Health Inspection Data Report

Analysis and Discussion

Team 12

STS 98
Assignment 5
June 2, 2016

The Most Dangerous Eateries in San Francisco, Alameda, and Yolo Counties

By Nancy Au, Shelby Innerst, and Jake Newman

Introduction

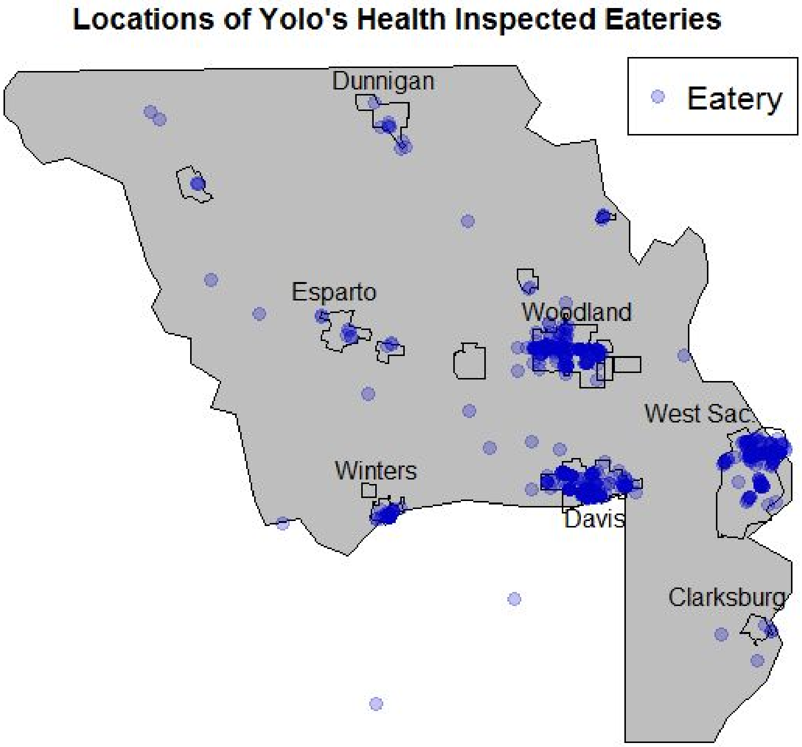
This report explores health inspection data from Alameda, San Francisco, and Yolo counties in search of the most dangerous eateries. For the purposes of this report, “dangerous” is defined as unsanitary. Yelp’s Local Inspector Value-Entry Specification (LIVES) format establishes a common standard for health inspection data and splits county records across three CSV files (labeled as `businesses.csv`, `inspections.csv`, and `violations.csv`). A major step in our analytical process was merging all three datasets by the `business_id` and `date` variables for each of the three counties. The resulting merged dataset for each county was then cleaned and reorganized in different ways depending on the initial format and status of each county’s data. Next, each county’s health inspection data was analyzed with an emphasis on geographic distribution and sanitary variance among cuisine types. However, the manner in which this information was measured differed by county and is discussed in further detail throughout the report. Another notable difference between each county’s dataset exists in the time frame included, varying from a range of 3.5 years in SF to a range of 9+ years in Yolo.

Yolo County

Description of Dataset

The original dataset for Yolo County health inspections had 23,548 rows and 13 variables; however, the data was organized by inspection with repeating business IDs. For example, the Ding How Restaurant (with business ID FA0001165) was repeated 198 times. Variables included “business_id,” “date”, “name,” “address,” “city,” “state,” geographic coordinates, “phone_number,” “score,” “result,” and “type.” Unfortunately, all of the information within the “score” and “result” variables were “NA.” Thus, the only information that implied positive or negative scores was the “type” variable which had the following kinds of health inspections: “follow up”, “complaint”, and “routine.” Also, the range of dates for each inspection in Yolo County was a span of about 9.3 years from January 2, 2007 to May 13, 2016.

The map to the right shows the geographic distribution of Yolo County's eateries (based on each eatery's latitude and longitude) that were inspected at least once during the 9.3 year range. 32.7% (259) are in Davis, 28.7% (227) are in Woodland, 27.8% are in West Sacramento, and 5.8% are in Winters. The remaining 5% are not in Yolo County's cities. This distribution of inspected eateries mirrors population differences between cities and unincorporated areas.



Data Cleaning & Reorganization

Using the `summary()`, `head()`, `table()`, and `sort()` functions, I determined that the data needed a little cleaning and extensive reorganizing. To clean the data, I made R recognize that the "date" variable was a date--not an integer. I also removed all rows listing "Sacramento" within the city variable, because Sacramento isn't in Yolo County.

To organize the data, I did the following: 1. created three subsets of the original data organized by inspection "types" (follow up, complaint, routine) keeping only the business ID and the type columns, 2. created a "total_counts" variable which summed the three types per business ID, and 3. merged these four new subsets together by business ID. Next I made a percentage by adding the newly created "follow up" and "complaint" columns and then dividing this sum by the total number of inspections per restaurant. This new percent column, titled "per_negative", creates kind of score for each restaurant where numbers closer to 1 (or 100%) are the least sanitary and numbers closer to 0 are the most sanitary based on health inspections. For the last reorganization step, I used the business IDs as the key to merge additional information which I subsetted from the original dataset--such as "name", "latitude", "longitude", and "city." In this final step, I made sure to select only unique rows using the `unique()` function (to remove repeating businesses). This resulted in 792 unique businesses and 9 variables.

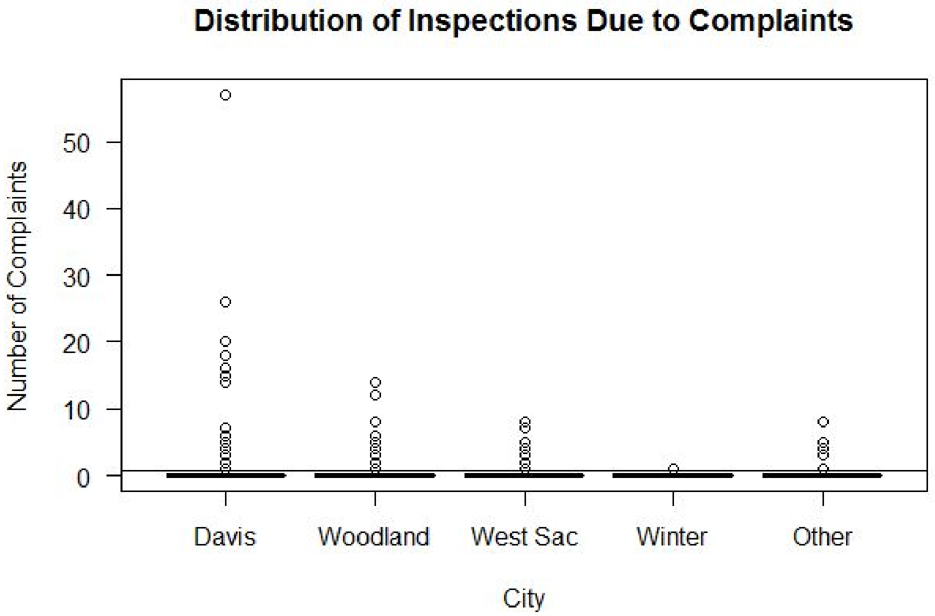
Analysis

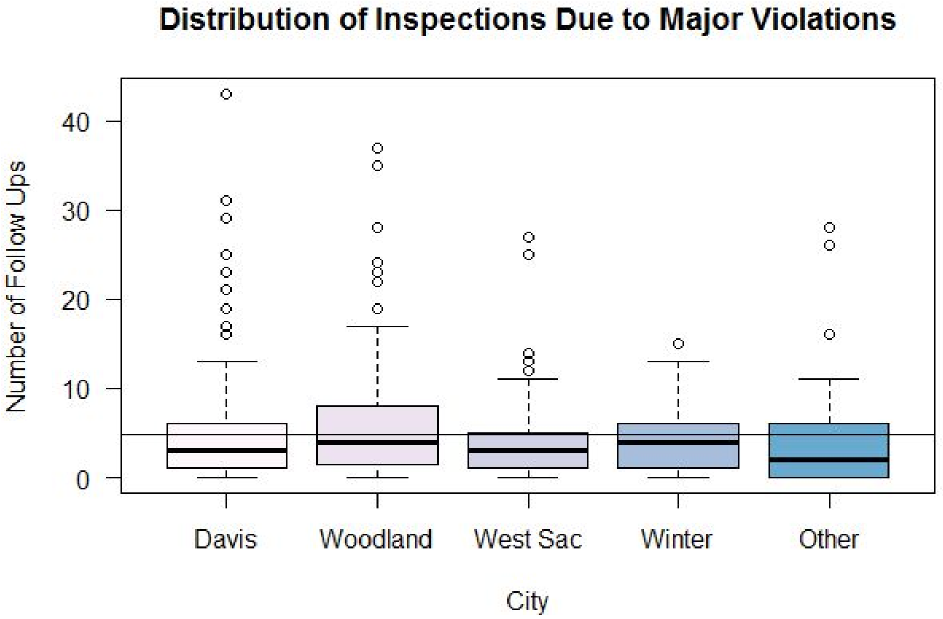
Focus and Underlying Assumptions

My analysis of health inspections in Yolo County primarily examines food serving establishments or “eateries” in the county’s 4 cities: Davis, Woodland, West Sacramento, and Winters. The “Other” category includes all eateries outside of these cities. According to the Health and Human Services page on the Yolo County website, eateries in Yolo that have “*critical violations usually require mandatory re-inspection by Yolo County Environmental Health to confirm that the violation has been addressed*” (yolocounty.org). Thus, a business having follow up inspections imply that it previously had a major violation. In contrast, health inspectors ensure that minor violations are corrected through routine inspections--which are inconveniently combined with normal routine inspections, making minor violations impossible to analyze (yolocounty.org). These health inspection practices informed my creation of the “per_negative” variable which looks at the percent of an eatery’s total inspections that were based on negative issues--namely, major violations or customer complaints.

