**Group Name:** ETLiens

**Members:** Kim Hogan, Daniel Emeri, Sebastian Di Francesco, Eddie Buckalew, JJ Krehbiel

**Project Title:** State-level Obesity and Fast Food Prevalence

**Project Description:**

Obesity is a major health concern in the United States, increasing people’s risk of several other chronic health conditions. Both dietary and economic factors can have major impacts on obesity levels, with cheaper food often having less nutritional value and a higher content in empty calories, fast food being a prime example. Therefore, for this project we compiled a database that includes state-level statistics related to obesity and the prevalence of fast food restaurants by geographic area.

The data we compiled for this project was gathered using a variety of methods. We scraped an HTML table on Wikipedia (sourced from the [*World Population Review*](https://en.wikipedia.org/wiki/Obesity_in_the_United_States#cite_note-tfah-1-70)) containing the obesity rates by state and downloaded two csv files from [Kaggle](https://www.kaggle.com/datafiniti/fast-food-restaurants) which came from Datafiniti’s Business Database containing a sample of 10,000 entries of fast food restaurants across the US.

**Project Report:**

**Extract**: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

1. The first source was a table we scraped from wikipedia (<https://en.wikipedia.org/wiki/Obesity_in_the_United_States>). The second and third sources were csvs (fastfood\_data1.csv and fastfood\_data2.csv).

**Transform**: what data cleaning or transformation was required.

1. For the table from wikipedia we filtered the data frame to show only the 50 states. Then we renamed the "States, district, & territories" column to “States” and the "Obese adults (2020)[78][73][79]" column to "Obese Adults (2020)".
2. For the two csvs after reading them into two data frames we then concatenated them together vertically. Then we had to find the extra state and remove it (District of Columbia). Next we created a dataframe that has states full names and states abbreviated names. After that we merged that data frame with the data frame that has fast food data to get the full names in there. Finally we dropped the “States” and “Province” column.

**Load**: the final database, tables/collections, and why this was chosen.

1. To load the data into mongoDb we had to merge the states\_obe\_rnk\_and\_Obe\_Adults\_2020 data frame with the cleaned\_combined\_fastfood\_data\_df\_no\_DC. Then we had to drop the column called “States Full Name”. To finally insert the data we converted the final\_data data frame into a dictionary (.to\_dict(orient='records') and inserted it into the all\_data collection (db.all\_data.insert\_many(data).
2. We chose to load the data into mongoDB because it simplified the data collections process by creating the schema for us so we did not have to create an external schema, and we wanted to strengthen our mongoDB skills since it is a newer database for us.

**Future Use:**

This new dataset could be used to compare rates of obesity per state and examine the correlations between obesity rates, fast food density, and median income by state.