Rating Factors: Which Restaurant Features Influnece Yelpers' Rating?

Shan Shan |Supervisor:Professor Masanao Yajima

Fall term 2017 (MA 684)

Introduction

Background: People argue that "Yelpers" have written over 33 million reviews for local businesses, with some sporting more than 1,000 reviews each", and over time, Yelp's crowd-sourced reader rating system can make or break a restaurant."(Forbes, 2013). This altering demand effect has a far-reaching product promotion power. Thus researchers are inquirying and concerning "How Yelp Could Create More Accurate Reviews" and "Are Consumer Reviews Good for Business?" (HBS Working Knowledge). Following up their ideas, my project goal is to determine the key attributes and features that influnece ratings on Yelp. I use the Yelp online dataset and adopt data visualizations in conjunction with multi-level regression models to identify attributes and features correlated with Yelpers' ratings. I also used natural language processing to extract a frequency of information from review texts.

Project Domain: consumer rating; product ratings; user-generated content

Project Work Cycle

- **Analysis**
- First, in order to define the data mining goals to summarize data and extract better conclusions, I analysed the online Yelp dataset provided, reviewing the most questions present in pertinent domains. In order to grab the basic variables and tasks in Yelp analysis, a brief review of the different solutions in both visualization and analytics are explored.
- **Design**
- In this section, I initial the project as a result of the analysis stage (data management), and try to find general rating and review behavior and the the influential components based on Yelp data sources.
- **Development**
- In this section, I finalize those variables with detail of each subtask with EDA.
- **Validation**
- In this part, I adopt the statistics regression and ANOVA analysis to approach the validation of my developed model (described in the analysis stage)

I.Analysis

1.1 Dataset and constructs

My application sources from the Yelp 2016 Dataset (https://www.yelp.com/dataset).The dataset is provided in SQL format and its data tables includes: businesses, reviews, tips (shorter reviews), user information and check-ins.

- **Business**: Localization(postcode, longitude, latitude,state, neighborhood,city), business category, reviews, starts and open hours
- **Review** : Business, users, starts, review text, date and votes
- **User** : User, reviews, votes, average starts, friends, antiquity, compliments and fans
- **Tip** : Tip text, business, user, date
- **Check In**: Business and check in info (hours)

1.2 Project domain

In this part, I try to identify the specific questions (with the associated underlying variables) that I will develop corresponds to the behavior of targeted users: restaurants and customers(yelp users).

Restaurants

Features: price (the price range), payment method(accept credit or not), restaurant types, facilities (wifi, tv, outdoor,seating, parking), and environment (smoking, alcohol, noise level)

Questions: 1. Is the price of the restaurant affecting the rating ? 2. Is the food type affecting the rating ? 3. Are the facilities, such as with TV, Wifi,etc, affecting the rating ? 4. Is the environment affecting the rating? 5. Is the state location affecting the rating? 6. Is there any difference of between each state per se that influnece the rating? 7. Is there any difference between each reviewer per se that influnece the rating?

Reviewer/Customer

Features: rating behavior

Questions: 1. What is the configuration of the review? 2. What is the review for each restaurant types, the price, the payment method, the environment, respectively?

II.Design

2.1 Data Mining Goal

After the examination of the data structure and the research, I select the following dimensions of restaurants to measure the rating behavior online: price (the price range), payment method(accept credit or not), restaurant types, facilities (wifi, tv, outdoor,seating, parking), and environment (smoking, alcohol, noise level).

The rating behavior here could be measured via "stars" and "average stars" exchangeblly.

I select the business table lists includes a restaurant's name, location, opening hours, category, average star rating, the number of reviews about the business and a series of attributes like noise level or reservations policy: alchohol and smoking.

The review table lists a restaurant's star rating, the review text, the review date, and the number of votes that the review has received.The texts from those restaurants reviews will form the corpus of this project, with the text mining analysis.

III. Development

3.1 Data prcoessing

*Data collection The data I used here is from the public dataset http://www.yelp.com/dataset_challenge .

*Transform SQL and load the data in R For this purpose I used the SQL read and save in Rds objects.

*Clean and filter the data

In the cleaning data process, I omit variables with NA value. Merge the business, reviews and categories into "myyelp.Rds" file, with the variables I need.

*Please refer to my Yelp_DataClean.Rmd file for details.

3.2 EDA and Visualization

- I read the prepared RDS file, use the SQL and R language to do the EDA and Visualization.
- Toolket: R,SQL,NPL
- Package and functions: ggplot2::ggplot,RMySQL,dplyr,"tm"" is the text mining package; wordcloud(used as self-explanatory)

business_id	restaurant_style	name.x	neighborhood	address
fN7ds9Dk4IfMsh4RT9x82w	Japanese	J's Kaiyo Sushi & Bar		4412 N Miller Rd
fN7ds9Dk4IfMsh4RT9x82w	Japanese	J's Kaiyo Sushi & Bar		4412 N Miller Rd
fN7ds9Dk4IfMsh4RT9x82w	Japanese	J's Kaiyo Sushi & Bar		4412 N Miller Rd
fN7ds9Dk4IfMsh4RT9x82w	Japanese	J's Kaiyo Sushi & Bar		4412 N Miller Rd
fN7ds9Dk4IfMsh4RT9x82w	Japanese	J's Kaiyo Sushi & Bar		4412 N Miller Rd
fN7ds9Dk4IfMsh4RT9x82w	Japanese	J's Kaiyo Sushi & Bar		4412 N Miller Rd

city	state	postal_code	latitude	longitude
Scottsdale	AZ	85251	33.5005	-111.918
Scottsdale	AZ	85251	33.5005	-111.918
Scottsdale	AZ	85251	33.5005	-111.918
Scottsdale	AZ	85251	33.5005	-111.918
Scottsdale	AZ	85251	33.5005	-111.918
Scottsdale	AZ	85251	33.5005	-111.918

stars.x	review_count.x	is_open	pricerange	creditcard
4.5	66	yes	2	yes
4.5	66	yes	2	yes
4.5	66	yes	2	yes
4.5	66	yes	2	yes
4.5	66	yes	2	yes
4.5	66	yes	2	yes

parking	wifi	tv	outseating	alcohol
{“garage”: false, “street”: false, “validated”: false, “lot”: false, “valet”: false}	free	yes	yes	full_bar
{“garage”: false, “street”: false, “validated”: false, “lot”: false, “valet”: false}	free	yes	yes	full_bar
{“garage”: false, “street”: false, “validated”: false, “lot”: false, “valet”: false}	free	yes	yes	full_bar
{“garage”: false, “street”: false, “validated”: false, “lot”: false, “valet”: false}	free	yes	yes	full_bar
{“garage”: false, “street”: false, “validated”: false, “lot”: false, “valet”: false}	free	yes	yes	full_bar
{“garage”: false, “street”: false, “validated”: false, “lot”: false, “valet”: false}	free	yes	yes	full_bar

3.2.1 Content of Yelp Reviews

The text analysis displays the content of Yelp reviews that has the most frequency as “food” place time great as the sensitive words.It captures the holistic picture of the reviews on food types(eps. for “drink”, “meat,”sandwich“,”burger“) and location (”place“) reviews often discussed.

3.2.2 Number of Yelp Reviews

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

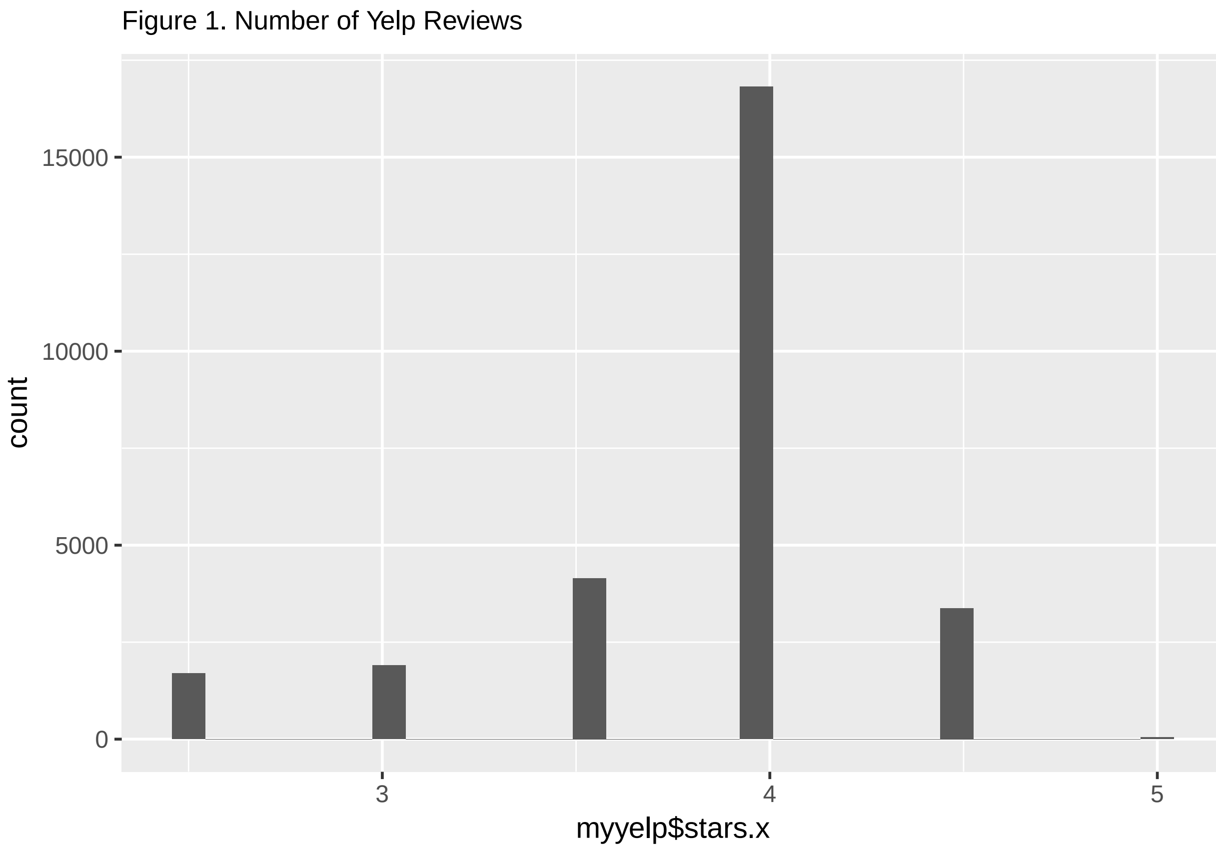
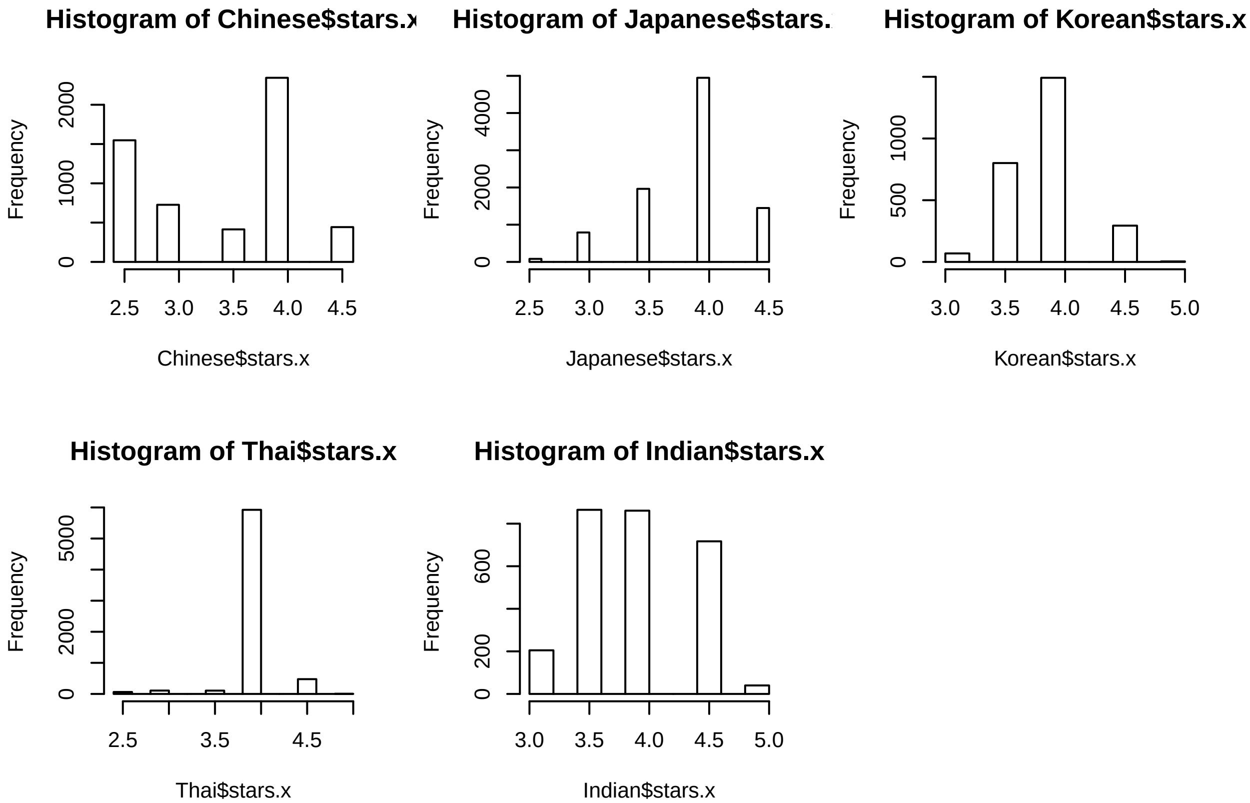


Figure 1 presents that the number of Yelp Reviews shows the four-star-review occupies the largest amount reviews, approximately 16000 pieces; and we could assume the sour-star-review restaurants' popularity based on the reviews frequency; and also the possibility that consumers are attracted by these four star restaurants. The five-star reviews has the minimum amount of reviews.

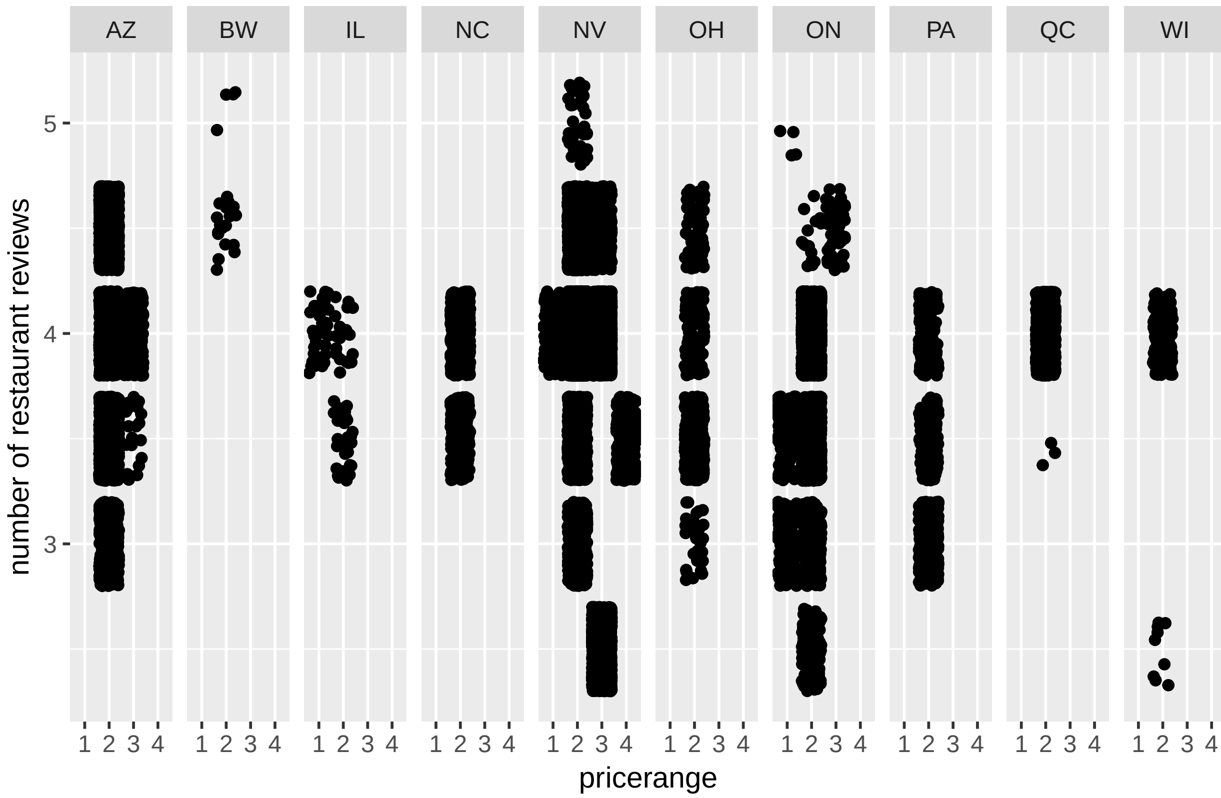
3.2.3 Restaurants rating amount by type



The figures of 3.2.3 display that the rating distribution varied through restaurants type. I examine five types of restaurants: Japanese, Chinese, Korean, Thai and Indian. Among them, in regards to the highest five-star reviews, Japanese ranks the first, and Chinese restaurants are the second. Of the four-star reviews has the same rating distribution: Japanese restaurants and Thai restaurants have the most substantial rating amount as around 5000, and Chinese reaches 4000 as well. From such observations, we could assume the connection between the rating behavior and the restaurant types.

