Capstone Project - Places most Covid-19 infected cases and deaths in New York (Week 2)

Applied Data Science Capstone by Rafael Mejia S.

June 2020

Table of contents

1. Introduction

1.1. Background

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The origin was first identified in Wuhan, China, in December 2019. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January 2020, and a pandemic on 11 March. As of 19 June 2020, more than 8.62 million cases of COVID-19 have been reported in more than 188 countries and territories, resulting in more than 458,000 deaths; more than 4.22 million people have recovered.

The virus is primarily spread between people during close contact, most often via small droplets produced by coughing, sneezing, and talking. Less commonly, people may become infected by touching a contaminated surface and then touching their face It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms.

Common symptoms include fever, cough, fatigue, shortness of breath, and loss of sense of smell.

Recommended preventive measures include hand washing, covering one's mouth when coughing, maintaining distance from other people, wearing a face mask in public settings, and monitoring and self-isolation for people who suspect they are infected. Authorities worldwide have responded by implementing travel restrictions, lockdowns, workplace hazard controls, and facility closures. Many places have also worked to increase testing capacity and trace contacts of infected persons.

1.2. Problem

The pandemic Covid-19 is a tragedy that affected a million of people around the world.

United States was not exception, especially New York City. It's a place where the illness growth exponentially during the last three months.

For this reason, we need to know the places where the cases and deaths for Covid-19 were most affected. Which are the features of the neighborhood and how can prevent for the future the cases will not raise quickly.

1.3. Interest

This study is for State Government of New York to know which are the features of every neighborhood with most people infections. How similar are they neighborhoods and their respective near venues. Also, we believe that is important to all people to know what place need to take prevention when visit the city in the future.

2. Data acquisition and cleaning

2.1. Data sources

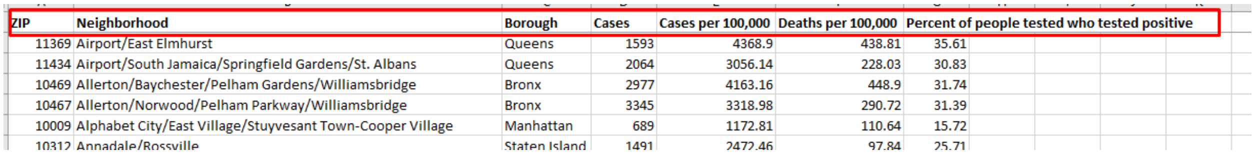
To solve the problem, we will use the dataset of Covid-19 belong to New York City governance. This data contains the Zip, Neighborhood, Borough, Cases, Cases per 100.000, Deaths per 100.000 and the Percent of people tested who tested positive.

This data I will combine with the geolocation information of foursquare to get the features of each neighborhood and make a clustering to define the relationships between the positive cases/deaths and the neighborhoods in New York.

Finally, I will detect the clustering of neighborhoods most optimal and then give the suggestions to reduce the amount of infections in another similar pandemic.

2.2. Data cleaning

The dataset got from https://health.data.ny.gov/browse was not cleaning, only rename the columns to make more easy the analysis.

