Factors that Influence Beauty Shops Stars

Jiawei Li

2017/11/18

Data Cleaning

First load the packages that are needed.

```
library(ggplot2)
library(data.table)
library(RMySQL)
library(foreign)
library(dplyr)
library(stringr)
library(tidyr)
library(ggmap)
library(maps)
library(lme4)
library(arm)
library(MASS)
library(VGAM)
library(kableExtra)
library(sjPlot)
```

Second, read the data.

1.Connecting to MySQL:

```
mydb = dbConnect(MySQL(), user='mssp', password='mssp2017', dbname='yelp_db', host='45.63.90.29')
```

2.Listing Tables and Fields:

```
#Return a list of the tables in our connection.
dbListTables(mydb)
```

```
##  [1] "attribute"   "business"    "category"    "checkin"     "elite_years"
##  [6] "friend"      "hours"       "photo"       "review"      "tip"
## [11] "user"
```

I choose three table and check their columns. The tables are business, category and hours.

```
dbListFields(mydb, 'business')
```

```
##  [1] "id"            "name"          "neighborhood" "address"
##  [5] "city"          "state"         "postal_code"  "latitude"
##  [9] "longitude"     "stars"         "review_count" "is_open"
```

```
dbListFields(mydb, 'category')
```

```
## [1] "business_id" "category"
```

```
dbListFields(mydb, 'hours')
```

```
## [1] "hours"       "business_id"
```

3.Retrieving data from MySQL:

```
business.sql = dbSendQuery(mydb, "select * from business ")  # still in mysql
business = fetch(business.sql, n = -1)    # fetch back to R
dbClearResult(business.sql)
```

```
## [1] TRUE
```

```
catergory.sql = dbSendQuery(mydb, "select * from category ")
category = fetch(catergory.sql, n = -1)
dbClearResult(catergory.sql)
```

```
## [1] TRUE
```

```
hours.sql = dbSendQuery(mydb, "select * from hours ")
hours = fetch(hours.sql, n = -1)
dbClearResult(hours.sql)
```

```
## [1] TRUE
```

Data cleaning.

First, 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 fro each shop.

Second,as the original dataset of hours is not good enough to do the analysis, I do some the 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).

At last,I eliminate those shops that are not open,merge three dataframe together and select the shops in the US. What's more, the number of review counts are divided it by 10, since it varies a lot.

```
#1.
category2 <- filter(category, category %in% c("Beauty & Spas","Acne Treatment",
                                              "Cosmetics & Beauty Supply","Day Spas",
                                              "Erotic Massage","Eyebrow Services",
                                              "Eyelash Service","Teeth Whitening","Tanning",
                                              "Tattoo","Skin Care","Piercing",
                                              "Permanent Makeup","Perfume","Nail Salons",
                                              "Medical Spas","Massage","Makeup Artists",
                                              "Hot Springs","Hair Salons","Hair Extensions",
                                              "Hair Loss Centers","Hair Removal",
                                              "Laser Hair Removal","Sugaring",
                                              "Threading Services","Waxing","Blow Dry/Out Services",
                                              "Hair Extensions","Hair Stylist","Men's Hair Salons",
                                              "Nail Technicians","Spray Tanning",
                                              "Tanning Beds","Barbers"))

category3<-data.frame(table(category2$business_id))
names(category3)[1]<-paste("business_id")
```

```r
names(category3)[2]<-paste("numbers_category")

#2.
hours2 <-separate(hours, hours, c("w","start","t","endt","t2"))
hours3 <-dplyr::select(hours2, w, start, endt, business_id)

hours3 <-tbl_df(hours3)
hours3$endt <- as.numeric(hours3$endt)
hours3$start <- as.numeric(hours3$start)
hours3$worktime <- abs(hours3$endt-hours3$start)
hours4 <- dplyr::select(hours3, worktime,business_id)
options(digits=3)
hours4 <- hours4 %>% group_by(business_id) %>% summarise(mean(worktime))
names(hours4)[2]<-paste("avg_worktime")
hours4 <- hours4[-which(hours4$avg_worktime == 0), ]

hours5<-data.frame(table(hours3$business_id))
names(hours5)[1]<-paste("business_id")
names(hours5)[2]<-paste("workday")
hours6 <- merge(hours4,hours5,by=c("business_id"))
hour.category <- merge(hours6,category3,by=c("business_id"))

#3.
business1 <- business[-which(business$is_open == 0), ]
total <- merge(hour.category,business1,by.x="business_id",by.y = "id")
total$review_count <- total$review_count/10
total <- filter(total, state %in% c("AL", "AK","AZ","AR","CA","CO","CT","DE","FL","GA","HI","ID","IL",
                             "IN","IA","KS","KY","LA","ME","MD","MA","MI","MN","MS","MO","MT","NE",
                             "NV","NH","NJ","NM","NY","NC","ND","OH","OK","OR","PA","RI","SC","SD",
                             "TN","TX","UT","VT","VA","WA","WV","WI","WY"))
dim(total)
```

```
## [1] 8586   15
```

