Inside Airbnb Boston: Modeling Price of Airbnb listings

Neha Agarwala, Shreya Khurana, Ying Liu

Final Project Report for STAT 425: Applied Regression and Design

Under the guidance of Professor Feng Liang, Department of Statistics
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
19 December 2017

Contents

Introduction

Our dataset comes from Aribnb website, which describes the listing activity of Housing data in Boston, MA. It includes three parts, Listings, Reviews and Calendar. We mainly use Listings for our analysis and Reviews for visualisation.

The objective is to explore the data and study the relationship between the price and other features shown in the dataset.

The dataset Listings contains some of the following important variables:

1. Host attributes: Host response time, Host acceptance rate, superhost or not, host identity verified or not, ways that host has been verified (like email, Facebook etc.)
2. Listing Attributes: Price per night, price per month, Neighbourhood, property type (house, apartment etc.), room type (like private room or entire home being rented out), number of beds, number of bathrooms, cancellation policy, the amenities available in the listing (TV, AC, washer-dryer etc.), availability in a given time frame (30 days, 60 days, 90 days, 1 year)
3. Reviews for Listings: Average ratings of listings on various parameters like cleanliness, communication, check-in, location and value

Our first step was data cleaning since the raw data was not fit for modeling. Based on the cleaned data, our next step was exploratory and graphical analysis. We discovered some interesting visualizations and results of descriptive features and the relationship between them. We also answered some interesting questions based on ANOVA. In our last step, we modeled the price based on all the desciptive features using multiple linear regression.

Data Cleaning

The original data required a lot of cleaning primarily because of three reasons:

1. Missing values
2. Text information
3. Categorical variables

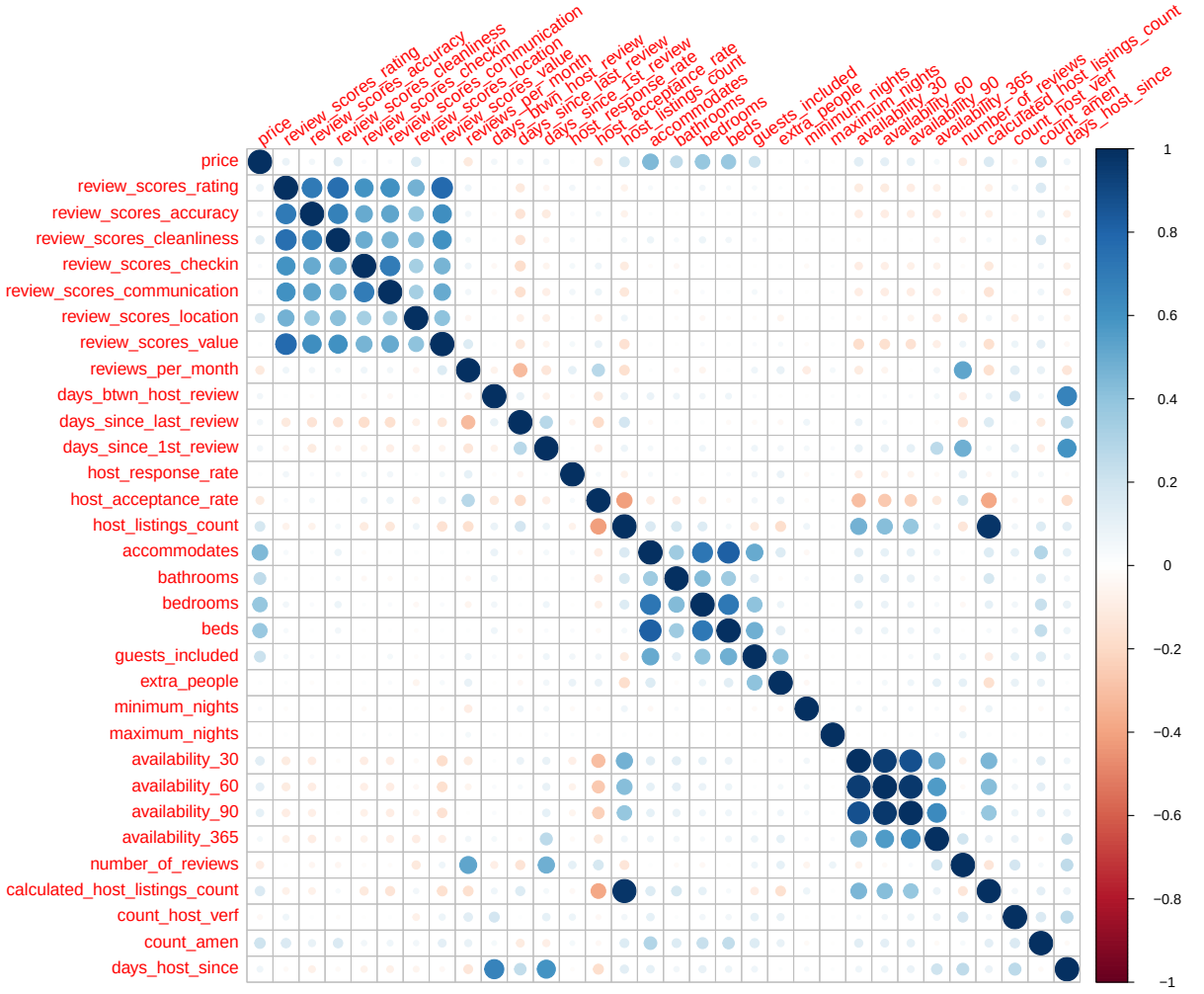
For missing values, we only had variables with less than 25% values or more than 80% values as missing. We treated NA's in numerical features with their median if their missing rates was smaller than 25%; while for categorical features we created a new group named "NA". For those variables whose missing rate were more than 80%, we dropped those variables.

The text variables contained categories and hence the text information was extracted by parsing the data and creating dummy variables for each of them. For example the variable called "Amenties", was formatted like *{"TV", "AC", "24-Hour Check-In"}*. We created a dummy binary variable for *"TV", "AC" and "Check-in"*. Also, some numerical variables were formatted as char type because of a symbol like *%* or *$* in them. We converted them to numerical using text parsing. We created dummy variables for all levels for each categorical variable like room type, apartment type, cancellation policy etc.

Exploratory Data Analysis

Our dataset contains many variables. So we want to see if we can find some relationship among them, specifically their relationship with Price.

