Where Best to Start Your Cafe in the City of Chicago

Christopher Weaver

August 11, 2021

1. Introduction

1.1 Background

The majority of startups will fail within the first couple years of opening up. Considering the amount of money, effort, and time involved in starting a business; entrepreneurs need every advantage they can get. Obviously the success of a business will largely be contingent upon how good your coffee is, price, atmosphere, and lots of the other tangibles. But another important factor for success may be the location of your business. With everything else being equal, the location for your cafe can be another choice that may help lead to the success or failure of business.

1.2 Problem

A entrepreneur is looking to open a new cafe in the city of Chicago. What information can we find that may help us decide a neighborhood to open up in within the city of Chicago.We will review different data points for each of Chicago's 77 neighborhoods and see if we can find any corollaries between aspects of these neighborhoods and positive metrics from Foursquare to help predict success.

2. Data acquisition and cleaning

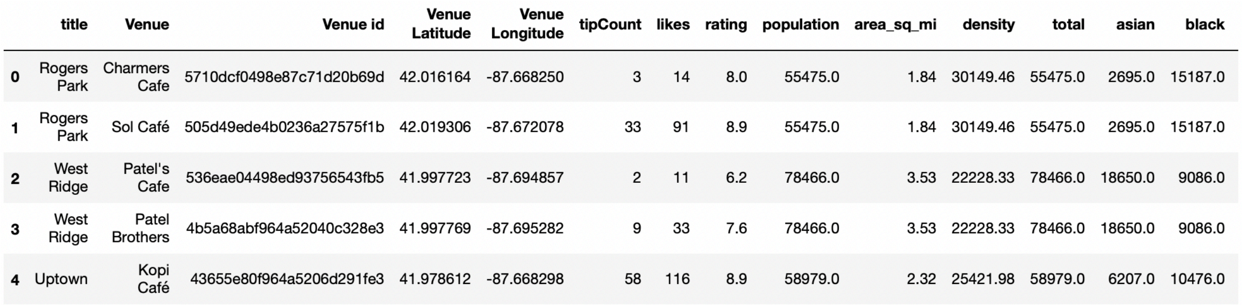
2.1 Data sources

Getting basic information about the 77 Chicago neighborhoods will come from Wikipedia. I will build a web crawler using BeautifulSoup that can grab neighborhood name, latitude, longitude, size, and density from https://en.wikipedia.org/wiki/Community_areas_in_Chicago. Further demographic information for each neighborhood including total population and racial distributions was found here https://www.cmap.illinois.gov/data/community-snapshots and imported as a csv. Finally, metrics to help us determine if a cafe is successful or not will be found from

Foursquare. Two Foursquare api's specifically will be utilized. The first is the search api to get all cafe's in each of the 77 neighborhoods. The second is the venue detail api which we will use to get information like the number of "likes" and overall rating for each cafe.

2.2 Data cleaning and feature selection

Accumulating all data for each neighborhood involved combining data frames for both the wikipedia data and the csv data from cmap. Lots of the numerical data from cmap ended up in the data frame as text with commas to help represent each data point. This required removing the commas from all those data points and converting to a float. The real challenge came from gathering success metrics from the foursquare api's. The first call quite easily could be entered into a data frame with columns for neighborhood, venue, venue id, venue latitude, and venue longitude. The second call returned "like" counts, overall rating, and how many "tips" the venue has, all of which seemed pertinent measures of success. This data is captured and inserted into the same data frame based off of the venue id.

