National Research University

Higher School of Economics

Faculty of Computer Science

Econometric study: determinants of Airbnb Prices in Paris

Made by students:

Sergazin Iskander, Spirina Mayya, Tomaily Tatiana

Course title: Econometrics

Supervisor: Stankevich Ivan Pavlovich

Moscow, 2023

Table of contents

1 Introduction

In recent years, the tourism industry has been reshaped by platforms like Airbnb which allow users to rent flats, dormitories and other types of property overnight. Airbnb property pricing is challenging because each listing is unique and presents a multitude of factors that may influence the overall experience of the customer. At the same time, determining the optimal price for the listing remains essential for the landlords. By understanding connections between various factors and final price, hosts can establish correct pricing. Having insight into what consumer preferences are and what contributes to the price will help hosts to find out which services and room characteristics are worth improving to get extra profit.

Econometric tools are well-designed to address the aforementioned task, therefore, in this study we will use regression analysis to support our investigation of price determinants on the Airbnb platform in Paris at the weekends.

We first review the literature on related studies, and in section 2, we examine the source of the Kaggle dataset and provide important visual data. The 3rd part will then cover the rationale behind the selected model and hypothesis. The results interpretation, variable discussion, and compatibility with the original hypothesis will all be covered in the upcoming chapter. Finally, we will discuss possible problems, provide a summary of the results, and suggest ways to improve the model in section 5.

2 Literature review

The rise of peer-to-peer accommodation platforms like Airbnb has sparked significant interest in understanding the determinants of pricing within this dynamic market. Existing literature has delved into various factors influencing Airbnb prices.

Similar works have focused on studying how prices are affected by spatial factors among a broad number of European cities [1], using a number of regression techniques like SAR and SLX. Our study, in contrast, narrows its focus to a single European city and will employ simple regression models.

In a parallel study [2], the point was the pricing of Parisian hotel rooms. Intriguingly, their findings revealed a statistically significant correlation between proximity to the Eiffel Tower and premium room rates. Contrarily, our study focuses on the distance from the heart of Paris, which is approximately near Notre-Dame de Paris Cathédrale. While their investigation segmented the analysis into low and high-priced subsamples, our paper deviates from this approach, concentrating on a unified examination of Airbnb pricing determinants in Paris without such stratification.

In the other paper [3], the regressions in the linear, semilog, and log-linear forms were tested. Based on its fitting coefficient, it was determined that the third regression was the best one. In our instance, a hedonic price equation will be expressed in semi-log form. Remarkably, they

also conducted a geographically weighted regression analysis and concluded that the Xicheng Borough's lodging prices were more fair.

Lastly, Fallis et al [5] have focused on studying the controlled and the uncontrolled rental units in Vancouver. In their study they performed OLS regression using a semi-logarithmic and a linear model, and their main findings were that distance from city centre, building renovation and other factors influence the uncontrolled sector. This suggests that in property pricing there may be transitive relationships (where one variable affects the other), which could result in a confounding effect. Nonetheless, the key difference between [5] and our study is that we focus on short-term renting by tourists, while long term rentals were considered by Fallis et al.

3 Data description

3.1 Data Source

The dataset has been borrowed from Kaggle [6], which was gathered from TripAdvisor. It is important to mention that even while undergoing quality checks, the chosen dataset still may have some potential biases or limitations.

The chosen dataset contains information about Airbnb listings in Paris during weekends. The following columns were chosen for the study at hand:

1) realSum - a numeric column giving the overnight price in an airbnb listing.
2) Room type - a categorical variable that is split into three dummy variables (room_shared, room_private and home entire). Room_shared implies that the entire place (including the bedroom) is shared between people, whereas in room_private the customer shares only the public space while retaining a private bedroom, lastly, entire_place means that the whole listing (apartment/house) is given to the customer.
3) Host other listings - a categorical variable which conveys information about how many airbnb listings the host has. This variable is also a dummy consisting of single (host has one listing), multi (from two to 4 listings) and biz (more than 4 listings).
4) Host is super host - a binary variable which indicates the status of the host. A super host status is given to hosts that provide excellent airbnb listings.
5) Cleanliness rating - a subjective score for cleanliness of the apartment given by previous tenants.
6) Guest satisfaction overall - a subjective rating given by previous renters.
7) Bedrooms - number of bedrooms in the listing.
8) Person capacity - the number of people who could comfortably stay in the apartment, a variable determined by the host.