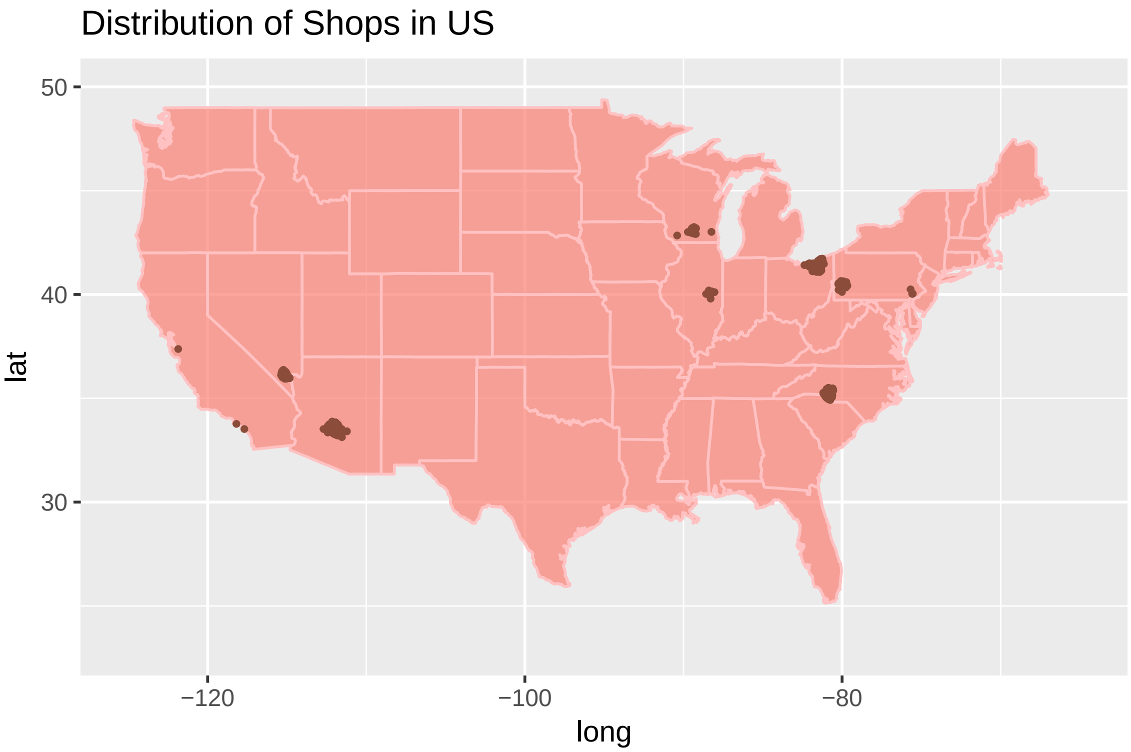
```r
write.csv(total, file="Beauty_Shops_data.csv")
```

