Eötvös Loránd University
Faculty of Social Sciences
BACHELOR OF ART PROGRAM

Predicting New York City Airbnb Prices

Obtaining an accurate predictive model with business implications

Rebeka Éva Cook & Jenő Tóth

Data Science

December 15, 2021

Table of Contents

Table of Figures

1. Introduction

The housing market is a network that has been around for centuries. Individuals buying and selling their homes, renting, and letting accommodation for both short and long term. All these exchanges are crucial to analyse and understand, however, analysts have had a long time to gain insight into the nooks and crannies of the varying phenomena the market has to offer. However, over the past decade, there has been a shift in trends, as many people have decided to rent out their apartments – or at least part of them – as Airbnb's.

Airbnb is a company that provides lodging through an online system. With access to the platform, users can rent out their space or find accommodation for themselves in a few easy steps. Thus, Airbnb is primarily used for holiday rentals and tourism. Due to the simplicity and flexibility of the platform, many people who traditionally rented out housing for long term have recently changed their behaviour, turning to short-term letting. With 5.6 million active listings throughout 100,000 cities worldwide interesting 150 million users, the market for Airbnb's is currently flourishing (Airbnb, 2021). Considering that, in the grand scheme of things, the relevant market is relatively new, and yet it concerns many due to increasing popularity, gaining insight into its work is an interesting topic to analyse. The most important outcome of such research is the prediction of how much a given accommodation could be let out for, as this would be a large help for the average user. What should the go-to price of a given rental be?

The prices of Airbnb's rely on many observable and unobservable characteristics. In our research, our main goal will be to provide a simple, yet efficient model to accurately predict the price of Airbnb's based on typically observable variables. Such a model could have great business value for both amateurs wanting to gain additional income, and professionals investing in Airbnb's, similarly to those who used to invest in traditional real estate.

Although our model could be useful for all Airbnb letters, professionals investing in Airbnb's are likely to know what they are doing. Therefore, our predictive model is targeted at amateurs newly entering the market. Our model will provide a basic guideline regarding where to begin pricewise. For example, consider an elderly married couple living in a large house after their children have all left for college. Through the system of Airbnb, they can simply let out the empty rooms, filling the extra space while also earning extra income during retirement. With the help of our model, they can do this expeditiously, asking the most efficient price based on the utilities they have to offer, and consequently making the most of their given situation. Thanks to our predictive model, an average supplier on the short-term housing market can become a professional, harvesting manifold rewards in the meantime.

To obtain the best possible predictive model, we made numerous steps. We will introduce the core components in this paper based on the following structure: first, we will give a brief overview of the dataset we used; second, we will provide some insight into the most crucial points in our methodology; third, we will introduce our results regarding our variables and predictive models; and finally, we will delineate potential business implications and draw a conclusion based on the best predictive model. This paper aims to provide a thorough overview, but we may go into little detail regarding the actual analytical steps. For more information, please refer to our R code.

2. Dataset of Airbnb listings

To create our baseline predictive model, we rely on New York City Airbnb listings data obtained from the website 'Inside Airbnb' (2021). 'Inside Airbnb' provides openly accessible databases on listings globally. This website is especially useful for future users of our model, since manifold databases can be found on large cities around the world, including but not limited to Amsterdam, Bangkok, Malta, and Sydney. Therefore, our model is in no way exclusive to New York or the United States, and we can confidently recommend the implementation of this database to other international users.

Our database of choice originally included 74 variables over more than 35,000 observations. While exhaustive, the data left significant room for cleaning, which we conducted through the data analysis tools of R. Such steps included the conversion of relevant strings to integers and Boolean variables to dummy variables and dealing with missing observations. We also created some additional important variables based on our intuition and hypotheses. Our main goal here was to save as many rows as possible while remaining in line with reason. In cases where the listing price was missing, we had no choice but to drop the variable, since price is the dependent variable in our predictive model. Once our data was clean, we choose 20 variables to continue our analysis. To do this, we used our intuition and some analytical methods. We understand that some other variables may also have an impact on the price, and perhaps more could be added. However, examining all influential factors, especially considering the potentially high number of unobservable's, is out of scope of this model. We want to obtain a clearly interpretable and transparent model, for which 20 variables is plenty. Again, for more detailed steps please refer to our R code.

