Where should we take the kids? An analysis of family-friendly businesses in the Yelp dataset

Introduction

The amount of time that parents devote to their children has been increasing steadily since the 1990s[1]. The amount of money parents spend on their children has also increased by more than 50% from 1972 to 2007[2]. So, targeting families could be a good strategy to woo customers. Identifying what makes a business family-friendly may help attract families to a new business or guide improvements to an established business. Yelp makes a dataset of crowd-sourced reviews about businesses available for educational and academic purposes. The dataset includes information about more than 200,000 businesses worldwide, including business type, location and amenities as well as customer reviews. The proposed project is an exploratory analysis of businesses that market themselves as family-friendly. Using machine learning, I aim to identify common features, explore customer sentiment and predict business ratings.

Literature Review

Parents may invest more in their children than in earlier generations to help prepare them for college[1]. Taking kids to the theatre, to art lessons or to basketball camp are enriching, skill building opportunities that may set kids apart from their peers. In dual income families, leisure time is family time[3]. So when parents aren't working, they are with their kids. They may prefer restaurants or gyms that welcome children. Businesses, both big and small, are capitalizing on this "kidfluence"[4]. That is the direct or indirect influence that children exert on their parents spending. Whatever the reason, there is a strong market for family-friendly businesses, so attracting or retaining families could be critical to a company's success.

---

[1] Ramey, G., Ramey, V.A., Hurst, E. & Sacks, D.W. "The Rug Rat Race." *Brookings Papers on Economic Activity, (Spring 2010)*, pp. 129-199.

[2] Kornrich, S., & Furstenberg, F. "Investing in children: Changes in parental spending on children, 1972–2007." *Demography*, vol. 50, 2013, pp. 1–23.

[3] Fromm, J. & Vidler, M. "Marketing to This Powerful and Surprisingly Different Generation of Parents." New York, New York : American Management Association, 2015

[4] Williams, K.C. & Page, R.A "Marketing to the Generations." *Journal of Behavioral Studies in Business,* vol. 3, 2011, http://www.aabri.com/manuscripts/10575.pdf .

Dataset

The Yelp dataset was downloaded from https://www.yelp.com/dataset on May 5, 2020. The data is stored in JSON format in 6 separate collections: 1) Business, 2) Review, 3) User, 4) Tip, 5) Check-in, and 6) Photos. The current analysis focuses on only the business and review collections.

The complete business collection is 149 MB and contains 209,393 records. The complete review collection is 6.1 GB and contains 8,021,122 records. Only a subset of the businesses and reviews will be included in the current analysis. Business records with the attribute "good for kids" equal to true will be included. "Good for kids" is a self-selected attribute that each company has the option to select when they create or update their Yelp profile. It indicates that they are a family-friendly business. There are 55,527 businesses in the "good for kids" subset (18.6Mb). Each business is identified with a business_id, which is a unique identifier that is used to link to records in the other collections. Additional attributes that will be used in the analysis are described in Table 1 below.

Table 1. Attributes Description

Attribute	Datatype	Description	Example(s)
business_id	Character	22 character unique identifier for each business and primary key for linking with the review collection	--DaPTJW3-tB1vP-PfdTEg TvuwCoKIttCtx6kGYziziQ
name	Character	The business's name	Sunnyside Grill, Pioneer Community Park
city	Character	The city where the business is located	Phoenix, Toronto
state	Character	The state where the business is located	AZ, ON
stars	Factor	Star rating from 1 to 5, rounded to half-stars	2, 4.5, 5
review_count	Integer	Number of reviews	3, 7, 52
is_open	Integer	Indicator that the business is currently open	0 = open 1 = closed
categories	Character	Character string or array of business categories	Restaurants, Breakfast & Brunch Active Life, Parks, Dog Parks, Fishing
attributes	Character	Characteristics of the business where true indicates it is present and false indicates it is absent.	Wifi, RestaurantsTakeOut, Ambiance.Trendy, BusinessParking

review_id	Character	22 character unique identifier for each review	j0XdD-b_z2rJ7LYvpxgJrw e-ogzvZazshsG0vUlbxiWA
text	Character	A string with the full review text	"Wonderful experience! Totally exceeded all of my expectations. My daughter is six and loves her coaches. I myself now have confidence that I will reach my goals." "Great service, awesome food relaxed atmosphere. I've been going with my family since I was a kid."

In the original JSON format the attributes and hours fields are nested objects with key-value pairs (e.g., "Wifi" : True, or "Business Parking" : ["garage": True, "street": false]). To use the data in R Studio, the data must be unnested and stored in columns. Selecting attributes or hours for your Yelp profile is optional. The semi-structured JSON format is useful for this type of data as it does not have a fixed schema[5]. The implication for the current analysis is that the total number of attributes for each record varies and they are sparsely populated. The JSON files did not contain NULL values. Field values that are not available because they were not provided, were missing or were not relevant, are marked as NA[6].

Approach

A flow diagram of the analysis plan is shown in Figure 1. There are 5 stages to the data analysis:

1) Stating the question
2) Data cleaning and exploration
3) Building statistical models
4) Testing the models
5) Interpreting and communicating results

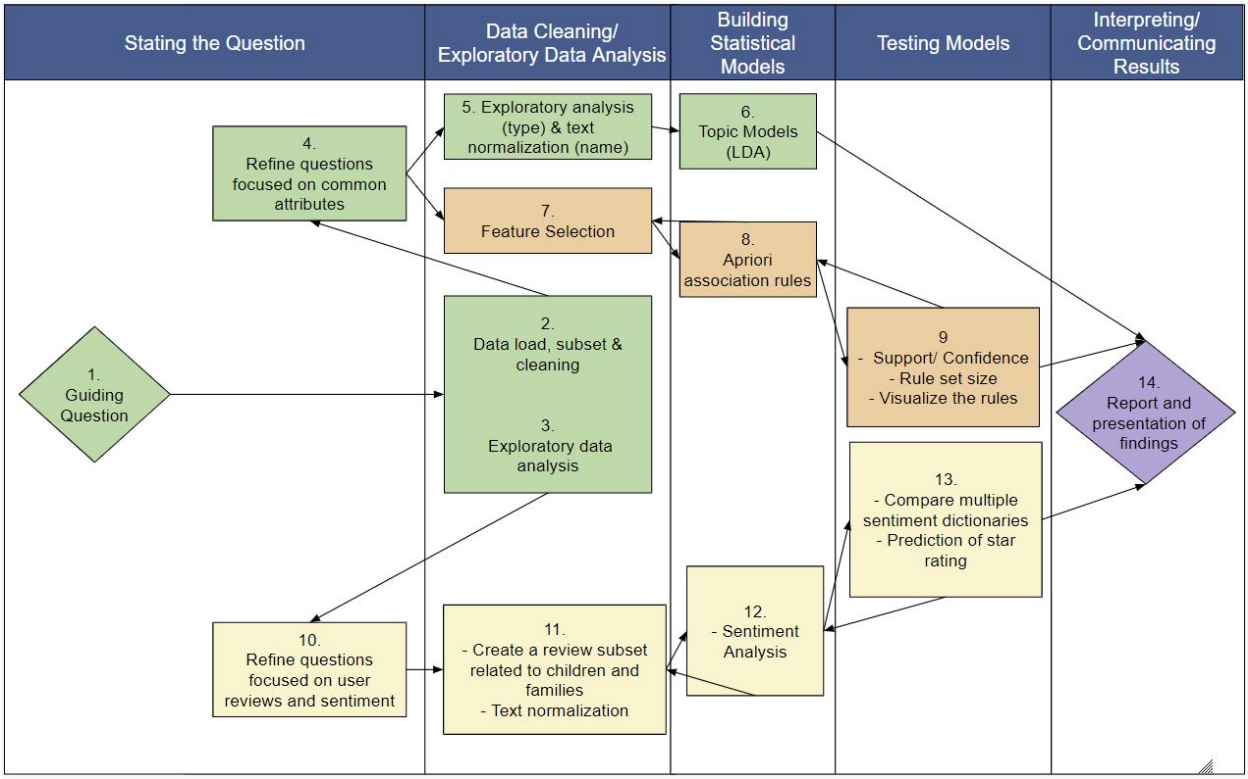
Each stage is iterative, and as new information becomes available, some steps need to be refined, repeated and retested[7]. This is reflected in Figure 1 by arrows that point back to earlier stages of the

[5] R. Elmasri, S.B. Navathe (2016). Fundamentals of Database Systems, 7th Edition, Addison-Wesley.
[6] Jonge, E., & van der Loo, M. (2013). An introduction to data cleaning with R. Statistics Netherlands, 53. Accessed on July 10, 2020 at: http://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
[7] Peng, R.D. & Matsui, E. "The Art of Data Science: A Guide for Anyone Who Works with Data." *Leanpub*, Sept. 2015.

analysis. RStudio (Version 1.2.5033) will be the software used for data cleaning, analysis and visualization.

Project datafiles and code can be found at: https://github.com/levinemi/capstone-yelp.git or https://drive.google.com/drive/folders/1kqAkVwOsOIuiqytZ5zPSFHH0UdN-5ui0?usp=sharing.

Fig 1. Overview of Analysis Methodology



Step 1: Guiding Question

The main objective of the proposed project is to conduct an exploratory data analysis of family-friendly businesses using the Yelp academic dataset. It is worth knowing if there are patterns, trends or relationships in this data, because it could inform families' choices about which businesses to frequent. Or it could be useful to businesses that are trying to attract families by identifying popular business types and common features. Sentiment analysis of the reviews will help us understand how customers feel about their experiences, so businesses can tailor their services to things customers enjoy.

Step 2: Data load, subset & cleaning

The business collection was small (149MB) and was loaded directly to RStudio, flattened and stored as an RStudio object using the jsonlite package. Family-friendly businesses were being identified using the "good for kids" attribute (see Table 2). 26.5% of all businesses in the Yelp dataset indicated that they were "good for kids".

Table 2 - Frequency count of businesses that are "good for kid" (N=209,393)

False	None	True	NA
12,932	76	55,527	140,858

The review collection was large (6.1GB). The Ryerson Data Science Tutor converted the file to CSV using a combination of Apache Spark and Python. The CSV (4.6GB) file was loaded to RStudio using the data.table package, which is designed for faster processing of large data. Given the large number of reviews and the size of the file, only a subset were used for analysis. Additional details on the sampling methodology for the reviews are provided in step 11.

Step 3: Exploratory data analysis

Initial exploration of the family-friendly businesses dataset revealed that several fields contain no NA values (e.g., name, city, state, stars, review_count and is_open). The businesses are located in 765 different cities in 22 states or provinces (see Figure 2). But more than 40% are concentrated in just a few cities, namly Las Vegas, Toronto and Phoenix, and their suburbs (e.g., Scottsdale, Mesa and Mississauga). The businesses are rated on a 5 star scale, rounded to the nearest half star. Most businesses get 3.5 or 4 stars. There is a fairly normal distribution for star ratings with a slight left skew as very few businesses receive 1 or 1.5 stars (see figure 3). 72.6% of the businesses are currently open.

