Analysis of Restaurants in New York

Ana Parčina

5.11.2020.

Content

1. Introduction

1.1 Background

New York is big city with high diversity of food that restaurants offer. When someone is looking to open a restaurant there, it is difficult to break into market because there are many restaurants. And this problem is getting even bigger in New York where you can find a lot of restaurants that are popular, and people are used to going there. To make restaurant as successful as possible, it would be great if future owner would know what the best place is for opening new restaurant according to kind of food that it would offer.

1.2 Problem

The problem that will be solved in this case is finding answer on question where to open a restaurant. In this case I will consider one of the most popular borough of New York, Manhattan. Also, I will consider three types of restaurants that are the most popular: Italian, American and Mexican. I will take in account what ratings have these restaurants and where are they located.

1.3 Interest

According to analysis of ratings and types of restaurants, client will see on map where he/she can open a restaurant without good competition nearby. Considering that Manhattan is big, if someone wants to eat something good, proximity will play an essential role in success of restaurant at the beginning.

Figure 1. An example of NY restaurant

2. Data acquisition and cleaning

2.1 Data sources

First, I will use data that we used for our Foursquare Lab about neighborhoods in New York. Another data that will be used is from Foursquare locator. In next chapter I will explain why I used each data, how I cleaned it and what features are important for this case.

2.2 Data cleaning

So, neighborhood data has four columns: borough, neighborhood, latitude and longitude. I cleaned that data so that I removed all data that is not about Manhattan. What I got was a dataset with boroughs on Manhattan and their latitude and longitude.

Cleaning second data that I got using Foursquare Locator was a little bit harder. This data was gathered using function that gives nearby venues of each borough. After finding venues, it is necessary to exclude venues that are not restaurants. Also, after analyzing the data, I exclude all restaurants that are not American, Italian or Mexican because this three are the most common, and there is not enough data for other restaurants. I labeled Italian restaurants as 1, American as 2 and Mexican as 3.

Rating for each restaurant was difficult to find by name of restaurant, unfortunately I was not able to combine Foursquare locator information so I gave random ratings from 1 to 5 for each restaurant using *numpy.random* function.

2.3 Feature selection

After cleaning data, I got 98 restaurants of 325. There are some features that are not necessary for later analysis, so I dropped it.

First, I dropped neighborhood latitude and longitude because this data is not essential for recommendation of location for future restaurant.

Second unnecessary data is related with borough names. We only included Manhattan borough, so this data is irrelevant.

Also, as I mentioned before, data about restaurants that are not Italian, Mexican or American needs to be dropped.

Table 1 explains simple feature selection during data cleaning.

Table 1. Simple feature selection during data cleaning.

Kept features	Dropped features	Reason for dropping features
Venue Latitude, Venue Longitude	Neighborhood Latitude, Neighborhood Longitude	In this case we are interested in location of each restaurant, not the coordinates of its neighborhood.
Neighborhood	Borough	Borough is the same for every neighborhood.
Italian, American, Mexican restaurants	All other types of restaurants	Only three type of restaurants are considered in this case due to amount of data.

3 Exploratory Data Analysis

Figure 2 shows the map of New York with boroughs and neighborhoods. As I mentioned I later considered only area of Manhattan showed on figure 3.

