Final Project:

Chicago Restaurants Health Inspection Analysis

BAIS:3250, Data Wrangling

Aidan Stanik, Kylie McCabe, Alyssa Smith

1. Introduction

Chicago restaurants have a high volume of foot traffic. What may not be known by the large majority is the public health inspections in respect to highly recommended restaurants. Our project aims to integrate the two. The city of Chicago has made their inspections available to the public. We will condense this data and align it with the top restaurants. Customers can utilize this dataset to find the best rated restaurants in the city and also evaluate their health inspection data. The health inspections will measure compliance with sanitary standards, food handling and other regulations. Online reviews like blog posts are customer-driven evaluations and will highlight numerous attributes of the restaurant. Here, our project will create a comprehensive understanding of the cleanest and best Chicago restaurants. Consumers could use this to ensure their most loved and recommended restaurants are consistently passing inspections and are up to the health code. Many students at the Tippie College of Business migrate to the Chicagoland area for their post-graduate plans. This information, particularly to us, can affect not only our eating habits but also our friends as they move to the windy city.

In this project, we utilized the City of Chicago Health and Human Services as our csv. Then scraped from the article, *The Infatuation Best Restaurants Chicago Article.* This will allow us to see what restaurants in the area are highly rated amongst consumers and also their ratings with health inspection data.

Link to GitHub Repository: kyliemccabe1/Project-DW

Questions:

1. **Are certain types of cuisine more likely to have critical violations and health scores, among highly rated restaurants?**
    a. Exploratory: This investigates different health scores and outcomes among different cuisine types in Chicago restaurants.
2. **Is there a relationship between restaurant neighborhoods and their health inspection risk levels and outcomes?**
    a. Hypothesis: H_0 (Null): Neighborhood is independent of risk levels and inspection results.
    b. H_1 (Alt): Certain neighborhoods have higher risk restaurants or worse inspection outcomes.

3. **How stable are restaurants' health inspection outcomes and critical violation counts over multiple inspections and years? Do past inspection results predict future inspection performance?**

2. Data

Data Dictionary

Column Name	**Data Type**	**Description**
Name	**Object**	**Name of the restaurant or food establishment.**
Address_x	**Object**	**Full street address of the restaurant.**
Price	**Object**	**General pricing tier, represented by dollar signs ($–$$$$).**
Cuisine	**Object**	**Type of cuisine served at the restaurant.**
Neighborhood	**Object**	**Chicago neighborhood where the restaurant is located.**
Inspection ID	**Float**	**Unique ID assigned to a specific health inspection.**
AKA Name	**Object**	**Alternate or registered name of the restaurant.**
License #	**Float**	**Business license number registered with the city.**
Facility Type	**Object**	**Type of food service establishment (e.g., Restaurant, Bakery).**
Risk	**Object**	**The risk category from the health department is based on potential food safety hazards.**
City	**Object**	**City in which the business is located (usually "Chicago").**
State	**Object**	**State abbreviation (e.g., IL).**
Zip	**Float**	**Zip code of the restaurant’s address.**
Inspection Date	**Object**	**Date when the health inspection was conducted.**
Inspection Type	**Object**	**Type of inspection (e.g., Canvass, Re-Inspection).**
Results	**Object**	**Outcome of the inspection (e.g., Pass, Fail).**
Violations	**Object**	**Description of violations cited during inspection.**

Latitude	**Float**	**Latitude coordinate of the restaurant location.**
Longitude	**Float**	**Longitude coordinate of the restaurant location.**
Location	**Object**	**Combined latitude and longitude in tuple format.**

We primarily used data from the City of Chicago Health and Human Services website, Food Inspections | City of Chicago | Data Portal and scraped from a website Best Restaurants in Chicago, from the Infatuation.com.

The food inspection data, from the city of Chicago, gave the restaurant, inspection ID, DBA name (doing business as), AKA Name (also known as), license number, facility type, risk, address, city, state, zip code, inspection date, inspection type, results, violations, latitude, longitude, location. The csv is called, “Food_Inspections_2050331.csv.”

From the article, Best Restaurants in Chicago, from the Infatuation.com, provides information on not only Chicago, but other major cities. It serves as a hub for travelers to research the best attractions and restaurants in their destinations. The website allowed us to scrape the following: Restaurant Name, Address, Price (on range of $), cuisine type, and the neighborhood it resides in. We were able to scrape 30 web pages. And gain information on the top 25 restaurants in the Chicago area. The csv is called, “infatuation_chicago_restaraunts.csv.”