Figure 2. Cities with the highest number of businesses

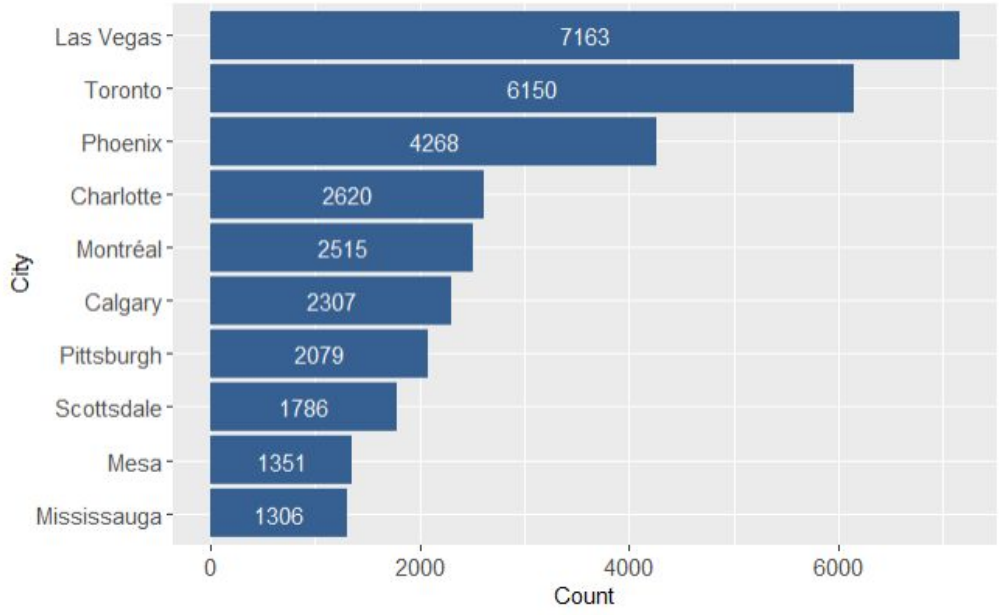
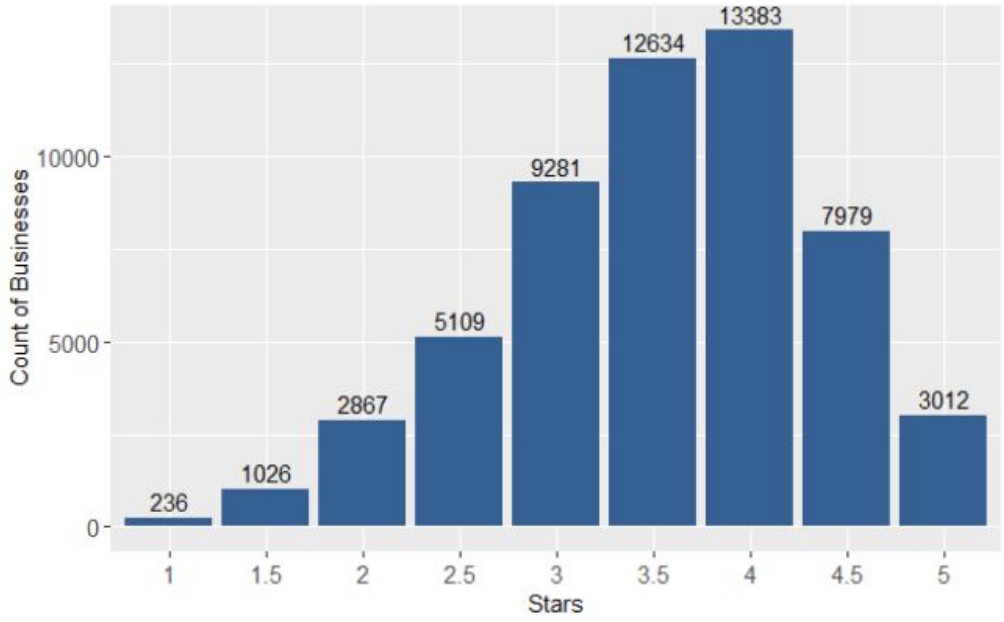
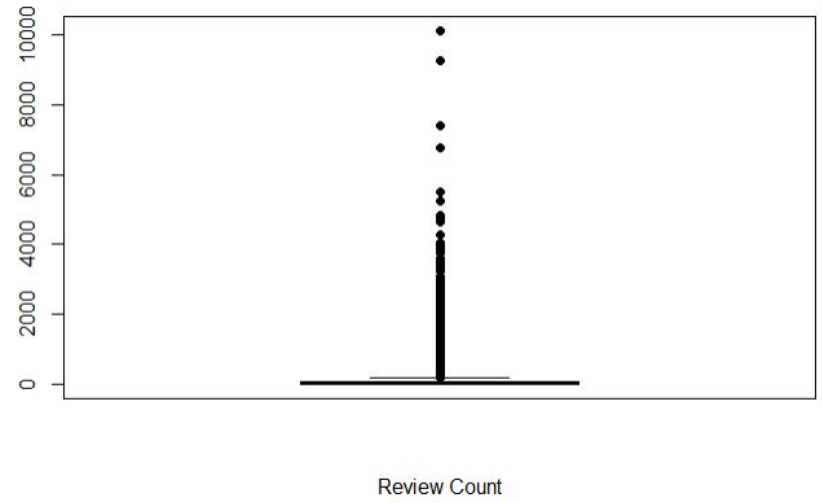


Figure 3. Histogram of star ratings



Based on the review count attribute there are 4,279,746 reviews for the businesses that are "good for kids". The distribution of reviews is extremely positively skewed. The mean number of reviews (77 per business) is greater than the median number of reviews (24 per business). There are 6119 outliers. Further analysis and treatment of these outliers is described in step 11.

Figure 4. Boxplot of review count for family friendly businesses



Businesses that create a profile on Yelp select one or more categories from a standard list to help with search filtering[8]. They are organized in a hierarchy of 1344 categories and subcategories. The family-friendly business subset contains 22 of the categories and 906 of the subcategories. The vast

[8] https://www.yelp.ca/developers/documentation/v3/category_list

majority of businesses select 1 to 3 categories (98.3%). While a small number select 4 or more categories as shown in Table 3.

Table 3. The number of categories listed for each business

# of Categories	1	2	3	4	5	6	7	8	9	10	11
# of Businesses	35481	15851	3263	701	186	31	10	3	0	0	1

"Restaurants" was by far the most popular category. More than half of the businesses selected "restaurants" as one of their categories (see Figure 4). The next most common categories are "food", "active life", "nightlife" and "beauty and spas". The subcategories show a similar pattern with 9 of the 10 most popular subcategories being food related (see Figure 5).

Figure 4. The number of businesses by Yelp category for family-friendly businesses

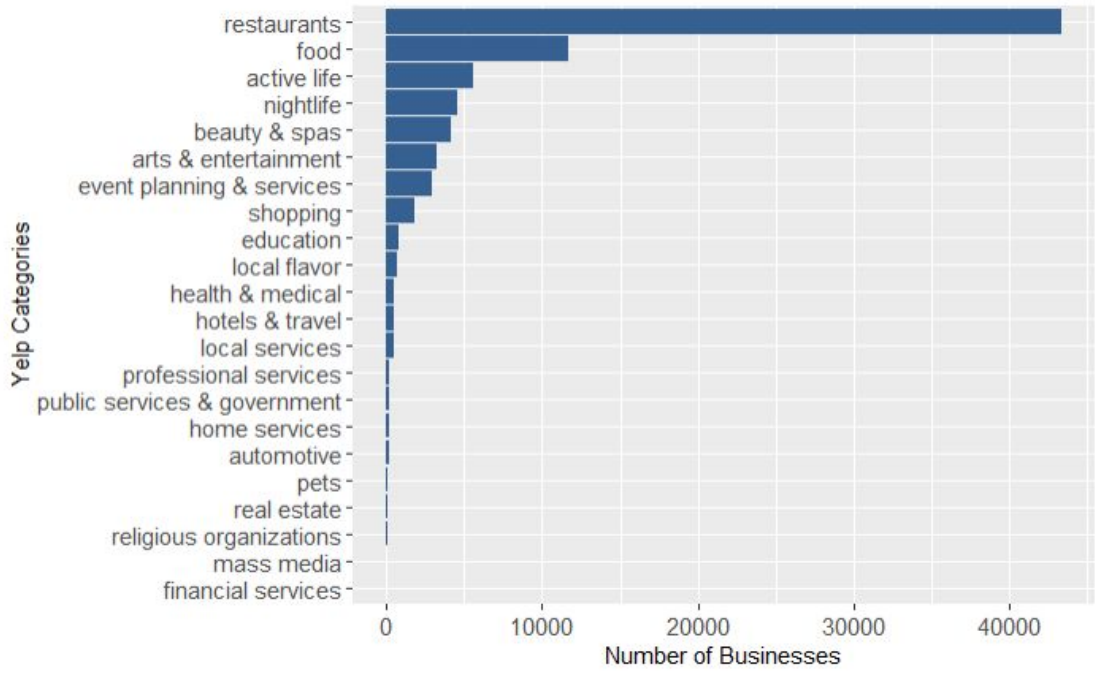
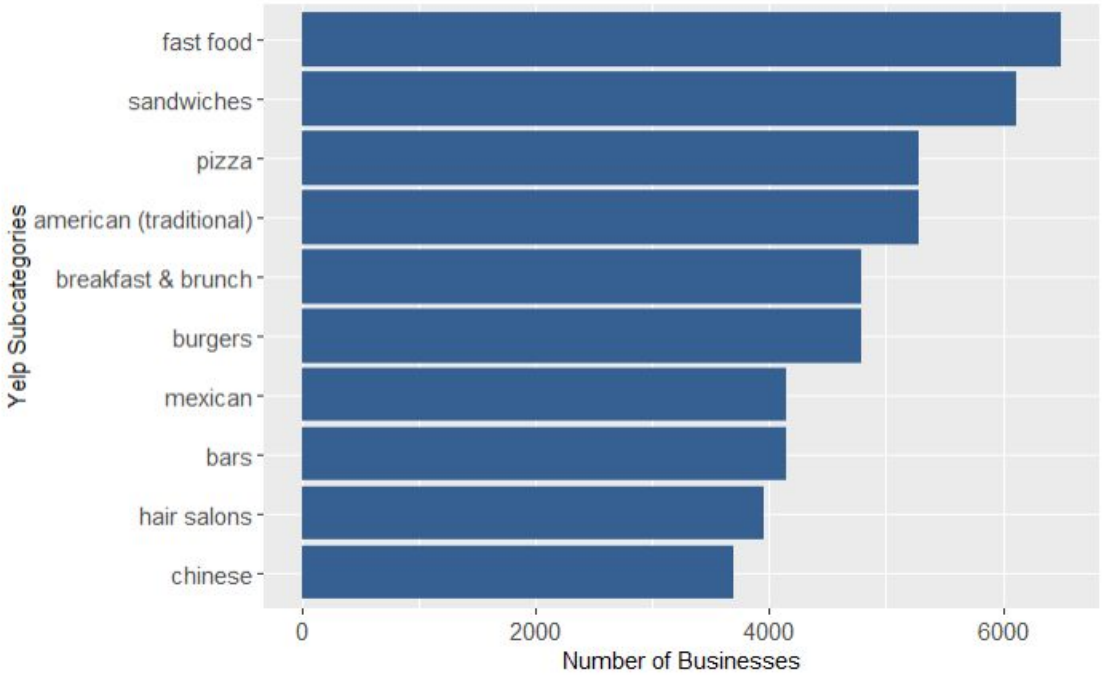


Figure 5. The 10 most popular Yelp subcategories for family friendly businesses



Step 4: Refine questions focused on business attributes

The remaining analysis centres on the 43,368 businesses that selected “restaurants” as one of their Yelp categories. Narrowing the scope of the exploratory analysis is appropriate because more than three quarters of the businesses in the “good for kids” subset are restaurants. This overrepresentation could skew results of the exploratory analysis to reflect restaurants more than other business types. A separate analysis is needed to learn about the features or customer preferences for other business types (e.g., "religious organizations”, “shopping” or “arts & entertainment”).