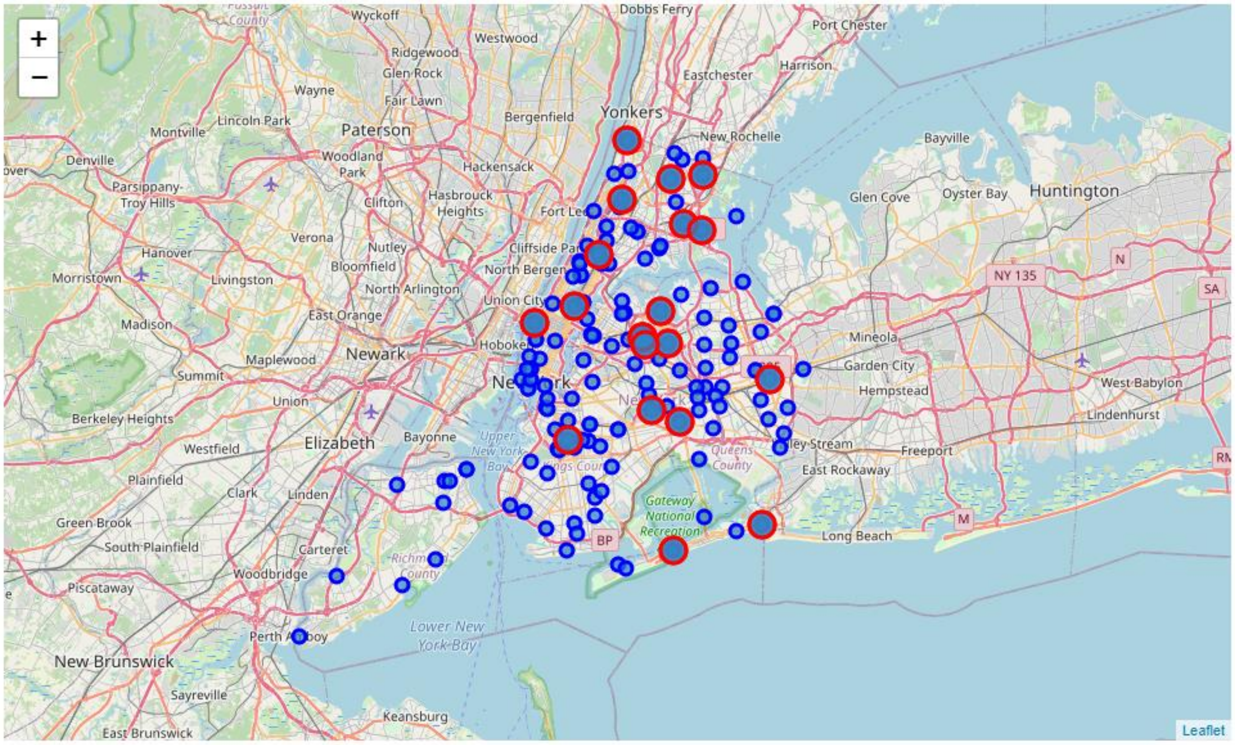


ZIP	Neighborhood	Borough	Cases	Cases per 100,000	Deaths per 100,000	Percent of people tested who tested positive
11369	Airport/East Elmhurst	Queens	1593	4368.9	438.81	35.61
11434	Airport/South Jamaica/Springfield Gardens/St. Albans	Queens	2064	3056.14	228.03	30.83
10469	Allerton/Baychester/Pelham Gardens/Williamsbridge	Bronx	2977	4163.16	448.9	31.74
10467	Allerton/Norwood/Pelham Parkway/Williamsbridge	Bronx	3345	3318.98	290.72	31.39
10009	Alphabet City/East Village/Stuyvesant Town-Cooper Village	Manhattan	689	1172.81	110.64	15.72
10312	Annadale/Rossville	Staten Island	1491	2472.46	97.84	25.71

Also, the geolocation got for every neighborhood was successfully, without null values.

2.3. Feature Selection

I process the data to focus only in the top 20 most cases/death of Covid-19 because there are the current places more critical.



Graph 1: Map of New York with top 20 cases/deaths Covid-19 (red circle)

Also, we identify the venues categories more common near these neighborhoods in the radius of one kilometer.

3. Exploratory Data Analysis

3.1. Target Analysis

First, I grouping the neighborhoods with the venue categories got from the Foursquare. I focus on the top 5 most frequent venues categories for each neighborhood.

```
----Airport/East Elmhurst----
             venue  freq
0  Airport Service  0.15
1   Airport Lounge  0.15
2     Burger Joint  0.10
3        Hotel Bar  0.10
4      Coffee Shop  0.10


----Allerton/Baychester/Pelham Gardens/Williamsbridge----
```

```
                  venue  freq
0  Caribbean Restaurant  0.15
1               Brewery  0.05
2            Soup Place  0.05
3           Supermarket  0.05
4             Nightclub  0.05


----Belle Harbor-Neponsit/Rockaway Park----
                venue  freq
0               Beach  0.40
1       Deli / Bodega  0.10
2                 Pub  0.10
3               Trail  0.05
4  Mexican Restaurant  0.05


----Co-op City/Edenwald----
                  venue  freq
0              Pharmacy  0.10
1           Pizza Place  0.10
2         Shopping Mall  0.10
3  Caribbean Restaurant  0.10
4          Cocktail Bar  0.05


----Corona/North Corona----
                       venue  freq
0  Latin American Restaurant  0.10
1     Argentinian Restaurant  0.10
2         Italian Restaurant  0.10
3                Pizza Place  0.10
4             Ice Cream Shop  0.05


----Country Club/Throgs Neck----
                 venue  freq
0  American Restaurant  0.15
1          Pizza Place  0.15
2       Ice Cream Shop  0.05
3                 Café  0.05
4   Mexican Restaurant  0.05


----Cypress Hills/East New York----
           venue  freq
0    Pizza Place  0.25
1  Grocery Store  0.17
2    Supermarket  0.08
3     Donut Shop  0.08
4           Bank  0.08


----East Harlem----
```

