Assignment: Exploratory Data Analysis (EDA) and Data Cleaning Report

**Assignment Title:**

**EDA and Data Cleaning Report for Capstone Project**

**Due Date: Week 3**

**Objective:**

The purpose of this assignment is to prepare a thorough Exploratory Data Analysis (EDA) and Data Cleaning report as a foundational step for your capstone project. This report will help you understand the structure, patterns, and issues in your data before moving on to advanced analysis and modeling. Through this assignment, you will gain insights into your dataset, clean it for better usability, and ensure that it's ready for the next phases of your project.

**Instructions:**

**1. Dataset Overview:**

* Provide a brief description of the dataset you are working with.
  + **What is the source of the dataset?**
  + **What are the key features (columns)?**
  + **What does each feature represent?**
  + **Is there any metadata associated with the dataset?**

**Deliverable:**  
A summary of your dataset including its source, size, number of features, and an initial assessment of its relevance to your capstone project.

**2. Initial Data Inspection:**

* Perform an initial inspection of the dataset:
  + **Print the first few rows of the data (e.g., using .head() in Python).**
  + **Inspect the data types of each column.**
  + **Check for missing values.**
  + **Identify outliers or unusual entries (e.g., using .describe()).**

**Deliverable:**  
A concise inspection of your data structure including data types, missing values, and outliers.

**3. Univariate Analysis:**

* Analyze each feature independently:
  + **For numerical features:** Calculate summary statistics (mean, median, standard deviation, min/max values).
  + **For categorical features:** Count the frequency of each category.
  + **Visualize distributions:** Use histograms, boxplots, or bar charts to visualize numerical and categorical distributions.

**Deliverable:**  
Summary statistics for each feature, along with visualizations (histograms, boxplots, or bar charts) of the most important features.

**4. Bivariate/Multivariate Analysis:**

* Investigate relationships between pairs or sets of features:
  + **Correlation matrix** for numerical features (e.g., using .corr()).
  + **Scatterplots, pair plots, or heatmaps** to visualize relationships between features.
  + **Cross-tabulation** for categorical data to understand interactions between categories.

**Deliverable:**  
A correlation matrix and at least two visualizations exploring relationships between key features.

**5. Data Cleaning:**

* Perform data cleaning to prepare the dataset for further analysis:
  + **Handle missing data:** Impute missing values or drop rows/columns with excessive missing values.
  + **Handle outliers:** Discuss and decide whether to remove or modify outliers.
  + **Correct data types:** Ensure that each feature has the correct data type.
  + **Create new features** if necessary (e.g., aggregating categories, creating dummy variables for categorical features).

**Deliverable:**  
A description of the cleaning steps taken, including the treatment of missing values, outliers, and any new feature engineering. Provide code snippets where appropriate.

**6. Data Quality Report:**

* Provide a summary of the overall quality of the dataset:
  + **What are the key issues identified?**
  + **How did you resolve these issues (e.g., cleaning, feature engineering)?**
  + **Is the dataset now ready for analysis? What further cleaning steps may be needed later?**

**Deliverable:**  
A final report summarizing the data quality and any remaining issues or limitations.