**ETL Project Report**

***COVID-19 & INFLUENZA IN CALIFORNIA***

Introduction

Our goal for this project was to use two different publicly available data sets on COVID-19 infection rates and Influenza infection rates and use a SQL database that will act as a space where the relationship between COVID-19 rates in California and last year’s flu season can be explored.

EXTRACT

Kaggle was used to find a COVID-19 data set**1**. It’s a csv file that hosted a large amount of information that gave us a great starting point, with COVID-19 cases by county, state, per day, and number of deaths.

After finding a great COVID-19 virus dataset on Kaggle that included cases by county, we decided to investigate the regional flu data available. Although Sacramento County did have some data there wasn’t a full dataset, so we looked a little bit larger regionally. This brought us to the CDC website. The CDC has a weekly tracking map that tracks flu rates per state per week. From this we were able to pull data from the past year into a CSV File**2**.

We chose to focus on the columns: State, Cases, Deaths, Week, and Year.

TRANSFORM

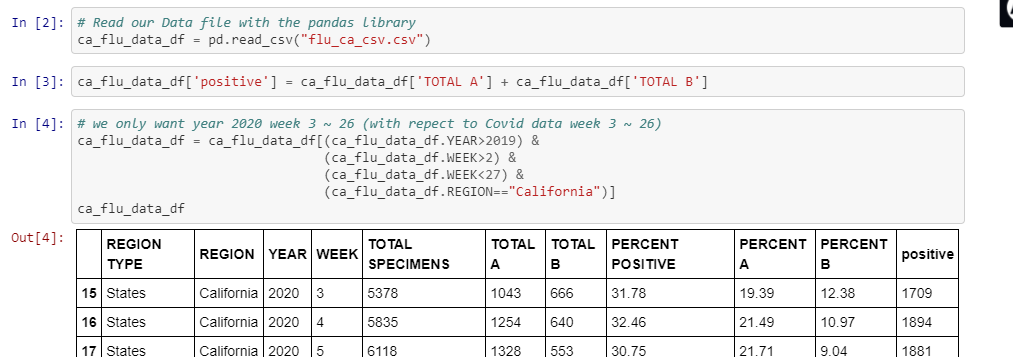
Pandas was used in Jupyter Notebook to load and create the CSV files for Flu and COVID Data. The CSV Files were reviewed and transformed to data frames.

*Influenza*

COVID-19 is fairly new. Although it is well documented, the data only goes back so far. For the flu, the data was the opposite, the data had been tracked for many years. This meant transforming the flu data to fit the COVID-19 timeframe and region of our needs was important. Pandas was used for this process.

The CDC provided this data, and had a column “Total Specimens” and “Percent Positive”. In order to get the number of flu cases we added the two columns that depicted the percent of the total specimens that were positive. It was interesting to note that “Flu Season” was something that really showed itself in the data set, as there were much greater numbers of cases in the winter than in the summer or spring. The actual number of cases the CDC lists surprised us, but it does seem like cases of the flu aren’t always recorded and don’t often require hospitalization. The percent positive column gives a good idea of what potential percentage of the overall population may have had the flu during that time.

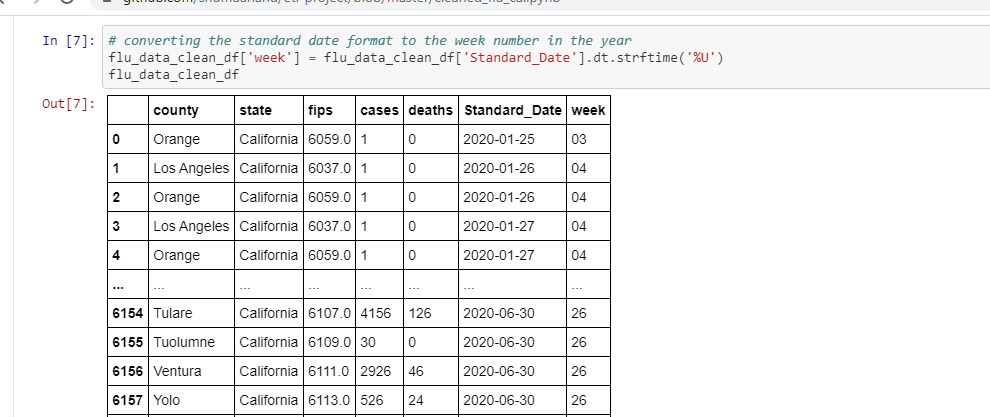
The process of changing the time frame, region, and cases column is shown in the screenshot below:

