DSO 545: Statistical Computing and Data Visualization
Analysis of Health Inspection Data of San Francisco Restaurants

---

Charles Gregart, Niranjan Kasi, Miles Nash, Thomas Wu, Andrew Yau

Executive Summary

In our project proposal, we stated that we wanted to analyze a dataset of health code violations for San Francisco restaurants and determine if there were any interesting patterns. For example, we wanted to display the data against a map of San Francisco to see if the violations were clustered around specific areas. To explore different correlations, we also pulled in income data from the U.S. Census Bureau.

From the geographic maps that we created, we found that most of the violations were clustered around the "Tenderloin" area. These violations also seemed to be independent of external factors: the vermin problem did not seem to correlate with pockets of violations. Additionally, we found a slightly positive correlation between the median income of an area and the average inspection score of its restaurants. However, we need more granular income data to draw better conclusions. Most of the violations were "low risk", related to improper cleaning procedures. Finally, the inspection score distribution was normally distributed around a mean in the 80s, which seems to suggest that most restaurants in San Francisco are safe to eat at. The authors of this paper, however, would be cautious of eating at any of the restaurants in the "Tenderloin".

Business Objectives

The overall business objective of this analysis is to try and understand the different the different inspection outcomes of various establishments across SF and drill down on the factors which have the biggest impact on the inspection scores. By doing so we want to help the various business owners understand improve their overall inspection statistics. The insights gleaned from this data analysis can also be used by the SF Department of Health in allocating resources to the direst locations in the city, as well as the most pressing issues and violations. The analysis can also be shared with customers to make more informed decisions for their meals.

Data Mining Goals

- Understand the relation between location and inspection outcomes by using spatial analysis
- Use various word scraping methods to understand what kind of violations lead to more severe outcomes
- Use additional data sources such as median income to find other extrinsic relationships.

Data Description

Our primary resource is a dataset of thousands of health inspection results of restaurants in the city of San Francisco. The dataset was obtained from HealthData.gov and was published by the City of San Francisco. There are 50,458 observations, and the 17 variable fields for each observation which include:

- Business_id: a unique numeric identifier for each individual physical restaurant (numeric)
- Business_name: the moniker of each restaurant; generally, chain restaurants include a store number in the business name (text)
- Business_address (text)
- Business_city: all observations are within the city of San Francisco (text)
- Business_state: all observations are within the state of California (text)
- Business_postal_code: aka zip code (numeric)
- Business_latitude: geographic coordinate (numeric)
- Business_longitude: geographic coordinate (numeric)
- Business_location: combined string of latitude and longitude (text)
- Business_phone_number: 11-digit number including country code (1), 3-digit area code, and 7-digit phone number (numeric)
- Inspection_id: unique numeric identifier for each unique inspection performed at a single location on a single date; combination of business_id, underscore separator, and date in YYYYMMDD format (numeric)
- Inspection_date: in MM/DD/YYYY H:MM format; all times left as 0:00 (date)
- Inspection_score: numeric score given for each unique inspection on a scale of 0-100; range is 48-100 for this set of observations (numeric)
- Inspection_type: reason for the inspection; examples include: routine (unscheduled), complaint, new ownership, new construction (text)
- Violation_id: unique identifier for the recorded violation and the key unique observation on which the dataset is compiled; combination of business_id, date of inspection in YYYYMMDD format, and a code for the type of violation (e.g. “high risk vermin infestation” = 103114; with underscore separators (text)
- Violation_description: specific violation that the restaurant was cited for examples include “unclean hands or improper use of gloves” and “unapproved or unmaintained equipment or utensils” (text)
- Risk category: a categorization of the violation as low, moderate, or high risk (text)

To reiterate: the 50,458 observations are comprised of 50,458 unique entries for violation_id. Inspections can occur at the same location across multiple dates, but even a single

inspection can have multiple observations in the form of unique violations (identified by violation_id).

This data is continuously updated; we downloaded the dataset on 17 April 2018. The dataset is in CSV format and is entitled "SFRestaurants."

Additional Datasets:

Median Income and Population Data from the US Consensus Bureau website, this data was then linked to the SF restaurants data using the Postal Code as the unique identifier.

Data Cleaning

Because one of our main objectives was to map all the restaurants using ggmap and provide some visualizations as to where certain violations were occurring, it was important that we had as much complete geolocation (latitude and longitude) data as possible. Unfortunately, approximately 20,000 of the 50,000 observations were missing latitude and longitude data. However, virtually all the data was complete in terms of having an address (there were a few dozen observations for mobile (e.g. food truck) restaurants that didn't have associated addresses).

To fill in the missing 40% of observations without latitude and longitude coordinates, we wrote code that utilized a *for* statement to go through all the observations, and for those that were missing latitude and longitude, to use Google's API to search for the address and then fill in the latitude and longitude. To make the search as accurate as possible, we created a new field that combined the address, city (San Francisco), and state (California) into a single string that would be used for the search.

