How to Obtain High Customers' Review Scores in Airbnb

Airbnb Customers' Reviews Rating Analysis

Qixuan, Zhang

Contents

Abstract

This study investigates key factors that influence Airbnb customer experiences by analyzing big dataset of online review comments through the process of text mining and sentiment analysis. Based on the key attributes, we will give some suggestions to future Airbnb hosts. Compared with traditional study, the innovation of this study is using AFINN dataset to calculate sentiments scores of Airbnb customers' comments to study customers' opinions about Airbnb services. From our analysis, we found price is not the crucial factor influence customers' experience, and customers' review scores are based on location, accuracy, communication, room types and cleanliness. Methodologically, this study contributes to improve the satisfaction of customers and illustrate how big data can be used and visually interpreted in marketing studies.

Key Words: Customers' Reviews, Latent Rating Regression, Multilevel Regression

1.Introduction

1.1 Background

With advantages of price, there is no doubt Airbnb is becoming the best choice of more and more people. Due to high volume of customers, it is crucial for hosts to provide a comfortable and high-quality room for customers. Based on those previous studies, most of researchers paid more attention to listing dataset, they analyzed customers experience by using reviews rate as response variable and took economy-based variables as predictors.

In this study, we used sentiment analysis get the mean sentiments scores of each apartment in the cleaned listing dataset. And then we will use sentiment scores as response to fit our models and compare the results with the results based on traditional response (review rating/value).More specifically, we will compare key factors of customer reviews from different models. And then we figure out the key attributes that influence customers reviews and give some suggestions to improve customer reviews rate.

2 Previous work

2.1 Cleaning Dataset and summary variables

Data-source Description:

The Boston dataset was released by Airbnb itself on 08 October, 2018 to show people that how Airbnb is really being used and is affecting their neighborhood. The original data set had 6041 rows and 96 variables (columns) in Boston listing dataset, Boston reviews data had 173818 rows and 6 columns.

Cleaning Dataset:

We kept the variables that are relevance to customers reviews in listing dataset. Furthermore, we deleted NA values and filtered numbers of reviews more than three. In this study, we focused on price interval of \$0-\$1000 rental Airbnb apartments in Boston. The cleaned listing dataset we used had 3611 rows and 23 variables. And for Boston reviews dataset, we only deleted NA comments and non-English comments.

2.2 Summary Variables:

Summary Variables

Although there are 100+ variables in the 2 datasets, only 21 are used for the analysis. We briefly describe such variables is given below:

TABLE 2.1

ID	Unique Hosts
NEIGHBOURHOOD_CLEANSED	Defined apartment area in Boston
PROPERTY TYPE	The type of house that hosts provided
ROOM TYPE	The type of room that hosts provided
BATHROOMS	The number of bathrooms
BEDS	The number of beds
BEDROOMS	The number of bedrooms
PRICE	Price
REVIEW SCORES RATING	Overall scores rating of customers
REVIEW SCORES CLEANLINESS	Rating of cleanliness
REVIEW SCORES ACCURACY	Rating of accuracy
REVIEW SCORES LOCATION	Rating of location
REVIEWSCORESCOMMUNICATION	Rating of communication
REVIEW SCORES CHECKIN	Rating of check-in
NUMBER OF REVIEWS	Number of reviews of a unique host
COMMENTS	Comment of customers
SENTIMENT	Sentiment scores of customers
DESCRIPTION	Description of host’s house
ACCOMODATES	The number of accomodates

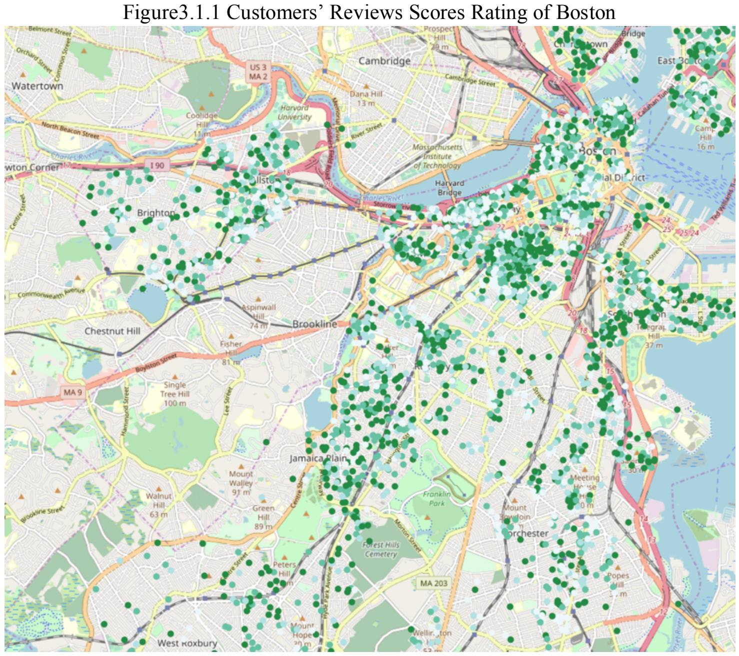
3. Exploratory Data Analysis

3.1 EDA of Boston customers reviews

3.1.1 Distribution of Airbnb customers' reviews in Boston

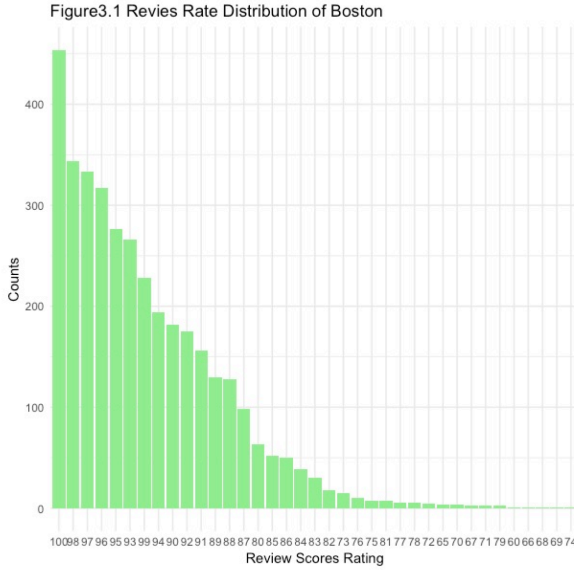
To give our reader a more observable visualization, we made a leaflet to describe the review scores distribution in different area in Boston at first. With stronger and stronger of the green color, the review scores rating are higher.

Figure3.1.1 Customers' Reviews Scores Rating of Boston



Following leaflet plot of Boston customers' reviews, we made a histogram plot to describe reviews of Boston.

