Analysis of Seattle's AirBnB Data

https://github.com/ryan-odea/678-AirBnB-Analysis

Ryan O'Dea
Boston University GRS MA678

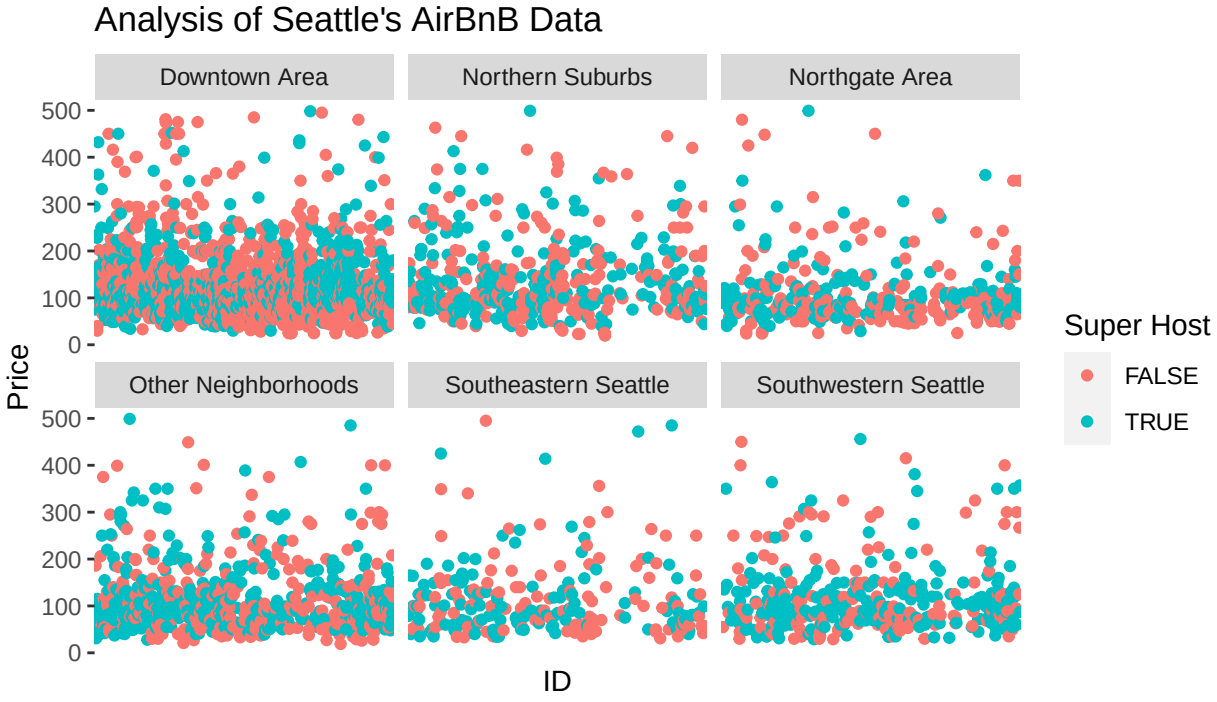

Figure 1: General overview of price vs host ID if AirBnB's in the Seattle Area (King County)

Abstract

Begun in August of 2008, AirBnB is an American vacation rental marketplace. Without owning any of the real estate listings, AirBnB connects hosts, users that are willing to share their home for payment, to guests. With data scraped from InsideAirbnb This project seeks to understand the relationship between how a host will price their home or single room in the Seattle area compared to factors of location, type of room being offered, and if the host is considered a Super Host, someone who AirBnB has designated as providing a "shining example for other hosts." In figure 1, we can observe a very noisy data set when simply plotting of listing ID vs price; however, there are some key takeaways. It appears many of the listings are in the Downtown Area and we generally see a trend around $100 per night in most of the areas. Later in this paper we will explore relationships via grouping to make better sense of our noisy data set.

