STAT 515 Final Project

Sarah Hayes, Sean Lei, Kausik Valetta

The Data

According to their website (Cox, 2019), “inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.” The site operators use Python to scrape data from the Airbnb public website. With this data, they offer some basic tools to analyze and visualize the data. More importantly, the raw data is available for cities around the world, including the focus of our analysis, Washington, D.C. The available data includes attributes about each rental property, such as number of bedrooms, capacity, price, and amenities. It also includes the latitude and longitude of each rental property.

In order to also analyze the desirability of locations, we also obtained geographic data from Open Data DC, a website where the District of Columbia government shares hundreds of datasets. Of particular interest was ShotSpotter gun shot data, crime data, Capital Bike Share locations, Metro Station Entrance data, Museum data, and the Sex Offender Registry.

Using Microsoft Access, we first converted all latitude and longitude to radians. Then, calculated the distance of gun shots, crimes, bike shares, metro stations, museums, and sex offenders from each Airbnb listing using the haversine formula:

$$d = 2r \arcsin \sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\, sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}$$

$d = distance,$
$r = radius(6371\ for\ earth)$
$\varphi = latitude\ of\ points\ 1\ (rental\ property)\ and\ 2\ (museum, etc.)$
$\lambda = longitude\ of\ points\ 1\ (rental\ property)\ and\ 2\ (museum, etc.)$

Once distance was calculated, we added two columns for each quality considered: the distance to the nearest location and the number of applicable locations within walking distance (0.5 miles).

The Problem

To begin with, we want to determine what factors are most influential on the price of an Airbnb rental. In the data, the price varied by date, so we calculated an average nightly price, the highest and lowest prices in 2019, and the prices per person and room. Ultimately, we want to know for a given property, what amount would be reasonable to charge.

Exploratory Analysis

First, in our effort to evaluate prices, we produced summary statistics and histograms for each variable – average nightly price, lowest nightly price, highest nightly price, nightly price per bedroom, and nightly price person. These histograms are included in Appendix B for reference. Upon review, it was clear that our data was right skewed with several outliers far out to the right. After transforming each of these variables by taking the square root, we also eliminated several of the highest price rentals after determining that many of them had either never been rented or had features not captured in the dataset. For example, many of them turned out to be listed at rates specifically for inauguration weekend but had yet to be removed from the site. We further limited the data to only include properties that had at least one review. Ideally, we would have been able to limit this to at least one review within the last year, which would have resolved both issues, but that data was not available to us.

Finally, we normalized the data on a scale from 0 to 1. The resulting histograms are also shown in Appendix B and the distributions are much closer to normal than our original data. For reference, we ran diagnostic plots on our model variables without the normalization. This is discussed in more detail in the following section.

Principal Components Analysis

PCA is a dimension-reduction tool that can be used to reduce a large set of variables to a small set which still contains most of the information in the large set. Due to the number of columns, it is not plausible to plot all of the data elements simultaneously. To get a sense of trends within our dataset, we performed a Principal Components Analysis (PCA). Using PCA in our exploratory analysis allowed us to see the overall shape of the data and some trends, basically identifying observations that are similar to one another and those that are very different.

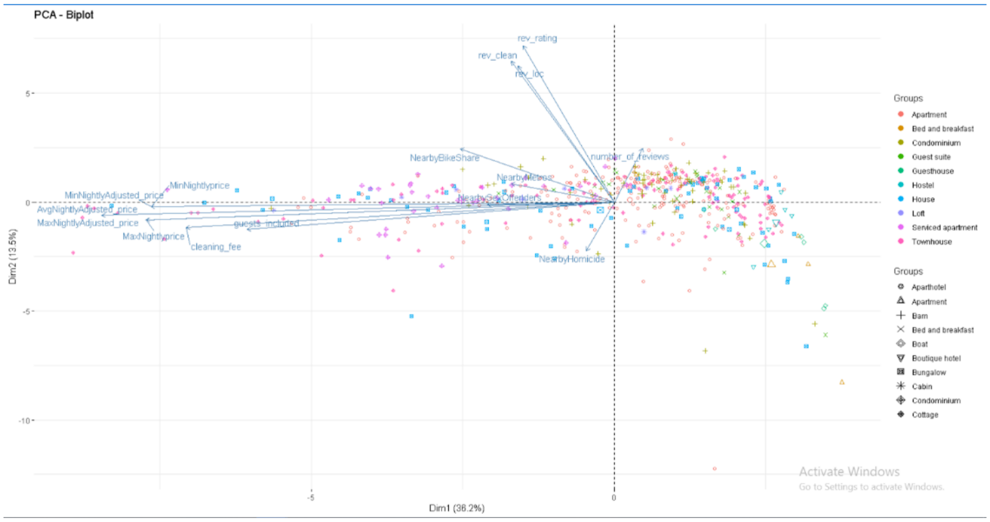

Figure 1 – Principal Component Analysis

The above figure displays the attributes in form of vectors and items in terms of samples or dots. 50% of the variance was explained in the first two principal components which is substantial. To achieve 80% of the variance we still need to use 6 principal components but most of the variance was explained in the first two principal components. From the graph, the prices, cleaning fee, guests included are highly correlated in terms of negative PC-1 whereas review for rating, cleaning, location are positively correlated in terms of positive PC-2. From the graph we can say that as the number of reviews per month is more the near by homicide is less. In PC-2 the sample points which have reviews for rating, cleaning, location, bike share, nearbymetro has less chance of nearby homicide. In PC-1 the samples which are on the positive side of PC-1 has less night prices, guests included, cleaning fee, Nearbysexoffenders. From this graph we can interpret that which variables are affecting on the property type and how these are aligned on PCA-Biplot.

After removing other price variables, we created an additional PCA. Please see Appendix C for results of this analysis. Ultimately, there was such a great volume of data that it was challenging to visualize. We did identify some relatively clear groupings related to room category and were able to confirm that nearby amenities and hazards had similar or more impact than just the number of bedrooms and bathrooms.

Method and Results

Linear Regression

Preprocessing

First, our decision variable did not have a normal distribution, with a skewness measure of 2.76. As any value above 1 indicates a significant amount of right skewness, we transformed it by taking the square root and removing outliers. As a result, we were able to reduce the skewness to 0.28, which is not a perfect normal distribution but is within the bounds of reasonableness.

We then performed a min-max normalization on all variables to bring all values into the range [0,1], using the formula below. This allows for a more directly comparison of each value of β in terms of its influence on the response variable.

$$norm(x_i) = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Our data was, then, divided between a training dataset and test dataset. Random sampling split 70% of the data into the training set with 30% in the testing set. It is important to divide the data because a model should not train and test on the same data. If the data was not split, the model would fit on all of the data and any measurement of accuracy would only pertain to the data with no indication on how well the model does on new data. The same split was used for the random forest model.

