Battle of the Neighborhoods
Visiting restaurants in San Francisco during the COVID 19 Pandemic

Matias Garib

July 27th, 2020

1. Introduction

1.1 Background

The COVID-19 pandemic has had a tremendous impact on our way of life. Today, hygiene has become a fundamental aspect in households, businesses and especially public places. Appropriate measures such as washing your hands, using a facemask and practicing social distancing have proved to be the most effective ways of avoiding the spread of the disease. While many day-to-day activities have been affected by the virus, going out to dinner is for many the most sorely missed thing to do. As restaurants start to re-open, families, foodies, couples and anyone going out for a meal will have to choose wisely where to go in order to avoid getting infected. Although hygiene has always been of utmost importance in restaurants, this matter is now even more relevant, and it would be very useful for people to know which places have the highest and lowest hygiene standards when choosing where to go.

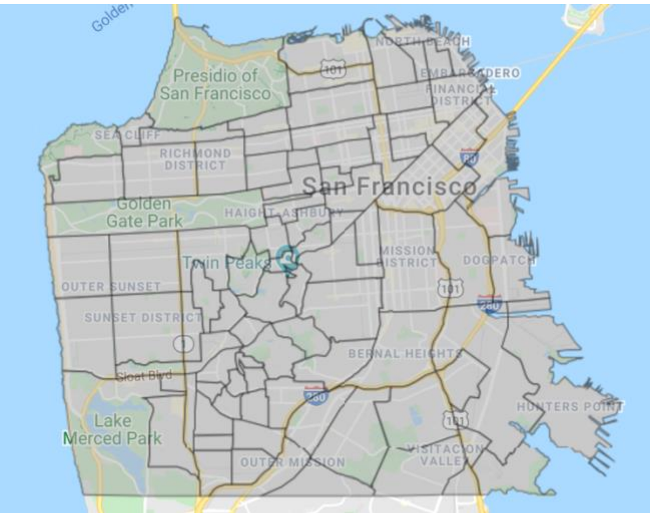
1.2 Problem

People want to start going out and visiting restaurants, but they want to visit places with the best hygiene practices. The questions we want to answer, for the city of San Francisco, are: which are the cleanest restaurants in each neighborhood? Which are the safest neighborhoods to go out to eat?

2. Data

2.1 Data Sources

The first dataset to be used consists of a GeoJSON file with the names and boundaries of 92 San Francisco neighborhoods. It is based on the real estate data and can be accessed here (DataSF, 2019). A view of the map is shown below.



Secondly, Foursquare APIs will be used to explore each neighborhood's restaurants. Foursquare contains massive datasets of accurate location data and works as social media as well. The main API that will be used is the Explore API, which returns popular spots around a certain location. The foursquare API Venue Search "Returns a list of venues near the current location, optionally matching a search term" (Foursquare Documentation). In our case, the current location used will not be an exact location, instead, the parameter "near" can be used to locate venues around a neighborhood. The URL used is shown below.

```
'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&near={},San Francisco, CA&categoryId={}&radius={}&limit={}'
```

The result of running the Foursquare API on all the neighborhoods results in 6.121 venues returned with data on the venue name, latitude, longitude, category, and ID.

Finally, the set of data used to perform the proposed restaurant hygiene analysis is provided by the City of San Francisco and is the result of the Health Department's inspection program. The data was last updated July 27th, 2020 and can be accessed here. Some of the relevant features include:

- business_name: Common name of the business.
- business_address: Street address of the business
- inspection_date: Date of the inspection in YYYYMMDD format
- inspection_score: Inspection score on a 0-100 scale. 100 is the highest score.
- inspection_type: String representing the type of inspection. Must be one of: initial, routine, followup, complaint
- violation_description: One-line description of the violation. 200-character max.
- risk_category: Low, medium or high risk depending on inspection score.

The dataset contains data for 53.973 inspections between 2016 and 2020, from which many are follow ups to each restaurant.

