Using Folium, DBSCAN, and Foursquare for spatial analysis

1. Introduction

1.1 Background

Whenever a stakeholder would like to do an investment and open a new venue (restaurant, hotel, etc ...) in a city, he should know the competition he will be going against in that new region. It is very important to understand how is the status of the neighbourhoods is and how they are viewed by the public in general; are they luxurious and satisfying the needs of the customers or in general it is a neglected and dangerous neighborhood.

The data of the venues in a city is very important to understand whether if a specific region is a hot spot and is a competitive one, and if different venue owners are working hard to satisfy the customers, maintain loyalty, and good fame, as a result becoming more successful in time. An initial comprehensive view of the market is important for new people who are planning to enter the market and open a new business. On the other hand, it is important to know if venue owners in specific regions are not caring that much about their businesses as a result becoming touristically undermined. One more question arises here, that what if there are venues who are so famous that they are not affected by their surrounding based on the type of service or commodities that they offer. This is also a point to be investigated for a new "business owner to be". Also if the new investor would open a new venue near another famous ones, he should expect a very strong challenge and a fierce competetion, at the same time our new venue investor would expect a lot of new intrigued customers whom their loyalty could be won.

1.2 Problem

By gathering the data of the rating of the venues, we will be able to determine the quality of each venue, the category that it belongs to, and be able to realize if venues and regions are grouped based on their quality and distance.

1.3 Interest

Such an analysis could be of interest of both the private and public sector. Whenever the government, a private industry stakeholders, or small business investors decide to open a new venue , the region where that business will start will effect the outcomes. This type of analysis also can help scientists to understand the behavior of the people in specific regions, prove if gentrification or segregation exitst between poor and rich people.

2. Data acquisition and preprocessing

2.1 Data sources

To gather the data, we used Foursquare to retrieve the venues within a specific city and gathered the ratings of those venues and their catagories based on venue id values. Our example was based on New York city(lat, lng = [40.7127281, -74.0060152]), where we gathered the venues within a vicinity of 1 mile and 100 venues.

2.2 Lack of Data

One issue with Foursquare is to get the rating of venues, a free account can get you only 50 venues. So either you wait for 2 days to gather you data or you pay for the service and get more.

3. Exploratory Data Analysis

3.1 Prilimenary visualization

At a first glance it is difficult to grasp the distrubution of the venues and detect the differences due to the closeness of the ratings and existence of noise. The distribution is not extremely Guassian but yet again they have a certain distribution. Viewing the histogram of it can show us how this are:

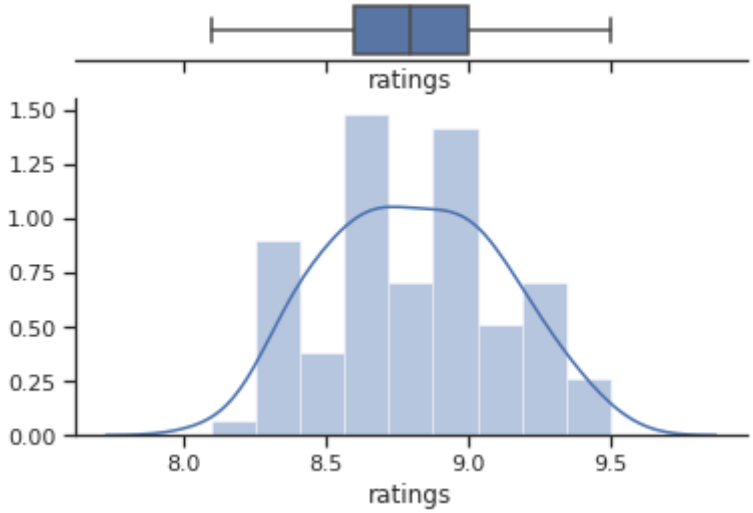

Fig1: Histogram of ratings with boxplot

Later if we visualize the distribution , again it is still difficult to understand the differences using scatter radius size or color as indicators of rating value due to it natural value and the noise:

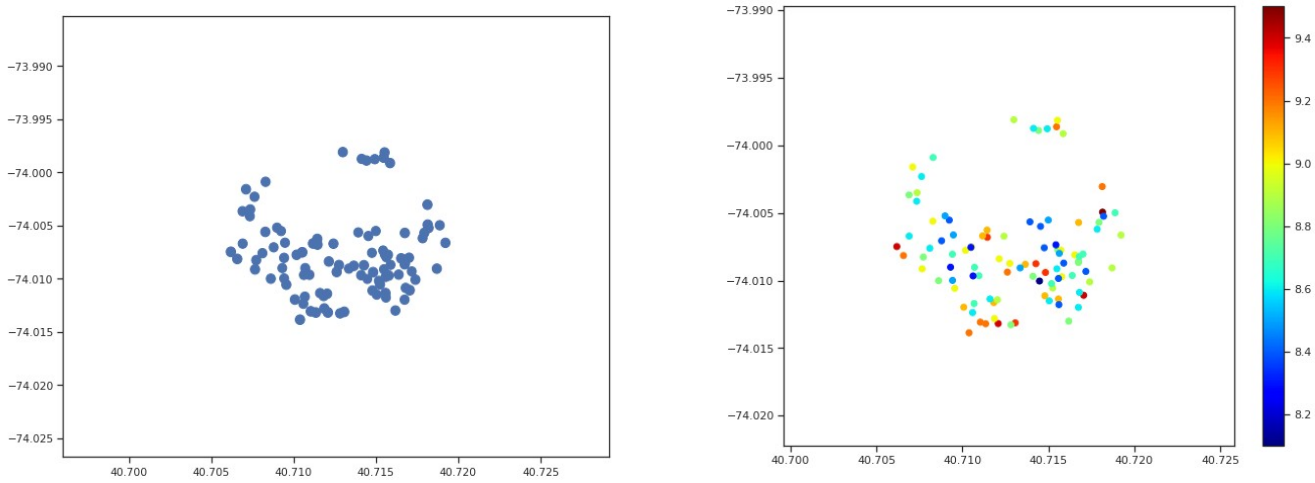

Fig 2: left: longitude versus latitude, and the radius of the dots being the ratings, right: plotted via coloration for rating

The solution is to normalize the data and apply a clustering technique that can detect the noise.

4. Normalization and Clustering

4.1 Normalization

To normalize the data we apply a min-max scaling:

$$Rating_{N_i} = \frac{Rating_i - Rating_{min}}{Rating_{max} - Rating_{min}}$$

Now the spatial spatial visualization can give a better sense of distribution where we can distiguish differences and the histogram has the same distribution structure.

