The Battle of Neighborhoods - Italian restaurants in NY City

Remigiusz Gębka

August 18, 2020

1. Introduction

1.1. Background

New York City (NYC/NY) is the most populous city in the whole United States, with an estimated 8.3 million of residents as of July 2019 – it's almost twice in comparison to the second most populous city (LA). In 2017, it had an estimated density of 11 thousands per square kilometer, so only few cities have greater density.

These numbers are only for NYC – because if we take these numbers for consideration in taking into account also numbers for the NY metropolitan area and state of NY – we will have 36% of metropolitan area and 43% of New York state population. Quite a lot, right?

What is behind these values? Mostly the fact that the city has higher immigration than outmigration since 2010 Census data – and it's built on fact, that NY has always been a major point for all immigrants – and of course a starting point for them, as a gateway to this country.

This city is a home for over 500 art galleries, 2000 arts and cultural organizations, and more than 27 thousands of restaurants – starting with Central and Eastern Europe cuisine, moving through Italian, Jewish, Irish, Middle-Eastern to Chinese cuisine. And it's all based, mostly, on immigrants from countries from specific regions of the world – and thanks to that, NY has a reputation of having the most diverse haute-cuisine, according to Michelin.

1.2. Problem

As I am a huge fan of Italian cuisine, I would like to answer following questions:

- Find areas with the best Italian restaurants (by overall quality of them) and visualize them on the map of NY.
- Find the best location for such a restaurant (Based on the number of them? Quality of them?)
- Find areas which has problem with quality of such restaurants
- Which part of NY will be best to stay in if you want to visit the best restaurants?

1.3. Interest

As a main group of potential users of outcome of this project:

- Tourists that are looking for cool area to stay, with great food nearby
- People that want to create a new restaurant and looking for place for it
- Owners of restaurants – to check, in which position is their district, maybe they can make something for their community to increase quality and promote?

2. Data acquisition and cleaning

2.1. Data sources

As a main dataset, we will use New York City dataset containing information about all neighborhoods and boroughs with their geolocalization. For that purpose we will use a simplified geojson file that can be downloaded here. For creating more detailed maps and visualizations we will use an enhanced geojson file that can be downloaded here.

To collect information about restaurants, we are using Foursquare API - both normal like also premium endpoints - where venues/search is responsible for gathering list of venues of specific type (italian restaurants), and where second endpoint venues/venude_id is responsible for collection elements such as: price tier, rating, number of likes and dislikes. Although, the number of restaurants is quite high, so for downloading all the data using free tiers of Foursquare API, we have to divide the dataset into 4 parts - each part will be gathered separately, on different days, but using the same API version.

2.2. Data processing

First, the initial dataset is created based on the minimal geojson file, using the "*features"* section inside this file. As a result of this step, we will have a list of neighborhoods inside all boroughs, with their latitude and longitude. Using this dataset and Foursquare API we can collect for each neighborhood top 50 nearest restaurants within 500 meters of the center of the neighborhood. After this step, there is a frame with about 3300 samples. Now there is a need to clear this table, first, starting with duplicates - after this step, the number of them is equal to 1669 samples. For few of them that will be only 1-3 restaurants, but for some of them - there will be much more.

