What's cooking!

A DATA DRIVEN STUDY

Turjo Mukherjee | Battle of Neighbourhoods | September 12, 2019

Introduction

As part of continuous improvement of the overall quality of life in the city of San Francisco, it is important for the City Council and the Mayor's office to manage high standards when it comes to restaurants. From ensuring that they are profitable to maintaining proper standards, from well-known cuisines to new ones, it is important to acknowledge that restaurants are a big part of San Francisco lifestyle. This has a direct impact to the lifestyle and enjoyment of the residents of the city as well as thousands of visitors that come to San Francisco to enjoy everything that this city has to offer. San Francisco has maintained a diverse and culturally rich reputation among all the major cities in the United States of America. From the proximity to Silicon Valley to the arts to the culinary experiences, Golden Gate Bridge to Senoma valley, coffee shops to restaurants – San Francisco has so much to offer to the people.

The City Council understands the importance of this and therefore it is important to ensure the safety standards of its restaurants and improve the reputation of the city.

Business Problem

In the upcoming 2020 financial budget, the City Council wants to allocate funds to set up **3 new food hubs in the city. These would feature restaurants with 3 different cuisines**.

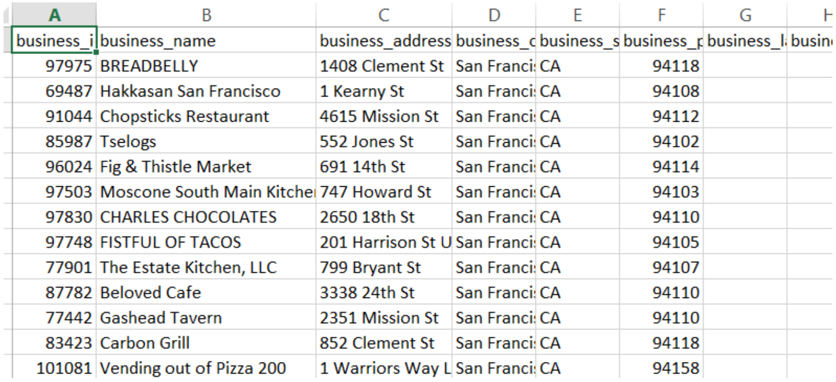
Additionally, the City wants to propose the shutting down of restaurants in the same neighbourhoods where the new hubs will be proposed as replacements. The aim of this study is to evaluate the current status of restaurants and their quality results. From there we need to identify the best neighbourhoods that should get the new food hubs and 3 cuisines for each. The study should give a full breakdown of the restaurants' data, provide clarity in identifying the right cuisine, recommend based on data the best neighbourhoods to have these new food hubs.

Required Data

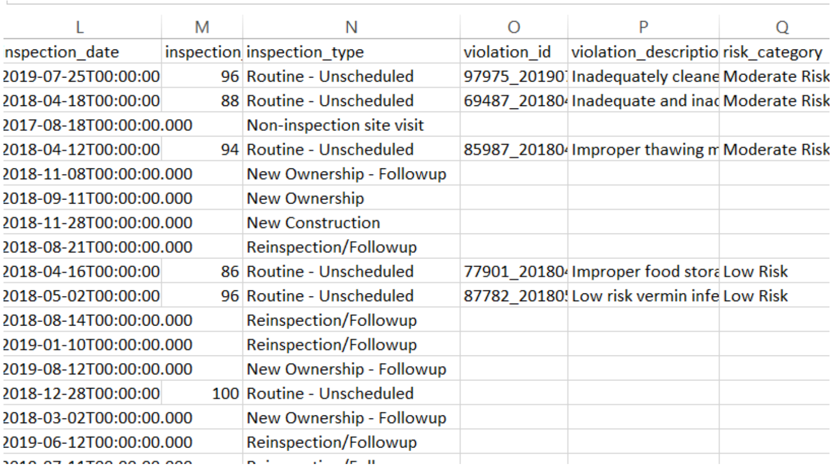
In order to make this analysis comprehensive, we are looking at available data of restaurants in SFO, retrieved from this website.

https://www.kaggle.com/san-francisco/sf-restaurant-scores-lives-standard

An extract of the data from the above source is shown below:



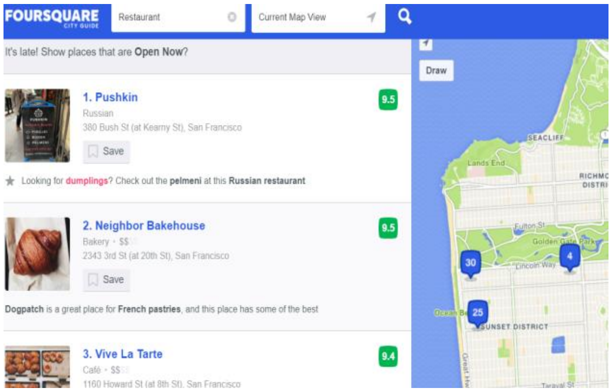
A	B	C	D	E	F	G	H
business_i	business_name	business_address	business_c	business_s	business_p	business_l	busin
97975	BREADBELLY	1408 Clement St	San Franci	CA	94118		
69487	Hakkasan San Francisco	1 Kearny St	San Franci	CA	94108		
91044	Chopsticks Restaurant	4615 Mission St	San Franci	CA	94112		
85987	Tselogs	552 Jones St	San Franci	CA	94102		
96024	Fig & Thistle Market	691 14th St	San Franci	CA	94114		
97503	Moscone South Main Kitche	747 Howard St	San Franci	CA	94103		
97830	CHARLES CHOCOLATES	2650 18th St	San Franci	CA	94110		
97748	FISTFUL OF TACOS	201 Harrison St U	San Franci	CA	94105		
77901	The Estate Kitchen, LLC	799 Bryant St	San Franci	CA	94107		
87782	Beloved Cafe	3338 24th St	San Franci	CA	94110		
77442	Gashead Tavern	2351 Mission St	San Franci	CA	94110		
83423	Carbon Grill	852 Clement St	San Franci	CA	94118		
101081	Vending out of Pizza 200	1 Warriors Way L	San Franci	CA	94158		



L	M	N	O	P	Q
nspection_date	inspection	inspection_type	violation_id	violation_descriptio	risk_category
2019-07-25T00:00:00	96	Routine - Unscheduled	97975_20190	Inadequately cleane	Moderate Risk
2018-04-18T00:00:00	88	Routine - Unscheduled	69487_20180	Inadequate and inac	Moderate Risk
2017-08-18T00:00:00.000		Non-inspection site visit			
2018-04-12T00:00:00	94	Routine - Unscheduled	85987_20180	Improper thawing m	Moderate Risk
2018-11-08T00:00:00.000		New Ownership - Followup			
2018-09-11T00:00:00.000		New Ownership			
2018-11-28T00:00:00.000		New Construction			
2018-08-21T00:00:00.000		Reinspection/Followup			
2018-04-16T00:00:00	86	Routine - Unscheduled	77901_20180	Improper food stora	Low Risk
2018-05-02T00:00:00	96	Routine - Unscheduled	87782_20180	Low risk vermin infe	Low Risk
2018-08-14T00:00:00.000		Reinspection/Followup			
2019-01-10T00:00:00.000		Reinspection/Followup			
2019-08-12T00:00:00.000		New Ownership - Followup			
2018-12-28T00:00:00	100	Routine - Unscheduled			
2018-03-02T00:00:00.000		New Ownership - Followup			
2019-06-12T00:00:00.000		Reinspection/Followup			

Based on this data, the idea is to do an analysis of the neighbourhoods, type of restuarants, number of quality violations based on cuisine in the neighbourhoods.

Foursquare Data will be used to get user reviews of the restaurants as well to get a balanced view of both Quality Inspectors and the general public.



It's late! Show places that are **Open Now**?

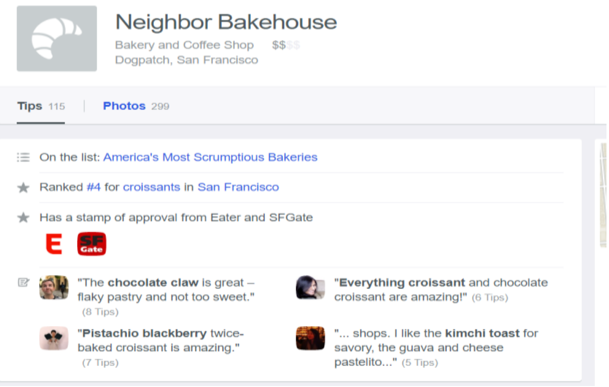
1. Pushkin — 9.5
Russian
380 Bush St (at Kearny St), San Francisco
Save

★ Looking for **dumplings**? Check out the **pelmeni** at this **Russian restaurant**

2. Neighbor Bakehouse — 9.5
Bakery · $$
2343 3rd St (at 20th St), San Francisco
Save

Dogpatch is a great place for **French pastries**, and this place has some of the best

