Eating Without E. Coli in Boston

Neil MacAloney

4 April, 2020

1. Introduction

Choosing a place to eat should not be so hard. One usually turns to either familiar, or the most popular or highest rated places on social media sites in order to pick a restaurant to eat at. Since your well being usually takes last place behind other companies trying to make money such as paid marketing and social media companies hyping their own products, it is nearly impossible to get an objective way to compare restaurants without going to the actual location and eating there.

At least annually, eating establishments in Boston are inspected by the city to ensure their standards are in accordance with food safety guidelines. However, the way scores are graded, any violations can be corrected with a re-inspection. The establishment uses the new grade (which is usually a passing grade), and all violations are buried in a violation log that is online. Even though the posted grade is the re-inspection grade, the hidden data from all violations is readily available in a public database from the city of Boston.

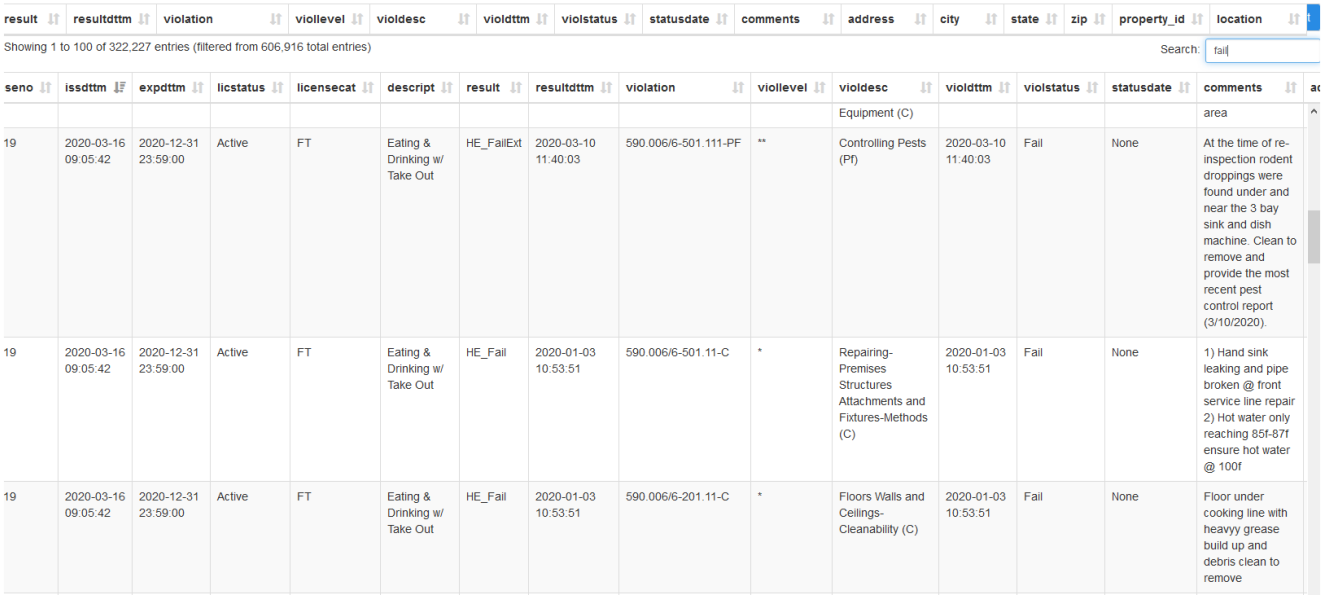
This project uses the database of violations that are not advertised to compare eating establishments in the Boston Area so people can choose places to eat that are based on how restaurants score on violations from initial inspections – which are more likely to represent actual eating conditions.

Anyone that regularly eats in Boston would be interested in seeing a graphic such as this. Ultimately, eating at these locations will reduce your chance of getting food poisoning. The food may not taste better, but maybe you'll try a place based on how much effort they put into their measured cleanliness and standard of care rather than some social media company's rating which may be based on nothing.

2. Data

The data used for this project is available from https://data.boston.gov/dataset/food-establishment-inspections.

