

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/318138987>

# Does Yelp Matter? Analyzing (And Guide to Using) Ratings for a Quick Serve Restaurant Chain

Chapter · May 2018

DOI: 10.1007/978-3-319-53817-4\_19

---

CITATION

1

---

READS

1,860

2 authors, including:



[Jennifer Priestley](#)

Kennesaw State University

63 PUBLICATIONS 537 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Sacrocolpopexy sexual function [View project](#)



Data Ethics [View project](#)

# **Does Yelp Matter? Analyzing (And Guide to Using) Ratings for a Quick Serve Restaurant Chain**

Bogdan Gadidov\*

Kennesaw State University, Kennesaw, GA, 30144, U.S.A.

[bgadidov@kennesaw.edu](mailto:bgadidov@kennesaw.edu)

Jennifer Lewis Priestley, Ph.D.

Kennesaw State University, Kennesaw, GA, 30144, U.S.A.

[jpriestl@kennesaw.edu](mailto:jpriestl@kennesaw.edu)

\*corresponding author

## **Abstract**

In this paper, we perform an analysis of reviews for a national quick serve (fast food) restaurant chain. Results show that the company-owned restaurants consistently perform better than franchised restaurants in numeric rating (1-5 stars) in states which contain both types of operations. Using sales data, correlations are used to evaluate the relationship between the number of guests or sales of a restaurant and the rating of the restaurant. We found positive correlations were present between the number of customers at a location and the numeric rating. No correlation was found between the average ticket size and numeric rating. The study also found that 5-star rated restaurants have frequent comments related to the cleanliness and friendliness of the staff whereas 1-star rated restaurants have comments more closely related to speed of service and temperature of food. Overall, the study found that in contrast to previous research, rating sites like Yelp are relevant in the quick serve restaurant sector and reviews can be used to inform operational decisions, leading to improved performance. Detailed explanations related to the process of extracting this data and relevant code are provided for future researchers interested in analyzing Yelp reviews.

*Key Words: Yelp Reviews, Word Clouds, Quick Serve Restaurants, R, SAS, Correlation*

# 1. Introduction

Many businesses, including restaurants, have access to data generated every day from customers on sites like Yelp, but do not take advantage of the data. This has been true for two main reasons. First, the rise of social media data is a recent phenomenon and the tools, skills and technology available to translate this data into meaningful information is evolving, but is still relatively nascent. Second, previous research has indicated that reviews of restaurants in the lowest price points have limited relevance. We challenge this premise.

In this chapter, we seek to answer the question - *Does Yelp Matter in the Quick Serve Restaurants Sector?* Within the context of this study, we also explore operational performance differences between company-owned and franchised outlets, the most frequently used terms associated with 5-star restaurants versus 1-star restaurants and provide future researchers with a guide on how to use the R Programming language to extract reviews for further analysis.

## 2. Literature Review

### 2.1. The Rise of Social Media Data

The Gartner Group defines the concept of “dark data” as “*the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes...*” (Gartner, 2016). The term is derived conceptually from dark matter in physics – matter which is known to exist but cannot be experienced directly.

Dark data has historically not been recognized as having value because (a) it was not viewed as data (e.g. security video files), (b) it was recognized as data, but too unstructured and therefore too difficult to translate into meaningful information (e.g., text narratives from surveys) or (c) it was truly “dark” and the target organization was not aware of its existence. Until recently, data derived from social media outlets met all of these conditions.

Because successful utilization of analytics to improve the decision making process is limited to the data available (Halevy, et al., 2009), the information embedded in dark data can be massively valuable. Researchers are now able to leverage new and evolving analytical techniques (e.g., machine learning, natural language processing and text mining) and scripting languages (e.g., R, Python) which enable access to and translation of this social media-generated dark data into information. This relatively new phenomena of extracting and leveraging social media analysis for organizational decision making is no longer a marginal “nice to have”, but rather a central informational asset.

Economic sectors across the U.S. economy are increasingly extracting previously “dark data” from outlets such as Yelp and Twitter to inform their decision making. For example, rather than waiting for impressions to be driven by traditional media or advertising, CEOs increasingly use Twitter as a medium through which to have direct communication with their customers –

sometimes millions of them at the same time. They can then receive immediate feedback reflected in responses, retweets and “likes” (Malhotra, et al., 2015).

One sector where social media analysis has become particularly critical is the restaurant sector. This is true because “*Restaurants are a classic example...where the consumer has to make a decision based on very little information*” (Luca, et al., 2011). The largest restaurant review site is Yelp with 135 million monthly visitors, followed by Open Table with 19 million monthly visitors (OpenTable, 2016).

Restaurants, particularly chains of restaurants which engage in national and large-scale regional advertising and product launches, can benefit from the immediacy of customer feedback from sites like Twitter and Yelp.

This chapter will examine the specific role of Yelp reviews for a national quick serve restaurant chain.

## **2.2. The Quick Service Restaurant Sector**

The quick service restaurant sector plays a significant role in the US economy. In 2015, 200,000 fast food restaurants generated revenue of over \$200 billion. Industry estimates indicate that 50 million Americans eat in a quick serve restaurant every single day. This sector also represents an important source of employment across the country – with over 4 million people employed in quick serve restaurant franchises in 2015, and one in three Americans worked in this sector at one point during their lives (Sena, 2016). While the food is often highly processed and prepared in an assembly line fashion, customers of these restaurants have placed value on consistency of service, value for money and speed (National Restaurant Association, 2014).

However, quick service restaurant failures are almost epidemic. Although a relatively modest 26 percent of independent restaurants failed during the first year of operation, quick service restaurants fail at substantively higher rate – failure of franchise chains have been reported to be over 57 percent (Parsa, et al., 2015). Failures of quick serve restaurants – like any small businesses – create negative externalities on local economies in the form of unemployment and lost local spending power.

While this failure rate has been attributed to macro factors like economic growth, federal and state legislation of minimum wage rates, new and different forms of competition, as well as to micro factors like access to capital, location, owner incompetence and inexperience (Gregory, et. al., 2011) only recently has any meaningful attention been paid to the role of online customer reviews (e.g., Hlee, et al., 2016; Taylor et al., 2016; Remmers, 2014).

Although some researchers have indicated that reviews do not matter in the quick service restaurant sector because of the low per ticket price point (Vasa, et al., 2016), this chapter challenges part of the premise of that perspective.

