Dish Detective

NYC Restaurant Inspection Analysis

Mihir Chhatre (mc9164)
Nachiket Khare (nk3559)
Amey Kolhe (apk9563)

Why?

Open portal that can foster collaboration between inspection officers, restaurant owners, regulators to improve diner's experience and ultimately the protect public health.

Objectives

- Understand how restaurant violation results have changed over the years and across seasons.
- Violation codes and their relationship with factors like cuisine type, borough, season.
- Discover patterns across a restaurant chains.

Dataset

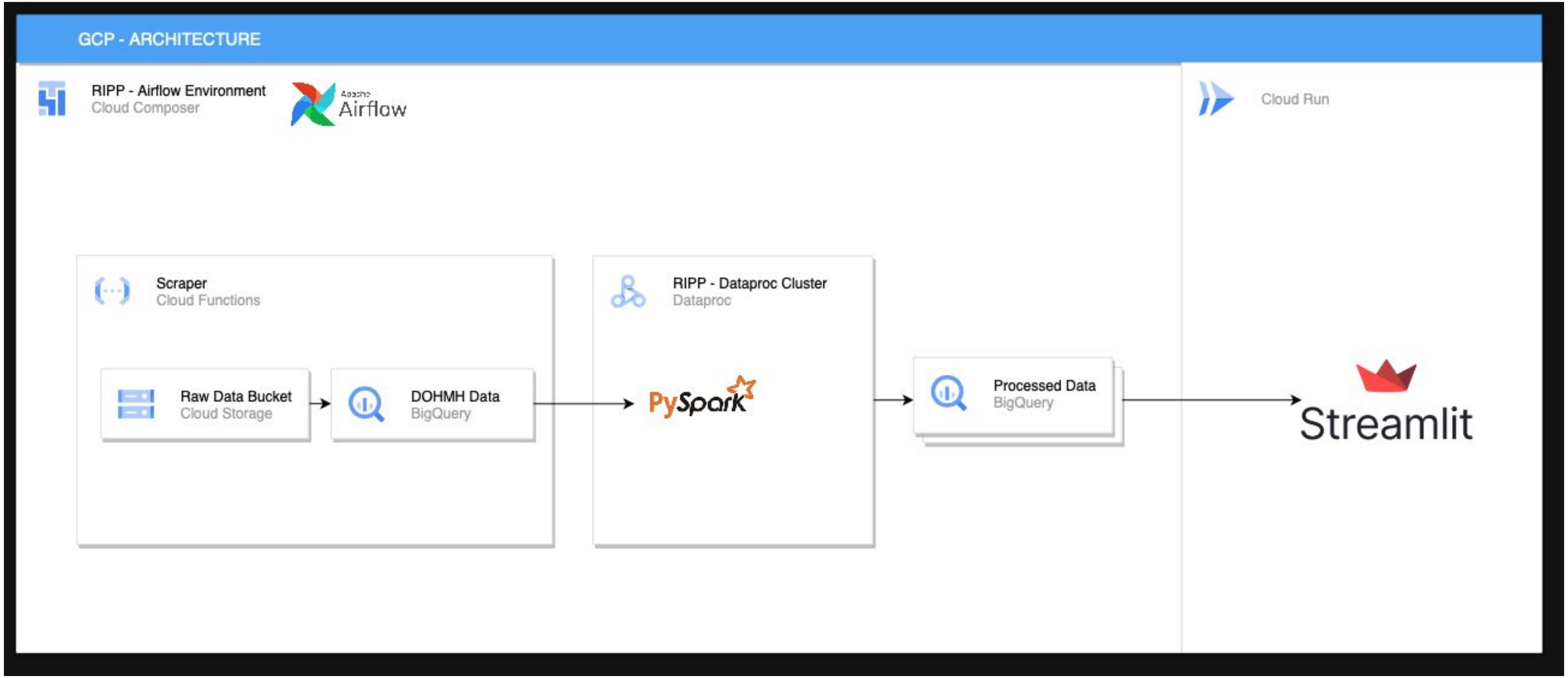
DOHMH New York City Restaurant Inspection Results:

- Inspection results for restaurants across NYC
- Source: NYC Open Data
- ~210K rows, 27 columns
- API endpoint available

