EE3801 Data Engineering

Laboratory Exercise (LAB-II)

Assignment release date: Sept 09, 2020 **Briefing: Sept 10, 2020**

Date submission due: Sept 16, 2020

Grading: Your ASSIGNMENT will be graded out of 100 marks and the final weight of this assignment is 10%. The guidelines explained in the template will also carry marks. So please adhere to the guidelines.

NOTE: This is an individual lab exercise and NO DISCUSSIONS AND EXCHANGE OF SOLUTION IDEAS BETWEEN ANY STUDENTS ARE ALLOWED. If we come to know of this in any form, we will be taking relevant disciplinary actions. Please follow this guideline strictly.

Concepts used: Reading and writing data from/to csv files, generating new data frames, data wrangling - extraction, handling missing data, transforming a DF to a target DF, results visualization by plotting (different types – bar charts, pie charts, etc), and interpretation of results.

Data required for this assignment: FAO.csv

Consider the food production dataset (***FAO.csv***) given to you. If you encounter issues reading the file, consider using: pd.read_csv(filename, encoding= 'unicode_escape')

Nomenclature: Dataframe – **DF**

Read the questions carefully. Answer all the following.

1) Compute the following. [22 Marks]
   (a) Total production by each country in every year *(e.g., collapsing across food items)*. Store your result **as a new DF**. Also, save this result **as a CSV file**.

(b) Overall production by a country (**OPC**), defined as the total of all the values in all the years *(eg., one value per country)*,

(c) Average Production per year by each country (APPYPC)[1] (See the footnote for the definition & use)

(d) Percentage of production by each country compared to the overall global production, referred to as, Global Average Production by each country (GAPPC)[2] (See the footnote for the definition & use)

For Part (b) to (d), store your result as a DF and display only the top 3 and bottom 3 results from this new DF.

Country	OPC	APPYPC	GAPPC
Afghanistan	12123	xxxxx	xxxxx
Albania			
. . .			

2) Using the DF or your newly created CSV file in Question (1), **plot a Bar Chart** capturing APPYPC. Label your x and y axes appropriately and give your bar chart a meaningful name with clear legends. Identify the countries that show the lowest and highest average productions. Display the bar chart **in the way you think is best**. [10 Marks]
3) Using the extracted data frame for the Question (1) above, create a **Pie-Chart** for capturing GAPPC. Your Pie-Chart must also capture the percentage and the country name **in the way you think is best to display**. [10 Marks]
4) Using the extracted data frame for the Question (1) above, create a **Pie-Chart** for capturing GAPPC for those whose average production is more than 5% and capture the rest under "Others" category. Your Pie-Chart must also capture the percentage and the country name **in the way you think is best to display**. [10 Marks]

[1] APPYPC = Overall production by a specific country / total number of years;
[2] GAPPC = Overall production by a specific country / Overall production by all the countries

5) Consider the honey production across all countries. Answer the following.

(a) Filter/extract the honey production data as a dataframe between the years 2010 and 2013 (both years inclusive) from your given dataset. Note that if a country does not produce honey, ignore that country. Print the top 3 and bottom 3 results from this honey DF. [5 Marks]

(b) Compute the total honey production between the years 2010 and 2013 from your honey DF for each country as a *Pandas series*. Compute the overall honey production globally. Append the above Pandas series to your honey dataframe and store as a csv file. Print the top 3 rows and bottom 3 rows results from this honey DF. [5 Marks]

(c) Create **a Pie-Chart** to capture the countries whose average honey production constitutes more than 5% of global honey production between the years 2010 and 2013. Capture the rest of the countries under "Others" category in your Pie-chart. Identify the country that has highest honey production from this Pie-Chart.

Your Pie-Chart must capture the percentage and the country name as legends in the way you think is best to display. [10 Marks]

6) Consider the **Sugar and allied sugar products** produced by Malaysia and France between the years 2010 and 2013. [22 Marks]

(a) [5 Marks] Create two new dataframes, one for each country that captures the sum of all the combined "sugar" products in each year between 2010 and 2013 (any product that has "sugar" string in it. Note that items with the same name but different feed/food classification **should be classified as separate items**). This means that you need to take the sum of all sugar-based products in the dataset given to you for two classes of sugar (Food and Feed). Note that your final extracted two data frames (one for each country) should look like:

Country	Food Type	2010	2011	2012	2013
Malaysia	Food				
Malaysia	Feed				

Country	Food Type	2010	2011	2012	2013
France	Food				
France	Feed				

(b) Merge the two data frames into one DF and display. [2 Marks]

(c) Plot a graph for each country (in the same canvas) that shows the trend over the years on the total sugar and allied sugar products produced. [5 Marks]

(d) Use the above data generated in the two DFs in Part (a) to plot 4 sub-plots (in one canvas) showing Bar-Charts for each country on the two types of food in each year (2010-2013). Decide on the way you wish to plot. Show the x and y labels and legends as you see fit. [5 marks]

(e) Describe and interpret the trends exhibited. [3 Marks]

(f) What is the percentage difference between the countries in the year 2012? Display this information with a meaningful message. [2 Marks]

- Overall results presentation of the charts and display messages: 4 Marks
- Comments towards proper documentation of your code: 2 Marks