SESUG Paper 229-2019

Analyzing Airbnb reviews using SAS® Text Miner and Predicting the factors contributing for higher ratings

Ashlesha Sampathi, Oklahoma State University

ABSTRACT

Airbnb, Inc. is a privately held global company headquartered in San Francisco that operates an online marketplace and hospitality service that is accessible via its website and mobile apps. Members can use the service to arrange or offer lodging, primarily homestays, or tourism experiences. It is the world's largest home sharing company and has over 4 million listings in more than 81,000 cities and 191 countries.

Airbnb projects the prospect of making money by renting out our home with the platform. But homeowners, especially those renting out their homes for the first time, may have many questions: What price should I set my home at? Can I trust my home to guests? How can I ensure I get a good rating?

Customer reviews and ratings play an important role in boosting a customer loyalty towards a brand. In fact, they can make or break a business .Therefore, it is very important for businesses to analyze the factors that are driving higher ratings and it is also equally important to have an overview of public opinion on the product.

In this paper, I have predicted the main factors that are driving higher ratings and analyzed the reviews from customers using SAS® text miner and compared them with numerical ratings to analyze the correlation between written reviews and numerical ratings. In addition, I've also performed descriptive analysis to explore some key points which would be very helpful for business such as:

1. What are the most popular times of the year for Airbnb rentals in Seattle?
2. Which locations in Seattle are most valued according to Airbnb customers?

An open dataset from insideAirbnb website was used for the research purpose. The dataset provides information on home features, review scores, comments and the availability of 8,460 listings in Seattle till the year 2019. SAS® Viya was used to conduct visual analytics on the Airbnb data and SAS® Studio to perform linear regression to predict the factors driving higher ratings. In addition, SAS® Text Miner was used for text mining of customer reviews.

INTRODUCTION

Today tourism has become integrated part of our life and within past few years, this industry has gained a great deal of popularity. People these days want to see world at minimum expenses. Hotels are one of the main concerns while planning for a vacation, business trips etc. Airbnb is a well-known company which has provided a platform for hosts and visitors to communicate. It is the world's largest home sharing company and has over 4 million listings in more than 81,000 cities and 191 countries. Members can use the service to arrange or offer lodging, primarily homestays, or tourism experiences.

Before Airbnb, it would have been a nerve-wracking prospect to let strangers stay in your home. Airbnb has changed this — with a mission that “connect people with places to stay and things to do around the world.”. The company has transformed the relationship between the homeowner and the renter. Most of us are familiar with the experience as guests, renting homes to stay in on Airbnb. But I was interested in the perspective of a homeowner. Homeowners, especially those renting out their homes for the first time, may have many questions on price to be set for the home and on obtaining good ratings.

According to a survey, 86% of people will hesitate to purchase from a business that has negative online reviews[1]. Fascinated by this fact, I did my research on predicting the factors contributing for higher ratings for Airbnb and also performed a text analysis of reviews from customers and compared them with numerical ratings to analyze the correlation between written reviews and numerical ratings. The purpose of this project is to analyze the important factors responsible for positive reviews in Airbnb and provide an insight to the hosts.

DATA DESCRIPTION

The datasets used for this research paper have been taken from insideairbnb.com website. The data collected from this source was a demographic listing file, outlining all Airbnb properties in Seattle, a calendar file providing information on availability of the listings and a review file providing customer reviews till the year 2019. Some of the important variables in the datasets are listed below:

1. **Reviews**

Variable	Description
Listing_ID	Listing Id of the property
Date	Date on which the review was given
Reviewer_ID	ID of the reviewer
Reviewer_name	Name of the Reviewer
Comments	Comments posted by reviewer

2. **Listings**

Variable	Description
name	Name of the property
Host_Id	Id of the property host
Host_response_time	Time taken by host to respond
host_total_Listings_count	Total listings the host has
Cancellation_policy	Cancellation policy of property
Weekly_price	Weekly price of the property
Monthly_price	Monthly price of the property
Review_scores_rating	Overall score given for the property by reviewer
review_scores_accuracy	Score given for accuracy of the property
review_scores_cleanliness	Score given for cleanliness of the property
review_scores_communication	Score given for communication by the host
review_scores_location	Score given for location of the property
review_scores_checkin	Score given for ease of checkin
Neighborhood group	Neighborhood of the property
Min_nights	Minimum number of nights required to book
Host is Superhost	Indicates whether a host is superhost or not
Reviews_per_month	Average Number of reviews in a month

3. **Calendar**

Variable	Description
Listing_id	Listing id of the property
Date	Date
availabilty	Availability of property on that date
price	Price of the property on that date

DATA CLEANING:

As part of data cleaning, the variables have been selected on the basis of data consistency, and relevancy towards the goal. For example, the dataset had variables with only one level or variables like zip code and ID's which weren't required for the analysis. To further prepare and analyze fields in the listings file, text analytics was performed on the text fields in the listings file to extract relevant text topics. Through this analysis, the significant text variables were determined to be interaction, access, Amenities and House Rules.

There is missing information for weekly price, monthly price, cleaning fee, security deposit and extra person fee. I've imputed them with zero as those fees are not applicable to certain listings. From my initial analysis, I suspected some extreme observations. So, I looked at the distribution of all continuous variables and removed all the extreme observations after three standard deviations. Some such observations were, listings that were charging $10,000 for one day and listings with maximum nights as a four-digit number. Observations with missing review scores rating(target variable) were deleted as they would not be helpful for the analysis.

DATA EXPLORATION

Most valued locations in Seattle:

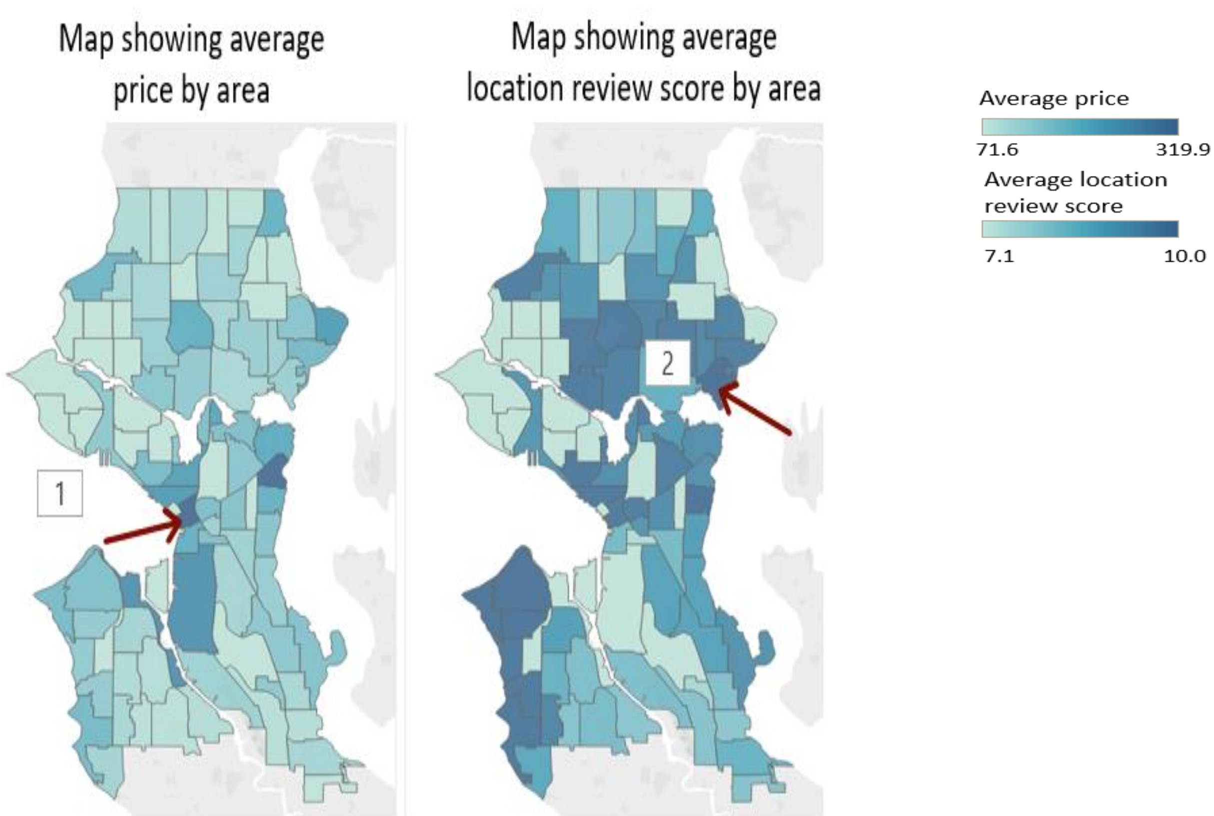

Figure1 Dashboard of Average price and location review score by neighborhood

1. The average price is seen to be highest for Central Business district i.e., downtown Seattle. This is expected due to obvious reasons. But the average location review score also seems to be on the high end for this neighborhood. This indicates that the people are seeing value worth the price in that neighborhood.
2. Average Location review score is seen to be highest for Laurelhurst neighborhood. Surprisingly, average price is on the lower end for this area. This indicates that people are considering it as a good deal at that price.

Most popular times of the year for Airbnb rentals in Seattle:

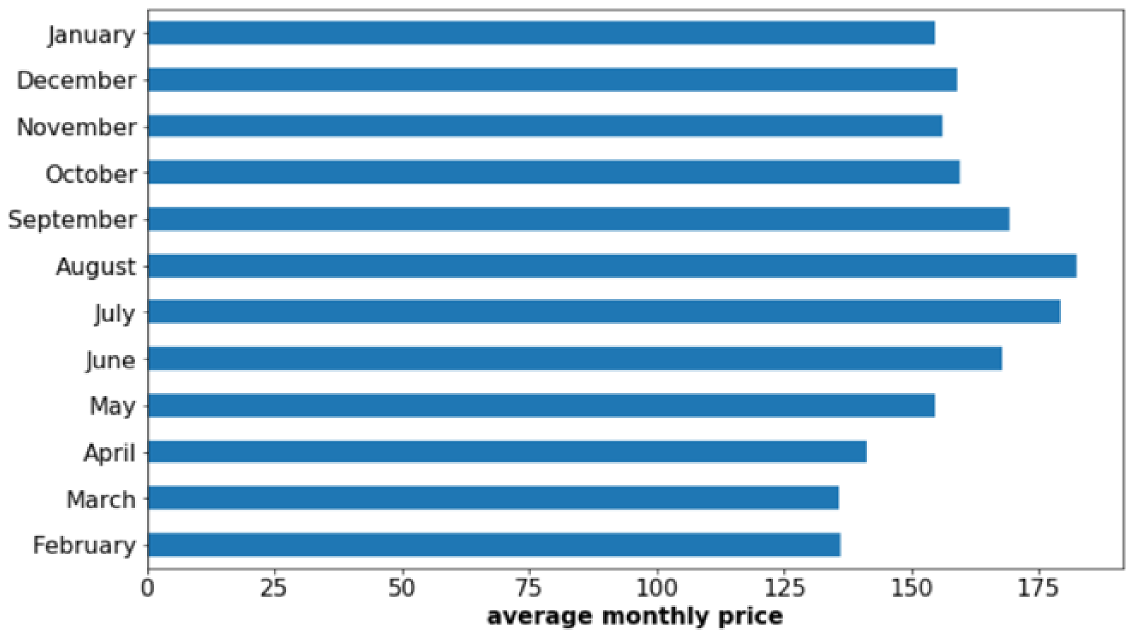

Figure2 Average monthly price of listings

Figure2 shows that average price of a listing is highest in the months of July and August which signifies that there is more demand for the hotels in Seattle in the month of August, then followed by July. To put it simply, July and August are the best months to visit Seattle as it usually has the best weather in these months[2]. This fact is reinforced with the analysis here.

Busiest Times of a Week for a listing in Seattle:

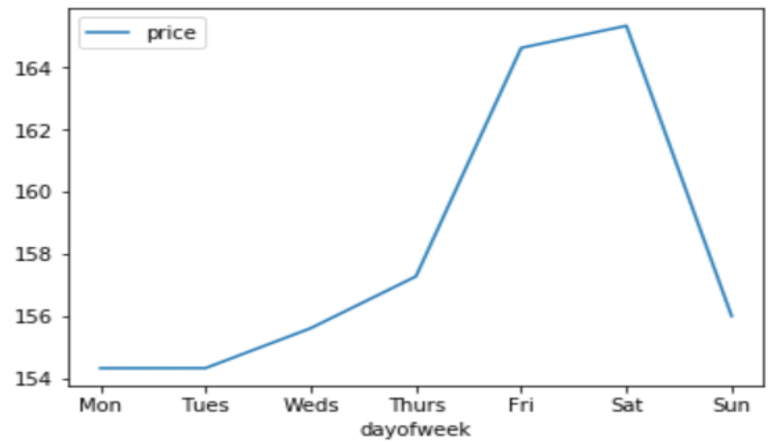

Figure3 Average Weekly price of listings

Figure 3 shows the variation in rental price for different days in a week. As expected, the prices show a spike on weekends (Friday and Saturday).

MODELING:

Predicting the factors contributing for higher ratings:

A new variable called "Rating category" was created to indicate if the overall rating given by the user was high or low. The original review_scores_rating variable is on a scale of 0-100. For the purpose of my analysis, if the rating is <50 it was considered a "low" rating and if the rating > 95 it was considered as a "high rating". Grouping of review scores was done in such a way in order to maintain balance in both the classes since there were very few observations which had original review rating scores on the lower side of the scale. To Predict the factors contributing for higher ratings, two classification models were built in SAS enterprise miner using rating category as dependent variables and all the other selected variables as independent variables.

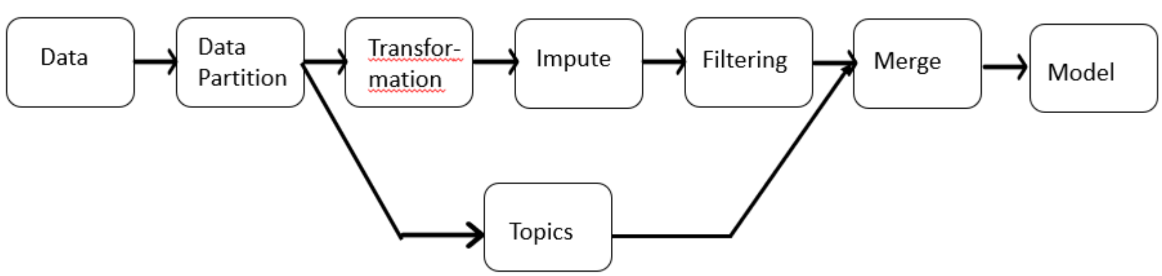

Figure4 Process flow of the modeling

At first, Listings dataset was imported into the enterprise miner. Data was partitioned into 70-30 split using stratified sampling method i.e., 70% of the data as used as training data and 30% was used as validation data. All the independent variables have been transformed to ensure normal distribution of the independent variables. This was done using Transformation node in enterprise miner. The missing values for variables like weekly price, monthly price, cleaning fee, security deposit and extra person fee have been imputed with zeroes since some of the listings may not have those properties. Outliers have been excluded from the analysis using filter node in the enterprise miner.

The data had 4 important text variables House rules, Amenities, interaction, access as mentioned under Data Preparation section. Text topics were created for each of these variables and used as input into the models. All the exported data from the text topics were merged and a final dataset with all the text topics and other non-text variables were used for modeling. Four Binary classification models have been built namely logistic regression using non-text variables, logistic regression using all the variables, Decision tree using non-text variables and Decision tree using all the variables. The Decision tree model in non-text approach was selected as the best model based on validation misclassification rate.

Model	Misclassification rate
Decision tree(non-text)	**15.71%**
Decision tree(all variables)	17.23%
Logistic regression(non-text)	16.89%
Logistic regression(all variables)	17.44%

Results:

Decision tree(non-text) model has been selected as the best model based on misclassification rate. Important variables based on the selected model are as follows:

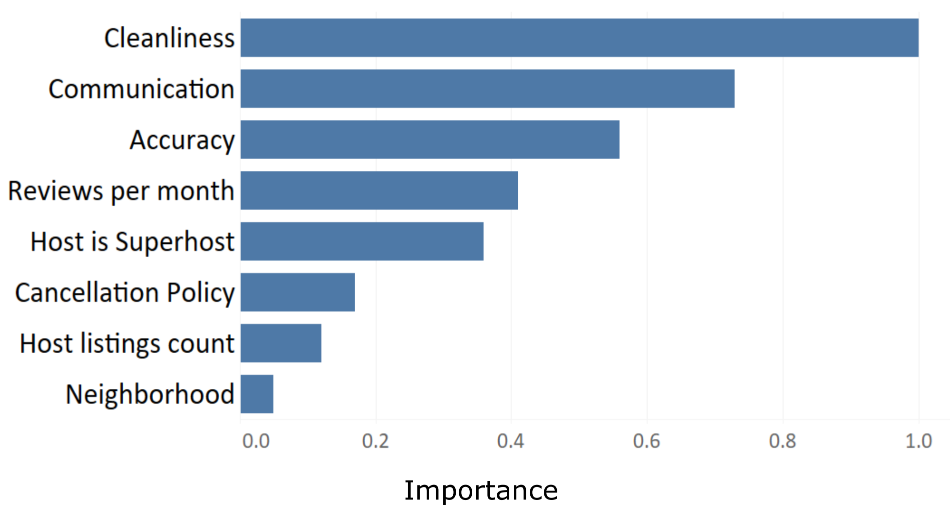

Figure5 Variable importance from the selected model

- According to the model, Cleanliness of the property is the most important variable for achieving a higher rating. This is true since guests usually prefer a hotel with good maintenance

- Communication by the host is considered as the second most important variable in obtaining a high rating
- Accuracy of the listing represented on the Airbnb website is considered as third important variable driving higher rating. This means it is better to provide accurate details and pictures of the listing on the website in order to achieve a good rating
- Reviews per month is the next important variable contributing for higher ratings. This can be explained by a simple fact. People generally check the number of ratings for a listing in order to estimate the authenticity of the property listed.So, it is better to ask the guests to leave a review after their stay to obtain a good rating in future
- Host is superhost and count of host listings are also seen to be the important variables for a good rating. Hosts with creditable service are generally rewarded with a Superhost status as a recognition. It would therefore not be a surprise that more experienced hosts tend to be more knowledgeable in what guests would appreciate in a home, hence leading to higher ratings.
- Cancellation policy is also seen to be important factors driving higher ratings. It is true that people tend to like a listing with flexible cancellation policy more than the one with strict cancellation policy leading to higher rating for the prior one.
- Neighborhood is also a important variable responsible for a high rating. This can be because of general tendency to like quite and nice neighborhoods more than the other ones leading to good ratings for the former one.

TEXT ANALYSIS ON REVIEWS:

The following figure shows the methodology used for the analysis:

After obtaining the variable importance from decision tree model, analysis of text reviews on same listings was done to learn the sentiment of people expressed in text reviews and to analyze the correlation between written reviews and numerical ratings.

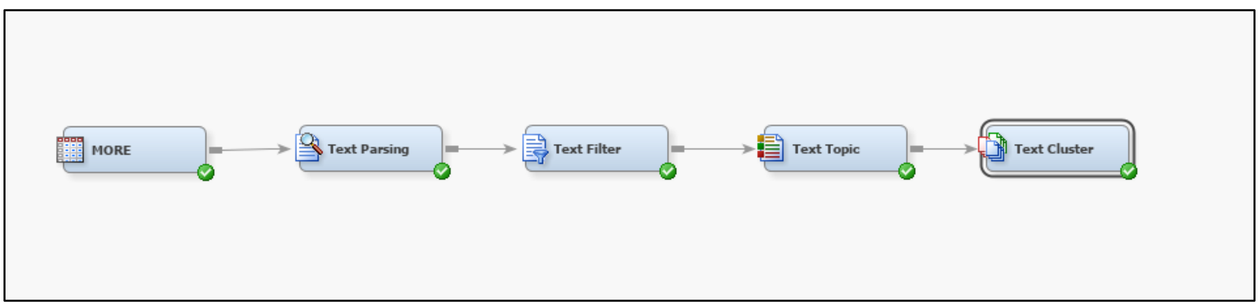

Figure6 Node diagram to generate clusters

TEXT PARSING

The text parsing node was connected to the data node and default settings are modified using properties panel before running the node.

Following parameters are used in properties panel:

• The 'detect different parts of speech option' is set to 'yes' to be able to treat the same words of different parts of speech as different.

• 'Abbr', 'Aux', 'Conj', 'Det', 'Interj', 'Num', 'Part', 'Prep', 'Pron', 'Prop' parts of speech have been ignored.

• 'Num', 'Punct' types of attributes have been ignored.

The text parsing node generated a term by document matrix helps to identify the most frequently occurring words and the number of comments in which each word occurred.

Below figure displays partial term by document matrix for Airbnb reviews dataset

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ be	...Verb	Alpha	72856	23698	N	+	65782	1
+ place	...Noun	Alpha	18380	13562	Y	+	19483	2
very	...Adv	Alpha	18297	12473	N		65580	3
+ stay	...Verb	Alpha	14567	11724	Y	+	697	4
+ great	...Adj	Alpha	14619	11534	Y	+	9376	5
+ have	...Verb	Alpha	17042	11429	N	+	65612	6
seattle	...Prop	Alpha	12150	9623	Y		36264	7
+ host	...Noun	Alpha	9405	8864	Y	+	25789	8
+ location	...Noun	Alpha	9176	8734	Y	+	24616	9
+ stay	...Noun	Alpha	8558	7370	Y	+	29549	10
+ clean	...Adj	Alpha	7438	7275	Y	+	27354	11
comfortable	...Adj	Alpha	7764	7186	Y		19992	12
+ recommend	...Verb	Alpha	7214	7088	Y	+	4671	13
again	...Adv	Alpha	7109	6803	N		65494	14
+ home	...Noun	Alpha	7255	5341	Y	+	46574	15
definitely	...Adv	Alpha	5410	5239	Y		3602	16
not	...Adv	Alpha	6664	5151	N		65644	17
+ neighborhood	...Noun	Alpha	5263	4928	Y	+	35892	18
here	...Adv	Alpha	5361	4891	N		65522	19
+ nice	...Adj	Alpha	5656	4852	Y	+	13366	20
+ make	...Verb	Alpha	5676	4808	N	+	65709	21
beautiful	...Adj	Alpha	5170	4774	Y		50222	22
+ house	...Noun	Alpha	6139	4635	Y	+	42362	23
+ space	...Noun	Alpha	5411	4581	Y	+	15896	24
+ need	...Verb	Alpha	5018	4514	N	+	65895	25
so	...Adv	Alpha	5122	4106	N		65690	26
+ love	...Verb	Alpha	4612	4068	Y	+	7242	27
wonderful	...Adj	Alpha	4473	4053	Y		35006	28
+ easy	...Adj	Alpha	4380	3968	Y	+	20834	29
+ time	...Noun	Alpha	4440	3904	Y	+	26145	30
+ restaurant	...Noun	Alpha	3961	3839	Y	+	3182	31
+ good	...Adj	Alpha	4218	3768	Y	+	4069	32
highly	...Adv	Alpha	3823	3744	Y		23773	33
really	...Adv	Alpha	4380	3619	N		65535	34
+ do	...Verb	Alpha	4532	3536	N	+	65857	35
+ bed	...Noun	Alpha	3729	3525	Y	+	17355	36
+ get	...Verb	Alpha	3964	3419	N	+	65534	37
+ area	...Noun	Alpha	3791	3306	Y	+	32890	38
+ room	...Noun	Alpha	4010	3305	Y	+	10428	39
+ enjoy	...Verb	Alpha	3597	3258	Y	+	6741	40
also	...Adv	Alpha	3721	3160	N		65480	41
+ apartment	...Noun	Alpha	4003	3092	Y	+	31850	42
+ feel	...Verb	Alpha	3346	3062	Y	+	19002	43
perfect	...Adj	Alpha	3284	3057	Y		35780	44
s	...Noun	Alpha	4093	3048	N		65531	45
+ view	...Noun	Alpha	3219	2963	Y	+	9378	46
+ quiet	...Adj	Alpha	3007	2945	Y	+	8600	47
+ thank	...Verb	Alpha	2741	2689	N	+	65922	48
+ coffee	...Noun	Alpha	2912	2678	Y	+	20639	49
+ go	...Verb	Alpha	2987	2646	N	+	65567	50
downtown	...Noun	Alpha	2740	2634	Y		37414	51
close to	...Adv	Alpha	2718	2621	Y		26156	52

Figure7 Term by document matrix result from text parsing node

Some of the most commonly used words in the reviewer comments were clean, comfortable, neighborhood, stay which describe the features related to housing.

TEXT FILTER

Further, parsing node is then connected to text filter node as shown in figure. This node filters out words that occur least number of times specified in the properties panel. Following parameters are changed from default settings in properties panel:

• The minimum number of documents is set to 4.

• Check spelling option is set to 'yes', which enables SAS® to create correctly spelled words in place of misspelled words.

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
place	18385	13566	✓	0.089	Noun	Alpha
place	17498	13182			Noun	Alpha
places	882	860			Noun	Alpha
placeâ	4	4			Noun	Alpha
plac	1	1			Noun	Alpha
stay	14581	11731	✓	0.1	Verb	Alpha
stayed	1	1			Prop	Alpha
stay	8842	8046			Verb	Alpha
staging	3	3			Noun	Alpha
stayi	1	1			Noun	Alpha
staying	2	2			Prop	Alpha
stacy	2	2			Prop	Alpha
ststaying	1	1			Noun	Alpha
stays	67	67			Verb	Alpha
stayed	2310	2167			Verb	Alpha
stayâ	4	4			Noun	Alpha
staying	3348	3093			Verb	Alpha
great	14626	11539	✓	0.102	Adj	Alpha
great	14569	11506			Adj	Alpha
gerat	1	1			Noun	Alpha
greatâ	2	2			Noun	Alpha
greate	1	1			Noun	Alpha
geat	2	2			Noun	Alpha
greater	25	25			Adj	Alpha
greatest	25	25			Adj	Alpha
grea	1	1			Noun	Alpha
seattle	12166	9637	✓	0.12	Prop	Alpha
seattles	5	5			Prop	Alpha
seattle	12150	9623			Prop	Alpha
seattel	1	1			Prop	Alpha
seatttle	1	1			Prop	Alpha
seattleå	1	1			Prop	Alpha
seatle	2	2			Noun	Alpha
seatttle	1	1			Noun	Alpha
seatel	1	1			Prop	Alpha
seattlites	1	1			Noun	Alpha
seattleì	1	1			Prop	Alpha
seattlite	2	2			Prop	Alpha
host	9414	8871	✓	0.122	Noun	Alpha
location	9198	8748	✓	0.123	Noun	Alpha
stay	8558	7370	✓	0.143	Noun	Alpha

Figure8 Grouping of word forms and misspelled words

Figure 8 from interactive filter viewer shows various forms of words ‘place’,’stay’,’great’ which are formed into groups along with misspelled words with the help of English dictionary.

Concept links

Concept links can be viewed in the interactive filter viewer from the properties panel of text filter node. Concept links are the type of association analysis between the terms used. They can be created for all the terms that are present in the comments, however, it is meaningful to create links for only a few important terms. Concept link diagram shows the word to be analyzed at the center and the words which are associated with that word are connected to it using links. The thickness of the link explains the strength of association between the two words in reviewer comments. Below are the Concept links for some of the most frequent terms in listings reviews with positive ratings:

The concept link for the word amazing shows that people are really amazed by things like hospitality, location, home view and amenities

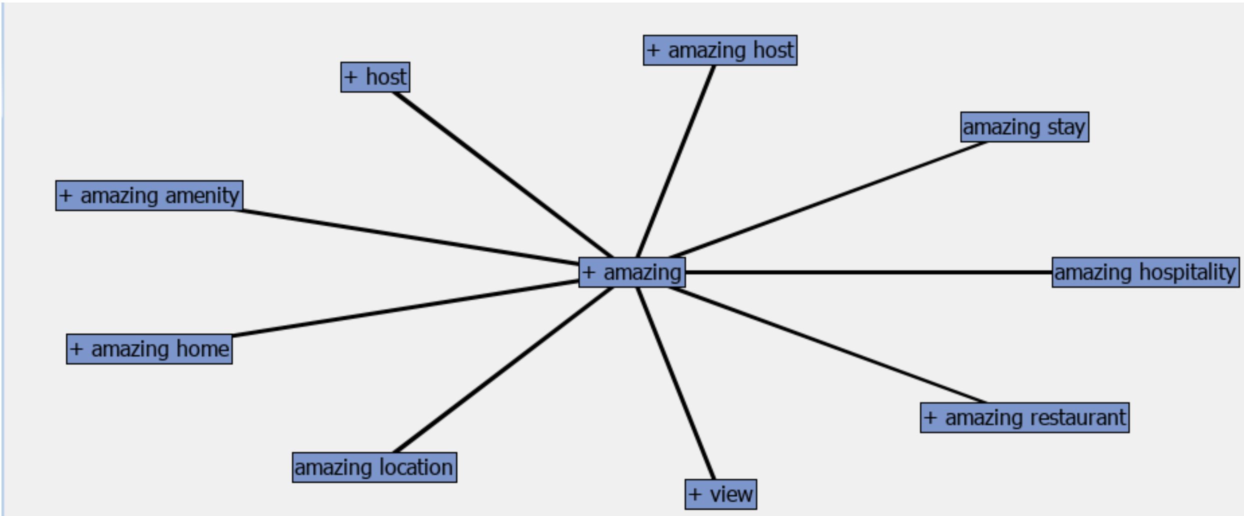

Figure9 Concept link for word 'amazing'

The concept link for the word 'cancel' shows that host was accommodating when the guests cancelled stay on a last minute due to some unforeseen reasons. This indicates that people tend to give higher rating for a listing with flexible cancellation policy.

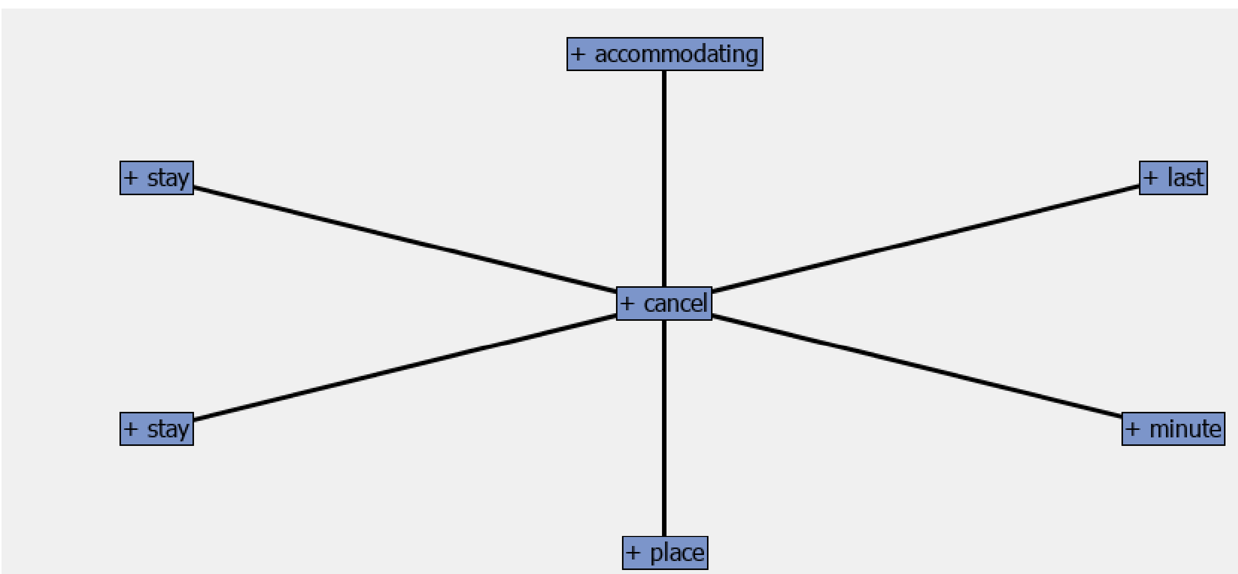

Figure10 Concept link for word 'cancel'

The concept link for communication shows that people give more importance to communication. They like the communication to be clear, fast and easy.

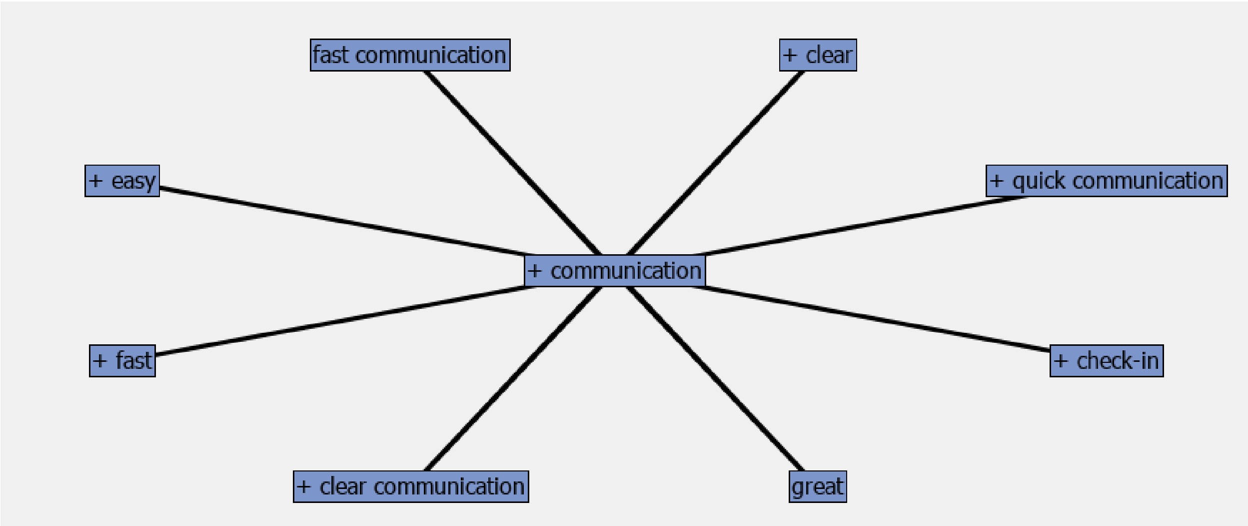

Figure11 Concept link for word 'communication'

Concept link for the word 'neighborhood' shows that people tend to give higher ratings for a safe, peaceful and a scenic view neighborhood.

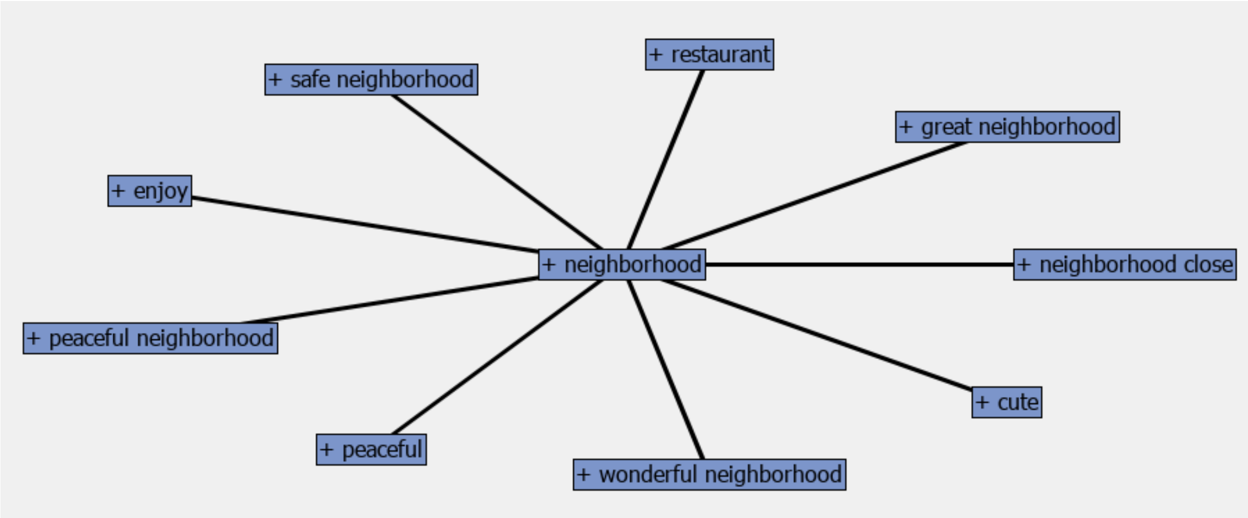

Figure12 Concept link for word 'neighborhood'

Following are the concept links for the most frequent terms in listings reviews with negative ratings:

Concept link for the word 'bad' shows that people tend to give lower ratings if the host is unresponsive, uncommunicative or if the listing is not maintained properly

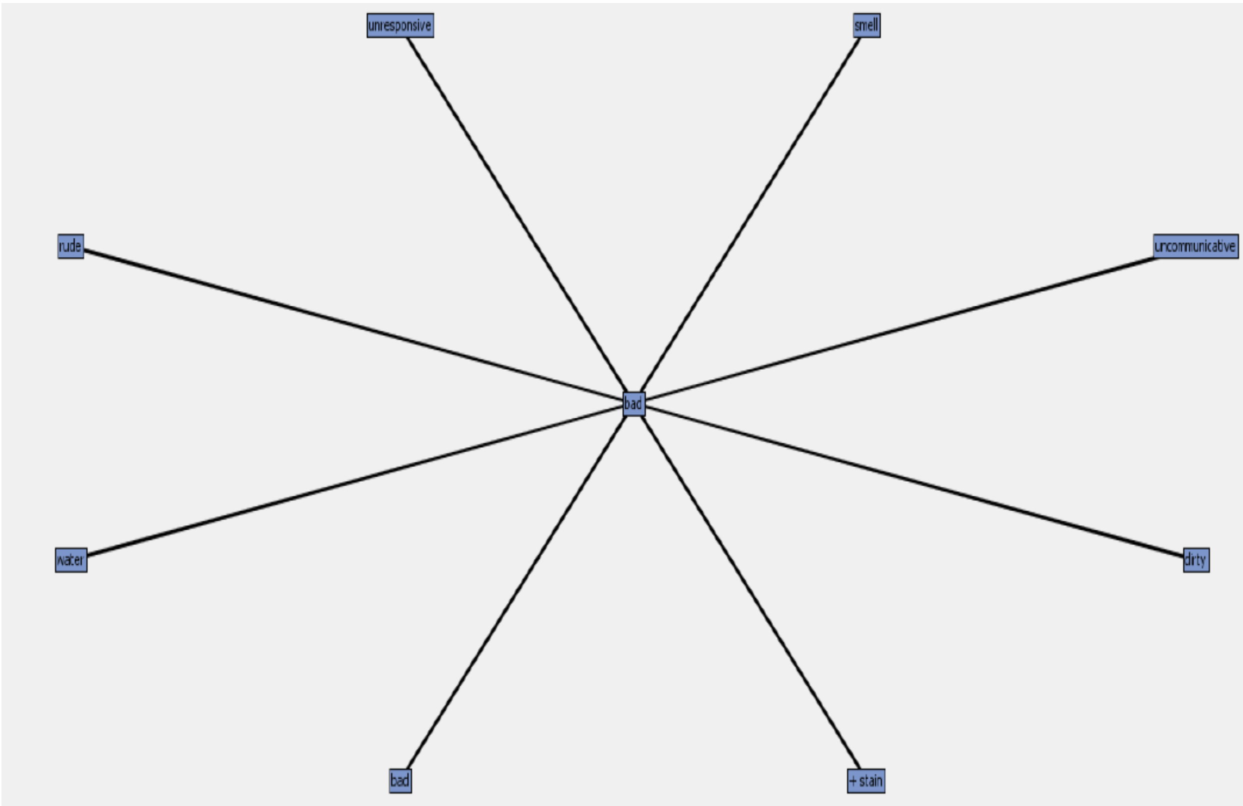

Figure13 Concept link for word 'bad'

TEXT TOPIC

After connecting the Text Filter node in SAS® Enterprise Miner to Text Parsing node, Text topic node is joined to the Text Filter node, which enables SAS® to combine terms into topics for obtaining valuable insights from data. The number of Multi-Term Topics has been set to 20 in properties panel to understand data and get the features that reviewers are more interested to comment about the products.

Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
1	0.067	0.012	+time,next,+seattle,+definitely,+visit	270	3721
2	0.081	0.010	+automate,+cancel,+post,+reservation,+arrival	10	414
3	0.072	0.011	å,æ,ä,é,è	72	189
4	0.071	0.011	+great place,+great,+place,+great host,+host	114	1951
5	0.072	0.012	+walk,close to,+downtown,+restaurant,+short	374	4417
6	0.069	0.013	+bed,+kitchen,+bathroom,+room,+space	453	4121
7	0.066	0.012	+home,+beautiful,+feel,+beautiful home,right	187	4018
8	0.071	0.011	+highly,+recommend,+visit,+fantastic,+apartment	138	4086
9	0.085	0.012	+great host,+host,+friendly,+amazing,+helpful	238	4248
10	0.076	0.011	+distance,+walking,+walking distance,+restaurant,+shop	121	1998
11	0.094	0.011	+great location,+location,+great,+space,+clean	113	2828
12	0.069	0.012	+view,+amazing,+amazing view,+beautiful,+apartment	318	3975
13	0.063	0.012	+experience,+airbnb,+good,+excellent,+first	352	3781
14	0.070	0.011	+place,+great,+stay,+seattle,+nice	114	3203
15	0.091	0.012	+place,+clean,+definitely,+nice,+s place	233	5018
16	0.069	0.012	+wonderful,+wonderful host,+wonderful stay,+wonderful place,...	182	3608
17	0.078	0.012	+quiet,+neighborhood,+quiet neighborhood,+nice,+lovely	274	4240
18	0.066	0.012	+easy,+access,+easy access,+park,+check	185	3538
19	0.066	0.012	+stay,+great stay,+enjoy,+comfortable,+apartment	175	3970
20	0.067	0.012	+house,+family,+nice,+beautiful,+enjoy	336	4073
21	0.059	0.012	+apartment,+nice,+perfect,+good,+spacious	281	2797
22	0.062	0.012	+perfect,+perfect location,+spot,+perfect place,+location	212	3157
23	0.066	0.012	+melissa,+room,+abil,+breakfast,+muffin	384	2403
24	0.061	0.012	+touch,+thoughtful,+nice,+little,+thoughtful touch	375	3441
25	0.057	0.012	+awesome,awesome place,+love,+awesome host,+spot	234	1909

Figure14 Text topic node output for Airbnb reviews dataset

Figure 14 shows 25 different topics with corresponding IDs. Topic 5 states that the house is very close to downtown and it is only a short walk to a restaurant. Topic 9 explains that the host is very friendly and helpful. Topic 11 shows that house is at a great location and is spacious and clean. Topic 17 tells that the neighborhood is quiet and lovely.

TEXT CLUSTER

After some trial-and-error, the properties settings for the Text Cluster node are set to generate well separated clusters in the cluster space. Default settings are used in the properties panel which results in the solution with 10 as the maximum number of clusters and 15 as the number of descriptive terms to describe clusters using Expectation-Maximization Cluster Algorithm.

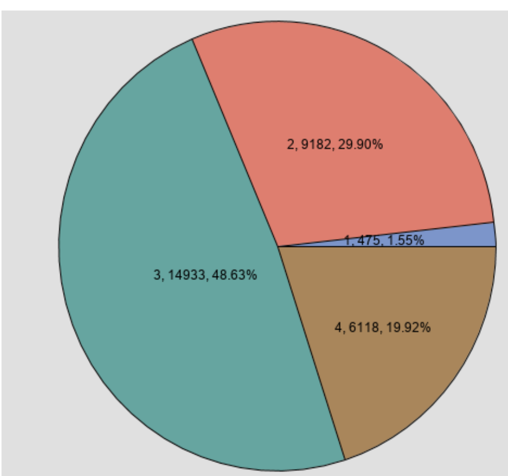

Figure15 Pie chart distribution of clusters obtained from text cluster node

Text cluster node generates 4 clusters from Airbnb reviews data set as shown in figure 16. Cluster 3 has the highest frequency from figure. Words in this cluster occur together, the maximum number of times in comments extracted from Airbnb.

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6	Coordinate 7	Coordinate 8	Coo ate
1	ï º å - æ ƒ é ä è ç ª ã œ çš žå ...	475	2%	0.0995...	0.33793	-0.06724	-0.03821	-0.0134	-0.26003	-.000689	-0.07659	0.0
2	+stay +place +host +perfect +recommend +wonderful +beautiful +home +space +love +definitely +view +time +highl...	9182	30%	0.4256...	-0.02766	0.0035...	-0.00172	-0.02915	0.0153...	0.1480...	-0.00633	0.1
3	+house +restaurant +room +bed +walk +area +coffee +park +kitchen +day +good +feel +easy +enjoy great ...	14933	49%	0.5269...	-0.03379	0.00387	.00079...	-0.02091	0.0028...	0.0200...	-0.00135	0.0
4	+great +place +location +stay +host +clean +'great location' +recommend +highly +nice +'great host' +'great place' +e...	6118	20%	0.3737...	-0.02438	0.0029...	-0.00257	-0.19173	0.0309...	0.1455...	-0.01254	0.1

Figure16 Text cluster node output for Airbnb reviews data

CONCLUSION

It is evident from both the analysis (text mining of reviews and decision tree model for predicting factors contributing for higher ratings) conducted above that good ratings can be ensured by the following factors:

- Good communication by the host
- Maintaining Cleanliness of the place
- Asking users to leave a review after their stay
- Representing accurate details and pictures of the listing on the website
- Maintaining a flexible cancellation policy

REFERENCES

1. https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic
2. https://santorinidave.com/best-time-to-visit-seattle
3. Goutam Chakraborty, Murali Pagolu, Satish Garla, Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® ® by SAS® Institute Inc 2014, Accessed 20 Mar 2017.

ACKNOWLEDGMENTS

I sincerely thank my professors Dr. Goutam Chakraborty and Dr. Miriam McGaugh for their constant support and guidance through every phase of this project

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ashlesha Sampathi
Oklahoma State University
Ashlesha.sampathi@okstate.edu