Analyzing NYC Restaurant Inspections: Patterns and Predictions

Name: Mihir Chhatre
NetID: mc9164
Course Name: Big Data
Section: D
Semester: III

Name: Nachiket Khare
NetID: nk3559
Course Name: Big Data
Section: D
Semester: III

Name: Amey Kolhe
NetID: apk9563
Course Name: Big Data
Section: D
Semester: III

Project Abstract

Our project aims to investigate the correlations between restaurant inspection results and a range of factors, including cuisine type, borough location, violation codes and other pertinent attributes. Additionally, we plan to conduct network analysis on restaurant chains operating multiple branches within New York City. This would help us delve into the interrelationships among these branches and identify significant patterns and anomalies within the network. Lastly, we aim to cluster restaurants based on their violation profiles to reveal overarching trends and patterns of non-compliance within the restaurant industry.

Problem statement

In New York City, where the restaurant industry thrives and forms an integral part of the city's culture, maintaining health and safety standards is paramount. While the Department of Health and Mental Hygiene (DOHMH) conducts regular inspections of these establishments, there remains a lack of in-depth analysis that ties together multiple factors contributing to restaurant inspection outcomes. This gap inhibits targeted regulatory interventions, restaurant management improvements, and effective public health campaigns.

Understanding the patterns and correlations between restaurant inspection results and multiple variables such as cuisine type, borough location, and specific violation codes is crucial. Additionally, for large restaurant chains that operate several branches within the city, there is a need to assess if and how violation patterns are consistent or divergent across their network.

This project aims to address these gaps. Specifically, the primary questions we seek to answer are:

- How do restaurant inspection results correlate with different factors like cuisine type, borough location, and violation codes?
- Are there discernible patterns or anomalies in violation results among restaurant chains across multiple branches within New York City?
- Can restaurants be categorized into clusters based on their violation profiles, and if so, what insights do these clusters reveal about non-compliance trends in the industry?

Answering these questions would provide stakeholders, from regulators to restaurant owners, with actionable insights to enhance compliance, improve restaurant safety, and ultimately protect public health.

Objective

The primary objective of our project is to leverage big data analysis to enhance transparency regarding restaurant safety in New York City. Our goal is to provide the general public with valuable insights into the safety conditions of restaurants in their vicinity. Through this initiative, we aim to educate and alert consumers about the safety and compliance status of restaurants, enabling them to make more informed dining choices.

Data

The '*DOHMH New York City Restaurant Inspection Results*' dataset includes NYC restaurant inspection results for up to three years prior to the most recent inspection. The purpose is to provide information on recent inspection results. Restaurants that go out of business are removed. In addition, restaurants can choose to go through the adjudication process, i.e., argue their case at an administrative hearing. Therefore, this dataset is not appropriate for historical analyses of NYC restaurant inspections that compare previous years of data to the current data. The data is refreshed on a daily cadence.

Name:

DOHMH New York City Restaurant Inspection Results

Link:

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Size:

87.3 Mb

Records:

208,000 records (each record is a restaurant citation)
27 attributes

Attributes:

1. CAMIS: A unique 10-digit identifier for each restaurant, static per permit.
2. DBA: Represents the public business name (or "doing business as") of the restaurant. This name can change at the restaurant owner's discretion.
3. BORO: The borough where the restaurant is located. Options include MANHATTAN, BRONX, BROOKLYN, QUEENS, and STATEN ISLAND. There may be discrepancies between the zip code and listed boro due to differences in mailing addresses and physical locations.
4. BUILDING: The building number where the restaurant is located.
5. STREET: The street name of the restaurant's location.
6. ZIPCODE: The zip code of the restaurant's location.
7. PHONE: Phone number provided by the restaurant owner or manager.
8. CUISINE DESCRIPTION: Describes the restaurant's cuisine type. This is an optional field.
9. INSPECTION DATE: Represents the date a restaurant was inspected. A date of 1/1/1900 means it hasn't been inspected yet.
10. ACTION: Describes the outcome or actions of the inspection. Possible outcomes include violations cited, no violations, reopening by DOHMH, re-closing by DOHMH, and others.
11. VIOLATION CODE: The specific code associated with any violations found during inspection.
12. VIOLATION DESCRIPTION: A detailed description of the violation associated with the given code.
13. CRITICAL FLAG: Indicates the severity of a violation. Options include Critical, Not Critical, and Not Applicable. Critical violations are more likely to lead to food-borne illnesses.
14. SCORE: The total score assigned after a particular inspection, which may be updated based on adjudication results.
15. GRADE: The grade associated with the inspection's outcome. Options include Not Yet Graded, Grade A, B, C, Pending, and Pending for re-openings after initial inspections resulting in closure.
16. GRADE DATE: The date when the current inspection grade was assigned to the restaurant.
17. RECORD DATE: The date when the dataset was last updated or extracted.
18. INSPECTION TYPE: Describes the nature of the inspection. This is a combination of the inspection program and the type of inspection carried out.

