Midterm Assignment: Data Science Analysis and Visualization

Objective:
The goal of this assignment is to assess your understanding and application of basic data science techniques. You will need to find a real-world dataset online, clean and prepare it for analysis, and then perform exploratory data analysis (EDA) using basic statistical techniques and visualizations. You will be required to submit your work as a Jupyter Notebook/R Markdown file.

Assignment Requirements

1. **Find a Dataset**
    - Choose a dataset of your interest from a reputable source (e.g., Kaggle, UCI Machine Learning Repository, government open data portals, or any domain-specific sources like healthcare, finance, sports, etc.).
    - Dataset should be rich enough to allow for basic data analysis (i.e., should have at least 5-10 columns and 100+ rows).
    - You are allowed to use publicly available datasets or APIs.
    - ***Place a link and brief description of your dataset on the class discussion board on blackboard. Verify that the dataset is unique to you. If someone else has already posted the dataset you plan to use, you need to choose another.***

Examples of sources:

- Kaggle Datasets
- UCI Machine Learning Repository
- Google Dataset Search

---

2. **Data Preprocessing**
    - **Load the dataset** into R or Python.
    - Inspect the dataset structure using str() or head() (in R) / info() or head() (in Python).
    - Handle missing data:

        - Identify missing values (use is.na() in R or isnull() in Python).
        - Perform one of the following based on the dataset:
            - Drop rows/columns with too many missing values.
            - Impute missing values with the mean/median (for numerical columns) or mode (for categorical columns).
    - Check for and handle duplicates in the dataset.
    - Remove any outliers that may affect your analysis. (e.g., using IQR or Z-scores for numerical columns).

---

3. **Exploratory Data Analysis (EDA)**
   Perform basic exploratory data analysis using summary statistics and visualizations.
    - **Summary Statistics**:
        - Calculate the following summary statistics for numerical columns:
            - Mean, Median, Mode, Standard Deviation, Min, Max, Range, and Quartiles.
        - For categorical columns:
            - Frequency of unique values (use table() in R or value_counts() in Python).
    - **Visualizations**:
        - **Histograms** for numerical variables to show the distribution of the data.
        - **Boxplots** for numerical data to identify outliers and visualize the distribution.
        - **Scatter Plots** to explore relationships between two numerical variables (if applicable).
        - **Bar Charts** for categorical variables to show the frequency of each category.
        - **Correlation Heatmap** to visualize correlations between numerical variables (if applicable).

        - **Pairwise Plots** or **Facet Grids** for multi-variable analysis.
    - **Observations**:
        - Based on your visualizations and summary statistics, describe any key patterns or trends you observe.
        - Identify potential relationships between variables.
        - Comment on any anomalies or potential outliers in the data.

---

4. **Data Visualization (Advanced)**
   Choose one of the following tasks to perform an advanced visualization:
    - **R**: Use ggplot2 to create a detailed and informative plot (e.g., a scatter plot with a regression line, bar chart with color, customized themes, etc.).
    - **Python**: Use matplotlib and seaborn for data visualization. Create customized plots such as a regression plot, heatmap, or a pairplot.

Note: Add appropriate labels, titles, and legends to your plots.

---

5. **Basic Statistical Analysis**
    - **For Numerical Variables**:
        - Perform a hypothesis test (e.g., t-test, ANOVA, or chi-squared test if comparing different groups). State your null hypothesis, alternative hypothesis, and interpret the results.
    - **For Categorical Variables**:
        - Perform a chi-squared test to assess independence between two categorical variables, if applicable.
    - Interpret the results of your statistical tests and discuss any insights.
    - Stats review:
        - https://www.khanacademy.org/math/statistics-probability/analyzing-categorical-data
        - https://www.youtube.com/@statquest/videos

6. **Create a Report or Notebook**
    - **R Markdown** or **Jupyter Notebook**:
        - Document all the steps in your analysis, including code and explanations.
        - Explain your choices for data cleaning, visualizations, and any transformations you made to the data.
        - Provide your observations and insights at each step of the analysis.
        - Discuss any challenges you faced during the analysis and how you overcame them.

Submission Guidelines

- Submit your **Jupyter Notebook (.ipynb)** or **R Markdown (.rmd)** file.
- Include all code, visualizations, and explanations in the file.
- Be sure to **comment** your code where appropriate and **explain** the logic behind each step.
- Write a **summary** of your analysis and conclusions at the end of the notebook. Try to address the exploratory questions below.

Grading Rubric (Total 100 Points)

1. **Dataset Selection (10 points)**
    - Chosen dataset is appropriate, clean, and sufficiently complex for basic data analysis.
2. **Data Cleaning and Preprocessing (20 points)**
    - Proper handling of missing data, duplicates, and outliers.
    - Clear explanation of data cleaning choices.
3. **Exploratory Data Analysis (20 points)**
    - Summary statistics calculated and explained.

- Visualizations are clear, appropriate, and well-labeled.
- Insightful interpretation of visualizations and statistics.

4. **Advanced Data Visualization (20 points)**
    - Well-designed, informative, and customized plot using ggplot2 (R) or matplotlib/seaborn (Python).
    - Visualizations are insightful and help communicate key findings.
5. **Statistical Analysis (10 points)**
    - Proper application and interpretation of hypothesis tests (t-test, ANOVA, chi-squared, etc.).
    - Clear explanation of statistical methods used and their interpretation.
6. **Report/Documentation (20 points)**
    - Clear and concise documentation.
    - Well-organized code with appropriate comments and explanations.
    - Insightful summary and discussion of findings.

Exploratory Questions

1. **Data Understanding**:
    - What does each column in the dataset represent? Are there any columns that need additional explanation or clarification?
    - What is the overall structure of the data? Are there any patterns or trends you notice from the summary statistics?
2. **Data Cleaning**:
    - Were there any missing values or duplicates in the dataset? How did you handle them?
    - Did you identify any outliers? How did you decide to handle them?
3. **Visualization**:
    - What do the histograms and boxplots tell you about the distribution of your variables?

- Do you notice any relationships between numerical variables? Can you visualize these relationships effectively?

4. **Statistical Analysis**:
    - What hypothesis tests did you use, and why did you choose them?
    - What were the results of the tests? What do they tell you about the data?
5. **Conclusions**:
    - What insights have you gained from the analysis? What recommendations or further analysis would you suggest based on your findings?