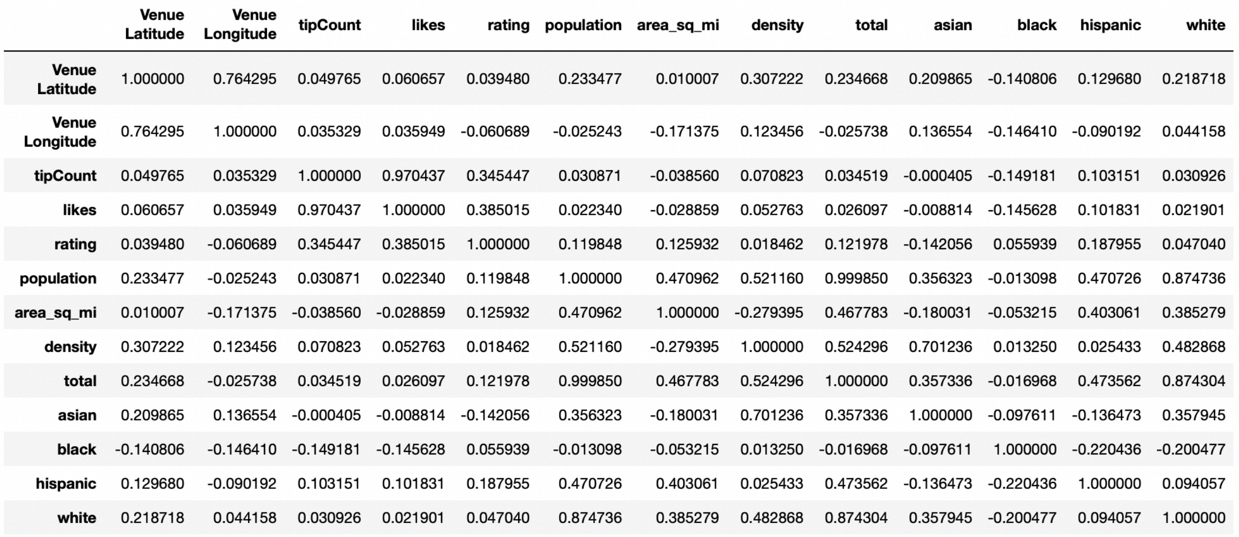


	title	Venue	Venue id	Venue Latitude	Venue Longitude	tipCount	likes	rating	population	area_sq_mi	density	total	asian	black
0	Rogers Park	Charmers Cafe	5710dcf0498e87c71d20b69d	42.016164	-87.668250	3	14	8.0	55475.0	1.84	30149.46	55475.0	2695.0	15187.0
1	Rogers Park	Sol Café	505d49ede4b0236a27575f1b	42.019306	-87.672078	33	91	8.9	55475.0	1.84	30149.46	55475.0	2695.0	15187.0
2	West Ridge	Patel's Cafe	536eae04498ed93756543fb5	41.997723	-87.694857	2	11	6.2	78466.0	3.53	22228.33	78466.0	18650.0	9086.0
3	West Ridge	Patel Brothers	4b5a68abf964a52040c328e3	41.997769	-87.695282	9	33	7.6	78466.0	3.53	22228.33	78466.0	18650.0	9086.0
4	Uptown	Kopi Café	43655e80f964a5206d291fe3	41.978612	-87.668298	58	116	8.9	58979.0	2.32	25421.98	58979.0	6207.0	10476.0

Exploratory data analysis

3.1 Correlations

Combining all neighborhood data into each cafe venue record allows us to quickly look at which fields might be correlated to other fields.