Exploratory Data Analysis

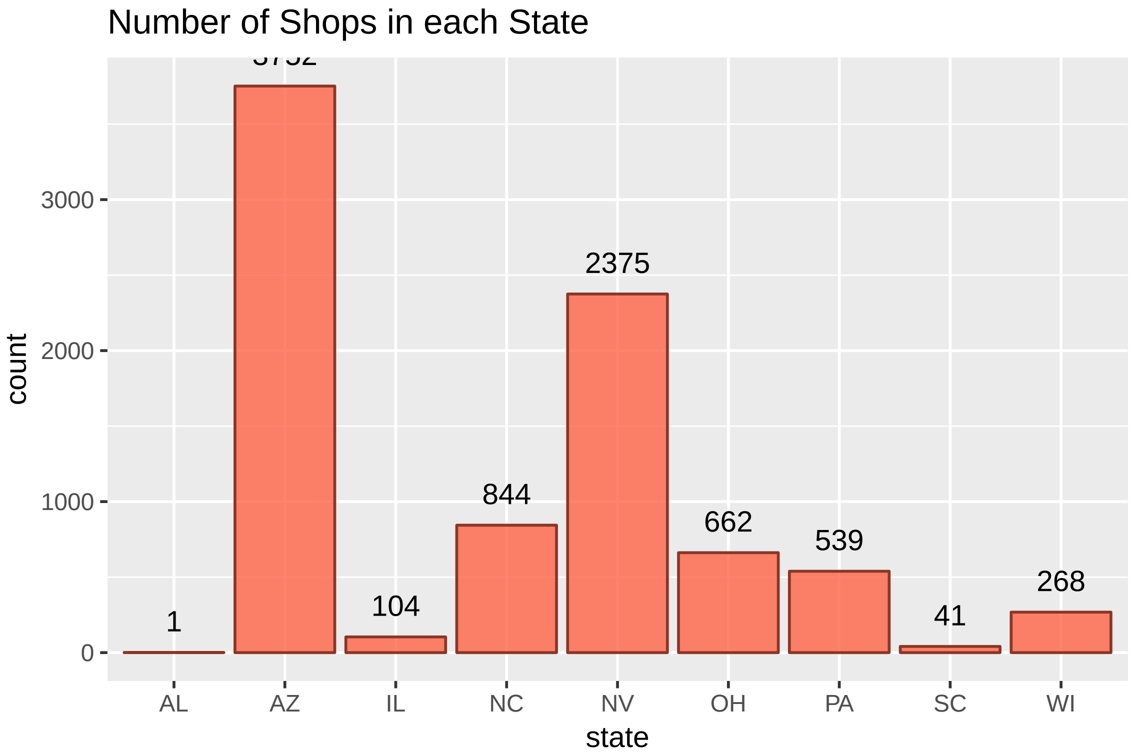
In this part, I would like to find out the dostribution of the beauty shpos in the US first, and then try to figure out the variables that may have impact on the stars of the shops.

We can see from the map below that after cleaning the data, 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.

