Exploring and Predicting Frequently Reviewed Businesses for Yelp's Dataset Challenge

John Fuller, Lauren Hubener, Jonathan Lin, Amy Trost

Abstract

Users are drawn to businesses with high review counts as a quick and easy-to-understand signifier of excellence on Yelp. In this paper we explore the attributes of businesses on Yelp and develop a series of statistical models to predict if a business is likely to receive more than 100 user reviews. We combined business information provided by Yelp with maps, census data, and calculations of review timing and frequency to conduct exploratory analysis and create statistical models. While many of our predictors--category, population density, and skewness, for example--showed some promise in the exploratory phase, we could only successfully classify businesses as highly reviewed when the number of reviews generated per day was considered. Our most successful model had a precision rate of 79% and accurately predicted 217 of 465 restaurants with more than 100 reviews. Future research could improve on these results by incorporating sentiment and topic analysis of user reviews and merging Yelp's dataset with local traffic and tourism information.

I. Introduction

Generally speaking, having a high number of reviews can help boost a business's profile on Yelp. In recent years, a number of "How To" articles have been published for the benefit of business owners ("5 Steps to Getting Your Business Ranked on Yelp", "6 Ways to Get Google and Yelp Reviews", "Yelp For Business – 6 Simple Steps To Success"), indicating that Yelp reviews are becoming increasingly relevant from the perspective of business owners. Users are drawn to businesses with high review counts – along with a good star rating – because they are quick and easy-to-understand signifiers of excellence. If a given number of people have rated and reviewed a particular business, others are likely to patronize that business over a similar business with a low review count (or a lower star rating).

In this report, we consider the possible factors that account for the high number of reviews for a relatively small number of businesses on Yelp, and why these businesses seem to generate many more reviews in comparison with the majority of businesses. Does the discrepancy in the high number of reviews for a small fraction of businesses indicate that these select businesses are of higher overall quality than their Yelp competitors, and can high review counts be used as a proxy for success? Furthermore, could a business's success be predicted based on certain attributes? These questions led to the formation of our research question: What are the attributes of frequently reviewed businesses on Yelp? Can we predict if a business will be highly reviewed based on particular attributes?

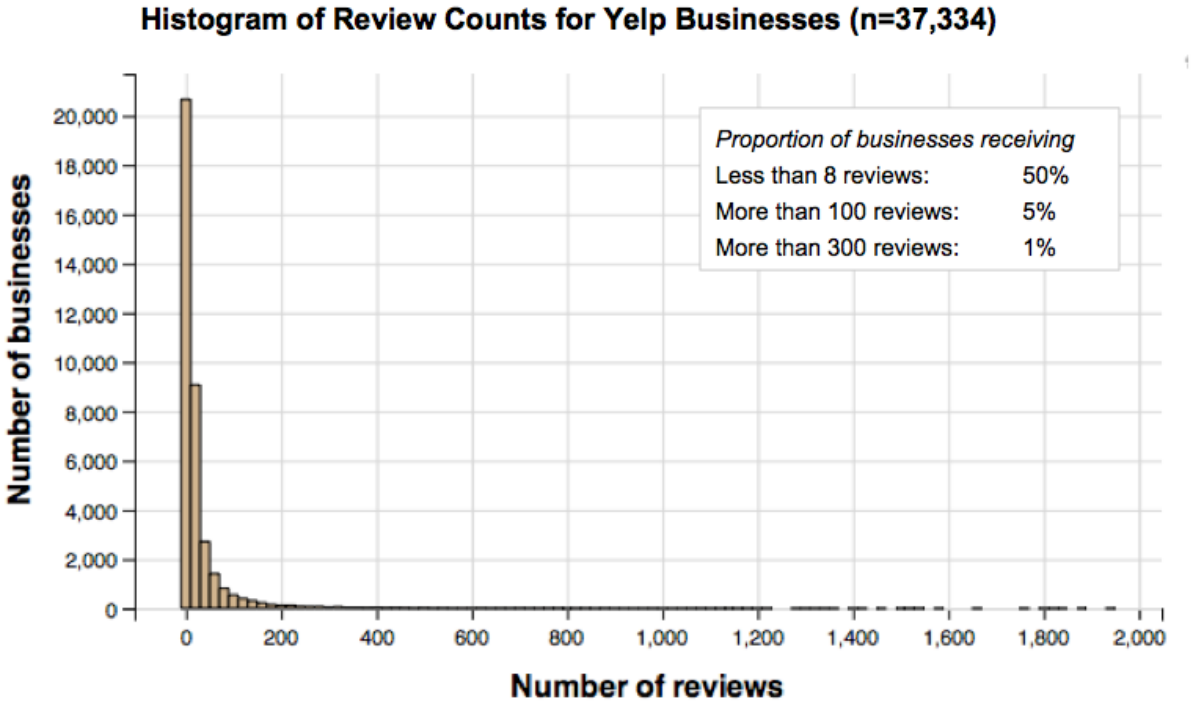
In recent years, researchers have designed prediction models related to online user behaviors in terms of rating prediction and product or service recommendations. Some methods have used a Bayesian framework to represent rating data via temporally constrained categorical mixture models, enabling prediction for newly occurring ratings (Günnemann, S., Günnemann, N., & Faloutsos, 2014). Sentiment analysis for summarizing or analyzing consumers reviews have used classification and regression models (Gupta, Di Fabbrizio, and Haffner, 2010), and collaborative filtering has become a common method to make recommendations to a target user based on the opinions and ratings of similar users (He and Chu, 2010). Hood, Hwang, and King (2013) conducted similar research for the Yelp Dataset Challenge and found that models incorporating review timing predicted business review counts most successfully (p. 7). However, the focus on these and other studies have been primarily concerned with *rating prediction*, whereas we are interested in understanding what possible business features – categorical and otherwise – can be used to determine a binary result, i.e., a review count of 100 or more for a given business.

II. Description of Dataset

The version of the Yelp Academic Dataset used in analysis was released by Yelp on August 1, 2014, coinciding with Round 4 of their ongoing Dataset Challenge. This dataset consists of five main objects encoded as JSON files: businesses, checkins, reviews, tips, and users. The recorded information spans a ten year period, with the earliest data from April 2004. The scope of our study utilized two of the objects, businesses and reviews, which provide over 1,000,000 reviews from more than 42,000 businesses. Important identifiers considered include business ids, locations, ratings, dates, and categories.

III. Methods

For this research project, highly reviewed businesses were defined as businesses with 100 reviews or more. Accounting for less than five percent of all businesses in the dataset (see Figure 1, below), each business in this group has generated enough attention to be considered a "destination location." Additionally, we limited our research to 37,334 businesses in the United States with their earliest Yelp review before January 1, 2014, in order to gain a stronger sense of how a business's reviews accumulate over a longer period of time.



One of the most challenging aspects of our project was choosing a set of predictor variables that could be feasibly compiled and analyzed in a 10-week time frame. In our initial research, we looked at the possibility of forming review topics via a natural language process model, such as latent Dirichlet allocation (LDA). Review topic and text analysis could potentially prove to be a useful indicator in our overall prediction analysis but was outside of the scope of our research at this time.

We relied on a number of predictors from the original Yelp business and review datasets for our analysis, including, most commonly, review count (the number of total reviews a business has received), stars (the star rating a business received), and business price range. Business categories were also included in our analysis in order to determine what types of businesses received a disproportionately high number of reviews. Four additional predictor variables were created from a combination of sources. Population per square mile was calculated using U.S. Census data. Review timing refers to the how businesses accumulate reviews over the time of the business's existence. The variable 'open' indicates the number of days a business has been open, and 'rate' refers to the average

number of reviews a business receives *per day*. How these predictor variables were calculated and used in our analysis will be described in more detail below.

Adding population density to Yelp's business data was a two-step process. First, we assigned each of our 37,334 businesses a census block number based on its latitude and longitude. This was accomplished using the Federal Communications Commission's Census Block Conversions call (FCC, 2014). Block numbers were simplified into block group numbers. The second phase of this process involved matching population and land area data to each business based on these census block groups. Data from 2010 were retrieved from the U.S. Census' American Fact Finder database (USCB, 2014) for "area(land)" (located in the geographic identifiers table) and "total population," spanning three metropolitan areas: Phoenix, AZ, Las Vegas, NV, and Madison, WI. Additional block group data for neighboring counties were added as needed.

In addition to a high review count, we recognized that we should also look at the frequency of business reviews, normalizing for time. We found this by calculating the businesses' rate of reviews *per day*. Since the Yelp business data does not provide opening or closing dates for businesses, the calculation for rate was restricted to the 37,145 businesses that remain open. First, two new variables were added to the data frame - the date of the earliest review for each business and the maximum date in the dataset (July 16, 2014). After converting the two dates to the correct object format in R, we then added an additional variable subtracting the earliest review date from the maximum date in order to calculate the number of days each business has been open. To find the rate of reviews for each business, the review count for the business was divided by the approximate number of days it had been open. The most frequently reviewed business in the dataset (of those open before January 1, 2014) has an average rate of 3.51 reviews per day.