2.2 Data Cleaning and Feature Selection

The approach taken towards Data Cleaning was first to select the important features for each of the two relevant data sets in hand (Foursquare restaurants and Hygiene inspections), clean them and select features separately, and then merge them into one "master" data frame. The steps are described in the following sections.

2.2.1 Feature Selection

Starting with the Foursquare API, only the following data was kept:

- Venue (name)
- Venue_category
- Venue_latitude
- Venue_longitude

There wasn’t a lot more to choose from, the only dropped column was “Venue ID”.

On the other hand, for the hygiene data frame, the selection was broader. The following columns were selected based on their relevance in answering the main question, which neighborhood has the best hygiene practices. The kept features were:

- Business_name
- Business_Address
- Inspection_Date
- Inspection_Type
- Violation_Description
- Risk_Category
- Inspection_Score

Although at first the intention was to use all of these columns, most of the analysis ended up relying on the last two columns, Risk Category and Inspection Score. The dropped columns were:

- Venue_ID
- City
- State
- Postal_Code
- Inspection_ID
- Phone_Number
- Location (already in the Foursquare API).

2.2.2 Data Cleaning

Cleaning up rows was also a fairly straightforward step and was only required for the Hygiene data set. As mentioned before, the hygiene data frame consisted of 53.973 inspections which were reduced to approximately 10% after the following three step cleanup:

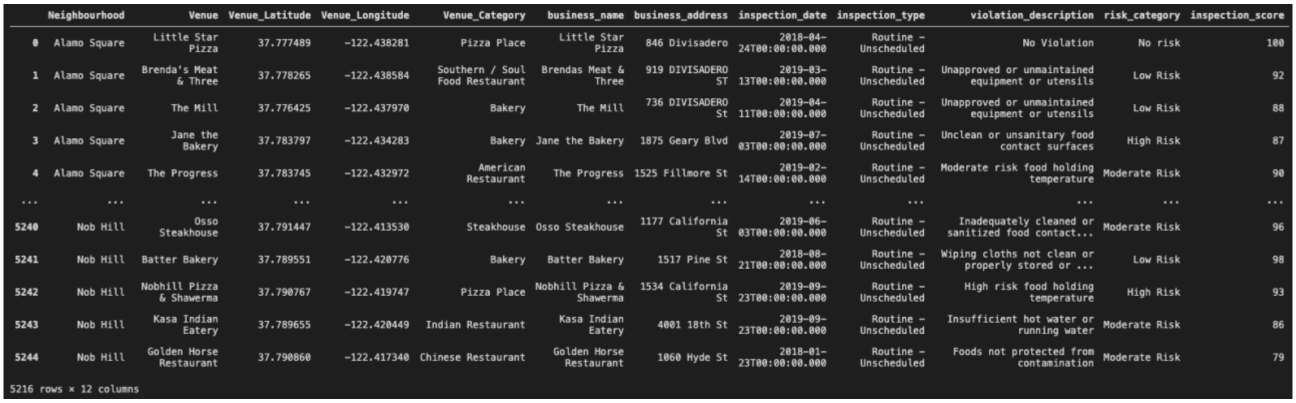
1. Approximately 12.000 rows with inspection_score = ‘NaN’ values were dropped.
2. All the data with inspection_score = 100, perfect score, has risk_category value equal to ‘Nan’. So, a new category was included and named ‘No Risk’.
3. The data was ordered by Inspection date and only the most recent inspection was considered for restaurants that were inspected more than once (some were inspected up to 6 times). This was done so that only the most updated hygiene practices were considered. Approximately 34.000 rows where dropped.

The resulting data consisted of the most recent inspection information for 5.277 restaurants. With the Foursquare API returning approximately 6.000 venues, it is fair to say there is a big overlap between the restaurants returned by Foursquare and the inspected venues.

2.2.3 Merging Data Frames

While feature selection and data cleaning were fairly straightforward, merging the data frames was not. The main issue faced here was that the merging process relied upon the venue names. However, between the hygiene data from SF and the venues data from Foursquare, there were various discrepancies in names, spelling, symbols (and v/s &) as well as caps. To solve this problem the fuzzy_merge function, from Fuzzy Pandas, was used. This library recognizes similar, but not equal strings and lets you connect strings according to different methods. More information about the library can be found here.