Overall, this dataset offers a comprehensive overview of restaurant inspections, including their outcomes, violation details, and grading.

To stay updated with the daily releases from the "DOHMH New York City Restaurant Inspection Results" dataset, we will use the Socrata Open Data API (SODA) as this would ensure a more stable data extraction process. The primary aim will be to configure the scraping job to discern and retrieve only the newest data, keeping in mind the dataset's sliding window characteristic. Then to store this data efficiently, implementing mechanisms for error management and elimination of duplicates, and consistently back up data to safeguard against the dataset's evolving nature.

Proposed technologies & Programming language

Firstly, data acquisition and scheduling scraping jobs will involve ***Python*** and ***Airflow*** or ***CRON*** jobs running on ***GCP Compute*** instances or ***Google Cloud Functions (serverless)***. For further preprocessing, cleaning and transforming the dataset into a format suitable for analysis we can utilize ***Apache Spark*** running on Google ***Dataproc*** cluster.

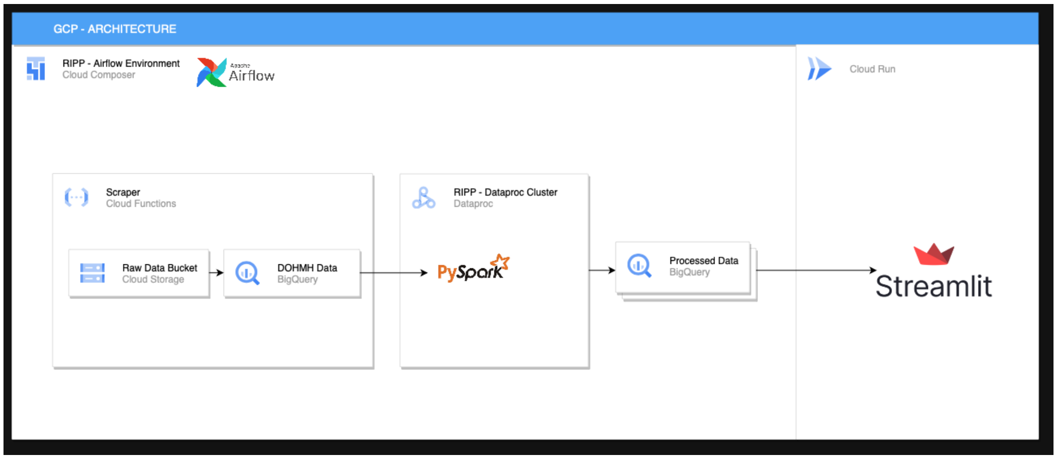
Coming to storage of raw data files we plan to use ***GCS buckets*** and later use Google ***BigQuery*** to store the processed data to be used for Dashboarding purposes.

To discern patterns or anomalies among restaurant chains and correlate inspection results with various factors we plan to use ***Spark ML*** Libraries.