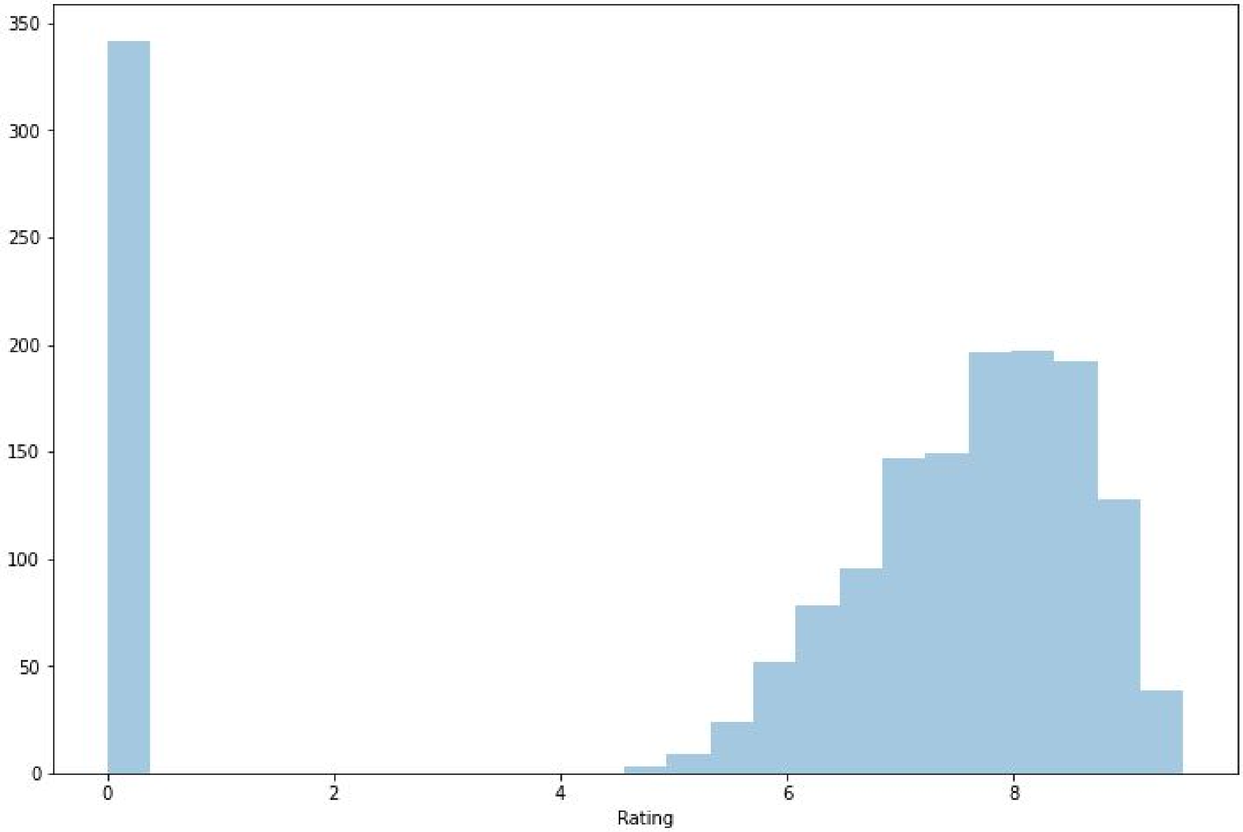
As a next processing step, we may want to exclude chain restaurants - e.g. Subway, The Meatball Shop. This step will remove 20 samples. Thanks to that step, the maximum number of restaurants in such chains is equal to 5 (3 samples), 4 (2 samples) - and the rest of them are below this value, so that will be our guarantee of unique experience.

Now, we have to split the dataset into 4 groups - due to the limit of Foursquare API for premium calls. Now, for each sample inside each group, using venue_id, we are calling endpoints that will return information such as: PriceTier, LikesCount, Rating, PhotosCount, ReasonsCount, TipsCount. After this we can join these 4 groups into one dataset. Now the dataset is ready for exploration.

