Investigating Airbnb listings in Edinburgh

Arnesh Saha and Abhay Maurya

April 11, 2023

1 Overview

Airbnb operates an online marketplace for short-term homestays and experiences. The report analyzes several Airbnb listings in the city of Edinburgh to identify the key factors that can potentially impact daily property prices. We perform thorough data wrangling and manipulation followed by several statistical analyses to produce detailed and informative conclusions. We explored different aspects and features of rental properties to study and analyse what factors impact the price of properties. In particular, we found that certain amenities provided in properties heavily impact the daily price. This allowed us to predict the price of properties in different neighbourhoods with accuracy, $R^2 \approx 0.70$. We also carefully evaluated and visualised the average daily price of different neighbourhoods in the City of Edinburgh.

2 Introduction

Context and motivation This study aims to identify the key factors that contribute to a desirable short-term rental property in Edinburgh and predict its daily price based on the amenities and features listed. Our analysis of data obtained by scraping Airbnb properties in the city provides valuable insights for future rental property owners who can use this information to optimize their services and amenities. Additionally, we analyzed how daily property prices fluctuate during different seasons and in various neighbourhoods, discovering that properties located in central areas and popular tourist attractions like Old Town and Princes Street tend to be more expensive with higher price fluctuations. Our study provides useful information for property owners looking to maximize their profitability by offering desirable amenities and targeting specific neighbourhoods in Edinburgh.

Previous work

- Airbnb Price Prediction Using Machine Learning and Sentiment Analysis [4]. A paper which examines the use of machine learning, deep learning, and natural language processing techniques to create a trustworthy model for predicting property prices.
- Machine Learning Prediction of New York Airbnb Prices [6]. This paper analyzes a sample of 48 896 listings in New York City from Airbnb.com and builds a price prediction model with natural language processing and machine learning techniques.
- Airbnb rental price modelling based on Latent Dirichlet Allocation and MESF-XGBoost composite model [2].
- Learning-based Airbnb Price Prediction Model [5]. This paper focuses on the Airbnb market in Beijing since China will be one of the main markets of Airbnb. The authors have developed a pricing prediction model based on machine learning approaches, i.e., XGBoost and neural network, for the Beijing Airbnb market.

- Leveraging multi-modality data to Airbnb price prediction [3]. In this paper, a variety of data are combined as inputs into machine learning algorithms and natural language processing framework to construct a reliable price prediction model for Airbnb rentals.

Objectives We are setting out to answer what makes a good property for Airbnb or short-term rental in the City of Edinburgh, in particular, we are trying to ask the following:

- How well can features of a property listing be used to predict its popularity or short-term rental price?
- Are particular areas or neighbourhoods more sought after or expensive than others?
- At which points during the year are notable variations in mean Airbnb prices observed across the neighbourhoods of Edinburgh?

3 Data

Data provenance InsideAirbnb is an independent project that provides free and public access to data and tools related to Airbnb rentals in different cities globally. The datasets used in this project were specifically obtained from the InsideAirbnb website for the city of Edinburgh. These datasets were downloaded as CSV files and pre-processed before analysis. We must give appropriate credit to the source and link to the license since the datasets are available under the Creative Commons Attribution 4.0 International License. This license allows for the data's unrestricted use as long as we credit the source appropriately.

Data description The data for this project was obtained from four files - listings.csv, calendar.csv, reviews.csv, and neighbourhood.geojson, all of which were sourced from InsideAirbnb. For this study, only the listings and calendar files were used. The listings file contains 75 columns with 7,389 entries and captures information about the host, property details (e.g., number of bedrooms/bathrooms, amenities), review ratings, coordinates, and availabilities. On the other hand, the calendar file contains 7 columns with 2,696,636 entries, including price, availability, and minimum/maximum stay duration for all listings from December 16, 2022, to December 15, 2023. Lastly, the neighbourhood file provides geographic information about Edinburgh's different neighbourhoods, enabling the creation of maps and visualizations to contextualize the listing data.