The process was somewhat tricky in that Google only allows 2,500 queries per user per day. However, with 5 of us in the group, and by running the code once each while at home, and then once while on the network at school (to get around Google recognizing the IP address), we were able to accomplish filling in all the missing entries in a single day.

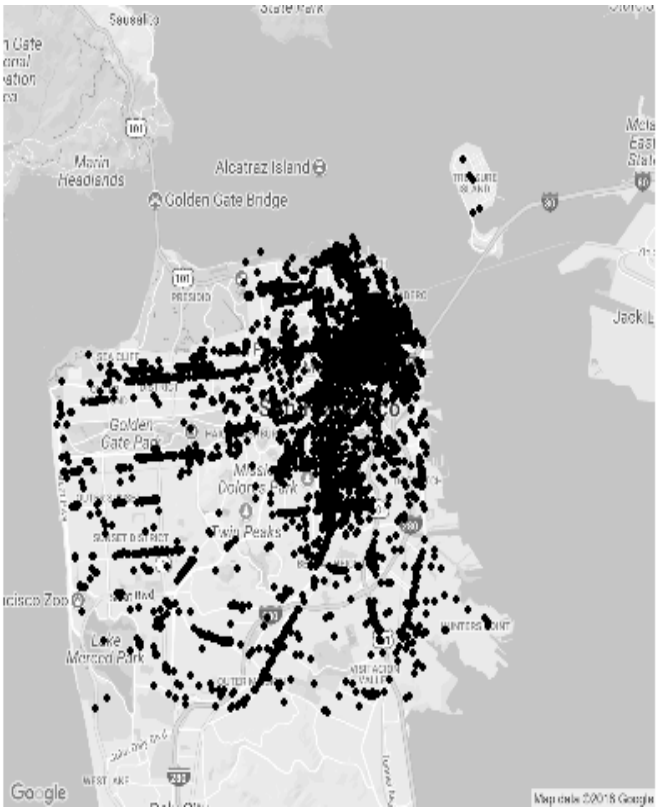
Once this process was complete, there were a few dozen queries that returned latitudes and longitudes in places like New York and Denver. Presumably, the search for the address brought up something similar in another city, despite having included "San Francisco, CA" in the string. So, to remove these incorrect locations, we filtered the dataset for a minimum and maximum latitude and longitude within 1 degree of downtown San Francisco.

For the initial regression analysis, we also used factor method to create numerical levels for categorical variables to have easier case for regression models and easy comparison of predictor variables

In addition to the data cleaning that we did above, we also wanted to tie in outside sources of data to see if there were any correlations with inspection scores. We pulled the U.S. Census Bureau data for median household income by zip code for San Francisco and integrated into our previous dataset. Since the income data is by zip code, it wasn't as granular as we would have liked.

