678_project

Jingning Yang

11/29/2019

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

Since our dataset too large to run local, this file is run in our MSSP server.

Import Yelp Data: Business + Review

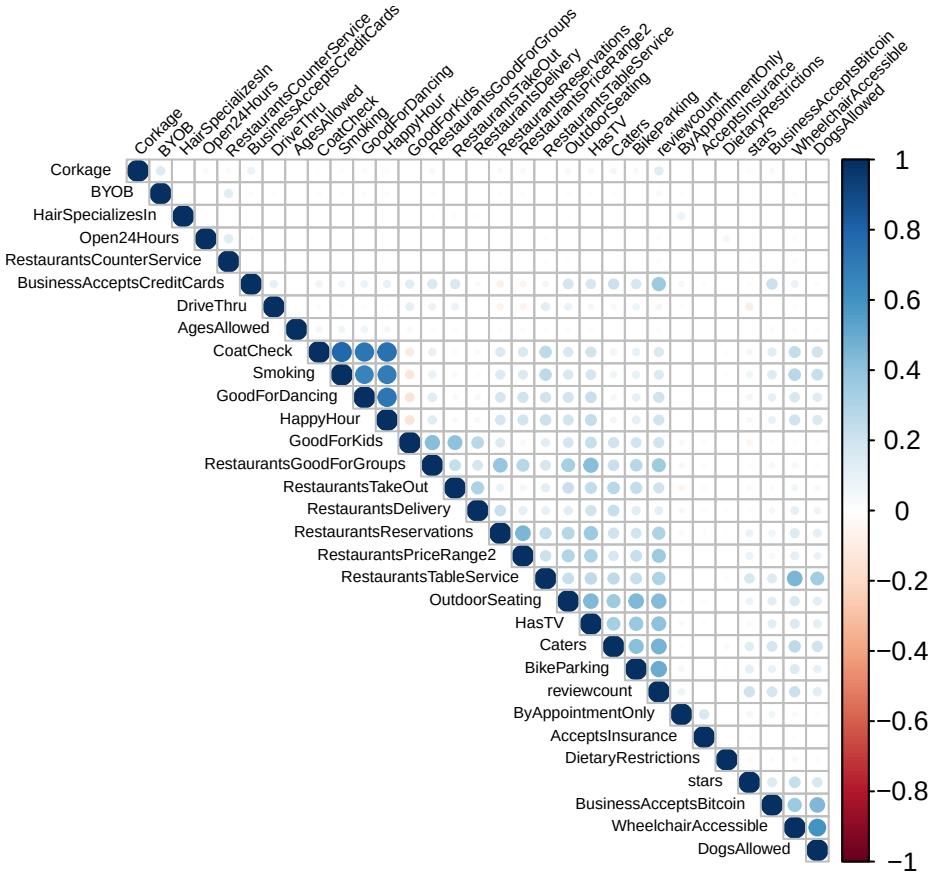
I will only use business.json and review.json for this report. And my goal is figuring out based on stars, is there any particular standard or requirement for a "good" restaurant(stars >=4) in different states? And it may help businesses change their stars in the short term by paying more attention to those standards/requirements in different states. Also, it may help visitors recognize a variety of tastes of each state so that they can avoid the situation: go to restaurants with high stars, but do not like its taste at all.

Clean Yelp Business Data

For business file, we only focus on stars, attributes, states and review counts, thus we delete other useless data.

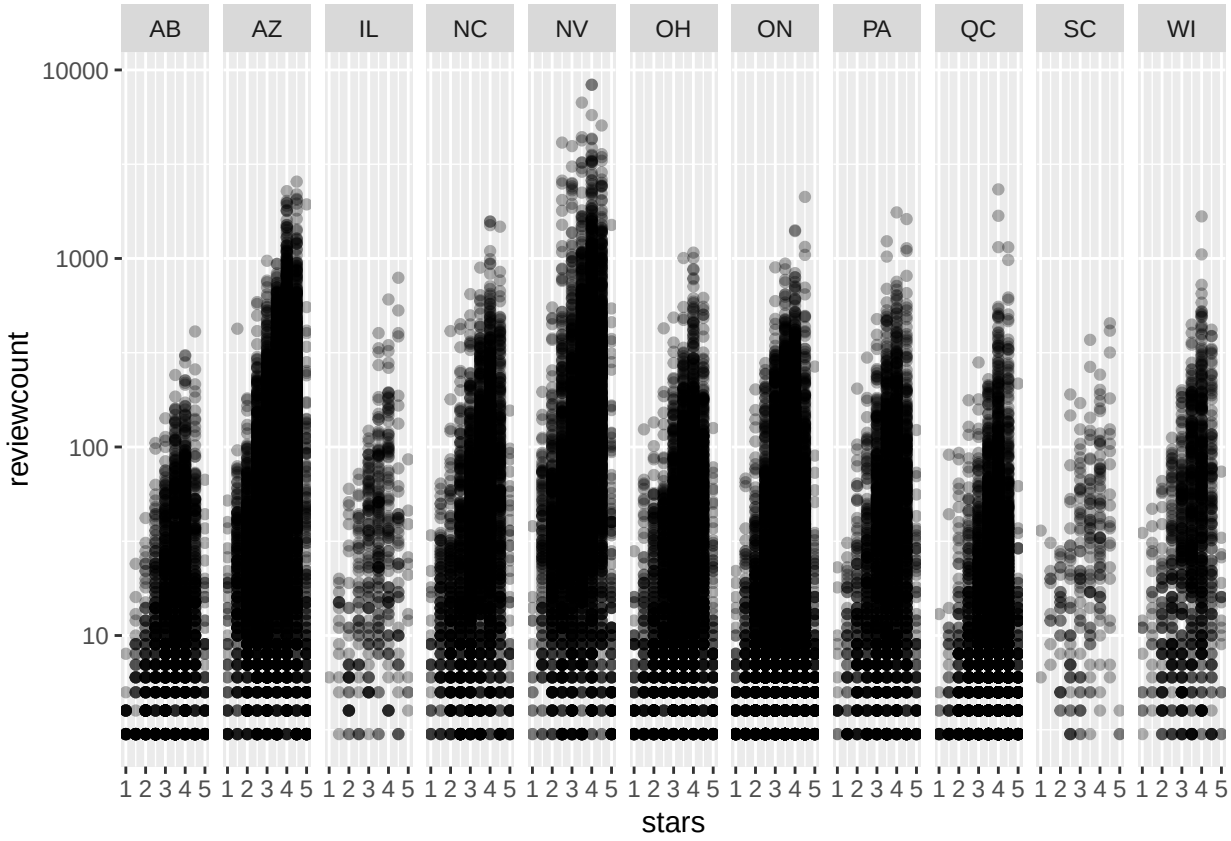
EDA for business data

By detecting correlated variables, the following plot can let us find some interesting correlations between variables:



We can see the relationship of other variables with stars is not too strong.

