Data Exploration and Visualisation

Coursework 1

Coursework Outline

This is the first coursework of the Data Exploration and Visualisation module. The coursework provides you with an opportunity to demonstrate the knowledge and skills that you have accumulated in data visualisation so far. The coursework is divided into three parts.

1. Data preparation & pre-processing: you need to explore and understand the dataset given to you (described below). You also need to prepare the data for visualisation by re-formatting it followed by identifying any missing values and proposing solutions to deal with the missing values.

2. Data visualisation – reproduction & discussion: you will be given a number of interesting plots from various publications and you will be asked to produce similar plots (sometimes with different parameters).

3. Data visualisation – investigation & analysis: you will be given a number of problems and you need to investigate the best way to visualise the data. You will be also asked to justify your choices and provide a brief explanation of the outputs.

Please note that a detailed break-down of what you need to do in each part is explained below. The project has a 50% weight of the overall mark.

All your answers, codes and figures must be submitted in **ONE** document (preferably a PDF document) via Moodle. Please be tidy and start each question on a new page.

For your implementation, you are free to use either Python or R.

Submission deadline: 26th June 2020

For any question, please contact the module leader at maysson.ibrahim@buckingham.ac.uk

Dataset description

In this coursework, you will be using the Oxford COVID-19 Government Response Tracker (OxCGRT) dataset [1]. In addition to recording the confirmed and death cases on daily basis, OxCGRT provides a systematic way to track the stringency of government responses to COVID-19 across countries and time. It uses a novel index that combines various measures of government responses. The data is collected and updated in real time by a team of dozens of students and staff at Oxford University.
To access the data, OxCGRT provides two web APIs:

[1] https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker

1) Data of all countries between two dates
   https://covidtrackerapi.bsg.ox.ac.uk/api/v2/stringency/date-range/{YYYY-MM-DD}/{YYYY-MM-DD}

   The output has JSON format of countries' stringency data, confirmed cases and deaths on a day by day basis.

2) Data of a country on a specific day
   https://covidtrackerapi.bsg.ox.ac.uk/api/v2/stringency/actions/{ALPHA-3}/{YYYY-MM-DD}
   The output has JSON format for the requested country/date combination where ALPHA-3 is the ISO 3166-1 alpha-3 country code.

Part 1: Data preparation & pre-processing

In the "Accessing OxCGRT.pdf", we explain how to access the above web APIs using Python. We also discuss some examples related to handling JSON output. Please check the slides for further details.

Q1. The part of OxCGRT dataset relevant to the practical elements of this coursework is given to you in "OxCGRT_summary.xlsx" accessible via Moodle. The summary has been presented in thee Excel sheets:

- Stringencyindex_legacy
- Confirmedcases
- Confirmeddeaths

As a data preparation exercise, write the code needed to produce an identical content to the above Excel file and save the output in a file called "My_OxCGRT_summary.xlsx". This should be done using data retrieved from the JSON responses using the web APIs explained earlier.

(10 marks)

Q2. Handling missing values in data before visualisation is a critical step in which different strategies can be applied (e.g. dropping observations, imputing by zero or other values such as mean, etc.). In the paper (Variation in government responses to COVID-19) [2], OxCGRT discusses the strategy chosen to handle missing value **when calculating the Stringency Index based on the indicators**

a) Review and briefly explain the chosen strategy (see Section 3 in the paper[2]). Highlight the pros and cons of their strategy (You may compare the chosen strategy to other strategies to support your answer).

[2] https://www.bsg.ox.ac.uk/sites/default/files/2020-05/BSG-WP-2020-032-v5.0_0.pdf

b) The three sheets in “OxCGRT_summary.xlsx” have some missing values. To prepare the data for visualisation, choose an appropriate strategy to handle the missing values in the data from the three sheets. Justify your choice and write the code needed to implement it.

(10 marks)

