Modeling Mid-term Project

Kerui Cao

11/28/2019

Data preparation Clean

I use *two dataset*, one is the dataset containing resord of the number imigrants of each states of the U.S, another one is the dataset downloaded from Yelp website, there are four sub datasets in total, I use “Business” dataset which contains the infomation of part of businesses listed on Yelp. There are 192609 businesses in this dataset, the main steps of data preparation and clean process are:

- Extract restaurant information from all these businesses;
- Extract restaurant with more than 100 reviews;
- Extract cities with more than 150 resturants, which can exactly give us 10 cities after filtering;
- Delete variables with more than 40% missing value;
- Reorganize some variables;
- Delete observations with missing values;

Variables “hours.Monday”, “hours.Tuesday”, “hours.Wednesday”, “hours.Thursday”, “hours.Friday”, “hours.Saturday”, “hours.Sunday” indicate the operational hours for each businesses, we don’t need these information, so we delete these variables.

Some variables whose value are list, for example, the value of variables “attributes.GoodForMeal” is ***{‘dessert’: False, ‘latenight’: False, ‘lunch’: True, ‘dinner’: True, ‘brunch’: False, ‘breakfast’: False}***, we simple re-define this variables as a numeric score by counting how many “True” contained, for the example above, the re-defined value is 2. We do the same for variables “attributes.BusinessParking” and “attributes.Ambience”.

For variable “categories”, we can see that some restaurant contains words like “Chinese”, “French” and so on, so I created new binary variables indicating those informations.

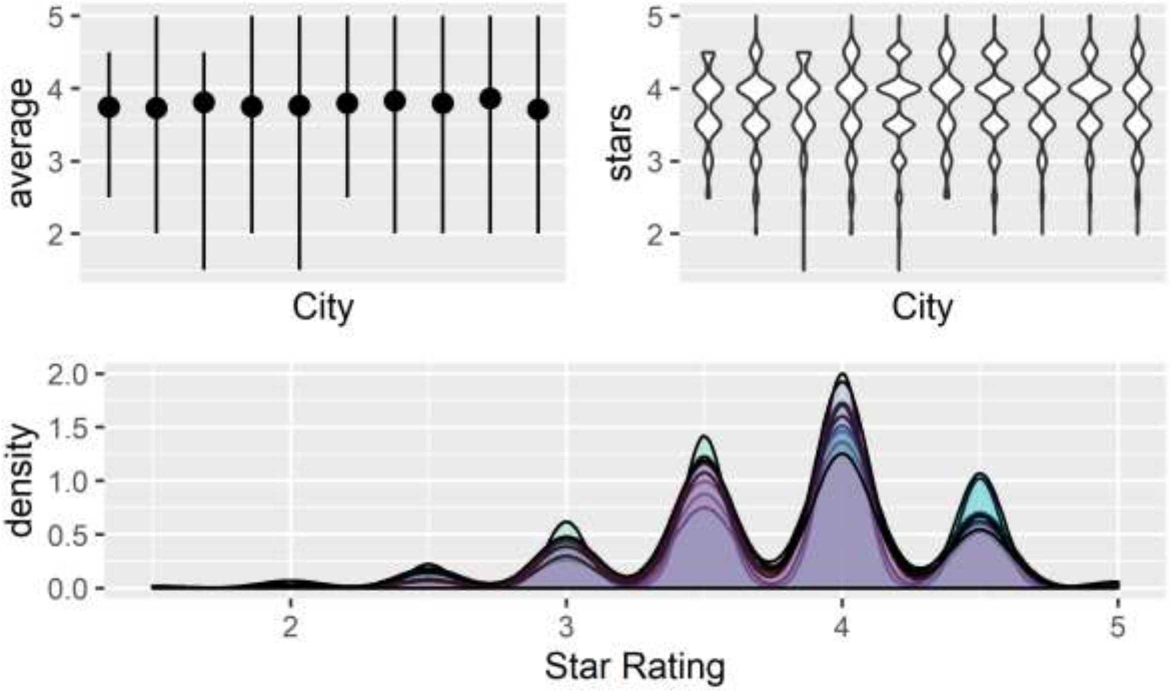
Some binary variables contain value “True” and “False” and “None”, but only a small part of them are “None”, so I simply delete them.

Some variables contain value like “u’average’”, we need to transforme it into only “average”.

EDA

Our research interest lies on the star rating for each restaurants, so we try to apply exploratory data analysis around star ratings of restaurants.

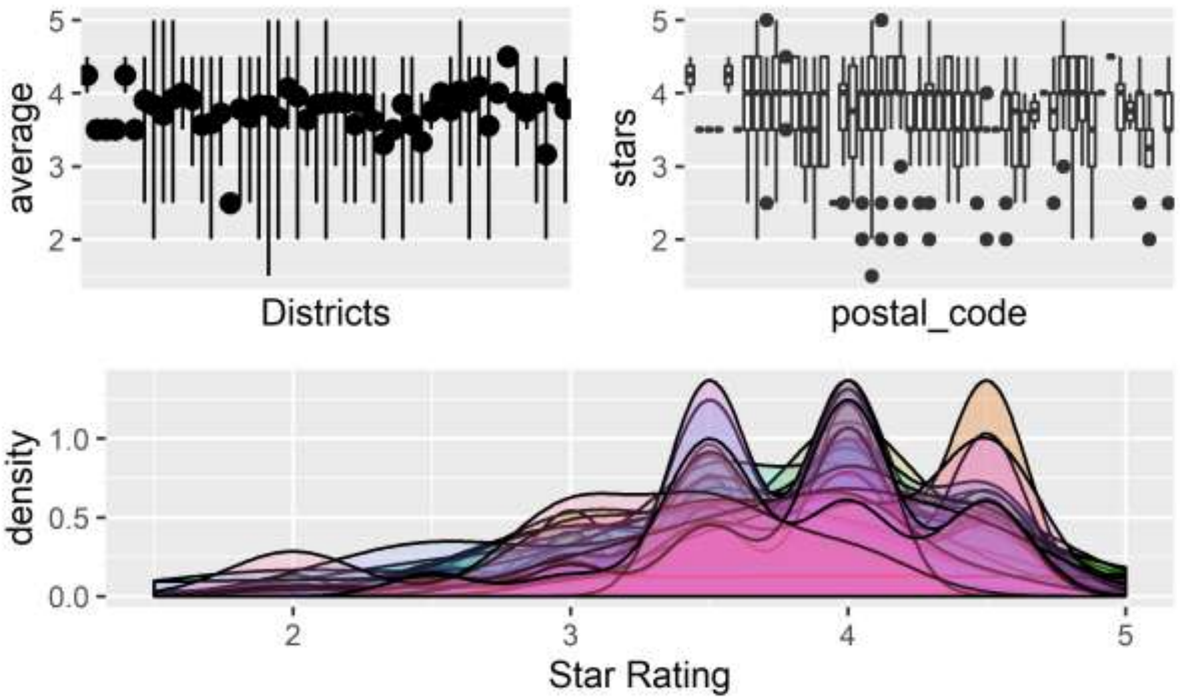
First we will see the distributions of star ratings of restaurants, and we are also interested in the difference of distributions across cities.



