Yelp Data Project

Jiawei Li

2017/11/18

1.Introduction

1.Problem Statements

There are many Beauty and Spas businesses in United states. The purpose of my project is try to find out what factors will have impacts on the rating of the beauty shops. Will the location impact the rating of the shops? Are there have any difference between shops that have longer opening hours and have shorter opening hours? If the Beauty shops provide many different items or products such as nails spa, hair salon, medical spas and so on, will the shops get higher rating? What about the influence from the number of the reviews that those shops have?

There might have other factors that will affect the ratings, but the questions above may be the main questions that I want to figure out in my project.

2. Data

Yelp challenge dataset: business, hours, category, attribute, checkin.

a.The business data contains 156639 observations of 12 variables.

b.The hours data contains 734421 observations of 2 variables.

c.The category data contains590290 observation of 2 variables

d.The attribute data contains 156639 observations of 12 variables

e.The ckeckin data contains 156639 observations of 12 variables

2.Data Cleaning

First we need to connect to the database and select the dataset we need (business/hours/category). After Listing those data into a table and checking their columns, I retrieved the data from MySQL and started to clean the data.

After getting data from MySQL, I select the beauty shops in the category dataset. Based on the Yelp data description of beauty shops, I filter all the categories that are related to the beauty shops, and then make it as a dataframe. And also calculate the number of category for each shops.

What's more, as the original dataset of hours is not in the formation that I want, I do some manipulation to get the information about the how many days each shop opens in a week and how long do the shops open in a day(Insted of using the specific time in different day, I use the average working time here, since the working time varies at different days).

Moreover, I calculate the total number of checkin per shop and pick five variables taht may have impacts on the stars from attribute dataset. To clean the attribute dataset, I first select the beauty shops and then change the form of the dataframe. There are a lot of variables in the attribute dataset, but some of these variables contain too much NA values. So I pick up 5 variables that have enough values (that means that the variables have at least 10,000 values in the attribute dataset after transformation[14,692 values]) and may

have impact on the stars of the shops. These five variables are BikeParking, BusinessAcceptsCreditCards, ByAppointmentOnly, RestaurantsPriceRange2,number_Pway.

At last, I eliminate those shops that are not open, merge three dataframe together and filter out the shops in the US.

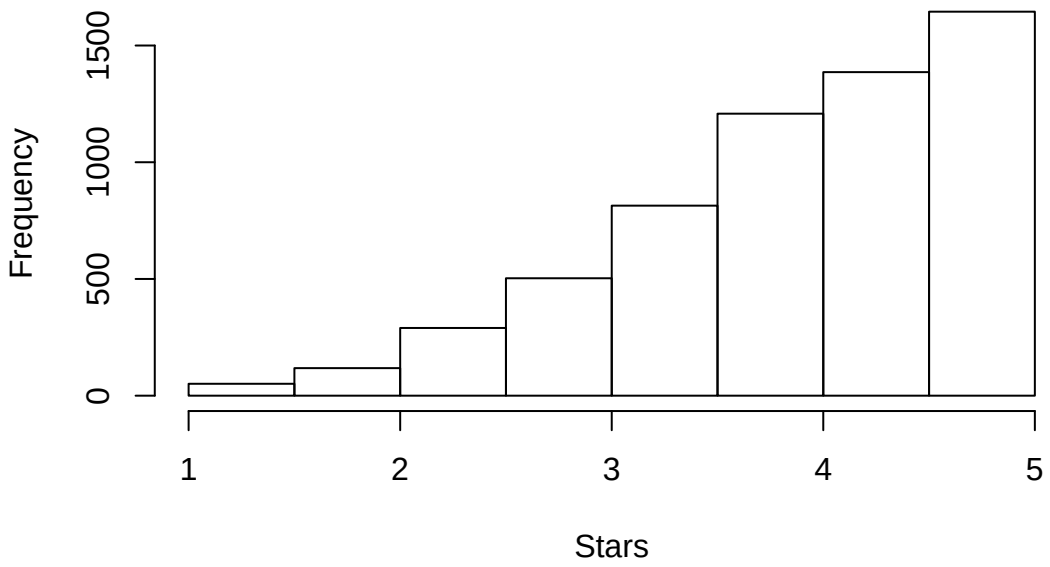
{“garage”: true, “street”: false, “validated”: false, “lot”: false, “valet”: false}

3.Exploratory Data Analysis

In this part, I would like to find out how many states have beauty shops records in yelp first, and then try to figure out the variables that may have impact on the stars of the shops. I find that there are only 9 state have the records of beauty shops in the yelp.And graph two shows the number of beauty shops in each state. AZ has the most higher number of beauty shops and AL has the most lower number of beauty shops.[see Appendix.1.(1)]