Part 2: Data visualisation – reproduction & discussion

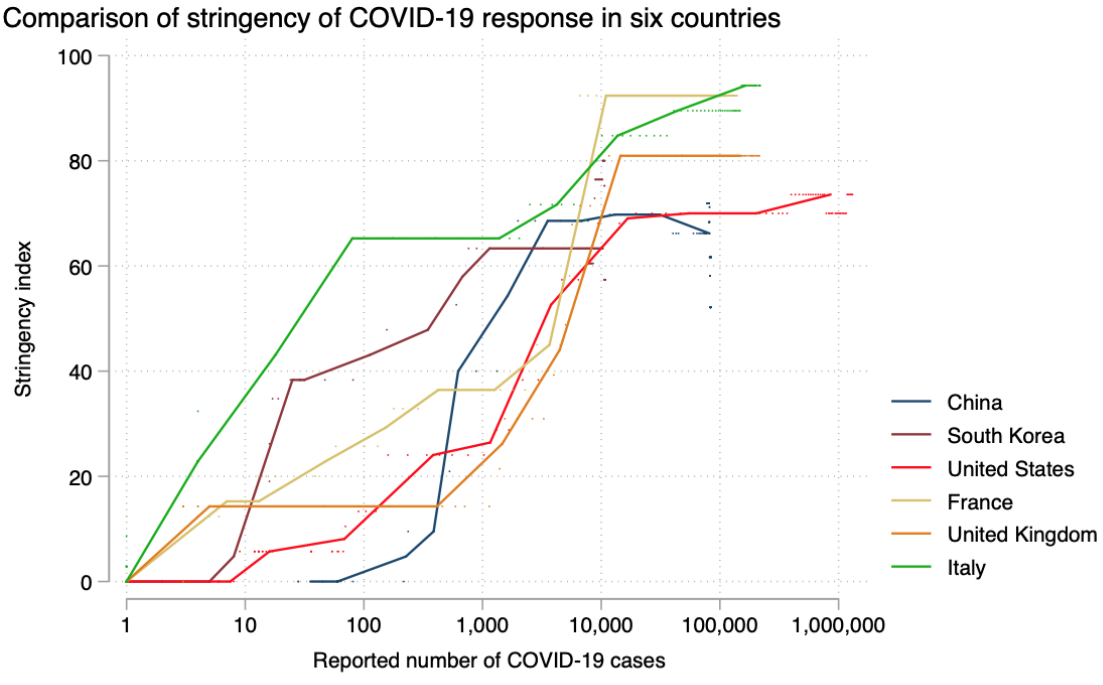
To create plots in this part, you need to use the data from “OxCGRT_summary.xlsx” after handling the missing values in Q2b. Please note that your plots might look different to the ones given in this part. You can use either python or R as a programming language.

Q3. Figure 1, below, shows a comparison of stringency of COVID-19 responses in six countries based on data up to 10th May 2020.

a) The logarithmic scale was used to show the number of COVID-19 cases on the X axis. Explain why log scale is preferable over the linear scale in this case.

b) Write the code needed to create a plot similar to Figure 1 based on data up to 10th May 2020 (different colours may be used).

(10 marks)



Data from 10 May 2020. Individual countries may be several days older.

Source: Oxford COVID-19 Government Response Tracker. More at: bsg.ox.ac.uk/covidtracker or github.com/OxCGRT/covid-policy-tracker

Figure 1

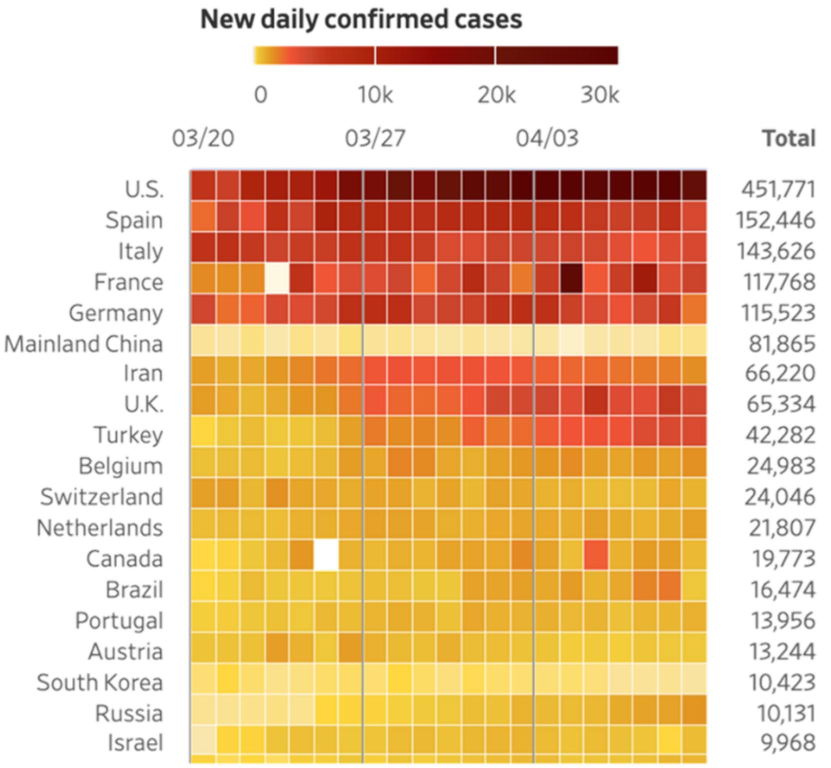
Q4. The heatmap in Figure 2 below shows the NEW DAILY confirmed cases of COVID-19 between 20th March and 10th April 2020.

a) Calculate the average number of confirmed cases for each of the 10 weeks between 2st March to 10th May e.g. for week 1 (from 2nd March to 8th March), calculate the average confirmed cases over the 7 days.

b) Create a heatmap to show the WEEKLY average number of confirmed cases of the 10 weeks calculated in (a) for the 10 countries that have the highest number of confirmed cases.

