title: "STA 220 Project: Insert Title" author: "Sandeep Nair, Alan Phan, Thommas Phan" date: "March 20, 2024" output: html_document: df_print: paged number_sections: yes

Introduction / Background

In today's landscape of heightened concern for food safety and public health, regulatory bodies like the Sacramento County Environmental Management Department (SCEMD) play a crucial role in ensuring community well-being. Driven by a commitment to enhancing quality of life, SCEMD employs various strategies including education, surveillance, enforcement, and community service, with a focus on mitigating foodborne illnesses—a significant health and economic burden.

Millions suffer from foodborne illnesses annually in the US, prompting organizations like the Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA) to identify key risk factors and interventions. These encompass issues such as improper temperature control, inadequate cooking, poor hygiene, and contaminated equipment, among others.

SCEMD conducts routine inspections of retail food establishments in Sacramento County, enforcing strict compliance with health and safety codes outlined in state and county regulations. These inspections are guided by comprehensive protocols, such as the Sacramento County Retail Food Code Inspection Guide, aimed at maintaining standardized practices across establishments.

Effective food safety regulation hinges on proactive measures to identify and mitigate risks within the food supply chain. SCEMD oversees a diverse range of establishments, implementing measures to uphold hygiene and sanitation standards while reducing the incidence of foodborne illnesses.

In addition to enforcement, SCEMD engages in community outreach and educational initiatives, fostering collaboration between stakeholders. These efforts aim to empower establishments and consumers with knowledge and resources for maintaining a safe food environment.

This project embarks on an exploratory data analysis, leveraging data scraped from SCEMD reports and Yelp. Through visualization and observation, the project aims to uncover insights into health code violations within Sacramento County’s food establishments over the past year, contributing to the broader discourse on food safety and regulatory enforcement.

Data Scraping / Organization

All data was scraped via Python and processed as either lists or dictionaries, and made into a data frame for visualizations and report format writeup in R Markdown, in the following sections. Below will outline the courses of action taken to scrape and organize the data and any problems we faced during the process.

Obtaining Links to Sacramento County Inspection Data for the Past 12 Months

To collect links to inspection data from the Sacramento County Environmental Management Department (SCEMD) for the previous 12 months, a systematic approach was devised, leveraging Selenium for web automation. Initially, the process involved selecting specific date ranges from the SCEMD's inspection database, which necessitated interaction with a calendar interface. Initially, the focus was on extracting data from the current and last month, facilitated by tabs on the calendar. However, this approach proved insufficient for capturing a comprehensive dataset.

Refinement of the date range selection process was crucial to obtaining a broader dataset, encompassing inspections from March 2023 to the present. This required a more intricate methodology. By identifying the XPath for the "Previous" and "Next" arrow buttons on the calendar interface, the navigation process was automated. By manually navigating through the calendar and observing the number of clicks required to reach specific dates, a systematic approach was devised. Subsequently, Selenium was utilized to programmatically perform the requisite number of clicks on the arrow buttons to navigate to the desired dates.

Handling dynamic page loading was another challenge encountered during the data collection process. Upon selecting specific dates, it was noted that the inspection data did not load entirely, necessitating interaction with a "Load More" button. To address this, a while loop was implemented to iteratively click the "Load More" button until all inspection data was loaded.

Once the complete inspection data was loaded, each inspection's URL was extracted for further processing. Utilizing CSS selectors, the "View" buttons corresponding to each inspection were targeted, as XPath did not consistently retrieve the necessary information. The attributes (links) associated with each "View" button were then saved into a list for subsequent analysis.

By meticulously executing the outlined steps, a comprehensive collection of links to Sacramento County inspection data for the past 12 months was obtained. This process laid the groundwork for subsequent data processing and analysis, enabling insights into health code violations within the county's food establishments.

Scraping Data from the Obtained Links to Inspection Data

In this phase of the project, the focus shifted towards extracting relevant information from the links obtained in the previous step, which led to summaries of observations and corrective actions from specific health inspections conducted at various restaurants. Each page was scraped using BeautifulSoup, a Python library for parsing HTML and XML documents. The process involved retrieving details such as the establishment name, inspection date, address, and health code violations.

One significant consideration was the possibility of multiple health inspections conducted at the same restaurant over the course of a year. Therefore, the scraping process needed to account for this potential repetition of data.

The scraping process was executed by iterating through each URL obtained from the previous step and extracting the required information. The key steps involved in the scraping process were as follows:

1. **Establishment Information Extraction:**
    - The BeautifulSoup library was employed to parse the HTML content of each page.
    - The establishment name and address were extracted from the designated HTML elements.
2. **Inspection Date Retrieval:**
    - Inspection dates were located within specific HTML elements and extracted accordingly.
3. **Health Code Violations Extraction:**
    - Health code violations were identified within the inspection summaries.
    - Each violation was associated with a specific health code, which ranged from 1 to 49. Sometimes the number had sub-letters, such as 1b and 1c.
    - To streamline analysis, only the primary code for each violation was retained.
    - A dictionary mapping each health code to its corresponding description was manually inputted from a PDF provided by the Sacramento County titled "Retail Food Inspection Guide."
4. **Handling Large Volume of Links:**
    - Given the substantial volume of links (nearly 10,000), the scraping process was divided into manageable chunks (500-1000 links at a time) to mitigate the risk of failure.
    - To prevent potential server blocks or interruptions, a sleep timer of 1 second was implemented between requests.
5. **Error Handling and Data Cleanup:**
    - Throughout the scraping process, error handling mechanisms were in place to identify and address any issues encountered.
    - Approximately 15 links were identified as faulty, lacking the establishment name, and subsequently discarded from the dataset.
6. **Data Storage and Merging:**

- Results from each scraping iteration were saved as JSON files for further processing.
- At the conclusion of the scraping process, the collected data from all iterations were merged to form a comprehensive dataset for subsequent analysis.

The provided code snippet serves as a reference for the scraping process, showcasing the implementation of BeautifulSoup to extract relevant data from individual URLs. By systematically executing the scraping process and meticulously handling data intricacies, a robust dataset was assembled, laying the groundwork for subsequent exploratory data analysis.

