Yelp_Report

Laura

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

Clean Data

```
- Choosed all restaurant data
- Choosed restaurant with reviews more than 30 reviews
- Cleaned data, deleted all NA rows also rows with all attributes with NA values
- Select potential variables
```

Modify Data

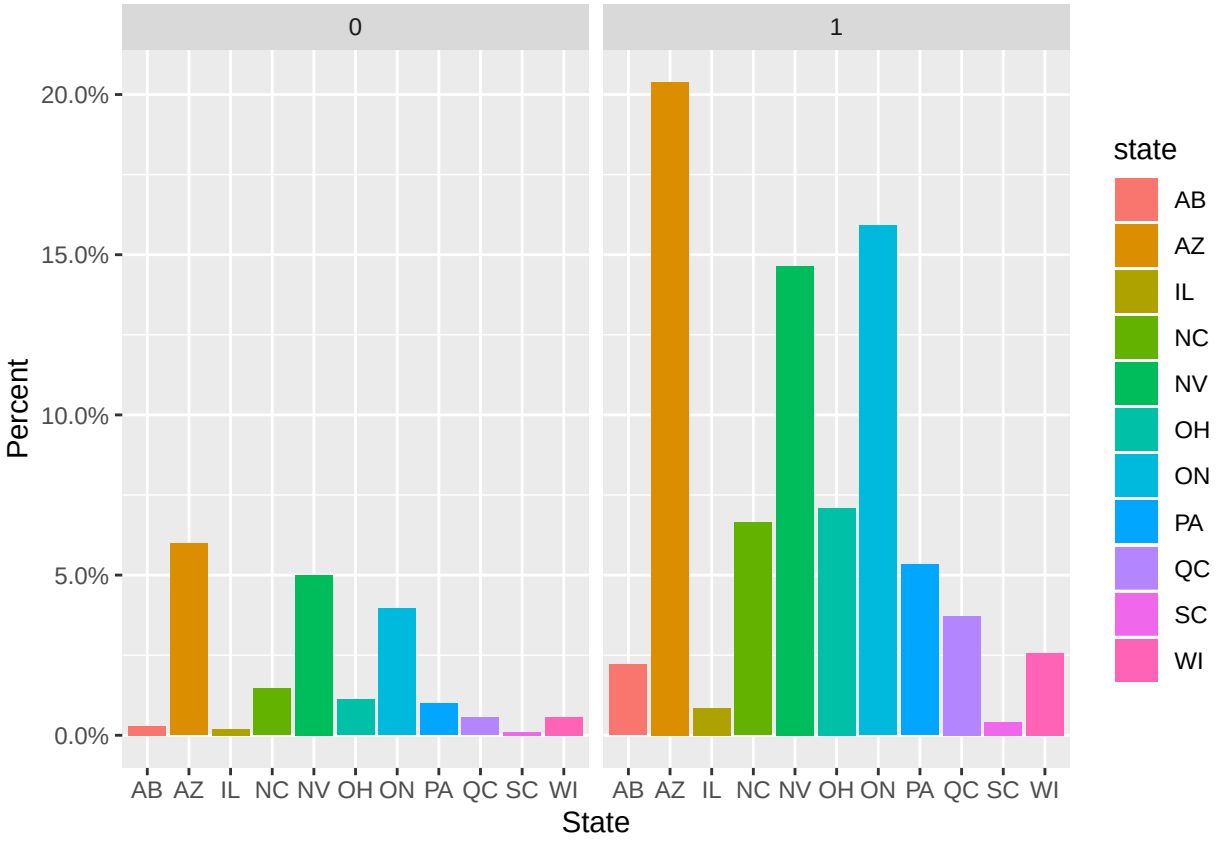
```
-Modify missing data in the subset dataset and add new columns as n_parks, Caters, Take out, PriceRange
 romantic, intimate, classy, hipster, divey, touristy, trendy, upscale, casual,
 Asian, American, Nightlife, Breakfast_Brunch, Italian,Mexican, Mediterranean
```

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`: no or yew for unavailable or available for WiFi
- `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`: Columns named as 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`: Columns named as Asian, American, Nightlife, Breakfast_Brunch, Italian,Mexican, Mediterranean. Each column with 0 or 1 value, indicates whether has the corresponsive attributes or not

