Airbnb Price Advisor

Group 11:

Adam Frink

Catherine Salgado

Daniel Kavuu

Micheal Bradley

Zhimin Zou

Final Project

DCS 423: Data Analysis and Regression

Depaul University

11/25/2019

TABLE OF CONTENTS

ABSTRACT

Over 2 million people currently choose to stay in an Airbnb per day. At the time of writing this paper Airbnb has over 7 million listings, in over 100,000 cities, in greater than 191 countries.[1] The pricing of Airbnb units is an important factor to the millions who rent and rent out their space. The purpose of this paper is to summarize the modeling creation process and analyze which attributes of an Airbnb unit drives its price using regression analysis. Our analysis focuses specifically on Airbnb listings in the city state of Singapore, using data from August 28th 2019 and limiting our analysis to listings less that $500 SGD. We found that proximity to the southernmost latitude (proximity to Singapore's southern coast), availability (in days), number of bathrooms, number of people a unit can accommodate, and increased privacy positively drive listing price. Whereas units located in the medium priced neighborhood subgroup, increased number of reviews, and being a shared room decrease the listing price.

INTRODUCTION

The objective of this project is to create a model to provide price estimates for Airbnb listings listed in Singapore on August 28th 2019. We hypothesized that we could use the data points from "Singapore Airbnb at 28 August 2019' dataset retrieved from www.kaggle.com we can create a regression model that will estimate Airbnb listing prices.

Such a model would be beneficial to Airbnb hosts and renters alike. A price estimator would benefit Airbnb hosts by providing them with a reference point of how much money their listing can command based on the attributes of their property. The same is true for Airbnb renters. Renters can use estimated price as a tool to assess the value of an Airbnb listing vs the estimated listing price. Such a tool is important because as stated by Zhang, Chen,Han, and Yang in their paper *Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach*

> *Price plays an important role in the shared economy in the hospitality industries such as Airbnb because price impacts guests' lodging selection and also significantly impacts hosts' profits. Thus, knowing factors affecting the price is of great value, and can help hosts ask a reasonable price so that both the hosts and guests can benefit from the sharing economy.*[2]

Literature Review

Gibbs et al. did a deep dive into what drives pricing for Airbnb units in 5 of Canada's largest cities in 2017 with their paper *Pricing In The Sharing Economy: A Hedonic Pricing Model Applied To Airbnb Listings.*They found that the basic characteristics of the airbnb unit such number of bedrooms, bathrooms, and number of people it can accommodate as well and whether or not the unit is private will all positively increase the listing price of an Airbnb unit.[3] These findings are intuitive and unsurprising. Other research has focused on less intuitive factors such as proximity to city centers and tourist attractions as well as well as the level of sophistication of Airbnb hosts effect price. Their findings interestingly are mostly conflicting.

Li, Pan, Yang, and Guo argue that distance from tourist landmarks or the coast will influence Airbnb rental prices in their research paper *Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering.* Li et al. assert that Airbnb units are unlike traditional housing rentals and are largely used for vacation travel and thus their price increases when they are in close proximity to tourist landmarks and transportation.[5] Gibs et al. had similar findings showing that closer proximity to the city center, defined as the locations of City Hall, had a positive effect on price.[3] Conversely, when Perez-Sanchez, Serrano-Estrada, Marti, and Mora-Garcia tested if proximity to tourist locations and landmarks affected Airbnb pricing in 4 coastal Spanish cities in their 2018 study they found a negative relationship between price and proximity to landmarks and the city center. They instead found that proximity to the coast drove a positive influence on price.[4] It will be interesting to see if we observe either of these trends in our data.

Another variable that may influence Airbnb listing price observed in existing research is if the number of units an Airbnb lister has under management. Li, Moreno, and Zhang in their paper *Agent Behavior in the Sharing Economy: Evidence from Airbnb* explores the professionalization of Airbnb listers. They found that listers in Chicago who oversee greater than one unit have higher daily revenue, higher occupancy rates and are less likely to exit the Airbnb platform.[7] Gibbs et al. have contrary take finding a negative correlation between number of unit reviews and price. They hypothesized that properties that were more actively managed where more likely to have lower prices in order to have higher occupancy rates.[3]

Data Information

Dataset Name: Singapore Airbnb at 28 August 2019

Singapore Airbnb listing information originally obtained from
http://insideairbnb.com/get-the-data.html

The dataset used by this project has been retrieved from:
https://www.kaggle.com/jojoker/singapore-airbn

This dataset is retrieved from AirBnB live listing on August 28th 2019. This dataset includes specific location, review information, neighborhood information and pricing of each AirBnB listing in Singapore.

There are 11 Numeric Variable., 5 text variables, 1 date variable, 2 indices, 2 descriptive columns, 11 Independent Variables and 1 Dependent Variable. This particular dataset contains 7907 entries. The data type and variable type associated with each variable is listed in Appendix F

Id: room id
Name: room names
Host_id: host id
Host_name: host names
Neighbourhood_group: Singapore regions
Neighbourhood: specific
Latitude: latitude
Longitude: longitude
Room_type: room type
Price: singapore dollar per night
Minimum_nights: minimum nights
Number_of_reviews: number of reviews
Last_review: last review
Reviews_per_month: number of reviews per monthly aggregate
Calculated_host_listings_count: total room or house in host catalog on Airbnb
Availability_365: availability

As the original dataset found on Kaggle is missing bedroom and accommodation information. Through further investigation on the origin of the Kaggle dataset, we found the complete set of scraped data, http://insideairbnb.com/get-the-data.html, that included the following attributes. We joined the attributes in Python with Id: room id column.

Accommodates: number of people the room can accommodate
Bathrooms: bathroom count
Bedrooms: bedroom count
Beds: bed count

Data Cleaning

The original *Singapore Airbnb at 28 August 2019* dataset pulled from Kaggle.com contained 7,907 observations. Before importing into SAS we executed analysis in Python/Jupyter notebooks. We chose to build our model for Airbnb listings under $500 SGD (Singapore Dollars), since we were only interested in property bookings under $500 SGD for this project. A sampling without replacement has been used in Python to randomly sample the dataset down to the restricted 3000 rows.

After limiting the scope of our dataset to observations with listing prices less $500 SGD, the next step was to take an early look for potential independent variables that affect price. We created a scatterplot to facilitate this. The scatter plot showed only a few independent variables with weak linear relationships to price (Appendix A). We hypothesized that this might be due to outliers or erroneous data points. We ran a univariate procedure on our independent variables as a starting point to identify extreme observations that were potentially erroneous. We found that most of the extreme observations were caused by unique property types or obvious data entry errors. Unique properties such as hostels, party room rentals, and hammock rentals for example possessed abnormal data. The hostel had 1 room with 16 beds and 5 bathrooms. The party room had 1 room with 2 bathrooms and accommodated 16 people. The hammock rental had 0 rooms, 0 beds, and 0 bathrooms. We choose to manually remove oddities such as these. In general, we manually removed listings that did not appear to have the bed and bedroom numbers needed to house the number of people the listing claimed it could accommodate. We also removed listings that appeared to have data entry issues, such as a listing that had a minimum stay of 1000 days.

Our analysis of the extreme observation analysis output also led us to remove an independent variable completely. Calculated_host_listings_count appeared to have a super host “Jay” who had 274 listings. Upon further analysis of “Jay” we found that the data was actually aggregating host listing count on host name instead of host id. Therefore, we decided to exclude calculated_host_listings_count from the data set since it appears to be incorrectly calculated.

A more detailed analysis and output form the univariate procedure can be viewed in Appendix B. The full list of the observations manually reviewed is available in Appendix C.

METHODOLOGY

In this section of the report, we will define our analysis workflow and structure. At the same time we will document the detailed pre-processing stage of our dataset.

Pre-Processing and Creation of Dummy Variables

Several variables of interest required preprocessing in order to be modeled:

Neighborhood: The neighborhood variable from the original dataset is comprised of 40 Singapore neighborhoods/regions. We were interested in the effect neighborhood might have on the price of an Airbnb listing, so to make it more manageable to analyze we used price aggregation to group the neighborhoods into 3 bins (low, medium, and high) based on average price per neighborhood. The split was based on equal depth binning. The price we used to divide the neighborhoods into 3 groups of neighborhood is $100 singapore dollar and $160. For listings in neighborhood lower than $100, we binned them into low_region_by_price. Neighborhood listing between $100 and $160 are binned into medium_region_by_price. Listing in neighborhood greater than $160 is binned into high_region_by_price. We then created the dummy bit variables for each bucket: low_region_by_price, medium_region_by_price, and high_region_by_price. Low_region_by price is used as the base case.

Amenities: The amenities variable in the original dataset is comma separated text list of all the amenities the Airbnb contains. In order to analyze amenities we used Pandas from Python the parse the amenities variable and output each discrete amenity as its own column in the dataframe. Each column is encoded by having the amenity 0 and not having the amenity 1. These columns can be used as dummy variables.

Room_type: Room_Type in the original dataset consists of 'Entire home/apt', "Private room', and 'Shared room'. We created binary dummy variables for each room type: Enitere_home_apt, Private_room, Shared_room. Private_room is used as base case.

Cor_downtown_area: The original dataset provides the latitude and longitude of the Airbnb property listing. We found previous research that suggests that proximity of the Airbnb to popular landmarks positively affects its rental price, Li, et.all.[7] According the the Singapore Tourism Board the Marina Bay area is the epicenter of tourism in Singapore and contains architectural icons and large land reclamation works.[6] For this reason we choose to use geographical center of Marina Bay as our most desirable tourism/landmark reference point. We then created a new variable *distance_from_core* stores the calculated euclidean distance of the Airbnb unit to the Marina Bay center using the latitude/longitude of the unit in the dataset and the latitude/longitude of the Marina Bay core/center which we found online.

Last_review_new: The original dataset provided last review as date format for data type. We took the difference in days between the date the data was scraped, August 28th 2019, and the last reviewed date and calculated the date elapsed from the last review of the Airbnb listing.

Missing Values

A number of columns from original data that is dominated by missing values has been cut from our attribute list selected as described in previous section. There are a number of missing attributes from last_review_new, reviews_per_month column. These missing values are due to 0 reviews on the Airbnb listing. We have replaced the missing NULL values with integer 0.

Model Approach

After our data has been pre-preprocessed, we will import the CSV formatted data into SAS in order to perform data exploration. During the data exploration phase of the analysis we will first examine the data with interaction coefficient and dummy variables. Once data has been reviewed and confirmed to be imported correctly, we will visually inspect the data in various plots, tables, and high level descriptives. We will generate a full model with all attributes described in the previous section. Pre-qualified assumptions will be examined on the full model. Any assumptions that are not satisfied will need to be met by using transformation and filtering to correct the issue. Outliers removal is one of the tools that we will use to correct the data to meet our assumptions. At the end of data exploration stage, our model will fit pre-qualified assumptions in order to perform our data analysis.

First step in our analysis stage is splitting the data tuples into training set and test set. A 60-40 training and test set split is chosen for the abundance of data we obtained. The structure of our data analysis is a two branch approach. We will split our analysis into two separate analysis, models with interaction term and models without interaction term. We will perform identical 5 model selection methods: Forward, Backward, Stepwise, Cp, RADJUSTED in SAS and select the candidate model for the validation section of the analysis for model with interaction term and model without interaction term. The selection of the candidate models out of the results of 5 selection methods will be based on performance metrics such as Adjusted R-Squared, RMSE, GOF, and number of terms. Each of the final model candidate will be re-examined for its model assumption to be met at the end of the analysis stage.

Once the pre-qualified assumptions are met, we will feed our two candidate final models into validation stage.

Validation Methods

At the validation stage, we will use our candidate model and evaluate the training set on the final model candidates. We will get a predicted value for each of the training set tuple. This predicted value will be used to calculate performance metrics such as RADJUSTED, RME, RMSE, CV-R2, and number of terms. Comparison will be drawn between final model candidate with interaction term, without interaction term, and their respective performance on training set. We will finalize on a single final model by examining these metrics. We will use ease of use, cohesive of terms, and domain knowledge to determine the final model.

Specific effect of each term will be examined on the final model. Variation of a prediction will be examined to evaluate the properties of the model to understand the model intuitively.

ANALYSIS, RESULTS AND FINDINGS

For organization purposes our analysis, results, and findings will be grouped into a series of steps.

Step One: Analyze the distribution of the dependent variable.

For the first step of the data exploration stage we examined the distribution of price as the dependent variable by visualizing it as a histogram.

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	22.0000	-81.6117
5.0	31.0000	-16.7990
10.0	42.0000	17.7525
25.0	65.0000	75.4864
50.0	119.0000	139.6330
75.0	188.0000	203.7796
90.0	283.0000	261.5135
95.0	319.0000	296.0650
99.0	450.0000	360.8777

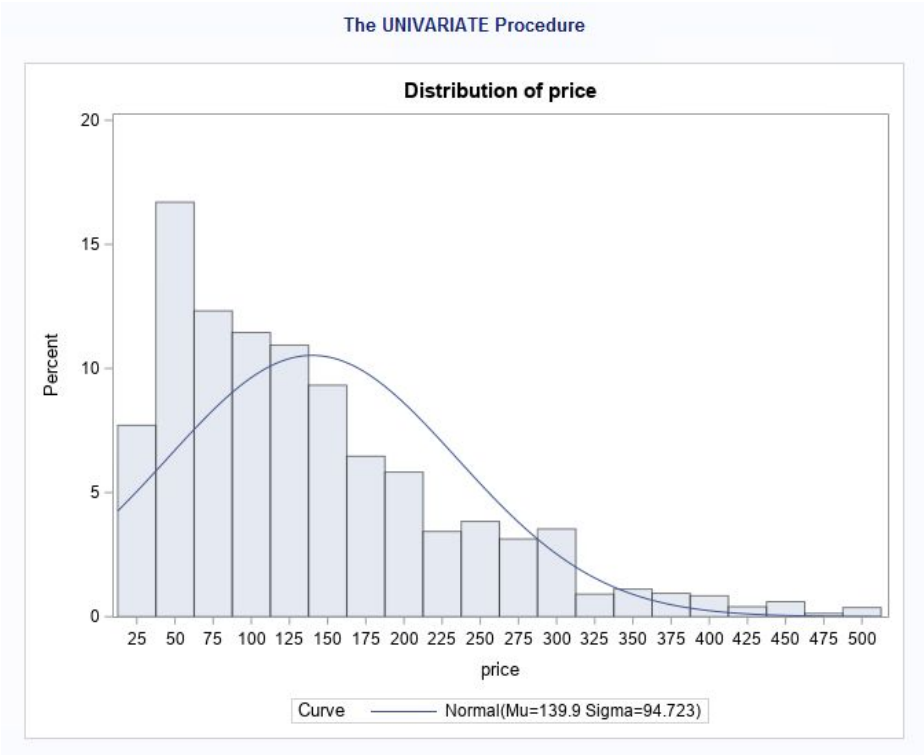


Figure 1.1

Price Distribution

Figure 1.1 shows that the price of airbnb room per night is highly skewed to the right. As we can see from the figure above, the median is 119.0, which is far from the mean of 139.9. This does not fit the normal assumption of linear regression. To correct this we elected to transform *Price* with the log function*. ln(price).*

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	3.09104	3.04629
5.0	3.43399	3.53179
10.0	3.73767	3.79061
25.0	4.17439	4.22308
50.0	4.77912	4.70359
75.0	5.23643	5.18410
90.0	5.64542	5.61658
95.0	5.76519	5.87539
99.0	6.10925	6.36090

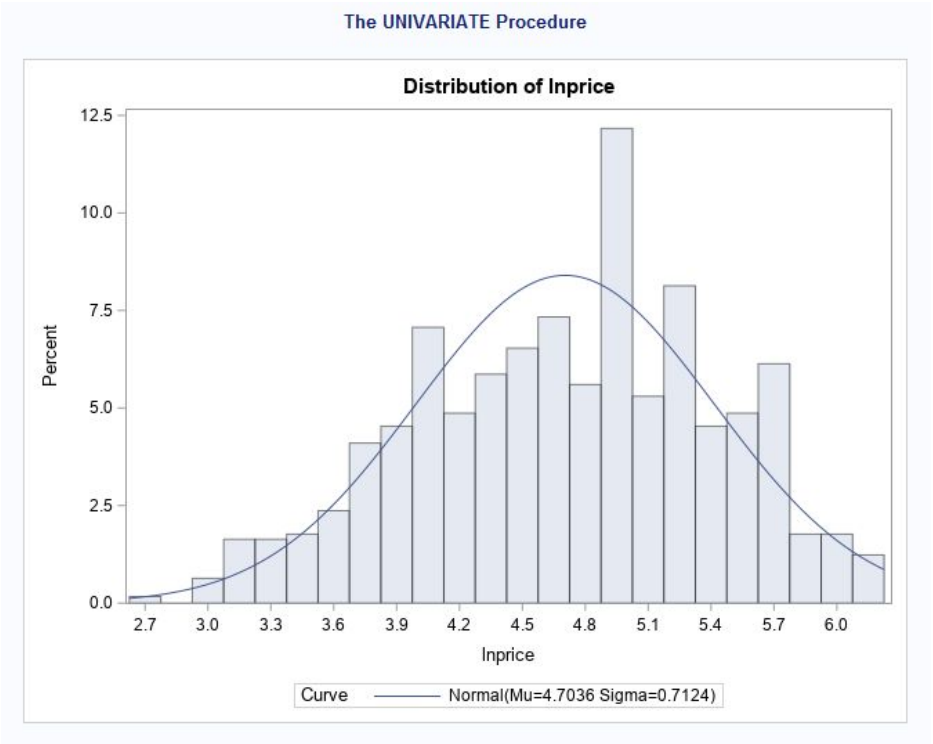


Figure 1.2

LnPrice Distribution

The median and mean are now much closer at 4.779 and 4.703 respectively. Transforming Price to *ln(price)* in Figure 1.2 yields a normal distribution for *ln(price)*.

Further exploring the SAS print out for the independent variables has shown that some independent variables such as *bathrooms*, *accommodation*, *bedrooms*, and *number_of_reviews* all are skewed to the right heavily. We will further investigate these independent variables' distributions after generating the full model.

Step two: Analyze the relationships between price and the independent variables

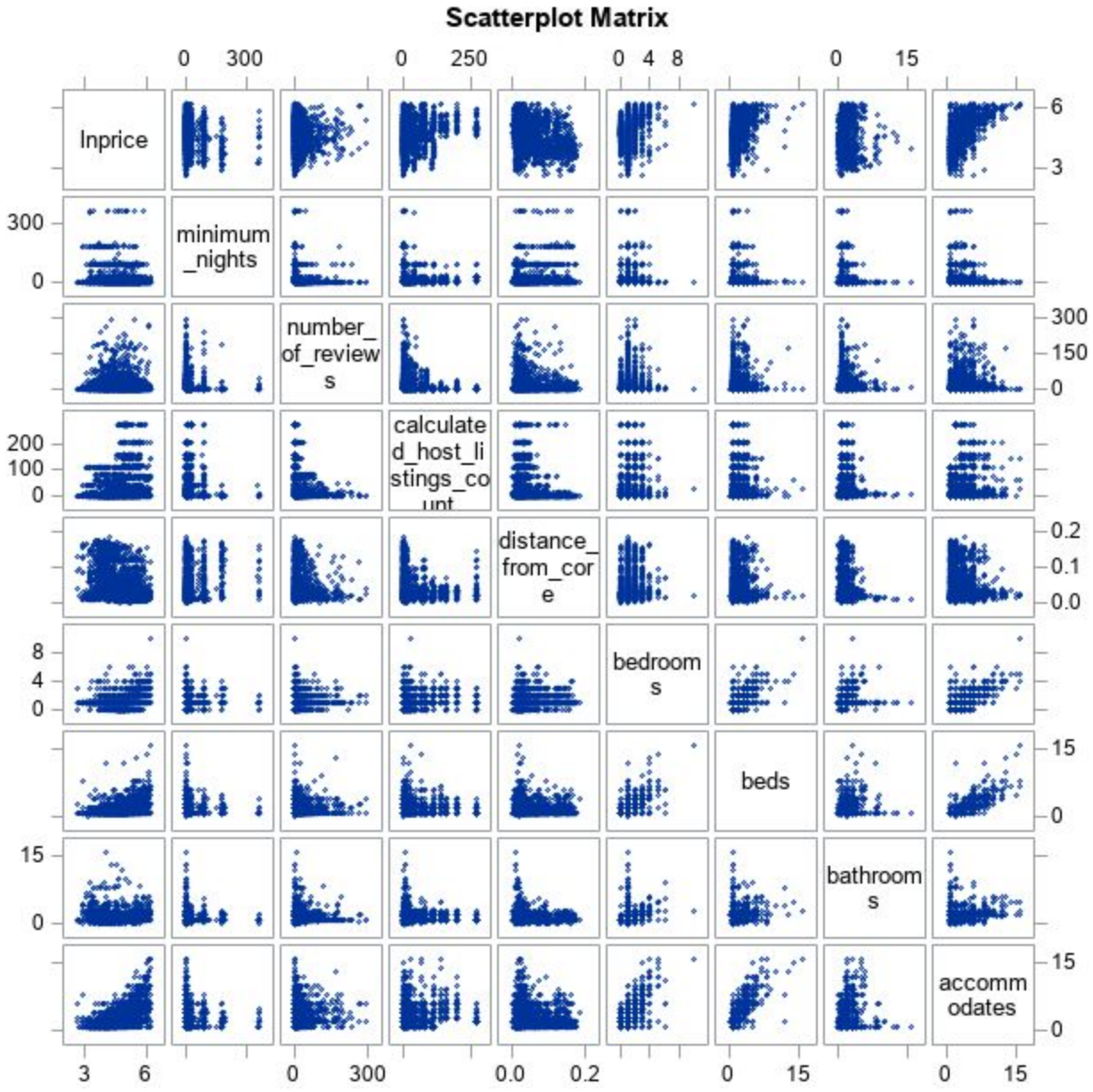


Figure 1.3

Scatter Plot Matrix

After examining the normality of our variables, we generated a scatterplot matrix, Figure 1.3, of all non-categorical or dummy variables. The matrix revealed there are some extreme observations or outliers that may need some further attention.

The matrix also revealed that there may be multicollinearity between *bedrooms*, number of *beds*, *bathrooms*, and *accommodates*.

