Regression Analysis on Seattle Airbnb Price

Seraphina Shi[1], Tianxin Song[2], Yijie Wu[1]

[1] Statistics, [2] Information

Analysis Motivation

Since 2008, guests and hosts have used Airbnb to travel in a more unique, personalized way. But what influences the price of Airbnb? In this project, we will focus on Airbnb prices in Seattle and examine the relationships between the listing prices and possible factors such as locations, room types, number of beds, number of reviews, review rating, availability, number of guests an Airbnb can host, and so on to find out which factors make the price of an Airbnb higher/lower than others.

Data Manipulation

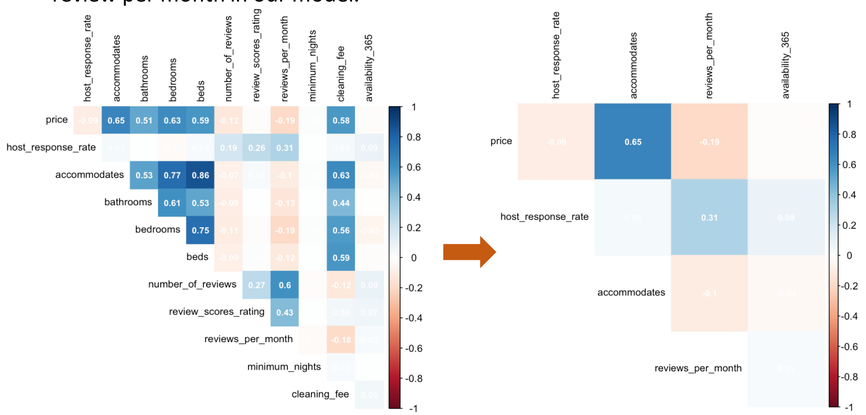
We found Seattle Airbnb open data from Kaggle that contains 3818 homestays. Among 92 variables, we selected 17 variables (13 numerical variables and 4 categorical variables) that we were most interested in and removed 75 variables that were not relevant to our topic (host ID, host picture url and so on).
We conducted data cleaning as follows:

- Removed predictor square-feet (95% observations miss square-feet)
- Removed special characters (such as '$' in prices, '%' in host response rate).
- Replaced missing values with 0 (review number, host response rate etc.)
- Changed categorical variables into factors.
- Transformed left-skewed data into categorical variables based on percentiles as 75% values are between 93 to 100, which has no big difference based on their numeric values(review score rating).
- Derived four categorical predictors (parking, washer, checkin24, pets_allowed) from variable "amenities"

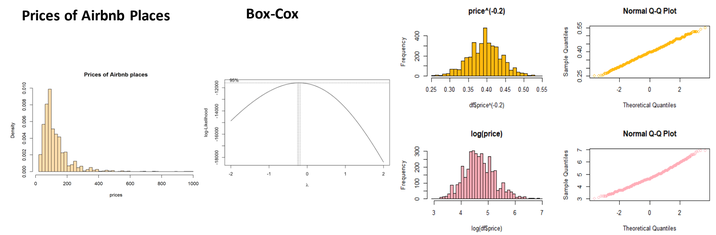
As all our predictors include zero, even though some are right-skewed, we cannot do log-transformation on all predictors.

Variable Selection

Accommodates, bathrooms, bedrooms, beds, and cleaning-fee are multicollinear. Since accommodates is the one mostly correlated with the other three, we decide to keep it, and exclude bathrooms, bedrooms, beds, and cleaning-fee in our model.
Review-per-month and number-of-reviews are collinear, so we only include review per month in our model.



Data Transformation



Because Airbnb prices are strictly positive and right-skewed, we used Box-Cox to find an appropriate transformation which is price^-0.2. But this transformation is hard to interpret. So we also tried log(price) and compared price^-0.2 to log(price), price^-0.2 does a better job to transform price to be normally distributed.

The opt value $\hat{\lambda}$ is -0.2, so we transform y to $\mathbf{y}^{-0.2}$ for fitting the regression model.

Since we have too many categorical variables, we do not do LASSO or Ridge transformations.

Modeling

Firstly, we build a multiple linear regression with all predictors and use stepwise model selection using AIC criterion. We tried both forward selection from empty model and backward selection from full model, and they end with same model as shown below:

- ***Fit1:***

price^(-0.2) ~
accommodates + room type + neighbourhood group + reviews per month + host response rate + review rating + availability 365 + cancellation policy + checkin24 + instant bookable

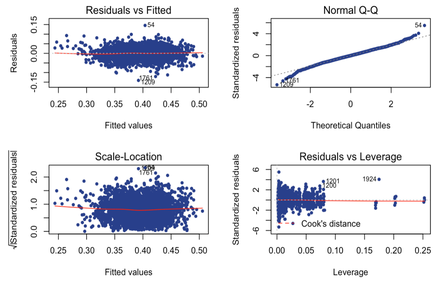
Next, we decided to test if there is any significant interaction relationship between any variables. We added all pairwise interaction and use stepwise model selection using AIC criterion. Again, both forward selection from empty model and backward selection from full model give us the same model.

