Data Mining

Ran Tao

December 1, 2017

Data pre-process

1) Import data

```
ParisAirbnb<-read.csv("Paris Airbnb.csv")
```

Remove unrelevant columns

```
Parisdat<-ParisAirbnb[, c(-2,-5,-6,-7,-13,-15,-16,-18,-21)]
```

Correct French character into English character, as french character cant be stored properly in csv file

```
Parisdat$neighborhood<-as.character(Parisdat$neighborhood)

Parisdat$neighborhood[Parisdat$neighborhood=="Amérique"]="Amerique"
Parisdat$neighborhood[Parisdat$neighborhood=="Champs-Elysées"]="Champs-Elysees"
Parisdat$neighborhood[Parisdat$neighborhood=="Chaussée-dAntin"]="Chaussee-dAntin"
Parisdat$neighborhood[Parisdat$neighborhood=="Folie-Méricourt"]="Folie-Mericourt"
Parisdat$neighborhood[Parisdat$neighborhood=="Grandes-Carrières"]="Grandes-Carrieres"
Parisdat$neighborhood[Parisdat$neighborhood=="Hôpital-Saint-Louis"]="Hopital-Saint-Louis"
Parisdat$neighborhood[Parisdat$neighborhood=="Père-Lachaise"]="Pere-Lachaise"
Parisdat$neighborhood[Parisdat$neighborhood=="Place-Vendôme"]="Place-Vendome"
Parisdat$neighborhood[Parisdat$neighborhood=="Saint-Germain-des-Prés"]="Saint-Germain-des-Pres"
Parisdat$neighborhood[Parisdat$neighborhood=="Salpêtrière"]="Salpetriere"
```

2) Check missing data

```
summary(Parisdat)
```

```
##     room_id            host_id                    room_type
##  Min.   :    2525   Min.   :     2626   Entire home/apt:61763
##  1st Qu.: 5351442   1st Qu.:  7528891   Private room   : 7880
##  Median :11320552   Median : 20706768   Shared room    :  515
##  Mean   :10753849   Mean   : 34352480
##  3rd Qu.:16404387   3rd Qu.: 47692925
##  Max.   :20144084   Max.   :143204619
##
##  neighborhood          reviews       overall_satisfaction  accommodates
##  Length:70158       Min.   :  0.00   Min.   :0.000        Min.   : 1.00
##  Class :character   1st Qu.:  0.00   1st Qu.:0.000        1st Qu.: 2.00
##  Mode  :character   Median :  4.00   Median :4.500        Median : 2.00
##                     Mean   : 14.61   Mean   :2.615        Mean   : 3.07
##                     3rd Qu.: 14.00   3rd Qu.:4.500        3rd Qu.: 4.00
##                     Max.   :529.00   Max.   :5.000        Max.   :16.00
##
##     bedrooms          price                property_type
##  Min.   : 0.000   Min.   :   10.0   Apartment      :66914
##  1st Qu.: 1.000   1st Qu.:   60.0   Loft           :  745
##  Median : 1.000   Median :   84.0   House          :  708
##  Mean   : 1.059   Mean   :  110.8   Condominium    :  494
##  3rd Qu.: 1.000   3rd Qu.:  121.0   Bed & Breakfast:  357
```

```
##  Max.   :10.000   Max.   :11323.0   Other          :  272
##                                     (Other)        :  668
##     latitude       longitude
##  Min.   :48.82   Min.   :2.225
##  1st Qu.:48.85   1st Qu.:2.324
##  Median :48.86   Median :2.348
##  Mean   :48.86   Mean   :2.345
##  3rd Qu.:48.88   3rd Qu.:2.370
##  Max.   :48.90   Max.   :2.468
##
```

there is no NA data, ratings are our dependent variable, it doesn't make sence to include properties with zero reviews

```
Parisdat<-filter(Parisdat, reviews>0)
```

3) Check category data

All the category data has been factored

```
table(Parisdat$room_type)
```

```
##
## Entire home/apt    Private room     Shared room
##           45046            5632             377
```

```
table(Parisdat$neighborhood)
```

```
##
##               Amerique               Archives                Arsenal
##                    682                    662                    489
##        Arts-et-Metiers                Auteuil             Batignolles
##                    676                    769                    908
##                Bel-Air             Belleville                  Bercy
##                    358                    777                    152
##         Bonne-Nouvelle               Chaillot         Champs-Elysees
##                    913                    557                    218
##               Charonne        Chaussee-dAntin           Clignancourt
##                    687                    215                   2490
##                Combat            Croulebarbe        Ecole-Militaire
##                    899                    315                    273
##         Enfants-Rouges              Epinettes                 Europe
##                    664                    909                    391
##      Faubourg-du-Roule    Faubourg-Montmartre        Folie-Mericourt
##                    422                    426                   1288
##                Gaillon                   Gare             Goutte-dOr
##                    134                    553                    678
##      Grandes-Carrieres               Grenelle           Gros-Caillou
##                   1972                   1173                    795
##                 Halles    Hopital-Saint-Louis              Invalides
##                    697                   1060                    157
##     Jardin-des-Plantes            Javel 15Art            La Chapelle
##                    510                    650                    396
##              Madeleine                   Mail         Maison-Blanche
##                    240                    519                    645
##                Monnaie           Montparnasse                 Muette
##                    512                    403                    724
```

```
##                 Necker             Notre-Dame  Notre-Dame-des-Champs
##                    802                    241                    601
##                 Odeon           Palais-Royal     Parc-de-Montsouris
##                    406                    219                    201
##          Pere-Lachaise        Petit-Montrouge                 Picpus
##                    869                    642                    772
##          Place-Vendome         Plaine-Monceau              Plaisance
##                    163                    442                    741
##        Pont-de-Flandre         Porte-Dauphine      Porte-Saint-Denis
##                    320                    429                    726
##     Porte-Saint-Martin          Quinze-Vingts           Rochechouart
##                    998                    689                    829
##               Roquette         Saint-Ambroise            Saint-Avoye
##                   1545                   1117                    597
##          Saint-Fargeau          Saint-Georges Saint-Germain-des-Pres
##                    454                    785                    272
##          Saint-Gervais          Saint-Lambert            Saint-Merri
##                    776                   1149                    554
##    Saint-Thomas-dAquin           Saint-Victor  Saint-Vincent-de-Paul
##                    382                    443                    784
##      Sainte-Marguerite            Salpetriere               Sorbonne
##                   1003                    263                    514
##  St-Germain-lAuxerrois                 Ternes           Val-de-Grace
##                    134                    804                    488
##               Villette               Vivienne
##                    765                    178
```

```
table(Parisdat$property_type)
```

```
##
##          Apartment    Bed & Breakfast              Boat
##              48835                270               35
##     Boutique hotel            Bungalow             Cabin
##                 83                   4                 9
##               Cave         Condominium              Dorm
##                  1                 267                54
##        Earth House         Guest suite         Guesthouse
##                  5                  10               102
##             Hostel               House            In-law
##                 28                 483                 6
##              Loft              Other Serviced apartment
##                580                 179                15
##          Timeshare                Tipi          Townhouse
##                  1                   1                80
##          Treehouse       Vacation home              Villa
##                  1                   1                 5
```

4) Check Numberic Data

we can characterlized id of room and host

```
Parisdat$room_id<-as.character(Parisdat$room_id)
Parisdat$host_id<-as.character(Parisdat$host_id)
```

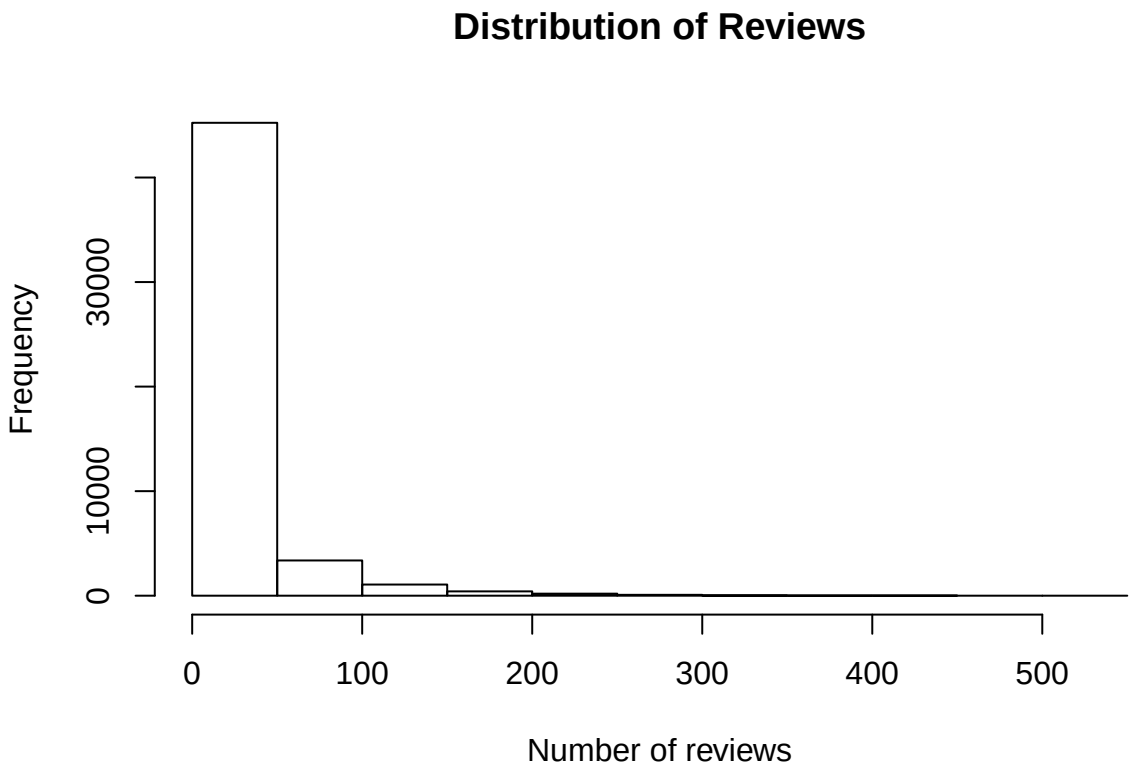
check outliers of numeric variables: number of accommodate, bedroom and price

```
mahal<-mahalanobis(Parisdat[, c(7:9)],
                   colMeans(Parisdat[, c(7:9)]),
                   cov(Parisdat[, c(7:9)], use = "pairwise.complete.obs"))

cutoff<-qchisq(1-0.001, ncol(Parisdat[, c(7:9)]))

paris<-Parisdat[mahal<cutoff, ]

length(unique(paris$room_id))
```

```
## [1] 50406
```

Check ratings (overal satisfication) and number of reviews

```
hist(paris$reviews, main = "Distribution of Reviews", xlab = "Number of reviews")
```




```
ecdf.rev<-ecdf(paris$reviews)
ecdf.rev(50)
```

```
## [1] 0.8974527
```

```
max(paris$reviews)
```

```
## [1] 529
```

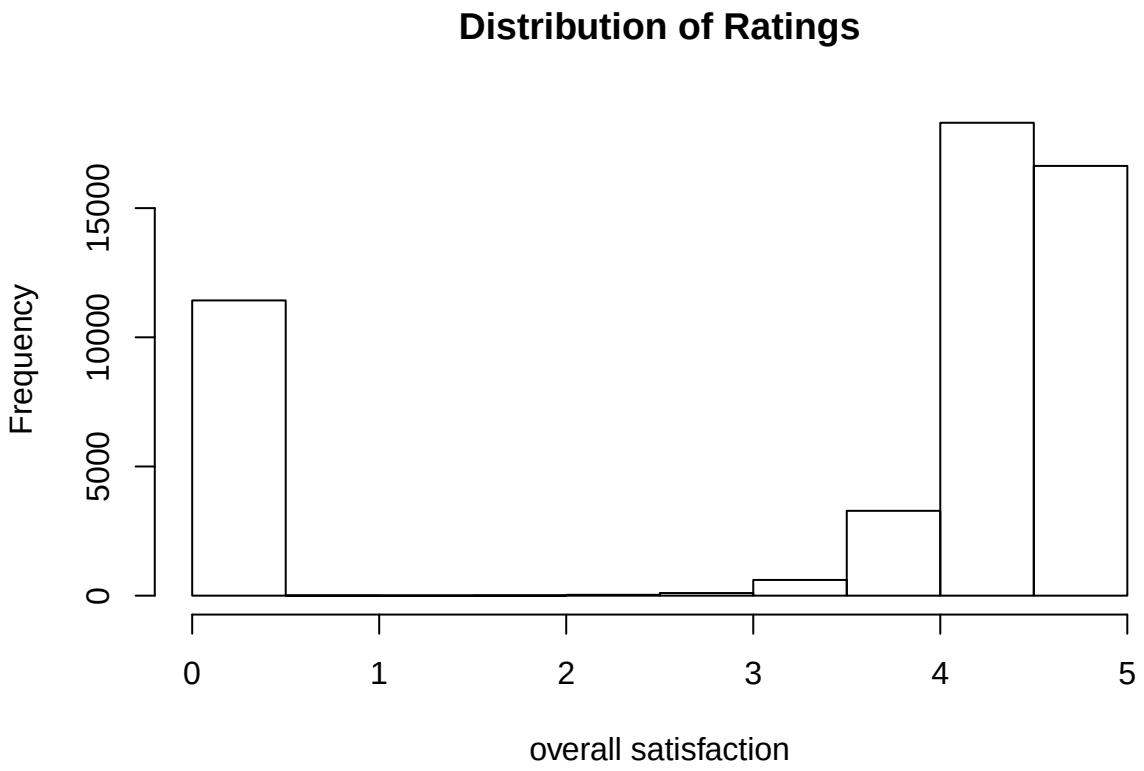
there are 50406 unique rooms in our data after cleaning, as we can see 89.7% of total Airbnb rooms, that are 45104 rooms have less than 50 reviews, while there is 1 room which has 529 reviews, 11 rooms have over 400 reviews, 46 rooms have over 300 reviews, 332 rooms have over 200 reviews, 1800 rooms have over 100 reviews

we can create a table of reviews with number of rooms

```
review<-sqldf("SELECT reviews, COUNT (room_id) FROM paris Group by 1")
colnames(review)<-c("reviews", "Num_room")
```

5) now let's explore the ratings

```
hist(paris$overall_satisfaction, main = "Distribution of Ratings", xlab = "overall satisfaction")
```




```
summary(paris$overall_satisfaction)
```

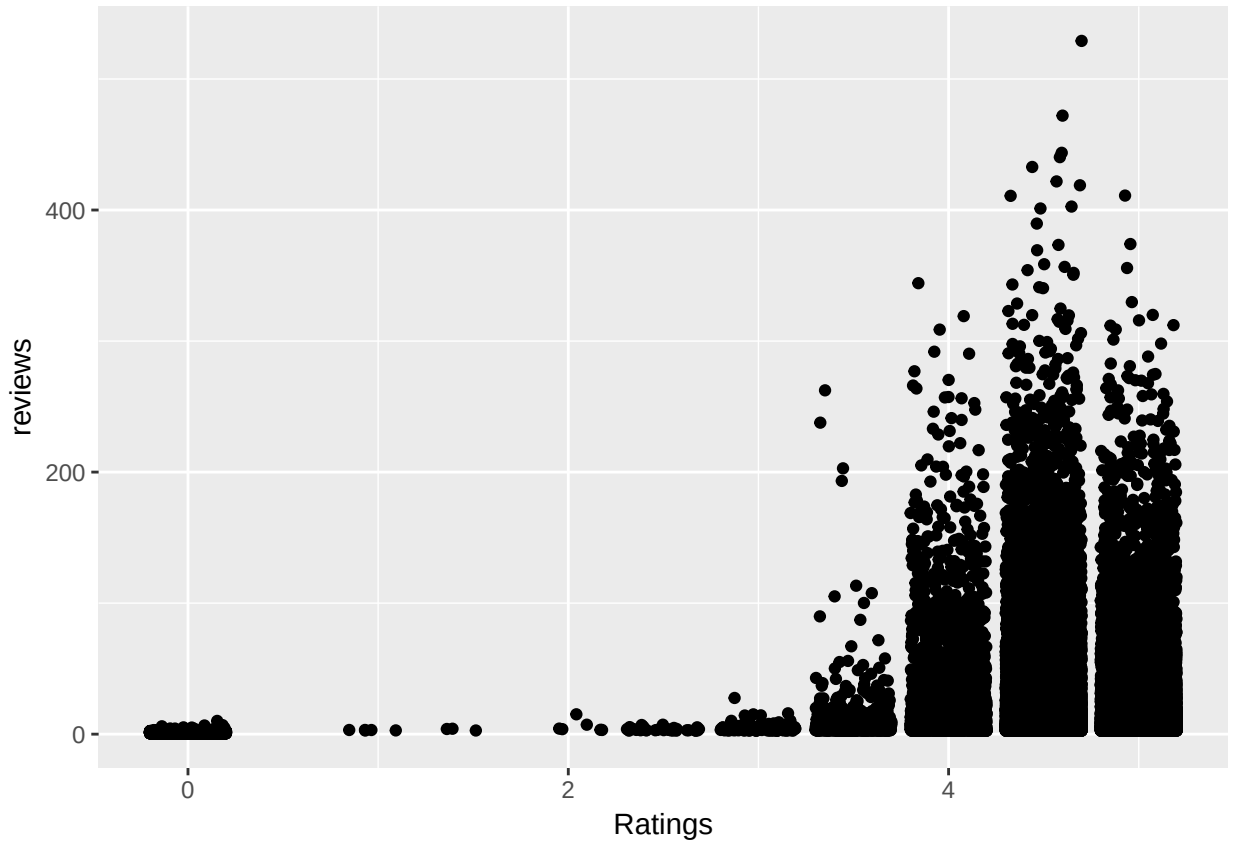
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   4.500   3.595   5.000   5.000
```

