Data Science Final Report: Predicting Airbnb Ratings

Section B - Team 36
Maggie Lundberg, Tingche Lyu, Niyati Mody, Vladislav Stanojevic, Kala Nandini Uppala

Statement of the Business Problem:

Airbnb is an online marketplace that started as a California startup in 2008. The purpose of Airbnb is to provide an alternative to hotels that feels more like home. Airbnb, which stands for "Air Bed and Breakfast," utilizes the internet to advertise affordable, safe places, from entire houses to single rooms, to stay in any city one could think of.

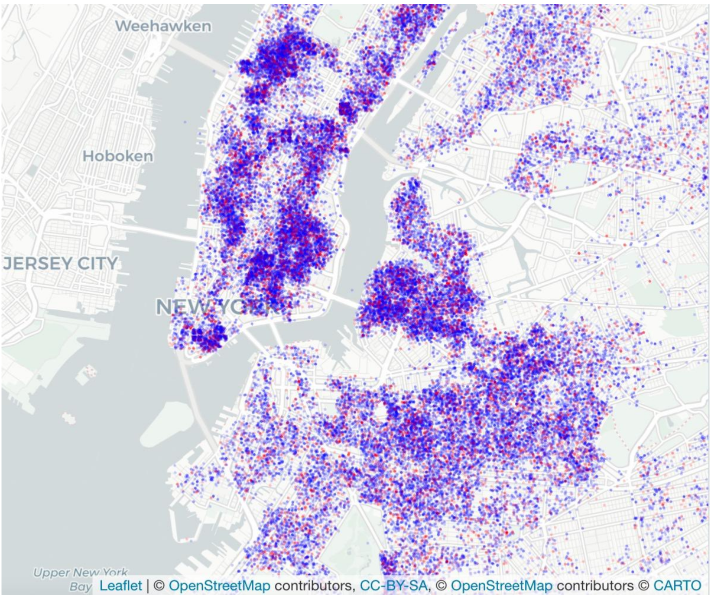
New York City is one of the most popular tourist destinations in the world, so Airbnb thrives in this area. However, due to recent laws imposed by New York City, Airbnb has a more challenging time generating revenue than in previous years.[1] In 2020, there was a large increase in short-term illegal rentals in New York City, leading to an increased housing crisis in New York and driving up rent for New York City residents. In 2021, the New York City Council approved a new bill requiring hosts to register with the city before making money off renting out their homes on a short-term basis, meaning less than thirty days. This proved to be a major obstacle for Airbnb because New York City was one of their largest domestic markets. According to Airbnb, there were over 37,700 Airbnb listings in New York City as of November 2021. This meant if not officially registered with the city, rentals would not appear online. Additionally, the fines for hosts who did not follow the city's new rule would be up to $5,000, and Airbnb would be fined $1,500 for every illegal transaction.[1]

This dataset focuses on Airbnb providers in New York City specifically. We would like to use the data from this dataset to understand what makes a New York City Airbnb listing good based on customer satisfaction. Our measure of a good Airbnb listing is customer ratings from reviews. We will use rental descriptors like New York City neighborhood, rental price, minimum nights stay, year constructed, room type, cancellation policy, and more. Through data visualization, clustering, and predictive modeling, we can help Airbnb understand what makes a positive customer experience. From this, we can give Airbnb recommendations for how hosts can improve guest satisfaction, and Airbnb can, in turn, recommend listings to advertise at the top of their website based on criteria we find to be the most important for a positive Airbnb experience. Our analysis for Airbnb, as a multi-sided platform, will be beneficial through predicting the rating, which is highly relevant to customer satisfaction and hosts' revenue. By performing this analysis, we can help improve Airbnb's reputation and increase its competitive edge over hotels in New York City. Additionally, this could help Airbnb bounce back from New York City's recent imposition of regulatory laws.

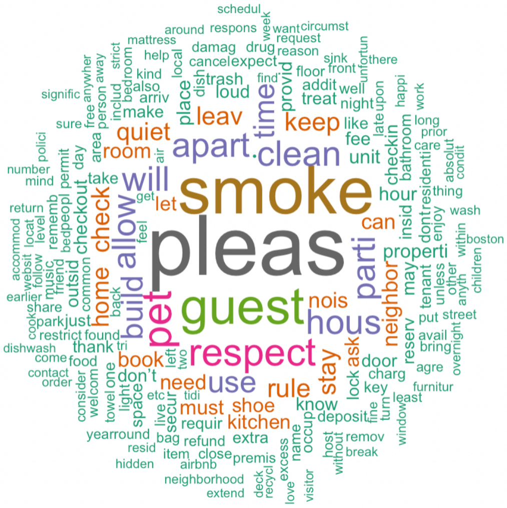
Data Understanding and Preparation:

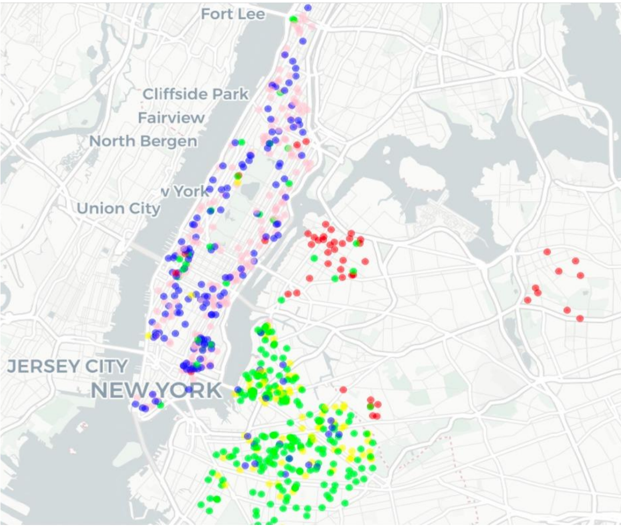
We took the dataset from Kaggle, containing over 100,000 observations and 25 variables. To clean our dataset, we started by changing all blank cells to missing values and deleting rows containing missing values in our target variable. Next, we checked for typos. A column called neighborhood groups had several entries with incorrect spellings (e.g., "brookln" instead of Brooklyn). The last_review column had one review dated in the year 2058, and the columns minimum_nights and availability had similar typing errors too. We then noticed that several column types needed to be changed to numerical ones, including price, construction_year, and service_fee. There were missing values in the column neighborhood_groups, but surprisingly not many missing values in the column neighborhood. Therefore, we built a dictionary to match the two. Following that, we identified several foreign names in the listing column, including Korean, Chinese, etc. To address that, we found a package called "cld2," using which we detected the language in these names and created several dummies for the different languages. Finally, we replaced all missing numeric values with their median. For other categorical missing values, we replaced them with the most frequent ones in each column.

Through exploratory data analysis, we attempted to uncover the relationships between some variables of interest in the data set. We looked at the relationship between the amount of Airbnb listings based on construction, the price of Airbnb listings based on the type of room, the total amount of Airbnb listings by price split by rating, the proportion of rental type by rating, and the total amount of Airbnb listings by neighborhood group split by price. All of these visualizations can be seen in the appendix. These visualizations gave us an idea of the distribution of our data (Appendix A: A-F). Additionally, the location of each Airbnb listing was visualized on a plot of New York City to show where the listings are spatially. To make these visualizations, we use the latitude and longitude variables, which allow us to determine the precise location of the listing. The listings rated high, five stars, are in red, and those not rated five stars are in blue.


Weehawken
Hoboken
JERSEY CITY
NEW YORK
Upper New York Bay
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap contributors © CARTO

We also created a word cloud to see the most frequent words used in “house rules” for guests staying at Airbnb. Most hosts’ policies mention their rules regarding smoking and pets.





Additionally, we performed a hierarchical clustering analysis of 728 samples from the Airbnb listings, and this analysis gave us five distinct clusters. It's important to understand that these 728 samples, on some level, represent the whole dataset. Given these 728 samples, we calculated gower's distance and implemented a hierarchical clustering for mixed data types (Appendix A: G). To ensure the interpretability of our unsupervised learning, we chose to keep five clusters. The red cluster consists primarily of listings in the Brox, the green cluster consists mainly of listings in Brooklyn that are higher prices and more instantly bookable, the yellow cluster consists of listings in Brooklyn that are lower prices and less instantly bookable, the blue cluster consists primarily of listings in Manhattan that are more likely to be booked, and the pink set consists of listings in Manhattan that are hard to be instant booked (Appendix A: H-J). We visualized our distinct clusters on the map of New York City again.

Modeling:

We first categorized reviews into good and bad based on the rating. In the data set, we distinguished five stars as good and four stars or below as bad. We then splitted the data into a testing and training set. We performed the following methods: logistic regression, lasso regression, post-lasso regression, classification tree, and random forest.

We started our Modeling process by converting our outcome variable, rating, into a binomial form, assigning 1 for ratings with five stars and 0 for ratings with less than five stars. Given the size of our cleaned dataset, we resampled the dataset to save time and for R not to crash. Specifically, we took 40% as the sample while keeping the distribution of each column the same. Then we divided the data into 75% train and 25% test, ensuring that most of the neighborhoods were kept in both sets. However, since the neighborhoods feature, as a categorical variable, had many different values, it would introduce new levels of a factor by coercion in cross validation. Therefore, we had to drop the column when building the model. We calculated the mean of our target variable in both data sets, and those values are similar.

