Airbnb Data Analysis - Santa Clara County

Hung Tong, Jelena Segan, Joshua Fontes, Nivedha Murugesan

December 5, 2018

I. Introduction

Airbnb was founded in San Francisco in 2008 as "AirBed & Breakfast", an online marketplace enabling people to list and rent a short-term lodging. It is the world's biggest accommodation-sharing site, as the company currently lists around 800,000 properties in 65,000 cities across 190 different countries.

The company has a project named "Inside Airbnb", an independent set of data that allows us to explore how the platform is being used around the world. In this analysis, we focus on listings in Santa Clara County which is known for its diverse neighborhoods. The county's Airbnb data give us an opportunity to register similarities and differences between neighborhoods.

The Santa Clara County dataset has 5,854 observations. Each observation is a full description of a booking, using 15 variables in total, including location information, review information, availability over a year and much more.

Using this dataset, we were able to build a regression model which takes a few features of the booking information as the input and returns the price (for one night) as the output. After several methods of variables selection to optimize our regression model, we finalized a model in which the chosen predictors have a significant influence on the price.

At first glance of the dataset, we noticed that it contains many irrelevant variables that would not have a significant effect on our response variable (the price), such as ID or host name. However, a lot of variables would very likely influence the price; we did further analysis on the dataset using these variables to select the best features for our model. The following list summarizes the variables and each definition:

Response Variable:

- Price: The price (in USD) for a night stay

Continuous Variables:

- longitude: The angular distance of a place east or west (degrees, minutes, seconds).
- latitude: The angular distance of a place north or south (degrees, minutes, seconds).
- minimum_nights: The minimum number of nights the host is willing to rent their property in a row

- number_of_reviews: The number of reviews that a listing has received
- reviews_per_month: The number of reviews that a listing has received in a month
- availability_365: The number of days for which a particular host is available in a year

Categorical Variables:

- neighborhood: A city in Santa Clara County for which the survey is carried out
- room_type: One of Entire home/apt, Private room, or Shared room
- last_review: Date when an Airbnb listing received its last review

II. Basic Data Cleaning

We noticed that the variable *reviews_per_month* has 1065 missing values, e.g. NA, in its record, and any Airbnb listing with missing *reviews_per_month* also has *number_of_reviews* equal to zero. Our assumption is that these are the new businesses with which renters are not familiar, causing them to have very few reviews. Specifically, 1059 observations have zero reviews, five listings have one review, and one listing has two reviews. Certainly, it is impossible for an observation to have a total number of reviews of zero concurrently with non-zero monthly reviews. For that reason, we decided to convert the missing values in *reviews_per_month* to zero.

Initially, we were reluctant to include the variable *last_review* because we thought that the last time at which an Airbnb listing received a review would have a weak connection to the listings price. However, looking deeper, one inference we could make from *last_review* is that if the last review was recorded not long ago, chances are that the corresponding Airbnb business is still active. This could mean the Airbnb property is a well-established one that has been doing business for a long time, or a new, competitive one that is attracting renters with many perks, possibly including price. We decided that perhaps *last_review* might in fact be relevant and wanted to examine further.

Although *last_review* includes year, month and day, we chose to represent it using just the year to avoid having many levels or creating hierarchy within a categorical variable, e.g. every year would have 12 months with about 30 different days. At the same time, we also observed 1065 empty records and coincidentally, these happen to have missing *reviews_per_month* values and approximately zero *number_of_reviews*. This makes our above assumption regarding new businesses seem more plausible. We decided to call these 1065 empty records in *last_review* "Unknown".

