**Yelp Me Eat**

By Kavitha Dhanukodi, Zuhair Bhatti, Debparna Pratiher and Sarah Rahman

**Overview**

For our final project, we chose to analyze the quality of food in restaurants in San Francisco by comparing restaurants in ten districts of San Francisco.

Our project overview is as follows:

**Introduction and Motivation:**Why did we choose to analyze food quality of restaurants? Why did we choose San Francisco as our city for the analysis? Why did we choose particular datasets for our analysis?

**Data acquisition:** We used four data sources to holistically analyze the quality of food in restaurants in San Francisco and arrive at our conclusion. Those are:

**Yelp API:**Ratings & Price - For average price and yelp rating of restaurants in each district.

**Yelp API:** Reviews - For quality of restaurants based on yelp reviews

**iwaspoisoned.com -** For amount of health complaints for each district.

**Health Inspection Rating** - For average health rating of each district.

**Conclusion and Considerations:** What did we learn from our analysis? What conclusion can be drawn from the four data sets that were analyzed? What future steps could be added to provide more in depth results or take this analysis to the next level?

This report goes over the findings of four datasets, the methodology used to analyze the dataset, the decisions we made to arrive at our conclusions and the results from the analysis. We have included our code in order to make it a consolidated notebook and report. Furthermore, each dataset has its own analysis and jupyter notebook included in the main directory of this project that can be run individually.

**Introduction**

In this project, we wanted to assess the quality of restaurants in different parts of a big city and to see if any comparisons could be drawn within the city itself. We chose to analyze San Francisco as it is one of the biggest, most populated cities in America, with many diverse districts with large ranges of population, median household income, ethnicities, etc. We have all visited San Francisco on multiple occasions and one of our favorite things to do in the city is try out different restaurants! However, since none of us are San Francisco natives, we never know where the locals like to eat and which areas are known for better quality food.

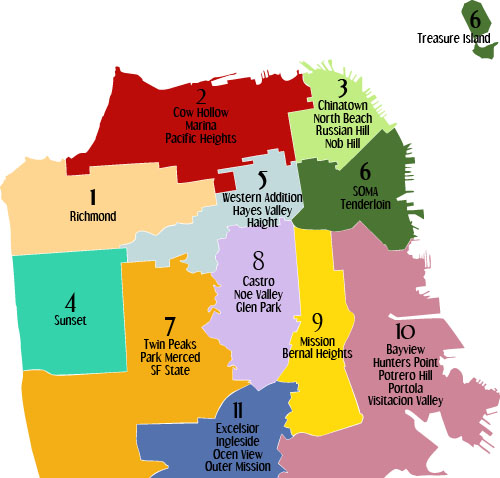
We essentially wanted to see if restaurant quality differed within the same city and whether demographic information of the area could be used to explain certain trends. We also had a few hypotheses we wanted to test during the project. We know there are certain neighborhoods that are generally lower income and less gentrified compared to others which we hypothesized would have lower quality restaurants. The districts we thought would fall under this category were Market and Mission. On the other hand, there are other neighborhoods which are thought to have higher-end average incomes and are more expensive such as Nob Hill and Sunset districts. We hypothesized these districts would lead in terms of high quality restaurants.

The first step was dividing the San Francisco neighborhoods into districts. While there are vague areas split into district, there are not any specific boundaries of districts. We started by collecting all the zipcodes of neighborhoods in San Francisco and then divided them into 10 districts according to area and other conventions used. Since we were comparing districts rather than specific restaurants in each district, we did not check to see the restaurants were similar between districts.

We identified a few critical factors when assessing a restaurant’s quality - food quality, price, health and hygiene. A difficulty we encountered initially in the project is identifying these factors as well as measuring them and figuring out how much weight to give to each. We concluded that these were the most important aspects of quality and looked for relevant data sources to collect this information from.