```
Rating<-sqldf("SELECT overall_satisfaction, COUNT (room_id) FROM paris Group by 1")
colnames(Rating)<-c("rating","Num_room")
```

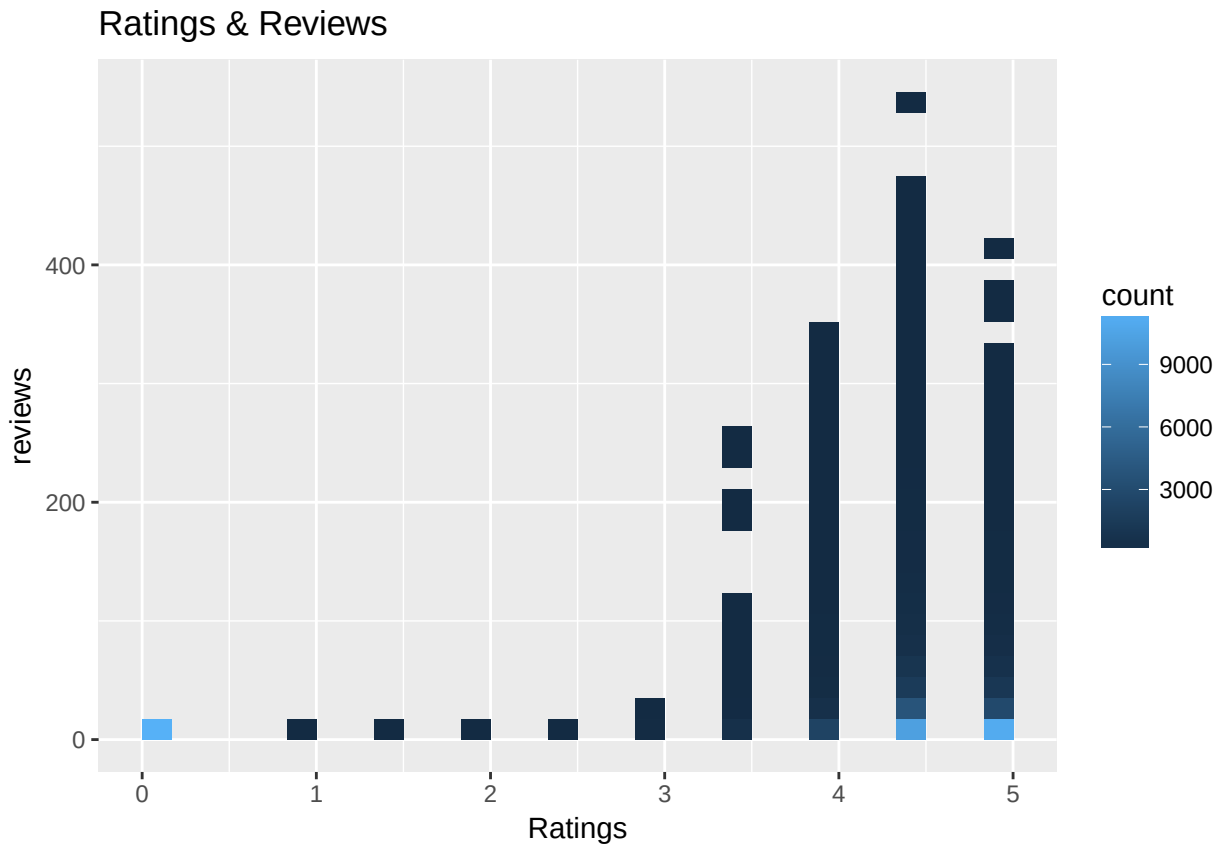
we can see only 22.7% of rooms with ratings less than 1, while 75% of room with ratings higher than 4, and the average rating is 3.6, 25% of rooms have ratings of 5, we can see generally the airbnb properties in paris have good ratings

6) now let's have a look of relationship between ratings and number of reviews

```
ggplot(data=paris, aes(x=overall_satisfaction, y=reviews))+geom_jitter()+xlab("Ratings")
```




```
ggplot(data=paris, aes(x=overall_satisfaction, y=reviews))+geom_bin2d()+xlab("Ratings")+
  ggtitle("Ratings & Reviews")
```



as we can see, generally higher ratings tend to have more reviews

Data manipulation

1) we are going to add a district indicator variable In total, there are 20 districts in Paris Through searching data online, we found the Violence rate per 1000 inhabitants of each district from http://www.lefigaro.fr/actualite-france/2017/01/02/01016-20170102ARTFIG00290-decouvrez-la-carte-des-crimes-et-delits-en-france-et-dans-le-php we will import this district info and add to our data

```
index<-read.csv("Index.csv")
colnames(index)<-c("neighborhood","district","violenceRate")
```

factor district

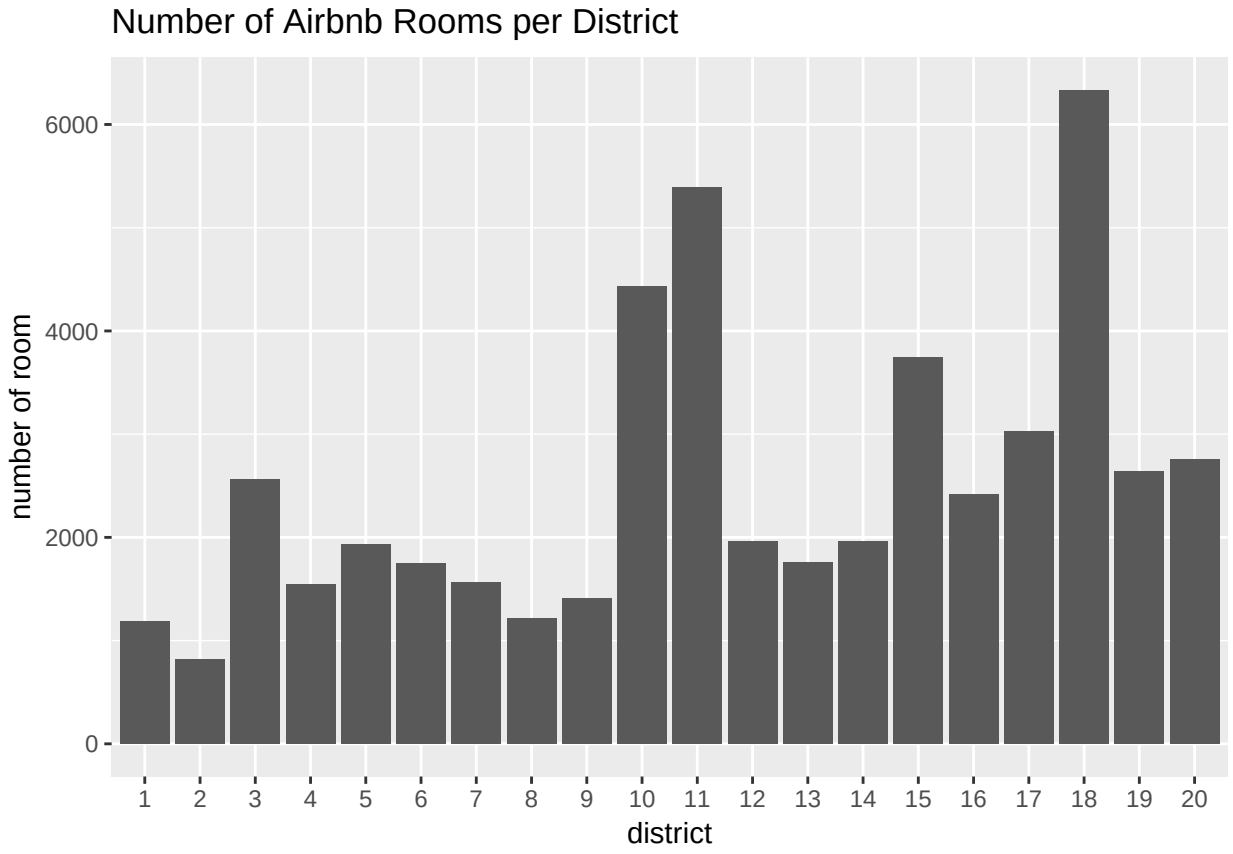
```
index$district<-as.factor(index$district)
```

characterized neighborhood to join two tables

```
index$neighborhood<-as.character(index$neighborhood)
paris$neighborhood<-as.character(paris$neighborhood)

parisdata<-left_join(paris, index, by = "neighborhood")
```

create a district table with number of rooms per district

```
district<-sqldf("SELECT district, COUNT (room_id), Avg(overall_satisfaction),
                Avg(longitude), Avg(latitude) FROM parisdata Group by 1")
colnames(district)<-c("district","Num_room", "Avg_rating", "lon", "lat")

ggplot(data=district, aes(x=district, y=Num_room))+geom_bar(stat = "identity")+
  ggtitle("Number of Airbnb Rooms per District")+ylab("number of room")
```



we also create the map with the circle size showin gthe number of rooms in each district

```
parismap<-qmap("paris", zoom=12, "osm")
```

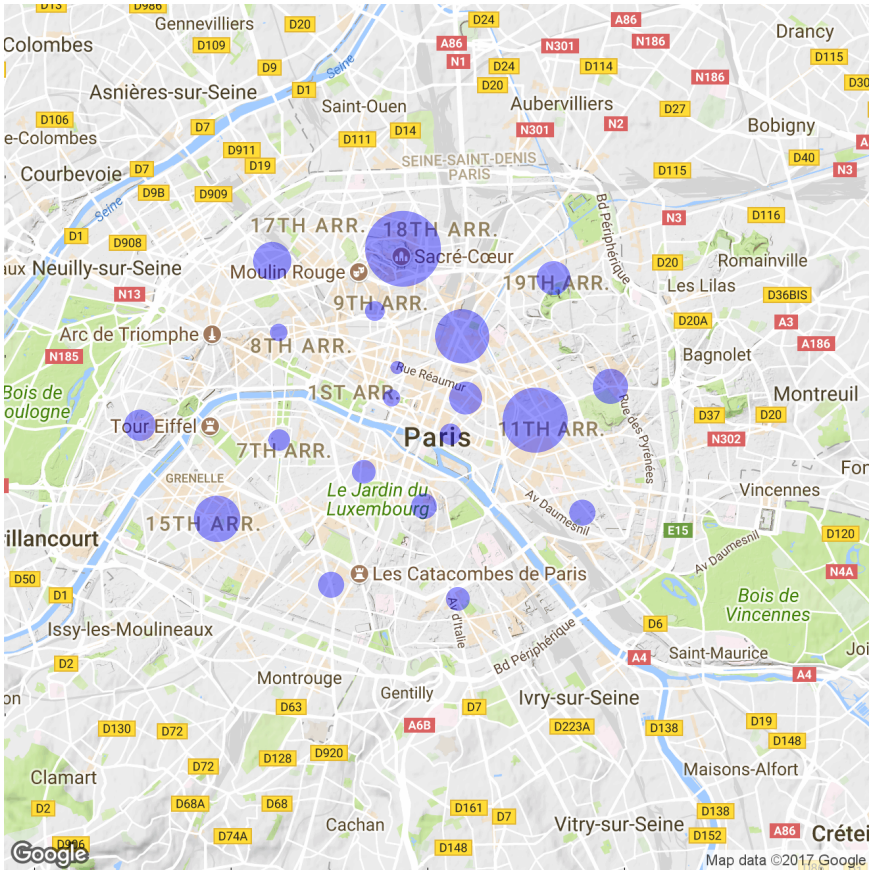
```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=paris&zoom=12&size=640x640&scale=
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=paris&sensor=false
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```

```
parismap+
  geom_point(aes(lon, lat), data=district, col="blue", alpha=0.4, size = (district$Num_room/500))+
  scale_size_discrete(range=range(district$Num_room))
```

```
## Warning: Using size for a discrete variable is not advised.
```



as we can see 2nd district has smallest number of rooms: 817 while 18th district has the highest number of rooms: 6334 10th and 11th have over 4000 room respectively

In conclusion, the sample sizes in all districts are relative large

2) Dependent variable as we know a rating of 5 with 100 reviews is different from a rating of 5 with 1 reviews, therefore, we are going to create the weighted rating as our response. the weight will be determined by the number of reviews, we use sigmoid function to determine the weight. as the increase of review can't indicate the linear increase of the weighted rating, we can't say a rating of 5 with 1000 reviews is 1000 times better than the rating of 5 with 1 reviews. we also found the previous study of yelp rating using sigmoid function, http://www.developintelligence.com/blog/2017/06/practical-neural-networks-keras-classifying-yelp-reviews/

```
sig.function<-function(data){
  y=1/(1+6.16*exp(-0.18*data))
  return(y)
}


weight<-sig.function(parisdata$reviews)
weight<-as.data.frame(weight)
colnames(weight)<-"weight"
```

add to our dataset

```
parisdata$weight<-weight$weight
```

our independent variable

```
parisdata$weightedRating<-parisdata$overall_satisfaction*parisdata$weight
```

3) Predictors our group level indicator: violence rate per district individual indicator: catergory: room type, property type numeric: price, accommodates, bedrooms

EDA

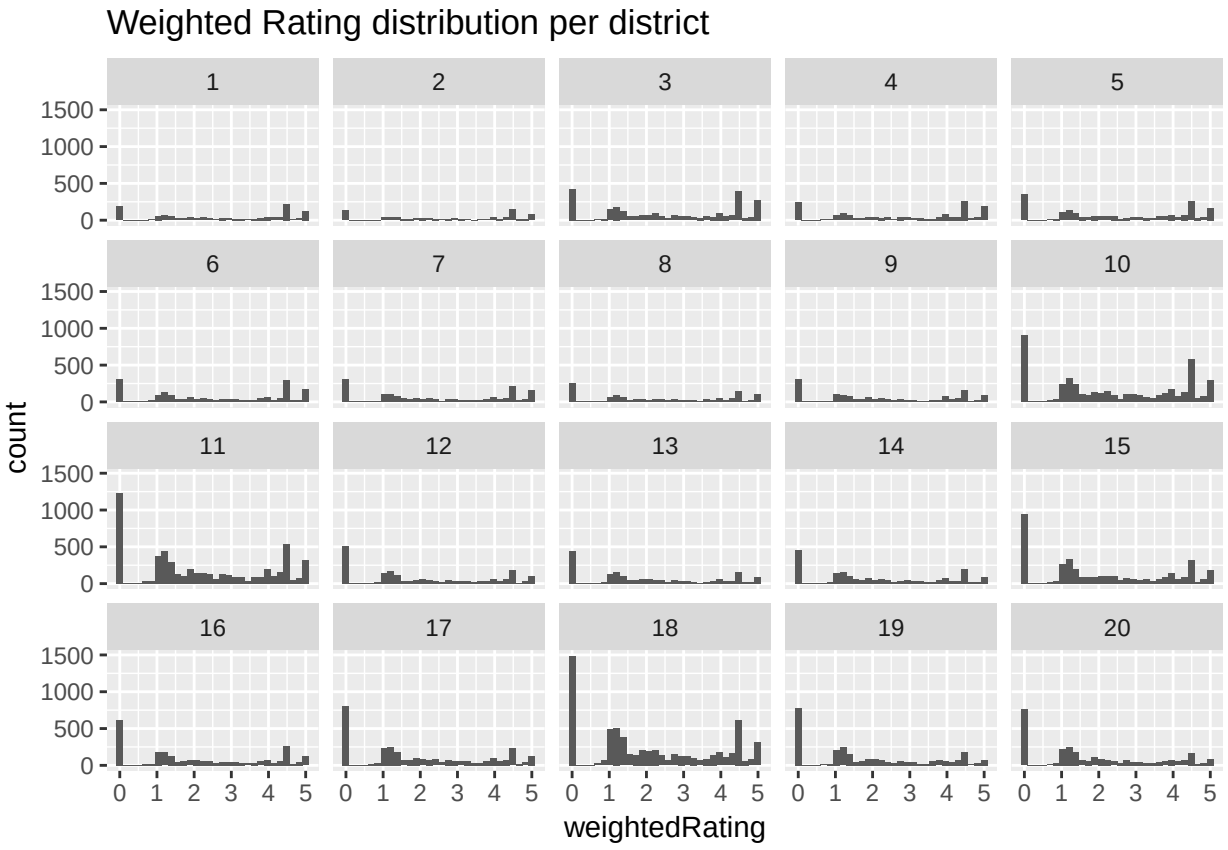
before we fit the multilevel model we need initial EDA to help us have a direct insight of the distribution of weighted ratings among and between districts and also the relationship between independent variables and dependents variables

