How Location Affects Choosing Airbnb

Jung Hwa Yeom

12/10/2020

Abstract

Customers' choices and reviews are important for companies to predict future customers' behaviors. Airbnb is a well-known company and shares its dataset of customers including reviews, location, etc. Based on the data, we can investigate what aspects impact customers to choose their accommodation during travel. In this project, I developed a linear mixed effects regression model to find how locations in Los Angeles affect the customers' reviews. Moreover, I found the crime rate has not impacted the customers to select Airbnb.

Introduction

Airbnb is one of the most popular accommodation companies. It is different from a hotel because customers can use lodgings or homestays from hosts, so customers can fully experience the country's culture. There are many aspects that affect customers to choose Airbnb such as locations, hosts, and previous reviews. In this project, I will mainly focus on locations. To be specific, I will investigate whether locations impact customers' choices or not. If so, I will figure out whether the crime rate will negatively affect or not when the customers select their Airbnb. In this report, I will use multilevel models to analyze Airbnb data. I will limit the location to Los Angeles in California and the data is from the Airbnb website. I will also use crime data during 2010-2019 on Los Angeles Open Data Organization website (Los Angeles Police Department).

Method

Data Cleaning

Airbnb and crime data were downloaded from each website. I excluded columns that were not related to review scores such as urlFurthermore, I removed not applicable rows. To be specific, there was over 30,000 data in the raw Airbnb data and about 11,000 data was included in the final Airbnb data. For the crime data, I cleaned over 2,000,000 data to about 350,000 data.

EDA and Linear Regression Model

Figure 1 shows a histogram of cleansed neighborhood groups. Half of the neighborhoods are in the City of Los Angeles, so we can see that the City of Los Angeles is a popular location for the customers. Figure 2 indicates a linear regression model between review scores location and review scores rating. There are outliers, but we can see it has a positive relationship between location scores and review scores rating.

