Project 6:

Loan Application Data Analysis

Project Description

In this project, we analyzed loan applications to identify factors influencing default risks among urban customers of a finance company. The goal was to assist in informed loan approval decisions to minimize financial losses while maintaining business growth by approving eligible applicants. A clean and accurate dataset was pivotal for making sound decisions based on meaningful insights.

Tech-Stack Used

MS Excel: Data cleaning, graphs, tables, dashboards.

MS Word: Documentation of tasks and results.

ChatGPT: Helped in summarizing insights clearly for stakeholders.

Tasks

1. Identify Missing Data and Handle It

Objective:

Ensure data accuracy by detecting and treating missing values appropriately.

Steps Taken:

- Analyzed application.csv for missing values.
- Added rows showing Count Values, Blank Values, and Blank Percentage.
- Columns with more than 62% missing data were removed (91 columns deleted)
- For important columns like NAME_TYPE_SUITE and OCCUPATION_TYPE, missing values were filled using Mode ("Unaccompanied") and "Unknown" respectively.

Row names	Formulas
Count values	=COUNTA(B4:B50003)
Count Blank Values	=COUNTBLANK(B4:B50003)
Blank Percentage	-(B2/B1)

Insight:

1. A total of 91 columns were deleted, some of which had missing value percentages of 62% or higher, while others were not relevant to the tasks at hand. The remaining columns after cleaning total 31 out of the original 122.
2. After handling missing values, we ensured that no major feature loss occurred, thus preserving the integrity of core information required for risk modeling.

2. Identify Outliers in the Dataset

Objective:

Detect and handle outliers to ensure reliable analysis.

Steps Taken:

1. Outlier Analysis Preparation: I created individual worksheets for each column containing numerical values and compared these attributes with the target column.

2. Scatter Plot Creation: For each numerical attribute, I plotted scatter graphs to visually identify outliers in relation to the target column. Key attributes like CNT CHILDREN, AMT INCOME TOTAL, and YEARS EMPLOYED exhibited noticeable outliers out of the eight attributes analyzed.

3. Outliers Dashboard: I compiled all the scatter plot graphs into a dedicated Outliers Dashboard for better visualization and analysis.
4. Dashboard Configuration: Since the target column (on the x-axis) was common across all graphs, I opted not to include slicers in the dashboard for simplicity.

Insight:

1. CNT CHILDREN: There are two notable outliers in the dataset—one with 9 children and another with 1 1 children. While such counts are rare, they may represent large families, which are uncommon but possible.
2. AMT INCOME: An outlier with an annual income of u was identified. Though exceptionally high, it might reflect data for a high-net-worth individual or a rare case, such as a business owner or corporate executive.
3. YEARS EMPLOYED: Two extreme outliers show employment durations of 1001 years. These are clearly impossible and likely result from data entry errors or incorrect calculations.

3. Analyze Data Imbalance

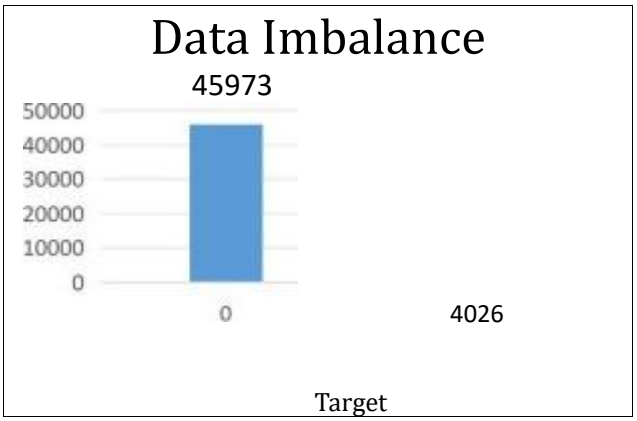
Objective:

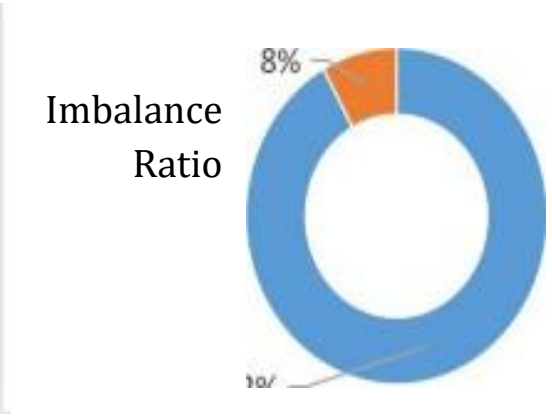
Understand imbalance in loan approval (default vs non-default cases).

