Capstone Project
Battle of Neighborhoods
Report

Diogo Quintão

May 2020

1. Introduction

1.1 Background

The United States of America is one of the countries with the highest rate of obesity in the world, both for adults and for children. It is estimated that about 71% of the American population over 15 years of age is obese or overweight. In children, the situation is more serious since the obesity rate is 43%, the highest value worldwide. An OECD report published in 2019 highlights obesity as an important risk factor for many chronic diseases, including diabetes, cardiovascular disease and cancer.
To promote a decrease in these values, it is necessary to implement some changes in lifestyle, namely with regard to the consumption of fast food, obesity and physical inactivity. Gyms are one of the "weapons" to combat these aspects, since they promote healthy eating habits and physical activity.

1.2 Problem

Having said that, and taking New York City as a sample, I intend to get an answer to the following problems:

- Compare the number of Fast Food restaurants with the number of Gyms in the different boroughs of New York;
- What are the neighborhoods with higher/less number of gyms?
- What are the neighborhoods with higher/less number of fast food restaurants?
- In how many neighborhoods there are no gyms?
- What is the best neighborhood to live in if you like going to gyms?
- What is the best neighborhood to open a new gym?

1.3 Interest

People or companies that want to open their business, must have prior knowledge about the possible place where they can or cannot proceed with their project. This work focuses on gyms, but with the necessary changes it can be adapted to another type of establishment or business.

2. Data

2.1. Data Acquisition

The data acquired for this final capstone was collected from two different sources. The first data source (https://cocl.us/new_york_dataset), contains data about New York City. Through this dataset it was possible to obtain a dataframe with the boroughs, neighborhoods of New York and the respective latitudes and longitudes.
The second source of data comes from Foursquare API. Through this API it was possible to obtain all venues in each of the neighborhoods and subsequently filter by gyms and fast food restaurants in order to make an exploratory analysis by neighborhood.
This dataset, by having the locations of every neighborhood of New York City and the respective points of interest, satisfies the requirements needed to perform the analysis above described.

2.2. Data Cleaning and Data Preparation

The exploratory analysis was achieved by answering to the exploratory questions defined in the Problem section. To do that the dataset needed to pass through some pre-processing tasks:

- Foursquare API

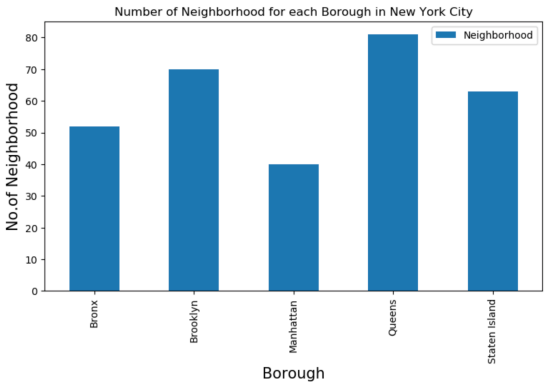
Using the dataset containing the neighborhoods in New York City along with the latitude and longitude, we were able to find all venues within a 1000 meters radius of each neighborhood by connecting to the Foursquare API. With this API we got all the venues in each neighborhood along with their coordinates and category.
We needed to get from the dataset all the points of interest related to the general category gym and fast food restaurant because for making the analysis easier we should have it in a unique column.
It was noticed that we had many types of fast food restaurants and gyms. We needed to ensure that they are all grouped in the same category in order to explore the neighborhoods with the certain characteristics of the two categories: Gyms and Fast Food Restaurant. Like described below:
Fast Food = Fast Food Restaurant + Pizza Place
Gym = Gym / Fitness Center + Sports Club + Gym Pool + Pilates Studio

- One hot encoding

Next, it was applied the one hot encoding technique because we had a categorical variable named 'Venue Category' that is related to the most valuable information in this dataset. Our analysis wouldn’t be rich if we haven’t transformed it into a numerical variable (or many) in order to perform some calculation. So, we transform the categorical column in many numerical columns (equal to the total number of unique categories). It was achieved by using the dummies function.

