# An Analysis of Food Establishment Inspections in Boston

*CS 591 Final Report*
*May 1, 2015*

COURTNEY PACHECO

Data from **yelp** and **cityofboston.gov**

# 1    Introduction

For my project, I chose to analyze the official Food Establishment Inspections dataset from the City of Boston.[1] This dataset is updated daily and contains a variety of detailed information regarding the health inspections of licensed food establishments in Boston. Overall, this dataset covers almost an entire decade's worth of health inspections, with some of the information dating back as far as December 27, 2006.

In addition to using this dataset, I also gathered Yelp user reviews for each restaurant in the dataset to explore possible relationships between restaurant health code violations and Yelp user reviews. Although Yelp does not have a physical dataset that can be downloaded, they offer a free API to extract information from their database.[2]

## 1.1    Project Goals

The goals of my project were as follows:

1. To investigate the possible relationship between number of violations (and the severity of such violations) and Yelp user ratings. I hypothesized that restaurants with high user ratings will have a lower number of health code violations than restaurants with low user ratings. However, it is possible that no relationship exists.

2. To investigate the possible relationship between number of violations (or type of violations) and restaurant location. This goal involved clustering the data using geolocation and violation data. (No datetime objects will be used here.)

3. (Just for fun) To investigate hospital food establishment inspections. For example, Massachusetts General Hospital had serious food contamination issues (they got caught having food touch the floor). I wanted to compare hospitals to one another in this respect, and again compare this data to Yelp's data.

## 1.2    Food Inspections Dataset: Format & Overview

The dataset can be downloaded in a variety of different formats, including JSON, CSV, and XML. I chose to work with the CSV format for the purposes of loading the dataset into a Pandas DataFrame object.

In total, there are 26 different columns and 316,889 rows of data, which makes the dataset quite large. On the next page is a table that describes each column of data:

| Column Title | Data Type | Description | Missing Values? |
| --- | --- | --- | --- |
| BusinessName | string | Name of the food establishment | No |
| DBA Name | string | Trade name of the food establishment. (DBA = "Doing Business As"). For example, the restaurant "Extreme Pita" on BU campus has its business name listed as "Extreme Pita", but its DBA name listed as "Trustees of Boston University" | Yes |
| LegalOwner | string | Legal owner of the business. (The legal owner is typically the parent company's name. e.g., ARAMARK) | Yes |
| NameLast | string | Owner's last name (Could be a person's last name OR a company name) | No |
| NameFirst | string | Owner's last name (if the owner is a company, then this field is blank) | Yes |
| LICENSENO | int64 | The food establishment's business license number | No |
| ISSDTTM | datetime | Issue date for food license | Yes |
| EXPDTTM | datetime | Expiration date for food license | Yes |
| LICSTATUS | string | License status. Takes on 1 of 3 values: {Active, Deleted, Inactive} | No |
| LICENSECAT | string | Food license category. Takes on 1 of 3 values: {MFW, FS, FT} · MFW = "Mobile Food Walk On" · FS = "Eating & Drinking" · FT = "Eating & Drinking W/ Take Out" | No |
| DESCRIPT | string | Description of license category. Takes on 1 of 3 values: (1) "Mobile Food Walk On", (2) "Eating & Drinking", and (3) "Eating & Drinking W/ Take Out" | No |

| Column Title | Data Type | Description | Missing Values? |
|:---:|:---:|:---:|:---:|
| RESULT | string | Result of inspection. Takes on 1 of 13 possible values:<br>(1) HE_fail = failed inspection<br>(2) Fail = failed inspection<br>(3) Failed = failed inspection<br>(4) HE_pass = passed inspection<br>(5) Pass = passed inspection<br>(6) HE_filed = "Pass w/ minor violations"<br>(7) HE_closure = closed<br>(8) HE_FAILNOR = ?<br>(9) HE_hearing = need to attend a hearing<br>(10) HE_NotReq = "not required"?<br>(11) HE_Misc = miscellaneous<br>(12) HE_OutBus = ?<br>(13) HE_TSOP = temporary suspension of permit | No |
| RESULTDDTM | datetime | Date and time "RESULT" was entered | Yes |
| Violation | string | Violation code from [3] | Yes |
| ViolDesc | string | Violation description, in words | Yes |
| VIOLDTTM | datetime | Date and time of violation | Yes |
| ViolStatus | string | Violation Status: NaN, Pass, or Fail | Yes |
| StatusDate | datetime | Sometimes information in the database is updated for certain restaurants. StatusDate marks when the data entry was updated. | Yes |
| Comments | string | Health inspector?s comments. (e.g., "elevate food 6 inches off floor") | Yes |
| Address | string | Physical address of the restaurant | Yes |
| City | string | These values are just the neighborhoods of Boston (e.g., Roxbury, Roslindale, Mission Hill, etc.) | Yes |

| Column Title | Data Type | Description | Missing Values? |
|---|---|---|---|
| State | string | State the restaurant is located in | Yes |
| Zip | int64 | Restaurant's zip code | Yes |
| Property_ID | float64 | Unique ID for the property. (Several properties have a Property_ID of 0, so this Property_ID column might not be very meaningful.) | Yes |
| Location | string | Geolocation of the restaurant | Yes |

# 2 Data Collection

As you have noticed, many of the columns in the Food Establishment Inspections dataset are missing data entries. Fortunately, I only needed to recover data in the "Location" column for my analysis. Therefore, my data collection piece consisted of two parts:

1. Recovering missing geolocation data from the main dataset

2. Extracting data from Yelp for each restaurant in the main dataset

## 2.1 Recovering Geolocation Data

To recover missing geolocation data, I wrote a script called `FixLocations.py`. This script loads the main CSV file to a Pandas DataFrame, then uses GeoPy [4] to make geolocator requests to OpenStreetMap[5] for each missing entry in the DataFrame. I could have used Yelp to obtain geolocation data, but I elected to use GeoPy because the geolocations returned from GeoPy appeared to be more accurate than Yelp's geolocations. (Each time you request restaurant information from Yelp, Yelp's response contains a geolocation for that restaurant and the accuracy of the geolocation, which is given in terms of `longitude delta` and `latitude delta`. Some of these delta values are rather large – i.e., > 0.1.)

Because some of the restaurant entries in the dataset are either missing address information or have incomplete address information, it was not possible to recover all of the missing geolocations. For example, one address I obtained from the dataset was "466 Centre". The actual address is "466 Centre St. Jamaica Plain, MA", and it belongs to a restaurant named Acapulco Mexican Restaurant. In `FixLocations.py`, you will see how I attempted to handle those situations.

Overall, I was missing `99613` out of `316889` geolocation entries in the original dataset, which meant I was missing a total of 31% of my geolocation data. After using GeoPy to recover this missing data, I was only missing `644` locations, or 0.2% of my geolocation data.