Obtaining Supplementary Yelp Data

In order to augment our dataset with additional information about food establishments in the greater Sacramento area, we utilized the Yelp Fusion API. This API allows for querying establishments based on various parameters such as location, cuisine type, and price level. Our objective was to gather supplementary data on restaurants, grocery stores, and convenience stores—any establishments selling food items.

Here's an overview of the process:

1. **API Querying:**
   - We employed the Yelp Fusion API to conduct search queries for restaurants, grocery stores, and convenience stores in the greater Sacramento area.
   - Each establishment's search query results provided information such as rating, review count, price level, cuisine type, address, and coordinates.
2. **Handling API Limits:**
   - The Yelp API imposes query limits, allowing only up to 1000 establishments per query. To circumvent this limitation and obtain a comprehensive list, we executed multiple queries with different parameters.
3. **Optimizing Query Results:**
   - The search queries return results based on location and a sorting algorithm. We utilized four types of sorting operations: distance, rating counts, ratings, and best matches.
   - To maximize the number of results, we conducted queries for multiple locations across the greater Sacramento region, including major cities like Elk Grove, Citrus Heights, and Folsom.
4. **Data Cleaning and Filtering:**
   - After aggregating results from all queries, we noticed that some establishments, particularly those offering only takeout and delivery services, did not have valid addresses listed. As these were not relevant to our analysis, we removed them from the dataset.

5. **Result Summary:**
    - In total, we gathered Yelp metadata for 4283 establishments, providing valuable supplementary information to enrich our dataset.

By obtaining supplementary Yelp data, we aimed to enhance our understanding of the food landscape in the greater Sacramento area, complementing the insights derived from SCEMD inspection reports. This additional information will facilitate a more comprehensive analysis of factors influencing food safety and consumer preferences within the region.

Creating a Key from our Datasets to Merge

In order to merge our datasets effectively, we needed to devise a unique identifier for each restaurant. This presented a challenge as traditional identifiers such as name, address, and zip code were not suitable due to the presence of multiple establishments with similar or identical attributes. After careful consideration, we settled on a combination of the first four numbers from the establishment's address and the first three letters of its name.

This approach offered a high degree of uniqueness, as it was highly improbable for two establishments with the same name to share the exact same four-number address. By incorporating both elements, we aimed to create a robust key that would facilitate accurate merging of our datasets.

Upon merging the datasets, we encountered a common issue where establishments appeared in multiple rows due to having multiple observations (e.g., health inspection data) associated with them. This occurred particularly because the health inspection data was scraped using dates, leading to multiple inspection records for the same restaurant.

To address this, we implemented a group-by operation followed by an aggregation function. By grouping the data by our established key, we were able to aggregate multiple observations for each establishment. Specifically, we retained the first observation for duplicate data points such as name, latitude, longitude, rating, price, etc., while consolidating different observations into a list. For instance, health code violations and inspection dates were aggregated into lists for each establishment.

This process resulted in a streamlined dataset with a reduced number of observations, from 3.6k to 2.3k. By consolidating duplicate entries and organizing the data systematically, we ensured the integrity and accuracy of our merged dataset, laying the groundwork for subsequent analysis and insights into the food landscape in the greater Sacramento area.

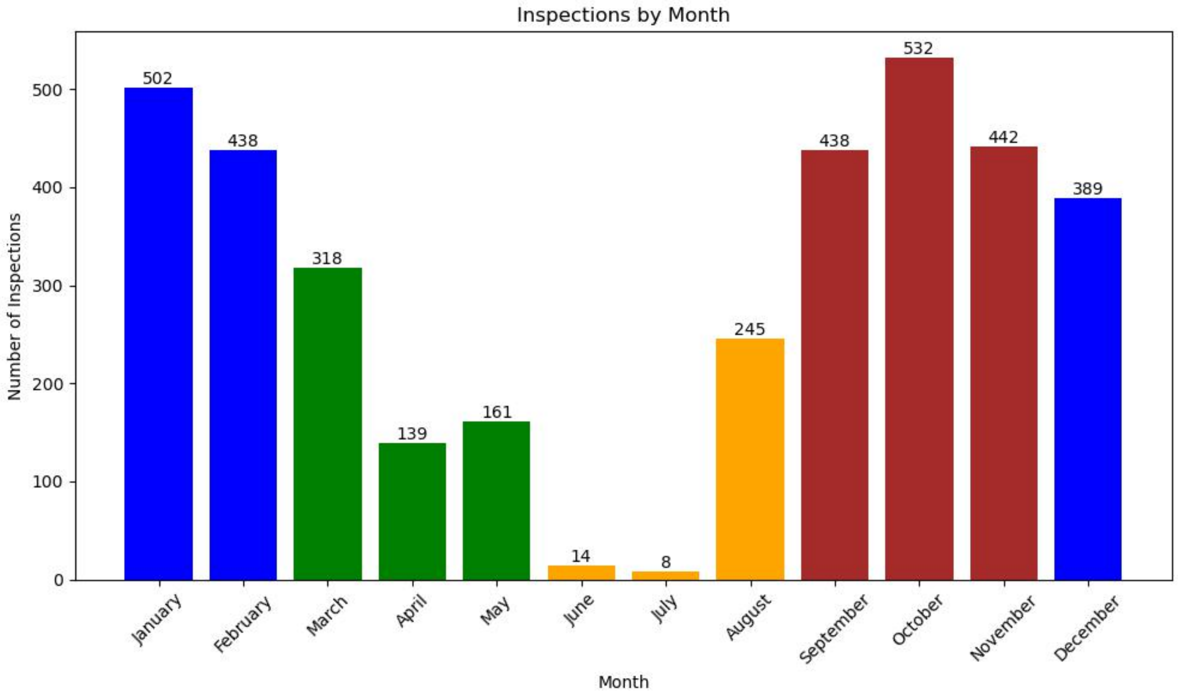
Data Visualization

Visualizing Inspections by Month

In this section, we present a visualization of the number of inspections conducted in Sacramento over a 12-month period. The graph is segmented by seasons to provide insights into seasonal variations in inspection activity.