Steps:

- Created a pivot table and visualized Target column distribution.

- Built a bar chart and donut chart showing 92% non-defaulters and 8% defaulters.





Insights:

1. Bar Graph (Data Imbalance):

    1. The bar chart highlights a significant imbalance in the Target column.
    2. Non-defaulters (Target = O) make up the majority of the dataset with 45,973 entries.
    3. Defaulters (Target = 1) are comparatively rare, with only 4,026 entries.
    4. This indicates that only 8% of the observations represent defaulters, while 92% are non-defaulters, showing a substantial imbalance in the data distribution.

2. Donut Chart (Imbalance Ratio):

    1. The donut chart further emphasizes the data imbalance by visually representing the ratio.
    2. Non-defaulters (92%) dominate the dataset, while defaulters constitute a small proportion (8%).
    3. This imbalance could lead to biased model predictions if not addressed, as the model may favor the majority class (non-defaulters) over the minority class (defaulters).

4. Perform Univariate, Segmented Univariate, and Bivariate Analysis

Objective:

Gain insights into customer behavior and risk patterns.

Steps :

1. For Income:

a. When creating income bins for analysis, I noticed an outlier in the dashboard: an applicant with an annual income of ₹11,70,00,000. While such an income is possible, it significantly skewed the analysis. To address this, I replaced the outlier value with 0, ensuring that it wouldn't distort the binning process or the overall analysis.

b. After removing the outlier, I observed that the income range was still highly variable, making it challenging to identify meaningful income groups. To resolve this, I tested multiple interval distributions:

c. First Attempt: I grouped incomes using an interval of ₹3,79,935, which resulted in 10 bins. However, this distribution was not effective, as the first income group alone contained 45,000 applicants, making the grouping unbalanced. ● Final Solution: I refined the grouping by reducing the interval to ₹1,08,55 which resulted in 35 bins. This approach provided much better clarity, as it balanced the distribution and helped identify the income groups with the highest number of applicants.

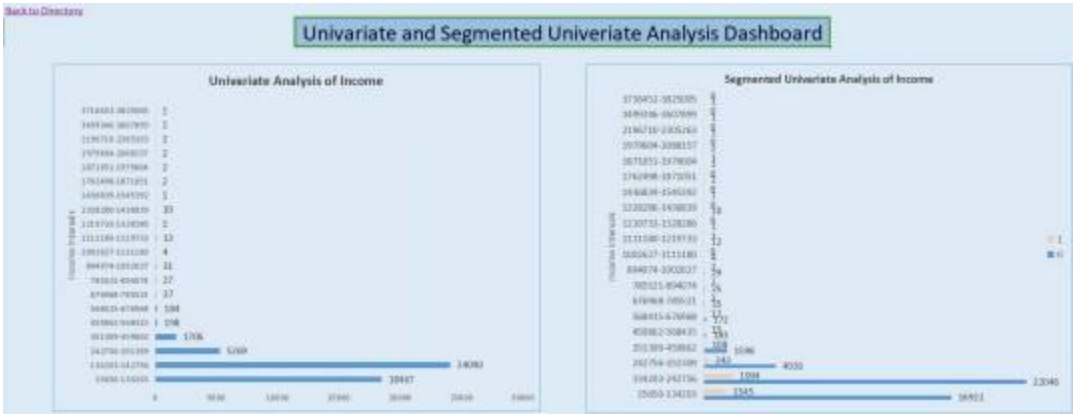
d. This iterative approach allowed me to achieve a more accurate and meaningful analysis of income groups.

2. For Other Factors:

a. Creating bins for the other factors was straightforward. For univariate analysis, appropriate groups were formed for each factor, and the data was visualized using a clustered bar chart.

b. To perform Segmented Univariate Analysis, each factor was analyzed in combination with the Target variable, providing insights into how different categories of each factor relate to the target groups (e.g., defaulters vs. nondefaulters).

1. Univariate and Segmented Univariate Analysis Dashboard of Income



Univariate Analysis of Income (Left Chart):

a. The income distribution shows a clear imbalance among applicants.
b. The majority of applicants fall into the lowest income group (kl ₹2,56,6with 24,040 applicants, followed by the next income group, which has 19,667 applicants.
c. Higher income groups have significantly fewer applicants, indicating that most applicants belong to lower income brackets.

