Bank Loan Case Study

Problem Statement: 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.

When a customer applies for a loan, your company faces two risks:

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 dataset you'll be working with contains information about loan applications. It includes two types of scenarios:

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.

2. All other cases: These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.

2. Cancelled: The customer cancelled the application during the approval process.

3. Refused: The company rejected the loan.

4. Unused Offer: The loan was approved but the customer did not use it.

Your goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

Tech Task Used:

Microsoft Excel 2016: Key Features

Pivot Tables:

- **Purpose:** Summarize, analyze, explore, and present data.
- **How to Use:**
    1. Select your data range.
    2. Go to the **"Insert"** tab.
    3. Click on **"PivotTable"**.
    4. Choose the data range and the location for the PivotTable.
    5. Drag and drop fields into the **Rows**, **Columns**, **Values**, and **Filters** areas to create your report.

Charts:

- **Purpose:** Visualize data trends and patterns.
- **How to Use:**
    1. Select the data you want to chart.
    2. Go to the **"Insert"** tab.
    3. Choose the type of chart you need from options like **Column**, **Line**, **Pie**, **Bar**, **Area**, **Scatter**, and more.
    4. Customize the chart using the **Chart Tools** that appear on the Ribbon.

Formulas and Functions:

- **Purpose:** Perform calculations and analyze data.
- **Key Functions**: All the formulae used for calculation are as follows -
    - **SUM**
    - **AVERAGE**
    - **MEDIAN**
    - **MODE**
    - **CORREL**
    - **COUNTIF**
    - **ISBLANK**
    - **IF**
    - **MAX**
    - **MIN**

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.

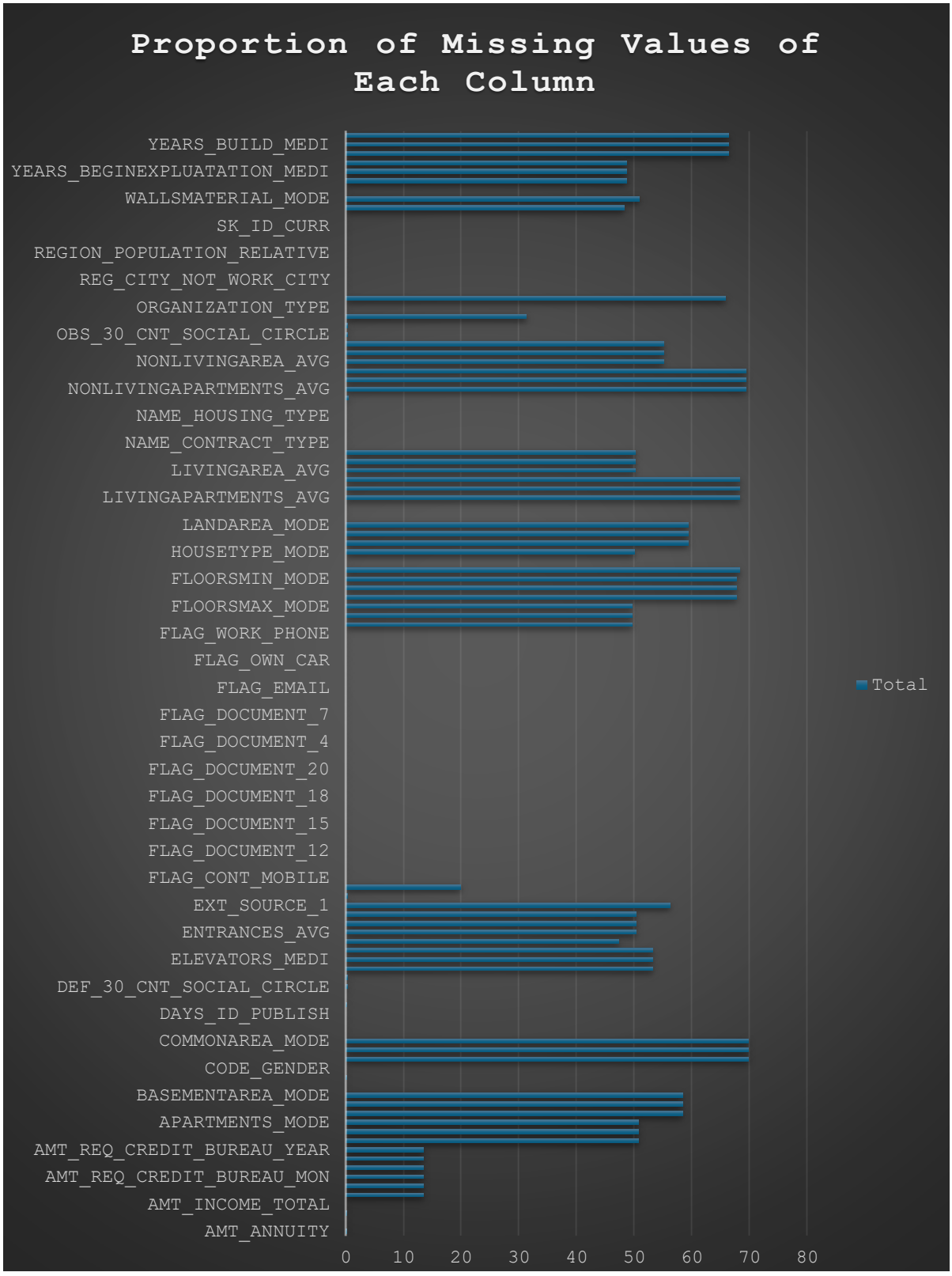
- **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