The matrix shows that a few of the independent variables show correlation with the dependent variable, price. Interpreting relationships through scatter plot is often subjective as it lacks quantitative judgements. Bearing this in mind, we continued to analyze the relationship between the variables, and searched for multicollinearity by examining the Pearson Correlation Coefficients. Using domain knowledge, we grouped attributes into types in order to reduce the large combination count of attributes pairs.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations							
	price	availability_365	minimum_nights	bedrooms	beds	bathrooms	accommodates
price	1.00000 2970	0.09492 <.0001 2970	-0.11543 <.0001 2970	0.54487 <.0001 2969	0.53891 <.0001 2965	0.14589 <.0001 2965	0.69305 <.0001 2970
availability_365	0.09492 <.0001 2970	1.00000 2970	0.17203 <.0001 2970	0.05266 0.0041 2969	0.06609 0.0003 2965	0.12938 <.0001 2965	0.06353 0.0005 2970
minimum_nights	-0.11543 <.0001 2970	0.17203 <.0001 2970	1.00000 2970	-0.04271 0.0199 2969	-0.07327 <.0001 2965	-0.06979 0.0001 2965	-0.13148 <.0001 2970
bedrooms	0.54487 <.0001 2969	0.05266 0.0041 2969	-0.04271 0.0199 2969	1.00000 2969	0.66680 <.0001 2964	0.31912 <.0001 2964	0.70485 <.0001 2969
beds	0.53891 <.0001 2965	0.06609 0.0003 2965	-0.07327 <.0001 2965	0.66680 <.0001 2964	1.00000 2965	0.28622 <.0001 2962	0.80121 <.0001 2965
bathrooms	0.14589 <.0001 2965	0.12938 <.0001 2965	-0.06979 0.0001 2965	0.31912 <.0001 2964	0.28622 <.0001 2962	1.00000 2965	0.23214 <.0001 2965
accommodates	0.69305 <.0001 2970	0.06353 0.0005 2970	-0.13148 <.0001 2970	0.70485 <.0001 2969	0.80121 <.0001 2965	0.23214 <.0001 2965	1.00000 2970

Figure 1.4

Pearson Correlation Room Info

Figure 1.4 shows that there are strong correlations between *beds*, *bedrooms* and accommodates. These correlations confirms the earlier observations made when observing Figure 1.3. Given the correlation between price and other independent variables and an understanding of how the variables, (*beds*, *bedrooms* and *accommodates*), are calculated, we

can remove the variables *beds*, and *bedrooms* since they are collinear with *accommodates*. We chose to keep *accommodates* since it has the higher correlation with the dependent variable than *beds* or *bedrooms.*

Pearson Correlation Coefficients, N = 2940 Prob > |r| under H0: Rho=0				
	price	latitude	longitude	distance_from_core
price	1.00000	-0.17827 <.0001	0.03857 0.0365	-0.21630 <.0001
latitude	-0.17827 <.0001	1.00000	-0.03732 0.0430	0.75768 <.0001
longitude	0.03857 0.0365	-0.03732 0.0430	1.00000	-0.34936 <.0001
distance_from_core	-0.21630 <.0001	0.75768 <.0001	-0.34936 <.0001	1.00000

Figure 1.5

Pearson Correlation Geographical Information

Distance from core is derived information from *latitude* and *longitude*, it represents the Euclidean Distance to the city core, i.e. *distance_from_core*. Therefore we can determine intuitively that there is some collinearity. The exact collinearity relationship between average price by region and geographic locations is unclear. As a result all variables will be examined closely in the full model.

Pearson Correlation Coefficients, N = 2970 Prob > |r| under H0: Rho=0					
	price	number_of_reviews	last_review_new	reviews_per_month_new	calculated_host_listings_count
price	1.00000	-0.04309 0.0188	-0.08125 <.0001	0.02476 0.1773	0.19594 <.0001
number_of_reviews	-0.04309 0.0188	1.00000	-0.09158 <.0001	0.68076 <.0001	-0.14753 <.0001
last_review_new	-0.08125 <.0001	-0.09158 <.0001	1.00000	-0.18109 <.0001	-0.11857 <.0001
reviews_per_month_new	0.02476 0.1773	0.68076 <.0001	-0.18109 <.0001	1.00000	-0.18378 <.0001
calculated_host_listings_count	0.19594 <.0001	-0.14753 <.0001	-0.11857 <.0001	-0.18378 <.0001	1.00000

Figure 1.6

Pearson Correlation Review Information

number_of_reviews is moderately to highly correlated with *reviews_per_month*. Intuitively we can ascertain that both variables are closely related and there is some collinearity. In this case, *number_of_reviews* total contains *reviews_per_month* information, so we will drop *reviews_per_month* from the independent variable list.

The original flow of our modeling process did not consider there may be a significant interaction term in our model. After some careful consideration between our list of independent variables, we suspect that there may be an interaction term between the number of bathroom and accommodation.

We suspect that a Airbnb listing may, at least in part, derive value from both the number of bathrooms the unit has and the amount of people the listing can accommodate. Generally, the larger and more luxurious units will have a higher bathroom and accomodation count. We suspect that the size of the unit, in square feet, is correlated with with the interactive term. However, Airbnb does not require hosts to provide this information about their units and most entries in our dataset don't have this information. So, we will be using the interaction term, *accommodates*bathrooms*

For the purpose of referencing the large set of models we will use the following indexing convention: Model 1.x represents models without the interaction term. Model 2.x represents models with the interaction term.

Before splitting the 3000 rows in the data set into training set and test set, we will first derive the full model and examine if the independent variables meets our model assumptions. We will examine these model assumptions based on the full model with interaction term. This model is defined as Model 2.1.

Based on previous data exploration procedures, we have full model for airbnb defined as:

> *ln(price) = B0 + B1*latitude + B2*longitude + B3*high_region_by_price + B4*medium_region_by_price + B5*distance_from_core + B6*availability_365 + B7*minimum_nights + B8*bathrooms + B9*accommodates + B10*number_of_reviews + B11*last_review_new + B12*Entire_home_apt + B13*Shared_room + B14*bathroom_accommodates*

Base case equation is encoded with dummy variable low_region_by_price for neighborhood variable and Private_room variable for the room type.

Dummy variables encoding for room_type:

> *Private_room: base case {The room is listed as a Private_room}*
>
> *Shared_room: {Shared_room = 1: The room is listed as shared room under room_type, Shared_room = 0: Not a shared room}*
>
> *Entire_home_apt: {Entire_home_apt = 1: The room is listed as entire home apt under room_type, Entire_home_apt = 1: Not listed as entire home/apt}*

Dummy variables for encoding for region_by_price:

Low_region_by_price: base case { This listing is in the neighborhood with average Airbnb price under $100}

Medium_region_by_price: { Medium_region_by_price = 1: This listing is in the neighborhood with average Airbnb price between $100 and $160,
Medium_region_by_price = 0: Listing is either higher $160 or lower than $100 }

High_region_by_price: { High_region_by_price = 1: This listing is in the neighborhood with average Airbnb price higher than $160,
High_region_by_price = 0: This listing is not in the neighborhood with average Airbnb price higher than $160}

Model Assumptions

In order for linear regression to be useful, assumptions about the data collected from Airbnb must be met.

Multicollinear check with VIF

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	10.86650	21.08773	0.52	0.6064	0	0
latitude	1	-2.52281	0.43380	-5.82	<.0001	-0.10994	3.11215
longitude	1	-0.03502	0.20480	-0.17	0.8643	-0.00218	1.41477
high_region_by_price	1	0.26090	0.04194	6.22	<.0001	0.14326	4.61755
medium_region_by_price	1	0.06934	0.03780	1.83	0.0667	0.04478	5.19100
distance_from_core	1	0.21982	0.39372	0.56	0.5767	0.01178	3.87764
availability_365	1	0.00037248	0.00005583	6.67	<.0001	0.07624	1.13742
minimum_nights	1	-0.00129	0.00019556	-6.62	<.0001	-0.07535	1.12900
bathrooms	1	0.00013005	0.01036	0.01	0.9900	0.00020737	2.37755
accommodates	1	0.12718	0.00799	15.92	<.0001	0.38962	5.21899
number_of_reviews	1	-0.00126	0.00025659	-4.92	<.0001	-0.05393	1.04745
last_review_new	1	-0.00000659	0.00002758	-0.24	0.8113	-0.00263	1.05720
Entire_home_apt	1	0.54749	0.01940	28.23	<.0001	0.38524	1.62209
Shared_room	1	-0.60020	0.03705	-16.20	<.0001	-0.19589	1.27346
bathroom_accomdate	1	0.00073319	0.00272	0.27	0.7877	0.00722	6.26552

Figure 2.1

Full Model 2.1

Multicollinearity was suspected in previous section of the data exploration phase when examining the correlation coefficient plot. With our SAS output from the full Model 2.1, we verified as shown in Figure 2.1 that all VIF are below the acceptable threshold value of 10.

With the residual plot from non-categorical independent variables from our full Model 2.1, we can check if our full model meets the linear and constant variance requirement for linear regression.

Constant Variance and Linear

The full set of residual plots can be found in the appendix. The following residual plots for bathrooms, accommodates, last_review_new, and bathroom_accomdate all shows strong sign of non-linear and non-constant variance. In fact, all of these residual plots have a funnel shape which resembles that they may need a log transformation.

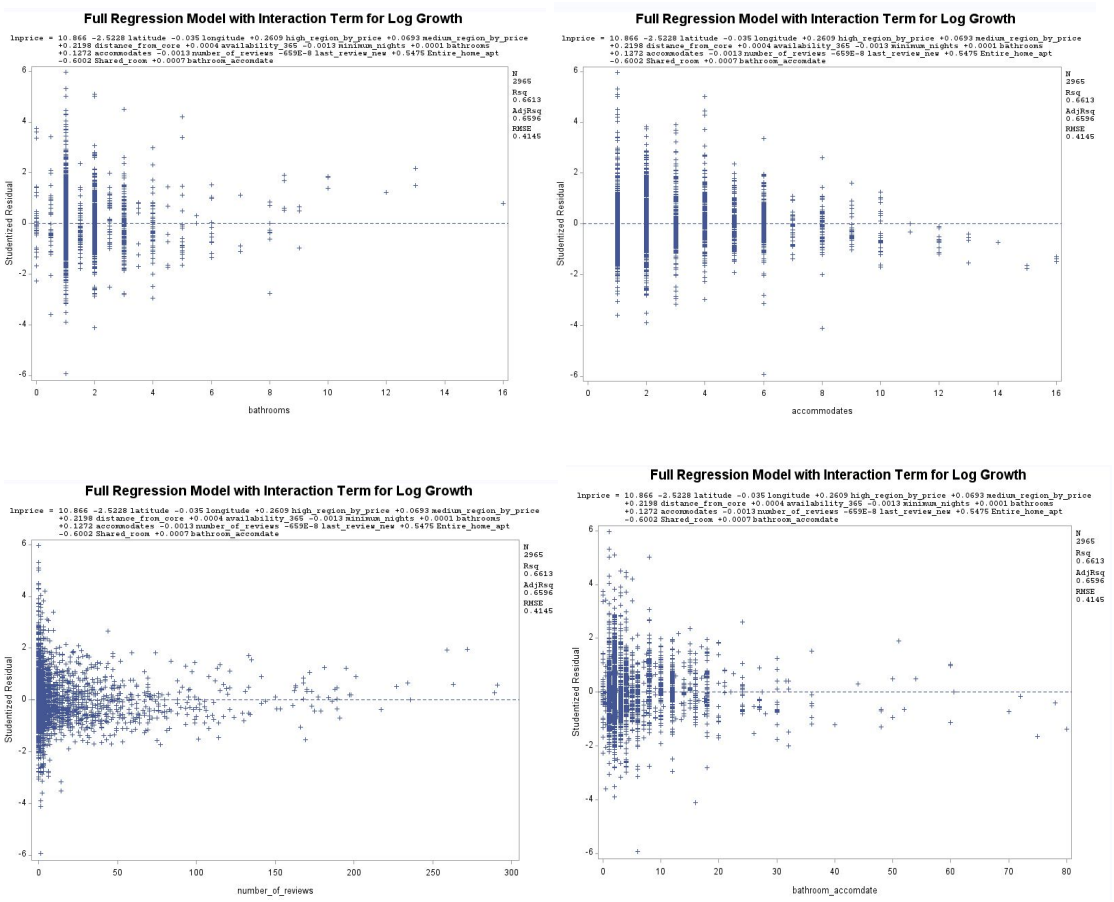


Figure 2.2

Full Model 2.1 Residuals

We transformed these independent variables with the natural log function and compared the MSE and Adj-Rsquared values of the transformed and original full model.

Before transformation

Root MSE	0.41450	R-Square	0.6613
Dependent Mean	4.70743	Adj R-Sq	0.6596
Coeff Var	8.80516		

After Transformation

Root MSE	0.37290	R-Square	0.7300
Dependent Mean	4.68713	Adj R-Sq	0.7280
Coeff Var	7.95591		

Figure 2.3

Full Model 2.1 Transformed

We can see from Figure 2.3, transforming the independent variables significantly reduced the RMSE from .41450 down to .37290 in the transformed full model and increased Adjusted R-squared from 0.6596 to 0.7280.

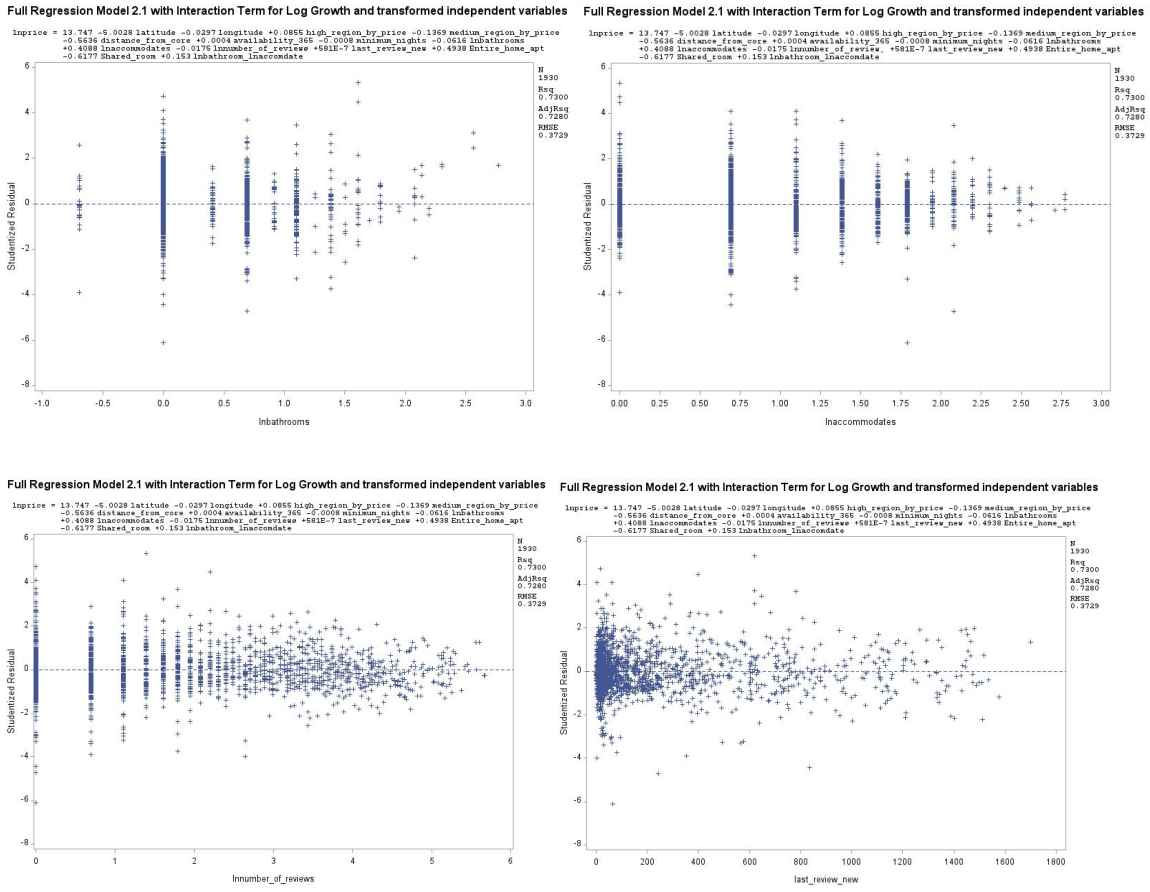


Figure 2.4

Full Model 2.1 Residuals After Transformation

As we can see from the newly transformed full model in Figure 2.4, the constant variance of residuals has been significantly increased. Funnel shapes has been eliminated.

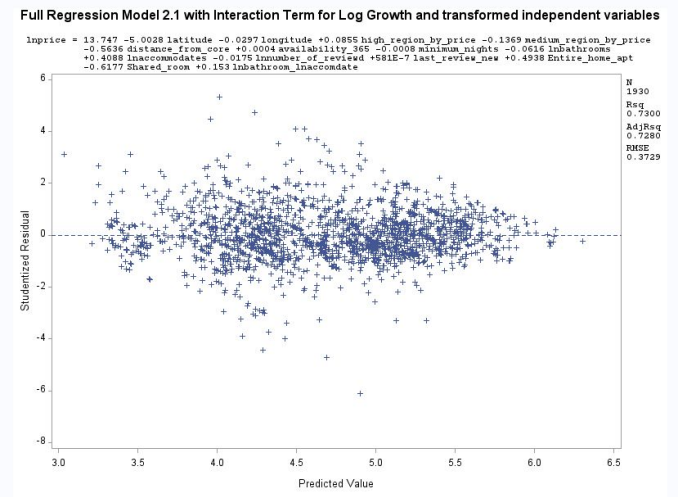


Figure 2.5

Full Model 2.1 Predicted Residual

A final look on the predicted value residual in Figure 2.5 reveals that the variance is fairly constant. Linearity at the same time is met. However, we do see some outliers and point of influence that may need to be examined carefully for a more robust model.

Normality

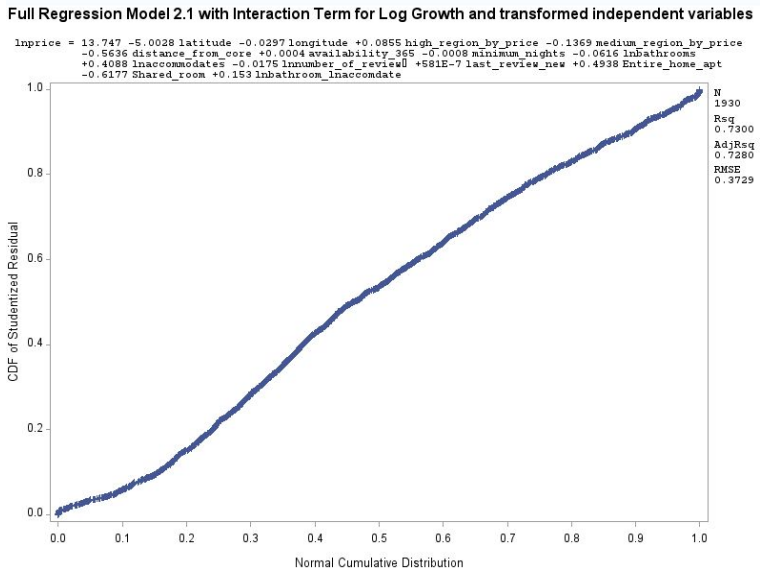


Figure 2.6

Full Model 2.1 Residual vs Normal Distribution

With our full Model 2.1 entered into SAS, we plotted the studentized residual again the normal distribution curve as shown in Figure 2.6. There is a mild S curve, this may be caused by

outliers that remain in the model that are not entry errors. Overall, the data set does satisfy the normal distribution assumption.

Remove Outliers and Influential Points

To further perfect our model we identified and removed 34 suspect outliers and influential points based on SAS Proc Reg influence analysis. This analysis determines outlier and point of influence based on r-student standardized residual and Cook's D. The observation index for this set of data point removal are: 2915, 2674, 2650, 2425, 2423, 2420, 2415, 2386, 2360, 2172, 2151, 2116, 2105, 2104, 2085, 2030, 1970, 1768, 1624,1318, 1190, 1187, 1177, 1165, 1154, 1144, 1060, 904, 481, 342, 201, 194, 95, 93. The removal of outliers and influential had a large positive impact on our model increasing Adj R^2 from .728 to .7745 and lowering Root MSE from .37290 to .33334.

Before removal

Root MSE	0.37290	R-Square	0.7297
Dependent Mean	4.68713	Adj R-Sq	0.7280
Coeff Var	7.95576		

After removal

Root MSE	0.33334	R-Square	0.7760
Dependent Mean	4.69016	Adj R-Sq	0.7745
Coeff Var	7.10726		

Figure 2.7

Full Model 2.1 Performance Before Removal **Full Model 2.1 Performance After Removal**

Full model 2.1 after checking assumptions:

*ln(price) = B0 + B1*latitude + B2*longitude + B3*high_region_by_price + B4*medium_region_by_price + B5*distance_from_core + B6*availability_365 + B7*minimum_nights + B8*lnbathrooms + B9*lnaccommodates + B10*lnnumber_of_reviews + B11*last_review_new + B12*Entire_home_apt + B13*Shared_room + B14*lnbathroom_lnaccomdate*

Data Analysis

Data Set Split for Validation

In order to ensure that the model we fit with our dataset is valid, we will run our model analysis only on the training set of data. This means we will need to split our cleaned data set with outliers removed and variable transformed into a training set and test set.

With 3000 entries in our dataset we can afford to use 100 attributes assuming that each attribute will require 30 data entries. Since our full model only contains 13 attributes, we have

the luxury to split the data where training set and test set have similar amount of rows. We chose a 60% to 40% between training and test set. 60% of the data set will be for training the model. 40% of the data will be for validating the model.

In the analysis phase of this project, we will be only using training set.

There is a lack of domain knowledge in what effect will the interaction term lnbathroom_lnaccomdate will bring into the full model. In order to fully understand the effect of this interaction term, we will independently run model selection process on the full model without interaction term, Model 1.1 and the full model with interaction term ln(bathrooms)*ln(accommodates) Model 2.1. We will run these two variations with 5 model selection methods each. The 5 model selection methods we have chosen are backward, forward, stepwise, Cp, and AdjRsquare. At the end of each attribute selection analysis, we will choose one candidate for the final model for Model 1 with interaction term and without interaction term.

Attribute selection:

Model 1.2 BACKWARD SELECTION without INTERACTION TERM

The results from the backward selection method on full Model 1.1 is attached in Appendix D.

Root MSE	0.33332	R-Square	0.7764
Dependent Mean	4.67884	Adj R-Sq	0.7748
Coeff Var	7.12398		

Root MSE	0.37636	R-Square	0.7248
Dependent Mean	4.68713	Adj R-Sq	0.7229
Coeff Var	8.02960		

Figure 3.1

Backward Selection on the left ***Full Model 1.1 output on the right***

Both RMSE and Adj R-Squared value has been significantly improved from the full model.

