Boston Area Research Initiative ● Northeastern & Harvard Universities
E-Mail: bari@northeastern.edu ● Web: www.bostonarearesearchinitiative.net

A Guide to Food Establishment Inspections Data for the City of Boston, MA

Overview

The document describes the structure and organization of the Food Establishment Inspections dataset gathered from the Analyze Boston website (data available here). The Food Inspections dataset is released by the Health Division of the Department of Inspectional Services of Boston which ensures that all food establishments in the City of Boston meet relevant sanitary codes and standards. The businesses serving food are inspected at least once a year, and follow-up inspections are performed on high-risk establishments. This dataset is released through the city's open data initiative.

The inspections data are linked to restaurant reviews scraped from Yelp. The data have been processed by the Boston Area Research Initiative (BARI) to generate metrics at three analytic levels: reviews (i.e. information about Yelp reviews posted for restaurants), restaurants (i.e. information about the restaurants), neighborhoods (i.e. aggregate measures describing neighborhoods). For the present connection Food Inspections – Yelp, the reviews are considered. For more details on how the reviews are scraped and structured please check the separated release on data scraped from "Yelp" on the Dataverse platform, under the Boston Area Research Initiative user.

Thus, this data includes two files:

- *Food.Inspections.Records.csv* contains information about food inspections at record level (i.e. each record for each restaurant is included).
- *Food.Inspections.Yelp.Restaurant.csv* contains information about food inspections at the restaurant level plus information from Yelp reviews also at the restaurant level.

Table of Contents

1. Food Establishment Inspections Records 3
1.1. Description of contents 3
1.2. Summary of Variables 3
1.2.1. General Characteristics 3
1.2.2. Geographical Information 4
2. Food Inspections per restaurant, and connection with Yelp restaurant reviews 5
2.1. Description of Contents 5
2.2. Summary of Variables 5
2.2.1. General Characteristics 6
2.2.2. Geographical Information 8

1. Food Establishment Inspections Records

1.1. Description of contents

The Food Establishment Inspection is a legacy dataset containing records of violations identified by inspections of establishments in Boston that serve food. The dataset is released by the Health Division of the Department of Inspectional Services, which ensures that all food establishments in the City of Boston meet relevant sanitary codes and standards, through the City of Boston's open data initiative data.boston.gov. The businesses serving food are inspected at least once a year, and follow-up inspections are performed on high-risk establishments.

The data has a total of 290,832 records (*note*: 52,807 cases from Retail Food Shops included in the City of Boston's original data release are excluded; also all the records without geographic *lat*, l*on* or *property_id* are excluded) and 33 variables related to business names, locations, inspections, violation levels, reasons for the violation, comments of improvement, etc. The data span January 1st, 2010 to August 17th, 2020.

1.2. Summary of Variables

The list of variables below includes the original ones from the Food Establishment Inspection dataset available at the Analyze Boston portal. In addition to them, we are adding nested geographical information from the Geographical Infrastructure for the City of Boston.

1.2.1. General Characteristics

- *Id* Indicates the Id number for individual inspection record.
- *businessname* gives the business name of food establishments.
- *legalowner* gives the name of legal owner of food establishments.
- *namelast* gives last name of contact.
- *namefirst* gives list name of contact.
- *licenseno* indicates License number of food establishments.
- *issdttm* gives issue date of license.
- *expdttm* gives expiration date of license.

- *licstatus* indicates status of license, active or inactive.
- *licensecat* indicates category of license.
- *descript* indicates type of food establishments.
- *result* indicates result of inspection.
- *resultdttm* gives the date on which inspection results were generated.
- *violation* gives coding of law regulation related to violations.
- *viollevel* indicates level of violations.
- *violdesc* indicates reason of violation.
- *violstatus* indicates status for violation, pass or fail.
- *statusdate* gives the date on which violation status was generated.
- *comments* indicates the comments given to the food establishment for improvement.

1.2.2. Geographical Information

- *address* indicates the addresses of food establishments.
- *city* indicates the city in which the food establishment is located in.
- *state* indicates the state of the restaurant.
- *zip* indicates the zip codes for the areas that food establishments located in.
- *property_id* indicates the property Id numbers for food establishments.
- *lon* gives the longitude where the restaurant is located.
- *lat* gives the latitude where the restaurant is located.
- *Land_Parcel_ID* is the unique identifier for the land parcel the restaurant is in.
- *GIS_ID* is the identifier for the land parcel the restaurant is in.
- *Blk_ID_10* is the 2010 census Block ID number.
- *BG_ID_10* is the 2010 census Block Group ID number.
- *CT_ID_10* is the 2010 census Tract ID number.