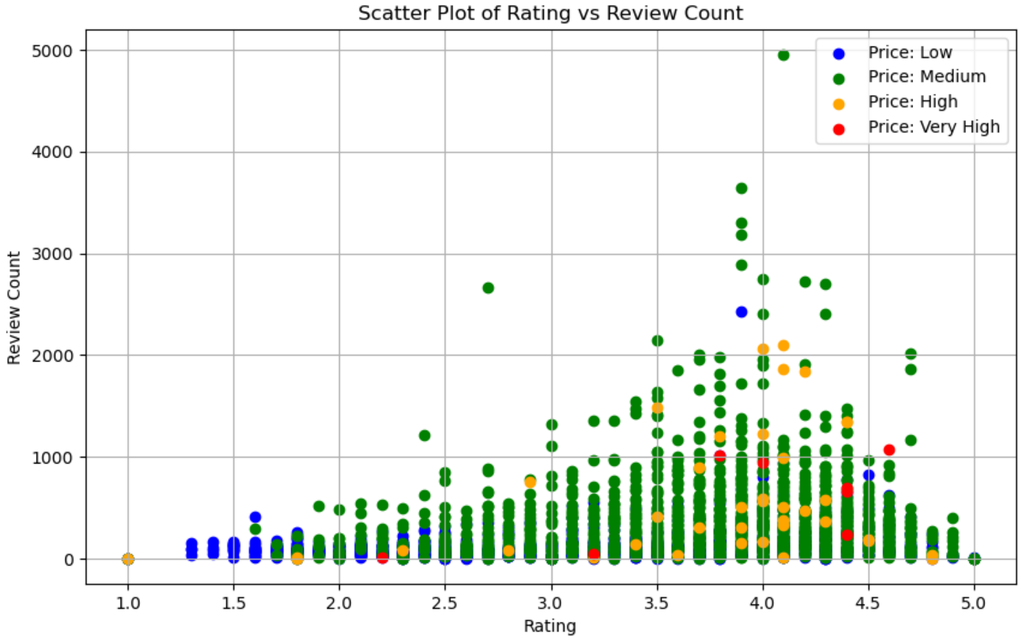
The visualization is created using a bar plot, with each bar representing the number of inspections conducted in a specific month. The color of each bar corresponds to the season in which the month falls: blue for winter, green for spring, orange for summer, and brown for autumn.

Upon examination of the graph, it's evident that inspection activity varies across seasons.



It's noteworthy that inspection activity tends to peak during the colder months (winter and autumn), with relatively lower activity observed during the warmer months (spring and summer). However, an anomaly is observed in August, where there's a significant increase in inspections compared to the preceding summer months.

This visualization offers valuable insights into seasonal patterns in inspection activity, which can inform resource allocation and planning for regulatory agencies and stakeholders in the food industry. Additionally, it underscores the importance of considering seasonal variations when analyzing inspection data and implementing targeted interventions to maintain food safety standards year-round.

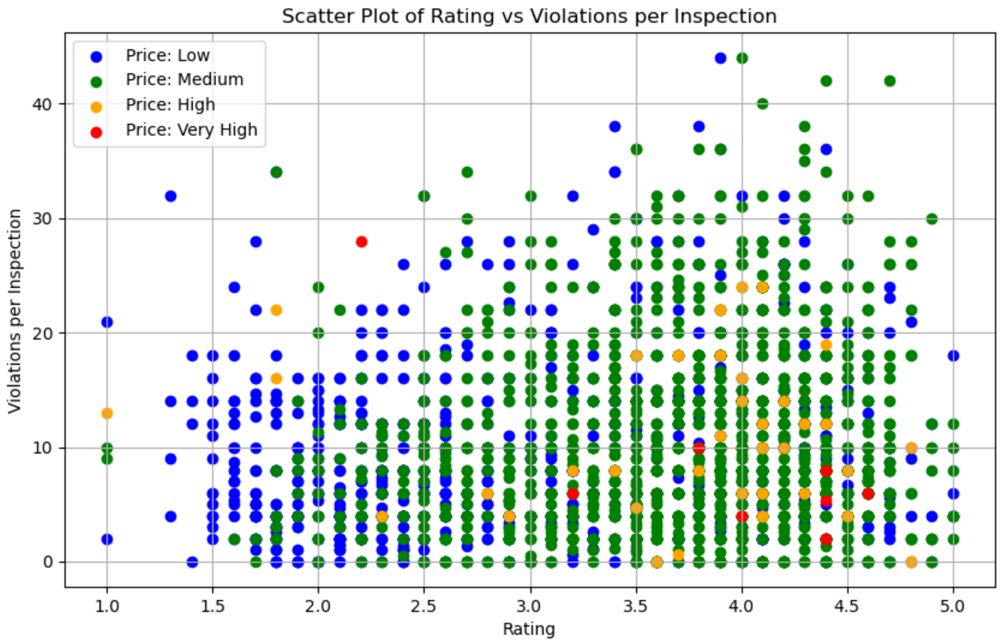


Visualizing Health Code Violations

In this section, we delve into visualizing health code violations within food establishments, aiming to glean insights into compliance trends and potential areas for improvement. To ensure a fair comparison across establishments, we chose to calculate health code violations per inspection, thereby mitigating the impact of restaurants with a higher number of inspections in the last 12 months.