Goodness-of-fit:

Ho: B1=B2=B3...B14 = 0
Ha: one of the Beta values is not equal to 0
F-value: 490.07
P-value: <0.001

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with ln(Price). Therefore, there is strong support for this model.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	14.53134	23.83259	0.61	0.5421	0	0
latitude	1	-5.04206	0.52245	-9.65	<.0001	-0.20336	3.09129
longitude	1	-0.03725	0.23144	-0.16	0.8722	-0.00228	1.39971
high_region_by_price	1	0.08229	0.05068	1.62	0.1046	0.04426	5.17349
medium_region_by_price	1	-0.14371	0.04567	-3.15	0.0017	-0.09124	5.85322
distance_from_core	1	-0.45838	0.44155	-1.04	0.2993	-0.02378	3.65161
availability_365	1	0.00039612	0.00006465	6.13	<.0001	0.07870	1.14827
minimum_nights	1	-0.00068298	0.00028437	-2.40	0.0164	-0.03131	1.18321
lnbathrooms	1	0.06666	0.01949	3.42	0.0006	0.04652	1.28777
lnaccommodates	1	0.49709	0.01832	27.14	<.0001	0.44552	1.87648
lnnumber_of_reviews	1	-0.02092	0.00616	-3.40	0.0007	-0.04379	1.15674
last_review_new	1	0.00005685	0.00002992	1.90	0.0576	0.02597	1.30125
Entire_home_apt	1	0.48677	0.02233	21.80	<.0001	0.34046	1.69769
Shared_room	1	-0.66514	0.04252	-15.64	<.0001	-0.22117	1.39135

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	12.27546	0.50849	24.14	<.0001	0	.	0
latitude	1	-6.23165	0.38092	-16.36	<.0001	-0.26053	0.78086	1.28064
medium_region_by_price	1	-0.21069	0.02371	-8.89	<.0001	-0.13676	0.83614	1.19597
availability_365	1	0.00032903	0.00007093	4.64	<.0001	0.06690	0.95202	1.05039
lnbathrooms	1	0.06222	0.02337	2.66	0.0079	0.04288	0.76368	1.30945
lnaccommodates	1	0.50660	0.02113	23.97	<.0001	0.46166	0.53390	1.87301
lnnumber_of_reviews	1	-0.02783	0.00680	-4.09	<.0001	-0.05830	0.97570	1.02491
Entire_home_apt	1	0.50737	0.02548	19.91	<.0001	0.36121	0.60175	1.66183
Shared_room	1	-0.67974	0.05105	-13.31	<.0001	-0.22139	0.71631	1.39604

Figure 3.2

Full Model 1.1 on the left ***Model 1.2 on the right***

Backward selection resulted in a model that has all its parameters well within the statically significant level of 0.05. Longitude, high_region_by_price, minimum nights, last_review_new, and distance from core all have been eliminated.

The backward selection method removed longitude and latitude as an important variable that explains the variance in *ln(price)*. From a scatter plot of longitude vs latitude vs. price that was generated during the pre-processing stage in Python, we can see higher priced and denser number of airbnb near the south shore line of Singapore. As we get further south the price tends to increase.

We can also conclude that there is little to no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the performance metrics for Model 1.2, we have R^2= 0.7764 Adj-R^2 = 0.7728 Using Adj-R2 77.28% of the y(Price) in the model can be captured by the predictors. 22.72% is unexplained

Model 1.2 Equation:

ln(Price) = 12.27546 -6.23165(latitude) - 0.21069(medium_region_by_price) + 0.00032903(availability_365) + 0.06222(lnbathrooms) + 0.50066(lnaccommodates) -0.02783(lnnumber_of_reviews) + 0.50737(Entire_home_apt) - 0.67974(Shared_room)

Model 1.3: Forward without Interaction

The results from the forward selection method on full Model 1.1 is attached in Appendix D.

Goodness-of-fit:

Ho: B1=B2=B3...B14= 0

Ha: one of the Beta values is not equal to 0
F-value: 357.44
P-value:<0.001

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

This forward selection model gave us a model where 8 parameters within the significant level of 0.05, and three were not. Comparing it to the full model, Longitude and distance from core all have been eliminated. We will later observe that all our models have eliminated these who as well.

As mentioned previously, this forward selection method removed longitude as an important variable that explains the variance in ln(price). As we get further south the price tends to increase, which ties into latitude measurement.

We can also conclude that there is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 1.3, we have R^2= 0.7774 Adj-R^2 = 0.7752

Root MSE	0.33305	R-Square	0.7774
Dependent Mean	4.67884	Adj R-Sq	0.7752
Coeff Var	7.11816		

Figure 3.3

Model 1.3 Performance

Using Adj-R2 77.74% of the y(Price) in the model can be captured by the predictors. 22.26% is unexplained.

Model 1.3 Equation:

ln(Price) = 11.67520 -5.81639(latitude) + 0.06607(high_region_by_price) - 0.15864(medium_region_by_price) + 0.00035971(availability_365) - 0.00051628 (minimum_nights) + 0.05971(lnbathrooms) + 0.50066(lnaccommodates) - 0.02534(lnnumber_of_reviews) + 0.00004694(last_review_new) + 0.50605(Entire_home_apt) - 0.68915(Shared_room)

In contrast to Model 1.2, this model added several new variables, (i.e. high_region_by_price, minimum_nights and last_review_new). Resulting in Model 1.3 having 11 coefficients.

Model 1.4 Stepwise Without Interaction term

Interestingly Stepwise selection resulted in the exact model as Model 1.2 Backward selection.

Model 1.5 C(p) without Interaction term

The results from the C(p) selection method on full Model 1.1 is attached in Appendix D.

Goodness-of-fit:

Ho: B1=B2=B3...B14= 0

Ha: one of the Beta values is not equal to 0

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

This Cp selection model gave us a model where all the parameters within the significant level of 0.05. Comparing it to the full model, Longitude, last_review_new, high_region_by_price, and distance_from_core all have been eliminated. The same elimination steps from Model 1.2 can be observed here and yield similar findings as well.

We can also conclude that is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 1.5, we have R^2= 0.7764 Adj-R^2 = 0.7748
Using Adj-R2 77.48% of the y(Price) in the model can be captured by the predictors. 22.52% is unexplained.

Model 1.5 Equation:

ln(Price) = 12.27546 - 6.23165(latitude) - 0.21069(medium_region_by_price) + 0.00032903(availability_365) + 0.06222(lnbathrooms) + 0.50660(lnaccommodates) - 0.02783(lnnumber_of_reviews) + 0.50737(Entire_home_apt) -0.67974(Shared_room)

Root MSE	0.33332	R-Square	0.7764
Dependent Mean	4.67884	Adj R-Sq	0.7748
Coeff Var	7.12398		

Figure 3.4

Model 1.5 Performance

Both RMSE and Adj R-Squared value has been significantly improved from the full model.

Model 1.6 Adjusted R^2 without Interaction term

The results from the Adjusted R^2 selection method on full Model 1.1 is attached in Appendix D.

Goodness-of-fit:

F-value: 436.06
P-value: <0.001

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

This ADJ-R2 selection model gave us a model where 8 of the parameters within the significant level of 0.05, and 1 was not. Longitude, last_review_new, and distance_from_core all have been eliminated. We we have observed here that this model also has eliminated the same longitude and distance_from_core as well.

We can also conclude that is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 1.6, we have R^2= 0.7767 Adj-R^2 = 0.7750

Using Adj-R2 77.50% of the y(Price) in the model can be captured by the predictors. 22.50% is unexplained.

Number in Model	Adjusted R-Square	R-Square	Variables in Model
11	0.7752	0.7774	latitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room
10	0.7751	0.7771	latitude medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room
10	0.7750	0.7770	latitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room

Figure 3.5

Model 1.6 Adjusted R-Squared Selection Results

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	435.76757	48.41862	436.06	<.0001
Error	1128	125.24968	0.11104		
Corrected Total	1137	561.01725			

Root MSE	0.33322	R-Square	0.7767
Dependent Mean	4.67884	Adj R-Sq	0.7750
Coeff Var	7.12189		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	11.67035	0.69174	16.87	<.0001	0	.	0
latitude	1	-5.80288	0.50549	-11.48	<.0001	-0.24260	0.44317	2.25648
high_region_by_price	1	0.06847	0.05308	1.29	0.1974	0.03740	0.23537	4.24869
medium_region_by_price	1	-0.15851	0.04689	-3.38	0.0007	-0.10289	0.21369	4.67973
availability_365	1	0.00032393	0.00007102	4.56	<.0001	0.06586	0.94906	1.05367
lnbathrooms	1	0.06049	0.02340	2.59	0.0098	0.04169	0.76118	1.31376
lnaccommodates	1	0.50423	0.02121	23.78	<.0001	0.45950	0.52988	1.88723
lnnumber_of_reviews	1	-0.02799	0.00680	-4.12	<.0001	-0.05863	0.97539	1.02523
Entire_home_apt	1	0.50216	0.02579	19.47	<.0001	0.35750	0.58700	1.70359
Shared_room	1	-0.67958	0.05104	-13.32	<.0001	-0.22134	0.71631	1.39605

Figure 3.6

Model 1.6 Performance

Model 1.6 Equation:

ln(Price) = 11.67035 -5.80288(latitude) + 0.06847(high_region_by_price) - 0.15851(medium_region_by_price) + 0.00032393(availability_365) + 0.06049(lnbathrooms) + 0.50423(lnaccommodates) - 0.02799(lnnumber_of_reviews) + 0.50216(Entire_home_apt) - 0.67958(Shared_room)

In contrast to Model 1.2, 1.4 and 1.5, this model added high_region_by_price, Resulting in Model 1.6 having 9 coefficients.

Model 1 Final Model Candidate without Interaction Term:

Of the 5 attribute selection methods we ran, backward, stepwise, and cp all resulted in a model with the same performance and attributes.

When comparing the RMSE and Adj R^2, the models highlighted are slightly less than its competition but no significant difference since the explained variance difference is only less than 1 percent. These three highlighted models have fewer attributes compared with selection result from forward and Adj R^2. This means the model will be more efficient or faster to execute estimations.

Without Interaction terms				Goodness of Fit				
Selection	RMSE	R2	Adj R2	F Value	Pr>F	Residuals	# Coef	Model
Backward	0.33332	0.7764	0.7748	490.07	<.0001	No violation (Ind, CV, Lin, Norm)	8	1.2
Forward	0.33305	0.7774	0.7752	357.44	<.0001	No violation (Ind, CV, Lin, Norm)	11	1.3
Stepwise	0.33332	0.7764	0.7748	490.07	<.0001	No violation (Ind, CV, Lin, Norm)	8	1.4
CP	0.33332	0.7764	0.7748	490.07	<.0001	No violation (Ind, CV, Lin, Norm)	8	1.5
Adj R2	0.33322	0.7767	0.775	436.06	<.0001	No violation (Ind, CV, Lin, Norm)	9	1.6

Figure 3.7

Model 1 Selection Results Comparison

Model 1

Model 1 Final Equation:

ln(Price) = 12.27546 - 6.23165(latitude) - 0.21069(medium_region_by_price) + 0.00032903(availability_365) + 0.06222(lnbathrooms) + 0.50660(lnaccommodates) - 0.02783(lnnumber_of_reviews) + 0.50737(Entire_home_apt) - 0.679745(Shared_room)

Further examining the model assumptions (constant variance, linear, normality) on Model 1 Figure 3.8 reveals that all previous met assumptions are still met.

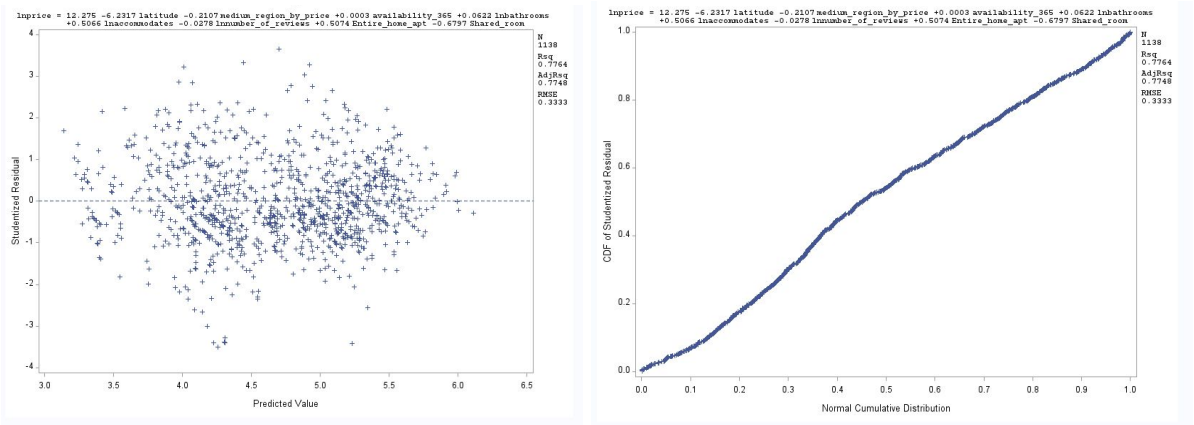


Figure 3.8

Model 1 Residual and Normality

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.27546	0.50849	24.14	<.0001	0	0
latitude	1	-6.23165	0.38092	-16.36	<.0001	-0.26053	1.28064
medium_region_by_price	1	-0.21069	0.02371	-8.89	<.0001	-0.13676	1.19597
availability_365	1	0.00032903	0.00007093	4.64	<.0001	0.06690	1.05039
lnbathrooms	1	0.06222	0.02337	2.66	0.0079	0.04288	1.30945
lnaccommodates	1	0.50660	0.02113	23.97	<.0001	0.46166	1.87301
lnnumber_of_reviews	1	-0.02783	0.00680	-4.09	<.0001	-0.05830	1.02491
Entire_home_apt	1	0.50737	0.02548	19.91	<.0001	0.36121	1.66183
Shared_room	1	-0.67974	0.05105	-13.31	<.0001	-0.22139	1.39604

Figure 3.9

Model 1 VIF

VIF is well within the threshold value of less than 10. Each attributes' parameter p-value are all under statistically significant threshold of 0.05. The residual plot of independent variables are attached in Appendix D.

Model 2.2 Backward with Interaction Term

The results from the Backward selection method with interaction attached in Appendix D.

Goodness-of-fit:

Ho: B1=B2=B3...B14= 0

Ha: one of the Beta values is not equal to 0

F-value: 447.31

P-value: <0.001

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

Backward selection model gave us a model that has 8 of it parameters within the significant level of 0.05, and 1 that is not. From the full model 2.1 longitude, high_region_by_price, minimum_nights, last_review_new, and distance_from_core all have been eliminated. Interestingly, as we saw in our model selections without interaction, this model that include the interaction term also has eliminated the longitude and distance_from_core as well. We observed that the interaction did turn out to be significant and it was not eliminated in this selection method.

We can also conclude that there is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 2.2, we have R^2= 0.7811 Adj-R^2 = 0.7794
Using Adj-R2 77.94% of the y(Price) in the model can be captured by the predictors. 22.06% is unexplained

Root MSE	0.32993	R-Square	0.7811
Dependent Mean	4.67884	Adj R-Sq	0.7794
Coeff Var	7.05159		

Figure 4.1

Model 2.2 Performance

Model 2.2 Equation:

ln(Price) = 12.35885 - 6.26269(latitude) - 0.20586(medium_region_by_price) + 0.00034182(availability_365) - 0.06460 (lnbathrooms) + 0.42478(lnaccommodates) - 0.02384(lnnumber_of_reviews) + 0.51392(Entire_home_apt) - 0.62372(Shared_room) + 0.14894(lnbathroom_lnaccomdate)

Model 2.3 Forward with Interaction

The results from the Forward selection method with interaction is attached in Appendix D.

Goodness-of-fit

Ho: B1=B2=B3...B14= 0
Ha: one of the Beta values is not equal to 0
F-value: 337.01
P-value:<0.001

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

The forward selection model gave us a model that has 10 of it parameters within the significant level of 0.05, and 2 are not. From model 2.1 Longitude and distance_from_core have both been eliminated. We observed that the interaction did turn out to be significant and it was not eliminated in this selection method.

We can also conclude that is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 2.3, we have R^2= 0.7824 Adj-R^2 = 0.7800

Using Adj-R2 78% of the y(Price) in the model can be captured by the predictors. 22% is unexplained

Root MSE	0.32944	R-Square	0.7824
Dependent Mean	4.67884	Adj R-Sq	0.7800
Coeff Var	7.04112		

Figure 4.2

Model 2.3 Performance

Model 2.3 Equation:

ln(Price) = 11.70374-5.80704(latitude) + 0.07162(high_region_by_price) - 0.14941(medium_region_by_price) + 0.00037739(availability_365) - 0.07107(lnbathrooms) + 0.41505(lnaccommodates) - 0.02113(lnnumber_of_reviews) + 0.51300(Entire_home_apt) - 0.63376(Shared_room) + 0.14894(lnbathroom_lnaccomdate)

Model 2.4 Stepwise with Interaction

The results from the Stepwise selection method with interaction is attached in Appendix D.

Goodness-of-fit

Ho: B1=B2=B3...B14= 0

Ha: one of the Beta values is not equal to 0

F-value: 403.37

P-value: <0.001

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

The stepwise selection model gave us a model that has 9 of it parameters within the significant level of 0.05, and 1 that is not. Comparing from model 2.1 Longitude, high_region_by_price, last_review_new, and distance_from_core have all been eliminated. We observed that the interaction did turn out to be significant and it was not eliminated in this selection method.

We can also conclude that is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 2.4 we have R^2= 0.7816 Adj-R^2 = 0.7797
Using Adj-R2 77.97% of the y(Price) in the model can be captured by the predictors. 22.03% is unexplained

Root MSE	0.32971	R-Square	0.7816
Dependent Mean	4.67884	Adj R-Sq	0.7797
Coeff Var	7.04687		

Figure 4.3

Model 2.4 Performance

Model 2.4 Equation:

ln(Price) = 12.30822 - 6.21633(latitude) - 0.20638(medium_region_by_price) + 0.00035924(availability_365) - 0.0052052(minimum_nights) - 0.06610 (lnbathrooms) + 0.41726(lnaccommodates) - 0.02433(lnnumber_of_reviews) + 0.51659(Entire_home_apt) - 0.63473(Shared_room) + 0.15189(lnbathroom_lnaccomdate)

Model 2.5 C(p) with Interaction

The results from the C(p) selection method with interaction is attached in Appendix D.

Goodness-of-fit

Ho: B1=B2=B3...B14= 0

Ha: one of the Beta values is not equal to 0

F-value: 337.01

P-value:<0.001

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

The Cp selection model gave us a model that has 10 of it parameters within the significant level of 0.05, and 2 that are not. Here, only Longitude and distance_from_core have both been eliminated. We observed that the interaction did turn out to be significant and it was not eliminated in this selection method. We did observe here similar findings to model 2.3.

We can also conclude that is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 2.5, we have R^2= 0.7824 Adj-R^2 = 0.7800
Using Adj-R2 78% of the y(Price) in the model can be captured by the predictors. 22% is unexplained

The third row model is chosen because Cp value is significantly less than Number of variable k + 1 = 13. Upon reviewing the Cp model selection method threshold value for Cp, this may not be the most optimal result in model selection. The model from row 1 already met this threshold and has higher R Squared value with lower Cp value. For the purpose of consistency with the

rest of the analysis data printout, we will continue the analysis with Model from row 3 as highlighted in Figure 4.4.

Number in Model	C(p)	R-Square	Variables in Model
10	11.0690	0.7816	latitude medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	11.0754	0.7820	latitude medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
12	11.2329	0.7824	latitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	11.3772	0.7819	latitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
10	11.5438	0.7815	latitude high_region_by_price medium_region_by_price availability_365 lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
9	11.5819	0.7811	latitude medium_region_by_price availability_365 lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
12	12.2103	0.7822	latitude medium_region_by_price distance_from_core availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	12.3371	0.7818	latitude medium_region_by_price distance_from_core availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	12.5526	0.7817	latitude high_region_by_price medium_region_by_price availability_365 lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate

Figure 4.4

C(p) Selection Result

Root MSE	0.32944	R-Square	0.7824
Dependent Mean	4.67884	Adj R-Sq	0.7800
Coeff Var	7.04112		

Figure 4.5

C(p) Selection Performance

Model 2.5 Equation:

ln(Price) = 11.70374 - 5.80704(latitude) + 0.07162(high_region_by_price) - 0.14941(medium_region_by_price) + 0.00037739(availability_365) - 0.07107 (lnbathrooms) + 0.41505(lnaccommodates) - 0.02113(lnnumber_of_reviews) + 0.0005027(last_review_new) + 0.51300(Entire_home_apt) - 0.63376(Shared_room) + 0.15353(lnbathroom_lnaccomdate)

Model 2.6 Adjust R^2 with Interaction

Goodness-of-fit

Ho: B1=B2=B3...B14= 0

Ha: one of the Beta values is not equal to 0

F-value: 403.37

P-value:<0.001.

Conclusion:

The P-value is very small. Much less than .05, therefore we can reject the null hypothesis and go with the alternate hypothesis. At least one or more predictor is significantly associated with y(Price). There is strong support for this model.

The ADJ_R2 selection model gave us a model that has 10 of it parameters within the significant level of 0.05, and 2 are not. Longitude, high_region_by_price, ln_num_of_reviews, and distance_from_core have both been eliminated. We observed that the interaction did turn out to

be significant and it was not eliminated in this selection method. These results are very similar to the findings we observed in Model 2.4

We can also conclude that there is no multicollinearity between Price and the other variables factors, since the VIF values are less than 10.

Looking at the values for Model 2.6, we have R^2= 0.7816 Adj-R^2 = 0.7797
Using Adj-R2 77.97% of the y(Price) in the model can be captured by the predictors. 22.03% is unexplained

Number in Model	Adjusted R-Square	R-Square	Variables in Model
12	0.7800	0.7824	latitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
13	0.7799	0.7824	latitude high_region_by_price medium_region_by_price distance_from_core availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	0.7799	0.7820	latitude medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
12	0.7798	0.7822	latitude medium_region_by_price distance_from_core availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
13	0.7798	0.7824	latitude longitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	438.50109	43.85011	403.37	<.0001
Error	1127	122.51616	0.10871		
Corrected Total	1137	561.01725			

Root MSE	0.32971	R-Square	0.7816
Dependent Mean	4.67884	Adj R-Sq	0.7797
Coeff Var	7.04687		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	12.30822	0.50429	24.41	<.0001	0	.	0
latitude	1	-6.21633	0.37799	-16.45	<.0001	-0.25989	0.77597	1.28871
medium_region_by_price	1	-0.20638	0.02347	-8.79	<.0001	-0.13396	0.83453	1.19828
availability_365	1	0.00035924	0.00007107	5.05	<.0001	0.07304	0.92799	1.07760
minimum_nights	1	-0.00052052	0.00032837	-1.59	0.1132	-0.02298	0.92240	1.08413
lnbathrooms	1	-0.06610	0.03458	-1.91	0.0562	-0.04556	0.34106	2.93200
lnaccommodates	1	0.41726	0.02711	15.39	<.0001	0.38025	0.31759	3.14869
lnnumber_of_reviews	1	-0.02433	0.00678	-3.59	0.0003	-0.05097	0.95984	1.04184
Entire_home_apt	1	0.51659	0.02530	20.42	<.0001	0.36778	0.59742	1.67386
Shared_room	1	-0.63473	0.05222	-12.15	<.0001	-0.20673	0.66977	1.49305
lnbathroom_lnaccomdate	1	0.15189	0.03025	5.02	<.0001	0.14619	0.22857	4.37502

Figure 4.6

Model 2.6 Adj R-Squared Result and Model Performance

Model 2.6 Equation:

ln(Price) = 12.30822 - 6.21633(latitude) - 0.20638(medium_region_by_price) + 0.00035924(availability_365) - 0.06610 (lnbathrooms) + 0.41726(lnaccommodates) - 0.02433(lnnumber_of_reviews) + 0.51659(Entire_home_apt) - 0.63473(Shared_room) + 0.15189(lnbathroom_lnaccomdate)

Model 2: Final Model Candidate with Interaction Term

Of the five selection methods we ran, backwards, (i.e. Model 2.2), selection produced the best model, while using the fewest coefficients. When comparing the RMSE and Adj R^2, Model 2.2 is slightly less than the others. However, since the explained variance difference is only less than 1 percent Model 2.2 having a lower RMSE and Adj-R2 is negligible.