Partial output, saved as a Pandas DataFrame:

```
                     Address        City State  Zip Property_ID Location
23322                    NaN         NaN   NaN  NaN         NaN      NaN
23374                    NaN         NaN   NaN  NaN         NaN      NaN
38465       50   Providence ST    Boston   NaN  NaN         NaN      NaN
38466       50   Providence ST    Boston   NaN  NaN         NaN      NaN
38467       50   Providence ST    Boston   NaN  NaN         NaN      NaN
...                      ...         ...   ...  ...         ...      ...
308607     334  Massachusetts AV  BOSTON   NaN  NaN         NaN      NaN
308608     334  Massachusetts AV  BOSTON   NaN  NaN         NaN      NaN
308609     334  Massachusetts AV  BOSTON   NaN  NaN         NaN      NaN
308610     334  Massachusetts AV  BOSTON   NaN  NaN         NaN      NaN
308611     334  Massachusetts AV  BOSTON   NaN  NaN         NaN      NaN

[644 rows x 26 columns]
```

You can see that some of the geolocations were not found because `State` had a value of `NaN`. In hindsight, I should have looked for indices in the DataFrame where `State` had a value of `NaN`, then replaced those `NaN` values with `MA` (for "Massachusetts"). Or even simpler, I should not have constructed each restaurant's address using the `State` entry in each row of the DataFrame. I should have simply assumed the state was `MA` for each entry. (Unfortunately, it literally took me 4 days to obtain these geolocations, so it was not realistic for me to rerun this code.)

Technically speaking, I could have gathered my Yelp data as I was recovering my Geolocation data, but I did not use Yelp at this time for a couple of reasons:

1. It took me a few days to to obtain my geolocations from GeoPy. (GeoPy is a bit slow...)

2. Yelp's API times out (A) after you have made too many requests in a short period of time (you need to wait at least 1 second between requests), or (B) when you have consistently made calls to their API over the course of a day or so. (Yelp measures how often you make calls to their API and for how long you have been making those calls.) When Yelp's API times out, all of the values returned are `NaN`.

Therefore, I ran `FixLocations.py` overnight while I worked on gathering Yelp data separately. I saved the results from `FixLocations.py` as `fixed_locations.csv`.

## 2.2   Gathering Yelp Data

The script I used to gather Yelp data is called `GetYelpData.py`. It takes about 3 hours and 15 minutes to scrape all of the necessary data from Yelp. This script imports another script called `Yelp.py`, which is the official Yelp API script for pulling data from Yelp (i.e., I did not write

`Yelp.py`.)

In order to run this code, you need these 4 authentication tokens from Yelp:

```
CONSUMER_KEY
CONSUMER_SECRET
TOKEN
TOKEN_SECRET
```

which I have removed from my `Yelp.py` script.

If those authentication tokens are inserted into `Yelp.py`, then `GetYelpData.py` will obtain the following information for each restaurant:

1. `LICSTATUS` = Status of the license (`Active` or `inactive`)

2. `Pass` = number of health inspections passed

3. `Fail` = number of health inspections failed

4. `Total` = total number of health inspections

5. `License Type` = one of 3 license types: `FS`, `FT`, or `MFW`

6. `Address` = address of the restaurant

7. `Location` = geolocation of the restaurant

8. `Yelp Rating` = average Yelp rating

9. `Review Count` = number of Yelp reviews

10. `Neighborhood` = neighborhood of Boston the restaurant is located in

11. `Categories` = categories of food the restaurant sells

The above information is stored in a dictionary called `RESTAURANTS`, which is saved as a Python pickle file `RESTAURANT_INFO.p`.

# 3  Data Analysis

## 3.1  Yelp User Ratings vs. % Health Inspections Failed

The script `AnalyzeRatings.py` loads the file `RESTAURANT_INFO.p` and analyzes its contents. (Note: This pickle file was obtained from `GetYelpData.py` in the previous section.) This script generates Figures 1-3 (which you see on the next couple of pages), and it takes under 5 seconds to run on my Macbook Air.

In Figure 1, we do not see a relationship between average Yelp user rating and percentage of health inspections failed. Likewise, we do not see a relationship between average Yelp user

rating and percentage of health inspections passed. In both cases, $R^2 \approx 0$. These results indicate that Yelp user rating is *independent* of health inspection pass/fail rates.

These results can be partially explained by the two plots on the right side of this figure. In the plot on the upper right, we see a relationship between average violation level and percentage of health inspections failed ($R^2 \approx 0.5$). However, from looking at the distribution of data points in this plot, we can see that health inspections are somewhat subjective. For example, some restaurants have failed a large percentage of health inspections while maintaining an average violation level of 1, whereas other restaurants have failed relatively few health inspections while having an average violation level of 2 or greater.
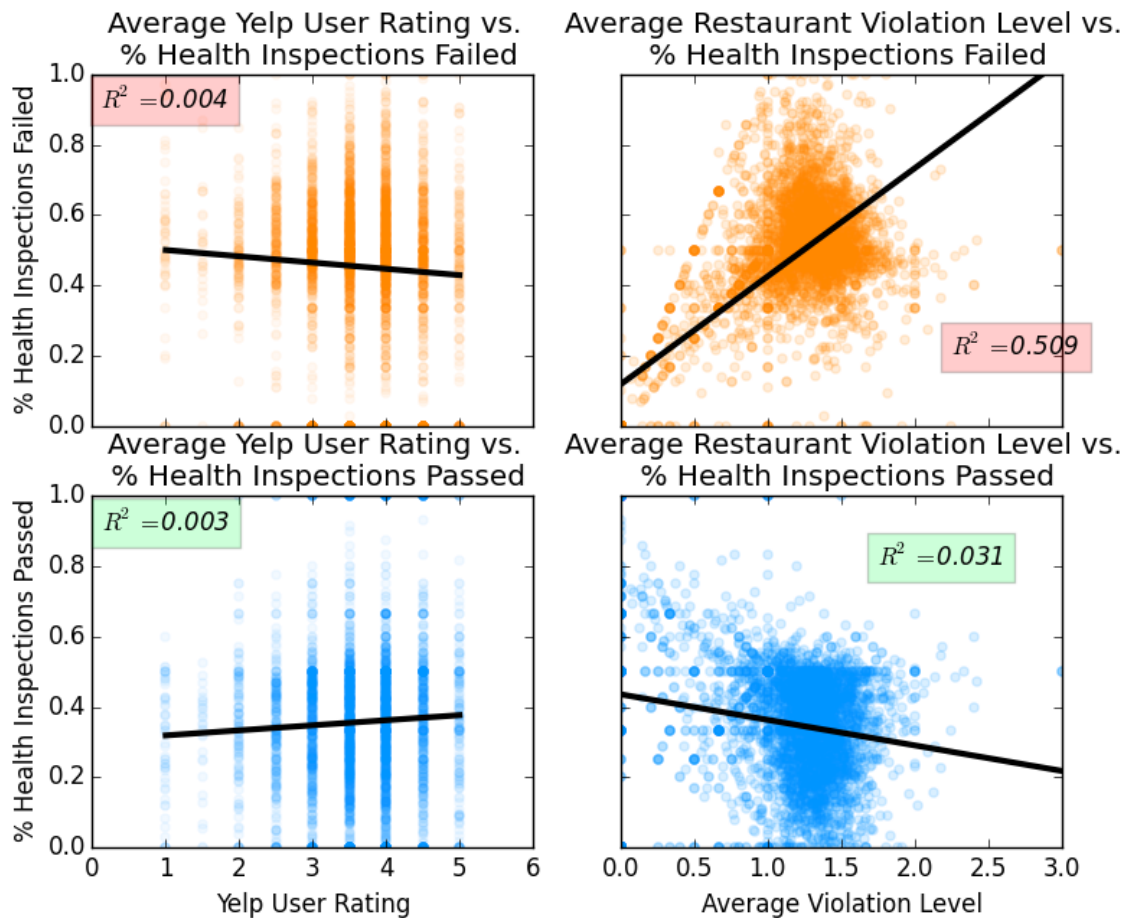


