Bank Loan Case Study

Final Project-2

Description:

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Drive Link:
https://drive.google.com/drive/folders/1Z1QcPLbcxUQAfgoFf1kwwFl7V4RHSRQX?usp=sharing

Loom Video Link:
https://www.loom.com/share/f1f623cda9284ae5afdff4499654efc2?sid=d9059c0d-f9b4-4eb9-9941-0021056215ef

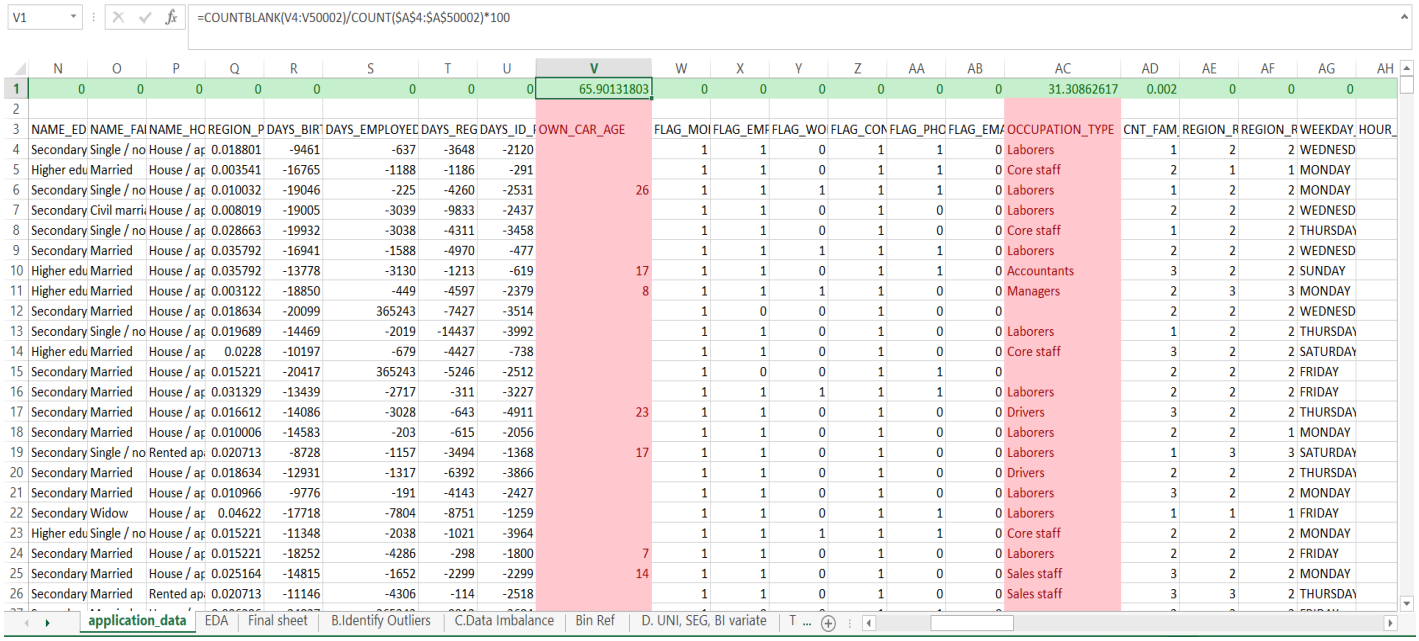
Data Analytics Tasks:

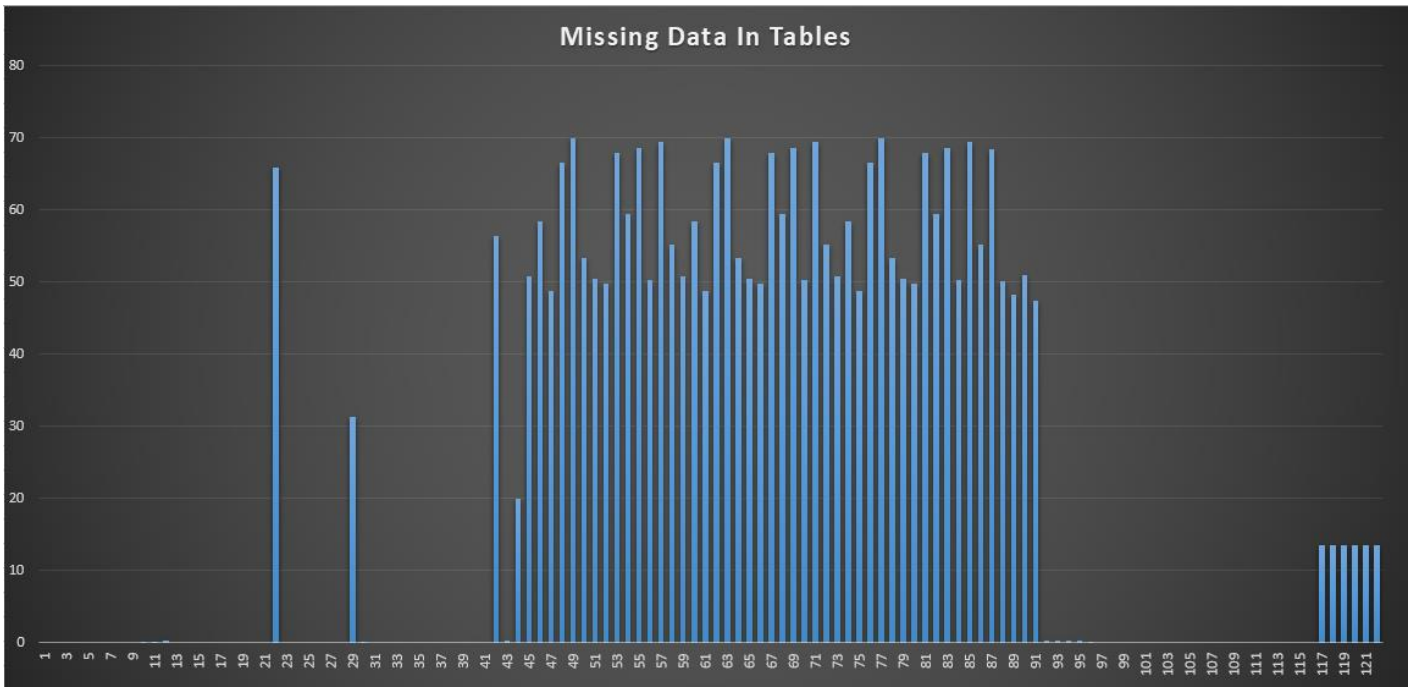
A. Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

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.

Hint: Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.

Graph suggestion: Create a bar chart or column chart to visualize the proportion of missing values for each variable.





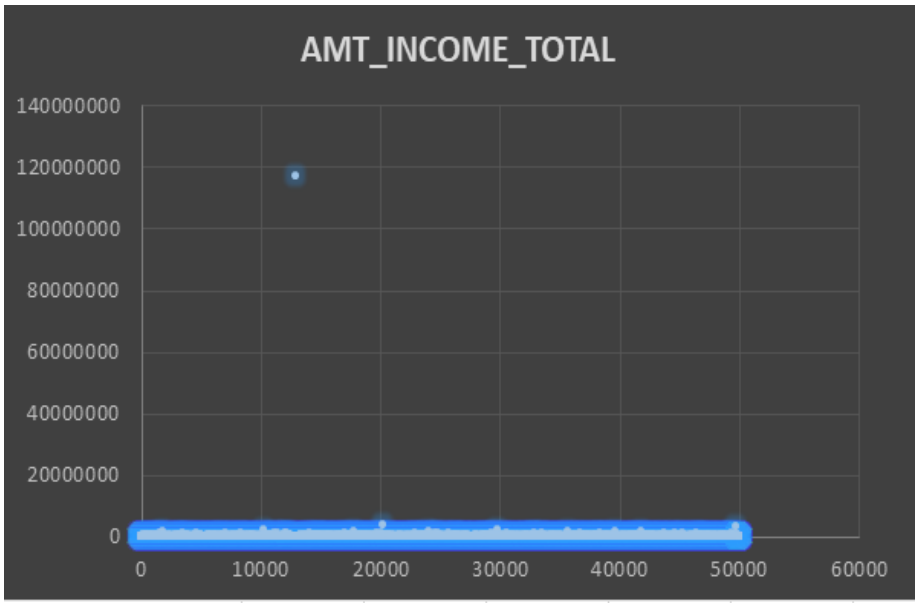
B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