Note: each cell in the heatmap should represent the average number of confirmed cases in one week. Your list of the 10 countries may be different from the ones in Figure 2.

(15 marks)



Sources: Johns Hopkins Center for Systems Science and Engineering, the Lancet, Associated Press

Figure 2

Q5. Figure 3 below shows the US compared with the rest of the world on May 1st. Write a code to create a similar figure based on data from 10th May 2020. The population can be taken from Figure 3 but the confirmed cases and deaths must be calculated based on the data from the excel file.

(10 marks)

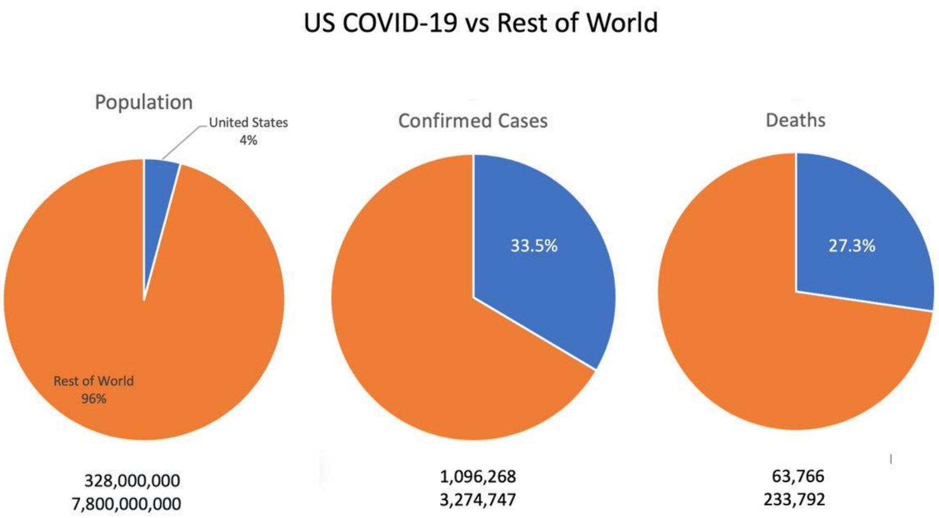


Figure 3

Part 3: Data visualisation – investigation & analysis

The data given to you in OxCGRT_summary.xlsx reports the accumulated number of confirmed cases and deaths on daily basis. This part uses the term "NEW DAILY" confirmed cases or deaths to refer to the new cases reported in each day (should be calculated from the data).

Q6. Visualise the number of confirmed **deaths** in the UK between 7st March and 10st May.

(10 marks)

Q7. Visualise and compare the change in the number of NEW DAILY confirmed **cases** between 1st March and 1st May for five countries (UK, Spain, Italy, France, and USA). Justify your choice of the plot and briefly discuss the results.

(15 marks)

Q8. Use scatter plot to visualise the correlation between the number of COVID-19 confirmed cases and the government response represented by stringency index for all countries based on data published on **4th May 2020.** Logarithmic scale should be used to show the number of COVID-19 cases on the X axis.

(10 marks)

Q9. Repeat Q8 on countries with more than 1000 confirmed cases where the dot size (bubble size) should represent the number of confirmed deaths based on data published on **4th May 2020**.

(10 marks)

Marking Matrix for this coursework

Work Aspects	**Mark Grades**					
	1st	**2.I**	**2.II**	**3rd**	**Pass**	**Fail**
Accuracy	Precise and correct terms used	Level of perfection close to the level for 1st	Sufficiently precise. Most terms used correctly	Reasonably accurate in context but not in words	substantial inaccuracy, but does not affect the whole work	Severe lack of precision and misunderstanding
Validity	Argument consistent and logical. Show strong critical reasoning	Most argument consistent and logical. Show critical reasoning	Good logical argument. Show limited critical thinking & reasoning	Sufficiently valid argument, but may not with proper reasoning	Limited valid argument. Limited critical reasoning	Little valid argument. Opinionated decisions
Completeness	All required elements covered	Nearly all elements covered	Majority elements covered	Sufficient elements covered	At least more than half of the work done	Severely incomplete work
Objectivity	Factual not opinionated	Factual not opinionated	Mainly factual	Limited or not well argued	Very limited	No objectivity
Clarity and Professionalism	Statements clearly made, diagrams, figures and references professionally presented	Statements carefully built. Diagrams, figures & references well presented.	Statements easy to follow, but may not be carefully built. reasonable use of figure reference, diagrams	Sufficiently clear to follow. Use of figures, diagrams and references is present	Difficulties in explanation. However, the work as a whole is still understandable.	Severely lack of clarity. Extremely limited in content. No sign of professional “look and feel”.