3. Methodology

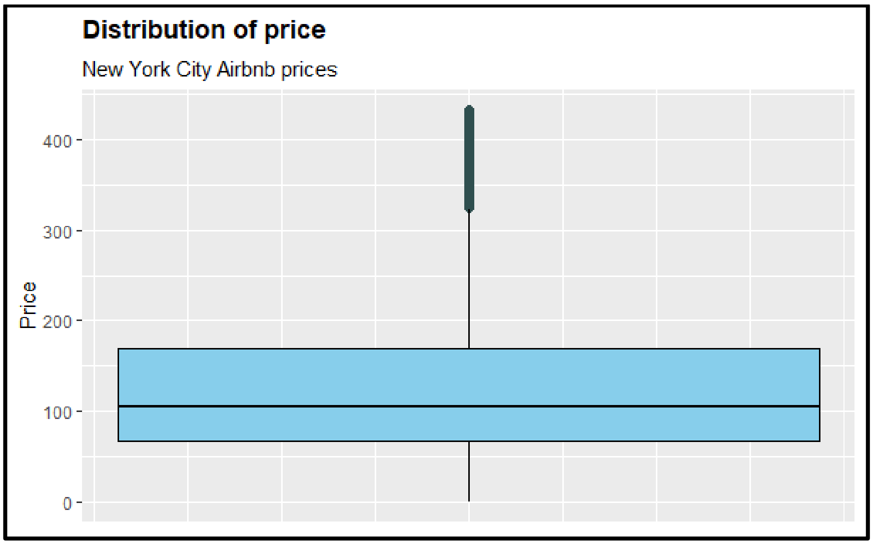
To analyse such large-scale data, we used the tools of R with the help of many packages. Once our main variables were chosen, we looked deeper into the most important contributors. We examined the distribution and works of the price variable and gave some good thought to what the best predictor of it would be. As a result, we came up with a baseline model, that certain areas of the city play a huge role in determining price. Consumers most likely use Airbnb's as cheap accommodation on their holiday, in this case shorter commuting time from the centre may be more important than the housing itself. We chose Times Square as the centre, and calculated the distance based on given coordinates. Working upwards from our baseline model, we observed the relationship price had with the other variables of interest, concluded with linear regression models. Once we gained ample insight, we moved on to the predictions. Using manifold data science skills, we compared different outcomes from different models. Our predictive models of choice were linear regressions, ridge regressions, lasso regressions, and a random forest. We made inference on the predictive power by comparing the root mean squared errors (RMSE), which are the differences between predicted values and observed values. We choose the model which minimized this measure, all to obtain the best possible prediction with our given variables. Once presented with our model of choice, we graphed the predictions to examine where the model is most efficient.

4. Results

4.1. Price of listings

As previously elaborated, our dependent variable is price. This is what we aim to accurately predict. All values should be interpreted as the cost of spending a night in the given accommodation. The average price of all listings is around $170, and 90% of listings are under $300. However, there are also some massive outliers, the maximum price lying at a shocking

$10,000 for single night. The distribution of these prices, excluding the top 5% of prices for transparency, can be seen below.

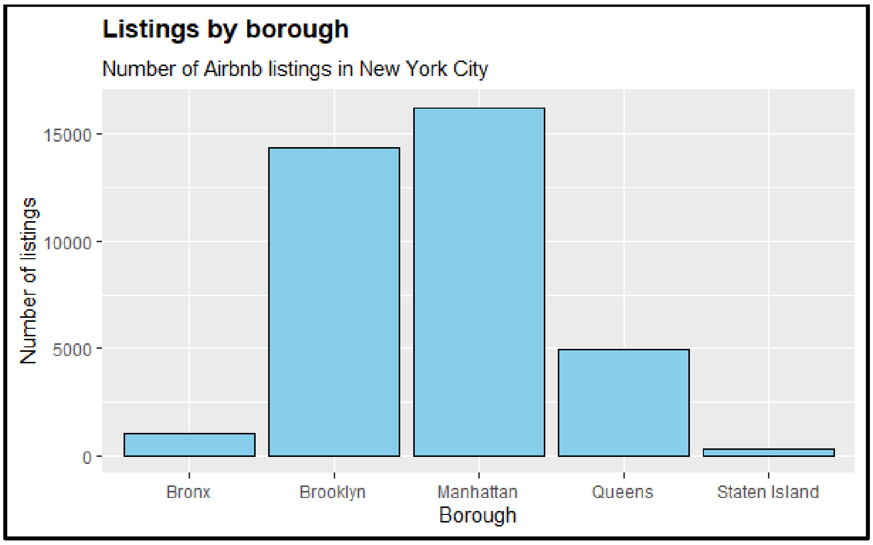


Graph 1: Distribution of New York City Airbnb prices

4.2. Explanatory variables

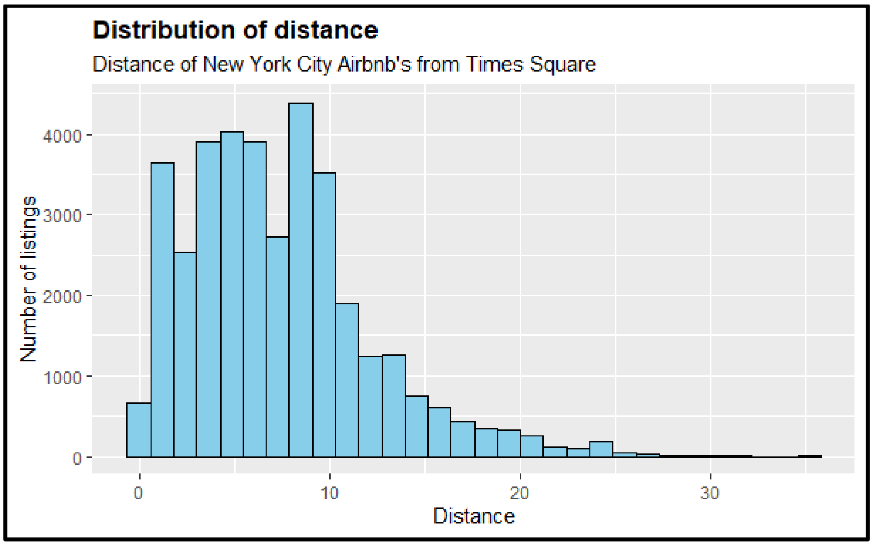
4.2.1. Distance

As described in the methodology section, our baseline model is that the distance of the accommodation from the city centre – Times Square in our case – has a large role in the formation of Airbnb prices. We expect that closer Airbnb's will be more popular since they are mainly used as holiday rentals. The borough of Manhattan is most centre located, and Brooklyn is the runner up. There are clearly more Airbnb listings in these areas, which indicates a higher demand for them. The grouping can be seen on the bar graph below.