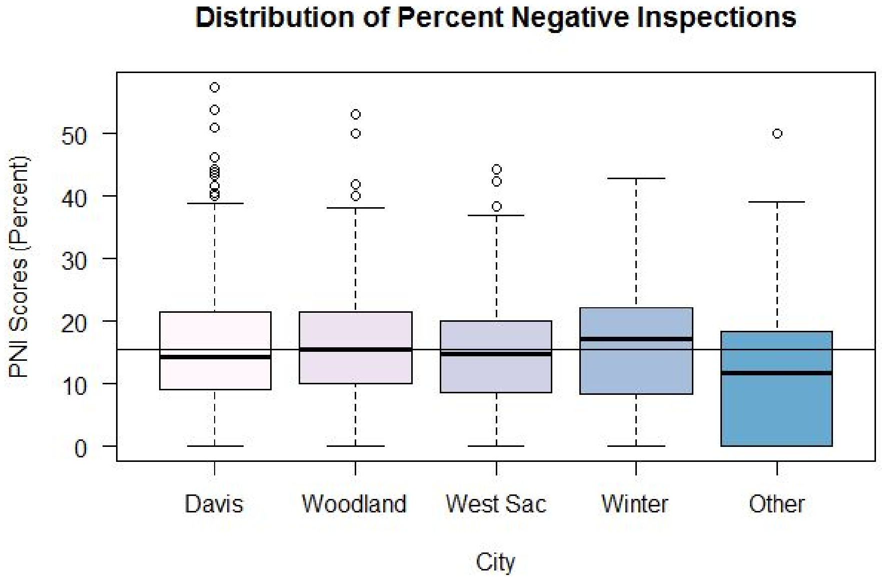
Methods

Before describing how I analyzed the data, I first wanted to explain why I did not analyze the data solely based on complaints. The two boxplots below show inspection distributions of eateries by city in Yolo County. The first boxplot shows distributions of complaints, while the second boxplot shows distributions of follow up inspections which were due to major violations. The number of complaints are clearly lower than the number of follow up inspections with mean of 0.7 and 4.8 respectively as indicated by the lines. Due to a lack of information necessary to conduct a meaningful analysis, I chose to not examine complaints in too much depth. Notably, the highest outlier on the complaints barplot is the Ding How Restaurant in Davis with a record breaking 57 complaints (which is 6.4 times Yolo County’s standard deviation)!





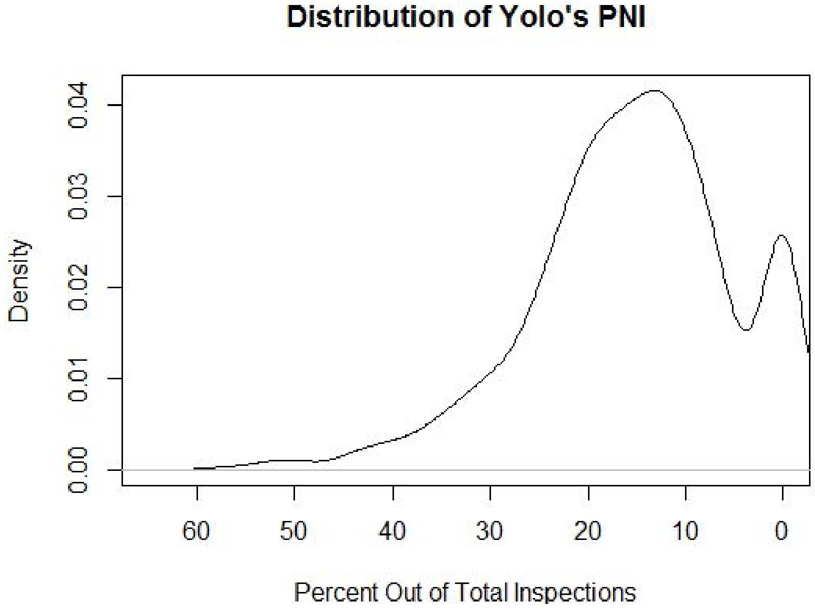
I analyzed the data by looking at the number of follow up inspections per eatery and also by looking at the percent of negative-based inspections (PNI) per eatery. Each method has strengths and weaknesses. Only looking at the total number of follow up inspections (which imply that the eatery had major violations) can be misleading since certain eateries were inspected more overall than others. For example, Wingstop in Woodland was inspected only 4 times while Ding How Restaurant in Davis was inspected 198 times. Of those 4 Wingstop inspections, only one was due to a major violation (25%), yet of those 198 inspections at Ding How Restaurant 29 were due to major violations (14.6%). Thus, this focus on follow ups fails to take into account the proportions of each type of inspection relative to the total number of inspections per eatery. I also looked at the percent of negative-based inspections (PNIs) per eatery which divides the number of inspections due to follow ups/major violations and complaints by the total number of inspections. Compare the boxplot below of PNIs with the boxplot above of "Inspections Due to Major Violations." These plots show how the different types of analysis can vary. For example, West Sacramento appears to be well below the county mean when looking at the number of violations per eatery (thus, more safe for consumers) whereas the city is close to the county mean when looking at the PNI scores (thus, less safe for consumers).



Analysis Across Yolo County and Between Cities

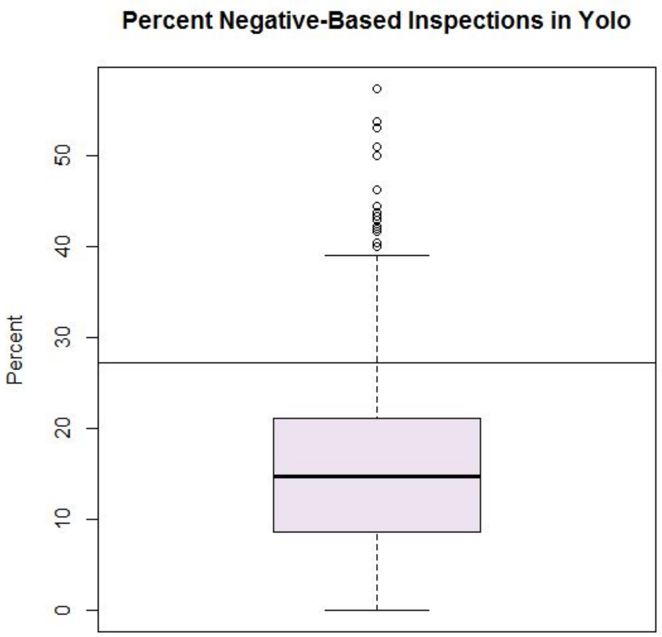
Yolo's density plot of PNI scores mirror the food health score frequencies of Alameda County and San Francisco County despite the measures of health safety varying dramatically between counties. This plot shows that food establishments in the county are generally safe. A moderately high amount of eateries across Yolo County have a score of 0, meaning that all of the inspections at these eateries were simply routine visits.

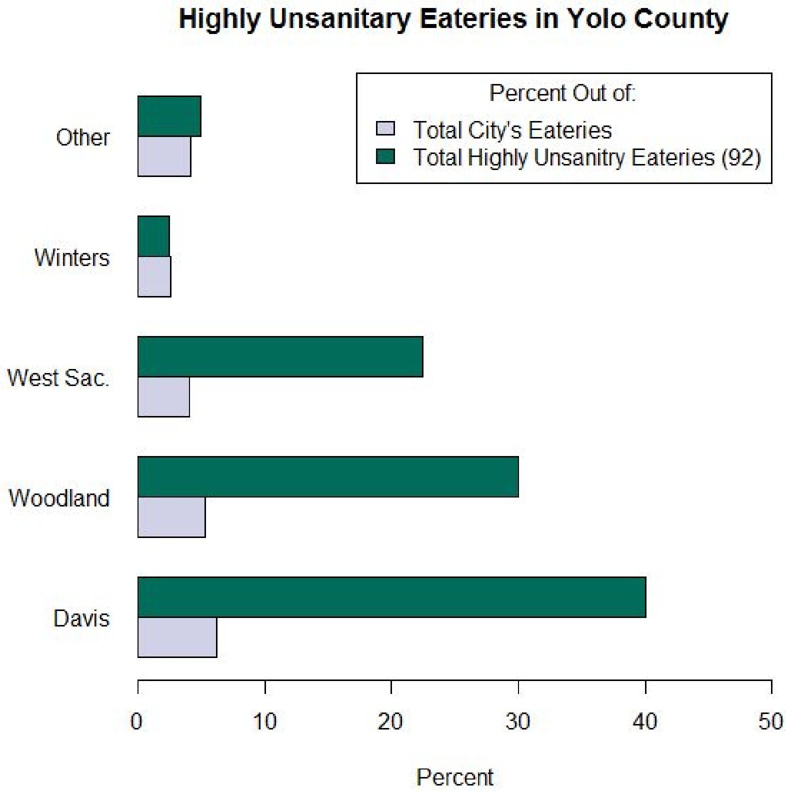
The most frequent score of eateries are in the 10-20% range with the tail decreasing steadily--and comfortingly--until



reaching the highest score of 57.4%.

The boxplot to the right examines the distributions of the PNIs for Yolo, where 0% is the highest "perfect" score and 100% is the worst possible score per eatery. The median score is 14.8% while the mean score is 15.5%. This higher mean reflects the high outliers. For my analysis, I consider scores above 1.5 the IQR (which is indicated by the line on the boxplot) to be "highly unsanitary" eateries which are ultimately more dangerous for customers.