Correlation between Variables



In the correlation plot, we can see the first column to find out the degree of correlation that price has with other variables. This will be useful in our model ahead. The darkness of the colour represents the degree and hence we can see that number of beds, bedrooms and the guests that the listing can accommodate has a linear correlation with price. The count of amenities like TV, AC, presence of 24-hour check in option etc (mentioned in the earlier section) also has some effect on price.

One cluster of variables which have a high correlation with each other is this set: {Number of beds, number of bedrooms, guest capacity, number of bathrooms}.

The availability of the listing in a month, in two months and a year are also correlated since very few listings are seasonal and try to give out their homes/ rooms uniformly in a year.

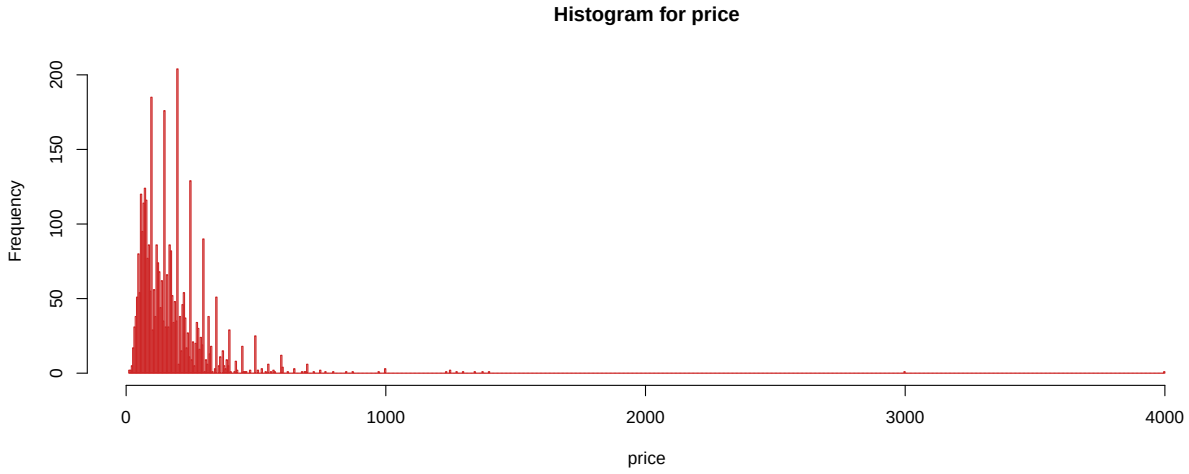
Another interesting correlation can be seen between number of amenities and the guest capacity. This could be because bigger homes tend to provide more options. For example,

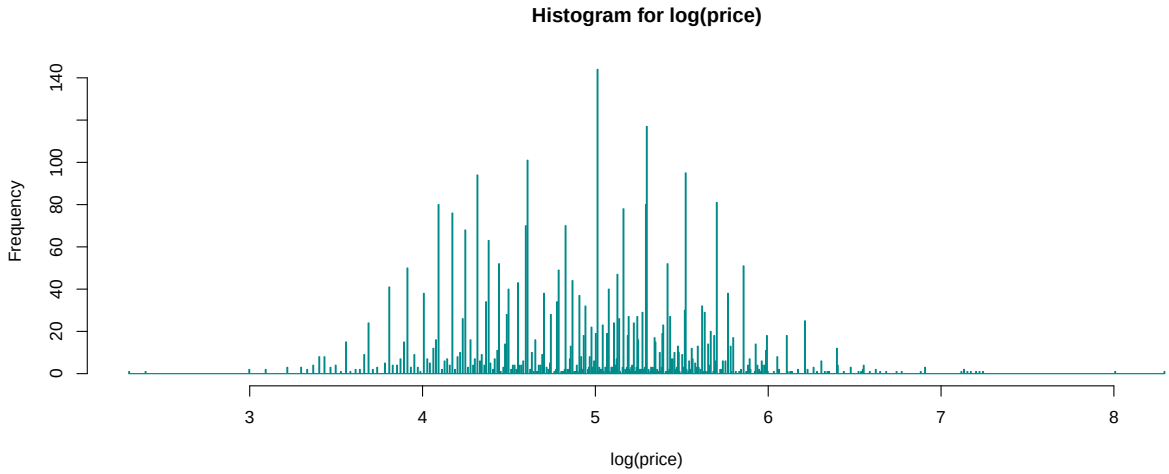
We see that the host listing count and the host acceptance rate are negatively correlated. This is interesting because it means that the fewer the listings a host rents out, the less selective they are about the guests.

One unintuitive result we see is the correlation between availability variables and the host acceptance rate, which is negative i.e. the greater the availability of the listing, the lower the acceptance rate.

Log Transformation of Price

We need to model price as a function of the other variables we see in the data (as well as our own features) and so we should see how the distribution of the response variable looks like. We used log transformation of the price variable and the histograms below show that the transformed variable appears more normal than the original variable. Hence for the furhter EDA and also for modeling, we shall use this transformed response variable.