Figure3.1 Revies Rate Distribution of Boston



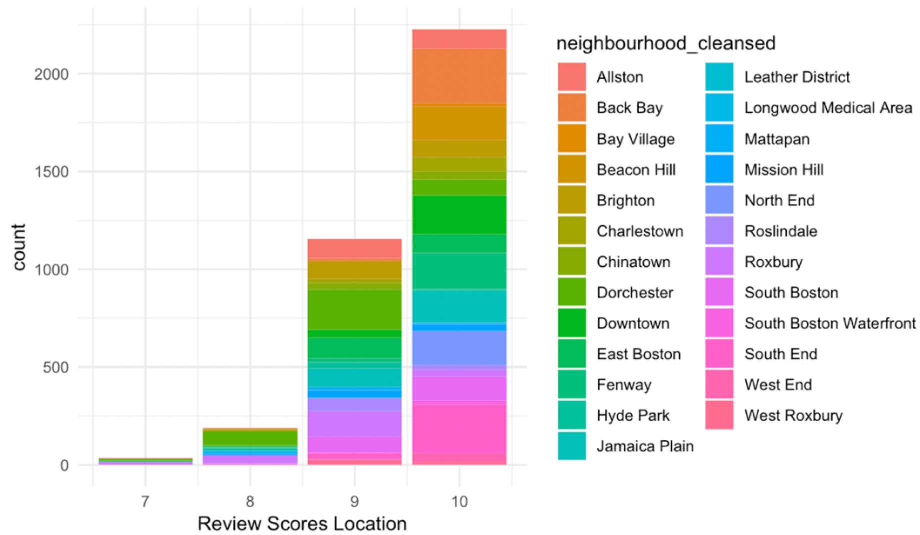
From Figure 3.1, we could get conclusion that most people prefer to give positive feedback. That made me more curious about what caused people to give negative feedback. Back to the Boston listing dataset, we found that reviews scores rating related to several aspects, therefore we made several plots to visualize and analyze different aspects related to customers' reviews.

3.1.2 VARIABLES ANALYSIS

There are several categorical variables, from cleansed Boston listing table, therefore we made several plots to illustrate different relationship between customers reviews scores and variables.

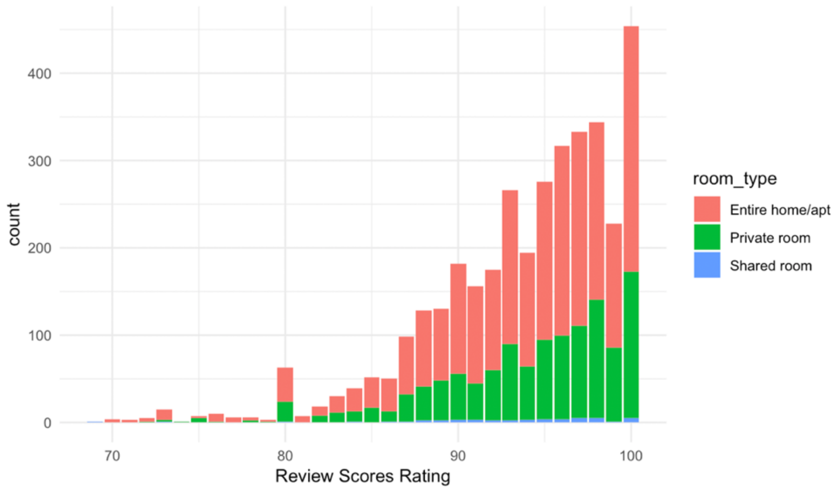
Initially, we want to figure out review scores distribution of different neighborhoods. We made a bar plot to illustrate counts of different location related to their customers reviews scores for location, in Figure 3.2 we could find the best location of Airbnb apartments is located in Allston. Because more than 2000 apartments located in Allston was given 10 scores by customers.

Figure 3.2 Review Scores Location Of Different Neighbourhood



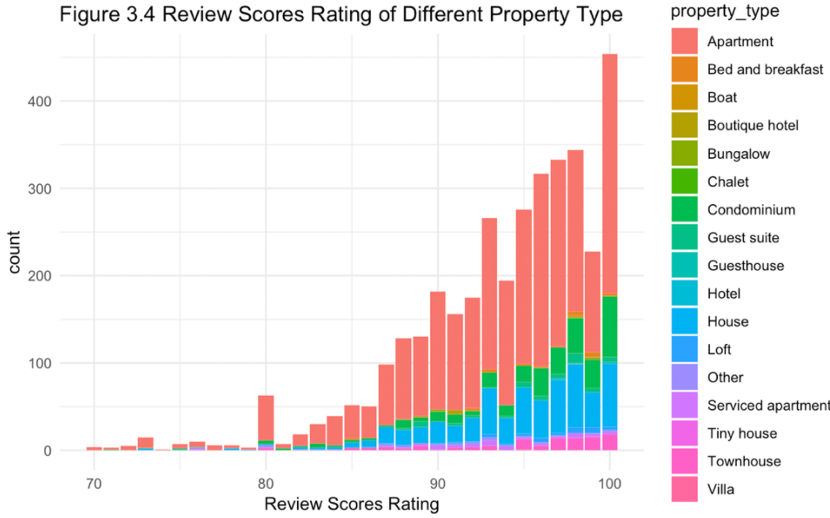
Based on the inspiration of neighborhood related to customers' reviews scores, we analyzed plots of property type, room type and price related to customers' reviews. From Figure 3.3, we found that the most common and popular room type is Entire home/apt, and there are few Shared room in Airbnb.

Figure 3.3 Review Scores Rating of Different Room Type



And we found similar conclusion when we observed Figure 3.4, we found the most popular property type is Apartment. As for price analysis, we made

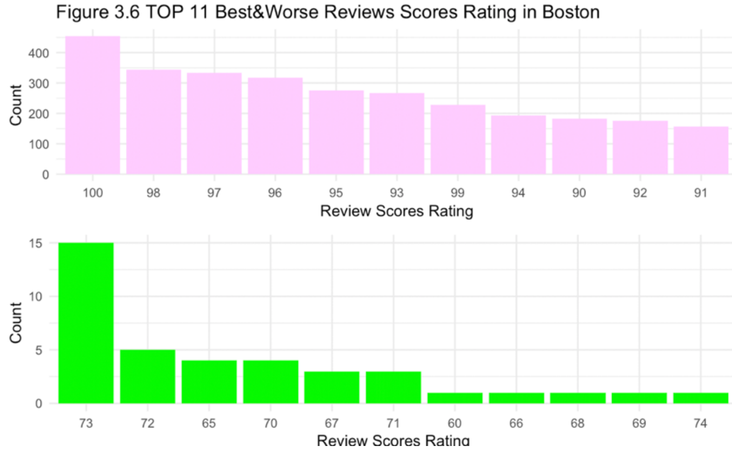
Figure 3.4 Review Scores Rating of Different Property Type



3.1.3 Compared different reviews scores rating in Boston

Based on above distribution of customers' review scores in Boston, we classified the review scores to three parts, good, normal and bad. As for good part, it included review scores rating from customers that more than 90. And for the bad part, it included review scores rating less than 70. And within the interval of 70-90, we classified it marked by normal. And we added evaluation to listing table to illustrate the different. After classified review scores rating, we made a plot to compare the distribution of top 11 best reviews and top 11 worse reviews in Boston.

