Yelp User Review Analysis

Delaney H.
Wrote a review

Data Description: This dataset is a subset of Yelp's businesses, reviews, and user data consisting of 5,200,000 user reviews, information on 174,000 businesses and data spanning over 11 metropolitan areas (4 GB).

Data Source:
https://www.yelp.com/dataset/challenge

Kyle S.
Wrote a review

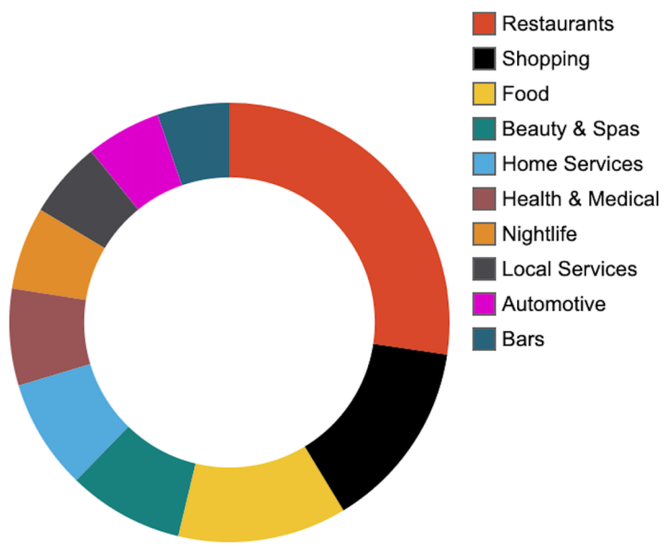
state	total_business	total_reviews	avg_rating
NC	500	8147	3.84
TX	1000	23935	3.72
IN	393	3041	3.71
CA	4000	141119	3.69
ON	228	1014	3.69
MI	500	11634	3.66
RI	500	11086	3.64
WA	500	17998	3.64
MA	1298	54477	3.60
GA	500	15455	3.58
VA	189	1503	3.56
PA	1000	20078	3.55
IL	500	7188	3.51
NY	1382	23675	3.49
NJ	500	7904	3.37
MD	500	5813	3.36

What is the average user rating per state?

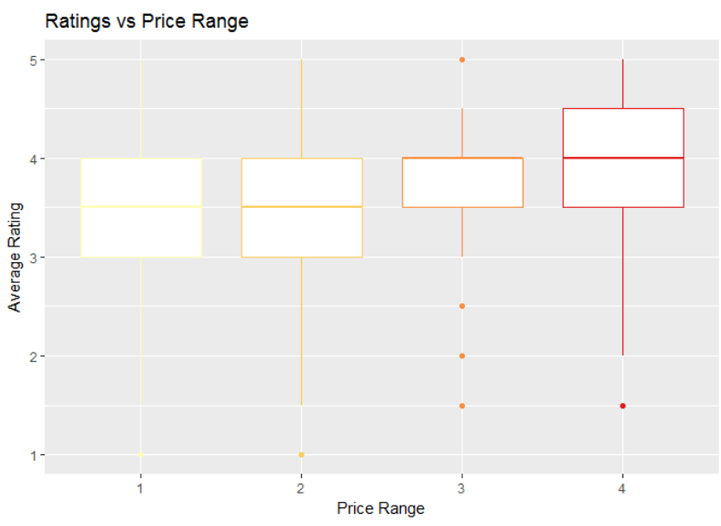
The first exploratory analysis began with looking at average review ratings by each state. Total 348,256 reviews on 13,730 business in 16 states were included in the analysis. After preprocessing the data, it was loaded into a table and created a leaflet map.

