NYC Food Inspection

Lets you decipher the actual restaurant health status!

Deepak Atal (deepak.atal@nyu.edu, da1722)
Richie Daniel Johnson (richie.johnson@nyu.edu, rdj259)
Vignesh Gawali (vignesh.gawali@nyu.edu, vg975)
Scott Lee (sl3998@nyu.edu, sl3998)

Project page (on Github): https://github.com/NYU-CS6313-Fall16/NYC-Food-Inspection-5/
Video: https://vimeo.com/196814827
Working demo: https://nyu-cs6313-fall16.github.io/NYC-Food-Inspection-5/dashboard.html

What is the problem you want to solve and who has this problem?

Most of us when visiting a new restaurant usually just check their grade to validate if it is safe to eat at the restaurant. But there is more than just a “Grade” to a food inspection. With this visualization we want to help people understand a restaurant’s inspection score history and how they have improved or deteriorated over the years.

What are the driving analytical questions you want to be able to answer with your visualization?

- How do restaurant inspection results change over time? Are restaurants improving based on the given feedback?
  The visualization should be able to convey information about how a restaurant has improved or deteriorated over time. It should be able to show if the previous inspection scores have influenced the later/current scores.

- Are certain violations especially common in certain areas? Is there an area that is particularly susceptible to vermin based problems?
  There should be an option to see each individual violations spread out through the country. It should be easy to identify areas that are particularly susceptible to particular violations.

- What kind of restaurants (cuisines) suffer from certain kinds of violations and how it changed over time? Do bakeries have different problems from burger restaurants?

The visualization should be able to show if there are particular cuisines that have specific violations. It should be easy to compare each cuisine against each of the violation and identify if there is a particular pattern for violations against cuisines.

What does your data look like? Where does it come from? What real-world phenomena does it capture?

Data was obtained from NYC Open Data which was provided by Department of Health and Mental Hygiene as Open Data. It contained 18 different attributes originally. Since we are using map for visualizing Food Inspection results of all restaurants in Brooklyn, we had to prune the data for our specific need.

But for visualization purpose, we decided to restrict the columns to following attributes :

Attribute Name	Attribute Type	Description	Value Range/Categories
Address	Categorical	Combination of attributes : Building, Street, Zipcode	All addresses where Borough is Brooklyn
Cuisine	Categorical	Type of Cuisine served in Restaurant	68 Categories were shortlisted/merged with similar ones
Inspection Date	Ordinal	Date of Inspection	Year 2013- 2016
Average Score	Quantitaive	Derived Data Type from Score	<=13 points = A grade >13 & <28 points = B grade >=28 points = C grade
Violation Type	Categorical	Specific codes of violation were merged to simplify Violation type	Vermin; Contamination; Regulation; Hygiene; Facility; Temperature;
Critical Flag	Categorical	Defines the critical nature of violations	Critical/ Not Critical
Grade	Ordinal	Grade assigned to Restaurant according to Violation Points (more the violation points, lesser the grade and vice-versa)	A,B,C,P(Pending)
Latitude	Quantitaive	Derived from Restaurant address	near 40.6782° N
Longitude	Quantitaive	Derived from Restaurant address	near 73.9442° W

What have others done to solve this or related problems?

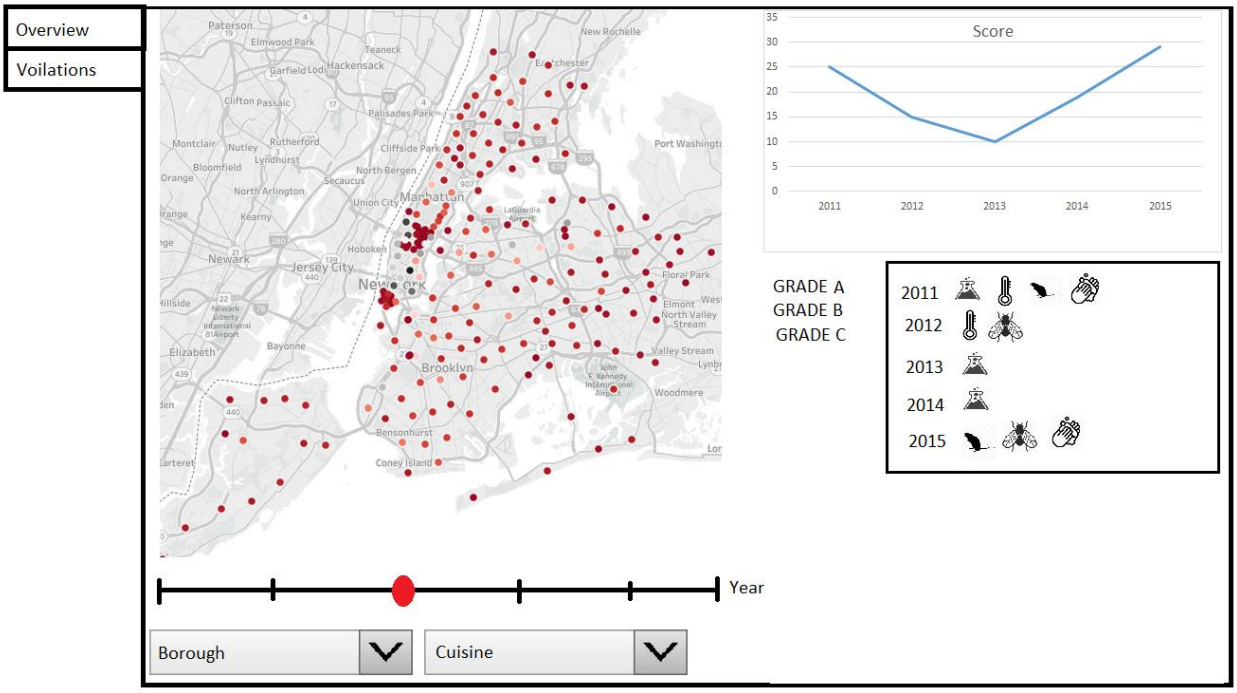
Open Health Scores attempts to give users data on the best and worst restaurants in any given week in terms of their health score and can be found here. The New York Times maintains an interactive map displaying restaurant health inspection grades across the city, and can be found here. The aesthetic design of the map portion of our visualization is heavily inspired by this. Kaiser Fung attempted to use data analytics to find the most predictive factors of a restaurant's