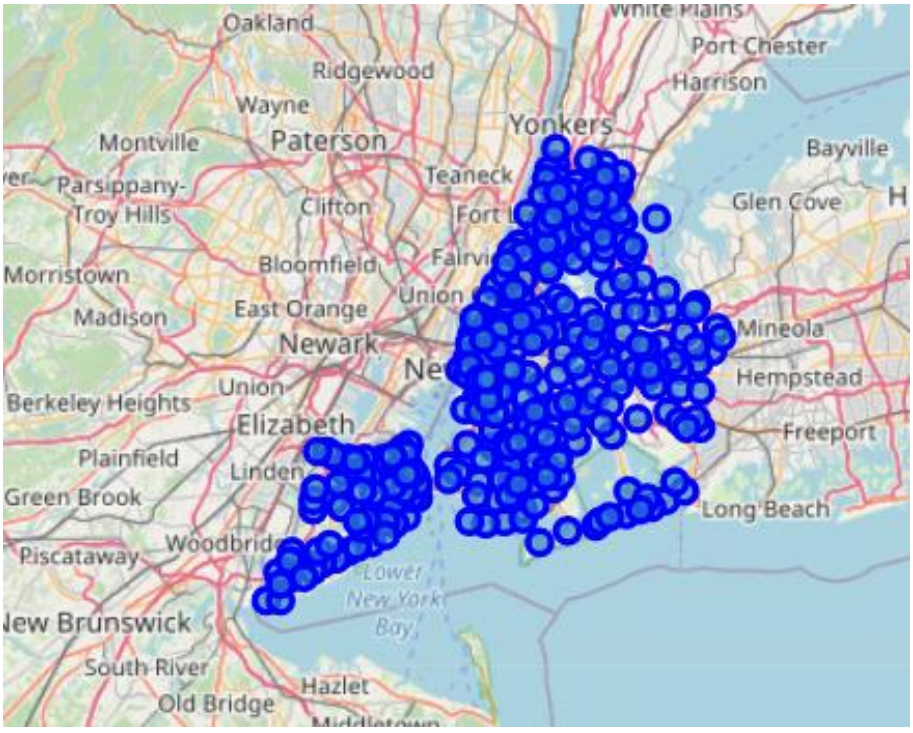

Figure 2. Map of New York and its boroughs and neighborhoods

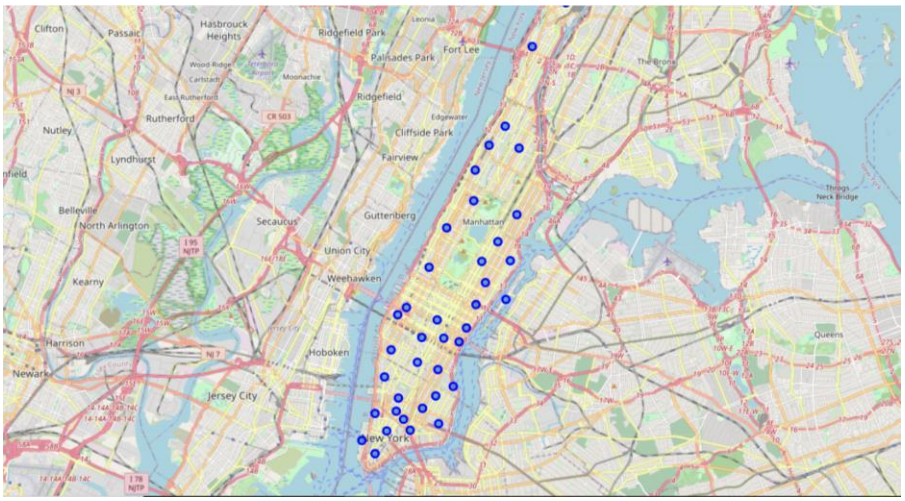

Figure 3. Neighborhoods on Manhattan

Using function *get nearby venues,* I found venues in each neighborhood and filter data to find restaurants that are showed on figure 4.

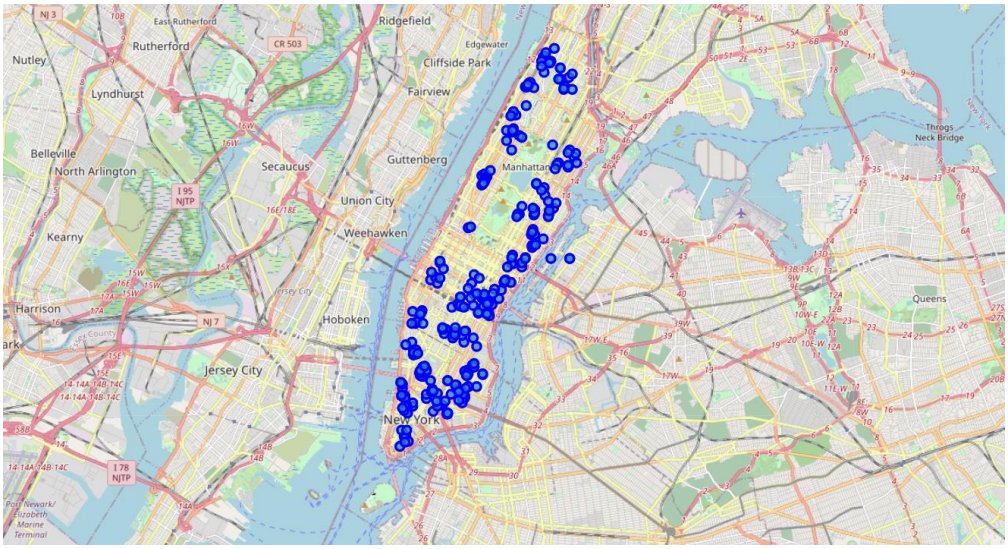

Figure 4. Restaurants on Manhattan

I made analysis of types of restaurants that showed me that most of restaurants are American, Italian and Mexican. For better visualization of data, I made pie plot showed on figure 5. Number of other types of restaurants is not very big, so I decided to exclude it because the need of bigger data.

