Airbnb listings in Edinburgh

Adam Chapman-Dodd, Ian Little, Jack Taylor

April 4, 2023

1 Overview

The report aims to investigate whether the features of a property listing can be used to predict its short-term rental price, and if some neighbourhoods are more expensive than others. Furthermore, it will analyse the differences in price trends during summer and winter, and whether or not the prices of the Airbnb's have been affected by inflation or the cost of living crisis. We used machine learning to predict the price with the given features of that respective property. We found that there were four standout features that contributed most to the models we created; the number of people the property accommodates, the type of room to rent, the number of beds at the property and the reviews of the surrounding location of the property. To see the price difference between neighbourhoods we plotted the median price per person on a map of Edinburgh. We found that in general, the centre of Edinburgh is more expensive however Cramond and The Pentlands are outliers to this. To analyse the difference in price between summer and winter, we plotted the price during each time. We found that summer is generally much more expensive however winter has a much larger peak for New Years Eve. Finally, to analyse whether the prices of Airbnb's have been affected by inflation or the cost of living crisis, we plotted the graph of price trends for the entire year and found that there was roughly a 12% price increase which follows the inflation in Scotland which was 10.1% [5].

2 Introduction

Context and motivation Our study is in the field of predictive modelling and exploratory data analysis. We use machine learning and statistical methods as well as visualisation techniques to analyse trends over time. Our study could be beneficial to Airbnb owners to help them optimise their pricing strategy, for example, hosts will be able to adjust their prices based on seasonal trends. On the flip side, our study will help to those looking to rent the Airbnbs as it will give them an idea of the price they should be paying for the accommodation they are looking to rent. Furthermore, our study has impacts on wider society as it identifies the neighbourhoods that are most attractive to tourists. This can be used to inform urban planning decisions and to benefit local businesses.

Previous work Researchers have conducted studies to examine the pricing of Airbnb listings in specific areas:

- The impact of Airbnb on San Francisco's housing market and the prices of rent [10]; The study of harvested data from the Airbnb website, finds that the neighbourhood is a strong predictor of the price of a listing.
- The factors affecting the ratings of listings and how Airbnb is affecting the sharing economy [4]; The study of the same harvested data from the Airbnb website observes room type and area are important features in how customers rate their stays.

- The factors affecting the price of Airbnb listings in Davidson County, Tennessee [15]: This study looks heavily into the geographical effect and the distances to certain areas in the change of prices of Airbnb locations. Again, this data was scraped from the publicly available Airbnb cite information.

Objectives We have four main topics of interest:

- How well can features of a property listing be used to predict its short-term rental price?
- Are particular neighbourhoods more expensive than others?
- What are the differences in price trends of properties between summer and winter?
- Are the prices of properties affected by inflation/cost of living crisis?

3 Data

Data provenance The mission of our data source *Inside Airbnb*, is to offer information and advocacy regarding the effects of Airbnb on residential communities. This study utilises publicly available data from the Airbnb website to curate a distinct data set, which has been made available on the platform. We obtained the data by reading in the CSV and geoJSON files provided. While we analysed the effect of different factors on price, *Inside Airbnb* uses this data to highlight the problems Airbnb is causing legally and financially in some areas. Their website goes into more depth as to where these problems arise. The guidelines state 'Only take the data you need' which means their T&Cs allow us to use the data for our project [2].

Data description The data [1] we used came from two CSV files and a geoJSON file.

The first CSV gave us information about the listing with 75 columns and 7389 entries in total. To answer our objectives, we manually selected 40; *host is superhost, host listings count, host total listings count, host has profile pic, host identity verified, neighbourhood cleansed, room type, accommodates, bathrooms text, bedrooms, latitude, longitude, beds, amenities, price, minimum nights, maximum nights, minimum nights avg ntm, maximum nights avg ntm, has availability, availability 30, availability 60, availability 90, availability 365, number of reviews, number of reviews ltm (last twelve months), number of reviews l30d (last 30 days), review scores rating, review scores accuracy, review scores cleanliness, review scores checkin, review scores communication, review scores location, review scores value, instant bookable, calculated host listings count, calculated host listings count entire homes, calculated host listings count private rooms, calculated host listings count shared rooms and reviews per month*.

The calendar CSV gave us information about each listing's price on every date from mid-December 2022 to mid-December 2023, whether it was still available to book on specific dates, the minimum and maximum periods it could be booked for. It contained 2696636 entries and 7 columns. The columns were *listing id, date, available, price, adjusted price, minimum nights and maximum nights*.

The geoJSON file gave us data about the geographical features of Edinburgh. It contained polygons that can be graphed to outline a map of Edinburgh with clear borders for all of its neighbourhoods.