1 Independent and Dependent

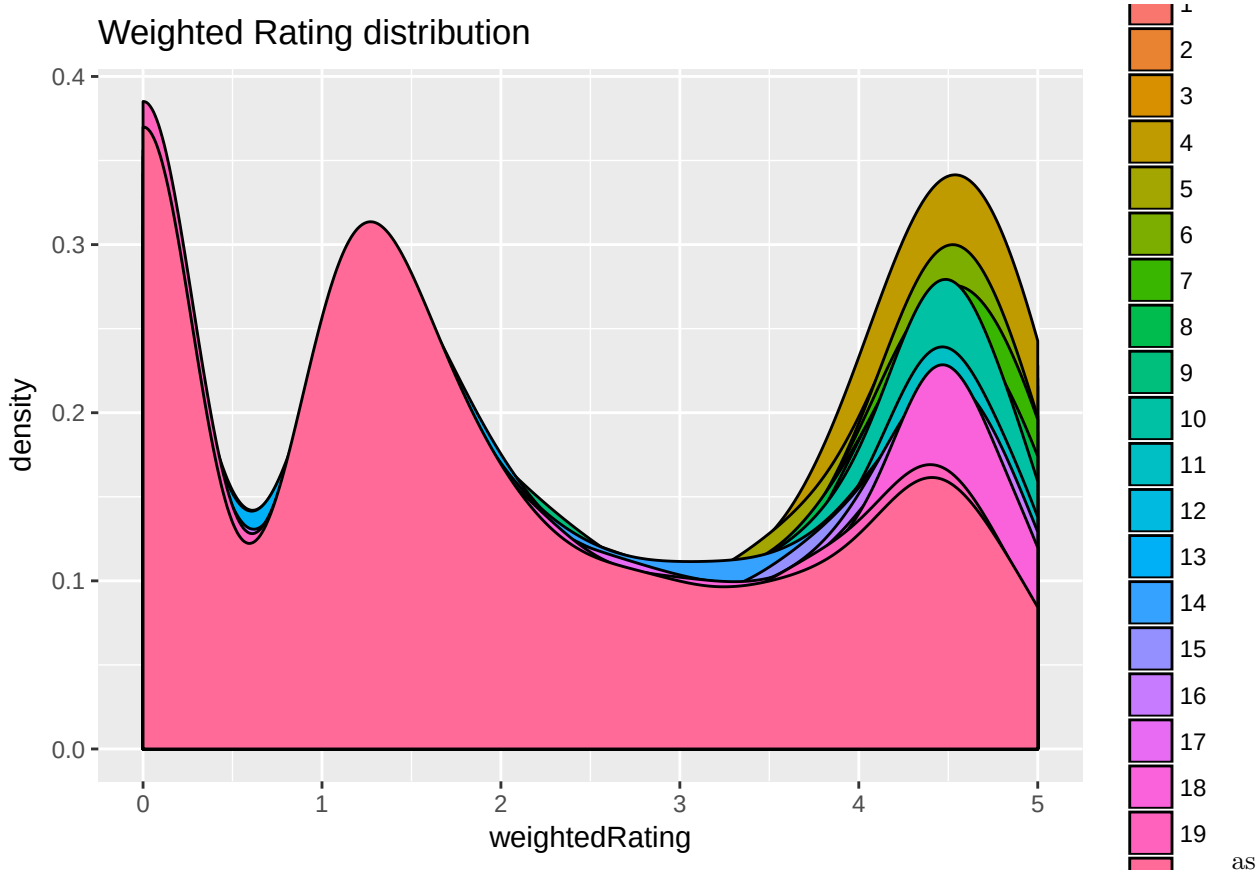
1) Distribution of ratings in different districts

```
ggplot(data=parisdata, aes(x=weightedRating))+
  geom_histogram()+facet_wrap(~district) + ggtitle("Weighted Rating distribution per district")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



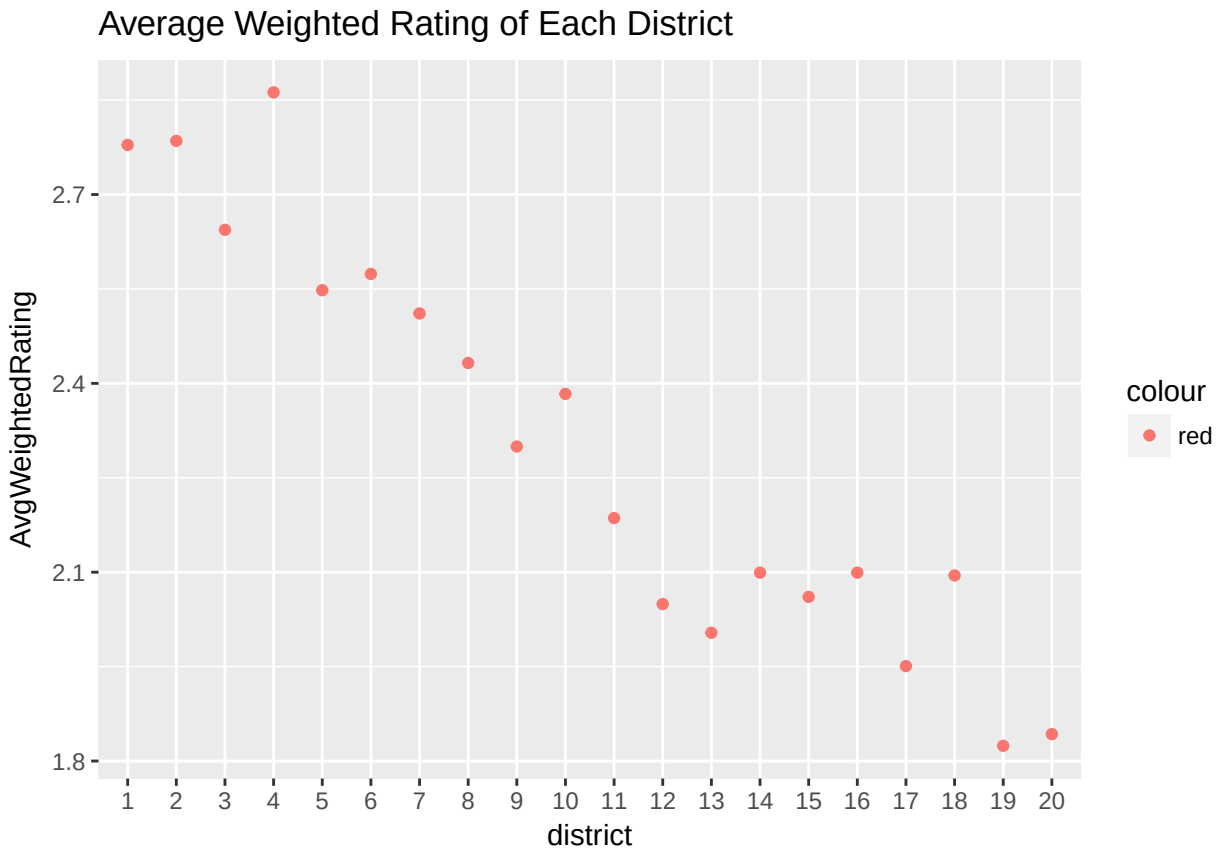

```
ggplot(data=parisdata, aes(x=weightedRating, fill=district))+geom_density()+
  ggtitle("Weighted Rating distribution")
```



as we can see the distribution of weighted ratings is different between districts

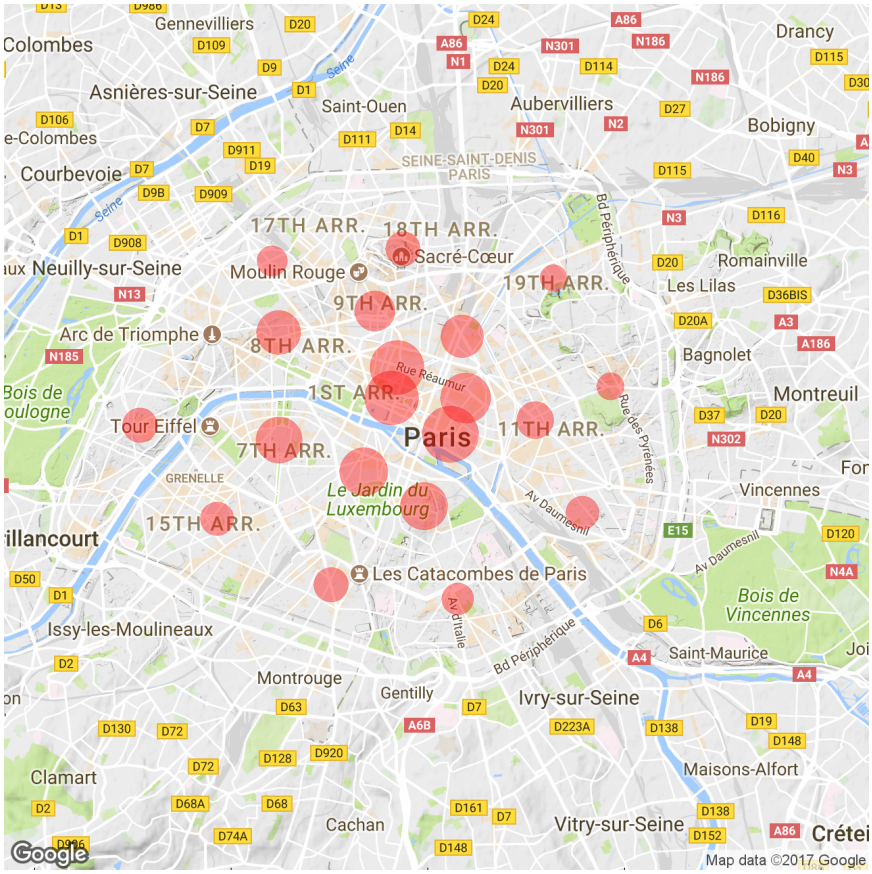
2) Average weighted rating per district

```
districtdat<-sqldf("SELECT district, Avg(longitude), Avg(latitude),
                Avg(weightedRating) FROM parisdata Group by 1")

colnames(districtdat)<-c("district", "lon", "lat", "AvgWeightedRating")

ggplot(data=districtdat, aes(x=district, y=AvgWeightedRating, col="red"))+geom_point()+
  ggtitle("Average Weighted Rating of Each District")
```



we also plot the average rating into map

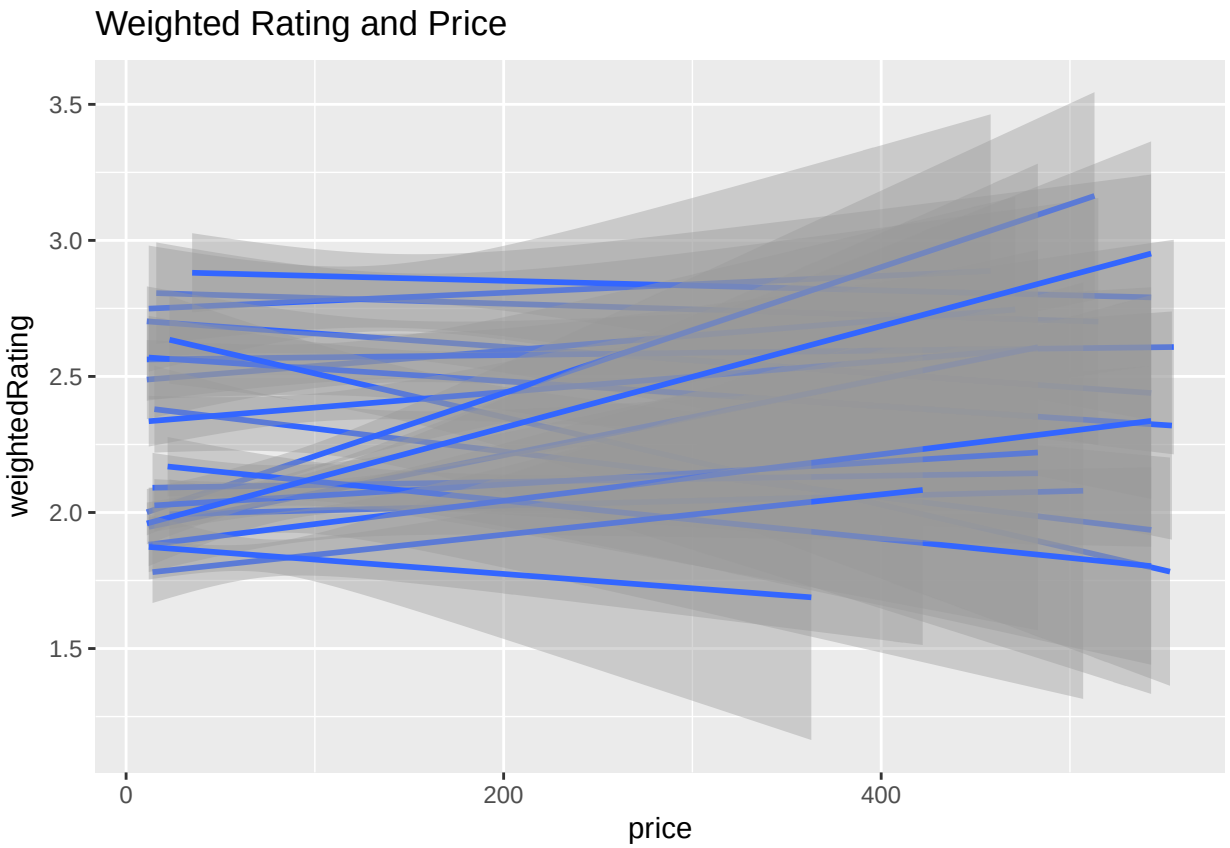
```
parismap+
  geom_point(aes(lon, lat), data=districtdat, col="red",
             alpha=0.4, size = ((districtdat$AvgWeightedRating-1)*5))
```



as we can see, the average weighted rating is different between districts

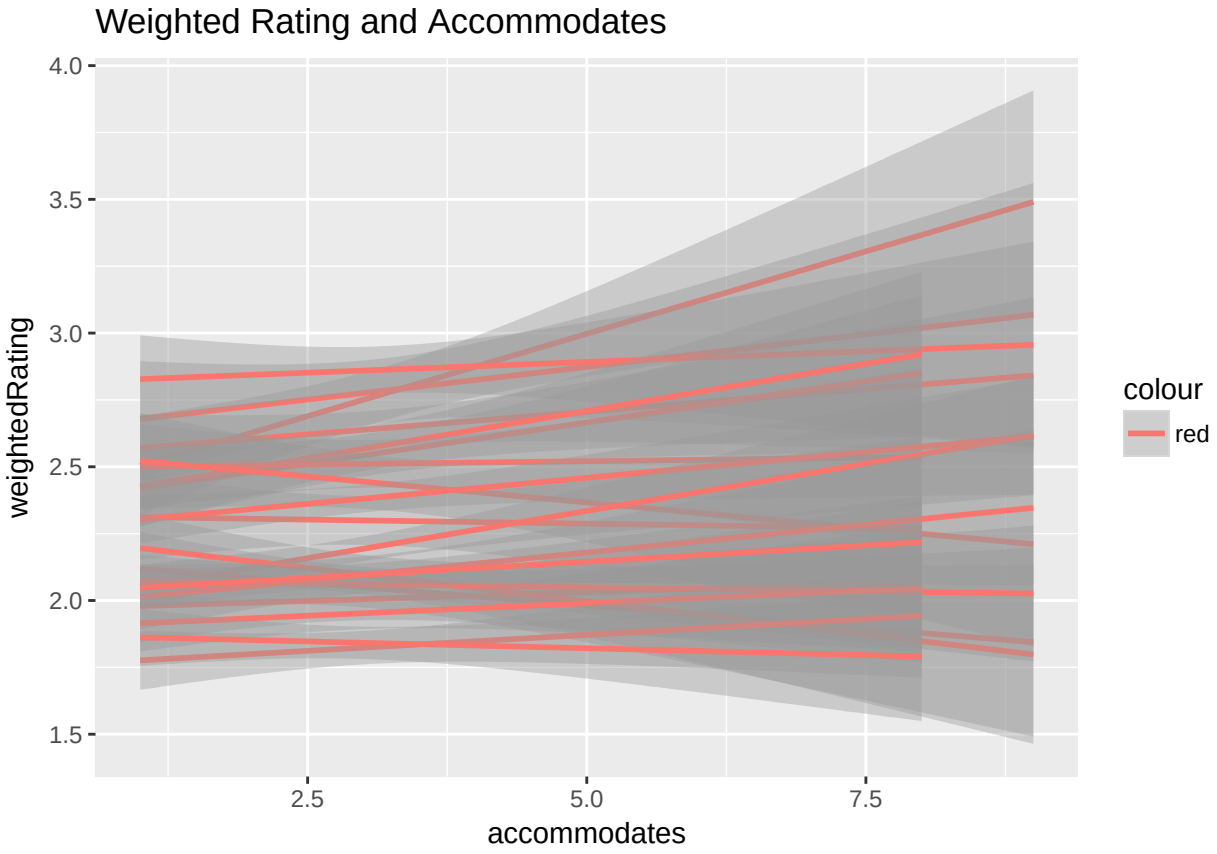
3) weighted rating and price obvious variablity of slope between district

```
ggplot(data=parisdata, aes(x=price, y=weightedRating, group=district))+
  geom_smooth(method = "lm")+ggtitle("Weighted Rating and Price")
```



4) weighted rating and accommodates varing slope betweeen district but less variability compared with that of weighted rating and price

```
ggplot(data=parisdata, aes(x=accommodates, y=weightedRating, group=district, col="red"))+
  geom_smooth(method = "lm")+ggtitle("Weighted Rating and Accommodates")
```

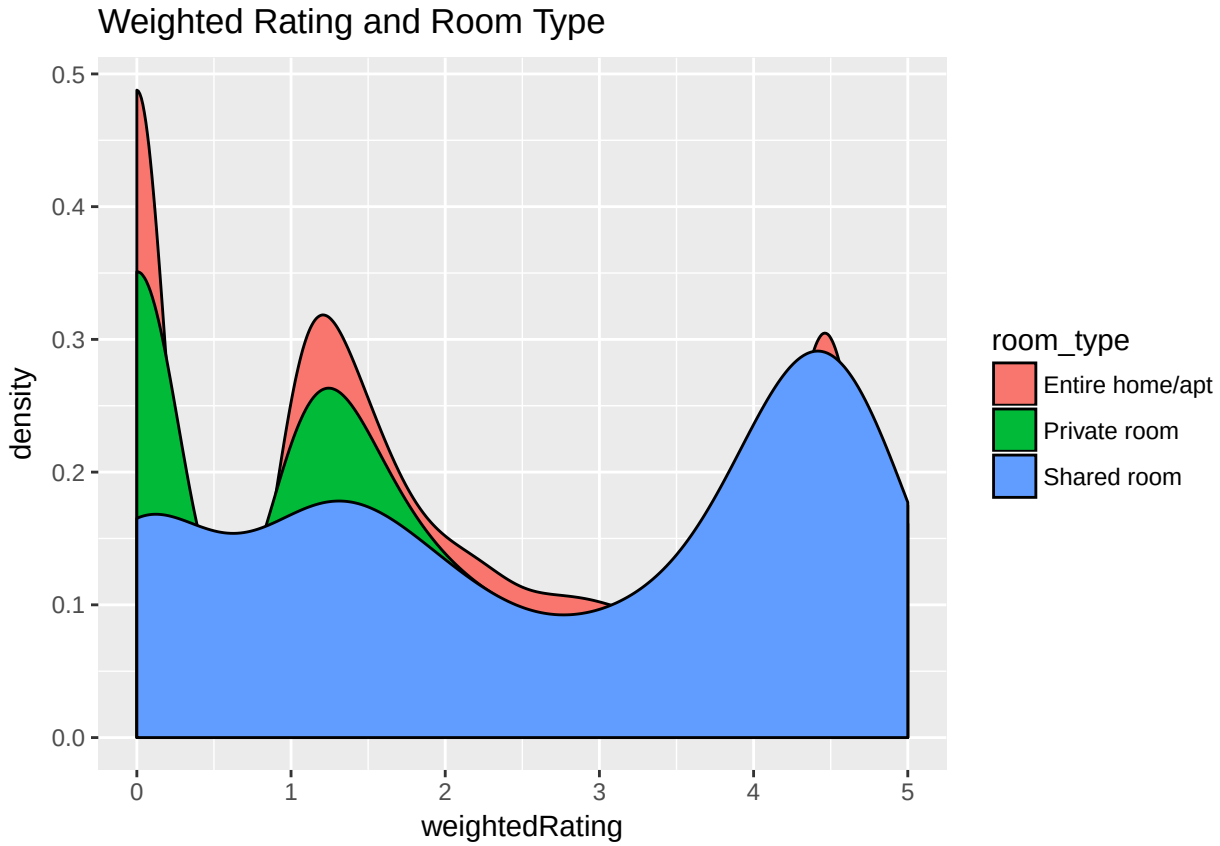


5) weighted rating and bedrooms varing slope betweeen district

```
ggplot(data=parisdata, aes(x=bedrooms, y=weightedRating, group=district))+
  geom_smooth(method = "lm")+ggtitle("Weighted Rating and Bedrooms")
```

6) weighted rating and room type varing slope betweeen district