Now, we try to illustrate is there any relationship between review count and stars in different stars:



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

Using PCA reduce our dimension of attributes

Because there is a correlation between each variable as our correlation plot shows previously, we try PCA here to reduce the dimension so that we can choose independent variables into our model as predictors.

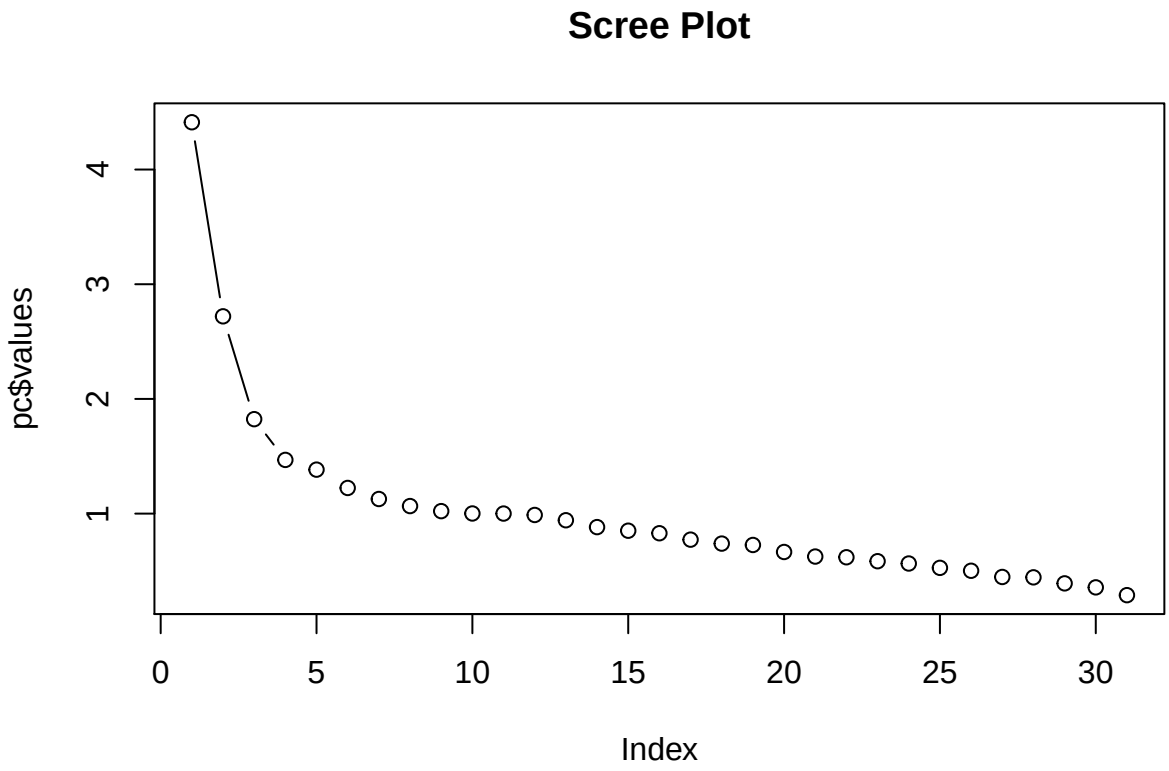
```
## R was not square, finding R from data

## $chisq
## [1] 235979.1
##
## $p.value
## [1] 0
##
## $df
## [1] 465

## [1] 0.0004861903
```

From output data, Bartlett's test is highly significant, asymptotically chi-square is 235979.1, and the P-value of chi-square smaller than 0.001, therefore factor analysis is appropriate.

And since the determinant is 0.001 larger than 0.00001, so our determinant does not seem problematic.

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

Factor extraction by using PCA:



From the Scree plot, x-axis: component number, y-axis:eigenvalues of principal components and factor analysis.

Since the elbow part is about the 4th point from the left, so the evidence from the scree plot and the eigenvalues suggests 4 component solutions may be the best.

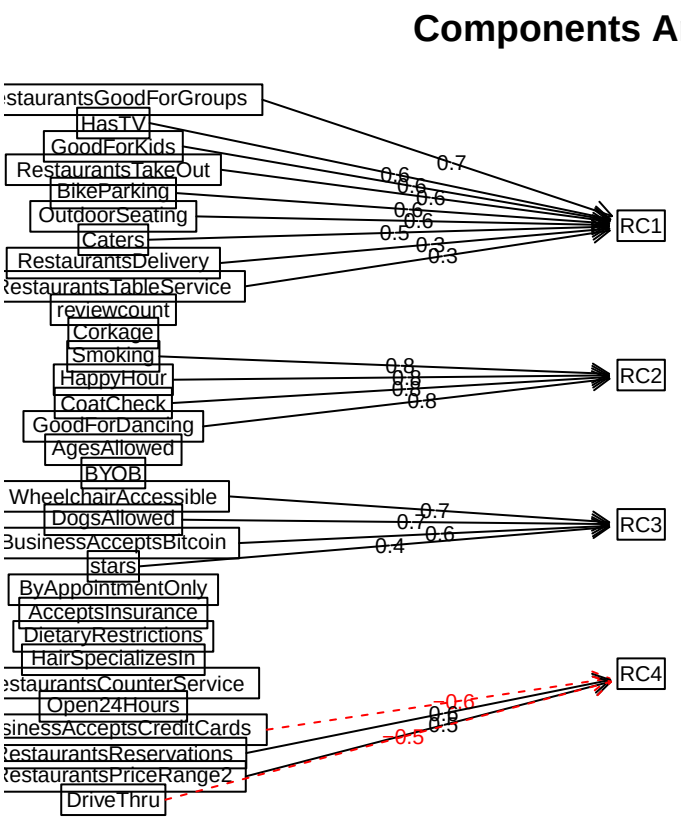
Thus, we choose 4 as our number of factors.