Data processing Each objective required slightly different cleaning. We started by cleaning the data in general, then cleaning specifically to answer each question. As mentioned above, we only read in columns that could be used to help answer the questions. We had to reformat the price and bathroom columns using regex and then parsing to a float. Then we dropped duplicates and NaN values. After some manual searching, we found an entry with an accommodates value of 0, which we removed.

When considering the feature importance for predicting the price of a short-term stay, we removed any entries that had a minimum stay of over 30 nights. We standardised the price column using a new variable we created called ‘price per person’ which was the total price divided by the respective entry in the accommodates column, this gave us a better understanding of what would qualify for an outlier. These outliers were removed using the 95% Empirical Rule which removes the top and bottom 2.5% of the data and then we took the log of the price column to normalise the distribution, Figure 1. To clean the amenities we found out which amenities had the most occurrences and then for each of the top eight appearing amenities we used regex to create a new column. As we are using regression, we utilised sci-kit learns ‘LabelEncoder’ function which numerically encoded the columns with categorical variables. Since our numerical data is measured in different units, we utilized the ’StandardScaler’ tool from scikit-learn, which standardizes the data by removing the mean and scaling it to have a unit variance.

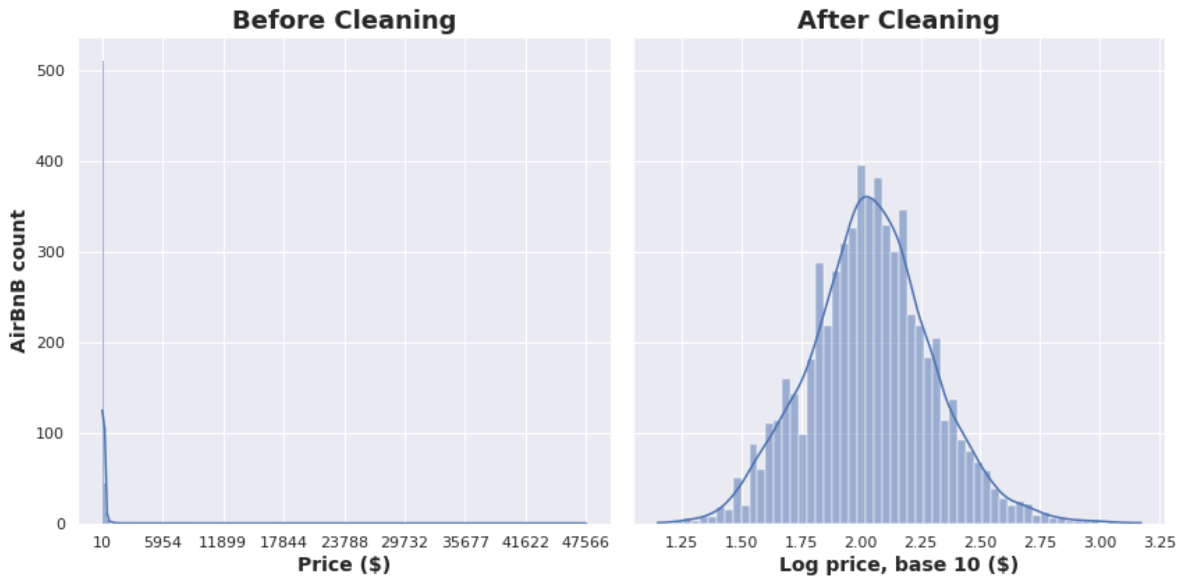

Figure 1: Cleaning of price feature

Considering the price difference between neighbourhoods, we used ’price per person’ which is defined above. This was to give a clearer display of how expensive a stay is. The geoJSON file contained a column called ’neighbourhood’ that contained the same neighbourhood areas as the column ’neighbourhood cleansed’ in the listings file. By renaming the geoJSON column to ’neighbourhood cleansed’, we were able to merge them seamlessly. This ensured that every listing had its corresponding neighbourhood geographical map interpretation. The ’price per person’ values were heavily skewed to the right so outliers were removed via the empirical rule. The data was still skewed to the right but by much less.

To compare the trends in price between summertime and wintertime. We created two new dataframes containing the dates of the UK winter and summer time. We define the UK summertime as beginning on the 21st of June 2023 and ending on the 23rd of September [8] and the UK wintertime as beginning on the 21st of December 2022 and ending on the 20th of March 2023 [9]. Similar to before, an analysis was conducted to assess the skewness of the price data, revealing a rightward skew in both the winter and summer prices. As a result, the empirical rule was applied to separately remove the outliers causing the skew in both datasets. We then ran an updated analysis and found that the data was still skewed to the right but by a significantly less amount.

Analysis of the effect of inflation/cost of living on price required no further cleaning.

4 Exploration and analysis

First, we want to investigate how well features of an Airbnb property can be used to predict its short-term rental price. After the previously mentioned data cleaning, we plotted the correlation between the 'log price' column and the features as seen in Figure 2.