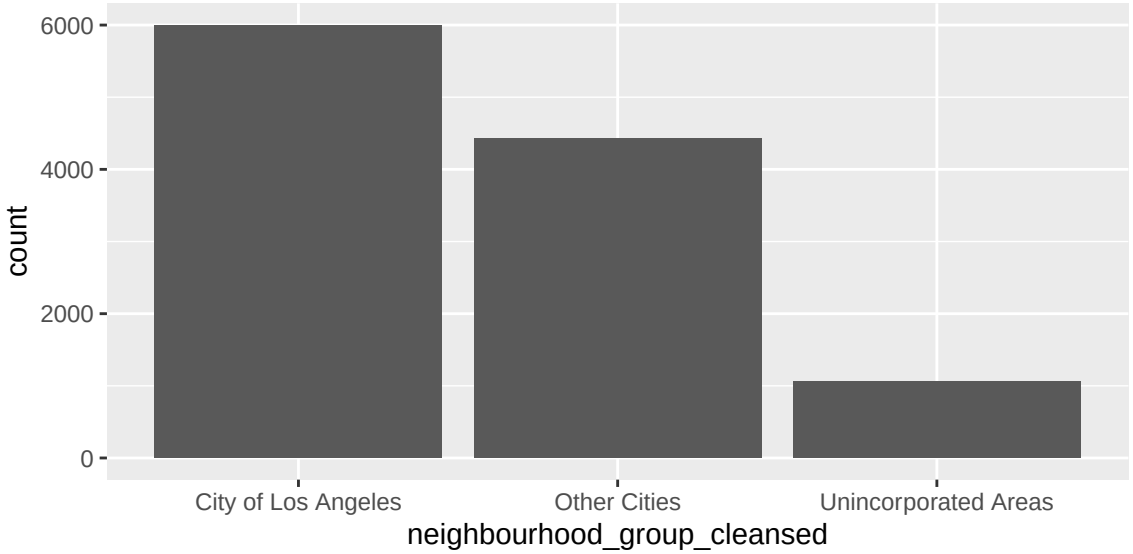

Figure 1: Histogram of cleansed neighborhood group

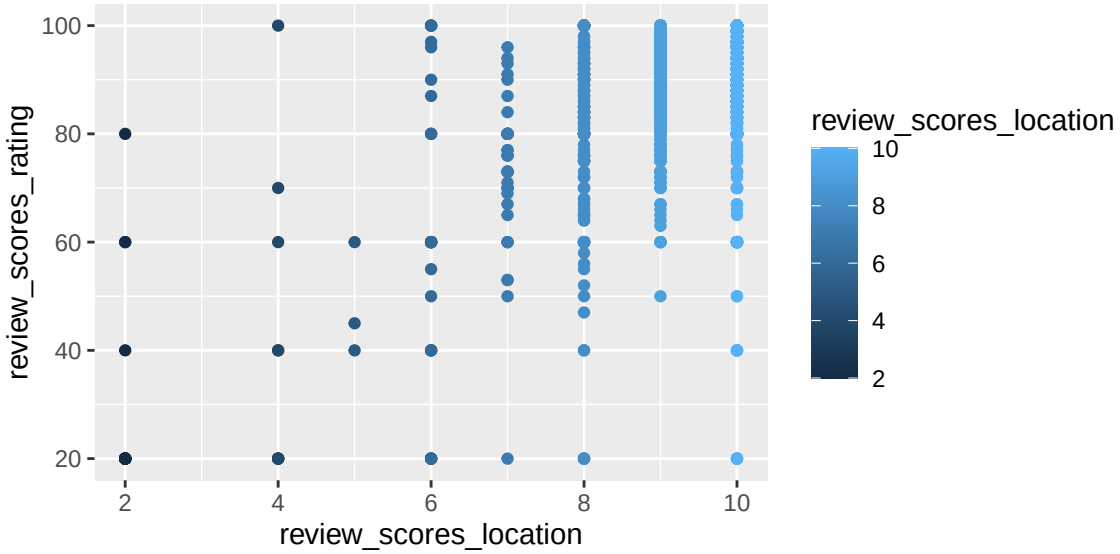

Figure 2: Linear regression model between review scores location and review scores rating.

Results

I used linear mixed effects regression to make model 1 and model 2. Model 1 predictor is neighborhood and model 2 predictor is cleansed neighborhood. Review scores rating include check-in scores, cleanliness scores, accuracy scores, communication scores, location scores, and value scores. I compared the AIC of model 1 and model 2. Model 2 has a smaller AIC and a larger degree of freedom, indicating model 2 is better. To be specific, a smaller AIC means it has a better fit. According to UT Austin, a larger degree of freedom suggests that it has "more power to reject a false null hypothesis and find a significant result" (2015). Therefore, I decided the cleansed neighborhood as a predictor. For the validation, I used residual plots to compare the fitted model and the residual model. We can observe that Figure 3 and Figure 4 have similar plots, meaning model 2 is a proper model for the prediction.

I would like to investigate whether the crime rate is an important aspect for customers to choose Airbnb. I listed 20 locations that have the most highest crime rate and limited the data to Broadway, Hollywood, Sherman, and Wilshire because only those locations are included in the Airbnb dataset. I expected customers to tend to avoid high crime rate locations, but Figure 5 and Figure 6 show that the crime rate did not affect the customers' ratings on locations.

```
## fixed-effect model matrix is rank deficient so dropping 912 columns / coefficients
```

```
## fixed-effect model matrix is rank deficient so dropping 459 columns / coefficients
```

```
##           df      AIC
## model1   770 58391.31
## model2 1293 56222.28
```