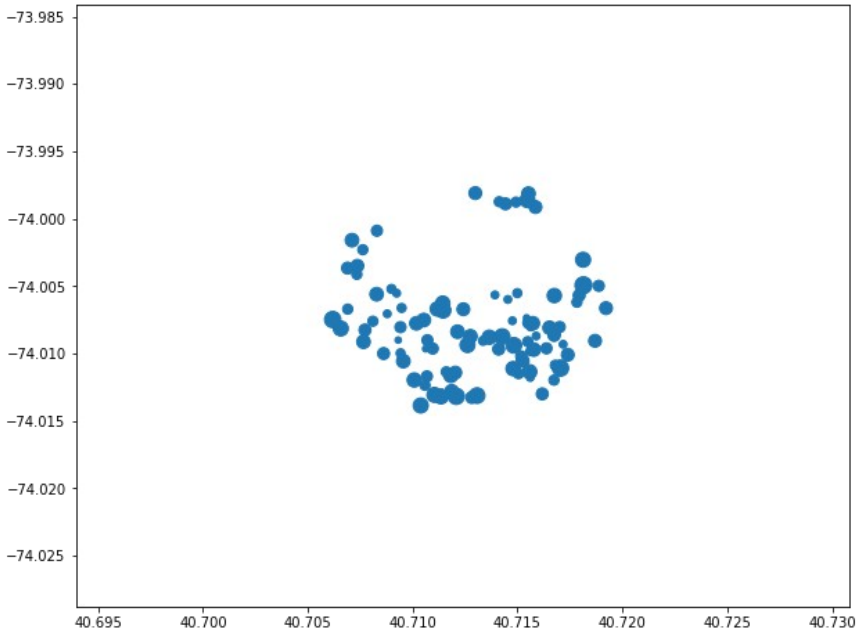

Fig4: Histogram of ratings with boxplot after normalization

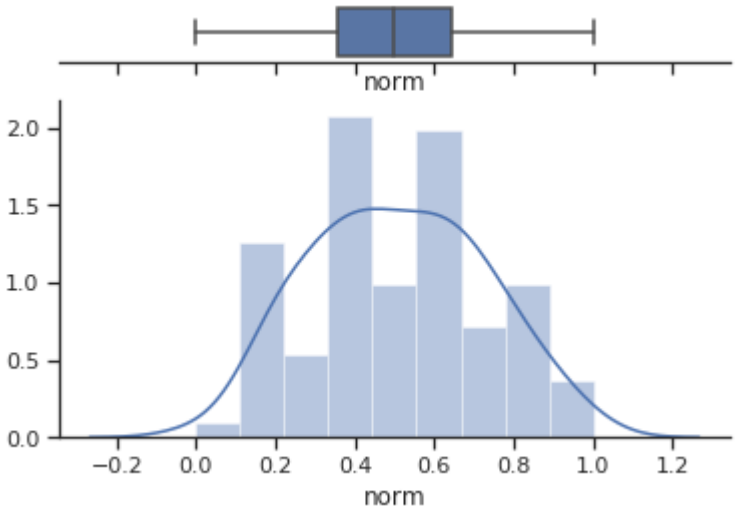

Fig4: Histogram of ratings with boxplot after normalization

4.2. Clustering using DBSCAN

To analyze each region, we will cluster the venues using the latitude, longitude, and rating. As a result, one would realize how the venues are close to each other and how similar the ratings are. This will give us a strong idea about each region and how tough the competition is.

For applying our clustering, we will be using DBSCAN (density based spatial clustering with application of noise) which is an unsupervised clustering algotithm. As you can see from it's name it clusters groups with similar characteristics and closeness and shows the outlier who do not belong to the group. This will be a good indicator to see if there are special cases and venues which have randomly opened. Also the outlier even would be indicator for the influence of the venues.

DBSCAN takes 2 parameters:
- “m” which is the minimum number of points within a neighborhood,
- “n” which the radius of the neighborhood

What happens is each point searches the vicinity arround it based on the passed parameters, and we get a cluster. Points which do not have any points arround them are assumed as noise or outliers.

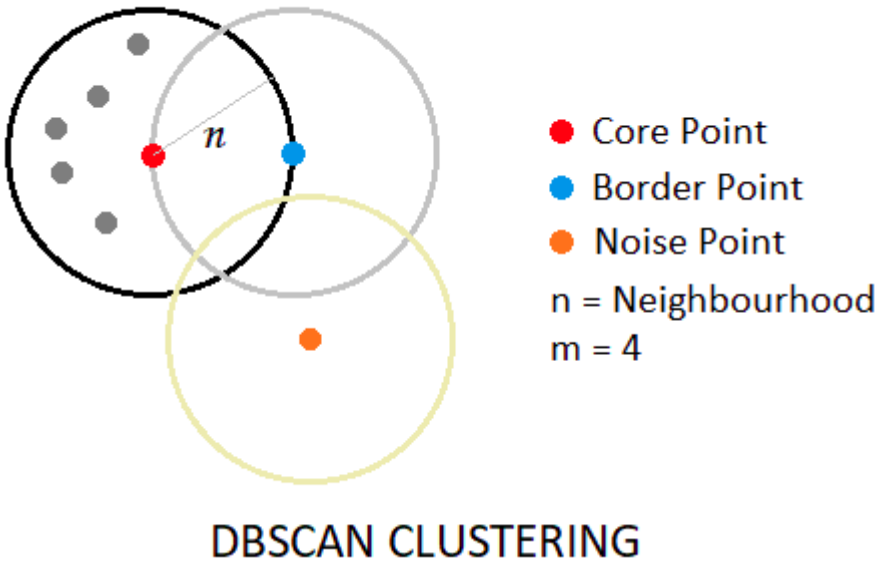

Fig4: DBSCAN

For our clustering we used m=4 and n=0.7:

- m=4 since venues could best exist at right, left, front, back direction, and having up a downstairs or updstairs direction is already assumed as within the neighborhood.
- 0.7 since after normalizing latitude and longitude at 0.7 clusters started to form we adopted it
- * Based on the clustering return a class from 0 to N (labels) will be returned. Note that a -1 is an outlier

5. Results

Let’s look at the results of the city of new york:

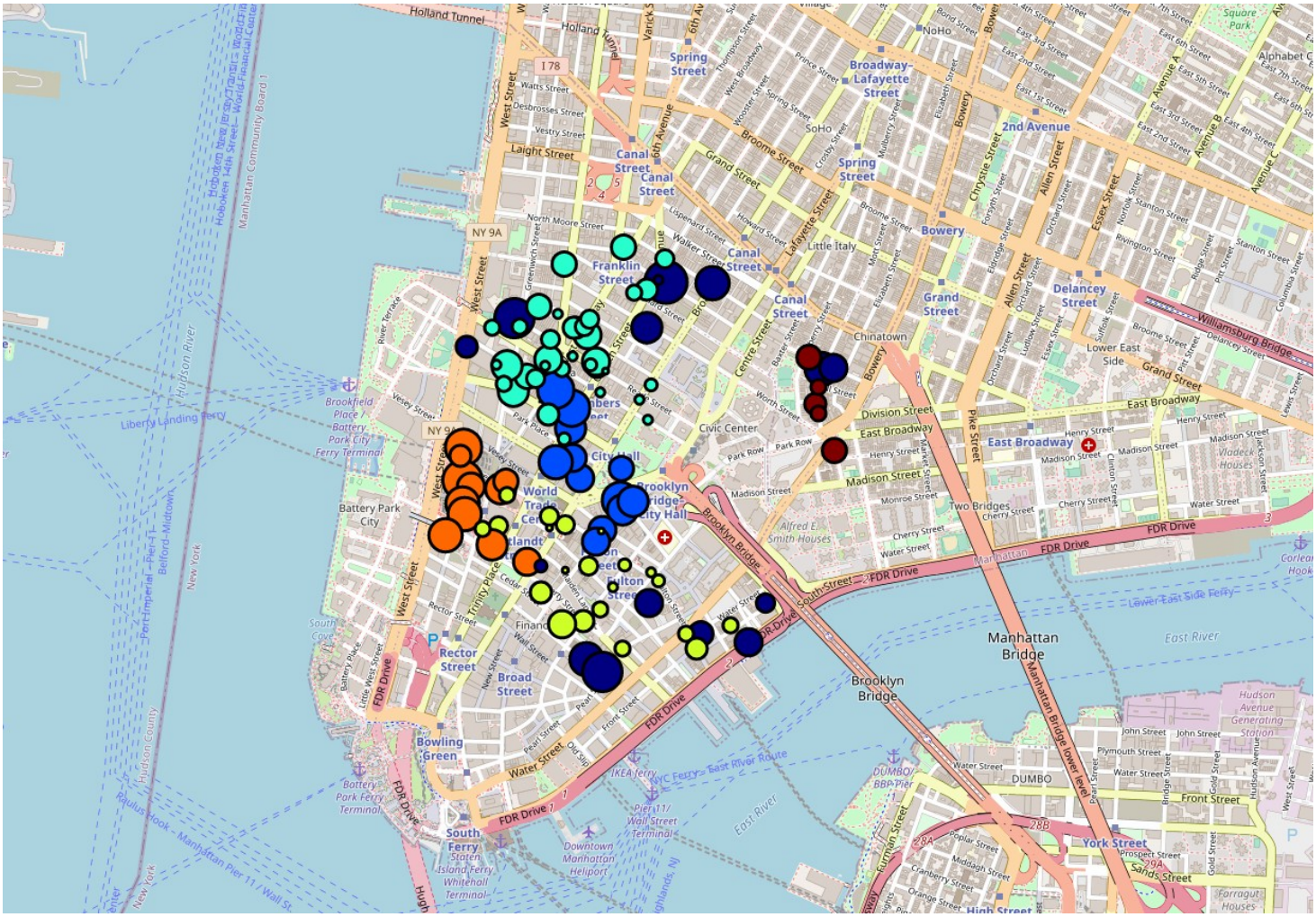

Fig4: Using Folium we overplot the results on the map of New York

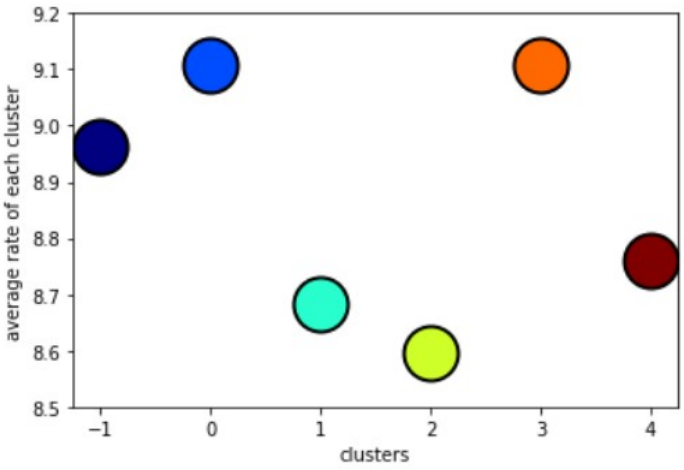

Fig5: indicators for clusters

- Each cluster has a color and encoded by it based on the label
- The size of the circle show how high the ratings is
- We also calculated the average of the classes to gave a global idea of the general nature of the results

	categories
Cocktail Bar	2
Gym / Fitness Center	2
Gym	2
French Restaurant	2
Café	2
Coffee Shop	2
American Restaurant	2
Bookstore	1
Burger Joint	1
Monument / Landmark	1

Table1: This table presents the top ten most frequent venues in the biggest cluster , we can see at the tail there are unique venues as well. The largest cluster has 35 venues with the a relatively low average. (Northern cluster)

	categories
Memorial Site	3
Auditorium	1
Building	1
Electronics Store	1
Plaza	1
Park	1
Coffee Shop	1
Scenic Lookout	1
Gourmet Shop	1

Table2: This table presents the top ten most frequent venues in the best cluster with the highest average (Centeral cluster)

Results and Discussion

In this project, we requested from the Foresquare API the venues and associated ratings of the city of New York within a radius of 1000 and limit of 200 venues. We recieved arround 100 venues. Since Foursquare has quota of 50 venues per day, externally the data was gathered using a paid version to ensure ratings for all the venues.