Joe H.
Added 1 photo

Top 10 Categories of Business



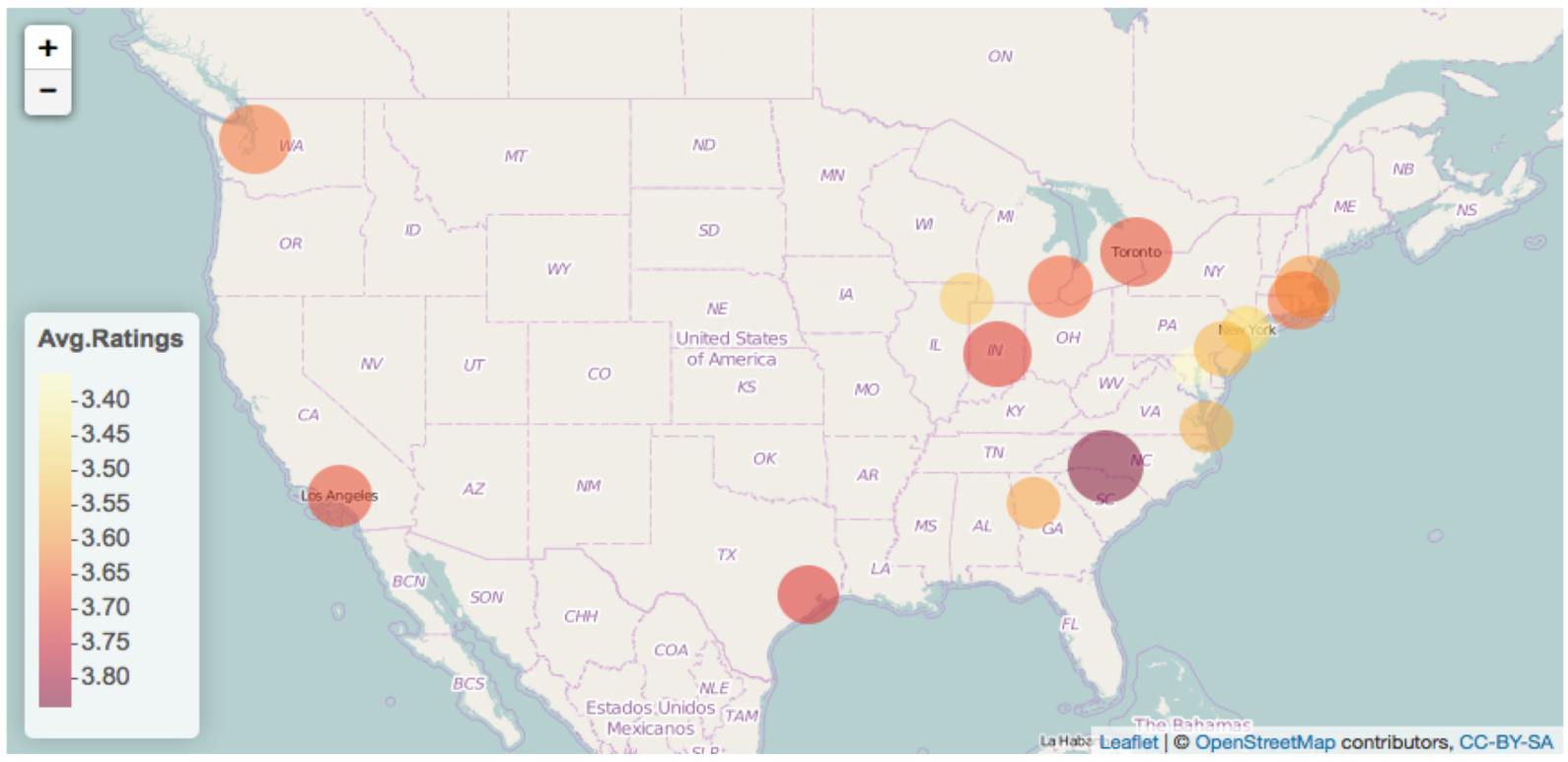
Jess M.
Wrote a review



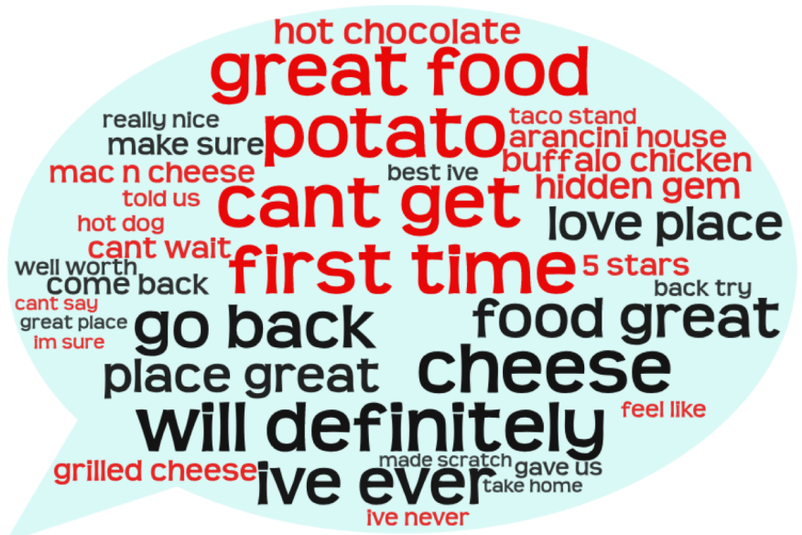
What is the correlation of price on ratings?