3.2.4 Distribution of average stars of restaurants by states

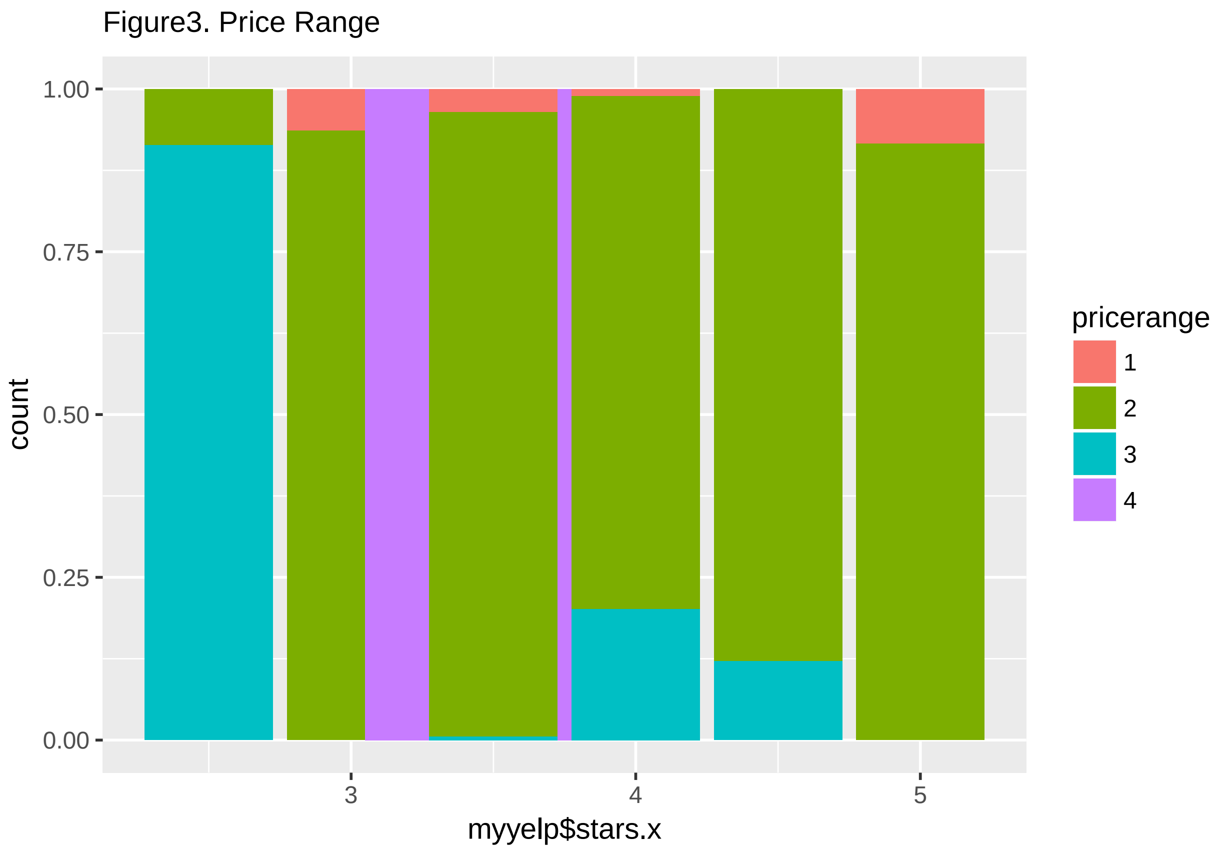
Figure2. Number of Restaurant Review by State



Rating and restaurat type by states

Physically, restaurants have the “space” and “field”, thus, the geographic location should be considered. As in my analysis question, I consider the domain of the “restaurants” and regard the “state” variable as a potential influence. Is there any state-influence that effects consumer rating? The figure above presents indicates the geolocation of the business reviews come from. Based on the location, we could assume the restaurants that traced by Yelp are clustered in the northeast part and the southwest part as presented: NV, OH, OR, PA, AZ, have the wide rating actions. Within each state, the rank of stars are also varied: AZ’s restaurants’ evaluation are comparatively has a 4-star-level, NV has four stars, ON has 3-4 star-level frequently.

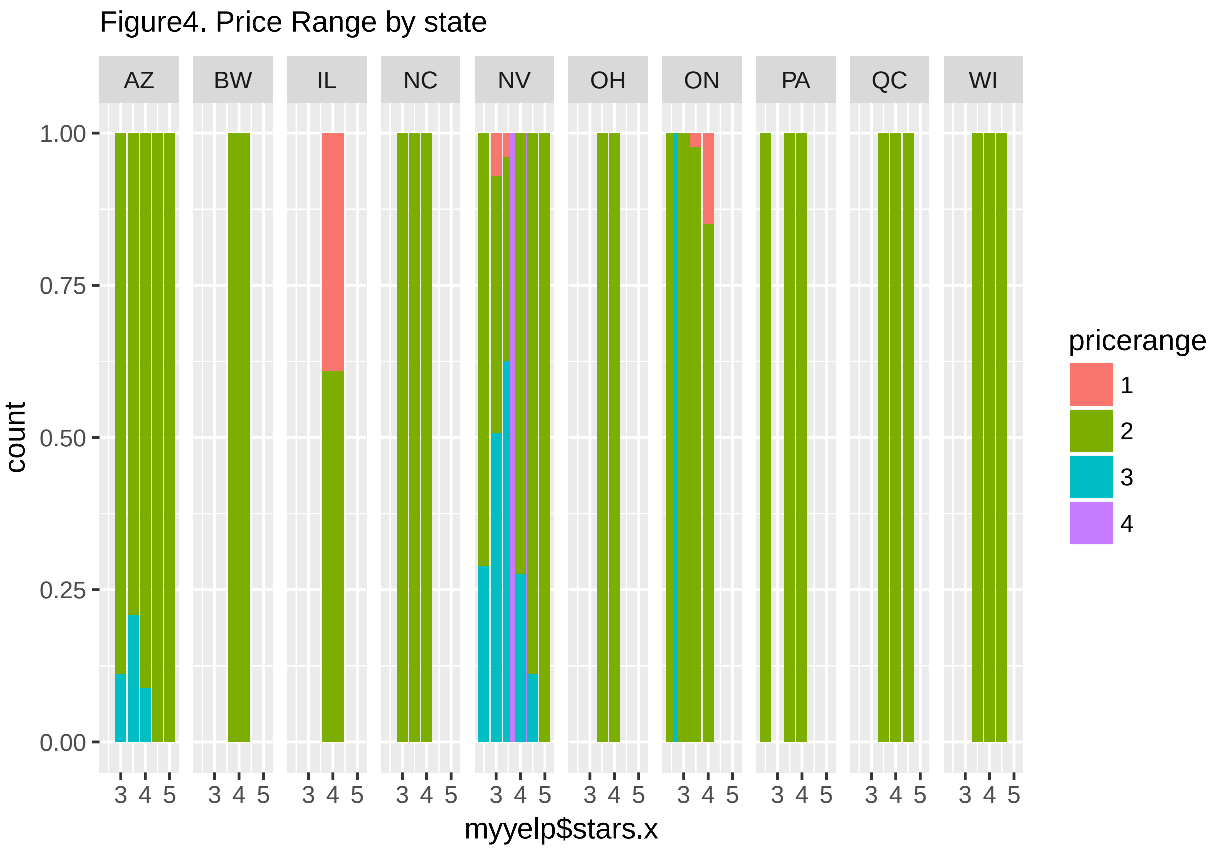
3.2.5 Price Range and Rating



Price range and rating

Price range's interaction with the rating is discussed above. Price range as 2 is most welcomed with the high rating. Some price range 4 (the higher price range one) generally has the review among 3-4. Price-range 3 has the lower rating below 3 stars. Thus, generally, to speak, people prefer the lower price range.

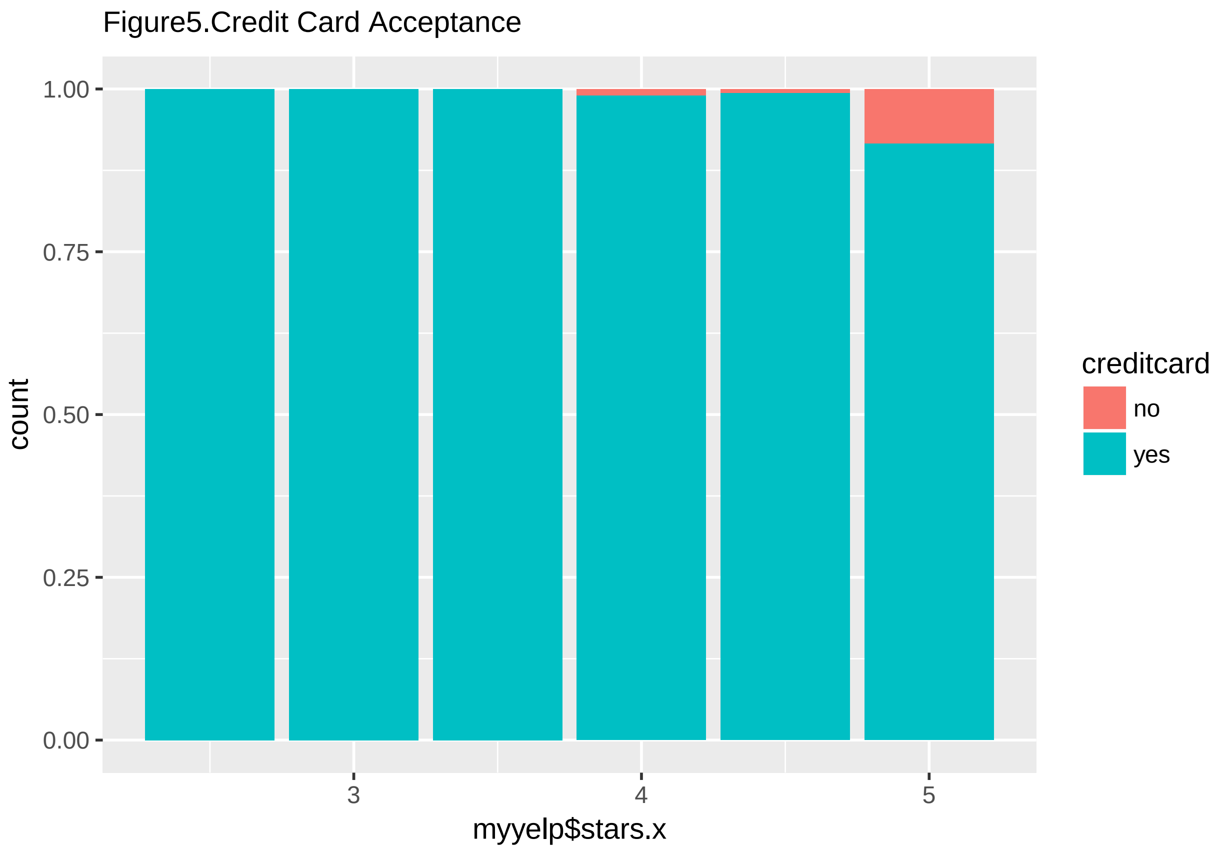
3.2.5 Price Range and Rating by state



Price Range and rating by state

Zoom into each state; the similar rating pattern displayed: consumer like lower price range, particularly price-range 2.

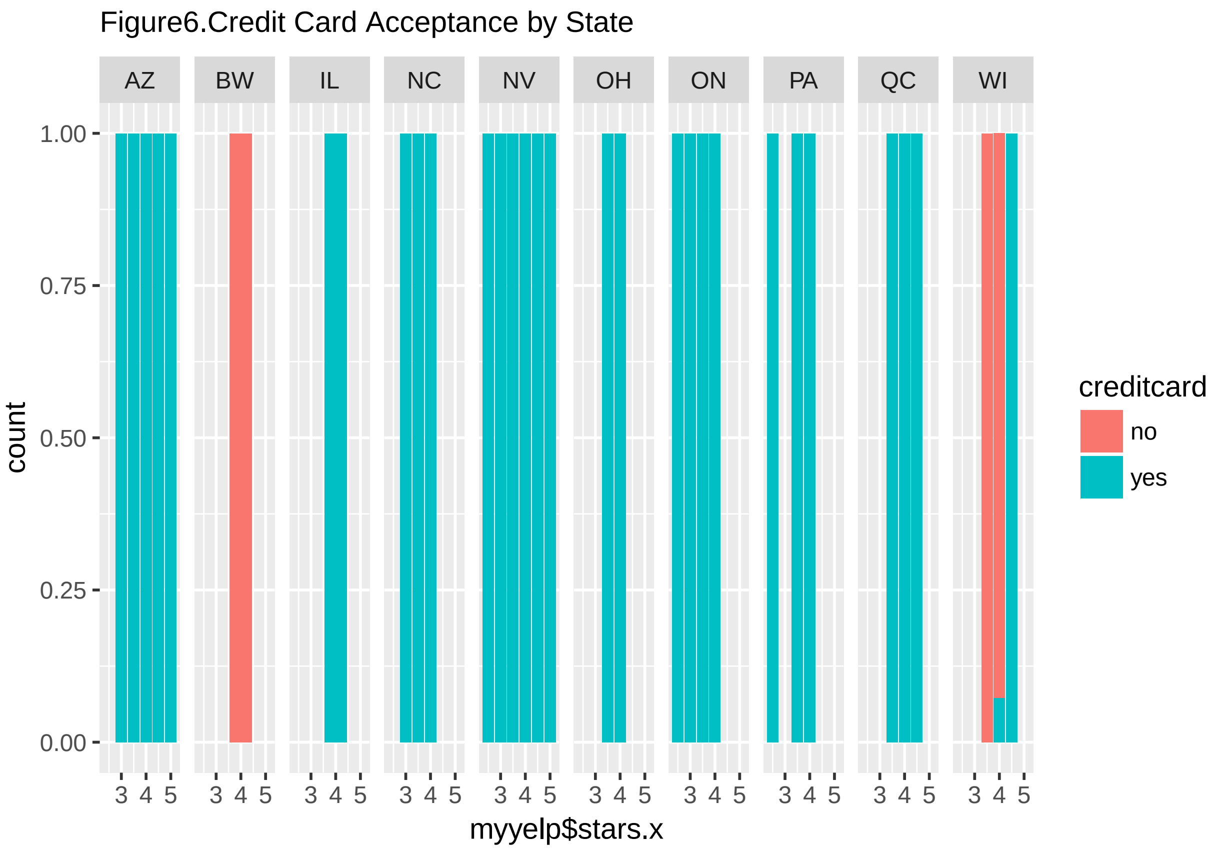
3.2.6 Payment Method and Rating



Payment method and rating behavior

Figure 5 shows most restaurants adopt the credit card accepted.

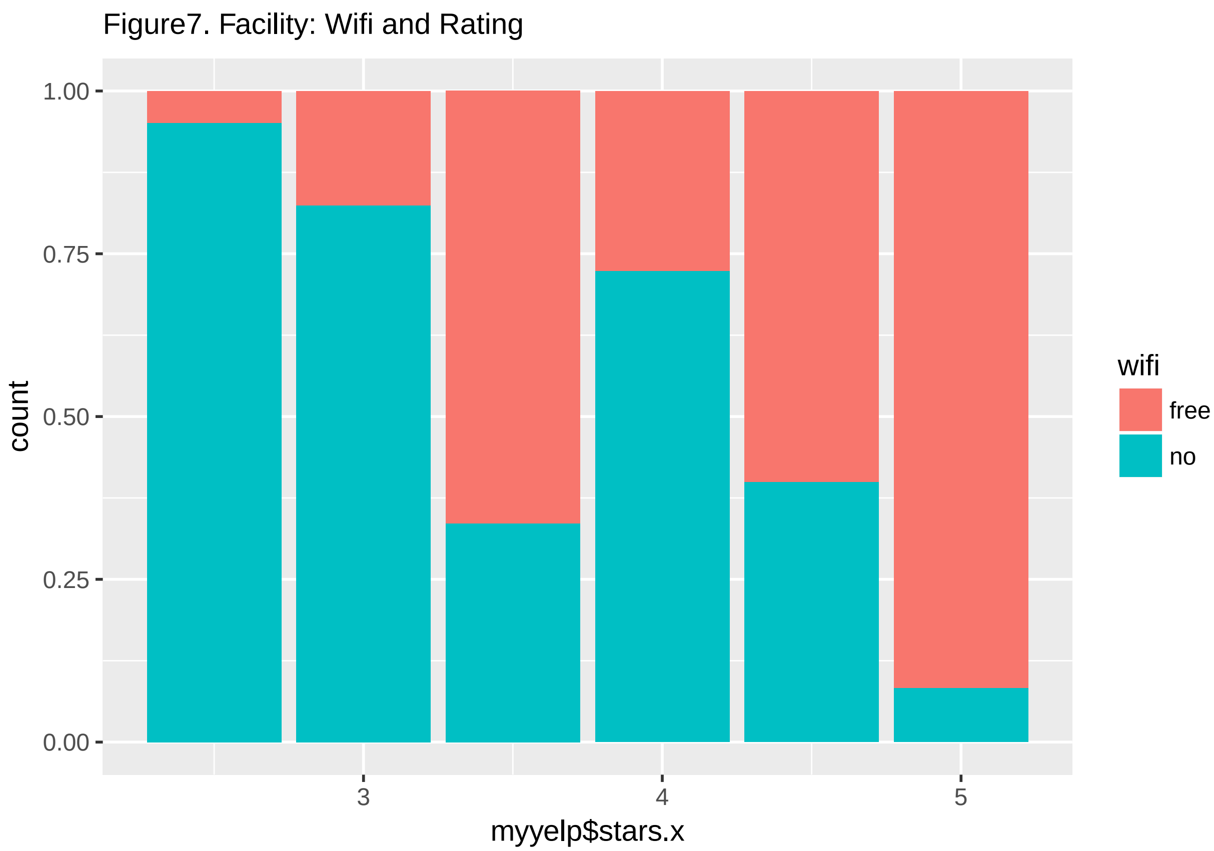
3.2.6 Payment Method and Rating by state



Payment method and rating behavior by state

Figure 6 displays the similar phenomenon within the states (like BW, WI). However, if the payment method is a common thing and could not tell the difference between each restaurant, then, it is hard to say that it will influence the rating.