grade. The results of his work can be found here. Some of the methodology relating to data grouping is replicated in this work in order to make the data easier to read.

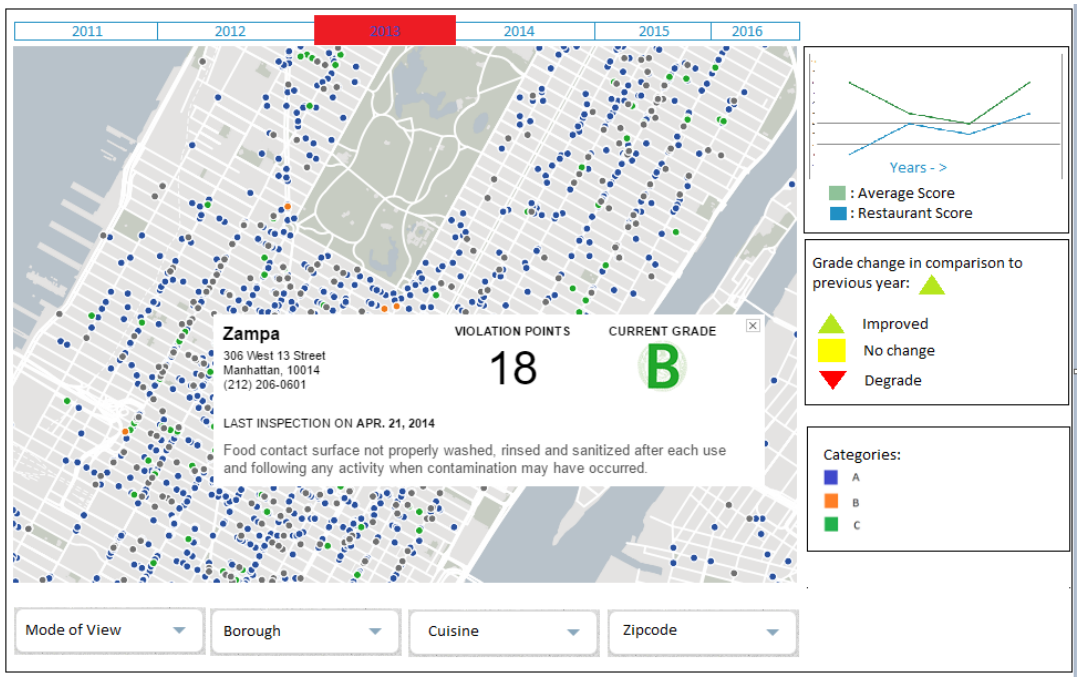
Design Iterations

The design process began with individual mockups.

This mockup was designed to show restaurants on the map colored either by grade or by violation, and selectively filtered out by cuisine. A user could look at the distributions for a given point in time using the slider. There were several problems with this mockup such as the slider (a continuous system for a discrete variable), and the fact that grades and violations were not simultaneously visible.



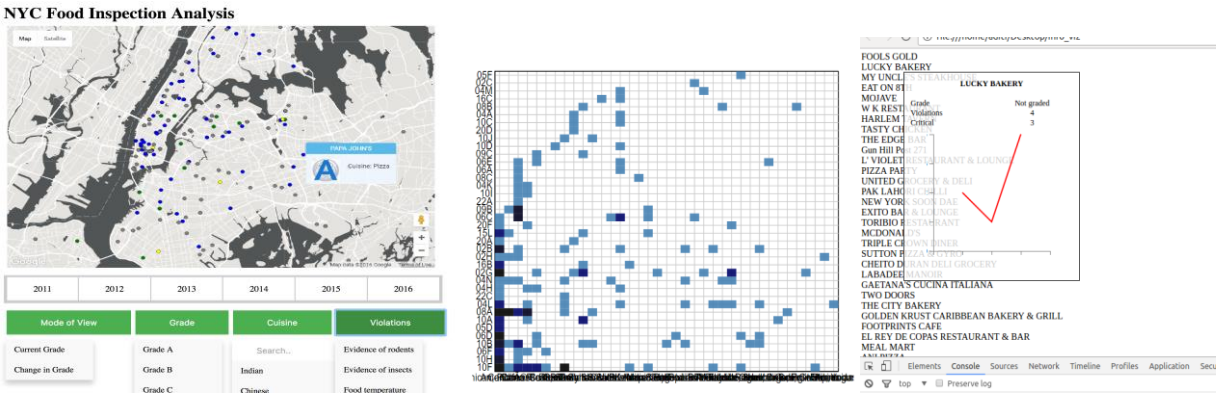
This mockup was intended to encompass all of the necessary information by making use of multiple tabs. The overview tab displayed grade information, while the violations tab would display the distribution of violations across the city. While it does convey all of the data necessary, it still makes simultaneous viewing somewhat difficult, and we decided that multiple tabs are an inconvenience that may not be necessary.



