New York City Food Venues Analysis

Peiwen Liu

June 29, 2020

Introduction

As many might not expect, the restaurant industry in New York City has been suffering for a while. Despite the large population of New York City (6% of the total population in the US), there were only around 27,000 restaurants (4% of the total restaurants in the US) by 2018 (Statista, 2019).[1] This number had not increased much since 2013 due to the intense competition and slim profit margin. Worse enough, as the increase of minimum wage (from $10.50 in 2016 to $15 in 2019) which resulted from the victory of "Fight for $15" labor movement, the restaurants jobs in New York City decreased by 5,900 (3.4%) in 2018 (David, 2019). As the outbreak of the COVID-19 pandemic, survival of restaurants in New York City becomes undoubtably even tougher. Although consumer rating is not the sole factor that determines the survival of a restaurant, it does positively correlate with survival. For example, although the increase of minimum wage decreases the overall survival of restaurants at city level, 3.5-star (out of 5) restaurants are 14% more likely to exit the market as $1 increase in minimum wage while 5-star restaurants are largely unaffected (Luca and Luca, 2017).[2] In addition, high-quality restaurants with higher ratings benefit from the increased Yelp exposure (i.e., 7-19% more likely to survive) but low-quality restaurants with lower ratings are undermined by the increased Yelp (i.e., 7-19% more likely to exit the market) (Fang, 2019).[3]

Therefore, it is critical for restaurant owners and investors to understand customer rating which can further interact with other factors and jointly predict the survival of a particular restaurant. Using a uniquely constructed dataset, this project aims to understand the correlations between customer rating and several other factors such as the number of customer rating, deviation of customer rating. Furthermore, this project explores the aforementioned relationships not only at city level, but also at food category level (e.g., bar, bakery, café) and price level and reveals several interesting findings. For example, the correlation between customer rating and number of customer rating is stronger for hedonic food category than for utilitarian food category. Also, this correlation is stronger for higher price tiers than for lower price tiers.

Data Description

- The list of all community boards in New York City with relevant information is fetched from the main table of ***Neighborhoods in New York City*** Wikipedia page.[4] The main libraries I used to build the function for fetching the table are **urllib.request** and **BeautifulSoup**.

  The database has main components of ***Community Boards***, ***Area***, ***Population*** and ***Population per Area*** data in New York City. To combine it with the following geometric database, an extra column of ***boro_cd*** is added as the codes of community boards, which serves as the primary key for the two databases.

	Community Board	Area km^2	Pop.	Pop. /km^2	Neighborhood	boro_cd
0	Bronx CB 1	7.17	91,497	12,761	Melrose, Mott Haven, Port Morris	201
1	Bronx CB 2	5.54	52,246	9,792	Hunts Point, Longwood	202
2	Bronx CB 3	4.07	79,762	19,598	Claremont, Concourse Village, Crotona Park, Mo...	203
3	Bronx CB 4	5.28	146,441	27,735	Concourse, Highbridge	204
4	Bronx CB 5	3.55	128,200	36,145	Fordham, Morris Heights, Mount Hope, Universit...	205

- The geometric and boundary information of all the community boards is found in the source geojson file of ***The NYC Boundaries Map***, which is an open source tool for viewing and querying overlapping administrative boundaries in New York City.[5]

  The key information of the database is in the ***geometry*** column, where the polygon or multipolygon data sets are presented for individual ***boro_cd***. The geometric information is necessary for the function folium.GeoJson that is incorporated in the choropleth map.

	boro_cd	shape_area	shape_leng	geometry	Latitude	Longitude
0	311	103177785.365	51549.5578567	MULTIPOLYGON (((-73.97299 40.60881, -73.97259 ...	40.607216	-73.993636
1	313	88195686.2748	65821.875577	MULTIPOLYGON (((-73.98372 40.59582, -73.98305 ...	40.580649	-73.982140
2	312	99525500.0655	52245.8304843	MULTIPOLYGON (((-73.97140 40.64826, -73.97121 ...	40.631227	-73.983260
3	304	56663217.0191	37008.1004432	MULTIPOLYGON (((-73.89647 40.68234, -73.89653 ...	40.694549	-73.916836
4	206	42664311.28	35875.7111716	MULTIPOLYGON (((-73.87185 40.84376, -73.87192 ...	40.849604	-73.887530

- The search for food venues near each community board and the acquisition of venue details are performed with **Foursqure Places API**.[6] In the process of data collection and cleaning, several issues are worth stressing. First, it only allows user to get the venues near specific locations rather than administrative districts. To make sure certain venue obtained belongs to the right community board, it is important to filter out venues by checking their coordination with the boundaries of community boards during search.

  Second, since there is a limit of 50 for number of venues to return for each community board, the resulting dataset is incomplete in terms of counting all food venues in New York City. However, the dataset can be regarded as evenly random samples for these administrative districts and have significant value of statistics.

  Third, it is found in the later investigation that venues corresponding to some category such as Flower Shop and Clothing Store are included in the database. To avoid misleading information in categorical analysis, category list is manually verified for the purpose of cleaning.

  Last, there are venues which don’t deliver either rating scores or price tiers message (None value). During the analysis of rating- or price-based information, database is wrangled accordingly to be ready for use.

