Assignment 1

The assignment is based on three datasets : exposure.csv, Countries.csv, Countries-Continents.csv (https://drive.google.com/drive/folders/1XeG7NP94bdbFmYUUdPwply_AWI5DBoQs?usp=sharing) these datasets are from Kaggle and we are not responsible for the political correctness of the data as this is just for the purpose of the exercise of working with data not to learm about geography.

You are supposed to **explore** these datasets and answer the following questions

---

- **Question 1 (1 Mark)**

  Perform the following preprocessing steps:

  - Drop all rows from the "exposure" dataset without country name
  - Join the two datasets (exposure.csv and Countires.csv) based on the "country" columns in the datasets, keeping the rows as long as there is a match between the country columns of both dataset (do not concatenate the datasets)
    You must ensure the countries with name issues match (e.g., USA and United States) but ignore if a country does not exist in either of datasets (e.g., Sudan and "South Sudan" are not the same)
  - Keep only a single country column
  - Set the index of the resultant dataframe as 'Country'
  - Sort the dataset by the index (ascending)

- **Question 2: (based on the dataframe created in Question-1) (1 Mark)**
  The "Cities" column is a complex string, containing information about cities (e.g., latitude and longitude) of the corresponding country; you should **explore** the content of this column for each country and ***add two new columns*** to the dataframe called: avg_latitude and avg_longitude ."avg_latitude" is the average latitude for all cities of the corresponding country, and "avg_longitude" is the average longitude for all cities of the same country.

- **Question 3: (based on the dataframe created in Question-2) (1 Mark)**
  Given that the first case of COVID has been found in Wuhan with the following coordinates: (30.5928° N, 114.3055° E), sort the dataframe based on how close they are to Wuhan . You can use "avg_latitude" and "avg_longitude"; the countries close to Wuhan should be ranked first in the final dataset. The final dataset should also contain ***a new column*** called distance_to_Wuhan , showing the distance to Wuhan in km.
  Update: You can assume the earth's radios is R = 6373.

- **Question 4: (based on the dataframe created in Question-2) (1 Mark)**
  Using the continent dataset **(Countries-Continets.csv)** calculate the average covid_19_Economic_exposure_index for each continent. The output should be a dataframe with two columns: " Continent ", " average_covid_19_Economic_exposure_index ". Rank the continents based on average "Covid_19_Economic_exposure_index" (ascending), with "Continent" being the index.

- **Question 5: (based on the dataframe created in Question-2) (1 Mark)**
  What is the the average "Foreign direct investment" and "Net_ODA_received_perc_of_GNI" for each income class (e.g., HIC, MIC, LIC). The output should be a dataframe with three columns: " Income Class ", " Avg Foreign direct investment ", " Avg_ Net_ODA_received_perc_of_GNI ", with "Income Class" being the index. If a country has no value for the given column, ignore the row. UPDATE: the column name changed by adding "Avg"

- **Question 6: (based on the dataframe created in Question-2) (1 Marks)**
  List top 5 most **Populous** cities located in Low Income Countries; ignore cities without population information. The output is a **python list** , containing the names of the cities.

- **Question 7: (based on the dataframe created in Question-2) (2 Marks)**
  Find cities which are located in different countries but have the same name. The result dataset should contain 2 columns:
  " city ", " countries ", with city being also the index. The result dataset should not have duplicate records. For instance, either "Melbourne, {Florida, Australia}" or "Melbourne, {Australia, Florida}" should be present in the final dataframe, not both. The countries column is a list of countries separated by comma (',').

- **Question 8: (based on the dataframe created in Question-2) (2 Marks)**
  - In a visualization show what percentage of the world population is living in each South American country. You can use the continent dataset **(Countries-Continets.csv)** to answer this question.
  - Choose an appropriate visualization type, presenting the requested information in the best way (check the lecture for Data Visualization about selecting the right paradigm)
  - Plot human-readable visualization; it should be self-explanatory and its elements (e.g., labels, legends) must be clear.

