**Team9-ETL-Project**

**Introduction**

Collaborators: Brandon V., Kyle H., Mark P.

Los Angeles Restaurant & Market Health Data

Obtained from [www.kaggle.com](http://www.kaggle.com) <https://www.kaggle.com/cityofLA/la-restaurant-market-health-data/version/22>

Data is maintained by Kaggle and last updated 3 months ago.

json file for Inspections

csv file for Violations

Data Sources need to be stored in a dataframe, cleaned up and merged.

**Methodology**

In jupyter, we imported pandas, json, and pymysql.

Set up a config.py file to store our mysql passwords in.

Both files were originally in CSV. Went to website they were pulled from and accessed Inspection file in json and created the file.

In mysql, we created a database. Did a “use” command to access database. Then constructed tables to utilize in the merger of the two datasets.

Utilized create\_engine from sqlalchemy to establish a connection with the sqlite database.

Created variables in jupyter to call out the location of each dataset. Then utilized pandas in translating the files into dataframes, with a .head() to view/verify content.

Called the engine table names in jupyter to verify connection to mysql.

Analyzed data and deleted unnecessary columns/information, renaming columns in preparation for merger.

Performed merger of the csv and json files utilizing unique information “serial\_number” under column names.

Used connection\_string and create\_engine to establish connection to mysql database and the tables within for merging

In jupyter, using pandas, performed a query to read the information in sql.

Most difficult part we had was in performing the merger:

The restaurants had multiple rows of information regarding both inspections and violations. We dropped duplicate rows of serial numbers in both files. Renamed columns, then used only those columns being utilized in the merger.

Had difficulty in moving the final datasets from jupyter notebook into mysql, because of the file size (even after deleting a lot of information). Ended up pulling parts of it at a time using chunksize and pool\_recycle to finally move it to sql so the files could be merged together.

There were several obstacles we encountered, which were new to all of us. Together, as a team, they were overcome.