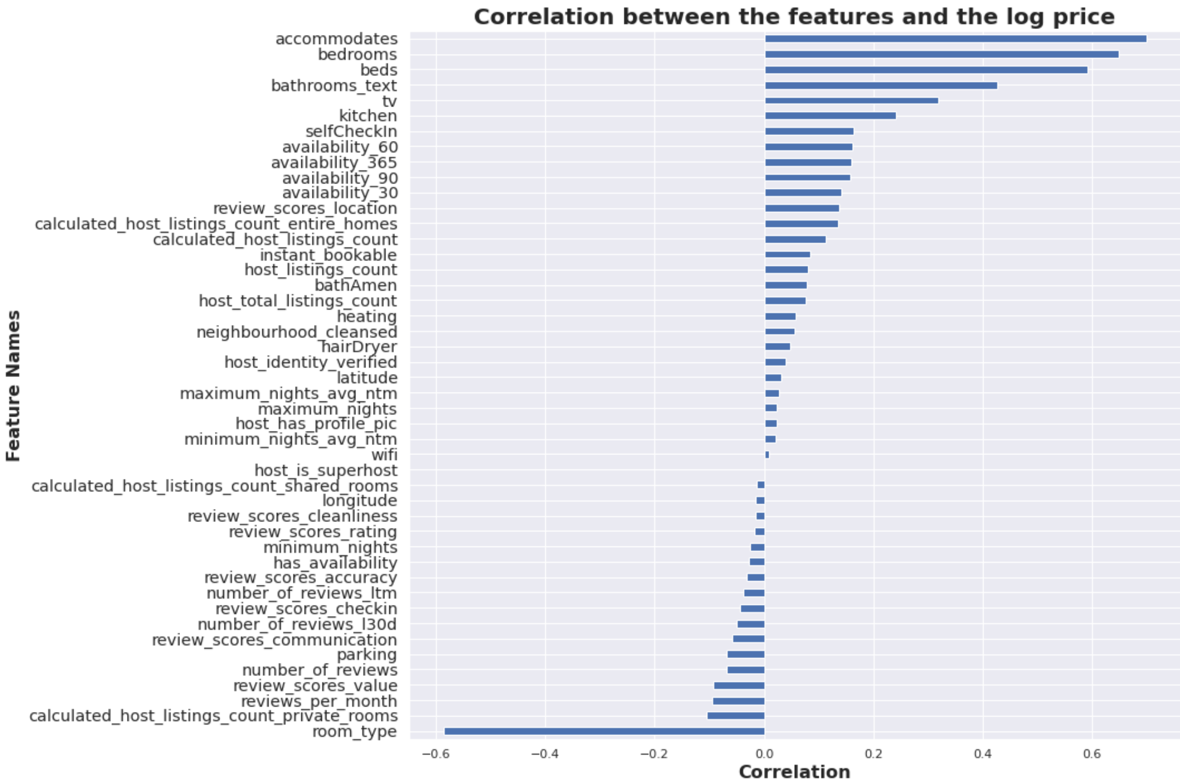

Figure 2: Correlation plot of all the features against log price

We can see four variables with a moderate to strong correlation; accommodates, bedrooms, beds and the room type. Therefore, a preliminary observation would be that the most important features to predict price of the Airbnb would be the number of people that it accommodates, the room type (entire place, private rooms, hotel rooms, and shared rooms), the number of bedrooms and the number of beds it has. However we must consider the fact that *correlation does not imply causation* so we need to apply a more rigorous method to obtain the most important features.

We chose to train five different regression models with the price as the dependent variable and the features of the Airbnb as the independent variables. The types of regression included; multiple linear regression (models the linear relationship between a dependent variable and two or more independent variables), k-nearest neighbours regression (determines the classification of a new observation by looking at the classification of its k-nearest neighbours in the training data), random forest regression (multiple decision trees that predict continuous variables), gradient boosting regression (ensemble learning with iteratively improved models for prediction) and XGBoosting regression (distributed gradient boosting with decision trees for prediction). To evaluate the performance of the model, we employed adjusted r^2 and mean squared error (MSE) as the performance metrics. Adjusted r^2 assesses the amount of variance in the response variable that can be explained by the predictors, whereas MSE quantifies the average of the squared differences between the predicted and actual values.

Upon analysis of the results from each model, it was observed that the top-performing models were the random forest, XGBoost and gradient boosting regression as they all had adjusted r^2 values greater than 0.7 (this indicates that each model explains a significant proportion of the variance in the target variable) and MSE values less than 0.02 (each model's predictions are relatively close to the actual values) which can be seen in Table 1. We utilised the feature importance function provided by the scikit-learn library. The feature importance function measures the contribution of each feature in the model's prediction accuracy, and visualises their relative importance in a plot. To evaluate the level of consistency between the models, we refer to Figure 3, which displays the plotted feature importance values for each model.