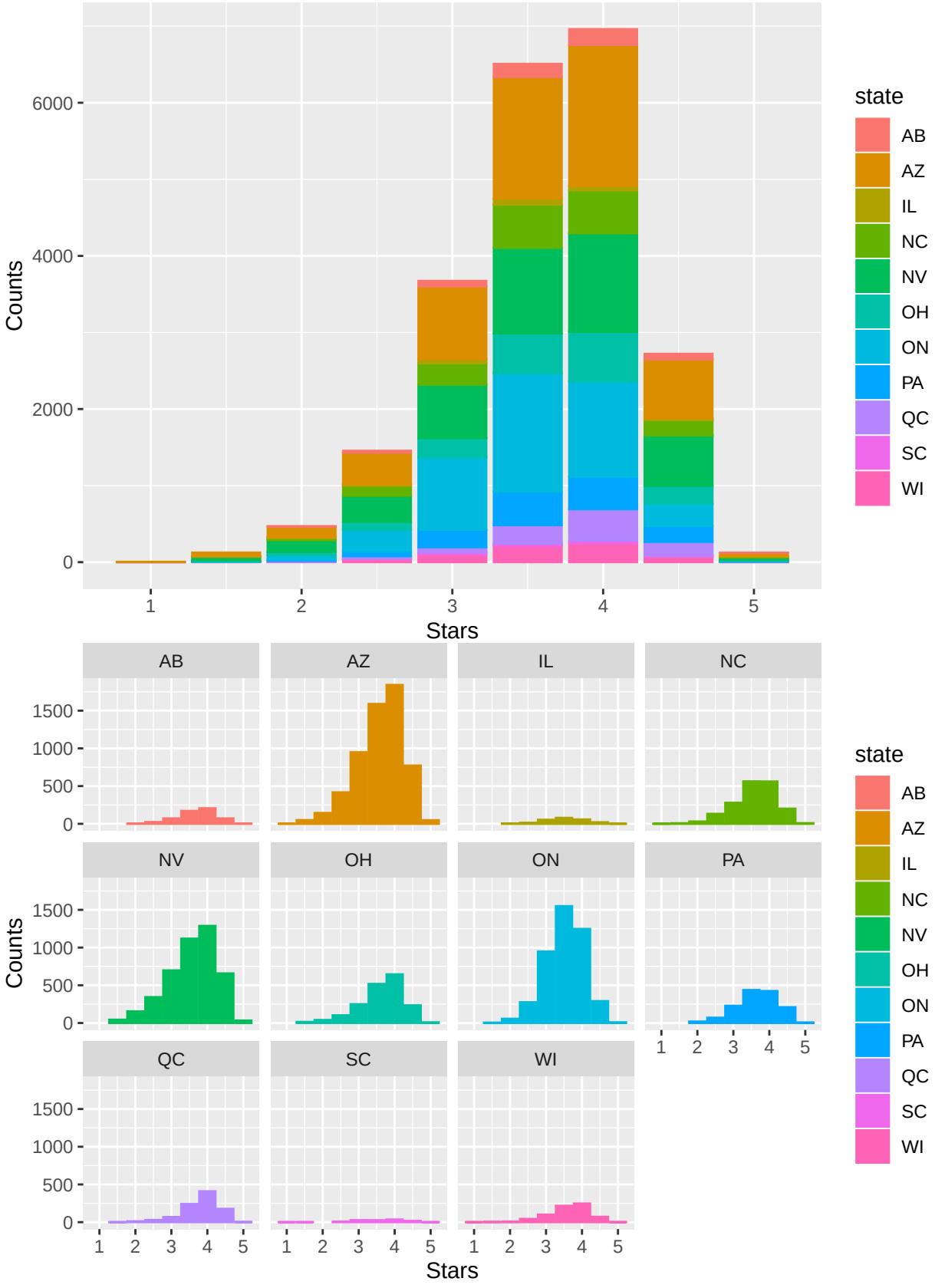
EDA

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

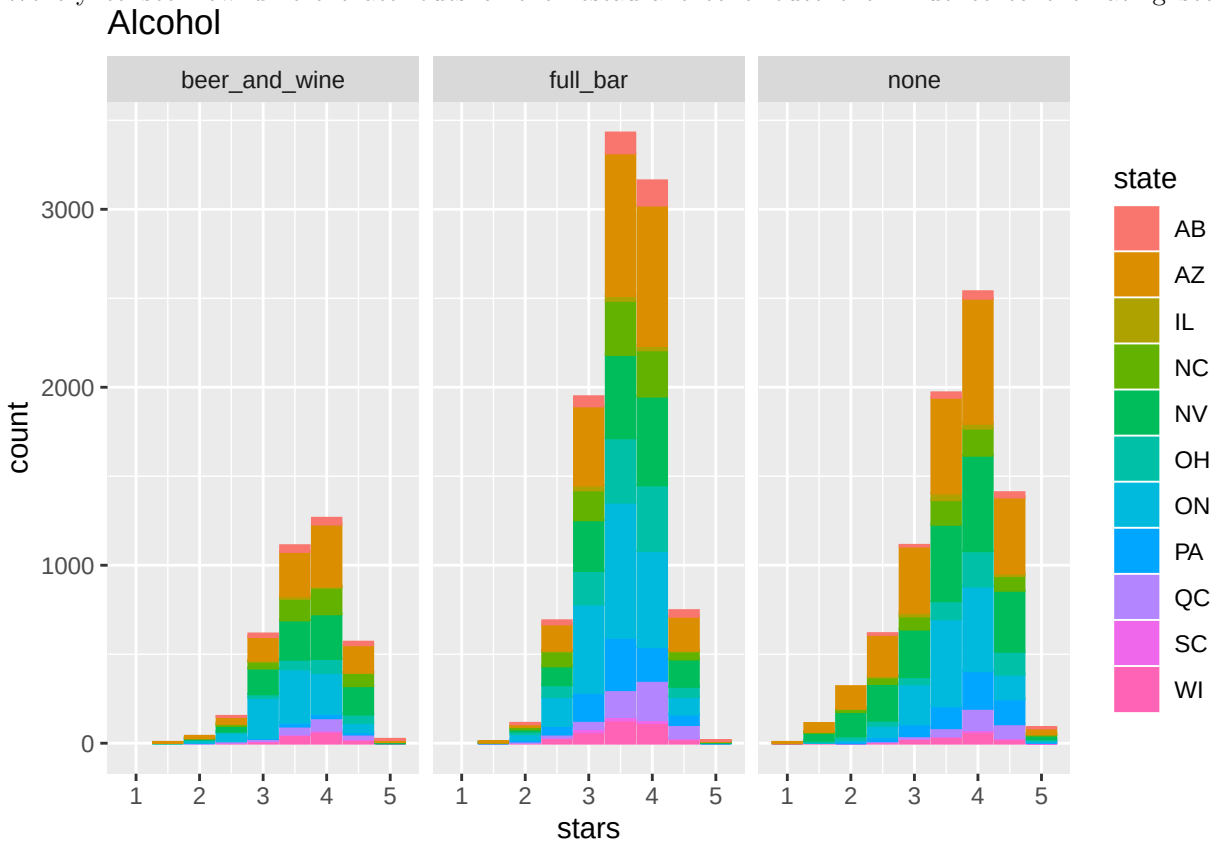


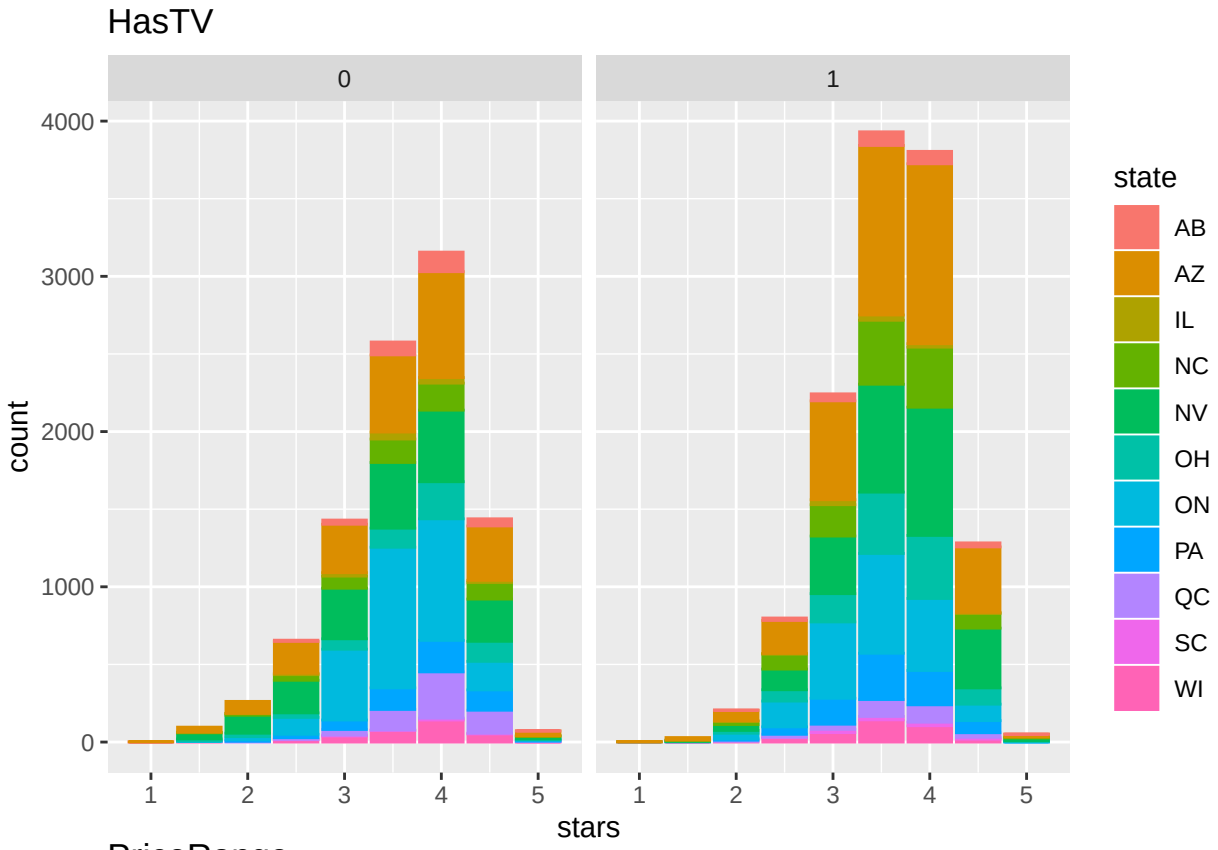
As we can see from the plot, the overall stars distribution are concentrated on 3.5-4 scale. From each state, we

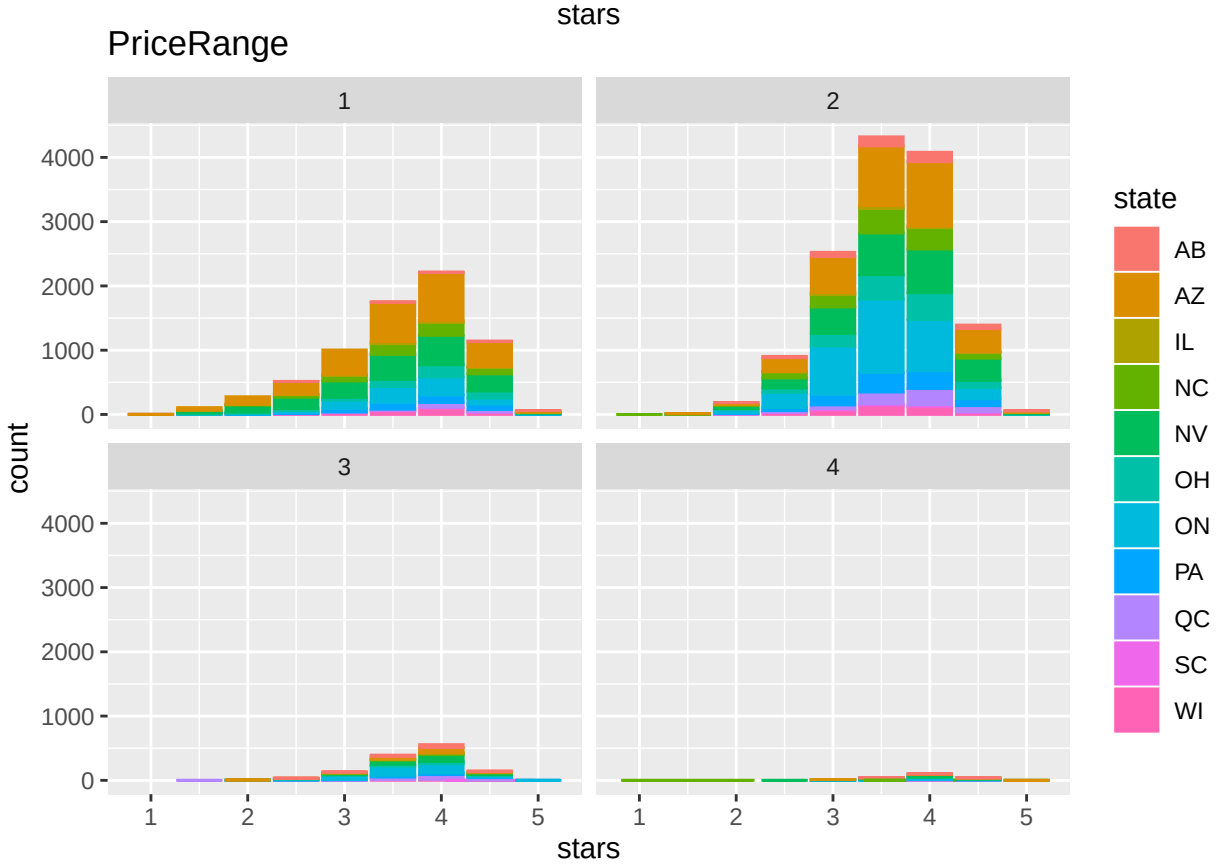
can see in AZ,NV and OH, the stars are concentrated on score 4, and in ON the score is more concentrated on 3.5.

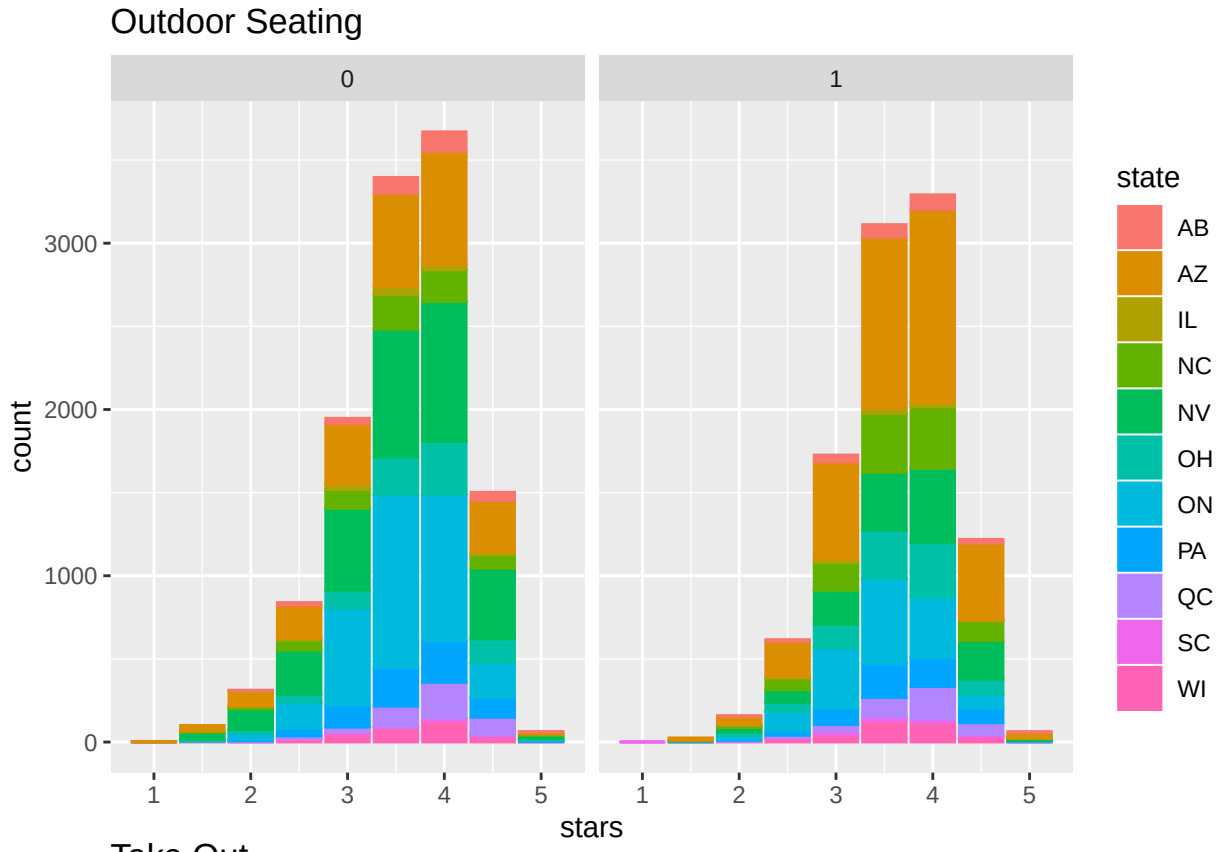
Rating Distributions in all state in different attributes

