Restaurant Clusters

IBM Course: Applied Data Science Capstone Project

Martin Monsch

March 28, 2019

Abstract

In this data science project, I clustered restaurants in New York using the Scikit-learn's density-based clustering algorithm DBSCAN and visualized the restaurants in the region of Lower Manhattan and Williamsburg based on the average rating of restaurants in the different clusters. The method used can easily be replicated with little extra work for any other city and for other venues such as bars or coffee shops. The resulting maps can help individuals find the neighborhoods and venues they are looking for. As such, they are especially helpful for tourists, but the maps can also help locals stay up-to-date. Knowing the average rating of a venue type in different clusters is especially helpful if the individuals' goal is to explore neighborhoods or if they plan to go to different venues during an evening by foot or bicycle. For businesses, the clusters can help determine where to locate a restaurant, a bar, or a related venue, investors can retrieve relevant information from these maps for their investments, and city planners, tourist guide editors, and other media companies can also make use of these maps. The map of New York showed that the larger clusters had better average ratings than the smaller clusters. Further research is required to analyze whether this is a consistent pattern.

Table of Content

1. Introduction

1.1. Background

Restaurants and bars are usually not evenly distributed in cities. Typically, there are areas with a lot of restaurants and bars, and both locals and tourists crowd these regions in the evenings. Some of these clusters are well-established, some hip, and others are filled with tourist traps. For individuals — especially for tourists — it is not always evident where the restaurant and bar clusters are and how they qualify. City guides or blogs provide information, but the acquisition of the required information is troublesome and reliable information rare. Even many locals are often not up-to-date and do not know where the good restaurants are located since restaurant clusters that were hip once might have already become well-established or even an area filled with tourist traps.

1.2. Problem

The first goal of the capstone project is, therefore, to find a way to cluster restaurants based on their location to identify and visualize restaurant clusters for a given city. As a second goal, the average rating of restaurants in a specific cluster shall be determined and the different clusters visualized again based on the different average ratings to make it easy to find restaurant clusters with a high (or low) average rating.

1.3. Interested Parties

The restaurant clusters are of interest for the following groups: (a) consumers, (b) venues, (c) investors and investment companies, (d) city planners, and (e) media companies. For consumers, the clusters are especially interesting if they do not know the city well, which is the case for tourists, other travelers, and newly settled individuals, but also locals are interested in staying up-to-date and knowing where new clusters emerge. For venues such as restaurants the clusters can help determine where to locate a restaurant or a bar, whether the goal is to be within a well-rated cluster, a not so well rated-cluster for the sake of less competition, or in an area in between clusters with an undersupply of restaurants. Investors can potentially get relevant information from these clusters as well, especially if they are compared over time to identify trends, and city planners might find the clusters useful for their work. Finally, media companies could be interested and companies that edit tourist guides.

2. Data

To build the desired restaurant clusters, location data and ratings for restaurants are required. This data was collected through the Foursquare API. New York was chosen as the location for the clusters, but it is possible to apply the same algorithms to any other city. Since both the search and explore queries of the Foursquare API are limited to fifty venues, a relatively dense grid with location points was required to collect sufficient data for restaurants in New York.

