615 Final

Jingning Yang

12/12/2019

Project Introduction

I use the dataset provided by Yelp as part of their Dataset Challenge 2019. The dataset includes data from 36 states in United States. It contains information about 158525 business and business attributes, 6685900 reviews and so on. Summarize, the dataset consists of five json files: business, review, user, check-in and tip.

Since our dataset too large to run in local, I cleaned the dataset on server, separate into 3 different datasets about restaurants business, saved them and will be read in this rmd file.

Import and clean Yelp JSON Data: Business + Review

I separated business.json into 2 csv file: business.csv and att.csv, selected useful information in review.json into one csv file: review.csv.

And my goal is try to figure out based on stars, is there any particular standard or requirement for a "good" restaurant(stars >=4) in state OH and PA? And it may help business change their stars in short term by paying more attention on those standard/requirement in these two states. Also it may help visitors recognize variety taste of each states so that they can avoid the situation: go to restaurants with high stars, but do not like its taste at all.

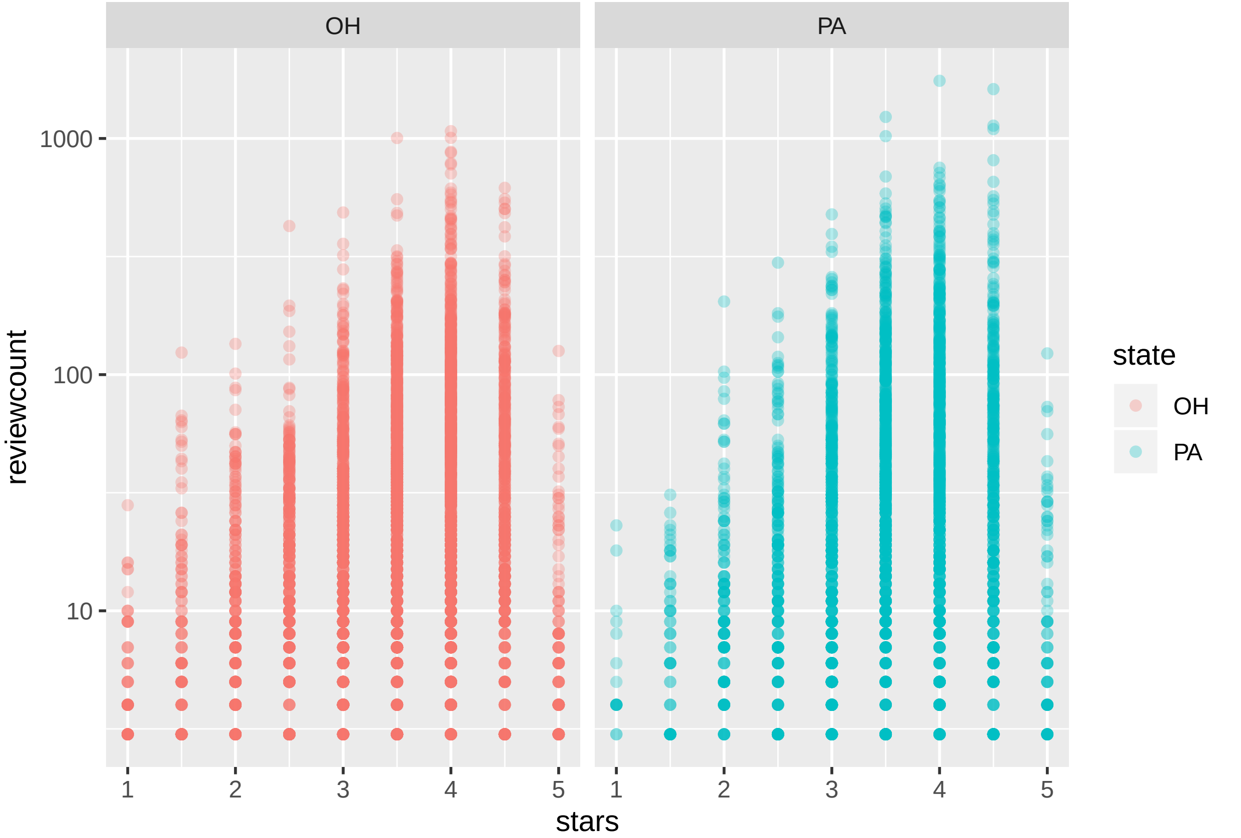
Clean imported Data

For business file, we focus on address, postal code, city, name, stars, states, review counts, longitude and latitude, thus we delete other useless data. For att file, we only focus on whole useful attributes for each restaurant in business file.

