SCHOOL OF COMPUTER SCIENCES,
UNIVERSITI SAINS MALAYSIA

CPC152 Foundations and Programming for Data Analytics

Semester 2, 2023/2024

Assignment (20%)
(Individual Work)

INSTRUCTION:

The assignment will be evaluated based on individual performance via written report. Every student must submit a written report (softcopy).

DEADLINE:

Wednesday 8th May 2024 (5:00 pm), submit the softcopy of your report & source codes (Python) in appendix through CPC152 e-learning portal.

ASSIGNMENT BACKGROUND:

Over time, Python has developed as one of the most suitable languages for data analysis. It is a high-level programming language that is easy to understand and contains more popular libraries such as NumPy, Pandas, Matplotlib, Bokeh etc., that are much needed for analysing the datasets.

In this assignment, you have to analyse the `car_prices.csv` dataset. It explains the selling details of cars. The dataset contains historical car auction sales prices, scraped from the outside internet sources. The dataset has been collected in 2015, and will not be updated.

ASSIGNMENT TASKS:

Data analysis process has six steps: Problem Definition, Data Collection, Data Cleaning and Data Processing, Data Analysis, Data Visualization and Data storytelling.

Task 1: Clearly state the problem definition (Student's choice).

Task 2: Collect the data that is relevant to solve the defined problem.

Task 3: Clean the data if any empty space or duplicates and Process the cleaned data

Task 4: Analyse the entire dataset using exploratory data analysis (EDA) technique.

a. The general information about the entire data set `car_prices.csv`
b. Find the most expensive car with details of make, model, seller and selling price.
c. Find the number cars sold in each model of BMW
d. Find the total number of cars sold in 2000
e. Do additional EDA steps to explain your problem definition in detail.

Task 5: Create the visuals using Matplotlib or Bokeh

a. Choose the suitable type of chart to plot an interactive chart for the number of cars sold over year after 2000.
b. Choose the suitable type of chart to plot an interactive chart for the top 10 seller to their total selling price.
c. Plot multiple graphs to explain your problem definition in detail.

Task 6: Make the report in a manner that's understandable for all types of audiences.

ASSIGNMENT CONTENT:

1. Abstract. a summary of the contents of your assignment work
2. Data Analysis Process: Follow the assignment tasks

    2.1 Problem Definition

    2.2 Data Collection

    2.3 Data Cleaning and Data Processing

    2.4 Data Analysis

    2.5 Data Visualization

    2.6 Data storytelling

3. References (at least 3 reference), Follow IEEE format.
4. Appendix (Add your jupyter notebook code and output)

GENERAL GUIDELINES

Grading of written assignment is based on appropriate assignment tasks, the quality of writing of your assignment and the format of the assignment. The assignment must be free from grammatical error and spelling mistakes.

The following guidelines for formatting assignment must be followed when you are submitting a softcopy. Use a standard font 12 pt Arial, apply 1.5 line spacing, use bold for

headings, use font 12 pt Courier New for code only, align the content with justify and add bottom-centered page number.

Include the screenshots for the assignment contents 2.3, 2.4 and 2.5. The screenshots must be clear.

In Jupyter notebook, each line in the program must be explained by using suitable comment statements.

MARKING SCHEME AND DATA SET (PROVIDED):

- Marking Scheme: refer to the rubrics posted on the e-learning page
- Source of the data set – posted on e-learning page

UPLOAD THE FOLLOWING TO THE LINK PROVIDED IN CPC152 E-LEARNING PORTAL BEFORE THE DEADLINE:

- Soft copy – Report with Source codes (Python) in Appendix

Note: You must write the report yourself according to assignment tasks and not simply cut and paste from your references. List all references in the references. If the part of this assignment has been copied from any other source or person, F grade is given.