In the scope of our project, skewness was used as an indicator of how a business accumulated reviews over time. As a rule, negative skewness indicates that the mean of the data values is less than the median, and the data distribution is left-skewed. Positive skewness would indicate that the mean of the data values is larger than the median, and the data distribution is right-skewed (Yau, 2014). In this context, a business with positive skewness signifies that a given business received a greater proportion of reviews earlier in its existence and has a skewness value of greater than zero. Conversely, a business with negative skewness received a greater proportion of its reviews later in its existence, and has a skewness value of less than zero (see Appendix, Figures 3a. and 3b. for examples of review skewness). A business with a skewness value very near or of zero indicates that the business received a fairly steady accumulation of reviews throughout its existence. To determine a skewness measure, we used the D'Agostino test for skewness provided in the R package moments.

With the exception of our categorical data (discussed below), all of the other independent variables (predictors) for our analysis were combined into a single file for exploratory data analysis and statistical modelling. This final predictor file included:

- selected data from Yelp's business table (review_count, latitude/longitude, price range, star rating, nine ambience attributes, and the business name);
- a designation for each business into one of three metropolitan areas;
- population density (from the 2010 Census);
- review density over time (measured by skewness); and
- review frequency (average number of reviews per day).

During this data combination exercise, we filtered the businesses to include only those with their earliest Yelp review before January 1, 2014.

For each business, Yelp provides up to 10 category tags from a pool of 703 total categories. Some of these, such as "Restaurants," "Nightlife," and "Health & Medical," are used to tag thousands of businesses. Others, such as "Ukranian" or "Buses," occur only once. We utilized the category data in two ways. First, we calculated the total number of times each tag was applied to a business for both our study population (37,334 businesses) and the 2417 businesses in this group that receive more than 100 reviews. This allowed us to use tables and graphics to explore which categories were over- and underrepresented among frequently reviewed restaurants. Next, we created a matrix of businesses (1 row for each) and categories (1 column for each). If a category was included as a tag for a specific business, we assigned a value of "1" to that cell in the matrix; otherwise, the cell received a value of "0." Categories were then filtered so that we only included the 124 categories that were represented in at least 100 businesses. In this way we created a large set of binary predictors for our statistical modelling.

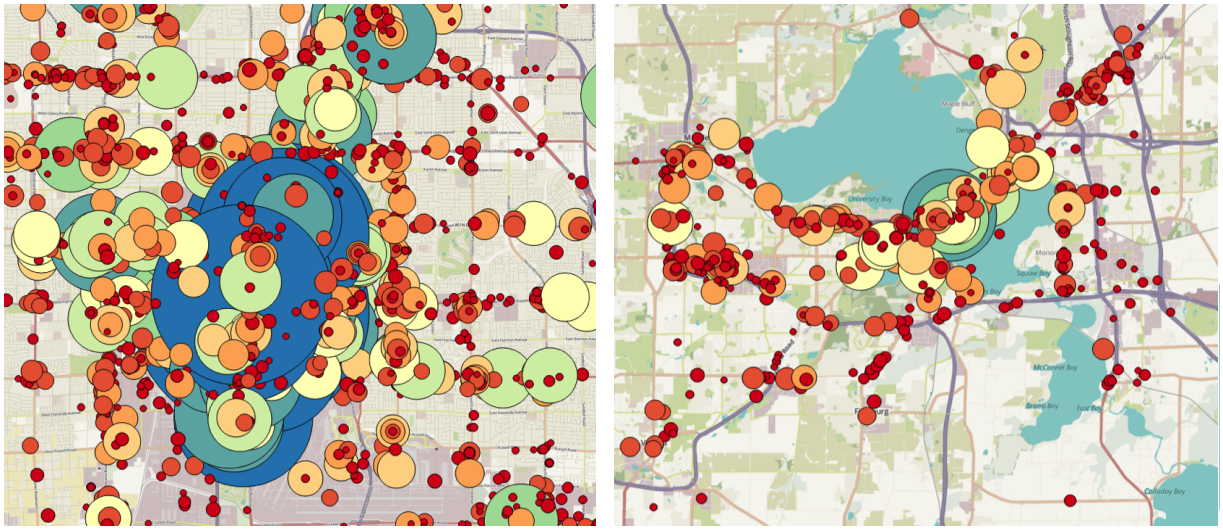
Maps of the three U.S. metropolitan areas – Las Vegas, Madison, and Phoenix – were created using QGIS, an open source geographic information system. The maps allow us to better visualize aspects of neighborhood population density and the location of frequently reviewed businesses. Tables were exported from R into QGIS though csv files. Variables used in plotting to geographic maps include the rate of reviews per day, total review count, positional longitude/latitude coordinates, and business names. The data was paired with the American Community Survey 5-Year Estimates - Geodatabase Format TIGER/Line Shapefiles for 2007-2011 Block Group Data (USCB, 2014). This allowed us to visualize businesses positioned within the local total population by census block group.

VI. Primary Results

i. Exploratory analysis

Using GIS and population density information, we were able to visualize where frequently reviewed businesses tend to locate in each of the three metro areas. While there doesn't seem to be an overall correlation between population density and location (see Appendix A, Figure 2), when isolating the metro area as a variable, we can examine trends within each city.

In Las Vegas, businesses with more than 100 reviews cluster along the Las Vegas Strip, a stretch of South Las Vegas Boulevard, located immediately south of the Las Vegas city limits (see below, left). This area exhibits a strong negative correlation between review count and population density. In Madison, most of the businesses with high review counts are clustered downtown on the isthmus between Lake Mendota and Lake Monona (see below, right). This area is located between Madison's Northeast side to the east and the University of Wisconsin campus to the west. In this small city, frequently reviewed businesses do tend to be positioned in more densely populated areas. Phoenix, on the other hand, has a vast spread of frequently reviewed businesses distributed throughout the Greater Phoenix Area, indicative of urban sprawl.



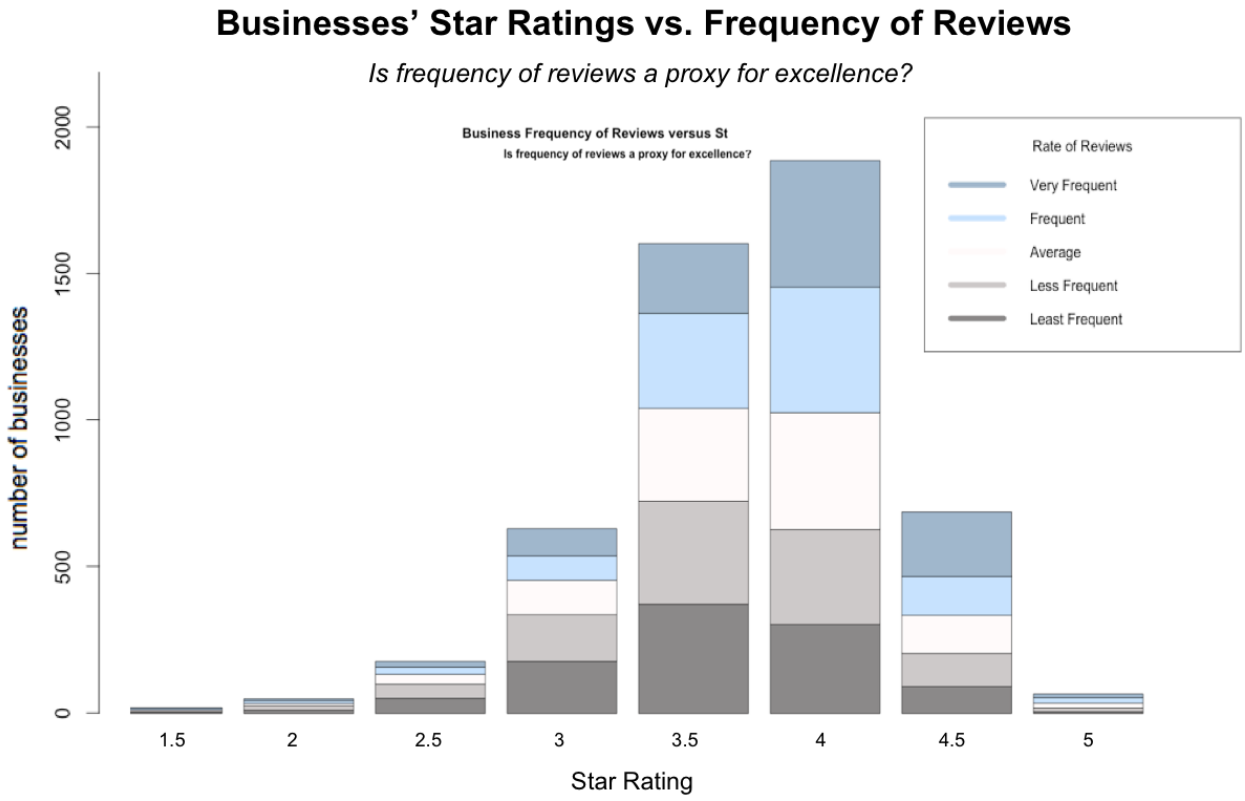
Total business Review Counts in Las Vegas (left) and Madison (right). Each point represents a business. The more reviews, the larger and bluer the circle.