Figure 1: *Average Yelp Ratings vs. Health Inspection Results.* The plot in the upper left corner compares the average Yelp user rating to the percentage of health inspections failed. Similarly, the plot in the lower left corner compares the average Yelp user rating to percentage of health inspections passed. (Note that not all health inspections result in a pass or fail, as indicated by the table in section 1.2 – hence these two plots are necessary.) The two plots on the right side of this figure compare the pass and fail rates to the average violation level. In all of these figures, each dot represents a restaurant.

Perhaps more interesting is the fact virtually no relationship exists between percentage of health

inspections passed and average violation level (as seen in the figure on the bottom right corner).

While the results in Figure 1 suggest there is no relationship between average Yelp user rating and inspection pass/fail rates, I wanted to investigate whether or not geographical location had an influence on pass/fail rates. For example, is there a relationship between Yelp user rating and pass/fail rates for *some* neighborhoods? I also wanted to investigate whether or not geographical location had an influence on Yelp user ratings. Figures 2-4 on the next couple of pages show the distribution of data for Yelp rating, violation level, and % health inspections failed for each neighborhood. (For a brief overview of box plots, see [6].)
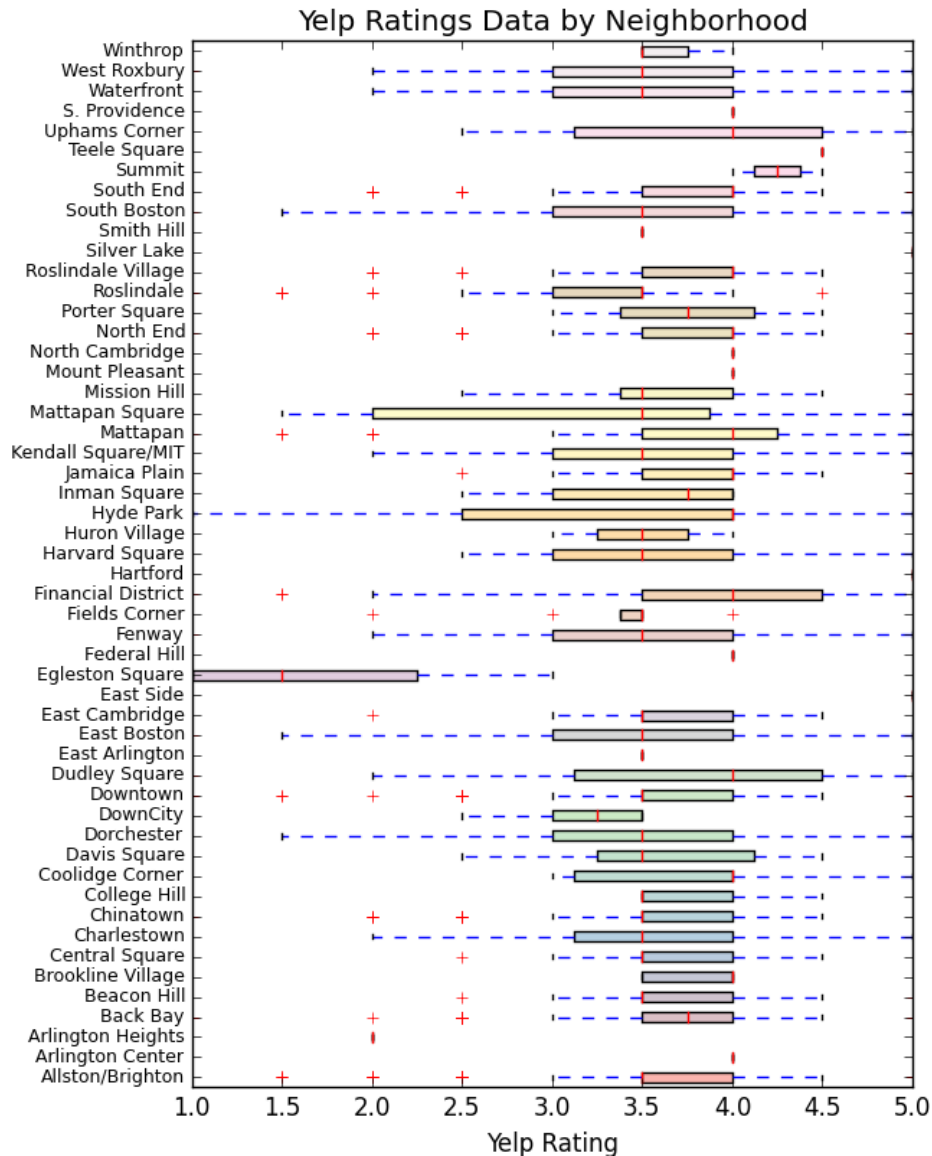