3. Data exploration

3.1. Taking a look into distribution of ratings

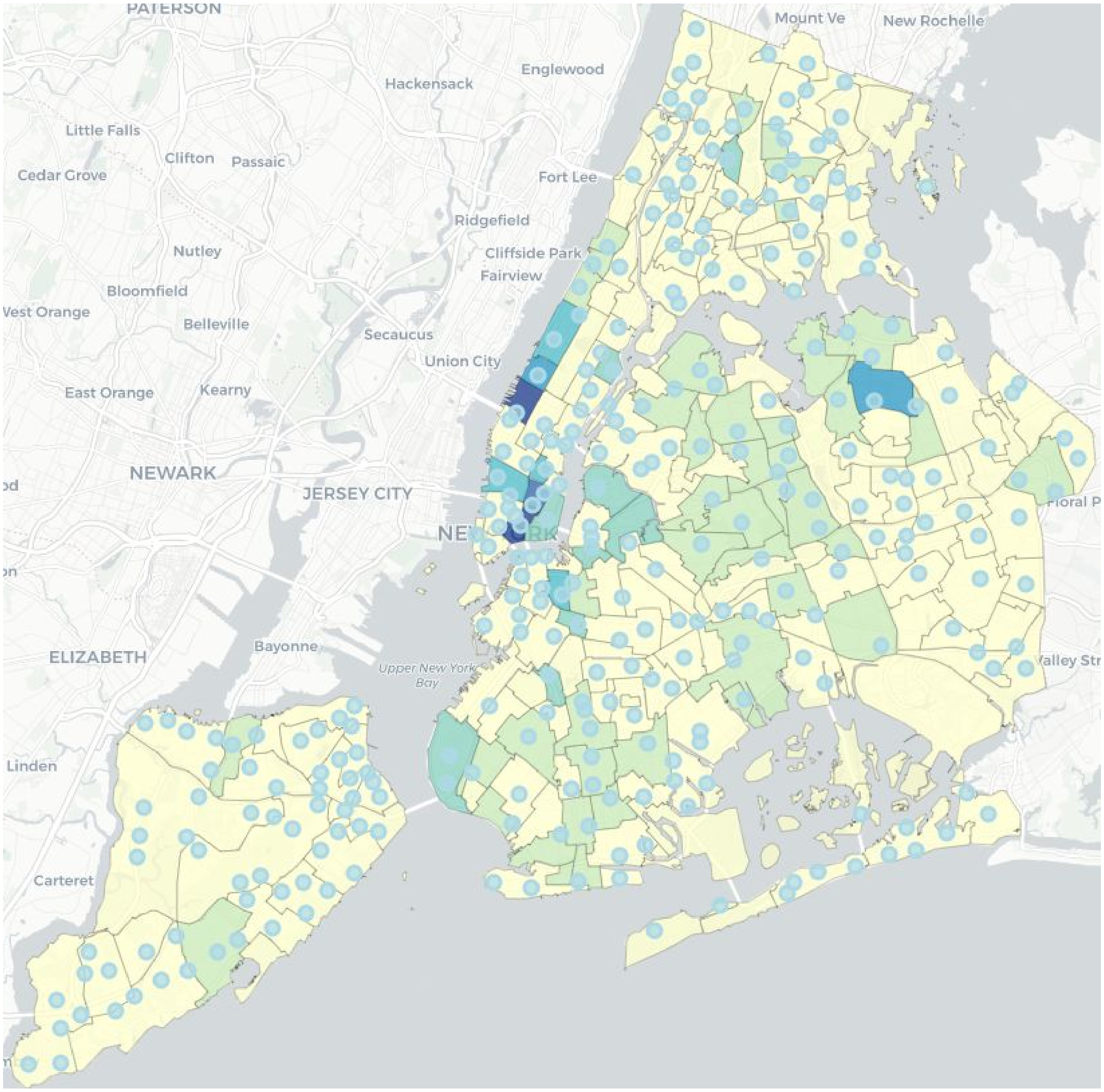
Almost everybody knows better and worse restaurants - but still we have to except, that most of them will be somewhere in the middle. In short, we are expecting a normal distribution of our ratings, with a low number of very low and very high ratings, and higher values in medium.



After taking a look into that distribution, we see that even if we cut off these with 0.0 ratings, as an outlier (even if there is a significant amount of them), distribution will not be so close to normal distribution. That's a surprising finding - because there is a bigger amount of them inside range 7.0 to 9.0 - and the ninth quantile is equal to 8.8, so 90% of samples are equal or below this value. So there are quite a lot of them with better ratings, but we should be interested in these best ones, best of the best - so these 10% better ones.