A business' category designations in Yelp showed promise in predicting review count (see Figure 3 in Appendix A for more details). Businesses in some categories, such as New American cuisine, Bars, and Nightlife, are 2-3 times more likely to receive more than 100 reviews than the general population of businesses. Businesses labelled as "Shopping," on the other hand, were nearly ten times *less* likely to receive more than 100 reviews. The ten categories that appeared most significant were included to varying degrees in our logistic regression and rpart models (described below).

Assigning a skewness measure (D'Agostino's) to each business allowed us to determine if businesses with a high review count tend to accumulate reviews in a particular pattern over time. From our sample of 37,334 U.S. businesses, 24,388 businesses (65.3%) were negatively skewed, 11,349 businesses (30.4%) were positively skewed, and 1,432 businesses (3.8%) had zero skew. The average skew for all businesses in the sample was -0.2555, indicating a slight overall negative skew, meaning that a majority of businesses tend to receive their reviews more recently in time throughout their respective existences. When exploring how review counts skew over time by city, no distinguishable differences were evident, with the majority of review skews for all cities falling between -1.0 to 0 (see Appendix, Figures 5a, 5b, and 5c for graphs of review skewness versus review counts by city).

Calculating the rate of reviews for businesses allowed us to explore businesses with a high review count, controlling for the approximate time that each business has been open. After examining the most frequently reviewed restaurants, we wondered whether there was a correlation between a high rate of reviews and highly rated businesses in terms of their average star rating. (See Appendix Figures 3a, 3b, and 3c for tables of most frequently reviewed businesses and their star ratings for each city.) The graph below shows the distribution of star ratings for all businesses (limited to businesses that have received at least 30 reviews, to allow for a statistically significant accumulation of reviews). The bars are colored according to the rate of review frequency for the businesses. The visual demonstrates that

more frequently reviewed businesses are slightly more likely to have a higher average rating as opposed to those businesses that are less frequently reviewed.



ii. Business classification with logistic regression and rpart

We attempted to predict the likelihood that a business would receive more than 100 reviews on Yelp with two statistical models: the recursive partitioning and regression trees (rpart) package and the binomial family of models within R's glm function. We incorporated a selection of both scaled and binary predictors, predicting the most effective independent variables from our exploratory analysis (see previous section). Our models were only successful, however, when they incorporated the number of reviews generated per day for each business.

The results of five separate logistic regression models are shown in Table 1. We trained the first two models with 80% of the business data, treating categorical and noncategorical predictors separately. However, neither of these models were able to predict the presence of any frequently (>100) reviewed businesses. We tried two additional models with biased training data: our third model was trained with a higher proportion of businesses with >100 reviews, and the fourth model was trained with a higher proportion of category designations. The precision of these models only improved to 5%. Incorporating the number of reviews generated per day improved the precision of our model to 79%, although the model was able to locate less than half of the restaurants in our test set with more than 100 reviews (217 of 465 for a sensitivity of 47%).

Predicting businesses with more than 100 reviews: Results of binomial modelling						
	Sample Size (n)					
	training data	testing data	Predictors	Accuracy	Precision	Sensistivity
1	19294	7467	poulation density, price range, skewness, star count	93%	0%	0%
2	29867	7467	Category Designations: restaurants, nightlife, bars, New American, Mexican, Beauty/Spas, shopping, Active Life, Fast Food, Traditional American	93%	0%	0%
3	5119 *	1959	star count, price range, skewness	88%	5%	6%
4	15147 **	4208	All category designations from test #2; price range, skewness, metro area	93%	5%	0%
5	19550	7467	price range, skewness, **rate of reviews (number per day)**	96%	79%	47%

Notes

* Training set was biased to include more frequently rated businesses.

** Training set to was biased to include more businesses with pertinent category designations.

Table 1. Evaluating logistic regression models.

Four decision-tree models were created with R's rpart package based on a mix of predictor variables (see Appendix figure X). The models varied by including or excluding the business attributes (price range, romantic, touristy, etc.) in our predictor list and the frequency of reviews per day (rate).

The most successful rpart model included our entire set of non-categorical predictor variables: price range, skewness and frequency, star rating, metropolitan area, and nine ambience attributes. Using three ambience attributes (casual, divey, hipster) as the most significant variables in tree construction, this model was able to correctly predict 1,596 of the 1,981 businesses with over 100 reviews, a sensitivity of 80%. However, when plotting this tree, we noticed that a large percentage (around 70%) of businesses were easily filtered from the first decision branch which hinged on the rate of review to be less than 0.044 per day (or about 16 reviews per year, see Appendix A figure X for details). Rerunning our model but removing our rate variable gave only 807 out of 1,981 (40%) correct predictions that a business will get over 100 reviews.

VII. Significance

While many of our predictors--category, population density, and skewness, for example--showed some promise in the exploratory phase, we were only able to successfully classify businesses as highly reviewed when the number of reviews generated per day was considered. This supports the results of Hood, Hwang, and King in the first Yelp dataset challenge, where indicators such as attention time and the maximum number of reviews in a day most accurately predicted the total number of reviews a business would receive (2013, p. 7).

The rate of reviews can be a useful predictor of emerging "hot spots" (see Table 2, below). A cursory survey of frequently reviewed businesses in Las Vegas shows an exponential increase in review counts between July and December 2014. However, temporal indicators only show predictive power for a limited set of businesses that are "high-visibility." An ideal predictive model would find emerging businesses that open to little fanfare, but develop into prominent fixtures in their communities due to the excellence of their products and services.

Up and Coming Businesses in Las Vegas			
Business	**Days Open***	**Review Count (then*)**	**Review Count (now*)**
Mercadito	26	85	201
Giada	44	85	488
Carson Kitchen	34	75	245
Guy Fieri	90	278	548
High Roller	111	177	332

* Days Open (at time of dataset compilation)
* Then (as of July 16, 2014)
* Now (as of December 1, 2014)

Table 2: Predicting emerging "hot spots" from the rate of reviews.

While we have not yet found the optimal mix of variables that will predict these less obvious "up and coming businesses," we have identified some future directions for this research which could improve the sensitivity and precision of our model. Both sentiment and topic analysis of initial reviews of a business could highlight unique characteristics of those likely to receive more than 100 reviews. Local traffic and tourism data would improve the model, as well. We are particularly interested in stratifying business characteristics based on the level of tourism that a geographic area receives. In Las Vegas, for example, tourism accounts for roughly 20% of local GDP (Applied Analysis, 2009, p. 10). The effects of tourism need to be considered more closely when evaluating the way that reviews are generated for individual businesses.

VIII. Conclusion

A variety of variables were taken into account in our attempts to predict successful businesses, defined as those having 100 or more reviews on Yelp. However, when excluding the predictor variable of rate of reviews generated per day, our generalized linear models and recursive partitioning and regression tree models proved unsuccessful in predicting the presence of frequently reviewed businesses with high precision. Further research is required to determine more effective predictor variables in order to identify emerging businesses new to Yelp or those aiming to accumulate a high review count. Sentiment and topic analysis of user reviews as well as merging Yelp's dataset with local traffic and tourism data might also be considered in formulating a better predictor model.

References

6 Ways to Get Google and Yelp Reviews. (2013). Yola Inc. Retrieved from: http://www.yola.com/blog/6-ways-to-get-google-and-yelp-reviews/

Applied Analysis. (2009). The Relative Dependence on Tourism of Major U.S. Economies. Published by the Las Vegas Convention and Visitor's Authority. Retrieved from http://www.appliedanalysis.com/projects/lvcvaeis/EIS%201.8%20Tourism%20Economies.pdf.

Federal Communications Commission. Census Block Conversions API: Census Block Methods Call. Retrieved from: http://www.fcc.gov/developers/census-block-conversions-api.

