2020 Career Fair Programming Competition
Part A – Code Challenge

Problem Revealed: Thursday, 4 March 2020
Submissions due: **Monday, 9 March 2020** at **midnight**

Contents

Problem Description

At the end of December 2019, there were reports of an acute respiratory syndrome in the Chinese Wuhan municipality. The outbreak spread quickly to other parts of China as well as to other countries in Asia, Australia, Europe, North America and Africa. In February, the novel coronavirus was named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the disease associated with it is now referred to as COVID-19. Since 31 December 2019 and as of 4 March 2020, **93,076 cases** of COVID-19 have been reported, including **3,202 deaths**. The outbreak is still rapidly evolving. For comparison, in the 2014-2015 Ebola outbreak in West Africa, over 28,000 cases were reported and more than 11,000 people were reported to have died of the disease.

In this coding challenge, you will process some data related to the ongoing COVID-19 outbreak.

Task 1

You have been provided with a log file (in CSV format) of data on the number of newly confirmed cases and deaths of COVID-19 by country and by date since December 2019. You have also been provided with a log file (in CSV format) of population data of countries. You are to write a programme which determines answers to the following questions:

a) Which country has the highest number of infections to date? How many infections have been recorded in that country?
b) Which country has the second highest number of infections to date? How many infections have been recorded in that country?
c) Which country has the highest infection rate (ratio of number of infections to population) to date? What is the infection rate?
d) What is the overall death rate (ratio of number of deaths to number of infections) for COVID-19?
e) Which country has the highest death rate (ratio of number of deaths to number of infections)? What is the death rate for that country?

f) In which countries is the number of new infections per day on the rise (i.e. a positive trend over the most recent 1 week of data)?

g) Of the countries above (from *f*), which country has the steepest increase?
h) In which countries is the number of new infections per day decreasing (i.e. a negative trend over the most recent 1 week of data)?
i) Of the countries above (from *h*), which country has the steepest decrease?
j) Of the countries above (from *h*), for which country did the number of infections per day peak the earliest? When did the number of infections per day peak, for this country?

Task 2

You have been given a third file with a list of time series data corresponding to a portion of one of the country's new confirmed cases data. However, the data is not labelled so you do not know which country the data corresponds to, or what dates it corresponds to. Write a programme to determine which country the data corresponds to, and what the starting date of the data is.

Note that, for both tasks, your program will be tested with different data from what has been provided to you. As such, it is important that you write your program such that it adheres to the guidelines described in this document, so that your code can be run with a different input file without re-compiling your code.

Program Execution Specifications

Your programs can be written in either Python or Java. In either case, you should write your program such that it can be executed from the command-line, taking the filenames as command-line parameters.

For **task 1**, your programme should be run from the command line as follows:

```
<interpreter> <programme> <covid_data> <population_data>
```

where:

`<interpreter>`	is the name of the Python interpreter or Java interpreter,
`<programme>`	is the name of your programme,
`<covid_data>`	is the name of the file containing the time series data of Covid cases and deaths by country, and
`<population_data>`	is the name of the file containing the population of various countries

For **task 2**, your programme should be run from the command line as follows:

```
<interpreter> <programme_name> <covid_data> <partial_time_series>
```

where the first two parameters have the same meaning as above, and

`<covid_data>`	is the name of the file containing the time series data of Covid cases and deaths by country, and
`<partial_time_series_file>`	contains the partial time series data that you are trying to match to the data in the `<covid_data>` file.

The format of the input files are described in the next section.

Your programme should write its results to an output file whose format is described later in this document.

Input File Formats

All input files are CSV (comma-separate-value) files. Note that CSV files are simply plain text files in which the data values on each line are separated by commas. You can read such files in your Java or Python program like you would read any text file. You can view such files with a text editor (such as Notepad). You can also use Excel to view such files in a nice tabular format.

The **covid_data** file is structured as follows:

- Line 1 has the column headers
- Line 2 onwards has the data. The columns are as follows:
    1. ***DateRep*** – the date
    2. ***CountryExp*** – the name of the country. All the data is linked with a specific country, with one exception described as "Cases on an international conveyance Japan".
    3. ***NewConfCases*** – the number of new confirmed cases of COVID-19 on that date in that country.
    4. ***NewDeaths*** – the number of deaths due to COVID-19 on that date in that country.
    5. ***GeoId*** – a code representing the country

The **population_data** file is structured as follows:

- Line 1 has the column headers
- Line 2 onwards has the data. The columns are as follows:
    1. ***Country*** – the country
    2. ***Country_Code*** – A code for the country
    3. ***Population*** – the population of the country
    4. ***Year*** – the year that the population data dates from

For task 2, the **partial time series** file also has a .csv extension, but it has only one column of data, representing the number of newly confirmed cases of COVID-19 in an unspecified country, for an unspecified period of time.

Note that the data files provided are simply samples. Your code will be tested on these and other data files with the same format, but which could have fewer or more data points than the sample input files that have been provided. As such, you need to write your code to be general enough to work with any number of rows of data.

Output File Formats

For task 1, your programme should output a file named *task1_solution-<input_file>.txt*, where *<input_file>* is the name of the input file. E.g. if your input file was ***data1.csv***, the output file will be named ***task1_solution-data1.txt***.

- The file *task1_solutions-<input_file>* should have answers to the questions (a) to (j) of task 1, one answer per line. If a given question has multiple answers, the various answers should be on the same line, separated by commas.

For task 2, your programme should output a file named *task2_solution-<input_file>.txt*, where *<input_file>* is the name of the file with the partial time-series data.

- The first line of file *task2_solution-<input_file>* should have the name of the country matching the partial time series data provided.

- The second line of the file should indicate the starting date of the partial time series data provided.

What to submit

You should submit:

1. Your code and data files
2. A readme.txt file indicating which tasks your code supports and any additional information needed to run your code.
3. A 1-page document briefly describing your high-level approach, and listing any references (see rules below).

Rules of the Challenge

1. Participation in the competition is by Ashesi students, either as **individuals** or as a **team** made up of **two** students from the same yeargroup. **No collaboration** is allowed except between two members of a team. Individuals who are not Ashesi students may not participate.
2. In addition to overall winners, there will be winners for each year group, so members of all year groups are encouraged to participate and not feel intimidated by the idea that they may be "competing" against more experienced programmers.
3. Solutions must be coded in Python or Java. Standard built-in classes and libraries can be used (e.g. file reading libraries, the math library in Python and classes in the java.util package in Java). In general, no external/third-party libraries can be used in either language. If planning to use any non-standard packages, it is best to enquire first about whether the package will be allowed.
4. All submitted code must be your own. Code may not be copied from any source.
5. You are allowed to do research if needed. However, you **must** cite **all** resources you consult.

Evaluation Criteria

We reserve the right to disqualify submissions that do not follow the specified code structure or do not comply with the submission instructions.

Solutions will be evaluated according to the following criteria. The order below will be used to rank the various submissions.

1. Completion & correctness (the number of questions answered correctly).
2. Efficiency (the time it takes to execute your solution approach). Accommodation will be made for the inherent difference in speed between Java and Python. However, no accommodation will be given for unnecessary I/O, so your programme should output only the information that is necessary.
3. Code formatting and structure: code should be modular, well-structured, well-formatted, and well-commented.
4. Clarity of 1-page documentation of approach