Figure 2: *Distribution of Yelp Ratings by Neighborhood.* This box plot shows the distribution of average Yelp user ratings for each neighborhood. The vertical red line on each bar represents the median ("middle quartile") Yelp rating for that neighborhood, and the red "+" signs represent the outliers for that neighborhood. Yelp user ratings can take on any value between 1 and 5, with 1 being abysmal and 5 being excellent.
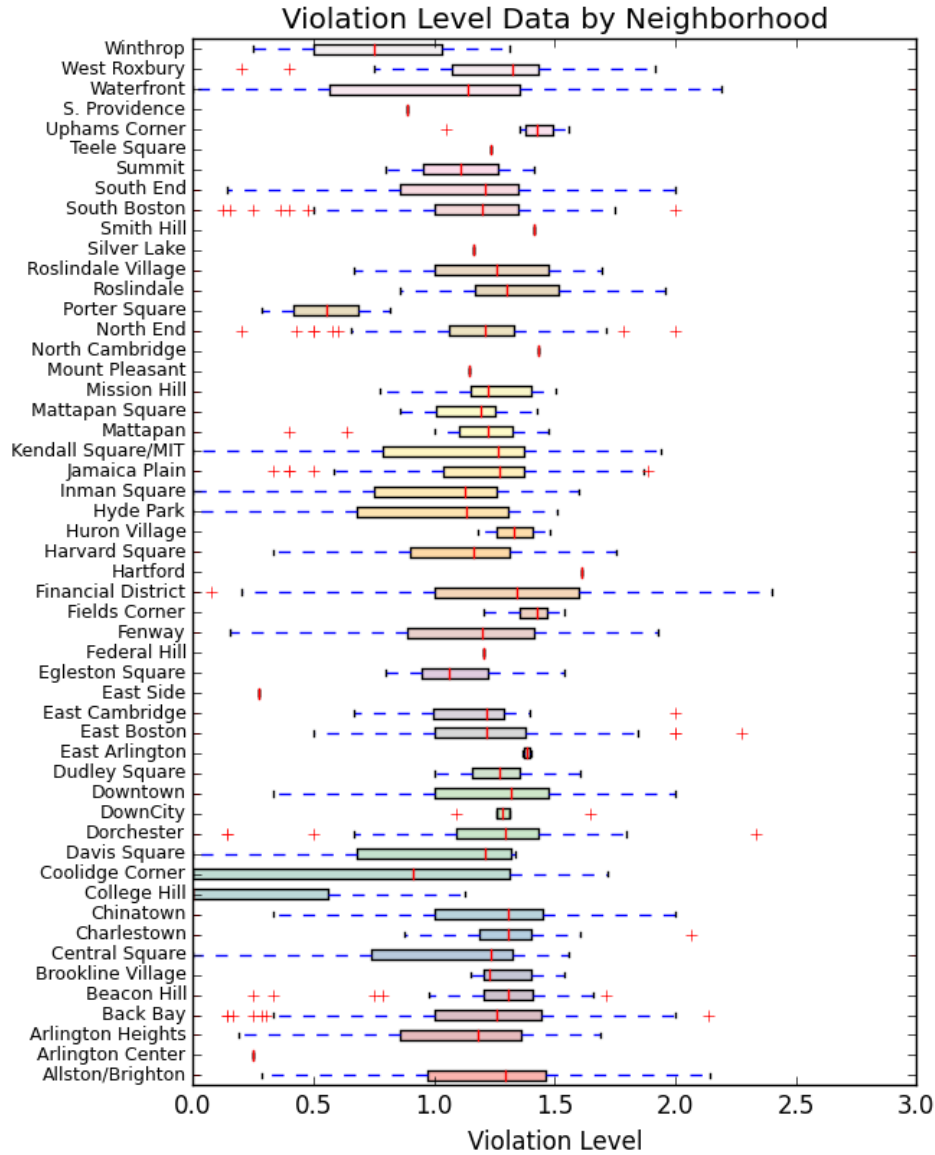
Figure 3: *Distribution of Average Violation Level by Neighborhood.* This box plot shows the distribution of violation levels for each neighborhood, where "violation level" can take on any value between 0 and 3, with 0 representing "no violation" and 3 representing the worst possible violation. Similar to Figure 2, the vertical red line on each bar represents the median ("middle quartile") violation level for that neighborhood, and the red "+" signs represent the outliers for that neighborhood.
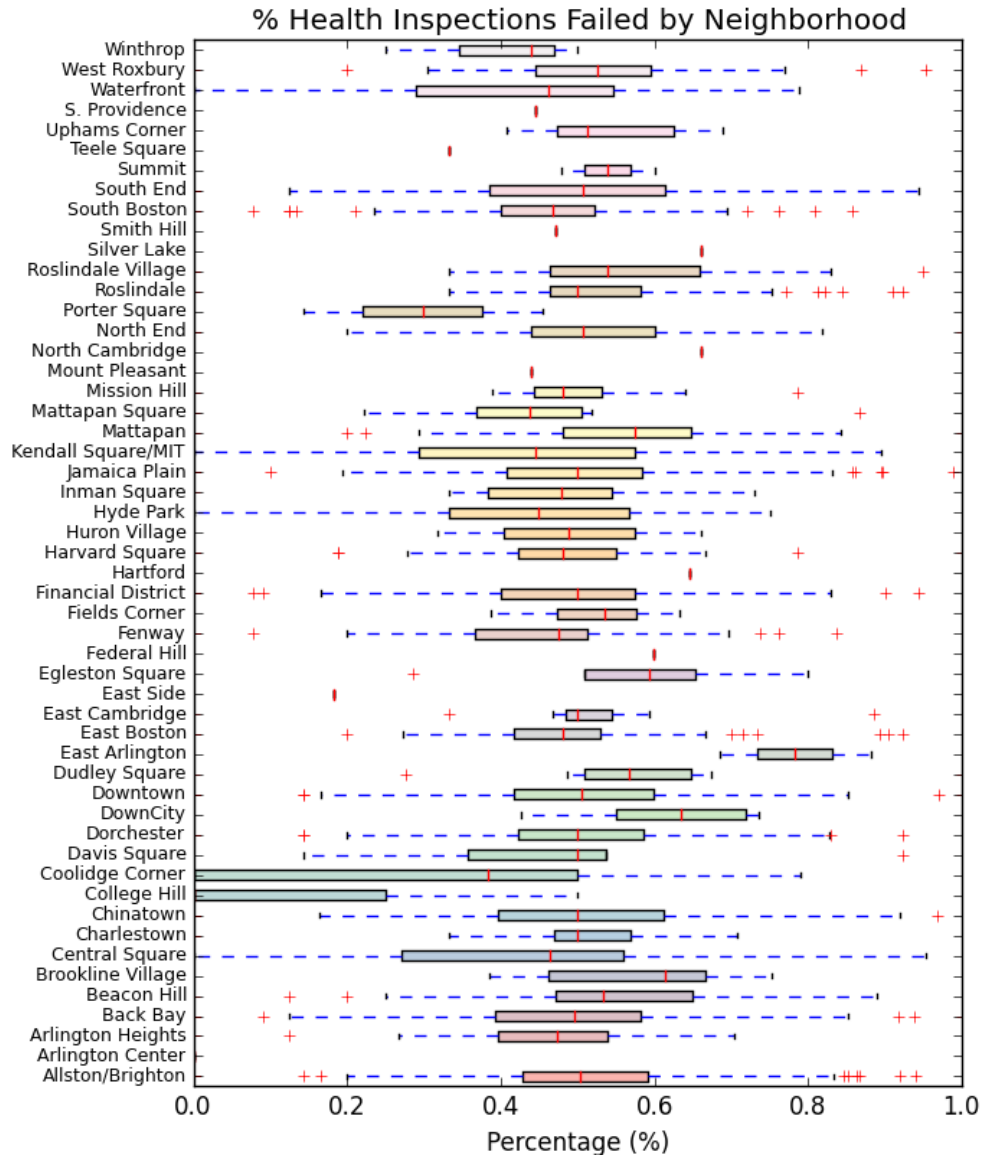
Figure 4: *Percentage of Health Inspections Failed by Neighborhood.* This box plot shows distribution of failed health inspection data for each neighborhood. Again, the vertical red line on each bar represents the median ("middle quartile") value for % health inspections failed for that neighborhood, and the red "+" signs represent the outliers for that neighborhood.

From the above box plots, we can make a few observations:

1. There are more than 23 neighborhoods in this dataset, which means some of these neighborhoods are *not* located in Boston. (Boston only has 23 neighborhoods.)