First, in 2016, there were millions of reviews associated with quick serve restaurants on Yelp. This is an indication that some customers are willing to provide feedback related to fast food experiences. Importantly, this is consistent with the point that other researchers have demonstrated that satisfaction with a restaurant experience is strongly related to perceived, rather than to absolute, value for price paid – across the range of price points (e.g., King, 2016; Dwyer, 2015). Second, the study highlighted in this chapter, provides some initial evidence for correlation between the number of guests, sales and numeric ratings on Yelp for a quick serve restaurant. Both of these points provide at least directional evidence that analysis of Yelp reviews for the quick serve restaurant sector could improve operational performance and provide meaningful feedback regarding customer experiences, thereby helping to mitigate the high rate of outlet failure.

### **3. Analysis of Numeric and Text Reviews in Yelp**

The current study examined both numeric and text reviews for over 2,000 locations of a quick serve restaurant chain across the United States. The results associated with the numeric and text results are provided in the following sections.

#### **3.1. Description of Numeric Ratings**

Numeric ratings taken from Yelp can range from 1-5 stars, where 1 star represents the worst possible rating and 5 stars represents the highest possible rating. Ratings are aggregated at the business (restaurant) level, and can be pulled by calling for the specific rating parameter in the code provided. Unfortunately, only the current numeric ratings can be extracted – meaning that researchers cannot extract ratings at specified past periods to ascertain changes in ratings over time.

Using these ratings, comparisons are made at the state level, and the GMAP procedure in SAS is used to create illustrative maps. Furthermore, data is used to compare franchise owned to non-franchise (corporate) owned restaurants. Transaction data is used for the non-franchise locations, to assess whether the numeric ratings of restaurants correlate to the number of sales, number of customers, or total order size. The transaction data set includes over 4.7 million transactions at over 400 non-franchise locations of this restaurant chain from January through June 2015. A sample of the transaction data can be seen in Table 1 below.

Table 1. Sample of Transaction Dataset

DateOfBusiness	StoreID	Order Number	GuestCount	ItemCount	NetSales
1/2/2015	20115	Order #214	1	1	1.99
1/2/2015	20116	Order #215	1	15	26.17
1/2/2015	20117	Order #216	1	5	15.65
1/2/2015	20118	Order #217	1	4	6.29
1/2/2015	20119	Order #218	1	1	1.29
1/2/2015	20120	Order #219	1	1	1
1/2/2015	20121	Order #219	1	1	1.39
1/2/2015	20122	Order #221	1	1	1
1/2/2015	20123	Order #222	1	8	11.68

In Table 1, the variables from the restaurant are provided: the date of the transaction, the store (restaurant location) ID, the order number, the number of guests served, the number of items ordered and to total ticket value are provided. The StoreID column was used to identify the phone number of the location, which in turn was used to differentiate a franchise versus non-franchise location. This transaction data was a separate component to the analysis and obtained directly from the quick serve restaurant company. This information is not found publicly through Yelp, and is not readily available for reproduction of analysis performed in this chapter.

### 3.2. Comparison between Non-Franchise and Franchise Locations

Limited formal research has been conducted evaluating the differences in online customer reviews between corporate owned and franchised outlets in the quick serve food sector. Research from the hospitality sector, indicate that franchised outlets typically outperform corporate owned outlets (e.g., Lawrence, 2015). This study found evidence that the reverse could be true in the quick serve food sector.

Corporate owned locations for this restaurant chain are located primarily in Southeastern and Midwestern states. A total of 462 locations for which there is also transaction data were found on Yelp. These locations spanned 17 states, and the average rating is calculated at the state level. A map of the U.S. with these 17 states highlighted is shown in Figure 1 below. Note that the numbers within the states represent the sample size of restaurant chains drawn from these states, while the color code represents the average rating within the state; dark blue represents the highest rated states followed by cyan followed by orange, which represents the lowest rated states.

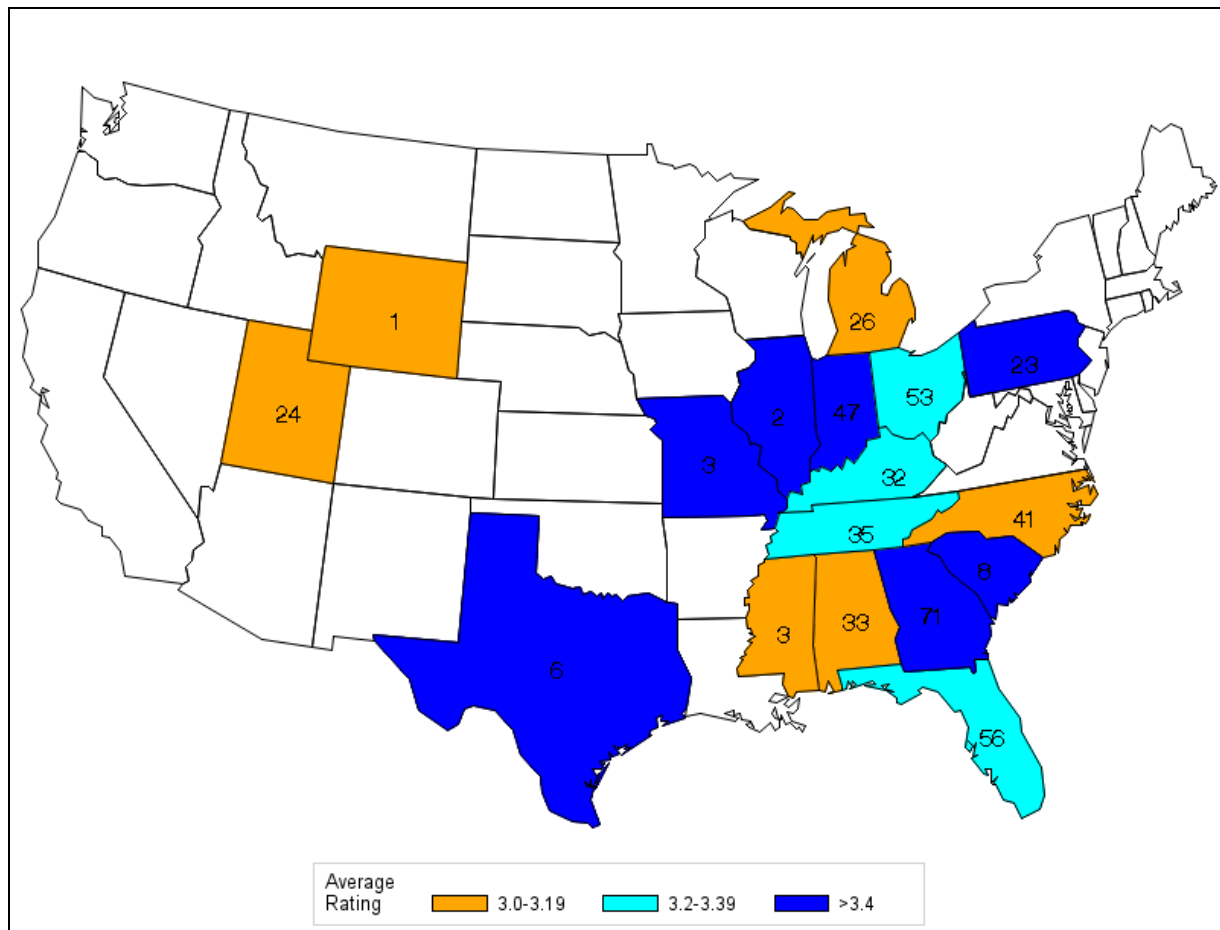


Figure 1. Map of Average Rating by State for Non-Franchise Locations

It is important to note in the figure above that the sample size for some of the states is small. Results from these states should be considered to be directional rather than statistical: Wyoming, Illinois, Missouri, Mississippi, South Carolina and Texas. None of the states in Figure 1 have an average rating below 3, which will draw a sharp contrast to the map in Figure 2 below. The map in Figure 2 shows the average rating of restaurant chains for franchised locations and displays the average ratings in the same 17 states as the above map in Figure 1.