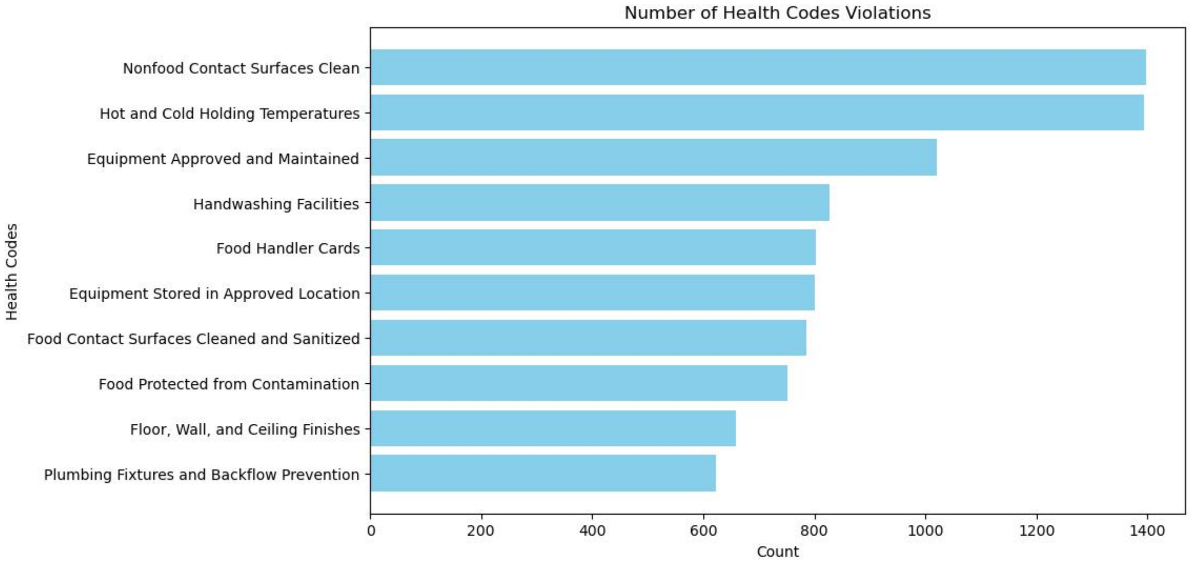
One of the visualization techniques employed is a scatter plot, leveraging data on establishment ratings and the number of violations per inspection. Each data point in the scatter plot represents an individual establishment, with the x-axis indicating the establishment's rating and the y-axis denoting the number of violations observed per inspection.

Through this scatter plot, we can discern any potential correlations between establishment ratings and health code violations. Additionally, by observing patterns and outliers, we can identify establishments that warrant closer scrutiny or intervention from regulatory agencies. This visualization aids stakeholders in the food industry and regulatory bodies in prioritizing inspections and implementing targeted interventions to enhance food safety standards and safeguard public health.



From this Scatter Plot, it is noteworthy that there is a general trend with more violations and higher ratings across all price points in restaurants. Also, there is a clear correlation between price point and rating, where lower priced restaurants tend to have lower ratings and higher priced restaurants generally have higher ratings.

In the next figure, Number of Health Codes Violations, we visualize the most common health code violations in Sacramento in the past 12 months.



Logic Behind Bayesian Averages

In this data visualization section, we aimed to identify the top categories with the highest ratings. However, a challenge arose when utilizing average ratings as a metric, as it can be heavily skewed by differing sample sizes. To address this issue and ensure more reliable ratings, we employed Bayesian averaging.

Here's how our Bayesian weighted averages were calculated:

1. **Prior Rating and Sample Size Calculation:**

   Prior ratings were determined differently based on our goal. For restaurants we did a sample size prior of 50 (we believe that a minimum of 50 reviews is a good estimate for gauging a restaurant) and a prior average rating of 4 (generally a common number for an average rating).

   For the Categories calculation, since we are grouping by categories, it is more difficult to determine a reasonable prior that intuitively made sense. For these we utilized the average rating of the categories and then the average number of reviews for the priors. However, specifically for the top 10 most rated categories graph we ignored the prior because there are a lot of reviews for these categories.

2. **Application of Bayesian Average Formula:**

   - Bayesian average is calculated using the formula:

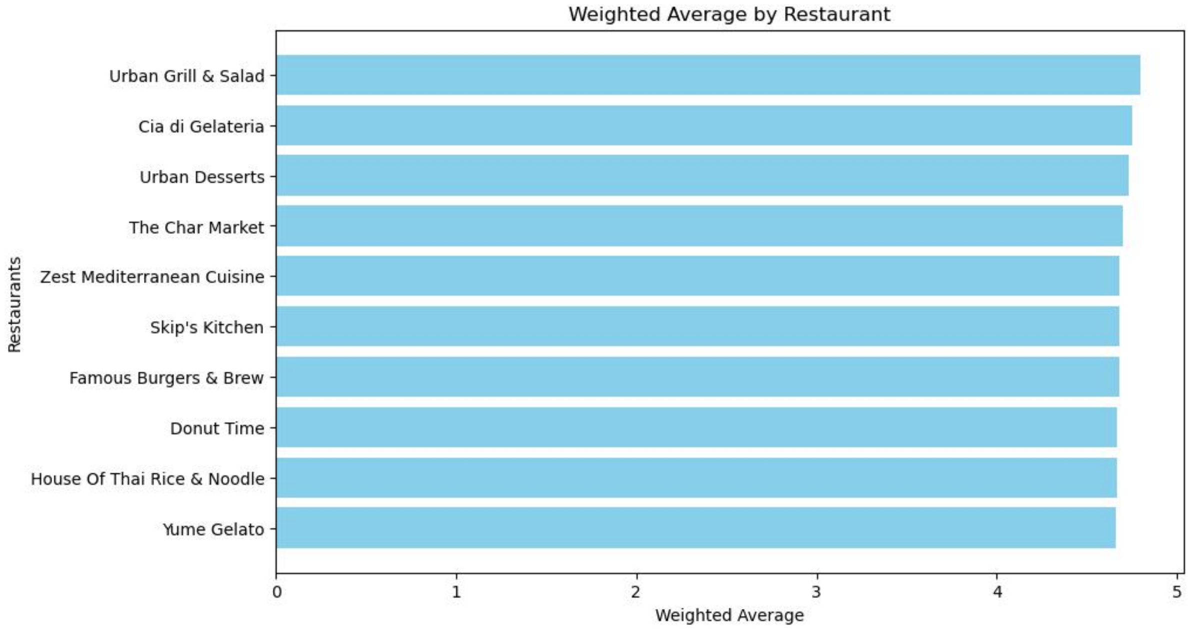
     ```
     Bayesian Average = ((c * m) + (number of ratings * review)) / (c +
     ```

     Where:
     - 'c' is the prior sample size
     - 'm' is the prior rating
     - 'number of ratings' is the number of ratings for a specific category
     - 'review' is the average rating for that category

By applying this Bayesian average formula, we aimed to mitigate the effects of small sample sizes on the calculated average ratings. This approach allows for more robust and reliable ratings by incorporating both the prior rating and the actual ratings from the sample, thus providing a more accurate representation of the category's overall rating.