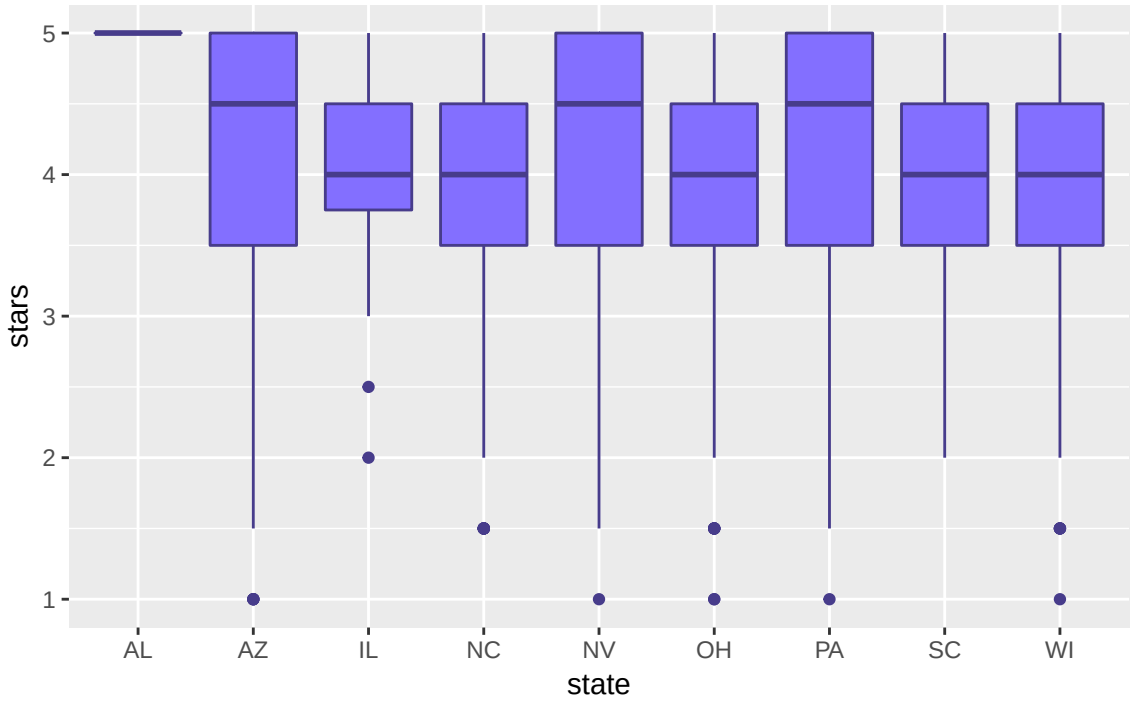
Let’s look at the histogram of stars and the boxplot of stars that group by states. The histogram of stars shows that the count of shops increase with the increase of the ranking.

Figure.1 Histogram of Stars



By comparing the stars in different state in Figure.2, we can see that the mean of different states are within 4-5 stars. Since in AL, there is only 1 observation, so the mean is equal to the value of stars in that observation. We can see from the plot that these boxplot is comparatively tall, which means that people hold quite different opinions in rating the beauty shops. Besides, the long lower whisker of these states means that stars are varied amongst the most least positive quartile group. What’s more, there are some outliers in some states.

Figure.2 Comparison among States−Stars



After that, we can look at the distribution of average worktime/ workday/ number of category / the review count.

Figure.3 Comparison among St

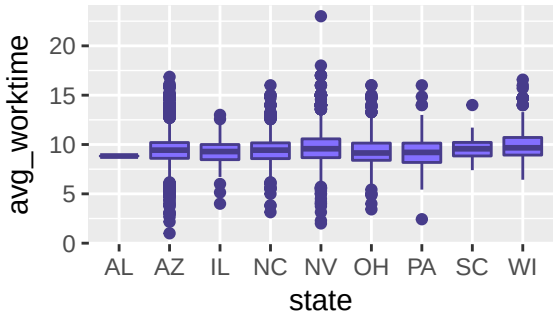


Figure.4 Comparison among Sta

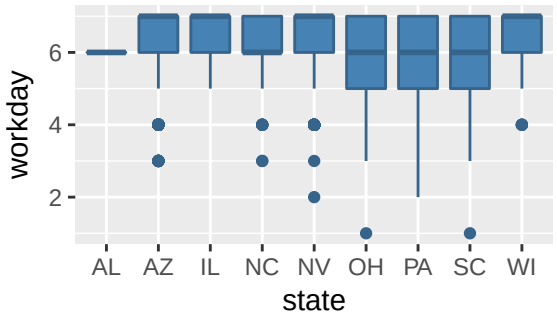


Figure.5 Comparison among St

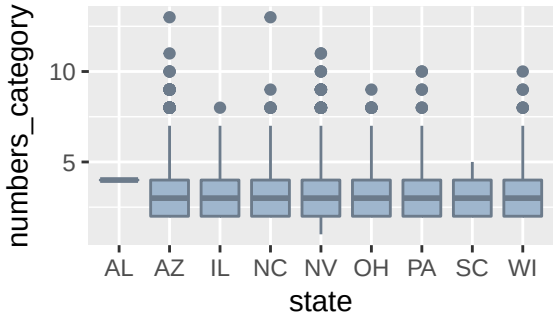
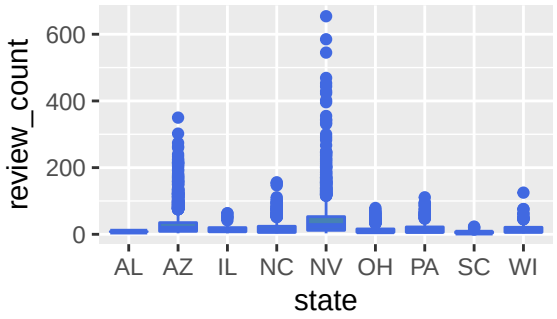


Figure.6 Comparison among S



We can see a common problems from the plot that each variables contains many outlier, especially the review_count plot. In the worktime graph, we can see that the mean of average worktime in different staes are quite similar, and the box plot is comparatively short. In the workday plot, we can see that some states have long lower whisker, which means that workdays are varied amongst the most least positive quartile

group. In the numbers of category plot, we can see that the mean values are different in different states. In the review plot, there might are not showing an obvious patterns.