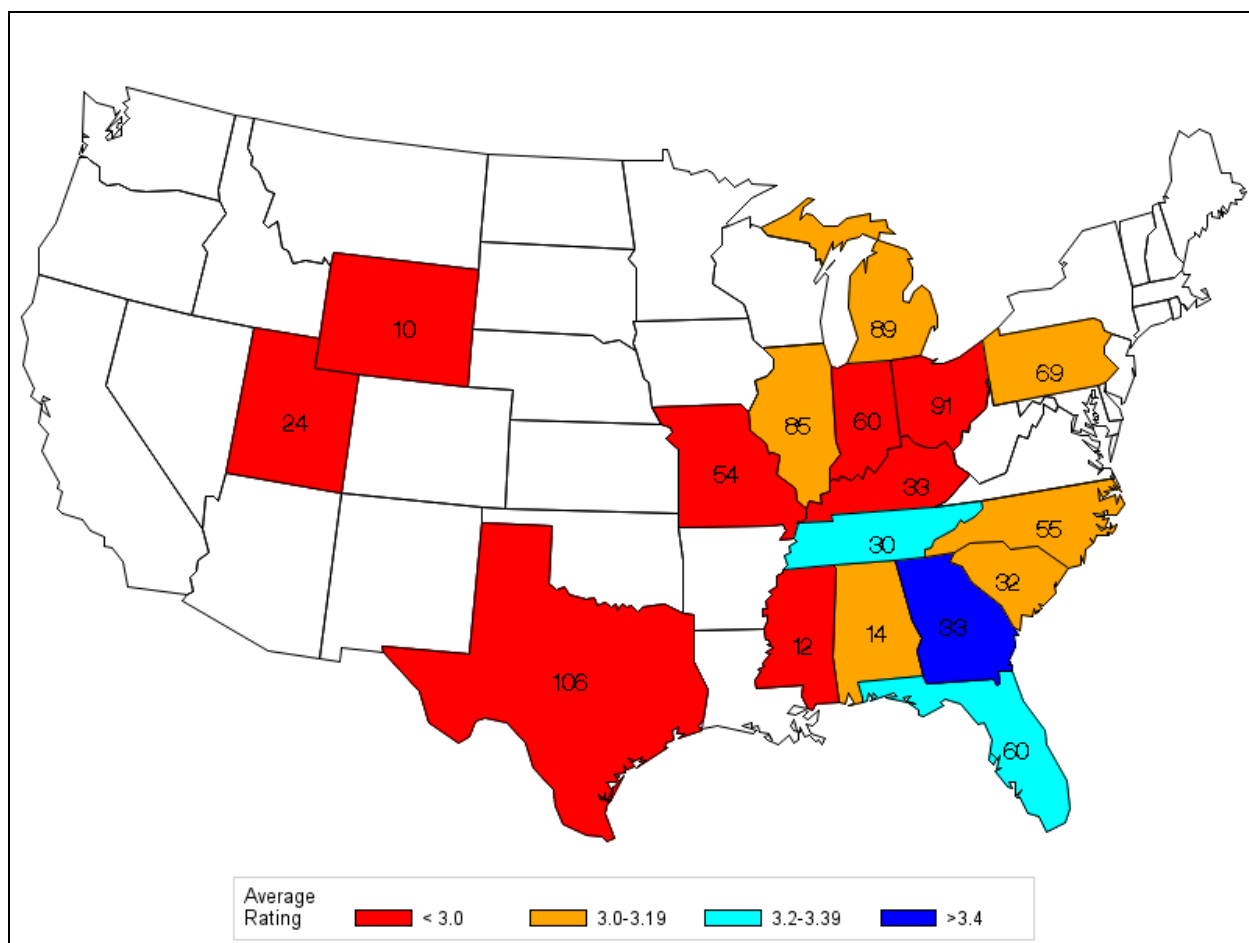


Figure 2. Map of Average Rating by State for Franchise Locations

The colors represent the same intervals of average rating as the map in Figure 1, but there is now a fourth color used – red – which represents states with an average rating below 3 stars. Nearly half the states have an average rating of less than 3 for the franchised locations.

Many franchise locations perform worse than their non-franchise counterparts in each state. This is true of states such as Indiana, Kentucky, Ohio, Pennsylvania, and Utah. Each of these states is colored differently between Figure 1 and Figure 2, and the coloring in Figure 2 shows a lower average rating of non-franchise locations as compared to franchise locations in these states. Indiana, for example, has an average rating of over 3.4 for its franchise locations (blue coloring in Figure 1), but has an average rating of under 3 for its non-franchise locations (red coloring in Figure 2). A more detailed state level comparison between the average rating of non-franchise and franchise locations is shown in Figure 3.

To better ascertain the state-by-state differences in average ratings between franchise and non-franchise locations, Figure 3 below shows a bar graph illustrating these differences. It should be noted that states which have single digit sample sizes in either Figure 1 or Figure 2 are not included in this graph. For example, a state like Texas which has 106 franchise restaurants and performed very poorly (average rating less than 3) only has six non-franchise locations for



comparison. While the six non-franchise locations have an average rating greater than 3.4, it is difficult to draw statistically meaningful conclusions with this imbalance in sample sizes.

It can be seen that only franchise locations in Georgia, North Carolina, and Alabama perform better than their non-franchise counterparts, and the differences are rather small. There are some rather large differences in favor of the non-franchise locations in states such as Indiana, Pennsylvania, Kentucky, Ohio, and Utah, which is confirmed in Figure 3.

