Final Report for CMSC12300 by WACC

1. Introduction

Yelp is one of the most popular website/mobile application in the world. Yelp runs “Yelp Dataset Challenge” every year, and it generously offers abundant data (not the entire dataset obviously) to the public. The main purpose of our project is to understand how competition within a neighborhood affect restaurants. We constructed measures that describe the general features and competition pressure of neighborhoods and built models to figure out how these features related to the success of restaurants.

2. Data

Yelp public dataset available on: https://www.yelp.com/dataset/download. The dataset is a 6.52 gigabytes JSON file that contains information about 174,000 businesses (over 1.2 million business attributes like opening hours, parking, availability, and ambience) and 5,200,000 reviews contributed from 1,300, 000 users. We converted all the json files to csv format files for our convenience. After pairing restaurants that are less than 1.5 km from each other, we get more than 26,914,288 pairs of merely id of restaurants in a 2.6 GB csv file. The 2 GB review csv file is used to train our Natural Language Processing models. Because of the our limited access to Google Cloud, our main analysis bases on a subset. We used data of all restaurants in four cities (Phoenix, Cleveland, Madison and Markham). The total number of restaurants in these four cities is 6,797. After pairing restaurants that are less than 1.5 km from each other, we get 629,670 pairs of restaurants.

3. Hypothesis

In this project, we use Yelp data to analyze factors that lead to the success of a restaurant from the perspectives of restaurant characteristics (star rating, price range, number of reviews) and spatial characteristics such as how many other restaurants are nearby and how does it compare to nearby restaurants based on price, ratings, etc. Hence, we proposed the following hypotheses:

1. Identifying the neighborhood effects:
    a. The performance of a restaurant is positively correlated with the restaurants nearby
2. Exploring the factors to success:
    a. Successful restaurants face less competition from outside

4. Methodology

We define the neighborhood of a restaurants in two ways:

1. The region of a circle with a radius of 1.5km, where the restaurant is at the center of the circle
2. For each city we study, we use k-means clustering to classify the restaurants into ten groups based on their location.

Then, we define the following indices to measure the competition level within a neighborhood:

1. Number of restaurants nearby
2. Average star rating for the restaurants nearby
3. Average price range for the restaurants nearby

4. Average number of reviews for the restaurants nearby. The number of reviews is treated as a proxy for the popularity of a restaurants.
5. Homogeneity for the restaurants nearby. Specifically, we examine the overlapping in business of restaurants. Yelp classifies restaurants by various dimensions including the type of food offered, the ambience and dining environment. These classifications are stored as a list in the dataset. We first construct a dummy variable to indicate where there is at least one same category in the defined the category list. Furthermore, we use a more complex measure, called Levenshtein distance to measure the similarity in categories for any two restaurants. The definition of Levenshtein distance is defined as the minimum step required to make the two lists identical to each other. It takes into account of the differences in length efficiently. Thought the ordering of elements in list is important in this measure, which might be essential to our case, generally speaking, the similarity in the business of two restaurants is higher with smaller distances.

- Neighborhood defined by K-means clustering model

1. Algorithm

   The algorithm operates iteratively between two steps:

   a. Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if u_i is the centroids in set S_i, then each data point x is assigned to a cluster based on

   $$\operatorname{argmin}_s \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

   b. The centroids are recomputed in this step by taking the mean of all data points assigned to that centroid's cluster.

   $$\mu_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

   We set the number of clusters to be 10 for each city, and we use latitude and longitude as the features to the model. The model is trained separately for the eleven cities with the most recorded restaurants in our dataset. We first train our model using a random draw of 20% of the restaurants and then predict the remaining restaurants to which cluster it belongs to based on its location. By each cluster, we calculate the total number of restaurants, average price range, average star rating and average number of reviews. A total 108 observations are obtained (two omitted since two cities are unable to identify 10 clusters). Then we run a regression to test the significance of "competition", the number of restaurants in the cluster i denoted as x_i, to different performance measure y_i, which are the average star rating, price level and popularity (number of reviews). The model is specified as follows:

   $$y_i = \beta_0 + \beta_1 x_i + u_i$$

   The regression results for rating, price level and popularity are:

1. Rating: the p-value is 0.831 and the R^2 is extremely close to 0, suggesting that competition does not affect the rating of a restaurant.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      1   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.009
Method:                 Least Squares   F-statistic:                   0.04599
Date:                Mon, 04 Jun 2018   Prob (F-statistic):              0.831
Time:                        04:38:42   Log-Likelihood:                 20.258
No. Observations:                 108   AIC:                            -36.52
Df Residuals:                     106   BIC:                            -31.15
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                            3.4554      0.026    132.798      0.000       3.404       3.507
number_of_restaurants_in_cluster 1.863e-05   8.69e-05      0.214      0.831      -0.000       0.000
==============================================================================
Omnibus:                        5.690   Durbin-Watson:                   0.655
Prob(Omnibus):                  0.058   Jarque-Bera (JB):                5.836
Skew:                           0.556   Prob(JB):                       0.0540
Kurtosis:                       2.758   Cond. No.                         400.
==============================================================================
```

2. Price Level: again, the p-value for the number of restaurants in the cluster is 0.635, we should reject the null hypothesis that the competition affects the pricing of a restaurant.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      2   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                 -0.007
Method:                 Least Squares   F-statistic:                    0.2266
Date:                Mon, 04 Jun 2018   Prob (F-statistic):              0.635
Time:                        04:39:07   Log-Likelihood:                 15.407
No. Observations:                 108   AIC:                            -26.81
Df Residuals:                     106   BIC:                            -21.45
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                            1.6376      0.027     60.173      0.000       1.584       1.692
number_of_restaurants_in_cluster 4.325e-05   9.09e-05      0.476      0.635      -0.000       0.000
==============================================================================
Omnibus:                       12.409   Durbin-Watson:                   0.774
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               13.114
Skew:                           0.823   Prob(JB):                      0.00142
Kurtosis:                       3.449   Cond. No.                         400.
==============================================================================
```

