Analysis the businesses in San Francisco during covid-19.

Konstantin Kovalev

July 29, 2020

1. Introduction

One of the most important problem in our days is covid-19 and his backwash. There were and are huge financial losses, and at the moment the situation has not been corrected. Many business owners and employees are in a difficult situation.

Data that can help identify a problem may include lists of locations in the city, information about the current state of business for each venue, geodata and etc. All data is public and free. This project aims to show the problems for the business and the scale of the consequences from the virus.

Obviously, everyone would be very interested to look to conclusions from this work

2. Data acquisition and cleaning

2.1 Data sources

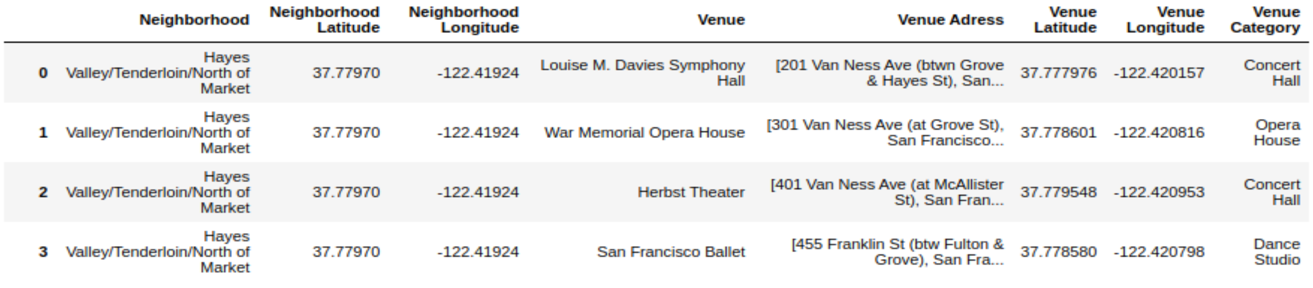
One part of data presented in the form of html tables or similar. There will be no problems extracting it. All necessary geo data and postal data are in the public domain. It could be government websites or similar.

Next part of data will be received thru Foursquare API. In this case there will be no problem too.

Last part is getting data from web about current business status. I will pars Google Maps for this. There can be difficulties for getting clear data. First it can take too much time for parsing. Second it can get holes for tables because source of venue is one company and source of status is other. Also there can be difficulties with names of places during creating request to Google.

2.2 Data cleaning and selection

Eventually, I have data frame with 1820 rows with name of venues, coordinates, addresses, types of place, business status grouped by neighborhoods. There was no missed data for every step except getting venue status. Borough data was downloaded from web. It was not much size.