Let's explore more information about the variables and its impacts in the stars by the plot below. Figure.7 shows the relations between review counts and stars. Most of the beauty shops have little review counts. In NV and AZ, it shows that more review counts may result in higher stars. In Figure.8 , comparing the average worktime to 10, it seems that those shops with longer worktime will get a lower stars, especially in state AZ and WI. Figure.9 and Figure.10 are not clearly enough to show the relations, but these plots show that workday and the number of category may have some impacts on the stars. As a result, we need to fit the model in order to make a more clearly and reasonable conclusion.

Figure.7 The Relations Between Review_counts and Stars

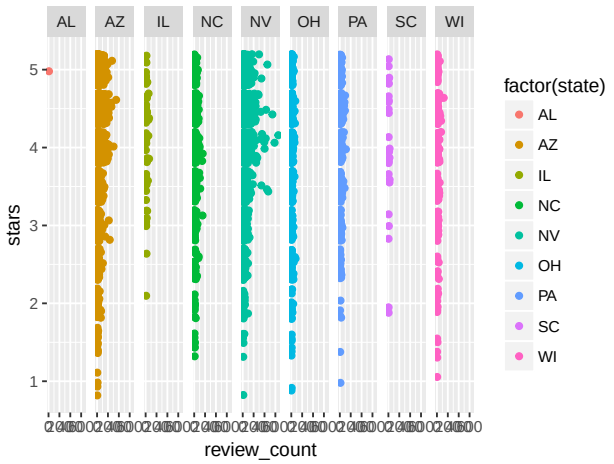


Figure.8 The Relations Between Avg−worktime and Stars

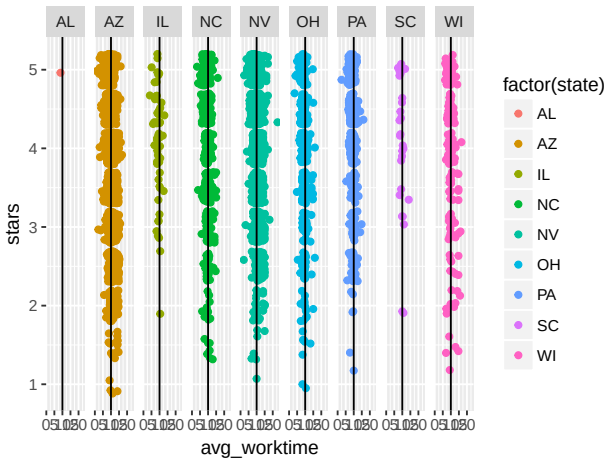


Figure.9 The Relations Between Workday and Stars

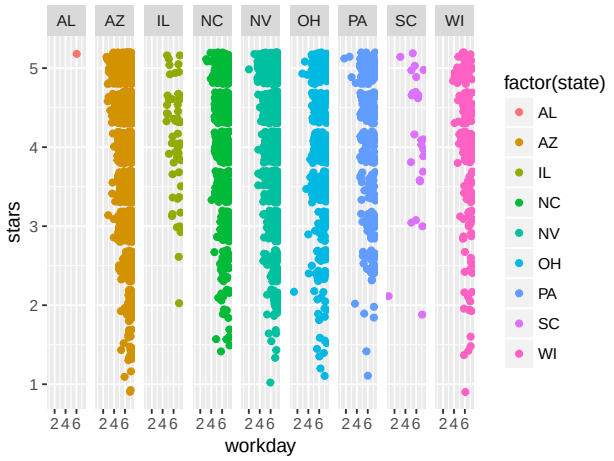
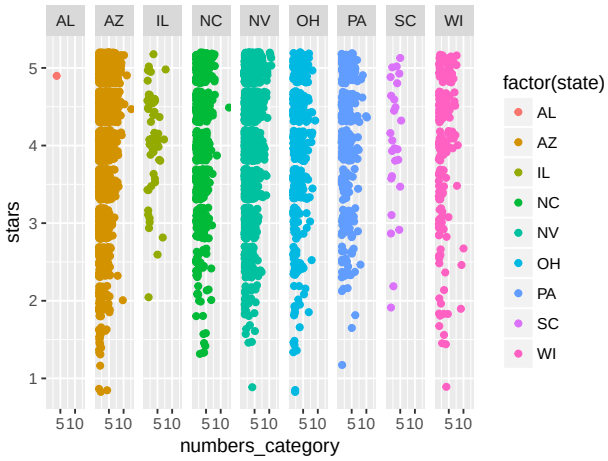


Figure.10 The Relations Between Workday and Stars



Besides those plots above, I also make other plots that contains other variables' information. For example, we can see from the plot of bike parking (Figure.3)that most of the high ranking beauty shops provides bike parking services. What's more, look at the density plot of Appointmentonly Data(Figure.5), it shows that most shops with 5 stars needs an appointment. [see Appendix.1.(2)] As these variables are not the major ones that I want to explore, so I put these plot in the Appendix.

4. Model Fitting

The goal of the project is to figure out what variables would have impact on the stars in yelp. We should define the varibles first. -Outcome variable: stars. -Predictors: Average worktime in a week, Workday, Review

counts, Numbers of category, BikeParking, BusinessAcceptsCreditCards, ByAppointmentOnly, PriceRange, number_Pway. -Group: State.