Figure 3: Feature importance graph for different regression types

Figure 3 shows that the three models agree on the main predictors of short-term rental price. By inspection, we can see the gradient rapidly declining after the fourth predictor in 3 so is clear that the most important features according to the regression analysis are the number of people the Airbnb accommodates, the room type, the number of bedrooms and the reviews of the surrounding location. We will consider the rest of the features to be unimportant. Using a form of dimension reduction, we then reran the three well-performing regression models with only these important features as predictors to see how the model performs. We can see the MSE and adjusted r^2 after performing dimension reduction in Table 1 is only marginally lower which shows that the unimportant features were contributing very little and that our choice of important features are contributing the most to the accuracy of the model.

Regression type	**MSE before dimension reduction**	**Adjusted r^2 before dimension reduction**	**MSE after dimension reduction**	**Adjusted r^2 after dimension reduction**
Multiple Linear	0.024440	0.631940	n/a	n/a
K Neighbors	0.033129	0.501102	n/a	n/a
Random Forest	0.018309	0.724273	0.024284	0.605803
XGBoosting	0.019518	0.706073	0.025155	0.59202
Grad. Boosting	0.018898	0.715402	0.002644	0.609171

Table 1: Model evaluations before and after dimension reduction

Next, we want to investigate if some neighbourhoods are more expensive than others. To do this we used the geoJSON data detailed previously, which allowed us to add a map of Edinburgh to our visualisation. We set the hue of each individual neighbourhood area to be the median price per person for all the properties that reside within that area, this can be seen in Figure 4. The decision was made to utilise the median price per person considering its lower sensitivity to outliers when compared to other measures such as the mean.

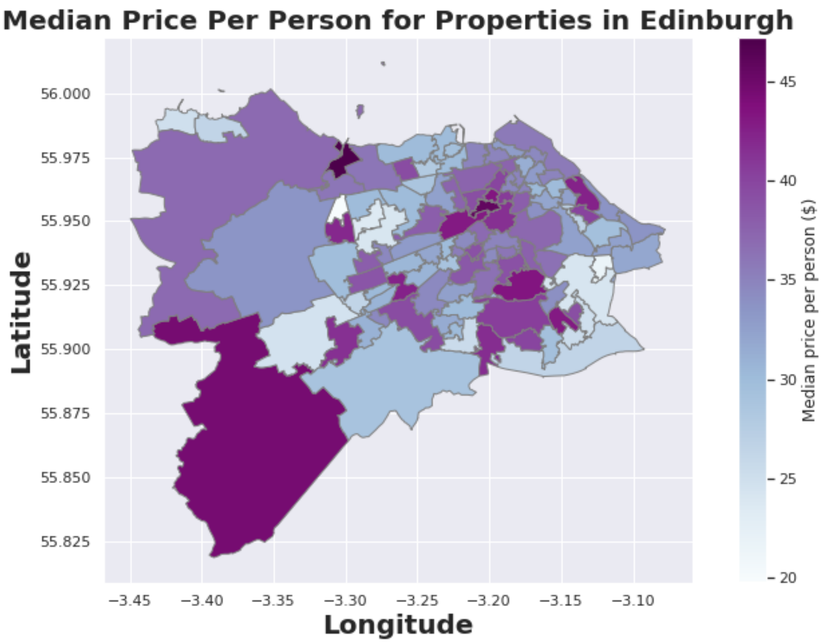

Figure 4: Property values on map of Edinburgh

This visualisation shows us that the centre of Edinburgh is one of the most expensive neighbourhoods, specifically the 'Old Town, Princess Street, and Leith Street' area. As we leave central Edinburgh, we generally encounter areas that are less expensive, with a few exceptions. We can see that the most expensive neighbourhood is 'Cramond', located North West of the city centre. Cramond is an ideal location for a family getaway as the primary rental accommodations are beach-side cottages [11]. The Pentlands area (the largest area on the map) is also an expensive area to rent a property. The properties here are cottage-like properties that tend to be larger and more luxurious, also intended for family stays [12]. The Heriot-Watt campus is another expensive area, individuals visiting the university and students who are unable to secure accommodation on-campus are directed to rent Airbnb properties in the vicinity [13]. The cheapest areas (e.g. 'East Craigs North') are residential neighbourhoods with smaller apartment housing. Thus they are less desirable as they aren't near any Edinburgh attractions.