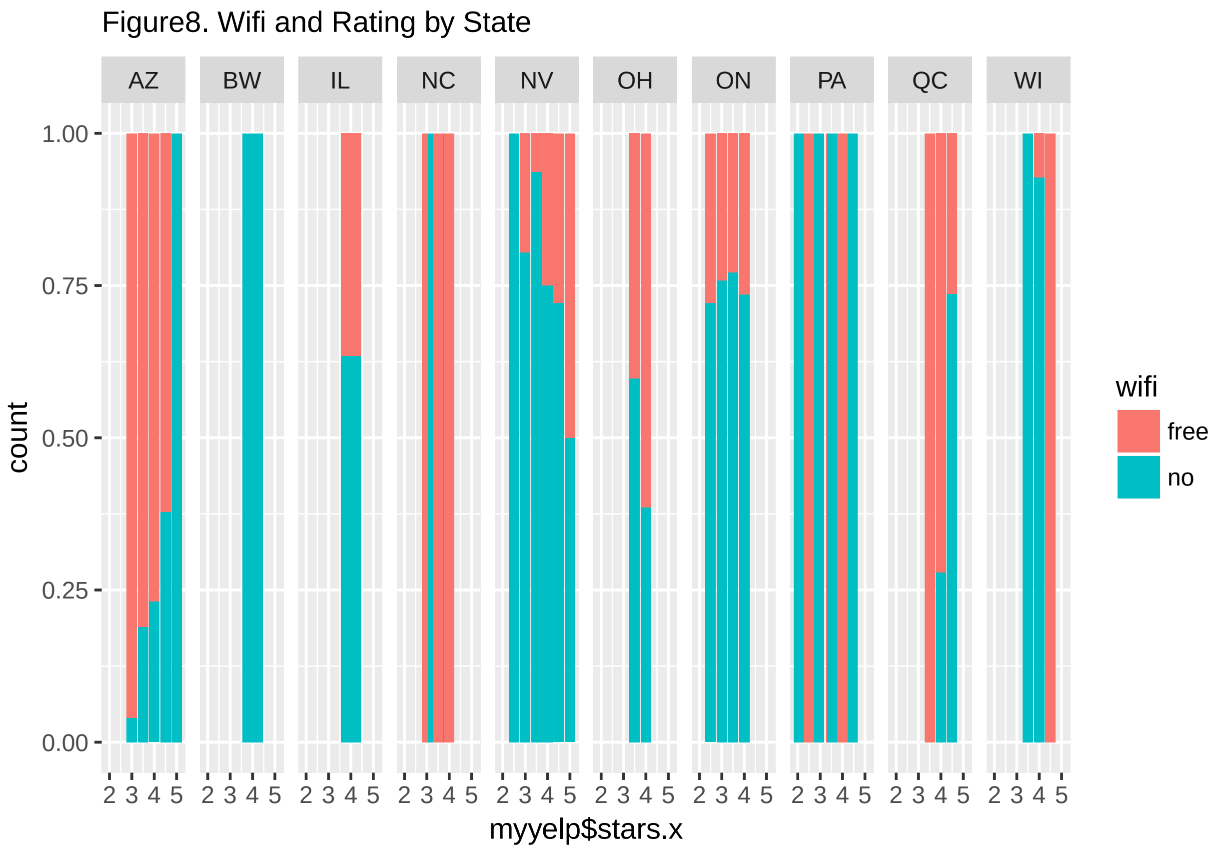
3.2.7 Wifi and Rating



Wifi and rating

Figure 7: Restaurants with free wifi have more higher level reviews amount than wifi without free wifi.

3.2.7 Wifi and Rating by state



Wifi and rating behavior by states

Fig 8: Within each state, the similar situations happened in OH, WI, NC states: restaurants with free wifi has higher level reviews amount than wifi without free wifi.

3.2.8 Rating and Parking

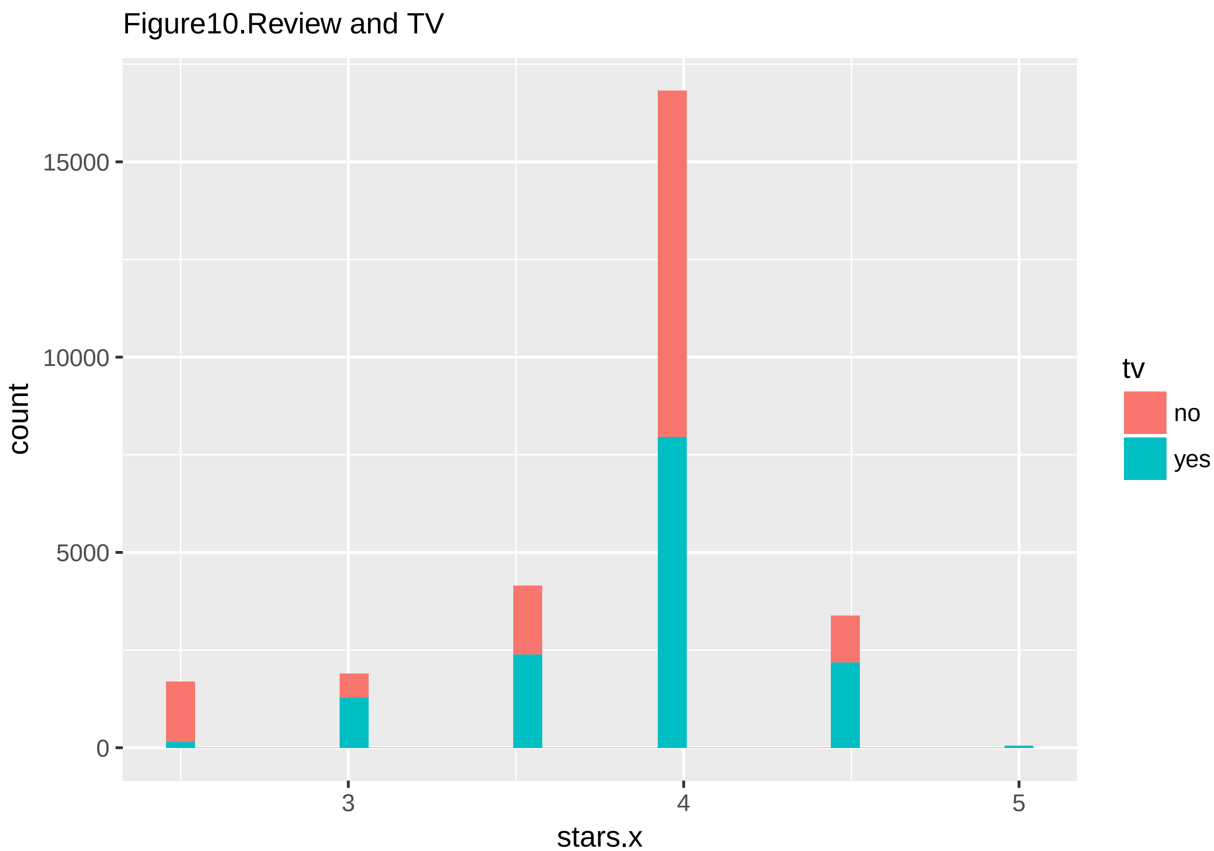
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Rating and Parking

Among the parking types, "lot" parking is the most popular rated one in all five categories. The parking type as garage and valet seems to relate to low rating level. However, this observation needs to be further discussed and tested, as parking and without parking is the first condition; and within "has parking", what is the different influences of each parking type?

3.2.9 Rating and TV facility

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

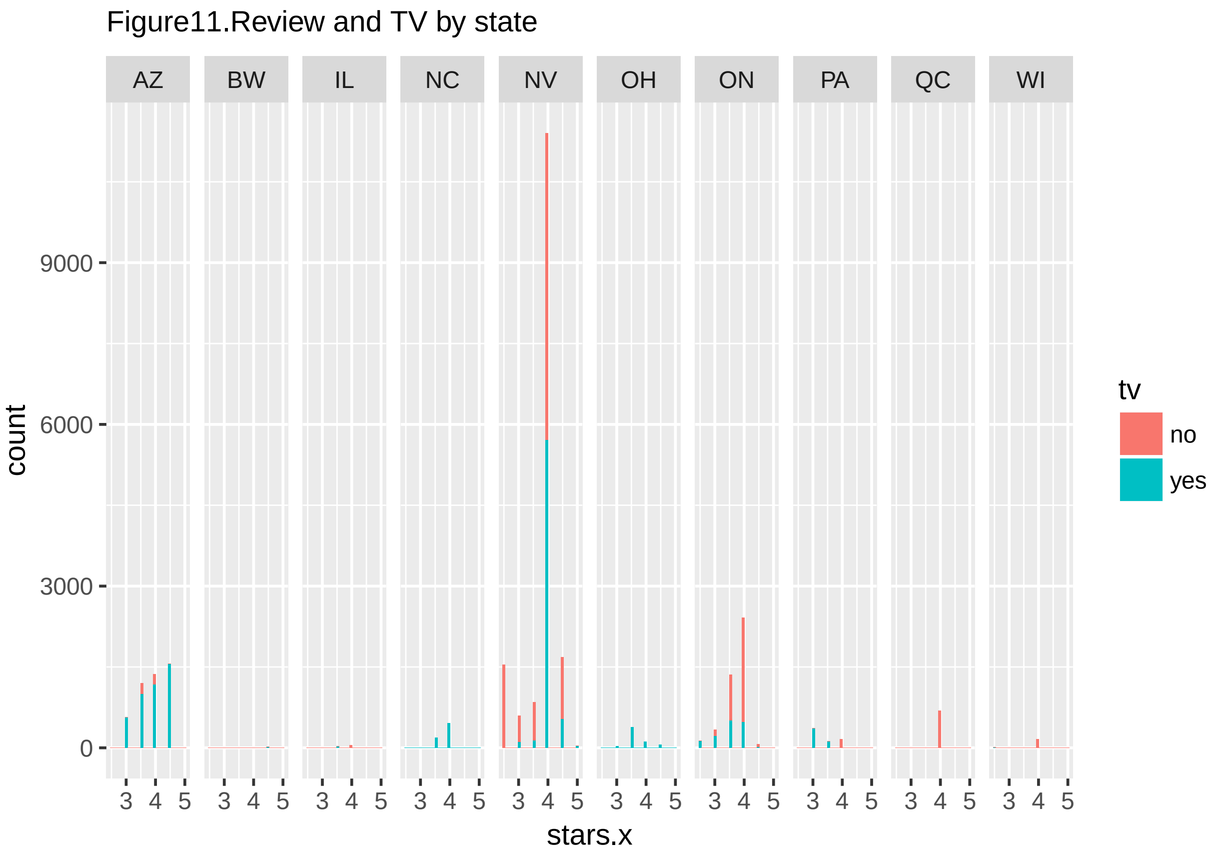


Rating and TV

It seems that restaurants with TV and without TV have no significant influence towards the reviews. But generally to speak, at the low-level rating (below three-star-rating), the no-tv phenomenon is visible.

3.2.9 Rating and TV facility by state

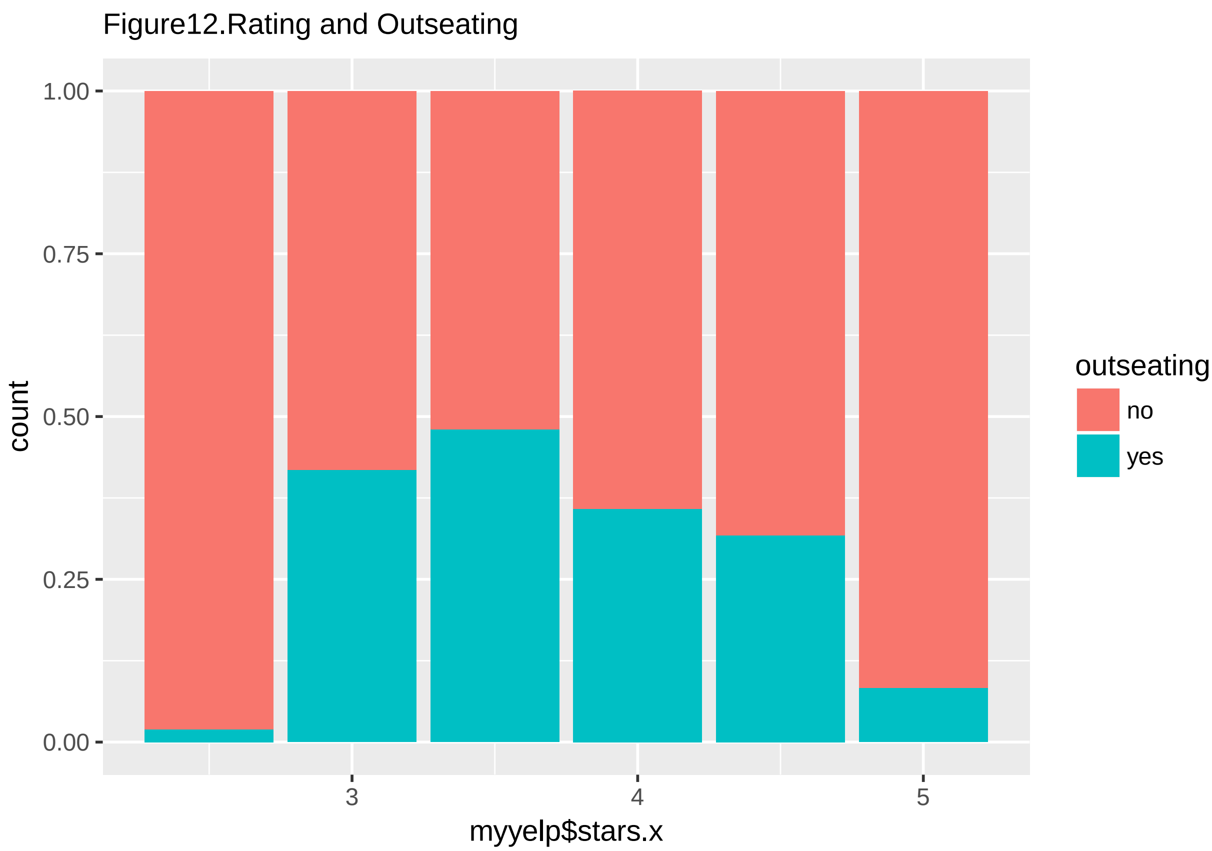
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Review and TV by state

It seems the rating and TV have no obvious relation within states as well: only NV and ON, AZ presents specific patterns, other states' situation is not apparent.

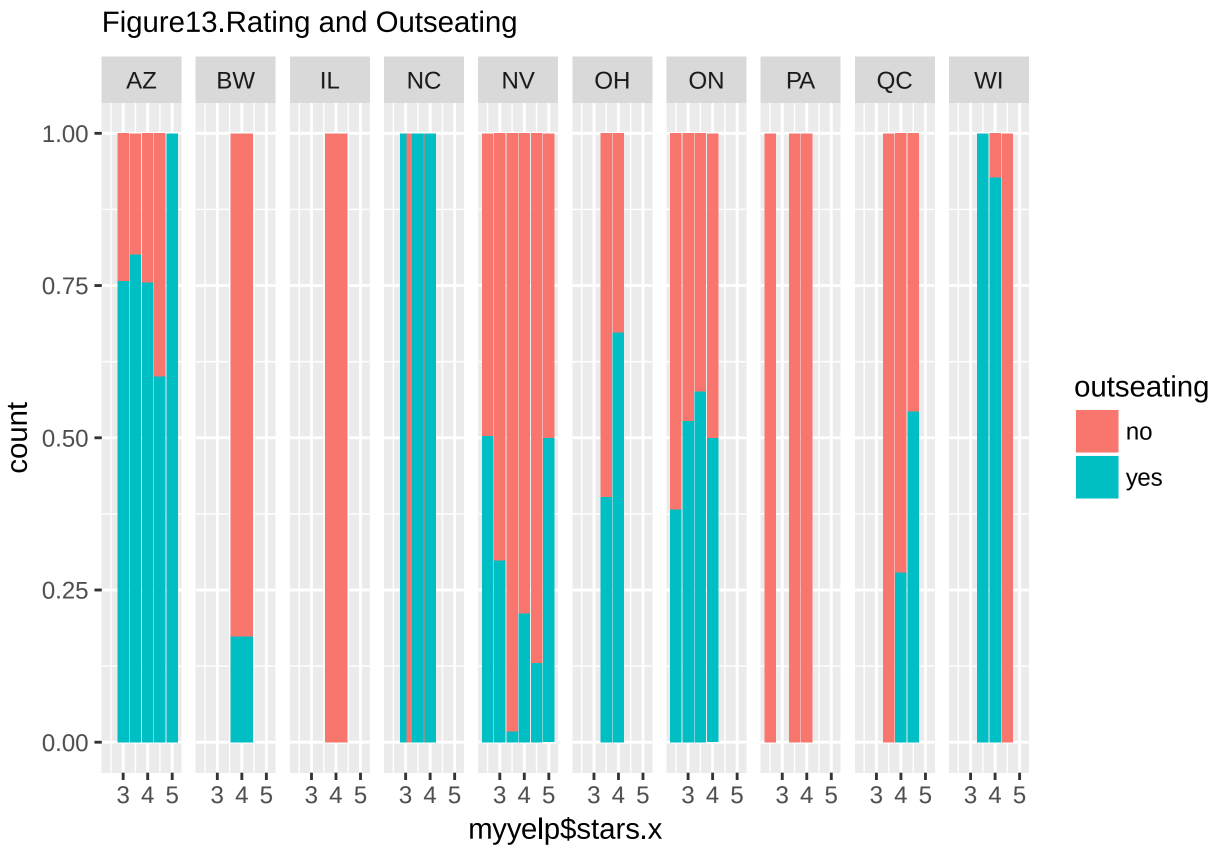
3.2.10 Rating and Outseating



Rating and Outseating

Restaurant with out eating or not does not influence too much of the restaurant rating.