Gupta, N., Di Fabbrizio, G., & Haffner, P. (2010, June). Capturing the stars: predicting ratings for service and product reviews. In Proceedings of the NAACL HLT 2010 Workshop on Semantic Search (pp. 36-43). Association for Computational Linguistics.

Günnemann, S., Günnemann, N., & Faloutsos, C. (2014, August). Detecting anomalies in dynamic rating data: a robust probabilistic model for rating evolution. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 841-850). ACM.

He, J., & Chu, W. W. (2010). A social network-based recommender system (SNRS) (pp. 47-74). Springer US.

Hood, B., Hwang, V., and King, J. Inferring Future Business Attention.(2013). In *Yelp Dataset Challenge Round One Winners*. Retrieved from: http://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_InferringFuture.pdf

Kumar, A. (2013). 5 Steps to Getting Your Business Ranked on Yelp. Entrepreneur Media, Inc. Retrieved from: http://www.entrepreneur.com/article/225570

Ma, M. (2014). Classification and Regression Trees (CART) with rpart and rpart.plot. Retrieved from http://rpubs.com/minma/cart_with_rpart

Statistical Consulting Group. (2013). Classification Trees (R). Retrieved from http://scg.sdsu.edu/ctrees_r/

United States Census Bureau (USCB). 2010 Census. In *American Fact Finder*. Retrieved from: http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml

United States Census Bureau (USCB). (2014). Tiger/Line with Selected Demographic and Economic Data. Retrieved from: https://www.census.gov/geo/maps-data/data/tiger-data.html.

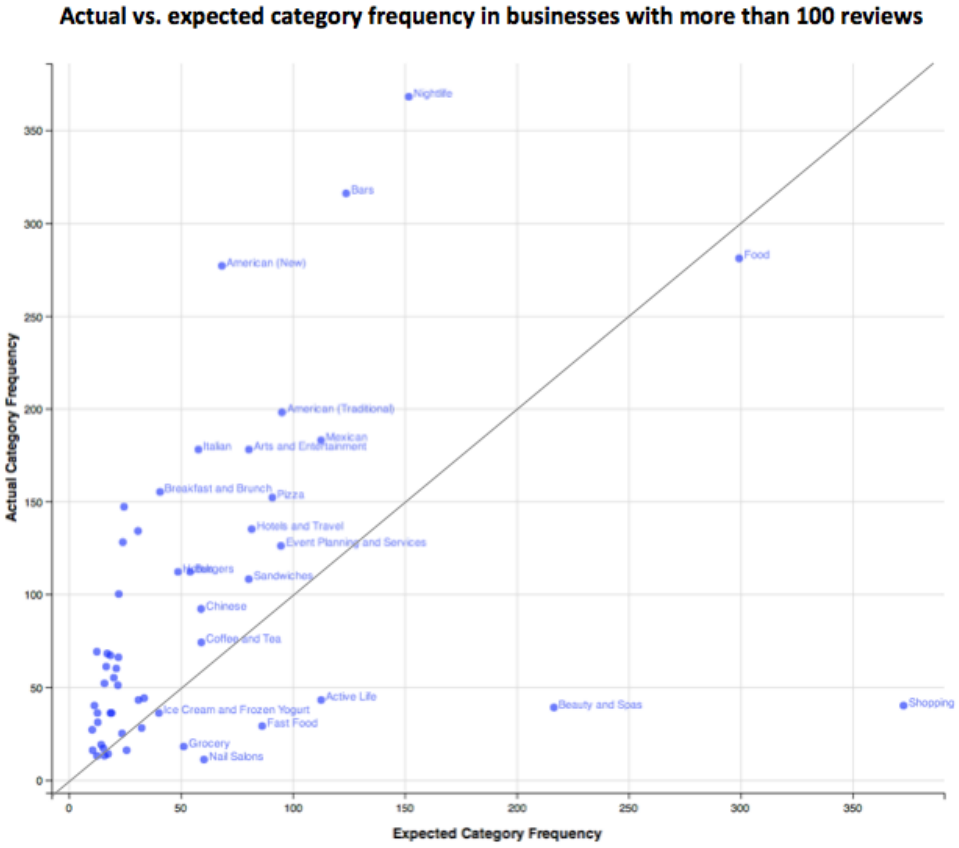
Waring, D. (2013). Yelp for Business - Six Simple Steps to Success. Fit Small Business. Retrieved from: http://fitsmallbusiness.com/yelp-for-business/

Wesley. (29 April 2013). A Brief Tour of the Trees and Forests. Retrieved from: http://www.r-bloggers.com/a-brief-tour-of-the-trees-and-forests/

Yau, C. (2014). Skewness. *R Tutorial: An Introduction to Statistics.* Retrieved from http://www.r-tutor.com/elementary-statistics/numerical-measures/skewness

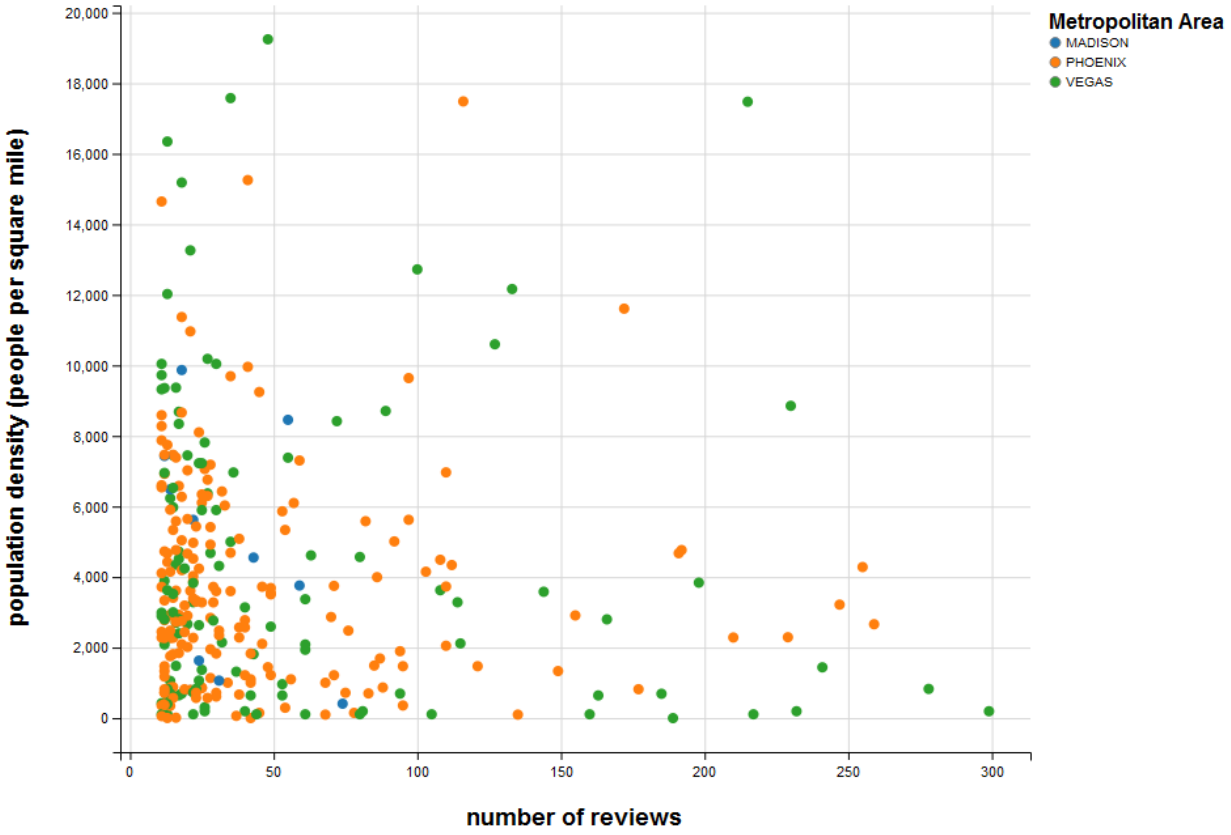
Appendix A: Additional Figures

Figure 3.



Note how some categories are disproportionately represented in businesses with more than 100 reviews. For easier visualization, the "Restaurants" tag (which was used 1892 in businesses with more than 100 reviews) was excluded here. We also excluded tags which were represented in less than 10 businesses. The line shown here (y=x) represents the location in two-dimensional space where the actual category frequency is equivalent to expected category frequency.

Population Density vs. Review Count



Here is the relationship between population density and review count for a random sample of businesses that received between 10 and 300 reviews. In all three metropolitan study areas, no strong correlation was found.

