Bank Loan Case Study

Project Description

This project analyzes loan application data to identify factors influencing loan defaults. The goal is to help a finance company minimize risks by distinguishing high-risk applicants while ensuring capable borrowers are not rejected. Using Exploratory Data Analysis (EDA) in Microsoft Excel, we uncover patterns in customer and loan attributes that indicate repayment difficulties. These insights support data-driven decisions to reduce financial losses, improve approval strategies, and maintain customer satisfaction.

Approach

1. **Data Cleaning**:
    - Identified missing data in the dataset using Excel functions like COUNT, ISBLANK, and conditional formatting.
    - Dealt with missing values by applying appropriate imputation methods such as replacing with the median, mean, or other business-relevant values, ensuring data consistency and accuracy.
    - Verified the dataset for inconsistencies, duplicates, and formatting issues, ensuring a clean and analyzable dataset.
2. **Outlier Detection**:
    - Used statistical measures like QUARTILE and the Interquartile Range (IQR) method to detect outliers in numerical variables.
    - Highlighted outliers with conditional formatting and analyzed their potential impact on the results.
    - Validated if outliers were valid data points or errors requiring correction/removal. Box plots were used to visualize and interpret the outlier distributions.
3. **Finding Insights**:
    - Performed **Univariate Analysis** to understand the distribution of individual variables, using histograms and descriptive statistics.

- Conducted **Segmented Univariate Analysis** by comparing variable distributions across different scenarios, such as payment difficulties vs. all other cases, with grouped bar charts and pivot tables.
- Explored relationships between variables and the likelihood of default using **Bivariate Analysis**, employing scatter plots, correlation matrices, and heatmaps to identify strong indicators of repayment difficulty.

4. **Report Creation**:
    - Compiled insights, trends, and visualizations into a clear and structured report.
    - Included sections highlighting key factors influencing loan default, data distribution, and recommendations for risk mitigation.
    - Presented actionable findings in charts and tables for ease of understanding by the leadership team.

Tech-Stack Used

The project utilized Microsoft Excel 2022 for data cleaning, analysis, and visualization. Excel's built-in functions like COUNT, ISBLANK, AVERAGE, MEDIAN, QUARTILE, and CORREL were employed for identifying missing data, imputing values, detecting outliers, and analyzing correlations. Pivot tables and conditional formatting were used for segmented and bivariate analysis, while charts such as histograms, bar charts, box plots, and scatter plots helped visualize insights effectively. Excel’s versatility made it ideal for this exploratory analysis.

INSIGHTS

TASK A. Identify Missing Data and Deal with it Appropriately:

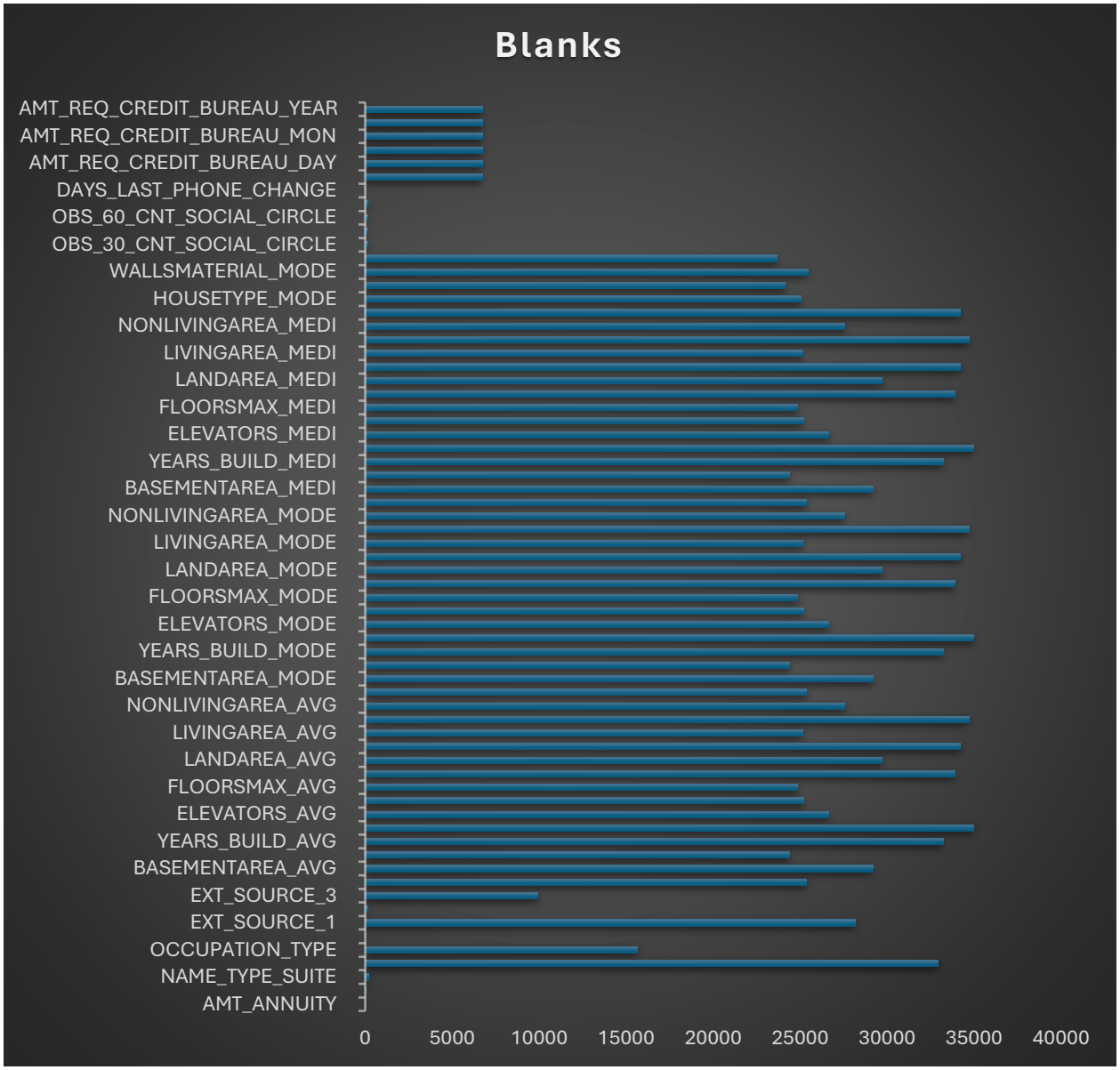
Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

S.NO.	Column Names	Blanks	Blanks Percentage	Dealing Method
10	AMT_ANNUITY	1	0%	Repaired
11	AMT_GOODS_PRICE	38	0%	Repaired
12	NAME_TYPE_SUITE	192	0%	Repaired
22	OWN_CAR_AGE	32950	66%	Deleted
29	OCCUPATION_TYPE	15654	31%	Repaired
30	CNT_FAM_MEMBERS	1	0%	Repaired
42	EXT_SOURCE_1	28172	56%	Deleted
43	EXT_SOURCE_2	126	0%	Repaired
44	EXT_SOURCE_3	9944	20%	Repaired
45	APARTMENTS_AVG	25385	51%	Deleted
46	BASEMENTAREA_AVG	29199	58%	Deleted
47	YEARS_BEGINEXPLUATATION_AVG	24394	49%	Deleted
48	YEARS_BUILD_AVG	33239	66%	Deleted
49	COMMONAREA_AVG	34960	70%	Deleted
50	ELEVATORS_AVG	26651	53%	Deleted
51	ENTRANCES_AVG	25195	50%	Deleted
52	FLOORSMAX_AVG	24875	50%	Deleted
53	FLOORSMIN_AVG	33894	68%	Deleted
54	LANDAREA_AVG	29721	59%	Deleted
55	LIVINGAPARTMENTS_AVG	34226	68%	Deleted
56	LIVINGAREA_AVG	25137	50%	Deleted
57	NONLIVINGAPARTMENTS_AVG	34714	69%	Deleted
58	NONLIVINGAREA_AVG	27572	55%	Deleted
59	APARTMENTS_MODE	25385	51%	Deleted
60	BASEMENTAREA_MODE	29199	58%	Deleted
61	YEARS_BEGINEXPLUATATION_MODE	24394	49%	Deleted
62	YEARS_BUILD_MODE	33239	66%	Deleted
63	COMMONAREA_MODE	34960	70%	Deleted
64	ELEVATORS_MODE	26651	53%	Deleted
65	ENTRANCES_MODE	25195	50%	Deleted
66	FLOORSMAX_MODE	24875	50%	Deleted
67	FLOORSMIN_MODE	33894	68%	Deleted
68	LANDAREA_MODE	29721	59%	Deleted