2. Some neighborhoods do not have a lot of data, but these particular neighborhoods appear to be neighborhoods that are not in Boston. For example, Silver Lake, Smith Hill, and Federal Hill only appear to have 1 data point in each figure. (Perhaps some restaurants were once located in Boston, but relocated to an area outside the city?)

3. *If* we exclude the neighborhoods that are not in Boston – i.e., the ones with very few data points – the distribution of data for all the neighborhoods are similar. (For example, neighborhoods like `Coolidge Corner`, `East Arlington`, `College Hill`, and `Porter Square` have odd distributions for Figures 2-4, but they also have very few data points to work with and are not located in Boston.) Notably, we see that the values for median Yelp Rating, the median violation level, and the median % health inspections failed are similar across all neighborhoods with in Boston itself.
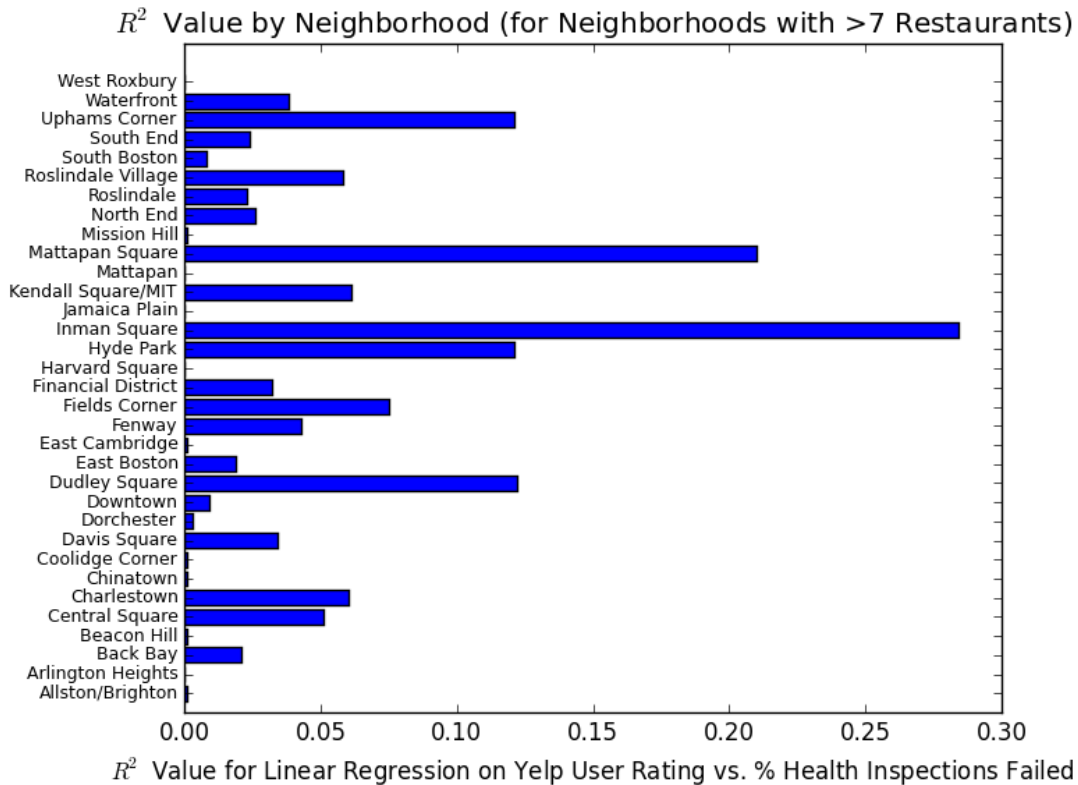


Figure 5: $R^2$ *Value by Neighborhood.* For each neighborhood, I extracted the Average Yelp Rating data and the % Health Inspections Failed data. I then performed a linear regression on the combined data, per neighborhood. (i.e., I used the same procedure as I did in Figure 1 to obtain the plot in the upper left corner, except instead of performing linear regression on the entire dataset, I performed linear regression on each neighborhood separately.) This figure reports the resulting $R^2$ value for each linear regression performed.

From Figure 5, we can see that only 5 neighborhoods (`Uphams Corner`, `Inman Square`, `Mattapan Square`, `Hyde Park`, and `Dudley Square`) appear to have a relationship between Yelp User Rating and % Health inspections Failed, and the relationship in both cases is weak. Moreover, of those 5 neighborhoods, only `Mattapan Square` and `Hyde Park` belong to Boston. We can hereby conclude that there really *isn't* a relationship between Yelp User Rating and % Health Inspections failed for Boston as a whole, nor is there a strong relationship between Yelp User Rating and % Health Inspections for certain neighborhoods of Boston. (Technically, an $R^2$ value of 0.1 to 0.3 indicates a very weak relationship, but given that some of these linear fits were performed on very small datasets (sometimes only 8 data points), we might expect our $R^2$ value to be artificially high.) Below is a list of the 33 neighborhoods I performed separate linear fits on. Reported next to each neighborhood is the $R^2$ value of the linear fit and the number of restaurants (data points) the linear fit was performed on.

|    | Neighborhood | R^2 Value | Num Restaurants |
|----|--------------|-----------|-----------------|
| 0  | Allston/Brighton | 0.001 | 206 |
| 1  | Arlington Heights | 0.0 | 18 |
| 2  | Back Bay | 0.021 | 394 |
| 3  | Beacon Hill | 0.001 | 51 |
| 4  | Central Square | 0.051 | 14 |
| 5  | Charlestown | 0.06 | 50 |
| 6  | Chinatown | 0.001 | 124 |
| 7  | Coolidge Corner | 0.001 | 38 |
| 8  | Davis Square | 0.034 | 8 |
| 9  | Dorchester | 0.003 | 172 |
| 10 | Downtown | 0.009 | 157 |
| 11 | Dudley Square | 0.122 | 14 |
| 12 | East Boston | 0.019 | 182 |
| 13 | East Cambridge | 0.001 | 11 |
| 14 | Fenway | 0.043 | 84 |
| 15 | Fields Corner | 0.075 | 8 |
| 16 | Financial District | 0.032 | 157 |
| 17 | Harvard Square | 0.0 | 50 |
| 18 | Hyde Park | 0.121 | 34 |
| 19 | Inman Square | 0.284 | 8 |
| 20 | Jamaica Plain | 0.0 | 123 |
| 21 | Kendall Square/MIT | 0.061 | 31 |
| 22 | Mattapan | 0.0 | 39 |
| 23 | Mattapan Square | 0.21 | 10 |
| 24 | Mission Hill | 0.001 | 20 |
| 25 | North End | 0.026 | 163 |
| 26 | Roslindale | 0.023 | 38 |
| 27 | Roslindale Village | 0.058 | 28 |
| 28 | South Boston | 0.008 | 138 |
| 29 | South End | 0.024 | 168 |
| 30 | Uphams Corner | 0.121 | 10 |
| 31 | Waterfront | 0.038 | 199 |
| 32 | West Roxbury | 0.0 | 54 |