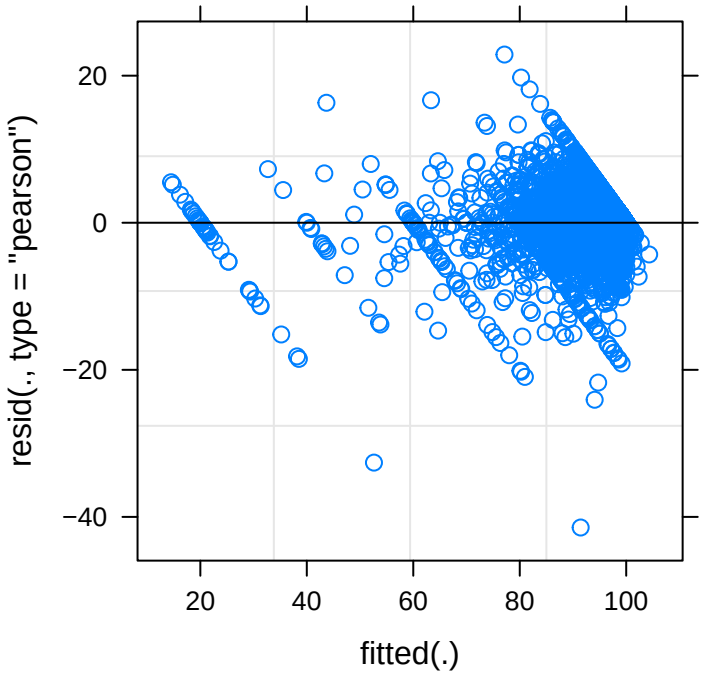

Figure 3: The visualization of model 2

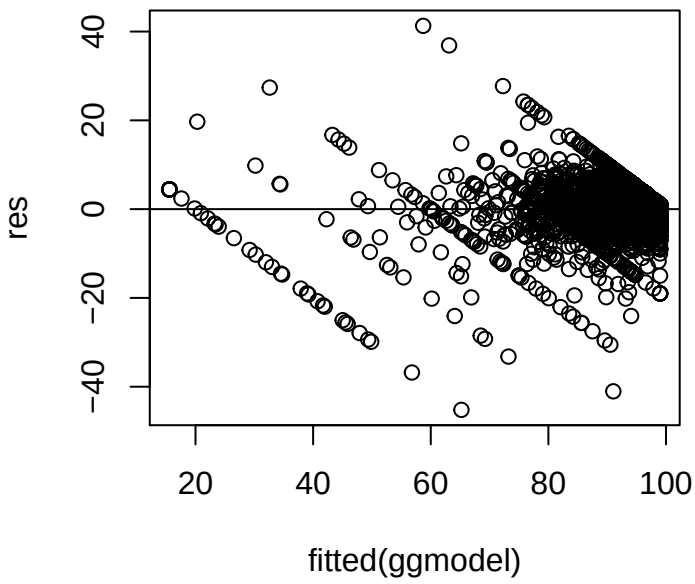

Figure 4: Validation. Residual plots comparing fitted model and residual model

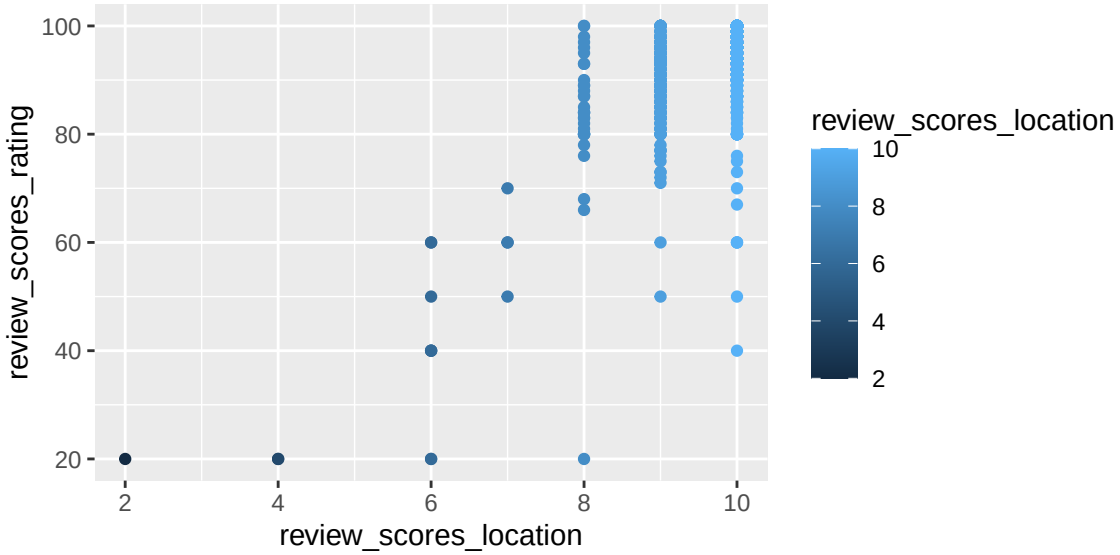

Figure 5: Linear regression model of Hollywood

Figure 6: Linear regression model of Wilshire

Discussion

Model 2 has a smaller AIC and a larger degree of freedom, indicating model 2 is better. To be specific, a smaller AIC means it has a better fit. A larger degree of freedom suggests that it has "more power to reject a false null hypothesis and find a significant result" (UT Austin, 2015). Therefore, I chose model 2 to validate my model. For the validation, we can observe that Figure 3 and Figure 4 have similar plots, meaning model 2 is a proper model for the prediction. I would like to investigate whether the crime rate is an important aspect for customers to choose Airbnb. I expected customers to tend to avoid high crime rate locations. Additionally, I thought customers would not be satisfied with their Airbnb located in Broadway, Hollywood, Sherman, and Wilshire. However, there was no significant relationship between the crime rate and review scores. In future research, I will figure