

University of New Hampshire
University of New Hampshire Scholars' Repository

Honors Theses and Capstones

Student Scholarship


Fall 2017

Predicting Yelp Food Establishment Ratings Based on Business Attributes

Peter Mark Shellenberger Jr.

University of New Hampshire, Durham, pms2000@wildcats.unh.edu

Follow this and additional works at: <https://scholars.unh.edu/honors>

 Part of the [Business Administration, Management, and Operations Commons](#), [Business Intelligence Commons](#), and the [Entrepreneurial and Small Business Operations Commons](#)

Recommended Citation

Shellenberger, Peter Mark Jr., "Predicting Yelp Food Establishment Ratings Based on Business Attributes" (2017). *Honors Theses and Capstones*. 374.

<https://scholars.unh.edu/honors/374>

This Senior Honors Thesis is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Honors Theses and Capstones by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.

Predicting Yelp Food Establishment Ratings Based on Business Attributes

Prepared by: Peter Shellenberger

Advised by: Professor Kholekile Gwebu

Fall 2017

Introduction

Social networking and business review sites play an integral role in the buying behavior of the modern consumer. An individual with access to the internet may observe millions of reviews on almost every type of product and service on the market. Websites such as Angie's List, Facebook, Yellow Pages, and Yelp all facilitate this process, acting as crowd-sourced hubs for review data. This has major implications for the modern business, as past research shows that reviews and ratings play a direct role on the demand a business receives. With the vast dissemination of review data to the potential buyer market, it is imperative for businesses to focus on receiving high reviews from all of their customers. Hence, one important question for many businesses is whether there is a way to encourage higher ratings through targeted alterations to their offerings.

This study seeks to create models that predict business ratings based on various business specific attributes. The study will focus on food establishments in the United States and analyze data provided by Yelp through their 9th Dataset Challenge. With 4.1 million reviews, 947,000 tips, 1 million users and 144k businesses, the Yelp dataset is large, robust, and permits the development of a testing sample that accurately portrays the average rating behavior of consumers across the country. The attributes that the model will be based on will include those that individuals note in their reviews.

The discovery of an accurate model to predict ratings based on Yelp reviews would be beneficial to food establishment owners and consumers alike. With access to an accurate predictive model, an owner would be able to tailor the features of their business to match those seen as promoting positive reviews within the model. In addition to helping raise review ratings,

alteration of features would provide the average customer with a positive experience. Using insights from the models generated here, business owners will be able to determine the features that carry the most influence in business ratings. Business owners would then be able to add or remove attributes that statistically tend to increase average review ratings.

The remainder of this study is organized as follows. The next section reviews related literature. These articles include studies involving both numeric and sentiment data analysis for predicting review ratings. They also include studies relating the impact of reviews on business demand. Next, hypotheses are developed. Thereafter, the methodology used to collect, clean, and analyze the data is presented. The methodology section also includes a description and classification of variables used in analysis, along with other data descriptors. Subsequently, the results of analysis are presented. Finally, the study concludes with a section that explains the findings and discusses their implications, followed by commentary on recommendations for future research.

Literature Review

This section reviews prior studies that have examined review data, particularly the Yelp dataset, to create predictive models associated with user behavior and preferences. While some studies also create models based on review attributes, many look at review text for script analysis purposes, behavior and stats relating to the user, geographical location in relation to the user and business, and specific qualities listed about the business. Table 1 summarizes the literature on user reviews.

As shown in Table 1, two studies (Mathieu et al. 2016; Sunil et al. 2017) focus on the physical position on the location of the business. Mathieu et al. (2016) create an application that predicts the chance of success for a business based on its location and type. To accomplish this, they gather data from Yelp and OpenStreetMap. Sunil et al. (2017) analyze the correlation between the distance from a business to a user’s “activity center” (the geographic location where the user had the most review activity) and the rating the user provided, anticipating higher ratings for greater distance. In addition, they looked at the relationship between users listed as friends and their behavior in relating similar items.

Summary of Referenced Texts				
Title	Authors	Journal	Focus	Summary
Uncovering Business Opportunities from Yelp and Open Street Map Data	Mathieu, Grillet, Passerini, Tiwari (2016)	Technische Universität Berlin	Location; Business Analysis	The authors created a model that would predict the success of a business based on business type and location. From Yelp, aspects such as ratings and check-ins were utilized, while mapping data was drawn from OpenStreetMap. Data mining and machine learning algorithms were utilized in creating potential models. A user-interface was successfully developed to identify business opportunities on a map, but it was difficult to draw accuracy from the Yelp data. Still, further iterations, using various modeling methods, drew better results.
Prediction of Rating by Using Users’ Geographical Social Factors	Sunil, M, Bari, Shetty (2017)	International Journal of Soft Computing and Engineering	Location; User Analysis	The authors performed analysis on the Yelp dataset to identify a correlation between rating and the social factors of a user’s physical location. They worked to identify a relationship between rating and the distance a user is from their home; they also investigate the connection between the ratings of similar items by users that are friends. They initially inferred that users

				tend to rate items higher the further away they are from their home, or “activity center”. They also noted that friends tend to rate similar items similarly, regardless of geographic distance.
User Modeling with Neural Network for Review Rating Prediction	Tang, Qin, Liu, Yang (2015)	International Joint Conferences on Artificial Intelligence Organization (Proceedings)	Text; User Analysis	The authors developed a neural network for review rating prediction based on review text, as well as information about the user. The potential interpretation of a word by user is assessed through a "user-word composition vector model". Results of the model indicate that the incorporation of user data when analyzing review text is more effective than several models that strictly consider the textural elements.
Predicting the Helpfulness of Online Reviews Using a Scripts-Enriched Text Regression Model	Ngo-Ye, Sinha, Sen (2017)	Expert Systems with Applications: An International Journal	Text Analysis	The authors aimed to utilize script analysis to predict helpfulness rating of online reviews. Human annotators were asked to highlight important phrases of reviews; phrases were added to a script lexicon. Text regression was performed to predict review helpfulness based on words in the lexicon. Compared against a Baseline model and a bag-of-words model, the scripts enriched model was found to predict helpful reviews quicker and more accurately.
Predicting Yelp Restaurant Reviews	Farhan (2014)	UCSD Jacobs School of Engineering: Computer Science and Engineering	Review Attribute Analysis	The author developed a linear regression model to predict business reviews based on restaurant attributes. Classification was conducted using Naive Bayes and neural networks, while regression testing was conducted through random forest and linear regression. The linear regression model was found to be the most accurate, but only worked a small amount better than taking the average, and would not serve well in a production environment.
Recommendation	Jayasimhan, Rai,	International	User;	The authors worked to create a

System for Restaurants	Parekh, Patwardhan (2017)	Journal of Computer Applications	Review Attribute Analysis	software application that would effectively recommend a restaurant to a user. The software utilized machine learning techniques and algorithms to establish recommendations based on data from the Yelp dataset. Models employed included a linear support vector machine (SVM), SVM with radial basis function (RBF) Kernel, and a decision tree.
Reviews, Reputation, and Revenue: The Case of Yelp.om	Luca (2016)	HBS Working Paper Series	Review Rating and Correlation	The author attempted to identify a significant connection between Yelp reviews and restaurant revenue. Ultimately, the author identified a positive correlation between Yelp reviews and revenue of independent restaurants. The author also noted a specific lack of correlation between Yelp reviews and the revenue of chain restaurants.

Table 1: Summary of related studies.

Other studies have focused on review text to develop their models. For instance, the study by Tang et al. (2015) focuses on the development of a neural network to predict review rating based on review text. To account for the difference between how users interpret different words, Tang et al. (2015) create a “user-word composition vector model”. Ngo-Ye et al. (2017) try to predict the helpfulness rating of a review based on review text in paper. Their approach involves the use of human annotators to highlight phrases of text reviews, then adding these highlighted phrases to a “text lexicon”. Regression is then performed on the finalized lexicon in order to create the predictive model.