```
ggplot(data=parisdata, aes(x=weightedRating, fill=room_type))+geom_density()+
  ggtitle("Weighted Rating and Room Type")
```



2. Between predictors ==> check Collinearity

1) correlation calculation

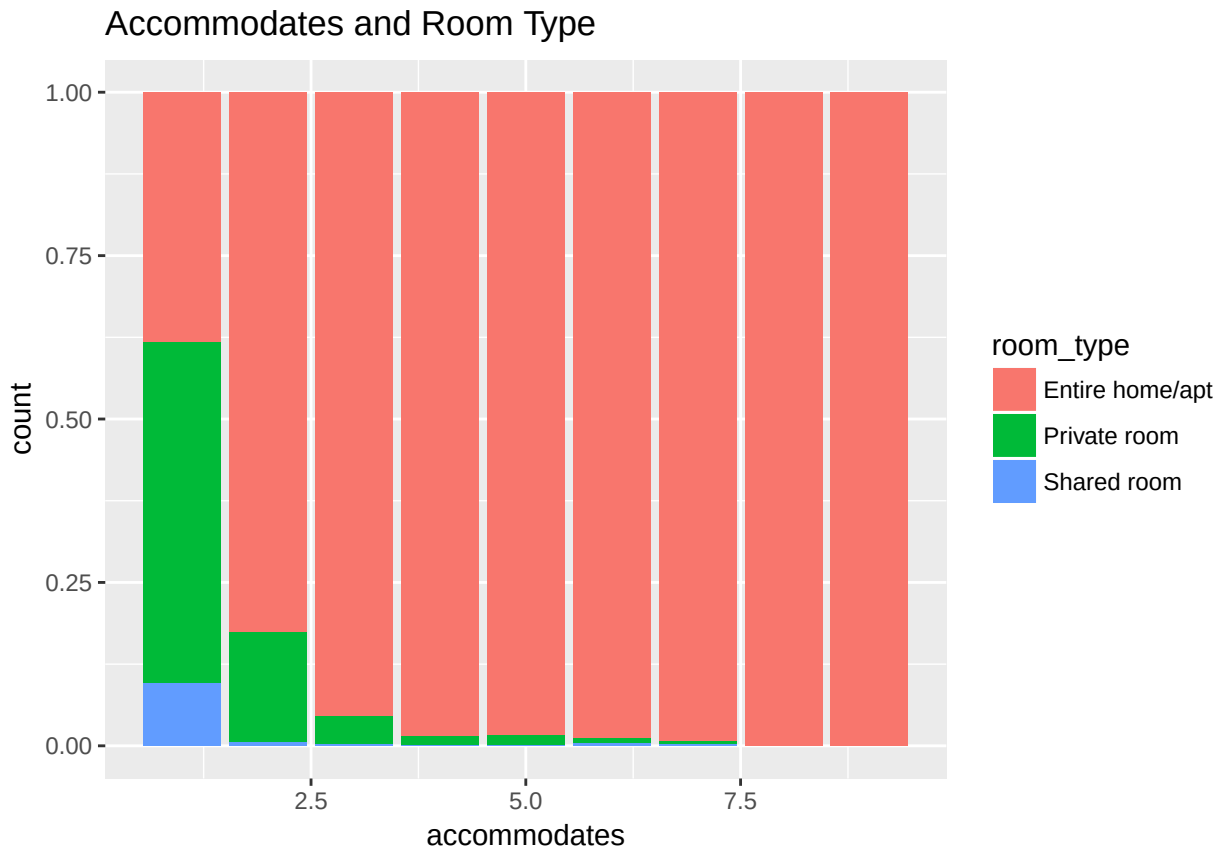
```
correlation<-cor(parisdata[,c(7,8,9,14)])
symnum(correlation)
```

```
##              a b p v
## accommodates 1
## bedrooms     , 1
## price        . . 1
## violenceRate       1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

we can see there is a moderate correlation between accommodates and bedrooms

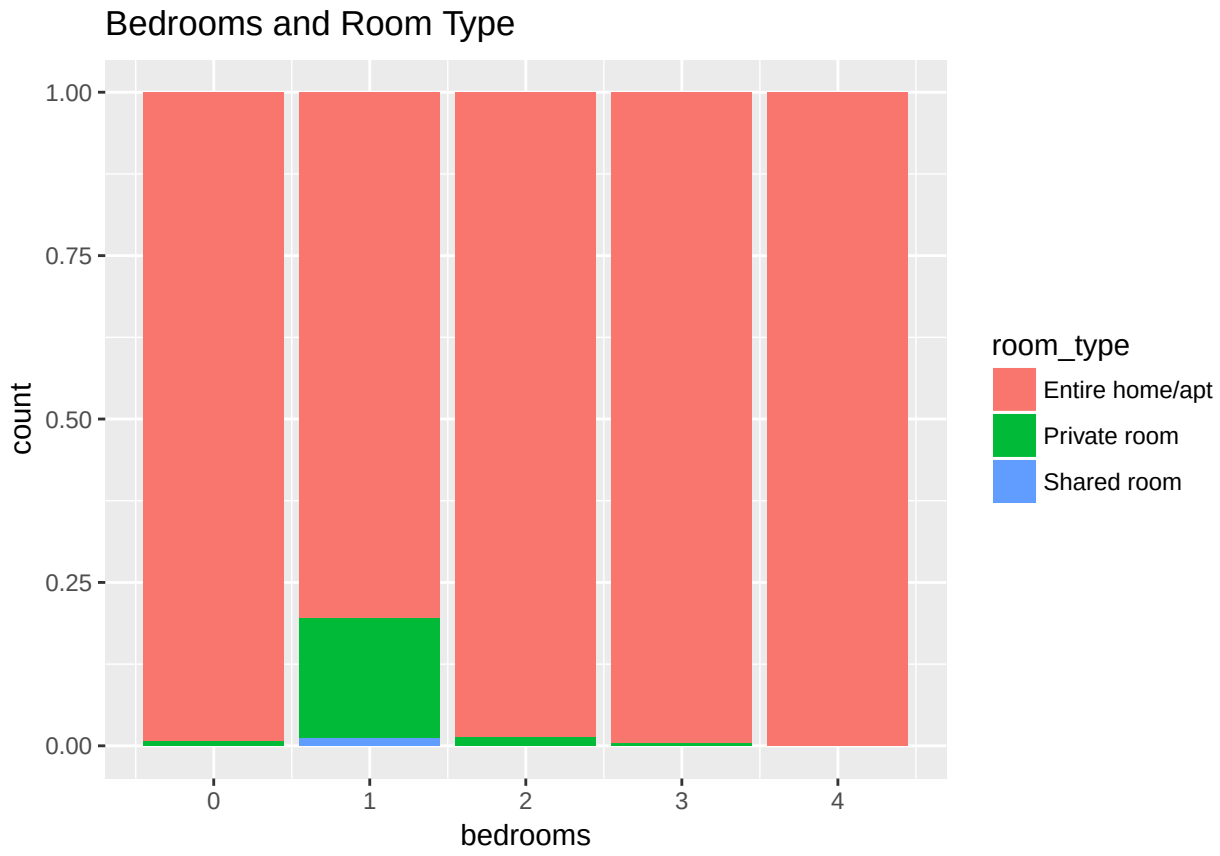
2) room type and accommodates: entire home tends to allow more accommodates

```
ggplot(data=parisdata, aes(x=accommodates, fill=room_type))+geom_bar(position = "fill")+
  ggtitle("Accommodates and Room Type")
```



3) room type and bedrooms: most of airbnb listes are entire home/apt or private room

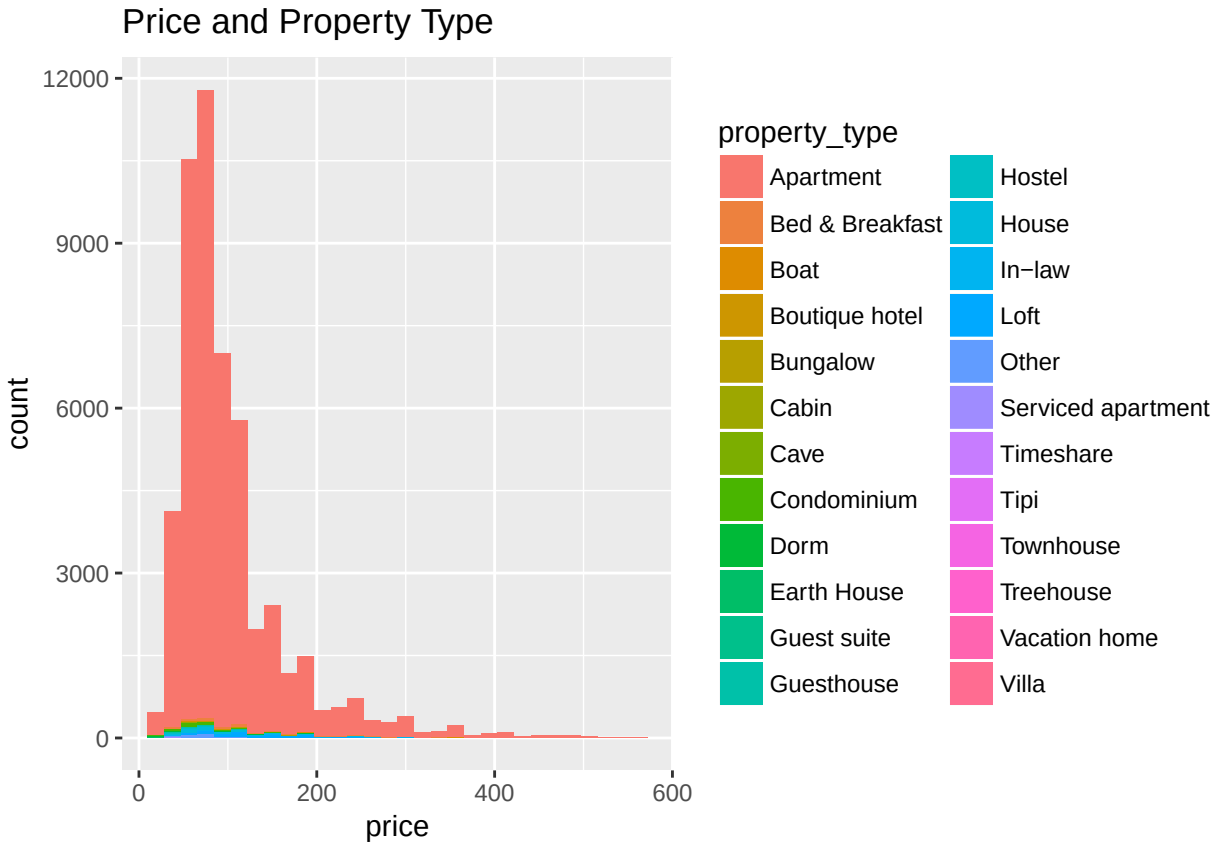
```
ggplot(data=parisdata, aes(x=bedrooms, fill=room_type))+geom_bar(position = "fill")+
  ggtitle("Bedrooms and Room Type")
```



4) price and property type: most of airbnb listed are apartment

```
ggplot(data=parisdata, aes(x=price, fill=property_type))+geom_histogram()+
  ggtitle("Price and Property Type")
```

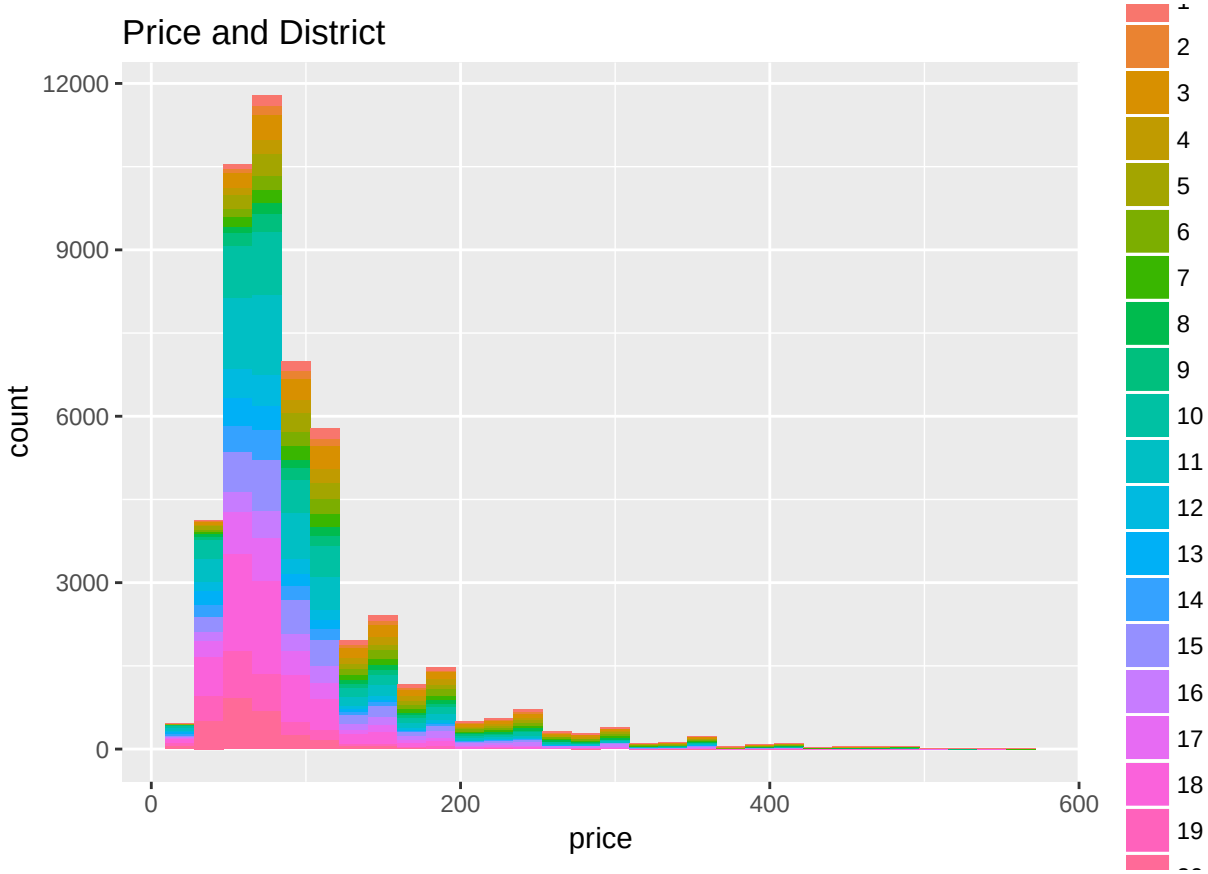
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



5) price and district: variability between districts

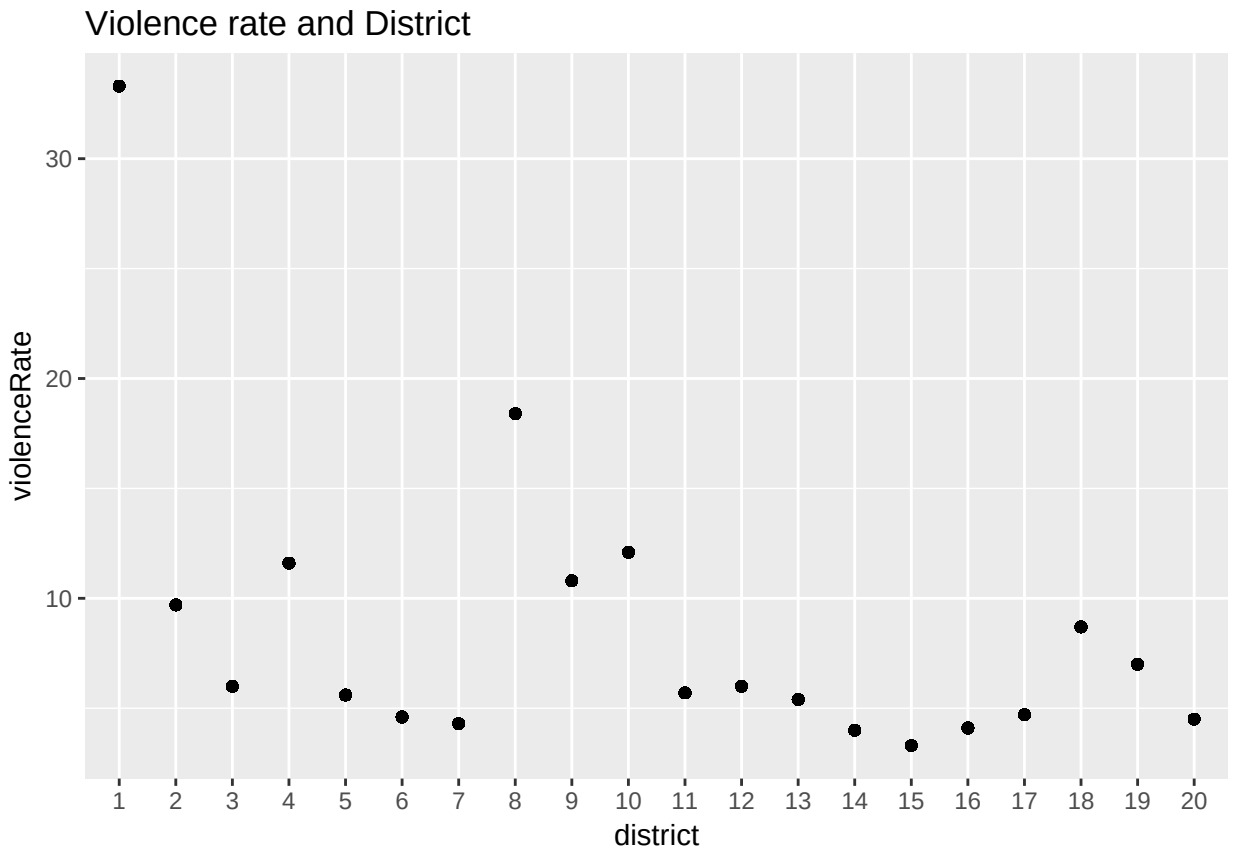
```
ggplot(data=parisdata, aes(x=price, fill=district))+geom_histogram()+
  ggtitle("Price and District")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



6) violence rate and district: variability between districts

```
ggplot(data=parisdata, aes(x=district, y=violenceRate))+geom_point()+
  ggtitle("Violence rate and District")
```



Model Fit

```
parisdata$cprice<-scale(parisdata$price, center = TRUE, scale = TRUE)
```

0) Model0: no random effect

```
model0<-lm(weightedRating~room_type+accommodates+
              bedrooms+cprice+violenceRate, data=parisdata)

display(model0)
```

```
## lm(formula = weightedRating ~ room_type + accommodates + bedrooms +
##     cprice + violenceRate, data = parisdata)
##                       coef.est coef.se
## (Intercept)            2.06     0.03
## room_typePrivate room  0.14     0.03
## room_typeShared room   0.60     0.09
## accommodates           0.06     0.01
## bedrooms              -0.20     0.01
## cprice                 0.13     0.01
## violenceRate           0.02     0.00
## ---
## n = 50406, k = 7
## residual sd = 1.73, R-Squared = 0.01
```

as we can see from the output of the model, all the coefficients are statistically significantly with each unit increase of accommodates, weighted rating increases by 0.06; with every unit increase of bedrooms, weighted rating decreases by 0.02; with every unit increase of violence rate, weighted rating increases by 0.02, as the districts in the downtown have higher rate of violence which indicate airbnb users prefer properties near

downtown with each unit increase of price deviating from mean price over standard deviation weighted rating increases by 0.13 with other variable constant, shared room has 0.6 higher weighted rating than that of entire room, while private room has 0.14 higher weighted rating than that of entire room

1) Model1: random intercept group predictor of violence rate and individual predictor of room type, accommodate, bedrooms, price

