Influence of physical features and location of an Edinburgh property on the price of its Airbnb listing

Pavel Khudov

April 9, 2023

1 Overview

The following data analysis investigates the impact of property features on the price of Airbnb listings in Edinburgh, addressing two main questions: (1) how well can physical features of a property predict its price, and (2) whether certain areas or neighbourhoods are more expensive than others. Results indicate that property size and listing type significantly influence price. Hotel rooms and unconventional properties, such as boats and windmills, are generally more expensive, while shared rooms are the cheapest. Larger properties, with more bedrooms and bathrooms, also demand higher prices. While exploratory analysis suggests that properties closer to the city centre are more expensive, regression analysis does not support this finding, likely due to methodological limitations. The low R-squared values indicate that the analysed features alone are insufficient for accurately predicting listing prices. Improvements to the analysis could include incorporating additional data, such as amenities and more information on existing features. Future research could also explore the impact of host information and reviews on listing prices. Comparing these findings to related work, this analysis supports Wang and Nicolau's [1] study on the impact of room type and property size on price, but differs on the influence of proximity to the city centre.

2 Introduction

Context and motivation Edinburgh, a city renowned for its thriving tourism industry, creates a high demand for short-term rental accommodations such as those offered on Airbnb. In this context, investigating the impact of property features on the price of Airbnb listings becomes a compelling and relevant research topic. The rise of internet technologies has made leasing properties through platforms like Airbnb easier than ever. However, increased market access has also intensified competition, making it challenging for property owners to determine optimal listing prices. With this data analysis, the aim is to investigate the factors influencing Airbnb pricing, so that property owners could make better-informed decisions, while potential renters could use this information to find the best deals on listings meeting their specific needs.

Previous work Multiple relevant analysis have been performed in the past, most notably the following study has been carried out: Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb. com. *International Journal of Hospitality Management*, 62, 120-131.[1]. It has concluded that proximity to the city centre and the occupancy are some of the factors that make a listing more expensive. Nevertheless, the study focuses on 33 cities, and not Edinburgh specifically.

Objectives This data analysis aims to answer two main questions related to the listing's price:

- How well can the **physical** features of the property be used to predict the price of the property?
- Are particular areas or neighbourhoods more expensive than others?

3 Data

Data provenance The source of the dataset is Inside Airbnb[2], and its snapshot was imported using University of Edinburgh's link: `https://www.inf.ed.ac.uk/teaching/courses/fds/data/project-2022-2023/airbnb`. Based on Inside Airbnb's website, it is allowed to obtain and analyse the data as long as it is with good intentions and it is not re-published: `http://insideairbnb.com/data-policies`.

Data description The dataset consists of 75 columns, and 7389 entries, each representing an Airbnb listing in Edinburgh at the time of 16th of December of 2022.

Data processing The raw dataset required significant cleaning before analysis could begin. It involved dropping all columns related to data scraping and those containing 20% or more missing values. The decision to delete columns with 20% missing values was based on optimization after trying different values. Information related to the host was dropped since the analysis focused solely on physical features. Although the reviews are not physical features, two of them are associated with them, these were left in the dataset: review_score_location and review_score_cleanliness. Description and amenities were also dropped due to high variability and the difficulty in analysing and grouping them sensibly. Data types were converted accordingly, repeated information was deleted, and the bathroom column was separated into two categories: the number of bathrooms and whether they were shared or not. Missing values were replaced with the most frequent value, except for reviews, where the mean was used due to greater variability. Z-score > 3 was used as a filter to remove outliers. Some prices were too high because some hosts set it in order to avoid the property being booked, as Airbnb does not allow setting the listing "on pause". Finally, the cleaned dataset was separated into two data frames: one containing physical features and cleanliness review score and another related to location, including location review score. Dummy variables were created in the former data frame to perform PCA later, while the latter data frame included a column indicating the Euclidean distance from the centre of Edinburgh (55.9533° N, 3.1883° W). The details of these operations can be found in the Jupyter notebook.