1. Number of reviews: the p-value is 0.013 so we can conclude significance for “competition” at 5% confidence level. The result suggests that people visit more frequently for the restaurants that are surrounded by many other restaurants.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      3   R-squared:                       0.300
Model:                            OLS   Adj. R-squared:                  0.293
Method:                 Least Squares   F-statistic:                     45.34
Date:                Mon, 04 Jun 2018   Prob (F-statistic):           8.81e-10
Time:                        04:39:34   Log-Likelihood:                -521.18
No. Observations:                 108   AIC:                             1046.
Df Residuals:                     106   BIC:                             1052.
Df Model:                           1
Covariance Type:            nonrobust
====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                               32.5716      3.913      8.323      0.000      24.813      40.330
number_of_restaurants_in_cluster     0.0880      0.013      6.733      0.000       0.062       0.114
==============================================================================
Omnibus:                       16.800   Durbin-Watson:                   0.764
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               43.922
Skew:                           0.476   Prob(JB):                     2.90e-10
Kurtosis:                       5.975   Cond. No.                         400.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

- Fit success score using neighborhood defined by distance < 1.5 km

1. OLS regressions:
   Model: Explore the relationship between success score and average rating, average price range, average number of reviews, average review similarity, average category similarity and average sentiment score for each restaurant (treat the restaurant i as the center and then compute average value of these attributes for the neighborhood around i):

$$success\_score_i = constant + avr\_rating_i + avr\_price\_range_i + avr\_num\_reviews_i \\ + avr\_review\_sim_i + avr\_category\_sim_i + avr\_sent\_score_i$$

   Where success score is computed by restaurant rating times restaurant review score. The review score for each restaurant is determined by its quartile rank (one of [0.25, 0.50, 0.75, 1.0]) of number of reviews within the city. For example, if a restaurant's number of reviews falls in the first quartile, it will receive review score 1, etc. This definition aims to alleviate the situation that a restaurant receives a relatively low rating but has a large volume of reviews.

   In this model, a neighborhood is defined by the region of a circle with a radius of 1.5km, where the restaurant is at the center of the circle. The dependent variable of our OLS model is the "success score", which is defined in the previous paragraph. We calculated neighborhood-based average rating, average price rank, average number of reviews, average reviews similarity, average categories similarity and average sentiment score of reviews. These average variables are explanatory variables in the model. For example, there are 10 restaurants are less than 1.5 km away from restaurant A. The average rating of the neighborhood is thus the average price of these 10 restaurants, and so forth.

2. Similarity:
   a. Category similarity

Every restaurant on Yelp has a list of categories (e.g. ['Breakfast & Brunch', 'Diners', 'Restaurants', 'Cafes', 'British']). We use an algorithm to calculate the the similarity of reviews of a pair restaurants. Actually, the similarity score we get measures how many differences should be done to make two list of categories identical. Therefore, the higher the value of variable ("avg_cate_similarity") is, the less similar the categories of a paired restaurants are.

b. Reviews similarity using Natural Language Processing
We fully utilized Yelp reviews data to analyze features of restaurants. Specifically, we analyze sentiment of reviews and calculate similarities of reviews of each pair of restaurants
    1) Sentiment Analysis -- using NLTK
    Calculate the general sentiment score (-1 to 1) of each restaurant's reviews, and take the median as the sentiment score of the restaurant. A high score indicates positive sentiment of reviews.
    2) Reviews Similarity -- using Bag of Words model
    We first construct bag-of-words sparse matrix of each restaurant's reviews. Then we calculate the reviews similarity score (cosine similarity) between each pair restaurant within the neighborhood. Take the average similarity as the reviews similarity of the neighborhood.

5. Results
    1. Regression outputs for Phoenix, Madison, Markham, and Cleveland respectively:

<u>Phoenix:</u>

OLS Regression Results

Dep. Variable:	score	R-squared:	0.145
Model:	OLS	Adj. R-squared:	0.144
Method:	Least Squares	F-statistic:	101.8
Date:	Mon, 04 Jun 2018	Prob (F-statistic):	1.14e-118
Time:	13:52:01	Log-Likelihood:	-10538.
No. Observations:	3606	AIC:	2.109e+04
Df Residuals:	3599	BIC:	2.113e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.3776	1.334	-0.283	0.777	-2.994	2.238
avr_rating	0.7309	0.330	2.216	0.027	0.084	1.378
avr_price	0.9365	0.491	1.909	0.056	-0.025	1.898
avr_num_review	0.0117	0.002	5.701	0.000	0.008	0.016
avr_re_sim	0.5114	0.667	0.766	0.444	-0.797	1.820
avr_cate_sim	0.8893	0.053	16.709	0.000	0.785	0.994
avr_sent_score	-0.0090	0.419	-0.022	0.983	-0.832	0.813

Omnibus:	220.939	Durbin-Watson:	2.055
Prob(Omnibus):	0.000	Jarque-Bera (JB):	113.926
Skew:	0.265	Prob(JB):	1.83e-25
Kurtosis:	2.310	Cond. No.	2.09e+03

The regression output shown above is for all restaurants in Phoenix. The estimated coefficients for average rating, average num of reviews, and average category similarity are statistically significant at 5% significance level. The value of R^2 is 0.145, which is acceptable.

