Technical Report: Data Cleaning, Preparation, and Analysis of Food Safety in Chicago

Ali Alhakam – DAB11

Project: Capstone – Food Safety in Chicago

Dataset: [2010 – Present Chicago Food Inspection Data](https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/about_data)

# 1. Data Dictionary

The dataset contains inspection records collected by the Chicago Department of Public Health (CDPH). The main variables include:  
- Inspection ID: Unique identifier for each inspection.  
- DBA\_Name: Business name(s).  
- License Number: Establishment license (0 = Unlicensed).  
- Facility Type: Type of establishment (e.g., Restaurant, Grocery Store, School, Daycare).  
- Risk: Health risk classification (High, Medium, Low).  
- Address / ZIP: Location details.  
- Latitude / Longitude: Geographic coordinates.  
- Inspection Date: Date of inspection.  
- Inspection Type: Reason for inspection (Complaint, License, Routine).  
- Results: Outcome (Pass, Fail, Pass w/ Conditions).  
- Violations: Codes and descriptions of health code breaches.  
- Violations Description: Extracted from the Violations Column.

# 2. Data Cleaning Process

The raw dataset required extensive cleaning before analysis. The following steps were implemented using Python (pandas, numpy, fuzzywuzzy(an attempt was made)):

**2.1 Import and Initial Review**  
- Loaded CSV files.  
- Reviewed column types, missing values, and data distributions.

**2.2 Standardization**  
- DBA\_Name Cleaning: an attempt was made to use fuzzy matching to unify brand names, but I don’t think I succeeded in doing it right, I kept the columns I created but didn’t really use them.- Risk Field: Removed ambiguous “All” category.  
- License Number: Rows with “0” flagged as Unlicensed.  
  
**2.3 Violation Parsing**  
- Split “Violations” into three new fields: Code (numeric), Description (short text), Comments (details).

- I had help from AI for the code for this step.  
  
**2.4 Missing Data**  
- Inspections missing core identifiers were removed.  
- Missing ZIP codes flagged separately for geospatial analysis.  
- Inspection dates reformatted to consistent datetime.  
  
**2.5 Geospatial Preparation**  
- Standardized ZIP codes as string type for joining with shapefiles in tableau.  
  
**2.6 Duplication**  
- Removed exact duplicates by Inspection ID.

# 3. Other Datasets

**3.1 (311 Service Requests post-2018)**

<https://data.cityofchicago.org/Service-Requests/311-Service-Requests/v6vf-nfxy/about_data>

**3.2 (311 Sanitation Service request pre-2018)**

<https://data.cityofchicago.org/Service-Requests/311-Service-Requests-Sanitation-Code-Complaints-Hi/me59-5fac/about_data>

# 4. Introduction

Food safety is a critical public health concern. According to the World Health Organization, unsafe food causes millions of illnesses and hundreds of thousands of deaths globally each year. In large urban centers such as Chicago, where thousands of food establishments operate across diverse neighborhoods, ensuring safe food handling and sanitary conditions is both a public health priority and a logistical challenge.

This project was undertaken to better understand trends in Chicago’s food inspection outcomes and the factors associated with failure. The analysis integrates inspection results with contextual data, such as weather and socioeconomic conditions, to provide deeper insight into the drivers of compliance. By identifying hotspots of risk, recurring violations, and differences across establishment types, the findings aim to support more efficient inspections, targeted outreach, and policies that ultimately improve food safety and public health outcomes in Chicago.

# 5. Problem Statement

Since 2010 to the present, the Chicago Department of Public Health(CDPH) has inspected thousands of food establishments, 75% of which are considered high-risk. However, about 59% of those pass their inspections. This persistent failure rate poses a serious risk to public health and suggests the need for better understanding of the factors that contribute to these outcomes.

# 6. Goal & Objectives

The goal is to analyze inspection results, violation types, location, and establishments to identify the most common factors or trends linked to inspection failures and to improve the risk and result of the inspections.

## Objectives:

What are the trends in inspection outcomes over time?

How do different types of food establishments compare in compliance?

Which high-risk establishments are repeatedly failing, and how can they be supported?

Which violations are most strongly associated with failure?

Where are the highest concentrations of failed inspections geographically?

# 7. Target Audience

## Primary:

Chicago Department of Public Health (CDPH)

## Secondary:

Health inspectors

Local government officials

Restaurant managers/Establishment owners.

Public health researchers

# 8. Analytical Findings

## 8.1 Overview of Inspections

Since 2010, CDPH has inspected 16,000+ establishments. About 75% are classified as high-risk, but only 59% pass inspections. This indicates systemic weaknesses in compliance among high-risk facilities.