Figure 3.6 TOP 11 Best&Worse Reviews Scores Rating in Boston

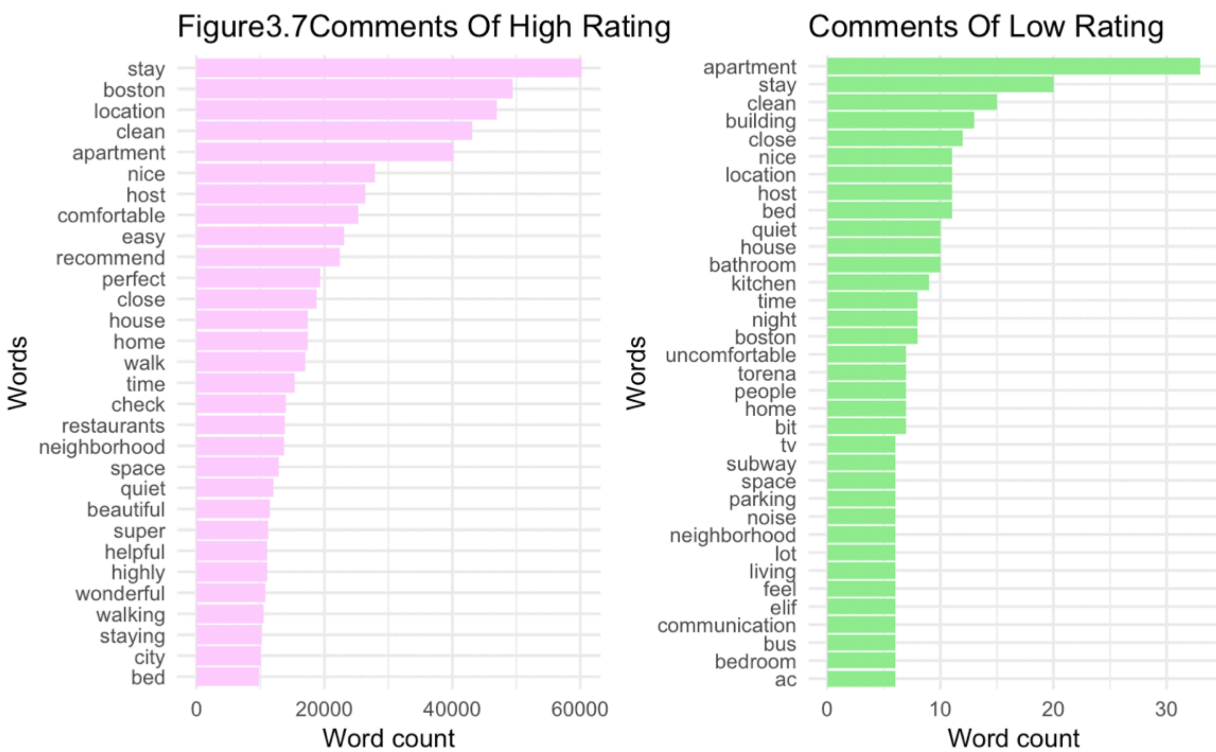


3.2 Text mining Customers' Review

3.2.1 word counts analysis of customers' comments

Based on visualization plots of different variables, we want to figure out some meaning information that latent in customers' reviews. We believe that these comments made Airbnb customers give positive and negative feedback. And we also think based on comments sentiments analysis, we could get more specific idea that made us to get close to what happened during the customers' Airbnb trip. Initially, we want to see the description combined with review scores. We divided our rating scores rating to two groups, first one is high rating group. We filtered the review scores rating more than 90 as the high rating group. And then we filter the review scores rating less than 70 as low rating group. After that, we apply these following methods to each city.

We get the plot of Top frequency words of high and low customers' scores rating in Boston as following(Figure3.7). From comments words counts of low customers' rating, we found some common word like noise, uncomfortable, etc. Compared with comments of low rating words clouds plot, we can find lots of positive words easily like clean, perfect, easy, beautiful, super, highly, super in comments of high rating words plot.



Furthermore, we would like to build two would clouds to illustrate the specific frequency words of these above two parts. In Figure 3.8 and Figure 3.9, we displayed word clouds of high scores rating comments and low scores rating comments.

Figure 3.8 High Rating Word Cloud

Figure 3.9 Low Rating Word Cloud

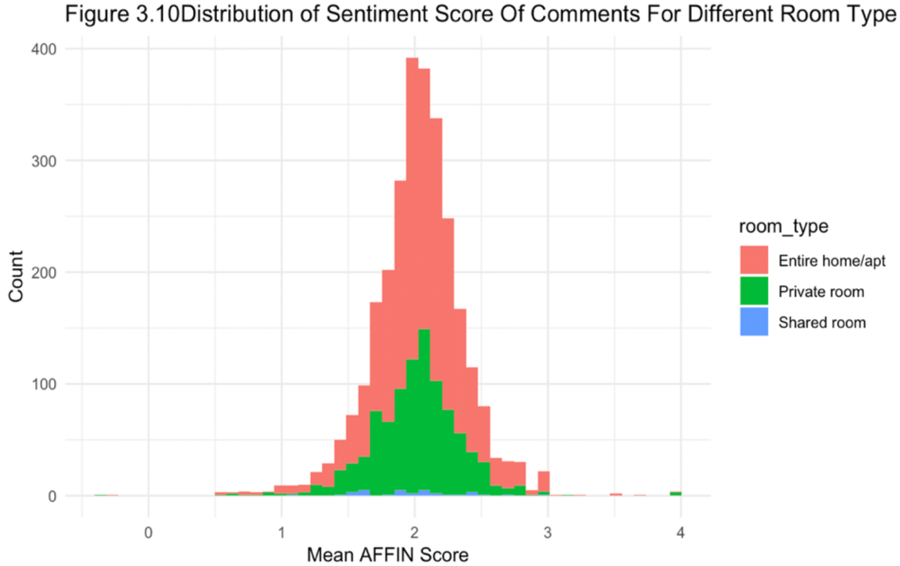
When we complete the part of checking frequency words and words cloud, we move forward to sentiments analysis of Airbnb customers' comments and we want to figure out the latent factors that influence customers' review scores rating.