3.2.10 Rating and Outseating



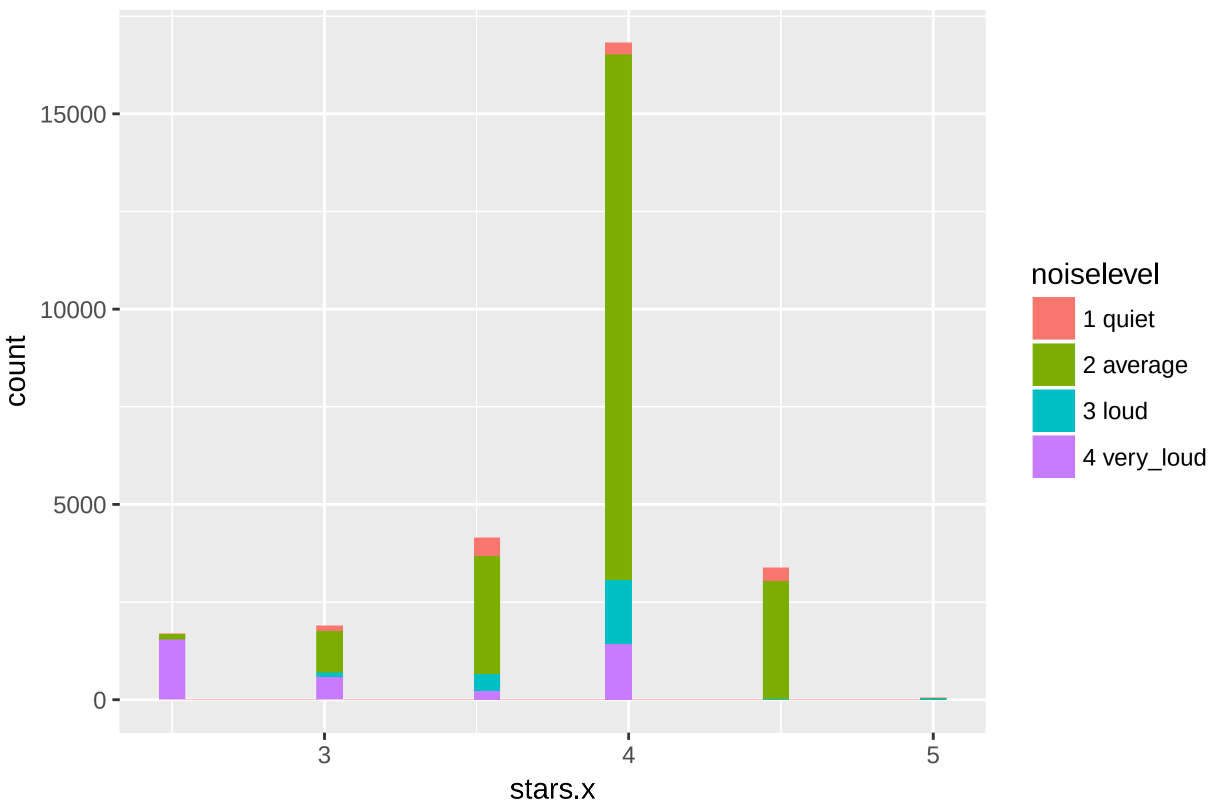
Rating and Outseating by state

Though, restaurant with out eating or not does not influence too much of the restaurant rating. However, the state has some influence on the evaluation that considerably change the pattern of the rating behavior. For instance, in NV it seems that higher restaurant could have no out seating, while in NC it seems that restaurant ranking has a positive trend with the out seating.

3.2.11 Rating and Noise Level

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 14.Rating and Noise Level

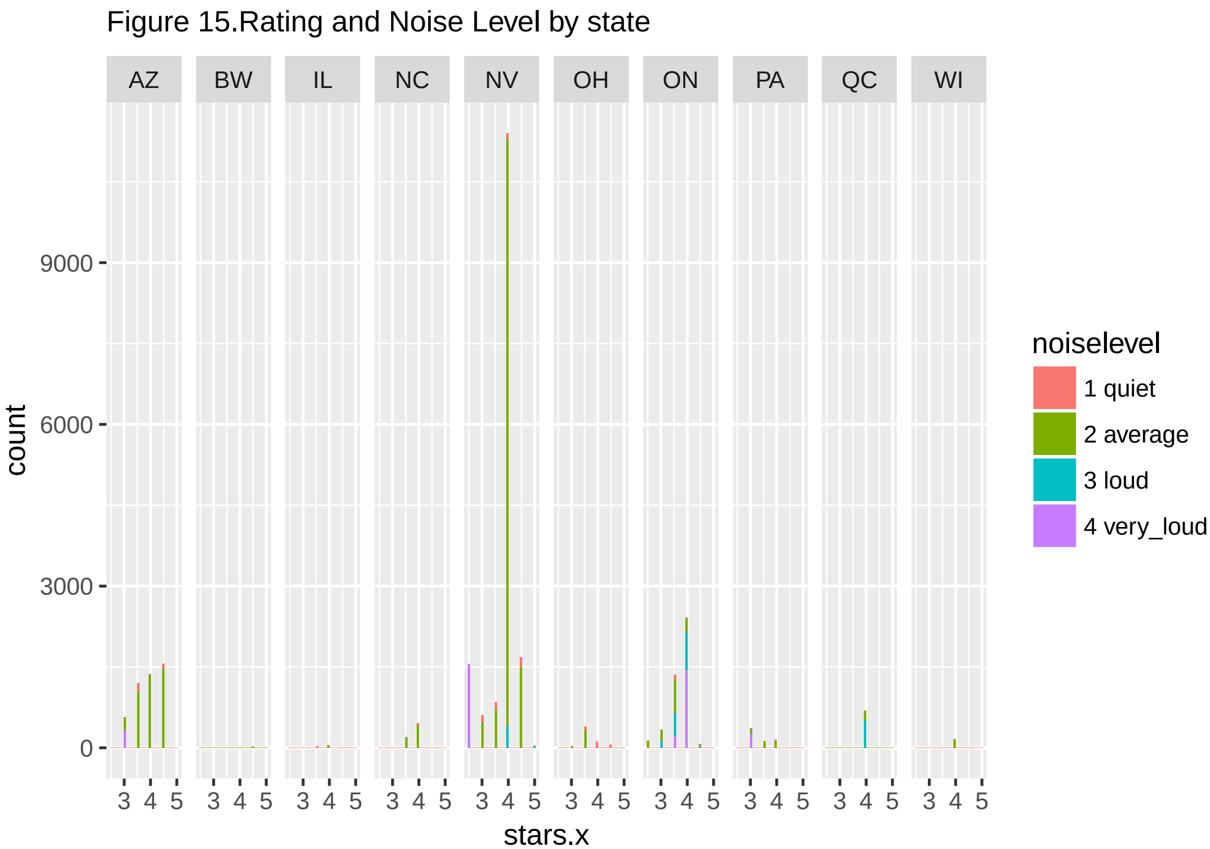


Rating and noise level

It seems the average noise level has the highest amount in both top rating and low rating. Relatively to speak, restaurant with average noise level seems to obtain the higher rating. However, we need to think about the specific restaurant types: such as bars or a fine dinner.

3.2.11 Rating and Noise Level by state

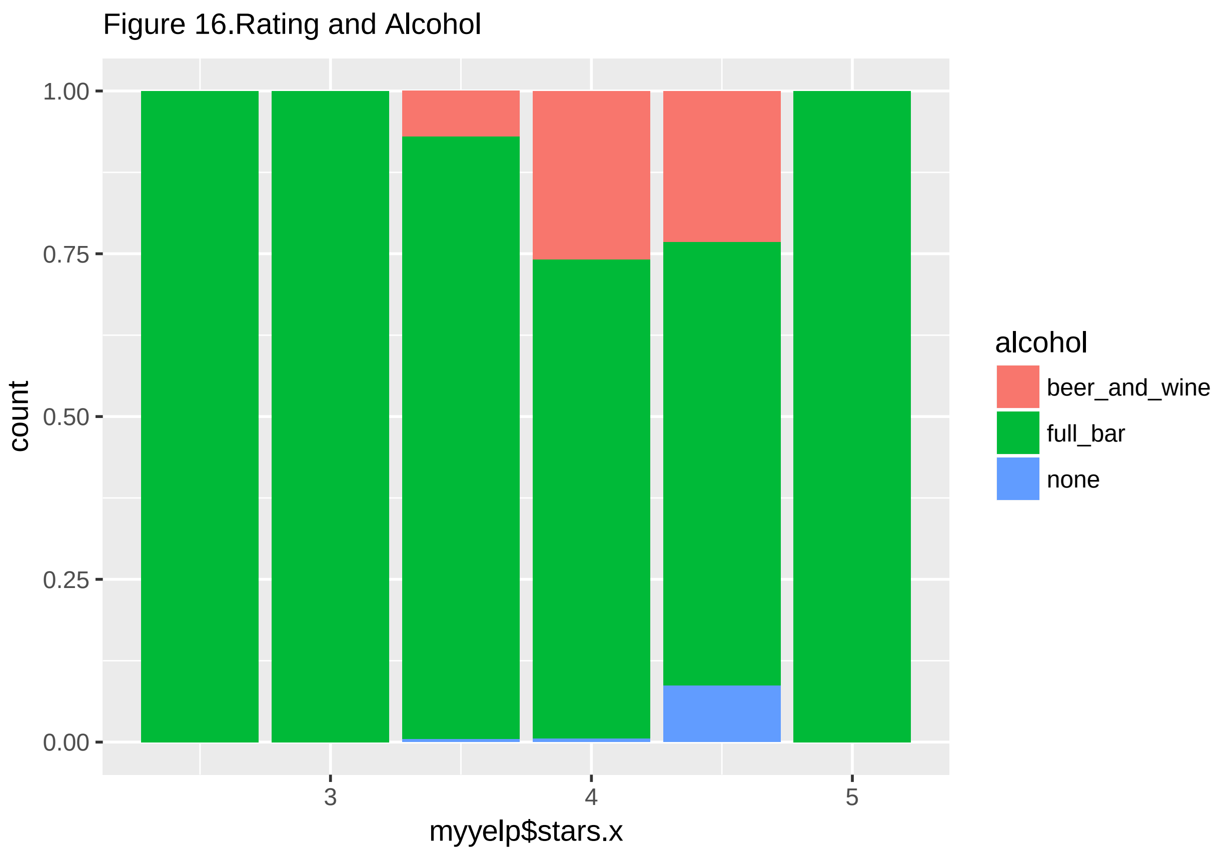
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Rating and noise level by state

Within the state, relatively to speak, restaurant with average noise level seems to obtain the higher rating. Still, we need to think about the specific restaurant types: such as bars or a fine dinner.

3.2.12 Rating and Alcohol



Rating and alcohol level

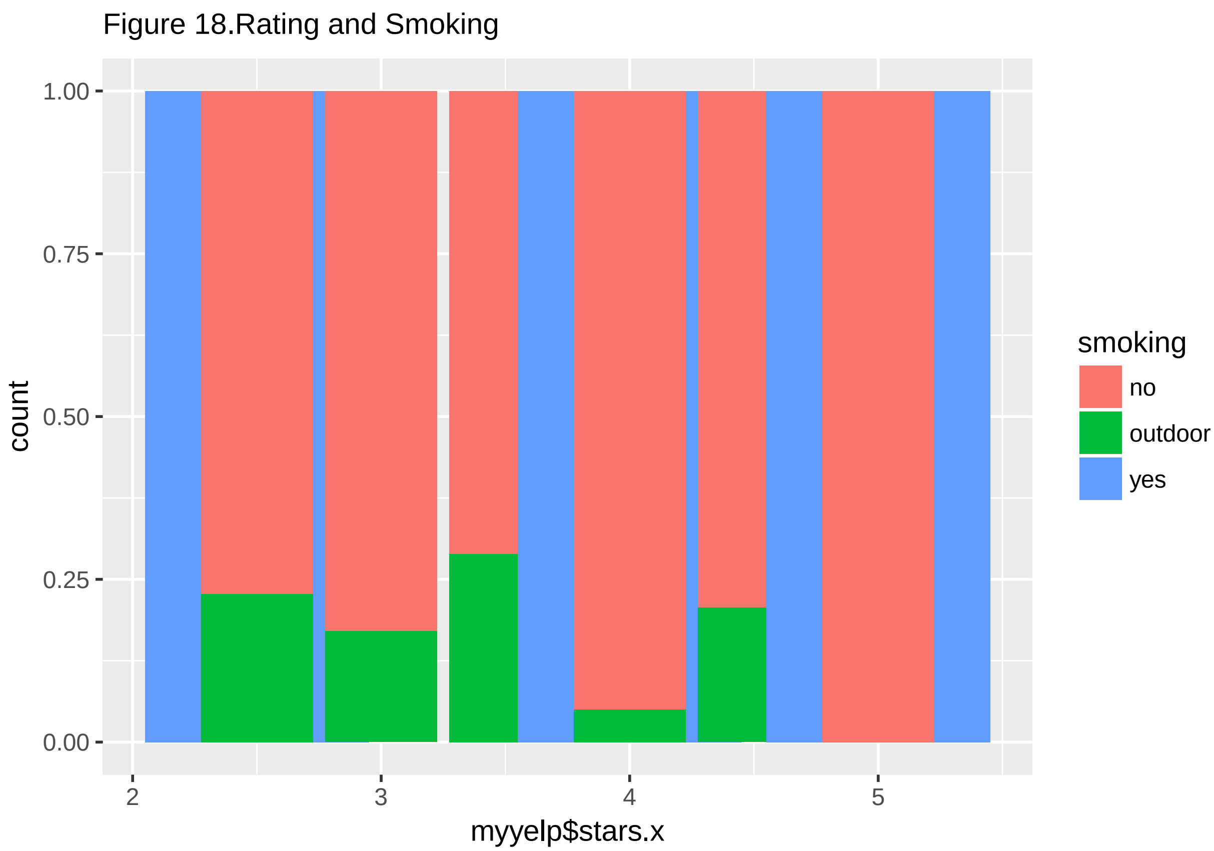
From the rating figure, it seems that full_bar is the most popular one among the yelp reviewers. However, does this character relates to the individual preference? Perhaps, the reviewers on the Yelp, happens to be the persons who enjoy the alcohol. For those that happen to enjoy the non-alcohol restaurant does not participate in the rating. Also, people with the religious practice such as Muslim, they do not drink alcohol. The figure may indicate that people that do not drink alcohol or do not like alcohol might not participate in the rating. Or perhaps, they are the person who relates to the “nono-alcohol” in the 4.5-star level rating.

3.2.12 Rating and Alcohol by state

Rating and alcohol by state

When observing each state, the change of the pattern of alcohol and rating indicates that state might influence this relationship. For instance, IL has the obvious "none-alcohol" restaurant with the relatively larger proportion of four stars comparing with "beer" and "full_bar" types. Other states such as QC, WI, BW seems to prefer the "full_bar" types. Therefore, the above discussed individual preference should be considered as well.

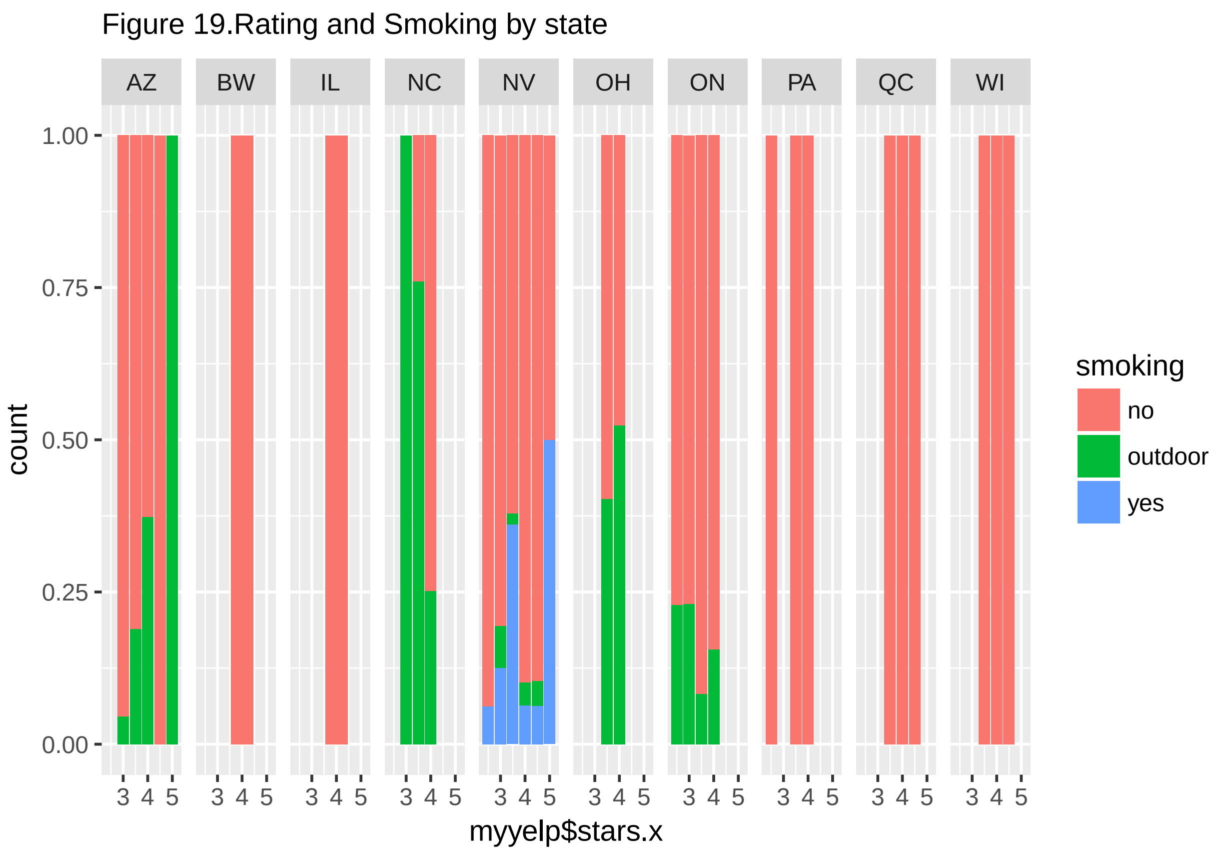
3.2.13 Rating and Smoking



Rating and smoking

Smoking seems to have the positive relationship with restaurants ranking: “No smoking”" restaurants have the higher rank than the rest.

3.2.13 Rating and Smoking by state



Rating and smoking by state

This phenomenon that "No smoking" restaurants have the higher rank than the rest is more evident within each state, except for AZ which prefer outdoor smoking, the rest states' restaurants have the apparent positive rating of no-smoking restaurants.