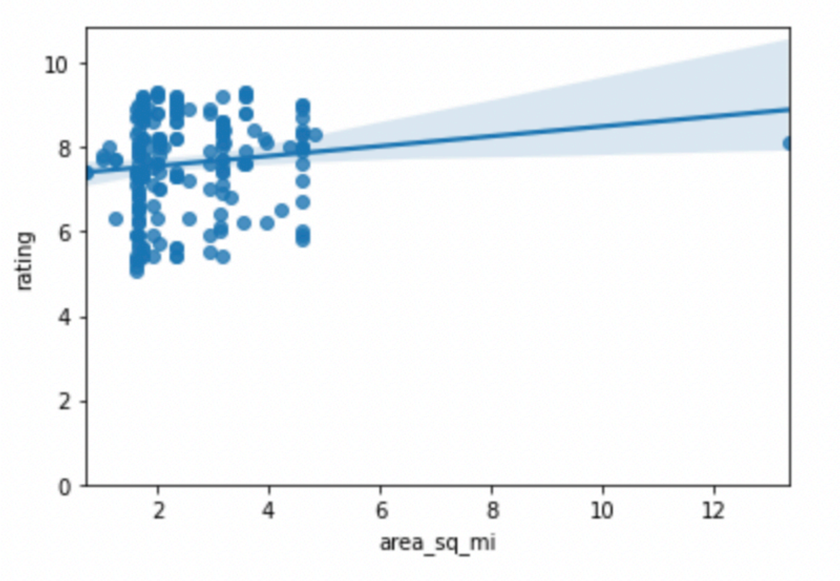


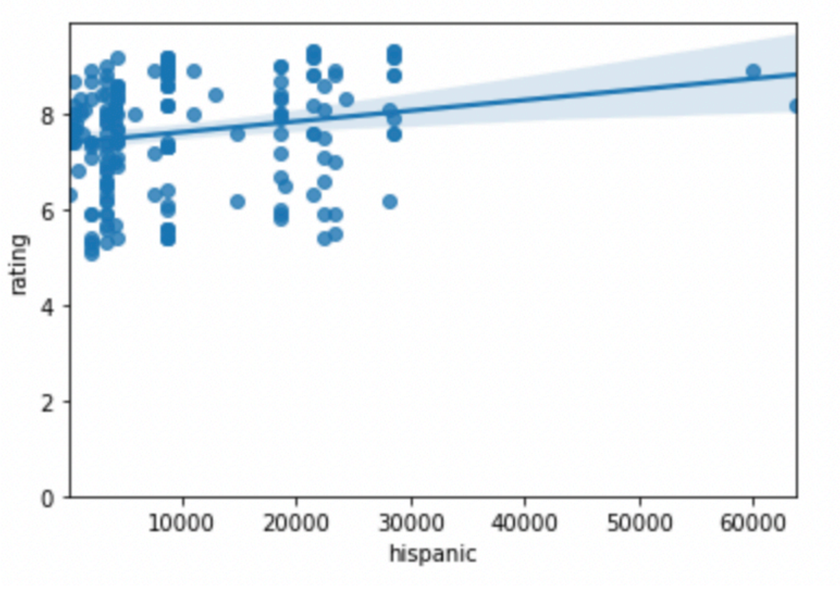
	Venue Latitude	Venue Longitude	tipCount	likes	rating	population	area_sq_mi	density	total	asian	black	hispanic	white
Venue Latitude	1.000000	0.764295	0.049765	0.060657	0.039480	0.233477	0.010007	0.307222	0.234668	0.209865	-0.140806	0.129680	0.218718
Venue Longitude	0.764295	1.000000	0.035329	0.035949	-0.060689	-0.025243	-0.171375	0.123456	-0.025738	0.136554	-0.146410	-0.090192	0.044158
tipCount	0.049765	0.035329	1.000000	0.970437	0.345447	0.030871	-0.038560	0.070823	0.034519	-0.000405	-0.149181	0.103151	0.030926
likes	0.060657	0.035949	0.970437	1.000000	0.385015	0.022340	-0.028859	0.052763	0.026097	-0.008814	-0.145628	0.101831	0.021901
rating	0.039480	-0.060689	0.345447	0.385015	1.000000	0.119848	0.125932	0.018462	0.121978	-0.142056	0.055939	0.187955	0.047040
population	0.233477	-0.025243	0.030871	0.022340	0.119848	1.000000	0.470962	0.521160	0.999850	0.356323	-0.013098	0.470726	0.874736
area_sq_mi	0.010007	-0.171375	-0.038560	-0.028859	0.125932	0.470962	1.000000	-0.279395	0.467783	-0.180031	-0.053215	0.403061	0.385279
density	0.307222	0.123456	0.070823	0.052763	0.018462	0.521160	-0.279395	1.000000	0.524296	0.701236	0.013250	0.025433	0.482868
total	0.234668	-0.025738	0.034519	0.026097	0.121978	0.999850	0.467783	0.524296	1.000000	0.357336	-0.016968	0.473562	0.874304
asian	0.209865	0.136554	-0.000405	-0.008814	-0.142056	0.356323	-0.180031	0.701236	0.357336	1.000000	-0.097611	-0.136473	0.357945
black	-0.140806	-0.146410	-0.149181	-0.145628	0.055939	-0.013098	-0.053215	0.013250	-0.016968	-0.097611	1.000000	-0.220436	-0.200477
hispanic	0.129680	-0.090192	0.103151	0.101831	0.187955	0.470726	0.403061	0.025433	0.473562	-0.136473	-0.220436	1.000000	0.094057
white	0.218718	0.044158	0.030926	0.021901	0.047040	0.874736	0.385279	0.482868	0.874304	0.357945	-0.200477	0.094057	1.000000

