Mustapha Tidoo Yussif and Samuel Atule

TASK ONE

Our approach maintains a series of initial variables and calculates the answers to all questions within a single *for-loop* when reading from the *covid_data.csv* file.

We answered (a) and (b) by initializing four variables to hold the *highest number of infections* and its *corresponding country*, the *second-highest number of infections* and its *corresponding country* with the first country values from the file. When we encounter a sum of a country's *new confirmed cases* which is larger than the *highest number of infections*, we update the *highest number of infections* with the *new confirmed cases*. The previous *highest number of infections* becomes the *second-highest number of infections*. We update the countries at the same time.

We answered (c) and (d) by first assigning the value 0, and the first country read from the file to the initial variables that will hold the *highest infection (and death) rate* and their countries accordingly. Next, we calculate the infection (and death) rates for *each* country. When a l*arger infection (and death) rates* are found, we update the *highest infection (and death) rates accordingly as well as their corresponding* countries.

We answered (e) by finding the total confirmed infection cases in a running sum. A similar thing is done to find the total deaths. The overall death rate is calculated by dividing the total number of deaths by the total number of confirmed infections.

We answered (f), (g), (h), and (I) by calculating the correlation coefficient over the recent 7 (1-week data) data points for each country. The countries with a positive correlation coefficient are considered as the countries with a positive trend while those with negative correlation are retrieved as countries with a negative trend. Among all the countries with the positive trends, the country with the highest positive correlation has the steepest increase while the country with the lowest negative correlation among the negative trend countries has the steepest decrease.

(j) To find the country whose number of infections per day peak the earliest, we first found the most initial peak point for each country and compare it with a global earliest peak point. If a country peaked earlier than the global one, we update the global earliest peak and keep that country until another peak is found.

TASK TWO

Our approach to solving this task uses pattern matching. We use the KMP and Boyer Moore string matching algorithms combined to find the pattern, that is the data points from the *partial_time_series.csv,* from the *covid_data.csv* file (Tsarev et al., 2016). We first read the partial time series data into a list. Based on the size of the *partial_time_series.csv* we ether call KMP or Boyer-Moore algorithm. Generally, KMP performs better than Boyer-Moore when the pattern to find is shorter. In this program, if the number of values in parital_time_series is greater than 15 we employ Boyer-Moore and use KMP is it is less than or equal to 15. Both algorithms process the pattern. The preprocessing is done so that the pattern can be shifted by more than one. Unlike KMP, Boyer-moore is based on backward pattern matching. This compelled as to store the data from the *covid_data.csv* in a list. But for the KMP, we search for the pattern while reading the file, *covid_data.csv.*

Reference.

Tsarev, R. Y., Chernigovskiy, A. S., Tsareva, E. A., Brezitskaya, V. V., Nikiforov, A. Y., & Smirnov, N. A. (2016). Combined string searching algorithm based on knuth-morris- pratt and boyer-moore algorithms. *IOP Conference Series: Materials Science and Engineering*, *122*, 012034. https://doi.org/10.1088/1757-899X/122/1/012034