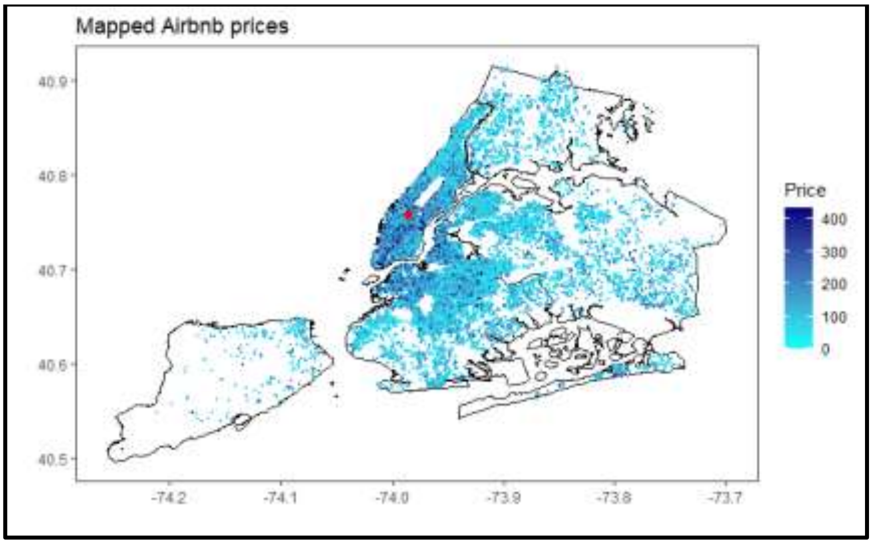
Graph 2: New York City Airbnb listings by borough

While the high presence of Airbnb listings in centre located boroughs is interesting and provides insight, it is not analytically useful in our research. Once we calculated the distance of the listing from Times Square, more accurate inferences could be drawn. The distribution of the distance is left peaked and right skewed. The majority of listings are within 10 kilometres from the chosen centre, and hardly any listings are found over 20 kilometres away. A histogram can be found below to visualize the distribution of the distance variable.



Graph 3: Distribution of the distance from the listing to Times Square

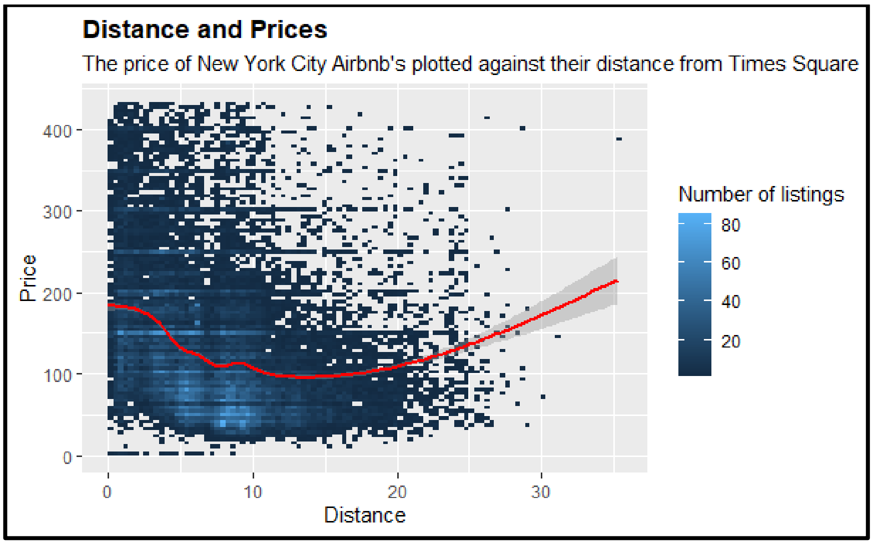
Considering our hypothesis that holidaymakers want to minimize their commute and consequently closer listings will be more popular, we can also infer that such accommodations will also be more expensive due to higher demand. The heatmap below presents this nicely, since there are more listings close to Times Square and the prices are also higher in general.



Graph 4: Heat map of Airbnb prices in New York City

Analysing the distances with respect to price, this is true for the closest 15 kilometres. The farther we get from the centre, the cheaper the listings. However, over 15 kilometres, this changes and prices begin to increase again. We can think of two explanations for this phenomenon. Firstly, over this given threshold we only have a few observations, which happen to be relatively expensive, thus bias our results due to small sample size. Secondly, listings farther from the centre are probably not cheap accommodations used for touristic reasons, but perhaps larger houses or villas with other redeeming qualities which dictate such high prices.