4 Exploration and analysis

4.1 How well can the physical features of the property be used to predict the price of the property?

Observing that the physical features data had many variables, the first obvious step would be to perform PCA to deal with collinearity and to reduce the number of dimensions. After standardising the data and applying the algorithm, 5 PCs resulted as optimal based on the knee. However, each of them did not explain significant variance (PC1: 11.3 %; PC2-5: <5%) and the similar loadings of initial features in different PCs indicated that we cannot associate any PC with a particular feature. Thus, the grouping of the features had to be done manually using reason and by considering high correlation ($0.4 < r < 0.8$) between beds, bedrooms, bathrooms and accommodates (visualisation in the Jupyter notebook). All of them could be put under the category of "size", and the features "room type", "property type" and whether bathroom is shared or not under "listing type". With that information, a number of visualisations were made for each of the two categories: size and listing type. "Review score cleanliness" will not be considered as it has very low correlation with a price ($r = -0.03$).

Firstly, the distribution of number of listings for each property type, room type and number of people it accommodates was visualised, from which it could be concluded that the majority of properties listings are for 2 and 4 people, they are entire homes/apt and either condos or apartments. (These plots not included into analysis due to space constraint, but they can be found in the Jupyter Notebook).

Then, with the Figure 1, the importance of listing type for a price was explored, letting its size apart. It can be seen that hotel rooms are the most expensive, since they provide other services. Entire home/apt and private rooms are about the same, since private rooms tend to be part of a home/apt. On the expensive side we can also see properties like boats and windmills, whether on the cheap side there are those in the countryside like villa, campsite and farm stay.

Figure 1: Price/person vs listing type

Lastly, the Figure 2 shows a positive trend in price with respect to number of guests it accommodates, number of bedrooms, beds and bathrooms. It makes sense as there is a high correlation between them, as can be seen on a heatmap in Jupyter notebook. We also see that the slopes of bedrooms' and bathrooms'graphs are steeper. Because those are associated more with the physical size of a property, we can conclude the price depends on it more than on the amount of people it can accommodate.

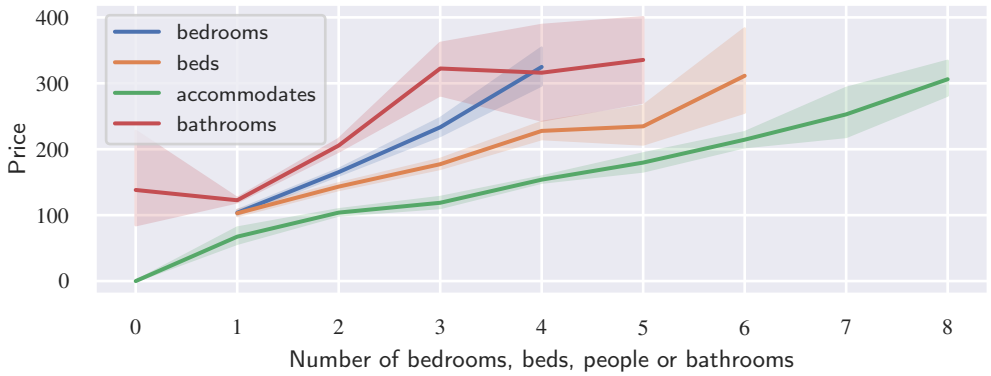

Figure 2: Price vs variables determining size

The plots described above resulted to be useful in concluding that the size of the property its type has influence on the price, however a multiple linear regression had to be performed in order to evaluate explicitly how well can the price be predicted. Firstly, a single variable linear regression was run on each of the variables in order to compare R squared to that of multiple regression in order to filter out variables

that do not give much information. The results are in the following Table 1, where it is separated into "listing type" and "size" categories:

Table 1: R-squared values of single and multiple regression

Price vs	R-squared
property_type	0.02
room_type	0.06
shared_bathroom	0.05
Multiple regression	0.08

Price vs	R-squared
accommodates	0.13
beds	0.01
bedrooms	0.14
bathrooms	0.08
Multiple regression	0.16

We see that for type of listing, "room type" has the largest R-squared, which means that the largest amount of variance can be explained with it. The second largest is that of "shared bathroom", however adding it with the previous one would result in a value bigger than that of the multiple regression (0.05+0.06 > 0.08). That means they are correlated, and it makes sense, since those listings where only a room is rented tend to have a shared bathroom. Hence, "shared bathroom" is dropped, and only "property type" and "room type" are considered. For the variables related to size, there is a similar situation: "bedrooms" explain the most variance, so "accommodates" is dropped because it is highly correlated (0.13+0.14 > 0.16). Because there tends to be a same number of beds as people, "beds" is dropped as well. "Bathrooms" remain in the analysis.

Afterwards a new multiple regression was performed, predicting price with the selected by the process described above variables "property type" and "room type" (type of listing), and "bedrooms" and "bathrooms" (listing's size). The resulting value of R-squared is $R^2 = 0.18$ and $R^2_{adj} = 0.18$.

4.2 Are particular areas or neighbourhoods more expensive than others?

In order to see whether location has influence on the price, we check whether it is correlated with the review score of location, as it is reasonable to assume that the higher the property was rated on this rating, the better the location is. The correlation coefficient turned out to be low ($r = 0.08$), however on the plot on the Figure 3 it can be seen that there is a slight positive overall trend. Thus, it is worth analysing which concrete areas tend to be more expensive.

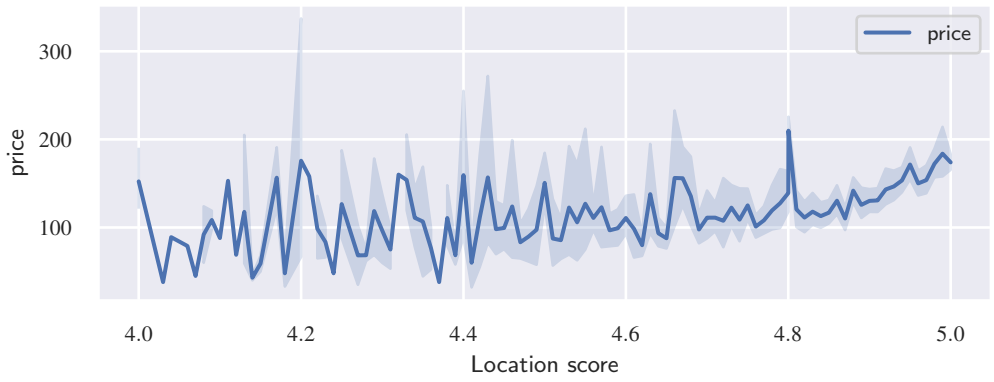

Figure 3: Price vs Location Score

Using latitude and longitude data, the properties are shown on the map on the Figure 4, with warmer colours and bigger circle size representing more expensive property. It can be concluded that the listings

in the centre of Edinburgh have higher price per night than those on the outskirts, with few exceptions. For example, in the south-east there is a cluster of very expensive listings. If we look at the bar plot in Figure 5 representing the listings' median price against the neighbourhood, we see that it matches the map: Old Town and New Town are on the expensive edge, those being the centre of the city. Fairmilehead and Carrick Knowe are the most expensive areas, the latter being the cluster referred to earlier. Median was chosen for the plot, not mean, to reduce the impact of few properties with high price that are an exception.

