Report for Capstone Project, “The Battle of Neighborhoods”

Applied Data Science Capstone

1/2/2021

1. Introduction

Pizzerias are one of the most emblematic icons of New York City, and one of the most lucrative business opportunities: a popular venue can easily earn millions of dollars in profits per year. However, the pizza market is rather crowded, with hundreds of pizza restaurants calling the city home. In addition, the potential downside of a failed restaurant is high. With startup costs running in the hundreds of thousands of dollars for ovens, kitchen fixtures, HVAC modifications, and dining area improvements. Few if any of these startup costs can be recouped in the event that the restaurant must be closed, creating the potential for large losses. How can an aspiring restauranteur maximize their chances of success in this challenging and high-stakes market?

Location is the most important feature of any business. Starting a pizzeria in a location that has little interest in pizza, or in a place that is already well-served by popular established venues, leaves little chance for success. Conversely, finding a location that has high demand for pizza, but is poorly served by existing establishments, creates great opportunities for a new pizzeria.

The business problem addressed by this report is to find the best NYC neighborhoods to start a new pizza location. The first criterion is that the neighborhoods should have high demand for pizza, as measured by the number of existing pizzerias in the neighborhood. The second criterion is that the neighborhoods should be poorly served by existing pizzerias, as measured by low ratings for existing venues. By finding neighborhoods with high demand but poor existing options, we will locate the most profitable potential locations for a startup pizzeria.

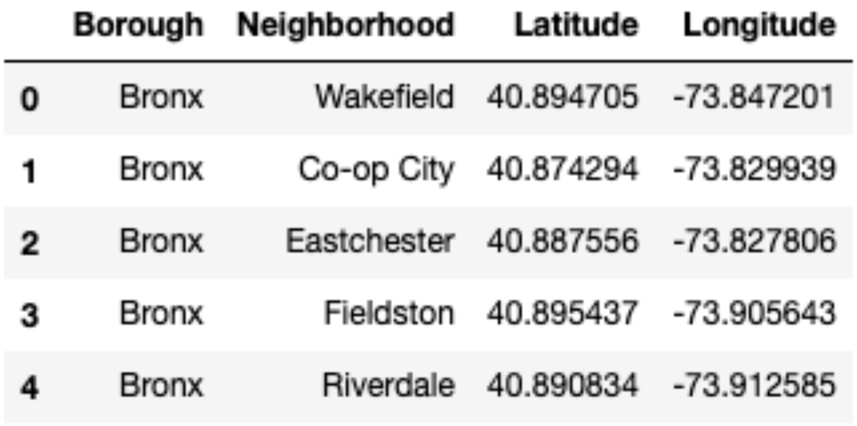
This report will be of interest to entrepreneurs who are interested in starting a new pizza retail business in New York City, but who do not yet know where their business should be located.

2. Data

This project draws on three sources of data. The first source gives the locations of all the neighborhoods in New York City. The second identifies, for each neighborhood, what venues belong to that neighborhood. The third contains the ratings for each venue.

The first data set, containing the neighborhood location data, has been supplied by the Coursera instructor for the purposes of completing a lab assignment. They originally came from a publicly available source such as Wikipedia. The data file contains the neighborhood name (unique identifier), borough, latitude, and longitude for each of the 306 neighborhoods in New York City. This data file is located at: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json")

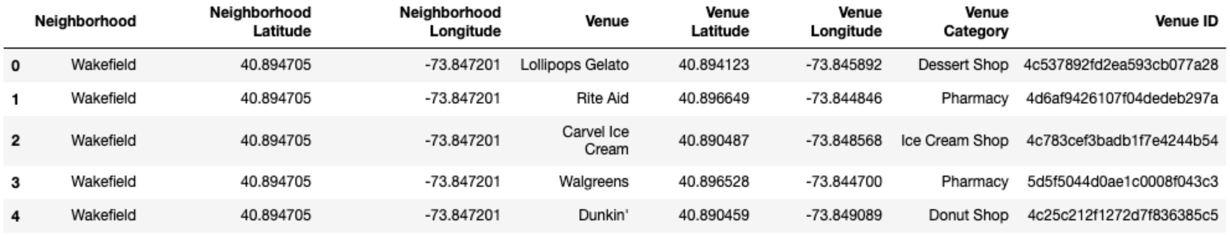
Here is what the first few lines of the first data file looks like:



	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

The second data set, which identifies what venues belong to each neighborhood, is downloaded from Foursquare, which is a local search-and-discovery app developed by Foursquare Labs. Foursquare Labs is a privately-held technology company with approximately 400 employees. Foursquare City Guide has over 50 million users and has been operating since 2009. It provides information on millions of venues around the world. This data set contains the venue name, latitude, longitude, category, and venue ID (unique identifier) for 10,159 venues in NYC. It is downloaded using an "explore" endpoint, which is a non-premium API request.