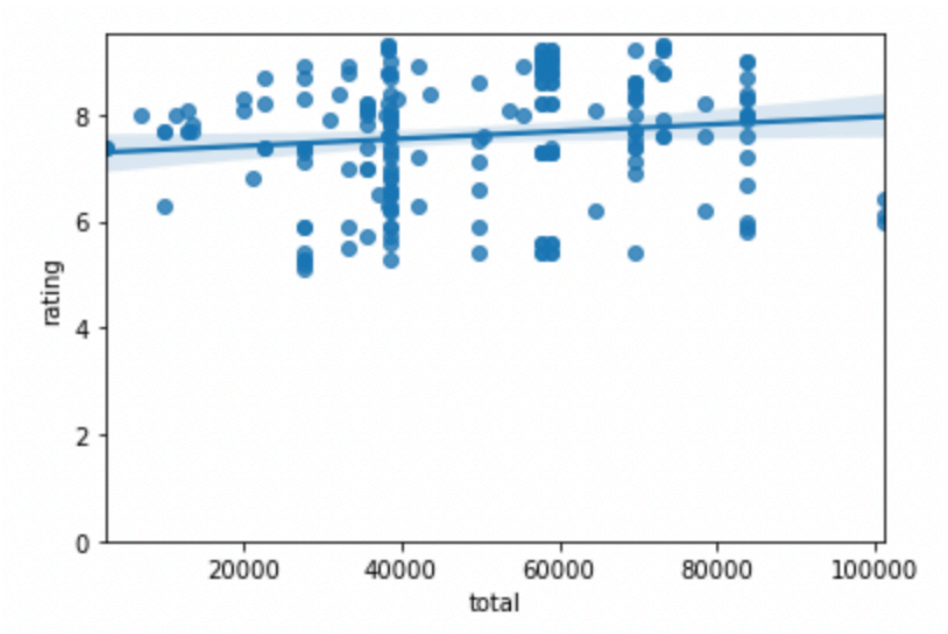
The first thing that stood out was that no fields tended to have much of a correlation with other fields. This does not bode well for us being able to give entrepreneurs helpful insight into which neighborhood a cafe should be started in. But still attempting to move forward, three things observations were made:

1. A small correlation between hispanic population and good rating for a cafe exists. This is also seen in total population.
2. larger square miles of a neighborhood seems to correlate with higher ratings a little.
3. ratings, likes, and tips all seem to be correlated