3. Vive La Tarte — 9.4
Café · $$
1160 Howard St (at 8th St), San Francisco



Neighbor Bakehouse

Bakery and Coffee Shop $$$$

Dogpatch, San Francisco

Tips 115 | Photos 299

On the list: America's Most Scrumptious Bakeries

Ranked #4 for croissants in San Francisco

Has a stamp of approval from Eater and SFGate

"The **chocolate claw** is great – flaky pastry and not too sweet." (8 Tips)

"**Everything croissant** and chocolate croissant are amazing!" (6 Tips)

"**Pistachio blackberry** twice-baked croissant is amazing." (7 Tips)

"... shops. I like the **kimchi toast** for savory, the guava and cheese pastelito..." (5 Tips)

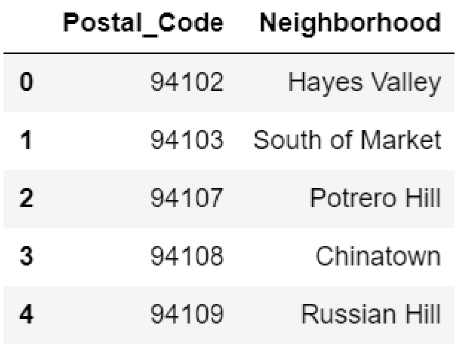
We need to give a balanced recommendation factoring user experience as well as inspection scores to provide the final recommendation on the neighbourhoods as well as cuisines for the food hubs that would **replace** existing restaurants.

UPDATED (3rd Oct) – Based on a forum discussion with the course instructor, I have shared the limitation of getting Foursquare data based on the limited daily API calls restricted. Therefore NO Foursquare data has been used for this study.

I have also referred the GeoJSON file for San Francisco neighbourhoods to add another layer of visualization for the neighbourhoods as shown later below.

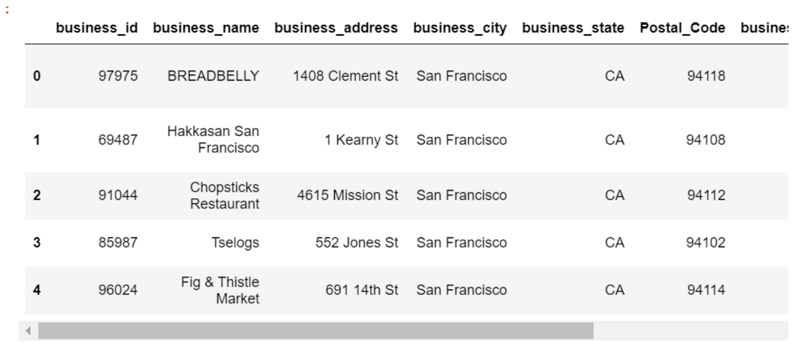
Methodology

To begin with, there was an extensive data preparation and cleaning exercise to achieve the intended objective. The dataset from Kaggle for restaurant inspection details contained street address and postcode but didn't have an obvious link from postcode to neighbourhood. This was prepared using the online data available at www.healthysf.org. It resulted in a csv file as shown below in a snapshot.

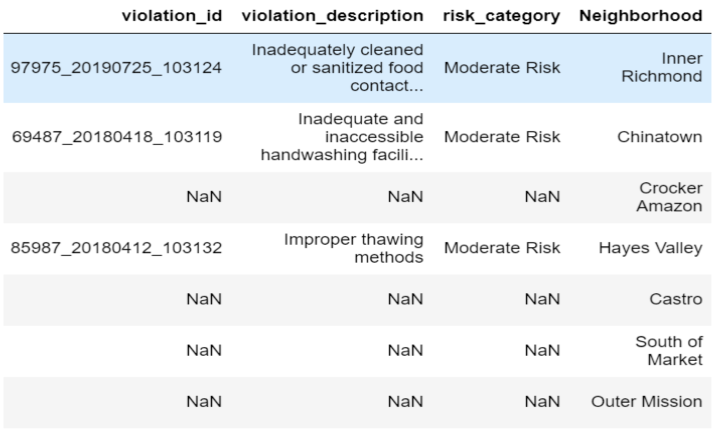


	Postal_Code	Neighborhood
0	94102	Hayes Valley
1	94103	South of Market
2	94107	Potrero Hill
3	94108	Chinatown
4	94109	Russian Hill

Based on the postcode data in the inspection data and this new file, each entry was associated with a neighbourhood. That resulted in a data set shown below in a snapshot.



