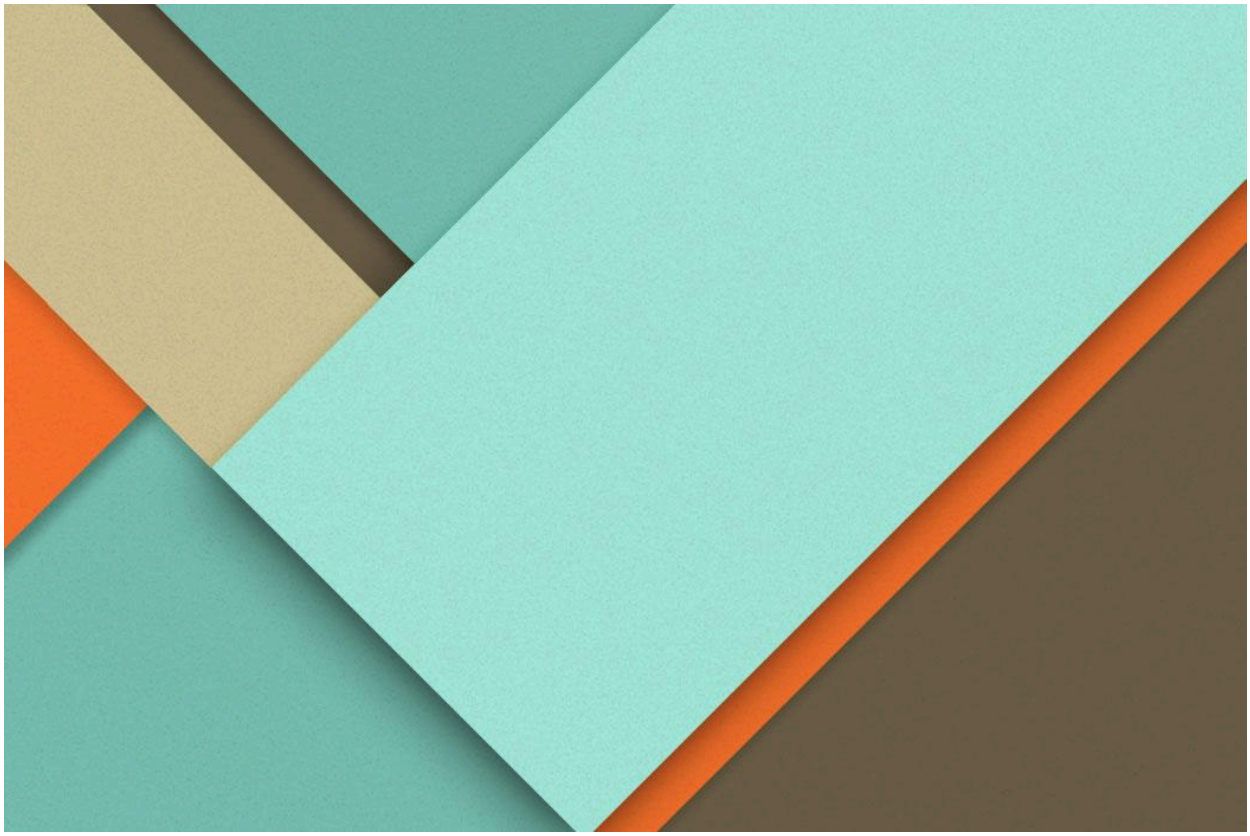


Bank Loan Case Study

Final Project-2

09.12.2024

—

Vinay Rana

Overview

- 1. Evaluate loan application data to identify the reasons behind customer loan defaults.
- 2. The primary goal is to determine the key factors that increase the likelihood of a customer defaulting, enabling more informed loan approval decisions.
- 3. Assist individuals with no prior credit history who may be at risk of defaulting.
- 4. Minimize financial losses by avoiding the rejection of reliable applicants and reducing approvals for those unlikely to repay.

Goals

- **Understand Loan Defaults**: Analyze loan application data to identify the main reasons customers fail to repay loans.
- **Enhance Loan Approval Accuracy**: Develop methods to improve loan approval decisions by identifying customers more likely to default.
- **Assist First-Time Borrowers**: Provide support for customers with no credit history to lower their chances of defaulting.
- **Reduce Financial Risks**: Prevent losses by avoiding the rejection of reliable applicants and minimizing approvals for high-risk borrowers.

The Provided dataset include:

- **Applications_data**: Contains comprehensive details about clients, including whether they have encountered payment difficulties.
- **Previous_data**: Provides information on clients' previous loans, including whether the loan was approved, canceled, refused, or unused.
- **Columns_description**: Describes the various columns present in the datasets, offering clarity on the data structure.
- **Important_notes**: Offers guidance on how to approach and analyze the case study effectively.