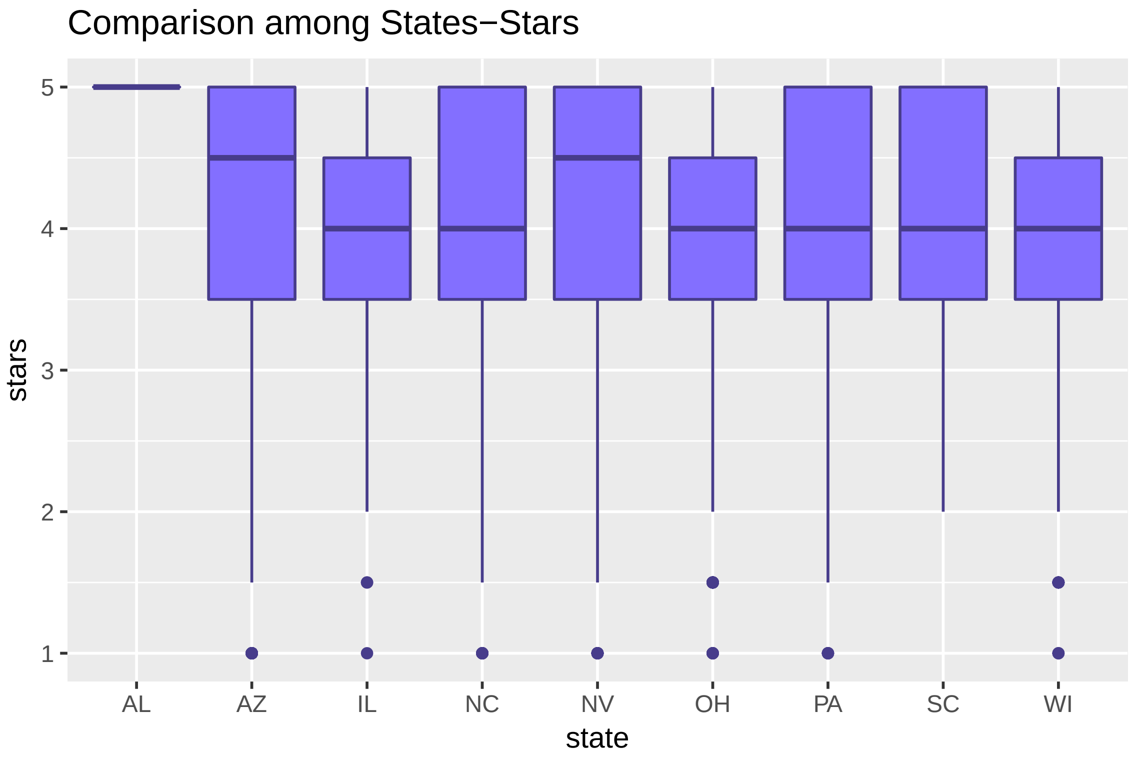
```r
#The distribution of the beauty shops in the US.
us <-map_data('state')
ggplot()+
  geom_polygon(data = us,aes(x=long,y=lat,group=group),color='rosybrown1',fill='salmon',alpha=.7)+
  geom_point(aes(x=longitude,y=latitude),size=.7,color='salmon4',data = total)+
  xlim(-125,-65)+ylim(23,50)+labs(title="Distribution of Shops in US")
```



```
#The number of the beauty shops and the average stars in each state.
ggplot(data.frame(total),aes(x=state))+geom_bar(fill = "tomato",alpha=0.8,colour= "tomato4")+
  geom_text(stat='count',aes(label=..count..),vjust=-1)+labs(title="Number of Shops in each State")
```



```
#The average stars in different state.#Boxplot of the stars, group by state.
ggplot(total)+geom_boxplot(aes(state,stars),fill="slateblue1",colour="slateblue4")+
  ggtitle("Comparison among States-Stars")
```

