Bank Loan Case Study

Project Description:

The project involves analyzing a loan application dataset to understand patterns influencing loan default, aiming to aid decision-making on loan approvals. The objective is to identify key factors behind loan default to make better decisions about loan approval, denial, or adjustment of terms.

Approach:

Utilizing Microsoft Excel 2022 for data analysis, I followed a structured approach to address the tasks outlined, leveraging Excel's functions and features for data cleaning, outlier detection, imbalance analysis, and various exploratory analyses.

Tech-Stack Used:

Software: Microsoft Excel 2021

Purpose: Excel was chosen for its robust functionalities in data manipulation, analysis, and visualization, making it suitable for this EDA task.

Insights:

1. Missing Data Handling:

- Identified missing values using functions like COUNT, ISBLANK, and IF.
- Employed appropriate imputation techniques (e.g., median imputation) for missing numerical values.
- Visualized missing data proportions using a bar/column chart.

2. Outlier Detection:

- Detected outliers using Excel's statistical functions (QUARTILE, IQR) and conditional formatting.
- Utilized box plots or scatter plots for visualizing and identifying outliers in numerical variables.

3. Data Imbalance Analysis:

- Determined class frequencies using COUNTIF and SUM functions to assess data imbalance.
- Visualized class distributions via pie or bar charts, highlighting any imbalance.

4. Univariate, Segmented Univariate, and Bivariate Analysis:

- Conducted univariate analysis to understand variable distributions.
- Utilized filters, sorting, and pivot tables for segmented and bivariate analysis.
- Visualized distributions and relationships using histograms, bar charts, box plots, and scatter plots.

5. Correlation Analysis:

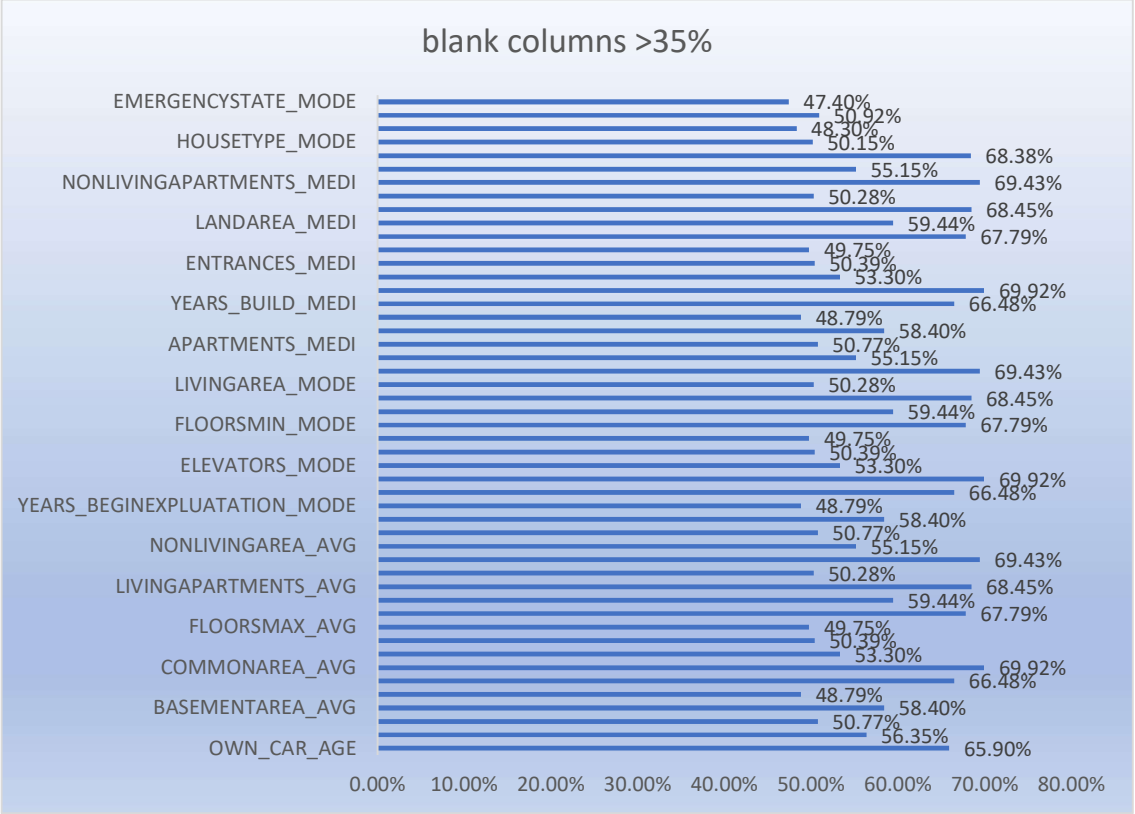
- Segmented dataset based on scenarios to identify top correlations for each segment.
- Calculated correlation coefficients using Excel's CORREL function.
- Visualized correlations with matrices or heatmaps to showcase strong indicators of loan default.