3. Exploratory Data Analysis

Number of neighborhoods by borough



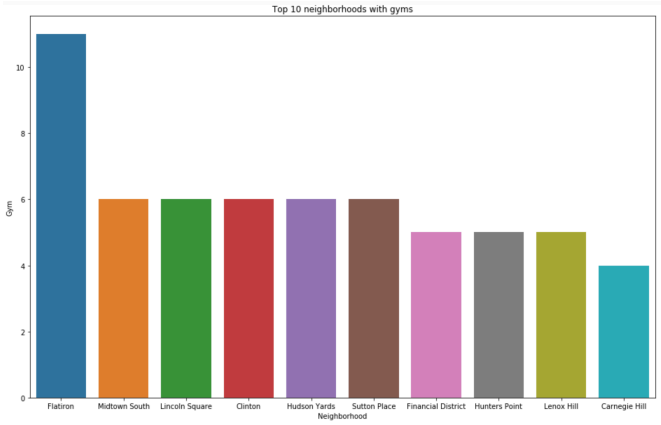

```
new_york_data['Borough'].value_counts()

Queens          81
Brooklyn        70
Staten Island   63
Bronx           52
Manhattan       40
Name: Borough, dtype: int64
```

Fig. 1 - Number of neighborhoods by borough

From the above analysis, we can see that Queens is the borough that has highest number of neighborhoods.

Top 10 neighborhoods with gyms

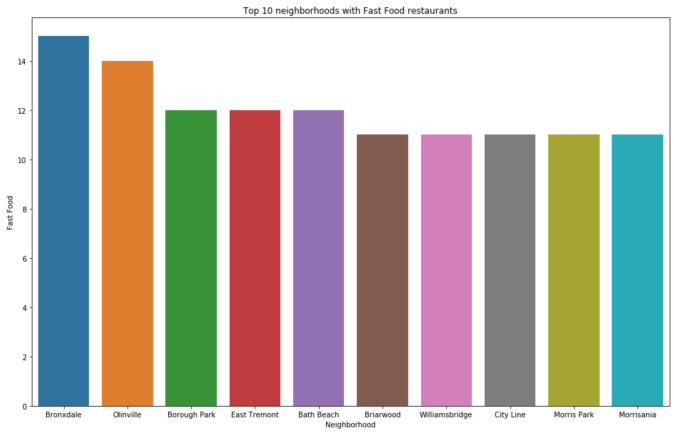


	Neighborhood	Gym
97	Flatiron	11
175	Midtown South	6
153	Lincoln Square	6
57	Clinton	6
135	Hudson Yards	6
267	Sutton Place	6
95	Financial District	5
137	Hunters Point	5
151	Lenox Hill	5
43	Carnegie Hill	4

Fig. 2 - Top 10 neighborhoods with gyms

From the bar chart shown above, we can see that Flatiron is the neighborhood with the largest number of gyms, eleven in total.

Top 10 neighborhoods with Fast Food restaurants



	Neighborhood	Fast Food
34	Bronxdale	15
203	Olinville	14
28	Borough Park	12
80	East Tremont	12
9	Bath Beach	12
30	Briarwood	11
292	Williamsbridge	11
52	City Line	11
181	Morris Park	11
182	Morrisania	11

Fig.3 - Top 10 neighborhoods with Fast Food restaurants

From the bar chart shown above, we can see that Bronxdale is the neighborhood with the largest number of fast food restaurants, fifteen in total.

Comparison between gyms and fast food restaurants by neighborhoods

	Neighborhood	Fast Food	Gym
34	Bronxdale	15	0
203	Olinville	14	1
28	Borough Park	12	0
80	East Tremont	12	1
9	Bath Beach	12	1
30	Briarwood	11	0
292	Williamsbridge	11	0
52	City Line	11	0
181	Morris Park	11	1
182	Morrisania	11	1

Fig.4 - Comparison between gyms and fast food restaurants by neighborhoods

The table illustrates the number of gyms in the top 10 neighborhoods with highest number of fast food restaurants and we can see that the contrast that exists between the number of fast food restaurants and the number of gyms is huge.
We discovered that in 302 neighborhoods only in 20 of them the number of gyms is superior to the number of fast food restaurant. And, even more shocking from the 302 neighborhoods there are 157 neighborhoods that don't have gyms.