Sr.No.	Column Name	Missing Values	Proportion of Missing Values
1	SK_ID_CURR	0	0
2	TARGET	0	0
3	NAME_CONTRACT_TYPE	0	0
4	CODE_GENDER	0	0
5	FLAG_OWN_CAR	0	0
6	FLAG_OWN_REALTY	0	0
7	CNT_CHILDREN	0	0
8	AMT_INCOME_TOTAL	0	0
9	AMT_CREDIT	0	0
10	AMT_ANNUITY	1	0.00200004
11	AMT_GOODS_PRICE	38	0.07600152
12	NAME_TYPE_SUITE	192	0.38400768
13	NAME_INCOME_TYPE	0	0
14	NAME_EDUCATION_TYPE	0	0
15	NAME_FAMILY_STATUS	0	0
16	NAME_HOUSING_TYPE	0	0
17	REGION_POPULATION_RELATIVE	0	0
18	DAYS_BIRTH	0	0
19	DAYS_EMPLOYED	0	0
20	DAYS_REGISTRATION	0	0
21	DAYS_ID_PUBLISH	0	0
22	OWN_CAR_AGE	32950	65.90131803
23	FLAG_MOBIL	0	0
24	FLAG_EMP_PHONE	0	0
25	FLAG_WORK_PHONE	0	0
26	FLAG_CONT_MOBILE	0	0
27	FLAG_PHONE	0	0
28	FLAG_EMAIL	0	0
29	OCCUPATION_TYPE	15654	31.30862617
30	CNT_FAM_MEMBERS	1	0.00200004
31	REGION_RATING_CLIENT	0	0
32	REGION_RATING_CLIENT_W_CITY	0	0
33	WEEKDAY_APPR_PROCESS_START	0	0

