**ETL Project**

***Introduction/Explanation***

Every year, approximately 200,000 amputations of the lower extremities occur. It is popular belief that the most common form of amputation is secondary to trauma; however, the most common form of amputation is secondary to type II diabetes mellitus coupled with neuropathy. Type II diabetes is known as adult-onset or acquired diabetes which is largely attributed to lifestyle choices, and the densest populations with type II diabetes are in the southern states. Since type II diabetes is based on lifestyle choices, one can assume that those with higher incomes can afford to conduct “healthier” lives thus have lower incidents of type II diabetes. Given the above information, is there a correlation between the rate/count of amputations and the median income?

***Extract Phase***

Given this overarching question, I began to search for amputation data by state. I came across a CSV file that had the rate and count of amputation secondary to type II diabetes by county for Florida from 2016 – 2018. I then needed to find the median income data by county for the state of Florida for 2016 -2018; such data was located as a CSV file as well. Both sources came from http://www.flhealthcharts.com/charts/default.aspx.

To get a better look at the data, each source was read into a data frame utilizing Pandas and the read\_csv function. Once the data was read in, the data was then transformed.

***Transform Phase***

The first data source, Florida income data, was relatively “clean” after extraction and contained zero null values. However, the column names were generic, and thus were renamed to include “income” in the column name for increased clarification. The second data source, Florida amputation data, was less “clean” than the first data source, for there were a significant number of unneeded columns, rows with null values, and generic column names. The unneeded columns were removed, the rows with null values were removed, and the columns were renamed resulting in a clean data frame. Both data frames had a primary key column named “county” with identical values in each. This column was utilized as the key to join both data frames together into one large data frame. The combined data frame was then saved as a CSV file.

***Load Phase***

Since the resultant CSV file contained structured data, the best database to utilize was SQL. A new table was created in DBeaver/PostgreSQL and the CSV file was then imported into the table to create a table. If analysis were to be conducted on this data, one would use SQL to pull the desired data from the table and then conduct detailed analysis to determine whether or not there is a correlation between median income and the rate/count of amputation per county.