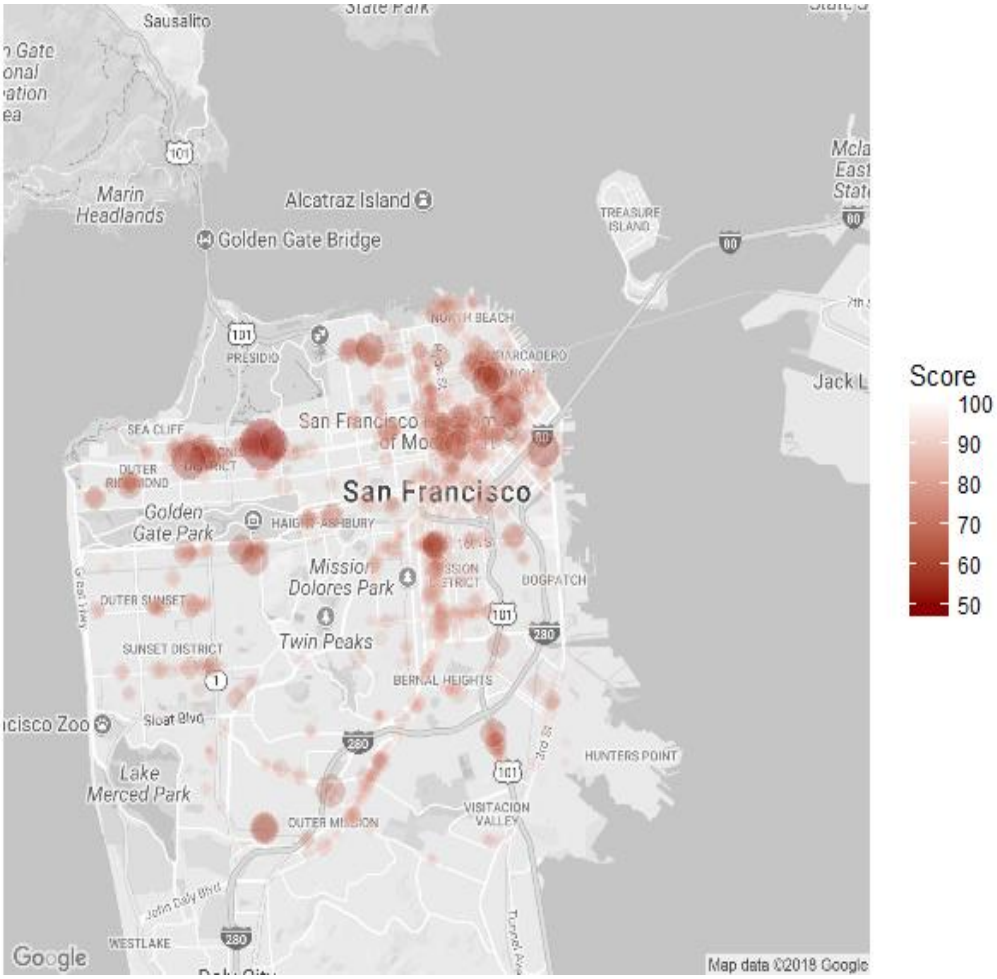
Exploratory Data Analysis

One of our primary objectives with this dataset was to map the locations of the violations to glean some insights as to where the problem areas were in the city. To generate the visualization, we plotted all the inspection observations on a map, using the latitude and longitude fields.



To highlight the problem areas (low scores), we set the fill of the observations on the high end of the inspection score range (100) to transparent, and the low end (48) to dark red. We also used exponential transformations of the inspection score to disproportionately enlarge

the size of the low-score observations, and to disproportionately reduce their transparency (higher alpha value) to darken them on the map and make them stand out.

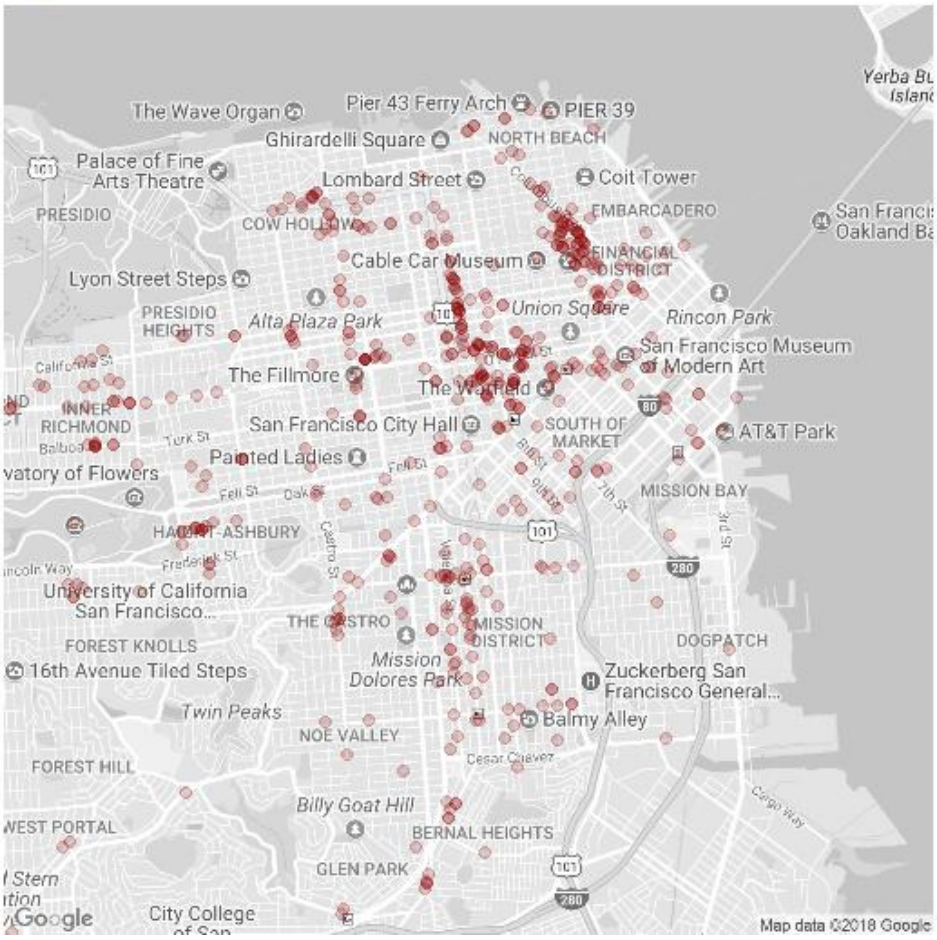


From this map, we can see that the major problem-areas include:

- Chinatown
- Nob Hill
- The Tenderloin
- Richmond District (south of the Presidio)

Additionally, we wanted to visualize where one of the most egregious violations - high risk vermin infestations - is occurring within the city. We created a map by filtering the dataset for violations of this type.