This mockup attempts to use a single tab view, with additional aggregations on the right hand side. It also allows for filtering by location. This one attempts to display violations by allowing the user to filter them out. However that makes it difficult to compare violations against one another. Additionally, it was decided that given the number of restaurants that come in and out of existence over the course of a year, grade change was not a viable variable for aggregation.

We as a group believed that this was the strongest initial mockup, as it effectively conveys the necessary information to answer the question in a single view, leveraging tooltips to give the data the level of granularity it needs. Several UI tweaks were made to create our group solution seen below.

Once we reached the execution phase of the project, we realized there were several problems with this mockup. The map shows the grade of a restaurant in the year chosen on the slider. The heatmap displays the frequency with which a cuisine was penalized for a violation, and the barchart shows the distribution of grades for the cuisines. Menus at the bottom are supplied to allow the user to filter by grade, cuisine, and violation.

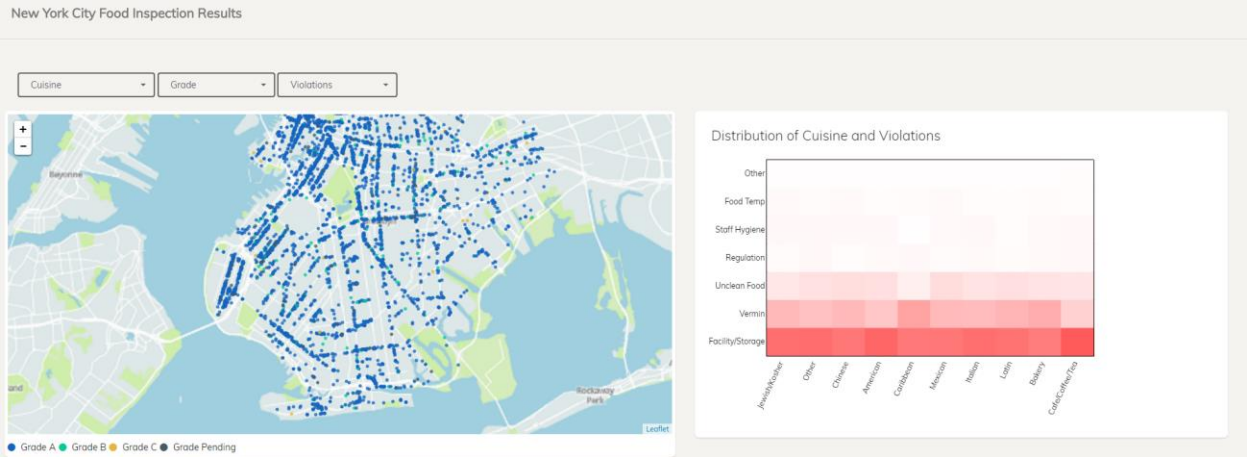


The above images show the components of the project while they were in progress, and they point to two major issues. First, the scope of the dataset was too large to effectively display on a map. Areas such as Chinatown and Little Italy have extraordinarily high restaurant densities, and there were simply too many restaurants in general. This resulted in long load times and severe clutter in the map. We eventually decided to limit the scope of this project to the Brooklyn area.

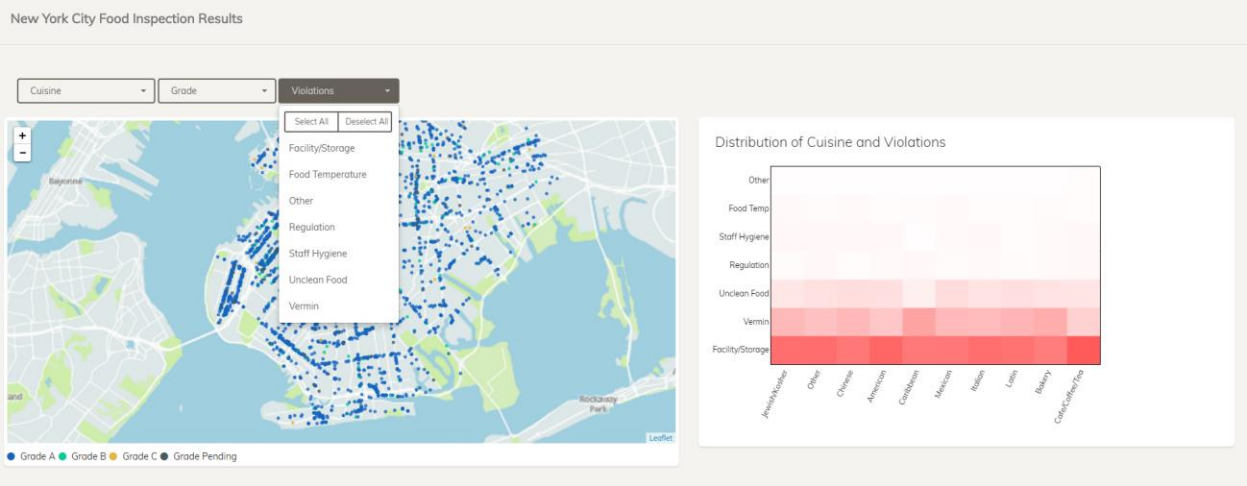
Second, the dataset was too varied. In all, there are roughly 22 different classes of violation, each with 1-15 subclasses, alongside 84 different cuisine types. As such, a heatmap attempting to display them all would be difficult to read and very sparse, and the menus for the filtration

steps would become difficult to navigate. To deal with this, we decided to have the heatmap only show the most populous cuisines plus the ones selected by the user. We also categorized the violation types into 7 broad categories.
Additionally, during this step, we found that the bar chart did not offer much information that the map was not already and it was therefore deemed unnecessary. The year slider also did not do much to answer the original questions presented, and was also eliminated. Time based data is instead displayed in the line charts showed through mouseover. These design decisions brought us to our final product.