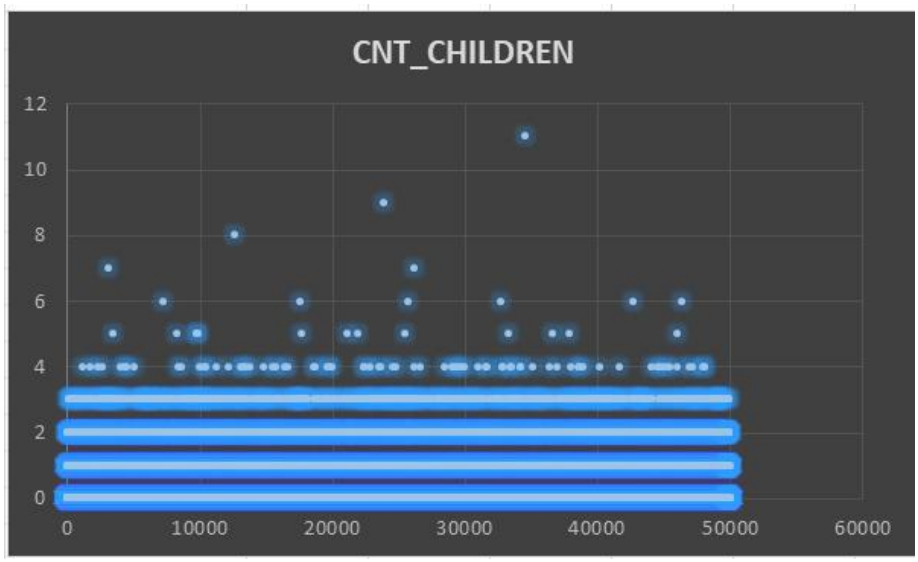
Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

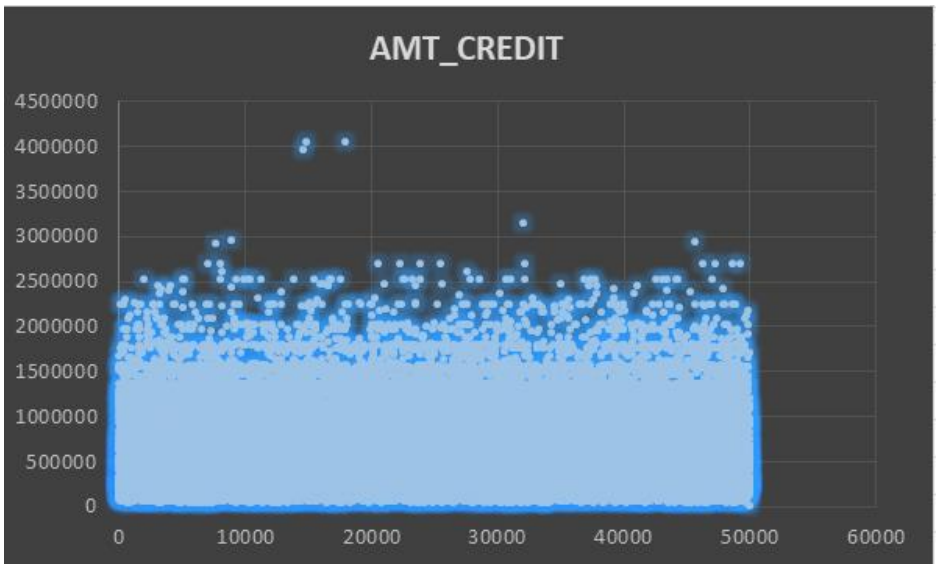
Hint: Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.