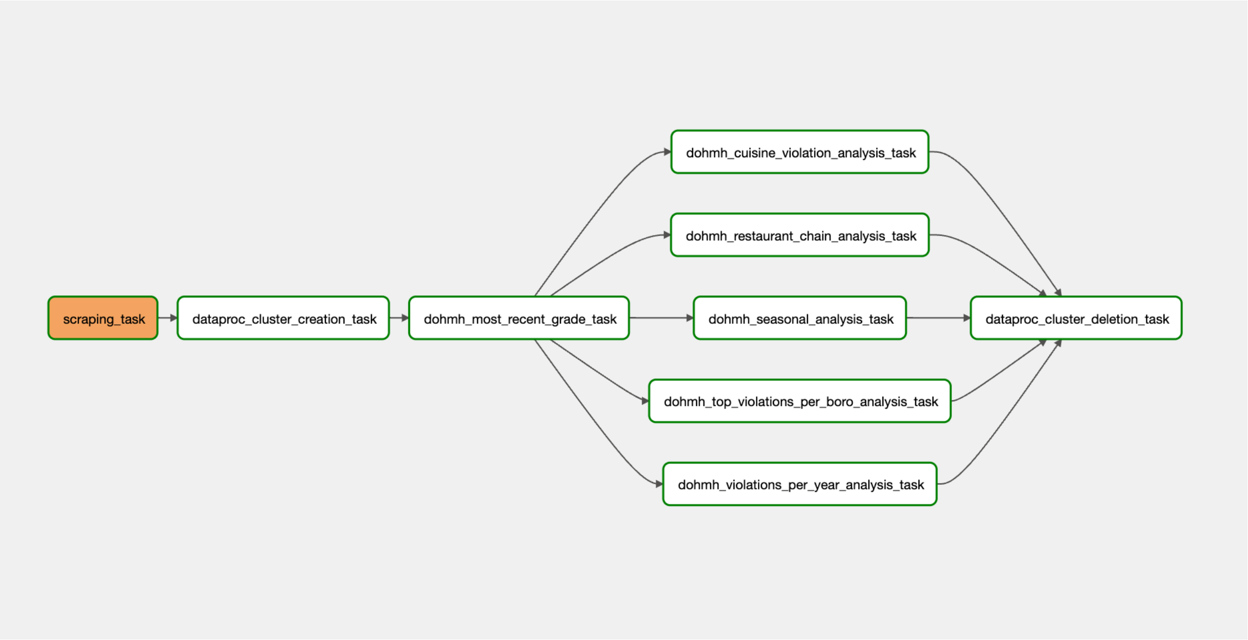
Lastly, an interactive dashboard can be developed using tools like ***Tableau, Power BI or Streamlit,*** allowing stakeholders to explore the data, see real-time correlations, and make informed decisions. The integration of machine learning models into this dashboard will provide predictive capabilities and insights on demand.

Tools and Technologies to be used - **Python, Cloud Composer, Cloud Compute, Cloud Functions, Cloud Run, GCS Buckets, BigQuery, Apache Spark, Streamlit for Data Visualization.**

System 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 Cases

1. Geospatial analysis of the most recent grade given to a restaurant.
    a. 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).
    b. Using this ‘most_recent_letter_grade’ dataframe build a map using ‘folium’
    c. ‘Most_recent_letter_grade’ is used as the reference dataframe for all future analysis.
2. Restaurant chain analysis
    a. Chain-Specific Data Collection
    b. Compiling data for each restaurant chain (DBA) to analyze inspection patterns across multiple locations.
    c. Violation Code Frequency per Chain Assessing which violation codes were inspected and how frequently they occurred across various locations within each restaurant chain
3. Ranking top 5 violation codes across different cuisines
    a. Quantifying occurrences for each specific violation code across different cuisine types, providing insights into the frequency of particular violations.
    b. Ranking violations for cach cuisine to identify most common health code infractions for cuisine type

4. Tracking trends across years for violation code
    a. Grouping data by violation code and inspection year
    b. Calculating the total number of occurrences for each violation code in each year
5. Seasonal trends across violations
    a. Categorizing months to seasons - Winter, Spring, Summer, and Autumn.
    b. Categorization is applied through a User Defined Function (UDF), adding a "SEASON" column to the dataset, which maps each inspection month to its corresponding season.
    c. Finally group by season for the entered violation code.
6. Identifying top five violations across Boroughs
    a. Grouping by violation code and borough.
    b. Window function to rank the top five most common violations partitioned by borough.

Project Demo

Link - https://dishdetective-lkdh4wts6q-ue.a.run.app/
Code - https://github.com/Ameykolhe/restaurant-inspections/tree/main

Conclusion

Our project successfully analyzed the complex factors affecting restaurant inspection outcomes in New York City, utilizing the DOHMH dataset with over 210k records. We identified correlations between inspection results and variables like cuisine type, borough location, and violation codes, and conducted network analysis on restaurant chains to detect patterns and inconsistencies across branches. Our approach of clustering restaurants based on violation profiles revealed significant non-compliance trends, offering valuable insights for regulatory bodies, restaurant owners, and consumers. Leveraging Python, Apache Spark, and advanced visualization tools, we transformed extensive data into actionable insights, enhancing transparency and safety in the NYC restaurant industry.