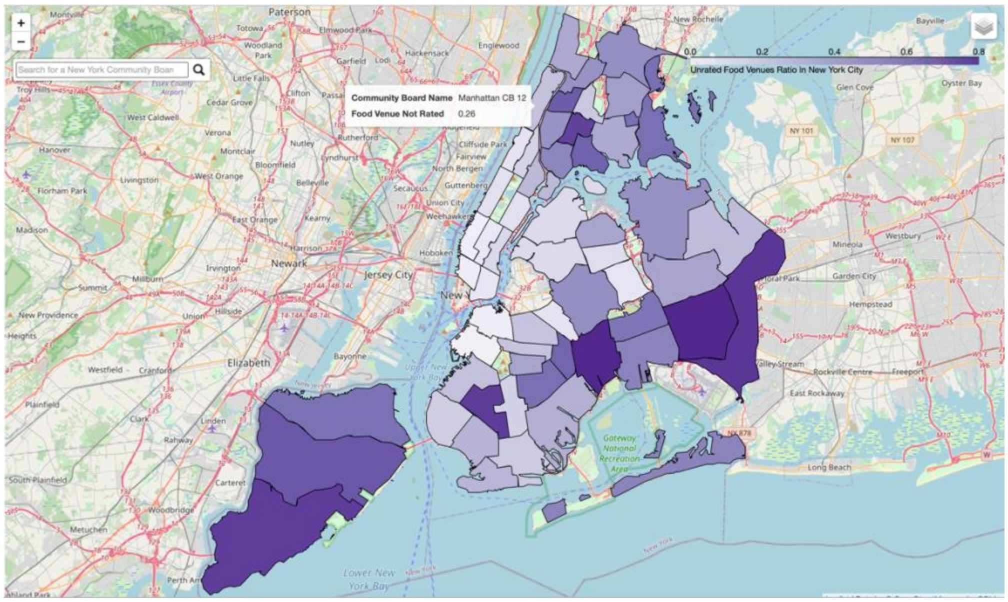
The final database which contains the information of **venue id, name, latitude, longitude and category**, **rating, rating signal, price tier, tips, likes and listed count** is shown below:

	boro_cd	Community Board	Community Board Latitude	Community Board Longitude	Venue Id	Venue Name	Venue Latitude	Venue Longitude	Venue Category	Rating	Rating Signal	Price	Tips Number	Likes	Listed Number
0	201	Bronx CB 1	40.810909	-73.91647	4dae69211e7207bbeb106d82	Cypress Corner Deli	40.807903	-73.913446	Café	None	None	None	0	0	0
1	201	Bronx CB 1	40.810909	-73.91647	4c0c07676071a593902be232	Primos	40.808725	-73.915402	Deli / Bodega	None	None	None	1	1	0
2	201	Bronx CB 1	40.810909	-73.91647	4c0952f0bbc676b0b30148d5	La Morada	40.810670	-73.921758	Mexican Restaurant	8.9	84	None	29	57	223
3	201	Bronx CB 1	40.810909	-73.91647	4cf2eb9c1c158cfa95b9dab5	Fresco Pizza & Pasta	40.814358	-73.913183	Pizza Place	7	20	None	9	10	5
4	201	Bronx CB 1	40.810909	-73.91647	4bf1b1ef3fa220a150991820	Mexicocina	40.811837	-73.909512	Mexican Restaurant	9	84	2	24	55	71
5	201	Bronx CB 1	40.810909	-73.91647	4e4e242dbd4101d0d7a333fc	KFC	40.816524	-73.918461	Fried Chicken Joint	6.2	8	1	2	4	2
6	201	Bronx CB 1	40.810909	-73.91647	4d2e531979dd6ea8795588d3	Checkers	40.815628	-73.917990	Fast Food Restaurant	4.8	12	None	11	0	2
7	201	Bronx CB 1	40.810909	-73.91647	5db1caab98c98d0008f537f9	Starbucks	40.816376	-73.918473	Coffee Shop	7.1	0	None	0	0	0
8	201	Bronx CB 1	40.810909	-73.91647	56eaefc4498e0c20d8bda9f1	Capri Cakes	40.816880	-73.920876	Bakery	None	None	None	1	3	0
9	201	Bronx CB 1	40.810909	-73.91647	5b54e70182a750002c111916	Taco Bell	40.818109	-73.923674	Taco Place	6	0	None	0	0	0

Data Analysis and Discussion

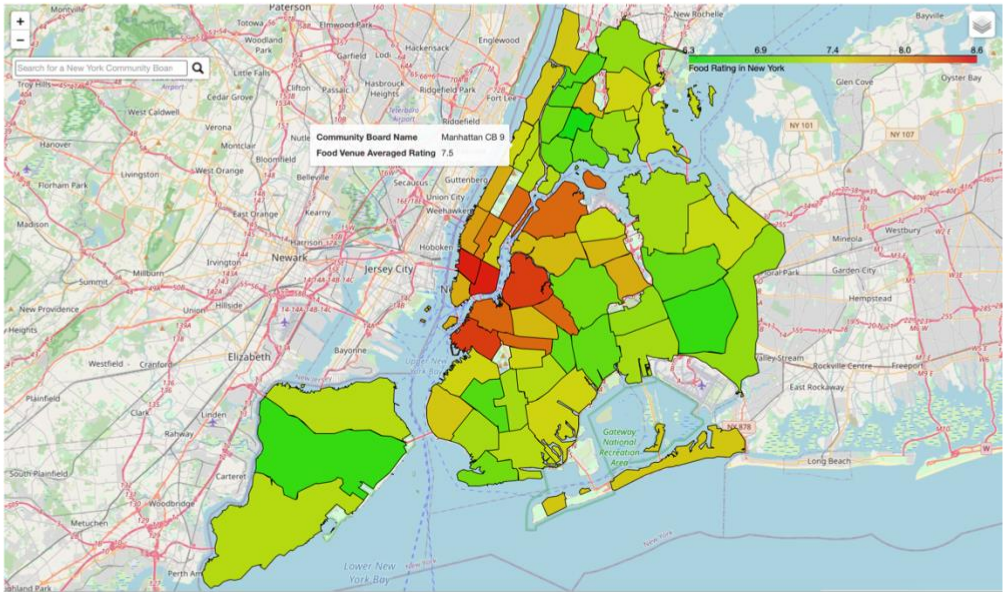
1. Choropleth maps

- Choropleth map for not rated food venues: The database for this choropleth map stems from counting the unrated venues for each community board. The result is shown as follow:



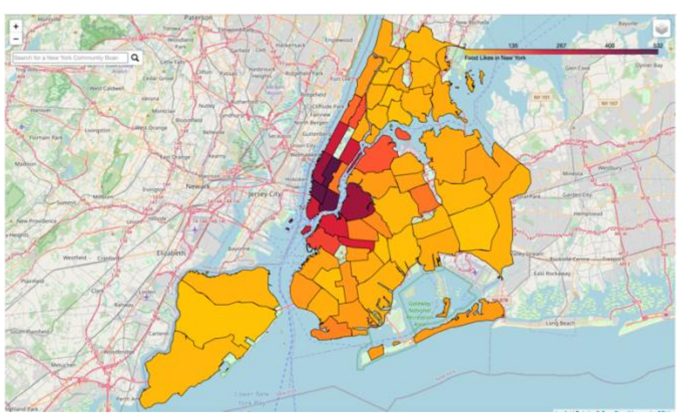
The community board with the high unrated food venue ratio is **Brooklyn CB 5 and Queens CB 12 (ratio = 0.85)**, followed by **Queens CB 13 (ratio = 0.8)**. Additionally, it is shown that there are more unrated food venues overall in the community boards which locate away from core area of New York City. Those food venues could be new in the area or belong to certain category that people are of less interest in rating for.

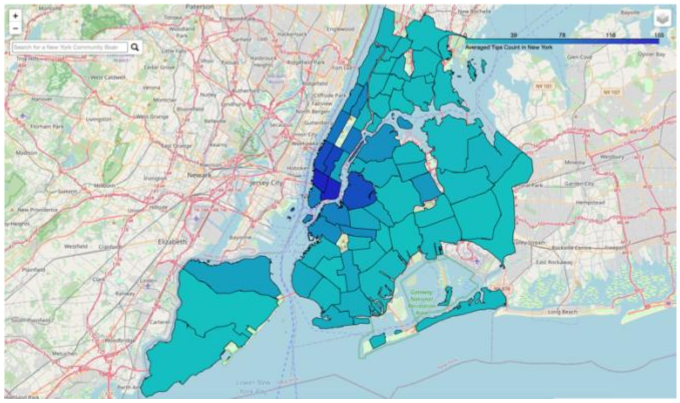
- Choropleth map for food rating: The database for this choropleth map is obtained by averaging the valid rating scores of all the food venues for each community board. The result is shown as follow:

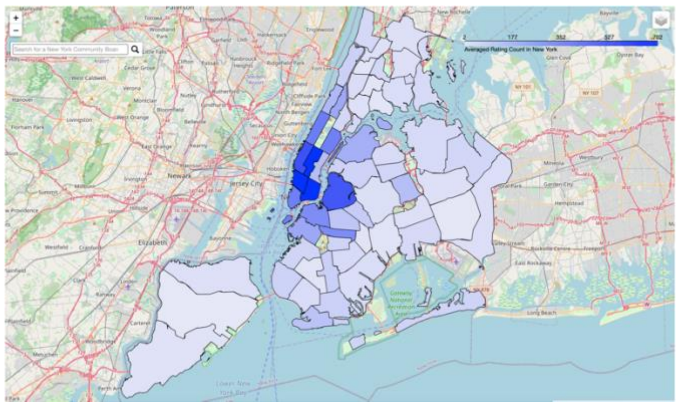


The community boards with highest average rating for food venues are **Manhattan CB 2 and 3 (average rating score: 8.6)**, closely followed by **Brooklyn CB 1 and 6 (average rating score: 8.4)**. Most community boards with high unrated food venues ratio to have low food ratings in general. On the other hand, community boards with fewer unrated food venues essentially show higher food rating.

- Choropleth maps for most ratings received, likes and listed count: the database for this choropleth map comes from summing up the rating signal, likes and listed count of the rated venues for each community board. The result is shown as follow:

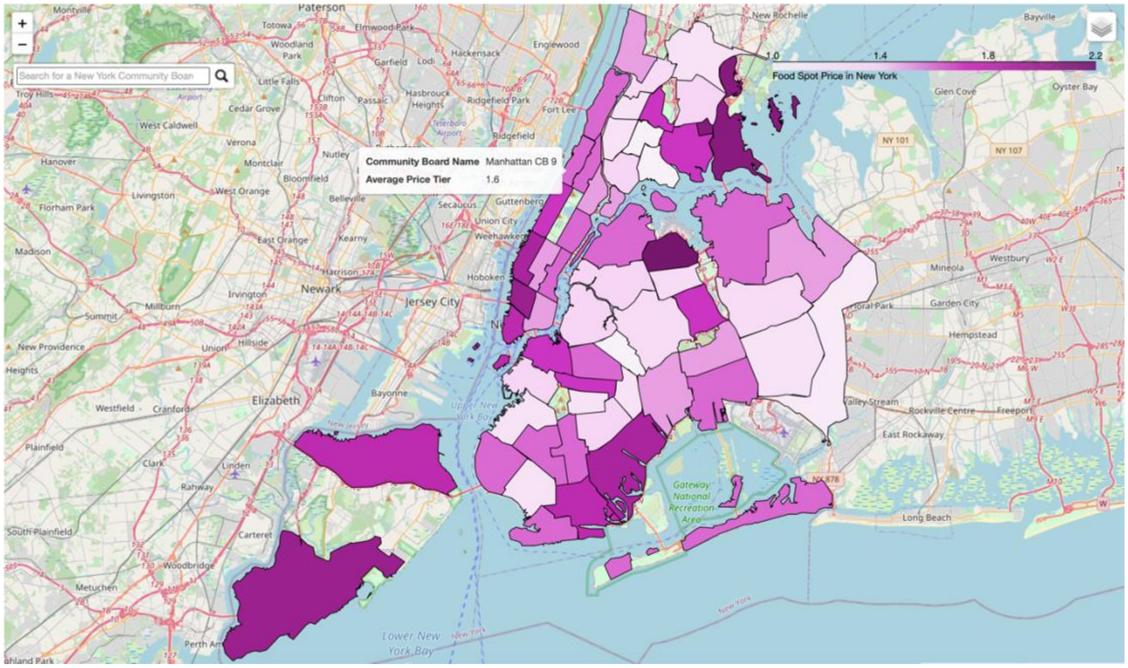






The community boards with most ratings received are **Manhattan CB 2 (average rating signals: 702)**, followed by **Manhattan CB 3 (average rating signals: 691) and 5 (average rating signals: 688)**. The rating signals of core areas in New York City significantly exceeds other regions. **Brooklyn CB 1, 2, 6** and **8** are exactly the community boards in Brooklyn with the highest food rating. The investigation on the correlation between rating and rating signals is thus very important and interesting, which is discussed in the following sections. Furthermore, choropleth maps for most ratings receive, likes and listed count look quite similar to each other. Therefore, correlation among these three features is meaningful to study as well.

- Choropleth map for food venue price: the database for this choropleth map comes from averaging the price tier of the food venues for each community board. The result is shown as follow:

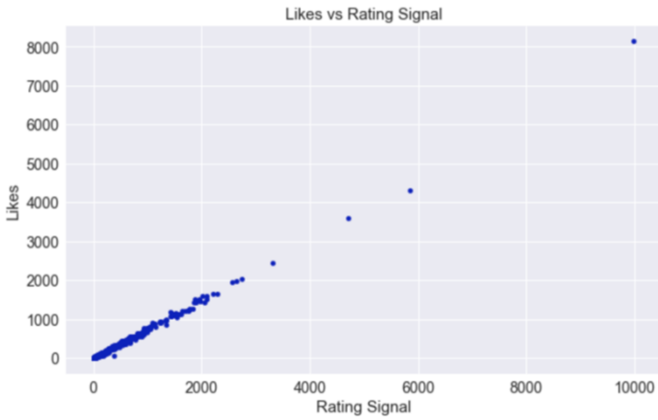


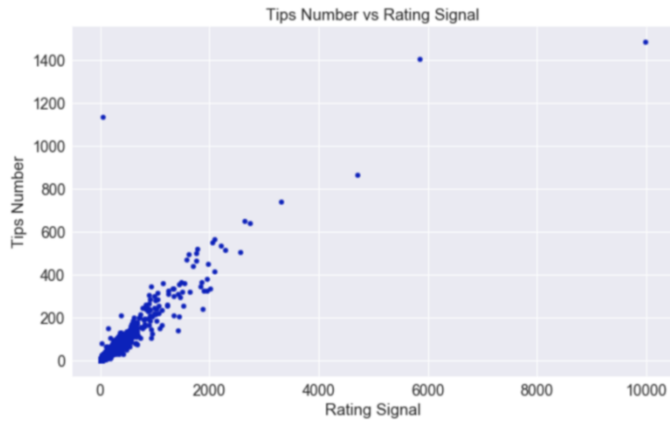
The community boards with most expensive food venues in average are **Queens CB 3 (average price tier: 2.2)**, followed by **Bronx CB 10 (average price tier: 2.1)**.

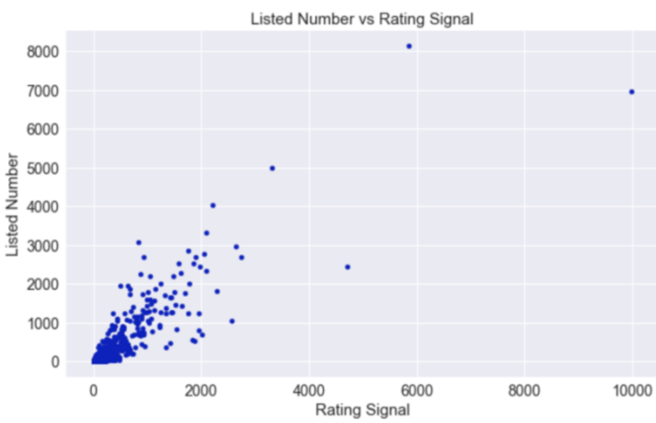
The choropleth maps above present a variety of information across community boards in New York City. Before we visualize the similar features across food categories in the next, it is a good idea to have a look at the relationship among rating signal, likes and listed count first so that we may simplify our analysis by reducing highly correlation features.