Data Analytics Tasks:

A. Identify Missing Data and Deal with it Appropriately:

1. Identify Missing Data:
    - In cell A2, enter the header "Missing Percentage."
    - In cell A3, use the formula `=(COUNTBLANK(B:B)/COUNT(B:B))*100` and drag it right for other columns.
    - use conditional formatting to highlight cells with a percentage greater than 35 of missing values.
2. Handling Missing Data:
    - If you choose to remove rows with missing values:
    - In a new column, use the `FILTER` or `IF` function to exclude rows with missing values.
    - If you choose to impute missing values:
    - For numerical columns, use functions like `AVERAGE` or `MEDIAN` to fill in missing values.
3. Visualization:
    - Create a bar chart:
    - Select the range B2:DS3.
    - Go to the 'Insert' tab and choose 'Bar Chart'.

This visualization will provide a quick overview of the proportion of missing values for each variable, helping you make informed decisions on handling missing data.



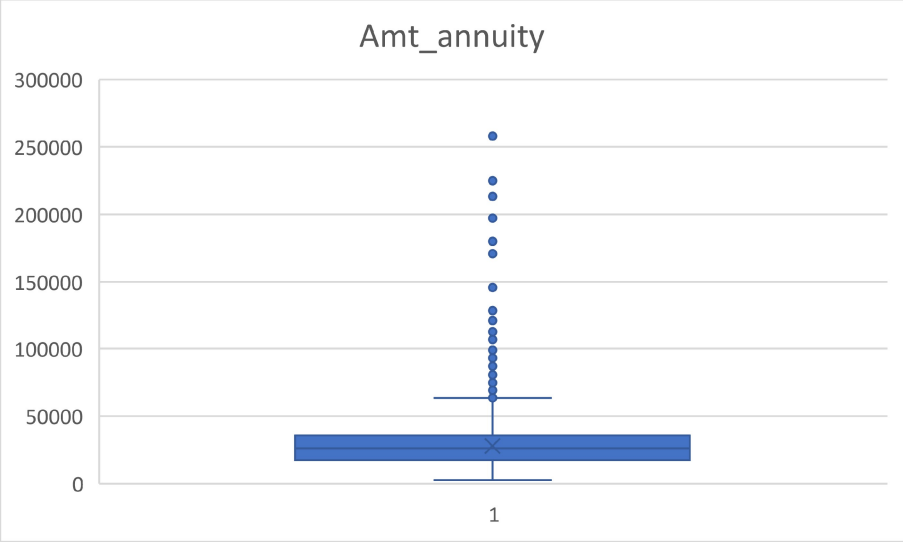
B. Identify Outliers in the Dataset:

1. Detecting Outliers:
   - Use the `QUARTILE` function to calculate the first (Q1) and third (Q3) quartiles for each numerical variable.
   - Calculate the Interquartile Range (IQR) using the formula `IQR = Q3 - Q1`.
   - Identify potential outliers as values beyond the range
     Lower Bound= [Q1 - 1.5 * IQR]
     Higher Bound= [Q3 + 1.5 * IQR]
   - Use conditional formatting to highlight cells containing potential outliers.

QUARTILE 1	17348.63
QUARTILE 2(median)	26145
QUARTILE 3	35814.38
Interquartile Range (IQR):	18465.75
Lower Bound:	-10350
Upper Bound:	35814.38

2. Validating Outliers:
   - Consider applying business rules or thresholds to determine if the identified outliers are valid or require further investigation.
   - You may choose to exclude extreme values if they are not consistent with the expected data range.

3. Visualization:
    - Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.
    - Box plots provide a visual representation of the quartiles and help identify outliers.
    - Scatter plots allow you to visually inspect individual data points and their positions relative to the expected range.

