Unlocking Insights: A Data Science Analysis of Airbnb Listings

April 4, 2023

1 Overview

This paper describes a data science project focused on investigating the characteristics of Airbnb listings and identifying the key factors that influence their price[1]

The results of applying regression analysis revealed that while the given features are limited in predicting price and popularity, the model successfully identified some factors that influence the price and popularity of an Airbnb listing. As an extension, we have also endeavoured to develop a methodology aimed at distinguishing authentic listings from counterfeit ones.

2 Introduction

Context and motivation

1. **What is this area of study?**

   The area of study for this data science project is Airbnb listings, which involves analysing a large dataset of properties in Edinburgh listed on the Airbnb platform. The analysis will focus on understanding the factors that influence property rental rates and occupancy levels, such as location, amenities, property type, and seasonal fluctuations.

2. **Why is it interesting to investigate?**

   - Airbnb has rapidly disrupted the traditional hospitality industry, providing travellers with more options and flexibility when it comes to choosing their accommodations.
   - By analysing the types of properties that are most popular and the factors that drive rental decisions, we can gain a better understanding of the evolving travel preferences of modern consumers.
   - Understanding of pricing is critical from practical as well as theoretical perspectives, as the pricing for hotel room rentals drives consumer decision-making and hotel profitability.
   - This paper [2] described hosts as being confounded when prompted by the platform to set a price and not being able to determine the real market value of their offerings.

Previous work Here are some of the previous research papers we've looked at:

- Jiao[3] has found that Airbnb rentals were more likely to be in better neighborhoods closer to the city center and with good transit service.

- According to Mao's research [4], the process of identity verification on the Airbnb platform does not have a substantial impact on future bookings. In other words, guests do not make judgments about the quality of hosts based on the verification of their identities provided by Airbnb.

[1] We used the Hedonic Price Model as mentioned in this paper[3] and popularity. Using data cleaning and transformation techniques, exploratory data analysis, and machine learning models such as Ridge Regression and NLP LDA modelling, we've uncovered relationships between different features.

- In a previous study, [1] used segmentation analysis to examine the motivations of tourists for selecting Airbnb over traditional hotels. The study identified five distinct tourist segments with varying motivations for choosing Airbnb, including those who prioritize authenticity, local experiences, social interaction, value for money, and convenience.

Objectives Here are the questions that we've addressed in this report:

1. What factors primarily determine price/ popularity?
   (a) Make a model to exclude the importance of different features
   (b) In what way do the two factors, price and popularity, vary based on location? This can be analyzed using both mapping and modeling techniques, and then the results from the two approaches can be compared and checked for consistency.
2. How do various features correlate with each other, as revealed by a heatmap?
3. **Extension:** What methodology can be employed to identify genuine and fraudulent listings by utilizing factors such as predicted versus actual prices, the date of the last review, the number of reviews, and availability within the dataset?

3 Data

Data provenance The dataset was originally created by Murray Cox, and has been obtained from insideairbnb.com. He scraped publicly available data from Airbnb's website and compiled it into a dataset to shed light on the impact of Airbnb on housing affordability and availability in cities around the world.

Data description The Airbnb listings data for Edinburgh contains information about the following categories:

1. **Physical Characteristics**: `neighbourhoods_cleansed, property_type, room_type, number_of_bedrooms, bathrooms,` and `amenities` etc.
2. **Info about the host**: `name, about, host_response_rate, host_response_time, host_total_listings_count, host_identity_verified` etc.
3. **Info about the bookings**: `availability, minimum_nights, maximum_nights, price` etc.
4. **Info about reviews**: `review, number_of_reviews, review_score_rating` etc

Data processing

1. **Handling NaN values**: Most of features had some NaN values. We replaced NaN values with the mean value or zeros depends on the situation.
2. **Dealing with values the are not float**: The linear regression model only accept float type. We managed to convert the values in some columns to floats whilst keeping most information remains.
3. **Dealing with categorical features** Features like `room_type` and `property_type` were categorical. We applied one hot encoding to quantify them, that each category becomes a feature.
4. **Dealing with the price feature** We converted the price feature from a string type to a float type to enable arithmetic operations. Although the currency notation differed between our dataset and the Inside Airbnb website, the numerical values were equivalent and it should be in pounds. We also addressed the skewness of the price feature in a later stage.

5. **Encode amenities feature** We encode amenities feature with the a column amenities_count which is the count of amenities list.

6. **Dealing with outliers** After identifying outliers in our listings dataset with unreasonably high prices given their facilities by visiting the Airbnb website, we dropped rows with outlier prices (threshold use = 3000).

4 Exploration and analysis

We explored the dataset by first examining each feature in the dataset individually, that is univariate exploration. Then, to identify more complex relationships and patterns, we did bivariate and multivariate analysis.