After feature selection, data cleaning and merging, the final data frame containing all the relevant data for the analysis looks like this:



	Neighbourhood	Venue	Venue_Latitude	Venue_Longitude	Venue_Category	business_name	business_address	inspection_date	inspection_type	violation_description	risk_category	inspection_score
0	Alamo Square	Little Star Pizza	37.777489	-122.438281	Pizza Place	Little Star Pizza	846 Divisadero	2018-04-24T00:00:00.000	Routine - Unscheduled	No Violation	No risk	100
1	Alamo Square	Brenda's Meat & Three	37.778265	-122.438584	Southern / Soul Food Restaurant	Brendas Meat & Three	919 DIVISADERO ST	2019-03-13T00:00:00.000	Routine - Unscheduled	Unapproved or unmaintained equipment or utensils	Low Risk	92
2	Alamo Square	The Mill	37.776425	-122.437970	Bakery	The Mill	736 DIVISADERO St	2019-04-11T00:00:00.000	Routine - Unscheduled	Unapproved or unmaintained equipment or utensils	Low Risk	88
3	Alamo Square	Jane the Bakery	37.783797	-122.434283	Bakery	Jane the Bakery	1875 Geary Blvd	2019-07-03T00:00:00.000	Routine - Unscheduled	Unclean or unsanitary food contact surfaces	High Risk	87
4	Alamo Square	The Progress	37.783745	-122.432972	American Restaurant	The Progress	1525 Fillmore St	2019-02-14T00:00:00.000	Routine - Unscheduled	Moderate risk food holding temperature	Moderate Risk	90
...	...	...	...	...	...	...	...	...	...	...	...	...
5240	Nob Hill	Osso Steakhouse	37.791447	-122.413530	Steakhouse	Osso Steakhouse	1177 California St	2019-06-03T00:00:00.000	Routine - Unscheduled	Inadequately cleaned or sanitized food contact...	Moderate Risk	96
5241	Nob Hill	Batter Bakery	37.789551	-122.420776	Bakery	Batter Bakery	1517 Pine St	2018-08-21T00:00:00.000	Routine - Unscheduled	Wiping cloths not clean or properly stored or ...	Low Risk	98
5242	Nob Hill	Nobhill Pizza & Shawerma	37.790767	-122.419747	Pizza Place	Nobhill Pizza & Shawerma	1534 California St	2019-09-23T00:00:00.000	Routine - Unscheduled	High risk food holding temperature	High Risk	93
5243	Nob Hill	Kasa Indian Eatery	37.789655	-122.420449	Indian Restaurant	Kasa Indian Eatery	4001 18th St	2019-09-23T00:00:00.000	Routine - Unscheduled	Insufficient hot water or running water	Moderate Risk	86
5244	Nob Hill	Golden Horse Restaurant	37.790860	-122.417340	Chinese Restaurant	Golden Horse Restaurant	1060 Hyde St	2018-01-23T00:00:00.000	Routine - Unscheduled	Foods not protected from contamination	Moderate Risk	79

5216 rows × 12 columns

***Figure 1**. 'Master' data frame.*

And it can be mapped out as follows:

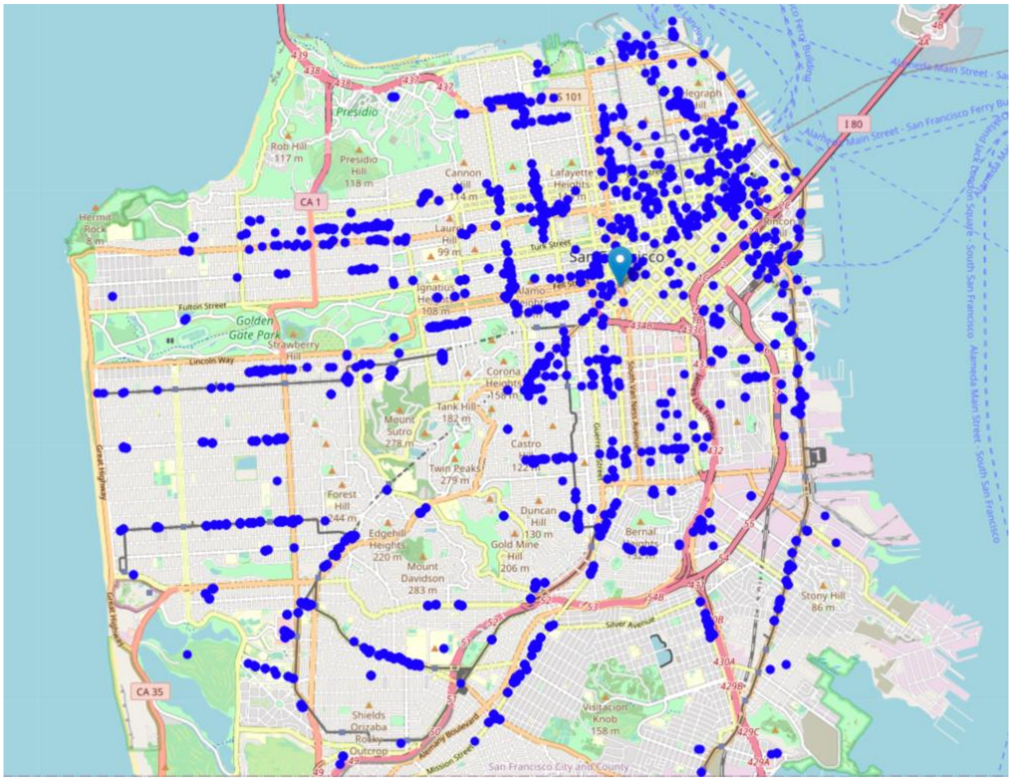


***Figure 2**. Location of analyzed restaurants.*

3. Methodology

Keeping in mind that the project's main goal is to classify each of San Francisco's neighborhoods according to their restaurant's infection risk, the following step by step approach will be executed:

1. *Geographic Visualizations*: We first want to understand the big picture and the best way to do this is without any numbers. The output from this first part will aim to locate where the best/worst restaurant zones are using both heat a choropleth maps. If there are clearly defined zones of better performance, this analysis should give us those hints.

2. Clustering and Analysis: This second part will mainly focus on separating the good/bad neighborhoods and understanding how, numerically, they differ. Two clusters will be used as we expect to get "GO" and "DON'T GO" groups of neighborhoods. The analysis will then focus on the following criteria to choose which cluster is best:

    a. Average inspection score value
    b. Average number of restaurants per neighborhood
    c. Percentage of No Risk restaurants
    d. Percentage of Low Risk restaurants
    e. Percentage of Moderate Risk restaurants
    f. Percentage of High-Risk restaurants

As a result, both analyses will hopefully give us hints on which areas of the city have higher hygienic standards than others and secondly, which specific neighborhoods have the best and worst inspection reviews. This should assist people in choosing where to go out to eat when in San Francisco, if they want to avoid Coronavirus.