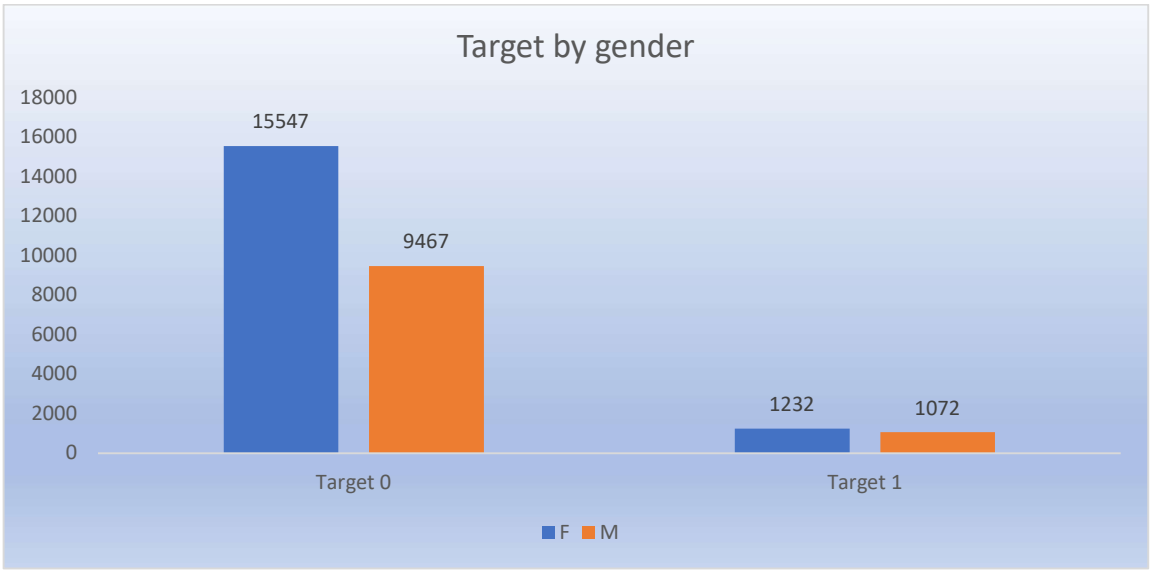


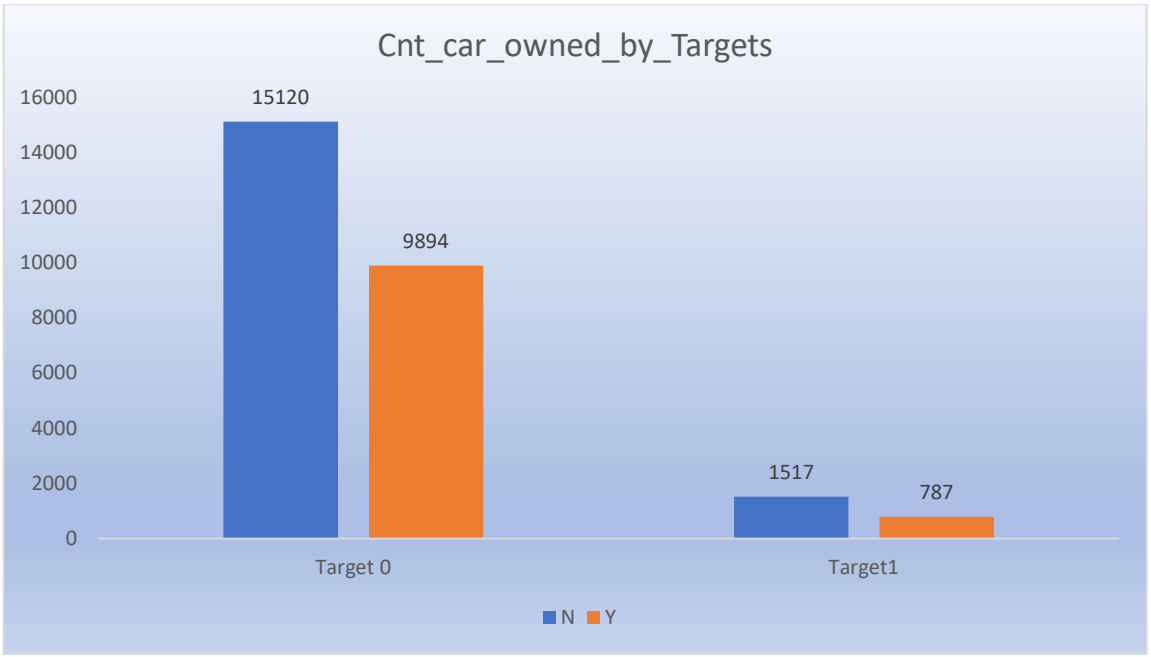
These visualizations will help you better understand the distribution of numerical variables and identify potential outliers for further investigation.

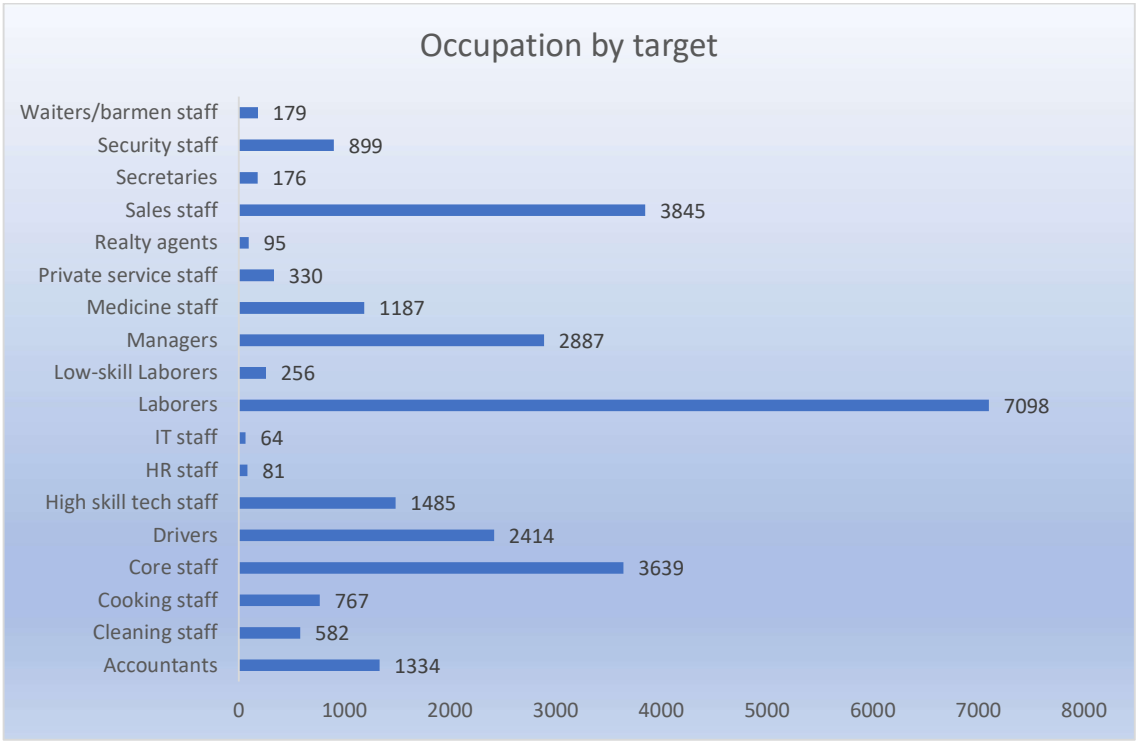
C. Analyze Data Imbalance:

1. Use the `COUNTIF` function to count the occurrences of each class in the target variable. Let's assume your target variable is in column A:
2. Calculate the proportions of each class by dividing the class frequency by the total number of samples:
3. Compare the class proportions to assess data imbalance. If the proportions are significantly different, there may be an imbalance issue.
4. Create a pie chart or bar chart to visually represent the distribution of the target variable. Highlight the class imbalance using different colors or annotations.
5. Select the class frequencies and proportions, including labels.
6. Go to the "Insert" tab and select either "Pie Chart" or "Doughnut Chart" from the Chart options.
7. Customize the chart to make it visually informative, such as adding data labels or a legend.

By following these steps, you can determine if there is data imbalance in your loan application dataset and create a visual representation of the class distribution. Adjust the formulas and chart options based on your specific dataset and requirements.





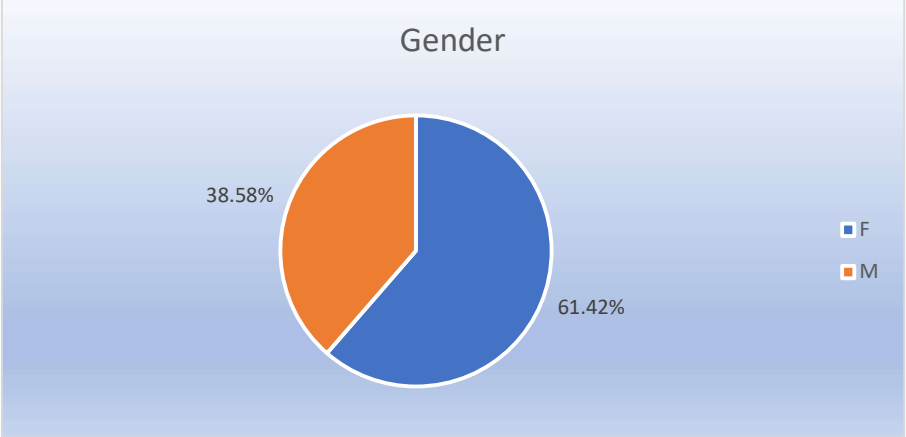

Occupation by target
Waiters/barmen staff 179
Security staff 899
Secretaries 176
Sales staff 3845
Realty agents 95
Private service staff 330
Medicine staff 1187
Managers 2887
Low-skill Laborers 256
Laborers 7098
IT staff 64
HR staff 81
High skill tech staff 1485
Drivers 2414
Core staff 3639
Cooking staff 767
Cleaning staff 582
Accountants 1334
0 1000 2000 3000 4000 5000 6000 7000 8000