Approach

- **Data Acquisition**:
  - Download the datasets: application_data and previous_data.
- **Data Understanding**:
  - Explore the datasets to understand the structure, types of variables, and relationships between them.
- **Data Cleaning**:
  - Remove columns with more than 40% missing values.
  - Replace missing values in other columns using appropriate measures, such as mean or median.
  - Eliminate irrelevant or redundant columns that do not contribute to the analysis.

- **Outlier Detection**:
  - Identify and handle outliers in the datasets using statistical methods or visualization techniques.
- **Data Visualization**:
  - Create meaningful charts and graphs to visualize trends, distributions, and patterns in the data.
- **Insights and Analysis**:
  - Use the cleaned and visualized data to derive insights that inform strategies for reducing loan defaults and improving approval decisions.

Tools and Software

I. Software: Microsoft Excel Version 16.91

Reason:Microsoft Excel Version 16.91 is chosen for its advanced data analysis, manipulation, and visualization capabilities. It provides essential tools for cleaning, processing, and visualizing large datasets, such as pivot tables, formulas, and charts, making it ideal for analyzing loan application data and identifying patterns.

II. Google Docs:

Reason:I used Google Docs to create the presentation and save it as a PDF because it offers a comprehensive set of tools and templates that make project creation efficient and straightforward.

Data Analytics Tasks:

A. Identify Missing Data and Deal with it Appropriately:

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.

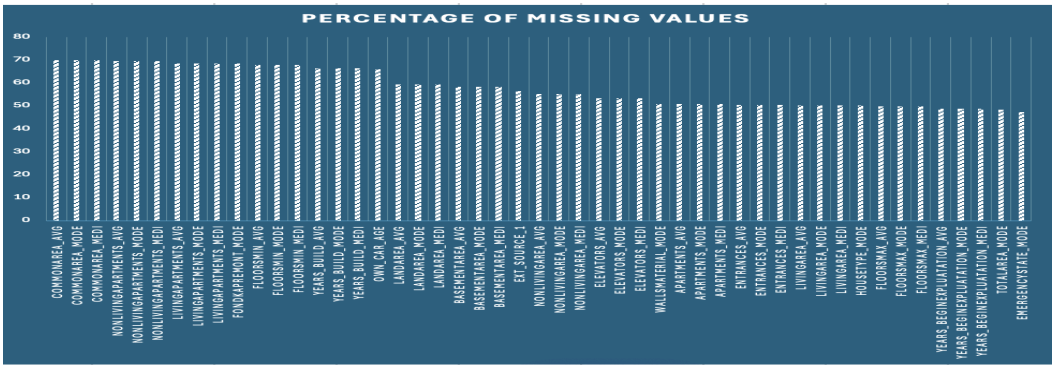
A. Handling Missing Data:

1. **Identifying Missing Values**:

   Calculated the percentage of null values for each column using the formula:
   excel
   Copy code
   `=(COUNTBLANK(range)/COUNTA(range))*100`

2. **Removing Columns with High Null Values**:
   - Eliminated all columns with more than 40% missing data.

3. **Imputing Missing Values in Remaining Columns**:
   - For columns with less than 40% missing data, filled the gaps using appropriate methods:
     - **Median** for numerical data.
     - **Pivot Table** for categorical data.



The dataset contains several columns with high percentages of missing values, ranging from approximately 47% to nearly 70%.

B. Identify Outliers in the Dataset:

Task: Detect and identify outliers in the dataset by utilizing Excel's statistical functions and features, with a focus on numerical variables.

Outliers

Dataset: application_data

Steps to Identify Outliers:

Use Quartiles to Detect Outliers:

Quartile 1 (Q1):

- =QUARTILE.INC(application_data!H2:H50000, 1)

Median (Q2):

- =QUARTILE.INC(application_data!H2:H50000, 2)

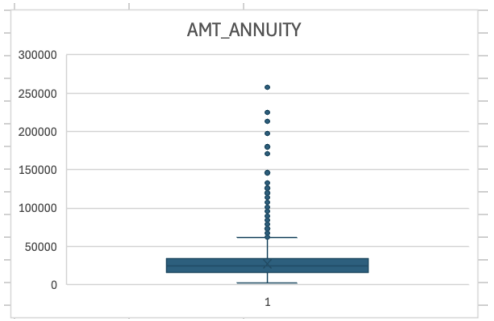
Quartile 3 (Q3):

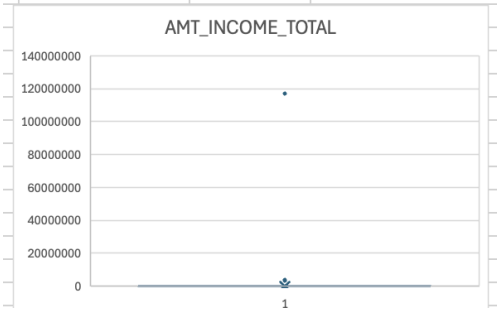
- =QUARTILE.INC(application_data!H2:H50000, 3)

Update the range (H2:H50000) for the desired column.

