Restaurants in New York, NY

Bryan J. Hernandez

May 17, 2020

1. Introduction

1.1. Background

New York, NY is known for their food and especially for their Italian food. With the growing market of restaurants around the world and especially in New York city, it is important to analysis the market before deciding on a location to open. For many businesses a fatal flaw in their business model is entering into a market that is already too dense to penetrate. Another barrier to entry for businesses is the competitions value proposition, thus it is important to analysis and take into consideration attributes of the venues around the area.

1.2. Audience

The target audience for this report is someone looking to enter into the restaurant industry within New York. Another target audience could be a market researcher who is looking to understand the demographic in more detail, having dining establishments being an attribute of their analysis.

2. Data acquisition and cleaning

2.1. Data Sources

The primary data source will be geographical data acquired via the foursquare API. The data that will be pulled starts from at the epicenter, New York, NY and is expanded 25 mile radius. An example of the dataset would include the longitude/latitude, distance from center, and other features describing the individual venues.

2.2. Data Cleaning

The data is pulled using foursquare api and comes down in a json format. This base dataset is cleaned, transformed and filtered for the applicable venue categories. To add onto the dataset another dataset that pulls the venues detailed information is retrieved, cleaned and joined with the base dataset to add the additional features that are going to be used to analysis the venues further.

2.3. Feature Selection

The features were heavily dependent upon the foursquare API limitations. As there was a limitation on the number of features, the feature selection was limited in nature. However, the features that were able to be pulled included:

a. Distance

b. Rating

c. Likes

d. Tier

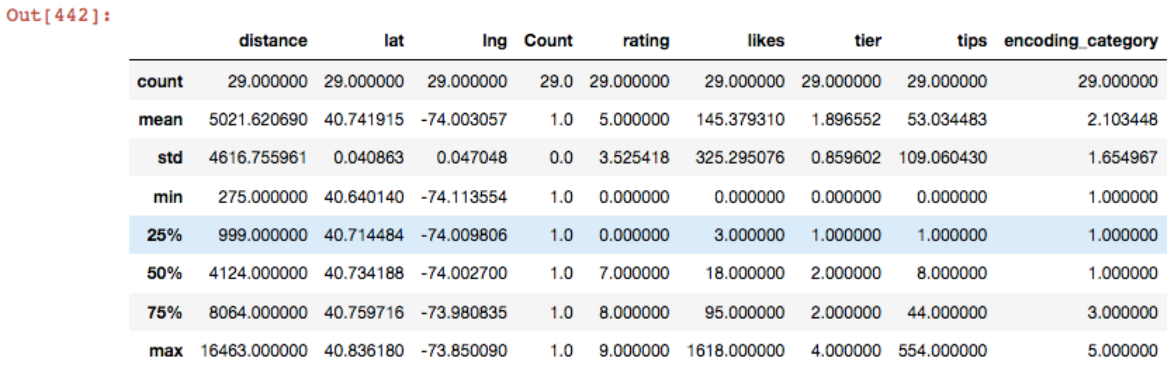
e. Tips

The features will provide some basic insights into the market of the venues surrounding New York, NY. We were able to look into the proximity of the venues from the desired location, analyze/bin the ratings, likes and tips to get a better idea regarding customer feedback. Furthermore, we were able to look into the different tiers of venues to get insights regarding level of quality.

3. Methodology

3.1. Exploratory Data Analysis

The exploratory analysis started with encoding the categories for the venues. Some example categories included: Food, Restaurant, Pizza Place. After the encoding the categories we ran some basic descriptive statistics to understand the features a little more.