Basic EDA

Introductory Analysis

With the beginning exploratory data analysis, we can confirm that many rooms are in the Downtown Area (figure 2) and renting an entire home/apartment is generally more expensive than renting a private room with the average entire home/apt with a mean of $136 per night while private rooms are about half with a mean of $68. (Figure 3) We also observe that most listings are full home/apt options with few private rooms available. As conjecture, this could be attributed to the COVID-19 pandemic - less hosts want to invite users into their homes while they are there and would like to have little contact with them.

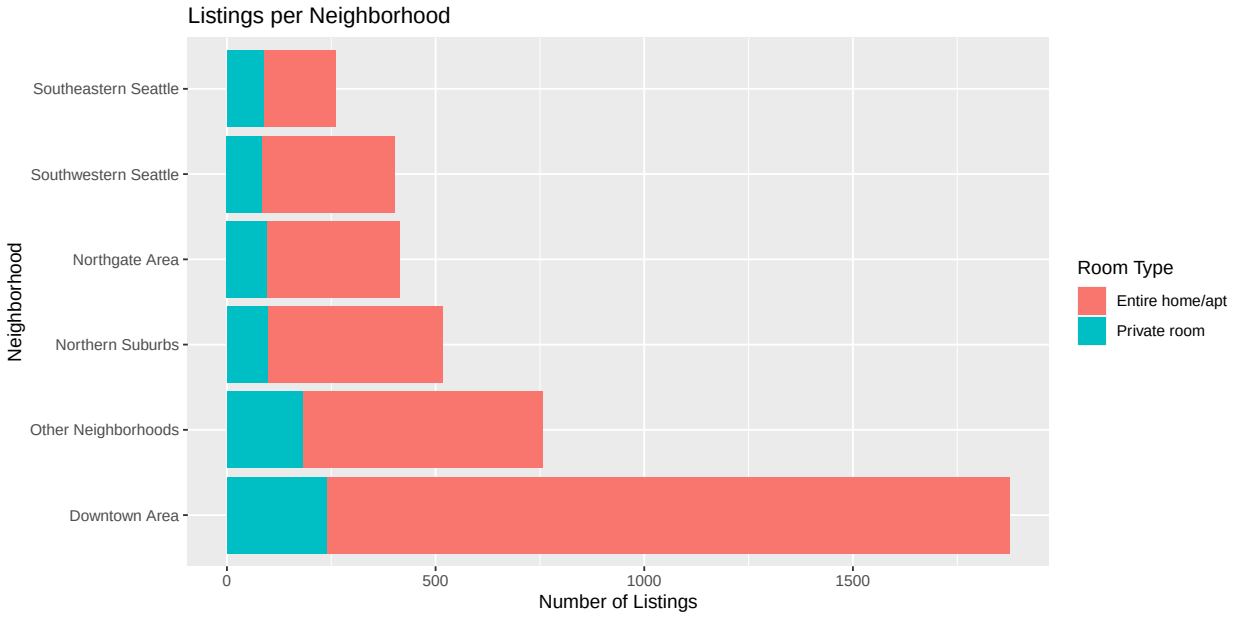

Figure 2: Overview of listings per neighborhood, split into Entire Home/Apt and Private Room

Spatial Analysis EDA

As expected, the location of the home also plays an important role in determining the price per night. Sorting by zipcode, the downtown area is generally more expensive than the other areas in King Country. We also see Lake City (categorized into "Northgate Area") in the north falls into the high average price as an outlier for zip codes in it's similar grouping. The area around SEATAC is the lowest average priced. (See Figure 8)

Exploring Data Relationships

Relation Between Superhost

Comparing points and violin plots, there is no apparent relation between being a Super Host and the price per night of the AirBnB. The violin plots (Figure 4) appear to have mostly even means; the point plot also shows Super Hosts are interspersed with non-Super Hosts when grouped by neighborhood (Figure 5).