If we take a look closer to our data, we can see the number of restaurants of 24 states are not stable, and for compare them in states, I will keep state OH and PA which has more than 20 restaurants record in this data.

Basic EDA

we try to illustrate is there any relationship between review count and stars in different stars?



Through this graph, we can see there is about positive correlation between review count and stars in different states.

Using PCA reduce our dimensions

Because of there are correlation between each variables in our dataset, we try PCA here to reduce the dimension so that we can explore potential relationship between variables.

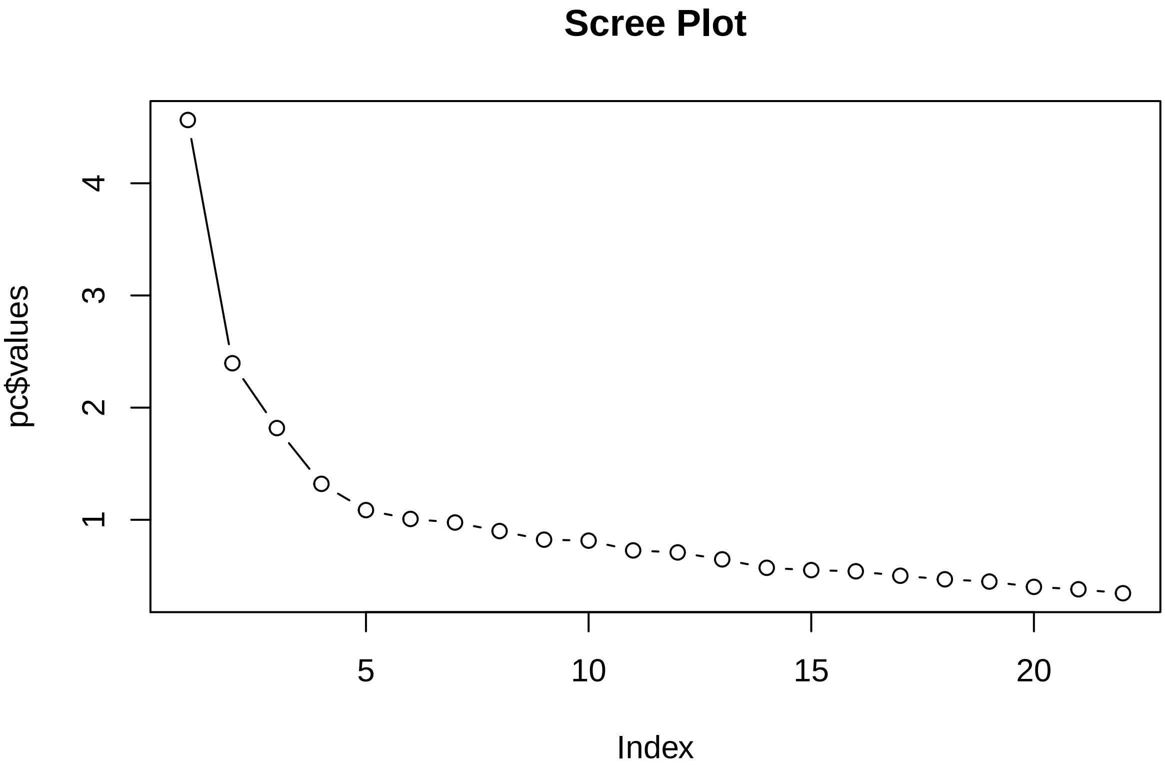
```
## R was not square, finding R from data
```

```
## $chisq
## [1] 35748.42
##
## $p.value
## [1] 0
##
## $df
## [1] 231
```

From output data, Bartlett's test is highly significant, asymptotically chi-square is 35748.42, and the P-value of chi-square smaller than 0.001, therefore factor analysis is appropriate.

Since my goal is reduce number of variables in my data by extracting important one from the data, thus, I will using PCA to do factor extraction.

Factor extraction by using PCA:



From the Scree plot, x axis: component number, y axis:eigenvalues of principal components and factor analysis. Since the elbow part is about the 3rd point from the left, so the evidence from the scree plot suggests 3 component solution may be the best.