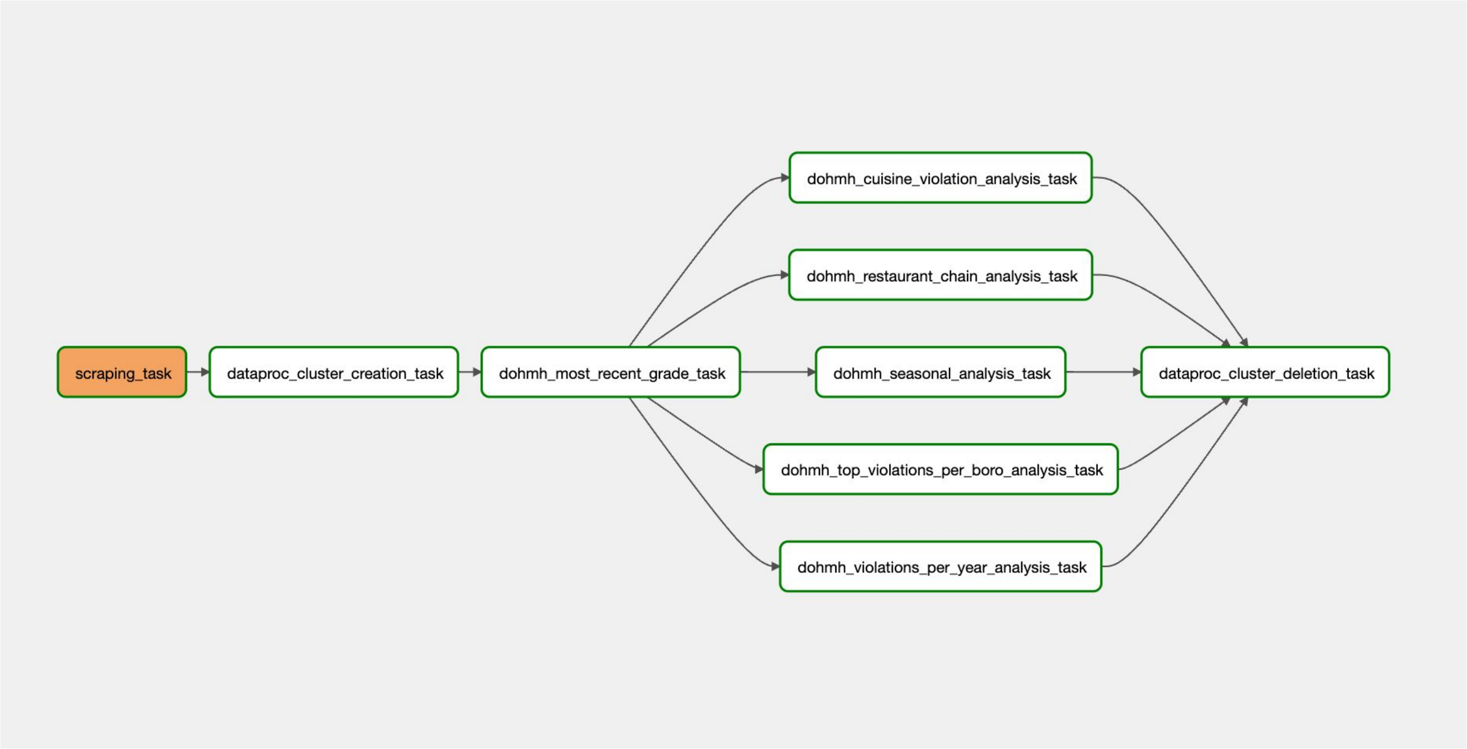
Keys points to note:

- *Conditional* temporal analysis is allowed.
- Violation codes, scores & grades(may) are assigned after inspection.
- API record limit of 50K per request (Requires Pagination)

