NYC Restaurant Inspection

Abstract

The goal of this project is to incorporate concepts on building an end-to-end data storage along with a processing pipeline to create an interactive web application to showcase users on information about NYC restaurants on how they are graded and scored in terms of food safety. It aims to provide any local or a potential traveler who wants to locate areas of interest when deciding on where to eat out or take out from a restaurant. The pipeline consists of data acquisitions from API calls, preprocessing by cleaning data using pandas and numpy, storing data using SQL database, processing data by loading and reading queries onto pandas, and deploying locally using Streamlit for the web app. The web application includes two pages: the main page provides an interactive map to get results on location and its information, the second page provides an interactive barchart to select a specific year where top 5 cuisines are visualized depending on the scores received from inspection.

Design

The project utilizes NYC Open Data Source on DOHMH New York City Restaurant Inspection Results. This project seeks to give awareness to the public on the potential dangers when eating at a food establishment that has had violations previously and ongoing. Users of interest are those who have had a bad experience eating at a restaurant or take out where they contracted symptoms related to foodborne illnesses. It is noted by CDC, roughly 48 million individuals get sick from foodborne illness and about 128,000 are hospitalized[1]. Therefore the project aims to establish a user friendly web application to visualize food establishments that have been inspected with information on grade and score received.

Data

The dataset comprises 248,020 restaurant violations with 26 features from the years of 2015-2022. Out of the 26 features only 9 features were analyzed to create interactive visualizations. It is noted on the DOHMH New York City Restaurant Inspection Results page, where the data was retrieved that a large portion of data contains either missing or information that does not give value toward analysis. After removing these discrepancies a total of 66,174 rows of data were retained.

Algorithms

Data Ingestion:

Data was collected using SODAPY API from the NYC Open Data Source. Google Map Platform API and Geopy API were applied to obtain latitude and longitude coordinates that were not available from the primary data source.

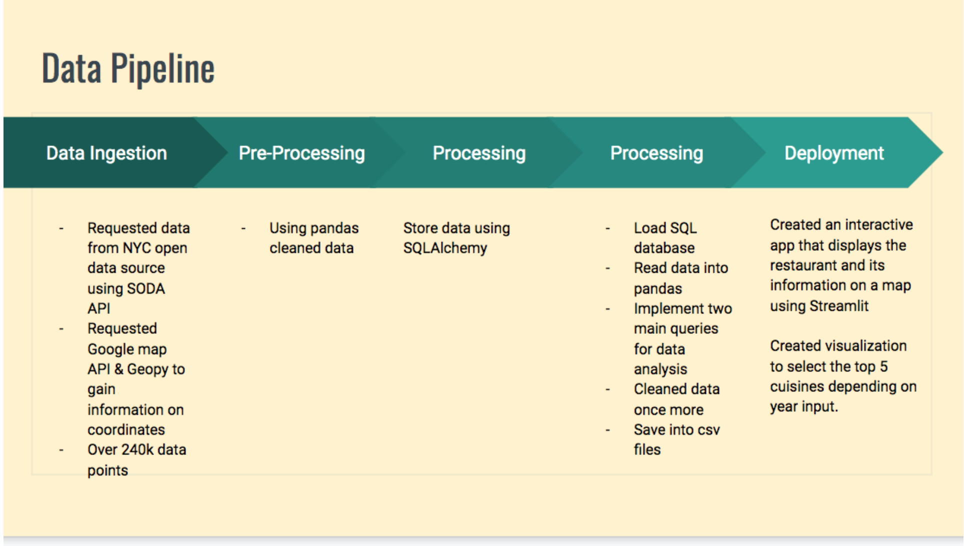
[1] https://www.cdc.gov/foodsafety/foodborne-germs.html

Data Preprocessing:
Data was cleaned accordingly by pandas and numpy. Cases with missing values were imputed using median values appropriately.

Data Processing:
The cleaned data was stored using a relational database, SQL and its Object Relational Mapper SQLAlchemy which later was loaded and read by pandas. Queries were implemented for analysis and visualizations in the web application.

Deployment:
Streamlit, folium, matplotlib and seaborn were used to create visualizations on the web application.

Pipeline Diagram:



Data Pipeline

Data Ingestion

- Requested data from NYC open data source using SODA API
- Requested Google map API & Geopy to gain information on coordinates
- Over 240k data points

Pre-Processing

- Using pandas cleaned data

Processing

Store data using SQLAlchemy

Processing

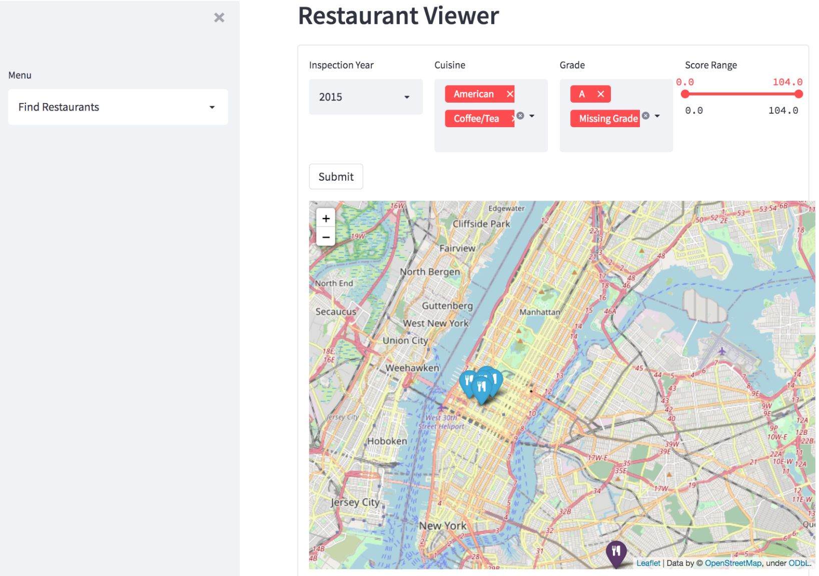
- Load SQL database
- Read data into pandas
- Implement two main queries for data analysis
- Cleaned data once more
- Save into csv files

Deployment

Created an interactive app that displays the restaurant and its information on a map using Streamlit

Created visualization to select the top 5 cuisines depending on year input.