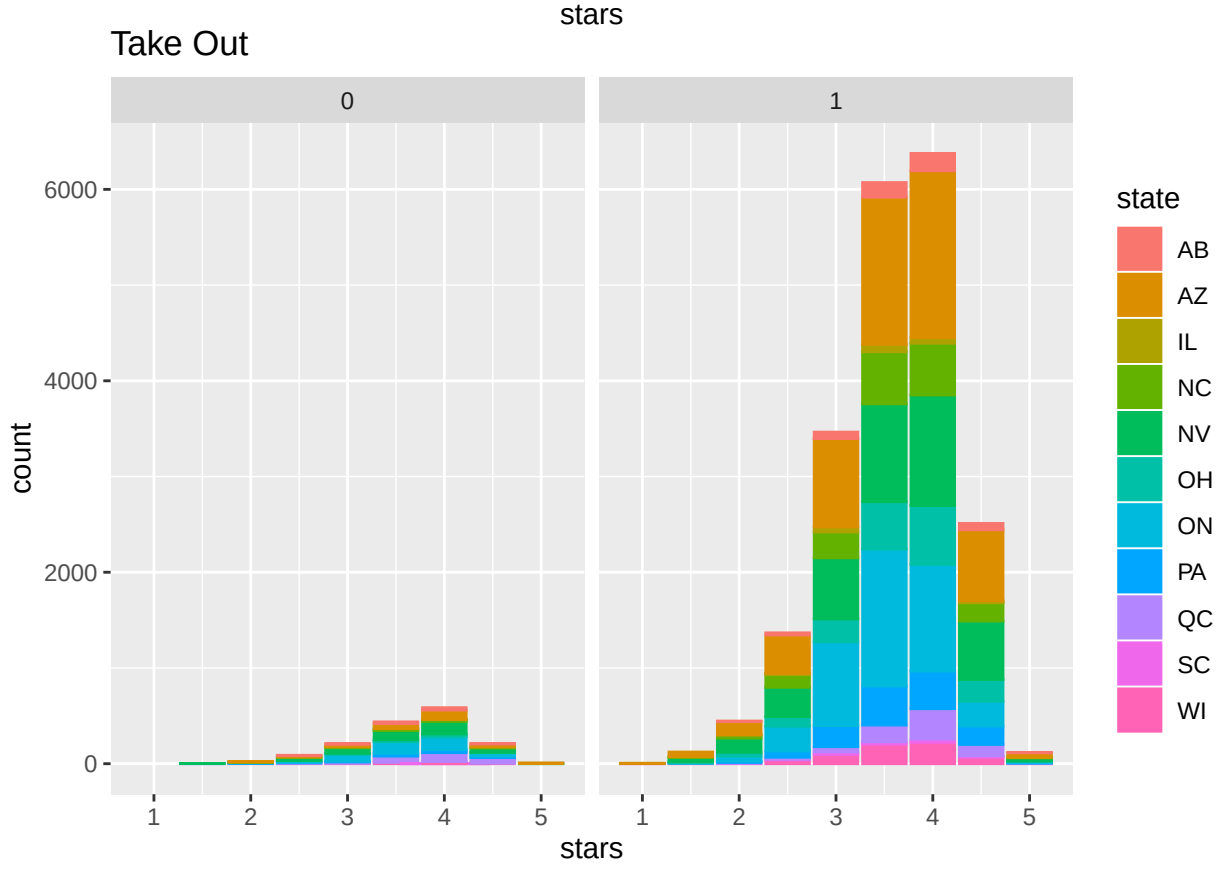
We try to see how different attributs of the restaurant contribute the influence to the rating scores.

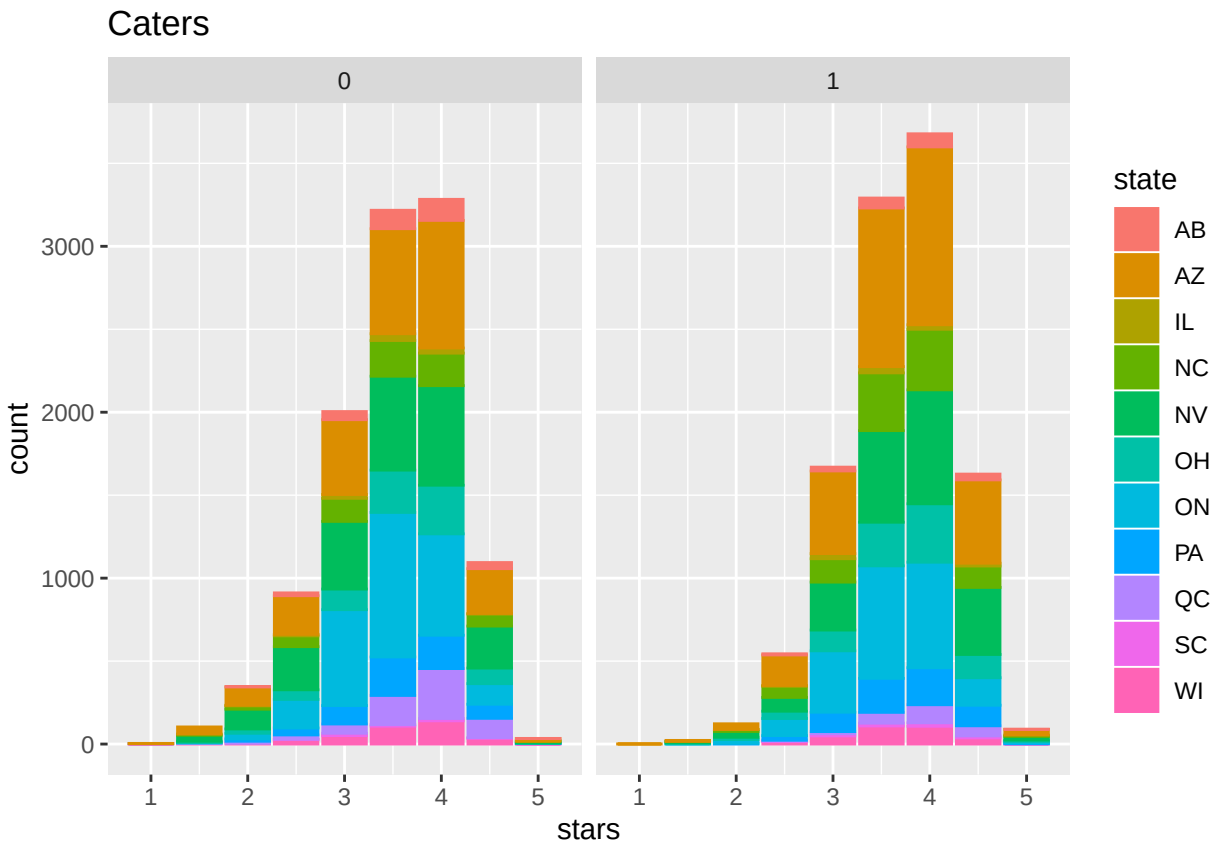



HasTV
0
1
4000
3000
2000
1000
0
count
1
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI


PriceRange
1
2
3
4
4000
3000
2000
1000
0
count
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