Since the review counts varies a lot in different shops, I do a linear tranformation of it(Centering by subtracting the mean of the data). And since state AL has only one observation, I change AZ as the baseline of the model. As the outcome variable ordinal catergory variables, I use multinomial logistic regression to fit the model.First, I put part of the predict variables into the model.(I choose Average worktime in a week, Workday, Review counts, Numbers of category as variables in the first model.)

```
total$review_count <- total$review_count-mean(total$review_count)
total <- within(total, state <- relevel(factor(state), ref = "AZ"))

fit1 <-polr(ordered(stars) ~ review_count+numbers_category+avg_worktime+workday+factor(state),
            data=total)
#display(fit1)
fit2 <-polr(ordered(stars) ~ review_count+numbers_category+avg_worktime+workday+factor(state)+
              factor(BikeParking)+factor(BusinessAcceptsCreditCards)+factor(Pricerange)+
              factor(ByAppointmentOnly)+number_Pway+sumcheckin, data=total)
#display(fit2)
fit3 <- lm(stars ~ review_count+numbers_category+avg_worktime+workday+factor(state),
           data = total)
#summary(fit3)
fit4 <-lm(stars ~ review_count+numbers_category+avg_worktime+workday+factor(state)+
            factor(BikeParking)+factor(BusinessAcceptsCreditCards)+factor(Pricerange)+
            factor(ByAppointmentOnly)+number_Pway+sumcheckin, data=total)
#summary(fit4)
```

The coeffients of Fit 1 model have been shown below. For review counts (and other continuous variables), the interpretation is that when review counts moves 1 unit, the odds of moving from "1"" applying to "1.5" or "1.5" applying (or from the lower and middle categories to the high category) are multiplied by exp(0.001). What's more, we can find that, only review counts,number of category and state AL and NV have positive impacts on the stars. Others are have negative impacts on stars. When looking at the linear model, it is the same as multinomial model that only review counts,number of category and state AL and NV have positive impacts on the stars. Others are have negative impacts on stars.In the linear model, we can find that one unit of review counts changes can result in 0.001 changes of the satrs.One unit of the number of category can result in 0.060 changes in the stars. Other variables' impacts on the stars can be found in the table below.

Table 1: Coefficients of Multinomial Model

	round(coef(fit1, 5), digit = 4)
review_count	0.001
numbers_category	0.149
avg_worktime	-0.259
workday	-0.617
factor(state)AL	9.625
factor(state)IL	-0.349
factor(state)NC	-0.354
factor(state)NV	0.080
factor(state)OH	-0.633
factor(state)PA	-0.325
factor(state)SC	-0.370
factor(state)WI	-0.211

Table 2: Coefficients of Linear Model

	round(coef(fit3, 5), digit = 4)
(Intercept)	6.368
review_count	0.001
numbers_category	0.060
avg_worktime	-0.107
workday	-0.223
factor(state)AL	0.700
factor(state)IL	-0.072
factor(state)NC	-0.133
factor(state)NV	0.037
factor(state)OH	-0.266
factor(state)PA	-0.118
factor(state)SC	-0.183
factor(state)WI	-0.107

Table 3: AIC Values Table

	df	AIC
fit1	20	19994
fit2	28	19558
fit3	14	13666
fit4	22	13289

```
##      obs
## pred     1  1.5    2  2.5    3  3.5    4  4.5    5
##   1      0    0    0    0    0    0    0    0    0
##   1.5    0    0    0    0    0    0    0    0    0
##   2      0    0    0    0    0    0    0    0    0
##   2.5    0    0    1    1    2    4    2    3    1
##   3      0    4    7   26   23   20   12    4    8
##   3.5    2   12   21   36   52   52   53   38   15
##   4      3   14   58  150  244  386  486  390  262
##   4.5    1    5   13   38  105  184  273  279  271
##   5      3    7   18   39   77  168  382  672 1088
```

After fitting the multinomial model with variables that I am interested in, I found that the model are not fitting well. Then I add some other variables that may help explain the variations and fit a new model(fit2) . Besides, I fit the linear model with 5 variables (fit3) and 10 variables (fit4) to compare the difference between multinomial model and linear model.

