ETL Project Report by: Aidan McMurry, Amos Fung, Oyeniyi D Ilugbuhi, & William Vo

What we’re looking to do is compare the amount of scientific research conducted in response to the Corona Virus (CO-VID19) and Ebola disease. Starting from day 1 of the virus, we want to compare how many cases of each virus there were, and how many articles of published research was released by that time. The purpose of doing this is to see if there was a change in the rate of how quickly each disease was spreading. We also want to compare how many confirmed cases there are and the amount of confirmed deaths.

A web scraping will be done on PubMed, which is a public online database of scientific papers maintained by the NIH. We want to scrape only articles that relate to the Ebola virus and Corona virus. Because the corona virus is relatively new, we want to get as many articles as we can. However, because we’re doing a comparison, with the Ebola virus, we want to limit the range of dates that are similar to the time frame of the Corona Virus.

The data source for both the Corona Virus and the Ebola Virus can be found on Kaggle. They have updated CSV files where the resources were pulled from various sources such as WHO, CDC, NHC, DXY. From the data, we want to clean the data so that we only show the Date, total # of cases, and total # of deaths. Because we’re looking at this from a global standpoint, we need to group everything by each date.

As a result, the final table should have a column for date, confirmed cases, and confirmed deaths. This will tell us how the virus is progressing globally. We would use the web scraped data to see if there is an increase in research being done to fight against the virus.

In summary, this information combines data for the following indicators: data, # of confirmed cases, and # of confirmed deaths. Both pieces of information highlight the progression of each disease globally and the data that we scraped can provide insight into correlations between an increase in research and the indicators we extracted from the data. The process simplifies and highlights the significant portions of the data that exclude location and virus strain specific information in an effort to provide a more accessible outlook into the research of these diseases. A global summary of the data is paralleled by the world response to outbreaks and pandemics and the individual response by each of the member-states that are a part of the WHO. Highlighting correlation and insight into such diseases can indicate the response and research of them globally and how that affects the outcome of fighting the disease.