In order to assess food quality and price, we used the Yelp API for a combination of the Yelp ratings and price ratings. We also did a separate analysis on the Yelp reviews where we analyzed the content of the reviews per district and whether they were largely positive or negative. To assess health and hygiene, we used two different data sources. The first is the official health inspection ratings given to restaurants in each district. The second is an independent website, iwaspoisoned.com, where users report any food poisoning incidents or other unhygienic experiences at restaurants. We used a combination of all this data to compare overall quality of restaurants in different districts.

**Districts and Demographics**



**Part 1: Yelp Ratings & Prices**

**Context**

In an effort to find a better understanding of the holistic breakdown of the qualities of food from district to district, we asked ourselves a very simple question: Do the average prices of resturanst in San Francisco determine the overall rating of the resturants?

To perform this analysis we chose to use Yelp's rating and price measure and compare them to one another in an effort to better understand the relationship of ratints on the price of the food. Originally we were hypothesizing that the better rating a resturant had, the higher stars would follow, since the quality would derive a higher price. To explore this data we turned to Yelp's Fusion API.

**Yelp**

Yelp is an American multinational corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations. The company also trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses, including health inspection scores.

The two features used in this project are :

1) /businesses/price - This returned for each resturant a price measure which was originally formatted in dollars signs ranging from one dollar sign to 4, which was converted 1 equaling one dollar sign up to 4 equaling four dollar signs.

2) /businesses/ratings - This returns the star ratings for individual resturants. Each star in the data frame was written as 1 for one star up to 5 for five stars.

**Methodology**

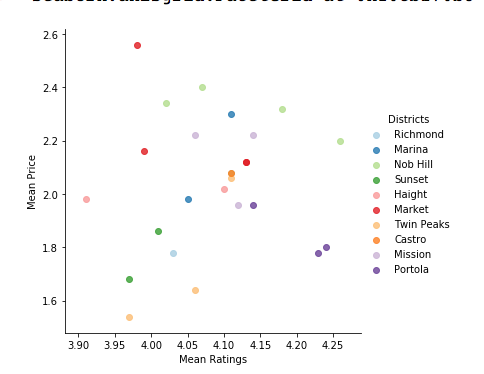
To perform this analysis we accessed Yelp's Fusion API and extracted all the features for the top 50 relevant businesses in a particular zipcode. After extracted this information it was further narrowed down to extracting simply the ratings and price, and appending them all to a list that would eventually be averaged and the function that called the API simply returned the zipcode in question, average ratings, and the average price of resturanrs in the respective zipcode.

Running this function for all zipcodes and obtaining all the zipcodes and related average price/rating these were compiled into a dataframe. The zipcodes were then matched to their appropriate districts (there are 10 districts) and plotted.

The plot we used was a seaborn scatter plot, and the hue was set to districts to show all the zipcodes within a district, while still grouped together. This was done in order to obtain a clearer understanding of the resturant's price and ratings relationships, by adding more data points to be analyzed and avoiding generalizing.

**Results**

Here we used seaborn plotting to plot all the zipcodes paired with their respective districts. Underneath that there is a fitted regression plot and the amount of variance around the fitted line.



**Observations and Comments**

From the two plots above, one can see that there does not seem to be a strong relationship between the average price of resturants in each district with the average ratings. But, this may be due to the outliers such as the Portola district as well as Market, which seem to be completely contridictory to what we have had believed. What this does tells us is that when it comes to ratings there seems to be no strong increase in price as ratings increase. From this we can determine that in order to get better quality food, does not mean that necessariy you have to pay the most either. While this information is not as helpful in determining what factor may lead to better quality food, it can have interesting to dive deeper and find out why this is the way that it is.

Some districts that stood out were the Portola district for having some of the highest reviews, while maintaining some of the lowest prices. While Sunset and Nob Hill had some of the best quality food, while maintaining true to our argument that they would premium prices for this quality food. Taking a step further we conducted a regression these points and found that the variance around line of best fit was suprisingly not too high according to the "rms" of the regression. While there is no strong relationship, there may be something worth investigating further down the road. That would be great to probe into, as well as investigating how socio-economic situations effects

**Part 2: Yelp Reviews**

**Context**

As a way to measure the quality of food in the restaurants of the ten different districts of San Francisco with other metrics we chose to analyze the resturaunt feedback of people in each of these districts. For this analysis, we chose the Yelp Review Dataset to perform a qualitative analysis of the reviews of each district in San Francisco. We chose to analyze the reviews of the restaurants because they collectively provide evidence to what people think of the food, if they are happy with it or if they do not like the place. The reviews from each district were collected using the Yelp Fusion API.