3.2 Summary of Proeject Development Stage

- 1. A set of restaurants' variables and their relations with the rating has been visualized
- 2. Price range is a common phenomenon and could not influence too much at the state level. Thus this variable has been removed for validation. The credit card has the similar issue.
- 3. EDA presents that "state" has the random effect that influences: smoking restaurant type, wifi, alcohol, smoking, outskating and noise level.
- 4. EDA presents that "individual" has the random effect that influences constructs such as alcohol, noise level.

IV. Validation: the statistical model selection

Why use multi-level model

As a random effects model assumes that the data being analyzed are drawn from a hierarchy of different populations whose differences relate to that hierarchy. In my Yelp project, the rating behavior is embedded in the state, geographically, and each identity, physically. Thus, I use the random effects model to shed light on the high-level context. Among the random effects model, I choose the random intercepts model, which is equivalent to assume that the correlation between two distance measurements (at state j,k such that j not equal to k) for the same subject (i) is constant no matter what the difference between state j and state k.

4.1. Model fitting (stepwise selection)

Model1 linear regression (full model)

$Stars_{[i]} = \beta_0+\beta_1 restaurant_style_i+\beta_2 wifi_i+\beta_3 tv_i+\beta_4 Parking_i+\beta_5 outseating_i+\beta_6 noiselevel_i+\beta_7 alcohol_i+ \beta_8 smoking_i + \beta_9 state_i + \epsilon_i$

Model2 linear regression within state effect

$Stars_{[i]} = \beta_0+\beta_1 restaurant_style_i+\beta_2 wifi_i+\beta_3 tv_i+\beta_4 Parking_i+\beta_5 outseating_i+\beta_6 noiselevel_i+\beta_7 alcohol_i+ \beta_8 smoking_i + \beta_9 State_{j[i]} + \epsilon_i$

Model3: consider the random effect within the state and user effect

$Stars_{[i]} = \beta_0+\beta_1 restaurant_style_i+\beta_2 wifi_i+\beta_3 tv_i+\beta_4 Parking_i+\beta_5 outseating_i+\beta_6 noiselevel_i+\beta_7 alcohol_i+ \beta_8 smoking_i + \beta_9 State_{j[i]} + \beta_1 0user_{j[i]} + \epsilon_i$

Model4 the random effect within the state and user effect, and their interaction effect

$Stars_{[i]} = \beta_0+\beta_1 restaurant_style_i+\beta_2 wifi_i+\beta_3 tv_i+\beta_4 Parking_i+\beta_5 outseating_i+\beta_6 noiselevel_i+\beta_7 alcohol_i+ \beta_8 smoking_i + \beta_9 State_{j[i]} + \beta_1 0user : State_{j[i]} + \epsilon_i$

4.2. Model Checking

Figure4.1.1ResidualPlot(model1)

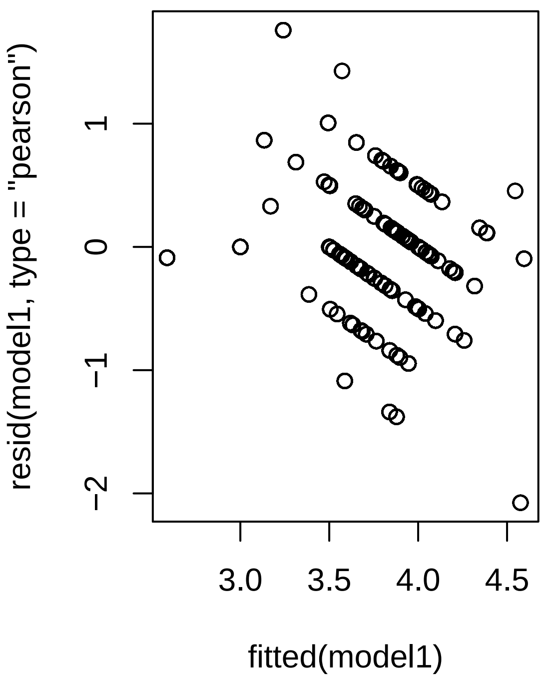


Figure 4.1.1 NormalQQ−Plot(model1)

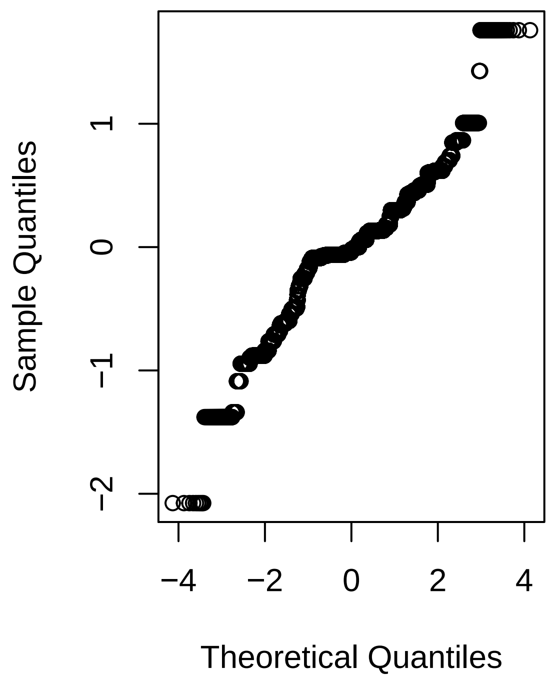


Figure4.2. Binned residual plot(model2)

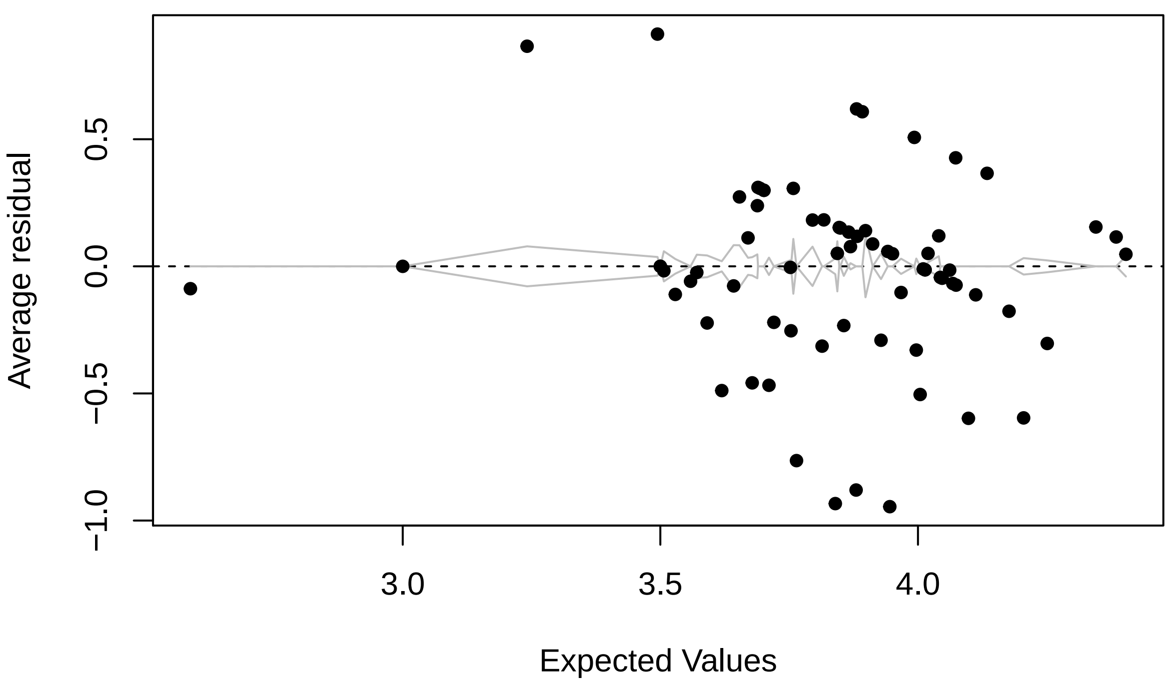


Figure4.3. Binned residual plot(model3)

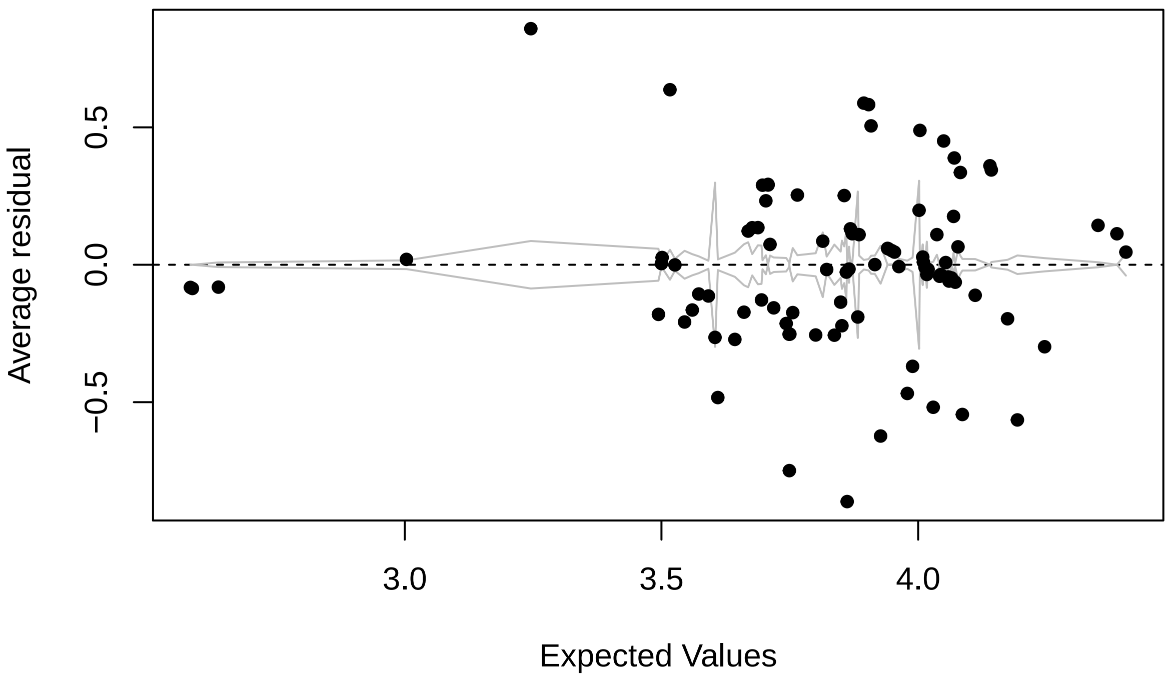


Figure4.4. Binned residual plot(model4)

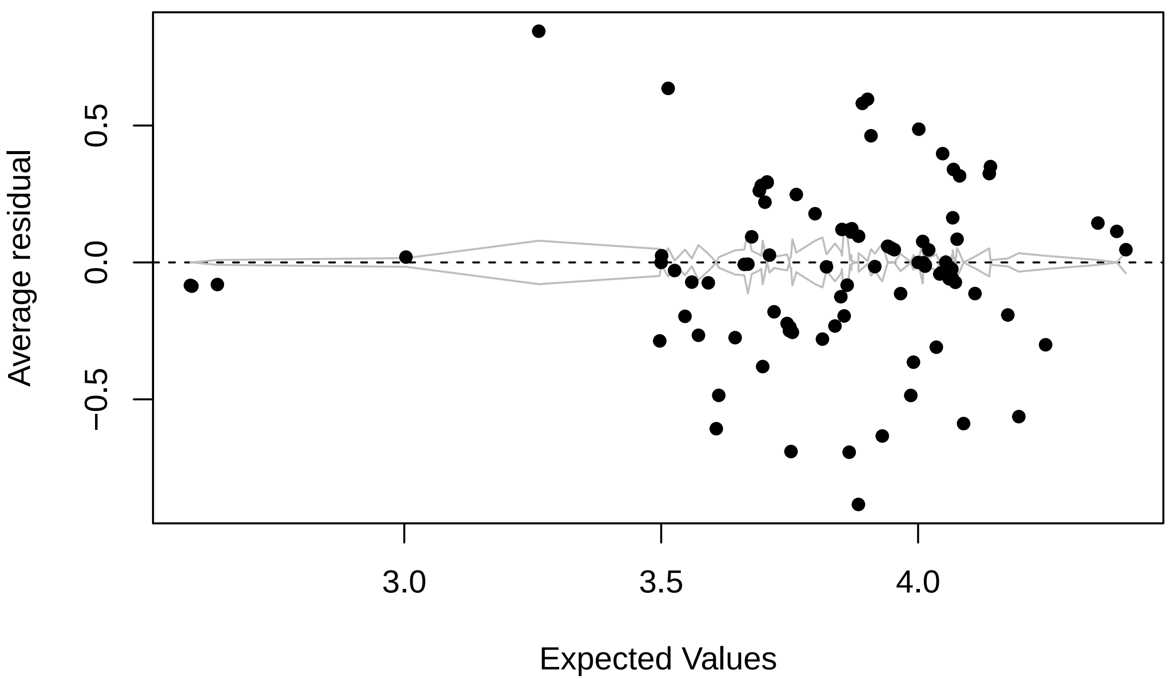


Figure4.1.1 to figure4.4 capture the residual for each model. Basically, for model 2, model 3 and model 4 the residual plots look relatively good since most residuals are within -0.5,0.5 range and the plots are symmetrical. The model 3 with the state intercept variance alone seems to be better than model 2 and model 4, since its fluctuation is not that much and do not display a kind of "scatter" below -0.5 line: that is to say, except several point residuals "fan out" from left to right the most are spread around the residual = 0 line on the right side. Thus, in regards to the residuals plot, model 3 seems to have the relatively good fitness.

4.3. Compare the coefficients

Table 4.3 Coefficients Table

	model2.beta	model3.beta	model4.beta
Intercept	3.5970416	3.5972926	3.5972699
restaurant_Indian	0.0575658	0.0572556	0.0573141
restaurant_Japanese	0.2761570	0.2757332	0.2758281
restaurant_Korean	0.1322884	0.1323274	0.1323178
restaurant_Thai	0.0868604	0.0867942	0.0868075
restaurant_Vietnamese	0.0541036	0.0534094	0.0535571
no-wifi	-0.0037106	-0.0039260	-0.0038850
have tv	0.0505797	0.0504360	0.0505021
parking_valet	0.1747290	0.1749431	0.1748729
parking_lot	0.3164394	0.3162885	0.3163120
parking_lot and valet	-0.3446250	-0.3440786	-0.3442878
parking_street	0.1850883	0.1852999	0.1852130
parking_street and valet	-0.3902287	-0.3898190	-0.3899176
parking_street and lot	0.5006465	0.5008603	0.5007964
parking_garage	0.1994762	0.1996161	0.1995419
parking_garage and valet	-0.0293060	-0.0290588	-0.0290649
parking_garage and lot and valet	-0.2767101	-0.2764860	-0.2765357
parking_garage and street	0.1619776	0.1623732	0.1622392
parking_garage and street and valet	0.4709990	0.4711306	0.4711410
outseating_yes	0.0003354	0.0000432	0.0001274
average noise	-0.0462588	-0.0461594	-0.0461479
loud	0.0031800	0.0033709	0.0033627
very_loud	-0.1923217	-0.1923125	-0.1922844
alcoholfull_bar	-0.0396753	-0.0396545	-0.0396895
alcohol_none	0.4051769	0.4048282	0.4048994
smoking_outdoor	-0.1206728	-0.1202973	-0.1203892
smoking_yes	-0.8552621	-0.8553812	-0.8553771

Comparing the coefficients of each model in the coefficients' table, we could find:

For restaurant style, comparing with Chinese restaurant (baseline), Japanese restaurant has a 0.28 stars rating higher than Chinese

For Wifi variable, the models show no-wifi decreases 0.004 of stars compared to restaurants with free WiFi.

For TV variable, the regression models show that "with TV" increases 0.05 of stars compared to restaurants without TV.

For variables of parking, restaurants who have valet, street parking, parking lot, garage and street parking, street and lot parking,garage and street and valet parking tend to have higher ratings than without these conditions especially, while the restaurants with lot and valet parking, garage and valet parking, garage and lot and valet parking tend to have lower ratings than restaurants without these parking conditions

For "with out-seating" variable, the three model consistently shows that it has very small influence towards restaurant rating.

For the "noise level", the baseline is quiet level, the coefficients of the models present that some kind of louder noise level could lead to having higher rating comparing with quiet level restaurant;

however, the very loud noise has an apparent negative influence towards the restaurant rating: the very-loud noise level decreases 0.19 of rating.

For alcohol variable with baseline “beer_and_wine”, restaurants with bar leads to the decrease of 0.04 point of rating comparing with beer_and_wine; non-alcohol could increase 0.41 point of rating comparing with beer_and_wine.

For variable smoking with baseline “no-smoking”, both outdoor and “could smoking” leads to the decrease of rating, comparing with “no-smoking”. Thus, smoking has a negative influence on restaurant rating.