The merged dataset was able to merge as a left join, on Name. To do this, we had to rename, the column in the initial inspections data frame from “DBA Name” to “Name.” As we were only looking for the health inspections of the restaurants listed in the article, we used a left join. In addition, after joining, we dropped the rows with NA values. This gave us 101 rows and 21 columns. There are 101 rows, as most restaurants have more than one visit, which allows us to see the progression of the cleanliness of the restaurant through time. The oldest inspection year is from 2010. All inspections are represented from the years 2010-2025. This merged data frame is saved as “merged_cleaned.csv.”

3. Analysis

3.1 Question 1

Are certain types of cuisine more likely to have critical violations and health scores, among highly rated restaurants?

The city of Chicago food inspection data categorizes restaurants into different cuisine types. The top ten includes American, Experimental (These restaurants use unique techniques and science to provide an experience for the customer), Filipino, Italian, Japanese, Mexican, Nepali, Scandinavian, and Ukrainian. A critical violation is referred to as a specific infraction that is an immediate threat to public health. Things like improper food storage temperatures, insect infections, and cross-contamination. An inspection fails when one or more critical violations are identified and cannot be corrected during the inspection. Pass with conditions is when an establishment remains open because it met the minimum requirements for safe operation. The inspector found non-critical or lower-risk violations such as labeling issues or small cleanliness issues.

Based on the analysis certain types of cuisine are more likely to receive critical violations and lower health scores, even among highly rated restaurants. Figure 1 displays a bar chart of the top 10 cuisines by average critical violations. Italian and American cuisines top the list, with Italian cuisine showing the highest average number of critical violations per inspection. Further, Figure 2 complements this.

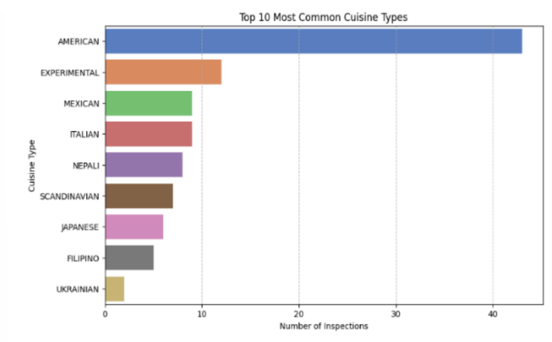


Figure 1

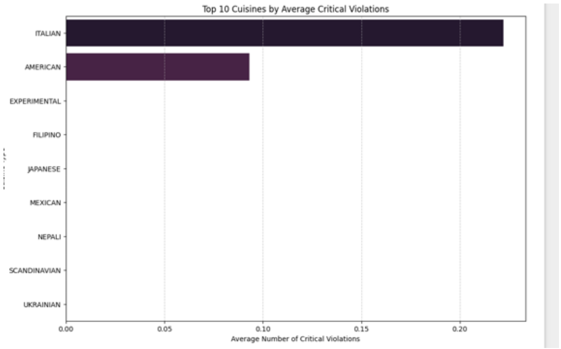


Figure 2

3.2 Question 2:

Is there a relationship between restaurant neighborhoods and their health inspection risk levels and outcomes?

This part of the analysis examined whether the neighborhood where a restaurant is located has any bearing on its health inspection outcomes and risk levels. Neighborhood options include West Loop, Wicker Park, River North, Fulton Market, Archer Heights, University Village, Logan Square, Ukranian Village, and Avondale. *Figure 3* shows the number of inspections per neighborhood; West Loop and Wicker Park emerged as the most frequently inspected neighborhoods. Interestingly, neighborhoods like Archer Heights and Logan Square showed relatively high failure rates, around 33%, whereas River North and Avondale had perfect pass records in the data reviewed. Despite these visible differences in descriptive trends, the Chi-Square Test between neighborhood and inspection outcomes yielded a **p-value of 0.2599**, indicating no statistically significant relationship at the 0.05 significance level. This suggests that, **although some neighborhoods appeared to have higher risk or worse outcomes, these differences could be due to chance**. Therefore, while geographic patterns may hint at localized issues, the data does not provide strong statistical evidence that a restaurant's neighborhood is a key determinant of health inspection performance. Text analytics, *seen in Figure 5,* was performed on the violation descriptions to enrich the analysis further. A word cloud visualization and keyword frequency analysis revealed that common violation terms across neighborhoods included food, temperature, sanitizer, and clean, pointing to recurring issues related to food handling and sanitation practices. While the neighborhoods did not statistically differ in overall inspection outcomes, the text analysis highlighted, consistent themes in health code violations, suggesting that while location may not be a decisive factor in pass/fail rates, the type of violations tends to follow common patterns across the city.