Data processing A substantial degree of data cleaning was required for this study. To refine the dataset for model training in predicting listing prices, we eliminated superfluous columns, transformed boolean values to 0s and 1s, and discarded rows containing missing values from the listings file. Moreover, we implemented one-hot encoding for the amenities column by generating distinct binary columns for each amenity. Furthermore, we employed z-score normalization on the price column to manage outliers. This method entails subtracting the data's mean and dividing the result by the standard deviation, thus yielding a transformed dataset with a mean of 0 and a standard deviation of 1. This approach assists in identifying and eliminating extreme values that deviate substantially from the mean, which could otherwise impede model performance.

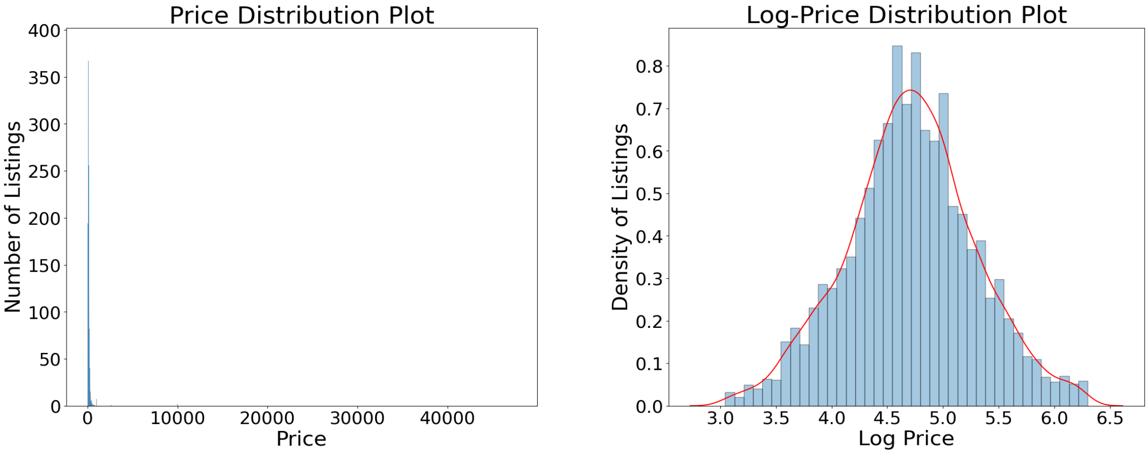
In order to examine the variation in listing prices across neighbourhoods, we combined the listings and neighbourhood files, 'listings.csv' and 'neighbourhoods.csv' respectively. This yielded a new data frame comprised of 7 columns and 111 rows, with each row representing a distinct neighbourhood group. The columns in this data frame encompass the average price of listings within each neighbourhood, alongside availability metrics such as availability for the upcoming 30 days and availability over the next 365 days. Additionally, the data frame features a count of properties within each neighbourhood, offering contextual information for the other metrics.

We explored the monthly price fluctuations for neighbourhoods throughout the City of Edinburgh.

Utilizing the available datasets, we generated a new data frame by merging daily price information for each property listing with the geometry of each neighbourhood group. We also assessed the presence of NaN values within our data frame. As an exploratory step, we removed all NaN values (empty entries in the date column for prices) and retained only 2 rows. This experience highlighted the challenges in obtaining every data point during the data scraping process. To preserve the overall data distribution and central tendency, we filled the empty values with the median value for each column. Subsequently, we grouped the properties based solely on their respective neighbourhoods and computed the daily price fluctuations and monthly price fluctuations (the mean of daily prices for each month) which allowed us to gain valuable insights into the dynamics of Airbnb prices across various neighbourhoods.

4 Exploration and analysis

Airbnb prices in Edinburgh exhibited a wide range, spanning from \$11 to \$47,566. Figure 1a illustrated a right-skewed distribution for the price, indicating positive skewness. A log transformation was applied to reduce the skewness of this feature, facilitating more straightforward interpretation and improved statistical analysis. Upon applying the log transformation, we obtained Graph 1b, where the good fit suggested that normality was a reasonable approximation. Consequently, we utilized log price as the target variable instead of the raw price for training and testing our price prediction models.



(a) Price Distribution plot for Airbnb Listings

(b) Normalized log-price distribution plot

Figure 1: Comparison of Raw-Price and Log-Price Distributions - Demonstrating the effect of log transformation on reducing skewness for improved interpretation and statistical analysis.

