School of Computer Sciences

CPC351/CPM351 Principles of Data Analytics

Academic Session: Semester 1, 2021/2022

Assignment 02 – Data Exploration and Visualization

I. Dataset

Download the following files from eLearn@USM:

- “Government_Schools_Pupils_Teachers_2017-2018.xlsx”
- “healthcare-dataset-stroke-data.csv”

The first dataset is based on the statistics provided by Department of Statistics of Malaysia, via https://www.dosm.gov.my/v1/index.php?r=column3/accordion&menu_id=amZNeW9vTXRydTFwTXAxSmdDL1J4dz09. The use of data shall abide by the terms and conditions, i.e. Terms of Use Government Open Data 1.0. The second dataset is taken from Kaggle, via https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

II. Pupils and Teachers in Government Schools

“Government_Schools_Pupils_Teachers_2017-2018.xlsx” contains number of pupils and teachers by states recorded for the years 2017 and 2018. Using R, answer the following questions.

1. Load the data into R environment. Then, perform data pre-processing and data cleaning:
   a) Some of the variables are not defined with correct data type. Convert these variables such that they are with a suitable data type.
   b) Are there any missing values? Identify the variables that contains missing values and state the total number of missing values. Remove all the missing values.
   c) Rename the variables name to "school_stage", "state", "district", "year", "school_type", "gender", "number_of_pupils", "number_of_teachers".
2. Create a pie chart to show the number of primary school pupils by years, i.e. 2017 and 2018. Then, create two bar charts to show the number of primary school pupils by states in 2017 and 2018 respectively. Perform the same steps for the secondary school pupils.
3. For the years 2017 and 2018, identify the states which have more female pupils. Explain your answer with appropriate visuals.
4. Show the number of pupils in Perak by year. Is there any increase in the number of female pupils from 2017 to 2018 in Perak? Explain your answer with appropriate visuals.
5. Create a visual to identify the top 10 districts which have the highest number of pupils in 2018.

III. Stroke Prediction Dataset

“healthcare-dataset-stroke-data.csv” is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient. Using R, answer the following questions.

6. Load the data into R environment. Then, perform data pre-processing and data cleaning:
   a. State the data type of each variable (ignore variable “id”). Some of the variables are not defined with correct data type. Convert these variables such that they are with a suitable data type. Show the summary of the dataset.
   b. After variables are with their suitable data type. Based on the variables of gender and BMI, explain the issues of these variables that can affect the analysis.
   c. Based on the issues that you explained in 1 (b), propose solutions to address the issues.
7. What is the relationship between age and stroke? Also, what is the relationship between BMI and stroke. Explain your answer with appropriate visuals.
8. Visualize the number of patients with heart disease, hypertension, and stroke, based on gender.
9. Create a visual that can help you to identify the work types in which the total number of males is higher than the number of females. Explain your results.
10. Create an analysis to visualize the relationship between average glucose level and age.

IV. Submission:

This is a group assignment (a group of three members). The member grouping will be done via the eLearn@USM.

You are required to submit a zip/rar package which consists of the following items to the eLearn@USM:

- R script (in .R format).
- An assignment report not more than 8 pages (in pdf format). Only the sample output screen shots and relevant explanation/write-up/description are expected. Also, a cover page which contains your details must be included in your assignment report.

The zip/rar package must be named according to the following notation: CPC351_CPM351_[Matric]_A02. For example, for a group of three students with matric number of 112211, 112222, and 112233 respectively, they must name the zip/rar package as CPC351_CPM351_112211_112222_112233_ A02.

One of the group members is required to submit the zip/rar package. Kindly communicate with your group member before the submission to avoid any miscommunication.

The submission deadline 09 January 2022 (Sunday), 23:59 p.m. Failure to submit the assignment will be a disadvantage to you.

Reference: Kindly state any source of reference in your assignment script should you refer to various sources to complete this assignment.

IMPORTANT: Students who copied or plagiarized other’s work or let their work be copied or plagiarized will be given an F grade. The student may be barred from sitting for final exam and reported to the university’s disciplinary board.

V. Grading Rubric

This assignment will be graded according the grading rubric as shown in Table 1. The total will be scaled to 8% of your overall grade.

Table 1: Assignment 02 grading rubric.

	Good (3)	Satisfactory (2)	Poor (1)	Fail (0)
Question 1 (10%)	• Meet all the requirements and contain all the required visuals. The requirements are as follows: ○ The choice of visual type. ○ Correctness of information display. ○ Visual title, colour scheme, visual legend, axis labels, and measurement units. • The R program can be executed, and correct outputs are shown. • Clear and detailed comments are added to scripts with excellent clarity. • The report includes the screen shots, and explains the results with excellent clarity, comprehensiveness and organization. The description is supported by the visuals created or additional visuals • Discussion are well focused and all important points are included.	• Partially meet the requirements. • The R program can be executed, and partially correct outputs are shown. The requirements are as follows: ○ The choice of visual type. ○ Correctness of information display. ○ Visual title, colour scheme, visual legend, axis labels, and measurement units. • Adequate comments are added to scripts with satisfactory clarity. • The report includes the screen shots, and explains the results with satisfactory clarity, comprehensiveness and organization. The description is partially supported by the visuals created or additional visuals. • Discussion are not comprehensive and it misses some important points.	• Fail to meet the requirements and incorrect outputs are shown. The requirements are as follows: ○ The choice of visual type. ○ Correctness of information display. ○ Visual title, colour scheme, visual legend, axis labels, and measurement units. • The R program cannot be executed, and incorrect outputs are shown • Minimal or no comments is added to the scripts. • The report includes the screen shots, and unclearly or loosely explains the results. The description is not supported by any visuals • Discussion is not well focused and it misses the important points.	• No submission or late submission.
Question 2 (10%)				
Question 3 (10%)				
Question 4 (10%)				
Question 5 (10%)				
Question 6 (10%)				
Question 7 (10%)				
Question 8 (10%)				
Question 9 (10%)				
Question 10 (10%)				

~~END OF ASSIGNMENT 02~~