Analysis of Airbnb Prices in Amsterdam

Eric Cheng, Nina Sun, Jeanny Zhang

Introduction

Prior to the Covid-19 pandemic, Amsterdam along with many other major European hubs had a bustling tourism industry. In 2019, city officials estimated that 20 million people had visited the city. This is despite the fact that the city has less than a million inhabitants (Kakissis). Amsterdam has joined the ranks of European cities debating the merits and cons of tourism. This debate has made a clear appearance in the Dutch legal system.

The Amsterdam city council in an effort to combat over tourism recently, introduced a ban against short term Airbnb rentals in its city center. This mandate however did not go unnoticed by the city's homeowners associations however. In this case, the court ruled that this ban must be lifted as it has no legal basis. The city in response now allows homeowners to apply for the same permit to rent their primary residence for up to 30 days a year for groups no larger than 4.

Despite the overturning of this ban, this skirmish illustrates not only the sheer amount of tourism in Amsterdam but also the controversies it has brought. In a city where officials estimate that roughly one in fifteen dwellings are placed on rental platforms such as Airbnb, there is undoubtedly a bustling rental market (Reuters Staff). With this highly active rental market in mind, this paper hopes to address the factors that predict the price of rental bookings in Amsterdam. We especially were interested in seeing if distance from the city affected the price of airbnb bookings in Amsterdam. This would help us shed light on the debate in Amsterdam regarding banning airbnb's in the city center. If there is no strong relationship between pricing and distance from the center we may support the city's actions.

Based on information found on Zenodo, we analyzed multiple factors that have the potential to influence rental booking pricing. These include nightly pricing of Airbnbs bookings for two nights for two people as a function of room type, room capacity, if the host is a super host, if the host owns 2-4 or over four properties, guest dissatisfaction,number of bedrooms, distance from city center, distance from the metro, attraction index of location, and restaurant index of the location. We used these factors to achieve the following:

1.Identify statistically significant variables that impact the pricing of rental properties-especially distance from city center.

2.Create an MLR to identify the effects of these significant variables on price while other factors are held constant.

Data

This data was taken from zenodo, an open access repository developed by the European OpenAire program to foster research in the European Research Area. This specific repository is managed by CERN also known as the European Organization for Nuclear Research. This specific data set is titled "Determinants of Airbnb Prices in European Cities: A Spatial Econometrics Approach" (https://zenodo.org/record/4446043#.YZhxj9nMLh8). It was submitted to zenodo by researchers Łukasz, Nawaro and Kristóf Gyódi of the University of Warsaw. Data was collected over the year 2019 for Airbnb bookings within the city boundaries of 10 commonly visited European cities four to six weeks prior to the intended day of travel. For each city, the total price of booking (including reservation and cleaning fees) for two nights were compiled into two data sets (one for weekends and one for weekday stays). We choose to analyze the data for Amsterdam's weekend bookings.

There were a total of 977 bookings possible bookings in the data set, with net price in Euros as a function of eighteen potential explanatory variables. Note that we did not include the longitude and latitude in our analysis of the data, this will be explained further in our analysis. In addition, we did not fit normalized restaurant index and normalized attraction index given their redundancy. The variables are as the following:

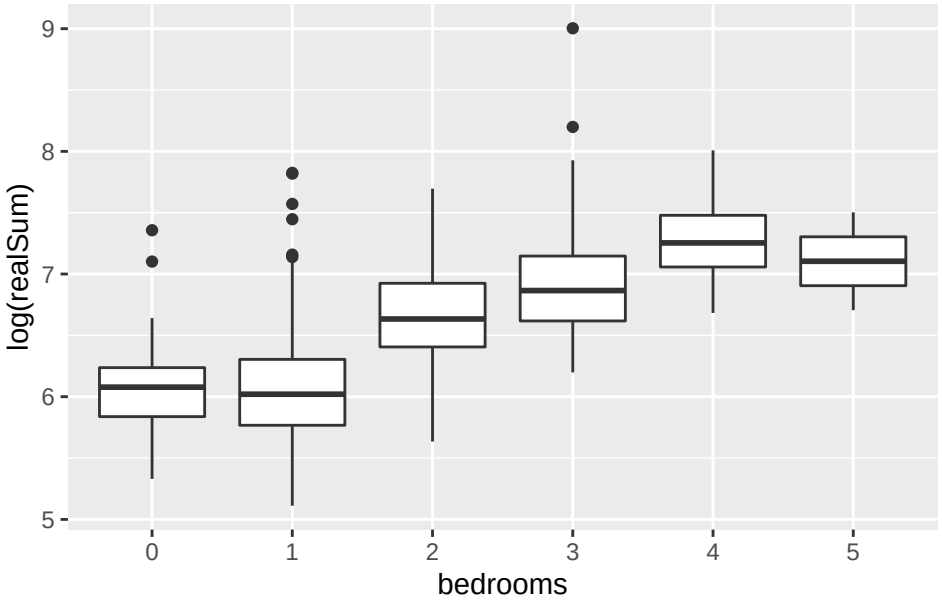
- realSum: the full price of accommodation for two people and two nights in EUR
- room_type: the type of the accommodation
- room_shared: dummy variable for shared rooms
- room_private: dummy variable for private rooms
- person_capacity: the maximum number of guests
- host_is_superhost: dummy variable for superhost status
- multi: dummy variable if the listing belongs to hosts with 2-4 offers
- biz: dummy variable if the listing belongs to hosts with more than 4 offers
- cleanliness_rating: cleanliness rating
- guest_satisfaction_overall: overall rating of the listing
- bedrooms: number of bedrooms (0 for studios)
- dist: distance from city centre in km
- metro_dist: distance from nearest metro station in km
- attr_index: attraction index of the listing location. Calculated via the following formula $attr\_index_j = \sum_{k=1}^{K} \frac{R_k}{d_{jk}}$. R is the number of reviews for an attraction k, divided by djk which represents the distance between the booking number j and attraction k.
- attr_index_norm: normalised attraction index (0-100). Attraction index for each booking is divided by the largest possible value for each city and multiplied by 100. Thus this is scale from 0-100 based on the largest attraction value.
- rest_index: restaurant index of the listing location. Calculated via the same method as the attraction index.
- attr_index_norm: normalised restaurant index (0-100)
- lng: longitude of the listing location
- lat: latitude of the listing location

Results

Our analysis of the Airbnb data set follows the standard order of finding a best fitting multiple linear regression model. The general steps are described as the following: exploratory data analysis, check for model transformation, fit a "rich" model, model assumptions, check for outliers, check for collinearity, remove statistically insignificant terms, and repeat the process above when necessary.

We first conducted an exploratory data analysis on the Airbnb data set and checked if we needed to transform any of the predictors. The response variable, which is the Airbnb price, was plotted against each explanatory variable. Some of the categorical explanatory variables were recognized by R as numeric variables, and we changed them into factor variables in order to fit the real situation better. For instance, the variable `bedrooms` was recognized as a numeric variable and it was not statistically significant in the model. However, after converting it to a factor variable, it is now statistically significant in the model. Figure 1 shows the boxplot of the log of Airbnb prices versus different bedroom sizes. Scatterplots were plotted for quantitative explanatory variables and boxplots were plotted for categorical explanatory variables.

Figure 1. log of realSum versus bedrooms



Based on Figure 2, 3, 4, 5, and 6, we found that the following variables need a logarithmic transformation: the overall rating of the listing (`guest_satisfaction_overall`), the attraction index of the listing location (`attr_index`), the normalized attraction index (`attr_index_norm`), the restaurant index of the listing location (`rest_index`), the normalized restaurant index (`rest_index_norm`), and the price of the Airbnb (`realSum`).

Due to the fact that the data for the variable `guest_satisfaction_overall` are clustered on the right side of the graph ranging from 80 to 100, a simple logarithmic transformation does not make the data satisfying the linearity assumption. We calculated the "guest dissatisfaction rating" by computing $101 - guest\_satisfaction\_overall$ since the guest satisfaction rating has a range of 0-100, and we used 101 instead of 100 to avoid $log(0)$. We then logged the "guest dissatisfaction rating" and plotted the response variable against it, yielding a graph demonstrating a more linear relationship between the predictor and the response variable.

We also found the following variables needing a square root transformation as shown in Figure 7 and 8: the distance from the city center in kilometers (`dist`) and the distance from the nearest metro station in kilometers (`metro_dist`).

The other variables remain the same form. However, we found the longitude of the listing location (`lng`) and the latitude of the listing location (`lat`) cannot be explained well by a multiple linear regression. In the city of Amsterdam, there does not seem to be a linear relationship between the longitude and latitude of a listing location and its Airbnb price. Therefore, we decided to remove these two variables from our model.