With interaction terms				Goodness of Fit				
Selection	RMSE	R2	Adj-R2	F Value	Pr>F	Residuals	Coef	Model
Backward	0.32993	0.7811	0.7794	447.31	<.0001	No violation (Ind, CV, Norm, Line)	9	2.2
Forward	0.32994	0.7824	0.78	337.01	<.0001	No violation (Ind, CV, Norm, Line)	12	2.3
Stepwise	0.32971	0.7816	0.7797	403.37	<.0001	No violation (Ind, CV, Norm, Line)	10	2.4
CP	0.32994	0.7824	0.78	337.01	<.0001	No violation (Ind, CV, Norm, Line)	12	2.5
Adj-R2	0.32971	0.7816	0.7797	403.37	<.0001	No violation (Ind, CV, Norm, Line)	10	2.6

Figure 4.7

Model 2 Comparison

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	438.22792	48.69199	447.31	<.0001
Error	1128	122.78933	0.10886		
Corrected Total	1137	561.01725			

Root MSE	0.32993	R-Square	0.7811
Dependent Mean	4.67884	Adj R-Sq	0.7794
Coeff Var	7.05159		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.35885	0.50361	24.54	<.0001	0	0
latitude	1	-6.26269	0.37711	-16.61	<.0001	-0.26182	1.28099
medium_region_by_price	1	-0.20586	0.02349	-8.76	<.0001	-0.13363	1.19805
availability_365	1	0.00034182	0.00007026	4.86	<.0001	0.06950	1.05183
lnbathrooms	1	-0.06460	0.03459	-1.87	0.0621	-0.04452	2.92981
lnaccommodates	1	0.42478	0.02671	15.91	<.0001	0.38710	3.05228
lnnumber_of_reviews	1	-0.02384	0.00678	-3.52	0.0005	-0.04995	1.03971
Entire_home_apt	1	0.51392	0.02526	20.35	<.0001	0.36588	1.66645
Shared_room	1	-0.62372	0.05179	-12.04	<.0001	-0.20314	1.46664
lnbathroom_lnaccomdate	1	0.14894	0.03022	4.93	<.0001	0.14335	4.35849

Figure 4.8

Model 2 Performance and VIF

Model 2 Final Equation:

ln(Price) = 12.35885 - 6.26269 (latitude) - 0.20586 (medium_region_by_price) + 0.00034182 (availability_365) - 0.06460 (lnbathrooms) + 0.42478 (lnaccommodates) - 0.02384 (lnnumber_of_reviews) + 0.51392 (Entire_home_apt) - 0.62372 (Shared_room) + 0.14894 (lnbathroom_lnaccomdate)

We observed that Model 1 has a more significant Beta parameter value for lnbathrooms than Model 2. This is very likely due to the introduction of the interactive term since the only difference between Model 1 and Model 2 is the presence of interaction term. We will leave lnbathrooms as it is on the borderline of being significant and reserve our judgement until we compare the two models in test set during validation.

Model 2 SAS outputs are attached in Appendix D.

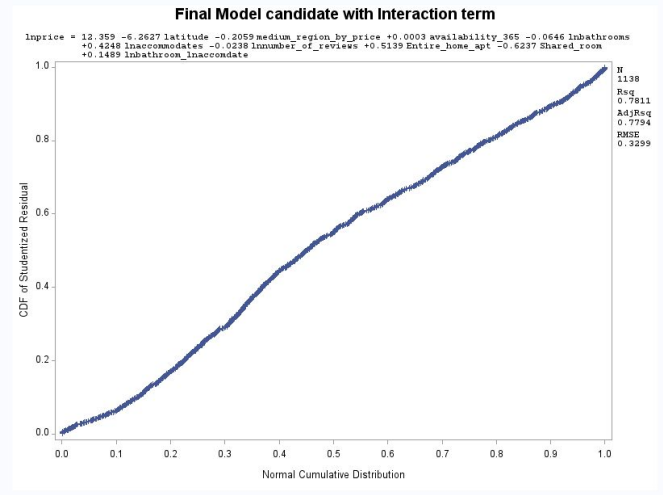


Figure 4.9

Model 2 Normality

At this stage of data analysis, the normality remains acceptable.

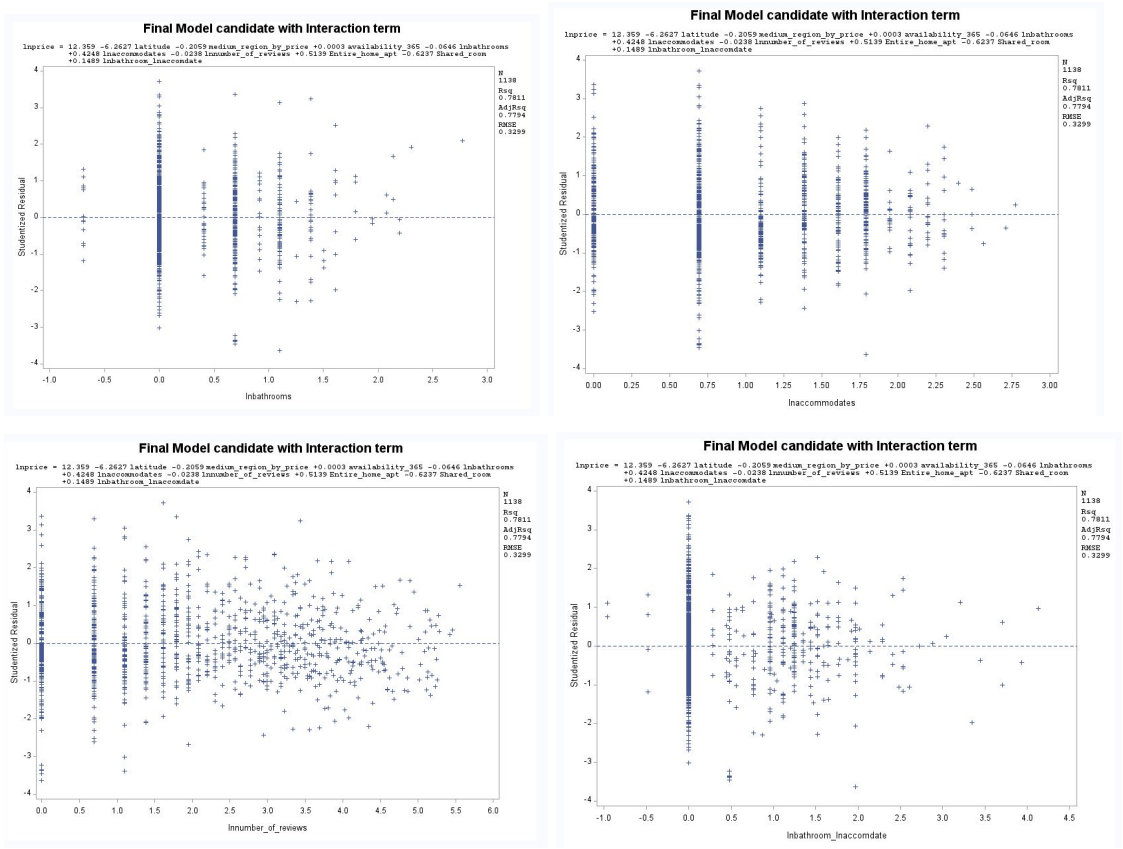


Figure 4.10

Model 2 Residuals

Previously transformed independent variables remains fairly constant and linear. From the above parameter table, we can also see VIF for each parameter all remains to be under acceptable threshold value of 10.

Final Model Validation

Fitted Final Model 1 without Interaction term Final model candidate and Model 2 with Interaction term Final model candidate.

Final Models Validation

The REG Procedure
Model: MODEL1
Dependent Variable: new_lnprice

Number of Observations Read	2936
Number of Observations Used	1138
Number of Observations with Missing Values	1798

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	435.58284	54.44785	490.07	<.0001
Error	1129	125.43441	0.11110		
Corrected Total	1137	561.01725			

Root MSE	0.33332	R-Square	0.7764
Dependent Mean	4.67884	Adj R-Sq	0.7748
Coeff Var	7.12398		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.27546	0.50849	24.14	<.0001	0	0
latitude	1	-6.23165	0.38092	-16.36	<.0001	-0.26053	1.28064
medium_region_by_price	1	-0.21069	0.02371	-8.89	<.0001	-0.13676	1.19597
availability_365	1	0.00032903	0.00007093	4.64	<.0001	0.06690	1.05039
lnbathrooms	1	0.06222	0.02337	2.66	0.0079	0.04288	1.30945
lnaccommodates	1	0.50660	0.02113	23.97	<.0001	0.46166	1.87301
lnnumber_of_reviews	1	-0.02783	0.00680	-4.09	<.0001	-0.05830	1.02491
Entire_home_apt	1	0.50737	0.02548	19.91	<.0001	0.36121	1.66183
Shared_room	1	-0.67974	0.05105	-13.31	<.0001	-0.22139	1.39604

Figure 4.11

Final Model 1

Model 1 Final Equation:

ln(Price) = 12.27546 - 6.23165(latitude) - 0.21069(medium_region_by_price) + 0.00032903(availability_365) + 0.06222(lnbathrooms) + 0.50660(lnaccommodates) - 0.02783(lnnumber_of_reviews) + 0.50737(Entire_home_apt) - 0.679745(Shared_room)

Final Models Validation

The REG Procedure
Model: MODEL2
Dependent Variable: new_lnprice

Number of Observations Read	2936
Number of Observations Used	1138
Number of Observations with Missing Values	1798

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	438.22792	48.69199	447.31	<.0001
Error	1128	122.78933	0.10886		
Corrected Total	1137	561.01725			

Root MSE	0.32993	R-Square	0.7811
Dependent Mean	4.67884	Adj R-Sq	0.7794
Coeff Var	7.05159		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	12.35885	0.50361	24.54	<.0001	0	.	0
latitude	1	-6.26269	0.37711	-16.61	<.0001	-0.26182	0.78064	1.28099
medium_region_by_price	1	-0.20586	0.02349	-8.76	<.0001	-0.13363	0.83469	1.19805
availability_365	1	0.00034182	0.00007026	4.86	<.0001	0.06950	0.95073	1.05183
lnbathrooms	1	-0.06460	0.03459	-1.87	0.0621	-0.04452	0.34132	2.92981
lnaccommodates	1	0.42478	0.02671	15.91	<.0001	0.38710	0.32762	3.05228
lnnumber_of_reviews	1	-0.02384	0.00678	-3.52	0.0005	-0.04995	0.96181	1.03971
Entire_home_apt	1	0.51392	0.02526	20.35	<.0001	0.36588	0.60008	1.66645
Shared_room	1	-0.62372	0.05179	-12.04	<.0001	-0.20314	0.68183	1.46664
lnbathroom_lnaccomdate	1	0.14894	0.03022	4.93	<.0001	0.14335	0.22944	4.35849

Figure 4.12

Final Model 2

Model 2 Final Model Equation:

ln(Price) = 12.35885 - 6.26269 (latitude) - 0.20586 (medium_region_by_price) + 0.00034182 (availability_365) - 0.06460 (lnbathrooms) + 0.42478 (lnaccommodates) - 0.02384 (lnnumber_of_reviews) + 0.51392 (Entire_home_apt) - 0.62372 (Shared_room) + 0.14894 (lnbathroom_lnaccomdate)

Test Set Validation Statistics

Validation statistics for Model 1

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1173	0.34694	0.26860

Validation statistics for Model 2

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1173	0.34036	0.26216

Figure 5.1

Validation Comparison

Validation performance metrics are obtained by using test set data. Training set is used to generate the model, test set is used to generate prediction, phat, from the model. The sum of the absolute difference of all non-missing data is aggregated as mean absolute error and root mean squared error.

Model 1 and Model 2 both have very similar **rmse** and **mae**. Model 2 have a slight edge on error terms based on test set.

	Without interaction		With Interaction	
	Final Model 1		Final Model 2	
Train	RMSE	0.3333	RMSE	0.3299
Set	R2	0.7764	R2	0.7811
	Adj-R2	0.7748	Adj-R2	0.7794
	# Coef	8	# Coef	9
	GOF	Good	GOF	Good
	Residuals	Good	Residuals	Good
Test	RMSE	0.3469	RMSE	0.3404
Set	R2	0.7764	R2	0.7811
	Adj-R2	0.7748	Adj-R2	0.7794
	MAE	0.2686	MAE	0.2822
	CV-R2	0.0208	CV-R2	0.0163
	# Coef	8	# Coef	9

Figure 5.2

Validation Comparison

Validation set R^2 for Model 1

The CORR Procedure

2 Variables: lnprice phat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
lnprice	1173	4.72644	0.71166	5544	2.63906	6.21461	
phat	758	4.70651	0.62068	3568	3.13235	6.01276	Predicted Value of new_lnprice

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	lnprice	phat
lnprice	1.00000 1173	0.86927 <.0001 758
phat Predicted Value of new_lnprice	0.86927 <.0001 758	1.00000 758

Validation set R^2 for Model 2

The CORR Procedure

2 Variables: lnprice phat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
lnprice	1173	4.72644	0.71166	5544	2.63906	6.21461	
phat	758	4.70494	0.62280	3566	3.22854	6.30790	Predicted Value of new_lnprice

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	lnprice	phat
lnprice	1.00000 1173	0.87451 <.0001 758
phat Predicted Value of new_lnprice	0.87451 <.0001 758	1.00000 758

Figure 5.3

Validation Comparison

Conclusion on performance on validation or test set

Performance of Model 1 and Model 2 are similar, no significant difference between having a model without interaction terms and with interaction term. However, ln(bathroom) coefficient flip to negative when interaction term is present. This means as bathroom count increase listing price decrease. Additionally, the significance of Beta on ln(bathrooms) is slightly outside the significance threshold. For these practical and technical reasons, the final model chosen by this analysis is the final model candidate Model 1.

Final Model Attribute Analysis

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.27546	0.50849	24.14	<.0001	0	0
latitude	1	-6.23165	0.38092	-16.36	<.0001	-0.26053	1.28064
medium_region_by_price	1	-0.21069	0.02371	-8.89	<.0001	-0.13676	1.19597
availability_365	1	0.00032903	0.00007093	4.64	<.0001	0.06690	1.05039
lnbathrooms	1	0.06222	0.02337	2.66	0.0079	0.04288	1.30945
lnaccommodates	1	0.50660	0.02113	23.97	<.0001	0.46166	1.87301
lnnumber_of_reviews	1	-0.02783	0.00680	-4.09	<.0001	-0.05830	1.02491
Entire_home_apt	1	0.50737	0.02548	19.91	<.0001	0.36121	1.66183
Shared_room	1	-0.67974	0.05105	-13.31	<.0001	-0.22139	1.39604

Figure 5.4

Validation Model 1: Final Model

Final Model Equation:

ln(Price) = 12.27546 - 6.23165(latitude) - 0.21069(medium_region_by_price) + 0.00032903(availability_365) + 0.06222(lnbathrooms) + 0.50660(lnaccommodates) - 0.02783(lnnumber_of_reviews) + 0.50737(Entire_home_apt) - 0.679745(Shared_room)

Significance of dummy variables

- 0.21069 (medium_region_by_price)

The final model retained medium_region_by_price as a dummy variable. While keeping all other attributes constant, when the active listing is in medium region the price of Airbnb decrease by 23.452% compared with the base case of low_region_by_price. This conflicting result maybe a product of using average to aggregate neighborhood Airbnb price. When re-examining the original neighborhood data, we can see from the plot below that there are extreme outliers that is skewing the average to the higher end while most neighborhood listing is much lower.

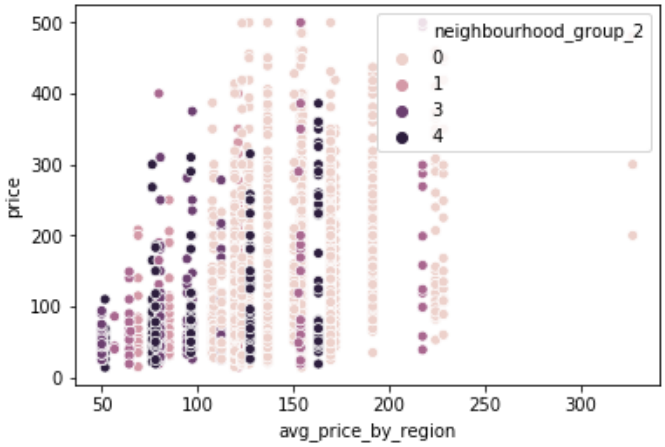


Figure 5.5

Price vs Avg Price by Neighborhood vs Neighborhood Group

+ 0.50737(Entire_home_apt) - 0.679745(Shared_room)

Both room_type dummy variables remained in the final model. When listing is the entire home/apt, one can expect the price to increase 66.091 percent from a private room listing. When the unit listing is entered as shared room there is a significant decrease in the listing price of 97.337 percent from private room listing. This finding is in-line with what Gibbs et al. found when they observed that “hosts charge a huge premium for privacy” when they analyzed Airbnb data for five major Canadian Cities in 2017.[8]

Significance of location variables

- 6.23165(latitude)

A single degree increase in latitude which is 69 miles results in price decrease of 50759.40%. When looking at the map of singapore, this term of model does make sense since the coast of Singapore is on the South side of the island. There is a significant increase in price as the Airbnb unit location is closer to the Southern coast line. Visualized in Appendix E. We did not find that proximity to tourist or landmarks, measured as distance_corre in our model, was significant. Our findings are similar to those of Perez-Sanchez et al. where they observed that proximity to the cost and not tourist locations and landmarks positively influenced Airbnb prices in 4 coastal Spanish cities in 2018.[4]

Significance of availability and number of reviews variables

0.00032903(availability_365) - 0.02783(lnnumber_of_reviews)

A single day increase in availability results in about .0329 percent increase in the price on airbnb. Having the unit available in a wider range of days generally increases the value of the Airbnb listing. This does intuitively make sense since more customers are competing to book the listing when available at a wider range of days.

The number of review is transformed with natural log function. When finding the significance of independent variable, we would need to treat the independent variable as a % increase or decrease. A single review increase is equivalent to the following while other attributes are kept constant: Percentage decrease in Price = EXP((1+(percentage increase/100)*abs(-0.02783)) - 1) * 100. Doubling the number_of_reviews or otherwise put increasing the number_of_reviews by 100% decrease of 1.947%. Our findings are consistent with Gibbs et al. analysis where they observed that the number of reviews of an Airbnb unit had a negative correlation between price. Gibbs et al. hypothesized that more actively managed properties used lower prices as a mechanism to increase occupancy rates.[3]

Significance of bathroom count and accommodation count variables

+ 0.06222(lnbathrooms) + 0.50660(lnaccommodates)

Using the same method of evaluating the contribution of natural log of reviews:

Percentage increase in *Price = EXP((1+(percentage increase/100)*abs(0.06222)) - 1) * 100.*
With all other attributes constant, a 100% increase or doubling in bathroom count increase the listing price by 4.407%. This is intuitive to what is expected which is an increased count in bathroom increase in price of listing.

Percentage increase in *Price = EXP((1+(percentage increase/100)*abs(0.50660)) - 1) * 100.*
With all other attributes constant, a 100% increase or doubling in the amount of people the listed unit can accommodate results in a 42.069% increase in listing price.

The number of people an Airbnb unit can accommodate in both physical capacity and number of bathrooms increase the listing price of the unit. Gibbs et al. found the same observing that capacity and number of bathrooms where associated with higher price.[8]

Issues and Limitations of the Model

During the end of the analysis stage, we discovered that there was an issue in transforming the number of reviews. When we applied natural log transformation of number_of_reviews due to non-constant residual issues on the full model. There were a number of entries that has 0 reviews which means the log transformation resulted NULL.

After we derived our final model, we went back and attempted to resolve this issue by offsetting the number_of_reviews by adding 1.

lnnumber_of_reviews = ln(number_of_reviews + 1)

Although by doing this, we resolved the issue of missing values for this predictor, the models chosen by attribute selection methods are generally significantly worst in R-squared values by a factor of 5 - 10% decrease. We suspect this significant decrease is due to the fact that unreviewed Airbnb units maybe new listings. New listers may not be familiar with the pricing for their Airbnb units.

This means our models is only usable at estimating the pricing of Airbnb units when there is at least 1 review given for the unit. For future work, we may derive a separate model for Airbnb with number of reviews greater than 0.

Prediction

We generated a prediction with the final model by first considering a sample similar to one of our actual observations, see attributes and values below. The predicted ln(price) = 5.2265, when transformed equals **$183.93** with a 95% C.I. (4.8226,5.6304) and a 95% prediction interval (4.4349,6.0180). **See Predicted Price Output in Appendix G.**

Latitude = 1.3, Medium_region_by_price = 1, Availability_365 = 0, lnBathrooms = 3, lnAccomodations = 2, lnNumber of reviews = 25, Entire_home_apt = 1, Shared_room = 0.

Additional predictions

- When slightly changing latitude from 1.3 to 1.35 while keeping other attributes constant, the predicted price decreases to **$128** transformed. The predicted ln(price)= 4.9002, (before transformation) with a 95% C.I. (4.4946,5.3059) and a 95% prediction interval (4.1078,5.6927)

- When slightly changing latitude from 1.3 to 1.28 (only), the predicted price would increase to **$212** transformed. The predicted ln(price)= 5.3570, before transformation with a 95% C.I. (4.9518,5.7622) and a 95% prediction interval (4.5648,6.1492)
- When changing latitude to 1.28, bathrooms to 4, accommodations to 2 and 10 reviews, the predicted price increases to **$280.00** (transformed). The predicted ln(price) 5.6331, before transformation with a 95% C.I. (5.3873,5.8789) and a 95% prediction interval (4.9094,6.3569) - **See Predicted Price 4 in Appendix G**

Predicted Price 4

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	5.6331	0.1252	5.3873	5.8789	4.9094	6.3569	.
2	3.56	.	.	.	.	.	.	.
3	3.74	4.1395	0.0381	4.0647	4.2143	3.4547	4.8243	-0.4018

Figure 6.1 Prediction

Future Work

Some different analysis methods could have been employed to further explore the dataset:

A cluster analysis in sync with linear regression could have been used to increase prediction accuracy. Airbnb units of a similar type could be grouped into clusters. Such a model should give better performance results since variance from outside clusters, or other dissimilar types of listings, would be eliminated.