Neighborhoods with and without gyms

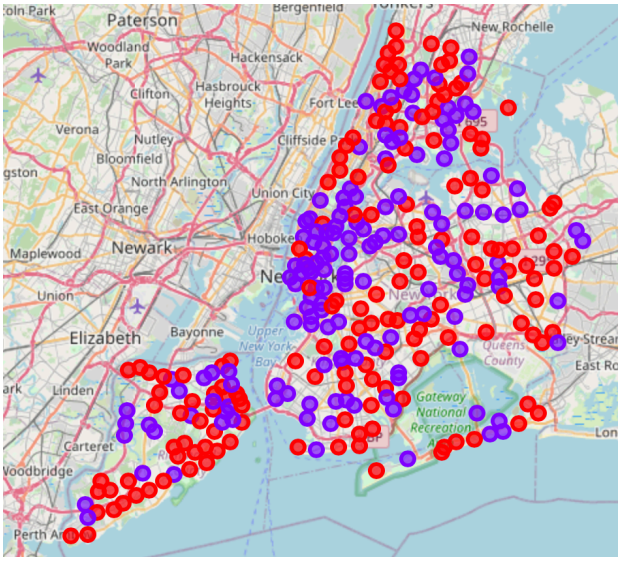
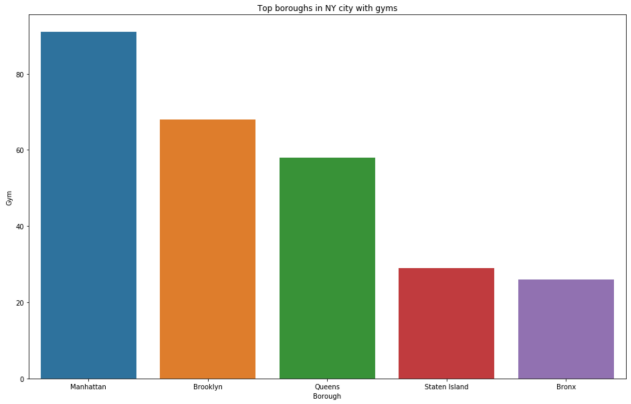


Fig. 5 - Neighborhoods with and without gyms

After this, we decided to explore where the neighbors with gyms and without gyms are located by mapping it geographically. In the right figure, in purple we can visualize all the neighbors that have gyms and in red all the neighborhoods that don't have gyms.
We can conclude that close to the center of NY there is a higher number of neighborhoods with gyms and close to the periphery there is a lower number of neighborhoods with gyms. We also see that there are some neighbors that don't have gyms and aren't near to any neighbor that have gyms, specially in Staten Island and close to the sea.

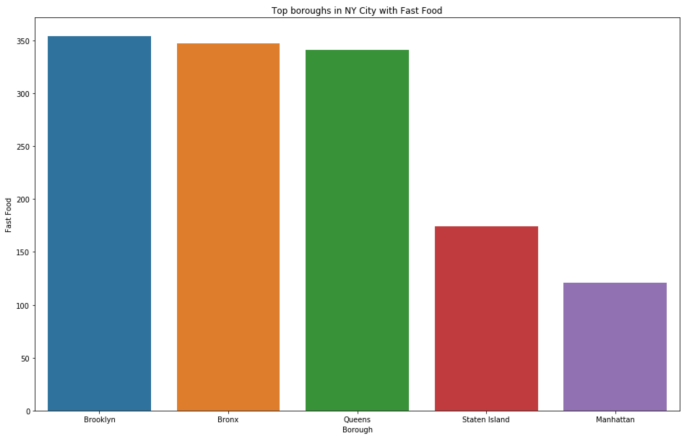
Top boroughs with gyms



	Borough	Gym
2	Manhattan	91.0
1	Brooklyn	68.0
3	Queens	58.0
4	Staten Island	29.0
0	Bronx	26.0

Fig. 6 - Top boroughs with gyms

Top boroughs with Fast Food restaurants



	Borough	Fast Food
1	Brooklyn	354.0
0	Bronx	347.0
3	Queens	341.0
4	Staten Island	174.0
2	Manhattan	121.0