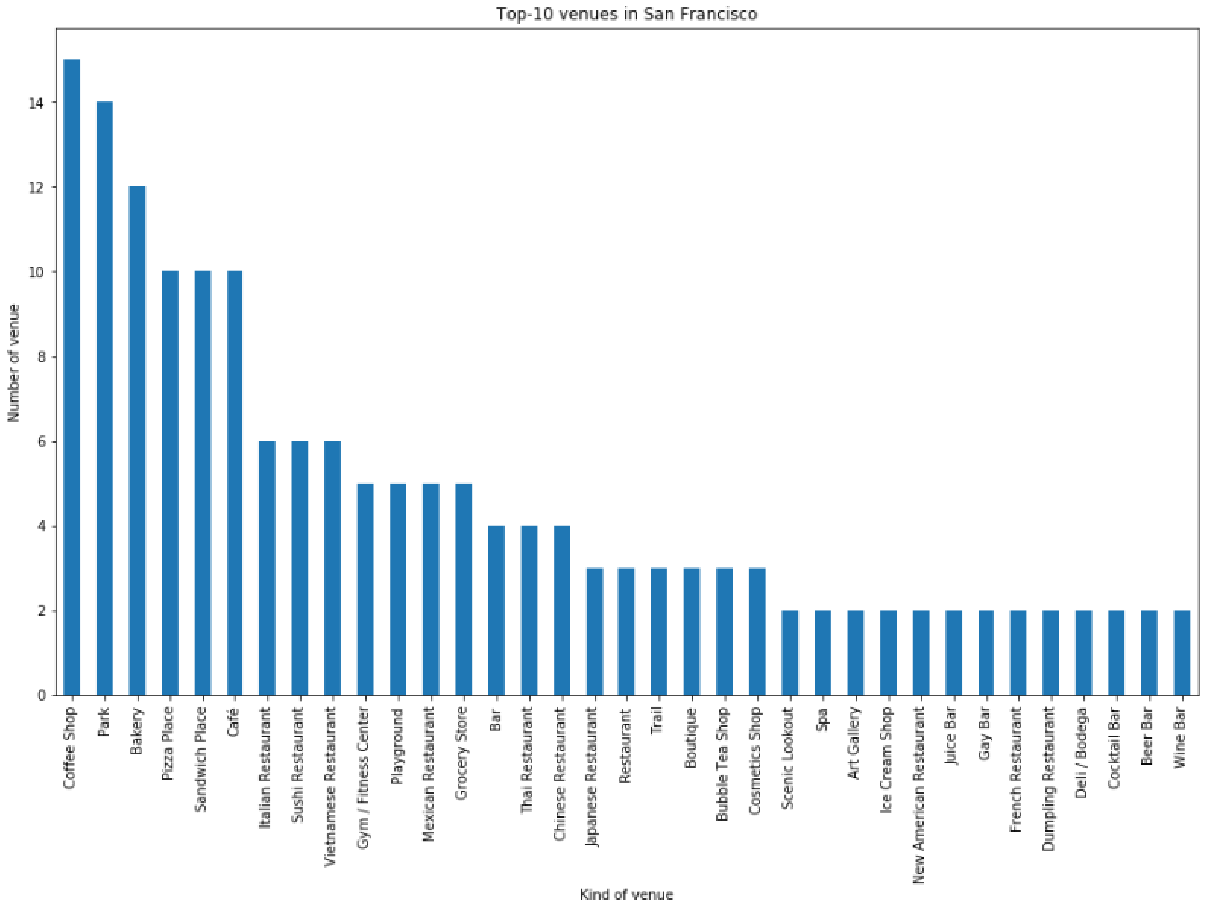


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Adress	Venue Latitude	Venue Longitude	Venue Category
0	Hayes Valley/Tenderloin/North of Market	37.77970	-122.41924	Louise M. Davies Symphony Hall	[201 Van Ness Ave (btwn Grove & Hayes St), San...	37.777976	-122.420157	Concert Hall
1	Hayes Valley/Tenderloin/North of Market	37.77970	-122.41924	War Memorial Opera House	[301 Van Ness Ave (at Grove St), San Francisco...	37.778601	-122.420816	Opera House
2	Hayes Valley/Tenderloin/North of Market	37.77970	-122.41924	Herbst Theater	[401 Van Ness Ave (at McAllister St), San Fran...	37.779548	-122.420953	Concert Hall
3	Hayes Valley/Tenderloin/North of Market	37.77970	-122.41924	San Francisco Ballet	[455 Franklin St (btw Fulton & Grove), San Fra...	37.778580	-122.420798	Dance Studio