Architecture



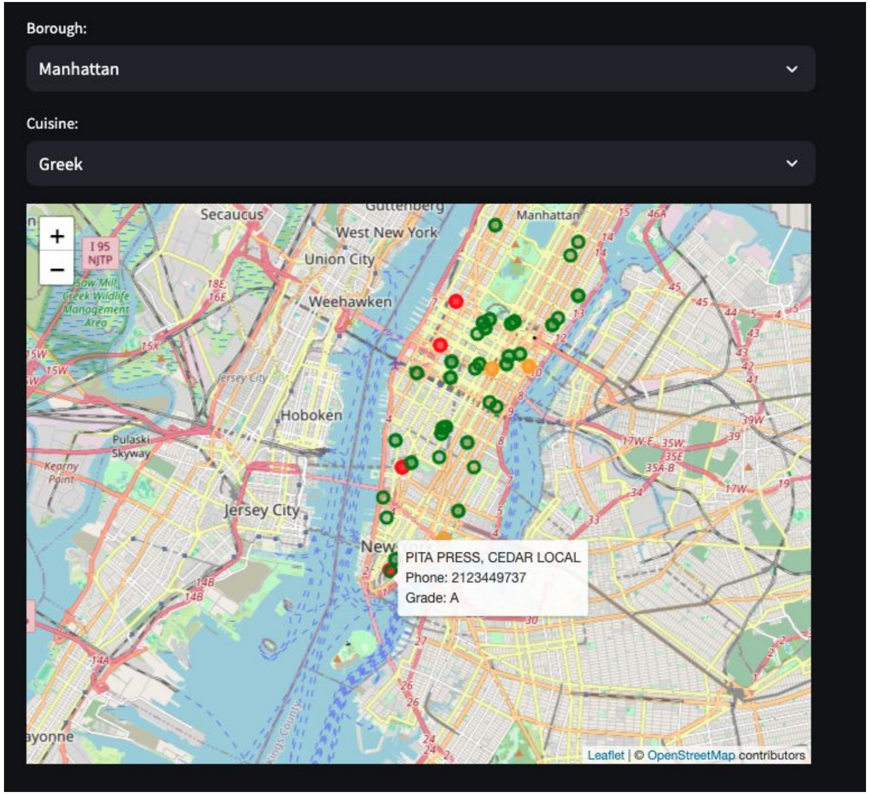
Airflow DAG


scraping_task
dataproc_cluster_creation_task
dohmh_most_recent_grade_task
dohmh_cuisine_violation_analysis_task
dohmh_restaurant_chain_analysis_task
dohmh_seasonal_analysis_task
dohmh_top_violations_per_boro_analysis_task
dohmh_violations_per_year_analysis_task
dataproc_cluster_deletion_task

Analysis

Geospatial analysis of the most recent grade given to a restaurant.

- Find the most recent inspection results for each restaurant based and filter using instructions defined by NYC Open Data (such as score value, type of instruction and grade allowed).
- Using this ‘most_recent_letter_grade’ dataframe build a map using ‘folium’
- ‘Most_recent_letter_grade’ is used as the reference dataframe for all future analysis.



Borough: Manhattan

Cuisine: Greek

PITA PRESS, CEDAR LOCAL
Phone: 2123449737
Grade: A

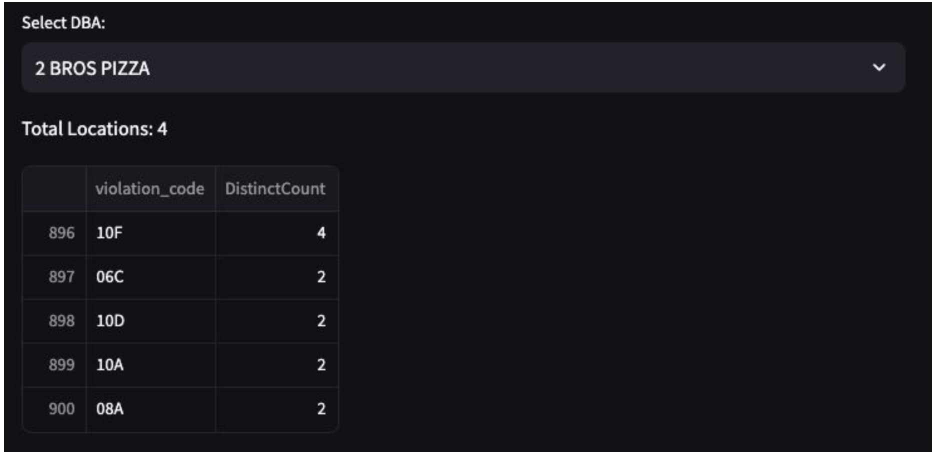
Analysis

Restaurant chain analysis

- Chain-Specific Data Collection

- Compiling data for each restaurant chain (DBA) to analyze inspection patterns across multiple locations.