Madison:

OLS Regression Results

Dep. Variable:	score	R-squared:	0.082
Model:	OLS	Adj. R-squared:	0.076
Method:	Least Squares	F-statistic:	15.06
Date:	Mon, 04 Jun 2018	Prob (F-statistic):	1.45e-16
Time:	13:52:02	Log-Likelihood:	-2973.7
No. Observations:	1020	AIC:	5961.
Df Residuals:	1013	BIC:	5996.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4600	3.755	0.123	0.903	-6.908	7.828
avr_rating	1.2602	0.921	1.369	0.171	-0.547	3.067
avr_price	-0.3346	1.061	-0.315	0.752	-2.416	1.747
avr_num_review	0.0348	0.010	3.450	0.001	0.015	0.055
avr_re_sim	1.8832	1.261	1.494	0.136	-0.591	4.357
avr_cate_sim	0.5668	0.106	5.370	0.000	0.360	0.774
avr_sent_score	-1.5812	0.829	-1.908	0.057	-3.208	0.045

Omnibus:	153.401	Durbin-Watson:	1.947
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47.923
Skew:	0.272	Prob(JB):	3.92e-11
Kurtosis:	2.087	Cond. No.	1.82e+03

Compared to OLS regression for restaurants in Phoenix, only the OLS estimates for average number of reviews and average category similarity are statistically significant at 5% level, using data for restaurants in Madison. Besides, the value of R^2 drops to 0.082, which means that the majority variation of success score should be explained by factors outside the model.

Markham:

OLS Regression Results

Dep. Variable:	score	R-squared:	0.041
Model:	OLS	Adj. R-squared:	0.033
Method:	Least Squares	F-statistic:	5.312
Date:	Mon, 04 Jun 2018	Prob (F-statistic):	2.23e-05
Time:	13:52:02	Log-Likelihood:	-2139.9
No. Observations:	757	AIC:	4294.
Df Residuals:	750	BIC:	4326.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.2953	3.662	1.446	0.149	-1.894	12.484
avr_rating	-1.2293	1.225	-1.004	0.316	-3.634	1.175
avr_price	2.7998	1.163	2.407	0.016	0.516	5.084
avr_num_review	0.0159	0.015	1.037	0.300	-0.014	0.046
avr_re_sim	-0.9221	1.336	-0.690	0.490	-3.544	1.700
avr_cate_sim	0.5090	0.120	4.227	0.000	0.273	0.745
avr_sent_score	-0.5519	0.733	-0.753	0.452	-1.991	0.887

Omnibus:	151.743	Durbin-Watson:	1.969
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.940
Skew:	0.321	Prob(JB):	4.74e-10
Kurtosis:	2.025	Cond. No.	1.11e+03

For Markham, the estimated coefficients for average price range and average category similarity are statistically significant at 5% level. However, the value of R^2 is only 0.041, illustrating a poor performance of the proposed model for fitting the success score of restaurants in Markham.