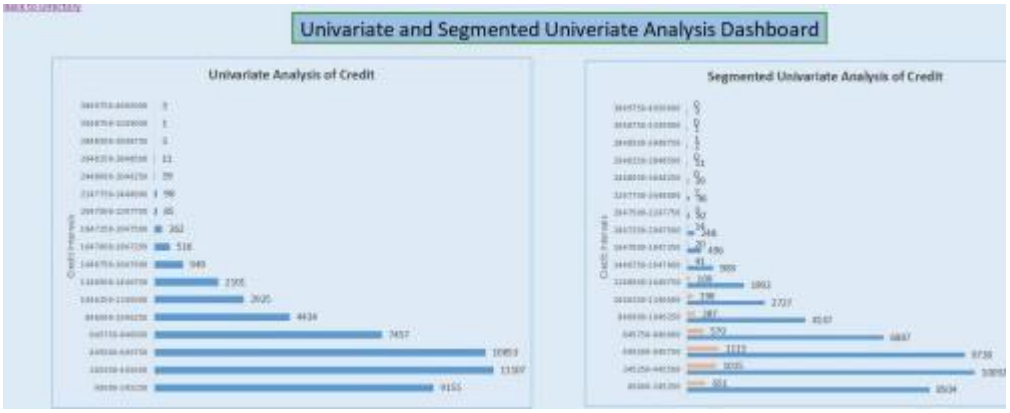
Segmented Univariate Analysis of Income (Right Chart):

a. When segmented by Target (O: Non-defaulters, 1: Defaulters):
b. Non-defaulters (Target = O) dominate across all income groups, as expected due to the overall data imbalance.
c. Defaulters (Target = 1) are primarily concentrated in the lower income groups, with 4,026 defaulters in the first income range ₹1 - ₹2,56,630),decreasing significantly in higher income groups.
d. This pattern suggests that applicants in the lower income brackets are more likely to default on loans compared to those in higher income brackets.

Insight:

The analysis indicates a strong correlation between lower income groups and higher default rates. This insight can guide risk assessment strategies, such as introducing stricter credit checks or customized loan terms for applicants in these income brackets.

2.Univariate and Segmented Univariate Analysis Dashboard of Credit



Univariate Analysis of Credit (Left Chart):

a. Credit Distribution: The credit distribution shows a clear imbalance among applicants.
b. Lowest Credit Group: The majority of applicants fall into the lowest credit group - with 11,107 applicants, followed by the next credit group - with 10,853 applicants and then with 9,155 applicants.

2. Higher Credit Groups: Higher credit groups have significantly fewer applicants, indicating that most applicants belong to lower credit brackets.

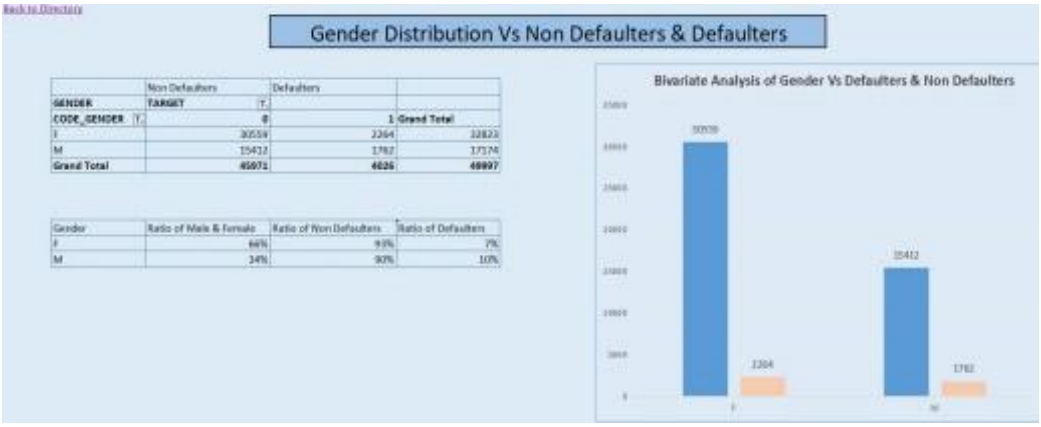
Segmented Univariate Analysis of Credit (Right Chart):

a. When segmented by Target (O: Non-defaulters, 1: Defaulters):
b. Non-defaulters (Target = O): Non-defaulters dominate across all credit groups, as expected due to the overall data imbalance.
c. Defaulters (Target = 1): Defaulters are primarily concentrated in the lower credit groups, with overall 3,966 defaulters in the credit range ₹16,47,0 decreasing significantly in higher credit groups.