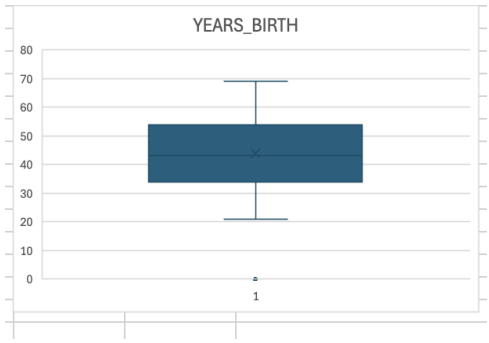
Use Box-and-Whisker Charts:

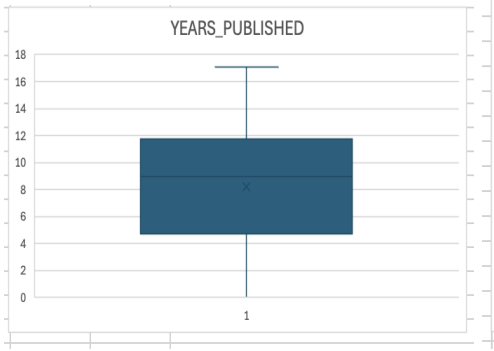
Visualize outliers using this chart to highlight the interquartile range (IQR) and extreme values.

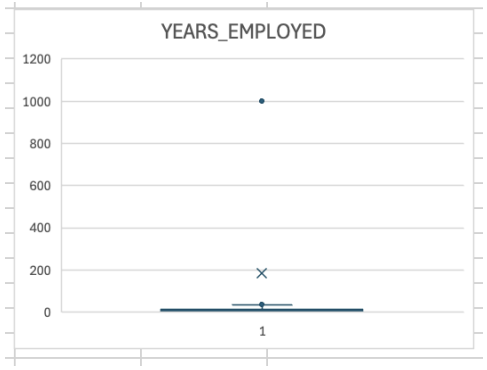

AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
1

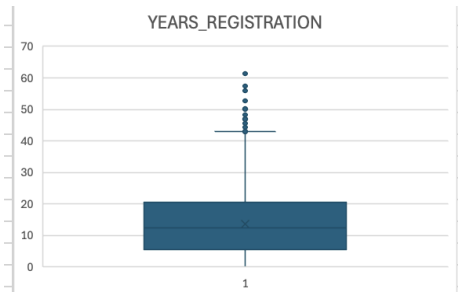

AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1

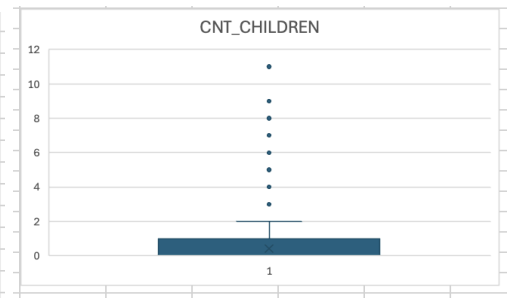

AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

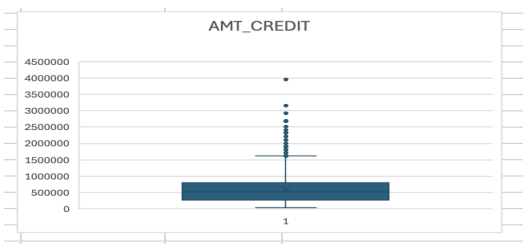

YEARS_BIRTH
80
70
60
50
40
30
20
10
0
1


YEARS_PUBLISHED
18
16
14
12
10
8
6
4
2
0
1


YEARS_EMPLOYED
1200
1000
800
600
400
200
0
1


YEARS_REGISTRATION
70
60
50
40
30
20
10
0
1


CNT_CHILDREN
12
10
8
6
4
2
0
1


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

Output:

During the analysis, I identified several legitimate data points across various columns. Additionally, some columns exhibited no outliers, indicating a relatively uniform data distribution.

However, outliers were occasionally detected and flagged as invalid based on predefined criteria:

Invalid Outliers:

One notable case involved an individual reported to have eleven children, which appeared unrealistic when assessed against contemporary norms.

Unrealistic Salary: An anomaly was identified with an exceptionally high salary of ₹11,70,00,000, which is implausible under most circumstances.

Improbable Work History: Another outlier indicated nearly 1,000 years — an extraordinarily unlikely scenario.

