How to Succeed as an Airbnb Host

——Evidence from Machine Learning and Text Analysis Approaches

Yutian Lai

06/09/2020

Abstract

The prevalence of the sharing economy greatly boosts the vibrancy of Airbnb, but also poses a new challenge to Airbnb hosts to strategically position themselves to attract and serve travelers. To provide better guidance for hosts, this paper explores factors that affect the review score rating, which could infer the satisfaction level of guests, through natural language processing and machine learning framework. The geospatial differences have also been detected by comparing the three localities of London, New York, and Tokyo, respectively containing 14676, 8427, 5288 properties' information in their datasets. The findings suggest cleanliness, communication between hosts and guests, location and verification as super host are the main influencing factors in the final rating score across regions, and the three localities also display significant differences in terms of factors framing the results.

Keywords: Airbnb, sharing economy, rating score, listing attributes, host marketing

1. Introduction

Airbnb, one of the most outstanding marketplace in the sharing economy that facilitates peer-to-peer communication and trust (Airbnb, 2020), provides an alternative accommodation experience for its consumers that cannot be replicated by the conventional hotel industry (Bridges & Vásquez, 2018). With its distinctive operation model and the loosely-regulated sharing-economy condition, it also challenges the theories and practices in consumers' needs identification, which are of great importance in marketing research.

Such a challenge has drawn the attention of many researchers. By analyzing and mining Airbnb data from **insideairbnb.com,** which provides information(including house attributes, consumers' reviews, etc.) about Airbnb's listings around the globe, many researchers have gained unique insight that would help Airbnb refine its business model (Cheng & Jin, 2019; Kalehbasti et al., 2019; Ma et al., 2017; Zhang et al., 2018)

My research project would also employ data from **insideairbnb.com.** I would use datasets of three cities: London, New York and Tokyo, to identify what customers value most behind the score rating system on Airbnb. More specifically, I want to investigate the attributes that become the driving force for Airbnb properties' ratings, so as to assist both the hosts and Airbnb to condition their service to satisfy guests' needs.

2. Literature Review

2.1 Airbnb Ratings

Customer review scores are an approximation of the host/property' s performance. It is regarded as essential in business as it helps offer information when consumers make purchasing decisions, thus promoting trade on the Airbnb platform and increasing the trust of buyers. It is also highly valued in academia as it is a good metric to infer the personal and diverse staying experience of consumers (Zervas et al., 2015).

As previous studies on Airbnb ratings suggested, customer review scores are important "signals" consumers refer to when making decisions (Zmud et al., 2010).

When the sharing-economy marketplace raises the problem of information asymmetry and lack of trust-building since consumers have difficulty discerning the true quality of properties when the only information source they have is the hosts' descriptions (Tsai & Huang, 2009), antecedent guests' staying experiences become an important indicator that is closely related to the service/quality of the host/property, and can boost trust and affect buyers' final decision (Zmud et al., 2010).

Though significant, the customer review score is not a reliable variable to depend solely on when making purchasing decisions. This is because, according to current literature, that the rating scheme loses its effectiveness in reflecting the intrinsic quality of properties as the rating distribution becomes dramatically skewed towards high scores, possibly owing to lack of standard in accommodation rating, the bias of antecedent ratings and underreporting of negative reviews (Tussyadiah & Zach, 2017; Salganik et al. 2006). Researchers have embarked on the investigation into what is driving behind these rating scores. And the findings imply what attributes are shaping the satisfaction level of consumers are highly context-specific (Aiken & Boush, 2006). One group of researchers find consumers put more emphasis on the practical attributes that relate to the condition of the property, including price, location, amenities and cleanliness, etc. (Bridges & Vásquez, 2018; Tussyadiah & Zach, 2017). The other group of researches argue customers value experiential attributes more, which refer to attributes of the hosts and interaction with the hosts/neighborhood, involving whether the host is local, whether the host is super host, whether the host completes host verification, and host response rate/time, etc. (Xie & Mao, 2017; Festila & Müller, 2017).

Based on preceding research, I aim to incorporate together the factors explored by researchers before and other factors that are beyond current literature (e.g. whether the property requires security deposit) which might impact review ratings into machine learning algorithms and feature importance analysis, so as to find the decisive factors in consumers' satisfaction-level. Besides, since the literature suggests the importance of "context" (Aiken & Boush, 2006) in determining the ranking of factors that affect the ratings, I would apply the analysis to three different regions(Tokyo, New York,

London) and compare the results. Hopefully, this piece of research could add more evidence to the debate of what factors are impacting ratings the most.

2.2 Host Description and Text Analysis

The advance of social media offers the public the opportunity to be "generators" of text instead of pure "consumers" (Piryani et al., 2017). Airbnb platform displays a tremendous amount of text data written by hosts/consumers for researchers to capture useful insights into the Airbnb business (Piryani et al., 2017). With the advent of corresponding algorithms and technology, such as natural language processing techniques, which move text analysis to a new stage beyond simple statistical analysis (Edwards et al. , 2017), Airbnb text data analysis draws the attention of more and more researchers. They realize text could be supplementary to the analysis of the macro-level attributes as text focuses on the nuanced aspects of Airbnb experiences. For instance, when macro-level analysis claims that “location” would significantly affect review rating, text analysis could answer what specific elements in the "location" attribute comprise its significance and thus helping researchers have a coherent and detailed view of consumers’ accommodation experiences and satisfaction level (Piryani et al., 2017).

Previous studies on Airbnb dataset text mainly focus on analyzing online review comments (Tussyadiah & Zach, 2017; Cheng & Jin, 2019). Researchers believed in the significance and usefulness of online review comments to identity consumers’ accommodation experience and how they make the judgment (Tussyadiah & Zach, 2017). Cheng & Jin (2019) applied text mining and sentiment analysis to Sydney Airbnb review comments and claimed consumers emphasized certain key influencers of their Airbnb experiences, including “host”, “amenities”, and “location.” Host text, on the other hand, attracted insufficient research focus compared with consumer text. Ma et al. (2017) explored how hosts in 12 major U.S. cities could be more trustworthy by strategically disclosing themselves in their profile. This study adopted text length analysis and topic modeling through latent Dirichlet allocation (LDA), aiming to find strategies behind trust-building on Airbnb. Zhang et al. (2018) also extracted textual

features from New York host descriptions, and analyzed the information amount, sentiment, readability, and semantic topics of the hosts to find what hosts would like to convey to the consumers.

My research would concentrate on the Airbnb host description analysis. The same as other attributes, the description text influences rating but the link is not fully addressed in past literature. To fill this research gap, I would apply bigram analysis and machine learning models with text input to discover the underlying topics in descriptions, aiming to find how hosts with high consumer satisfaction-level are strategically positioning their homes/themselves, and hopefully to discover geospatial patterns in host descriptions from the three cities accordingly.

2.3 Machine Learning Algorithm Used in Airbnb Study

One obvious feature of Airbnb data is its “bigness” (Cheng & Edwards, 2019). For instance, in my research, I would deal with three datasets of in total 40,000 properties’ information. Traditional statistical approaches would be unable to capture insights from the data effectively and efficiently (McAbee et al., 2017). Airbnb researchers now become more proficient in using new techniques that are suitable for the representation of "big" data, among which machine learning algorithms are frequently employed (Cheng & Edwards, 2019). Past research suggested researchers mainly used machine learning models for Airbnb price and trust prediction. Kalehbasti et al. (2019) took host/property attributes, and customer reviews as predictors to predict Airbnb price in New York, aiming to aid both the hosts and the guests with pricing and price evaluation in face of the information asymmetry problem (Tsai & Huang, 2009). Zhang et al. (2018) predicted perceived trust in New York based on numeric features (rating score, host response time, number of reviews, etc.), textual features (information readability, amount, etc.), and image features (facial emotions) of the hosts/properties. Both studies adopted multiple machine learning models, including linear regression, support vector regression, tree-based methods, neural networks, and K-means clustering, etc. and analyzed important impacting factors from the best-performing model.

