Fighting Obesity in New York

The Battle of Neighborhoods – Part 2

Agenda

1. The problem

The **United States of America is one of the countries with the highest rates of obesity in the world**. 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 like diabetes, cardiovascular disease and cancer.

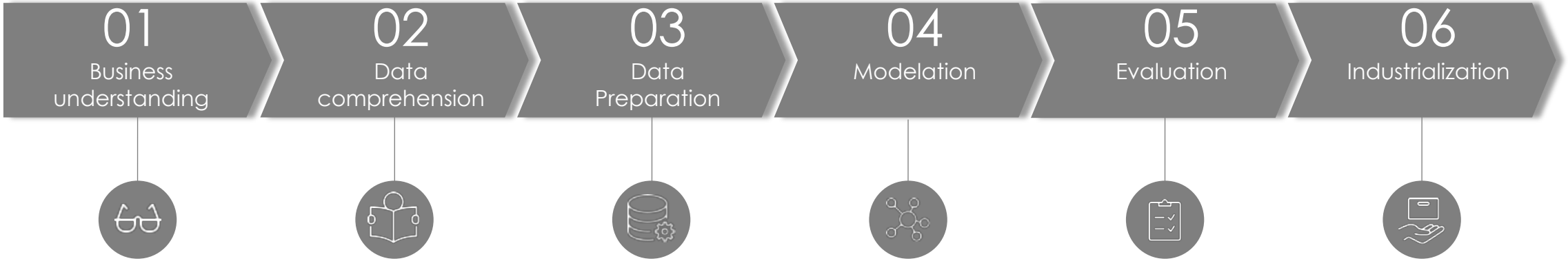
To promote a decrease in these values, it is necessary to implement some changes in lifestyle, namely with regard to the consumption of fast food, obesity and physical inactivity. **Gyms are one of the "weapons"** to combat these aspects, since they promote healthy eating habits and physical activity.

Having said that, and taking **New York City** as a sample, the goal of this project is 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 are the 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?

2. Approach

The approach used for the development of this project was based on the CRISP-DM methodology presented above:



Understanding the objectives of the project to convert the knowledge acquired into a Data Science problem

Initial data assessment and quality analysis, using descriptive and exploratory techniques

Data transformation to build the final dataset to be presented to the models

Model development

Evaluation and quality revision, validation if the results are the expected regarding the project objectives

Industrialize to turn the acquired knowledge into a solution (this step wasn't performed in this project example)

2.1. Dataset

The dataset used for this project is composed by the following data:

- **New York City data** that contains list Boroughs, Neighbourhoods along with their latitude and longitude.
  - Data source : https://cocl.us/new_york_dataset
  - Description: This data set contains the required information. And we will use this data set to explore various neighbourhoods of New York City.

- **Venues data** per neighbourhood
  - Data source : Foursquare API
  - Description: By using this API we will get all the venues in each neighbourhood. We can filter these venues to get only Gyms and Fast Food Restaurants.

2.2. Exploratory Analysis

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 de 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

