THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

ST309 Group Project Report

How can a business be successful?

Candidate numbers: 18141 Contribution: 50%
Candidate numbers: 14459 Contribution: 50%

TABLE OF CONTENTS

INTRODUCTION OF THE PROBLEM TACKLED

When opening a business, a major concern for the investor or entrepreneur is “will my business be successful?”. On average, 12% of businesses fail every year making it a risky proposition (Rhodes 2018). There are many aspects of a business that could lead to its failure. However, trying to account for all of them is inefficient and costly. So what aspects should a business owner concentrate on to guarantee business prosperity?

In this project, we use R to analyse a large Yelp business and reviews dataset in order to answer the question ***“How can a business be successful?”***. This project wants to generate a comprehensive model that will allow the use of Yelp business data to predict business failure and understand the reasons for success. In this scenario, a failed business will be defined as one that is no longer open in the dataset, while a successful business will be identified with open businesses with a star rating of 4 or above. We will attempt to examine a range of variables, from business location to more specific factors (like the quality of service and overall operations). For the latter, we intend to use the information provided by the reviews of the said businesses. Different data analytics method will be employed, including cv-trees, logistic regressions, random-forests, and text analysis techniques.

Given the large dataset size (6.5GB) and the limited computing power accessible to conduct this project, the analysis will be carried out on a subset of the data. In particular, data from the ‘Restaurant’ category located in Phoenix, USA, will be used for the analysis. The code generated can be easily modified to include different subsets of data, and thus shed light on different business categories and/or regions. The choice of focusing on a subset of the data, not only solves the computing limitations faced during the analysis, but also strengthens the validity of the result in uncovering the reasons for failure and success of restaurants in Phoenix.

Several academic papers have explored the theme of business success/failure, but have mostly focused on industries other than hospitality (Youn & Gu 2010). When focusing on the hospitality industry, financial metrics have been the primary input of analysis (Youn & Gu 2010). Other reasons for success have been found to be previous owner’s experience, organisational structure, startup and human capital, among the most notable ones (Kalnins & Mayer 2004; Isaack 1958; Robb & Fairlie 2009). While these are all clearly important factors, they offer limited visibility into the more practical business aspects that matter for success.

Previous analyses of this dataset put emphasis on narrow areas of the data, only including regression or text analysis, for example. Our project aims at a comprehensive and exhaustive understanding of the business characteristics that are significant for business success. Moreover, a focused analysis on a single business category in a particular city has not been attempted in such extent, if at all.

This project will conclude that location, number of reviews and parking features are the most significant factors for success. Text analysis will highlight the importance of food as part of customers' satisfaction, as well as the crucial role of quality, tempestive service, and atmosphere. While many of these attributes can be easily guessed, the analysis empowers restaurant owners of Phoenix with a simple and achievable recipe for success.

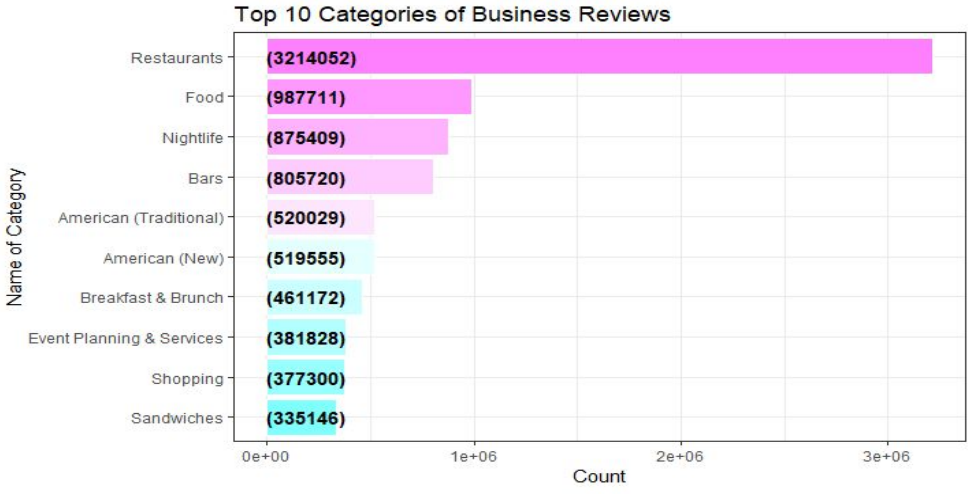
DATA DESCRIPTION

The dataset includes multiple subsets of Yelp data, originally made public for the Yelp Dataset Challenge and accessible on Kaggle (Kaggle 2018). Our analysis merged different datasets: business data, business attributes, business hours, and users reviews. The resulting merged dataset spans 11 metropolitan areas in 4 different countries, with information on 151,605 businesses and 106 variables, as well as over 5 million reviews. The data is well suited to answer the question at hand since it provides access to business data which includes business characteristics, rating levels (also a proxy for success or failure), whether it is open or closed and reviews.

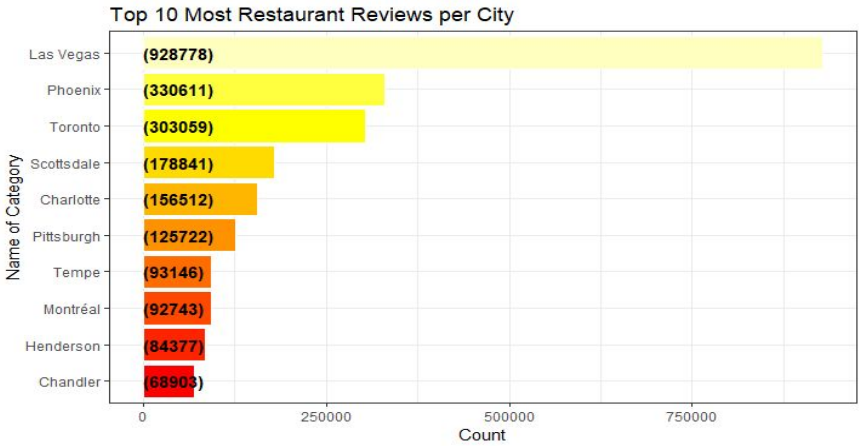
Data cleansing and missing values

Data cleansing was crucial and led to a reduction in the dataset size. Firstly, the different data subsets were merged associating businesses to attributes and reviews using business id, thus obtaining 5,059,341 observations with 106 variables. Missing values and their nature were subsequently explored. *Neighbourhood* presented a large number of NAs and was thus removed as a variable. However, this emission did not impact the analysis as city and geographical coordinates were also available. *User_id* and *Review_id* from the review dataset were also omitted as variables, as not useful for the analysis. 22,526 business observations with important missing values were also emitted.

