COMP517–Data Analysis

ASSIGNMENT ONE

Data Exploration and Analysis

Semester 2, 2024

Due Date: Midnight Friday 30th August 2024.
Total Marks: 100

Late submissions will incur a 10 marks penalty per day.

Submission:

When submitting the assessment, the name and student ID must be indicated on the front page of the report.

Note: This assignment must be completed individually, and all work submitted must be entirely your own.

Introduction

The purpose of this assignment is to provide you with a hands-on opportunity to explore the world of data analysis through the lens of exploratory data analysis (EDA). By working with a provided dataset, the assignment aims to empower you to become proficient in analysing and understanding data, thereby enabling them to extract valuable insights and make data-driven decisions.

The dataset for this assignment is provided in the "BikeSharing Dataset.csv" file available under the Assignment One section on Canvas. This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. For more information about the dataset see 'Readme' file available under the Assignment One section on Canvas.

1. **Dataset** (10 marks)

Download and load the dataset (BikeSharing Dataset.csv) and use it for the analysis.

- Explain the purpose of your report and provide a short summary of the dataset.
- Use appropriate Python libraries to load the dataset into a DataFrame.
- Display the first few rows of the dataset to get a quick overview.
- Describe your observations and provide information about the dataset attributes (shape, data types, etc.).

2. **Data Pre-processing** (20 marks)

You are required to report your preprocessing steps:

a) Handling Missing Values:
   - Identify columns with missing values and decide how to handle them (e.g., imputation, removal). Justify your choice.
   - Use appropriate methods to fill in missing values.

b) Handling Duplicates:
   - Check for duplicate rows in the dataset.
   - Describe your observations and decide whether to keep or remove duplicate rows based on the analysis context. Justify your choice.

c) Handling Outliers:
   - Identify potential outliers in numerical data using both statistical and visualization methods.
   - Identify the attributes with outliers and plot the outlier/non-outliers using scatterplots for those attributes.
   - Decide whether to keep, remove, or transform outliers based on their impact on the analysis. Justify your choice.

3. **Explore the Visualize Clean Dataset** (25 marks)

- Calculate basic summary statistics (e.g., mean, median, standard deviation) for numerical columns and explain your findings.
- Use visualization methods to understand the distribution of numerical and categorical data.
- Use techniques like histograms, box plots, scatter plots, pie plots, etc.
- Provide suitable labels, titles, and legends for each plot.
- Describe your findings.

4. **Multivariate Analysis** (20 marks)
   a) Correlation Analysis:
      - Calculate and visualize correlations between numerical features.
      - Discuss any strong positive or negative correlations between variables.
      - Provide appropriate labels, titles, and legends for the plot.
      - Discuss any insights or patterns observed during the analysis.
   b) Perform multivariate analysis of data to visualize relationships, in one plot, between two categorical variables of your choice (the choice of these variables should be rationale) AND the 'count' variable.
      - variables, allowing you to understand how the average 'count' or distribution of 'count' varies across different categories.
      - Provide appropriate labels, title, and legend for the plot.
      - The plot should summarize the 'count' values within each category of the categorical Aggregation Analysis
      - Explain your findings.
   c) Perform aggregation analysis to calculate two common statistics (Mean and Median) for the 'count' column in the dataset for your choice of attribute (the choice of this variable should be rationale).
      - Provide your results in both tabular and graphical format.
      - Provide appropriate labels, title, and legend for the plot.
      - Explain your findings.
   d) Perform analysis to show how different 'season' affect the 'count' on average and how much variation exists using an appropriate plot.
      - Provide appropriate labels, title, and legend for the plot.
      - Discuss two significances of performing such analysis.

5. **Conclusion** (15 marks)
      - Summarize the key findings and insights gained from the EDA process.
      - Discuss any challenges faced during the analysis and how you addressed them.
      - Suggest possible next steps for further analysis or data preprocessing.

6. **Report presentation** (10 marks)
   - Your report must include the following elements:
     - **Title, Full Name, and Student ID:** Clearly state your title, full name, and student ID at the beginning of the report.
     - Table of Contents: Include a table of contents to provide an overview of the report's structure.
     - **List of Figures/Tables:** Provide a list of figures and tables used in your report for easy navigation.
     - **Answers to Questions:** Present your answers to the questions asked. Explain your findings, insights, and observations in a clear and concise manner.
     - **Figures (Plots) and Tables:** Include all relevant figures and tables that support your answers. However, DO NOT include the actual code used to generate these visualizations and tables.
     - **Informative Labels and Captions:** Ensure that all visualizations and tables have informative labels and captions with suitable resolution to help the reader understand their significance.
   - Ensure that your code is clean, well-organized, and properly commented.
     - The code must be ready to execute without errors.

Submission Instructions

Please submit the following two files **separately** as part of your assignment:

1. Python Notebook or Code File (.ipynb, .py):

2. Report File (PDF Format)

Note: the report should focus on presenting your findings and insights, rather than including the code itself. Please refrain from including the code file in your report, as including code in the report will result in a penalty.