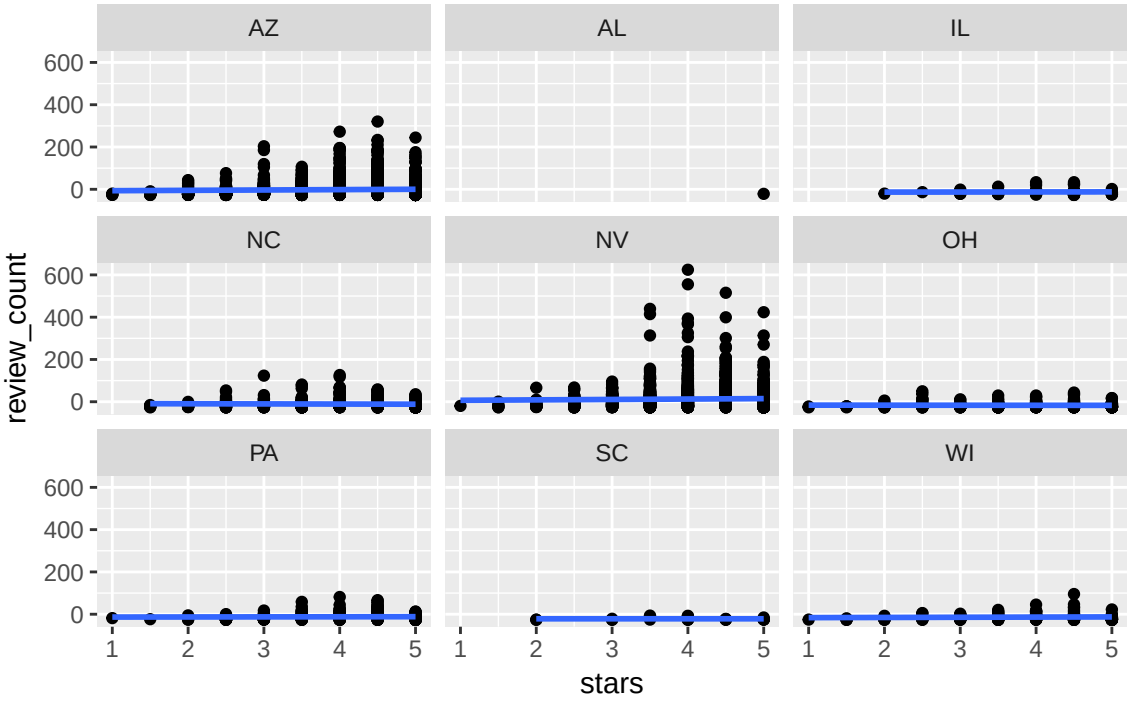
By comparing the AIC of the model, I find that with the increase of predicted variables, the variations of model can be reduced. What's more, it is surprising that linear models are fitted better than the multinomial logistic model.

The predicted values of multinomial model(fit1) are shown in the last table. It seems that the preicted values are not really similar to the observed values.

Mixed Effect Model

In this part, I want to figure out the random effect in different state with different review_count, so I make a plot of review count, group by different state.

Figure.11 The Plot of Random Effect in Each State with Review Cou



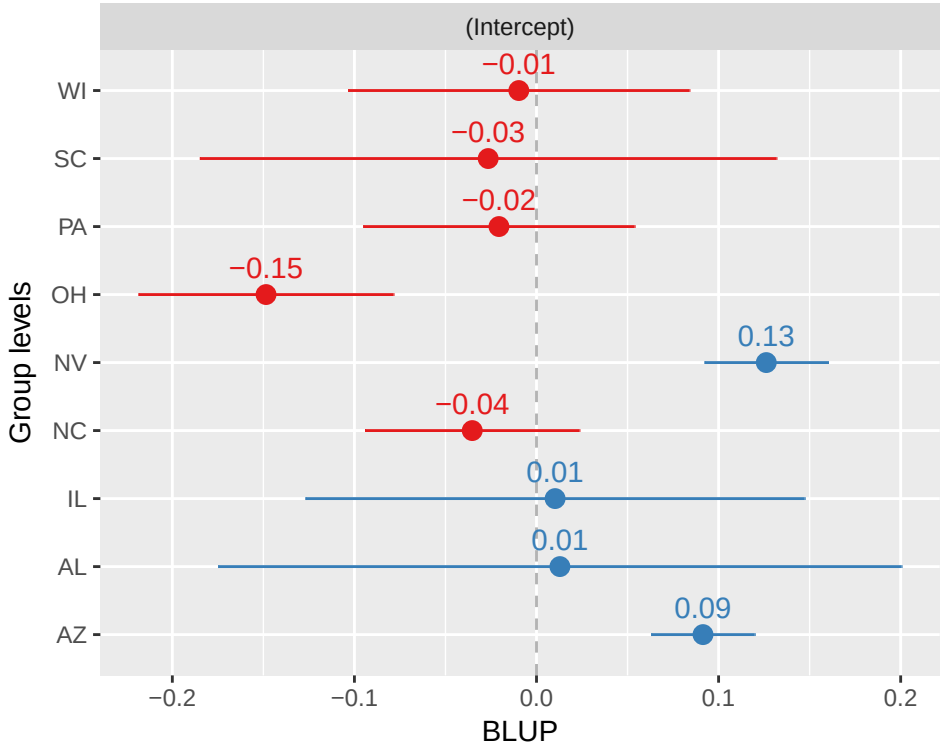
From the plot above, we see little random effect on each states with difference of review count. So I decide to find the random effect of intercept and fit a model with random intercept.

```
mod1 <- lmer(stars ~ review_count+numbers_category+avg_worktime+workday+(1|state),data=total )
knitr::kable(as.data.frame(fixef(mod1)),caption = "Fixed Effect of Model")
```

Table 4: Fixed Effect of Model

	fixef(mod1)
(Intercept)	6.268
review_count	0.001
numbers_category	0.060
avg_worktime	-0.107
workday	-0.222

```
sjp.lmer(mod1, y.offset = .4,title = "Figure.12 Random Effect of Model")
```



The fixed effect shows that review counts and category counts have positive effect on the stars, average worktime and workday have negative effect on the stars. In addition, according to the plot above, we can see that shops in SC,NV,IL,AZ,Al are getting higher stars than other states.

5.Conclusion

In this project, I fit four regression model and a mixed effect model. According to the result of the multinomial regression, I find that review counts,number of category and state AL and NV have positive impacts on the stars while others are not. What's more, By increasing predict variables in the same model, we can have better understanding of the variations of the model. Moreover, it is surprising that using simple linear regressionis better to fit the data than using a multinomial regression. Besides, considering the random effect of different states, the beauty shops in SC,NV,IL,AZ,Al are getting higher stars than those in WI,PA,OH and NC. So there have a random effect in each states. The limitations of my project is model and predict variables' selection. Thus, future studies could focus on these two direction to improve the outcome.