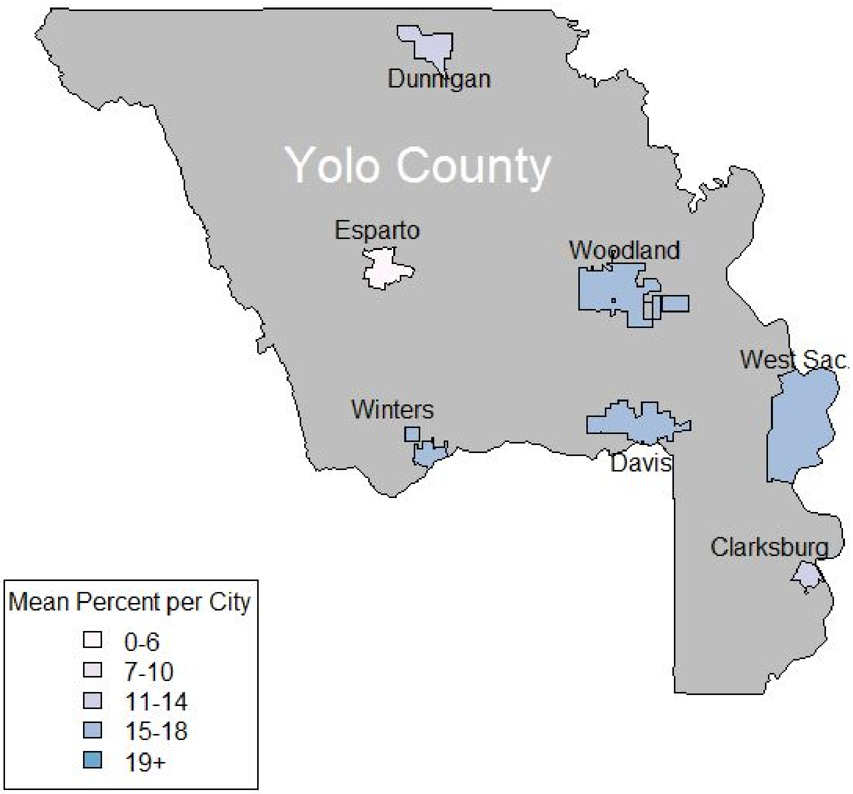




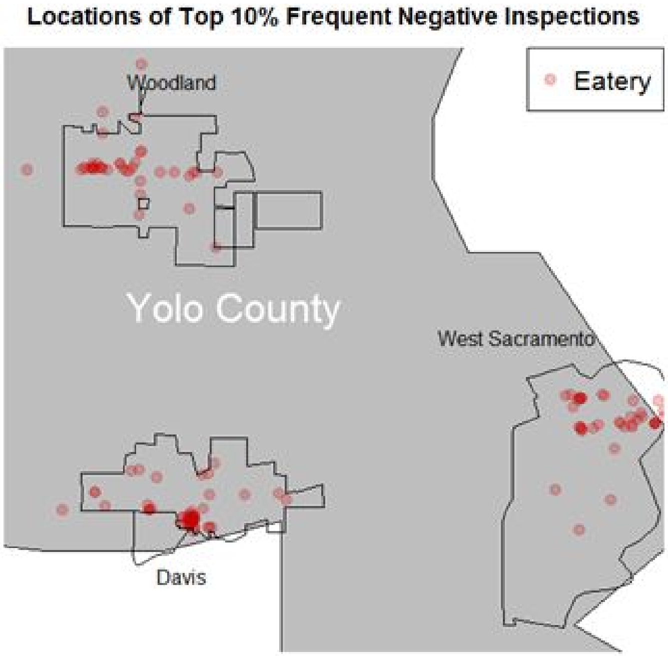
The subgroup of highly unsanitary eateries had a median of 9 and mean of 11 major violations (in contrast with the overall dataset's median of 3 and mean 4.8). These highly unsanitary eateries are organized by city and shown on the bar graph to the left. The grey bars indicate the percent of highly unsanitary eateries compared to the total in each city. Davis has the highest percent of highly unsanitary eateries compared to the other cities and the unincorporated areas ("Other").

The green bars show the percent of highly unsanitary eateries compared to the total number of highly unsanitary eateries (92). Again, Davis has the highest percentage. Both bars imply that Davis's college town food establishments are more dangerous to customers than non-college towns. Notably, population size and/or number of eateries per city could be confounding variables.

Percent Means of Negative Inspections in Yolo County



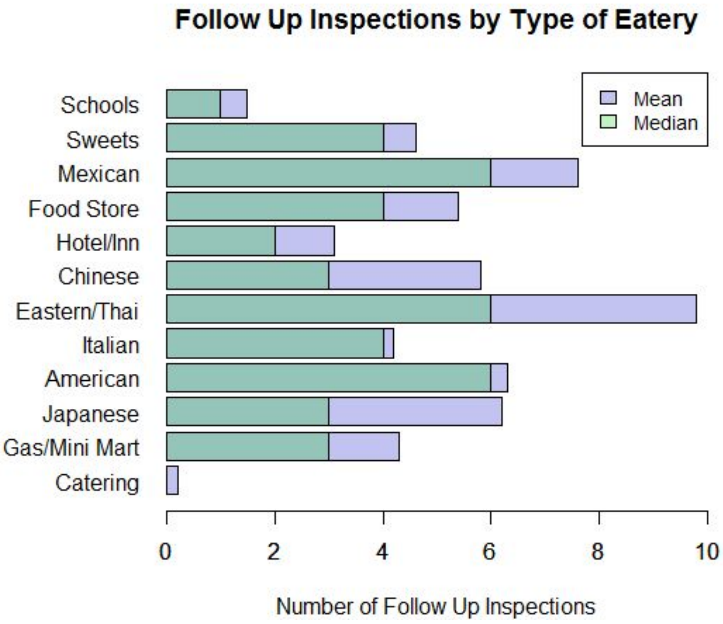
Although Davis might have the highest percentage of highly unsanitary eateries, the map above shows that its overall average PNI score is similar to the other major cities in Yolo. All of the cities had higher average PNI scores compared to unincorporated areas such as Dunnigan, Clarksburg, and Esparto. Again, for reference, higher PNI scores mean that the eateries in these cities are less sanitary because more of their health inspections were due to major violations and/or customer complaints. From this it could be implied that higher and more dense the population is, the more likely the food establishments will be dangerous and unsanitary for consumers. As a disclaimer, a more thorough statistical analysis is necessary to determine if the differences in mean PNI scores are significant.



The map to the left shows the locations of the eateries in Davis, Woodland, and West Sacramento that had the highest PNI scores, which refers to the worst and most unsanitary eateries. Each eatery is indicated by one semi-transparent red point. This map shows that in Woodland and Davis these eateries have clusters in the downtown areas. In contrast, West Sacramento's worst scoring eateries are more spread out.

Analysis of Eatery Type Differences

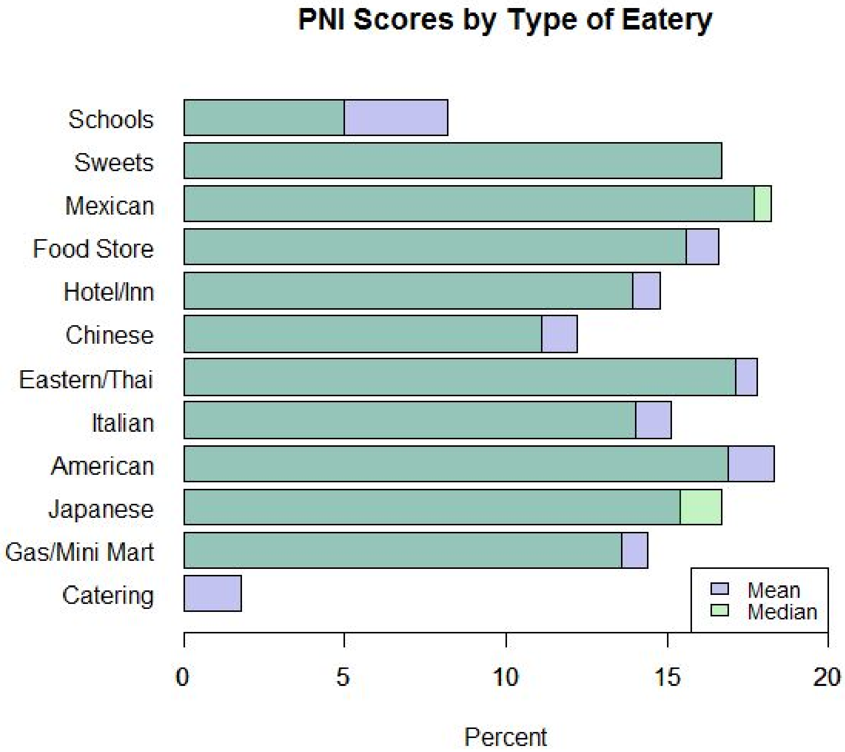
In addition to comparing cities and unincorporated areas, I looked at differences between eatery types based on number of major violations/follow up inspections and based on FNI scores. The possible types of food serving establishments include: school cafeterias, dessert places ("sweets"), various ethnic foods, grocery stores ("food store"), and catering businesses. Each of the bar charts below look at the mean and median for each type of eatery; the mean is indicated by the blue bar and the median is indicated by the green bar.