34	HOUR_APPR_PROCESS_START	0	0
35	REG_REGION_NOT_LIVE_REGION	0	0
36	REG_REGION_NOT_WORK_REGION	0	0
37	LIVE_REGION_NOT_WORK_REGION	0	0
38	REG_CITY_NOT_LIVE_CITY	0	0
39	REG_CITY_NOT_WORK_CITY	0	0
40	LIVE_CITY_NOT_WORK_CITY	0	0
41	ORGANIZATION_TYPE	0	0
42	EXT_SOURCE_1	28172	56.3451269
43	EXT_SOURCE_2	126	0.25200504
44	EXT_SOURCE_3	9944	19.88839777
45	APARTMENTS_AVG	25385	50.77101542
46	BASEMENTAREA_AVG	29199	58.39916798
47	YEARS_BEGINEXPLUATATION_AVG	24394	48.78897578
48	YEARS_BUILD_AVG	33239	66.47932959
49	COMMONAREA_AVG	34960	69.92139843
50	ELEVATORS_AVG	26651	53.30306606
51	ENTRANCES_AVG	25195	50.39100782
52	FLOORSMAX_AVG	24875	49.75099502
53	FLOORSMIN_AVG	33894	67.78935579
54	LANDAREA_AVG	29721	59.44318886
55	LIVINGAPARTMENTS_AVG	34226	68.45336907
56	LIVINGAREA_AVG	25137	50.2750055
57	NONLIVINGAPARTMENTS_AVG	34714	69.42938859
58	NONLIVINGAREA_AVG	27572	55.1451029
59	APARTMENTS_MODE	25385	50.77101542
60	BASEMENTAREA_MODE	29199	58.39916798
61	YEARS_BEGINEXPLUATATION_MODE	24394	48.78897578
62	YEARS_BUILD_MODE	33239	66.47932959
63	COMMONAREA_MODE	34960	69.92139843
64	ELEVATORS_MODE	26651	53.30306606
65	ENTRANCES_MODE	25195	50.39100782
66	FLOORSMAX_MODE	24875	49.75099502
67	FLOORSMIN_MODE	33894	67.78935579
68	LANDAREA_MODE	29721	59.44318886
69	LIVINGAPARTMENTS_MODE	34226	68.45336907
70	LIVINGAREA_MODE	25137	50.2750055
71	NONLIVINGAPARTMENTS_MODE	34714	69.42938859
72	NONLIVINGAREA_MODE	27572	55.1451029
73	APARTMENTS_MEDI	25385	50.77101542
74	BASEMENTAREA_MEDI	29199	58.39916798
75	YEARS_BEGINEXPLUATATION_MEDI	24394	48.78897578
76	YEARS_BUILD_MEDI	33239	66.47932959
77	COMMONAREA_MEDI	34960	69.92139843

78	ELEVATORS_MEDI	26651	53.30306606
79	ENTRANCES_MEDI	25195	50.39100782
80	FLOORSMAX_MEDI	24875	49.75099502
81	FLOORSMIN_MEDI	33894	67.78935579
82	LANDAREA_MEDI	29721	59.44318886
83	LIVINGAPARTMENTS_MEDI	34226	68.45336907
84	LIVINGAREA_MEDI	25137	50.2750055
85	NONLIVINGAPARTMENTS_MEDI	34714	69.42938859
86	NONLIVINGAREA_MEDI	27572	55.1451029
87	FONDKAPREMONT_MODE	34191	68.38336767
88	HOUSETYPE_MODE	25075	50.15100302
89	TOTALAREA_MODE	24148	48.29696594
90	WALLSMATERIAL_MODE	25459	50.91901838
91	EMERGENCYSTATE_MODE	23698	47.39694794
92	OBS_30_CNT_SOCIAL_CIRCLE	168	0.33600672
93	DEF_30_CNT_SOCIAL_CIRCLE	168	0.33600672
94	OBS_60_CNT_SOCIAL_CIRCLE	168	0.33600672
95	DEF_60_CNT_SOCIAL_CIRCLE	168	0.33600672
96	DAYS_LAST_PHONE_CHANGE	1	0.00200004
97	FLAG_DOCUMENT_2	0	0
98	FLAG_DOCUMENT_3	0	0
99	FLAG_DOCUMENT_4	0	0
100	FLAG_DOCUMENT_5	0	0
101	FLAG_DOCUMENT_6	0	0
102	FLAG_DOCUMENT_7	0	0
103	FLAG_DOCUMENT_8	0	0
104	FLAG_DOCUMENT_9	0	0
105	FLAG_DOCUMENT_10	0	0
106	FLAG_DOCUMENT_11	0	0
107	FLAG_DOCUMENT_12	0	0
108	FLAG_DOCUMENT_13	0	0
109	FLAG_DOCUMENT_14	0	0
110	FLAG_DOCUMENT_15	0	0
111	FLAG_DOCUMENT_16	0	0
112	FLAG_DOCUMENT_17	0	0
113	FLAG_DOCUMENT_18	0	0
114	FLAG_DOCUMENT_19	0	0
115	FLAG_DOCUMENT_20	0	0
116	FLAG_DOCUMENT_21	0	0
117	AMT_REQ_CREDIT_BUREAU_HOUR	6734	13.46826937
118	AMT_REQ_CREDIT_BUREAU_DAY	6734	13.46826937
119	AMT_REQ_CREDIT_BUREAU_WEEK	6734	13.46826937
120	AMT_REQ_CREDIT_BUREAU_MON	6734	13.46826937
121	AMT_REQ_CREDIT_BUREAU_QRT	6734	13.46826937
122	AMT_REQ_CREDIT_BUREAU_YEAR	6734	13.46826937