out how customers choose their Airbnb locations because in this report I found the crime rate is not an important aspect. The customers might select locations near to public transportations, tourist attractions, or price. Not limited to locations, I will investigate how other aspects such as a type of Airbnb impact their selections.

Bibliography

Airbnb. insideairbnb.com/get-the-data.html.

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Jared E. Knowles and Carl Frederick (2020). merTools: Tools for Analyzing Mixed Effect Regression Models. R package version 0.5.2. https://CRAN.R-project.org/package=merTools

Los Angeles Police Department. Los Angeles Open Data, data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.

Statistics Online Support, The University of Texas at Austin, sites.utexas.edu/sos/degreesfreedom/.

Van P, Jiang W, Gottardo R, Finak G (2018). "ggcyto: Next-generation open-source visualization software for cytometry." *Bioinformatics*. <URL: https://doi.org/10.1093/bioinformatics/bty441>.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendix

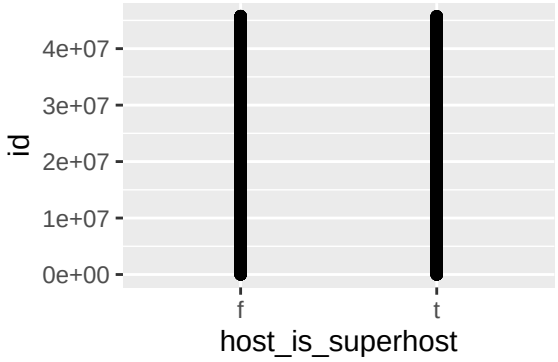

Figure 7: Counting the number of superhost and non-superhost.

```
## `geom_smooth()` using formula 'y ~ x'
```