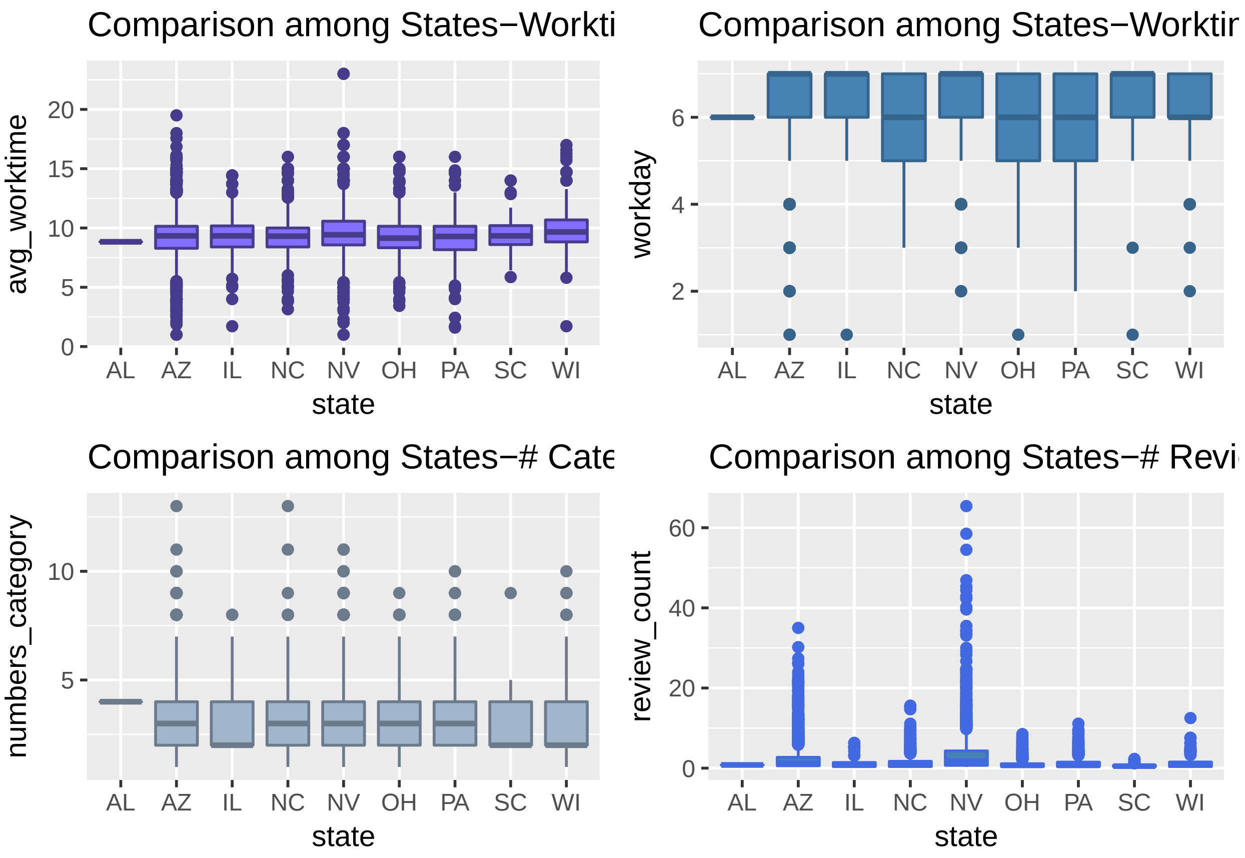


By comparing the stars in different state, we can see that the mean of different states are within 4-5 stars. Since in AL, there is only 1 observation, so the mean is equal to the value of stars in that observation. We can see from the plot that these boxplot is comparatively tall, which means that people 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.

The plots below shows the comparison among states with different variabes.

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.

```
gridExtra::grid.arrange(
  ggplot(total)+geom_boxplot(aes(state,avg_worktime),fill="slateblue1",colour="slateblue4")+
    ggtitle("Comparison among States-Worktime"),
  ggplot(total)+geom_boxplot(aes(state,workday),fill="steelblue",colour="steelblue4")+
    ggtitle("Comparison among States-Worktime"),
  ggplot(total)+geom_boxplot(aes(state,numbers_category),fill="slategray3",colour="slategray4")+
    ggtitle("Comparison among States-# Category"),
  ggplot(total)+geom_boxplot(aes(state,review_count),fill="steelblue",colour="royalblue")+
    ggtitle("Comparison among States-# Review"),
  ncol=2)
```



Model Fitting

Since I want to figure out what variables would have impact on the stars in yelp, the outcome variable in this project should be stars, and the predictors are average worktime in a week, workday, review counts, numbers of category and the state. As the outcome variable is count numbers, I used possion regression and negative binomial regression to fit the model. Pay attention, in order to use the poisson model and negative binomial model best, the stars of the shops has been scaled by multiplying 2. The scale table is listed below.

```r
stars <-matrix(c(1,1.5,2,2.5,3,3.5,4,4.5,5,2,3,4,5,6,7,8,9,10),ncol=2,byrow=F)
colnames(stars)<-c("Original Stars","Scaled Stars")
stars <- as.table(stars)
knitr::kable(stars) %>%
  kable_styling()
```

	Original Stars	Scaled Stars
A	1.0	2
B	1.5	3
C	2.0	4
D	2.5	5
E	3.0	6
F	3.5	7
G	4.0	8
H	4.5	9
I	5.0	10