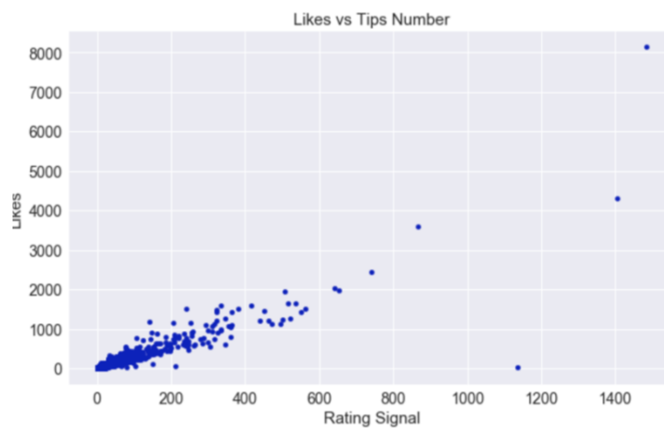
2. Correlation among details of venues

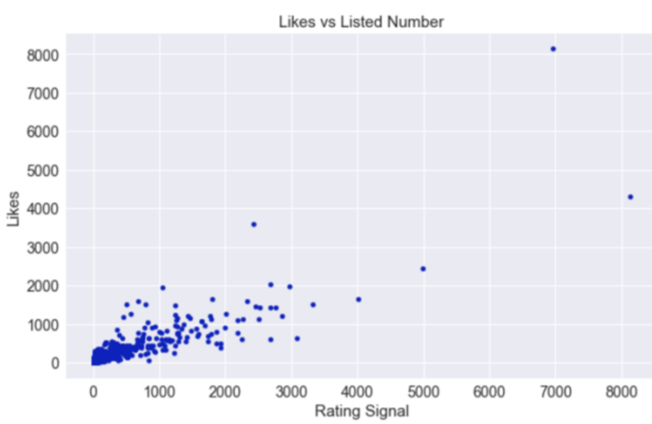
- The following six figures are scatter plot for different sets of features to reveal the relationship amongst them:

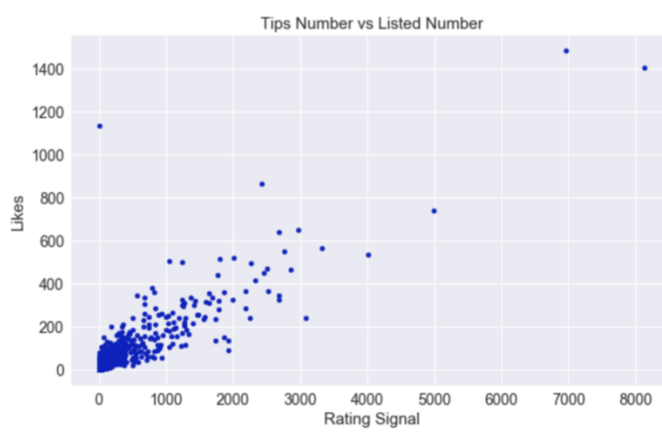












- We generate two tables based on the correlation matrix for rating signal, tips number, likes and listed number to show their correlations and corresponding p-values.

Correlation Table:

	Rating Signal	Tips Number	Likes	Listed Number
Rating Signal	1.000	0.921	0.997	0.884
Tips Number	0.921	1.000	0.905	0.882
Likes	0.997	0.905	1.000	0.874
Listed Number	0.884	0.882	0.874	1.000

Corresponding p values (<):

	Rating Signal	Tips Number	Likes	Listed Number
Rating Signal	1.0	0.0	0.0	0.0
Tips Number	0.0	1.0	0.0	0.0
Likes	0.0	0.0	1.0	0.0
Listed Number	0.0	0.0	0.0	1.0

The two features with the highest correlation are **Rating Signal and Likes (correlation = 0.997, P < 0.001)**. The two features with the lowest correlation are **Tips Number and Listed Number (correlation = 0.882, P < 0.001)**. These four features are highly correlated with each other as the values of correlation among them are all greater than 0.88. In the following analysis, we thus mainly concentrate on **Rating Signal** in these features.

3. Analysis rating signal and rating score across prices

The bar plot shown below averages rating signal, likes and tips count in terms of each price tier. The data comes from mean aggregation with respect to price tier for all venues.

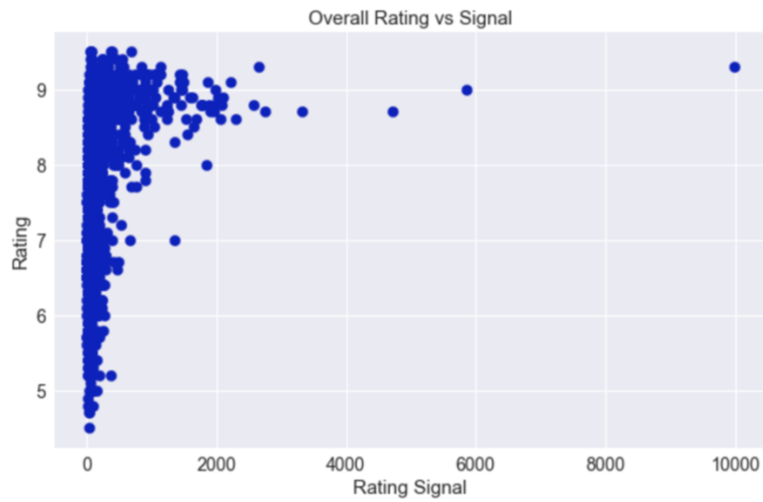
Venues with the most average rating signal received is of price tier 2, followed by price tier 3, then price tier 4 and last price tier 1.

Next, to the study of the relationship between rating and price tier of food venues. All rating information associated with their price tiers are extracted and the database with price tier 1-4 as feature columns is generated for our analysis.

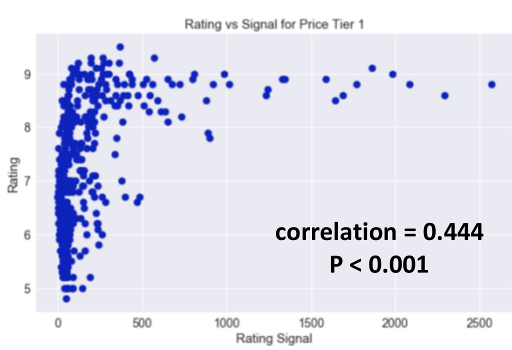
The correlation between rating and price is **correlation = 0.307 (P < 0.001)**. The result indicates that a food venue with a higher price tier is likely to receive a higher rating in general. In addition, when the average rating score of the food venues across price tiers becomes higher, it looks less deviated and differentiated.