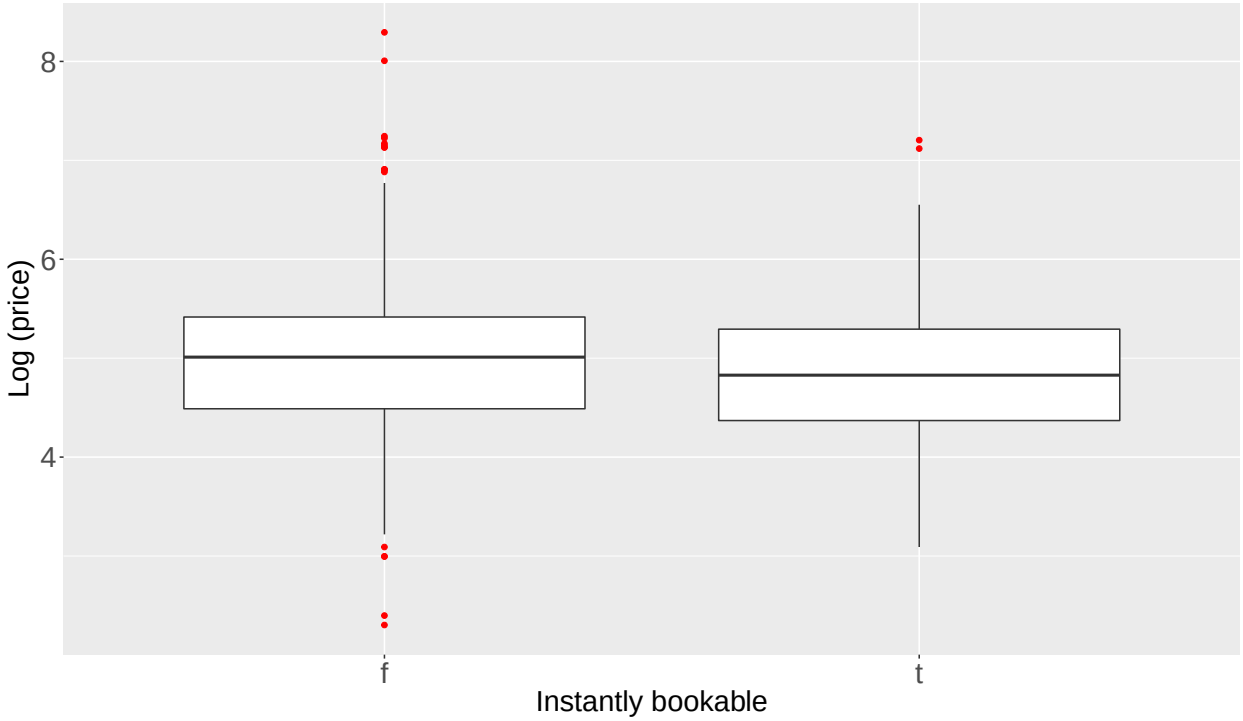




The next plots will show some of the interesting analyses we discovered.

Log Price and Instantly Bookable Listings

We tried to see the relationship of various categorical variables and log price using a boxplot. Instantly bookable listings are those that the customer can book any time and does not need any prior permission or time frame (like minimum number of days required before reservation) to book.

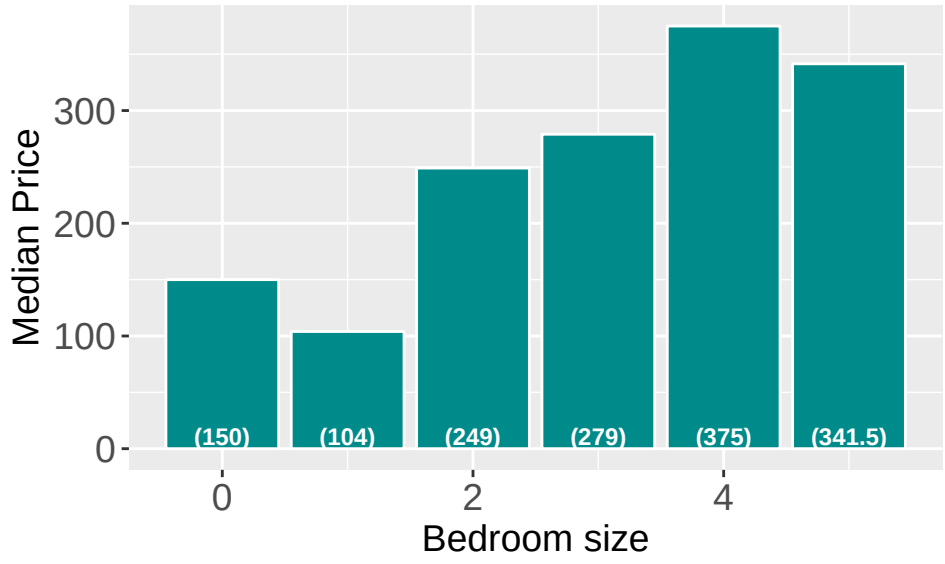


The boxplot of instant bookable shows that the price for when it is not instantly bookable is higher because these hosts are much more selective and hence charge higher in order to filter out applicants. It makes sense because the hosts of instantly bookable listings are not looking for any certain criteria among their customers and charge a lower price in order to get more bookings.

Log Price and Host Response Time

In this figure we can see a very interesting trend. It is not when the response time is the least, that a listing tends to have a higher price, but it is maximum when it is a few hours. This means that responding at the earliest means that the host is very eager to rent out the listing and hence the price is low. On the other end of the spectrum the price decreases as the response time increases i.e. hosts that respond in a few days or more tend to have the least price, maybe because they have a lot more offers and so take their time in deciding who should be their guests.

Median Price and Number of bedrooms

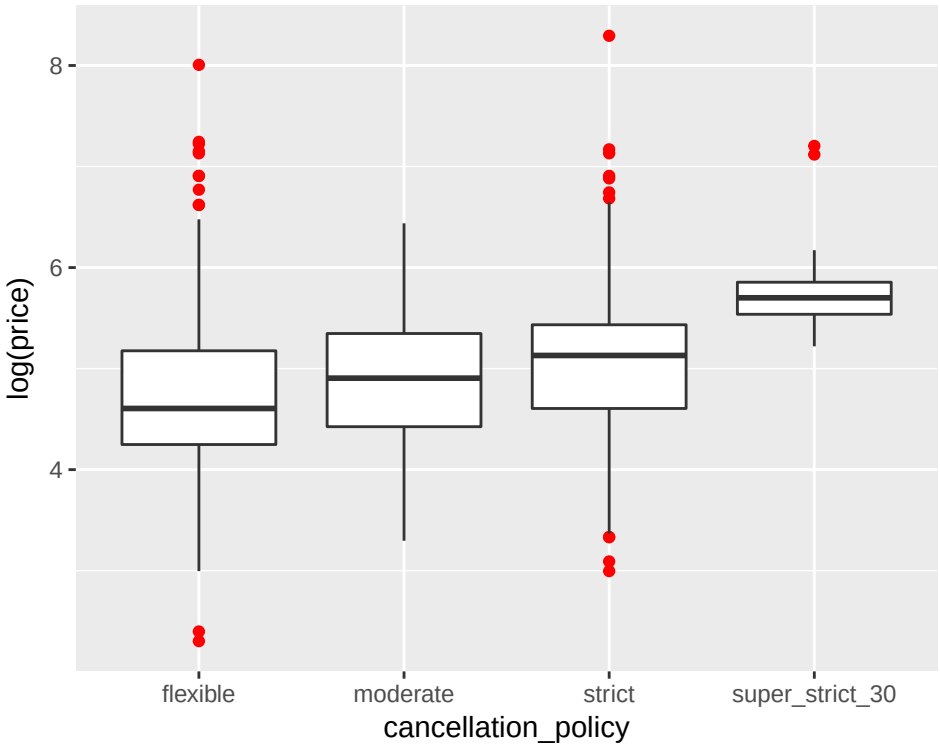


In this plot, we can see a linear trend in the middle portion where number of bedrooms is between 1 and 4, but when we have a studio apartment we see the median price is higher than that of a 1-bedroom and when the number of bedrooms is very high(5), we see economies of scale take place and the price is actually lower than that of a 4-bedroom home.