Since our outcome is binomial, the first model we tried was a logistic regression. We expected a logistic model to be a better fit than the linear regression because the result would range from 0 to 1. We also wanted to implement lasso and post-lasso as well. We believed that the real-world dataset was noisy, and penalization on coefficients might affect the performance. Similarly, we considered the classification tree, and random forest candidates that potentially gave more accurate results.

We prepared the data for k-fold cross validation. In the process of dropping variates for lasso and post-lasso, we observed that the deviance is almost the same for different values of log(lambda) (see Appendix A: K). With regards to the optimal lambda value, theoretically, for a binomial, the valid choice is using lambda theory. However, using lasso theory did not work when using this value because it would take zero coefficients. Similarly, neither lasso/post lasso one standard error nor lasso/post lasso minimum worked. In our case, using regularization to drop all the variates was not optimal either. Even so, we kept all these models to compare their accuracies, but we expected random forest to be the best candidate.

We then performed logistic regression, lasso regression with lambda min, lasso regression with lambda one standard error, lasso theory, random forest, and classification tree. With these models, we can determine which variables are most significant in predicting whether an Airbnb would be ranked high or low, and identify any trends or patterns in the same. By doing so, Airbnb can keep track of the listings customers love and provide benefits to owners of high-rated rentals, such as a subsidy to get registered in NYC.

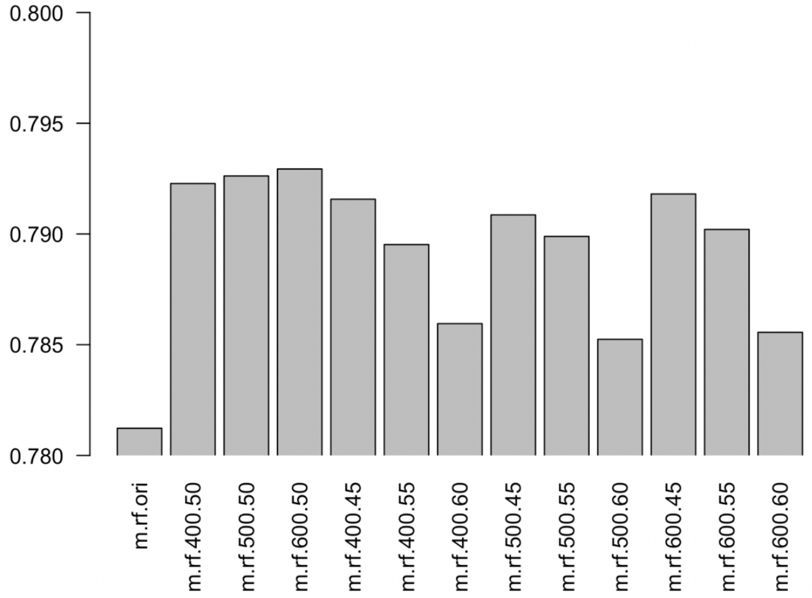
Evaluation:

To compare the performance of each model, we carried out K-fold cross-validation to compute out-of-sample (OOS) accuracy to evaluate the predictive capabilities of each model. From the business side of the problem, we want the prediction to be as precise as possible, which makes accuracy a proper performance metric over all other indicators.

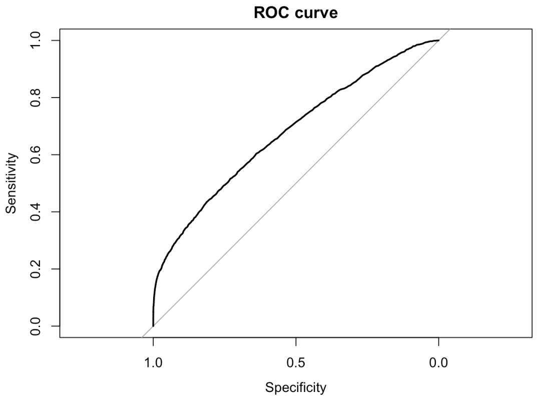
In logistic regression, we found the OOS accuracy for this model to be 0.7763192. For lasso and post lasso, we found the accuracy score similar in the cases of minimum, one standard deviation, and theory, at 0.7753281. As mentioned above, they are not good models in our case, given the variates they dropped. We also performed a classification tree, giving us a similar accuracy score to lasso. This is not a surprise because noisy real-world data would be predicted well by a single tree learning algorithm. Indeed, the random forest had the best performance of all models, with its accuracy value being 0.7850394 (see Appendix A: L).