Redo PCA by using 4 factors:

Through output data, the Cumulative variable shows these 4 principle components explain 34% data with 31 variables. Specific output table placed in the Appendix.

For easier to explain the output of factor extraction, we can use orthogonal rotation to decreasing noise for factors as much as possible.

Orthogonal rotation(varimaax):



From the graph, we can see 31 variables can be separated as 4 groups, and "stars" belongs to the RC4.

Thus, we can choose the other 3 variables as a predictor in our model from RC1, RC2, and RC3 which has the highest correlation coefficient in its RC.

Model multilevel model in business file based on our EDA graphs:

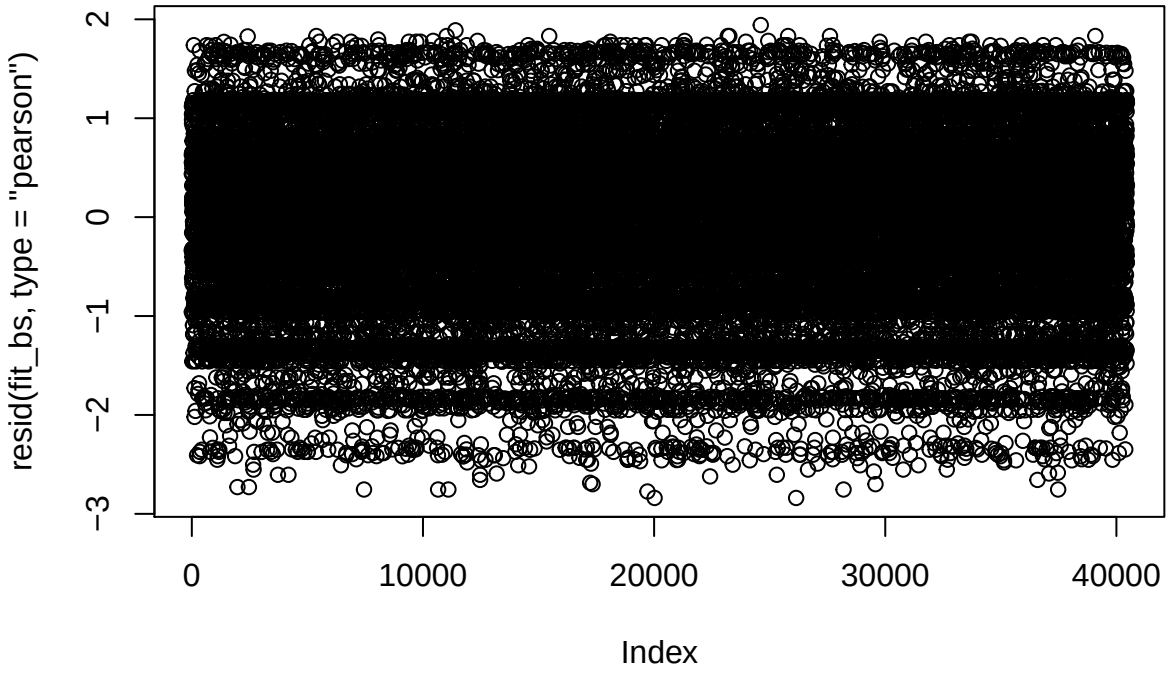
```
## boundary (singular) fit: see ?isSingular
```

```
##    (Intercept) RestaurantsReservations RestaurantsGoodForGroups HappyHour
## AB        3.32                    0.09                    -0.01      0.02
## AZ        3.35                    0.16                    -0.06      0.04
## IL        3.26                    0.12                    -0.02      0.02
## NC        3.23                    0.12                     0.00      0.02
## NV        3.26                    0.21                    -0.05      0.06
## OH        3.31                    0.11                    -0.01      0.02
## ON        3.35                    0.05                    -0.02      0.00
## PA        3.52                    0.08                    -0.05      0.02
## QC        3.75                    0.03                    -0.05      0.02
## SC        3.06                    0.18                    -0.01      0.04
## WI        3.44                    0.07                    -0.02      0.01
```

As the table shows, we can get functions for each states:
For state AB, stars = 3.32 + 0.09*RestaurantsReservations - 0.01*restaurantsGoodForGroups + 0.02*Happy-

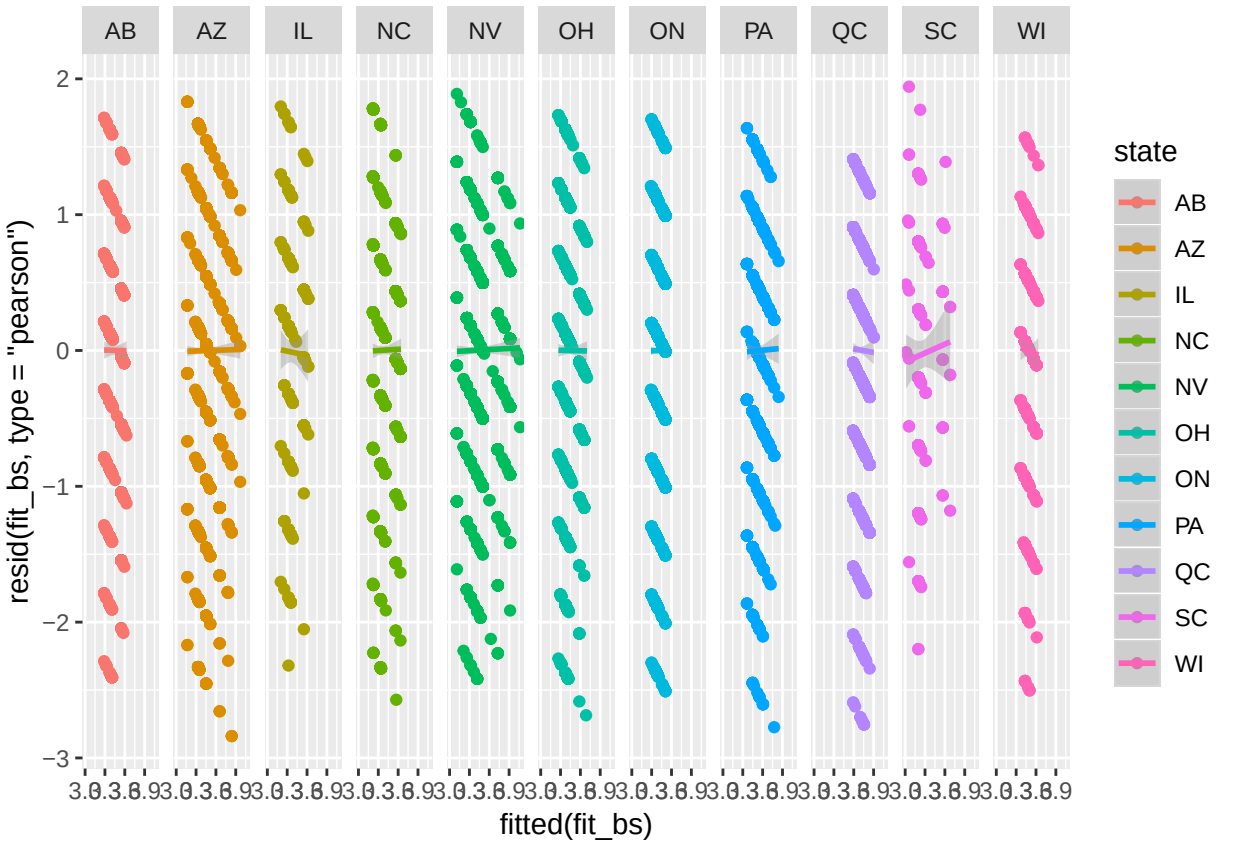
Hour.
And functions for other states are the same logic with state AB.