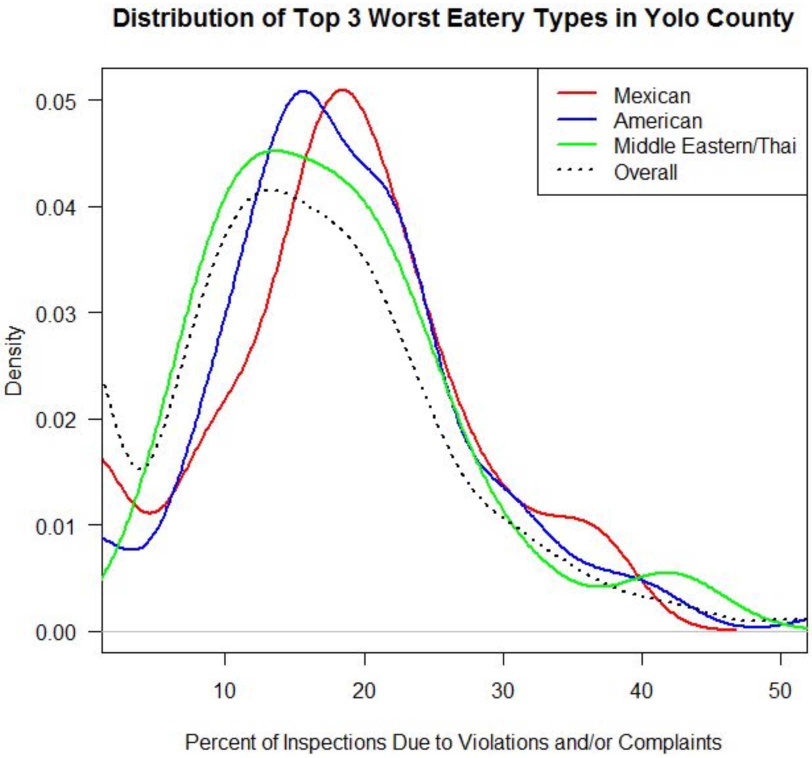


Examining types of eateries by number of follow up inspections, it appears that Chinese, Japanese, and Middle Eastern/Thai restaurants have very high outliers which cause their means to be higher than their medians. Meanwhile, Mexican, Middle Eastern/Thai, and American restaurants appear to have the highest median number of follow up inspections (aka, major health inspection violations). In short, these kinds of restaurants are the most dangerous in

Yolo County.

The bar graph to the right which looks at the same types of restaurants but examines their PNI scores. Similar to the previous bar graph, Mexican, Middle Eastern/Thai, and American restaurants scored the worst with PNI scores of 17% or higher. Interestingly, this perspective of analysis reveals that dessert places also scored relatively poorly compared to the other types of eateries.



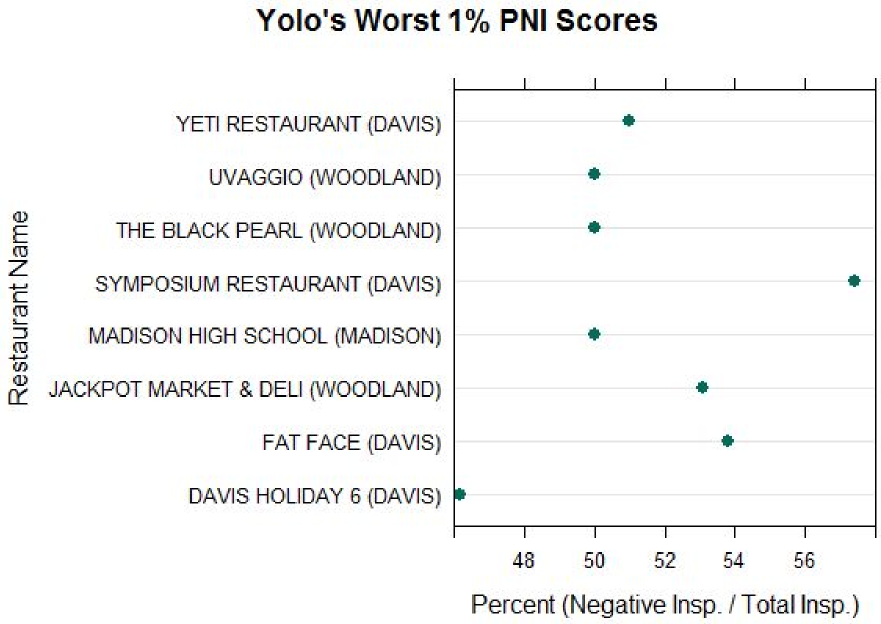


The density plot to the left examines the three worst eatery types in Yolo County based on PNI scores: Mexican, American, and Middle Eastern/Thai restaurants. These distribution lines show that Middle Eastern/Thai restaurants are more evenly distributed with a lower peak between the 10% and 20% PNI score range. Yet, Middle Eastern/Thai restaurants have the highest number of eateries scoring more than 40%. Mexican restaurants have relatively narrow peaks (indicating the highest number of eateries) scoring close to 20% while American restaurants have relatively narrow peaks at the 15% mark. In short, on average, a

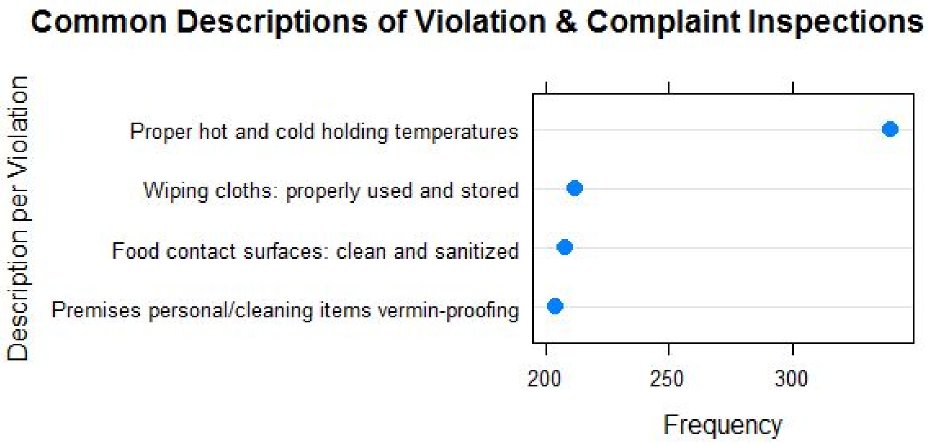
Middle Eastern/Thai restaurant in Yolo County are likely safer than the American or Mexican restaurants down the street--unless you were unlucky and happened to dine at the very unsanitary and unsafe Middle Eastern/Thai one.

Examining the Worst Eateries and Why

Having examined the data based on county-wide, city, and eatery type only two things remain: 1) looking at the worst individual eateries and 2) taking a cursory look at the reasons why eateries were scored poorly. The dotplot below looks at Yolo County's 1% worst eateries based on PNI scores. Following in suit with the bar graph on Page 6, all but one of these restaurants are in either Davis or Woodland.



Lastly, it is worth exploring the details of inspections that were in response to major violations and/or complaints to understand why some of these eateries were receiving low scores. Taking a subset of the original Yolo County dataset (where one row referred to one inspection), I used the `sort()` and `table()` functions to determine what the most common issue were--as described in the dotplot below.



Analysis Conclusion

In summary, I focused my analysis on the number of follow up inspections per eatery (which indicated a major violation occurred) and also created a percent score called the Percent of Negative-based Inspections or PNI (which takes into account the total number of inspections per eatery) to thoroughly examine the Yolo County health inspection data. Yolo County's distribution of PNI scores mirror the other counties, with more eateries having a positive health score overall. The highly populated cities of Davis, Woodland, and West Sacramento have a greater amount of the highly unsanitary eateries than smaller cities and unincorporated areas, perhaps implying a positive relationship between population/density and more dangerous, less sanitary eateries. In terms of the geographic distribution of unsanitary eateries within cities, they tend to be in the downtown areas. Lastly, taking both kinds of analysis into account, Mexican, American, and Middle Eastern/Thai restaurants in Yolo County are the most dangerous overall. Despite this similarity, these kinds of restaurants do have varying frequencies.

Alameda County

Description of the Dataset

The original Alameda county dataset had 129,922 observations with 12 variables, but I ended up cleaning and sorting it into two separate data sets. One had 119,064 observations of 12 variables, and the other data set had 7,481 observations of 10 variables. I needed these two separate data sets to look at different things. The data was collected for the eateries at various dates, ranging from July 2, 2012 to May 13, 2016, a span of roughly four years.

Alameda county's data was also different from the others in that everything in the score category was listed as NA. Instead of actual number results, there was just a ranking of "green" meaning everything was fine, "yellow" meaning be cautious, and "red" meaning the restaurant had so many violations that it was shut down (meanings from acgov.org/aceh/food/grading.htm). One thing I noticed while looking at these results was that there were a few columns other than green, yellow, and red. Two had extremely low observations, but one that had no name had nearly 10,000 observations. My hypothesis is that those in the blank category were factors that do not apply to the eatery and what they sell, but still are supposed to be checked by the health inspector. Therefore, they just leave them blank, as there is no reason to give a score for something not applicable.

Data Cleaning & Reorganization

My first step cleaning the data was to convert the results R, r, Y, y, G, and g into three categories: red (R & r), yellow (Y & y), and green (G & g). With those result I created the data set a_clean, which I used primarily for the descriptions of the violations.

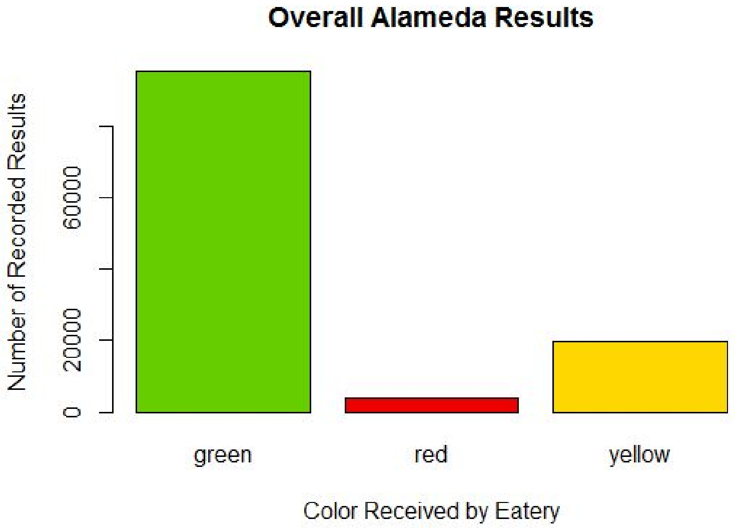
I then wanted to use this data and combine the business ids, in order to see the number of results in each violation. To do this, I created subsets of just red, yellow, green, and total results. Then I merged together these data sets by business ID, resulting in a single observation for each restaurant that shows the number of green, yellow, and red ratings they received. I then added back in the business ID, name, latitude, longitude, and city with the unique function, and merged all of that as well to create my second data set, a_clean2. I primarily used this data set to look at the restaurants in each city that had the highest amount of red violations.