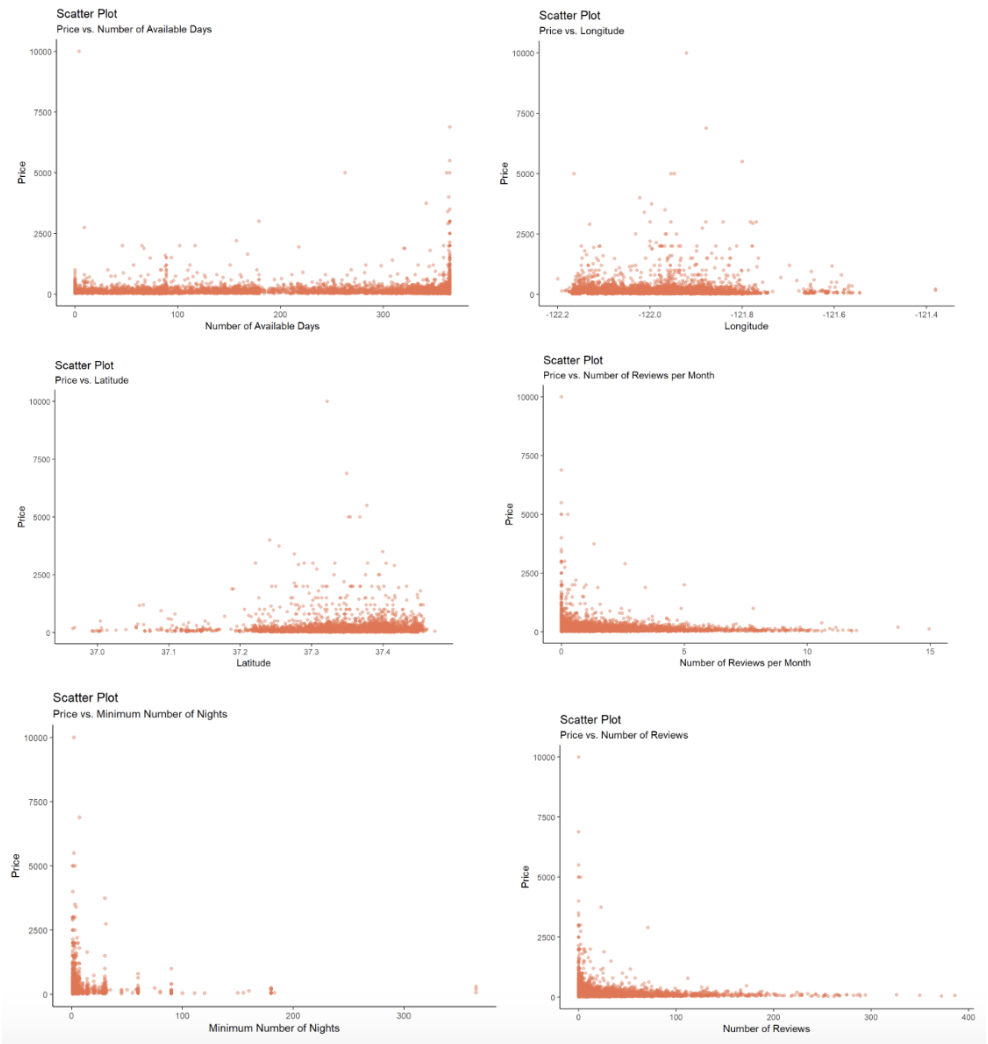
III. Main Objective

Ultimately, our main goal in this project was to learn about the relationship between the rental price and other variables of an Airbnb property in the Santa Clara County. Understanding if such a relationship exists could help customers estimate their base rental price based on the rental factors that they effectively control. Furthermore, it could potentially

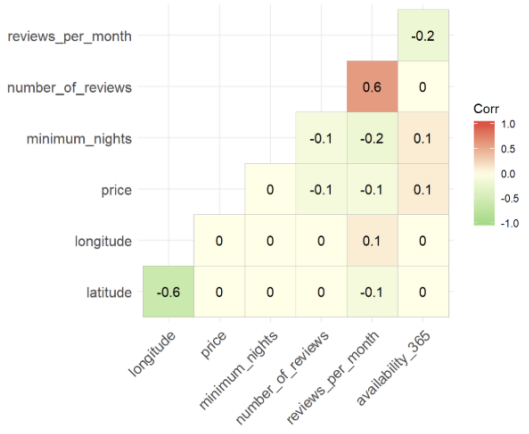
help Airbnb regulate their prices so they could compete with similar accommodation-based businesses, such as Expedia, Booking.com, etc. Lastly, it could help Airbnb hosts maximize profits by offering more accommodation perks.

IV. Exploratory Data Analysis

The following plots are scatterplots of the response variable, price, against each of the six continuous predictor variables.