9) Distance - the kilometre distance from the city centre

10) Metro distance - distance in kilometres to the nearest underground station.

11) Attraction index - an index which measures the number of tourist attractions in the area.

12) Restaurant index - an index which quantifies available restaurants in the area.

Overall, the data contains listing specific information, which can be further split into an objective component (bedrooms, room_type and person_capacity) and a subjective component (cleanliness_rating and guest_satisfaction). Additionally, the data considers location specific factors (metro_dist, dist, attr_index and rest_index). Finally, platform specific variables and information is also taken into account (host_superhost and host_other_listings). Overall, this creates a well-shaped dataset for further econometric study.

3.2 Data analysis

First and foremost, from the statistics of the variables it can be noticed that in general guest satisfaction is quite high, most of the data lies in the interval between 89 and 98. Also, all the listings are close to the metro, with the maximum of 1.15 and 75% of the data below 0.29. This may affect the results.

Statistic	Min	Pctl(25)	Median	Pctl(75)	Max	St. Dev.
realSum	95.30	240.99	316.20	461.13	4,188.41	260.08
room_shared	0	0	0	0	1	0.11
room_private	0	0	0	0	1	0.41
person_capacity	2	2	2	4	6	1.21
host_is_superhost	0	0	0	0	1	0.35
multi	0	0	0	0	1	0.41
biz	0	0	0	0	1	0.42
cleanliness_rating	2	9	9	10	10	0.96
guest_satisfaction_overall	20	89	94	98	100	8.66
bedrooms	0	1	1	1	5	0.64
dist	0.07	1.82	2.96	4.04	7.70	1.46
metro_dist	0.003	0.14	0.21	0.29	1.15	0.12
attr_index_norm	5.65	12.86	16.53	22.20	100.00	7.78
rest_index_norm	11.93	27.61	35.87	47.71	100.00	13.05

Table 1: Table summarising descriptive statistics for columns in the Airbnb dataset.

What is more, after a thoughtful review of histograms for the dependent and the independent variables, it was decided to log-transform the realSum variable [Figure 1]. This led to a non-skewed distribution for the respective variable.

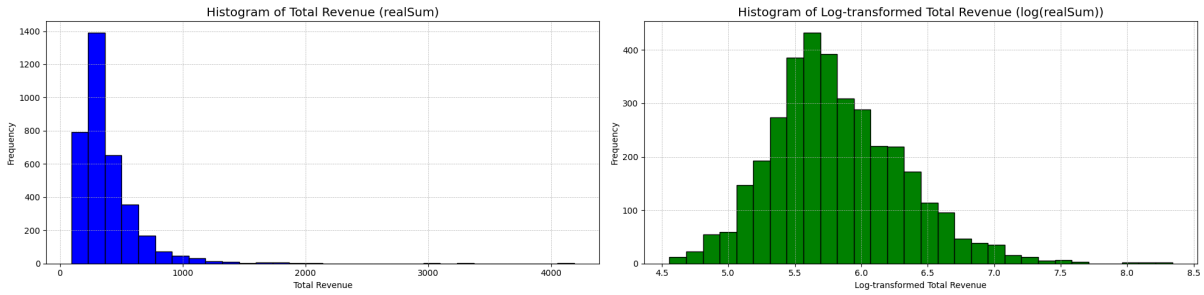

Figure 1: Histograms for realSum and log(realSum)