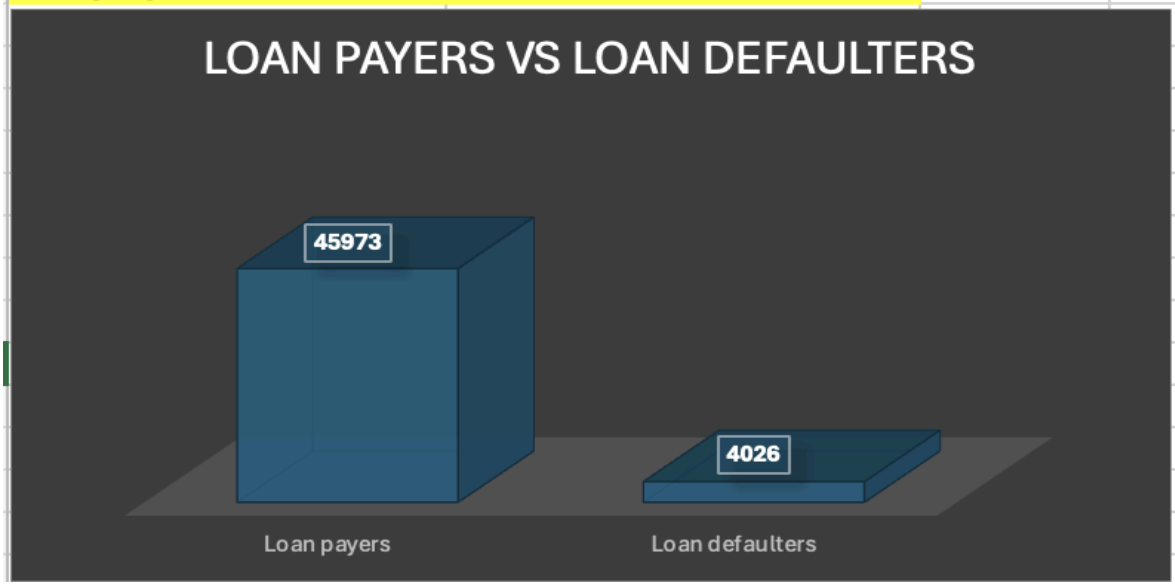
Action Taken:

To ensure accurate and reliable analysis, anomalies were identified as invalid. These entries may need further review and cleanup, and in some cases, removal might be necessary to maintain the quality of the data. Steps were taken to handle these outliers and ensure the analysis is trustworthy.

C. Analyze Data Imbalance:

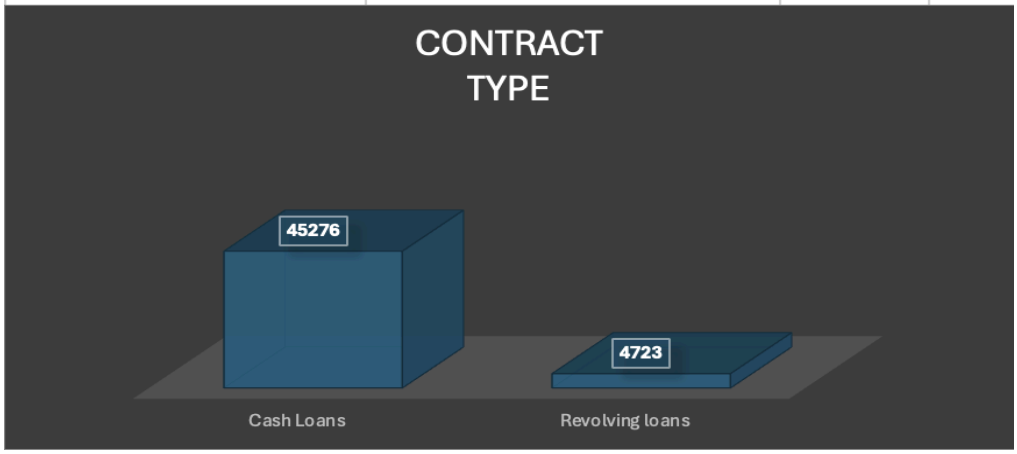
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Loan payers	45973	92
Loan defaulters	4026	8
Total people	49999	



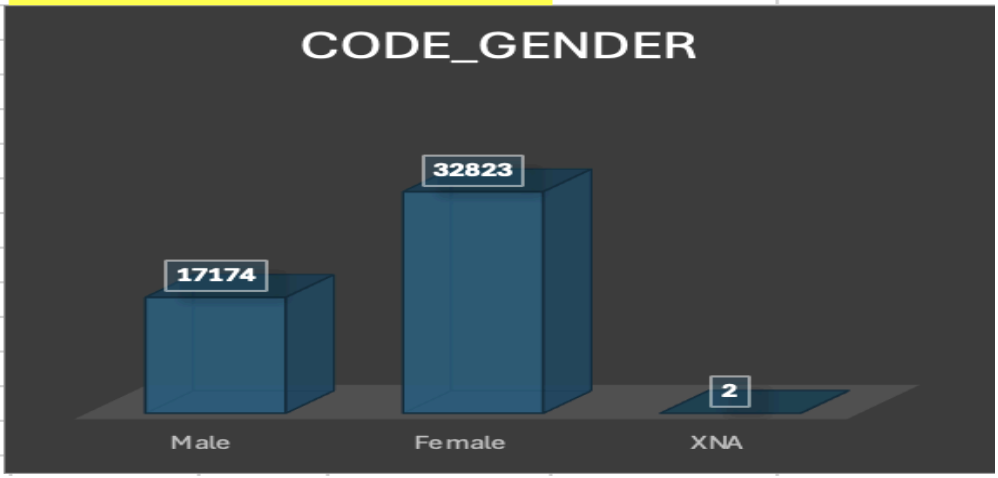
From the bar graph above, it is evident that 92% of individuals repay their loans on time, while only 8% face payment difficulties. This indicates a significant imbalance in the dataset.