4. Results and Discussion

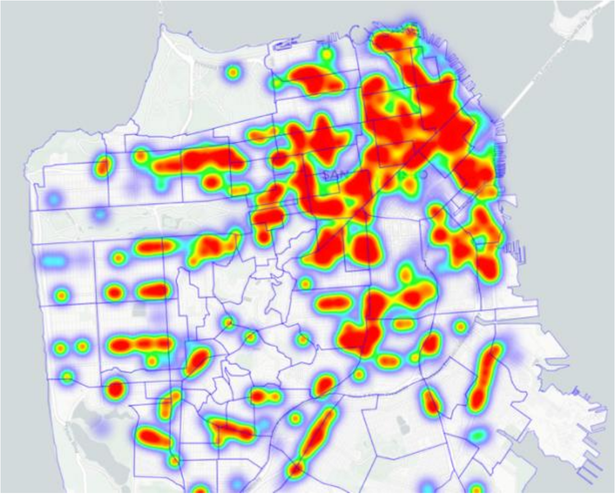
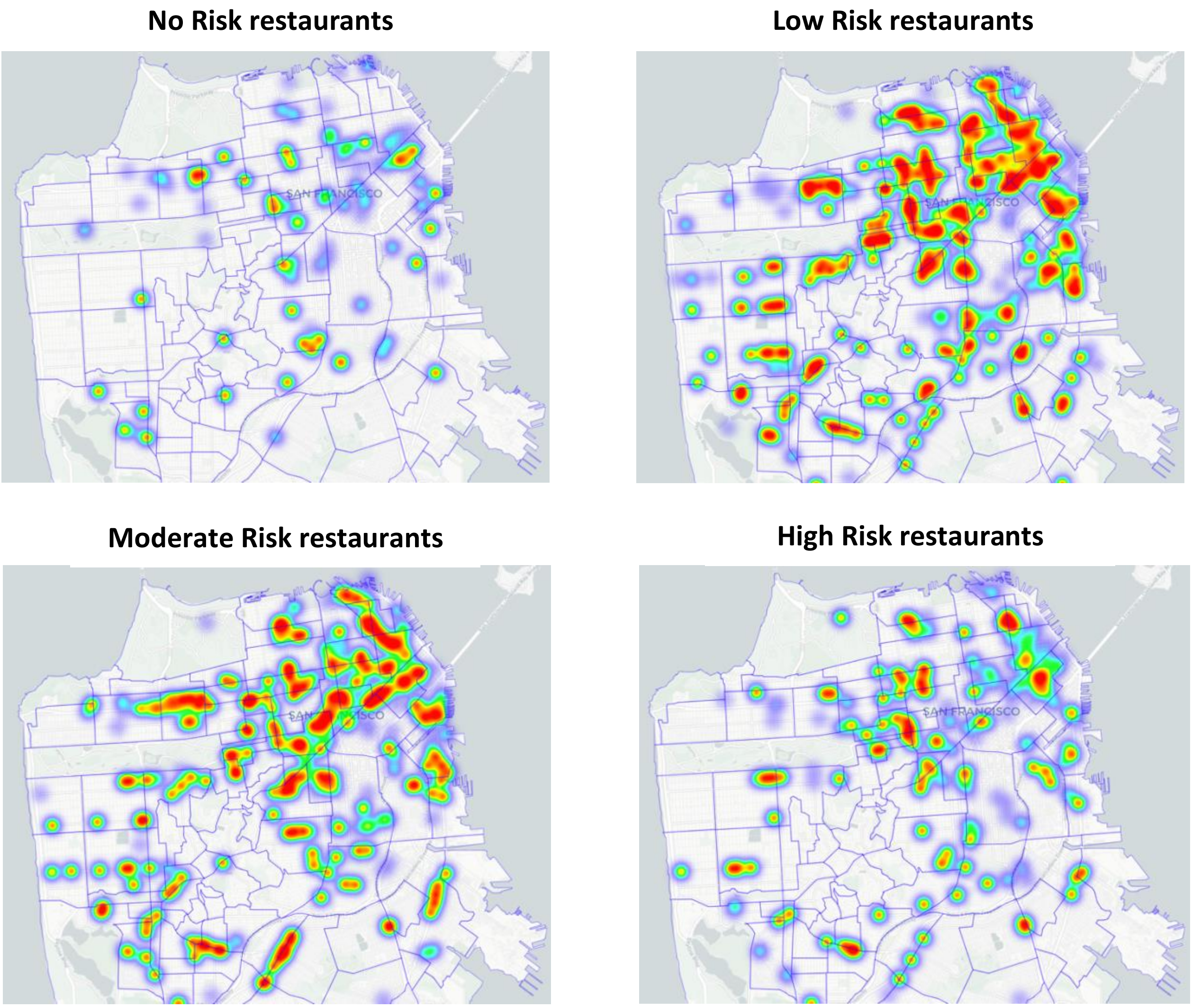


Figure 3. *Restaurants Heat Map in SF*



No Risk restaurants

Low Risk restaurants

Moderate Risk restaurants

High Risk restaurants

Figure 4. *Risk Categories Heat Maps*

Figure 3 results immediately tell us that, in terms of numbers, restaurants are mostly located towards the NE side of SF, with highest concentrations towards Nob Hill and the Financial District. We can also notice large patches of land with no restaurant presence. At a closer look, we notice that large park areas have little or no restaurants within or around. The most noticeable of these patches are the Golden Gate Park, Presidio of SF and Lake Merced Park.