Time series forecasting can be performed on airbnb price listing. Seasonal forecasting analysis can be performed with yearly aggregated airbnb listing data. This analysis may be performed in conjunction with cluster analysis and linear regression analysis as an extra dimension regression.

The Airbnb help center provides a list of some basic, or essential, amenities “that a guest expects in order to have a comfortable stay.” (Airbnb Help Center 2019). This list, which includes fundamental items such as toilet paper, pillows and linens, when compared to some of the entries in our original dataset, (e.g. Gym, Pool, Doorman, etc...), lends credence to the existence of different categories of amenities. Defining this distinction in the types of amenities could give insight into what kinds of amenities do consumers value. By grouping the amenities into different buckets

Comparative analysis between the factors that predict price in Airbnbs vs the factors that predict price in hotels. The model is able to predict price for airbnb units in Singapore. However, most of the attributes used in the analysis of those units are also present in hotels in Singapore.

These attributes, (e.g. Bedrooms, Bathrooms, Latitude, etc...), are not unique to airbnb and so comparison between the factors that affect price in hotels and airbnbs would give better insight into whether the factors in accommodation pricing are consistent across mediums.

A recent paper by Georgios Zervas, Davide Prosperio and John W. Byers examines the emergence of peer-to-peer markets and the effect that Airbnb has on the hotel industry. They found that there is a determinable effect of Airbnb on the Texas hotel industry. They studied over 10,000 hosts and 15,000 listings to discover that, "...at Airbnb adoption rates exceeding 1,000 room, the estimate (−:085, p < :05), indicates (because we are now working with a log-level specification) an average impact of 8.5% on hotel room revenue. ", (Zervas 2016). An adoption rate of over 1,000 rooms was seen by Zervas in cities like Austin, Texas. Given that there are few studies on the competition between sharing economy businesses, analyzing the effect that hotels and airbnb have on one another is a good starting point.

RESEARCH REFERENCES

1) Fast Facts (n.d.) Retrieved November 11, 2019, from https://news.airbnb.com/fast-facts/.

2) Zhihua Zhang, Rachel J. C. Chen, Lee D. Han, and Lu Yang. "Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach." *Sustainability* 9.9 (2017): 1635. Web.

3) Gibbs, C., D. Guttentag, U. Gretzel, J. Morton, and A. Goodwill.2017. Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. Journal of Travel & Tourism Marketing 35 (1): 46–56.

4) V. Raul Perez-Sanchez, Leticia Serrano-Estrada, Pablo Marti, and Raul-Tomas Mora-Garcia. "The What, Where, and Why of Airbnb Price Determinants." *Sustainability* 10.12 (2018): 4596. Web.

5) Yang Li, Quan Pan, Tao Yang, and Lantian Guo. "Reasonable Price Recommendation on Airbnb Using Multi-Scale Clustering." *2016 35th Chinese Control Conference (CCC)* 2016 (2016): 7038-041. Web.

6) "Marina Bay." *Singapore*, Singapore Tourism Board, www.visitingapore.com/see-do-singapore/places-to-see/marina-bay-area.

7) Li, J., A. Moreno, and D. Zhang. 2015. Agent behavior in the sharing economy: Evidence from Airbnb. Ross School of Business Working Paper Series (1298).

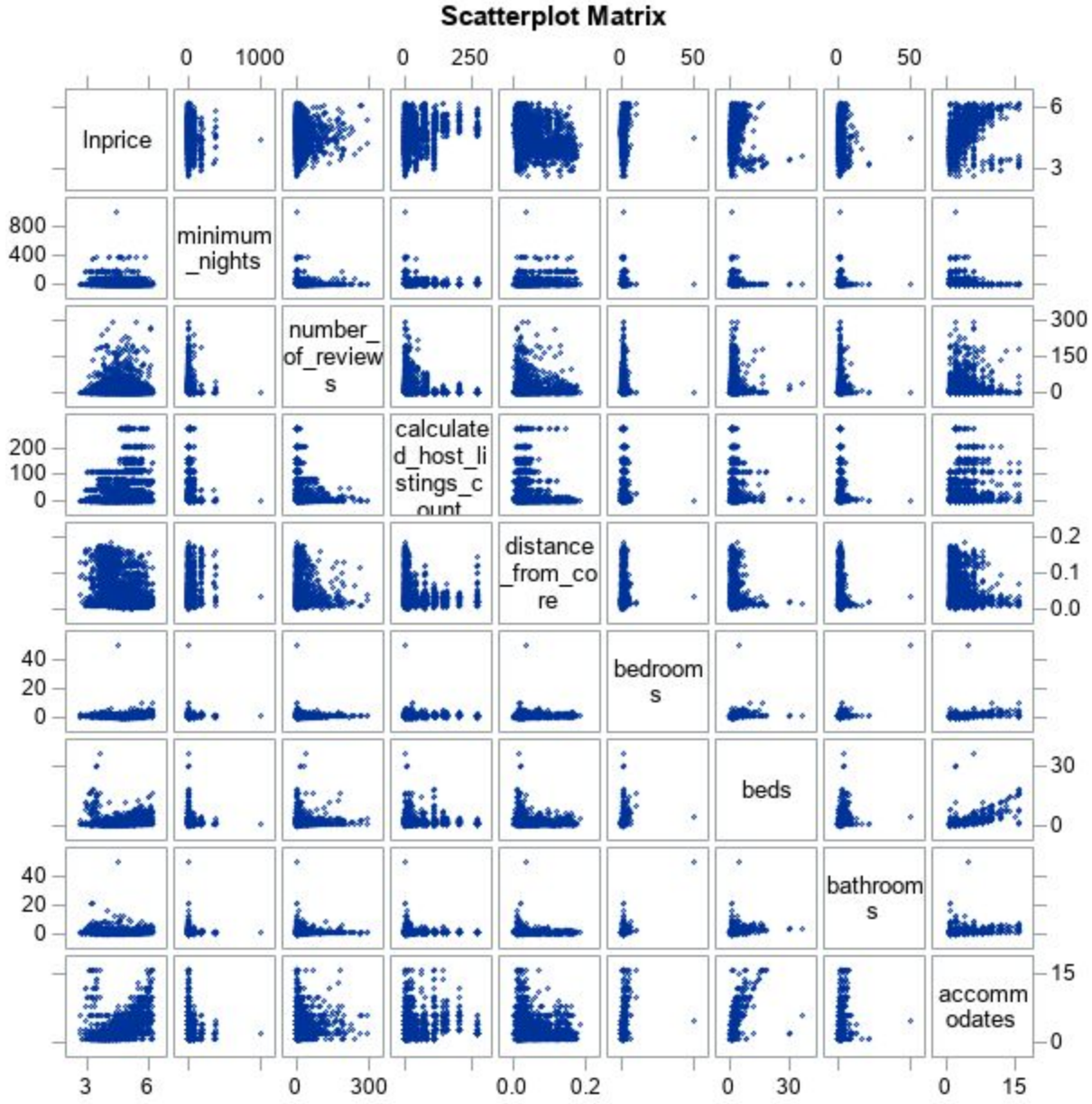
8) Kaggle.com. (2019). *Singapore Airbnb*. [online] Available at: https://www.kaggle.com/jojoker/singapore-airbnb [Accessed 31 Oct. 2019].

9) Zervas, Georgios and Proserpio, Davide and Byers, John, The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry (Nov 18, 2016). Boston U. School of Management Research Paper No. 2013-16. Available at SSRN: https://ssrn.com/abstract=2366898 or http://dx.doi.org/10.2139/ssrn.2366898

10) "What Are Essential Amenities?" Airbnb Help Center, 2019, www.airbnb.com/help/article/2343/what-are-essential-amenities.

Appendix

Appendix A: Scatterplot Matrix



Appendix B: Univariate Extreme Observations

The UNIVARIATE Procedure
Variable: price

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14	2675	500	1488
14	2384	500	1949
14	142	500	2233
15	2029	500	2952
15	1158	500	2966

The UNIVARIATE Procedure
Variable: lnprice

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2.63906	2675	6.21461	1488
2.63906	2384	6.21461	1949
2.63906	142	6.21461	2233
2.70805	2029	6.21461	2952
2.70805	1158	6.21461	2966

The UNIVARIATE Procedure
Variable: minimum_nights

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	2999	365	2537
1	2997	365	2880
1	2996	365	2981
1	2994	365	2983
1	2991	1000	1538

The UNIVARIATE Procedure
Variable: number_of_reviews

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	3000	259	2064
0	2987	263	2041
0	2986	272	2056
0	2985	289	1562
0	2981	291	1446

The UNIVARIATE Procedure
Variable: calculated_host_listings_count

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	3000	274	2818
1	2999	274	2830
1	2997	274	2905
1	2996	274	2960
1	2994	274	2970

The UNIVARIATE Procedure
Variable: distance_from_core

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.00230972	2563	0.173736	2085
0.00248224	2493	0.174313	2080
0.00291707	2534	0.174693	2091
0.00339247	2536	0.175817	2114
0.00345023	2643	0.186210	3000

The UNIVARIATE Procedure
Variable: bedrooms

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2999	6	413
0	2981	6	460
0	2965	10	305
0	2964	10	532
0	2962	50	967

The UNIVARIATE Procedure
Variable: beds

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2958	30	282
0	2916	30	283
0	2821	30	438
0	2790	30	595
0	2666	36	1244

The UNIVARIATE Procedure
Variable: bathrooms

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	2999	13	1641
0	2978	16	1602
0	2947	21	402
0	2904	21	553
0	2488	50	967

Appendix B: Univariate Extreme Observations

The UNIVARIATE Procedure
Variable: accommodates

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	2998	16	1609
1	2987	16	1618
1	2985	16	1639
1	2982	16	2219
1	2976	16	2267

Univariate Analysis Extreme Observations

Minimum_nights

Observation 1538 has an obvious entry error for minimum_nights of 1000 nights needed to book. This observation will be removed from the dataset as an outlier

Calculated_host_listings_count

Some data points shared the same 274 input for host listings. After some examination, we observed the data that reported 274 host listings and found that hostname “Jay” had 274 listings. To further understand and comprehend the data under this host, we looked at the variable Host ID and determined that there was a different host id for the name “Jay”. Here we concluded to exclude this variable calculated_host_listings_count from the data set.

Bedrooms

Extreme observations for bedrooms: Observation 967 has an obvious entry error for bedrooms at a count of 50. This observation will be removed from the dataset, as the data for this entry reads: accommodates 5 people, 50 bedrooms, 50 bathrooms, 5 beds. Observation 967 has an obvious entry error for bedrooms at a count of 50. This observation will be removed from dataset. Observation 967 is already removed from dataset. Previously noted as an entry error.

Appendix B: Univariate Extreme Observations

Bed

For properties under “hostels”, we concluded to replace all bed counts that have a count of <1 with a count of “1” under beds. Reason being is that many of these hostel properties are including the total amount of beds in the total property, which implies that not all beds listed are included for price.

Bathrooms

Observation 137 is a missing value. This observation will be defaulted to 1. Observations 402 and 553 were entry errors for bathrooms at a count of 21. These observations will be removed from the dataset.

Extreme outliers

Lowest Values (observations) -

3000, 2999, 2997, 2996, 2994, 2987, 2985, 2981. We identified all of the lowest values that were deemed to be extreme observations, but retained them for purposes of our model.

Highest Values (observations) - 967

Dropped (402, 553, 895, 1073, 1744, 2219, 2267,374, 1339, 2966, 1538, 2958, 121, 2790, 321, 1827, 1747, 1848, 1085, 147, 501, 564, 524, 2100, 1193, 2821, 240, 2916, 2666, 967)

Appendix C: Variables Manually Removed

row	name	minimum_nights	number_of_reviews	calculated_host_listings_count	bedrooms	beds	bathrooms	accommodates
2958	Relaxing Quiet Spacious Lo	2	0	2	2	0	2	4
1827	å„ä¸å¿ƒæ¥è¡ŒBraddell MR	3	0	1	1	0	1.5	2
1848	Quiet room with private bath	2	3	1	1	0	1.5	2
2916	Brand New Master Room Be	90	9	14	1	0	1	1
501	lavLoftbed Rm1, no-sharing,	18	0	112	1	0	3	1
564	1960s Queenbed Rm1 @ lav	18	0	83	1	0	3	1
2100	Cozy Flat @ Jurong West	3	2	1	0	0	1	1
1747	Amazing room with private b	90	0	7	1	0	1.5	2
2790	666	3	0	3	0	0	3	2
147	Beachfront Hammocking at	1	0	2	0	0	0	2
1193	kEnsuiteM 1* Brickwall2 @ (	28	1	112	1	0	1	1
1085	3mins walk to MRT- Aircon I	2	48	1	1	0	1	2
2666	*Female only* Masterbedroc	5	1	2	1	0	1	1
121	Beachfront Budget Glamping	1	0	2	0	0	0	4
240	Heritage apartment: Master	18	0	112	1	0	1	1
2821	Little pandan	2	13	2	1	0	1	1
524	lavLoftbed Rm2, no-sharing,	18	0	112	1	0	3	1
2966	Nice attic room	1	0	1	1	1	1	2
402	Single Bed in 10-Bed Mixed	1	0	3	1	1	21	1
2267	Unique spacious loft located	1	1	1	0	1	1	16
374	Comfy & Relaxing Studio AF	30	0	274	0	1	1	2
553	Female Pod	1	0	3	1	1	21	1
321	Capsule Family Room | Mins	1	1	3	1	1	3	1
1538	Where Luxury City Living Re	1000	0	3	1	1	1	2
1073	Birthday, Parties, Corporate	1	0	1	1	2	1	16
1744	New Comfy 2 Queen Beds F	2	13	48	1	2	1	5
1339	Prominent location, cozy roo	1	1	2	1	2	1	3
967	family room 5pax	1	4	1	50	5	50	5
895	Idyllic suites æ•´å¥—å...¬å¯	1	3	1	1	15	8	16
2219	Central 65 Hostel & Cafe	1	0	1	1	16	5	16

Appendix D: Data Exploration and Analysis SAS Outputs

Model 1.1 Full Model without interaction

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	2970
Number of Observations Used	1930
Number of Observations with Missing Values	1040

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	714.70824	54.97756	388.14	<.0001
Error	1916	271.39249	0.14165		
Corrected Total	1929	986.10073			

Root MSE	0.37636	R-Square	0.7248
Dependent Mean	4.68713	Adj R-Sq	0.7229
Coeff Var	8.02960		

Model 1.2 without Interaction

Backward Elimination: Step 5

Variable minimum_nights Removed: R-Square = 0.7764 and C(p) = 8.9132

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	435.58284	54.44785	490.07	<.0001
Error	1129	125.43441	0.11110		
Corrected Total	1137	561.01725			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	12.27546	0.50849	64.74810	582.78	<.0001
latitude	-6.23165	0.38092	29.73403	267.63	<.0001
medium_region_by_price	-0.21069	0.02371	8.77354	78.97	<.0001
availability_365	0.00032903	0.00007093	2.39051	21.52	<.0001
lnbathrooms	0.06222	0.02337	0.78784	7.09	0.0079
lnaccommodates	0.50660	0.02113	63.83878	574.59	<.0001
lnnumber_of_reviews	-0.02783	0.00680	1.86070	16.75	<.0001
Entire_home_apt	0.50737	0.02548	44.04643	396.45	<.0001
Shared_room	-0.67974	0.05105	19.69679	177.29	<.0001

Appendix D: Data Exploration and Analysis SAS Outputs

Model 1.2 without Interaction

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	435.58284	54.44785	490.07	<.0001
Error	1129	125.43441	0.11110		
Corrected Total	1137	561.01725			

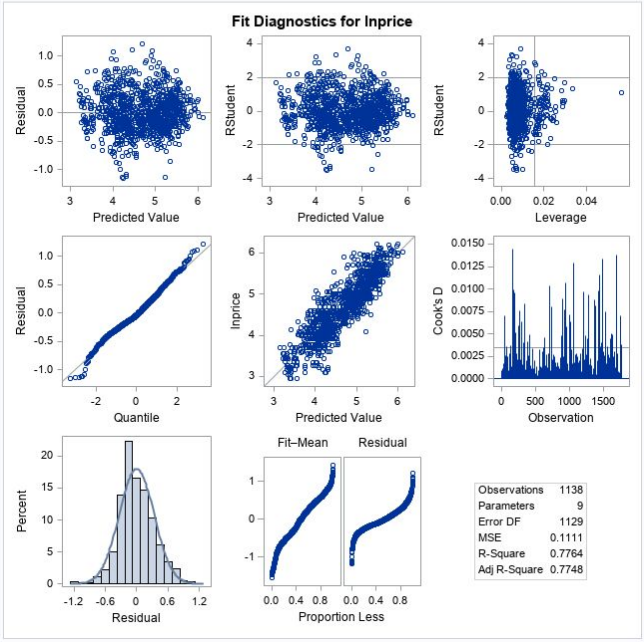
Root MSE	0.33332	R-Square	0.7764
Dependent Mean	4.67884	Adj R-Sq	0.7748
Coeff Var	7.12398		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	12.27546	0.50849	24.14	<.0001	0	.	0
latitude	1	-6.23165	0.38092	-16.36	<.0001	-0.26053	0.78086	1.28064
medium_region_by_price	1	-0.21069	0.02371	-8.89	<.0001	-0.13676	0.83614	1.19597
availability_365	1	0.00032903	0.00007093	4.64	<.0001	0.06690	0.95202	1.05039
lnbathrooms	1	0.06222	0.02337	2.66	0.0079	0.04288	0.76368	1.30945
lnaccommodates	1	0.50660	0.02113	23.97	<.0001	0.46166	0.53390	1.87301
lnnumber_of_reviews	1	-0.02783	0.00680	-4.09	<.0001	-0.05830	0.97570	1.02491
Entire_home_apt	1	0.50737	0.02548	19.91	<.0001	0.36121	0.60175	1.66183
Shared_room	1	-0.67974	0.05105	-13.31	<.0001	-0.22139	0.71631	1.39604

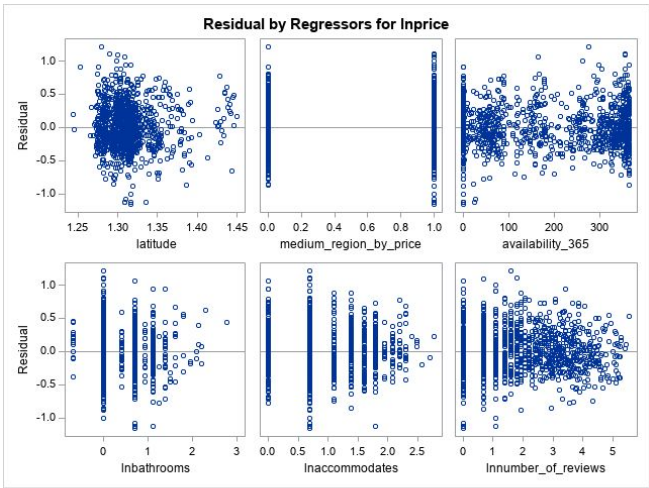
Model Selection Result from backward without Interaction term

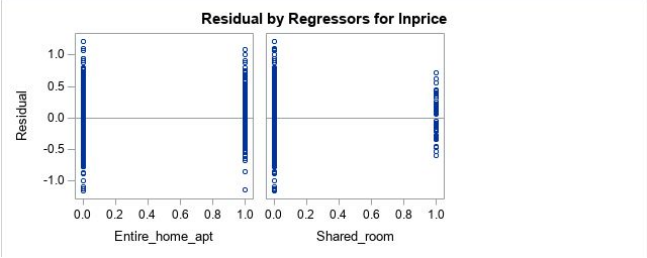
The REG Procedure
Model: MODEL1
Dependent Variable: lnprice



Appendix D: Data Exploration and Analysis SAS Outputs

Model 1.2 without Interaction





Appendix D: Data Exploration and Analysis SAS Outputs

Model 1.3 Forward without Interaction

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

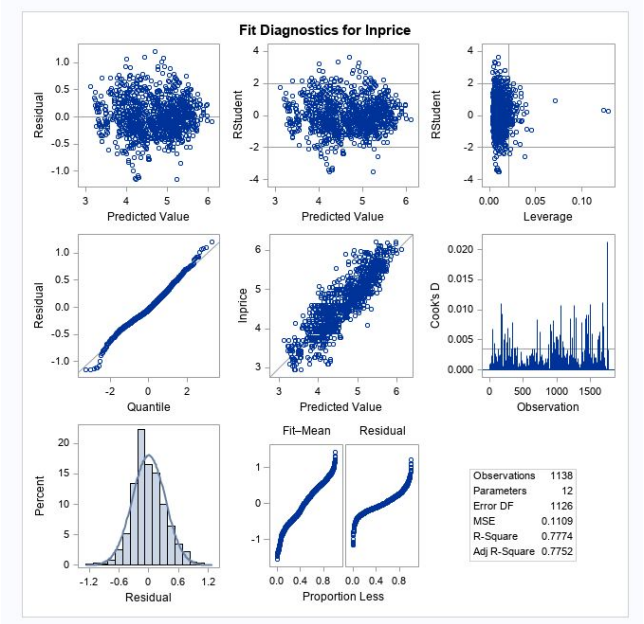
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	436.12055	39.64732	357.44	<.0001
Error	1126	124.89669	0.11092		
Corrected Total	1137	561.01725			

Root MSE	0.33305	R-Square	0.7774
Dependent Mean	4.67884	Adj R-Sq	0.7752
Coeff Var	7.11816		