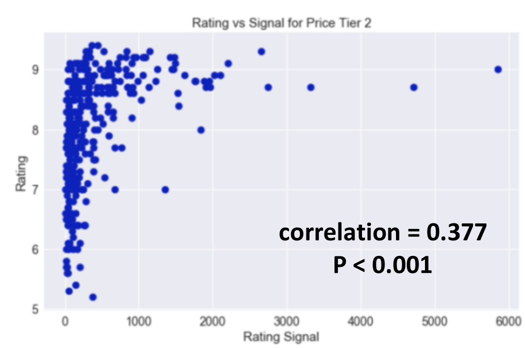
4. Analysis of rating and rating signal

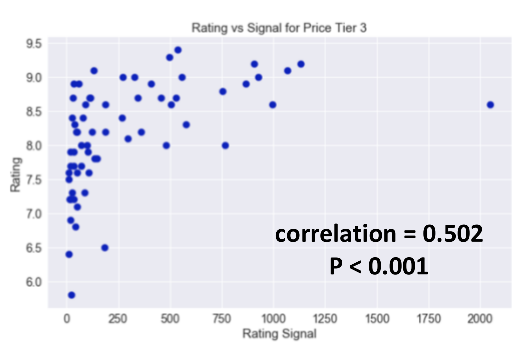
It is shown previously in choropleth maps that rating and rating signal across community boards have similar distribution. Now, we analyze all the datapoints of these two properties from our database.

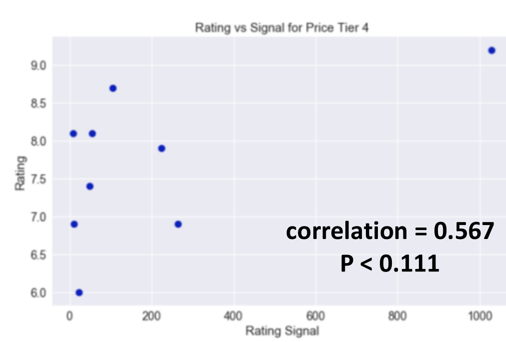


The correlation between overall rating scores and number of rating signal received is **correlation = 0.307 (P < 0.001)**. The result perfectly describes the phenomenon that food venues with more rating received (more popular in common sense) tend to have higher rating scores. The same studies across different price tiers are performed as well. The results are:

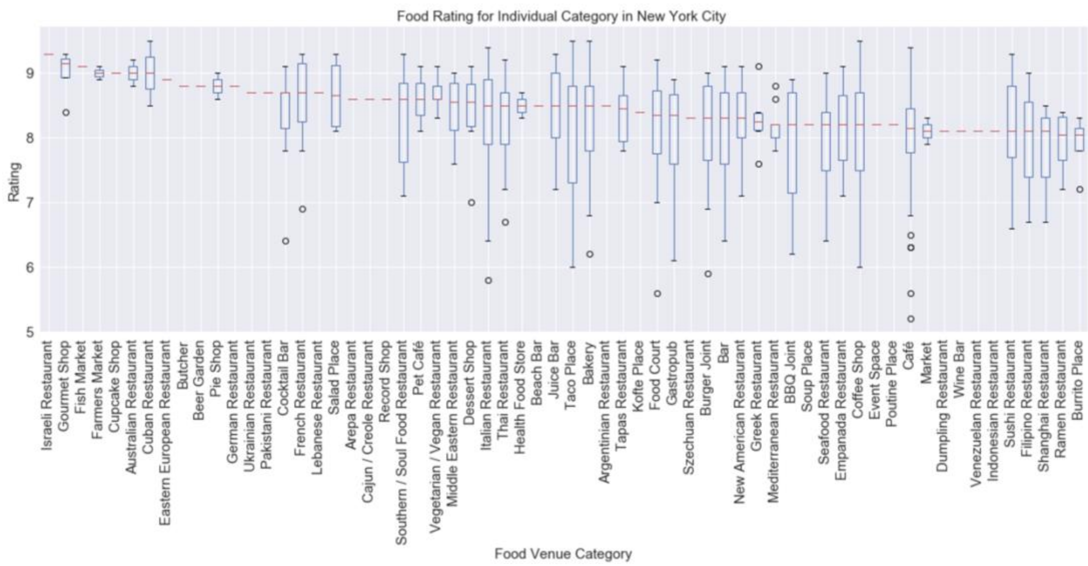


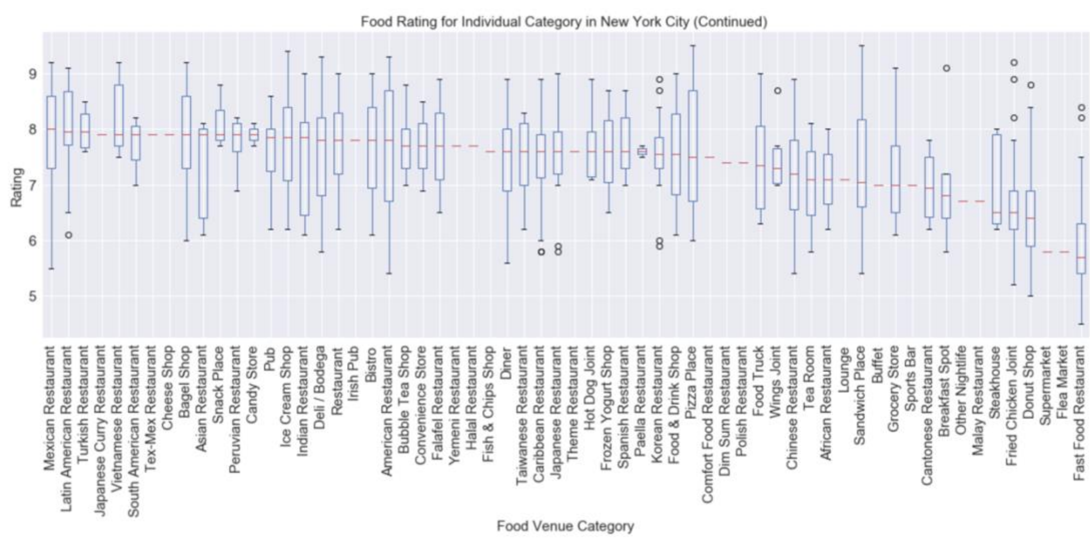






As we already see the significant correlation between rating score and rating signal, it is also very interesting to figure out the correlation across food venue categories. To achieve a throughout investigation, we first analyze the statistics of rating scores across categories. The box plot is generated after value of rating scores of all the venues are extract and put into data frame with columns consisting of categories.





