Final Project

Jack Cunningham & Ali Fazl

2023-04-23

Abstract:

In this project, we aimed to predict the price of New York City AirBnB listings using various machine learning models. The random forest model performed the best, with an RMSE of 0.326, despite the inherent challenges in predicting a variable with many upper outliers. Our model identified four main categories of features that best predict price: location, property type, minimum stay requirements, and host-specific features.

Although the random forest model outperformed other methods, it still had a relatively high RMSE, likely due to potential issues with both the model itself and the data. Nonlinear relationships, missing interactions between features, and issues in feature selection might have contributed to the model's error. Additionally, limitations in capturing location information and seasonality may have hindered the model's predictive power.

The model performed particularly well in most of Manhattan above Midtown and in most of Brooklyn outside of Downtown Brooklyn, while it struggled in sparse areas on the outskirts of the city and in the most expensive areas of Lower Manhattan and Downtown Brooklyn. The analysis showed that seasonal factors and certain amenities had minimal impact on price.

Our findings provide valuable insights for both AirBnB hosts and travelers. Hosts should focus on property location, type, minimum stay requirements, and maintaining a good reputation to optimize their pricing strategy. Travelers can use this information to compare similar listings, prioritize important features, and find the best value for their money. However, due to the limited number of price data points and the relatively high error rate of our model, both hosts and travelers should interpret these findings with caution.

Introduction:

Introduce the question you're trying to answer at a reasonable level of detail. Give background and motivation for why it's important.

Our question is: What factors best predict AirBnB prices in New York City, and how can hosts and travelers use that information?

The rise of the sharing economy has transformed the way people travel and seek accommodations. Platforms such as AirBnB have gained popularity by allowing property owners to rent out their homes or rooms to travelers, offering an alternative to traditional hotels. New York City sees over 50 million visitors per year; the city has experienced significant growth in its AirBnB market, with tens of thousands of active listings at any given time. Accurate prediction of AirBnB prices is essential for hosts to optimize their revenue and for travelers to make informed decisions when selecting accommodations. This project aims to examine the factors connected to AirBnB prices in New York City, focusing on the role of geography.

The audience for this project is hosts and travelers who want to understand the factors that predict New York's AirBnB prices. Travelers might be interested in saving money; understanding the factors that are most important for price, and the spatial locations of the most and least expensive AirBnBs, could help in that goal. And hosts, seeking to maximize revenue, will be curious about the factors that predict price for the

same reason travelers are. The model in this project gives insight into pricing for a wide range of properties across the city; hosts could use the predicted prices of similar properties to set their own pricing strategies.

Spatial geography turns out to be a key predictor of New York's AirBnB prices; as the saying goes, location, location, location. To better understand how well our predictive model performs in each neighborhood, we map the predicted prices of New York's AirBnBs against their actual prices, and we map the percent error of our model. This helps the reader understand where in the city our model does a good job predicting prices, and thus how well calibrated our model is to their neighborhood of interest.

In this project, we employ machine learning techniques, including LASSO regression, random forest, and gradient boosting, to best predict AirBnB prices based on a comprehensive set of variables. We leverage clustering algorithms, ie. DBSCAN, to capture spatial patterns in the data and map predicted vs. actual prices. By doing so, we aim to enhance our understanding of the factors driving AirBnB prices in New York City and provide valuable insights for hosts and travelers alike.

Methods:

Data

For this project, we use four datasets combined into one. Each dataset contains the entire set of scraped NYC AirBnB listings at the following dates: June 15, 2022; September 15, 2022; December 15, 2022; and March 15, 2023. The data contain 70+ variables and more than 160,000 total listings (roughly 40,000 per quarter). These data come from InsideAirbnb: http://insideairbnb.com/get-the-data/.

We made some important modifications to the dataset in order to meet our needs:

- Creating dummy variables for each season (June is summer, September is fall, December is winter, and March is spring), depending on which initial dataset the observation came from.
- Modifying the host_since variable to provide a number of years since the start of the host's presence on AirBnB.
- Removing roughly half the columns which we did not use in our analysis.
- Dropping all observations with host_since < 1 year. This ensures that all the listings come from hosts who have been on the platform for at least one year, helping to ease concerns about seasonal effects.
- Dropping all observations with NA values in any of the fields. Among other effects, this ensures that our dataset includes only listings with at least one review, host-provided descriptions, and complete information about amenities, bedrooms, etc.
- Manipulating variables to be easier to work with, e.g. adding dummy variables for room_type and changing f/t format to 0/1.
- Using the DBSCAN clustering algorithm to spatially cluster the listings, to create another spatial variable beyond latitude and longitude to use as a predictor in our models.
- Extracting six important terms from the "amenities" list to use as predictors in our models.
- Filtering price outliers (those over $1500, roughly 0.4% of our dataset).

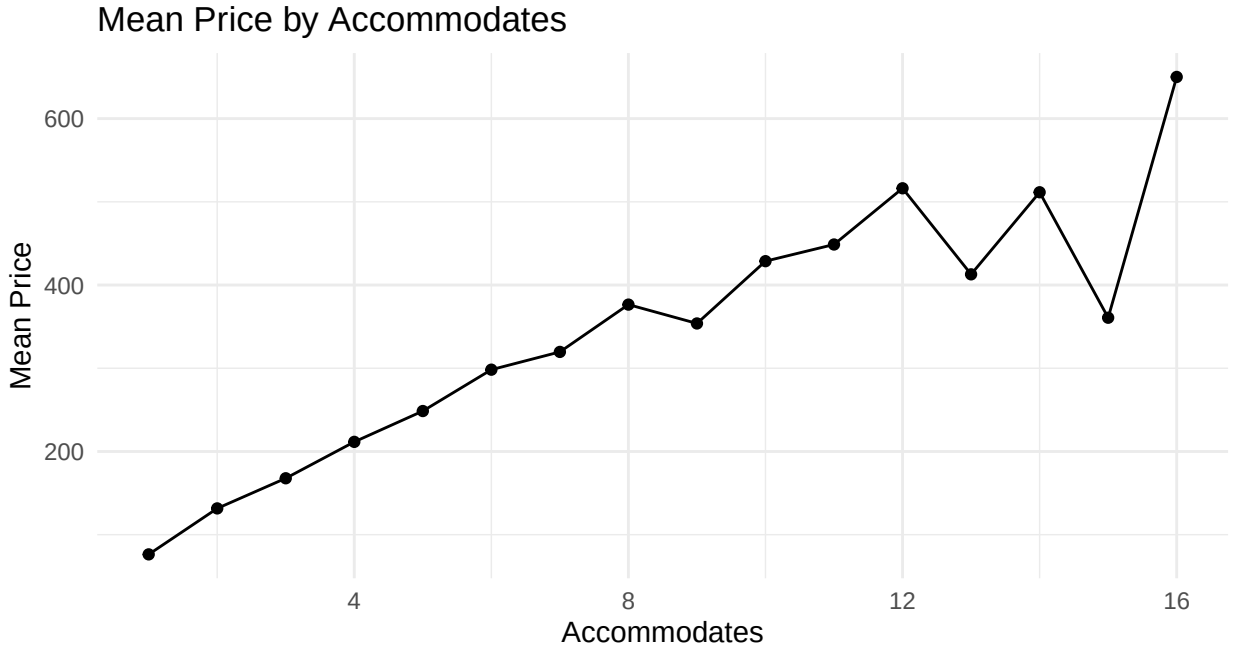
We then took a random sample of 25% of the cleaned data to use for our analysis. This sample still has over 25,000 observations of 42 variables. Reducing the dataset via random selection makes it easier to work with computationally while not sacrificing much accuracy, especially since we are working with full data on all of NYC's AirBnBs.