- ***Fit2:***

price^(-0.2) ~
accommodates + room type + neighbourhood group + reviews per month + host response rate + review rating + availability 365 + cancellation policy + checkin24 + instant bookable
+ accommodates:reviews_per_month + room_type:host_response_rate + accommodates:room_type + reviews_per_month:review_rating + accommodates:availability 365 + room_type:cancellation_policy + review_rating:cancellation_policy + accommodates:cancellation_policy +availability_365:cancellation policy + reviews per month:availability_365+ room_type:availability_365 + reviews_per_month:checkin24 + accommodates:instant_bookable + checkin24:instant_bookable + reviews_per_month:instant_bookable + cancellation_policy:instant_bookable + availability 365:checkin24 + host response rate:availability 365

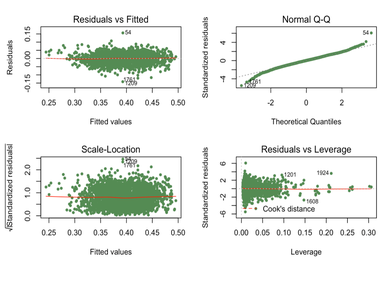
Model Comparison

Fit 1 (Without Interaction)



Residual standard Error: 0.02653
Adjusted R square: 0.6437

Fit 2 (With Interactions)



Residual standard Error: 0.02587
Adjusted R square: 0.6526

The residual normality with zero-mean and constant variance assumptions are not violated for both models.
As both models are significant predicting the airbnb prices in Seattle, and there are no big difference between these two models' statistics, we decide to choose ***Fit1*** as our final model.

Conclusion

$$\frac{1}{\text{price}^{0.2}} = 0.3996 - 0.0010 * \text{accommodates} + 0.0028 * \text{reviews\_per\_month} + 0.0148 * \text{host\_response\_rate} - 0.0001 * \text{availability\_365}$$

$$+\begin{cases} 0 & \text{if not instant\_bookablet} \\ 0.0025 & \text{if instant\_bookablet} \end{cases}$$

$$+\begin{cases} 0 & \text{if cannot check in 24h} \\ 0.0042 & \text{if can check in 24h} \end{cases}$$

$$+\begin{cases} 0 & \text{if cancellation\_policy is flexible} \\ -0.0008 & \text{if cancellation\_policy is moderate} \\ -0.0041 & \text{if cancellation\_policy is strict} \end{cases}$$

$$+\begin{cases} 0 & \text{if room\_type is Entire home/apt} \\ 0.0318 & \text{if room\_type is Private room} \\ 0.0726 & \text{if room\_type is Shared room} \end{cases}$$

$$+\begin{cases} 0 & \text{if review\_rating is bad} \\ 0.0003 & \text{if review\_rating is fair} \\ -0.0048 & \text{if review\_rating is good} \\ -0.0089 & \text{if review\_rating is excellent} \end{cases}$$

$$+\begin{cases} 0 & \text{if neighbourhood is Alki} \\ 0.0247 & \text{if neighbourhood is Arbor Heights} \\ 0.0322 & \text{if neighbourhood is Bitter Lake} \\ 0.0139 & \text{if neighbourhood is Briarcliff} \\ 0.0325 & \text{if neighbourhood is Broadview} \\ 0.0043 & \text{if neighbourhood is Broadway} \\ -0.0038 & \text{if neighbourhood is Central Business District} \\ 0.0249 & \text{if neighbourhood is Columbia City} \\ 0.0281 & \text{if neighbourhood is Crown Hill} \\ 0.0273 & \text{if neighbourhood is Dunlap} \\ 0.0253 & \text{if neighbourhood is Greenwood} \\ 0.02467 & \text{if neighbourhood is Haller Lake} \\ -0.0115 & \text{if neighbourhood is Industrial District} \\ -0.0042 & \text{if neighbourhood is International District} \\ -0.0043 & \text{if neighbourhood is Lower Queen Anne} \\ 0.0294 & \text{if neighbourhood is Maple Leaf} \\ -0.0019 & \text{if neighbourhood is Montlake} \\ 0.0204 & \text{if neighbourhood is North College Park} \\ 0.0332 & \text{if neighbourhood is North Delridge} \\ -0.0027 & \text{if neighbourhood is Pike-Market} \\ 0.0388 & \text{if neighbourhood is Pinehurst} \\ -0.0220 & \text{if neighbourhood is Pioneer Square} \\ -0.0081 & \text{if neighbourhood is Portage Bay} \\ 0.0323 & \text{if neighbourhood is Rainier Beach} \\ 0.0102 & \text{if neighbourhood is Other} \end{cases}$$

- We analyzed relationships between many variables and listing prices. The variables above are the ones that can significantly influence the price. This model also tells us how these variables affect the prices. For example, for 1 person increase in accommodates, the price^-0.2 will decrease by 0.001 when other variables holding constant, which means the price will increase as accommodates increase. Most of the coefficients do align with our intuitions.
- Our project's limitation is that we include several categorical variables in the model and each categorical variable contains many categories, which limits the number of methods that we can use in our analysis. The future analysis could include fewer variables and try other methods of transformation and other models, such as LASSO and Random Forest.

Reference:

Airbnb. (2018, June 26). Seattle Airbnb Open Data. Retrieved from https://www.kaggle.com/airbnb/seattle#listings.csv