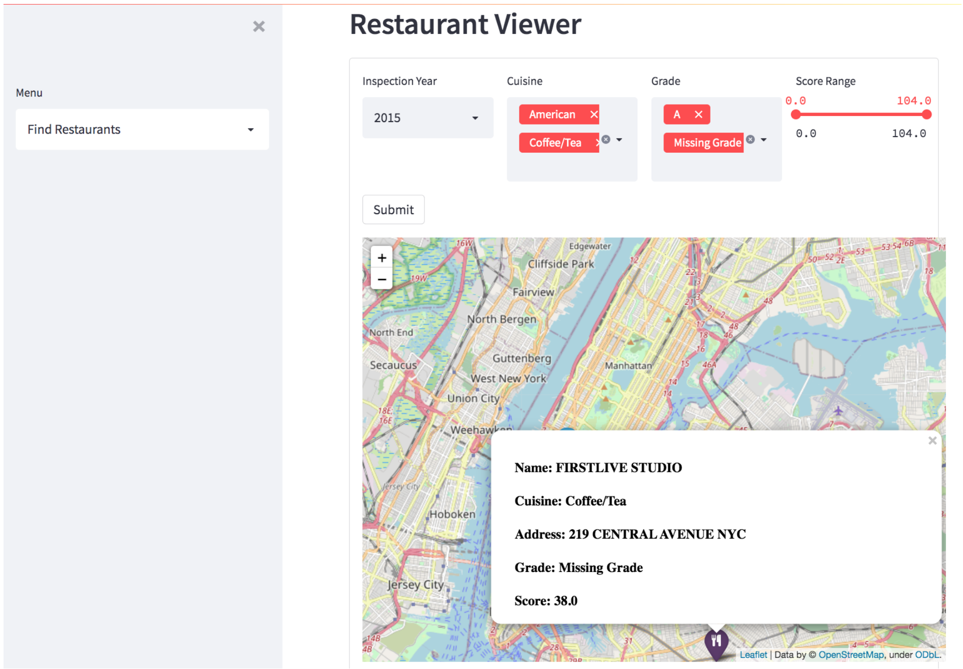
Web Application:



Menu
Find Restaurants

Restaurant Viewer

Inspection Year: 2015
Cuisine: American, Coffee/Tea
Grade: A, Missing Grade
Score Range: 0.0 – 104.0



Restaurant Viewer

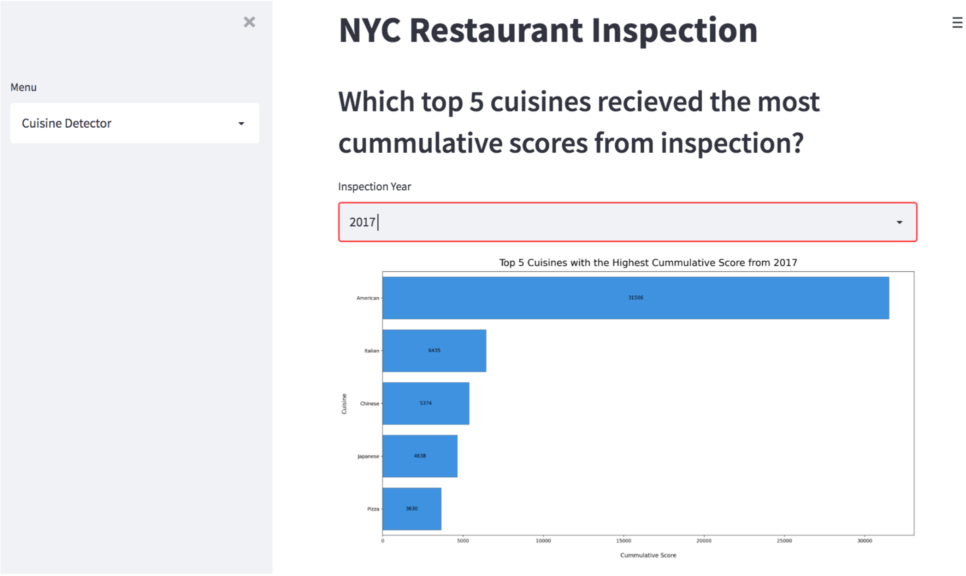
Name: FIRSTLIVE STUDIO

Cuisine: Coffee/Tea

Address: 219 CENTRAL AVENUE NYC

Grade: Missing Grade

Score: 38.0



NYC Restaurant Inspection

Which top 5 cuisines recieved the most cummulative scores from inspection?

Tools

- Used SODAPY API
- Google Map Platform API
- Geopy API
- Pandas and Numpy for data cleaning/ manipulation
- SQLAlchemy for data storage
- Folium, Matplotlib, Seaborn for data visualization
- Streamlit for web app deployment and interactive visualization

Communication

After local deployment, I will use the recommended steps from Streamlit to deploy the web app onto my github.