Utilizing Bayesian averaging helped to address the inherent limitations of relying solely on average ratings, particularly when dealing with categories with varying sample sizes. With this approach, we aimed to ensure a more equitable representation of each category's rating, enabling us to identify the top categories with confidence in our data visualization analysis.

The following 4 Figures utilize the aforementioned Bayesian Averaging methods.

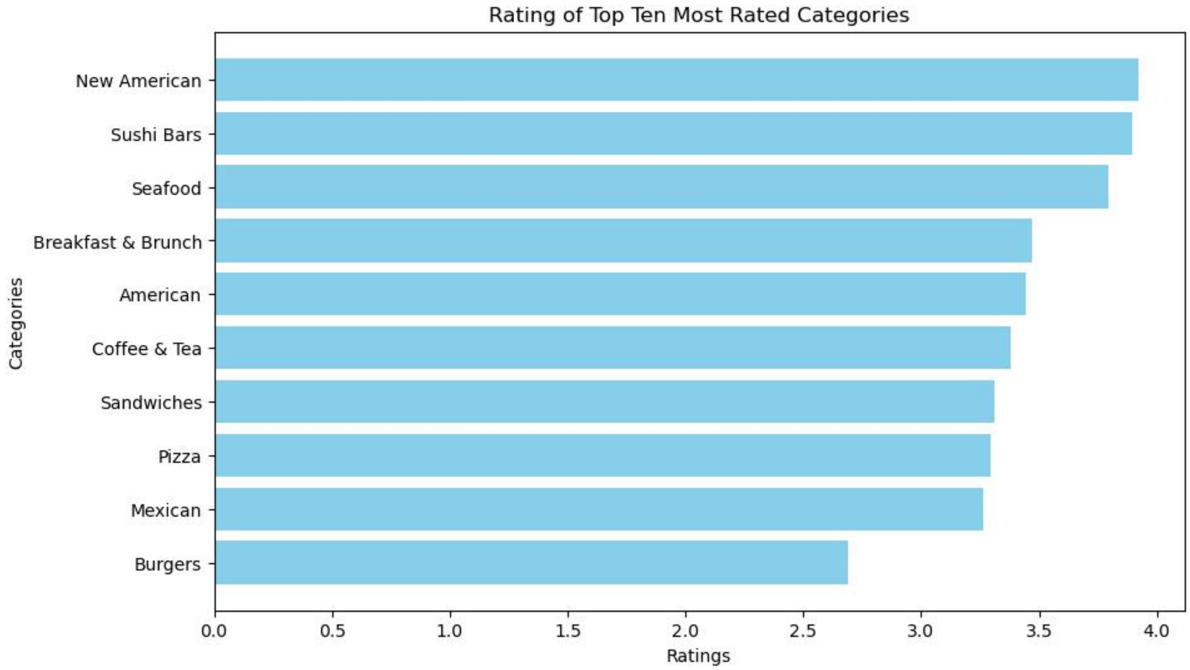


The graph above shows the top ten highest weighted average restaurants in the greater Sacramento Region.

alt text

The graph above shows the top ten highest rated categories. In this order the top ten are:

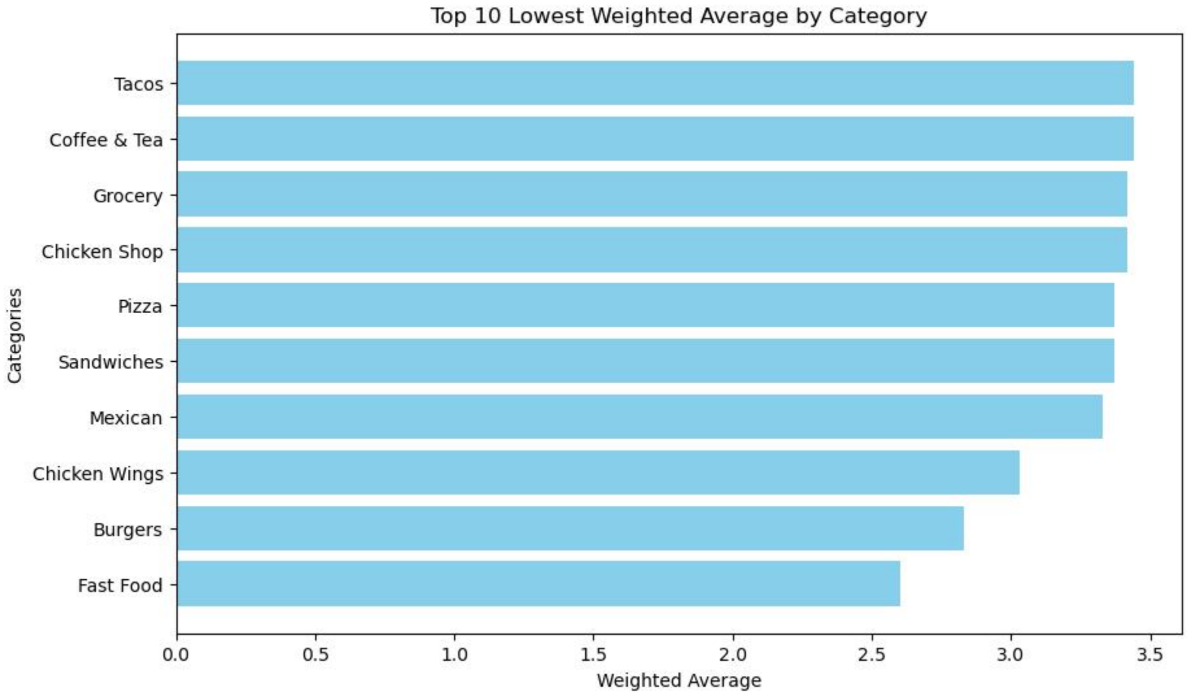
1. Mediterranean
2. Vegan
3. Halal
4. Coffee Roasteries
5. Wine Bars
6. Greek
7. Thai
8. Korean
9. Bubble Tea
10. Indian