Fitting the model

Using the lm() or linear model function in R on our training dataset, we created a multiple linear regression model with average nightly price as the response variable and 47 independent variables as denoted in Appendix A. Based on the results, we were able to determine that several variables did not have a statistically significant relationship with our response variable.

To fit our model, we used R to perform stepwise model selection by minimizing the Akaike Information Criterion (AIC) to identify the deviance (*n*log(*RSS/n*)). Larger models like ours tend to have better fit and, therefore, smaller RSS but use more parameters. Ideally, we hoped to include 20 or fewer variables in the final model, which required us to balance fit with model size. Ultimately, we identified 19 variables and coefficients as listed in Table 1 with the following linear equation:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{19} x_{19}$$

Variable Description (x_i)	$\boldsymbol{\beta_i}$	**Variable Description (x_i)**	$\boldsymbol{\beta_i}$
Intercept ($\boldsymbol{\beta_0}$)	-0.1437	Cleanliness Rating-User Review	0.0915
# Reviews per month	-0.0919	# of Basic Amenities (pool, etc.)	0.0919
Distance from nearest arson	-0.0504	# of bathrooms	0.0923
# of Assaults within 0.5 miles	-0.0483	Distance from nearest gunshot	0.1012
Flexible cancellation policy	0.0093	Location Rating-User Review	0.1049
Hosted by "Super Host"	0.0124	Capacity (max # of guests)	0.1572
Designated Suitable for Families	0.0165	Cleaning Fee	0.1692
# of Museums within 0.5 miles	0.0542	# of Bike Shares within 0.5 miles	0.1822
# of Basic Amenities (A/C, etc.)	0.0557	Property Category	0.2223
Distance from nearest sex offender	0.0593	# of bedrooms	0.3321

Table 1 – Summary of Linear Regression Model

Evaluating the fit

The R-squared for our model was 0.63, which means that the model explains about 63% the variability of the response data around its mean. While this value is not especially high, all of our predictors are statistically significant predictors (See Appendix E for details), which allows us to draw conclusions about how changes in the predictor values are associated with changes in the response value. As was the objective of the analysis, the coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant regardless of the R-squared value. This does indicate, however, that predictions based on our model may not have a high degree of precision.

Our standard error is 0.097 and represents the average distance that the observed values fall from the regression line. This means that regression model is on average deviates from the actual values in our training data set by 0.097. To further evaluate the model we ran summary diagnostic plots, which did not indicate any assumption had been violated. We also ran our model on the data without preprocessing to demonstrate the differences.

Residuals vs Fitted

The residuals vs fitted plot is intended to identify residuals with non-linear patterns, such as a non-linear relationship between predictor variables and the outcome variable. If the red line is close to flat and oriented around 0, the linear assumptions are not violated. Otherwise, the model doesn't account for that non-linear relationship. You can see in below that our line is fairly flat except for the highest values. There is a significant improvement over that shown in the unaltered data.

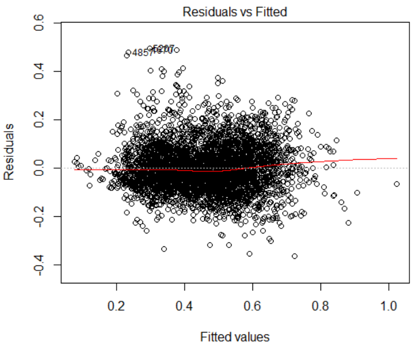

Figure 2 – Residuals vs Fitted (Final)

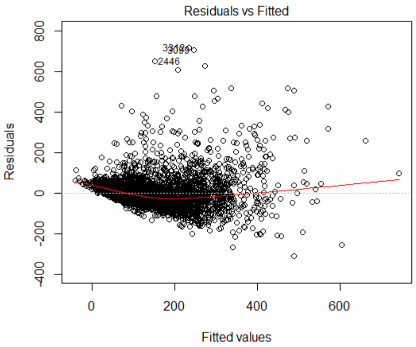

Figure 3 – Residuals vs Fitted (Pre-transformation)

Normal Q-Q

This plot shows if residuals are plausibly normally distributed. It is created by plotting two sets of quantiles against one another. If they are from the same distribution, we would see the points forming a straight line. The heavy tails in our plots below indicate, as previously noted, there is some skewness in our data, which has been significantly improved but not eliminated through transformation of the response variable and removal of outliers.

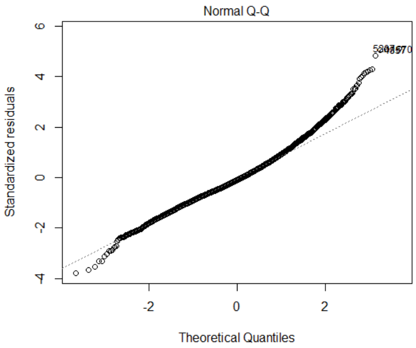

Figure 4 – Normal Q-Q (Final)

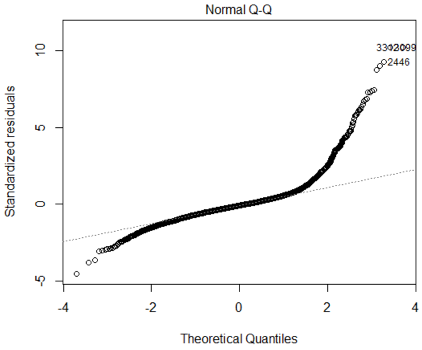

Figure 5 – Normal Q-Q (Pre-transformation)

Scale Location

This plot checks for homoscedasticity (uniform variance), showing if residuals are spread equally along the ranges of predictors. We would expect to see a horizontal line with randomly spread points roughly equal around the line. As shown, this has also been improved by our pre-processing steps. There are fewer points at the high and low ranges of our data, but our variance is fairly constant.