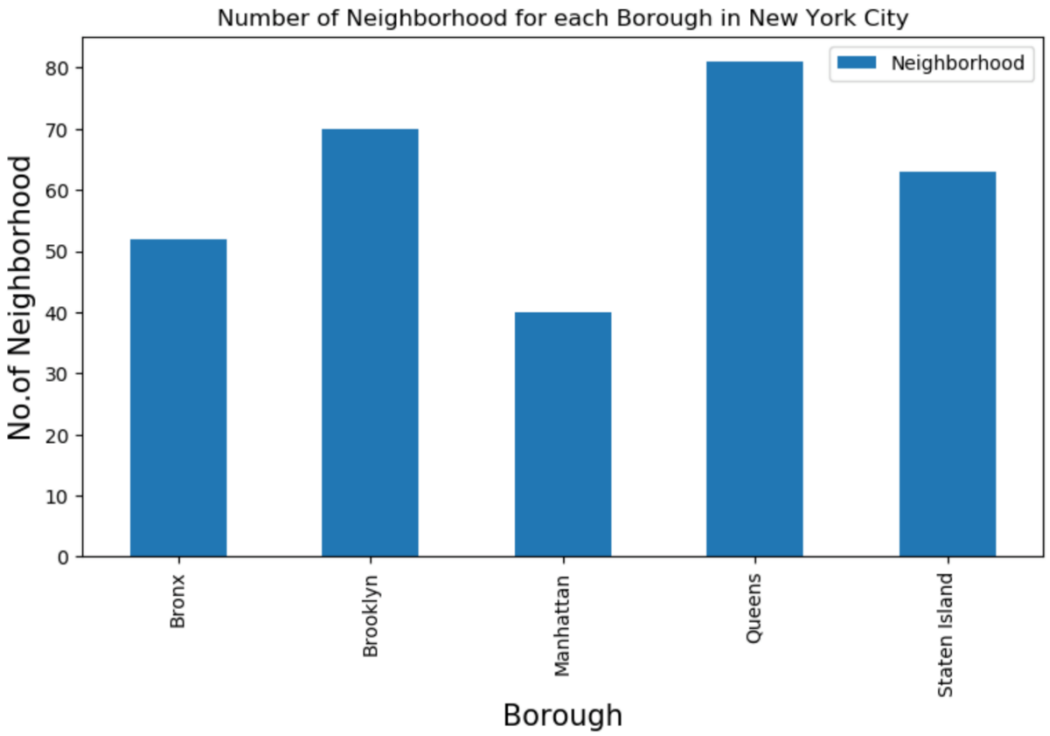
- ***One hot encoding***

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.

2.2. Exploratory Analysis

After preparing the dataset, the exploratory Analysis started.

Number of neighborhoods by borough



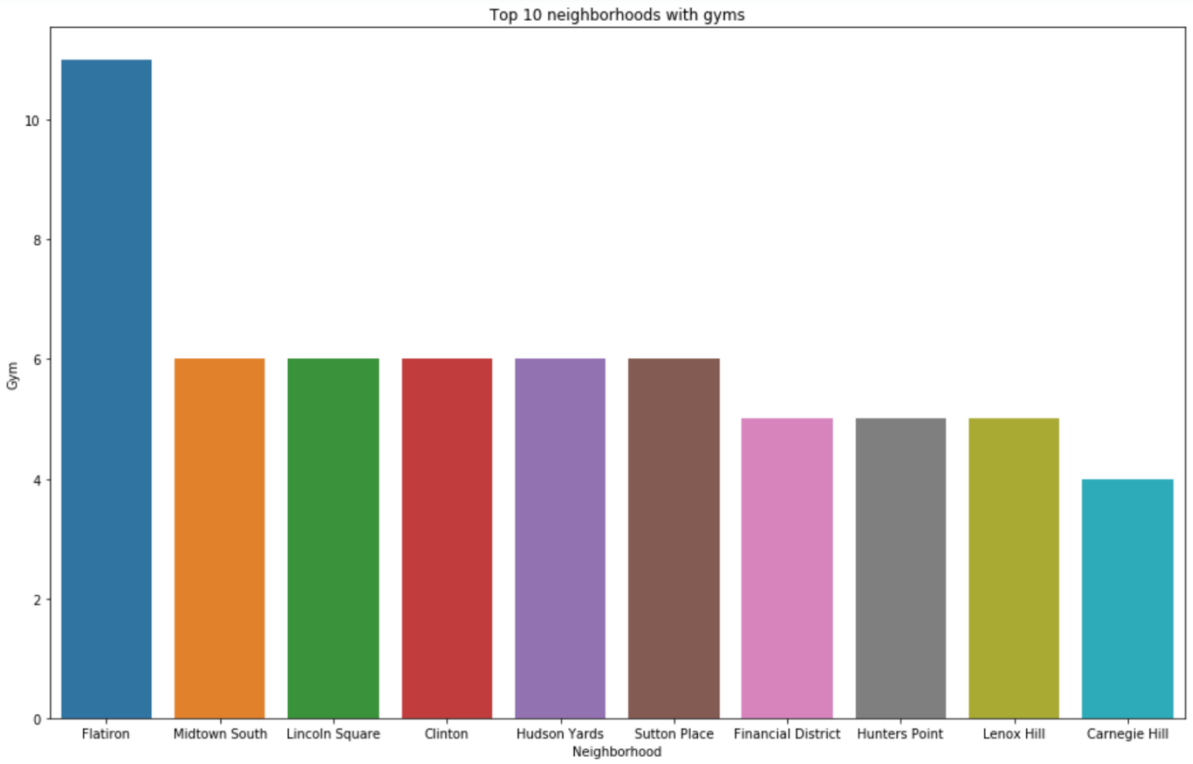

```
new_york_data['Borough'].value_counts()
```