From the scatterplots, we notice that there is not much of an obvious linear relationship between *price* and each of the individual continuous predictor variables. We can also keep in mind that a transformation on one or more variables might be useful, since many of the observations in each scatterplot are bunched up in a corner. Naturally, our next step would be to consider a multiple regression linear model, since the response variable might be better explained using many predictors simultaneously. Thus, we should take a look at the correlation between the continuous predictors.



Based on the correlation matrix above, we notice that the only two pairs of variables that are relatively highly correlated are (1) *longitude* and *latitude*, and (2) *number_of_reviews* and *reviews_per_month*. As high correlations may signify multicollinearity, we expect that our final model(s) may not include the two variables from either of these two pairs. However, further analysis on variance inflation factor is required to verify our expectation. We keep in mind that we want to look further into this later on.

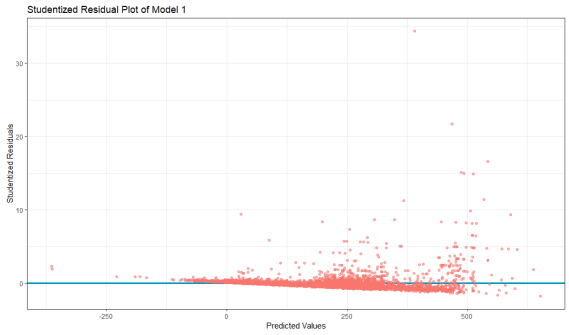
V. Model Fitting & Transformations

The goal of the model building process was to create a simple model that best explains the response variable, the price of an Airbnb property in the Santa Clara County, using the necessary predictor variables. This entails creating a model that satisfies linear regression assumptions and displays a high R^2 value. In this process, many models were explored using different transformations on the response variable and a different combination of predictor variables. To check that the assumptions for linear regression were satisfied for each model, studentized residual plots were observed to determine if the residuals were normally distributed and displayed constant variance. Each model was compared to each other to determine which model fit the data best.

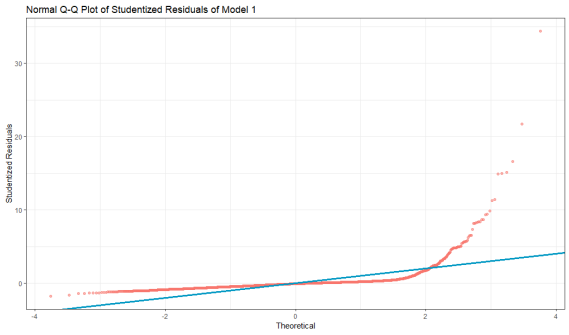
The first model created used the predictor variables *latitude*, *longitude*, *number_of_reviews*, *reviews_per_month*, *availability_365*, *minimum_nights*, *room_type*, *neighborhood*, and *last_review*. For categorical variables, alphabetical order was used to determine the reference.

Model (1): price ~ room_type + last_review + neighbourhood + reviews_per_month + availability_365 + minimum_nights + longitude + latitude + number_of_reviews

The plots below show that the assumptions of linear regression are not satisfied in Model 1. The residual plot (Figure 1) depicts an undesired pattern. The residuals increase in magnitude as the predicted price values increase, forming an arrowhead shape. This means that the residuals of the model do not have constant variance. The qqplot (Figure 2) shows that the quantiles of the residuals do not closely follow the plotted line. They form a C-shape