## 3.2 Hospitals: Yelp Rating vs. % Health Inspections Failed
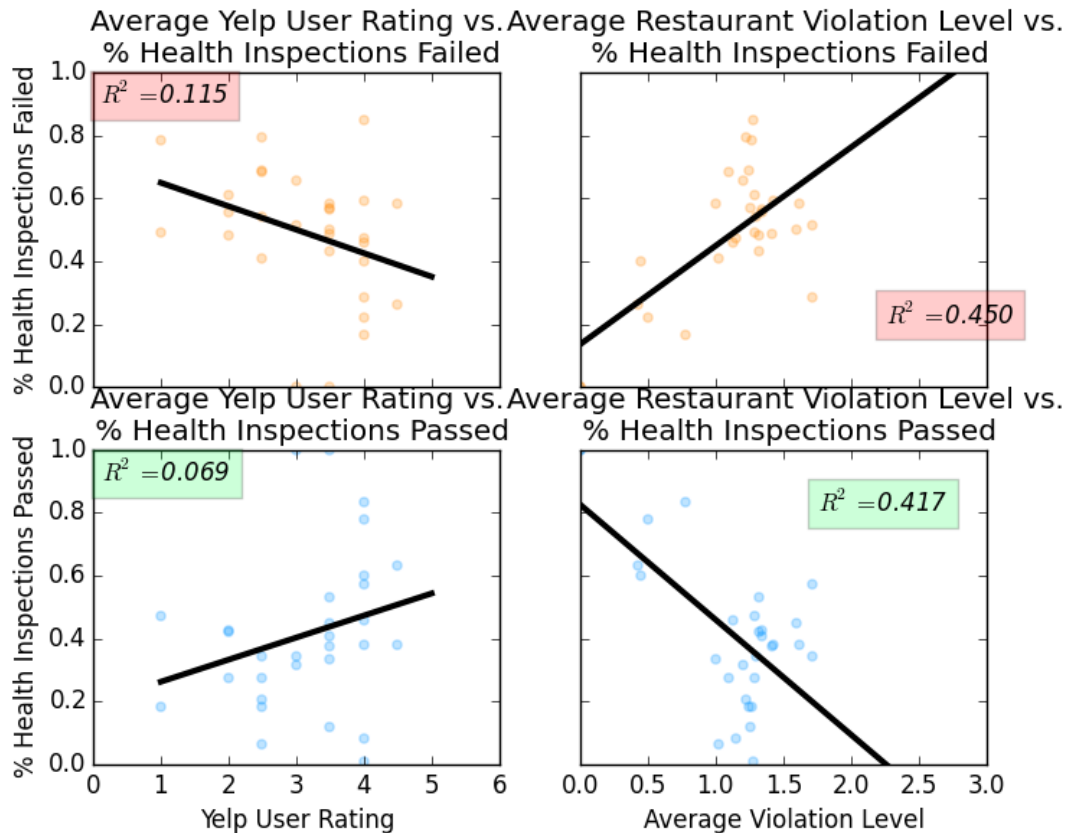


Figure 6: *Average Yelp Ratings vs. Health Inspection Results for Hospitals in Boston.* The plot in the upper left corner compares the average Yelp user rating to the percentage of health inspections failed for all the hospitals in Boston. Similarly, the plot in the lower left corner compares the average Yelp user rating to percentage of health inspections passed for all the hospitals in Boston. The two plots on the right side of this figure compare the pass and fail rates to the average violation level. In all of these figures, each dot represents a hospital.

As we expected in Figure 6, there isn't much of a relationship between average Yelp user rating and inspection pass/fail rates for hospitals. Though, it's worth noting that the $R^2$ values for all of the plots, except the plot in the upper right corner, are larger in this figure than the $R^2$ values in the corresponding plots in Figure 1. (However, it is certainly possible that these $R^2$ values are higher than the $R^2$ values in Figure 1 simply because there are much less data points to work with in this case and thus less possibility of "error".)

To see these results for yourself, run `AnalyzeHospitals.py`. This script, like `AnalyzeRatings.py`, loads a pre-generated dictionary of data (`RESTAURANT_INFO.p`), so it does not take very long to run.

## 3.3 Restaurant & Hospital Rankings

## Restaurant Rankings (Based on % of Health Inspections Failed)

| | Restaurant | % of Health Inspections Failed | Average Yelp Rating | License Status |
|---|---|---|---|---|
| 1 | Dunkin Donuts Space no. 86 | 100% (7/7) | (88) | Active |
| 2 | MDM Noodles | 100% (8/8) | (22) | Active |
| 3 | Dudley Square Grille | 100% (10/10) | (22) | Active |
| 4 | Boston Indian Kitchen, Inc. | 100% (8/8) | (1277) | Active |
| 5 | King's Pizza and Grill | 100% (14/14) | (3) | Active |
| 6 | Blue Fuji | 100% (3/3) | (261) | Active |
| 7 | Rita's Catering | 100% (3/3) | (3) | Active |
| 8 | Papa Rino's Pizza | 100% (30/30) | (3) | Inactive |
| 9 | Casa Columbia Restaurant & Bakery | 100% (62/62) | (357) | Active |
| 10 | Soup's On | 100% (5/5) | (6) | Inactive |

Figure 7: *Top 10 restaurants by % health inspections failed.* The first column represents the ranking of the restaurant in terms of % health inspections failed. The second column is the restaurant name. The third column is the % health inspections failed and in parenthesis is the ratio of health inspections failed to total number of health inspections (e.g., 7/7 means a restaurant has failed 7 out of 7 health inspections). The fourth column is the average Yelp user rating and in parenthesis is the number of ratings (which in a way represents foot traffic to the restaurant). Finally, the last column represents the food license status. Either the restaurant has an active license or it doesn't.

# Hospital Rankings (Based on % of Health Inspections Failed)

| | Hospital | % of Health Inspections Failed | | yelp Rating | |
|---|---|---|---|---|---|
| 1 | Boston Medical Center (BU) | 78.9% | (221/280) | ⭐⭐ | (43) |
| 2 | Tufts | 76.5% | (110/134) | ⭐⭐⭐ | (740) |
| 3 | Carney Hospital | 69.0% | (98 / 142) | ⭐⭐ | (17) |
| 4 | Brigham & Women's Faulkner Hospital | 65.8% | (52 / 79) | ⭐⭐ | (22) |
| 5 | Carter Fuller Mental Hospital | 60.9% | (67 / 110) | ⭐⭐⭐⭐ | (3) |
| 6 | St. Elizabeth Hospital | 58.2% | (32 / 55) | ⭐⭐⭐⭐ | (72) |
| 7 | Jewish Memorial Hospital | 56.7% | (38 / 67) | ⭐⭐⭐ | (71) |
| 8 | Mass. Eye & Ear Infirmary | 56.6% | (82 / 145) | ⭐⭐⭐ | (37) |
| 9 | Spaulding Rehabilitation Hospital | 54.2% | (27 / 52) | ⭐⭐ | (6) |
| 10 | Beth Israel Deaconess Hospital | 49.0% | (148 / 299) | ⭐⭐⭐ | (128) |

Figure 8: *Top 10 hospitals by % health inspections failed.* The first column represents the ranking of the restaurant in terms of % health inspections failed. The second column is the restaurant name. The third column is the % health inspections failed and in parenthesis is the ratio of health inspections failed to total number of health inspections (e.g., 7/7 means a restaurant has failed 7 out of 7 health inspections). The fourth column is the average Yelp user rating and in parenthesis is the number of ratings. (Note that some hospitals actually have multiple restaurants inside them. Therefore, the results displayed in this table represent the average across all food establishments within that hospital. For example, Boston Medical Center has 2 food establishments inside it, which means the % health inspections failed for that hospital represents the average % health inspections across 2 food establishments.)