The studies by Farhan (2014) and Jayasimhan et al. (2017) bear some resemblance to this study because they develop predictive models based on attributes listed for a business. The paper by Farhan (2014) creates a regression model to predict business rating by assigned business attributes. Classification was conducted using Naive Bayes and neural networks, while

regression testing was conducted through random forest and linear regression. Jayasimhan et al. (2017) carried a similar focus on business attributes with the addition of user characteristics. Unlike the models used by Farhan (2014), Jayasimhan (2017) used a linear support vector machine (SVM), SVM with radial basis function (RBF) Kernel, and a decision tree as test predictive models in a software meant to effectively recommend restaurants to users.

Finally, Luca (2016) focuses on the identification of a positive correlation between business rating on Yelp and business revenue. This paper is particularly important in justifying the research in this study, as it reflects a significant purpose for the creation of a model that may predict how businesses may achieve the highest business rating through alterations to their business attributes.

While the papers discussed above encompass predictive models based on analysis of the Yelp dataset, almost all of them seek to predict a different value, utilizing different factors and different models than those used in our piece. The study by Farhan (2014) exhibits the most similar approach to the current study because it creates predictive models for business rating based on business attributes. However, the methodology used for prediction is different. The current study uses linear regression over two categories of business attributes, compared across several states, to attain an accurate prediction. Additionally, this study does not solely focus on the accuracy of the entire models, but attempt to draw additional value from attributes that consistently have significant influence within models.

Hypothesis

If a model that calculates business rating based on business features listed in Yelp.com can be created, then the model can be used to significantly predict the ratings another establishment will receive, because the features of a business have the greatest impact on how a business is rated. It is expected that the model will not be accurate in predicting ratings for every geographic region across the United States, as local cultures will likely have different tastes for certain business features. We expect physical amenities, such as whether the business has a parking lot or serves alcohol, will have more of an impact than atmospheric variables, such as whether or not the business is romantic or casual.

Methodology

Data was first collected from the .json file provided by the Yelp.com page for the dataset challenge. This file was initially opened in RStudio for analysis. Unfortunately, technical limitations of the equipment being used for processing was too limited to handle the entirety of the dataset. Specifically, the computers used were not able to convert the entire reviews section of the dataset into a dataframe in R without RStudio crashing. It was determined that the computing capabilities on hand would only be able to analyze a sample of the dataset for purposes of this project, leaving room for follow-up analysis in the future (given that better computing equipment would become available).

The Yelp data used in this study originally contained 4.1 million reviews. To draw the sample for analysis, first all non-US reviews were eliminated from the sample. Thereafter the sample was narrowed down to 77,168 reviews with 54 total variables to facilitate analysis.

Seven of the initial variables (user_id, business_id.business, attributes, state, stars.business, stars.review, and name.user) were selected for analysis. Since sentiment analysis was not conducted in this study data, the review and tip variables were eliminated from the dataset. In making this choice, it was noticed that the rows were not unique by business or user, but instead by tips (text.tip; business_id.tip). To eliminate the possibility of an unequal number of tips influencing the weighting of scores for each business, the repetition of business_id.business within the dataset was eliminated. Additionally, the dataset was filtered to only include records belonging to the “Food” and “Restaurants” categories. This narrowed the dataset considerably, bringing it down from the initial 77,167 records to 2265.

After narrowing the data set, it was necessary to decide a way to access the data in the attributes section. The attributes variable possessed a sort of listing of data. In summary, certain attributes such as alcohol, were split up into types. For instance, the variable, “Alcohol”, would have the type, “none”, “beer_and_wine”, or “full_bar”. Some variables, such as “Ambience”, had subtypes, like “hipster”, “divey”, and “trendy”. Subtypes listed within these variables had binary options, true or false. The overall list of these variables was contained within a single cell. To extract these data points and format them in a way that was more feasible for analysis, it was necessary to create an individual column for each variable (or in the instance of variables with subtypes, only creating columns for the subtypes). Each column would portray the variables as either having or lacking the certain attribute in a binary format (0 or 1). To accomplish this, the COUNTIF function was used, inputting a one in the column if the corresponding attributes variable contained appropriate text signaling possession of said attribute. For example, cell, L2, in the column acknowledging whether the establishment was noted for having a divey ambience, IsDivey, possessed the formula “=COUNTIF(C2, “*’divey’: True*)”, where C2 was the

corresponding cell in the column for Attributes. After conducting this for each attribute of interest, 39 variables remained for analysis.

The analysis began with the use of the 39 variables. User_id and business_id.business are the unique identifiers for Yelp user accounts and business on Yelp, respectively. State is the state where the business was located, star.business is the average star rating the business received, and stars.review is the average star rating the review itself had received. Star.business is a numeric variable on a scale of one to five with one-half point increments, while star.rating is a numeric variable with one point increments. Name.user is the name of the reviewer. IsRomantic, IsIntimate, IsClassy, IsHipster, IsDivey, IsTouristy, IsTrendy, IsUpscale, IsCasual, IsGoodforKids, and GoodforGroups are all characteristics the user may have picked to describe the atmosphere of the business. These variables are all binary in this analysis, meaning that a business may or may not possess any one of these traits without any middleground. AcceptsCreditCards, HasGarage, StreetParking, IsValidated, HasParkingLot, HasValet, DoesCater, HasDessert, HasLateNight, HasLunch, HasDinner, HasBreakfast, HasBrunch, HasTV, HasOutdoorSeating, AllowsReservations, HasTableService, WheelchairAccessible, and NoWiFi are amenities that the reviewer has acknowledged the business possesses. Like the “atmospheric” variables, these are all binary. Again, for purposes of this analysis, all binary variables are listed as “0” if a business does not possess said feature, while they are listed as “1” if they do. Table 2 provides a description of each variable, while Table 3 provides summary statistics of the sample regarding variable frequency.

Variables Used in Study	
Variable	Description
stars.business	The star rating (1-5, at .5 point increments) a user gives a restaurant in a review.
stars.review	The star rating (1-5, at 1 point increments) given to a review by other users to reflect the review's quality.
IsRomantic	A binary variable that reflects whether or not the user listed the business as having a romantic atmosphere (1=True; 0=False).

IsIntimate	A binary variable that reflects whether or not the user listed the business as having an intimate atmosphere (1=True; 0=False).
IsClassy	A binary variable that reflects whether or not the user listed the business as having a classy atmosphere (1=True; 0=False).
IsHipster	A binary variable that reflects whether or not the user listed the business as having a hipster atmosphere (1=True; 0=False).
IsDivey	A binary variable that reflects whether or not the user listed the business as having a divey atmosphere (1=True; 0=False).
IsTouristy	A binary variable that reflects whether or not the user listed the business as having a touristy atmosphere (1=True; 0=False).
IsTrendy	A binary variable that reflects whether or not the user listed the business as having a trendy atmosphere (1=True; 0=False).
IsUpscale	A binary variable that reflects whether or not the user listed the business as having an upscale atmosphere (1=True; 0=False).
IsCasual	A binary variable that reflects whether or not the user listed the business as having a casual atmosphere (1=True; 0=False).
AcceptsCreditCards	A binary variable that reflects whether or not a business accepts credit cards for payment (1=True; 0=False).
HasGarage	A binary variable that reflects whether or not a business offers a garage for parking (1=True; 0=False).
StreetParking	A binary variable that reflects whether or not a business requires street parking (1=True; 0=False).
IsValidated	A binary variable that reflects whether or not a business offers validated parking (1=True; 0=False).
HasParkingLot	A binary variable that reflects whether or not a business offers a parking lot (1=True; 0=False).
HasValet	A binary variable that reflects whether or not a business has valet parking (1=True; 0=False).
DoesCater	A binary variable that reflects whether or not a business offers a catering service (1=True; 0=False).
IsGoodforKids	A binary variable that reflects whether or not a business provides a good environment for kids (1=True; 0=False).
HasDessert	A binary variable that reflects whether or not a business offers a dessert menu (1=True; 0=False).
HasLateNight	A binary variable that reflects whether or not a business offers a late night menu (1=True; 0=False).
HasLunch	A binary variable that reflects whether or not a business offers a lunch menu (1=True; 0=False).
HasDinner	A binary variable that reflects whether or not a business offers a dinner menu (1=True; 0=False).
HasBreakfast	A binary variable that reflects whether or not a business offers a breakfast menu (1=True; 0=False).
HasBrunch	A binary variable that reflects whether or not a business offers a brunch menu (1=True; 0=False).
HasTV	A binary variable that reflects whether or not a business possess a television (1=True; 0=False).
HasOutdoorSeating	A binary variable that reflects whether or not a business offers outdoor seating (1=True; 0=False).
GoodForGroups	A binary variable that reflects whether or not a business provides a good environment for groups (1=True; 0=False).
AllowsReservations	A binary variable that reflects whether or not a business allows reservations for seating (1=True; 0=False).
HasTableService	A binary variable that reflects whether or not a business offers table service for dining (1=True; 0=False).

