CS598 Final Report

Wenyan Zhang

Project Introduction and Motivation

New York City is a global hub for innovative restaurants and diverse cuisines from around the world. There are approximately 18,000 restaurants in NYC (Restroworks), so ensuring food safety for both New Yorkers and millions of visitors is crucial. The New York City Department of Health and Mental Hygiene (DOHMH) publishes detailed restaurant inspection records containing every violation citation for active restaurants and college cafeterias in NYC on NYC Open Data to encourage people to engage with information produced and used by the City government (NYC Open Data).

As a New Yorker myself, I am motivated to explore the data to understand patterns in food safety across different boroughs. The goal of this project is to curate, clean, and analyze this inspection dataset using methods from the course. This dataset presents real-world problems in data collection and curation processes, including inconsistent text, missing values, and a lack of data context. Therefore, this project will use data curation techniques to transform raw data provided by the government agency into a clean dataset for data analysis.

The main research questions are:

1. Borough-level comparison: How do food safety results differ in the five boroughs? Is it related to demographic profiles such as population density?
2. Violation patterns: Do restaurants generally improve after multiple inspections?
3. Cuisine-specific analysis: Are some cuisines more likely to receive critical violations or bad grades?
4. Seasonality: Are there seasonal or holiday-related spikes in food safety violations?

The questions are exploratory analysis, and the final curated dataset and workflow will enable reproducibility and standardization for more in-depth research in the future.

Dataset Profile

The main dataset used throughout the project is the DOHMH New York City Restaurant Inspection Results (NYC Open Data). It contains 27 columns and this project will provide a data dictionary with definition to support independent understanding. It also contains approximately 293K rows that are growing and updated daily, which can be downloaded as Excel file from the NYC Open Data website or via API. The data in the project is extracted via API so that the data can be updated everything the code is executed.

Another dataset used to supplement the data analysis is the Population FactFinder dataset from NYC Planning (NYC City Planning). This data will support the borough-level analysis since it contains borough-level population data.

In terms of ethical considerations, both datasets are public and contain non-sensitive information. However, the DOHMH data includes restaurant addresses, requiring considerations around inference of business owners. Therefore, the analysis of this project will focus on granular levels, such as borough and cuisine-level.

Data Curation Workflow

The data curation workflow for this project is organized as a modular folder structure in the GitHub repository. The workflow integrates raw data acquisition, cleaning, enrichment, and analytical processing to ensure that each step is reproducible and aligned with best practices in data curation.

All raw and curated data are stored in the “data” folder. The data.txt contains the original dataset from the NYC Open Data API and the curated dataset. The data is stored in .txt format with Box links to download the dataset due to the size is too large to be supported by GitHub. Data cleaning and transformation procedures are implemented within the workflow notebooks stored in the “workflow” directory. The data_pipeline.ipynb notebook performs core operations such as standardizing categorical fields, resolving missing values, and deriving analytic features like seasonality variable and critical violation flag.

Documentation and metadata are stored in the “documentation” directory. This includes a comprehensive data dictionary (data_dictionary.csv) with all fields and value types if applicable. It also includes a metadata.json file following the Schema.org Dataset specification. These documentations ensure that future users understand the provenance and transformations applied to the curated dataset.

The analysis and reporting stage is recorded in the data_analysis.pdf file in the “workflow” directory. It summarizes key findings from all the research questions, such as borough-level inspection patterns, cuisine-specific trends, grade improvement behavior, and seasonality patterns.

All the artifacts together form a complete workflow from acquisition to documentation and analysis while remaining reproducible through the provided notebooks and metadata.

Relation to Data Lifecycle Model

The project uses the USGS Science Data Lifecycle Model (SDLM) to guide an end-to-end data curation workflow, as it follows the guidelines of SDLM:

1. Plan: Come up with research questions to guide data curation and analysis, understand constraints of selected datasets, and design the workflow.
2. Acquire: Use the provided API and CSV download to source data.
3. Processing: Cleaning, standardization, and validation of data using the data_pipeline.ipynb notebook on Google Colab.
4. Analysis: Perform exploratory data analysis and other data analysis to answer all the research questions.
5. Preserve: Store the raw and cleaned datasets externally on Box for user reuse and download

6. Publication: Publish all artifacts on the GitHub repository and document key requirements in README.md.

Summary

From the data analysis, these are some major findings:

1. Queens stands out as the borough with the lowest food safety grades.
2. There exist temporal patterns in inspections by season. Summer is the season with the worst ratings.
3. The cuisine with the worst rating is inconclusive because food inspection results have different dimensions - grade and critical violation rate. The analysis in this project presents rankings for both aspects so that readers can evaluate themselves.

However, to achieve these conclusions, I also faced some challenges:

1. Both the raw and cleaned data CSV files are too big (>140MB), which caused GitHub push issues. This was addressed by using the API to retrieve raw data and document the full steps of cleaning up data so the user can reproduce the clean data themselves. In addition, the CSV files were uploaded to Box as a public shared file to allow user downloads. In this way, the user can retrieve the datasets in both ways, regardless if they will run the full script to perform data cleanup.
2. Another challenge was finding a supplemental demographic dataset to support more analysis. Initially, I was focusing on official census data and found it hard to map zip codes to community districts and back to boroughs, but it was not possible because a zip code can be in multiple community districts. After conducting more research, I found a website that contains population data grouped by borough, published by the NYC Department of City Planning, which is also a government agency. This is a credible data source with borough-level data matching my research question, so it was the perfect dataset to use.

From this project, I learned that:

1. Even though the dataset is on the public NYC Open Data website with a data dictionary, data curation is still necessary since it is important for user needs time to understand how to retrieve the data and the refresh frequency. The metadata and code script are essential for future reproduction.
2. The data cleaning step was crucial since early standardization of text and new category creation reduced variation and helped with data analysis.
3. There is a lot more that can be done for this research-ready dataset. There is a lot of geographical information, so more analysis of geospatial clustering and socioeconomic status of unsafe hotspots can be done.

References

1. Restroworks. *New York City restaurant industry statistics*. https://www.restroworks.com/blog/new-york-city-restaurant-industry-statistics/
2. New York City Open Data. *DOHMH New York City restaurant inspection results*. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
3. New York City Department of City Planning. *Population FactFinder: New York City*. https://popfactfinder.planning.nyc.gov/explorer/cities/NYC