	business_id	business_name	business_address	business_city	business_state	Postal_Code	busine
0	97975	BREADBELLY	1408 Clement St	San Francisco	CA	94118	
1	69487	Hakkasan San Francisco	1 Kearny St	San Francisco	CA	94108	
2	91044	Chopsticks Restaurant	4615 Mission St	San Francisco	CA	94112	
3	85987	Tselogs	552 Jones St	San Francisco	CA	94102	
4	96024	Fig & Thistle Market	691 14th St	San Francisco	CA	94114	



violation_id	violation_description	risk_category	Neighborhood
97975_20190725_103124	Inadequately cleaned or sanitized food contact...	Moderate Risk	Inner Richmond
69487_20180418_103119	Inadequate and inaccessible handwashing facili...	Moderate Risk	Chinatown
NaN	NaN	NaN	Crocker Amazon
85987_20180412_103132	Improper thawing methods	Moderate Risk	Hayes Valley
NaN	NaN	NaN	Castro
NaN	NaN	NaN	South of Market
NaN	NaN	NaN	Outer Mission

By dropping some columns and removing the rows with no data, we had a clean dataset of all inspection records for restaurants in San Francisco and also associated with neighbourhoods.

There were a few assumptions made in this step that are important to highlight:

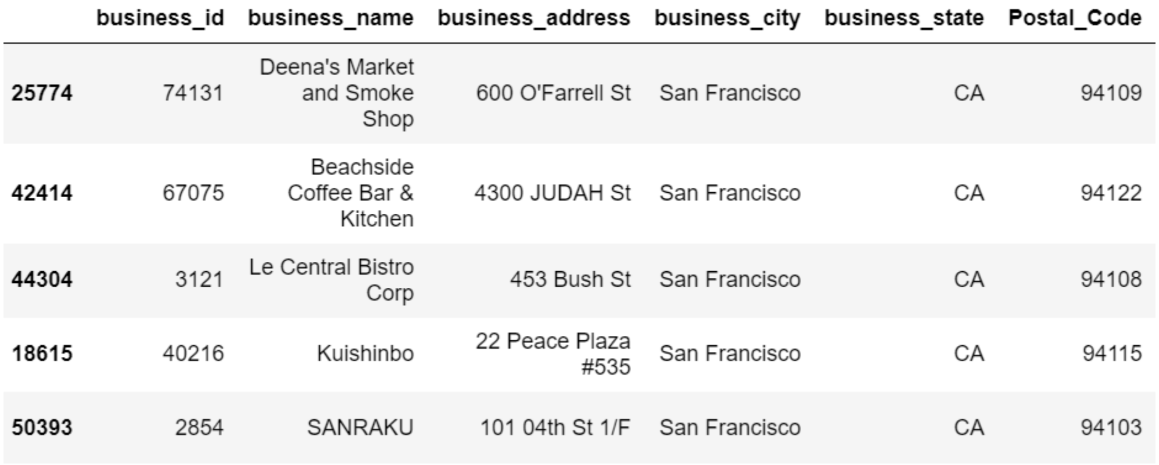
1. There are overlapping post codes in San Francisco that belong to more than one neighbourhood. This is a key challenge because that means the segregation is most likely at a complete address level. While this may not be unique to San Francisco, when we are dealing with a problem statement that looks at neighbourhood, this leads to some complexity.
2. Based on point 1, as part of this, we only factored one neighbourhood for each postcode. That meant that it is possible that while ALL restaurants that had a postcode was mapped to a neighbourhood, some of them may not be the official neighbourhood but an estimate.
3. There were cases where restaurants didn’t have postcode data in their address. This was removed from the dataset.
4. All the restaurants that had blank inspection score, blank risk category and blank neighbourhood were removed from the dataset

The problem statement revolved around identifying the most relevant neighbourhood where it would make most sense to set up the new food hubs. **The original concept was to take a combination of official inspection data and combine it with user review data from**