Fig. 7 - *Top boroughs with Fast Food restaurants*

Comparison of the number of restaurants with the number of gyms by Borough

	Borough	Fast Food	Gym
1	Brooklyn	354.0	68.0
0	Bronx	347.0	26.0
3	Queens	341.0	58.0
4	Staten Island	174.0	29.0
2	Manhattan	121.0	91.0

Fig. 8 - *Comparison of the number of restaurants with the number of gyms by Borough*

From the graphs, we can see that the Borough with the largest number of gyms is Manhattan, with a total of 97 gyms. In terms of fast food restaurants, Brooklyn is the Borough with the highest number, 354 in total and Manhattan is the borough with a smaller number of fast food restaurants which is a good finding because it was the borough with highest number of gyms. The Bronx case is more critical because is one of the boroughs with higher number of fast food restaurants but is the borough with a smaller number of gyms.

4. Clustering

To find similar neighborhoods we will use K-means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size.
A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data is going to be clustered. The Elbow Method is one of the most popular methods to determine the optimal value of k. It performs k-means for a range of Ks, and we can see the K for which the Euclidean distance between each example and the corresponding centroid is minor.

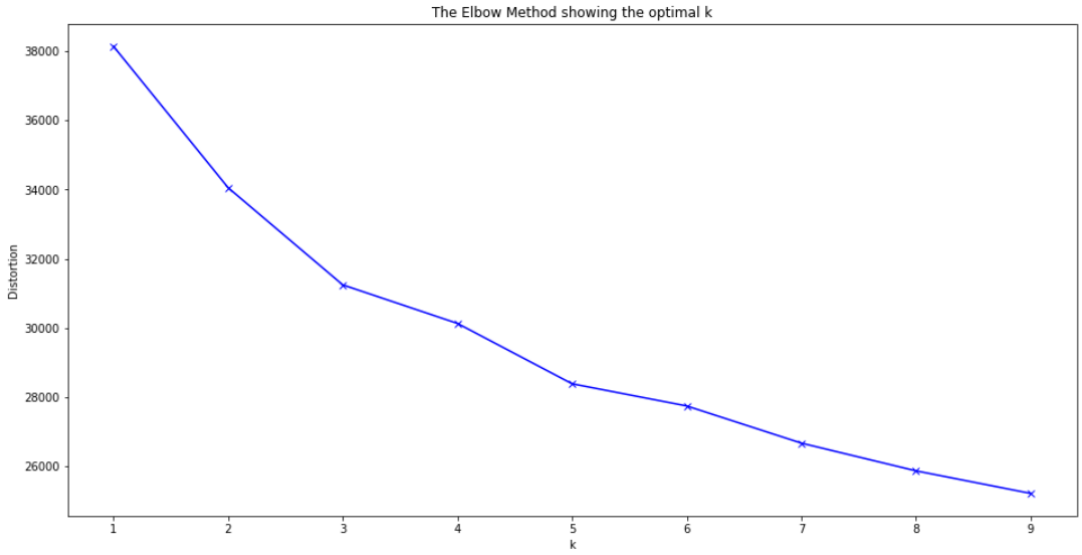


Fig. 9- *Elbow Method*

From this step we choose a K = 5 to start the analysis. The K =3 seemed also like a good approach but we wanted to have a more uniform distribution of all our 302 neighborhoods.