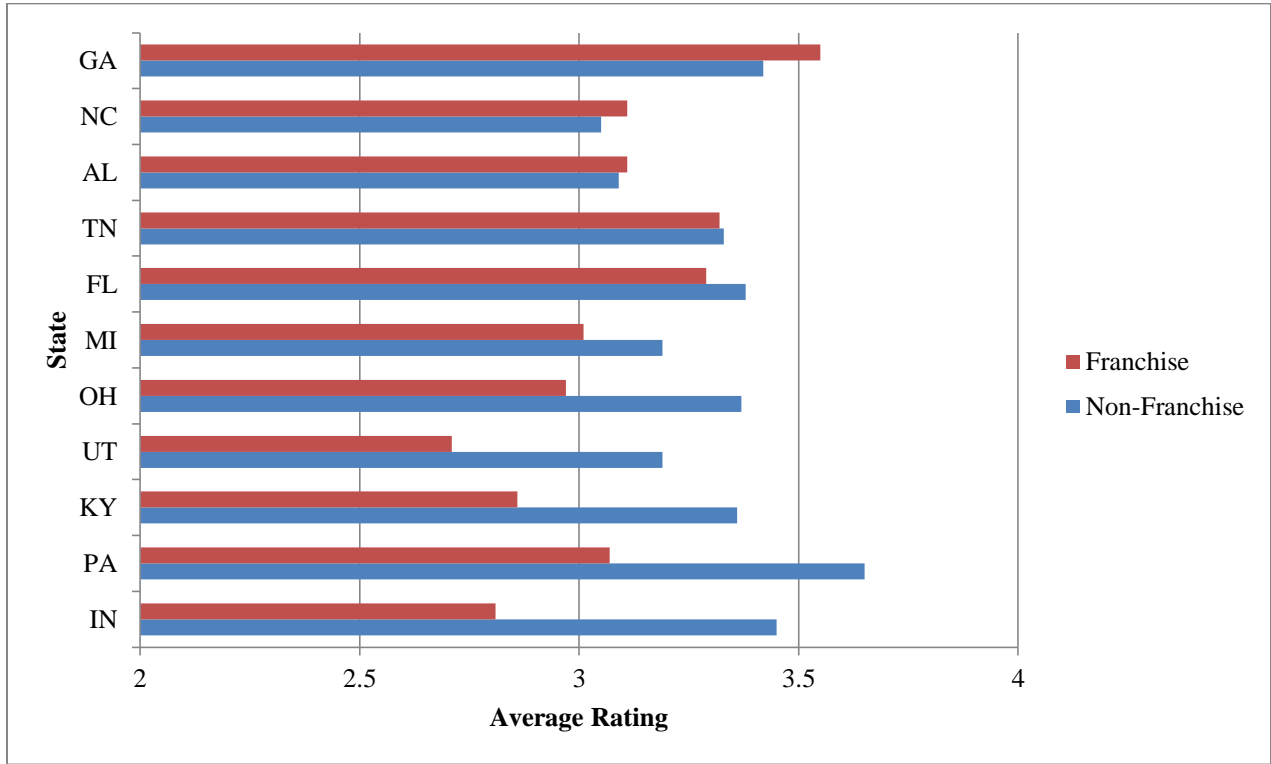


Figure 3. Comparison of Average Ratings between Franchise and Non-Franchise Locations

The differences in the means between franchise and non-franchise locations can be tested using a two-sample t-test. The sample sizes for the states in Figure 3 are generally large enough (most have at least 30) to make meaningful comparisons. The standard deviations for the testing groups were found to be sufficiently similar to allow for the use of a pooled standard deviation.

The test statistic for a two-sample t-test:

$$\text{test statistic} = \frac{\bar{x}_{NF} - \bar{x}_F}{s \sqrt{\frac{1}{n_{NF}} + \frac{1}{n_F}}} \quad (\text{eq. 1})$$

$$\text{where } s = \sqrt{\frac{(n_{NF}-1)s_{NF}^2 + (n_F-1)s_F^2}{n_{NF} + n_F - 2}} \quad (\text{eq. 2})$$

In the equations above,  $\overline{x}_{NF}$  and  $\overline{x}_F$  represent the mean rating of the non-franchise and franchise locations in a given state, respectively,  $n_{NF}$  and  $n_F$  represent the number of non-franchise and franchise locations in a given state, respectively and finally,  $s_{NF}$  and  $s_F$  represent the standard deviation of ratings for non-franchise and franchise locations in a given state, respectively.

The results of the tests are shown in Table 2. The test statistics can be used to calculate a corresponding p-value by using a t-distribution with the corresponding degrees of freedom ( $n_{NF} + n_F - 2$ ). The differences in average ratings are statistically significant for Indiana and Pennsylvania. The differences in the average rating between non-franchise and franchise locations in Ohio, Kentucky, and Utah are not statistically significant, but given the relatively large differences (0.4, 0.5, and 0.48, respectively), may be considered to be practically significant.

The remaining states in Table 2 have relatively small differences between the non-franchise and franchise locations, and the corresponding p-values suggest that there is no evidence of a statistically significant difference.

Table 2. Results for Comparisons between Non-Franchise and Franchise Locations

State	Non-Franchise Average Rating	Franchise Average Rating	Test Statistic
IN	3.45	2.81	2.87***
PA	3.65	3.07	2.06**
OH	3.37	2.97	1.79*
KY	3.36	2.86	1.66*
UT	3.19	2.71	1.51
MI	3.19	3.01	1.50
GA	3.42	3.55	0.60
FL	3.38	3.29	0.51
NC	3.05	3.11	0.28
TN	3.33	3.32	0.04

\*  $p < 0.1$ , \*\* $p < 0.05$ , \*\*\*  $p < 0.01$

These results provided the quick serve company with insight they had not previously understood – specifically that there were differences in customer perceptions and experiences between non-franchise and franchised restaurants.

These findings were quick, inexpensive and easy to extract.

### 3.3. Analysis of Reviews and Transaction Data

The second stage of the analysis was to determine if the ratings of restaurants in a state had any impact on the number of guests or the total amount of dollars spent at a restaurant outlet.

Table 3 contains the correlations between numeric ratings (i.e., 1 – 5) and the number of guests which visited the restaurant, where correlations measures the strength of the linear relationship between the two variables. It can range between -1 and +1, with negative values indicating a negative association between the variables, and positive values indicating a positive association. Correlations closer to -1 or +1 indicate strong correlation while values near 0 indicate weak correlation.

There is moderately positive correlation between the ratings and number of guests in states such as Ohio, Pennsylvania, and Michigan. For all other states, the correlations are weak. This may indicate that, at least in these three states, Yelp does “matter” for customers selecting a quick serve restaurant option.

There was little correlation found with the average check amount. This is likely due to the limited scale of the check values.