The two endpoints used in this project are :

1) /businesses/{id} - This endpoint returns detailed business content. Normally, you would get the Business ID from /businesses/search, /businesses/search/phone, /transactions/{transaction\_type}/search or /autocomplete. To retrieve review excerpts for a business, please refer to our Reviews endpoint (/businesses/{id}/reviews)

2) /businesses/{id}/reviews - This endpoint returns up to three review excerpts for a given business ordered by Yelp's default sort order

**Methodology**

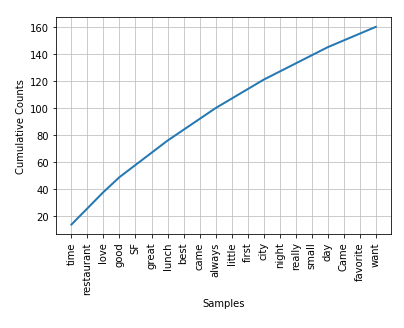
For this analysis, we took Yelp Business API and joined it with the Yelp Reviews API to get the reviews of 50 restaurants of each zip code that falls in that district. Therefore, each district's review analysis is based on 50 resturants of all the Zip codes that fall under it.

The reviews are then put in a dataframe, after some text processing including removing stop words, and supplied to create word clouds. After generating word clouds, the corpus is then used to perform natural language processing using the NLTK library to find the 30 highest frequency words and a frequency distribution of the most used 20 words.

Based on this data, the most frequently used words of each district is used to find the general quality of the restaurants of these districts.

**Results**

**District 1: Market**



**…The rest of the districts are not shown and are included in the jupyter notebook.**

**Observations and Comments**

From these graphs, charts and word clouds for the restaurants of the ten districts of San Francisco we can draw a few conclusions about the quality of food in each of these districts based on the Yelp reviews. I have provided a district by district analysis of the most frequently appearing words in each of 50 restaurants in each district.

The analysis was performed using techniques taught in class. I created frequency distributions and word counts and manually analyzed them. There were some ambiguities as it is not always obvious what the connotations of certain words are. Some words could have either been positive or negative. In terms of deciding if a word is positive or negative, I did research on positive words from reviews and these are some of the words that indicate positivity in reviews[ delicious, great, amazing, excellent, good, best, awesome, perfect, love, friendly, stars, nice, fantastic, tasty, loved, favorite, yummy, bit, wonderful, definitely]. The negative words were words such as [never, bad, unhealthy, poisonous, hazardous, rotten, close, not].

Also it should be noted that only 50 restaurants from each zip code were used for this analysis.

For district market, we can see most common words such as "Favourite", "loooove", "always", "best time ever". Most of the words are positive. There are some negative words such as "nervous", "never" and "crowded".

For district Richmond we can see most common words such as "good", "live", "best", "pretty" and "delicious". Most of the words are positive. There are not as many occurrences of negative words with high frequency or any words that stood out.

For district Haight we can see most common words such as "amazing", "must", "going", "service" and "recommend". Most of the words are positive. There are some negative words such as "last time" which is ambiguous because this could either be last time I came here or last time I am coming here.

For district Twin Peaks had more negative words than positive words with some negative words really standing out such as "hazard", "long" and "never".

For district Castro we can see most common words such as "love" and the rest of the most occurring words were "rude", "fasting" and "wrong.

From this analysis, we can conclude that based on the reviews of each district of San Francisco, Sunset, Nob Hill and Mission stood out as the ones with the districts with the most positively worded reviews. Districts Twin Peaks, Market and Castro had the most negatively worded reviews.