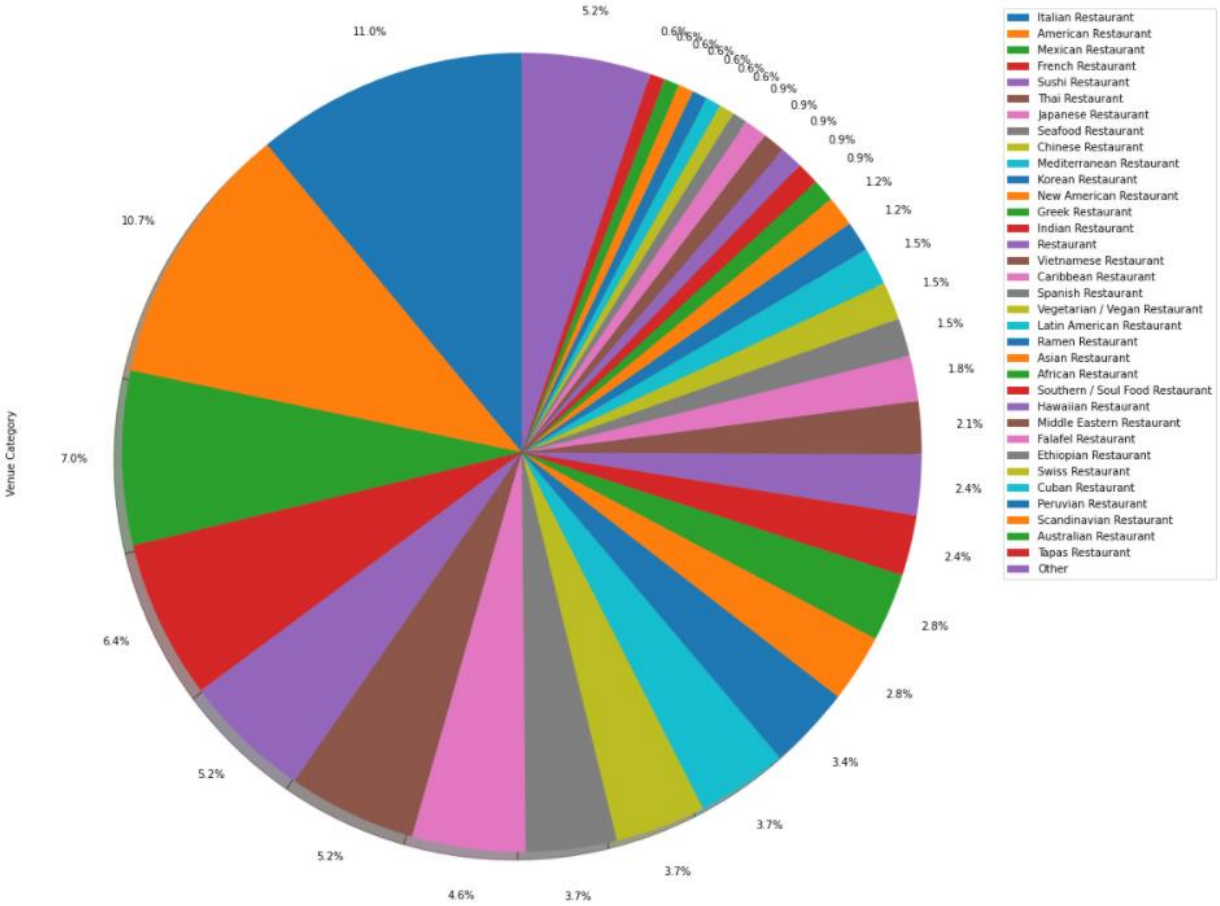

Figure 5. Types of restaurants on Manhattan

American, Italian and Mexican restaurants on Manhattan are shown on figure 6.