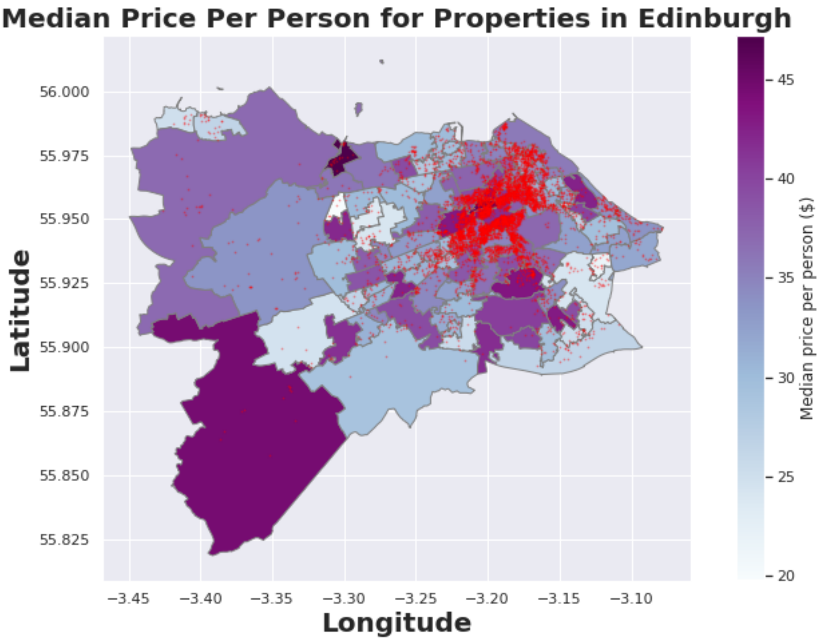

Figure 5: Geographical distribution of Edinburgh properties

We can see the distribution of Airbnb properties in Figure 5. It shows us that the majority of properties listed on Airbnb reside within the main city of Edinburgh. The outskirts of Edinburgh are lightly polluted with Airbnb properties.

We then conducted an analysis to assess the difference between the price trends during the summer and winter periods. To accomplish this, we calculated the median price of each property for every date within each respective season. We chose the median because both the summertime and wintertime prices were skewed to the right after removing outliers. The median is therefore a better representative of the price in an area when data is skewed. In our subplot, we can see that both plots experience waves in price. This translates directly to weekends and weekdays with weekends being the peaks and weekdays being the troughs.

In Figure 6, we can see that the summer period is generally more expensive than the winter period, most likely due to vacations being more popular in summer. It's worth noting that winter marks the highest spike in median price, particularly on New Year's Eve (Hogmanay). During this time the median price can reach as high as $150 a night, which can be attributed to Edinburgh's world-renowned New Year's Eve celebrations, particularly the Edinburgh street party [7]. It brings in many tourists from around the world which gives context to the price of Airbnb properties increasing on this day.

Another spike occurs just before Christmas Eve, which can be attributed to the popularity of Edinburgh during the holiday season. Numerous seasonal attractions are set up during this time of year, and people often travel back to the city to spend Christmas with their families [6].

We have observed three other prominent spikes in prices, occurring specifically on weekends when the Scotland National Rugby Team played their Six Nations home games at Murrayfield in Edinburgh. Traveling fans will have to find accommodation in Edinburgh resulting in an increased demand for Airbnbs, thus becoming more expensive.