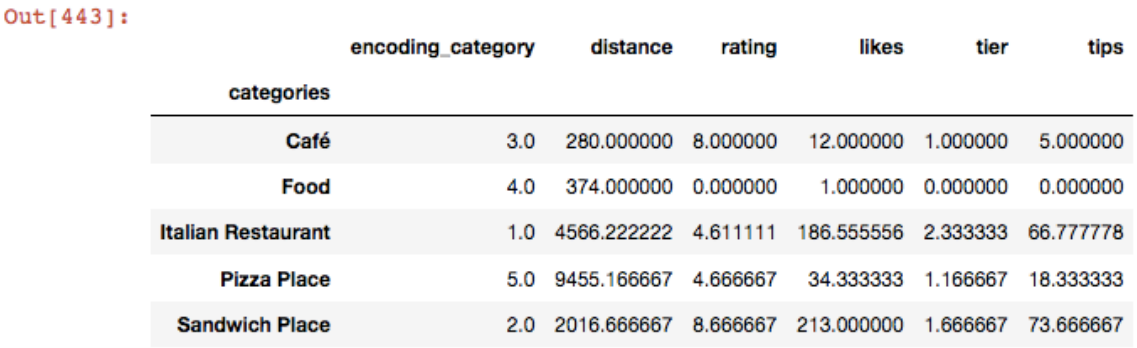


Out[442]:

	distance	lat	lng	Count	rating	likes	tier	tips	encoding_category
count	29.000000	29.000000	29.000000	29.0	29.000000	29.000000	29.000000	29.000000	29.000000
mean	5021.620690	40.741915	-74.003057	1.0	5.000000	145.379310	1.896552	53.034483	2.103448
std	4616.755961	0.040863	0.047048	0.0	3.525418	325.295076	0.859602	109.060430	1.654967
min	275.000000	40.640140	-74.113554	1.0	0.000000	0.000000	0.000000	0.000000	1.000000
25%	999.000000	40.714484	-74.009806	1.0	0.000000	3.000000	1.000000	1.000000	1.000000
50%	4124.000000	40.734188	-74.002700	1.0	7.000000	18.000000	2.000000	8.000000	1.000000
75%	8064.000000	40.759716	-73.980835	1.0	8.000000	95.000000	2.000000	44.000000	3.000000
max	16463.000000	40.836180	-73.850090	1.0	9.000000	1618.000000	4.000000	554.000000	5.000000

We can see that the average distance from the epicenter is approximately 5,000 meters or 3 miles. Furthermore, we see that ratings have bounds of [0,9] with an average rating for the 29 restaurants settled at 5/10.

After running some basic descriptive statistics, next was grouping the average values by the categories of the venues.

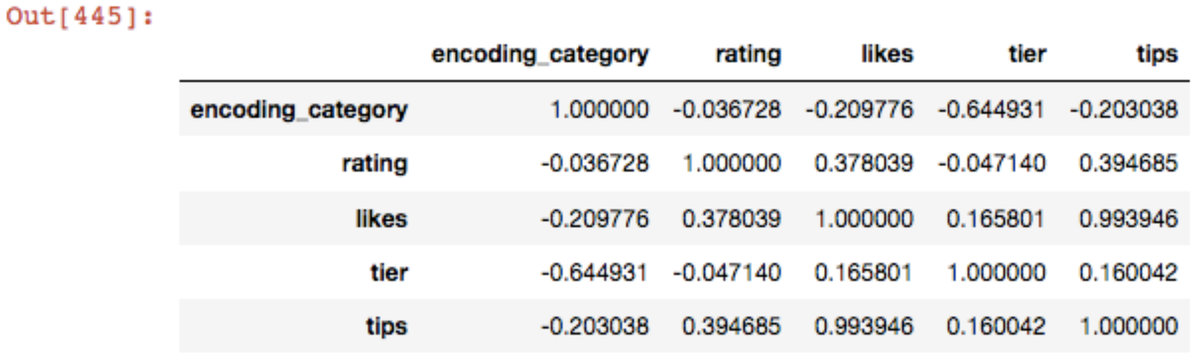


Out[443]:

	encoding_category	distance	rating	likes	tier	tips
categories						
Café	3.0	280.000000	8.000000	12.000000	1.000000	5.000000
Food	4.0	374.000000	0.000000	1.000000	0.000000	0.000000
Italian Restaurant	1.0	4566.222222	4.611111	186.555556	2.333333	66.777778
Pizza Place	5.0	9455.166667	4.666667	34.333333	1.166667	18.333333
Sandwich Place	2.0	2016.666667	8.666667	213.000000	1.666667	73.666667

We can notice that Italian Restaurants have an average rating of 4.61, which can indicate a potential opening in the market by the venues having a low value proposition. It can also be seen that the pizza place category is the furthest away in terms of average distance. We could look into and focus in on this subset to identify if there are gaps in the market closely surrounding the epicenter.