- **Question 9 : (based on the dataframe created in Question-2) (2 Marks)**
  - Plot a visualization to compare the high, middle, and low income level countries based on the following metrics: Covid_19_Economic_exposure_index_Ex_aid_and_FDI
  Covid_19_Economic_exposure_index_Ex_aid_and_FDI_and_food_import
  Foreign direct investment, net inflows percent of GDP
  Foreign direct investment
  - Choose an appropriate visualization type, presenting the requested information in the best way.
  - Plot human-readable visualization; it should be self-explanatory and its elements (e.g., labels, legends) must be clear.

- **Question 10: (based on the dataframe created in Question-2) (3 Marks)**
  - Plot a scatter plot with y axis being "avg_latitude" and x axis being "avg_longitude". Each point in this plot indicates a labelled country.
  - Ink each country (e.g., red, green) based on its continent (e.g, Asia, Africa). You can pick any colour for each continent. You can use the continent dataset **(Countries-Continets.csv)** to answer this question.
  - The size of each point must represent the population of its country. For example, the points representing China and India should be bigger than that of Australia.
  - Add a legend showing the name of continents and their associated colours.
  - Plot human-readable visualization; it should be self-explanatory and its elements (e.g., labels, legends) must be clear.
  UPDATE: page last update on 17 March
  - The code template is updated on 17 March, 10:15 AM [renmaing the lst to cities_lst]
  - You should not overwrite any CSV files; keep your changes in Dataframes
- The graphs should not pop up. The files will be marked
- Please read all highlighted text

What not to forget!

Due Date: Thursday the 18th of March 2021 19:59

Submit your script named " YOUR_ZID .py" (z2123232.py) which contains your code.
You are required to use the following code template (it is not complete; please download the file) for your submission.

- You can download the code template from : (https://github.com/mysilver/COMP9321-Data-Services/blob/master/z1111111.py) https://github.com/mysilver/COMP9321-Data-Services... (https://github.com/mysilver/COMP9321-Data-Services/blob/master/z1111111.py)

```
import json
import matplotlib.pyplot as plt
import pandas as pd
import sys
import os
import numpy as np
import math
studentid = os.path.basename(sys.modules[__name__].__file__)
def log(question, output_df, other):
    print("--------------- {}----------------".format(question))
    if other is not None:
        print(question, other)
    if output_df is not None:
        df = output_df.head(5).copy(True)
        for c in df.columns:
            df[c] = df[c].apply(lambda a: a[:20] if isinstance(a, str) else a)
        df.columns = [a[:10] + "..." for a in df.columns]
        print(df.to_string())
def question_1(exposure, countries):
    """
    :param exposure: the path for the exposure.csv file
    :param countries: the path for the Countries.csv file
    :return: df1
            Data Type: Dataframe
            Please read the assignment specs to know how to create the output dataframe
    """
    ################################################
    # Your code goes here ...
    ################################################
    log("QUESTION 1", output_df=df1, other=df1.shape)
    return df1
...
if __name__ == "__main__":
    df1 = question_1("exposure.csv", "Countries.csv")
    df2 = question_2(df1.copy(True))
    df3 = question_3(df2.copy(True))
    df4 = question_4(df2.copy(True), "Countries-Continents.csv")
    df5 = question_5(df2.copy(True))
    lst = question_6(df2.copy(True))
    df7 = question_7(df2.copy(True))
    question_8(df2.copy(True), "Countries-Continents.csv")
    question_9(df2.copy(True))
    question_10(df2.copy(True), "Countries-Continents.csv")
```

- If you do not follow this structure, you will not be marked.
- You can only add codes in the specified lines (do not edit the rest of the lines):

```
################################################
# Your code goes here ...
################################################
```