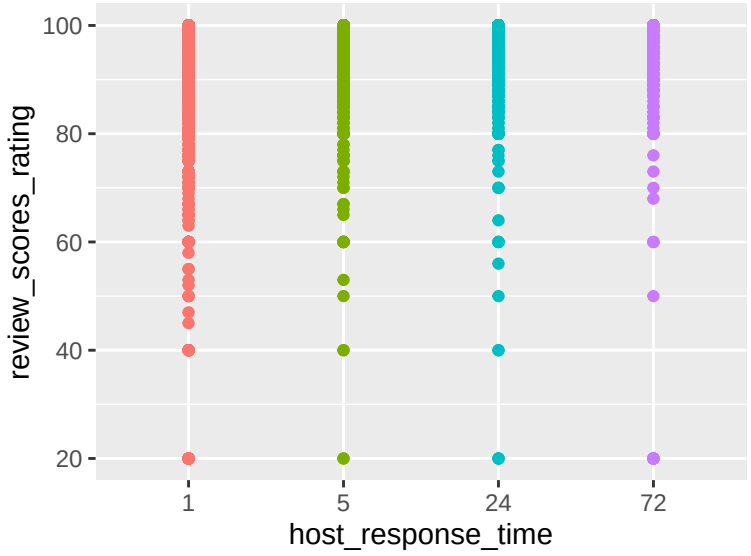

Figure 8: Relationship between host response time and review scores rating

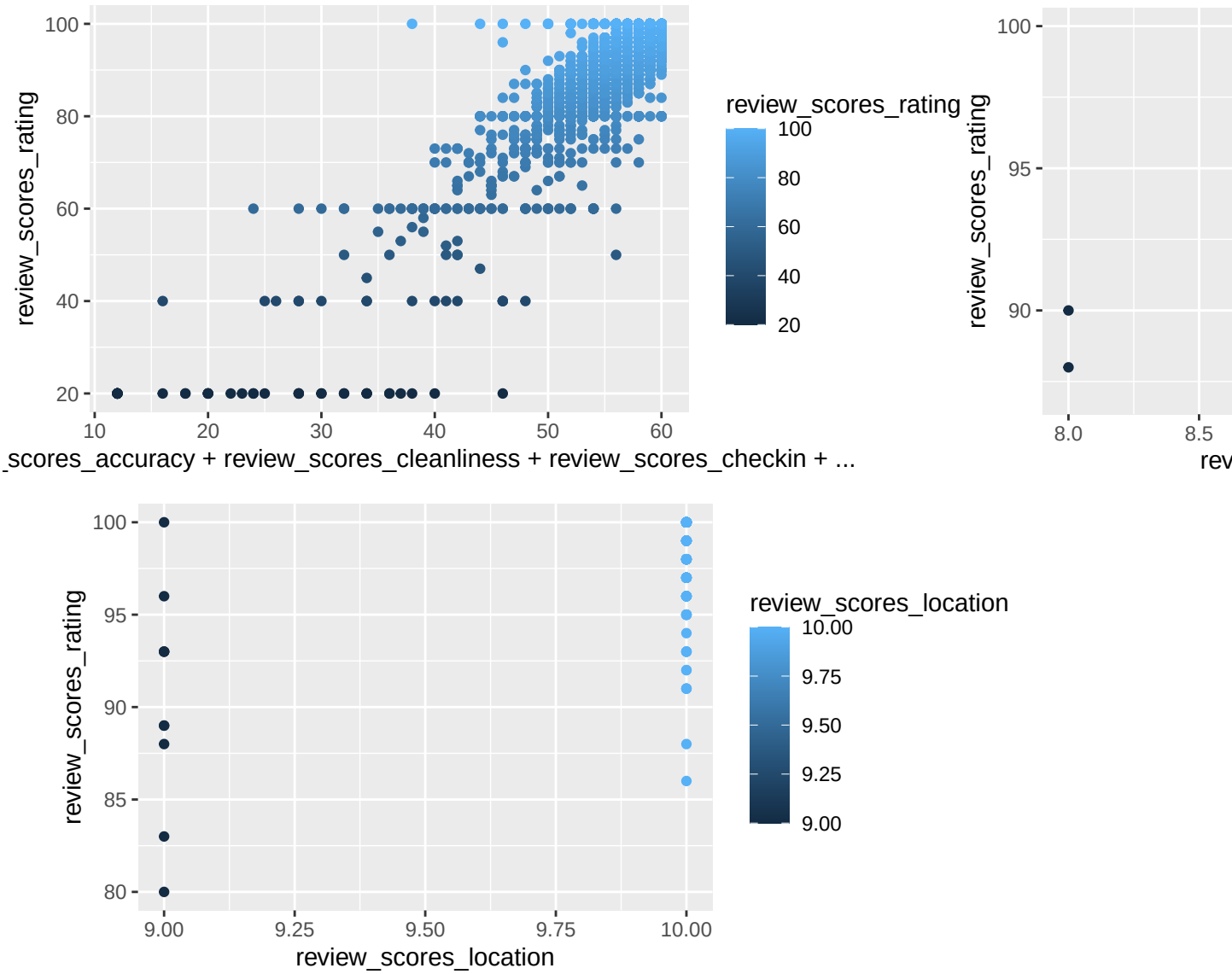

Figure 9: Linear regression model of Sherman