysis must be conducted and insights needs to be drawn to help bank identify the pattern which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (too risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.

APPROACH

I have used COUNTA function to count the total rows in each column. After that I have found the percentage of null values in each column using the formula 1- (Total Row Counts for each columns / Total Row Counts). After that I have removed all the columns having null value percentages more than 30%. For column having less than 30% null value percentages I have done mean, median and mode imputations for the missing values for columns having null value percentages less than 30%. I have also found the outliers using interquartile range method considering relevant columns. After going through each column description, I have kept only relevant columns to bring out the insights. The columns having days are converted in to years by simply dividing the days by 365.Click on the below link to open the excel file. The excel file contain all the analysis.

TECH STACK USED

Purpose – All the analysis has been performed in excel. This tool is also used to create graphical representation of the results and to understand the result set better.

Microsoft Excel

DATA ANALYTICS TASKS:

❑ 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.

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

I have used COUNTA function to count the total rows in each column. After that I have found the percentage of null values in each column using the formula 1- (Total Row Counts for each columns / Total Row Counts). After that I have removed all the columns having null value percentages more than 30%. For column having less than 30% null value percentages I have done mean, median and mode imputations for the missing values for columns having null value percentages less than 30%. I have also found the outliers using interquartile range method considering relevant columns. After going through each column description, I have kept only relevant columns to bring out the insights. The columns having days are converted in to years by simply dividing the days by 365.Click on the below link to open the excel file. The excel file contain all the analysis.

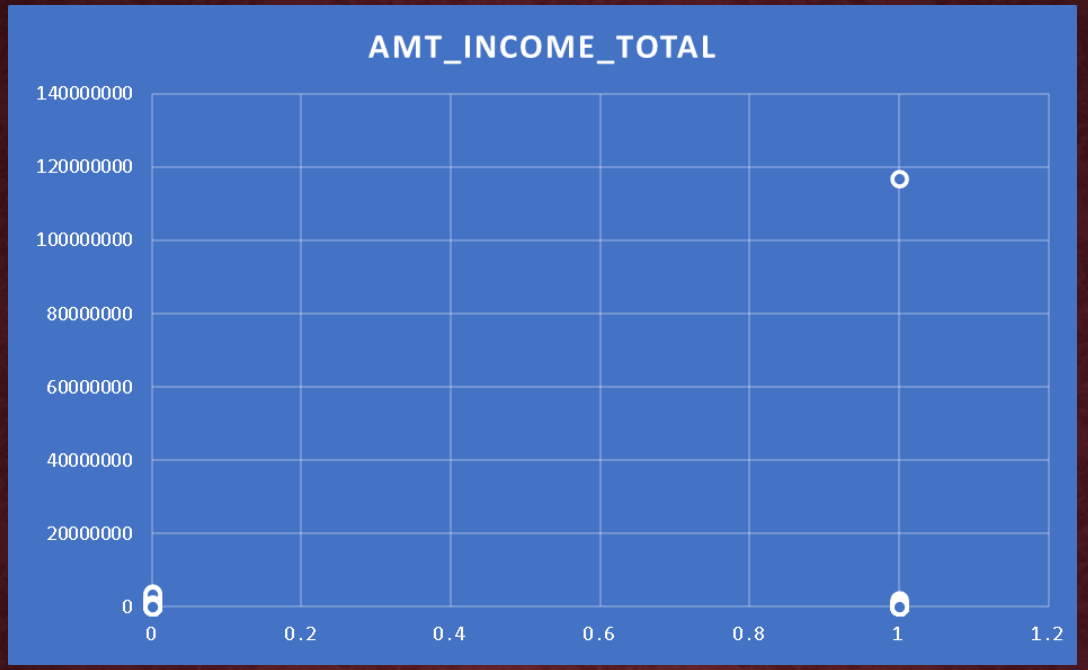
DATASET- https://drive.google.com/file/d/1Y4yv3pi3T293R-sB1WKlZzfUWVthGjOU/view?usp=drive_link