69	LIVINGAPARTMENTS_MODE	34226	68%	Deleted
70	LIVINGAREA_MODE	25137	50%	Deleted
71	NONLIVINGAPARTMENTS_MODE	34714	69%	Deleted
72	NONLIVINGAREA_MODE	27572	55%	Deleted
73	APARTMENTS_MEDI	25385	51%	Deleted
74	BASEMENTAREA_MEDI	29199	58%	Deleted
75	YEARS_BEGINEXPLUATATION_MEDI	24394	49%	Deleted
76	YEARS_BUILD_MEDI	33239	66%	Deleted
77	COMMONAREA_MEDI	34960	70%	Deleted
78	ELEVATORS_MEDI	26651	53%	Deleted
79	ENTRANCES_MEDI	25195	50%	Deleted
80	FLOORSMAX_MEDI	24875	50%	Deleted
81	FLOORSMIN_MEDI	33894	68%	Deleted
82	LANDAREA_MEDI	29721	59%	Deleted
83	LIVINGAPARTMENTS_MEDI	34226	68%	Deleted
84	LIVINGAREA_MEDI	25137	50%	Deleted
85	NONLIVINGAPARTMENTS_MEDI	34714	69%	Deleted
86	NONLIVINGAREA_MEDI	27572	55%	Deleted
87	FONDKAPREMONT_MODE	34191	68%	Deleted
88	HOUSETYPE_MODE	25075	50%	Deleted
89	TOTALAREA_MODE	24148	48%	Deleted
90	WALLSMATERIAL_MODE	25459	51%	Deleted
91	EMERGENCYSTATE_MODE	23698	47%	Deleted
92	OBS_30_CNT_SOCIAL_CIRCLE	168	0%	Repaired
93	DEF_30_CNT_SOCIAL_CIRCLE	168	0%	Repaired
94	OBS_60_CNT_SOCIAL_CIRCLE	168	0%	Repaired
95	DEF_60_CNT_SOCIAL_CIRCLE	168	0%	Repaired
96	DAYS_LAST_PHONE_CHANGE	1	0%	Repaired
117	AMT_REQ_CREDIT_BUREAU_HOUR	6734	13%	Repaired
118	AMT_REQ_CREDIT_BUREAU_DAY	6734	13%	Repaired
119	AMT_REQ_CREDIT_BUREAU_WEEK	6734	13%	Repaired
120	AMT_REQ_CREDIT_BUREAU_MON	6734	13%	Repaired
121	AMT_REQ_CREDIT_BUREAU_QRT	6734	13%	Repaired
122	AMT_REQ_CREDIT_BUREAU_YEAR	6734	13%	Repaired

Proportion of missing values	Dealing Method
<35%	Numerical value- Average or Median imputations
	Categorical value- Mode imputation
>35%	Removing the columns

Dealing with Numerical Values	Median
AMT_ANNUITY	24939
AMT_GOODS_PRICE	450000
CNT_FAM_MEMBERS	2
OBS_30_CNT_SOCIAL_CIRCLE	0
DEF_30_CNT_SOCIAL_CIRCLE	0
OBS_60_CNT_SOCIAL_CIRCLE	0
DEF_60_CNT_SOCIAL_CIRCLE	0
EXT_SOURCE_2	0.58859369
EXT_SOURCE_3	0.53527625
DAYS_LAST_PHONE_CHANGE	-742
AMT_REQ_CREDIT_BUREAU_HOUR	0
AMT_REQ_CREDIT_BUREAU_DAY	0
AMT_REQ_CREDIT_BUREAU_WEEK	0
AMT_REQ_CREDIT_BUREAU_MON	0
AMT_REQ_CREDIT_BUREAU_QRT	0
AMT_REQ_CREDIT_BUREAU_YEAR	1



1.) Identified Missing Data and Blank Percentages:

- Checked each column for missing values and calculated the percentage of blanks by using COUNTBLANK function.
- Found that several columns had missing values, ranging from very low percentages (e.g., AMT_ANNUITY at 0%) to very high percentages (e.g., COMMONAREA_AVG at 70%).

2.) Handling Missing Data:

- **For Columns with <35% Missing Data:**
  - Imputed missing values using the following strategies:
    - Numerical Columns (Median Imputation):
      - Columns: AMT_ANNUITY, AMT_GOODS_PRICE, EXT_SOURCE_2, etc.
    - Categorical Columns (Mode Imputation):
      - OCCUPATION_TYPE (Mode: "Labourers").
      - NAME_TYPE_SUITE (Mode: "Unaccompanied").
- **For Columns with >35% Missing Data:**
  - Removed columns where more than 35% of the data was missing, as imputing such data could introduce noise or bias.
  - Removed Columns: OWN_CAR_AGE, EXT_SOURCE_1, APARTMENTS_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, FLOORSMIN_AVG, LIVINGAPARTMENTS_AVG, and others with similar proportions.

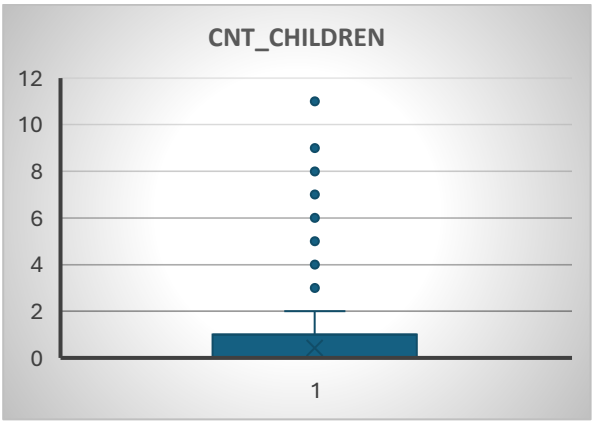
TASK B. Identify Outliers in the Dataset

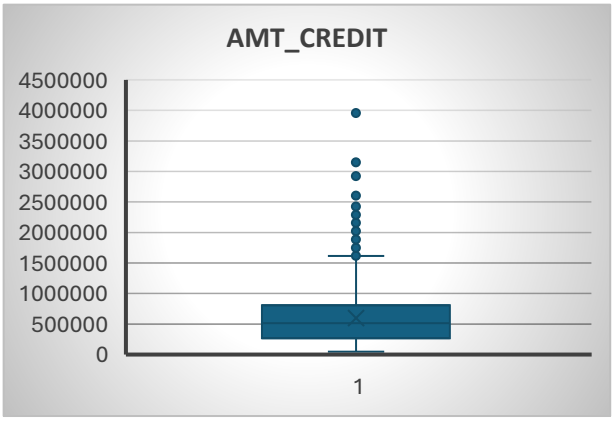
Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

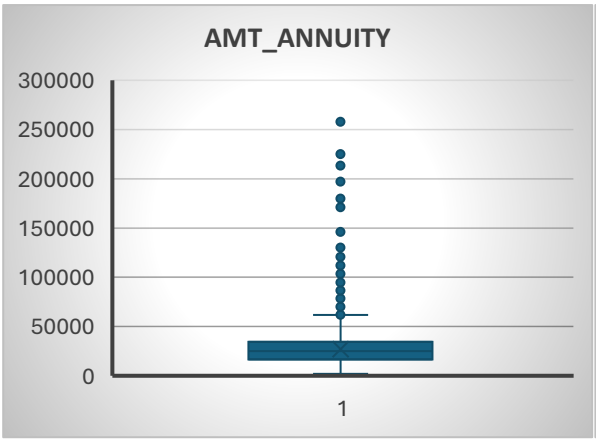
Columns	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION
Q1	0	112500	270000	16456.5	238500	0.010006	33.9123288	2.556164384	5.473972603
Q3	1	202500	808650	34596	679500	0.028663	53.8191781	15.66575342	20.44931507
IQR	1	90000	538650	18139.5	441000	0.018657	19.9068493	13.10958904	14.97534247
Upper Limit	2.5	337500	1616625	61805.25	1341000	0.0566485	83.6794521	35.33013699	42.91232877
Lower Limit	-1.5	-22500	-537975	-10752.75	-423000	-0.0179795	4.05205479	-17.10821918	-16.9890411
Outlier Below	0	0	0	0	0	0	0	0	0
Outlier Above	723	2295	1063	1188	2387	1329	0	9082	96
Total Outliers	723	2295	1063	1188	2387	1329	0	9082	96
Max	11	117000000	4050000	258025.5	4050000	0.072508	68.9972603	1000.665753	61.34794521
Outlier Validity	Not valid	Not valid	Valid	Valid	Valid	Valid	No Outlier	Not Valid	Valid