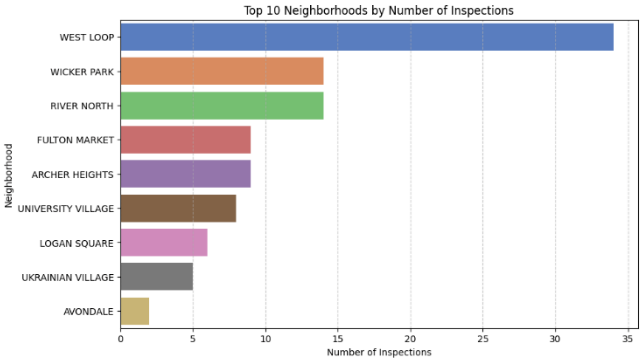


Figure 3

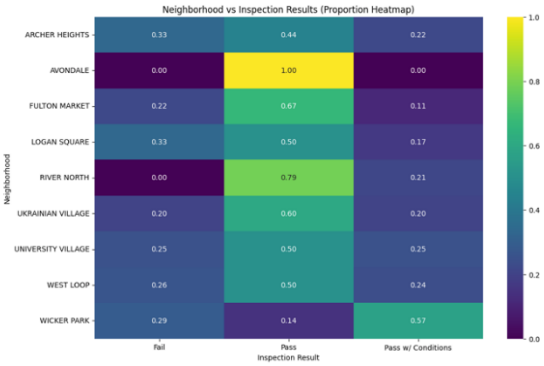


Figure 4

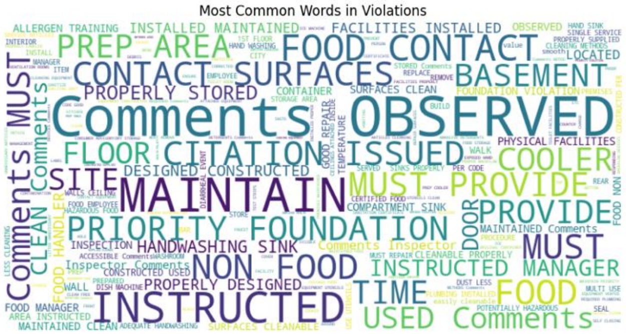


Figure 5

3.3 Question 3

How stable are restaurants' health inspection outcomes and critical violation counts over multiple inspections and years? Do past inspection results predict future inspection performance?

The final analysis focused on the stability of restaurants' inspection results over time, evaluating whether poor inspection results tend to persist across multiple inspections. The number of inspections has increased notably over time, peaking in 2014, indicating growing regulatory oversight. Critical violations showed minor fluctuations over the years but remained relatively low on average, often hovering around zero. To assess whether past inspection results predict future performance, a crosstab of first vs. last inspection results for each restaurant was analyzed. This revealed that most restaurants either improved or maintained their inspection status over time, suggesting some level of self-correction after poor performance. However, the Chi-Square Test between past and next inspection results produced a p-value of 1.0, indicating no significant predictive relationship between previous and future inspection outcomes. **This implies that while some individual restaurants may show improvement or consistency, there's no strong, consistent pattern across the broader dataset that allows us to predict future performance solely based on past inspections**. To further investigate whether past performance could predict future outcomes, a logistic regression machine learning model was implemented. The model used past inspection results, critical violation counts, and risk levels as features to predict the next inspection outcome. The model demonstrated high accuracy in predicting Pass results, largely due to the data's class imbalance (few Fail cases), but it struggled to effectively predict Fail outcomes. The confusion matrix highlighted this imbalance, confirming that while machine learning can effectively model patterns in the data, it is limited by the available features and the lack of strong signals that link past inspections to future failures. **This supports the conclusion that inspection stability is complex and likely influenced by factors beyond what is captured in the dataset.**