After looking into the categories, the next area to look at was the correlation between the features. In order to accomplish this Pearson's correlation test was run. The results were as follows:

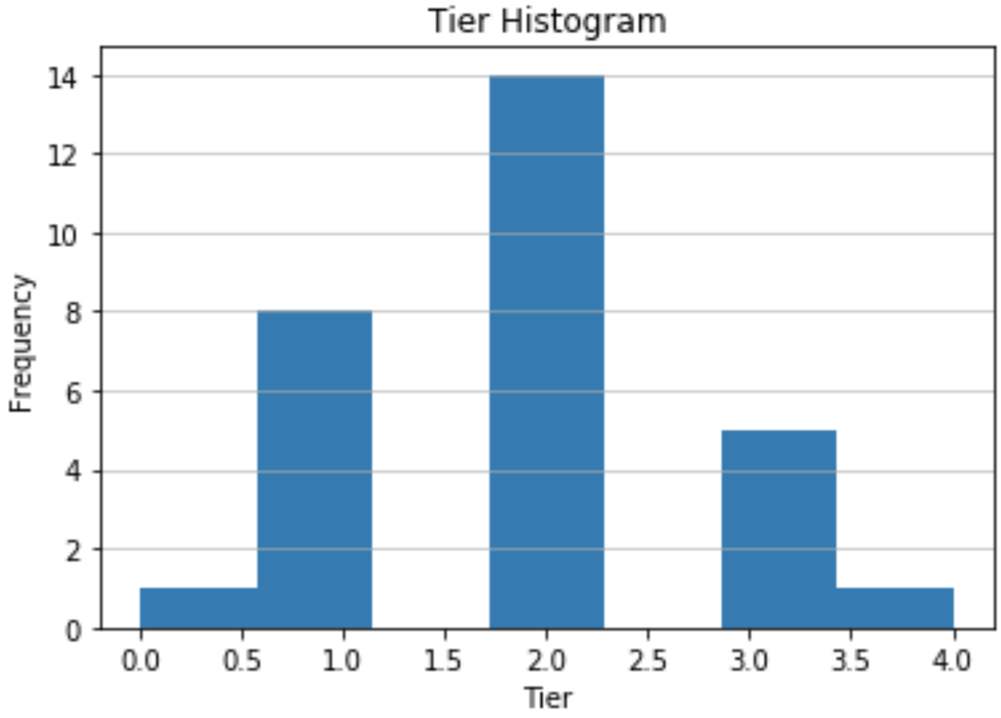


Out[445]:

	encoding_category	rating	likes	tier	tips
encoding_category	1.000000	-0.036728	-0.209776	-0.644931	-0.203038
rating	-0.036728	1.000000	0.378039	-0.047140	0.394685
likes	-0.209776	0.378039	1.000000	0.165801	0.993946
tier	-0.644931	-0.047140	0.165801	1.000000	0.160042
tips	-0.203038	0.394685	0.993946	0.160042	1.000000

