ETL Project Report

08/10/2019  
Bryce Davis  
Amy Cleveland

Extraction   
For this project we looked into how US life expectancy and leading causes of death changed over time. We used two data sets; once NCHS table from Data.gov which listed life expectancy from 1900 -2015, broken down by race and gender. The second set is a CDC table from Kaggle which lists the leading cause of death in the US in 1980 and 2009. This was also broken down by race and gender. Both were in CSV format.   
  
Transformation

Both datasets were pulled into Jupyter Notebook using Pandas dataframes.

For the NCHS dataset we updated the data frame was pulled in with the index set to Year. We parsed out rows for the years 1980 and 2009. We then dropped all rows with any missing values and reset the index. Next we removed the all races and all sexes rows and dropped the age adjusted death rate column. Finally, we renamed columns to lowercase and removed all spaces.

For the CDC dataset, the race and gender information was in one column, called group. We set the index to the group column, then removed the Flag column and dropped all rows with missing values. We then used the split function to parse the group column into two separate race and gender columns, we kept the new columns and dropped the original group column. Next we standardized the race and gender names to match the NCHS data and reset the table index. Finally we, updated the column names to lowercase, with no spaces.

Load

We chose to store the datasets in a relational database because the two datasets have related data; both have year, race, and sex. Each row for each database is unique and the data within and between the tables is related to each other.

So in Postgress we created new database and two tables to hold the CDC and NCHS data frames. Then in Jupyter Notebook we created a connection to the database, and called a table names function to make sure we were connected to the right database and Postgres tables. We then appended the Postgres tables with the CDC and NCHS data frames. Next, in Jupyter Notebook and in Postgres we ran sample queries to make sure all data was imported to Postgres correctly.

Thoughts

The Postgres tables could be used to see how life expectancy and leading cause of death changed/didn’t change between 1980 and 2009. A query for that information could also include if the same factors changed for men and/or women, as well as black, and or white people. With more time, we could have pulled and cleaned a similar table of leading cause of death from WHO, so that we would have more years available to assess trends. We would also cast the year column in each data frame to date.