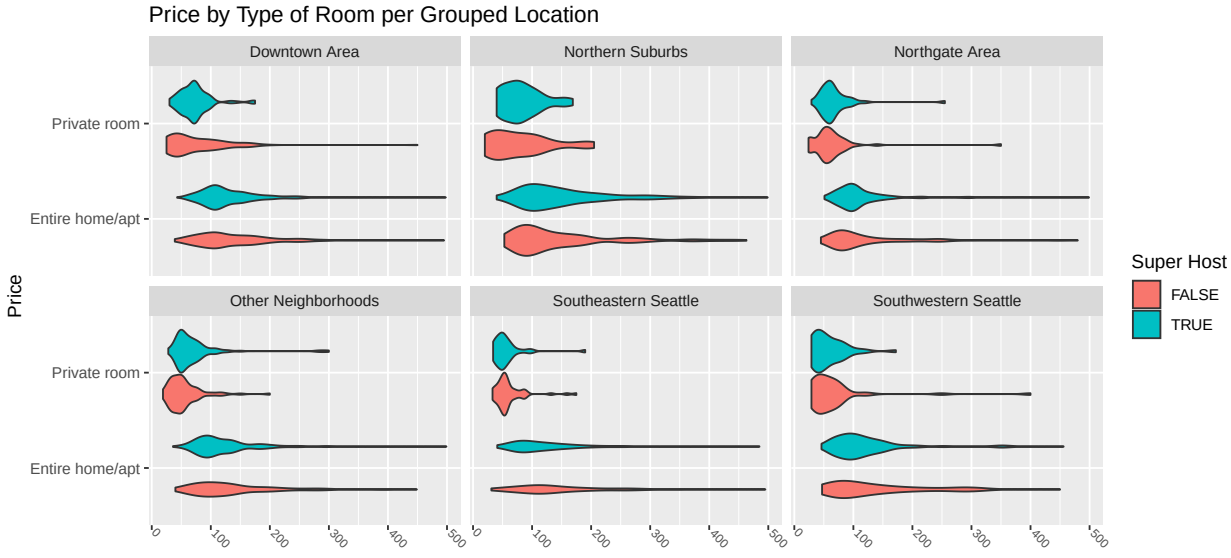

Figure 3: Violin plot overview of SuperHost pricing per neighborhood

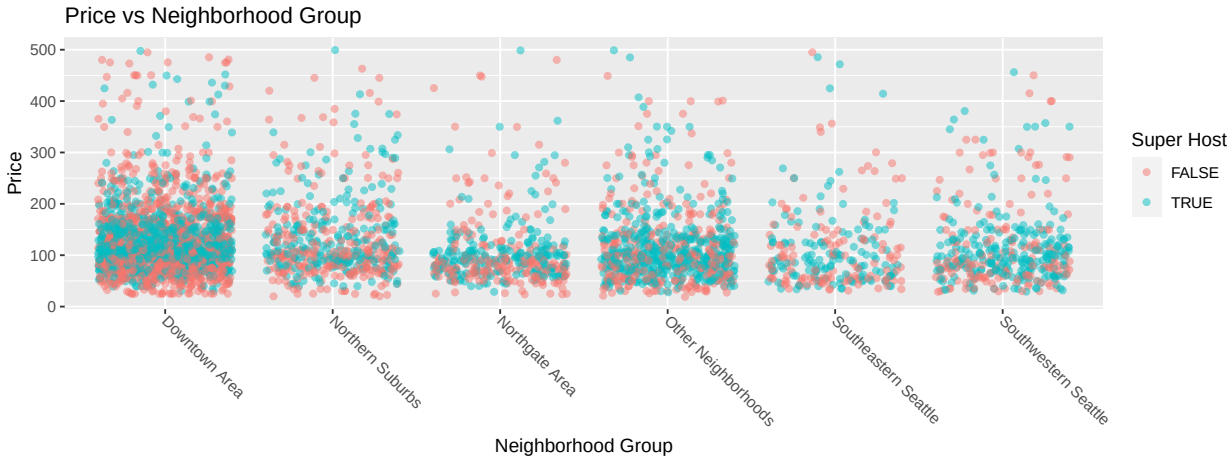

Figure 4: Point plot overview of SuperHost pricing per neighborhood

Model Fitting

In observation of how the model fit would behave, I used the Superhost, neighborhood group, and room type variables to predict the price as an outcome. The AIC and RMSE were quite high, even after fitting a lmer model to the raw price. To lower AIC and RMSE, the outcome was binned into 5 different groups (Figure 9). When fitting a model using these groups as an outcome; however, as observed in the binned plot (figure 2) most of the price data falls into the (0-200] range, therefore developing a model to predict the binned outcome would shed little light to our question. The log of the price was then taken as a variance stabilizing measure and two glm models were tested against a lmer model (seen in table 1).