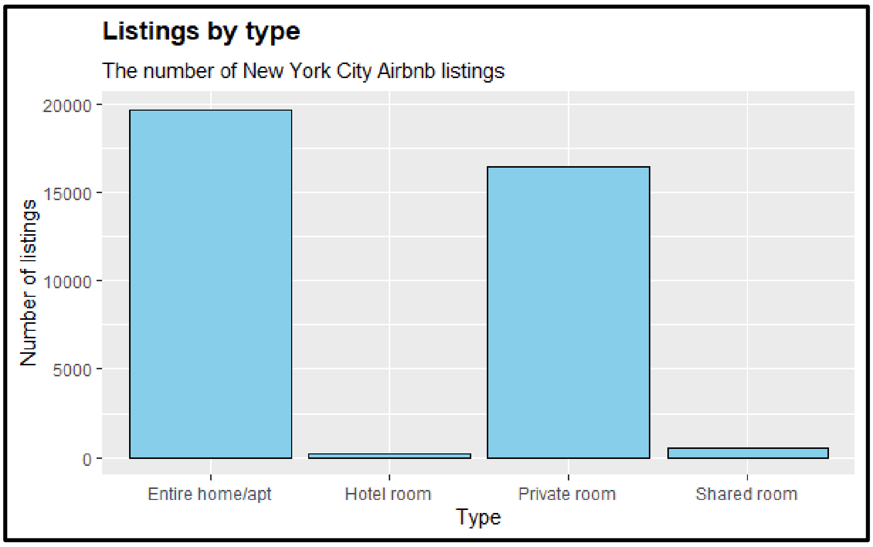
The graph below represents the relationship between distance from the centre and prices on a binned scatter plot. The points are represented with a lighter colour given a higher density of observations. Most listings are between 5 and 10 kilometres from the centre, with a slightly below average price. This shows that in the distance-price trade-off, the average consumer leans slightly towards a longer commuting time but a lower price. This can be explained by the fact that most renters cannot afford the highest prices dictated in the city centre. This should be kept in mind by letters creating their prices, and by us when creating our price predicting model.



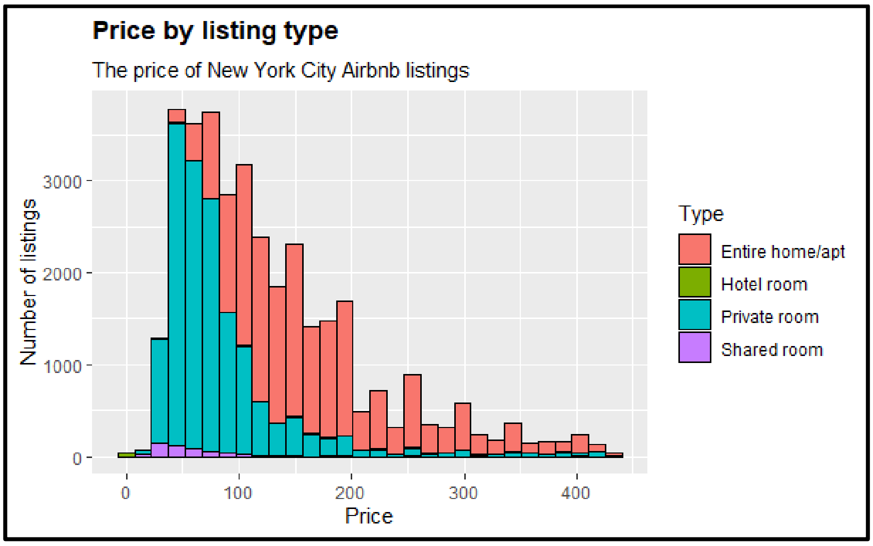
Graph 5: The relationship between price and distance from centre

4.2.2. Type of listing

Building our model further from the baseline, a further crucial aspect is the characteristics of the listing. Whether it is an entire apartment or simply a room, whether it fits 2 or more guests, and how many bathrooms the accommodation has makes all the difference. Therefore, these factors will also play a big role in our predictive model. In the previous section we concluded that the most popular listings are a little farther from the centre, thus a bit cheaper. We expect that such listings will be smaller, perhaps just rooms. To gain insight into the type of accommodations offered, and their distribution by price, consider the graphs below.



Graph 6: Number of Airbnb listings by type



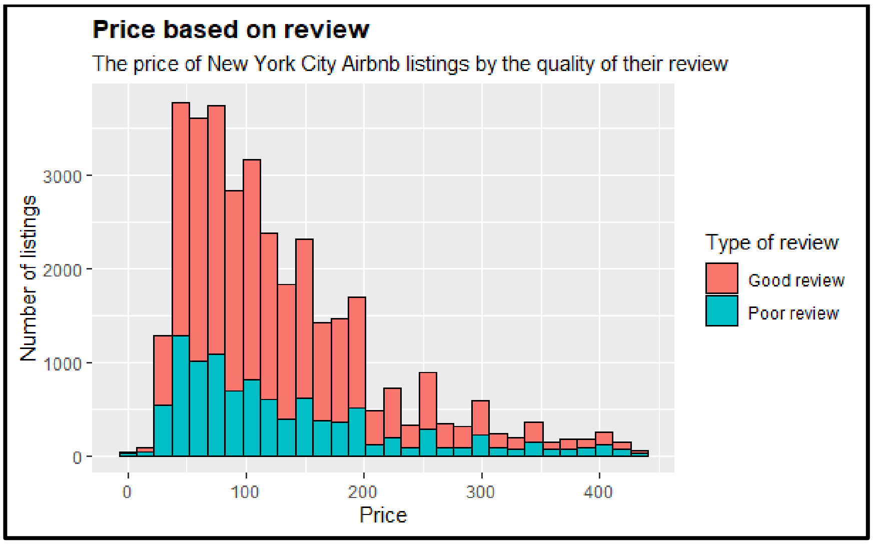
Graph 7: Price of Airbnb listing by its type

Clearly entire homes and apartments are most common, and they are also the most expensive. The higher price of entire homes is evident due to their nature compared to rooms, as they typically have more utilities and extra space. Additionally, they work in similar fashion to traditional real estate, therefore a larger portion of professional investors may be involved. Such professionals may set higher prices compared to amateurs, who are more likely to have only private rooms as Airbnb's. One of the goals of our price predicting model will be to close this gap, so armatures can gain as much as professionals through the Airbnb system.