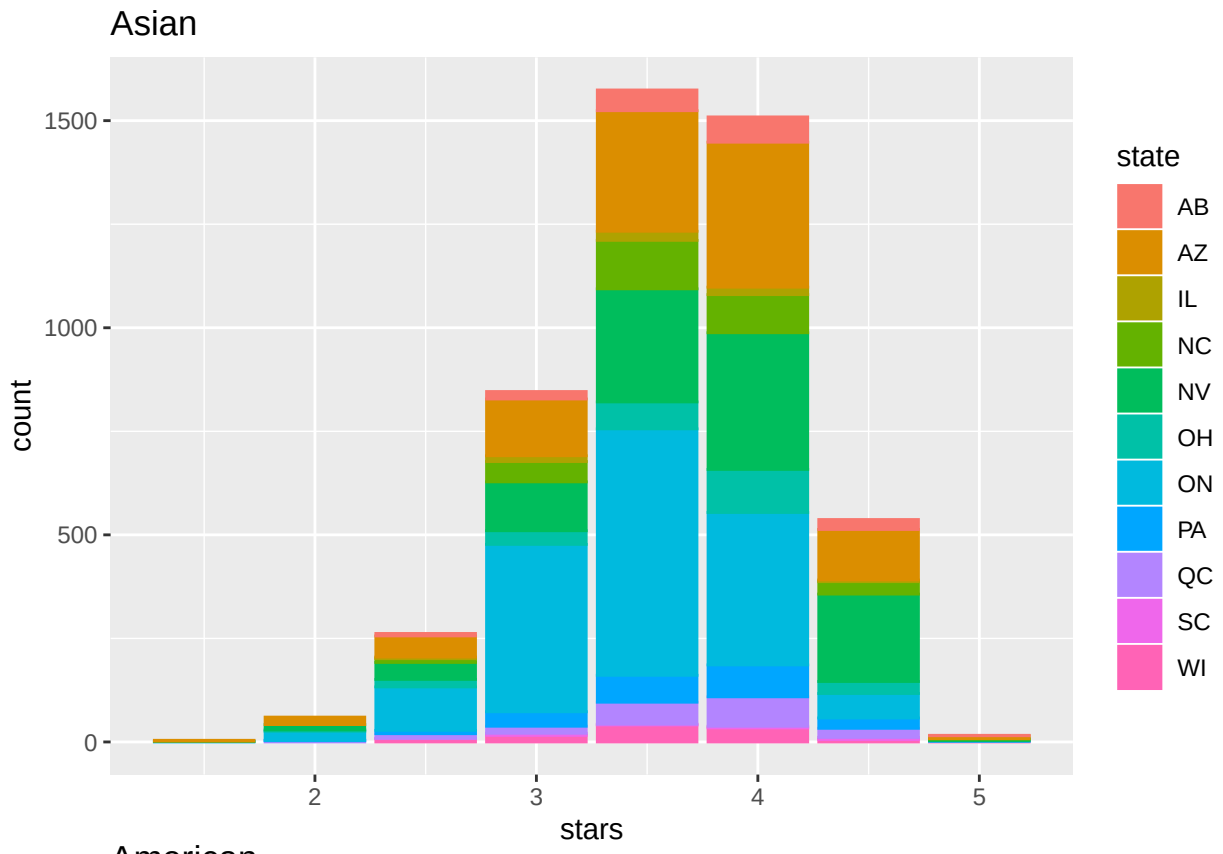

Take Out
0
1
count
6000
4000
2000
0
1
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI

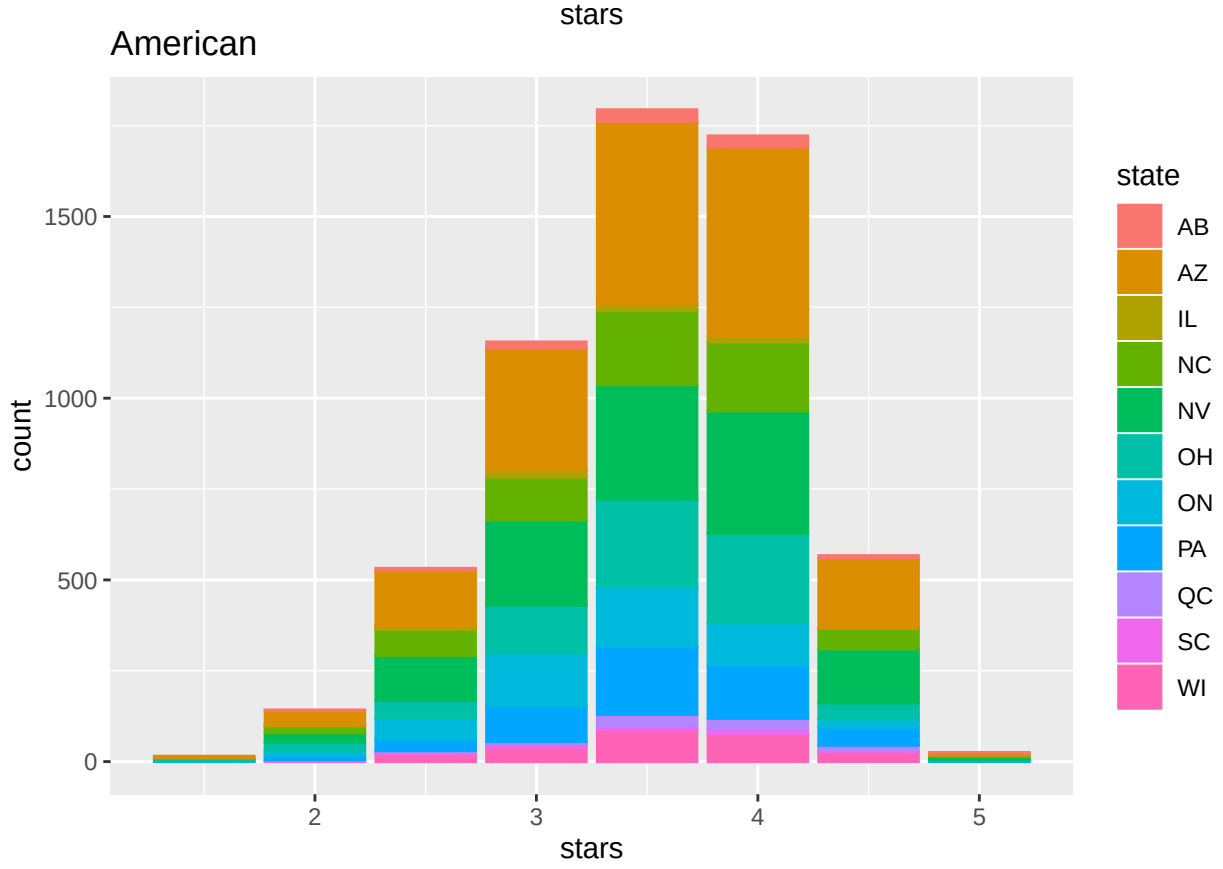


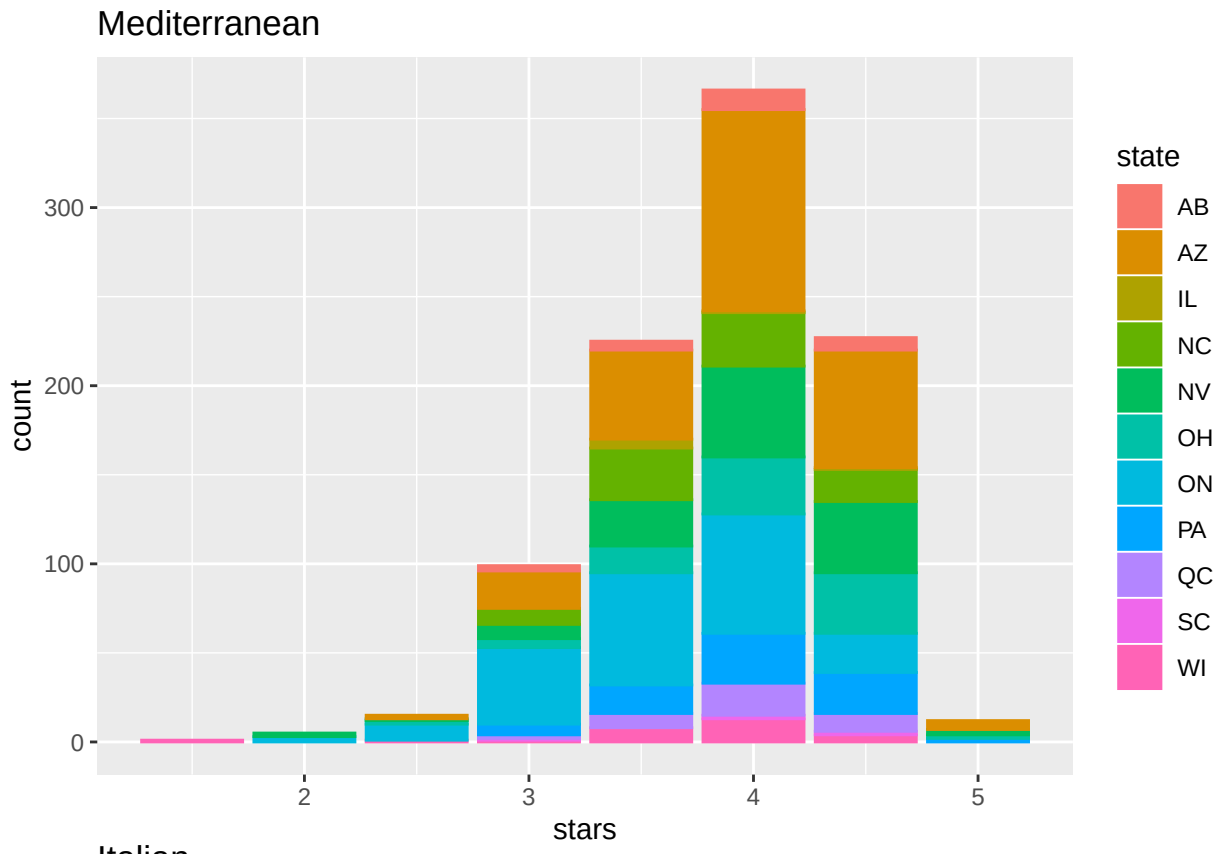
From above plots, we can see that the overall star distribution in each cateogries of the attributes didn't show much difference, however, we still can see the star distribution differences in difference state.