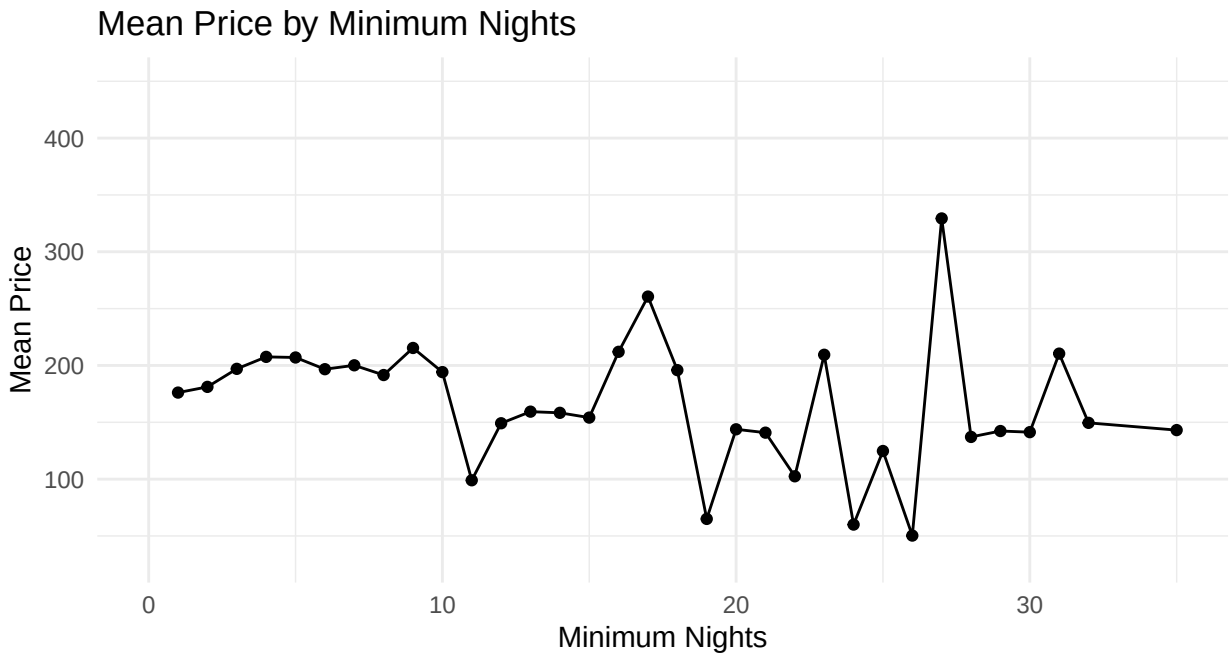
The most important reason we combined the four datasets into one is that it gives us a snapshot of seasonality. We only have price data for the dates that each dataset was scraped (6/15/22, 9/15/22, 12/15/22, and 3/15/23). One limitation of our dataset is that we don't have daily price data for a year; this would have been desired in order to uncover seasonal trends and variation with more fidelity. However, quarterly price

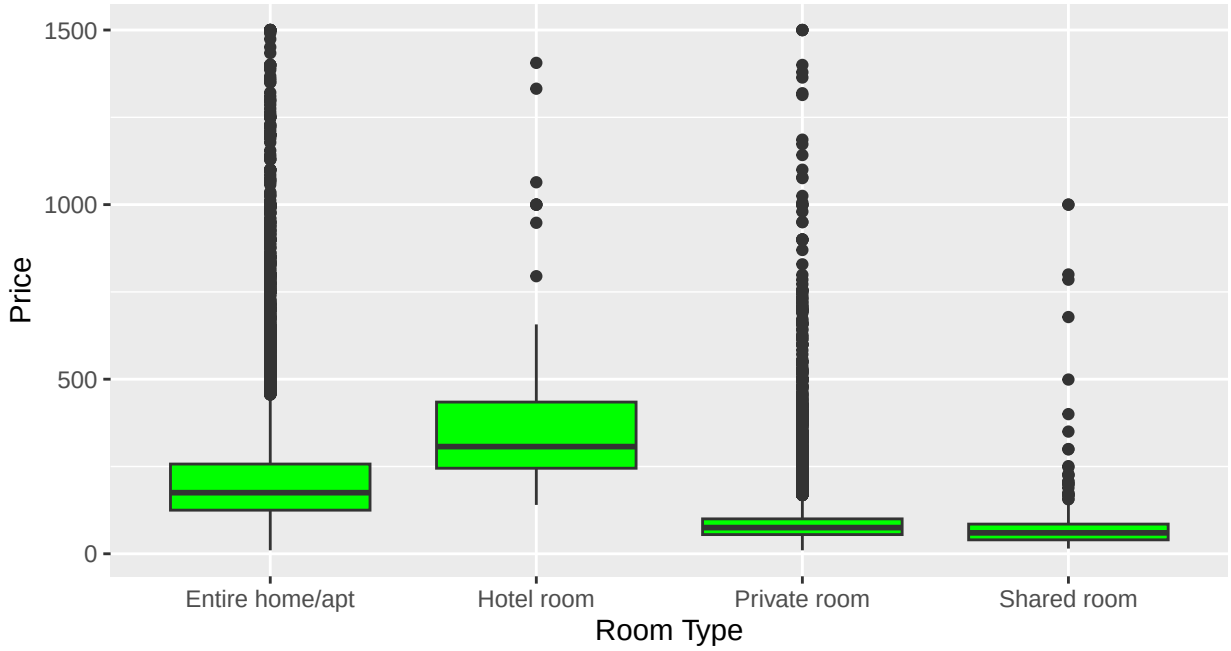
data provides a rough proxy of seasonal price trends. As our results will show, season is not a particularly important predictor of NYC AirBnB prices.