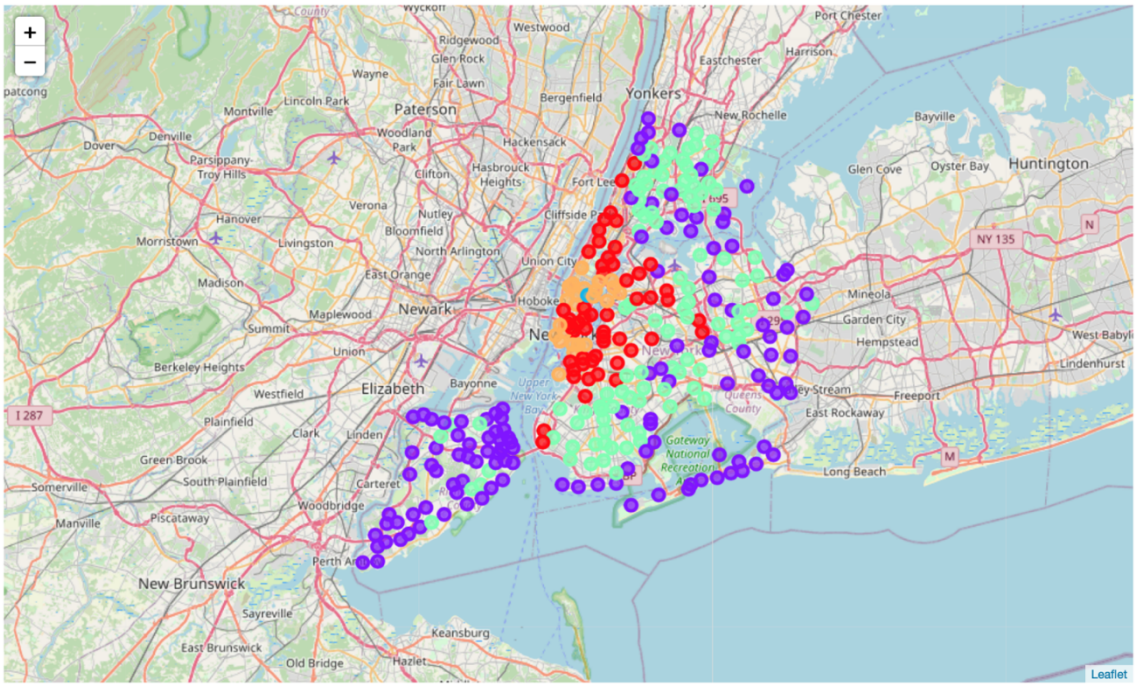


Fig. 10- Distribution of Neighbourhoods by cluster label: In red it is represented cluster 0, in purple cluster 1, in blue cluster 2, in green cluster 3 and in orange cluster 4

To analyze each cluster in order to answer all our initial questions we need to find the mean number of gyms and fast food in general and compare to the values achieved for each cluster. To compare we use the describe function on which cluster.

Mean and standard deviation of the number of gyms in New York city:

```
In [72]: ny_grouped['Gym'].mean()
Out[72]: 0.8675496688741722

In [73]: ny_grouped['Gym'].std()
Out[73]: 1.3624069802804275
```

Mean and standard deviation of the number of fast food restaurants in New York city:

```
In [74]: ny_grouped['Fast Food'].mean()
Out[74]: 4.347682119205298

In [75]: ny_grouped['Fast Food'].std()
Out[75]: 3.0195890776096364
```

Starting from the first cluster, which comprise 53 neighborhoods has a mean of gyms (1,28) superior comparing to the mean of New York City. Regarding the mean of fast food restaurants (3,91), this is lower when comparing to the mean of New York City.

```
In [78]: describe_cluster0['Gym']

Out[78]: count    53.000000
         mean      1.283019
         std       1.349891
         min       0.000000
         25%       0.000000
         50%       1.000000
         75%       2.000000
         max       5.000000
         Name: Gym, dtype: float64
```

```
In [79]: describe_cluster0['Fast Food']

Out[79]: count    53.000000
         mean      3.905660
         std       1.832024
         min       1.000000
         25%       3.000000
         50%       3.000000
         75%       5.000000
         max       8.000000
         Name: Fast Food, dtype: float64
```

Cluster 1, which is the largest of the clusters comprising 123 neighborhoods has a mean of gyms (0,37) and fast food restaurants (2,48) inferior comparing to the mean of New York City.

```
In [81]: describe_cluster1['Gym']

Out[81]: count    123.000000
         mean       0.373984
         std        0.657845
         min        0.000000
         25%        0.000000
         50%        0.000000
         75%        1.000000
         max        3.000000
         Name: Gym, dtype: float64
```

```
In [82]: describe_cluster1['Fast Food']

Out[82]: count    123.000000
         mean       2.479675
         std        1.780355
         min        0.000000
         25%        1.000000
         50%        2.000000
         75%        4.000000
         max        6.000000
         Name: Fast Food, dtype: float64
```