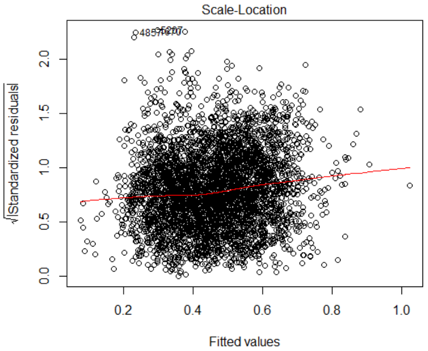

Figure 6 – Scale Location (Final)

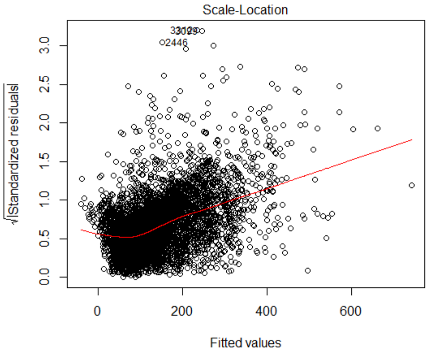

Figure 7 – Scale Location (Pre-transformation)

Residuals vs Leverage

This plot helps us to find influential observations. Not all outliers are influential; even though data have extreme values, the results may be about the same whether we include or exclude them from analysis. In the reverse, other cases could be very influential even if they look to be within a reasonable range of the values. In this plot patterns are not relevant. Ideally, there will be no values in the upper or lower right corners outside the line marking Cook's distance. This is the case in both plots shown below.

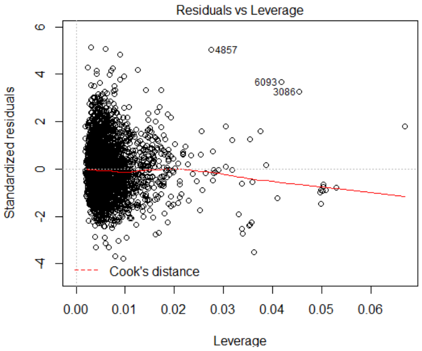

Figure 8 – Residuals vs Leverage (Final)

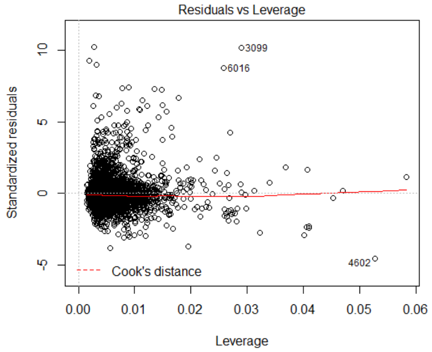

Figure 9 – Residuals vs Leverage (Pre-transformation)

Predictive Value

We calculate mean squared error (MSE) on our test dataset (1803 records) to assess the quality of our model as a predictor using the formula below.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The resulting value was 0.0090 in the units of the response variable (normalized to the range [0,1]). By converting both the predicted average nightly price and actual average nightly price to the original units using the formula below, we were able to calculate the MSE in the units of the pre-processed response variable as $1067.16.

$$X = x^2(\max(AvgNightlyPrice) - \min(AvgNightlyPrice)) + \min(AvgNightlyPrice)$$

Random Forest Regression

On top of a linear regression model, a random forest regression was constructed to predict the average nightly price of an Airbnb unit. The response variable is the average nightly price (AvgNightlyPrice) and the predictors used were all variables listed in Appendix A with a few exceptions. The independent variables left out of model training were the ids and prices (AvgNightlyPrice, AvgNightlyAdjusted_price, MinNightlyprice, MinNightlyAdjusted_price, MaxNightlyprice, MaxNightlyAdjusted_price). The id is the unique identifier used by Airbnb and was removed because it does not have any correlation with any of the other variables. The other price variables are not reasonable predictor variables and thus removed. The objective is to predict the price and therefore knowledge of the upper and lower bounds of the price would not be present before the prediction. In addition to the data, various numbers of trees and variables randomly chosen at each split were adjusted to find the optimal amount.

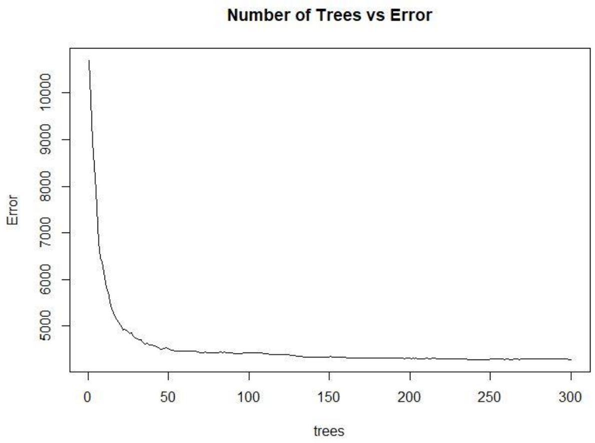

Figure 10 – Number of Trees vs Error

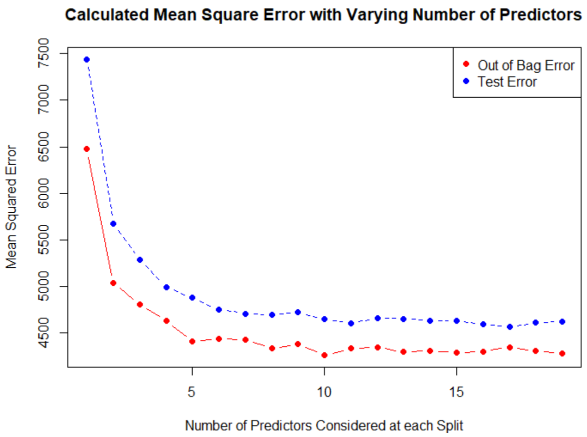

Figure 11 – MSE with Varying Number of Predictors

Figure 10 displays how the increase of trees effects the error. Initially the increase of trees drastically diminishes error, which is desired. At the certain point, there are diminishing returns to increasing the number of trees. While the mean squared error will continue to decrease, more computational power will be utilized as you add more trees. At a certain point the improvement to fit is marginal. Diminishing returns starts between 20 and 50 trees. The marginal gain from increasing the number of trees over 200 is marginal. Considering the volume of data and the computing power present, we will include 200 trees in the random forest. While creating the trees, the algorithm tries a specified number of variables at each split. This is identified by the mtry parameter. Different values for mtry were used for random forests to calculate the mean square error of each one.

As the number of predictors at each split increase, mean square error decreases. Similar to the number of trees, at a certain point there is marginal benefit in increasing the number of predictors. Although increasing the predictor number lowers error slightly, the computational cost of it increases drastically. Due to computation costs and limited mean square error improvement, the random forest regression will be constructed with an mtry number of 7.

Some variables affect the price prediction more than others. The random forest model determined the most important variables for predicting average nightly price was bedrooms, room_cat, cleaning_fee, accommodates, and host_listings_count (figure 12).