For further and more accurate analysis, I could have implemented NLP techniques such as word tagging and classification of positive and negative words. I could have also combined review data from other sources to account for people other than just yelp users. More data from each district would have provided a better picture.

**Part 3: Health Inspection Data**

**Context**

Now we are going into another aspect of quality which is health but from the perspective of a public health agency instead of a consumer. The Department of Public Health preforms annual health inspections and uses as inspection scoring system to rate each restaurant. The health inspector will calculate a score based on the violations observed. According to sfdph.org, health violations fall into three categories:

High Risk: Violations that directly relate to the transmission of food borne illnesses, the adulteration of food products and the contamination of food-contact surfaces.

Moderate Risk: Violations that are of a moderate risk to the public health and safety.

Low Risk: Violations that are low risk or have no immediate risk to the public health and safety.

The inspections also provided an inspection score for each food establishment. The scores were a number from 0 - 100. The following explains the scores:

Greater than 90: GOOD (Only lower-risk health and safety violations observed. May have high-risk violations.)

Between 89-90: ADEQUATE (Several violations observed. May have high-risk violations.)

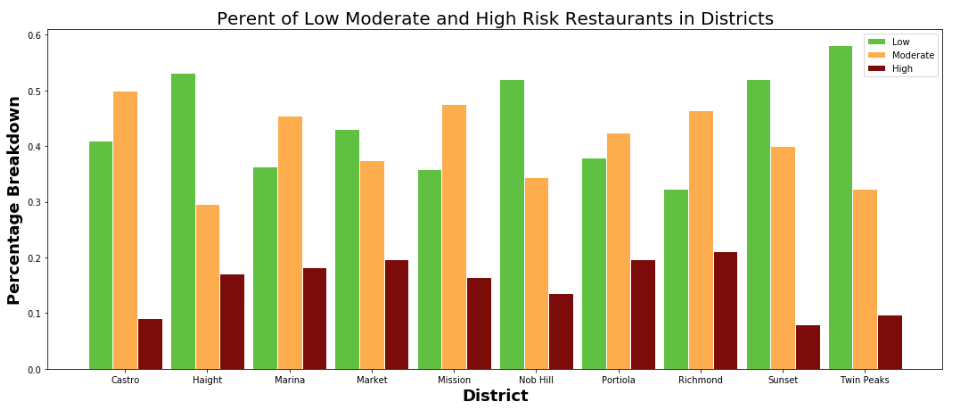
Between 71-85: NEEDS IMPROVEMENT (Multiple violations observed. Typically, several high-risk violations.)

Less than 71: POOR (Multiple violations observed. Typically, several high-risk violations.)

**Methodology**

In terms of assessing health risk, we need to choose a metric to use (like inspection score or the risk\_category) To choose a category , we decided to look at the category that fits the data better, i.e. the category that most resturants had.

We can see that there are 44 more restuarants that have the risk cateory assigned, we we will start off using that as a metric to determine the health of the food at a restauarant.



**Observations and Comments**

Which districts were the cleanest?

The district that had the highest percentage of 'low risk' violations were twin peaks. Looking at the highest percent of low risks is important because it means less moderate and high risks. Furthermore, it has the second lowest moderate risks as well, making it a very 'safe' district to eat at.

When looking at only the high risks however, we can see that the Sunset district and Castro district and Twin Peaks follow as the 3 districts with the lowest high risks respectively. These three districts have populations of 99, 74 and 31 thousand which comparatively are lower population areas. Also it is noteabe that castro is a very small area in San Francisco (only consist of one zip code) while sunset and twin peas have 2 -3 zip code areas.

Which districts weren't the cleanest?

Now the districts had bad health inspection ratings are Richmond and Market district. Almost 22% of food establishments in Richmond fall in the 'high risk category', and compared to all other districts, this category is the highest. Similarly Marlet has around 1 in 5 food establishments rated as 'high risk' places. Recall that high risk category means the restaurant has violations that directly relate to the transmission of food borne illnesses, and the contamination of food-contact surfaces. In other words it might be a good idea to advise consumers to be cautions when eating in the Richmond and Market district.