3.2.2 sentiment analysis

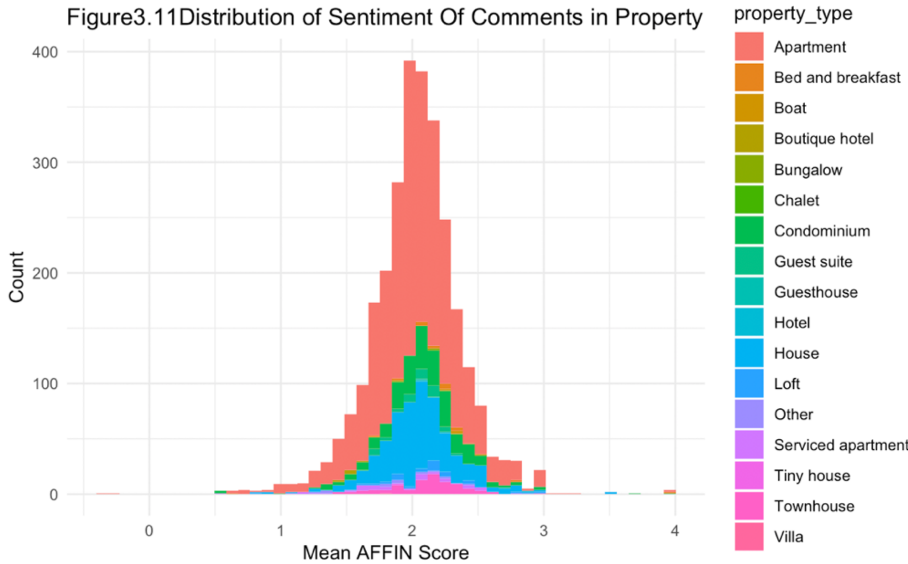
There is no doubt that comments of customers reflect why they give the positive or negative feedback to hosts. However, Airbnb did not provide a review system which could evaluate customers' comments. In this study, we calculated sentiment scores by using Airbnb customers' comments words joined with AFINN sentiment dataset, and then we create a new column to named sentiment which recorded sentiments scores corresponding to unique host ID. More specifically, we want to confirm if review scores rating is perfectly reflecting customers' reviews.

After adding sentiment scores column, we want to analyze several customer behavior-related variables combined with review scores. More specifically, we want to confirm if review scores rating is perfectly reflecting customers' reviews. After getting sentiment scores of every host, we made the distribution of comments in different room type, plot that shown in Figure 3.10, by using AFFIN Lexicon sentiment scores.

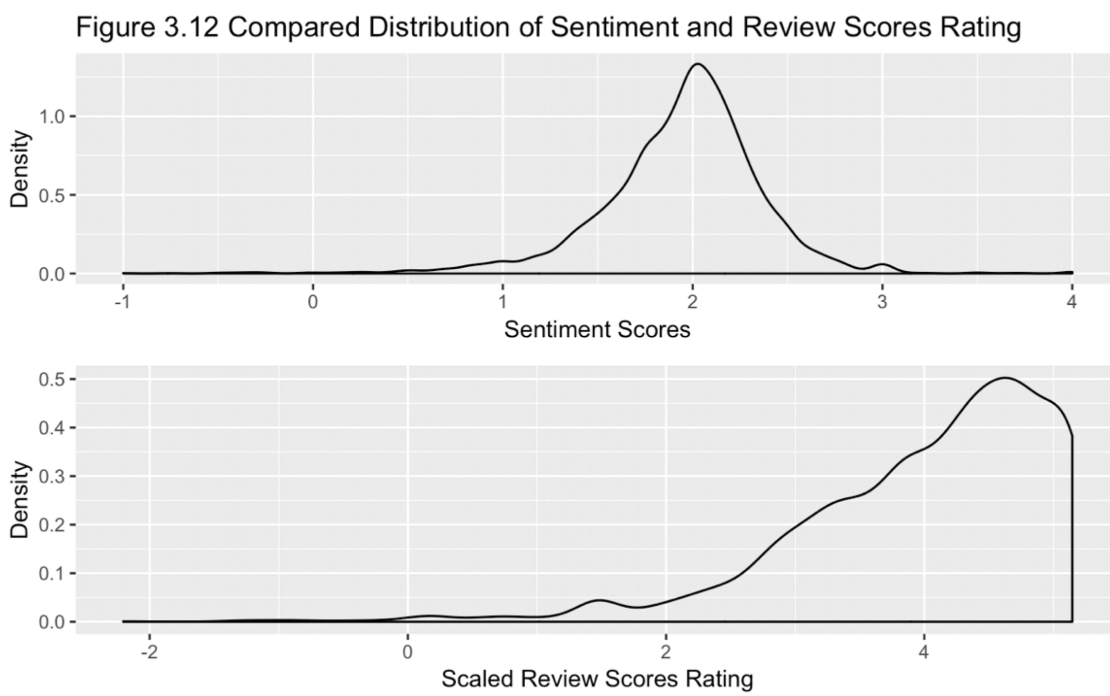
Figure 3.10Distribution of Sentiment Score Of Comments For Different Room Type



From Figure 3.11, we could get the conclusion that the distribution of sentiment scores follow normal distribution and Apartments is most popular property type.



Comparing with reviews scores rating, from Figure 3.12, the distribution of sentiment scores is different from customers scores rating. It demonstrates that there is some useful information that embedded in customers' comments.



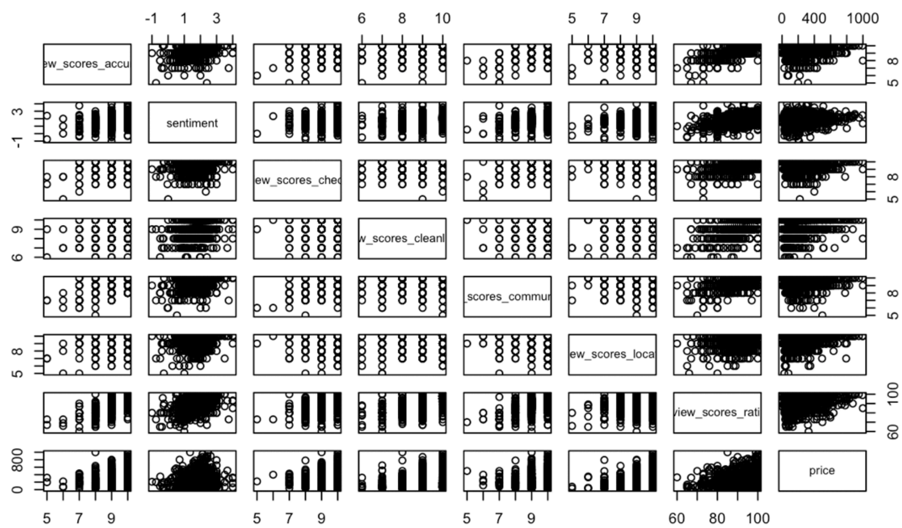
4. Method

4.1 Linear Regression Model

After exploratory of data analysis and text mining analysis process for Airbnb customers’ reviews and listing dataset, we move forward to model building step.