Analysis

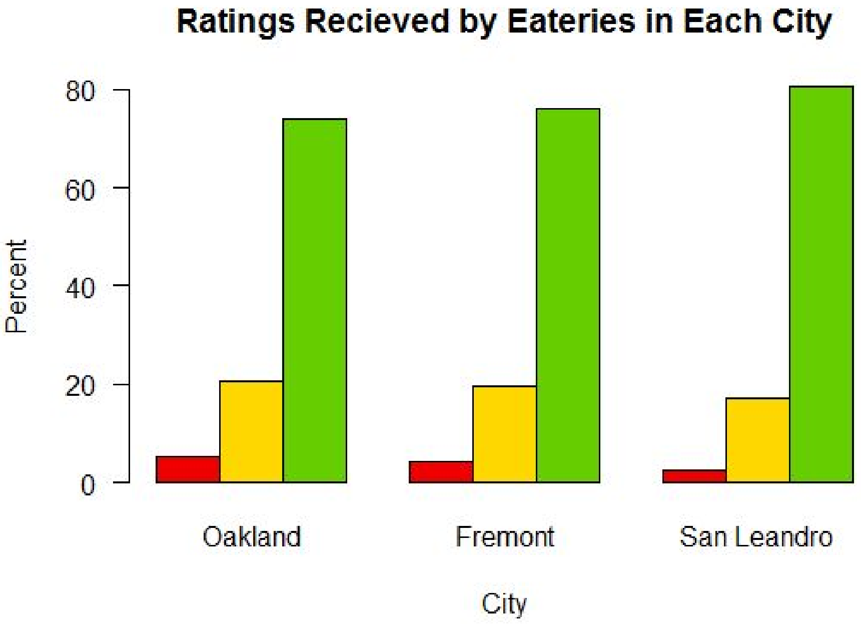
For my data I specifically looked at the three cities with the highest amount of observations: Oakland, Fremont, and San Leandro. I decided on these three cities, because I saw that they had the most observations.

As previously stated, I used two different cleaned data sets: a_clean and a_clean2. The reason I had to do this was because I could not merge the various observations of each restaurant together without getting rid of the descriptions, but I still wanted the overall description to look at and work with later, to see which violations were most common to receive red results for. So I used a_clean to see the descriptions, and a_clean2 to look at overall result totals from eateries.

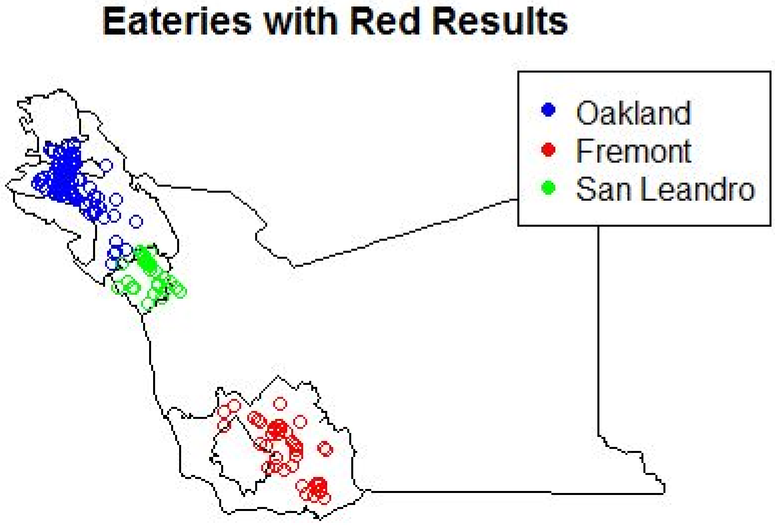
First, I wanted to look at Alameda county overall to compare later to the cities and see if there were any discrepancies. I found that a large majority of the results were green, and that only a small portion were red, and although yellow had more results than red, it was still a lot smaller than the green results. I took a look at the description of just the red results, and I saw that the most common issues at the eateries with red results was that the didn't pass the "No rodents, insects, birds, or animals" portion of the inspection.



Then I wanted to compare the difference between the results from each city, and how many of their eateries got green, yellow, or red ranking. What I saw in all of them was that a large majority of restaurants were green, and that only a small portion of them were red, similar to what I saw when I looked at Alameda county as a whole. However, Oakland had the most reds, at 5.4%, whereas the lowest, San Leandro, had half as many red eateries at 2.4%, showing that there definitely was some difference between the cities. Fremont was in the middle, with 4.4% of their eateries receiving red ratings. I also looked at the descriptions for the eateries with red results, but this time just the ones in each respective city. When doing so, I found that the results for this were also similar to the overall Alameda county results, and the most common was not passing the “No rodents, insects, birds, or animals” part of the inspection.

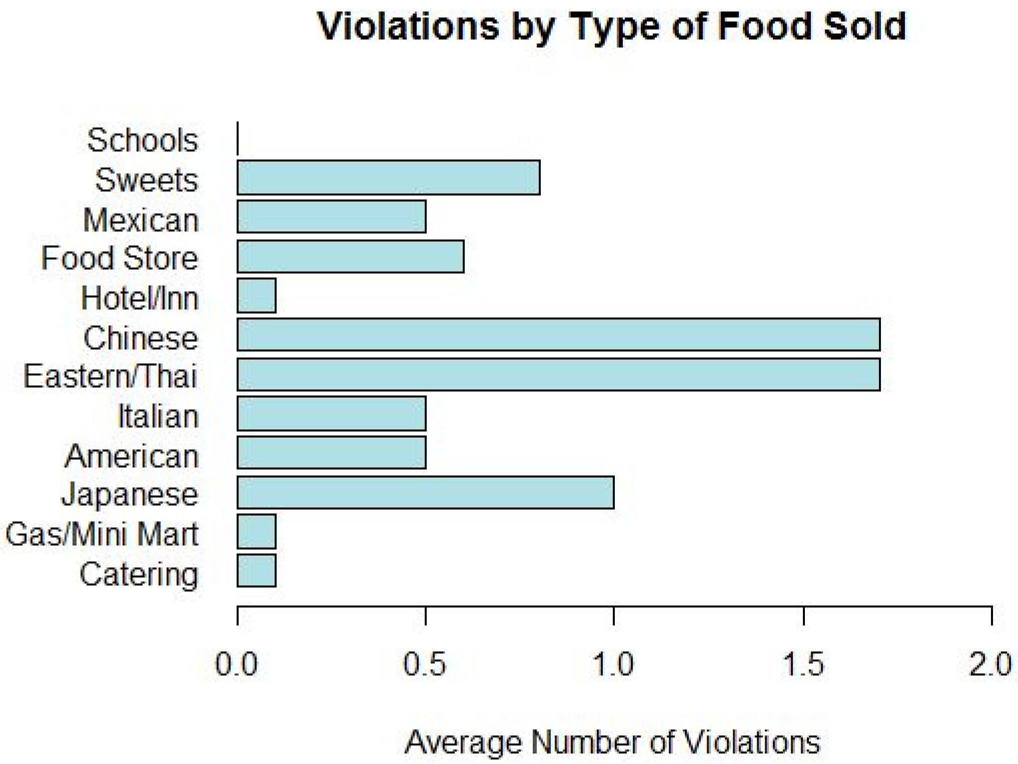


I also graphed the red results in the three cities I focused on. It showed that there were a lot of them all around certain areas, which I am assuming is downtown in the respective cities. As you move away from downtown, the restaurants are more spaced out. While further examining these subsets, I also found that the mean amount of red violations per city among those eateries that had received red results were 12.3 in Oakland, 11.3 in Fremont, and 9.7 in San Leandro. It is clear that Oakland had a significant amount more than the other two cities, and that San Leandro had the least. That could potentially be related to the population in each city, as Oakland is one of the larger cities in the Bay Area, but at the same time it could also be partially as a result of the overall city conditions.

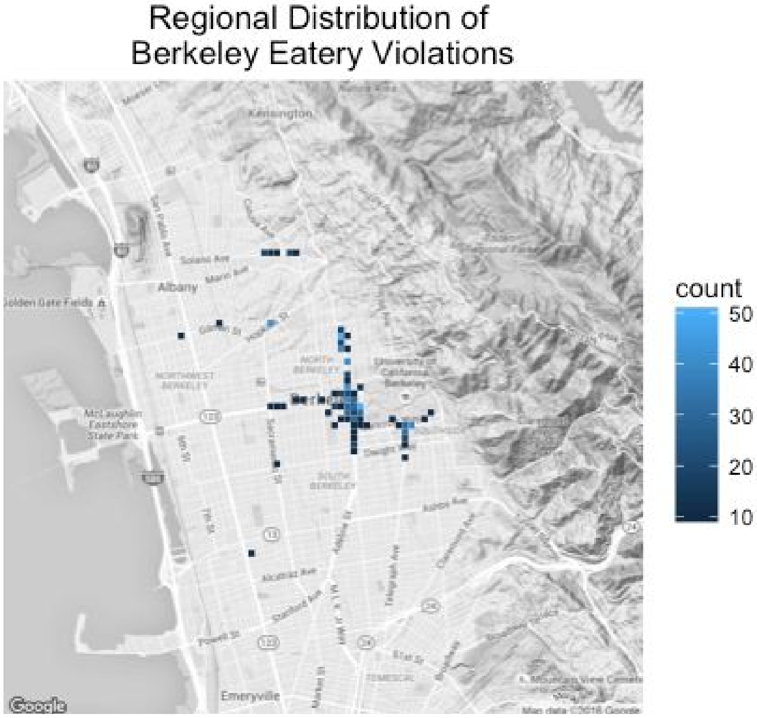


The highest number of red results in each city followed the same pattern, Oakland being the highest and San Leandro being the lowest. Oakland's highest was 44 at Vege House & Spices, Fremont's highest was 35 at General Pot, and San Leandro's highest was 27 at Foodnet Supermarket.