The accuracy score of the random forest is given by the model with hyperparameters at default. We then roughly tuned the hyperparameters by out-of-bag (OOB) error to optimize the model because it reduced the computational cost. Out-of-bag (OOB), similar to out-of-sample

(OOS), can provide some implications for the performance of our models, even if it is not necessarily as precise as OOS. We only looked at "ntree" ranging from 400 to 600, and we could continue testing until the improvement of OOB error is less than 10^{-4}. With that idea, we found the relative best model rather than the actual best. The model has the smallest error when "mtry" is around 16; therefore, we used a 5-folder cross validation to finally tune the threshold and tree numbers and compare the accuracy. By using different combinations of tree sizes of 400, 500, and 600 at threshold levels of 0.45, 0.50, 0.55, and 0.60, we realized that the accuracy of a random forest with 600 trees, "mtry" being 16, and a threshold of 0.5 was the highest (Appendix A: M-P). Specifically in the 5-folder cross validation, the average accuracy performance of this model reached 0.793.



This was the final model we used to make predictions on the test dataset regarding which Airbnb rentals in New York City are good, with a high rating of 5, and which are not as good, with any rating lower than 5. The accuracy of predictions in the test set reached 0.7977278, and the AUC-ROC curve is exhibited below.



Confusion Matrix		Actual Rating	
		High	Low
Predicted Rating	High	410	79
	Low	2466	9632

We did not apply this model to a cost-benefit analysis because the business value of this model is more about indirect benefits through recommendations and customer experience rather than a direct impact on the profitability of each Airbnb listing. For example, it would be hard to define how much benefit we can bring by encouraging high-rating listings to get registered or updating the recommendation system.

Instead, we compare the feature importance and realize that the price and availability of listings would be the most important ones to predict the customer rating. (Appendix A: Q). On the contrary, languages of listings' names seem to be the least irrelevant factors, and a possible explanation is that the package we used did not detect languages perfectly, which makes the data even more noisy to be analyzed. Some similar limitations will be discussed in the next section.

Deployment:

The main findings of our analysis suggest that it is possible, but with limited out-of-sample accuracy, to predict whether a future listing will be excellent, given that the available real-world data is noisy. Our analysis could be used in providing Airbnb recommendations for promoting types of listings that possess the aspects of what our analysis deemed as excellent qualities of a listing. Airbnb could, in turn, advertise these types of listings more, leading to more customers choosing listings that are more likely to provide an excellent

Airbnb experience. Additionally, Airbnb could give recommendations to listings having qualities our analysis deemed as poor qualities of a listing. These recommendations could include how to improve their listing for a better customer experience. Overall, this would lead to more positive reviews and revenue for Airbnb.

The prediction accuracy suggests that it is possible to estimate the quality of Airbnb roughly. Our best model is a random forest with an accuracy at 0.798. This is reasonably better than the random guess and may be used as a temporary screening instrument for potential renters by Airbnb if such a system is not already in place. In its current state, the model allows us to predict the quality of the listings that renters make and give them feedback on what they should change in order for their listing to be considered excellent. This feature within the app would help renters make better listings, and hopefully, this would lead to higher guest satisfaction.

The main drawback of the model is that it suffers from unobserved features. We only used the listing information posted by the hosts to predict the rating. Other factors such as host friendliness, which is difficult to observe and quantify unless written by visitors as review, the convenience of public transport, proximity to tourist locations of interest, price-to-quality ratio, and more could all be used to improve prediction accuracy. Adding these variables could also allow us to better examine the relationship of the ratings and amenities with price. If we could somehow quantify the utility of amenities observed from the keywords, and combine them with our predictive model, we may see a significant improvement. An extension for the NLP techniques would be to acquire data of guest reviews and analyze it. This would allow for better approximation of guest satisfaction, as in our case we can only use house rules provided by the hosts. Another suggestion for further research is using a neural network to analyze the photographs provided by the renters. If our model could estimate for instance the presence of a bed, tv, and other important items for visitors and perhaps go even as far to estimate the size of the apartment, we could have a very powerful model.

In the ideal scenario, the deployment would be an ensemble model featuring textual, photographic and demographic data capable of ranking listings according to their estimated quality. This system would then post "flags" to renters on the app and advise them on how to improve their listing, if it does not satisfy all the "optimal" criteria suggested by the domain experts. This allows for better arrangement of listings on the app, as we would be sure that the best listings always appear first. Further, the model may suggest improvements for the listing to the renters such as "house rules not in line with Airbnb policy" or "photographs do not show enough information to guests." We may even establish a model that takes price into account in order to suggest to renters what is realistic to expect when putting up their apartment. In this way we improve the experience of both the renters and the guests and therefore boost the image and quality of experience for Airbnb.