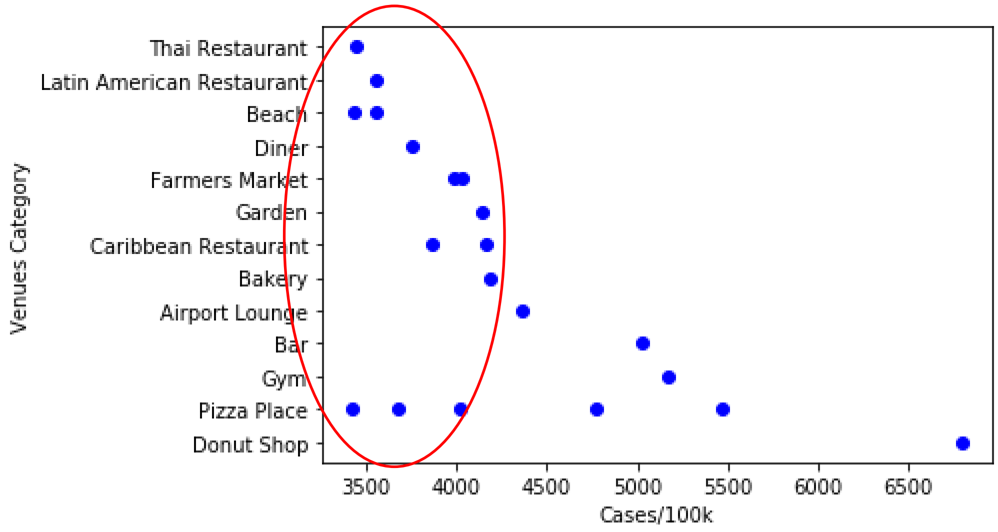
```
                  venue  freq
0                   Bar  0.15
1    African Restaurant  0.05
2               Theater  0.05
3               Library  0.05
4  Gym / Fitness Center  0.05


----East New York----
                        venue  freq
0                      Garden  0.20
1                        Park  0.10
2            Botanical Garden  0.10
3  Tourist Information Center  0.05
4                  Playground  0.05


----Edgemere/Far Rockaway----
            venue  freq
0     Supermarket  0.15
1           Beach  0.15
2            Bank  0.10
3  Sandwich Place  0.10
4      Donut Shop  0.10


----Elmhurst----
                venue  freq
0     Thai Restaurant  0.30
1  Mexican Restaurant  0.15
2         Pizza Place  0.05
3  Chinese Restaurant  0.05
4          Food Truck  0.05


----Fieldston/North Riverdale/Riverdale----
                venue  freq
0         Pizza Place  0.10
1                Pool  0.10
2  Italian Restaurant  0.10
3        Burger Joint  0.10
4      Ice Cream Shop  0.05


----Fordham/Kingsbridge/University Heights----
                       venue  freq
0  Latin American Restaurant  0.10
1         African Restaurant  0.05
2                        Gym  0.05
3                Music Venue  0.05
4         Spanish Restaurant  0.05


----Hell's Kitchen/Midtown Manhattan----
```

```
            venue  freq
0             Gym  0.15
1  Ice Cream Shop  0.05
2         Dog Run  0.05
3     Comedy Club  0.05
4     Coffee Shop  0.05


----Jackson Heights----
              venue  freq
0        Food Truck   0.1
1            Bakery   0.1
2   Thai Restaurant   0.1
3  Arepa Restaurant   0.1
4    Farmers Market   0.1


----Jackson Heights/Rikers Island----
              venue  freq
0        Food Truck   0.1
1            Bakery   0.1
2   Thai Restaurant   0.1
3  Arepa Restaurant   0.1
4    Farmers Market   0.1


----Lenox Hill/Upper East Side----
                venue  freq
0  Italian Restaurant  0.10
1         Coffee Shop  0.10
2              Bakery  0.10
3        Dessert Shop  0.05
4           Hotel Bar  0.05


----Morris Park/Pelham Bay/Westchester Square----
                       venue  freq
0                      Diner  0.10
1                Pizza Place  0.10
2  Latin American Restaurant  0.05
3                        Bar  0.05
4                 Restaurant  0.05


----Ozone Park----
               venue  freq
0           Pharmacy  0.10
1        Pizza Place  0.10
2                Gym  0.10
3     Ice Cream Shop  0.05
4  Health Food Store  0.05


----Queens Village----
```

```
            venue  freq
0     Pizza Place  0.10
1      Donut Shop  0.10
2  Discount Store  0.10
3        Bus Stop  0.10
4  Ice Cream Shop  0.05
```