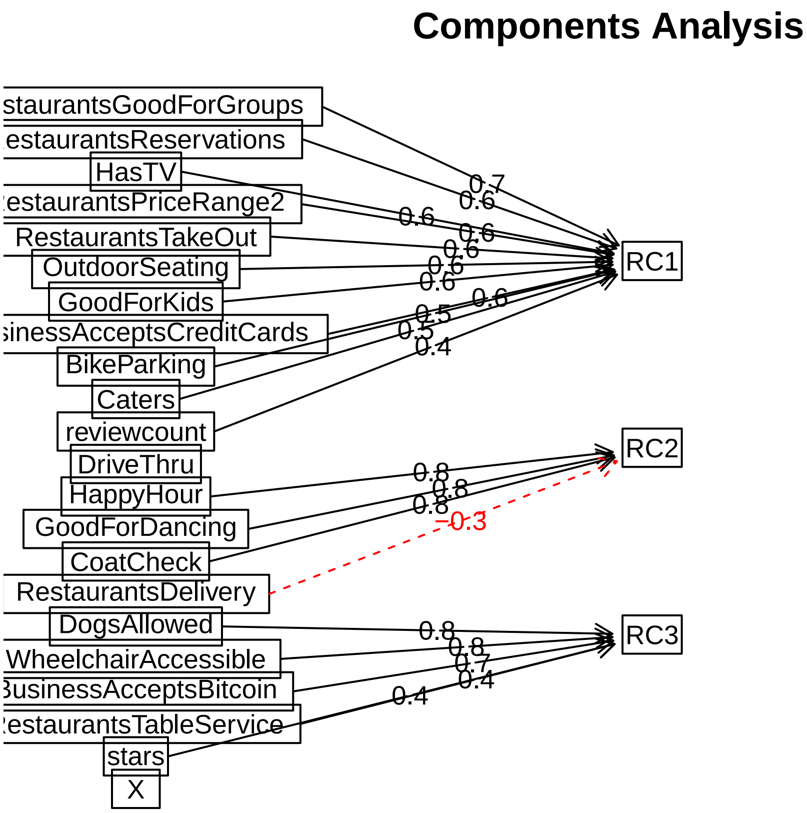
Thus, we choose 3 as our number of factors.

Redo PCA by using 3 factors:

Through output data, Cumulative variable shows these 3 principle components explains 40% data with 22 variables. Specific output table placed in Appendix slides.

Plot result:

For easier to explain the output of factor extraction, we can using orthogonal rotation to decreasing noice for factors as much as possible and get the plot illustrate correlationships between variables.