The database is approximately 600,000 entries back to 2007 of over 3000 establishments. As snapshot of some of the columns and example violations are listed below



result	resultdttm	violation	viollevel	violdesc	violdttm	violstatus	statusdate	comments	address	city	state	zip	property_id	location

Showing 1 to 100 of 322,227 entries (filtered from 606,916 total entries) Search: fail

seno	issdttm	expdttm	licstatus	licensecat	descript	result	resultdttm	violation	viollevel	violdesc	violdttm	violstatus	statusdate	comments
										Equipment (C)				area
19	2020-03-16 09:05:42	2020-12-31 23:59:00	Active	FT	Eating & Drinking w/ Take Out	HE_FailExt	2020-03-10 11:40:03	590.006/6-501.111-PF	**	Controlling Pests (Pf)	2020-03-10 11:40:03	Fail	None	At the time of re-inspection rodent droppings were found under and near the 3 bay sink and dish machine. Clean to remove and provide the most recent pest control report (3/10/2020).
19	2020-03-16 09:05:42	2020-12-31 23:59:00	Active	FT	Eating & Drinking w/ Take Out	HE_Fail	2020-01-03 10:53:51	590.006/6-501.11-C	*	Repairing-Premises Structures Attachments and Fixtures-Methods (C)	2020-01-03 10:53:51	Fail	None	1) Hand sink leaking and pipe broken @ front service line repair 2) Hot water only reaching 85f-87f ensure hot water @ 100f
19	2020-03-16 09:05:42	2020-12-31 23:59:00	Active	FT	Eating & Drinking w/ Take Out	HE_Fail	2020-01-03 10:53:51	590.006/6-201.11-C	*	Floors Walls and Ceilings-Cleanability (C)	2020-01-03 10:53:51	Fail	None	Floor under cooking line with heavyy grease build up and debris clean to remove

The site has a row for each entry which consists of violation, re-inspection results, filed-violations, and many others. The data we are concerned with in this project is the failed violations from inspections that are recent (within the past year) for each and every one of the food establishments in Boston. The data is cleaned as described in the Methodology section below.

3. Methodology

A fair bit of data cleaning was done in order to display the information it in a format that is useful and easy to compare establishments to each other:

1. The data was downloaded from the link above in the Data section, and a number of extra columns that were not needed were dropped.
2. A new dataframe of a list of food establishments was extracted from the list. Each row represented a single food establishment for which scores, geographic coordinates, and Foursquare data would be added to.
3. For the food establishments that did not have geolocation data, the address was run through Nominatim (https://nominatim.openstreetmap.org/) in order to get lat/long data to plot. For establishments that did not have lat long data in the database and could not fetch a coordinates from Nominatim, the entry was dropped (this was approximately 235 of the 3285 total establishments)
4. From the original violation list, all violations more than 365 days were removed from the dataframe. The conditions at food establishments are always changing and annual inspections allow for a recent snapshot of restaurant cleanliness. It didn't make sense to have violations more than a year old. Next, violations that weren't a "fail", violations from inactive establishments, and duplicate entries were dropped.
5. Each establishment then had their violations totaled by violation level, from non-critical violations, critical violations, and foodborne illness risk factor critical violations which have been identified by the Centers for Disease Control and Prevention (CDC) as the

