EE101 Class of year 2024 Final project

2024. 6. 05(revised)

1. 본인이 참조할 수 있는 모든 자료를 사용하여 수행 가능
2. 2명의 팀원이 서로 협력해서 수행할 것
3. 문항별 코드를 설명하는 Markdown 및 주석 작성
   - 작성하지 않거나 내용이 미흡할 시 감점 사유 (문제 별로 5~10 % 감점)

4. 제출 시에는 ipynb 파일과 더불어 본인이 사용한 Dataset도 함께 제출하기
   (Extra Credit 문제에서 자신이 사용한 New dataset)

6. Due date : 2024. 6. 12. 10:00PM
*** 1일 delay 30% 감점. 이후는 0점 처리됨**

A. Data Analysis using PANDAS & actual data (1)

* [4 Datasets] needed for this project

1. WHO-COVID-19-global-data.csv:
- 2020년 초부터 2024년 초까지 세계 각국의 COVID-19 감염자 및 사망자 수 추이

(8 columns: Date_reported, Country_code, Country, WHO_region, New_cases, Cumulative_cases, New_deaths, Cumulative_deaths)

2. country-capital-lat-long-population.csv:
- 국가 별 주요 도시의 위도 경도 (단위: degree) 정보

3. country-population.csv:
- 연도에 따른 국가 별 인구수 정보

4. list_of_countries_by_area.csv:
- 국가 별 면적 정보

A. Data Analysis using PANDAS & actual data (2)

* Prep1~4: Preparation works for data analysis (5점)

Prep1. Load data

# Load WHO-COVID-19-global-data.csv, country-capital-lat-long-population.csv, country-population.csv, list_of_countries_by_area.csv file as pandas.DataFrame

Prep2. View Data Configuration

# Print first 5 rows of data in 4 csv files

Prep3. Check for missing values in WHO-COVID-19-global-data.csv and fill them

# Check and fill with zeros where numerical values are supposed to be

# Non-numerical values should not be modified.

A. Data Analysis using PANDAS & actual data (3)

Prob1. Data analysis (45 점)

1) Extract row index value where New_cases <0 and delete it and delete the row index value where the number of New_deaths < 0 (cnt <0) in **WHO-COVID-19-global-data.csv**

2) Remove **countries** that do not exist in the **WHO-COVID-19-global-data.csv** file from the **country-capital-lat-long-population.csv**, **list_of_countries_by_area.csv**, & **country_population.csv**

3) Change the 'Alpha-2 code' from **list_of_countries_by_area.csv** and the 'Country Code' from **country_population.csv** to the 'Country_code' in **WHO-COVID-19-global-data**.**csv** where the '**Country / Dependency**' in **list_of_countries_by_area.csv** and '**Country Name**' in **country_population.csv** matches the '**Country**' in **WHO-COVID-19-global-data.csv**

4) Clear 'values including ()' fro m the 'Total in km (mi)', 'Land in km (mi)', and 'Water in km (mi)' inside **list_of_countries_by_area.csv**.

ex) 1,000,000 (500,000) -> 1,000,000

A. Data Analysis using PANDAS & actual data (4)

Prob1. Data analysis (continued)

5) Find countries where the annual number of **New_cases in 2021** is 50,000 or more, while simultaneously having a **population density** less than 1000 persons/km

('2021' in country-population.csv / 'Total in km (mi) in list_of_countries_by_area.csv)

6) Find the countries with the highest annual increase in New_cases in 2021 compared to 2020 by WHO region

7) Find the date with the largest decrease in weekly New_cases globally, compared to the previous day, over the entire period

8) Find the longest period during which the number of New_cases globally remained above 10,000 continuously

9) Divide the **population density based on the number of population density** into 10 sections and calculate the average cumulative number of confirmed cases in 2021 per section (based on the year 2021 only) – (hint: use pandas.qcut() method)

A. Data Analysis using PANDAS & actual data (5)

Prob2. Visualizing Data using Matplotlib (20 점)

- Ensure that the axis labels on the graph do not overlap and are clearly visible.

1. Graph showing the monthly cumulative changes in the total number of infections for the top 10 countries with the highest cumulative_cases (as of the most recent date).

2. Bar graph depicting the mortality rate (number of New_deaths/number of New_cases) for the top 10 countries with the highest mortality rates, as of the most recent date.