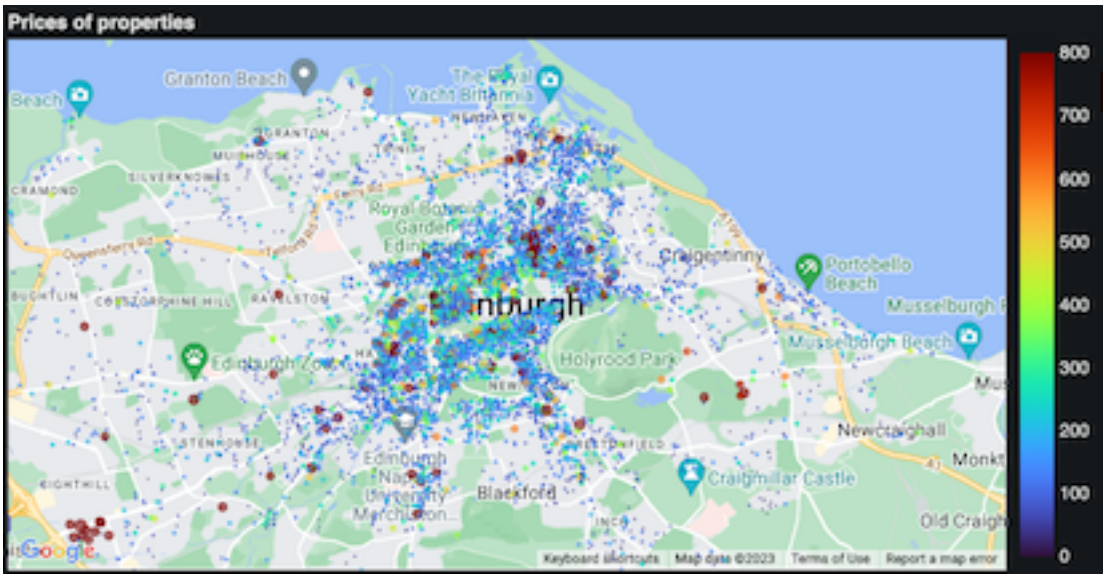

Figure 4: Prices of properties based on location

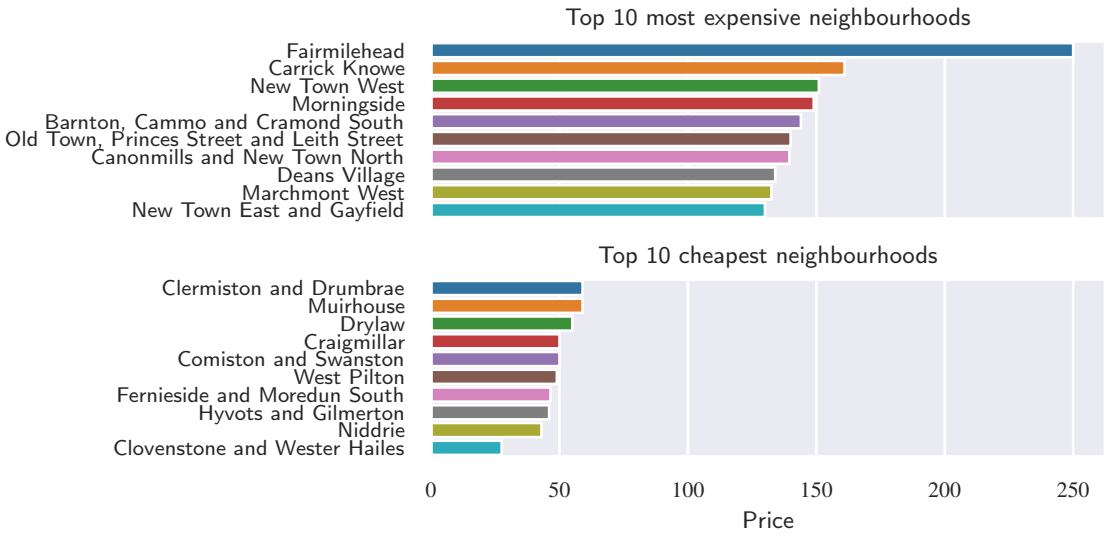

Figure 5: Prices of properties based on neighbourhood

Furthermore, a single variable regression predicting price was performed, resulting in an R-squared of $R^2 = 0.06$ for "neighbourhood", $R^2 = 0.006$ for "location review score" and surprisingly very low $R^2 = 0.004$ for "distance to city centre". Because the two latter are very small, it is unnecessary to perform multiple regression.