<u>Cleveland:</u>

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  score   R-squared:                       0.268
Model:                            OLS   Adj. R-squared:                  0.226
Method:                 Least Squares   F-statistic:                     6.341
Date:                Mon, 04 Jun 2018   Prob (F-statistic):           1.04e-05
Time:                        13:52:02   Log-Likelihood:                -317.06
No. Observations:                 111   AIC:                             648.1
Df Residuals:                     104   BIC:                             667.1
Df Model:                           6
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const             -3.6970      8.057     -0.459      0.647     -19.675      12.281
avr_rating        -1.7445      1.755     -0.994      0.323      -5.225       1.736
avr_price          8.3090      2.985      2.784      0.006       2.389      14.228
avr_num_review    -0.0023      0.025     -0.090      0.929      -0.052       0.048
avr_re_sim         0.1753      3.865      0.045      0.964      -7.489       7.839
avr_cate_sim       1.1761      0.376      3.131      0.002       0.431       1.921
avr_sent_score     1.4963      2.458      0.609      0.544      -3.377       6.370
==============================================================================
Omnibus:                        2.022   Durbin-Watson:                   2.333
Prob(Omnibus):                  0.364   Jarque-Bera (JB):                1.749
Skew:                          -0.179   Prob(JB):                        0.417
Kurtosis:                       2.501   Cond. No.                     1.31e+03
==============================================================================
```

This is the OLS regression output for restaurants in Cleveland. The R-squared is 0.268, which is the best among all 4 cities we compared. The OLS estimates of average number of reviews and average category similarity are statistically significant at the level of 5%. Again, although the R-squared is much better, most of the coefficients are not significant for predicting success score of restaurants in Cleveland.

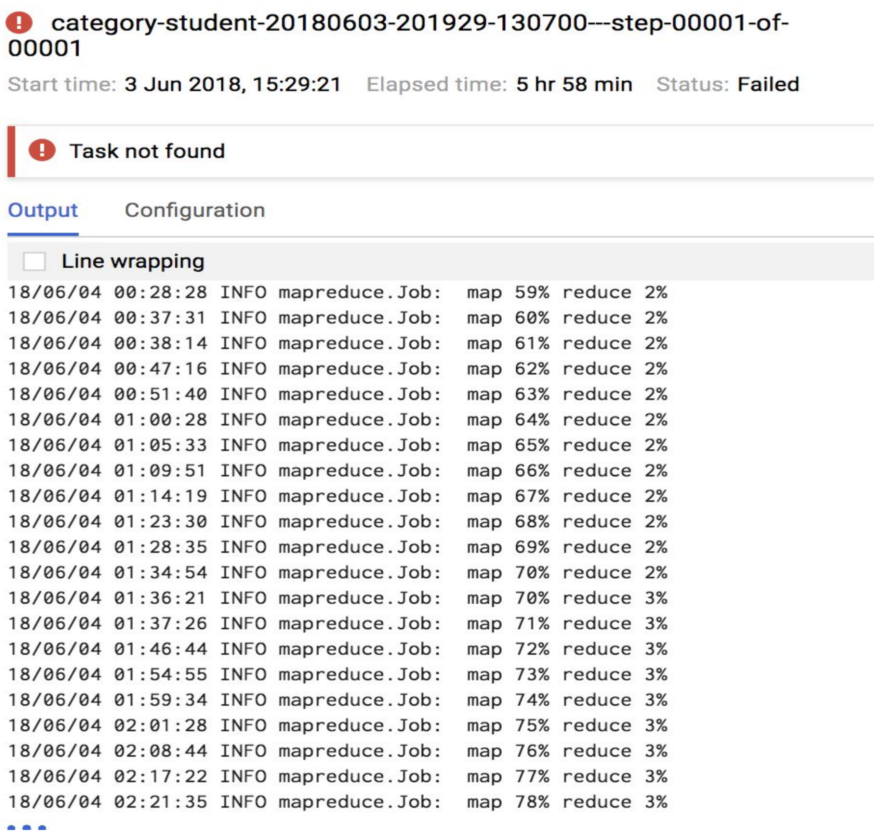
2. Summary for coefficients:

```
Coefficient Report
                     PNX       MAD       MAK       CLV
const          -0.377624  0.459997  5.295250 -3.696987
avr_rating      0.730900  1.260181 -1.229257 -1.744524
avr_price       0.936502 -0.334633  2.799759  8.308972
avr_num_review  0.011704  0.034832  0.015858 -0.002260
avr_re_sim      0.511410  1.883192 -0.922069  0.175297
avr_cate_sim    0.889283  0.566770  0.509000  1.176106
avr_sent_score -0.009030 -1.581171 -0.551870  1.496286
```

3. Plots for results (see Appendix):

In Appendix, we display plots for OLS regressions if we only have one independent variable (e.g. avr_rating, avr_price, etc.) in the model.

6. Challenge Limitations



One of our biggest challenges for this project comes from Google Cloud Dataproc. At the beginning, we had trouble setting up the configuration. While running our code using Google Cloud Dataproc, we met different kinds of error such as Dataproc exception error and broken pipe error, and it's difficult for us to infer what happened exactly. We tried to run a small set on Google dataproc and it takes about 10 times longer when we run on our own laptop. The Google dataproc failed when we run our full-data, which are several files with around 2GB each. Then we tried to partition our data into some small sets, it seems that Google Dataproc is unable to handle our function using a file takes about 400MB. The