The graph above shows top 10 most rated categories ordered by highest ratings. In this order the top ten are:

1. New American
2. Sushi Bars
3. Seafood
4. Breakfast and Brunch
5. American
6. Coffee & Tea
7. Sandwiches
8. Pizza
9. Mexican
10. Burgers

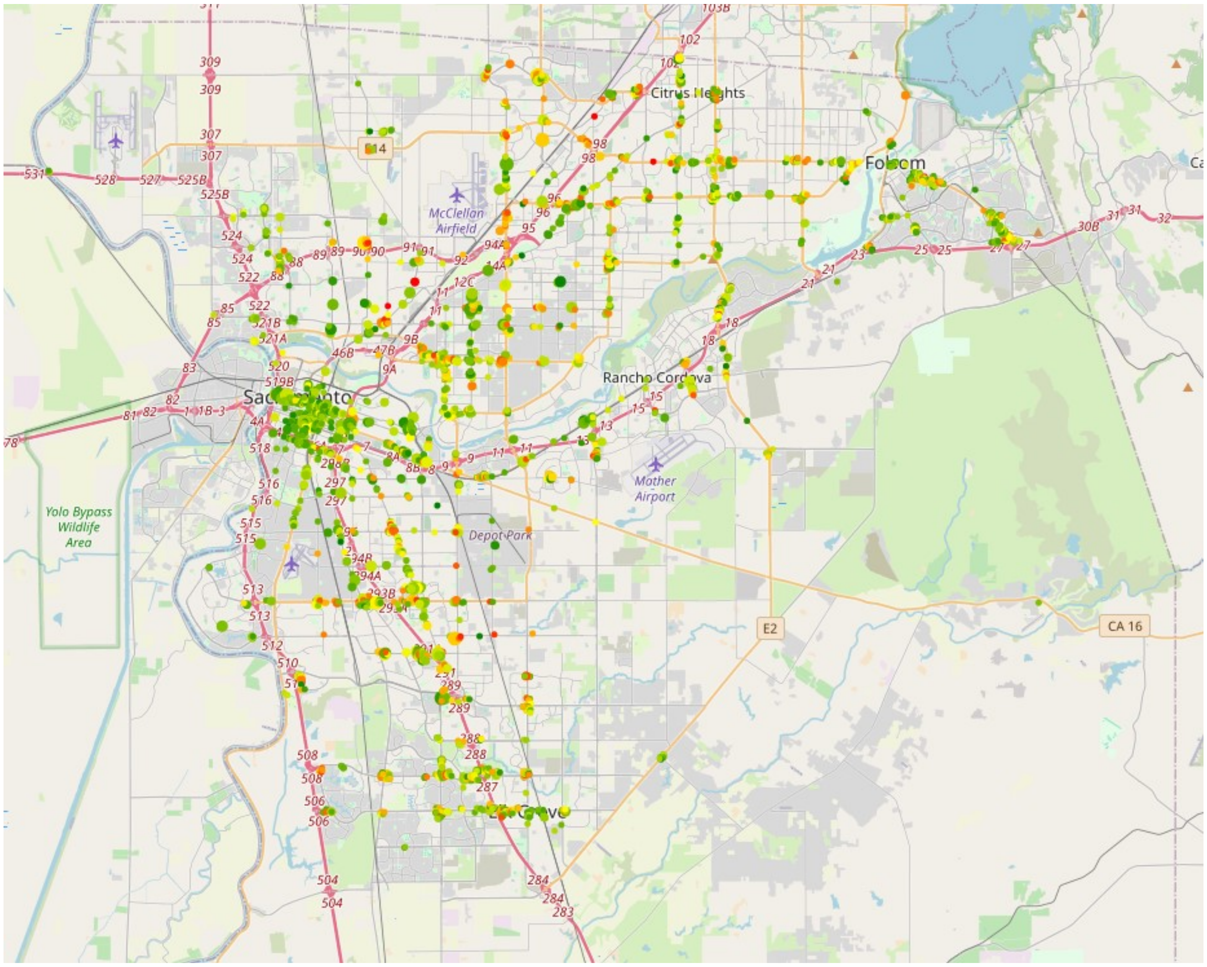
Intuitively these categories being the top ten makes sense. There are a lot of restaurants that cook at least one of these categories which would increase the odds of that category being rated.



The Image above shows the top 10 lowest rated categories. In this order the lowest ten are:

1. Tacos
2. Coffee & Tea
3. Grocery
4. Chicken Shop
5. Pizza
6. Sandwiches
7. Mexican
8. Chicken Wings
9. Burgers
10. Fast Food