Additionally, for the guest_satisfaction and the cleanliness_rating variables boxplots were plotted to detect outliers. The figures indicate that the data is shifted right near its maximum value (10 for cleanliness_rating and 100 for guest_satisfaction). Extreme outlier bounds were calculated using the formula:

$$Q1 - 3*IQR \text{ and } Q3 + 3*IQR$$

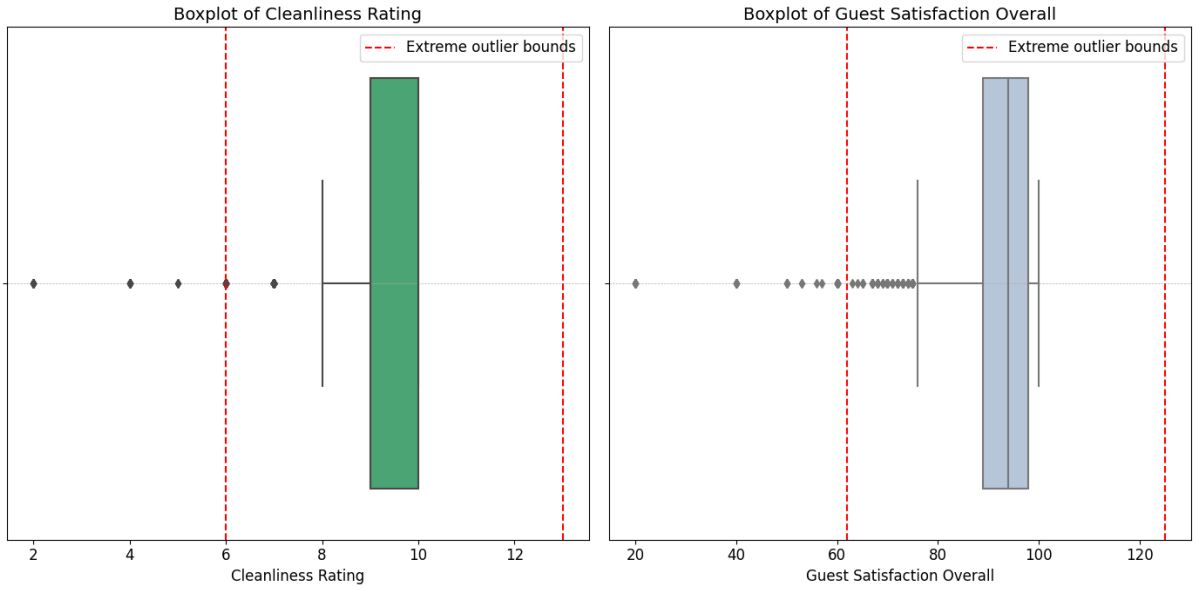

Figure 2: Boxplots for cleanliness_rataing and guest_satisfaction_overall.

The red dotted lines on [Figure 2] give a projection of the lower boundary for the extreme outliers (please note that the upper boundary goes over the maximum value for the two statistics). On the one hand, these outliers will affect the results and the coefficients of our regression model. At the same time it is probable that the nature of the outliers is not erroneous, instead they contain valuable information about the structure of the data and removing them could introduce bias. For instance, it is highly probable that the apartments with low cleanliness and guest_satisfaction will offer prices that are significantly lower, and we want our regression line to highlight and take into account this effect.

The correlation matrix [Figure 3] demonstrates inter variable correlation within our model. Notably, there is a strong positive correlation between attr_index_norm and rest_index_norm and between bedrooms and person_capacity. This is expected, because the two indices give

information about attractions and restaurants, it is quite likely that restaurants will be located near attractions, since there will be more potential customers there. Moreover, the two indices are negatively correlated with dist and metro_dist. This also follows economic intuition because as we move away from the city centre, the number of attractions and restaurants should decrease, because city centres are the most developed parts of the city, with high daily human traffic. Nonetheless, the correlations between variables could suggest multicollinearity, this in turn could affect our regression coefficients and introduce unreliable coefficient estimates.