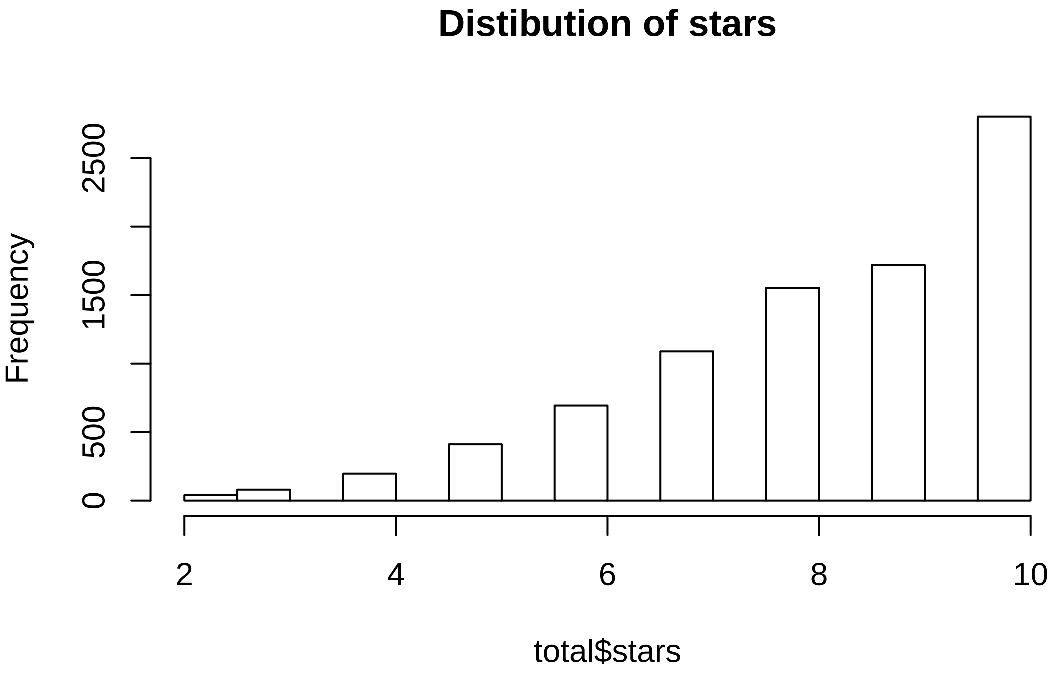
```
total$stars <- total$stars*2
```

In the first possion model, consider state as a factor, I fit a poisson regression model.

```
names(total)
```

```
##  [1] "business_id"      "avg_worktime"     "workday"
##  [4] "numbers_category" "name"             "neighborhood"
##  [7] "address"          "city"             "state"
## [10] "postal_code"      "latitude"         "longitude"
## [13] "stars"            "review_count"     "is_open"
```

```
hist(total$stars,main = "Distibution of stars")
```



```
fit1 <-glm(stars ~ review_count+numbers_category+avg_worktime+workday+factor(state),
          data=total,family=poisson)
summary(fit1)
```

```
##
## Call:
## glm(formula = stars ~ review_count + numbers_category + avg_worktime +
##     workday + factor(state), family = poisson, data = total)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.266  -0.338   0.103   0.409   1.400
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.70792    0.31754    8.53  < 2e-16 ***
## review_count       0.00176    0.00111    1.58     0.11
## numbers_category   0.01457    0.00244    5.97  2.4e-09 ***
## avg_worktime      -0.02073    0.00223   -9.29  < 2e-16 ***
## workday           -0.04699    0.00435  -10.80  < 2e-16 ***
## factor(state)AZ  -0.15171    0.31629   -0.48     0.63
## factor(state)IL  -0.18006    0.31812   -0.57     0.57
## factor(state)NC  -0.19641    0.31647   -0.62     0.53
## factor(state)NV  -0.13833    0.31633   -0.44     0.66
## factor(state)OH  -0.22460    0.31654   -0.71     0.48
## factor(state)PA  -0.19907    0.31660   -0.63     0.53
## factor(state)SC  -0.17013    0.32094   -0.53     0.60
## factor(state)WI  -0.20103    0.31699   -0.63     0.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3562.6  on 8585  degrees of freedom
## Residual deviance: 3148.0  on 8573  degrees of freedom
## AIC: 37013
##
## Number of Fisher Scoring iterations: 4
```

```
qchisq(0.95, df.residual(fit1))
```

```
## [1] 8790
```