High Risk Vermin Infestations



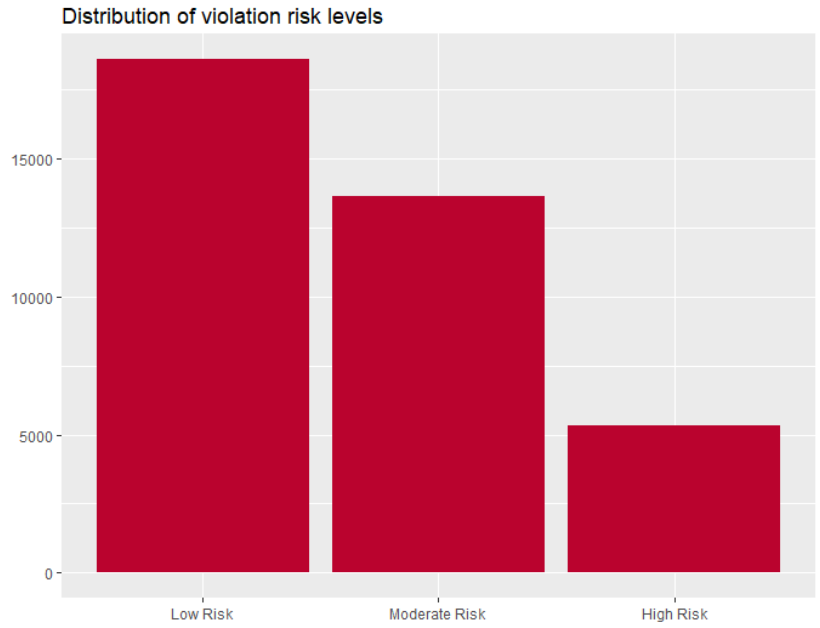
From this map, we can see clear concentrations of vermin infestations in two areas:

- Columbus Avenue northwest of the financial district, extending into the blocks south into Chinatown
- From Nob Hill, north up Polk Street

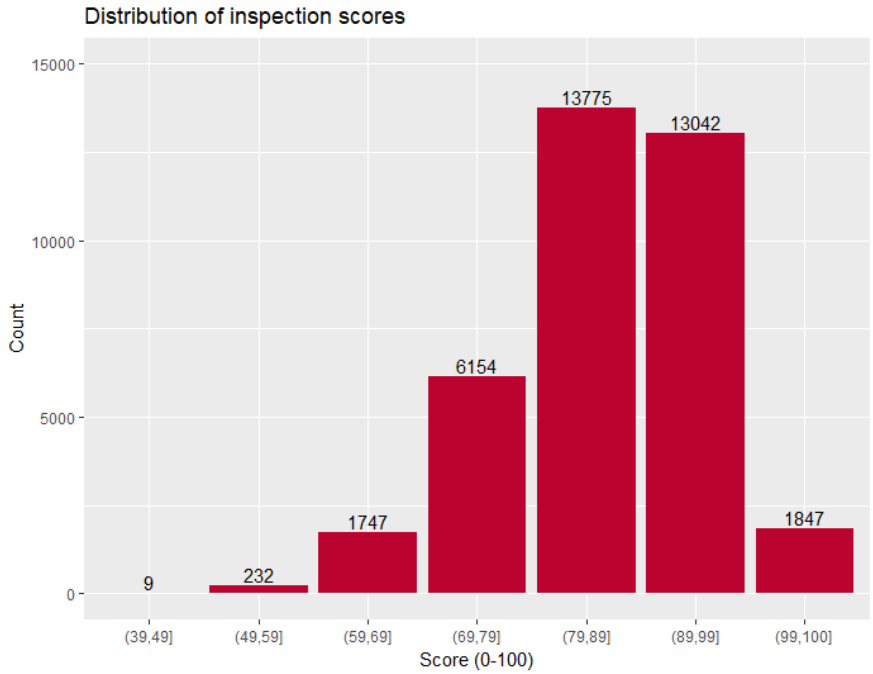
With this information, we would recommend focusing extermination and prevention measures on these two streets. There may be a connection to something like older sewer or subway tunnels that allow rats and cockroaches to breed and get into restaurants. Both areas appear to be central to adjacent locations that spread out from their cores. By addressing these cores, it will likely improve surrounding areas.

While these are certainly the areas to focus on in terms of improving performance and allocating resources, it is important to note that most of the observations indicate that the city is in an

overall good status of health code adherence. We prepared a bar chart that shows that most of violations fall into the low-risk category:



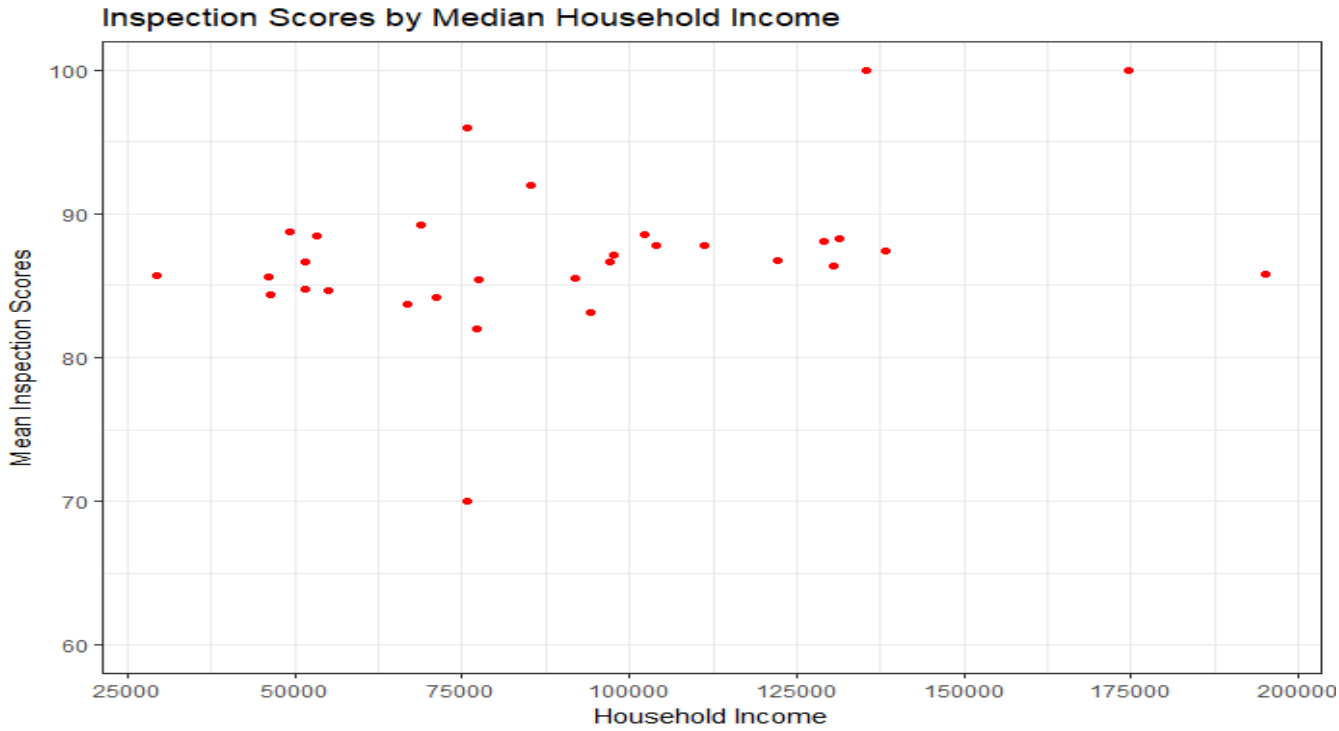
We also prepared a histogram of the distributions of all inspection scores:



These results indicate that most of inspection results fall into high-score categories (>80). The data follows a left-skewed normal distribution. It is important to note that the final category is

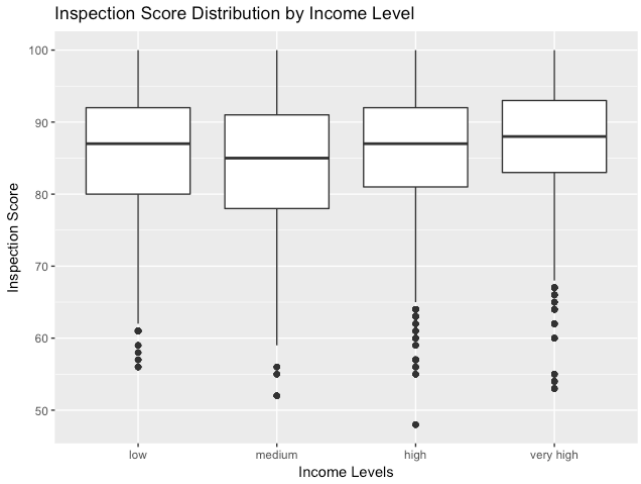
only for perfect scores (100), and thus the width of the final bin should be 1/10 of what it is (range of only 1 as opposed to 10 for the other categories); this would further demonstrate the skew favoring more observations on the higher end of the inspection score spectrum.

From the onset, we had the hypothesis that the inspection score might be positively correlated to income levels. Using the U.S. Census Bureau data, we were able to make some general comparisons. To accomplish this, we grouped the data by zip code and summarized the mean of the inspection scores and income levels. From the plot below, although there is no clear conclusion, there appears to be a slight positive correlation with the mean inspection score and income levels.



To further examine the relationship between income level and inspection score, we generated a variety of charts to illustrate the differences between each income level, which we defined somewhat arbitrarily as low ($50-70k), medium ($70-90k), high ($90-120k), and very high ($120-200k). First, we generated a faceted view of inspection score distribution by income level.

