Location, Location, Location! Quantifying the True Impact of Location on Business Reviews Using a Yelp Dataset

Abu Saleh Md. Tayeen
Department of Computer Science
New Mexico State University
Las Cruces, NM, USA
tayeen@nmsu.edu

Abderrahmen Mtibaa
Department of Computer Science
New Mexico State University
Las Cruces, NM, USA
amtibaa@cs.nmsu.edu

Satyajayant Misra
Department of Computer Science
New Mexico State University
Las Cruces, NM, USA
misra@cs.nmsu.edu

***Abstract*—Today, with the emergence of various business review sites such as Yelp, Trip Advisor, and Zomato, people can write reviews and provide an assessment (often as 1-5 score rating). The success of a business on the crowd-sourced review platform has taken the form of positive reviews and high star ratings (failure are associated with negative reviews and low star ratings). We often claim that location plays a major role in determining the success or the failure of a given business. This paper attempts to verify this claim and quantifies the impact of location, solely, on business success, using two data sets; a Yelp dataset for business information and reviews, and another Location dataset that gathers location-based information in a city or an area. We perform an empirical study to quantify the impact of (i) relative location to well known landmarks and (ii) parameterized location (such as cost of living in a given zip code), on the success of restaurants. In our study, we found that parameterized location using location characteristic parameters such as housing affordability correlate highly with restaurant success with more than 0.81 correlation ratio. We also observe that the closer the restaurant to a landmark (relative location) the more likelihood it succeeds.**

***Index Terms*—Location, Restaurant, Business graph, Yelp**

I. Introduction

We often hear the cliché, used by business experts, that the three most important factors in determining the success of a business are "location; location; location". Studies have investigated recommendation systems for optimal location to expand an existing business utilizing online customer reviews [1] and check-in patterns [2] or labeling business popularity using customers online social connections and geographical proximity to businesses [3]. However, none of these works have tried to evaluate the impact of location on the success of a business. *In this work, we choose restaurant as representative business and perform empirical investigation to quantify how much of a restaurant's success can be affected by its location.*

In this paper, we investigate the impact of (i) *parameterized location* and (ii) *relative location* on the success of restaurants measured as combination of its review counts and star ratings. While we refer to *relative location* as the proximity of a given location to properties, landmarks, parks, etc., the *parameterized location* is the geographic location devoid of its surroundings. *Parameterized location* is represented by a zip code associated with its characteristic parameters such as the cost of living in the given zip code. We adopt a data driven approach to quantify and compare the impact of parameterized locations for different categories (*e.g.,* population demographics, tourism potential, safety) of parameters and relative locations with respect to other restaurants and landmarks on the success of restaurants.

This paper attempts to answer two main research questions, RQ1 and RQ2:

RQ1: Can we determine the potential success of any restaurant solely from its location (parameterized and/or relative)?

RQ2: Which (parameterized and/or relative location) contribute the most to the success of restaurants?

To address these research questions, we perform two data driven experiments, namely parameterized location and relative location based experiments. The parameterized location based experiments consist of measuring the location-by-location (with zip code granularity) correlations between the success of restaurants and the characteristic parameters of the locations those are situated in. In the relative location based experiment, we examine the impact of restaurants' success with respect to distance to landmarks or other restaurants. We construct a restaurant vicinity graph that interconnects restaurants and city landmarks in a given city.

Our empirical experiments have shown that, parameterized location has significant effect in determining restaurants' success. We achieve a 0.81 correlation ratio between location parameters and restaurant success. Moreover, we observe that proximity to landmarks (relative location) can increase a

ASONAM '19, August 27–30, 2019, Vancouver, BC, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6868-1/19/08...$15.00
https://doi.org/10.1145/3341161.3345334

restaurant's success by at most 72%. We also observe that parameterized location influences success of restaurants the most compared to relative location.

The rest of this paper is organized as follows. Section II includes relevant research works. Section III summarizes the methodology, restaurant success metric and datasets. In Section IV, we define metrics to evaluate an absolute location, and provide results of correlations between restaurant success and location metric values. In Section V, we show the impact of relative location on restaurants' success. Finally, Section VI concludes the paper and suggest future work.

II. Related Work

In this section, we briefly overview different streams of studies related to predicting business reviews and ratings based on business services [4], features [3], [5]–[7], and the location of business [1], [2]. Researchers have used features *e.g.,* degree centrality and clustering coefficient, derived from graph model of user rating and business category data in Yelp, to predict user ratings [7]. Using a Yelp dataset, Lu *et al.* [5] have shown that non-text features, such as review star loss and return guest count, are more significant than text (*e.g.,* unigram, bigram) features to predict the future success (star rating) of restaurants. None of these have investigated the impact of location on business success.

More relevant to our work, Hu *et al.* [8] used matrix factorization model and evaluated the influence of geographical neighbors to business ratings prediction. Another study showed that time dependent features, such as the number of days since the first and last review, provide the best result to infer future attention (number of reviews) of businesses [6]. Wang *et al.* [1] proposed a solution to find an optimal location from some candidate locations for expanding a chain restaurant business to a new branch. They found that market attractiveness and competitiveness features which are generated solely from user review text are more accurate than geographic features to predict the number of likely visits to a restaurant in a prospective location. Eravci *et al.* [2] performed experiments using check-in data for New York from Foursquare and presented two solutions: Bayesian inference-based and collaborative filtering using neighborhood similarity to recommend a set of city neighborhoods as venues to successfully invest in a new specific business category. In another study [3], researchers proposed an approach to label popular and unpopular businesses based on region-wise popularity metrics. They also demonstrated the influence of local and foreign customers on the popularity of businesses and proposed a model to predict popularity of businesses with an accuracy of 89%. Hegde *et al.* [4] identified restaurant properties such as "accept credit card", "ambience" to be the most interesting to customers, found Monday as the most crowded day of the week for restaurants, and explored other restaurant properties (*e.g.,* Wi-Fi, parking lot) that are missing in nearest restaurants to help setup a new restaurant business.

These studies, while investigating the impact of location on business success in general, they mostly focus on recommending venues to investors for setting up new businesses utilizing multiple features including the location. However, and to the best of our knowledge, this paper is the first to investigate and quantify the impact of location, solely, on the success of a business or its failure.

III. Methodology & Datasets

In this section, we briefly describe our methodology, define our restaurant success evaluation metric, and introduce the two datasets: *Restaurant*, *RD*, and *Location*, *LD*, used in our study.

A. Methodology

We adopt two different methodologies to empirically quantify the impact of a given location (parameterized or relative) on the success of a restaurant. For the parameterized location based experiments, our methodology consists of finding a zip-by-zip[1] correlation between the restaurant success and the location characteristic parameters. We are mainly interested in the interrelationship of restaurant success and its location parameters rather than the corresponding cause and effect relationship. In the relative location based experiments, we explore the variation of average success of all restaurants when the distance to other restaurants or landmarks increases.

Parameterized Location based Experiments:

In this experiment, we first define parameters illustrated in subsection IV-A to evaluate location characteristics. Second, we make use of a Location dataset, *LD* outlined in subsection III-D to calculate the parameter values for each zip code. Finally, we utilize the values of a location parameter for each zip and restaurants' average success values for those zips to compute standard correlations. We conduct this experiment per state for different categories of restaurants and every location parameter separately.

Relative Location based Experiments: In this experiment, we build a restaurant vicinity graph depicted in subsection V-A to find restaurants located at different distances to some landmarks or other types of restaurants. Then, we plot the average success of restaurants as a function of the maximum distance to a landmark (in V-B), minimum number of nearby fast-food franchises (in V-C), and same restaurant category density (in V-D). We performed this experiment for specific type restaurants (*e.g.,* cuisine), located in a city for different states. For this experiment, we did not consider zip codes per state as we did in the first experiment.

B. Restaurant Success Metric

Online user reviews and star ratings are crucial factors determining the reputation or success of restaurants. While business success metrics may include income, service evaluation, awards, etc., we focus solely on the social media success, limited to reviews and ratings. Because customers decision to visit a restaurant heavily depends on how many positive reviews and star rating that restaurant has.

Following our restaurant business success definition, we will quantify the success of a given restaurant based on both: (i)

[1] Locations are represented by their corresponding zip codes

its star rating and (ii) the users review counts it received. We, therefore, define $SB(r)$, the success metric of a given restaurant business r using the following equation:

$$SB(r) = S(r)^{\log(RC(r))}, \tag{1}$$

where $S(r)$ is r's average star rating and $RC(r)$ is the number of reviews users gave to restaurant r. For each restaurant r, we extract $S(r)$ and $RC(r)$ values from our *RD* dataset and compute the restaurant success metric ($SB(r)$).

C. Restaurant Dataset

We collected our Restaurant dataset, *RD*, from Yelp Dataset Challenge 2018 [9]. The Yelp data contains information of total 174,567 businesses, 5,261,669 reviews and 1,326,101 users. Each business in this dataset has unique identifier, name, address, city, state, GPS coordinate. As attribute information, a business also has categories (such as Restaurants, Hotels, Shopping etc.), average rating, review count, and review text associated with it. Since in this work, we are focusing on the success of restaurant businesses only, we filtered out all non-restaurant businesses and compiled our *RD* dataset with restaurants from three different states of the United States: Arizona (AZ), North Carolina (NC), and Ohio (OH), which are among the top five states with large number of businesses in the Yelp data.

Our *RD* dataset consists of two types of restaurants: *cuisine* and *fast-food* franchise. Cuisine restaurants are businesses that serve a well known cooking style or quality of food from a given region or country. Examples include *Chinese* and *Mexican* cuisines. Fast-food franchise restaurants are part of a chain operation that serves partially prepared food and has minimal table service (*e.g.,* Subway).

We consider four different categories of cuisine restaurants: *American* (*Am*), *Italian* (*It*), *Mexican* (*Mx*), and *Chinese* (*Ch*) for our *RD* dataset. We chose restaurants from these cuisines based on the business category attribute values found in the Yelp data. Thus, if a restaurant business category includes the keyword "Mexican", we classify it as Mexican cuisine restaurant. For American cuisines, we used the keywords "American Traditional" and "American New", for Chinese cuisine we used the keyword "Chinese", and for Italian cuisines we looked for the coexistence of "Pizza" and "Pasta Shops" keywords in addition to the keyword "Italian". We use St_C to represent the restaurants of cuisine type C from state St. For example, AZ_{Mx} denotes the set of all Mexican cuisine restaurants of state Arizona.

We also choose top 10 fast food franchise restaurants [10] in the United States for our *RD* dataset based on the business name attribute values from the Yelp data. For example, if a restaurant business has the name "McDonald's" (*McD*) or "Subway" (*Sub*), we classify them as fast-food franchise restaurants. This left us with total 8,954 restaurants in the *RD* dataset. In Table I, we summarize the number of both cuisine and fast-food franchise restaurants (state wise) in our *RD* dataset.

TABLE I: Number of both cuisine and fast food franchise restaurants for the three states in our *RD* dataset

State	**Cuisine**				**Fast Food Franchise**		
	Am	*Ch*	*Mx*	*It*	*McD*	*Sub*	*Others*
AZ	1,159	528	1,279	567	158	248	508
NC	550	241	259	167	62	52	204
OH	664	277	208	253	93	55	262
Total	**3,533**	**1,046**	**1,746**	**987**	**313**	**355**	**974**

D. Location Dataset

We constructed the Location dataset, *LD* by crawling the City-Data website[2] which contains demographics and tourism related information pertaining to United States cities and regions. Our *LD* dataset includes location-based information such as race, income, education, housing, transportation, crime etc. Data is gathered for different zip code locations from the three states mentioned earlier. We gathered data for up to 141 zip codes in AZ, 64 zip codes in NC, and 107 zip codes in OH.

IV. Effect of Absolute Location on Restaurant Success

In this section, we will introduce several parameters to evaluate parameterized location of a restaurant and use the *LD* dataset to calculate the values of these parameters. Then, we will discuss result of correlations between the values of restaurants' success and location parameters.

A. Parameterized Location Categories and Parameters

Different categories of parameterized location such as population demographics, tourism, accessibility, safety might have a profound impact on the rise or fall of a restaurant business [11]. Note that our lowest location granularity is the zip code as per our Location dataset. However, we argue that the methodology taken in this paper applies for a finer granularity, such as streets or full addresses.

Our analysis focus on four main location parameter categories, namely (i) the *living standard*, (ii) the *tourism significance*, (iii) the *business convenience*, and (iv) the *safeness* of a given location. Next we introduce these four categories in more detail and propose few location parameters per category.

1) Living Standard: The living standard of a location can be determined by looking at the job prospects, the trend in education, the cost of living and the housing affordability of that location. We present four living standard parameters, namely *Education Index*, *Housing Affordability Index*, *Cost of Living Index*, and *Life Style*.

Education Index: To measure job prospects and education standard in a zip code location z, we multiply the weighted average of the percentage of people educated in different levels such as high school, graduate, by the percentage of people employed in z. We name this parameter EI^z as defined by the following equation.

$$EI^z = (u_1 \times G^z + u_2 \times B^z + u_3 \times H^z) \times E^z \tag{2}$$

[2] www.city-data.com

where, u_1, u_2, and u_3 are weights such that $\sum u_i = 1$; H^z, B^z, and G^z are the percentage of people educated up to High school or higher, educated up to Bachelor's degree or higher, and educated up to Graduate or professional degree respectively; and E^z is the percentage of people employed. We choose $u_i \geq u_{i+1}$ as we assume that the higher the degree of education of a person, the higher his/her income thus the restaurant affordability. In our experiments, $u_1 = 0.5, u_2 = 0.3,$ and $u_3 = 0.2$.

Housing Affordability Index: It is a parameter used by the US National Association of Realtors (NAR) to assess a typical middle income family's qualification for a mortgage loan on a median-priced home [12]. This parameter is a ratio between the median household income and the annual qualifying income needed to own a median-priced home. To calculate the qualifying income, NAR assumes that the homeowner uses no more than 25 percent of monthly household income for the mortgage payments. We use this parameter as an indicator for the living standard in a zip code location, z. We denote this parameter by HAI^z as defined by the following equation.

$$HAI^z = \frac{M^z}{MP^z \times 12 \times 4} \times 100, \tag{3}$$

where M^z is median household income, and MP^z is median monthly payment for housing units with a mortgage.

Cost of Living Index: This index [13] is often used to quantify the expenses to live in a given area. We use Cost of Living Index, CLI^z as another parameter to estimate the living standard of a location z.

Life Style: We also combine the parameters mentioned above to produce a new metric and name it *Life Style*, LS^z. It is defined by the following equation.

$$LS^z = \frac{EI^z \times HAI^z}{CLI^z}, \tag{4}$$

2) Tourism Significance: Tourism significance expresses how compelling a location is in terms of attracting tourists. Tourists often visit areas such as outdoor scenery (*e.g.,* landscape, wildlife), historic or cultural venues, and recreation facilities. Next, we present two tourism related parameters, namely *Tourism Attraction Density* and *Neighboring Tourism Attractions*.

Tourism Attraction Density: We argue that the number of touristic sights or attractions in a given location can be a good indicator for its tourism significance. Thus, we define the tourism attraction density parameter of a location z as:

$$TD^z = \frac{e^{TC^z}}{Area(z)}, \tag{5}$$

where, $Area(z)$ is the land area of location z, and TC^z is the weighted sum of touristic place counts of location z, defined by the following.

$$TC^z = v_1 \times (NP^z + BE^z) + v_2 \times TL^z + v_3 \times RV^z + v_4 \times LR^z + v_5 \times PK^z, \tag{6}$$

where NP^z is the number of National Parks, BE^z is the number of Beaches, TL^z is the number of Tourist Locations, RV^z is the number of Rivers, Streams or Creeks, LR^z is the number of Lakes and Reservoirs, PK^z is the number of Parks; v_i are weights such that $\sum v_i = 1$. We choose $v_i \geq v_{i+1}$ to give more weight to national parks and beaches compared to other tourism attractions such as rivers. In our experiments, $v_1 = 0.3, v_2 = 0.2,$ and $v_3 = v_4 = v_5 = 0.1$.

Neighboring Tourism Attractions: We argue that tourists may visit touristic attractions in a given location z and walk to neighboring locations nz for a meal. Thus, measuring the tourism attraction density only in the current location z may not be sufficient and can be augmented by looking at neighboring location, nz's attraction density as well.

$$NT^z = \alpha TD^z + (1-\alpha)\frac{\sum_{nz} TD^{nz}}{t_z}, \tag{7}$$

where, TD^{nz} is the tourism attraction density of nz; t_z is the total number of neighboring locations for z and α is the influence factor between z and it's neighboring locations such that $\alpha > 0.5$. In our experiments, we let $\alpha = 0.8$.

3) Business Convenience: We refer to business convenience of a given location z, as the relative ease of access to z. While there maybe multiple factors to determine the business convenience of a given restaurant r, we consider only two: the presence of shopping malls and the availability of public transportation. Next, we present two business convenience related parameters, namely *Business Accessibility* and *Neighboring Business Accessibility*.

Business Accessibility: To measure the business accessibility of z, we define the BA^z parameter by the following equation.

$$BA^z = \frac{e^{NS} \times TA^z}{Area(z)}, \tag{8}$$

where NS is the number of shopping centers, $Area(z)$ is the land area of z, and TA^z is the weighted average of the percentage of people using different transportation in location z, defined by the following.

$$TA^z = w_1 \times BR^z + w_2 \times CT^z + w_3 \times CM^z + w_4 \times WB^z, \tag{9}$$

where BR^z is the percentage of people using subway or bus or railroad, CT^z is the percentage of people using carpool or taxi, CM^z is the percentage of people using car or motorcycle, WB^z is the percentage of people who walks or use bicycle; w_i are weights such that $\sum w_i = 1$. We choose $w_i \geq w_{i+1}$ to give more weight to public transportation convenience compared to other means that can be costly and/or inconvenient (*e.g.,* finding parking). In our experiments, $w_1 = 0.7,$ and $w_2 = w_3 = w_4 = 0.1$.

Neighboring Business Accessibility: We argue that people may visit neighboring shopping malls while shopping from a center in a given location z. Thus, we can augment the business accessibility of z with the business accessibility of its neighboring location, nz using the following equation.

$$NB^z = \beta BA^z + (1-\beta)\frac{\sum_{nz} BA^{nz}}{t_z}, \tag{10}$$

where, BA^z is the business accessibility of z, BA^{nz} is the business accessibility of the neighboring location nz of z, t_z is the total number of neighboring locations for z and β is the influence factor between z and it's neighboring locations such that $\beta > 0.5$. We choose $\beta = 0.7$.

4) Safeness: People prefer to live and travel in a safe area where they can escape their fear of being victimized by violent crimes. Safeness of a location expresses how safe a location is in terms of crime rates. To estimate the safeness of a location z, we define the following parameter.

$$SM^z = \frac{1}{CI^z}, \tag{11}$$

where CI^z is the crime index based on the crime rate per 1,000 population for all crimes in a specific location.

5) Combined Parameters: Motivated by the success of all these parameters, we proposed a combination metric that aims at merging all these parameters. In order to choose the best combination we have performed roughly 100 experiments with various combinations and weights for each parameter. Based on these empirical experiments, we have selected the best combination which is represented by:

$$CA^z = 0.6LS^z + 0.2NB^z + 0.1NT^z + 0.1SM^z \tag{12}$$

B. Correlation Metrics

Our objective is to measure the degree of statistical association between the restaurants' success values (measured by both star rating and number of reviews) and the parameter values of the areas where these restaurants are located. To do that, we first create two vectors: $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_n)$, where x_i represents the value of a location parameter in the i-th zip code, y_i may represent the value of the success metric of restaurant located in the i-th zip code, and n, the size of both vectors X and Y, is the number of zip codes in a given state. Note that, since multiple restaurant may be co-located in the same zip code location, we consider the average restaurant success in location i (*i.e.,* which represent a given zip code z_i) as y_i.

In our experiments, we are interested in the ranks of the data points rather than the value of the data points. Therefore, in our experiments we adopt two rank-order based correlation metrics, Spearman's correlation coefficient [14] and Kendall's correlation coefficient [15] to measure the correlation between location parameter and restaurant success metric values.

1) Spearman's correlation: The Spearman's correlation coefficient can detect the strength and direction of the monotonic relationship between two variables. The idea is to first rank the data points of each variable separately. Ranking is obtained by assigning a rank of 1 to the lowest value, 2 to the next lowest and so on. Then for each pair (x_i, y_i) the difference between ranks of x_i and y_i are calculated. If the variables are correlated, then the sum of the squared difference between ranks of the data points will be small. The magnitude of the sum is related to the significance of the correlation. For example, if we consider LS^z parameter as X variable and success metric SB as Y variable, then from the correlation coefficient of the values of X and Y, we can infer that how strongly living standard of location affects the success of restaurants. The Spearman's correlation coefficient, ρ is computed according to the following equation:

$$\rho = 1 - \frac{6 - \sum^n {d_i}^2}{n(n^2-1)}, \tag{13}$$

where d_i is the difference between ranks for each (x_i, y_i) data pair and n is the number of data pairs.

2) Kendall's correlation: Unlike Spearman's correlation coefficient, Kendall's correlation does not take into account the difference between ranks, only directional agreement. The definition of Kendall's correlation relies on the notion of concordance. Two pairs of observations (x_i, y_i) and (x_j, y_j) where $i < j$, are called concordant pairs if the ranks (the sort order by x and y) for both elements agree. That is, if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. The pairs are called discordant, if $x_i > x_j$ and $y_i < y_j$; or if $x_i < x_j$ and $y_i > y_j$. The equation to compute the Kendall's correlation coefficient, τ, is as follows:

$$\tau = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \tag{14}$$

where, $n_0 = n(n-1)/2$, n is the size of the X and Y, $n_1 = \sum_i t_i(t_i - 1)/2$, $n_2 = \sum_j u_j(u_j - 1)/2$, t_i is the number of tied values in the i-th group of ties for X variable, u_j is the number of tied values in the j-th group of ties for Y variable, n_c is the number of concordant pairs, and n_d is the number of discordant pairs.

The coefficient value varies between -1 and 1. A coefficient of 1 denotes that the rankings of the two variables are in perfect agreement. A coefficient equal to -1 signifies that one ranking is the opposite of the other. If the coefficient is 0, then X and Y are independent variables.

C. Correlation Results

We created X and Y vectors for each state using all its zip codes that exist in our *LD* dataset. We removed the zip codes that did not have any restaurants based on our *RD* dataset. We computed both Spearman's and Kendall's correlation coefficient by choosing values of each location parameter separately to build X. We present the Spearman's and Kendall's correlation coefficient values between average restaurant success values (SB) and location parameter values for cuisine restaurants of three states in Table II. Note that we have shown only the cuisine categories from each state such that for at least 75% zip codes z in each state, there exists at least one restaurant r in z.

In Table II, each cell displays a correlation coefficient between values of a location parameter for all zip codes in a particular state and average values of success metric for all restaurants of particular cuisine located in those zip codes. For example, the Spearman's correlation coefficient between values of EI^z parameter for all zip codes in NC state and average success (SB) values of all "American" cuisine restaurants located in those zip codes is 0.57.

TABLE II: Correlation results of all three states for different categories of cuisine restaurants

Category	Parameter	NC_{Am}		NC_{Ch}		OH_{Am}		OH_{Ch}		AZ_{Mx}		AZ_{Am}		AZ_{Ch}		Avg		Max	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Living St.	EI^z	0.57	0.40	0.73	0.54	0.41	0.27	0.49	0.33	0.51	0.37	0.44	0.30	0.38	0.27	0.50	0.36	0.73	0.54
	LS^z	0.49	0.34	0.70	0.51	0.27	0.18	0.41	0.28	0.44	0.30	0.33	0.22	0.45	0.31	0.44	0.30	0.70	0.51
Tourism Sig.	TD^z	0.47	0.33	0.49	0.34	0.23	0.15	0.00	-0.004	0.11	0.07	0.20	0.13	-0.16	-0.11	0.19	0.13	0.49	0.34
	NT^z	0.43	0.29	0.44	0.30	0.23	0.16	-0.02	-0.01	0.09	0.07	0.20	0.14	-0.16	-0.11	0.17	0.12	0.44	0.30
Business Con.	BA^z	0.37	0.27	0.17	0.12	0.21	0.14	0.21	0.14	0.09	0.06	0.21	0.13	-0.15	-0.10	0.16	0.11	0.37	0.27
	NB^z	0.22	0.15	-0.01	0.03	0.21	0.14	0.20	0.13	0.17	0.12	0.26	0.18	0.04	0.03	0.16	0.11	0.26	0.18
Safeness	SM^z	0.43	0.33	0.27	0.18	0.01	0.01	0.11	0.07	0.18	0.14	0.21	0.15	0.36	0.28	0.23	0.17	0.43	0.33
Combined	CA^z	0.72	0.57	0.81	0.62	0.24	0.16	0.43	0.29	0.45	0.31	0.41	0.28	0.50	0.35	0.51	0.37	0.81	0.62
	Avg	0.46	0.33	0.45	0.33	0.23	0.15	0.23	0.15	0.26	0.18	0.28	0.19	0.16	0.11	0.30	0.21	0.53	0.38
	Max	0.72	0.57	0.81	0.62	0.41	0.27	0.49	0.33	0.51	0.37	0.44	0.30	0.50	0.35	0.51	0.37	0.81	0.62

From Table II, we find that among four different categories, parameters from living standard category achieved the highest average $\rho = 0.50, \tau = 0.36$ (EI^z) and $\rho = 0.44, \tau = 0.30$ (LS^z) across all three states and different cuisine categories. This implies that living standard of a location has the most effect on the success of restaurants. Because people, residing in locations having greater job prospects, better education, and low living cost, can frequently dine out in restaurants and provide positive reviews. On the other hand, location parameters BA^z and NB^z from business convenience category obtained the lowest average correlation coefficients across all three states and different cuisine categories. This may be because people don't care about the public transportation when they are visiting restaurants and giving reviews. We also observe that education index, EI metric obtained high values consistently compared to other location metrics for different cuisine restaurants across all three states. Thus, we can infer that highly educated people visits restaurants more compared to people with low education level and like to provide favorable reviews and ratings, which contribute greatly to the success of the restaurants.

Among all location parameters, the combination of LS^z, NT^z, NB^z, and SM^z parameters, CA^z has achieved the highest average Spearman's $\rho = 0.51$ and Kendall's $\tau = 0.37$ as shown in Table II. In our experiments, we found LS^z parameter to be the most contributing parameter towards obtaining the highest value for CA^z. This suggests that the living cost, education level and housing affordability of the people in a location have the most influence on the success of restaurants.

We also note that NC state has the highest $\rho = 0.81$ and $\tau = 0.62$ for CA^z parameter across all cuisine categories compared to OH and AZ states. One of the reasons is AZ and OH state has 68 and 46 more zip code locations respectively than NC state. As a result, there is large variation in parameter values and restaurant success values for AZ and OH state compared to NC state and it causes the correlation values of both of these states to go lower.

From the findings explained above we can conclude that parameterized location as captured by the defined parameters indeed has vital effect on the success of restaurants. Now, in addition to the parameterized location of a restaurant, we will also find out whether the relative location of a restaurant affects its success or failure in the next section.

V. Effect of Relative Location on Restaurant Success

In this section, we will further investigate the impact of nearby popular landmarks, fast-food franchise restaurants, and competing restaurants on the success of cuisine restaurants. For this experiment, we first build a graph to retrieve the restaurants located within specific distances to landmarks, franchises or other restaurants.

A. Restaurant Vicinity Graph

We construct a restaurant vicinity graph, an undirected weighted graph $G_c = (V_c, E_c)$ consisting of restaurants $r \in \mathcal{R}_c$ and landmarks $l \in \mathcal{L}_c$ for a given city c, such that $V_c = \mathcal{R}_c \cup \mathcal{L}_c$. For any two nodes $\{p, q\} \in V_c$, $e_{pq} \in E_q$ denotes the weighted link connecting p and q. Edges $e_{pq} \in E_q$ have the Euclidean distance between p's and q's locations as e_{pq}'s weight, which we refer to as W_{pq} as shown in Figure 1. We measure weights as per the Haversine function [16] using the latitudes and longitudes of the nodes (restaurants and landmarks).

Restaurant nodes have three attributes, Name denoting the name and the address of the restaurant, Type which takes cuisine or franchise as values to denote cuisine and fast-food restaurants respectively, and the Category to indicate the food cuisine type, *e.g.,* Mexican, Italian.

Figure 1 shows an instance of graph G_c in a city c. In this graph, four nodes consisting of two restaurants named r_1 and r_2 with same cuisine category "Mexican", one American fast-food franchise restaurant, r_3, and a landmark l_1 interconnected with weight distances such as $W_{r_1 r_2}$ which connects restaurants r_1 and r_2.

B. Impact of Proximity to Popular Landmarks

Landmarks of a city (*e.g.,* the Eiffel Tower in Paris, or the Millennium Park in Chicago) serve as popular gathering place which attracts residents and visitors of the city. Being able to tour the landmarks as well as have lunch or dinner at restaurants in the vicinity of those landmarks, may motivate the visitors and city dwellers to provide favorable reviews and ratings for the restaurants.

We choose one popular landmark from each city. For example, we picked "Her Secret is Patience" as landmark from

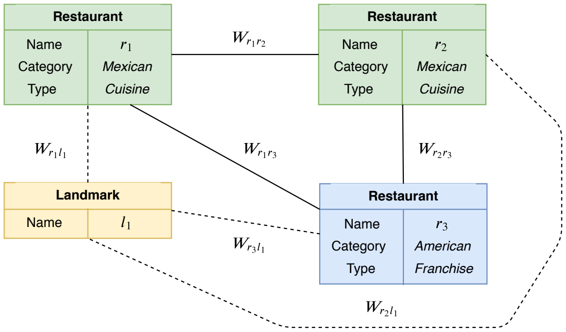


Fig. 1: Example restaurant vicinity graph with two cuisine restaurants, one fast food restaurant, and one landmark

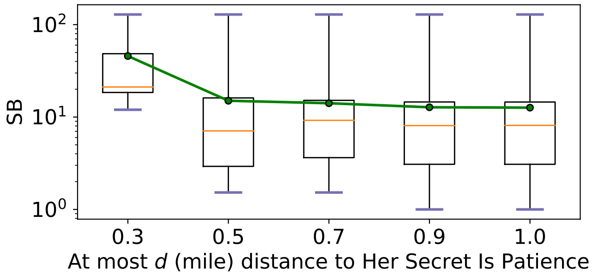


Fig. 2: Distribution of success of restaurants located at most d distance from "Her Secret is Patience" landmark in Phoenix

the city of Phoenix, "Bechtler Museum of Modern Art" as landmark from city of Charlotte, and "Soldiers and Sailors Monument" as landmark from city of Cleveland.

In Figure 2, we display the distribution of success values of all four categories of cuisine restaurants that are located at most d mile distances to landmark "Her Secret is Patience". In Figure 2, the line that divides each box marks the median success, the lower and upper whiskers are minimum and maximum success respectively and green data points represent the average SB values. For example, in the second box from left, the average SB is approximately 10.5 for all cuisine restaurants that are located within 0.5 mile distance to "Her Secret is Patience". From Figure 2, we note that as the distance to the landmark increases, average SB values for restaurants decreases. As shown in Figure 2, the average SB is 25% lower for cuisine restaurants that are located at most 1 mile of the "Her Secret is Patience" compared to those that are within 0.3 mile distance.

In Figure 3, we plot the average restaurant success SB as a function of the maximum distance to a landmark in all three cities. From Figure 3, we notice that as the distance to the landmarks increases, the restaurant success decreases. We also find that the average SB value of cuisine restaurants of Phoenix that are located at most 0.3 mile distance to the "Her Secret is Patience" landmark gains up to 83% compared to that of cuisine restaurants of Charlotte and Cleveland respectively.

Summary: Proximity to city landmarks increases the potential success of restaurants up to 72% as verified in the city of Phoenix.

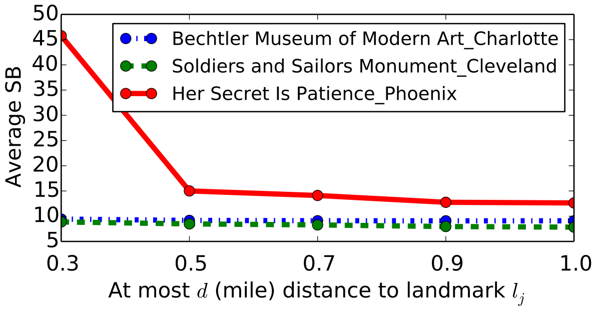


Fig. 3: Average success of restaurants located at most d distance from landmarks of Phoenix, Charlotte, and Cleveland

C. Impact of Proximity to Fast-Food Franchises

Cuisine restaurants that are in proximity of fast food franchise restaurants often gain easy visibility and draw customers. We argue that customers may get tempted to eat in a given cuisine restaurant if they have to choose between it and a neighboring fast-food restaurant. This can increase the number of check-ins, reviews, and potential number of stars given by these customers, as they compare the food and the service to the fast food option.

In Figure 4, we plot the average restaurant success SB as a function of the minimum number of neighboring fast-food restaurants (*i.e.,* at least k neighboring restaurants), where neighboring restaurants are defined by a distance radius of 1, 1.5, and 2 miles for the city of Phoenix, Cleveland, and Charlotte. We show that when the number of neighboring fast-food restaurants increases, the restaurant success increases. In fact, as shown in Figure 4a, the average SB of cuisine restaurants that are located near at least 4 fast-food franchises is up to 4% higher than the total average of SB for all restaurants (*i.e.,* at least zero neighboring fast food). We observe similar trend in Figure 4b and 4c for the city of Cleveland and Charlotte respectively. We also note that the increase in restaurant success is significant for dense fast food areas (*i.e.,* at least 4) such in malls, downtown, etc.

Summary: Proximity to one or many fast-food franchise restaurants can make cuisine restaurants up to 16% more successful as verified in the city of Cleveland.

D. Impact of Same Restaurant Category Density

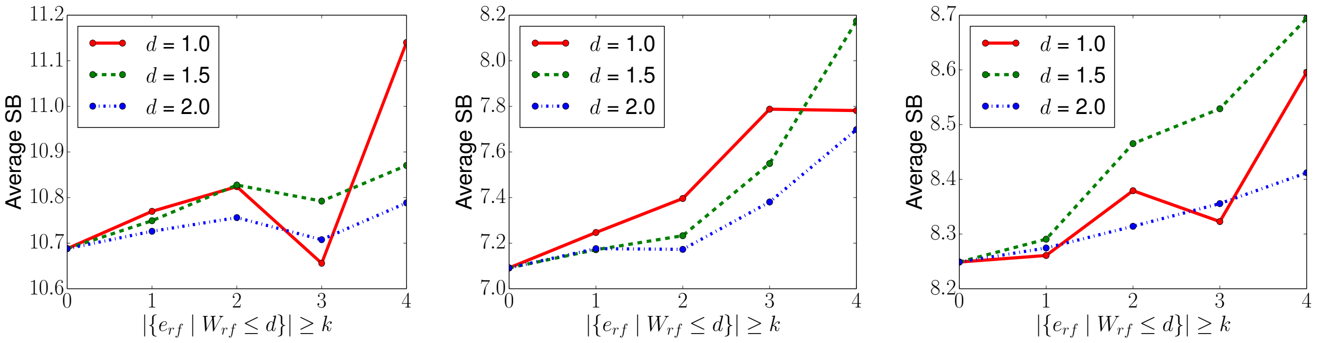
We argue that proximity to same or similar restaurants increases the competition among them which drive better service and food quality resulting in higher customer satisfaction.

To assess this intuition, we define the competing score metric as a function of the density of same category restaurant within a distance radius, d.

$$CP(r,d) = \frac{|SC(r,d)|}{|AC(r,d)|}, \tag{15}$$

where $SC(r,d) = \{r_j \mid r_j$ is the same cuisine category as restaurant r and $W_{r_jr} \leq d\}$, and $AC(r,d) = \{r_j \mid W_{r_jr} \leq d\}$.

