GT Data Analytics Bootcamp Project 2

**Are Michelin star restaurant prices affected by the city they’re in – either due to population size or average income in the surrounding area?**

Data Extraction:

We found three data sources in Kaggle and extracted them into CSV files:

1. US Cities by Population (2020 census) - https://www.kaggle.com/datasets/axeltorbenson/us-cities-by-population-top-330?select=us\_cities\_by\_pop.csv
2. US Household Income Statistics - https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations
3. Michelin Restaurants for restaurant information (broken up into 3 files, one for each star rating) - https://www.kaggle.com/datasets/jackywang529/michelin-restaurants

As a result, there were 5 total CSV files (Kaggle\_Income.csv, us\_cities\_by\_pop.csv, one-star-michelin-restaurants.csv, two-star-michelin-restaurants.csv, and three-star-michelin-restaurants.csv).

Data Transformation:

We used Pandas to insert all the data into dataframes. We combined the three Michelin dataframes into a single dataframe, adding the appropriate star ratings.

Next, we inspected the regions in our Michelin dataset and realized there was more than United States data. We created a new dataframe that filtered to only United States data. We exported the consolidated Michelin dataframe for the US only into a new CSV file.

Data Loading:

We chose PostgreSQL to load and analyze our data because it lends itself to best handling the data available to us both for handling and analysis.

We began by creating a new database called Restaurant\_Analysis on our server using pgAdmin. We then created tables and columns based on our datasets – an income table, a population table, and a michelin\_restaurants table. We imported data into the tables using the corresponding CSV files.

In order to relate the tables, we created a fourth table for every city and state in our datasets. We created primary keys for each table. We added the city\_id based on the unique combination of city and state to each of the other tables in order to join properly and efficiently.

We created a query to look at the data from all tables. The results of the query are shown below:

****

There are many additional queries that could be run against this data because of the way we structured our tables. The above is just one example.