Research using such a computational framework for rating predicting is relatively rare. My research intends to fill the gap, incorporating host/property attributes and host text data as predictors, employing machine learning algorithms used in previous research, to find what factors are significantly affecting review scores.

To conclude, my study contributes to the Airbnb literature through the investigation into the critical but less addressed topic of rating prediction by combining text analysis and machine learning techniques, which have been proved capable of deriving crucial data-driven insights. Through this study, I aim to obtain a holistic and detailed knowledge of the attributes determining the satisfaction-level of consumers and to generate critical and useful marketing and academic implications.

3. Data

3.1 Data Collection

The Airbnb dataset of the three cities was obtained at http://insideairbnb.com/get-the-data.html , recently updated on April, 2020. For each locality, the dataset contains information about 15000 host homes with 80 features describing almost all aspects of them, including host self-introduction, neighborhood description, number of bathrooms/bedrooms, etc.

3.2 Data Preprocessing

I divided the data preprocessing into four steps. First, I dropped features without sufficient data or unrelated to the response variable. Second, I created new variables based on current variables for the machine learning algorithm application. For instance, I extracted the text-form amenities information into several binary variables to indicate whether the host home has certain amenities. Third, I modified current variables to help them perform more effectively in machine learning algorithms. For example, I encoded certain numerical variables into binary variables (for instance, the variable "security_desposit" changes from "cost of the security deposit" to "whether the host requires security deposit") as the availability of such information could be vital deal-

breakers in guests' decision. Finally, I split and create gram features of the text of "self_about" (hosts' self-introduction) and "description" (description of the host home) so bi-gram analysis can be performed.

4. Methods

I plotted the distribution of the response variable: review_scores_rating in the three localities, confirming previous research regarding the skewness of rating towards high scores(Tussyadian & Zach, 2017). In all three localities, the mean review score is around 93. Such dramatically high score is especially evident in New York City, with its low-rating (<=85) homes covering only 5% of all properties, while the other two cities have more than 10% low-rating homes. Since there are still homes receiving scores relatively lower, the problem I aim to address in my paper is to understand what features can interpret the ratings of Airbnb homes.

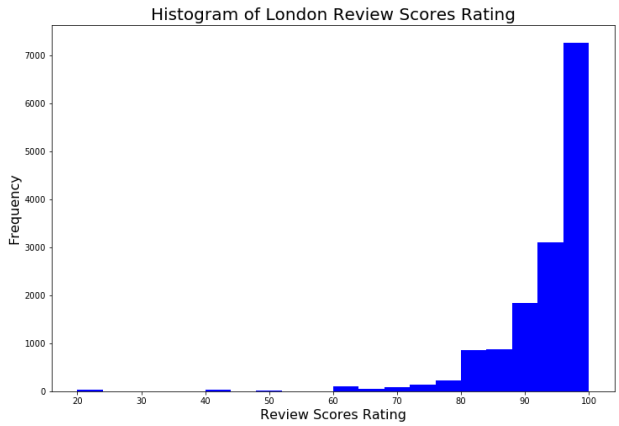

Figure1. Histogram of London Review Scores Rating

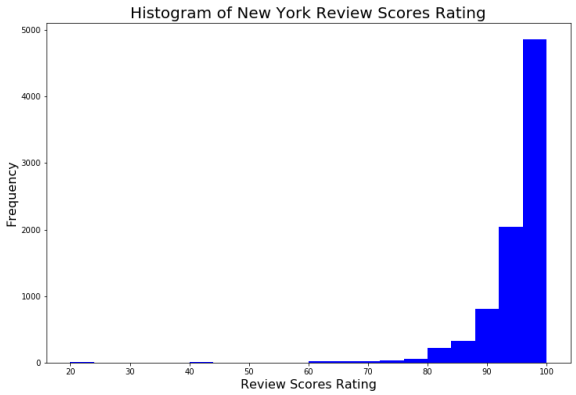

Figure 2. Histogram of New York Review Scores Rating

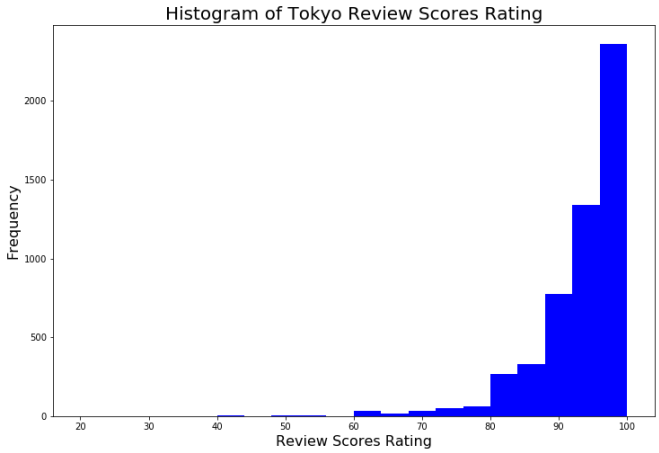

Figure 3. Histogram of Tokyo Review Scores Rating

4.1 Machine Learning Techniques

4.1.1 Preprocessing

To understand the variable correlation of my datasets, I drew heatmaps of the numerical features and the response variables of the three localities respectively (See Appendix 2). The interdependency of different features is not too strong. I have tested removing some features to reduce collinearity and keeping all of them, and it proves out that the performance of the latter is better. So no feature is moved after the collinearity analysis.

For the machine learning algorithms to perform more accurately, I conducted data normalization using standard scaler for numerical variables, deployed one hot encoding method for categorical features. And after tokenizing the house description and host

self-introduction text data into both unigrams and bigrams, I built two pipelines based on the numerical, categorical, and text data. These two pipelines are the same except in the first pipeline, the text data is passed into bag of words function, while in the second pipeline, the text is processed with term frequency–inverse document frequency (TF-IDF) vectorizer. After these pre-processing steps, the three localities respectively have 10051(London), 10052(New York), 10049(Tokyo) features, and 14676(London), 8427(New York), 5288(Tokyo) samples. Finally, I split each dataset into two parts: 80% training set, and 20% test set.

4.1.2. Model Selection

I deployed 4 machine learning models: ridge regression, lasso regression, random forest, and gradient boosting, aiming to take advantage of both linear regression and ensemble methods. Among the 4 models, ridge and lasso regression were applied to both pipelines, whereas random forest and gradient boosting were only applied for the first pipeline, since it proves by the two former models that the first pipeline generates better performance.

a. Linear Regression: Ridge (TFIDF) /Ridge(Bag of Words) / Lasso(TFIDF) / Lasso(Bag of Words):

I adopted these two methods, aiming to approximate how linear combination of all the attributes could predict the rating of Airbnb houses. Since I did not remove features based on heatmaps, ridge and lasso could respectively help me conduct feature selection using l2 and l1 regulation, thus reducing overfitting.

b. Random Forest Regressor and Gradient Boosting

I also used regression tree to solve this problem. Random forest and gradient boosting are used owing to the effectiveness of ensemble methods. Better than standalone regression trees, random forest improves predictive accuracy by finding the best predictors among a random subset of predictors, and gradient boosting outperforms single tree by training predictors sequentially and finally form the best learner out of all the weak learners.