It appears that there is some evidence of this hypothesis from the data. Especially, the restaurants in the Price Bracket of 4 (highest) have a higher average Rating.

Andrea B.
Wrote a review

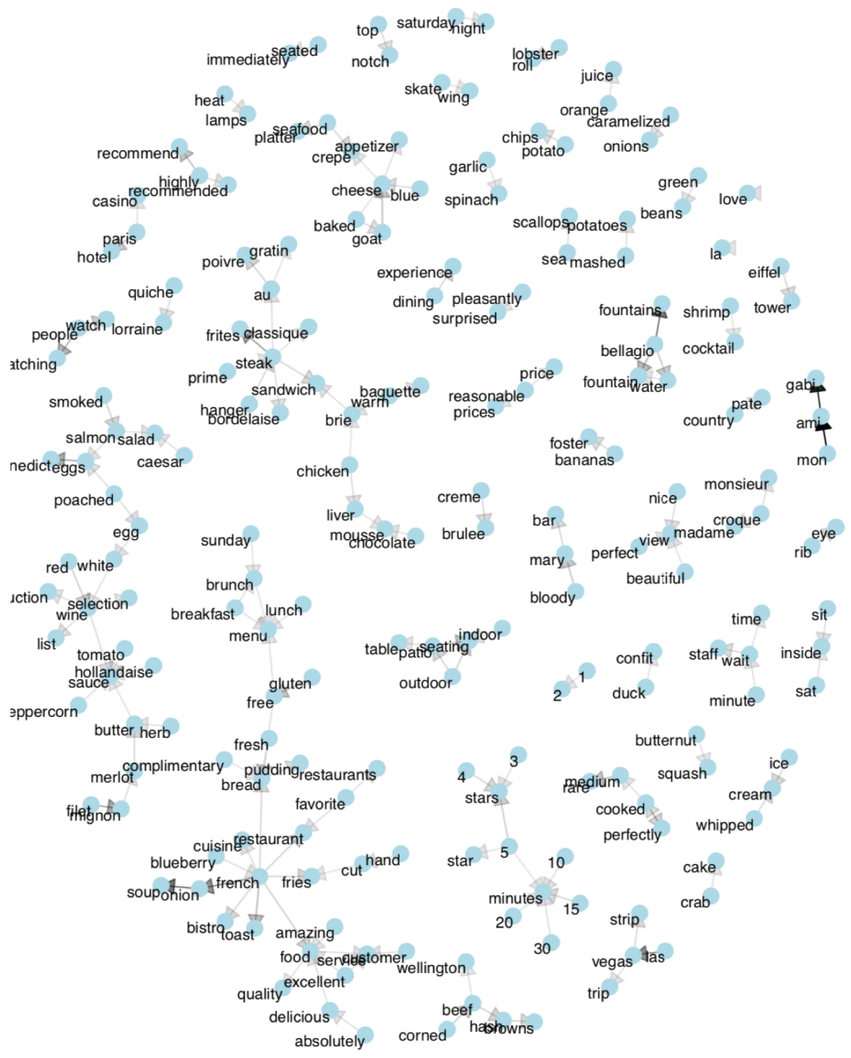


Geoff D.
Wrote a review



The word clouds indicate with great clarity the difference between reviews for restaurants with low ratings versus those with higher ratings. We can see that phrases like 'will never', 'go back', 'one star', 'waste time', '2 stars', 'dont think' etc are used frequently for lower-rated restaurants. Whereas phrases like 'really good', 'great food', 'will definitely', 'great place', 'best', 'can't wait' etc are used with the highest frequency in reviews for highly rated restaurants.