```
model1<-lmer(weightedRating~room_type+accommodates+
               bedrooms+cprice+violenceRate+(1|district), data=parisdata)

display(model1)
```

```
## lmer(formula = weightedRating ~ room_type + accommodates + bedrooms +
##     cprice + violenceRate + (1 | district), data = parisdata)
##                        coef.est coef.se
## (Intercept)             2.01     0.10
## room_typePrivate room   0.16     0.03
## room_typeShared room    0.59     0.09
## accommodates            0.08     0.01
## bedrooms               -0.15     0.01
## cprice                  0.03     0.01
## violenceRate            0.02     0.01
##
## Error terms:
##  Groups   Name        Std.Dev.
##  district (Intercept) 0.28
##  Residual             1.71
## ---
## number of obs: 50406, groups: district, 20
## AIC = 197377, DIC = 197274.7
## deviance = 197317.0
```

as we can see from the output of the model, all the coefficients are statistically significantly except the coefficient of violence rate with each unit increase of accommodates, weighted rating increases by 0.08 on average; with every unit increase of bedrooms, weighted rating decreases by 0.15 on average; with each unit increase of price deviating from mean price over standard deviation weighted rating increases by 0.03 on average with other variable constant, shared room has 0.59 higher weighted rating than that of entire room on average, while private room has 0.16 higher weighted rating than that of entire room

the unexplained within-county variation has an estimated standard deviation of 1.71 the estimated standard deviation of the district intercept is 0.28

as we can see from the correlation table: there is high correlation between bedrooms and accommodates, violence rate and intercept

2) Model2: random slope

```
model2<-lmer(weightedRating~accommodates+room_type+
               bedrooms+violenceRate+cprice+(0+cprice|district), data=parisdata)

display(model2)
```

```
## lmer(formula = weightedRating ~ accommodates + room_type + bedrooms +
##     violenceRate + cprice + (0 + cprice | district), data = parisdata)
##                        coef.est coef.se
## (Intercept)             2.08     0.03
## accommodates            0.06     0.01
## room_typePrivate room   0.17     0.03
```

```
## room_typeShared room   0.62     0.09
## bedrooms              -0.21     0.01
## violenceRate           0.02     0.00
## cprice                 0.16     0.03
##
## Error terms:
##  Groups   Name   Std.Dev.
##  district cprice 0.12
##  Residual        1.72
## ---
## number of obs: 50406, groups: district, 20
## AIC = 198005, DIC = 197892.4
## deviance = 197939.8
```

as we can see from the output of the model, all the coefficients are statistically significantly with each unit increase of accommodates, weighted rating increases by 0.06 on average; with every unit increase of bedrooms, weighted rating decreases by 0.21 on average; with every unit increase of violence rate, weighted rating increases by 0.02 on average, as the districts in the downtown have higher rate of violence which indicate airbnb users prefer properties near downtown with each unit increase of price deviating from mean price over standard deviation weighted rating increases by 0.16 on average with other variable constant, shared room has 0.62 higher weighted rating than that of entire room on average, while private room has 0.17 higher weighted rating than that of entire room

the unexplained within-county variation has an estimated standard deviation of 1.71 the estimated standard deviation of the district price slope is 0.12

as we can see from the correlation table: there is high correlation between bedrooms and accommodates, accommodate and intercept

3) Model3: random slope and intercept

```
model3<-lmer(weightedRating~accommodates+room_type+
               bedrooms+violenceRate+cprice+(1+cprice|district), data=parisdata)

display(model3)
```

```
## lmer(formula = weightedRating ~ accommodates + room_type + bedrooms +
##     violenceRate + cprice + (1 + cprice | district), data = parisdata)
##                       coef.est coef.se
## (Intercept)            2.07     0.10
## accommodates           0.07     0.01
## room_typePrivate room  0.17     0.03
## room_typeShared room   0.61     0.09
## bedrooms              -0.15     0.01
## violenceRate           0.02     0.01
## cprice                 0.04     0.02
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  district (Intercept) 0.28
##           cprice      0.06     -0.41
##  Residual             1.71
## ---
## number of obs: 50406, groups: district, 20
## AIC = 197346, DIC = 197241.9
## deviance = 197283.1
```

as we can see, the coefficients of violence rate is not statistically significant the intercept doesn't change much compared with the previous models the coefficent of accommodates is between that of model2 and model1 the bedroom coefficient is equal to that of model1, indicating each unit increase of bedroom will decrease the weighted rating by 0.15 with each unit increase of price deviating from mean price over standard deviation weighted rating increases by 0.04 on average

the residual within district is 1.71, which is smaller than those of model1 and model2 indicating model3 is a better fit the sd of district price slope is 0.06, and its correlation with intercept is -0.41 while the sd of district intercept is 0.28

```
model3.1<-lmer(weightedRating~cprice+room_type+
               bedrooms+violenceRate+accommodates+(1+accommodates|district), data=parisdata)

display(model3.1)
```

```
## lmer(formula = weightedRating ~ cprice + room_type + bedrooms +
##     violenceRate + accommodates + (1 + accommodates | district),
##     data = parisdata)
##                         coef.est coef.se
## (Intercept)              2.03     0.10
## cprice                   0.03     0.01
## room_typePrivate room    0.15     0.03
## room_typeShared room     0.59     0.09
## bedrooms                -0.15     0.01
## violenceRate             0.02     0.01
## accommodates             0.07     0.01
##
## Error terms:
##  Groups   Name         Std.Dev. Corr
##  district (Intercept)  0.26
##           accommodates 0.02     0.09
##  Residual              1.71
## ---
## number of obs: 50406, groups: district, 20
## AIC = 197376, DIC = 197270.6
## deviance = 197312.3
```

as we can see, the coefficients of violence rate is not statistically significant the coefficient of bedrooms, accommodates, and violence rate in model3.1 are same as those in model3 while the coefficeint of private room, share room and price decrease slightly compared with model3

the residual is same as that of model3 while the standard error of district intercept and accommodates slope decreased the correlation between accommodates and intercept is 0.09

```
model3.2<-lmer(weightedRating~cprice+room_type+
                 accommodates+violenceRate+bedrooms+(1+bedrooms|district), data=parisdata)

display(model3.2)
```

```
## lmer(formula = weightedRating ~ cprice + room_type + accommodates +
##     violenceRate + bedrooms + (1 + bedrooms | district), data = parisdata)
##                         coef.est coef.se
## (Intercept)              2.02     0.10
## cprice                   0.04     0.01
## room_typePrivate room    0.16     0.03
## room_typeShared room     0.59     0.09
## accommodates             0.08     0.01
```

```
## violenceRate           0.02     0.01
## bedrooms              -0.15     0.02
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  district (Intercept) 0.28
##           bedrooms    0.03     -0.25
##  Residual             1.71
## ---
## number of obs: 50406, groups: district, 20
## AIC = 197379, DIC = 197273.4
## deviance = 197315.4
```

the coefficients of bedroom, violence, and price remain unchanged compared with model3 while the intercept decreases by 0.05 to 2.02 the coefficients of private and shared room decrease slightly and the accommodate coefficient increases by 0.01

the residual is 1.71, same as that of model3 and model3.1 the sd of district intercept is same as that of model3 while the district bedroom slope is between that of model3 and model3.1 correlation between bedroom slope and district intercept is -0.25

IN CONCLUSION: random slope and random intercept model is a better fit compared with only random slope or only random intercept

4) Model4: random slope and intercept with interaction

as we noticed in the previous model and also shown in our correlation table there is a moderate collinearity between accommodates and bedrooms we can add interaction of those two into our model

```
model4<-lmer(weightedRating~accommodates*bedrooms+
                room_type+violenceRate+cprice+(1+cprice|district), data=parisdata)

display(model4)
```

```
## lmer(formula = weightedRating ~ accommodates * bedrooms + room_type +
##     violenceRate + cprice + (1 + cprice | district), data = parisdata)
##                       coef.est coef.se
## (Intercept)            2.05     0.10
## accommodates           0.08     0.01
## bedrooms              -0.14     0.03
## room_typePrivate room  0.17     0.03
## room_typeShared room   0.62     0.09
## violenceRate           0.02     0.01
## cprice                 0.05     0.02
## accommodates:bedrooms -0.01     0.01
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  district (Intercept) 0.28
##           cprice      0.06     -0.41
##  Residual             1.71
## ---
## number of obs: 50406, groups: district, 20
## AIC = 197356, DIC = 197232.9
## deviance = 197282.4
```

the coefficients of price and share room increase slightly compared with model3 the coefficients of private room and violence remain same while the coefficients of accommodate and bedroom increase, which is offset

by the coefficient of those two interactions

the sd of district intercept, slope and their correlation are same as model3 the residual is same as model3, while the deviance decreases

```
model4.1<-lmer(weightedRating~accommodates*room_type+bedrooms+
               violenceRate+cprice+(1+cprice|district), data=parisdata)

display(model4.1)
```

```
## lmer(formula = weightedRating ~ accommodates * room_type + bedrooms +
##     violenceRate + cprice + (1 + cprice | district), data = parisdata)
##                                    coef.est coef.se
## (Intercept)                         2.05     0.10
## accommodates                        0.08     0.01
## room_typePrivate room               0.41     0.07
## room_typeShared room                1.04     0.17
## bedrooms                           -0.16     0.01
## violenceRate                        0.02     0.01
## cprice                              0.04     0.02
## accommodates:room_typePrivate room -0.12     0.03
## accommodates:room_typeShared room  -0.24     0.08
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  district (Intercept) 0.28
##           cprice      0.06     -0.41
##  Residual             1.71
## ---
## number of obs: 50406, groups: district, 20
## AIC = 197338, DIC = 197213.1
## deviance = 197262.5
```

the coefficients of price and violence remain same the coeffienct of accommodate, private and share increase obviously, which are offset by the interaction the bedroom coefficient decreases

the sd of district intercept, slope and their correlation are same as model3 and model4 the residual is same as model4, while the deviance decreases

Model Diagnose

1) Anova Analysis

```
anova(model1, model2, model3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: parisdata
## Models:
## model1: weightedRating ~ room_type + accommodates + bedrooms + cprice +
## model1:     violenceRate + (1 | district)
## model2: weightedRating ~ accommodates + room_type + bedrooms + violenceRate +
## model2:     cprice + (0 + cprice | district)
## model3: weightedRating ~ accommodates + room_type + bedrooms + violenceRate +
## model3:     cprice + (1 + cprice | district)
##        Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## model1  9 197335 197414 -98659   197317
## model2  9 197958 198037 -98970   197940   0.00      0          1
```

```
## model3 11 197305 197402 -98642   197283 656.68      2     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we can see model3 is a better fit compared with model1 and model2

```
anova(model3, model3.1, model3.2)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: parisdata
## Models:
## model3: weightedRating ~ accommodates + room_type + bedrooms + violenceRate +
## model3:     cprice + (1 + cprice | district)
## model3.1: weightedRating ~ cprice + room_type + bedrooms + violenceRate +
## model3.1:     accommodates + (1 + accommodates | district)
## model3.2: weightedRating ~ cprice + room_type + accommodates + violenceRate +
## model3.2:     bedrooms + (1 + bedrooms | district)
##          Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model3   11 197305 197402 -98642   197283
## model3.1 11 197334 197431 -98656   197312     0      0          1
## model3.2 11 197337 197434 -98658   197315     0      0          1
```

we can see pvalue is 1 bigger than 0.05, indicating no difference of model3, model3.1, model3.2

```
anova(model3, model4, model4.1)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: parisdata
## Models:
## model3: weightedRating ~ accommodates + room_type + bedrooms + violenceRate +
## model3:     cprice + (1 + cprice | district)
## model4: weightedRating ~ accommodates * bedrooms + room_type + violenceRate +
## model4:     cprice + (1 + cprice | district)
## model4.1: weightedRating ~ accommodates * room_type + bedrooms + violenceRate +
## model4.1:     cprice + (1 + cprice | district)
##          Df    AIC    BIC logLik deviance   Chisq Chi Df Pr(>Chisq)
## model3   11 197305 197402 -98642   197283
## model4   12 197306 197412 -98641   197282  0.6505      1     0.4199
## model4.1 13 197288 197403 -98631   197262 19.9441      1  7.974e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

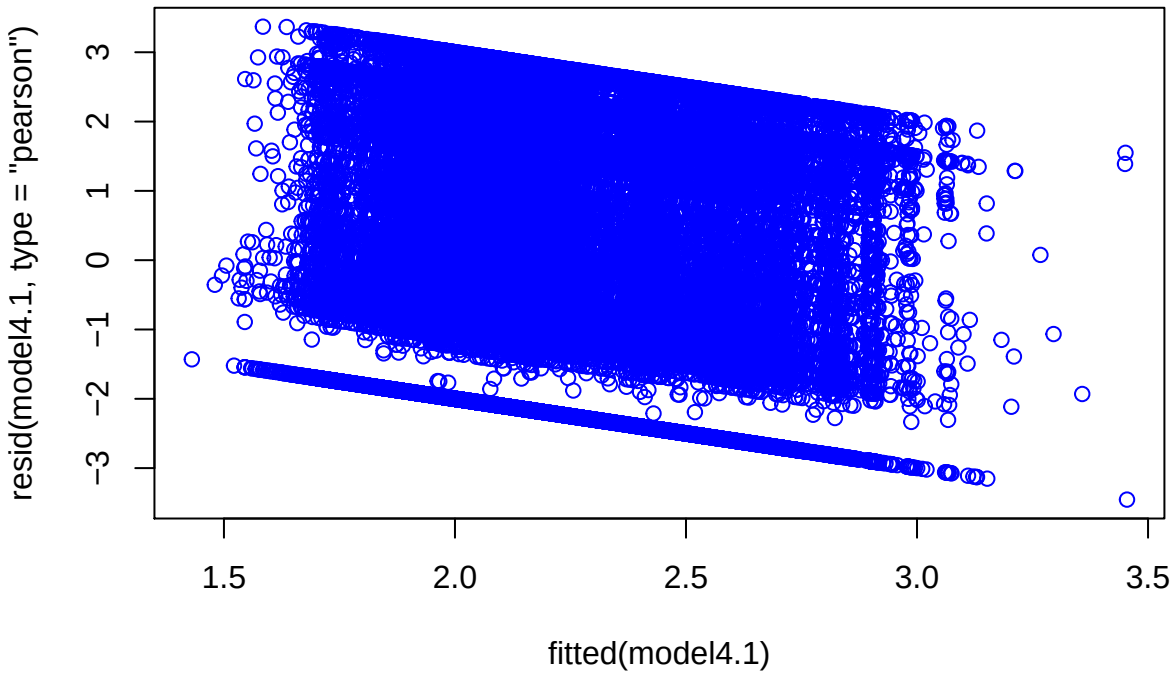
we can see pvalue of model 4.1 is smaller than 0.05, indicating it is a better fit compared with model4 and model3

In Conclusion, model4.1 is the best fit among our models

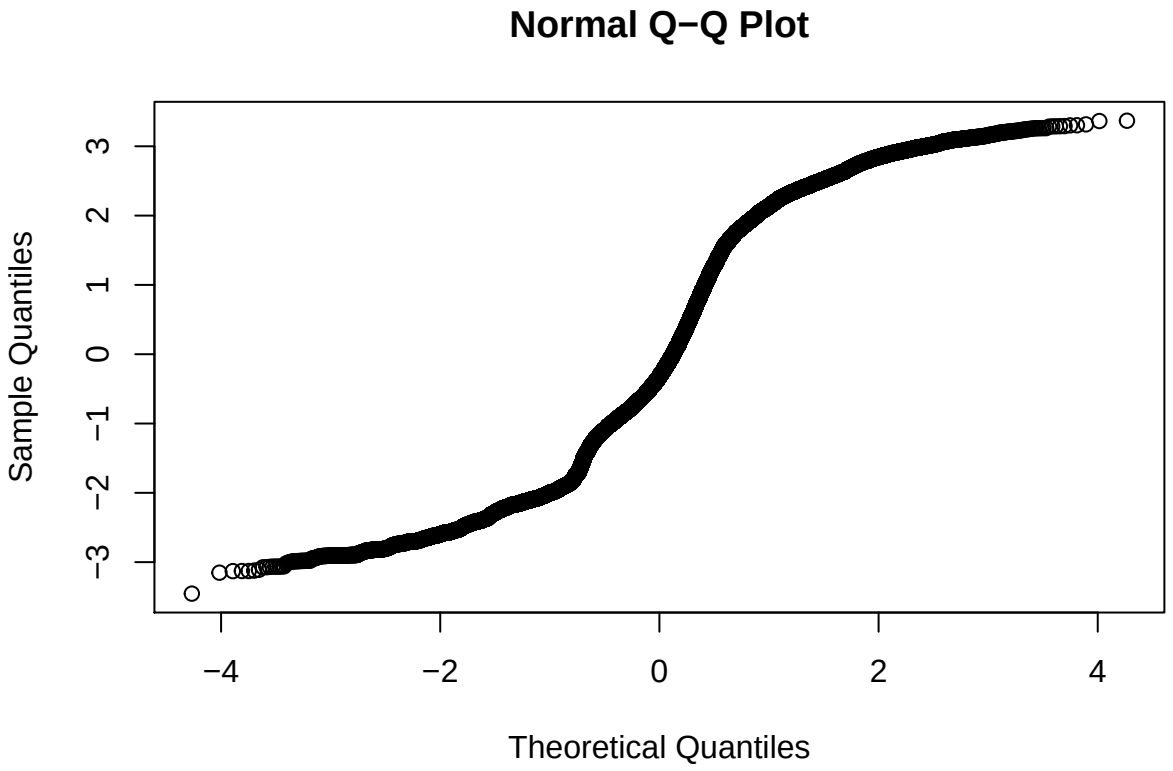
2) Residual

check the constant standard deviation of model4.1