For each model fitting, I tuned the hyperparameters by a randomized search method, with "r2(proportion of variance in the response variable that is explained by the predictors)" as the main scoring metric to judge results of different models. Mean squared error (MSE) and Mean absolute error (MAE) were also calculated as additional references. I optimized these models using 5-fold cross-validation with 10 iterations on the training set. For ridge and lasso, the parameter I optimized is alpha; in random forest model, I optimized the number of estimators, minimum sample split, and minimum sample leaf; in gradient boosting model, I optimized the number of estimators, learning rate, max depth, and alpha.

4.2 Bigram Analysis

I chose bigram analysis because rather than unigram/simple word counts, it allows researchers to investigate the context in which certain words are used. And to make comparative text analysis so as to understand how "successful" Airbnb hosts strategically market themselves, I divide the three datasets respectively into two groups of high review scores home(review_scores_rating=100) and low review scores home (review_scores_rating<=85), and perform bigram analysis on the host description and host self-introduction text of these sub-datasets.

5. Results

5.1 Machine Learning Algorithms Results

Model Name	Training R2	Test R2	MSE	MAE
Ridge1	0.84	0.61	29.1	3.66
Ridge2	0.81	0.66	25.55	3.3
Lasso1	0.68	0.66	25.33	3.2
Lasso2	0.67	0.65	25.86	3.24
Random Forest	0.88	0.68	23.69	2.96
Gradient Boosting	0.84	0.69	23.09	2.94

Table 1. Model Performance of London Airbnb Homes

Model Name	Training R2	Test R2	MSE	MAE
Ridge1	0.74	0.68	12.56	2.39
Ridge2	0.69	0.67	12.73	2.37
Lasso1	0.66	0.65	13.4	2.43
Lasso2	0.65	0.64	14.02	2.47
Random Forest	0.87	0.62	14.84	2.38
Gradient Boosting	0.85	0.66	13.34	2.31

Table 2. Model Performance of New York Airbnb Homes

Model Name	Training R2	Test R2	MSE	MAE
Ridge1	0.74	0.71	14.05	2.63
Ridge2	0.7	0.71	14.24	2.68
Lasso1	0.67	0.7	14.72	2.72
Lasso2	0.67	0.7	14.85	2.73
Random Forest	0.87	0.66	16.4	2.56
Gradient Boosting	0.86	0.69	14.95	2.61

Table 3. Model Performance of Tokyo Airbnb Homes

Table 1-3 display the performance metrics for the models. These six models have very similar performance in terms of r2 values in all three cities. Owing to the high dimensionality of the datasets, random forest and gradient boosting models are usually substantially overfitted as they have very high r2 around 0.9 in the training set, and the metric falls back around 0.7 in the test set. In linear regression models, the overfitting problem is slighter. In the London dataset, though encountering the problem of overfitting, gradient boosting still proves out to be the model performing the best, with the highest test r2 of 0.69, and lowest mean squared error of 23.09 among all models. New York and Tokyo both have ridge regression (bad of words) as the best model, with r2 respectively being 0.68 and 0.71, and mean squared error being 12.56 and 14.05.

5.2 Interpreting the Models

	feature	gradient boosting
0	review_scores_cleanliness	0.477479
1	review_scores_communication	0.313059
2	review_scores_location	0.034324
3	number_of_reviews	0.008099
4	x11_f	0.007872
5	x11_t	0.007456
6	street	0.003941
7	host_acceptance_rate	0.003153
8	x9_t	0.002494
9	area bedroom	0.002468
10	home please	0.002236
11	much offer	0.002031
12	boast large	0.001997
13	home hand	0.001946
14	original	0.001921

Table 4. London Airbnb Homes Feature Importance (gradient boosting)

	feature	ridge(bad of words)
0	review_scores_cleanliness	2.719609
1	review_scores_communication	1.881774
2	review_scores_location	0.874736
3	x11_f	-0.481701
4	x11_t	0.481701
5	flat	-0.373484
6	also give	-0.254097
7	number_of_reviews	-0.223108
8	exquisite	-0.188712
9	floor first	0.188407
10	anywhere manhattan	-0.180656
11	designer	-0.179872
12	kitchen equipped	-0.175243
13	host_acceptance_rate	-0.171852
14	yankee	-0.168897

Table 5. New York Airbnb Homes Coefficient Weights (ridge with bag with words)

	feature	ridge(bad of words)
0	review_scores_cleanliness	3.061630
1	review_scores_communication	1.961779
2	review_scores_location	1.157020
3	x11_f	-0.589857
4	x11_t	0.589857
5	meter	-0.188597
6	x3_Asakusa/Ueno	-0.184333
7	good time	0.183026
8	shared room	0.182400
9	tokyo min	-0.176283
10	place people	-0.168128
11	他可是一个不错的顾问	-0.165896
12	tasting	0.165703
13	narita	-0.162606
14	specialized	-0.161084

Table 6. Tokyo Airbnb Homes Coefficient Weights (ridge with bag of words)

For each locality, I performed feature importance/coefficient weights analysis of the model with the best performance in terms of the r2 value. The results of the three cities have a lot in common, as their top features all suggest review score rating is mainly determined by review scores of cleanliness, communication, and location of the Airbnb home, and among the three common factors, cleanliness is valued most by guests. More interestingly, all three localities imply that being a super-host, as indicated by x11_t, significantly helps promote review scores. In terms of other crucial contributing factors, London and New York have much in common, where host acceptance rate and the number of reviews have high rankings, shedding light on what aspects guests pay the most attention to when they choose houses. These two cities also display differences, as New York guests show dislike for certain types of homes, with words indicating special features, such as “exquisite”, “designer” and “yankee” all negatively correlated with review score rating. Distinctive from the two localities, Tokyo guests emphasize more on “location” of Airbnb homes. Certain locations, including “Narita” and “Asakusa/Ueno”, are negatively affecting review scores, indicating guests might prefer to live in other districts.

Since all the three models, especially the gradient boosting model for London Airbnb homes, suffer from overfitting problem that might reduce their reliability, I supplement the current conclusion by analyzing other models(See Appendix 3,4,5). The

results again reveal that review scores of cleanliness, communication, and location are the most influencing features. This time, more conspicuous differences among the three localities emerge. London guests highlight the importance of house amenities and neighborhood around by mentioning “beautiful kitchen”, “modern room”, “furnished room”, and “lovely garden”, etc., and do not value host attributes that much. In New York, the host attributes gather the most attention. Time as host and host acceptance rate rank high in almost every model. In Tokyo, “location” still proves to be the most vital concern, with one group of districts, such as “Shibuya” negatively correlated with the review score, while the other group of districts, such as “Ginza” positively associated with the review score, exhibiting guests' preference regarding where to stay when they are in Tokyo.

5.3 Text Analysis Results

The top 10 bi-gram comparison between high-score houses and low-score houses across the three cities (See Appendix 6,7) indicates the way “successful” and “failed” hosts position themselves are not substantially different. In terms of house/home description, all hosts addressed the equipment of the houses and convenience of the location. In New York, the high-score homes seem to be more “newly renovated”, so that might be a reason guests are in favor of them. In Tokyo, there appears to be an obvious difference that high-score houses are near “Shinjuku station” while the low-score homes are near “ Ikebukuro station”, corresponding with previous machine learning algorithm results that “location” is an important discriminating factor in Tokyo.