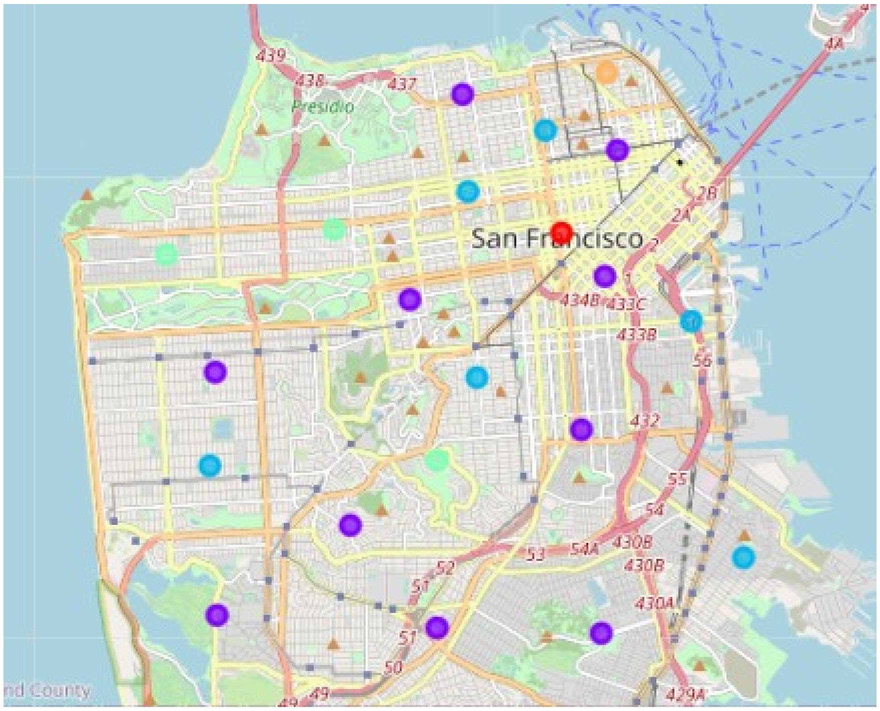
On the step of parsing I took decision just check for one condition and just skipping non-standard moments that can be some a little error. I think it is no problem.

3. Methodology

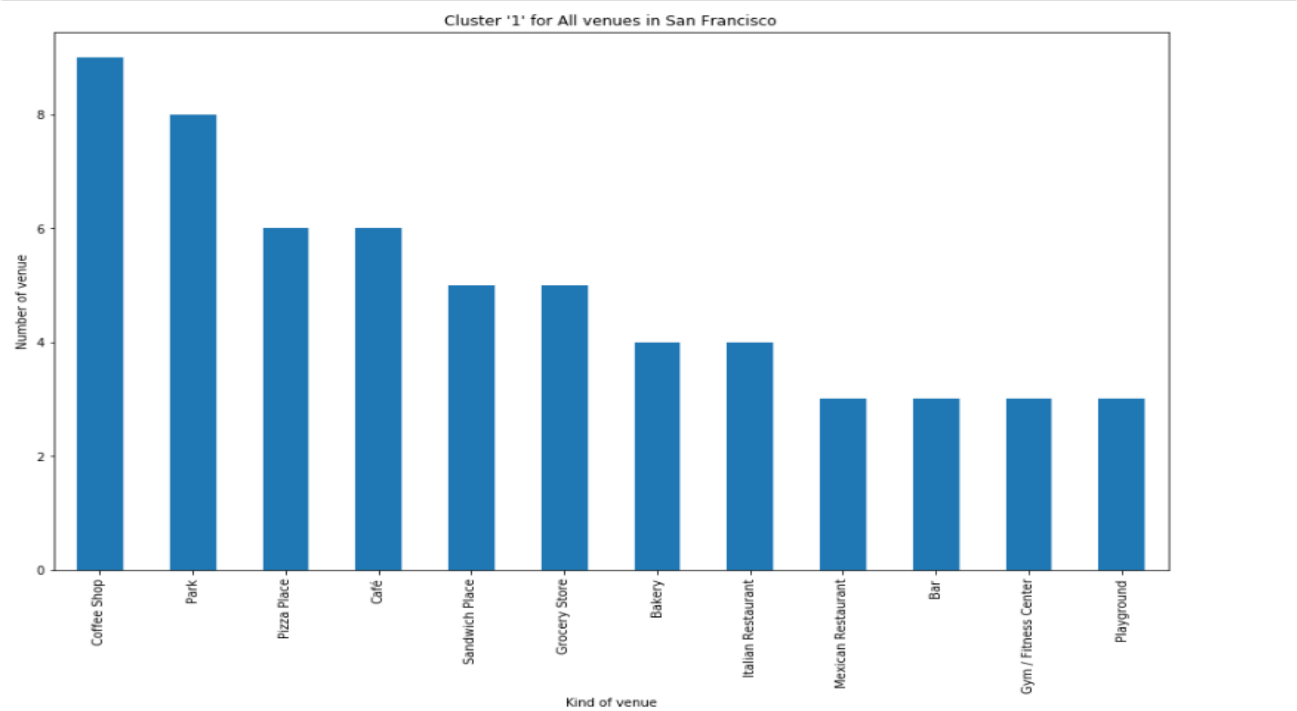
The idea is as follows, I take all data for open and closed venues. I should group data by district and calculate average value for every type of place. Next I write top 10 types of venue to one table.
I can estimate top venues for San Francisco.