above job was among one of the jobs that was killed by a broken pipe error. The Dataproc job details were also not very helpful for interpreting the errors (taking up to many temporary memory for our case?). Thus, we were not able to run the whole dataset to find all the pairs of restaurants.

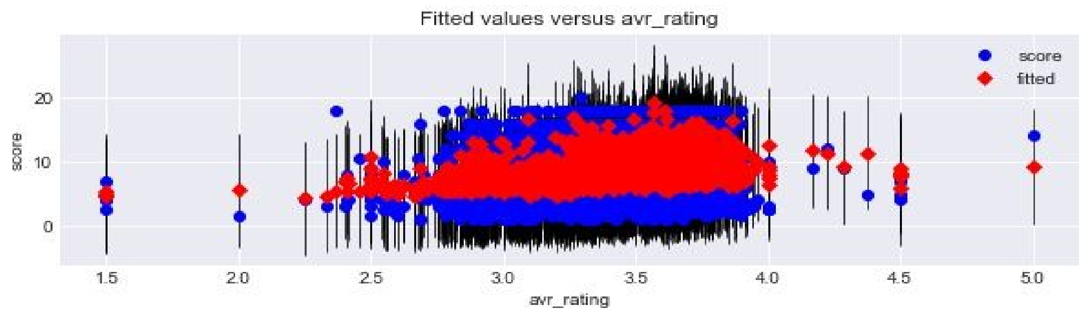
Yelp data is not comprehensive, and we don't know how the samples were collected by yelp, so we were unable to know whether the data are representative or not. More in-depth analysis can be conducted if we have access to some economic datasets at city level, for example, the traffic for each road, the population density and the average income within each neighborhood. One possible source of data is Uber dataset or city taxi dataset. We can look at the frequency of pick-up and drop-off spots to identify the actual human traffic for each area, and examine the impact of site selection to the success of a restaurant. Moreover, the current definition of neighborhood is based on haversine distance, while, it might be more appropriate to use the driving or walking distance.

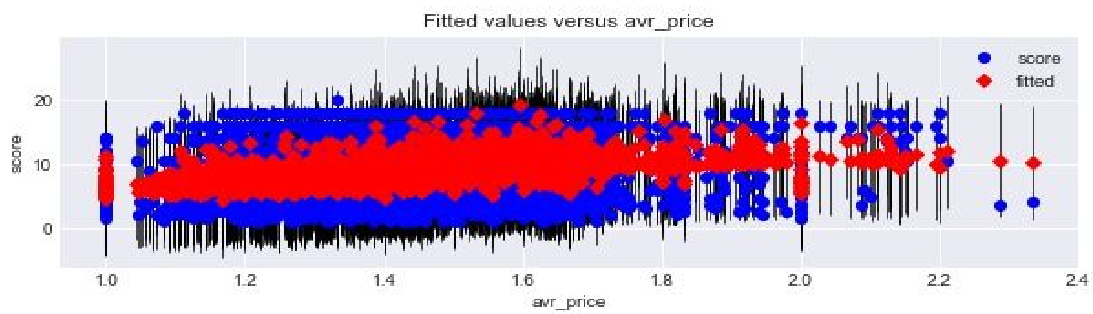
We have noticed that when we ran our jobs using Google Cloud Dataproc, the CPU usage is around 7% and 8%, so how to use CPU more sufficiently is also one of the problems and challenges we met, and unable to find a better solution.