The dataset obtained following the merging process was so large in size (6.5GB) that it made the code execution extremely lengthy and several analyses were not supported by the computer's memory. It has therefore been necessary to generate a subset on which to carry out the analysis. After exploring the top business categories across reviews we decided to focus on the largest one, restaurants.



The restaurant category was then examined to identify a suitable city to focus the analysis on. The decision to focus on a single city was made for two reasons: i) the restaurant subset was still very large (3,214,052 reviews), thus not allowing a marked improvement in data processing; ii) each city and customer group present unique aspects that should be accounted for. Given the dataset includes different metropolitan areas in different countries, analysing it in its entirety would have led to a loss of local specificity. Phoenix, the city with the second largest number of restaurant reviews, was selected because offering a better dataset size (330,611 reviews) than the first city, Las Vegas, about 3 times larger. Moreover, Las Vegas is a peculiar city for its extravagant tourism activities, thus not representative of the most common city setting.



The selection of the restaurants motivated the removal of several business attributes (e.g. those relating to hair treatments) and other variables that were not deemed relevant for this category or that presented more than a third of NA values, such as opening hours.

Transformations

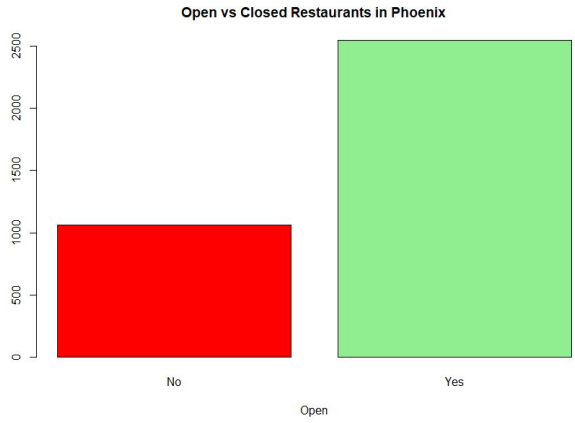
Some transformations were necessary when preparing the data. Several observations in the Phoenix dataset had NA values in the parking variables (*parking validated, parking valet* and *parking lot*). Ideally, these observations should have been completely removed, however, this was not feasible since we would have lost a large chunk of the dataset. The NA values were therefore treated as 'No' values instead. This was a required transformation to avoid complications in the regression analysis. Following this simplification, the parking variables only took two values: "Yes" and "No". It was then further transformed into a dummy variable with "Yes" being represented by 1 and "No" by 0. In addition to this transformation, *longitude, latitude*, and *review counts* variables were transformed to natural logs. The longitude and latitude transformations will improve the interpretability of the results, since the scale of variation in these variables is very small in this subset. The review count transformation is intended to smooth out the distribution of the review count variable.

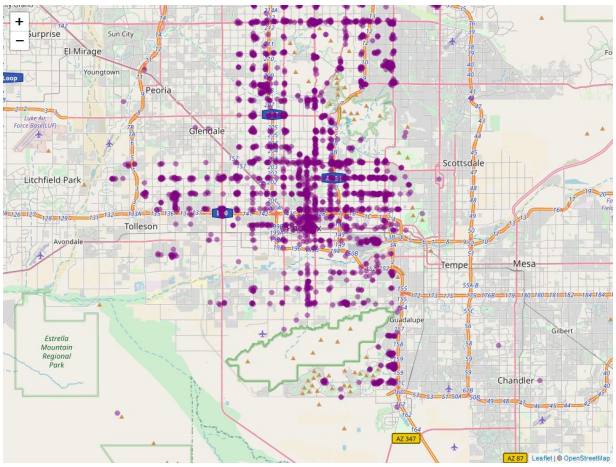
Data description

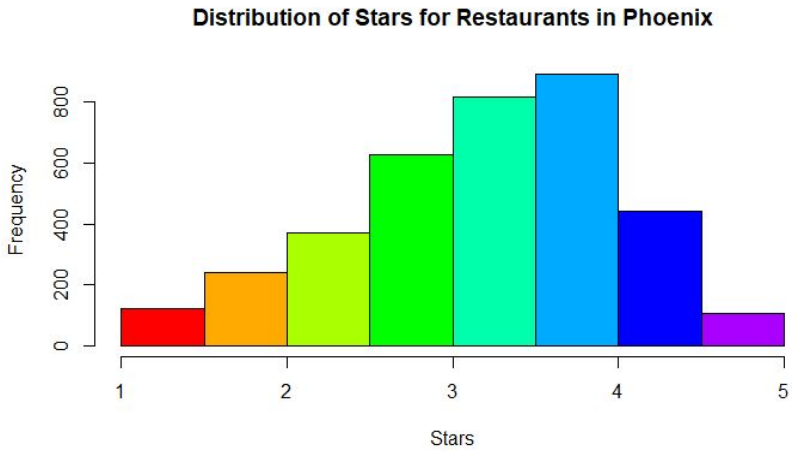
The Phoenix dataset used for statistical analysis consisted of 3,608 restaurants and 9 variables: *latitude, longitude, stars, review count, open, parking validated, parking valet, parking lot* and *parking street*. The

relevant characteristics of the Phoenix dataset used for the text analysis will be addressed in the text analysis section.

The following graphs visualise the distribution of open and closed restaurants in Phoenix, the distribution of Yelp stars for the restaurants, and their geographical distribution. We notice that 29.46% of the restaurants are closed, that they are concentrated around the city centre and along major roads, and that the stars are nicely distributed.







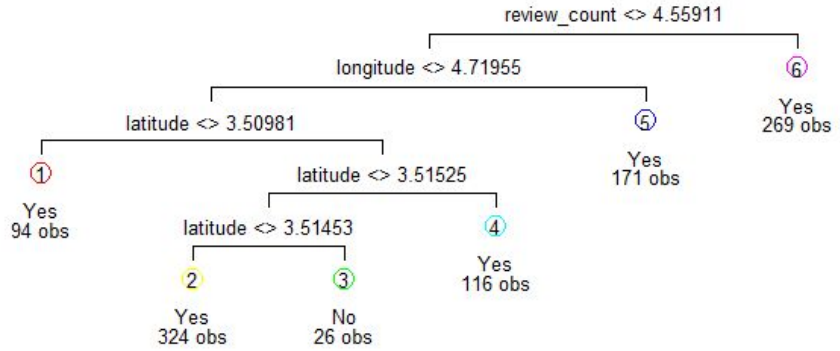
DATA ANALYSIS & MAIN RESULTS

The analysis conducted can be broken down into two sections: the first involves statistical methods to identify the reasons for failure while the second section includes a text investigation of the customer reviews to extrapolate the primary characteristics of successful businesses.