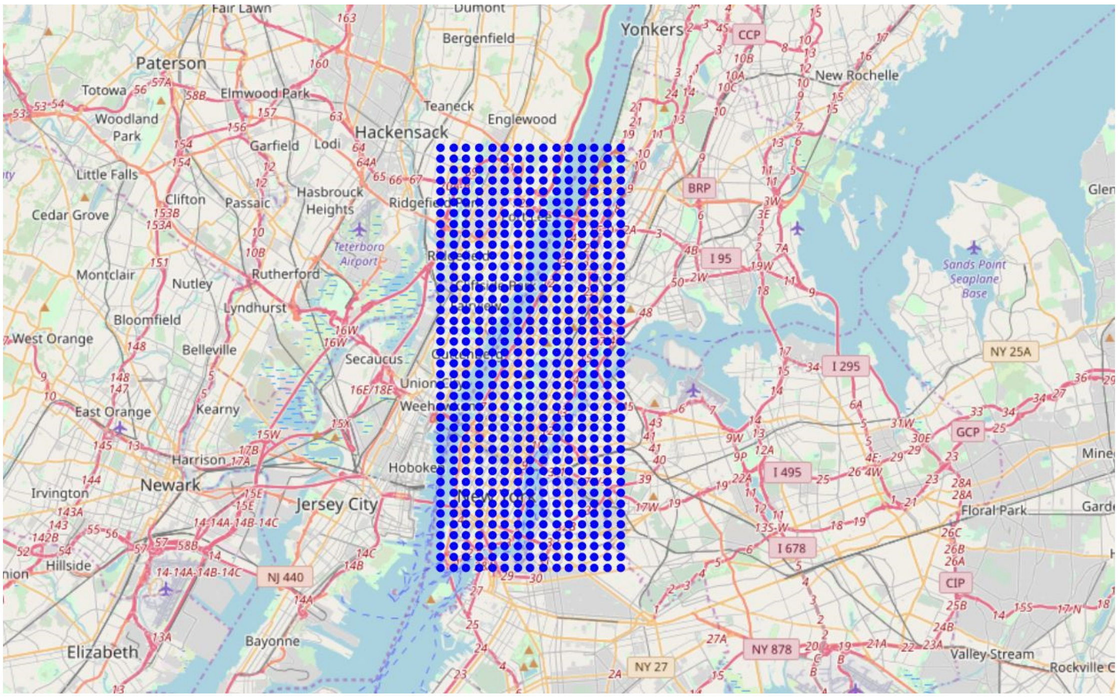
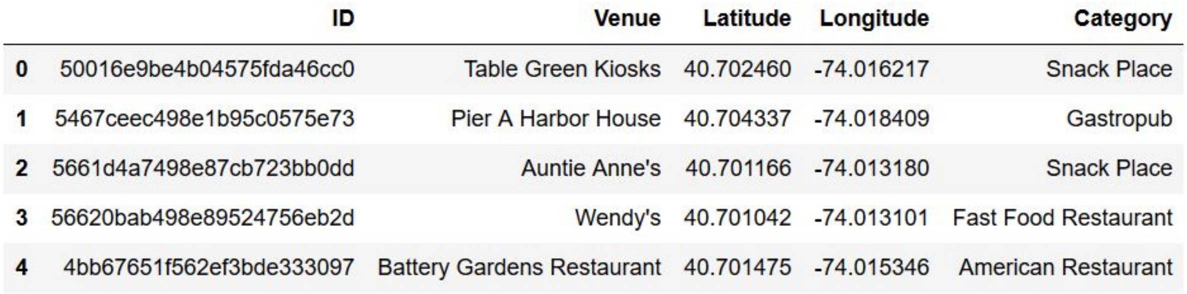


Figure 1: Grid with location points to access restaurant data for Manhattan through Foursquare.

To collect the data through the Foursquare API, a URL is required for each data point. I defined an explore query based on the latitude and longitude of a given point with a radius of 500 meters and a limit of 100. In addition, the section “food” was used to narrow the search to food related venues. With the URL, I then defined a function to collect the results for all the grid points. I used the try/except mechanism since some grid points were located in areas with no food venues close. This creates an error without the use of try and except.

The collected data was then merged into one data frame, and duplicates were eliminated. The following feature columns were chosen for the merged data frame with food venues in New York: “ID” (to get the ratings later), “Venue Name”, “Venue Category” (since filters may be applied at a later stage), “Latitude”, and “Longitude”.

Table 1: Merged data frame with restaurants and other food related venues in New York (first 5 rows).



	ID	Venue	Latitude	Longitude	Category
0	50016e9be4b04575fda46cc0	Table Green Kiosks	40.702460	-74.016217	Snack Place
1	5467ceec498e1b95c0575e73	Pier A Harbor House	40.704337	-74.018409	Gastropub
2	5661d4a7498e87cb723bb0dd	Auntie Anne's	40.701166	-74.013180	Snack Place
3	56620bab498e89524756eb2d	Wendy's	40.701042	-74.013101	Fast Food Restaurant
4	4bb67651f562ef3bde333097	Battery Gardens Restaurant	40.701475	-74.015346	American Restaurant