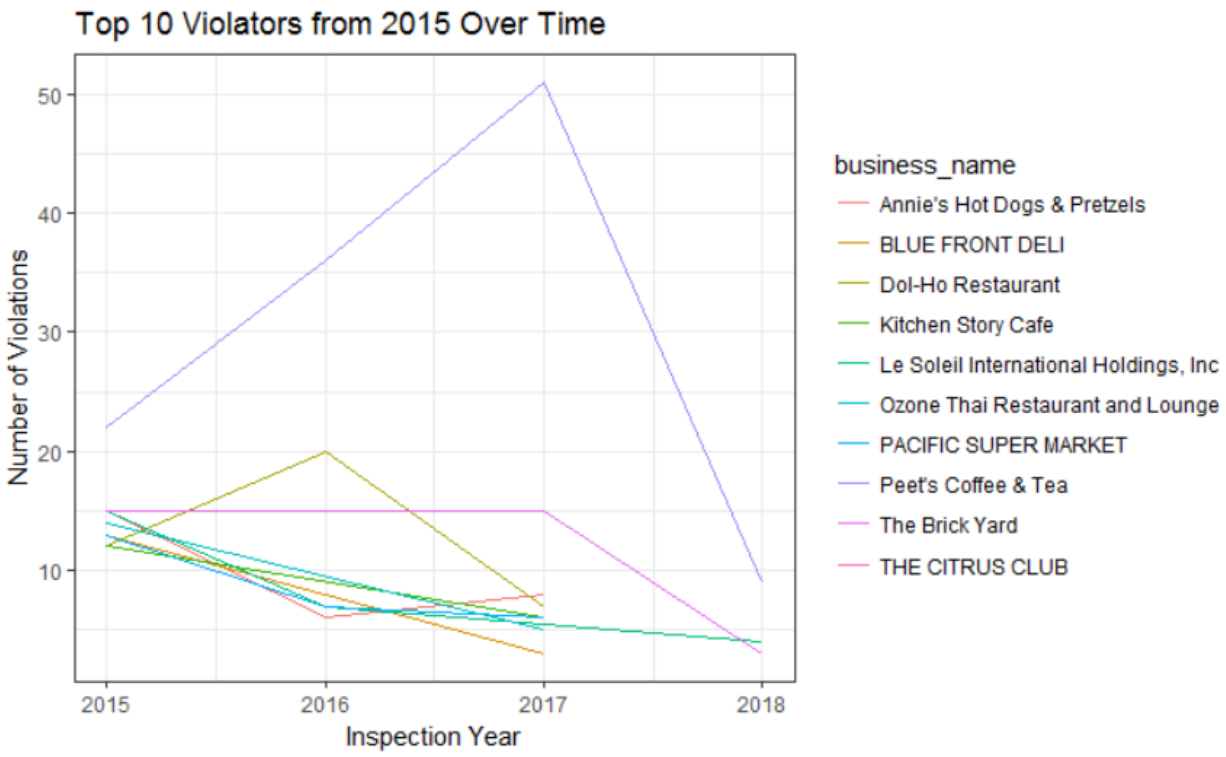
We originally developed this view as a single chart, but even after adjusting opacity levels it was difficult to parse between the income levels. As was quantifiable confirmed through our regression analysis, the trends between income levels are somewhat weak, but you can see that the higher income levels tend to peak higher, and with greater counts at the top inspection score levels. To further reinforce our regression findings, we also generated a basic boxplot graph, with a trend of increasing scores with income levels, but with a notable exception: the lowest income level broke with the trend with unexpectedly high scores.



Next, we developed a stacked bar chart to illustrate the share of inspection scores by income levels:

While this graph is useful in displaying the overall distribution of scores, and though you can see a trend of increasing share of inspections by higher income levels, we were dissatisfied with the ability to highlight share. Therefore, we modified data (using group_by, summarise, and mutate) to develop a view that focuses on % share of inspection scores. This plot, below, clearly illustrates the increasing share of inspection scores by higher income level area restaurants, especially within the typical range of 70 to 100.

Another trend we were curious about was how the top violators did over time. Did they improve or regress? To determine this, we first found the restaurants with the most violations in 2015. Then we plotted the number of violations they had from 2015 to 2018. From the plot below, we can see that most of the top violators from 2015 improved over time. The major outlier is Peet's Coffee & Tea which doubled in violations in 2017.