ANOVA

One-Way ANOVA

Price vs cancellation_policy

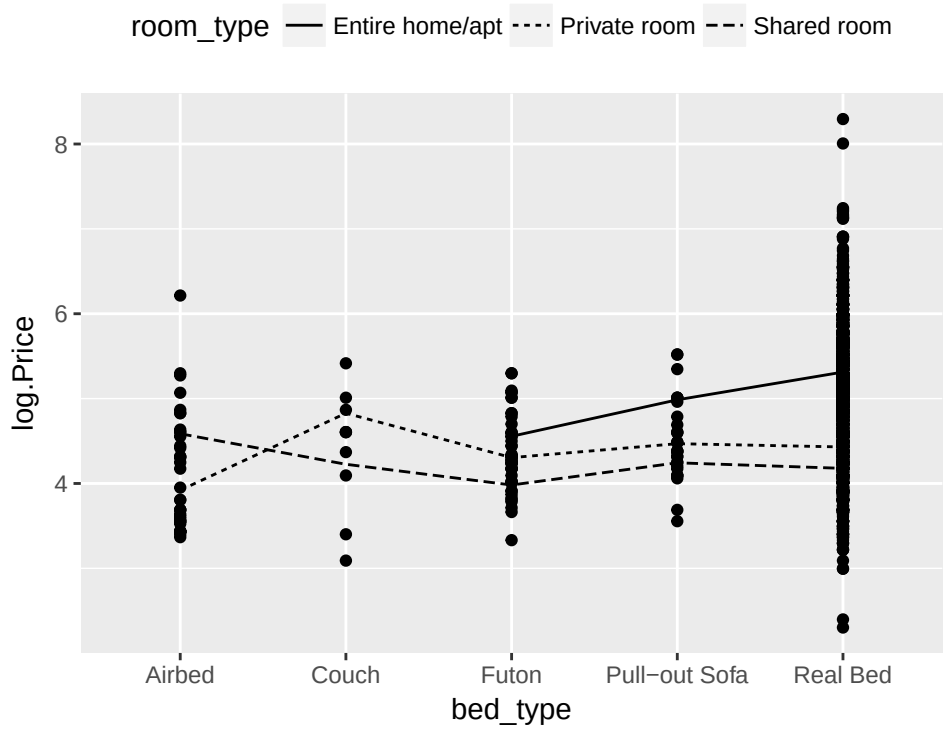


Df Sum Sq Me	an Sq	F value	Pr(>F)		
cancellation_policy	3	116.0709	38.6902905	98.30845	0
Residuals	3581	1409.3390	0.3935602	NA	NA

Rooms with strict cancellation policy would be much more expensive than those with flexible policy. Maybe those hosts who made strict policies are really mean. Or maybe those rooms are so nice and so popular that they should be booked in advance. Flexible cancellation policy may bring some unexpected and serious loss to the hosts. So they make strict cancellation policy to avoid this kind of loss.

Two-Way ANOVA

Price vs bed_type & room_type

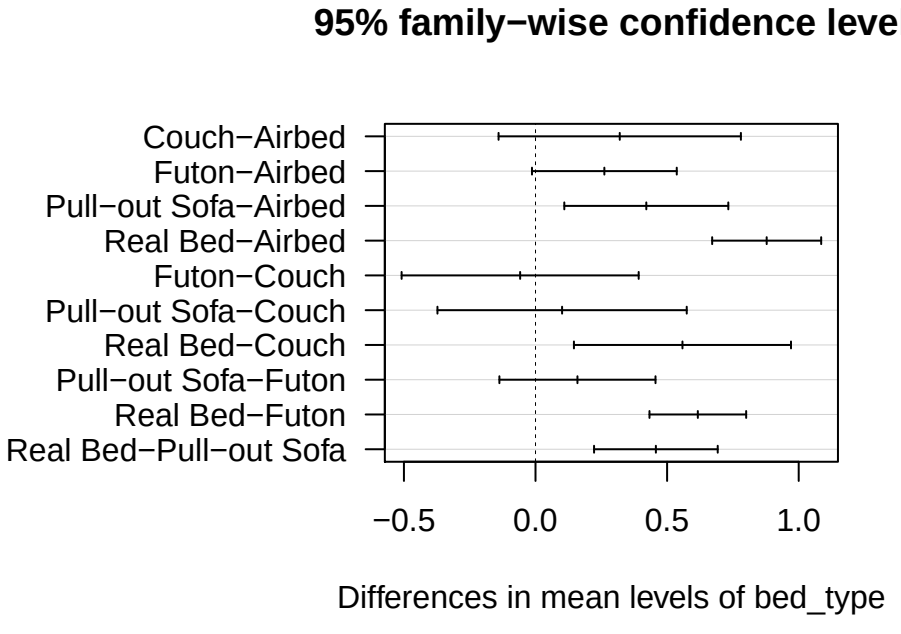


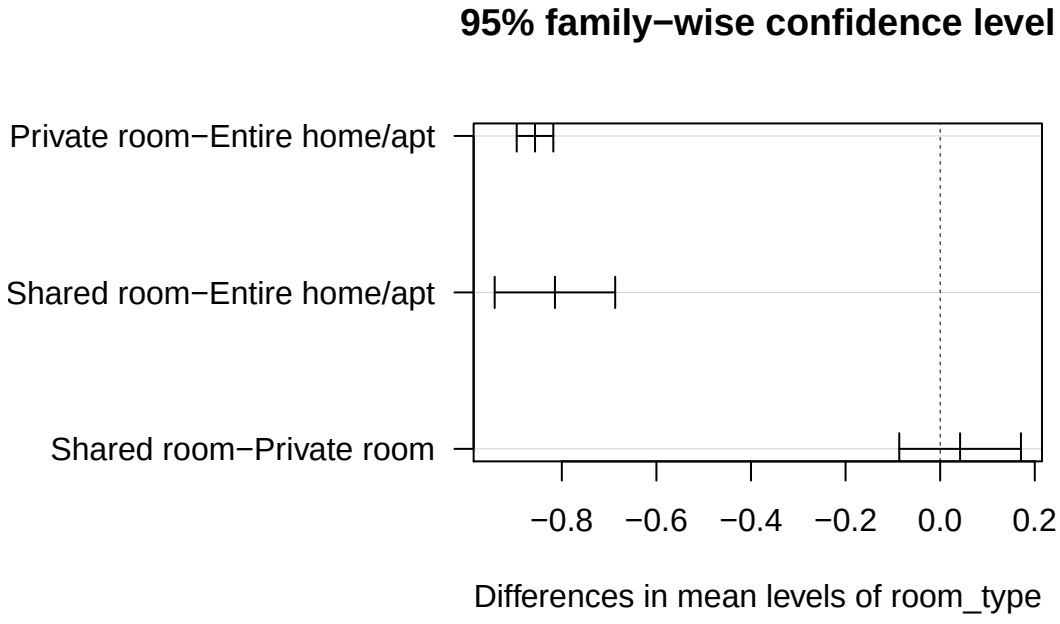
Df Sum Sq	Mean Sq	F val	ue Pr(>F)		
bed_type	4	57.77488	14.4437191	64.160748	0
room_type	2	653.09602	326.5480096	1450.565778	0
bed_type:room_type	6	10.41862	1.7364370	7.713463	0
Residuals	3572	804.12037	0.2251177	NA	NA

Since the slopes of lines are not parallel, we can conclude that the interaction effect is significant.

Tukey's Honest Significant Difference (HSD)

The differences of price between pull-out sofa and airbed, real bed and all other types of bed are quite significant, as can be seen below in the figure.

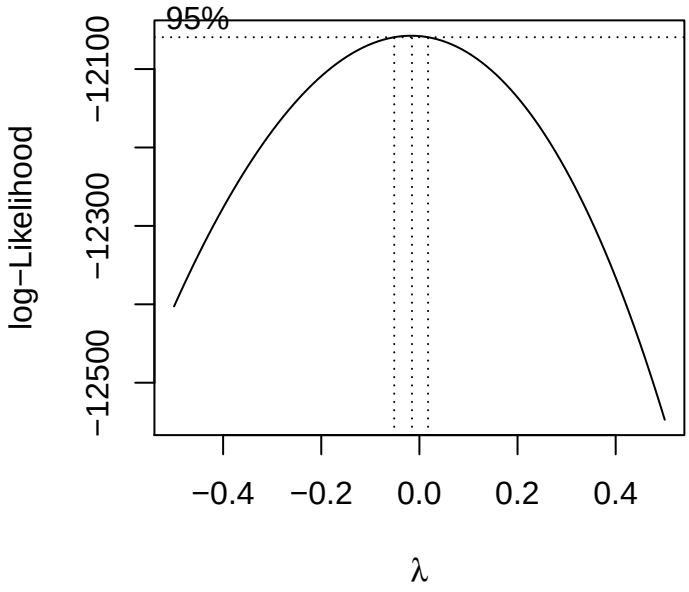




Differences of the price between entire home and shared room, private room and entire home are quite significant, but there's no significant difference between the price of shared room and private room .

Modeling - MLR & Ancova

Suppose we want to find out the factors that drive the price of airbnb rental homes. Of course we can build complex models with high prediction accuracy. However our goal is to develop a classical ancova model which is easy to interprete and understand using complex models to reduce the number of predictors. ## Box-cox transformation



The idea is to use mlr model with all numerical predictors to validate whether the transformation of price to log(price) is supported by box-cox transformation which is quite evident here as 0 is included in the 95% confidence interval.

Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.3403309	0.1126756	20.770519	0.0000000
count_amen	0.0237782	0.0026386	9.011576	0.0000000
review_scores_cleanliness	0.0639188	0.0098462	6.491741	0.0000000
review_scores_location	0.1724240	0.0109492	15.747696	0.0000000
review_scores_value	-0.0653879	0.0120500	-5.426376	0.0000001
reviews_per_month	-0.0271987	0.0045242	-6.011854	0.0000000
days_since_last_review	0.0003325	0.0000590	5.633113	0.0000000
accommodates	0.1956779	0.0084658	23.113982	0.0000000
beds	-0.0348697	0.0140315	-2.485100	0.0129971
guests_included	0.0281441	0.0090121	3.122912	0.0018051
availability_90	0.0008421	0.0002661	3.164481	0.0015667
calculated_host_listings_count	0.0021463	0.0003211	6.684455	0.0000000
days_host_since	0.0000617	0.0000134	4.598653	0.0000044

The multiple linear regression model is optimum with respect to the BIC criterion and has r-square 0.46. We will see later how the inclusion of categorical predictors increases the r-square to 0.74. We use this model to check the assumptions of linear regression and conduct other diagonistics(outliers or high influential points). Many observations were declared as outliers using bonferroni correction. However we removed only "some" which were not falling in pattern of the overall price range.

Ancova

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9486977	0.1323468	14.724180	0.0000000
count_amen	0.0231372	0.0026211	8.827271	0.0000000
review_scores_cleanliness	0.0642364	0.0097883	6.562538	0.0000000
review_scores_location	0.1669000	0.0108989	15.313529	0.0000000
review_scores_value	-0.0614341	0.0119872	-5.124979	0.0000003
reviews_per_month	-0.0275713	0.0044932	-6.136183	0.0000000
days_since_last_review	0.0003325	0.0000587	5.668359	0.0000000
accommodates	0.1898629	0.0084511	22.466160	0.0000000
beds	-0.0308664	0.0139541	-2.211998	0.0270302
guests_included	0.0266847	0.0089480	2.982189	0.0028814
availability_90	0.0009139	0.0002646	3.453867	0.0005591
calculated_host_listings_count	0.0020922	0.0003191	6.556561	0.0000000
days_host_since	0.0000609	0.0000133	4.567371	0.0000051
bed_typeCouch	0.2871571	0.1683079	1.706142	0.0880690
bed_typeFuton	0.1346691	0.1013614	1.328603	0.1840642
bed_typePull-out Sofa	0.2281365	0.1144303	1.993671	0.0462643
bed_typeReal Bed	0.4364732	0.0771151	5.660023	0.0000000

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3558	811.4668	NA	NA	NA	NA
3554	798.6646	4	12.8022	14.24222	0

Here we try to illustrate the problem of including the categorical variable, say bed_type with all its levels. Though the F-test for the inclusion of categorical variable is significant, bed_typeFuton and bed_typeCouch are not significant at 1% level. Therefore, we need to create dummy variables for all the categorical predictors.

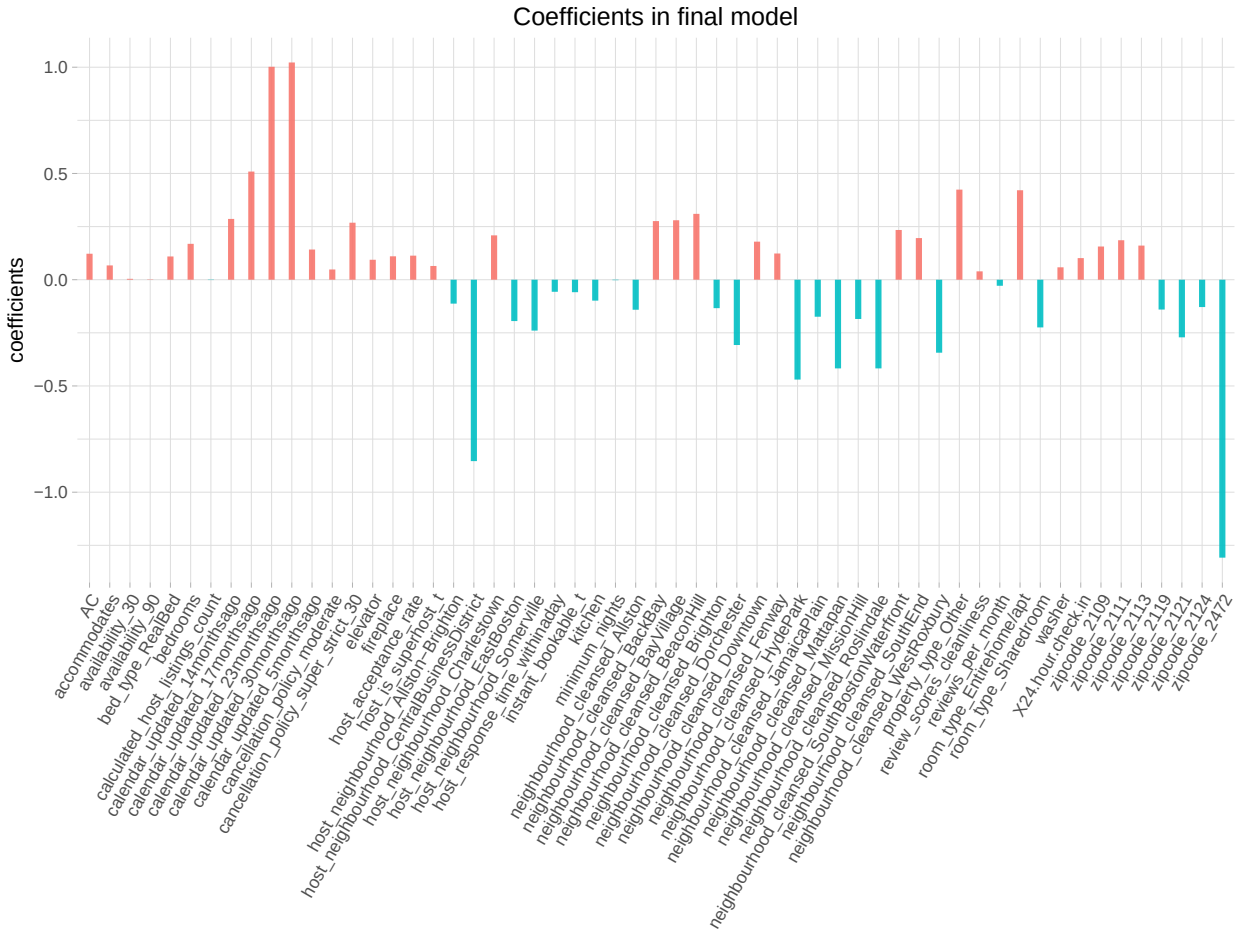
Final Model

Using Lasso, we reduced the number of variables including dummy variables (from 275 to 152) and finally used this set to obtain our Ancova model using the BIC criterion and response as log(price). The final model displayed below has r-square 0.74.

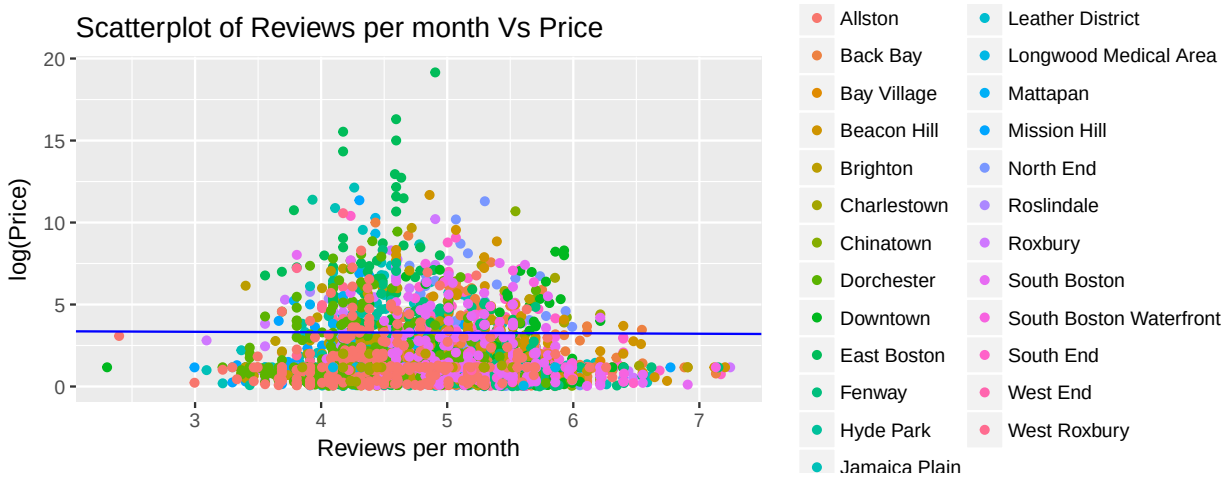
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	3.5249243	0.0727180	48.473858	0.0000000
review_scores_cleanliness	0.0393335	0.0054253	7.249967	0.0000000
reviews_per_month	-0.0285500	0.0033538	-8.512639	0.0000000
host_acceptance_rate	0.1131790	0.0336356	3.364854	0.0007741
accommodates	0.0672394	0.0055402	12.136702	0.0000000
bedrooms	0.1688468	0.0114868	14.699166	0.0000000

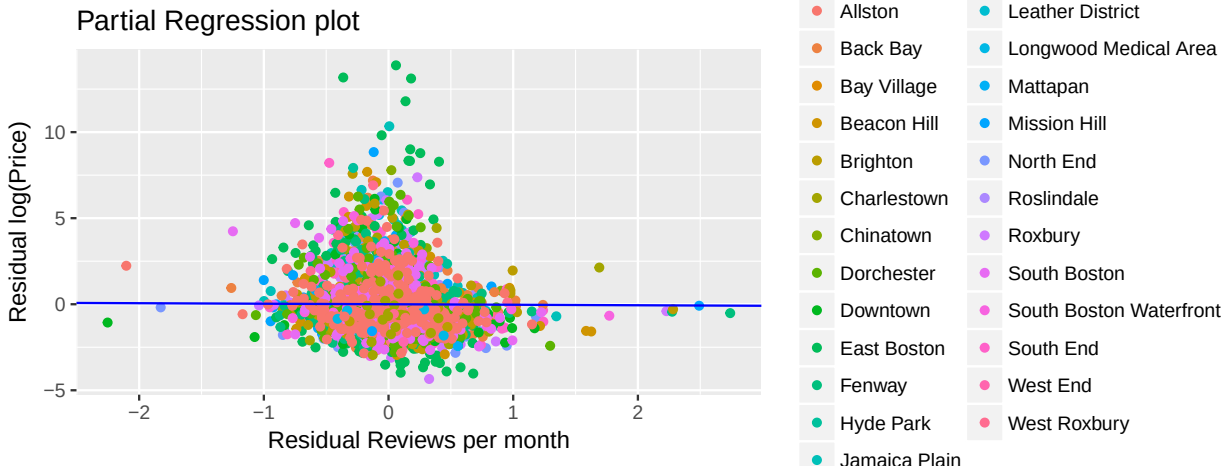
	Estimate	Std. Error	t value	Pr(>|t|)
minimum_nights	-0.0023076	0.0006350	-3.633992	0.0002831
availability_30	0.0041903	0.0011742	3.568574	0.0003637
availability_90	0.0016623	0.0003573	4.651894	0.0000034
calculated_host_listings_count	-0.0015383	0.0002559	-6.011235	0.0000000
host_response_time_withinaday	-0.0567861	0.0176918	-3.209731	0.0013406
host_neighbourhood_CentralBusinessDistrict	-0.8536849	0.0713040	-11.972463	0.0000000
host_neighbourhood_Charlestown	0.2087156	0.0447352	4.665574	0.0000032
host_neighbourhood_EastBoston	-0.1945971	0.0358705	-5.424984	0.0000001
host_neighbourhood_Somerville	-0.2394263	0.0705870	-3.391933	0.0007017
zipcode_2109	0.1562135	0.0499245	3.128998	0.0017684
zipcode_2111	0.1856714	0.0371622	4.996249	0.0000006
zipcode_2113	0.1605645	0.0374311	4.289598	0.0000184
zipcode_2119	-0.1403838	0.0367980	-3.814982	0.0001385
zipcode_2121	-0.2712199	0.0650153	-4.171629	0.0000310
zipcode_2124	-0.1286764	0.0447592	-2.874863	0.0040663
zipcode_2472	-1.3079810	0.3361646	-3.890895	0.0001017
host_is_superhost_t	0.0645796	0.0186393	3.464697	0.0005372
neighbourhood_cleansed_BackBay	0.2758926	0.0258534	10.671406	0.0000000
neighbourhood_cleansed_BayVillage	0.2797622	0.0702371	3.983111	0.0000694
neighbourhood_cleansed_BeaconHill	0.3098554	0.0294992	10.503866	0.0000000
neighbourhood_cleansed_Brighton	-0.1338892	0.0422112	-3.171890	0.0015276
neighbourhood_cleansed_Dorchester	-0.3073109	0.0300145	-10.238741	0.0000000
neighbourhood_cleansed_Downtown	0.1791139	0.0312004	5.740760	0.0000000
neighbourhood_cleansed_Fenway	0.1233369	0.0265407	4.647085	0.0000035
neighbourhood_cleansed_HydePark	-0.4699568	0.0624269	-7.528114	0.0000000
neighbourhood_cleansed_JamaicaPlain	-0.1741825	0.0248697	-7.003796	0.0000000
neighbourhood_cleansed_Mattapan	-0.4173557	0.0707219	-5.901367	0.0000000
neighbourhood_cleansed_MissionHill	-0.1847909	0.0346257	-5.336809	0.0000001
neighbourhood_cleansed_Roslindale	-0.4174219	0.0487477	-8.562911	0.0000000
neighbourhood_cleansed_SouthBostonWaterfront	0.2340310	0.0417212	5.609402	0.0000000
neighbourhood_cleansed_SouthEnd	0.1959276	0.0251648	7.785771	0.0000000
neighbourhood_cleansed_WestRoxbury	-0.3435317	0.0521851	-6.582944	0.0000000
property_type_Other	0.4240065	0.0831432	5.099710	0.0000004
room_type_Sharedroom	-0.2244794	0.0406735	-5.519060	0.0000000
bed_type_RealBed	0.1094138	0.0319275	3.426950	0.0006174
calendar_updated_14monthsago	0.2860850	0.0773789	3.697196	0.0002213
calendar_updated_17monthsago	0.5089690	0.1117114	4.556108	0.0000054
calendar_updated_23monthsago	1.0020986	0.3349130	2.992116	0.0027898
calendar_updated_30monthsago	1.0222823	0.3328840	3.070986	0.0021499
calendar_updated_5monthsago	0.1419857	0.0453475	3.131058	0.0017561
instant_bookable_t	-0.0587648	0.0163740	-3.588914	0.0003366
cancellation_policy_moderate	0.0480965	0.0133150	3.612190	0.0003079
cancellation_policy_super_strict_30	0.2682789	0.0403411	6.650270	0.0000000
washer	0.0583520	0.0132146	4.415723	0.0000104
X24.hour.check.in	0.1016986	0.0151244	6.724128	0.0000000
kitchen	-0.0985601	0.0221029	-4.459139	0.0000085
elevator	0.0938826	0.0168709	5.564760	0.0000000
fireplace	0.1101264	0.0183946	5.986872	0.0000000
AC	0.1222319	0.0155374	7.866934	0.0000000
host_neighbourhood_Allston-Brighton	-0.1125425	0.0360140	-3.124967	0.0017928
neighbourhood_cleansed_Allston	-0.1413224	0.0383751	-3.682656	0.0002343
room_type_Entirehome/apt	0.4210127	0.0163416	25.763193	0.0000000

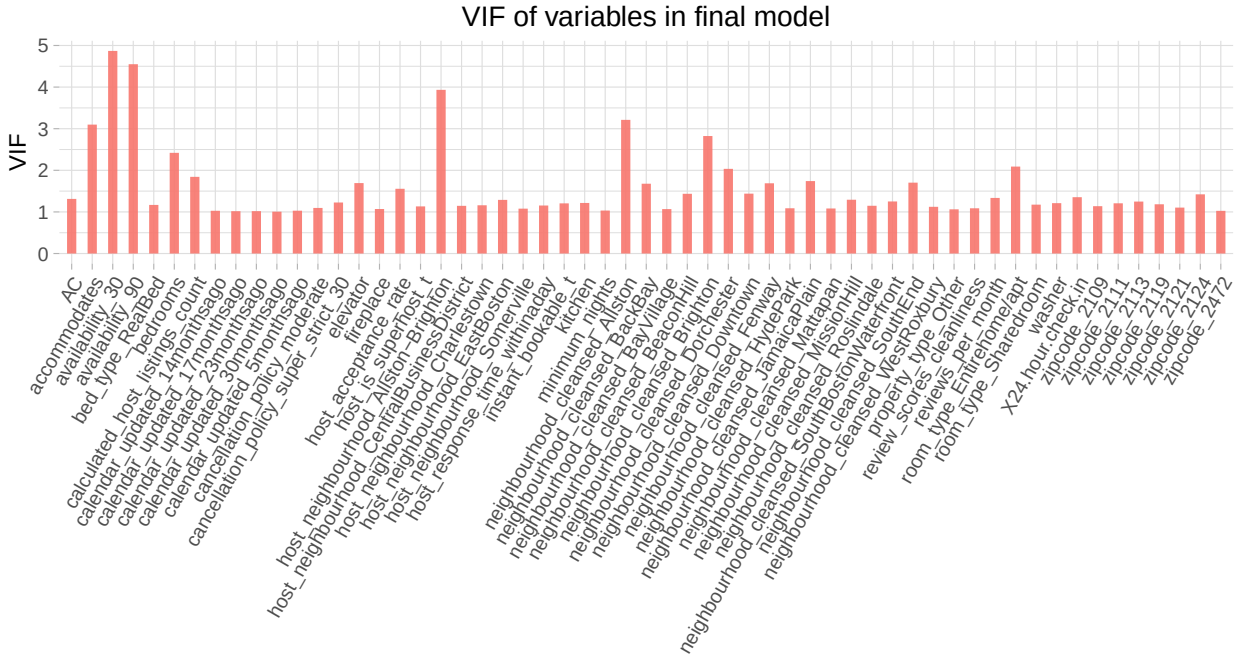
Interpretation of coefficients: For homes with other variables fixed, the ones with room type Entirehome/apt have higher price and room type Shared room has lower price. For an extra unit of bedroom, all types of homes charge higher price.



It is interesting to see that dummy variables AC, washer, room-type_Entirehome/apt have positive coefficient whereas instant_bookable_true, room-type_SharedRoom,etc have negative coefficients for response log(price). Though it was not totally clear from the original scatterplot, partial regression plot reflected the true association between the review per month and price which is practically 0, also evident from the coefficient for Reviews-per-month which is -0.029.


Scatterplot of Reviews per month Vs Price
log(Price)
Reviews per month
Allston
Back Bay
Bay Village
Beacon Hill
Brighton
Charlestown
Chinatown
Dorchester
Downtown
East Boston
Fenway
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury


Partial Regression plot
Residual log(Price)
Residual Reviews per month
Allston
Back Bay
Bay Village
Beacon Hill
Brighton
Charlestown
Chinatown
Dorchester
Downtown
East Boston
Fenway
Hyde Park
Jamaica Plain
Leather District
Longwood Medical Area
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury


VIF of variables in final model
VIF
accommodates
availability_30
availability_90
bed_type_RealBed
bedrooms
calculated_host_listings_count
calendar_updated_14months ago
calendar_updated_17months ago
calendar_updated_23months ago
calendar_updated_30months ago
calendar_updated_5months ago
cancellation_policy_moderate
cancellation_policy_super_strict_30
elevator
fireplace
host_acceptance_rate
host_is_superhost_t
host_neighbourhood_Allston-Brighton
host_neighbourhood_CentralBusinessDistrict
host_neighbourhood_Charlestown
host_neighbourhood_EastBoston
host_neighbourhood_Somerville
host_response_time_within a day
instant_bookable_t
kitchen
minimum_nights
neighbourhood_cleansed_Allston
neighbourhood_cleansed_BackBay
neighbourhood_cleansed_BayVillage
neighbourhood_cleansed_BeaconHill
neighbourhood_cleansed_Brighton
neighbourhood_cleansed_Dorchester
neighbourhood_cleansed_Downtown
neighbourhood_cleansed_Fenway
neighbourhood_cleansed_HydePark
neighbourhood_cleansed_JamaicaPlain
neighbourhood_cleansed_Mattapan
neighbourhood_cleansed_MissionHill
neighbourhood_cleansed_Roslindale
neighbourhood_cleansed_SouthBostonWaterfront
neighbourhood_cleansed_SouthEnd
neighbourhood_cleansed_WestRoxbury
property_type_Other
review_scores_cleanliness
reviews_per_month
room_type_Entire home/apt
room_type_Shared room
washer
X24.hour.check.in
zipcode_2109
zipcode_2111
zipcode_2113
zipcode_2119
zipcode_2121
zipcode_2124
zipcode_2472

From the plot, it is very intuitive to see variables with high VIF values: are high for variables: availability_30 and availability_90 and also host neighbourhood and neighbourhood of the home for Allston. Nevertheless the lasso selection of variables criteria made sure that VIF is less than 5.

We use random forest to obtain the importance of variables significant in the final model. It is interesting to note that room-type "Entire home/apt" is the most important variable followed by accomodates.

Conclusion

- One-way Anova: There is a significant difference in the price of airbnb homes in Boston with respect to the cancellation policy.
- Two-way Anova: Prices seem to be affected by bed-type, room-type and their interaction. Differences in the price between entire home and shared room, private room and entire home are quite significant, but there's no significant difference between the price of shared room and private room. The differences in price between pull-out sofa and airbed and between real bed and all other types of bed are quite significant.
- Modeling: Interesting variables like bedrooms, accommodates, bed_type_RealBed, room_type_Entirehome/apt, AC, host_is_superhost_t seem to have a positive effect on the price whereas room_type_Sharedroom, calculated_host_listings_count, instant_bookable_t seem to have negative effect on the price.

Acknowledgement

We would like to thank Professor Feng Liang and TA Yujia Deng for their support and guidance in making this project (hopefully) a success.