In Figure 5, we plot the average success SB as a function of maximum competing score of the cuisine restaurants for



(a) Average success of cuisine restaurants (r) within d miles of at least k number of Franchises (f) in Phoenix

(b) Average success of cuisine restaurants (r) within d miles of at least k number of Franchises (f) in Cleveland

(c) Average success of cuisine restaurants (r) within d miles of at least k number of Franchises (f) in Charlotte

Fig. 4: Average success of cuisine restaurants located at different distances to at least k number of Franchises in Phoenix, Cleveland, and Charlotte

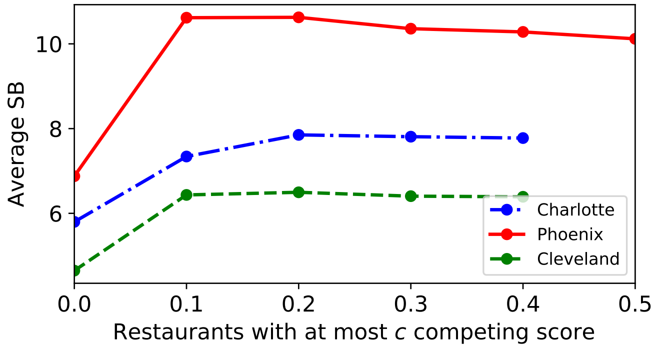


Fig. 5: Average success of restaurants with at most c competing score in city Phoenix, Charlotte, and Cleveland

the city of Phoenix, Cleveland and Charlotte. We show that often with the increase of competition, the restaurant success increases. In fact, as shown in Figure 5, we observe that the average SB value of cuisine restaurants is higher by 53.62% in Phoenix, by 34.04% in Cleveland, by 25.86% in Charlotte when competing score is 0.1, compared to the case when computing score is 0.0.

Summary: Competition can increase restaurants' success up to 53.62% as verified in the city of Phoenix.

VI. Conclusion

To the best of our knowledge, this is the first study to quantify the impact of location on the success of restaurants. We empirically show that when a location is considered singularly, different determinant categories such as demographics, tourism, business convenience, safeness can have positive impact on restaurant success. The success is particularly higher when the restaurant caters to a location with a large proportion of educated individuals. We also demonstrate that nearby landmarks, fast food franchises, and competing restaurants affect the success of restaurants. There are other factors such as restaurant's ambience and service quality which can also influence the success of a restaurant in addition to its location. In future work, we will try to quantify the impact of these factors on the success of restaurants.

References

[1] F. Wang, L. Chen, and W. Pan, "Where to place your next restaurant?: Optimal restaurant placement via leveraging user-generated reviews," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 2371–2376.

[2] B. Eravci, N. Bulut, C. Etemoglu, and H. Ferhatosmanoğlu, "Location recommendations for new businesses using check-in data," in *Proceedings of the 16th IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2016, pp. 1110–1117.

[3] A. K. Bhowmick, S. Suman, and B. Mitra, "Effect of information propagation on business popularity: A case study on yelp," in *Proceedings of the 18th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2017, pp. 11–20.

[4] S. Hegde, S. Satyappanavar, and S. Setty, "Restaurant setup business analysis using yelp dataset," in *Proceedings of the 6th International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 2342–2348.

[5] X. Lu, J. Qu, Y. Jiang, and Y. Zhao, "Should i invest it?: Predicting future success of yelp restaurants," in *Proceedings of the Practice and Experience on Advanced Research Computing*. ACM, 2018, p. 64.

[6] B. Hood, V. Hwang, and J. King, "Inferring future business attention," *Yelp Challenge, Carnegie Mellon University*, 2013.

[7] A. Tiroshi, S. Berkovsky, M. A. Kaafar, D. Vallet, T. Chen, and T. Kuflik, "Improving business rating predictions using graph based features," in *Proceedings of the 19th International Conference on Intelligent User Interfaces*. ACM, 2014, pp. 17–26.

[8] L. Hu, A. Sun, and Y. Liu, "Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 2014, pp. 345–354.

[9] "Yelp Dataset challenge," https://www.yelp.com/dataset/challenge, accessed: 11-15-2018.

[10] "Business-Insider biggest fast-food chains in america," https://www.businessinsider.com/biggest-fast-food-chains-in-america-2018-6#6-wendys-15, accessed: 11-20-2018.

[11] H. Parsa, A. Gregory, M. Terry *et al.*, "Why do restaurants fail? part iii: An analysis of macro and micro factors," *Institute for Tourism Studies*, 2011.

[12] "HAI: housing affordability index," https://www.nar.realtor/research-and-statistics/housing-statistics/housing-affordability-index/methodology, accessed: 11-15-2018.

[13] "Cost of Living index," https://worldwidecostofliving.com/asp/wcol_HelpIndexCalc.asp, accessed: 11-15-2018.

[14] E. C. Fieller, H. O. Hartley, and E. S. Pearson, "Tests for rank correlation coefficients. i," *Biometrika*, vol. 44, no. 3/4, pp. 470–481, 1957.

[15] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938. [Online]. Available: http://dx.doi.org/10.1093/biomet/30.1-2.81

[16] F. Cajori, *A History of Mathematical Notations: Vol. II*, ser. A History of Mathematical Notations. Cosimo, Incorporated, 2007, no. v. 2. [Online]. Available: https://books.google.com/books?id=bT5suOONXlgC