Model checking



From this plot, we can see the residuals randomly around y=0 and this suggests that the assumption that the relationship is linear is reasonable.

For more specific information for each states:



This difference pattern means our response might be the categorical number in the dataset, and the response in our model is continuous, therefore this pattern occurs. As we can see in the plot, most lines are horizontal at y=0. That means our assumption that the relationship is linear is reasonable as well.

QQ-plot of this model in the Appendix part.

Try to test accuracy of multilevel model by using 10-fold cross-validation

```
## [1] "the mean our our 10 MSE is: "
```

```
## [1] 0.6334376
```

Since our accuracy of the model by using 10-fold cross-validation is 63.4%, thus our model is kind of not good enough to cover most information in the whole dataset.

After model checking, try to explain the result of the model and explore it in the original dataset.

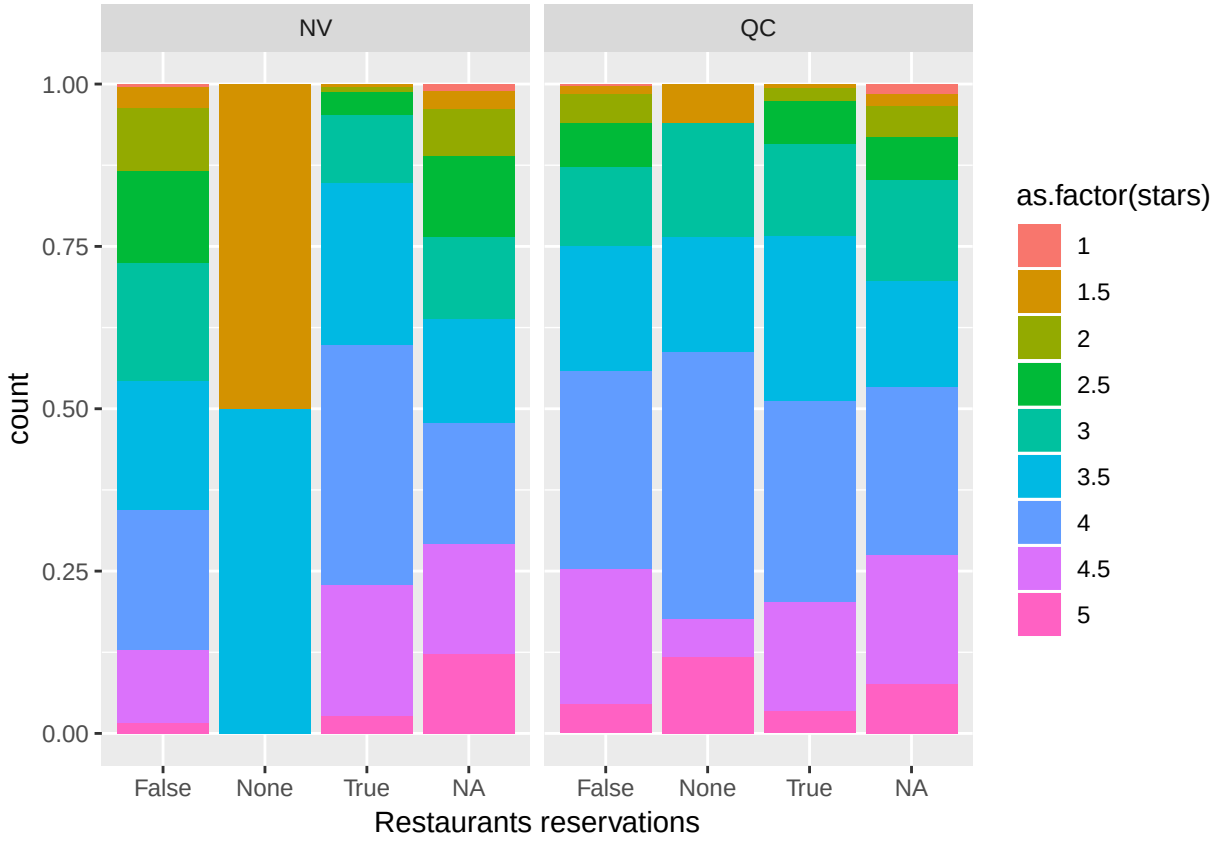
As previously table shows, we can get functions for each state:

```
##    (Intercept) RestaurantsReservations RestaurantsGoodForGroups HappyHour
## AB        3.32                    0.09                    -0.01      0.02
## AZ        3.35                    0.16                    -0.06      0.04
```

```
## IL       3.26                0.12                -0.02       0.02
## NC       3.23                0.12                 0.00       0.02
## NV       3.26                0.21                -0.05       0.06
## OH       3.31                0.11                -0.01       0.02
## ON       3.35                0.05                -0.02       0.00
## PA       3.52                0.08                -0.05       0.02
## QC       3.75                0.03                -0.05       0.02
## SC       3.06                0.18                -0.01       0.04
## WI       3.44                0.07                -0.02       0.01
```