4.4. Confidence interval: hypothesis testing

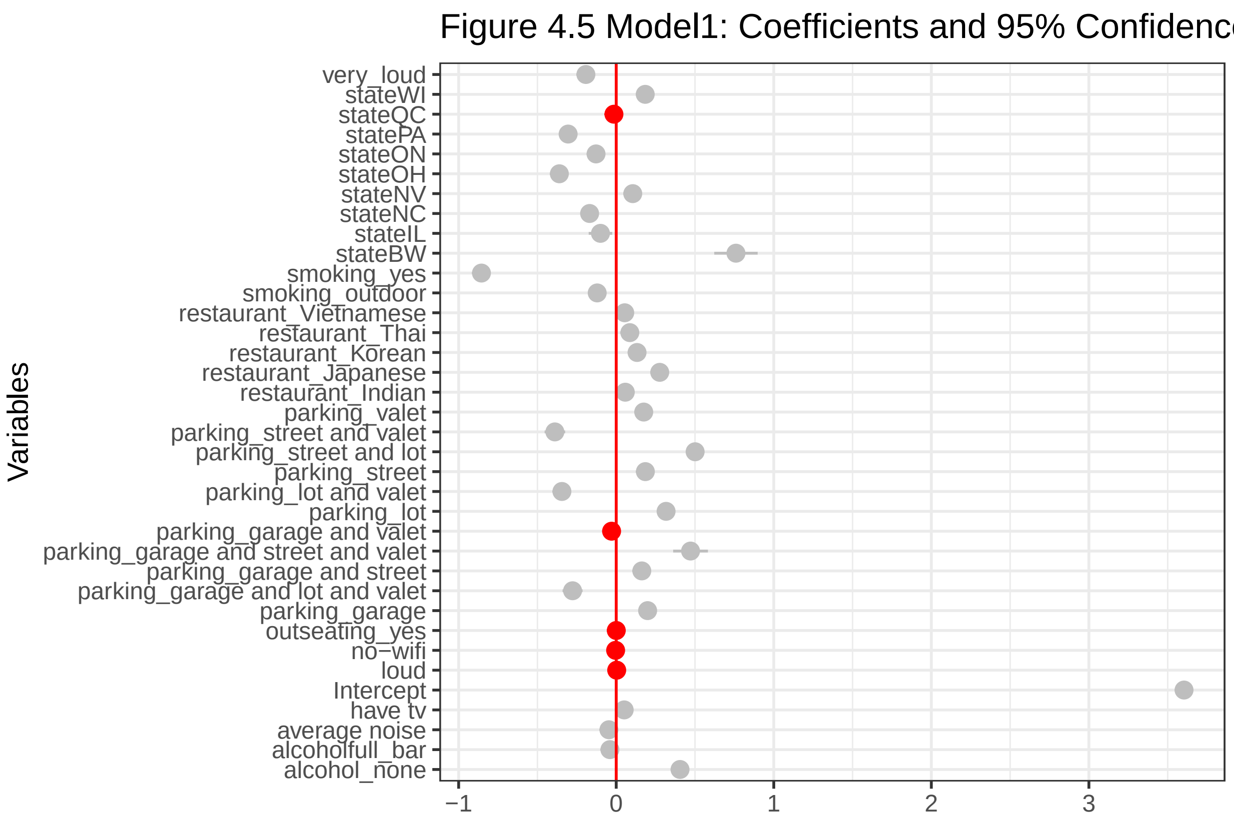


Fig 4.5 :The confidence interval of the combined effect size in Figure 4.5 include zero, i.e. in case of a confidence level of 95% the p-value is smaller than .05. In traditional terminology, this means that the meta-analytic effect is not statistically significant. Specifically, from figure4.5 Model1 presents the 95% confidence interval each coefficients, and whether it includes the point zero where the vertical red line is the line of y = 0. Visually, “parking_garage and valet”, “outseating_yes”, “no-wifi” and “loud” might not significant since its confidence interval includes 0 but rest of the variables are significant at level 5%.

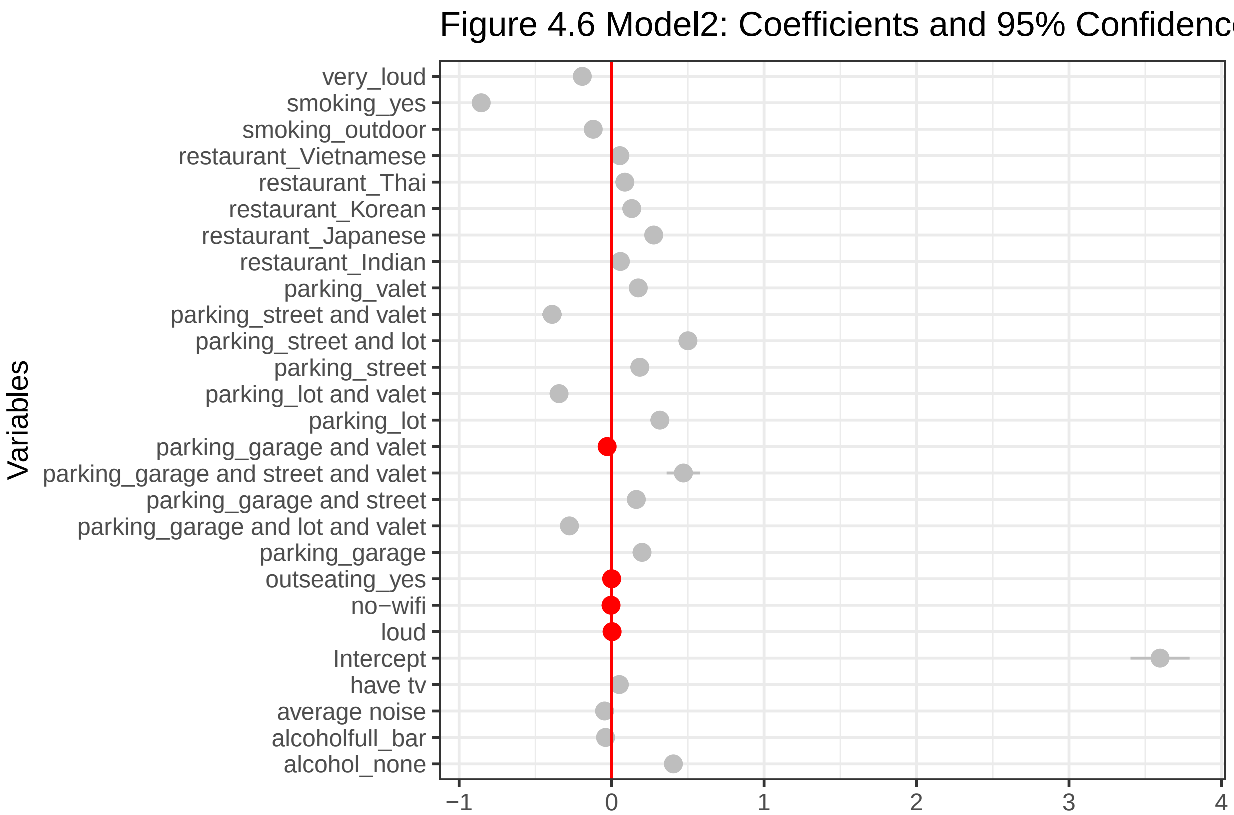


Fig 4.6: The confidence interval of the combined effect size in Figure 4.6 include zero, i.e. in case of a confidence level of 95% the p-value is smaller than .05. Specifically, from figure 4.6 Model2 presents the 95% confidence interval each coefficients, and whether it includes the point zero where the vertical red line is the line of y = 0.Visually, Model 2 and Model 1 has the same result: "parking_garage and valet" "outseating_yes", "no-wifi"and "loud" might not significant since its confidence interval includes 0 but rest of the variables are significant at level 5%.

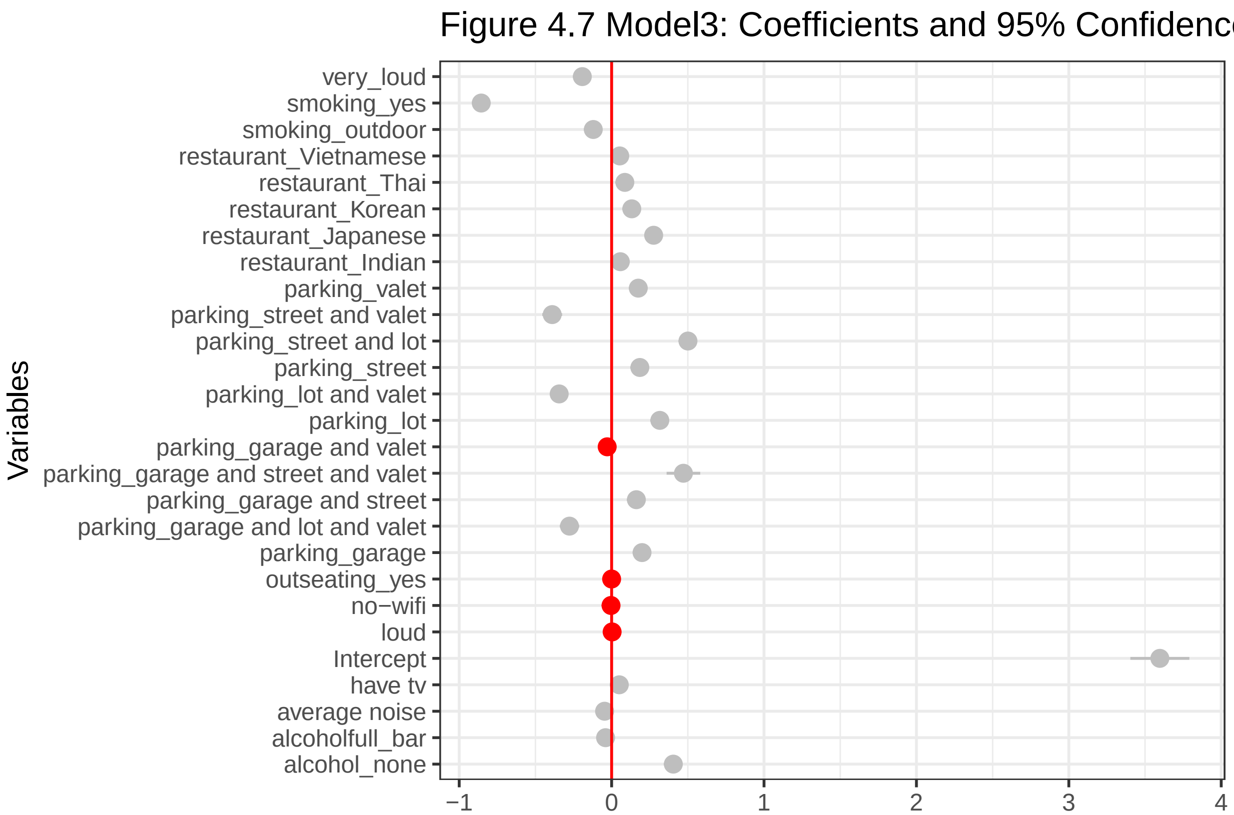


Fig 4.7: The confidence interval of the combined effect size in Figure 4.6 include zero, i.e., in case of a confidence level of 95% the p-value is smaller than .05. Specifically, from figure4.6 Model2 presents the 95% confidence interval each coefficients, and whether it includes the point zero where the vertical red line is the line of y = 0.Visually, Model 3 and Model 2 and Model 1 have the same result: "parking_garage and valet", "outseating_yes", "no-wifi" and "loud" might not significant since its confidence interval includes 0 but rest of the variables are significant at level 5%.

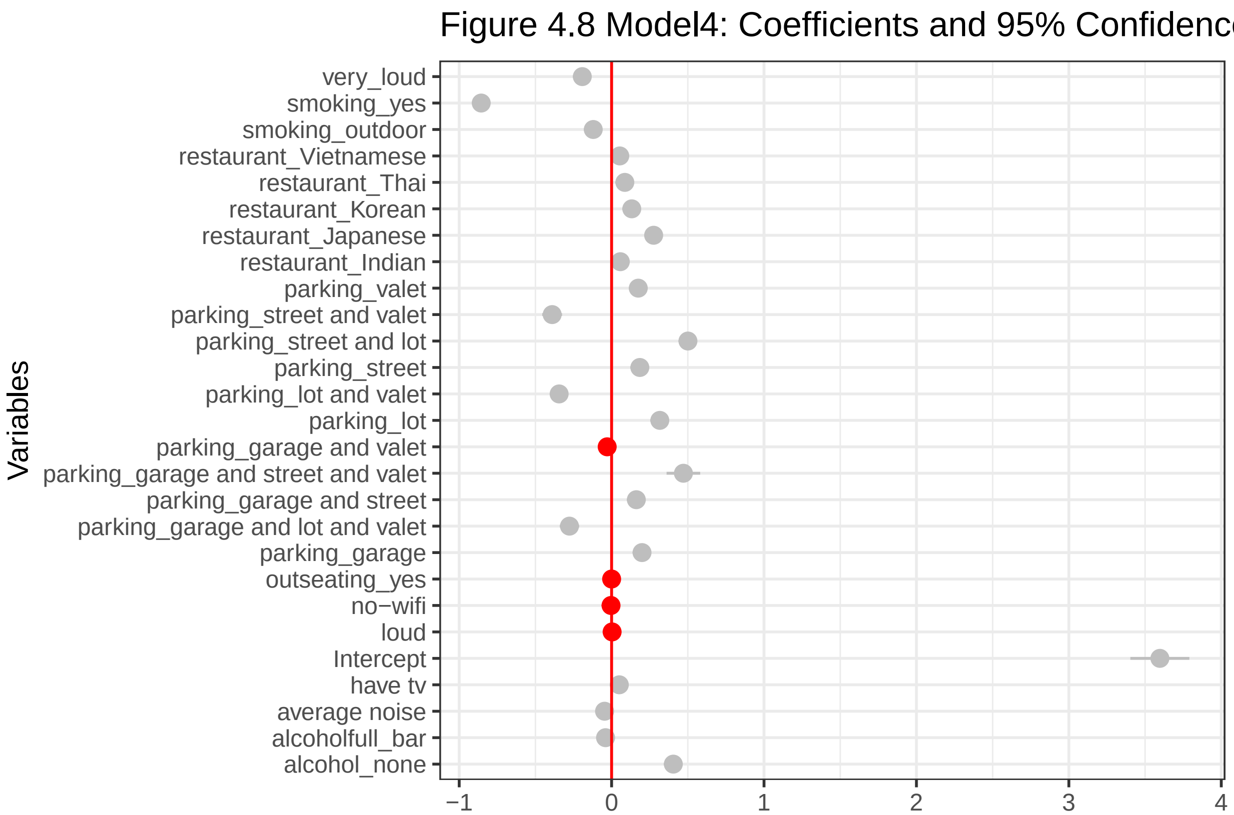


Fig 4.8: Visually, Model4 alongside Model3 and Model2 and Model1 have the same result "parking_garage and valet" "outseating_yes", "no-wifi" and "loud" might not significant since its confidence interval includes 0 but rest of the variables are significant at level 5%.

4.5 Abridged Models

Based on the multi-level regression and the CI hypothesis testing, I decide to remove the "parking", "with wifi" and "outseating" variables and form abridged models as follows.

Model 5: linear regression with fixed effect

$$Stars_{[i]} = \beta_0 + \beta_1 restaurant_s tyle_i + \beta_2 tv_i + \beta_3 noiselevel_i + \beta_4 alcohol_i + \beta_5 smoking_i + \beta_6 State_i + \epsilon_i$$

Model 6: Random effect within state

$$Stars_{[i]} = \beta_0 + \beta_1 restaurant_s tyle_i + \beta_2 tv_i + \beta_3 noiselevel_i + \beta_4 alcohol_i + \beta_5 smoking_i + \beta_6 State_{j[i]} + \epsilon_i$$

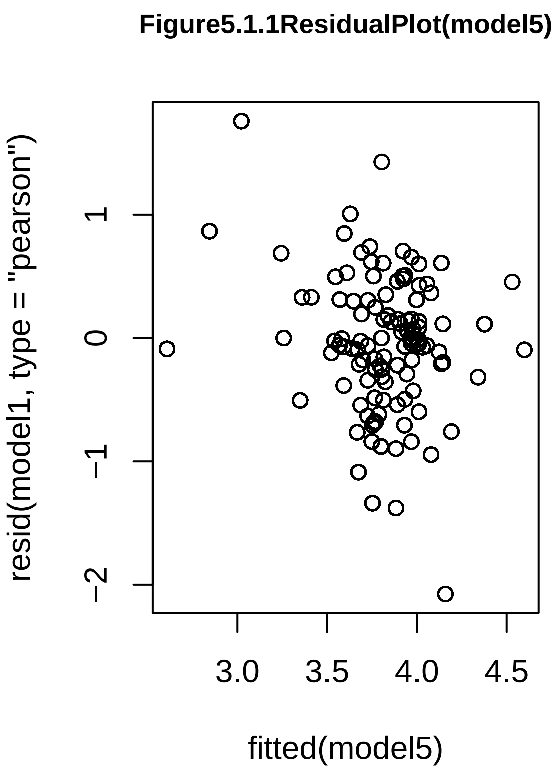
Model7: The random effect within the state and the user

$$Stars_{[i]} = \beta_0 + \beta_1 restaurant_s tyle_i + \beta_2 tv_i + \beta_3 noiselevel_i + \beta_4 alcohol_i + \beta_5 smoking_i + \beta_6 State_{j[i]} + \beta_7 user_{j[i]} + \epsilon_i$$

Model8 The random effect within the state,the user and user and state interaction effect

$Stars_{[i]} = \beta_0 + \beta_1 restaurant_s tyle_i + \beta_2 tv_i + \beta_3 noiselevel_i + \beta_4 alcohol_i + \beta_5 smoking_i + \beta_6 State_{j[i]} + \beta_7 user_{j[i]} + \beta_8 business_{j[i]} + \beta_9 user : State_{j[i]} + \epsilon_i$

4.6 Abrridged Models Checking



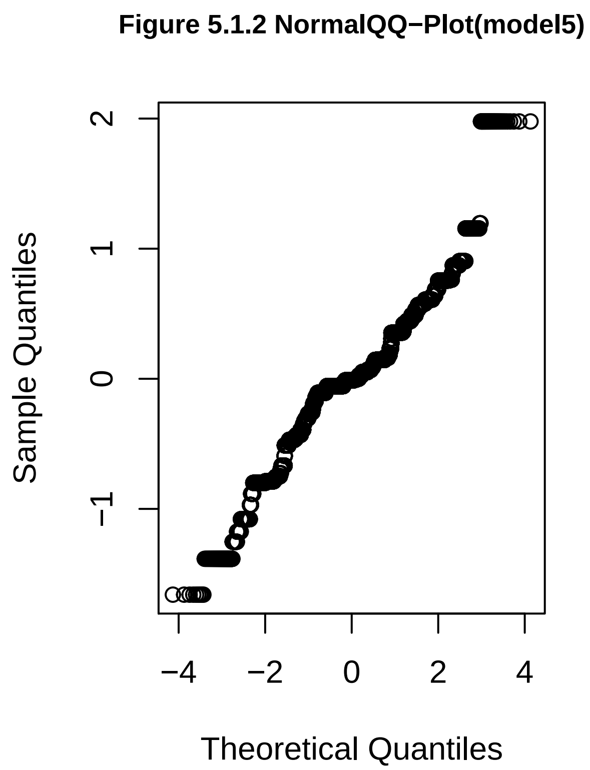


Figure. Binned residual plot(model6)