Figure 4 then provide us a good sense of how No, Low, Moderate and High-Risk restaurants are distributed. At first glance, one can see that most restaurants are categorized either as Low or Moderate risk restaurants. The No risk and High-risk restaurants, i.e. the extremes, are fairly small in numbers. What’s most concerning is that the No Risk category has less coverage compared to the others. In terms of distribution, it is fairly even, with no areas having noticeably better results than the others.

The big picture tells us that wherever we go in SF, there will be a wide variety of restaurants to choose from and it’s hard to tell which zone is better than the other. There’s everything everywhere. This images also tell us that most restaurants in San Francisco present either Low or Moderate risk of infection.

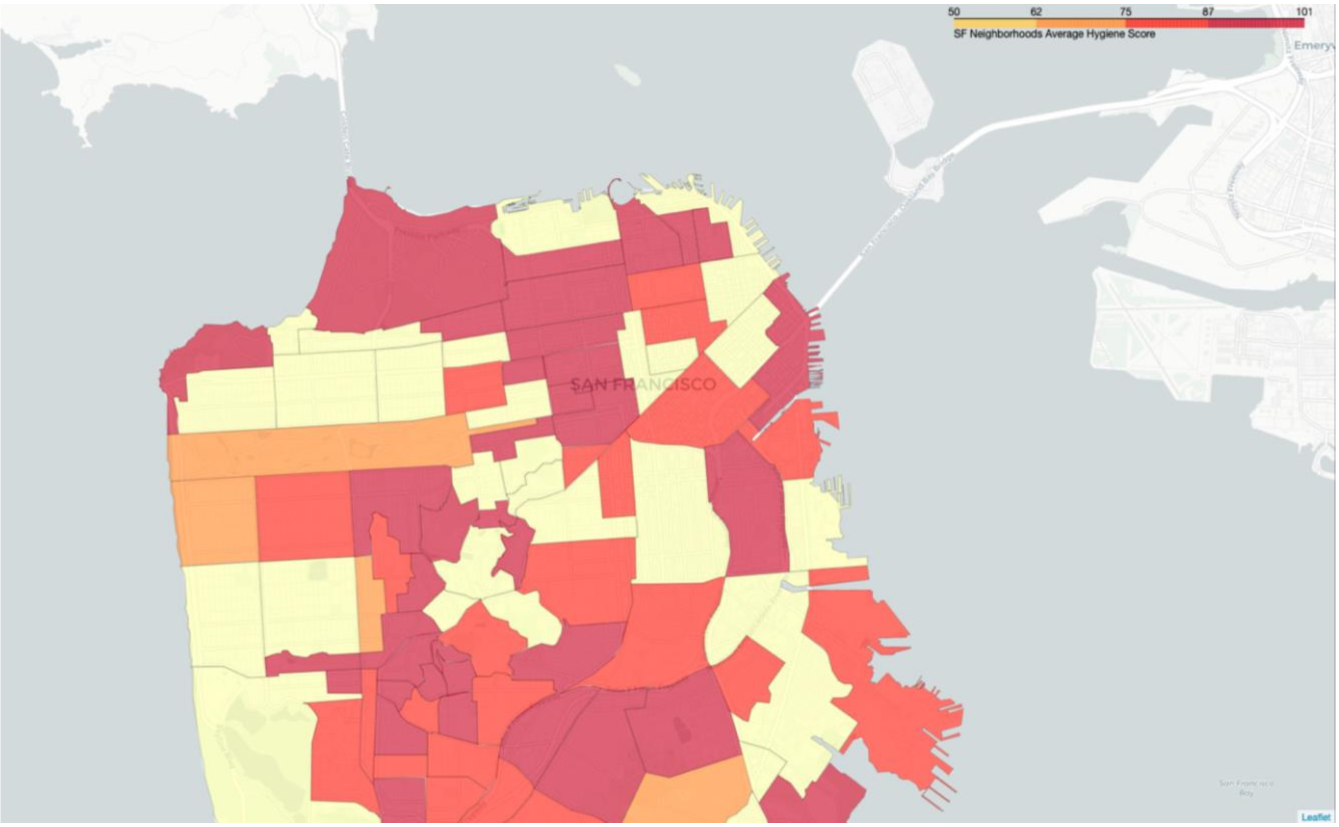


Figure 5. *Choropleth map of average inspection scores per neighborhood*

Note: Yellow neighborhoods correspond to those which the Foursquare API couldn’t get any data. However, there are restaurants considered in all of those neighborhoods, as the radius of the search was set to include them.