3.2. Taking a look into their distribution in map

Let’s find where we can find most restaurants - using folium library to create such a map.



After taking a quick look at this - we can see that there are a lot of them in the Manhattan area, because the top 10 neighborhoods, with highest numbers of restaurants (in overall) are in Manhattan.

3.3. Relationship between parameters and ratings

One of the ideas that could be valuable income and source of insights, is the relationship between ratings - so our target variable. One of the theories that I've been thinking about - are there any regions, neighborhoods with high ratings, so their localization has an influence on ratings. Another, about the number of reviews/likes/tips - if a place is more popular, we can expect that rating could be higher.
We will use in this and following sections two correlation metrics (methods). First of them is most popular, Pearson correlation

"The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable."

And Spearman correlation coefficient

"The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data."[1]

Driver	Pearson	Spearman
Venue latitude	0.09	0.007
Venue longitude	0.03	-0.02
Price Tier	0.21	0.18
Likes count	0.27	0.79
Photos count	0.21	0.68
Reasons count	0.6	0.64
Tips count	0.31	0.71

[1] https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/#what-is-correlation

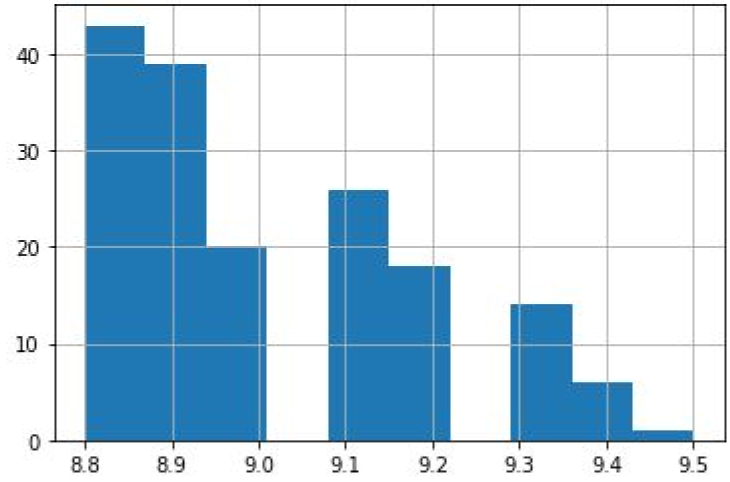
First theory failed - unfortunately, there is a lack of linear correlation between latitude and longitude and our target. Second theory is more optimistic - we see that likes, photos, reasons and tips have this positive influence on rating. What is surprising - lower influence has a price tier, looking both on Pearson and Spearman correlation coefficients.

What we should point out there, that this dataset is not filtered in any way - there could be some domain knowledge outliers, e.g. total new samples, some closed places etc.