Based on this table I make segmentation.
I use K-means clustering one of the popular method. K-means clustering is a method of vector quantization, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.
After I got clusters I can add this data to table and draw the map for visualization.



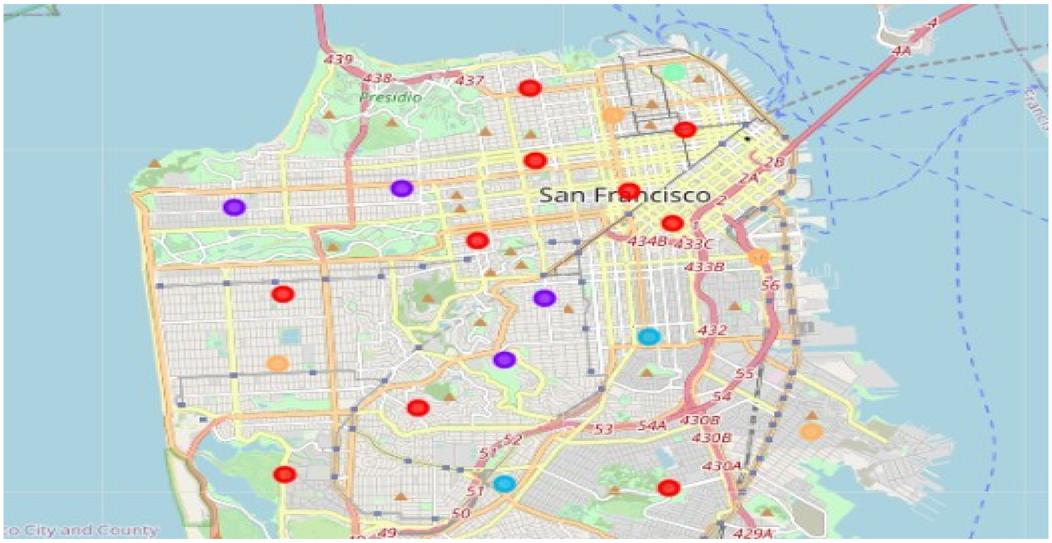
Next I can estimate the groups of clusters and give them names. Foe example, most popular one might be called [Coffee, Park, Pizza, Cafe] .