The restaurants were selected for the current analysis because most of the business features/attributes in the dataset relate to restaurants (e.g., Ambience, Reservations, Happy Hour, Outdoor Seating). This means that more features/attributes are potentially relevant for the association analysis. Narrowing the focus to “restaurants” may also reduce some of the noise in the analysis of customer reviews.

The revised guiding questions for the project and methods for answering them are:

- What are the characteristics of family-friendly restaurants?
    - Exploratory analysis of family friendly restaurant characteristics
    - Explore subcategory frequencies and correlations
    - Analyse word-pairs in restaurant titles
    - Topic modelling, using latent dirichlet allocation, of restaurant titles
- Are there combinations of business features related to good reviews?
    - Use dimensionality reduction, including a comparison of feature selection techniques, to find features related to high star ratings
    - Identify association rules for feature sets that relate to having a star rating of 4 or more.

- What matters most to customers that are satisfied or dissatisfied with their experience?
    - Sentiment analysis of customer reviews
    - Explore frequency and type of positive and negative terms in reviews
    - Compare the type of positive and negative terms in reviews with high and low star ratings
    - develop a classification model for star rating using review sentiment and features identified in the association rules analysis

Step 5: Exploratory Analysis (Type) & Text Normalization (Name)

Trends in restaurant types

Most family-friendly restaurants serve fast food or foods that kids like, such as sandwiches, pizza and burgers (Table 4). But looking at individual subcategories doesn't provide a very clear picture of the most popular restaurant types because businesses can select multiple categories. Identifying popular combinations of subcategories may help identify trends in family friendly restaurants. Unfortunately the most popular subcategory pairs are also the most popular subcategories so frequency counts alone don't tell the whole story. To understand restaurant types, I need to find the yelp subcategories that are more likely to occur together than with other subcategories in the dataset.

Table 4. The top 10 subcategories for family-friendly restaurants

Yelp Subcategory	Number of Businesses
fast food	6503
sandwiches	6110
pizza	5284
breakfast & brunch	4790
burgers	4788
mexican	4151
chinese	3695
italian	3470
salad	2346
cafes	2326

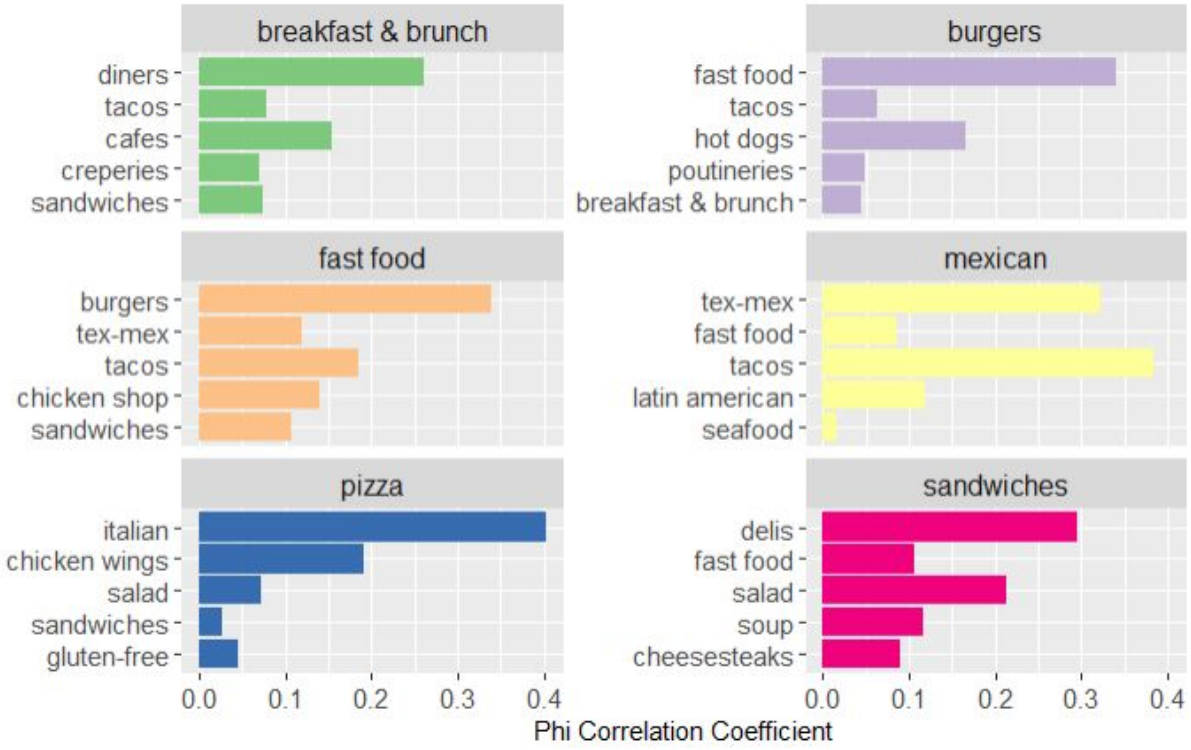
Figure 6 shows the subcategories that are most strongly associated within a restaurant (using phi coefficients). The subcategories had to appear together in at least 100 restaurants to be included in this analysis. The first finding is that the associations between subcategories are not strong. The highest

correlations are between “pizza” and “italian” (ϕ = 0.40), “mexican” and “tacos” ((ϕ = 0.38) and “burgers” and “fast food”(ϕ = 0.34). Given the popularity of those combinations at fast food restaurants, I would have expected stronger associations. This may indicate that offering familiar combinations of food types is good, but not necessary for a family-friendly restaurant.

Second, some subcategories appear multiple times. For example, “tacos” is associated with “breakfast & brunch”, “burgers”, “fast food”, and “mexican”. And “fast food”, in addition to being a popular type on its own, is associated with “sandwiches”, “mexican”, and “burgers”. So, if restaurants want to attract families, they should consider offering food quickly and including tacos on the menu.

Finally, a few niche subcategories appear with negligible but positive associations. For example, “breakfast & brunch” and “creperies” (ϕ = 0.07), “burgers” and “poutineries” (ϕ = 0.05), “pizza” and “gluten-free” (ϕ = 0.05), and “sandwiches” and “cheesesteaks” (ϕ = 0.09). While the associations are very weak, they occur in at least 100 restaurants in the dataset. Which means that some family-friendly restaurants find success by offering less common fare.

Figure 6. Top 6 subcategories combinations for family-friendly restaurant subcategories



Restaurant name cleaning and normalization

The next step in the analysis was to determine if restaurant names can provide additional information about restaurant themes.

The name field is free-text and required additional cleaning and normalization for analysis. First, restaurant names were tokenized, including removal of punctuation and extra white-space. Second, stop

words were removed. Stop words are very frequent words, like *in, a,* or, *the.* These words are removed during bag of words analysis because they are very common in text but are not typically meaningful for analysis[9] English, French and Spanish stop words (e.g., the, and, le, el, des) were all present in the dataset. The Snowball stopword lexicon was used, because it provides stop words for multiple multiple languages beyond English[10, 11]. Third, numeric digits and terms that were 2 characters or shorter were removed. Finally, terms with very high or very low frequency were removed. The term "restaurant" was the most frequent word. It appeared 11052 times in the dataset. It was removed because of its frequency and because it did not add new information. All the businesses in the analyses were restaurants. There were 630 terms that each appeared only 1 time in the dataset. These low frequency words were also removed. Infrequent words are removed because their association to other words adds noise to the analysis[12].

Step 6: Topic Modelling

Word Frequencies

The 10 most common terms in the restaurant tiles are shown in Table 5. Some of the terms were similar to the most common YELP subcategories (e.g., pizza, mexican). But other terms were new types of food (e.g., sushi) or were terms that describe the experience (e.g., grill, bar, kitchen, house, cuisine).

Table 5. Ten most frequent words in the restaurant titles

Word	Count
pizza	9063
grill	7230
cafe	5433
bar	4071
sushi	3772
kitchen	3030
house	2793
mexican	2678

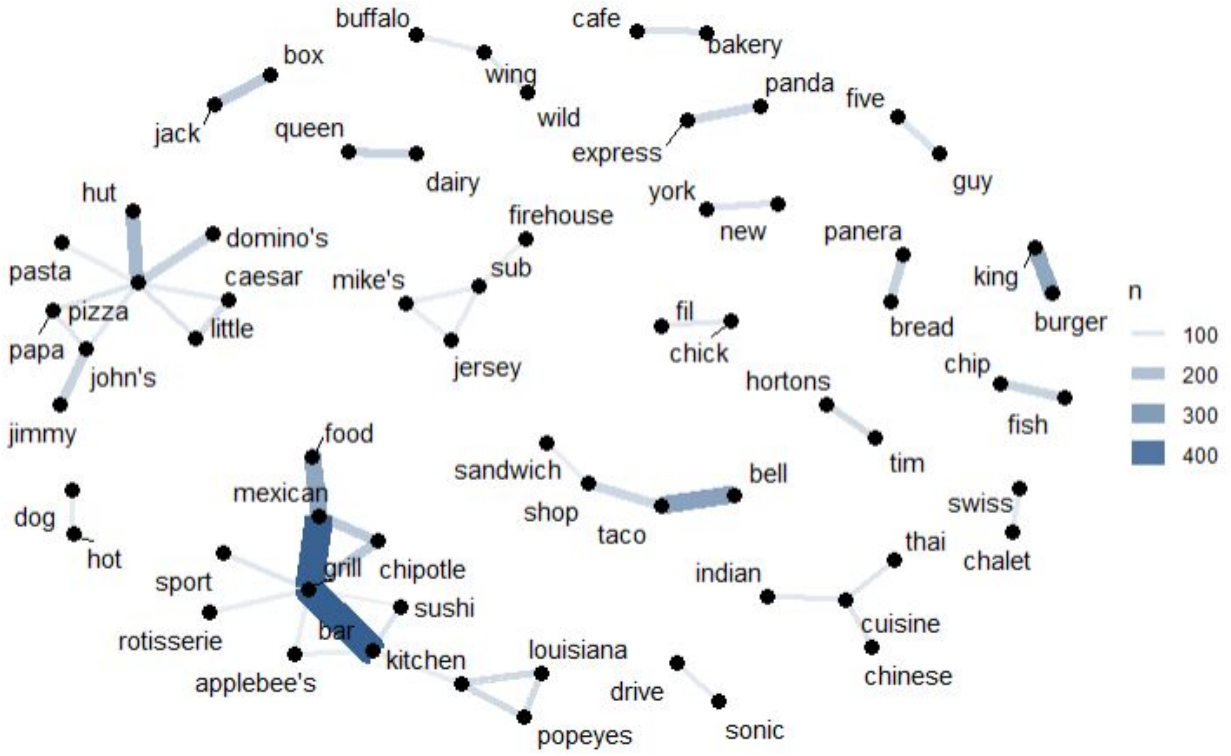
[9] Silge, J. & Robinson, D. "Text Mining with R: A Tidy Approach." O'Reilly, March 2020, accessed on May 18, 2020 at https://www.tidytextmining.com/.
[10] Porter, M.F. "Snowball: A language for stemming algorithms." Snowball.org, accessed on July 20, 2020 at https://snowballstem.org/texts/introduction.html.
[11] Benoit, K. Muhr, D., & Watanabe, K. "Package 'stopwords'." cran.r-project.org, accessed on July 20, 2020 at https://cran.r-project.org/web/packages/stopwords/stopwords.pdf.
[12] Ganesan, K. "Tips for Constructing Custom Stop Word Lists." Accessed on July 20, 2020 at http://kavita-ganesan.com/tips-for-constructing-custom-stop-word-lists/#.Xx8P855KiUk.