Insight:

- Prioritize Risk Management for Lower Credit Groups:
- Since defaulters are predominantly concentrated in the lower credit brackets (₹45,000 - ₹16,47,000), focus on strengthening risk assessment processes for applicants in these ranges. This could involve tighter credit checks, more frequent monitoring, and adjusting lending criteria to minimize defaults.
- Tailored Lending Products for Non-Defaulters in Higher Credit Groups:
- With higher credit groups showing fewer defaulters, consider creating premium loan products that appeal to these applicants, offering competitive interest rates or rewards. This strategy can help attract more applicants to higher credit tiers while maintaining low default risks.

2. Bivariate Analysis Dashboard of Gender



Bivariate Analysis of Gender VS Defaulters and Non Defaulters:

1. Higher Proportion of Non-Defaulters Among Females:
   66% of the applicants are female, and among them, 93% are non-defaulters. This suggests that female applicants are more likely to repay loans and are a lower-risk demographic for lenders.

2. Higher Default Rate Among Males:
   Although males make up only 34% of the total applicants, they account for 10% of the defaulters, which indicates a higher default rate compared to females (7%). This suggests that males are relatively riskier borrowers.

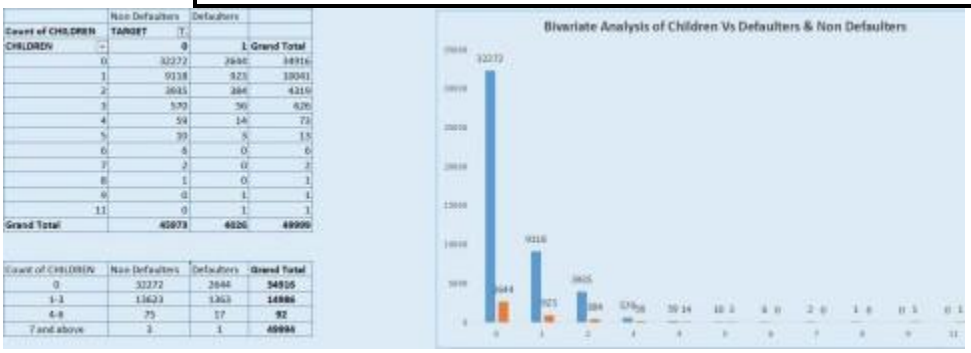
Insight:

- Targeted Risk Mitigation for Male Borrowers:
  With a higher default rate among male borrowers, introduce stricter lending criteria or personalized loan management strategies for male applicants, such as credit education programs or tailored repayment plans, to reduce defaults.

- Leverage the Low Risk of Female Borrowers:
  Given that females have a higher proportion of non-defaulters, consider offering attractive financial products or incentives (e.g., lower interest rates or loyalty rewards) to female applicants to expand your customer base while maintaining low risk.

4. Bivariate Analysis Dashboard of Children

Count Of Childred vs Non Defaulters & Defaulters



Bivariate Analysis of Children VS Defaulters and Non Defaulters:

- Higher Default Rate for Applicants with Fewer Children (0-3):
  Applicants with no children (O children) make up the majority of non-defaulters (32,272 out of 34,916), but also a significant portion of defaulters (2,644). Additionally, those with 1-3 children show a notable level of default risk (1,363 defaulters out of 13,623). This suggests that applicants with fewer children are more likely to default, which might indicate financial stress or other socioeconomic factors.

- Lower Default Rate for Applicants with More Children (4+):
  Applicants with 4 or more children (especially those with 4-6 children) have much lower numbers of defaulters (17 out of 92 total). This suggests that applicants with larger families may be more cautious or have more stable financial situations, potentially due to higher financial responsibility or assistance.

Insight:

- Focus Risk Management on Applicants with Fewer Children (0-3):
  Given that applicants with fewer children (especially those with none or 1-3 children) are more likely to default, it may be beneficial to adjust risk management strategies for this group. This could include stricter credit assessments, tailored financial advice, or offering more flexible repayment plans to help prevent defaults.

- Target Applicants with Larger Families for Stability-Oriented Products:
  Since applicants with 4 or more children show a lower default rate, consider offering products specifically designed for families with more children, such as familyoriented loan packages, lower interest rates, or long-term financial planning tools that align with their stability and needs.

