**Lillian**

**Joseph**

**Rahul**

**Group 3 Summary**

**Extract** The data was initially in CSV format from Kaggle (COVID 19 Containment measures data). I used Pandas to read directly from the CSV file.

**Transform** I grouped by country, city, and keyword to see what appeared most frequently. I then tokenized the keywords to see what came up most frequently. Then I made a frequency distribution of trigrams in the keywords. I used the NTLK package for doing NLP analysis with Python. I wanted to combine similar descriptions, but lacked the time (considering analysis was not a required step). Mostly wanted to practice with Python NLP tools.

**Load** I used SqlAlchemy and PsycoPG to push the data to PostgreSQL. I kept the tables very simple, storying the text along with value counts as 5 separate tables (keyword\_table, country\_table, city\_table, keyword\_count\_table, and trigram\_table).

**Extract**:

Our data was selected on Kaggle.com. We wanted to work with COVID related data. We worked with two datasets: COVID-19 bed preparedness and COVID cases. Our analysis with these datasets would look at how prepared states were for handling COVID cases. Both datasets were in CSV format.

**Transform**:

Irrelevant columns were deleted from both datasets. We also grouped by state to summarize our data.

**Load**:

The data was then loaded into two tables (same database) in postgres using sqlalchemy/psycopg2.