WheelchairAccessible	A binary variable that reflects whether or not a business is accessible by wheelchair (1=True; 0=False).
NoWiFi	A binary variable that reflects whether or not a business lacks a WiFi connection (1=True; 0=False).
NoAlcohol	A binary variable that reflects whether or not a business does not provide alcohol (1=True; 0=False).

Table 2: Description of each variable used for analysis in study.

Sample Characteristics		
State	Frequency	Percent
Arizona	758	33.47%
Illinois	9	0.40%
North Carolina	259	11.43%
Nevada	952	42.03%
Ohio	97	4.28%
Pennsylvania	123	5.43%
South Carolina	26	1.15%
Wisconsin	41	1.81%
Variable (Value=1)		
IsRomantic	23	1.02%
IsIntimate	22	0.97%
IsClassy	64	2.83%
IsHipster	62	2.74%
IsDivey	51	2.25%
IsTouristy	12	0.53%
IsTrendy	157	6.93%
IsUpscale	23	1.02%
IsCasual	1339	59.12%
AcceptsCreditCards	2224	98.19%
HasGarage	268	11.83%
StreetParking	353	15.58%
IsValidated	13	0.57%
HasParkingLot	1477	65.21%
HasValet	152	6.71%
DoesCater	947	41.81%
IsGoodforKids	1603	70.77%
HasDessert	120	5.30%
HasLateNight	184	8.12%
HasLunch	1176	51.92%
HasDinner	1216	53.69%
HasBreakfast	240	10.60%
HasBrunch	239	10.55%
HasTV	1206	53.25%

HasOutdoorSeating	974	43.00%
GoodForGroups	1824	80.53%
AllowsReservations	812	35.85%
HasTableService	1372	60.57%
WheelchairAccessible	1543	68.12%
NoWiFi	1011	44.64%
NoAlcohol	642	28.34%

Table 3: Characteristics of sample based on variable frequencies.

Data analysis, after cleansing using Excel, was conducted with IBM SPSS. The methods used to analyze the data included correlation analysis, independent sample t-tests, and multiple regression analyses.

The first of these tests conducted was the correlation analysis. All variables previously listed, excluding user_id, business_id.business, state, and name.user were part of the sample. After correlation analysis was conducted, results were observed in order to detect variables that had a statistically significant impact on the star.business variable. Statistically significant results were considered to be those that were found to have p-values less than or equal to 0.05 (two-tailed). Next, a t-test was conducted with each of the variables found to be significant from correlation analysis. In each t-test conducted, star.business was used as the dependent variable. Independent samples t-tests were conducted to explore whether there are any statistically significant differences between different variables. Data on these t-tests is provided in Appendix A. Finally, we used linear regression to build a predictive model.

Since the business rating was the variable we were trying to predict, star.business is the model's dependent variable. To create a more accurate predictive formula, two different models which were subsequently named Model 1 and Model 2 were created. Model 1 includes mostly amenity attributes while Model 2 includes ambience attributes. To test accuracy of the models across different geographic segments, we create separate regression models for each location Arizona, Nevada, North Carolina, and Pennsylvania in the data set. Pertinent statistics from

regression are reported in the tables presented in the next section, including beta values, standard error coefficients, and p-value. Significance (p-value) was noted for those variables with p-values below 0.1, 0.05, 0.01 and 0.001.

Results

Using the Yelp dataset, the study attempts to create a model that accurately predicts business rating for food establishments based on business attributes. Additionally, it aims to find attributes that have a significant correlation to business rating. The analysis began with testing correlation between business attributes and business rating. Table 4 presents the correlation between the variables. Based on this matrix, it is evident that stars.review, IsRomantic, IsIntimate, IsClassy, IsHipster, IsDivey, IsTrendy, IsUpscale, StreetParking, HasParkingLot, HasValet, DoesCater, HasDessert, HasLunch, HasDinner, HasBrunch, AllowsReservations, WheelchairAccessible, NoAlcohol are positively correlated with the star rating. The variables IsTouristy, IsCasual, AcceptsCreditCards, HasGarage, IsValidated, IsGoodforKids, HasLateNight, HasBreakfast, HasTV, HasOutdoorSeating, GoodForGroups, HasTableService, NoWiFi are negatively correlated with star rating.

Attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33		
1. stars/business	1	.468	.098	.071	0.036	.093	0.010	-0.040	.049	0.022	-0.017	-.082	-.064	.120	-0.003	.120	0.017	.110	-.089	.042	-0.011	.053	-0.019	.060	-.077	-0.014	-.082	.048	-.043	.125	-0.002	.065			
2. stars/review	.468	1	0.019	0.028	0.035	.074	-0.037	-0.024	0.010	0.008	0.001	-0.031	-.079	.078	-0.012	.067	-0.011	.056	-0.030	0.020	-0.084	0.018	.047	-0.030	-0.040	-0.045	-0.005	-0.040	-0.040	0.003	.051				
3. IsRomantic	.098	0.019	1	.293	.142	-0.017	-0.015	-0.007	-0.028	.078	-.122	0.014	0.017	0.006	.051	-0.009	.078	-0.006	-0.100	-0.004	-0.030	-.096	.094	-0.035	-0.036	-0.002	0.010	0.039	.128	.082	0.041	0.033			
4. Ismeme	.071	0.026	.293	1	0.037	0.011	-0.015	-0.007	-0.027	.080	-.110	-0.020	0.006	.082	-0.008	-.091	.045	-0.038	-.124	0.017	0.004	-0.076	.074	-0.034	-0.034	-0.024	0.014	0.037	.095	.080	0.010	.047	-.052		
5. IsClassy	0.036	0.035	.142	0.037	1	-0.029	-0.026	0.024	-0.028	.195	-.183	0.023	.086	0.037	.093	.274	-0.026	-.195	-0.017	-0.031	-.135	.105	-.041	-0.024	0.021	.051	.077	.211	.121	.059	0.013	-.107			
6. IsHesher	.093	.074	-0.017	0.011	-0.029	1	-0.007	-0.012	0.039	-0.017	-.081	-0.038	-0.011	.162	-0.013	-.048	.052	0.017	-0.005	.057	0.039	-0.023	-0.040	.083	.101	-.043	.108	.066	-0.018	-0.020	0.022	-0.083	-0.021		
7. IsDivey	0.010	-0.037	-0.015	-0.015	-0.026	-0.007	1	-0.011	-.041	-0.015	-.156	-0.024	-0.037	0.028	-0.012	0.017	-0.041	0.010	.045	.044	0.031	0.039	-0.014	0.035	-0.004	0.023	-0.067	0.037	-.051	-0.012	0.008	.097	.083		
8. IsFronzy	-0.040	-0.024	-0.007	-0.007	0.024	-0.012	-0.011	1	0.004	-0.007	-.063	0.010	.124	-0.031	-0.008	-.082	0.029	-0.025	0.034	.064	0.001	-0.015	-0.018	0.034	.054	0.007	-0.014	0.021	.047	0.034	0.024	0.008	-0.019		
9. IsTrendy	.046	0.010	-0.028	-0.027	-0.026	0.039	-0.041	0.004	1	-0.010	-.219	0.037	.137	.084	-.094	-.115	.170	0.012	-.115	0.013	.078	-.096	.124	-.071	0.008	-.102	.135	.117	.195	.156	0.030	-0.028	-.135		
10. IsUpscale	0.022	0.008	.078	.080	.195	-0.017	-0.015	-0.007	-0.010	1	-.113	0.014	.072	-0.019	.051	-.083	.166	-.050	-.119	-0.024	-0.030	-.088	.085	-0.035	-0.020	-0.002	-0.017	0.039	.135	.082	0.031	.042	-.064		
11. IsCasual	-0.017	0.001	-.122	-.110	-.183	-.081	-.156	-.083	-.219	-.113	1	.103	-0.040	-.056	-.044	.296	-.147	.321	.565	0.032	.096	.541	.337	.155	.113	.320	.145	.455	.051	.298	.123	.174	.117		
12. AcceptsCreditCards	-.082	-0.031	0.014	-0.020	0.023	-0.038	-0.024	0.010	0.037	0.014	.103	1	0.039	-0.015	0.010	.068	0.023	.075	.066	0.017	0.004	.068	.080	0.004	0.014	.112	.044	.151	.095	.107	.056	0.029	-0.025		
13. HasGunge	-.064	-.079	0.017	0.036	.088	-0.011	-0.037	.124	.137	.072	-0.040	0.039	1	0.005	.117	-.401	.355	-.136	-.089	.086	0.016	-0.072	.086	0.016	-0.072	.086	0.016	.043	0.031	-0.037	.070	.159	.145	.082	-0.028
14. StreetParking	.120	.078	0.006	.082	0.037	.152	0.025	-0.031	.094	-0.019	-.066	-0.015	0.005	1	.048	-.363	0.031	0.008	-.155	-0.009	.073	-.047	0.001	-0.033	-0.009	-0.002	.119	0.002	.049	0.036	0.017	-0.080	-0.043		
15. IsValedict	-0.003	-0.012	.051	-0.008	.093	-0.013	-0.012	-0.008	.054	.051	-.044	0.010	.117	.048	1	-.092	.143	-0.017	-.093	-0.018	-0.001	-.056	0.035	-0.007	-0.026	0.038	0.028	0.037	.077	.061	.052	-0.033	-0.035		
16. HasParkingLot	.120	.067	-0.009	-.091	-.096	-.048	0.017	-.082	-.115	-.083	.296	.086	-.401	-.383	-.092	1	-.167	.194	-.160	-0.034	-0.014	.202	.154	0.032	0.006	.096	.078	.046	-0.015	0.016	.252	0.039	.050		
17. HasValet	0.017	-0.011	.078	.045	.274	.052	-0.041	0.029	.170	.166	-.147	0.023	.355	0.031	.143	-.167	1	-.045	-.185	0.015	-0.002	-.145	.146	-.052	-0.012	.057	.101	.230	.166	.138	-0.038	-.157			
18. DoesCar	.110	.056	-0.006	-0.038	-0.026	0.017	0.010	-0.025	0.012	-.050	.321	.075	-.136	0.008	0.017	.194	-.045	1	.364	0.039	-.046	.336	.227	.207	.306	.236	.206	.206	.305	.193	.125	.128	0.026	.094	
19. IsGoodForKids	-.088	-0.030	-.107	-.124	-.195	-0.005	.045	0.034	.115	-.119	.565	.066	-.089	-.155	-.093	.160	-.185	.354	1	.091	-0.026	.542	.207	.208	.151	.246	.080	.522	0.005	.197	-0.021	.218	.294		
20. HasDessert	.042	0.020	-0.004	0.017	.057	.044	.064	0.13	0.024	0.032	0.017	.066	-0.008	-0.018	-0.034	0.015	0.039	.091	1	.080	0.028	0.010	.079	.060	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	.061		
21. HasLateNight	-.072	-.064	-0.030	0.004	-0.031	0.039	0.031	0.001	.078	-0.030	.096	0.004	0.016	.073	-0.001	-0.014	-0.002	-.046	-0.026	.080	1	-0.008	.095	.045	-0.013	.113	0.036	.125	0.000	.078	0.009	0.013	-0.018		
22. HasLunch	0.011	0.018	-.096	-.078	-.135	-0.023	0.039	-0.015	-.096	-.088	.541	.068	-.072	-.047	-.056	.202	-.145	.336	.542	0.028	-0.008	1	.308	0.007	-0.038	.235	.095	.366	-0.020	.153	.072	.187	.199		
23. HasDinner	.053	.047	.094	.074	.105	-0.040	-0.016	.124	.085	.337	.080	.066	.031	0.030	.154	.145	.227	.207	.207	0.010	.095	.306	1	-.253	-.191	.356	.113	.433	.416	.497	.229	.196	-.204		
24. HasBreakfast	-0.019	-0.030	-0.035	-0.034	-0.041	.083	0.035	0.034	-.071	-0.035	.155	0.004	0.016	-0.033	-0.007	0.032	-.052	-0.013	.208	.077	.045	0.017	-.253	1	.554	-.086	0.031	.108	-.159	0.014	.042	-0.041	.229		
25. HasBunch	.060	0.003	-0.035	-0.034	-0.004	.103	-.004	.067	0.008	0.008	.113	0.043	0.031	0.028	0.006	-0.012	.050	.151	.208	.077	.080	-0.013	-.038	-.191	.564	1	-.055	.079	.136	-0.035	.104	-0.046	-.060	.119	
26. HasTV	-.077	-.044	-0.002	-0.024	0.021	-.043	0.023	0.007	.102	-0.002	.320	.112	.043	-0.001	0.028	.096	.057	.236	.246	0.017	0.004	0.000	-0.020	.113	.235	.356	-.086	-.045	1	-.190	.442	.302	.428	-.105	-.196
27. HasOutdoorSeating	-0.014	0.021	0.010	0.014	.051	.105	-.066	-0.014	.135	-0.017	.145	.044	-0.037	.119	0.028	.078	0.027	.205	.060	0.029	0.036	.095	.113	0.031	.079	.190	1	.233	.119	.082	.067	-.123	-.060		
28. GoodGroups	-.082	-.045	0.039	0.037	.077	.095	0.037	0.021	.117	0.039	.455	.151	.070	0.020	0.037	.046	.101	.306	.522	.057	.126	.395	.433	.108	.136	.442	.233	1	.335	.518	.118	.179	0.040		
29. AlwaysReservations	.048	-0.005	.126	.095	.211	-0.018	-.051	.047	.195	.135	.051	.095	.145	.095	.049	.077	-0.015	.230	.193	0.005	0.010	0.000	-0.020	.416	.150	.497	0.014	.104	.428	.082	.515	.556	1	.185	-.327
30. HasTableService	-.043	-0.040	.082	.080	.121	-0.020	0.034	.156	.086	.298	.107	.145	.035	.061	.081	.018	.188	.125	.197	.125	.197	.125	.197	.125	.197	.125	.197	1	.185	-.106	-.327				
31. WheelchairAccessible	.125	0.028	0.041	0.010	-.199	0.022	0.008	0.024	0.030	0.031	.123	.056	.092	0.017	.062	.252	.136	.125	-.021	-0.029	0.009	.072	.229	.042	-0.006	.105	.087	.118	.233	.186	1	-.003	-.097		
32. NoWiFi	-0.002	0.003	0.033	.047	0.013	-.085	.097	0.008	-0.028	.042	.114	0.029	-0.028	-.060	-0.033	0.039	-0.038	0.028	.218	-.066	0.013	.187	.195	-0.041	-.060	-.056	-.123	.179	0.010	.106	-0.003	1	.172		
33. NoAlcohol	.055	.051	-.084	-.052	-.107	-.021	.083	-0.019	-.133	-.084	.117	-0.025	-.152	-.043	-0.035	.050	-.157	.094	.294	.081	-0.018	.199	-.204	.229	.119	-.196	-.080	0.040	-.327	-.097	-.172	1			

Pearson Correlation

Table 4: Correlation table of attributes used in analysis.