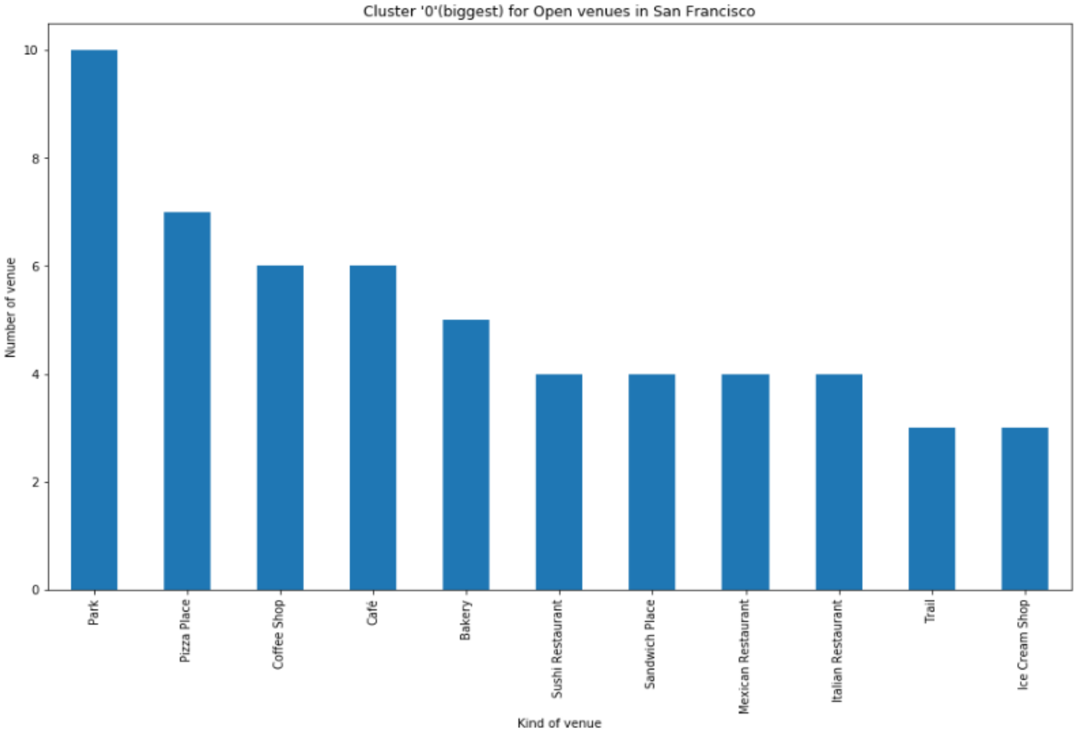
4.Results

As a result of my work, I divided data to opened venues and closed and made same manipulation as previous part.

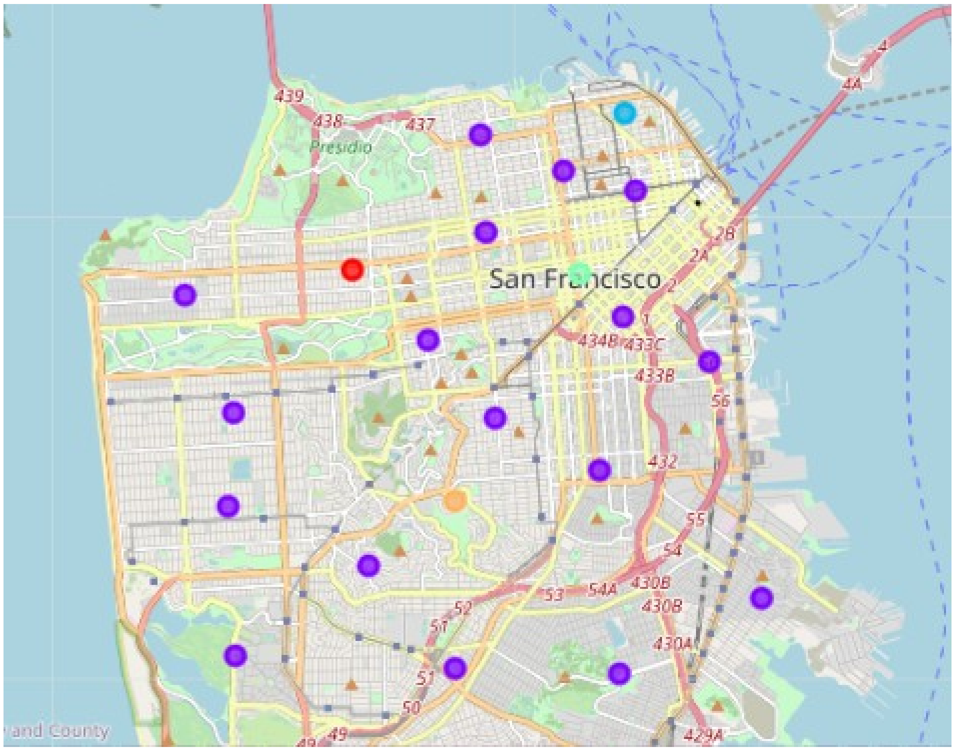
As a percentage differences is about 12% for closed.
As you can see, the distributions of clusters for Opened places differs from the initial state in the number of cluster elements