Most Frequently Reviewed Businesses in Las Vegas		
Rank	**Business**	**Rating**
1.	Bacchanal Buffet	
2.	Gordon Ramsay BurGR	
3.	Wicked Spoon	
4.	The Cosmopolitan of Las Vegas	
5.	Hakkasan Nightclub	
6.	Earl of Sandwich	
7.	Gordon Ramsay Steak	
8.	Soho Japanese Restaurant	
9.	Gordon Ramsay Pub & Grill	
10.	Secret Pizza	

Figure 3a. This table shows the top ten most frequently reviewed businesses in Las Vegas based on their average rate of reviews per day and each businesses' average star rating.

Most Frequently Reviewed Businesses in Phoenix		
Rank	**Business**	**Rating**
1.	Snooze AM Eatery	
2.	The Henry	
3.	Hopdoddy Burger Bar	
4.	Taco Guild	
5.	Angry Crab Shack	
6.	Joyride Taco House	
7.	Matt's Big Breakfast	
8.	Portillo's Hot Dogs	
9.	Rehab Burger Therapy	
10.	The Clever Koi	

Figure 3b. This table shows the top ten most frequently reviewed businesses in Phoenix.

Most Frequently Reviewed Businesses in Madison		
Rank	**Business**	**Rating**
1.	Bassett Street Brunch Club	
2.	Graze	
3.	Heritage Tavern	
4.	DLUX	
5.	Paul's Pel'meni	
6.	Journey Sushi & Seafood Buffet	
7.	The Old Fashioned	
8.	A Pig In a Fur Coat	
9.	Grampa's Pizzeria	
10.	La Taguara	

Figure 3b. This table shows the top ten most frequently reviewed businesses in Madison.

Figure 1a and 1b?

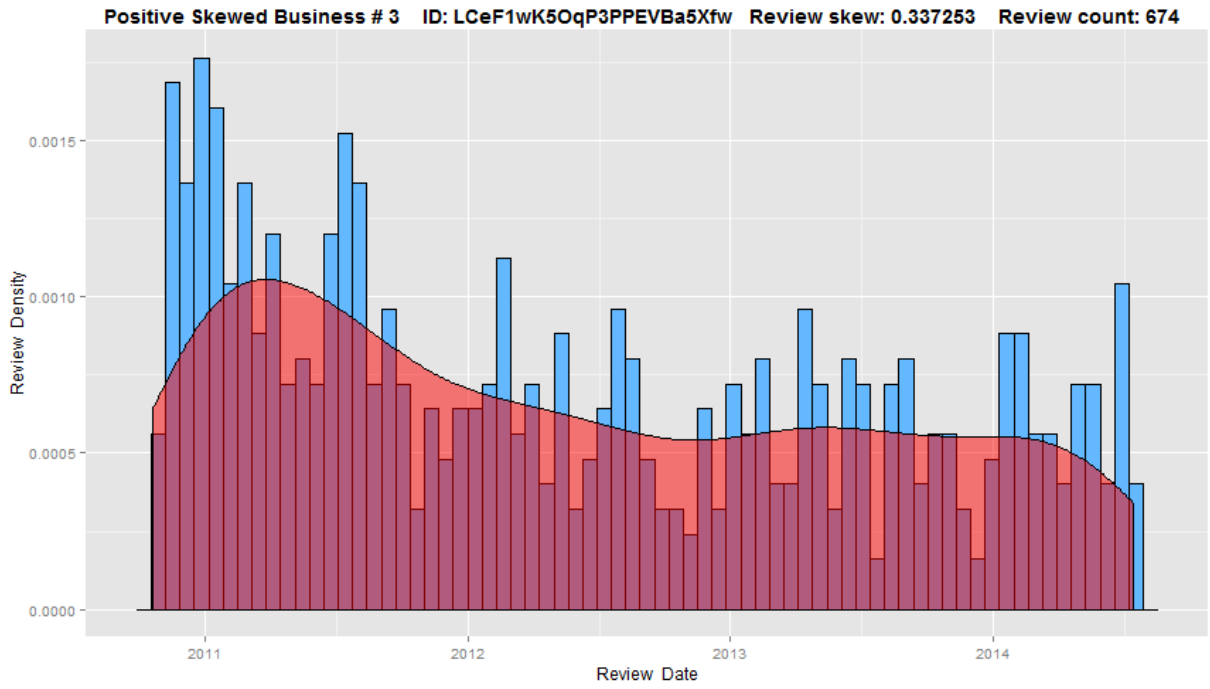


Figure 4a. This histogram illustrates review accumulation over time for a business called The BabyStacks Cafe (additional information: metro area: Las Vegas, earliest review: 2010-10-20, average star rating: 4.0). The BabyStacks Cafe has generated a majority of its reviews earlier in its existence, giving it a positive review skew.

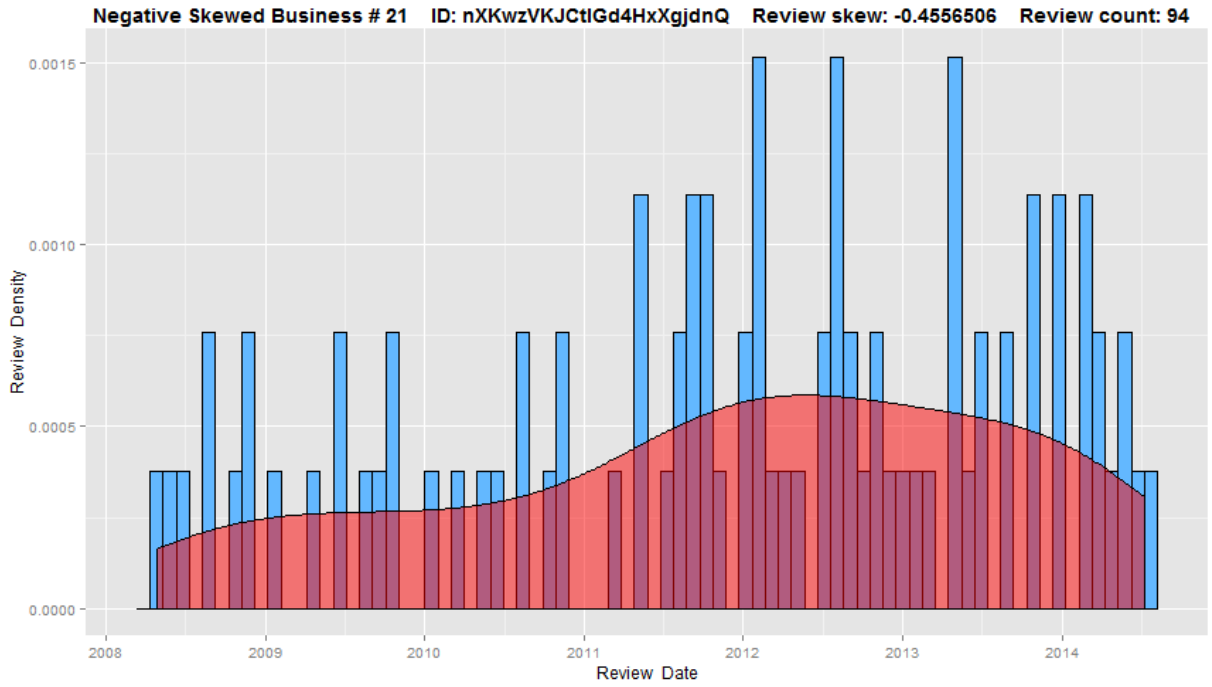


Figure 4b. This histogram illustrates review accumulation over time for a business called Saketini Japanese Sushi Bar and Lounge (additional information: metro area: Phoenix, earliest review: 2008-04-26, average star rating: 3.5). The Saketini Japanese Sushi Bar and Lounge has generated a majority of its reviews later in its existence, giving it a negative review skew.