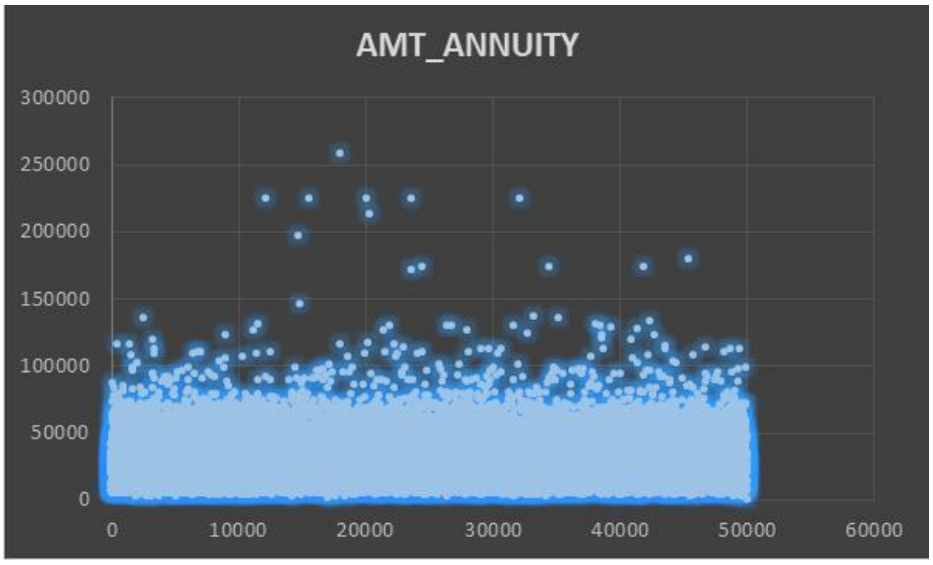
Graph suggestion: Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

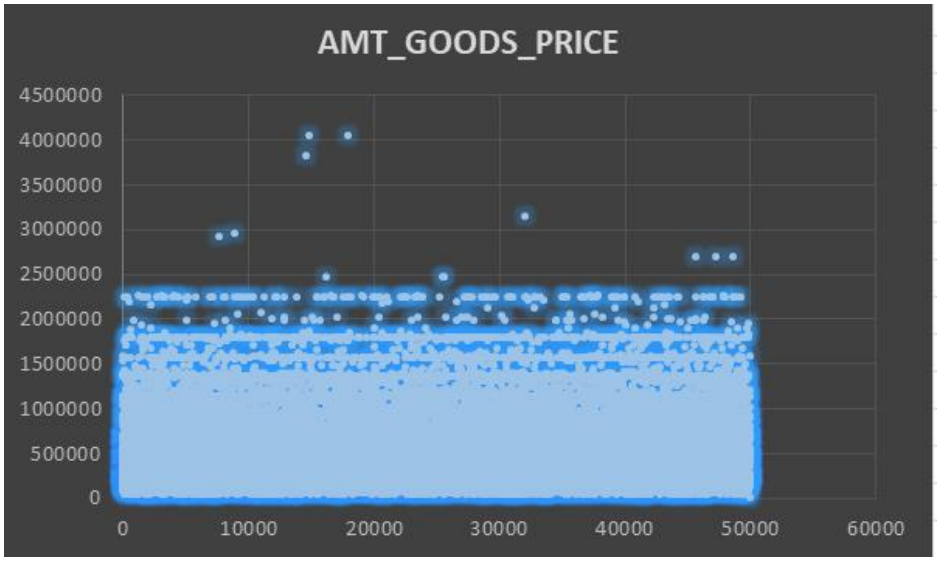
Quartile 1	112500
Quartile 3	202500
Inter Quartile Range	90000
Lower limit	-22500
Upper Limit	337500