Color	Proportion of Missing Values
	<50% Categorical Values
	<50% Numerical Values
	Special Case
	>50% Missing Values



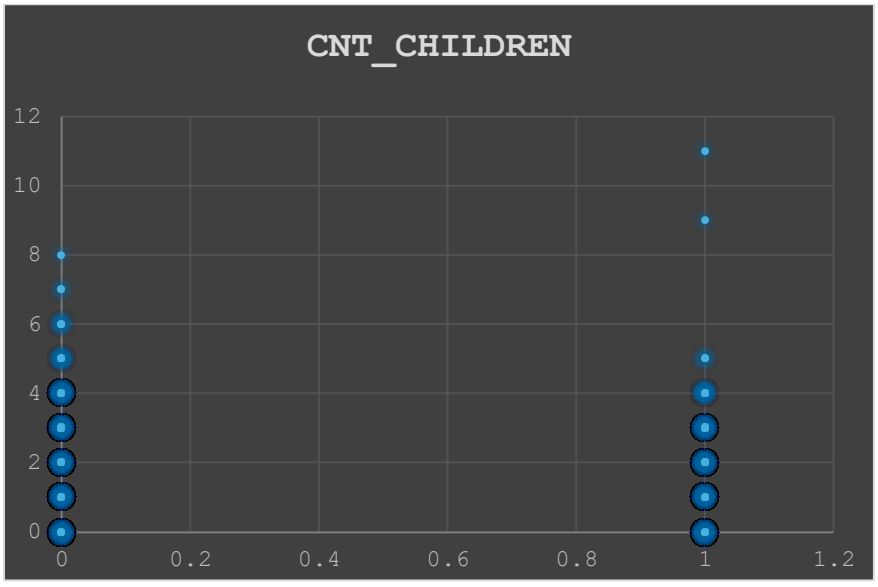
B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

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

Result:

1. Outliers in CNT_CHILDREN

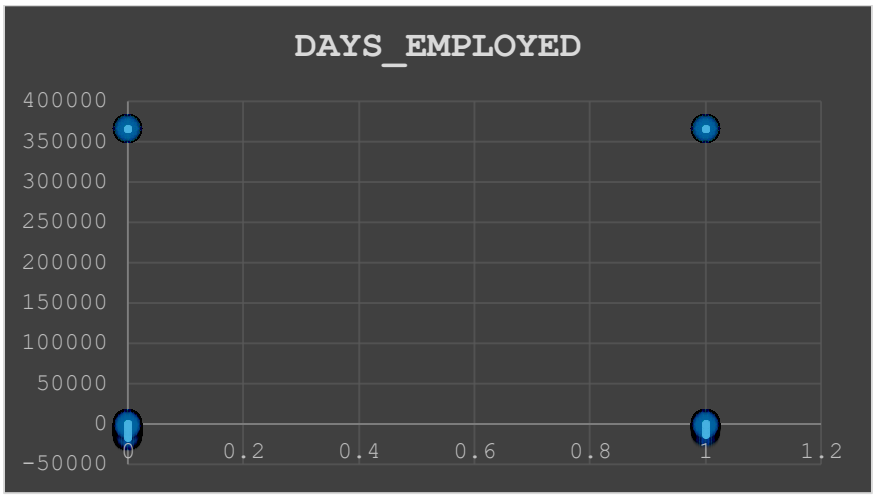
Column1	Column2
QUARTILE Q1	0
QUARTILE Q3	1
IQR	1
LOWER BOUND	-1.5
UPPER BOUND	1.5