After cleaning the data we have 101 columns and 5152 rows of entries in total. There are 63 columns generated from amenities columns after the one hot encoding. Our Null Hypothesis, H_0 will be that we cannot predict the price of a listing from a list of features and our Alternative Hypothesis, H_1 will be that we can predict the price of a listing from its features.

We first created the heatmap to look at the possible correlations with our data. However, with 101 columns it is not feasible. So, plotted the top 10 highest correlations between the data. We immediately noticed that the 'accommodates', 'bedrooms' and 'entire home' had a direct impact on the price. However, the other factors seemed too weak for adding to our regression.

So, we then used Random Forest Regressor for feature selection and prediction. Random Forest can handle both continuous and categorical variables and is able to identify complex, nonlinear relationships between features and the target variable. We chose this model because there could be a lot of variables which has nonlinear relationships from price. In addition, we have continuous values for columns related to ratings and scores, whereas we have categorical values for columns such as room type and host response rate num. We used a test size of 0.25; this returned $R^2 \approx 0.708$ as seen in Figure 2, $mse \approx 0.1$, which is a good indication that our predictors are strong.

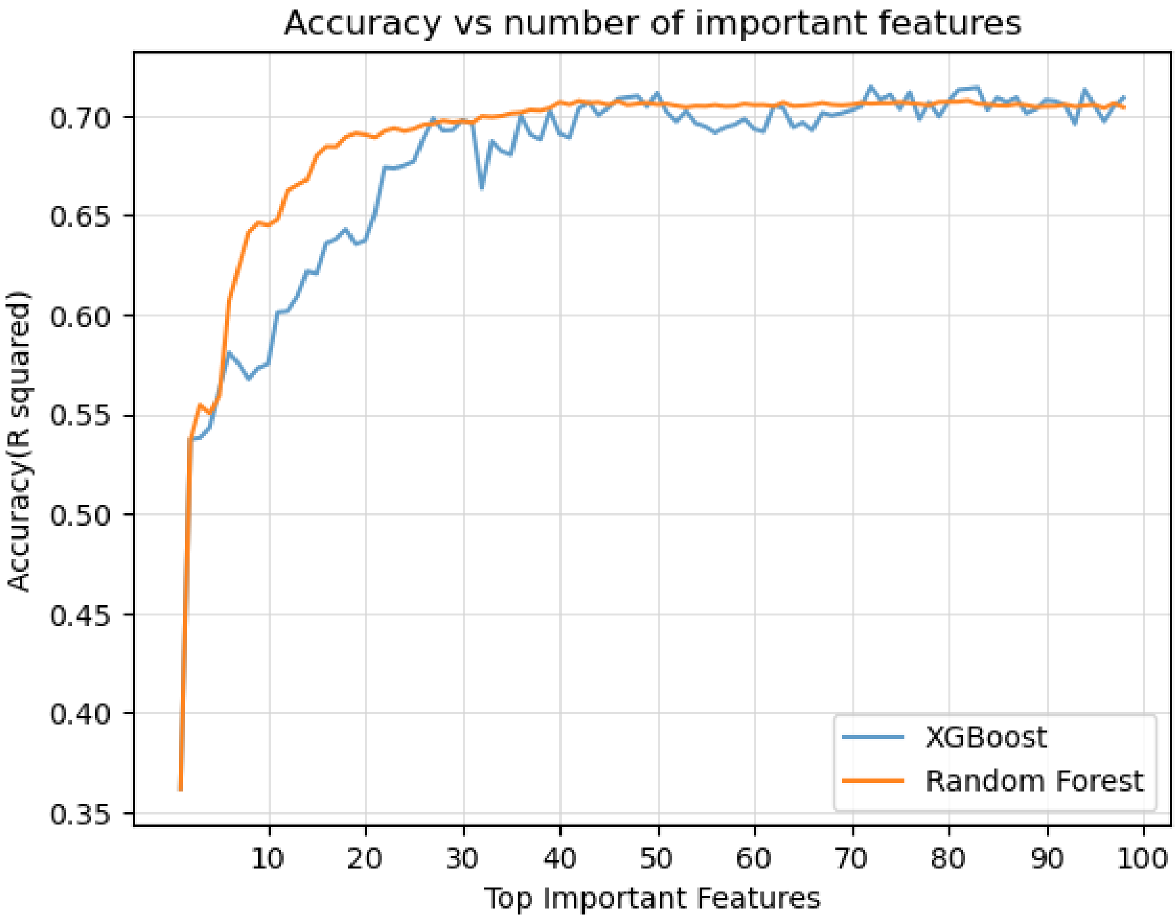


Figure 2: Accuracy vs. Number of Top Important Features - A comparative analysis of Random Forest and XGBoost models, illustrating the optimal feature subset for price prediction based on peak accuracy scores.

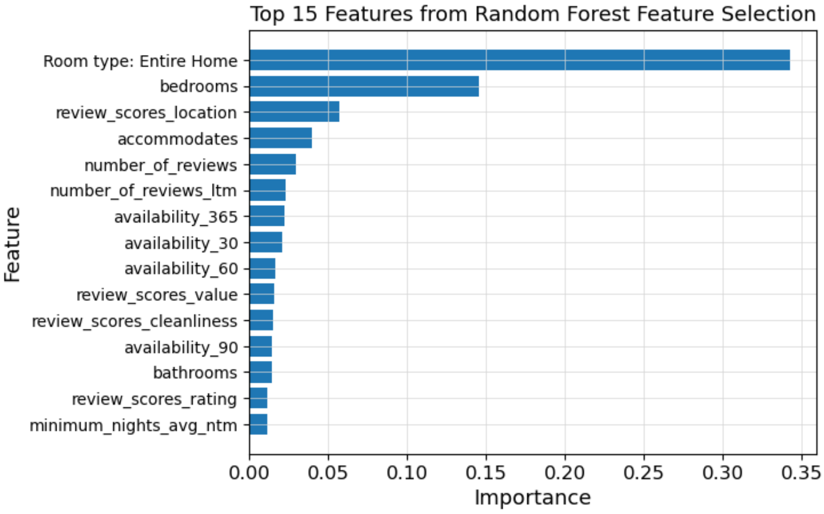


Figure 3: Top 15 features from the Random Forest Regressor which can predict the price of Airbnb listings in Edinburgh with accuracy $R^2 \approx 0.68$

We ran an additional model, XGBoost Regressor, to verify the accuracy of our prediction. It produced an $R^2 \approx 0.708$ and an $MSE \approx 0.1$, which is comparable to the results of the random forest model. Therefore, we reject the Null Hypothesis as it is clear that features play a significant role in predicting the listing price. Specifically, we can use the top 15 features from either model to accurately predict the price of a listing.

We conducted an analysis to determine if certain neighbourhoods in Edinburgh are more popular or expensive than others. Our analysis involved creating visualizations of