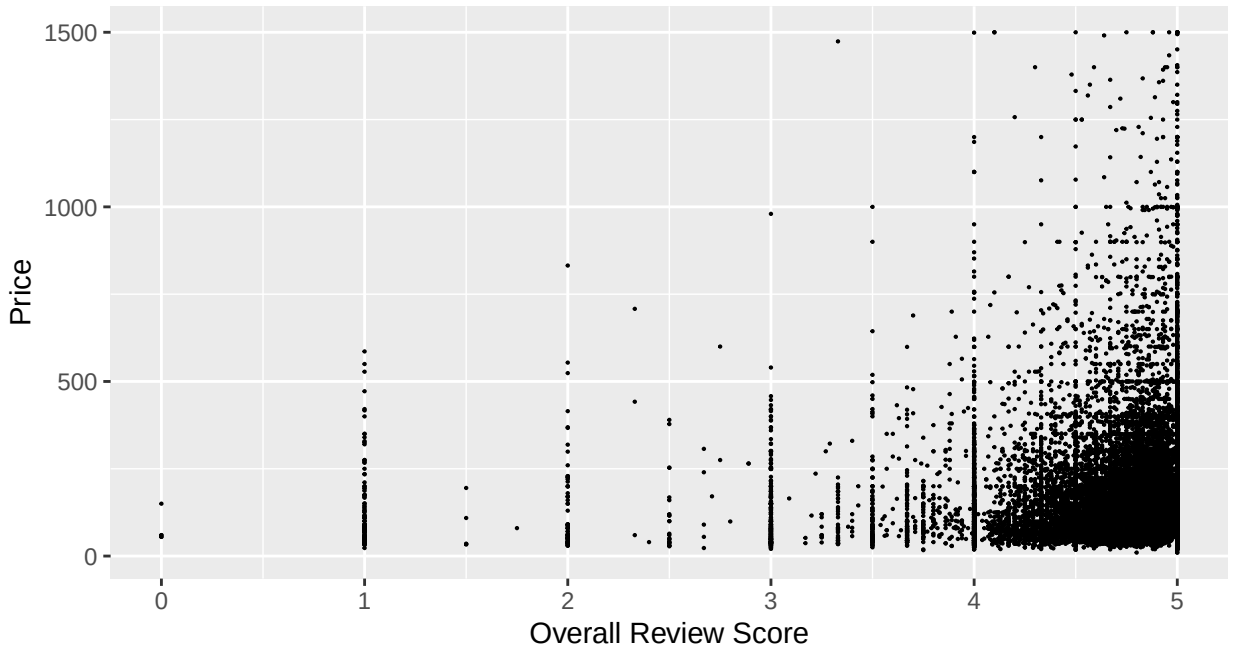
One other important note about our dataset is that we are treating each listing as unique, despite the fact that many of the listings are run by the same hosts in each period. The reason for this approach is that many of the variables we use as predictors can change from one quarter to the next; e.g., review scores, availability data, and number of reviews, among many more. Hence, it makes more sense to treat each observation as unique, rather than simply extracting the seasonal prices for each listing and attaching them to one of the seasonal datasets.

The plots that follow give a sense of how price relates to some of the important variables in this dataset, including how many guests the listing can accommodate, minimum length of stay, listings of each room type, and overall review scores.




Mean Price by Minimum Nights
Mean Price
400
300
200
100
0
10
20
30
Minimum Nights


Price
1500
1000
500
0
Entire home/apt
Hotel room
Private room
Shared room
Room Type



These plots demonstrate that some of the variables that we think have a big impact on price, like accommodates and room type, have a clear (if noisy) relationship, while others are much less clear. For example, it's hard to tell if there's a downward relationship between minimum price and nights or if that's just noise in the dataset. And in the review score plot, there is so much mass in the square between 4 and 5 review score and $0 and $500 price that it's hard to see any relationship, other than with outliers.

Given this initial glance at the dataset, it's clear that we will have our work cut out for us in trying to predict price. In the next section, we will describe the approach we took in order to build the best predictive model of price possible, despite the noise.

Approach

We used a selection of methods in order to create the best price prediction model possible with this dataset. We trained each model on 80% of the cleaned dataset, then tested its root mean squared error using the remaining 20% of the dataset. The outcome of interest here is log price instead of price, in order to help normalize the distribution; as we will demonstrate below, the price variable is heavily left-skewed. This transformation also helps in the case that the relationship between our predictors and the price is non-linear.

We first tried a simple linear model, including just a few factors we thought would likely prove important, including location, property type, and review score. We then used the LASSO method with all the variables in the cleaned dataset in order to determine the most important ones; then, we ran a linear model using only the variables selected in the LASSO process.

Then, we turned to more sophisticated machine learning techniques: random tree, random forest, and gradient boosting. We tweaked each of these models via trial-and-error, adding and removing predictor variables as appropriate for the model. The random forest model turned out to have the best performance; we plotted the importance of each variable in this model in order to determine the most important factors affecting price.

Finally, in order to map our selected model's performance in each neighborhood, we first added the predicted price values from the random forest model to the dataset for each listing. Then, we calculated the mean predicted price in each of the 235 distinct New York City neighborhoods in our dataset, along with the actual mean prices for each neighborhood, and calculated the error rate. Finally, we plotted all three of these

measures for each listing by latitude and longitude, along with NYC's geographic boundaries.

Excluded Approaches

As an aside, we'd like to mention a couple of approaches to this question we tried that did not work, or did not prove useful, and are thus not included below. They include:

- Completing all the same analysis, but for occupancy rate: Our original question was about predicting New York's AirBnB occupancy rate, in addition to price. However, occupancy rate was not a variable we had access to in the dataset, so we sought to construct the occupancy rate for each listing. We realized that the algorithm that others who have studied AirBnB's occupancy rates have used (monthly reviews x expected review rate x average stay length) relied on the faulty assumptions that expected review rate and stay length are constant for all listings in the dataset. Rather than using an algorithm based on such faulty assumptions, we decided to forgo our analysis of occupancy rate and instead limit our analysis to price.
- Topic Modeling: We tried using the topic modeling technique of Latent Dirichlet Allocation (LDA) to extract key words and topics from the names and descriptions of the listings. Unfortunately, upon performing this analysis, the topics were insufficiently differentiated from each other (e.g., four topics under the heading "apartment" in the top 10), and the coefficient of each topic was so low, that they yielded very little insight into predicted price and we excluded the results from this report. However, in a similar vein, we did pull out six key words from the amenities column and created new dummy variables to use as predictors in our models: pets_allowed, kitchen, heating, air_conditioning, elevator, and long_term_allowed.
- PCA:

Principal component analysis (PCA) is a technique used to obtain reduced-dimensional representations of high-dimensional data sets. In the case of our model, which has 48 features, we were motivated to explore the effectiveness of PCA in summarizing the data. We experimented with PCA using two different approaches.