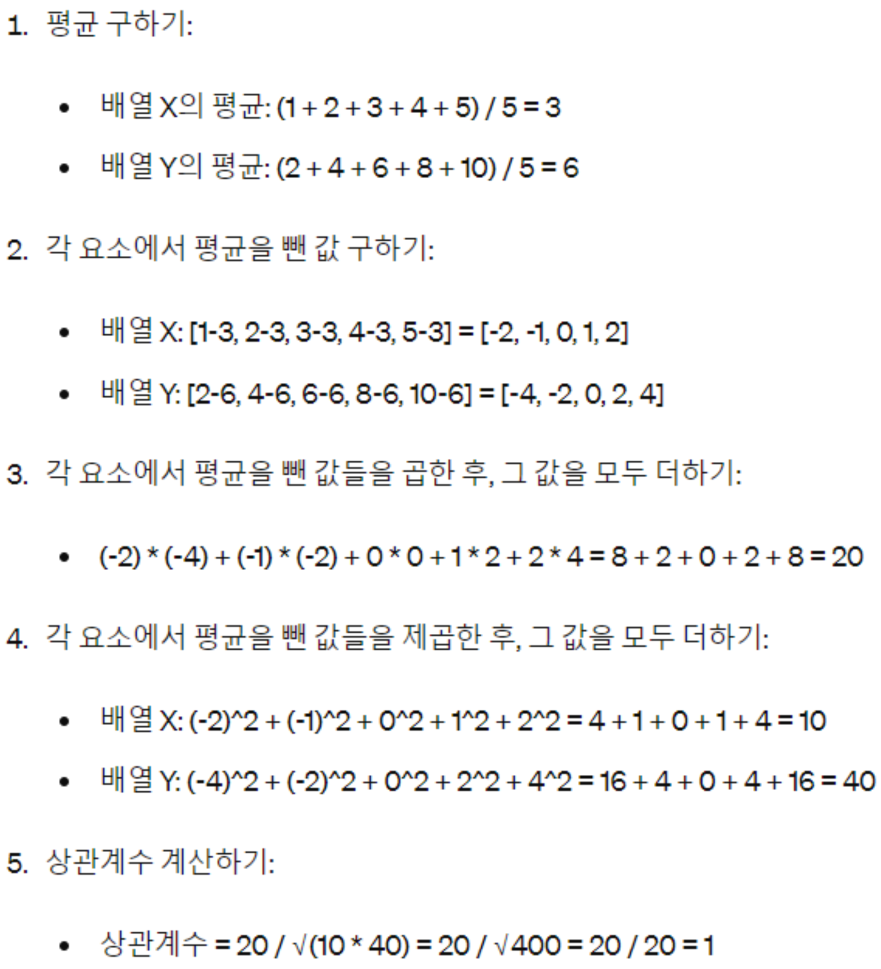
3. Create a bar graph showing **the correlation coefficients** of cumulative COVID-19 cases between the Democratic Republic of the Congo and the nearest four countries, as well as the furthest four countries within the 'AFRO' region. The countries should be ordered by proximity to the Democratic Republic of the Congo.

4. Perform the same task as 2-3 for the country of Spain in the ‘EURO’ region

A. Data Analysis using PANDAS & actual data (5)

HINT. Correlation coefficient의 의미 (두 데이터가 얼마나 강한 선형적 관계 y=ax+b를 가지는가)

Ex. 두 1차원 배열 (시계열 데이터) X: [1,2,3,4,5], Y: [2,4,6,8,10]이 있다고 할 때,



1. 평균 구하기:
   - 배열 X의 평균: (1 + 2 + 3 + 4 + 5) / 5 = 3
   - 배열 Y의 평균: (2 + 4 + 6 + 8 + 10) / 5 = 6
2. 각 요소에서 평균을 뺀 값 구하기:
   - 배열 X: [1-3, 2-3, 3-3, 4-3, 5-3] = [-2, -1, 0, 1, 2]
   - 배열 Y: [2-6, 4-6, 6-6, 8-6, 10-6] = [-4, -2, 0, 2, 4]
3. 각 요소에서 평균을 뺀 값들을 곱한 후, 그 값을 모두 더하기:
   - (-2) * (-4) + (-1) * (-2) + 0 * 0 + 1 * 2 + 2 * 4 = 8 + 2 + 0 + 2 + 8 = 20
4. 각 요소에서 평균을 뺀 값들을 제곱한 후, 그 값을 모두 더하기:
   - 배열 X: (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 4 + 1 + 0 + 1 + 4 = 10
   - 배열 Y: (-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2 = 16 + 4 + 0 + 4 + 16 = 40
5. 상관계수 계산하기:
   - 상관계수 = 20 / √(10 * 40) = 20 / √400 = 20 / 20 = 1

B. Object Oriented Programming Project (1)

Prob3. Object Oriented Programming Project (30 점)

1. Build ‘Country’ Class using the following information.

# columns to be imported from the CSV files:

- **Country, Country_code, WHO_region - WHO-COVID-19-global-data.csv**
- **Capital city - country-capital-lat-long-population.csv**
- **Land in km (mi), Water in km(mi) - list_of_countries_by_area.csv**
- **1960- country-population.csv**