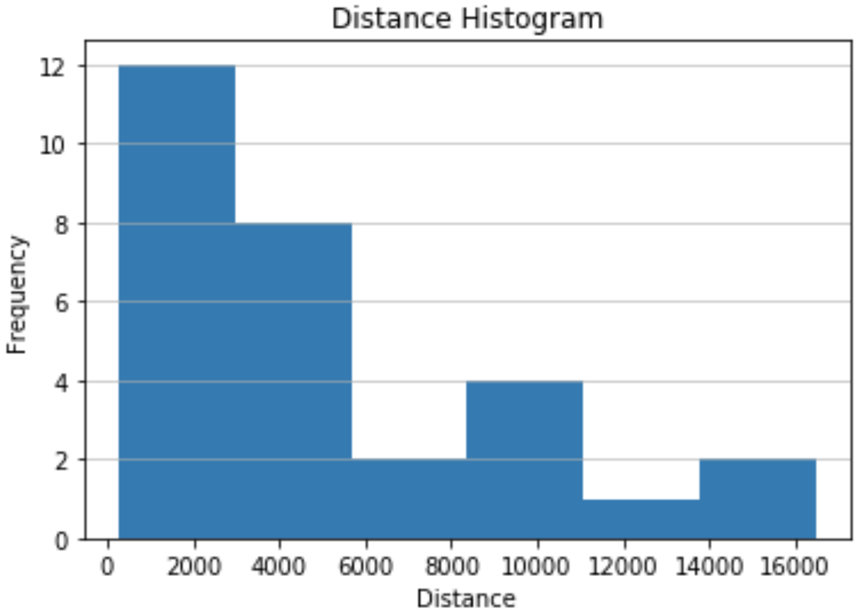
We notice right off of the bat a high correlation between tips and likes. This might imply that the increased number of tips for a venue might influence the number of likes that people leave regarding the venue.

Next is to create some histograms to get some insights into the distributions. As expected, the distributions do not look complete as there is a low number of n within the population. Thus, the histograms do not resemble a normal distribution since n is low.



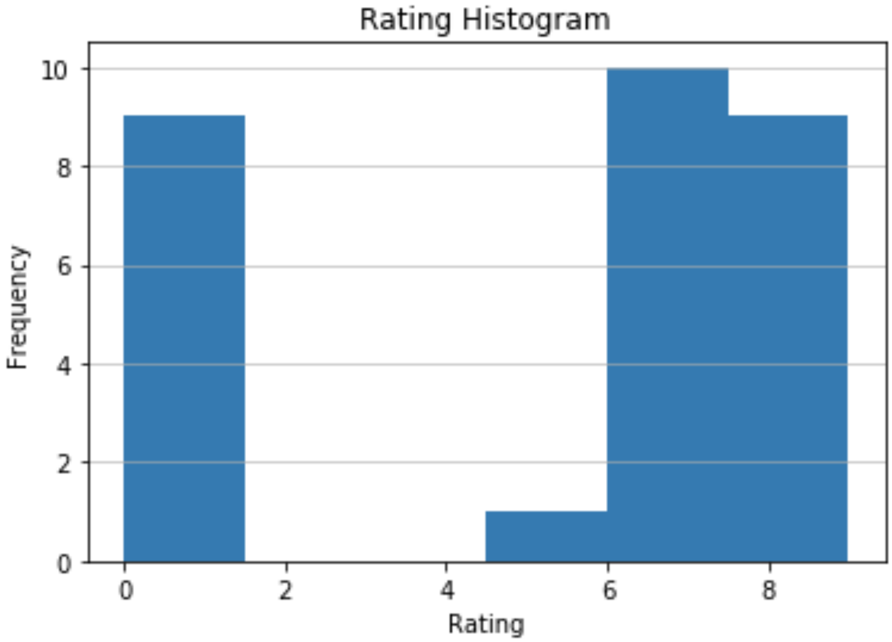
```
The average tier:  1.896551724137931
```

We can see from the histogram that there is a high frequency binned around 1.75-2.25 suggesting that there is an increased number of low-mid restaurants in the surrounding area. Furthermore, we can see that there is a low frequency at the edge around 3.25-4.0 with only 2/27 restaurants being in a tier above 3.25. This is good news for the higher end restaurant owner as the market is not saturated with a lot of upper end venues.