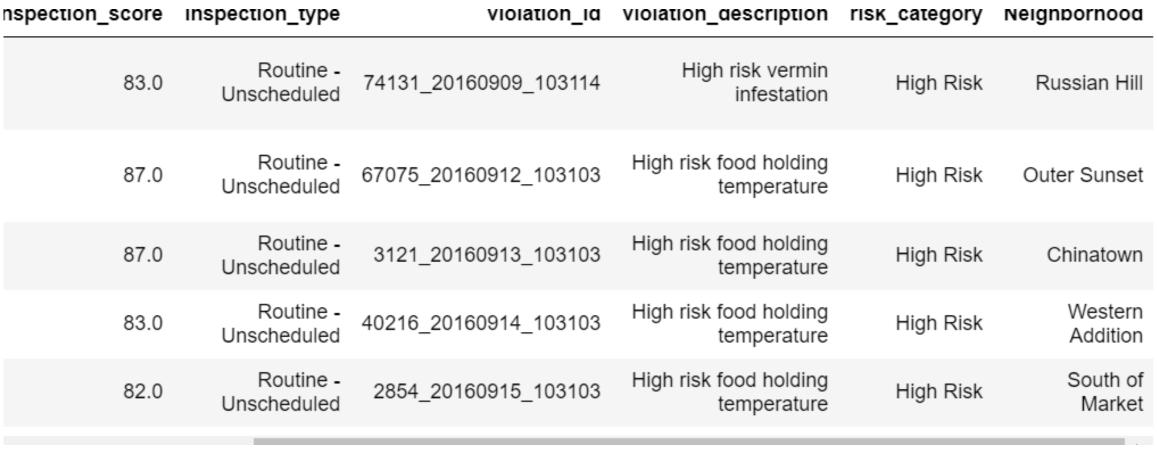
Foursquare to assign a weight score to the restaurants. From there we would continue the study.

Unfortunately as stated above, there were challenges to get Foursquare data and therefore the analysis was limited to inspection related data alone.

In that context, the approach was to focus on restaurants that are deemed as High Risk in terms of category as part of the inspection process.



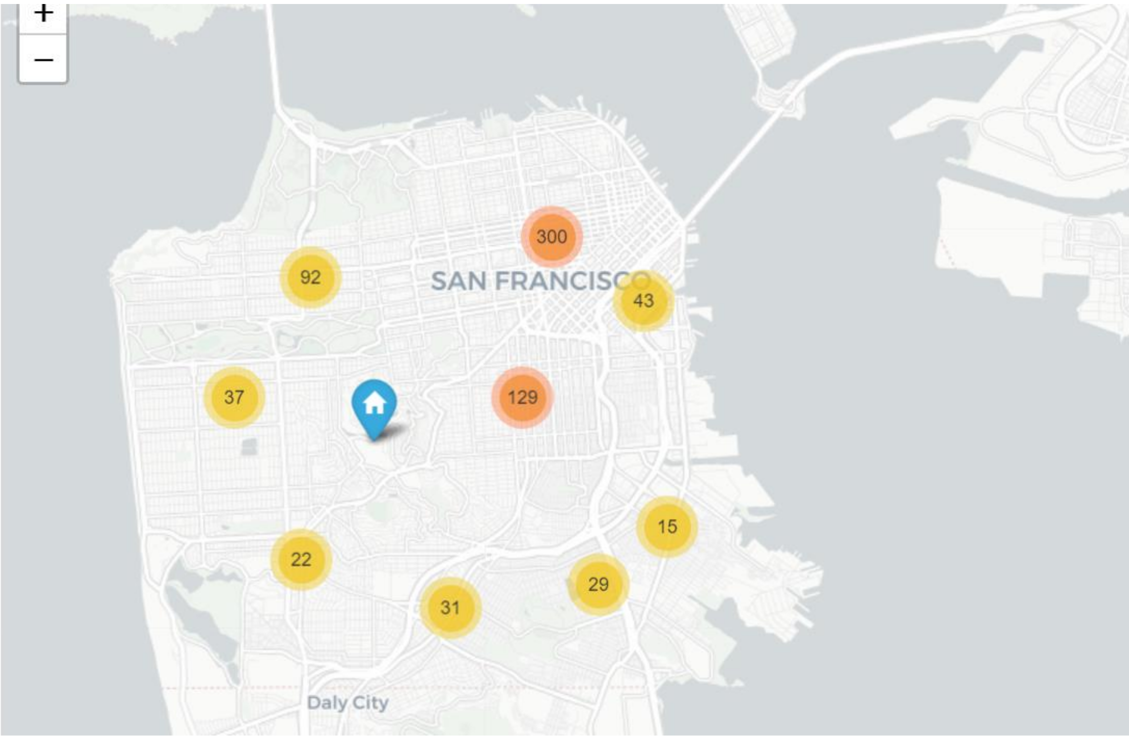
	business_id	business_name	business_address	business_city	business_state	Postal_Code
25774	74131	Deena's Market and Smoke Shop	600 O'Farrell St	San Francisco	CA	94109
42414	67075	Beachside Coffee Bar & Kitchen	4300 JUDAH St	San Francisco	CA	94122
44304	3121	Le Central Bistro Corp	453 Bush St	San Francisco	CA	94108
18615	40216	Kuishinbo	22 Peace Plaza #535	San Francisco	CA	94115
50393	2854	SANRAKU	101 04th St 1/F	San Francisco	CA	94103