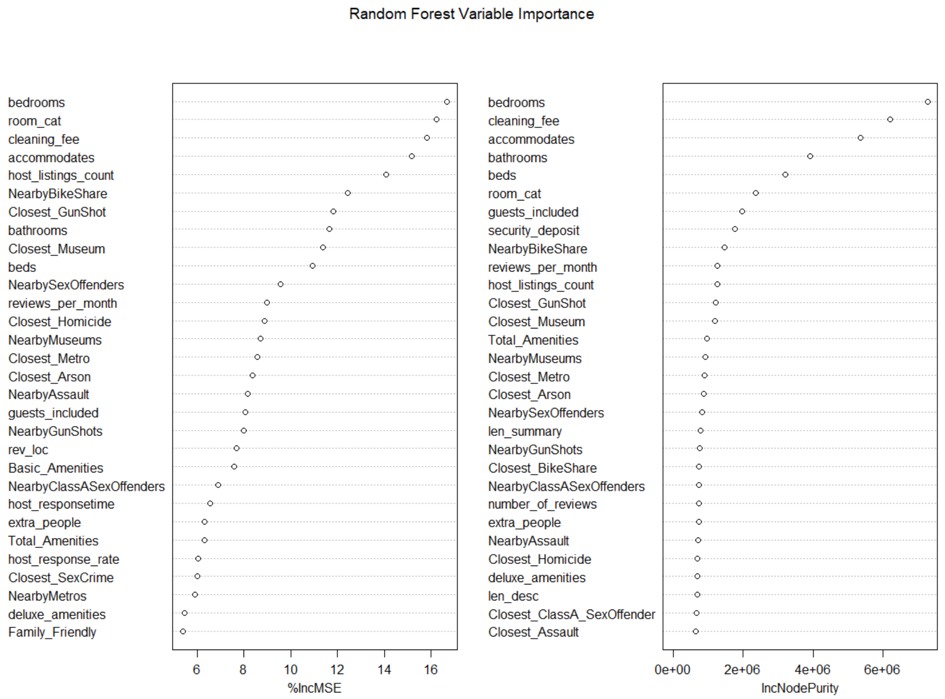

Figure 12 – Important Variables for Predicting Price

Figure 12 shows how much the MSE would increase is a variable was permuted, or randomly shuffled. The higher the %IncMSE, the more important the variable.

A random forest comprised of 200 trees and where 7 variables were randomly selected at each split, the mean squared error on the testing set was 4766.409. The root mean squared error is 69.04. The standard deviation of the unexplained variance, that is the average distance measured along the average nightly price axis, from a data point to the random forest prediction is $69.04.

Comparison of different models

A key benefit of decision trees, such as random forest, is that it can accurately divide the data based on categorical data. Because even our categorical variables (e.g. property category) could be expressed numerically given that a hierarchy exists (private rooms are preferable to shared rooms), linear regression was still a reasonable solution, though it would not have accounted well for the degree of benefit derived from each potential value (a private room is not necessarily twice as valuable as a shared room even though the numeric values would indicate this pattern). The random forest model also was able to identify certain variables which were much more important than others. It was also able to handle nonlinear relationships without the need to transform variables. Alternatively, the linear model was able to quantify the mean impact of an incremental change to each of the relevant independent variables, which could be useful in discussing differences in prices across properties or how to price a particular rental property.

The random forest and linear model shared 7 out of the 8 most important variables in each regression. One interesting thing about our particular linear regression model and random forest was that the number of listings for the particular host did not factor into the linear regression model but was relatively high importance in the random forest model. In Table 2, we ranked the importance of each variable by β for the linear regression model and %incMSI for the random forest. Other variables, such as NearbyBikeShare, Closest_GunShot and bathrooms were similar in importance. Due to the nature of the random forest, all variables were selected during the training process. For the linear model, only statistically significant variables were kept, and thus explains the n/as in table 2 for the linear regression model.

Variable	RF	LR	Variable	RF	LR	Variable	RF	LR
bedrooms	1	1	NearbySexOffenders	11	n/a	Basic_Amenities	21	13
room_cat	2	2	reviews_per_month	12	9	NearbyClassASexOffenders	22	n/a
cleaning_fee	3	4	Closest_Homicide	13	n/a	host_responsetime	23	n/a
accommodates	4	5	NearbyMuseums	14	14	extra_people	24	n/a
host_listings_count	5	n/a	Closest_Metro	15	n/a	Total_Amenities	25	n/a
NearbyBikeShare	6	3	Closest_Arson	16	15	host_response_rate	26	n/a
Closest_GunShot	7	7	NearbyAssault	17	16	Closest_SexCrime	27	n/a
bathrooms	8	8	guest_included	18	n/a	NearbyMetros	28	n/a
Closest_Museum	9	n/a	NearbyGunShots	19	n/a	deluxe_amenities	29	10
beds	10	n/a	rev_loc	20	6	Family_Friendly	30	17

Table 2 – Ranked Importance of Variables by Model

The mean squared error (MSE) of the random forest is 4766.41 and the MSE of the linear regression is 1067.16. Converting the MSE to the root mean squared error (RMSE) is the standard deviation of the unexplained variance, which allows us to easily interpret the statistic since it has the same units as the dependent variable. The RMSE of the random forest is $69.04 and the RMSE of the linear model is $32.66. This would indicate that the linear regression model has a better absolute fit than the random forest. However, the random forest is not as susceptible to problems when dealing with non-normal distributions, so it was run on the raw data. When run on the normalized and scrubbed data, MSE for random forest is 0.0076 ($906.65) compared to 0.0090 ($1067.16) and the RMSEs are $30.11 and $32.67 respectively. To calculate the performance difference in percentage, the residual mean squared percentage error (RMSPE) was calculated for both models. RMSPE for the random forest was 37.0% and the RMSPE for the linear model was 41.9%. Based on RMSPE, the random forest performance 4.9% better than the linear model. Using a different performance metric, the average percentage error (APE) for the random forest was 26.4% and the APE for the linear model was 29.2%. Based on APE, the random forest performed 2.8% better than the linear model.

Conclusion

The random forest model provided a reasonably good fit for all of the data available, including outliers that would have caused key assumptions to be violated in linear regression. Still, when precision is required, it is best to analyze the data and remove observations that are not relevant to the analysis (e.g. over-priced rentals that have never been rented or were for special events only are not relevant to our question). By performing this step, we were left with a model for a vast majority of Airbnb rentals with a much better absolute measure of fit - less than half the RMSE.

From the linear regression, we were able to determine that only a few factors have an inverse relationship with price. The most significant, reviews per month, may indicate that properties that are consistently in use have lower average properties. This is reasonable when you consider that the properties used most often are likely to be run by management companies that have the ability to undercut competing prices and are less likely to include properties rented out for significant events in the area, like inaugurations or protests. What was surprising, though, was that arson was the only crime with a real negative association with price. Upon further review, we determined that all of the reported arsons were committed in a (subjectively) less desirable area, near Anacostia, that is far from most of the more desirable areas as shown in Figure 13.