Cluster 2, which covers only one neighborhood has a mean of gyms (4) and fast food restaurants (9) superior comparing to the mean of New York City.

```
In [84]: describe_cluster2['Gym']

Out[84]: count    1.0
         mean     4.0
         std      NaN
         min      4.0
         25%      4.0
         50%      4.0
         75%      4.0
         max      4.0
         Name: Gym, dtype: float64
```

```
In [85]: describe_cluster2['Fast Food']

Out[85]: count    1.0
         mean     9.0
         std      NaN
         min      9.0
         25%      9.0
         50%      9.0
         75%      9.0
         max      9.0
         Name: Fast Food, dtype: float64
```

Cluster 3, which comprise 102 neighborhoods has a mean of gyms (0,71) inferior comparing to the mean of New York City. The mean of fast food restaurants in this clusters (7,35) is higher when compared to the mean of New York City.

```
In [87]: describe_cluster3['Gym']

Out[87]: count    102.000000
         mean       0.705882
         std        0.839475
         min        0.000000
         25%        0.000000
         50%        1.000000
         75%        1.000000
         max        4.000000
         Name: Gym, dtype: float64
```

```
In [88]: describe_cluster3['Fast Food']

Out[88]: count    102.000000
         mean       7.352941
         std        2.476212
         min        3.000000
         25%        5.000000
         50%        7.000000
         75%        9.000000
         max       15.000000
         Name: Fast Food, dtype: float64
```

Last cluster, cluster 4, which comprise 23 neighborhoods has a mean of gyms (3,1) superior comparing to the mean of New York City. The mean of fast food restaurants in this clusters (1,83) is lower when compared to the mean of New York City.

```
In [90]: describe_cluster4['Gym']

Out[90]: count    23.000000
         mean      3.130435
         std       2.784769
         min       0.000000
         25%       1.000000
         50%       2.000000
         75%       5.500000
         max      11.000000
         Name: Gym, dtype: float64
```

```
In [91]: describe_cluster4['Fast Food']

Out[91]: count    23.000000
         mean      1.826087
         std       1.266785
         min       0.000000
         25%       1.000000
         50%       2.000000
         75%       3.000000
         max       4.000000
         Name: Fast Food, dtype: float64
```

By analyzing the results, we can see that the worst cluster is number 3 because it has a low mean number of gyms and a high mean number of fast food restaurants. The best cluster would be 4 because it has high mean number of gyms and low mean number of fast food restaurants. We can see that most of the neighborhoods in cluster 3 are distributed in Bronx, Queen and Brooklyn and most neighborhoods in cluster 4 are distributed in Manhattan. The cluster with lower mean number of gyms is cluster 1 that have most of its neighborhoods distributed in Staten Island.
For a person that likes gyms and lives a healthy lifestyle, the best neighborhoods to live are from neighborhoods from the cluster 4, especially in Manhattan and if we want to open a new gym cluster 1 would be a good option, especially in Staten Island.

5. Conclusion

In resume:

- **Compare the number of Fast Food restaurants with the number of Gyms in the different boroughs of New York;** The contrast that exists between the number of fast food restaurants and the number of gyms is huge. We discovered that in 302 neighborhoods only in 20 of them the number of gyms is superior to the number of fast food restaurant.
- **What are the neighborhoods with higher/a smaller number of gyms?** Flatiron is the neighborhood with the largest number of gyms, ten in total. We cannot name one neighborhood with a smaller number of gyms because there are a lot of neighborhoods without gyms.
- **What are the neighborhoods with higher/a smaller number of fast food restaurants?** Bronxdale is the neighborhood with the largest number of fast food restaurants, fifteen in total. For the smaller number of fast food restaurants it happens the same as before because there are 23 neighborhood with no fast food restaurants.
- **In how many neighborhoods there are no gyms?** There are 157 neighborhoods that don't have gyms. The number is higher than from the neighborhoods without fast food restaurants.
- **What is the best neighborhood to live in if you like going to gyms?** We conclude that for a person that likes gyms and lives a healthy lifestyle, the best neighborhoods to live are from neighborhoods from the cluster 4, especially in Manhattan.
- **What is the best neighborhood to open a new gym?** If we want to open a new gym, neighborhoods from cluster 1 would be a good option, especially the ones distributed in Staten Island.

A future analysis that should be interesting to be done to get a more conclusive result should be to add demographic and economic data about New York to this dataset, such as number of inhabitants per neighborhood, scholar degree, economic status.