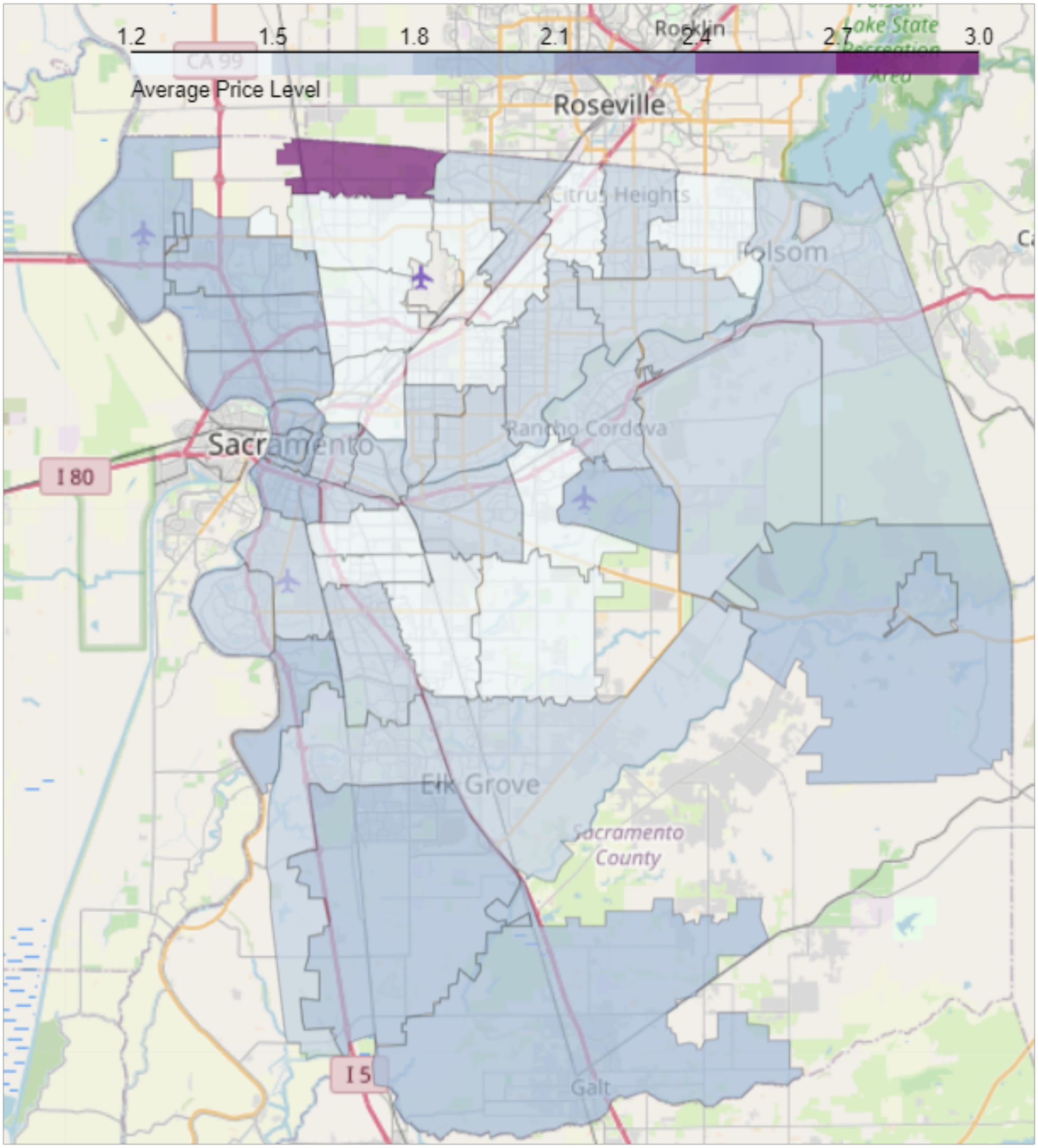
These ratings make intuitive sense as these categories has a fast food restaurant tied to them which would drastically lower the average ratings.



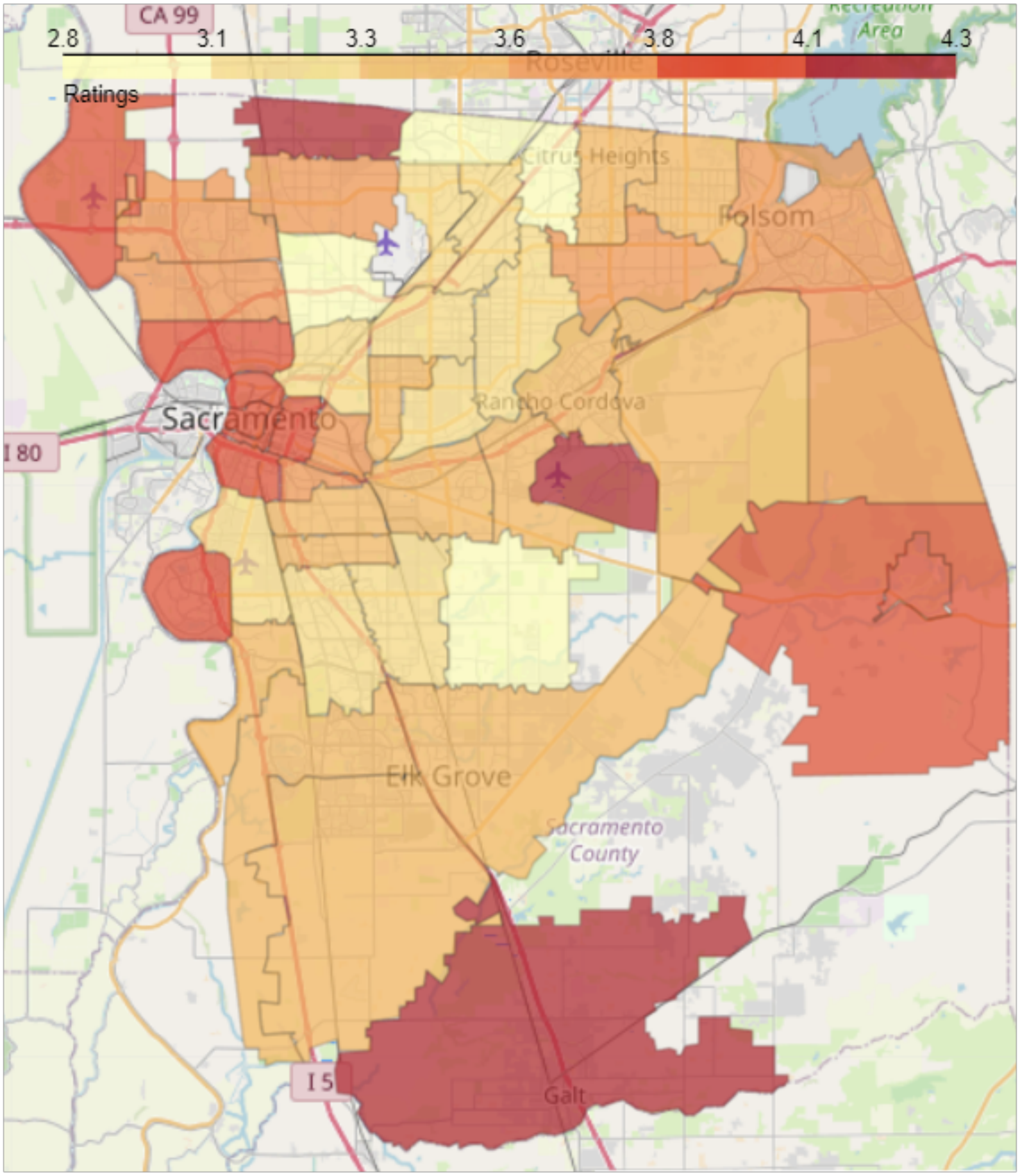
The plot above shows a map of Sacramento where each point is a restaurant. The color represents the average waiting of the restaurant (green being highest rated, yellow being mid rated, and red being low rated). The size of the point represents the number of health code violations found by the restaurant's health inspections in a year. Interestingly it appears that downtown Sacramento has decently rated restaurants while restaurants further from downtown sacramento tends to have lower rated restaurants.