- Violation Code Frequency per Chain Assessing which violation codes were inspected and how frequently they occurred across various locations within each restaurant chain



Select DBA:

2 BROS PIZZA

Total Locations: 4

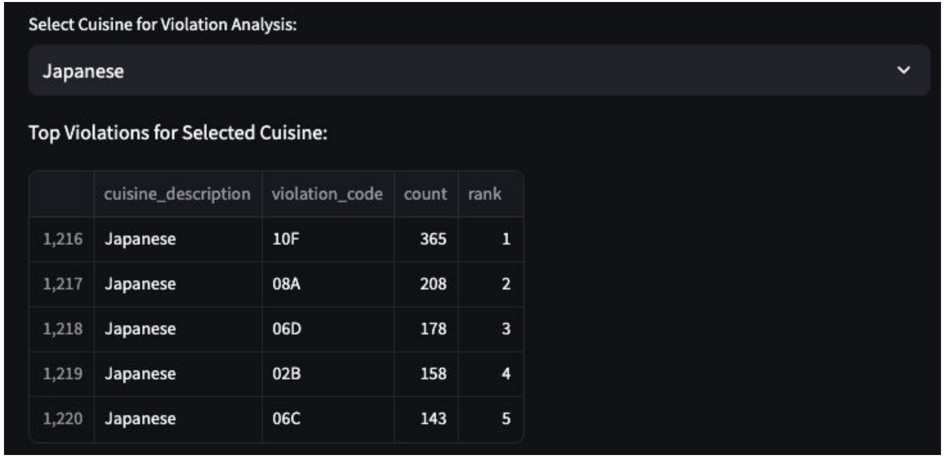
	violation_code	DistinctCount
896	10F	4
897	06C	2
898	10D	2
899	10A	2
900	08A	2

Analysis

Ranking top 5 violation codes across different cuisines

- Quantifying occurrences for each specific violation code across different cuisine types, providing insights into the frequency of particular violations.

- Ranking violations for cach cuisine to identify most common health code infractions for cuisine type



Select Cuisine for Violation Analysis:

Japanese

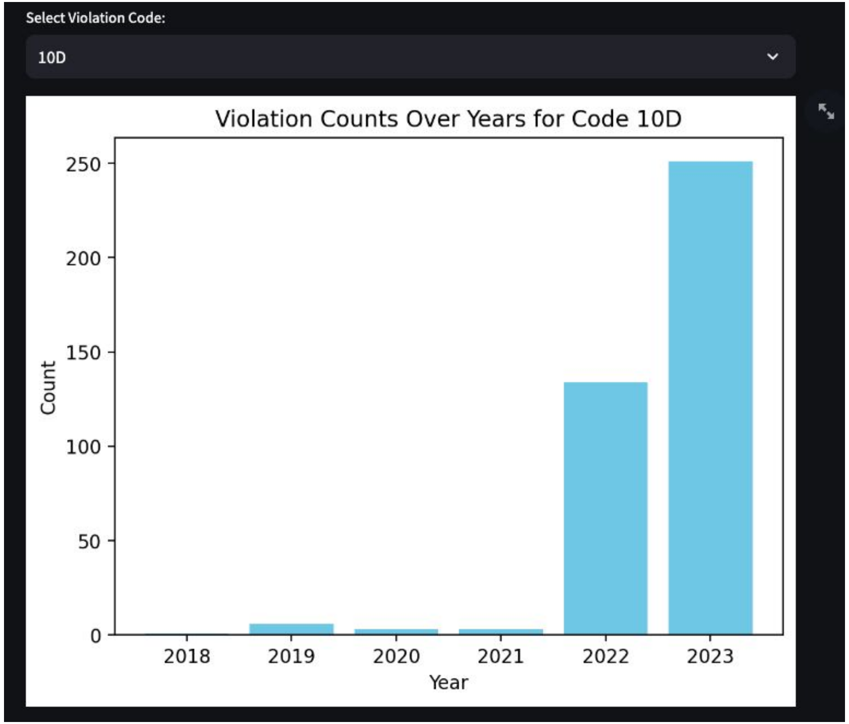
Top Violations for Selected Cuisine:

	cuisine_description	violation_code	count	rank
1,216	Japanese	10F	365	1
1,217	Japanese	08A	208	2
1,218	Japanese	06D	178	3
1,219	Japanese	02B	158	4
1,220	Japanese	06C	143	5

Analysis

Tracking trends across years for violation code

- Grouping data by violation code and inspection year
- Calculating the total number of occurrences for each violation code in each year