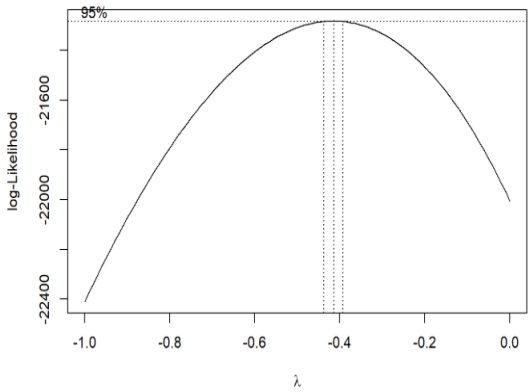
(a) Figure 1



(b) Figure 2

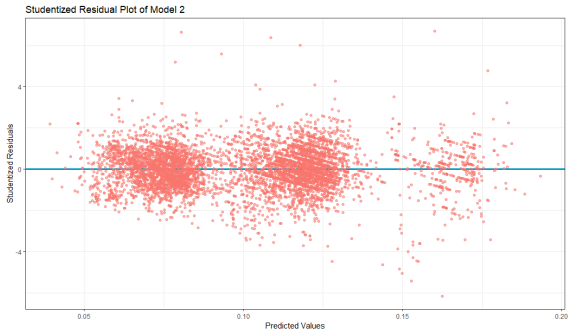
with extreme steepness at the right-end of the plot. This implies that the residuals form a right-skewed distribution and are not approximately normally distributed.

To correct these violated assumptions, a variance-stabilizing transformation was applied to the response variable of this model. Using the Box-Cox Method, the maximum-likelihood estimate for the power transformation of the response variable is approximately -0.415, which is depicted in the Box-Cox normality plot shown below.

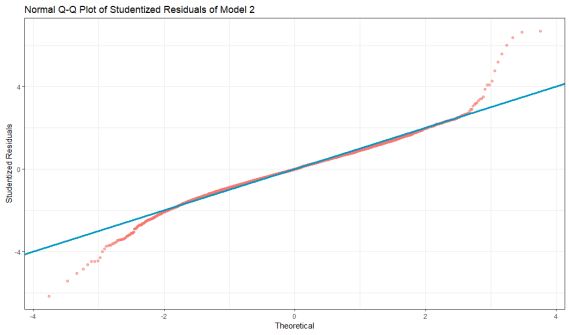


We had tried several Box-Cox transformations using different lambdas that are near the optimal lambda, -0.415. The residual plots and R^2 value of each model were compared to each other to determine the best model.

This model used a reciprocal square root transformation on the response variable. We chose this model because we can see the residuals display roughly constant variance (Figure 3) and are approximately normally distributed (Figure 4).



(c) Figure 3



(d) Figure 4

This model also has the highest R^2 value of any other model created using this process. Another reason for choosing this model is because a reciprocal square root transformation is easier to interpret and understand than using the exact estimate obtained from the Box-Cox Method. We did not perform any transformations on any of the individual predictors. The residual plots for each simple linear model of each predictor individually fitted with the response did not show any functional pattern between the residuals and the predicted values of each model.

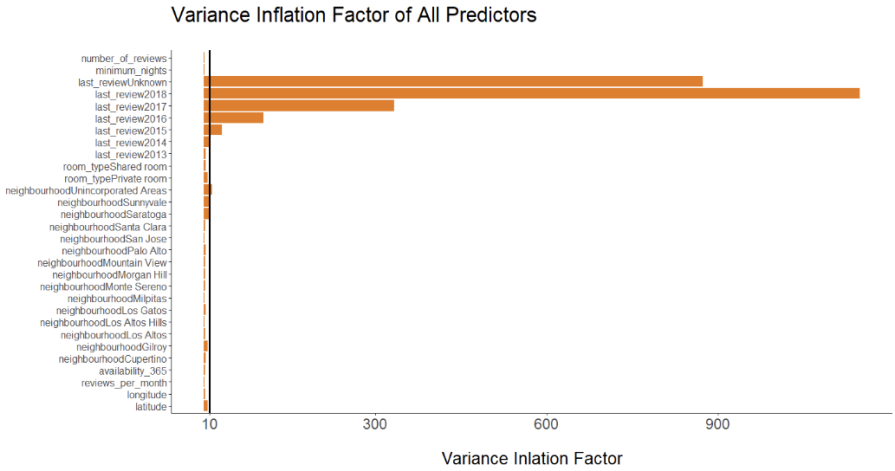
At this point, our model is

Model (2): $\text{price}^{-0.5} \sim \text{room\_type} + \text{last\_review} + \text{neighbourhood} + \text{reviews\_per\_month} + \text{availability\_365} + \text{minimum\_nights} + \text{longitude} + \text{latitude} + \text{number\_of\_reviews}$

VI. Multicollinearity

Multicollinearity is an issue in many regression problems. When multicollinearity is present, any small change in the model can easily alter the coefficient estimates, and increase their standard errors. This means lower precision and less statistical significance is needed to explain the response-predictor relationships.