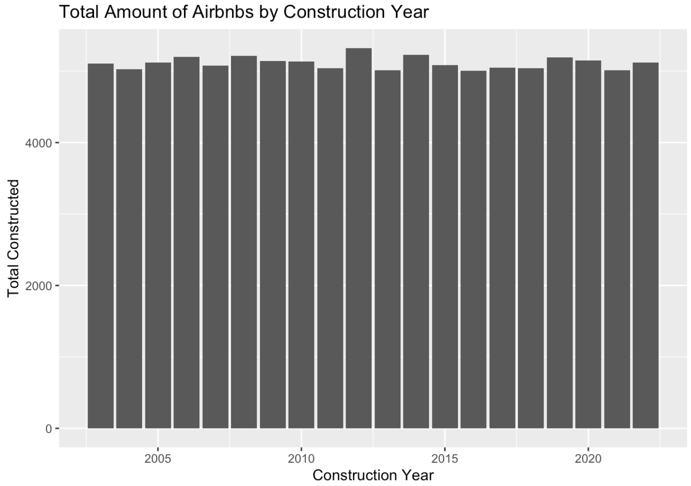
In terms of ethical issues, the main concern here is data privacy. While all of the information we used in our analysis is public, in order to have a deeper, more accurate analysis of what makes an excellent Airbnb, customer data that is considered private would be needed (if we want to make suggested apartments for different groups of guests). In terms of risks associated with our proposed plan, if our model predicts a false negative, meaning an excellent Airbnb is deemed poor, it would flag this user as a bad Airbnb. This host might be unhappy about this, because it could reduce their income and amount of potential visitors, and subsequently remove their listing from Airbnb, potentially moving their business to another site. On the other hand, if our model predicted a certain Airbnb to be excellent, but in reality, a customer had a bad experience there, this could result in lower star ratings for the listing.

Works Cited:

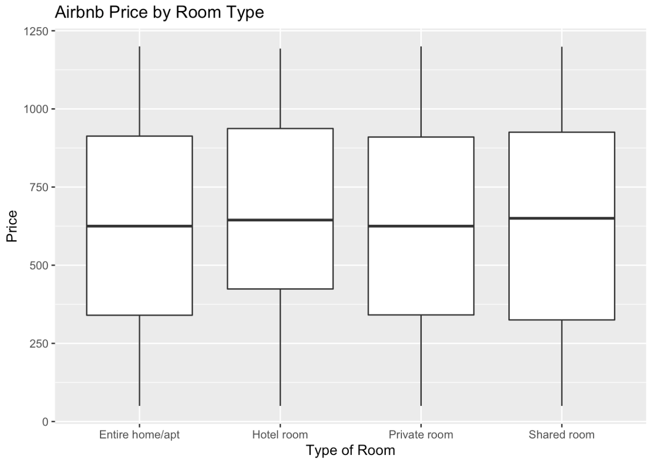
1. *nytimes.com*. (2021, December 9). Retrieved October 14, 2022, from https://www.nytimes.com/2021/12/09/nyregion/nyc-illegal-Airbnb-regulation.html

Appendix A: Supplementary Materials

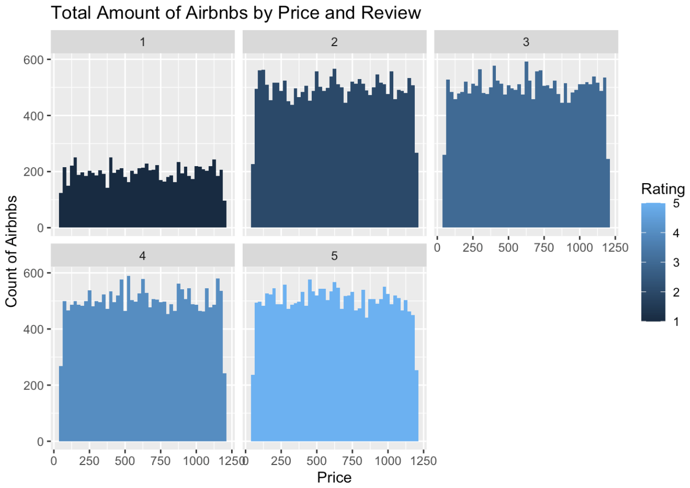
A) EDA: number of airbnbs by construction year



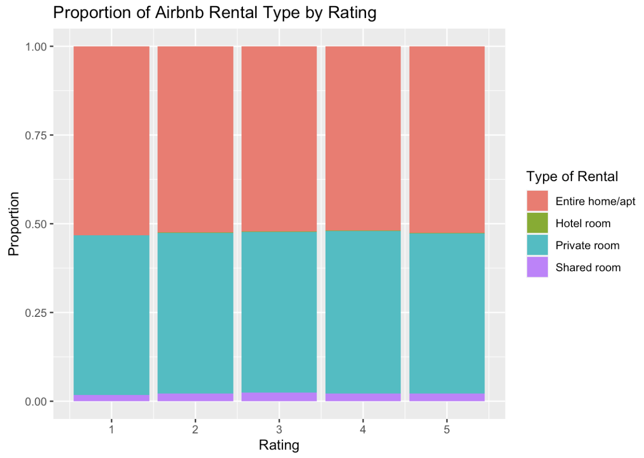
B) EDA: airbnb price range by different types of rooms



