# OPENING A NEW PIZZA PLACE IN NYC

IBM Applied Data Science Capstone – Week 5

## Introduction

New York, among other things, is famous for its pizza that is being sold in many pizza places. Therefore if someone was willing to open a new pizza place they would have to face a tough competition. With this in mind, the use of data analysis in order to determine which could be the best spot to open such restaurant is a crucial step for its success.

## Data

The data analysis will be based on data from the following sources:

1. Geographical data of the city of New York (https://cocl.us/new_york_dataset)

2. Foursquare API

From 1. we retrieve the list of neighborhoods of the city of New York with their geographical coordinates.
We will use this information to retrieve venues in each neighborhood that fall under the *Food* category in the Foursquare database, then we will further restrict ourselves to the *Pizza Place* category to pursue our goal.

## Methodology

The first step is to retrieve the geographical data of the city of New York at the aforementioned link and then encode it in a pandas dataframe.
In particular we will obtain a dataframe whose columns are:

1. Neighborhood

2. Latitude

3. Longitude

Such dataframe is then the basis to retrieve all the venues that are elements of the Foursquare food category, obtained by specifying the categoryId=4d4b7105d754a06374d81259

We limit ourselves to food venues because we assume that they are the most relevant data for grouping different neighorhoods into different clusters.

To do so we use the KMeans method and we establish, via the elbow method, that the most suitable number of clusters (more on this in the conclusion section).

Once we have put all the neighborhoods into food-based clusters we go on to analyze the frequency of pizza places in each neighborhood.

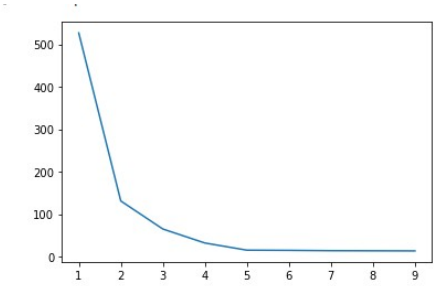
We finally use box plots to identify the presence of eventual outliers where the frequency appears to be far lower than the distribution constructed out of the cluster.

## Results and discussion

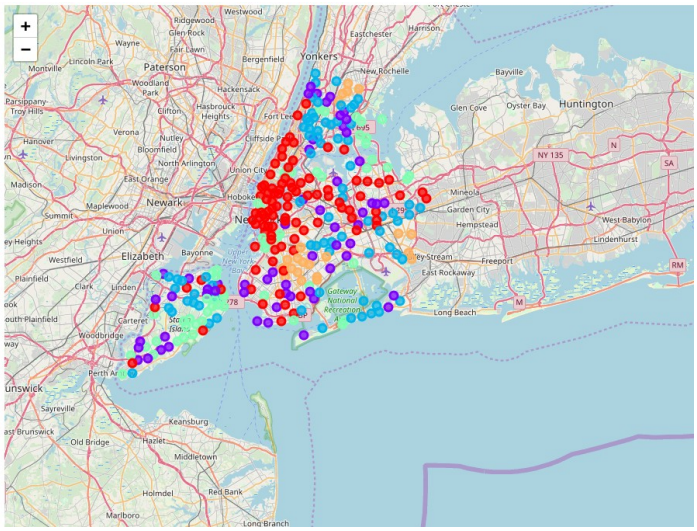
We obtain the following dataframe for the New York neighborhoods

| | Neighborhood | Latitude | Longitude |
|---|---|---|---|
| 0 | Wakefield | 40.894705 | -73.847201 |
| 1 | Co-op City | 40.874294 | -73.829939 |
| 2 | Eastchester | 40.887556 | -73.827806 |
| 3 | Fieldston | 40.895437 | -73.905643 |
| 4 | Riverdale | 40.890834 | -73.912585 |
| ... | ... | ... | ... |
| 301 | Hudson Yards | 40.756658 | -74.000111 |
| 302 | Hammels | 40.587338 | -73.805530 |
| 303 | Bayswater | 40.611322 | -73.765968 |
| 304 | Queensbridge | 40.756091 | -73.945631 |
| 305 | Fox Hills | 40.617311 | -74.081740 |

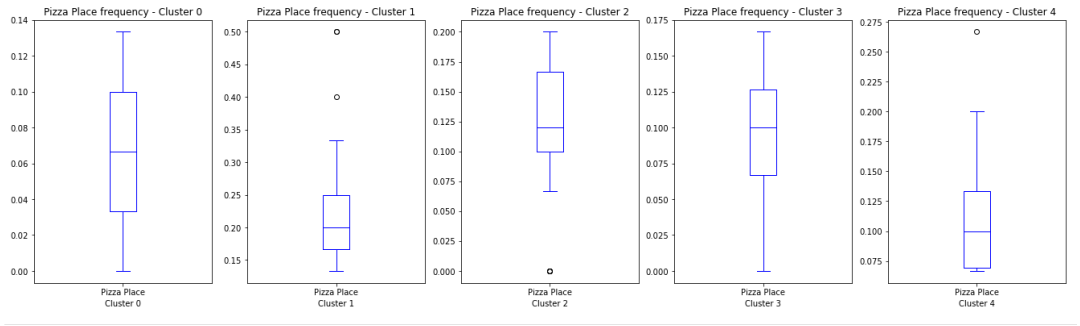
We determined, through the elbow method, that the best number of clusters for our analysis is 5.