```
plot(fitted(model4.1),resid(model4.1,type="pearson"),col="blue")
```




```
qqnorm(resid(model4.1))
```



as we can see from the Q-Q plot, it doesn't show a normal distribution pattern but a bimodal pattern

Therefore linear multilevel model may not be appropriate we can differentiate the airbnb weighted rating into two category above 2: then satisfied below 2: then unsatisfied Then we can fit a logistic model

Model Refit

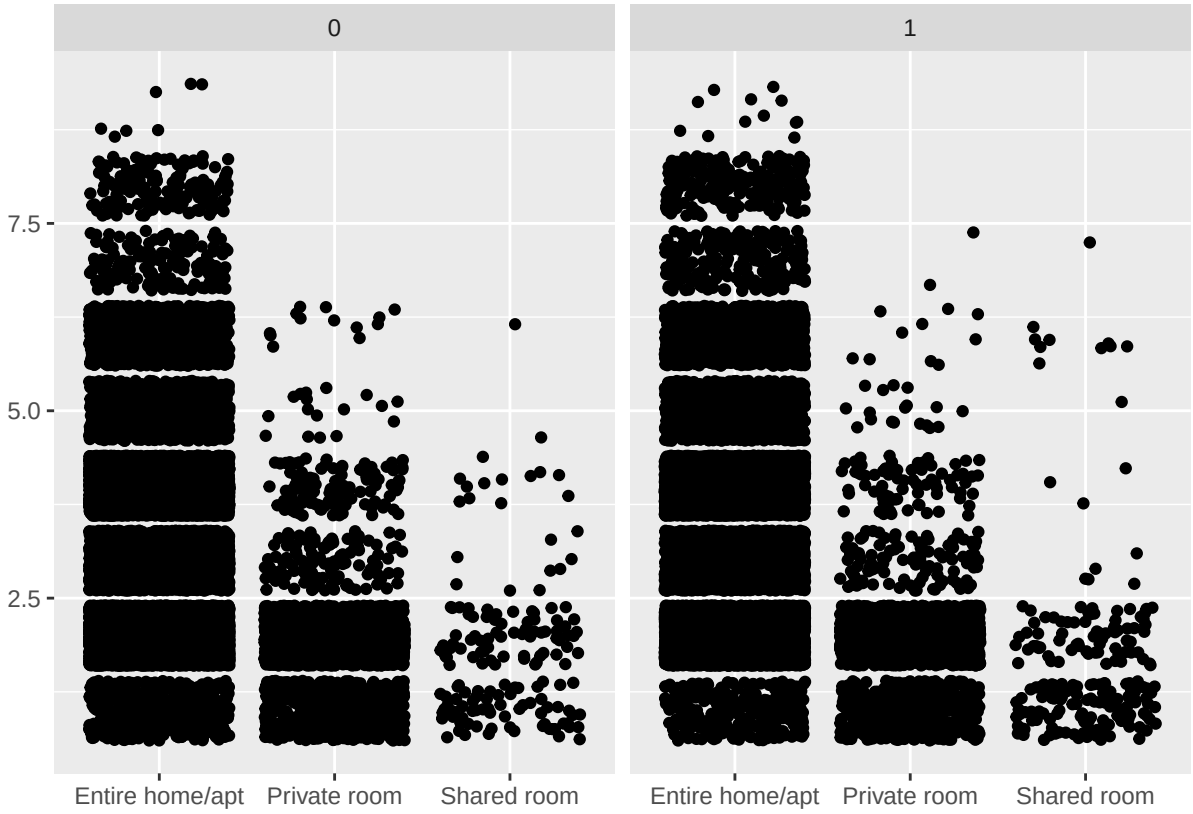
1) we will start to add one binary variable into our data as our new response

```
AddSatisfaction<-function(data){
  result=data
  for (row in 1:nrow(result)){
    for (n in names(result)[16]){
      if (result[row,n]>=2){
        result$Satisfaction[row] = 1
      }else{
        result$Satisfaction[row] = 0
      }
    }
  }
  return(result)
}

pardata<-AddSatisfaction(parisdata)
```

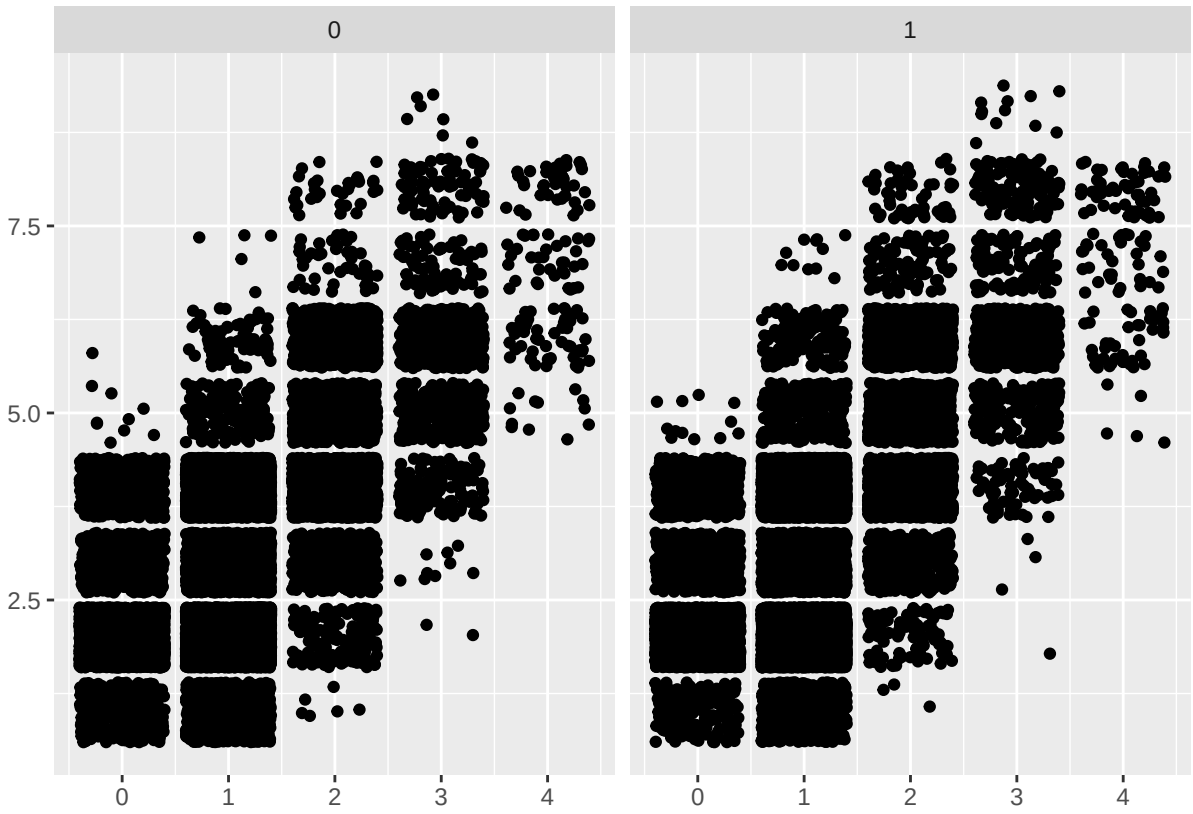
2) we can start from fitting simple logistic model

data visualization

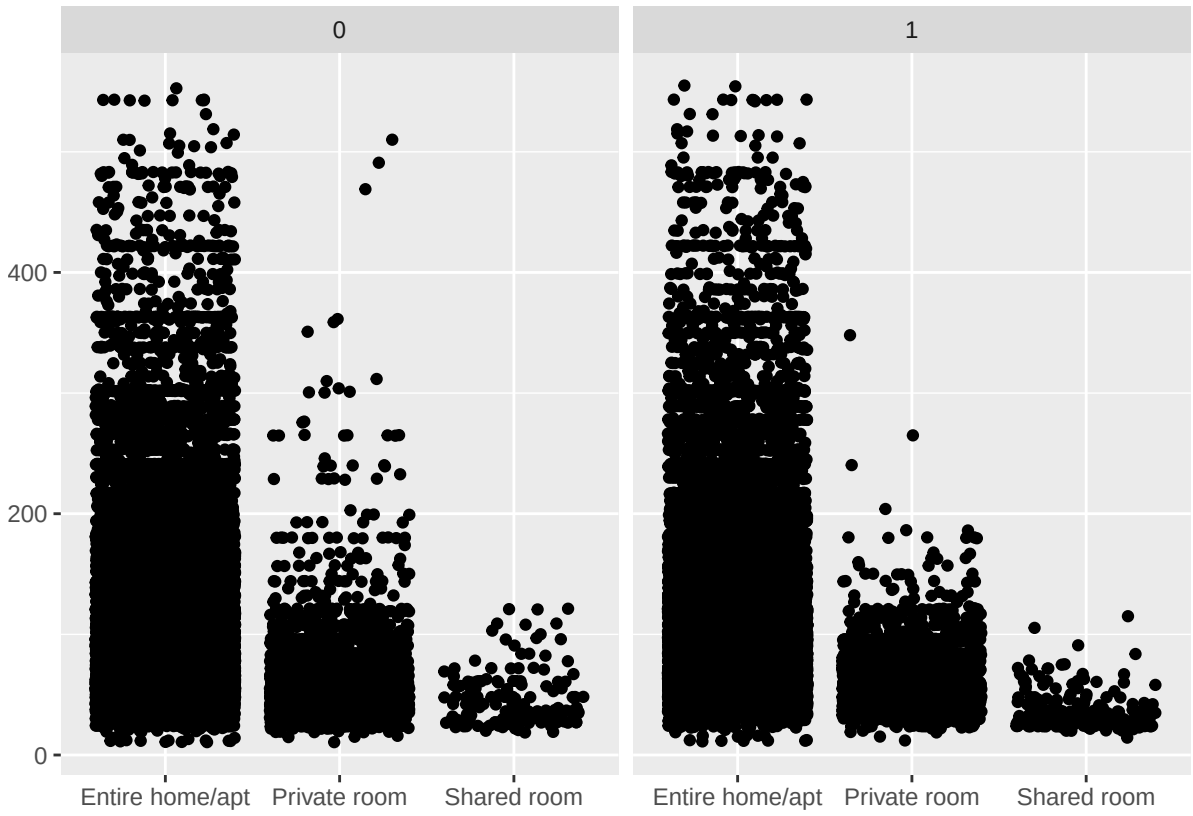
```
ggplot(pardata)+aes(x=room_type,y=accommodates)+
  geom_jitter()+facet_grid(.~Satisfaction)+
  scale_fill_manual(values=c("blue","red"))+ylab("")+xlab("")
```




```
ggplot(pardata)+aes(x=bedrooms,y=accommodates)+
  geom_jitter()+facet_grid(.~Satisfaction)+
  scale_fill_manual(values=c("blue","red"))+ylab("")+xlab("")
```




```
ggplot(pardata)+aes(x=room_type,y=price)+
  geom_jitter()+facet_grid(.~Satisfaction)+
  scale_fill_manual(values=c("blue","red"))+ylab("")+xlab("")
```



model fit1

```
fit1<-glm(Satisfaction~room_type+accommodates+
            bedrooms+cprice+violenceRate, family = binomial, data=pardata)
display(fit1)
```

```
## glm(formula = Satisfaction ~ room_type + accommodates + bedrooms +
##     cprice + violenceRate, family = binomial, data = pardata)
##                       coef.est coef.se
## (Intercept)           -0.25     0.03
## room_typePrivate room  0.10     0.03
## room_typeShared room   0.59     0.11
## accommodates           0.07     0.01
## bedrooms              -0.20     0.02
## cprice                 0.11     0.01
## violenceRate           0.02     0.00
## ---
##   n = 50406, k = 7
##   residual deviance = 69406.1, null deviance = 69839.2 (difference = 433.2)
```

model fit1.1 as we have seen in the EDA, there is correlation between room type and accommodate we can add interaction to our model

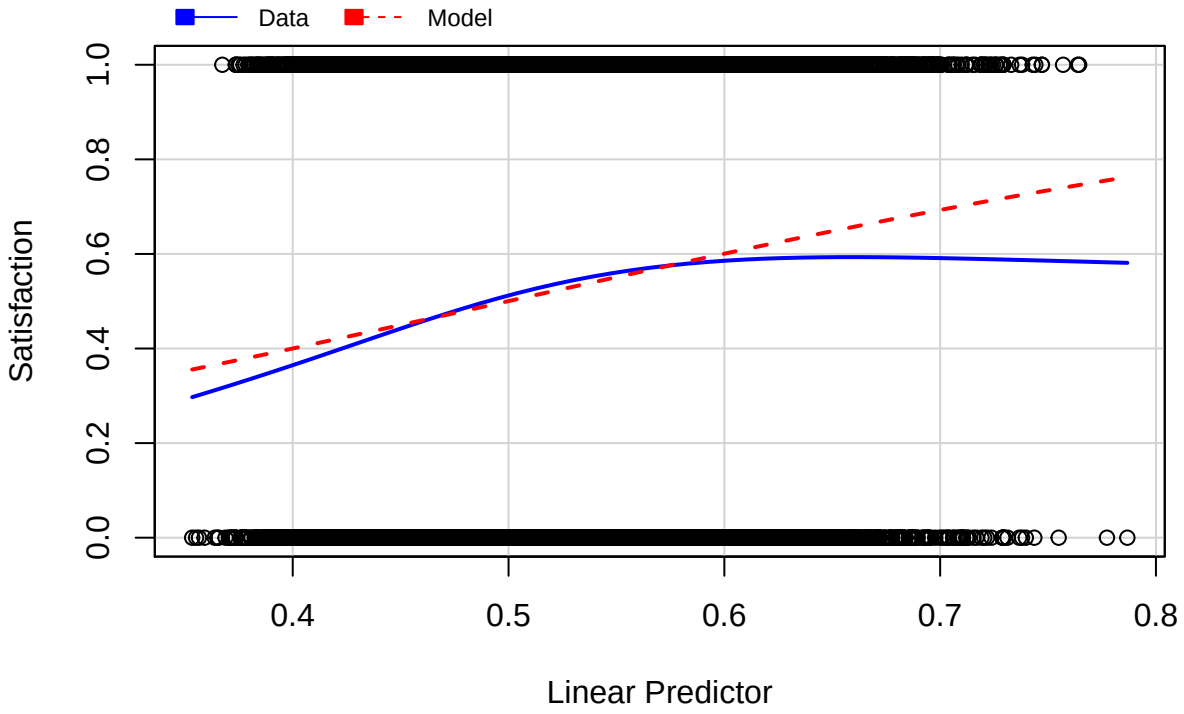
```
fit1.1<-glm(Satisfaction~room_type*accommodates+
            bedrooms+cprice+violenceRate, family = binomial, data=pardata)
display(fit1.1)
```

```
## glm(formula = Satisfaction ~ room_type * accommodates + bedrooms +
##     cprice + violenceRate, family = binomial, data = pardata)
##                                     coef.est coef.se
```

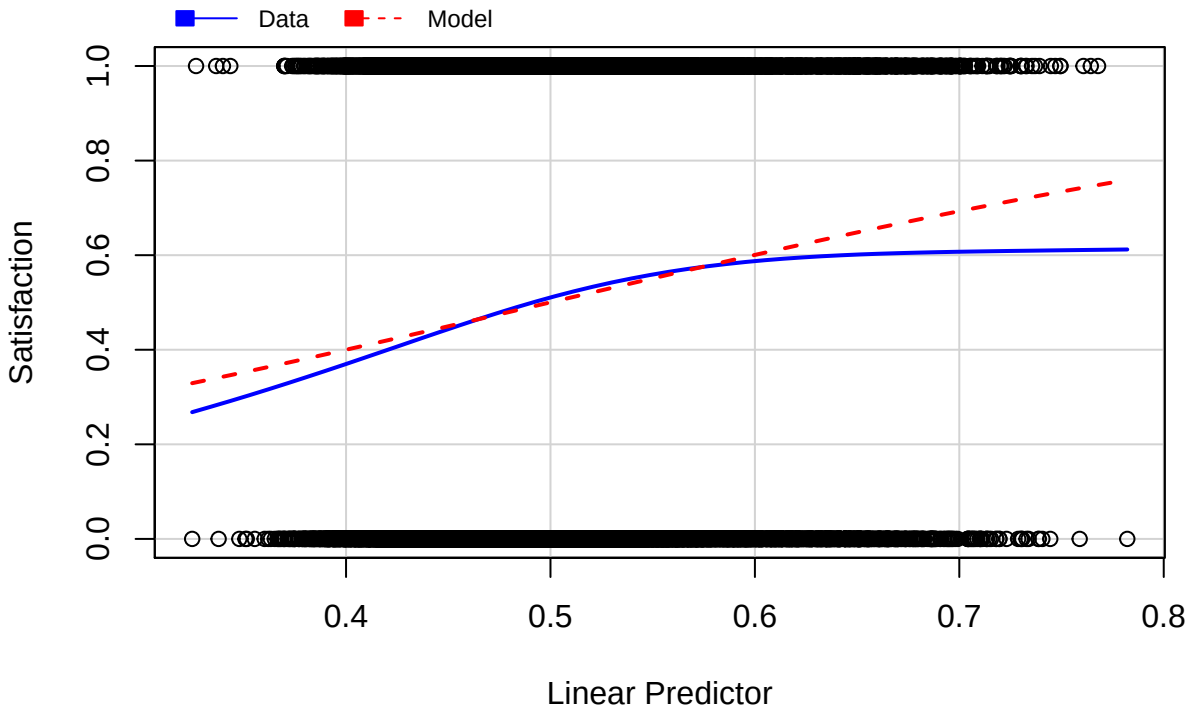
```
## (Intercept)                           -0.27     0.03
## room_typePrivate room                  0.32     0.09
## room_typeShared room                   0.99     0.20
## accommodates                           0.08     0.01
## bedrooms                              -0.21     0.02
## cprice                                 0.11     0.01
## violenceRate                           0.02     0.00
## room_typePrivate room:accommodates -0.11     0.04
## room_typeShared room:accommodates  -0.22     0.10
## ---
##   n = 50406, k = 9
##   residual deviance = 69393.2, null deviance = 69839.2 (difference = 446.0)
```

model check

