Wrangling details

Data sources

COVID-19 Government Measures Dataset (XlSX/CSV, 3.8MB):
https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset

This dataset contains COVID-19 related data for each country, in particular, what actions has each country taken to contain COVID and the corresponding time each action has been taken. The time start from January 2020 to December 2020.

There is a total of 23925 rows of data in total, 14 columns. Some rows have missing values but is not real concern, some date values are missing that those rows will need to be deleted because date is one of the keys for merging. The most important column "measure" has around 30 unique values. There are some null values for implemented_date and that must be addressed since date will be the key column for merging. The date is in mm/dd/yy format.

The dataset will be reliable because the it comes from a reliable source.

Coronavirus (COVID-19) Death Datasets (JSON, 50.8MB):
https://covid.ourworldindata.org/data/owid-covid-data.json

https://ourworldindata.org/covid-deaths (Main page with description)

This dataset contains COVID-19 data for each country, there are around 60 fields containing a lot of relevant information including number of populations infected and number of deaths as well as date. This dataset is updated daily so there will be minor differences between the obtained dataset and dataset freshly downloaded from the source page. The time starts from January 2020 to early 2021.

There are around 200 objects each represents the symbol of a country (e.g. NZL, USA), each has many sub-objects which are the records. There is a total of 85377 records, 23 fields. Some of the fields contain information that is not useful such as life expectancy and hospital admission rate. Some of the fields contain mostly null values (e.g. total number of vaccinated). The date is in yy-mm-dd format which is different than the first dataset.

This dataset is also reliable because it's from a rather well-known website.

Both two datasets contain COVID related info for every single country and both files are way too large and both datasets contained information that is not relevant to the questions that I imposed and will not be useful for my audiences. Therefore, we decided to use subset for both datasets.

Combination steps

Subsetting

The subset was created by selecting specific rows, in particular, rows for the countries of

China, New Zealand and United states. This step can be easily done for the CSV file by tabularizing the data and using filter. For the json file, this can be carried out manually using regex and replace that we learned in class. (For example, replace everything from start to the first object with country of China to empty space, from last object with country of China to first object with country of New Zealand to empty space)

Now we have done row wise subset, there is also need for column wise subset, it is very simple for the csv file, we simply delete the columns that we do not want. For the json file we can load the file to mongodb and only project the fields with useful info then export. Alternatively, I found it easier to convert json to csv first then carry out this step.

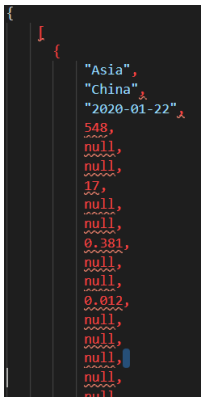
Fields selected for first dataset: Country, Category, Measure, Comment, Date_implemented

Fields selected for second dataset: Location, Date, total_cases, total_deaths, total_cases_per_million, total_vaccination, stringency_index, new_death_per_million_smoothed and new_cases_per_million_smoothed

Converting and Preprocessing

First, I converted the json file to csv, The regex methods we learned at class can achieve this. Using regex method we will first record field names.

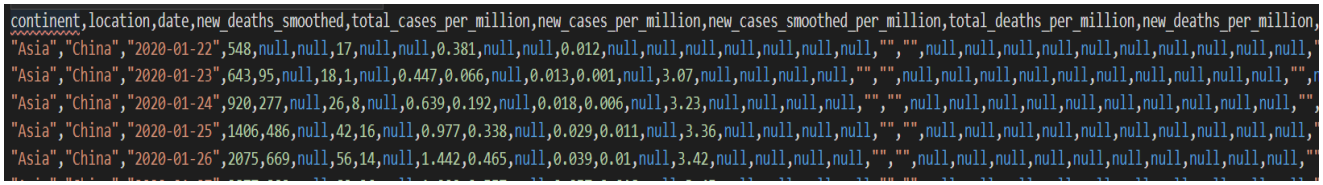
Then find and replace ".+: with nothing to remove field names.



Then replace existing \n with nothing, and find all /},{/ to replace them with \n so that each record are one a different line,

Then replace brackets and curly brackets between the start and end of each the three main obejcts) /\}\],\[\{/ to \n

Lastly, add the header and replace spaces and we have it converted to csv file.



After both files are in csv format we load both files in excel.

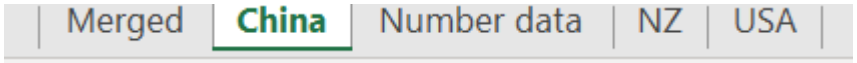
There are some rows in the first dataset being removed because these has no date (Blank).