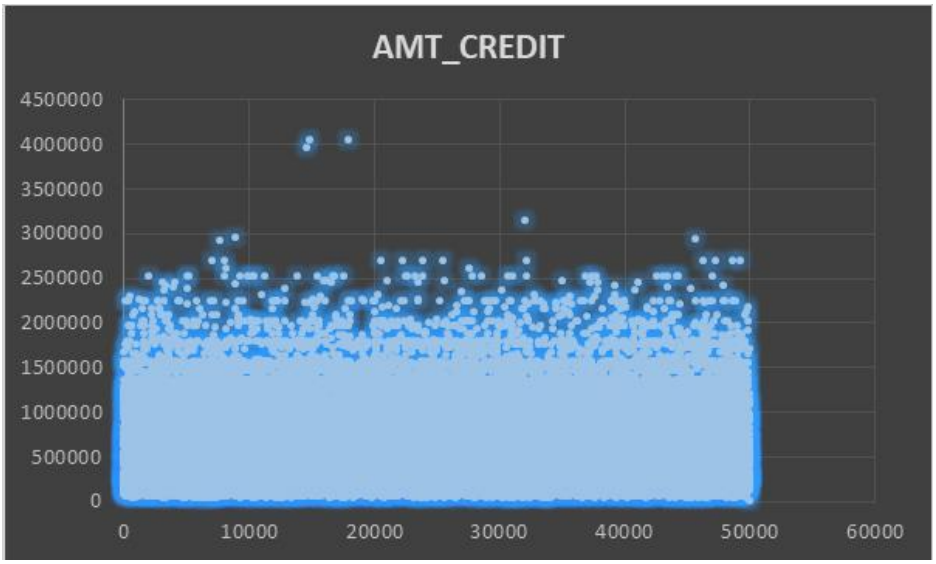

AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
0
10000
20000
30000
40000
50000
60000


CNT_CHILDREN
12
10
8
6
4
2
0
0
10000
20000
30000
40000
50000
60000


AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000
60000


AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
0
10000
20000
30000
40000
50000
60000


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000
60000


AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
10000
20000
30000
40000
50000
60000

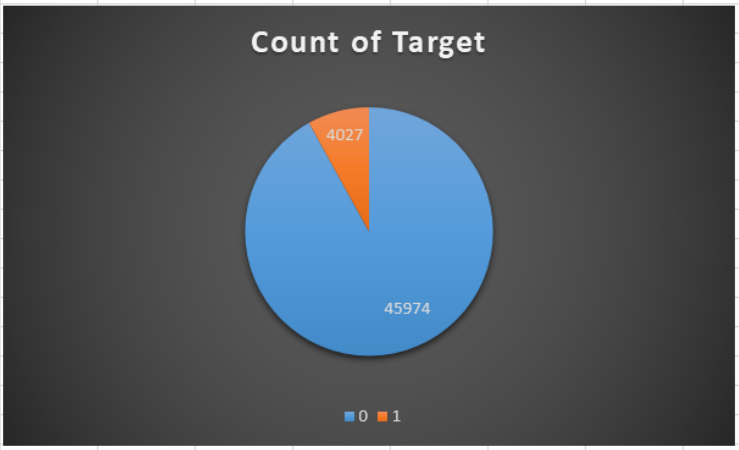
C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models**.**

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

Hint: Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.

Graph suggestion: Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

Target	Count of Target
0	45974
1	4027
Grand Total	50001
Ratio	0.087592987



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.

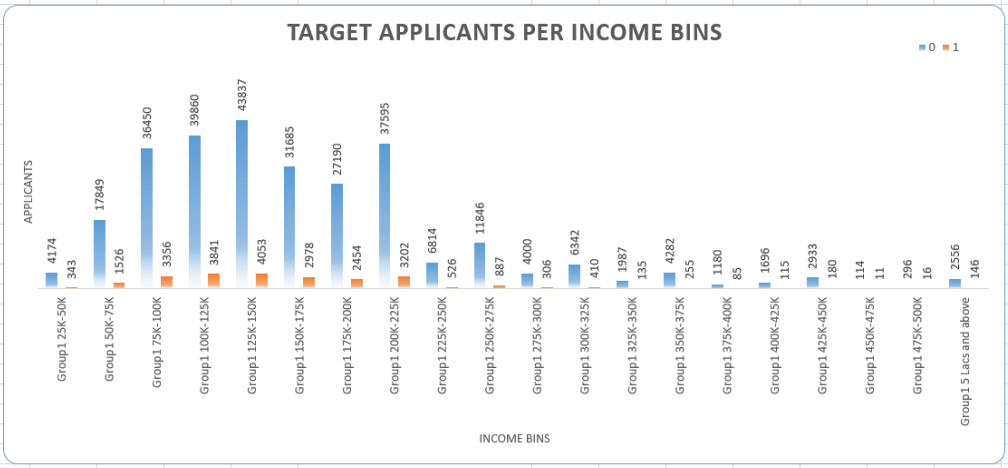
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Hint: Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.