Above plot shows the distribution of star ratings, as for the upper left plot, black points are the average star ratings for selected ten city, vertical lines shows the range of star ratings, upper right plot is the violin plot of star ratings of each cities, lower plot shows the density of star ratings in for each cities. We can tell that restaurants in different cities have similar didtribution, which is centering at 4 stars, and barely seeing restaurants with lower than 2 stars. So we may consider that there is no difference between cities.

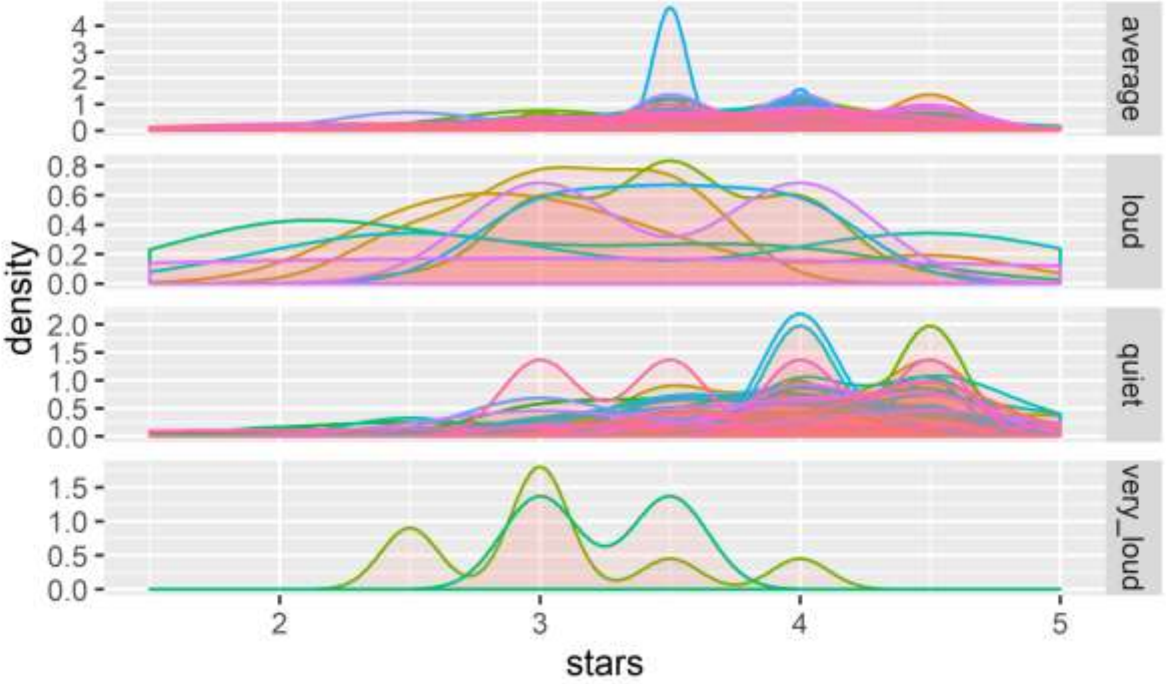
We consider that maybe city is not a good standard to separate and group restaurants, so I tried to separate and group restaurants by districts, which can be indicated by variable "postal_code", here I first tried restaurants in Las Vegas, because we have more date from restaurants in Las Vegas, which is 1916 restaurants, and I only pick districts with more than 30 restaurants, below is part of the list of chosen districts:

Do the same as grouping restaurants by cities, we drew the same plot shown below, from the plot we can tell that restaurants from different district have quite different distributions, which is more significant than the difference between grouping by cities, so district is a better standard to separate and group restaurants.



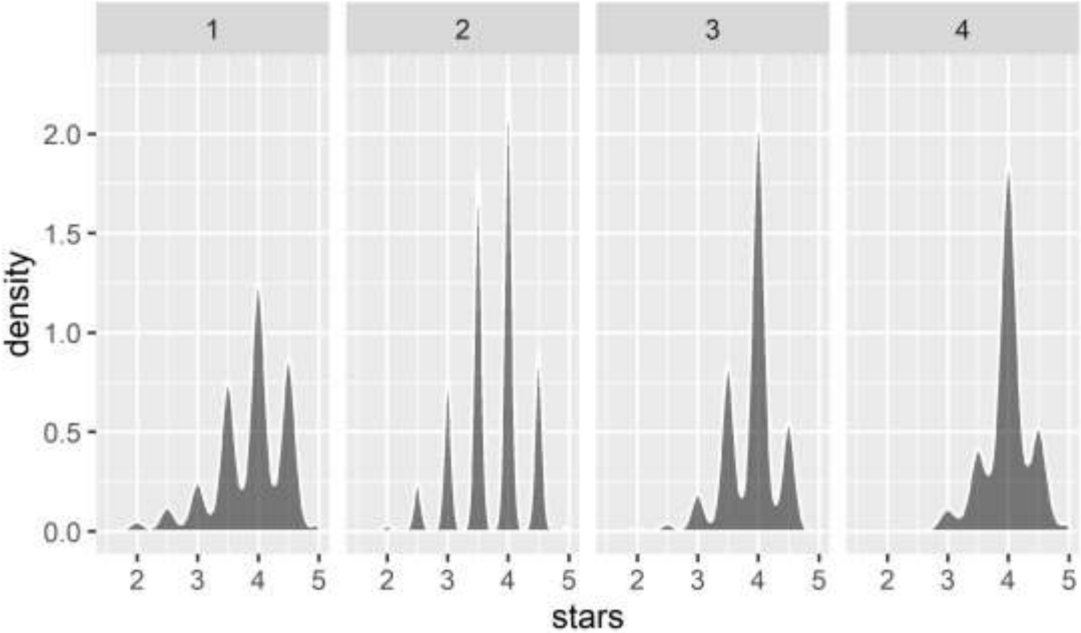
Than we will dive deeper into the data, we try to see the difference between distribution of star ratings between restaurants distinguished by features. I examined all 33 features, below are part of the result.