The date in second dataset is in the format of yy-mm-dd, I used Excel formula to convert it to mm/dd/yy so the formats are the same now.
=REPLACE(RIGHT(B2,5)&"/"&LEFT(B2,4),3,1,"/").

There are also rows in the first dataset that does not have corresponding row in the second dataset since the second dataset starts at 1/22/2020 (dd/mm/yy). Therefore, Rows with date before 1/22/2020 were removed.

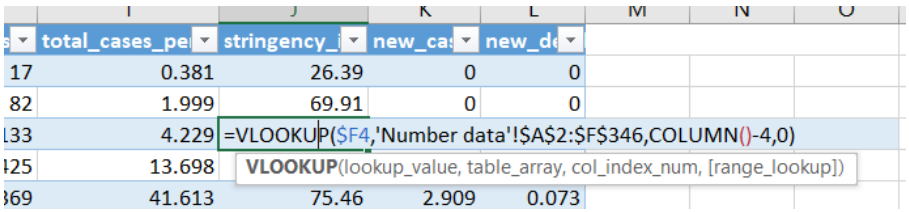
Merging

I created a new excel file, included both subsets as sheets. Using Vlookup with the second dataset being lookup table, we were able to merge the two sheets into one but we need to consider that there are three countries and vlookup assumes all observations belong to the same country so I have created three separate sheets for the three chosen countries and performed Vlookup merge to the three work sheets individually.



The row range used for Vlookup for China is 2:346, for New Zealand is 347:654 and for USA is 655:999, these were observed manually.

Lastly, I combined the merged data of the three countries into one sheet using copy and paste values only, we have finally obtained the final merged data in excel format with 12 columns and 968 rows.



Google drive link for the final merged dataset:
https://drive.google.com/drive/folders/132S16pkC2QnClOdbqaqKRRzoSgJp5yKC?usp=sharing

Questions and Answers

When did China implement the first international flight suspension, how many cases was there in China, United States and New Zealand respectively?

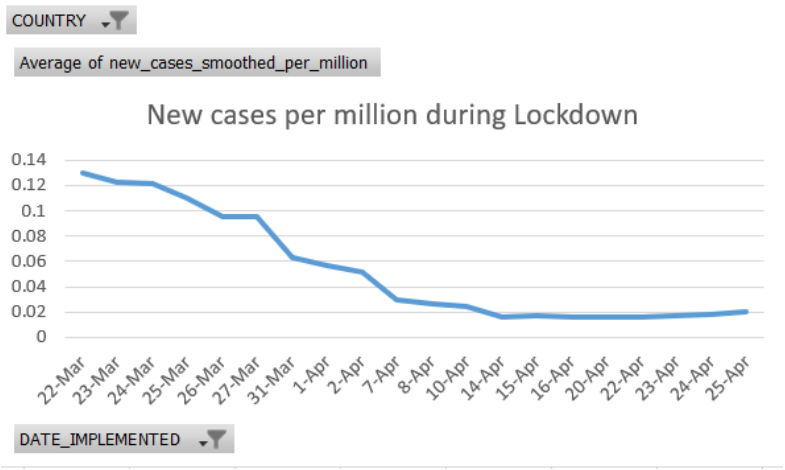
Apply the same filter and also filter the measure to “International flight suspension”, the first record has a date of 3/20/2020. Clear the filters, apply a filter to date. China had 81250 cases, NZ had 39 cases and USA had 20030 cases at the time of first international flight suspension announced by China.

COUNTRY	ADMIN_LEVEL_NAME	CATEGORY	MEASURE	COMMENTS	DATE_IMPLEMENTED	total_cases
China		Movement restric	International flights suspension	Beijing will redirect some inbound flights scheduled to land at its Beijing Capital Internatio	3/20/2020	81250
New Zealand		Public health mea	Isolation and quarantine policies	All returning residents and citizens must isolate themselves for 14 days upon arrival inc NZ	3/20/2020	39
New Zealand		Governance and s	Economic measures	NZ Govt and AirNZ have agreed a debt funding agreement through commercial 24-month l	3/20/2020	39
United States		Movement restric	Border closure	The United States has reached mutual agreements with Canada and Mexico to restrict no	3/20/2020	20030

When did New Zealand implement the first lock down and did it have a significant effect on reducing number of new cases? Show the effect of lockdown on number of new cases in New Zealand.

First find out the date of first lockdown using filter, filter country to “NZ”, filter measure to “lockdown”. Then create a pivot table, setting the range of date from first lockdown to one month after, use value as average of new cases per million. Lastly create the pivot chart below. (attached in the data file)

At the time of first lockdown the new cases per million was 0.123, after four weeks of lockdown, the number of new cases per million dropped to 0.016, Using excel formula we calculated the decrease was 87% which was greatly significant.