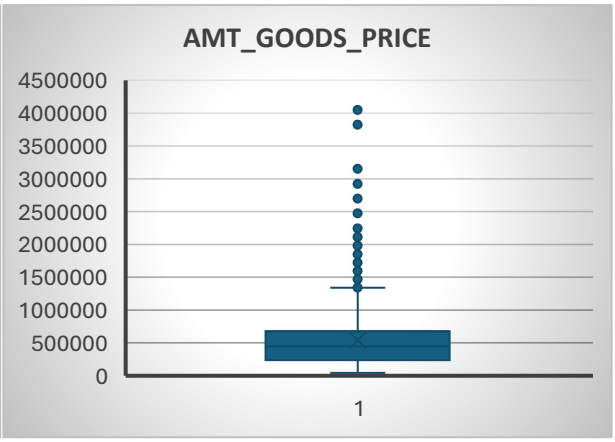
I have used the **Interquartile Range (IQR) Method** to identify outliers in the numerical columns of the dataset. This method is robust and effective in detecting outliers by analyzing the spread of the middle 50% of the data. Here's how the calculations were performed for each numerical column:

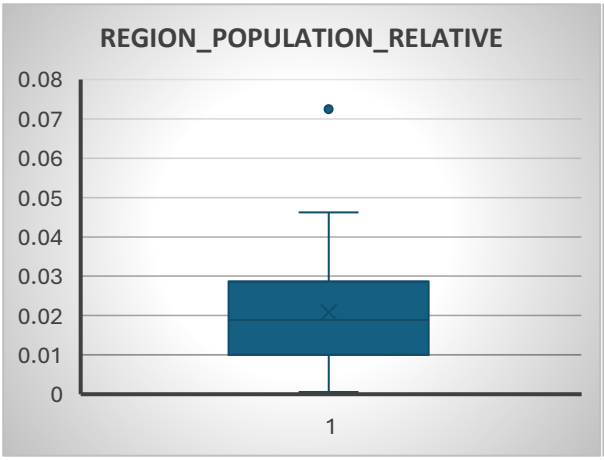
1. **Quartiles Calculation:**
    - Calculated the first quartile (Q1, 25th percentile) and third quartile (Q3, 75th percentile) using the QUARTILE.EXC function in Excel.
        - Q1 Formula: =QUARTILE.EXC(range, 1)
        - Q3 Formula: =QUARTILE.EXC(range, 3)
2. **Interquartile Range (IQR):**
    - Determined the IQR as the difference between Q3 and Q1.
        - Formula: IQR = Q3 - Q1
3. **Upper and Lower Limits:**
    - Calculated thresholds to detect outliers:
        - Upper Limit: Q3 + 1.5 * IQR
        - Lower Limit: Q1 - 1.5 * IQR

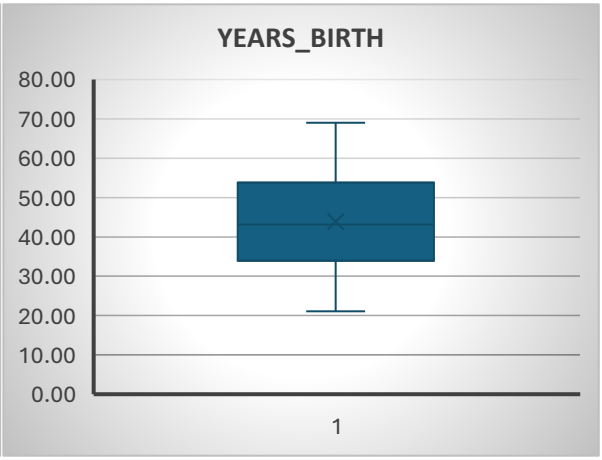

CNT_CHILDREN
12
10
8
6
4
2
0
1

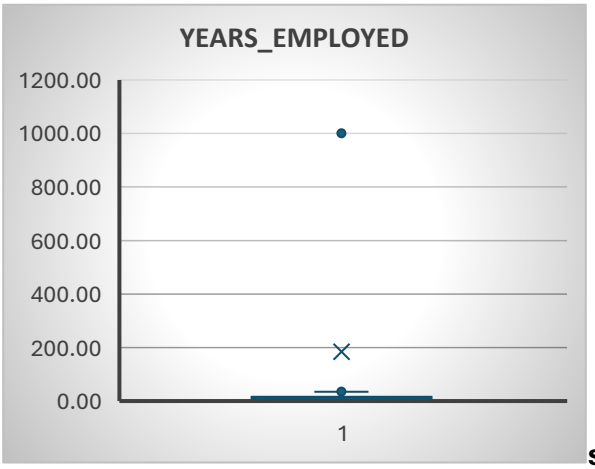

AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

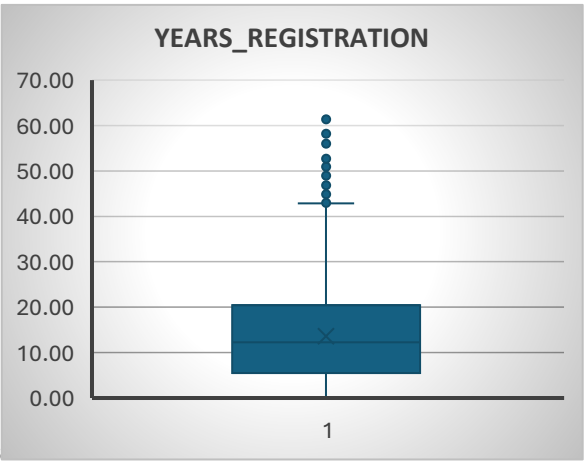

AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
1


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


REGION_POPULATION_RELATIVE
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1


YEARS_BIRTH
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1


YEARS_EMPLOYED
1200.00
1000.00
800.00
600.00
400.00
200.00
0.00
1


YEARS_REGISTRATION
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1

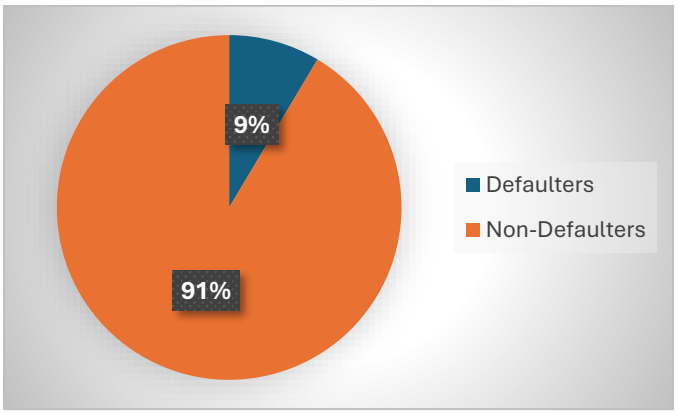
s

1. **CNT_CHILDREN (Number of Children)**
    - Outliers (723): Invalid (max = 11). Likely data errors.
    - Action: Clean data to avoid misclassification of affordability.
2. **AMT_INCOME_TOTAL (Total Income)**
    - Outliers (2,295): Invalid (max = $117M). Unrealistic values.
    - Action: Cap or clean to prevent income skewing risk analysis.
3. **AMT_CREDIT (Loan Amount)**
    - Outliers (1,063): Valid (max = $4M). High-value loans.
    - Action: Scrutinize for income-to-loan ratio risks.
4. **AMT_ANNUITY (Loan Annuity)**
    - Outliers (1,188): Valid. High annuities = significant obligations.
    - Action: Monitor applicants with high annuities vs. income.
5. **AMT_GOODS_PRICE (Goods Price)**
    - Outliers (2,387): Valid. Reflects high-value assets.
    - Action: Assess affordability for high-priced purchases.
6. **REGION_POPULATION_RELATIVE (Population of Region)**
    - Outliers (1,329): Valid. Represents densely populated areas.
    - Action: Consider economic opportunities in urban areas.
7. **YEARS_BIRTH (Age)**
    - No Outliers. Data is valid.
8. **YEARS_EMPLOYED (Employment Duration)**
    - Outliers (9,082): Invalid (max = 1,000). Data errors.
    - Action: Clean for accurate stability analysis.
9. **YEARS_REGISTRATION (Registration Duration)**
    - Outliers (96): Valid. Indicates stability.
    - Action: Positive factor for risk evaluation.