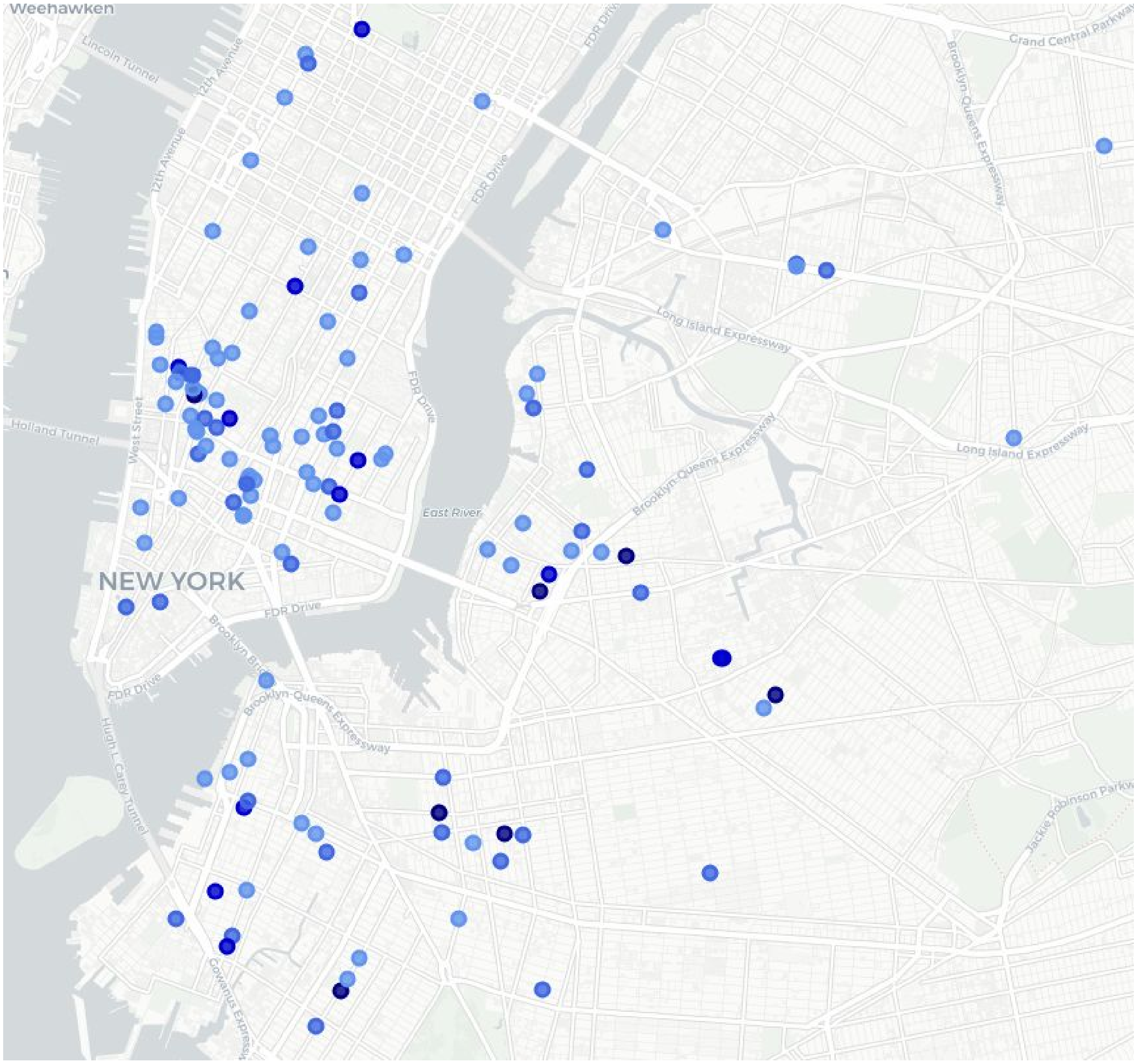
4. Results

4.1. Areas with best italian restaurants

First, let’s find where - in which neighborhood - we can find the best italian restaurants. We should start with selecting only these 10% of restaurants, the best ones.Thanks to that filtering, we will be able to just find these more interesting restaurants from our perspective - as we can see on distribution, there is significant amount of medium and high grade restaurants, so we have to select a good cut-off value - in this particular test, that will be score equal to 8.8.
We can now take a look into their distribution after this filtering



As we can expect, there is a tail for best values, with maximum value equal to 9.5 - thanks to that we know, that probably there is no bias connected with having a new restaurant, with one score equal to 10.
In our interest there will be about 167 restaurants. Where can we find most of them? Are they in every borough? Let’s take a look into that map



What we can see there, that significant amount of them is in Manhattan, as we probably can expect, but also - there are a lot of them in Williamsburg, Fort Greene. But what we can see there in broader perspective is, that accumulation of them we have in the central area of NY city, but these outside of this accumulation (e.g. in Brooklyn) can have better ratings, especially these ones in Fort Greene. Unfortunately, we can see that there are only a few of them in Queens, so that will not be a good place for eating Italian food.

