3031_A5_Complete_Template_v0

November 4, 2021

1 Assignment 5

Instructions:

a. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and not as a group activity.

b. Do not write your name on the assignment. (1 point)

c. Please include each question (or question number) followed by code and your answer (if applicable). Write your code in the 'Code' cells and your answer in the 'Markdown' cells of the Jupyter notebook. Ensure that the solution is written neatly enough to understand and grade. (1/2 point value of each question)

d. Export your Jupyter notebook as a PDF file. If you get an error, make sure you have downloaded the MikTex software (for windows) or MacTex (for mac). Note that after installing MikTex/MacTex, you will need to check for updates, install the updates if needed, and re-start your system. Submit the PDF file. (1 point)

This assignment is due at 11:59pm on Wednesday, November 10th. Good luck! *(30 points overall – 28 points for code & answers, 2 points for anonymity and proper formatting)*

1.1 Part 1

(5 points total)

Read FIFA world cup attendance data from the page: https://en.wikipedia.org/wiki/FIFA_World_Cup . Use 'attendance' as the matching string to find the table.

(a) Find the number of levels of column labels and row labels in the data *(2 points for code)*

(b) Reduce the multiple levels of column labels to a single level as follows. If the column names at all the levels are different, then concatenate the names together. Otherwise, keep the name at the highest level. For example, if the column name is ('Hosts','Hosts'), it should change to 'Host'. If the column name is ('Highest attendances †','Number'), it should change to 'Highest attendances †Number'. Do not rename each column manually. Use a method that will work efficiently if there were a large number of columns, say 10,000 columns. *(3 points for code)*

1.2 Part 2

(7 points total)

(a): Read the GDP per capita data from https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_

Perform the following operations on this datatable:

(i) Drop all the columns except Country and GDP per capita estimate by IMF. *(1 point for code)*

(ii) The country names contain some special characters (characters other than letters) and need to be cleaned. The following code can help clean country names:

```
import re
```

```
f = lambda x: re.sub(r'[^A-Za-z]', '', x)
```

Apply the above lambda function on the country column to clean country names. *(1 point for code)*

(b) Read the population data from https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_N
Drop all columns except Country and Population (1 July 2019). *(1 point for code)*

(c) Merge the datasets obtained in (a) and (b) such that the merged dataset contains each observation of the GDP per capita data (dataset obtained in (a)), but not necessarily each observation of the population data (dataset obtained in (b)). *(2 points for code)*

(d) For how many countries in the GDP per capita data was the population not available in the population data? Note that you don't need to clean country names in the population table. *(1 point for code, 1 point for answer)*

1.3 Part 3

(16 points total)

The dataset *Real GDP.csv* contains the GDP of each US State for all years starting from 1997 until 2020. The data is at State level, i.e., each observation corresponds to a unique State.

The dataset *Surplus.csv* contains the surplus of each US State for all years starting from 1997 until 2020. The data is at year level, i.e., each observation corresponds to a unique year.

The dataset *Compensation.csv* contains 'Compensation' and 'Chain-type quantity indexes for real GDP' for each US State and year starting from 1997 to 2020. The dataset is at Year-State-Description level, where 'Description' refers to either 'Compensation' or 'Chain-type quantity indexes for real GDP'.

Q1) Combine all these datasets to obtain a dataset at State-Year level, that contains the GDP, surplus, 'Compensation', and 'Chain-type quantity indexes for real GDP' for each US State and all years starting from 1997 until 2020. *Note that each observation must contain the name of the US State, year, and the four values (GDP, surplus, compensation, and Chain-type quantity indexes for real GDP).*

Hint: Here is one way to do it: 1) Melt the GDP dataset to year-State level
2) Melt the Surplus dataset to year-State level
3) Pivot the compensation dataset to year-State level

Now that all the datasets are at the year-State level, merge them!

(4 points for code)

Q2) Use a single plot to answer all three questions below by visualizing:
(a) How does the mean GDP (mean over all States) change with year? *(1 point for visualization)*
(b) How does the mean compensation (mean over all States) change with year? *(1 point for visualization)*
(c) How does the mean surplus (mean over all States) change with year? *(1 point for visualization)*

Also show the 95% confidence interval for the mean GDP, mean compensation, and mean surplus in the plot.

Hint: Use the *seaborn* function *lineplot()* . No calculations are needed. Just use *lineplot()* three times.

Q3) The mean GDP (over all States) seems to have decreased in 2020 as compared to 2019 *(you know why!)*. How many States observed a decrease in GDP in 2020 (as compared to 2019)? For which States did the GDP increase in 2020 (as compared to 2019)? *(2 points for code, 2 points for answers)*

Q4) Merge the file *State_region_mapping.csv* with the dataset obtained in Q1. Make a lineplot showing the mean GDP for each of the five regions with year. Do not display the confidence interval. Which two regions seems to have the least growth in GDP over the past 24 years? *(2 points for code, 1 point for answer)*

Q5) Identify the States contributing the most to the total GDP of their region, in 2020. Also, find the percentage contribution of these States to the total GDP of their region. *(2 points for code)*

Hint: You may use *DataFrameGroupBy.idxmax()* for the first part of this question.