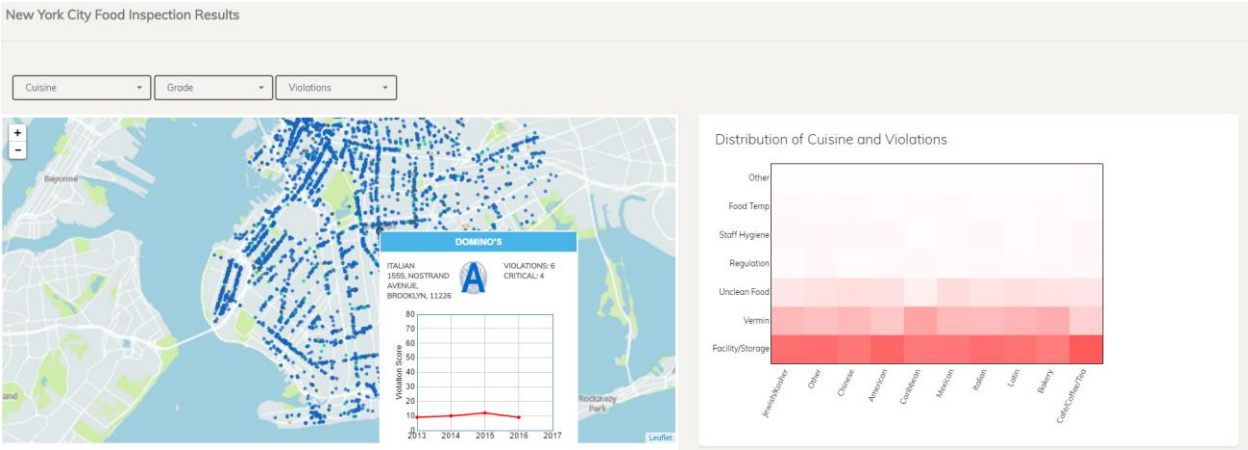
Final Visualization



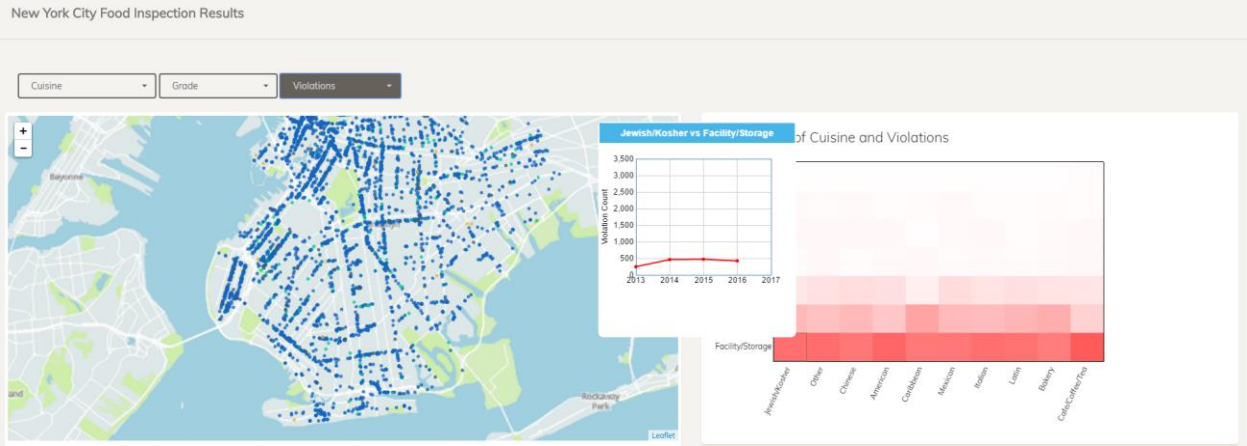
The visualization is a minimalist single-view system with a map to display geographic data and a heatmap on the right to display data based on cuisines. The dot color indicates the inspection grade corresponding to the current year (2016). The heatmap displays the 10 largest cuisines in the dataset, as well as those specifically chosen by the user. White indicates that a small portion of a cuisine suffers from a violation while red indicates that a large portion of a cuisine suffers from a certain type of violation. More information can be obtained by utilizing filtering options and mouseover interactions.



Clicking on the three dropdown menus give the ability to filter out dots on the map based on cuisine, grade and violations. The heatmap will also update to show the selected cuisines, although it will not update according to grade and violation filtering.