When did United States announced state of emergency? How has the stringency index and number of cases changed during that month? (Stringency index measures restrictions on travel and closure on facilities, the higher number the more restrictions and more closure)

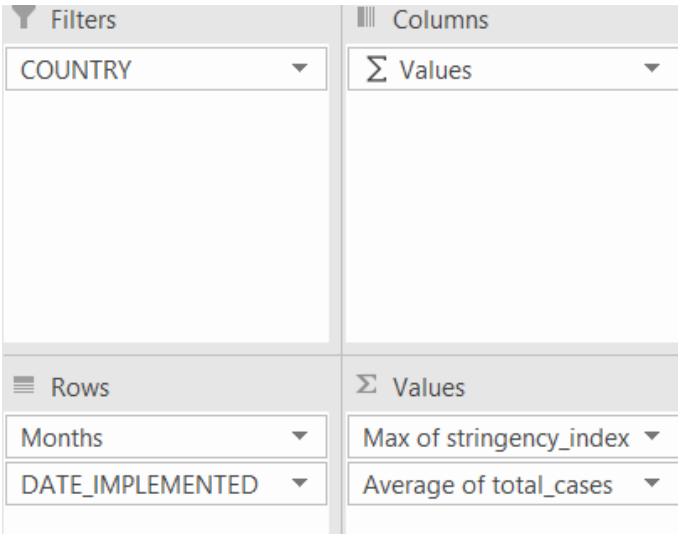
First filter the measure to “State of emergency declared” and country to “United States”, found the date was 3/1/2020.

Created a pivot table and filtered the country to “United States”, Use date as rows and max of stringency index and average of total cases as value, we obtained the information in the next page:

The stringency index drastically increased from 8.33

To 72.69 and the number of cases increased from 32

To 192301, an increase of 192269 in one month.



Row Labels	Max of stringency_index	Average of total_cases
Jan	**0**	**7.5**
Feb	**5.56**	**15.3**
Mar		
1-Mar	8.33	32
2-Mar	11.11	55
4-Mar	11.11	107
5-Mar	20.37	184
6-Mar	20.37	237
8-Mar	20.37	519
9-Mar	20.37	594
10-Mar	20.37	782
11-Mar	21.76	1147
13-Mar	30.09	2219
14-Mar	35.65	2978
15-Mar	41.2	3212
16-Mar	52.31	4679
17-Mar	55.09	6512
18-Mar	55.09	9169
19-Mar	67.13	13663
20-Mar	67.13	20030
21-Mar	72.69	26025
22-Mar	72.69	34898
23-Mar	72.69	46136
24-Mar	72.69	56755
25-Mar	72.69	68837
26-Mar	72.69	86693
27-Mar	72.69	105383
28-Mar	72.69	125013
29-Mar	72.69	143912
30-Mar	72.69	165987
31-Mar	72.69	192301

Project Summary

The purpose of my project is to combine two datasets that are in two different formats as one and gather insight from the combined dataset for New Zealand or International audiences.

The two dataset contains data on government actions on COVID and relevant numbers such as number of new cases, total cases and deaths. I chose this topic because I believe insights that will be useful and interesting for New Zealand and international audiences could be obtain from these two datasets, the insights will be especially valuable to countries that are suffering from COVID.

I have successfully combined the two datasets based on the two common key values: date and countries. And I have converted the merged file to the final intended format of excel file.

Lastly, I have imposed three question that can only be answered with the combined dataset.

First, When did China implement the first international flight suspension, how many cases was there in China, United States and New Zealand respectively?

We found that at the time of first international flight suspension from China (20th of march), China had 81250 cases, NZ had 39 cases and USA had 20030 cases. We can see that NZ has only 39 cases when China implemented the restriction, that is to say, most cases in NZ are either induced from countries that hasn't restrict its international flight or are spread domestically rather than induced from China.

Second, When did New Zealand implement the first lock down and did it have a significant effect on reducing number of new cases? Show the effect of lockdown on number of new cases in New Zealand.

We found that there was 87% reduced number of new cases, as this decrease was greatly significant, we may claim that this result outweighs the negative effects on economics and the implementation of lockdown was appropriate. See previous Q&A section for the pivot graph.

Third, When did United States announced state of emergency? How has the stringency index and number of cases changed during that month?

We found that the United states announced state of emergency at first of march, the stringency index (Index measuring restrictions on travel and closure on facilities) increased from 8.33 to 72.69 the number of cases increased from 32 to 192301. With this information we may predict that the travel restriction and closure on public facilities will likely take place shortly after an announcement of state of emergency.