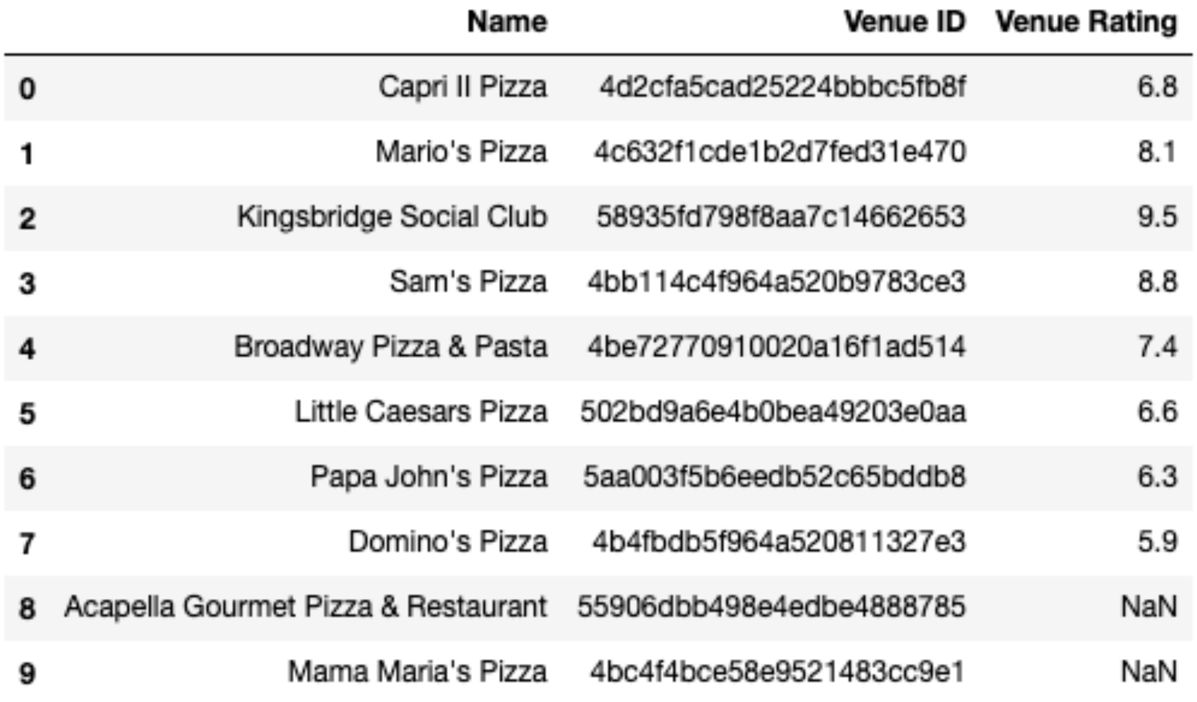
Here is what the first few lines of the second data set look like:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop	4c537892fd2ea593cb077a28
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy	4d6af9426107f04dedeb297a
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop	4c783cef3badb1f7e4244b54
3	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy	5d5f5044d0ae1c0008f043c3
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop	4c25c212f1272d7f836385c5

The third data set, which contains user-provided ratings for the pizza venues in New York City, is also downloaded from Foursquare. It contains the name, venue ID (unique identifier), and average user rating for the 440 pizza venues in New York City. Because not all pizza venues have user ratings on Foursquare, the dataset contains 84 missing ratings values, leaving 356 valid data points. This dataset is downloaded using a “details” endpoint, which is a premium API request.

Here is what the first few lines of the third data set look like:

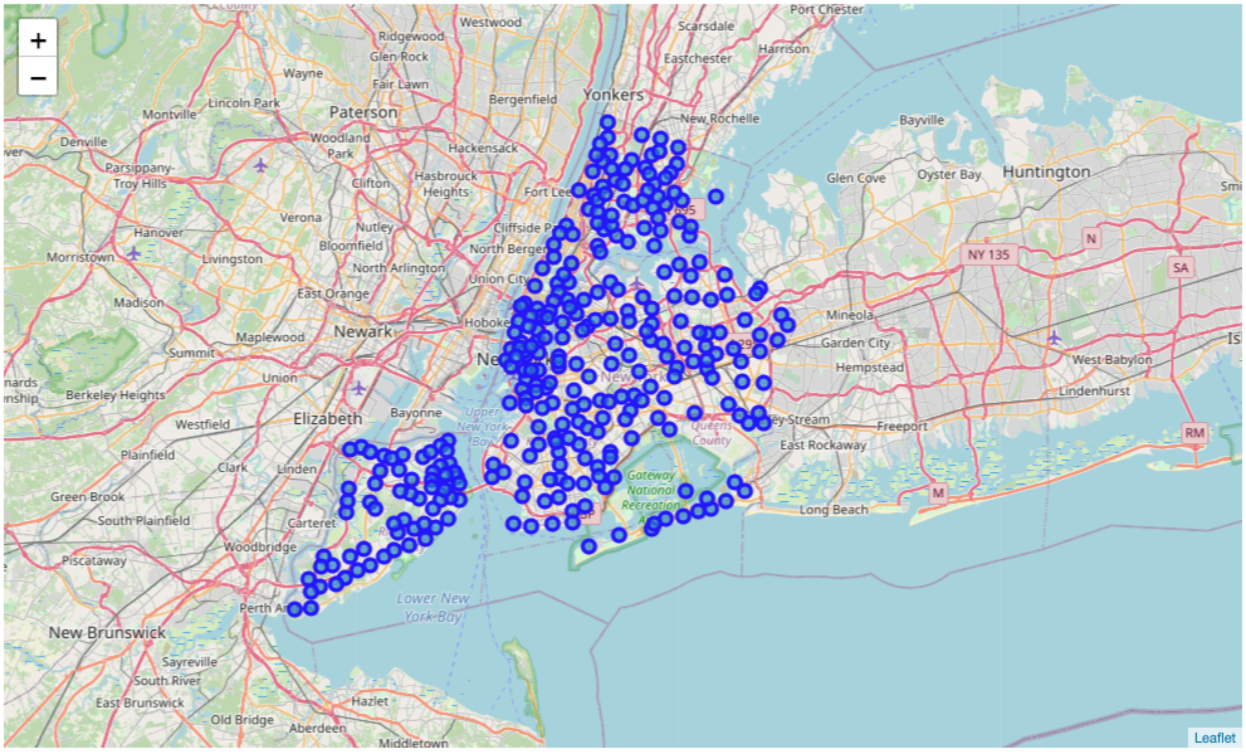


	Name	Venue ID	Venue Rating
0	Capri II Pizza	4d2cfa5cad25224bbbc5fb8f	6.8
1	Mario's Pizza	4c632f1cde1b2d7fed31e470	8.1
2	Kingsbridge Social Club	58935fd798f8aa7c14662653	9.5
3	Sam's Pizza	4bb114c4f964a520b9783ce3	8.8
4	Broadway Pizza & Pasta	4be72770910020a16f1ad514	7.4
5	Little Caesars Pizza	502bd9a6e4b0bea49203e0aa	6.6
6	Papa John's Pizza	5aa003f5b6eedb52c65bddb8	6.3
7	Domino's Pizza	4b4fbdb5f964a520811327e3	5.9
8	Acapella Gourmet Pizza & Restaurant	55906dbb498e4edbe4888785	NaN
9	Mama Maria's Pizza	4bc4f4bce58e9521483cc9e1	NaN

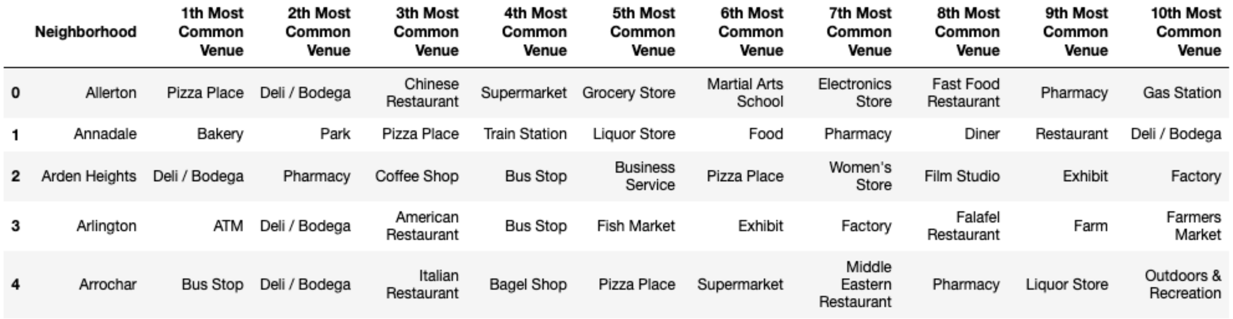
To formulate our business recommendations based on this data, we merge the three data files. First, the pizza venues are assigned to neighborhoods using the latitude and longitude coordinates. Second, the ratings are merged with the venue data using the venue ID as the unique identifier.

3. Methodology

In this section of the report, I discuss my methodology and descriptive data analysis used for generating business recommendations from the data. First, after loading the neighborhoods data, I mapped it using Folio to check that all the neighborhoods are located in New York City and that everything appears correct:

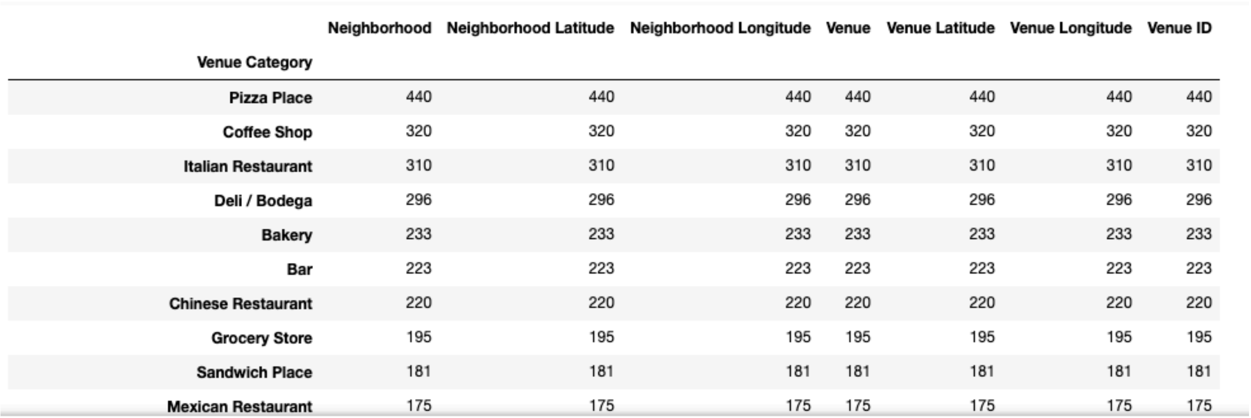


As we can see, all the neighborhoods appear to be located in New York City, and are evenly distributed across the city. Next, to get a feel for what types of venue are located in which neighborhood, I created a data frame showing the top ten most common venue types for each neighborhood:



	Neighborhood	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allerton	Pizza Place	Deli / Bodega	Chinese Restaurant	Supermarket	Grocery Store	Martial Arts School	Electronics Store	Fast Food Restaurant	Pharmacy	Gas Station
1	Annadale	Bakery	Park	Pizza Place	Train Station	Liquor Store	Food	Pharmacy	Diner	Restaurant	Deli / Bodega
2	Arden Heights	Deli / Bodega	Pharmacy	Coffee Shop	Bus Stop	Business Service	Pizza Place	Women's Store	Film Studio	Exhibit	Factory
3	Arlington	ATM	Deli / Bodega	American Restaurant	Bus Stop	Fish Market	Exhibit	Factory	Falafel Restaurant	Farm	Farmers Market
4	Arrochar	Bus Stop	Deli / Bodega	Italian Restaurant	Bagel Shop	Pizza Place	Supermarket	Middle Eastern Restaurant	Pharmacy	Liquor Store	Outdoors & Recreation