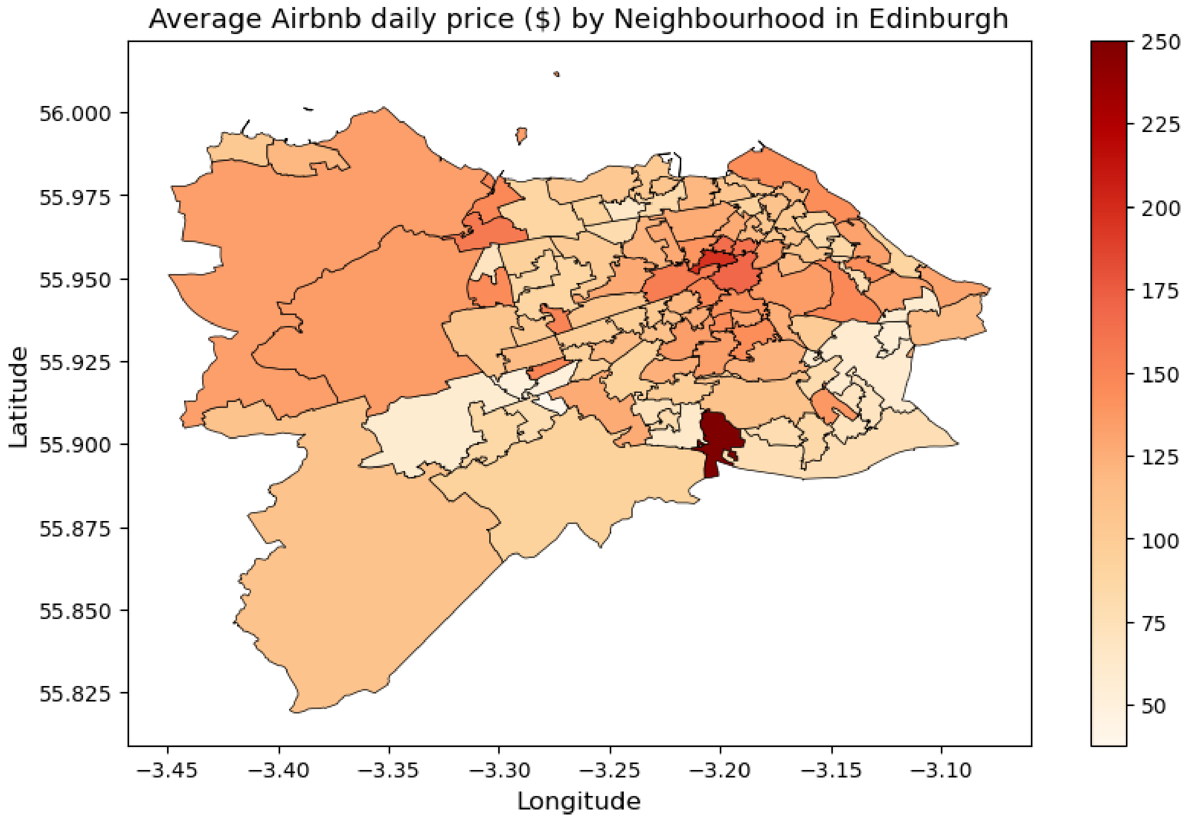


Figure 4: Average Airbnb daily price by Neighbourhood in Edinburgh

neighbourhood groups in the city based on various factors such as the average daily Airbnb price in USD, price per person based on the number of bedrooms, availability of properties during different time periods, and the number of properties in each neighbourhood. This allowed us to gain a deeper understanding of the issue at hand.

The analysis of average daily prices of Airbnb properties in different neighbourhoods is depicted in Figure 4. This visualization enables us to observe that certain neighbourhoods have higher prices compared to others. Specifically, the neighbourhoods located near Old Town and Princes Street tend to be more expensive than those situated farther away from the city centre.

In Figure 4, we can observe an outlier in the average daily price of Airbnb properties in the neighbourhood of Fairmilehead, which stands out from the other neighbourhoods. Surprisingly, the spread of average prices across neighbourhoods seems to contradict the usual trend of prices decreasing as we move away from the city centre. This finding prompted us to investigate further into the reasons behind these higher prices. To gain more insights, we visualized the number of Airbnb property listings in each

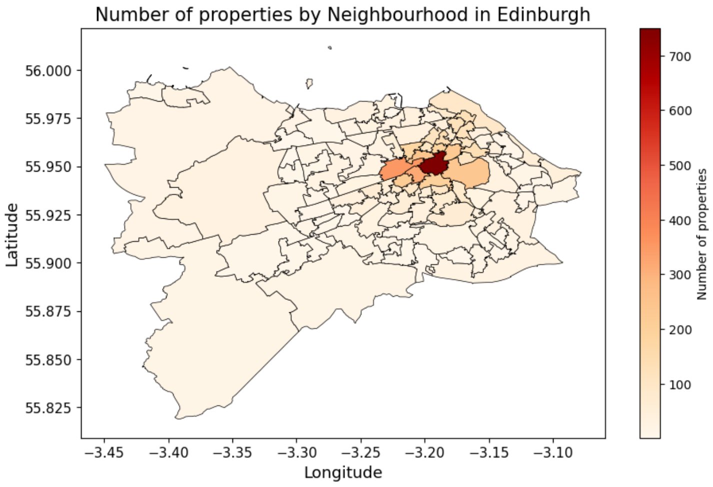


Figure 5: Number of properties in different neighbourhoods

neighbourhood in Edinburgh and observed a high number of listings in central neighbourhoods, including Old Town and Princes Street, which are popular tourist destinations featuring attractions like Edinburgh Castle, National Museum of Scotland, etc. These neighbourhoods had the highest number of listings, ranging from 400 to over 700 properties. However, in the case of Fairmilehead, we found that there was only one property contributing towards the high average price, which was located in a beautiful natural setting.