In the first approach, we applied PCA to all of the numerical variables, reducing the dimensionality to rank=8. We then used these 8 variables to estimate prices using a random forest model. However, the RMSE of this model tended to be higher than 0.4, which is not acceptable.

In our second approach, we categorized the features into two groups based on their relevance to hosting and the physical place. The first category included host-specific features such as "host_since", "host_total_listings_count", "host_identity_verified", and others. The second category included place-specific features such as "bedrooms", "room_type", "accommodates", and others. We then applied PCA separately to each category, with rank equal to 3 for the host-related features and rank equal to 5 for the place-related features. We used these reduced-dimensional representations in a random forest model to estimate prices. According to our simulations, the RMSE results were between 0.36 to 0.38.

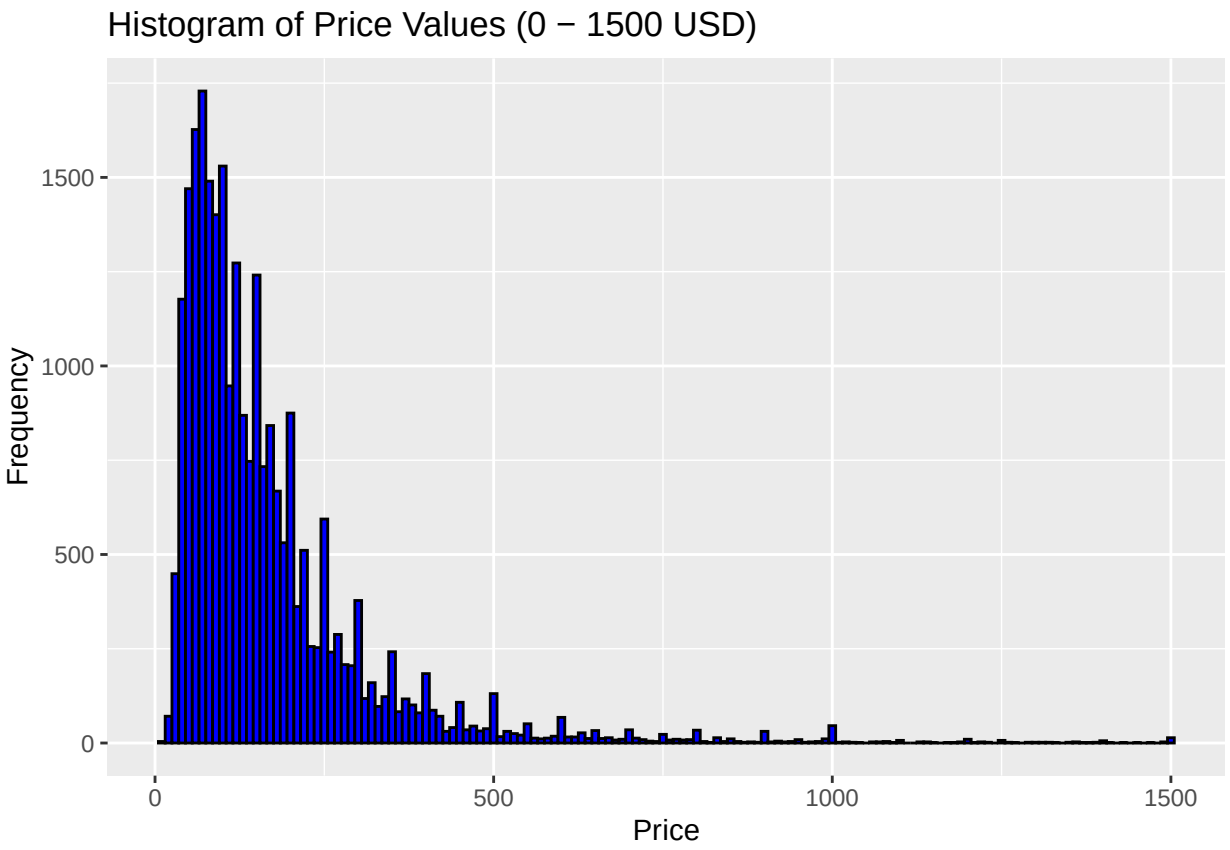
Although the accuracy of neither approach was better than the overall model, the second method, which utilized only 8 variables, yielded a RMSE that was relatively close to the best result. This suggests that reducing the dimensionality of the place-related features using PCA may have some potential in improving the model's performance. (If you'd like to reproduce our analysis, we included our PCA code as comments in the code block below; simply remove the first # from each line of code.)

Results:

Understanding Price

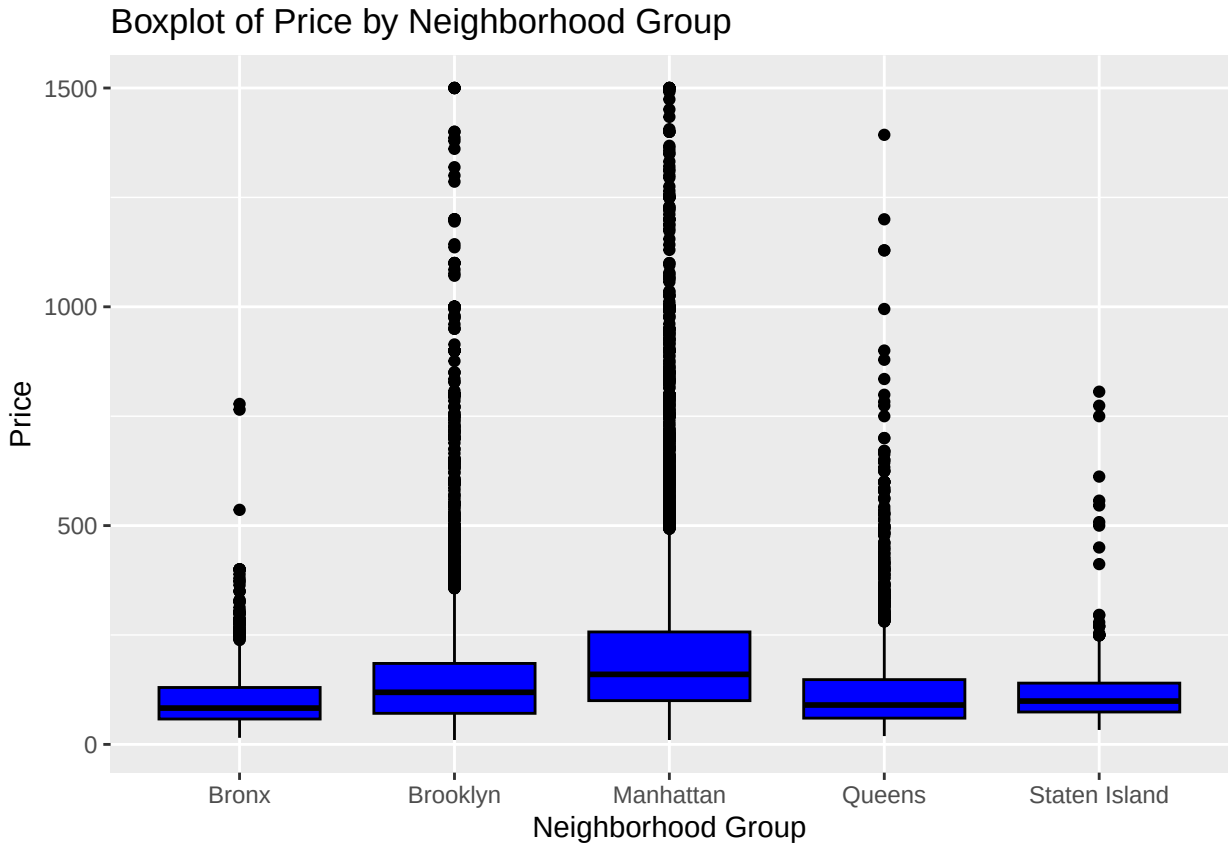
Price in our dataset is left-skewed; while some listings exceed $1000 per night, 75% of listings are $200 or less.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0    75.0   125.0   167.9   200.0  1500.0
```



Manhattan has the most expensive listings, with a median nightly price of $160. Brooklyn follows at $119 per night; listings in the other boroughs have median nightly prices between $80 and $100. Each borough has quite a few upper outliers in price, as demonstrated by the boxplot below.

neighbourhood_group_cleansed	median_price
Bronx	83
Brooklyn	119
Manhattan	160
Queens	90
Staten Island	99



Below, we've included the median listing price across the city by room type and season. Hotel rooms are by far the most expensive, although they are the least common. Entire homes are more than twice as expensive as private rooms, and nearly three times as expensive as shared rooms.

room_type	median_price
Entire home/apt	177
Hotel room	307
Private room	75
Shared room	60

Season appears not to affect citywide median price much. The median price in each season ranges from $120 in spring to $130 in winter. Again, since we only have price data for one date per season, we lack the fine-grained price data that would be useful to identify bigger seasonal shifts, or price trends for each neighborhood (especially during e.g. big events or holidays). But for many listings in our dataset, the listed price was identical for each of the four dates of scraped data. It's possible that many AirBnB hosts don't change their prices much or at all from the default, although without better data this hypothesis is just speculation.

At the borough level, there is more seasonal variation in median price in some boroughs, and less in others. The range of Staten Island's median price by season is 25, while the range of Queens's median price by season is just 4.5.

season	median_price
fall	128
spring	120
summer	121
winter	130

neighbourhood_group_cleansed	season	median_price
Bronx	fall	90.0
Bronx	spring	79.0
Bronx	summer	75.0
Bronx	winter	86.0
Brooklyn	fall	120.0
Brooklyn	spring	115.0
Brooklyn	summer	115.0
Brooklyn	winter	123.0
Manhattan	fall	167.0
Manhattan	spring	150.0
Manhattan	summer	160.0
Manhattan	winter	175.0
Queens	fall	90.0
Queens	spring	92.5
Queens	summer	88.0
Queens	winter	90.0
Staten Island	fall	100.0
Staten Island	spring	90.0
Staten Island	summer	95.5
Staten Island	winter	115.0

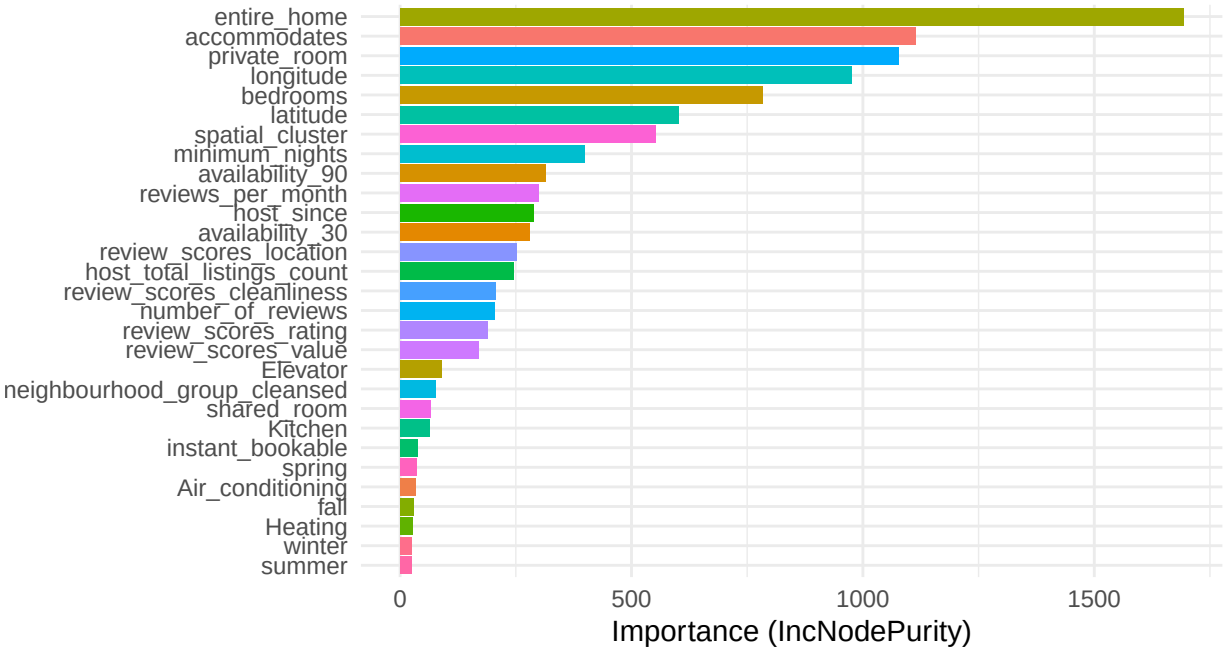
Building a Prediction Model

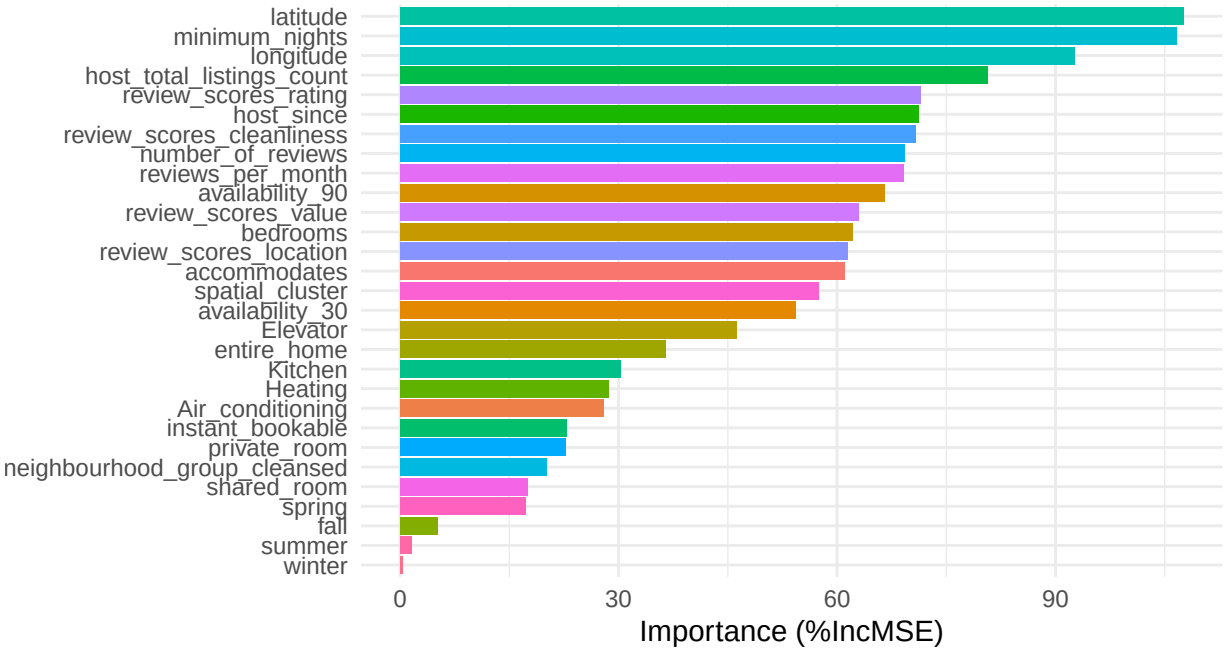
The table below gives the RMSE of each model we tested. The best-performing model is the random forest model, which has a RMSE of 0.326; on average, this model is roughly 32.6% off of the true price of a given listing.

Table 5: RMSE of each model (log price)

model	RMSE
lm_null	0.7150350
lm2	0.4858065
lm3	0.7041360
lm_lasso1	0.4336333
BNB.tree1	0.4419975
BNB.forest1	0.3260509
BNB.boost1	0.3321262