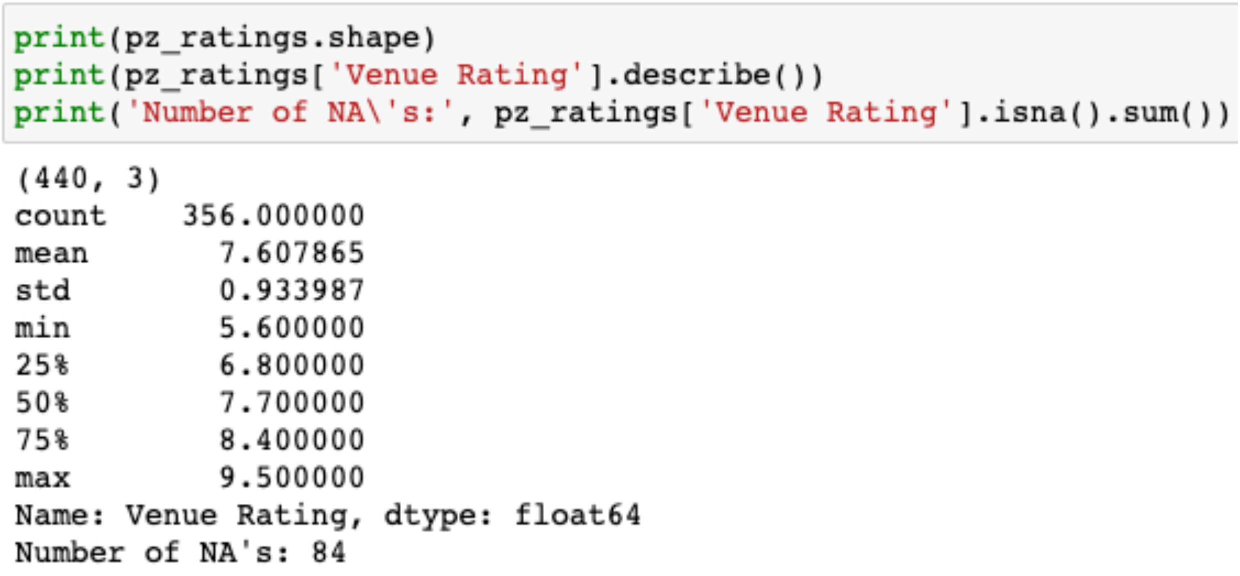
As we can see, pizza places are the most common venue in some neighborhoods. Other common business venue types are bakeries, delis, and pharmacies. Next I created a data frame showing the most common venue types for the city as a whole, irrespective of neighborhood, sorted from most to least common:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue ID
Venue Category							
Pizza Place	440	440	440	440	440	440	440
Coffee Shop	320	320	320	320	320	320	320
Italian Restaurant	310	310	310	310	310	310	310
Deli / Bodega	296	296	296	296	296	296	296
Bakery	233	233	233	233	233	233	233
Bar	223	223	223	223	223	223	223
Chinese Restaurant	220	220	220	220	220	220	220
Grocery Store	195	195	195	195	195	195	195
Sandwich Place	181	181	181	181	181	181	181
Mexican Restaurant	175	175	175	175	175	175	175

As we can see, pizza places are by far the most common venue type in the city, with 440 venues. This demonstrates the social, economic, and business impact of this type of venue. Pizza places are more than 30% more common than the next most common business type, coffee shops.

Next I used the venue ID's for all the pizza venues in the city to scrape the pizza venue ratings from Foursquare. I then did some descriptive statistics on the ratings, to show the percentiles, min, max, and number of missing values:

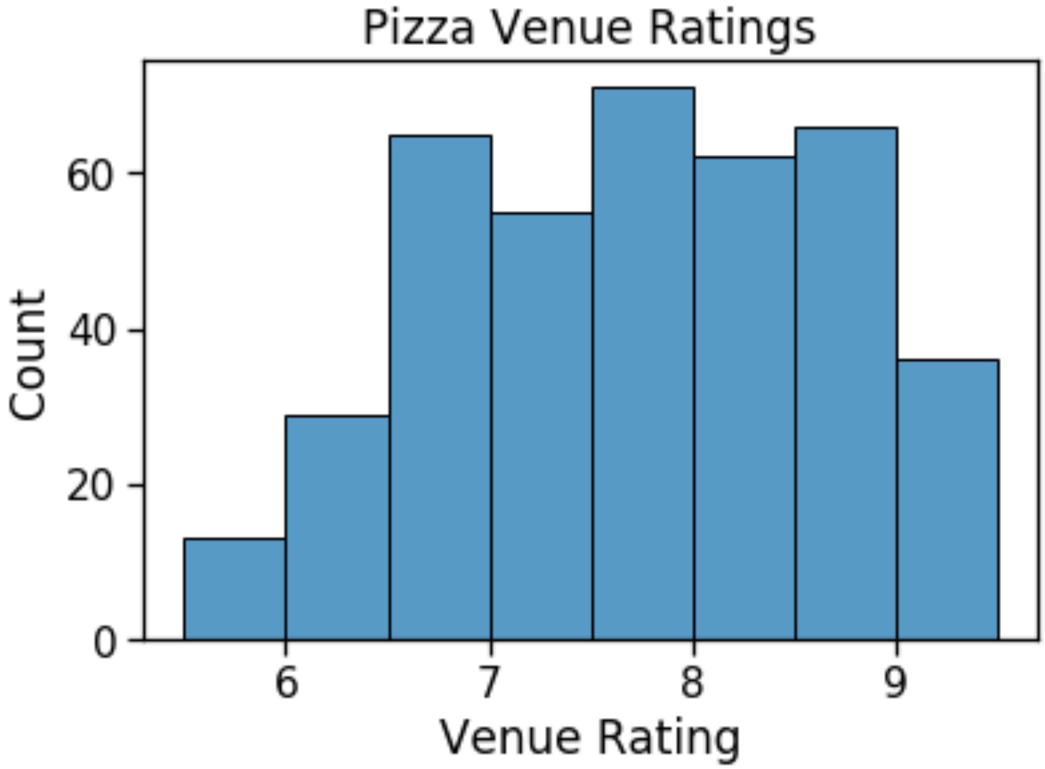


```
print(pz_ratings.shape)
print(pz_ratings['Venue Rating'].describe())
print('Number of NA\'s:', pz_ratings['Venue Rating'].isna().sum())
```

```
(440, 3)
count    356.000000
mean       7.607865
std        0.933987
min        5.600000
25%        6.800000
50%        7.700000
75%        8.400000
max        9.500000
Name: Venue Rating, dtype: float64
Number of NA's: 84
```