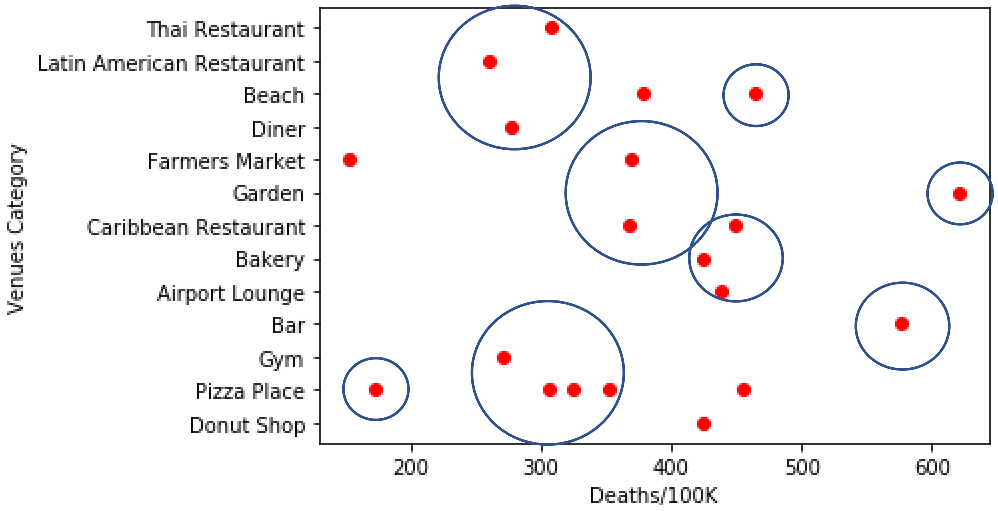
3.2. Relationship between Cases/100k and Venue Categories

We watch in the graph that the infections are more in the beach, farmers market and food places where are people conglomerations.



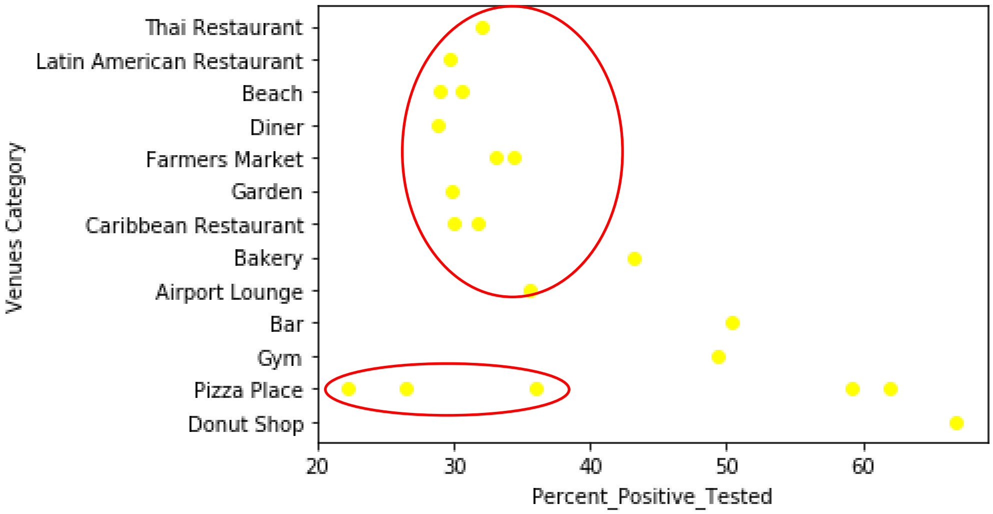
3.3. Relationship between Deaths/100k and Venue Categories

We watch in the graph not patterns in this case, in all categories are people death for the illness.