Next, two linear regression models were constructed, with one based on business amenities and the other based on business attributes. The variables used in linear regression modeling were limited to those identified with a statistically significant correlation to business rating after correlation analysis. Table 5a shows a model summary of our linear regression model, Model 1. This model was created through linear regression conducted upon the entirety of our dataset (2265 observations) using our amenity attributes.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.344	0.118	0.112	0.5805

Table 5a: Summary statistics for Model 1.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.948	0.094		41.924	0.000
	AcceptsCreditCards	-0.391	0.093	-0.085	-4.196	0.000
	HasGarage	0.007	0.044	0.003	0.150	0.881
	StreetParking	0.278	0.038	0.164	7.364	0.000
	HasParkingLot	0.205	0.033	0.159	6.144	0.000
	DoesCater	0.134	0.028	0.108	4.752	0.000
	IsGoodforKids	-0.167	0.036	-0.123	-4.645	0.000
	HasDessert	0.152	0.055	0.055	2.757	0.006
	HasLateNight	-0.159	0.046	-0.071	-3.473	0.001
	HasDinner	0.147	0.032	0.119	4.593	0.000
	HasBrunch	0.198	0.043	0.099	4.585	0.000
	HasTV	-0.071	0.029	-0.058	-2.415	0.016
	GoodForGroups	-0.140	0.045	-0.090	-3.103	0.002
	AllowsReservations	0.089	0.033	0.070	2.697	0.007
	HasTableService	-0.058	0.037	-0.046	-1.585	0.113
WheelchairAccessible	0.084	0.029	0.064	2.909	0.004	
NoAlcohol	0.126	0.032	0.092	3.903	0.000	

Table 5b: Linear regression coefficients of Model 1 attributes over entire sample.

Table 5b shows the coefficients of our linear regression model, Model 1. Based on Table 5b, it is evident that having street parking has a significant ($\beta = 0.278, p \leq 0.05$) positive impact on business review, along with possession of a parking lot ($\beta = 0.205, p \leq 0.05$), offering catering ($\beta = 0.134, p \leq 0.05$), offering dessert ($\beta = 0.152, p \leq 0.05$), offering a dinner menu ($\beta = 0.147, p \leq 0.05$), offering a brunch menu ($\beta = 0.198, p \leq 0.05$), allowing reservations ($\beta = 0.089, p \leq 0.05$), being wheelchair accessible ($\beta = 0.085, p \leq 0.05$), and not offering alcohol ($\beta = 0.126, p \leq 0.05$). The results also reveal that accepting credits cards has a significant ($\beta = -0.391, p \leq 0.05$) negative impact, along with being good for kids ($\beta = -0.167, p \leq 0.05$), offering a late night menu ($\beta = -0.159, p \leq 0.05$), possessing a TV ($\beta = -0.071, p \leq 0.05$), and being good for groups ($\beta = -0.14, p \leq 0.05$).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
2	.133	0.018	0.016	0.6110

Table 6a: Summary statistics for Model 2.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
2	(Constant)	3.691	0.014		271.962	0.000
	IsRomantic	0.276	0.133	0.045	2.079	0.038
	IsIntimate	0.377	0.136	0.060	2.782	0.005
	IsHipster	0.344	0.079	0.091	4.372	0.000
	IsTrendy	0.116	0.051	0.048	2.296	0.022

Table 6b: Linear regression coefficients of Model 2 attributes over entire sample.

Table 6a shows a model summary of our linear regression model, Model 2. This model was created through linear regression conducted upon the entirety of our dataset (2265 observations) using our ambience attributes. Table 6b shows the coefficients of our linear regression model, Model 2. The linear regression results of the ambience variables presented in Table 6b shows that a romantic atmosphere has a significant ($\beta = 0.276, p \leq 0.05$) positive impact on business review, along with an intimate atmosphere ($\beta = 0.377, p \leq 0.05$), a hipster atmosphere ($\beta = 0.344, p \leq 0.05$), and a trendy atmosphere ($\beta = 0.116, p \leq 0.05$).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
3	.356	0.127	0.108	0.5755

Table 7a: Summary statistics for Model 3.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
3	(Constant)	3.928	0.209		18.771	0.000
	AcceptsCreditCards	-0.373	0.207	-0.063	-1.800	0.072
	HasGarage	-0.015	0.108	-0.005	-0.141	0.888
	StreetParking	0.269	0.074	0.146	3.609	0.000
	HasParkingLot	0.140	0.061	0.098	2.286	0.023
	DoesCater	0.174	0.048	0.142	3.650	0.000
	IsGoodforKids	-0.262	0.070	-0.189	-3.761	0.000
	HasDessert	0.038	0.109	0.012	0.352	0.725
	HasLateNight	-0.383	0.089	-0.157	-4.311	0.000
	HasDinner	0.209	0.053	0.171	3.949	0.000
	HasBrunch	0.221	0.078	0.104	2.830	0.005
	HasTV	-0.115	0.051	-0.094	-2.236	0.026
	GoodForGroups	0.039	0.080	0.025	0.484	0.629
	AllowsReservations	0.007	0.058	0.005	0.118	0.906

	HasTableService	-0.118	0.061	-0.096	-1.936	0.053
	WheelchairAccessible	0.075	0.053	0.055	1.426	0.154
	NoAlcohol	0.108	0.055	0.083	1.957	0.051

Table 7b: Linear regression coefficients of Type 1 attributes over sample restricted to Arizona state.

Table 7a shows a model summary of our linear regression model, Model 3. This model was created through linear regression conducted upon our dataset, limited to the state of Arizona (758 observations) using our amenity attributes. Table 7b shows the coefficients of our linear regression model, Model 3. Data from Table 7b reveals that having street parking has a significant ($\beta = 0.269$, $p \leq 0.05$) positive impact on business review, along with possession of a parking lot ($\beta = 0.140$, $p \leq 0.05$), offering catering ($\beta = 0.174$, $p \leq 0.05$), offering a dinner menu ($\beta = 0.209$, $p \leq 0.05$), and offering a brunch menu ($\beta = 0.005$, $p \leq 0.05$). The results also show that being good for kids has a significant ($\beta = -0.262$, $p \leq 0.05$) negative impact, along with offering a late night menu ($\beta = -0.383$, $p \leq 0.05$), and possessing a TV ($\beta = -0.115$, $p \leq 0.05$).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
4	.150 ^a	0.023	0.017	0.6041

Table 8a: Summary statistics for Model 4.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
4	(Constant)	3.657	0.023		157.872	0.000
	IsRomantic	0.180	0.356	0.019	0.505	0.614
	IsIntimate	0.490	0.206	0.087	2.373	0.018
	IsHipster	0.368	0.134	0.099	2.748	0.006
	IsTrendy	0.158	0.087	0.066	1.819	0.069

Table 8b: Linear regression coefficients of Type 2 attributes over sample restricted to Arizona state.

Table 8a shows a model summary of our linear regression model, Model 4. This model was created through linear regression conducted upon our dataset, limited to the state of Arizona (758 observations) using our ambience attributes. Table 8b shows the coefficients of our linear regression model, Model 4. Data from Table 8b reflects that an intimate atmosphere has a significant ($\beta = 0.49, p \leq 0.05$) positive impact on business review, along with a hipster atmosphere ($\beta = 0.368, p \leq 0.05$).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
5	.368	0.135	0.121	0.5628

Table 9a: Summary statistics for Model 5.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
5	(Constant)	3.815	0.143		26.732	0.000
	AcceptsCreditCards	-0.304	0.140	-0.067	-2.170	0.030
	HasGarage	-0.018	0.059	-0.012	-0.301	0.764
	StreetParking	0.312	0.072	0.137	4.328	0.000
	HasParkingLot	0.246	0.052	0.196	4.707	0.000
	DoesCater	0.125	0.044	0.102	2.833	0.005
	IsGoodforKids	-0.175	0.055	-0.132	-3.177	0.002
	HasDessert	0.203	0.070	0.090	2.887	0.004
	HasLateNight	-0.040	0.060	-0.021	-0.673	0.501
	HasDinner	0.069	0.050	0.057	1.382	0.167
	HasBrunch	0.156	0.060	0.090	2.603	0.009
	HasTV	0.020	0.043	0.016	0.457	0.648
	GoodForGroups	-0.190	0.073	-0.122	-2.609	0.009
	AllowsReservations	0.125	0.051	0.103	2.440	0.015
	HasTableService	0.048	0.056	0.038	0.852	0.394

	WheelchairAccessible	0.098	0.044	0.074	2.211	0.027
	NoAlcohol	0.186	0.048	0.139	3.886	0.000

Table 9b: Linear regression coefficients of Type 1 attributes over sample restricted to Nevada state.