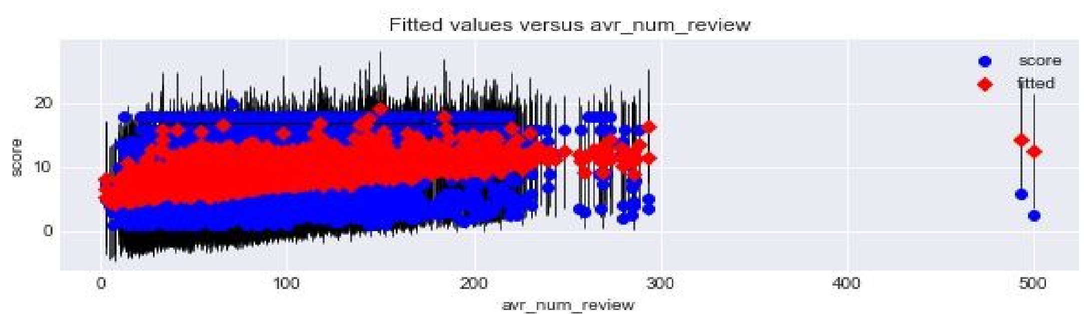
7. Conclusion

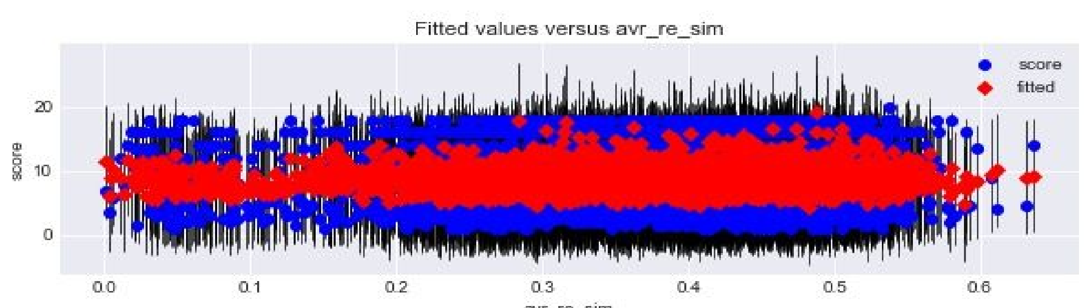
Based on the OLS regression results across different cities, the average category similarity is the most powerful variable to predict a restaurant's success score. Since we define the category similarity score as how many differences should be done to make two list of categories identical, the larger average category similarity means higher diversity of restaurants within the neighborhood. However, there is no coincident conclusion about the explanatory power of other proposed variables. One potential reason could be that the data collected by Yelp is not representative. Additionally, the variables included in our model likely suffer from the issue of multicollinearity.

Appendix:
Plots for each city (OLS results):
<u>Phoenix:</u>

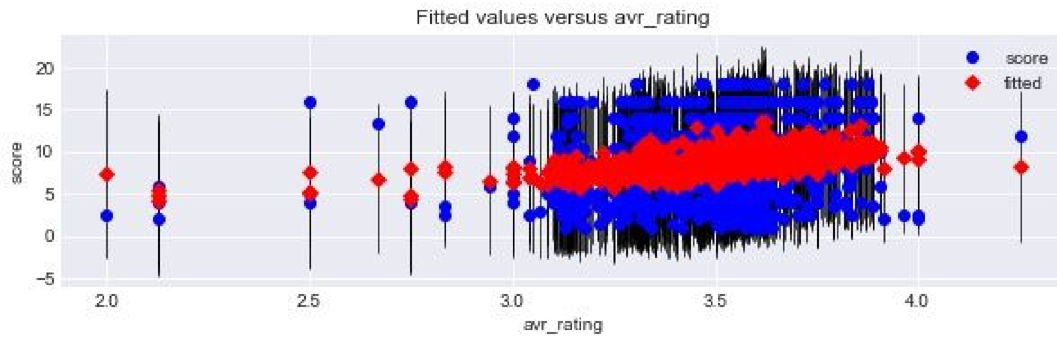


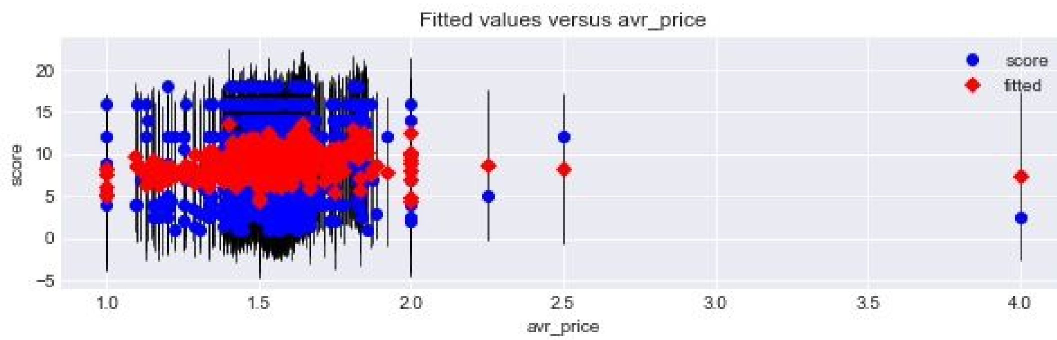


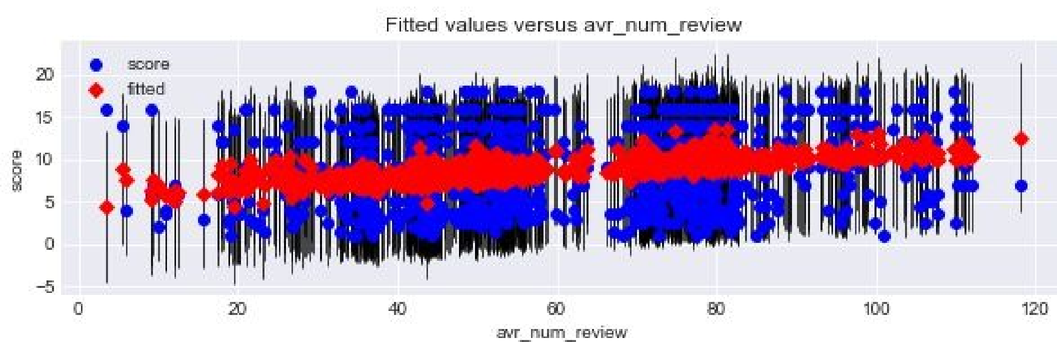


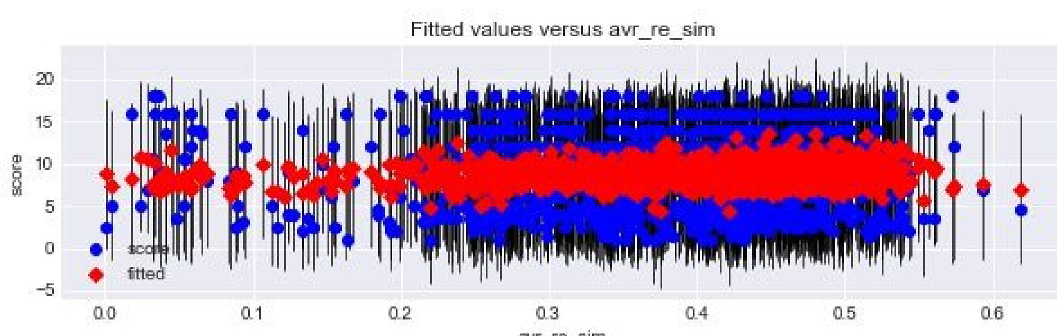


Madison:

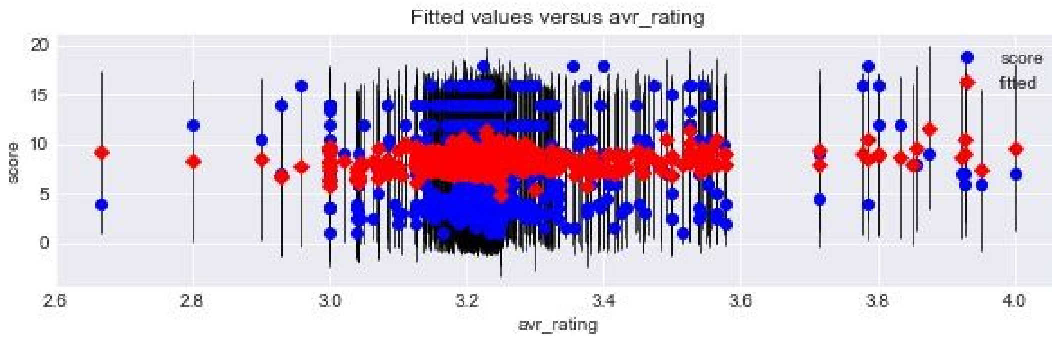

Fitted values versus avr_rating
score
fitted
score
avr_rating
20
15
10
5
0
-5
2.0
2.5
3.0
3.5
4.0

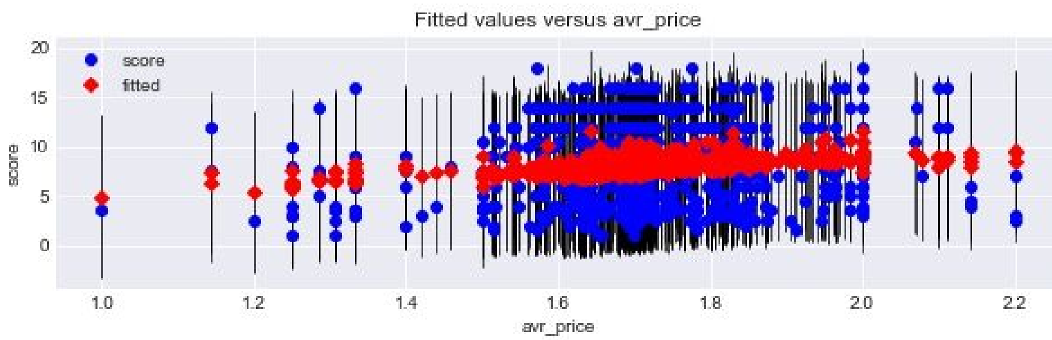

Fitted values versus avr_price
score
fitted
score
avr_price
20
15
10
5
0
-5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

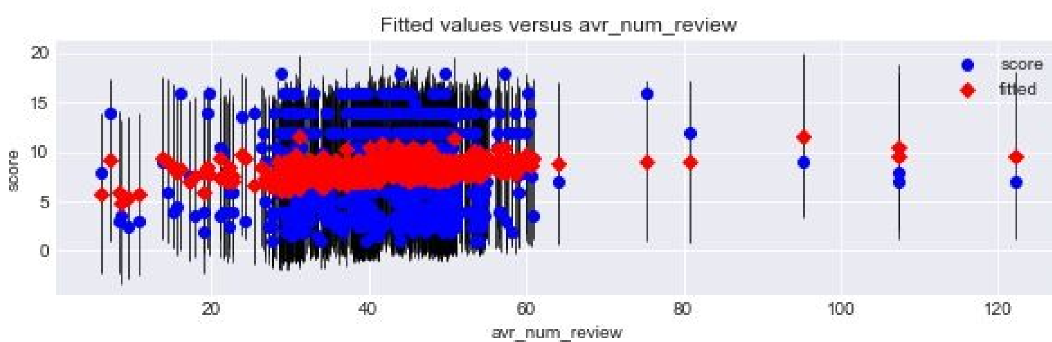

