Group 3: James Min, Cam Nguyen, Kim Nguyen, Anh Nguyen, Eddie Pham

AirBnB Final Project: Austin, Texas

Quick Recap: Our original research question was to help the host predict the optimal price listing for their Airbnb. To add on to our research question, we want to help the host figure out the most important factors in receiving good ratings. As we previously found, ratings are important for predicting Price and, in general, are beneficial for drawing in customers.

Role: Our role for this project is to be in the perspectives of hosts and help them predict whether a guest would leave a 4.9-or higher star review or not. This is important for AirBnB hosts because if they are able to have a better idea on what factors/variables take into account in predicting whether the guest will leave a 5-star review, it will help hosts work on factors that are more important. By having a higher star rating, hosts will look much more attractive compared to other hosts.

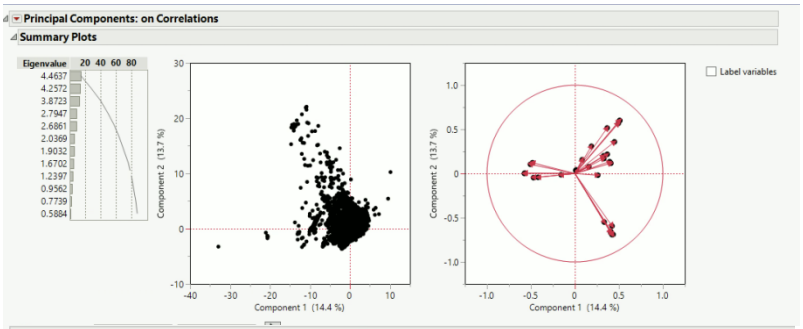
Research Question: What are some factors or components that guests consider important when leaving a 4.9 or higher -star review?

Phenomenon of Interest: We want to understand how to help hosts be more attractive by helping them gain more 4.9 or higher star reviews. This could be done by perfecting the factors that guests find to be most important when leaving a 4.9 or higher -star review and by using that, hosts are able to be better prepared.

Dataset: We will continue to use the same dataset and we have also used the reviews dataset

Unsupervised Learning:

PCA/Factor Analysis



Principal Components: on Correlations

Factor Analysis: Maximum Likelihood / Varimax

Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
host_listings_count	0.04983	-0.04742	0.00513	0.09434	0.98654	-0.05616	-0.09467	0.03088
accommodates	0.11593	-0.00020	0.00236	-0.05416	-0.03249	0.90661	0.01025	0.00859
bedrooms	0.07235	0.02834	-0.01918	-0.00555	-0.02835	0.92484	-0.02327	0.00987
beds	0.08864	-0.00583	-0.00374	-0.02764	-0.01784	0.84725	0.02161	0.01010
price	0.07039	0.02813	0.00232	-0.00426	0.00179	0.38040	-0.04767	-0.01808
minimum_nights	-0.00042	0.01352	-0.00638	0.96624	0.02651	-0.02812	-0.05809	-0.11465
maximum_nights	-0.06255	-0.02881	-0.01063	0.06296	0.14691	0.01228	0.02737	0.01254
minimum_minimum_nights	-0.01080	0.01390	-0.00500	0.90147	0.14823	-0.03353	-0.05564	-0.03002
maximum_minimum_nights	0.04472	0.02530	-0.00143	0.69396	0.56848	-0.05125	-0.05269	0.42009
minimum_maximum_nights	-0.00759	-0.00422	0.97390	-0.00363	-0.00781	-0.00287	-0.02049	-0.00059
maximum_maximum_nights	0.00367	-0.00667	0.90265	-0.00647	-0.00801	-0.00772	-0.01686	-0.00180
minimum_nights_avg_ntm	0.04144	0.02509	-0.00136	0.70878	0.56449	-0.05655	-0.05602	0.41280
maximum_nights_avg_ntm	-0.00772	-0.00215	1.01675	-0.00412	-0.00840	-0.00437	-0.02133	-0.00066
availability_30	0.96476	-0.03195	-0.01256	-0.01307	-0.12752	0.04766	-0.02604	-0.01260
availability_60	1.02705	-0.02738	-0.00545	-0.00008	-0.09768	0.03137	-0.00923	-0.00489
availability_90	1.00124	-0.01834	-0.00183	0.00644	-0.07700	0.03631	0.01630	0.00164
availability_365	0.69273	-0.02269	0.00742	0.04967	0.04580	0.13151	0.00891	0.02578
calendar_last_scraped	0.08888	-0.01895	-0.00760	0.01387	0.02548	0.02567	0.04495	0.02031
number_of_reviews	0.04785	0.03985	-0.00418	-0.01817	0.01641	-0.03228	0.98961	-0.04688
number_of_reviews_ltm	0.16691	0.05132	-0.00715	-0.08553	-0.03153	0.00573	0.67340	-0.01005
number_of_reviews_l30d	0.18764	0.06155	-0.01201	-0.09578	-0.05188	-0.02128	0.40033	0.00880
first_review	0.27036	-0.01964	0.02907	-0.09738	0.04003	0.05554	-0.42238	-0.03427
last_review	0.45262	0.06130	0.01673	-0.06427	0.04372	0.09732	0.15463	-0.06005
review_scores_accuracy	-0.00383	0.89982	0.00096	0.00032	-0.01019	0.01044	0.03085	-0.00473
review_scores_cleanliness	0.03025	0.79106	0.00917	-0.01239	0.02353	0.03202	0.03294	-0.02465
review_scores_checkin	-0.02330	0.78557	-0.01656	0.01560	-0.06720	0.04364	0.03637	0.02388
review_scores_communication	-0.03220	0.79961	-0.00941	0.01555	-0.08193	0.01514	0.02958	0.03156
review_scores_location	-0.00073	0.64674	0.00773	0.00890	0.00673	-0.01210	0.02370	-0.03348
review_scores_value	-0.04927	0.83794	-0.00687	-0.00060	-0.06750	-0.00572	0.04030	0.01389

Response review_scores_rating

Effect Summary

Source	LogWorth	PValue
Review Score	3907.560	0.00000
Accomodate	20.185	0.00000
Host Listing Count	9.512	0.00000
Number of reviews	8.190	0.00000
Availability	1.428	0.03729
Minimum Nights	0.536	0.29106
Average Min Max nights	0.406	0.39277
Maximum Nights	0.339	0.45782

Remove Add Edit ☐ FDR

Lack Of Fit

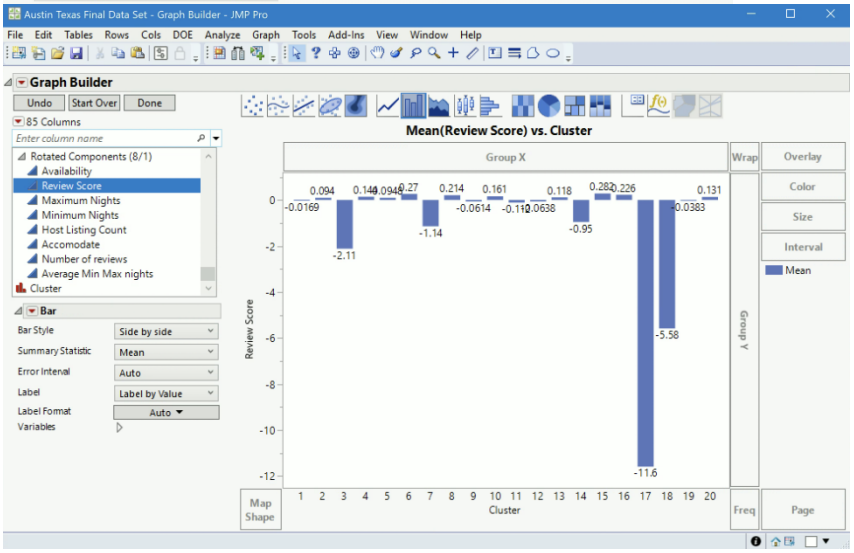
Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	10556	264.17171	0.025026	.
Pure Error	1	0.00000	0.000000	Prob > F
Total Error	10557	264.17171		.
				Max RSq
				1.0000

Summary of Fit

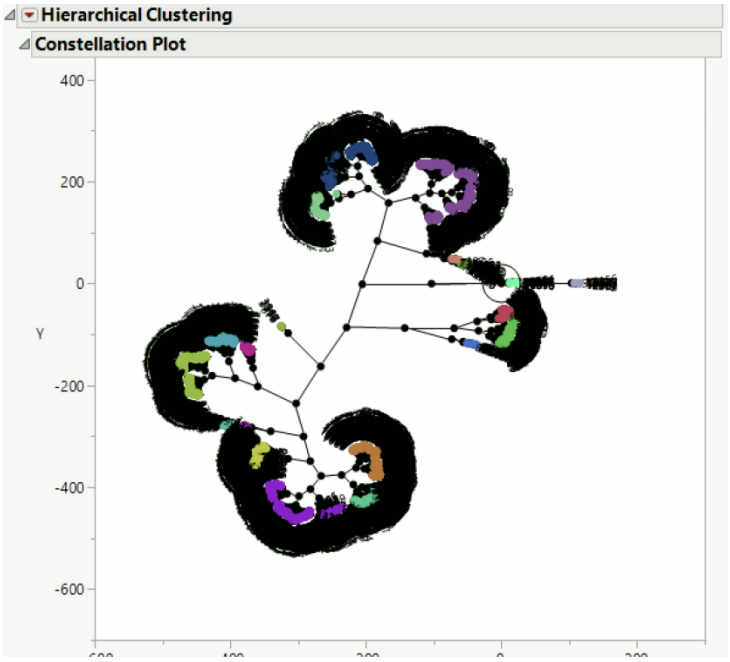
RSquare	0.820001
RSquare Adj	0.819864
Root Mean Square Error	0.158188
Mean of Response	4.816994
Observations (or Sum Wgts)	10566

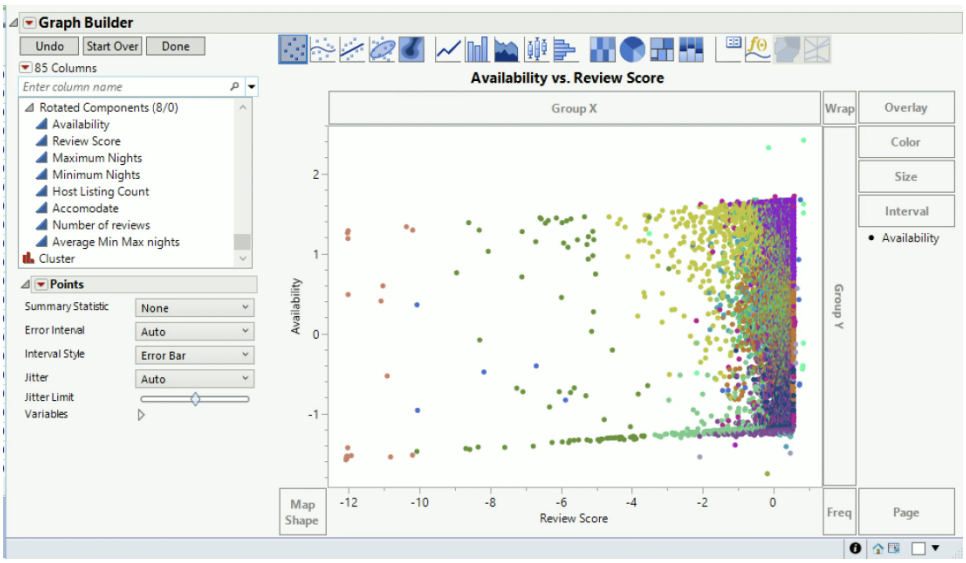
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.8127819	0.001637	2940.5	<.0001*
Availability	-0.003238	0.001555	-2.08	0.0373*
Review Score	0.360218	0.001654	217.81	<.0001*
Maximum Nights	-0.001146	0.001544	-0.74	0.4578
Minimum Nights	0.0016754	0.001587	1.06	0.2911
Host Listing Count	-0.022305	0.00354	-6.30	<.0001*
Accomodate	0.0148519	0.00158	9.40	<.0001*
Number of reviews	0.0088286	0.00152	5.81	<.0001*
Average Min Max nights	0.0030744	0.003597	0.85	0.3928



- There are 20 clusters and 8 factors.





- The pattern that we detected is unclear. It's concentrated from point 1 to –2 and scattered until –12. For recommendation, we should focus on increasing review score ratings and availability.

Logit Regression:

After running the Logit regression model, we were able to find the misclassification rate to be 40.57%. The reason why we believe the misclassification rate is so high is because when creating the dummy variable, we classified all 4.9-star ratings or higher to be =1 and anything else to be =0. Because the predictable range is so small (4.9+), the predictability goes down which then raises the misclassification rate. When looking at the profiler, we were able to find that when the price is $500, 4 bedrooms, 50 reviews and 4 accommodates, there is a 79.6% probability that a customer would leave a 4.9 or higher star rating. Something interesting we realized was that, when the price went down, the

probability also went down and we believe that is because price likely reflects quality and therefore, even if an Airbnb is able to accommodate the same number of people, the quality may differentiate how customers may leave reviews. Additionally, we found that when accommodation is higher than the number of bedrooms, the probability of leaving a 4.9 or higher rating goes down and we believe that may be because customers feel less value or space. We also found that increasing the number of bedrooms has a significant impact on the predictability of highly rated reviews. Interestingly, when price, bedrooms, and accommodation all go up, the probability of predicting a 4.9 star or higher rating goes up. This could be reflecting the higher quality which matches price. Lastly, the reason why it may have a high predictive ability is because of its high misclassification rate. When looking at the log worth, since accommodates seem to have the most worth, it would be best to focus on bedrooms. This could be by either increasing the number of bedrooms in comparison to accommodation or it could be by improving the quality, which may also impact the rating. When looking for methods of improvement, it may be best to look at customers/groups who have the highest likelihood of leaving a lower-star rating review such as the first group with 57% of them leaving a 4.9-star rating or lower. This is because by focusing on these groups, we are able to have a better understanding of how to improve the quality of their stay.

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	6060	4057.0011	8114.002
Saturated	6064	245.4907	Prob>ChiSq
Fitted	4	4302.4919	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.31261899	0.0533827	34.29	<.0001*
price	0.00081227	0.0001263	41.35	<.0001*
bedrooms	0.35226376	0.0451124	60.97	<.0001*
number_of_reviews	-0.0018189	0.0003352	29.45	<.0001*
accommodates	-0.1693189	0.0169919	99.29	<.0001*

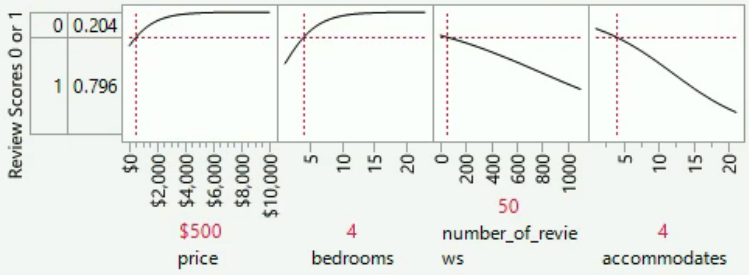
For log odds of 1/0

Covariance of Estimates

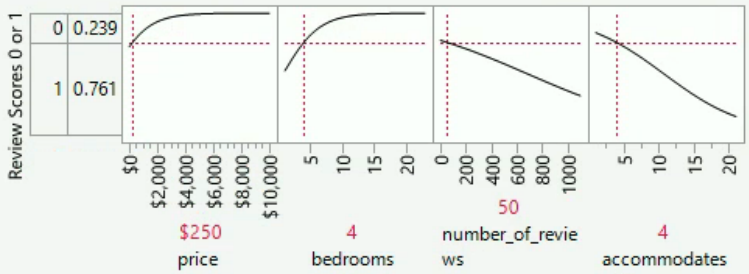
Effect Likelihood Ratio Tests

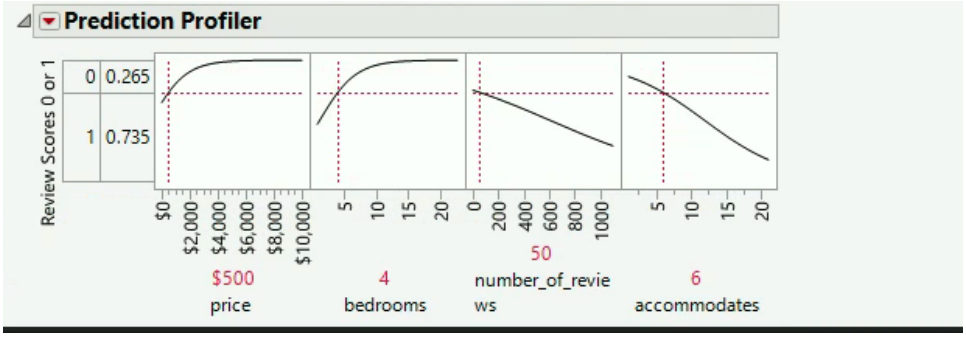
Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
price	1	1	53.37472	<.0001*
bedrooms	1	1	63.9535748	<.0001*
number_of_reviews	1	1	31.0308678	<.0001*
accommodates	1	1	105.151834	<.0001*

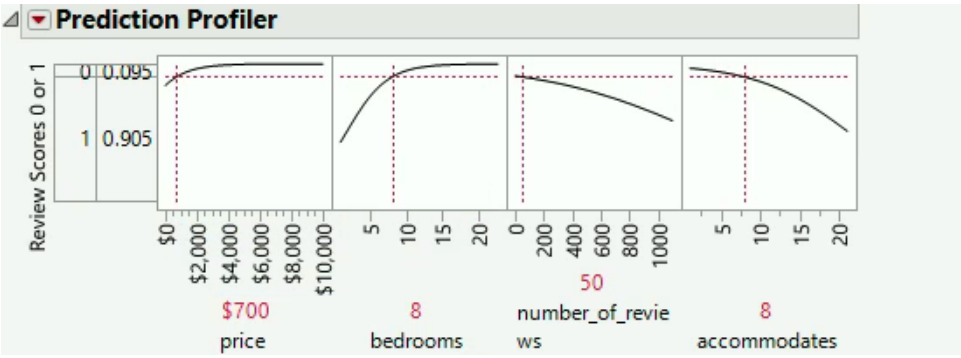
Prediction Profiler



Prediction Profiler



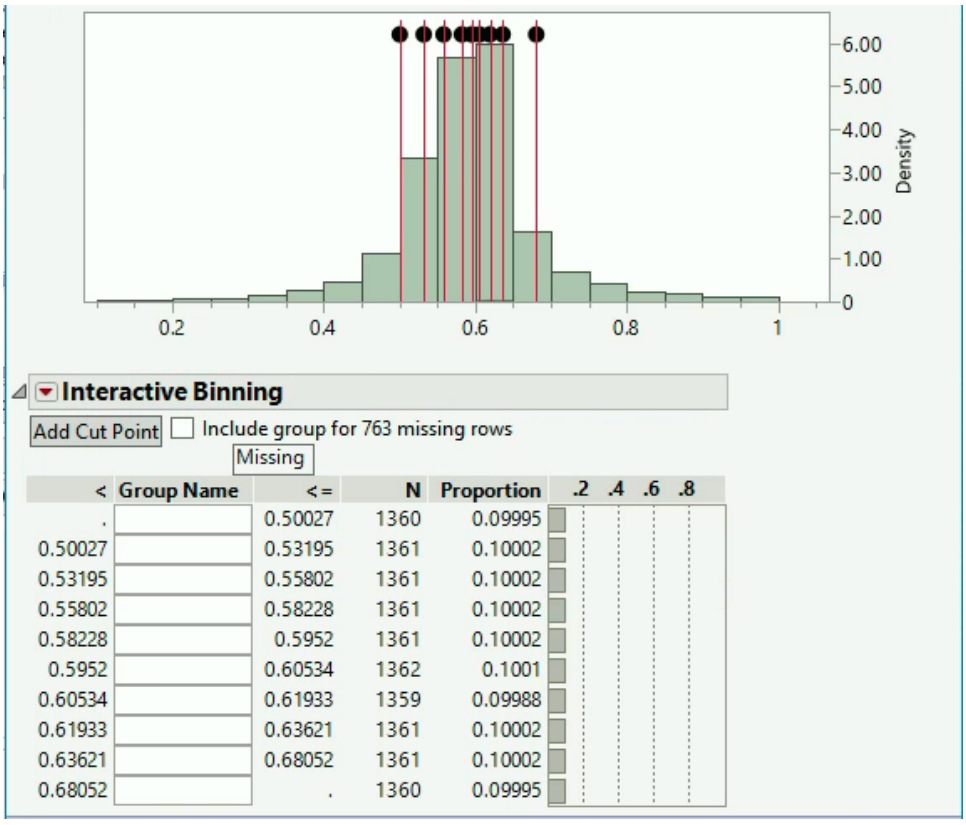




Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.0231	0.0187	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0416	0.0339	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6645	0.6696	∑ -Log(ρ[j])/n
RASE	0.4851	0.4869	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.4716	0.4727	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.4028	0.4057	∑ (ρ[j]≠ρMax)/n
N	6475	4240	n

Source	LogWorth	PValue
accommodates	23.946	0.00000
bedrooms	14.895	0.00000
price	12.560	0.00000
number_of_reviews	7.595	0.00000

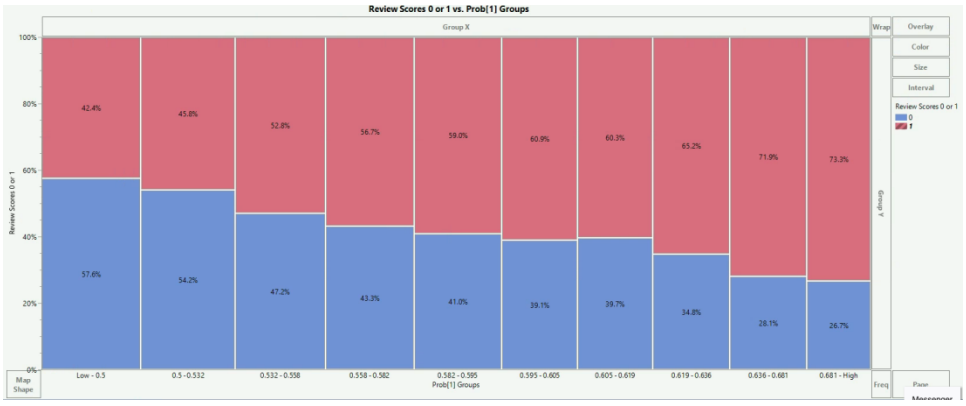


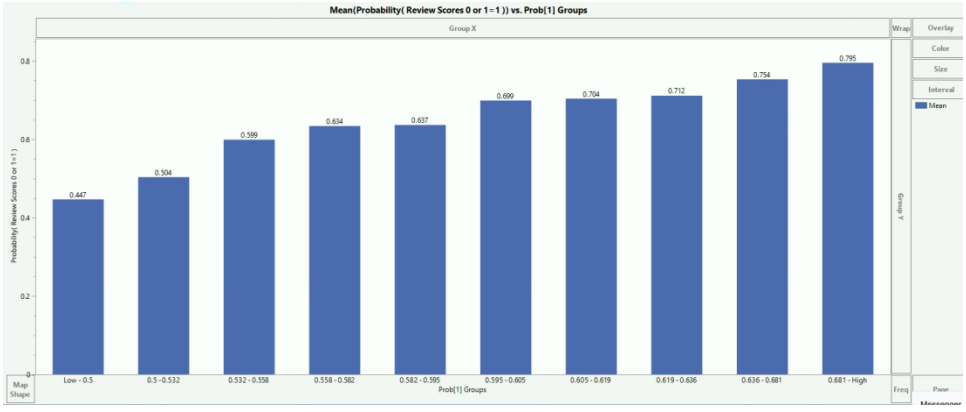
Interactive Binning

Add Cut Point ☐ Include group for 763 missing rows

Missing

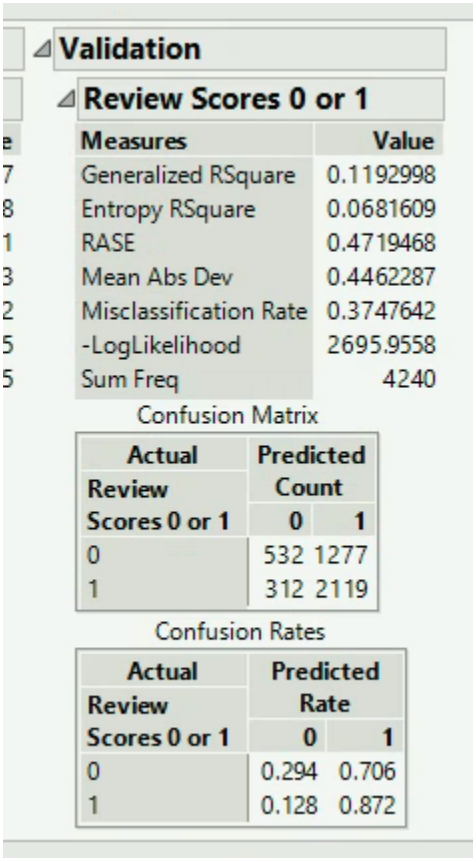
<	Group Name	<=	N	Proportion	.2 .4 .6 .8
.		0.50027	1360	0.09995	
0.50027		0.53195	1361	0.10002	
0.53195		0.55802	1361	0.10002	
0.55802		0.58228	1361	0.10002	
0.58228		0.5952	1361	0.10002	
0.5952		0.60534	1362	0.1001	
0.60534		0.61933	1359	0.09988	
0.61933		0.63621	1361	0.10002	
0.63621		0.68052	1361	0.10002	
0.68052		.	1360	0.09995	





Neural Model:

After running the neural model, we were able to find the misclassification rate to be 37.47%. Again, while this is not the best, since the likelihood of a customer leaving a review star rating of 4.9- or higher is low, it makes sense that it has a relatively high misclassification rate. If our goal was to predict customers who would leave a 4-star rating or higher, then our misclassification rate should go down. Additionally, when looking at the profiler, we were able to see that when using the same predictors, the neural model has a lower probability (63% vs 79.6%) of a customer leaving a 4.9 or higher rating but this could be because it has a lower misclassification rate, the model tries to not over predict its prediction. There was as 63% chance a customer would leave a 4.9 or higher star rating review if the price were $500, 4 bedrooms, 50 reviews, and 4 accommodations.



Validation

Review Scores 0 or 1

Measures	Value
Generalized RSquare	0.1192998
Entropy RSquare	0.0681609
RASE	0.4719468
Mean Abs Dev	0.4462287
Misclassification Rate	0.3747642
-LogLikelihood	2695.9558
Sum Freq	4240

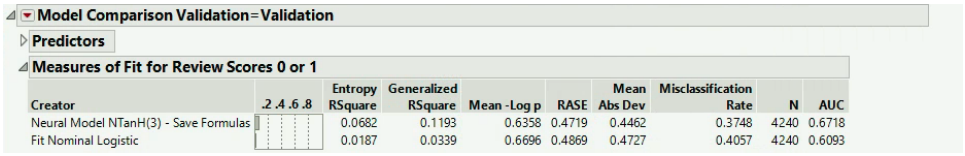
Confusion Matrix

Actual Review Scores 0 or 1	Predicted Count 0	1
0	532	1277
1	312	2119

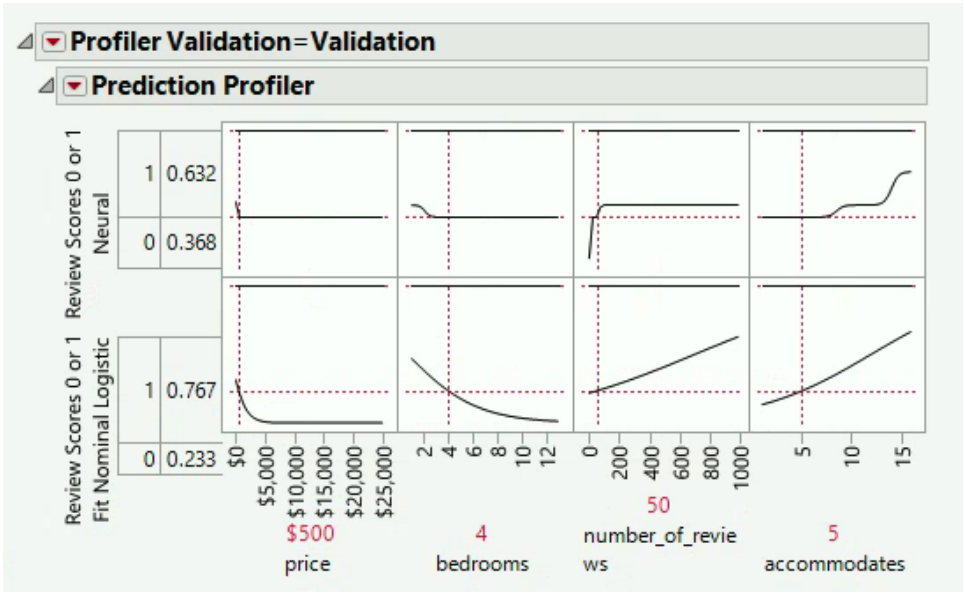
Confusion Rates

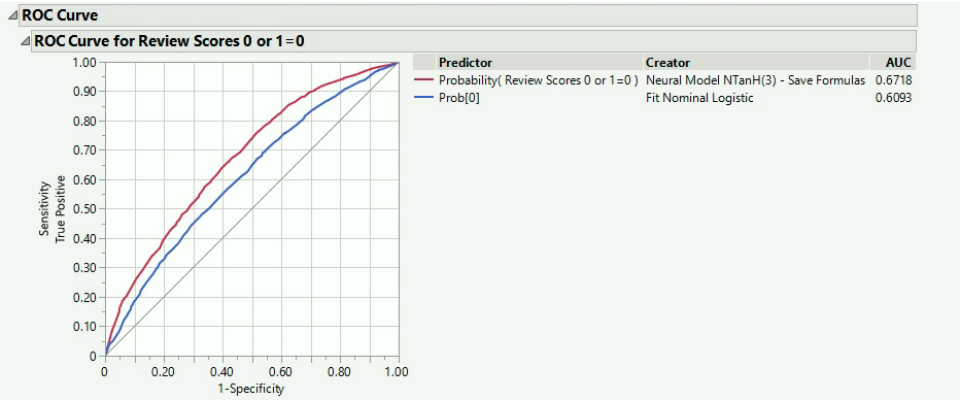
Actual Review Scores 0 or 1	Predicted Rate 0	1
0	0.294	0.706
1	0.128	0.872

Model Performance:



Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N	AUC
Neural Model NTanH(3) - Save Formulas		0.0682	0.1193	0.6358	0.4719	0.4462	0.3748	4240	0.6718
Fit Nominal Logistic		0.0187	0.0339	0.6696	0.4869	0.4727	0.4057	4240	0.6093





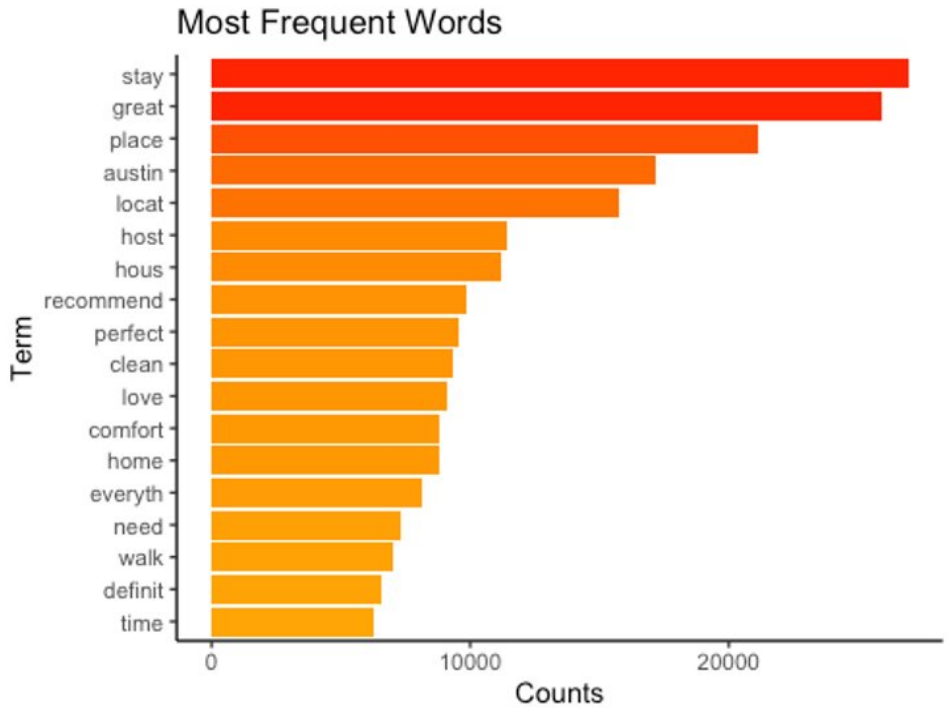
Model Comparison: The two ML models we used are Logit Regression and Neural Network. When comparing the two ML models, we were able to compare the misclassification rate and the AUC. In this scenario, neural network model can be seen superior to Logit Regression since it has a lower misclassification rate (lower the better). Furthermore, because the neural model has a higher AUC which means that it has a lower false positive rate while having a higher true positive rate. Since the neural model has a higher AUC, it can be seen superior. AUC measurement is critical since it takes into consideration the class imbalance problem when modeling. Since Neural is better in both categories, it can be seen as the better model.

Text Mining

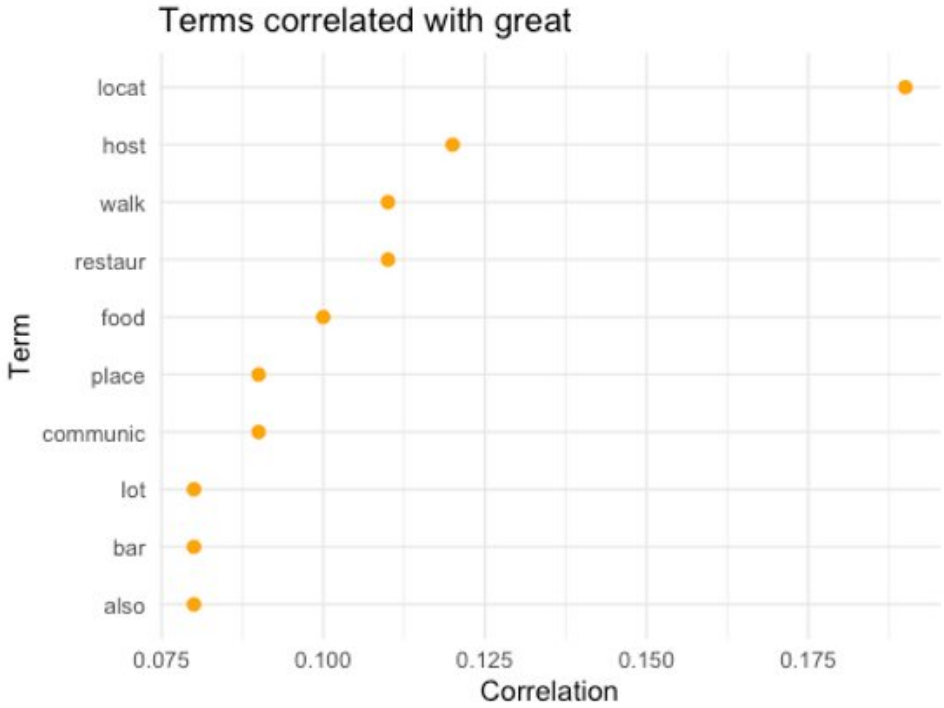
Word Frequency and Association

We used text mining in R to find the most frequent words in reviews for hosts that have a score of 4.9 or above. After finding the most frequent words, we found the words that are most associated with the most frequent words. By finding the most frequent words we can get insight into what most customers care about when they book an Airbnb. We focused on the first 5 most frequent words except for Austin. From the “Most Frequent Words” graph, the words suggest that customers are focused most on their

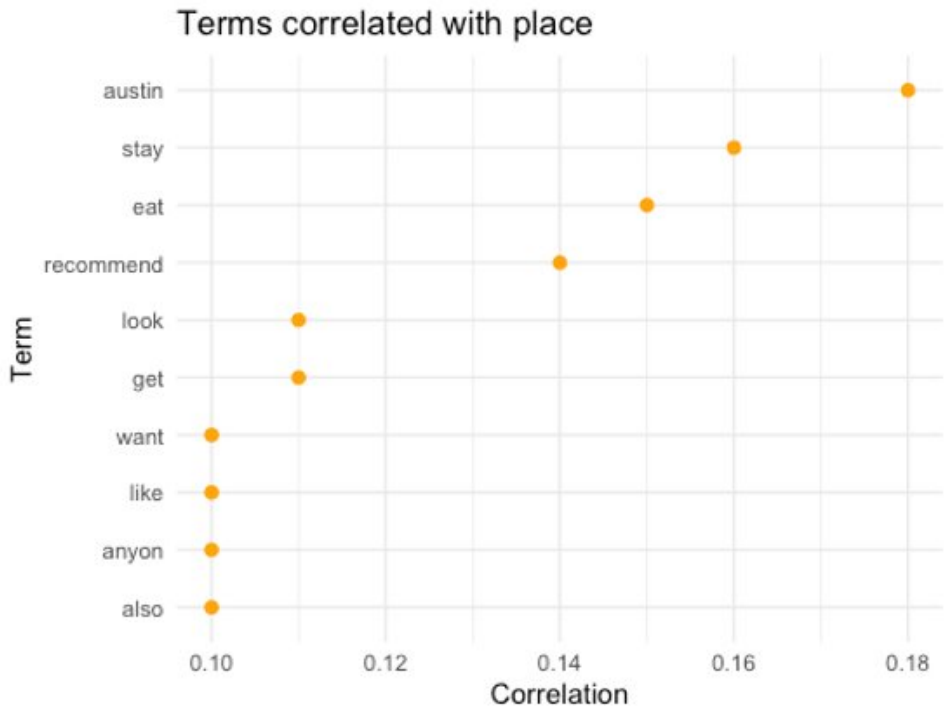
stay, the place, the location, the host, and the house. The associated words with each of these words give more specific insight. For example, the most correlated term with great is locat. We can infer that having the Airbnb in a good location will result in positive feedback. Some words most correlated with location are central, walk, and conveni. Therefore, places that are “great” seem to be in locations that are convenient for walking or travel and are central in the city.

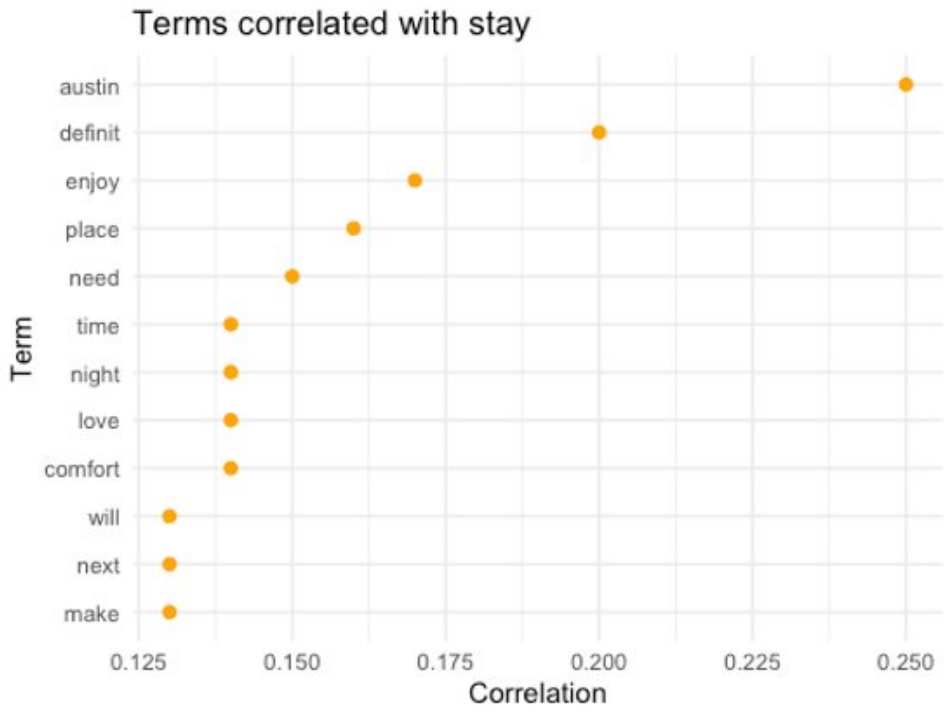


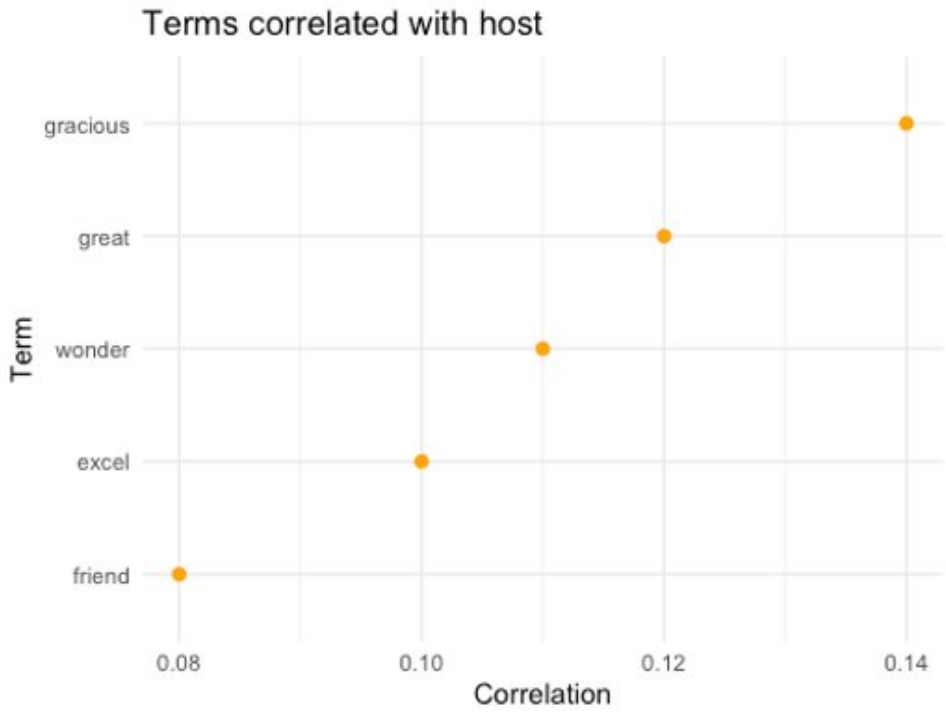

great
stay
place
host
locat
austin
perfect
love
recommend
hous
clean
comfort
space
downtown
home
neighborhood
everyth
definit
walk
nice
close
time
easi
realli
amaz
helpeven
bed
quiet
awesom
wonder
enjoy
also
excel
cottag
friend
back
high
super
well
need
beauti
trip
away
room
accommod
provid
get
book
conveni
come
respons
cozi
around
will
cute
thing
area
restaur
night
made
minut
thank
one
kitchen
decor
check
congress
much
thought
weekend
welcom
absolut
next
spot
south
feel
just
like
visit
mani
apart
fantast
want
good
quick
felt
make
look
fun
park
coffe
street
experi
communic
distanc
best
day
airbnb
arriv
littl

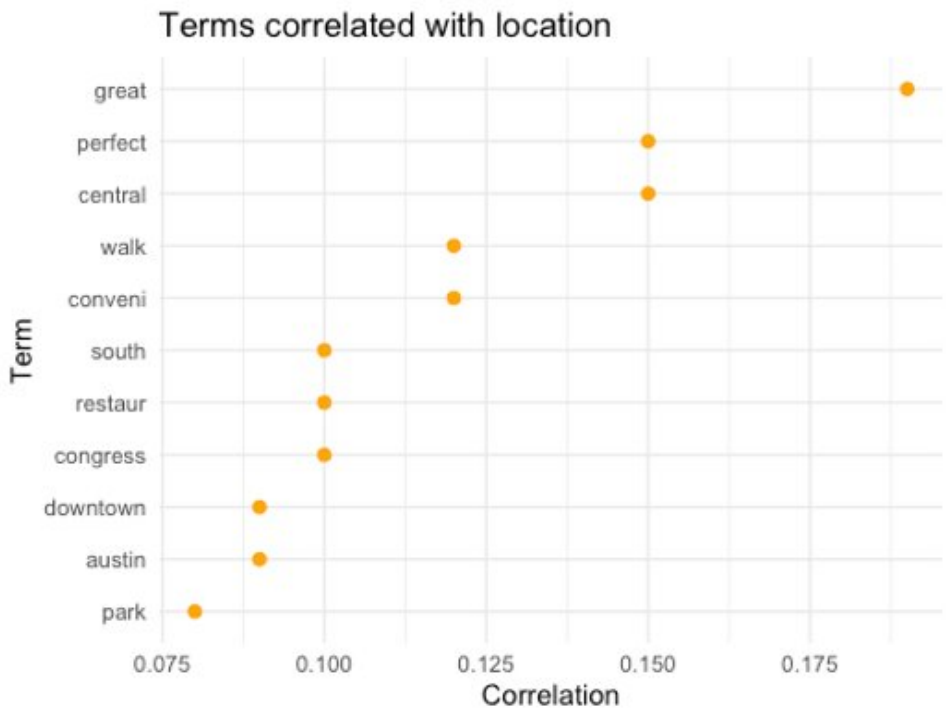

Terms correlated with great
locat
host
walk
restaur
food
place
communic
lot
bar
also
Term
0.075
0.100
0.125
0.150
0.175
Correlation

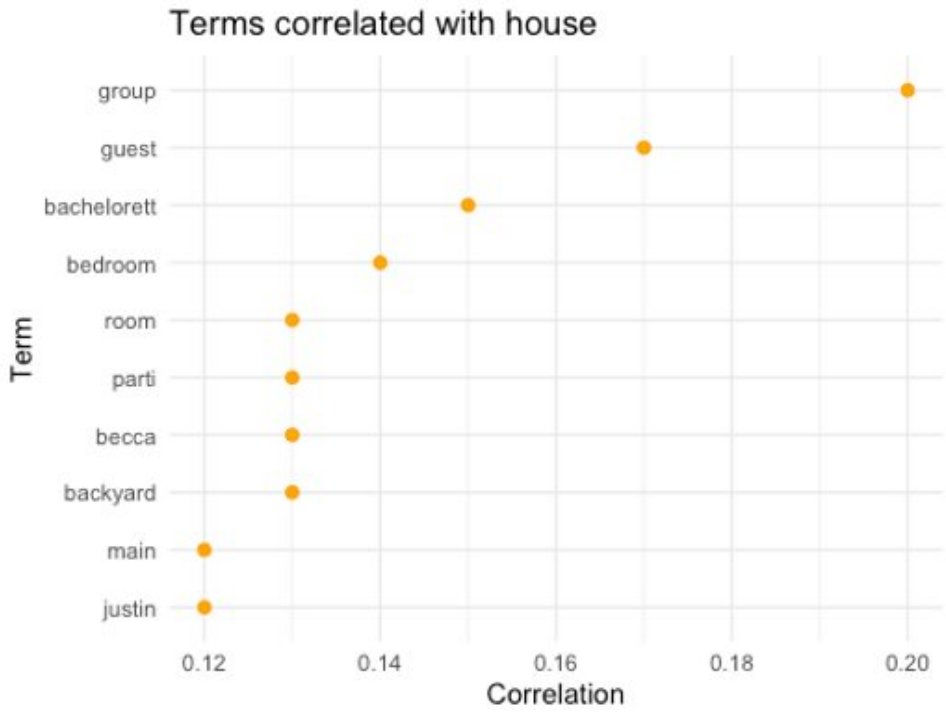
Terms correlated with great signify that people took note of the location, host, walkability, restaurants, and food. Highly rated Airbnbs seem to have a great location and host. The terms walk, restaur, and food could mean that the area is walkable and has good food and restaurants.




Terms correlated with stay
austin
definit
enjoy
place
need
time
night
love
comfort
will
next
make
Term
0.125
0.150
0.175
0.200
0.225
0.250
Correlation


Terms correlated with host
gracious
great
wonder
excel
friend
Term
0.08
0.10
0.12
0.14
Correlation

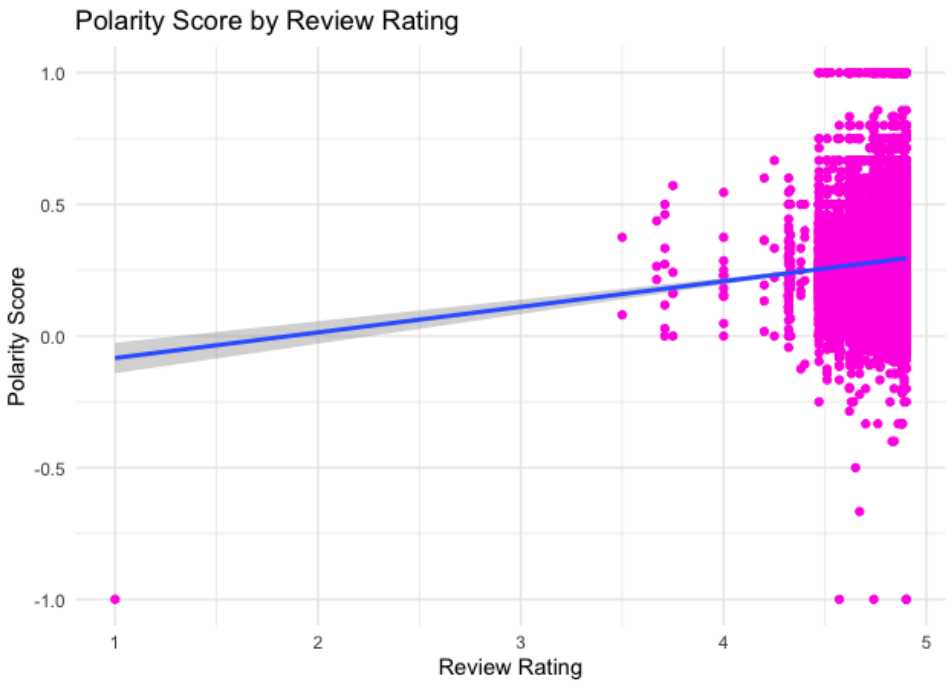


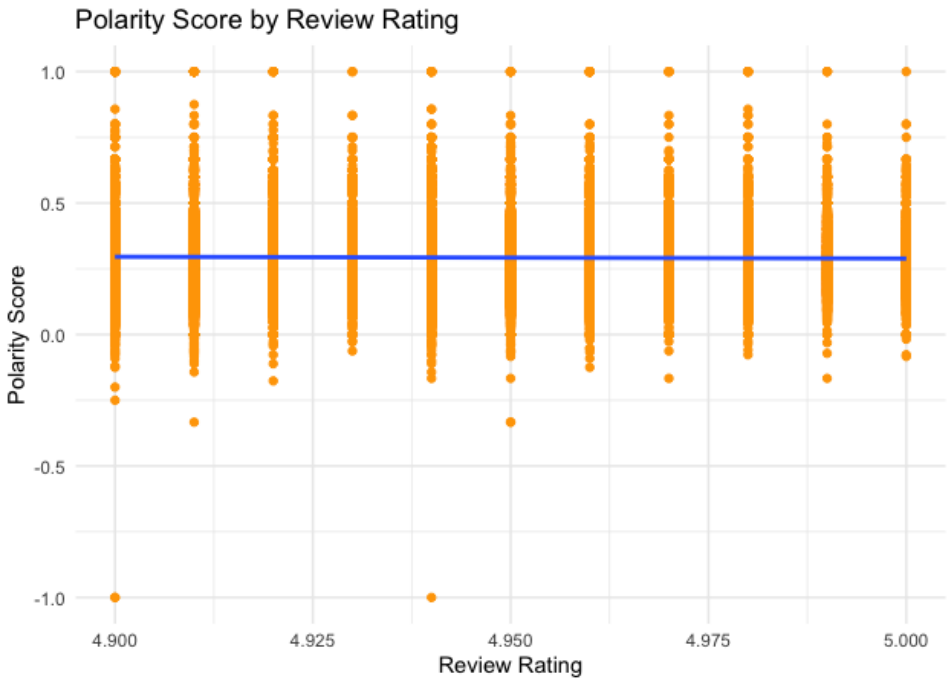


The info we gathered from word frequency and word association suggests that high reviews are given to listings that provide a great stay. A great stay encompasses the location, host, and house. A great location

is in a walkable area, has great restaurants, and is centrally located. A great host is most correlated with "gracious", so the interaction with the host is important. The term house is most correlated with "group" so it is important provide a great stay for groups of people. A recommendation would be to make a guest's stay as enjoyable as possible by providing the guest with an overview of great places to go during their stay and to help them with tips on getting around the city. Not only is this related to location, but this will also make a good impression as a host. The other important correlation is between house and group. Besides providing a clean and organized house, we recommend that the host provide amenities for groups or a welcome gift.

Sentiment Analysis





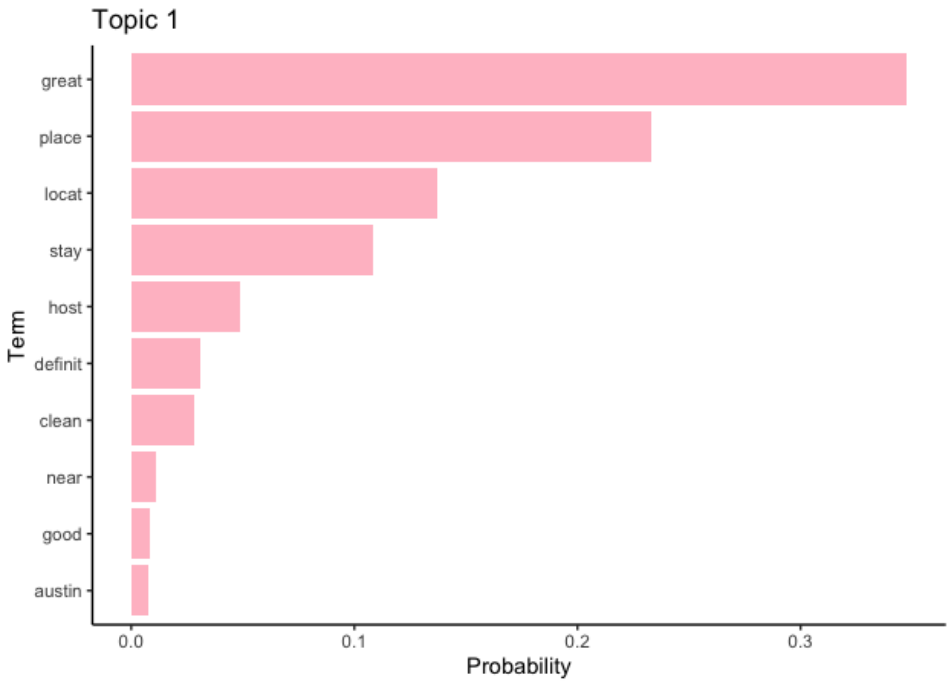
These charts display the mean polarity scores ranging from -0.1 to 0.3 for review score ratings between 1 star and 4.9 stars. The scatter plot points represent the mean polarity scores for each rating, and the line connects the points, showing the overall trend of mean polarity scores across the range of review score ratings.

The line connecting the scatter plot points slopes upwards, it suggests a positive correlation between higher-rated reviews and more positive sentiment. In other words, as the review score ratings increase, the mean polarity scores also tend to increase, indicating a stronger positive sentiment in the reviews.

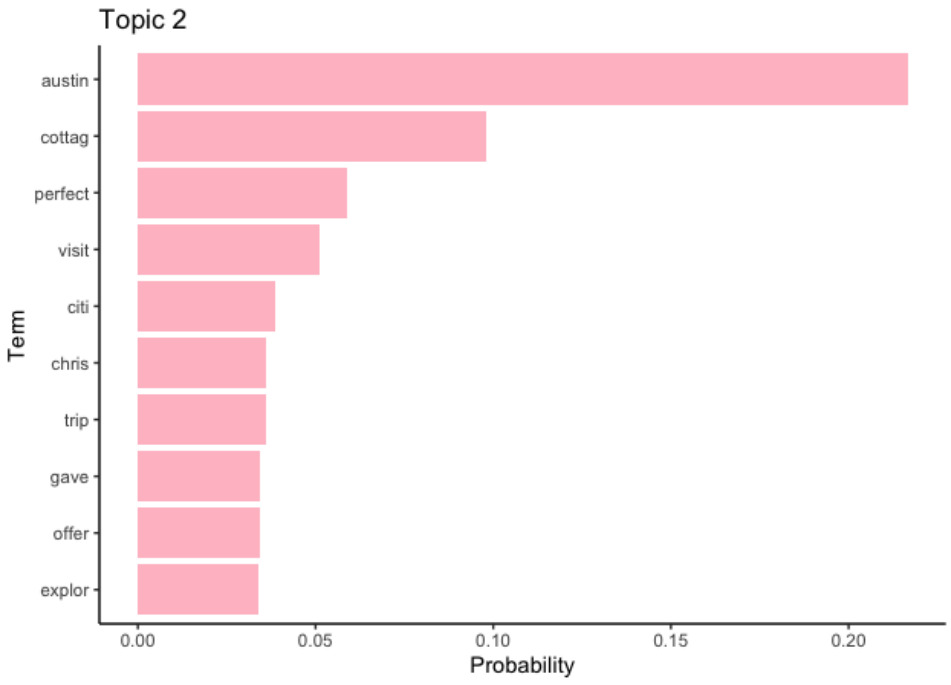
This positive relationship between higher review score ratings and more positive sentiment aligns with the common expectation that higher-rated reviews generally reflect more favorable experiences or opinions. It implies that customers who have had positive experiences or hold positive opinions are more likely to give higher ratings to the product, service, or experience being reviewed.

Topic Word Distributions

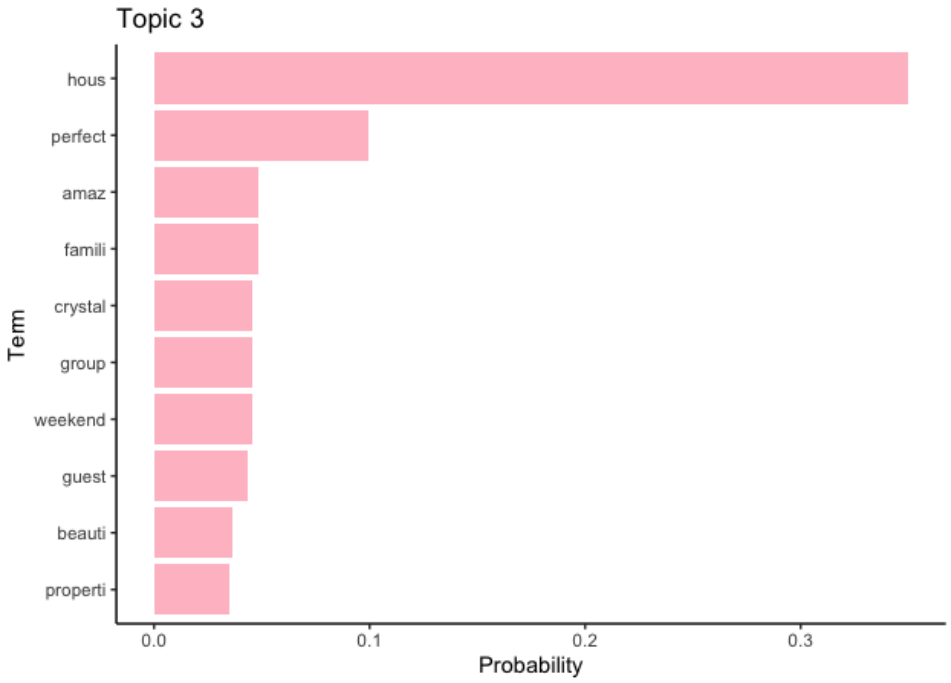
Positive Experience - This topic could represent the positive experiences of guests at Airbnb. It may include reviews that mention the great overall experience, positive comments about the place and cleanliness, satisfactory stays, and good hospitality from the host.



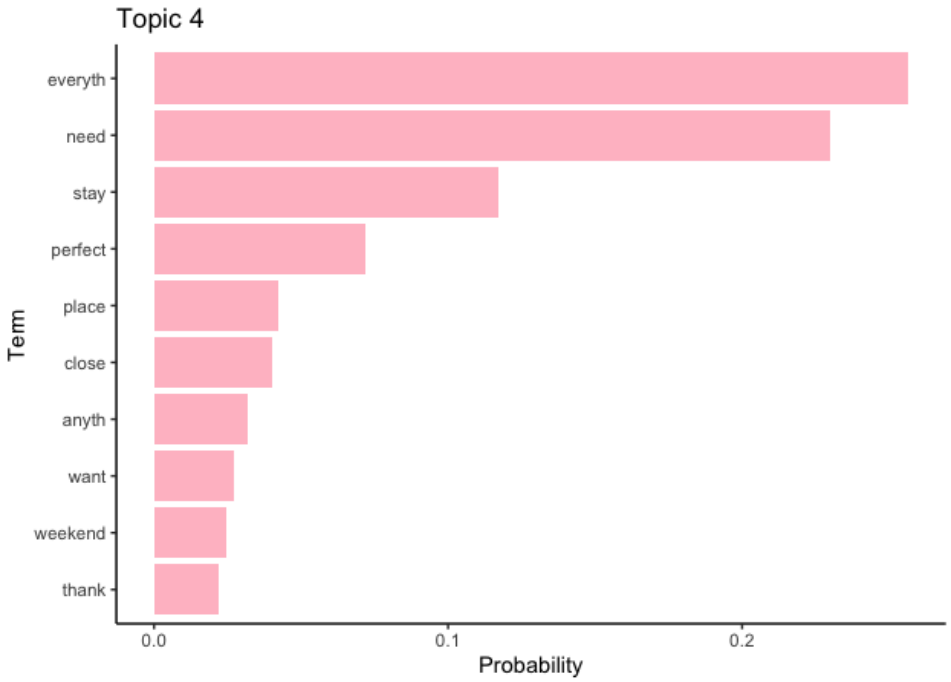
Austin Exploration - This topic could represent reviews or discussions related to exploring and experiencing the city of Austin. It might include mentions of visiting Austin, the attractions or activities available in the city, recommendations for things to do or places to visit, and offers or suggestions for exploring different aspects of Austin.



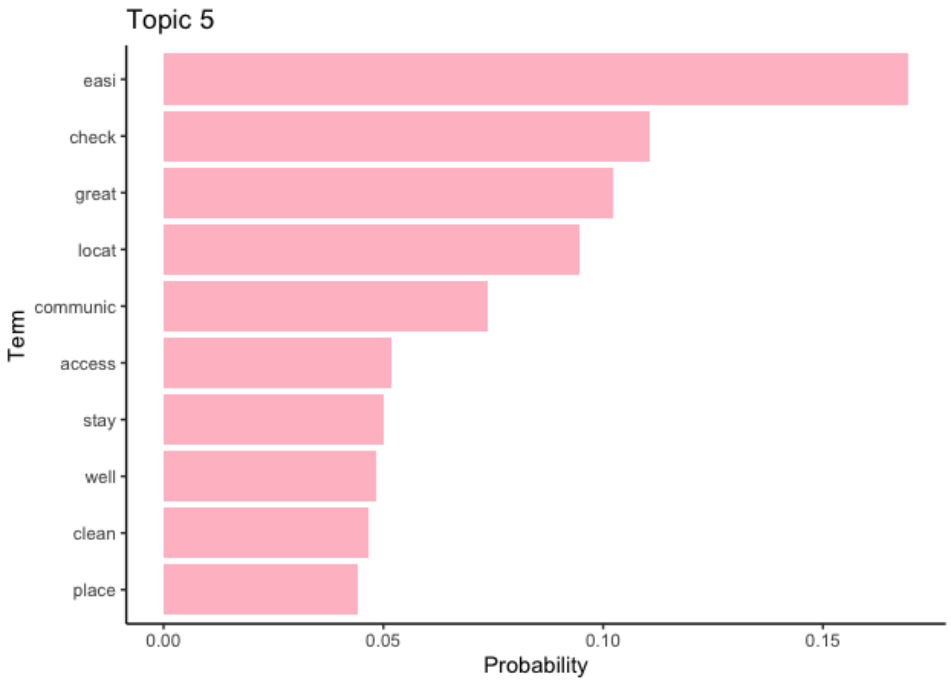
Perfect Vacation Rental - This topic could represent reviews or discussions about vacation rental properties.



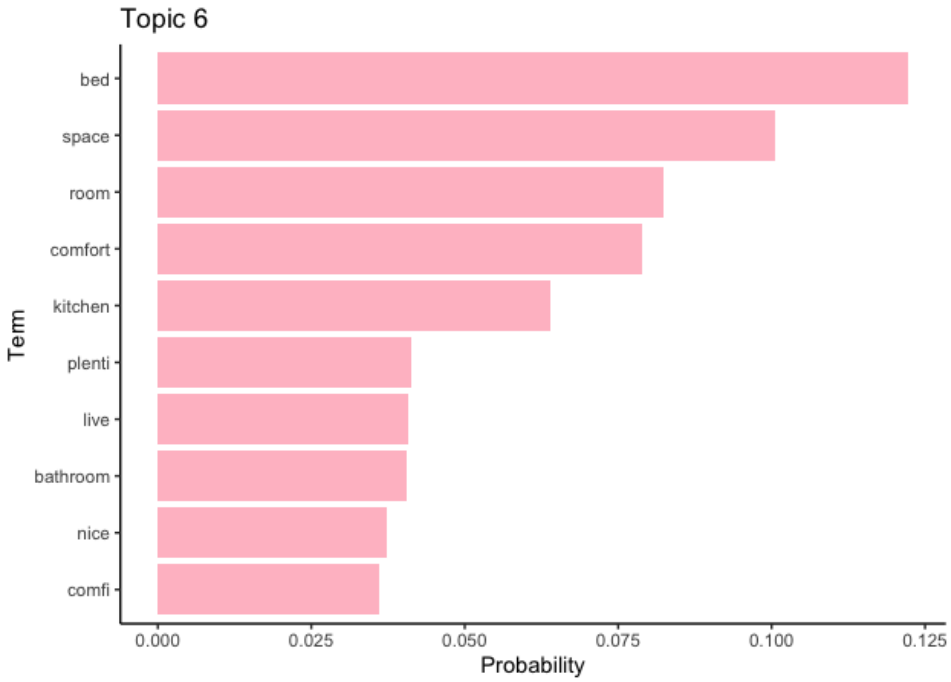
Perfect Accommodation - This topic could represent reviews or discussions about a perfect accommodation or place to stay



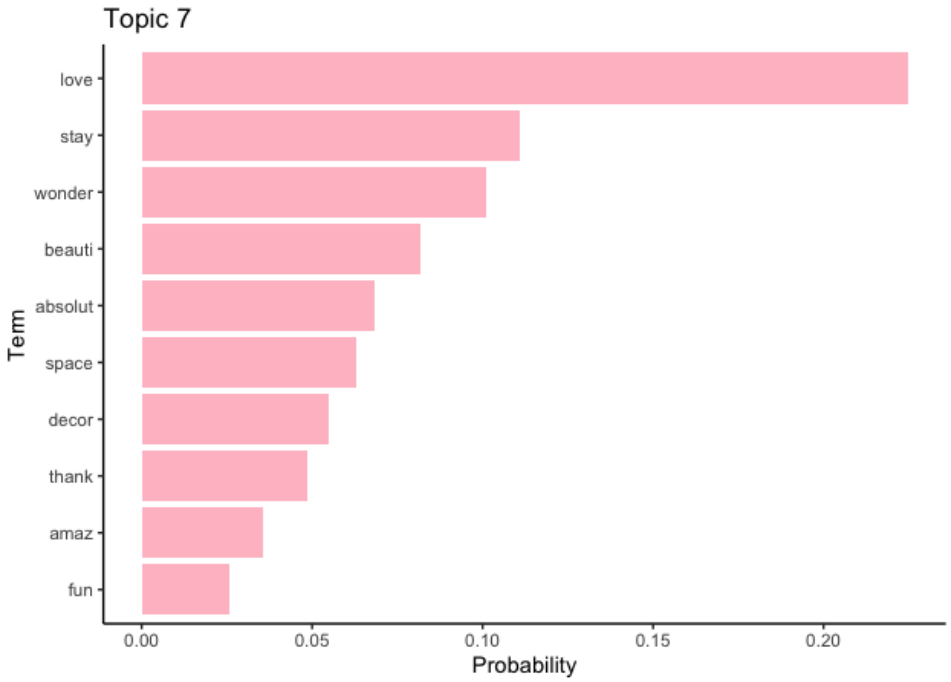
Convenient and Accessible Stay - This topic could represent reviews or discussions about the convenience and accessibility of the stay



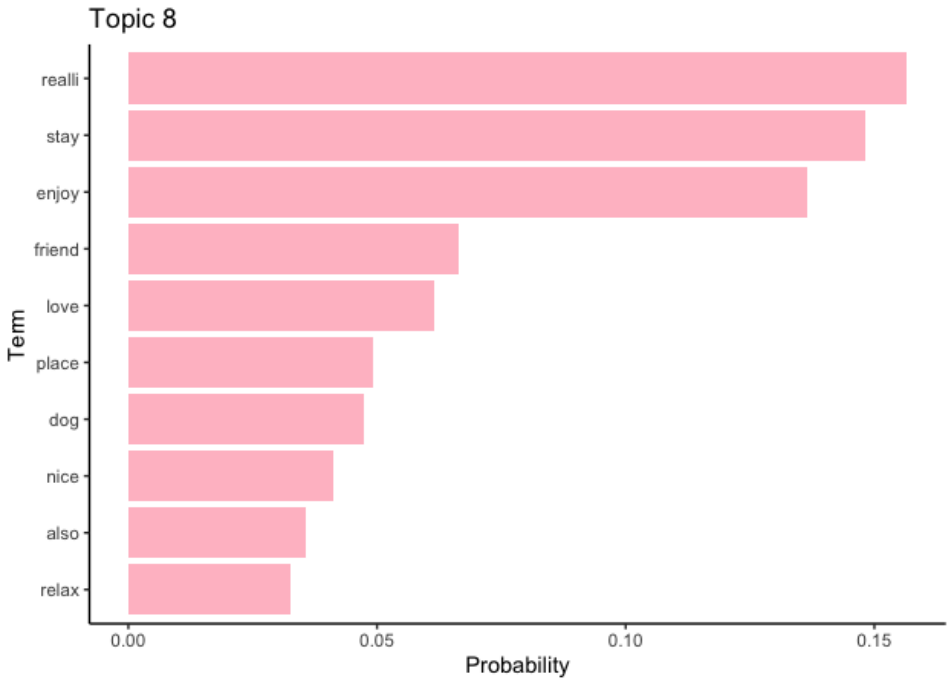
Living Spaces and Amenities - This topic might focus on the living spaces and amenities provided. It could include reviews or comments that highlight the quality of the bathroom facilities, positive remarks about the overall niceness of the accommodations, and the comfort of the living areas



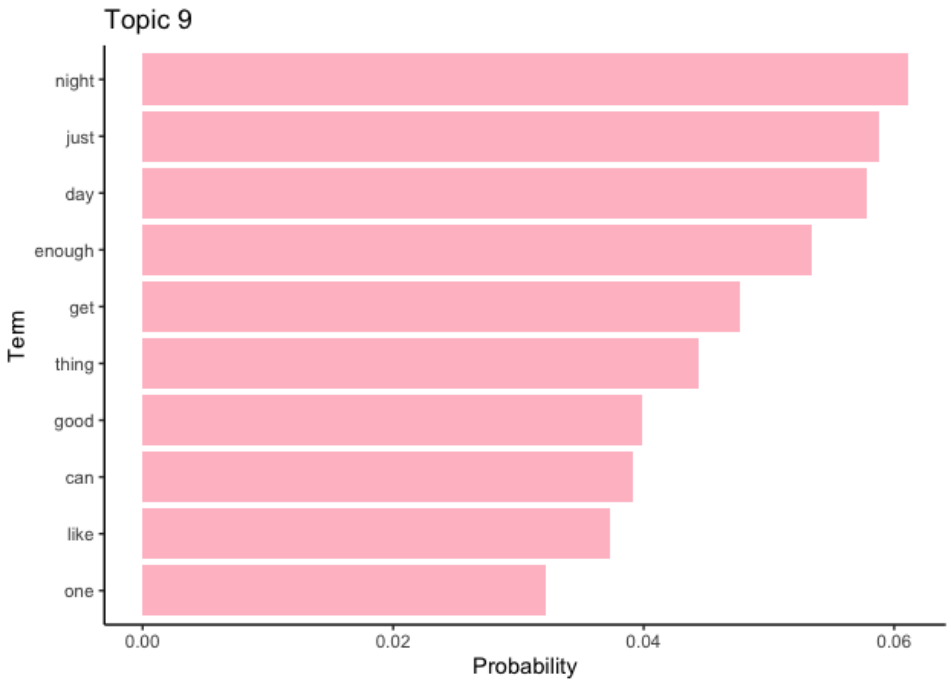
Enthusiastic Experience - This topic could represent reviews or discussions expressing enthusiasm and positive experiences.



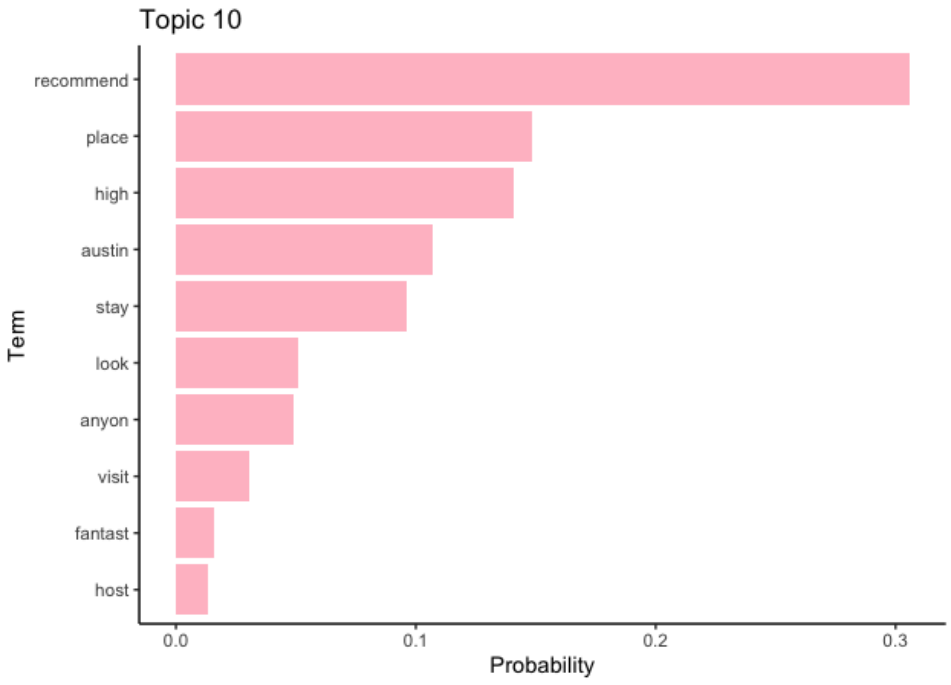
Enjoyable - This topic could represent reviews or discussions about an enjoyable stay experience



Pet-Friendly Accommodations - This topic might focus on pet-friendly accommodation. It could include reviews or comments from guests who stayed with their dogs, highlighting the friendliness and acceptance of pets in the place



Fantastic Hosts - This topic could revolve around fantastic hosts and their impact on the guest experience



Overall, understanding the potential topics that could emerge from a topic model can be helpful in several ways:

+ Understanding Customer Preferences: By identifying the topics that are frequently discussed or mentioned in customer reviews or discussions, businesses can gain insights into the preferences and priorities of their customers. This understanding can inform decision-making processes related to product improvements, service enhancements, or marketing strategies.

+ Improving Customer Experience: Analyzing the topics can reveal common positive or negative experiences reported by customers. This information can help businesses identify areas of improvement or areas where they excel, allowing them to focus on enhancing the aspects that matter most to their customers and addressing any pain points.

+ Competitive Analysis: Analyzing the topics discussed by customers can also provide insights into how a business compares to its competitors. By understanding the topics associated with positive sentiments

or differentiation, businesses can identify opportunities to stand out in the market and identify areas where they can improve to gain a competitive advantage.