Graph suggestion: Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

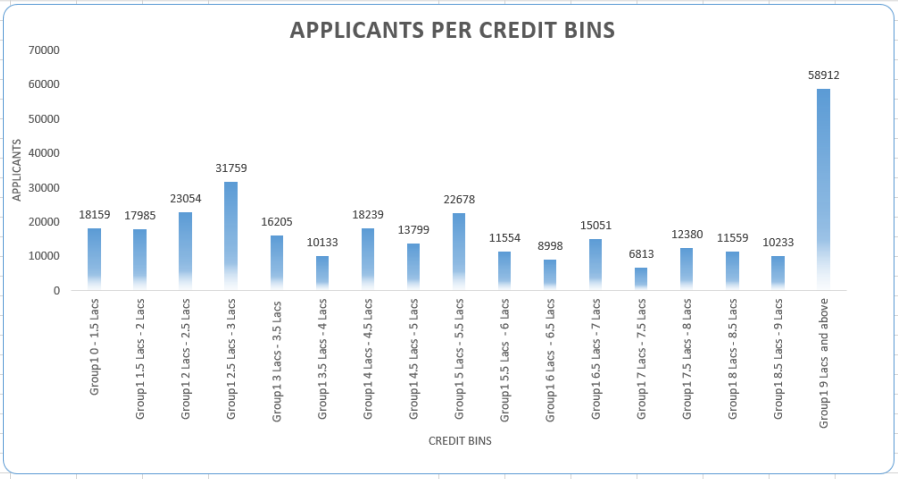
SEGMENTED UNIVARIATE ANALYSIS

.	TARGET	
INCOME BINS	0	1
Group1	282686	#####
25K-50K	4174	343
50K-75K	17849	1526
75K-100K	36450	3356
100K-125K	39860	3841
125K-150K	43837	4053
150K-175K	31685	2978
175K-200K	27190	2454
200K-225K	37595	3202
225K-250K	6814	526
250K-275K	11846	887
275K-300K	4000	306
300K-325K	6342	410
325K-350K	1987	135
350K-375K	4282	255
375K-400K	1180	85
400K-425K	1696	115
425K-450K	2933	180
450K-475K	114	11
475K-500K	296	16
5 Lacs and above	2556	146



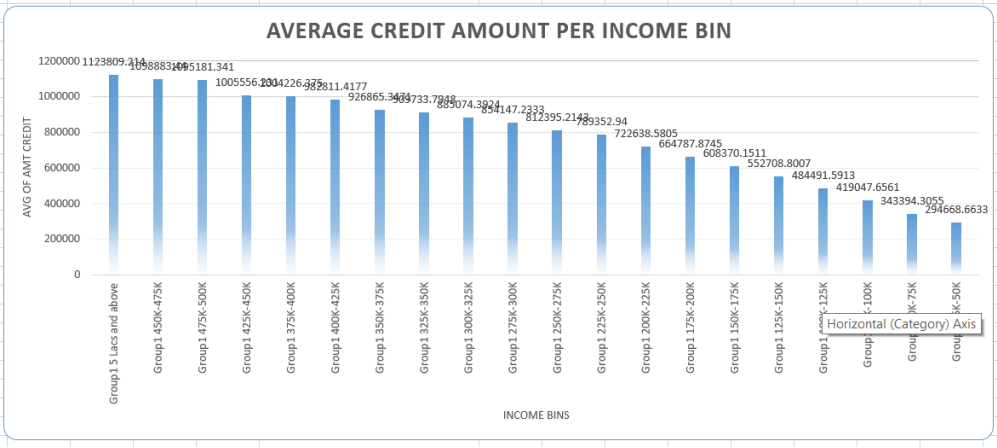
UNIVARIATE ANALYSIS

CREDIT BINS	APPLICANTS
Group1	307511
0 - 1.5 Lacs	18159
1.5 Lacs - 2 Lacs	17985
2 Lacs - 2.5 Lacs	23054
2.5 Lacs - 3 Lacs	31759
3 Lacs - 3.5 Lacs	16205
3.5 Lacs - 4 Lacs	10133
4 Lacs - 4.5 Lacs	18239
4.5 Lacs - 5 Lacs	13799
5 Lacs - 5.5 Lacs	22678
5.5 Lacs - 6 Lacs	11554
6 Lacs - 6.5 Lacs	8998
6.5 Lacs - 7 Lacs	15051
7 Lacs - 7.5 Lacs	6813
7.5 Lacs - 8 Lacs	12380
8 Lacs - 8.5 Lacs	11559
8.5 Lacs - 9 Lacs	10233
9 Lacs and above	58912



INCOME BINS	Average of AMT_CREDIT
Group1	5,99,026
5 Lacs and above	11,23,809
450K-475K	10,98,883
475K-500K	10,95,181
425K-450K	10,05,556
375K-400K	10,04,226
400K-425K	9,82,811
350K-375K	9,26,865
325K-350K	9,09,734
300K-325K	8,85,074
275K-300K	8,54,147
250K-275K	8,12,395
225K-250K	7,89,353
200K-225K	7,22,639
175K-200K	6,64,788
150K-175K	6,08,370
125K-150K	5,52,709
100K-125K	4,84,492
75K-100K	4,19,048
50K-75K	3,43,394
25K-50K	2,94,669

BIVARIATE ANALYSIS



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.