As for using Linear Regression Model to demonstrate customers’ reviews rating, we built a correlation plot to observe relationship between these variables. And then, we divided our dataset to training dataset and test dataset. We randomly pick up 2400 rows as training dataset and use it to predict our model.

Figure4.1 Plot numeric variables correlation



By using training dataset, the results of model1 (lm1) display as following:

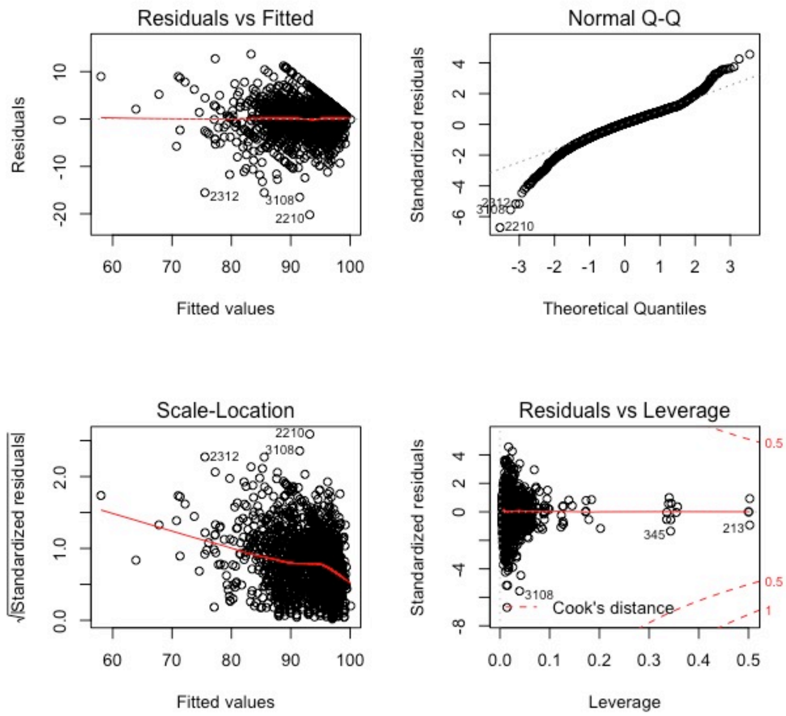
Figure 4.2 Linear Regression Model

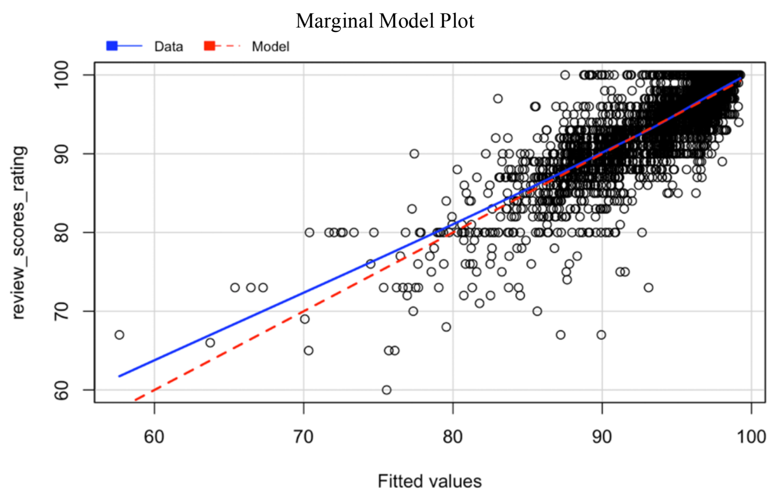
```
Response: review_scores_rating
                              Df Sum Sq Mean Sq  F value    Pr(>F)    
scale(price)                   1      5       5   0.5403 0.4623788    
neighbourhood_cleansed        24   5312     221  24.1185 < 2.2e-16 ***
accommodates                   1     43      43   4.7205 0.0298964 *  
bathrooms                      1     63      63   6.8186 0.0090741 ** 
bedrooms                       1      0       0   0.0205 0.8862023    
property_type                 15   1409      94  10.2367 < 2.2e-16 ***
room_type                      2     12       6   0.6533 0.5204282    
beds                           1     39      39   4.2609 0.0390994 *  
review_scores_accuracy         1  37008   37008 4032.5261 < 2.2e-16 ***
review_scores_location         1   1396    1396 152.0895 < 2.2e-16 ***
review_scores_communication    1   2921    2921 318.3339 < 2.2e-16 ***
review_scores_cleanliness      1   4775    4775 520.2905 < 2.2e-16 ***
review_scores_checkin          1    114     114  12.4098 0.0004346 ***
Residuals                   2555  23448       9                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the Linear Regression Model summary, the model p-value and predictor’s p-value are less than the significance level, so we know we have a statistically significant model. For now, key attributes of customers’ reviews are comments, neighborhood, property type, beds, bathrooms, accommodates, review scores accuracy, review scores cleanliness, review scores communication and review scores check in.

After fitting linear regression model, we will diagnose and check our linear models to get more accuracy result under the correct hypothesis. Figure 4.3 displays the residual plot and marginal model plots of linear model1.

Figure 4.3 Linear Regression Model Residual plots





4.2 Latent Linear Regression Model

Going beyond the scores rating to know the opinions of a reviewer on different aspects is important because different reviewers may give a host the same scores rating for very different reasons. For example, under the same 80 review scores rating, one customer may be unsatisfied with location and the other one may be unhappy with the bed in the room. Therefore, it is important for us to add this latent information in customers' comments to get more accuracy and reasonable customers' review scores related to their Airbnb experience.

Furthermore, even if we can reveal the rating on an aspect such as "price", it may still be insufficient because "cheap" may mean different price ranges for different reviewers. Even the same reviewer may use a different standard to define "cheap" depending on how critical other factors (e.g. location) are; intuitively, when a reviewer cares more about the location, the reviewer would tend to be more willing to tolerate a higher price. To understand such subtle differences, it is necessary to further reveal the relative importance weight that a reviewer placed on each aspect when assigning the overall rating.