The results seen in Figures 7 and 8 confirm our conclusion earlier that there is no relationship between Yelp user ratings and % health inspections failed.

# 4 Violations vs. Restaurant Location

As mentioned in section 1.1, one of my project goals was to explore the possible relationship between restaurant location and violation data. Therefore, I performed 2 different clusterings on the data, and the results are seen in Figures 9-14 below. In both clustering attempts, I was not able to find a good clustering for the data, which leads me to conclude that there is no relationship between violation data and restaurant location.

To cluster my data, I ran the script `ClusterData.py`. As with `AnalyzeRatings.py` and `AnalyzeHospitals.py`, this script loads the dictionary `RESTAURANT_INFO.p` and analyzes its contents. (Note that this script takes several minutes to generate the plots in this section.)

In my first clustering (seen in Figure 9 below), I attempted to cluster violation level, fail rate, and pass rate. However, from the heat map in Figure 11, we can see that my clustering is not very good. In fact, just from looking at Figure 9, we can see that the data is not clustered by location at all.
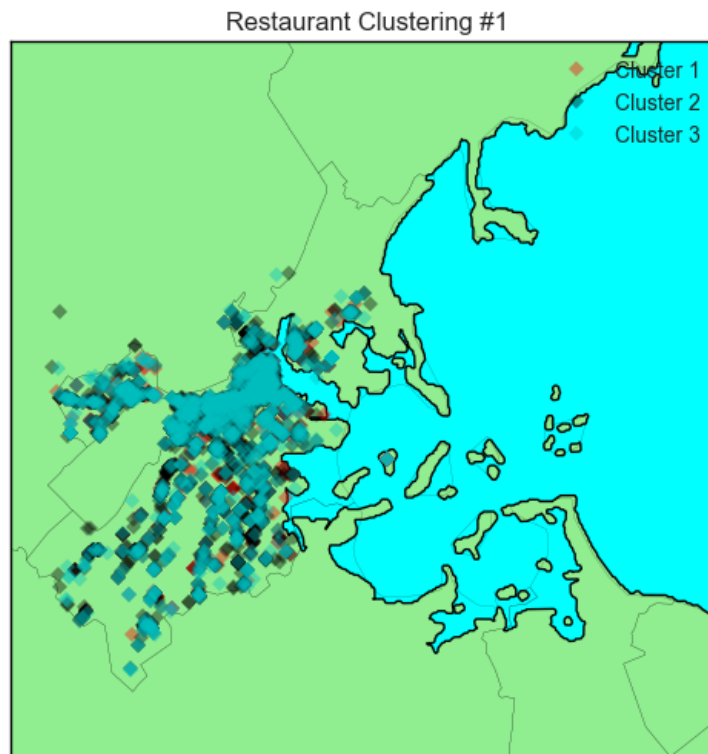


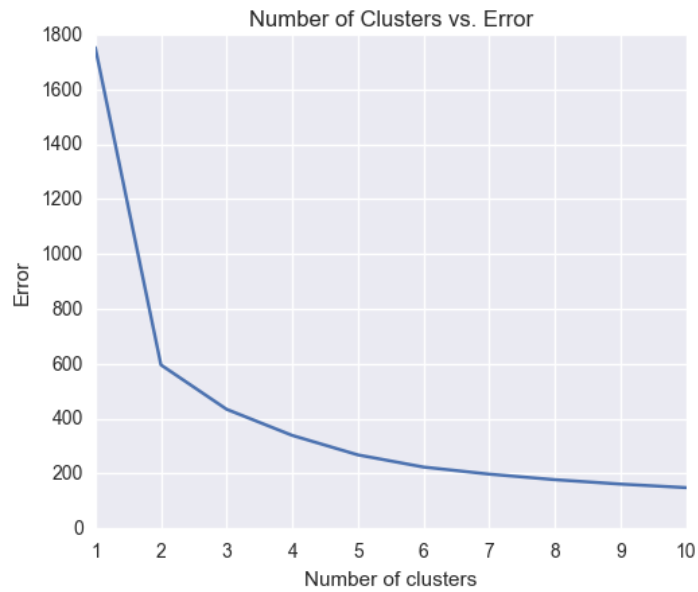Figure 9: *Clustering #1.* Clustering of violation level, fail rate, and pass rate.

Figure 10: *Error vs. number of clusters for clustering #1.* Plot of error vs. number of clusters. The optimal number of clusters is 3 for this particular clustering. (We start experiencing diminishing returns in error after 3 clusters.)
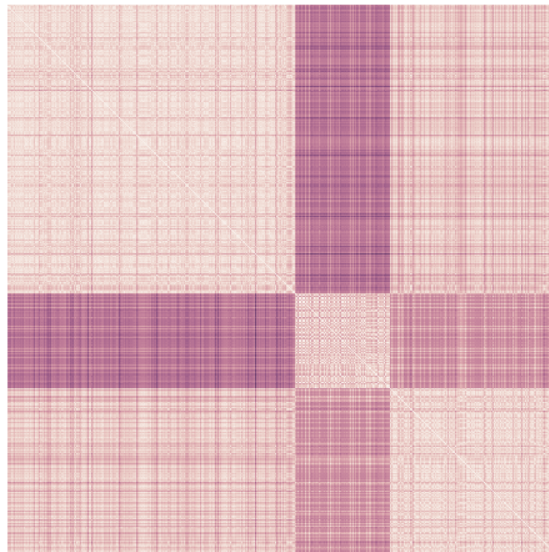


Figure 11: *Clustering results for #1.* Heat map of the clustering results seen in Figure 9.