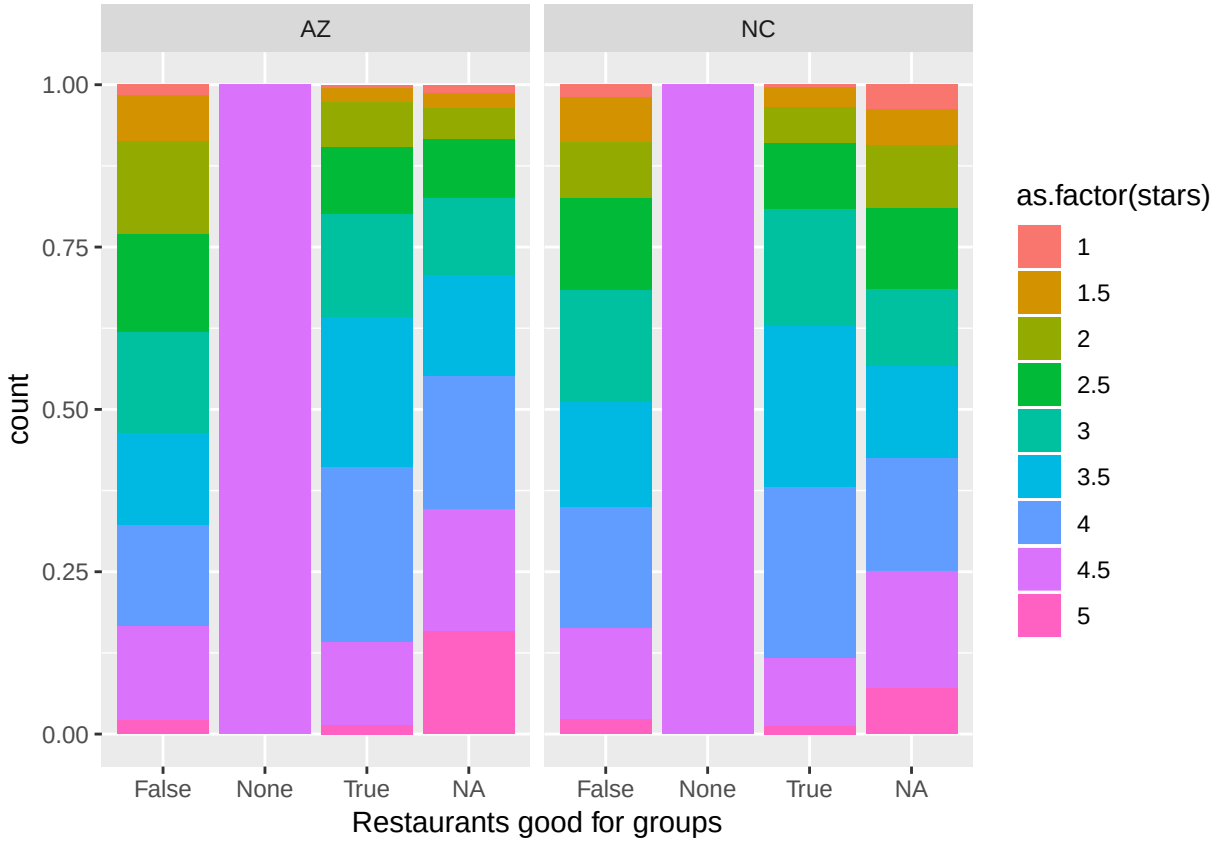
First, we can see the largest coefficient of Restaurant reservations is in state NV which is 0.21, and the lowest coefficient of it is in state QC which is 0.03. That means restaurant reservations affect the stars of restaurants in AZ more than in QC.

And we can explore it in the original dataset:



From this graph, we can see the difference between NV and QC, in QC, no matter which kind of restaurants reservations are, the largest portion always blue(star=4), and in NV, only when restaurants reservations is "True", we can see the largest portion is blue(star=4) apparently. That's why the result of our model said different kind of restaurants reservations affect restaurant stars in NV more than in QC.

Second, for the coefficient of RestaurantsGoodForGroups, it's kind of similar to each state since the maximum is 0 in NC and the minimum is -0.06 in AZ. From the following plot, we can see the ratio of stars in a different kind of "restaurants good for groups" in 2 states are pretty close, and it's hard to find a pattern between them that can explain why the difference of coefficient of RestaurantsGoodForGroups from our model is small.