We put the food-based clustered neighborhoods on a map:



We produced a box plot for every cluster obtaining the following picture:

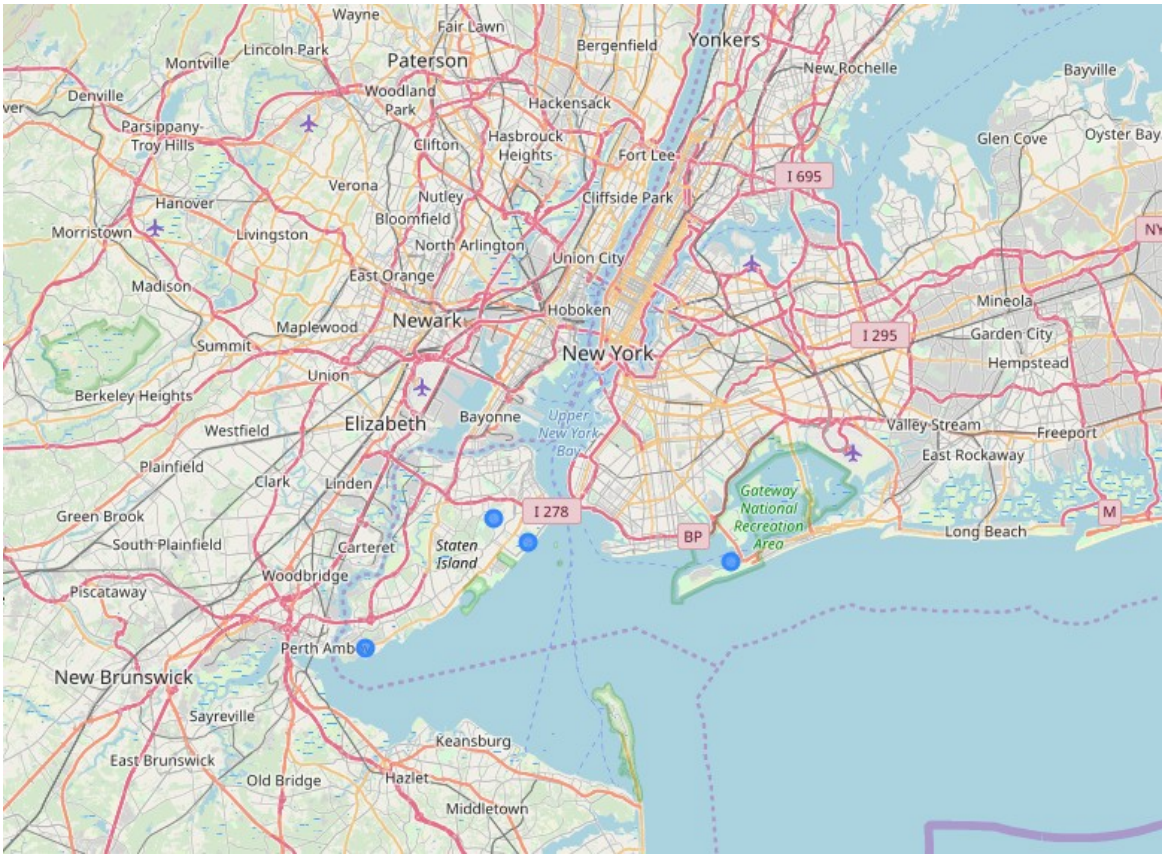



We see the presence of outliers only for Cluster 2.

We go check to what neighborhoods these outliers correspond, finding the following:

| | Neighborhood | Pizza Place |
|---|---|---|
| 12 | Butler Manor | 0.0 |
| 60 | Roxbury | 0.0 |
| 63 | South Beach | 0.0 |
| 67 | Todt Hill | 0.0 |

And we finally show them on a map:



# Conclusions

There are two interesting takeaways from this analysis. First, under our clustering assumptions, having zero pizza places in a neighborhood doesn’t necessarily mean that there should be one. Moreover, as could have been expected, it seems that the only neighborhoods in which a new pizza place wouldn’t face much competition is in quite peripheral areas.

Some observations are in order. We should note that there is some arbitrariness in the choice of the number of clusters. We chose 5 as it seemed to be the more sound but also 2,3,4 could be considered, based on the plot we showed above. Moreover, we should stress considering different clustering methods from the K-means could be a path worth taking.
Finally, there is a lot of room for improvement for this analysis, for example adding features like the price range for the pizza place to be opened and the one retrieved via the Foursquare API.