__Init__ method : The arguments of the __init__ method should be received as **attributes**.

Attribute:

the attribute names should be consistent as mentioned below.
Creation of additional attributes is not allowed

- **name (str), country_code(str), who_region(str) - WHO-COVID-19-global-data.csv**
- **capital(str) - country-capital-lat-long-population.csv**
- **land_area(int), water_area(int) - list_of_countries_by_area.csv**
- **population(based on 1960 data) - country-population.csv**

B. Object Oriented Programming Project (1)

Prob3. Object Oriented Programming Project (30 점)

1. Build ‘Country’ Class using the following information.

# Other Methods in Class

def info: print the country's information using **dictionary**

def population_density: return **population density(str)** = population / total area

def update_population:

- use the annual population growth rate based on **country_population.csv** to predict and update the population for 2024

- The **annual population growth rate** should be calculated based on the **growth rates between consecutive years** from 1960 to 2022. The **average of these growth rates** will be used to predict the population growth rate for the year **2024**.

def compare: Among countries with the same **WHO_region**, return the **top 10 countries** with the **highest population density**.

B. Object Oriented Programming Project (2)

1. Build ‘Country’ Class(continue)

1) Create instances for countries in the dataset with WHO_region as EURO.

2) Use **compare** method to compare population density in EURO and rank them.

3) Use **update_population** to update the predicted population(2024).

4) Use **compare** method again to rank the countries in EURO.

B. Object Oriented Programming Project (3)

2. Build ‘Covid’ Class using the following information.

# Inheritance: Inherit from ‘Country’ class
# columns to be imported from the CSV files

- **Cumulative_cases per year (up to 2024.02.18)**
- **Cumulative_deaths per year (up to 2024.02.18)**
- **average weekly New_cases (up to 2024.02.18)**
- **average weekly New_deaths (up to 2024.02.18)**

# Init Methods: The arguments of the **__init__** method should be received as **attributes**.

Attribute: the attribute names should be consistent as mentioned below.

- **cumulative_cases**
- **cumulative_deaths**
- **avg_weekly_new_cases**
- **avg_weekly_new_deaths**

-

B. Object Oriented Programming Project (3)

2. Build ‘Covid’ Class using the following information.

# Other Methods

def info: print the Country’s information(inheritance) and the Covid's information using the **dictionary**

def mortality_rates: return mortality_rates

- Mortality rate = (cumulative deaths / cumulative cases) * 100 (%)

def infection_rates: return infection_rates

- Infection rate = (cumulative cases / total population) * 100 (%)

def composite_risk_index: return composite_risk_index

- Composite risk index = 0.4 * mortality rate + 0.6 * (population – cumulative cases) * infection rate

def rank_by_composite:

- Take the **first alphabet of the capital** as an argument and return the countries starting with that letter, sorted by their **composite risk index in ascending order**.

B. Object Oriented Programming (OOP) Project (4)

2. Build ‘Covid’ Class(continue)

1) Create instances for EURO countries in the dataset.

2) Use **update_population** to update the predicted population(2024).

3) Use **rank_by_composite** to return results for countries whose capitals start with the letter B.

B. Object Oriented Programming Project (5)

3.Estimation of Vaccine Efficacy method.

Given the new vaccine developed by Company A, estimate the efficacy based on the assumption that it would have been used since the start of the reporting period.

def calculate_vaccine_effect: find the greatest vaccine effect country

- **Estimated average weekly New_cases**

=Min(0.5 * existing average weekly New_cases + deviation score, existing average weekly New_cases)

- **Estimated average weekly New_deaths**

= Min(0.4 $*$ existing average weekly New_deaths + deviation score, existing average weekly New_deaths)

- Definition of the greatest vaccine effect:

the country where the value of

$$\frac{\text{existing average weekly New\_cases}}{\text{Estimated average weekly New\_cases}} + \frac{\text{existing average weekly New\_deaths}}{\text{Estimated average weekly New\_deaths}}$$

is the highest.

1) Find the country where WHO_region is AFRO with the greatest vaccine effect when the deviation score is 10.

2) Find the country with Land in Km greater than 1,000,000 that shows the greatest vaccine effect when the deviation score is 20.

C. Extra credit Problem (Max 20 점 추가)

4. Searching & Integrating New Datasets for Comprehensive Analysis

1. First, obtain at least one OECD statistical dataset.
2. Then integrate it with existing datasets.
3. Create a your own problem and solve it based on the data.

(Example problem: Estimating infection routes, Identifying significant factors for predicting next year's confirmed cases, etc.)

# You can use any type of dataset from https://stats.oecd.org/