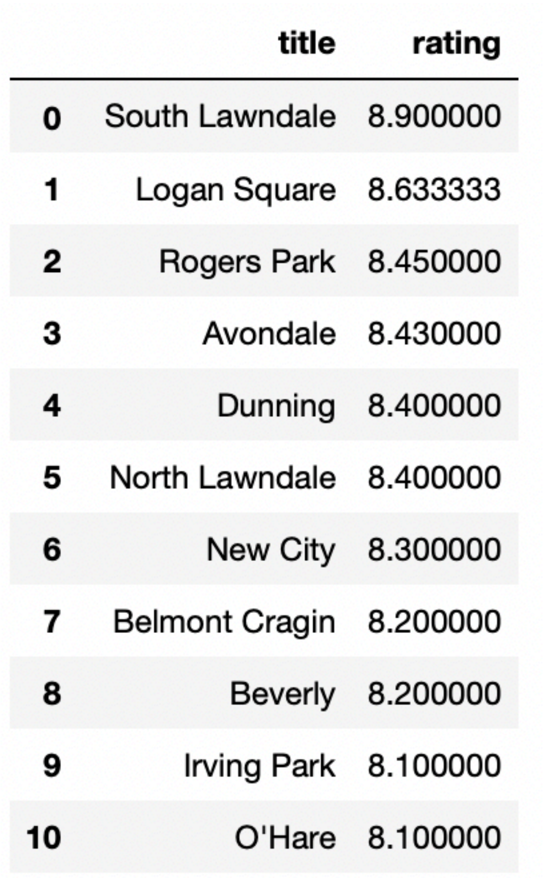
Mapping some of these shows just how weak the correlation is:



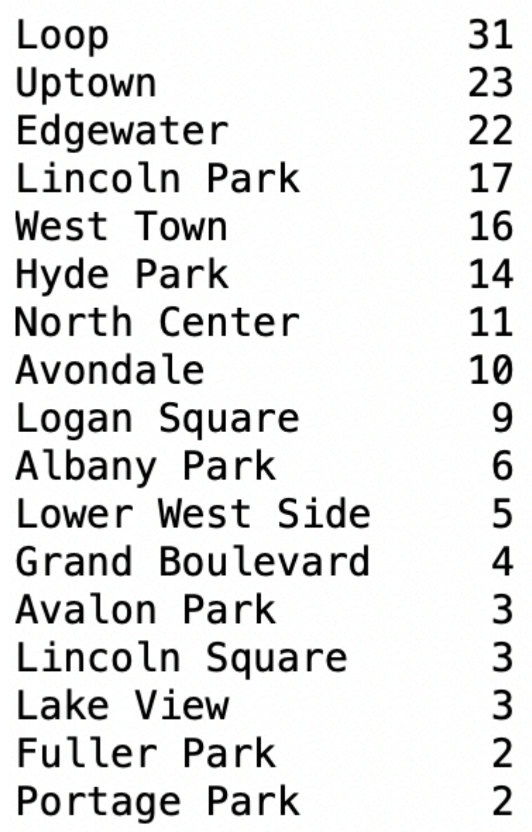




Next I wanted to look further into which neighborhoods had a higher average among all the cafe’s within them.



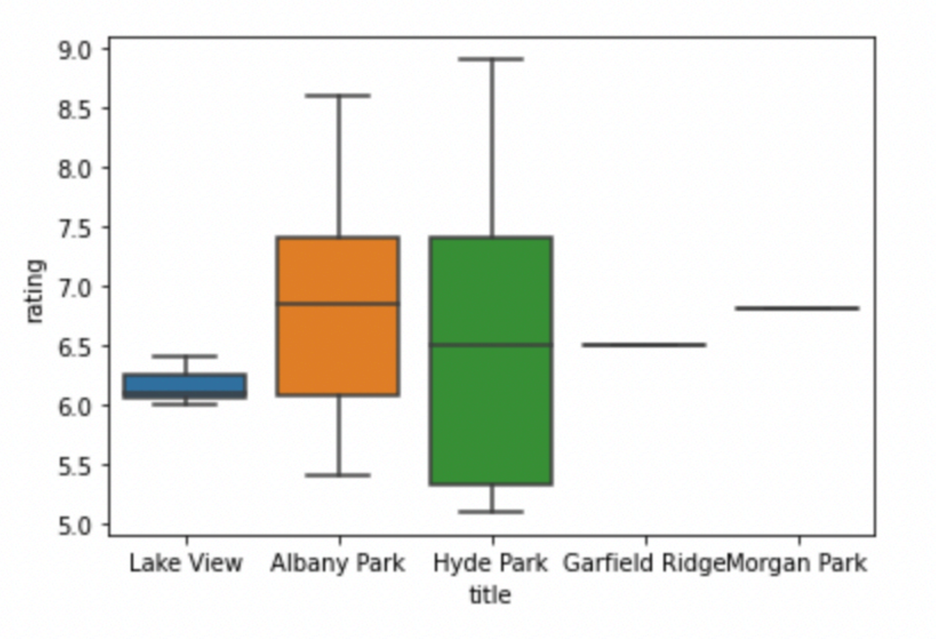
	title	rating
0	South Lawndale	8.900000
1	Logan Square	8.633333
2	Rogers Park	8.450000
3	Avondale	8.430000
4	Dunning	8.400000
5	North Lawndale	8.400000
6	New City	8.300000
7	Belmont Cragin	8.200000
8	Beverly	8.200000
9	Irving Park	8.100000
10	O'Hare	8.100000