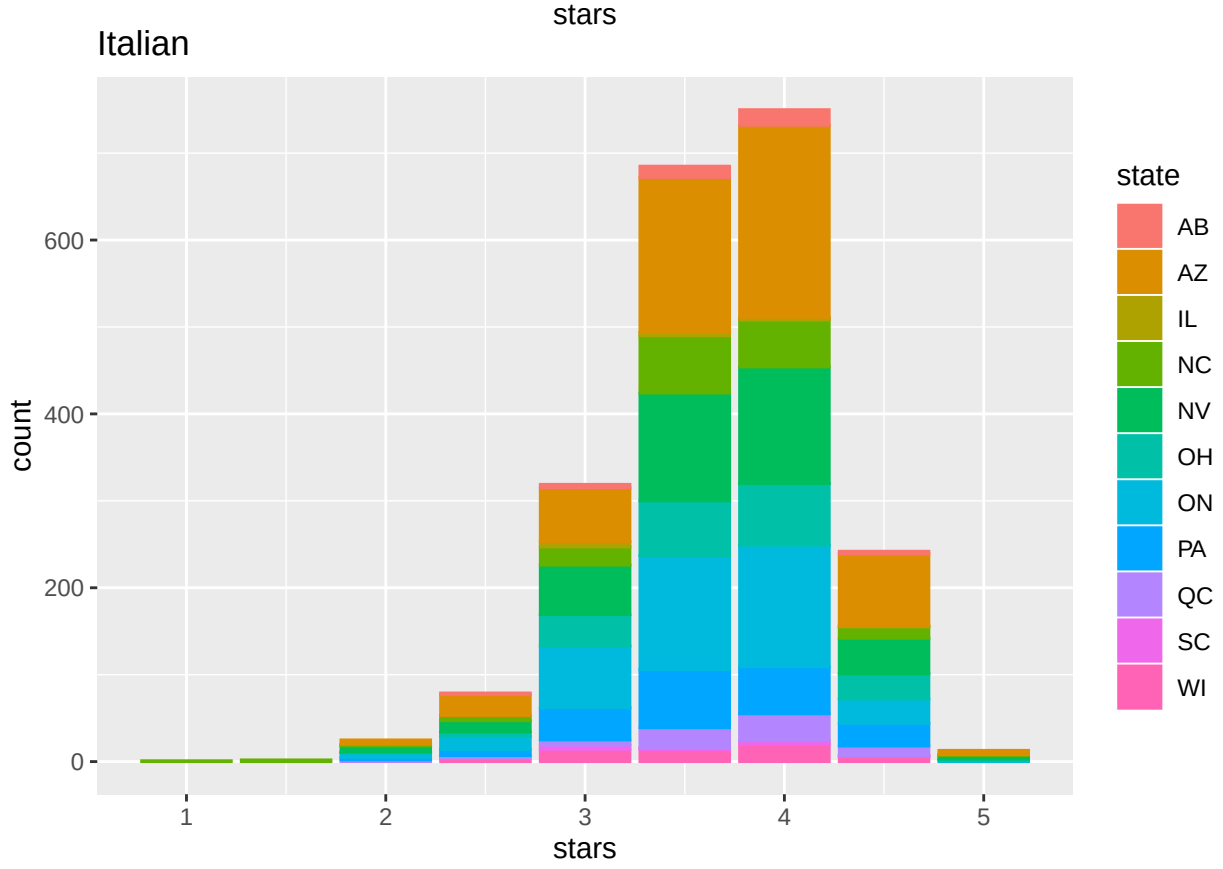
Rating stars in different types of cuisine- Asian, American, Mediterranean, Italian, Mexican

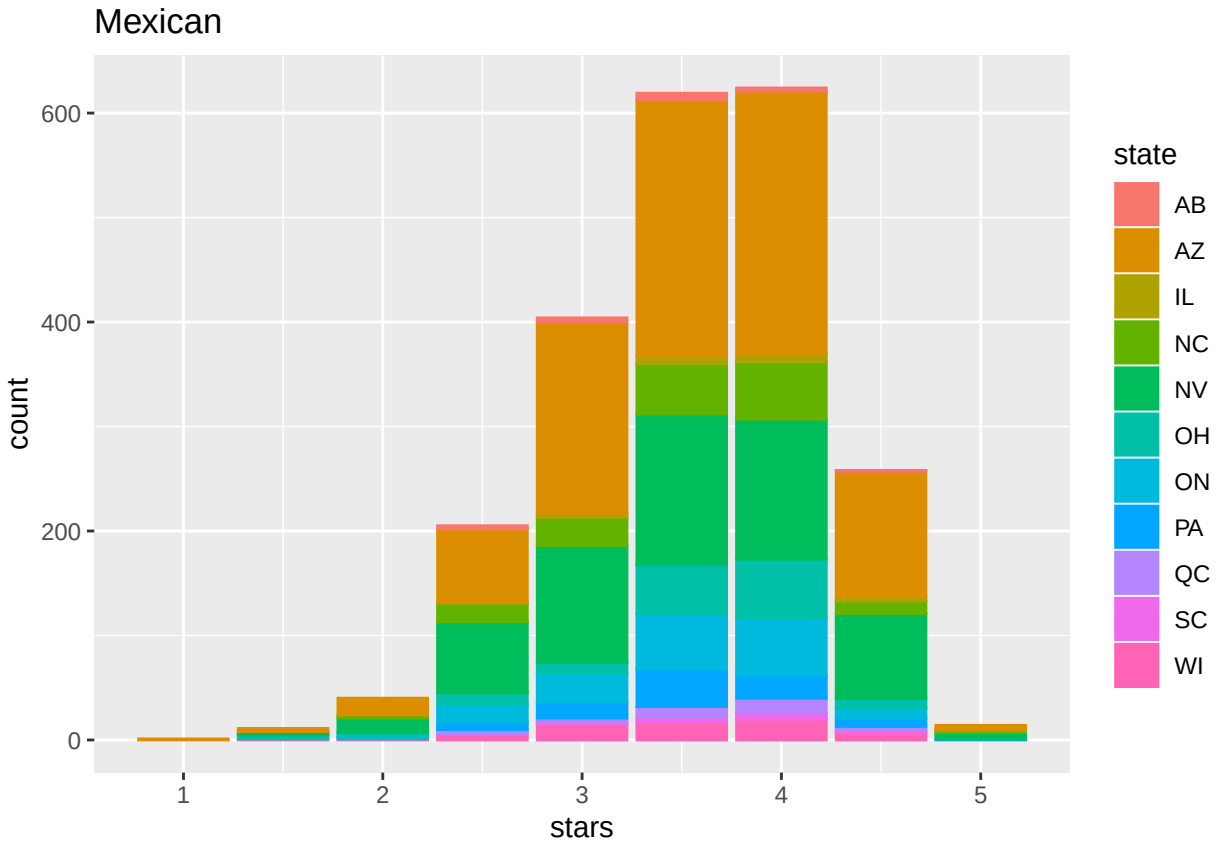
We want to see if the rating distribution would be different in different restaurant categoires.


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


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

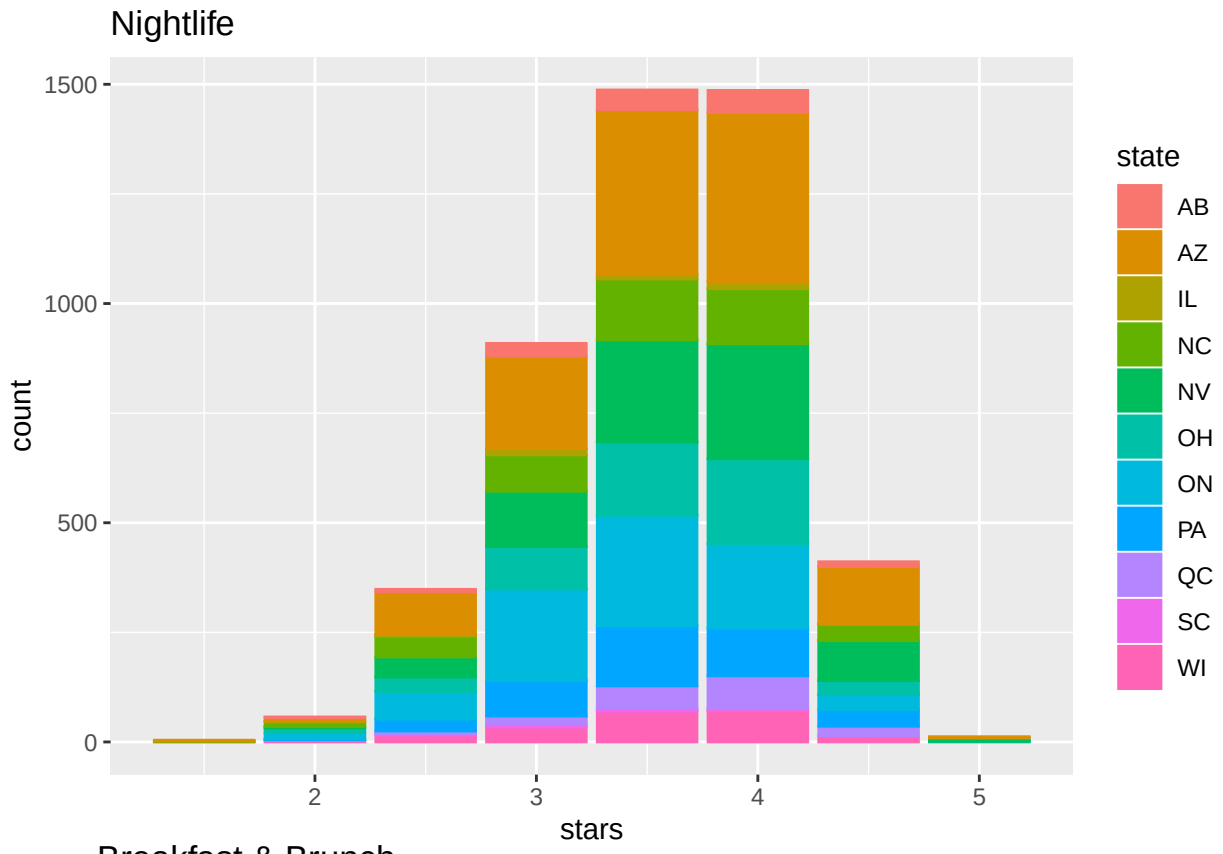

Mediterranean
count
300
200
100
0
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI

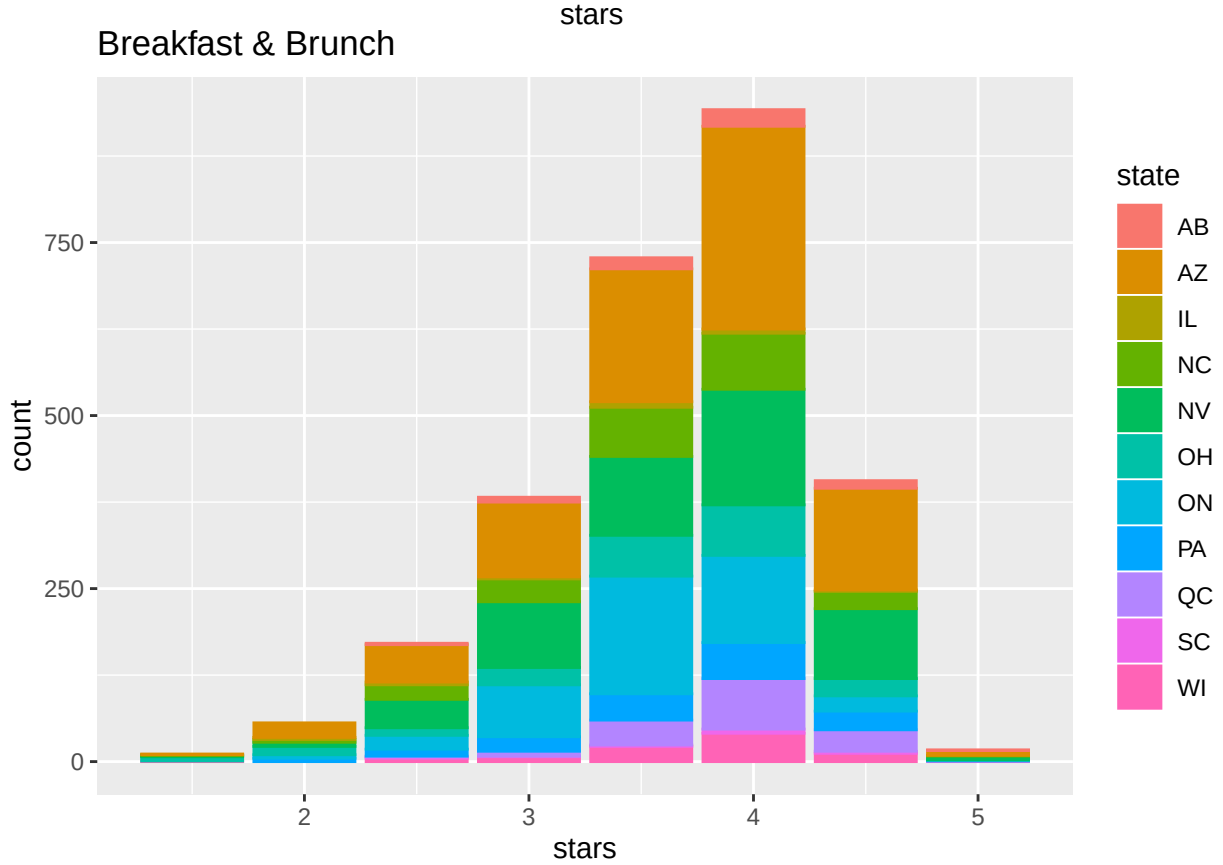

Italian
count
600
400
200
0
1
2
3
4
5
stars
state
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI



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. 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 ~ NoiseLevel + review_count + trendy + Caters +
##     state + hipster + intimate + Alcohol + Mediterranean + n_parks +
##     classy + casual + PriceRange + divey + American + touristy +
##     is_open + Nightlife + romantic + Italian + TakeOut + Breakfast_Brunch,
##     data = aic_table)
```

From the above result, we choose NoiseLevel , review_count, trendy, Caters, state, hipster, intimate, Alcohol, Mediterranean, n_parks, classy, casual, PriceRange, divey, American, touristy, is_open, Nightlife, romantic, Italian, TakeOut, Breakfast_Brunch as independent varaibles

Fit lmer model to treat state as group

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

```
fit1 <-  lmer (data= bs_model2, stars ~ NoiseLevel + review_count + trendy + Caters + hipster + intimat
    is_open + Nightlife + romantic + Italian + TakeOut + Breakfast_Brunch + (1|state))
display(fit1)
```

```
## lmer(formula = stars ~ NoiseLevel + review_count + trendy + Caters +
##     hipster + intimate + Alcohol + Mediterranean + n_parks +
##     classy + casual + PriceRange + divey + American + touristy +
##     is_open + Nightlife + romantic + Italian + TakeOut + Breakfast_Brunch +
##     (1 | state), data = bs_model2)
##                       coef.est coef.se
## (Intercept)            3.36     0.04
## NoiseLevelloud        -0.24     0.02
## NoiseLevelquiet        0.18     0.01
## NoiseLevelvery_loud   -0.45     0.03
## review_count           0.00     0.00
## trendy                 0.25     0.01
## Caters                 0.14     0.01
## hipster                0.32     0.02
## intimate               0.34     0.03
## Alcoholfull_bar       -0.17     0.01
## Alcoholnone           -0.03     0.01
## Mediterranean          0.22     0.02
## n_parks                0.06     0.01
## classy                 0.18     0.02
## casual                 0.16     0.01
## PriceRange2            0.00     0.01
## PriceRange3            0.14     0.02
## PriceRange4            0.30     0.04
## divey                  0.24     0.03
## American              -0.08     0.01
## touristy              -0.24     0.04
## is_open                0.05     0.01
## Nightlife              0.06     0.01
```

```
## romantic             0.12     0.03
## Italian              0.05     0.01
## TakeOut             -0.06     0.02
## Breakfast_Brunch     0.04     0.01
##
## Error terms:
##  Groups   Name        Std.Dev.
##  state    (Intercept) 0.10
##  Residual             0.57
## ---
## number of obs: 22041, groups: state, 11
## AIC = 37707.4, DIC = 37272.6
## deviance = 37461.0
```

```
summary(fit1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: stars ~ NoiseLevel + review_count + trendy + Caters + hipster +
##     intimate + Alcohol + Mediterranean + n_parks + classy + casual +
##     PriceRange + divey + American + touristy + is_open + Nightlife +
##     romantic + Italian + TakeOut + Breakfast_Brunch + (1 | state)
##    Data: bs_model2
##
## REML criterion at convergence: 37649.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.1428 -0.6013  0.0650  0.6958  3.7363
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.01039  0.1019
##  Residual             0.32018  0.5658
## Number of obs: 22041, groups:  state, 11
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          3.3584246  0.0396393  84.725
## NoiseLevelloud      -0.2371172  0.0152034 -15.596
## NoiseLevelquiet      0.1798720  0.0110574  16.267
## NoiseLevelvery_loud -0.4489167  0.0285904 -15.702
## review_count         0.0004002  0.0000161  24.849
## trendy               0.2497292  0.0139313  17.926
## Caters               0.1368002  0.0080216  17.054
## hipster              0.3201336  0.0209101  15.310
## intimate             0.3372148  0.0277555  12.149
## Alcoholfull_bar     -0.1650547  0.0121501 -13.585
## Alcoholnone         -0.0272059  0.0120295  -2.262
## Mediterranean        0.2239448  0.0189347  11.827
## n_parks              0.0637756  0.0072604   8.784
## classy               0.1813524  0.0220142   8.238
## casual               0.1562028  0.0106221  14.705
## PriceRange2          0.0024323  0.0104293   0.233
## PriceRange3          0.1403548  0.0225217   6.232
## PriceRange4          0.2953966  0.0442623   6.674
```

```
## divey                0.2419891  0.0254895   9.494
## American            -0.0763846  0.0097611  -7.825
## touristy            -0.2409211  0.0425165  -5.667
## is_open              0.0536029  0.0097904   5.475
## Nightlife            0.0571229  0.0113070   5.052
## romantic             0.1169466  0.0289355   4.042
## Italian              0.0502717  0.0135087   3.721
## TakeOut             -0.0634805  0.0172857  -3.672
## Breakfast_Brunch     0.0400576  0.0120636   3.321