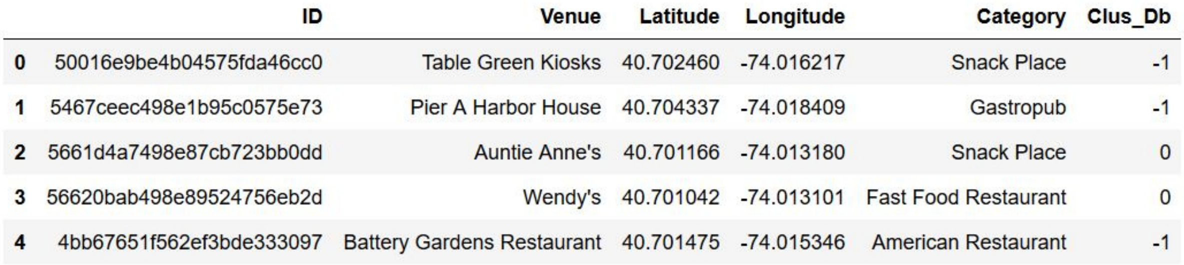
The ratings will be added at a later stage and just for a sample of restaurants since my access to rating data is limited to 500 venues per day.

3. Methodology

To cluster the restaurants, only the latitude and longitude of the restaurant data frame was required. As a clustering algorithm, DBSCAN, a density-based algorithm from Scikit-learn, was chosen. The DBSCAN algorithm first classifies the different data points into three categories: core points, border points, and outliers. This is done based on a predefined radius *R* and a predefined number *m*. Core points are data points with at least *m* data points within its radius, border points are data points with less than *m* data points but with a core point within its radius, and outliers are the remainder. The DBSCAN algorithm was mainly chosen because it finds clusters with an arbitrary shape and is also able to identify outliers. Furthermore, it is not necessary to specify the number of clusters in advance.

As parameters, a radius of 0.05 and an *m* of 28 were used to identify the clusters. The cluster labels were then introduced as a new column in the restaurants data frame.

Table 2: Restaurants and other food related venues in New York and a clustering label (first 5 rows).



	ID	Venue	Latitude	Longitude	Category	Clus_Db
0	50016e9be4b04575fda46cc0	Table Green Kiosks	40.702460	-74.016217	Snack Place	-1
1	5467ceec498e1b95c0575e73	Pier A Harbor House	40.704337	-74.018409	Gastropub	-1
2	5661d4a7498e87cb723bb0dd	Auntie Anne's	40.701166	-74.013180	Snack Place	0
3	56620bab498e89524756eb2d	Wendy's	40.701042	-74.013101	Fast Food Restaurant	0
4	4bb67651f562ef3bde333097	Battery Gardens Restaurant	40.701475	-74.015346	American Restaurant	-1

The outliers (Clus_Db label of -1) were then eliminated and the result visualized using Folium with a rainbow color map.