Adding the "neighbourhood" into multiple regression performed in the analysis of physical features as another variable allowed to not significantly increase the R-squared from $R^2 = 0.18$ to $R^2 = 0.19$. Nevertheless, the adjusted R-squared remained the same $R^2_{adj} = 0.18$, suggesting that the location data in the form that is presented in the dataset, does not help to predict the price of the property.

5 Discussion and conclusions

Summary of findings

- Hotel rooms and similar (hostel, aparthotel etc.) are the most expensive, possibly because they provide additional services. They are followed by unconventional properties, such as boat and windmill. This could be explained by the law of supply-demand, as there are only a few of them on the market. Shared rooms are the cheapest, possibly because the cost is divided between the accommodation sharers.

- The size of the property plays a big role in determining its letting price, a parameter which can be the best determined by the amount of bedrooms and bathrooms rather than the occupancy. The larger the property, the more expensive. It can be explained by the fact that it usually provides more comfort and initially required a higher investment by the owner.

- On the first sight, it seems like the location of the property should play an important role based on the exploratory analysis. Those properties closer to the centre (e.g. in New Town and Old Town) tend to be more expensive as they are considered more prestigious due to being close to the main attractions. There are a few that are exceptions (e.g. Fairmilehead and Carrick Knowe). Nevertheless, the regression analysis did not support the exploratory analysis, a fact that suggests inadequate methodology rather than reflection of reality.

Evaluation of own work: strengths and limitations The low values of R^2 (<0.2) indicated above suggest that the analysed features alone are not sufficient to accurately predict the price of the Airbnb listing in Edinburgh. Nevertheless, the analysis concluded that they still play a role, a fact that is supported mainly by the exploratory analysis. This lack of accuracy is due to not inclusion of sufficient data in the study: both of additional features and of more information about the existing ones. For example, the amenities' column was not analysed, and a distance between neighbourhoods was not represented. Moreover, the lack of distance between "property type" categorical variables did not allow performing PCA accurately. Nevertheless, the strength of the study is that still some additional information was calculated manually from the existing data, for instance, price per person, that allowed to compare listing types fairly, and distance to the city centre, that allowed to make the latitude and longitude data suitable for quantitative analysis.

Comparison with any other related work This data analysis supports the findings of the paper by Wang, D., & Nicolau, J. L[1] that the room type and size of the property impact the price. However, their study found that the price decreases the further the listing is from the centre, a conclusion that was not reached in this analysis.

Improvements and extensions The dropped "amenities" column contains variables that are good predictor of price, such as "free parking" and "Wi-Fi", according to Wang, D., & Nicolau, J. L's paper [1]. Including them into analysis would improve the accuracy of the model. Moreover, a thoughtful association between "neighbourhood" and "distance to city centre" and distance to other neighbourhoods would allow reducing collinearity and making location a better predictor of price. Lastly, the analysis could be extended by going beyond physical features, including information about the host and investigating the impact of reviews on the price of the listing.

References

[1] Wang D. Nicolau J. L. "Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com". In: *International Journal of Hospitality Management* (2017). Retrieved on 9 April 2023. URL: https://www.researchgate.net/publication/312544103_Price_determinants_of_sharing_economy_based_accommodation_rental_A_study_of_listings_from_33_cities_on_Airbnbcom.

[2] *Inside Airbnb: Edinburgh*. Last accessed 9 April 2023. 2023. URL: http://insideairbnb.com/edinburgh.