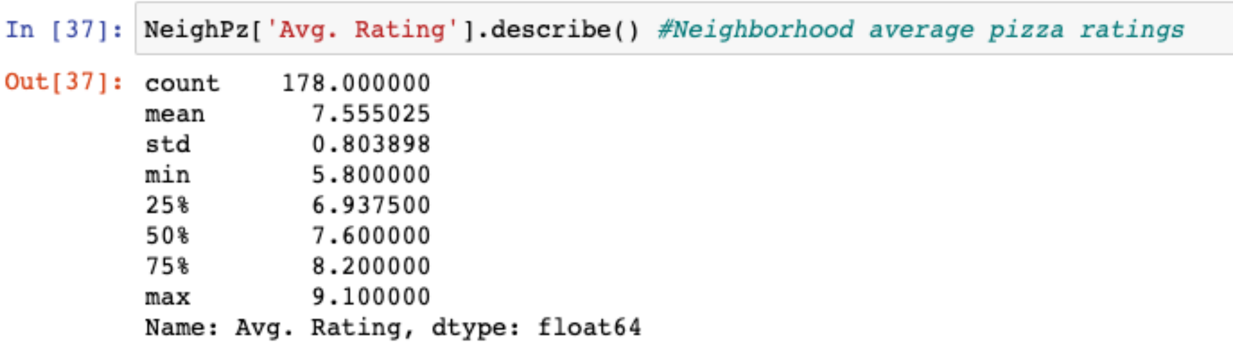
This analysis shows that the median pizza venue rating was 7.7. This is a relatively high rating, giving a sense of the competitiveness of the pizza market and the high quality of many of the existing venues. This is not surprising, given that NYC is famous for its pizza! However, the 25th percentile rating is only 6.8, a mediocre rating indicating that there are many relatively low-quality pizza venues in the city. This indicates that there is room for improvement and a possibility of opening a new pizza restaurant that improves on existing offerings.

The same pizza venue ratings data can be visualized in a histogram:



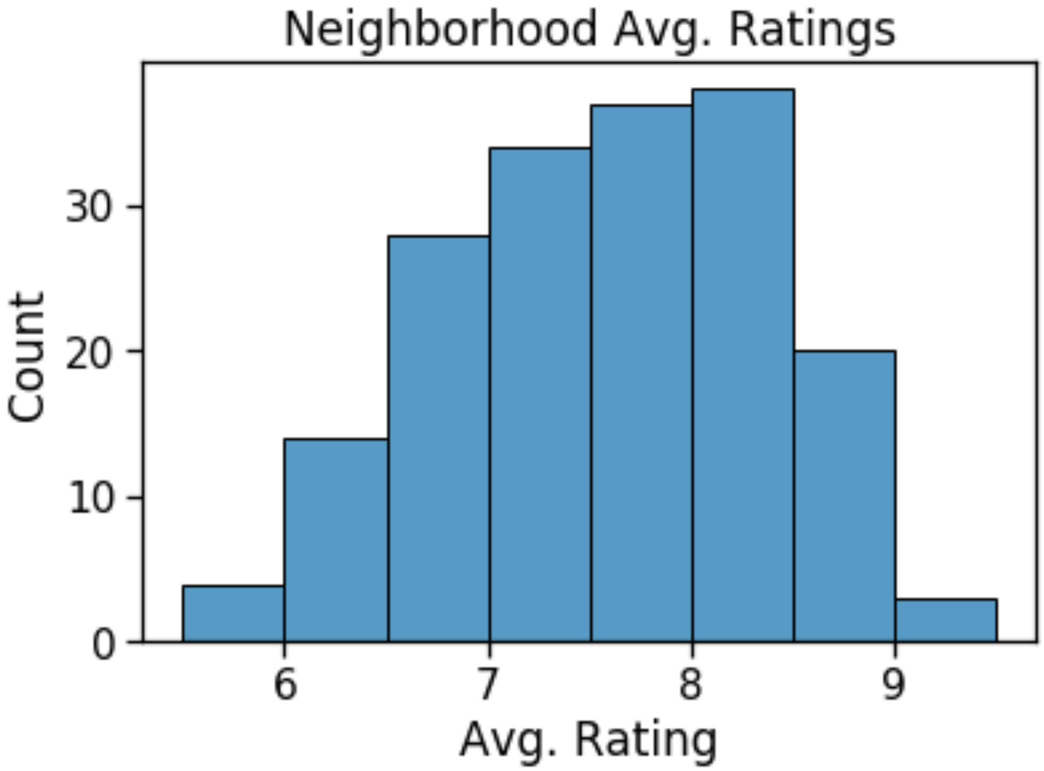
The histogram indicates that the ratings are approximately normally distributed, with most of the ratings falling between 7 and 9. Again, we see a substantial tail of venues with ratings below 7, indicating that there is room for disruption in the NYC pizza market.