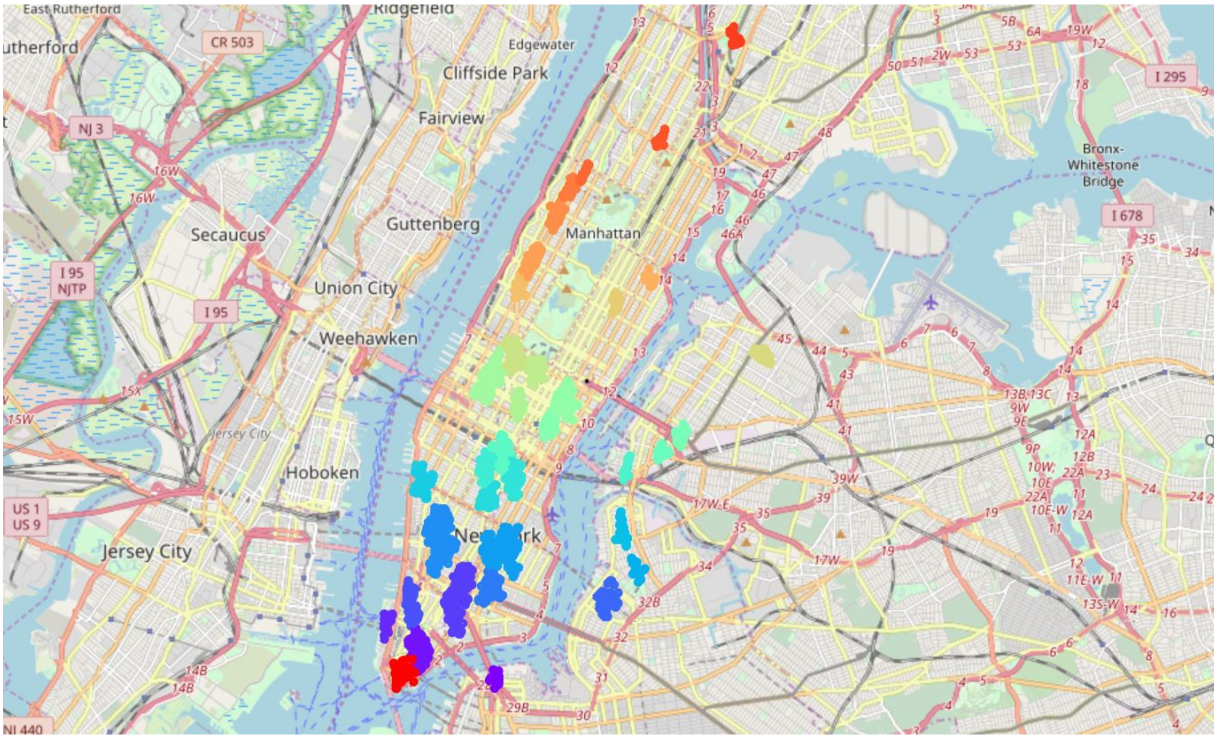


Figure 2: Restaurant Clusters in New York created with Scikit-learn's DBSCAN.

As a next step, a sample of clusters for lower Manhattan and Williamsburg was selected based on the following map.

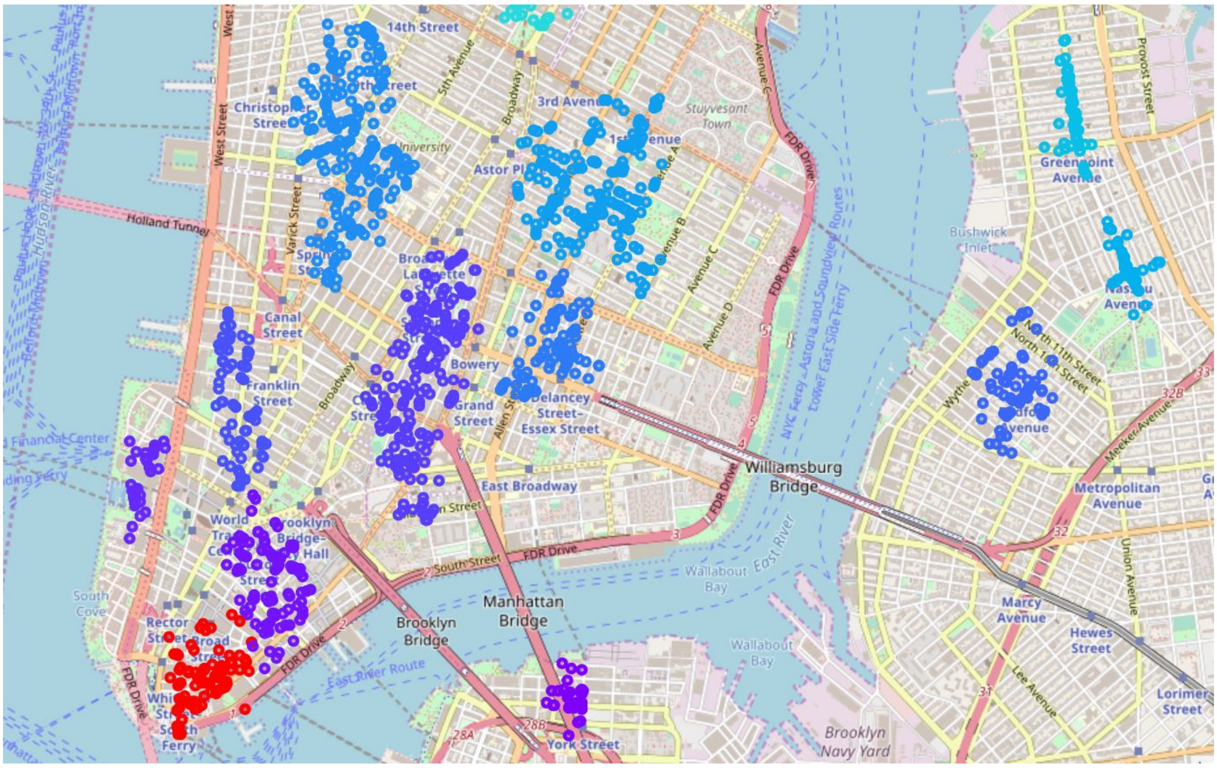
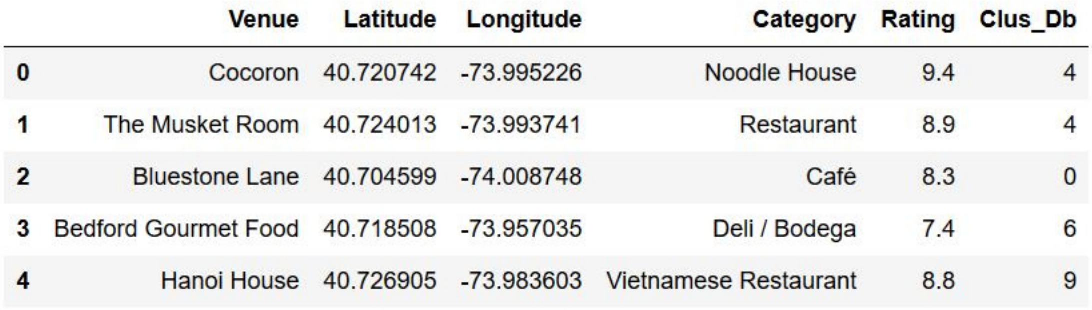