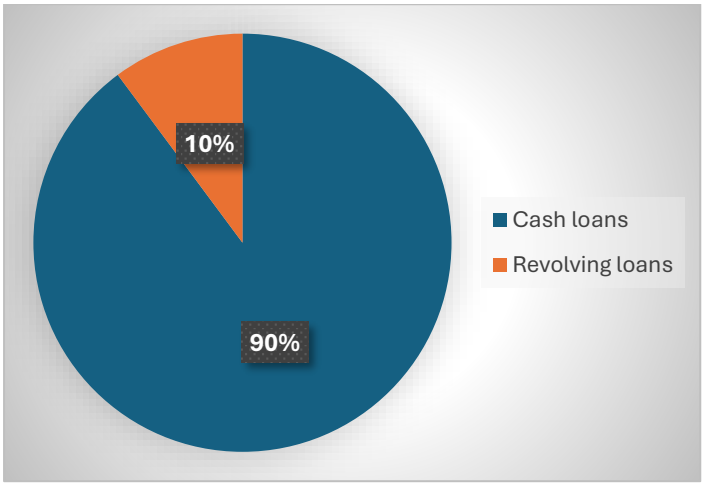
TASK C. Analyze Data Imbalance

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

Target	%
Defaulters	9%
Non-Defaulters	91%
Ratio of Data Imbalance(Min/Max)	0.09



Loan Type	%
Cash loans	90%
Revolving loans	10%
Ratio of Data Imbalance(Min/Max)	0.11



1. Defaulters vs. Non-Defaulters

- **Distribution:** Defaulters = 9%, Non-Defaulters = 91%.
- **Imbalance Ratio:** 0.09 (9:91).
- **Impact:** The dataset is highly imbalanced, with far fewer defaulters. This could bias models toward predicting non-defaults and reduce the ability to identify risky applicants.

2. Loan Types: Cash Loans vs. Revolving Loans

- **Distribution:** Cash Loans = 90%, Revolving Loans = 10%.
- **Imbalance Ratio:** 0.11 (10:90).
- **Impact:** Cash loans dominate the dataset. This may limit the model's ability to learn patterns specific to revolving loans.

TASK D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

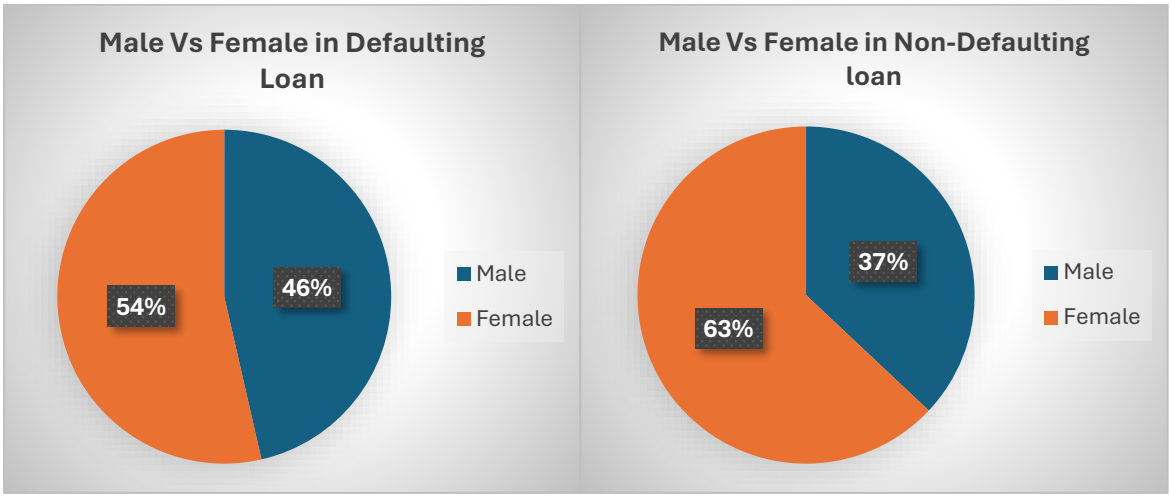
Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

1.) UNIVARIATE ANALYSIS

a) Analysis Based on Gender

% Defaulters	
Male	**Female**
46%	54%

% Non-Defaulters	
Male	**Female**
37%	63%



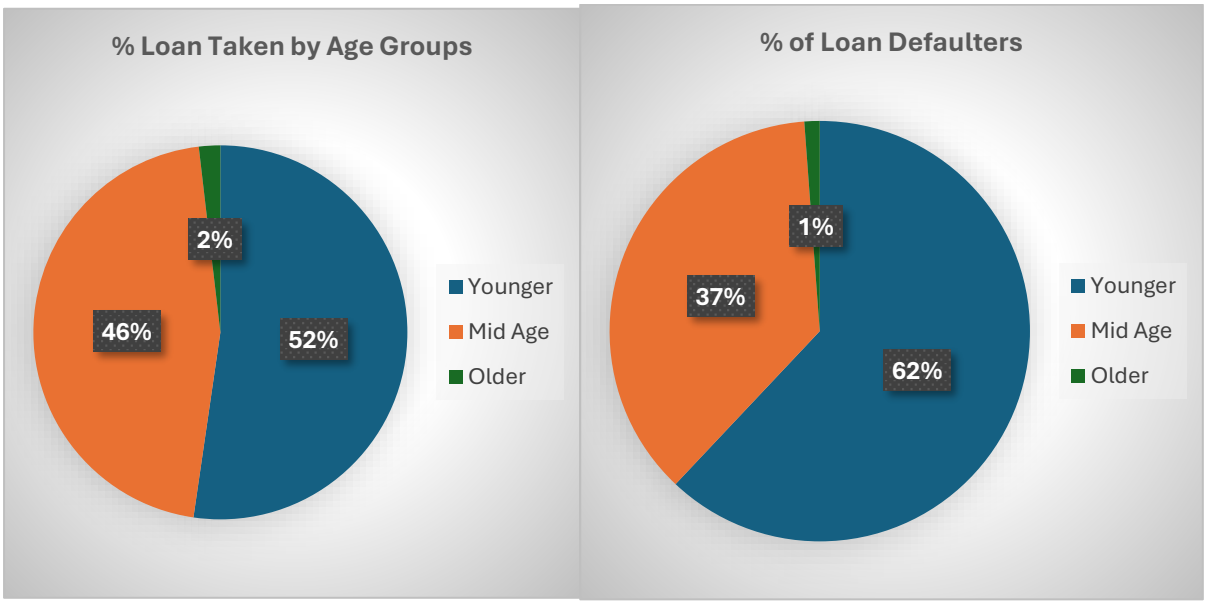
- A higher proportion of female applicants (54%) default compared to males (46%).
- However, females also dominate the non-defaulters category, indicating they make up a larger portion of the applicant pool.

b) Analysis Based on Age

Age Group	Younger	Mid Age	Older
Age Range(Years)	20-40	41-60	>61
Total	20616	18074	728
Defaulters	2114	1254	40
Non-Defaulters	18502	16820	688

% Loan Taken by Age Groups		
Younger	Mid Age	Older
52%	46%	2%

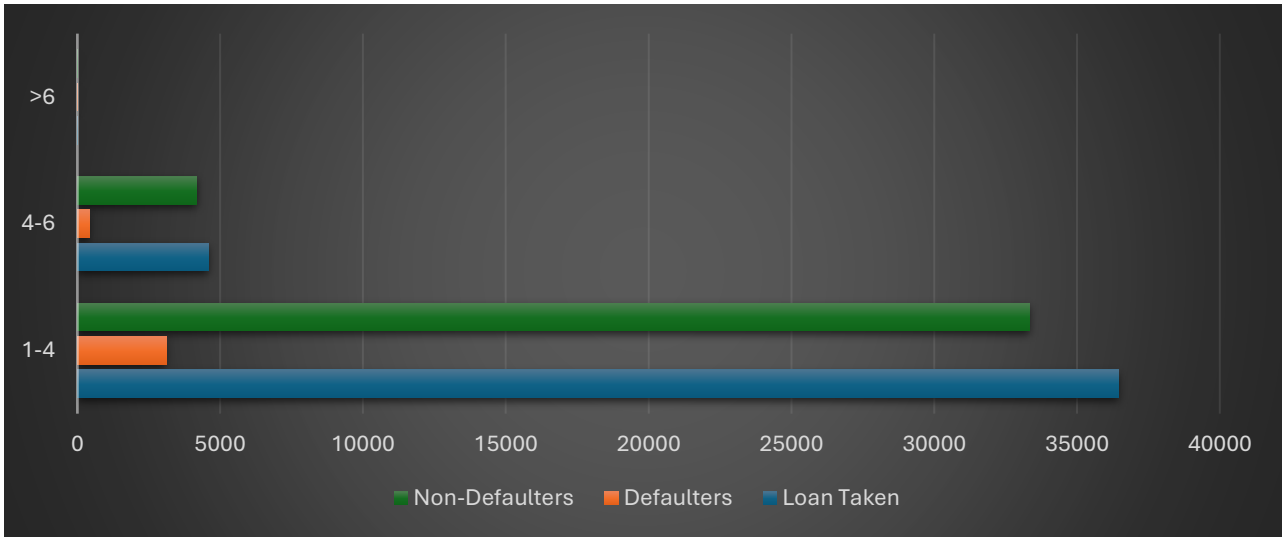
% of Loan Defaulters		
Younger	Mid Age	Older
62%	37%	1%