Below, we see two variable importance plots; the first represents the increase in node purity from each variable, and the second represents the percent increase in MSE if that variable is omitted. In general, the variables that have the most impact on the model are latitude & longitude (location), room type, minimum length of stay, number of guests the listing can accommodate, and reviews per month. The least important variables in our model are, in general, the seasonal variables and the amenity variables.

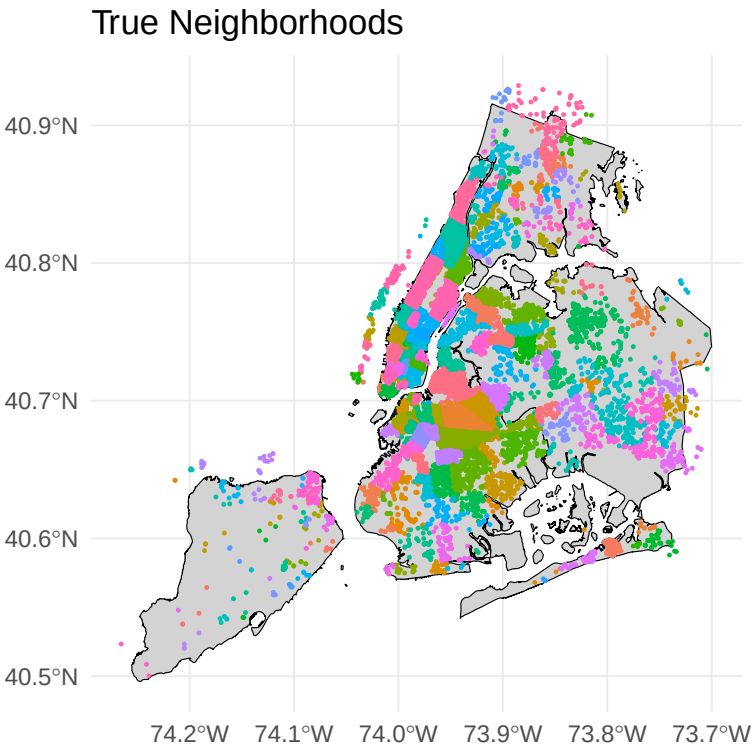


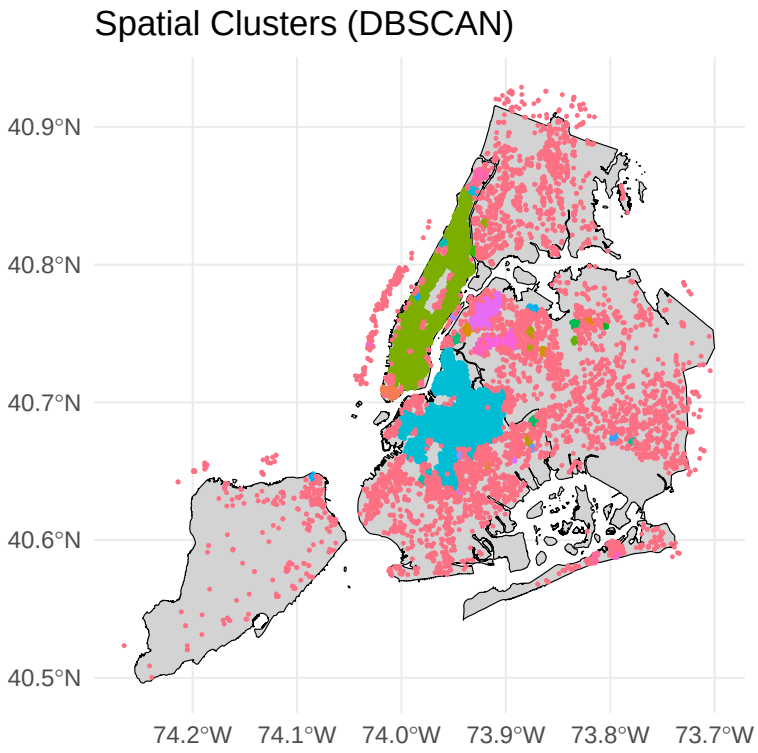


Mapping

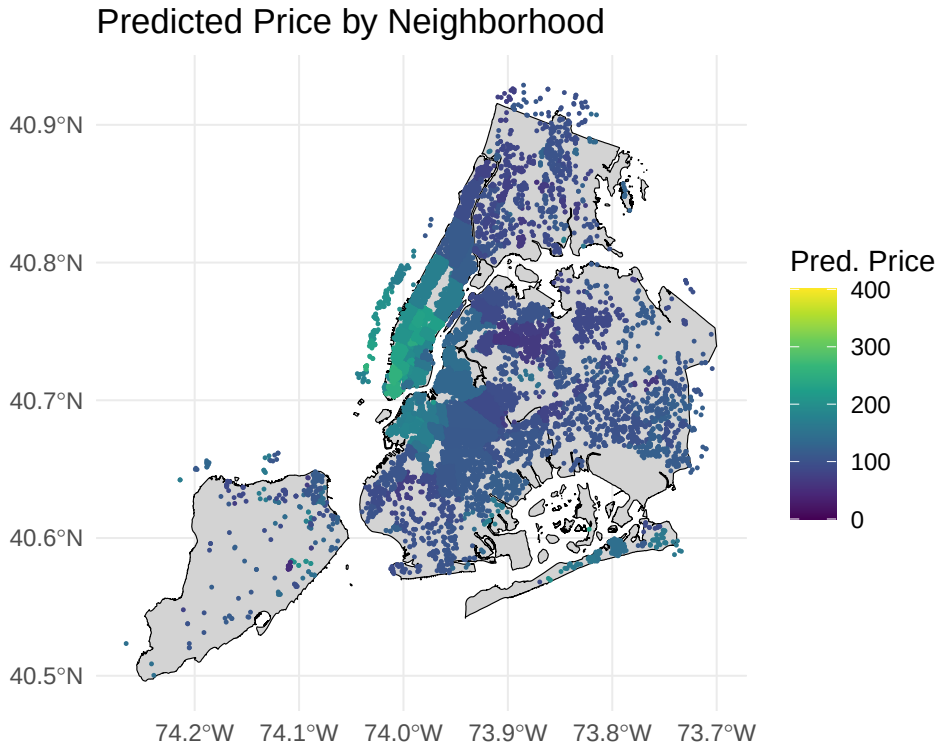
Here we see the maps of true neighborhood (not borough; recall that, while there are 5 boroughs in NYC, there are over 250 neighborhoods in our dataset) as coded in the dataset, versus spatial cluster as predicted by the DBSCAN algorithm.

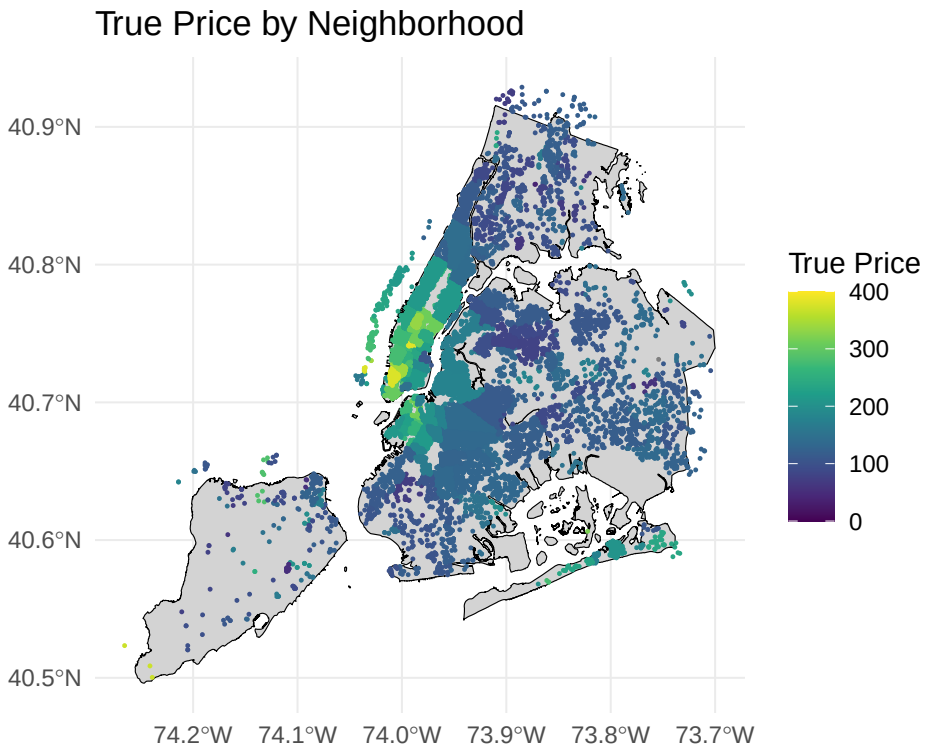
The spatial clusters do a reasonably good job of telling the difference between Manhattan, Brooklyn, and everywhere else, but they are not nearly as fine-grained as the true neighborhood from the dataset. This might explain why they are not a particularly important predictor of our random forest model.



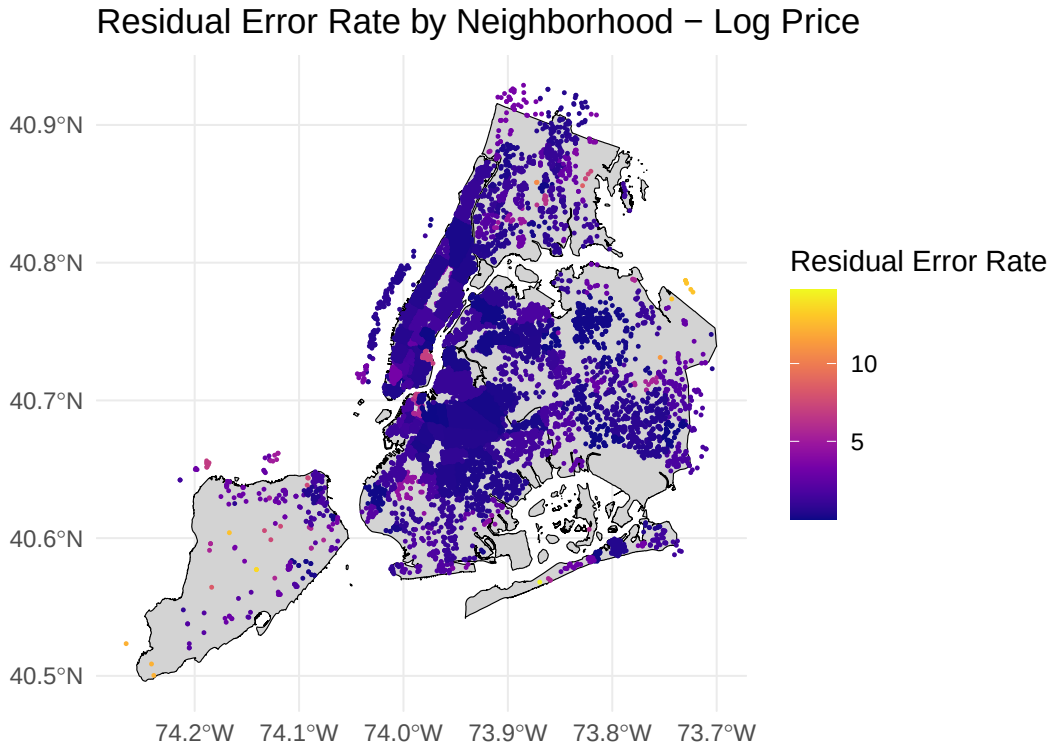


Below, we see first the map of mean predicted price by neighborhood (converted back from log form for interpretability), followed by the map of actual mean price by neighborhood.





And below, we see the mean residual error rate of our model's predicted price in each neighborhood:



Conclusion:

Of all the models we examined to predict the price of New York City's AirBnBs, the random forest model produced the best results, with an RMSE of about 0.326. There are a few potential reasons random forest worked better than other methods. By using ensemble learning to average the predictions of individual trees, the random forest model reduces the risk of overfitting. Also, by bootstrap aggregating multiple decision trees, the random forest method increases the model's diversity and robustness and reduces the variance in the predictions. Finally, by selecting a random subset of features at each split, random forest introduces additional diversity into the model, which helps to minimize overfitting and improve the generalization of the model to new data.

In spite of these factors, we ended up with a relatively high RMSE: our model's predicted price was, on average, 32.6% off the true price of the listing. There are several possible reasons for this high RMSE. The first subset of potential issues is with the model itself. There might be a nonlinear relationship between some or many of the predictors and price, which is not accounted for in our model. We only included one interaction term, latitude x longitude; if there are other important interactions between features, our model does not capture them. And while we tried several different combinations of features in our random forest model, there could still be issues in feature selection that increase the model's error.