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

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

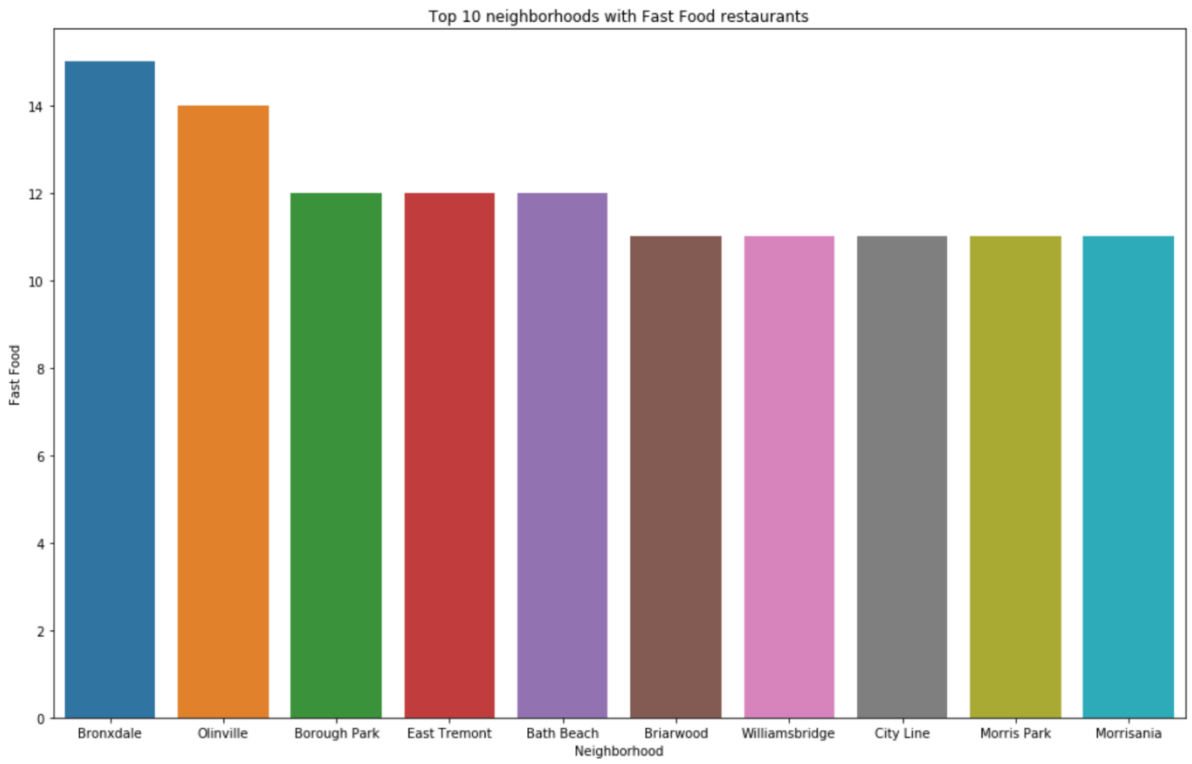
2.2. Exploratory Analysis

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



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.