Figure 3: Restaurant Clusters in Lower Manhattan and Williamsburg.

The sample included the clusters 0 to 11 with 1170 venues. While the biggest cluster (Greenwich Village) included 217 food venues, the smallest one enclosed only 31 venues. Since the number of venues was still too big, a subset of 480 food venues was selected for the purpose of accessing rating data.

As a next step, a function was defined to access the ratings for all restaurants. The URL just requires the unique id of each venue to access the details. A try/except mechanism was used again since some venues do not have a rating. The restaurants data frame was supplemented with a “rating” column.

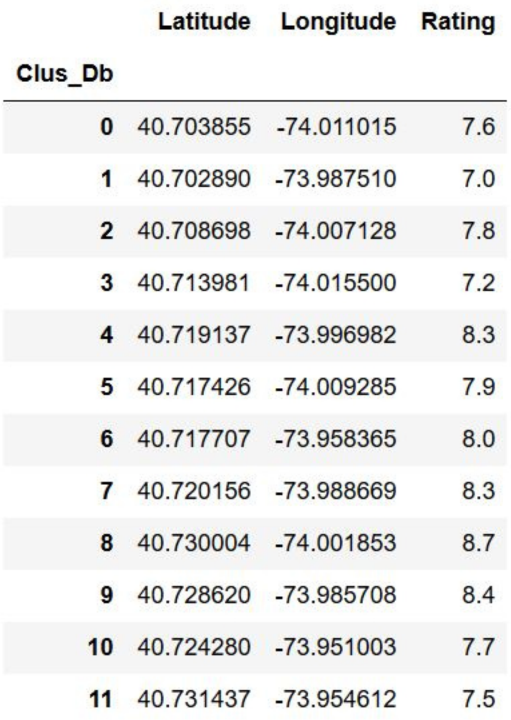
Table 3: Restaurants and other food related venues in New York rated and clustered (first 5 rows).