taco	2428
cuisine	2203

To introduce more context to the analysis of titles, word pairs (or bigrams) within a restaurant title were explored. It's difficult to see patterns in the word pairs from counts alone. So a network diagram was created to show the most common word pairs. Only pairs that appeared at least 75 times in the dataset were included in the diagram for visual clarity.

The diagram in Figure 7 highlights common words as nodes (e.g., grill, pizza, cuisine). The diagram also highlights that the most common word pairs are names of fast food chains (e.g., dairy queen, pizza hut, burger king, panera bread)) or common food pairings (e.g., fish and chip, bakery and cafe). The terms "grill" and "bar" seem to be popular for many types of food (e.g., rotisserie, sushi, mexican) and ambiance (e.g., sport). Excluding less frequent pairs could be seen as a limitation of the analysis, because it does favors restaurant chains that have multiple locations with the same title.

Figure 7. Network diagram



Latent Dirichlet Allocation

Topic modeling techniques are unsupervised machine learning algorithms used to automatically discover themes in a set of documents. I used the topicmodels[13] package in R to do LDA with Gibbs sampling.

LDA is an algorithm for inferring latent or underlying topics from a set of documents[14]. In the current analysis each restaurant name is a document. The goal is to infer the unknown topics from the words in restaurant names. LDA does this by breaking down the joint probability distribution of words and documents into the dot product of the joint distributions of word and topic and topic and document[15]. Over many iterations, the relative importance of a topic in a document and a word in a topic is adjusted. Eventually, you reach a steady state for the relative importance values and the steady state estimates indicate the distribution of words to topics and topics to documents.

Gibbs sampling is one type of sampling that can be used for the LDA iterations. Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method for approximating a probability distribution when direct sampling is difficult. You start by randomly assigning words to topics and topics to documents. For each iteration, you are looking to assign one word to a topic given the current assignment of all the other words. Because you start with a random assignment of relative importance, you don't use the early iterations as part of the sample. The burn in value indicates the number of iterations to throw away before iterations are sampled for your estimate[16].

My analysis used a total of 6000 iterations, with a burn in of 4000. I selected every 500th iteration from the remaining 2000 iterations. I used 5 different starting points and had the program return the one with the highest posterior probability.

One parameter that needs to be set upfront in LDA is k, the number of topics. I tested values of k from 5 to 20 to find a useful set of topics. I inspected the results to see if they produced practical categories. In all cases, the LDA topics developed from the business names were not clear or usable.

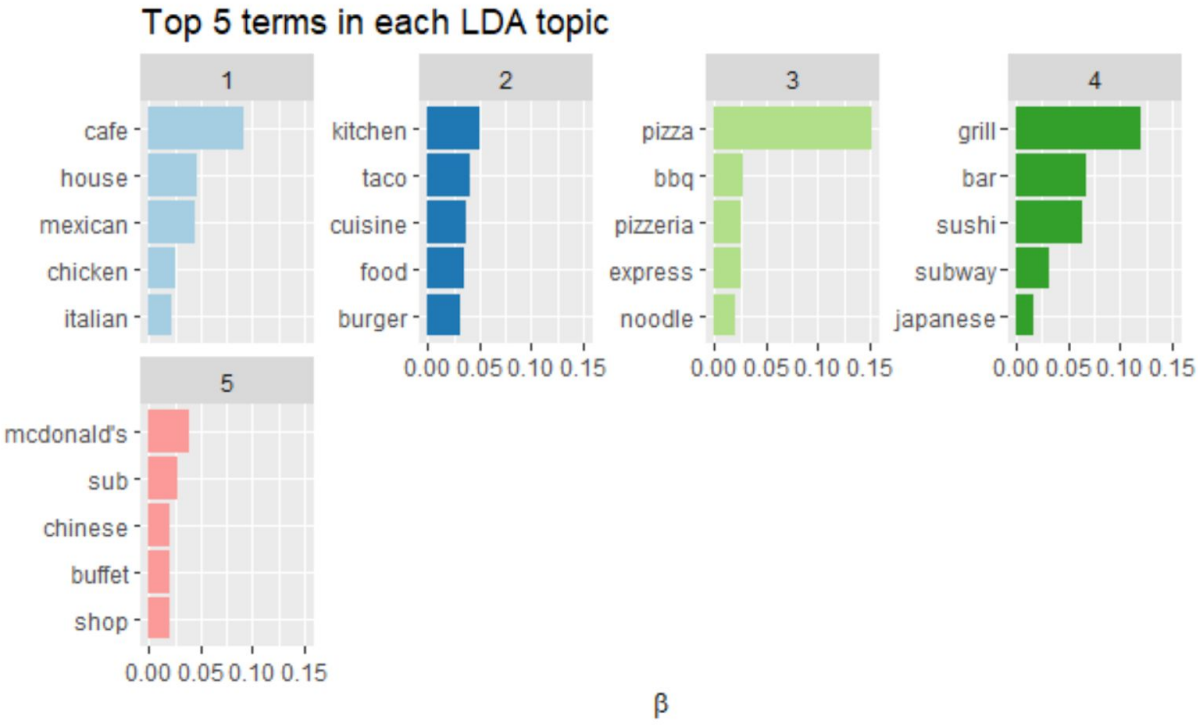
I had anticipated the higher values of k to produce better results. Lower values of α (defined as 50/k[17]), are appropriate when each document contains a mixture of just a few or only one topic. The names of restaurants are short and should likely contain only one topic, so I expect lower levels of α to be more successful. The LDA approach uses sparse vectors, long vectors containing mostly 0s because many

[13] Grün et al. "Package 'topicmodels'." April 19, 2020. Accessed on June 5, 2020 at https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf.
[14] Awati, K. "A gentle introduction to topic modeling using R." eight2late.wordpress.com, accessed on June 15, 2020 at https://eight2late.wordpress.com/2015/09/29/a-gentle-introduction-to-topic-modeling-using-r/.
[15] Tomar, A. "Topic modeling using Latent Dirichlet Allocation(LDA) and Gibbs Sampling explained!" Medium, accessed on June 15, 2020 at https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045 .
[16] Grün, B. et al. "Package 'topicmodels'." cran.r-project.org, accessed on June 16, 2020 at https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf.
[17] Grün, B. & Hornik, K. "topicmodels: An R package for fitting Topic Models". *Journal of Statistical Software,* 40(13), 2011, pp. 1-30.

words only occur in a handful of documents[18]. And while the number of distinct terms was not high (14597), I ran into performance issues at higher levels of k. Scalability is a known issue with LDA modelling[19].

One metric for LDA is beta, the probability of each term being generated from each topic. Figure 8 shows the top 5 terms for each topic in a model with k=5 topics. At a value of k= 5, the probabilities of each term per topic are quite low, less than 15%. The LDA reveals some topics that are similar to those identified by the categories. Topic 3 seems to relate to fast food pizza and topic 4 seems to relate to japanese restaurants. But both topics include words that seem unrelated to those topics (e.g., noodle and subway, respectively). For the other topics, the beta values are all less than 10% and the terms don't seem to hang together well.

Figure 8. Top 5 terms in each LDA topic, where k=5



Gamma is another metric of LDA. It is the per-restaurant-per-topic probabilities. Gamma is useful for seeing how well the LDA identifies a topic for each restaurant. You can assess if the model does a good job of assigning restaurants to topics by comparing the relative importance of the assigned topic (e.g., topic with the maximum gamma probability) with the next highest topic probability for each restaurant. If the relative importance of the assigned topic is high, that means that the probability of the assigned

[18] Jurafsky, D. & James H., M. " Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition." 3rd ed. Stanford University, UK: Online, 2019, chapter 6, accessed on June 27, 2020 at https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

[19] Baroni, M., Dinu, G. & Kruszewski, G. " Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." Association for Computational Linguistics , Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 238–247.

topic is much greater than the probability of the other topic. If the relative importance is close to 1, that means that the probability of a restaurant being assigned to the two topics is similar. Table 6 shows that three-quarters of restaurants in the sample had a relative importance near 1.

Table 6. Count and Percent of Restaurants by Relative Importance of Assigned Topic

Relative Importance	Frequency	% Total
1.00 - 1.49	31,951	74%
1.50 - 1.99	8,139	19%
2.00 - 2.49	2,263	5%
2.50 -or more	723	2%

One of the benefits of LDA is that each document can exhibit multiple topics. But this analysis didn't take advantage of this property because each restaurant title was short and it's assumed to contain only topic. In the end, topic modelling using LDA did not add new information to the exploratory analysis. No clear topics can be identified by business name. Moreover, for all values of k that were tested, more than 60% of restaurants were equally likely to be assigned to more than 1 topic. The self-identified categories are a more useful way of grouping than the topics created through LDA.

Step 7. Feature Selection (Data Cleaning)

The next step in the analysis was to identify common business features or attributes among family-friendly restaurants with high star ratings. The features selected will be used in an association analysis and in the classification models.

In the YELP dataset the business features are a series of binary attributes where true means present, false means absent and NA indicates the information was not provided. There were 70 total features relevant to family-friendly restaurants. Many had high numbers of records with NA values. Forty-two features with NA for more than 30% of their values were removed. The feature "good for kids" had no variance as it was the original inclusion criteria. It was also removed.

The remaining 27 features are listed in Table 7.