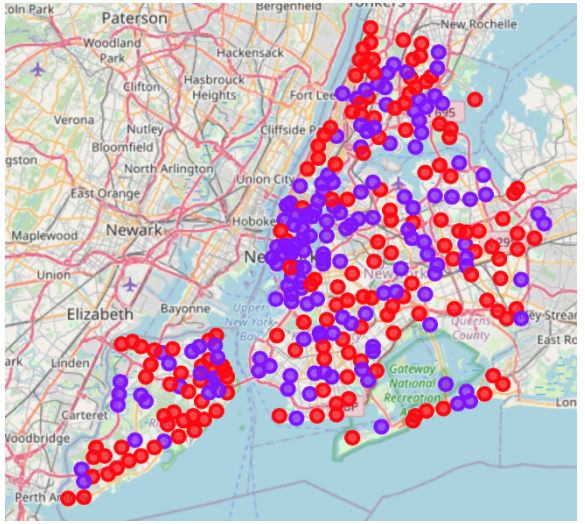
2.2. Exploratory Analysis

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

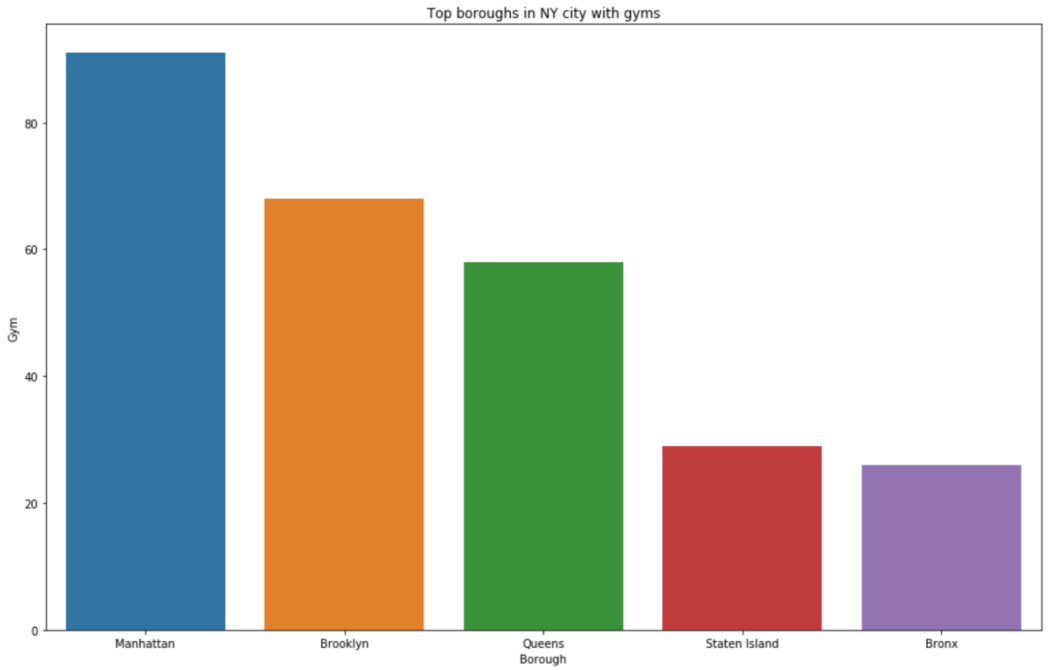
The table illustrates the number of gyms in the top 10 neighborhoods with highest number of fast food restaurants and we can see that the contrast that exists between the number of fast food restaurants and the number of gyms is huge.
We discovered that **in 302 neighborhoods only in 20 of them the number 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**.

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.