	Venue	Latitude	Longitude	Category	Rating	Clus_Db
0	Cocoron	40.720742	-73.995226	Noodle House	9.4	4
1	The Musket Room	40.724013	-73.993741	Restaurant	8.9	4
2	Bluestone Lane	40.704599	-74.008748	Café	8.3	0
3	Bedford Gourmet Food	40.718508	-73.957035	Deli / Bodega	7.4	6
4	Hanoi House	40.726905	-73.983603	Vietnamese Restaurant	8.8	9

Then, the data frame was grouped by the different clusters, the average rating calculated for each cluster and rounded.

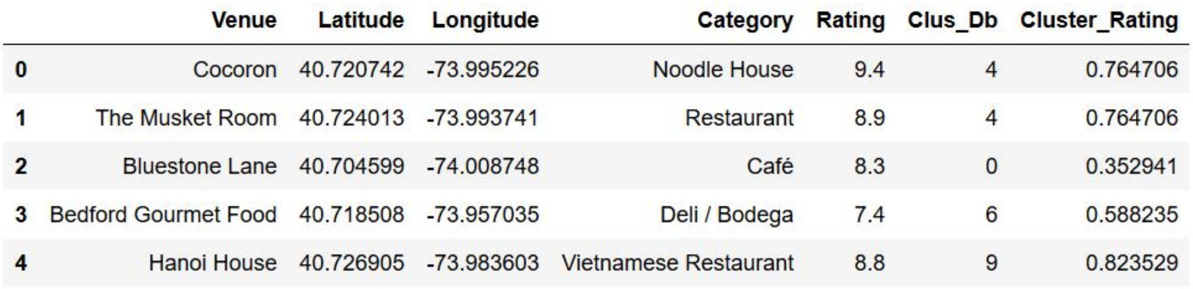
Table 4: Restaurants in New York grouped by cluster with the average rating.



Clus_Db	Latitude	Longitude	Rating
0	40.703855	-74.011015	7.6
1	40.702890	-73.987510	7.0
2	40.708698	-74.007128	7.8
3	40.713981	-74.015500	7.2
4	40.719137	-73.996982	8.3
5	40.717426	-74.009285	7.9
6	40.717707	-73.958365	8.0
7	40.720156	-73.988669	8.3
8	40.730004	-74.001853	8.7
9	40.728620	-73.985708	8.4
10	40.724280	-73.951003	7.7
11	40.731437	-73.954612	7.5

This average was then inserted as a new column in the data frame with the sample of restaurants and the values normalized to give a clearer image (dark-red for an average rating of 7.0 and dark-green for an average rating of 8.7).

Table 3: Restaurants and other food related venues in New York with normalized cluster rating (first 5 rows).



	Venue	Latitude	Longitude	Category	Rating	Clus_Db	Cluster_Rating
0	Cocoron	40.720742	-73.995226	Noodle House	9.4	4	0.764706
1	The Musket Room	40.724013	-73.993741	Restaurant	8.9	4	0.764706
2	Bluestone Lane	40.704599	-74.008748	Café	8.3	0	0.352941
3	Bedford Gourmet Food	40.718508	-73.957035	Deli / Bodega	7.4	6	0.588235
4	Hanoi House	40.726905	-73.983603	Vietnamese Restaurant	8.8	9	0.823529

The color map RdYlGn (red-yellow-green) was used to visualize the restaurant cluster with the best ratings dark-green and cluster with the worst ratings dark-red.

4. Result

The following figure shows a map with the clusters from lower Manhattan and Williamsburg with the restaurants visualized based on the average rating per cluster with dark-green for the restaurants in the best-rated cluster (Greenwich Village; average rating of 8.7) and dark-red for the cluster with the worst average rating (7.0).