Regression analysis

Initially, we will try to understand what factors contribute to business closure. This will be done by analysing the binary variable *is_open* in our dataset, which was transformed into a factor variable *Open* for this analysis. Our exploration of the subset begins with a classification tree analysis. First, we split the data into training and testing data subsets. We generated our training subset by randomly sampling 1000

observations from our Phoenix dataset; the rest of the observations were grouped into the testing data subset. We generated a classification tree with our training data subset, the following was the result:



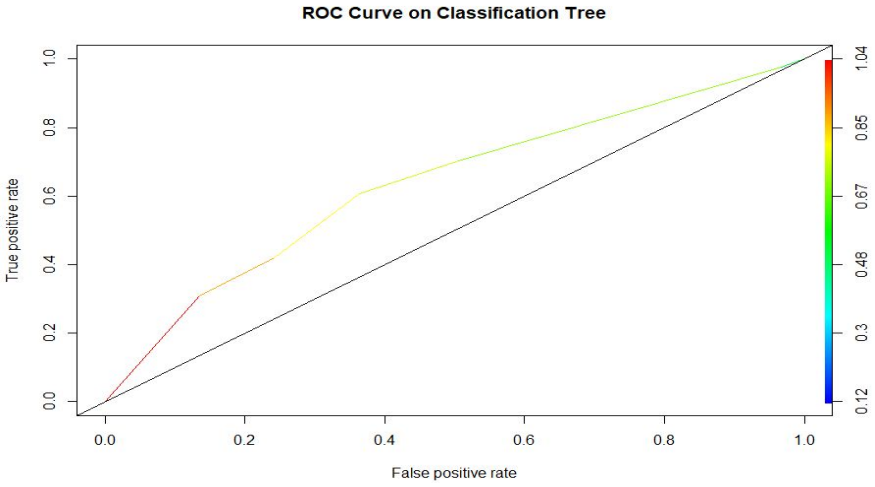
This tree has a misclassification rate of 0.2991 on testing data, which is higher than our base rate of 0.2946. We try to improve our tree's performance by using cross-validation in our node selection to prune the tree. We find that after pruning the tree, our number of terminal nodes remains unchanged and, as a result, so does our misclassification rate. Finally, we use random forests, a form of bagging, as a way to improve our misclassification rate. In theory, random forests should generate a decorrelated average tree with a lower variance and a better fit. We achieved the following results:

Confusion Matrix

	No	Yes	Error Rate
No	*93*	*207*	*0.69*
Yes	*85*	*615*	*0.1214286*

OOB estimate of error rate: 29.2%

From these results, we can see that the random forests method has managed to produce a model with a lower misclassification rate than the base. However, it should be noted that the difference in error rates is very minor. Since the values generated through random forests only showed a slight improvement in the misclassification rate overall, we can deduce that our tree is quite effective in predicting our testing data. This is further supported by the ROC curve generated from our classification tree below:

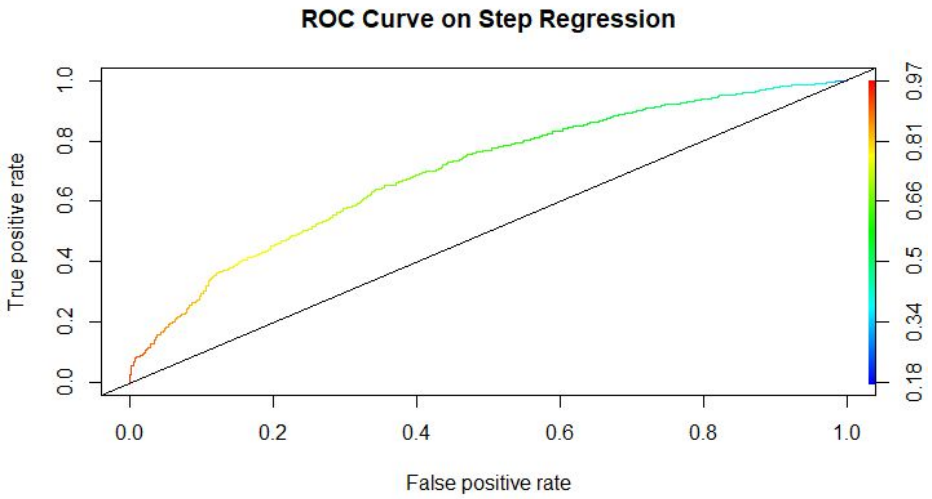


This ROC curve clearly shows how well the model performs in predicting the testing data. The classification tree yields an AUC value of 0.6355546 which is higher than the 0.5 baseline.

We continued exploring the subset through the use of a logistic regression. We used it to determine the statistical significance of each variable in the subset in determining the likelihood of a restaurant being open for business. We use a step regression approach to remove variables which are not significant at a 5% significance level or below. This approach gives us a logistic regression with 4 explanatory variables.

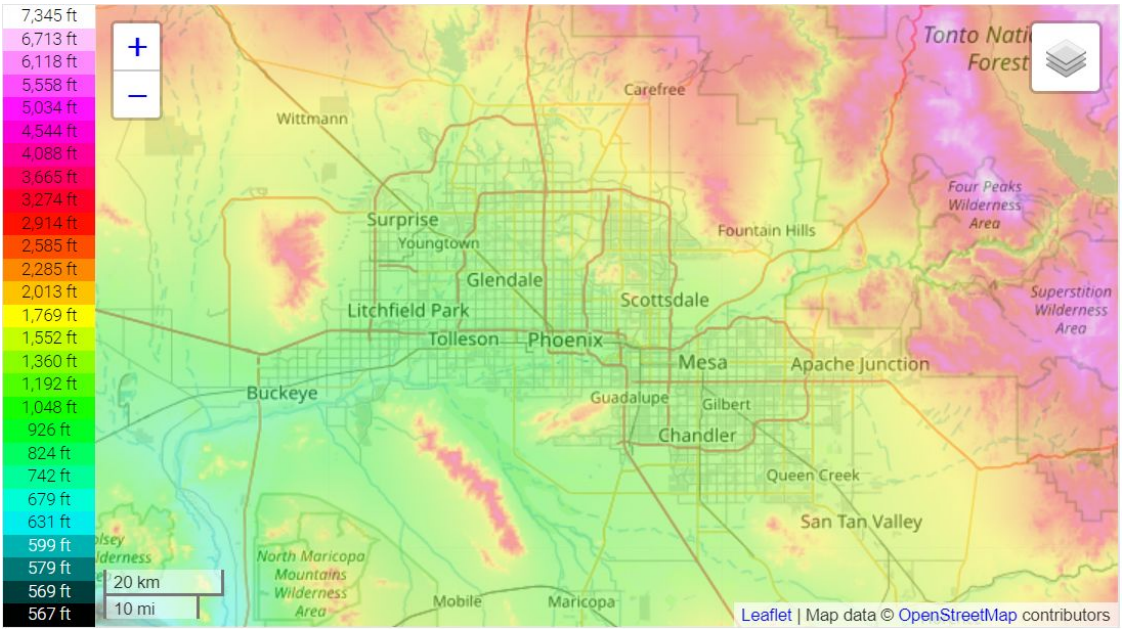
	Estimate	Std. Error	Z Value	Pr(>|z|)
Intercept	*-2.21E+03*	*6.24E+02*	*-3.54*	*0.000400 ****
Parking Valet	*-6.50E-01*	*1.70E-01*	*-3.834*	*0.000126 ****
Parking Validated	*-8.93E-01*	*3.28E-01*	*-2.726*	*0.006416 ***
Number of Reviews	*5.82E-01*	*6.46E-02*	*9.002*	*< 2e-16 ****
Longitude	*4.68E+02*	*1.32E+02*	*3.539*	*0.000402 ****