Table 7. Business Features

Ordinal	Binary (0 or 1)		
Restaurants Price Range (1-4)	Restaurants Reservations	Ambience Casual	Ambience Hipster
Noise Level (1-4)	Restaurants Take Out	Ambience Intimate	Ambience Trendy
Nominal	Bike Parking	Ambience Classy	Ambience Upscale
WiFi	Outdoor Seating	Ambience Romantic	Ambience Touristy
Restaurants Attire	Restaurants Good for Groups	Ambience Divey	Business Parking Garage
Alcohol	Has TV	Business Parking Street	Business Parking Validated
Stars	Restaurants Delivery	Business Parking Lot	Business Parking Valet

The ordinal and nominal features were converted to rank and dummy variables, respectively. The star attribute was simplified into two levels. Those restaurants with "3.5 or less" and those with "4 or more" stars.

Three different feature selection techniques were used to identify the business features that are most relevant for the association rules analysis. The techniques were selected to represent different feature selection families (i.e. filter, wrapper) and for use with categorical variables. Since the goal of the apriori association analysis is to find features common to restaurants with 4 or more stars, stars is the dependent variable in each of the feature selection analyses.

Filter techniques use statistical measures to score the relatedness between features, which can then be sorted or filtered to choose the most relevant. They are calculated independently from model construction[20]. This can be an advantage if the model or subsequent steps in the analysis have not been finalized. It can also be seen as a disadvantage, because the features selected aren't optimized for your model, possibly increasing error in the model. Another issue with filter techniques is that depending on the statistic used, different features may be identified as the most important.

Information gain and chi-squared techniques were the filter techniques selected for the current analysis. Information gain is a measured based on the entropy of a system[21]. Entropy is the amount of

[20] Jiarpakdee, J., Tantithamthavorn, C., & Treude, C. "AutoSpearman: Automatically Mitigating Correlated Software Metrics for Interpreting Defect Models." Proceedings - 2018 IEEE International Conference on Software Maintenance and Evolution, ICSME 2018, 92–103.

[21] Alhaj TA, Siraj MM, Zainal A, Elshoush HT, Elhaj F (2016) Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. PLoS ONE 11(11): e0166017. https://doi.org/10.1371/journal.pone.0166017

information gained by knowing an attribute's value. Features were ranked from highest to lowest entropy and the 10 features with highest score were identified. Chi-square is a test of the independence between two independent factors [22]. In feature selection, the dependent variable is compared with each feature and the features that are more dependent are more relevant to include in the analysis. Features were ranked from highest to lowest and the 10 features with highest χ^2 were identified.

Wrapper techniques use a specific machine learning classification algorithm. Features are selected based on their classification performance. Multiple combinations of features are evaluated to find the best subset metrics based on an evaluation criteria.[23] Because wrapper techniques are model oriented, they can improve performance if you plan to use the model for analysis. A key challenge of wrapper techniques is they are more computationally expensive than filter techniques[24]. The wrapper technique selected for the current analysis is stepwise regression.

The top features identified by each technique are listed in Table 8.

Table 8. Top business features by feature selection method

Information Gain	Chi-square	Stepwise Regression
Business Parking Street †	Business Parking Street †	Noise Level 2 †
Noise Level †	Noise Level †	Noise Level 3 †
Ambience Trendy †	Ambience Trendy †	Noise Level 4 †
Ambience Classy †	Has TV †	Has TV †
Ambience Hipster ✦	Ambience Classy †	Ambience Casual †
Ambience Casual †	Ambience Hipster ✦	Ambience Classy †
Has TV †	Ambience Casual †	Ambience Trendy †
Bike Parking ✦	Bike Parking ✦	Business Parking Street †
WiFi Free	Alcohol Full Bar ✦	Alcohol Full Bar ✦
WiFi No	Ambience Intimate	--

† Feature identified in 3 feature selection techniques; ✦ Feature identified in 2 feature selection techniques

[22] Mendenhall, William, Robert J. Beaver, and Barbara M. Beaver. Introduction to Probability and Statistics. Belmont, CA: Thomson/Brooks/Cole, 2006. Print.

[23] Jiarpakdee, J., Tantithamthavorn, C., & Treude, C. "AutoSpearman: Automatically Mitigating Correlated Software Metrics for Interpreting Defect Models." Proceedings - 2018 IEEE International Conference on Software Maintenance and Evolution, ICSME 2018, 92–103.

[24] Brownlee, J. "How to Choose a Feature Selection Method For Machine Learning". *Machine Learning Mastery, Nov 27, 2019*, accessed on May 5, 2020 at https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/.

When comparing the top 10 features across the three feature selection methods, 6 attributes were identified in all three (BusinessParking Street, NoiseLevel, Ambience Trendy, Ambience Classy, Ambience Casual, Has TV). And 3 attributes appeared in 2 of the 3 models (Ambience Hipster, Bike Parking, Alcohol Full Bar).

The association rules analysis focused on the attributes that appeared in 2 or 3 of the feature selection methods. Attributes that were rated highly by only 1 model were not included.

Steps 8 & 9 : Build and test a statistical model - Apriori association rules analysis

Association analysis is a method of finding sets of items that often occur together in a sample[25]. My goal was to find patterns between restaurant features and high star ratings (i.e., 4 or more stars). The output of association analysis are rules such as {Item A} ⟶ {Item B}. Meaning when {Item A} is present there is a strong likelihood that {Item B} is also present. {Item A} is the left-hand-side (LHS) or antecedent and {Item B} is the right-hand-side (RHS) or consequent[26]. In the current analysis, the antecedents were the combinations of restaurant features that commonly co-occur with the consequent of having a "4 or more" star rating.

The first step of apriori analysis is to inspect the item frequency. Figure 9 shows the percent of restaurants with particular features. Not having a hipster, trendy or classy ambience were the three most frequent features. Having a TV, bike parking and an average noise level (Noise Level = 2) were present in more than 50% of restaurants. Examples of less frequent features were quiet (1) or loud (3) noise level, street parking and a trendy ambience. Thirty-eight percent of restaurants had 4 or more stars. This is noteworthy, because this feature is the consequent. All the rule sets will be drawn from this subset of records.

More than half of the restaurants have between 1 and 3 features and about 40 percent have between 4 and 6 features (Figure 10). The distribution of features per restaurant was the same for businesses with and without high star ratings. This means that rules with fewer than 6 features are most likely.

[25] Kotu, V. & Deshpande, B. "Data Science: Concepts and Practice - Chapter 6: Association Analysis". *Elsevier Science & Technology,* 2019, pp 199-220.
[26] Kotu, V. & Deshpande, B. "Data Science: Concepts and Practice - Chapter 6: Association Analysis". *Elsevier Science & Technology,* 2019, pp 199-220.

Figure 9. Relative Item Frequency Plot

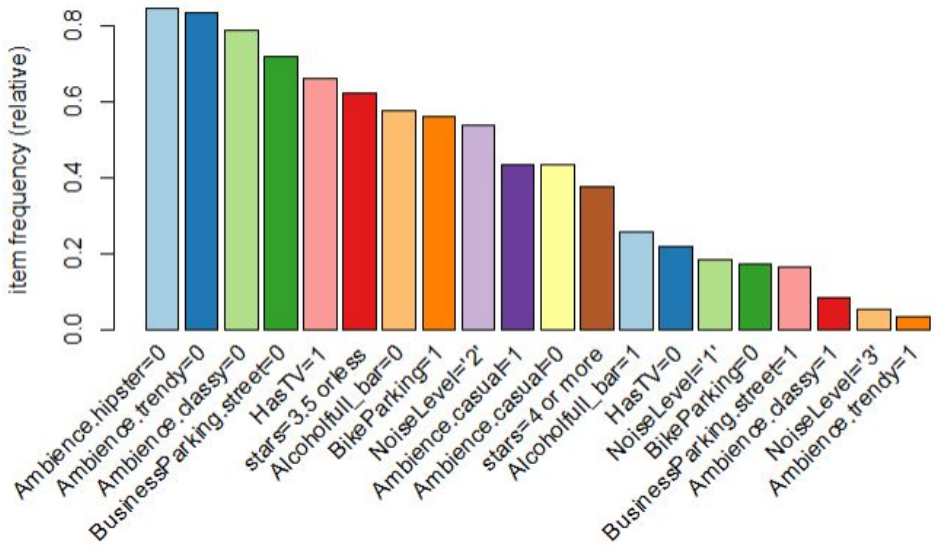
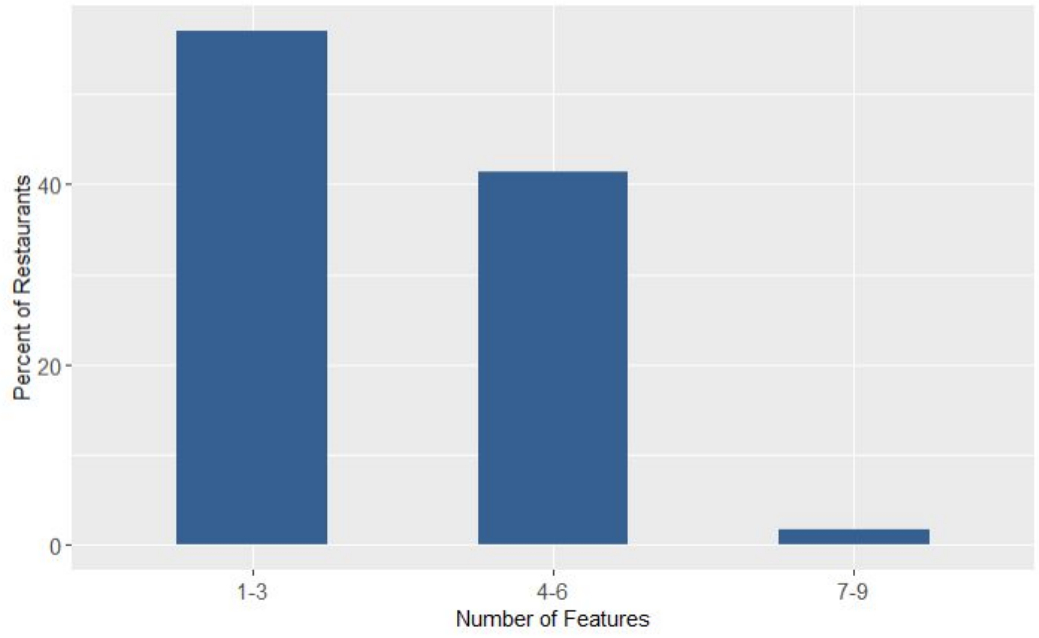


Figure 10. Features per restaurant



The next step in apriori analysis is to mine the rules. Support and confidence are measures of the strength of association rules[27]. In apriori analysis, the support and confidence parameters are set at the

[27] Kotu, V. & Deshpande, B. "Data Science: Concepts and Practice - Chapter 6: Association Analysis". *Elsevier Science & Technology,* 2019, pp 199-220.

outset for computational efficiency[28]. Finding the optimal support and confidence threshold is a challenge in apriori association rule analysis[29]. The consequent for all rules was set to "4 or more" stars, because we are interested in finding feature sets common to those establishments. The number of rules generated under different support and confidence levels is listed in Table 9. Support of 0.05 and confidence of 0.4 were used in the analysis.