**Part 4: iwaspoisoned.com**

**Context**

One of the data sources we thought would be relevant to assessing the health aspect of quality of restaurants in districts in San Francisco is [iwaspoisoned.com](https://www.iwaspoisoned.com). Iwaspoisoned.com is a consumer led website for diners to report suspected food poisoning or bad food experiences. The website collects and screens reports of foodborne illness across the United States and the world, and shares that information with public health departments and industry.

The intention behind examining these health reports was to see from which district the most number of negative health experiences were reported. Rather than considering the number of reports on a restaurant-level, we would aggregate all the data of any restaurant in a given district. This could help us compare the general quality of restaurants in each district, specifically the health aspect in this case.

**Methodology**

The metric to compare the different San Francisco districts using the website is the number of reports of restaurants in each district. The content or description of the health report is not important to us since we know the sentiment will be negative if it is reported on the site. This is in contrast to yelp reviews where the content of the review is helpful to our analysis since it can be positive or negative. In the case of iwaspoisoned.com, users only report negative health experiences to the website so an analysis of the content will not produce any significant results. As a result, we are only considering the total number of reports per district to determine the quality of restaurants in each area.The layout of the website was such that not all the reports contained the full address of the restaurant on the homepage. While location is part of the report preview as the restaurant name is required, not all of them listed the zipcodes of the restaurant. The main part of the web scraping takes place in a loop to go through all the pages of reports. Since there was no way to scrape the total number of pages of reports, we realized that when the end of the reports was reached, the website displays "No incidents". So we added a check to our function that gets the report links to see if this string was returned. If it is, then the function returns an empty list and in our main function we maintain a while loop which continues while the list returned is not empty.

In order to keep track of the number of reports per zipcode, we create and populate a dictionary where the keys are the zipcodes and the values are the counts of reports for restaurants in that zipcode.

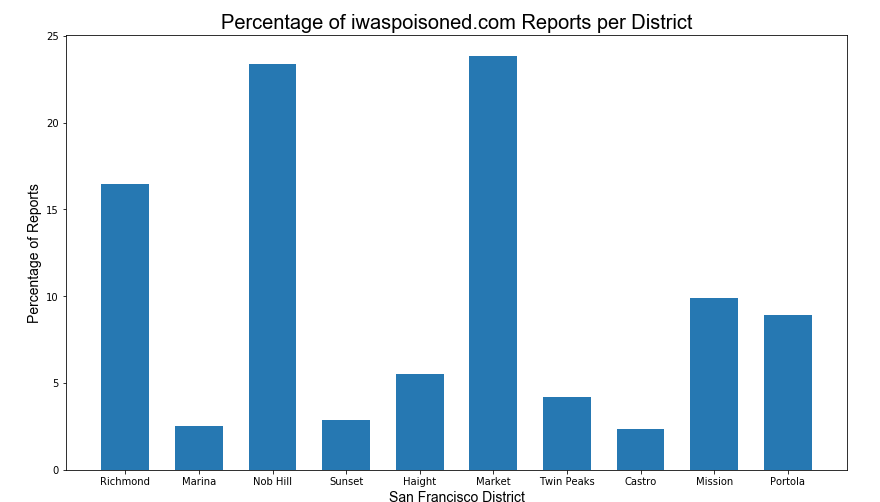
We noticed that not all the zipcodes of the neighborhoods in San Francisco were included in the iwaspoisoned reports. As a result, we listed all the zipcodes we were considering for San Francisco and added those not found in the reports to the dictionary with a value of 0.

In order to get the number of reports per district, we added the counts of the reports of each zipcode in that district. We created a new dictionary with the key as the district and the value as the total count of reports in the district.

During this analysis so far, we did not account for certain factors that might skew the data such as population. In order to make the data more balanced, we analyzed the percentages of reports per district instead of just a raw count of reports. To do this we summed all the reports of all districts to get a total count and then calculated the percentage of each district using this total. We then added this value back to the dictionary to be analyzed.