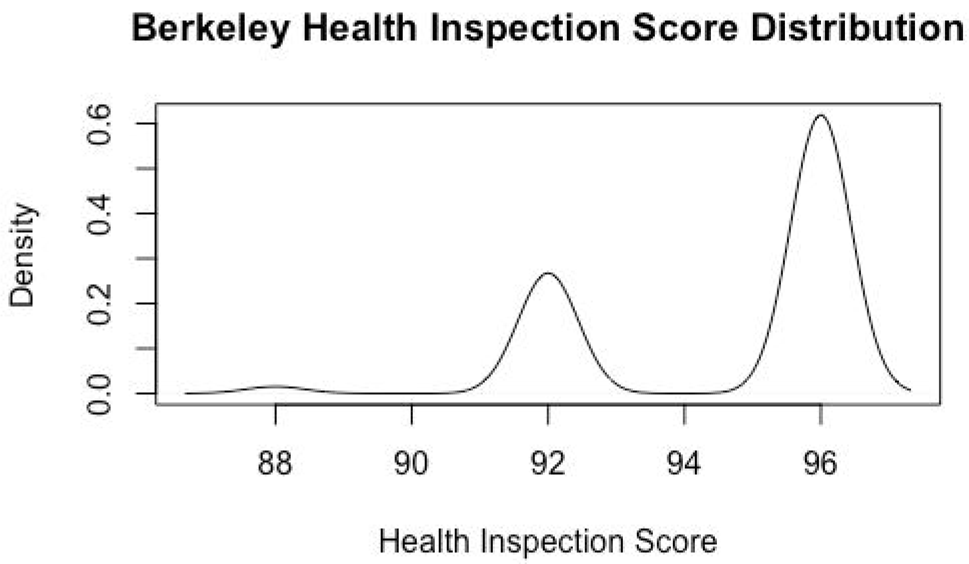
One of the things our team wanted to find out was if different types of food meant more or less violations. In order to do this and keep it consistent across our individual counties, we created similar subsets: Schools, sweets, Mexican, food store, hotel/inn, Chinese, Eastern/Thai, Italian, American, Japanese, gas/mini mart, and catering. Using these subsets, I graphed the mean number of violations in each subset. What I found was that both Chinese and Eastern/Thai had the most, and nearly the same, average violations overall. When looking at Japanese restaurants and the types of violations they had, I found that it was "Food contact surfaces: clean and sanitized" that was the most common among them. This could come from the use of raw fish, and not being able to properly clean up after them. Personally, the most shocking I find is how low the average violations are at gas stations or mini marts. I always go into them expecting the worst, but after seeing this I feel a bit relieved. I am also happy seeing that schools are the best when it comes to not violating codes.



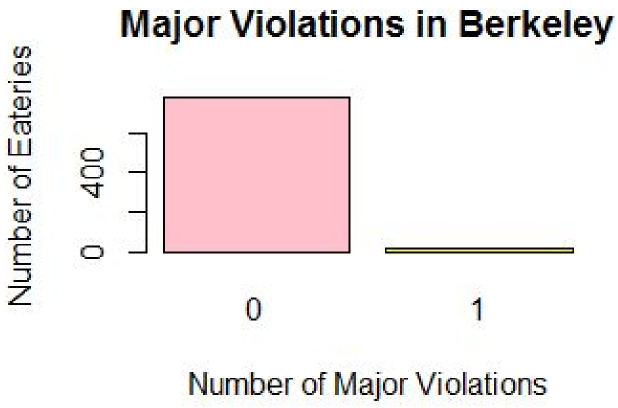
In Berkeley, the vast majority of health code violations are tightly clustered near the UCB campus, suggesting a high density of eatery locations nearby. This seems reasonable enough, considering the enormous foot traffic in the area.

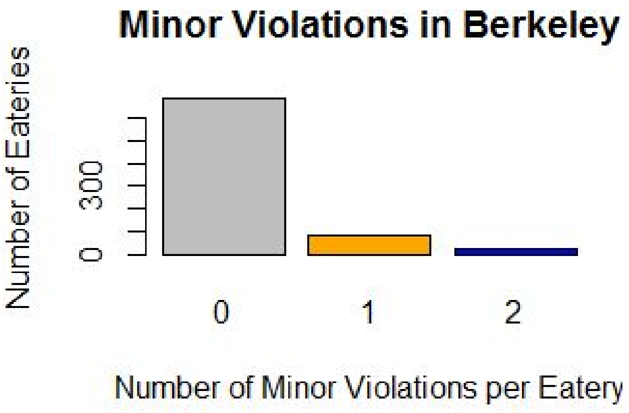


Overall, the distribution of Berkeley's health inspection scores (computed in Part 1 of this assignment) was relatively high, ranging from 88 to 96. Within this distribution, there are two major spikes at score values of 92 and 96, shown in the density plot below.



When looking at how many eateries had any minor violations, I saw that out of 791 observations only 23 eateries had minor violations and 83 had one. To compare, I looked at overall major violations, and saw that only 14 eateries had one, and no other had more than that. These results could be related to the rest of Alameda county. In the county, very few restaurants had red or yellow ratings, similar to how Berkeley had few restaurants with any type of violation.





Conclusion

Overall, in Alameda, a very small percentage of eateries have red ratings. There were obvious similarities between the ratings of the three cities I looked at and Alameda county as a whole. We also saw that, across Alameda county, the worst types of food places were Chinese and Eastern/Thai. Similar to Yolo, I noticed that the most populated city, Oakland, had more red results than the other two less populated cities. This could be just due to population, or it could be as a result of crowding in downtown. As for Berkeley, the data was not ranked in the same way as the rest of Alameda county. However, while looking at the places with violations, we can see that, similarly to the rest of Alameda, there are few places that actually have violations or low ratings. Even though we can see these similarities, we cannot conclude for sure that this Berkeley is perfectly related to Alameda, because they were not given the exact same data to work with.

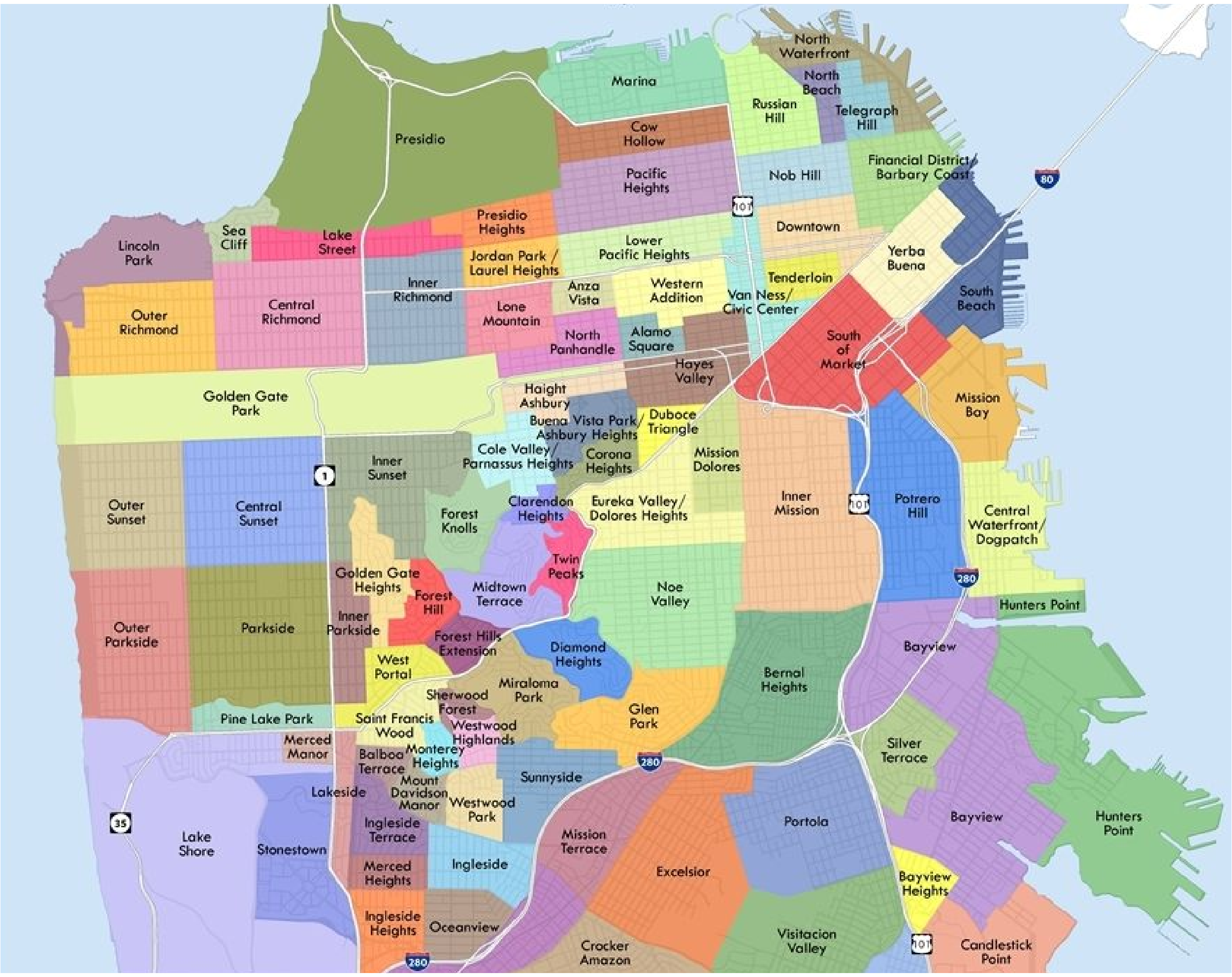
Description of SF County Dataset

San Francisco County's health inspection data followed Yelp's LIVES format, provided in three separate csv files. Merging these files by `business_id` and `date` yields a comprehensive health inspection dataset containing 59,381 observations of 22 variables. Each row of the merged dataset corresponds to an individual violation, explaining why a high number of restaurants are featured multiple times in different rows. Categorical features of the merged dataset include the city in which the business is located, violations' assessed risk type (split into low, medium and high), and written text descriptions of the violations themselves. The most significant numerical variable present defines the assessed health inspection score according to the sanitation of a given restaurant. In comparison to the other counties featured in this report, San Francisco's original health inspection dataset was presented in a more detailed, coherent and well-organized manner. Furthermore, the inclusion of a predefined health inspection score greatly assisted in the following analysis. However, some anomalies do exist. For example, the SF dataset contains observations of violations from cities like Oakland, which is not technically part of San Francisco County, located instead in Alameda - observations of which were largely excluded from analysis (Wikipedia).