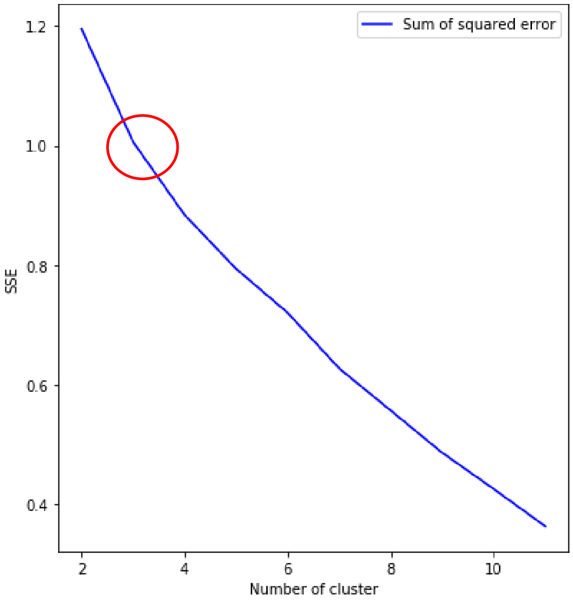
3.4. Relationship between Percent Positive Tested and Venue Categories

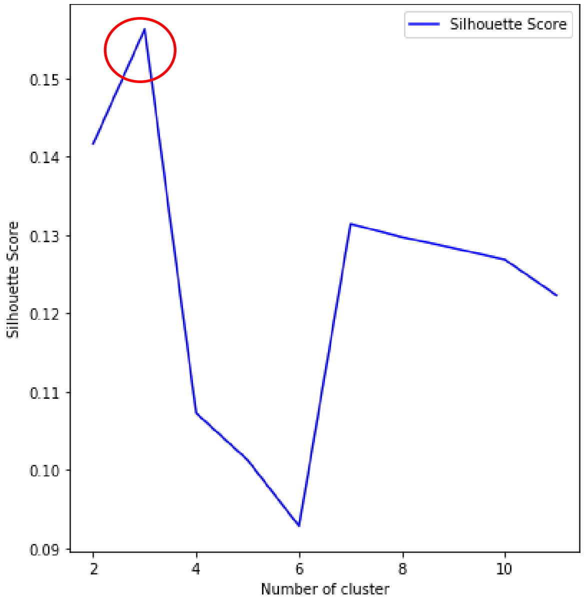
We watch in the graph similar pattern with Cases/100k chart. The concentration of people with percent positive teste are in the food places and beach.



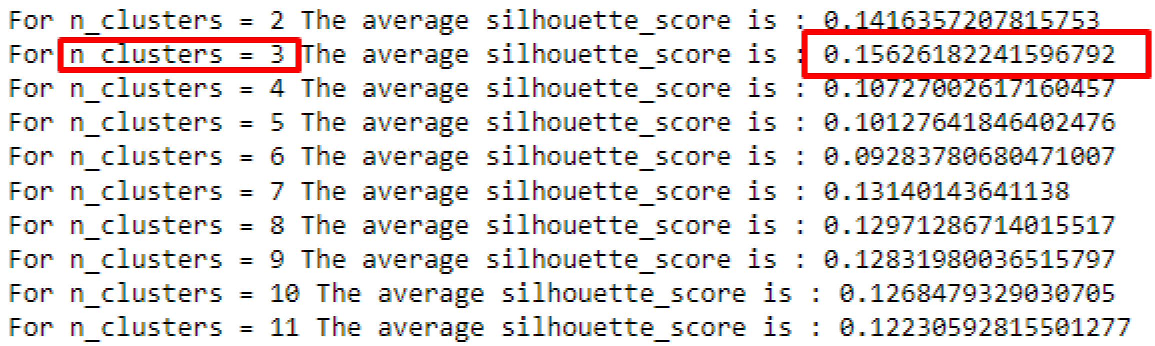
4. Clustering Data

I used the K-means algorithm to clustering the information because I can use the specific data for segmentation, I don't have predictive nothing, only clustering the neighborhood with the more optimal value for K. For this reason, I used two methods to find the best value of K: silhouette score and elbow (sum of squared error) algorithm.