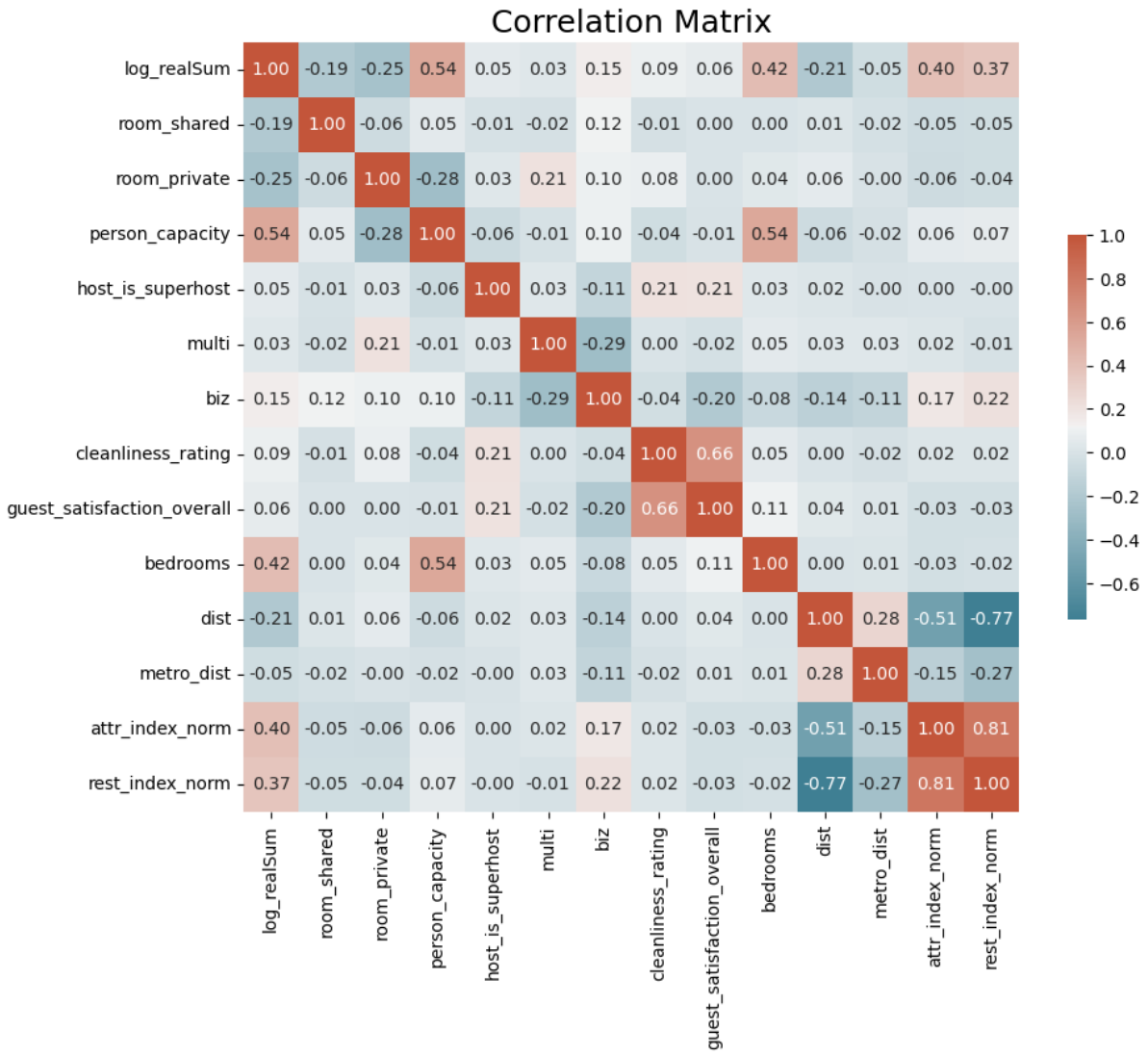

Figure 3: Correlation matrix for the dataset

4 Economic model

4.1 Functional form

In the study, it was decided to log-transform the dependent variable (realSum) for several reasons. Firstly, upon examining [Figure 1], we observed that the histogram for realSum exhibited a right-skewed distribution, and log-transforming this variable helped us attain a more symmetric and normally distributed form. Our decision to employ log-transformations aligns with similar methodologies adopted in related studies [1,3,4,5]. Lastly, we wanted to account for the magnitude of the price change and not the change itself.