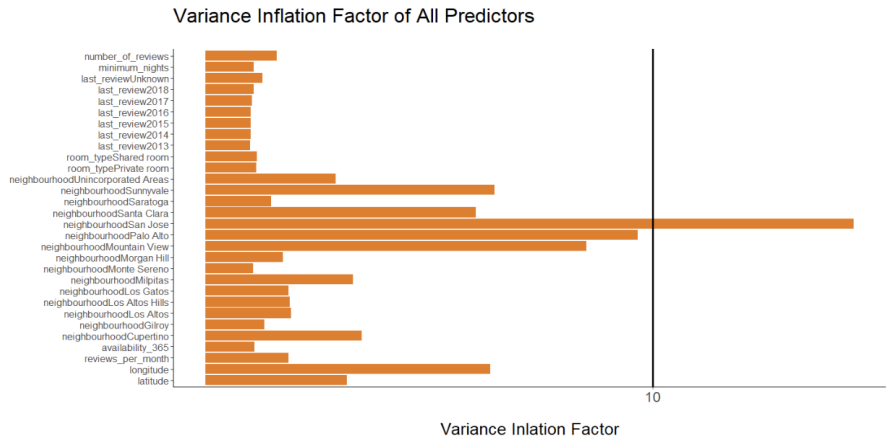
In this section, we are concerned about detecting multicollinearity in our model. As mentioned earlier, we have some high pairwise correlations among all predictors, namely (1) *longitude* and *latitude*, and (2) *number_of_reviews* and *reviews_per_month*. This could be a warning sign of the existence of multicollinearity. To verify that, we calculated each predictors variance inflation factor (VIF), using a cutoff of 10 to decide whether a predictor may be the source of multicollinearity. We visualized the results as below



Notice that the dummy variable *last_review* has exceptionally high VIFs, while most of other variables, even continuous variables, have VIFs less than 10. We suspected the cause of this phenomenon is that the dummy coding for *last_review* chose a level that only accounts for a small portion of the whole categorical variable to be the reference. Lets take a look at the frequency table of all levels within *last_review*.

2011	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Unknown**
1	2	8	31	105	352	4290	1065

By default, R relies on alphabetical order to choose the reference, e.g. 2011 in our case. However, as there is only 1 observation that falls into 2011 category, it is understandable to observe high VIFs among the *last_review* dummy variables. To fix this issue, we forced R to use level 2018 as the reference, since it appears the most within *last_review*. After this change, we notice the change:



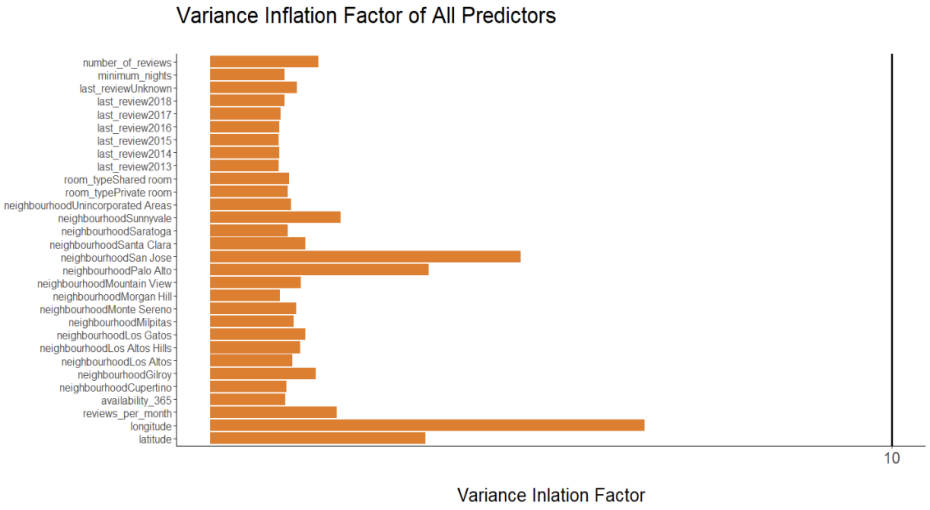
Therefore, we see that recoding dummy variables for *last_review* truly reduced their VIFs. However, we still observed a VIF greater than 10 in dummy 'San Jose' of variable neighbourhood. As neighbourhood is a categorical variable, we expected the same phenomenon that happened to *last_review* before. Here is its frequency table