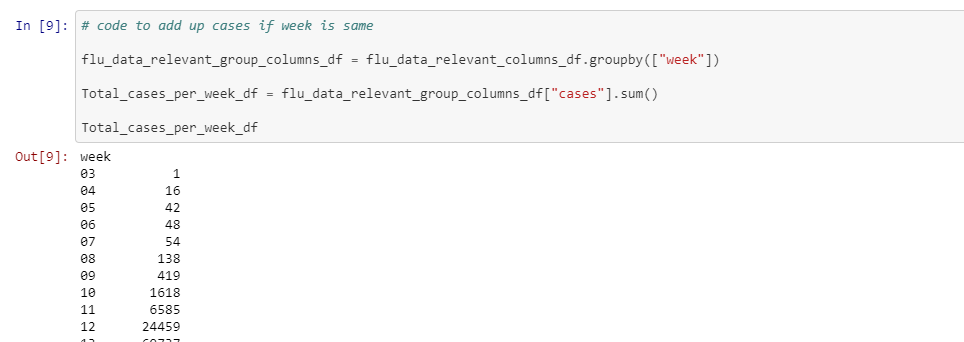


*COVID-19*

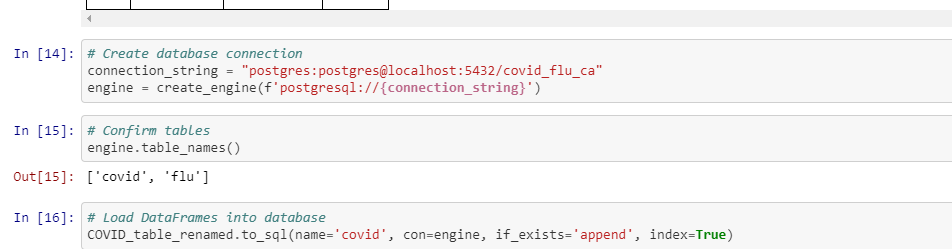
The COVID-19 Dataset had an extensive amount of data. It had cases and deaths per day, per county for the entire country. The challenge with this dataset was transforming the dates in a way that would fit the flu data from the CDC which listed its cases by week of the year (Ex. 2019 Week 1 -52) . Pandas, SQL Alchemy and Datetime were used to accomplish this.

The Standard Date was converted to what week in the year it was associated with, which is shown in the screenshot below:

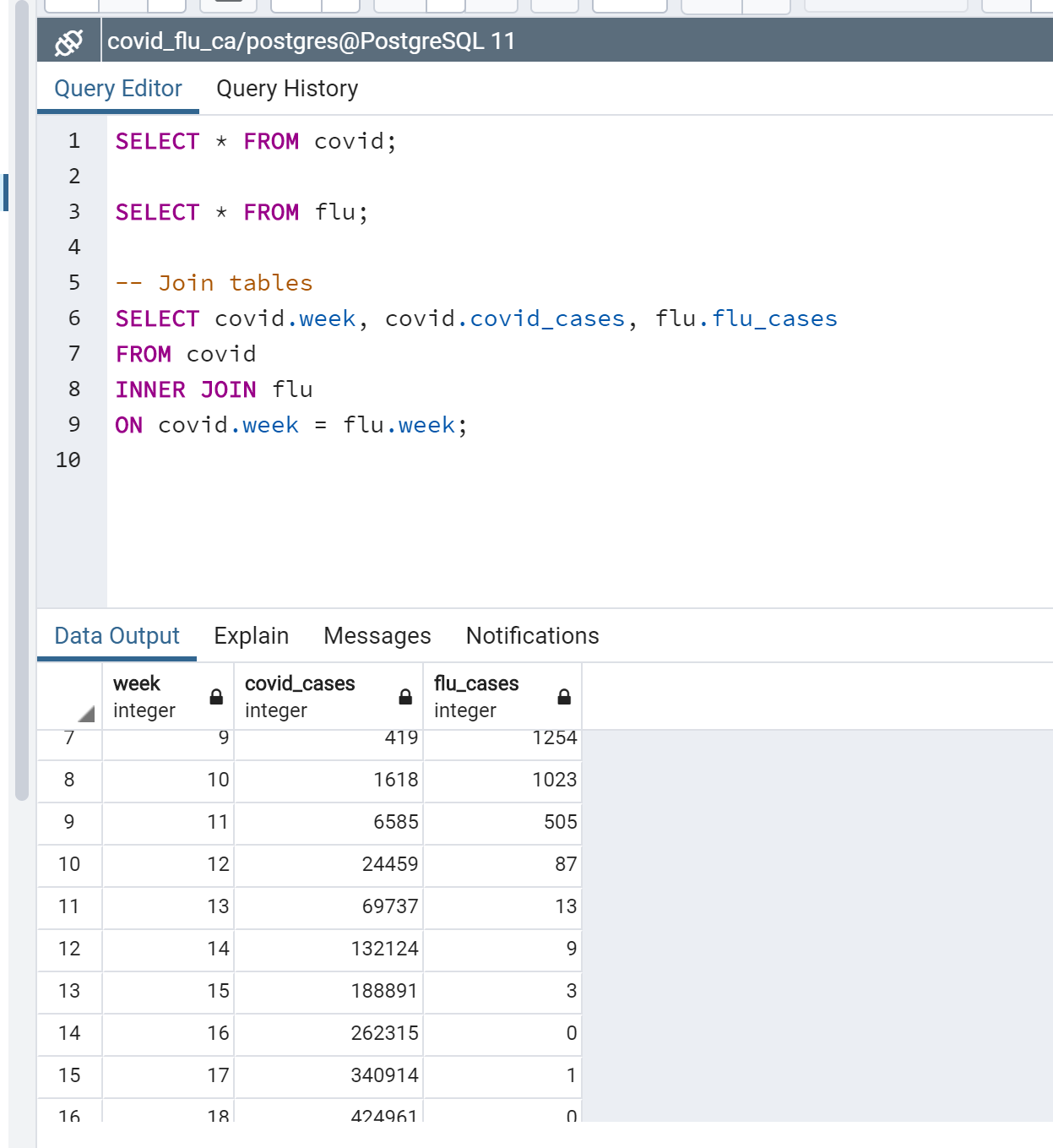


From there, the next step was to transform the data that now was framed as what week was associated with what number of cases and reframe it to number of cases per week of the year. 

Once the data was cleaned it was thenlinked to create a database connection:

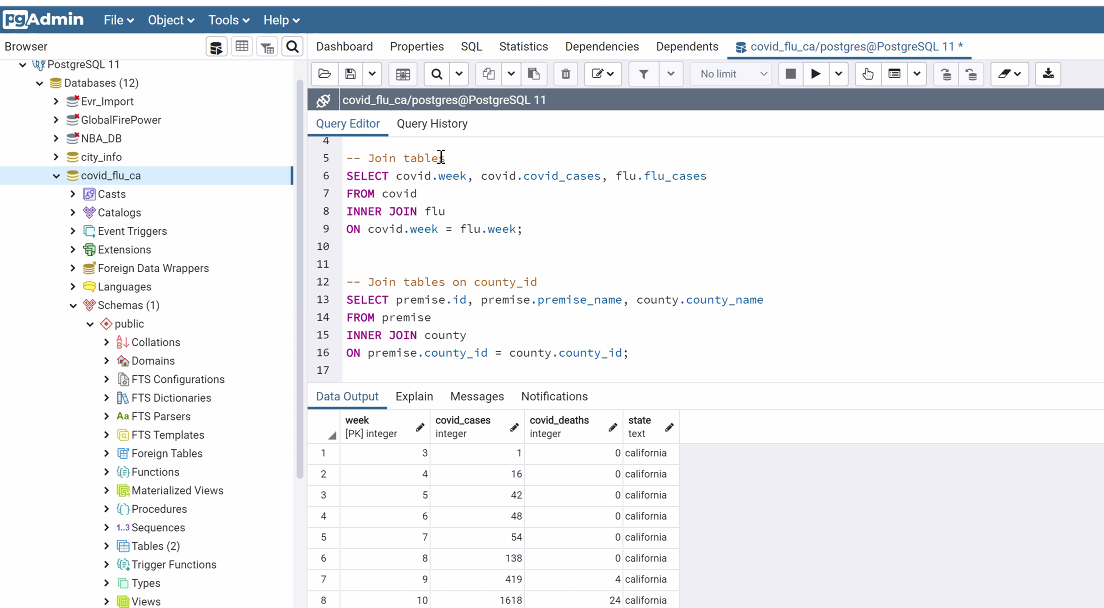


LOAD

The CSV Files were loaded into data frames and then loaded into a SQL Database. We created a SQL connection to the Postgres database using PG admin to store the cleaned data sets. After running the queries and creating the new tables with only the relevant information we needed, we then reconnected to the database and generated additional tables for the data frames. 

Initially when connecting to the SQL database, we encountered an issue where it did not read the pandas dataframe. We realized that we needed to exactly match the tables we created in the SQL database. Once we cleaned the pandas the dataframe, the connection successfully went through.

The two files were joined together by week and cases:



Conclusion

It is clear from the two datasets that tracking influenza has much more established tracking data than COVID-19, a lot of the datasets tell more of a story. As COVID-19 is continued to be studied, we can see an evolution in the data that comes out of this. When transforming the COVID-19 dataset in association with flu data, it created a really interesting presentation of the rapid increase in COVID-19 cases and deaths. Using SQL and putting these datasets in database format really allows for trends in the data to show.