4.2. Finding new place for restaurant

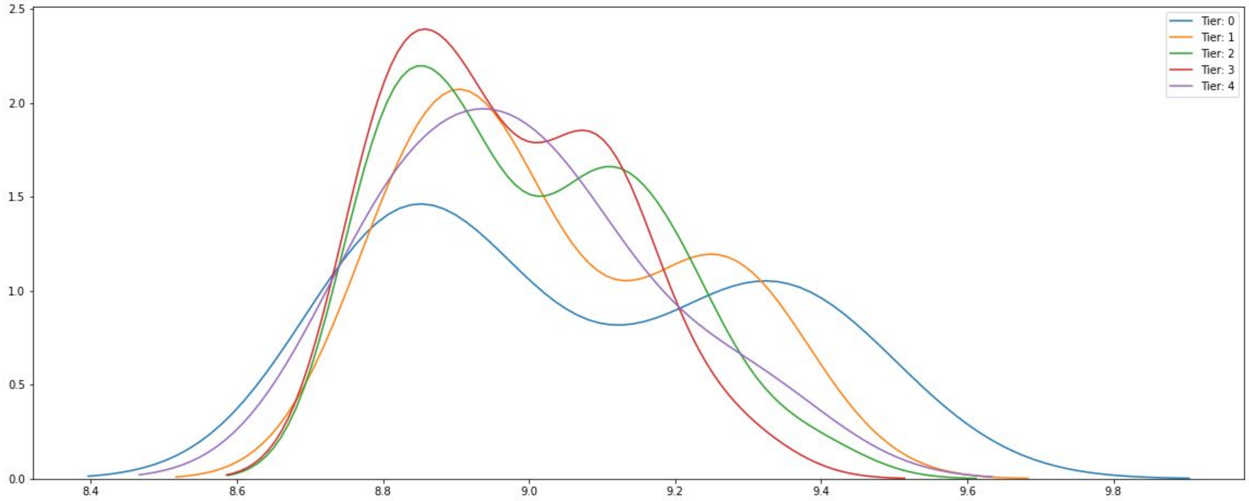
Using previous map, knowledge and insights that we have extracted from the previous section, we can think about possible areas for new restaurants. We have to consider there few factors:

- There has to be market for that restaurant, so that would be great if there are few restaurants, but the market cannot be overpopulated with restaurants of this type, because in that situation, that will be harder for owner to reach customer
- Availability of good restaurants in area - people very often look for similar restaurants in area of these good ones
- Not so far from city center
- Near communication nodes

According to that, we can say that perfect placing for that will be in following areas:

- Williamsburg
- East Williamsburg
- Fort Greene
- Brooklyn Heights
- Greenpoint

What we should also consider is that, what should be pricing and price/quality ratio.Take a look into this distribution for particular price tiers (0 - cheapest, 4 - most expensive ones)



What we can see there - that there can be some kind of cross-correlation between price Tier and scoring, but after checking Pearson correlation factor, we see value 0.15 - weak positive correlation. But what we see there, that this bimodal distribution of Tier 0 is pretty interesting from the perspective of fact, that better restaurants can be from cheaper tiers.

Let's review a few top restaurants and their price tier.

Borough	Neighborhood	Name	Rating	Price Tier
Brooklyn	Fort Greene	Mekelburg’s	**9.5**	**0**
Brooklyn	East Williamsburg	Lella Alimentari	9.4	1
Brooklyn	Gowanus	L'Albero dei gelati	9.4	2
Brooklyn	Williamsburg	L’Industrie Pizzeria	**9.4**	**0**
Brooklyn	Bushwick	Carmenta’s	**9.4**	**0**
Manhattan	West Village	Faicco’s Italian Specialities	9.4	2
Brooklyn	Fort Greene	Evelina Restaurant	**9.4**	**0**
Brooklyn	Carroll Gardens	Court Street Grocers	**9.3**	**0**
Manhattan	Lincoln Square	Marea	9.3	4
Manhattan	Gramercy	Eataly Flatiron	**9.3**	**0**

Half of the top 10 restaurants have Price Tier 0 - even if there are no trends behind this behaviour, that good insights and knowledge - maybe people that are making reviews and rating are including price tier as one of the elements?