2. Outliers in DAYS_EMPLOYED

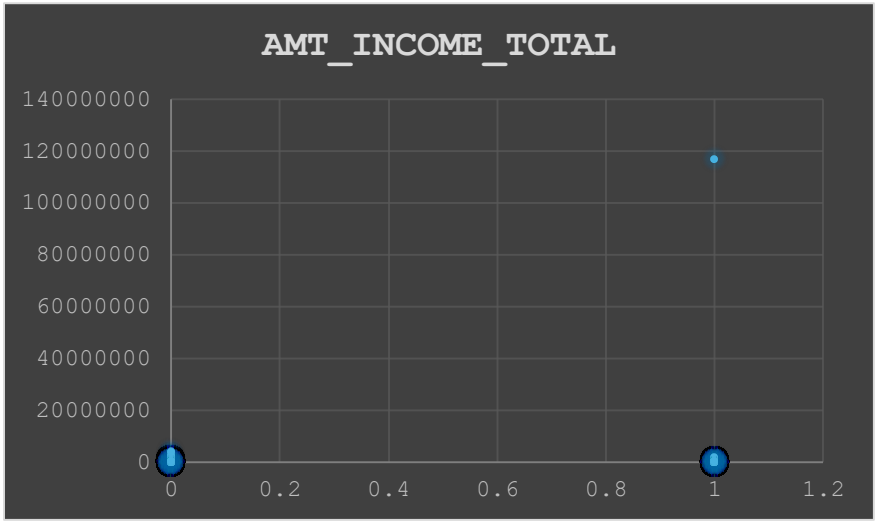
Column1	Column2
QUARTILE Q1	-2786
QUARTILE Q3	-292
IQR	2494

LOWER BOUND	-6527
UPPER BOUND	955



3. Outliers in AMT_INCOME_TOTAL

QUARTILE Q1	112500
QUARTILE Q3	202500
IQR	90000
LOWER BOUND	-22500
UPPER BOUND	247500



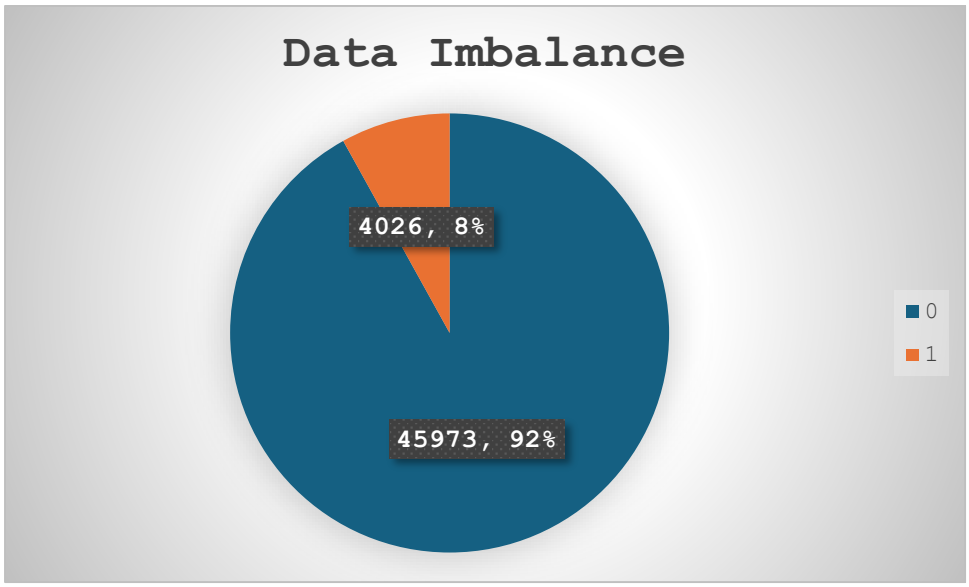
C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

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

Result:

Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	**49999**

Data Imbalance Ratio	**11.42**
Class 0 (No Payment Difficulties) **Class 1 (Data Imbalance/ Payment Difficulties)**	