Category	Count	Percentage
Cash Loans	45276	91
Revolving loans	4723	9
Total	49999	



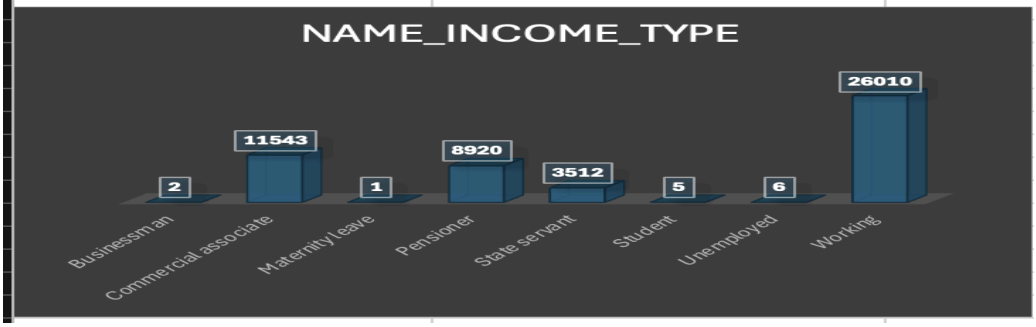
The bar graph shows that 91% of loans are cash loans, while only 9% are revolving loans, indicating a strong preference for cash loans in the dataset.

Category	Count	Percentage
Male	17174	34
Female	32823	66
XNA	2	0.004
Total	49999	



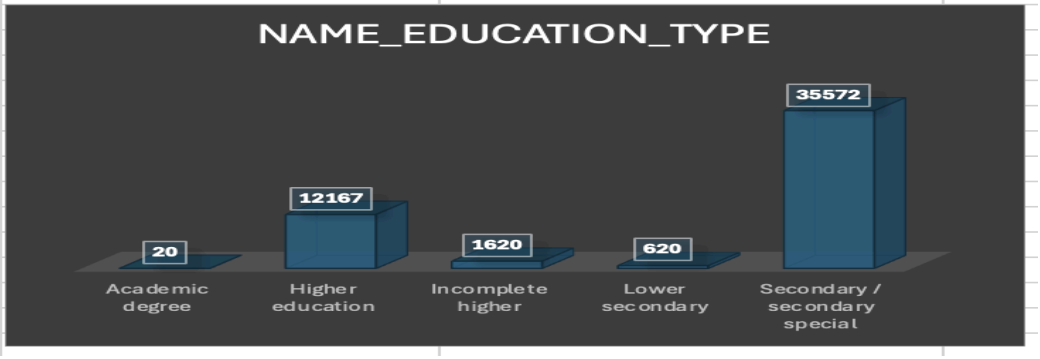
The graph shows that 66% of loan borrowers are female, while 34% are male, indicating a higher participation of women in taking loans compared to men.

NAME_INCOME_TYPE	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010
Grand Total	49999



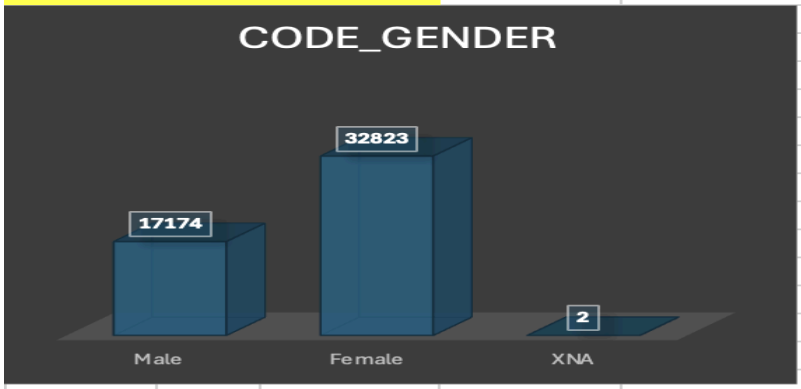
The data indicates that the majority of individuals are working (26,010), followed by commercial associates (11,543) and pensioners (8,920). Smaller groups include state servants (3,512), with very few students, unemployed, and businessmen.