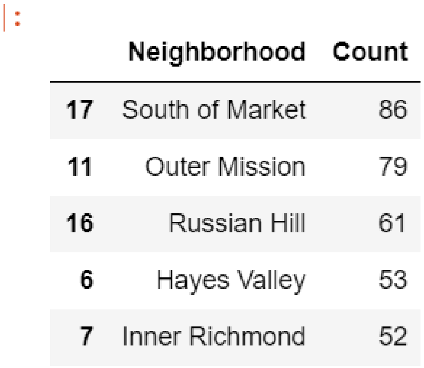
nspection_score	inspection_type	violation_id	violation_description	risk_category	Neighborhood
83.0	Routine - Unscheduled	74131_20160909_103114	High risk vermin infestation	High Risk	Russian Hill
87.0	Routine - Unscheduled	67075_20160912_103103	High risk food holding temperature	High Risk	Outer Sunset
87.0	Routine - Unscheduled	3121_20160913_103103	High risk food holding temperature	High Risk	Chinatown
83.0	Routine - Unscheduled	40216_20160914_103103	High risk food holding temperature	High Risk	Western Addition
82.0	Routine - Unscheduled	2854_20160915_103103	High risk food holding temperature	High Risk	South of Market

Given the dataset was an ongoing repository of all inspections and the scores, a lot of restaurants had multiple entries. For the purposes of this analysis, we only looked at the latest inspection when there were multiple. The idea here was that enabled the possibility of restaurant's intent to improve their score was captured and factored.

Based on the streamlined data set, we started the analysis. First we did a quick simple visualization of the high risk restaurants geographically on a Folium map with cluster marker. This is shown below as a snapshot.



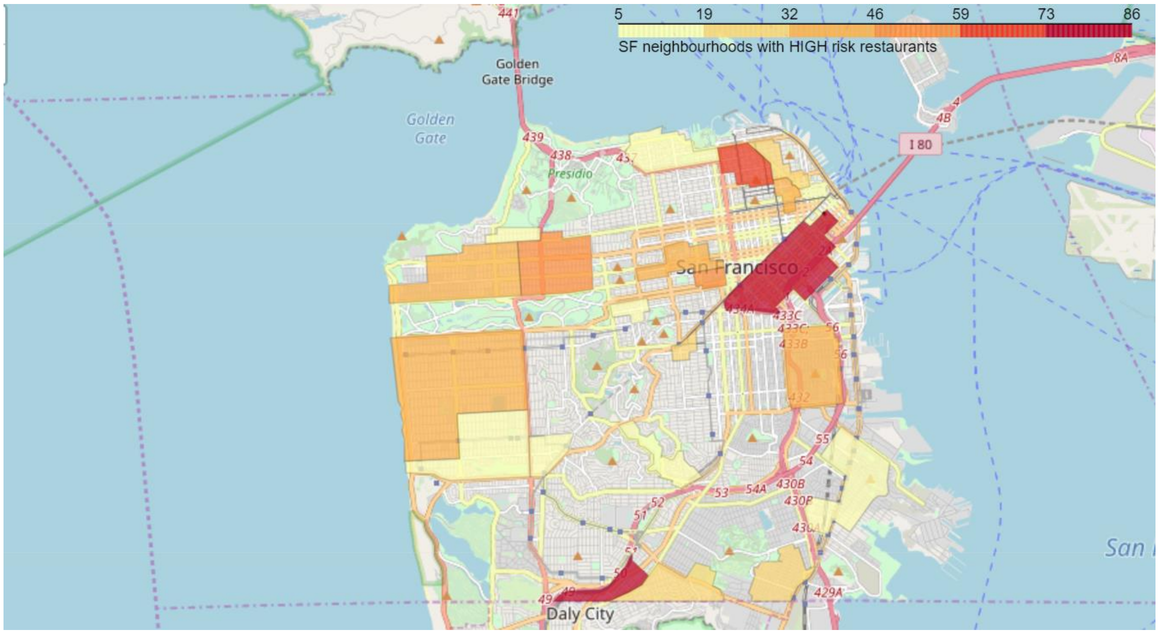
The next part of the analysis focused on getting the neighbourhoods with the maximum number of High Risk restaurants and identify the neighbourhoods. That resulted in a summary as shown in the snapshot below.



