EDA Python Project Plan: Credit Card Payment Analysis

Objective:

The goal is to analyze the dataset of customer details and their credit card payment history (April 2005 to September 2005) to extract insights on factors that distinguish credible from non-credible customers.

General Submission Instructions:

General Submission Instructions:

- **Submission Format**:
    - Jupyter notebook (.ipynb) containing all code, visualizations, and explanations.
    - PDF report summarizing the findings, methodologies, visualizations, and insights derived from the analysis.
- **Code and Comments**:
    - Ensure code is well-organized into logical sections (e.g., data loading, preprocessing, exploratory data analysis, feature engineering, and model evaluation).
    - Include comments throughout the code to explain the purpose of each step, making it clear and understandable for anyone reviewing your work.
    - Use meaningful variable names and function names to enhance code readability.
- **Visualizations**:
    - Use relevant plots (e.g., histograms, box plots, bar charts, heatmaps) to support your analysis and effectively communicate findings.
    - Include captions and brief descriptions for each visualization to explain what is being shown and how it relates to your insights.
    - Ensure that visualizations are clear, well-labeled, and aesthetically pleasing, utilizing consistent color schemes and fonts.

- **Insights**:
  - Provide clear and concise insights based on the analysis performed.
  - Summarize key findings related to the factors that distinguish credible customers from non-credible customers.
  - Use bullet points or numbered lists for clarity and ease of reading.
- **Data Description**:
  - Include a brief description of the dataset, including the source, size, and any assumptions made during the analysis.
  - Refer to the data dictionary to explain the significance of each feature used in the analysis.
- **Deadlines**:
  - Ensure that each day's tasks are completed and submitted by **11:59 PM** on that day.
  - Keep a checklist to track the completion of tasks and submission requirements to avoid any last-minute issues.
- **Final Review**:
  - Before submission, review both the Jupyter notebook and the PDF report for clarity, grammar, and completeness.
  - Ensure that all required elements are included and that the work adheres to any specific guidelines provided by your instructor or organization.

Data Dictionary:

A **data dictionary** is a comprehensive description of the data elements within a dataset, outlining their attributes, types, and relationships. It serves as a reference to help users understand the structure, meaning, and constraints of the data. Below is a detailed explanation of the data dictionary for your dataset on customer credit card payment details:

Notes on Data Dictionary:

- **Data Type**: Specifies the nature of the data (e.g., numeric, categorical), which is essential for data processing and analysis.
- **Possible Values**: Lists the acceptable values for each feature, which helps ensure data integrity and consistency during analysis.
- **Description**: Provides a concise explanation of what each feature represents, aiding in understanding the data and guiding the analysis process.

#	Feature	Description	Data Type
1	ID	Unique Identifier	Numeric
2	LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family credit.	Numeric
3	GENDER	Gender	Categorical
4	EDUCATION	Education	Categorical
5	MARRIAGE	Marital status	Categorical
6	AGE	Age	Numeric
7	PAY_1	the repayment status in September, 2005	Categorical
8	PAY_2	the repayment status in August,2005	
9	PAY_3	the repayment status in July,2005	
10	PAY_4	the repayment status in June,2005	
11	PAY_5	the repayment status in May,2005	
12	PAY_6	the repayment status in April,2005	
13	BILL_AMT1	amount of bill statement in September, 2005	Numeric
14	BILL_AMT2	amount of bill statement in August, 2005	
15	BILL_AMT3	amount of bill statement in July, 2005	
16	BILL_AMT4	amount of bill statement in June, 2005	
17	BILL_AMT5	amount of bill statement in May, 2005	
18	BILL_AMT6	amount of bill statement in April, 2005	
19	PAY_AMT1	amount paid in September, 2005	
20	PAY_AMT2	amount paid in August, 2005	
21	PAY_AMT3	amount paid in July, 2005	
22	PAY_AMT4	amount paid in June, 2005	
23	PAY_AMT5	amount paid in May, 2005	
24	PAY_AMT6	amount paid in April, 2005	
25	DEFAULT	default payment, as the response variable	Categorical

Day 1: Project Setup and Data Understanding

Objectives:

- Set up the project environment.
- Gain an initial understanding of the dataset and its features.
- Perform basic data loading and inspection.