4.2.3. Listing reviews

A great feature of the Airbnb platform is the opportunity to leave reviews. Former renters can voice their opinions on the listings they have visited, which is an accurate and generally reliable signal to future customers. If a listing or a host has good reviews, they may have the opportunity to increase prices, and vice versa. As a result, reviews are likely to be a good predictor for Airbnb prices.

Separating listings into two categories, those with good reviews – with an average rating over 4 out of 5 – and those with poor reviews, we conducted a t-test on their average price respectively. We found that listings with good ratings had a higher price, significant at the 1%. Also delineating the distribution of prices by the two review groups separately, a clear difference can be seen.



Graph 8: Price of Airbnb listing by quality of review

4.2.4. Host related features

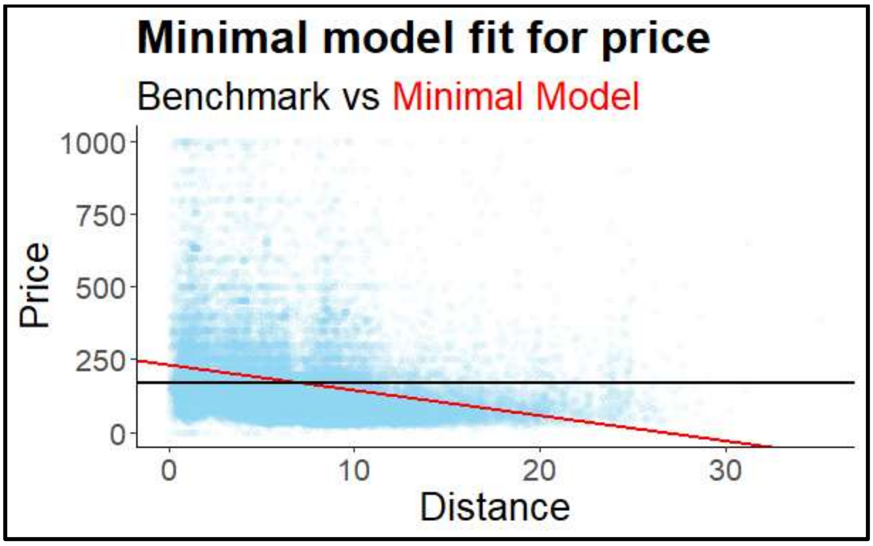
We also had ample variables about the characteristics of Airbnb hosts. Intuitively, we believe that if a host has better characteristics, they will be more trustworthy, leading to a higher demand for their listings and a consequential higher price. We determined 4 host-level variables which could be useful predictors in our model. How long someone has been a host, what percentage of users the host accepts into their Airbnb, whether the host is a professional based on the number of their listings, and the quality of their listing description, which we calculated by assigning Bing sentiment scores to each description.

To gain insight into the relationship between host related variables and the price of the listing, we ran a simple linear regression with the price as the dependent variable and the previously described 4 variables as explanatory variables. All variables were significant at the 5% level. Price increases significantly with the acceptance rate, professionalism, and the quality of the description. Interestingly, the longer an individual has been a host, the lower the price of the listing. However, finding adequate reasoning for this is out of the scope of this paper. The output table for the linear regression can be found in Appendix 1.

We ran some further linear regressions, playing around with the different variables described in the past sections of our paper. The results of these are not directly relevant to our price predicting model, but they are interesting for insight, nevertheless. Therefore, the output table can be found in Appendix 2. Each coefficient should be interpreted as the marginal change in price given a unit change in the given explanatory variable, all things held constant.

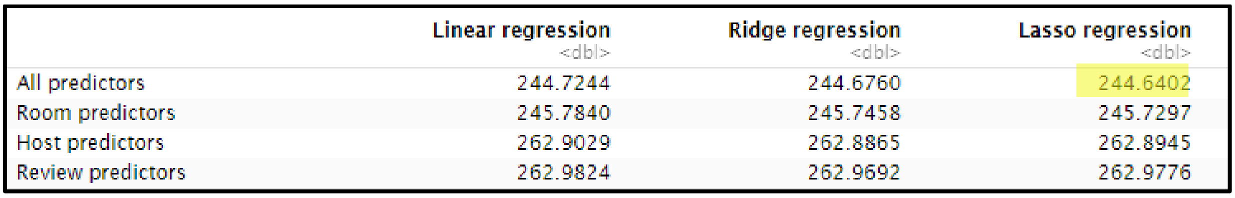
4.3. Predictions

To begin our investigation, we trained a baseline prediction model with a simple linear regression with price as the dependent variable and distance from the centre as the explanatory variable. This prediction model had a RMSE of 296.29, which we want to minimize. Plotting price against distance and adding our regression line, we can get an overall idea of the direction of our data, but the linear model does not fit the data points well.



Graph 9: Minimal baseline model for prediction

Therefore, to obtain the minimal RMSE from our given variables and make the most accurate prediction model possible, we explored numerous model types Firstly, we separated our dataset into a test and train dataset. This is a rational step considering the abundance of observations. Then we defined four different model formulae, all with price as the dependent variable: a model with all explanatory variables, a model with only listing related variables such as the number of rooms, a model with host-related characteristics such as whether the host is a superhost, and a model with review related variables including the sentiment score of the description. Next, we trained our predictive models with 3 types of regressions: linear regression, ridge regression, and lasso regression. It is important to mention that in the case of regularization, we used cross-validated regressions to obtain models with the best lambdas. Then, we made our predictions. As a result, we obtained 12 different prediction models, which we then compared based on their root mean squared error. This can be done based on the table below.