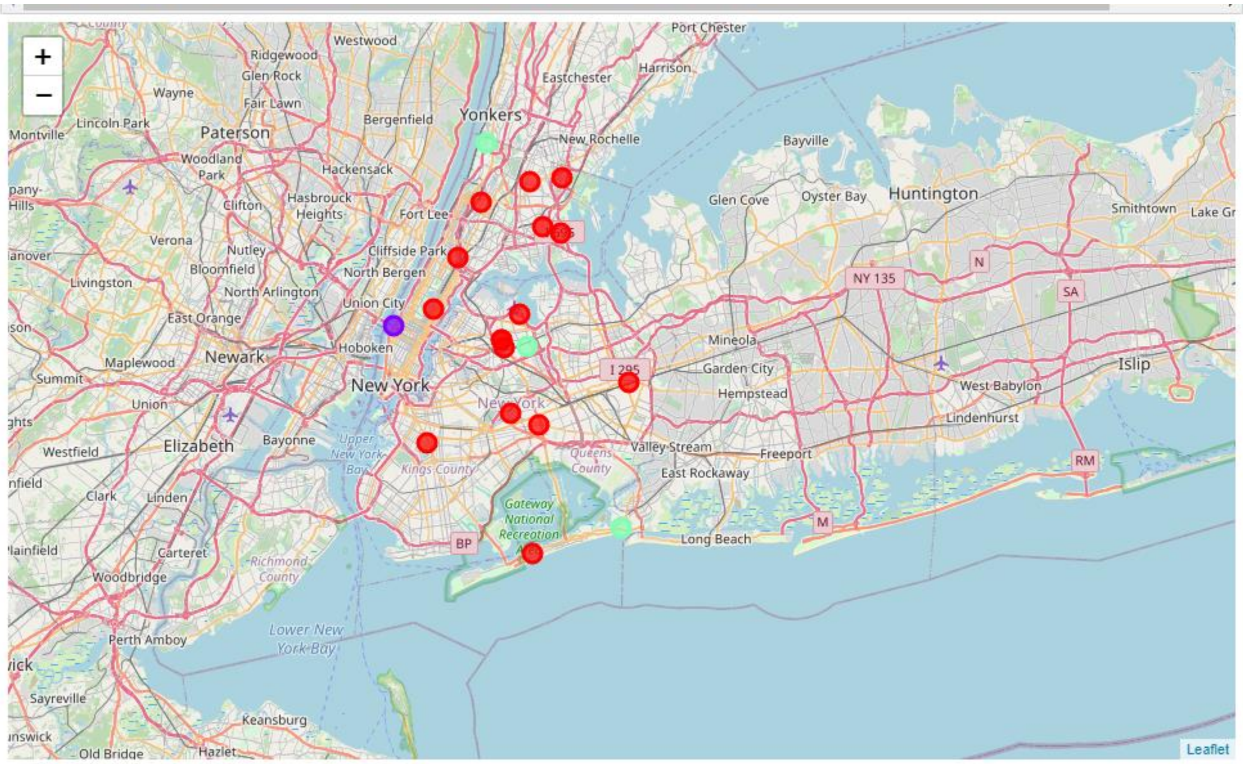


Watching the graphs and according to the results, the best value for K is n_cluster = 3 because the average silhouette scores the max value is 0.156.



```
For n_clusters = 2 The average silhouette_score is : 0.1416357207815753
For n_clusters = 3 The average silhouette_score is : 0.15626182241596792
For n_clusters = 4 The average silhouette_score is : 0.10727002617160457
For n_clusters = 5 The average silhouette_score is : 0.10127641846402476
For n_clusters = 6 The average silhouette_score is : 0.09283780680471007
For n_clusters = 7 The average silhouette_score is : 0.13140143641138
For n_clusters = 8 The average silhouette_score is : 0.12971286714015517
For n_clusters = 9 The average silhouette_score is : 0.12831980036515797
For n_clusters = 10 The average silhouette_score is : 0.1268479329030705
For n_clusters = 11 The average silhouette_score is : 0.12230592815501277
```