4.3. In which area we have lack of good italian restaurants

Now, we have to look at the opposite site of our distribution - to these low rated restaurants. Are there any particular neighborhoods that have such low levels of ratings? What can be behind that? Take a look into the following table, which will contain top 10 worst neighborhoods with their boroughs.

Borough	Neighborhood	Number of restaurants	Average rating
Staten Island	Woodrow	1	0.0
Staten Island	Concord	2	0.0
Staten Island	Castleton Corners	2	0.0
Brooklyn	Homecrest	1	0.0
Staten Island	Bay Terrace	4	0.0
Staten Island	Annandale	1	0.0
Staten Island	Midland Beach	1	0.0
Staten Island	New Drop Beach	3	0.0
Brooklyn	Brighton Beach	3	0.0
Staten Island	Pleasant Plains	1	0.0

That can tell us a lot - but there are almost 30 of neighborhoods with this average rating. We have to build a bigger picture of that - let’s take a look at these scores, but with taking into account not only average value - but also median value.

Borough	Number of restaurants	Average rating	Median rating
Bronx	124	5.19	7.10
Brooklyn	393	6.16	7.50
Manhattan	925	6.49	7.40
Queens	116	5.28	7.15
Staten Island	**94**	**3.62**	**0.0**

And now we have a bigger picture of whole situation - and we have a confirmation, that Staten Island is worst borough in this topic - number of restaurants is lowest, average rating is lowest, but what is also very important - median is giving us information about fact, that over half of restaurants there have rating equal to 0.0.
Also, we see that Queens and Bronx are also not a good choice for such a place, like also a place for looking for these best ones - but that a look into the number of them after this 10% best filtering.

Borough	Number of restaurants	Average rating	Median rating
Bronx	**10**	**8.9**	**8.9**
Brooklyn	56	9.0	9.1
Manhattan	88	8.98	8.9
Queens	**8**	**8.98**	**8.95**
Staten Island	**5**	**8.98**	**9.0**

As we see, there are few rare pearls at Staten Island, Bronx and Queens - and they have good opinions - but in comparison to the full picture, these are places for one-night visits, but not for longer stays.

4.4. Which area would be best for longer stay, to visit multiple neighborhoods

And when we are talking about the worst and best places to visit, can we say clearly what will be the best place, borough, to visit for a few days, to explore new places?
According to all previous tables, we can say clearly, that this boroughs are:

- Brooklyn
- Manhattan

And what is the best place for that? In that situation, we have to find a place in the middle, near communications nodes that allow us to travel between points relatively fast, with nearest restaurants not so far, even for evening walks. In that situation, I would like to recommend one neighborhood - **Brooklyn Heights**, which are in perfect distance to all the best restaurants in the area, which has few good restaurants.

5. Discussion

In this experiment, I've analyzed only the restaurants itself, based on few most important parameters - but there are also few additional factors that could be considered, e.g. number of reasons, tips, photos. Also, the dataset is a little bit outdated - it is based on the 2018 version of API, so part of information could not be relevant now.

What should be included there also, we can think about adding some demographic data there - to add demographic context to each borough/neighborhood if that will be possible, like also we can take into account sentiment analysis of reviews/tips - that could be also useful in that type of analysis.

What should be also considered there, is also visibility of places, where by visibility I mean what people from the nearest area know and what they think about particular places.

6. Conclusion

This analysis shows us huge disparities between neighborhoods and boroughs, which are visible on level of amount and quality of restaurants - and which can be also confirmed and determined from demographic data, economic and financial information, about quality of neighborhoods. We know where to find the best food, we know what places should be avoided - but should we support these smaller places that are in not so good neighborhoods? Even if the price of that visit is the quality of food, we should consider giving them a chance - maybe they are so niche, that they cannot break the wall and gather more customers? There are so many questions and issues, and so many other drivers that are influencing which restaurants can be considered as the best ones.

Or maybe it all depends on individual taste?