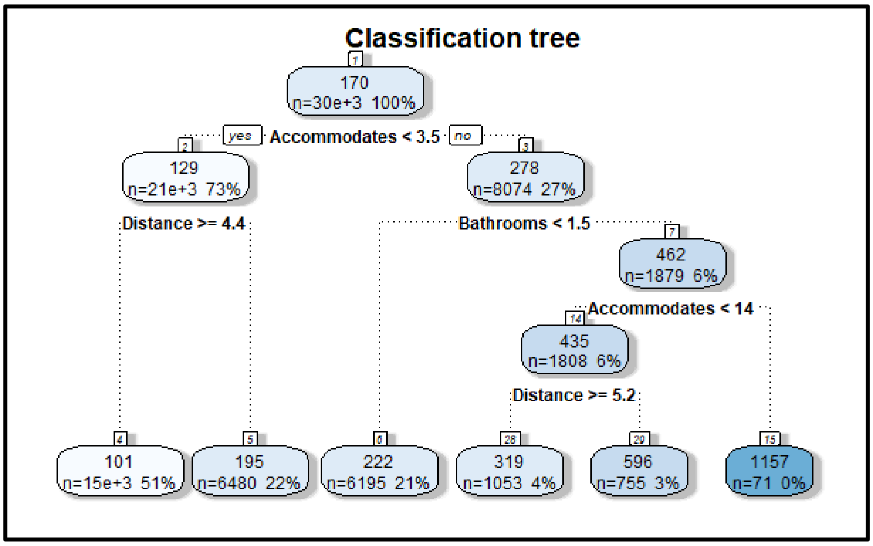


	Linear regression	Ridge regression	Lasso regression
	<dbl>	<dbl>	<dbl>
All predictors	244.7244	244.6760	244.6402
Room predictors	245.7840	245.7458	245.7297
Host predictors	262.9029	262.8865	262.8945
Review predictors	262.9824	262.9692	262.9776

Graph 10: Root mean squared error of our 12 price prediction models

Examining the above table, the three regression types conclude very similar root mean squared errors within each model formula. The lowest RMSE can be obtained by including all chosen explanatory variables. While the cross-validated lasso regression has the lowest RMSE at present, since the three models provided such similar results – with 244.72, 244.68, and 244.64 respectively –, we must wonder whether we can reduce this measure even further.

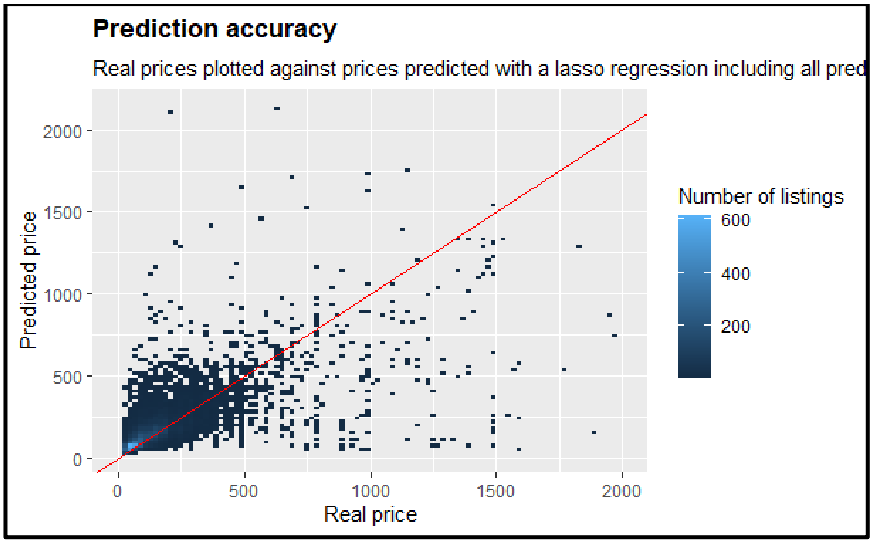
Hoping to obtain an even better model, we trained a prediction model with random forest. Random forest models are essentially a collection of decision trees with aggregated results. To provide an example of a decision tree, we created one based on our explanatory variables. This can be found below. Similar trees can be made for further insight and analysis; however, these are out of scope at present.



Graph 11: An example of a decision tree based on our variables

Since we have a very large dataset, the running time of a random forest model is crucially slow. Therefore, we decided to limit the tried models at one, and to restrict the number of created decision trees at 100. Since the formula including all explanatory variables clearly led to the lowest RMSE in all previous models, this is the framework we used here also. This model resulted in a root mean squared error of 233.25, which is the smallest yet. Therefore, we concluded that the price prediction model based on the random forest including all explanatory variables as predictors is our model of choice.

It is also important to consider the circumstances under which our chosen model predicts accurately, or perhaps less so. To analyse this, we plotted the predicted prices against the observed prices to see where they are equivalent.



Graph 12: Predicted price and real price

Based on the above graph, we can say that the model works best for lower prices. This can be seen in the lighter blue area (indicating a higher density), as ideally, the data points should be along the red 45° line. While the prediction model is not perfect, we argue that it is good enough for an average Airbnb user, since typically new-comers or amateurs will not start their careers with high-value real estate. Therefore, this model gives a relatively accurate guideline to new, average hosts who are uncertain about the price they should list.