Next, we create the similar database for analyzing rating signal across food venue category. Since the range and difference of rating signal is very large. It is not a good way to demonstrate it in box plot, though the data is used in the following correlation test.

The result of correlation between rating scores and signal across food venue categories is shown below:

The two highest correlation exists in **Fried Chicken Joint (correlation = 0.636, P < 0.001)** and **Ice Cream Shop (correlation = 0. 611, P < 0.001)**, which are both very hedonic categories. The effect of popularity on rating (users' experience) is thus stronger. On the other hand, the two lowest correlation exists in **Coffee Shop (correlation = 0.237, P < 0.002)** and **Bakery (correlation = 0.230, P < 0.04)**. Bakery are quite a utilitarian category, which is opposite to hedonic categories. The effect of popularity on rating (users' experience) is thus weaker.

Finally, we test the correlation between mean and standard deviation of rating scores and signal. Each set of mean and standard deviation is calculated on individual category.

Correlation Table

	rating mean	rating std	signal mean	signal std
rating mean	1.000	-0.246	0.336	0.184
rating std	-0.246	1.000	-0.039	0.231
signal mean	0.336	-0.039	1.000	0.652
signal std	0.184	0.231	0.652	1.000

P < Table

	rating mean	rating std	signal mean	signal std
rating mean	1.000	0.006	0.000	0.041
rating std	0.006	1.000	0.671	0.010
signal mean	0.000	0.671	1.000	0.000
signal std	0.041	0.010	0.000	1.000

The correlation between mean and standard deviation of rating scores is **correlation = -0.246 (P < 0.006)**. This result confirms our observation that food venues across categories or price tiers with higher average rating show less difference and polarization in rating. It also agrees well with the rating analysis across price tiers shown in the previous section. In another aspect, the correlation between mean and standard deviation of rating signal is **correlation = 0.652 (P < 0.001)**. To a great extent, such a high correlation is due to numbers of extreme large rating signal for some food venues along with large variances in their categories.

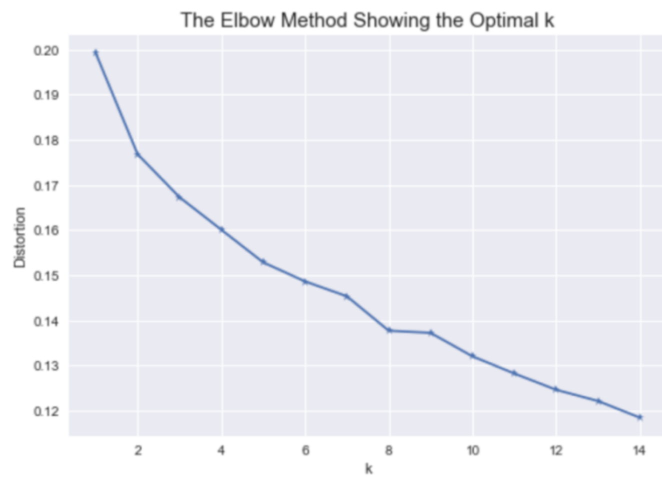
The **correlation = 0.336 (P < 0.001)** between means of rating and rating signal received is consistent with our analysis on overall rating and rating signal previously (correlation = 0.307, P < 0.001) and further confirm their positive correlation. Interestingly, though the mean of rating score is negatively correlated with its standard deviation, it is found to be positively correlated with the standard deviation of rating signal (**correlation = 0.184, P < 0.04**). However, the other way around, there is no correlation between the mean of rating signal and the standard deviation of rating.

5. Clustering community boards based on frequency of food categories

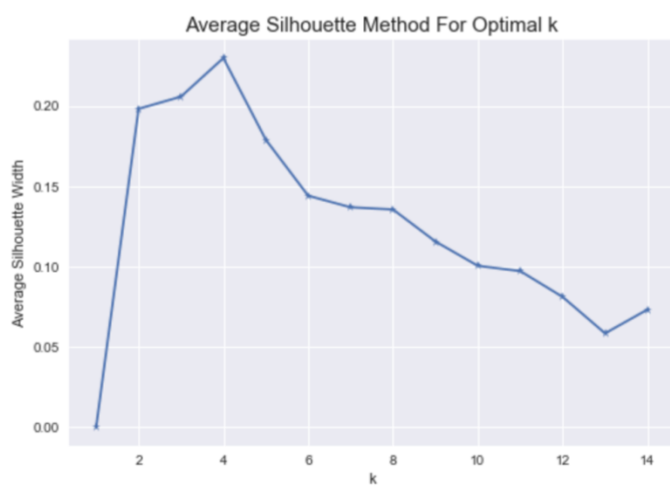
Before performing clustering analysis, a table showing top 10 venue category for each community boards are created based on one-hot encoding of venues in terms of category.