5. Identify Top Correlations

- Filtered the data in the 'Target' column to include only rows where the value is O (non-defaulters).
- Selected all columns with numeric values in the filtered data and copied them.
- Paste the copied data into a new worksheet for analysis.
- In another new worksheet, created a table where both the column names and rows represent the numeric columns I selected.
- For each pair of columns (one as the row and the other as the column), added the correlation function =CORREL('Target O' $B:$B,'Target 0'!B:B) where the first and second factors were chosen columns.
- Repeated this process for all pairs of numeric columns to calculate the correlation for each pair.
- This allowed me to analyze the correlations between the different numeric columns in my dataset.

Correlation for Non Defaulters

	CNT_CHILDREN	AMT_INCOME_TC	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PR	Age	YEARS_EMPLOYE	YEARS_REGISTRATIO	YEARS_ID_PUBLISH
CNT_CHILDREN	-CORREL(Target	0.036319722	0.005705458	0.02638217	0.001550025	-0.335876268	-0.245521512	-0.183072478	0.032537221
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.451135696	0.384675092	-0.073769425	-0.161680938	-0.06893375	-0.032286356
AMT_CREDIT	0.005705458	0.377965752	1	0.770772965	0.987244066	0.051084182	-0.074733443	-0.008053758	0.008290189
AMT_ANNUITY	0.02638217	0.451135696	0.770772965	1	0.776141898	-0.009915685	-0.111294243	-0.034609089	-0.009426496
AMT_GOODS_PRICE	0.001550025	0.384675092	0.987244066	0.776141898	1	0.048700977	-0.072505216	-0.011290011	0.009304005
Age	-0.335876268	-0.073769425	0.051084182	-0.009915685	0.048700977	1	0.623474675	0.335028046	0.270073313
YEARS_EMPLOYED	-0.245521512	-0.161680938	-0.074733443	-0.111294243	-0.072505216	0.623474675	1	0.208846476	0.274516224
YEARS_REGISTRATION	-0.183072478	-0.06893375	-0.008053758	-0.034609089	-0.011290011	0.335028046	0.208846476	1	0.103548902
YEARS_ID_PUBLISH	0.032537221	-0.032286356	0.008290189	-0.009426496	0.009304005	0.270073313	0.274516224	0.103548902	1

	CNT_CHILDREN	AMT_INCOME_TC	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRI	Age	YEARS_EMPLOYE	YEARS_REGISTRATIO	YEARS_ID_PUBLISH
CNT_CHILDREN	1	-0.003115055	0.007601905	0.029172977	-0.001116682	-0.2496732	-0.189773227	-0.152113117	0.042360717
AMT_INCOME_TOTAL	-0.003115055	1	0.306757363	0.377333333	0.306696699	-0.009740795	-0.116816139	-0.029760823	-0.000559639
AMT_CREDIT	0.007601905	0.306757363	1	0.749665201	0.982432318	0.142506035	0.018782223	0.042844404	0.043771901
AMT_ANNUITY	0.029172977	0.377333333	0.749665201	1	0.749705184	0.008751713	-0.078113894	-0.021581654	0.02132109
AMT_GOODS_PRICE	-0.001116682	0.306696699	0.982432318	0.749705184	1	0.140996151	0.023159154	0.043371319	0.049784603
Age	-0.2496732	-0.009740795	0.142506035	0.008751713	0.140996151	1	0.588242824	0.288437837	0.247896571
YEARS_EMPLOYED	-0.189773227	-0.116816139	0.018782223	-0.078113894	0.023159154	0.588242824	1	0.19243569	0.232661912
YEARS_REGISTRATION	-0.152113117	-0.029760823	0.042844404	-0.021581654	0.043371319	0.288437837	0.19243569	1	0.09029149
YEARS_ID_PUBLISH	0.042360717	-0.000559639	0.043771901	0.02132109	0.049784603	0.247896571	0.232661912	0.09029149	1