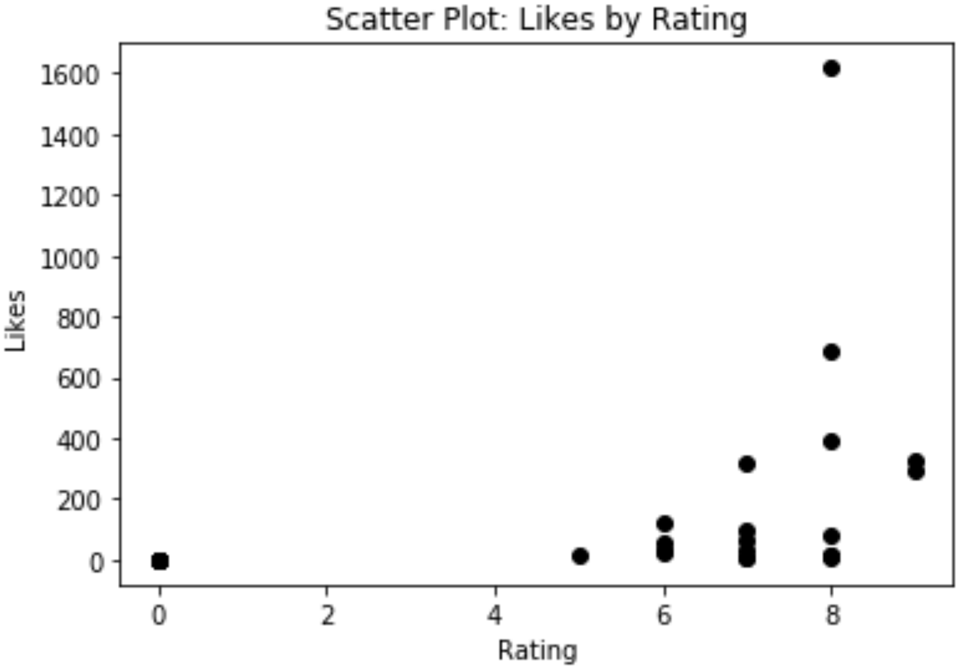
```
The average distance from center:  5021.620689655172
```

The major point to note about the distance histogram is the skew to the graph leaning towards the lower end of the histogram. Out of the 29 venues 20 of the venues are within a distance of 5,750 meters or 3.5 miles of the center. This is identifying that the market is fairly dense when looking at a close surrounding proximity.