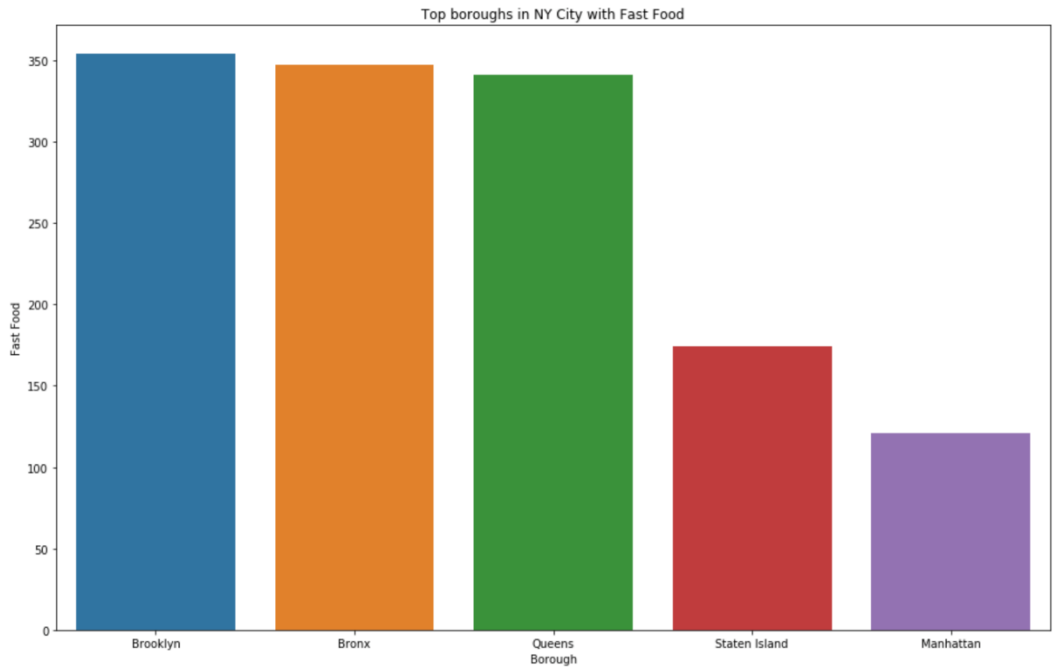


2.2. Exploratory Analysis

Top boroughs with gyms



Top boroughs with fast food restaurants



From the graph on the left, 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 less 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 less number of gyms.

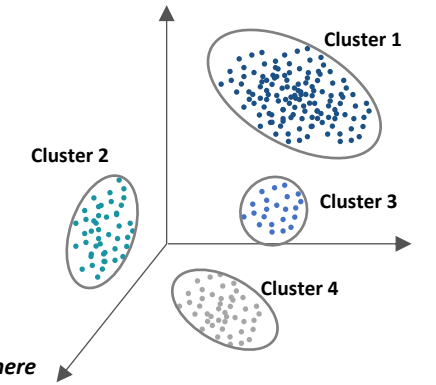
2.3. Clustering

Next, it is presented the explanation for developing the clustering task:

Clustering algorithms can be applied on top of the collected data in order to identify groups or patterns in the data presented. For example - the segmentation of neighbors based on their vendor categories can be considered an unsupervised learning problem and as such, the *k-means* clustering method can be applied. This helps to answer to the questions described on the right panel:

Methodology:

1. *Define the nº of clusters (K)*
2. *Randomly assign K centroids*
3. *Each exemple from the dataset is assigned to the closest centroid*
4. *Centroids are recalculated until there are no more changing regarding the mean point of the cluster*