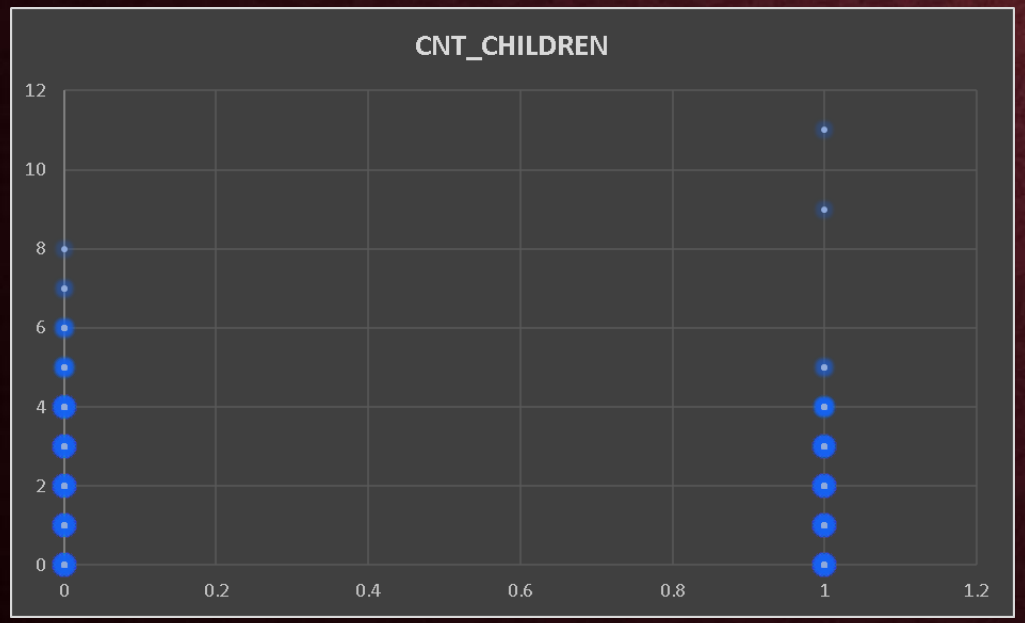
❑ Identify Outliers in the Dataset:

Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

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

Q1				
1				
Q2				
1.883135239				
Q3			OUTLIERS BELOW THE LOWER BOUND	0
3			OUTLIERS BELOW THE UPPER BOUND	1169
IQR			NO OF OUTLIERS	1169
2			MAX	25
LOWER BOUND	-2			
UPPER BOUND	6			


AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
0
0.2
0.4
0.6
0.8
1
1.2


CNT_CHILDREN
12
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
1.2


DAYS_EMPLOYED(YEARS)
1200
1000
800
600
400
200
0
0
0.2
0.4
0.6
0.8
1
1.2

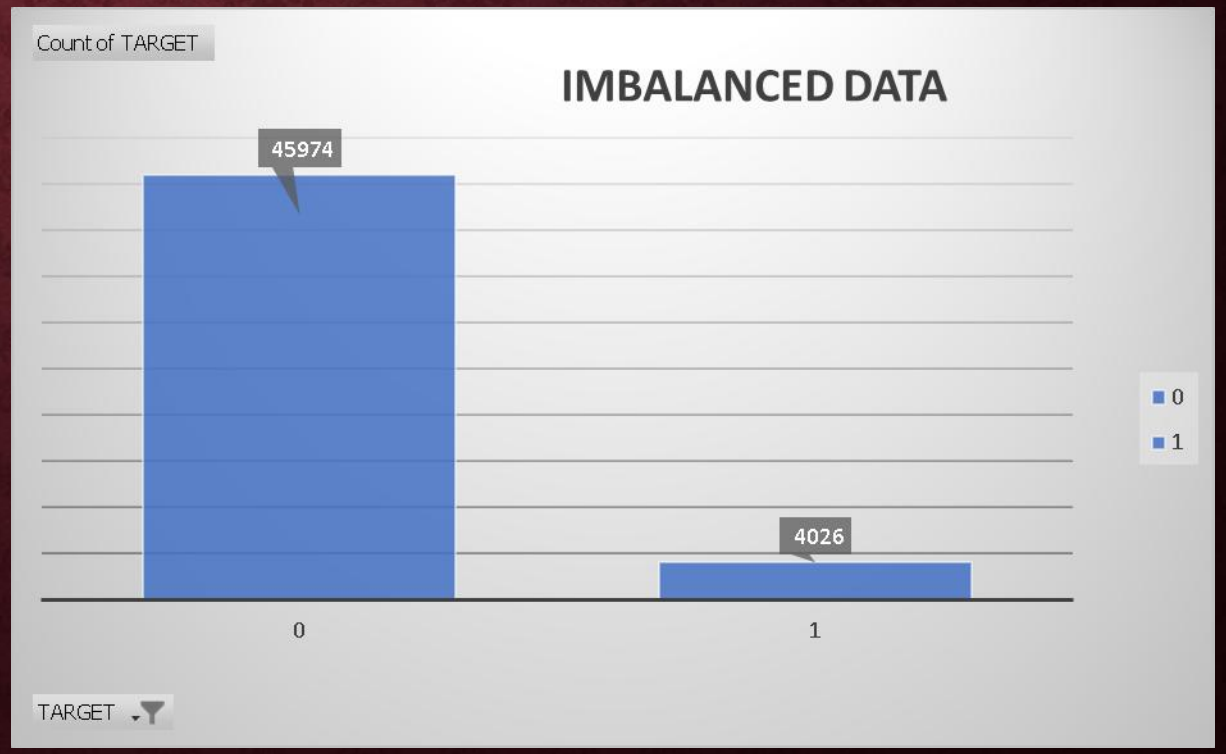
❑ Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