We applied the above sentiment scores to our linear regression model as weights attributes of influence customers' reviews.

Figure 4.4 Latent Linear Regression Model

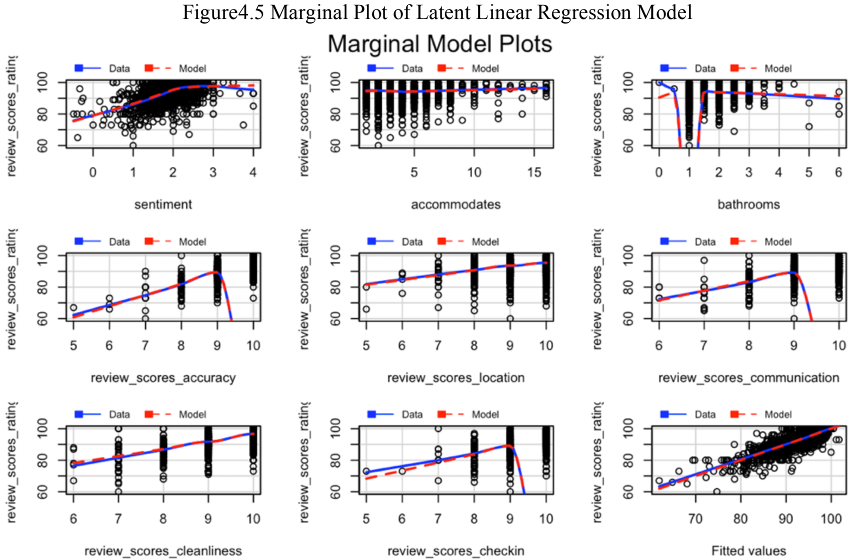
Response: review_scores_rating

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
scale(price)	1	5.0	5.0	0.5749	0.44841	
sentiment	1	18311.1	18311.1	2122.9020	< 2.2e-16	***
neighbourhood_cleansed	24	3422.4	142.6	16.5325	< 2.2e-16	***
accommodates	1	24.7	24.7	2.8629	0.09076	.
bathrooms	1	145.8	145.8	16.9081	4.047e-05	***
bedrooms	1	0.8	0.8	0.0912	0.76267	
property_type	15	598.9	39.9	4.6287	7.676e-09	***
room_type	2	10.1	5.0	0.5832	0.55820	
beds	1	31.8	31.8	3.6848	0.05502	.
review_scores_accuracy	1	24652.0	24652.0	2858.0384	< 2.2e-16	***
review_scores_location	1	884.4	884.4	102.5340	< 2.2e-16	***
review_scores_communication	1	2155.3	2155.3	249.8740	< 2.2e-16	***
review_scores_cleanliness	1	4191.5	4191.5	485.9444	< 2.2e-16	***
review_scores_checkin	1	81.8	81.8	9.4886	0.00209	**
Residuals	2554	22029.5	8.6			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The results display in Figure 4.4, from the above chart we could get the conclusion that only the review scores of different aspects did not include all the sufficient and reliable information. As we can see, combined with sentiment scores, we could get more accuracy predict results.

Figure4.5 Marginal Plot of Latent Linear Regression Model

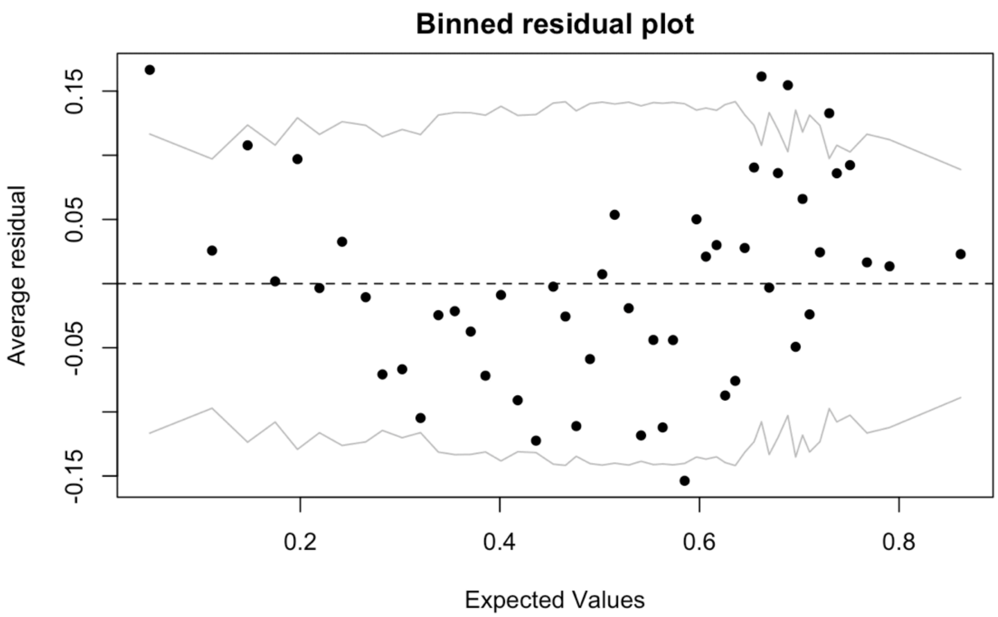


From Figure 4.5, the marginal plots of Latent Linear regression model, we could find that sentiments scores of comments are perfectly relevant to customers' rating scores. It means that we could use sentiment scores to get more accuracy results. More specifically, we will create Airbnb customers' comments table in the future. We could use that sentiment table to get more accuracy sentiment scores and use that scores to predict review scores rating.

4.3 Logistic Regression Model

To fit Logistic Regression Model, we use the same evaluation variable as we mentioned before. Figure 4.5 displays the residual binned plot from Logistic Regression Model.

Figure 4.5 Binned residual plot



The accuracy of Logistic Regression Model is only 65.29%. Figure 4.6 shows ROC plot of Logistic Regression Model. From our confusion matrix(Figure 4.7) we find that no information is 66%, the p-value of Logistic Regression Model is only 0.66. Therefore, we do not recommend applied this model to classify key attributes and predict the customers' reviews.

Figure 4.6 ROC plot of Logistic Regression Model

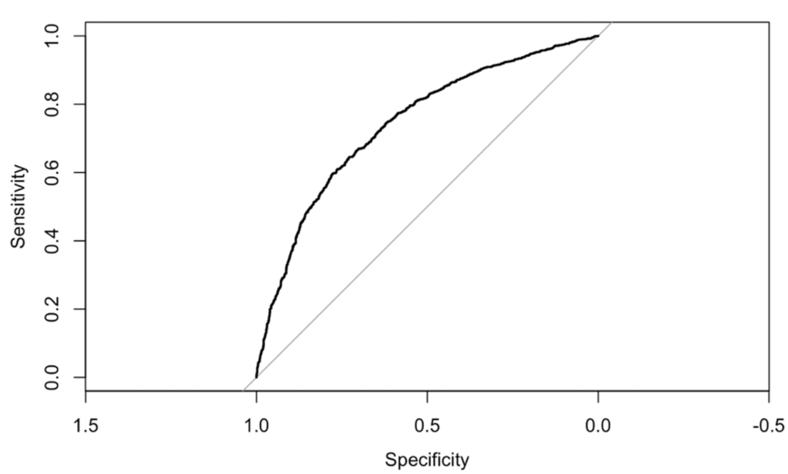


Figure4.7 Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 366  89
         1 209 206

               Accuracy : 0.6575
                 95% CI : (0.6249, 0.689)
    No Information Rate : 0.6609
    P-Value [Acc > NIR] : 0.6003

                  Kappa : 0.3047
 Mcnemar's Test P-Value : 5.444e-12

            Sensitivity : 0.6365
            Specificity : 0.6983
         Pos Pred Value : 0.8044
         Neg Pred Value : 0.4964
             Prevalence : 0.6609
         Detection Rate : 0.4207
   Detection Prevalence : 0.5230
      Balanced Accuracy : 0.6674

       'Positive' Class : 0
```