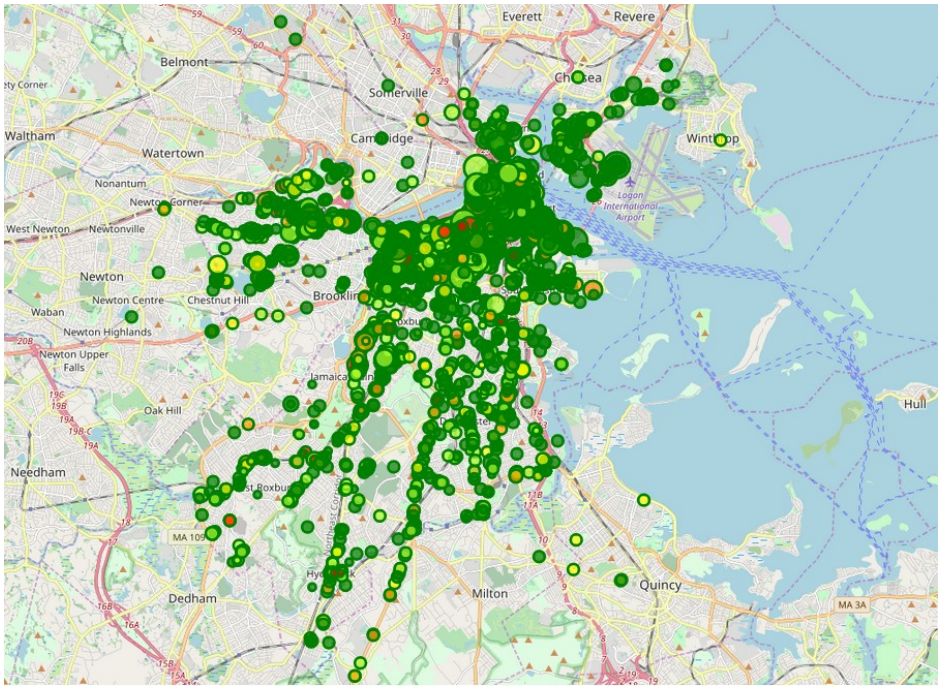
most prevalent factor to foodborne illnesses. The violations for each establishment were then used to create a score by deducting the following points from 100 for each violation type.

* = Non-Critical Violation, ½ point penalty
** = Critical Violation, 3-½ points penalty
*** = Foodborne Illness Risk Factor Critical Violation,5 point penalty

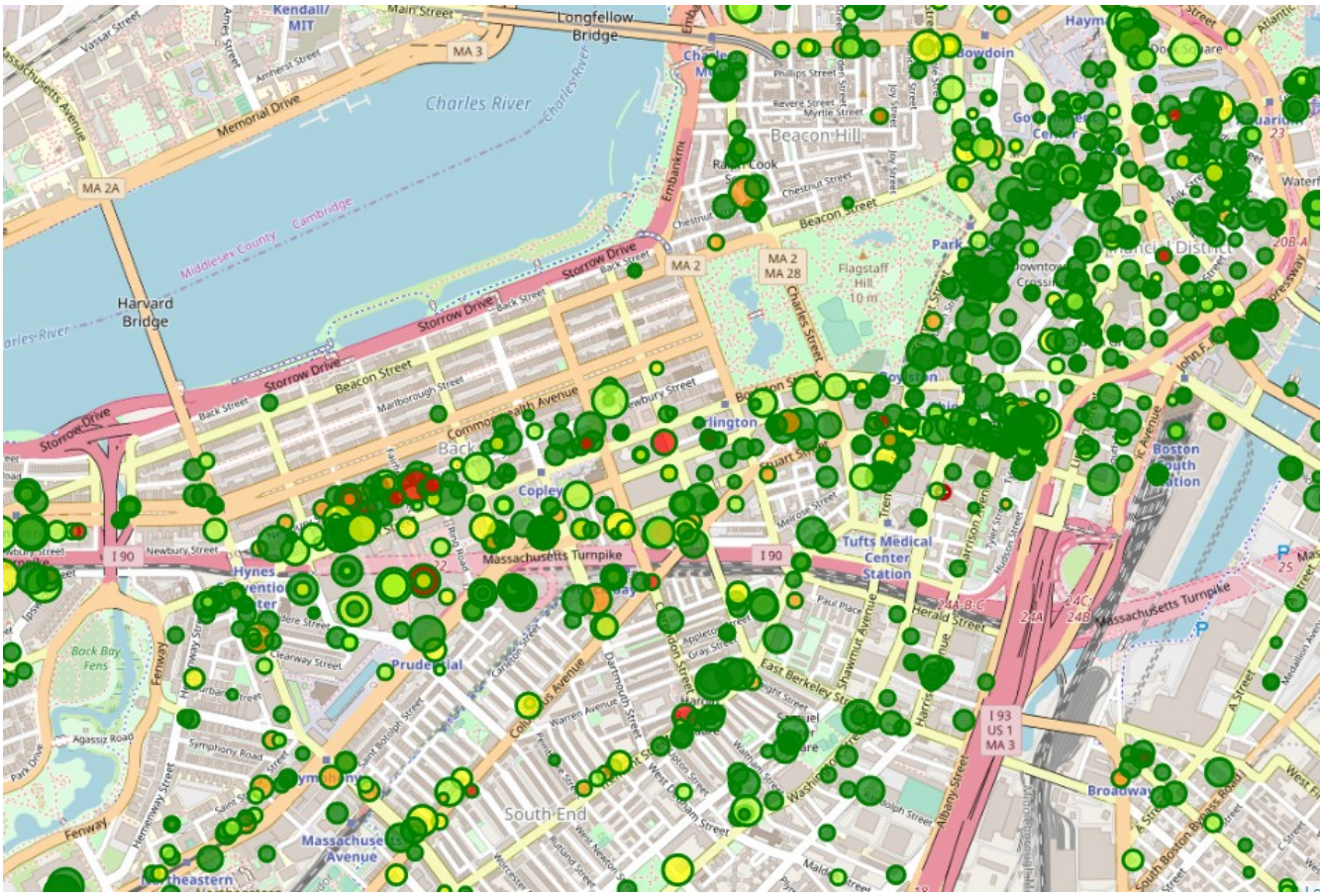
6. Each of the establishments was run through the Foursquare database twice. Once in order to obtain the ID of the establishment using the name of the establishment and its lat/long address. Next the ID of each site was again run through Foursquare in order to get the number of likes for each.
7. The results were plotted with folium. The color of each establishment's marker was related to it's number of violations recorded, and the size of each marker was related to the number of likes the venue received through Foursquare.
8. One final step that was taken was that all violation comments from the past year were added to the marker, so that users of the map can click on the marker and see what violations were recorded (optional but very informative).