We then perform descriptive statistics on the average ratings for the pizza venues in a particular neighborhood, and display as a table:



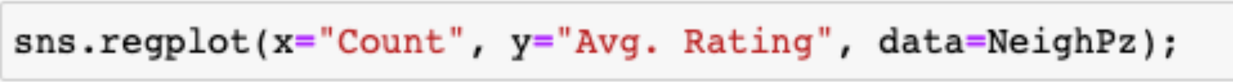
```
In [37]: NeighPz['Avg. Rating'].describe() #Neighborhood average pizza ratings

Out[37]: count    178.000000
         mean       7.555025
         std        0.803898
         min        5.800000
         25%        6.937500
         50%        7.600000
         75%        8.200000
         max        9.100000
         Name: Avg. Rating, dtype: float64
```

We see that among the 178 neighborhoods with pizza venues, the median neighborhood has an average rating of 7.6. We also see that 25% of neighborhoods have a median rating below 7.0, indicating that many neighborhoods have room for improvement. Here is a histogram of the same data, showing the distribution of average neighborhood ratings:

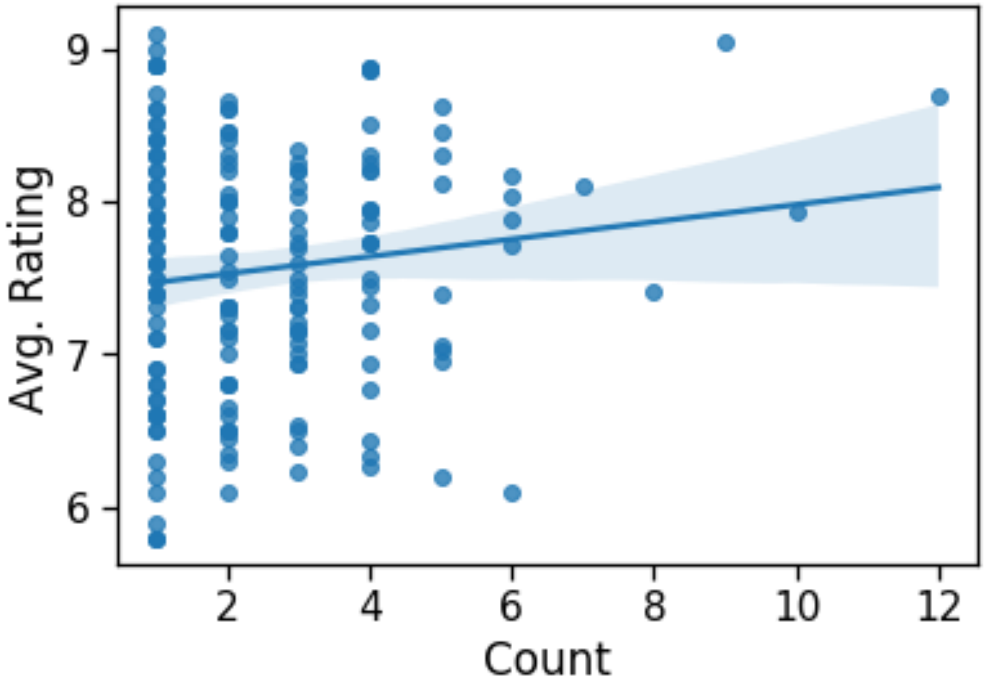


Similarly to the venue ratings, the neighborhood average ratings are approximately normally distributed, with a substantial tail of neighborhoods with average ratings below 7. This indicates that there are many neighborhoods where a new pizza venue could substantially improve upon existing offerings.

What we are looking for is a neighborhood with many existing pizza venues, indicating substantial demand for pizza in the neighborhood, but with a low average rating, indicating an opportunity for disruption by a new entry. To get an idea for how difficult it will be to find neighborhoods that meet our two criteria, we perform a linear regression of the average pizza rating in a neighborhood, against the count of pizza venues in the neighborhood:

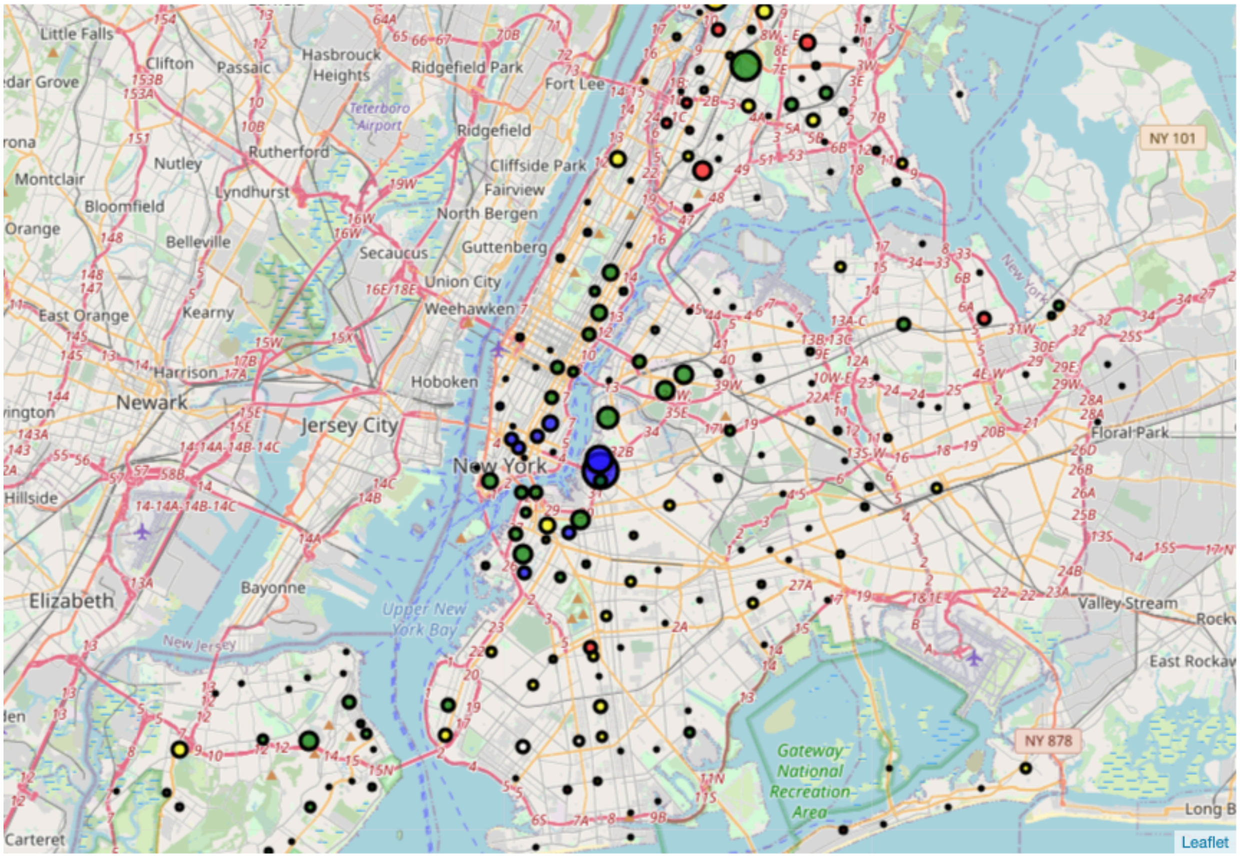


```
sns.regplot(x="Count", y="Avg. Rating", data=NeighPz);
```



We see from the plot that there is a slight positive relationship, not quite statistically significant at 95% confidence, between the number (count) of pizza venues in a neighborhood and their average quality. We see that there is some support in the bottom center of the graph, with several neighborhoods with a substantial number of pizza places (> 4), but a relatively low average quality (< 7). These will be good places to start looking for the ideal location of our new pizza place.

To visualize the density and quality of pizza locations in New York City, we create a map, with circle markers with size corresponding to the number of pizza locations in a neighborhood, and color corresponding to their average quality:

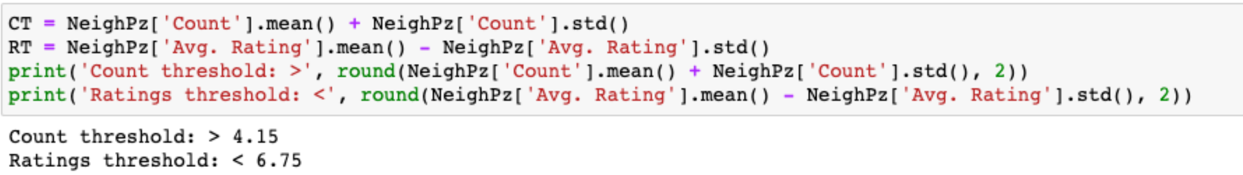


In this map, the larger circles correspond to neighborhoods with more pizza venues. The color indicates average quality, with blue circles having the best average quality (above 8.5), green coming next in quality (between 7.5 and 8.5), yellow having mediocre quality (between 6.5 and 7.5), and red indicating poor quality (below 6.5).

We see from the map that downtown Manhattan and west Brooklyn are already well-served by a large number of high-quality pizza establishments, which we can see from the large blue and green circles in these locations. Conversely, we can see neighborhoods in the Bronx with large red circles, indicating a high demand for pizza, but low quality of existing venues. These will be the places to look for ideal locations for our new pizza restaurant.

4. Results

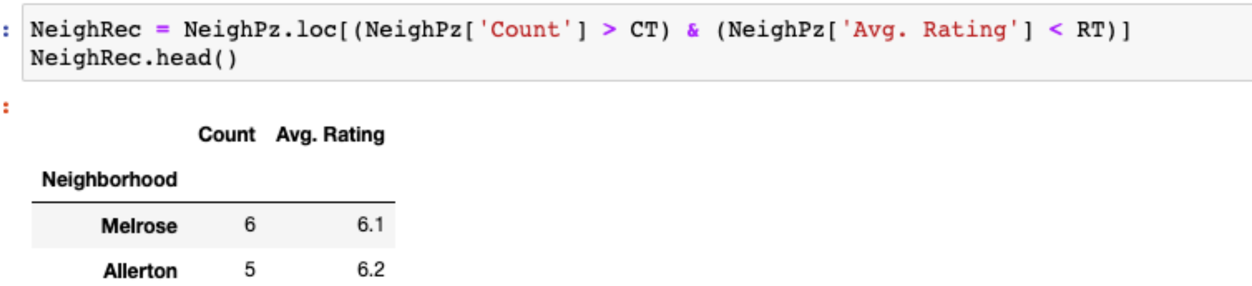
To narrow down our search for the best neighborhoods to start a new pizza restaurant, we establish formal criteria based on the number of existing venues and their quality. We define a neighborhood with a high demand for pizza as one where the number of existing venues is more than one standard deviation from the mean. Furthermore, we define a neighborhood with poor-quality existing venues as one in which the average quality is more than one standard deviation below the mean. This gives rise to thresholds for count and ratings that we can use to narrow down our list of neighborhoods:



```
CT = NeighPz['Count'].mean() + NeighPz['Count'].std()
RT = NeighPz['Avg. Rating'].mean() - NeighPz['Avg. Rating'].std()
print('Count threshold: >', round(NeighPz['Count'].mean() + NeighPz['Count'].std(), 2))
print('Ratings threshold: <', round(NeighPz['Avg. Rating'].mean() - NeighPz['Avg. Rating'].std(), 2))
```

```
Count threshold: > 4.15
Ratings threshold: < 6.75
```