4.4 Multilevel Models

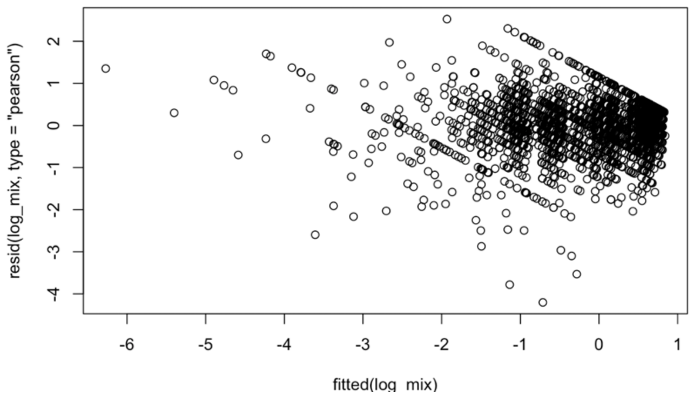
Based on above model analysis, we consider utilizing Multilevel Regression Model to compare differences between different group. We use neighborhoods as our first group, it contains 21 kinds of different neighborhoods. And then we consider using property type as our second group. We build four Multilevel models listed as following, in the following chart(Figure4.8), the log_mix3 model is Multilevel Logistic Model.

Figure 4.8 Compare Different Multilevel Models

```
refitting model(s) with ML (instead of REML)
Data: reviews_sentiment3
Models:
log_mix1: Evaluation2 ~ review_scores_location + review_scores_accuracy +
log_mix1:     review_scores_cleanliness + review_scores_communication +
log_mix1:     (0 + review_scores_accuracy | property_type) + (1 | property_type)
log_mix2: scale(review_scores_rating) ~ bathrooms + review_scores_location +
log_mix2:     review_scores_accuracy + review_scores_cleanliness + review_scores_communication +
log_mix2:     (0 + review_scores_accuracy | neighbourhood_cleansed) + (1 |
log_mix2:     property_type)
log_mix3: Evaluation2 ~ bathrooms + review_scores_location + review_scores_accuracy +
log_mix3:     review_scores_cleanliness + review_scores_communication +
log_mix3:     (0 + review_scores_accuracy | neighbourhood_cleansed) + (1 |
log_mix3:     property_type)
log_mix: Evaluation2 ~ bathrooms + review_scores_location + review_scores_accuracy +
log_mix:     review_scores_cleanliness + review_scores_communication +
log_mix:     (1 + review_scores_accuracy | neighbourhood_cleansed)
         Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
log_mix1  8 2334.5 2383.7 -1159.2   2318.5
log_mix2  9 5998.0 6053.4 -2990.0   5980.0    0.00      1          1
log_mix3  9 2335.9 2391.3 -1159.0   2317.9 3662.11      0     <2e-16 ***
log_mix  10 2230.8 2292.3 -1105.4   2210.8  107.13      1     <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From Figure 4.8, we could get the conclusion that log_mix is the best model of Multilevel Models. And the residual plot of log-mix model is shown in Figure 4.9. From Figure 4.9 we could find customers' reviews different from neighborhoods location.

Figure 4.9Residual Plot of Log-mix



4.5 Multinomial Model

Before building Multinomial Model for Airbnb customers' reviews, we build a new column named Evaluation2 at first. We divided the review scores rating to 4 levels and we listed as following:

	Review scores rating	Evaluation2
Excellent	>90	4
Good	(80,90]	3
Normal	(75,80]	2
Bad	<75	1

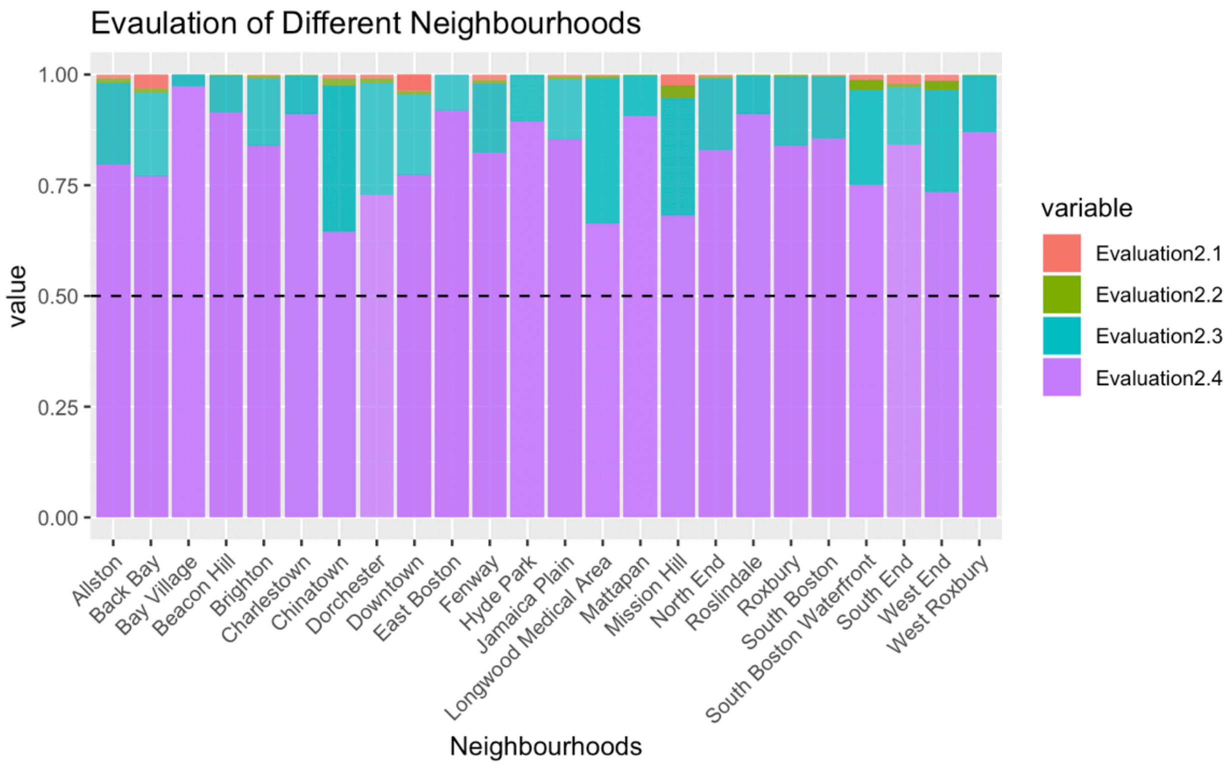
And then we build the Multinomial Model for Airbnb customers' reviews, from Figure 4.10, we get the coefficient of different evaluation levels. More specifically, the variables listed in Figure 4.10 are key factors that influence customers' reviews scores.

Figure 4.10 Result of Multinomial Model