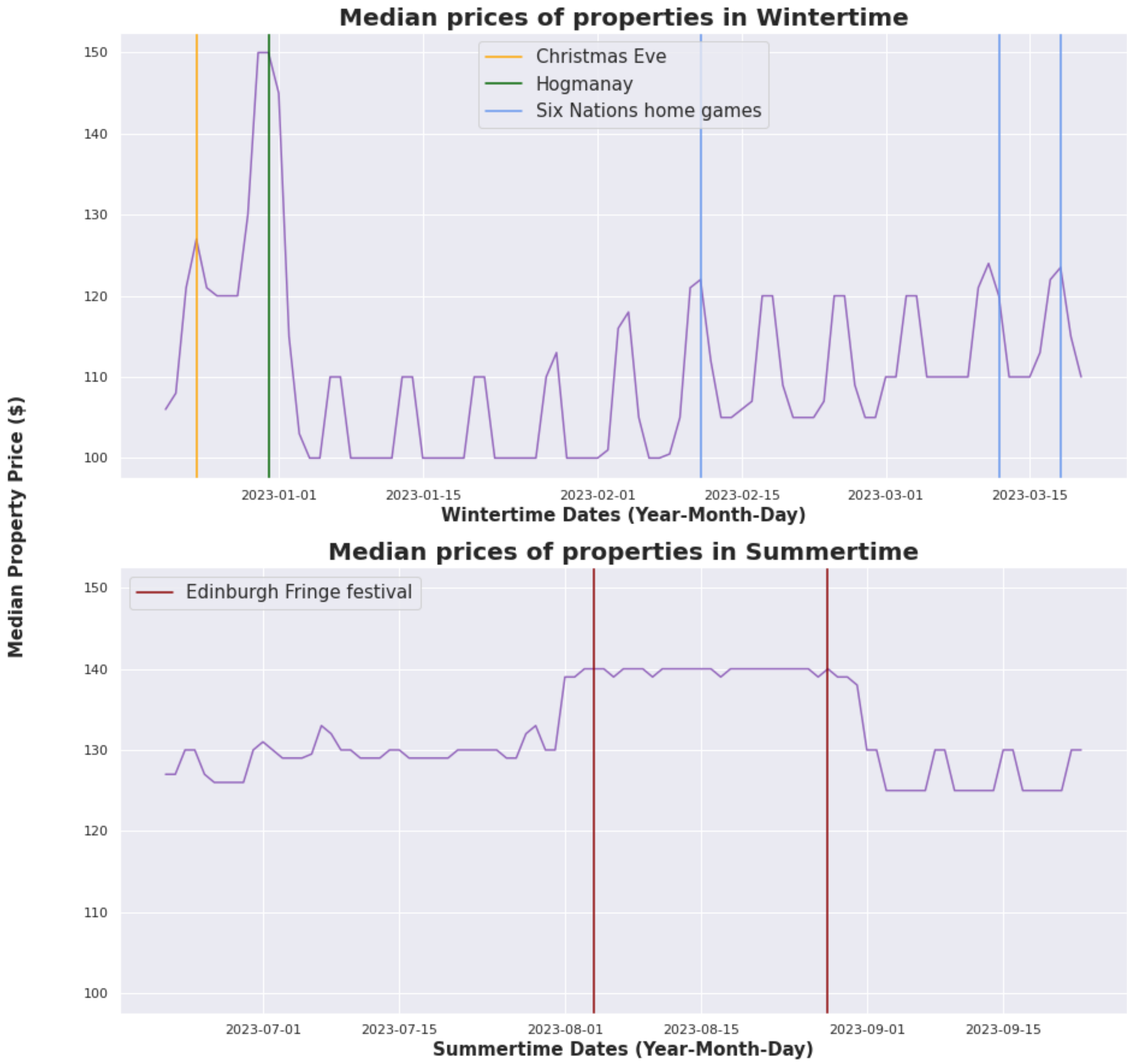

Figure 6: Summer and winter price trends

During summertime, the median prices are much more expensive than in winter. There is a noticeable period in which prices experience a significant increase. This is primarily due to the Edinburgh Fringe Festival, a massive comedy festival that draws tourists from all over the world [14]. During this period the waves seen previously become nullified, implying that there is a constant expensive price during The Fringe period. Not only do tourists intending on going to the festival need a place to stay, but also the performers and their crews. After The Fringe ends, we can see the price begin to fall off to become cheaper than at the beginning of summer goes back to displaying waves.

Lastly, we wanted to extend our analysis of price trends in summer and winter periods alone and simply conduct an analysis of the price trends across the year of data provided (16/12/2022 - 15/12/2023). To accomplish this, we looked at the percentage increase (or decrease) of prices of each property compared to the price on the first date provided. The mean for the price change across all listings was taken for each date and this was plotted in Figure 7. As some of the listings did not change their price throughout the year, we applied the mean. This ensured we could capture all the necessary data. As the prices were all standardised to show proportions from their respective starting prices, the skewness that originally affected these prices has been eliminated.