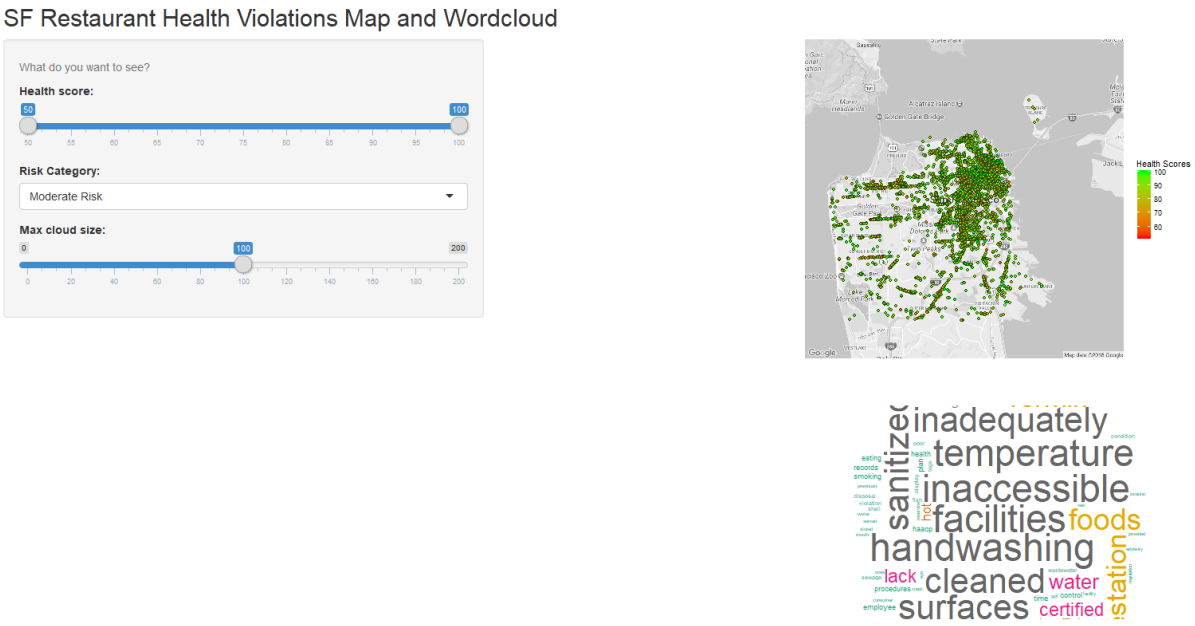
When creating the Word Cloud, we wanted to examine the key words that correspond to certain types of risk categories (High, Medium, and Low). After removing the words "food", "risk", "inadequate", "improper", "contact", "holding", "moderate", and "safety". We found the following frequency of violation descriptive words:

1. unclean 6842
2. surfaces 4647
3. temperature 4022
4. equipment 3815
5. ceilings 3465

6. degraded 3465
7. floors 3465
8. walls 3465
9. storage 3443
10. infestation 3385
11. vermin 3385
12. utensils 3316
13. facilities 3155
14. unmaintained 2580
15. handwashing 2468

We found that in terms of violations, outside of general cleanliness (unclean), that physical attributes of the restaurant dominate the violations (surfaces, temperature, equipment, ceilings, floors, walls, storage). Infestations was ranked #10 for violations which is with vermin. This would tell restaurants that focusing on the cleanliness of their physical space is paramount to passing inspections. These issues speak to fixes that can be accomplished through proper upkeep by restaurant staff and should be incorporated into daily maintenance routines. Examining the next five violations we noticed that infestation and vermin are usually paired as a violation type. These violations can be tied to the basic lack of cleanliness in the facilities which would attract vermin. Utensils, facilities, unmaintained, and handwashing also round out the top 15 violation types which further speaks to the lack of restaurant upkeep.

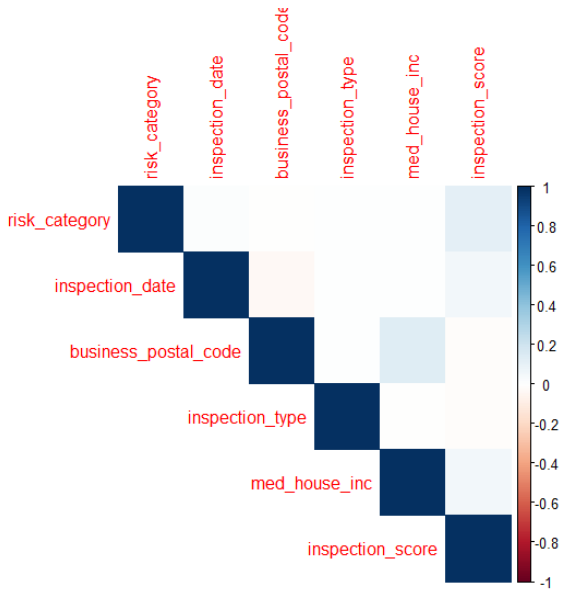
Finally, we developed a dashboard to allow the user to visualize changes through a dynamically updating map and wordcloud (+ a chart quantifying the content in the wordcloud). The dashboard allows the user to simultaneously control the range of health scores and risk categories he/she wants to see, as well as to configure the density/volume of the wordcloud. Notably, we adjusted the visual treatment of individual locations on the map (relative to our static analysis from earlier) to highlight the differences within the currently selected range, and the chart quantifying the content of the wordcloud made it easier to parse information as the user adjusts the conditions.



As a downside that we were unable to overcome, the dashboard has severe performance issues. While the map by itself is calculated and drawn quickly, the wordcloud is extremely resource intensive and often takes upwards of 20-30 seconds to redraw

Predictive Modelling

To understand some basic relationships between the data variables we ran a correlation plot and a linear regression. We filtered down to only the categorical variables which can be used to possibly predict inspection scores based on risk categories, location, median income and inspection type.



Although there is not a very strong relationship we can see that some of the variables do have linear relationships. There is small relationship between inspection scores and median income as well as the risk categories and inspection score but not as high as one would expect it to be.