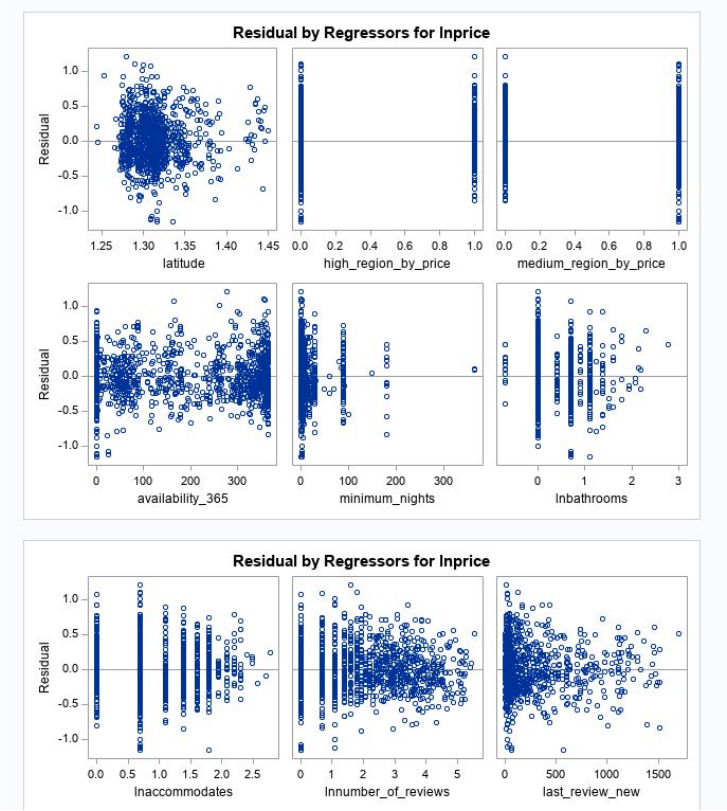
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	11.67520	0.69141	16.89	<.0001	0	.	0
latitude	1	-5.81639	0.50536	-11.51	<.0001	-0.24317	0.44293	2.25770
high_region_by_price	1	0.06607	0.05328	1.24	0.2152	0.03610	0.23334	4.28561
medium_region_by_price	1	-0.15864	0.04712	-3.37	0.0008	-0.10298	0.21133	4.73203
availability_365	1	0.00035971	0.00007380	4.87	<.0001	0.07314	0.87799	1.13896
minimum_nights	1	-0.00051628	0.00034591	-1.49	0.1358	-0.02279	0.84808	1.17913
lnbathrooms	1	0.05971	0.02343	2.55	0.0109	0.04115	0.75828	1.31877
lnaccommodates	1	0.50066	0.02151	23.28	<.0001	0.45625	0.51470	1.94286
lnnumber_of_reviews	1	-0.02534	0.00717	-3.53	0.0004	-0.05309	0.87654	1.14085
last_review_new	1	0.00004694	0.00003467	1.35	0.1760	0.02161	0.77651	1.28781
Entire_home_apt	1	0.50605	0.02588	19.55	<.0001	0.36027	0.58238	1.71709
Shared_room	1	-0.68915	0.05160	-13.36	<.0001	-0.22445	0.70012	1.42833



Appendix D: Data Exploration and Analysis SAS Outputs

Model 1.3 Forward without Interaction



Model 1.4 Stepwise Without Interaction term

See model 1.2 (backward selection method with interaction)

Appendix D: Data Exploration and Analysis SAS Outputs

Model 1.5 Cp without Interaction term

Model Selection Result from Cp without Interaction term

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	435.58284	54.44785	490.07	<.0001
Error	1129	125.43441	0.11110		
Corrected Total	1137	561.01725			

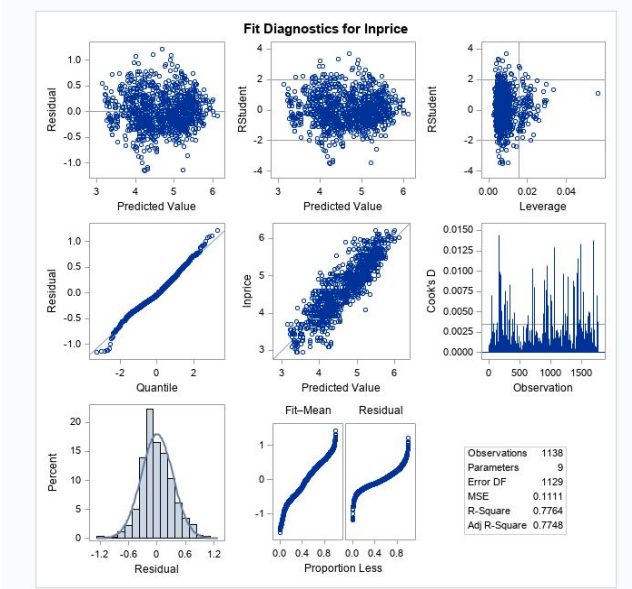
Root MSE	0.33332	R-Square	0.7764
Dependent Mean	4.67884	Adj R-Sq	0.7748
Coeff Var	7.12398		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	12.27546	0.50849	24.14	<.0001	0	.	0
latitude	1	-6.23165	0.38092	-16.36	<.0001	-0.26053	0.78086	1.28064
medium_region_by_price	1	-0.21069	0.02371	-8.89	<.0001	-0.13676	0.83614	1.19597
availability_365	1	0.00032903	0.00007093	4.64	<.0001	0.06690	0.95202	1.05039
lnbathrooms	1	0.06222	0.02337	2.66	0.0079	0.04288	0.76368	1.30945
lnaccommodates	1	0.50660	0.02113	23.97	<.0001	0.46166	0.53390	1.87301
lnnumber_of_reviews	1	-0.02783	0.00680	-4.09	<.0001	-0.05830	0.97570	1.02491
Entire_home_apt	1	0.50737	0.02548	19.91	<.0001	0.36121	0.60175	1.66183
Shared_room	1	-0.67974	0.05105	-13.31	<.0001	-0.22139	0.71631	1.39604

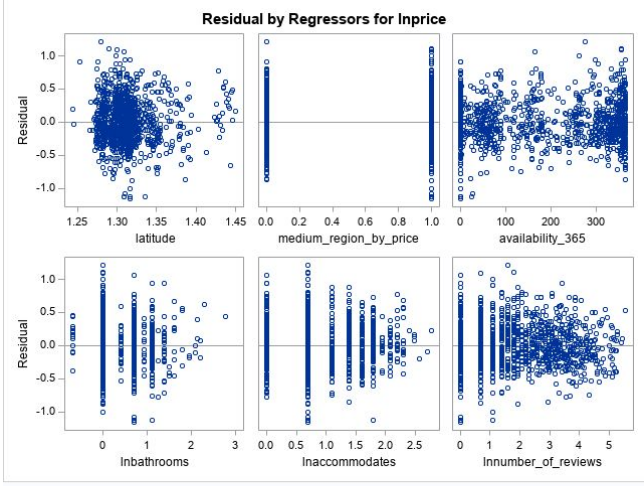
Appendix D: Data Exploration and Analysis SAS Outputs

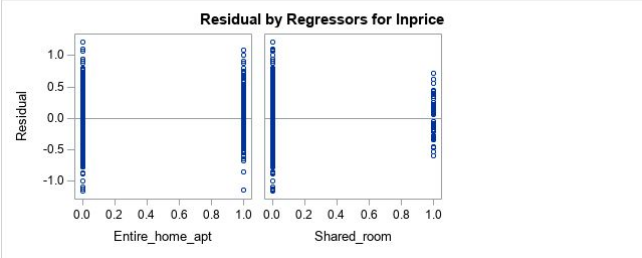
Model 1.5 Cp without Interaction term



Appendix D: Data Exploration and Analysis SAS Outputs

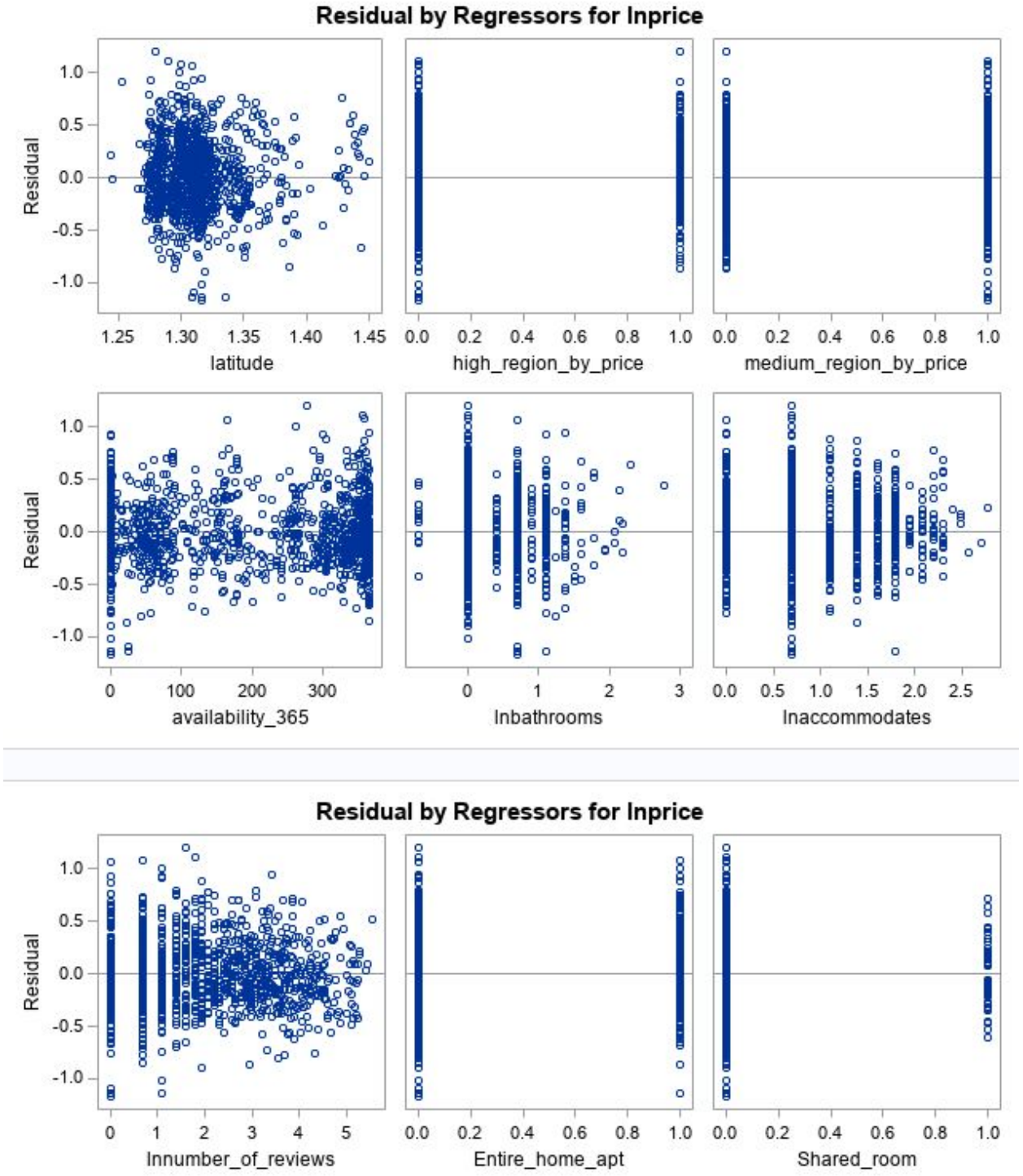
Model 1.5 Cp without Interaction term





Appendix D: Data Exploration and Analysis SAS Outputs

Model 1.6 Adj-R^2 without interaction term



Appendix D: Data Exploration and Analysis SAS Outputs

Model 1 Final Model Candidate

Final Model candidate without Interaction term

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	435.58284	54.44785	490.07	<.0001
Error	1129	125.43441	0.11110		
Corrected Total	1137	561.01725			

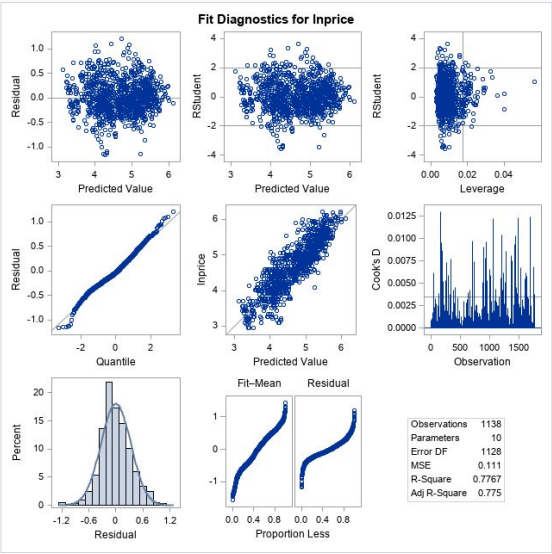
Root MSE	0.33332	R-Square	0.7764
Dependent Mean	4.67884	Adj R-Sq	0.7748
Coeff Var	7.12398		

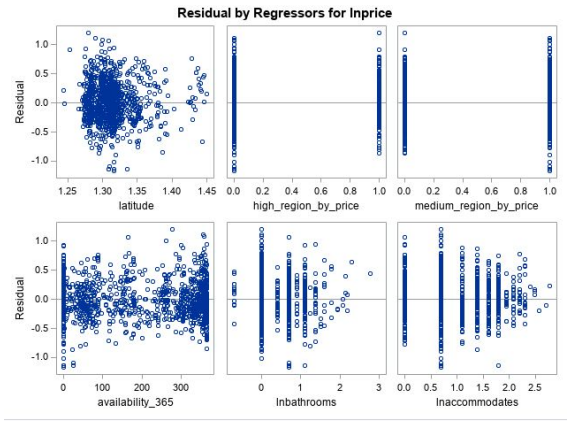
Parameter Estimates

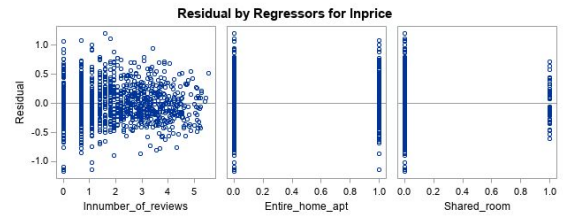
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	12.27546	0.50849	24.14	<.0001	0	0
latitude	1	-6.23165	0.38092	-16.36	<.0001	-0.26053	1.28064
medium_region_by_price	1	-0.21069	0.02371	-8.89	<.0001	-0.13676	1.19597
availability_365	1	0.00032903	0.00007093	4.64	<.0001	0.06690	1.05039
lnbathrooms	1	0.06222	0.02337	2.66	0.0079	0.04288	1.30945
lnaccommodates	1	0.50660	0.02113	23.97	<.0001	0.46166	1.87301
lnnumber_of_reviews	1	-0.02783	0.00680	-4.09	<.0001	-0.05830	1.02491
Entire_home_apt	1	0.50737	0.02548	19.91	<.0001	0.36121	1.66183
Shared_room	1	-0.67974	0.05105	-13.31	<.0001	-0.22139	1.39604

Appendix D: Data Exploration and Analysis SAS Outputs

Model 1 Final Model Candidate

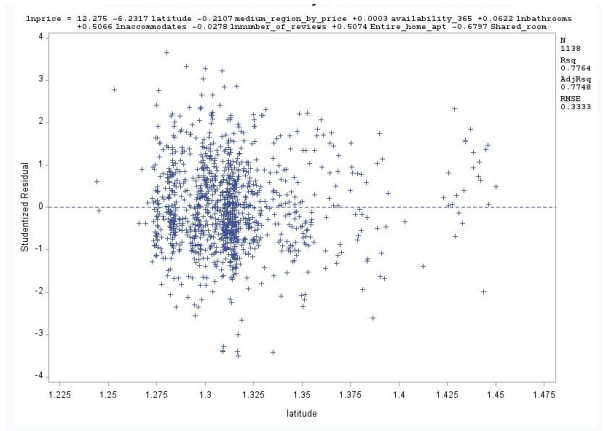


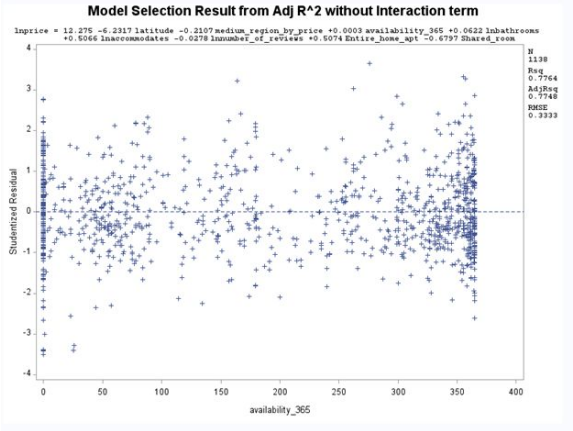


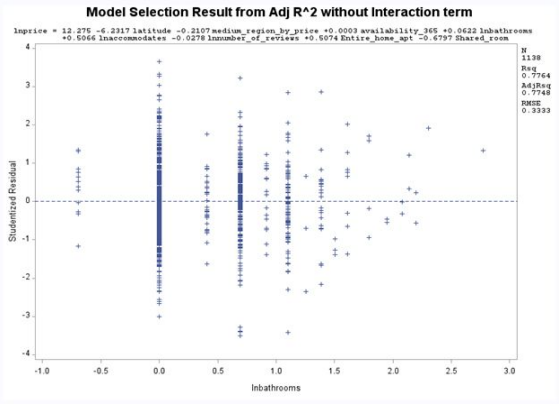


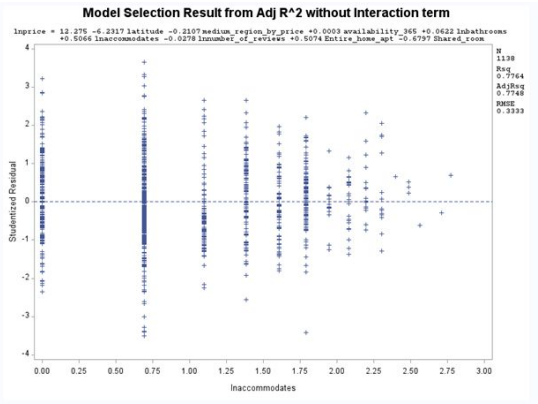
Appendix D: Data Exploration and Analysis SAS Outputs

Model 1 Final Model Candidate



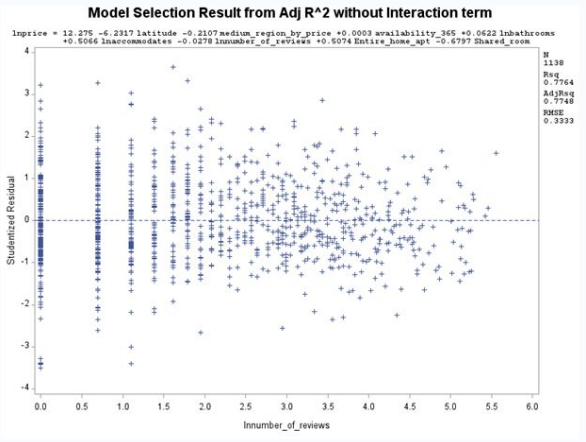






Appendix D: Data Exploration and Analysis SAS Outputs

Model 1 Final Model Candidate



Full Model 2.1

Full Regression Model 2.1 with Interaction Term for Log Growth and transformed independent

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	2970
Number of Observations Used	1930
Number of Observations with Missing Values	1040

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	719.80565	51.41469	369.74	<.0001
Error	1915	266.29508	0.13906		
Corrected Total	1929	986.10073			

Root MSE	0.37290	R-Square	0.7300
Dependent Mean	4.68713	Adj R-Sq	0.7280
Coeff Var	7.95591		

Appendix D: Data Exploration and Analysis SAS Outputs

Full Model 2.1

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	13.74664	23.61423	0.58	0.5605	0	0
latitude	1	-5.00283	0.51770	-9.66	<.0001	-0.20178	3.09177
longitude	1	-0.02968	0.22932	-0.13	0.8970	-0.00182	1.39975
high_region_by_price	1	0.08551	0.05022	1.70	0.0888	0.04599	5.17407
medium_region_by_price	1	-0.13693	0.04526	-3.03	0.0025	-0.08694	5.85680
distance_from_core	1	-0.56360	0.43784	-1.29	0.1982	-0.02923	3.65737
availability_365	1	0.00041292	0.00006411	6.44	<.0001	0.08203	1.15042
minimum_nights	1	-0.00076451	0.00028208	-2.71	0.0068	-0.03505	1.18591
lnbathrooms	1	-0.06164	0.02867	-2.15	0.0317	-0.04301	2.83821
lnaccommodates	1	0.40877	0.02329	17.55	<.0001	0.36636	3.08859
lnnumber_of_reviews	1	-0.01745	0.00613	-2.85	0.0044	-0.03654	1.16691
last_review_new	1	0.00005808	0.00002965	1.96	0.0503	0.02654	1.30131
Entire_home_apt	1	0.49378	0.02215	22.29	<.0001	0.34537	1.70235
Shared_room	1	-0.61768	0.04285	-14.42	<.0001	-0.20539	1.43954
lnbathroom_lnaccomdate	1	0.15299	0.02527	6.05	<.0001	0.14710	4.18615

Model 2.2 backward with interaction term

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

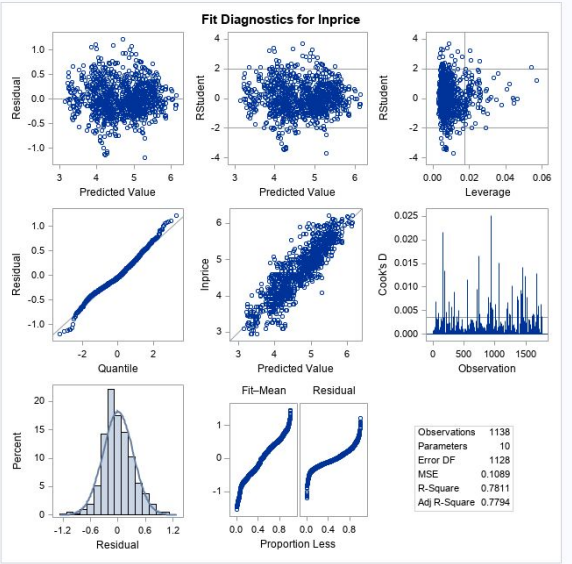
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	438.22792	48.69199	447.31	<.0001
Error	1128	122.78933	0.10886		
Corrected Total	1137	561.01725			

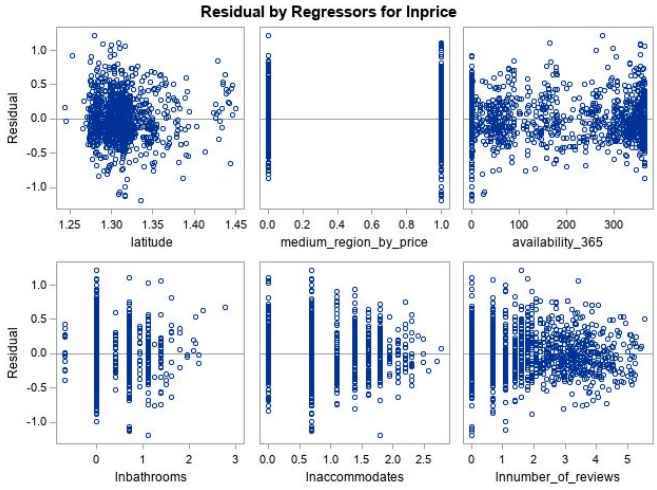
Root MSE	0.32993	R-Square	0.7811
Dependent Mean	4.67884	Adj R-Sq	0.7794
Coeff Var	7.05159		