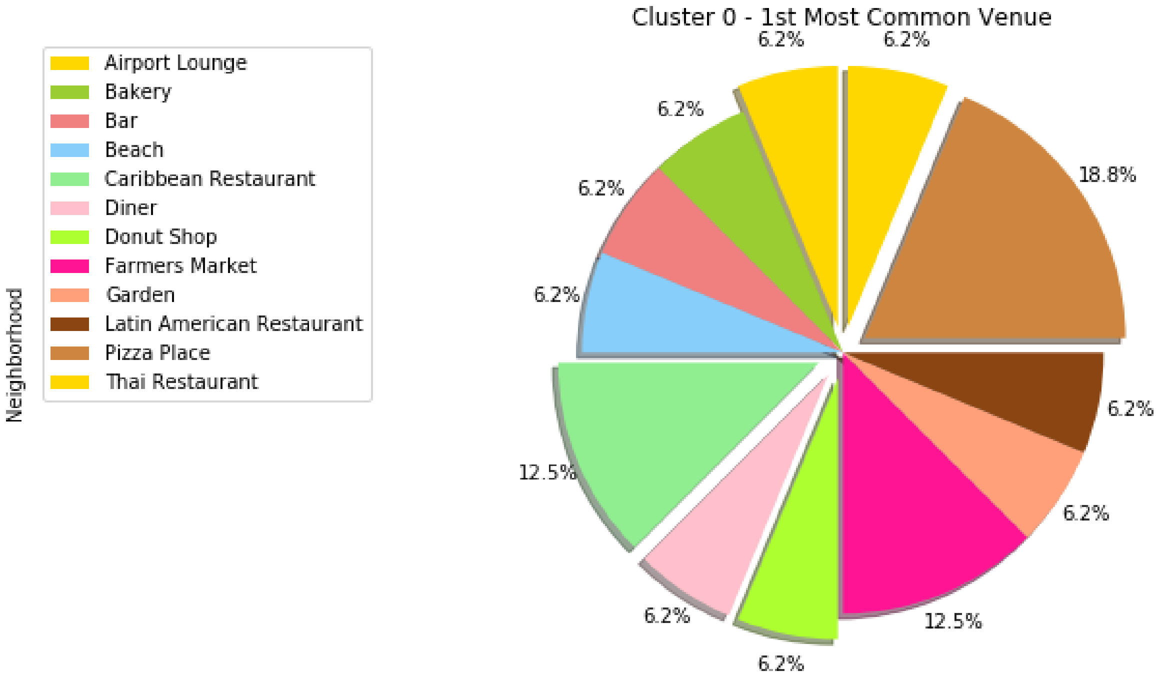
Now the clustered neighborhoods are distributed on the map in this way:



Where the red color is the cluster 0, red cyan is belong to cluster 1 and the color purple if for cluster 3.

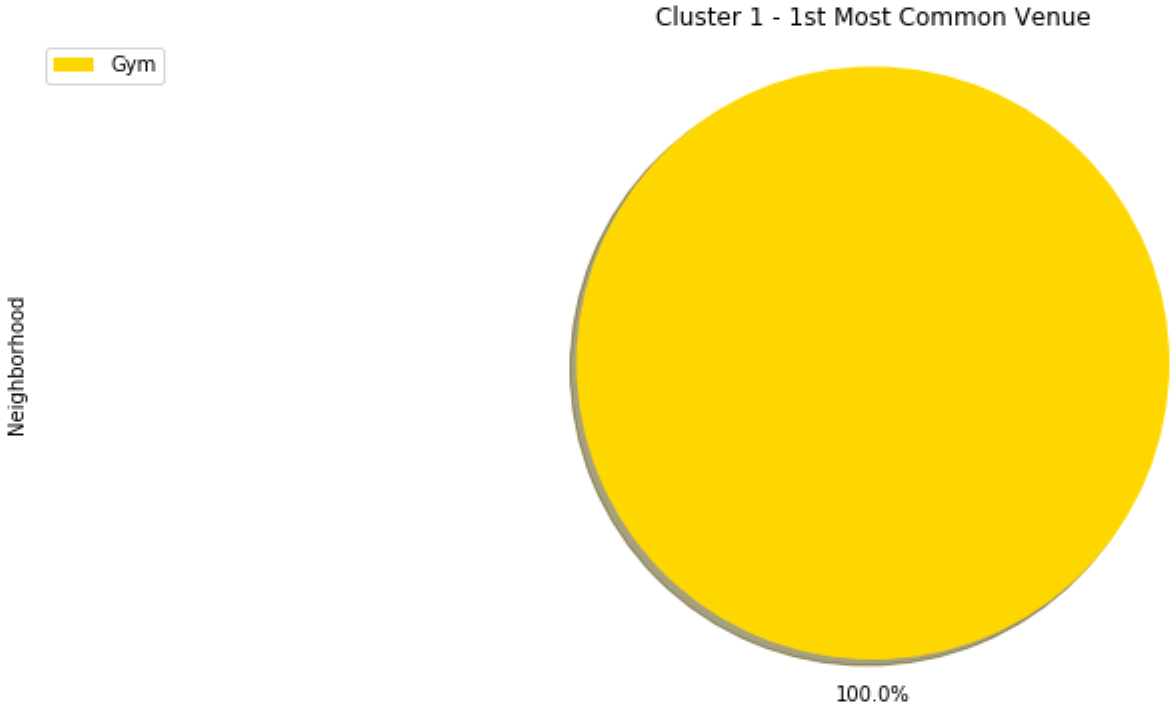
4.1. Cluster 0 – Places Foods

The most critical places of cases are the cluster 0 because in those neighborhoods there are many venues to conglomeration people, this in a concentration of food places like restaurants, bars, pubs and beach with more than 68.000 cases confirmed and more than 5.900 deaths per 100.000 persons.