Table 3. Correlations between Ratings and Number of Guests

State	Correlation with Number of Guests	Correlation with Average Check Amount
OH	0.43**	0.03
PA	0.42**	-0.11
MI	0.31*	0.16
AL	0.23	0.05
UT	0.23	0.11
GA	0.22	0.13
NC	0.18	0.03
KY	0.06	0.37*
IN	0.01	-0.09
FL	-0.02	0.08
TN	-0.15	0.29

\*  $p < 0.1$ , \*\* $p < 0.05$

### 3.4. Analysis of All U.S. Locations

Analysis of numeric ratings was completed at the state level, regardless of whether the location was a franchise or non-franchise restaurant. The map of results is shown in Figure 4.

Almost every state is represented in this map, except for Rhode Island, Vermont and New Hampshire, where no restaurant outlets were located with reviews on Yelp. Again, the numbers on the states represent the sample size of restaurants drawn from that state and the colors reflect the average numeric ratings.

The Midwest and Mountain states have the lowest ratings on average, with many of these states in red, signifying an average Yelp rating of less than 3.0 stars. States in the Southeast and Northeast regions generally have higher numeric Yelp ratings than the Midwest and Mountain states. These two regions exhibit states with similar average ratings, with numerous blue and cyan states in the Southeast and Northeast, signifying average ratings of 3.2 or greater. The Pacific Coast states, including Alaska and Hawaii all have average numeric ratings of slightly above 3.0 stars.

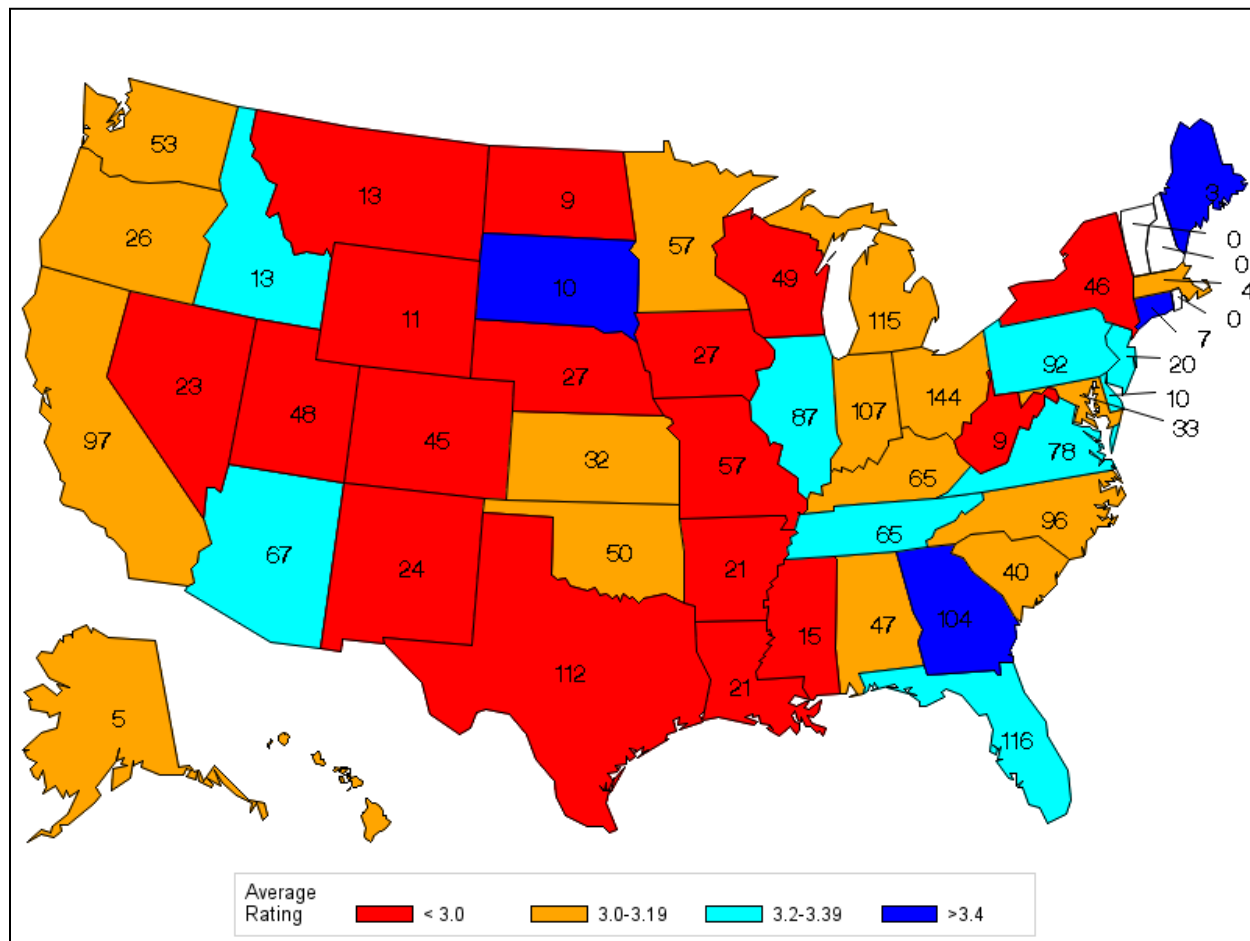


Figure 4. Map of Average Rating by State across United States

### 3.5. Description of Text Reviews

In addition to the numeric values that are included in a Yelp review, records also include a substantive amount of text. Yelp reviews can include up to 5,000 characters. This is in contrast to the 140 character limit for Twitter. Whereas numeric ratings, once extracted, can be analyzed using relatively traditional statistical techniques like t-tests and correlations, text analysis requires a new generation of analytical techniques.

A word cloud is simply a spatial image of extracted words where the size of the word indicates its frequency or importance in a text narrative.



In the study, word clouds were created to compare franchise locations to non-franchise locations, similar to the comparisons in the actual numeric ratings of the restaurants. Results showed that the word clouds generally had the same pattern, with many customers talking about food items. Word clouds were also created to compare reviews based on the rating of the restaurant (i.e., 1-star rated to 5-star rated restaurant locations).



The word clouds shown measure the frequency of single words, but it is also possible to plot combinations of words, referred to as n-grams. In an n-gram, the frequency of "n" consecutive words is measured and then plotted in a word cloud. While not used in this study, a bigram or trigram word cloud is also another simple alternative method for generating word clouds (Garcia-Castro, 2016).

of messaging and advertising for quick serve restaurant chains – to determine which menu items are most frequently mentioned in reviews, and whether those mentions are associated with low ratings or high ratings.

### **3.6. Caution Related To Analysis of Reviews**

There is no question that some percentage of reviews on Yelp, like any other social media site, are “fake”. This is true because either the owner/manager of the restaurant is providing his own “glowing” reviews or because competitors are unfairly “slamming” the restaurant in an effort to drive traffic to their own restaurant.

