Week4.Datasets

Daniel Vogel

2/14/2020

2019-nCoV Datasets

Assignment: Find Three biological data sources that you can use for your first project. For each write about why you find the dataset interesting, anticipated difficulties, and a question you would like to answer. Perform a precursory exploration (types and amount of different variables and a simple graph) for each dataset and turn in an R Markdown document.

Why is this interesting?

The nCoV virus is currently spreading at alarming rates. There are efforts to track the spread to learn the nature of the virus. It seems like a good demonstration of the capabilities of R, with respect to plotting graphs and Geo-locational information. There is also a good opportunity to show progress of the virus over time.

I am also interested in developing dashboards to show this kind of data for other uses, such as CyberSecurity. In fact, my website was hacked and as a result, I plan to develop a real-time R plot, to show where access is coming from in an effort to block access from bad actors.

I also would like to show on a dashboard, work being done by technicians in each location. Working with these nCoV datasets will help me develop the skills to create these real-time dashboards.

Questions that could be answered from this data are:

```
Rate of change in number of affected cases over time
Mortality rate per country
Recovery rate per country
Confirm Rate of Spread which is supposed to be about 2
```

Issues with the data

The main issue is how lopsided the data is per country. With 60,000+ cases in China and less than 100 cases in other regions, it is difficult to show this information to scale on a plot. To tackle this, for this assignment, I had to filter China data from the other regions.

Another issue with the data is that the tests for nCoV are not simple and accurate so data is being under-reported. DNA tests are accurate but cannot be completed for all sick people and becuase of quarantines and movement restrictions, people sick at home, are not being recorded.

Description of the Data from the source, JHU via Kaggle.

Source: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Content

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it's unclear how easily or sustainably this virus is spreading between people - CDC

This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. Please note that this is a time series data and so the number of cases on any given day is the cumulative number.

The data is available from 22 Jan, 2020. Column Description

2019_ncov_data.csv

Sno - Serial number Date - Date and time of the observation in MM/DD/YYYY HH:MM:SS Province / State - Province or state of the observation (Could be empty when missing) Country - Country of observation Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised currently. So please clean them before using it) Confirmed - Number of confirmed cases Deaths - Number of deaths Recovered - Number of recovered cases

Acknowledgements Johns Hopkins university has made the data available in google sheets format here. Sincere thanks to them. Thanks to WHO, CDC, NHC and DXY for making the data available in first place.

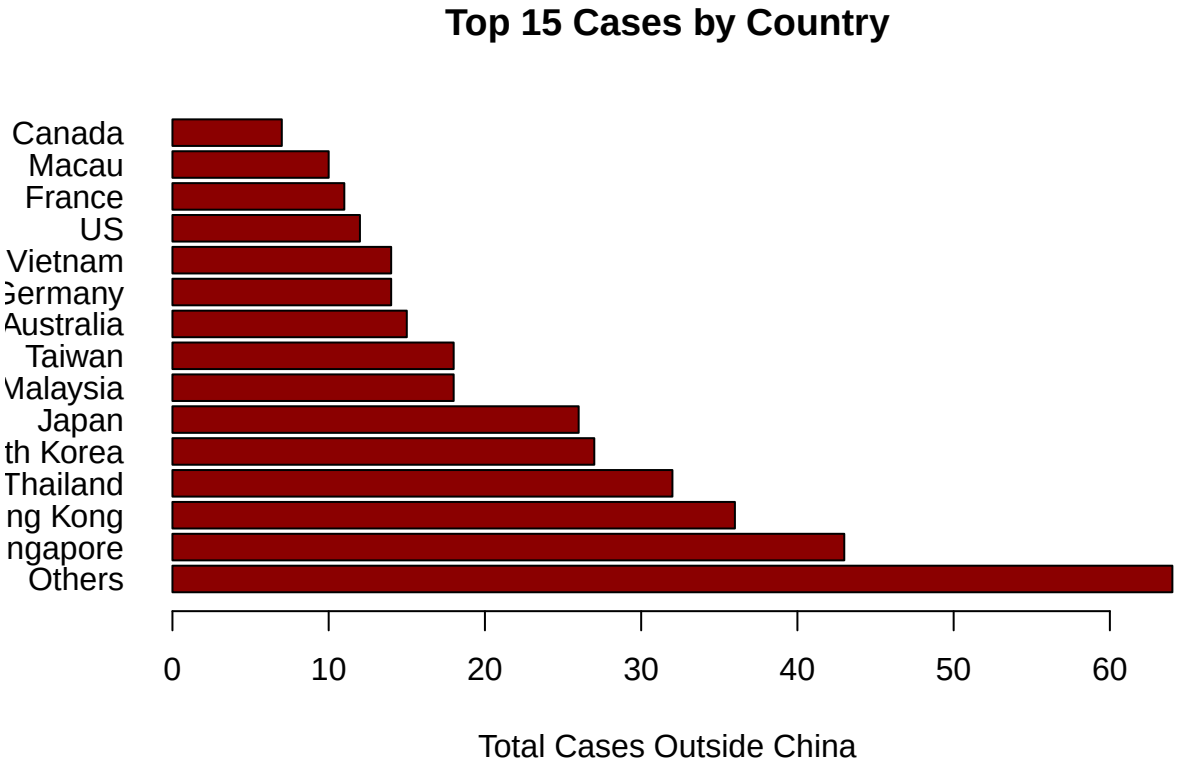
Each of these datasets has 72 rows for location and 40 variables. 36 of the columns have daily statistics for the date range 1/20/2020 - 2/9/2020 Country data is included in the first 4 columns that can be used for plotting