We also ran a basic regression model to see if we could build a predictive system. For the analysis we also used factor method create numerical levels for categorical variables to have easier case for easy comparison of predictor variables

- Risk Categories were converted into 1,2,3.
- Each of the Postal Codes were assigned a number between 1-36
- The Inspection types were assigned a number between 3-15

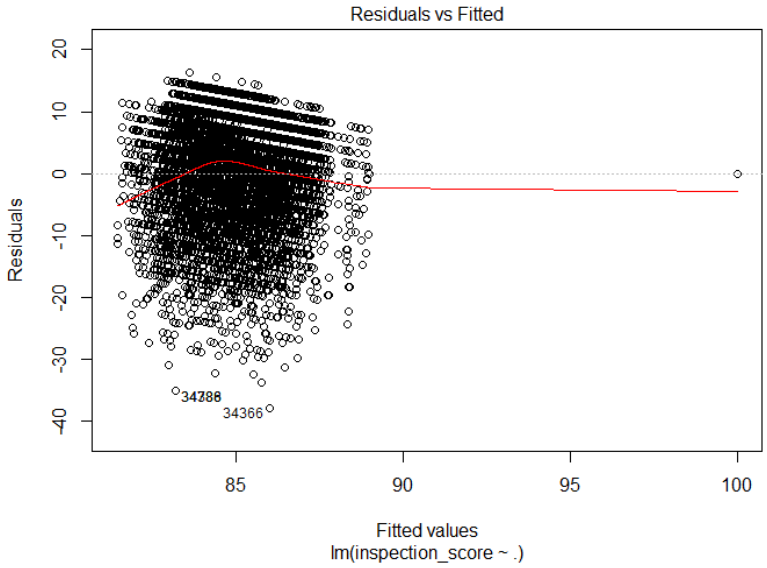
We used a training set with 70% of the data and the remaining to test. Although the individual P Values indicated significant predictors, the overall regression results were disappointing with high RMSE when used to predict inspection scores and very low overall Adj R^2. Maybe we need to use more sophisticated unsupervised models using more variables to be able to achieve a more stable predictive model.

```
Call:
lm(formula = inspection_score ~ ., data = test.set)

Residuals:
    Min      1Q  Median      3Q     Max
-36.609  -4.979   1.525   6.151  16.341

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.506e+01  1.172e+01   2.139   0.0325 *
risk_category         1.362e+00  1.211e-01  11.243  < 2e-16 ***
inspection_date       1.284e-03  2.741e-04   4.685 2.84e-06 ***
business_postal_code -2.518e-02  1.006e-02  -2.503   0.0123 *
med_house_inc         1.399e-05  2.362e-06   5.923 3.26e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.374 on 10351 degrees of freedom
Multiple R-squared:  0.01801,   Adjusted R-squared:  0.01763
F-statistic: 47.46 on 4 and 10351 DF,  p-value: < 2.2e-16
```



Conclusions from Data

Examining the location of the worst violations we see that they are clustered in the portion of San Francisco known as the "Tenderloin". In San Francisco, poor nutrition contributes to 6 out of the top 10 causes of death (heart failure, stroke, hypertension, colon cancer etc.). With a concentration of health violations in this one area city planners can do much to assist in the health standards of restaurants. With 48% of residents spending their dollars in food prepared away from home, the lack of cleanliness can pose a threat to community health. We recommend that San Francisco target these areas and focus on Operational Discipline and Facility Design.

Since our results found that most of the issues arose from basic cleanliness of the facilities rather than external factors (vermin) the city can examine employee onboarding techniques. This examination would determine whether the restaurants that perform low in these areas have formalized programs to teach sanitization to their employees. City ordinances that regulate the standards for how a restaurant sanitizes the problem areas that we identified can fix the top ten violations and impose Operational Discipline on the restaurants. The last issue in the top #15, hand-washing also speaks to Operational Discipline and the lack of employee training or enforcing standards. Examining Facility Design can show why certain portions of the restaurant (ceiling, walls, floors, surfaces etc.) are flagged during inspections. Also, if the buildings themselves are older than this can contribute to the temperature issues mentioned in the violations. If the restaurant is poorly designed or constructed, then it might be difficult for employees to clean all the necessary locations. Facility design can also lead to vermin infestations by providing vermin with places to hide or breed. This also leads to the top 15 issues of "facilities" and "unmaintained" as the building themselves might not be up to code.

Future Work

Over the course of this project, we came up with several ideas that we would have liked to implement seemed very beyond the scope of the project and our ability to execute at present to implement. One example is trying to integrate Yelp review data for each of these restaurants. We suspect that there would be a correlation between the Yelp reviews and the number of violations or the inspection score. However, when we attempted to use the Yelp dataset, we found that it was not in an easily accessible format. Upon consulting subject matter experts, we learned that Javascript would likely be necessary to tie into the APIs and reliably map Yelp data to our inspection data, given the messiness of the data (due to variation in name and address formats in addition to potentially stale inspection data listings). If we were able to get the data into a format we could clean with R, we would be able to tie this in with our existing data.