These variables are longitude, review count and types of parking (BPv and BPva). These variables are significant at least at the 1% significance level, implying they have significant predictive power of whether a restaurant is open for business. We can test this claim by generating a ROC curve for our logistic regression to test its predictive power.



The ROC curve is clearly above the threshold line, hence implying that the model has significant predictive power. This is confirmed by the AUC value of 0.6969284, which is significantly above the 0.5 threshold and higher than the tree's. Although, it should be noted that the AUC value is still relatively low if we are attempting to design a comprehensive model, as we still get incorrect results around 30% of the time, which falls in line with the base misclassification rate. This is most likely a reflection of the limits of our subset. Despite these limitations, this analysis has still managed to yield some results.

We can see that both our classification tree and logistic regression highlight the importance of reviews and location in determining the success of a restaurant. In terms of location, it is interesting to see that longitude is considered significant in our logistic regression, while latitude is not. This could be a reflection of the topography of Phoenix (Topographic-map.com 2019). In general, the city is quite flat and surrounded by mountains, with wide main roads (interstates) that are complemented by a grid-based street layout, as shown below:



We can see that there is one major road (I-10) that runs through the centre of Phoenix. Notably, we demonstrated that the concentration of restaurants was highest at this point in our data description section. Furthermore, we can see that the distribution of restaurants seems to be dictated by proximity to interstate routes (I-17 and I-10) that lead to the city center. This could imply that ease of access (transport links) are a key factor in determining a successful restaurant.

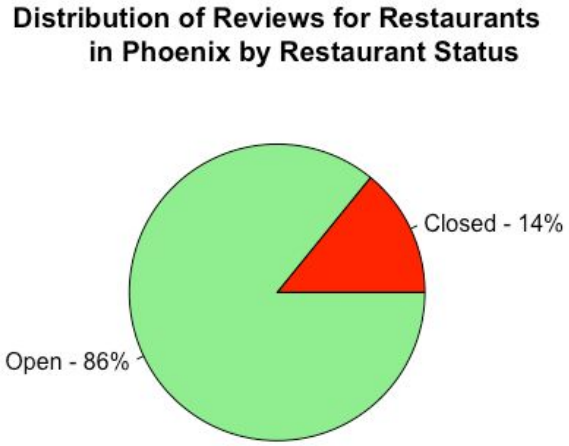
This is further supported by the other significant variables related to the types of parking available at the restaurants. Restaurants that offer validated or valet parking are more likely to succeed. Our analysis also included variables for the availability of bike and street parking, however, these variables proved to be insignificant. From these results, we can deduce that the majority of transportation in the city of Phoenix is done by car, therefore restaurants that are more accessible to cars through proximity to interstates and availability of “hassle-free” parking are more likely to succeed.

Finally, we turn our attention to the review count variable. Our logistic regression implies that an increase in the number of reviews for a restaurant is likely to increase the probability that the restaurant is successful. The model also defines stars as insignificant, implying that the quantity is more important than quality. This is valuable insight, although we should be careful with the direction of causality, since it could be that the higher review count is a byproduct of a successful restaurant and not its determinant. Furthermore, we should not discount the effects that reviews can have on a restaurants success without examining them more closely.

Text analysis

The statistical methods employed in detecting which variables affect business success the most seem to point to review count as a significant variable. While stars seem not to be significant, previous academic papers have largely demonstrated the influence of ratings and reviews, quantifying the impact of an extra half-star on Yelp as to lead to a reduction in restaurant availability by approximately 19 percentage points (Anderson & Magruder 2012, pp.957-989). A deeper understanding of customer reviews is therefore necessary. Although reviews might be subject to manipulation, existing academic usage of similar Yelp data has been found robust to such risk (Anderson & Magruder 2012, pp.957-989). The below text analysis unpicks sentiment, term frequency and importance of the review content, as well as the most important aspects customers focus on when judging their dining experience. Particular attention is given to the variation in such measures across restaurants of different Yelp rating.

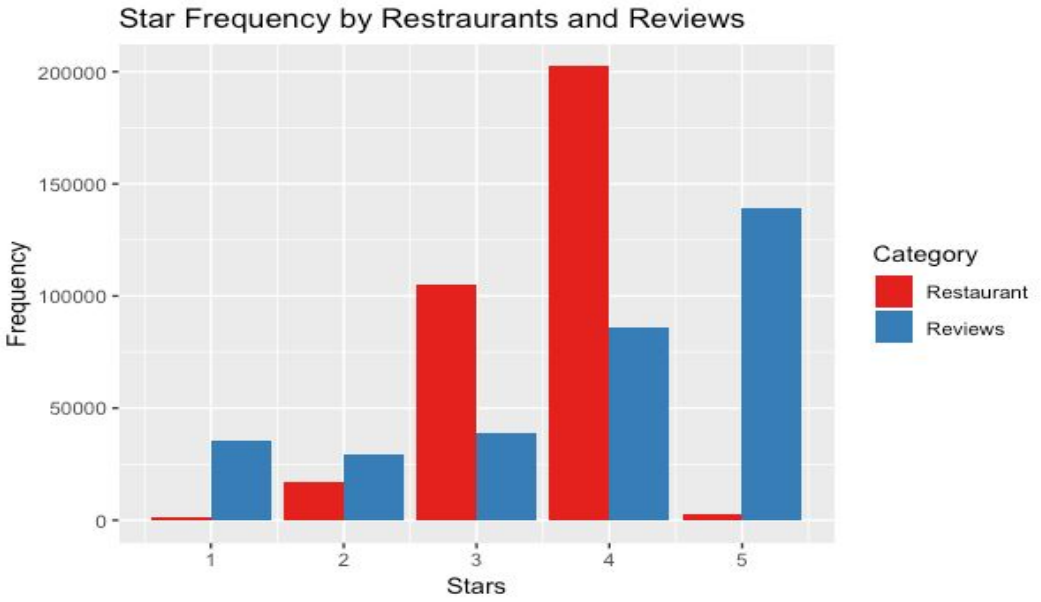
During the merging process, reviews were matched with businesses in the business dataset using *bussiness_id*. Reviews for restaurants in Phoenix were then extracted and the text language was detected. Only reviews categorised as in 'English' were kept, amounting to 330,615 entries spanning a period of more than 12 years: from March 2005 to December 2017. Out of the 330,615 reviews, 283,970 applied to open restaurants, while the remaining 46,645 referred to closed restaurants.