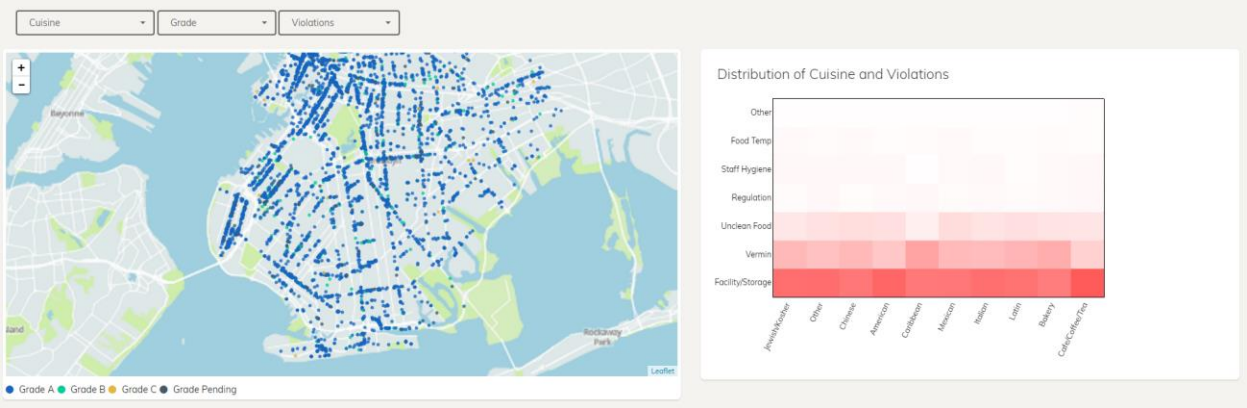


Mousing over a restaurant brings up a tooltip that shows the restaurant's current grade and its score for all years in which the relevant data is available. It is important to note that score in food inspections operate similarly to points on driver's licenses. That is, lower scores indicate fewer violations, and therefore cleaner restaurants.



Similarly, mousing over a rectangle in the heatmap brings up a tooltip with historical data for the number of restaurants in a certain cuisine suffering from a certain violation.

Findings

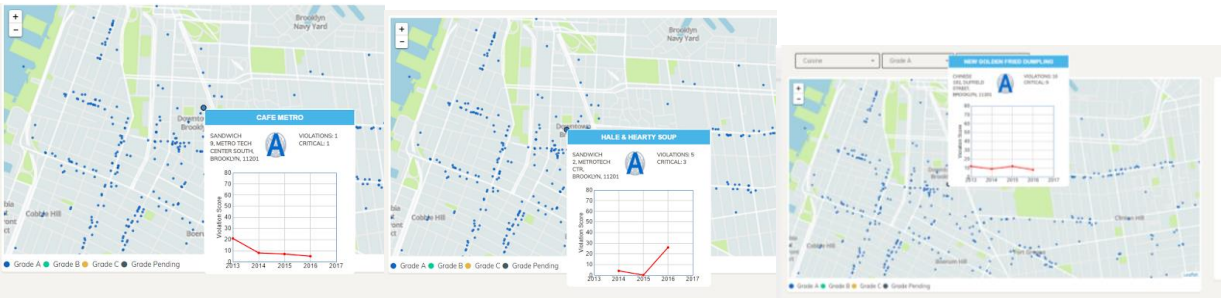


Upon first glance, several things become immediately clear about the state of restaurants in Brooklyn. First, the A grade is exceedingly common across Brooklyn. Also, Facility/Storage based violations are the most common type across all cuisines. Facility/Storage refers to violations that involve proper storage of food and facility maintenance such as “plumbing not properly installed or maintained”. We can surmise that as a whole, restaurants in Brooklyn have problems with the facilities more than anything else. On the other hand, we can see clearly that Food Temperature, Regulation, and Staff Hygiene are fairly uncommon. Given this, it seems almost as if the restaurants are trying their best, but are ultimately victims of the buildings they exist in.

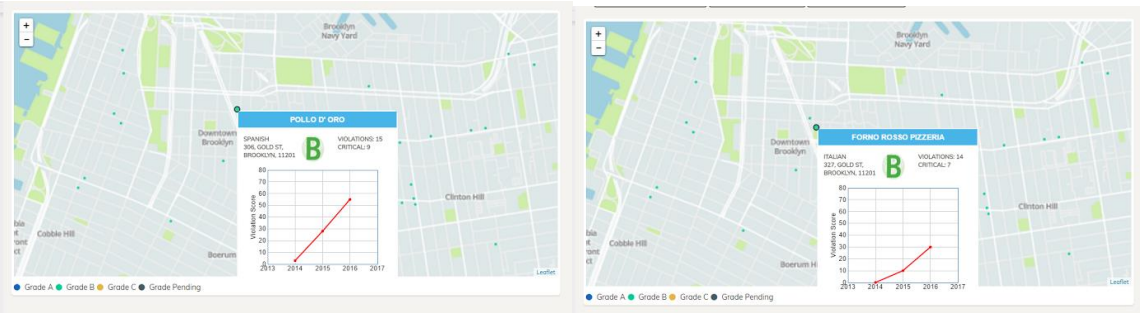
Having established a general feel for what our data looks like, we can begin looking to answer our driving analytical questions.

The question of whether restaurants improve is one that must really be answered at the individual level. The level of noise and variance introduced by the high volatility of the

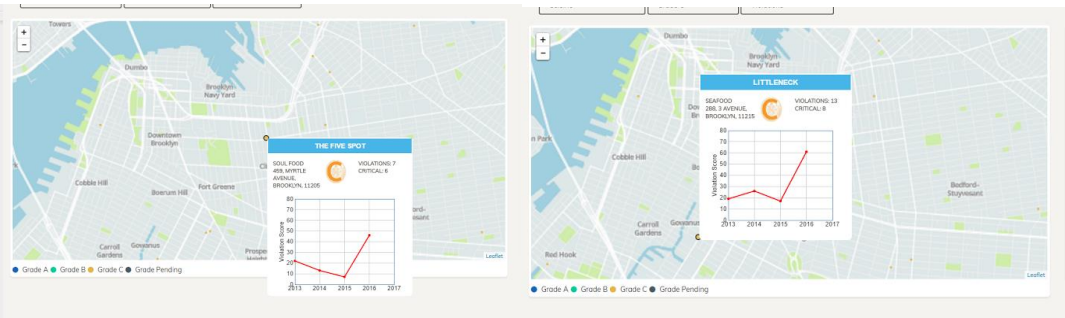
restaurant market makes most aggregation-over-time techniques unhelpful. We will look at several case studies of restaurants in the Metrotech area, and for robustness, we will look at a few restaurants of each grade level, to see how they changed.