Table 9a shows a model summary of our linear regression model, Model 5. This model was created through linear regression conducted upon our dataset, limited to the state of Nevada (952 observations) using our amenity attributes. Table 9b shows the coefficients of our linear regression model, Model 5. Data from Table 9b indicates that having street parking has a significant ($\beta = 0.312, p \leq 0.05$) positive impact on business review, along with possession of a parking lot ($\beta = 0.246, p \leq 0.05$), offering catering ($\beta = 0.125, p \leq 0.05$), offering dessert ($\beta = 0.203, p \leq 0.05$), offering a brunch menu ($\beta = 0.156, p \leq 0.05$) being wheelchair accessible ($\beta = 0.098, p \leq 0.05$), and not offering alcohol ($\beta = 0.186, p \leq 0.05$). The results also show that accepting credits cards has a significant ($\beta = -0.304, p \leq 0.05$) negative impact, along with being good for kids ($\beta = -0.175, p \leq 0.05$), and being good for groups ($\beta = -0.190, p \leq 0.05$).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
6	.138	0.019	0.015	0.5956

Table 10a: Summary statistics for Model 6.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
6	(Constant)	3.727	0.020		182.061	0.000
	IsRomantic	0.284	0.171	0.055	1.667	0.096
	IsIntimate	0.429	0.250	0.057	1.714	0.087
	IsHipster	0.391	0.126	0.100	3.109	0.002

IsTrendy	0.097	0.073	0.043	1.332	0.183
----------	-------	-------	-------	-------	-------

Table 10b: Linear regression coefficients of Type 2 attributes over sample restricted to Nevada state.

Table 10a shows a model summary of our linear regression model, Model 6. This model was created through linear regression conducted upon our dataset, limited to the state of Nevada (952 observations) using our ambience attributes. Table 10b shows the coefficients of our linear regression model, Model 6. Data from table 10b shows that a hipster atmosphere has a significant ($\beta = 0.391$, $p \leq 0.05$) positive impact on business review.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
7	.389	0.151	0.095	0.6119

Table 11a: Summary statistics for Model 7.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
7	(Constant)	4.397	0.260		16.920	0.000
	AcceptsCreditCards	-0.682	0.270	-0.160	-2.526	0.012
	HasGarage	-0.231	0.180	-0.084	-1.281	0.201
	StreetParking	0.192	0.122	0.109	1.582	0.115
	HasParkingLot	0.100	0.100	0.076	0.998	0.319
	DoesCater	0.161	0.090	0.125	1.799	0.073
	IsGoodforKids	-0.235	0.118	-0.168	-1.985	0.048
	HasDessert	-0.104	0.207	-0.031	-0.503	0.616
	HasLateNight	-0.292	0.196	-0.096	-1.491	0.137
	HasDinner	0.102	0.106	0.079	0.959	0.338
	HasBrunch	0.134	0.169	0.052	0.796	0.427
	HasTV	-0.008	0.111	-0.006	-0.074	0.941
	GoodForGroups	-0.089	0.151	-0.056	-0.589	0.556
	AllowsReservations	-0.117	0.105	-0.085	-1.117	0.265
HasTableService	-0.117	0.130	-0.088	-0.897	0.370	

	WheelchairAccessible	0.126	0.092	0.097	1.367	0.173
	NoAlcohol	0.058	0.119	0.035	0.486	0.627

Table 11b: Linear regression coefficients of Type 1 attributes over sample restricted to North Carolina state.

Table 11a shows a model summary of our linear regression model, Model 7. This model was created through linear regression conducted upon our dataset, limited to the state of North Carolina (259 observations) using our amenity attributes. Table 11b shows the coefficients of our linear regression model, Model 7. Data from Table 11b reveals that none of the attributes have a significant positive impact on business review. The results also reveal that accepting credits cards has a significant ($\beta = -0.682$, $p \leq 0.05$) negative impact, along with being good for kids ($\beta = -0.235$, $p \leq 0.05$).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
8	.094	0.009	-0.007	0.6455

Table 12a: Summary statistics for Model 8.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
8	(Constant)	3.644	0.042		86.144	0.000
	IsRomantic	0.356	0.647	0.034	0.550	0.583
	IsIntimate	0.106	0.458	0.014	0.230	0.818
	IsHipster	0.187	0.248	0.047	0.756	0.450
	IsTrendy	0.178	0.158	0.071	1.130	0.260

Table 12b: Linear regression coefficients of Type 2 attributes over sample restricted to North Carolina state.

Table 12a shows a model summary of our linear regression model, Model 8. This model was created through linear regression conducted upon our dataset, limited to the state of North Carolina (259 observations) using our ambience attributes. Table 12b shows the coefficients of

our linear regression model, Model 8. Data from Table 12b indicates that none of the attributes have a significant impact on business review.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
9	.506	0.256	0.144	0.5198

Table 13a: Summary statistics for Model 9.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
9	(Constant)	4.167	0.274		15.231	0.000
	AcceptsCreditCards	-0.401	0.281	-0.127	-1.425	0.157
	HasGarage	-0.144	0.192	-0.071	-0.753	0.453
	StreetParking	0.264	0.123	0.220	2.150	0.034
	HasParkingLot	0.263	0.132	0.197	1.986	0.050
	DoesCater	-0.041	0.125	-0.034	-0.328	0.744
	IsGoodforKids	-0.106	0.124	-0.095	-0.856	0.394
	HasDessert	0.065	0.280	0.020	0.231	0.818
	HasLateNight	-0.164	0.175	-0.087	-0.938	0.350
	HasDinner	0.109	0.142	0.096	0.764	0.446
	HasBrunch	-0.008	0.239	-0.003	-0.033	0.974
	HasTV	-0.194	0.150	-0.173	-1.299	0.197
	GoodForGroups	-0.181	0.175	-0.149	-1.031	0.305
	AllowsReservations	0.042	0.149	0.036	0.285	0.777
	HasTableService	-0.101	0.178	-0.090	-0.567	0.572
	WheelchairAccessible	0.095	0.105	0.085	0.901	0.370
	NoAlcohol	0.210	0.151	0.167	1.389	0.168

Table 13b: Linear regression coefficients of Type 1 attributes over sample restricted to Pennsylvania state.

Table 13a shows a model summary of our linear regression model, Model 9. This model was created through linear regression conducted upon our dataset, limited to the state of

Pennsylvania (123 observations) using our amenity attributes. Table 13b shows the coefficients of our linear regression model, Model 9. Data from Table 13b suggests that having street parking has a significant ($\beta = 0.264$, $p \leq 0.05$) positive impact on business review, along with possession of a parking lot ($\beta = 0.263$, $p \leq 0.05$). The results also indicate that none of the attributes have a significant negative impact on business review.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
10	.142	0.020	-0.004	0.5629

Table 14a: Summary statistics for Model 10.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
10	(Constant)	3.826	0.053		71.924	0.000
	IsIntimate	0.674	0.565	0.108	1.192	0.236
	IsHipster	0.174	0.329	0.048	0.529	0.598
	IsTrendy	-0.183	0.219	-0.076	-0.835	0.406

Table 14b: Linear regression coefficients of Type 1 attributes over sample restricted to Pennsylvania state.

Table 14a shows a model summary of our linear regression model, Model 10. This model was created through linear regression conducted upon our dataset, limited to the state of Pennsylvania (123 observations) using our ambience attributes. Table 14b shows the coefficients of our linear regression model, Model 10. Data from Table 14b shows that none of the attributes have a significant impact on business review. The binary “IsRomantic” variable was excluded from regression, as there was only one record in the Pennsylvania sample where an establishment had a romantic atmosphere.

Discussion and Conclusion

Analysis of a pooled set of food establishment records and their attributes produces important insights about the general rating behavior of consumers. Certain establishment attributes tend to have a statistically significant impact on the average of reviews the establishment receives. This information is valuable to businesses, as they may be able to identify the most impactful features on rating, effectively implement or remove them, and potentially raise future ratings. The boost in ratings may have a significant impact on demand, ultimately generating greater revenues for the business (Luca, 2016).