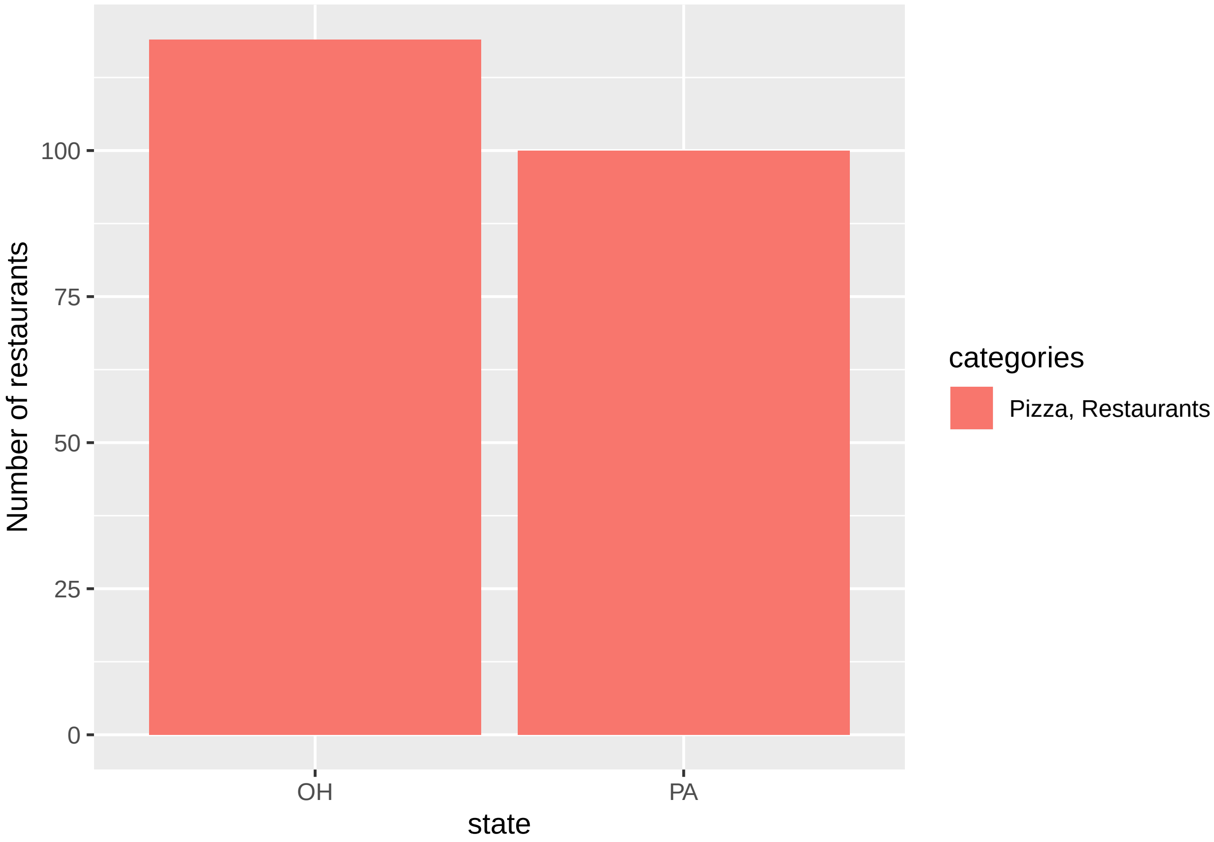


From the graph, 22 variables can separated as 3 groups: RC1, RC2, RC3. RC1 including restaurants good for groups with highest factor loadings, has tv, good for kids and so on. We can consider RC1 as "Suitable for Family/group". RC2 including smoking, happy hour, coat check and so on that we may recognize RC2 as "Suitable for social/networking". RC3 including wheel chair accessible, dogs allowed and so on that can be called "Humanistic care".

Therefore, we can summarize our variables from 3 parts: suitable for family/group, suitable for social/networking and humanistic care.

Explore any preference for a "good" restaurants in different states

Try to solve this question, we need to use review.csv based on cleaned review.json. I merge review.csv and business.csv together and use it to do text mining for exploring any preference. Before we do text mining, We try to figure out the top number of restaurants in every states:

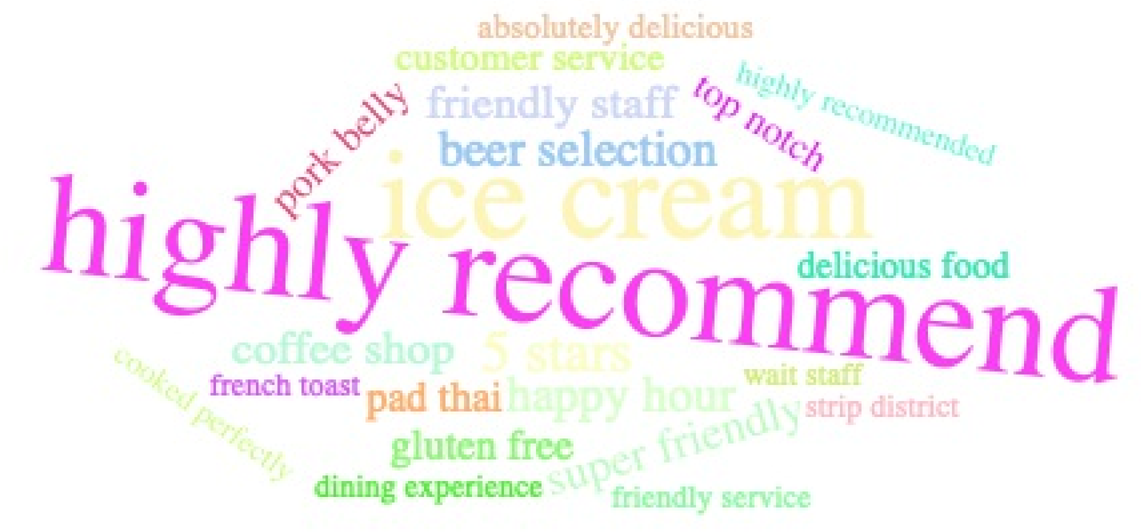


From this plot, we can see there is 1 top number of restaurants: Pizza restaurants.
Therefore, we can make a guess:
Pizza restaurants are popular in both state OH and PA

Next, we can check our guess through text mining through review information in different states.

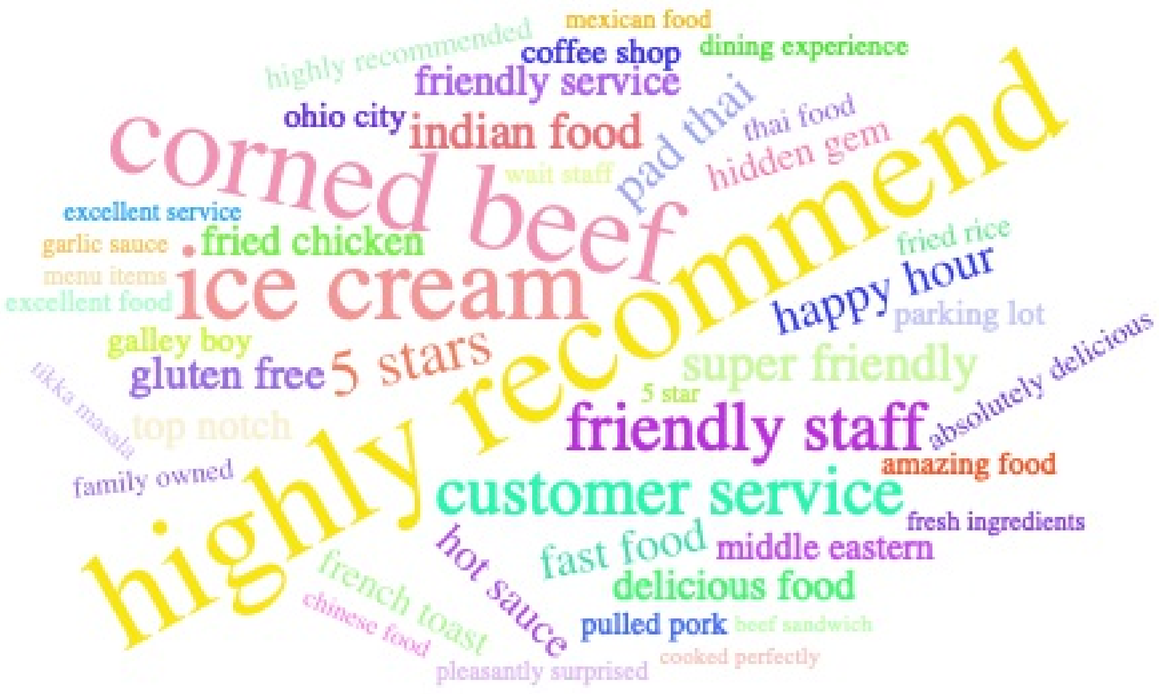
Checking our guess by using Text mining for feature reviews of restaurants with high stars(4~5) in different states.

Review features of good restaurants in PA:



From the plot, the highest frequency of words are: beer selection, friendly staff, ice cream, coffee shop, friendly service,happy hour, pad thai, pork belly and so on. Looks like people in PA love Thai food.

Rreview features of good restaurants in OH:



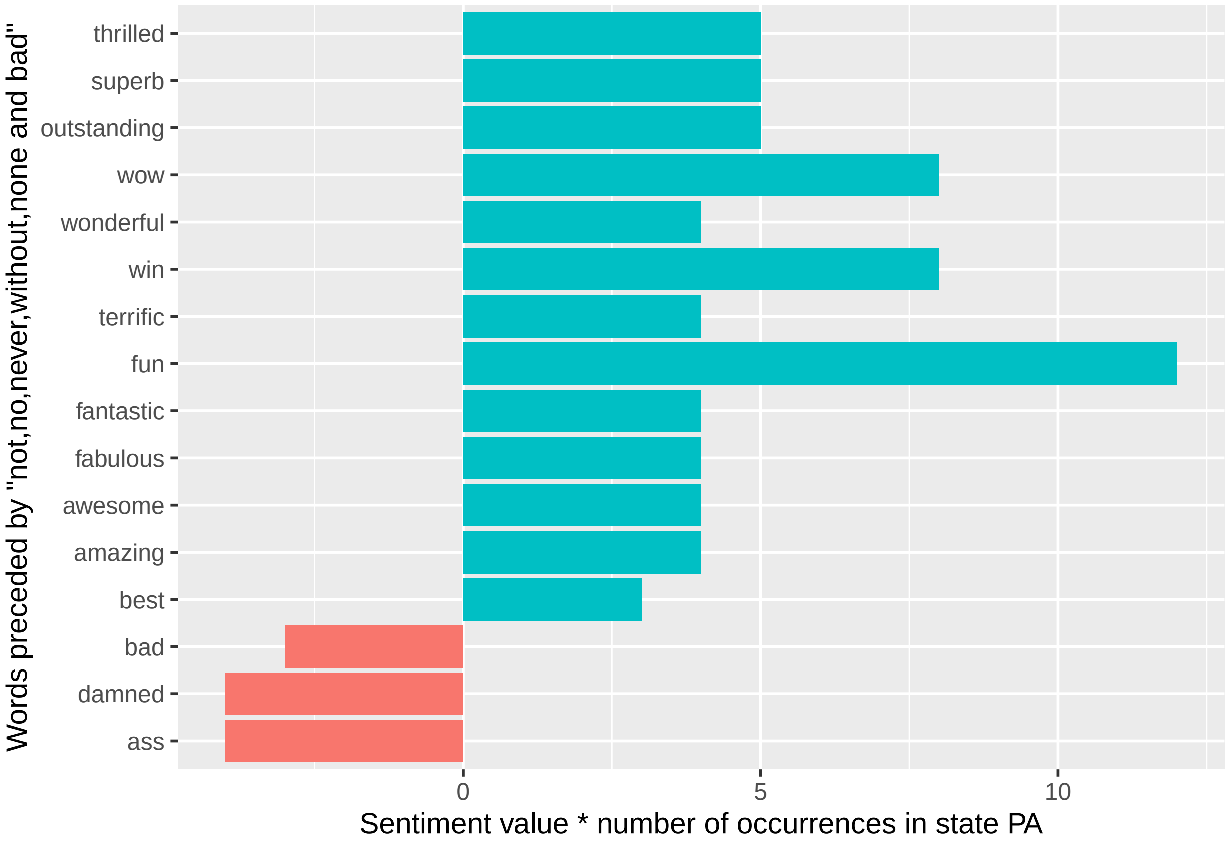
From the plot, the highest frequency of words are: corned beef, customer service, ice cream, pad thai, happy hour, friendly staff, indian food, fast food and so on. Looks like people in OH attention on customer service, corned beef, indian food and so on.

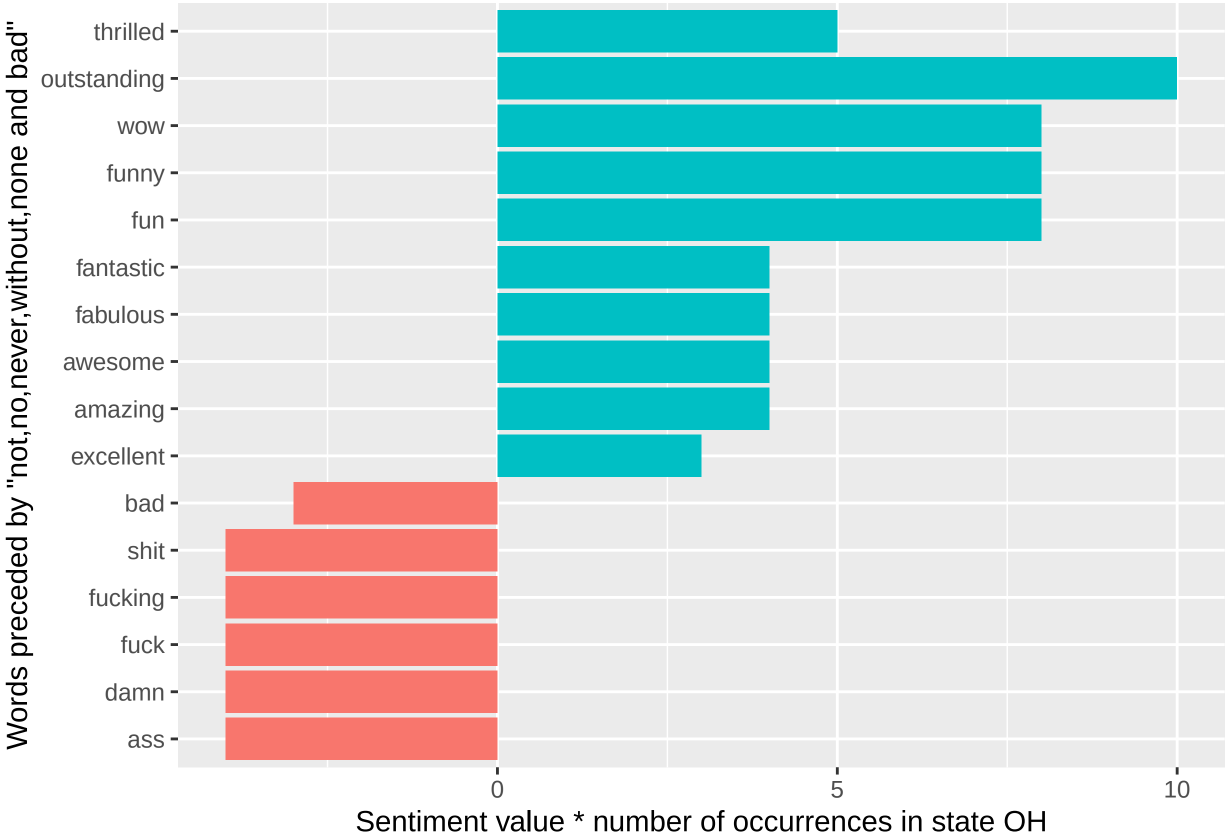
Same logic for rest states. And through previously analysis, we know that the top number of restaurants not represent the taste preference of people. And it shows the importance of the frequency of review word that illustrate what customers thought, feeling and focusing on. From previously text mining for 2 states, we know customers love ice cream, focus on customer service no matter which states they are coming from.

Thus, maybe restaurants business can pay more attention on these general points, and improving more based on specific preference of different states.

Sentiment analysis

Performing sentiment analysis on the bigram review data of 2 states: we examine how often sentiment associated words are preced by some negative words by using AFINN lexicon, and we will get a numeric sentiment value for each word, with positive or negative numbers indicating the direction of the sentiment.


Words preceded by "not,no,never,without,none and bad"
thrilled
superb
outstanding
wow
wonderful
win
terrific
fun
fantastic
fabulous
awesome
amazing
best
bad
damned
ass
0
5
10
Sentiment value * number of occurrences in state PA


Words preceded by "not,no,never,without,none and bad"
thrilled
outstanding
wow
funny
fun
fantastic
fabulous
awesome
amazing
excellent
bad
shit
fucking
fuck
damn
ass
0
5
10
Sentiment value * number of occurrences in state OH

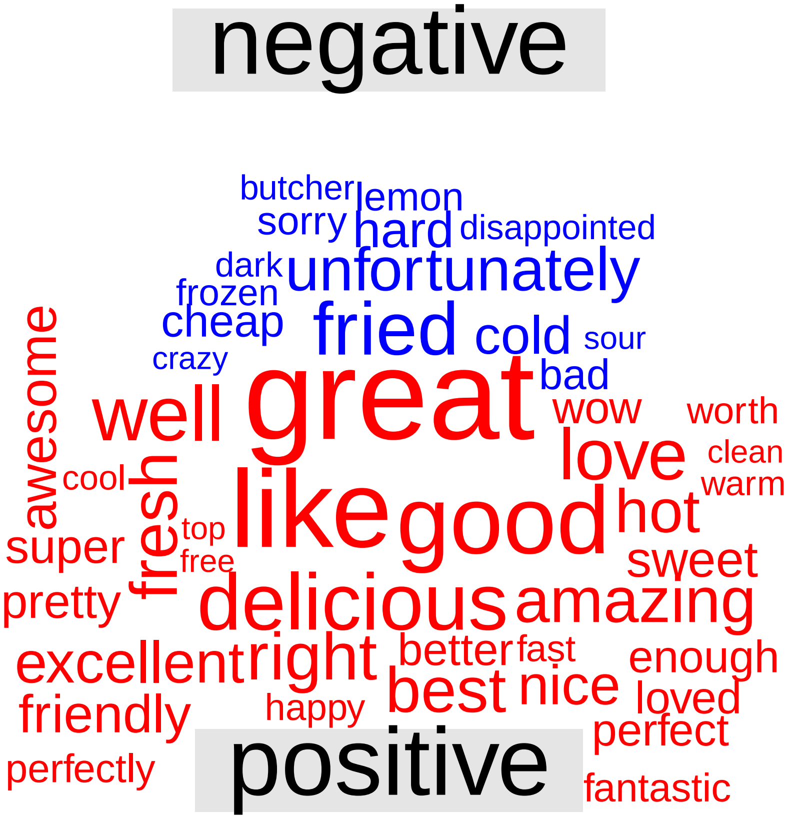
Sentiment analysis summary

In general, review words in both states are more positive than negative, and that might because I only choose rating scores over 4.0.

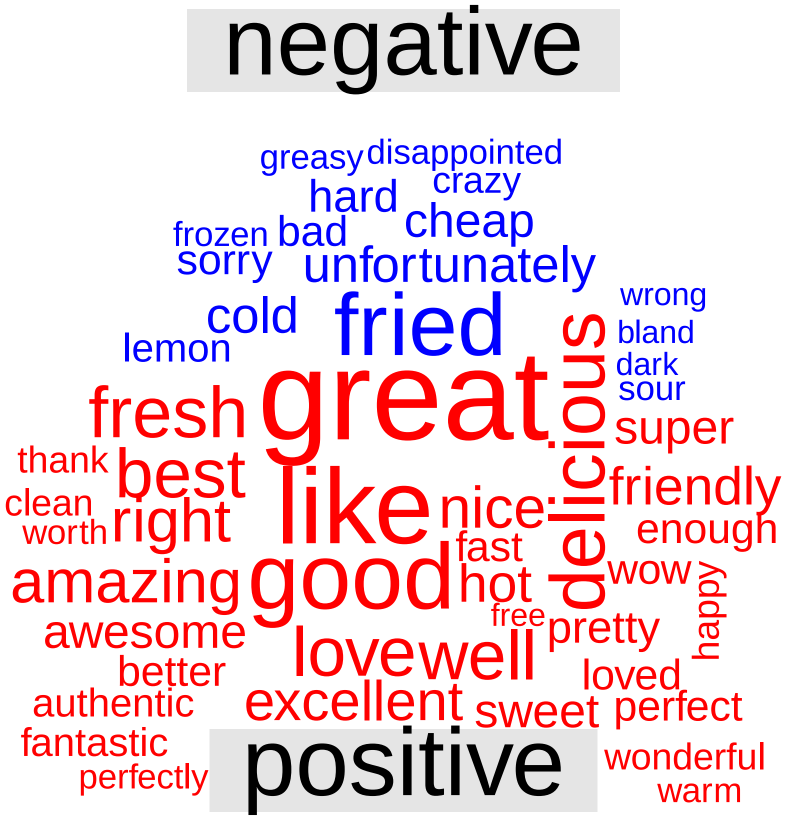
Also, we observe that the weight of positive review words for good restaurants in PA is more than in OH. That might because people in OH have higher demanding on food than people in PA.

Another way sentiment analysis for state PA:

Do the sentiment analysis to tag positive and negative words using an inner join, then find the most common positive and negative words.



Another way sentiment analysis for state OH:



Appendix

Summary information of redo PCA:

```
## Principal Components Analysis
## Call: principal(r = att, nfactors = 3, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                              PC1   PC2   PC3     h2   u2 com
## X                          -0.01  0.03 -0.03 0.0019 1.00 2.5
## GoodForKids                 0.30  0.59 -0.01 0.4324 0.57 1.5
## RestaurantsReservations     0.64  0.18 -0.10 0.4545 0.55 1.2
## Caters                      0.49  0.20  0.31 0.3770 0.62 2.1
## RestaurantsTableService     0.56 -0.25  0.10 0.3879 0.61 1.5
## RestaurantsTakeOut          0.39  0.49  0.05 0.3986 0.60 1.9
## RestaurantsPriceRange2      0.59  0.13 -0.18 0.3986 0.60 1.3
## OutdoorSeating              0.61  0.11 -0.08 0.3897 0.61 1.1
## BikeParking                 0.55  0.17  0.07 0.3393 0.66 1.2
## HasTV                       0.67  0.11 -0.15 0.4828 0.52 1.2
## RestaurantsGoodForGroups    0.57  0.37 -0.24 0.5231 0.48 2.1
## RestaurantsDelivery         0.14  0.33  0.25 0.1862 0.81 2.2
## BusinessAcceptsCreditCards  0.44  0.35 -0.06 0.3155 0.68 2.0
## BusinessAcceptsBitcoin      0.33 -0.22  0.62 0.5470 0.45 1.8
## GoodForDancing              0.47 -0.55 -0.38 0.6585 0.34 2.8
## CoatCheck                   0.45 -0.59 -0.28 0.6241 0.38 2.4
## HappyHour                   0.49 -0.56 -0.37 0.6885 0.31 2.7
## WheelchairAccessible        0.46 -0.35  0.53 0.6064 0.39 2.7
## DogsAllowed                 0.30 -0.31  0.63 0.5879 0.41 1.9
## DriveThru                   0.07  0.05 -0.09 0.0151 0.98 2.5
## stars                       0.23 -0.13  0.27 0.1430 0.86 2.4
## reviewcount                 0.46 -0.04 -0.06 0.2197 0.78 1.1
##
##                        PC1  PC2  PC3
## SS loadings           4.56 2.40 1.82
## Proportion Var        0.21 0.11 0.08
## Cumulative Var        0.21 0.32 0.40
## Proportion Explained  0.52 0.27 0.21
## Cumulative Proportion 0.52 0.79 1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.06
##  with the empirical chi square  12182.14  with prob <  0
##
## Fit based upon off diagonal values = 0.9
```