The host self-introduction proves to be quite similar between high-score hosts and low-score hosts, as the top bigrams in both groups suggest a sense of wish/welcome message sending by adopting phrases like “feel free” and “look forward.” This is most applicable to Tokyo hosts as there is no noticeable difference between high and low-score Tokyo hosts’ self-introduction. The other two cities, on the other hand, show that high-score hosts would address their own hobbies (especially the love for “traveling”) in the self-introduction to arouse resonance with the guests, whereas low- score hosts focus more on neutrally introducing the amenities and neighborhood of their houses,

using phrases including “walking distance” and “subway ride” without the sense of enthusiasm radiating from effective self-introduction.

6. Conclusions

Methodologically, my research corresponds to the call for deploying social media data (Zhou et al., 2014) in social sciences research. Theoretically, this paper contributes to the Airbnb literature through a comprehensive and detailed understanding of what influences Airbnb guests’ satisfaction level. I build six machine learning models and use bigram analysis to predict Airbnb home rating scores based on property and host attributes, together with the text description of them. This research is innovative in incorporating more than 10000 attributes in the models for a thorough analysis, and use three localities to test the generalizability of predictor results(Brochado et al., 2017). From the above result analysis, it is concluded that the review score of cleanliness, location, and communication, and being a super-host stand out to be the most critical predictors across regions. And the three cities exhibit essential differences in terms of the order of importance of key predictors. London Airbnb guests focus on home attributes, such as facilities and surroundings; New York guests put more emphasis on host attributes, such as host response rate and acceptance rate. Tokyo guests, on the other hand, highlight the importance of the location of houses, as some certain districts, such as Ginza, receive higher review scores while others do not. Interestingly, the three cities respectively contribute to different aspects of the "amenities, host and location" debate regarding the main contributing factors to review scores(Cheng & Jin, 2019), and the geographical difference could possibly be the explanation.

The bigram analysis points out that Airbnb hosts generally passionately sending welcome messages to guests, and high-rating hosts would also strategically introduce their own background and hobbies to arouse resonance in the guests, so as to increase their review score. This corresponds to the high ranking of tourist-host communication in the machine learning models, and might also be a reason to partially explain the positivity bias in Airbnb rating (Bridges and Vásquez, 2016), since the hosts work as

traveling experiences facilitators to increase guests' satisfaction level, and use their welcome message to create a home environment, which is a competitive advantage against traditional traveling industry. Another important finding from the bigram analysis is the importance of location in Tokyo, again confirming the findings from the machine learning models.

In conclusion, we gain unique data-driven insight from the study about what factors are affecting Airbnb guests' staying experiences. It offers a useful guideline for Airbnb hosts to condition attributes identified in this research to high review-score, and for Airbnb to incentivize its hosts to modify certain features, improve its service quality, and boost business growth accordingly.

7. Limitations and Future Work

My research still suffers from several limitations that warrant future exploration. First, since my result indicates the importance of locality when analyzing the influencing factors of the models, I could take a step further to comparatively analyze datasets from different, or even cross-cultural regions that might yield new sights into how local culture/context would influence guests' preference for staying, adding to the literature regarding the convergence/divergence of Airbnb guests' traveling experiences (Brochado et al., 2017). Second, it could be a useful step forward to adopt time series analysis if I could get Airbnb listing information in previous years. Then I could analyze how the importance ranking of Airbnb features has changed over the years based on the enriched datasets to infer how Airbnb users' concentration has transformed. Finally, the Tokyo dataset contains a substantial amount of text information written in Chinese or Korean. To translate them into English for further analysis and to understand why there is such a language issue particularly in Tokyo would be of great significance in enhancing the research findings.

8. Acknowledgement

Part of my code was modified from the project of review number prediction from https://github.com/activerabbit/Forecasting-Review-Counts-for-Airbnb-Hosts. This project inspired me to incorporate text information into machine learning algorithms.

Thanks for all guidance from Dr. Evans in this project. His insightful ideas for better visualization of the research results helped me a lot in polishing the final paper.

References

Aiken, K. D., & Boush, D. M. (2006). Trustmarks, objective-source ratings, and implied investments in advertising: investigating online trust and the context-specific nature of internet signals. *Journal of the Academy of Marketing Science*, *34*(3), 308-323.

Airbnb (2020). About us. Retrieved from https://www.airbnb.com.au/about/about-us.

Bridges, J., & Vásquez, C. (2018). If nearly all Airbnb reviews are positive, does that make them meaningless?. *Current Issues in Tourism*, *21*(18), 2057-2075.

Brochado, A., Troilo, M., & Aditya, S. (2017). Airbnb customer experience: evidence of convergence across three countries. *Annals of Tourism Research*, *63*, 210-212.

Cheng, M., & Edwards, D. (2019). A comparative automated content analysis approach on the review of the sharing economy discourse in tourism and hospitality. *Current Issues in Tourism*, *22*(1), 35-49.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58-70.

Edwards, D., Cheng, M., Wong, I. A., Zhang, J., & Wu, Q. (2017). Ambassadors of knowledge sharing. *International Journal of Contemporary Hospitality Management*.

Festila, M., & Müller, S. (2017). The impact of technology-mediated consumption on identity: The case of Airbnb. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *arXiv preprint arXiv:1907.12665*.

Ma, X., Hancock, J. T., Lim Mingjie, K., & Naaman, M. (2017). Self-disclosure and perceived trustworthiness of Airbnb host profiles. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 2397-2409).

McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review*, *27*(2), 277-290.

Piryani, R., Madhavi, D., & Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, *53*(1), 122-150.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *science*, *311*(5762), 854-856.

Tsai, H. T., & Huang, H. C. (2009). Online consumer loyalty: Why e-tailers should seek a high-profile leadership position. *Computers in Human Behavior*, *25*(6), 1231-1240.

Tussyadiah, I. P., & Zach, F. (2017). Identifying salient attributes of peer-to-peer accommodation experience. *Journal of Travel & Tourism Marketing*, *34*(5), 636-652.

Xie, K., & Mao, Z. (2017). The impacts of quality and quantity attributes of Airbnb hosts on listing performance. *International Journal of Contemporary Hospitality Management*.

Zervas, G., Proserpio, D., & Byers, J. (2015). A first look at online reputation on Airbnb, where every stay is above average. *Where Every Stay is Above Average (January 28, 2015)*.

Zhang, L., Yan, Q., & Zhang, L. (2018). A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb. *Decision Support Systems*, *115*, 105-116.

Zmud, R. W., Shaft, T., Zheng, W., & Croes, H. (2010). Systematic differences in firm's information technology signaling: implications for research design. *Journal of the Association for Information Systems*, *11*(3), 1.