Researchers engaged in examining questions related to falsified reviews have found that roughly 16% of restaurant reviews on Yelp are fake. They also found that restaurants are more likely to commit review fraud when its reputation is weak, i.e., when it has few reviews, or it has recently received genuine bad reviews in an effort to “balance out” the ratings. Restaurants are also more likely to be victims of unfavorable fake reviews when they face increased competition. However, these same researchers found that chain and quick serve restaurants are less likely to have fake reviews on Yelp (Luca M. et al., 2015).

While fake reviews do not negate the value of performing analysis of numeric ratings and text comments for restaurants, researchers should be mindful of their presence – particularly when sample sizes are small (Lim, 2015).

.

## **4. Conclusion**

This study sought to determine whether ratings on Yelp are relevant to a quick serve restaurant's performance, with particular attention paid to differences between franchised and non-franchised outlets. Both numeric ratings and text reviews were analyzed. The numeric ratings indicated that non-franchise locations of restaurants for this company generally performed better in terms of Yelp ratings relative to the franchise locations. Results were plotted in a series of maps to highlight differences by state.

Given the volume of reviews, combined with the detected variation between franchised and non-franchised outlets and the correlation between numeric ratings with the number of guests, this study also provided evidence for the position that Yelp reviews are relevant to operational performance and evaluation of customer satisfaction in the quick service restaurant sector. This is contrary to previous findings.

Overall, accessing information from review sites like Yelp can provide practitioners with quick, inexpensive (effectively free), valuable information regarding operational productivity, customer perceptions, efficacy of messaging and advertising. The flexibility provided to researchers by Yelp to extract specific data related to location, time period, identified words related to products or menu items, allows for the translation of previously “dark data” into meaningful information to improve decision making.

## Guide To Using R to Extract Yelp Data

In addition to providing research results related to a study in the quick serve restaurant sector, this chapter is also written as an instructional “how to” guide for researchers and practitioners to access and extract Yelp data.

In the current study, the programming language R is used to connect to the Yelp API (Application Program Interface). The Yelp API allows researchers to search and query Yelp for information about rated businesses. In the current study, the R packages used for working with the Yelp API are stringr, httr, jsonlite, and RCurl.

The first step in connecting to the API is creating an account on Yelp. Creating an account provides the researcher with access to the Yelp developers’ page (<https://www.yelp.com/developers>). To initialize access through R, the researcher needs to generate a unique consumer key, consumer secret, token, and token secret. These elements are analogous to a password to connect to Yelp's API. A sample of what these tokens look like is shown in Figure 7 below (Yelp Developers Search API). Some of the characters in the tokens are shaded out in the figure due to the nature of the data, but notice that these tokens can be generated freely by creating an account on Yelp.

### API v2.0

Consumer Key	Rea2EEtgYc3VP	
Consumer Secret	v-g9WTFENLkM	
Token	wNidw1XQfWp1Z	
Token Secret	NSkpHidrRCx_oj	

Generate new API v2.0 token/secret

Figure 7. Consumer Tokens from Yelp API

To create the connection with the API, two lines of code are needed in R. The first of the two lines takes the consumer key and consumer secret, and registers an application. The second of the two lines takes the token and token secret, and creates a signature which can then be used to generate requests from the Yelp API. This process can be thought of as a "handshake" between the R console and Yelp API.



Sample code is shown below:

```
myapp = oauth_app("YELP", key=consumerKey, secret=consumerSecret)
signature = sign_oauth1.0(myapp, token=token, token_secret=token_secret)
```

Once the connection is made, the researcher can now search Yelp for a particular business. This is accomplished through identification of the desired search parameters. The general form of the URL string is `http://api.yelp.com/v2/search/`. Then, depending on specified parameters such as location, name, or category of the business, the URL string is expanded to contain these search criteria.

In the example of searching for the quick serve restaurant used in this analysis, the search string is created as follows:

```
yelpurl <- paste0("http://api.yelp.com/v2/search/?location=", city, "&term=restaurant_name")
```

The location input into the search string took the variable "city". Notice that the term "city" does not appear in quotes in the search string. This is because it was not static, and varied throughout the process. In order to obtain results for this quick serve restaurant nationally, a list of approximately 2,000 U.S. cities is used as arguments in this search string. With each iteration, a different city is passed into this line of code, retrieving results when searching that city. The second part of the search string, coined "term", is the name of the quick serve restaurant. Since the name of the actual restaurant will remain anonymous throughout this analysis, the term "restaurant\_name" will signify where the researcher should input the name of the business of interest.

Each time a given search string is passed through the API, the results are restricted to only 20 search results returned for each iteration. While this is a challenge, there is an alternative to using a list of thousands of cities to perform the search. Specifically, there is an optional "offset" feature which can be used; instead of passing a list of cities through the search strings, the offset parameter can be utilized to retrieve the first 20 search results in the first iteration, followed by the 21st to 40th search results in the second iteration, and so on. Either way, a loop is required to either cycle through a list of cities, or through all the search results. In using the former option, the results should be stripped of duplicates. This is true because frequently searching neighboring cities may yield the same business in the returned search results. Up to 25,000 calls can be made through the Yelp API daily.

Once the search string is built for each iteration, the "GET" function can be used in R. The relevant and required arguments in this function include the search string (called "yelpurl") and the signature variable from above. Sample code is shown below:

```
data = GET(yelpurl, signature)
datacontent = content(data)
yelp.json = jsonlite::fromJSON(toJSON(datacontent))
yelp.df = yelp.json$'businesses'
```

The first line creates an R object which contains the data pulled from Yelp for the given search string. The second through fourth lines of the code above transform the data into a structured file which is more readily analyzable. After the fourth line of code, the object "yelp.df" is a data.frame, which in R is similar to a matrix or table of data. Importantly, this is where the data converts from being “unstructured” to becoming “structured”.

The next line of code allows the user to specifically choose the desired attributes about the business or restaurant of interest. For example, if the researcher wants the phone number, rating, name and indicator of whether the business is closed or open, then the following line of code can be used:

```
ScrapeOutput = yelp.df[1:20, c("phone", "rating", "name", "is_closed")]
```

In this line of code, the researcher needs to select which parameters should be kept from the list of returned parameters from the scrape. Parameters should be listed in quotations just as they appear above (such as "phone" or "rating"). It is useful to get the name of the business, as sometimes search results yield names of different businesses, which can then be removed. Additional parameters which can be selected are shown below in Figure 8. A more comprehensive list of all search parameters, including location data such as the longitude and latitude coordinates of the business, can be found on the Yelp website under the “Developers” section.