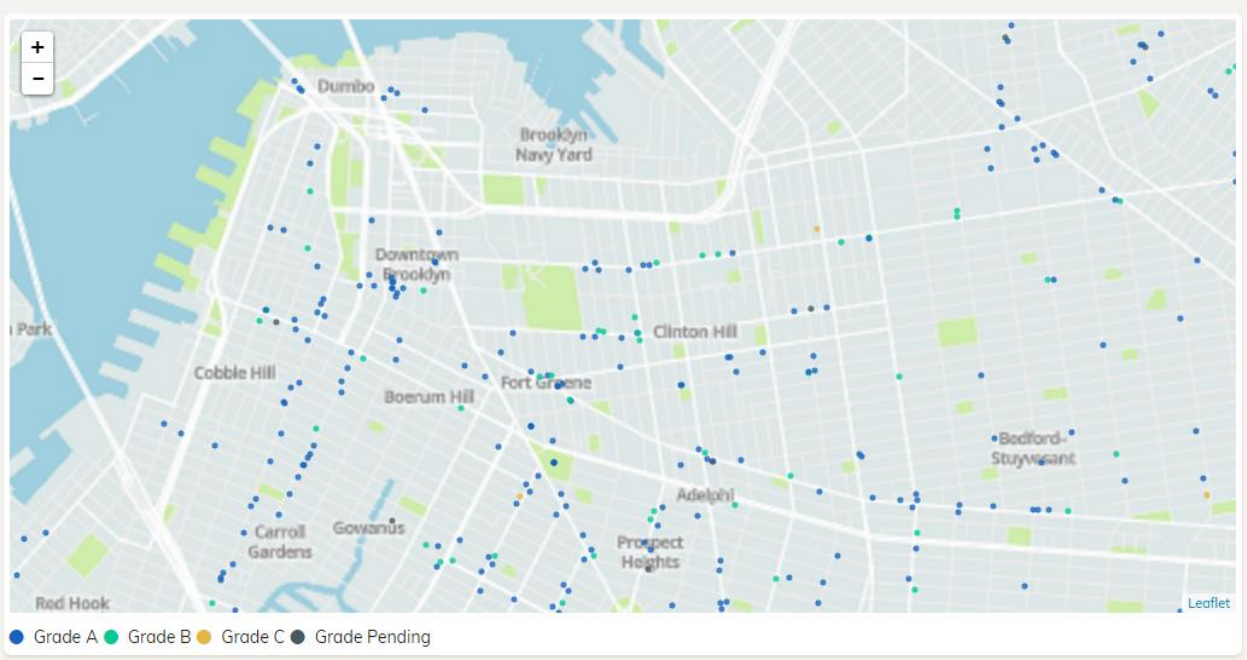
Above are the tooltips for three restaurants in the Metrotech Center area with A grades. When reading these charts, it is important to keep in mind that lower scores are better. There are of course many more restaurants, but these three show the general trends that a restaurant can follow. Cafe Metro shows a restaurant that started with a mediocre score, but managed to improve it in the last few years. Hale & Hearty Soup, on the other hand, seems to tell the opposite story - one of a clean restaurant perhaps getting sloppy. Then there are the consistently high-performing restaurants such as New Golden Fried Dumpling. What we see is that the history of Grade A restaurants can vary greatly, and because there are so many of them, it is exceedingly difficult to make any sweeping statements about this category.



The first thing that is apparent when looking at Grade B restaurants is their general sparsity in the Metrotech area. Whether this is because Brooklyn is a great place with many clean restaurants or because the grading system is too lenient is a debate for another time. However, of the few restaurants that do have a B grade in this area, one trend is exceedingly clear. Some B grade restaurants start as A grade restaurants, but then have their scores increase for some reason or another. As would be expected, there are some variations and exceptions among the B Grade restaurants, but many of them conform to this trend. Additionally, because of how sparse these data points are, we can rule out the possibility that some large-scale change in restaurant culture or inspection methodology was the cause of this pattern. If there were, there would be far more restaurants displaying this behavior.