Campbell	Cupertino	Gilroy	Los Altos	Los Altos Hills	Los Gatos	Milpitas
106	265	13	81	70	89	188

Monte Sereno	Morgan Hill	Mountain View	Palo Alto	San Jose	Santa Clara
7	42	752	740	2112	553

Saratoga	Sunnyvale	Unincorporated Areas
48	586	202

Initially, R chose 'Campbell' to be the reference. We also tried forcing R to to use level 'San Jose', the one that appears the most within *neighbourhood*, as the reference. In the end, here is what we got

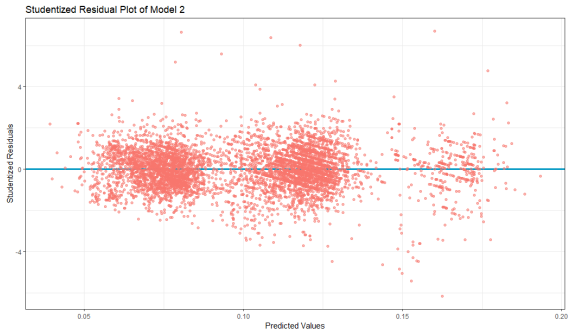


At this point, we had successfully resolved multicollinearity. The good thing is that we did not have to remove any variable. Therefore, we ensured we have not missed any useful information from the predictors. Furthermore, without multicollinearity, the coefficient estimates will be less sensitive, making the result of variable selection more reliable.

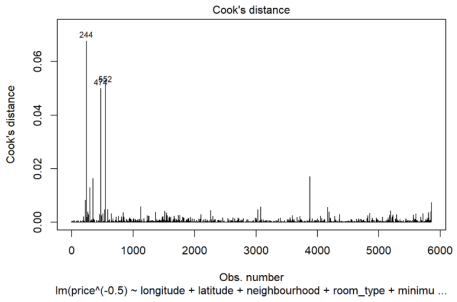
VII. Influential Observations

The coefficient estimates can change a lot if influential observations exist in the dataset. Looking back at our residual plot of Model 2 (Figure 5), we actually had some outliers, in the sense that they have studentized residuals greater than 3. They could potentially be the sources of influence in our model.

To verify, we consulted Cooks distance plot (Figure 6) considering any value of Cooks distance greater than 1 to be influential. However, since even the highest Cooks distance is just slightly greater than 0.06, we were assured that not any observation in our dataset is particularly influential.



(e) Figure 5



(f) Figure 6

VIII. Variable Selection

The next step in our process was variable selection. For variable selection, we utilized the "OLSRR" package. This package contains methods for forward, backward, and stepwise selection based on p-values, which definitely was more convenient than manual variable selection. To start, we used forward selection on a linear model fitting all the predictor variables and the response variable with the reciprocal square root transformation. Based on a p-value of 0.1, this was the result:

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	room_type	0.5394	0.5392	927.3995	-26442.2880	0.0253
2	last_review	0.5718	0.5711	452.8167	-26855.1278	0.0244
3	neighbourhood	0.5875	0.5858	222.6133	-27044.9404	0.0240
4	reviews_per_month	0.5945	0.5928	122.0200	-27142.7516	0.0238
5	availability_365	0.5987	0.5969	61.9126	-27202.0123	0.0236
6	minimum_nights	0.6031	0.6013	-0.7139	-27264.4457	0.0235
7	longitude	0.6042	0.6023	-14.1987	-27277.9902	0.0235

Out of the the nine predictor variables, seven variables were added in the order above. We see that the first predictor to be added is *room_type*. As each of the remaining six predictors are added, the R^2 value increases and the $C(p)$ value decreases until the final model with all seven predictors. The RMSE (root mean square error) also decreases with the addition of each predictor, but we notice there is not a significantly large change between the RMSE values of each subsequent model.