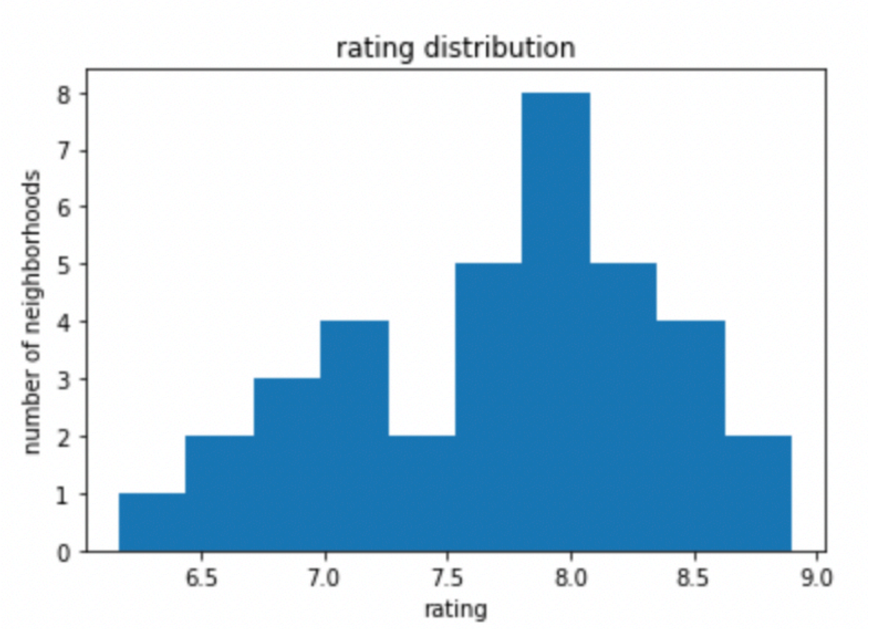
```
Loop                31
Uptown              23
Edgewater           22
Lincoln Park        17
West Town           16
Hyde Park           14
North Center        11
Avondale            10
Logan Square         9
Albany Park          6
Lower West Side      5
Grand Boulevard      4
Avalon Park          3
Lincoln Square       3
Lake View            3
Fuller Park          2
Portage Park         2
```

First thing I checked was to see if the same neighborhoods with higher average ratings also tended to have lots of review or very few. I did not see a relationship which was confirmed as the correlation is -0.109242.