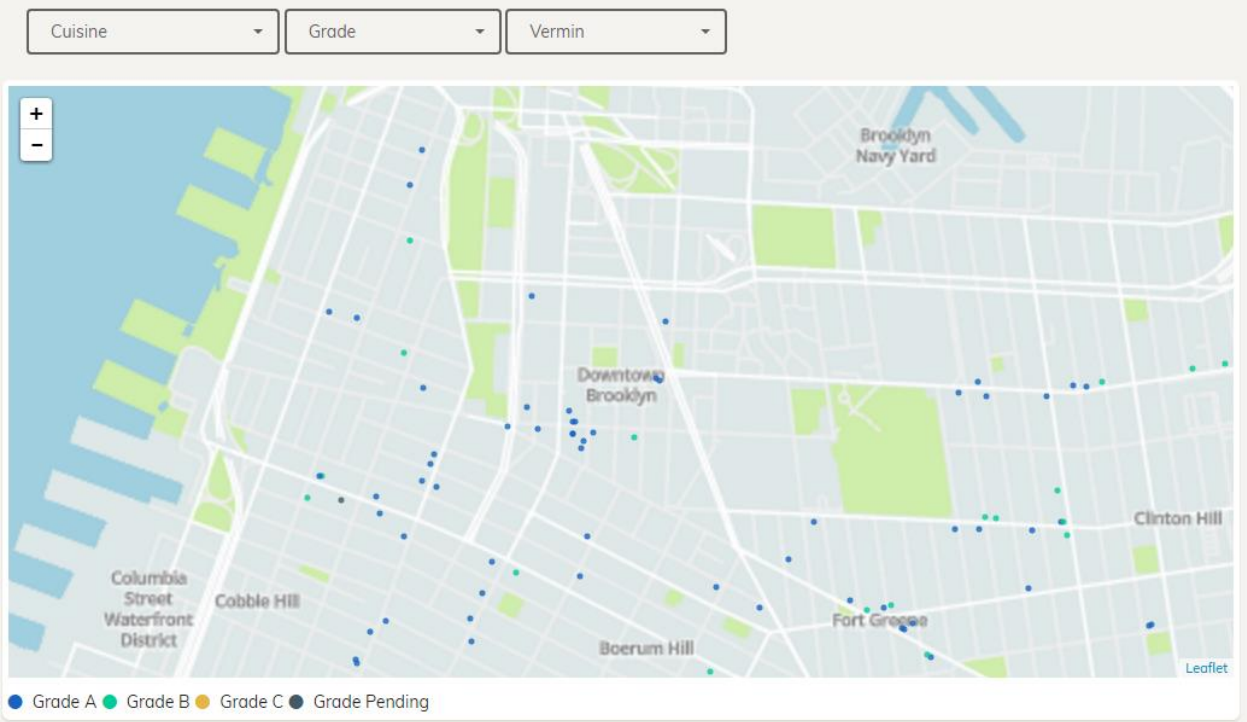


C Grade restaurants tell a very interesting story. One notable trend among restaurants is that they rarely stay in very high scores that would justify a C grade for very long. This makes sense, as a restaurant with a high score will either improve or shut down. More notable, perhaps, is how these restaurants ended up with a C grade. These restaurants generally had reasonable scores up until a massive spike in their violation count brough their score up significantly. We can only guess what happened to cause this, but we can only assume that some major and sudden change happened in these restaurants that caused a significant decline in cleanliness.



We previously mentioned that a significant portion of restaurants obtain an A grade, but that may not be as informative as we would perhaps want it to be. The above shows all of the restaurants with vermin-related violations, and many of these still have a grade of A. Having seen this, we can begin digging into trends as they relate to violations.

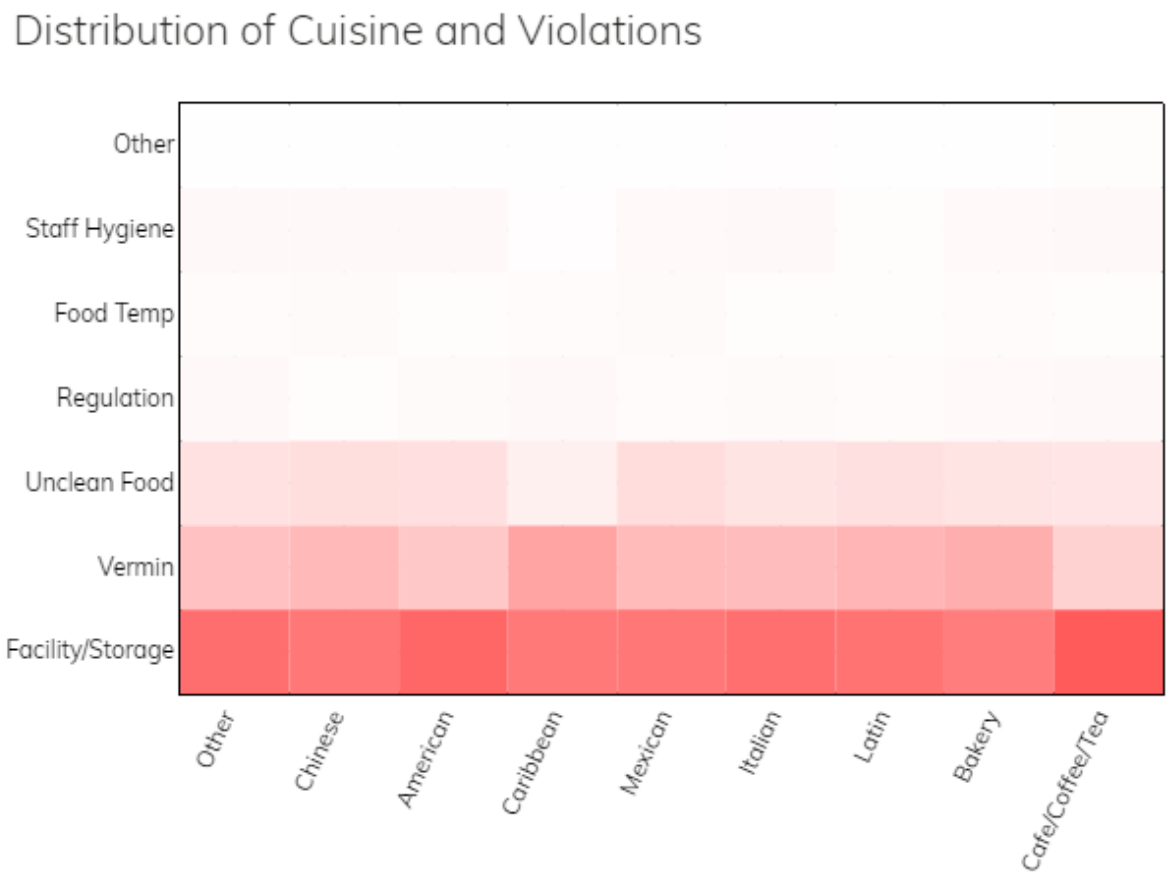
One question of particular interest is seeing whether or not certain violations are more common in some areas than others. In particular, if there is a large cluster of restaurants has a vermin problem, then perhaps we can interpret that as a vermin infestation in that area.



