# Chicago and Dallas Food Inspection Data Profiling

## Introduction

The Chicago and Dallas Food Inspection Datasets are public datasets that offer detailed insights into food establishment inspections conducted by local health departments. These datasets include inspection details such as inspection dates, risk levels, inspection results, violation descriptions, establishment names, and location-related fields. Profiling and analyzing these datasets enables insights into food safety compliance, violation trends, and operational risks.

## Overview

• Combined Dataset Size: Approximately 200K+ records after cleaning and integration.  
• Update Frequency:  
 ◦ Chicago: Updated daily.  
 ◦ Dallas: Updated weekly.  
• Columns:  
 ◦ Chicago: ~17 core fields.  
 ◦ Dallas: ~114 fields (including multiple violation fields).  
• Source: Cleaned and unified staging tables exported as .tsv files.

## Data Quality Assessment

1. NULL and Missing Values:  
  
• Chicago: Fields like AKA Name, License, Risk, and Violations contain NULLs or blanks.  
• Dallas: Violation-related fields (Violation\_1 to Violation\_25) and Score/Result fields often contain missing values.

2. Inconsistent Representations:  
  
• Chicago: 'Risk' values vary in formatting ('Risk 1 (High)', 'Risk 1-High').  
• Dallas: Free-text violation fields and inconsistent casing.  
  
Actions Taken:  
• Standardized casing, removed whitespaces, mapped risk levels and results.

3. Precision Standardization:  
  
• Chicago has reliable latitude/longitude.  
• Dallas has embedded addresses within location fields.  
  
Actions Taken:  
• Standardized lat/long to float; split composite location in Dallas.

4. Unwanted / Redundant Columns:  
  
• Chicago: Dropped duplicate geolocation columns.  
• Dallas: Dropped 100+ wide violation fields.  
  
Actions Taken:  
• Transformed violation columns in Dallas using pivot. Created fields:  
 - Violation\_Description

5. Duplicate Records:  
  
Both datasets contained duplicates based on inspection date, establishment, and result.  
  
Actions Taken:  
• Defined composite uniqueness key and deduplicated exact matches.

6. Violation Normalization in Chicago (flattned):  
  
Chicago stores multiple violations in one text field.  
  
Actions Taken:  
• Parsed entries using delimiters (|, :)  
• Separated into Violation\_Description and Violation\_Comments

## Detailed Profiling Findings by Dataset

### Chicago Dataset

The Chicago dataset contains approximately 41,000 records with key fields including Inspection ID, DBA Name, Facility Type, Risk, Results, and Violations. Profiling reveals that the 'Violations' column has over 40,000 nulls, indicating most inspections resulted in passes. The 'Facility Type' field is missing in 661 records, and geolocation fields (Latitude, Longitude, Location) are missing in 476 records. 'Risk' is missing in 43 records, and minor nulls exist in 'AKA Name', 'City', and 'Zip'.

### Dallas Dataset

The Dallas dataset contains about 119,000 records and includes fields such as Restaurant Name, Inspection Type, Inspection Score, Violation Descriptions 1-25, and geolocation. Profiling reveals that most 'Violation Description' columns are null in over 100,000 records each, due to the spread of violations across 25 fields. 'Inspection Score', 'Inspection Date', and 'Street Name/Number' are missing in over 51,000 rows. This suggests a large portion of the data may be incomplete or redundant and may require filtering or structural transformation.

|  |  |  |
| --- | --- | --- |
| Issue | Chicago Dataset | Dallas Dataset |
| Missing Violations | Violations (40k+) | Violation Desc 1-25 (100k+ each) |
| Missing Risk/Score | Risk (43) | Score (51k) |
| Missing Location | Lat/Long (476) | Lat Long (minor) |
| Date Nulls | Minor | 51k rows |
| Business Name Gaps | AKA Name (188) | Restaurant Name (60) |

## Conclusion

This profiling effort resolved key data quality issues—missing values, redundancy, duplicates, and inconsistencies. Transformations such as unpivoting violation columns and normalizing text fields made the dataset analysis-ready. This cleaned and standardized dataset supports accurate reporting and comparative insights across cities.