Geoplot Depicting Variation of Average Price Level and Average Ratings of Different Zipcodes

Visualization of restaurant data by zip code reveals notable patterns in both price levels and ratings.



Price levels, denoted by the number of dollar signs ($ to $$$$), exhibit a predictable trend across neighborhoods. Generally, areas with higher affluence, such as downtown, Folsom, and Elk Grove, tend to boast higher average price levels. This suggests a correlation between neighborhood socioeconomic status and restaurant pricing.



Similarly, the distribution of ratings follows a similar pattern, with affluent neighborhoods generally displaying higher average ratings. An exception to this trend is observed in Galt, where despite having a relatively average price level, the average ratings are notably high. This divergence suggests that factors beyond price alone influence the perceived quality of restaurants in certain areas.

Map of Sacramento Restaurants

Discussion

1. Seasonal Inspection Activity

The analysis of inspections by month unveils intriguing seasonal nuances in inspection patterns. During the colder months, specifically winter and autumn, there is a discernible uptick in inspection activity. This trend aligns with common expectations, as colder weather often brings about increased concerns regarding food safety and hygiene practices. Factors such as temperature control, storage procedures, and handling practices may be more closely scrutinized during these periods, contributing to higher inspection rates.

However, what stands out as particularly noteworthy is the anomaly observed in June and July. Typically considered a warmer summer months, June and July experiences a significant lack in inspection numbers, deviating from the usual monthly trend. This deviation prompts questions regarding potential catalysts for lessened inspection activity during this period. Potential factors could include seasonal events, festivals, or increased tourist activity, all of which may impact food establishments' operations and require less stringent regulatory oversight.

Further exploration into the specific drivers behind the Summer season anomaly could provide valuable insights into dynamic regulatory needs across different seasons and events within Sacramento County.

2. Correlation Between Establishment Ratings and Violations

The scatter plot analysis delving into the relationship between establishment ratings and health code violations uncovers a surprising positive correlation. Contrary to conventional expectations, higher-rated establishments tend to exhibit a higher frequency of health code violations per inspection. This finding challenges the assumption that higher ratings consistently equate to better compliance with health regulations.

Possible explanations for this correlation could stem from various factors. Higher-rated establishments may attract larger customer volumes, leading to increased operational complexities and potential lapses in compliance. Conversely, lower rated restaurants may be under higher scrutiny and are more careful with their health code adherance. Additionally, stringent adherence to health codes may not always be correlated to customer ratings, as ratings often encompass broader aspects such as ambiance, service quality, and menu variety, which may overshadow food safety considerations.

This observation underscores the complexity of interpreting establishment ratings in isolation and highlights the need for multifaceted assessments when gauging food safety and regulatory compliance.

3. Top-Rated Categories and Geographic Variations

The utilization of Bayesian averaging offers valuable insights into the top-rated categories within Sacramento's food landscape. Categories such as Mediterranean, Vegan, and Halal emerge as consistently highly rated, indicating consumer preferences for diverse and healthier culinary options. These findings reflect evolving dietary trends and heightened awareness of healthconscious dining choices among Sacramento residents.

Moreover, the geoplot analysis uncovers notable geographic variations in restaurant quality and compliance. Downtown and affluent areas exhibit higher-rated establishments with fewer violations, indicative of robust food safety practices and regulatory adherence. In contrast, outlying regions demonstrate lower-rated establishments with higher violation frequencies, suggesting potential disparities in regulatory oversight and resource allocation across different neighborhoods.

The observed geographic variations emphasize the importance of targeted interventions tailored to specific regions, considering local demographics, economic factors, and regulatory dynamics.

4. Price Levels and Consumer Perception

The analysis of price levels across zip codes unveils intriguing insights into consumer perception and dining preferences. Higher-income neighborhoods tend to host higher-priced restaurants, reflecting a perceived correlation between price and quality. This association aligns with common consumer behavior, where higher prices often signify premium offerings and elevated dining experiences.

However, the presence of exceptions, such as Galt with relatively high ratings despite average pricing, introduces a layer of complexity to consumer perceptions. Factors beyond price, such as unique menu offerings, exceptional service, or community reputation, may significantly influence how consumers evaluate restaurant quality.

This nuanced understanding of consumer perception underscores the multifaceted nature of dining experiences and the diverse factors that contribute to establishment ratings and success.

5. Data-driven Insights and Decision Making

The comprehensive data scraping, organization, merge, and analysis provides valuable insights into food safety trends, consumer preferences, and regulatory dynamics within the Sacramento food landscape. These insights serve as a foundation for informed decision-making, targeted interventions, and ongoing evaluation of regulatory practices.

By leveraging publicly available data, regulatory agencies and stakeholders can adopt proactive measures to enhance food safety standards, allocate resources effectively, and address localized challenges and opportunities within Sacramento County's diverse culinary ecosystem. Continuous monitoring, analysis, and adaptation based on publicly available, data-driven insights are crucial for maintaining robust food safety protocols, safeguarding public health, and fostering a safer food industry.

https://www.openstreetmap.org/export#map=13/38.5620/-121.4736 too big went to here http://download.geofabrik.de/north-america/us/california/norcal.html