Yelp_Report

Laura Wang

12/7/2019

Introduction

From the Yelp Data Set Challenge, I choose the business data for this project. The goal of this project is to see the association between restaurant's attributes(such as the ambience of restauran,noise level,parking availability..etc), categories(such as cuisine type and also serving type) with their rating stars. For the multilevel modeling, I will use city and state as random effects and to compare the difference of the outcome between when using only state and use both state and city as random effects. This project will also explained the fixed effect of the model.

Data Preparation

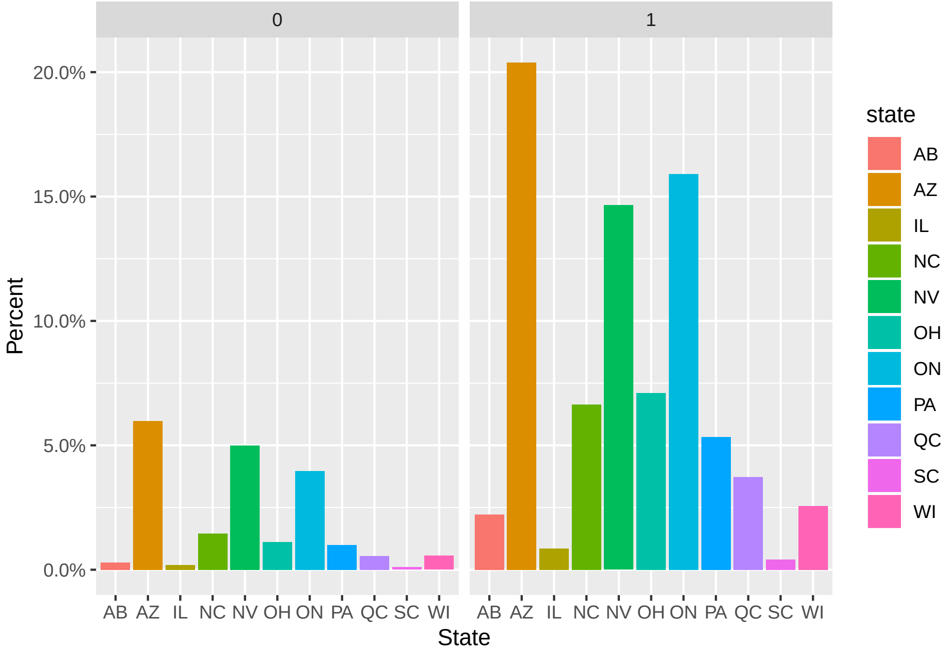
- – Choosed all restaurant data
- – Choosed restaurant with reviews more than 30 reviews
- – Cleaned data, deleted rows with all NA values and rows that all attributes with NA values
- – Encode/convert categorical and logical variables
- – Normalize continues variableas

Variable Explanation

- `business_id` : Business(retaurant)'s unique ID
- `name`: Business(retaurant)'s name
- `city`: City of the business located
- `state`: State of the business located
- `postal_code` : Post code of the business
- `stars`: The rating stars of the business, rounded to half-stars
- `review_count`: Number of reviews
- `is_open`: 0 or 1 for closed or open
- `n_parks`: Number of ways of parking the restauran is avaliable
- `Caters`: 0 or 1 for without or with caters
- `TakeOut`: 0 or 1 for can takeout or can't takeout
- `PriceRange`: 1-4 for low to high price level
- `OutdoorSeating`: 0 or 1 for unavailable or available for ourdoor seating
- `HasTV`: 0 or 1 for unavailable or available for TV
- `NoiseLevel`: Categorized as average, loud, quiet, very_loud
- `WiFi`: 0 or 1 for unavailable or available for WiFi
- `Alcohol`: Categorized the restauran as avaliable for beer_and_wine,full_bar or none
- `Ambience`: Column separated to column romantic, intimate, classy, hipster, divey, touristy, trendy, upscale, casual. Each column with 0 or 1 value, indicates whether has the corresponsive attributes or not
- `Categories`: Column separated to column American, Nightlife, Breakfast_Brunch, Italian,Mexican, Mediterranean. Each column with 0 or 1 value, indicates whether has the corresponsive attributes or not
- `longitude`: Restaurant lontitude
- `latitude`: Restaurant latitude