Noise Level



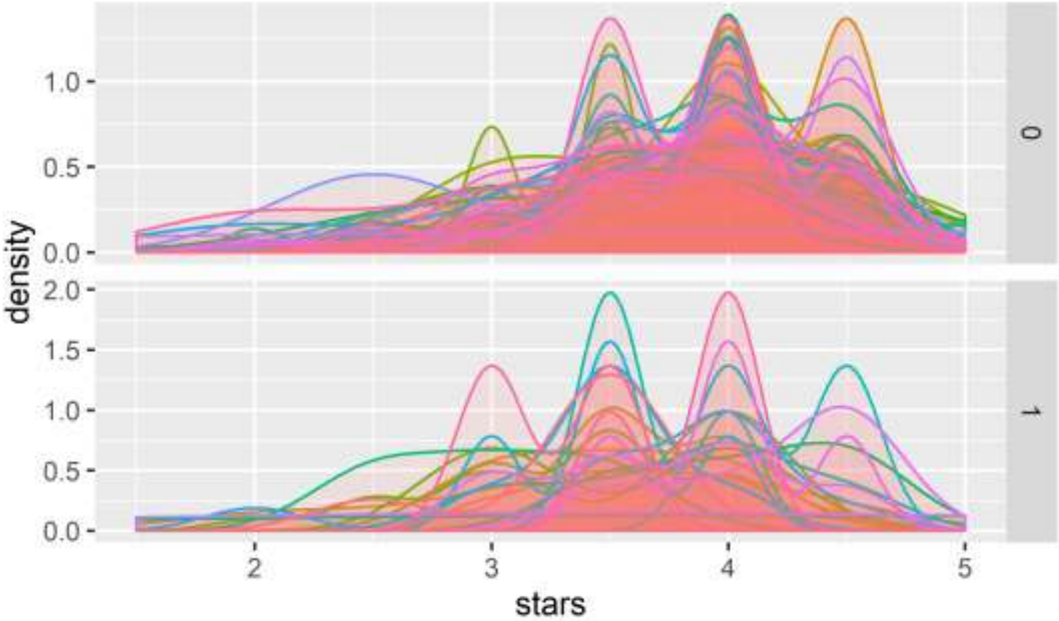
Above plot shows the distribution of stars of restaurants with different noise level across cities, we can see clearly that, ignore difference between districts, as the noise level increase, the stars center ar lower score, quiet restaurants center at 4, loud restaurants center at 3, as for the deifference between cities, we can see that most cities have similar distribution at each noise level.

Price Range



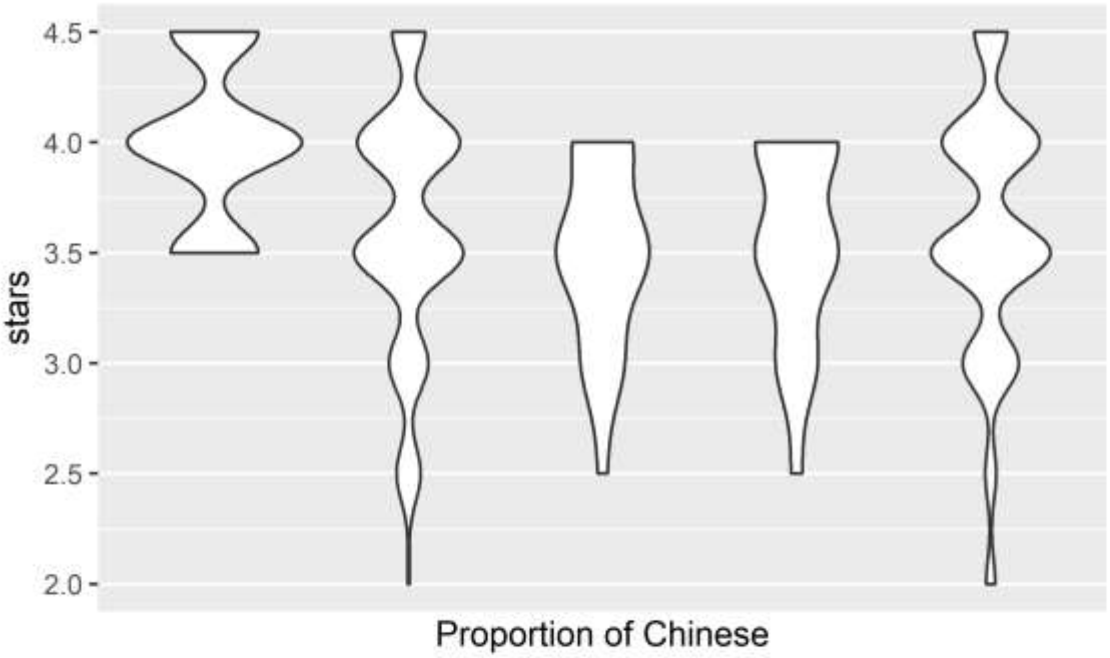
Above plot shows the distribution of stars of restaurants with different price range, expensive restaurants are more concentrating than cheap restaurants, but this phenomenon may be caused by small sample size of luxury restaurants.

Chinese Restaurants



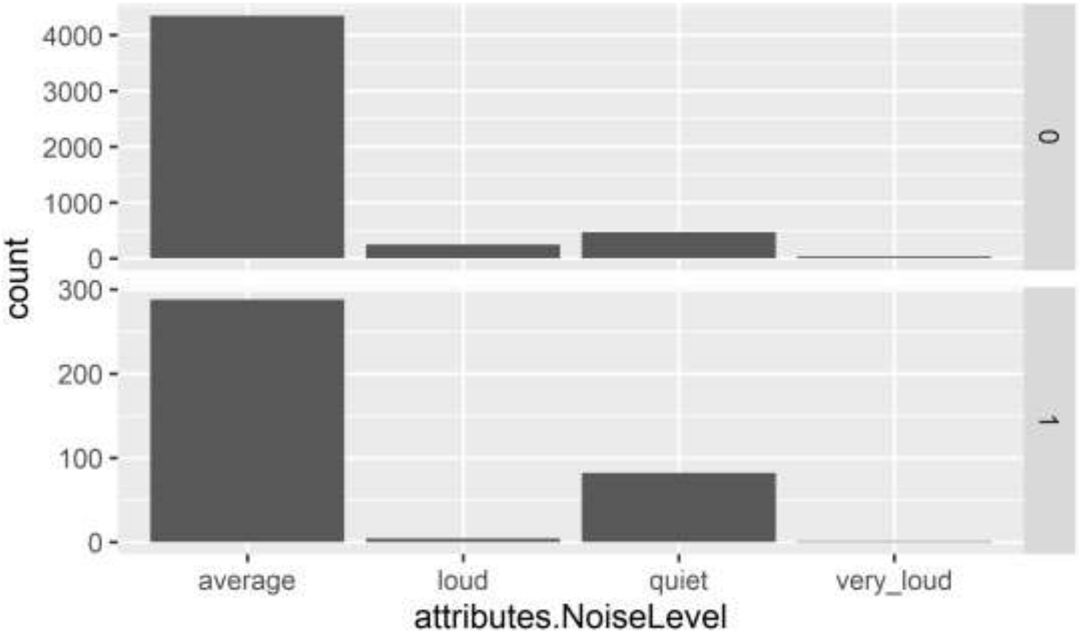
Above plot shows the distribution of star ratings of restaurant providing Chinese dishes and not providing Chinese dishes, we can see clearly that Chinese restaurants tend to have lower stars, as for the difference between districts, we can see that for some districts, Chinese restaurants center at around 4 stars rather than 3.5 stars.

I am a little interested about this phenomenon, so I will explore further. One explaination could be the proportion of Chinese in that city may influence the star ratiings of Chinese Restaurants, so I included a new dataset containing records of imigrants numbers of each states, below is the scatter ponit plot of the proportion of Chinese in each cities and star ratings of restaurants.