Table 9. Number of rules with different support and confidence thresholds

Support	Confidence	Rules
0.35	0.5	0
0.25	0.5	0
0.15	0.5	0
0.05	0.5	40
0.05	0.55	5
0.05	0.6	0
0.05	0.4	295

The minimum support threshold is small but it represents 1,988 restaurants in the dataset, which is large enough to detect patterns. The minimum confidence threshold of 0.4 is less ideal. It indicates that having a rating of 4 or more stars occurs less than half the time that the LHS feature set (the antecedent) appears in the dataset. This may mean that the rules are not reliable.

Making a recommendation to *not* do something isn't a very actionable insight. In order to generate advice for customers or proprietors about common features in highly rated restaurants, rules with at least one feature present in the antecedent were selected, leaving 196 rules.

Apriori analysis is based on the property that every subset of a frequent itemset is also frequent[30]. The subset of rules were as follows:

1) {Business Parking Street = 1} ⟶ {4 or more stars} (support = 0.09; confidence = 0.52)

2) {Ambience casual = 1} ⟶ {4 or more stars} (support = 0.21; confidence = 0.44)

[28] Hahsler, M. et al. "arules: Mining Association Rules and Frequent Itemsets". Accessed on June 7, 2020 at: https://cran.r-project.org/web/packages/arules/index.html

[29] Plasse, M. et al. "Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set." *Computational Statistics & Data Analysis,* 52(1), 2007, pp. 596-613.

[30] Plasse, M. et al. "Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set." *Computational Statistics & Data Analysis,* 52(1), 2007, pp. 596-613.

3) {Bike Parking = 1} ⟶ {4 or more stars} (support = 0.25; confidence = 0.41)

4) {Noise Level = 1, Has TV = 1} ⟶ {4 or more stars} (support = 0.05; confidence = 0.41)

The features that were most commonly present in the feature sets were having street parking, having a casual ambience, having bike parking and having a quiet noise level and a TV. This indicates that restaurant owners who offer these features have a 40 - 50% chance of receiving a rating of 4 or more stars. See the limitations section below for descriptions of issues and proposed enhancements to this analysis.

Step 10: Refine questions focused on user reviews and sentiment

The final stage in the analysis focused on user reviews of the restaurants. User reviews are one of the richest sources of information in the Yelp dataset because they provide the unrestricted opinions of customers. What matters most to customers with kids that are satisfied or dissatisfied with their restaurant experiences? Using natural language processing techniques, the analysis below explores this question by finding common positive and negative terms in the reviewsm; assigning each restaurant an aggregate review sentiment and developing a classification model for star rating using both review sentiment and restaurant features.

Step 11: Data preparation

There were more than 4 million reviews for family-friendly restaurants available for analysis in the Yelp dataset. A random sample of 25% of those reviews was selected for analysis. In the 993,038 sampled reviews, 41,253 restaurants were represented. That means 95% of the businesses in the family-friendly restaurant subset were represented in the reviews sample.

The average number of reviews per restaurant (92.63) was higher than the median (33.00) indicating some outliers identified in step 3 were also part of the review sample. There were 4,340 businesses with review counts that were higher than expected in the sample. That is more than 10% of sampled restaurants. Moreover, those restaurants had a total of 554,103 reviews, which was more than half the review sample (55.80%). To prevent the reviews for these businesses from skewing the analysis, they were removed.

Many customers of family-friendly restaurants do not have kids or do not have children with them when they go out. So, the next step was to identify the child- or family-specific reviews within the remaining 438,935 reviews. A custom list of words related to children and families (e.g., kids, daughter, niece, grandson, baby, family, mom) was created. After converting all the text to lowercase, removing punctuation and extra whitespace and tokenizing at the word level, the subset of reviews with at least one word in the children and families list was identified. There were 64,806 reviews, or 14.76% of the cleaned sample, that mentioned children or families. Therefore, the first insight from the analysis of reviews is that the vast majority of reviews for family-friendly restaurants do not relate to children or families.

The other data cleaning steps were:

- Remove words that are numbers

- Remove words that are special characters (e.g., _)
- Remove words that are less than 2 characters long
- Remove stop word using the same dictionary as the LDA analysis above
- Lemmatize words
- Remove the 10 most popular words
- Remove the least popular words (27,869 words that appeared only once)

Steps 12 & 13: Build and test a statistical model using sentiment analysis

Sentiment Analysis

There are many possible techniques for carrying out sentiment analysis. I used the simplest form of document-level sentiment analysis using unsupervised machine learning[31]. Each review was considered a document. Each word in a review was compared to a lexicon of sentiment words and assigned a value. The word values were aggregated to determine the sentiment orientation of each review[32].

The results of a sentiment analysis can vary based on the lexicon used as each one contains different terms and sentiment classifications. I compared the results for three lexicons before selecting one for use in the star rating classification model. The lexicons were three general purpose lexicons available in the R tidytext package:

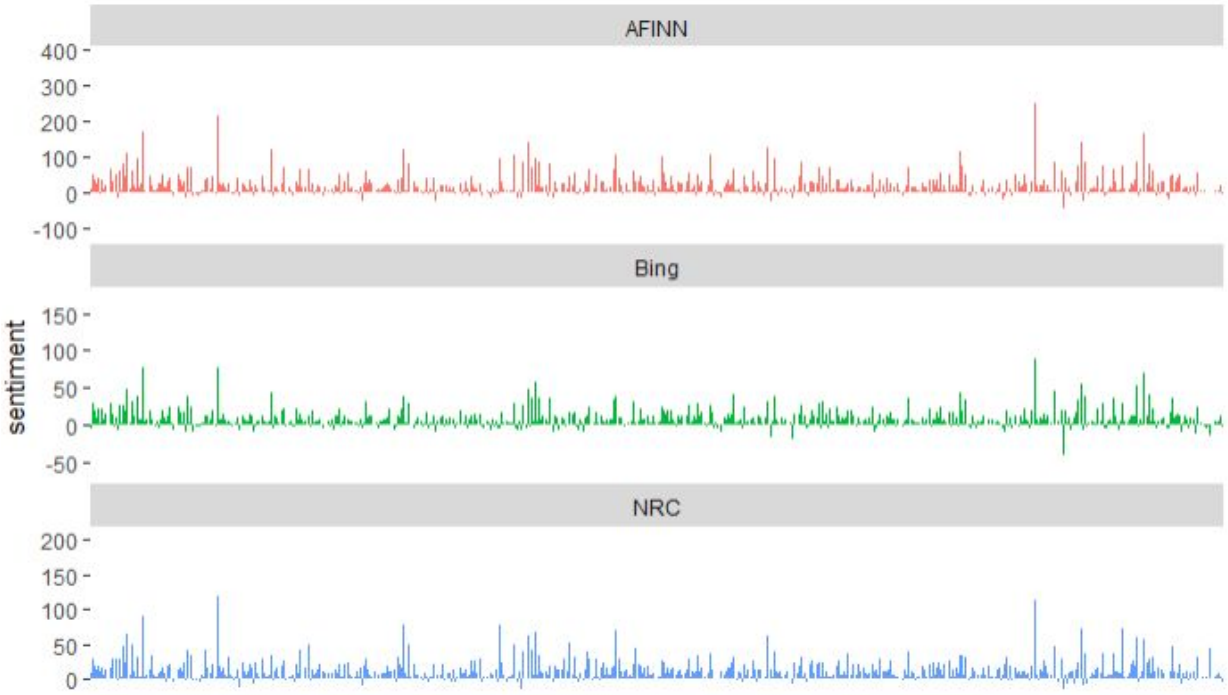
- Afinn from Finn Årup Nielsen;
- Bing from Bing Liu and collaborators; and
- NRC from Saif Mohammad and Peter Turney[33].

The lexicons vary in how they classify sentiment. Afinn uses a likert scale from -5 to +5. Bing classifies words as positive or negative. And NRC categorises words in a binary fashion ("yes" or "no") across 10 categories. Only the positive and negative categories for NRC were used.

The specific method of sentiment aggregation varied across the lexicons. For Afinn, the likert scale scores were summed for each review for a restaurant. For Bing and NRC, the total number of positive and negative sentiments for each review was counted and then the number of negative sentiments was subtracted from the number of positive sentiments. For all three lexicons, the restaurant was tagged as negative if the aggregated value was a negative number and the restaurant was tagged as positive if it was a positive number[34].

[31] Feldman, R. "Techniques and applications for sentiment analysis". *Communications of the ACM, 56(4),* April 2013, pp 82 - 89.
[32] Feldman, R. "Techniques and applications for sentiment analysis". *Communications of the ACM, 56(4),* April 2013, pp 82 - 89.
[33] Silge, J. & Robinson, D. "Text Mining with R: A Tidy Approach." O'Reilly, March 2020, accessed on May 18, 2020 at https://www.tidytextmining.com/.
[34] Removing the reviews near the threshold to maximize sentiment polarity was tested. It was not used in the final analysis as it decreased the performance of the classification models.

Figure 11. Sentiment polarity by restaurant and by lexicon



As shown in Figure 11, the Afinn sentiment scores are objectively higher than the Bing and NRC. That's not unexpected as Afinn scores were calculated by adding together the word sentiment values for each business. While Bing and NRC were aggregated by subtracting the total number of positive words from the total number of negative words per business.The sentiment orientation for most restaurants across all lexicons was positive.

The positive trend is seen at the word level as well. While all three lexicons contain more negative than positive terms, the majority of terms most frequently found in the reviews were positive (Table 10). One limitation of the word-level analysis highlighted by the NRC lexicon is that terms may not be correctly classified by a general purpose lexicon. 'Bite' and 'serve' are classed as negative by NRC, but in the context of restaurant reviews those terms may not be negative. If many of the common terms were miss-classified, it could skew the restaurant level sentiment orientation.

Table 10. Ten most frequent sentiment words by lexicon

Afinn	NRC	Bing
great	eat	great
love	love	love
want	wait◦	nice
nice	friendly	friendly

friendly	delicious	delicious
bad◦	bad◦	bad◦
fresh	bite◦	fresh
pretty	customer	work
leave◦	serve◦	pretty
enjoy	pretty	right

◦ indicates negative term

The question is whether the three lexicons appear to be consistent in their classification of restaurants? The overall number of reviews was similar. NRC's lexicon classified the fewest restaurants as negative. Afinn had more negative restaurants and Bing had the most. The sentiment scores across the lexicons were compared using a Kruskal-Wallis test (H = 2981.9, df =2, p-value <2.2e-16) and there was a significant difference between them in the overall sentiment. Post-hoc Nemenyi tests indicate that each of the lexicons differs significantly from the others at p-value < 2.2e-16.