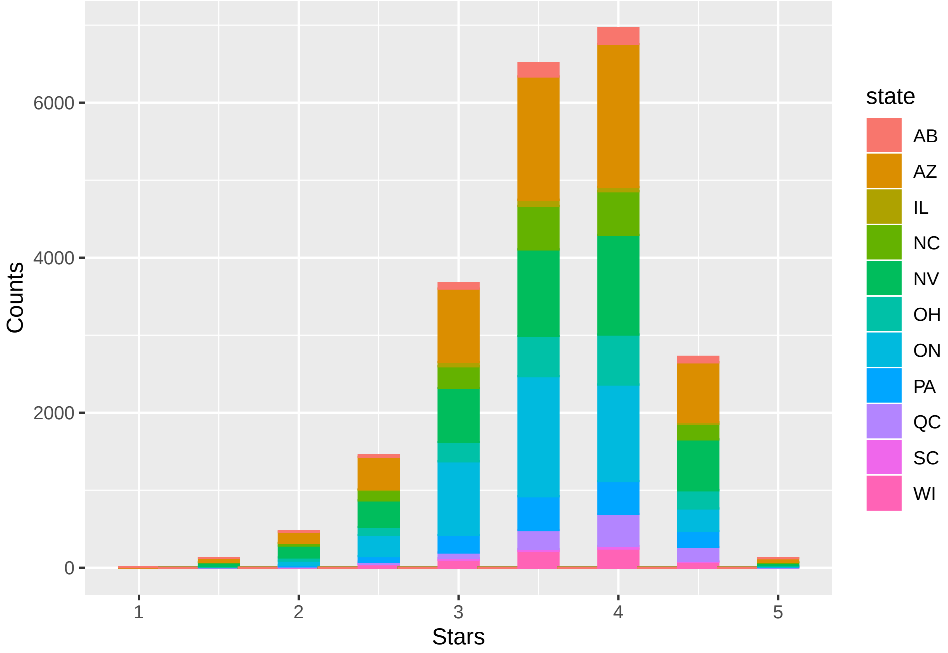
EDA

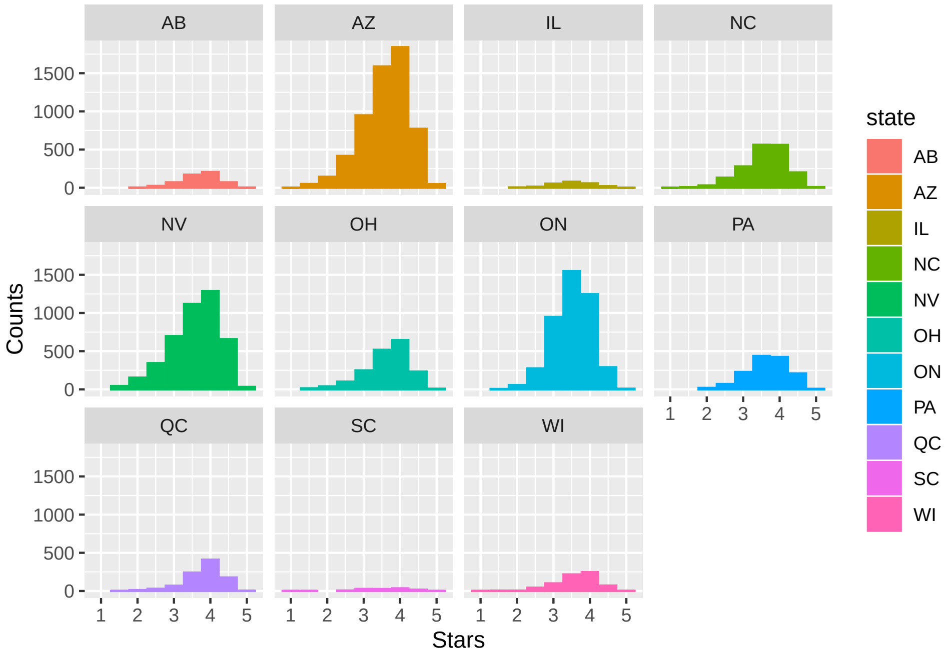
Restaurant observations in each states, grouped by is_open, 0 for closed, 1 for open



As we can see in this data set, 26% of data are from Arizona, 20% from Nevada and 20% from Ontario (Canada). Most of the restaurants in the dataset are open.

Rating Distributions in all state

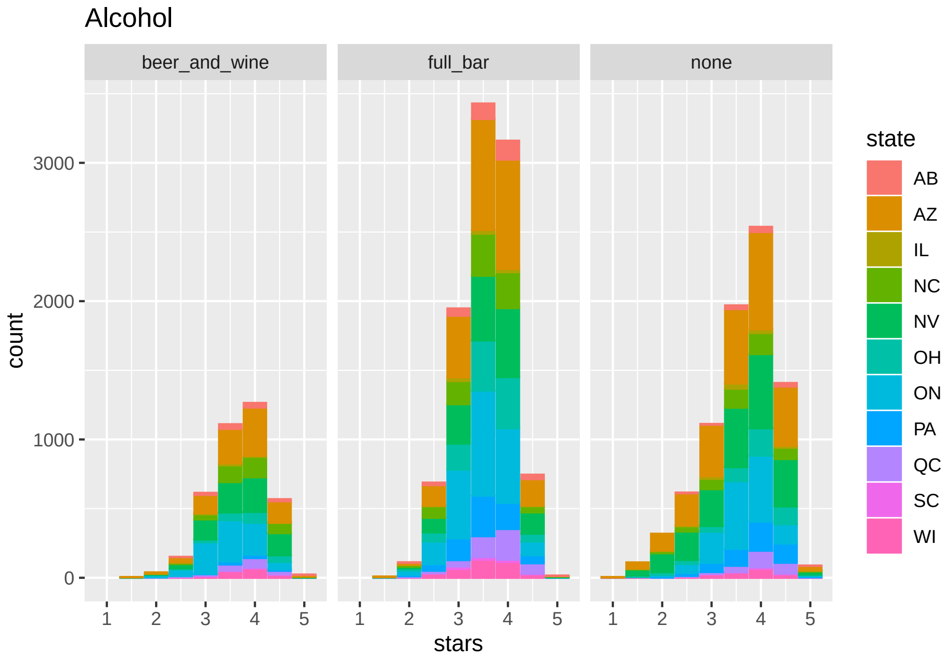


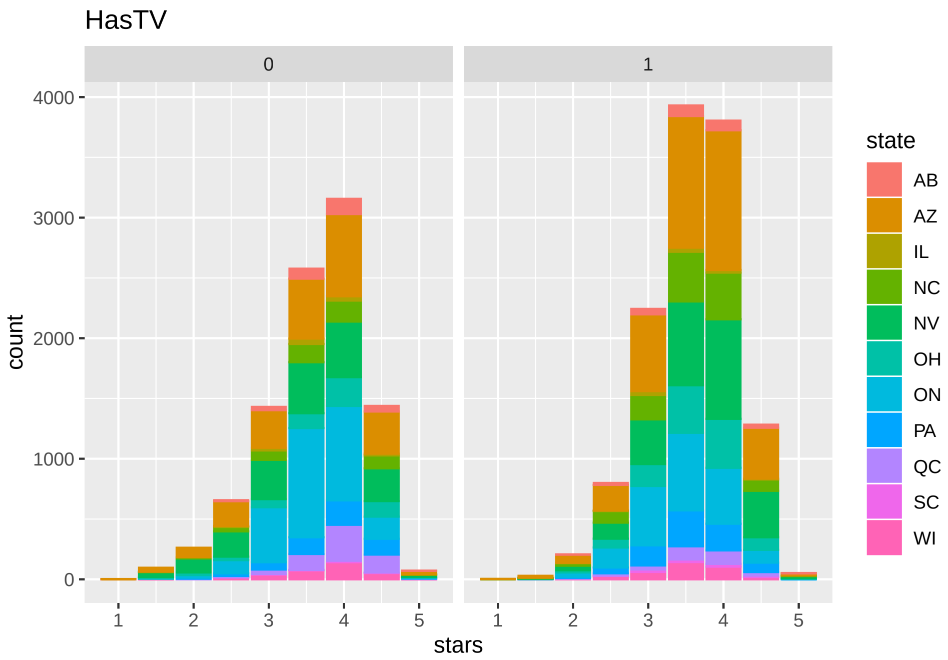


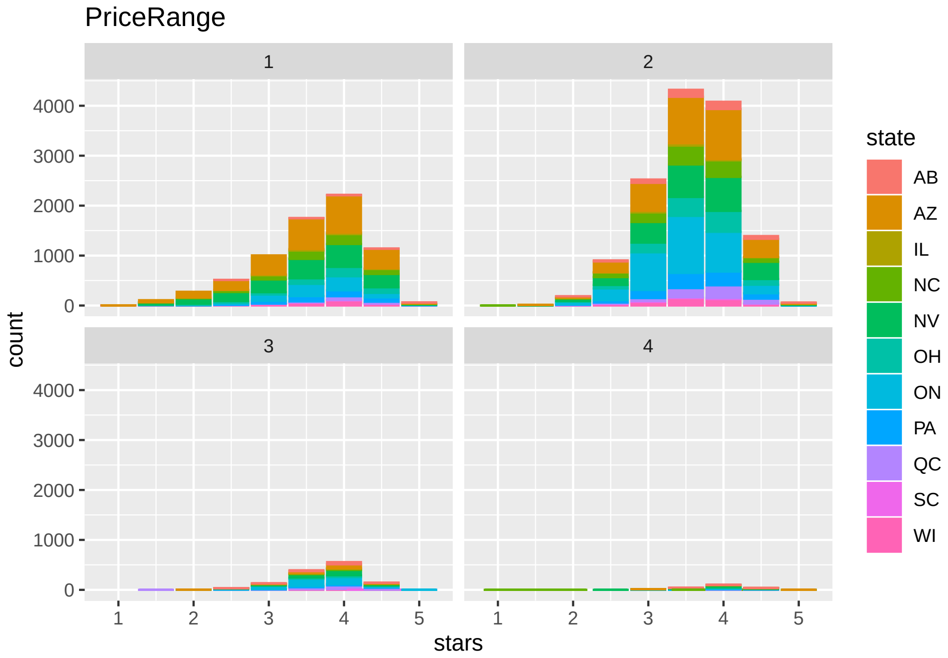
As we can see from the plot, the overall stars distribution are concentrated on 3.5-4 scale. From each state, we can see in AZ,NV and OH, the stars are concentrated on score 4, and in ON the score is more concentrated on 3.5.