Because we have restaurants data from only 5 states, so the number of unique values of proportion of Chinese is only 5, from above plot we can see that there is no clear pattern that indicating that the star ratings of Chinese restuarants are related to the proportion of Chinese imigrants.

To further explore the reason Chinese restaurants having lower stars, and based on previous findings, we suspect that noisy restaurants will have lower stars, below plot shows the distribution of noise level of restaurants providing Chinese dishes and not providing Chinese dishes, we can see that Chinese restaurants have similar distribution and even higher proportion of quiet restaurants, so there may be some other factors. To further analyze those factors affecting star ratings, we need some models to do this.

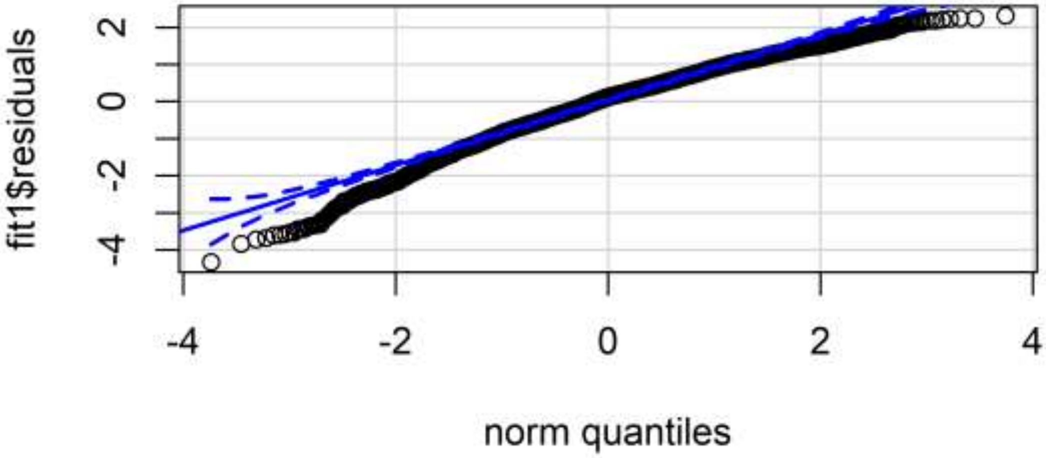


CDA

First we need a rough understanding of the relationships between variables and stars, so we construct linear regression first, as for model selection, the main goal of this research is find out factors other than food quality that affect the star ratings of a resturant, and adjust the star ratings of restaurants so that it can purely reflect the food quality of a restaurants. So as for the model selection, we may have concerns that whether we should include all the potential attibutes or only those attibutes having strong evidence suuporting they will affect the star ratings of restaurants. So we will construct both models and compare the prediction result from both model, as for attibutes selection, we use backward selection.

One more important thing to notice is that I made little transformation about the attibutes indication whether a restaurant is Chinese restaurant or French or others, I did not directly put those attibutes into tht model, instead I put the interaction of those sttibutes and corresponding proportion of imigrants to the model, for example, instead of just variable "China", I put "China:Chinese.p + China" into the model.

We fit the simple linear model with all attibutes, since for this model we do not care the significance of variables, the R^2 is 16.9%, which is not very good is for an model whose aim is to predict, which means our model can not capture too much information of the data, but from another perspective, we can say that all these attributes can only affect a small part of the star ratings, which means the star ratings is less bised by these attibutes. Below plot shows the QQ plot of the residual, which indicating that the residuals are not normally distributed, but it is close to normal distribution.

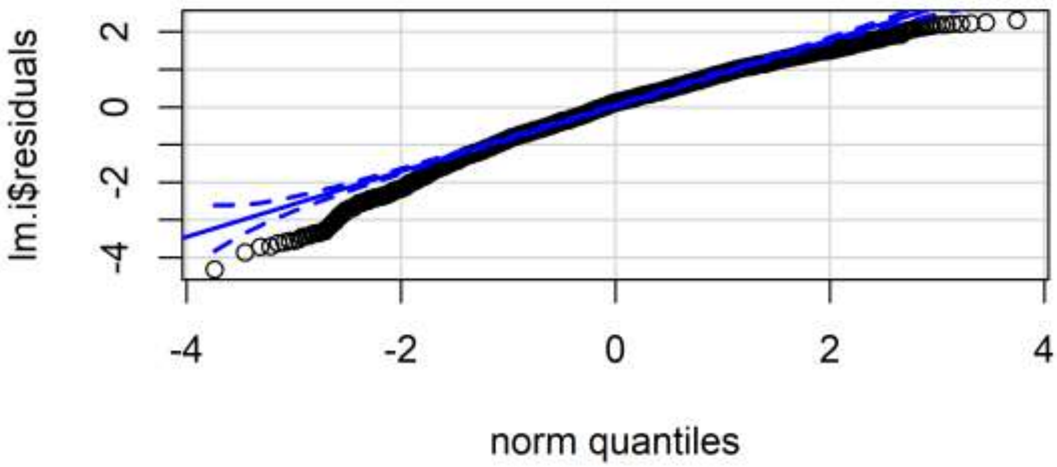


We fit the model after variables selection, below is the summary of this model:

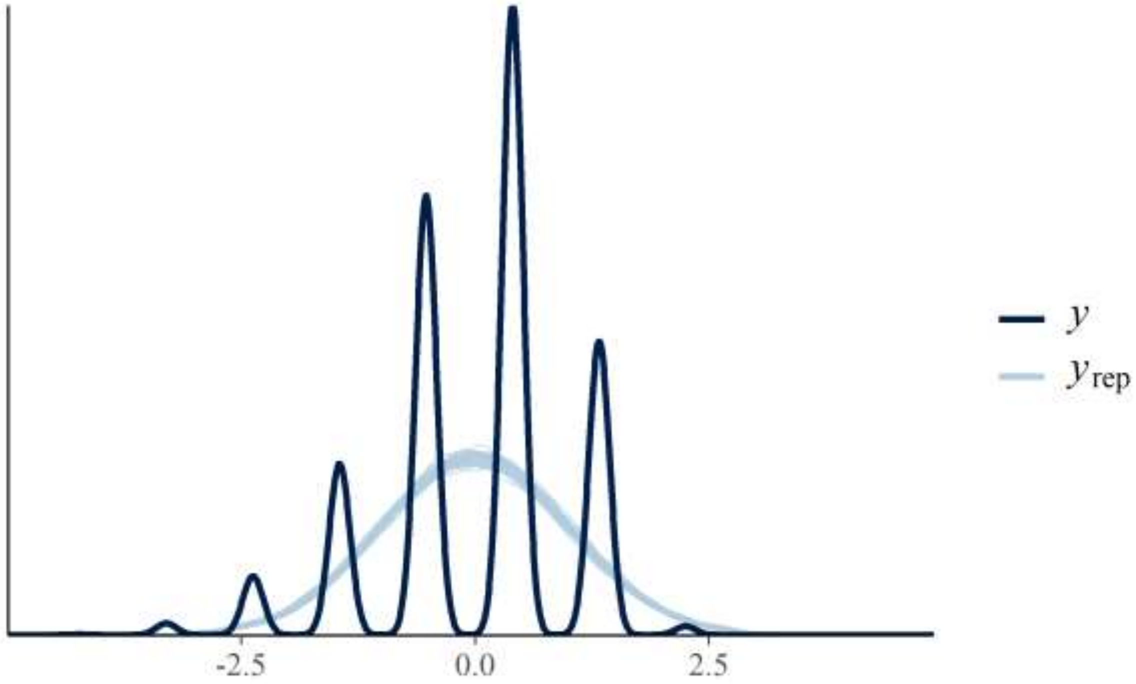
```
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -0.21066    0.09105  -2.314 0.020730 *
## attributes.GoodForKids                -0.19753    0.03979  -4.964 7.12e-07 ***
## attributes.RestaurantsReservations     0.20361    0.03011   6.762 1.50e-11 ***
## attributes.GoodForMeal                -0.05057    0.01488  -3.398 0.000685 ***
## attributes.BusinessParking             0.10929    0.02155   5.071 4.08e-07 ***
## attributes.Caters                      0.12148    0.02701   4.498 6.99e-06 ***
## attributes.NoiseLevelloud             -0.54500    0.06087  -8.954  < 2e-16 ***
## attributes.NoiseLevelquiet             0.25441    0.04237   6.004 2.05e-09 ***
## attributes.NoiseLevelvery_loud        -1.08730    0.15776  -6.892 6.11e-12 ***
## attributes.BikeParking                 0.16371    0.03406   4.807 1.57e-06 ***
## attributes.Ambience                    0.36742    0.02100  17.499  < 2e-16 ***
## attributes.WiFino                     -0.07750    0.02658  -2.916 0.003562 **
## attributes.WiFipaid                   -0.07798    0.19151  -0.407 0.683893
## attributes.Alcoholfull_bar            -0.43700    0.03719 -11.750  < 2e-16 ***
## attributes.Alcoholnone                 0.19217    0.03751   5.124 3.10e-07 ***
## attributes.RestaurantsAttiredressy     0.10516    0.07220   1.456 0.145321
## attributes.RestaurantsGoodForGroups   -0.18901    0.06093  -3.102 0.001931 **
## Chinese                               -0.44927    0.05009  -8.969  < 2e-16 ***
## French                                 0.23919    0.09560   2.502 0.012384 *
## Italian                               -0.03053    0.06210  -0.492 0.622960
## Japanese                               0.27132    0.07541   3.598 0.000324 ***
## Italian:Italy.p                      745.31294  307.02728   2.428 0.015235 *
## Japanese:Japan.p                    -551.96873  322.18058  -1.713 0.086728 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9106 on 5442 degrees of freedom
## Multiple R-squared:  0.1742, Adjusted R-squared:  0.1709
## F-statistic: 52.19 on 22 and 5442 DF,  p-value: < 2.2e-16
```

According to the regression result, most of the selected variables are statistic significantly, which meets our requires, the result also confirmes what we find in the EDA part, load restaurants have lower average star, Chinese restaurants have lower stars, and more funny facts are shown in the result, we can see that Italian restaurants also have lower average star ratings, just like Chinese restaurants, but what is different is that if a city has higher proportion of Italian imigrants, the star ratings of Italian restaurants will be even lower, but Japanese restaurants are just the opposite, higher average star ratings and can be even higher if there are a lot of Japanese imigrants in a city.

As for the R^2, which is 17% and is slightly bigger than the model without variable selection, below is the QQ plot of this model:



Observe this plot, we can see the it is quite similar to the one belonging to model without variable selection, so for simlicity, we want to keep the attibutes selection process. We further exam the model, below is the replicated star ratings from this model:



Observe the plot above, our model caputures the general trend of the star ratings, which is certering at 4 star and nearly normal distributed, but because the star ratings are disctete, so the limitation of model makes it lost a lot information of the data, this also suggests that we may switch from Gaussian distribution to Multinomial distribution, which is categorical regression.

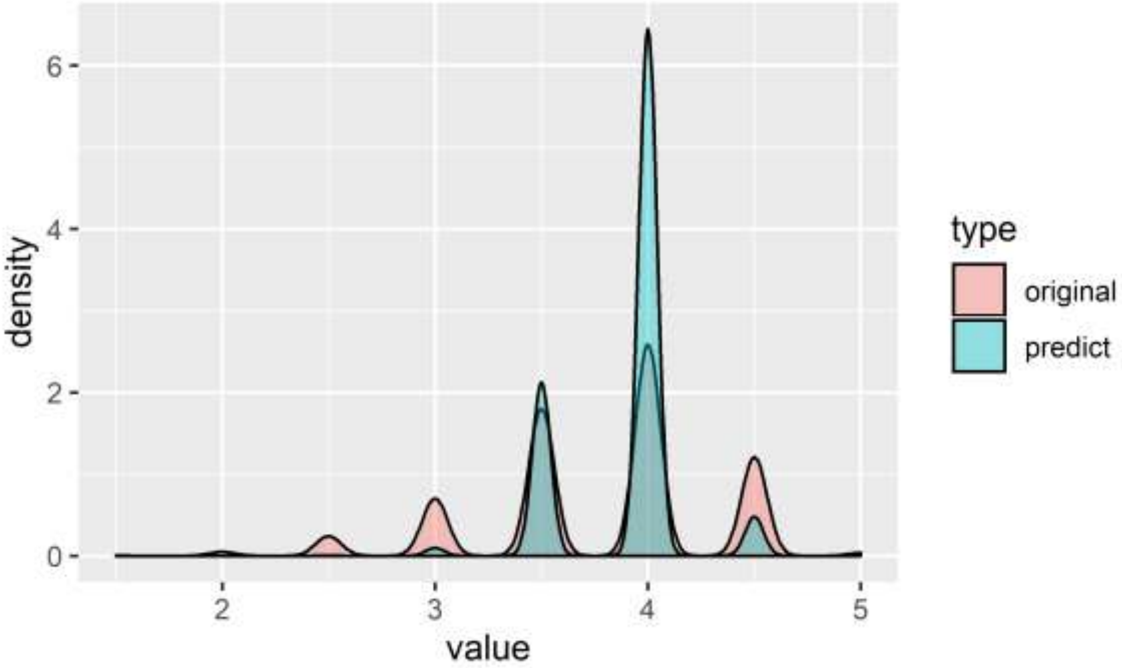
As for fitting the categorical regreesion, we still keep the variable selection process that based on backward and forward selection. Below is the result of categorical:

```
## Coefficients:
##                                         Value Std. Error    t value
## attributes.GoodForKids               -0.44580  7.599e-02 -5.867e+00
## attributes.RestaurantsReservations    0.38853  6.094e-02  6.376e+00
## attributes.GoodForMeal               -0.11138  3.021e-02 -3.686e+00
## attributes.BusinessParking            0.18560  4.342e-02  4.275e+00
## attributes.Caters                     0.21140  5.471e-02  3.864e+00
## attributes.NoiseLevelloud            -1.05028  1.221e-01 -8.600e+00
## attributes.NoiseLevelquiet            0.57300  8.733e-02  6.561e+00
## attributes.NoiseLevelvery_loud       -2.08319  3.217e-01 -6.476e+00
## attributes.BikeParking                0.24772  6.820e-02  3.632e+00
## attributes.Ambience                   0.79327  4.425e-02  1.793e+01
## attributes.WiFino                    -0.18011  5.382e-02 -3.347e+00
## attributes.WiFipaid                  -0.06464  3.876e-01 -1.667e-01
## attributes.Alcoholfull_bar           -0.91987  7.611e-02 -1.209e+01
## attributes.Alcoholnone                0.47838  7.740e-02  6.181e+00
## attributes.RestaurantsGoodForGroups  -0.57948  1.331e-01 -4.354e+00
## Chinese                              -0.89680  1.000e-01 -8.968e+00
## French                                0.53380  1.936e-01  2.757e+00
## Italian                               0.04881  8.208e-02  5.946e-01
## Japanese                              0.32244  9.853e-02  3.273e+00
## Italian:Italy.p                     695.23513  1.515e-05  4.588e+07
```

```
##
## Intercepts:
##       Value         Std. Error    t value
## 1.5|2       -8.1860        0.7323      -11.1777
## 2|2.5       -5.1263        0.2438      -21.0300
## 2.5|3       -3.3017        0.1995      -16.5460
## 3|3.5       -1.8677        0.1923       -9.7108
## 3.5|4       -0.2690        0.1906       -1.4112
## 4|4.5        1.7719        0.1917        9.2425
## 4.5|5        5.7143        0.2645       21.6045
##
## Residual Deviance: 14837.14
## AIC: 14891.14
```