Figure 2. log of realSum versus log of guest dissatisfaction rating

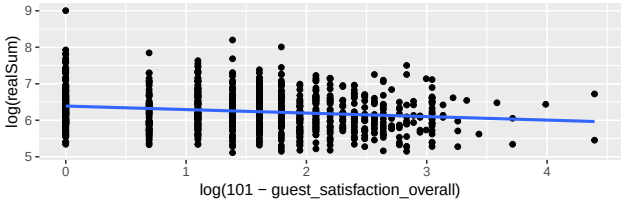


Figure 3. log of realSum versus log of attraction index

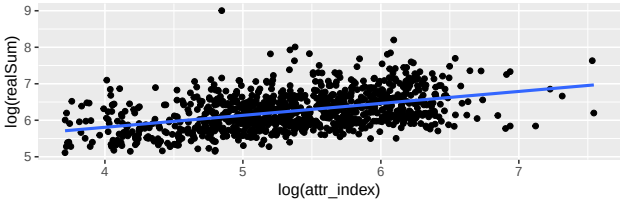


Figure 4. log of realSum versus log of normalized attraction index

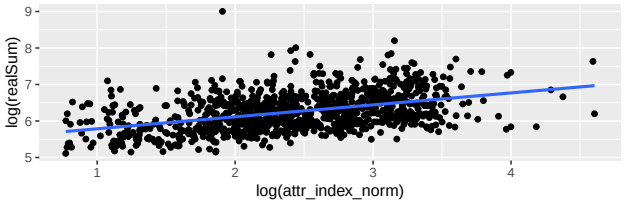


Figure 5. log of realSum versus log of restaurant index

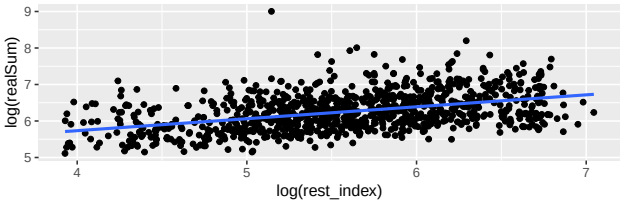


Figure 6. log of realSum versus log of normalized restaurant index

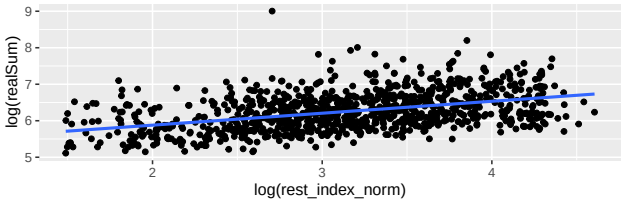


Figure 7. log of realSum vs sqrt of distance from city center in km

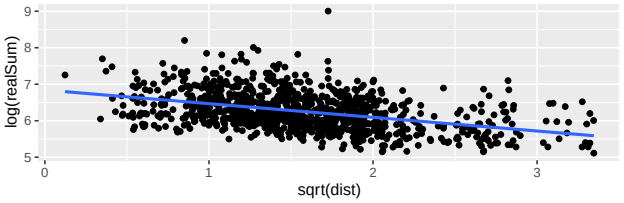
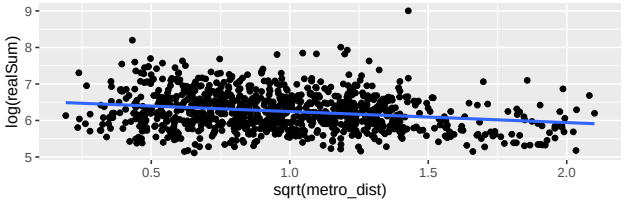


Figure 8. log of realSum vs sqrt of distance from metro station in km



A "rich" model with all transformed terms was fitted. We noticed that the estimated coefficients for the dummy variable for shared rooms (`room_shared`), the dummy variable for private rooms (`room_private`), the logged normalized attraction index (`attr_index_norm`), and the logged normalized restaurant index (`rest_index_norm`) are null in the model summary. This indicates that these variables are linearly related to other variables, so we decided to remove all four of them from the model.

We checked the resulting model's assumptions by plotting its residual plot, QQ plot, and partial residual plot. We found that linearity and constant variance are not violated as all residual plots are randomly scattered and there is no curvature in sight. The QQ plot suggests a slightly longer right tail because it is concave up. However, the sample size n is big enough (almost 1000). Therefore, we can still assume normality. Even though we have spatial information on latitude and longitude, we lack the information on and ability to assess whether the spatial correlation has a significant effect on the response. Therefore, we lack the information on judging if the independence assumption is violated. We will discuss the independence assumption of our model in further detail in discussion.