Appendix

1. Exploratory Data Analysis

(1).Distribution about the

Figure.1 Distribution of Shops in US

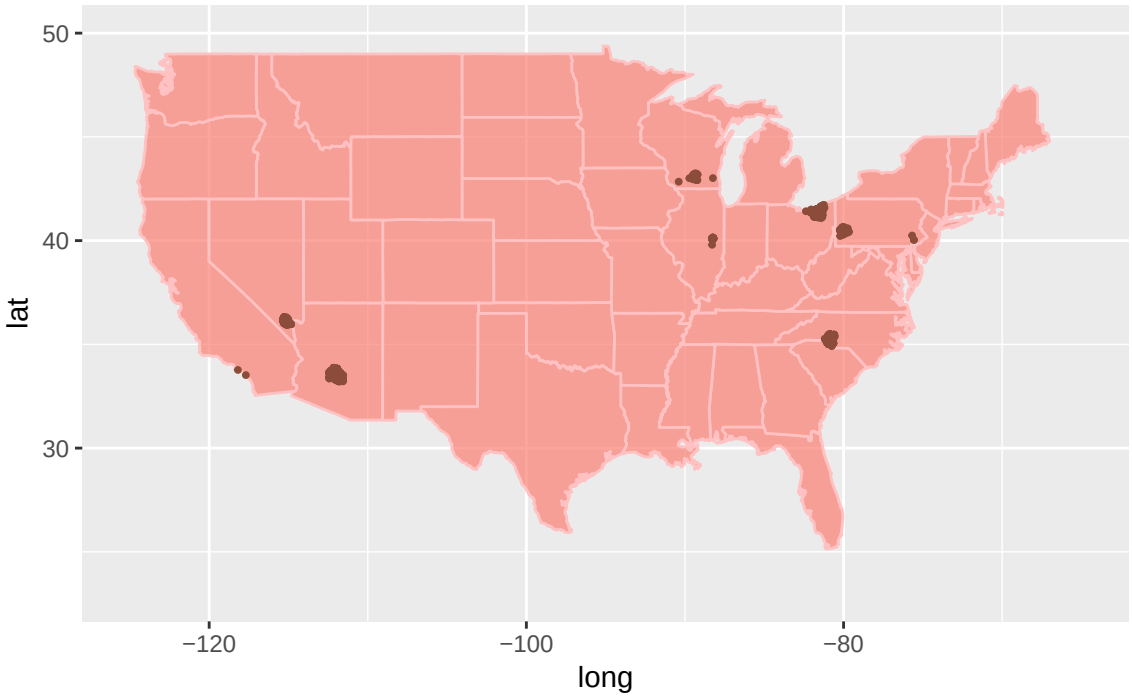
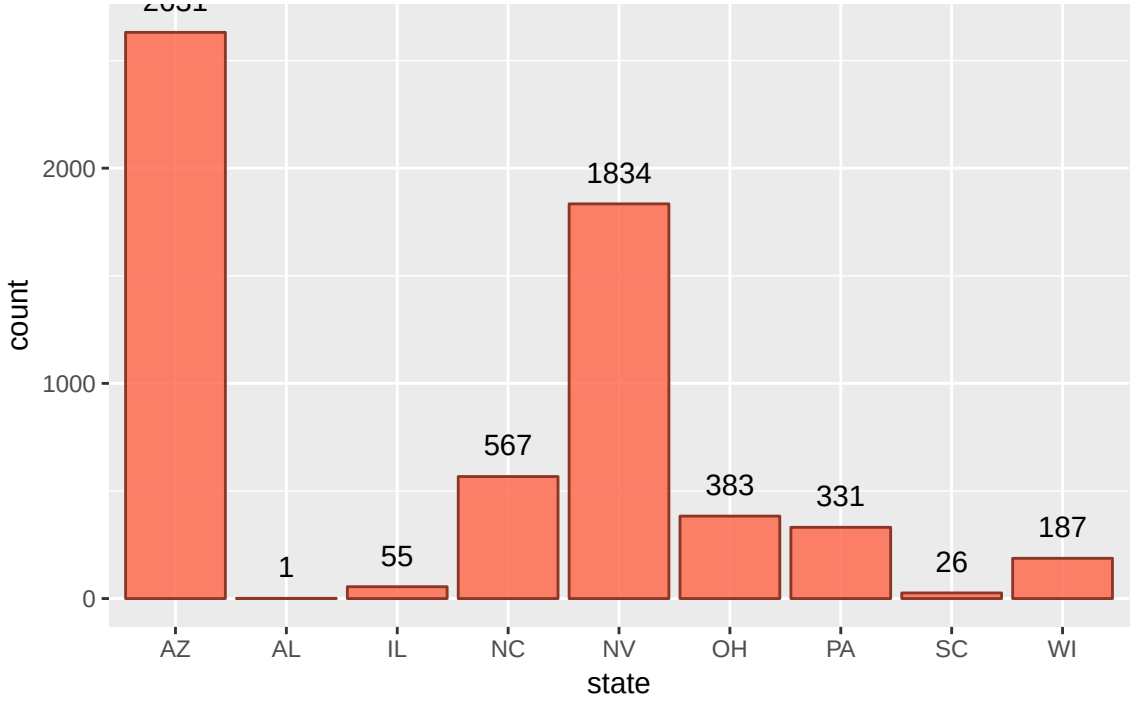