We can see that the result of multinomial regression still confirms our findings in EDA, load restaurants have lower average star, Chinese restaurants have lower stars, the difference between categorical regression and simple linear regression is for Japan restaurants, no strong evidence will support that higher proportion of Japanese imigrants will increase the star ratings of Japanese restaurants.

Below is the model checking, we use categorical regression model to predict the star ratings of each resturants and compare it to the observed star ratings, we can see that predicted stars from multinomial regression model are more concertrating at 4 star, original data are more evenly separated.



Based on our EDA findings, restaurants from same district will have different distributions of star ratings, so it is better to consider multilevel model with random effect related to districts, because we also find that different cities will have similar star rating distributions, so we just construct Mixed Effect Model with in one single city, here I still choose Las Vegas for the same reason, it has most restaurants, since in a single city, the proportion of imigrants are the same for all restaurants, so I did not include them in the model. Below is the summary of fitted model:

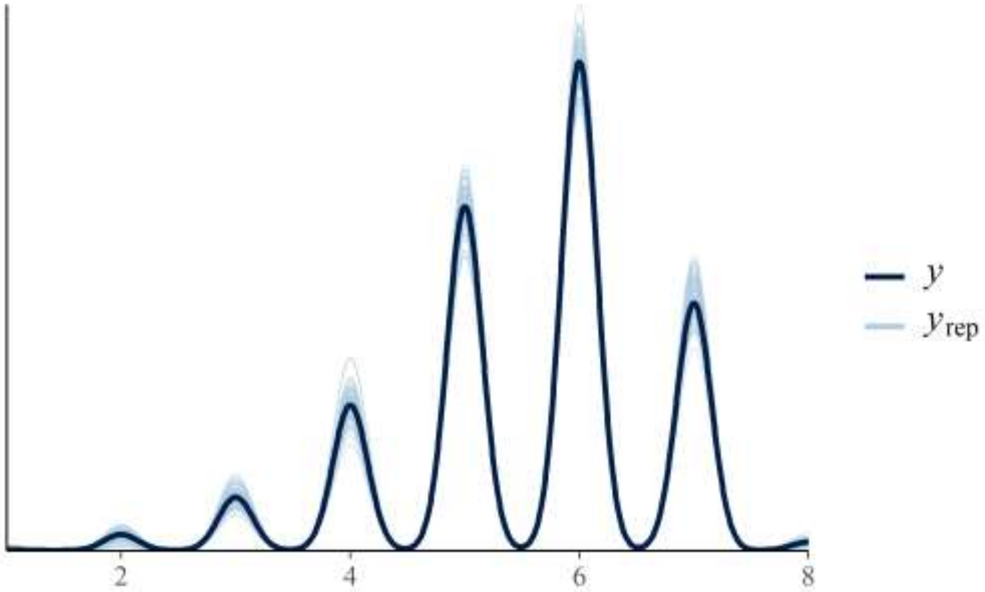
```
## Group-Level Effects:
## ~postal_code (Number of levels: 21)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.20      0.05     0.12     0.32 1.00     2040     2750
## Population-Level Effects:
##                                     Estimate Est.Error l-95% CI u-95% CI Rhat
## Intercept[1]                           -4.15      1.29    -7.16    -2.18 1.00
## Intercept[2]                           -1.64      0.32    -2.27    -1.02 1.00
## Intercept[3]                           -1.20      0.23    -1.65    -0.75 1.00
## Intercept[4]                           -0.82      0.22    -1.25    -0.39 1.00
## Intercept[5]                           -0.08      0.20    -0.46     0.31 1.00
## Intercept[6]                            1.25      0.21     0.86     1.67 1.00
## Intercept[7]                            4.45      0.37     3.72     5.20 1.00
## attributes.GoodForKids                 -0.26      0.08    -0.41    -0.11 1.00
## attributes.RestaurantsReservations      0.34      0.06     0.21     0.46 1.00
## attributes.GoodForMeal                 -0.05      0.03    -0.10     0.00 1.00
## attributes.BusinessParking              0.05      0.04    -0.03     0.14 1.00
## attributes.Caters                       0.26      0.06     0.14     0.37 1.00
## attributes.NoiseLevelloud              -0.59      0.11    -0.81    -0.38 1.00
## attributes.NoiseLevelquiet              0.29      0.10     0.10     0.49 1.00
## attributes.NoiseLevelvery_loud         -1.06      0.24    -1.54    -0.61 1.00
## attributes.BikeParking                  0.10      0.07    -0.03     0.23 1.00
## attributes.Ambience                     0.51      0.05     0.42     0.60 1.00
## attributes.WiFino                      -0.16      0.05    -0.27    -0.07 1.00
## attributes.WiFipaid                    -0.13      0.32    -0.73     0.51 1.00
## attributes.Alcoholfull_bar             -0.24      0.07    -0.38    -0.09 1.00
## attributes.Alcoholnone                  0.17      0.07     0.03     0.31 1.00
## attributes.RestaurantsGoodForGroups    -0.26      0.14    -0.54     0.01 1.00
## Chinese                                -0.44      0.09    -0.61    -0.26 1.00
## French                                  0.37      0.19     0.01     0.75 1.00
## Italian                                 0.04      0.09    -0.13     0.21 1.00
## Japanese                                0.47      0.10     0.29     0.66 1.00
##                                     Bulk_ESS Tail_ESS
## Intercept[1]                            5582     2154
## Intercept[2]                            5224     2969
## Intercept[3]                            7399     3309
## Intercept[4]                            7566     3379
## Intercept[5]                            7189     3130
## Intercept[6]                            7036     3236
## Intercept[7]                            7669     3197
## attributes.GoodForKids                  6681     3100
## attributes.RestaurantsReservations      6175     3143
## attributes.GoodForMeal                  6285     3429
## attributes.BusinessParking              7721     3253
## attributes.Caters                       6730     2852
## attributes.NoiseLevelloud               8421     2914
## attributes.NoiseLevelquiet              8761     2752
## attributes.NoiseLevelvery_loud          8693     3150
## attributes.BikeParking                  7182     2978
## attributes.Ambience                     6529     3024
## attributes.WiFino                       8036     2881
## attributes.WiFipaid                    10142     2775
## attributes.Alcoholfull_bar              5539     3341
## attributes.Alcoholnone                  5823     3196
## attributes.RestaurantsGoodForGroups     7840     2745
## Chinese                                 8035     3373
## French                                  9598     2829
## Italian                                 8943     3136
## Japanese                                8401     3035
```