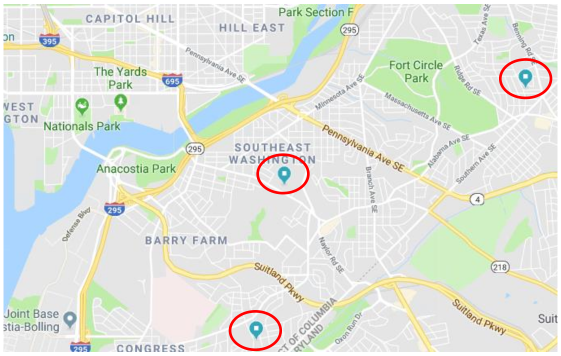

Figure 13 – Locations of Arsons

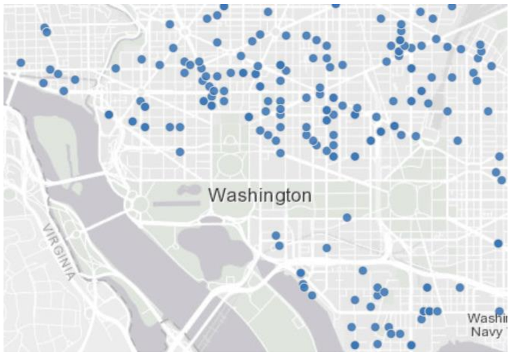

Figure 14 – Locations of Sex Offenders
Source: DCGISopendata. (n.d.)

Another surprise was that the distance from nearest sex offender and gunshot were positively correlated. As shown in Figure 14, sex offenders live near many desirable locations, including the white house. Georgetown, and DuPont circle. Also we identified 29 records in the Airbnb data where a host listed their property as Washington DC, but it was actually in a suburb; since we only had DC crime data, lower priced properties in the suburbs of Virginia and Maryland would be further away from the nearest one. This coupled with the fact that there are likely other relationships driving some of the crime data, makes interpreting results challenging.

Most of the remaining variables matched our expectations, with the number of bedrooms, property category, capacity, proximity to bike share locations, number of bathrooms, and cleaning fees all having strong positive relationships with price. What was surprising was that amenities, while positively correlated with price, did not have nearly as strong a relationship. At the end of the day, it seems, you pay the most for location, size/capacity, and cleanliness when it comes to these rentals.

From the random forest model, we were able to see which factors were most important. For the most part, this corresponded to the linear model. We were not able to see, though, the direction of the impact (which factors are associated with higher vs lower prices). The model was, however, able to be executed on the data in its raw form and did not require the pre-processing needed to perform the linear regression.

As mentioned previously, it would be useful to cleanse the data of all one-time, special event rentals. It may also be valuable to pull data near the end of a year to see what proportion of the time a property was rented; as our data was pull in April and had only 2019 dates, it is not reflective of a full cycle given DC tourist patterns. It would also be helpful to either remove the 29 rentals outside of DC or include additional data from other states for sex offenders and crimes. It would also be useful to explore whether the variables are truly independent since user reviews for location and cleanliness were relevant. Are location ratings impacted by safety concerns or the presence of nearby amenities? With some further exploration, it appears that this may be the case. Another plausible hypothesis is that properties with higher cleaning fees get better ratings in cleanliness. If so, adjustments should be made to ensure independence of the variables.

References

1. Cox, M. (2019, April 15). Inside Airbnb: Get the Data: Washington, D.C., District of Columbia, United States. Retrieved April 26, 2019, from http://insideairbnb.com/get-the-data.html
2. DCGISopendata. (n.d.). Sex Offender Registry. Retrieved April 26, 2019, from http://opendata.dc.gov/datasets/sex-offender-registry
3. DCGISopendata. (2018, November 16). Museums in DC. Retrieved April 26, 2019, from http://opendata.dc.gov/datasets/museums-in-dc
4. DCGISopendata. (2018, December 31). Crime Incidents in 2018. Retrieved April 26, 2019, from http://opendata.dc.gov/datasets/crime-incidents-in-2018
5. DCGISopendata. (2019, February 14). Shot Spotter Gun Shots. Retrieved April 26, 2019, from http://opendata.dc.gov/datasets/shot-spotter-gun-shots
6. DCGISopendata. (2018, August 20). Capital Bike Share Locations. Retrieved April 26, 2019, from http://opendata.dc.gov/datasets/capital-bike-share-locations
7. DCGISopendata. (2012, July 12). Metro Station Entrances in DC. Retrieved April 26, 2019, from http://opendata.dc.gov/datasets/metro-station-entrances-in-dc

Appendix A – Data

Column Name	Description	Source[1]	LR[2]
id	Unique ID for the AirBnb Rental Property	1	
bathrooms	Number of bathrooms	1	I,F
bedrooms	Number of bedrooms	1	I,F
beds	Number of beds	1	
guests_included	Number of guests included in the nightly price	1	
extra_people	Cost of each additional guest in excess of "guests_included"	1	
cancel_pol	Value 0 - 3 based on strictness, 0 is least strict	1	I,F
room_cat	Value 1 - 3 based on privacy, 1 is least private	1	I,F
Family_Friendly	Boolean, 1 if property is tagged as suitable for families	1	I,F
Pets_Allowed	Boolean, 1 if property is tagged as allowing pets	1	I
Smoking_Allowed	Boolean, 1 if property is tagged as allowing smoking	1	I
Events_Allowed	Boolean, 1 if property is tagged as allowing parties/events	1	I
rev_rating	Average of all user reviews for the "Overall" rating	1	I
rev_acc	Average of all user reviews for the "Accuracy" rating	1	I
rev_clean	Average of all user reviews for the "Cleanliness" rating	1	I,F
rev_check	Average of all user reviews for the "Check-in experience" rating	1	I
rev_comm	Average of all user reviews for the "Communication" rating	1	I
rev_loc	Average of all user reviews for the "Location" rating	1	I,F
rev_val	Average of all user reviews for the "Value" (best value for the $) rating	1	I
host_responsetime	Value 0-4 rating of responsiveness, 0 is least responsive	1	I
host_response_rate	Percent of inquiries host responded to	1	I
host_is_superhost	Boolean, 1 if host is certified "Superhost" by AirBnb	1	I,F
host_listings_count	Number of properties that the host has listed on AirBnb	1	I
accommodates	Number of people that can sleep at the property	1	I,F
Basic_Amenities	Number of basic amenities listed (AC, heating, wifi, TV)	1	I,F
deluxe_amenities	Number of deluxe amenities listed (e.g. pool, hot tub, etc.)	1	I,F
Total_Amenities	Number of amenities listed	1	I
security_deposit	Cost of the refundable security deposit	1	I
cleaning_fee	Cost of the cleaning fee (charged once per stay, not nightly)	1	I,F
number_of_reviews	Number of reviews given for the property	1	I
reviews_per_month	Average monthly reviews	1	I,F
NearbyArson	Average number of annual arson crimes within 0.5 miles of the property	4,1	I
NearbyAssault	Average number of annual assault crimes within 0.5 miles of the property	4,1	I,F

