Harper Hahn

Data Analytics Bootcamp

Date: March 7, 2020

ETL Project Proposal

This project analyzes the occurrence of SARS (2003), TB (2007), and H1N1 Virus (2009) in different countries around the world.

Extract: The data sources for this project are taken from Kaggle.com. All three datasets were extracted in a CSV format.

Transform: Basic cleaning took place in Excel (i.e. reformatting dates), but Pandas was the primary tool for data clean up and transformation. I had to start by figuring out which columns to keep on each CSV file and rename them to have the same column header; ‘country’, ‘year’, ‘number of cases’ and ‘number of deaths.’ After this, each dataset showed the values for the number of cases, number of deaths and countries where the presence of illnesses were documented. When I configured the columns, I could then merge my CSV files on ‘country’. The result will show the number of cases and deaths a country has experienced with respect to all three outbreaks.

Load: I connected to PostGres using SQL-Alchemy, a relational database, for the data loading component. I needed to create a table within my database before making a connection between PostGres and Pandas. When the connection was made, I made a query within PostGres (SELECT \* FROM pandemics\_data;) which displayed my cleaned Pandas dataframe in SQL. I also performed the same query from within Pandas, which is shown within my project notebook. I chose PostGres because I wanted the user to be able access country, disease, or case-specific data, and SQL enables the user to perform queries and retrieve the desired information.

My main challenge was encountering the “case sensitive error” when connecting my final Pandas dataframe to PgAdmin. I had to reset my index within Pandas and remove all capital letters, numbers, and parenthesis from my column headers in order for my data to connect properly. After doing this, the connection was successful. I also struggled with whether or not I should include the year/date of each occurrence. I ultimately decided not to because I just wanted to see how many cases/deaths each country had. The downside of not including the dates is the vast amounts of duplicate data; while not duplicates, without the dates attached to them, they seem repetitive.