Predicting the log price (Figure 4) proved better for both our RMSE and the fit AIC, and a glm with call "logprice ~ neighborhood_group + rm_type" was chosen due to the lowest AIC and a simpler fit – Superhost was removed as a predictor because it was not detected as significant in determining the price of a listing. Checking our residuals and R2, we can observe that the residuals fall within normal range and that the model explains about 73% of the variance in our data.

Figure 5: Log price transformation of SuperHost pricing per neighborhood

In an analysis of the coefficients, with a baseline intercept of exp(4.8) - the average price of a full house/apt in the Downtown Area. The other areas are generally less expensive, with the exception of the northern suburbs which are approximately the same, as seen by the summary, additionally private rooms are approximately exp(0.68) less than their whole house/apt equivalents. Bootstrapping (figure 10) was done to test the coefficients and yielded that the true coefficients are close to the approximate. The full table of coefficients can be found in the appendix (table 1). These coefficients show the log relationship between our baseline Downtown full home/apt the other respective groupings. The residuals of our fit fall within normal guidelines per groups (Figures 6, 7)

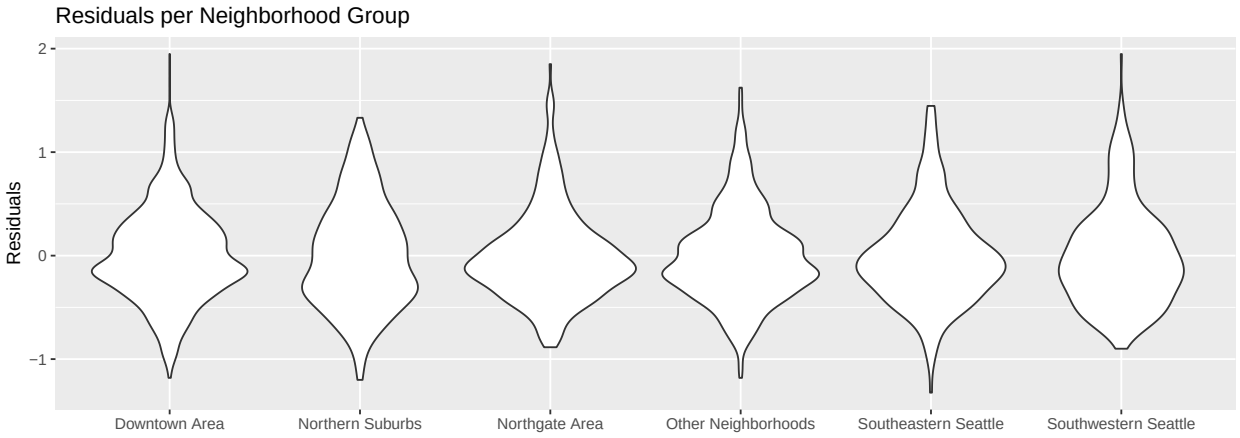

Figure 6: Residuals from selected fit, plotted by neighborhood

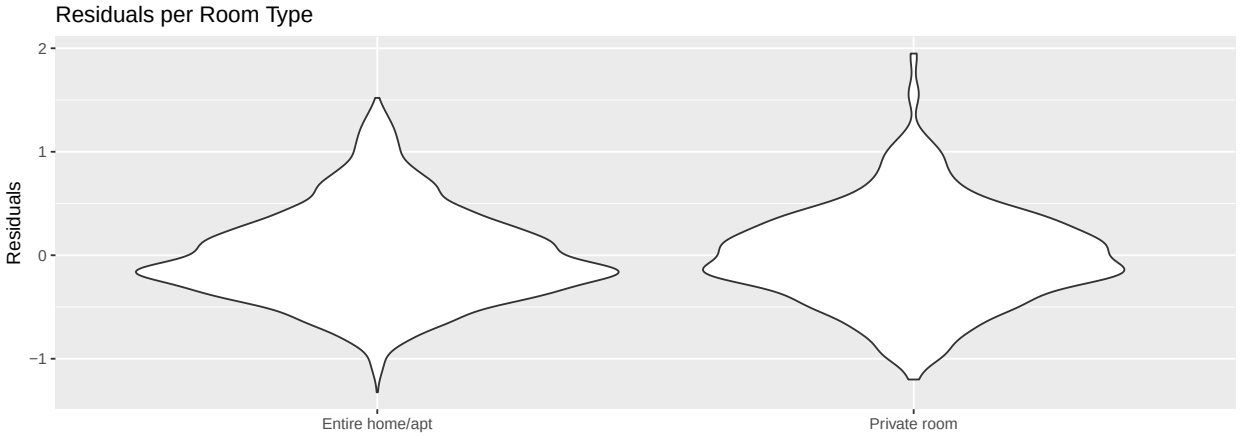

Figure 7: Residuals from selected fit, plotted by listing type

Discussion

Originally, I thought that location and type of room would be the most influential in determining the price of a room. After learning about the Superhost status, I thought this would also allow a host to increase the price of their listing as Superhosts are more likely to be featured in search results and promotional emails. I was surprised to see that the Superhost status had little effect in determining the price - a little further digging explained why.

Superhosts are more likely to get featured and therefore attract more guests, so they often have less need to undercut their pricing to look for more visitors. As a hypothetical, someone is coming into Seattle for a convention. They want to stay in the Downtown area because of the proximity to the event, and they might be coming with a couple friends so splitting the cost for a whole apartment isn't too bad. The first few posts they see within the price range their searching for will be Superhosts because of priority, and their decision for one of the first couple listings is further confirmed by the "SUPERHOST" icon on the first image. Superhosts don't need to compete with price, because of their priority already nets them more on average.