Figure.2 Number of Shops in each State



(2) Plot of other variables

Figure.3 The Plot of BikeParking

Figure.4 The Plot of BusinessAcceptsCreditCards

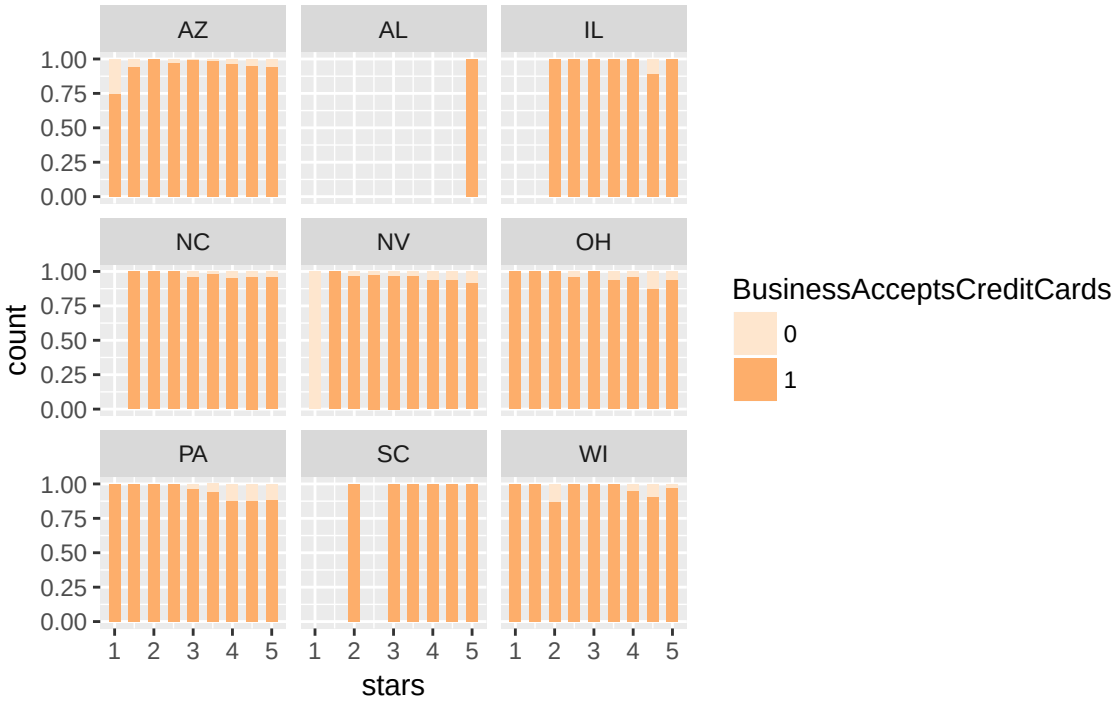


Figure.5 The Plot of Appointmentonly