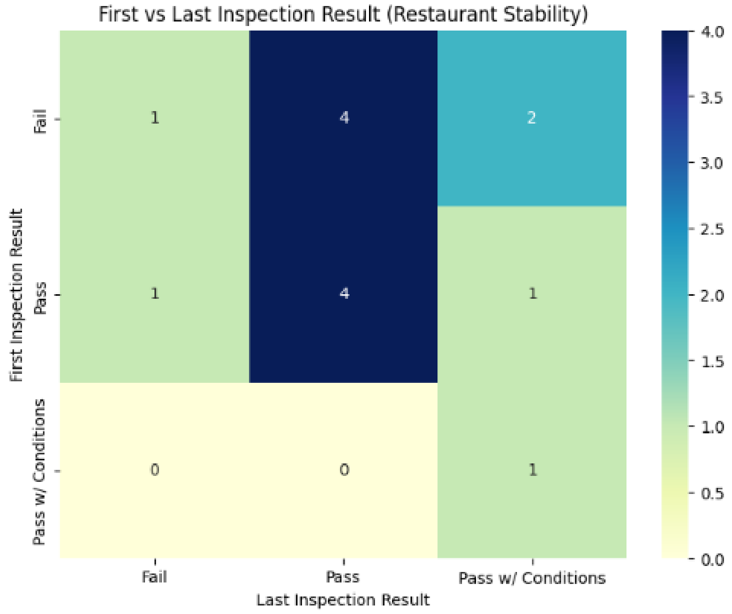
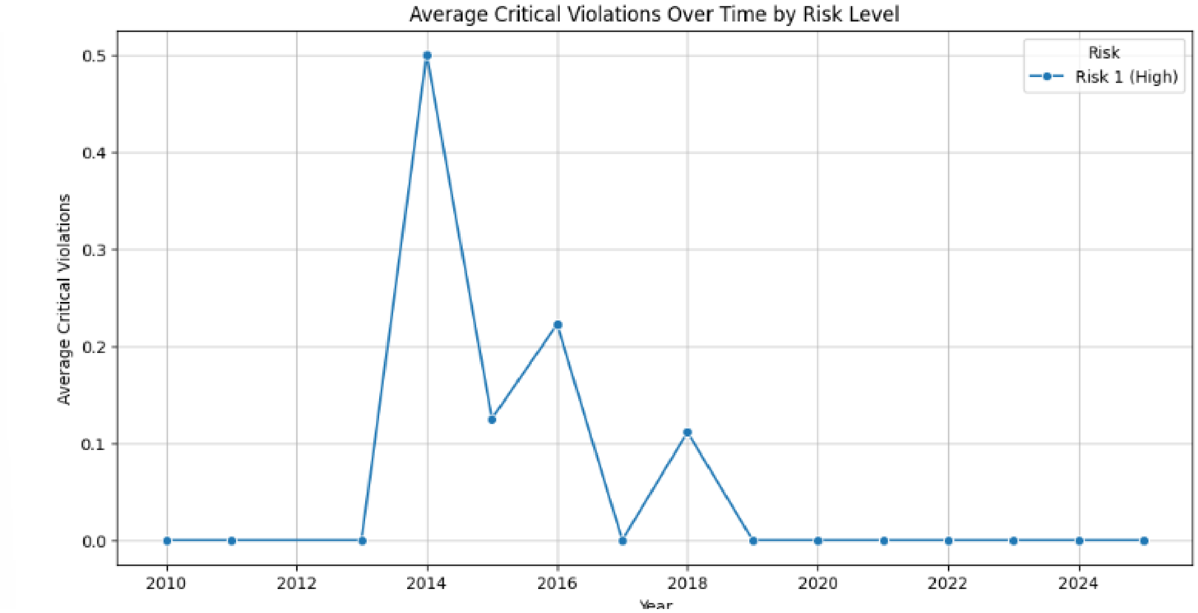
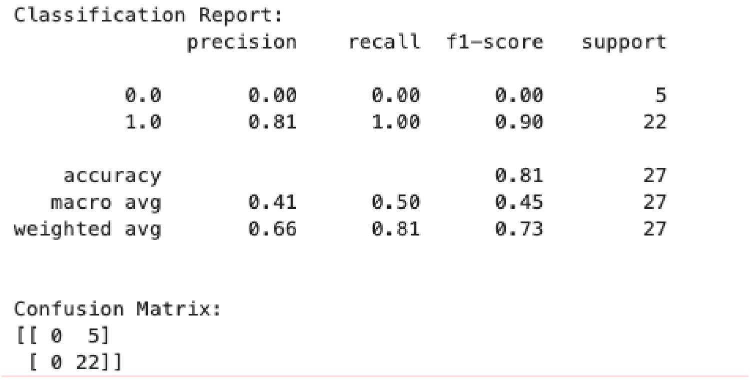


Figure 6



tex

Figure 7



```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         5
         1.0       0.81      1.00      0.90        22

    accuracy                           0.81        27
   macro avg       0.41      0.50      0.45        27
weighted avg       0.66      0.81      0.73        27


Confusion Matrix:
[[ 0  5]
 [ 0 22]]
```

Figure 8

4. Conclusion

In this project we concluded with this analysis:

1. Italian and American cuisines showed the highest average critical violations, suggesting that these cuisines may be more prone to serious health issues than others.
2. There is no significant evidence to show that there is a relationship between health inspection rate and the Chicago neighborhood
3. There is no strong, consistent pattern across the broader dataset that allows us to predict future health inspections performance solely based on past inspections.

In this project, we had a few limitations. To start, there were missing health scores, we adjusted the analysis to focus on pass/fail outcomes and critical violations instead of numeric scores. Then, imbalanced data: more 'pass' results than 'fail'; used stratified sampling to balance model training. We also had a small sample size, with only 101 rows to analyze.

Future work on this project could be to find a more robust website that allows scraping. Being able to have access to this data would allow for deeper analysis and hopefully some recommendations for the Chicago restaurant community.