```
car::marginalModelPlot(fit1)
```




```
car::marginalModelPlot(fit1.1)
```

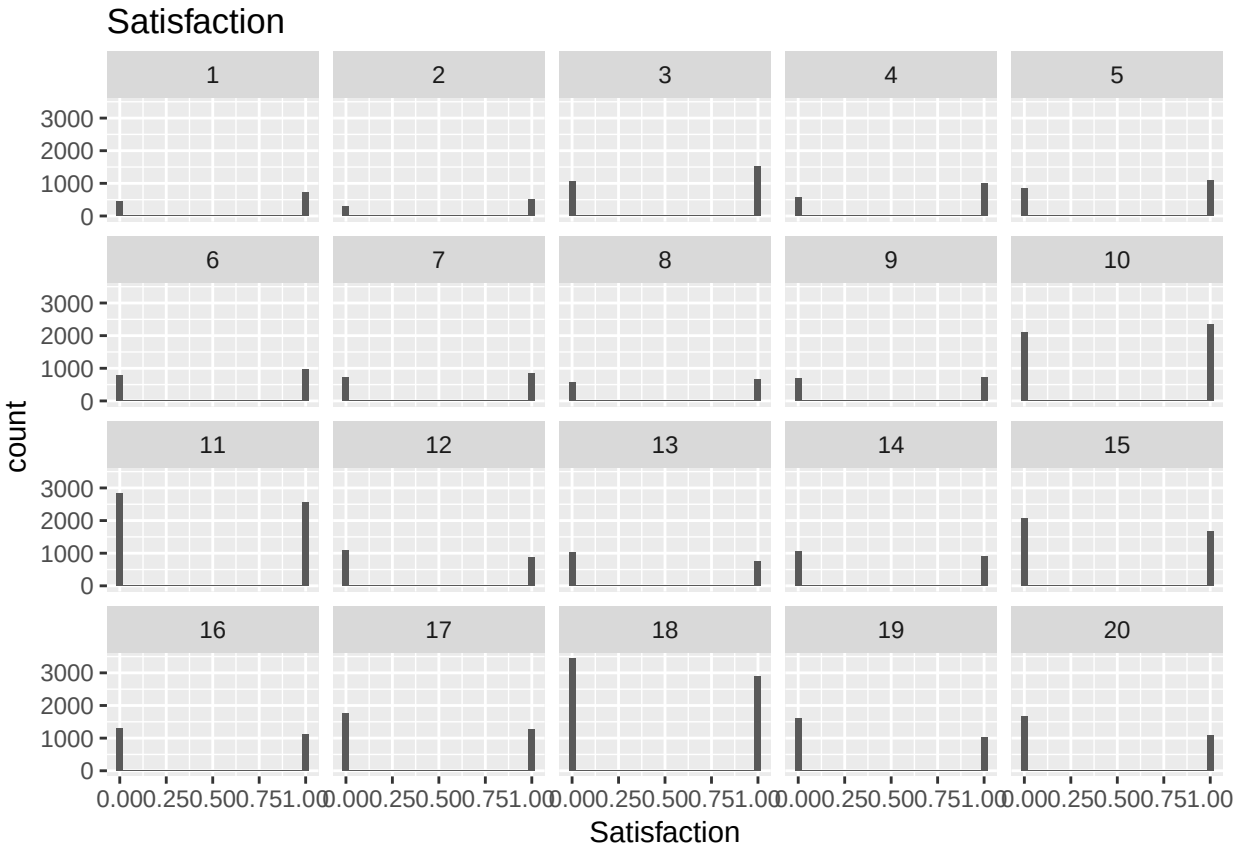


as we can see from the marginal model plot model fit1.1 is a better fit compared with model fit1

3) so we can fit multilevel logistic model

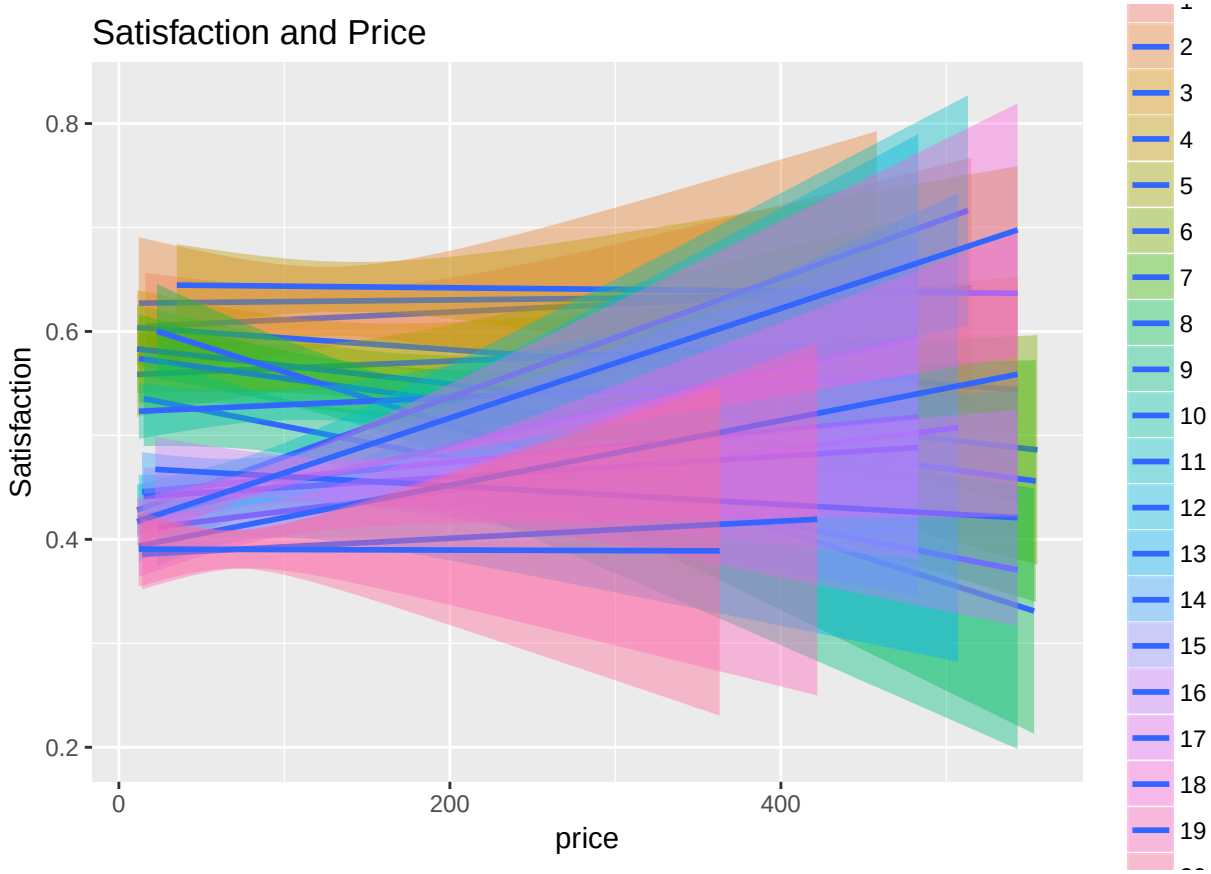
```
ggplot(data=pardata, aes(x=Satisfaction))+geom_histogram()+
  ggtitle("Satisfaction")+facet_wrap(~district)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

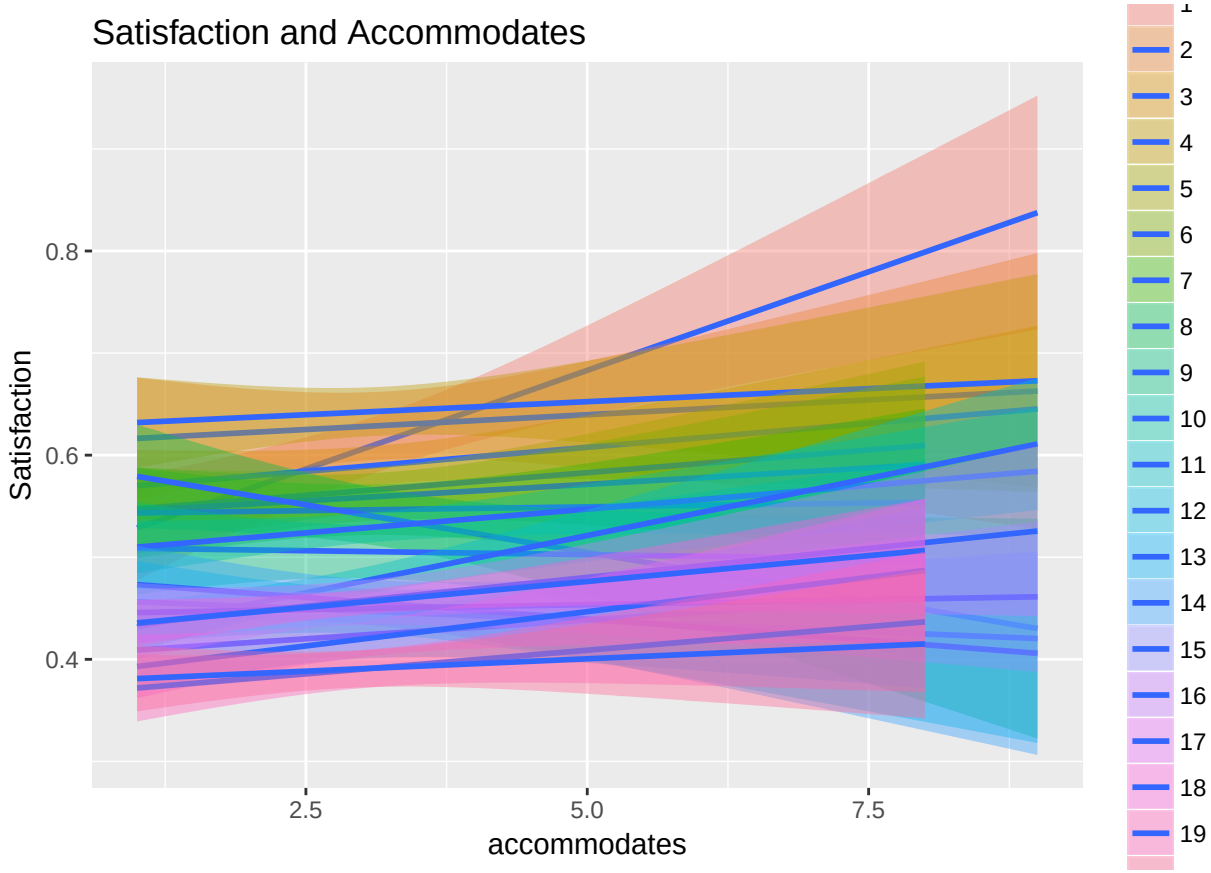


first we can visualize the relationship between predictors and single variable group by district

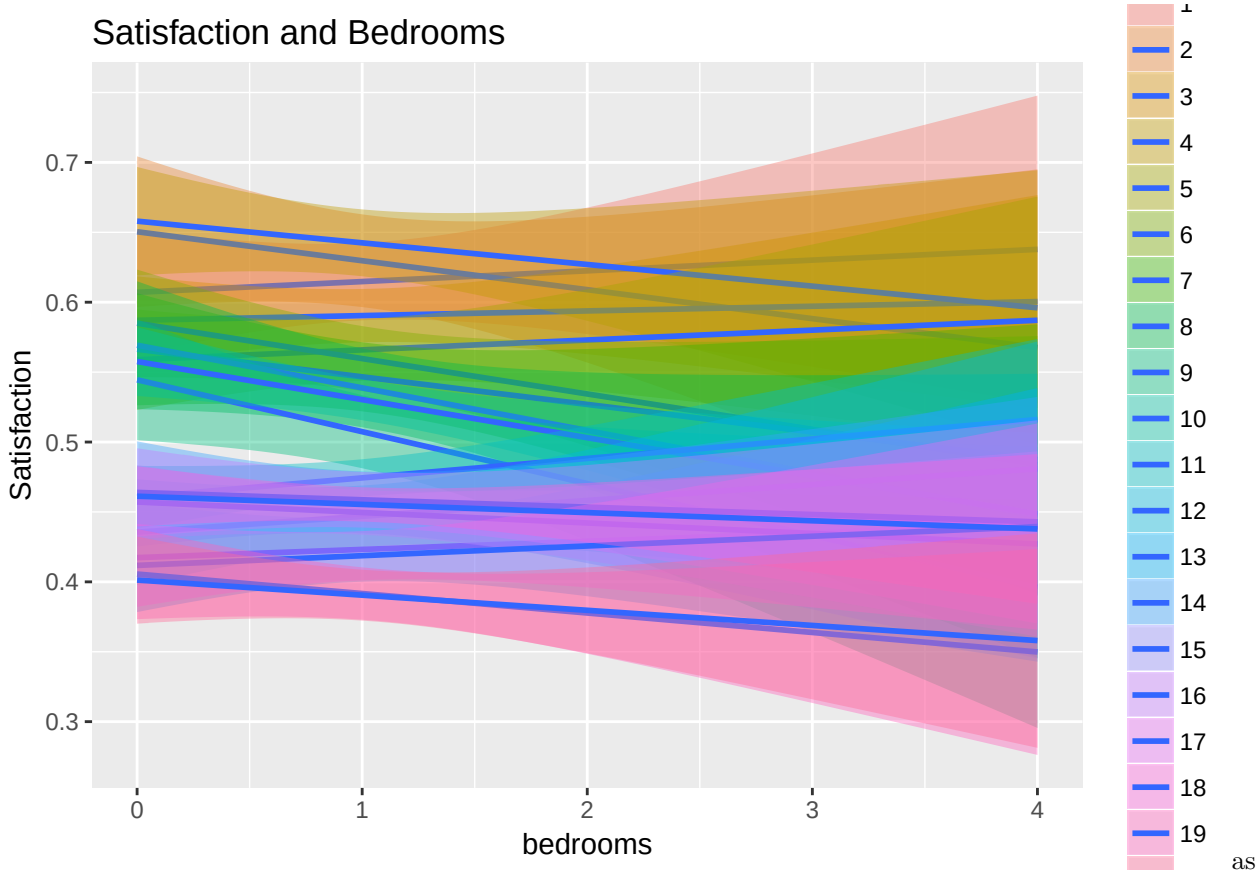
```
ggplot(data=pardata, aes(x=price, y=Satisfaction, fill=district))+
  geom_smooth(method = "glm")+ggtitle("Satisfaction and Price")
```




```
ggplot(data=pardata, aes(x=accommodates, y=Satisfaction, fill=district))+
  geom_smooth(method = "glm")+ggtitle("Satisfaction and Accommodates")
```




```
ggplot(data=pardata, aes(x=bedrooms, y=Satisfaction, fill=district))+
  geom_smooth(method = "glm")+ggtitle("Satisfaction and Bedrooms")
```



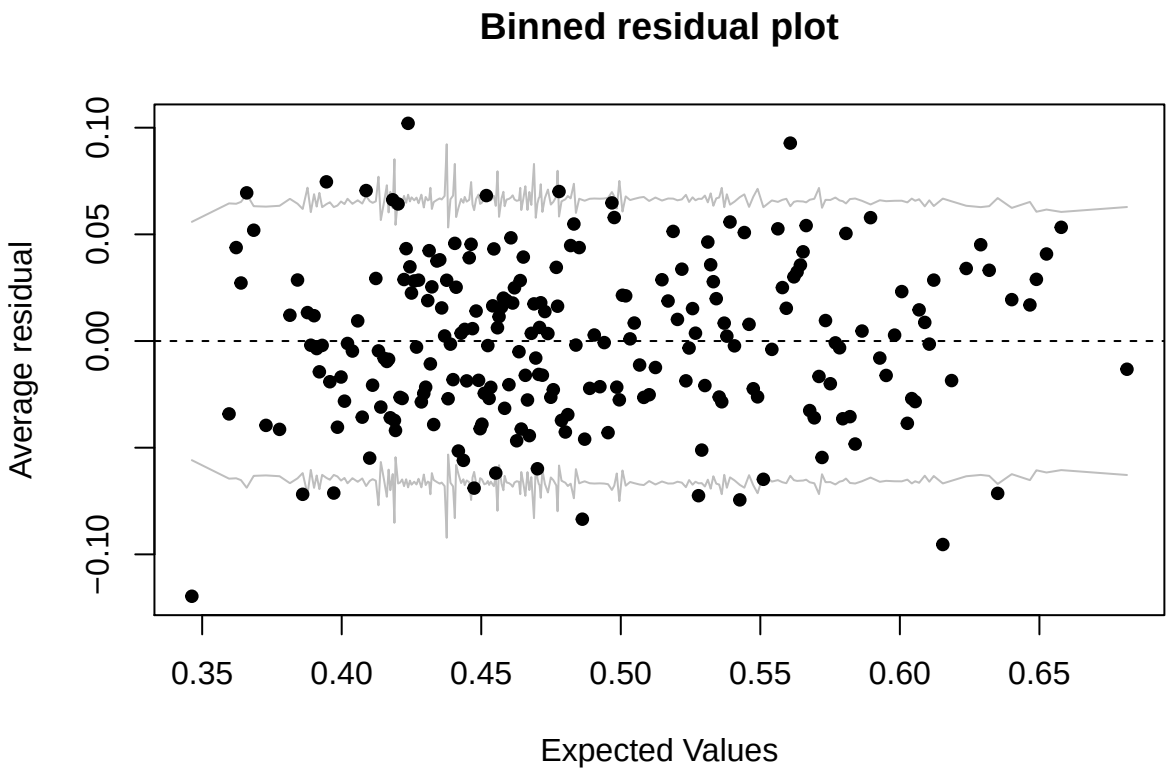
as

we can see from the graphics the slope varies between district so we will start with varying slope models

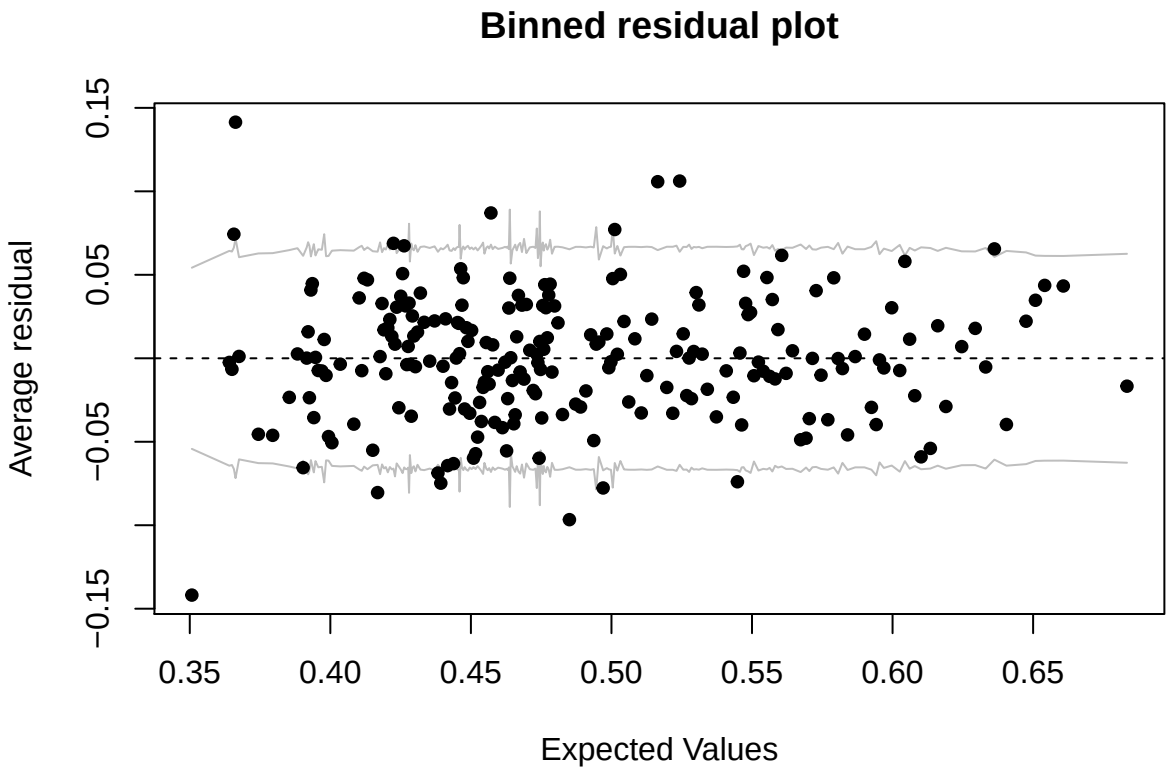
```
fit2<-glmer(Satisfaction~accommodates*room_type+bedrooms+
              violenceRate+cprice+(1+cprice|district), family = binomial, data=pardata)

fit2.1<-glmer(Satisfaction~accommodates*room_type+bedrooms+
              violenceRate+cprice+(1+bedrooms|district), family = binomial, data=pardata)
```

we can check this model by looking at residual plot

```
binnedplot(fitted(fit2),residuals(fit2, type="response"))
```




```
binnedplot(fitted(fit2.1),residuals(fit2.1, type="response"))
```



we can see model fit2 is a better fit compared with model fit2.1

4) model diagnosis 4.1) Deviance and residual analysis now we are going to compare model fit1.1 and model fit2