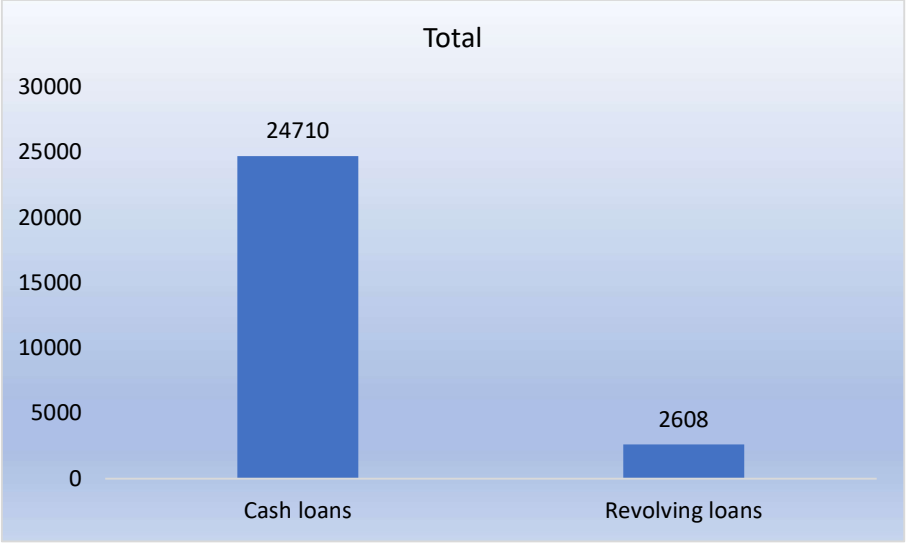

Ogranization by target
Transport: type 3
Trade: type 7
Trade: type 4
Trade: type 1
Self-employed
School
Realtor
Other
Medicine
Insurance
Industry: type 7
Industry: type 4
Industry: type 13
Industry: type 10
Hotel
Electricity
Cleaning
Business Entity Type 1
Advertising
0 1000 2000 3000 4000 5000 6000 7000
Target 1
Target 0

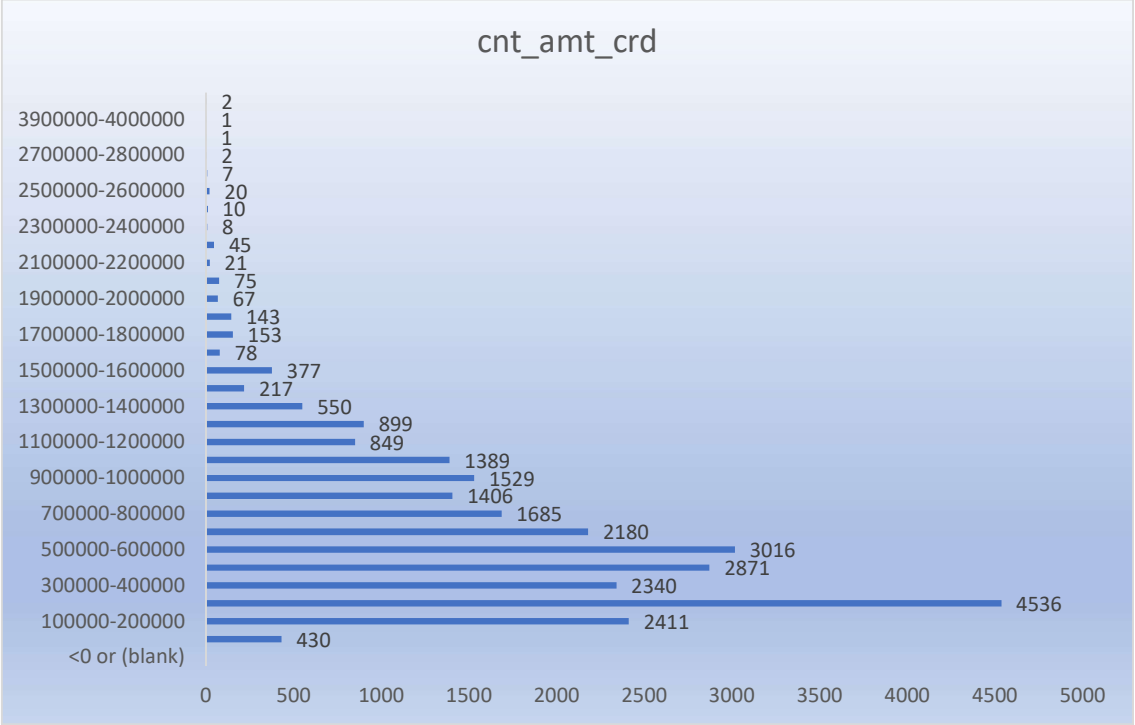
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

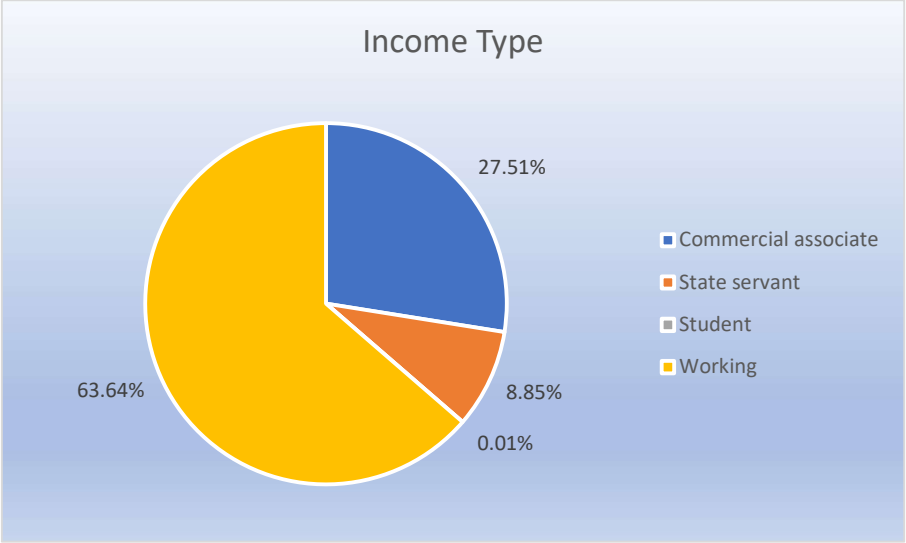
Univariate Analysis:

- List the variables you want to analyze (e.g., income, loan amount, credit score).
- Use functions like `COUNT`, `AVERAGE`, `MEDIAN`, `MIN`, `MAX`, and `STDEV` to calculate descriptive statistics for each variable.
- Visualize the distribution of each variable using histograms or box plots. You can use the "Insert" tab and select the appropriate chart type.
- Determine the criteria for segmenting the data (e.g., age groups, income brackets).
- Use Excel filters to segment the data based on your criteria.
- Repeat the univariate analysis steps for each segment separately.
- Create stacked bar charts or grouped bar charts to compare variable distributions across different segments. This can be done by selecting the relevant data and using the "Insert" tab to create the desired chart.





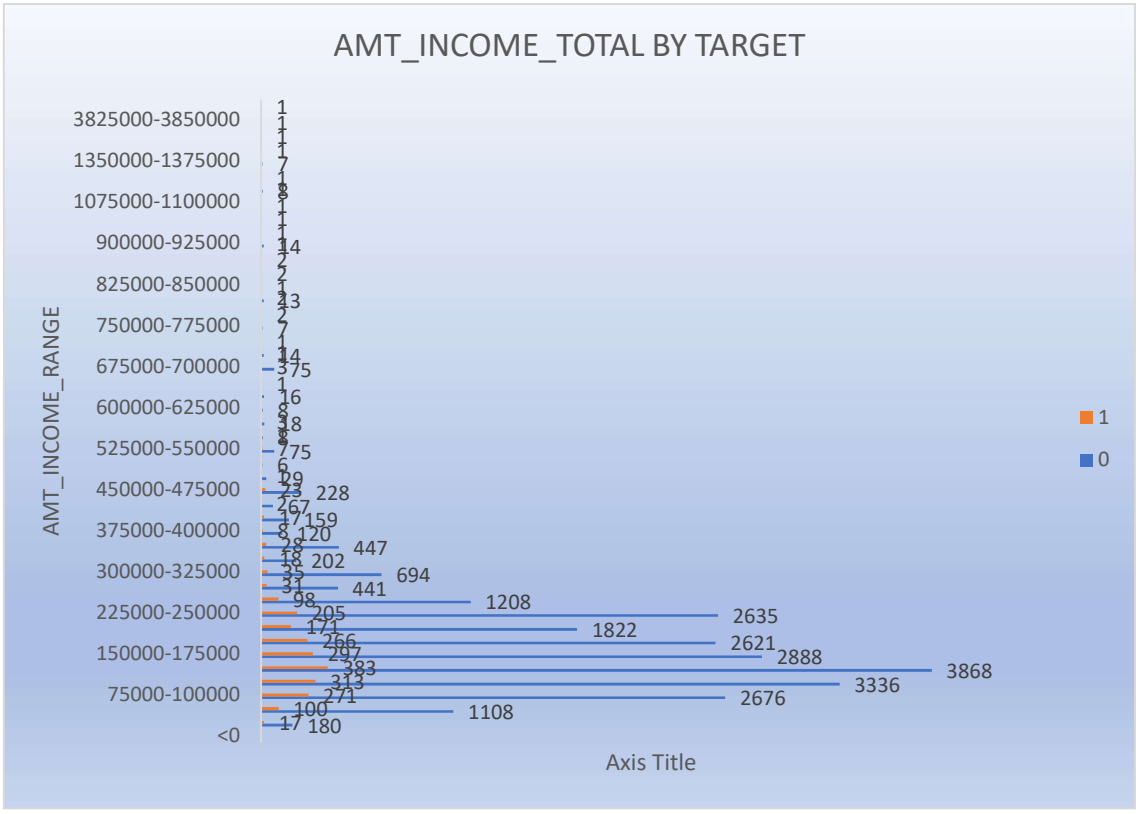


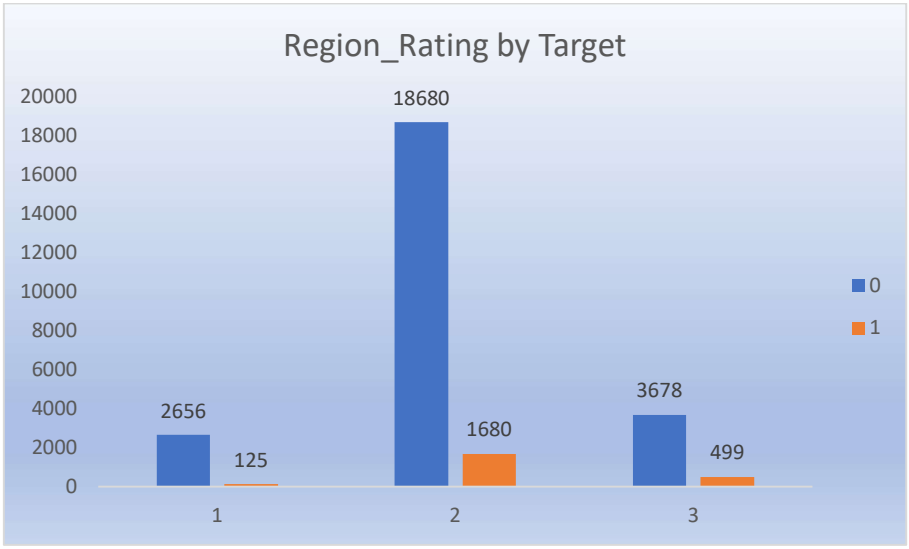


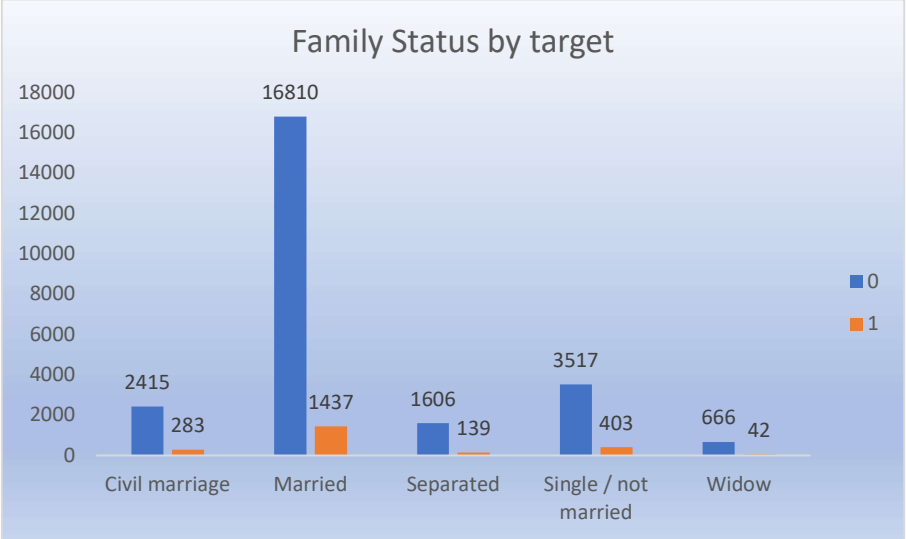
Bivariate Analysis:

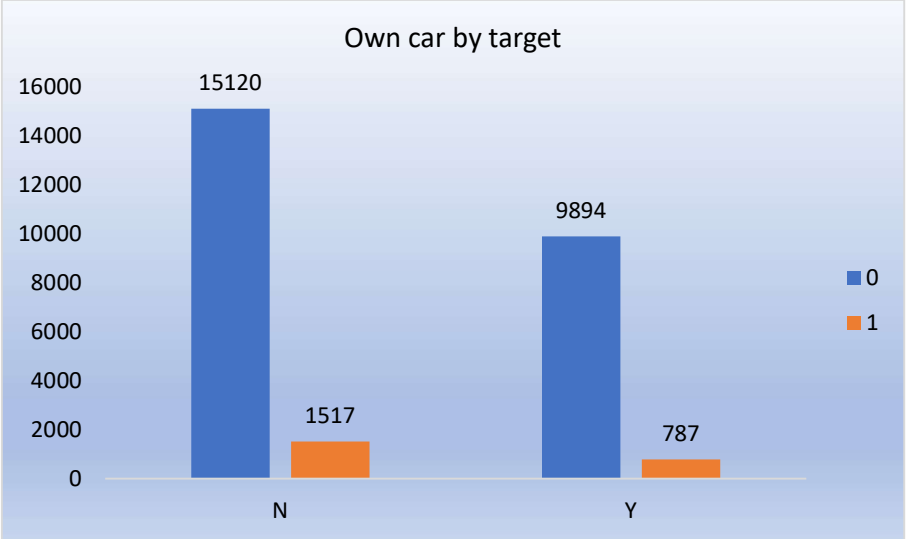
- Determine the relationships you want to explore (e.g., relationship between income and loan default).
- Use scatter plots to visualize the relationship between two continuous variables. Select the data and use the "Insert" tab to create a scatter plot.
- For categorical variables, use heatmaps to visualize the correlation. You can create a heatmap by using conditional formatting on a correlation matrix.
- Use the `COUNTIF` function or pivot tables to create cross-tabulations that show how two categorical variables relate to each other.

- Create graphical representations like stacked bar charts or grouped bar charts to compare the distribution of the target variable across different levels of another variable.
- Utilize Excel's pivot tables for dynamic and interactive analysis, especially in segmented and bivariate analyses.









By following these steps, you can perform a comprehensive analysis of consumer and loan attributes, gaining insights into the driving factors of loan default. Adjust the steps based on your specific dataset and analysis goals.