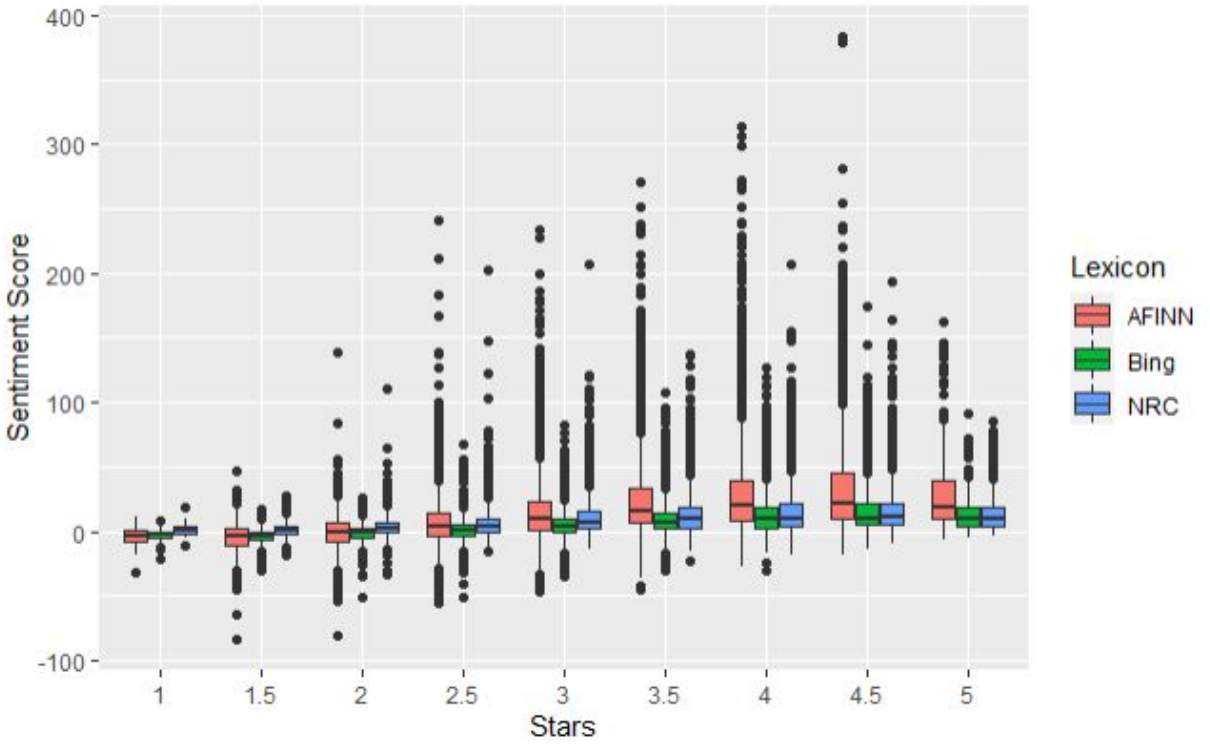
Table 11. Count of restaurants by aggregate sentiment orientation across three lexicons

	Positive	Negative	Total
Afinn	18,659	3600	22,259
Bing	17,910	4409	22.319
NRC	19,945	2429	22,374

There was also a relationship between a restaurant's star rating and their review sentiment. Figure 12. shows a linear upward trend for all three lexicons. As the number of stars increases the sentiment score increases. Spearman correlations indicate that the strongest relationship was for the Bing lexicon (0.41), followed by Afinn (0.37) and NRC (0.25).

The Bing lexicon was used for the remaining analysis. While it's correlation with stars is only moderate, it was the strongest of the three lexicons and it had a smaller imbalance between positive and negative sentiment orientation. A Kruskal-Wallis test comparing Bing sentiment score by stars confirmed there was a significant difference (H = 71413, df =8, p-value < 2.2e-16). Post-hoc Nemenyi tests indicated that the sentiment ratings differed significantly between all ratings levels, except 1 and 1.5 and 4.5 and 5.

Figure 12. Sentiment score by star rating and lexicon



The statistical difference in sentiment orientation held when the star ratings were grouped as '3.5 or less' and '4 or more'. Figure 13. shows the 50 most popular words for the two star categories. Negative words are blue and positive words are orange and the size of the words is proportional to the number of times they appear in the reviews. The most frequent words are similar across all the reviews. But the relative frequency of negative words is higher in the lower star reviews (e.g., words like 'wrong', 'rude' and 'slow') and the relative frequency of positive words is higher in the higher star ratings (e.g., 'amaze', 'recommend', 'fresh').

Figure 13. Word Clouds for the Top 50 Words Grouped by Sentiment Orientation and Number of Stars

Overall, the sentiment analysis of reviews aligned with star ratings, which is another measure of restaurant experience in the Yelp dataset. Restaurants with higher star ratings also had reviews with more positive sentiment. In fact, the restaurant reviews were primarily positive. Both of these patterns were consistent across lexicons, which suggests that the findings are reliable.

The most frequent positive and negative terms were the same for the high and low star restaurants, suggesting that the satisfied and dissatisfied customers care about the same things. They just report those terms at different rates

Classification models

The final step in the analysis was to create a classification model to predict star ratings for restaurants. The models incorporated the business features identified in the association rules analysis as well as the Bing sentiment ratings from the sentiment analysis. Models that were appropriate for categorical data were selected and the results compared before selecting a preferred model.

Since two machine learning algorithms were being compared, a consistent training and test set to be used with both procedures were created. A 75/25 split was used to randomly split the complete dataset. As shown in table 12, the distribution of the dependent class variable (stars) was consistent across the full dataset, training subset and test subset.

Table 12. Number of records by the dependent class (stars) used to train and test models

	3.5 or less stars	4.0 or more	Total Number	Percent 4 or more
Training	10,595	6,145	16,740	37%
Test	3,531	2,048	5,579	37%
Total	14,126	8,193	22,319	37%

The independent variables included in the two models were:

- Sentiment (numeric value between -51 and 175)
- Bike Parking (binary - present or absent)
- Noise Level (factor with 4 levels - quiet, average, loud, very loud)
- TV (binary - present or absent)
- Ambience Casual (binary - present or absent)
- Ambience Classy (binary - present or absent)
- Ambience Hipster (binary - present or absent)
- Ambience Trendy (binary - present or absent)
- Street Parking (binary - present or absent)
- Full Bar (binary - present or absent)

All records were complete for the star value (i.e., dependent variable) and sentiment. There were 6,651 records with one or more values missing for the other attributes. The missing values were imputed using K-nearest neighbors. The data set with imputed values was used for modelling.

The first type of model was binomial logistic regression. This method is appropriate when the dependent variable is categorical and binary, as is the case with the star rating variable (e.g., “3 or less” or “4 or more”). No transformations were required as the key assumptions of logistic regressions were met (e.g., linearity between logit of the dependent variable and independent variable; no multicollinearity between independent variables). One benefit of a logistic regression model is that the coefficients describe the relevance of each variable on the star rating.

The second type of model was a C5.0 decision tree. This method is useful when there are a mix of continuous and categorical variables in the dataset. Decision trees are beneficial because they require minimal data preparation as there are no assumptions about normality or scale of variables. They are transparent, as logical branches can be traced and visualized. And, they can be used when there is a mix of continuous and categorical variables.

There are several decision tree algorithms available. The C5.0 implementation, which uses information entropy computation to split the data, was selected for its efficiency and because there is an easy-to-use

package available for r[35]. One key limitation of decision trees is the risk of over-fitting. This was primarily mitigated by using cross-validation and pruning, see the methods below for additional details.

Binomial Logistic Regression

The logistic regression model was created on the training set using 10-fold cross validation. The coefficients of the model show the change in the log odds of the star rating (Table 13). This can also be expressed as an odds ratio (Table 13). The coefficients indicate that if restaurants offer bike or street parking the odds of having a 4 star rating increase 1.16 and 1.84 times, respectively. It also shows that as the noise level increases, the odds of having a 4 star rating decrease slightly (between 0.21 and 0.66) with each level. Another notable result is that restaurants that offer a hipster ambience, the odds of having a 4 star rating increase by 2.67.

One surprising finding is that there was not a statistical relationship between offering a casual ambience and star rating. At first, this is surprising, because one would assume that a casual environment is a key feature of a family-friendly restaurant. But that may be that feature's undoing in the model. Many restaurants, regardless of star rating, offer a casual ambience so that feature may no longer be an indicator of quality.

Table 13. Coefficients, Odds Ratio, Wald's Statistic and p-Value of the logistic regression model

	Estimate	Odds Ratio	z Value	p Value
Sentiment	0.05	1.05	33.551	< 2e-16***
Bike Parking	0.14	1.16	3.374	0.000741
Noise Level 2	-0.41	0.66	-10.049	< 2e-16***
Noise Level 3	-1.23	0.29	-12.619	< 2e-16***
Noise Level 4	-1.55	0.21	-8.70	< 2e-16***
TV	-0.20	0.82	-4.807	1.53e-6***
Ambience Casual	-0.7	0.93	-1.837	0.066161.
Ambience Classy	0.22	1.25	3.255	0.001136**
Ambience Hipster	0.98	2.67	6.462	1.04e-10
Ambience Trendy	0.34	1.41	3.41	0.000647
Street Parking	0.61	1.84	13.60	< 2e-16***

[35] Pandya, R. & Pandya, J. "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning ." *International Journal of Computer Applications, Volume 117 (*16), May 2015, pp. 18-21.

Full Bar	-0.64	0.53	-15.09	< 2e-16***

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

While the coefficients seem to offer interesting insight into the utility of business features. When the model was used to predict the star class for the test dataset, it did not have a very high accuracy rate (68%). Indicating that just over two-thirds of restaurants were correctly classified using this method.

Decision Tree

The decision tree was created using 3 boosting iterations. While the tree was not winnowed (as feature selection was applied earlier in the analysis), the tree was pruned to limit the number of cases in a leaf to 100.

As shown in Figure 14, sentiment was the first decision node. If the sentiment is negative (or nearly negative <=1), the restaurant likely has a low rating. If the sentiment is not negative and the restaurant is noisy, it also likely has a lower rating. A higher star rating is more probable for restaurants with moderate noise and a hipster ambience.

The next major branch of the tree (7) also begins with sentiment. If sentiment is high and it's a quiet restaurant or it has no full bar, then the restaurant likely has a higher star rating. If the sentiment rating is lower, then the restaurant needs to be quiet, trendy or offer street parking to increase the chance of a higher star rating.

The overall accuracy of the decision tree (69%) is only 1 percent higher than the logistic regression model. The negligible difference in diagnostic ability is visible in the two ROC curves shown in Figure 15. A Wilcoxon rank sum test comparing the two models' performance on each fold of the training set was not statistically significant (W = 34, p-value =0.2408).

