**Bank Loan Case Study**

**Final Project-2**

Difficulty Level:

**Description:**

Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

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

1. If the applicant can repay the loan but is not approved, the company loses business.
2. If the applicant cannot repay the loan and is approved, the company faces a financial loss.

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

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.
2. All other cases: These are cases where the payment was made on time.

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

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

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

**Business Objectives:**

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.  
  
**Note:** To better understand this project, you might want to research a bit about risk analytics in banking and financial services. Understanding the types of variables and their significance should be enough.

**Data Analytics Tasks:**

A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

* **Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
* **Hint:**Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.
* **Graph suggestion:**Create a bar chart or column chart to visualize the proportion of missing values for each variable.

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

* **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
* **Hint:**Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.
* **Graph suggestion:** Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

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

* **Task:**Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
* **Hint:** Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.
* **Graph suggestion:** Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

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

* **Task:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.
* **Hint:** Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.
* **Graph suggestion:** Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

E.**Identify Top Correlations for Different Scenarios:**Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

* **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.
* **Hint:**Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario.
* **Graph suggestion:** Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.