- If your code does not run on CSE machines for any reasons (e.g., hard-coded file path such as C://Users/), you will be penalized at least by 5 marks. We assume that the csv files are located in the same directory of your script, and the name is the same as the one in the template (e.g., exposure.csv and Countires.csv)
- Please look at the documentation for each question method; it describes the inputs (e.g., a dataframe) and output (e.g., dataframe, list of cities) of the method.

```
"""
:param df7: the dataframe created in question 7
:return: df8
          Data Type: Dataframe
          Please read the assignment specs to know how to create the output dataframe
"""
```

- Please use the same variable names as mentioned in the comments
- You are supposed to use Pandas library for all questions. That being said, it is forbidden to use regular python codes to process data. However, you can use lambda when required and user-defined functions for panda methods such as 'apply'.
- In the last three questions, you need to plot charts; please do not use "plt.show()" function to pop up charts. The code template will automatically save the chart on the disk. What you need to do is to just call the plot functions of the dataframe (e.g., df.plot.pie()). We highly recommend you go through the lab activities to know how to plot charts.
- You may need make sure country names in both dataframes match. If they do not match (e.g., USA and United States), you must apply appropriate pre-processing techniques to make sure country names match.
- You should not edit the dataset files. You can only submit your code.
- You cannot use other python libraries unless it is already listed in the template file
- Use the latest version of the python libraries

FAQ:

- **Can I pass extra variables to functions?**
  No
- **Can we create our own functions besides the question functions (e.g., question_1)?**
  Yes
- **Can I call another function inside the question functions? e.g., calling question_1 inside question_2**
  Yes
- **What should I do if my charts are not shown automatically?**
  Look at the lab sample codes; if still need a help, ask your tutor during the labs.
- **How are our submissions marked?**
  They are marked manually by tutors, by running the following command: python3 z{YOUR_ZID}.py
- **What python packages can I use in my assignment?**
  You can only use packages imported in the template file to do the assignment.
- **What version of python should I use?**
  Python 3+
- **What version of pandas should I use?**
  the latest version and you can update the version on you cse account to make sure you can test your code
- **How I can submit my assignment?**
  Go to the assignment page click on the "Make Submission" tab; pick your files which must be named

"YOUR_ZID.py". Make sure that the files are not empty, and submit the files together.

- **Can I submit my file after deadline?**
  Yes, you can. But 25% of your assignment will be deducted as a late penalty per day. In other words, if you be late for more than 3 days, you will not be marked.
- If the coordinates of two cities matches and they have the same name, they are duplicates [this is just a clarification not a spec change]
- If you have been asked to return a dataframe with n columns and set one of them as index; the total columns of the dataframe will be n-1
- Please do not change the file names (figures); keep the file names as it is.

Plagiarism

This is an *individual assignment* . The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this course. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted, you may be penalized, even if the work was submitted without your knowledge or consent. Pay attention that is **also your duty to protect your code artifacts** . if you are using any online solution to store your code artifacts (e.g., GitHub) then make sure to keep the repository private and do not share access to anyone.

Reminder: Plagiarism is defined as (https://student.unsw.edu.au/plagiarism) using the words or ideas of others and presenting them as your own. UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties as severe as being excluded from further study at UNSW. There are several on-line sources to help you understand what plagiarism is and how it is dealt with at UNSW:

- Plagiarism and Academic Integrity (https://student.unsw.edu.au/plagiarism)
- UNSW Plagiarism Procedure (https://www.gs.unsw.edu.au/policy/documents/plagiarismprocedure.pdf)

Make sure that you read and understand these. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible for ensuring that your assignment files are not accessible by anyone but you by setting the correct permissions in your CSE directory and code repository, if using one (e.g., Github and similar). Note also that plagiarism includes paying or asking another person to do a piece of work for you and then submitting it as your own work.

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW staff and students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW.

Resource created 21 days ago (Saturday 27 February 2021, 03:15:42 PM), last modified 3 days ago (Wednesday 17 March 2021, 10:22:25 PM).

Comments

Loading...