| **MATH 118: Statistics and Probability** | | **(Due: 26/04/21)** |
|---|---|---|
| | **Homework #1** | |
| *Instructor:* Dr. Zafeirakis Zafeirakopoulos | *Name:* | *Student Id:* |
| *Assistant: Gizem Süngü* | | |

**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
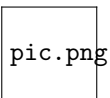
Figure 1: The complete dataset links

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)* (Write your observations here.)

Solution:

- Hypothesis : Even some countries's new cases or deaths become 0 sometime, new cases always come up again .
- How did you obtain the hypothesis: I checked many successful countries like Israel and New Zealand, sometimes there aren't new cases for days/weeks but then new Corona virus cases always come up.

- Hypothesis : If vaccinated people are not the most of the population, new cases are not shrinking that much.
- How did you obtain the hypothesis: I checked some countries like Vietnam and Taiwan, I thought when vaccinations begin, new cases will be reduce but new cases increases. And when I check population, vaccinated people are not most of the population.

- Hypothesis : In the countries that vaccinated people are many of the population, deaths are decreasing.
- How did you obtain the hypothesis: I checked some countries like Israel that many of the population were vaccianted, and after months there were no deaths.

- Hypothesis : Death rates are less in countries that aged 65 or 70 older are small part of country.
- How did you obtain the hypothesis: I checked countries that has big population (Vietnam 96 million and Taiwan 23 million) and small percent of aged 65/70 older people and I saw that new deaths are almost 0 and new cases are not big even before of the vaccination.

- Hypothesis : New cases are less in countries that aged 70 older are small part of country.
- How did you obtain the hypothesis: Like hyphotesis before, Vietnam and Taiwan's aged 70 older population are less that % 10. And when the death rates are small, the cases are small too.

- Hypothesis : Most of the non-old people survive from corona without any test etc.
- How did you obtain the hypothesis: If aged 70 older are small part of country, death rates are small but at the same time new cases are small too even that country has big population. Also, when we divide total deaths to total coronavirus cases, we can see that, aged people's population in country are not changing death rates. So there is a ratio between death/case. Most of the countries' rate are between %1.5 and %3.5.(There are some exceptions like Turkey). For example 2 countries like United Kingdom (aged 70 older is %12.53) and Iran(aged 70 older is %3.18) and even the difference between age ratio that much, death/case are both around ∼2.9%.

- Hypothesis : Diabetes prevalence and cases/death rate are irrelevant.
- How did you obtain the hypothesis: I took 2 countries like United States of America(%10 diabetes prevalence) and Benin(%0.99 prevalance) that has big difference. And when I checked the death rates I saw that rates are really close like %1.3 ∼ %1.8.

- Hypothesis : The countries that smoking is high, the death rates are more.
- How did you obtain the hypothesis: There are 2 countries Cuba and Greece that has close population(10 million and 11 million) and Cuba's male and female smokers average is %35 and for Greece, it is more than %43. So I checked the death numbers and Cuba was 569 but Greece was sadly 9864. So I think smoking has really bad effects for coronavirus.

- Hypothesis : The countries that don't wash their hands most and the countries that wash their hands corona virus death rate difference is irrelevant.
- How did you obtain the hypothesis: There are 2 countries called Bosnia and Herzegovina and Zimbabwe. In Zimbabwe handwashing facilities are 36.79 and in Bosnia and Herzegovina it is 97.16. But the death rates(corona cases / corona deaths ) is ∼ %4.2 for Bosna and Herzegovina, and ∼%4.1 for Zimbabwe.