We then checked the resulting model's outliers by looking at the Cook's distance of each data point. We did not find any data point with a Cook's distance value over 1, and the highest Cook's distance is only around 0.11. However, we found out that the categorical variable `bedrooms` only contains two data points for the level of five bedrooms. We decided to remove both of the data points, case 45 and case 975, from our model as the sample size is too small to have any statistical significance. As a result, the *bedrooms*5 term was removed from the model. We refitted the model again omitting the other potentially influential cases, and none of them turned out to be influential outliers.

The collinearity of the resulting model was checked, and we found a high collinearity between the square root of distance from city center in kilometers (`dist`), the logged attraction index of the listing location (`attr_index`), and the logged restaurant index of the listing location (`rest_index`). The variance inflation factors calculated were 12.346 for `sqrt(dist)`, 23.807 for `log(attr_index)`, and 32.702 for `log(rest_index)`. We first did an ANOVA test comparing models with and without the `dist` variable, which turned out to be statistically insignificant. Thus, we removed the `dist` variable from the model. The resulting model still exhibited a high collinearity between `log(attr_index)` and `log(rest_index)` with VIFs high as 23.586 and 23.444. We then did an ANOVA test comparing models with and without the `attr_index`, which turned out to be statistically insignificant. Thus, we removed the `attr_index` variable from the model too. The resulting model does not exhibit a high collinearity between explanatory variables anymore.

We eventually attempted to remove all statistically insignificant predictors from the model by conducting an ANOVA test. We created a reduced model removing the dummy variable for superhost status (`host_is_superhost`), the dummy variable if the listing belongs to hosts with 2-4 offers (`multi`), the dummy variable if the listing belongs to hosts with more than 4 offers (`biz`), the cleanliness rating (`cleanliness_rating`), and the square root of the distance from nearest metro station in kilometers (`sqrt(metro_dist)`). We used ANOVA to compare the reduced model and the "rich" model, and we yielded a f-statistic of 1.173 and a p-value of 0.301, indicating that all of the removed variables are not statistically significant.

We checked the resulting model's assumptions again. Linearity and constant variance are not violated. We can also still assume normality since the sample size is large enough and the cases removed don't change the scale of the data. We also lack information on judging if independence is violated.

We also checked the resulting model's outliers again by looking at Cook's distance. There was no Cook's distance larger than 1 in this model. The highest Cook's distance was about 0.11. We refitted the model without the top five cases with the highest Cook's distance, and the refitted model demonstrated no significant change compared to the old one.

Lastly, the collinearity of the resulting model was checked, and we found no significant collinearity between the explanatory variables. Therefore, the best-fitting multiple linear regression model we found for estimating Airbnb prices in Amsterdam is shown below. Table 1 shows the summary for our fitted multiple linear regression model.

$$\begin{aligned}\widehat{log(realSum)} = {} & \beta_0 + \beta_1 room\_typePrivateroom + \beta_2 room\_typeSharedroom + \beta_3 person\_capacity \\ & + \beta_4 log(101 - guest\_satisfaction_overall) \\ & + \beta_5 bedrooms1 + \beta_6 bedrooms2 + \beta_7 bedrooms3 + \beta_8 bedrooms4 + \beta_9 log(rest\_index)\end{aligned}$$

$$\begin{aligned}\hat{med}(realSum) = {} & e^{\beta_0 + \beta_1 room\_typePrivateroom + \beta_2 room\_typeSharedroom + \beta_3 person\_capacity} \\ & \times e^{+\beta_4 log(101 - guest\_satisfaction_overall) + \beta_5 bedrooms1 + \beta_6 bedrooms2 + \beta_7 bedrooms3 + \beta_8 bedrooms4 + \beta_9 log(rest\_index)}\end{aligned}$$

Table 1. Model Statistics

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.594	0.093	49.366	0.000	4.411	4.776
room_typePrivate room	-0.324	0.021	-15.610	0.000	-0.365	-0.283
room_typeShared room	-0.630	0.147	-4.298	0.000	-0.917	-0.342
person_capacity3	0.199	0.041	4.888	0.000	0.119	0.279
person_capacity4	0.328	0.031	10.709	0.000	0.268	0.388
person_capacity5	0.536	0.096	5.554	0.000	0.347	0.725
person_capacity6	0.775	0.084	9.196	0.000	0.610	0.940
log(101 - guest_satisfaction_overall)	-0.044	0.010	-4.366	0.000	-0.064	-0.024

term	estimate	std.error	statistic	p.value	conf.low	conf.high
bedrooms1	0.051	0.039	1.329	0.184	-0.024	0.127
bedrooms2	0.226	0.048	4.667	0.000	0.131	0.320
bedrooms3	0.458	0.062	7.429	0.000	0.337	0.579
bedrooms4	0.769	0.107	7.202	0.000	0.560	0.979
log(rest_index)	0.287	0.014	20.563	0.000	0.260	0.315