First we have to validate this result, because we obtained this result through MCMC sampling method, so we have to check if the sampling converges, here all the $\hat{R}$ are 1, and we achieved a relatively large Efficient Sample Size, so we can tell that the MCMC sampling converges well.
Second, From the summary we can see that 95% confidence interval of most coefficients do not contain 0, so we can conclude strong evidence supporting that most of these variables do affect the star ratings of a restaurants, and most of them are consistant with the result of our previous model, due to this is a categorical regression, so the interpretation of model will be based on possibility or odds. Example interpretation is shown below:

We take the mean of posterior distribution as the point estmator:

- $Log[\frac{P(star>1.5)}{P(star=1.5)}]$ = -4.13 + $X\beta$
- $Log[\frac{P(star>2.0)}{P(star\leq 2.0)}]$ = -1.63 + $X\beta$
- $Log[\frac{P(star>2.5)}{P(star\leq 2.5)}]$ = -1.19 + $X\beta$
- $Log[\frac{P(star>3.0)}{P(star\leq 3.0)}]$ = -0.82 + $X\beta$
- $Log[\frac{P(star>3.5)}{P(star\leq 3.5)}]$ = -0.07 + $X\beta$
- $Log[\frac{P(star>4.0)}{P(star\leq 4.0)}]$ = +1.25 + $X\beta$
- $Log[\frac{P(star=5.0)}{P(star\leq 4.5)}]$ = +4.46 + $X\beta$

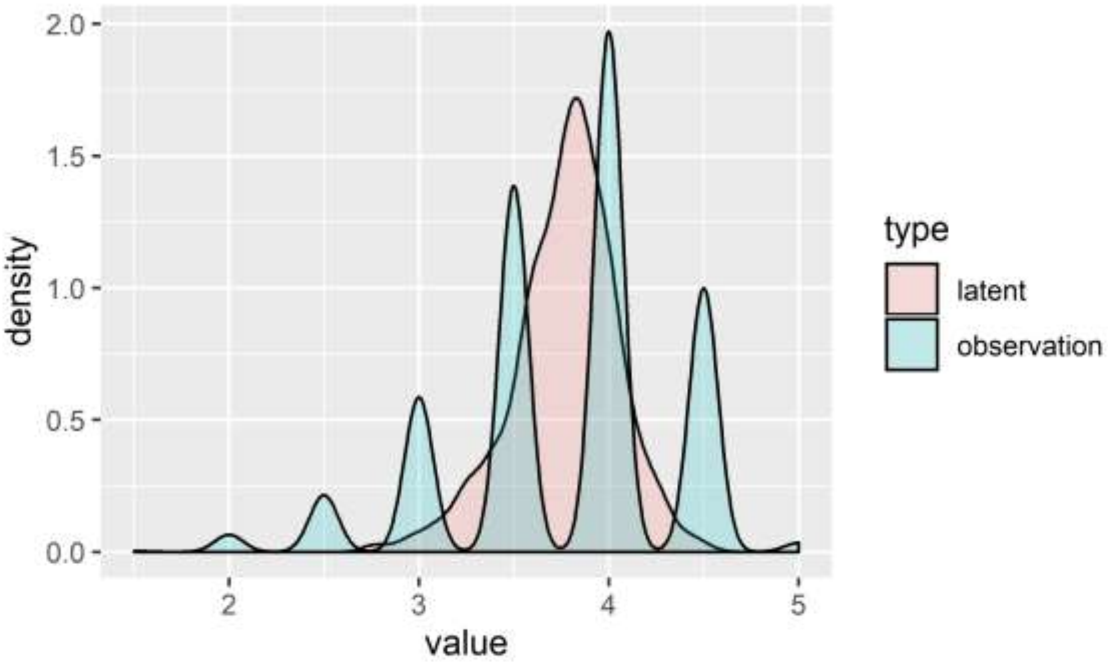
Next we can use the model to replicate a star ratings and compare it to observed star ratings, we can treat the result of comparison as model checking:



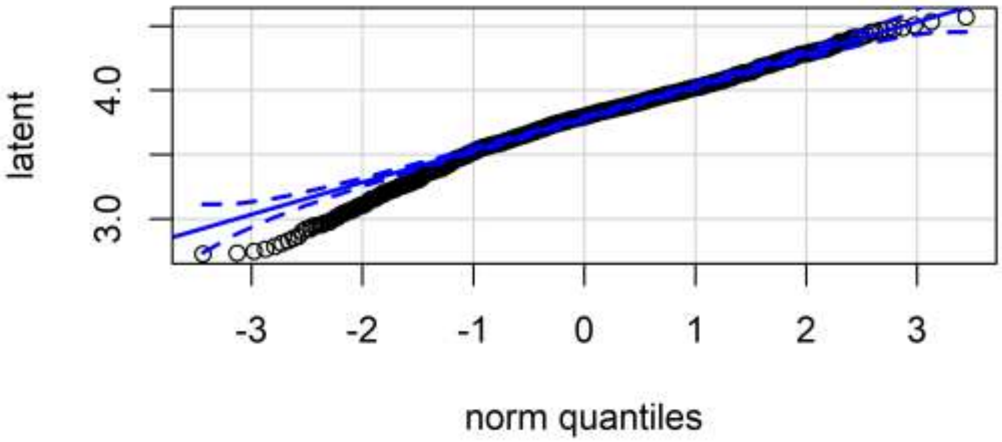
We can see that compared to non-mixed effect multinomial model, new mixed-effect multinomial model performs better, which captured most of informations of dataset.

Based on above interpretation, we can also calculate the possibility of each restaurants receiving each star ratings, below is part of the result:

Here our response variable are star ratings, which is ordinal, so according to theory, behind the observed star ratings, there is a continuous latent variable that dominate the observed star ratings, so here we use the expectation of above predictions as the estimation of continuous latent variable:



Observe above plot, the estimated latent variable is consistent with the observed star ratings, which means our model is valid. Below plot shows that the latent variable are nearly normally distributed, but left-skewed.