[1] Reflects the reference number for the related source data

[2] I indicates variable included in initial linear regression model, F indicates variable still included in final

Column Name	Description	Source[1]	LR[2]
NearbyBikeShare	Number of bike share drop offs within 0.5 miles of the property	6,1	I,F
NearbyClassASexOffenders	Number of Class A sex offenders living within 0.5 miles of the property	2,1	I
NearbyGunShots	Average number of semi-annual gun shots within 0.5 miles of the property	5,1	I
NearbyHomicide	Average number of annual homicide crimes within 0.5 miles of the property	4,1	I
NearbyMetros	Number of metro stations within 0.5 miles of the property	7,1	I
NearbyMuseums	Number of museums within 0.5 miles of the property	3,1	I,F
NearbySexCrime	Average number of annual sex crimes within 0.5 miles of the property	4,1	I
NearbySexOffenders	Number of sex offenders living within 0.5 miles of the property	2,1	I
Closest_BikeShare	Distance from the property to the nearest arson crime within the last year	6,1	I
Closest_Metro	Distance from the property to the nearest assault crime within the last year	7,1	I
Closest_Museum	Distance from the property to the nearest bike share drop off as of 12/31/18	3,1	I
Closest_SexCrime	Distance from the property to the nearest class A sex offender as of 12/31/18	4,1	I
Closest_SexOffender	Distance from the property to the nearest gun shot within the last 6 months	2,1	I,F
Closest_Arson	Distance from the property to the nearest homicide crime within the last year	4,1	I,F
Closest_Assault	Distance from the property to the nearest metro entrance as of 12/31/18	4,1	I
Closest_GunShot	Distance from the property to the nearest DC museum as of 12/31/18	5,1	I,F
Closest_Homicide	Distance from the property to the nearest sex crime within the last year	4,1	I
Closest_ClassA_SexOffender	Distance from the property to the nearest registered sex offender as of 12/31/18	2,1	I
AvgNightlyPrice	Average (mean) nightly rental price excluding fees	1	RV
AvgNightlyAdjusted_price	duplicate column with the above	1	
MinNightlyprice	Lowest nightly rental price in 2019 excluding fees	1	
MinNightlyAdjusted_price	duplicate column with the above	1	
MaxNightlyprice	Highest nightly rental price in 2019 excluding fees	1	
MaxNightlyAdjusted_price	duplicate column with the above	1	
square_feet	Size of property in squarefeet (only populated for 57)	1	
property_type	Type of rental (e.g. room, apartment, townhouse)	1	
len_desc	Number of characters in the description of the property	1	
len_experiences	Number of characters in the description of "experience" near the property	1	
len_summary	Number of characters in the summary of the property	1	
len_rules	Number of characters in the "house rules" of the property	1	
Cost_room	The average cost per bedroom	1	
Cost_person_incuded	The average cost per person included in the price	1	

Appendix B – Histograms

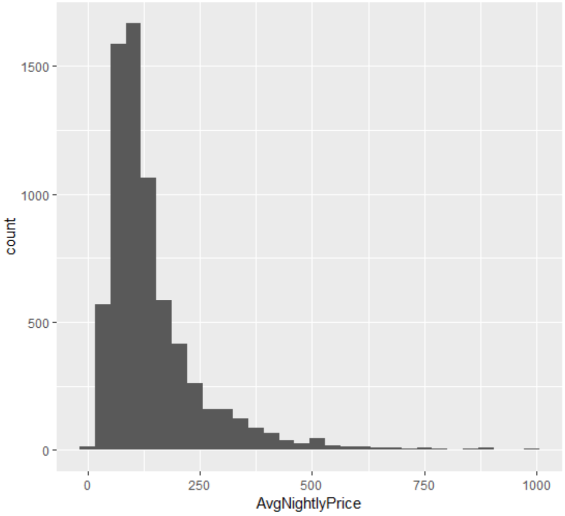

Figure 15 – Average Nightly Price Distribution (Pre-transformation)

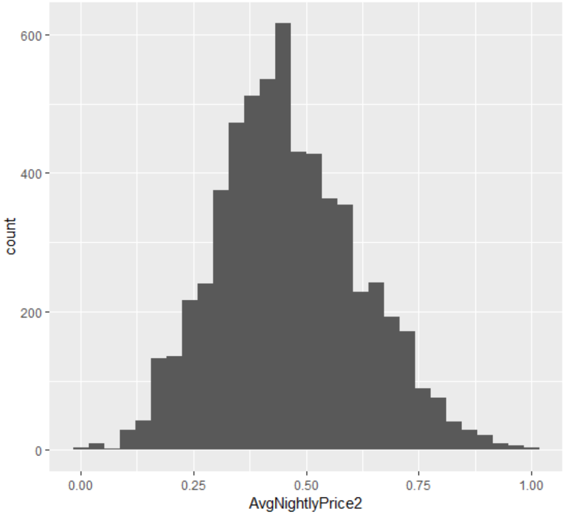

Figure 16 – Average Nightly Price Distribution (Post-transformation)