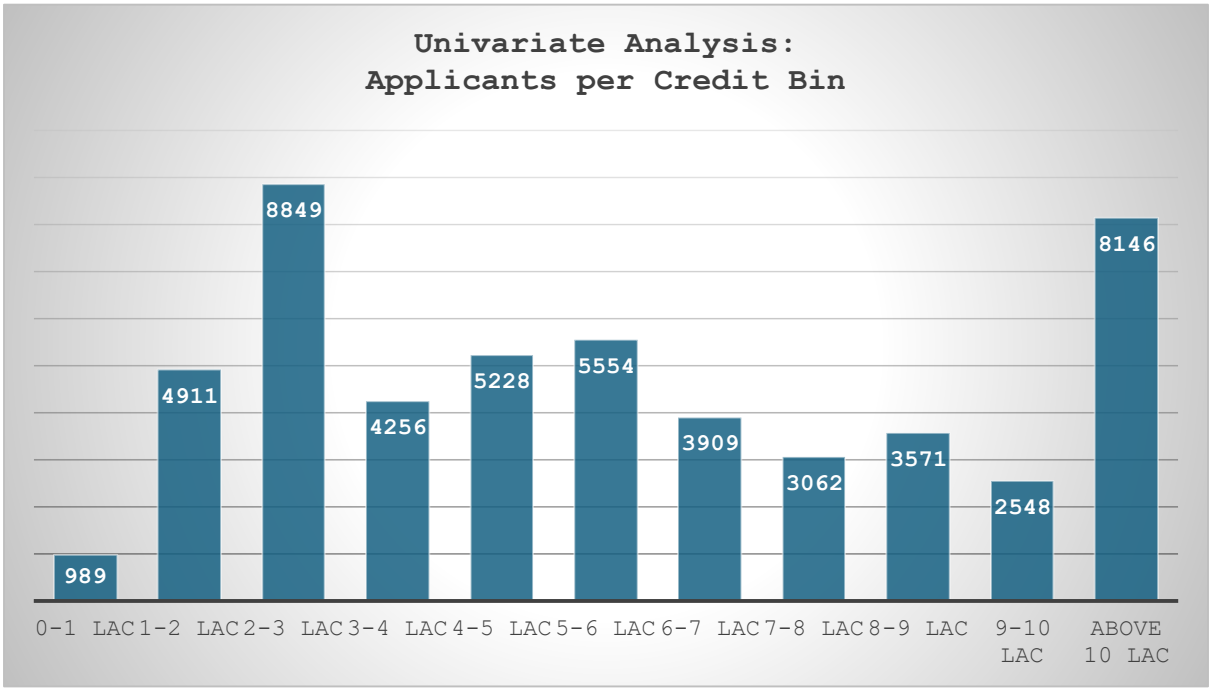
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

- **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.

Result:

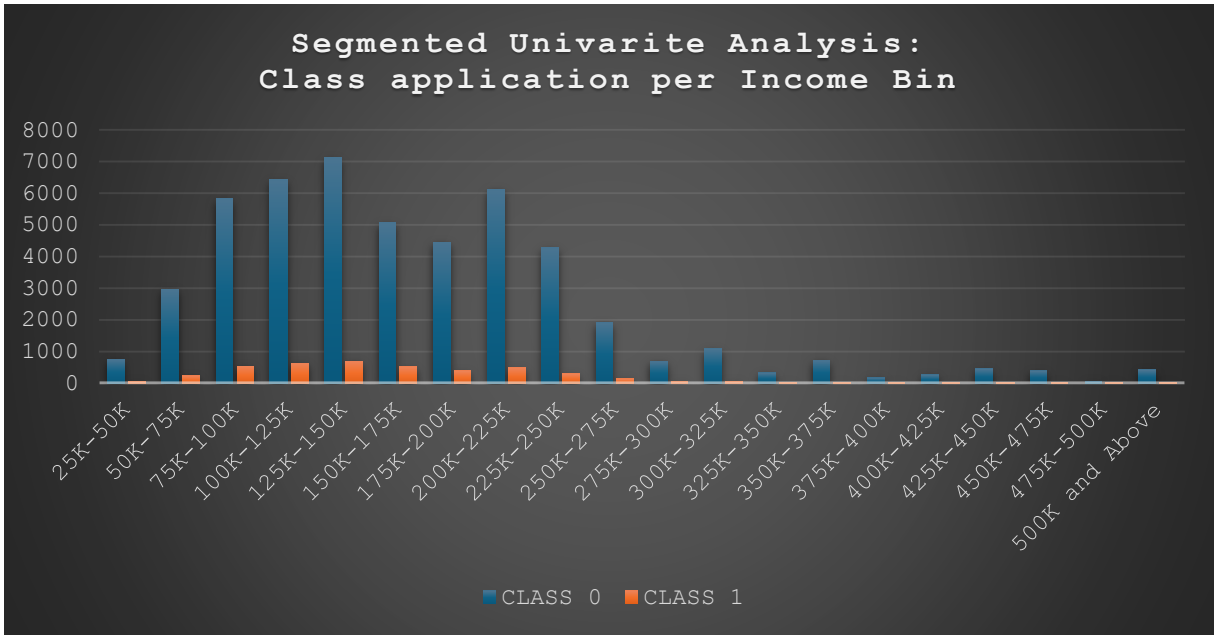
1. **Univariate Analysis: Applicants per Credit Bin**