Summary of datasets (commented because they were verbose)

```
#summary(ncov_deaths_df)
#summary(ncov_confirmed_df)
#summary(ncov_recovered_df)
```

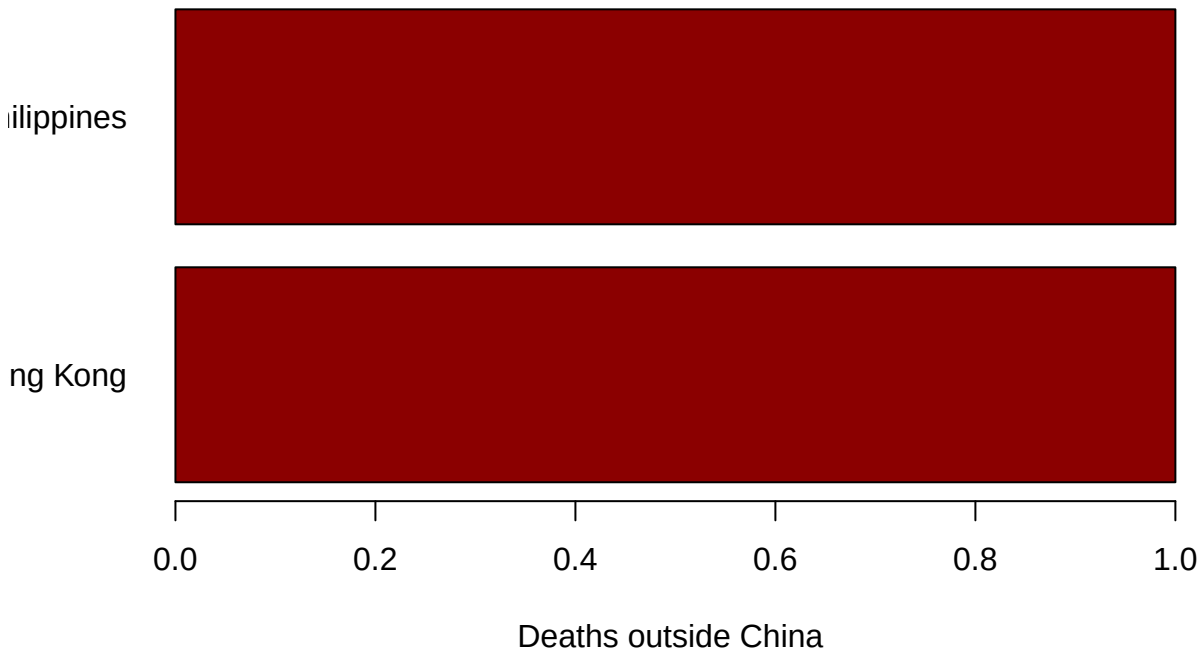
Including Plots

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## [1] "Locations outside China infected:" "28"
```



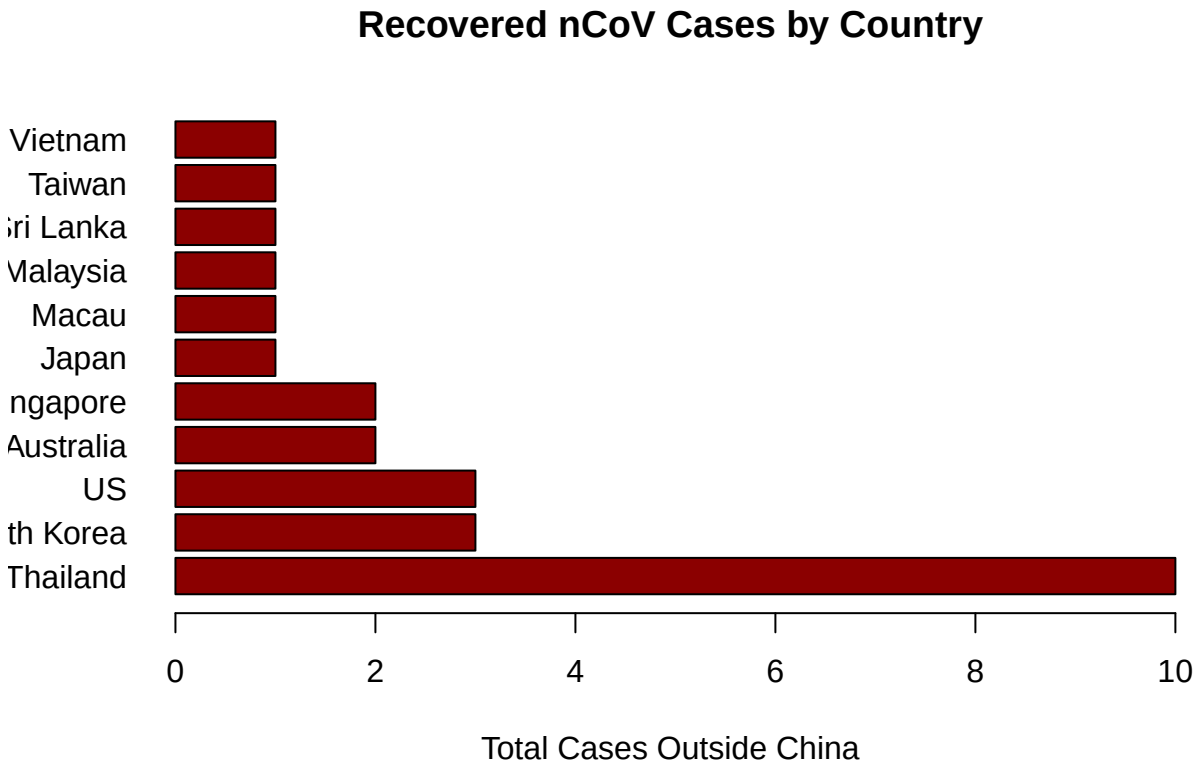

```
## [1] "Locations with deaths"
```

```
## # A tibble: 3 x 2
##   country        count
##   <chr>          <dbl>
## 1 Mainland China   908
## 2 Hong Kong          1
## 3 Philippines        1
```




```
## [1] "Recovered cases"
```

```
## # A tibble: 12 x 2
##    country        count
##    <chr>          <dbl>
##  1 Mainland China  3286
##  2 Thailand          10
##  3 South Korea        3
##  4 US                 3
##  5 Australia          2
##  6 Singapore          2
##  7 Japan              1
##  8 Macau              1
##  9 Malaysia           1
## 10 Sri Lanka          1
## 11 Taiwan             1
## 12 Vietnam            1
```


Recovered nCoV Cases by Country
Vietnam
Taiwan
ri Lanka
Malaysia
Macau
Japan
ngapore
Australia
US
th Korea
Thailand
0
2
4
6
8
10
Total Cases Outside China