<b>Business:</b>		
<b>Name</b>	<b>Type</b>	<b>Definition</b>
id	string	Yelp ID for this business
is_claimed	bool	Whether business has been claimed by a business owner
is_closed	bool	Whether business has been (permanently) closed
name	string	Name of this business
image_url	string	URL of photo for this business
url	string	URL for business page on Yelp
mobile_url	string	URL for mobile business page on Yelp
phone	string	Phone number for this business with international dialing code (e.g. +442079460000)
display_phone	string	Phone number for this business formatted for display

Figure 8. Sample Parameters Available through Yelp API

The first step is to collapse all the separate reviews gathered from the previous section into one string of words (McNeill, 2015). The next step is to use built in functions to make all the letters lowercase, remove any punctuation, and erase extra whitespace which appears in the reviews. Some sample code which performs these actions is shown below:

```
r1 <- paste(nonfranchise, collapse=" ")
review_source <- VectorSource(r1)
corpus <- Corpus(review_source)
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, stripWhitespace)
```

In analyzing the text reviews, it is also useful to suppress common terms which are not of interest. For example, pronouns such as "I", "my", or "we" are not typically of interest. There is a built-in function in the tm package which contains "stopwords". Stopwords is an already built list of 174 terms which include common pronouns and verbs typically ignored in analysis. Some of these terms are shown in Figure 9 below.

```
> stopwords("english")
[1] "i"      "me"      "my"      "myself"  "we"
[6] "our"    "ours"    "ourselves" "you"     "your"
[11] "yours"  "yourself" "yourselves" "he"      "him"
[16] "his"    "himself" "she"      "her"     "hers"
[21] "herself" "it"      "its"      "itself"  "they"
[26] "them"   "their"   "theirs"   "themselves" "what"
[31] "which"  "who"     "whom"    "this"    "that"
[36] "these"  "those"   "am"      "is"      "are"
[41] "was"    "were"    "be"      "been"    "being"
[46] "have"   "has"     "had"     "having"  "do"
```

Figure 9. List of Stopwords in Text Mining Library of R

It is also useful to add to this list. For example, in searching for a restaurant, it is not of particular interest to see the term "food" appear in a review. To add to this list, one can use the concatenate feature in R to append more words to the existing list of stopwords. A sample line of code which adds the words "food", "get", and "will", followed by code to apply the stopwords, is shown below.

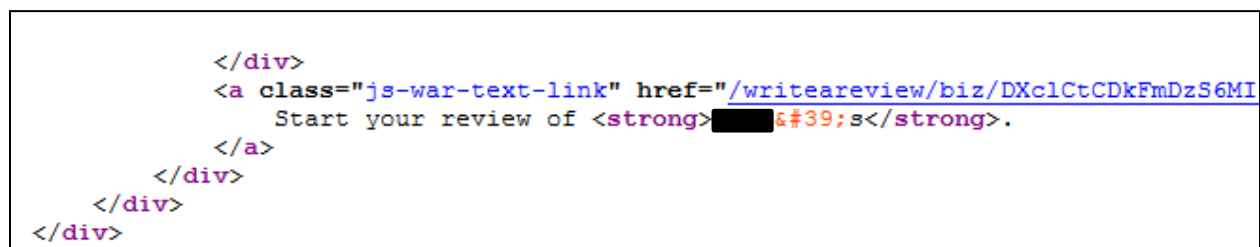
```
mystopwords <- c(stopwords("english"), "food", "get", "will")
corpus <- tm_map(corpus, removeWords, mystopwords)
```

The remaining code pertains to finding the frequency of the individual terms after all unwanted terms have been removed. In the last line, the wordcloud function in R takes the top words with

their corresponding frequencies to plot them. The number of words which are to be included can be adjusted, as the code shown below will take the top 50 words.

```
dtm <- DocumentTermMatrix(corpus)
dtm2 <- as.matrix(dtm)
frequency <- colSums(dtm2)
frequency <- sort(frequency, decreasing=TRUE)
words <- names(frequency)
wordcloud(words[1:50], frequency[1:50], colors=brewer.pal(8, "Dark2"))
```

The second part of this analysis involves using the text reviews given by customers. As described above, the "snippet\_text" parameter can be returned to view written comments by customers of the business. However, this only contains the first text review of a business on Yelp. As a result if there are multiple reviews, the remaining reviews are ignored. To work around this, R can be used to perform an HTML scrape. To perform the HTML scrape in R, the httr and XML packages are needed. For the HTML scrape, it is necessary to view the source code behind the web page. This can be done by "right clicking" on a web page and selecting the "View Page Source" option. When doing this, a new page will open with the HTML code. In looking at this page, one needs to find where the comments start. For example, in Figure 10 below, the text "Start your review of" signifies that comments will begin (the name of the quick serve restaurant in this study is blacked out, but would appear in place of the black box). Once this is identified, there are a series of R functions which can be used to parse this information and translate it into a data frame for analysis.



```
</div>
<a class="js-war-text-link" href="/writeareview/biz/DXclCtCDkFmDzS6MI
  Start your review of <strong>[REDACTED]&#39;s</strong>.
</a>
</div>
</div>
</div>
```

Figure 10. Sample View of HTML Source Code

To create the HTML scrape, the "url" parameter (shown in Figure 8) must be returned from the original Yelp API call. This parameter contains the URL of the business of interest. Using this parameter, the HTML functions can be used to go to this specific webpage, and retrieve each of the individual comments left by customers. Some sample code is provided below. In this code, the dataset "test" contains the URL of each individual restaurant searched, along with a corresponding phone number and state to uniquely identify the location. A for loop is run for each URL in the dataset, to iterate through all of them and output the reviews for each individual business which is searched. The functions "htmlParse" and "xpathSApply" are needed to turn the HTML source code into a data frame in R which can then be manipulated by the user. The

"grep" function is also required to find the specific location on the webpage where the comments begin. Refer to the example in Figure 10, where the words "Start your review" signify a new review. The "grep" function is used to locate where on the page this begins, and the subsequent line begins retrieving results only once this string has been located. The "grep" function is also useful in removing lines which contain something other than a review. For instance, as part of each review, there is a line which asks if the review was helpful. By using the statement "-grep('Was this review helpful')", all lines containing this statement can be removed. Ultimately, the lines selected in this code should each reflect a separate review from a customer. The "cbind" function (column bind) is used to append the individual phone and state identifier to each review. The "rbind" function at the end is used to append results from each iteration to create one master data frame object which contains all the reviews for searched businesses. Notice that in the code below, a "#" symbol reflects a comment in R, which is ignored in processing of code. Sample code is shown below here:

```
finaldata <- NULL # initialize a data frame named finaldata
for (i in 1:nrow(test)){ # data frame test contains the url, phone and state for business
  tempurl <- test$url[i] # tempurl is the url in each iteration
  doc <- htmlParse(tempurl) # htmlParse function
  y <- xpathSApply(doc,'//p', xmlValue, encoding="UTF-8") # turns result into a table
  n <- grep('Start your review', y) # grep is used to search for strings
  y2 <- y[-c(1:n, (length(y)-2):length(y))] # further subset results
  y3 <- y2[-grep('Was this review helpful', y2)] # further subsetting
  y4 <- cbind(y4, as.character(test$phone[i]), as.character(test$state[i])) # combine data
  finaldata <- rbind(finaldata, y4) # append results
}
```