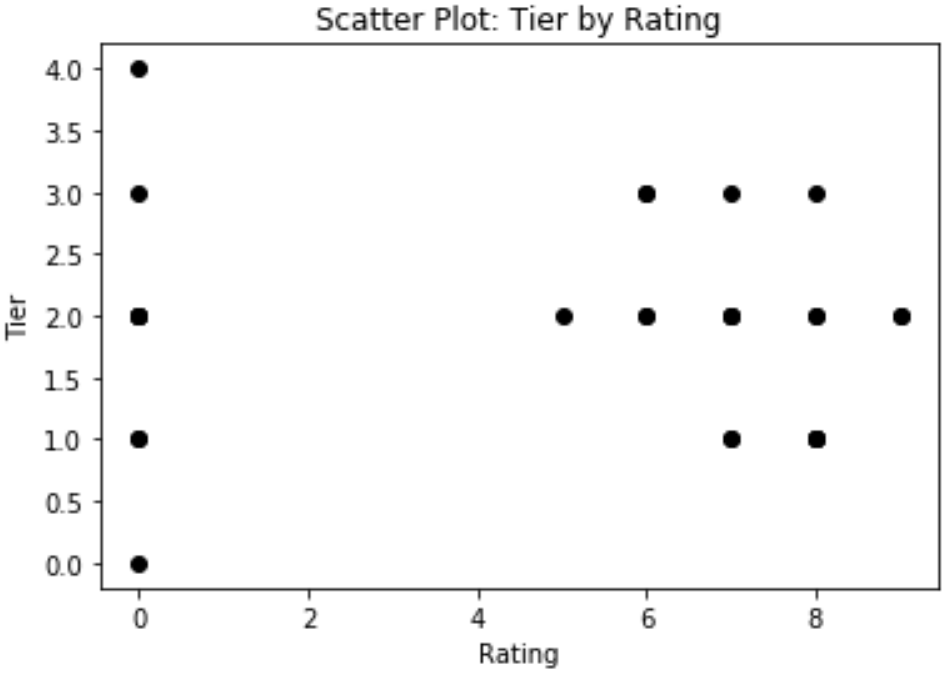


```
The average rating:  5.0
```

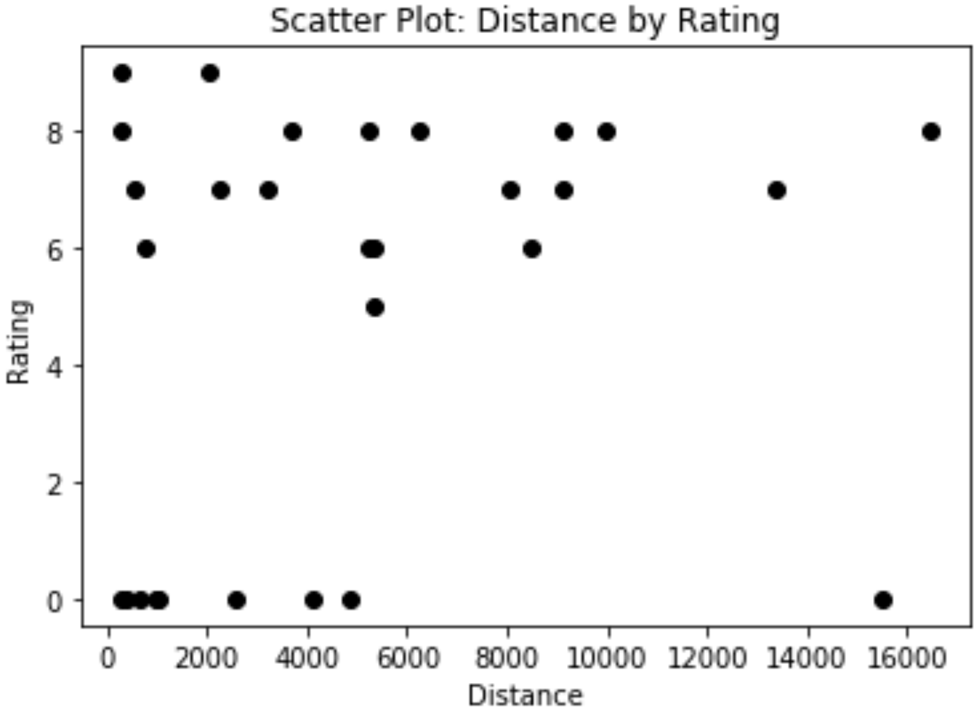
The histogram of the rating is fairly distributed in terms of the rating either being very high or very low. This can indicate that the reviewers are reviewing due to having either a strong positive opinion on the restaurant or the direct opposite, in which they have a negative one.