4. Results

The result of this exercise is an interactive map that is able to provide a comparison of food inspection scores using a standard ranking system across all establishments in Boston.



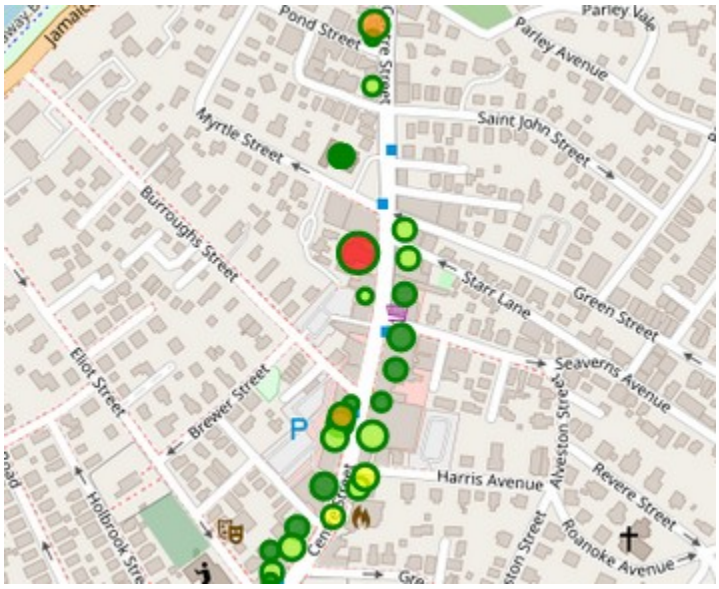
From this, a user can scroll into different areas of Boston to look at different neighborhoods and how the venues look in terms of violation history. From the next picture, if you were looking at going to a neighborhood to eat, a cluster of red/orange marks in an area should keep you on edge.