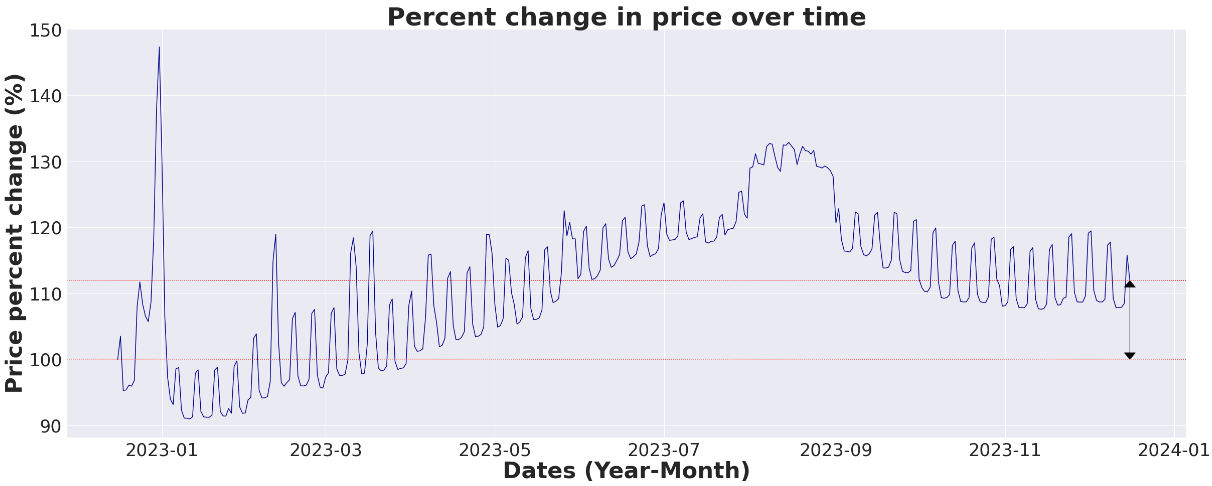

Figure 7: Mean percentage change of listing price over time

In Figure 7, we can see that the average percentage of price change spiked for Hogmanay [7] and rose drastically for the summer months, The Fringe in particular [14], while the spring and autumn months brought the price changes back down. This is analysed in greater detail in the previous section.

Previously, the prices across this period have already been discussed with regards to the spikes and general increases (summer generally has higher prices with drastic increases for The Fringe, Six Nation rugby games, and the festive period in winter), however, we now look at the impact inflation has had on the prices of stays in Airbnb properties. We can see that, as expected, the cost is rising throughout the year. While we cannot be sure if this is due to the increasing cost of living and general inflation, we can study the differences between the start and the end of this plot.

Shown by the arrow in Figure 7, there is a large average price increase across the year. If the percentage changes were purely due to the events in Edinburgh (The Fringe, Rugby matches, etc.), then we would expect to see this price difference to drop back to around the same it was at the start. However, this is not the case. The average cost of an Airbnb rental rose by approximately 12%.

As we are looking at the same dates a year later, we can inculpate this increase on the current rising cost of living. As of winter 2022, inflation in Scotland was at approximately 10.1% [5]. Furthermore, we can also blame an overall increase in living costs on the growing number of short-term rentals [3]. Airbnb is enabling more and more property owners to increase the costs of long-term rent and open up their homes to short-term rentals. Edinburgh does indeed rely heavily on tourism but the number of short-term lets in the city is unsustainable and the rise in costs portrays this clearly.

5 Discussion and conclusions

Summary of findings The differing prices of Airbnb listings are the result of a large range of factors. The findings in this study show that there are particular aspects that influence costs more than others such as the number of people a property is for, the neighbourhood the property is located within, and the time of year the booking is for.

Using the data acquired from the Inside Airbnb website [1], we have focused on analysing the individual features of Airbnb listings and the correlation with price. While it may seem obvious, this validation tests between regressions and an initial look at correlations confirms that the most important feature to predict the price was the number of people a property accommodates. We also see that the room type and the number of bedrooms were important factors in determining the price of a listing.

Furthermore, the location of a given listing had a large impact on the price. Locations closer to the city centre and in luxury holiday areas (e.g. The Pentlands) cost more. In general, the price decreases as one goes further away from the city centre. We can also see that residential areas seem to be the cheapest.

There is a distinct difference in price trends between winter and summer. The prices in winter are much more volatile with many spikes, whereas prices in summer are higher but much more steady - only increasing significantly during the Edinburgh Fringe Festival [14].

The time of year is an influential factor in the price of listings, as can be seen by the large differences in price. Edinburgh is a tourist destination and thus there are events (The Fringe, Hogmanay, international rugby games) that cause spikes in prices.

Moreover, there is a clear increase in the price of a listing which could be due to the impact of inflation across Scotland and the rest of the UK and the current cost of living crisis [5]. This has a not-so-instantaneous but substantial impact.

In short, it is trivial that the number of people a property can accommodate will be the most influential factor in the cost of a stay as more people means more room required, increased cleaning time and higher utility costs. Of course this is naturally going to be the case; even though the time of year and the room type do sway the price, it would be farcical to deny the relation between number of guests and price.

Evaluation of own work: strengths and limitations The strengths of our work are our regression models. We have strong adjusted r^2 and MSE values. Furthermore, we have crossed checked these models and they agree on the feature selection. Our knowledge of machine learning and Matplotlib allowed us to create and visualise these models and decide on the weight each factor had on the price of listings. Our visualisations are clear and concise which helped in pulling insights to discuss our objectives.

Limitations include only using data from Edinburgh. If we had designed the model using data from multiple cities we would have achieved a model that could have been applied to any city. In addition to this, Airbnb is not the only site for renting properties in Edinburgh which may have shown more of an insight into how competitive these listing prices are. As of the publication of this project (April 2023), the Edinburgh Fringe Festival has yet to happen. Prices can therefore still be altered which could cause an even bigger increase in price than we have been able to show.

Comparison with any other related work Previous papers have drawn conclusions as to how price varies for different Airbnb listings, our findings align with these other studies. Similar to the study in NYC [4] and San Francisco [10], we found neighbourhood and room type were important features for predicting the price of a listing.