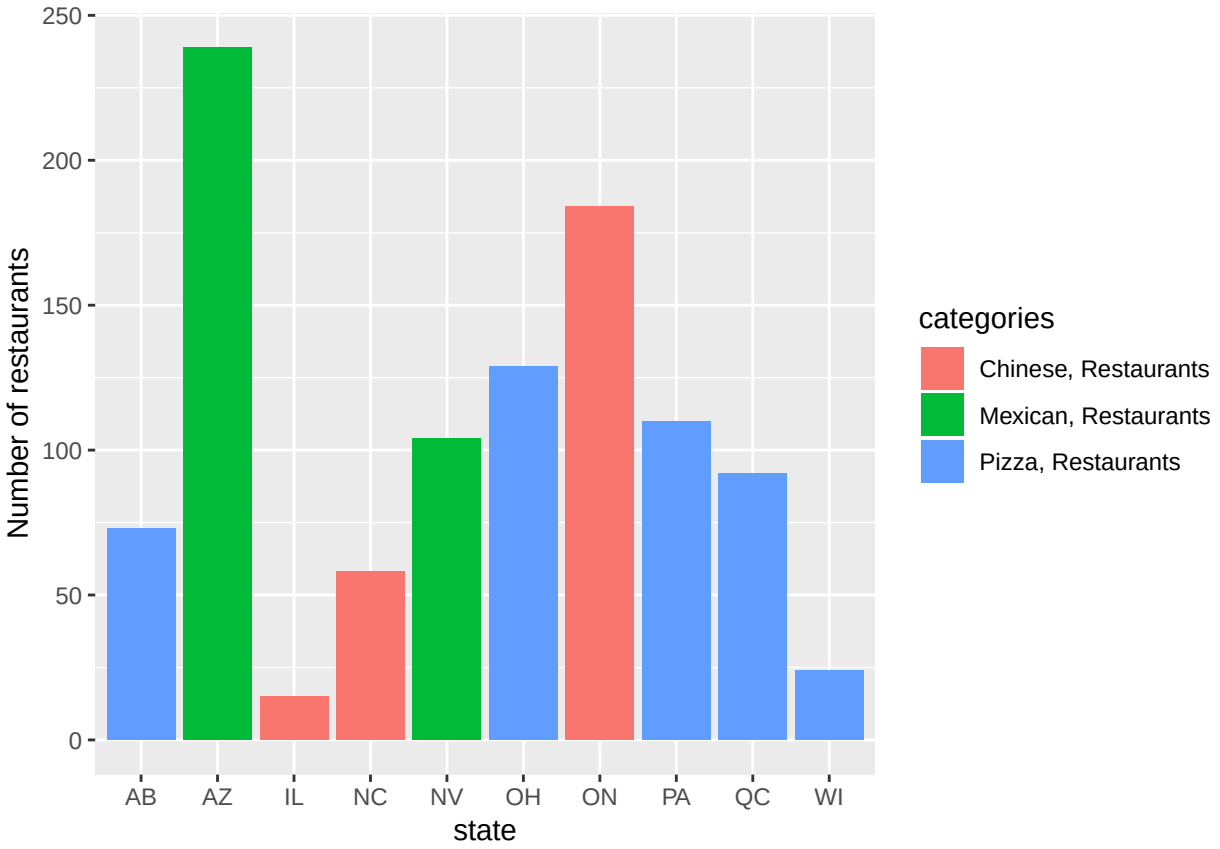
Third, the difference of coefficient in HappyHour is small as well, the maximum is 0.06 in NV and the minimum is 0 in ON. And I will place related plots in the Appendix part.

Consequently, for most states, there is no standard/requirement for "good" restaurants from the attribution part.

Next, is there any taste preference for a "good" restaurant in different states from our reveiw.json file?

Match reviews to each restaurant in the business file:

We try to figure out the top number of restaurants in every state:



From this plot, we can see there is 3 top number of restaurants: Chinese, Mexican, and Pizza.
Therefore, we can make a guess: Mexican restaurants are popular in-state AZ and NV. Pizza restaurants are popular in-state AB, OH, PA, QC, and WI. Chinese restaurants are popular in-state IL, NC and ON.

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

Text mining for feature reviews of restaurants with high stars(4~5) in different states.

Wordcloud for review features of good restaurants(stars > 4) in IL:

From the plot, the highest frequency of words are Mexican food, black dog, pulled pork, Chinese food, ice cream, sweet potato and so on. Therefore, we can say people in IL love Mexican and Chinese food that same with our guess.
And the interesting thing is why "black dog" will become one of the top frequency review word in restaurants with high stars?
After we find some reviews including Balck Dog:
Big fan of Black Dog.
Black Dog all you need to know is it is scary good
Black Dog clearly does bbq right.
Black Dog has so much good food & good beer.
and so on. . .
From these reviews, we have an idea that Black Dog is the name of a restaurant in IL, and that restaurants is pretty popular.

Wordcloud for review features of good restaurants in ON:

From the plot, the highest frequency of words are ice cream, pad thai, banh mi, pork belly, customer service, jerk chicken, super friendly and so on. It looks like people in ON love Thai food and attention to customer service.

We will go back to our dataset and have a look at those reviews including pad thai:
Yummy yummy Thai food! We ordered take-out.
Yummy Thai food but super hard to get into
Yum!!! Love their chicken pad thai!
and so on. . . . . .
Most reviews including pad thai are say love Thai food, therefore, people in ON love Thai food as well, and since the top number of restaurants is Chinese food, maybe new businesses can consider open thai restaurants.

Wordcloud for review features of good restaurants in AB:

From the plot, the highest frequency of words is super friendly, friendly staff, spring rolls, gluten-free, friendly service, french toast, customer service, Indian food and so on. It looks like people in AB love Indian food and french toast.

Let's see specific reviews including these words:
Delicious stuffed French toast!
Probably the best French toast I have tried in Calgary
Wow! I thought I wouldn't find better Indian food.
We have had Indian food here which is good and so on...
Therefore, we may say people in AB love Indian food and french toast even though the number of pizza restaurants is the most.

Wordcloud for review features of good restaurants in NV:

From the plot, the highest frequency of words are las vegas, customer service, ice cream, happy hour, fried chicken, friendly staff, super friendly, fried rice and so on. Looks like people in NV attention on customer service, love las vegas and fried chicken.

From some reviews including these words:
Yummy! The fried chicken is so delicious.
Yummy Korean fried chicken.
You gotta love fried chicken!!!
This is the best Korean food in North Las Vegas
The Best Sushi Restaurant in Las Vegas
THE NUMBER ONE BUFFET IN THE LAS VEGAS
and so on. . .
Therefore, we can make a hypothesis that many people eat food in las vegas and love fried chicken and so on food.

Same logic for rest states. And through previously analysis, we know that the top number of restaurants does not represent the taste preference of people. And it shows the importance of the frequency of review words that illustrate what customers thought, feeling and focusing on. From previously text mining for some states, we know customers love ice cream, focus on customer service no matter which states they are coming from. Thus, maybe restaurants business can pay more attention to these general points, and improving more based on the specific preference of different states.

PS: results of the top frequency of review words of rest states are in the Appendix part, welcome to check them if interested in.

