School of Computer Sciences

CPC351/CPM351 Principles of Data Analytics

Academic Session: Semester 1, 2022/2023

Assignment 02 – Data Exploration and Visualization

I. Dataset

Download the following files from eLearn@USM:

- "StudentsPerformance.csv"
- "SuperStoreOrders.csv"

II. Student Performance Dataset

"StudentsPerformance.csv" is a performance record for three courses in a university. 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 may not be 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) The variable "lunch" is not an important variable and can be omitted. Remove this variable column instances from the dataset.
   d) Rename the variables name to "STU_Gender", "STU_Ethnic", "PAR_Education", "PREP_CourseStatus", "Maths", "Reading", "writing".
2. Create a pie chart to show the distribution of students according to their ethnicity. and 2018. Then, create for each ethnicity, a bar chart to show distribution of gender in each ethnic. Perform the same steps for the preparation course status and parental level of education.
3. Using appropriate visualization chart, show the distribution of marks for the three courses. Compare the performance of students from different ethnicity in each of the courses. Explain your answer with appropriate visuals.
4. Does parental level of education affect students' performance in general context? Explain your answer with appropriate visuals.
5. Does parental level of education affect students' performance in specific context (based on ethnicity)? Explain your answer with appropriate visuals.
6. How do students course preparation status affect their performance? Explain your answer with appropriate visuals.
7. Other than the above, show two more trends / patterns / relationship that can be deduce from the data. Explain your answer with appropriate visuals.

III. Super Store Order

"SuperStoresOrder.csv" is sample Dataset includes data for the Sales of multiple products sold by the store along with subsequent information related to geography, Product categories, and subcategories, sales, and profits, segmentation amongst the consumers, etc. Using R, answer the following questions.

8. Load the data into R environment. Then, perform data pre-processing and data cleaning:
    a. State the data type of each variable. 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. Maintain only the top 1000 instances and remove the others.
9. From the reduced dataset, using an appropriate visualization:
    a. Group customer based on segments and determine the category distribution for each.
    b. Find the Top 10 categories.
    c. For each category determine the 3 most frequent-bought sub-categories.
    d. Where does most customer come from? Highlight the Top10 countries.
10. Other than the above, show four more trends / patterns / relationship that can be deduce from the data. Explain your answer with appropriate visuals.

IV. Submission:

This is a group assignment (a group of three members). The member grouping is as in assignment 1.

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 **08 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.

<table>
<tr><th></th><th>Good
(3)</th><th>Satisfactory
(2)</th><th>Poor
(1)</th><th>Fail
(0)</th></tr>
<tr><td>Question 1
(10%)</td><td rowspan="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.</td><td rowspan="10">• 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.</td><td rowspan="10">• 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.</td><td rowspan="10">• No submission or late submission.</td></tr>
<tr><td>Question 2
(10%)</td></tr>
<tr><td>Question 3
(10%)</td></tr>
<tr><td>Question 4
(10%)</td></tr>
<tr><td>Question 5
(10%)</td></tr>
<tr><td>Question 6
(10%)</td></tr>
<tr><td>Question 7
(10%)</td></tr>
<tr><td>Question 8
(10%)</td></tr>
<tr><td>Question 9
(10%)</td></tr>
<tr><td>Question 10
(10%)</td></tr>
</table>

~~END OF ASSIGNMENT 02~~