CREDIT BIN	APPLICANTS
0-1 LAC	989
1-2 LAC	4911
2-3 LAC	8849
3-4 LAC	4256
4-5 LAC	5228
5-6 LAC	5554
6-7 LAC	3909
7-8 LAC	3062
8-9 LAC	3571
9-10 LAC	2548
Above 10 lac	8146



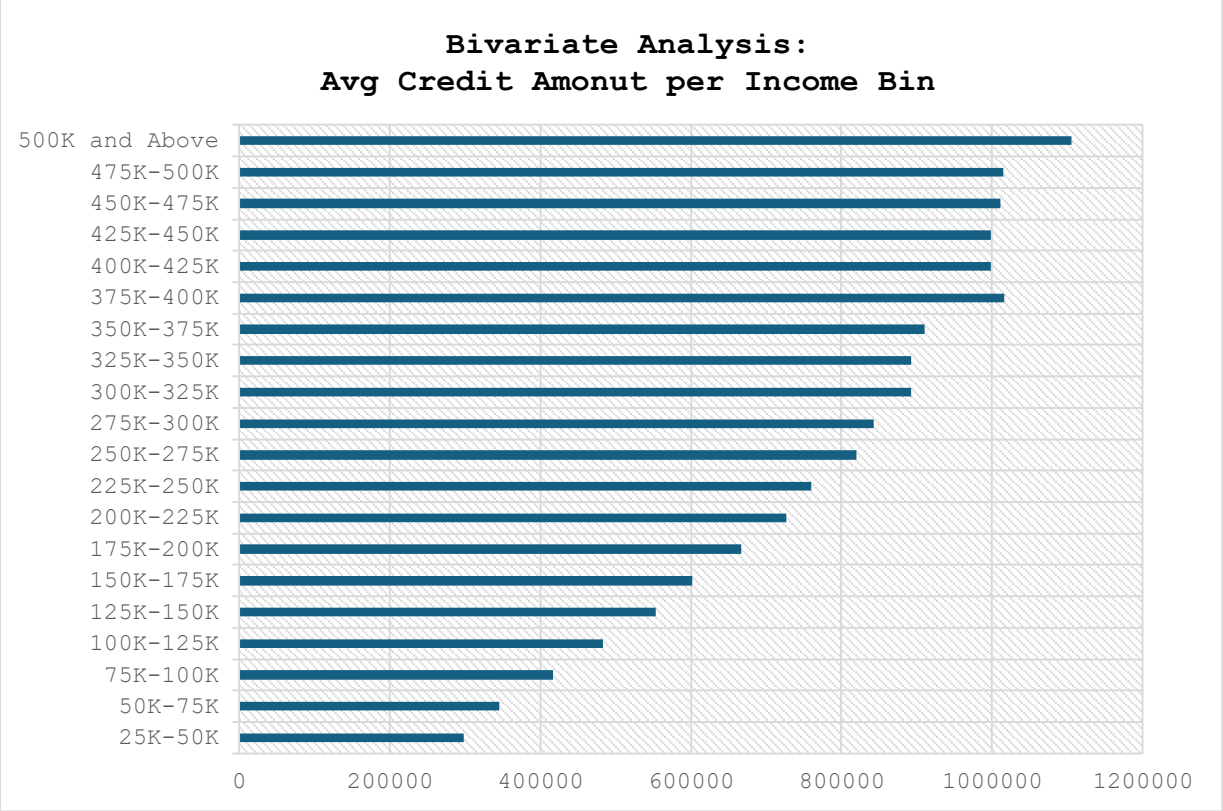
2. Segmented Univariate Analysis: Class Application per Income Bin

