

1. Data source

COVID-19 Government Measures Dataset (CSV, 3.8MB):

<https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset>

<https://data.humdata.org/event/covid-19> (Main page)

This dataset contains COVID-19 related data for each country, in particular, what measures has each country used to contain COVID and the corresponding time each measure has been taken.

Coronavirus (COVID-19) Death Datasets (JSON, 13MB):

<https://covid.ourworldindata.org/data/owid-covid-data.json>

This dataset contains COVID-19 data for each country, there are around 60 columns containing a lot of relevant information including number of populations infected and number of deaths as well as date.

I am planning to use a subset of this dataset due to its large size.

2. Merge technique

The most important common variable in both datasets is time(date), which comes in different formats (mm/dd/yy vs yy-mm-dd). The common field need to be changed to same format then we will join the rows based on the date.

1. I will try import the two files to MongoDB and merge the two datasets in MongoDB then export the resulting dataset as excel workbook or MongoDB store.

2. I will also try transforming the json file to CSV file first and then work in Excel workbook to merge the two CSV files.

3. Intended final format

Excel workbook OR MongoDB store

4. Backup plan

If I am not able to successfully combine the two datasets by week 11, I will:

A. Manually select a few important rows and columns that can be used to answer proposed question and produce a mini-subset and work with the mini-subset.

B. Submit as much as what is being done to obtain some partial marks.