Using linear regression over food establishment attributes, this study attempts to find a statistically significant model for predicting business rating. To increase the accuracy of the model, only attributes that were found to have a statistically significant correlation to business rating (identified through initial correlation testing in IBM SPSS) were utilized. In an attempt to find greater accuracy, attributes were split into two model types, with the first type consisting mainly of amenities the establishment provided while the second type focused on the ambience the establishment reflected.

Certain attributes exhibited a consistently significant influence on business rating. Whether or not a business accepts credit cards, provides street parking, has a parking lot, provides catering services, is good for kids, offers a brunch menu, or serves alcohol all have a significant influence on business rating results in the states of Arizona and Nevada. Unfortunately, there was much less consistency in attribute findings for North Carolina and Pennsylvania. The ultimate lack of findings in these states may largely be a result of the notably smaller sample size compared to Arizona and Nevada. Given a larger sample size it is likely that

those attributes found statistically significant in Arizona and Nevada would also be found to be statistically significant in other states.

The differences in local preferences may create too much inconsistency for an accurate predictive model based on attributes alone. Difference in local influence on attribute preference is reflected by the state-to-state comparison of statistical significance in linear regression model attributes. It is also likely that individuals have different opinions on which attributes are important, and they may also allow attributes to influence their final ratings differently.

There were some unexpected but interesting findings that warrant additional examination. For example, a food establishment was more likely to have a higher rating if they did not serve alcohol. One would assume that the provision of alcohol is an amenity that most adults would appreciate in a restaurant. Based on the data, star rating was inversely affected by the presence of alcohol in an establishment. To understand reasoning, we considered the potential that it was families that did not prefer the presence of alcohol in a restaurant, as it may not provide a positive environment for children to be around. To identify a correlation between these factors, we reviewed the correlation table, specifically looking at the variables, “NoAlcohol” and “IsGoodforKids”. We found that there was a Pearson correlation value of 0.294 between NoAlcohol and IsGoodforKids at the 0.01 significance level (2-tailed).

Another unexpected finding was the negative correlation between star rating and having a good environment for kids. It would seem that this would be a positive aspect for an establishment, as it would surely result in positive reviews from the vast population of families who desire to maintain a positive atmosphere for their children while dining out. Looking further at correlation data, it becomes increasingly interesting that a number of other attributes that have an inverse relationship with star rating also have a statistically significant positive correlation

with the `IsGoodforKids` variable. These variables include `IsCasual`, `AcceptsCreditCards`, `HasTV`, `GoodForGroups`, and `NoWiFi`. This information, particularly regarding the variables `IsGoodforKids`, `IsCasual`, `HasTV`, and `GoodForGroups`, reflects that the majority of Yelp users may prefer a more mature, intimate environment. Based on data posted on Yelp's website, 36 percent of users are aged 18-34 years old, 35.5 percent are aged 35-54 years old, and 27.7% are aged 55+ years old (Yelp 2017). Assuming that most families with children occupy the 35-54-year-old category, it makes sense that the impact of their reviews on the average is diminished by the 18-34 and 55+ year old age brackets.

Other peculiar findings include the inverse correlations found for `HasGarage` and `AcceptsCreditCards` with star rating. In both instances, we lack statistical evidence to reasonably justify causation for these relationships. One may assume that requiring garage parking may negatively impact review due to the negative experience of having to walk a relatively further distance from parking to destination (compared to parking options like a parking lot or street parking). Negative experience in parking within the garage, regarding finding available spots and driving past other vehicles on narrow routes, may also contribute to a negative rating. In the case of the variable, `AcceptsCreditCards`, the negative correlation may have to do with skewed data resulting from a small sample size. With an overwhelming 98.19 percent of the observations in our sample accepting credit cards, the positive reviews associated with the less than 2 percent of observations may distort analysis.

Ultimately, an accurate predictive model for business rating in the United States, based on attributes, is unlikely using linear regression. An accurate predictive model using attributes may be possible on a smaller regional scale (e.g. by town), as this may help to reduce differences in local preferences. A predictive model that flexes to exclusively include locally significant

attributes may have an even greater chance for accuracy. Attributes appear to carry more consistent significance over larger areas than linear regression models. This may be due to the fact that an attribute is less dynamic as a single variable, being less vulnerable to changes in local preferences than a multi-variable model.

Recommendations and Future Research

There are many topics that we could move towards to further our analysis on the Yelp dataset, as evident by research referred to in the literature review section. However, much could be done to better the analysis conducted on our subject of focus, the prediction of restaurant ratings based on review attributes, through revision of methodology and greater external research. An obvious improvement would be to use a larger sample in testing. This would require improved technology, specifically regarding hardware, from that used in our research. We were notably limited by the processing speed and capabilities of our resources. Another improvement may include a more heavily diversified and equally distributed sample regarding physical location.

A vast majority of the observations that we analyzed were from two states, Arizona and Nevada. Assuming that local preferences and behaviors significantly affect review rating, the uneven distribution of observations by state may affect the results. The number of states represented in the sample is an additional point for development. The dataset only included observations from eight of 50 states. Inclusion of more regions in the sample may provide greater insight into the significance of physical location on review behavior. A similar case could be made for the inclusion of analysis based on the city where the observation occurred.

This study took a relatively high level view of the sample set used, leaving opportunity for more focused review in the future. Analysis by city is just one instance of a more detailed approach, but further potential lies in the variables that were provided by the Yelp dataset that were excluded from our sample. Such variables may include review text, hours of operation, a variety of details regarding the user (e.g. number of friends, numbers of reviews; years active), and other variables in the attribute section that were not considered in this piece. Finally, looking beyond the sample, research could be enhanced by the use of other modeling techniques.

References

- Farhan, W. (2014). Predicting Yelp Restaurant Reviews. UC San Diego, La Jolla. Chicago
- Jayasimhan, A., Rai, P., Parekh, Y., & Patwardhan, O. (2017). Recommendation System for Restaurants. *International Journal of Computer Applications*, 167(6).
- Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. com.
- Mathieu, P. H., Grillet, A., Passerini, G., & Tiwari, I. (2016). Uncovering Business Opportunities from Yelp and OSM Data.
- Ngo-Ye, T. L., Sinha, A. P., & Sen, A. (2017). Predicting the helpfulness of online reviews using a scripts-enriched text regression model. *Expert Systems with Applications*, 71, 98-110.
- Sunil, M. G., Abhishek, H. M., Bari, A. K., & BR, B. K. S. (2017). Prediction of Rating by Using Users' Geographical Social Factors. *International Journal of Engineering Science*, 12517.
- Tang, D., Qin, B., Liu, T., & Yang, Y. (2015, July). User Modeling with Neural Network for Review Rating Prediction. In *IJCAI* (pp. 1340-1346).
- Yelp (2017). An Introduction to Yelp Metrics as of September 30, 2017. (n.d.). Retrieved December 11, 2017, from <https://www.yelp.com/factsheet>

Appendices

Appendix A Independent Samples T-Test

Appendix A.1 T-Test: IsRomantic = 1 vs. IsRomantic = 0

Group Statistics					
IsRomantic		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	2242	3.711	0.6174	0.0130
	1	23	4.065	0.2740	0.0571

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-2.745	0.006	-0.3538	0.1289
	Equal variances not assumed	-6.036	0.000	-0.3538	0.0586

Appendix A.2
T-Test between stars.business and IsIntimate

Group Statistics					
IsIntimate		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	2243	3.711	0.6164	0.0130
	1	22	4.159	0.3581	0.0764

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-3.406	0.001	-0.4484	0.1316
	Equal variances not assumed	-5.790	0.000	-0.4484	0.0775

Appendix A.3
T-Test between stars.business and IsHipster

Group Statistics					
IsHipster		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	2203	3.705	0.6188	0.0132
	1	62	4.056	0.3633	0.0461

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-4.445	0.000	-0.3510	0.0790
	Equal variances not assumed	-7.315	0.000	-0.3510	0.0480

Appendix A.4
T-Test between stars.business and IsTrendy

Group Statistics					
IsTrendy		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	2108	3.707	0.6249	0.0136
	1	157	3.825	0.4664	0.0372