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

TARGET	Count of TARGET
0	45974
1	4026
Grand Total	50000

TARGET	COUNT OF TARGET	PERCENTILE(%)
0	45974	92
1	4026	8

RATIO OF IMBALANCE
11.42



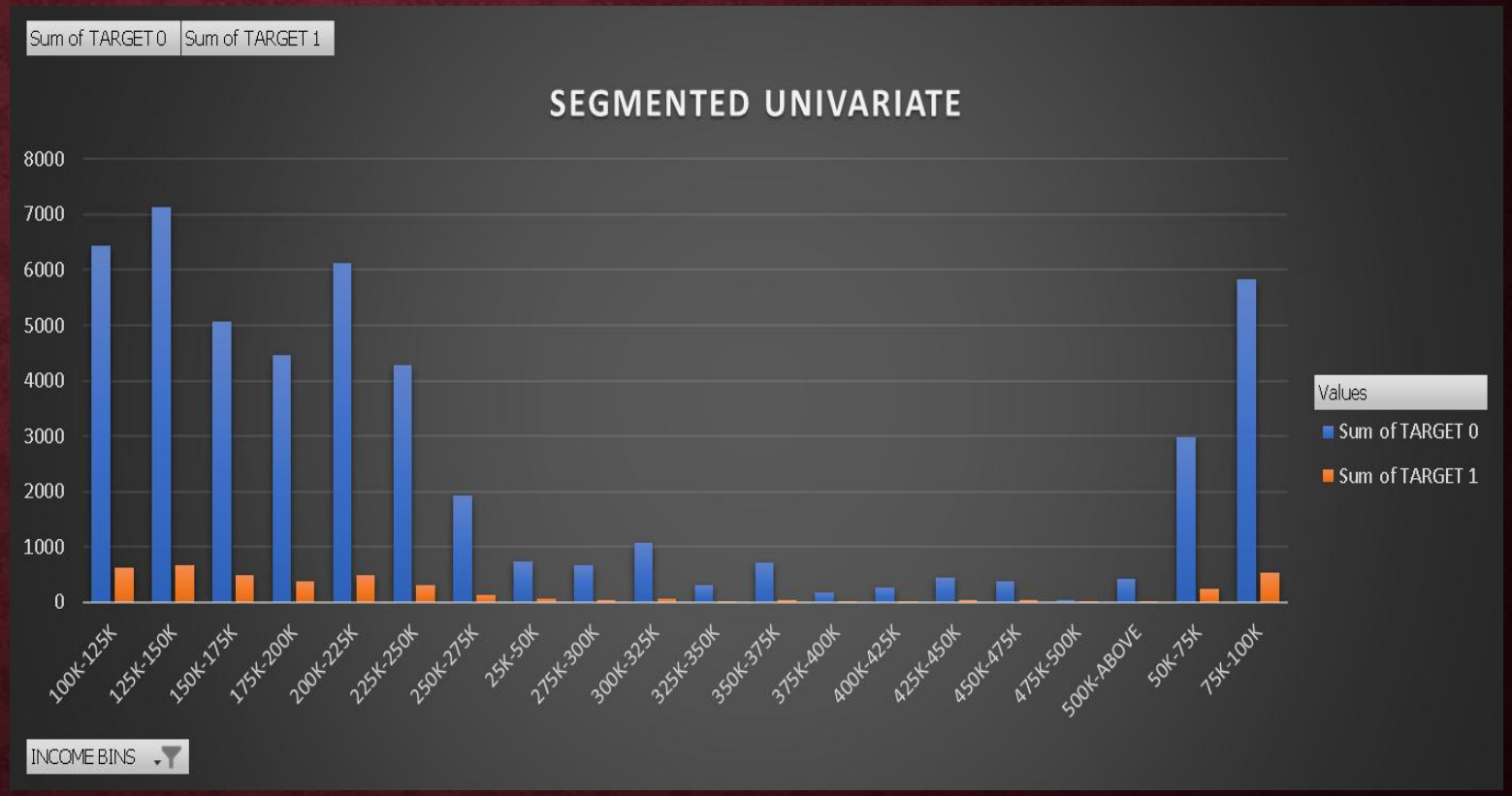
❑ Perform Univariate, Segmented Univariate, and Bivariate Analysis:

To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

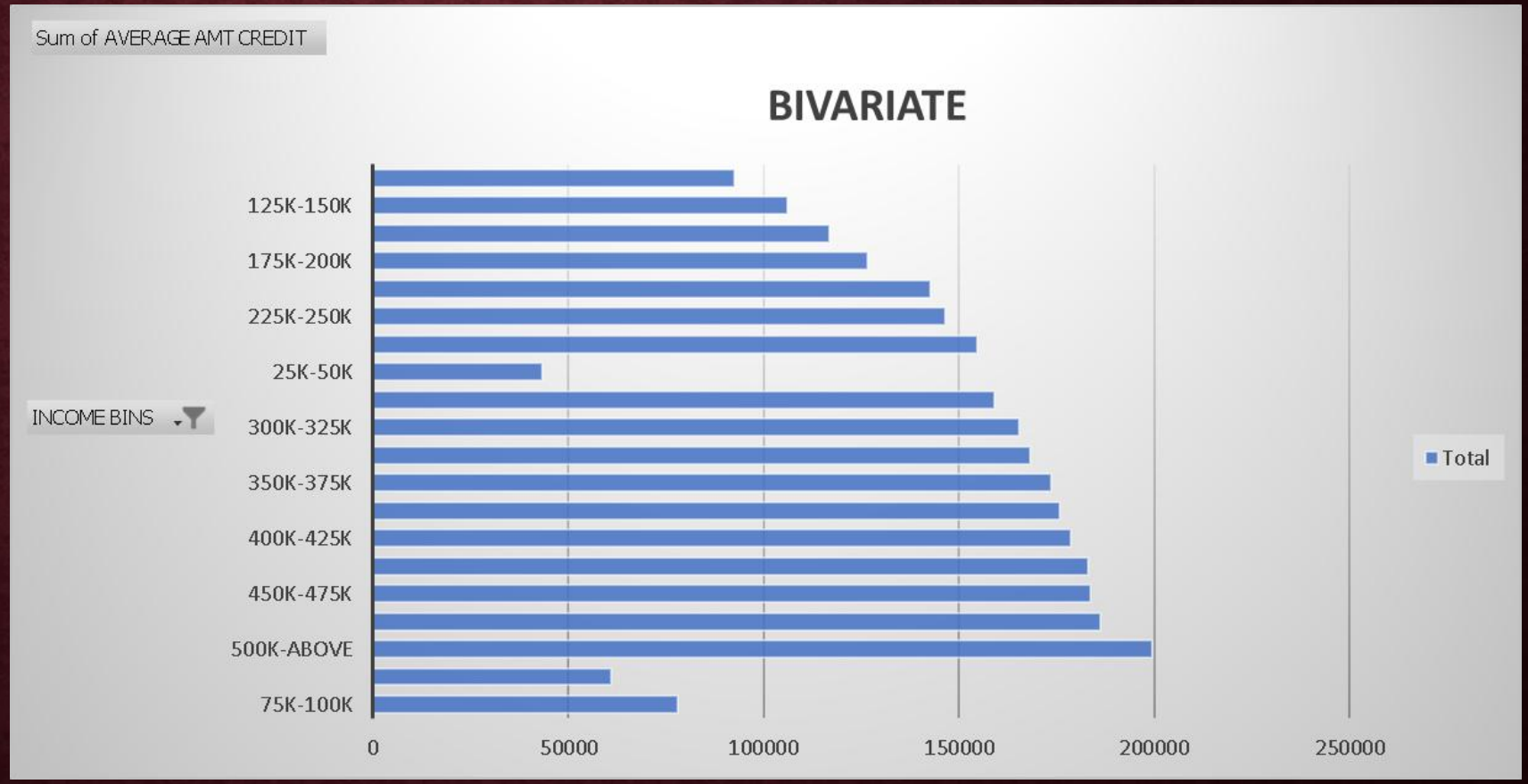
- **Task:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and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