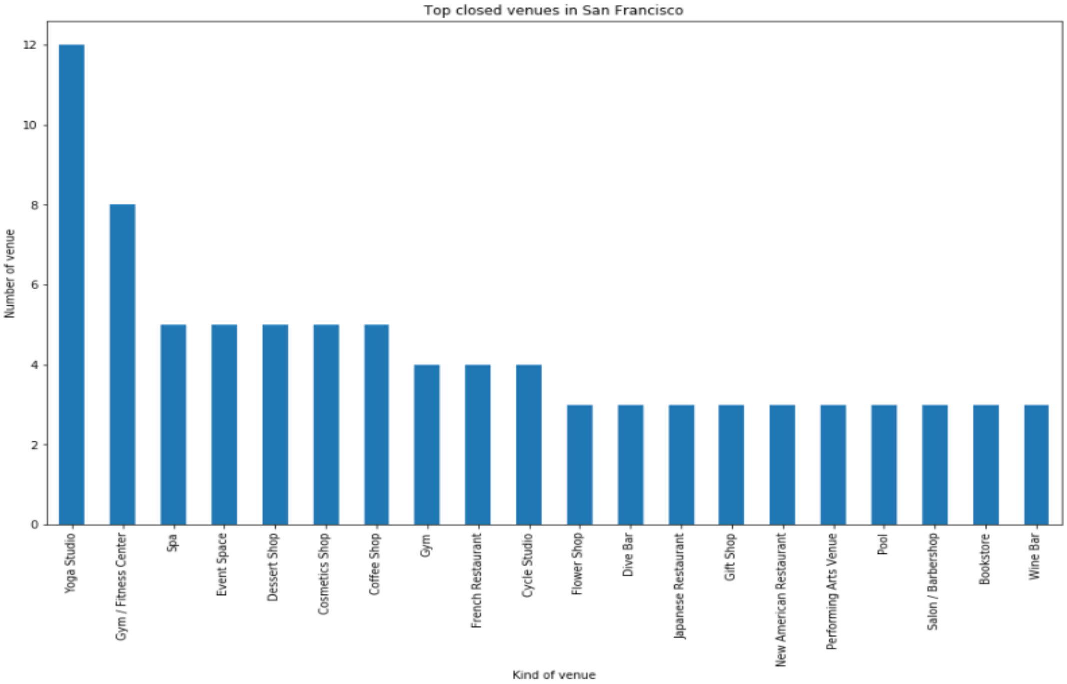
But main and biggest cluster also can name [Coffee, Park, Pizza, Cafe]



For Closed venues situation is radically opposite. There is one big cluster spread on whole city.



We can cal it [Sport] or [Yoga, Gym, Fitness]



5.Discussion

Certainly, for fuller and more accurate analysis I had to have more data. I mean, I had to got more venues. The best way is have all venues for city. Also, I could more accurate grab data from web. For example, checking merchant not for only one field "Temporary closed" and for another field as "Permanently closed".
Additionally, I could divide each available neighborhood for smaller parts

6.Conclusion

Definitely for this data set I can make accurate conclusion. A lot of business still have problems, someone more, someone less. The main group of venue with greater consequences is sport-venues as fitness and yoga centers, gyms. I can not highlight some borough, problems spread thru all city and does not depend for location.