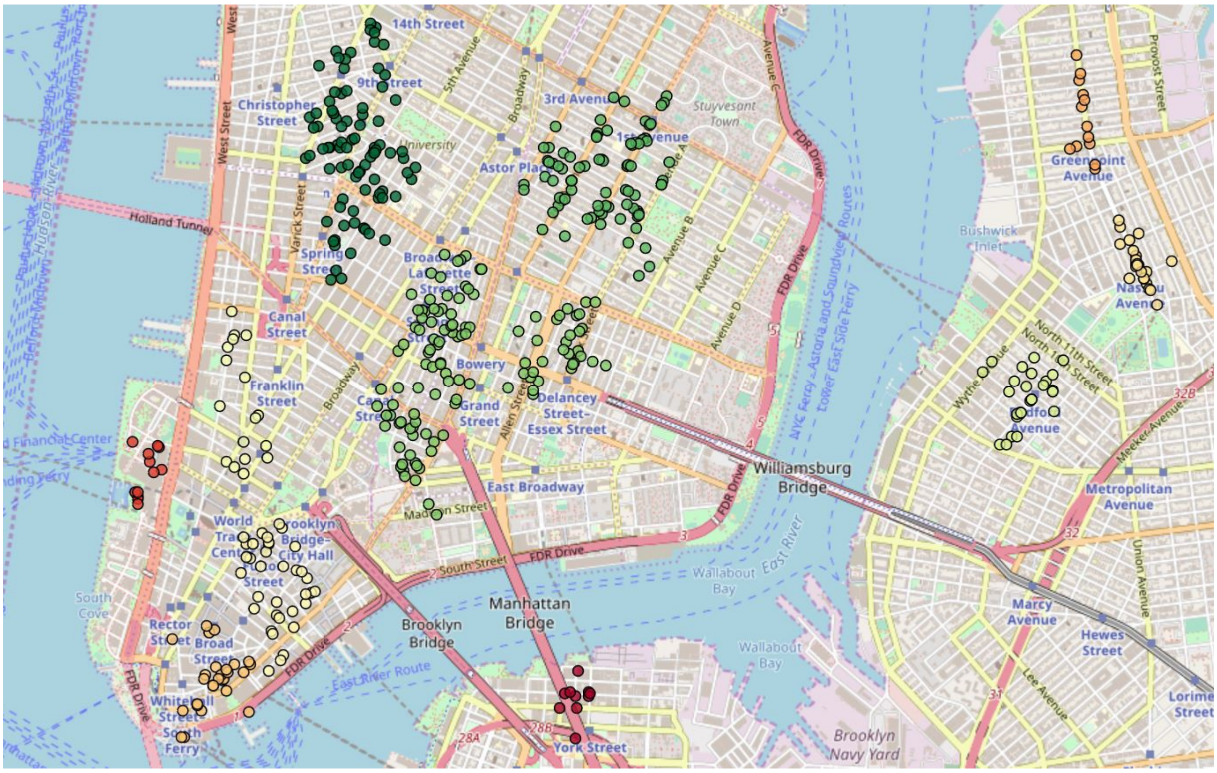


Figure 4: Restaurant Clusters in Lower Manhattan and Williamsburg colored based on average cluster rating.

The production of this map was the main goal of this project. However, the higher goal is the whole methodology and the code that allows replicating the same process with any other city and with other venues such as bars or coffee shops.

5. Discussion

Based on this map, as a tourist I can find now the regions in New York with a lot of good restaurants. For a person who wants to open a restaurant, this map could be valuable as well, as this map could help this person find out where opening a restaurant makes sense. That person might choose to open a restaurant within a cluster with many well-rated restaurants since this is likely the place where many informed customers crowd to, or on the contrary, open a restaurant at a place with only poorly rated competitors around. Investors, editors of tourist guides, the media, and city planners might also gain valuable information from maps like this one.

Most importantly, it is easy to replicate this process with any other city in the world. Mainly, the coordinates need to be changed and probably the parameters of DBSCAN adjusted, but the work involved in creating new maps like the one above is very limited, and thus the possible cost-benefit relationship very good. Also, it would be very easy to create maps for bars or coffee shops.

However, such maps could have the effect that everybody — even more than nowadays — crowd to the same regions. Restaurants located in areas with a bad average rating might be adversely affected even if the restaurants are good. Therefore, I deem it appropriate to show also a map where the restaurants are visualized individually based on their rating.