The estimated intercept of our model is 4.105 with a standard error of 0.085. The private room type is associaed with an estimated 27.675% decrease on the median Airbnb price compared to the full house/apartment room type (95% CI 24.648% to 30.580%), holding all other predictors fixed. The shared room type is associated an estimated 46.741% decrease on the median Airbnb price compared to the full house/apartment room type (95% CI 28.965% to 60.028%), holding all other predictors fixed. Airbnbs with person capacity of 3 is associated with a 22.018% increase on median Airbnb price compared to person capacity of 2 (95% CI 12.637% to 32.181%), holding all other predictors fixed. Airbnbs with person capacity of 4 has 38.819% increase on median Airbnb price compared to person capacity of 2 (95% CI 30.734% to 47.403%), holding all other predictors fixed. Airbnbs with person capacity of 5 is associated with a 70.916% increase on median Airbnb price compared to person capacity of 2 (95% CI 41.482% to 106.473%), holding all other predictors fixed. Airbnbs with person capacity of 6 is associated with an estimated 117.0592% increase on median Airbnb price compared to person capacity of 2 (95% CI 84.043% to 155.998%), holding all other predictors fixed. A doubling of guest dissatisfaction overall score is associated with an decrease in the median Airbnb price by an estimated 3.00381% (95% CI 1.650% to 4.339%), holding all other predictors fixed. The Airbnbs with two bedroom is associated with an estimated 25.358% increase on median Airbnb price compared to Airbnbs with no bedroom (95% CI 13.997% to 37.713%), holding all other predictors fixed. The Airbnbs with three bedroom is associated an estimated 58.091% increase on median Airbnb price compared to Airbnbs with no bedroom (95% CI 40.073% to 78.425%), holding all other predictors fixed. The Airbnbs with four bedroom is associated with an estimated 115.761% increase on median Airbnb price compared to Airbnbs with no bedroom (95% CI 75.068% to 166.179%), holding all other predictors fixed. A doubling of restaurant index is associated with an estimated 22.010% increase effect on median Airbnb price (95% CI 19.748% to 24.401%), holding all other predictors fixed.

Discussion

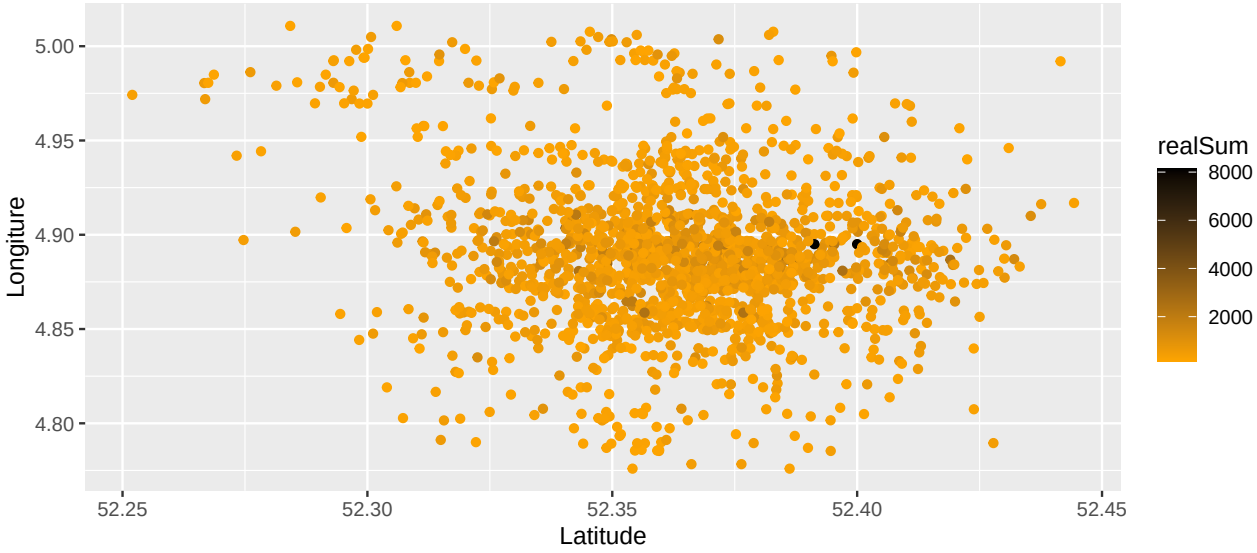
Through the creation of our multiple linear regression model, we found that the mean of the log of total price is linearly related to the following factors: the type of accommodation (`room_type`), the maximum number of guests (`person_capacity`), the log of the overall rating of guest dissatisfaction (`log(101-guest_satisfaction_overall)`), the number of bedrooms (`bedrooms`), and the log of restaurant index of the listing location (`log(rest_index)`). We thus have an exponential model relationship between the estimated median of realSum and the explanatory variables `room_type`, `person_capacity`, and `bedrooms`. In addition, we have a power model relationship between the estimated median of realSum and the restaurant index as well as guest dissatisfaction. The relationship of our significant predictors and realSum confirmed our previous beliefs. For instance, we had a decrease in price when going from a full apartment to a private room and a shared room. This is likely due to guests valuing personal space and being willing to pay more for it. In addition we had a negative relationship for our guest dissatisfaction, which likewise seems plausible as demand for an unsatisfying apartment would likely be low. Our positive relationships between the estimated median of realSum and the explanatory variables `person_capacity`, `bedrooms`, and `log(rest_index)` appear reasonable. We can thus conclude that host pricing for their booking is a function of privacy of the booking, space, and the diversity of restaurant options in the vicinity.