## 8.2 Trends in Results

2010–2017: Pass/Fail outcomes dominated; conditional passes were rare.  
2018–2019: Conditional passes rose sharply due to policy changes, peaking in 2019.  
2020–2024 (COVID): Inspections declined, passes rose, failures stayed constant.

## 8.3 Seasonality

Higher failure rates observed in Q3 (summer months), likely linked to hotter weather and food storage challenges.

## 8.4 Facility-Type Insights

Restaurants: Account for ~70% of inspections and dominate failure counts.  
Grocery Stores: Mixed outcomes, moderate failures.  
Schools: Mostly pass, but high-risk failures concerning since they serve children.  
Daycares & Long-Term Care: Few inspections, but failures here are highly critical due to vulnerable populations.

## 8.5 High-Risk Failures (Chronic Offenders)

Repeat failures found at establishments like Soho House and Luigi’s Pizza (24 fails per site). Popeyes stands out as the only large national chain in the top five offenders.

## 8.6 Violation Drivers

Top contributors to failures:  
- Basic hygiene lapses (unclean floors, food equipment).  
- Pest control issues (entry points, infestations).  
- Maintenance problems (ventilation, plumbing).  
- Critical risks like wastewater mismanagement.

## 8.7 Geographic Patterns

Downtown & North Side: Highest clustering of establishments → most inspections and failures.  
South & West Sides: Moderate density, fewer establishments.  
Edges: Sparse, residential/industrial zones.

## 8.8 Linking Inspections & Sanitation Complaints

Overlay of 311 sanitation requests with inspection failures shows clear overlap in hotspot areas. Suggests that neighborhood-level sanitation drives establishment failures, not just internal hygiene practices.

# 9. Policy Recommendations

Based on the analysis:  
1. Target High-Risk Establishments: Focus inspections and training on recurring high-risk offenders.  
2. Address Core Violations: Hygiene, pest control, and equipment upkeep should be priority.  
3. Seasonal Preparedness: Implement Q3-focused training (temperature control, storage).  
4. Geographic Programs: Deploy integrated sanitation + inspection initiatives in hotspot ZIP codes.  
5. Protect Vulnerable Populations: Special oversight for schools, daycares, and long-term care facilities.

# 10. Conclusion

The cleaned dataset provided a reliable foundation for time-series, geospatial, and categorical analysis. Insights reveal chronic compliance problems in restaurants, seasonal peaks in failures during hot months, repeat offenders driving failure counts, and strong links between community sanitation issues and food inspection failures. These findings highlight the need for integrated interventions combining inspections, sanitation programs, and tailored support for vulnerable facilities.

# 11. Limitations & Assumptions

## Data Quality Issues

Some ZIP codes were missing, miscoded, or contained placeholder values. During mapping, a few points appeared outside of Chicago’s boundaries; these were excluded to ensure geographic accuracy.

## Inspection Frequency Bias

High-risk establishments are inspected at least twice as often as medium- or low-risk facilities. This increases the likelihood of identifying failures in high-risk establishments compared to lower-risk ones, which may introduce a bias in comparisons.

## Policy and External Disruptions

The introduction of new Food Code Rules in July 2018 altered inspection outcomes, most notably increasing 'Pass with Conditions.' Similarly, inspection volumes and outcomes were disrupted during the COVID-19 period, affecting trend consistency.

## Violation Data Limitations

Violation text was unstructured and varied by inspector. Although significant cleaning was performed, inconsistencies may remain and could have influenced the accuracy of violation-level analysis.

## Assumption

On the 311 requests dataset the sanitation one was for data pre-2018, and the data that is after 2018 the 311 recording system became that the data was not categorized but all 311 requests for Chicago were recorded in one dataset, so the data set was too big and even the cleaning was taking to long to load using python. So, the assumption I made for making 1 of the visuals using the 311 Sanitation request data was that the pre-2018 data was enough.

# 12. Citations

**World Health Organization (WHO**) – Food safety statistics used in the introduction.  
https://www.who.int/news-room/fact-sheets/detail/food-safety

**City of Chicago Data Portal** – Primary inspection dataset and supporting datasets.  
https://data.cityofchicago.org/

**City of Chicago Official Website** – Food Code Rules (effective July 1, 2018).  
https://www.chicago.gov

**U.S. Census Bureau** – Shapefiles used for geographic mapping.  
https://census.gov

**Visual Crossing Weather** – Weather data used to examine seasonal and temperature-related inspection effects.  
https://www.visualcrossing.com/