	Community Board	1st Most Common Venue	2nd Most Common Venue	3rd 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	Bronx CB 01	Fast Food Restaurant	Mexican Restaurant	Donut Shop	Deli / Bodega	Café	Pizza Place	Bakery	Coffee Shop	Fried Chicken Joint	Juice Bar
1	Bronx CB 02	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Food	American Restaurant	Juice Bar	Restaurant	Café	Cafeteria	Deli / Bodega
2	Bronx CB 03	Deli / Bodega	Seafood Restaurant	Fast Food Restaurant	Pizza Place	Bakery	Food	Chinese Restaurant	Mexican Restaurant	Food Truck	Donut Shop
3	Bronx CB 04	Deli / Bodega	Fast Food Restaurant	Chinese Restaurant	Donut Shop	Sandwich Place	Hot Dog Joint	Seafood Restaurant	Steakhouse	Mexican Restaurant	Pizza Place
4	Bronx CB 05	Mexican Restaurant	Fast Food Restaurant	Pizza Place	Bakery	Deli / Bodega	Fried Chicken Joint	Donut Shop	Restaurant	Chinese Restaurant	African Restaurant

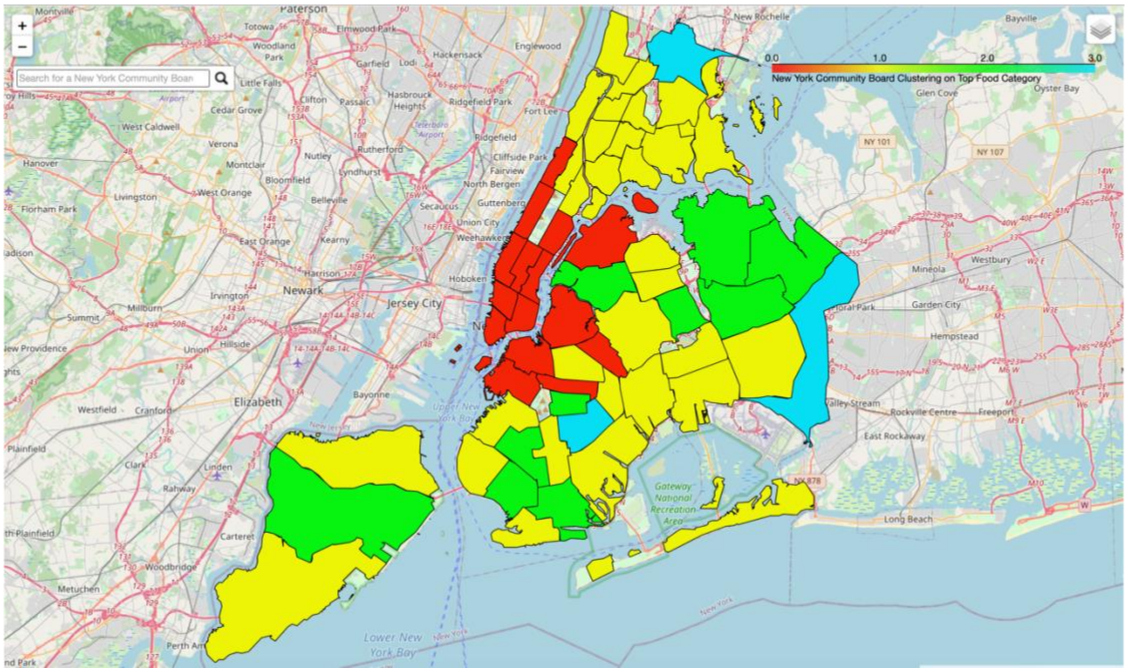
Unsupervised learning K-means algorithm is used to cluster the community boards. For choosing the optimal k number, elbow method is applied first, and the result is:



Apparently, the improvements of the distortion decline relatively steadily and evenly so that no elbow point is found for the optimized k-clustering number. Hence, we apply a different method of average silhouette, with result shown below. In this case, we are able to determine an optimal **k = 4**



After choosing the k-clustering number, we conduct clustering, label each community boards and update the information in our database. The clustering map is generated afterwards based on the result:



For community boards in the 1st cluster, the 1st most common venues are all Coffee Shop. The 2nd and 3rd ones are mainly Pizza Place, Deli / Bodega, Italian Restaurant and Bakery. For community boards in the 2nd cluster, the 1st most common venue is mostly Deli / Bodega, followed by mainly Pizza Place as the 2nd most common venue. For community boards in the 3rd cluster, the 1st, 2nd and 3rd most common venues are quite diverse. For community boards in the 4th cluster, the 1st most common venues are all Caribbean Restaurant and the 1st most common venues are all Bakery.

Conclusion

In the study, we analyze the data of food venues in New York City. We first visualize different properties of food venues for each community boards by generating corresponding choropleth maps. We then found strong positive correlation among details of venues (rating signal, likes, tip and listing count). More importantly, we analyze the correlation between rating and price. The result shows that food venues with higher price tiers tend to receive higher rating in general. Furthermore, we investigate the relationship between rating and signal across categories. We found that food venues with more rating received is likely to have higher rating scores. Then, we analyze rating/signal correlation in terms of food venue category. We observe that hedonic venues show much stronger rating/signal correlation than utilitarian ones. Meanwhile, it is shown that there are less difference and polarization for higher rated food venues. Finally, we perform k-mean clustering on community boards in New York City based on their most common venue categories. Our results are both informative and conclusive for entrepreneurs, investors, and city managers to enhance the survival of a particular restaurant and to improve the survival rate of restaurants in certain community.

Reference

1. David, Greg. **NYC restaurant jobs decline for the first time in a decade.** Crain's New York Business, March 2019

2. Luca, Dara Lee, and Michael Luca. **Survival of the Fittest: The Impact of the Minimum Wage on Firm Exit.** Harvard Business School Working Paper, No. 17-088, April 2017.

3. Fang, Limin. **The Effects of Online Review Platforms on Restaurant Revenue, Survival Rate, Consumer Learning and Welfare.** UCDavis Economics Event Paper, January 2019

4. **Neighborhoods in New York City**

5. **The NYC Boundaries Map**

6. **Forsquare API**