It was decided to maintain the semi-log functional form, which meant that the independent variables would remain unchanged. Nonetheless, the addition of the categorical variables “room_shared” and “room_private”, multiplied by person_capacity is a reasonable action. The rationale behind this is to capture the relationship between the type of room and its occupants: in the case of shared rooms, a smaller number of people residing in the same room makes it better, and in the case of a flat owned by a single person, conversely, a larger capacity makes it more attractive.

4.2 Omitted and excessive variables

To address the issue of excessive variables, it was decided to omit ‘Unnamed: 0’ column because it was used for indexing, furthermore we have removed ‘lng’ and ‘lat’, the coordinate variables of the rented apartment, since these are irrelevant for the study at hand, and there is no economic background for which these variables should affect the rent price. Similarly, a decision was made to omit the room_type variable since it contained string values which were difficult to analyse. Finally, the rest_index and the attr_index variables were removed because they gave duplicate non-normalised information for the variables norm_rest_index and norm_attr_index. Omitting these variables does not introduce omitted variables bias, since it is unlikely that these variables affect our regression model.

On the other hand, it is likely we have underspecified our model, which would make our OLS estimates inconsistent. It is possible for variables like season or the respective design or district in Paris to affect our OLS regression, however, the project work at hand was limited by the dataset. For future studies it is recommended to collect a dataset with a greater number of explanatory variables. Also, an important note should be made regarding model specification, overspecification, while leading to inefficient estimators, is generally preferred to underspecification, as the latter results in inconsistent estimators.

4.3 Hypothesis on the influence

Independent variable	**Hypothesis on the influence**
Room shared	Price should decrease if this dummy is true due to the fact that the coefficient contains a dummy variable responsible for the whole apartment. Shared room is valued much less than an apartment, because people usually do not like to live with people they do not know.

Room private	Price should decrease if it is true, but less than for a shared room. The reasoning is similar.
Person_capacity/Person_capacity*room_shared/Person_capacity*room_private	From the empirical experience it is clear that if the room is shared, smaller person capacity is valued more. However, for entire homes and private rooms the opposite holds. Hence, is person_capacity increases, person_capacity*room_shared should decrease price, on the opposite, person_capacity*room_private should increase it.
Host_is_superhost	In case a host is marked as superhost, people understand immediately that living conditions are likely to be good. Thus, the price charged ought to be higher.
Multi	If it is true, the price should increase, because hosts are more experienced. They know how to attract guests, can offer better living conditions, and have higher income for renovation purposes.
Biz	Biz is similar to multi, but the effect ought to be even stronger, as biz hosts have even more flats or apartments to rent.
Cleanliness rating	Higher cleanliness rating should attract tourists and such flats should have higher pricing.
Guest satisfaction overall	Before choosing a place to stay at airbnb people take into account reviews. Thus, greater guest satisfaction leads to higher prices.
Bedrooms	More bedrooms, the higher the price should be, as more people can sleep separately at this place.
Dist	With a greater distance from the city centre, prices should drop. It seems to be the case, as usually all main tourist attractions, restaurants are located there, thus, more people want to live close to these places and the price is higher.
Metro_dist	Metro is a form of public transport, which reduces travel time. Living close to the subway would save time. Thus, listings closer to the subway should have greater prices.
Attr index norm	Attraction index in this dataset is calculated by the formula $\sum_{k=1}^{K} \frac{R_k}{d_{jk}}$, where R_k is the number of reviews and d_{jk} is the distance from attraction to the listing. The hypothesis is that more tourist attractions are located in more expensive areas, as many people want to stay closer to all the amusing and historical places. Thus, with higher index price, also, ought to be greater.
Rest index norm	Calculated in a similar way as attraction index, but for restaurants. More restaurants should be closer to the city center and higher indexes must increase the price of the listing.