Cynthia N.
Wrote a review



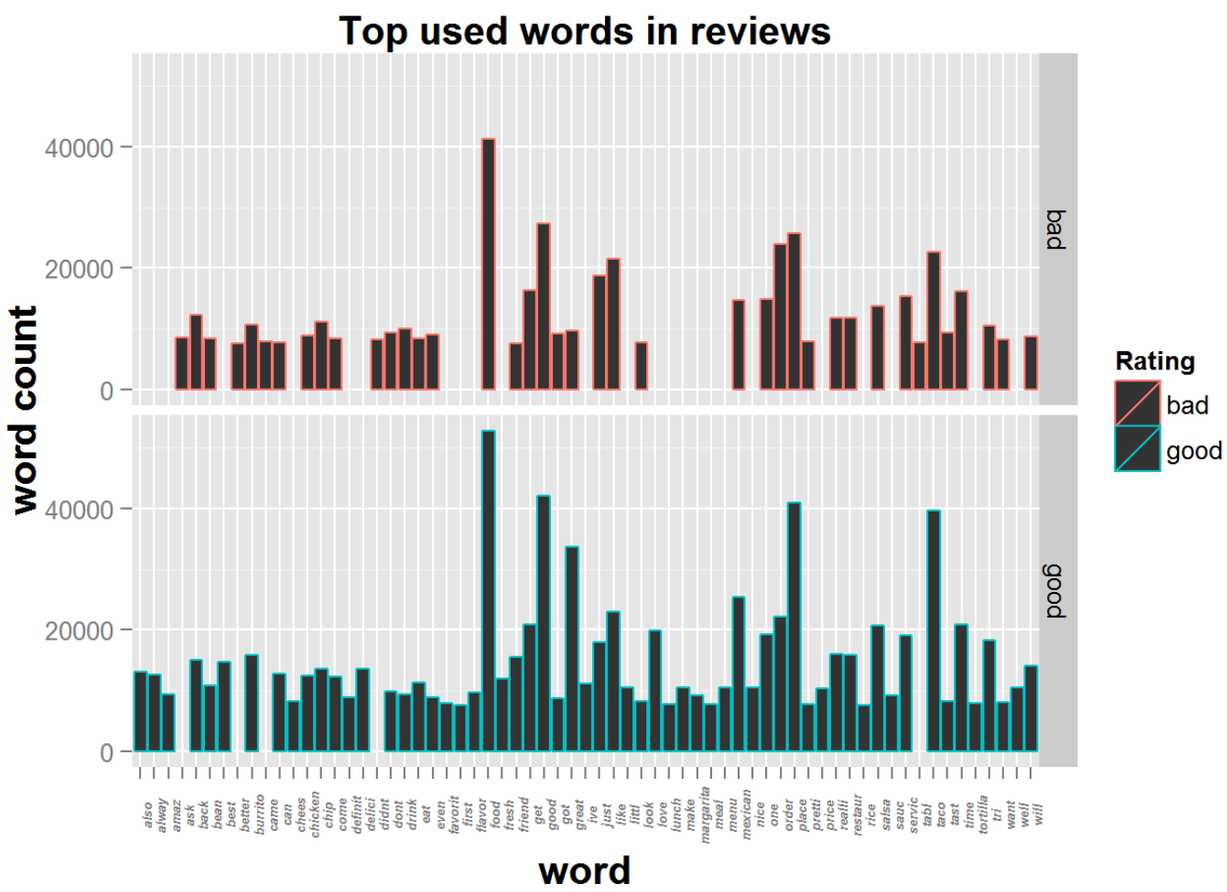
What is the relationship among the words used?

Explored different relationship among various words in top reviewed restaurant through a network graph. Entities are displayed as round nodes and lines show the relationships between them. The vivid display of network nodes can highlight non-trivial data discrepancies that may be otherwise be overlooked. For instance, stars is associated with numerical characters used in the review

Hayley B.
Wrote a review

In this analysis, a model was identified that can identify the rating of the review based on words used in review's texts. It was found the some words like love, favorite, flavor, well, margarita appears in good rating reviews. The word good was used two times more in good reviews compared to bad reviews.

R-Package used: tidyverse, stringr, lubridate, wordcloud, tidytext, DT, leaflet, igraph, ggraph, topicmodels, tm, snowballc, textcat.



Author: Vaibhav Kamal Nigam
From: IST 719
Information Visualization