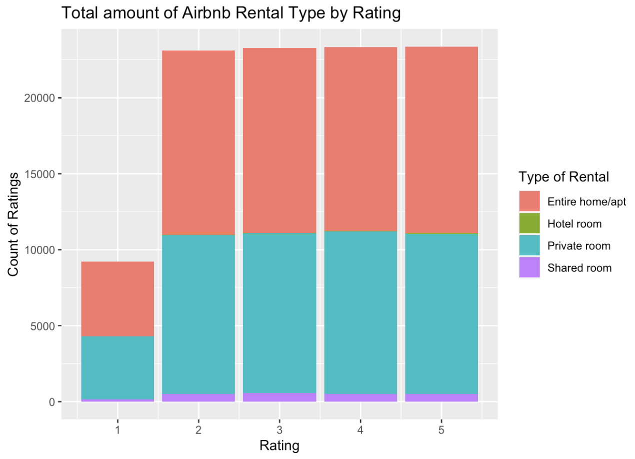
C) EDA: total amount of airbnbs by prices and review



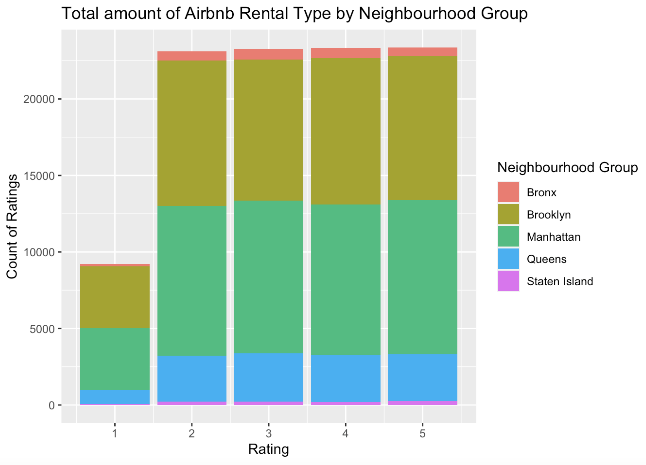
D) EDA: proportion of airbnb rental type by rating



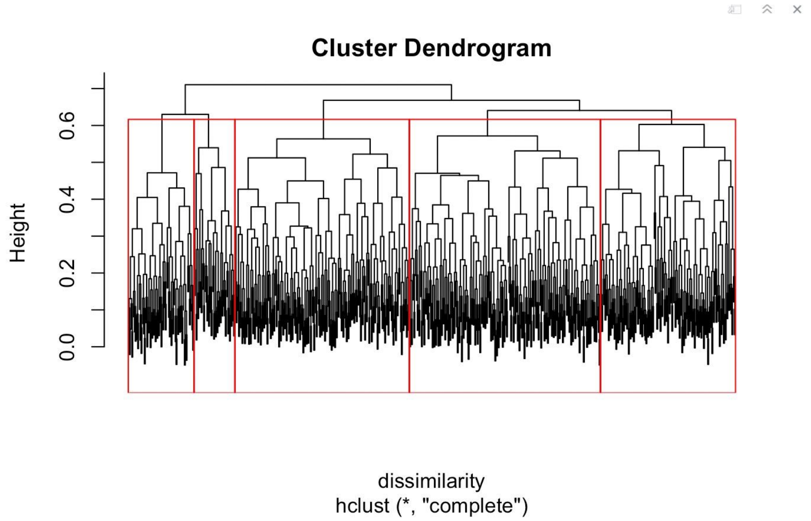
E) EDA: total amount of airbnb rental type by rating



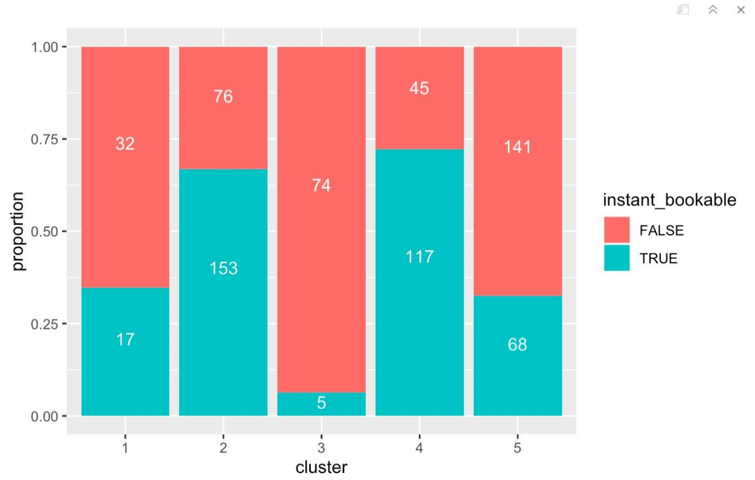
F) EDA: total amount of airbnb rental type by neighbourhood group