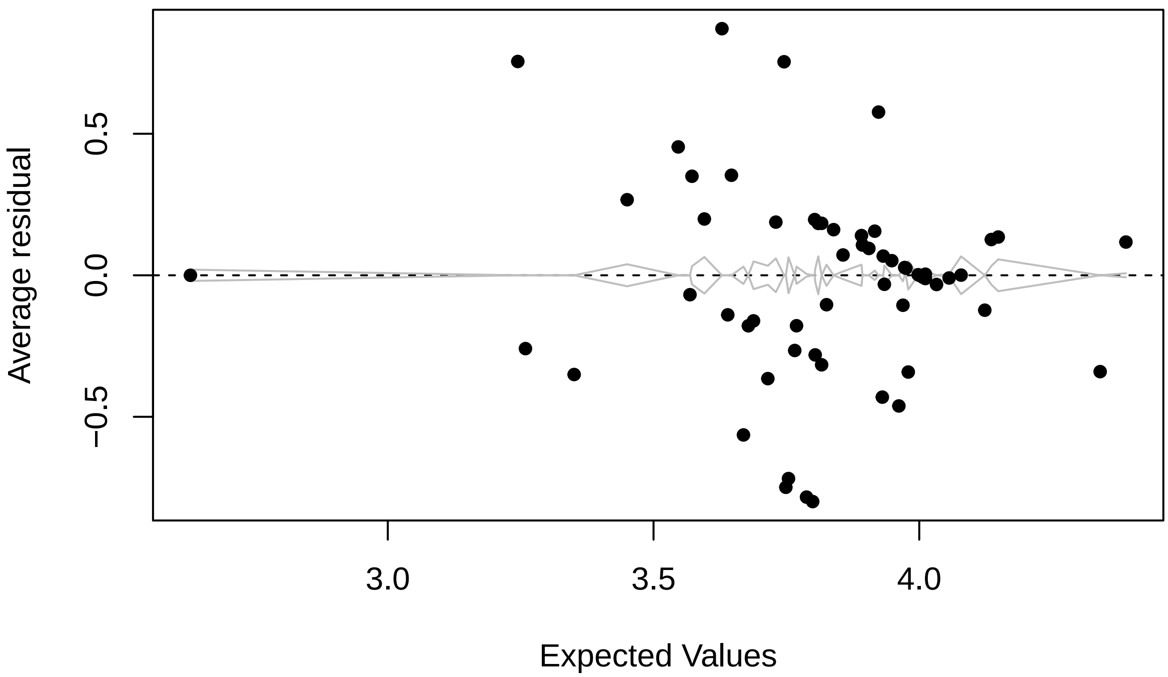


Figure. Binned residual plot(model7)

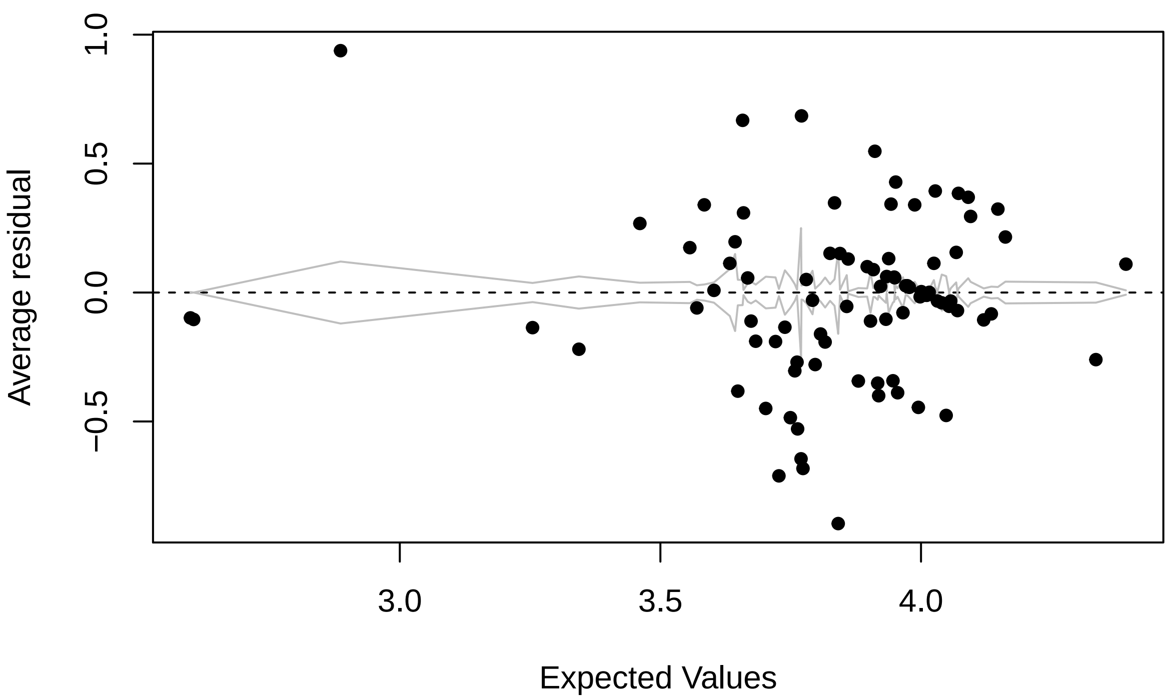
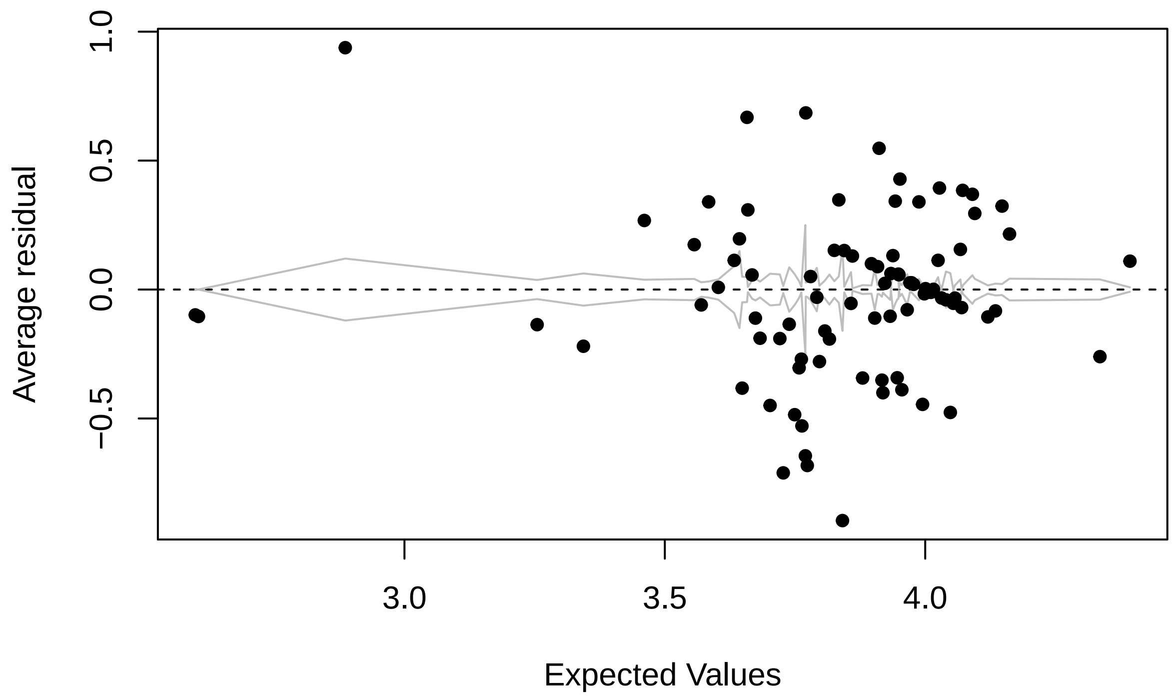


Figure. Binned residual plot(model8)



4.7. Compare the coefficients table

	model6.beta	model7.beta	model8.beta
Intercept	3.8243841	3.8246808	3.8246810
restaurant_Indian	0.1454853	0.1447449	0.1447449
restaurant_Japanese	0.2237291	0.2232393	0.2232393
restaurant_Korean	0.1443389	0.1439823	0.1439823
restaurant_Thai	0.0932224	0.0928648	0.0928648
restaurant_Vietnamese	0.1564485	0.1551580	0.1551580
have tv	0.0671549	0.0671321	0.0671321
average noise	0.0388436	0.0388802	0.0388802
loud	0.0706400	0.0709093	0.0709093
very_loud	-0.1309965	-0.1315412	-0.1315412
alcoholfull_bar	-0.1751025	-0.1750035	-0.1750035
alcohol_none	0.3099824	0.3095156	0.3095156
smoking_outdoor	-0.0780282	-0.0779002	-0.0779002
smoking_yes	-1.0102100	-1.0096379	-1.0096379

Comparing the coefficients of each abridged model in the coefficients' table, we could find:

For restaurant style, comparing with Chinese restaurant (baseline), Japanese restaurant has a 0.23 stars rating higher than Chinese

For TV variable, the regression models show that "with TV" increases 0.07 of stars compared to restaurants without TV; slightly larger comparing with former coefficients table

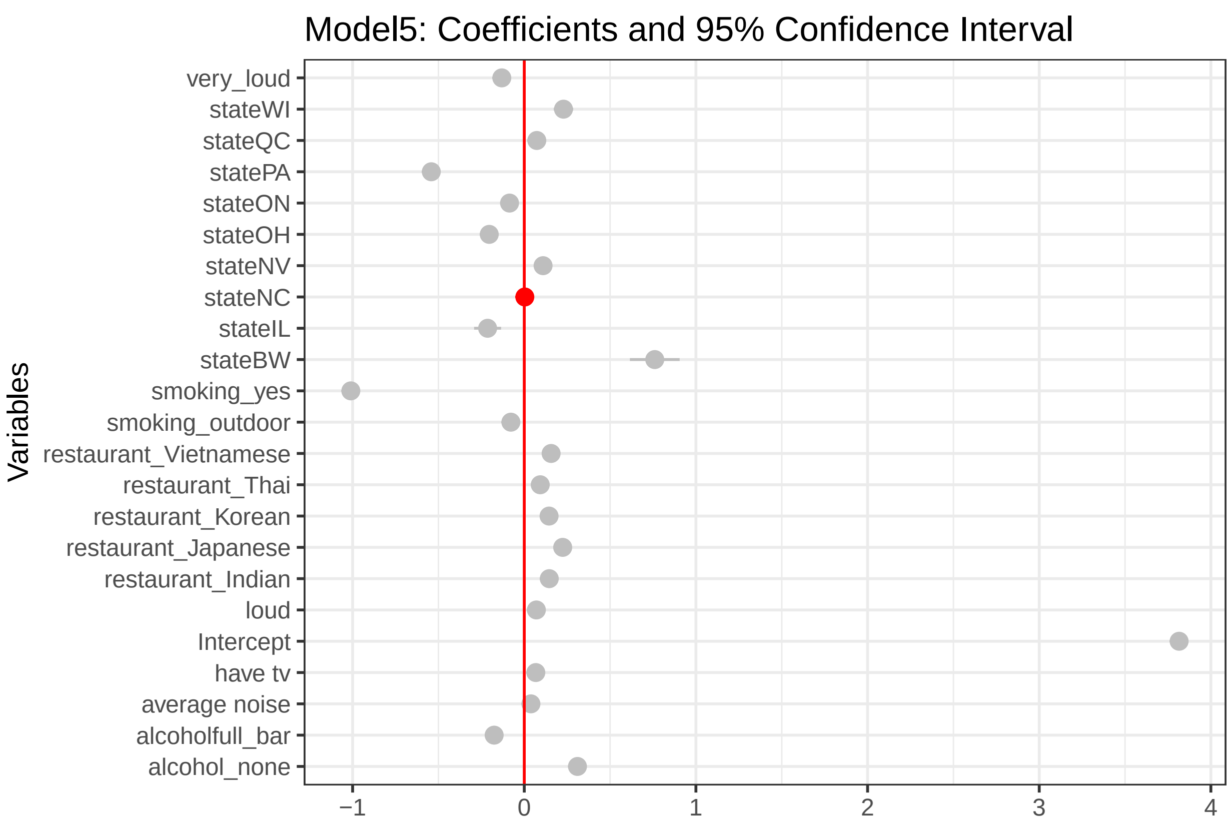
For the "noise level", the baseline is quiet level, the coefficients of the models present that some kind of louder noise level could lead to having higher rating comparing with quiet level restaurant;

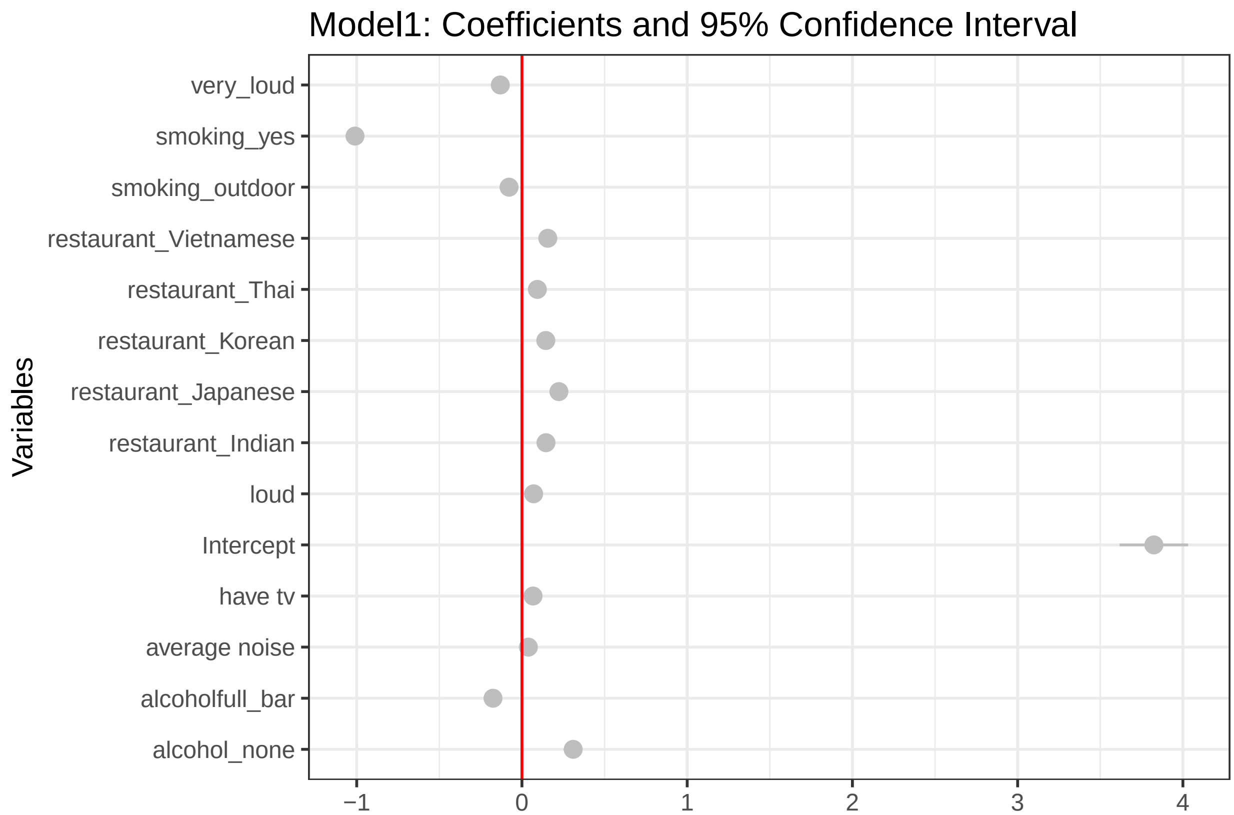
however, the very loud noise has an apparent negative influence towards the restaurant rating: the very-loud noise level decreases 0.13 of rating; slightly smaller comparing un-abridged models.

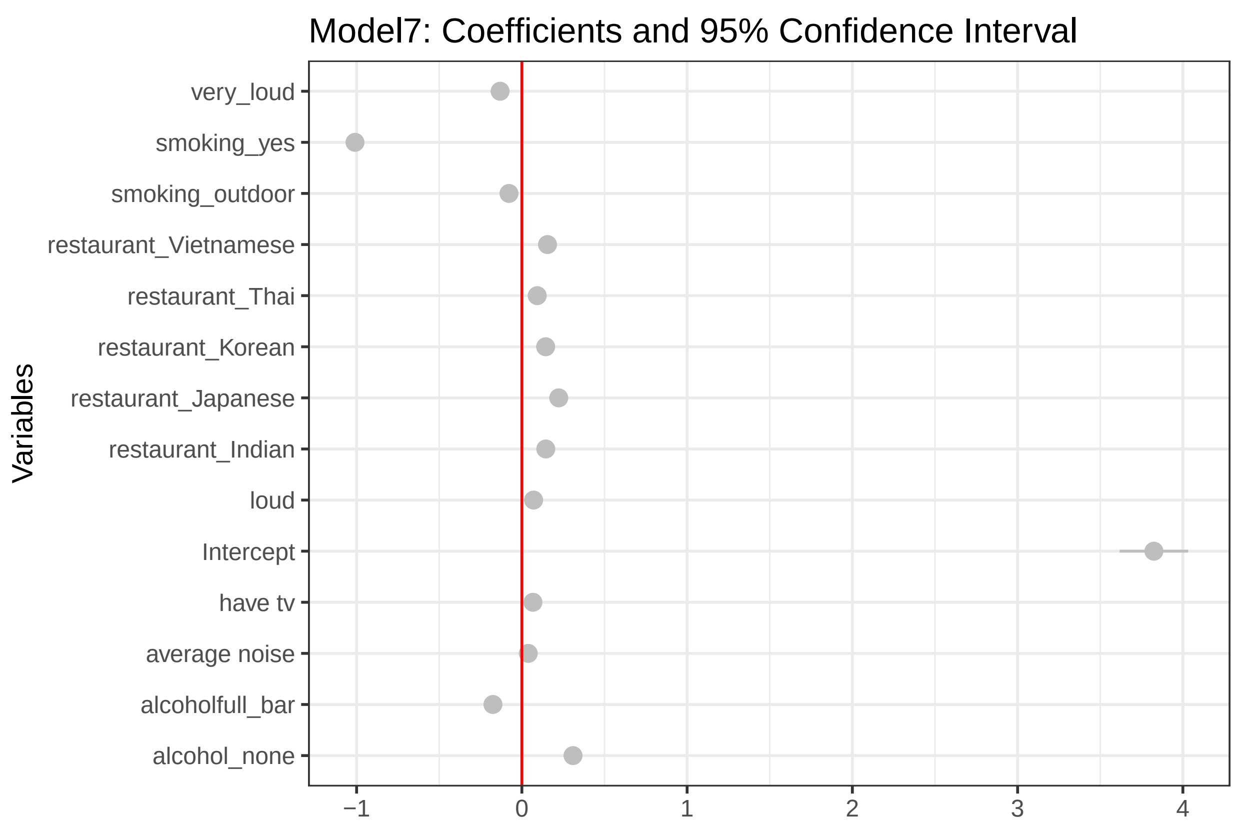
For alcohol variable with baseline “beer_and_wine”, restaurants with bar leads to the decrease of 0.18 point of rating comparing with beer_and_wine. This is larger than previous results. In addition, non-alcohol could increase 0.3 point of rating comparing with beer_and_wine; similar to previous one.