By filtering the restaurants down to only the ones with vermin violations, we can see that there is a small cluster of restaurants with vermin violations near the Metrotech Center area. At this point, we have to acknowledge that there are a lot of restaurants in that area, and this is perhaps statistically consistent. However, we can actually get a general feel for whether this is really proportional by comparing it to the base map.

A side by side comparison shows that the area to the far west, south, and east all have fairly large clusters of restaurants, but the area around Metrotech center still has a fair number of vermin violators relative to its restaurant density. This is somewhat of an eyeball measure, but it is still concerning.

Perhaps the distribution of violations in an area is somehow related to what kinds of restaurants are popular. Perhaps a certain cuisine is far more susceptible to a vermin violation than others.



For this information, we can turn to the heatmap. In the heatmap, the darker the color, the more restaurants there are of this cuisine having this violation. Above is the violation distribution data for the 10 most common types of restaurants. We have already established that Facility/Storage issues are the most prevalent, but what about others? Vermin is the second most common violation type for most cuisines. We can see here that Cafes are less susceptible to vermin issues than bakeries or Caribbean restaurants. This makes a lot of sense on a number of levels. Research indicates that caffeine may be a naturally occurring insecticide, and so it stands to reason that Cafes, which specialize in caffeine-rich beverages, would be less susceptible to insects. On the other hand, bakeries produce a large quantity of sugar-rich foods, which pests are naturally attracted to. It is somewhat difficult to see here, but Mexican, Latin, and Chinese restaurants appear to have slightly more problems with food sanitation than others. However, the difference here is small, and it is difficult to say whether this is statistically significant.

It is at least comforting to know that restaurants are generally free from problems relating to staff hygiene and food temperature. Brooklynites can rest easy knowing that they are fairly unlikely to contract salmonella, or eat food touched by unwashed hands. On the other hand, one could easily argue that vermin and food cleanliness are much more important things to be worried about in restaurants, in which case, this is good cause for concern.

Limitations and Future Works

While we are generally happy with the progress made in the short time given, this project currently has several major limitations. While this was ultimately a design choice to cut down on clutter and load times, it is unfortunate that the data had to be limited to the Brooklyn area. It is conceivable that different boroughs would have different food cultures, and with that, a different distribution of cuisines and violations. It would have been interesting to see whether these findings hold up in Manhattan or Queens. It would also make for interesting future work to be able to take borough or city level statistics and compare them to one another.

The current visualization also has several issues with regard to the way some of this data is presented. For example, the system's representation of score is somewhat misleading. In health inspections, score is representative of the number and severity of violations for a restaurant. Therefore a lower score is better. However, the way it is currently framed now represents a worse restaurant appearing higher in the line chart. This is problematic, as there is an implicit bias with line charts wherein a higher number is considered positive, especially when the term that is increasing is called "score". With more time to develop, we would have experimented with inverting the graph and removing the scale on the y-axis, to properly convey the semantics of the graph without confusing a viewer with the caveats that inevitably come with inverting a y-axis.

There is also a potential issue with the way cuisines and violations were preprocessed. Cuisines were generally left alone, save for a few odd edge cases. For example, Pizza, Italian, and Pizza/Italian are three different cuisines with a fair number of restaurants in each. These were consolidated to ensure the data was representative of what one would expect, but it is very possible information was lost in this step. Violations were much more aggressively consolidated into 6 major categories. This was done by hand, and so it is difficult to make any guarantees regarding the integrity of these categorizations. Several of them could arguably fit into multiple categories, but our design required that they be pushed to the one we felt was the best fit, which may not necessarily be the case.

One major issue deals with the dataset itself. A restaurant with one cockroach and a restaurant practically made of cockroaches will both get filed under the cockroach violation. Violation severity would have been a particularly interesting piece of data when investigating relations between investigation type and geographical region. However, severity data is unavailable in the dataset, and as such could not be represented.

Additionally, the filtration techniques this visualization allows suffer from the lack of an anchor point to compare to. Answering the question of whether certain areas suffer from a type of violation involves filtering for vermin violations and attempting to identify clusters of restaurants. However, because clusters of restaurants occur naturally in Brooklyn, it is difficult to tell if a cluster of vermin restaurants occurs because that area has a vermin problem, or if it is simply because there is a cluster of restaurants there. One possible fix for this is to dim restaurants that have been filtered out, rather than remove them entirely.

One last limitation of this system is the inability to see specifically what restaurant has which violation. This is in part by design, as this doesn't really answer any of the questions we had,

but in retrospect, it might have yielded some interesting information pertaining to whether restaurants improve over time.