Row Labels	Sum of APPLICANTS
0-1LACS	989
10LACS-ABOVE	8146
1-2LACS	4911
2-3LACS	8849
3-4LACS	4256
4-5LACS	5228
5-6LACS	5554
6-7LACS	3909
7-8LACS	3062
8-9LACS	3571
9-10LACS	2548
Grand Total	**51023**

Row Labels	Sum of TARGET 0	Sum of TARGET 1
100K-125K	6428	620
125K-150K	7126	678
150K-175K	5060	501
175K-200K	4458	389
200K-225K	6121	491
225K-250K	4279	304
250K-275K	1919	143
25K-50K	741	63
275K-300K	681	45
300K-325K	1076	59
325K-350K	322	24
350K-375K	723	34
375K-400K	186	14
400K-425K	263	26
425K-450K	456	36
450K-475K	375	34
475K-500K	44	3
500K-ABOVE	423	31
50K-75K	2980	246
75K-100K	5826	536
Grand Total	**49487**	**4277**



Row Labels	Sum of AVERAGE AMT CREDIT
75K-100K	77889.78064
50K-75K	60938.37955
500K-ABOVE	199436.5293
475K-500K	186206.1803
450K-475K	183785.4382
425K-450K	183022.5916
400K-425K	178541.9372
375K-400K	175727.0305
350K-375K	173508.3829
325K-350K	168219.9584
300K-325K	165318.4532
275K-300K	159049.266
25K-50K	43179.35032
250K-275K	154612.9794
225K-250K	146394.564
200K-225K	142505.4605
175K-200K	126665.2453
150K-175K	116820.3445
125K-150K	106165.4935
100K-125K	92359.98262
Grand Total	**2840347.348**



- **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- **Task:** 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	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	BIRTH_YEARS	DAYS_EMPLOYED(YEARS)	PUBLISH_YEAR	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.009588558	0.00497156	-0.025555665	-0.329263754	-0.241539565	0.032115773	0.025913889
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	0.029841469	-0.016002774	-0.03151033	-0.003506646	-0.038188511
AMT_CREDIT	0.00497156	0.069315897	1	0.095111221	0.059342658	-0.06773941	0.012228765	-0.100507425
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	1	0.032513748	-0.004158337	0.004345136	-0.532667302
BIRTH_YEARS	-0.329263754	-0.016002774	0.059342658	0.032513748	1	0.62172831	0.270825141	-0.016779196
DAYS_EMPLOYED(YEARS)	-0.241539565	-0.03151033	-0.06773941	-0.004158337	0.62172831	1	0.272766672	0.034558656
PUBLISH_YEAR	0.032115773	-0.003506646	0.012228765	0.004345136	0.270825141	0.272766672	1	0.002307011
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	-0.532667302	-0.016779196	0.034558656	0.002307011	1

RESULT

This project helps in handling the large datasets. How exploratory data analysis can be applied to large datasets. When dealing with the large datasets it is also important to select only those columns which are extremely useful to our analysis. Finding correlations columns can become very convenient while dealing with large datasets as it saves time selecting which columns should be considered for analysis. The project also helps in understanding the various terminologies used in the banking domain. The insight drawn from the project are as follows:

1. Applicants drawing higher income were offered higher loan amount by the bank.
2. Majority of applicants drawn an income range between 1 Lacs – 1.5 Lacs, also the defaults drawn income between the same range.
3. Majority of applicants were offered loans in the credit range of 9 Lacs and above.

THANK YOU!