Improvements and extensions To further improve the accuracy of our analysis, it would be beneficial to include more locations in our study. By incorporating places with diverse levels of tourism and attractions, we would be able to gain a better understanding of how various variables impact different areas. Moreover, given our study's limitations, examining more listings beyond Airbnb would provide a more comprehensive picture of how the data we have currently compares to that of alternative websites and accommodations in Edinburgh. Lastly, the same data but from other years would have allowed for an even more accurate analysis, specifically when looking at the Edinburgh price trends throughout the year.

References

[1] Inside Airbnb. "Data - Edinburgh". In: *Inside Airbnb* (2023). Retrieved on 03 April 2023. URL: http://insideairbnb.com/edinburgh.

[2] Inside Airbnb. "Data Policies - Community Guidelines". In: *Inside Airbnb* (2023). Retrieved on 03 April 2023. URL: http://insideairbnb.com/data-policies.

[3] Alison Campsie. "Save our city from out of control rents and protect those who call Edinburgh home". In: *Edinburgh News* (2021). Retrieved on 03 April 2023. URL: https://www.edinburghnews.scotsman.com/news/opinion/columnists/save-our-city-from-out-of-control-rents-and-protect-those-who-call-edinburgh-home-daniel-johnson-3464462.

[4] Jason Clampet. "Airbnb in NYC: The Real Numbers Behind the Sharing Story". In: *Skift* (2014). Retrieved on 28 March 2023. URL: https://skift.com/2014/02/13/airbnb-in-nyc-the-real-numbers-behind-the-sharing-story/.

[5] Scottish Government. "The Cost of Living Crisis in Scotland: analytical report". In: *Scottish Government: Publications* (2022). Retrieved on 03 April 2023. URL: https://www.gov.scot/publications/cost-living-crisis-scotland-analytical-report/#:~:text=The%5C%20report%5C%20has%5C%20been%5C%20produced,last%5C%20summer%5C%20inflation%5C%20was%5C%202%5C%25.

[6] Tasha Hall. "Edinburgh named one of the world's 10 best Christmas destinations". In: *Scottish Daily Express* (2022). Retrieved on 03 April 2023. URL: https://www.scottishdailyexpress.co.uk/news/scottish-news/edinburgh-named-one-worlds-10-28660830.

[7] Jessica Norah. "A Guide to Celebrating Hogmanay in Edinburgh Scotland". In: *Independent Travel Cats* (2022). Retrieved on 03 April 2023. URL: https://independenttravelcats.com/hogmanay-edinburgh-scotland-festival-guide/.

[8] Met Office. "When does summer start?" In: *Met Office* (2022). Retrieved on 03 April 2023. URL: https://www.metoffice.gov.uk/weather/learn-about/weather/seasons/summer/when-does-summer-start.

[9] Met Office. "When does winter start?" In: *Met Office* (2022). Retrieved on 03 April 2023. URL: https://www.metoffice.gov.uk/weather/learn-about/weather/seasons/winter/when-does-winter-start.

[10] Carolyn Said. "Window into Airbnb's hidden impact on S.F." In: *The San Fransisco Chronicle* (2014). Retrieved on 28 March 2023. URL: https://www.sfchronicle.com/business/item/airbnb-san-francisco-30110.php.

[11] TripAdvisor. "Cramond Village". In: *TripAdvisor* (2022). Retrieved on 03 April 2023. URL: https://www.tripadvisor.co.uk/Attraction_Review-g186525-d213427-Reviews-Cramond_Village-Edinburgh_Scotland.html.

[12] TripAdvisor. "Pentland Hills". In: *TripAdvisor* (2022). Retrieved on 03 April 2023. URL: https://www.tripadvisor.co.uk/Attraction_Review-g186525-d1099333-Reviews-Pentland_Hills-Edinburgh_Scotland.html.

[13] Heriot Watt University. "Temporary / short-term accommodation". In: *Heriot Watt University* (2023). Retrieved on 03 April 2023. URL: https://www.hw.ac.uk/uk/edinburgh/living/off-campus/temporary-short-term-accommodation.htm.

[14] Danyel VanReenen. "Edinburgh Fringe 2022: Airbnb prices in the capital more than double for August". In: *Edinburgh Live* (2022). Retrieved on 03 April 2023. URL: https://www.edinburghlive.co.uk/news/edinburgh-news/edinburgh-fringe-2022-airbnb-prices-24344947.

[15] Zhihua Zhang; Rachel J. C. Chen; Lee D. Han; Lu Yang. "Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach". In: *MDPI* (2017). Retrieved on 28 March 2023. URL: https://www.mdpi.com/2071-1050/9/9/1635#B4-sustainability-09-01635.