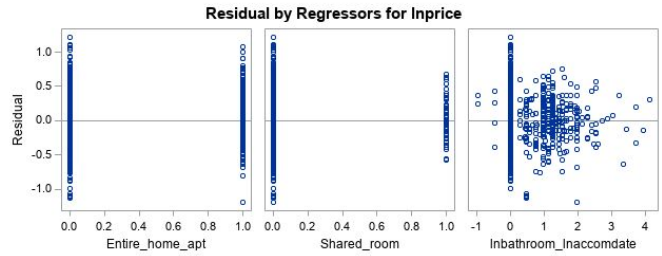
Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	12.35885	0.50361	24.54	<.0001	0	.	0
latitude	1	-6.26269	0.37711	-16.61	<.0001	-0.26182	0.78064	1.28099
medium_region_by_price	1	-0.20586	0.02349	-8.76	<.0001	-0.13363	0.83469	1.19805
availability_365	1	0.00034182	0.00007026	4.86	<.0001	0.06950	0.95073	1.05183
lnbathrooms	1	-0.06460	0.03459	-1.87	0.0621	-0.04452	0.34132	2.92981
lnaccommodates	1	0.42478	0.02671	15.91	<.0001	0.38710	0.32762	3.05228
lnnumber_of_reviews	1	-0.02384	0.00678	-3.52	0.0005	-0.04995	0.96181	1.03971
Entire_home_apt	1	0.51392	0.02526	20.35	<.0001	0.36588	0.60008	1.66645
Shared_room	1	-0.62372	0.05179	-12.04	<.0001	-0.20314	0.68183	1.46664
lnbathroom_lnaccomdate	1	0.14894	0.03022	4.93	<.0001	0.14335	0.22944	4.35849

Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.2 backward with interaction term







Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.3 Forward with Interaction

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

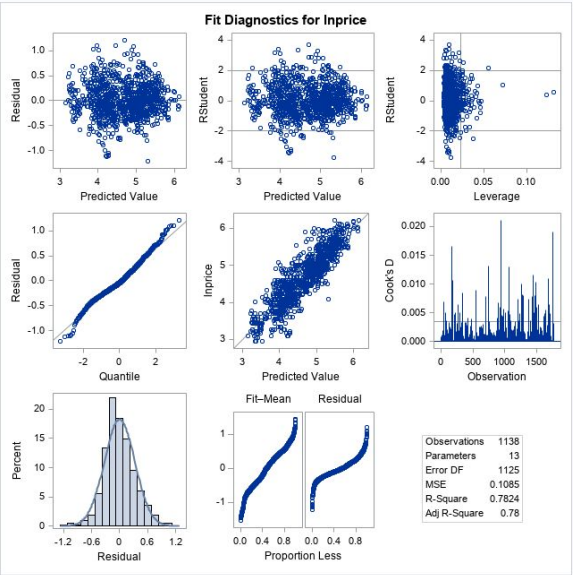
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	438.91808	36.57651	337.01	<.0001
Error	1125	122.09916	0.10853		
Corrected Total	1137	561.01725			

Root MSE	0.32944	R-Square	0.7824
Dependent Mean	4.67884	Adj R-Sq	0.7800
Coeff Var	7.04112		

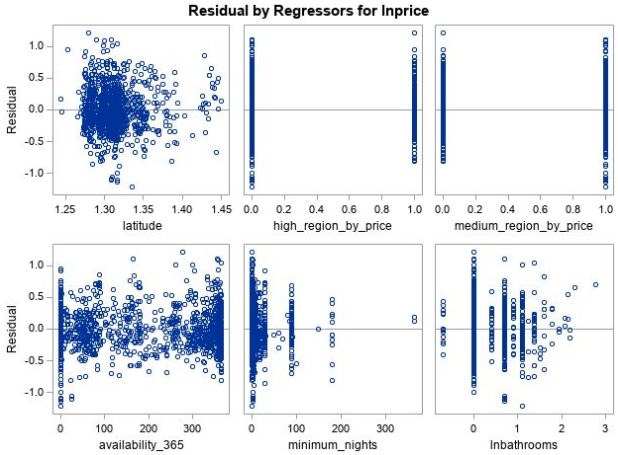
Parameter Estimates

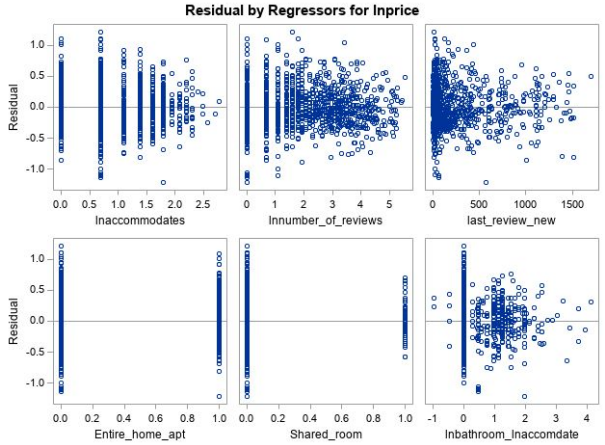
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	11.70374	0.68395	17.11	<.0001	0	.	0
latitude	1	-5.80704	0.49990	-11.62	<.0001	-0.24277	0.44292	2.25773
high_region_by_price	1	0.07162	0.05272	1.36	0.1746	0.03912	0.23324	4.28745
medium_region_by_price	1	-0.14941	0.04665	-3.20	0.0014	-0.09698	0.21100	4.73923
availability_365	1	0.00037739	0.00007309	5.16	<.0001	0.07673	0.87600	1.14155
minimum_nights	1	-0.00062515	0.00034284	-1.82	0.0685	-0.02759	0.84476	1.18377
lnbathrooms	1	-0.07107	0.03465	-2.05	0.0405	-0.04898	0.33921	2.94799
lnaccommodates	1	0.41505	0.02715	15.29	<.0001	0.37823	0.31611	3.16345
lnnumber_of_reviews	1	-0.02113	0.00714	-2.96	0.0031	-0.04428	0.86474	1.15642
last_review_new	1	0.00005027	0.00003430	1.47	0.1431	0.02314	0.77623	1.28828
Entire_home_apt	1	0.51300	0.02564	20.01	<.0001	0.36522	0.58072	1.72201
Shared_room	1	-0.63376	0.05219	-12.14	<.0001	-0.20641	0.66952	1.49361
lnbathroom_lnaccomdate	1	0.15353	0.03024	5.08	<.0001	0.14776	0.22840	4.37837



Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.3 Forward with Interaction





Model 2.4 Stepwise with Interaction

Model Selection Result from Stepwise with Interaction term

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	438.50109	43.85011	403.37	<.0001
Error	1127	122.51616	0.10871		
Corrected Total	1137	561.01725			

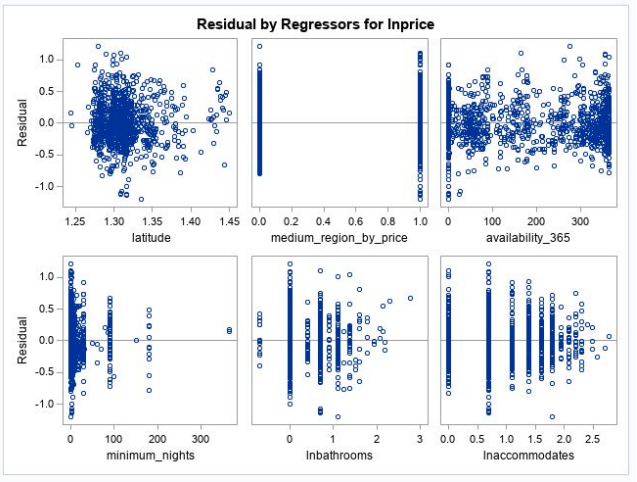
Root MSE	0.32971	R-Square	0.7816
Dependent Mean	4.67884	Adj R-Sq	0.7797
Coeff Var	7.04687		

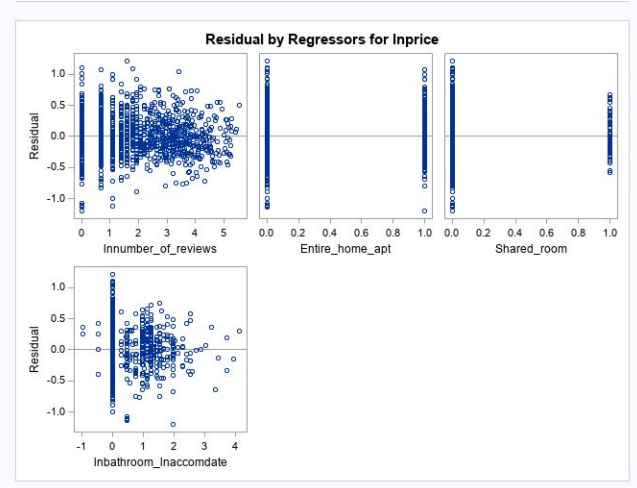
Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.4 Stepwise with Interaction

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	12.30822	0.50429	24.41	<.0001	0	.	0
latitude	1	-6.21633	0.37799	-16.45	<.0001	-0.25989	0.77597	1.28871
medium_region_by_price	1	-0.20638	0.02347	-8.79	<.0001	-0.13396	0.83453	1.19828
availability_365	1	0.00035924	0.00007107	5.05	<.0001	0.07304	0.92799	1.07760
minimum_nights	1	-0.00052052	0.00032837	-1.59	0.1132	-0.02298	0.92240	1.08413
lnbathrooms	1	-0.06610	0.03458	-1.91	0.0562	-0.04556	0.34106	2.93200
lnaccommodates	1	0.41726	0.02711	15.39	<.0001	0.38025	0.31759	3.14869
lnnumber_of_reviews	1	-0.02433	0.00678	-3.59	0.0003	-0.05097	0.95984	1.04184
Entire_home_apt	1	0.51659	0.02530	20.42	<.0001	0.36778	0.59742	1.67386
Shared_room	1	-0.63473	0.05222	-12.15	<.0001	-0.20673	0.66977	1.49305
lnbathroom_lnaccomdate	1	0.15189	0.03025	5.02	<.0001	0.14619	0.22857	4.37502





Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.5 Cp with Interaction

Number in Model	C(p)	R-Square	Variables in Model
10	11.0690	0.7816	latitude medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	11.0754	0.7820	latitude medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
12	11.2329	0.7824	latitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	11.3772	0.7819	latitude high_region_by_price medium_region_by_price availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
10	11.5438	0.7815	latitude high_region_by_price medium_region_by_price availability_365 lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
9	11.5819	0.7811	latitude medium_region_by_price availability_365 lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
12	12.2103	0.7822	latitude medium_region_by_price distance_from_core availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	12.3371	0.7818	latitude medium_region_by_price distance_from_core availability_365 minimum_nights lnbathrooms lnaccommodates lnnumber_of_reviews Entire_home_apt Shared_room lnbathroom_lnaccomdate
11	12.5526	0.7817	latitude high_region_by_price medium_region_by_price availability_365 lnbathrooms lnaccommodates lnnumber_of_reviews last_review_new Entire_home_apt Shared_room lnbathroom_lnaccomdate

Number of Observations Read	1763
Number of Observations Used	1138
Number of Observations with Missing Values	625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	438.91808	36.57651	337.01	<.0001
Error	1125	122.09916	0.10853		
Corrected Total	1137	561.01725			

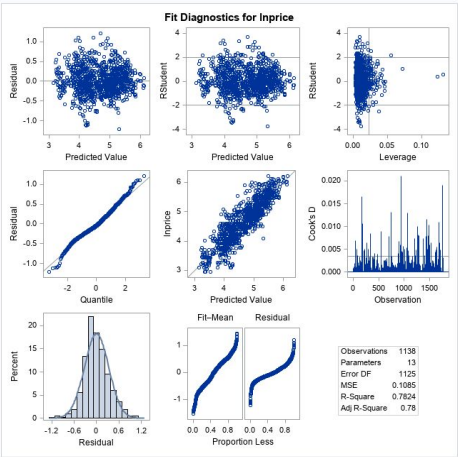
Root MSE	0.32944	R-Square	0.7824
Dependent Mean	4.67884	Adj R-Sq	0.7800
Coeff Var	7.04112		

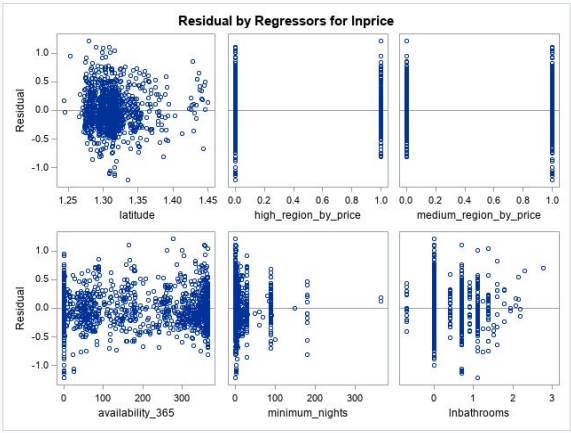
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	11.70374	0.68395	17.11	<.0001	0	.	0
latitude	1	-5.80704	0.49990	-11.62	<.0001	-0.24277	0.44292	2.25773
high_region_by_price	1	0.07162	0.05272	1.36	0.1746	0.03912	0.23324	4.28745
medium_region_by_price	1	-0.14941	0.04665	-3.20	0.0014	-0.09698	0.21100	4.73923
availability_365	1	0.00037739	0.00007309	5.16	<.0001	0.07673	0.87600	1.14155
minimum_nights	1	-0.00062515	0.00034284	-1.82	0.0685	-0.02759	0.84476	1.18377
lnbathrooms	1	-0.07107	0.03465	-2.05	0.0405	-0.04898	0.33921	2.94799
lnaccommodates	1	0.41505	0.02715	15.29	<.0001	0.37823	0.31611	3.16345
lnnumber_of_reviews	1	-0.02113	0.00714	-2.96	0.0031	-0.04428	0.86474	1.15642
last_review_new	1	0.00005027	0.00003430	1.47	0.1431	0.02314	0.77623	1.28828
Entire_home_apt	1	0.51300	0.02564	20.01	<.0001	0.36522	0.58072	1.72201
Shared_room	1	-0.63376	0.05219	-12.14	<.0001	-0.20641	0.66952	1.49361
lnbathroom_lnaccomdate	1	0.15353	0.03024	5.08	<.0001	0.14776	0.22840	4.37837

Appendix D: Data Exploration and Analysis SAS Outputs

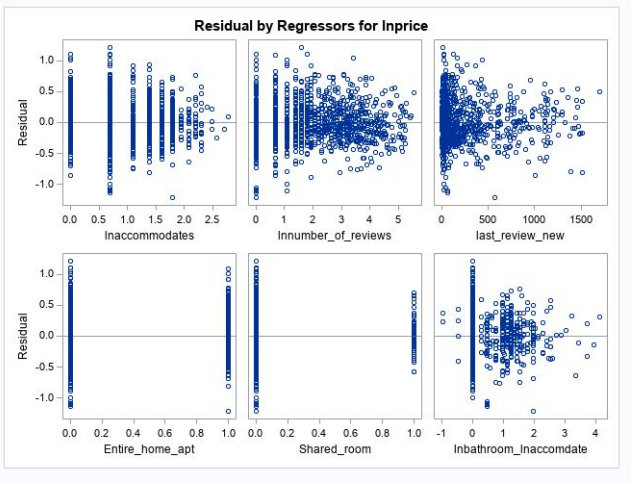
Model 2.5 Cp with Interaction



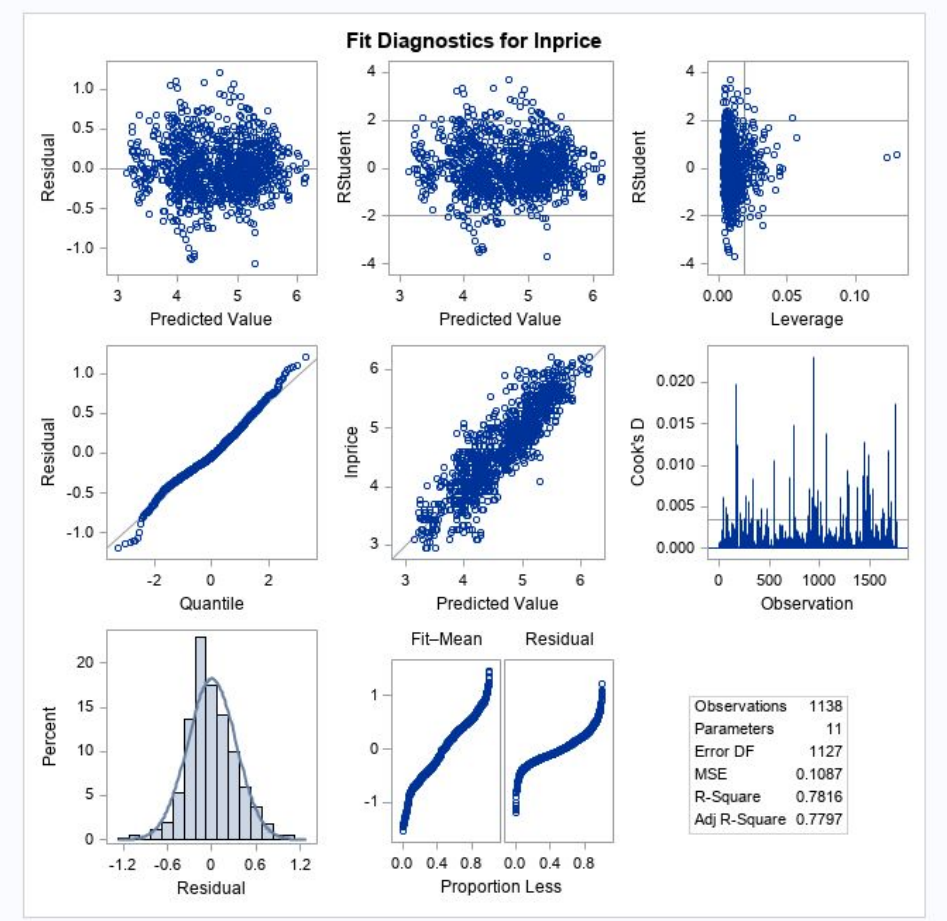


Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.5 Cp with Interaction

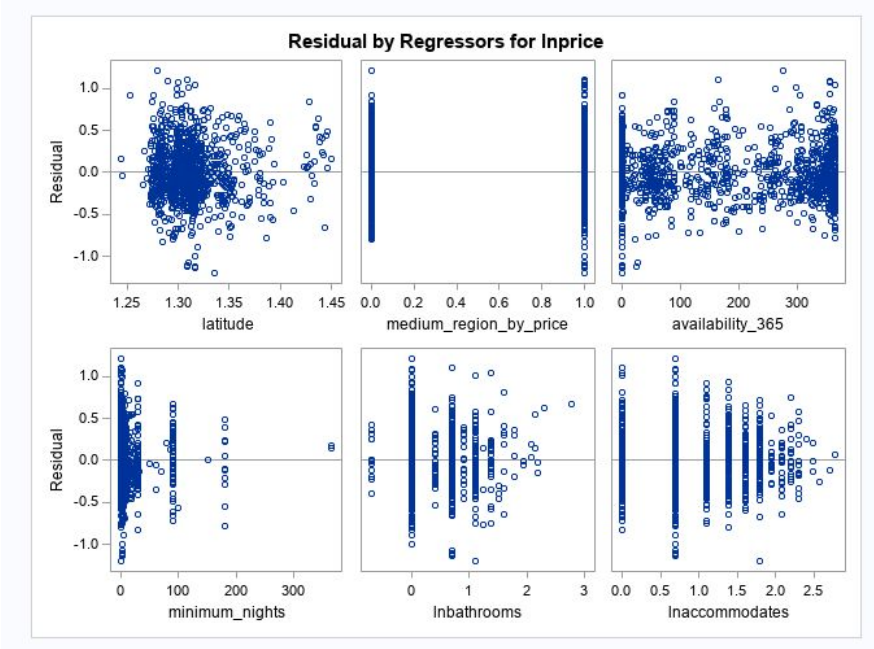


Model 2.6 Adj R square with Interaction Term



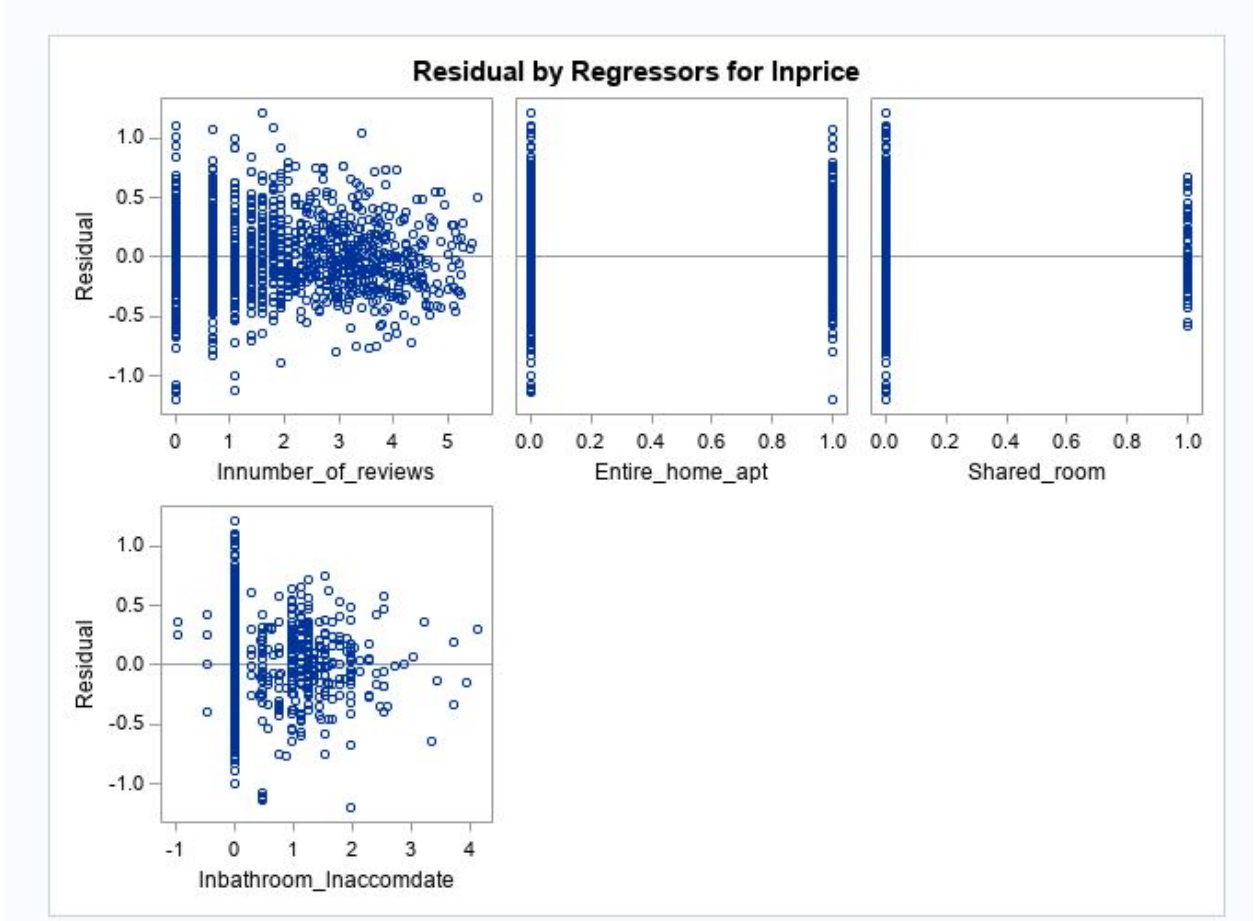
Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.6 Adj R square with Interaction Term