5 Model Results

5.1 Coefficient Interpretation

The results of our OLS regression model show that at 0.001 significance level, room_type, person_capacity, bedrooms, host_is_superhost, host_listings, cleanliness_rating, attr_index_norm, rest_index_norm, dist and room_shared*person_capacity are significant.

The coefficients of the property type imply that if a listing is shared (dummy variable room_shared is true), then the listing is 32% cheaper, when compared to an apartment with an exclusive private access. This is expected, because in a shared property the total price is split among unacquainted individuals, moreover the added discomfort of sharing property should reduce its price. Similarly when a customer books a private room, there is a decrease in price by 29%. In this case, the interpretation is similar to when a property is shared, however, the price shift is less severe because there is extra privacy specifically confined to a single room.

On the other hand, the coefficients for listing capacity and size generally increase the lending price of the property. The regression estimates show that for a unit increase in the capacity of an Airbnb property, there is a 14% increase in price. This follows economic sense, because properties which can house more people, are bigger and thus should be more expensive. Additionally, for each extra bedroom in the property there is a 20% increase in the rent price. The same economic interpretation can be drawn as to the previous case. Nonetheless, it is interesting that extra bedrooms have a greater impact on price than person capacity.

Furthermore, our regression estimate implies that another key factor influencing the price of a property

Regression with Robust Errors

	Dependent variable:
	log(realSum)
room_shared	-0.324***
	(0.107)
room_private	-0.287***
	(0.063)
person_capacity	0.137***
	(0.006)
host_is_superhost	0.088***
	(0.017)
multi	0.121***
	(0.016)
biz	0.191***
	(0.016)
cleanliness_rating	0.042***
	(0.008)
guest_satisfaction_overall	0.001
	(0.001)
bedrooms	0.201***
	(0.012)
attr_index_norm	0.015***
	(0.002)
rest_index_norm	0.007***
	(0.002)
metro_dist	0.108**
	(0.052)
dist	0.038***
	(0.007)
room_shared:person_capacity	-0.184***
	(0.025)
room_private:person_capacity	0.013
	(0.027)
Constant	4.032***
	(0.086)
Observations	3,558
R^2	0.566
Adjusted R^2	0.565
Residual Std. Error	0.329 (df = 3542)
F Statistic	308.502*** (df = 15; 3542)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Regression with Robust Errors

is the number of listings that the host has. In detail, when a host has from 2 to 4 listings on Airbnb (indicated by the dummy multi), the price of the listing is 12% higher and when a host has 4+ listings (indicated by the dummy variable biz) the price is 19% higher when compared to the case of a host with a single listing. This shows that on Airbnb the experience of the host plays a key part in the price of the listing. A possible explanation for this observation could be superior services to the customer, from an experienced host.

Regarding the platform specific variables, a unit increase in the cleanliness_rating is estimated to result in a 4% increase in the price of the apartment. A possible economic explanation is that clean flats are demanded more than unclean property. Additionally, the presence of the super host status increases the price of an airbnb listing by 9%, this effect could again be linked to the experience of the host.

Additionally, our model estimated that all of the location-specific factors are significant. Generally, the model estimates that being further away from the city centre and the underground results in a 4% and a 10% increase in the price, which contradicts our hypothesis. Contrastingly, the two indices (restaurant index and the attraction index) exhibit a clear and logical impact on the price of the apartment. For each unit increase in the respective index, there is a 1.5% and 0.7% increase in the price of the listing, aligning with expected economic principles and the fact that tourists are ready to pay more for areas with a large number of tourist attractions and restaurants.