- Younger applicants take more loans but also have a significantly higher default rate (62%).
- Older applicants rarely take loans and have the lowest default rate (1%).

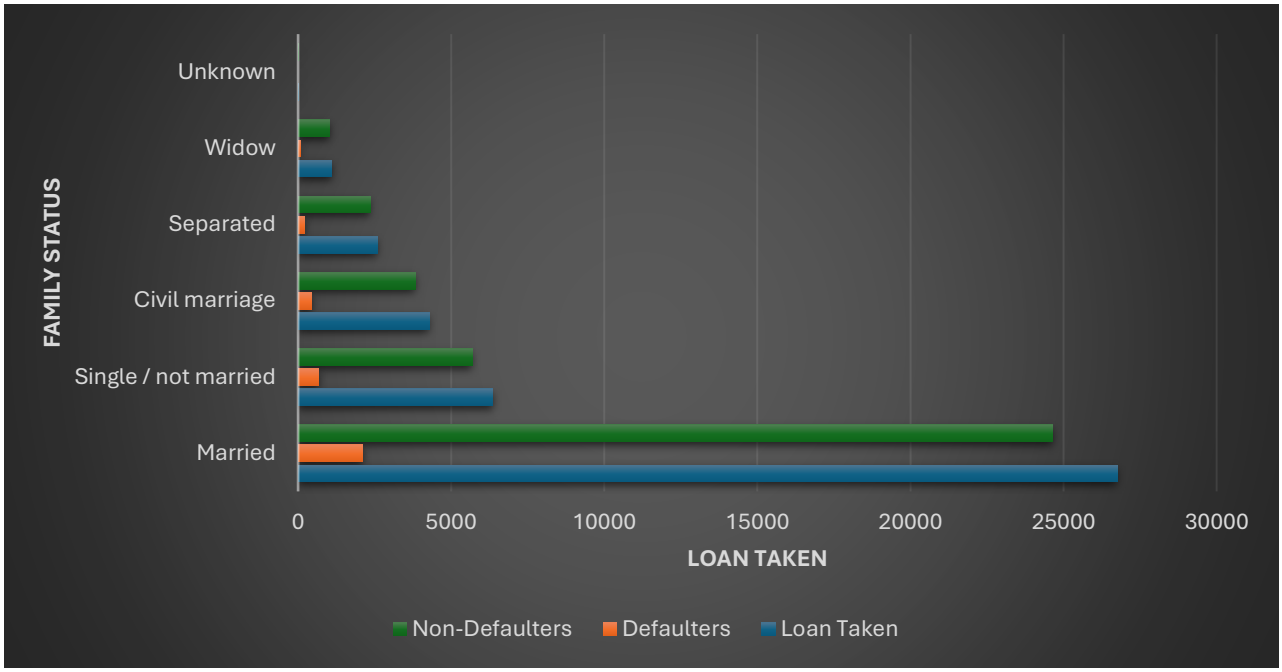
c) Analysis Based on Family

No. of Family Members	Loan Taken	Defaulters	Non-Defaulters	% Defaulters
1-4	36457	3109	33348	9%
4-6	4591	408	4183	9%
>6	21	3	18	14%



- Families with more than 6 members have a noticeably higher default rate (14%) despite accounting for a very small percentage of loans.

Family Status	Loan Taken	Defaulters	Non-Defaulters	%Defaulters
Married	26758	2103	24655	8%
Single / not married	6352	672	5680	11%
Civil marriage	4277	453	3824	11%
Separated	2578	225	2353	9%
Widow	1103	67	1036	6%
Unknown	1	0	1	0%