In our analysis, we have clustered these venues based on their latitude,longitude, and rating using DBSCAN. 6 clusteres were created and one is an outliers cluster. We have realized a spatial and rating wise clustering does exist where the top ratings cluster being the city center towards and its west, the worst being the south but includes some far outliers with much higher ratings. The best clusters has venues which are shops, and memorial cites, etc ... making it a great touristic hub. The largest cluster was cluster 1 and located in the North, although it had a low average rating, and checking it's categories you can see it's mainly composed of restaurants and cafe's which are common daily venues. One clusters remians which is the Eastern part of New York where we can see that the density is low based on the venue gathering radius and there is a gap for venues from the center to that region. to further understand that region one can analyze that section. One can extract the latitude and longitude and rerun the code. What we can also see that the outliers have larger ratings compared to their clusters, as a results being special and better than those regions which a new stakeholder should realize that these sorts of places have a specific shine to them and might monopolizing the market. For example the Washington market park where it is a public park and funded by the government which makes it a well groomed site.

Conclusion

The purpose of this project was to analyze different parts the city of New York with a specific vicinity or radius, in order to understand the quality of of each region and detect homogenous clusters and outlier based on the ratings. This aids stakeholders to make a final descions about opening a new venue. By clustering the venues using latitide, longitude, and rating from Foursquare data, we were able to detect the clusteres within each part of the city and calculate an average rating for each cluster, as a result giving the ability to stackholders to have a generic view of the city's quality of venues and supplying them with an insight regarding the quality,luxury, and expectation of customers for each region.

Final decission on optimal venue location will be made by stakeholders based on specific characteristics of taking into consideration the average rating and additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood, etc. Not to forget the amount of capital a stakeholder would be ready invest in the new venue.