Report for ETL

Group Members:

An Nguyen

Quaid Scrivens

Brett Fuller

For our Project, we researched our topic and retrieved our 5 csv datasets from Kaggle.com for Corona Virus (COVID-19):

[https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset#time\_series\_covid\_19\_deaths.csv](https://slack-redir.net/link?url=https%3A%2F%2Fwww.kaggle.com%2Fsudalairajkumar%2Fnovel-corona-virus-2019-dataset%23time_series_covid_19_deaths.csv)

**EXTRACTION:**

We retrieved 5 csv datasets for the project.

2019\_nCoV\_data.csv

time\_series\_covid\_19\_confirmed.csv

time\_series\_covid\_19\_deaths.csv

time\_series\_covid\_19\_recovered.csv

**TRANSFORM:**

For transforming, we read the csv’s into Jupyter notebook and cleaned the information using python pandas. Because there are lots of discrepancy in the information in terms of country names and province/state names among the csv. We cleaned the data to match up the names of countries and provinces/states in all csv using rename().

main\_df['Province/State'] = main\_df['Province/State'].replace(  
{'Cruise Ship': 'Diamond Princess cruise ship',  
'Chicago': 'Chicago, IL',  
'California':'Sacramento, CA',  
'San Diego County, CA':'San Diego, CA',  
'Washington': 'Seattle, WA',  
'Illinois': 'Chicago, IL',  
'Ontario': 'Toronto, ON',  
'Victoria':'Victoria, BC',  
'Arizona': 'Tempe, AZ'  
})

We also run some statistic analysis using sum() to collect the total numbers of cases in each category “confirmed”, “deaths”, and “recovered”

**LOAD:**

We set up the database layout using Quickdatabsediagram, link below:

<https://app.quickdatabasediagrams.com/#/d/aKZvAB>

Then, we exported the diagram to schema and set up the tables in PGAdmin

We have 4 tables:

* Case: is the summary/ big picture snapshot of the dataset – this table will be useful for future data updates with simple layout/format
* Geography: this table contains Latitude and Longitude of the locations in the database, where we can utilize the information for future mapping work (google heatmap for example). Location\_id is the primary key that connects to other tables for reference.
* TotalCase\_Location: this table includes the summary of “confirmed”, “deaths”, and “recovered” cases per location, utilizing the location\_id as the key to the Geography table
* Case\_country: this table includes the summary of the cases “confirmed”, “deaths”, “recovered” per country per date entry. This table can be useful for future graphing work that can display how cases are developed over time for each country.

We used sqlalchemy create\_engine, we were able to push the data to SQL database. With this database, we would be able to set up a dashboard in the future with heatmaps and time-lapse analysis of different cases in different countries around the world.