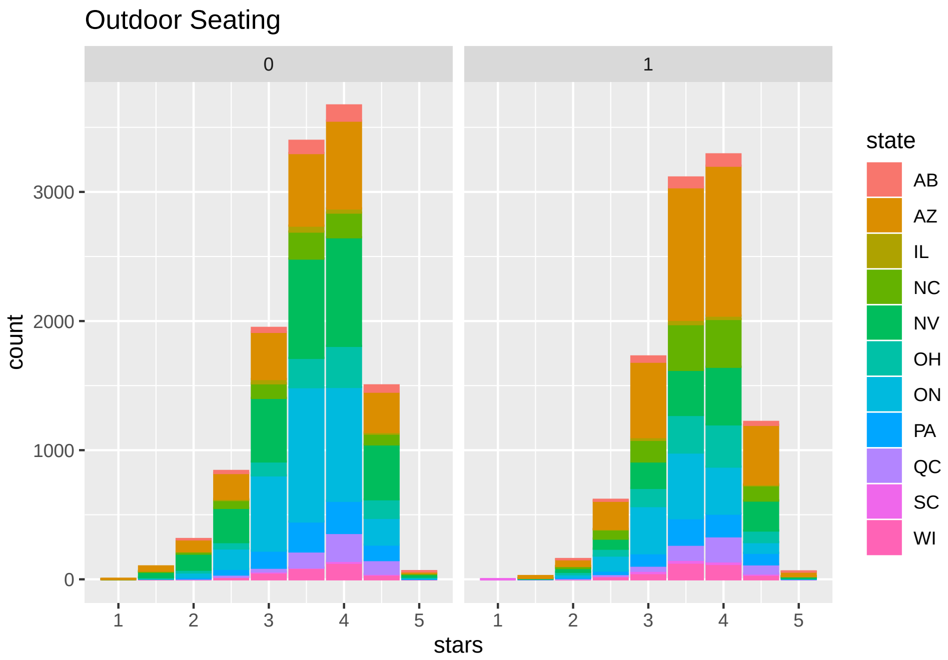
Rating Distributions in all state in different attributes

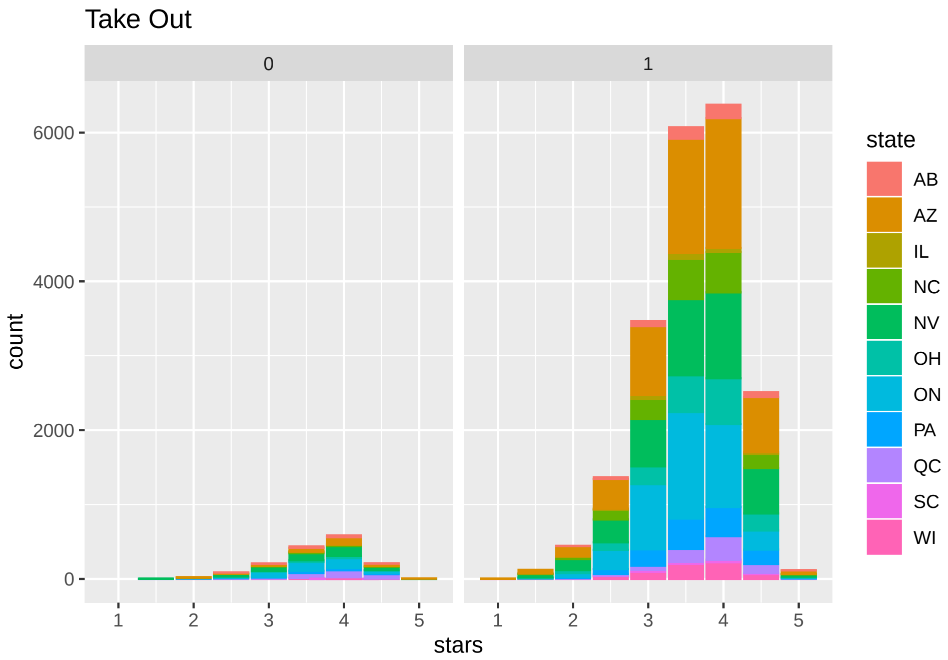
We try to see how different attributs of the restaurant contribute the influence to the rating scores.

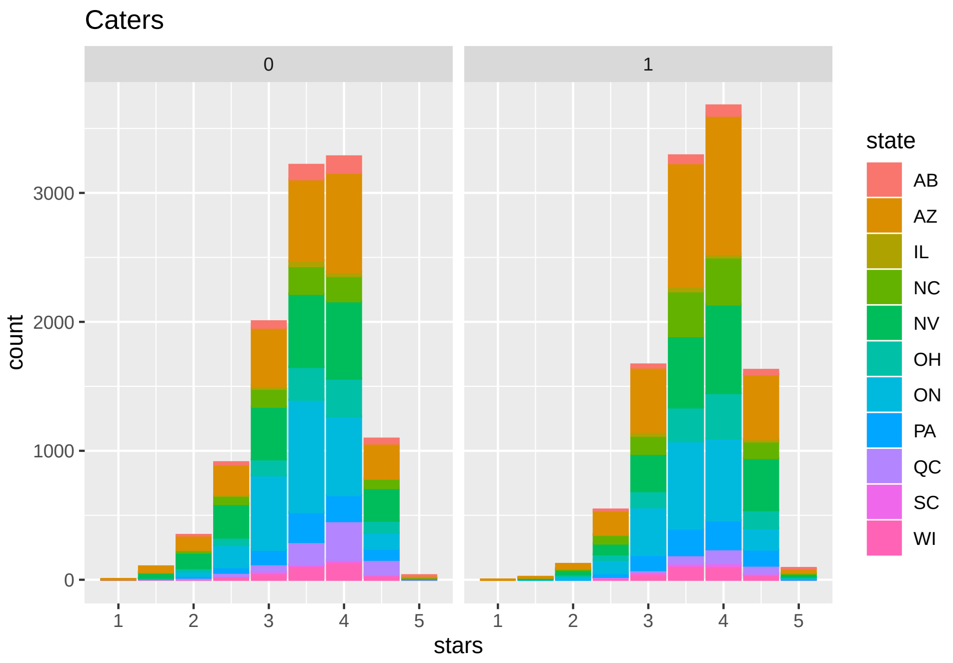



HasTV
0
1
count
4000
3000
2000
1000
0
1
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI


PriceRange
1
2
3
4
count
4000
3000
2000
1000
0
1
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI


Outdoor Seating
0
1
count
3000
2000
1000
0
1
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI

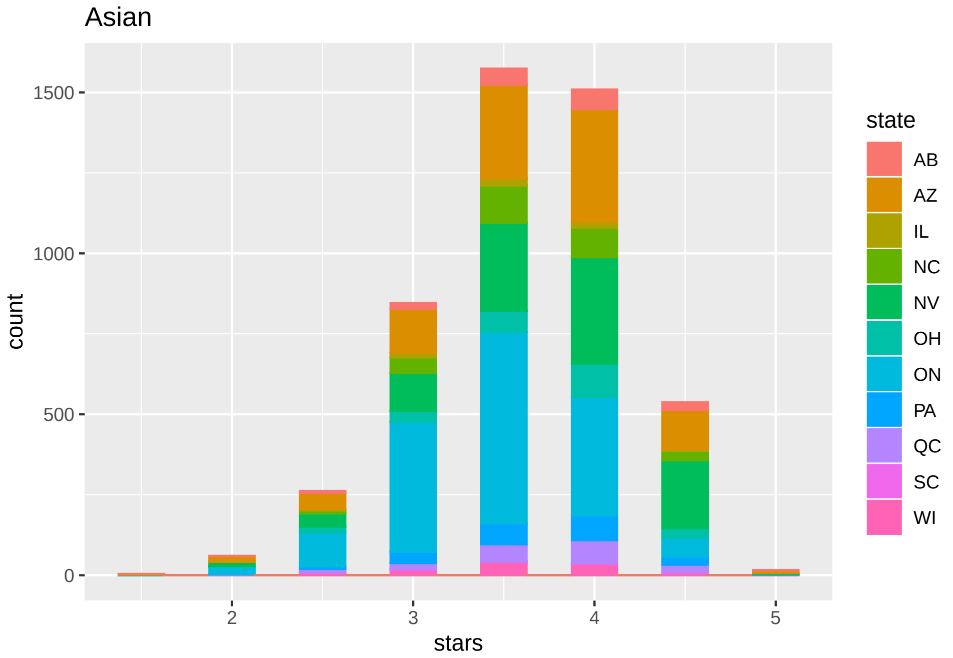

Take Out
0
1
count
6000
4000
2000
0
1
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI

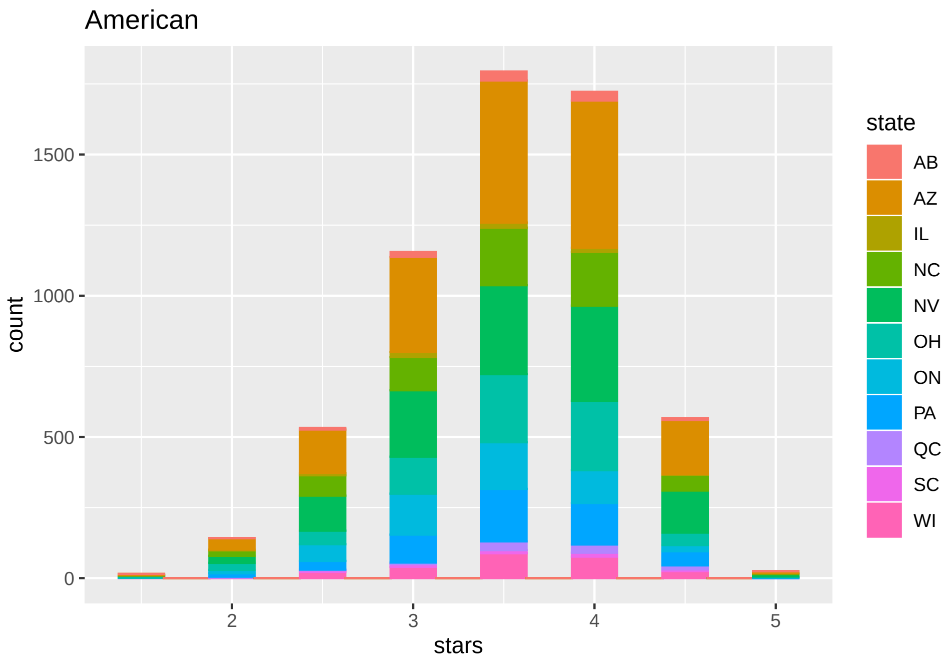


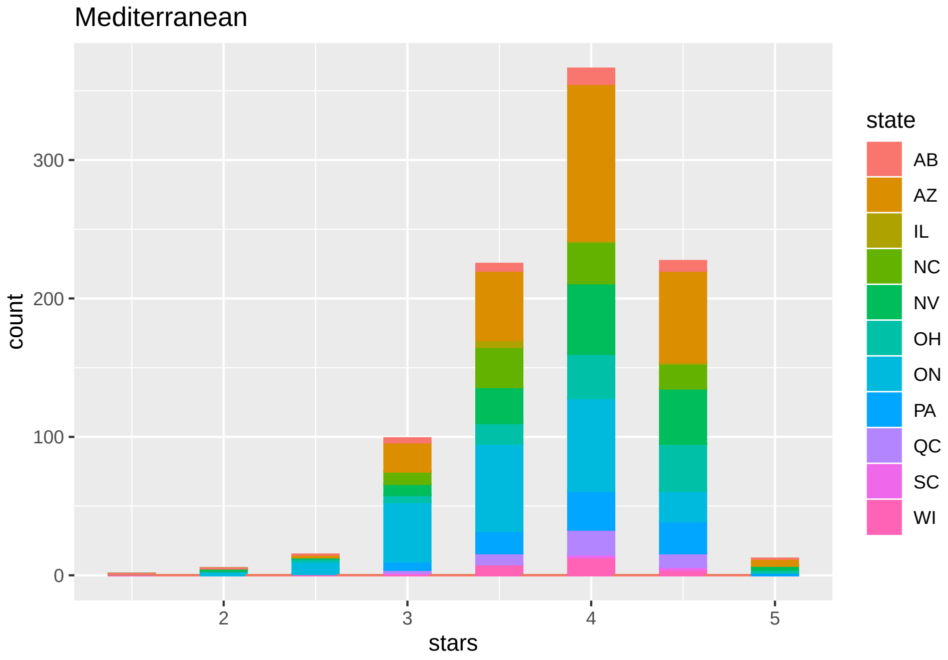
From above plots, we can see that the overall star distribution in each cateogries of the attributes didn't show much difference, however, we still can see the star distribution differences in different state.

