**Final Report ETL Project**

**Team Members:** Patricia Mayer, Benjamin Nelson, Logan Fitzpatrick, Devynn Moreno

**Database:** Heart disease by state and factors that could affect heart disease (fast food restaurants, vehicle ownership)

**Data Sources:**

1. Thrillist <https://www.kaggle.com/mazharkarimi/heart-disease-and-stroke-prevention)>
   1. Data/content: Number of fast food restaurants per capita by state
   2. Technique: web scrape
2. Wikipedia <https://en.wikipedia.org/wiki/List_of_U.S._states_by_vehicles_per_capita>
   1. Data/content: Number of vehicles per capita by state
   2. Technique: web scrape
3. Kaggle Heart Disease and Stroke Pevention data <https://www.kaggle.com/mazharkarimi/heart-disease-and-stroke-prevention>
   1. Data/content: Provides a comprehensive image for cardiovascular diseases & related prevention
   2. Technique: CSV to pandas DataFrame
   3. Third dataset, because we’re overachievers ;)

**ETL Process:**

1. Thrillist:
   1. Thrillist reported count as fast food restaurants per capita, but after deeper investigation of the original data sources, it was discovered that the count was per 10,000 people.
   2. Thrillist had what seemed to be a list of the top 50 states arranged from least to greatest by number of fast food restaurants per capita. However, this was not a list it was actually a paragraph that they had formatted to look like a list.
      1. Split strings to transform the paragraph into multiple lists (Jupyter Notebook)
      2. Lists joined into dataframe
      3. State ID column added to dataframe
      4. Created two CSVs
         1. State ID with states
         2. State ID with fast food restaurant count per 10,000 residents
   3. 51 states were included, including Washington DC. Washington dc had to be checked for existence in the other databases before it could be included
2. Wikipedia:
   1. Scraped table from Wikipedia using pandas read\_html function
   2. Split table into two tables (2015 and 2018)
   3. Exported tables into two separated CSV files
   4. Imported CSVs into SQL
   5. Added primary keys in SQL to both tables
3. Kaggle
   1. Imported CSV into pandas dataframe using jupyter notebook
   2. Read through kaggle documentation and dropped two columns as a result of redundancy (columns shown below)
      1. LocationAbbr
      2. GeoLocation
   3. Sifted through columns to replace empty values with zeros. Had to ensure no zeros already present in order to not mix up zeros with nulls and skew the data
   4. Resaved dataframe as a CSV
      1. Creating separate CSV file allowed for modularization in combining separate data sources into one database
   5. Used import data wizard in SQL to import CSV
4. Loading onto Heroku
   1. Struggled with understanding how Heroku worked after some assistance, we were able to create a remote server via mySQL Workbench, had difficulty connecting from another machine but realized we were using the wrong variables for the default schema.
   2. SQL Alchemy needs extra installs to connect to MySQL databases vs SQLite databases

**Final Database:**

* Why it was chosen:
  + We decided to store our data in SQL. Because our data has state (state ID) as the common link across different data sources, we needed a relational database to link the data sources together. In addition, because we knew what structure we wanted to use and we do not plan on adding states, there is not much benefit in having an unstructured, more flexible database.
* Schema of tables/collections:
  + Created a total of five separate tables, all linked by State ID (see below). Did not join them because wanted to leave granularity and flexibility for the user

1. Heart disease data
2. Restaurant count data
3. States with State ID
4. Wikidata 2015
5. Wikidata 2018

* Hypothetical use cases for database:
  + One could see if there is a correlation between heart disease and number of fast food chains per capita by state. One could also potentially determine what states are the ‘healthiest’. By importing the state vehicle count per capita, one could potentially determine if a higher vehicle count potentially leads to higher heart risk due to lower recurring mobility.