Directory

Top 10 Correlation of Non Defaulters and Defaulters

Non Defaulters

Rank	Variable 1	Variable 2	Correlation
1	AMT_CREDIT	AMT_GOODS_PRICE	0.987244066
2	AMT_ANNUITY	AMT_GOODS_PRICE	0.776141898
3	AMT_CREDIT	AMT_ANNUITY	0.770772965
4	Age	YEARS_EMPLOYED	0.623474675
5	AMT_INCOME_TOTAL	AMT_ANNUITY	0.451135696
6	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.384675092
7	AMT_INCOME_TOTAL	AMT_CREDIT	0.377965752
8	Age	YEARS_REGISTRATION	0.335028046
9	YEARS_EMPLOYED	YEARS_ID_PUBLISH	0.274516224
10	Age	YEARS_ID_PUBLISH	0.270073313

Defaulters

Rank	Variable 1	Variable 2	Correlation
1	AMT_CREDIT	AMT_GOODS_PRICE	0.982432318
2	AMT_ANNUITY	AMT_GOODS_PRICE	0.749705184
3	AMT_CREDIT	AMT_ANNUITY	0.749665201
4	Age	YEARS_EMPLOYED	0.588242824
5	AMT_INCOME_TOTAL	AMT_ANNUITY	0.377333333
6	AMT_INCOME_TOTAL	AMT_CREDIT	0.306757363
7	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.306696699
8	Age	YEARS_REGISTRATION	0.288437837
9	Age	YEARS_ID_PUBLISH	0.247896571
10	YEARS_EMPLOYED	YEARS_ID_PUBLISH	0.232661912

Insights for Non Defaulters:

- Target Larger Credit Amounts Based on Goods Price and Annuity:
  Given the strong correlations between credit amount, goods price, and annuity, consider refining lending policies to offer more personalized loan amounts based on the value of the goods being purchased and the applicant's ability to make regular payments.

- Use Age and Employment History for Credit Risk Assessment:

  The correlation between Age and Years Employed suggests that older applicants with more years of employment could be lower-risk borrowers. Financial institutions could leverage this information when assessing creditworthiness, especially for applicants in middle to late career stages.

Insights for Defaulters:

- Careful Assessment of Loan Affordability for High Credit Amounts:

The strong correlation between AMT CREDIT and AMT GOODS PRICE suggests that loans for higher-priced goods tend to have a higher default rate. Lenders should assess the borrower's ability to repay in relation to the goods' price and consider stricter affordability checks, especially for applicants requesting large loans for expensive goods.

- Focus on Income vs. Loan Amount Balance:

The weaker correlation between AMT INCOME TOTAL and the credit or goods price for defaulters highlights that income alone may not be sufficient for predicting default risk. Lenders should develop more nuanced risk models that not only account for income but also evaluate how well income aligns with the requested credit amounts and repayment terms, particularly for those with lower incomes or higher credit demands.

Skills gained through project

- Data Cleaning and Pre-processing: Developed expertise in identifying and handling missing or inconsistent data to ensure accuracy in analysis.
- Data Analysis Using Pivot Tables: Gained proficiency in creating and interpreting pivot tables to uncover key trends and insights, such as filling missing values based on mode or mean.
- Data Visualization: Improved skills in creating visual representations of data (graphs, charts) to communicate findings clearly and effectively.
- Problem-Solving: Learned how to approach and resolve common data issues, such as missing values or inconsistent formats.
- New Concepts Learned: Gained knowledge of new concepts, including outliers, univariate analysis, segmented univariate analysis, bivariate analysis and data
- imbalance analysis. Additionally, developed proficiency in creating correlation tables using the CORREL function.
- MS Excel Proficiency: Strengthened abilities in utilizing Excel for advanced data manipulation, graph creation, and table generation.

Approach

- **Define the Objective**: Understand the problem, set goals for analysing loan defaults and improving approval decisions.
- **Data Collection**: Gather the relevant dataset and ensure it’s in a usable format.
- **Exploratory Data Analysis (EDA)**: Visualize and analyse the data to identify patterns, correlations, and trends.
- **Data Cleaning**: Handle missing values, remove irrelevant columns, and address outliers.
- **Modelling and Analysis**: Apply statistical or machine learning models to identify key factors influencing loan defaults.
- **Interpret and Present Results**: Summarize insights, generate actionable recommendations, and present findings to stakeholders.

Results

- Successfully cleaned a highly imbalanced and noisy dataset by removing redundant or incomplete columns.

- Identified critical factors influencing loan defaults, such as low income, occupation type, gender, and age.

- Addressed extreme outliers to enhance data quality and ensure fair model predictions.

- Developed visual dashboards and pivot tables summarizing data patterns for stakeholder communication.

- Gained key insights that will allow more efficient credit risk management and model training in future predictive analytics projects.