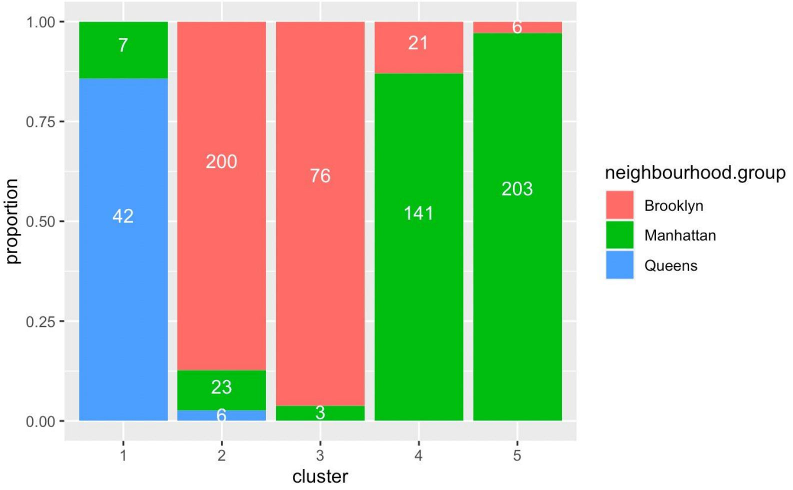
G) Hierarchical Clustering: the whole cluster dendrogram



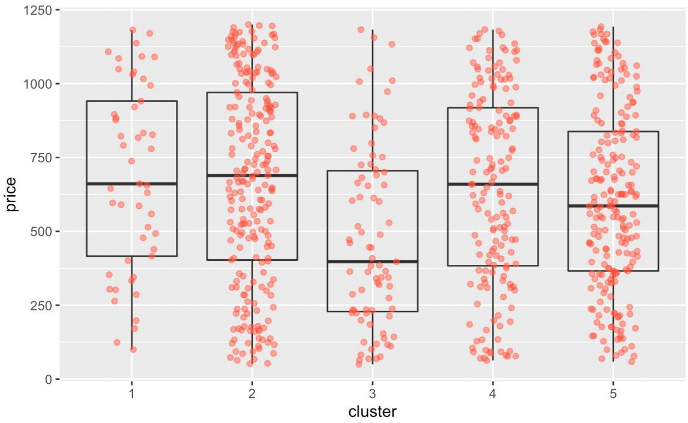
H) Cluster Analysis Classification based on Bookability



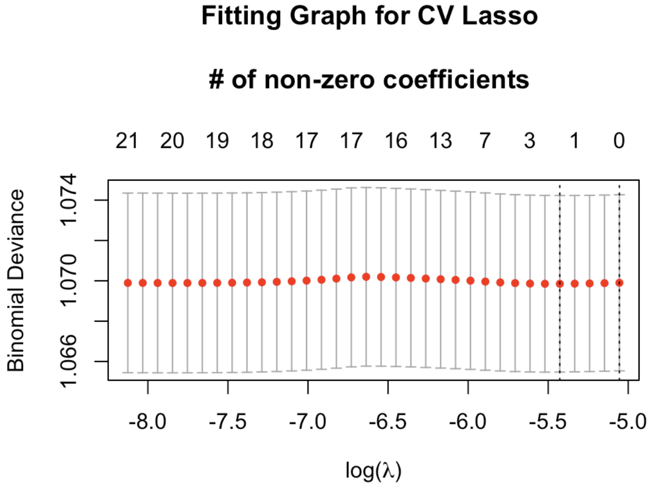
I) Cluster Analysis Classification based on Neighborhood Group



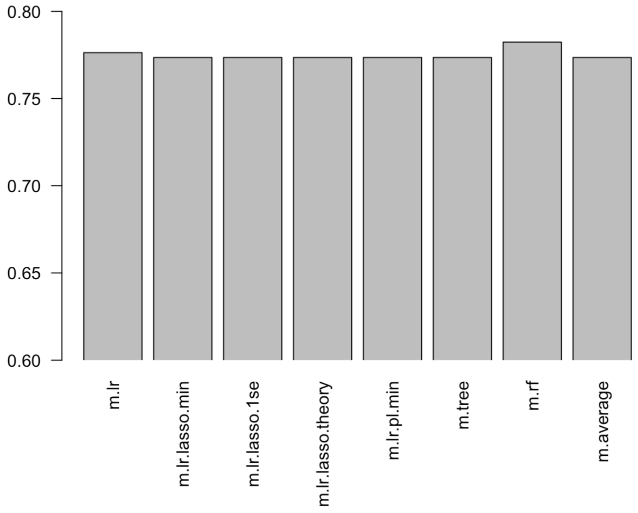
J) Distribution of Price in Each Cluster