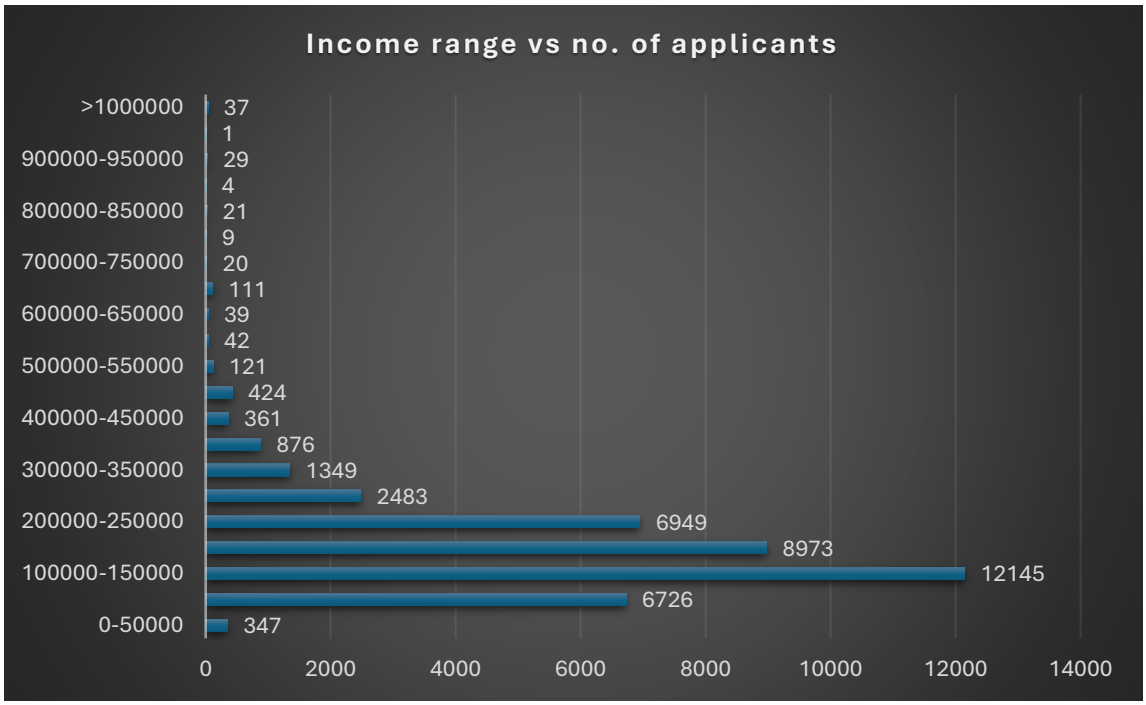


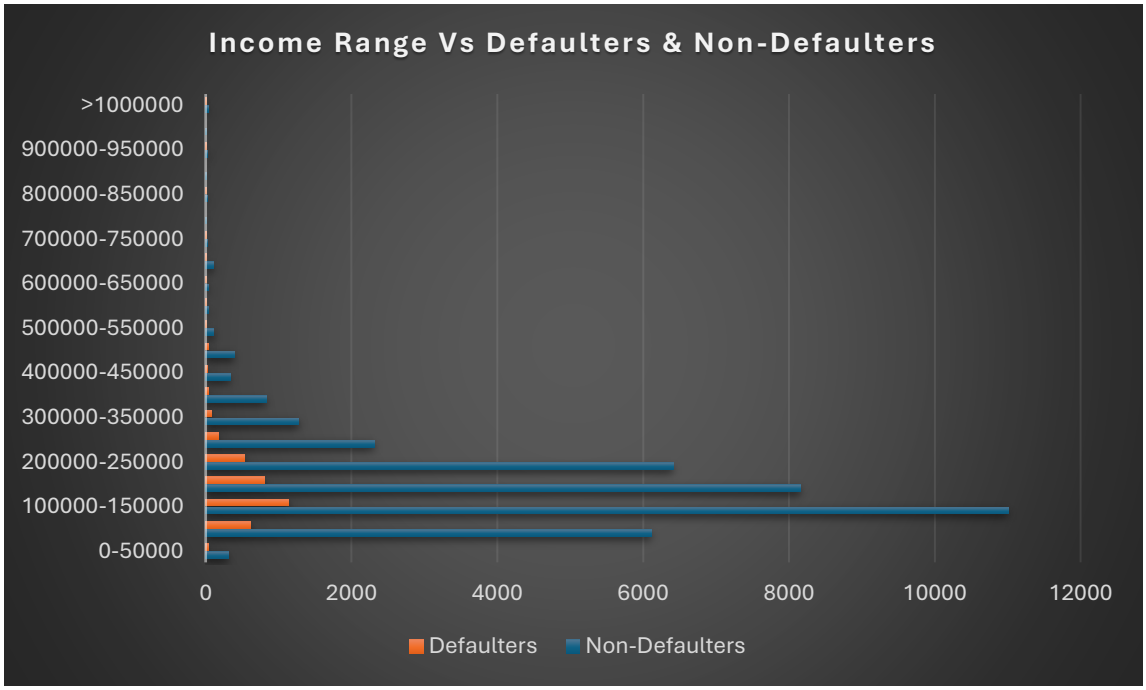
- Single or civil marriage applicants show the highest default rates (11%).
- Widows have the lowest default rate (6%), indicating better repayment reliability.

2.) SEGMENTED UNIVARIATE ANALYSIS

a) Analysis on the basis of Income Range

Income Range	Total Applicant	Non-Defaulters	Defaulters
0-50000	347	310	37
50000-100000	6726	6109	617
100000-150000	12145	11006	1139
150000-200000	8973	8162	811
200000-250000	6949	6414	535
250000-300000	2483	2312	171
300000-350000	1349	1273	76
350000-400000	876	831	45
400000-450000	361	335	26
450000-500000	424	391	33
500000-550000	121	112	9
550000-600000	42	37	5
600000-650000	39	38	1
650000-700000	111	104	7
700000-750000	20	19	1
750000-800000	9	9	0
800000-850000	21	19	2
850000-900000	4	4	0
900000-950000	29	27	2
950000-1000000	1	1	0
>1000000	37	36	1



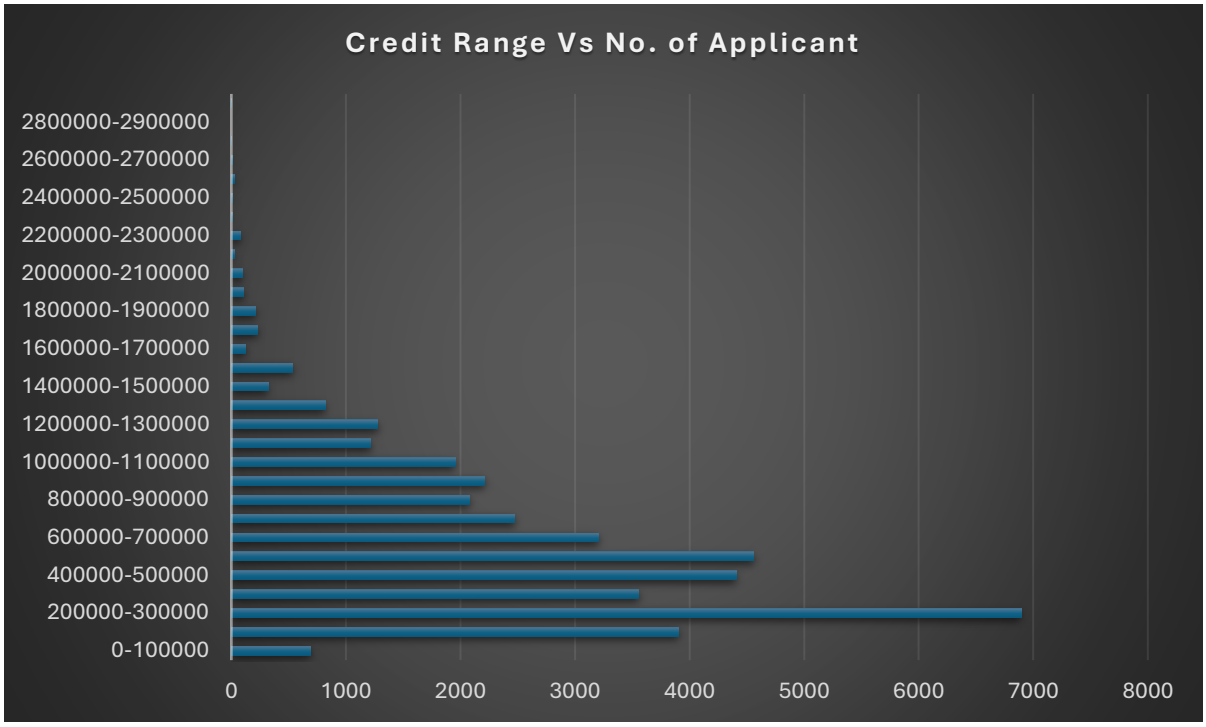


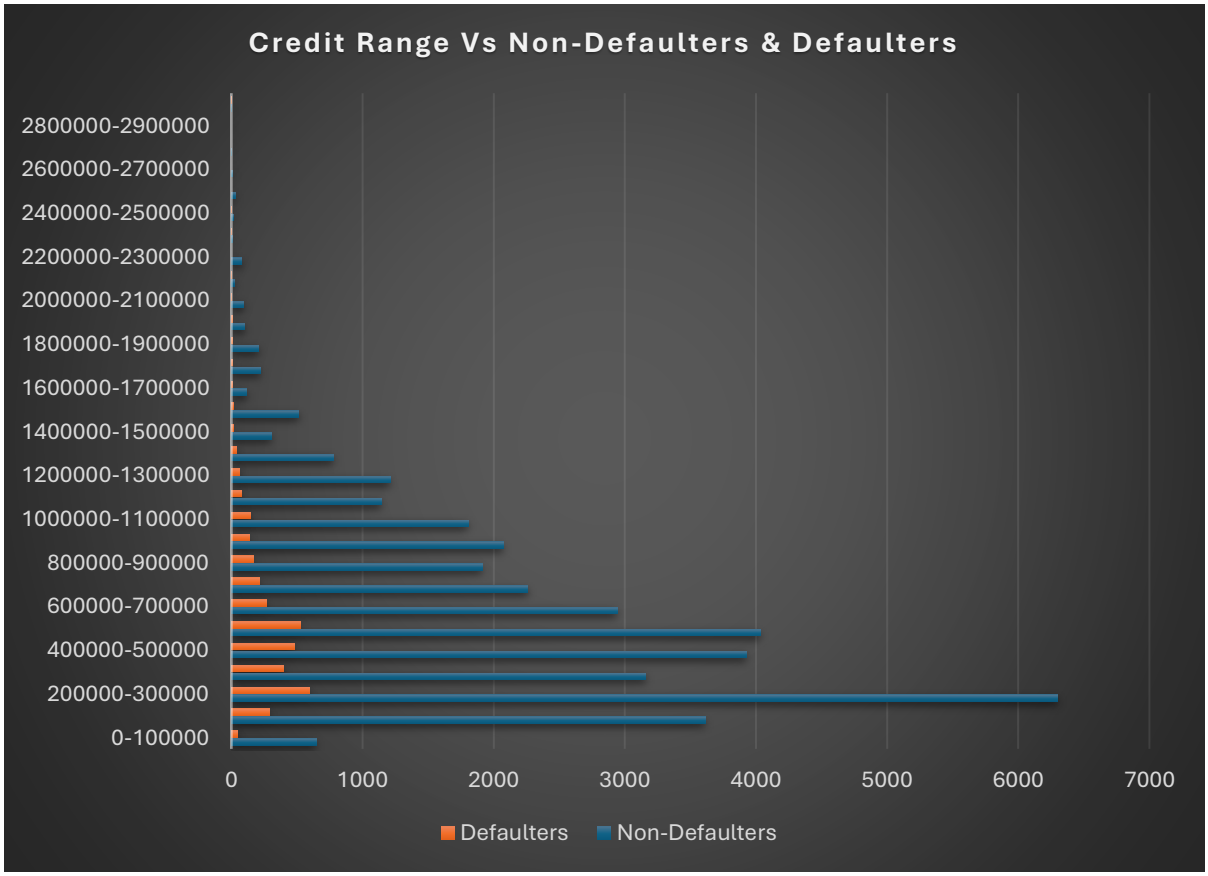
- **Lower Income (<50K):** High default rate (37 out of 347 = 10.7%), which could indicate financial instability.
- **Higher Income (>=500K):** Much lower default rates (e.g., 1 out of 37 for >$1M).