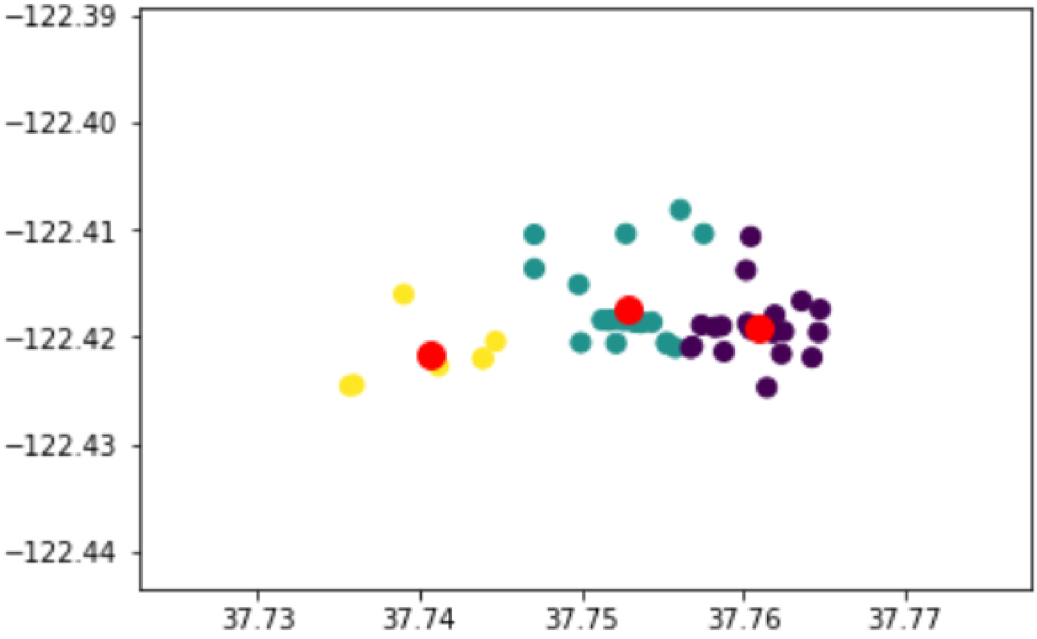
	Neighborhood	Count
17	South of Market	86
11	Outer Mission	79
16	Russian Hill	61
6	Hayes Valley	53
7	Inner Richmond	52

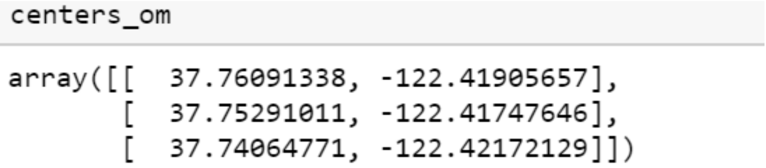
We wanted to then try out another visualization to indicate the spread of the high risk restaurants on the city map to get a sense of the concentration of the presence. A choropleth map was created to add that additional view.

This is where we made use of the SF neighbourhoods GeoJSON file that was available from https://data.sfgov.org website.



Based on the assumption that High Risk restaurants were most likely to be closed based on quality and health regulation compliance, we focused on the 3 neighbourhoods with the most High Risk restaurants to continue the study. The next step was to identify the 3 ideal locations in the neighbourhood to open the new restaurants. We made use of k-means clustering of the subset of geocodes of the high risk restaurants to identify the center points of each.

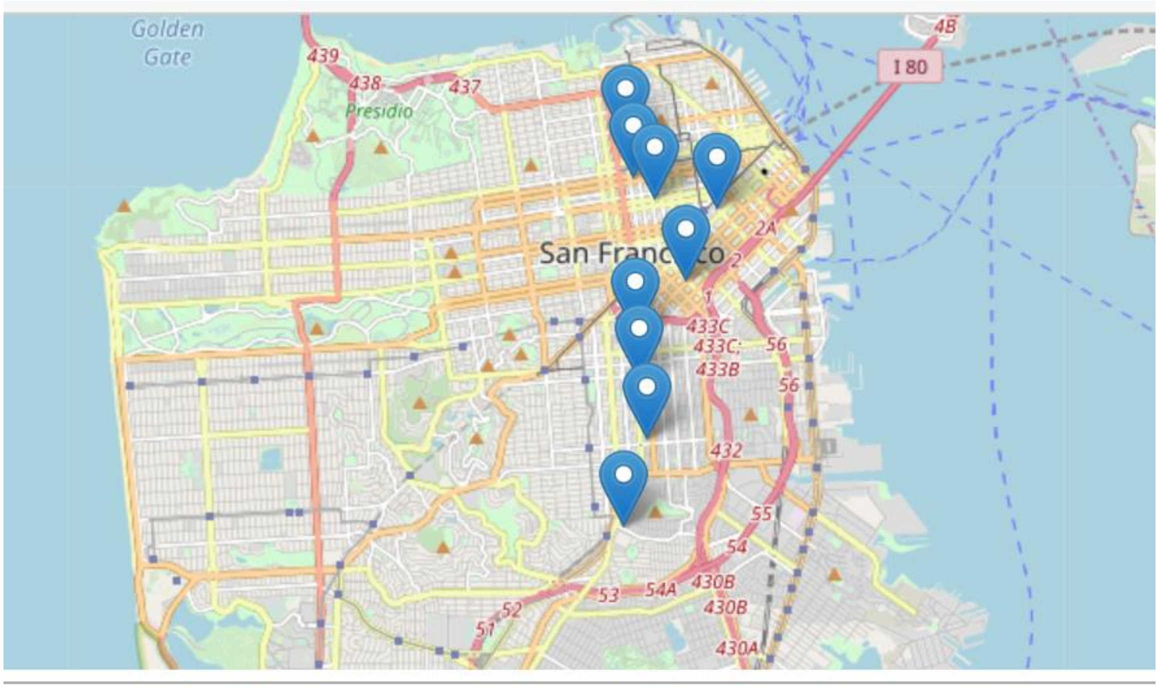




```
centers_om
```

```
array([[  37.76091338, -122.41905657],
       [  37.75291011, -122.41747646],
       [  37.74064771, -122.42172129]])
```