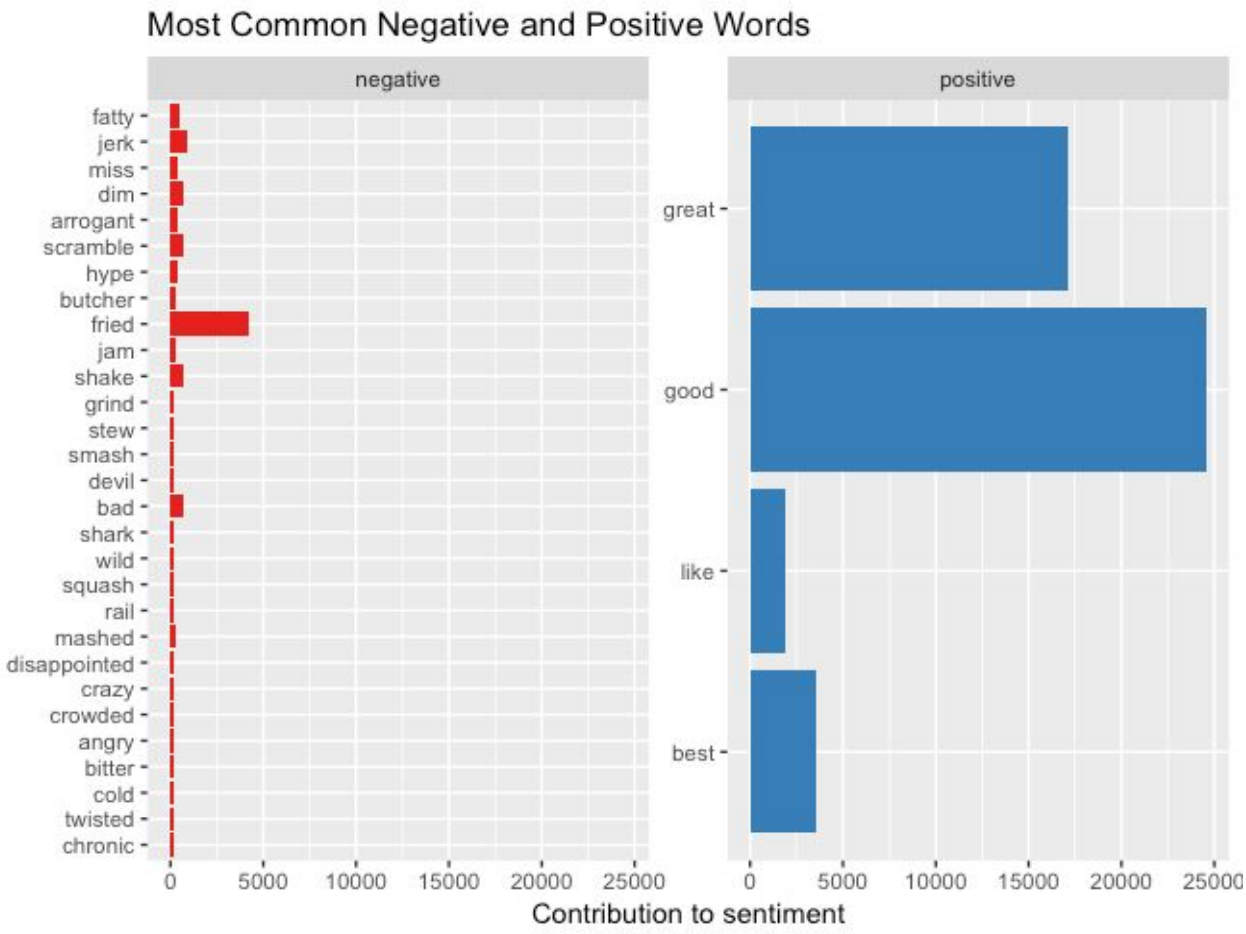
This second part of the analysis defines business success as having a higher star rating, more specifically, restaurants with at least 4 stars are considered successful. This decision of using stars rather than closure for the analysis wants to tackle a major limitation of the data: the fact that timing of reviews vs closures is not taken into consideration and the fact that several closed restaurants might have been removed from the website upon closure. This intuition is reinforced by the relatively weak correlation between stars and opening status of 0.1182327. Moreover, this paper wants to provide business owners with a recipe for success and there aren't any better ways to suggest success as in looking at customer satisfaction, and thus ratings and reviews.

Users on Yelp can leave a review as well as rate the business on the famous 1-5 star scale. The Phoenix dataset for the text analysis presents two variables with a star rating. *Stars.x* refers to the aggregate,

official Yelp star rating of the restaurant on the website, while *stars.y* refers to the rating the single customer gave when leaving a review. We have looked at the distribution of both. Note that in order to facilitate the comparison and to make the interpretation of further analysis easier, *stars.x* were combined when 0.5 values were present (e.g. *stars*=4 consists of both *stars.x*=4 and *stars.x*=4.5). As the below graph illustrates, the distribution of stars across restaurants vs reviews varies. While both similarly distributed with a mean of 4 stars, reviews are clearly more varied, with a greater number of reviews of 5 and 1 stars. This uncovers a major characteristic of reviews: customers tend to be vocal in such instances where they were extremely satisfied or unsatisfied. While this conclusion could be interpreted as a limitation of this analysis, it is important to recognise that the overall review count and official Yelp rating are critical factors in suggesting the success of a business. Since these variables are simply an aggregate of the single reviews, the latter are valuable in shading light on reasons for business success.