  As income increases, the **default rate decreases** significantly, with the highest income group showing almost no defaults.

b) Analysis on the basis of Amount Credit

Credit Range	Total Applicant	Non-Defaulters	Defaulters
0-100000	695	648	47
100000-200000	3903	3612	291
200000-300000	6895	6302	593
300000-400000	3555	3158	397
400000-500000	4407	3927	480
500000-600000	4559	4032	527
600000-700000	3209	2943	266
700000-800000	2473	2260	213
800000-900000	2079	1912	167
900000-1000000	2209	2074	135
1000000-1100000	1954	1807	147
1100000-1200000	1219	1143	76
1200000-1300000	1278	1213	65
1300000-1400000	822	779	43
1400000-1500000	320	306	14
1500000-1600000	535	515	20
1600000-1700000	125	116	9
1700000-1800000	229	220	9
1800000-1900000	215	209	6
1900000-2000000	107	102	5
2000000-2100000	96	92	4
2100000-2200000	28	26	2
2200000-2300000	79	79	0
2300000-2400000	13	12	1
2400000-2500000	14	13	1
2500000-2600000	29	29	0
2600000-2700000	11	11	0
2700000-2800000	2	2	0
2800000-2900000	0	0	0
2900000-3000000	3	2	1

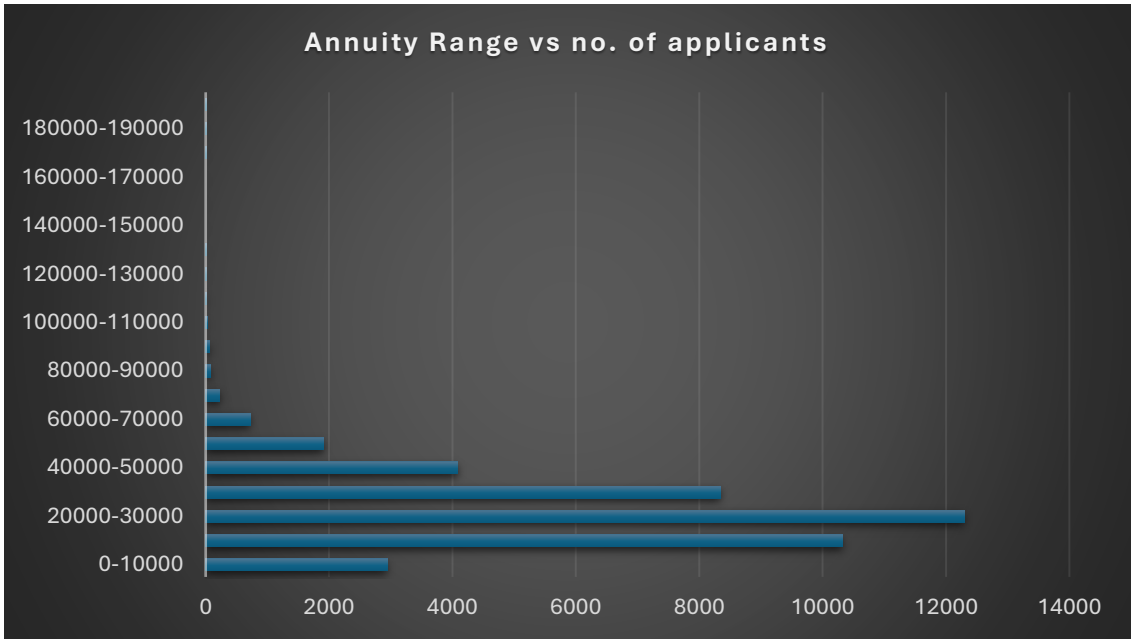


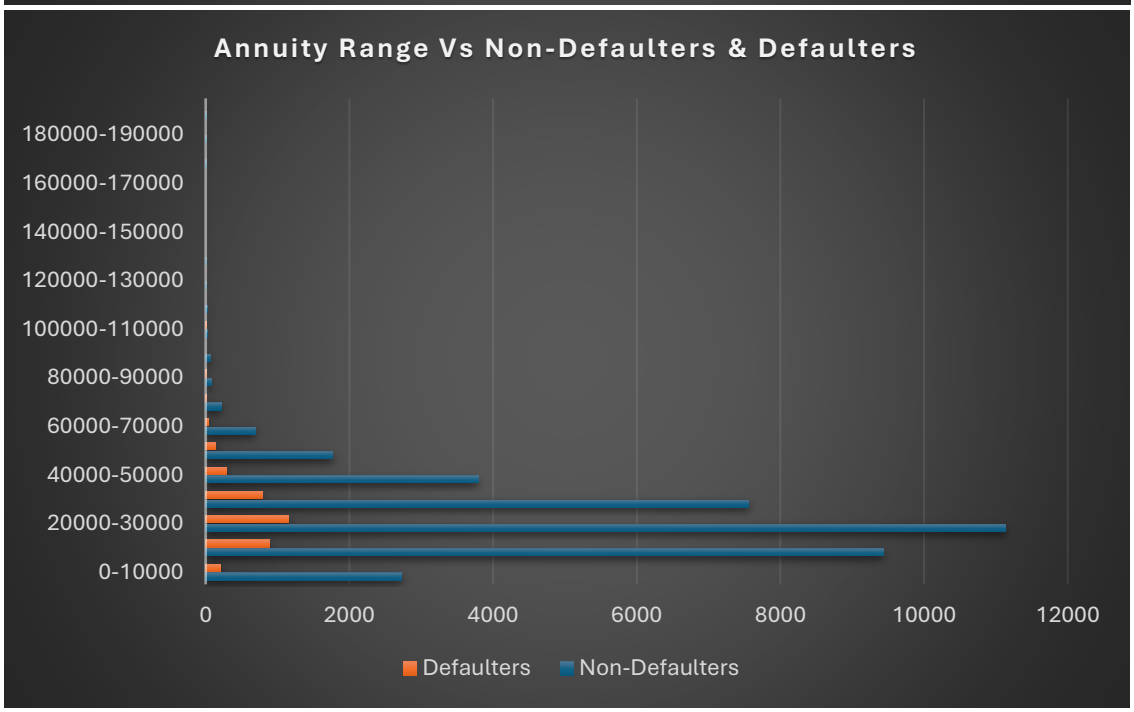


- **Lower Credit (<$100K):** Default rate is 47 out of 695 = 6.7%.
- **Higher Credit (>=$1.5M):** Default rates drop as the credit amount increases.

c) Analysis on the basts of Amount Annuity

Annuity Range	Total Applicant	Non-Defaulters	Defaulters
0-10000	2939	2730	209
10000-20000	10325	9436	889
20000-30000	12294	11139	1155
30000-40000	8348	7559	789
40000-50000	4088	3797	291
50000-60000	1908	1769	139
60000-70000	722	688	34
70000-80000	230	222	8
80000-90000	78	74	4
90000-100000	67	67	0
100000-110000	22	21	1
110000-120000	20	20	0
120000-130000	7	7	0
130000-140000	8	8	0
140000-150000	0	0	0
150000-160000	0	0	0
160000-170000	0	0	0
170000-180000	4	4	0
180000-190000	1	1	0
>190000	6	6	0