The choropleth map above paints a similar picture to what was mentioned before. When looking at average inspection scores for the analyzed neighborhoods (Orange to Red), we see that the distribution is fairly even. At a closer look, it’s evident that the darker red neighborhoods tend to be located NE, where the majority of restaurants are located. As one moves S-SW, lighter reds and orange start to appear, suggesting that areas with less restaurants also have worst average inspection scores.

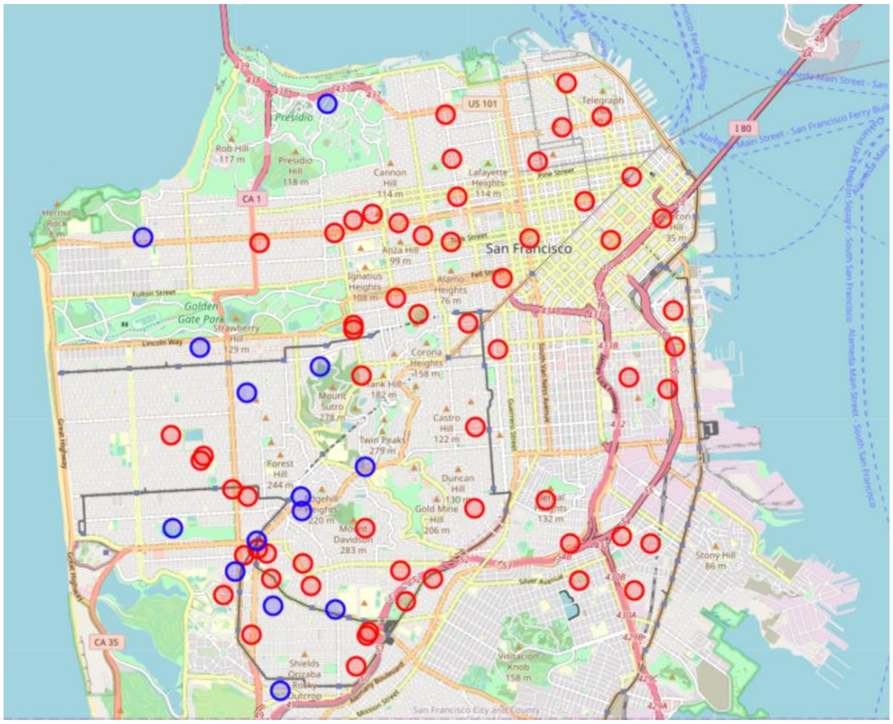


Figure 6. *Neighborhoods cluster analysis.*

When clustering the restaurants, there is a similar tendency of high restaurant density areas NE and lower or mixed areas S-SW. From that point of view, and comparing this result with the heat maps, it is reasonable to believe that the clustering analysis provided two distinct types of neighborhoods: high restaurant density areas and lower restaurant density areas. This visualization, however, doesn’t add anything to determine which cluster is better. Nevertheless, Figure 5 suggested lower restaurant density neighborhoods could have lower inspection scores. Figure 7 can provide insights by comparing each cluster inspection scores.