In order to carry out the text analysis, the reviews were “tidied” separating the words in a single vector, with a word in each row. A subset of the “tidy text” was created for stop and non-stop words. Stop words are defined as the most common words, such as “and”, that are not useful for the analysis. As a first step, we identified the most frequent positive and negative words in the entire set of reviews to understand what influences customers’ mood. Such a classification was conducted using the Bing dictionary of sentiment. A few main results can be driven from the graph below. Firstly, there is more homogeneity in term frequency for positive words compared to negative words. The most common positive words are: “great, good, like, best”. This suggests people reference good food and a pleasant experience. Secondly, the negative words are more varied in terms of connotations. Many seem to refer to the poor quality of food (“bad, fatty, fried”), while others to the poor quality of service (“jerk, arrogant, bitter, disappointed, angry”). Overall, while these findings are not surprising, they suggest that good quality of food and service are the key to customer satisfaction and topics customers share comments about.



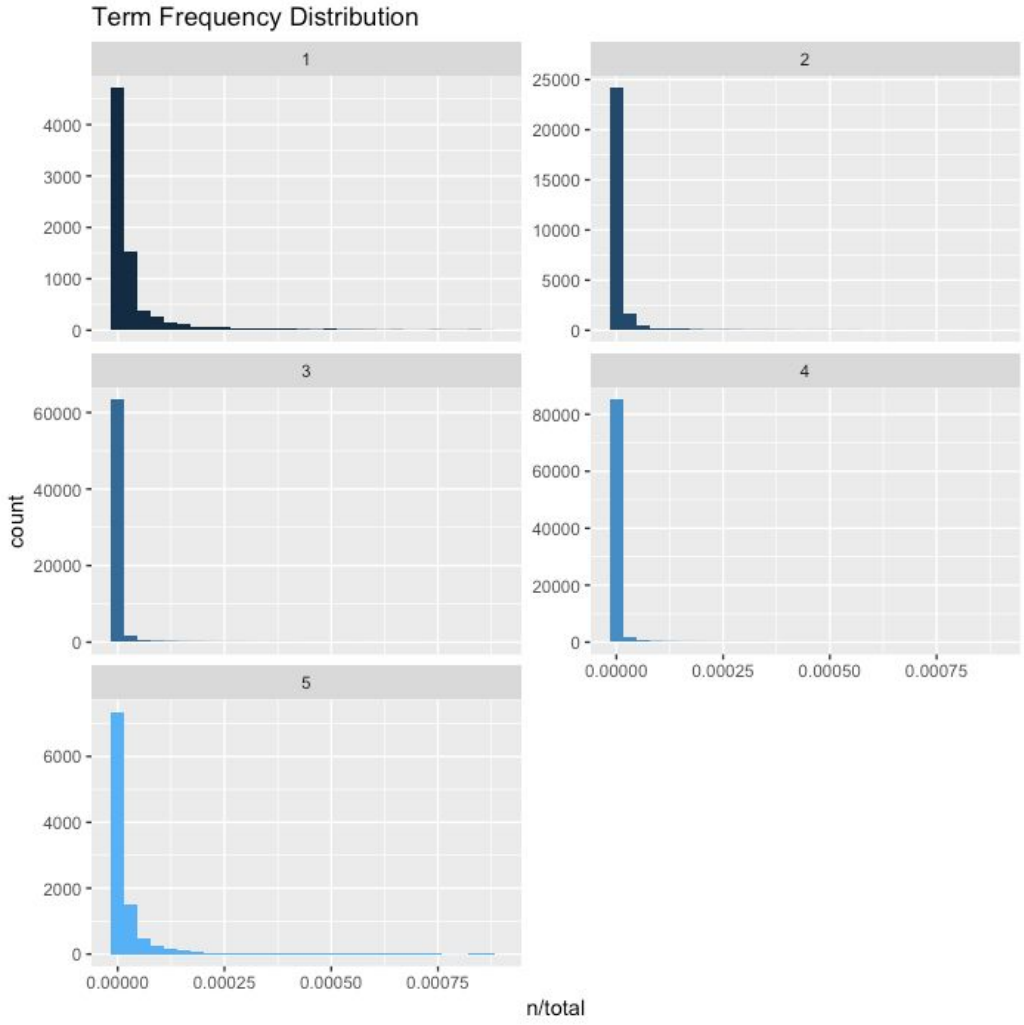
Some of the above insights can be illustrated through a word cloud portraying the words that contribute to negative and positive sentiment the most. The graph below illustrates such words, whose size is proportional to their frequency in the reviews by sentiment (sizes not comparable across sentiments). Negative sentiments are again driven by words that can be traced to service and food ("hard, cold, bad"). Positive words this time point to different aspects of a restaurant experience, including food, service, location, and potentially price ("fair"). These categories are definitely important and often used in the evaluation of restaurants, such as in a famous Italian television hospitality show for example (Borghese 2015).

We now explore net sentiment by star category to understand how sentiment changes across different star groups. Although customers are required to give a star rating when leaving a review for a restaurant, the businesses were segmented using their official star rating on Yelp. This precaution was taken not to introduce any bias. The aim of text analysis is to identify differences in review content among the different star ratings so to understand what makes a business successful or not. Business success is related to the cumulative rating on Yelp, not simply to the one of a single review.

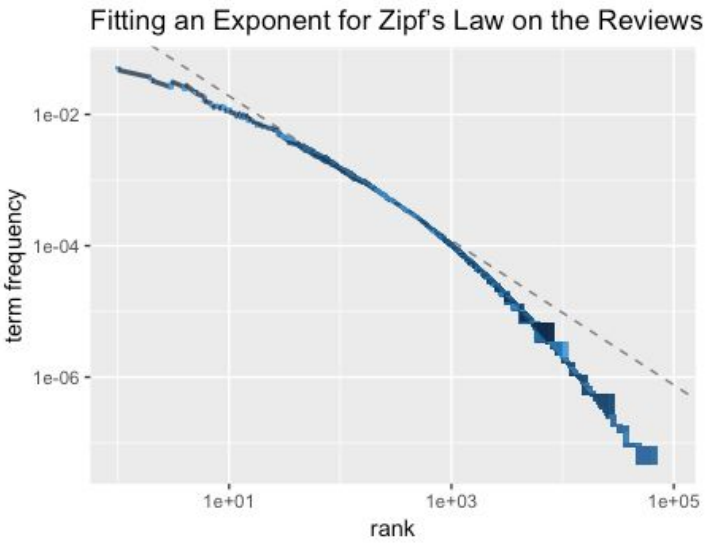
The graphs below plot sentiment scores for each star category. We use the reviewer number created when tidying the text as an index against which the sentiment was plotted. We can see that net sentiment is more positive in reviews for restaurants with 4 stars, while less positive in those for restaurants with 2. The sentiment for reviews for restaurants with 1 and 5 stars seem to be closer to the indifference line (the zero line), with a slightly negative direction for the 1 star. The small magnitude for the 1 and 5 star categories could be a consequence of the smaller number of reviews available for these restaurants.