One surprising factor that did not appear significant in our final model was the distance from the city center. Given the concentration of tourist activity in Amsterdam's as well as the high amount of controversy Airbnb's in that location has brought, we expected this variable to have a significant impact on price. This may justify the city's ban on Airbnbs in the city center. If there is no increase in revenue for bookings in the city center,

the positive externality/social benefit brought by a potential ban may offset revenue lost. In addition, we were somewhat surprised that metro distance did not impact pricing of Airbnbs. We hypothesize that this may be due to the well developed biking infrastructure of Amsterdam may disincentivize visitors from taking the metro.

We also have a problem examining the independence assumption of our model. By examining the scatterplots of the Airbnb price versus the longitude and latitude, we noticed that most of the data points are concentrated in the center of the plot with higher Airbnb prices. This could mean that there are more Airbnb listings in the city center of Amsterdam and their prices are higher than the other locations. We then created a scatter plot (Figure 9) with longitude versus latitude while displaying the Airbnb price of each data point in colors. This plot shows a higher density of data points with higher Airbnb prices in the center of the graph than other locations. Although this plot suggests a possible spatial correlation between Airbnb price and location, we would need a spatial model to explore the true relationship between the Airbnb prices and the longitude and the latitude of the location, as testing the significance of the spatial correlation exceeds our capacity.

Figure 9. Latitude vs Longitude faceted by Airbnb Price



Conclusion

In summary, there are a plethora of factors that influence the net price of an Airbnb in Amsterdam. Through our model analysis, we determined that these factors were `room_type`, `person_capacity`, `log(101-gues_satisfaction_overall)`, `bedrooms`, `log(rest_index)`. The relationship between these variables and the cost of a booking confirmed our previous assumptions. To our surprise however, we found that distance from the city center was not a significant variable in our multiple linear model. Given the high amount of tourist activity and attractions in the city center as well as recent government action to limit tourist congestion in that region, this finding defied our assumptions. If this is true, the social benefit of a ban may outweigh the revenue lost from bookings in the center of Amsterdam. A further step to confirm this would be to test a spatial model.

Bibliography:

Kakissis, Joanna. "In Amsterdam, Even the Tourists Say There Are Too Many Tourists." NPR, NPR, 7 Aug. 2018, https://www.npr.org/2018/08/07/632012775/in-amsterdam-even-the-tourists-say-there-are-too-many-tourists.

Gyódi, Kristóf, and Łukasz Nawaro. "Determinants of Airbnb Prices in European Cities: A Spatial Econometrics Approach." Tourism Management, Pergamon, 8 Apr. 2021, https://www.sciencedirect.com/science/article/pii/S0261517721000388?via%3Dihub#appsec1.

Reuters Staff. "Amsterdam to Allow Airbnb Rentals in City Centre after Court Order." Reuters, Thomson Reuters, 16 Mar. 2021, https://www.reuters.com/article/us-netherlands-airbnb-amsterdam-idUSKBN2B81NS.