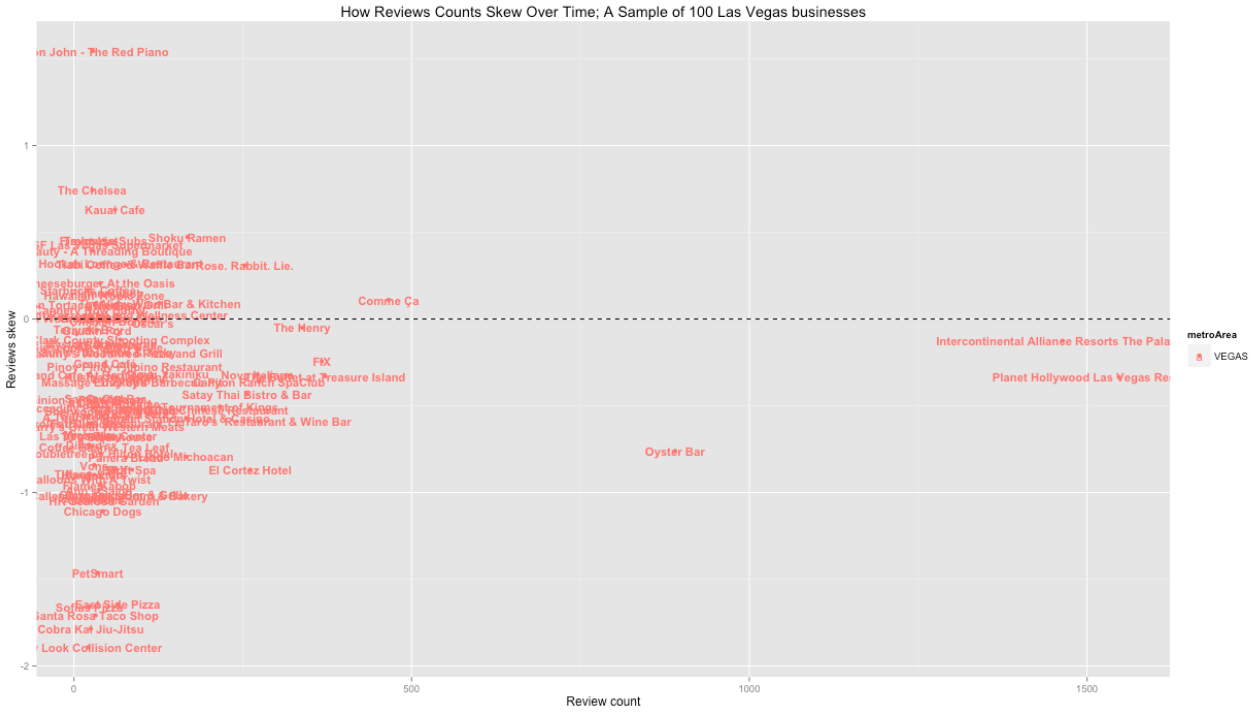


Figure 5a. Review skewness versus review counts for a random sample of 100 Las Vegas businesses. This and the following two graphs illustrate how the majority of businesses in Las Vegas, Madison, and Phoenix have a negative review skew, regardless of review counts. The majority of businesses have a review skew that falls between -1.0 and 0, meaning that these businesses have a received the majority of their reviews more recently in time during their respective existences. The lower the review skew of a business means the more recent the majority of that business's reviews have been received. The average review skew for our entire 37,334 business sample is -0.2555.

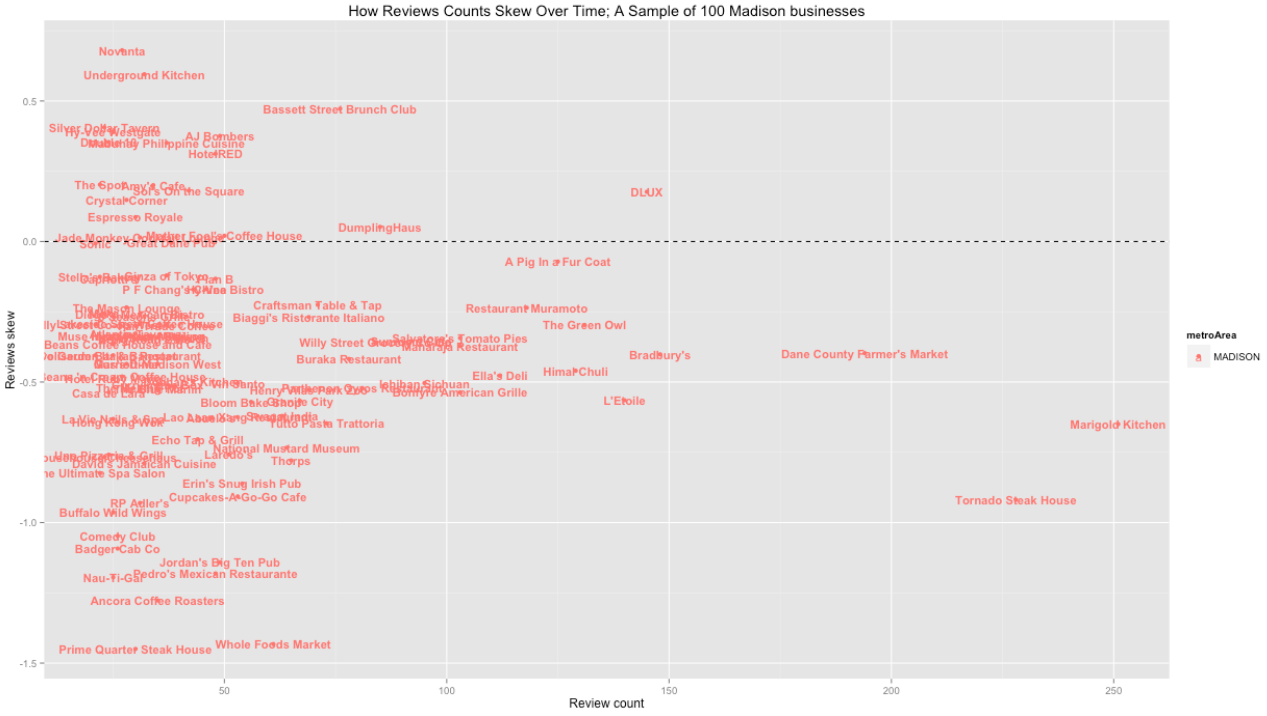


Figure 5b. Review skewness versus review counts for a random sample of 100 Madison businesses.

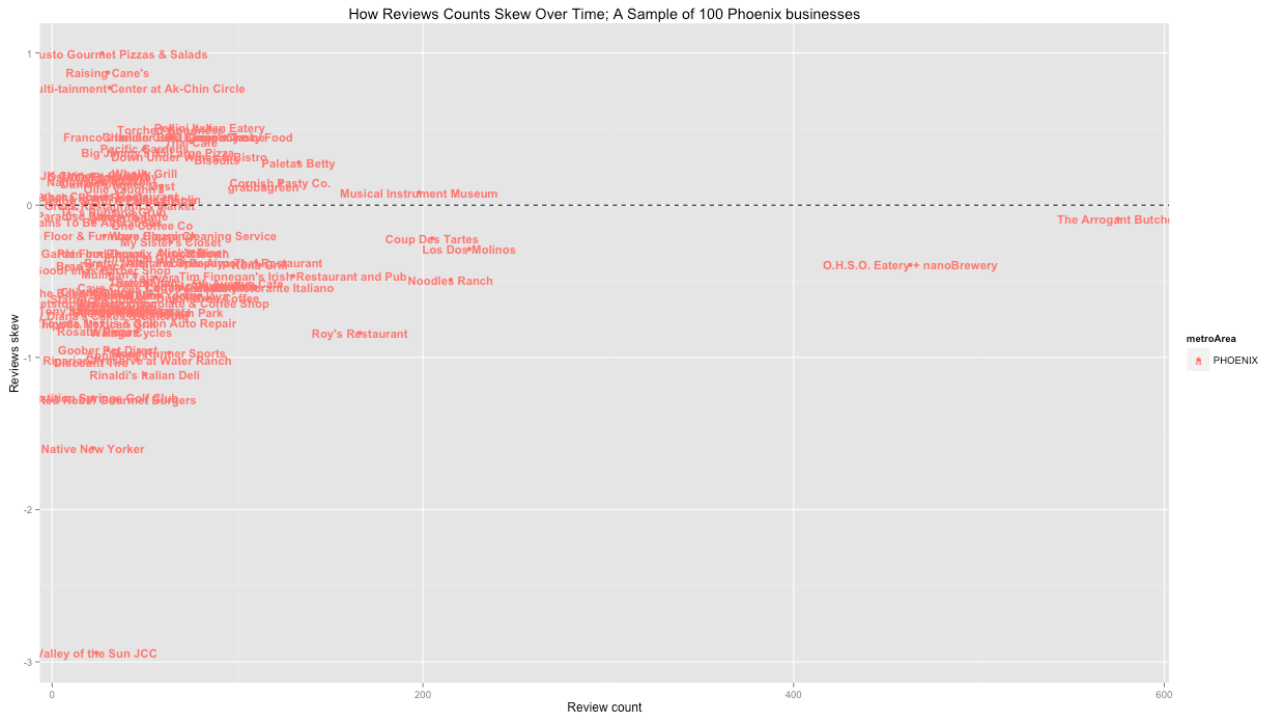


Figure 5c. Review skewness versus review counts for a random sample of 100 Phoenix businesses.