Appendix

Summary information of redo PCA:

```
pc2
```

```
## Principal Components Analysis
## Call: principal(r = att, nfactors = 4, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                              PC1   PC2   PC3   PC4     h2   u2 com
## GoodForKids                 0.22  0.64 -0.18  0.15 0.5088 0.49 1.5
## RestaurantsReservations     0.50  0.15 -0.17 -0.53 0.5860 0.41 2.4
## Caters                      0.46  0.37  0.25  0.05 0.4072 0.59 2.5
## RestaurantsTableService     0.57 -0.03  0.14 -0.19 0.3763 0.62 1.4
## RestaurantsTakeOut          0.29  0.52 -0.10  0.23 0.4247 0.58 2.1
## RestaurantsPriceRange2      0.48  0.15 -0.20 -0.45 0.5006 0.50 2.5
## OutdoorSeating              0.55  0.23 -0.06  0.07 0.3672 0.63 1.4
## BikeParking                 0.50  0.34  0.01  0.08 0.3710 0.63 1.8
## HasTV                       0.60  0.26 -0.18 -0.02 0.4613 0.54 1.6
## RestaurantsGoodForGroups    0.50  0.44 -0.33  0.00 0.5467 0.45 2.7
## RestaurantsDelivery         0.16  0.35  0.03 -0.11 0.1633 0.84 1.6
## BusinessAcceptsCreditCards  0.26  0.16  0.13  0.63 0.5031 0.50 1.6
## BusinessAcceptsBitcoin      0.34 -0.05  0.57  0.16 0.4704 0.53 1.8
## ByAppointmentOnly          -0.01 -0.08  0.23  0.03 0.0624 0.94 1.3
## AcceptsInsurance           -0.02 -0.04  0.11  0.01 0.0149 0.99 1.3
## GoodForDancing              0.53 -0.53 -0.27  0.11 0.6481 0.35 2.6
## CoatCheck                   0.56 -0.56 -0.22  0.10 0.6888 0.31 2.4
## HappyHour                   0.58 -0.53 -0.27  0.13 0.7102 0.29 2.5
## WheelchairAccessible        0.51 -0.11  0.56 -0.02 0.5796 0.42 2.1
## DogsAllowed                 0.40 -0.13  0.59  0.01 0.5183 0.48 1.9
## DriveThru                   0.01  0.12 -0.09  0.49 0.2625 0.74 1.2
## Smoking                     0.59 -0.58 -0.17  0.13 0.7283 0.27 2.3
## AgesAllowed                 0.08 -0.11 -0.06  0.05 0.0239 0.98 2.9
## HairSpecializesIn          -0.01 -0.02  0.06  0.00 0.0039 1.00 1.4
## Corkage                     0.10  0.03  0.03  0.05 0.0139 0.99 2.0
## BYOB                        0.04 -0.01  0.02  0.04 0.0034 1.00 2.4
## DietaryRestrictions         0.03  0.00  0.08 -0.03 0.0087 0.99 1.7
## Open24Hours                 0.01  0.00  0.04 -0.01 0.0020 1.00 1.3
## RestaurantsCounterService   0.02  0.00  0.04  0.00 0.0020 1.00 1.5
## stars                       0.24 -0.03  0.38 -0.34 0.3188 0.68 2.7
## reviewcount                 0.35  0.07  0.14  0.02 0.1480 0.85 1.4
##
##                        PC1  PC2  PC3  PC4
## SS loadings           4.41 2.72 1.82 1.47
## Proportion Var        0.14 0.09 0.06 0.05
## Cumulative Var        0.14 0.23 0.29 0.34
## Proportion Explained  0.42 0.26 0.17 0.14
## Cumulative Proportion 0.42 0.68 0.86 1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.05
```

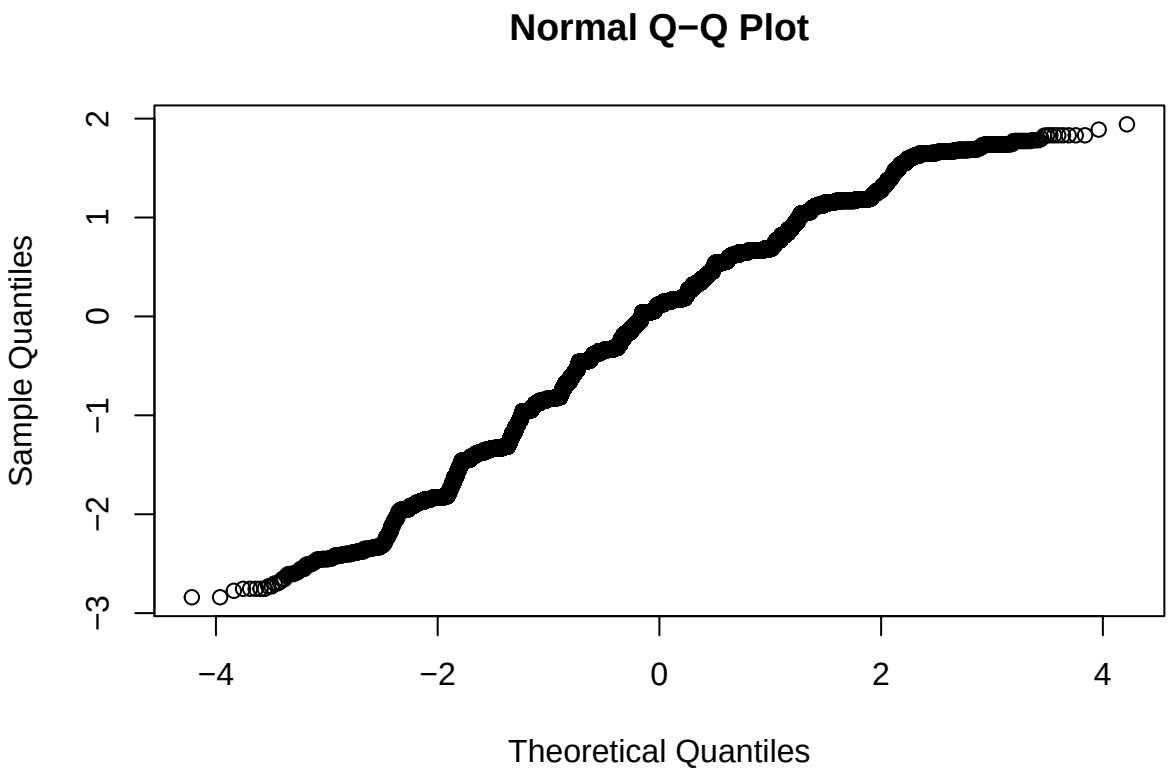
```
##  with the empirical chi square  91275.32  with prob <  0
##
## Fit based upon off diagonal values = 0.88
```