Appendix C – Principal Components Analysis

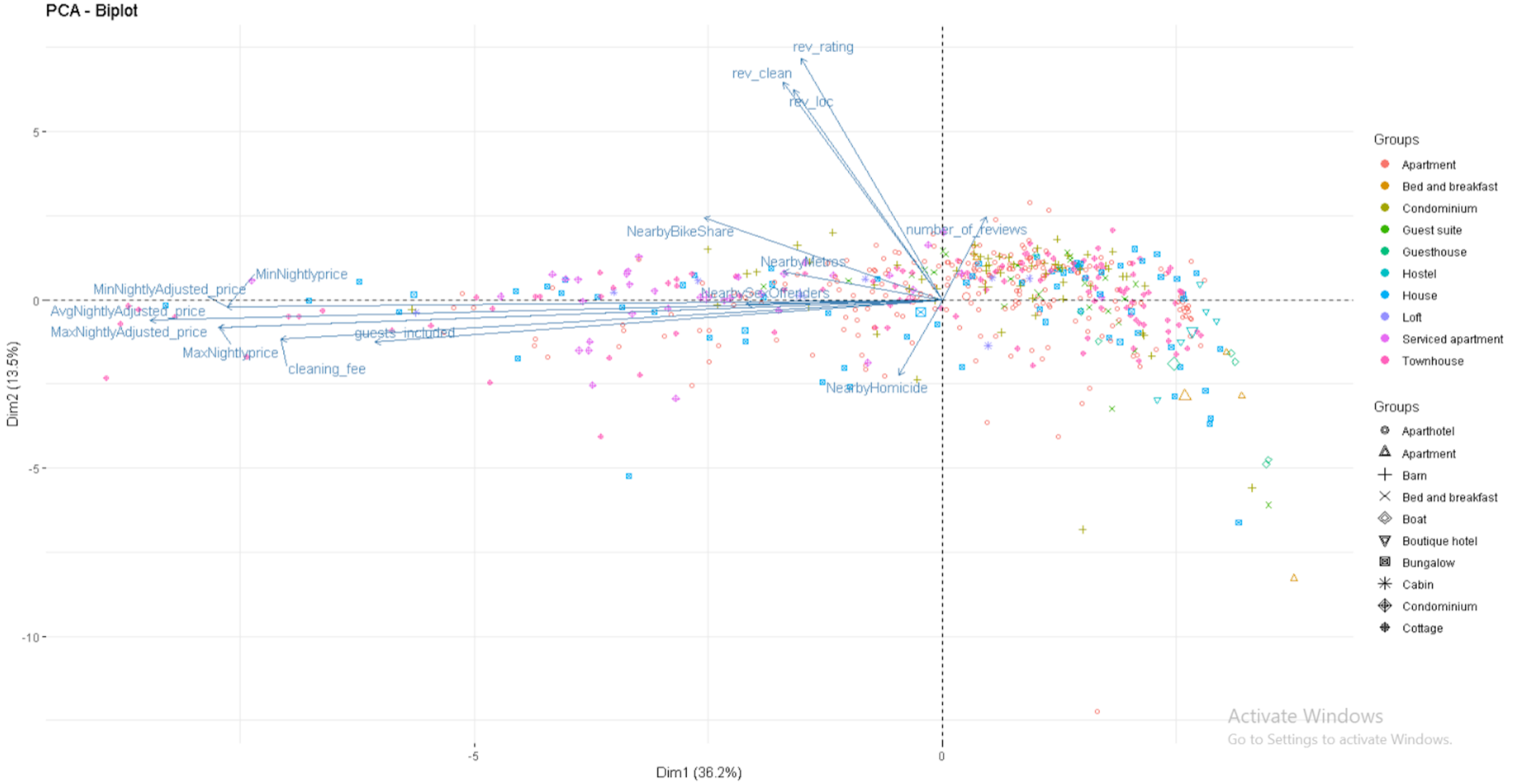

Figure 17 – Principal Component Analysis

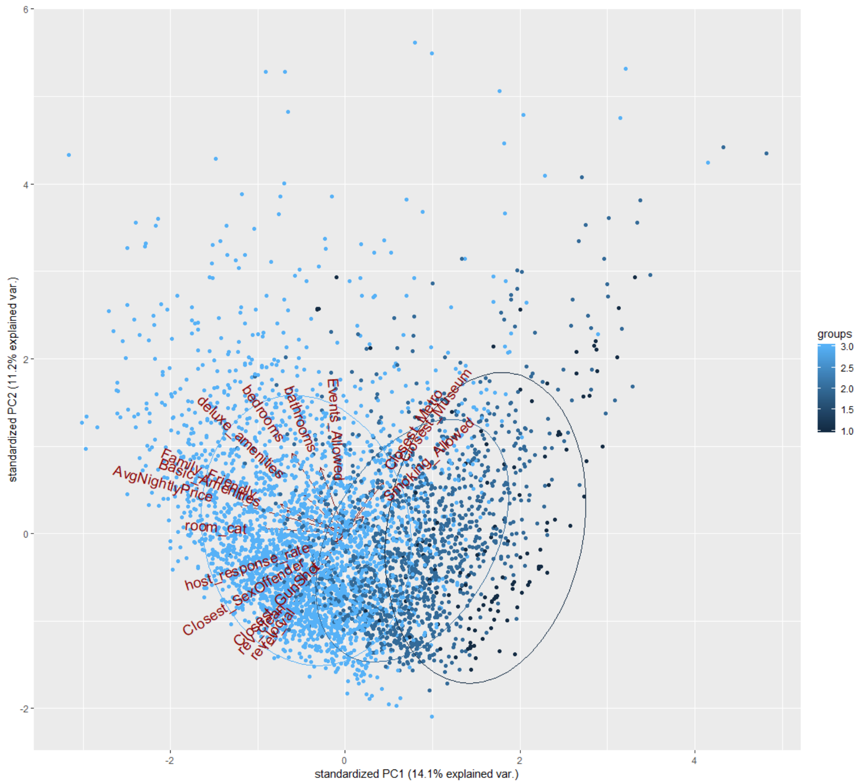

Figure 18 – PC1 amd PC2

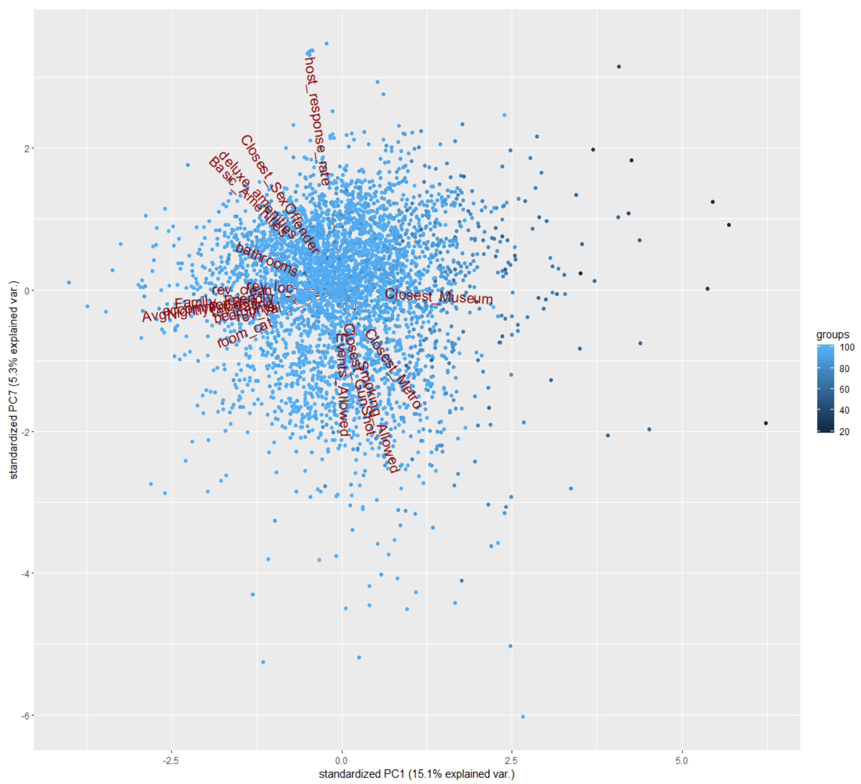

Figure 19 – PC1 and PC3

Appendix D – Initial Linear Regression Model Summary

```
     Min       1Q   Median       3Q      Max 
-0.37891 -0.06164 -0.00953  0.05219  0.48735 
Coefficients:                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -0.018552   0.028895  -0.642 0.520883    
bedrooms2                    0.339012   0.017563  19.302  < 2e-16 ***
bathrooms2                   0.099578   0.028334   3.514 0.000445 ***
cancel_pol2                  0.011276   0.003843   2.934 0.003366 ** 
room_cat2                    0.222208   0.006777  32.787  < 2e-16 ***
Family_Friendly2             0.014534   0.003355   4.333 1.51e-05 ***
Pets_Allowed2                0.003722   0.004402   0.845 0.397944    
Smoking_Allowed2             0.023505   0.010733   2.190 0.028572 *  
Events_Allowed2             -0.012407   0.008104  -1.531 0.125844    
rev_rating2                  0.075964   0.032697   2.323 0.020213 *  
rev_acc2                     0.022145   0.027271   0.812 0.416822    
rev_clean2                   0.121924   0.020818   5.857 5.09e-09 ***
rev_check2                  -0.013410   0.030593  -0.438 0.661172    
rev_comm2                   -0.087149   0.029154  -2.989 0.002813 ** 
rev_loc2                     0.101547   0.017350   5.853 5.20e-09 ***
rev_val2                    -0.100127   0.024758  -4.044 5.34e-05 ***
host_responsetime2           0.009410   0.012096   0.778 0.436640    
host_response_rate2         -0.021703   0.011425  -1.900 0.057560 .  
host_is_superhost2           0.016303   0.003603   4.525 6.21e-06 ***
host_listings_count2         0.025231   0.044162   0.571 0.567809    
accommodates2                0.159907   0.016887   9.469  < 2e-16 ***
Basic_Amenities2             0.053658   0.010640   5.043 4.78e-07 ***
deluxe_amenities2            0.094726   0.014916   6.351 2.37e-10 ***
Total_Amenities2            -0.007113   0.018596  -0.382 0.702120    
security_deposit2            0.030799   0.027436   1.123 0.261677    
cleaning_fee2                0.163363   0.017317   9.434  < 2e-16 ***
number_of_reviews2           0.046533   0.023644   1.968 0.049127 *  
reviews_per_month2          -0.096244   0.015855  -6.070 1.39e-09 ***
NearbyArson2                -0.005866   0.007468  -0.786 0.432177    
NearbyAssault2              -0.069257   0.014580  -4.750 2.10e-06 ***
NearbyBikeShare2             0.108485   0.016471   6.586 5.07e-11 ***
NearbyClassASexOffenders2   -0.107989   0.045937  -2.351 0.018779 *  
NearbyGunShots2              0.018842   0.019802   0.952 0.341397    
NearbyHomicide2             -0.011839   0.013586  -0.871 0.383577    