Data Cleaning, Reorganization, and Assumptions

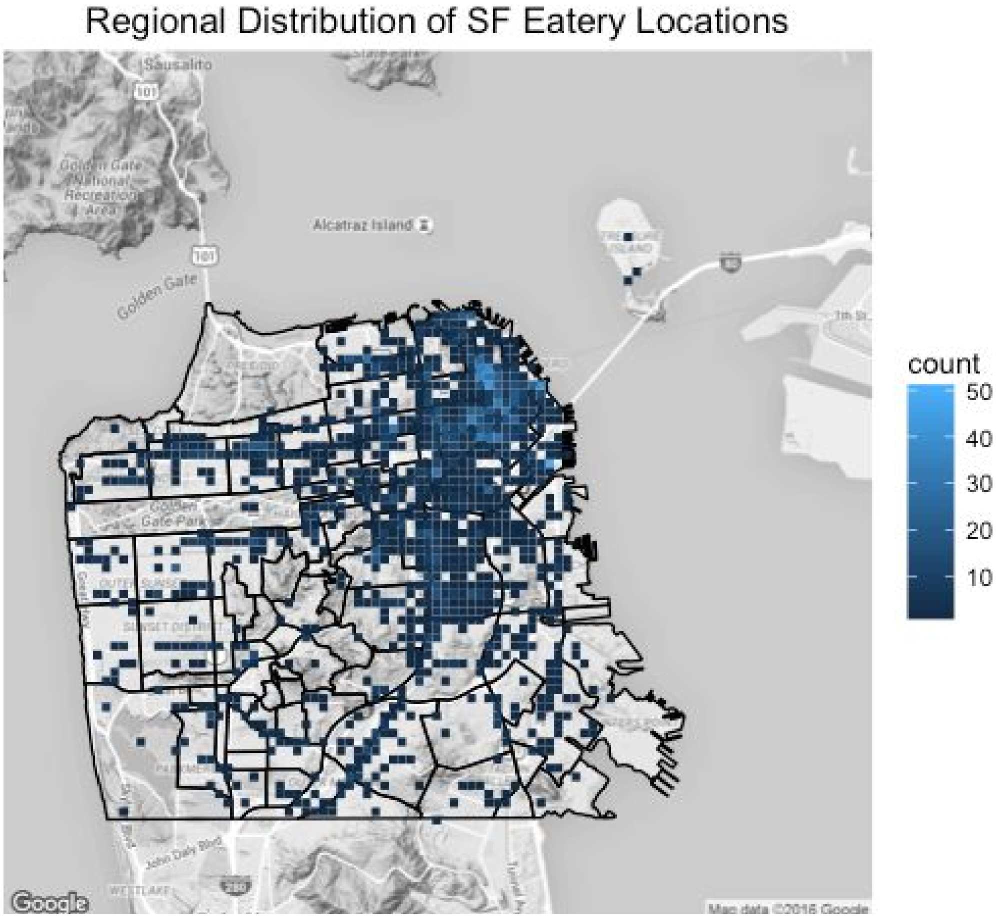
Using common exploratory methods, including `table()` and `sort()` among others, the merged SF data was reorganized to identify and emphasize significant trends. Each row of the merged dataset represented one violation, with no predefined variable alluding to which businesses were the most frequent offenders. In order to determine this information, a sorted table of `business_id` frequency counts was created, an effective tally of how many violations each business has been charged with. Likewise, there was no predefined variable regarding eateries' cuisine type or general function as an establishment. To mediate this lack of information, eateries' cuisine types were estimated by using the `grepl()` text search function and subsetting the original data into 10+ categories, including Chinese, Japanese, Thai / Indian, schools, convenience stores, hotels, and supermarkets, among others. The same general procedure was also used in estimating cuisine types for the other counties in this report. As a disclaimer, these breakdowns are not comprehensive and were determined by the presence of common keywords that our group manually defined by skimming through the `names` variable and identifying repeated strings. Using the stringr package, a new variable was then created to translate all the originally provided names into capitalized form for effective matching and consistency with the other counties analyzed in this report. This procedure assumes that eatery names accurately reflect the cuisine type provided, which is not necessarily true in all cases.

My analysis of health inspections in San Francisco focuses on geographic variations in health score of local food establishments, or “eateries”. There are 7,525 different eateries represented in the merged SF dataset, determined by applying the `length()` and `unique()` functions to the `business_id` variable. The map below from the Paragon Real Estate Group defines San Francisco’s neighborhood regions, informing the following discussion.

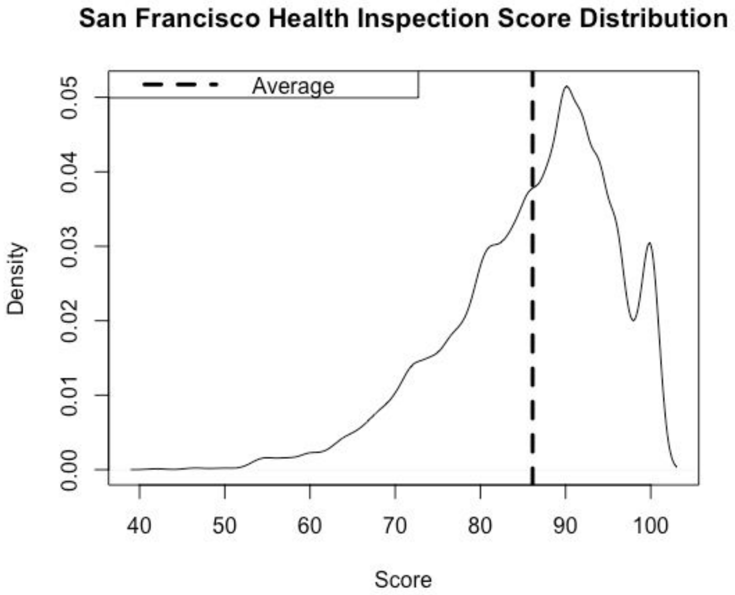


(Source: http://www.paragon-re.com/San_Francisco_Neighborhood_Map)

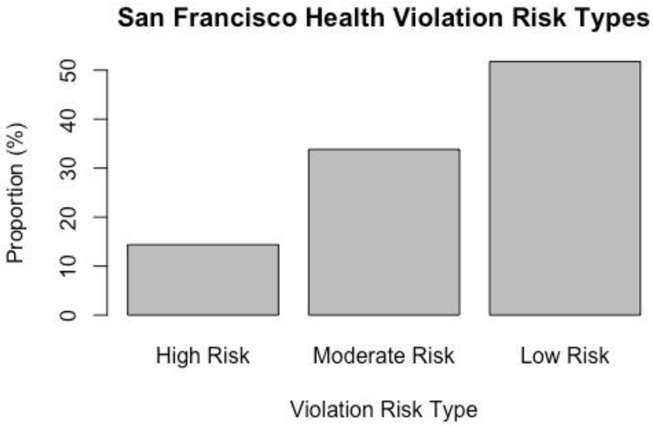
San Francisco has a lot of restaurants throughout the city, regardless of neighborhood. The map below was made by plotting the latitude and longitude coordinates of the SF business csv file. Downtown and the surrounding northwest region, including the Financial District, is an area with a particularly high density of eateries. This seems reasonable, since the location is a major economic and social hub of activity with high foot traffic and food vendors of all scales and reputability types seeking to conduct business here. Upon closer inspection, the Mission District and Marina also emerge as neighborhoods with a relatively high density of eatery locations. Hunter's Point stands out as a neighborhood with an usually low number of restaurants throughout. Perhaps this is due to the historic presence of industrial pollution in the area and its notoriously high crime rate being unconducive to and disruptive of an eatery's business operations. The Presidio is another anomalous neighborhood, with no eatery locations present, which seems reasonable considering how much of the area within the Presidio is undeveloped green space.



