Marketing Analytics: Airbnb Final Project - Hawaii

Group 6: Aditi, Airi, Pankhuri, Pooja

Quick Recap: *Hawaii, United States*

In our project, we picked ***Maui***, which includes the most popular and expensive neighborhoods, Kihei and Lahaina. Maui also has the highest average price per night in Hawaii.

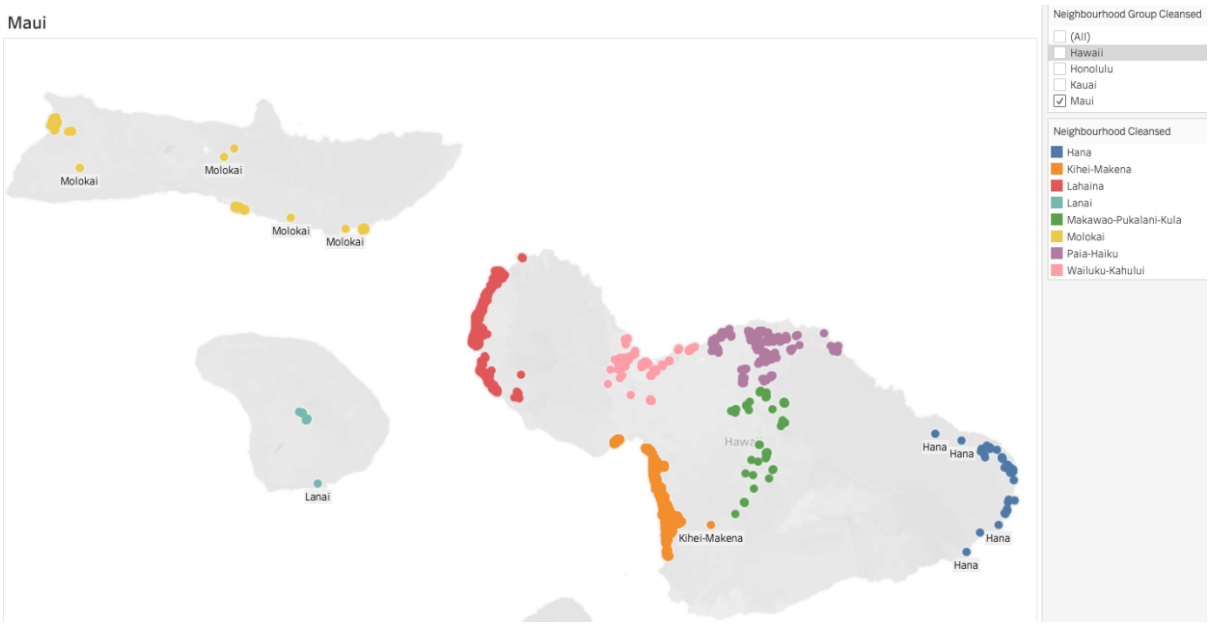
As a ***guest*** of Airbnb, they would like to know the ***average price per night***, which is the phenomenon of interest (DV/y variable). Guests may have questions, like how much is the average price per night in a certain neighborhood with certain number of people, room type, and the number of bedrooms. These questions are important in determining and predicting the budget for the stay in Maui.

Data: We used the detailed listing data from Hawaii, Inside Airbnb

- **Listings_detailed**
- **Reviews**

Based on intuition, the key variables/factors that affect and can predict your phenomenon of interest, (independent variables = x variables) are:

- *Room Type*
- *Instant Bookable*
- *Neighborhood in Maui*
- *Number of Bedrooms*
- *Number of People (Accommodates)*

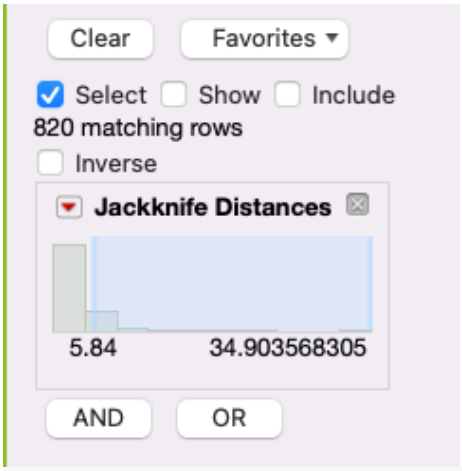


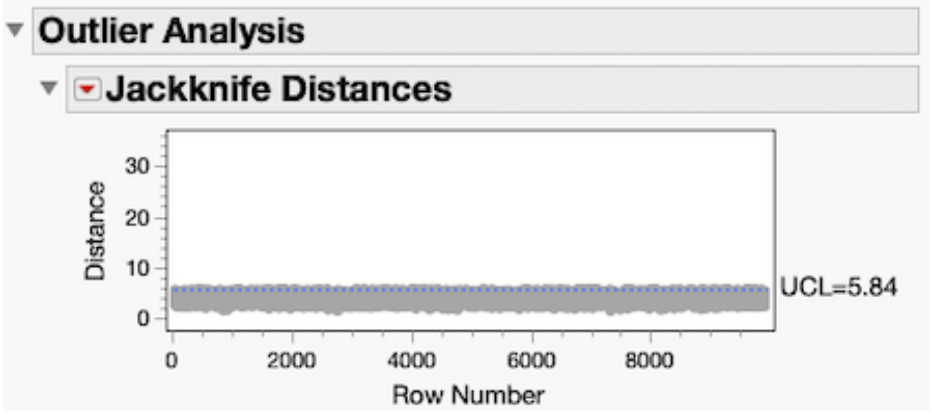
Recap: Reccomendation

We concluded that all the variables influence the price, but the neighborhood influences the price the most in our model.
Kihei and Lahaina are the most visited neighborhoods in Maui and that's where we can find the most Airbnb listings. As a guest, you must make adjustments to the variables in considering the budget.
By using the profiler, we were able to predict the average price per night using these variables.
Students Vacation Trip: Private room, Instantly Bookable True, 4 people, Molokai, 2 bedroom - on average, $270 price per night.
Couple Vacation: Private room, Instantly Bookable True, 2 people, Lahaina, 1 bedroom - on average, $493 price per night.
Family Get-Together: Entire Home, Instantly Bookable True, 20 people, Lahaina, 10 bedroom - on average, $18,033 price per night.

Jackknife analysis (to remove outliers):

We used jackknife analysis to remove outliers, and we excluded 820 observations.





Unsupervised Learning:

PCA/Factor Analysis:

a) Excluding your Y-variable, run the analyses with your potential continuous x-variables (the more the better)

We picked 22 potential continuous x-variables

Number of Bedrooms
Number of People (Accommodates)
Bathrooms_text (number of bathroom)
Number of Beds (bed)
Minimum Nights
Maximum Nights
Number of reviews per month
Number of Reviews
Number of reviews ltm
Number of reviews last 30 days
availability_30
availability_60
availability_90
availability_365
calculated_host_listings_count
Review_scores_rating
Review_scores_accuacy
Review_scores_cleanliness
Review_scores_checkin
Review_scores_communication
Review_scores_location
Review_scores_value

Some of these variables are highly correlated (# of bedrooms, # of bathrooms, availability, and # of reviews).
These may result in multicollinearity - variables have a higher correlation with one another and may undermine the statistically significant of an independent variable.

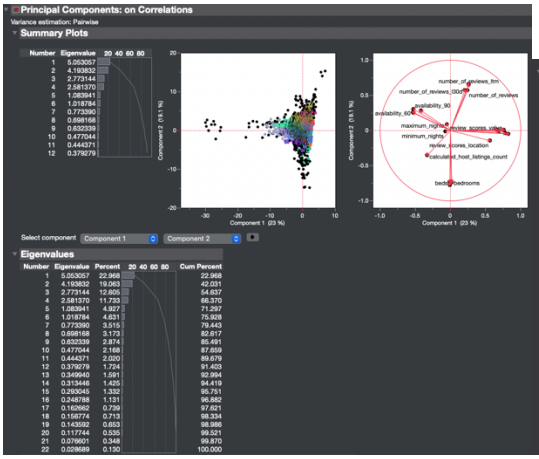
Multivariate

Correlations

	accommodates	number of bathroom	bedrooms	beds	minimum_nights	maximum_nights	availability_30	availability_60	availability_90	availability_365	number_of_reviews	number_of_
accommodates	1.0000	0.7688	0.8404	0.8397	-0.0140	-0.0386	-0.1092	-0.0689	-0.1040	-0.0749	-0.2031	
number of bathroom	0.7688	1.0000	0.8195	0.6986	-0.0097	-0.0149	-0.1176	-0.0704	-0.0929	-0.0515	-0.1713	
bedrooms	0.8404	0.8195	1.0000	0.8388	0.0317	-0.0081	-0.1127	-0.0609	-0.0912	-0.0386	-0.1196	
beds	0.8397	0.6986	0.8388	1.0000	-0.0161	-0.0174	-0.1100	-0.0602	-0.0906	-0.0517	-0.1590	
minimum_nights	-0.0140	-0.0097	0.0317	-0.0161	1.0000	0.0410	0.1032	0.0761	0.0593	0.0003	-0.0553	
maximum_nights	-0.0386	-0.0149	-0.0081	-0.0174	0.0410	1.0000	0.0223	0.0357	0.0520	0.0527	0.1483	
availability_30	-0.1092	-0.1176	-0.1127	-0.1100	0.1032	0.0223	1.0000	0.8864	0.8121	0.5228	-0.0991	
availability_60	-0.0689	-0.0704	-0.0609	-0.0602	0.0761	0.0357	0.8864	1.0000	0.9556	0.6491	-0.0727	
availability_90	-0.1040	-0.0929	-0.0912	-0.0906	0.0593	0.0520	0.8121	0.9556	1.0000	0.7150	-0.0423	
availability_365	-0.0749	-0.0515	-0.0386	-0.0517	0.0003	0.0527	0.5228	0.6491	0.7150	1.0000	0.0642	
number_of_reviews	-0.2031	-0.1713	-0.1196	-0.1590	-0.0553	0.1483	-0.0991	-0.0727	-0.0423	0.0642	1.0000	
number_of_reviews_ltm	-0.2250	-0.2003	-0.1376	-0.1673	-0.0770	0.0968	-0.0836	-0.0405	0.0029	0.1176	0.8060	
number_of_reviews_l30d	-0.1697	-0.1608	-0.1008	-0.1253	-0.0587	0.0378	-0.0232	0.0261	0.0594	0.1224	0.5416	
review_scores_rating	0.0137	0.0519	0.0299	0.0082	-0.1125	-0.0468	-0.1841	-0.1747	-0.1770	-0.1551	0.0423	
review_scores_accuracy	0.0130	0.0471	0.0193	0.0015	-0.0283	-0.0334	-0.1848	-0.1729	-0.1747	-0.1357	0.0674	
review_scores_cleanliness	-0.0073	0.0253	0.0096	-0.0115	0.0097	-0.0261	-0.1388	-0.1221	-0.1296	-0.1291	0.0446	
review_scores_checkin	0.0024	0.0287	0.0161	-0.0058	0.0006	-0.0395	-0.1824	-0.1586	-0.1499	-0.1106	0.1068	
review_scores_communication	0.0182	0.0184	0.0303	0.0107	-0.0185	-0.0379	-0.2139	-0.1886	-0.1900	-0.1580	0.1083	
review_scores_location	0.0618	0.0756	0.0350	0.0411	-0.0151	-0.0340	-0.1357	-0.1356	-0.1467	-0.1322	-0.0337	
review_scores_value	-0.0227	-0.0054	-0.0239	-0.0328	0.0197	-0.0218	-0.1998	-0.2086	-0.2133	-0.1878	0.1046	
calculated_host_listings_count	0.2183	0.2211	0.1529	0.1976	-0.0765	0.0178	0.0937	0.0885	0.0973	0.1415	-0.2719	
reviews_per_month	-0.2214	-0.2367	-0.1442	-0.1741	-0.0617	-0.0054	-0.0592	-0.0187	0.0202	0.0832	0.6451	

b) How many factors do you get? Label at least the first three (If you get less than 3 factors, label all your factors)

To avoid multicollinearity, we conducted the PCA to summarize large number of similar/redundant variables into several factors. As a result, we got *6 components* with their Eigenvalues > 1. These components capture approximately *75.93% (75.928%)* of information contained in our 22 variables.



Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
accommodates	0.007794	**0.922750**	-0.136175	-0.052945	-0.034504	-0.025898
number of bathroom	0.042029	**0.876468**	-0.131813	-0.043635	-0.046955	0.019968
bedrooms	0.012372	**0.949861**	-0.037465	-0.036982	0.032434	0.001633
beds	-0.006712	**0.910700**	-0.077337	-0.043158	-0.031461	-0.008906
minimum_nights	-0.036499	0.028149	-0.108391	0.071383	**0.886612**	0.102190
maximum_nights	-0.027394	-0.014899	0.069877	0.026947	0.031047	**0.967874**
availability_30	-0.132037	-0.093516	-0.095504	**0.880361**	0.084134	-0.017685
availability_60	-0.111580	-0.029640	-0.034667	**0.962429**	0.047110	-0.008976
availability_90	-0.117411	-0.056001	0.008321	**0.956844**	0.011204	0.011691
availability_365	-0.111777	-0.000901	0.129111	**0.776761**	-0.125187	0.054511
number_of_reviews	0.034503	-0.095313	**0.823457**	-0.053504	0.016509	0.168393
number_of_reviews_ltm	0.044146	-0.098766	**0.950465**	-0.001119	-0.013331	0.057688
number_of_reviews_l30d	0.042312	-0.055621	**0.831912**	0.067546	-0.020793	-0.038274
review_scores_rating	**0.918502**	0.011144	0.029902	-0.067110	-0.102113	-0.018720
review_scores_accuracy	**0.869607**	0.010516	0.034495	-0.067857	-0.029294	0.007418
review_scores_cleanliness	**0.772957**	-0.006508	0.030150	-0.032900	0.001988	0.017202
review_scores_checkin	**0.686948**	0.017649	0.111269	-0.080890	0.101978	-0.069910
review_scores_communication	**0.786605**	0.023701	0.121381	-0.108199	0.067272	-0.057394
review_scores_location	**0.658771**	0.032043	-0.124470	-0.066315	-0.034254	0.032544
review_scores_value	**0.798870**	-0.038780	0.062120	-0.128807	0.080772	0.017887
reviews_per_month	0.056986	-0.115745	**0.906731**	0.014075	0.022823	-0.086848
calculated_host_listings_count	-0.196797	0.207940		0.127849	**-0.490102**	0.177434

Renamed 6 Factors:

1. *Ratings*
2. *Number of Reviews*
3. *Instant Availability*
4. Room
5. Minimum Nights
6. Maximum Nights

Also, Multivariate data shows that these factors are not correlated.

Multivariate

Correlations

	Ratings	Number of Reviews	Instant Availability	Room Amenities	Minimum Nights	Maximum Nights
Ratings	1.0000	0.0027	-0.0298	-0.0098	0.1818	-0.0240
Number of Reviews	0.0027	1.0000	-0.0209	0.0333	0.1780	-0.0142
Instant Availability	-0.0298	-0.0209	1.0000	0.0224	-0.2551	-0.0148
Room Amenities	-0.0098	0.0333	0.0224	1.0000	-0.0396	-0.0071
Minimum Nights	0.1818	0.1780	-0.2551	-0.0396	1.0000	-0.1673
Maximum Nights	-0.0240	-0.0142	-0.0148	-0.0071	-0.1673	1.0000

c) Run a linear regression model on your y-variable using all the factors. What is the most important factor? If that factor was not labeled in Step b, give a name to that factor

From this linear regression, we can analyze that the most important factor is the Room, followed by Number of reviews, Ratings, Minimum nights and maximum nights.

Room amenities (# of bedrooms, beds, bathrooms, and accommodation), number of reviews and ratings do matter in predicting our y-variable, price per night.
When looking at the t-value, all values have probability of less than 0.05, except instant availability. In other words, these are all statistically significant with our y-variable, price per night. (Instant availability and maximum nights are not statistically significant in impacting the price per night.

Response price

Effect Summary

Source	Logworth	PValue
Room Amenities	95.802	0.00000
Number of Reviews	28.633	0.00000
Ratings	25.656	0.00000
Minimum Nights	4.196	0.00006
Maximum Nights	1.788	0.01630
Instant Availability	0.083	0.82632

Remove Add Edit FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	6184	4426204779	715751	.
Pure Error	3	0	0	Prob > F
Total Error	6187	4426204779		.
				Max RSq
				1.0000

Summary of Fit

RSquare	0.103672
RSquare Adj	0.102802
Root Mean Square Error	845.8156
Mean of Response	531.1723
Observations (or Sum Wgts)	6194

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	511945890	85324315	119.2673
Error	6187	4426204779	715404.04	Prob > F
C. Total	6193	4938150669		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	542.90167	11.0715	49.04	<.0001*
Ratings	136.53365	12.78748	10.68	<.0001*
Number of Reviews	-135.4464	11.97808	-11.31	<.0001*
Instant Availability	-2.658862	12.11686	-0.22	0.8263
Room Amenities	293.30376	13.82148	21.22	<.0001*
Minimum Nights	-89.46983	22.35814	-4.00	<.0001*
Maximum Nights	26.031577	10.83433	2.40	0.0163*

Effect Tests

Effect Details

Clustering:

a) Run a Cluster Analysis using the factors from PCA/Factor Analysis AND other variables that you would like to add (variables that were not used when running PCA/Factor analysis). How many clusters are there? (Feel free to visualize the clusters using a Constellation Plot)

6 Factors + Room type + Neighborhood_cleansed + host_is_superhost

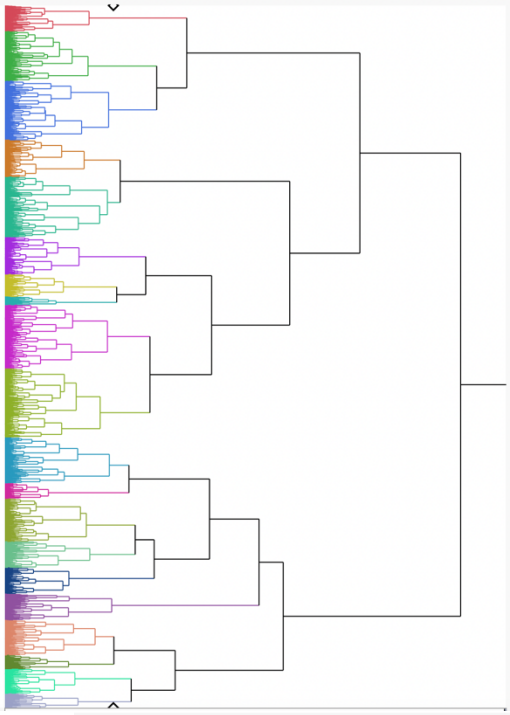
There are 20 clusters

Since we have many listings for each cluster, it created a condensed constellation plot.

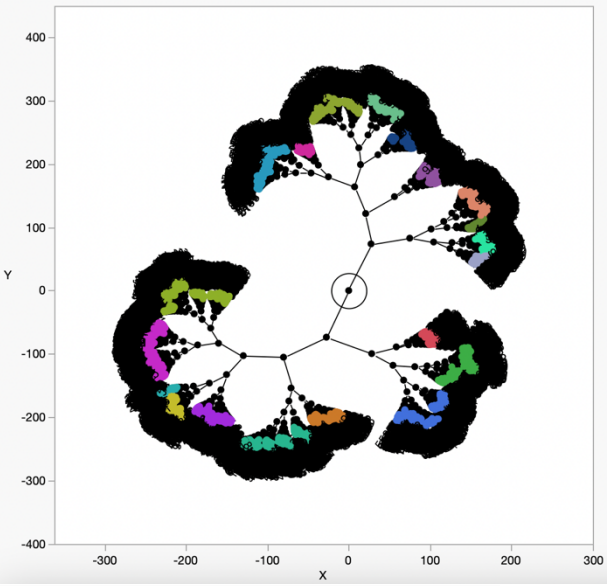
Hierarchical Clustering

Method = Ward, Standardized by Column

Dendrogram

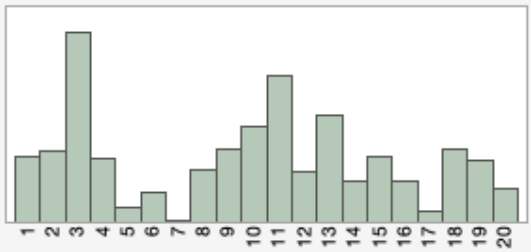


Constellation Plot



Distributions

Cluster



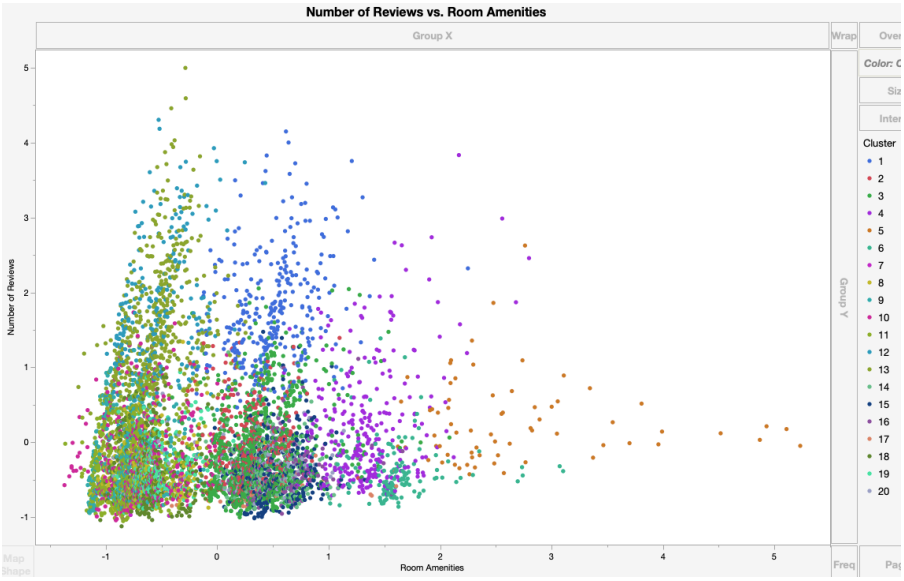
Frequencies

Level	Count	Prob
1	313	0.05053
2	341	0.05505
3	926	0.14950
4	303	0.04892
5	68	0.01098
6	140	0.02260
7	4	0.00065
8	249	0.04020
9	355	0.05731
10	463	0.07475
11	711	0.11479
12	241	0.03891
13	523	0.08444
14	190	0.03067
15	318	0.05134
16	196	0.03164
17	50	0.00807
18	349	0.05634
19	293	0.04730
20	161	0.02599
Total	6194	1.00000

N Missing 2931

20 Levels

b) Build a perceptual map using the two most important factors (the first two factors) and describe the patterns you detect.

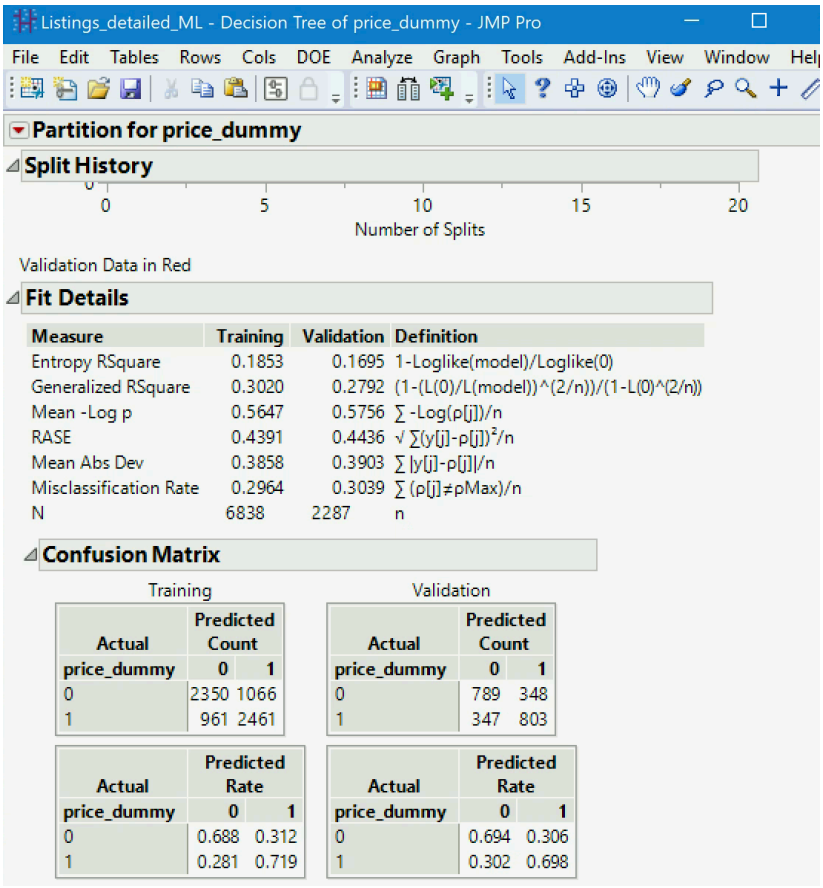


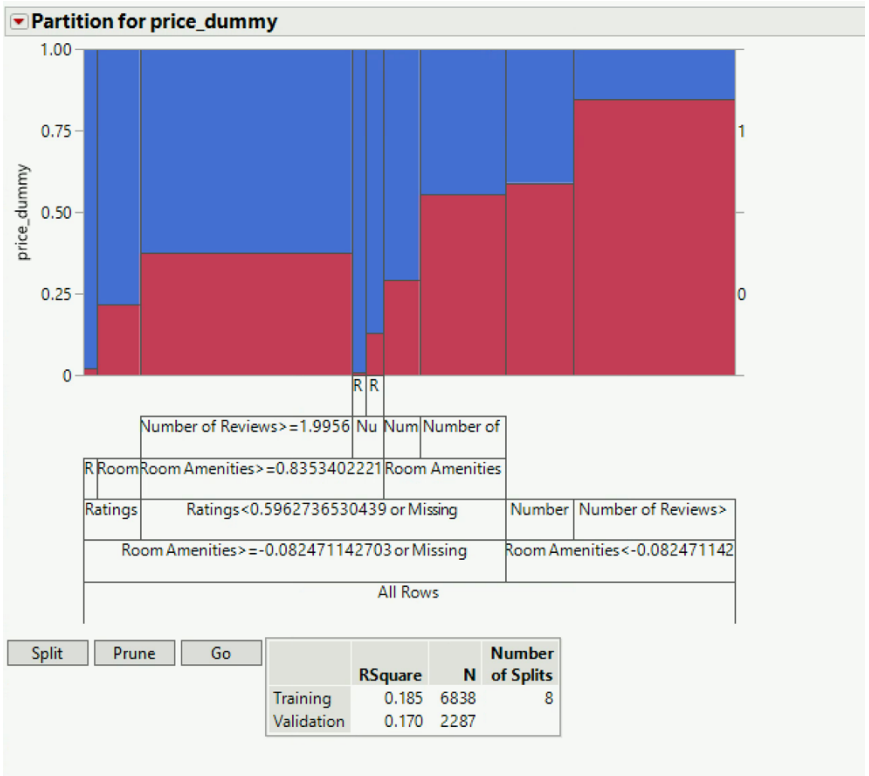
Each cluster (same colors) are in the same/similar area. Orange cluster tends to have higher number in room (# of bathrooms, bedrooms, beds, and accommodation).
Light blue cluster seems to have higher number of reviews.

Supervised Learning:

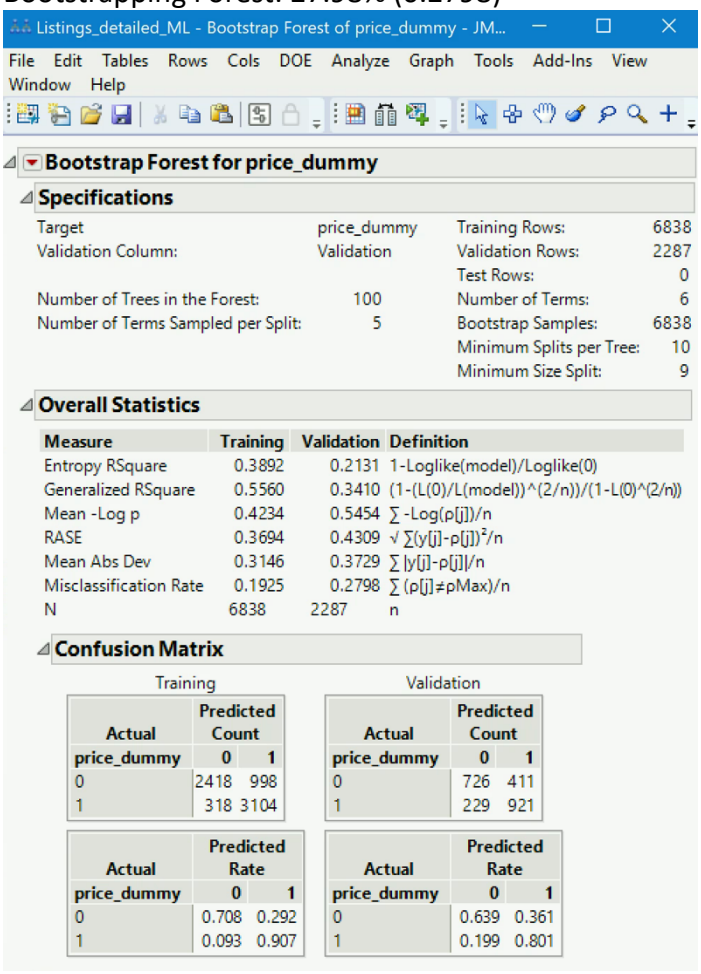
2 ML Models: Decision Tree, Bootstrapping Forest, and Neural Network
Y variable is price, 0 = below $387 and 1 = above $387

Decision Tree: *8 Splits*

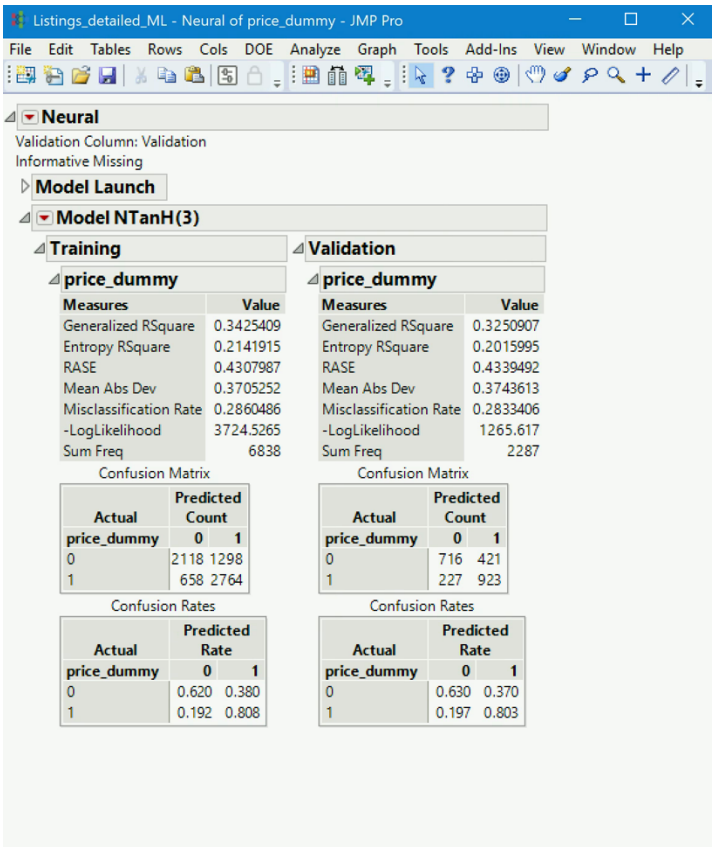

Listings_detailed_ML - Decision Tree of price_dummy - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help
Partition for price_dummy
Split History
0 5 10 15 20
Number of Splits
Validation Data in Red
Fit Details
Measure Training Validation Definition
Entropy RSquare 0.1853 0.1695 1-Loglike(model)/Loglike(0)
Generalized RSquare 0.3020 0.2792 (1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p 0.5647 0.5756 ∑ -Log(ρ[j])/n
RASE 0.4391 0.4436 √ ∑(y[j]-ρ[j])²/n
Mean Abs Dev 0.3858 0.3903 ∑ |y[j]-ρ[j]|/n
Misclassification Rate 0.2964 0.3039 ∑ (ρ[j]≠ρMax)/n
N 6838 2287 n
Confusion Matrix
Training
Actual Predicted Count
price_dummy 0 1
0 2350 1066
1 961 2461
Validation
Actual Predicted Count
price_dummy 0 1
0 789 348
1 347 803
Actual Predicted Rate
price_dummy 0 1
0 0.688 0.312
1 0.281 0.719
Actual Predicted Rate
price_dummy 0 1
0 0.694 0.306
1 0.302 0.698


Partition for price_dummy
1.00
0.75
0.50
0.25
0
price_dummy
1
0
R R
Number of Reviews>=1.9956 Nu Num Number of
R Room Room Amenities>=0.8353402221 Room Amenities
Ratings Ratings<0.5962736530439 or Missing Number Number of Reviews>
Room Amenities>=-0.082471142703 or Missing Room Amenities<-0.082471142
All Rows
Split Prune Go
RSquare N Number of Splits
Training 0.185 6838 8
Validation 0.170 2287

Bootstrapping Forest: 27.98% (0.2798)

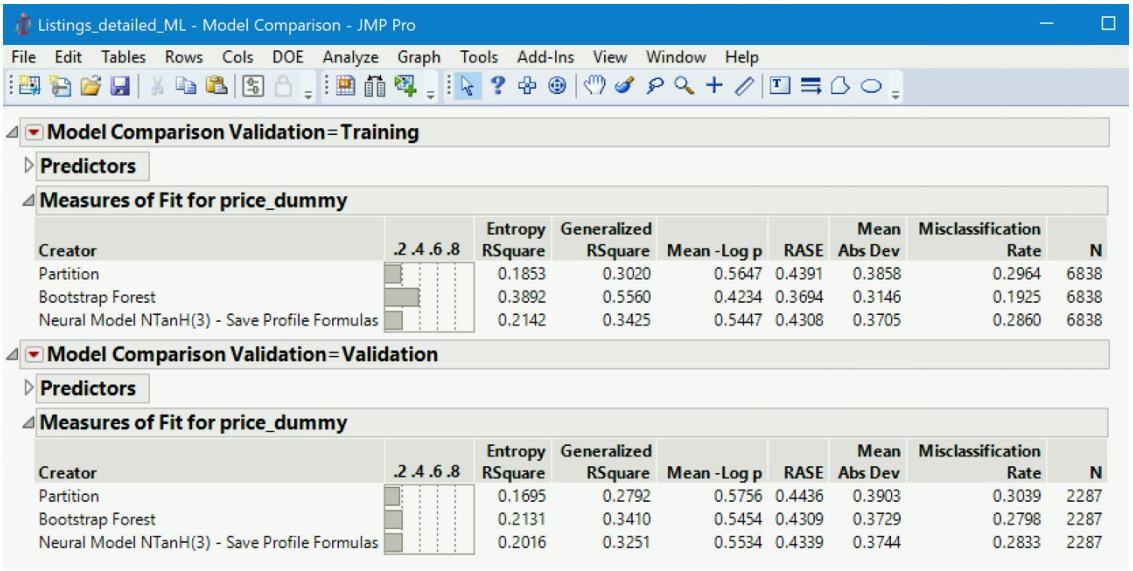


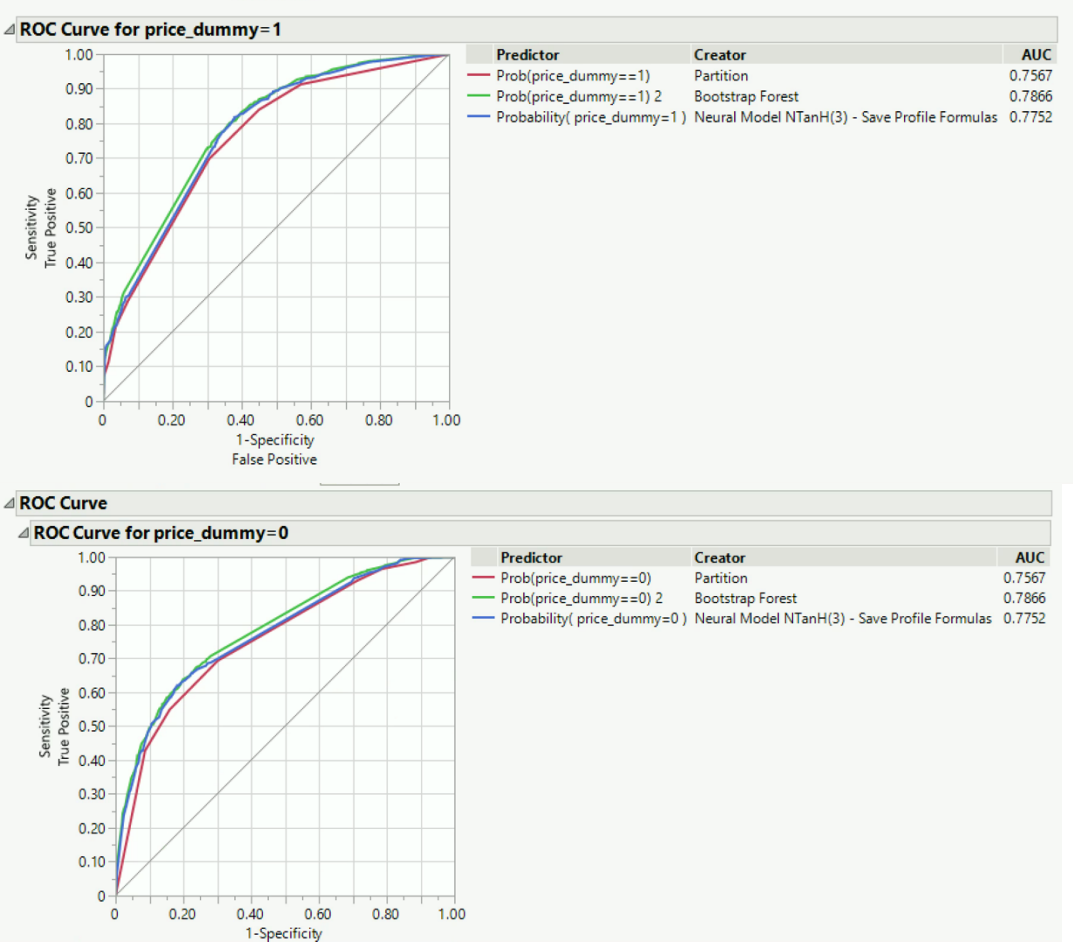
Neural Network: 28.33% (0.2833)



Model Performance Comparison:

The lowest Misclassification is *27.98%* and the highest AUC % is at *78.66%*: The best model is ***Bootstrap Forest***





Text Mining: on (r and jmp)

Word Count: As shown below, most of the reviews consist of terms and phrases, such as “have stay, great, place, locat, beach, clean, love, etc. (JMP had several unrecognizable letters (foreign letters), but we were able to generate better word cloud with R using cleaned up data).

Text Explorer for Comment(Review)

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
39795	77003	3460307	44.9373	66946	0.8694

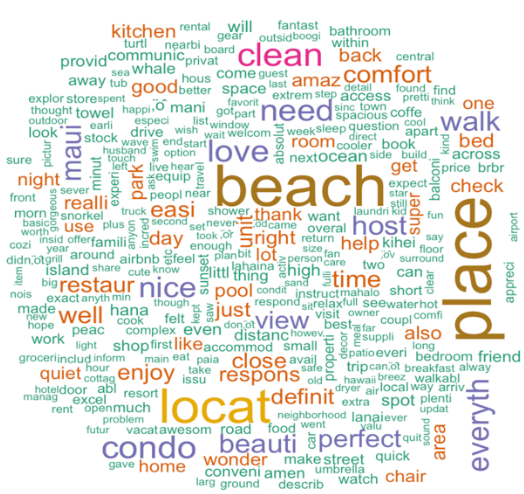
Term and Phrase Lists

Term	Count
stay·	43830
great·	38289
place·	37283
br·	35967
locat·	25506
beach·	25416
clean·	16503
love·	15593
host·	14573
need·	14424
everything	13688
nice·	13665
beauti·	13220
recommend·	12846
view·	12799
äôs·	12067
walk·	11613
us·	11306
comfort·	10942
perfect·	10873
well·	10000
time·	9945
enjoy·	9573
condo·	9571
close·	9104

Phrase	Count	N
br br	10420	2
great location	6188	2
great place	4309	2
highly recommend	4145	2
place to stay	4108	3
äôs place	3492	2
walking distance	3353	2
everything we needed	2962	3
great stay	2819	2
definitely stay	2799	2
come back	2353	2
enjoyed our stay	2287	3
beach chairs	2255	2
didn äôt	2081	2
great host	1939	2
great place to stay	1698	4
really enjoyed	1608	2
definitely recommend	1594	2
well stocked	1536	2
everything you need	1437	3
beach towels	1394	2
hot tub	1353	2
loved our stay	1328	3
big island	1315	2
within walking	1184	2

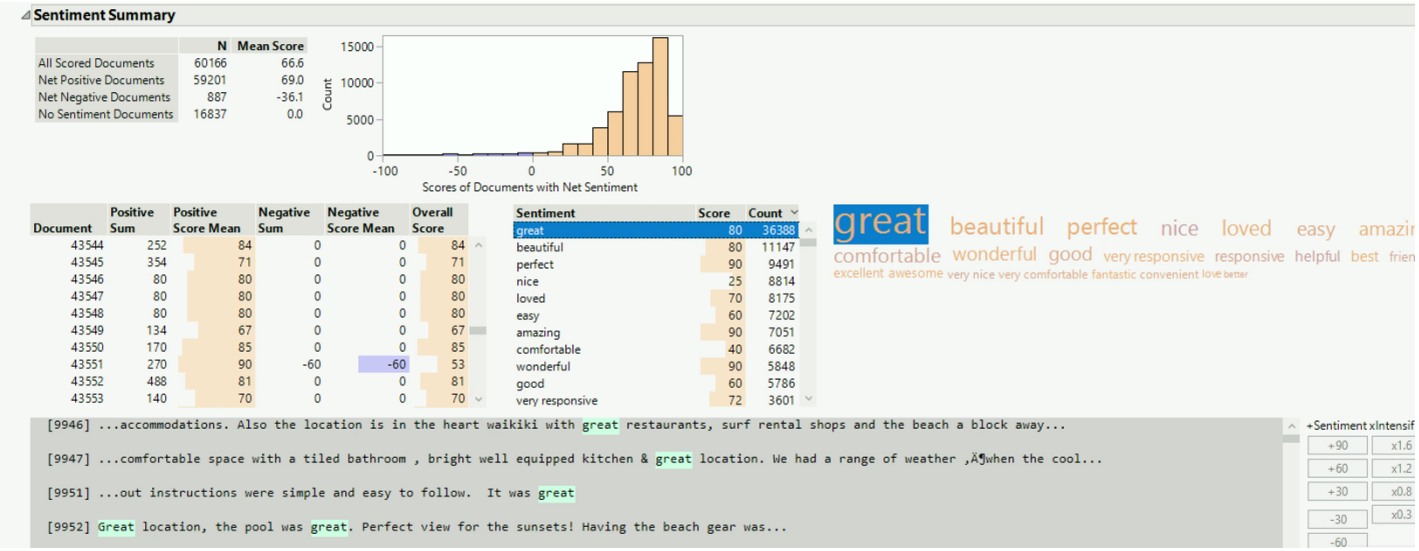
Word Cloud

<u>R- with cleaned data</u>



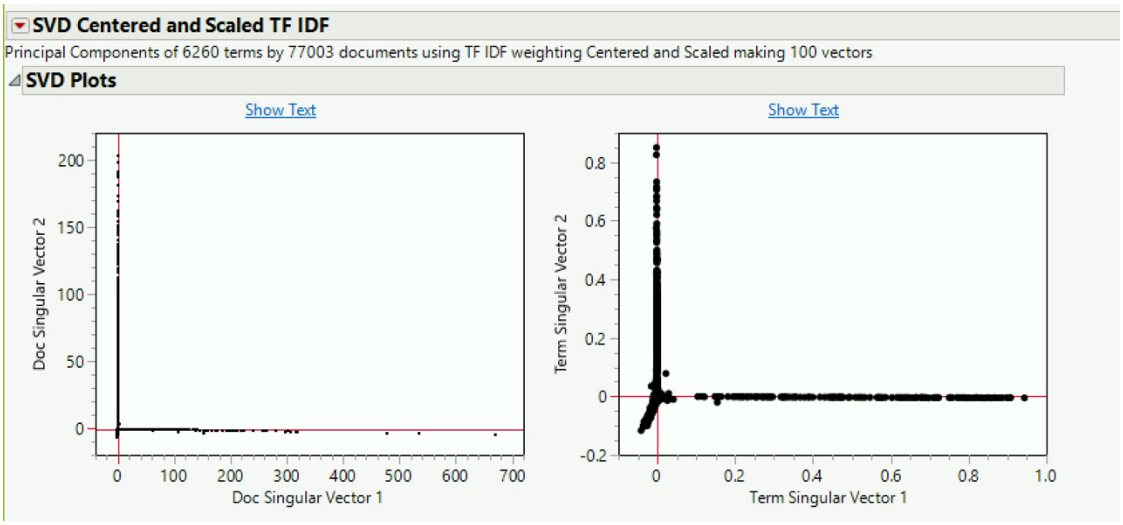
Sentiment Analysis

As we explained previously, these terms and phrases are very positive, it is proven by the sentiment analysis that almost 70% are positive, including the word, “great, beautiful, perfect, loved, easy, amazing, comfortable, wonderful, etc.”



Topic Modeling

With the topic modeling, Topic 1 to 5 has many encoded symbols where the system cannot translate into English. This means that the reviews are left with other foreign languages. (Topic 2 contains German and Topic 3 contains French, Topic 5 and 9 contains Spanish or Portuguese). Topic 6 contains amenities, and Topic 8 contains things like location and amenities and nearby activities).



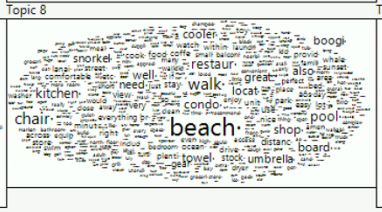
Top Loadings by Topic

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Term	Loading	Term	Loading	Term	Loading	Term	Loading	Term	Loading
åæ·	0.94361	und·	0.86780	et·	0.85140	µ·	0.99052	y·	0.81039
é°·	0.90982	die·	0.84462	le·	0.76915	ª·	0.98894	el·	0.80161
åß·	0.90733	der·	0.74897	est·	0.72907	å·	0.98840	muy·	0.70790
åü·	0.90537	ist·	0.73139	pour·	0.72051	æ·	0.98492	todo·	0.67102
åñ·	0.90040	das·	0.72653	√†·	0.71535	ω·	0.98100	lo·	0.64491
åó·	0.89744	ein·	0.72034	de·	0.71488	≤·	0.98095	una·	0.64312
å¶·	0.88591	war·	0.69604	nous·	0.68387	¥·	0.97814	para·	0.60221
åå·	0.87830	sehr·	0.69481	les·	0.67725	ç·	0.97698	las·	0.59856
å´·	0.87449	wir·	0.68151	tr√®s·	0.67159	∏·	0.97510	en·	0.58730
çå·	0.87316	mit·	0.65893	pas·	0.64631	è·	0.97309	que·	0.57554
äç·	0.86564	zu·	0.65506	il·	0.60831	ä·	0.97292	la·	0.57048
é≥·	0.85628	nicht·	0.65389	des·	0.60202	∞·	0.97223	tien·	0.56610
åă·	0.85070	f√°r·	0.63303	tout·	0.59736	∫·	0.94761	playa·	0.54976
åç·	0.84624	auch·	0.60403	avon·	0.59070	á·	0.94656	lugar·	0.52945
çä·	0.83951	man·	0.59288	la·	0.58836	ø·	0.94528	pero·	0.51864
åø·	0.83923	auf·	0.57522	avec·	0.56940	°·	0.94065	los·	0.50595
åô·	0.83152					π·	0.93717		
éé·	0.82804					∂·	0.93581		
						é·	0.92873		

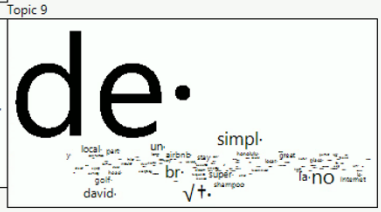
Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
Term	Loading	Term	Loading	Term	Loading	Term	Loading	Term	Loading
not	0.53772	ïà¤·	0.61930	beach·	0.64610	o·	0.80424	us·	0.31623
no	0.40702	ïûàïñ¥ïñú·	0.50575	walk·	0.44119	√©·	0.78416	fruit·	0.30630
äôt·	0.35682	ïûàïàî·	0.50261	chair·	0.43263	uma·	0.75327	home·	0.28915
br·	0.33159	ïñàï¨¥·	0.48032	restaur·	0.37231	e·	0.74904	so	0.27440
only	0.33086	î¤ïî°ú·	0.47011	kitchen·	0.36897	muito·	0.74262	island·	0.27288
dirti·	0.32102	ïûàí≥†·	0.44988	pool·	0.34977	n√£o·	0.69338	stay·	0.27086
door·	0.31944	ïûò·	0.44599	towel·	0.34203	bem·	0.68972	tree·	0.26307
bathroom·	0.29513	ï™®îî†·	0.44222	condo·	0.32185	com·	0.68753	frog·	0.25603
one·	0.29179	ïûàïóàïãµîâàïâš·	0.42943	shop·	0.32133	foi·	0.67559	peac·	0.24899
issu·	0.28791	ïàóïüåîàï·	0.42748	well·	0.31586	mas·	0.65521	feel·	0.24709
if·	0.28539	ïûàïãµîâàïâš·	0.42060	board·	0.31524	localiza√ß√£o·	0.60706	beautiful	0.24512
room·	0.27292	ï¢ãïïöïãµîâàïâš·	0.42059	snorkel·	0.31130	√≥tima·	0.60190	bird·	0.23804
stain·	0.27238	ïª'íîßê·	0.41724	locat·	0.30866	poi·	0.60001	hous·	0.23602
so	0.27227	ïàóïïüå·	0.41698	umbrella·	0.29714	praia·	0.53940	fresh·	0.23307
shower·	0.26954	îßšïö∞·	0.41317	also·	0.28930			recommend·	0.22724
use·	0.26923	ï¢ãïùä·	0.40824					experi·	0.22597
		ïù¥·	0.39699					natur·	0.22194

Topic 7

Topic 8

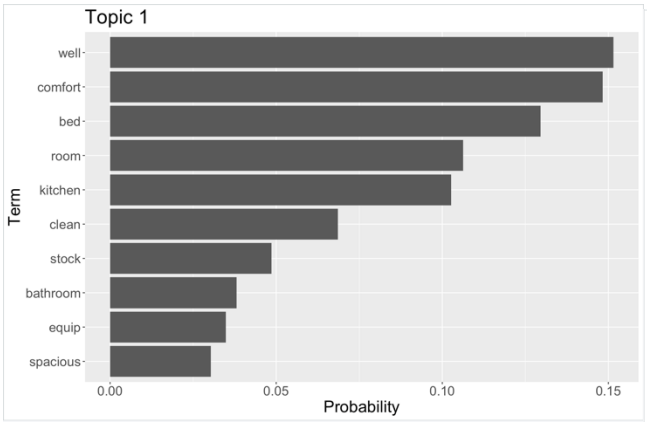


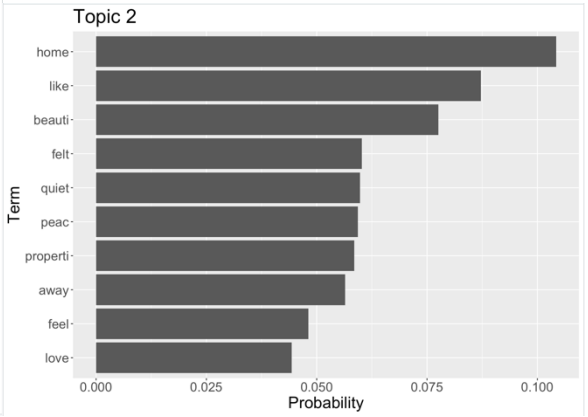
Topic 9

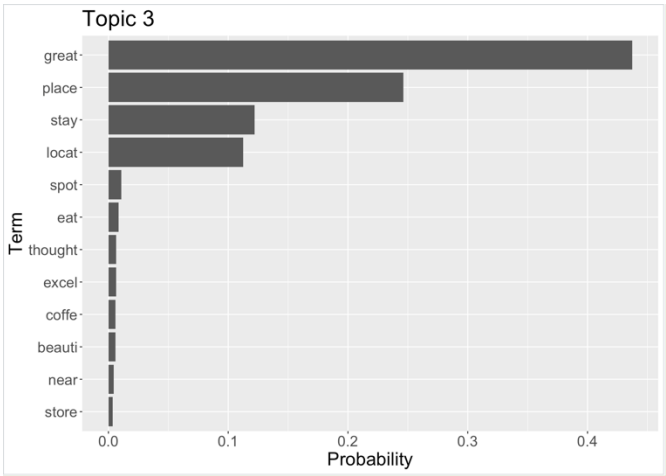


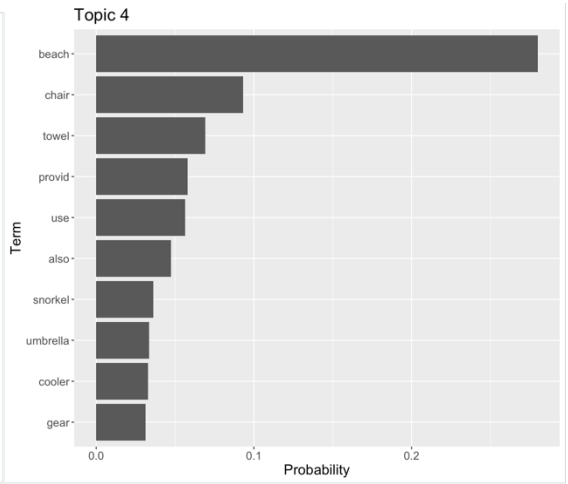
Topic 10

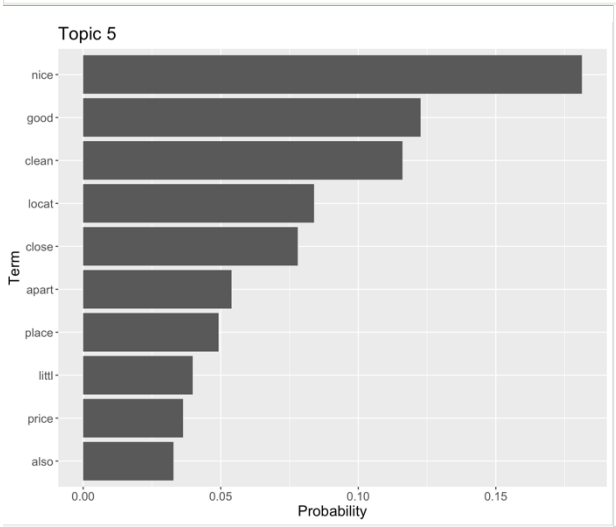
With cleaned data, we were able to generate better topic modeling with R

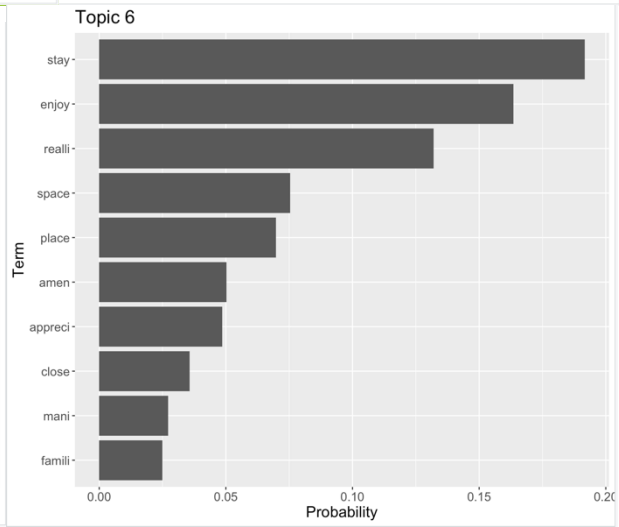


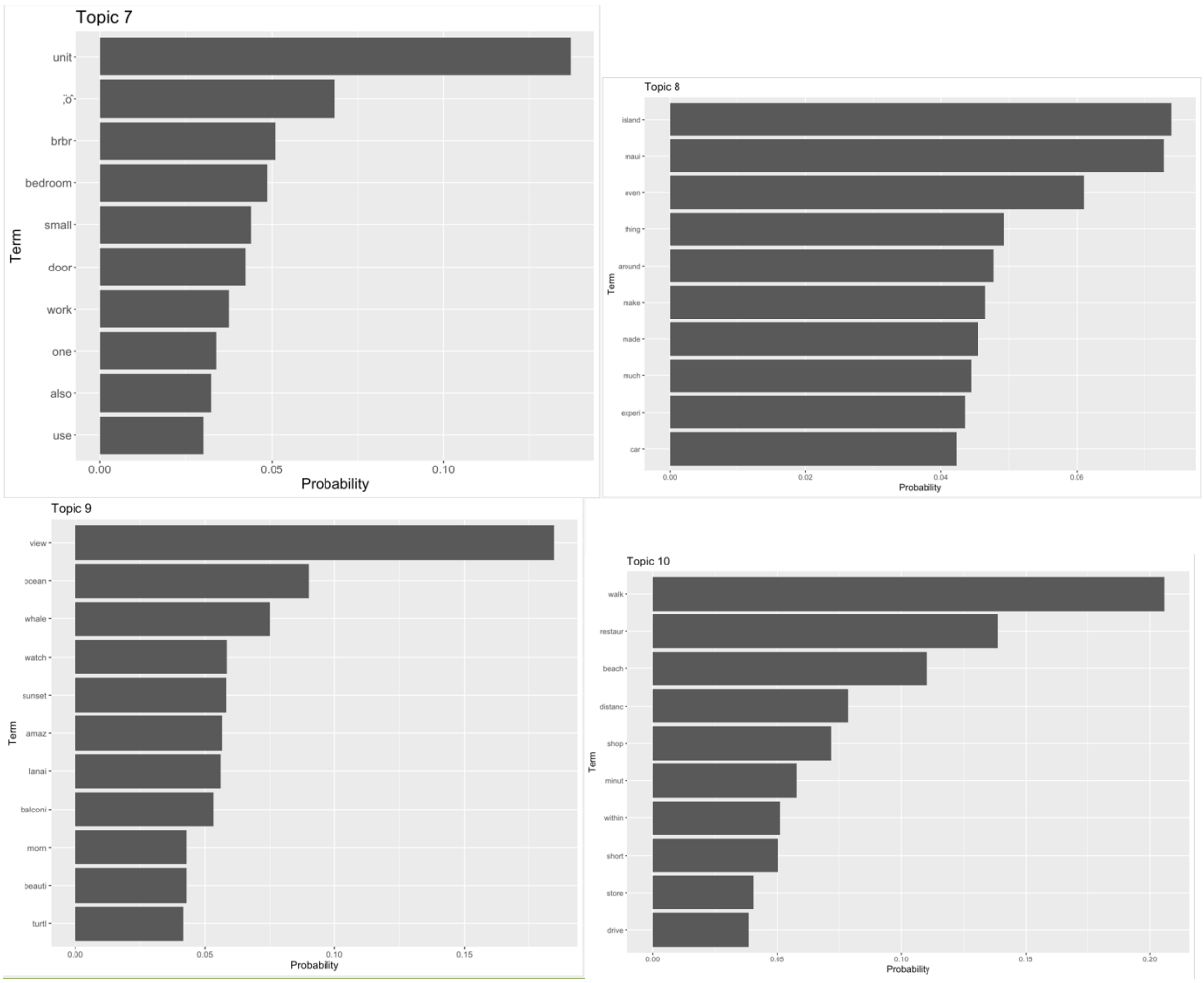








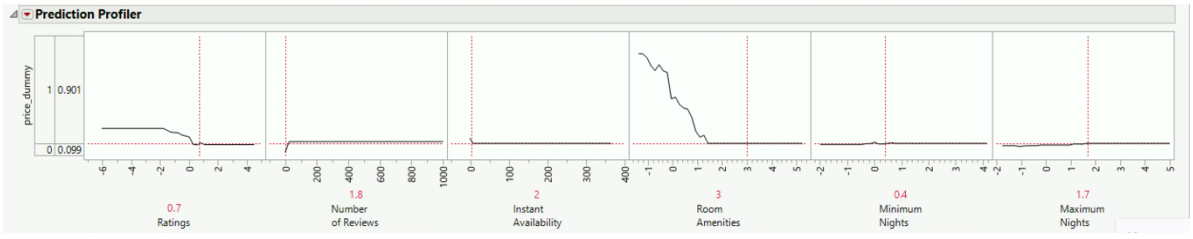




Price Prediction Profiler Comparision with Mid-term model

Bootstrap Forest Model – all 3 ML Scenarios
Price 0 = Below $387, Price 1 = Above $387

Scenario 1: More than 90% of our data is above $387 when we have a good rating, quite a number of reviews, have better room amenities (# of bathrooms, bedrooms, and beds), has many availabilities, and have higher minimum nights and higher maximum nights.



Scenario 2: More than 60% of data has listings below $387 when it has the lower ratings, number of reviews, less availability, room amenities (# of bathrooms, bedrooms, and beds), lower minimum nights, and lower maximum nights.