Figure 6. Italian, American and Mexican restaurants on Manhattan

4 KMeans model

I used *sklearn* library for KMeans model. After examination of data, I decided to use 6 clusters. Data that is used for this model is related with latitude, longitude, type of restaurant and rating. After fitting model, I got results of restaurants that have excellent grade(5) and they are Italian and Mexican, also I got one cluster of restaurants that have bad rating(1 and 2) and are Italian and American and so on. Figure 7 shows results of clustering on Manhattan map.

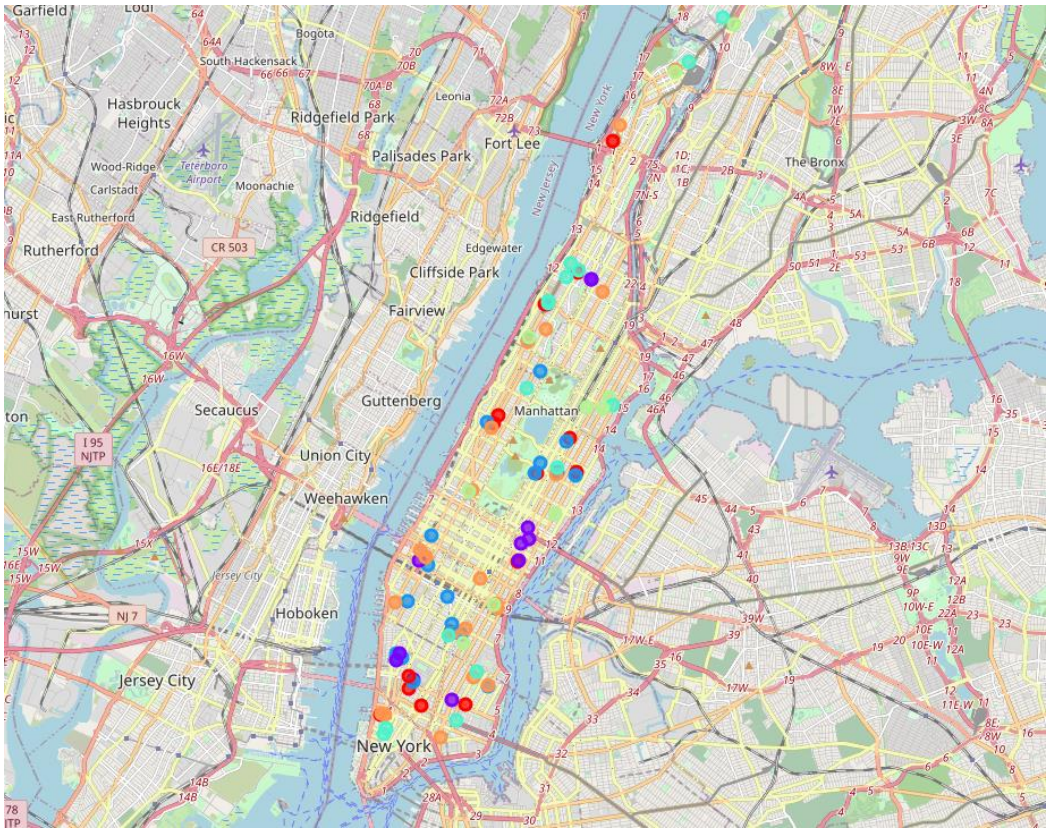

Figure 7. Clusters from KMeans model

Figure 8 shows restaurants that are grouped in first cluster. We see that here are only Italian restaurants with medium grades, but what is interesting here are restaurants with high grades and restaurants with bad grade. The goal is to open restaurant in area where there is no good competition which means it has to be area with no excellent restaurants in vicinity and with bad restaurants.

```
In [39]: restaurants.loc[restaurants['Cluster Labels'] == 0]
```

Out[39]:

	Cluster Labels	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	rating
64	0	Washington Heights	Saggio Restaurant	40.851423	-73.939761	1	3
122	0	Hamilton Heights	Fumo	40.821412	-73.950499	1	3
158	0	Manhattanville	Pisticci Ristorante	40.814015	-73.960266	1	3
170	0	Manhattanville	Bettolona	40.814084	-73.959574	1	4
241	0	Upper East Side	Sant Ambroeus	40.775328	-73.962819	1	3
282	0	Yorkville	Nica Trattoria	40.775688	-73.950570	1	4
365	0	Upper West Side	Celeste	40.786689	-73.975737	1	3
369	0	Upper West Side	Pizzeria Sirenetta	40.788640	-73.974282	1	4
553	0	Greenwich Village	Coco Pazzo Kitchen & Restaurant	40.726036	-74.001507	1	3
561	0	Greenwich Village	Dante NYC	40.728847	-74.001622	1	3
628	0	Lower East Side	Il Posto Accanto	40.722301	-73.984188	1	4
631	0	Tribeca	Locanda Verde	40.719981	-74.010002	1	3
717	0	Soho	Osteria Morini	40.721990	-73.997790	1	3
913	0	Carnegie Hill	Sfoglia	40.783419	-73.952768	1	3
1071	0	Turtle Bay	La Pecora Bianca	40.755116	-73.968509	1	3
1161	0	Flatiron	Rezdôra	40.738941	-73.988862	1	4