	A	B	C	D	E	F	G	H	I	J	K	L
1	CART Prediction Confusion Matrix											
2		yfit1			yfit2			yfit3			yfit4	
3		Pred:0	Pred:1		Pred:0	Pred:1		Pred:0	Pred:1		Pred:0	Pred:1
4	Actual:0	7962	335		7802	495		8167	130		7937	360
5	Actual:1	385	1596		329	1652		1770	211		1193	788
6												
7	base = open, stars, metroArea, popPersqMi, review_timing											
8	attribute = priceRange, romantic, intimate, touristy, hipster, divey, classy, trendy, upscale, casual											
9	rate = rate											
10	yfit1 = base + attribute + rate											
11	yfit2 = base + rate											
12	yfit3 = base											
13	yfit4 = base + attribute											

Figure 6. Of the four rpart model simulations, the first (yfit1) had the highest level of precision, accuracy and sensitivity.

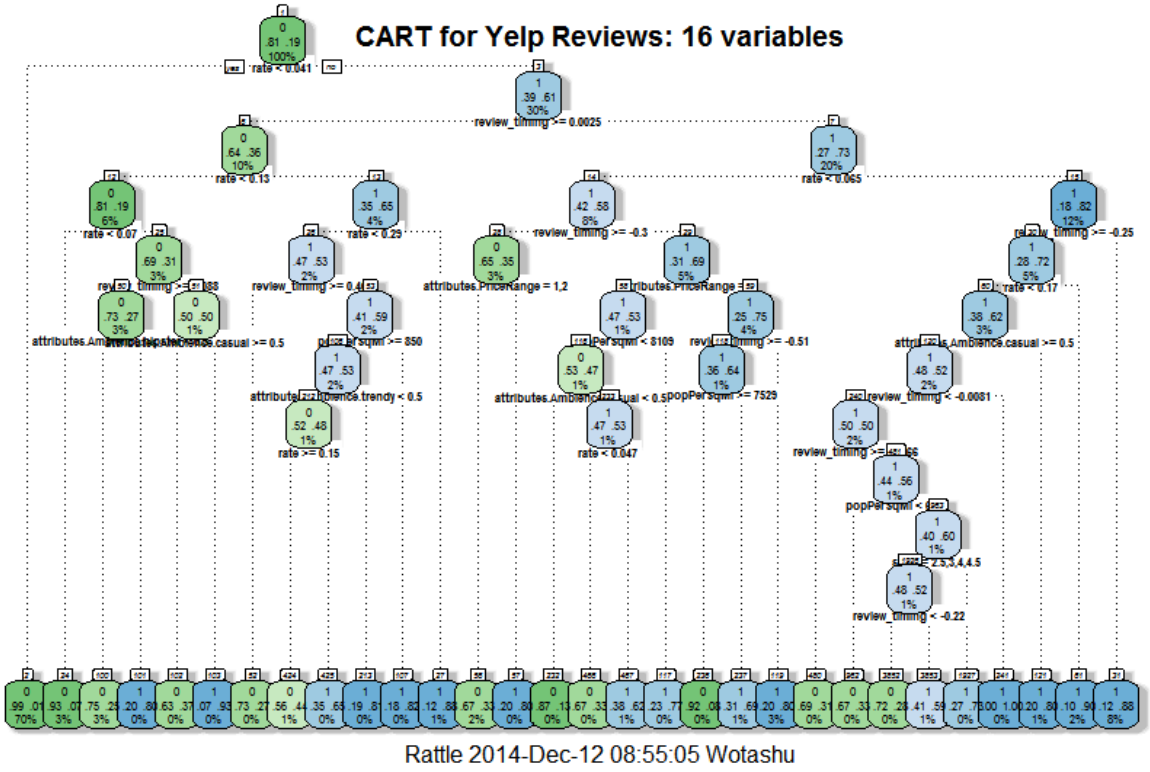


Figure 7a. *Decision tree for rpart model for 16 predictor variables (yfit1). The majority of businesses are filtered at the first branch based on the rate of reviews. The tree charts the determination of whether or not a particular business will receive more than 100 reviews.*

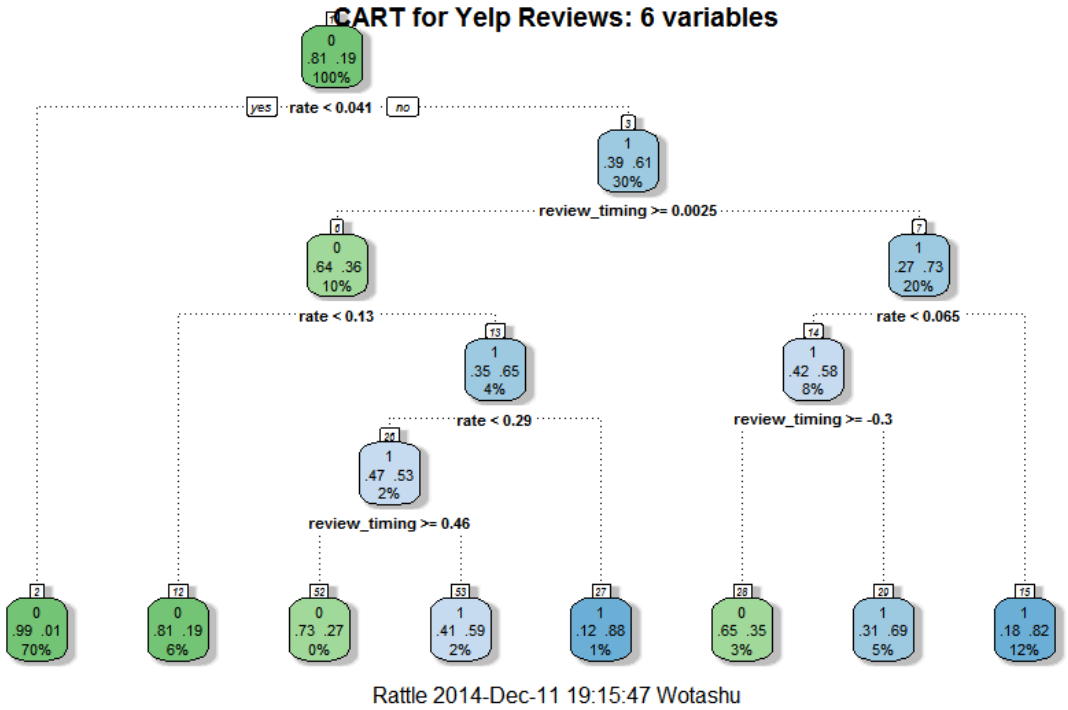


Figure 7b. *Decision tree for rpart model for 6 predictor variables (yfit2). The majority of businesses are filtered at the first branch based on the rate of reviews.The tree charts the determination of whether or not a particular business will receive more than 100 reviews.*

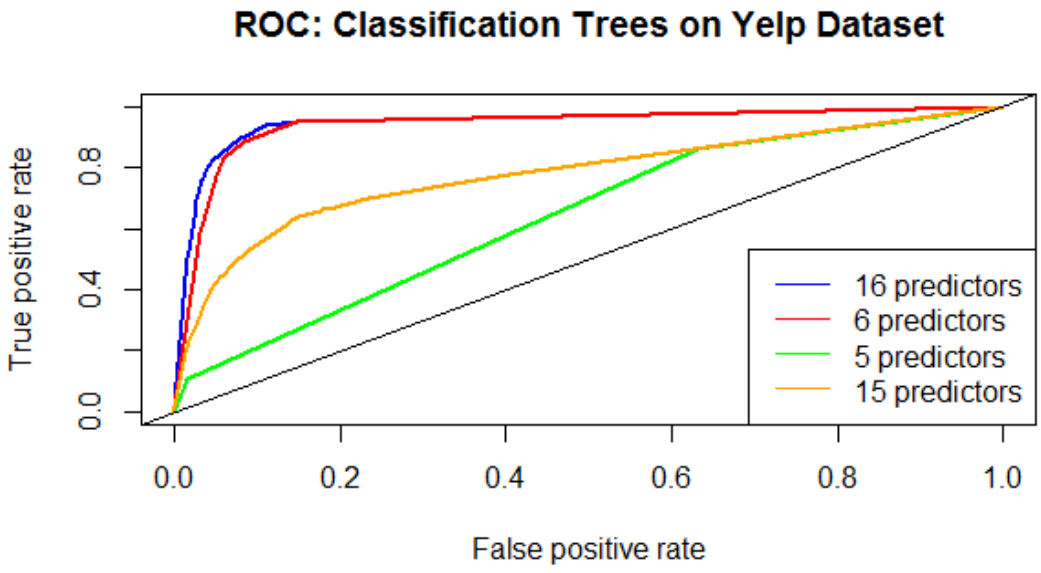


Figure 8. *ROC curves for rpart model. The inclusion of review rate (in the 16 predictor and 6 predictor case) yields the most predictive model.*