Fitted values versus avr_num_review
score
fitted
score
avr_num_review
20
15
10
5
0
-5
0
20
40
60
80
100
120

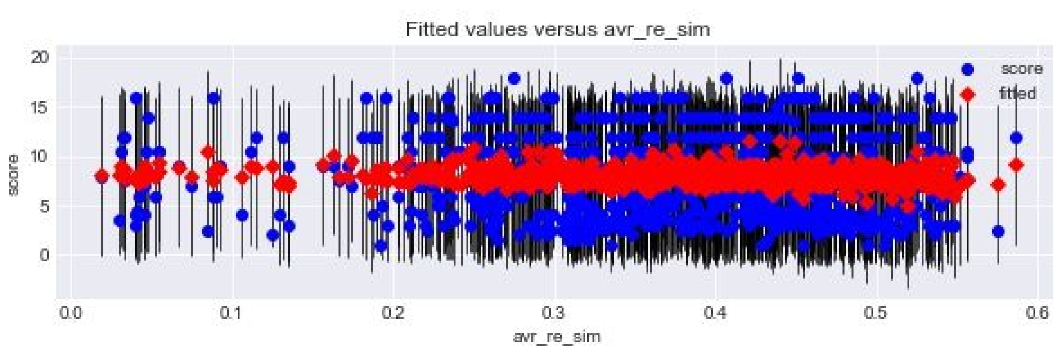

Fitted values versus avr_re_sim
score
fitted
score
avr_re_sim
20
15
10
5
0
-5
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Markham:

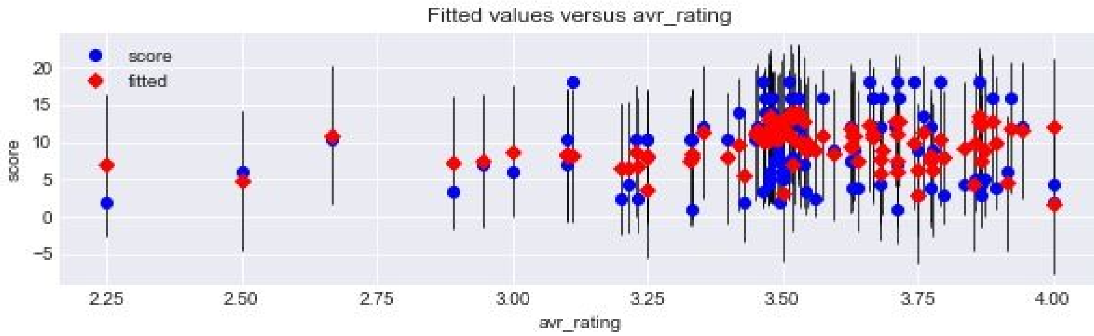

Fitted values versus avr_rating
score
fitted
score
avr_rating
0
5
10
15
20
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0

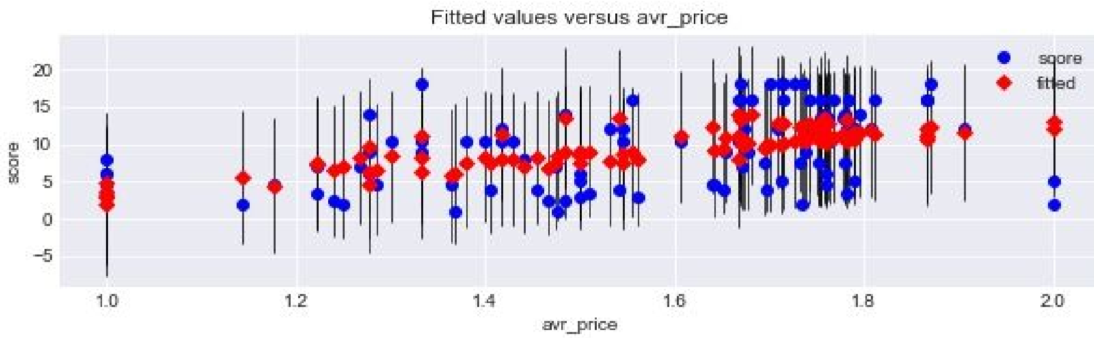

Fitted values versus avr_price
score
fitted
score
avr_price
0
5
10
15
20
1.0
1.2
1.4
1.6
1.8
2.0
2.2

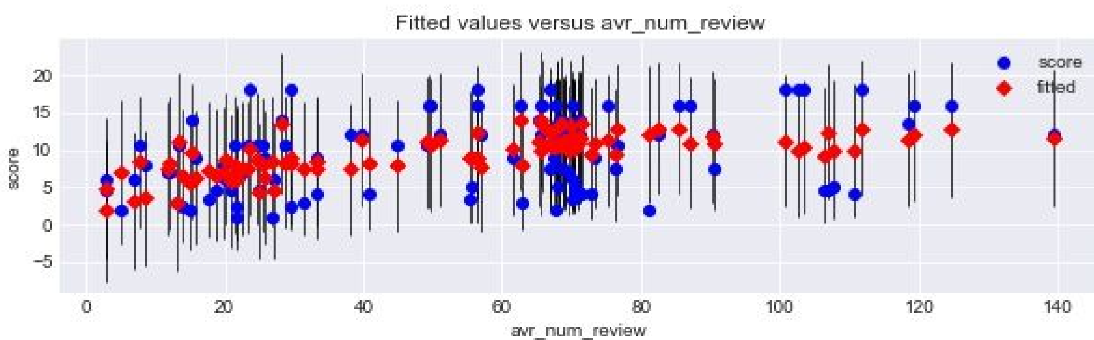

Fitted values versus avr_num_review
score
fitted
score
avr_num_review
0
5
10
15
20
20
40
60
80
100
120

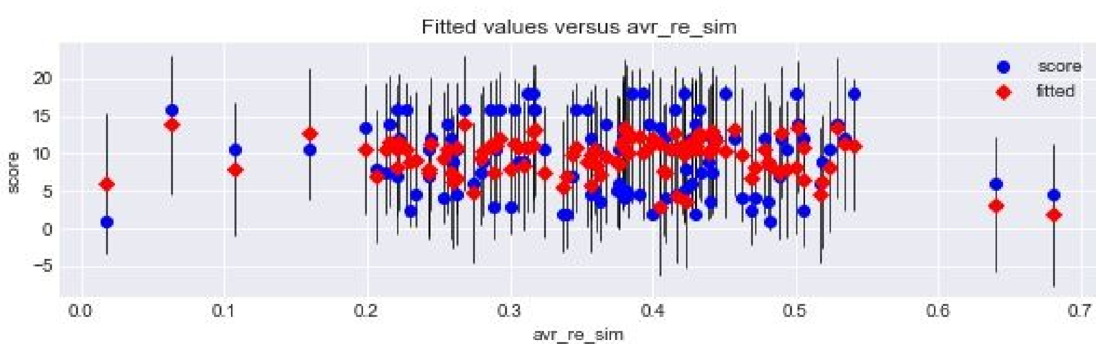

Fitted values versus avr_re_sim
score
fitted
score
avr_re_sim
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Cleveland:


Fitted values versus avr_rating
score
fitted
score
avr_rating
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
-5
0
5
10
15
20


Fitted values versus avr_price
score
fitted
score
avr_price
1.0
1.2
1.4
1.6
1.8
2.0
-5
0
5
10
15
20


Fitted values versus avr_num_review
score
fitted
score
avr_num_review
0
20
40
60
80
100
120
140
-5
0
5
10
15
20


Fitted values versus avr_re_sim
score
fitted
score
avr_re_sim
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
-5
0
5
10
15
20