2. Food Inspections per restaurant, and connection with Yelp restaurant reviews

2.1. Description of Contents

Measures from the Food Establishment Inspection Records were aggregated at restaurant level. In addition, the aggregate measures from were merged with content from Yelp to form a new, restaurant-level dataset. The merged dataset has 4,884 rows and 37 columns. From these, 1,044 restaurants exist in both databases (Food Inspections and Yelp Reviews), 3,319 are only present in the Food Inspections (i.e., had at least one inspection generating violations during the time period measured and no content on Yelp) and 2,610 only present in the Yelp Reviews[1] (i.e, they had minimum one review at the time the Yelp data was scraped but no violations generated over the time period measured).

Food Inspections and Yelp Reviews were linked based on a mixture of spatial joining locations (i.e. by using the provided latitude and longitude) and fuzzy string matching by: (1) using the Levenshtein distance between restaurant names in the two datasets; (2) finding similarity between a combination of restaurant names and Geographic Information (longitude and latitude, GIS_ID, Land_Parcel_ID). Manual checks were conducted following the joining process to ensure a better accuracy. Methodology advances may help improving the completeness of the data joined.

2.2. Summary of Variables

The first part of the dataset contains variables referring to the Food Inspections dataset (from the first column *Restaurant Name* to the column *Violation 2020 count*), followed by variables from the Yelp Reviews dataset (starting from the column *Price* to the column

[1] For more details on how the reviews are scraped and structured please check the separated release on data scraped from "Yelp" on Dataverse.

inactive_year)[2]. The Geographic Information variables are common for the two datasets. If a restaurant was found on the Food Inspections dataset, the column *Food.Inspection* will get the value "1". If a restaurant was found on the Yelp dataset, the column *Yelp* will get the value "1". Thus, the restaurants which are common in the two datasets will have value "1" in both columns.

2.2.1. General Characteristics

- *Restaurant Name* gives the name of the restaurant.
- *Address* gives the address of the restaurant.
- *Inspection Fail count* gives number of times restaurant failed an inspection.
- *Inspection Pass count* gives number of times restaurant passed an inspection.
- *Violation Level 1* gives number of times restaurants had violations of level 1.
- *Violation Level 2* gives number of times restaurants had violations of level 2.
- *Violation Level 3* gives number of times restaurants had violations of level 3.
- *Violation Pass count* gives number of times restaurant passed the violations.
- *Violation Fail count* gives number of times restaurant failed the violations.
- *Violation 2010 count* gives number of times violation happened in the restaurant for year 2010.
- *Violation 2011 count* gives number of times violation happened in the restaurant for year 2011.
- *Violation 2012 count* gives number of times violation happened in the restaurant for year 2012.
- *Violation 2013 count* gives number of times violation happened in the restaurant for year 2013.

[2] The merged dataset does not include retail food restaurants from the Food Establishment Inspections dataset.

- *Violation 2014 count* gives number of times violation happened in the restaurant for year 2014.
- *Violation 2015 count* gives number of times violation happened in the restaurant for year 2015.
- *Violation 2016 count* gives number of times violation happened in the restaurant for year 2016.
- *Violation 2017 count* gives number of times violation happened in the restaurant for year 2017.
- *Violation 2018 count* gives number of times violation happened in the restaurant for year 2018.
- *Violation 2019 count* gives number of times violation happened in the restaurant for year 2019.
- *Violation 2020 count* indicates the number of times violation happened in the restaurant for year 2020.
- *Price* indicates how expensive the restaurant is.
- *review_count_total* indicates the total count of reviews for the restaurant.
- *review_count_anuave* indicates the count of annual average reviews for the restaurant.
- *year_count* indicates the number of years the restaurant has been active.
- *ave_rating* gives the average rating for the restaurant.
- *latest_review* gives the latest year when the restaurant was reviewed.
- *earliest_review* gives the earliest year when the restaurant was reviewed.
- *inactive_year* gives the count of years the restaurant has been inactive.
- *Yelp is a* binary variable where "1" indicates that the restaurant has minimum one Yelp review, and "0" indicates the restaurant is not part of the Yelp dataset.

- *Food.Inspection is a* binary variable where "1" indicates that the restaurant has minimum one Food Inspection, and "0" indicates the restaurant is not part of the Food Inspection dataset.

2.2.2. Geographical Information

- *lon gives the longitude where the restaurant is located.*
- *lat gives the latitude where the restaurant is located.*
- *Land_Parcel_ID is the unique identifier for the land parcel the restaurant is in.*
- *GIS_ID is the identifier for the land parcel the restaurant is in.*
- *Blk_ID_10 is the 2010 census Block ID number.*
- *BG_ID_10 is the 2010 census Block Group ID number.*
- *CT_ID_10 is the 2010 census Tract ID number.*