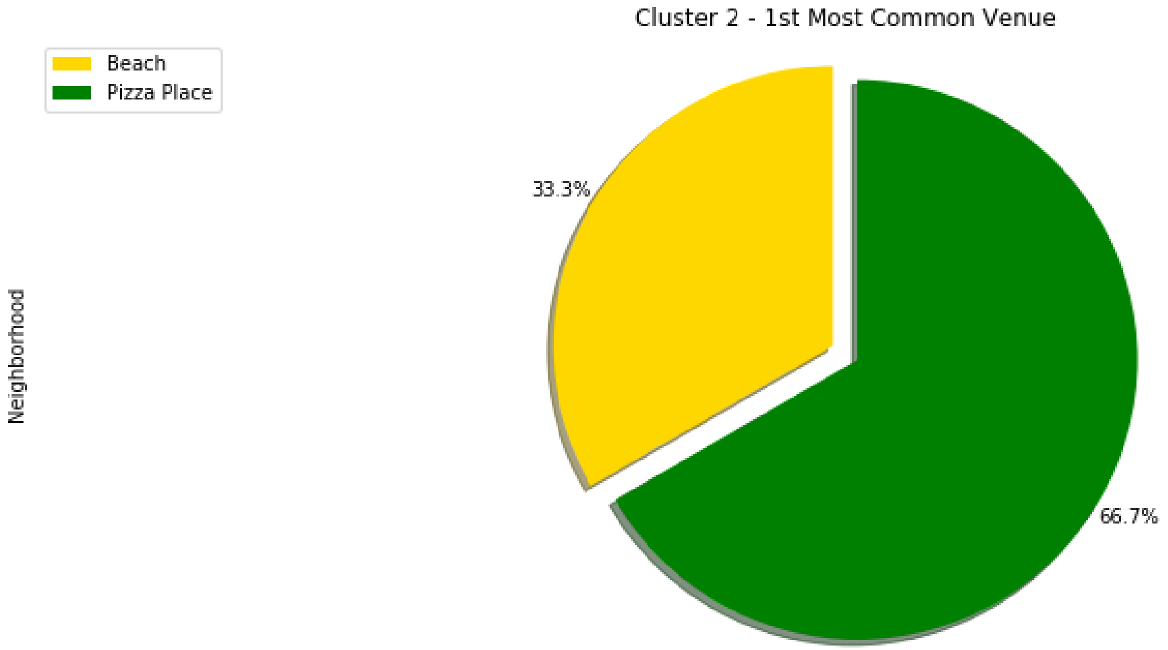
4.2. Cluster 1 – Gym/Bike Rental

In the cluster 1 is an important place: Gym, only in this neighborhood are more 5.000 cases per 100.000 persons. A gym is a cabin where the people produce perspiration and the manipulations of equipment is frequently. Another places like bike rental/bike shares is other focal point of infection for the manipulations of every bike was not with the best sanitation.



4.3. Cluster 2 – Beach/Food Place

For the last cluster, we can identify the beach as a focal venue of infection and also the restaurants. In this cluster we have more than 11.000 cases per 100.000 persons and more than 1.100 deaths.



5. Conclusions

The purpose of these study is to identify the common places of New York city with the most infections on the neighborhoods. This information can help to the government of the state to know which are the venues where can prevent other contagious with the measures of social distancing, biosecurity, sanitization and disinfection particularly for manipulations of any equipment.

Other important point is a strong campaign of informative in these venues to prevent more infected and deaths with covid-19.

Also, these analyses can help to others cities to know more the behavior of the specific neighborhood that can help to improve the public health system, education of social distancing preventions and get the best hygiene in every place.

6. Future Directions

The most important for this analysis is first to the New York State Government to do the best campaigns in the focal points infected places to social distancing, sanitizations, etc. The second is for the people, these persons who live in the neighborhoods can make a prevent measures to not increment the contagious. I think the best vacuum for Covid-19 is the education and the prevention.