Additionally, we noticed that the neighbourhood of Ratho, Ingliston, and Gogar, which includes the Edinburgh Airport, has a high average daily price ranging between 125 to 175 USD. The adjacent neighbourhoods of Dalmeny, Kirkliston, and Newbridge also have a high average daily price despite having fewer than 100 listed Airbnb properties. This is likely due to their proximity to the Edinburgh coast, which makes them an attractive location for visitors.

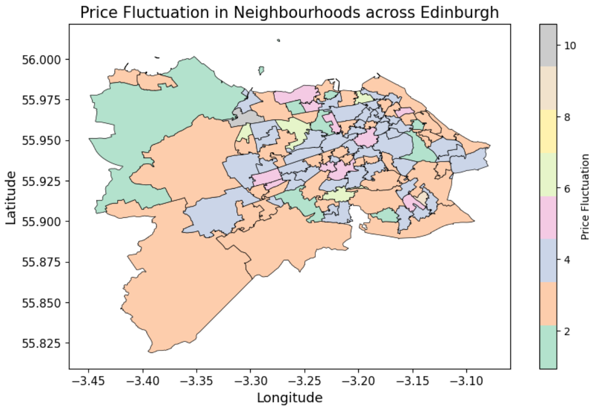


Figure 6: Daily fluctuation in price

Next, we conducted a more detailed analysis of the price of Airbnb properties. This involved examining the price fluctuations across various neighbourhoods in Edinburgh. We measured these fluctuations by calculating the standard deviation of the data, which involved merging the list of neighbourhoods with their respective mean daily prices from December 16th, 2022 to December 15th, 2023. To eliminate any outliers that could skew our analysis, we excluded any properties that had a daily price above 200 USD at any point during the year.

We also conducted experiments by setting the price cap at 300 USD, 400 USD, and 600 USD to study the impact of high-priced properties on our analysis. However, we did not find any significant fluctuations that could affect our goal of identifying the season with the maximum price deviation in Airbnb properties. Our analysis focused on identifying fluctuations or deviations in average daily price across different neighbourhoods in Edinburgh over a year. We observed that the most significant price fluctuations occurred in the neighbourhoods adjacent to central Edinburgh, such as Old Town and Princes Street, with fluctuations ranging between 4 to 6 units of the mean daily price. On the other hand, there was very little fluctuation in the outskirts of Edinburgh, with fluctuations ranging from 0 to 4 units of the mean daily price. Our careful observations suggest that the reasons for such price fluctuations could be the smaller number of properties in those neighbourhoods compared to other regions and the already high average daily price in USD.

We further analysed the fluctuation in average prices of properties in various neighbourhoods over the span of a year by considering the standard deviation in prices for neighbourhoods over 12 months. To gain a better understanding of the months in which there is a fluctuation in price, we grouped certain neighbourhoods on the basis of the value of fluctuation or deviation in their monthly price and created a visualization to understand the fluctuation over a certain span of months. From figure 7, we observed that despite the value of fluctuation, all neighbourhoods experience a positive hike in prices during the months between March and October(Prime tourist season) after which the property prices start gradually going down back to their initial values as winter approaches. We sought to find the reason for the high fluctuation in mean price across all neighbourhoods between the months of March and October. Our intuition of the cause of fluctuation during the summer months being the tourist season coupled with the Edinburgh Fringe Festival was confirmed by Edinburgh by Numbers 2021 released by the council of Edinburgh which provided quantitative data highlighting a spike in tourism by approximately 70% including both domestic and international tourists with the Edinburgh Fringe Festival having a footfall of over 3 million individuals. Our intuition of heavy footfall and a spike in Airbnb rental prices was confirmed through

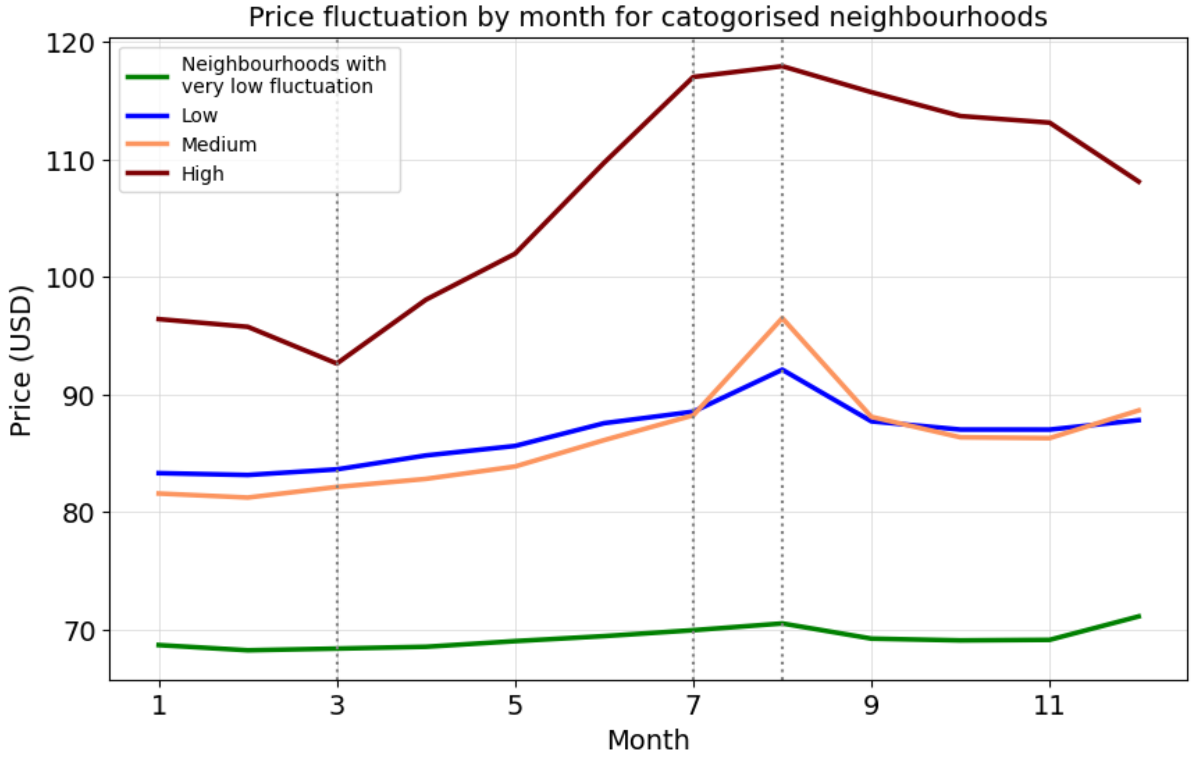


Figure 7: Monthly fluctuation in the average price of properties based on categorical neighbourhoods

several media publications by agencies like The Scotsman: Scotland's National Newspaper as well as EdinburghLive which recorded 50% hike in Airbnb rental prices during the Fringe Festival 2022.

5 Discussion and conclusions

Summary of findings There are almost the top 15 features which can predict the price of Airbnb. In addition to focusing on these top features, hosts can also use this information to set a competitive price for their listing. This information can be useful for people looking to book accommodations on Airbnb. By understanding which features contribute the most to the price of a listing, people can make more informed decisions about where to stay and what to prioritize in their search. Additionally, this analysis may be useful for Airbnb as a company, as it provides insights into what factors drive pricing and can help inform their business decisions.

The central region may have higher demand due to its convenient location and popular tourist attractions, which could be another reason for the higher prices. Additionally, the hosts with properties in the central region could potentially adjust their pricing strategy to take advantage of the higher demand and optimize their earnings. Alternatively, if a person is looking to rent out their property on Airbnb in the Central region, they may be able to charge a higher price due to the high demand for properties in this area. This information could also be useful for Airbnb competitors or other businesses in the hospitality industry looking to expand or compete in the Central region. Moreover, it is important for Airbnb hosts to take into account the location and features of their property when determining their pricing strategy, as these factors play a significant role in determining the demand and hence the price.

There is a surge in mean prices of properties in different neighbourhoods of Edinburgh during the summer months (March-September). This surge can be attributed to the renowned Edinburgh Fringe Festival and other cultural events during this period, which attract a significant influx of tourists to the city. Edinburgh's rich offering of galleries and historical landmarks further amplifies its appeal as a prime tourist destination.

The observed price fluctuations align with findings from previous research on Airbnb pricing dynamics, wherein special events and seasonal trends are known to impact demand and accommodation prices. Our analysis contributes to the understanding of the Edinburgh Airbnb market and supports strategic decision-making for hosts, guests, and policymakers.

Evaluation of own work: strengths and limitations One of the main strengths of our study is that we identified the importance of amenities in attracting guests and generating higher revenues for Airbnb rental units. This finding can help hosts and property managers improve the attractiveness of their listings and potentially increase their profitability.

However, one limitation of our study is that we relied solely on data from Inside Airbnb, which may not be representative of all Airbnb rentals. This limitation may affect the generalizability of our findings to other regions or countries. In future research, we plan to address this limitation by gathering data from other sources, such as Airbnb's official website, to provide a more diverse and comprehensive dataset. Additionally, we will work to refine and improve the accuracy of our predictive model to enhance the robustness of our findings.

Comparison with any other related work Our study on Airbnb rental units found some similarities with Islam et al.'s study [2] that highlighted the significant impact of geography on determining the prices of Airbnb rentals. However, we also discovered that the set of amenities offered in rental units plays a crucial role in attracting guests, resulting in higher revenues. This finding was backed by our observation that improving the amenity score significantly increased the accuracy of price prediction using Random Forest and XGBoost algorithms.

It is essential to note that Islam et al.'s study was focused on California, which may explain any differences in the findings. Additionally, we used different machine learning algorithms than Islam et al.'s LDA and MESF to analyze our data. These differences in methodology could have influenced the results of our respective studies.

Overall, these studies comparative analysis highlights the importance of considering both geography and amenities while determining Airbnb rental prices. However, differences in methodology and geographic focus may lead to distinct results that must be evaluated with care.

Improvements and extensions To extend this research, it would be beneficial to gather data from other sources, such as Airbnb's official website or other third-party rental platforms. This could provide a broader and more diverse dataset, which would improve the accuracy and generalizability of our findings. Future studies could also explore the impact of regulatory frameworks and legal restrictions on Airbnb rentals. Many cities and municipalities have implemented regulations on short-term rentals, which could impact rental prices and availability. It would be worthwhile to investigate how these regulations impact the profitability of Airbnb rentals in various locations.

Finally, it could also be valuable to explore how the COVID-19 pandemic has impacted Airbnb rental prices and demand. The pandemic has caused significant disruptions in the travel industry, and understanding its impact on Airbnb rentals could provide useful insights for hosts and property managers.

References

[1] *Inside Airbnb*. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit `http://creativecommons.org/licenses/by/4.0/`. 2022. URL: `http://insideairbnb.com`.

[2] Md Didarul Islam et al. "Airbnb rental price modeling based on Latent Dirichlet Allocation and MESF-XGBoost composite model". In: *Machine Learning With Applications* 7 (2022), p. 100208.

[3] Ningxin Peng, Kangcheng Li, and Yiyuan Qin. “Leveraging multi-modality data to airbnb price prediction”. In: *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*. IEEE. 2020, pp. 1066–1071.

[4] Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei. “Airbnb Price Prediction Using Machine Learning and Sentiment Analysis”. In: *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5*. Springer. 2021, pp. 173–184.

[5] Siqi Yang. “Learning-based airbnb price prediction model”. In: *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. IEEE. 2021, pp. 283–288.

[6] Ang Zhu, Rong Li, and Zehao Xie. “Machine learning prediction of new york airbnb prices”. In: *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*. IEEE. 2020, pp. 1–5.