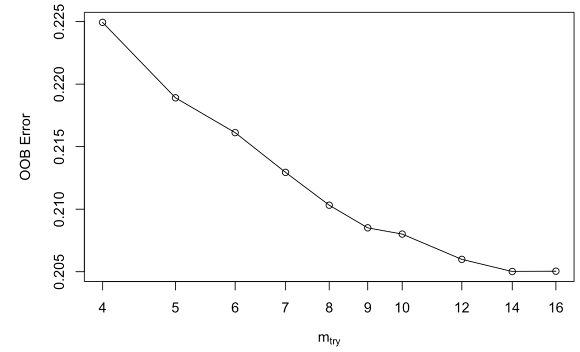
K) Lasso Binomial Deviance as the number of coefficients changes



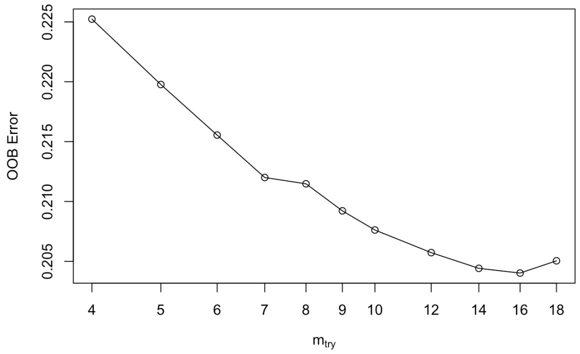
L) Out of Sample Accuracy Score in Cross Validation



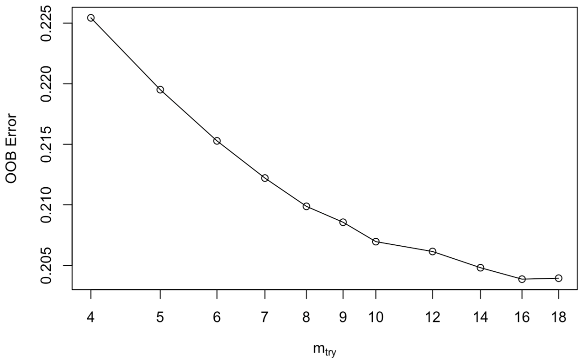
M) Out of Bag Error by Choosing Different mtry at ntree = 300



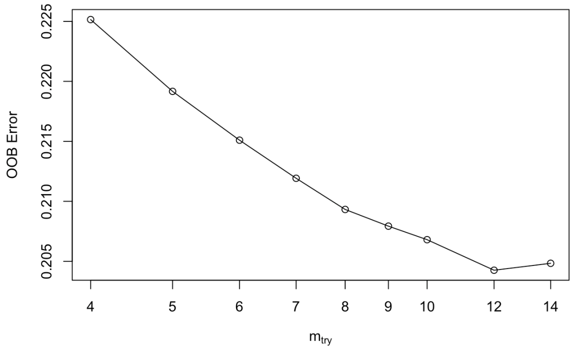
N) Out of Bag Error by Choosing Different mtry at ntree = 400



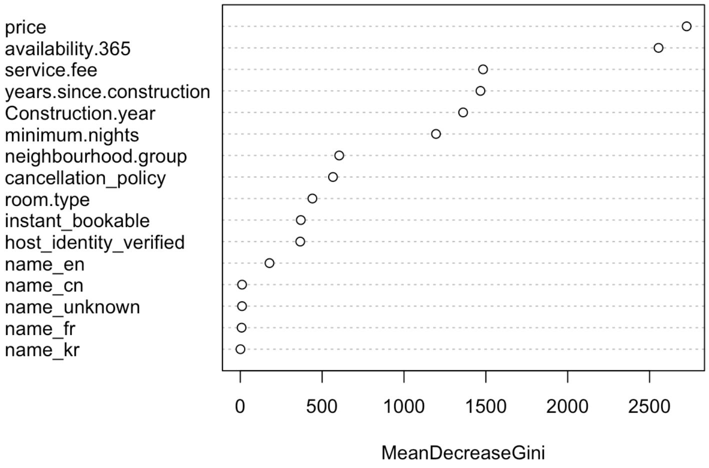
O) Out of Bag Error by Choosing Different mtry at ntree = 500



P) Out of Bag Error by Choosing Different mtry at ntree = 600



Q) Feature Importance



Appendix B: Team Contributions

Maggie Lundberg: exploratory data analysis, clustering map, write up, powerpoint creation

Tingche Lyu: Data cleaning, cross validation, lasso, post lasso, tune hyperparameters, write up

Niyati Mody: write up, powerpoint creation, logistic regression

Vladislav Stanojevic: SVM, word cloud, write up, powerpoint creation

Kala Nandini Uppala: clustering, random forest, write up