Another subset of potential issues is with the data itself. As seen above, price is a variable with a lot of upper outliers in our dataset; even after limiting the dataset to listings with price between $0 and $1500, 75% of listings fall into the range between $0 and $200, with the rest above. This is a significant challenge for any model to accurately predict price. Our model might also not be capturing location as well as necessary. The location of the listing is an extremely important factor in its price, as seen in the maps above. Our model considered neighborhood group, spatial cluster, and the interaction between latitude and longitude; still, these are relatively crude proxies for location, and a more precise geographic variable would have been desired.

The most important variables in our model, in terms of both node purity and impact on RMSE, included location (latitude & longitude, spatial_cluster), entire_home, private_room, bedrooms, minimum_nights, and host-specific features like review_scores_rating and host_since. We posit that there are four categories of features that best predict price:

- Location: There is a wide spread between the mean price for a neighborhood like Midtown Manhattan and one like, say, Jackson Heights, Queens or West Bronx. Proximity to tourist attractions and conference centers, crime rates, transit access, real estate cost, and other factors make some neighborhoods much more desirable than others; that comes with a price premium.
- Property type: Hosts charge much more for entire homes than they do for private or shared rooms, likely because travelers see having access to an entire home as an important amenity that they are willing to pay more for. And the number of bedrooms and guests that a listing accommodates are also important; travelers are willing to pay more per night for a listing that hosts more people.
- Minimum nights: Some hosts seem to aim their listings at tourists, with minimum stay lengths between 1 and 5 nights; others aim their listings at longer-stay travelers or subletters, with minimum stay lengths of 30 nights or longer. The latter listings likely have a lower price per night than the former, since these travellers are "buying in bulk."
- Host-specific features: Some features, like review scores and how long a host has been on AirBnB, are endogenous to the hosts. Travelers might be willing to pay more for a listing with higher review scores or a listing that has been active for longer. And hosts might "learn by doing," setting higher prices as they gain experience and better reviews.

The features that are least important to predicting price in our model, in general, are those relating to season (spring, summer, fall, and winter) and those relating to amenities (heating, air conditioning, kitchen, etc.). For seasonality, this could be related to the shortcoming of our data that we only have the price for one day in each season, rather than daily price every day for a year. We could be missing important intra-seasonal shifts in price. However, this concern is mitigated by the fact that most listings for which we have four