For my 2nd clustering attempt, I tried to cluster only violation level and Yelp user rating. However, as before, my clustering is not very good, as seen in Figure 14. However, interestingly enough, we do see 2 clear clusters in my data from this heat map, which suggests there is a relationship between Yelp user rating and violation level, though we cannot relate these results back to restaurant location (which is what I was hoping to do).
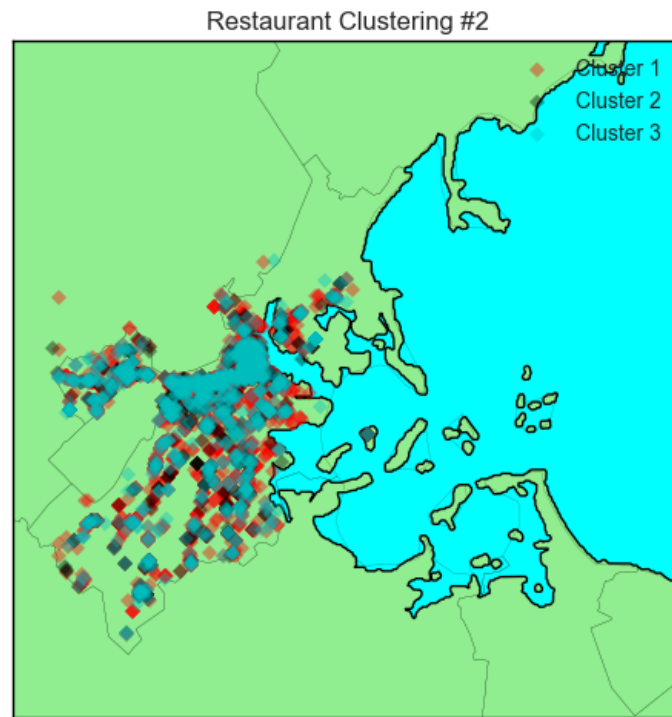


Figure 12: *Clustering #1.* Clustering of violation level and Yelp user rating.
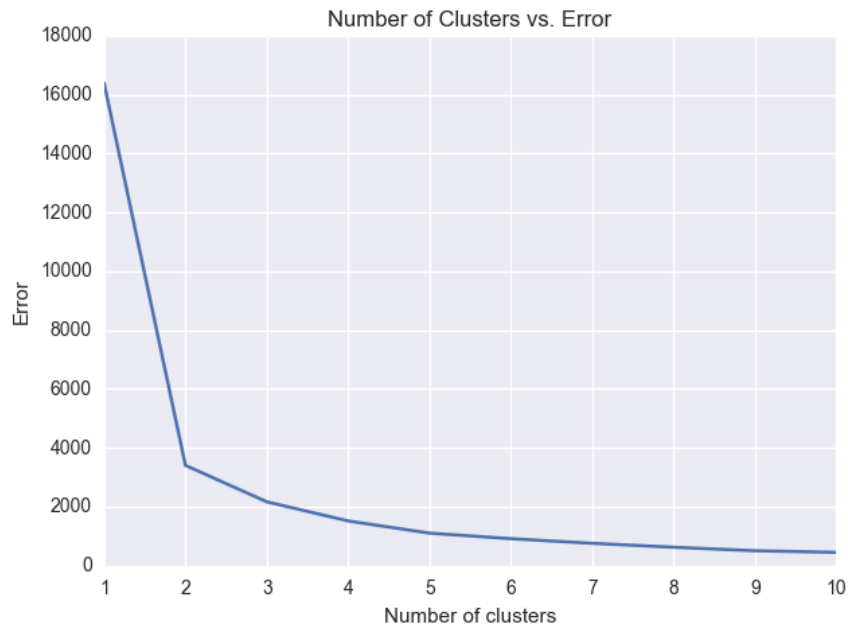
Figure 13: *Error vs. number of clusters for clustering #1.* Plot of error vs. number of clusters. The optimal number of clusters is 3 for this particular clustering. (We start experiencing diminishing returns in error after 3 clusters.)
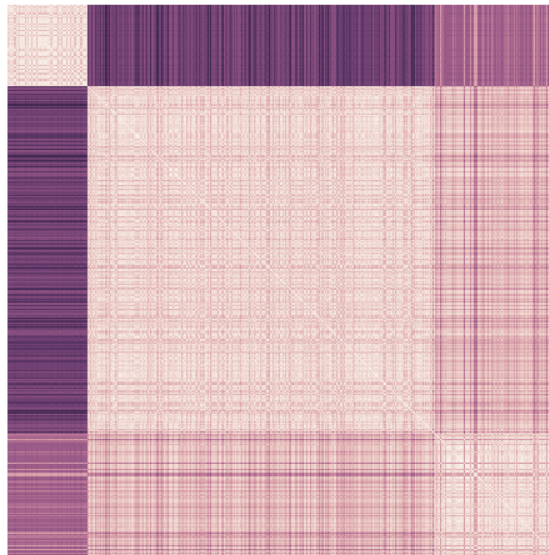


Figure 14: *Clustering results for #1.* Heat map of the clustering results seen in Figure 12.

# 5  Conclusion

Summarization of my results:

1. There is no relationship between average Yelp user rating and % health inspections failed. (See Figure 1.)

2. Even if we attempt to break the dataset down by neighborhood, we see a very weak relationship between average Yelp user rating and % health inspections failed, and that weak relationship only exists for about 5 neighborhoods, which only 2 of are actually located in Boston (`Hyde Park` and `Mattapan Square`). Though, these $R^2$ values may be artificially high due to a low number of points used for linear regression on these neighborhoods. So really, there isn't a relationship between average Yelp user rating and % health inspections failed. (See Figure 5.)

3. Given the fact this dataset contains restaurant information for restaurants *outside* of Boston, it is very possible that some of these restaurants used to operate in Boston, but relocated to outside the city. Therefore, it might be useful to know: (a) if restaurants relocated, and (b) when they relocated. This information may allow us to examine the data more intelligently.

4. I expected there to be a very strong relationship between % health inspections failed and average restaurant violation level, but it turns out that after performing linear regression on this data, $R^2 \approx 0.5$, which suggests that health inspections are somewhat subjective. (See Figure 1.)

5. For hospitals, there is a very weak, but noticeable correlation between Yelp user rating and % health inspections failed. (We saw no correlation between these 2 variables for the entire dataset overall.) However, as before, this $R^2$ value may be artificially high due to a lesser number of points being used, so it's very likely that there isn't a real relationship between average Yelp user rating and % health inspections failed of hospitals either. (See Figure 6.)

6. There really is no relationship between violation data and restaurant location (see Figures 9-14). However, there does appear to be a relationship between violation level and Yelp user rating (see Figure 14).

# References

[1] City of Boston, "City of Boston Data", *Food Establishment Inspections*, 2015. [Online]. Available: https://data.cityofboston.gov/Health/Food-Establishment-Inspections/qndu-wx8w [Accessed: 19-Apr-2015].

[2] Yelp, "Yelp API Documentation V2", 2015. [Online]. Available: https://www.yelp.com/developers/documentation

[3] City of Boston, "Inspectional Services", *Mayor's Food Court - Meaning of Violation Codes*, 2014. [Online]. Available: http://www.cityofboston.gov/isd/health/violationcodes.asp [Accessed 19-Apr-2015]

[4] GeoPy, "GeoPy Documentation", 2015. [Online]. Available: https://geopy.readthedocs.org/en/1.10.0/

[5] OpenStreetMap, "About OpenStreetMap", 2015. [Online]. Available: https://www.openstreetmap.org/about

[6] Wellbeing@School, "Understanding and interpreting box plots", 2015. [Online]. Available: http://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots