MA615 Final Project

Yiping Jiang

12/15/2019

Contents

1. Introduction

1.1 About Yelp

As a business directory service forum, Yelp develops hosts and markets the Yelp.com website and the Yelp mobile app, which publishes crowd-sourced reviews about businesses. With the continuous update, this application has brought more and more features, such as users can evaluate their experiences in different levels of stars and leave comments for various restaurants, which makes it easier and more efficient while looking for the right places to have great food. As a result, I am interested in comparing restaurants among different regions, stars, and the relationship between reviews and the quality of restaurants.

As our dataset contains 5 distinct files, we first separated them into two and analyzed them individually. Then we combined and compared our findings and finished this report together.

1.2 Dataset Description

The dataset is from Yelp Open Dataset which contains information of businesses, reviews and users:

- business.json: a dataset including names, locations, stars, number of reviews, attributes, and open hours of businesses.
- review.json: a text dataset including ratings, time, content, and votes of reviews for businesses.
- user.json: a user dataset including user id, name, review count, friends, and average rating of each user.

This dataset also has some other files, from which information not used in this report, such as check-in time, tips, and photos of each business.

2. Data Preparation

2.1 Read-in data

From the above dataset, I found four features that could plot in different patterns, which are star, state, categories, and review_count, from which the following analyses will base on.

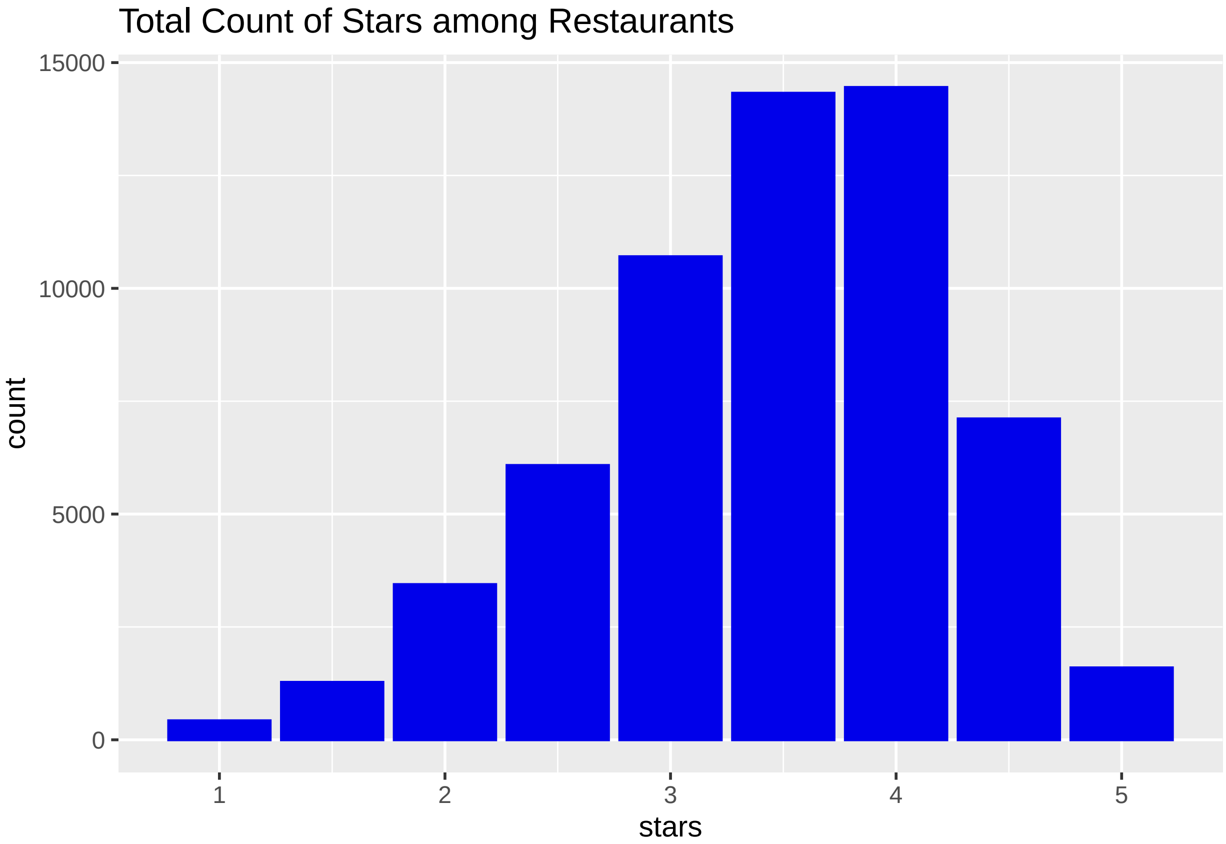
2.2 Data cleaning

In the above procedure, I eliminated some features (attributes and hours) to make the dataset cleaner, and I also unnested the variables to store as yp_un, which helps to analyze them further when one restaurant is attached to a different variable.

2.3 Variables checking

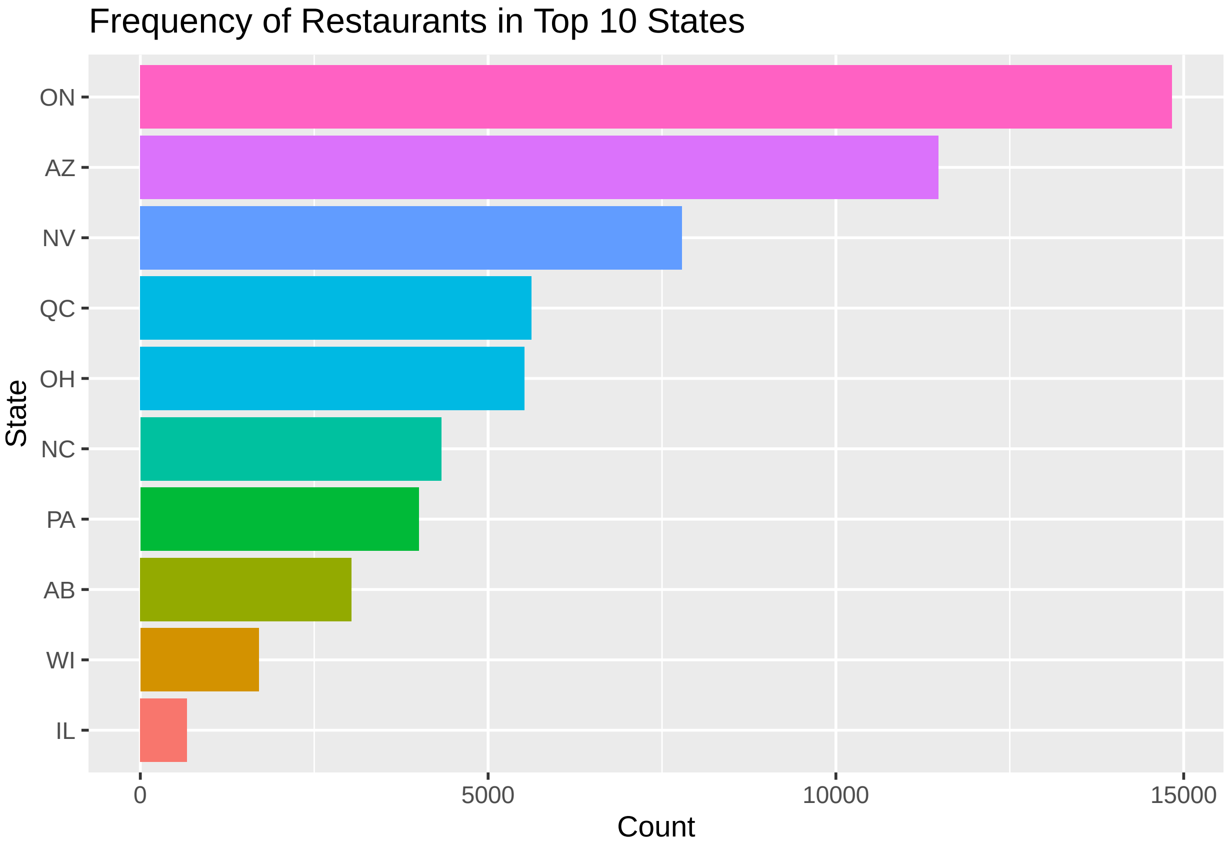
For starters, let us take a look at the four critical variables (star, state, category, and review_count) in the dataset.

2.3.1 Stars

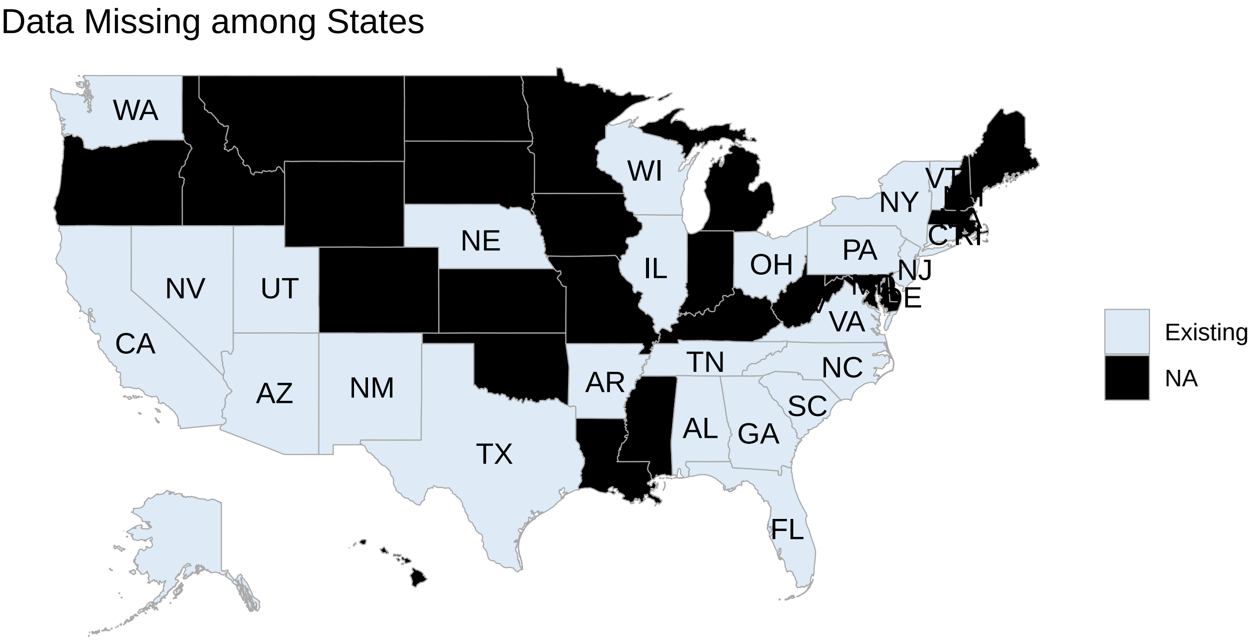


The figure above shows the distribution of stars, and we can tell that more restaurants tend to have a rating of stars between 3 to 4.

2.3.2 State

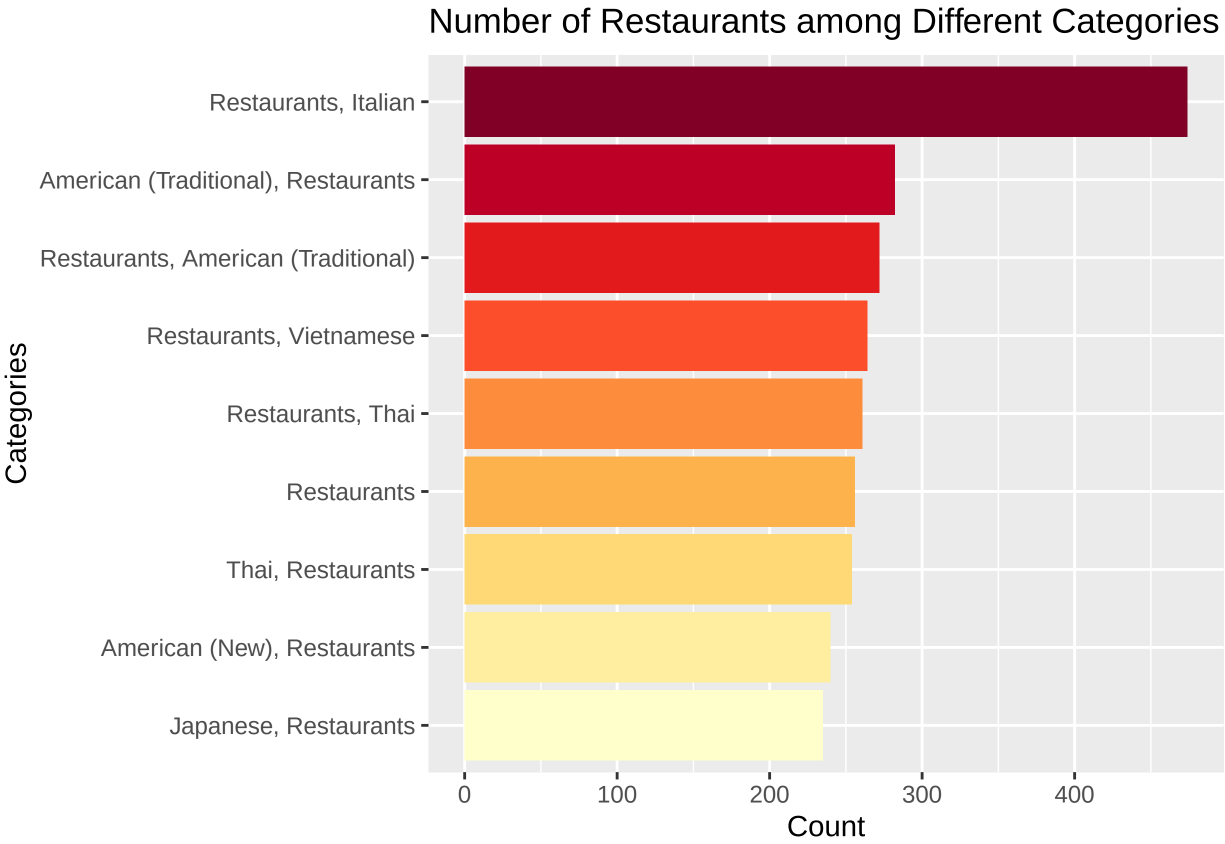


From the above plot, I eliminated states with missing values about businesses and showed only ten representative states in increasing order; and I did the following step to check their completeness.



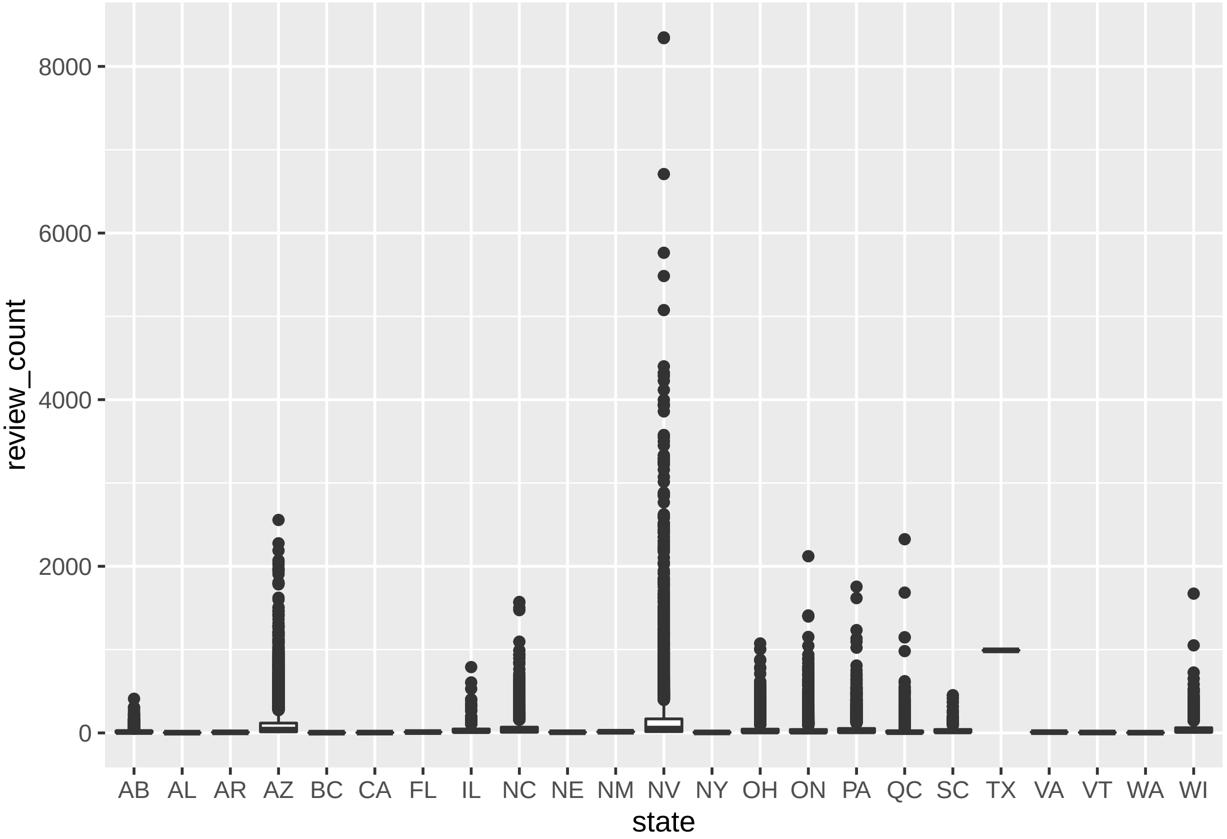
As the black area represents for states of the United States with missing value in the dataset, the analysis will be more persuasive if we focus on exploring specific states without missing values.

2.3.3 Category



From the above plot, we can tell the distribution of categories is quite dispersed, because the categories themselves not exclude from each other, and they follow a pattern of hierarchy. As a result, we choose high frequency categories that people used to have while introducing a restaurant with visual sense of distribution among various categories.

2.3.4 Review count



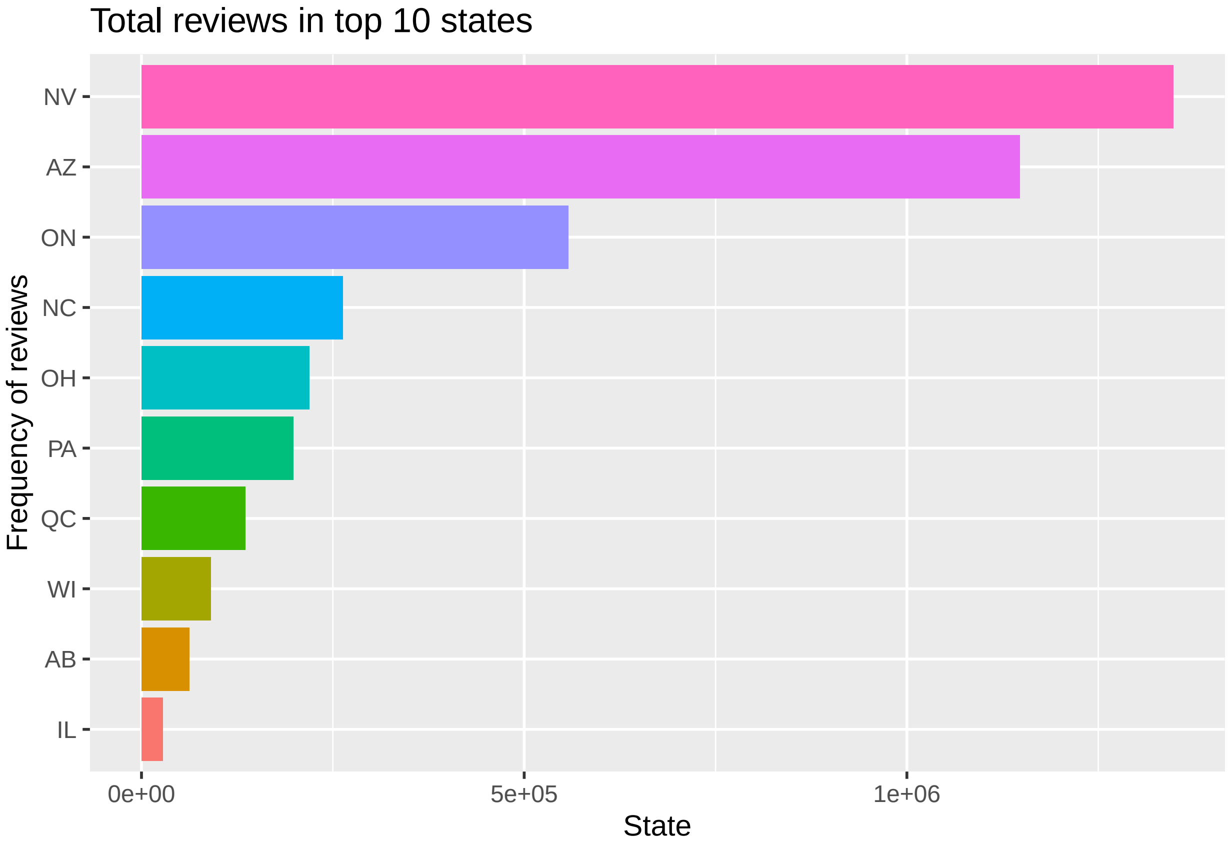
From the above plot, we can see there are only a few outliers of review count for most states, which may due to the small data size, so in order to get better distribution, we need to do further analyses on specific states with more information.

3. Exploratory Data Analysis

3.1 Review count (total)

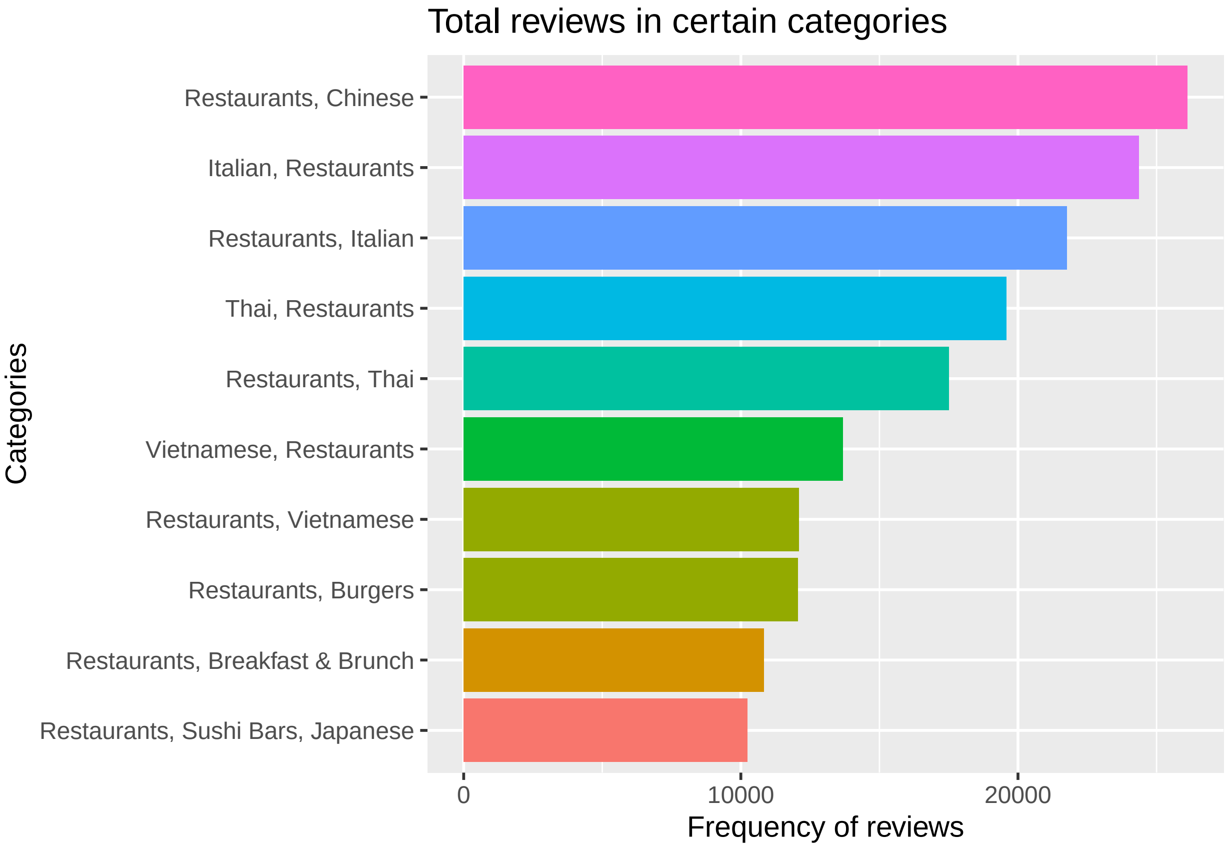
This section introduces the relationship between total review count and 3 factors: state, category, and star.

3.1.1 State



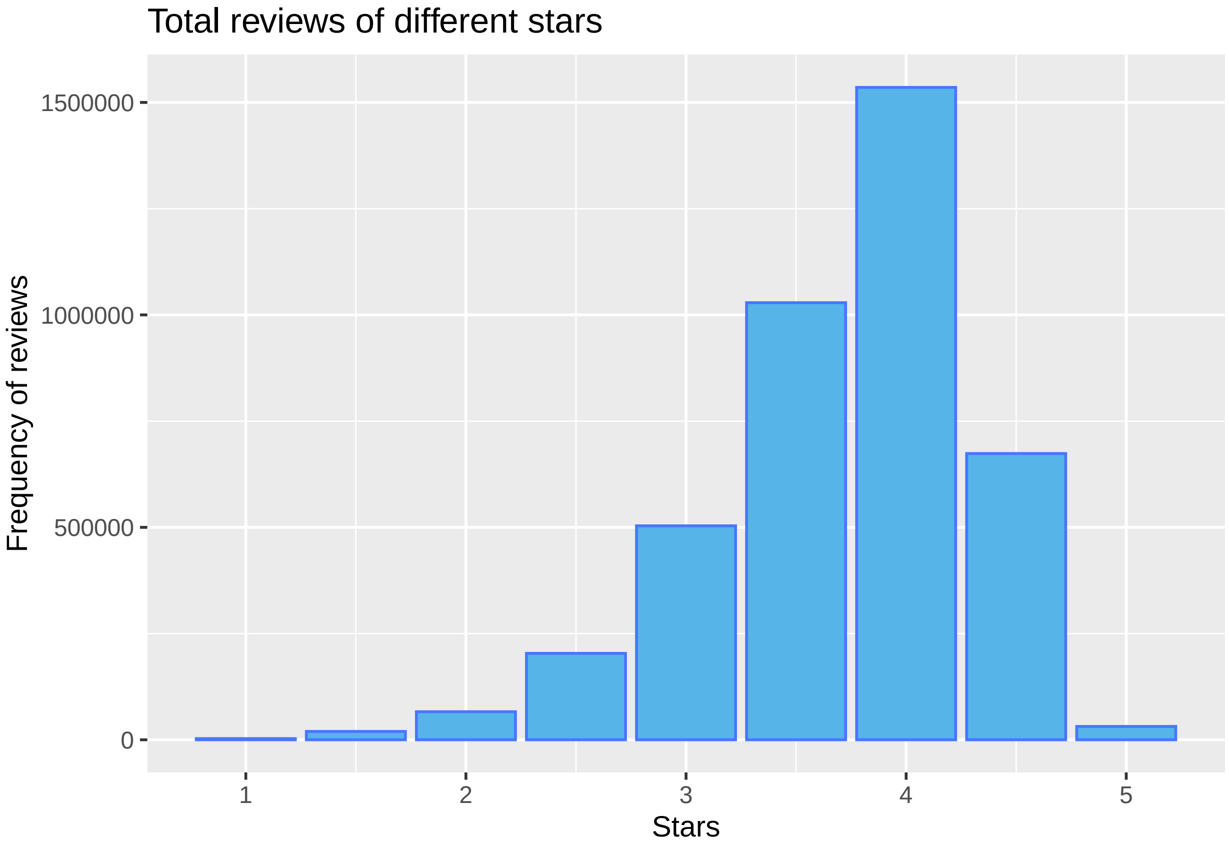
For the same reason we have mentioned in our data quality section, we only plot states with top 10 frequency. We can see that NV has the largest number of reviews, about 1.2 millions; and AZ is the second, about 1.05 millions.

3.1.2 Category



For this plot, we ignore categories like Restaurants, Nightlife, Foods which are too general, and then plot categories with top 10 review counts. It's unsurprising to see American foods take the most part, and the only Asian foods in the list is Japanese.

3.1.3 Stars

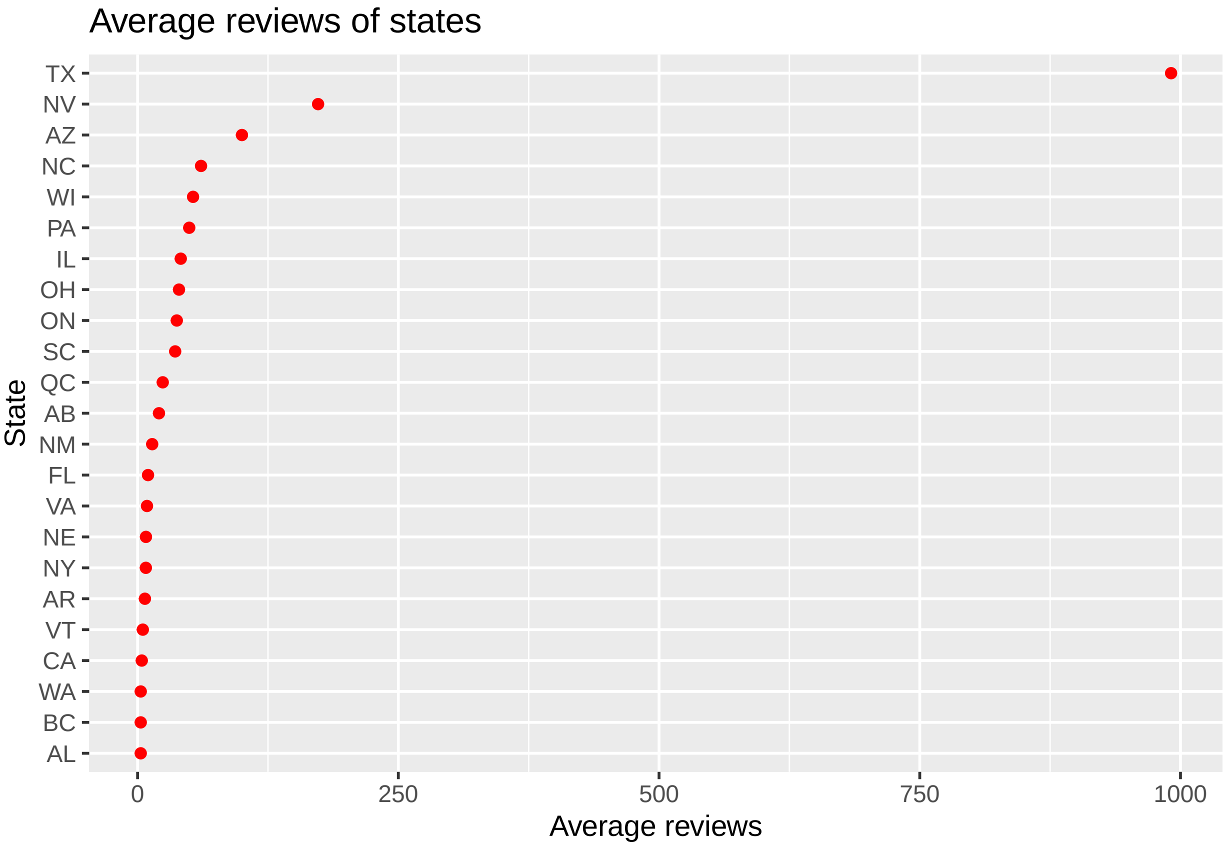


The histogram above shows that 4 stars restaurants receive far more reviews than other levels, over 6 millions. And restaurants under 2 stars and at 5 stars receive fewer reviews, probably because of the less amount of restaurants than other levels. In addition, the gap between different levels are larger than that of restaurants frequency, which means amount of restaurants is not the only reason for more reviews.

3.2 Review count (average)

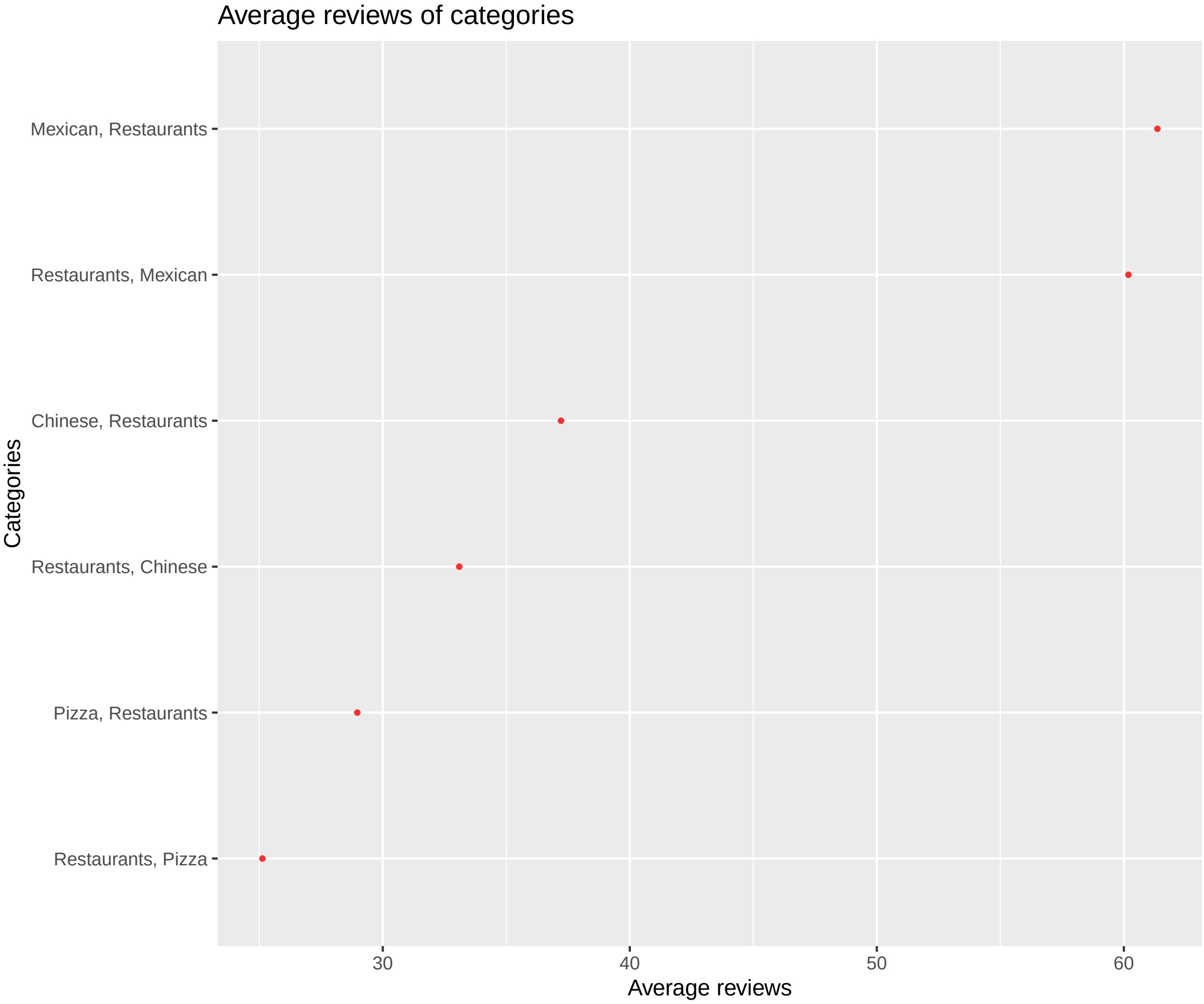
In this section, we introduce the relationships between average review count and 3 factors: state, category and star. And we will find a different pattern from total review. And we can regard average review as popularity of this level: people are more willing to give evaluations to them.

3.2.1 State



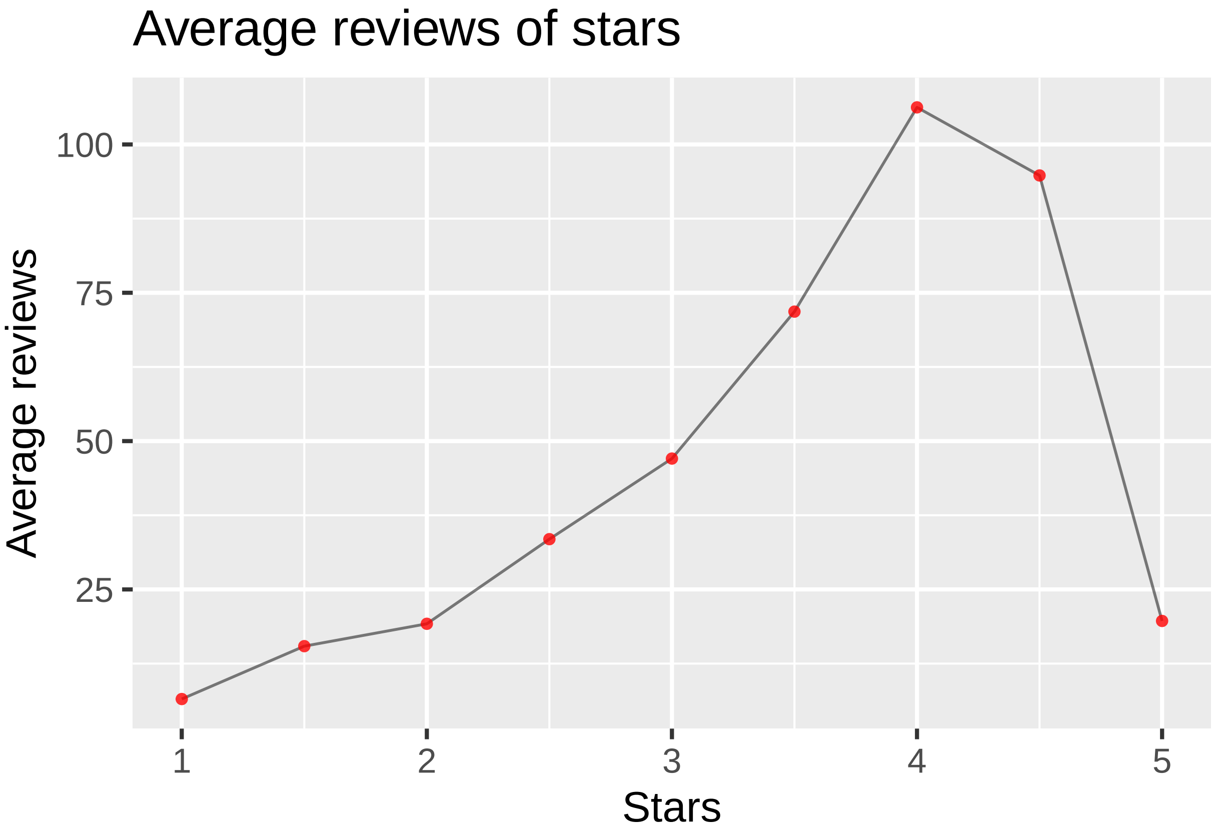
From the Cleveland plot above, we can still see NV has more average reviews than other states. Because of the bias of dataset, we can only conclude that the information about NV is more complete than other states, rather than people in NV are more likely to give reviews.

3.2.2 Category



In this figure, to only take a look of larger categories, we set retaurants frequency to be larger than 700, which can help to filter categories that are too specific. The plot indicates that American (New) and Cocktail Bar are more popular than other categories.

3.2.3 Stars

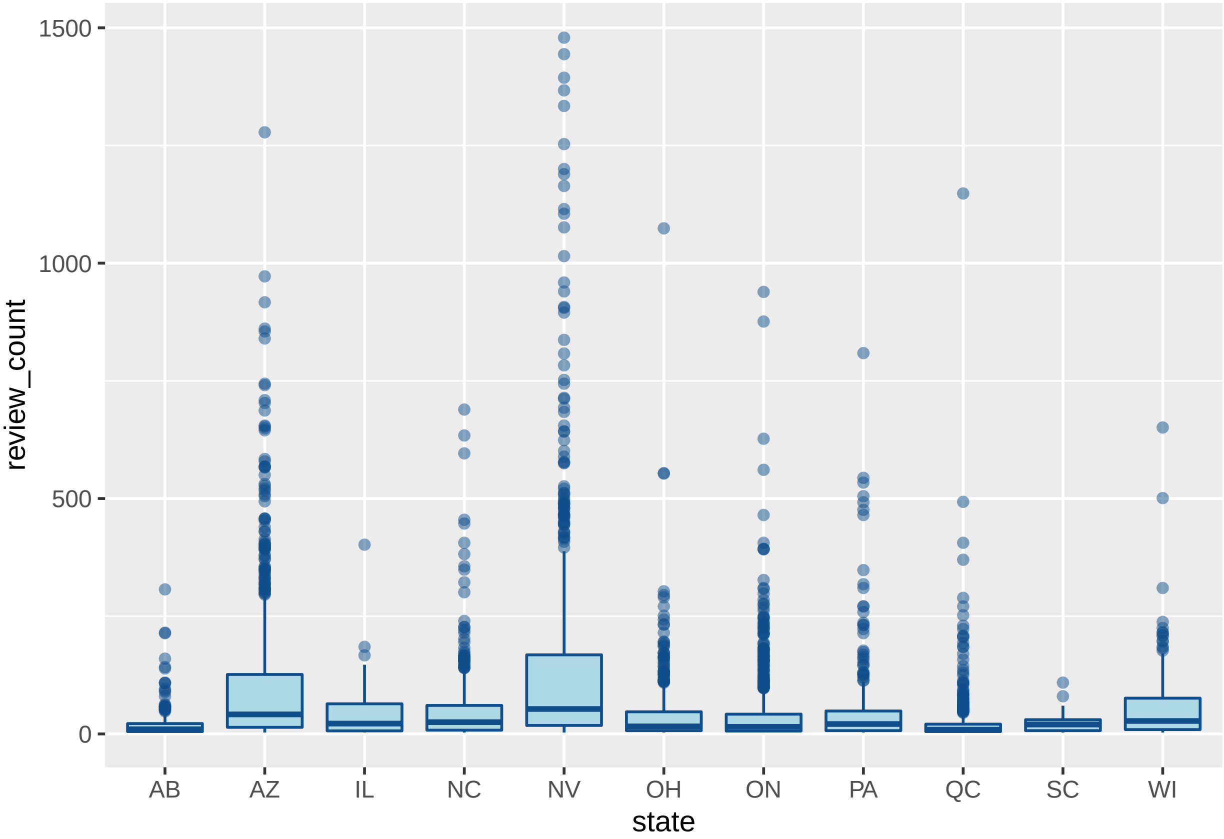


As we've discussed in the previous section, the number of the restaurants is not the only factor that leads to the variation of number of reviews. We can see from the dot plot that 4 stars and 4.5 stars are more likely to receive reviews than other levels. Also, in general, there's an increasing trend of receiving reviews from star 1 restaurants and star 4 restaurants, and decreasing trend afterwards. It makes sense because, according to our intuition, many 4 stars restaurants are "popular and good" restaurants; while higher stars, say 5, are likely to be very young restaurants: it's hardly possible for a restaurant to maintain a totally perfect feedback as long as it gets enough number of customers.

3.3 Review distribution

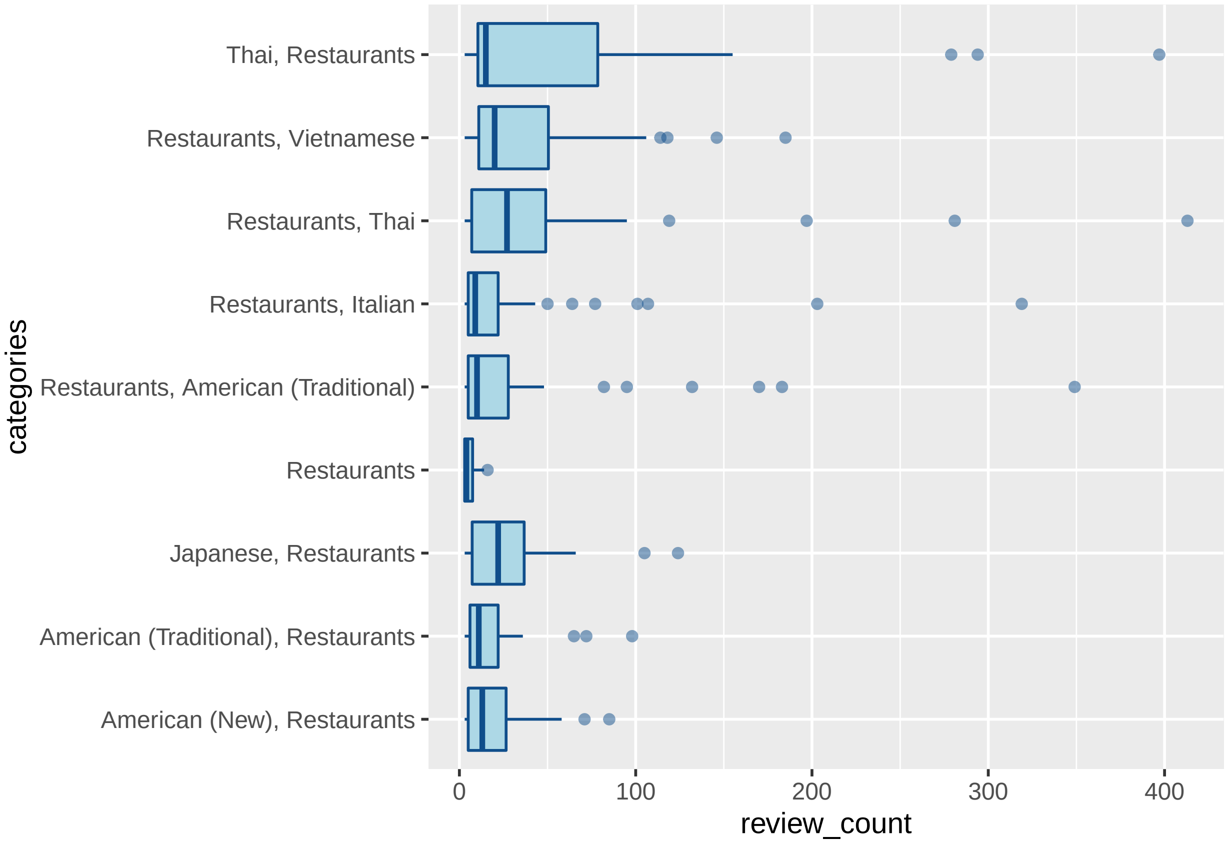
3.3.1 State

As we've mentioned in data quality section, we will only pick states with more information to see their distribution clearly. And for the reason of too much large outlier, we will get rid of outliers larger than 1500 in order to analyse the normal pattern of different states.



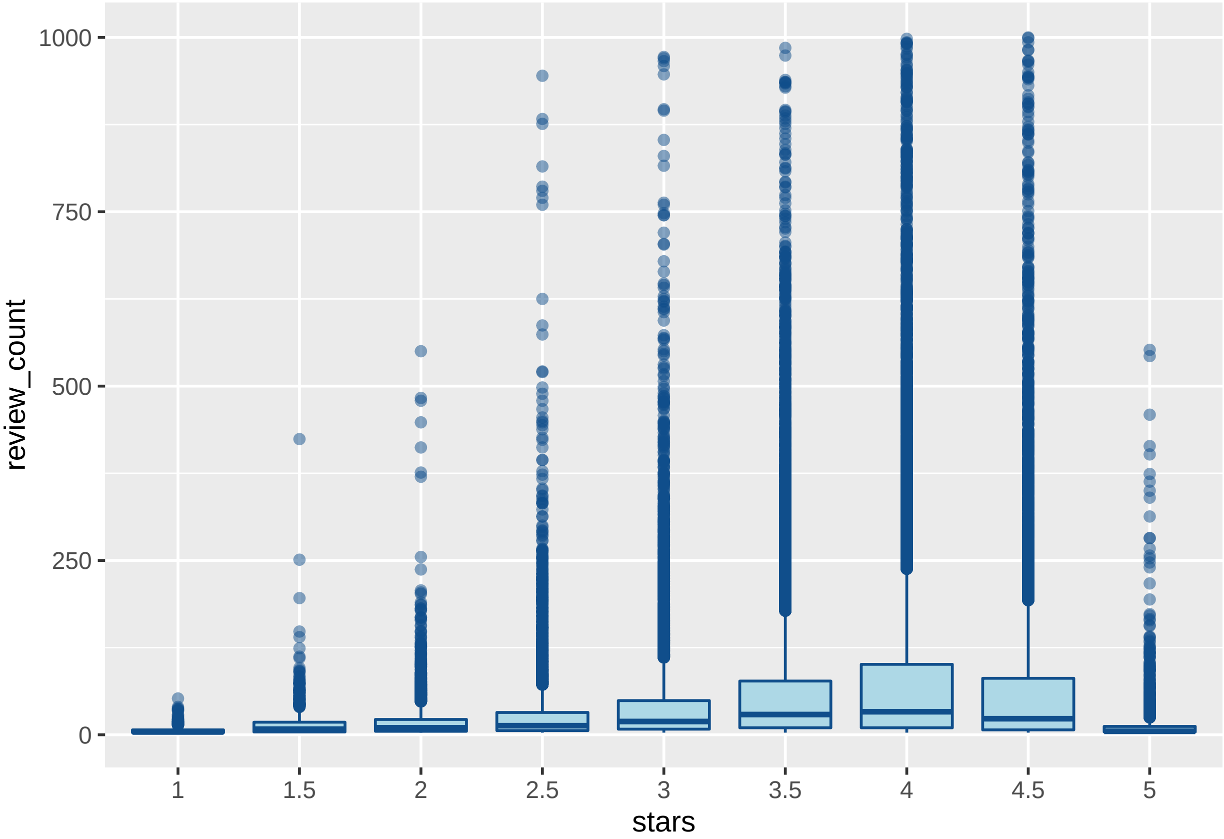
3.3.2 Category

Here we will only analyzed categories that have been chosen in the data quality section, which are more representative.



Clearly American food has a larger variation comparing to other kinds of food, while the median of these categories are quite similar to each other.

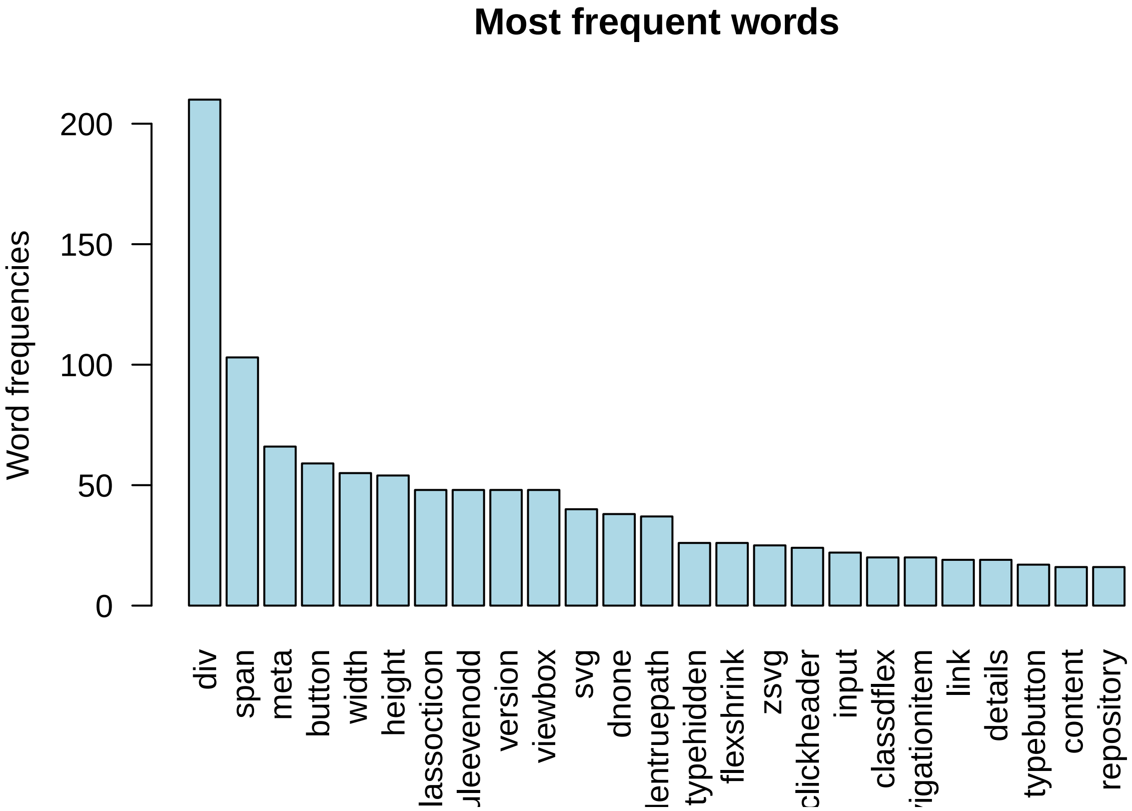
3.3.3 Star



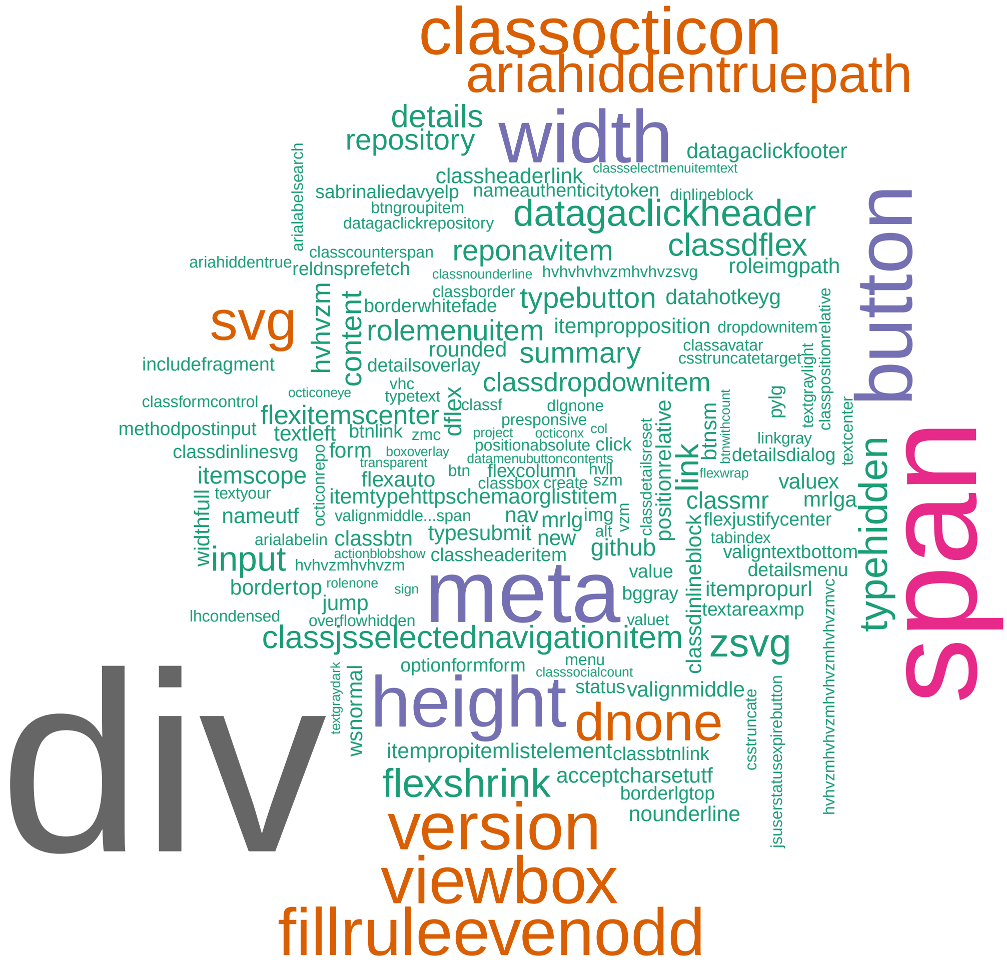
This boxplot shows that the trend of medians and Q3 of stars are same as that of average reviews.

4. Word Cloud

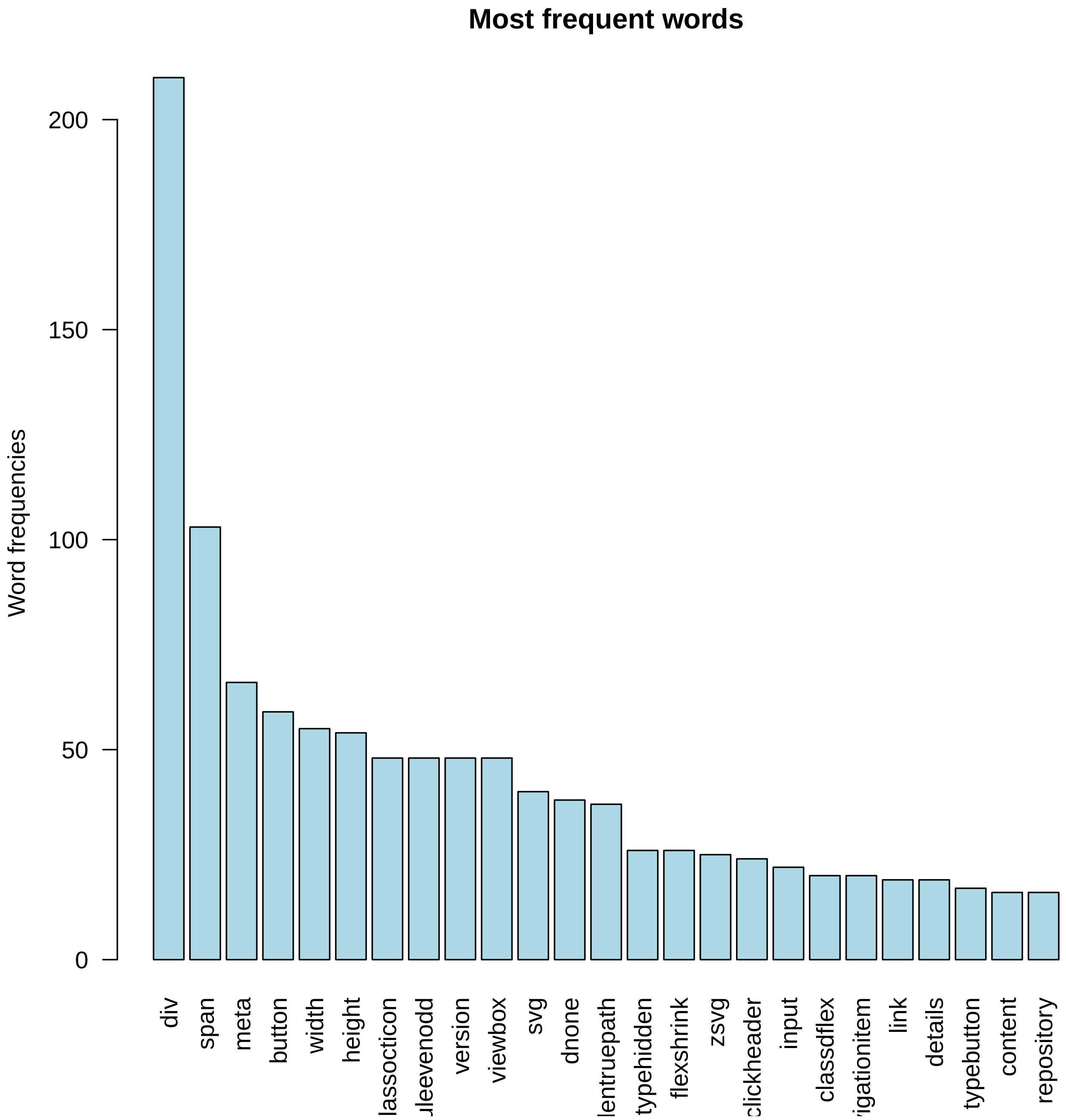
To further look at the review data, we would like to check if the words that show up more frequently in the reviews varies from different level of restaurants. Therefore, we generated two word clouds for restaurants with 1 star and 5 stars to see the difference.

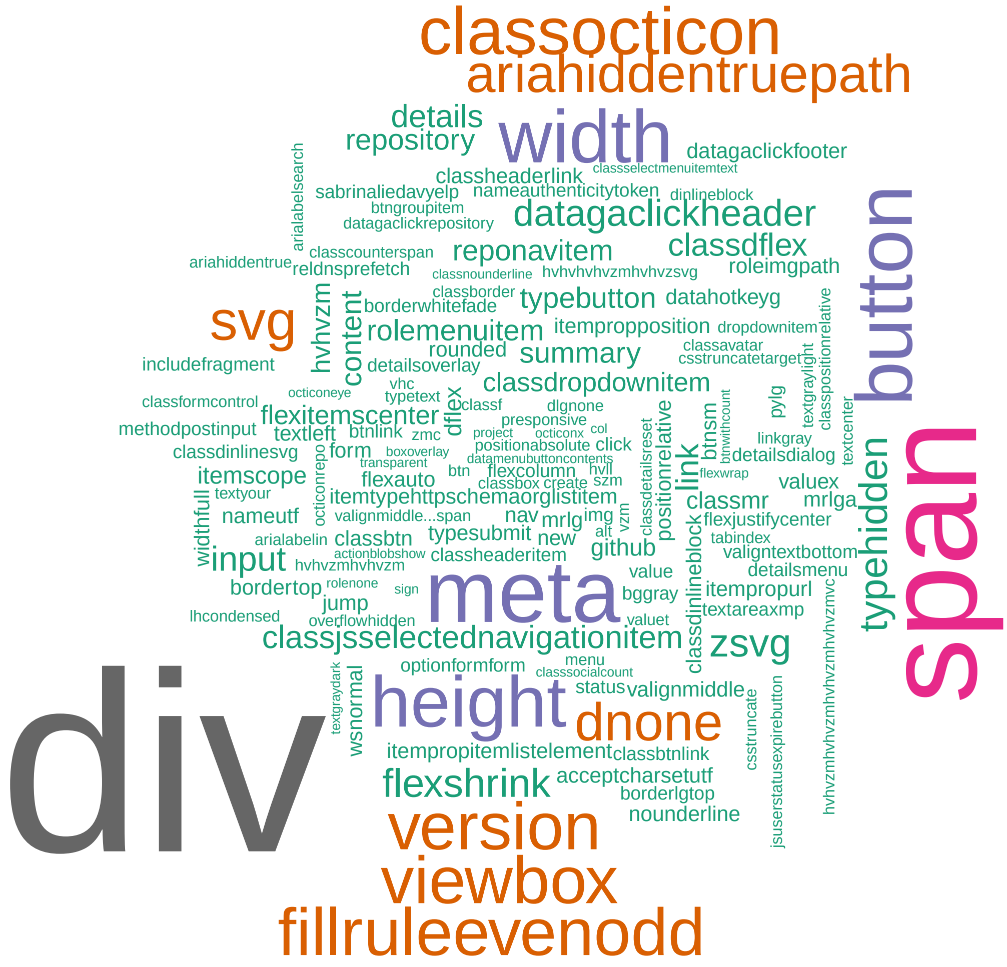


We first plot this bar plot to show to frequency of words in reviews, then plot the word cloud for a better view of the words.



From the word cloud, we clearly see that there are more nagetive words in 1-star restaurants, such as `dont`, `never`, and `horrible`.

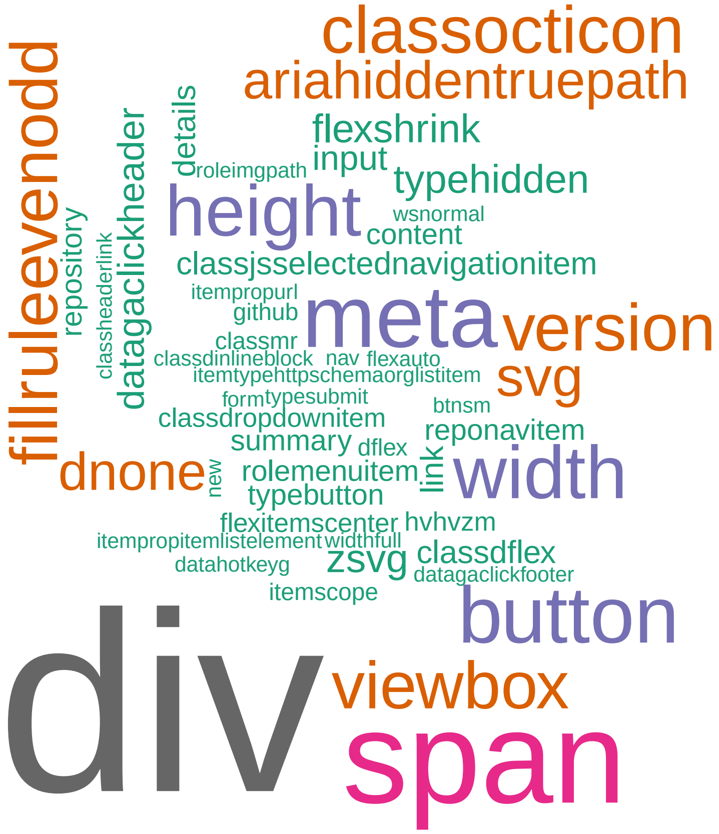

Most frequent words
Word frequencies
0
50
100
150
200
div
span
meta
button
width
height
lassocticon
ıleevenodd
version
viewbox
svg
dnone
lentruepath
typehidden
flexshrink
zsvg
clickheader
input
classdflex
rigationitem
link
details
typebutton
content
repository

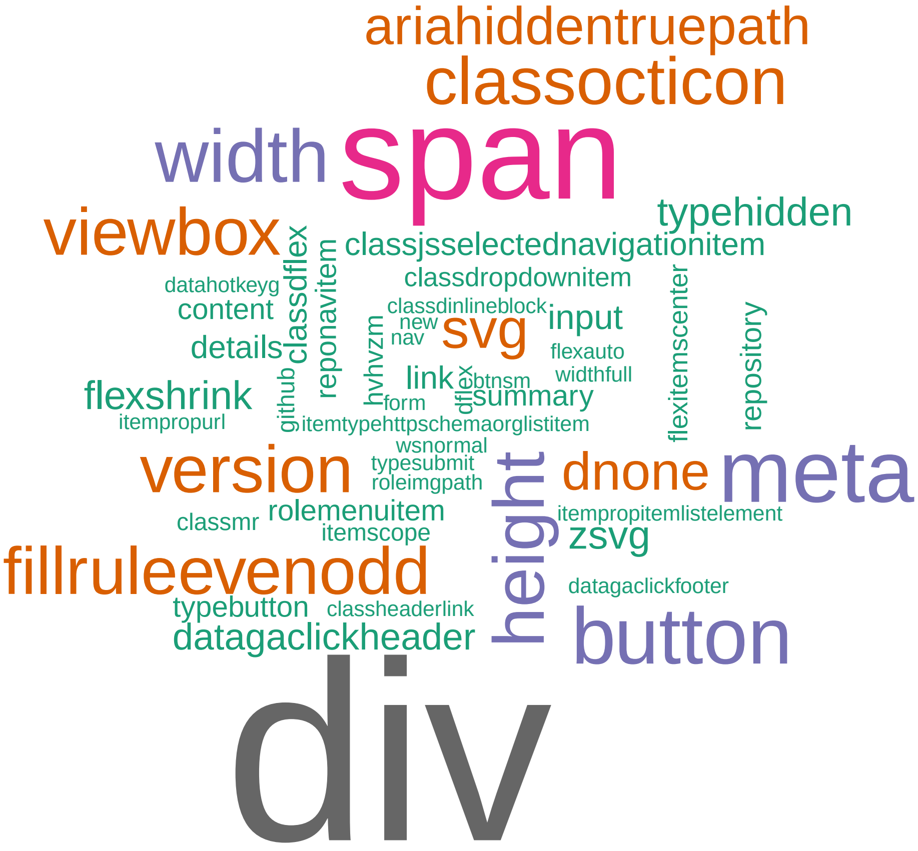


After we did the exact same thing to 5-star restaurants, we found that words such as `great`, `best`, `amazing` have a much larger frequency in the reviews. This signifigant difference shows that the reviews can actually reflect the ratings from the users.

4.3 Word cloud

In this section, we would like to analyze how stars are related to contents of review. Does some words tend to appear more in 5 stars than that in 1 star? To find out the patterns, we plotted word clouds and tried to figure out the difference.


classocticon
ariahiddentruepath
fillruleevenodd
details
datagaclickheader
flexshrink
input
roleimgpath
typehidden
repository
height
wsnormal
content
classheaderlink
classjsselectednavigationitem
itempropurl
github
meta
version
classmr
classdinlineblock
nav
flexauto
itemtypehttpschemaorglistitem
svg
form
typesubmit
btnsm
classdropdownitem
reponavitem
summary
dflex
link
dnone
new
rolemenuitem
width
typebutton
flexitemscenter
hvhvzm
itempropitemlistelement
widthfull
classdflex
datahotkeyg
zsvg
datagaclickfooter
itemscope
button
div
viewbox
span



The difference in these two word clouds shows that reviews do reflect the quality of restaurants. For example, words such as `dont`, `never`, and `horrible` clearly show up more frequently for 1-star restaurants, whereas 5-star restaurants have more reviews that include words such as `great`, `best`, and `amazing`.

4.4 Interactive Component

The word clouds that we have generated succesfully describe the difference between words that show up more frequently in terms of different ratings of restaurants. To further customize the results by controlling the number of words we would like to involve and the minimum frequency of words, we developed a Shiny app that allows users to interact with and find the patterns of words for restaurants in different levels. Here is the link to our app: https://sabrinali18.shinyapps.io/WordCloud/. Users can first choose a level of stars of restaurants, and then play around with the word cloud to see what it looks like when the minimum frequency and the maximum number of words are changed. The word cloud ranks the word frequency by the size and color of words, and the changes through different choices are clear to identify.

In addition to the word frequency that we have discovered, we can also calculate the association between two words using our results of textual analysis. For example, the we can generate the words that are highly associated with the word "service" for restaurants with different ratings. Future work can involve another interactive component that allows users to type in a word and find out the rank of the words that it is associated with.

5. Conclusion

In conclusion, we visualized the Yelp dataset with different plots to explore the patterns and relationships within and across variables that we were interested in. Throughout this project, we have gained experience of choosing certain plots for better visualizing the patterns of data. In addition, we learned the process of exploratory data analysis that starts from scratch, and how to deliver our findings to audience. More work can be done if we had a full dataset with information for every state in the US. Also, for the word cloud, we noticed that there are still some neutral words in it after we removed certain stopwords. This can be improved by taking out words that show up frequently in every level of restaurants. More analysis on text and user preference can be done in the future as well.