Exploratory Data Analysis (EDA) Task

Objective

The goal of this task is to perform a **complete Exploratory Data Analysis (EDA)** on the given dataset. You are expected to **understand, clean, analyze, visualize, and communicate insights** from the data.

This task is designed to evaluate:

- Your data preprocessing skills
- Your ability to extract insights using EDA
- Your visualization choices and reasoning
- Your clarity in explaining findings

Instructions

1. Notebook Requirements

- Perform **all work in a single Jupyter Notebook**.
- Use **Markdown cells extensively** to explain:
    - What you are doing
    - Why you are doing it
    - What insights you observe
- **Do NOT clear outputs** before pushing the notebook to the repository.
- The notebook should read like an **EDA report**, not just code.

---

2. Data Loading & Initial Inspection

- Load the dataset using appropriate libraries.
- Display:
    - First few rows
    - Shape of the dataset
    - Column names and data types
- Use functions like:
    - `.info()`
    - `.describe()`

- Write brief observations about:
    - Dataset size
    - Types of features (numerical / categorical)
    - Any immediate issues you notice

---

3. Data Cleaning & Preprocessing

Perform and **justify** the following steps:

a. Missing Values

- Identify missing values
- Decide how to handle them:
    - Drop
    - Impute (mean/median/mode/other)
- Explain **why** you chose that method

b. Duplicates

- Check for duplicate rows
- Handle them appropriately
- Mention their impact (if any)

c. Data Types

- Convert incorrect data types if required
- Explain why the conversion was necessary

d. Feature Engineering (if applicable)

- Create new features if they help analysis
- Explain the intuition behind them

---

4. Univariate Analysis

Analyze individual features using plots and statistics.

Mandatory plots:

- Histograms
- Box plots
- Count plots (for categorical variables)

Write observations such as:

- Distribution shape
- Skewness
- Presence of outliers
- Class imbalance (if any)

5. Bivariate & Multivariate Analysis

Explore relationships between variables.

Use and justify:

- Scatter plots
- Bar plots
- Correlation heatmaps
- Pair plots (if feasible)

Explain:

- Trends
- Correlations
- Interesting interactions between features

6. Outlier Detection & Handling

- Identify outliers using:
  - Box plots
  - IQR method
  - Z-score (if relevant)
- Decide how to handle them:
  - Keep
  - Cap
  - Remove
- Clearly justify your choice and its impact on the dataset

7. Advanced / Niche Visualizations (Mandatory)

You **must** include and use the following plots at least once:

- Box Plot
- Violin Plot

For each of these plots:

- Use them on meaningful features
- Write **1–2 lines** explaining:
    - Why you used this plot
    - What extra information it provides compared to simpler plots

---

8. Final Dataset Check

- Show the final shape of the dataset
- Summarize how the dataset changed after preprocessing
- Mention:
    - Rows removed/added
    - Columns modified/created

---

9. Key Insights & Summary

In a Markdown cell, summarize:

- 5–10 key insights from your EDA
- Patterns or anomalies discovered
- How these insights could help in modeling or decision-making

---

Submission Guidelines

- Push the notebook to the repository
- **Do not clear outputs**
- Ensure:
    - Clean code
    - Proper headings
    - Clear explanations

Bonus (Optional)

- Use interactive plots (Plotly, etc.)
- Compare distributions before vs after outlier handling
- Add assumptions or limitations of your analysis

Evaluation Criteria

- Completeness of EDA
- Quality of visualizations
- Clarity of explanations
- Correctness of preprocessing steps
- Overall presentation

Good luck, and treat this as a real-world EDA report!