# ETL Project Proposal

I will extract COVID-19 Data from Kaggle and Data.World, and store in postgreSQL to compare worldwide COVID statistics (case number, deaths, etc.) to the United States.

I will utilize postgreSQL to store my transformed data.

# ETL Project Report

I extracted COVID-19 Data from Data.World and The Covid Tracking Project.

<https://data.world/associatedpress/johns-hopkins-coronavirus-case-tracker>

<https://covidtracking.com/data/download>

This resulted in 6, mostly cleaned CSV files.

However, while each unique data set was moderately cleaned, compared to each other, they were not in agreement. The primary cleaning was dropping unnecessary columns, e.g., columns that tracked unwanted data, rows with null values or for a different state.

Finally, after each data set’s transformation, they were loaded into PostgreSQL tables. Since the date columns were not unique in each table, only the ones where it was used that column as a primary key. All of the tables contain the state and death count, while others contain information about the population or county. The point of this was to eventually compare the reporting standards of different agencies as it relates to tracking COVID-19 for the state of Texas.