INCOME BIN	CLASS 0	CLASS 1
25K-50K	741	63
50K-75K	2980	246
75K-100K	5826	536
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
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 and Above	423	31



3. Bivariate Analysis: Average Credit Amount per Income Bin

INCOME BIN	Avg AMT_CREDIT
25K-50K	297752.0765
50K-75K	345240.3585
75K-100K	417267.8771
100K-125K	483568.8073
125K-150K	553042.1642
150K-175K	602034.4016
175K-200K	667004.421
200K-225K	727198.4449
225K-250K	759541.3782
250K-275K	820255.3451
275K-300K	842725.6488
300K-325K	892300.0718
325K-350K	892332.6503
350K-375K	910363.0482
375K-400K	1016814.375
400K-425K	999208.199
425K-450K	999153.6402
450K-475K	1011521.839
475K-500K	1015150.404
500K and Above	1105365.122



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.

Result:

CNT_CHILDREN	1	0.009588558	0.00497156
AMT_INCOME_TOTAL	0.009588558	1	0.069315897
AMT_CREDIT	0.00497156	0.069315897	1
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221
DAYS_BIRTH	0.329263754	0.016002774	-0.059342658
DAYS_EMPLOYED	-0.239693041	-0.031615555	-0.070471393
DAYS_REGISTRATION	0.181217183	0.009952379	0.003448569
DAYS_ID_PUBLISH	-0.032115773	0.003506646	-0.012228765
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT

CNT_CHILDREN	-0.025555665	0.329263754	-0.239693041
AMT_INCOME_TOTAL	0.029841469	0.016002774	-0.031615555
AMT_CREDIT	0.095111221	-0.059342658	-0.070471393
REGION_POPULATION_RELATIVE	1	-0.032513748	-0.004101686
DAYS_BIRTH	-0.032513748	1	-0.613553972
DAYS_EMPLOYED	-0.004101686	-0.613553972	1
DAYS_REGISTRATION	-0.059322344	0.333632509	-0.204680611
DAYS_ID_PUBLISH	-0.004345136	0.270825141	-0.270382022
REGION_RATING_CLIENT	-0.532667302	0.016779196	0.034321673
	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED

CNT_CHILDREN	0.181217183	-0.032115773	0.025913889
AMT_INCOME_TOTAL	-0.031615555	0.009952379	0.003506646
AMT_CREDIT	-0.070471393	0.003448569	-0.012228765
REGION_POPULATION_RELATIVE	-0.059322344	-0.004345136	-0.532667302
DAYS_BIRTH	0.333632509	0.270825141	0.016779196
DAYS_EMPLOYED	-0.204680611	-0.270382022	0.034321673
DAYS_REGISTRATION	1	0.104298561	0.087517643
DAYS_ID_PUBLISH	0.104298561	1	-0.002307011
REGION_RATING_CLIENT	0.087517643	-0.002307011	1
	DAYS_REGISTRATION	DAYS_ID_PUBLISH	REGION_RATING_CLIENT

Conclusion:

In this analysis of the loan application dataset, we effectively handled missing data using Excel functions to ensure data integrity. Outliers were identified and addressed, safeguarding the accuracy of our results. We assessed data imbalance and visualized the class distribution to understand potential biases. Univariate, segmented univariate, and bivariate analyses provided insights into the driving factors of loan defaults. Lastly, we identified top correlations for different scenarios, highlighting key indicators of loan default. These comprehensive analyses form a solid foundation for making data-driven decisions and improving loan application evaluations.