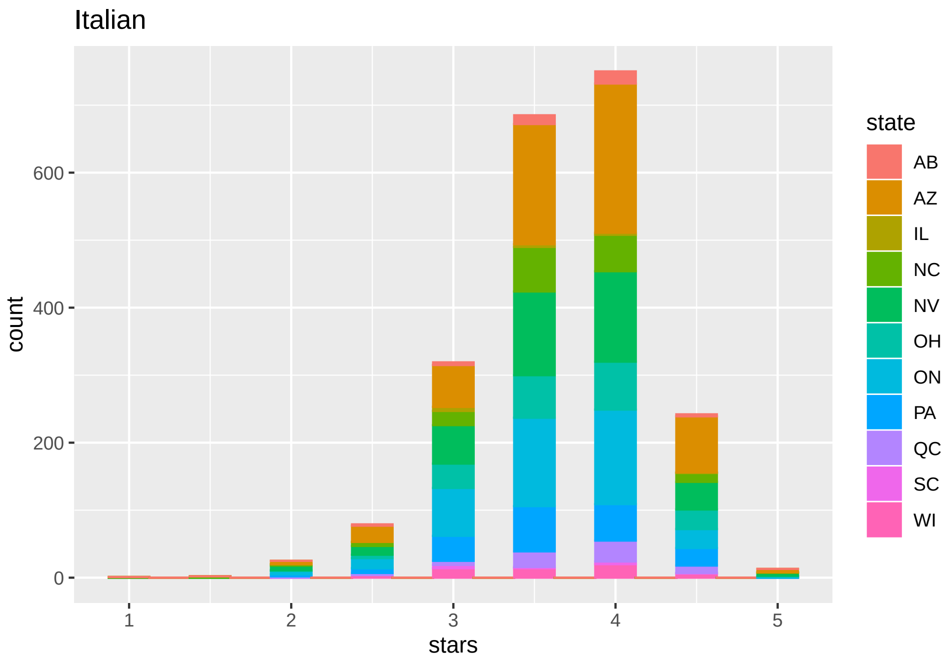
Rating stars in different types of cuisine- Asian, American, Mediterranean, Italian, Mexican

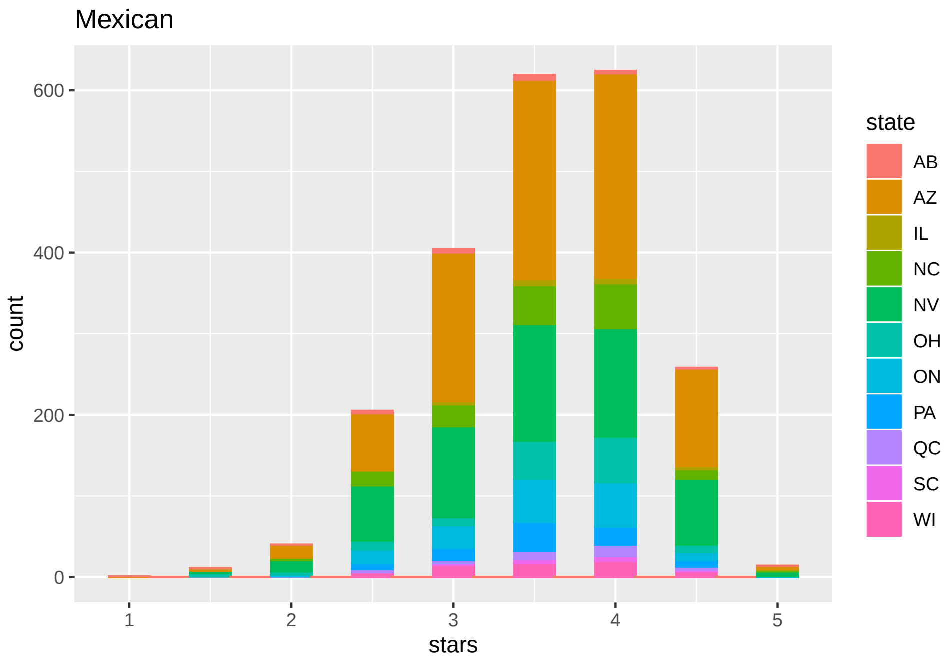
We want to see if the rating distribution would be different in different restaurant categories.




American
count
1500
1000
500
0
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI

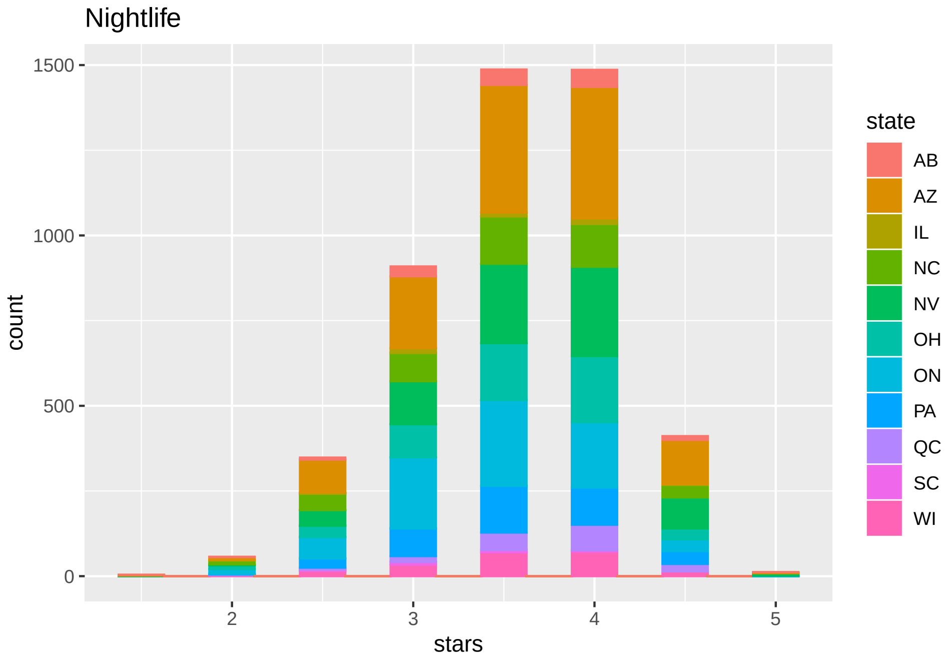

Mediterranean
count
300
200
100
0
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI

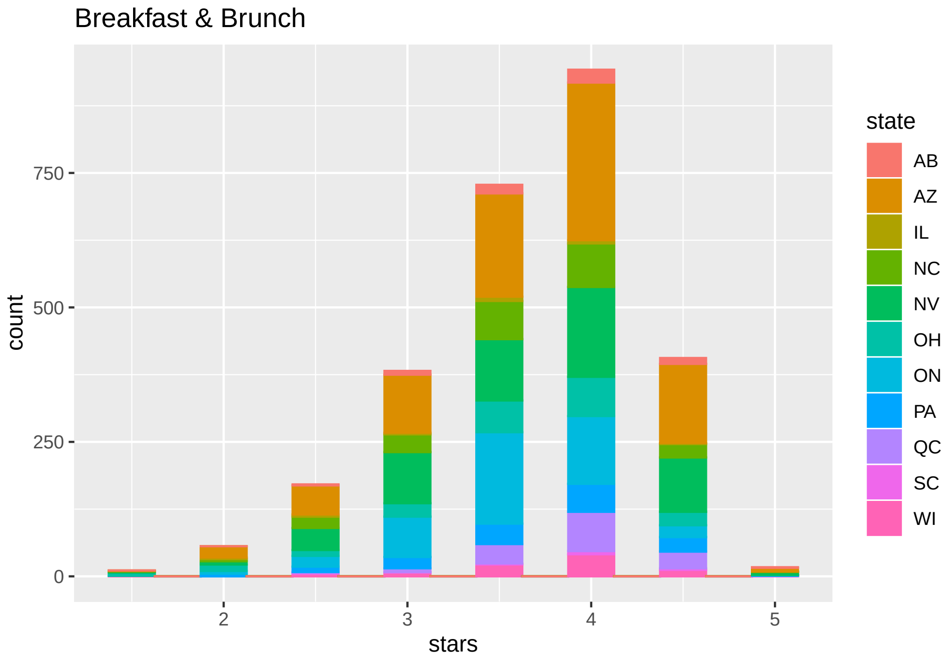




As we can see Mediterranean and Italian food restaurant are more concentrated on 4 stars. Asian and American food restaurant are more concentrated on 3.5 stars. Mexican food restaurant are distributed more equally on 3.5 and 4 stares. So maybe the different cuisine may have inflence on the rating stars.

Rating stars for categories tagged as Nightlife and Breakfast & Brunch





For restaurant categorized as Nightlife and Breakfast & Brunch. We can see that for Nightlife, the difference between sore 3.5 and 4 are not that much, however, for Breakfast & Brunch, scores are more concentrated on 4.

Modeling

We are trying to see how strong the association of each factors has with the restaurant ratings. ### Select variables for Predictor First we use the stepwise regression (or stepwise selection) to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

```
## lm(formula = stars ~ review_count + NoiseLevel + state + Caters +
##     Alcohol + trendy + intimate + hipster + Mediterranean + PriceRange +
##     divey + casual + classy + American + touristy + n_parks +
##     Nightlife + romantic + Italian + TakeOut + HasTV, data = aic_table)
```

From the above result, we're going to choose review_count, NoiseLevel, state, Caters, Alcohol, trendy, intimate, hipster, Mediterranean, PriceRange, divey, casual, classy, American, touristy, n_parks, Nightlife, romantic, Italian, TakeOut, HasTV as predictor variables

Fit lmer model to treat State as group. Random intercept with fixed mean.

Then we use lemr using state as group to fit the model-fit1

```
fit1 <-  lmer (data= bs_model2, stars ~ review_count + NoiseLevel + Caters +
    Alcohol + trendy + intimate + hipster + Mediterranean + PriceRange +
    divey + casual + classy + American + touristy + n_parks +
    Nightlife + romantic + Italian + TakeOut + HasTV + (1|state))
```

Fit lmer model to treat State and City as group. Intercept varying among State and City.

Try to use both city and state as group

```
fit2 <-  lmer (data= bs_model2, stars ~ review_count + NoiseLevel + Caters +
    Alcohol + trendy + intimate + hipster + Mediterranean + PriceRange +
    divey + casual + classy + American + touristy + n_parks +
    Nightlife + romantic + Italian + TakeOut + HasTV  + (1|state)+ (1|city))
```

Fit lmer model to treat State and City as group. Intercept varying among State and City within State.

```
fit3 <-  lmer (data= bs_model2, stars ~ review_count + NoiseLevel + Caters +
    Alcohol + trendy + intimate + hipster + Mediterranean + PriceRange +
    divey + casual + classy + American + touristy + n_parks +
    Nightlife + romantic + Italian + TakeOut + HasTV  +  (1|state/city))
```

Model Validation & Interpretation

AIC check

```
##      df      AIC
## fit1 28 36729.70
## fit2 29 36534.93
## fit3 29 36532.17
```

As can see from the result, using City and State as random effect (Intercept varying among State and City) improved the model. We choose fit3 (Intercept varying among State and City within State) as our model for interpretation and validation since it has the lowest AIC.

Random Effect

```
## [1] "The fixed effect intercept is 3.609"
```

```
## [1] "The random effect intercepts for each state are:"
```

```
##    (Intercept)
## AB       0.075
## AZ      -0.174
## IL       0.028
## NC       0.037
## NV      -0.187
## OH       0.019
## ON      -0.074
## PA       0.080
## QC       0.134
## SC       0.020
## WI       0.042
```

The intercept in each state is the state value plus the fixed effect intercept value (3.069). For example for Arizona, the intercept would be 3.069-0.174 = 2.895. The meaning is that, the overall average stars is 0.174 below the complete pooling(overall) mean.

Fixed Effect

```
## [1] "The top5 positive factors:"
```

```
##        intimate         hipster           divey NoiseLevelquiet 
##           0.343           0.282           0.241           0.220 
##   Mediterranean 
##           0.210
```

```
## [1] "The top5 negtive factors:"
```

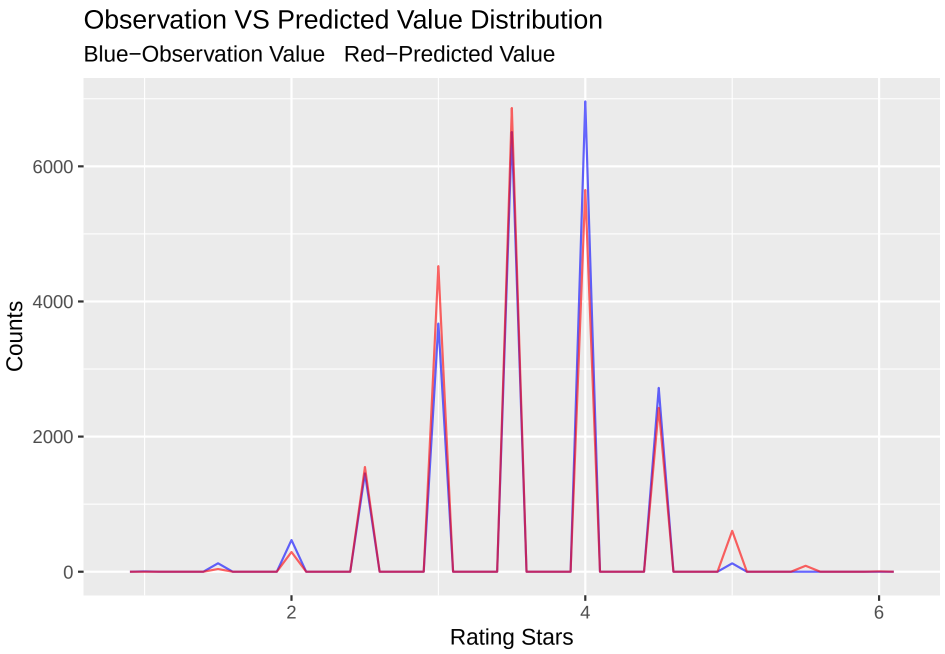
```
## NoiseLevelvery_loud            touristy      NoiseLevelloud 
##              -0.387              -0.246              -0.219 
##     Alcoholfull_bar            American 
##              -0.181              -0.072
```

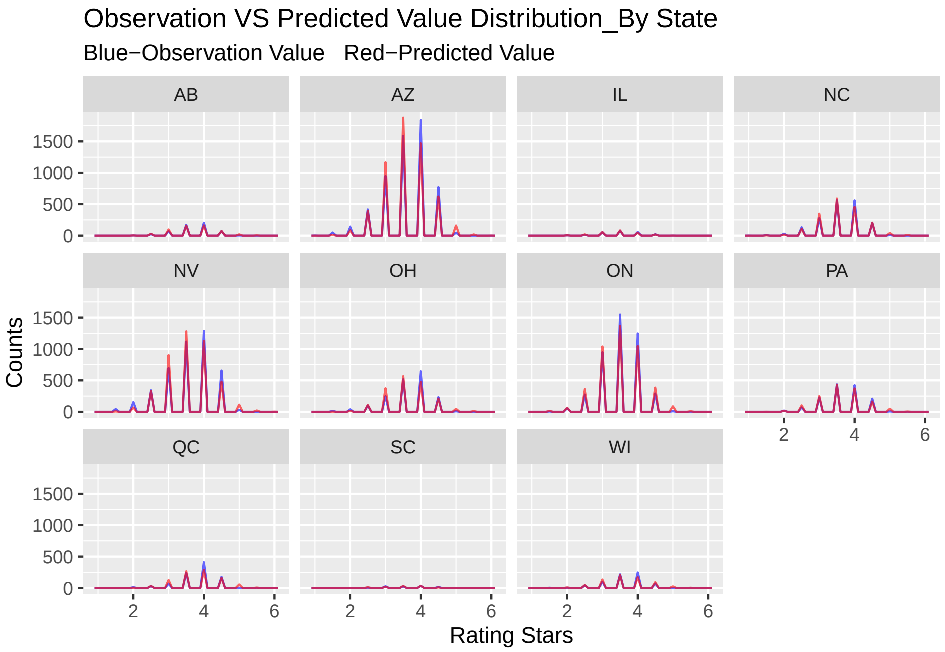
Since the variables are binary, so can tell from the above result that, intimate factor has the strongest positive association with rating score and the NoiseLevel_very_loud contributes the most negtive relationship. This means, if all other factors are the same between two restaurants, the one with ambience attribute coded as "intimate" will have 0.343 score higher on average then the restaurant without that attribute. And if all other factors are the same between two restaurants, the one with NoiseLevel labeled as "very_loud" will have 0.387 score lower on average then the restaurant without that attribute.

Model use State and City as Random Effect

Check Fitted Values Distribution

- Red plot is the observation value and the blue plot is the predicted value

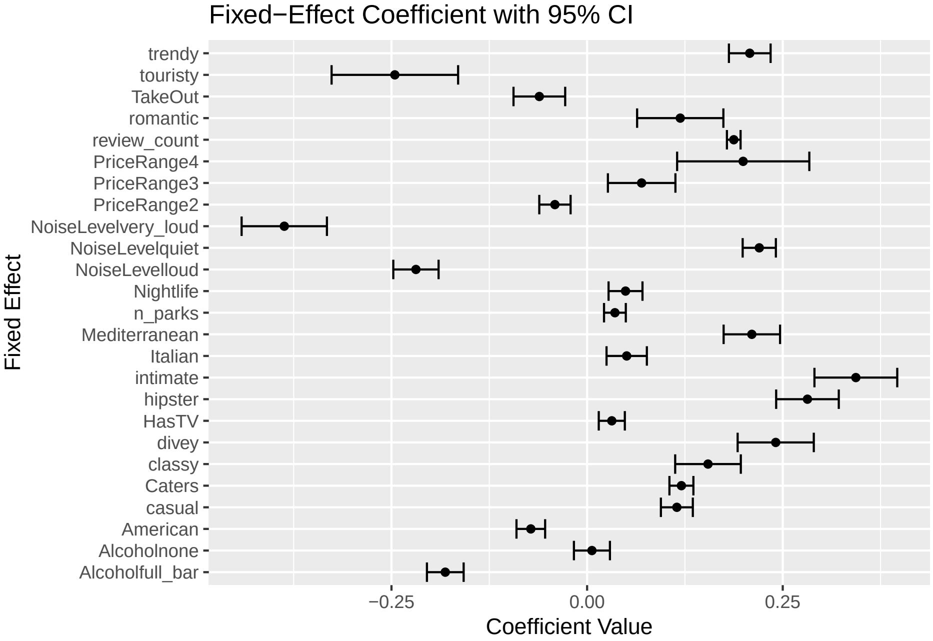

Observation VS Predicted Value Distribution
Blue−Observation Value   Red−Predicted Value


Observation VS Predicted Value Distribution_By State
Blue−Observation Value   Red−Predicted Value

From the plot we can see the multilevel model will pool the predicted value more toward 3.5. It will pull up the value of lower scores and will pull down the value of higher scores. For each state fitting, the effect can also see from each state.

Fix-Effect Coefficient for model fit3

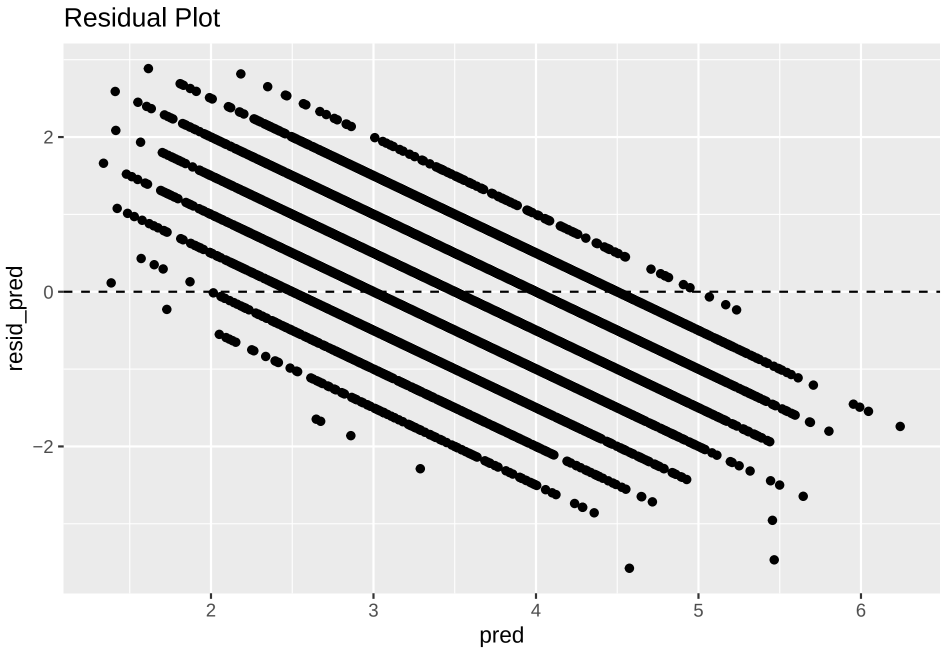
```
## Computing profile confidence intervals ...
```



From the fixed effect coefficeint value plot, we can see that the Alcohol labled as none doesn't have influence on the rating score, since the confidence interval is crossing 0. Among all the ambience factors, except touristy, all other ambience type have positive influence on the score, and touristy has negtive influence. From this we can infer that most of the restaurant with touristy ambience will lower the scores of the restaurant. Another interesting finding is that, American food would have a negative influence. This might be caused by the diversity of American food restaurant which could also include fast-food.

Problems and Limitaion

Since the outcome of the stars is ordinal, the model lmer is trying to fit the out come as continues, so we can see the residual plot is not as normal distributed. To improve the model, we need to considerr the the multinomial multilevel mnodels. This model can be implemented in brms package using brm function to choose the family equals to acat("logit"). This can be tried in future modeling.



Appedix

Display of Fitted Model

```
## lmer(formula = stars ~ review_count + NoiseLevel + Caters + Alcohol +
##     trendy + intimate + hipster + Mediterranean + PriceRange +
##     divey + casual + classy + American + touristy + n_parks +
##     Nightlife + romantic + Italian + TakeOut + HasTV + (1 | state),
##     data = bs_model2)
##                      coef.est coef.se
## (Intercept)           3.57     0.05
## review_count          0.19     0.00
## NoiseLevelloud       -0.22     0.01
## NoiseLevelquiet       0.22     0.01
## NoiseLevelvery_loud -0.39     0.03
## Caters                0.12     0.01
## Alcoholfull_bar      -0.17     0.01
## Alcoholnone           0.00     0.01
## trendy                0.21     0.01
## intimate              0.35     0.03
## hipster               0.29     0.02
## Mediterranean         0.21     0.02
## PriceRange2          -0.04     0.01
## PriceRange3           0.07     0.02
## PriceRange4           0.20     0.04
## divey                 0.24     0.02
## casual                0.11     0.01
## classy                0.15     0.02
## American             -0.07     0.01
## touristy             -0.24     0.04
## n_parks               0.04     0.01
## Nightlife             0.05     0.01
## romantic              0.11     0.03
## Italian              0.06     0.01
## TakeOut             -0.06     0.02
## HasTV                0.03     0.01
##
## Error terms:
##  Groups   Name        Std.Dev.
##  state    (Intercept) 0.14
##  Residual             0.55
## ---
## number of obs: 22041, groups: state, 11
## AIC = 36729.7, DIC = 36332.1
## deviance = 36502.9
```

```
## lmer(formula = stars ~ review_count + NoiseLevel + Caters + Alcohol +
##     trendy + intimate + hipster + Mediterranean + PriceRange +
##     divey + casual + classy + American + touristy + n_parks +
##     Nightlife + romantic + Italian + TakeOut + HasTV + (1 | state) +
##     (1 | city), data = bs_model2)
##                     coef.est coef.se
## (Intercept)          3.61     0.05
## review_count         0.19     0.00
## NoiseLevelloud      -0.22     0.01
## NoiseLevelquiet      0.22     0.01
## NoiseLevelvery_loud -0.39     0.03
## Caters               0.12     0.01
## Alcoholfull_bar     -0.18     0.01
## Alcoholnone          0.01     0.01
## trendy               0.21     0.01
## intimate             0.34     0.03
## hipster              0.28     0.02
## Mediterranean        0.21     0.02
## PriceRange2         -0.04     0.01
## PriceRange3          0.07     0.02
## PriceRange4          0.20     0.04
## divey                0.24     0.02
## casual               0.11     0.01
## classy               0.15     0.02
## American            -0.07     0.01
## touristy            -0.25     0.04
## n_parks              0.04     0.01
## Nightlife            0.05     0.01
## romantic             0.12     0.03
## Italian              0.05     0.01
## TakeOut             -0.06     0.02
## HasTV                0.03     0.01
##
## Error terms:
##  Groups   Name        Std.Dev.
##  city     (Intercept) 0.15
##  state    (Intercept) 0.11
##  Residual             0.55
## ---
## number of obs: 22041, groups: city, 362; state, 11
## AIC = 36534.9, DIC = 36134.5
## deviance = 36305.7
```

```
## lmer(formula = stars ~ review_count + NoiseLevel + Caters + Alcohol +
##     trendy + intimate + hipster + Mediterranean + PriceRange +
##     divey + casual + classy + American + touristy + n_parks +
##     Nightlife + romantic + Italian + TakeOut + HasTV + (1 | state/city),
##     data = bs_model2)
##                     coef.est coef.se
## (Intercept)          3.61     0.05
## review_count         0.19     0.00
## NoiseLevelloud      -0.22     0.01
## NoiseLevelquiet      0.22     0.01
## NoiseLevelvery_loud -0.39     0.03
## Caters               0.12     0.01
## Alcoholfull_bar     -0.18     0.01
## Alcoholnone          0.01     0.01
## trendy               0.21     0.01
## intimate             0.34     0.03
## hipster              0.28     0.02
## Mediterranean        0.21     0.02
## PriceRange2         -0.04     0.01
## PriceRange3          0.07     0.02
## PriceRange4          0.20     0.04
## divey                0.24     0.02
## casual               0.11     0.01
## classy               0.15     0.02
## American            -0.07     0.01
## touristy            -0.25     0.04
## n_parks              0.04     0.01
## Nightlife            0.05     0.01
## romantic             0.12     0.03
## Italian              0.05     0.01
## TakeOut             -0.06     0.02
## HasTV                0.03     0.01
##
## Error terms:
##  Groups     Name        Std.Dev.
##  city:state (Intercept) 0.15
##  state      (Intercept) 0.12
##  Residual               0.55
## ---
## number of obs: 22041, groups: city:state, 366; state, 11
## AIC = 36532.2, DIC = 36131.8
## deviance = 36303.0
```