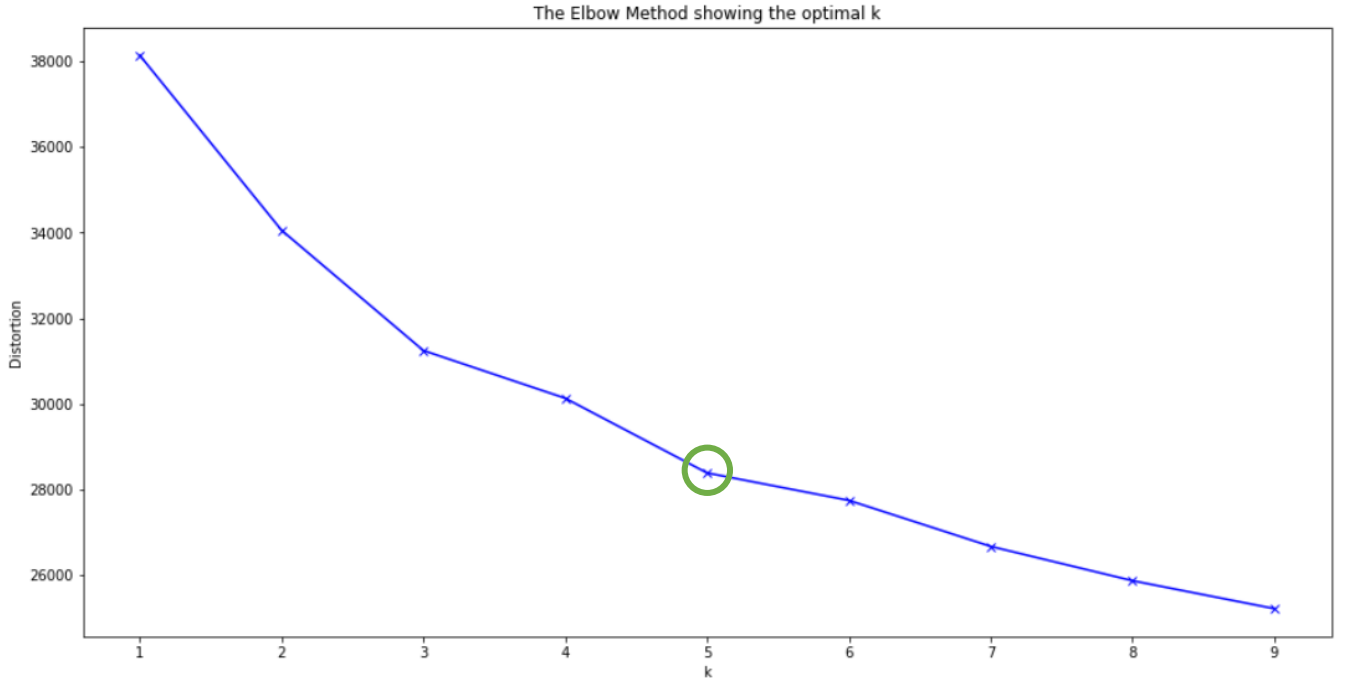


Ilustrative

1. Where people should stay in New York if they are gym fans?
2. What would be a good neighbor to build a new gym?

2.3. Clustering

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 the each example and the corresponding centroid is minor.

