**ETL Report**

**Roshni Chainani and Madeline Starr**

**Extract**

In order to compare the health ratings of the top five highest grossing fast food restaurants, as determined by QSR Magazine ([https://qsrmagazine.com/QSR-50](https://www.qsrmagazine.com/QSR-50)), across three cities, the publicly available data provided by each city in the form of CSV files were used as the primary sources (file sources listed below). These files were extracted into a Jupyter notebook using Pandas’ .read\_csv function.

CSV sources:

New York City: <https://opendata.cityofnewyork.us/>

LA: <https://catalog.data.gov/dataset/restaurant-and-market-health-inspections>

Austin: <https://catalog.data.gov/dataset/restaurant-inspection-scores>

**Transform**

After being extracted, each data file was transformed into a Pandas data-frame. Since each file contained large amounts of data that was not relevant to the project, all but three columns were dropped from each data-frame. The remaining columns contained the following information: restaurant name, the common name of the restaurant (ex: restaurant name: Subway #76432, common name: Subway), and the inspection score assigned to the restaurant. Using the .isin function, only the restaurants matching the common name of the 5 highest grossing fast food restaurants (McDonald’s, Burger King, Subway, Starbucks, and Taco Bell) were kept in each data-frame. All restaurants without a score were dropped from the data-frames used the .dropna function. Additionally, while most cities score their restaurants out of 100, subtracting points for infractions, New York City scores their restaurants by totaling up the infractions and only reporting that number. To bring the scores in line with the other cities, the NYC scores were subtracted from 100, creating a new Score column, to replace the old one. One data frame was created for each city with the restaurant name, the common name, and the inspection score. An index labelled, “ID,” was used as a unique identifier for each restaurant.

From this data, we then used the .groupby function on each table to calculate the mean inspection scores of each of the 5 fast food within each city as well as a count of the occurrences of each of the 5 restaurants in each city. These tables were then converted to data-frames and then merged for each city. Finally, the resulting three data-frames were merged, and the counts from each city were totaled to create a final data-frame that contained the common restaurant name, the average inspection scores for New York City, LA, and Austin, and the total count of restaurants included in the evaluation.

**Load**

After creating a database scheme to represent and match the columns of the data that had been transformed, a database was created in PostgreSQL using that schema. PostgreSQL was chosen because it is free and open source while still offering the standard SQL features.

The tables were all connected using the common name of the restaurant (“rest\_no”), which was the primary key of the summary table, and a foreign key in each set of the city-specific data.

Using SQLAlchemy in Jupyter Notebook, a connection was made to the existing database, and all four tables (top\_rest, NYC\_data, LA\_data, and Austin\_data) were pushed to the database using the .to\_sql function. pgAdmin was used to confirm the successful transfer of the data-frames into their respective tables in the final database.