```
polr(formula = ordered(Evaluation2) ~ bathrooms + review_scores_accuracy +
    review_scores_location + neighbourhood_cleansed + review_scores_communication +
    review_scores_cleanliness + review_scores_checkin, data = train_glm)
                                             coef.est coef.se
bathrooms                                    -0.05     0.12
review_scores_accuracy                        1.97     0.14
review_scores_location                        0.50     0.13
neighbourhood_cleansedBack Bay               -0.78     0.39
neighbourhood_cleansedBay Village            -0.28     1.20
neighbourhood_cleansedBeacon Hill            -1.36     0.44
neighbourhood_cleansedBrighton               -0.56     0.45
neighbourhood_cleansedCharlestown            -0.60     0.62
neighbourhood_cleansedChinatown              -2.48     0.49
neighbourhood_cleansedDorchester             -0.39     0.38
neighbourhood_cleansedDowntown               -1.19     0.38
neighbourhood_cleansedEast Boston            -0.89     0.41
neighbourhood_cleansedFenway                 -0.48     0.41
neighbourhood_cleansedHyde Park               0.32     0.85
neighbourhood_cleansedJamaica Plain           0.55     0.46
neighbourhood_cleansedLeather District       13.54     0.00
neighbourhood_cleansedLongwood Medical Area  12.95     0.00
neighbourhood_cleansedMattapan                0.34     0.75
neighbourhood_cleansedMission Hill           -0.78     0.49
neighbourhood_cleansedNorth End              -0.64     0.44
neighbourhood_cleansedRoslindale              0.32     0.75
neighbourhood_cleansedRoxbury                -0.61     0.40
neighbourhood_cleansedSouth Boston           -0.57     0.44
neighbourhood_cleansedSouth Boston Waterfront 14.34    0.00
neighbourhood_cleansedSouth End              -1.12     0.42
neighbourhood_cleansedWest End               -0.83     0.79
neighbourhood_cleansedWest Roxbury           -0.55     0.79
review_scores_communication                   0.83     0.13
review_scores_cleanliness                     1.27     0.10
review_scores_checkin                         0.55     0.13
1|2                                          40.04     1.97
2|3                                          41.20     1.98
3|4                                          46.20     2.11
---
n = 2607, k = 33 (including 3 intercepts)
residual deviance = 1482.2, null deviance is not computed by polr
```

After fitting model, we checked and visualized our model by using test dataset. Figure 4.11 shows the evaluation of different neighborhoods.

Figure4.11



5.Conclusion

5.1 Model Choice

In this study, based on the inspire of latent rating regression model (Wang,2011), I created a unique Airbnb sentiments scores of Airbnb customers' comments and reviews scores of location, cleanliness and description and etc. aspects from Airbnb to get more accuracy results. Based on above analysis, we will choose Latent Linear Regression model to analyze key attributes of customers' reviews and give some suggestions to improve future customers' reviews for Airbnb hosts.

The results display in Figure 4.4, from the above chart we could get the conclusion that only the review scores of different aspects did not include all the sufficient and reliable information. As we can see, combined with sentiment scores, we could get more accuracy predict results.

From the Latent Linear Regression Model summary, the model p-value and predictor's p-value are less than the significance level, so we know we have a statistically significant model. For now, key attributes of customers' reviews are comments, neighborhood, property type, beds, bathrooms, accommodates, review scores accuracy, review scores cleanliness, review scores communication and review scores check in.

5.2 Interpretation

Within the Latent Linear Regression Model, we summary coefficients of main variables in the model shown in Figure 5.1. From Figure 5.1, we find that customer's review scores strongly correlated with review scores location, cleanliness, communication, accuracy and check in. We listed the coefficient as following.

Figure 5.1 Coefficient of main variables

```
room_typePrivate room                        0.220210   0.189615   1.161 0.245607
room_typeShared room                         0.474351   0.544338   0.871 0.383603
review_scores_accuracy                       3.040789   0.142501  21.339  < 2e-16 ***
review_scores_location                       0.874835   0.115066   7.603 4.04e-14 ***
review_scores_communication                  1.774146   0.150683  11.774  < 2e-16 ***
review_scores_cleanliness                    2.291581   0.103451  22.151  < 2e-16 ***
review_scores_checkin                        0.827074   0.154709   5.346 9.79e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.939 on 2556 degrees of freedom
Multiple R-squared:  0.7187,    Adjusted R-squared:  0.7132
F-statistic: 130.6 on 50 and 2556 DF,  p-value: < 2.2e-16
```

5.3 Model Check

I mentioned the details model in above sections.

6.Discussion

6.1 Implications

The findings from the present study provide important theoretical and managerial implications for academicians and practitioners. Specifically, the implications may be beneficial to scholars who seek to evolve research in the areas of customer experience and rating behavior, and to practitioners (particularly in the lodging industry) when developing useful strategies to create and offer the ideal experiences that customers seek.

6.2 Limitations

In the Latent Linear Regression Model, we did not calculate the weights, we just use sentiment scores as a new variable. It may cause bias, though when we did VIF diagnose it has normal results, when we did BP-test, the model shows Heteroscedasticity.

Therefore, the further step for this study we plan to use Weighted Linear Regression Model to eliminate Heteroscedasticity and get more reliable results. As for the other model we studied in this project, they still have space to improve and get more accuracy results.

6.3 Future Direction

The first goal is to combine more dataset of major cities in United States and create an Airbnb sentiment words analysis dataset. When we want to calculate the total customers' review scores of one specific trip, we will combine the scores the customer given with his/her comment's sentiment scores. Based on these reliable information, we could build a more sufficient, reasonable and reliable customer reviews scores rating system. Both Airbnb and hosts in Airbnb could get feedback to improve their service from the scores given by customers.

7.Reference

Wang(2011) Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach

Hsieh, J. K., Hsieh, Y. C., & Tang, Y. C. (2012). Exploring the disseminating behaviors of eWOM marketing: persuasion in online video. Electronic Commerce Research, 12(2), 201-224.

Hsu, M. H., Ju, T. L., Yen, C. H., & Chang, C. M. (2007). Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. International journal of human-computer studies, 65(2), 153-169.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.