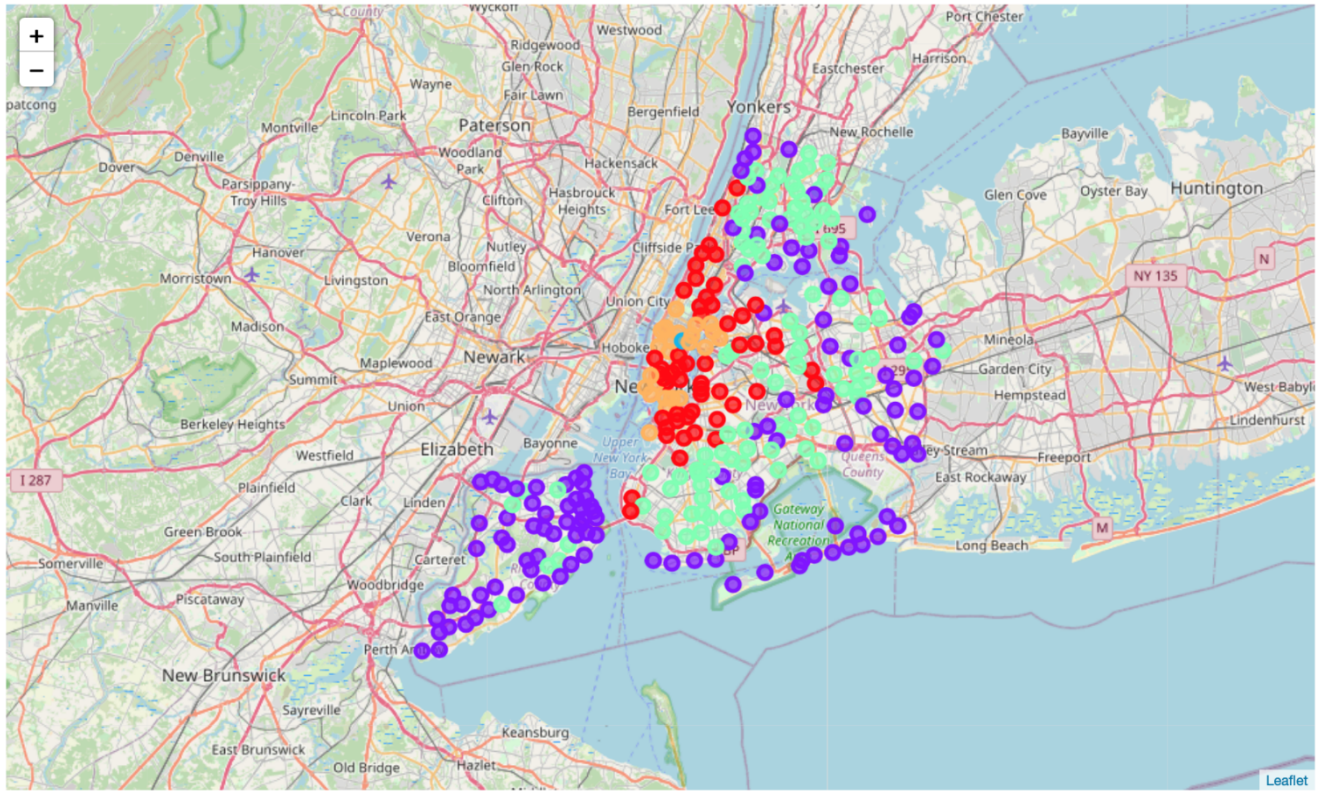


From this step we choose a K = 5 to start the analysis. The K =3 seemed also like a good approach but we wanted to have a more uniform distribution of all our 302 neighborhoods.

2.3. Clustering

After applying k-means clustering with k = 5, we got the following distribution of Neighborhoods by cluster label:

Label	Nº of Neighborhoods
0	53
1	123
2	1
3	102
4	23



In red it is represented cluster 0, in purple cluster 1, in blue cluster 2, in green cluster 3 and in orange cluster 4

2.3. Clustering

By filtering the dataset by the cluster label, it was possible to analyse each cluster individual and answer to the questions that were left. It was done by finding the mean number of gyms and fast food in general and compare to the values achieved for each cluster.

	Gym	Fast Food
General	0,87	4,35
Cluster 0	1,28	3,91
Cluster 1	0,37	2,48
Cluster 2	4,00	9,00
Cluster 3	0,71	7,35
Cluster 4	3,13	1,83

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.

3. Conclusions

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/less number of gyms?** Flatiron is the neighborhood with the largest number of gyms, ten in total. We cannot name one neighborhood with less number of gyms because there are a lot of neighborhoods without gyms.
- **What are the neighborhoods with higher/less number of fast food restaurants?** Bronxdale is the neighborhood with the largest number of fast food restaurants, fifteen in total. For the less 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 and economic status.