We decided to utilize backward selection to see if we arrive at a different model. Using the "OLSRR" package, we performed automatic backward selection on the full linear model again. Given a p-value of 0.1, this was the result:

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	latitude	0.6042	0.6023	-12.9567	-27276.7522	0.0235
2	number_of_reviews	0.6042	0.6023	-14.1987	-27277.9902	0.0235

We see that the final model we acquire from backward selection is the same model as the one acquired from forward selection. We found it interesting that the first predictor variable to be dropped was *latitude*, but the variable *longitude* was included in the final model. We were unsure of the exact reason why this happened. However we hypothesized that in the Bay Area, the *longitude* variable meant an Airbnb property was either closer to or further away from the coast, which could result in a significantly higher or lower rental price respectively. Just to be thorough, we also utilized stepwise selection. Using a p-value of 0.1 again, this was the result:

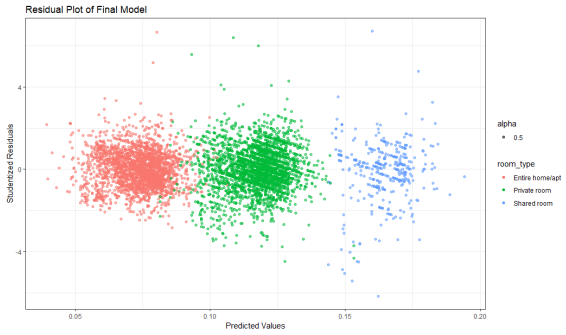
Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	room_type	addition	0.539	0.539	927.3990	-26442.2880	0.0253
2	last_review	addition	0.572	0.571	452.8170	-26855.1278	0.0244
3	neighbourhood	addition	0.588	0.586	222.6130	-27044.9404	0.0240
4	reviews_per_month	addition	0.595	0.593	122.0200	-27142.7516	0.0238
5	availability_365	addition	0.599	0.597	61.9130	-27202.0123	0.0236
6	minimum_nights	addition	0.603	0.601	-0.7140	-27264.4457	0.0235
7	longitude	addition	0.604	0.602	-14.1990	-27277.9902	0.0235
8	number_of_reviews	addition	0.604	0.602	-12.9570	-27276.7522	0.0235
9	number_of_reviews	removal	0.604	0.602	-14.1990	-27277.9902	0.0235
10	latitude	addition	0.604	0.602	-12.2370	-27276.0286	0.0235
11	latitude	removal	0.604	0.602	-14.1990	-27277.9902	0.0235

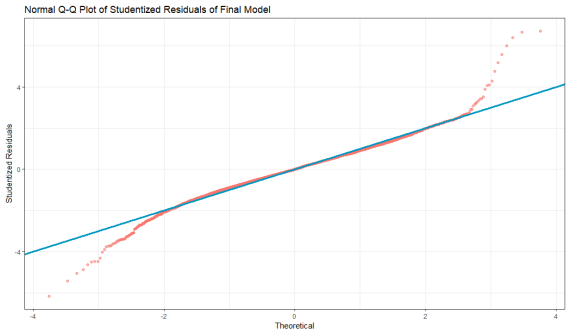
Based on the information from all three types of variable selection methods, we choose our model in this stage to be:

$$\text{Model (3): } \text{price}^{-0.5} \sim \text{room\_type} + \text{last\_review} + \text{neighbourhood} + \text{reviews\_per\_month} + \text{availability\_365} + \text{minimum\_nights} + \text{longitude}$$

Earlier, we noticed that the correlation was relatively high between the pairs of variables (1) *longitude* and *latitude*, and (2) *number_of_reviews* and *reviews_per_month*. We wanted to steer clear of any model that simultaneously used any of these two pairs to avoid issues of multicollinearity. Our final model does not use any of these pairs, so we can continue onto the next stage: model validation. Additionally, we can check normality assumptions once more by checking the residual plot (Figure 7) and qqplot (Figure 8) of Model 3.