Regarding our interaction terms, the interaction term between a shared_room and person_capacity turned out significant and means that in a shared property for every additional co-renter, there is a decrease in the cost of the listing by 18%.Contrarily, the interaction term between room_private and person_capacity was insignificant. This outcome could be credited to the nature of private rooms, where the renter’s private space diminishes the influence of additional occupants.

Finally, it should be noted that guest_satisfaction_overall was estimated as an insignificant variable. This is an interesting observation and contradicts the findings of [4], where customer satisfaction and reviews were of utmost importance. In our case, a possible explanation could be drawn from the distribution of guest_satisfaction, where the variable is shifted towards its max_value - 100.

Overall, the results generally follow our hypothesis. A major deviance was singled out in our predictions for the dist and metro_dist factors. Additionally, guest_satisfaction_overall and an interaction term between room_private and person_capacity was estimated as a non-significant variable.

6 Conclusion

In the tourism industry, determining the optimal prices for rented property is an essential objective. By constructing a semi-logarithmic linear regression model and investigating the variety of factors that may influence prices of Airnbnb listings in Paris, this study has

generated a comprehensive pricing framework, which may be readily applied to price listings within the Airbnb application.

With regard to our findings, it is important to note that, although the results of the OLS strongly correspond to our hypothesis, certain results were counterintuitive. For instance, the insignificance of `guest_satisfaction_overall` coefficient and the positive relationship between price and distance from the city centre, as well as distance from the underground are results which demand deeper exploration (perhaps even in a separate study).

Nonetheless, we postulate that a possible cause for the insignificance of the guest_satisfaction_overall coefficient could be the distributional shift in the data, wth 75% of the data lying above 89, as indicated by [Figure 2]. The reduced variability could in turn make it difficult for the OLS to establish a statistically significant relationship between our dependent and the independent variables. Furthermore, a possible reason behind the positive relationship between distance and price could be the locational peculiarity of Paris, where the main tourist attractions (the Eiffel Tower, the Arc de Triomphe and Louvre) are located to the East of the city. Since tourists would prefer being near the attractions we recommend recalculating the dist variable, taking one of these attractions as the centre.

Lastly, the possibility of violations of the Gauss-Markov theorem should not be neglected. Adding new functional forms, new interaction terms and variables like latitude, longitude or season could potentially diminish the effect of model underspecification. Moreover, to further improve our model, it is advisable to reduce the multicollinearity between variables as demonstrated by [Figure 3](we could try to create an interaction between correlated features).

7 References

[1] - Gyódi, Kristóf, and Łukasz Nawaro. “Determinants of Airbnb Prices in European Cities: A Spatial Econometrics Approach.” Tourism Management 86 (2021): 104319. https://doi.org/10.1016/j.tourman.2021.104319.

[2] - Meahan, Sam A. “Pricing Paris Hotel Rooms: A Hedonic Pricing Approach.” *Department of Economics of the University of Ottawa*, January 2016.

[3] - Zhang, Honglei, Jie Zhang, Shaojing Lu, Shaowen Cheng, and Jinhe Zhang. “Modeling Hotel Room Price with Geographically Weighted Regression.” *International Journal of Hospitality Management* 30, no. 4 (2011): 1036–43. https://doi.org/10.1016/j.ijhm.2011.03.010.

[4] - Soler, Ismael P., German Gemar, Marisol B. Correia, and Francisco Serra. “Algarve Hotel Price Determinants: A Hedonic Pricing Model.” *Tourism Management* 70 (2019): 311–21. https://doi.org/10.1016/j.tourman.2018.08.028.

[5] - Marks, Denton. “The Effect of Rent Control on the Price of Rental Housing: An Hedonic Approach.” *Land Economics*60, no. 1 (1984): 81. https://doi.org/10.2307/3146095.

[6] - “Airbnb Prices in European Cities.” Kaggle. Accessed December 2, 2023. https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities?select=paris_weekends.csv.