Instructions:

1. **Project Setup:**
    - Install necessary libraries such as pandas, numpy, matplotlib, seaborn, scikit-learn, etc.
    - Create a Jupyter notebook and set up your environment.
2. **Data Loading:**
    - Load the dataset into a pandas DataFrame.
    - Inspect the first few rows using .head() to understand the structure.
    - Check the column data types using .info().
    - Check for missing values using .isnull().sum().
3. **Initial Exploration:**
    - Generate basic summary statistics using .describe() for numerical and categorical features.
    - Identify the target variable (DEFAULT), which indicates whether a customer defaulted (1 = Yes, 0 = No).
    - Understand the range, possible values, and meaning of each feature.

Expectations:

- Clear explanations of each feature (what it represents and its potential influence on the target).
- Identification of any initial issues such as missing values or unusual data points.

Day 2: Data Cleaning and Feature Engineering

Objectives:

- Handle missing values and outliers.
- Engineer new features if relevant.

Instructions:

1. **Data Cleaning:**
    - Handle missing values (e.g., fill, drop, or impute based on the context).
    - Address any outliers (check distributions and decide on actions like capping, flooring, or transformation).
    - Review the unique values for categorical features (GENDER, EDUCATION, MARRIAGE, etc.) and decide whether you need to group or rename some categories (e.g., combine Unknown education levels).
2. **Feature Engineering:**
    - **Create new features** if they might provide additional insight. For example:
        - **Average 6 months Bill Amount**
        - **Average 6 months Payment Amount**
    - Transform categorical variables into numerical representations
3. **Exploratory Visualizations:**
    - Create initial visualizations like boxplots or histograms to understand feature distributions.
    - Correlation heatmap to identify relationships between numerical features.

Expectations:

- No missing values or inconsistent data.
- Clean and structured dataset with potential new features.
- Visualizations with interpretations of patterns and distributions.

Day 3: Exploratory Data Analysis (EDA)

Objectives:

- Conduct detailed analysis to find patterns between the features and the target variable (DEFAULT).
- Visualize the relationships and distributions.

Instructions:

1. **Univariate Analysis:**
    - Analyze individual features (distribution, outliers, etc.) using histograms, count plots, or boxplots.
    - Focus on how features like LIMIT_BAL, AGE, BILL_AMT1-6, and PAY_AMT1-6 behave across the dataset.
2. **Bivariate Analysis:**
    - Explore relationships between the target variable (DEFAULT) and individual features using:
        - Bar plots for categorical features like GENDER, EDUCATION, MARRIAGE, etc.
        - Box plots or violin plots for numerical features like LIMIT_BAL, AGE, BILL_AMT, and PAY_AMT values split by DEFAULT.
        - Correlation analysis between BILL_AMT, PAY_AMT, and DEFAULT.
3. **Multivariate Analysis:**
    - Pair plots (using seaborn.pairplot) to identify relationships between multiple numerical features.
    - Visualize the correlation matrix and identify highly correlated features.

Expectations:

- Insights on which features are positively or negatively correlated with DEFAULT.
- Well-annotated code and conclusions from the analysis.

Day 4: Conclusion, Final Insights, and Submission Preparation

Objectives:

- Summarize insights gained from the EDA.
- Prepare final deliverables for submission.

Instructions:

1. **Final Insights:**
    - Summarize the key takeaways from your EDA.
    - Identify the most important factors affecting whether a customer defaults.
    - Highlight any surprising or unexpected findings.
2. **Report Preparation:**
    - Create a clean final version of the Jupyter notebook.
    - Ensure all code cells are run and the notebook is properly formatted with markdown explanations.
    - Add a conclusion section summarizing your findings.
3. **Presentation (Optional):**
    - Prepare a few slides summarizing the EDA process, key insights, and potential next steps.
    - Include visualizations and conclusions for key stakeholders.