Appendix1: Attributes used in analysis

Numerical: 'accommodates', 'bathrooms', 'beds', 'bedrooms', 'price', 'guests_included', 'extra_people','minimum_nights', 'maximum_nights', 'number_of_reviews', 'review_scores_location', 'review_scores_cleanliness', 'review_scores_communication', 'time_as_host', 'host_response_rate','host_acceptance_rate'

Categorical: 'transit', 'host_has_profile_pic', 'host_identity_verified', 'neighbourhood', 'property_type', 'room_type', 'bed_type', 'security_deposit', 'cleaning_fee', 'instant_bookable', 'weekly_price', 'host_is_superhost', 'host_response_time' and the list of amenities in the corresponding locality

Text: 'self_about', 'description'

Response Variable: 'review_scores_rating'

Appendix 2.Heatmaps

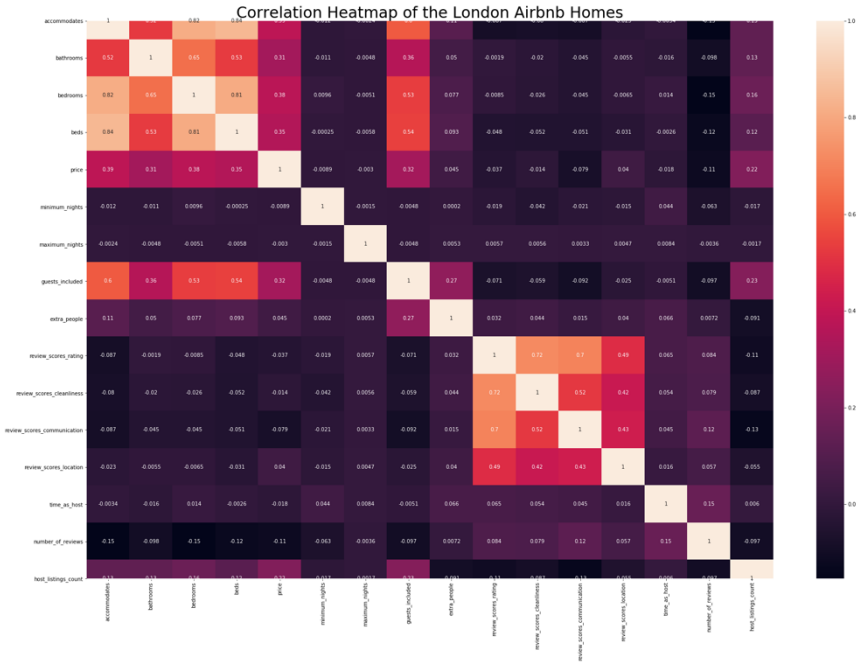

Figure1. Heatmap of London Airbnb Homes

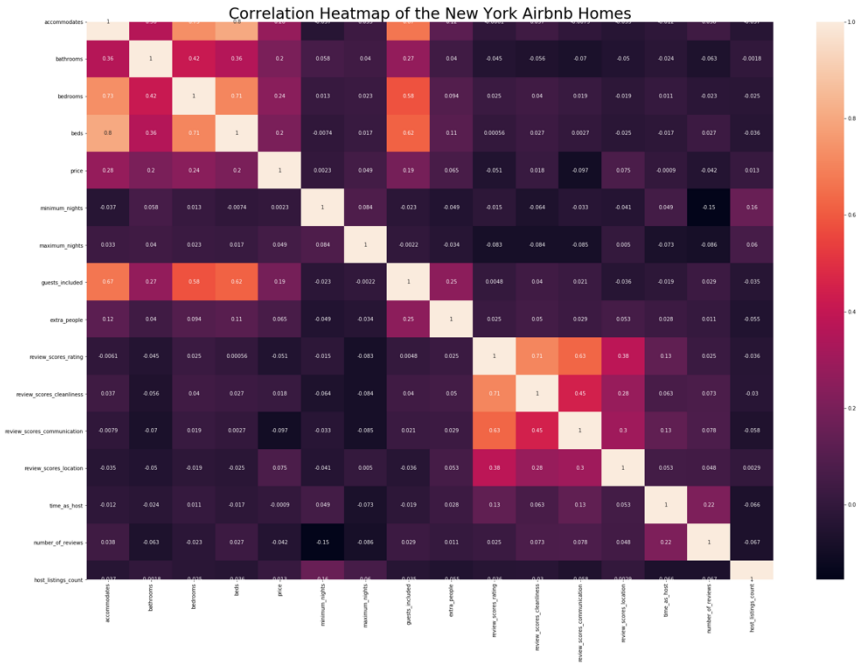

Figure 2. Heatmap of New York Airbnb homes

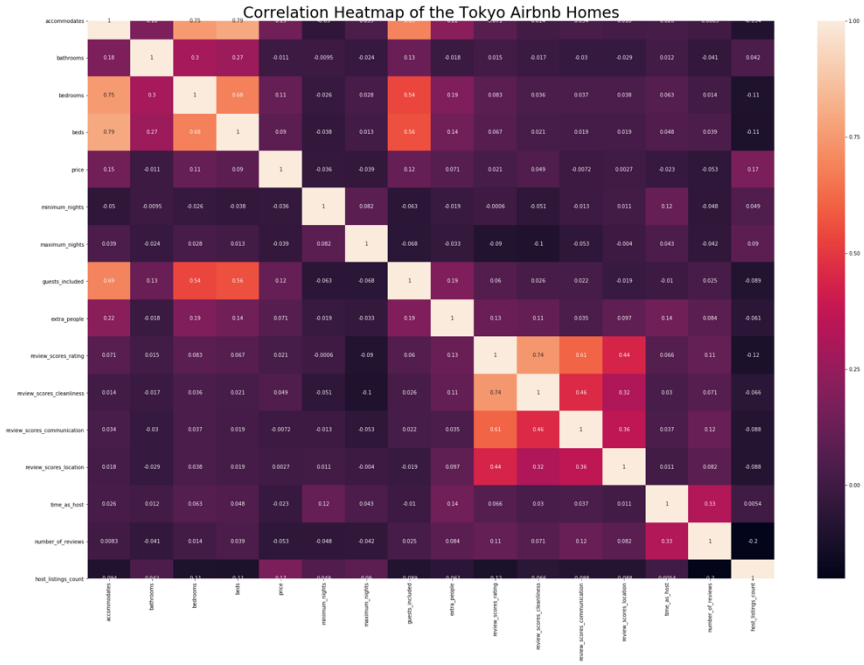

Figure 3. Heatmap of Tokyo Airbnb homes

Appendix 3. Coefficient weights/Feature importance tables of London

	feature	ridge(bad of words)
0	review_scores_cleanliness	3.484939
1	review_scores_communication	3.130619
2	kitchen beautiful	1.626500
3	flat perfectly	-1.536597
4	hob oven	-1.499507
5	walk camden	-1.479074
6	room modern	-1.450935
7	government	-1.410651
8	cost	-1.405379
9	share flat	-1.303574
10	back soon	-1.280412
11	review_scores_location	1.256319
12	furnished bedroom	1.252755
13	much offer	-1.239625
14	ha always	-1.229664

Table 1. London Airbnb Homes Coefficient Weights Table (ridge with bag of words)

	feature	ridge(TFIDF)
0	walk camden	-10.560887
1	government	-9.953712
2	hob oven	-9.893350
3	kitchen beautiful	8.854857
4	professional couple	-8.787529
5	back soon	-8.504729
6	cost	-8.412489
7	bathroom shared	-8.149296
8	hospital	-8.007084
9	queen size	-7.825816
10	ha always	-7.720148
11	able share	-7.655592
12	large comfortable	-7.636694
13	garden lovely	-7.563018
14	warmest regard	-7.400118

Table 2. London Airbnb Homes Coefficient Weights Table (ridge with TF-IDF)

	feature	lasso(bag of words)
0	review_scores_cleanliness	3.725473
1	review_scores_communication	3.242501
2	review_scores_location	0.860570
3	street	-0.279452
4	accommodates	-0.000000
5	bathrooms	0.000000
6	beds	0.000000
7	bedrooms	0.000000
8	price	0.000000
9	guests_included	0.000000
10	extra_people	0.000000
11	minimum_nights	0.000000
12	maximum_nights	0.000000
13	number_of_reviews	-0.000000
14	time_as_host	0.000000

Table 3. London Airbnb Homes Coefficient Weights Table (lasso with bag of words)

	feature	lasso(TFIDF)
0	review_scores_cleanliness	3.735104
1	review_scores_communication	3.382070
2	review_scores_location	0.820763
3	accommodates	-0.000000
4	bathrooms	0.000000
5	beds	-0.000000
6	bedrooms	0.000000
7	price	-0.000000
8	guests_included	-0.000000
9	extra_people	0.000000
10	minimum_nights	0.000000
11	maximum_nights	0.000000
12	number_of_reviews	-0.000000
13	time_as_host	0.000000
14	host_response_rate	0.000000

Table 4. London Airbnb Homes Coefficient Weights Table (lasso with TF-IDF)

	feature	random forest
0	review_scores_cleanliness	0.468316
1	review_scores_communication	0.271681
2	number_of_reviews	0.017424
3	review_scores_location	0.008488
4	x11_t	0.006655
5	host_acceptance_rate	0.006124
6	x11_f	0.004665
7	time_as_host	0.004613
8	price	0.004462
9	street	0.002932
10	maximum_nights	0.001605
11	ballet	0.001577
12	space ideal	0.001556
13	beds	0.001533
14	detached	0.001514

Table 5. London Airbnb Homes Feature Importance Table (random forest)

	feature	gradient boosting
0	review_scores_cleanliness	0.477479
1	review_scores_communication	0.313059
2	review_scores_location	0.034324
3	number_of_reviews	0.008099
4	x11_f	0.007872
5	x11_t	0.007456
6	street	0.003941
7	host_acceptance_rate	0.003153
8	x9_t	0.002494
9	area bedroom	0.002468
10	home please	0.002236
11	much offer	0.002031
12	boast large	0.001997
13	home hand	0.001946
14	original	0.001921

Table 6. London Airbnb Homes Feature Importance Table (gradient boosting)

Appendix 4. Coefficient weights /Feature importance tables of New York

	feature	ridge(bad of words)
0	review_scores_cleanliness	2.719609
1	review_scores_communication	1.881774
2	review_scores_location	0.874736
3	x11_f	-0.481701
4	x11_t	0.481701
5	flat	-0.373484
6	also give	-0.254097
7	number_of_reviews	-0.223108
8	exquisite	-0.188712
9	floor first	0.188407
10	anywhere manhattan	-0.180656
11	designer	-0.179872
12	kitchen equipped	-0.175243
13	host_acceptance_rate	-0.171852
14	yankee	-0.168897

Table 1. New York Airbnb Homes Coefficient Weights Table (ridge with bag of words)

	feature	ridge(TFIDF)
0	review_scores_cleanliness	3.023066
1	review_scores_communication	2.012767
2	flat	-1.306692
3	williamsburg jfk	-0.953300
4	exquisite	-0.939776
5	x5_Private room	0.933641
6	globe	-0.912901
7	review_scores_location	0.844676
8	also give	-0.830433
9	quiet comfortable	-0.826645
10	radio	-0.819729
11	fun fulfilling	0.801051
12	designer	-0.798140
13	stunning	-0.779048
14	listen	0.722349

Table 2. New York Airbnb Homes Coefficient Weights Table (ridge with TF-IDF)

	feature	lasso(bag of words)
0	review_scores_cleanliness	3.026230
1	review_scores_communication	2.028487
2	review_scores_location	0.763526
3	x11_f	-0.603290
4	flat	-0.329374
5	host_acceptance_rate	-0.199723
6	number_of_reviews	-0.180043
7	exquisite	-0.159862
8	globe	-0.143259
9	rare find	-0.134709
10	stunning	-0.127870
11	williamsburg jfk	-0.123344
12	time_as_host	0.112196
13	fun fulfilling	0.085956
14	importantly	0.069753

Table 3. New York Airbnb Homes Coefficient Weights Table (lasso with bag of words)

	feature	lasso(TFIDF)
0	review_scores_cleanliness	3.073856e+00
1	review_scores_communication	2.064565e+00
2	review_scores_location	7.147882e-01
3	x11_f	-6.423078e-01
4	host_acceptance_rate	-2.268486e-01
5	number_of_reviews	-1.797785e-01
6	time_as_host	1.441641e-01
7	price	-2.451104e-02
8	minimum_nights	2.121193e-02
9	x11_t	4.745758e-16
10	accommodates	-0.000000e+00
11	bathrooms	0.000000e+00
12	beds	-0.000000e+00
13	bedrooms	-0.000000e+00
14	guests_included	-0.000000e+00

Table 4. New York Airbnb Homes Coefficient Weights Table (lasso with TF-IDF)

	feature	random forest
0	review_scores_cleanliness	0.558610
1	review_scores_communication	0.149543
2	number_of_reviews	0.028101
3	review_scores_location	0.011344
4	time_as_host	0.008245
5	x11_f	0.007723
6	x11_t	0.006889
7	host_acceptance_rate	0.006522
8	price	0.005571
9	exquisite	0.004042
10	africa	0.003964
11	extra_people	0.003691
12	cheer	0.003606
13	williamsburg jfk	0.003526
14	fun fulfilling	0.002999

Table 5. New York Airbnb Homes Feature Importance Table (random forest)

	feature	gradient boosting
0	review_scores_cleanliness	0.595844
1	review_scores_communication	0.150562
2	review_scores_location	0.022379
3	number_of_reviews	0.019633
4	ac heat	0.012097
5	x11_f	0.011973
6	x11_t	0.009203
7	time_as_host	0.004842
8	host_acceptance_rate	0.004823
9	price	0.004626
10	minimum_nights	0.003629
11	microwave coffee	0.002767
12	tourism	0.002147
13	alumnus one	0.002036
14	travel around	0.001985

Table 6. New York Airbnb Homes Feature Importance Table (gradient boosting)

Appendix 5. Coefficient weights/Feature importance tables of Tokyo

	feature	ridge(bad of words)
0	review_scores_cleanliness	3.061630
1	review_scores_communication	1.961779
2	review_scores_location	1.157020
3	x11_f	-0.589857
4	x11_t	0.589857
5	meter	-0.188597
6	x3_Asakusa/Ueno	-0.184333
7	good time	0.183026
8	shared room	0.182400
9	tokyo min	-0.176283
10	place people	-0.168128
11	他可是一个不错的顾问	-0.165896
12	tasting	0.165703
13	narita	-0.162606
14	specialized	-0.161084

Table 1. Tokyo Airbnb Homes Coefficient Weights Table (ridge with bag of words)

	feature	ridge(TFIDF)
0	review_scores_cleanliness	3.665994
1	review_scores_communication	2.133918
2	review_scores_location	1.077961
3	x3_Asakusa/Ueno	-0.872699
4	takeko	-0.756255
5	x11_t	0.715077
6	x11_f	-0.715077
7	great length	-0.691162
8	여행하게된	0.638326
9	super	0.625475
10	x3_Shibuya District	-0.617687
11	thing give	-0.602274
12	situation hesitate	0.595667
13	place people	-0.590912
14	ginza tokyo	0.582963

Table 2. Tokyo Airbnb Homes Coefficient Weights Table (ridge with bag of TF-IDF)

	feature	lasso(bag of words)
0	review_scores_cleanliness	3.746943
1	review_scores_communication	2.146220
2	x11_f	-1.065800
3	review_scores_location	0.966697
4	accommodates	0.208355
5	bathrooms	0.105436
6	time_as_host	0.100183
7	time one	-0.096751
8	golden gai	0.079355
9	exclusive	0.077824
10	extra_people	0.076730
11	여행하게된	0.063049
12	host_acceptance_rate	-0.062873
13	land liberty	0.056137
14	super	0.053305

Table 3. Tokyo Airbnb Homes Coefficient Weights Table (lasso with bag of words)

	feature	lasso(TFIDF)
0	review_scores_cleanliness	3.762771
1	review_scores_communication	2.153907
2	x11_f	-1.102462
3	review_scores_location	0.961990
4	accommodates	0.239803
5	time_as_host	0.133888
6	bathrooms	0.109817
7	extra_people	0.089086
8	host_acceptance_rate	-0.080032
9	minimum_nights	0.057841
10	maximum_nights	-0.043366
11	guests_included	0.031561
12	x13_0	-0.031250
13	x15_0	0.021800
14	beds	0.002221

Table 4. London Airbnb Homes Coefficient Weights Table (lasso with TF-IDF)

	feature	random forest
0	review_scores_cleanliness	0.594552
1	review_scores_communication	0.095636
2	number_of_reviews	0.035284
3	review_scores_location	0.018731
4	x11_f	0.009893
5	x11_t	0.008377
6	price	0.008282
7	time_as_host	0.005673
8	accommodates	0.004267
9	extra_people	0.003839
10	minimum_nights	0.003453
11	shop around	0.003214
12	maximum_nights	0.002717
13	soon possible	0.002599
14	beds	0.002495

Table 5. Tokyo Airbnb Homes Feature Importance Table (random forest)

	feature	gradient boosting
0	review_scores_cleanliness	0.609788
1	review_scores_communication	0.122613
2	review_scores_location	0.038822
3	number_of_reviews	0.024917
4	x11_t	0.013624
5	x11_f	0.012151
6	yen	0.005134
7	time_as_host	0.004088
8	price	0.003710
9	directly connected	0.003631
10	minimum_nights	0.002522
11	accommodates	0.002401
12	accommodation	0.002401
13	theme	0.002207
14	sofa living	0.002190

Table 6. Tokyo Airbnb Homes Feature Importance Table (gradient boosting)