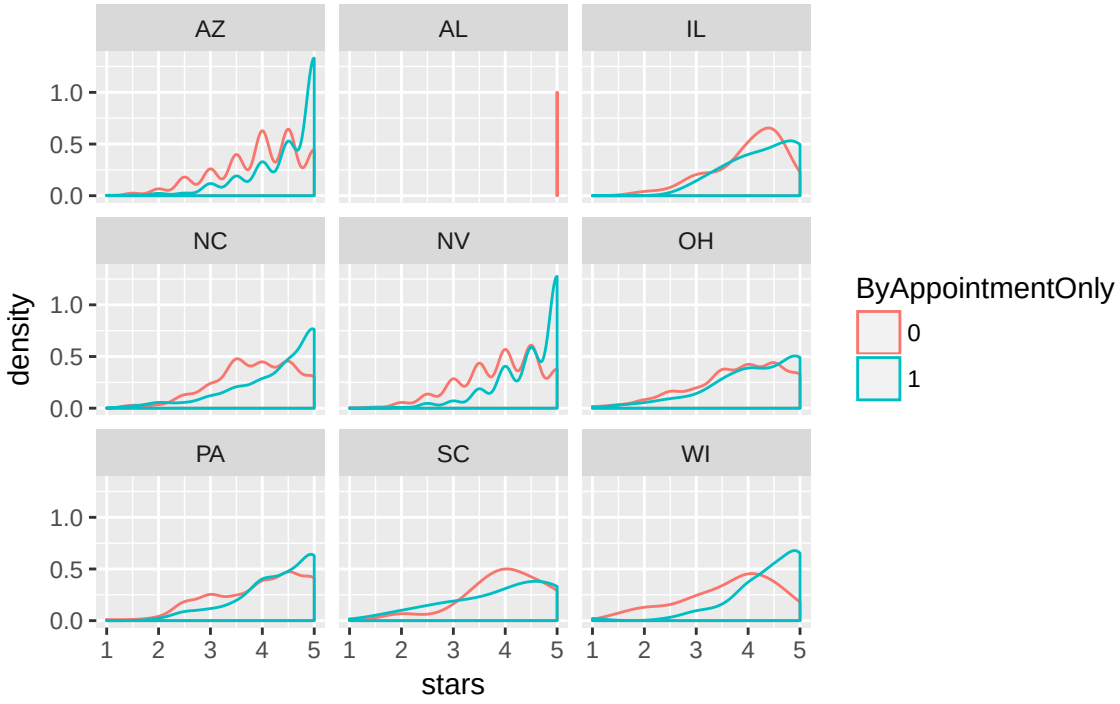


Figure.6 The Plot of Number of Parking Way

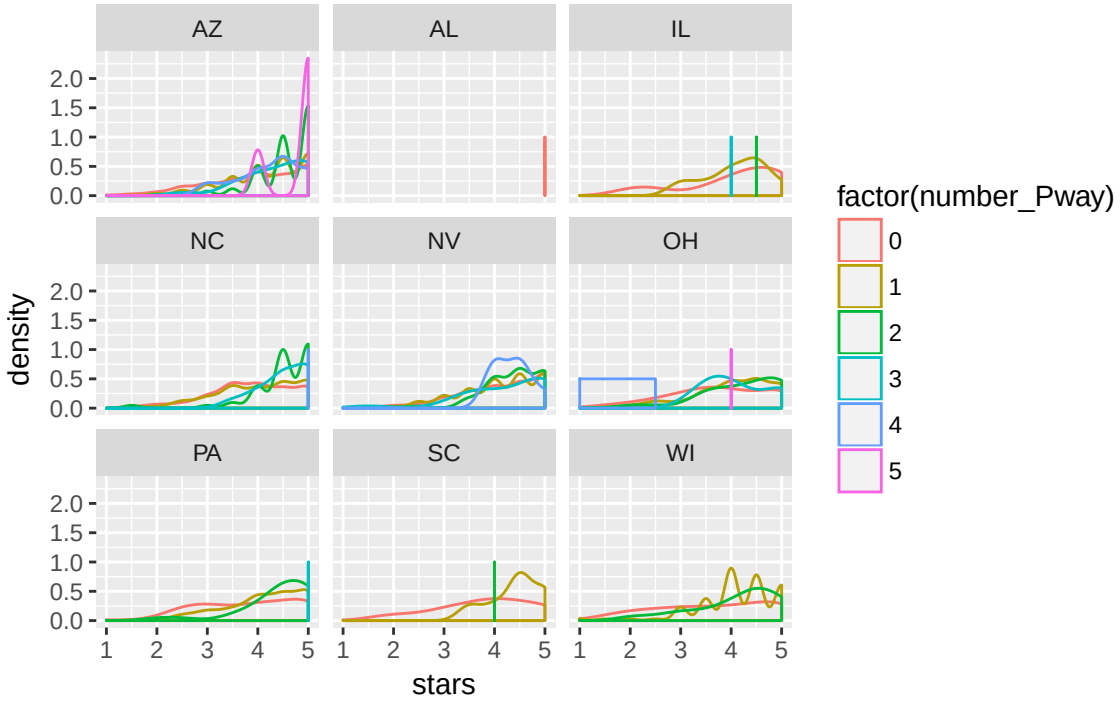


Figure.7 The Plot of Pricerange

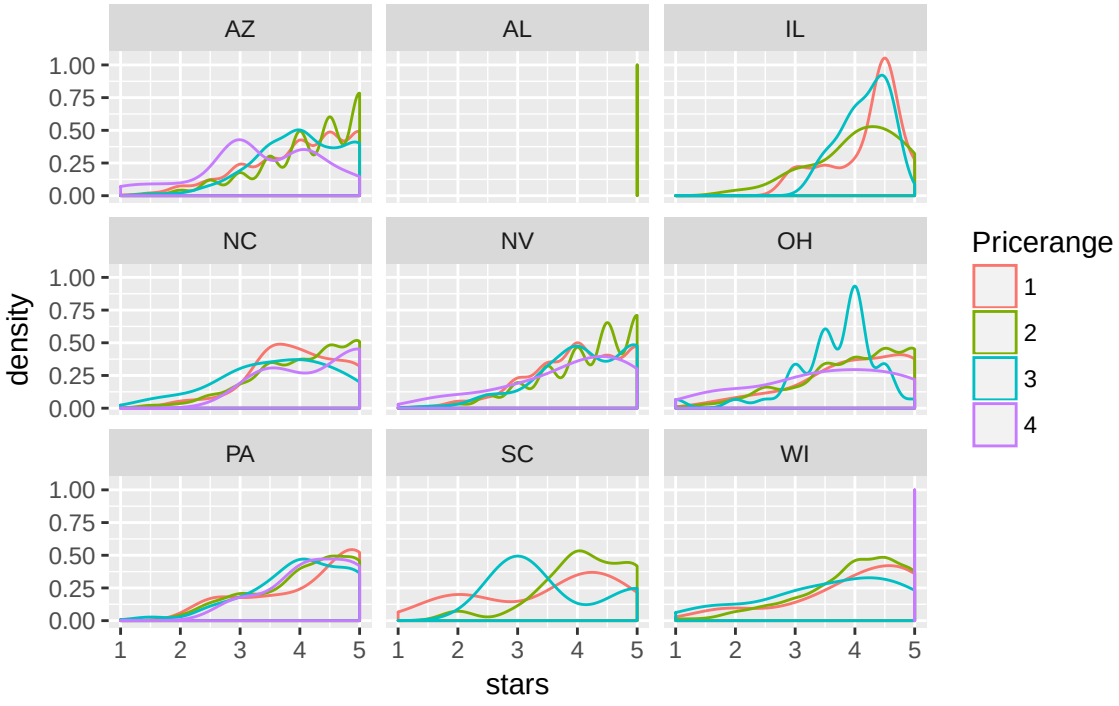


Figure.8 The Plot of Checkin