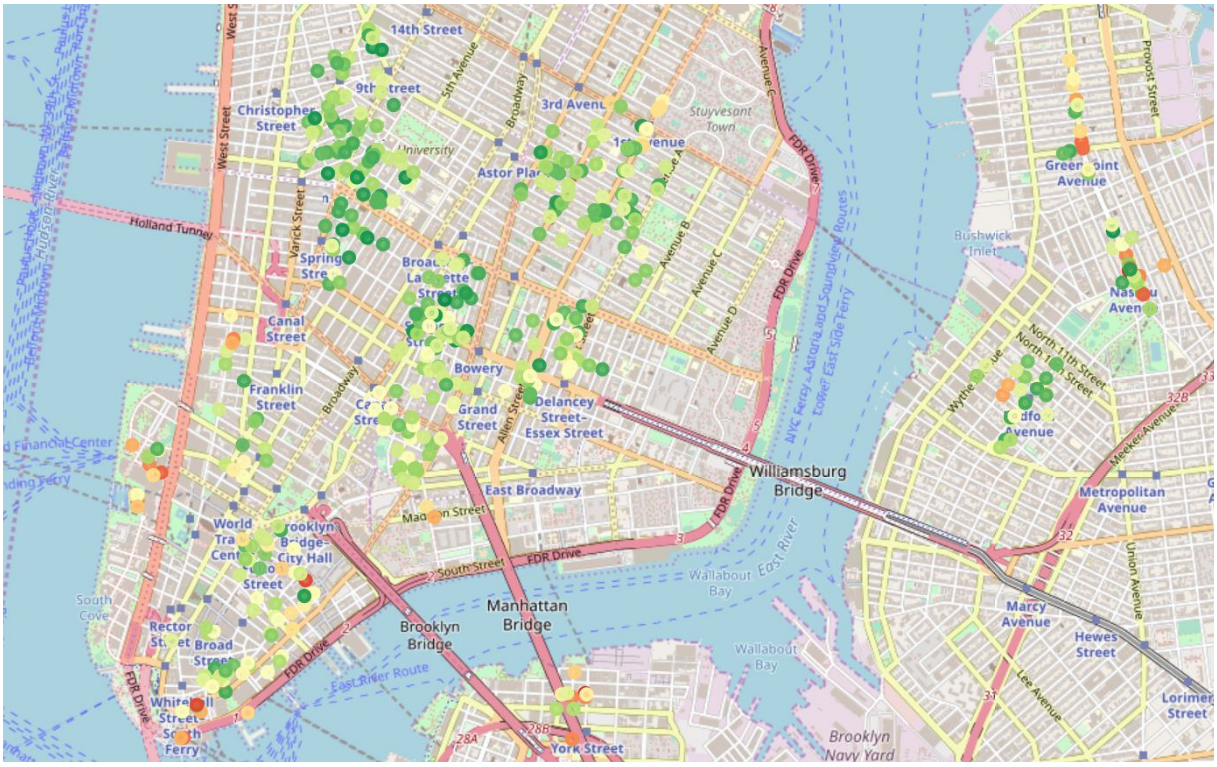


Figure 5: Restaurant Clusters in Lower Manhattan and Williamsburg colored based on restaurant rating.

This figure shows the exceptions within the different clusters quite well.

Furthermore, I used a general food query based on the classification of Foursquare. Possibly, it would be more adequate to use a restaurant query. Fast food venues could have a detrimental effect on the average rating of a cluster while the average ratings of restaurants in that cluster might be good.

Figure 4 and table 4 above indicate that size of a cluster might have an influence on the rating and thus the quality of a restaurant. This might be due to more competition in these areas. Further research is required to analyze this potential relationship.

6. Conclusion

In this data science project, I clustered restaurants in New York using the Scikit-learn's density-based clustering algorithm DBSCAN and visualized the restaurants in the region of Lower Manhattan and Williamsburg based on the average rating of restaurants in the different clusters. The method used can easily be replicated with little extra work for any other city and for other venues such as bars or coffee shops. The resulting maps can help individuals find the neighborhoods and venues they are looking for. As such, they are especially helpful for tourists, but the maps can also help locals stay up-to-date. Knowing the average rating of a venue type in different clusters is especially helpful if the individuals' goal is to explore neighborhoods or if they plan to go to different venues during an evening by foot or bicycle. For businesses, the clusters can help determine where to locate a restaurant, a bar, or a related venue, investors can retrieve relevant information from these maps for their investments, and city planners, tourist guide editors, and other media companies can also make use of these maps. The map of New York showed that the larger clusters had better average ratings than the smaller clusters. Further research is required to analyze whether this is a consistent pattern.