##
## Correlation matrix not shown by default, as p = 27 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

Fit lmer model to treat state and city as group

Try to use both city and state as group

```
fit2 <-  lmer (data= bs_model2, stars ~ NoiseLevel + review_count + trendy + Caters + hipster + intimat
    is_open + Nightlife + romantic + Italian + TakeOut + Breakfast_Brunch + (1|state) + (1|city))
display(fit2)
```

```
## lmer(formula = stars ~ NoiseLevel + review_count + trendy + Caters +
##     hipster + intimate + Alcohol + Mediterranean + n_parks +
##     classy + casual + PriceRange + divey + American + touristy +
##     is_open + Nightlife + romantic + Italian + TakeOut + Breakfast_Brunch +
##     (1 | state) + (1 | city), data = bs_model2)
##                     coef.est coef.se
## (Intercept)          3.36     0.04
## NoiseLevelloud      -0.23     0.02
## NoiseLevelquiet      0.18     0.01
## NoiseLevelvery_loud -0.44     0.03
## review_count         0.00     0.00
## trendy               0.25     0.01
## Caters               0.14     0.01
## hipster              0.31     0.02
## intimate             0.33     0.03
## Alcoholfull_bar     -0.17     0.01
## Alcoholnone         -0.03     0.01
## Mediterranean        0.22     0.02
## n_parks              0.06     0.01
## classy               0.18     0.02
## casual               0.16     0.01
## PriceRange2          0.00     0.01
## PriceRange3          0.14     0.02
## PriceRange4          0.29     0.04
## divey                0.24     0.03
## American            -0.08     0.01
## touristy            -0.25     0.04
## is_open              0.06     0.01
## Nightlife            0.06     0.01
## romantic             0.12     0.03
## Italian              0.05     0.01
## TakeOut             -0.06     0.02
## Breakfast_Brunch     0.04     0.01
```

```
##
## Error terms:
##  Groups   Name        Std.Dev.
##  city     (Intercept) 0.12
##  state    (Intercept) 0.07
##  Residual             0.56
## ---
## number of obs: 22041, groups: city, 362; state, 11
## AIC = 37546, DIC = 37108.2
## deviance = 37297.1
```

```
summary(fit2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: stars ~ NoiseLevel + review_count + trendy + Caters + hipster +
##     intimate + Alcohol + Mediterranean + n_parks + classy + casual +
##     PriceRange + divey + American + touristy + is_open + Nightlife +
##     romantic + Italian + TakeOut + Breakfast_Brunch + (1 | state) +
##     (1 | city)
##    Data: bs_model2
##
## REML criterion at convergence: 37486
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.1038 -0.5931  0.0694  0.6921  3.7503
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  city     (Intercept) 0.015497 0.12449
##  state    (Intercept) 0.005186 0.07202
##  Residual             0.315474 0.56167
## Number of obs: 22041, groups:  city, 362; state, 11
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          3.363e+00  3.630e-02  92.628