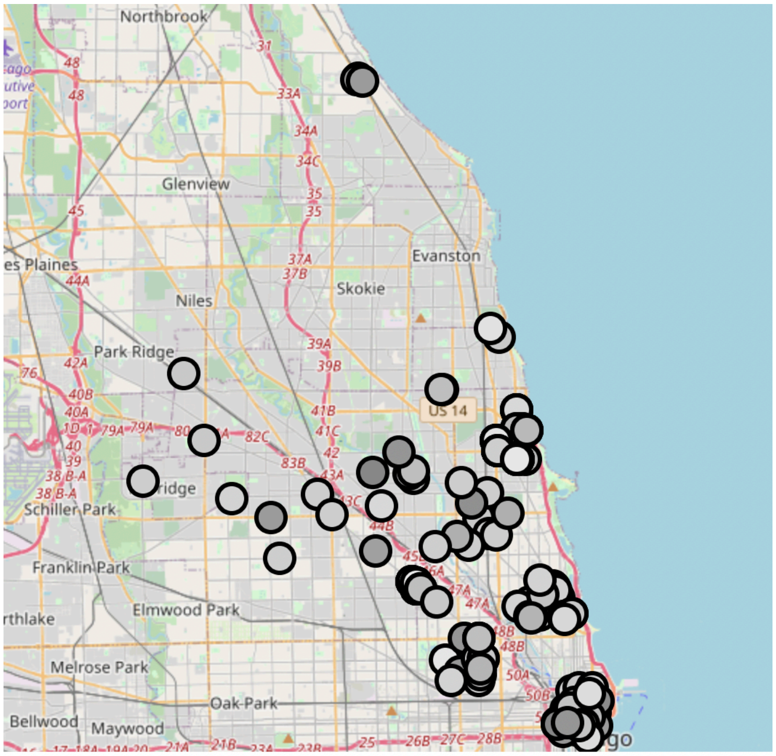
Looking at the top five neighborhoods for rating using a box plot also showed no uniformity in ratings.



Looking at the distribution of all ratings shows a Gaussian bell curve considering the low number of examples we have.



I also wanted to plot venues on a map to see if there is any geographical relationship between cafes and ratings. I plotted each venue at their location, the dot has a darker color the higher the rating was and lighter color for lower ratings. The presumption was that perhaps this might give us insight into if a specific location in the city tends to produce higher ratings. For example, maybe cafe's further north tend to get better ratings due to different mico-climates. The map does not appear to show any such relationship.



3**.2 Other observations**

One further observation to note that is about is the distribution of ratings among the neighborhoods. What I found is that the distribution was not equal at all. Some neighborhoods had ten or more venues that were rated, a handful had more than two, but fourteen of our neighborhoods had only one venue that was rated.

If we are not careful, this can seriously bias any further analysis that we may attempt to do. This also suggests that building a model that takes into account some of the characteristics of each neighborhood may allow us to generalize to the neighborhoods with few rated venues better than we could just by simple data analysis