NAME_EDUCATION_TYPE	Count of NAME_EDUCATION_TYPE
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary special	35572
Grand Total	**49999**



The data reveals that most individuals have secondary or secondary special education (35,572), followed by higher education (12,167), with smaller groups in other categories.

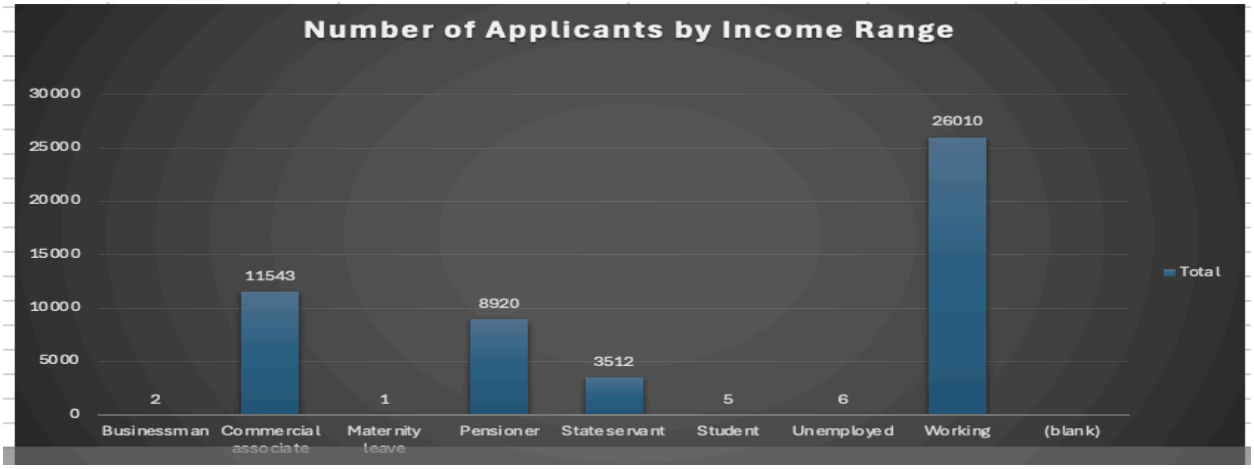
Category	Count	Percentage
Male	17174	34
Female	32823	66
XNA	2	0.004
Total	49999	