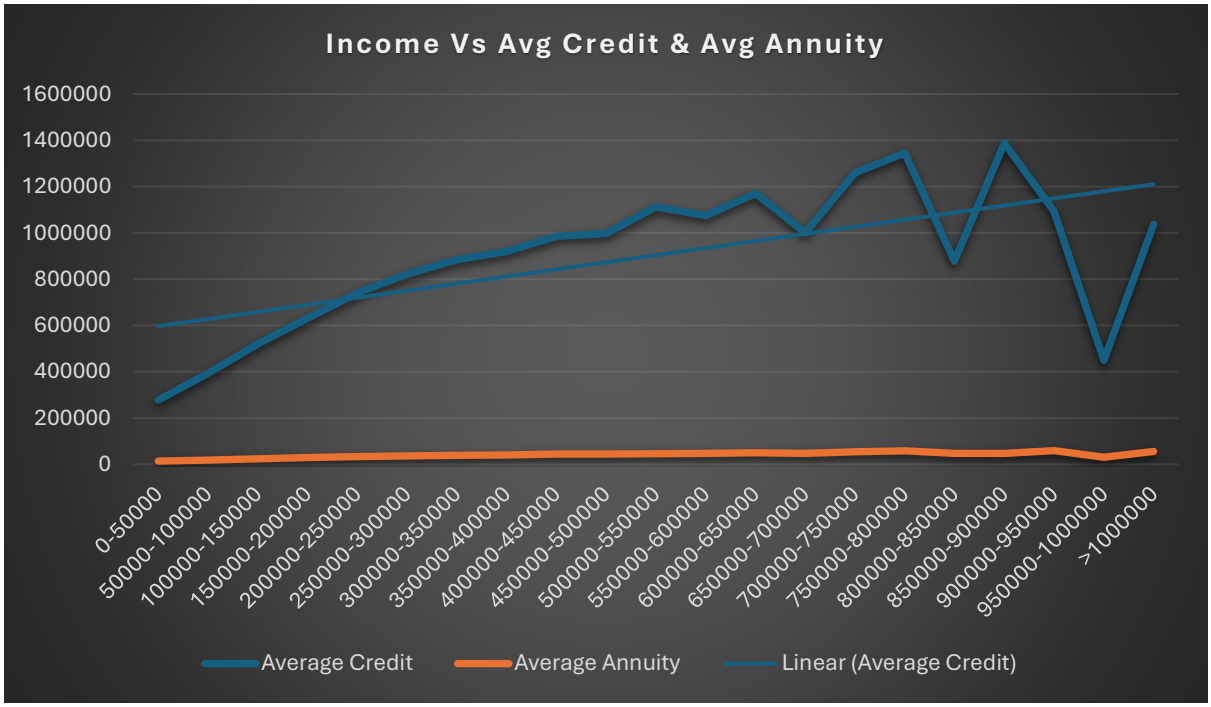




- **Lower Annuities (<10K):** Default rate of 209 out of 2,939 = 7.1%, possibly indicating affordability issues with smaller loans.
- **Higher Annuities (>=100K):** Very low default rates (only 1 or 0 defaults), which might suggest more reliable borrowers.

3) Bivariate Analysis

Income Range	Average Credit	Average Annuity
0-50000	277298.05	13640.21
50000-100000	393349.85	18704.74
100000-150000	519709.47	24009.44
150000-200000	630878.77	28654.87
200000-250000	740833.61	33007.89
250000-300000	821826.37	36100.31
300000-350000	884090.21	39301.63
350000-400000	920791.10	40810.46
400000-450000	985704.88	44097.51
450000-500000	997944.62	44784.76
500000-550000	1112433.21	45984.46
550000-600000	1074844.07	46788.96
600000-650000	1171325.88	49857.69
650000-700000	1000031.84	47379.41
700000-750000	1259983.35	53482.28
750000-800000	1344940.00	58803.00
800000-850000	876760.07	47799.86
850000-900000	1388400.75	47631.38
900000-950000	1093675.66	58976.22
950000-1000000	450000.00	30073.50
>1000000	1037054.80	54840.04



1. Relationship Between Income and Average Credit

As income increases, the **average credit amount** also increases.

Observations:

- This suggests that higher-income applicants tend to be approved for larger loan amounts, which is likely due to a better repayment capacity.
- The correlation appears to be positive, meaning the higher the income, the higher the average credit extended.

2. Relationship Between Income and Average Annuity

As income increases, the **average annuity** also increases, though not as consistently as average credit.

Observations:

- The average annuity rises steadily with income, but the increase is more gradual compared to the average credit.
- This indicates that higher-income applicants not only get larger loans but also have higher monthly repayment obligations.

Positive Correlation: Both credit and annuity amounts increase with higher income, but the growth in average credit is more pronounced.

TASK E. Identify Top Correlations for Different Scenarios

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.

I have created a **correlation matrix** for two cases (target=0 and target=1) using Excel's **Data Analysis Tool**. This helped identify the top correlations between variables for **non-defaulters** and **defaulters**, providing insights into the factors that influence loan repayment and default risk.

a) For target=0 (Non-Defaulters)

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH
AMT_INCOME_TOTAL	1								
AMT_CREDIT	0.360011781	1							
AMT_ANNUITY	0.431026919	0.760827873	1						
AMT_GOODS_PRICE	0.367729007	0.98635817	0.765201743	1					
REGION_POPULATION_RELATIVE	0.188785867	0.09654213	0.116108426	0.099641681	1				
YEARS_BIRTH	0.049536299	0.160878908	0.099056566	0.155143369	0.048987416	1			
YEARS_EMPLOYED	0.036221572	0.094943177	0.06213425	0.095918555	-0.005606334	0.352389434	1		
YEARS_REGISTRATION	-0.033500961	0.024294688	0.000851932	0.020392259	0.064621189	0.304733514	0.175692735	1	
YEARS_ID_PUBLISH	0.023115928	0.044246818	0.03222207	0.044495693	0.004355924	0.107692262	0.08250215	0.037023396	1

Top 5 Correlation (Non-Defaulters)		
Variable 1	**Variable 2**	**Correlation**
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
AMT_GOODS_PRICE	AMT_CREDIT	0.986
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.948
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.853

- **Social Circle:** Very high correlation between **OBS_60_CNT_SOCIAL_CIRCLE** and **OBS_30_CNT_SOCIAL_CIRCLE** (0.998), showing social network stability impacts loan repayment.
- **Loan Amount & Goods Price:** Strong correlation (0.986) between **AMT_GOODS_PRICE** and **AMT_CREDIT**, suggesting larger loans are tied to higher-value goods.
- **Region Rating: REGION_RATING_CLIENT_W_CITY** and **REGION_RATING_CLIENT** show high correlation (0.948), indicating location influences creditworthiness.

b) For target = 1(Defaulters)

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED
AMT_INCOME_TOTAL	1						
AMT_CREDIT	0.312173644	1					
AMT_ANNUITY	0.371245075	0.745132112	1				
AMT_GOODS_PRICE	0.313726831	0.981928143	0.746422447	1			
REGION_POPULATION_RELATIVE	0.096758897	0.055597704	0.06586731	0.061151451	1		
YEARS_BIRTH	0.087629893	0.194437534	0.086228175	0.18810843	0.013409076	1	
YEARS_EMPLOYED	0.022601082	0.105109669	0.054426768	0.113070145	-0.001640893	0.305741728	1

Top 5 Correlation (Defaulters)		
Variable 1	**Variable 2**	**Correlation**
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
AMT_GOODS_PRICE	AMT_CREDIT	0.982
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.951
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.891
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.806

- **Social Circle:** Similar strong correlation between **OBS_60_CNT_SOCIAL_CIRCLE** and **OBS_30_CNT_SOCIAL_CIRCLE** (0.998), indicating social support impacts defaults.
- **Loan Amount & Goods Price:** Strong correlation (0.982) remains, suggesting expensive loans are riskier.
- **Region Rating:** High correlation (0.951) between region ratings, showing geography's role in defaults.

RESULT

Through the project, I successfully analyzed various factors influencing loan defaults by performing Exploratory Data Analysis (EDA) and statistical evaluations. I achieved the following:

1. **Outlier Detection:** Identified significant outliers in the dataset that could distort the analysis, ensuring a more accurate representation of the data.
2. **Data Imbalance Assessment:** Recognized the imbalance in the data, particularly the disproportionate ratio of defaulters to non-defaulters, which helps in model building by addressing potential bias.
3. **Univariate and Segmented Analysis:** Gained insights into how customer attributes (e.g., gender, age, family status) and financial factors (e.g., income, credit amount) influence loan repayment.
4. **Bivariate Analysis:** Explored relationships between variables such as income, credit amount, and loan annuity, providing deeper insights into loan approval criteria.
5. **Correlation Matrix:** Created correlation matrices for both defaulters and non-defaulters, identifying strong relationships between variables like social circle count, loan amount, and region ratings.

This project enhanced my understanding of the complexities involved in loan approval processes, particularly the factors that predict defaults. By identifying key patterns and correlations, I am now better equipped to suggest improvements in the loan approval process, such as adjusting criteria based on customer attributes and financial data. Additionally, the analysis of data imbalance and outliers will help refine risk models for more accurate decision-making in the finance industry.

Link of Excel file:

Click