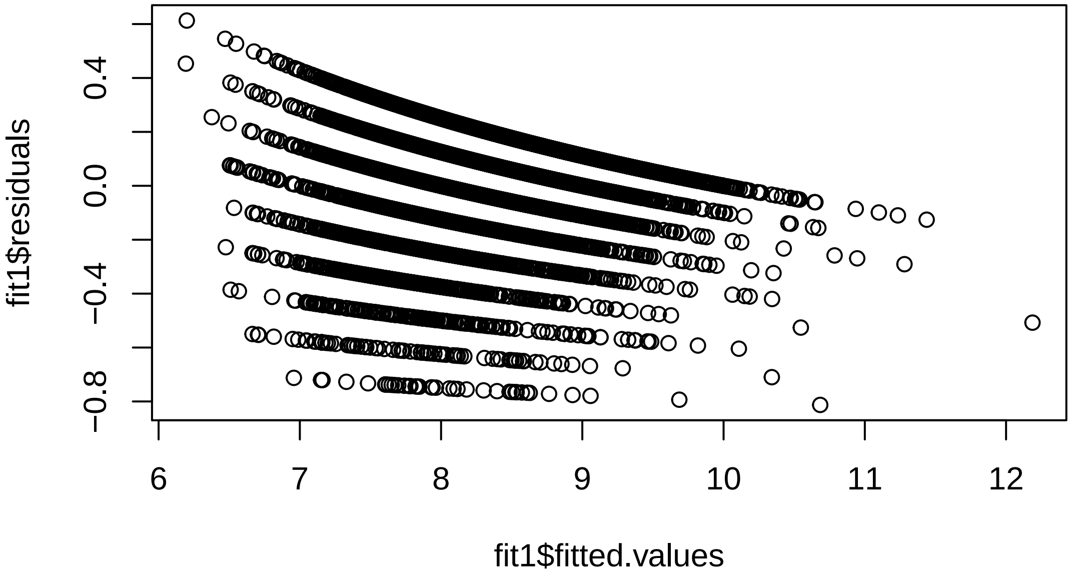
```
deviance(fit1)
```

```
## [1] 3148
```

```
pr <- residuals(fit1,"pearson")
sum(pr^2)
```

```
## [1] 2862
```

```
plot(fit1$fitted.values,fit1$residuals)
```



```
#Pearson residual
res <- residuals(fit1, "pearson")
```

We can see that the model does not fit the data well. The five-percent critical value for a chi-squared with 8573 degreee of freedom is 8790 and the devian is 3148 and Pearson's chi-squared is 2862.

Then I try the quasipoisson model and the negative binomial model. However the result is similar to the poisson result.

```
fit2 <-glm(stars ~ review_count+numbers_category+avg_worktime+workday+factor(state),data=total,family =
#summary(fit2)
fit3 <-glm.nb(stars ~ review_count+numbers_category+avg_worktime+workday+factor(state),data=total)
```

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached
```

```
#summary(fit3)
```

The estimates coefficients of quasipoisson and negative-binomial did not change at all, and the standard errors of negative-binomial and poisson model look similar. AIC of the negative-binomial model is 37015 which is close to the AIC of poisson model.

So I keep fitting the mixed effect model, group by states, and add the random effects on the intercept.

```
mod1 <- lmer(stars ~ review_count+numbers_category+avg_worktime+workday+(1|state),data=total )
summary(mod1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## stars ~ review_count + numbers_category + avg_worktime + workday +
##     (1 | state)
##    Data: total
##
## REML criterion at convergence: 32921
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -5.151 -0.575  0.186  0.733  2.457
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.0574   0.24
##  Residual             2.6937   1.64
## Number of obs: 8586, groups:  state, 9
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      11.74009    0.16712    70.2
## review_count      0.01435    0.00528     2.7
## numbers_category  0.12144    0.01171    10.4
## avg_worktime     -0.16605    0.01039   -16.0
## workday          -0.40170    0.02110   -19.0
##
## Correlation of Fixed Effects:
##             (Intr) rvw_cn nmbrs_ avg_wr
## review_cont  0.099
## nmbrs_ctgry -0.294 -0.120
## avg_worktim -0.324 -0.036  0.090
## workday     -0.569 -0.133  0.024 -0.362
```