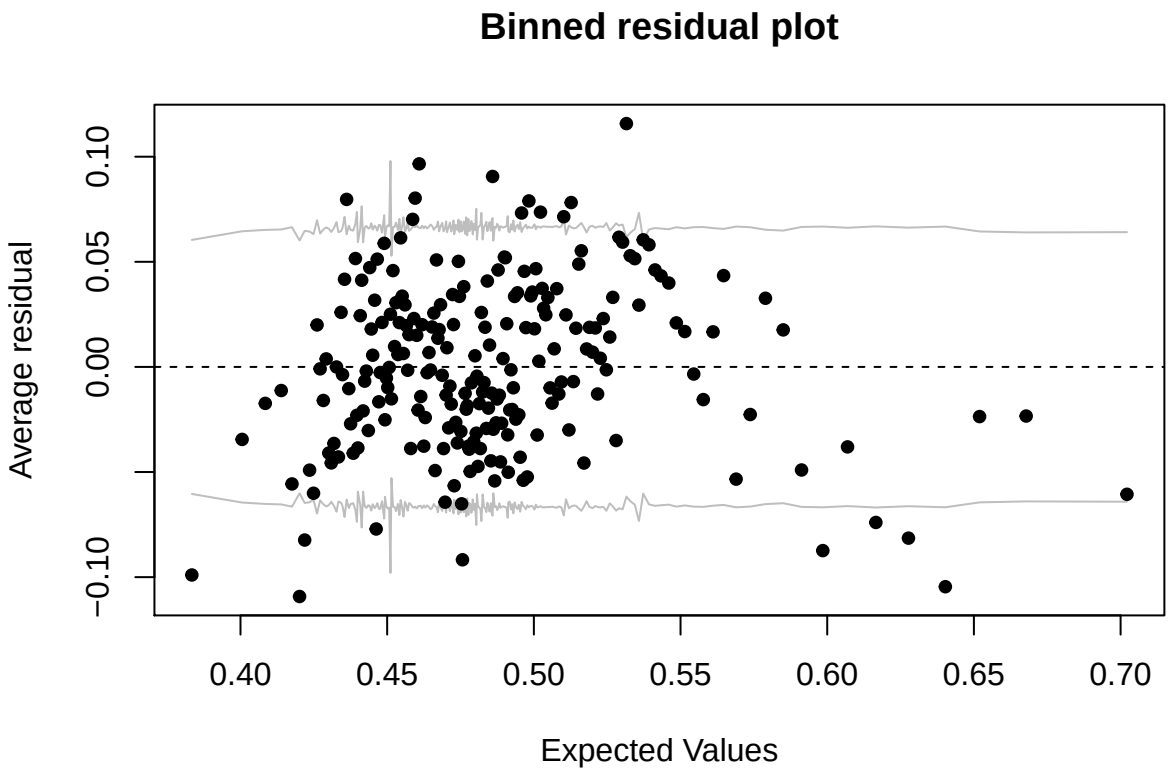
```r
deviance(fit1.1)
```

```
## [1] 69393.21
```

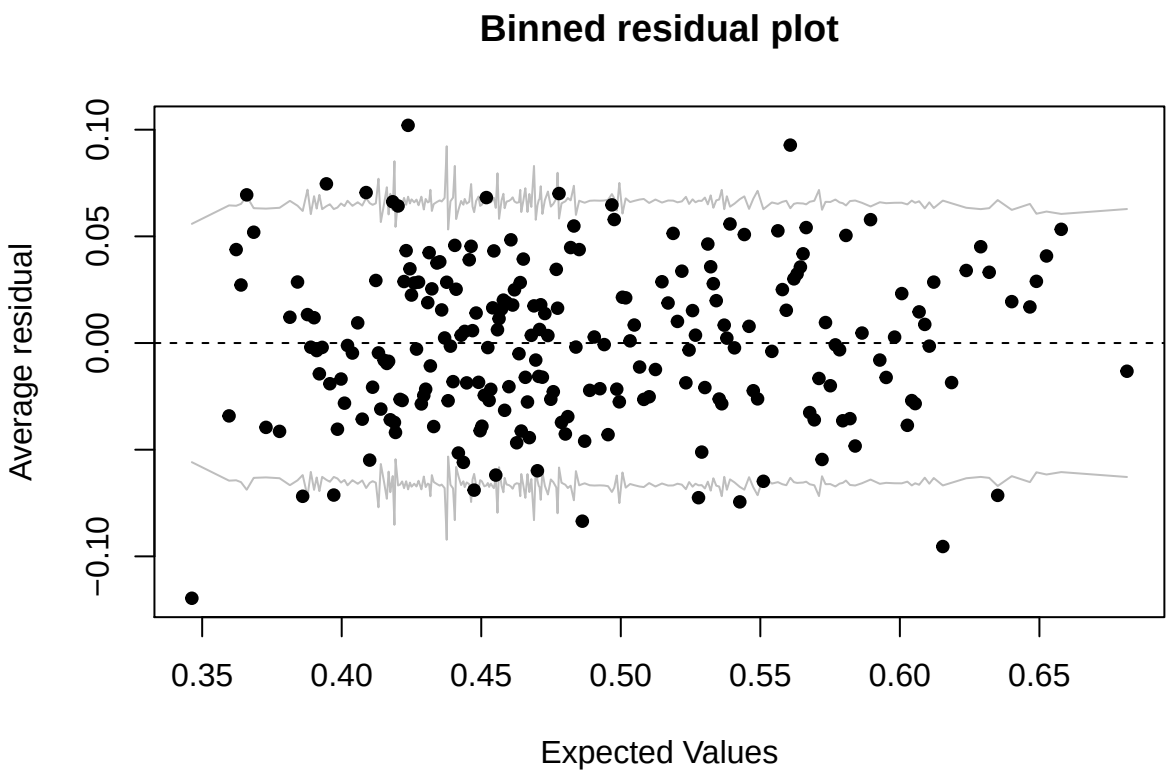
```r
deviance(fit2)
```

```
## [1] 68724.17
```

```r
binnedplot(fitted(fit1.1),residuals(fit1.1, type="response"))
```




```r
binnedplot(fitted(fit2),residuals(fit2, type="response"))
```



as we can see model fit2 is the best fit so far

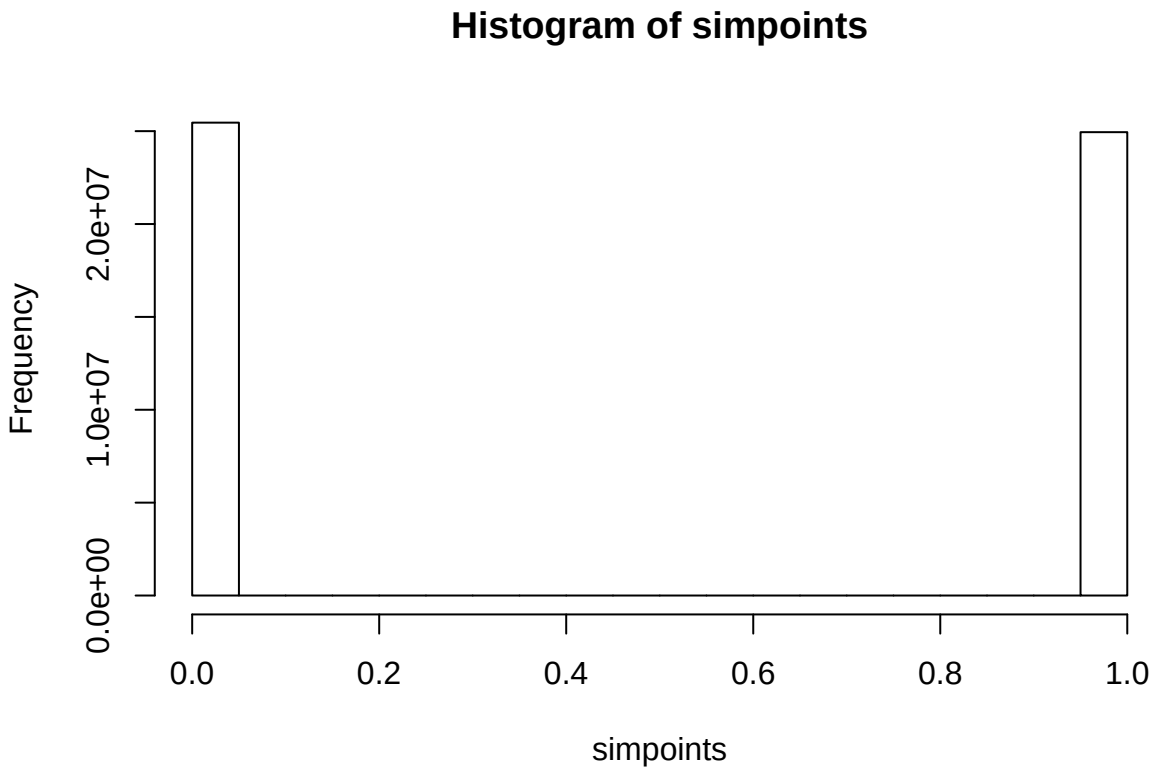
4.2) we are going to do predictive checking

```
display(fit2)
```

```
## glmer(formula = Satisfaction ~ accommodates * room_type + bedrooms +
##     violenceRate + cprice + (1 + cprice | district), data = pardata,
##     family = binomial)
##                                    coef.est coef.se
## (Intercept)                        -0.26     0.10
## accommodates                        0.09     0.01
## room_typePrivate room               0.34     0.09
## room_typeShared room                1.04     0.20
## bedrooms                           -0.16     0.02
## violenceRate                        0.02     0.01
## cprice                              0.03     0.02
## accommodates:room_typePrivate room -0.11     0.04
## accommodates:room_typeShared room  -0.24     0.10
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  district (Intercept) 0.26
##           cprice      0.06     -0.54
##  Residual             1.00
## ---
## number of obs: 50406, groups: district, 20
## AIC = 68870, DIC = 68602.3
## deviance = 68724.2
```

```
newdata<-model.matrix(~accommodates*room_type+bedrooms+
                        violenceRate+cprice,data=pardata)
```

```r
cf<-fixef(fit2)
coefhat<-as.matrix(coef(fit2)$district)
sigma.p.hat<-sigma.hat(fit2)$sigma$data
sigma.p.hat<-sigma.hat(fit2)$sigma$district

n<-nrow(newdata)
simpoints<-matrix(NA,n,1000)

for (i in 1:1000){
  ptide<-invlogit(newdata%*%t(coefhat))
  ytide<-rbinom(n,1,ptide)
  simpoints[,i]<-ytide
}

hist(simpoints)
```




```r
propsim<-sum(simpoints)/50406000
propdat<-sum(pardata$Satisfaction)/50406

print(propsim)
```

```
## [1] 0.4949163
```

```r
print(propdat)
```

```
## [1] 0.486212
```

we can see model fit2 is pretty a good fit for our data

Impplication

```
display(fit2)
```

```
## glmer(formula = Satisfaction ~ accommodates * room_type + bedrooms +
##     violenceRate + cprice + (1 + cprice | district), data = pardata,
##     family = binomial)
##                                     coef.est coef.se
## (Intercept)                         -0.26     0.10
## accommodates                         0.09     0.01
## room_typePrivate room                0.34     0.09
## room_typeShared room                 1.04     0.20
## bedrooms                            -0.16     0.02
## violenceRate                         0.02     0.01
## cprice                               0.03     0.02
## accommodates:room_typePrivate room -0.11     0.04
## accommodates:room_typeShared room  -0.24     0.10
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  district (Intercept) 0.26
##           cprice      0.06     -0.54
##  Residual             1.00
## ---
## number of obs: 50406, groups: district, 20
## AIC = 68870, DIC = 68602.3
## deviance = 68724.2
```

Implication of fixed effect

```
cf<-fixef(fit2)
cf<-as.data.frame(cf)
colnames(cf)<-"coefficient"
```

On average, the number of accommodates, room type, violence rate and price have positive impact on the probability of airbnb being satisfied, among them accommodates and room type are statistically significant (1) specifically, shared room tend to have higher the probability of being satisfied than private room and entire room (2) increase of accommodates could increase the probability of entire room being satisfied while decreasing the probability of being satisfied for shared room and private room (3) each unit increase of price deviate from average price over standard deviation will increase the probability of being satisfied on average; while price is not statistically significant (4) the increase of violence rate also could increase the probability of being satisfied. As the most popular districts generally have higher violence rate. While this predictor is not statistically significant (5) the number of bedrooms has negative impact on the probability of being satisfied, each unit increase of bedrooms will decrease the probability of being satisfied on average

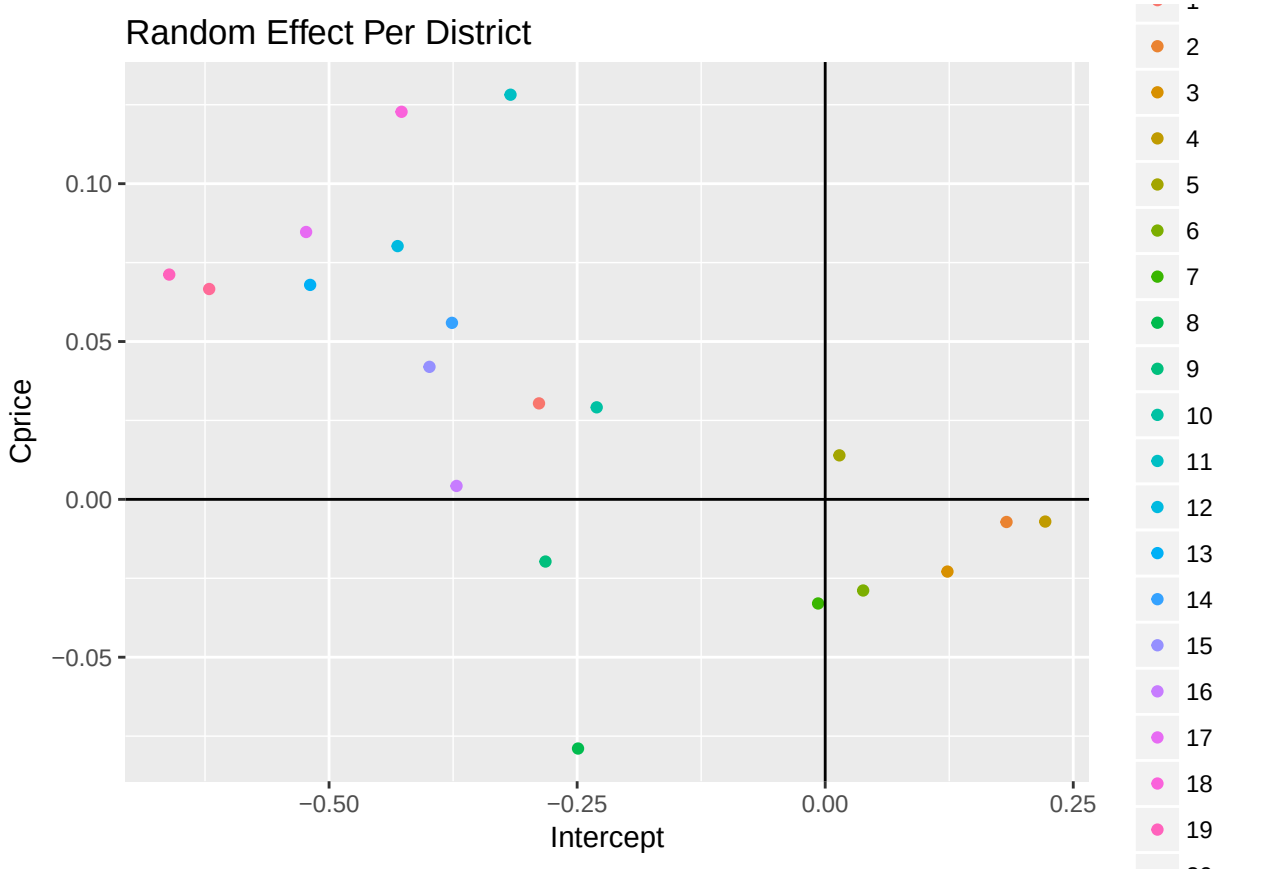
Implication of random effect

```
rand<-coef(fit2)$district
rand<-as.data.frame(rand)
randef<-rand[,c(1,7)]
colnames(randef)<-c("Intercept","Cprice")
randef$district<-c(1:20)
randef$district<-as.factor(randef$district)
```

random intercept and random slope

```
ggplot(data = randef, aes(x=Intercept, y=Cprice, col=district))+
  geom_point(size=1.5)+
  scale_color_discrete(randef$district)+ggtitle("Random Effect Per District")+
```

```
geom_vline(xintercept = 0)+geom_hline(yintercept = 0)
```



(1) random intercept: we can see 2,3,4,5,6 districts have positive intercept, which means they have higher probability of being satisfied with average price than other districts when other variables are same

(2) random slope: we can see 2,3,4,6,7,8,9 districts have negative slope, which means each unit increase of price deviating from average price over standard error will decrease probability of being satisfied, while other districts have positive slope

clustering we can cluster 20 districts into four categories

(1) Star District (positive random intercept and random slope): district 5th Airbnbs in this district have higher probability of being satisfied generally, and there is still potential for airbnb hosts to increase the price

(2) Golden District(positive random intercept but negative random slope): district 2nd, 3rd, 4th, 6th, 7th Airbnbs in this district have higher probability of being satisfied generally, but they are already highly priced, further increase in price will decrease the ratings

(3) Problem District (negative random intercept and random slope) district 8th, 9th Airbnbs in this district have lower probability of being satisfied generally, but they are already highly priced, further increase in price will decrease the ratings

(4) Potential District (negative random intercept and random slope) district 1st, 10th, 11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th Airbnbs in this district have lower probability of being satisfied generally, but there is still potential for airbnb hosts to increase the price