COMP615 – Foundations of Data Science

ASSIGNMENT ONE

Data Exploration and Regression Analysis

Semester 1, 2022

Due Date: Midnight Friday 15th April 2022.

Late submissions will incur a 10 marks penalty per day.

Weighting: 25% of the final course mark

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, all work submitted must be entirely your own.

This assignment assesses your skills for data exploration and regression analysis of a dataset using Python. You are required to describe characteristics of a given dataset with the help of diagrams and statistical metrics, build a regression model and provide inference for your regression model.

Preparation

Choose **one of the datasets** listed below. These datasets are selected from the UCI Machine Learning repository. More information about the data is provided by UCI Maching Learning repository website. You must carefully read this information to understand the dataset attributes.

1. Beijing PM2.5 Data Data Set (2010-2014)
2. Beijing Multi-Site Air-Quality Data Data Set (2013-2017)
3. Electric power consumption Data Set
4. Istanbul stock exchange data set
5. Italian Air Quality Data Set
6. QSAR fish toxicity Data Set
7. Student Performance Data Set

Project Report – Basic Structure

Your final report must include the following sections and information. The word count gives you an indication of the expected length for each section.

Title Page must include:

- Paper Code, Paper Name and Semester
- Project Title
- Student Name and ID
- Table of Contents, List of Figures and Tables

Deliverables:

Download your code as ‘Markdown’ and attach it as an appendix at the end of your report.

DO NOT use screenshots of your code.

Task 1: Introduction (200-400 words) **[10 marks]**

Provide a statement of the problem, outlining the problem that your chosen dataset addresses. The statement of the problem should briefly address the question: What is the problem that you will investigate in this assignment?

Your introduction must describe:

- The aim of your work, what are you trying to achieve, and research questions you attempted to answer.
- All assumptions that your data must meet.

Task 2: Data Exploration **[20 marks]**

This section of your report must discuss the dataset and any features which you consider to be relevant to the analysis and modelling task.

- What is the data about?
- How many features (attributes), instances and what data types are these?
- Provide summary statistics of the continuous numerical features.
- Illustrate the features of your dataset using **meaningful** boxplots, histograms and grouped scatter plots (remember, these plots allow you to analyse the individual distribution of features, as well as the relationship between them).
- Explain what you can learn about the dataset from the diagrams. Is there any sign of violation of assumption(s)? If yes, explain your approach to handle it before moving to the next tasks?

Task 3: Correlation Analysis **[15 marks]**

Perform correlation analysis and provide correlation matrixes and plots. Discuss your findings in terms of a) correlation between the independent variables (Multi-Collinearity) and b) between independent variables and dependent variables. Is there any sign of violation of assumption(s)? If yes, explain your approach to handling it before moving to the next task?

Task 4: Linear Regression [25 marks]

Taking the result of Task 3 into consideration, perform multiple regression using Python. Provide the results including regression results, statistical significance metrics, and coefficients tables from this model.

Task 5: Statistical Inference (300-400 words). [25 marks]

In this section, you are required to describe and analyse the results of your regression model. Are all the assumptions made in Task 1 satisfied? Provide evidence to support your answer. Compare your findings in Task 3's with the coefficients table results generated in Task 4 and discuss your findings.

There will be **5 marks** for the presentation of the assignment including spelling and grammar, layout, formatting, and readability of the figures.

Good luck!