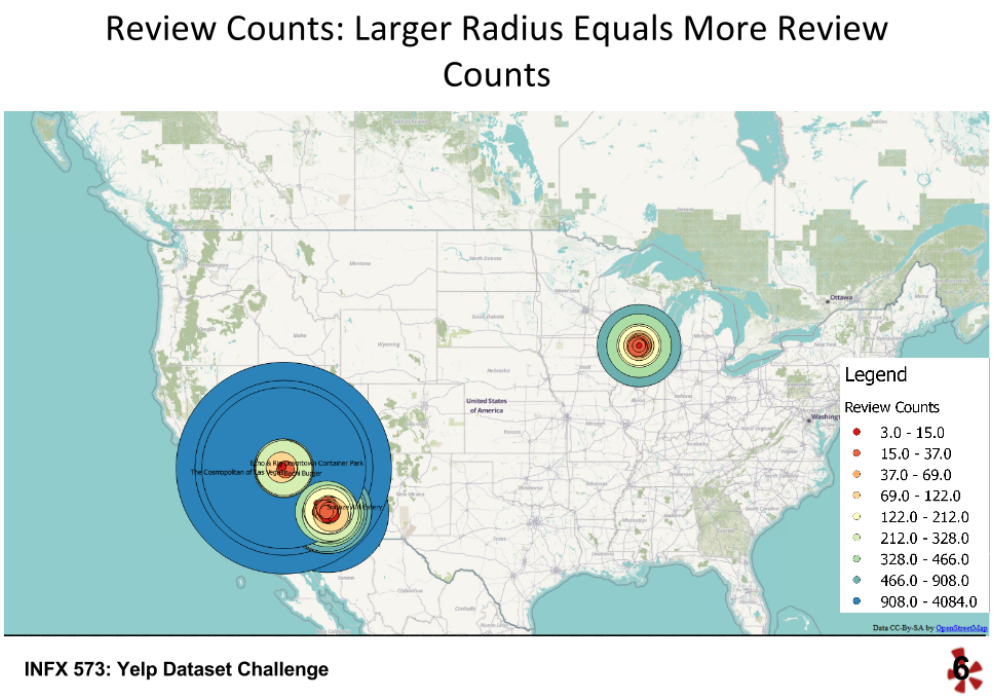


Figure 9a. *Review counts of businesses in the United States. A larger radius and bluer point indicates more reviews.*

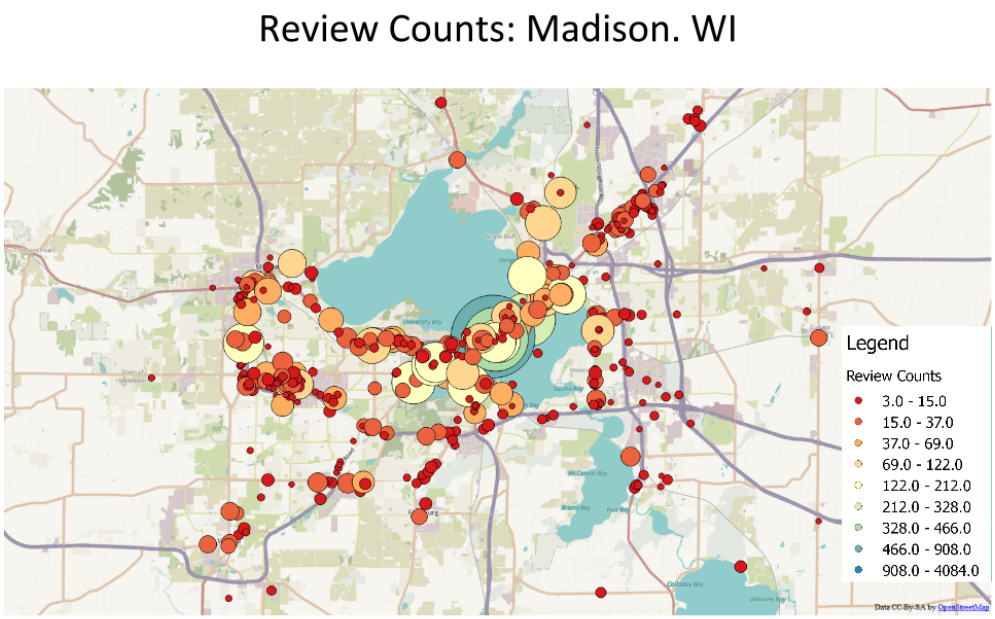


Figure 9b. *Map of Reviews Counts per business in Madison, WI.. A larger radius and bluer point indicates more reviews.*

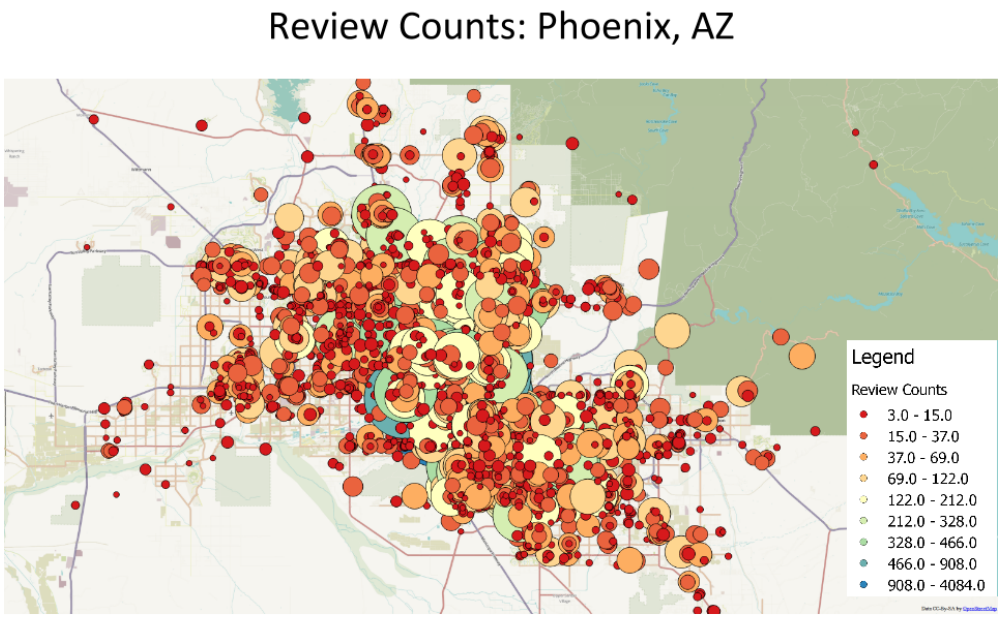


Figure 9c. *Map of Reviews Counts per business in Phoenix, AZ. A larger radius and bluer point indicates more reviews.*

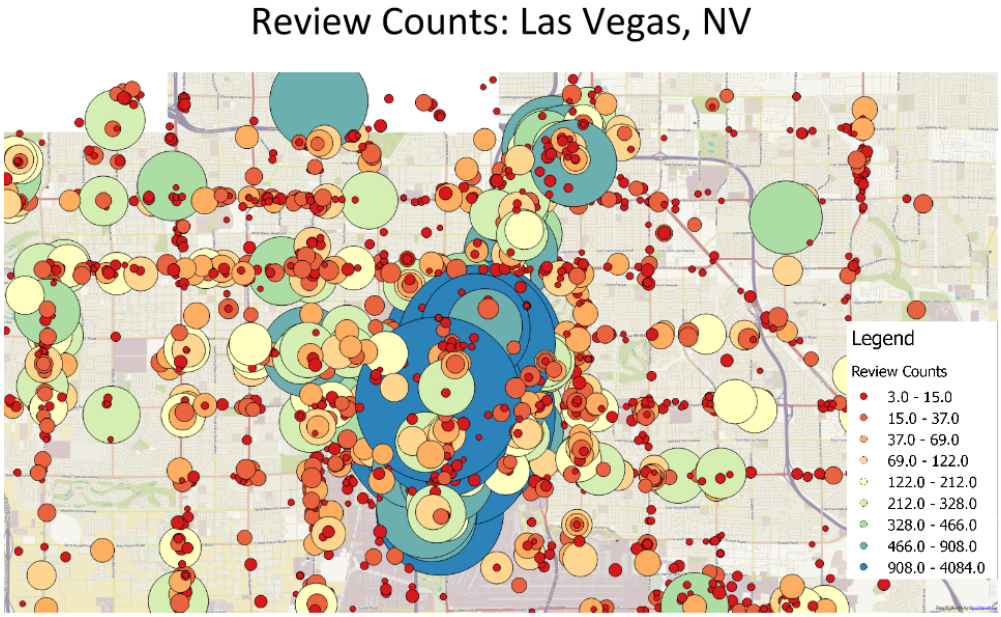


Figure 9d. *Map of Reviews Counts per business in Las Vegas, NV. A larger radius and bluer point indicates more reviews.*