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-2.318	0.021	-0.1180	0.0509
	Equal variances not assumed	-2.978	0.003	-0.1180	0.0396

Appendix A.5
T-Test between stars.business and AcceptsCreditCards

Group Statistics					
AcceptsCreditCards		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	41	4.085	0.5579	0.0871
	1	2224	3.708	0.6149	0.0130

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	3.898	0.000	0.3772	0.0968
	Equal variances not assumed	4.281	0.000	0.3772	0.0881

Appendix A.6
T-Test between stars.business and HasGarage

Group Statistics					
HasGarage		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	1997	3.729	0.6247	0.0140
	1	268	3.608	0.5348	0.0327

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	3.029	0.002	0.1211	0.0400
	Equal variances not assumed	3.409	0.001	0.1211	0.0355

Appendix A.7
T-Test between stars.business and StreetParking

Group Statistics					
StreetParking		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	1912	3.683	0.6298	0.0144
	1	353	3.887	0.5013	0.0267

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-5.740	0.000	-0.2034	0.0354
	Equal variances not assumed	-6.707	0.000	-0.2034	0.0303

Appendix A.8
T-Test between stars.business and HasParkingLot

Group Statistics					
HasParkingLot		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	788	3.614	0.6894	0.0246
	1	1477	3.769	0.5657	0.0147

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-5.766	0.000	-0.1555	0.0270
	Equal variances not assumed	-5.432	0.000	-0.1555	0.0286

Appendix A.9
T-Test between stars.business and DoesCater

Group Statistics					
DoesCater		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	1318	3.657	0.6677	0.0184
	1	947	3.795	0.5255	0.0171

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-5.280	0.000	-0.1377	0.0261
	Equal variances not assumed	-5.487	0.000	-0.1377	0.0251

Appendix A.10
T-Test between stars.business and IsGoodforKids

Group Statistics					
IsGoodforKids		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	662	3.800	0.6376	0.0248
	1	1603	3.680	0.6034	0.0151

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	4.229	0.000	0.1199	0.0283
	Equal variances not assumed	4.133	0.000	0.1199	0.0290

Appendix A.11
T-Test between stars.business and HasDessert

Group Statistics					
HasDessert		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	2145	3.709	0.6148	0.0133
	1	120	3.825	0.6273	0.0573

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-2.012	0.044	-0.1161	0.0577
	Equal variances not assumed	-1.976	0.050	-0.1161	0.0588

Appendix A.12
T-Test between stars.business and HasLateNight

Group Statistics					
HasLateNight		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	2081	3.728	0.6100	0.0134
	1	184	3.565	0.6621	0.0488

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	3.450	0.001	0.1630	0.0473
	Equal variances not assumed	3.222	0.001	0.1630	0.0506

Appendix A.13
T-Test between stars.business and HasDinner

Group Statistics					
HasDinner		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	1049	3.680	0.7157	0.0221
	1	1216	3.745	0.5128	0.0147

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-2.538	0.011	-0.0658	0.0259
	Equal variances not assumed	-2.478	0.013	-0.0658	0.0265

Appendix A.14
T-Test between stars.business and HasBrunch

Group Statistics					
HasBrunch		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	2026	3.702	0.6206	0.0138
	1	239	3.822	0.5639	0.0365

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-2.849	0.004	-0.1198	0.0421
	Equal variances not assumed	-3.072	0.002	-0.1198	0.0390

Appendix A.15
T-Test between stars.business and HasTV

Group Statistics					
HasTV		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	1059	3.765	0.6647	0.0204
	1	1206	3.671	0.5662	0.0163

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	3.655	0.000	0.0945	0.0259
	Equal variances not assumed	3.617	0.000	0.0945	0.0261

Appendix A.16
T-Test between stars.business and GoodForGroups

Group Statistics					
GoodForGroups		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	441	3.817	0.7463	0.0355
	1	1824	3.690	0.5775	0.0135

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	3.905	0.000	0.1272	0.0326
	Equal variances not assumed	3.346	0.001	0.1272	0.0380

Appendix A.17
T-Test between stars.business and AllowsReservations

Group Statistics					
AllowsReservations		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	1453	3.693	0.6732	0.0177
	1	812	3.754	0.4950	0.0174

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-2.272	0.023	-0.0613	0.0270
	Equal variances not assumed	-2.473	0.013	-0.0613	0.0248

Appendix A.18
T-Test between stars.business and HasTableService

Group Statistics					
HasTableService		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	893	3.748	0.7125	0.0238
	1	1372	3.694	0.5431	0.0147

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	2.061	0.039	0.0545	0.0265
	Equal variances not assumed	1.948	0.052	0.0545	0.0280

Appendix A.19
T-Test between stars.business and WheelchairAccessible

Group Statistics					
WheelchairAccessible		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	722	3.602	0.7259	0.0270
	1	1543	3.768	0.5494	0.0140

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-5.993	0.000	-0.1652	0.0276
	Equal variances not assumed	-5.429	0.000	-0.1652	0.0304

Appendix A.20
T-Test between stars.business and NoAlcohol

Group Statistics					
NoAlcohol		N	Mean	Std. Deviation	Std. Error Mean
stars.business	0	1623	3.694	0.5940	0.0147
	1	642	3.769	0.6655	0.0263

Independent Samples Test					
		t-test for Equality of Means			
		t	Sig. (2-tailed)	Mean Difference	Std. Error Difference
stars.business	Equal variances assumed	-2.612	0.009	-0.0749	0.0287
	Equal variances not assumed	1066.783	0.013	-0.0749	0.0301

Appendix B
Regression Coefficients by State

	Stars_Business (Type 1 Attributes)			
Attribute	Arizona	Nevada	North Carolina	Pennsylvania
AcceptsCreditCards	-0.373*	-0.304*	-0.682*	-0.401
Standard Error	0.207	0.140	0.270	0.281
HasGarage	-0.015	-0.018	-0.231	-0.144
Standard Error	0.108	0.059	0.180	0.192
StreetParking	0.269***	0.312***	0.192	0.264*
Standard Error	0.074	0.072	0.122	0.123
HasParkingLot	0.14*	0.246***	0.100	0.263*
Standard Error	0.061	0.052	0.100	0.132
DoesCater	0.174***	0.125**	0.161*	-0.041
Standard Error	0.048	0.044	0.090	0.125
IsGoodForKids	-0.262***	-0.175**	-0.235*	-0.106
Standard Error	0.070	0.055	0.118	0.124
HasDessert	0.038	0.203**	-0.104	0.065
Standard Error	0.109	0.070	0.207	0.280
HasLateNight	-0.383***	-0.040	-0.292	-0.164
Standard Error	0.089	0.060	0.196	0.175
HasDinner	0.209***	0.069	0.102	0.109
Standard Error	0.053	0.050	0.106	0.142
HasBrunch	0.221**	0.156**	0.134	-0.008
Standard Error	0.078	0.060	0.169	0.239
HasTV	-0.115*	0.020	-0.008	-0.194
Standard Error	0.051	0.043	0.111	0.150
GoodForGroups	0.039	-0.19**	-0.089	-0.181
Standard Error	0.080	0.073	0.151	0.175
AllowsReservations	0.007	0.125*	-0.117	0.042
Standard Error	0.058	0.051	0.105	0.149
HasTableService	-0.118*	0.048	-0.117	-0.101
Standard Error	0.061	0.056	0.130	0.178
WheelchairAccessible	0.075	0.098*	0.126	0.095
Standard Error	0.053	0.044	0.092	0.105
NoAlcohol	0.108*	0.186***	0.058	0.210
Standard Error	0.055	0.048	0.119	0.151
	Stars_Business (Type 2 Attributes)			
IsRomantic	0.180	0.284*	0.356	
Standard Error	0.356	0.171	0.647	

IsIntimate	0.49*	0.429*	0.106	0.674
Standard Error	0.206	0.250	0.458	0.565
IsHipster	0.368**	0.391**	0.187	0.174
Standard Error	0.134	0.126	0.248	0.329
IsTrendy	0.158*	0.097	0.178	-0.183
Standard Error	0.087	0.073	0.158	0.219

*P-value less than .1, **p-value less than .01; ***p-value less than .001