This was done for the remaining 2 neighbourhoods to identify the 9 locations of new restaurants in total to address the potential closures of the High Risk restaurants.



Results

As stated through the above section, with the intent of identifying the neighbourhoods that had to be the focus of the new hubs, the proposed neighbourhoods are:

1. Outer Mission
2. Russian Hill
3. South of Market

The restaurants with the lowest scores in Outer Mission are:

1. LA TRAVIATA
2. Duc Loi Supermarket
3. Yangtze Market

Based on Google search for these establishments, the **proposed new restaurants in Outer Mission should be Italian, Vietnamese and Chinese cuisine.**

The restaurants with the lowest scores in South of Market are:

1. Lers Ros 16th Street
2. Poc-Chuc Restaurant
3. Mission Beach Café

Based on Google search for these establishments, the **proposed new restaurants in South of Market should be Thai, Spanish and Californian cuisine.**

The restaurants with the lowest scores in Russian Hill are:

1. ITHAI
2. Jayhoon Fedaiy
3. Vietnam House

Based on Google search for these establishments, the **proposed new restaurants in Russian Hill should be Thai, Vietnamens and Indian cuisine.**

Discussion

This project was an excellent exercise to convey the message to all aspiring data scientists – there are tools and it is possible to get the training to use those tools but every analysis needs to have a clear purpose and a high degree of confidence before it can be considered as fit for purpose.

In today's world, data-driven decision making is getting more importance than before. Courses and projects like this one reinforce the relevance of that growing dependency and importance given to data science.

Keeping that in mind, this specific exercise has highlighted to me a few things. Firstly, this is a initial analysis that gives a brief glimpse of a possibility of how the objective could be achieved. While this helps narrowing down the scope of investigation based on restaurant quality, it doesn't necessarily define absolute terms a conclusion.

The lack of user reviews/ratings which was the original intent plays a big role in this analysis. A restaurant survives based on both basic health and safety compliance, but it remains profitable and sustainable based on consumer demands. There could be several Low Risk or Moderate Risk restaurants that may not be liked by consumers which is an aspect not present in this study.

Population demographic is another aspect that is not part of this study. Neighbourhoods may have over-representation of certain communities and ethnicities that naturally demand more restaurants of certain cuisines. For example, an authentic Italian or Spanish restaurant in a neighbourhood like Chinatown will not be a wise business decision unless there is factual evidence of such demands. So studying the patterns of that consumer data would make this study more complete.

Lastly, this should trigger a side study on the effectiveness of the inspection process as well. We ignored the repeated studies in this analysis and focused on the latest scores. However, the repeated inspections and obvservation of the score fluctuations could also be a factor to indicate if the ownership and accountability of such a critical procedure is effective. And if not, that should also play a role in deciding if a certain restaurant should be allowed to continue doing business.

Conclusion

In conclusion, this was not an easy exercise with regards to the fact that the original approach had to be modified well into the analysis. Therefore, I acknowledge that this study may come across as simple or incomplete and full of assumptions.

But the overall process of data analysis was followed in this project along with applications of concepts learnt along the course.

Special thanks to Alex Aklson for being flexible and allowing me to continue the exercise without Foursquare data.