For the multilevel summary information:

```
summary(fit_bs)
```

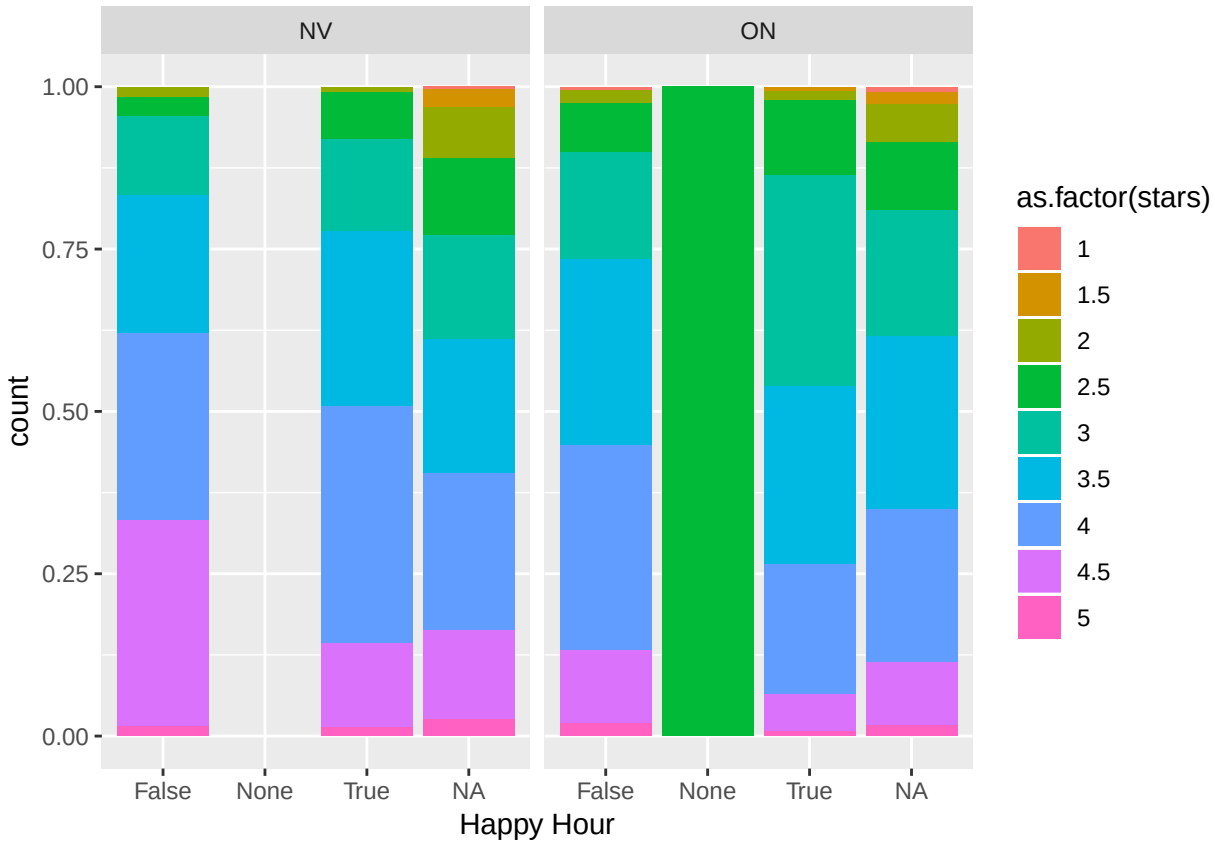
```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## stars ~ RestaurantsReservations + (1 + RestaurantsGoodForGroups +
##     HappyHour + RestaurantsReservations | state) + RestaurantsGoodForGroups +
##     HappyHour
##    Data: att
##
## REML criterion at convergence: 96328.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5720 -0.5771  0.1446  0.7882  2.4434
##
## Random effects:
##  Groups   Name                       Variance  Std.Dev. Corr
##  state    (Intercept)                0.0337965 0.18384
##           RestaurantsGoodForGroups 0.0005973 0.02444  -0.54
##           HappyHour                  0.0003364 0.01834  -0.31 -0.43
##           RestaurantsReservations  0.0034485 0.05872  -0.69 -0.17  0.88
##  Residual                           0.6317247 0.79481
## Number of obs: 40456, groups:  state, 11
##
## Fixed effects:
##                            Estimate Std. Error t value
## (Intercept)                3.349455   0.057327  58.427
## RestaurantsReservations    0.109980   0.018886   5.823
## RestaurantsGoodForGroups -0.028871   0.008940  -3.230
## HappyHour                  0.025635   0.007965   3.219
##
## Correlation of Fixed Effects:
##             (Intr) RstrnR RstGFG
## RstrntsRsrv -0.659
## RstrntsGdFG -0.519 -0.206
## HappyHour   -0.212  0.598 -0.294
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

QQ plot of our model:



This QQ-plot is kind of linear but like stair form means the difference of our continuous response with the categorical observation.

A plot about the difference of coefficient in HappyHour for state NV and ON:



Wordcloud for review features of good restaurants in AZ:

From the plot, the highest frequency of words are: ice cream, happy hour, customer service, friendly staff, super friendly and so on.

Wordcloud for review features of good restaurants in NC:

From the plot, the highest frequency of words are: ice cream, banh mi, pulled pork, customer service, jerk chicken, super friendly and so on.

Wordcloud for review features of good restaurants in QC:

From the plot, the highest frequency of words are: de la, à la, ice cream, le service, foie gras, tasting menu and so on.

Wordcloud for review features of good restaurants in PA:

From the plot, the highest frequency of words are: ice cream, happy hour, beer selection, friendly staff, super friendly, coffee shop and so on.

Wordcloud for review features of good restaurants in OH:

From the plot, the highest frequency of words are: corned beef, ice cream, customer service, friendly staff, super friendly, pad thai, happy hour and so on.

Wordcloud for review features of good restaurants in SC:

From the plot, the highest frequency of words are: fort mill, rock hill, dixie pig, pulled pork, pimento cheese, local dish, chinese food and so on.

Wordcloud for review features of good restaurants in WI:

From the plot, the highest frequency of words are: ice cream, pork belly, friendly staff, sour cream, cheese curds, indian food, hot pot and so on.