## NoiseLevelloud      -2.341e-01  1.513e-02 -15.477
## NoiseLevelquiet      1.773e-01  1.101e-02  16.111
## NoiseLevelvery_loud -4.446e-01  2.843e-02 -15.636
## review_count         3.957e-04  1.608e-05  24.599
## trendy               2.476e-01  1.387e-02  17.854
## Caters               1.359e-01  7.991e-03  17.003
## hipster              3.084e-01  2.085e-02  14.794
## intimate             3.336e-01  2.762e-02  12.080
## Alcoholfull_bar     -1.723e-01  1.212e-02 -14.213
## Alcoholnone         -2.736e-02  1.198e-02  -2.284
## Mediterranean        2.222e-01  1.884e-02  11.793
## n_parks              6.180e-02  7.250e-03   8.525
## classy               1.808e-01  2.192e-02   8.249
## casual               1.603e-01  1.058e-02  15.148
## PriceRange2          1.701e-03  1.039e-02   0.164
## PriceRange3          1.379e-01  2.241e-02   6.152
## PriceRange4          2.950e-01  4.401e-02   6.703
## divey                2.367e-01  2.541e-02   9.317
```

```
## American            -7.674e-02  9.724e-03  -7.892
## touristy            -2.459e-01  4.228e-02  -5.815
## is_open              5.748e-02  9.777e-03   5.879
## Nightlife            5.575e-02  1.125e-02   4.954
## romantic             1.190e-01  2.879e-02   4.134
## Italian              4.721e-02  1.345e-02   3.509
## TakeOut             -5.816e-02  1.720e-02  -3.382
## Breakfast_Brunch     3.510e-02  1.202e-02   2.920

##
## Correlation matrix not shown by default, as p = 27 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

From the model we can see that, the Noiselevel, Alcohol, American restaurant, touristy, and TakeOut will have negtive influence of the rating stars, among them NoiseLevel as very_loud is the most significant one. Among all positive influence, review_count will be the most significant one.

Model Validation

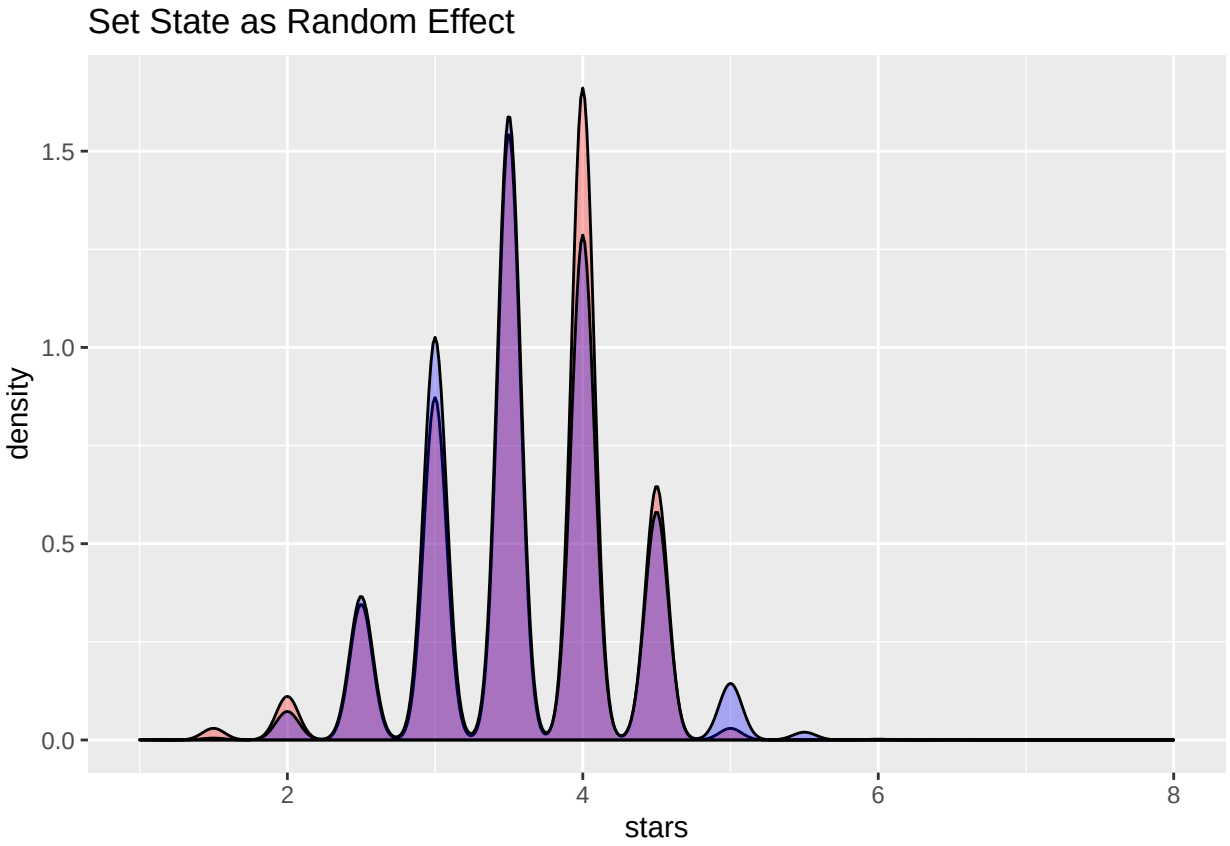
AIC check

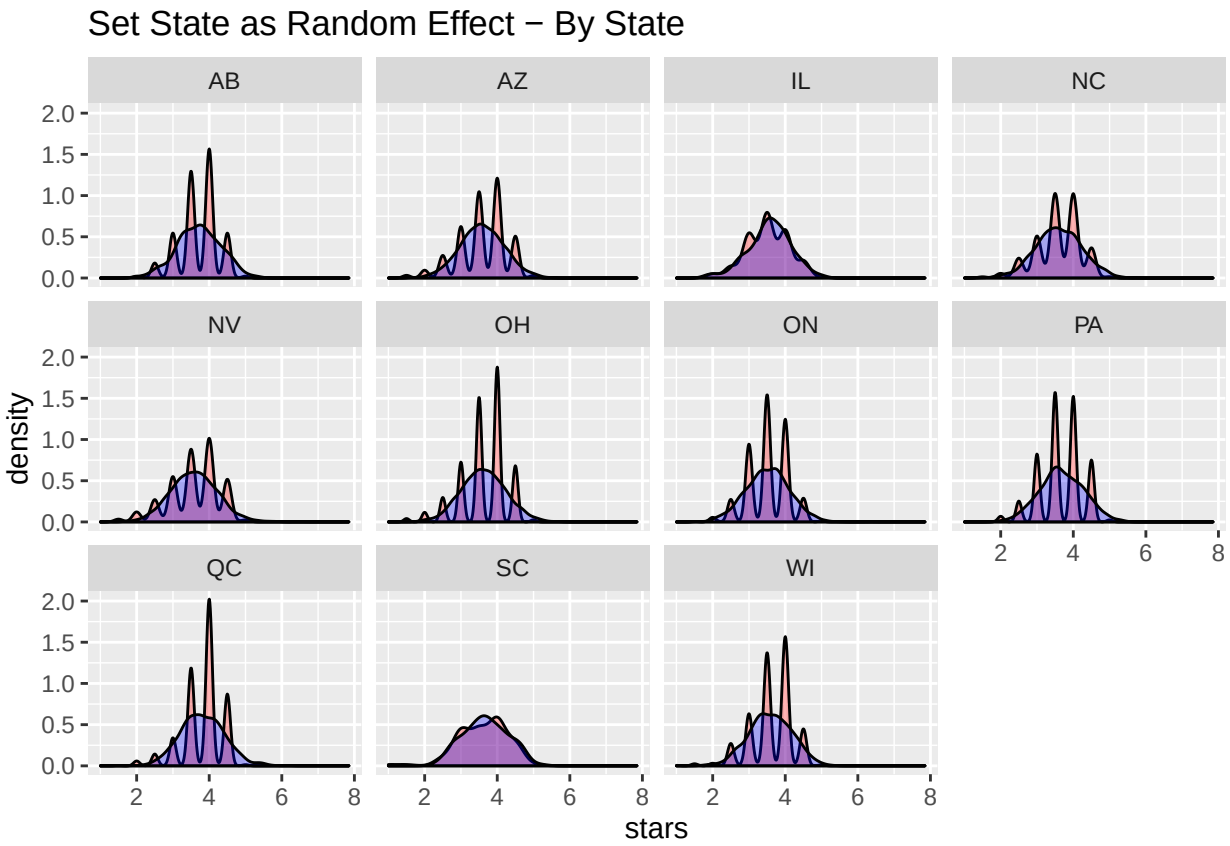
```
##      df      AIC
## fit1 29 37707.36
## fit2 30 37545.98
```

As can see from the result, using City and State as random effect improved the model.

Model use State as Random Effect

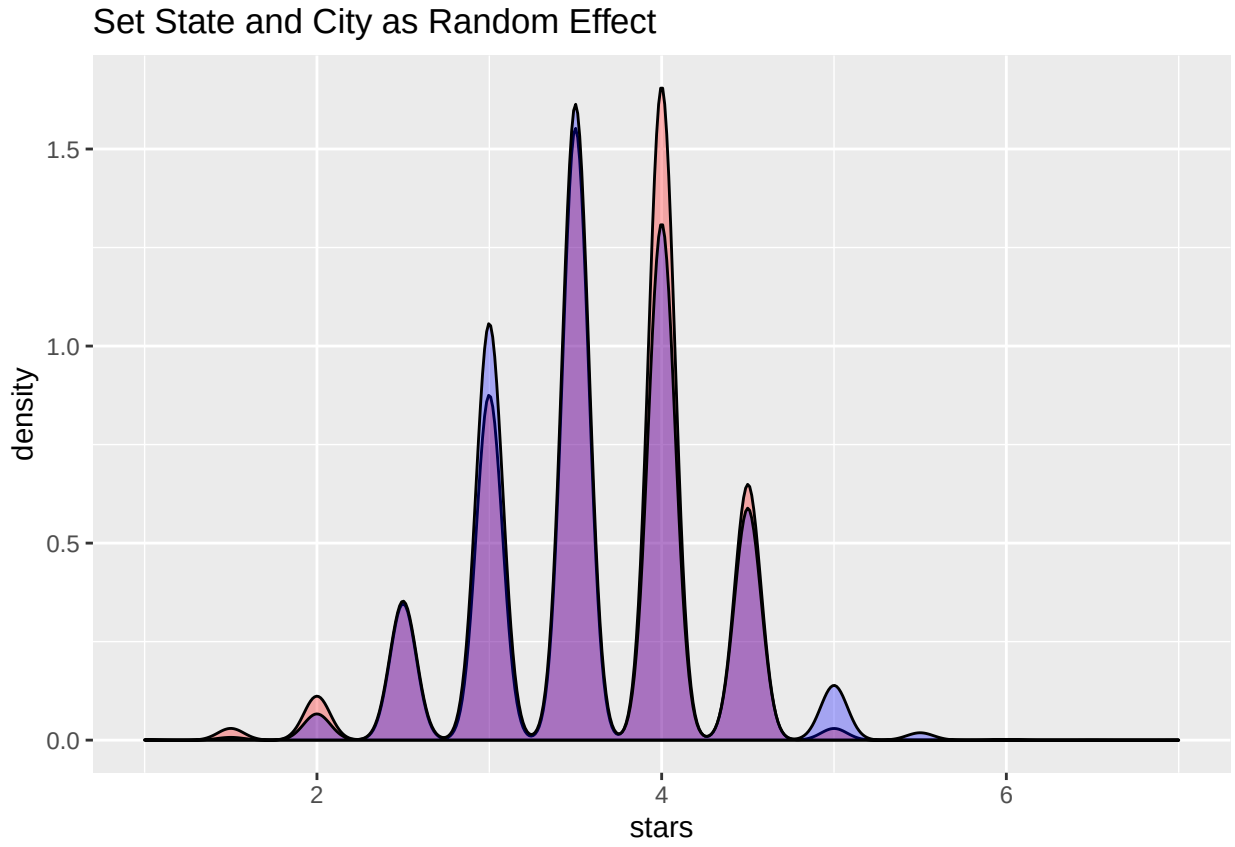
- Red plot is the observation value and the blue plot is for the prediction.


Set State as Random Effect
density
1.5
1.0
0.5
0.0
2
4
6
8
stars


Set State as Random Effect − By State
AB
AZ
IL
NC
NV
OH
ON
PA
QC
SC
WI
density
2.0
1.5
1.0
0.5
0.0
2
4
6
8
stars

Model use State and City as Random Effect

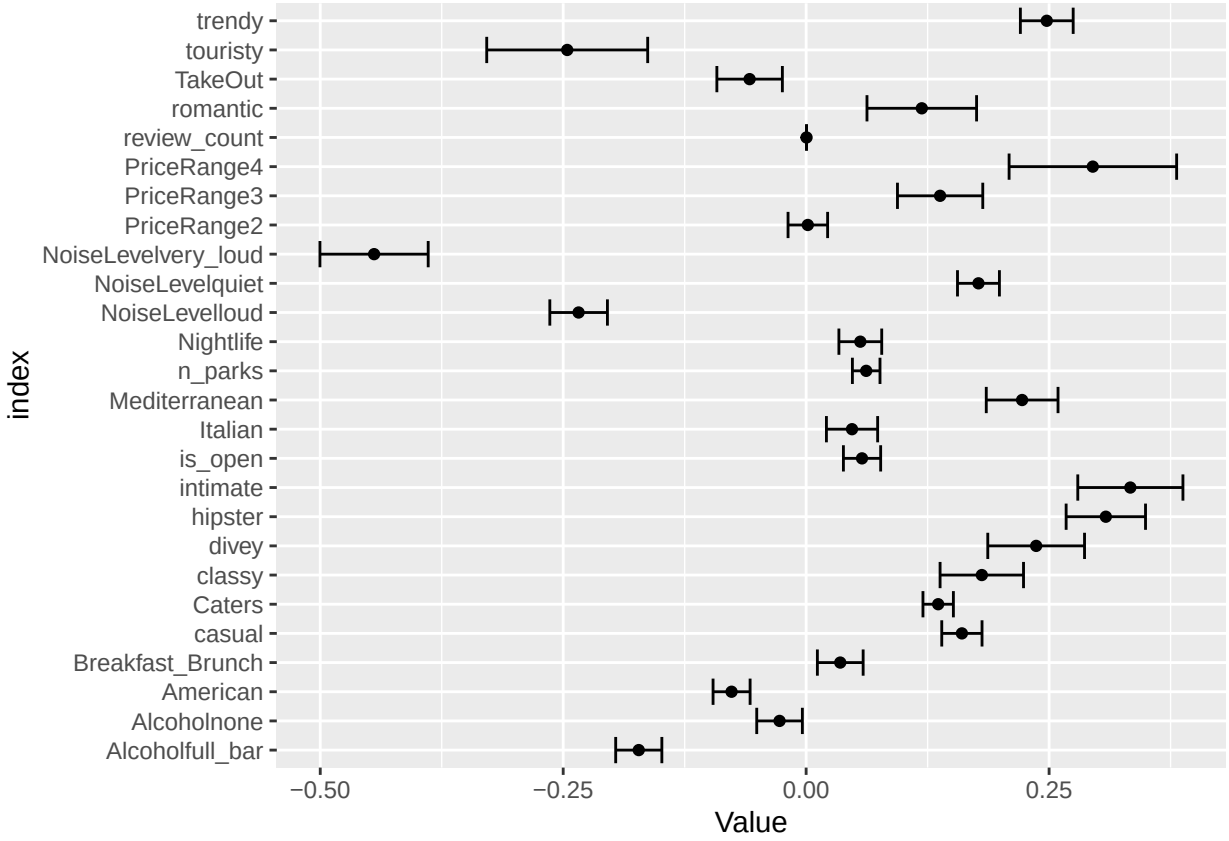
- Red plot is the observation value and the blue plot is for the prediction



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.

Validate Fix-effect for model fit2

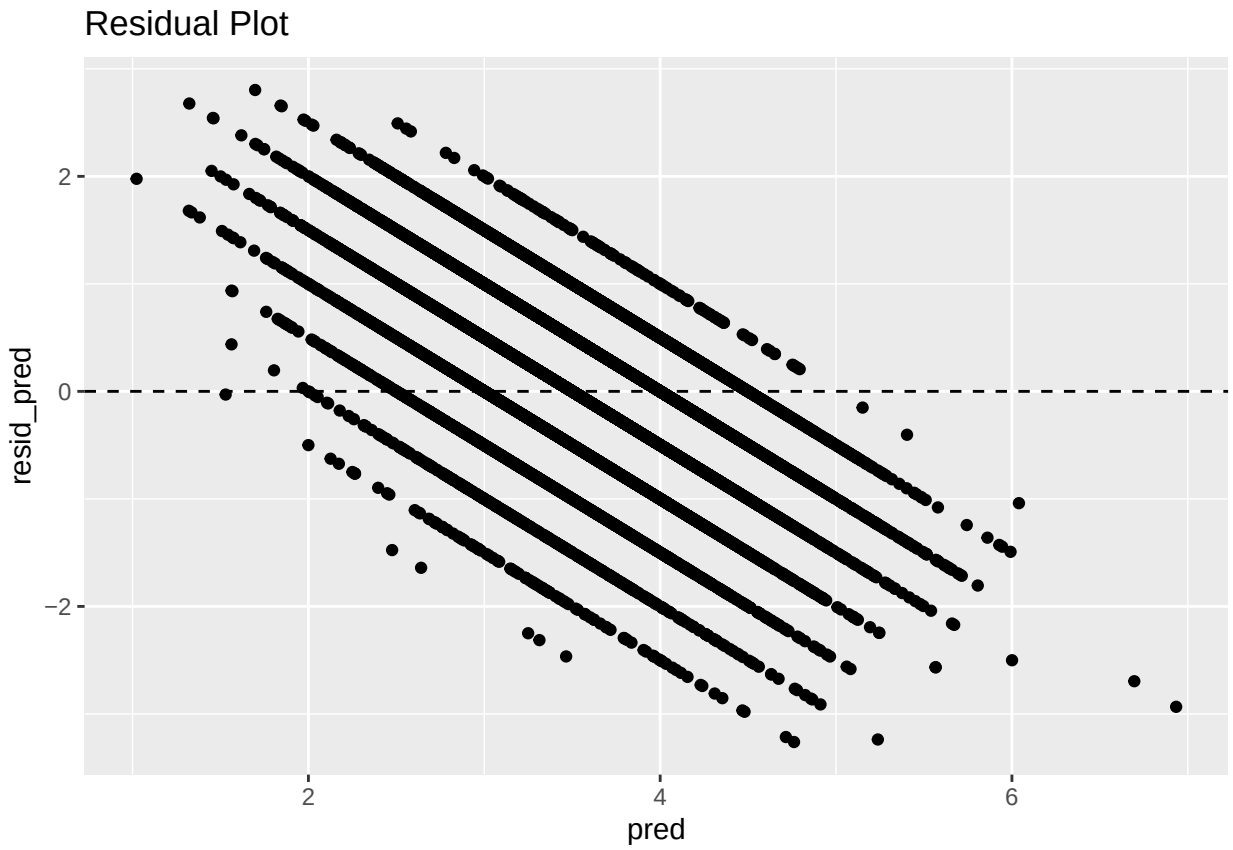
```
## Computing profile confidence intervals ...
```



From the fixed effect coefficeint value plot, we can see that PriceRange2 doesn't have influence on the rating score. Even thou the reveiw_count is significant, however it doesn't really help for predicting the score. 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 but Breakfast & Brunch would have a positive influence. This might be caught by the diversity of American food restaurant which would 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.


Residual Plot
resid_pred
2
0
−2
2
4
6
pred