```r
ranef(mod1)
```

```
## $state
##    (Intercept)
## AL      0.0342
## AZ      0.2466
## IL      0.0138
## NC     -0.1146
## NV      0.3538
## OH     -0.3221
## PA     -0.1318
## SC      0.0453
## WI     -0.1252
```

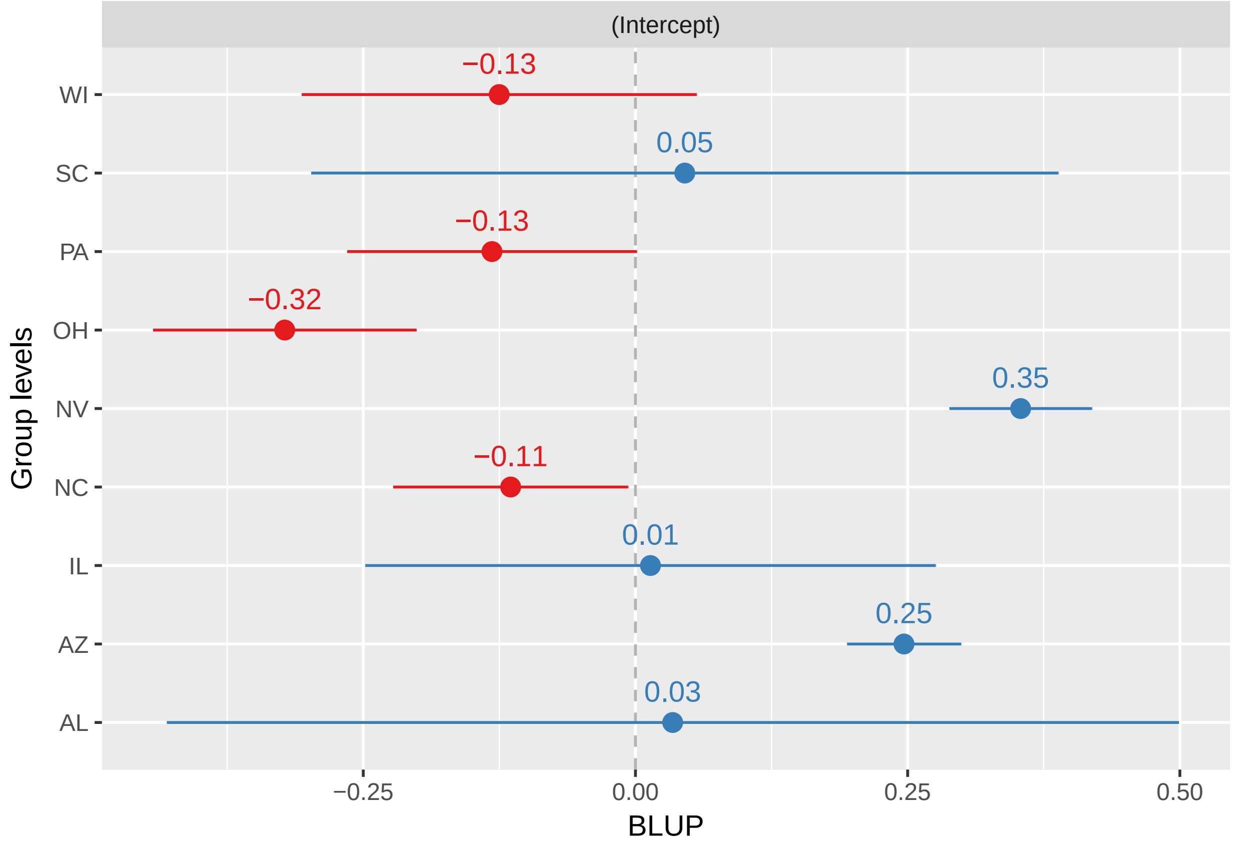
```r
fixef(mod1)
```

```
##     (Intercept)    review_count numbers_category     avg_worktime
##         11.7401          0.0143           0.1214          -0.1661
##         workday
##         -0.4017
```

```r
sjp.lmer(mod1, y.offset = .4)
```

```
## `sjp.lmer()` will become deprecated in the future. Please use `plot_model()` instead.
```

```
## Plotting random effects...
```



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 SC,NV,IL,AZ,Al are getting higher stars than other states.

Conclusion

In this project, I fit three regression model and a mixed effect model. Although the poisson model does not fit very well, it shows us that the increase of review counts and the numbers catergory may have a positive impact on the increase of stars, while the increase of working time and working day may cause a negative impact on the increase of stars. And according to the mixed-effect model, the situation is the same. 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. As there are limitations about the model and the variables I chose, it is better for further study to focus on selecting better model and variables.