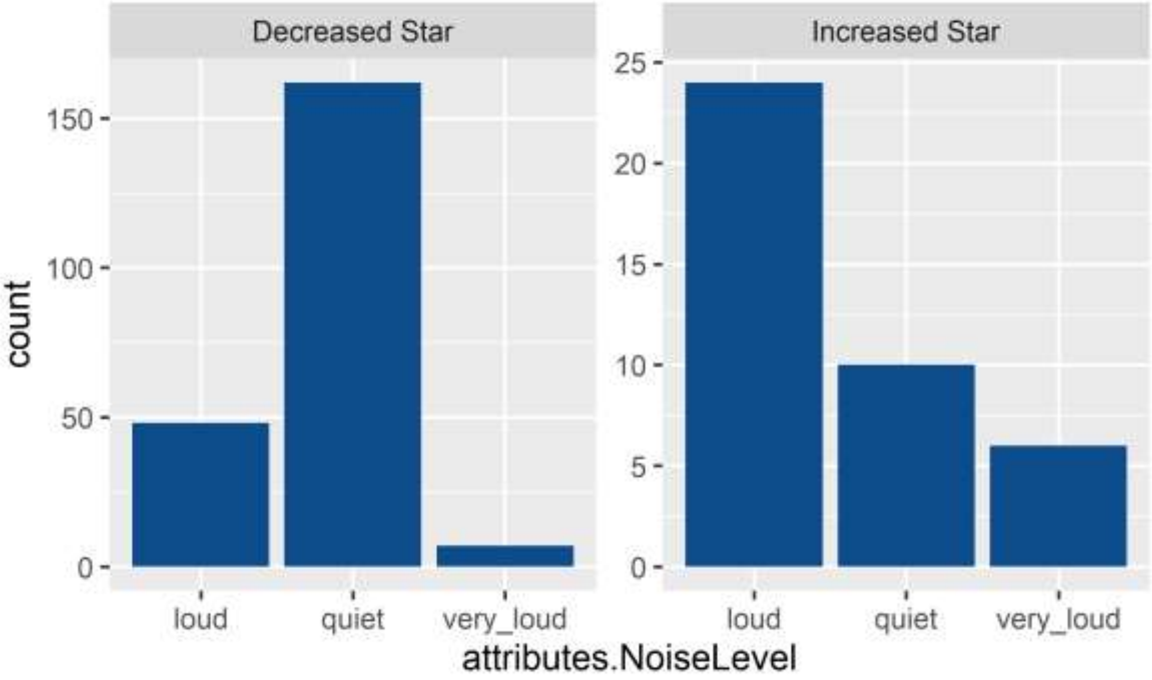


Since we ontained the latent variables, now we fit a regression model of this latent variable, note that we will also consider multi-level, Here we do one more step that we check if there is any mix-effet:

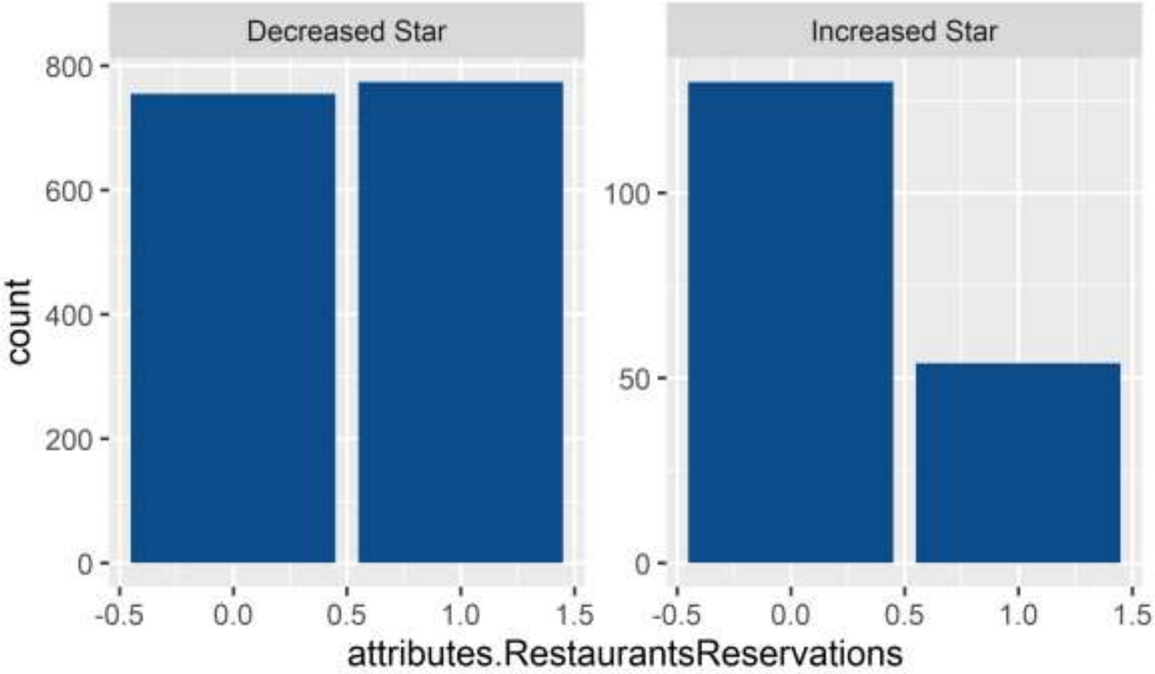
```
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 1579.1, p-value < 2.2e-16
```

The P-value above shows that there is mix-effect. The model we obtained tells us what attibutes will affect the latent variables, all those variables identified are not related to food quality, so we can assume that information related to food quality is put into residuals, so the the residual will be a good indicator of the food quality of a restaurants, so I calculated the residuals and cut it into 10 groups, the index of new group will be the new "star", below is part of the comparison between old "star" and new "star":

Look at above table, we can see that after adjustment, some restaurants have lower star ratings, some have higher star ratings, we can go further to check whether it is reasonable:



Look at above plot, based on the changes of star ratins, I divided restaurants into two groups, one is restaurants received lower new star ratings, another is restaurants received higher new star ratings, for those received lower star ratings, we can find that a lot of them are quite restaurants, so their original star ratings benefit from quite environment, but after we exclude the effect of quite environment, it is reasonable to receive lower star ratings, while load restaurants are just the opposite, so the result is consistent, which is good.



Whether a restaurants accept reservation will affect its star ratings, accept reservation will increase the star ratings, so after we exclude its effect, for those restaurants that accept reservations will receive lower star ratings, so in above plot, restaurants that received lower new star ratings have a larger proportion of accepting reservation, which is consistent.

Above two simple checks validate our model and result.

Conclusion

We successfully iddentified factors that affect the star ratings of resraurants, after that I adjusted the star ratings of restaurants according to those attibutes, the new star ratings are not biased by some attibutes that is not related to food quality, which can provide better guidance to people who value the food quality a lot.

Further research direction

There are still drawbacks of our research, below is the future directions we can improve our research.

- More sample from different cities, states and different kinds of restaurants;
- More attributes of restaurants;