CP2403: Project – Part 2 – 30%

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

This project consists of **Part A** and **Part B**.

Part A:

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
   (You can also access it via
   https://new.data.calcofi.org/index.php/database/calcofidatabase/bottle-field-descriptions)

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.

1. Select a categorical variable and a quantitative variable from the dataset to perform ANOVA analysis. What is 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

2. Select two categorical variables from the dataset to perform Chi-Squared Test. What is 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

3. Select three or more variables from the dataset to perform multiple regression. What is conclusion can you draw from the regression analysis?

   Hint: Refer to Modules 7-9 and Practicals 7-9 for help

Part B
You are provided with the Iowa Lottery Weekly Sales by Game Type Dataset (sales.csv). Select one game type and perform time series analysis (ARIMA). What conclusion can you draw from the ARIM analysis?
Hint: Refer to Module 10 and Practicals 10 for help

Submission

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

- '**CP2403 - Project – Part 2 – ANOVA - FirstNameLastName.docx**'
- '**Project-Part2-ANOVA- FirstNameLastName.ipynb**'

- '**CP2403 - Project – Part 2 – Chi_Squared - FirstNameLastName.docx**'
- '**Project-Part2- Chi_Squared- FirstNameLastName.ipynb**'

- '**CP2403 - Project – Part 2 – Regression - FirstNameLastName.docx**'
- '**Project-Part2-Regression- FirstNameLastName.ipynb**'

- '**CP2403 - Project – Part 2 – TS - FirstNameLastName.docx**'
- '**Project-Part2-TS- FirstNameLastName.ipynb**'

Project – Part 2 (30%) Rubric

Criteria	Exemplary (9, 10)	Good (7, 8)	Satisfactory (5, 6)	Limited (2, 3, 4)	Very Limited (0, 1)
ANOVA	Applied excellent ANOVA analysis Excellent interpretation of ANOVA analysis	Exhibits aspects of exemplary (left) and satisfactory (right)	Applied satisfactory ANOVA analysis Satisfactory interpretation of ANOVA analysis	Exhibits aspects of satisfactory (left) and very limited (right)	Applied limited or no ANOVA analysis Limited or no interpretation of ANOVA analysis
Chi Squared	Applied excellent Chi Squared analysis Excellent interpretation of Chi Squared analysis		Applied satisfactory Chi Squared analysis Satisfactory interpretation of Chi Squared analysis		Applied limited or no Chi Squared analysis Limited or no interpretation of Chi Squared analysis
Regression (Worth Double)	Applied excellent regression techniques (linear, multiple, polynomial) Excellent interpretation of regression analysis		Applied satisfactory regression techniques (linear, multiple, polynomial) Satisfactory interpretation of regression analysis		Applied limited or no regression techniques (linear, multiple, polynomial) Limited or no interpretation of regression analysis
Regression Model Validation	Created appropriate regression model validation graphs and excellent interpretation of validation graphs		Created appropriate regression model validation graphs and satisfactory interpretation of validation graphs		Created no regression model validation graphs and no interpretation of validation graphs

Time Series Analysis	Excellent time series analysis, applying all the steps in time series analysis Excellent interpretation of time series analysis		Satisfactory time series analysis, applying some steps in time series analysis Satisfactory interpretation of time series analysis		Limited or no time series analysis, applying non or limited steps in time series analysis Limited or no interpretation of time series analysis