Figure 14. Decision Tree

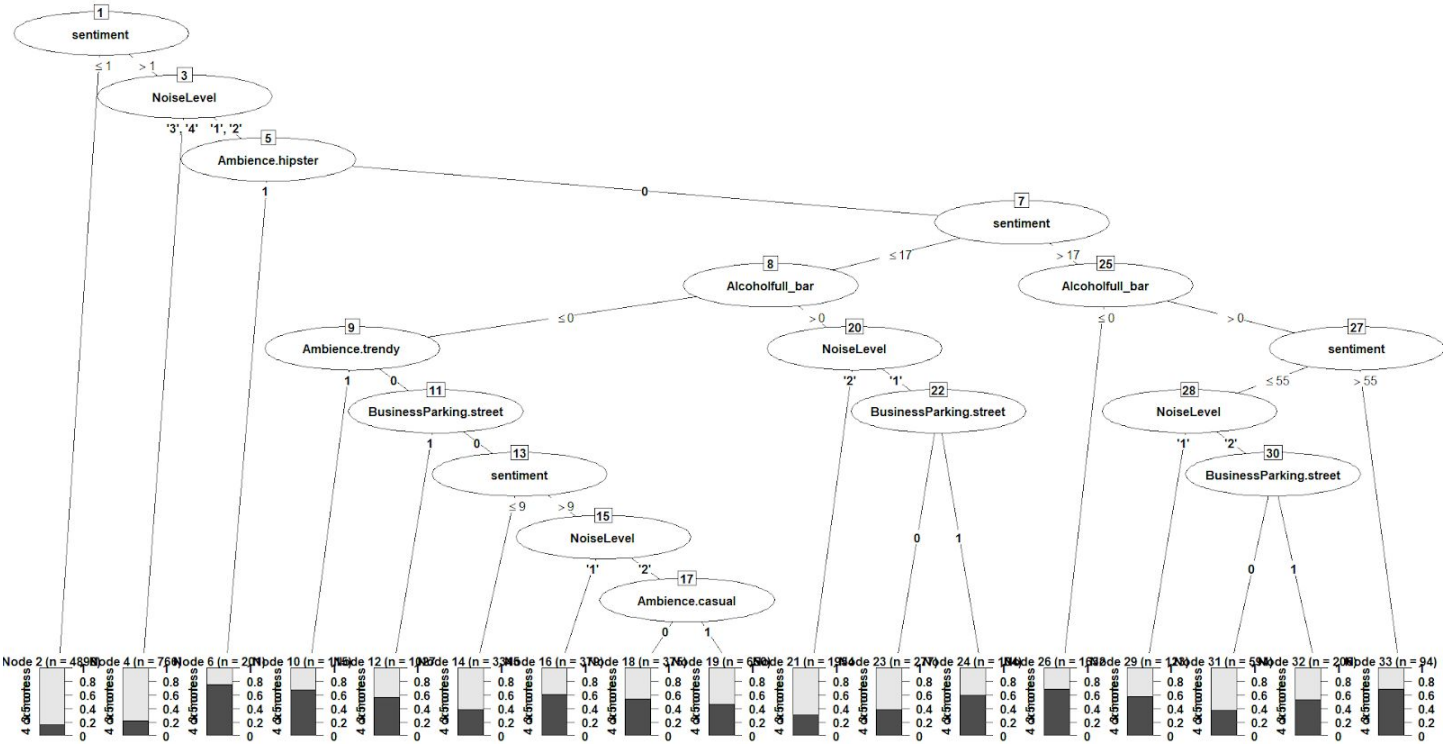
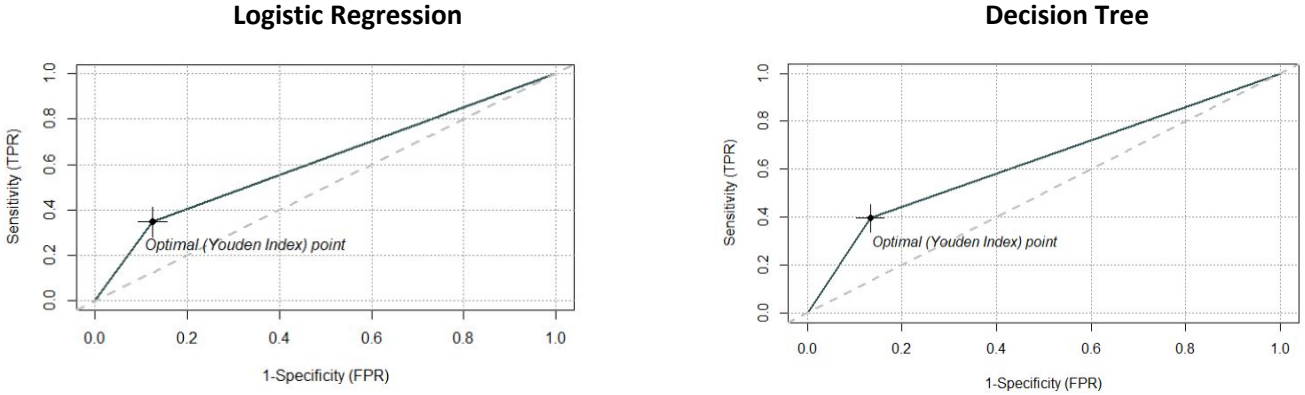


Figure 15. ROC curves of sensitivity and specificity of the two models



In the end, neither model is a very effective predictor of restaurant star rating. The two models do suggest that review sentiment, noise level, having a hipster or trendy ambience and street parking all have an impact on star rating. Perhaps with additional tuning (e.g., undersampling or oversampling, alternate star categories) or the introduction of additional variables (e.g., geography, yelp subcategory, alternative) then one or both models could be improved for real-word use.

Step 14: Interpret and communicate results

Conclusions

The main objective of the proposed project was to conduct an exploratory data analysis of family-friendly businesses using the Yelp academic dataset. The first observation was that the Yelp dataset mainly focuses on restaurants, so the analysis zeroed-in on family-friendly restaurants. Using multiple techniques a range of patterns and relationships emerged that could inform families' restaurant choices or could be used to attract families to a restaurant.

The restaurant subcategories that were the most frequent in the dataset were:

1. Breakfast & brunch,
2. Burgers,
3. Fast food,
4. Mexican,
5. Pizza, and
6. Sandwiches.

By exploring the restaurant titles, other types of cuisine common in the dataset were sushi, Indian, Thai and Chinese were identified. If families are looking to try a new restaurant, the ones that offer these foods or cuisines are likely to be family-friendly.

Using multiple feature selection methods, the following attributes were identified as having a relationship with a restaurant's star rating:

- Bike parking,
- Noise level,
- Has a TV,
- Ambience (casual, classy, hipster, trendy),
- Parking (street, bike), and
- Has a full bar.

The association analysis confirmed that the features common to highly rated restaurants were providing street parking, a casual ambience, bike parking or a quiet restaurant with a TV. Including these features in family-friendly restaurants may be related to customer satisfaction.

Reviews related to kids and families were overwhelmingly positive. And there was a moderate relationship between star ratings and review sentiment. For families, this means they can save time by

relying on the star rating rather than taking time to read the reviews when choosing a restaurant. For family-friendly restaurant owners, they may want to encourage or incentivise customers with kids to write reviews on Yelp. Customers seem to write positive things and provide high star ratings, which improves the restaurant's profile online.

Two classification models were developed for predicting the star rating of family friendly restaurants. Both a logistic regression and decision tree model correctly classified about two-thirds of the restaurants as having a high or low star rating based on each restaurant's sentiment orientation and the features identified through feature selection. While the models performed better than chance, the performance was not accurate enough to inform restaurant choice for families. The decision tree model did provide information about logical rule sets (i.e., moderate positive sentiment (1<=sentiment<=17, no full bar, and trendy ambience) related to higher star ratings that could be explored further. And the logistic regression model suggests features that increase the odds of a high star rating (e.g., hipster ambience or street parking) that may also be worth investigating further.

Limitations

There are several issues with the current analysis that could be addressed to improve the quality or utility of the results.

First, the analysis focuses on only family-friendly restaurants. To see if the findings differentiate family-friendly restaurants from other establishments, similar analyses could be done for restaurants in the Yelp dataset that are not "good for kids".

Next, the topic modeling and sentiment analysis could benefit from additional inspection and cleaning of the text. For example, converting characters to ASCII to remove remaining accents or special characters could improve text normalization. Or, normalizing the text for common misspellings of words (e.g., "yaaaaay!!", "yummmm").

Another improvement could be the use of an alternative topic modeling approach, such as skip-gram with negative sampling in the word2vec package. Word2vec is based on the assumption that related terms occur physically close to each other. It looks at the probability of a word's proximity to target words. This could be a useful approach for the restaurant titles because titles are short and semantically rich[36]. The results of a word2vec analysis could be used to predict words for a restaurant title, which could be a useful tool for business owners.

Next, the association rules analysis had several limitations. The most significant is that the rules generated could be spurious. The support and confidence levels were set very low to generate any rules at all from the data. And no steps were taken to check if rules could be reproduced or to check for false positives[37]. The association rules analysis should be redone by separating the data into training,

[36] Jurafsky, D. & James H., M. " Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition." 3rd ed. Stanford University, UK: Online, 2019, chapter 6, accessed on June 27, 2020 at https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

[37] Liu, G., Zhang, H., & Wong, L. "Controlling False Positives in Association Rule Mining." *Proceedings of the VLDB Endowment, Vol. 5, No. 2*, pp. 145-156.

validation and test subsets so that the reproducibility of the rules can be tested. Given the moderate class imbalance, it would be important to ensure consistency across the sets in the class ratio. Feature sets associated with "3.5 or less" star ratings could also be explored to check for overlap with feature sets identified for higher star ratings.

An alternative method for the association rules analysis may be to combine association rules analysis with clustering[38]. Instead of separately selecting features and then applying rules, one could do a clustering analysis on the sparse restaurant feature data. After clusters of restaurant clusters are identified, an association rules analysis can be done to mine for rules relevant to the each cluster. This type of analysis could incorporate information from the Yelp subcategories and location (e.g. city) along with the other features. Given the diversity of restaurant types (e.g., ambience, cuisine, franchise, independent) this approach may provide more nuanced, accurate and abundant feature sets.

Next, a key assumption of the document-level sentiment analysis is that the document contains a single opinion on a subject[39]. However, this assumption may not be accurate for the Yelp reviews. Users may express opinions on more than one subject or express mixed feelings in a review. This is a limitation of the current approach because mixed sentiments within a review were aggregated and may have canceled each other out. Sentence level sentiment analysis may be an alternative. At the sentence level additional context can be considered (e.g., phrases like "not bad" or "not my favorite" or "worth the wait"). The average or standard deviation of sentence-level, sentiment scores can be used to explore the valence and polarity of reviews or restaurants.

The current sentiment analysis also doesn't explore what about a customer's experience was positive or negative. Aspect-based sentiment analysis would provide more detailed information about the topics identified in each review (i.e., clean, friendly service, fresh food) and the sentiment for each topic[40]. Aspect-based sentiment is useful for reviews, because customers often comment on multiple topics and share different opinions about each.

The final limitation that should be addressed is for the classification models. The star rating classes were selected to try and split the data as evenly as possible using the existing adjacent groups. However, the result was a class imbalance of 63% to 37% for "3.5 or less" and "4.0 or more", respectively. The results of the decision tree, in particular, and possibly the logistic regression could be improved by introducing oversampling of the "4 or more" star class or undersampling the "3.5 or less" star class. Alternatively, the decision tree could be trained to predict raw star classes, rather than a binary grouping.

Addressing one or more of the issues above should be undertaken to provide more useful insights about family-friendly restaurants from the Yelp dataset.

---

[38] Plasse, M. et al. "Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set." Computational Statistics & Data Analysis, 52(1), 2007, pp. 596-613.
[39] Feldman, R. "Techniques and applications for sentiment analysis". *Communications of the ACM, 56(4),* April 2013, pp 82 - 89.
[40] Feldman, R. "Techniques and applications for sentiment analysis". *Communications of the ACM, 56(4),* April 2013, pp 82 - 89.