Appendix

Figures

Figure 1

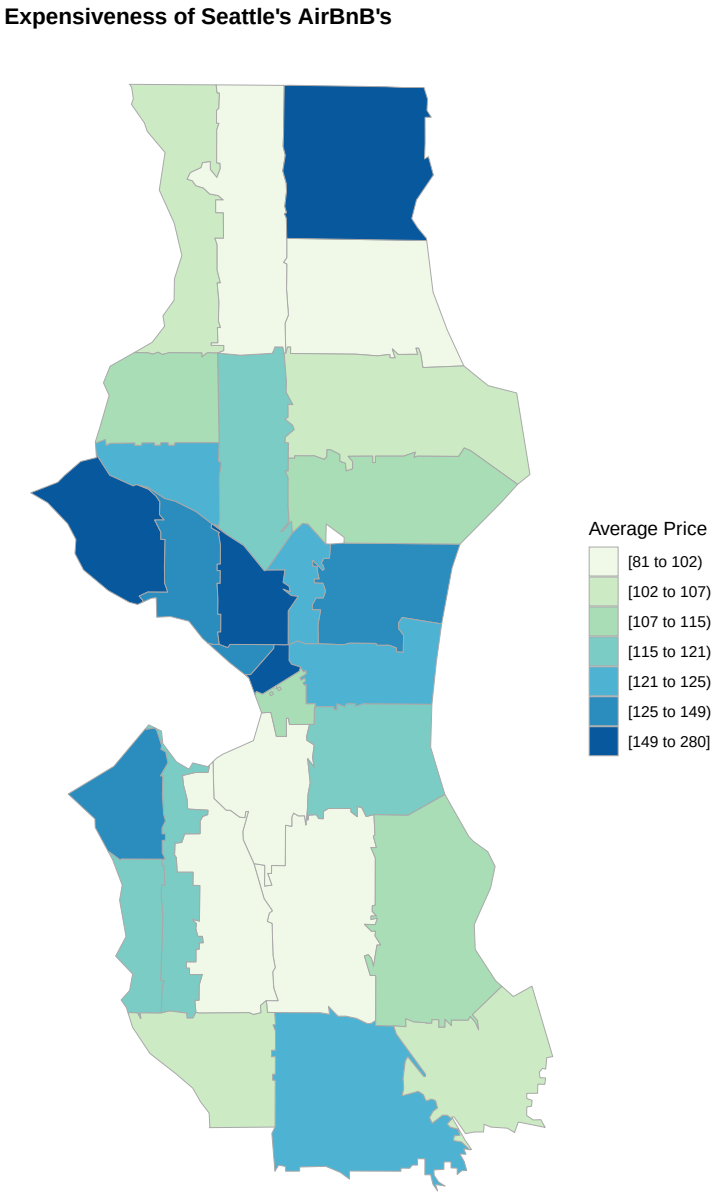

Figure 8: Average price per zipcode in King County

Figure 2

Figure 9: Binned prices per neighborhood

Figure 3

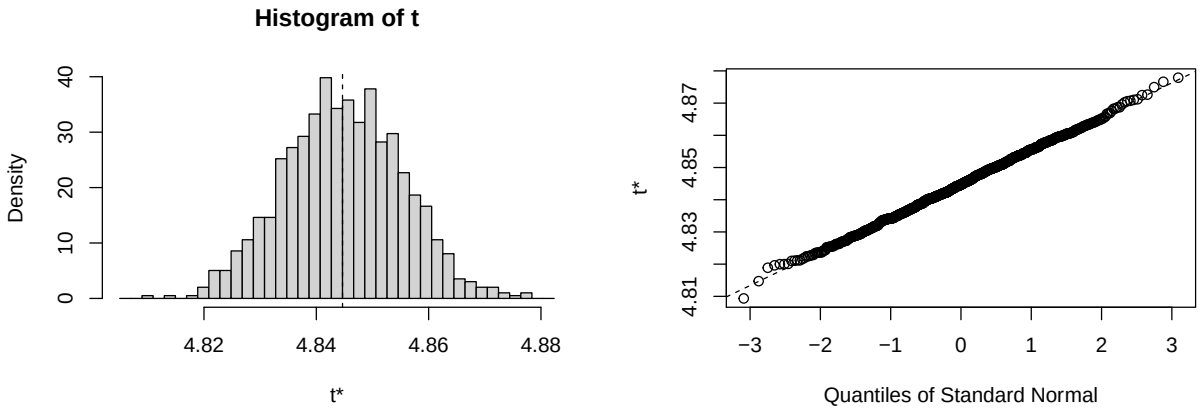

Figure 10: Results of bootstrapping our fit 1000 times

Table 1

Call	Fit AIC	Fit RMSE
Raw Price (GLM)		
price ~ neigh_grp + s_host + rm_type	47766.52	139.69
price ~ neigh_grp + rm_type	47775.94	139.69
Raw Price (LMER)		
price ~ neigh_grp + s_host + rm_type + (1 | neigh_grp)	47731.52	139.69
Log Price (GLM)		
logprice ~ neigh_grp + s_host + rm_type	5412.14	3.70
logprice ~ neigh_grp + rm_type	5410.19	3.70
Log Price (LMER)		
logprice ~ neigh_grp + rm_type + s_host + (1 | neigh_grp)	5463.30	3.70

Table 2

Predictor	Estimate	Std. Error
Intercept	4.844	0.011
Nor Suburb	0.036	0.023
Northgate	-0.153	0.025
Other	-0.079	0.019
Southeast	-0.087	0.030
Southwest	-0.117	0.025
Private Rm	-0.684	0.018

Bibliography

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.

Angelo Canty and Brian Ripley (2020). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25.

Murray Cox. Inside Airbnb, adding data to the debate. http://insideairbnb.com/get-the-data.html

Ari Lamstein (2015). choroplethrZip: Shapefile, metadata and visualization functions for US Zip Code Tabulated Areas (ZCTAs).. https://github.com/arilamstein/choroplethrZip, https://groups.google.com/forum/#!forum/choroplethr.

Thomas Lin Pedersen and David Robinson (2020). gganimate: A Grammar of Animated Graphics. R package version 1.0.7. https://CRAN.R-project.org/package=gganimate

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

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

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