**ETL of Recent Virus Outbreak Data**

**Introduction**

This document serves as a brief overview and as documentation for the UMN-Data Bootcamp ETL Project. Python cleaning and database setup/upload was performed by Ryan Zhang and Tyler Buhr. The data for the project was pulled from <https://www.kaggle.com/datasets>, and includes three separate CSV files for Novel Corona Virus 2019, SARS Virus 2003, and Ebola Virus 2014-2016, respectively.

* <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>
* <https://www.kaggle.com/imdevskp/sars-outbreak-2003-complete-dataset>
* <https://www.kaggle.com/imdevskp/ebola-outbreak-20142016-complete-dataset>

Python jupyter notebook was used for cleaning before creation of a PostgreSQL database to store the data. A link to the <https://github.com/> repository for the entire project can be found at the following link: <https://github.com/zhuoranzhang-ryan/ETL_project>.

**Data Used and Transformation**

The data for this project was pulled from the previously mentioned CSV files above. The SARS and Corona files contained an extra column of recovered individuals, while the Ebola file lacked this information, and thus discarded from the set. All files, in some form, contained data for a Date, Country, Number of Confirmed Cases, and Number of Confirmed Deaths; these became the common column values for the created database. Data transformation was done using python inside a jupyter notebook. Because the files were all CSV, cleaning and transformation was relatively straightforward utilizing pandas data frames to clean the data, and then utilizing a PostgreSQL connection to upload the three separate data frames to a common data base. In depth transformation processes can be viewed at the github link provided in the introduction. With these columns and datapoints, one can perform various data analyses to determine things such as mortality rate, survival/recovery rate, comparisons between viruses, project transmission potential, and more.

**Database**

The database used to load the cleaned data was PostgreSQL, nefariously named virus\_db. The database contains 3 tables, one each for Corona Virus, SARS, and Ebola; each is sequentially ID’d. The columns are country, total\_cases\_(ncov),(sars),(ebola), and total\_deaths\_(ncov),(sars),(ebola). The total\_cases and total\_deaths columns pertain to the total number of cases/deaths per country. We choose PostgreSQL due to the streamlined process of upserting data from python via a connection string, making future data set updates fast and easy to incorporate into the database. PostgreSQL is a relational database, which is more structured and rigid, because we want these important statistics such as confirmed cases and confirmed deaths to only contain those objects and be independent to avoid any confusion or misplacement of data.