MATH 118: Statistics and Probability **(Due: 26/04/21)**

Homework #1

Instructor: Dr. Zafeirakis Zafeirakopoulos *Name:* Esra Eryılmaz *Student Id:* 171044046
Assistant: Gizem Süngü *171044046*

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- It is not a group homework. Do not share your answers to anyone in any circumstance. Any cheating means at least -100 for both sides.
- Do not take any information from the Internet.
- No late homework will be accepted.
- For any questions about the homework, come to my office hour.
- After the office hour, no questions about the homework by email will be responded.
- Submit your homework (both your latex and pdf files in a zip file) into the course page of Moodle.
- Save your latex, pdf and zip files as "Name_Surname_StudentId".{tex, pdf, zip}.
- The deadline of the homework is 22/04/21 23:55.

Problem 1 (100 points)

Homework 1 considers a Covid-19 dataset which is published on Github. Please download any document type that you prefer of the dataset from the links which are shown in Figure 1. The dataset is updated daily and

pic.png

Figure 1: The complete dataset links

includes data on confirmed cases, deaths, hospitalizations, testing, and vaccinations as well as other variables of potential interest. The data set has the following basic columns:

- iso_code: Short name of the country
- continent: The continent where the country exists
- location: The country name
- date: The date when the data about various variables are taken.

You are responsible to implement a program which reads the given dataset from the file and computes the data for the following questions. Any programming language that you prefer will be accepted. Putting comments on your functions that you implement is must. Each question must be appended to a file which is called "output{.csv, .txt}". The file contains the first 18 questions listed below. The 18th question will be written in this document.

1. How many countries the dataset has?
2. When is the earliest date data are taken for a country? Which country is it?
3. How many cases are confirmed for each country so far? Print pairwise results of country and total cases.

Table 1: The format of the output for the questions 5, 6, 7, 8, 9, 10, 12, 13.

Country	minimum	maximum	average	variation
value	value	value	value	value

4. How many deaths are confirmed for each country so far? Print pairwise results of country and total deaths.
5. What are the average, minimum, maximum and variation values of the reproduction rates for each country?
6. What are the average, minimum, maximum and variation values of the icu_patients (intensive care unit patients) for each country?
7. What are the average, minimum, maximum and variation values of the hosp_patients (hospital patients) for each country?
8. What are the average, minimum, maximum and variation values of the weekly icu (intensive care unit) admissions for each country?
9. What are the average, minimum, maximum and variation values of the weekly hospital admissions for each country?
10. What are the average, minimum, maximum and variation values of new tests per day for each country?
11. How many tests are conducted in total for each country so far?
12. What are the average, minimum, maximum and variation values of the positive rates of the tests for each country?
13. What are the average, minimum, maximum and variation values of the tests per case for each country?
14. How many people are vaccinated by at least one dose in each country?
15. How many people are vaccinated fully in each country?
16. How many vaccinations are administered in each country so far?
17. List information about population, median age, # of people aged 65 older, # of people aged 70 older, economic performance, death rates due to heart disease, diabetes prevalence, # of female smokers, # of male smokers, handwashing facilities, hospital beds per thousand people, life expectancy and human development index.

Table 2: The format of the output for the question 17

Country	population	median age	# of people aged 65 older
value	value	value	value

18. Summarize all the results that you obtain by the first 17 questions (except question 2).

Table 3: The format of the output for the question 18

Country	q#3	q#4	q#5_min	q#5_max	q#5_avg	q#5_var
value	value	value	value	value	value	value

19. Comment the results based on your observations. Write your opinions about the reasons of increasing infection rates by giving examples from the results. Feel free to explain any situation that you observe. More observations more opportunities will bring you for the second homework.
(Solution)

- **-Hypothesis 1 :** Intensive care unit patients values causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Austria and Belgium which have the same situations in most of the variables.Country Belgium has higher corona death rates than country Austria. Country Belgium has also higher intensive care unit patients than country Austria.
- **-Hypothesis 2 :** Hospital patients values causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Denmark and Finland which have the same situations in most of the variables.Country Denmark has higher corona death rates than country Finland. Country Denmark has also higher hospital patients than country Finland.
- **-Hypothesis 3 :** Weekly intensive care unit admission values causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Cyprus and Estonia which have the same situations in most of the variables.Country Estonia has higher corona death rates than country Cyprus. Country Estonia has also higher weekly intensive care unit admission values than country Cyprus.
- **-Hypothesis 4 :** Weekly hospital admission values causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Belgium and Czechia which have the same situations in most of the variables.Country Czechia has higher corona death rates than country Belgium. Country Czechia has also higher weekly hospital admission values than country Belgium.
- **-Hypothesis 5 :** Higher middle age values causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Afghanistan and Poland which have the same situations in most of the variables.Country Poland has higher corona death rates than country Afghanistan. Country Poland has also higher middle age values than country Afghanistan.
- **-Hypothesis 6 :** New test values per day causes coronavirus cases.
  -How did you obtain the hypothesis : There are two countries Ghana and Peru which have the same situations in most of the variables.Country Peru has higher corona case rates than country Ghana. Country Peru has also higher new test values per day than country Ghana.
- **-Hypothesis 7 :** Positive rates of the test values causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Peru and Saudi Arabia which have the same situations in most of the variables.Country Peru has higher corona death rates than country Saudi Arabia. Country Peru has also higher positive rates of the test values than country Saudi Arabia.
- **-Hypothesis 8 :** Administered vaccination values decreases coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Turkey and Germany which have the same situations in most of the variables.Country Turkey has higher corona death rates than country Germany. Country Turkey has also lower administered vaccination values than country Germany.
- **-Hypothesis 9 :** Higher diabetes prevalence causes coronavirus deaths
  -How did you obtain the hypothesis : There are two countries Nepal and Madagascar which have the same situations in most of the variables.Country Nepal has higher corona death rates than country Madagascar. Country Nepal has also higher diabetes prevalence than country Madagascar.
- **-Hypothesis 10 :** # of people aged 70 older causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Thailand and France which have the same situations in most of the variables.Country France has higher corona death rates than country Thailand. Country France has also higher # of people aged 70 older than country Thailand.
- **-Hypothesis 11 :** Lower hospital bed values causes coronavirus deaths.
  -How did you obtain the hypothesis : There are two countries Sweden and Tajikistan which have the same situations in most of the variables.Country Sweden has higher corona death rates than country Tajikistan. Country Sweden has also lower hospital bed values than country Tajikistan.
- **-Hypothesis 12 :** Heart disease causes coronavirus deaths
  -How did you obtain the hypothesis : There are two countries Russia and Bangladesh which have the same situations in most of the variables.Country Russia has higher corona death rates than country Bangladesh. Country Russia has also higher cardiovasc death rate than country Bangladesh.
- **-Hypothesis 13 :** Lower handwashing facilities causes coronavirus cases.
  -How did you obtain the hypothesis : There are two countries Mali and Malawi which have the same situations in most of the variables.Country Malawi has higher corona case rates than country Mali. Country Malawi has also lower handwashing facilities than country Mali.

- **-Hypothesis 14 :** # of female smokers and # of male smokers rates causes coronavirus deaths
  -How did you obtain the hypothesis : There are two countries Ukraine and Uganda which have the same situations in most of the variables.Country Ukraine has higher corona death rates than country Uganda. Country Ukraine has also higher # of female smokers and # of male smokers rates than country Uganda.

Some notes:
→ Nan values in output.txt indicates that there is no value in excel.
→ I did not put the 18th question in output.txt because I have already printed each problem one by one.