observations had identical prices at all four times; this suggests that many hosts simply keep the same price throughout much or all of the year. And for amenities, 90% of listings included a kitchen, 78% had heating, and 68% had air conditioning; one hypothesis is that since so many listings include these amenities, on the margin they do not impact price much.

In terms of the relative predictive value of our model in different New York neighborhoods, it performed particularly well in most of Manhattan above Midtown, and in most of Brooklyn outside of Downtown Brooklyn. It performed less well in sparse areas on the outskirts of the city, including the furthest reaches of Staten Island and Queens, and in the most expensive areas of Lower Manhattan and Downtown Brooklyn. We hypothesize that the low density of listings in the outermost areas of the city gave our model a hard time, and that our model was also insufficiently sensitive to location to capture the high prices of the most desirable AirBnB locations in New York.

These findings might prove useful to NYC's AirBnB hosts and travelers alike. Hosts can learn that the location of their property, property type, minimum stay requirements, and their own reputation as a host are crucial factors in determining the price they can set for their listings. To maximize their earnings, hosts should focus on providing a clean, comfortable, and well-maintained property that caters to their target market, whether it be short-term tourists or longer-term renters. They should also strive to maintain high review scores and promptly address any concerns raised by guests. However, we caution against hosts reading too much into our model. It is most relevant for hosts who have had a NYC listing for at least a year, and the relatively high error rate of our model might indicate that there are factors beyond what our model considered that are important in setting price.

Travelers can learn from these findings that location, property type, and host-specific features are significant determinants of the price they can expect to pay for an Airbnb accommodation. When searching for a suitable place to stay, travelers should compare listings with similar attributes, such as the number of bedrooms or guest capacity to identify potential deals and make more informed decisions about where to stay. Finally, understanding that seasonal factors and certain amenities may not have a substantial impact on price can help travelers find the best value for their money. This is especially true if AirBnB listing prices are less responsive to season than hotels; in this case, AirBnBs might prove to be better deals on the margin than hotels during "high seasons", and worse deals on the margin during "low seasons." Again, travelers should be cautioned about the limited number of price data points, and the relatively high error rate of our model, and take these findings with a grain of salt.