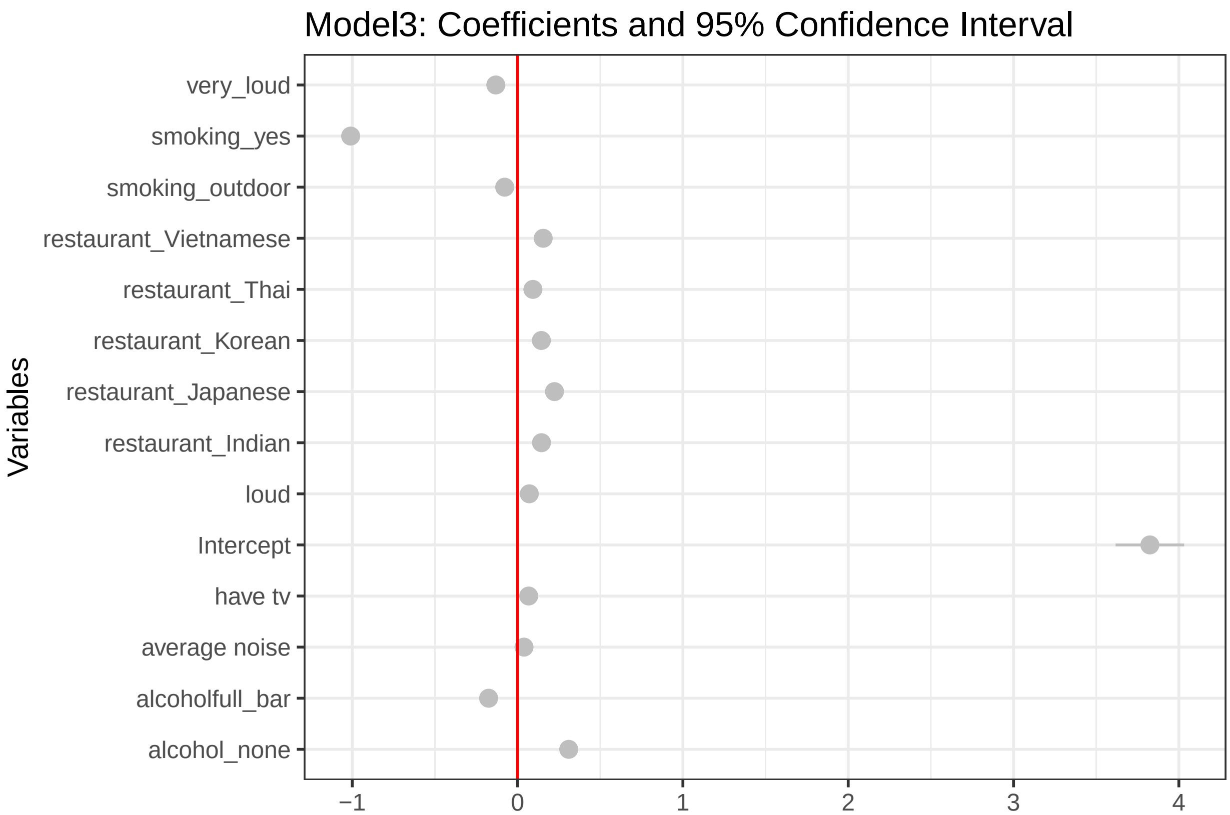
For variable smoking with baseline “no-smoking”, both outdoor and “could smoking” leads to the decrease of rating, comparing with “no-smoking”. Thus, smoking has a negative influence on restaurant rating. This is same to the previous observation.

4.8. Check the significance of coefficients




Model1: Coefficients and 95% Confidence Interval
Variables
very_loud
smoking_yes
smoking_outdoor
restaurant_Vietnamese
restaurant_Thai
restaurant_Korean
restaurant_Japanese
restaurant_Indian
loud
Intercept
have tv
average noise
alcoholfull_bar
alcohol_none
−1
0
1
2
3
4


Model7: Coefficients and 95% Confidence Interval
Variables
very_loud
smoking_yes
smoking_outdoor
restaurant_Vietnamese
restaurant_Thai
restaurant_Korean
restaurant_Japanese
restaurant_Indian
loud
Intercept
have tv
average noise
alcoholfull_bar
alcohol_none
−1
0
1
2
3
4



4.9 Abridged Model Hypothesis Testing Summary

The abridged models above shows that except model5's "state NC" variable, the rest are all sigficant at 95% CI level.

4.10 ANOVA Analysis

```
## refitting model(s) with ML (instead of REML)
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
model6	16	21074.94	21206.77	-10521.47	21042.94	NA	NA	NA
model7	17	21071.56	21211.63	-10518.78	21037.56	5.373686	1	0.0204427
model2	29	17727.64	17966.58	-8834.82	17669.64	3367.922181	12	0.0000000

```
## refitting model(s) with ML (instead of REML)
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
model6	16	21074.94	21206.77	-10521.468	21042.94	NA	NA	NA
model7	17	21071.56	21211.63	-10518.781	21037.56	5.373686	1	0.0204427
model3	30	17727.51	17974.69	-8833.753	17667.51	3370.056971	13	0.0000000

```
## refitting model(s) with ML (instead of REML)
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
model6	16	21074.94	21206.77	-10521.468	21042.94	NA	NA	NA
model2	29	17727.64	17966.58	-8834.820	17669.64	3373.29587	13	0.0000000
model3	30	17727.51	17974.69	-8833.753	17667.51	2.13479	1	0.1439902

```
## refitting model(s) with ML (instead of REML)
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
model7	17	21071.56	21211.63	-10518.781	21037.56	NA	NA	NA
model2	29	17727.64	17966.58	-8834.820	17669.64	3367.92218	12	0.0000000
model3	30	17727.51	17974.69	-8833.753	17667.51	2.13479	1	0.1439902

The anova analysis presents that relatively Model 2 and Model 3 are proper here in regards with the cross validation, especially for Model2

4.11 Validation summary

In the validation and test of the my development has addressed the following issues:

- The stepwise process of full model1 to the multi-level effect model4
- The hypothesis testing of the coefficients of model1, model2, model3 and model4 respectively
- The abridged full model 5 and the stepwise selection of the model 6, model 7 and model8
- ANOVA analysis of certain models (see appedix for further information)

V. Discussion

5.1 General Findings

- For restaurant style, comparing with Chinese restaurant as baseline, Japanese restaurant has a stars rating higher than Chinese
- For TV variable, the regression models show that "with TV" increases the stars compared to restaurants without TV.
- For "with out-seating" variable, the three model, consistently shows that it has the very small influence towards restaurant rating.
- For the "noise level", the baseline is quiet level, the coefficients of the models present that some kind of louder noise level could lead to having higher rating comparing with quiet level restaurant; however, the very loud noise has an apparent negative influence towards the restaurant rating: the very-loud noise level decreases the rating.
- For alcohol variable with baseline "beer_and_wine", restaurants with bar leads to the decreased rating comparing with beer_and_wine; non-alcohol could increase rating comparing with beer_and_wine.
- For variable smoking with baseline "no-smoking",both outdoor and "could smoking" leads to the decrease of rating, comparing with "no-smoking". Smoking has a negative influence on restaurant rating.
- Different states have influnece to the

5.2 Further Discussion

- The relationship between the restaurant's features and the consumer's rating behavior on the Yelp platform is a complex issue. I provide the prototype of the rating behavior and the restaurants via the multi-level models. I regard the state influence, and the individual random effect. However, the Yelp online platform per se also have the certain influence towards behavior: such as the existing rating influences to the successive rating?
- Also, for the individual variance, each consumer has the demographic features that influence the rating. Such as the religious and the restaurant that provided alcohol attitude.
- How important and significant is personal income related to business traits such as location, hours, business categories?
- How much is spatial environment influence the restaurant reviews, such as location, residents' income, schools, etc.? Namely, is there any third variable (i.e., ethnicity, gender, education, income) that influence the consumption and the review?
- When controlling the confounding variables, how large are location influences the review traits based on a multivariate analysis? These specific statistics questions influnced the evaluation of the rating and Yelp users.

Appendix

Model2

```
## lmer(formula = stars.x ~ restaurant_style + wifi + tv + parking +
##     outseating + noiselevel + alcohol + smoking + (1 | state),
##     data = myyelp)
##                                                                                              coef.est
## (Intercept)                                                                                   3.60
## restaurant_styleIndian                                                                        0.06
## restaurant_styleJapanese                                                                      0.28
## restaurant_styleKorean                                                                        0.13
## restaurant_styleThai                                                                          0.09
## restaurant_styleVietnamese                                                                    0.05
## wifino                                                                                        0.00
## tvyes                                                                                         0.05
## parking{"garage": false, "street": false, "validated": false, "lot": false, "valet": true}  0.17
## parking{"garage": false, "street": false, "validated": false, "lot": true, "valet": false}  0.32
## parking{"garage": false, "street": false, "validated": false, "lot": true, "valet": true}  -0.34
## parking{"garage": false, "street": true, "validated": false, "lot": false, "valet": false}  0.19
## parking{"garage": false, "street": true, "validated": false, "lot": false, "valet": true}  -0.39
## parking{"garage": false, "street": true, "validated": false, "lot": true, "valet": false}   0.50
## parking{"garage": true, "street": false, "validated": false, "lot": false, "valet": false}  0.20
## parking{"garage": true, "street": false, "validated": false, "lot": false, "valet": true}  -0.03
## parking{"garage": true, "street": false, "validated": false, "lot": true, "valet": true}   -0.28
## parking{"garage": true, "street": true, "validated": false, "lot": false, "valet": false}   0.16
## parking{"garage": true, "street": true, "validated": false, "lot": false, "valet": true}    0.47
## outseatingyes                                                                                 0.00
## noiselevel2 average                                                                          -0.05
## noiselevel3 loud                                                                              0.00
## noiselevel4 very_loud                                                                        -0.19
## alcoholfull_bar                                                                              -0.04
## alcoholnone                                                                                   0.41
```

```
## smokingoutdoor                                                                          -0.12
## smokingyes                                                                              -0.86
##                                                                                         coef.se
## (Intercept)                                                                              0.10
## restaurant_styleIndian                                                                   0.01
## restaurant_styleJapanese                                                                 0.01
## restaurant_styleKorean                                                                   0.01
## restaurant_styleThai                                                                     0.01
## restaurant_styleVietnamese                                                               0.01
## wifino                                                                                   0.01
## tvyes                                                                                    0.01
## parking{"garage": false, "street": false, "validated": false, "lot": false, "valet": true}  0.02
## parking{"garage": false, "street": false, "validated": false, "lot": true, "valet": false}  0.01
## parking{"garage": false, "street": false, "validated": false, "lot": true, "valet": true}   0.02
## parking{"garage": false, "street": true, "validated": false, "lot": false, "valet": false}  0.01
## parking{"garage": false, "street": true, "validated": false, "lot": false, "valet": true}   0.03
## parking{"garage": false, "street": true, "validated": false, "lot": true, "valet": false}   0.02
## parking{"garage": true, "street": false, "validated": false, "lot": false, "valet": false}  0.01
## parking{"garage": true, "street": false, "validated": false, "lot": false, "valet": true}   0.02
## parking{"garage": true, "street": false, "validated": false, "lot": true, "valet": true}    0.03
## parking{"garage": true, "street": true, "validated": false, "lot": false, "valet": false}   0.02
## parking{"garage": true, "street": true, "validated": false, "lot": false, "valet": true}    0.06
## outseatingyes                                                                            0.01
## noiselevel2 average                                                                      0.01
## noiselevel3 loud                                                                         0.02
## noiselevel4 very_loud                                                                    0.02
## alcoholfull_bar                                                                          0.01
## alcoholnone                                                                              0.02
## smokingoutdoor                                                                           0.01
## smokingyes                                                                               0.02
##
## Error terms:
##  Groups   Name        Std.Dev.
##  state    (Intercept) 0.31
##  Residual             0.33
## ---
## number of obs: 27986, groups: state, 10
## AIC = 17918.4, DIC = 17478.9
## deviance = 17669.6
```

ANOVA (model2,model6,model7)

```
## Data: myyelp
## Models:
## model6: stars.x ~ restaurant_style + tv + noiselevel + alcohol + smoking +
## model6:     (1 | state)
## model7: myyelp$stars.x ~ restaurant_style + tv + noiselevel + alcohol +
## model7:     smoking + (1 | state) + (1 | user_id)
## model2: stars.x ~ restaurant_style + wifi + tv + parking + outseating +
## model2:     noiselevel + alcohol + smoking + (1 | state)
##        Df   AIC   BIC   logLik deviance     Chisq Chi Df Pr(>Chisq)
## model6 16 21075 21207 -10521.5    21043
## model7 17 21072 21212 -10518.8    21038    5.3737      1    0.02044 *
```

```
## model2 29 17728 17967  -8834.8    17670 3367.9222     12    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA (model3,model6,model7)

```
## Data: myyelp
## Models:
## model6: stars.x ~ restaurant_style + tv + noiselevel + alcohol + smoking +
## model6:     (1 | state)
## model7: myyelp$stars.x ~ restaurant_style + tv + noiselevel + alcohol +
## model7:     smoking + (1 | state) + (1 | user_id)
## model3: myyelp$stars.x ~ restaurant_style + wifi + tv + parking + outseating +
## model3:     noiselevel + alcohol + smoking + (1 | state) + (1 | user_id)
##        Df   AIC   BIC   logLik deviance     Chisq Chi Df Pr(>Chisq)
## model6 16 21075 21207 -10521.5    21043
## model7 17 21072 21212 -10518.8    21038    5.3737      1    0.02044 *
## model3 30 17728 17975  -8833.8    17668 3370.0570     13    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA (model2,model3,model6)

```
## Data: myyelp
## Models:
## model6: stars.x ~ restaurant_style + tv + noiselevel + alcohol + smoking +
## model6:     (1 | state)
## model2: stars.x ~ restaurant_style + wifi + tv + parking + outseating +
## model2:     noiselevel + alcohol + smoking + (1 | state)
## model3: myyelp$stars.x ~ restaurant_style + wifi + tv + parking + outseating +
## model3:     noiselevel + alcohol + smoking + (1 | state) + (1 | user_id)
##        Df   AIC   BIC   logLik deviance     Chisq Chi Df Pr(>Chisq)
## model6 16 21075 21207 -10521.5    21043
## model2 29 17728 17967  -8834.8    17670 3373.2959     13     <2e-16 ***
## model3 30 17728 17975  -8833.8    17668    2.1348      1      0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA (model2,model3,model7)

```
## Data: myyelp
## Models:
## model7: myyelp$stars.x ~ restaurant_style + tv + noiselevel + alcohol +
## model7:     smoking + (1 | state) + (1 | user_id)
## model2: stars.x ~ restaurant_style + wifi + tv + parking + outseating +
## model2:     noiselevel + alcohol + smoking + (1 | state)
## model3: myyelp$stars.x ~ restaurant_style + wifi + tv + parking + outseating +
## model3:     noiselevel + alcohol + smoking + (1 | state) + (1 | user_id)
##        Df   AIC   BIC   logLik deviance     Chisq Chi Df Pr(>Chisq)
## model7 17 21072 21212 -10518.8    21038
## model2 29 17728 17967  -8834.8    17670 3367.9222     12     <2e-16 ***
## model3 30 17728 17975  -8833.8    17668    2.1348      1      0.144
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reference

Yelp Context and Analytics Reference

- Forbes 2013 https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/hbsworkingknowledge/2013/01/28/how-yelp-could-create-more-accurate-reviews/&refURL=https://www.google.com/&referrer=https://www.google.com/
- German, Andrew, and Jennifer Hill. "Data analysis using regression and multilevel/hierarchical models (analytical methods for social research)." (2006).
- HBS Working Knowledge https://www.forbes.com/sites/hbsworkingknowledge/#10fee2b230b7
- Luca, Michael, and Georgios Zervas. "Fake it till you make it: Reputation, competition, and Yelp review fraud." Management Science 62, no. 12 (2016): 3412-3427.
- Naatus, Mary Kate. "The Yelp Effect: Impact of Online Reputation in the Digital Era." International Journal of Business & Applied Sciences (2014): 49.

Code reference

- https://github.com/yolanda93/yelp_challenge_ui
- https://www.datacamp.com/community/blog/r-yelp-and-the-search-for-good-indian-food-an-open-course
- https://github.com/timjurka/sentiment/blob/master/sentiment/R/classify_emotion.R
- https://github.com/XIANG-Z/Yelp-Challenge.git
- https://github.com/juliasilge/tidytext/
- http://minimaxir.com/2015/12/lets-code-1/