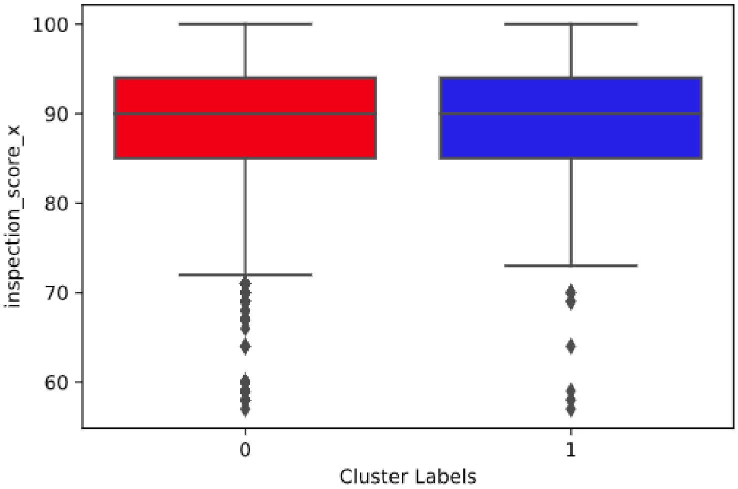


Figure 7*. Box plots for cluster's inspection scores*

Interestingly enough, both box plots are almost exactly the same, meaning each cluster's neighborhoods are equally distributed in terms of inspection scores. Because inspection scores don't tell us much about how to differentiate the clusters, the following table sums up other metrics comparing the red and blue clusters.

Table 1. *Cluster comparison*

Metric	Red Cluster (0)	Blue Cluster (1)
# of Neighborhoods	59	14
Average # Restaurants per Neighborhood	82	28
Average Inspection Scores	89	89
Max Inspection Score	100	100
Min Inspection Score	57	57
No Risk restaurants (%)	6%	4%
Low Risk restaurants (%)	43%	41%
Moderate Risk restaurants (%)	37%	45%
High Risk restaurants (%)	14%	10%

Table 1 provides us some more information to analyze. First of all, both clusters are extremely similar regarding inspection scores. As mentioned before, both have similar means, but also maximum and minimum values. This tells us that inspection scores are not a good metric to compare each cluster. Then how are the clusters different?

The clusters differ in two major things, first of all in the average number of restaurants per neighborhood. As seen before in the heat maps, there is a clear distinction between neighborhoods highly populated with restaurants and those with less options, usually closer to parks. The other distinction can be made when observing category percentages. We see how the red cluster has higher percentages of No Risk and Low Risk restaurants and lower percentages of Moderate Risk and High-Risk restaurants. Although not significant, it is a 4% difference in the top two and bottom two categories worth noticing. This would then suggest that higher density of restaurants results in more reliable places.

5. Conclusion

The analysis gives us the following conclusions:

Most restaurants in SF are located towards the NE part of the city.
The financial district and its surroundings have the highest restaurants concentrations. This concentration is reduced as one move S-SW to residential areas, usually near parks.

Most restaurants are considered to have either Low or Moderate Risk of infection.
When going out to eat in SF you should expect mostly Low to Moderate risks of catching COVID19. It's also concerning that there are more High-risk restaurants than No Risk restaurants.

When looking at the city as a whole, there is no clear distinction between good and bad areas, you can find everything everywhere
Probably the most relevant conclusion. In which neighborhood a restaurant is located is not a good indicator of how clean it is.

A better variable to determine Low or No risk of contamination is how restaurant-dense the neighborhood is.
You can expect high restaurant density neighborhood to be less risky than lower restaurant density neighborhoods.

When choosing where to go out to dinner, choose the neighborhoods in the red cluster.
Following the last conclusion, neighborhoods with more restaurants around will probably have better hygiene standards.

6. Future Considerations

For future considerations it would be great to have all neighborhoods covered in the Foursquare API. Also, comparing restaurant categories (i.e Chinese, American, fast food, etc.), their inspection scores and risk categories could be very insightful. Finally, using the Foursquare premium API, reviews could be scrapped to find any more suggestions of high/low hygiene standards.