Thus we conclude that we want a neighborhood with more than 4 pizza venues, with an average quality of less than 6.75. We apply these criteria to our list of neighborhoods to arrive at our final recommendations:

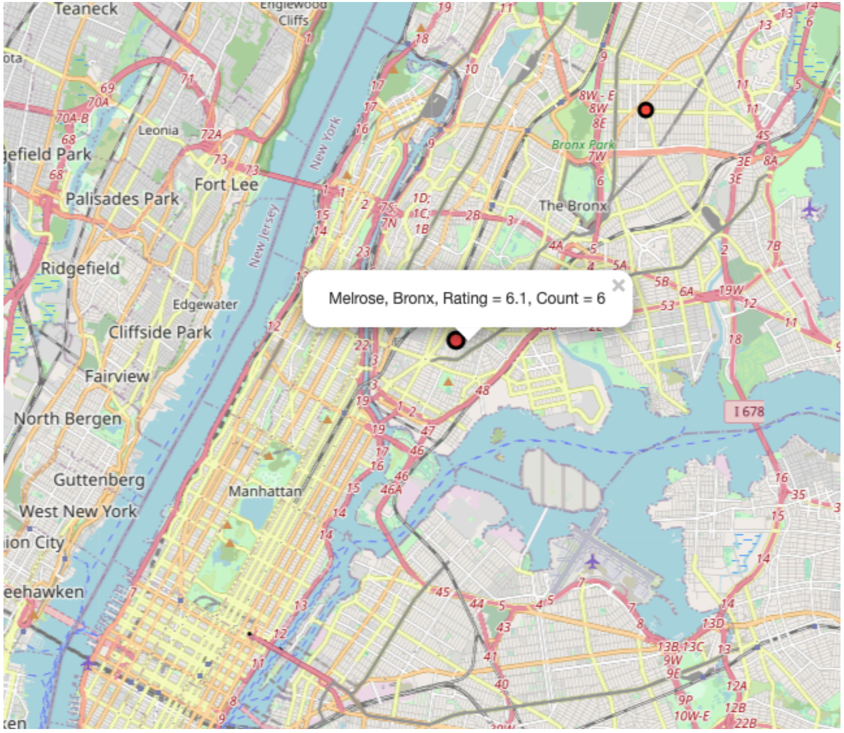


```
NeighRec = NeighPz.loc[(NeighPz['Count'] > CT) & (NeighPz['Avg. Rating'] < RT)]
NeighRec.head()
```

	Count	Avg. Rating
Neighborhood		
Melrose	6	6.1
Allerton	5	6.2

We see that the two neighborhoods in New York City that meet our criteria are Melrose and Allerton, both in the Bronx. The former has 6 pizza restaurants, with an average rating of 6.1, and the latter has 5 venues, with an average rating of 6.2. Either would be a fine place to start a new pizza restaurant, based on our criteria, and the final choice should be determined by an analysis of available commercial restaurant space and the associated rent and expenses.

Here is a map of our recommended neighborhoods, and their location in the Bronx:



Melrose, Bronx, Rating = 6.1, Count = 6

5. Discussion

We have seen above that, of the 440 established pizza restaurants in New York City, 397 have ratings available on Foursquare. Many of the established venues have high ratings, but 25% have ratings below 7, indicating substantial opportunity for disruption.

In addition, while the neighborhoods in downtown Manhattan and western Brooklyn are well served by existing venues, we have identified areas in the Bronx that have a high demand for pizza and are poorly served by existing venues. These two neighborhoods would be ideal places to open a new pizza restaurant, which was the purpose of this report.

6. Conclusion

Our analysis is driven by three data sources, including location data for neighborhoods, locations of venues, and ratings of venues. We analyzed the data using linear regression and exploratory descriptive statistics. Based on this analysis, we conclude that the best places in New York City to open a new pizza restaurant are the neighborhoods of Melrose and Allerton, both in the Bronx. These neighborhoods are ideal because they combine a high demand for pizza with low quality of existing venues, indicating high potential for disruption by a startup pizzeria.

Our analysis could be improved by the inclusion of more data. 84 of the city's pizzerias had no ratings information on Foursquare. One way to remedy this would be to include ratings information from other platforms, such as Yelp, to fill in the missing values. Our data could also be supplemented by an analysis of other features found to be predictive of pizza venue quality, such as hours of operation, foot traffic, square footage, and revenue.

Our business recommendations could also be improved by including data on the costs of doing business in each of the proposed locations. These costs could include commercial real estate rents, advertising, shipping of raw materials, and regulatory compliance. These data, particularly real estate rents, are easily available and could be included in subsequent analysis. By formulating concrete recommendations on the basis of publicly available data, this report has demonstrated the importance of data science and visualization for informing business decisions.