Figure 8. Results for one of clusters (medium grades)

Figure 9 shows restaurants that do not have good grades.

```
In [41]: restaurants.loc[restaurants['Cluster Labels'] == 2]
```

Out[41]:

	Cluster Labels	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	rating
260	2	Upper East Side	Sistina	40.777597	-73.961685	1	2
266	2	Upper East Side	Antonucci	40.775711	-73.956607	1	2
283	2	Yorkville	Felice 83	40.774867	-73.950868	1	1
521	2	Chelsea	The Meatball Shop	40.745988	-74.001686	1	1
555	2	Greenwich Village	Lupa	40.727577	-74.000095	1	1
720	2	West Village	L'Artusi	40.733888	-74.005114	1	1
721	2	West Village	I Sodi	40.733348	-74.004947	1	2
741	2	West Village	Fiaschetteria Pistoia	40.733203	-74.005426	1	1
742	2	West Village	Via Carota	40.733052	-74.003573	1	1
753	2	Manhattan Valley	Osteria 106	40.798825	-73.961793	1	1
821	2	Gramercy	Maialino	40.738442	-73.985610	1	1
926	2	Carnegie Hill	Lex Restaurant	40.782530	-73.953700	1	1
995	2	Midtown South	L'Amico	40.747096	-73.989805	1	2
1127	2	Stuyvesant Town	Fiaschetteria Pistoia	40.726534	-73.977358	1	1
1190	2	Hudson Yards	Sergimmo Salumeria	40.754388	-73.995642	1	1
252	2	Upper East Side	The Mark Restaurant by Jean-Georges	40.775219	-73.963381	2	1
362	2	Upper West Side	Maison Pickle	40.786990	-73.977787	2	1
439	2	Clinton	44 & X	40.761024	-73.994518	2	1
790	2	Morningside Heights	Community Food & Juice	40.805823	-73.965483	2	1
1166	2	Flatiron	Dig Inn	40.740698	-73.988315	2	1

Figure 9. Italian and Mexican restaurants with bad grades

Figure 10 shows restaurants with excellent grades.

```
In [40]: restaurants.loc[restaurants['Cluster Labels'] == 1]
```

Out[40]:

	Cluster Labels	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	rating
541	1	Greenwich Village	Carbone	40.727903	-74.000136	1	5
941	1	Noho	Lil' Frankie's	40.723445	-73.988328	1	5
1025	1	Sutton Place	Felidia	40.760267	-73.965277	1	5
1035	1	Sutton Place	Scalinatella	40.762761	-73.965566	1	5
1040	1	Sutton Place	amata	40.759236	-73.967515	1	5
140	1	Hamilton Heights	The Edge Harlem	40.819692	-73.946073	2	5
196	1	Central Harlem	The Edge Harlem	40.819692	-73.946073	2	5
734	1	West Village	Westville West	40.734055	-74.004279	2	5
739	1	West Village	The Little Owl	40.732441	-74.005424	2	5
1060	1	Turtle Bay	The Smith	40.755376	-73.968243	2	5
1173	1	Hudson Yards	Friedman's	40.755271	-73.998509	2	5

Figure 10. Restaurants with excellent grades

5 Conclusions

As a result of KMeans model we got the map of Manhattan restaurants with 6 clusters, each one representing Italian, Mexican and American restaurants with excellent, medium and low grades. When client looks at the map he/she can determine where should it be good place for opening the restaurant according to the type of food that it will serve. The restaurant owner will not have good competition which will be great tail wind in the beginning of his business.

6 Future directions

Projects like this one are nowadays very much needed in large cities with high variety of food like New York and other world metropolises. Analysis of big data is un advantage, but in this case, I did not have it. In future it would be grate to gather more data, from for example hundreds of restaurants which will then make it possible to do analysis for different kind of restaurants.