We now divert our attention to overall term frequency. Firstly, we plot the number of times (n) a word appears in the reviews for a given star category, over the total number of words that appear in that category. The plots show a similar distribution for all star categories: an extremely high frequency for certain words (“the, and, a”) and a long tail to the right. The 1 and 5 star categories differentiate themselves slightly from the others for having a much longer right tail. This indicates a more varied

vocabulary employed in these reviews and strengthens the fact that consumers tend to be more vocal in extreme cases.



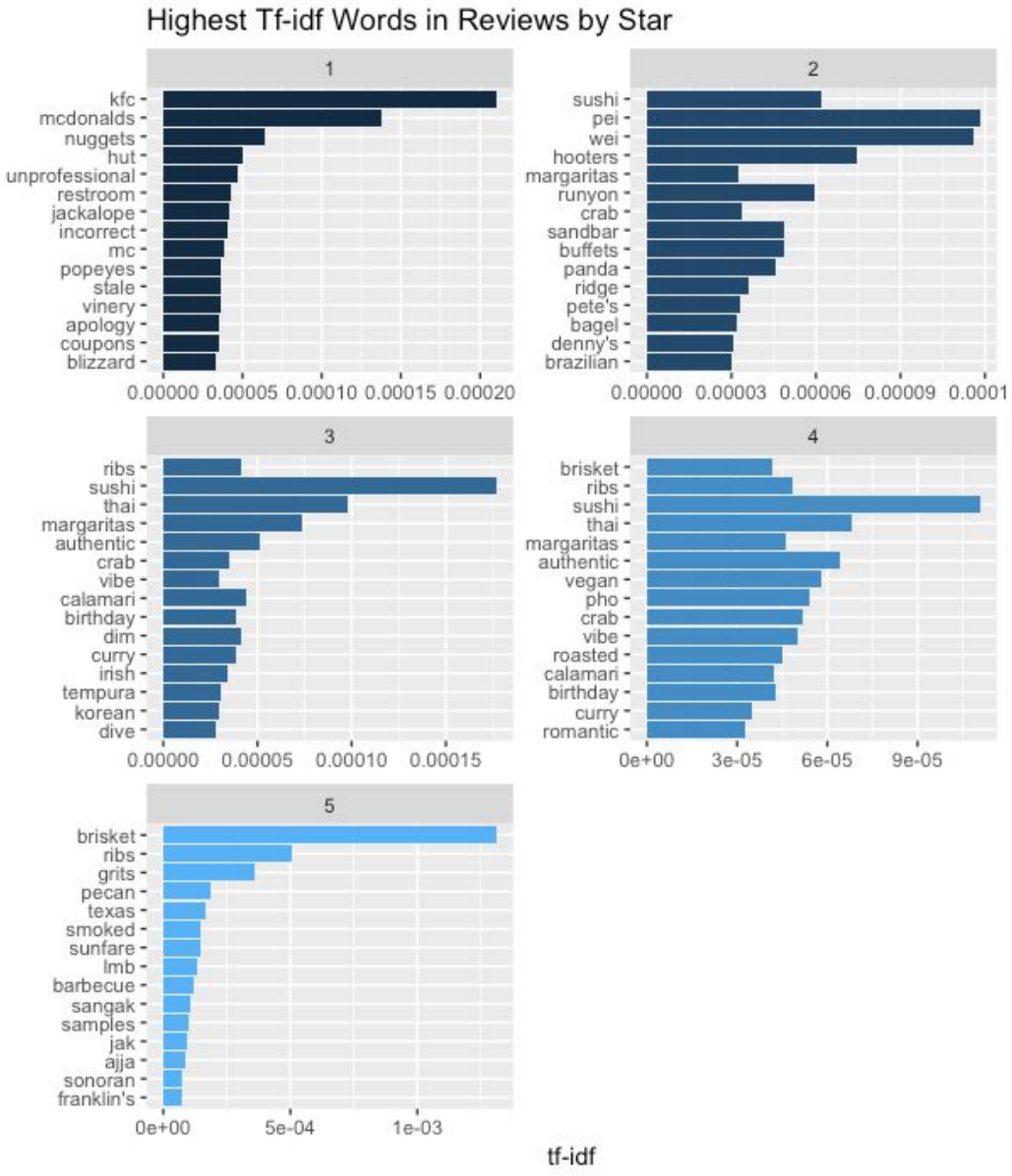
We utilise Zipf's law to confirm that a word's frequency is inversely proportional to its rank, where rank refers to the ordering of the words in a frequency table (Silge & Robinson 2019). We visualise Zipf's law by plotting the term frequency (y-axis) against the rank (x-axis) in a logarithmic scale. We notice that the relationship is very similar across all star ratings. The fitted power law below is similar to the findings of Zipf's law, with a slope coefficient close to -1, a perfect negative relationship. We also observe deviations at high ranks, which are expected given the simple and mono-thematic nature of the text. That is, we observe less rare words compared to the predictions of the power law. We also see some deviations at low ranks, which are less common. The reviews seem to be using less of the most common words than representative collections of text. This could again be driven by the mono-thematic nature of the reviews.