(g) Figure 7



(h) Figure 8

IX. Model Validation

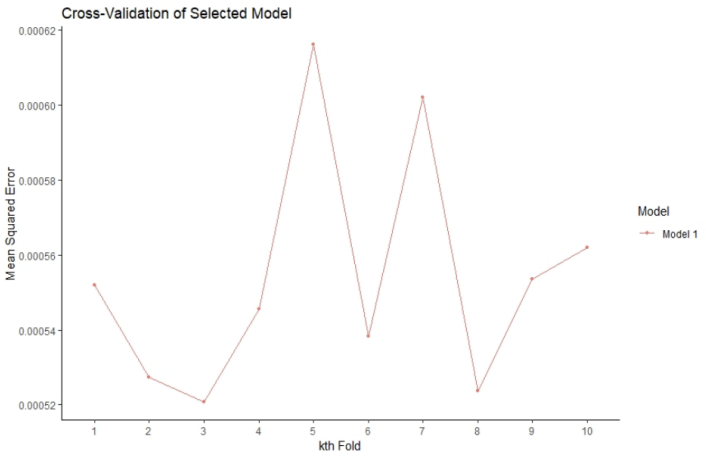
R^2 is a popular measure of goodness of fit in regression and for our final model R^2 is 0.604. Notice that R^2 is pretty close to Adj. R^2, 0.6023, which means this goodness of fit is caused by all useful predictors. However, it does not offer any significant insights into how well our regression model can predict future response. In this section, we want to use cross-validation on our final model to see how it can predict the best price for an apartment. Using our final model with transformed price and only significant variables, the data were split into 10 equally sized partitions. During the first fitting of the model, the first 10% of the data are considered the test set and the remaining 90% of the data are considered the training set. In the following iterations a different 10% of the data are considered the test set, while the remaining 90% of the data are considered the training set. The model is fit to the test/training data 10 times, and the prediction error from each model fitting is then averaged to determine the prediction statistics for the model. We can see that the value of R^2 for the whole sample is similar to the cross-validation result where R^2 is 0.6061.

The mean square prediction error as a result of cross validation, 0.000554 is just 5% greater than the MS_{Res} of Model (3), 0.00055131. We believe that our model predicts quite well. In the future if Airbnb releases more data for the Santa Clara County, we would be able to put our model to the test and use it to predict the rental prices. The following table shows the MSE in each of the ten folds:

1st fold	**2nd fold**	**3rd fold**	**4th fold**	**5th fold**
0.0005518953	0.0005272728	0.0005207722	0.0005456775	0.0006161870

6th fold	**7th fold**	**8th fold**	**9th fold**	**10th fold**
0.0005384117	0.0006021220	0.0005236355	0.0005535910	0.0005620016

This result can also be seen in the graph below:



X. Discussion

In summary, we started by fitting a model using all predictor variables we deemed would be effective and practical in predicting the price of a bed and breakfast. After residual plots showed inconstant variance and non-normality in the residuals, we used the Box-Cox method to select a variance stabilizing transformation on the response variable. Then, we checked and resolved issues of multicollinearity present with the predictor variables. Lastly, we conducted several different variable selections and concluded that the variables latitude and number of reviews were not needed in our model. The resulting model we decided was best is stated below, previously labeled as Model 3.

$$\text{Final Model: } \text{price}^{-0.5} \sim \text{room\_type} + \text{last\_review} + \text{neighbourhood} + \text{reviews\_per\_month} + \text{availability\_365} + \text{minimum\_nights} + \text{longitude}$$

This final model was chosen as the final model based on several different criteria. Our goal was to create a simple model with all assumptions of linear regression satisfied, a high R^2, low multicollinearity in the predictors, and that it was effective in predicting the price of an Airbnb in Santa Clara County.

The objective of this report was to understand the impact of several relevant variables on the base rental price of an Airbnb. We found that the most significant variables on price were the predictors included in our model.

One interesting point is the interpretation of the slopes of the individual predictor variables. Since we applied a reciprocal transformation on the response, the positive coefficients of the predictor variables in our model indicate a decrease in price per increase in one unit of the predictor variable. For example, since the variable for reviews per month has a positive coefficient, for every additional review per month, we predict a decrease in the base rental price per night of an Airbnb in Santa Clara County. However, negative coefficients in our model indicate an increase in the base rental price per night of an Airbnb in Santa Clara County.