ETL Project Summary

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

My goal was to find vegetarian restaurants, including name, address, city, rating, and number of reviews. My data was from:

Kaggle/Zomato - World – CSV

Yelp API – Select US regions - JSON

• Transform: what data cleaning or transformation was required.

The Kaggle Dataset had to be cleaned, there were non-alphanumeric characters. I used this:

veg_data_df["Address"] = veg_data_df["Address"].str.replace('[^a-zA-Z],', '')

on each column.

The Yelp data was clean. There were some missing addresses, but as long as there was a name, city, rating, and number of reviewers, I wanted the data as those were the salient pieces of information.

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

It worked best to import into a MySQL database…the number of columns and rows are tabular and don't need the flexibility of something like Mongo.