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.

Hint: Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario.

Graph suggestion: Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME								
	CNT_CHILDREN	AMT_INCOME_TOTAl	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH(Years)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.027397188	0.003081225	-0.024362658	-0.336966484	-0.245174065	0.028750653	0.022842107
AMT_INCOME_TOTAl	0.027397188	1	0.34279945	0.167850636	-0.062609158	-0.140392466	-0.022896393	-0.186573418
AMT_CREDIT	0.003081225	0.34279945	1	0.100603799	0.047377831	-0.070104314	0.00146417	-0.103336744
REGION_POPULATION_RELATIVE	-0.024362658	0.167850636	0.100603799	1	0.025244113	-0.007197856	0.001070788	-0.539004783
DAYS_BIRTH(Years)	-0.336966484	-0.062609158	0.047377831	0.025244113	1	0.626113878	0.271314395	-0.002332327
DAYS_EMPLOYED (Years)	-0.245174065	-0.140392466	-0.07010431	-0.007197856	0.626113878	1	0.27666316	0.038327694
DAYS_ID_PUBLISH(Years)	0.028750653	-0.022896393	0.00146417	0.001070788	0.271314395	0.27666316	1	0.00899835
REGION_RATING_CLIENT	0.022842107	-0.186573418	-0.10333674	-0.539004783	-0.002332327	0.038327694	0.00899835	1

CORRELATION FOR APPLICANTS WITH PAYMENT DIFFICULTIES								
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH(Years)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.004795787	-0.00167496	-0.0319749	-0.259108666	-0.192863828	0.032298597	0.040680482
AMT_INCOME_TOTAL	0.004795787	1	0.038131435	0.009134586	-0.003096245	-0.014977396	0.004214856	-0.021486257
AMT_CREDIT	-0.001674961	0.038131435	1	0.069161087	0.135316369	0.001930183	0.05232898	-0.059192754
REGION_POPULATION_RELATIVE	-0.0319749	0.009134586	0.069161087	1	0.048190366	0.015531849	0.015536882	-0.443235509
DAYS_BIRTH(Years)	-0.259108666	-0.003096245	0.135316369	0.048190366	1	0.582185148	0.252862836	-0.033927932
DAYS_EMPLOYED (Years)	-0.192863828	-0.014977396	0.001930183	0.015531849	0.582185148	1	0.229090254	0.003489989
DAYS_ID_PUBLISH(Years)	0.032298597	0.004214856	0.05232898	0.015536882	0.252862836	0.229090254	1	-0.001397237
REGION_RATING_CLIENT	0.040680482	-0.021486257	-0.05919275	-0.443235509	-0.033927932	0.003489989	-0.001397237	1

- **Approach:** I followed a structured approach to analyze the loan application dataset. First, I conducted data preprocessing to handle missing values, identify outliers, and address data imbalance. Then, I performed univariate, segmented univariate, and bivariate analyses to understand the relationships between customer attributes, loan attributes, and loan default likelihood. Finally, I identified top correlations for different scenarios to uncover strong indicators of loan default.

- **Tech-Stack Used:** Microsoft Excel 2022: Used for data analysis, visualization, and statistical calculations.Google Drive: Used to store and share project documents and reports.

- **Insights:** Through EDA, I gained valuable insights into the factors influencing loan default. I observed that customers with payment difficulties tend to have specific attributes such as late payment history and higher debt-to-income ratios. Additionally, I discovered correlations between certain customer attributes (e.g., credit score, income level) and the likelihood of default.

- **Result**: The project provided valuable insights into the drivers of loan default, enabling the company to make informed decisions about loan approval. By understanding the key factors behind loan default, the company can mitigate risks and improve the accuracy of loan approval processes.