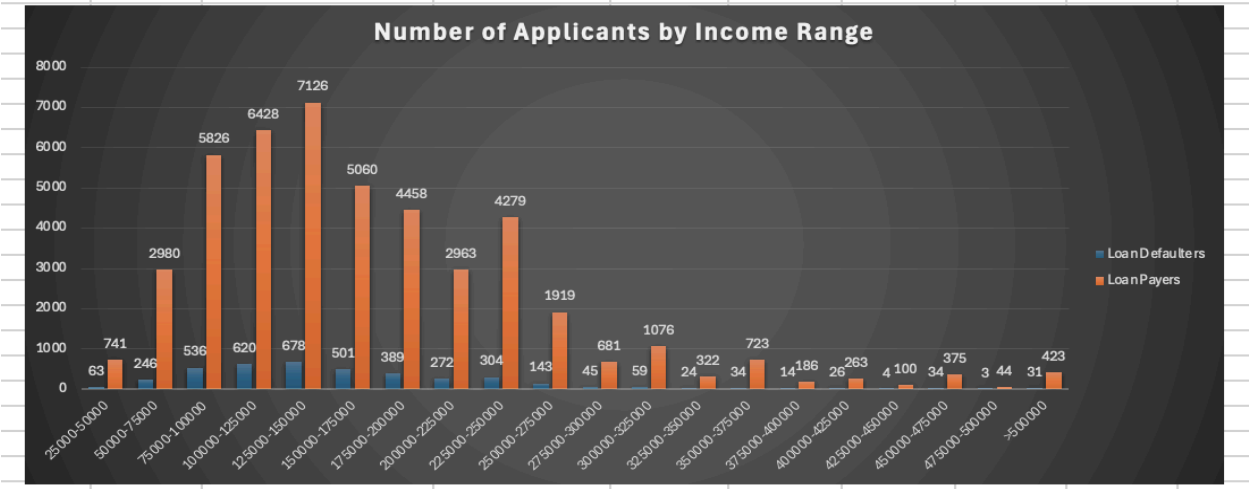


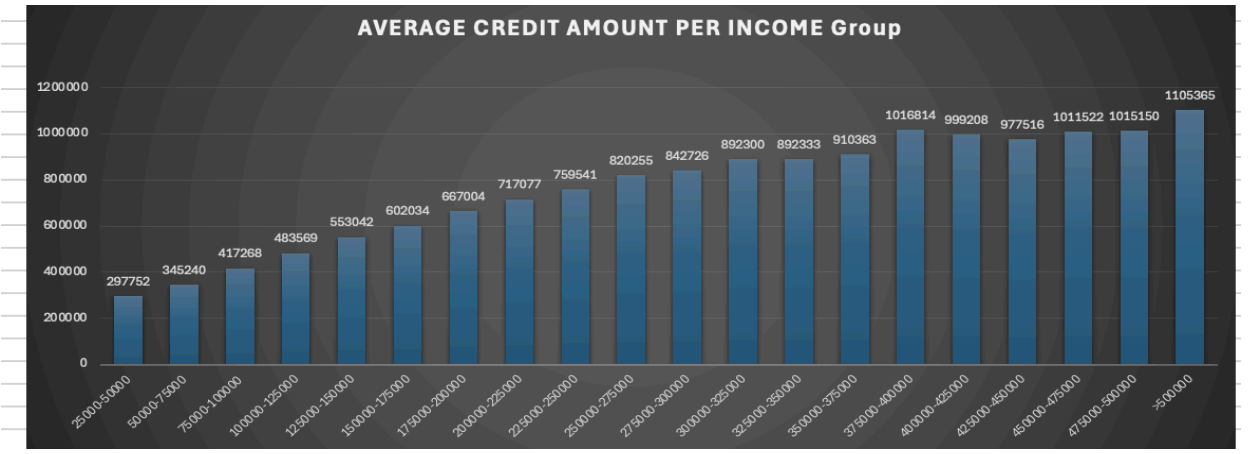
The data shows that the majority of individuals are female (32,823), followed by male (17,174), with only 2 entries categorized as XNA

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

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







The analysis highlights significant insights into income types, loan approval patterns, and credit trends:

Univariate Analysis:

The majority of applicants belong to the "Working" category (26,010), followed by "Commercial Associates" (11,543) and "Pensioners" (8,920). Categories such as "Businessman," "Maternity Leave," and "Student" have minimal representation.

Segmented Univariate Analysis:

Loan defaults are more common among lower income groups, with a noticeable decline as income levels increase.

Most applicants fall within the ₹1,25,000–₹1,50,000 income range, indicating a strong representation of middle-income groups.

Bivariate Analysis:

A positive correlation exists between income levels and average loan credit, with higher-income groups qualifying for larger loan amounts.

Applicants earning above ₹5,00,000 have the highest average credit amount of ₹11,05,365.

These findings emphasize the importance of targeted loan approval strategies, focusing on reducing default risks in lower-income groups while leveraging the potential of higher-income segments.

E. Identify Top Correlations for Different Scenarios:

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.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_PUBLISHED	CNT_FAM_MEMBERS	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.036319722	0.00570546	0.02638212	0.001369783	-0.024912809	-0.33587627	-0.245521512	-0.183072478	0.032537221	0.879238049	0.021288992
AMT_INCOME_TOTAL	0.036	1.000	0.378	0.451	0.384	0.182	-0.074	-0.162	-0.069	-0.032	0.042	-0.205
AMT_CREDIT	0.006	0.378	1.000	0.771	0.987	0.096	0.051	-0.075	-0.008	0.008	0.065	-0.103
AMT_ANNUITY	0.026	0.451	0.771	1.000	0.776	0.117	-0.010	-0.111	-0.035	-0.009	0.078	-0.130
AMT_GOODS_PRICE	0.001	0.384	0.987	0.776	1.000	0.099	0.049	-0.072	-0.011	0.010	0.063	-0.104
REGION_POPULATION_RELATIVE	-0.025	0.182	0.096	0.117	0.099	1.000	0.030	-0.007	0.059	0.002	-0.023	-0.539
YEARS_BIRTH	-0.336	-0.074	0.051	-0.010	0.049	0.030	1.000	0.623	0.335	0.270	-0.284	-0.009
YEARS_EMPLOYED	-0.246	-0.162	-0.075	-0.111	-0.072	-0.007	0.623	1.000	0.209	0.275	-0.235	0.041
YEARS_REGISTRATION	-0.183	-0.069	-0.008	-0.035	-0.011	0.059	0.335	0.209	1.000	0.104	-0.171	-0.083
YEARS_PUBLISHED	0.033	-0.032	0.008	-0.009	0.010	0.002	0.270	0.275	0.104	1.000	0.025	0.008
CNT_FAM_MEMBERS	0.879	0.042	0.065	0.078	0.063	-0.023	-0.284	-0.235	-0.171	0.025	1.000	0.022
REGION_RATING_CLIENT	0.021	-0.205	-0.103	-0.130	-0.104	-0.539	-0.009	0.041	-0.083	0.008	0.022	1.000

The correlation analysis highlights key relationships between variables:

- AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE are highly correlated, reflecting consistent patterns in loan amounts, annuities, and goods pricing.
- YEARS_BIRTH and YEARS_EMPLOYED have a moderate positive correlation (0.623), showing older individuals tend to have longer employment histories.
- Negative correlations, such as between REGION_POPULATION_RELATIVE and REGION_RATING_CLIENT (-0.539), suggest areas with higher population densities may receive lower region ratings.
- These findings provide insight into applicant profiles and relationships among financial and demographic variables, guiding targeted strategies.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_PUBLISHED	CNT_FAM_MEMBERS	REGION_RATING_CLIENT
CNT_CHILDREN	1.000	0.010	0.008	0.029	-0.001	-0.020	-0.250	-0.190	-0.152	0.042	0.893	0.056
AMT_INCOME_TOTAL	0.010	1.000	0.015	0.018	0.013	-0.006	-0.009	-0.012	0.010	0.009	0.013	-0.013
AMT_CREDIT	0.008	0.015	1.000	0.750	0.982	0.068	0.143	0.019	0.043	0.044	0.061	-0.045
AMT_ANNUITY	0.029	0.018	0.750	1.000	0.750	0.073	0.009	-0.078	-0.022	0.021	0.076	-0.062
AMT_GOODS_PRICE	-0.001	0.013	0.982	0.750	1.000	0.077	0.141	0.023	0.043	0.049	0.056	-0.052
REGION_POPULATION_RELATIVE	-0.020	-0.006	0.068	0.073	0.077	1.000	0.016	0.008	0.046	0.005	-0.017	-0.430
YEARS_BIRTH	-0.250	-0.009	0.143	0.009	0.141	0.016	1.000	0.588	0.288	0.248	-0.199	-0.045
YEARS_EMPLOYED	-0.190	-0.012	0.019	-0.078	0.023	0.008	0.588	1.000	0.192	0.233	-0.183	-0.009
YEARS_REGISTRATION	-0.152	0.010	0.043	-0.022	0.043	0.046	0.288	0.192	1.000	0.090	-0.152	-0.116
YEARS_PUBLISHED	0.042	0.009	0.044	0.021	0.049	0.005	0.248	0.233	0.090	1.000	0.044	-0.025
CNT_FAM_MEMBERS	0.893	0.013	0.061	0.076	0.056	-0.017	-0.199	-0.183	-0.152	0.044	1.000	0.057
REGION_RATING_CLIENT	0.056	-0.013	-0.045	-0.062	-0.052	-0.430	-0.045	-0.009	-0.116	-0.025	0.057	1.000

The correlation analysis reveals the following key insights:

- CNT_CHILDREN and CNT_FAM_MEMBERS exhibit a strong positive correlation (0.893), highlighting a direct relationship between family size and the number of children.
- AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE are closely related, with high correlations indicating consistency in loan credit, annuity payments, and goods prices.
- YEARS_BIRTH and YEARS_EMPLOYED show a moderate positive correlation (0.588), suggesting older individuals typically have longer employment durations.
- Negative correlations, such as between REGION_POPULATION_RELATIVE and REGION_RATING_CLIENT (-0.430), suggest denser populations may correspond to lower client ratings for regions.

These correlations provide valuable insights into applicant demographics and financial behaviors, supporting informed decision-making processes.

Findings:

- Demographics: Most applicants are "Working" professionals, with women and secondary education holders making up the majority.
- Loan Patterns: Defaults are higher in lower-income groups; higher-income individuals qualify for larger loans.
- Outliers: Identified unrealistic data points, such as extreme salaries and implausible work histories, which were removed for accuracy.
- Data Imbalance: A significant imbalance in loan repayments (92% repay on time, 8% default), and cash loans are preferred.
- Correlations: Strong links between family size and number of children, as well as between loan-related metrics. Older individuals tend to have longer employment histories.

What I Learned:

- Data Preparation: Importance of handling missing values and outliers to ensure data reliability.
- Patterns and Insights: Recognizing income-based trends and their impact on loan defaults.
- Effective Tool Use: Utilized Excel tools like pivot tables and quartile analysis to uncover insights.
- Correlations: Understanding relationships between variables for better decision-making.

Conclusion:

This project aim ed to analyze loan application data to identify factors influencing loan defaults, enhance approval accuracy, and minimize financial risks. Key insights include:

- Data Imbalance: The dataset showed a high imbalance, with 92% of applicants repaying loans on time, while only 8% defaulting. This skew impacts the analysis and requires careful handling of the minority class.

- Outliers: Several unrealistic data points, such as an individual with eleven children and an unusually high salary, were identified and excluded to maintain data integrity.

- Income and Loan Defaults: Income levels were found to correlate with loan defaults, with lower-income groups more likely to default. Higher-income applicants typically qualified for larger loan amounts.

- Demographic Insights: Women represented a larger portion of loan applicants, while most borrowers had secondary or special education. The majority were employed, and family size showed a strong relationship with the number of children.

- This analysis provides valuable insights for improving loan approval processes and targeting high-risk applicants, ultimately helping to optimize loan decision-making and reduce financial loss.