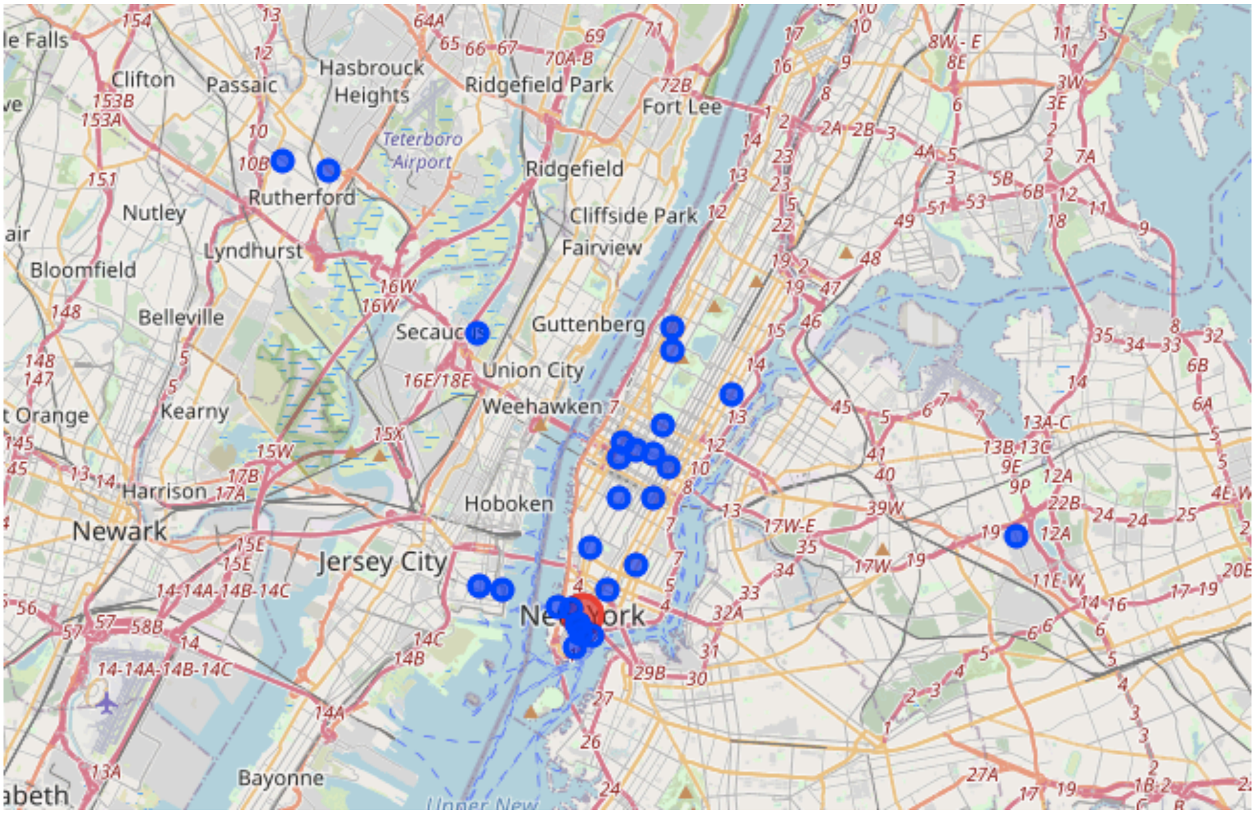
It can be seen from the scatterplot that the higher the rating of the venue the increased probability of there being a venue with a significant number of likes.



By looking at the scatter plot it can be seen that the ratings of tiers do not have a heavy influence on the rating. This is also affirmed by the correlation test between the features.



We can see that the distribution of venues within a close proximity have lower ratings between 6-8, suggesting that there is room for another venue.



For a general idea of the 29 different locations surrounding New York, NY. We can see that the majority of the locations are directly south within a close proximity or directly north a further distance away from the epicenter.

4. Results

   The results suggest that the market for higher end venues is relatively open for a new entry. This is due to the distance of upper tier restaurants are from the epicenter. Overall the ratings for the majority of the venue are on average a 5, leaving room for improvement. This leaves an opportunity for another venue to capture the market due to the venues not having high ratings, implying that the consumers are still looking for a place to eat.

5. Discussion

   5.1. Future

   In order for this to be more complete, the dataset would need to be supplemented with additional information. One example of an additional dataset that would provide further insights for the investor is evaluating the real estate pricing of the venues. Additionally, cross referencing some traffic data would provide insights into the areas with the most traffic.

6. Conclusion

   Overall, there appears to be an opportunity for a higher end restaurant to open up around New York, NY as the tiers of venues immediately around the area is on the lower end, making for a market opening. Furthermore, the ratings of the venues might imply that the consumers are not satisfied with the current options of venues in the surrounding area. Further, insights can be provided by valuing out the real estate tied with some projected capture of the market.