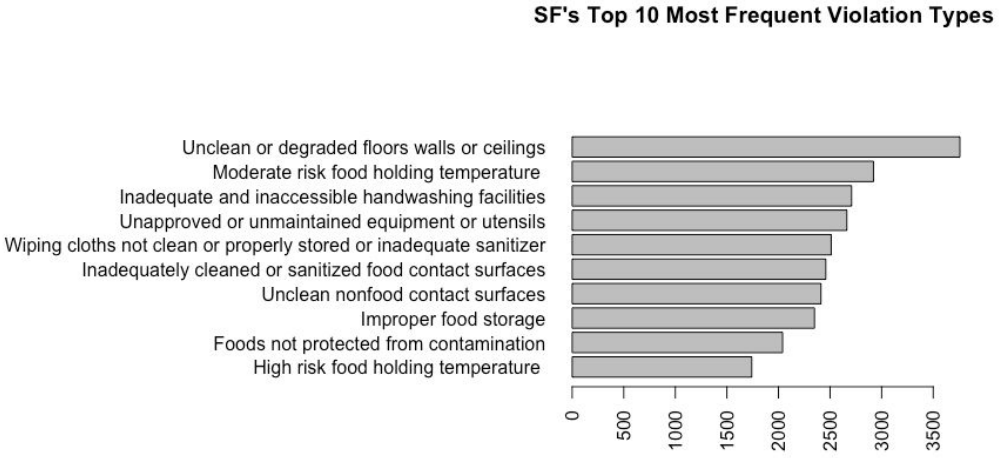
SF's health inspection scores ranged from 42 to 100 (a perfect score). The median health inspection score was 88, the average was 86 (s.d. = 9.5). Similarity between the mean and median suggests that few extreme outliers exist. According to the San Francisco Department of Public Health, eateries that score higher than 90 during inspections are classified as being in good "operating condition" (SFDPH). Eateries with a score of 86-90 are considered adequate, and those with a score between 71 and 85 "need improvement" (SFDPH). Locations with a score equal to or below 70 are considered as being in poor operating condition (SFDPH). SF's health inspection score distribution is visualized in the density plot below. In general terms, sanitary restaurants are more common than highly unsanitary restaurants in San Francisco, as defined by the SFDPH's established metrics.



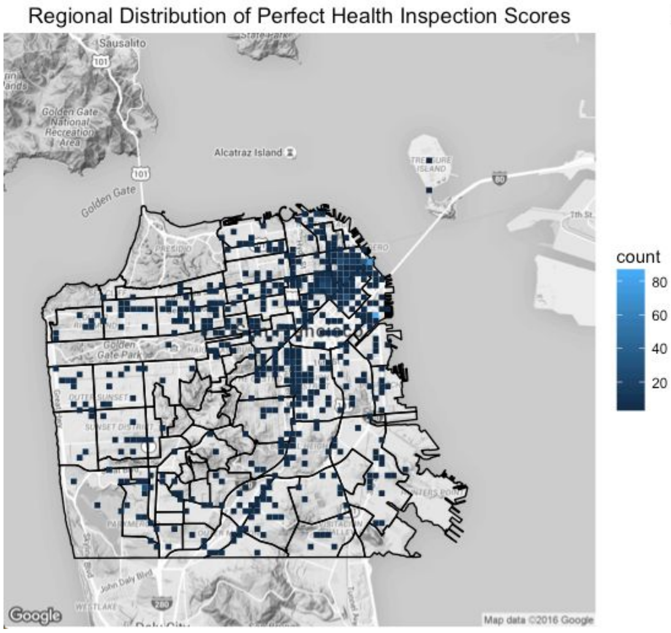
In a similar vein, low risk violations were by far the most common type of violation reported. High risk type violations were the least common, shown in the barplot below.

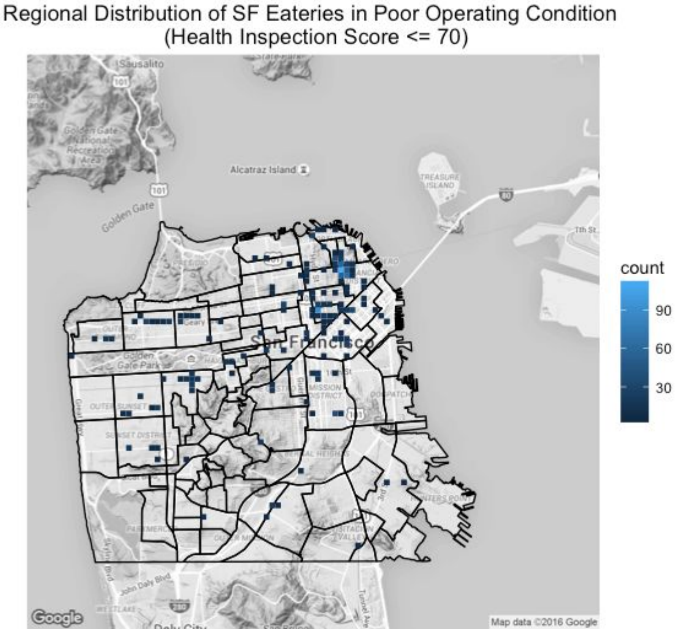


To investigate this concept further, the most common violation types, defined by included descriptions, are shown in the barplot below. Unclean or degraded floors, walls, or ceilings were the most frequent violation types present, supporting my conclusion that low risk violations are most common.


SF's Top 10 Most Frequent Violation Types

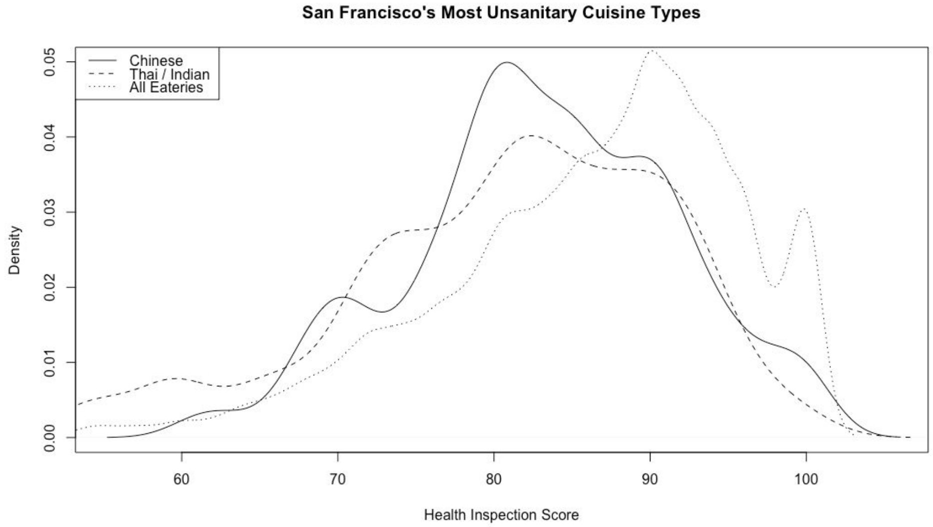
Mapping eatery locations subset by health inspection score reveals differences in the regional distribution of restaurants in terms of sanitary condition. This suggests that regional location is related to an eatery's cleanliness, but does not necessarily mean that location is a direct causal factor of inspection score.


Regional Distribution of Perfect Health Inspection Scores


Regional Distribution of SF Eateries in Poor Operating Condition (Health Inspection Score <= 70)

Shown in the map above on the left, the South Beach and Mission Bay neighborhoods have the highest density of locations with perfect scores - in stark contrast to the general distribution of all eateries, ensuring that this is not a mere byproduct of there being more restaurants within these neighborhoods in absolute terms. As predicted, SF eateries in poor operating condition were located in very different areas than those with perfect inspection scores. The extended downtown neighborhood, including Civic Center / Van Ness and the Tenderloin had the highest density of eateries in poor operating condition. This result seems reasonable considering the economically disadvantaged nature of these specific areas. However, this result is rather similar to the general distribution of eatery locations and therefore must be interpreted with skepticism. Perhaps there are simply more restaurants within this neighborhood in absolute terms - a potential underlying cause for these relatively high concentrations of eateries in poor operating condition. Another important distinction made apparent in the maps above is that eateries with perfect scores are far more prevalent than those in poor operating condition.

Evaluating health inspection scores subset by estimated cuisine type revealed the most dangerous and unsanitary eatery categories. Shown below, Thai / Indian eateries had the lowest score distribution of all 10+ categories our group defined. Chinese eateries had the second lowest distribution. A higher proportion of Thai / Indian eateries and Chinese eateries had health inspection scores lower than 85 in comparison to all eateries combined, suggesting that in general, these specific eatery types are more likely to be unsanitary. Furthermore, the most unsanitary eatery in San Francisco (determined by a sorted table of buisness_ids) is King of Thai Noodle, which is clearly a provider of Thai / Indian cuisine. With a similar trend also being present in most of the counties analyzed in this report, our group has concluded with reasonable certainty that these cuisine types are on average the least sanitary of those present. Of course, this interpretation is based on a number of significant assumptions, primarily in terms of defining the cuisine type categories in accordance with eatery names.



Despite the inconsistent structural composition of the original datasets for the counties of Yolo, Alameda, and San Francisco, a few common trends did emerge. Since the sanitary conditions of eateries were measured differently in each county, an exact and direct comparison cannot be made. Nevertheless, a holistic comparison is possible. Across all the counties, sanitary eateries were more common than extremely unsanitary ones. This was determined by comparing the distribution of estimated health inspection scores for eateries in Yolo and Alameda to San Francisco's distribution of predefined health inspection scores. Investigating the different cuisine types also yielded a shared trend; Middle Eastern/Thai eateries were generally among the lowest sanitary distributions of all defined cuisine types, regardless of county. As such, this type of eatery is one of the most "dangerous" across all three counties. Notably, this conclusion is based on the assumption that an eatery's title accurately reflects its cuisine type or general function as an establishment, which may not necessarily be true. Another caveat of our conclusions relates to the different date ranges included in each county's dataset. Yolo's data covered by far the longest time period, more than 9 years, which inflated its analysis of follow ups/major violations because there was simply more time for data to accumulate. However, this longer timespan did not necessarily affect its PNI scores which were weighted. If more time was available to complete this report, the discrepancy in dates covered by each county's data would have been controlled for by subsetting each dataset to have a common time interval.

Sources Cited:

Alameda County Health Department:
http://acgov.org/aceh/food/grading.htm

San Francisco Department of Public Health:
https://www.sfdph.org/dph/EH/Food/Score/

Paragon Real Estate Group:
http://www.paragon-re.com/San_Francisco_Neighborhood_Map

Yolo County Health Department:
http://www.yolocounty.org/health-human-services/health-department/restaurant-search/violations