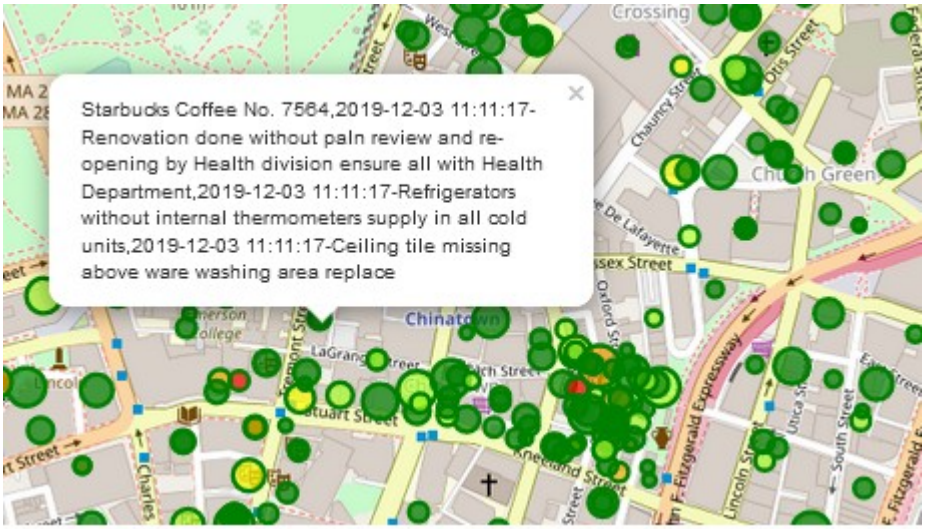
For example, Newbury street has a higher concentration of lower-scoring establishments in terms of food safety violations.

On the other hand, establishments just south of the Boston Common have a much higher concentration of good-scoring restaurants which may make you more willing to experience a new place even if you are not familiar with it.

Additionally, a large red marker in the image below shows that there is an establishment that is very popular but that received many food safety violations in previous inspections. This may indicate very active marketing, or a very popular place has recently had it's standards drop off. Since there are a number of establishments nearby that are small and have good inspection results, this particular place looks to be an outlier..



There is also a block of codes that applies each of the violation comments from the past year to the business name. By clicking on the marker, the business name comes up in addition to a datetime and comment from each violation from the past year.



Feel free to access the attached notebook to explore the data yourself.

5. Discussion

From the results above, it is clear that the data shows that there are many different ways to visualize the results of the data. The data provides a lens to the recent violation history at nearly all establishments inspected in Boston. This type of map would be very useful for consumers to look at new restaurants to get a sense of comfort going out to eat at a new place, or fact check a fancy marketing campaign to see if the restaurant is all that the ad says it is. This also gives business owners a good idea of their competition and can provide insights on where to open up a new business to either sell food in a “green” neighborhood, or perhaps open a business where your employees have walking access to clean, well-kept restaurants.

With the data, ~235 sites were dropped due to data issues. With more times, I would next develop ways to bring in all of the restaurants instead of dropping them because no lat/long data was provided or able to be gathered.

I would also like a way to incorporate the current grade (i.e. A, B, C) and also indicate whether the establishment is suspended or not. For this project it is a snapshot of the past-year’s data when the code is run and does not reflect current grade or status. Grades and suspension statuses change often so unless the code was run weekly or daily, it would give incorrect ratings of some sites.

It would be also good to have a way to signify a trending direction of each establishment. For example, if for many years, the restaurant scored very poorly and then received excellent results the past year, it may indicate a chance of management or turning point. The opposite may also be true where the safety and cleanliness fall off after many years of receiving high scores.

6. Conclusion

This project was able to successfully extract food establishment violation data from the Boston database and incorporate data from Foursquare to create an interactive map to compare establishments to each other. The data shown is truly objective in that it cannot be bought with marketing dollars and that it should be correlated to the chance of getting food poisoning at a particular venue. This map is useful for those people that don't want to put their hands over their eyes and truly want to see what happens behind the scenes in food safety. This kind of maps let users compare neighborhoods, look at the violations of their favorite venues, and be able to see if an online recommendation will end up putting them in the bathroom all night.