### Select SAS Code

SAS v9.4 was also used in this project. Specifically, SAS was used for the statistical analysis as well as the creation of the maps. Sample SAS code is shown below for creating the maps. The format for creating the ranges for the state colors is shown, followed by the PROC GMAP function which creates a map using one of the default maps (maps.us) in the SAS maps library. To put the sample size labels on the state, an annotate option is available within the GMAP procedure, which uses a separate dataset that contains the labels for the states.

```
proc format;
value rating_format low-3=< 3.0'
3.0-3.2 = '3.0-3.19'
3.2-3.4 = '3.2-3.39'
3.4-high = '> 3.4';
run;
```

```
proc gmap data=dataset map=maps.us; format average_rating rating_format.;  
id state;  
choro average_rating/ discrete coutline=black annotate=maplabel;  
run;  
quit;
```

## References

- Dwyer, E. A. 2015. Price, Perceived Value and Customer Satisfaction: A Text-Based Econometric Analysis of Yelp! Reviews. Scripps Senior Theses. Paper 715. [http://scholarship.claremont.edu/scripps\\_theses/715](http://scholarship.claremont.edu/scripps_theses/715) Accessed on Sep. 2, 2016.
- Garcia-Castro 2016. R. Example of creating n-gram clouds, [https://rstudio-pubs-static.s3.amazonaws.com/118348\\_a00ba585d2314b3c937d6acd4f4698b0.html](https://rstudio-pubs-static.s3.amazonaws.com/118348_a00ba585d2314b3c937d6acd4f4698b0.html) Accessed on Jan. 20, 2016.
- Gartner Group 2016. Dark Data. <http://www.gartner.com/it-glossary/dark-data/> Accessed on September 3, 2016.
- Gregory, A., Parsa, H. G., Terry, M. 2011. Why Do Restaurants Fail? Part III: An Analysis of Macro and Micro Factors. University of Central Florida, The Dick Pope Sr. Institute for Tourism Studies UCF Rosen College of Hospitality Management.
- Halevy, A. Norvig, P. and Pereira, F. 2009. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2):8–12.
- Hlee, S., Lee, J., Yang, S., Koo, C. 2016. An Empirical Examination of Online Restaurant Reviews (Yelp.com): Moderating Roles of Restaurant Type and Self-image Disclosure, Information and Communication Technologies in Tourism 2016. pp 339-353.
- King, B. 2016. Caught in the middle: franchise businesses and the social media wave, Journal of Business Strategy, Vol. 37 Iss: 2, pp.20 – 26.
- Lawrence, B. and Perrigot, R. 2015. Influence of Organizational Form and Customer Type on Online Customer Satisfaction Ratings. Journal of Small Business Management, 53: 58–74.
- Lim, Y, Van Der Heide, B. 2015. Evaluating the Wisdom of Strangers: The Perceived Credibility of Online Consumer Reviews on Yelp. Journal of Computer Mediated Communication, 20: 67–82.
- Luca, M. 2011. Reviews, Reputation, and Revenue: The Case of Yelp.com. Harvard Business School Working Paper, No. 12-016, September 2011. (Revised March 2016. Revise and resubmit at the American Economic Journal - Applied Economics.).
- Luca M. and Zervas, G. 2015. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. <http://people.hbs.edu/mluca/fakeittillyoumakeit.pdf> Accessed on Sep. 2, 2016.
- Malhotra, A. and Malhotra, C. 2015. How CEOs Can Leverage Twitter. <http://sloanreview.mit.edu/article/how-ceos-can-leverage-twitter/> Accessed on Sep. 3, 2016.
- McNeill, M. 2015. Text mining in R: how to find term frequency. <https://deltadna.com/blog/text-mining-in-r-for-term-frequency/> Accessed on Oct. 3, 2015.

Monk, B. and Navidi, W. 2014. Essential Statistics. New York, NY: McGraw-Hill.

National Restaurant Association 2014. Restaurant Industry Forecast  
<https://www.restaurant.org/Downloads/PDFs/News-research/research/RestaurantIndustryForecast2014.pdf> Accessed on Sep. 3, 2016.

Open Table 2016. Our Story. <https://www.opentable.com/about/> Accessed on Sep. 3, 2016.

Parsa, H.G., van der Rest, J.P., Smith, S., Parsa, R., Buisic, M. 2014. Why Restaurants Fail? Part IV The Relationship between Restaurant Failures and Demographic Factors. Cornell Hospitality Quarterly October 2014.

Remmers, M. 2014. 5 Ways to Get More Diners with Yelp: Leveraging the online review site can maximize your restaurant's exposure. <https://www.qsrmagazine.com/outside-insights/5-5ways-get-more-diners-yelp/> Accessed on Sep. 3, 2016.

Sena, M. 2016. Fast Food Industry Analysis 2016 - Cost & Trends. <https://www.franchisehelp.com/industry-reports/fast-food-industry-report> Accessed on Sep. 3, 2016.

Taylor, D, Aday, J. 2016. Consumer Generated Restaurant Ratings: A Preliminary Look at OpenTable.com, Journal of New Business Ideas & Trends. 2016, Vol. 14 Issue 1, p14-22. 9p.

Vasa, N., Vaidya, A., Kamani, S., Upadhyay, M., Thomas, M. 2016. Yelp: Predicting Restaurant Success. <http://www-scf.usc.edu/~adityaav/Yelp-Final.pdf> Accessed on Sep. 3, 2016.

Wang, Q., Wu, X., Xu, Y. 2016. Sentiment Analysis of Yelp's Ratings Based on Text Reviews. [http://cs229.stanford.edu/proj2014/Yun Xu, Xinhui Wu, Qinxia Wang, Sentiment Analysis of Yelp's Ratings Based on Text Reviews.pdf](http://cs229.stanford.edu/proj2014/Yun%20Xu,Xinhui%20Wu,Qinxia%20Wang,Sentiment%20Analysis%20of%20Yelp's%20Ratings%20Based%20on%20Text%20Reviews.pdf) Accessed on Oct. 20, 2016.

Yelp Developers Search API. [https://www.yelp.com/developers/documentation/v2/search\\_api](https://www.yelp.com/developers/documentation/v2/search_api) Accessed on Aug. 1, 2015.