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.

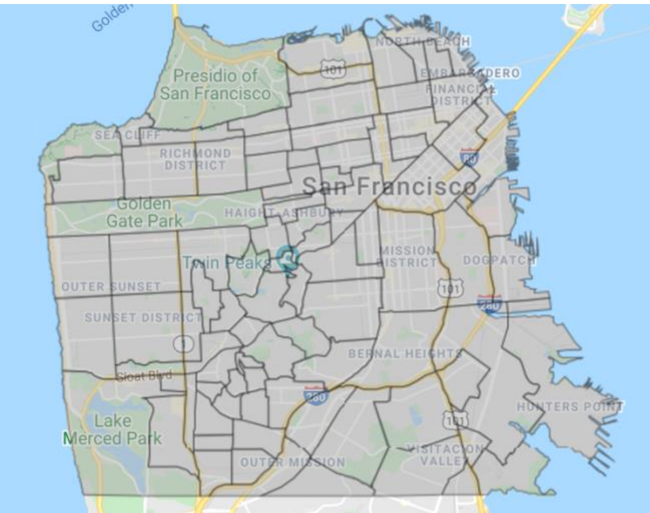
1.2 Problem

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 2010 census 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 and the input will be each neighborhood in SF taken from the above-mentioned data base. 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 data used 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 inspection.

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 as follows.

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% of the rows after the following two step cleanup:

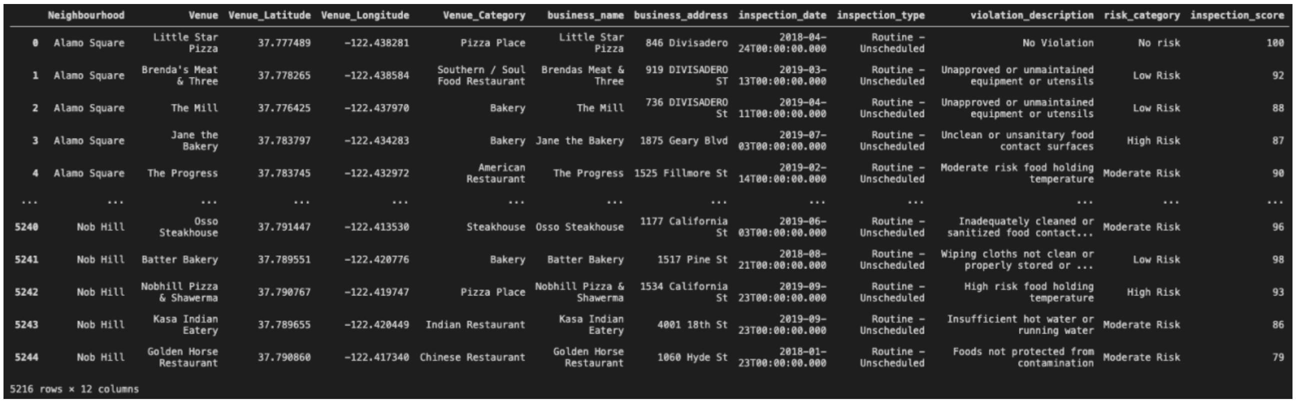
1. Approximately 12.000 rows with inspection_score = ‘NaN’ values were dropped.
2. All the data with inspection_score = 100, perfect score, has risk_category value equal to ‘Nan’. So, a new category was included and named ‘No Risk’.
3. The data was ordered by Inspection date and only the most recent 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 was fair there was a big overlap between the restaurants returned by Foursquare and the inspected venues.

3.2.3 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 was used, from Fuzzy Pandas. 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 the whole feature selection, data cleaning and merging, 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:

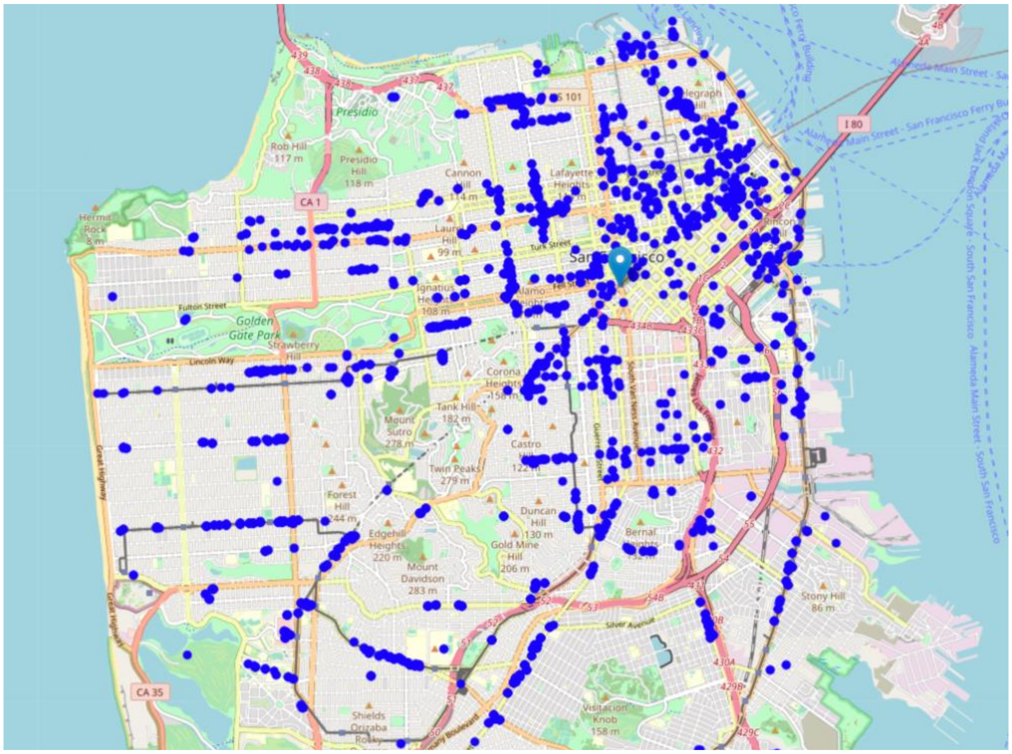


Figure 2. Location of analyzed restaurants.