5. Business Implications

As concluded, the best model to predict Airbnb prices is a random forest model using relevant observable variables, such as the distance from the city centre, the accommodation type and what it includes, the characteristics of the host, and available reviews. An exhaustive list of the relevant predictors we used can be found in our R code, based on which our model is easily replicable for further use on similar databases.

We recommend this model to all Airbnb beginners who are uncertain about the price they should list on their accommodation. With the help of our model, in a few simple steps, a host can fill out simple observable characteristics regarding their given space, and within minutes, an accurate price can be concluded. This can help all users be as efficient as professionals.

6. Conclusion

New York City has all the characteristics to be a useful base for our Airbnb price prediction model. There are ample residences with many different characteristics. The city also has a landmark centre, which attracts tourists. From our dataset we gathered multiple variables which can help us accurately predict the prices of Airbnb's. We found the most important predictor to be the distance from the city centre, and further variables were categorized into multiple groups: listing, host, and review related variables. Through the examination of numerous predictive models, we finally found the most efficient version, the random forest predictive model including all explanatory variables. Through this model, Airbnb hosts, especially armatures, can easily kickstart their career by obtaining a price guideline for their new listing within minutes. As a result, anyone can be a professional and collect many gains from the popular Airbnb platform.

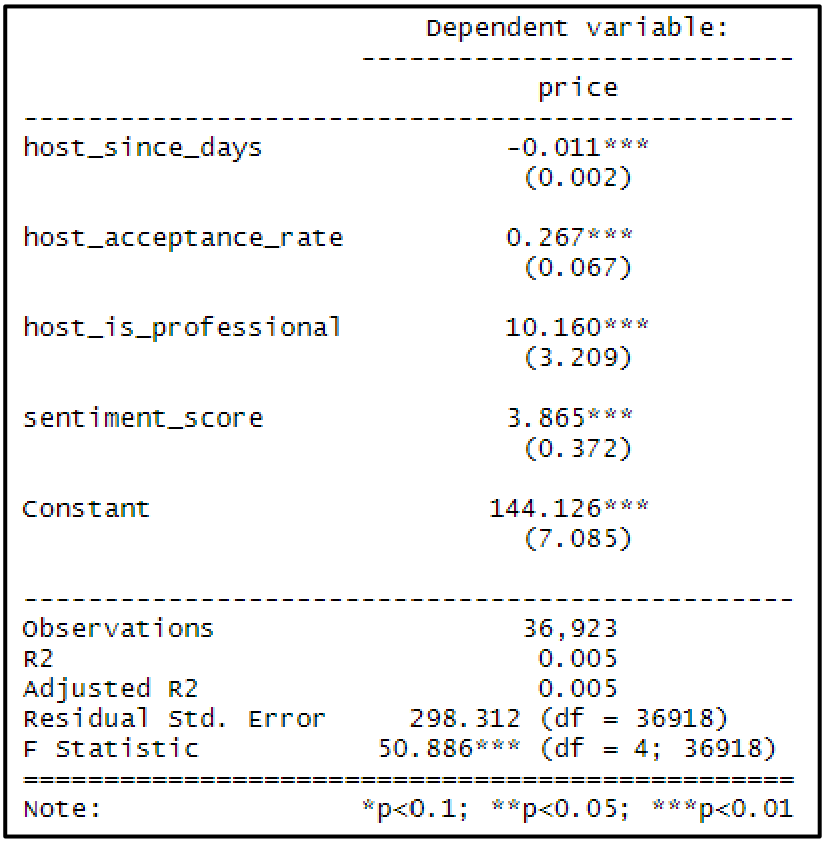
References

Airbnb. (2021, June 30). *About us.* Retrieved from Airbnb: https://news.airbnb.com/about-us/

Inside Airbnb. (2021). Get the Data. Retrieved from Inside Airbnb: http://insideairbnb.com/get-the-data.html