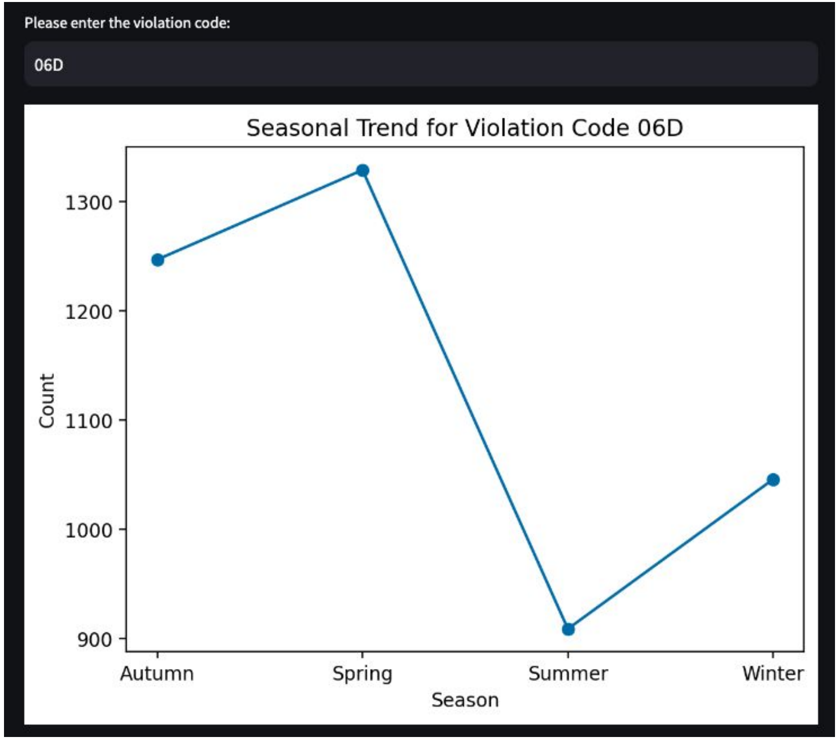

Select Violation Code:

10D

Analysis

Seasonal trends across violations

- Categorizing months to seasons - Winter, Spring, Summer, and Autumn.
- Categorization is applied to through a User Defined Function (UDF), adding a "SEASON" column to the dataset, which maps each inspection month to its corresponding season.
- Finally group by season for the entered violation code.



Please enter the violation code:

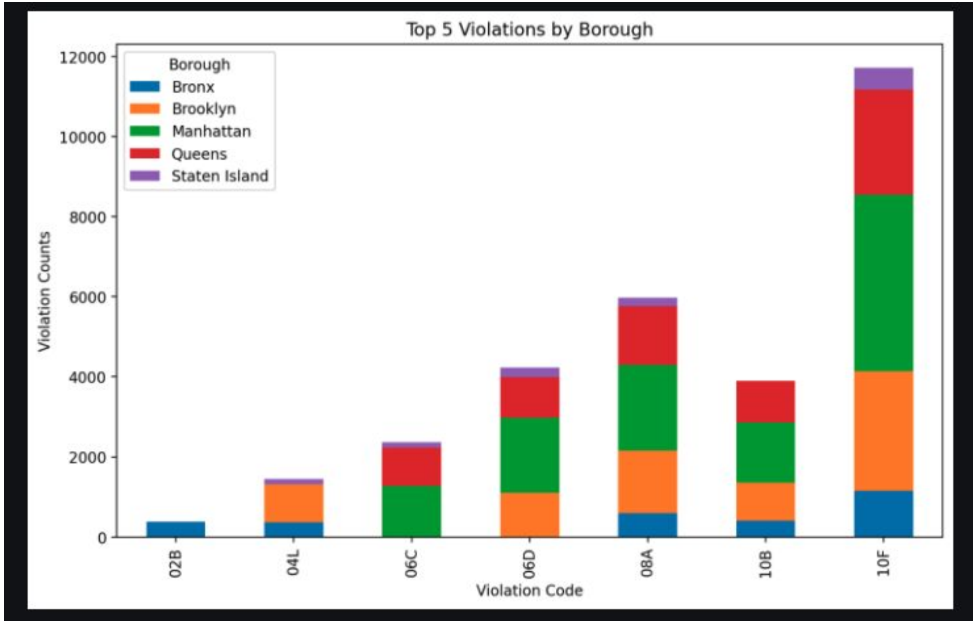
06D

Analysis

Identifying top five violations across Boroughs

- Grouping by violation code and borough.

- Window function to rank the top five most common violations partitioned by borough.



Demo

https://dishdetective-lkdh4wts6q-ue.a.run.app/

Future enhancements

1. Build fault tolerance into the pipeline.

2. Consider using different file formats (Parquet?) for archiving data.

3. Continued scrapping will allow us to build a unified historical data warehouse.

4. Address stop words within DBA and handle cuisine misclassification in source data such as ‘Pizza’ instead of ‘Italian’.

Thank you!

Q&A