E. Identify Top Correlations for Different Scenarios:

1. Identify the scenarios for segmentation (e.g., clients with payment difficulties and all other cases).
2. Use Excel filters or other methods to segment the data based on the identified scenarios.
3. Identify the target variable for loan default (e.g., 1 for default, 0 for non-default).
4. Use the `CORREL` function to calculate correlation coefficients between each variable and the target variable.
5. Create correlation matrices or heatmaps to visualize the correlations between variables within each segment:Highlight the top correlated variables using different colors or shading for better emphasis.
6. Select the range of correlation coefficients for the variables within each segment.

7. Go to the "Insert" tab and choose "Heatmap" or use conditional formatting to create a visually appealing correlation heatmap.
8. Adjust the color scale to make strong correlations stand out.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH_years	DAYS_EMPLOYED_year	DAYS_ID_PUBLISH_year	REGION_RATING_CLIENT
CNT_CHILDREN	1	-0.004911747	-0.01601349	-0.00448395	-0.019907888	-0.026197114	-0.2557575	-0.070314647	0.129553223	0.035395697
AMT_INCOME_TOTAL	-0.004911747	1	0.365279441	0.440204156	0.372806512	0.179846628	0.054125487	0.026995747	0.014928611	-0.20806829
AMT_CREDIT	-0.016013487	0.365279441	1	0.762470519	0.986611816	0.093268561	0.160451708	0.089589592	0.034060456	-0.107726011
AMT_ANNUITY	-0.00448395	0.440204156	0.762470519	1	0.766884412	0.110554518	0.101612037	0.054256785	0.024018313	-0.125911444
AMT_GOODS_PRICE	-0.019907888	0.372806512	0.986611816	0.766884412	1	0.096656435	0.15528024	0.090603251	0.034520555	-0.108551053
REGION_POPULATION_RELATIVE	-0.026197114	0.179846628	0.093268561	0.110554518	0.096656435	1	0.044620857	-0.010412732	0.000656732	-0.523154439
DAYS_BIRTH_years	-0.2557575	0.054125487	0.160451708	0.101612037	0.15528024	0.044620857	1	0.345553495	0.072472675	-0.045952464
DAYS_EMPLOYED_year	-0.070314647	0.026995747	0.089589592	0.054256785	0.090603251	-0.010412732	0.345553495	1	0.064585448	0.017966232
DAYS_ID_PUBLISH_year	0.129553223	0.014928611	0.034060456	0.024018313	0.034520555	0.000656732	0.072472675	0.064585448	1	-0.002768905
REGION_RATING_CLIENT	0.035395697	-0.20806829	-0.10772601	-0.125911444	-0.108551053	-0.523154439	-0.045952464	0.017966232	-0.002768905	1

a. Target 0

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH
CNT_CHILDREN	1	0.00952242	0.012408771	0.016697551	0.00458965	-0.017700172	-0.1790284	-0.028866522	-0.140729592	0.10657788
AMT_INCOME_TOTAL	0.00952242	1	0.011070999	0.011559155	0.008663754	-0.012250841	-0.0055071	-0.010187961	0.016877994	0.013724345
AMT_CREDIT	0.012408771	0.011070999	1	0.736843501	0.981090827	0.048988455	0.18628417	0.07853943	0.037521632	0.039868297
AMT_ANNUITY	0.016697551	0.011559155	0.736843501	1	0.74027537	0.05037331	0.0820702	0.033755336	-0.014784462	0.059835163
AMT_GOODS_PRICE	0.00458965	0.008663754	0.981090827	0.74027537	1	0.057351504	0.17661649	0.085359746	0.033589001	0.044063135
REGION_POPULATION_RELATIVE	-0.017700172	-0.012250841	0.048988455	0.05037331	0.057351504	1	0.02517085	0.011165554	0.058260731	0.020898197
DAYS_BIRTH	-0.179028358	-0.00550713	0.186284168	0.082070203	0.176616493	0.025170845	1	0.305038457	0.240980889	0.118073716
DAYS_EMPLOYED	-0.028866522	-0.010187961	0.07853943	0.033755336	0.085359746	0.011165554	0.30503846	1	0.141461758	0.109653148
DAYS_REGISTRATION	-0.140729592	0.016877994	0.037521632	-0.014784462	0.033589001	0.058260731	0.24098089	0.141461758	1	0.01983706
DAYS_ID_PUBLISH	0.10657788	0.013724345	0.039868297	0.059835163	0.044063135	0.020898197	0.11807372	0.109653148	0.01983706	1

b. Target 1

By following these steps, you can effectively identify and visualize the top correlations for different scenarios within your loan application dataset using Excel functions and features. Adjust the steps based on your specific dataset and analysis goals.

Result:

The analysis has provided significant insights into the loan application dataset:

- Identification of missing data and appropriate handling methods.

- Detection and understanding of outliers and their potential impact on analysis.

- Assessment of data imbalance and its potential implications for classification models.

- Uncovering relationships between variables and loan default, highlighting key factors influencing default.

Drive Link:

https://docs.google.com/spreadsheets/d/18ogiujClVHCsqdIQ56u18lST_4FVsYPy/edit?usp=sharing&ouid=118439998565682353976&rtpof=true&sd=true