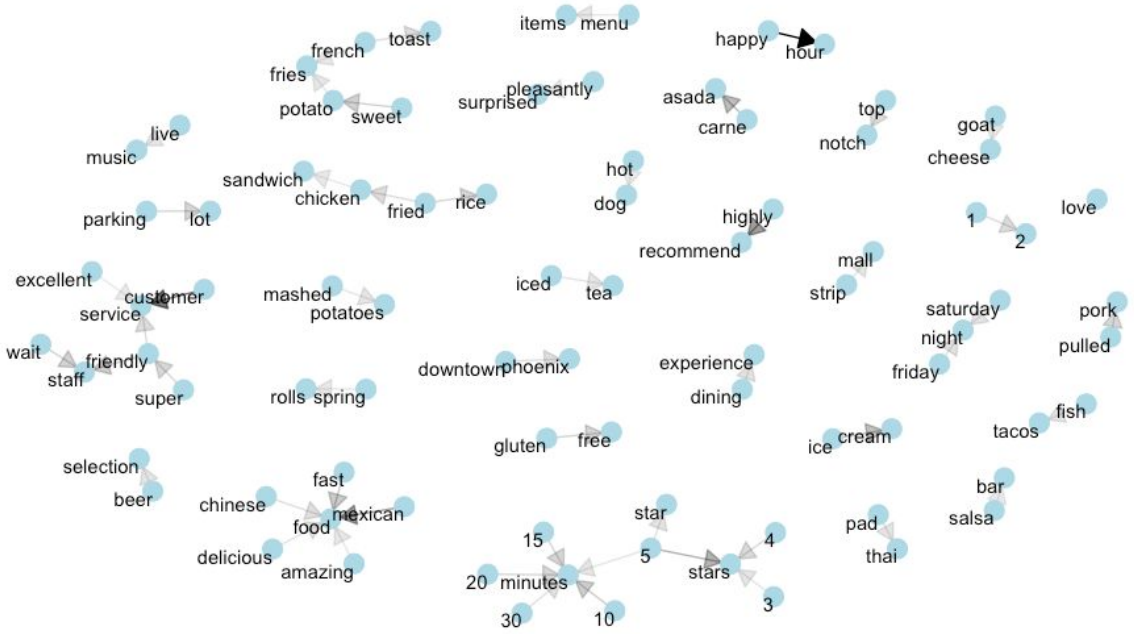
What are these most frequent words in the reviews? The word cloud below helps us answer this question through visual representation. The graph was plotted including only the non-stop words in the dataset. The size of the text is proportional to the frequency of the word in the reviews. Very frequent words refer to food ("food, pizza, pita, chicken"), while others refer to service ("wait, service"). This result strengthens our hypothesis that good food and service are factors that customers care about and on which they base their judgement of a restaurant.

Instead of plotting word frequency, the statistic of tf-idf is employed. Such measure multiplies together term frequencies with their inverse document frequency (idf) in order to analyse terms frequency adjusted for how common such word is (Silge & Robinson 2019). The advantage of doing so is to identify the most frequent, and thus important words in the reviews, but that are not too common. We hope to uncover deeper differences across star ratings by doing so. For the most part, the words refer to food or restaurant types. However, "unprofessional, incorrect, apology" emerge as important words in the 1 star

restaurants reviews. This suggests that restaurants with poor service are unsuccessful, while restaurants with a good atmosphere (“vibe, romantic” from 4 star restaurants’ reviews) are successful.



To conclude the text analysis, networks of the most frequent bigrams (groups of two related words) were plotted. The plot shows the most common relationships across words, where arrows indicate the sequence of the relationship, with a darker colour the more established the order of the words in the reviews is. Several words make up different bigrams and many refer to typical food combinations, such as “french fries” or “french toast”. Of greater significance is the frequent reference to time (“5- 10- 15- 20- 30-minutes”), service and staff (“customer-, friendly-service/staff”), and atmosphere (“live-music, dining-experience”). This result corroborates customer attention to food, service and atmosphere as the main elements on which a dining experience is judged.



CONCLUSION

This project aimed at uncovering the reasons for business success. This was done by combining statistical analysis of Yelp business data and text analysis of Yelp business reviews. A subset of the data including 3,608 restaurants in Phoenix and over 330,000 reviews was used for the analysis, but we predict reasons for success will be similar for restaurants in different regions and further analysis can confirm this. The code can also be easily modified to include different business categories and/or regions. We find that location, parking facilities, and review count are the most significant variables in predicting whether a business is open. Although the prediction has a relatively high misclassification rate and some of the parking data was transformed, the importance of location seems crucial in defining success for restaurants in Phoenix. Text analysis shifts the attention to the customer evaluation of the dining experience, where timely and friendly service, good food and a pleasant atmosphere are the unconditional factors that make a business successful. Phoenix restaurant owners should devote their attention to the above points to be worthy of a 5 star rating.

However, these findings are bounded by the data available, which did not include many critical business characteristics and business financial data. Macroeconomic effects, such as the 2008 economic crisis, and the belonging to restaurant chains could also help explain business success/failure, but were not taken into account here. The analysis is subject to the definition given to business success; an open business with at least a 4 star rating. Of course, business success can be defined in several other ways and many successful businesses do not have a good online presence. The focus on reviews could also be subject to the “extremes” mentioned in the analysis. Further analysis is necessary to investigate whether the reasons for restaurant success are indeed similar in other regions and to include different business categories, preferably using greater computing power than the one we had for this exercise. Heat maps, boosting, and exhaustive bigram analysis are some techniques that could be developed to

further enhance the understanding of the topic. In order to gain better insights into business success, a comprehensive dataset of business attributes would be preferred. While we did have access to such data, most of it consisted of NA values. This highlights the importance of having access to good and complete data and strengthens the need for thorough data cleaning and preparation before any analysis.

BIBLIOGRAPHY

Anderson, Michael and Jeremy Magruder “Learning from the crowd: regression discontinuity estimates of the effects of an online review database.” The Economic Journal, Vol. 122, No. 563 (September 2012), pp. 957-989.

Borghese, Alessandro. “4 Ristoranti.” Dry Media.

Isaack, Thomas S. “Organization Theory--Business Success Depends on It.” The Journal of the Academy of Management, Vol. 1, No. 3 (Dec., 1958), pp. 29-36.

Kalnins, Arturs and Kyle J. Mayer. “Franchising, Ownership, and Experience: A Study of Pizza Restaurant Survival.” Management Science, Vol. 50, No. 12 (Dec., 2004), pp. 1716-1728.

M. Robb, Alicia and Robert W. Fairlie. “Determinants of Business Success: An Examination of Asian-Owned Businesses in the USA.” Journal of Population Economics, Vol. 22, No. 4 (Aug., 2009), pp. 827-858.

Rhodes, Chris. “Business Statistics.” House of Commons Library, Number 06152, 12 December 2018. https://researchbriefings.files.parliament.uk/documents/SN06152/SN06152.pdf

Silge, Julia and David Robinson. “Text Mining With R.” O'Reilly, 2019-02-10. https://www.tidytextmining.com/

Topographic-map.com. “Phoenix.” http://en-gb.topographic-map.com/places/Phoenix-9209794

Yelp, Inc. “Yelp Dataset.” Kaggle. https://www.kaggle.com/yelp-dataset/yelp-dataset/version/6

Youn, Hyewon and Zheng Gu, Tourism and Hospitality. “Predict US restaurant firm failures: The artificial neural network model versus logistic regression model.” Tourism and Hospitality Research, Vol. 10, No. 3 (July 2010), pp. 171-187.