Appendix 6. Bigram of house description

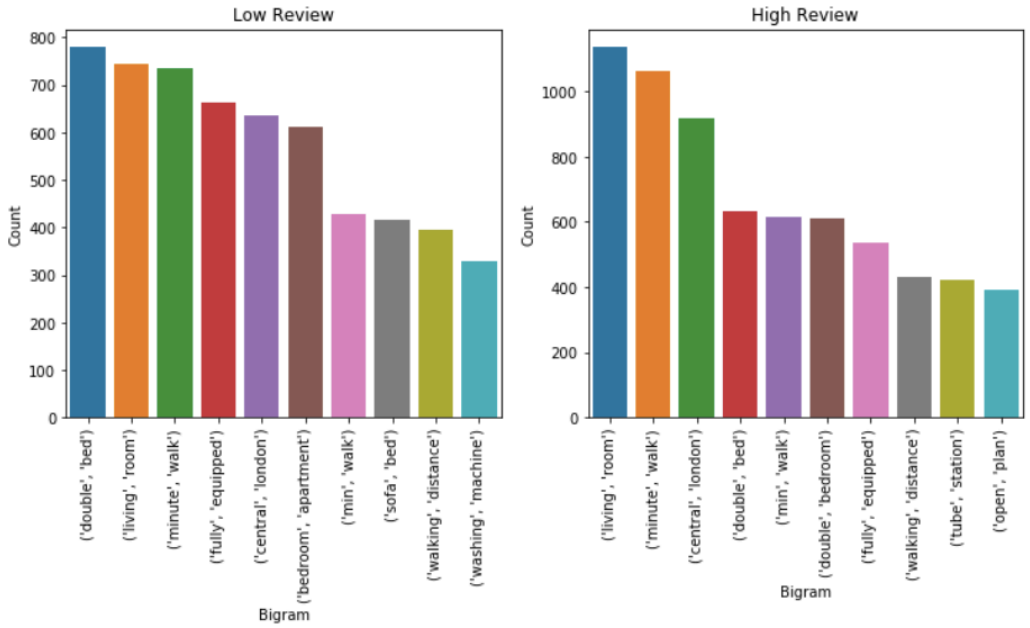

Figure 1. London House Description Bi-gram of Low-score and High-score Homes

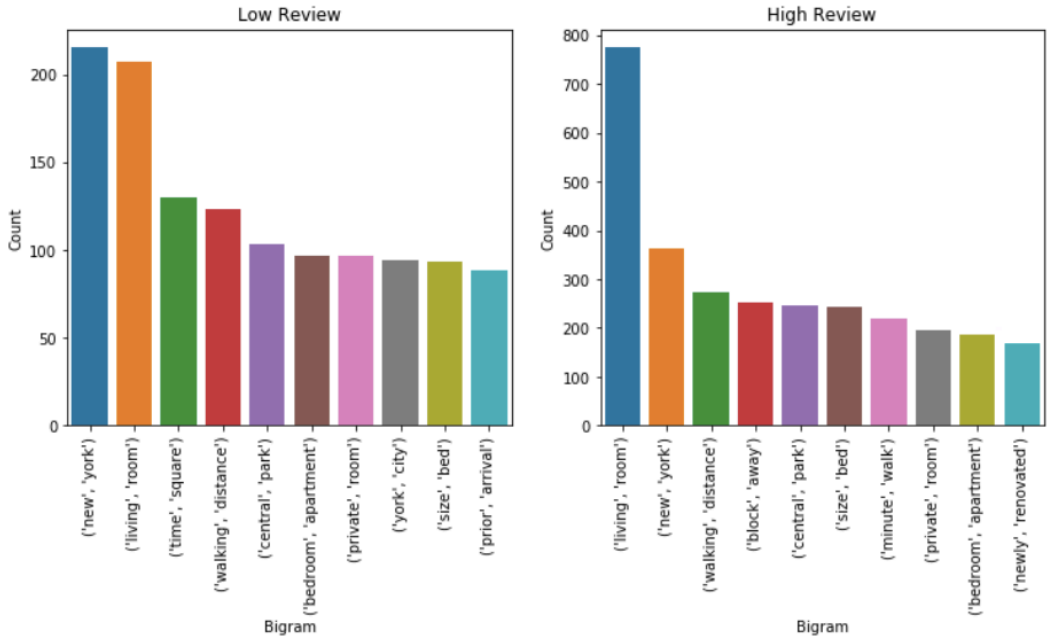

Figure 2. New York House Description Bi-gram of Low-score and High-score Homes

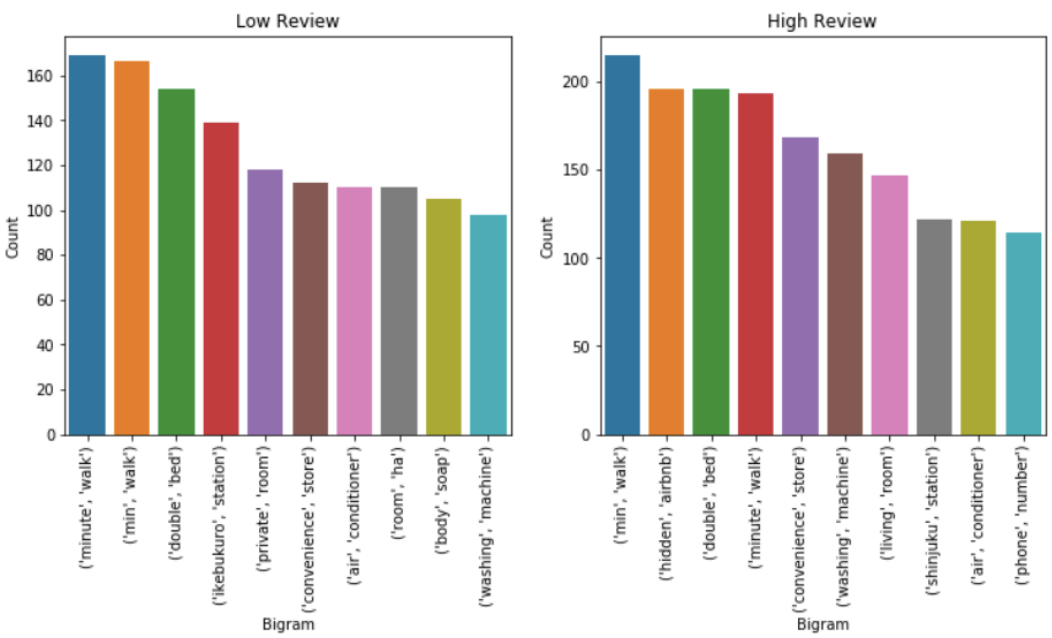

Figure 3. Tokyo House Description Bi-gram of Low-score and High-score Homes

Appendix 7. Bigram of host self-introduction

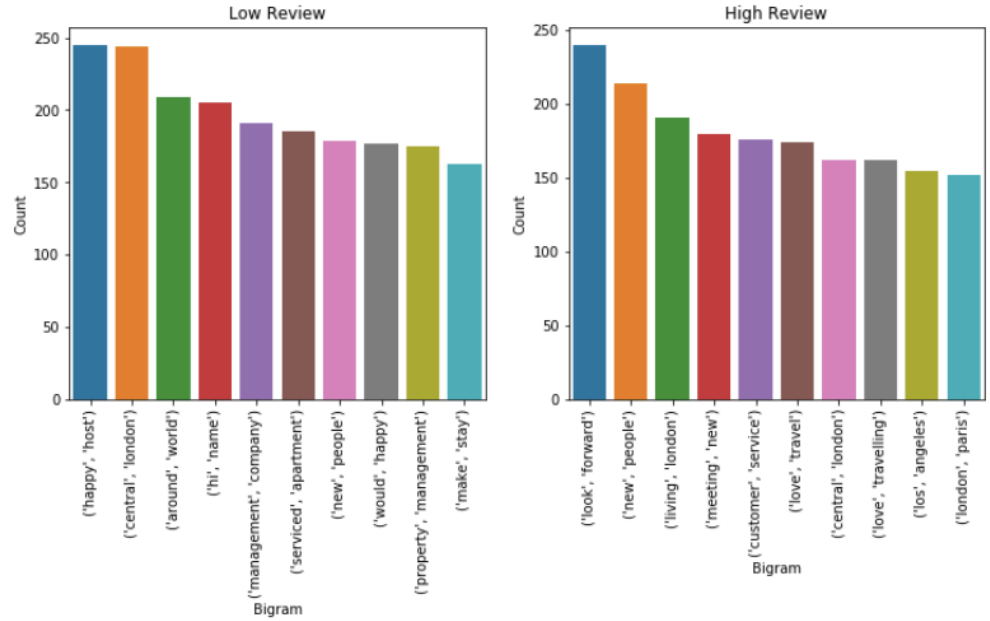

Figure 1. London Host-introduction Bi-gram of Low-score and High-score Homes

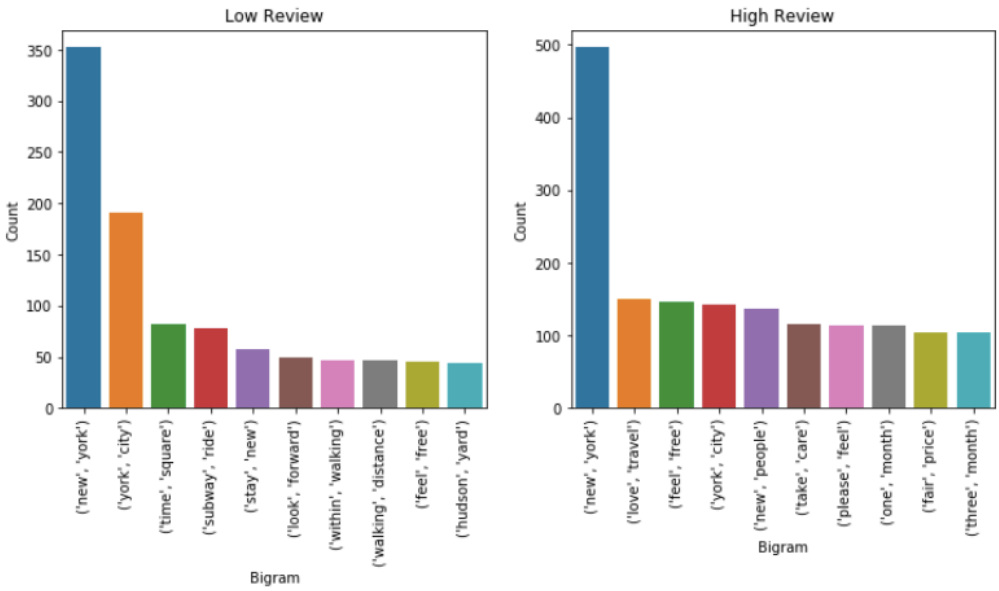

Figure 2. New York Host-introduction Bi-gram of Low-score and High-score Homes

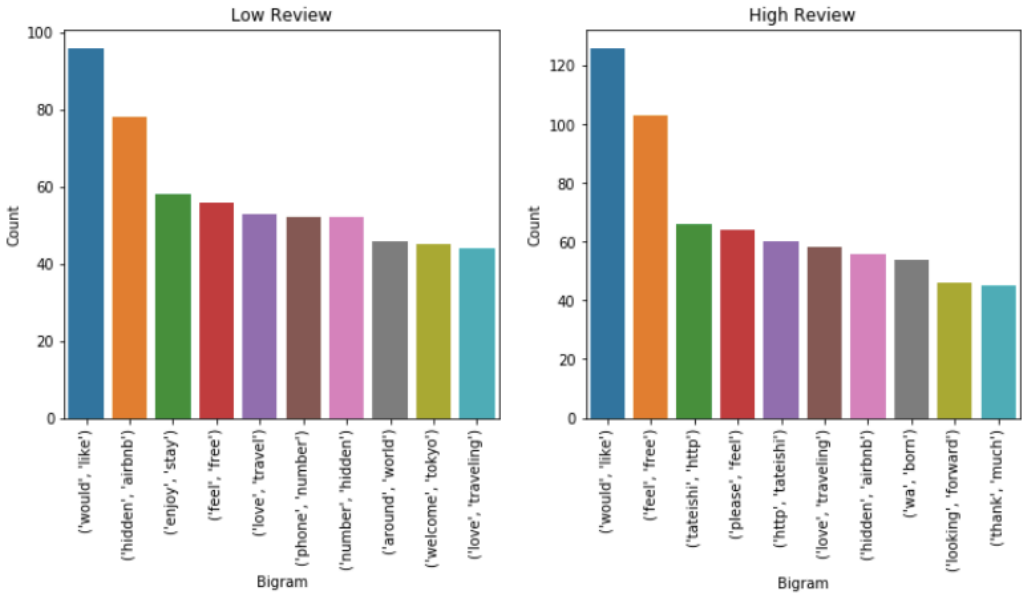

Figure 3. Tokyo Host-introduction Bi-gram of Low-score and High-score Homes