Appendix

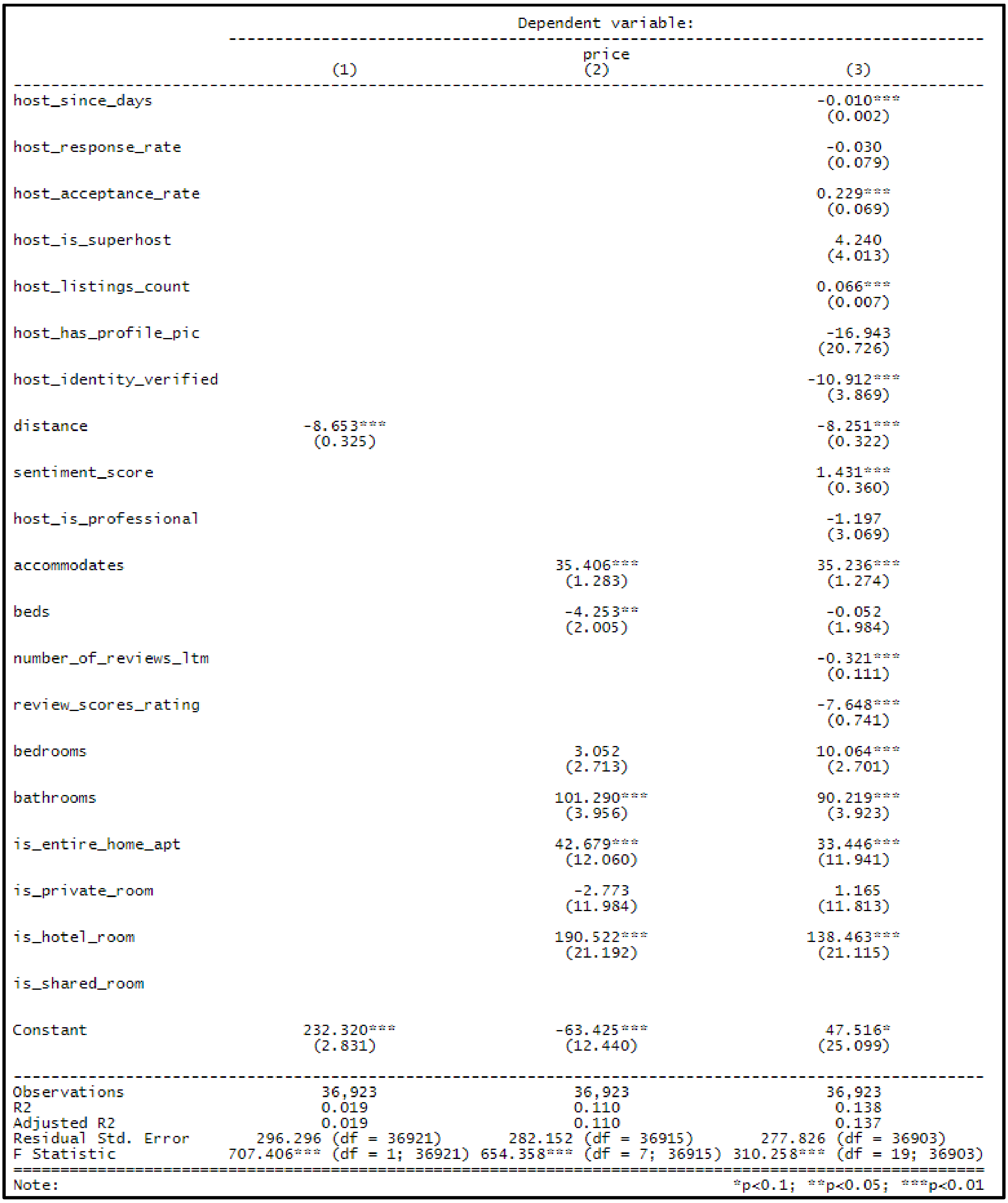
Appendix 1.



```
                           Dependent variable:
                       ----------------------------
                                  price
-----------------------------------------------
host_since_days                  -0.011***
                                  (0.002)

host_acceptance_rate              0.267***
                                  (0.067)

host_is_professional             10.160***
                                  (3.209)

sentiment_score                   3.865***
                                  (0.372)

Constant                        144.126***
                                  (7.085)

-----------------------------------------------
Observations                      36,923
R2                                 0.005
Adjusted R2                        0.005
Residual Std. Error       298.312 (df = 36918)
F Statistic            50.886*** (df = 4; 36918)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Appendix 1: Linear regression with prices explained by host characteristics

Appendix 2.



```
                                              Dependent variable:
                       -----------------------------------------------------------------------------
                                                     price
                                  (1)                      (2)                        (3)
--------------------------------------------------------------------------------------------------
host_since_days                                                                   -0.010***
                                                                                  (0.002)

host_response_rate                                                                -0.030
                                                                                  (0.079)

host_acceptance_rate                                                               0.229***
                                                                                  (0.069)

host_is_superhost                                                                  4.240
                                                                                  (4.013)

host_listings_count                                                                0.066***
                                                                                  (0.007)

host_has_profile_pic                                                             -16.943
                                                                                 (20.726)

host_identity_verified                                                           -10.912***
                                                                                  (3.869)

distance                        -8.653***                                         -8.251***
                                (0.325)                                           (0.322)

sentiment_score                                                                    1.431***
                                                                                  (0.360)

host_is_professional                                                              -1.197
                                                                                  (3.069)

accommodates                                          35.406***                   35.236***
                                                      (1.283)                     (1.274)

beds                                                  -4.253**                    -0.052
                                                      (2.005)                     (1.984)

number_of_reviews_ltm                                                             -0.321***
                                                                                  (0.111)

review_scores_rating                                                              -7.648***
                                                                                  (0.741)

bedrooms                                               3.052                      10.064***
                                                      (2.713)                     (2.701)

bathrooms                                            101.290***                   90.219***
                                                      (3.956)                     (3.923)

is_entire_home_apt                                    42.679***                   33.446***
                                                     (12.060)                    (11.941)

is_private_room                                       -2.773                       1.165
                                                     (11.984)                    (11.813)

is_hotel_room                                        190.522***                  138.463***
                                                     (21.192)                    (21.115)

is_shared_room


Constant                        232.320***           -63.425***                   47.516*
                                (2.831)              (12.440)                    (25.099)

--------------------------------------------------------------------------------------------------
Observations                     36,923               36,923                      36,923
R2                               0.019                0.110                       0.138
Adjusted R2                      0.019                0.110                       0.137
Residual Std. Error       296.296 (df = 36921)   282.152 (df = 36915)   277.826 (df = 36903)
F Statistic           707.406*** (df = 1; 36921) 654.358*** (df = 7; 36915) 310.258*** (df = 19; 36903)
==================================================================================================
Note:                                                                 *p<0.1; **p<0.05; ***p<0.01
```

Appendix 2: Linear regressions with price as the dependent variable and varying explanatory variables