**Results**

In order to visualize this data, we decided the best method would be to make a histogram of the percentage of reports of each district. We used the final dictionary computed so the x-axis of the plot was the keys which was the district name, and the y-axis was the values which was the percentage of the total reports from restaurants in each district.



**Observations and Comments**

The very nature of the website is to report specific unhygienic or negative experiences at restaurants. As a result, it could be argued that the data might be slightly biased or subjective as it is consumer-driven experiences. However, we accounted for the fact that the reports are all negative as we did not analyze the specific content or reason of each report. Instead, we counted the number of negative experiences of restaurants in any area.

Further, we realized certain areas could have more reports of food poisoning simply for the fact that the area is more populated than others and is more likely to have reports from consumers at its restaurants. In order to account for this, instead of simply comparing the raw counts of reports per district, we analyzed the percentages of reports of the total number of reports in San Francisco. While this may not account for differences population density exactly, it does get rid of some of the possible bias in the data to a certain extent.

Once all the factors had been accounted for, we analyzed the data using a histogram. It can be seen that the districts with the highest percentage of the total food poisoning reports were Market and Nob Hill. Both districts together had a combination of around 50% of the total reports. Market was expected to be one of the districts with the most reports as it includes certain lower income neighborhoods such as Tenderloin and Civic Center. However, we were surprised that Nob Hill, a relatively high-end district, had almost as many reports.

We can see the districts with the least percentages of reports are Marina, Sunset and Castro districts with a combination of less than 10% of the total reports on the site. This result was expected as these districts are all nicer, more hygienic neighborhoods with more residential areas than offices and other buildings.

**Conclusion and Considerations**

We accounted for certain factors that may affect the data is the manner in which we divided the zipcodes into the 10 districts. We revised this division of districts a couple times as initially, we realized certain districts included upto 5 neighborhoods while other districts just had one or two. As expected, these percentage of these districts were much higher compared to others simply because they encompassed more neighborhoods. We then rearranged the neighborhoods to make the districts more fairly divided so the zipcodes were evenly distributed. We also accounted for population differences between districts by considering percentages of data rather than raw values.

From each of our analyses, four in total, we found that there were some districts that seemed to be coming up continuously as the best districts for whatever metrics they were being analyzed for. The biggest winners being: Sunset, Castro, and Nob Hill. These came up at least 2+ times as the top winners in the total four sorts of analyses performed, while the Sunset district came up as a top district for all 4 analyses.

For the Sunset district we could see that it has the most positive number of words in its reviews, lowest levels of complaints from iwaspoisoned.com, as well as the lowest number of high risk and highest number of low risk restaurants from health inspections. Considering that it performed so well in our overall assessment it is surprising to see that when it comes to median households it is the median for all the districts in San Francisco, but it is clear to see that an overwhelming majority of people in this district are younger, and not many people under 18 reside here.

For the Castro district we see that it performed best for ratings to price and low number of iwaspoisoned.com complaints. However, this is probably due to the fact that the Castro district only has one of the smallest populations of 31,124. This falls more in line with our hypothesis as the Castro has the second largest median household income at $114,689. Something to note for the Castro district is that it had one of the lowest populations of all of the districts, have the second least to be exact. This, is something to consider for the iwaspoisoned.com complaints as having such a low population could result in lower amount of complaints as there are simply less people in total in this district.

And for Nob Hill we can find that it performed the best for the overall review sentiment analysis as well as ratings and price comparison. This district has an overall quite low median household income. This explains why it ranks well in the ratings and price comparison. While there is a huge disparity amongst highest performing districts we can take away that the best performing district was the one in between the two extremes in terms of median income. The district that was the median of the median incomes happened to be the best performing for the money, quality, health, and positive review feedback.

With further and deeper investigation it would be better to make links to socio-economic and economic trends to the quality of food, and find deeper underlying links between these factors and how food quality is effected by these factors. This helps to provide a deeper understanding on what the driving factors are for having a district for having higher quality food in order to best model it, rather to make observations on an exploratory level.