NearbyMetros2                0.043435   0.019213   2.261 0.023827 *  
NearbyMuseums2               0.024905   0.015097   1.650 0.099093 .  
NearbySexCrime2              0.002881   0.011206   0.257 0.797087    
NearbySexOffenders2          0.147590   0.055051   2.681 0.007370 ** 
Closest_BikeShare2          -0.053295   0.021478  -2.481 0.013128 *  
Closest_Metro2              -0.039176   0.015558  -2.518 0.011835 *  
Closest_Museum2             -0.076570   0.014824  -5.165 2.51e-07 ***
Closest_SexCrime2            0.008966   0.018574   0.483 0.629313    
Closest_SexOffender2         0.043066   0.020407   2.110 0.034888 *  
Closest_Arson2              -0.056560   0.011488  -4.924 8.83e-07 ***
Closest_Assault2            -0.019154   0.016923  -1.132 0.257763    
Closest_GunShot2             0.124494   0.016299   7.638 2.72e-14 ***
Closest_Homicide2           -0.010623   0.015387  -0.690 0.489976    
Closest_ClassA_SexOffender2  0.055851   0.028432   1.964 0.049556 *  
---
Residual standard error: 0.09559 on 4156 degrees of freedom
Multiple R-squared:  0.6416,	Adjusted R-squared:  0.6375 
F-statistic: 158.3 on 47 and 4156 DF,  p-value: < 2.2e-16
```

Appendix E – Refined Model Summary

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.36468 -0.06241 -0.00981  0.05383  0.50158 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -0.143657   0.019512  -7.362 2.16e-13 ***
bedrooms2             0.332127   0.017623  18.846  < 2e-16 ***
bathrooms2            0.092267   0.028595   3.227 0.001262 ** 
cancel_pol2           0.009324   0.003833   2.433 0.015023 *  
room_cat2             0.222312   0.006709  33.136  < 2e-16 ***
Family_Friendly2      0.016497   0.003283   5.025 5.25e-07 ***
rev_clean2            0.091518   0.015413   5.938 3.12e-09 ***
rev_loc2              0.104853   0.015970   6.566 5.81e-11 ***
host_is_superhost2    0.012372   0.003457   3.579 0.000349 ***
accommodates2         0.157163   0.016938   9.279  < 2e-16 ***
Basic_Amenities2      0.055727   0.009259   6.019 1.91e-09 ***
deluxe_amenities2     0.091913   0.011444   8.032 1.24e-15 ***
cleaning_fee2         0.169209   0.016511  10.248  < 2e-16 ***
reviews_per_month2   -0.091939   0.011540  -7.967 2.07e-15 ***
NearbyAssault2       -0.048292   0.010581  -4.564 5.16e-06 ***
NearbyBikeShare2      0.182230   0.012524  14.550  < 2e-16 ***
NearbyMuseums2        0.054157   0.013930   3.888 0.000103 ***
Closest_SexOffender2  0.059332   0.016538   3.588 0.000338 ***
Closest_Arson2       -0.050394   0.009762  -5.162 2.55e-07 ***
Closest_GunShot2      0.101182   0.013913   7.272 4.19e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09707 on 4184 degrees of freedom
Multiple R-squared:  0.6279,	Adjusted R-squared:  0.6262 
F-statistic: 371.6 on 19 and 4184 DF,  p-value: < 2.2e-16
```

Appendix E – R Code (Embedded File)

AirBnB_KV.r

final_SH_SL.R