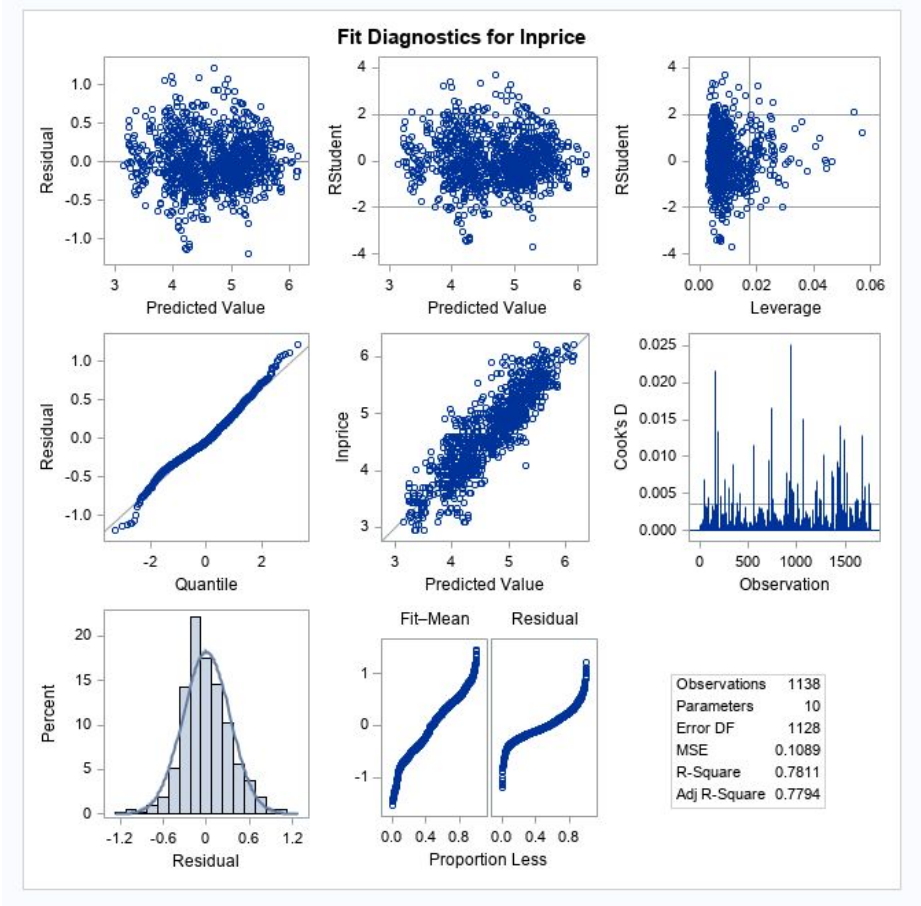
Appendix D: Data Exploration and Analysis SAS Outputs

Model 2.6 Adj R square with Interaction Term



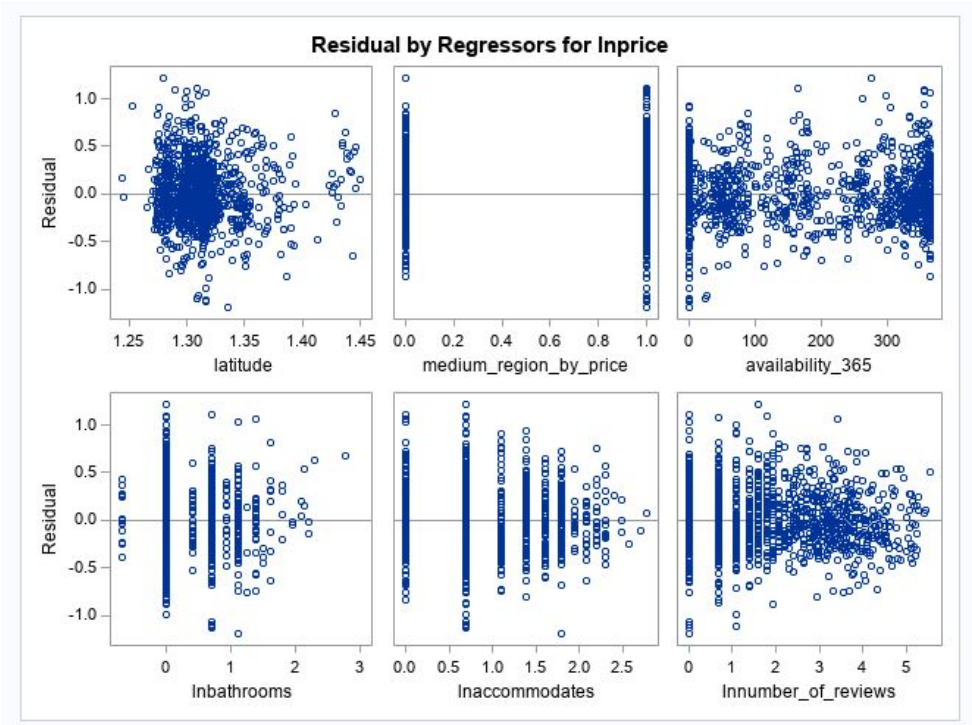
Appendix D: Data Exploration and Analysis SAS Outputs

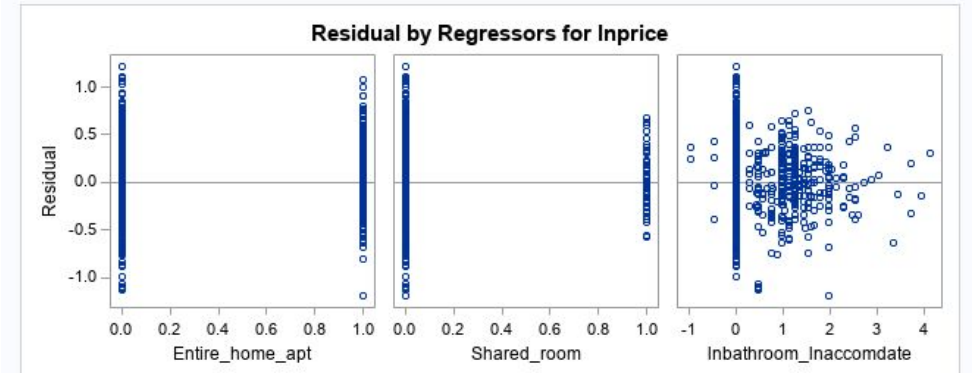
Model 2: Final Model Candidate with Interaction Term



Appendix D: Data Exploration and Analysis SAS Outputs

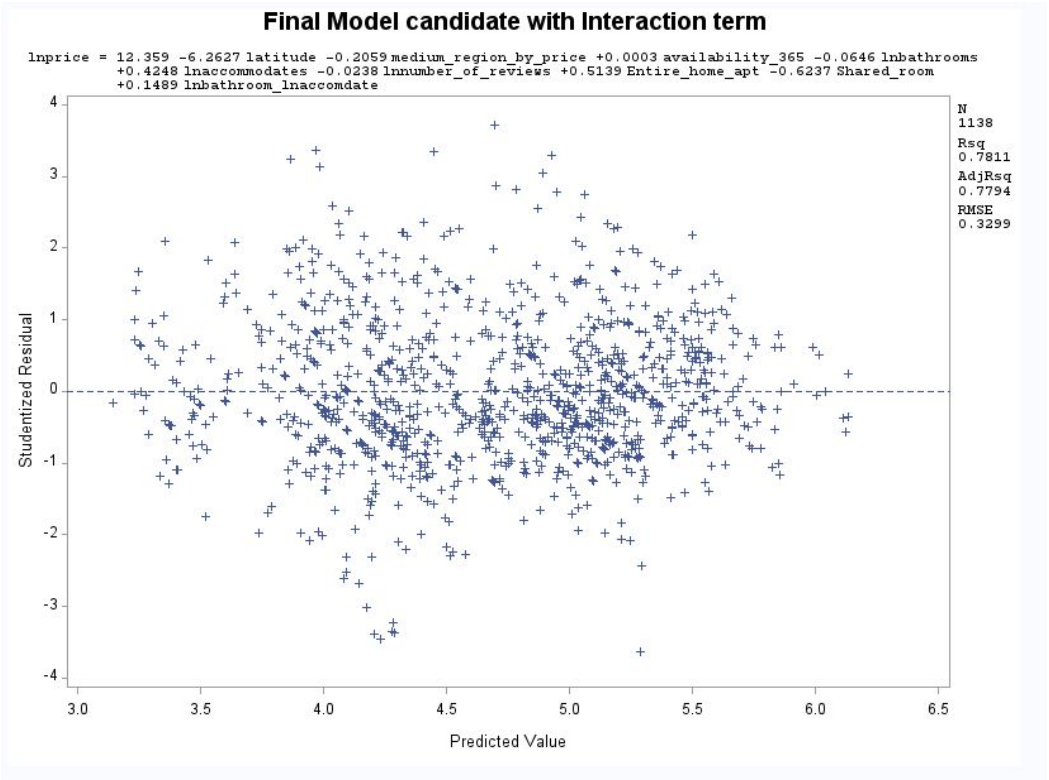
Model 2: Final Model Candidate with Interaction Term





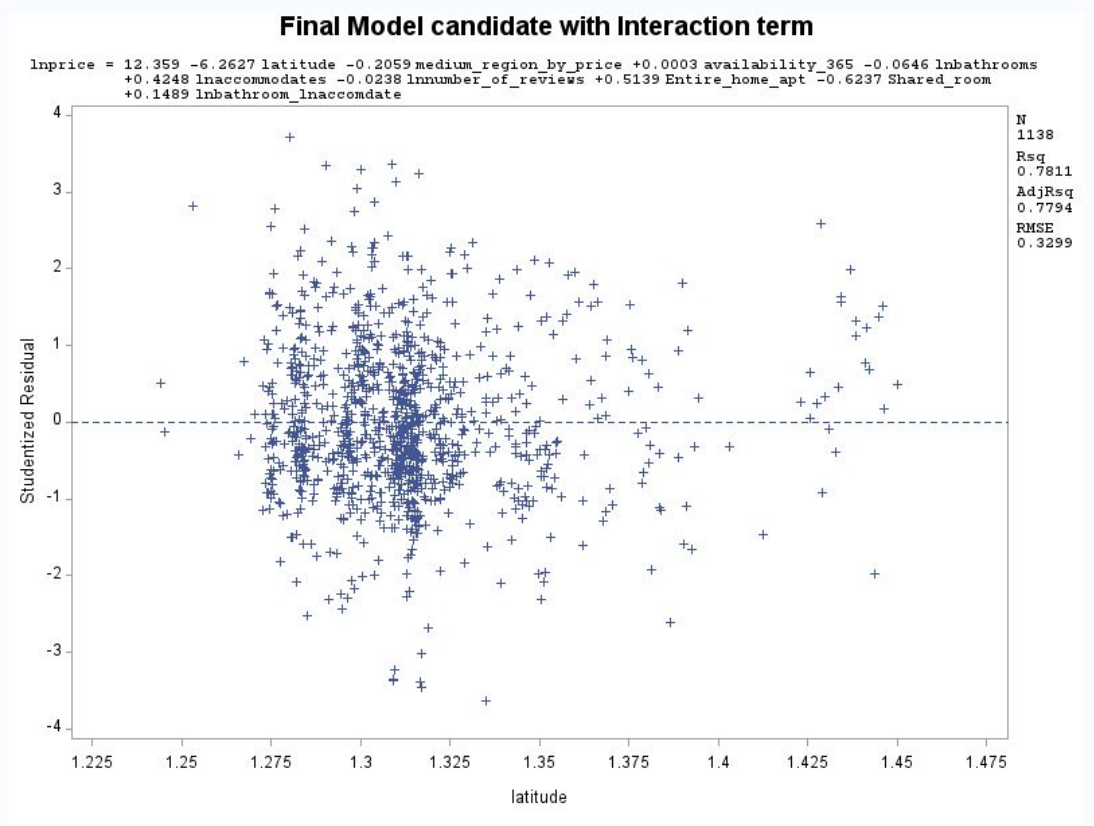
Appendix D: Data Exploration and Analysis SAS Outputs

Model 2: Final Model Candidate with Interaction Term



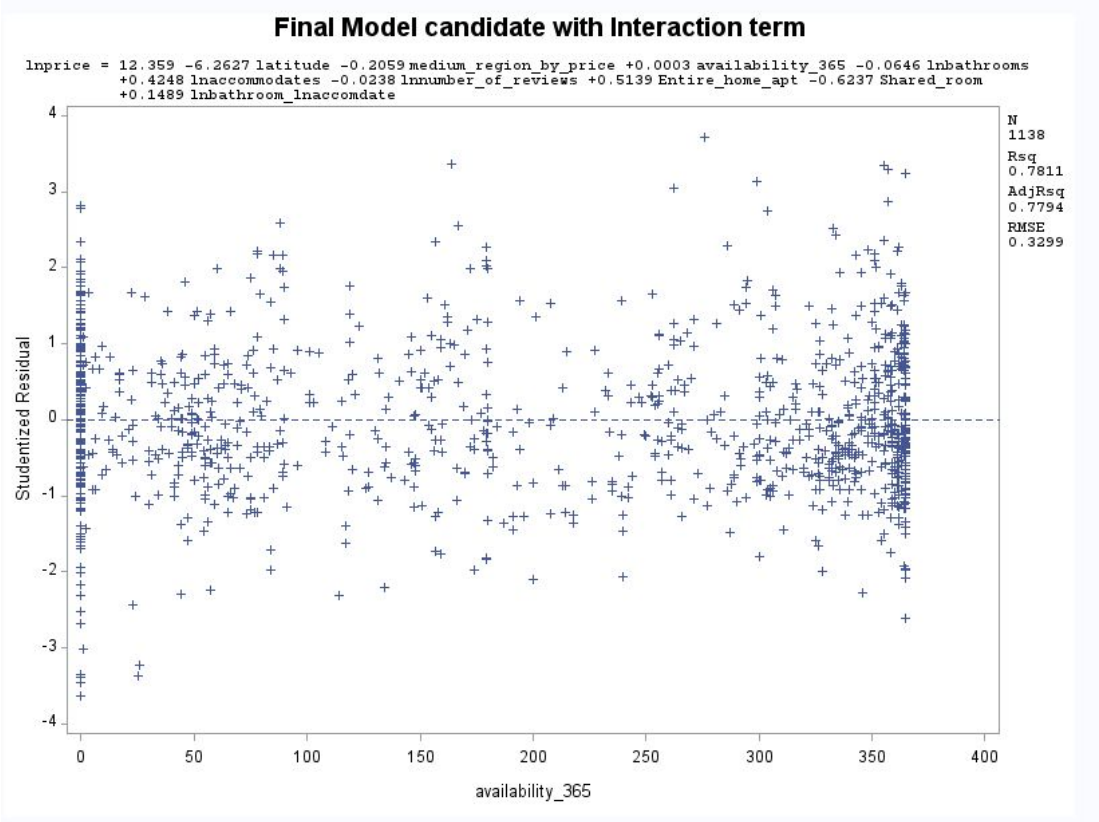
Appendix D: Data Exploration and Analysis SAS Outputs

Model 2: Final Model Candidate with Interaction Term



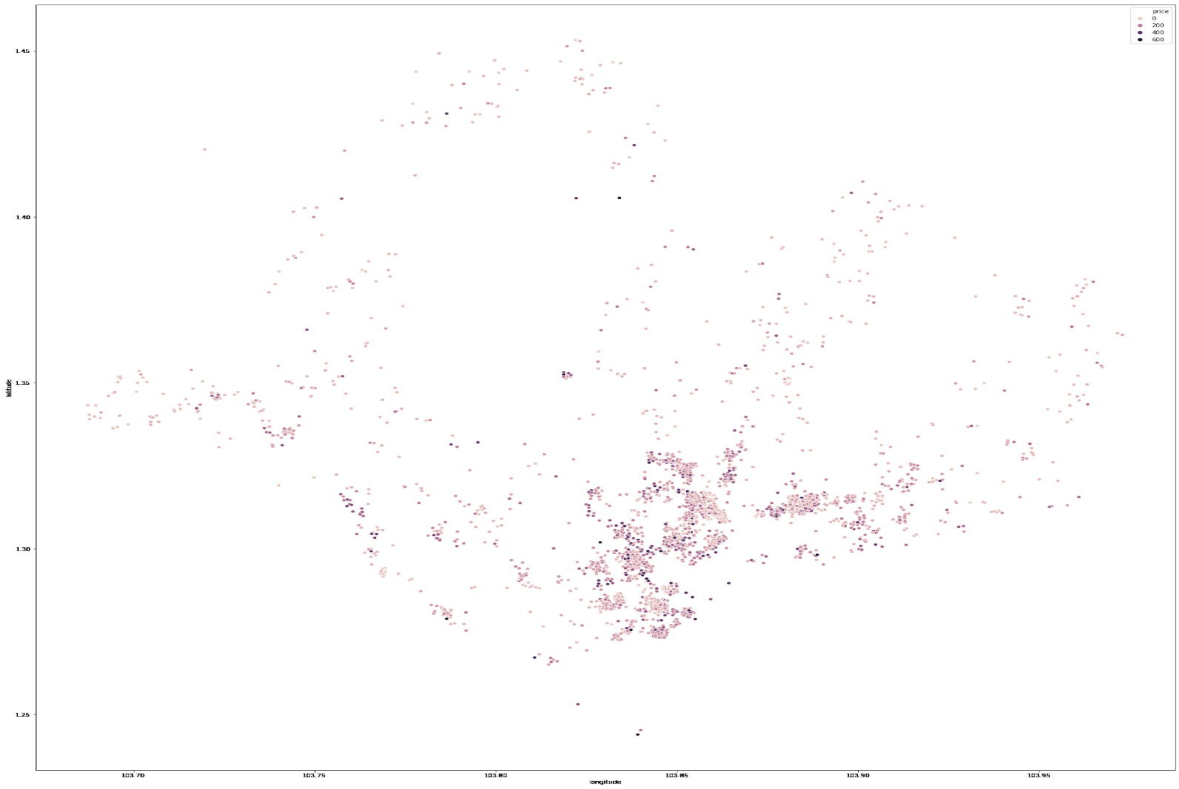
Appendix D: Data Exploration and Analysis SAS Outputs

Model 2: Final Model Candidate with Interaction Term


Final Model candidate with Interaction term
lnprice = 12.359 -6.2627 latitude -0.2059 medium_region_by_price +0.0003 availability_365 -0.0646 lnbathrooms +0.4248 lnaccommodates -0.0238 lnnumber_of_reviews +0.5139 Entire_home_apt -0.6237 Shared_room +0.1489 lnbathroom_lnaccomdate
N 1138
Rsq 0.7811
AdjRsq 0.7794
RMSE 0.3299
Studentized Residual
availability_365
4
3
2
1
0
-1
-2
-3
-4
0
50
100
150
200
250
300
350
400

Appendix E

Geo locations of Airbnb listings observations



Map of Singapore by way of Wikipedia

Appendix F *List of datapoints*

Id: room id, **data_type:** numeric, **variable_type:** index

Name: room names, **data_type:** text, **variable_type:** description

Host_id: host id, **data_type:** numeric, **variable_type:** index

Host_name: host names, **data_type:** text, **variable_type:** description

Neighbourhood_group: Singapore regions, **data_type:** text, **variable_type:** independent variable

Neighbourhood: specific, **place data_type:** text, **variable_type:** independent variable

Latitude: latitude **data_type:** numeric, **variable_type:** independent variable

Longitude: longitude, **data_type:** numeric, **variable_type:** independent variable

Room_type: room type, **data_type:** text, **variable_type:** independent variable

Price: singapore dollar per night, **data_type:** numeric, **variable_type:** dependent variable

Minimum_nights: minimum nights, **data_type:** numeric, **variable_type:** independent variable

Number_of_reviews: number of review, **data_type:** numeric, **variable_type:** independent variable

Last_review: last review, **data_type:** date, **variable_type:** independent variable

Reviews_per_month: number of reviews per monthly aggregate, **data_type:** numeric, **variable_type:** independent variable

Calculated_host_listings_count: total room or house in host catalog on Airbnb, **data_type:** numeric, **variable_type:** independent variable

Availability_365: availability, **data_type:** numeric, **variable_type:** independent variable

As the original dataset found on Kaggle is missing bedroom and accommodation information. Through further investigation on the origin of the Kaggle dataset, we found the complete set of scraped data, http://insideairbnb.com/get-the-data.html, that included the following attributes. We joined the attributes in Python with Id: room id column.

Accommodates: number of people the room can accommodate, **data_type:** numeric, **variable_type:** independent variable
Bathrooms: bathroom count, **data type:** numeric, **variable_type:** independent variable
Bedrooms: bedroom count, **data_type:** numeric, **variable_type:** independent variable
Beds: bed count, **data type:** numeric, **variable_type:** independent variable

Appendix G : Predicted Price Output

Predicted Price

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	1174
Number of Observations Used	758
Number of Observations with Missing Values	416

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	282.44057	35.30507	293.62	<.0001
Error	749	90.05967	0.12024		
Corrected Total	757	372.50024			

Root MSE	0.34676	R-Square	0.7582
Dependent Mean	4.70715	Adj R-Sq	0.7556
Coeff Var	7.36658		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.60853	0.71684	17.59	<.0001
latitude	1	-6.52539	0.53780	-12.13	<.0001
medium_region_by_price	1	-0.21592	0.03051	-7.08	<.0001
availability_365	1	0.00043247	0.00009262	4.67	<.0001
lnbathrooms	1	0.05177	0.02847	1.82	0.0694
lnaccommodates	1	0.53033	0.02705	19.61	<.0001
lnnumber_of_reviews	1	-0.01496	0.00832	-1.80	0.0725
Entire_home_apt	1	0.47486	0.03225	14.73	<.0001
Shared_room	1	-0.57932	0.06481	-8.94	<.0001

Predicted Price 4

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	5.6331	0.1252	5.3873	5.8789	4.9094	6.3569	.
2	3.56	.	.	.	.	.	.	.
3	3.74	4.1395	0.0381	4.0647	4.2143	3.4547	4.8243	-0.4018