CP2403/CP3413: Assignment – Part 2 – 25%
Data Management, Visualization & Data Analysis
Due: Friday of Week 11, 11.59pm (Study Vacation Week)

In Assignment Part 2, you are required to apply appropriate data management, data visualization and data analytic techniques for a given scenario. The techniques required to complete this assignment are covered in Module 1 to 10 of the subject. You will have to explain what conclusions you draw after completing the different data analytics.

Scenario
The California Cooperative Oceanic Fisheries Investigations (CalCOFI) was formed in 1949 to study the ecological aspects of the sardine population collapse off California. CalCOFI conducts quarterly cruises off southern & central California, collecting a suite of hydrographic and biological data on station and underway. The CalCOFI data set represents the longest (1949-present) and most complete (more than 50,000 sampling stations) time series of oceanographic in the world.

The physical, chemical, and biological data collected at regular time and space intervals quickly became valuable for documenting climatic cycles in the California Current and a range of biological responses to them. Data collected at depths down to 500 m include: temperature, salinity, oxygen, phosphate, silicate, nitrate and nitrite, chlorophyll, transmissometer, PAR and C14 primary productivity.

You are provided with the following:-
1. bottle.csv
2. CalCOFI Database Tables Description - Bottle Table.pdf

The following website provides relevant information and documents about the CalCOFI project and data source.
https://calcofi.org/data/oceanographic-data/bottle-database/

Using the dataset and codebook provided, complete the following tasks.

Note:
As this data set contains a big number of sample records, you may be required to apply some pre-processing on the original data to extract less-sized but more valid sample records for some tasks (depending on what variable you would select for each task). It would be your choice which method you will use (or not use) to reduce the sample size, but it should be logical and must result to generate valid reduced samples for further processing of each task. For example, reducing the size of the data by simple cutting-off or random sampling with no supporting reason would not be acceptable.

[Task 1]
Select one categorical variable and one quantitative variable from the dataset to perform ANOVA analysis. What conclusion can you draw from the ANOVA analysis?

(Note: for the selection of a categorical variable, you can either select an existing categorical variable directly or generate a new categorical variable by transforming an existing non-categorical variable).

Hint: Refer to Module 5 and Practical 5 for help

[Task 2]
Select two categorical variables from the dataset to perform Chi-Squared Test. What conclusion can you draw from the Chi-Squared Test?

(Note: for the selection of a categorical variable, you can either select an existing categorical variable directly or generate a new categorical variable by transforming an existing non-categorical variable.)

(Note: for this task, be careful not to select (or generate) a categorical variable having more than ten categories. Having too many categories may cause the post-hoc test (if necessary) not to make meaningful results.)

Hint: Refer to Module 6 and Practical 6 for help

[Task 3]

Select two quantitative variables from the dataset to perform linear regression. What conclusion can you draw from the linear regression analysis?

Hint: Refer to Module 7 and Practical 7 for help

[Task 4]

Select four quantitative variables (one respond variable and three explanatory variables) from the dataset to perform multiple regression. What conclusion can you draw from the regression analysis?

For this task, you have to perform pre-testing by performing a number of different multiple regression models by selecting various combinations of individual regression functions. You are required to demonstrate at least three different combinations to compose the multiple regression model. Finally select one best multiple regression model based on your justification if possible.

Hint: Refer to Module 7 and 9 and Practical 7 and 9 for help

Submission

Ensure you complete, and submit all the files below to LearnJCU. Ensure you add your FirstName and LastName inside the files and to the file names.

- **Ass-Part2-Task1-ANOVA - FirstNameLastName.docx**
- **Ass-Part2-Task1-ANOVA-FirstNameLastName.ipynb**

- **Ass-Part2-Task2-Chi_Squared - FirstNameLastName.docx**
- **Ass-Part2-Task2-Chi_Squared-FirstNameLastName.ipynb**

- **Ass-Part2-Task3-Linear_Regression - FirstNameLastName.docx**
- **Ass-Part2-Task3-Linear_Regression-FirstNameLastName.ipynb**

- **Ass-Part2-Task4-Multiple_Regression - FirstNameLastName.docx**
- **Ass-Part2-Task4-Multiple_Regression-FirstNameLastName.ipynb**

Assignment – Part 2 (25%) Marking Criteria (Rubric) – Total Raw Marks: 100

Marking Criteria details are provided on a separate document.