Predictive modeling

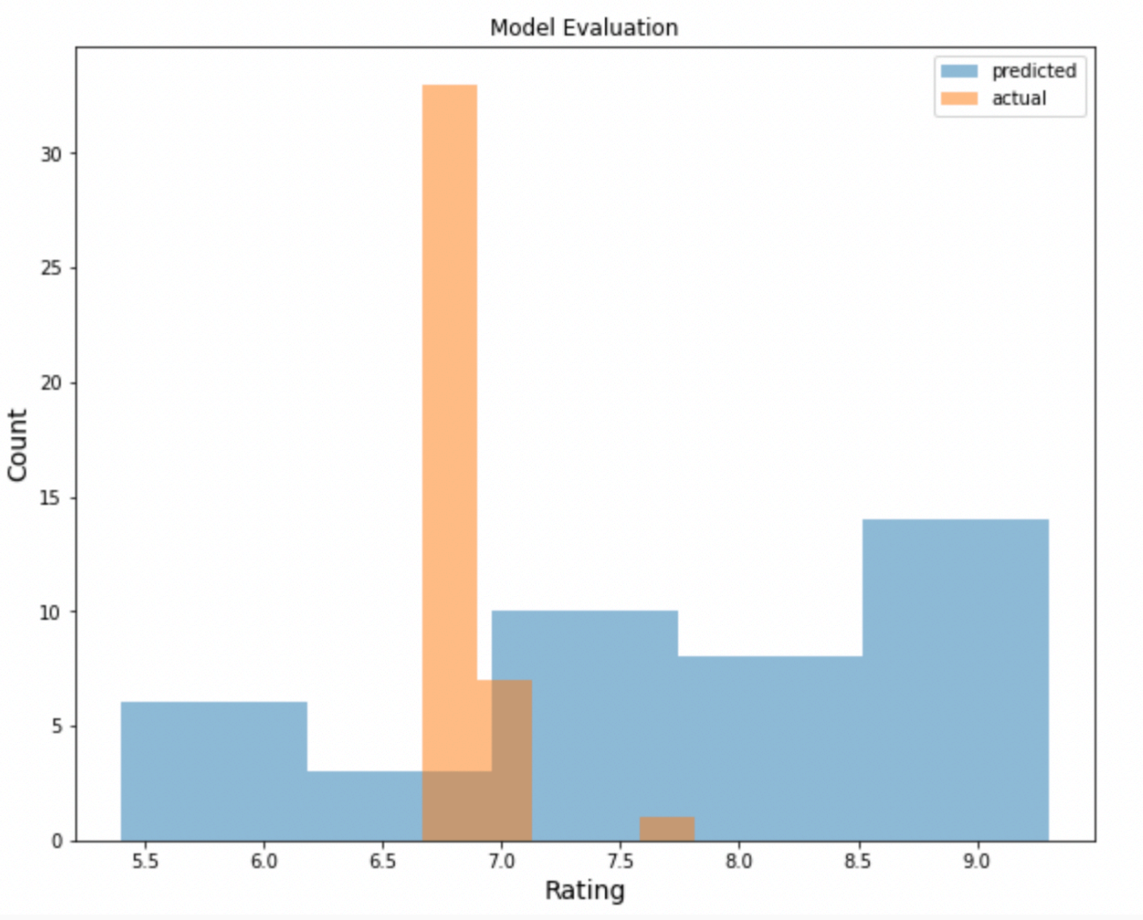
4.1 Regression model

I decided the best way to try help entrepreneurs is to help predict the rating they might receive on Foursquare based solely on the neighborhood they opened up in. Since rating is a continuous value, our model would need to output a continuous value. My first choice is simple linear model with the following features for each neighborhood:

total population, square miles, density, asian, black, hispanic, white. Each venue is passed as a training example with the features above which are always the same for the neighborhood they reside in. The model is asked to predict the rating for that specific venue. The hope is that if a relationship exists between any combination of these features and the venue, we could then help predict new rating levels for a cafe based solely on the neighborhood. What we found though was that this model performed very poorly with a r score of only 0.09 on the test data.

4.2 Polynomial Linear model

Next I attempted to use a polynomial linear model to ty and capture more complexity in the model. With some experimenting, I found the optimal Polynomial degree to be 5. But even in this case the r score was low, 0.24. Plotting predicted outcomes against actual outcomes in the test dataset shows the linear model could not use neighborhood data to find patterns that diverged the outcome outside a small range between 6.5 and 7.



5. Conclusion

The poor results for these models help bolster that nothing within the available data we have will help us build out a robust model that can help us predict how well a cafe is rated based on its neighborhood in Chicago. It could perhaps be that we are missing important information such as socio-economic status of a neighborhood or primary working industries for residents. Most likely though, ratings are probably more guided by how good of cafe you make or how cool your cafe is.

I suspect that expanding this analysis to a larger region of the country may result in a more fruitful analysis, as although the neighborhoods of Chicago can be culturally diverse, they still will be much more similar that that of regions further away from each other in the country. Different cultures will likely value and rate cafe houses different.

The results of this project were not a waste, as in science it is often just as important to show where correlation and causation are not as it is to show where they are. This study helped inform us that neighborhood differences such as size and ethinicity distributions will not impact ratings on Foursquare which was our indicator of success for a cafe. This frees entrepreneurs to pick any neighborhood within the city that they like best without fear that it was sub-optimal.