Univariate Exploration

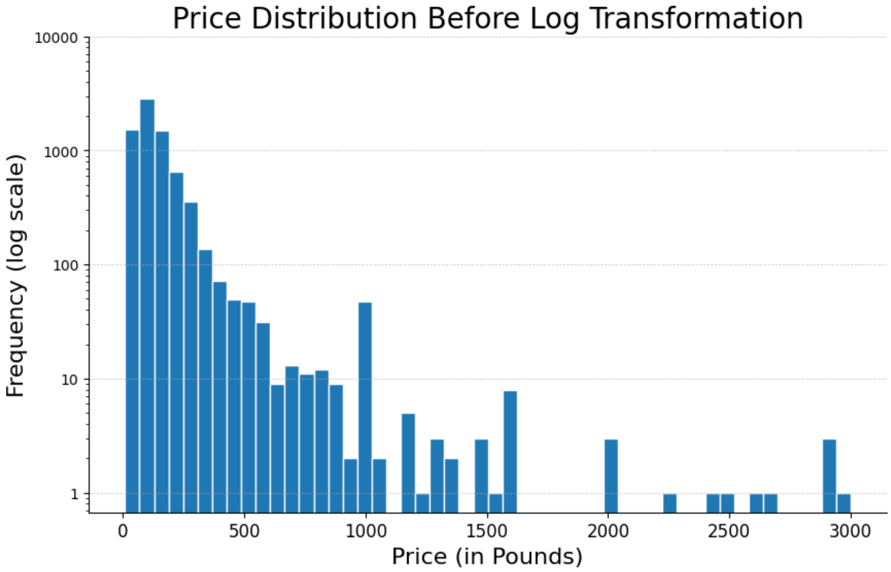

Figure 1: Skewness Analysis for Price Feature

1. `room_type` We found that the most common property type was an entire apartment(70.0%), followed by private rooms(29.4%). Shared rooms and hostel rooms have a joint proportion of 0.6%.

2. `price` The average price per night was £169. We observed that there are outliers before filtering out them, one outlier has value more than £47000. We looked into the distribution of price variable by plotting a histogram. 1 It turned out to be highly skewed, we addressed it later.

3. `review` We did sentiment analysis on review feature to find out that there were 7180 positive reviews, 53 negative reviews, and a total of 141 neutral reviews. Which coincides the fact that most review has ratings.

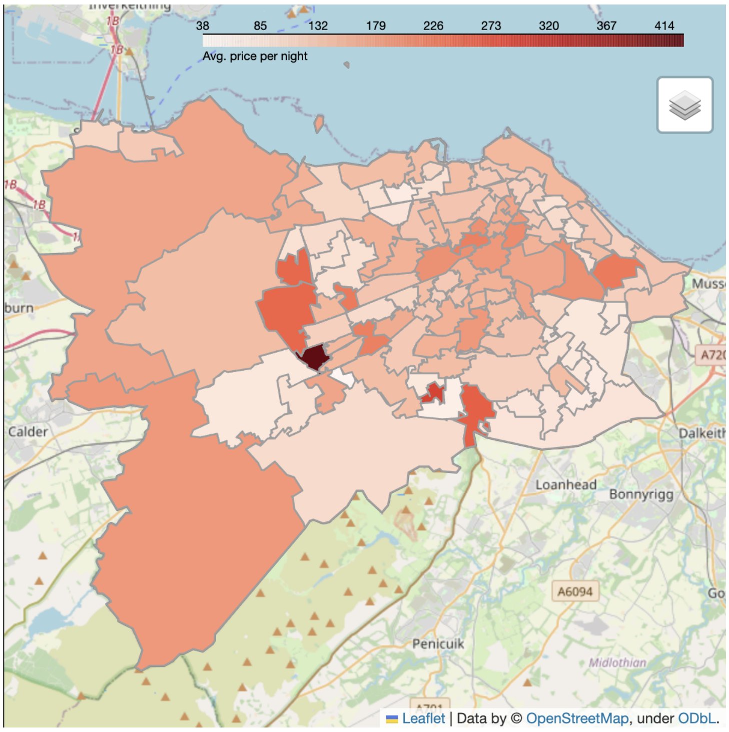

Figure 2: Average price per night according to neighbourhoods

4. `location` The neighbourhood with most listings - Old Town, Princes Street and Leith Street , a total of 802 listings, nearly 11% of all of Edinburgh Airbnb Listings.

Bivariate Exploration

One main methodology used in this section is correlation analysis, which is a statistical method. We have also created two visualisations (4 and 4) on maps by average price and average reviews per month to see if any neighbourhoods are more popular or expensive.

Outcomes of Correlation Analysis 3 :

- We found that the price per night had the strongest positive correlation with the number of bedrooms, bathrooms and people accommodated in the listing, followed by number of bathrooms which has a weaker correlation. Moreover, the four stated features are strongly correlated with each other. This is the multicollinearity of features, this may cause problem in our further explorations.

- `reviews_per_month` does not have very strong correlations with the selected features. A few features worth note that it has correlation with are `amenities_count`, `host_response_rate` and `host_acceptance_rate`. These correlations suggested that listings with an actively responding host and well equipped facilities tend to be more popular.

- `price` was negatively correlated with the reviews per month, suggesting that popular listings may offer a lower price without comprising service [insert correlation value].

- `host_identity verified` was positively correlated with `number_of_reviews`.

Findings and interpretations for the maps

- Price were widely distributed across various neighbourhoods.

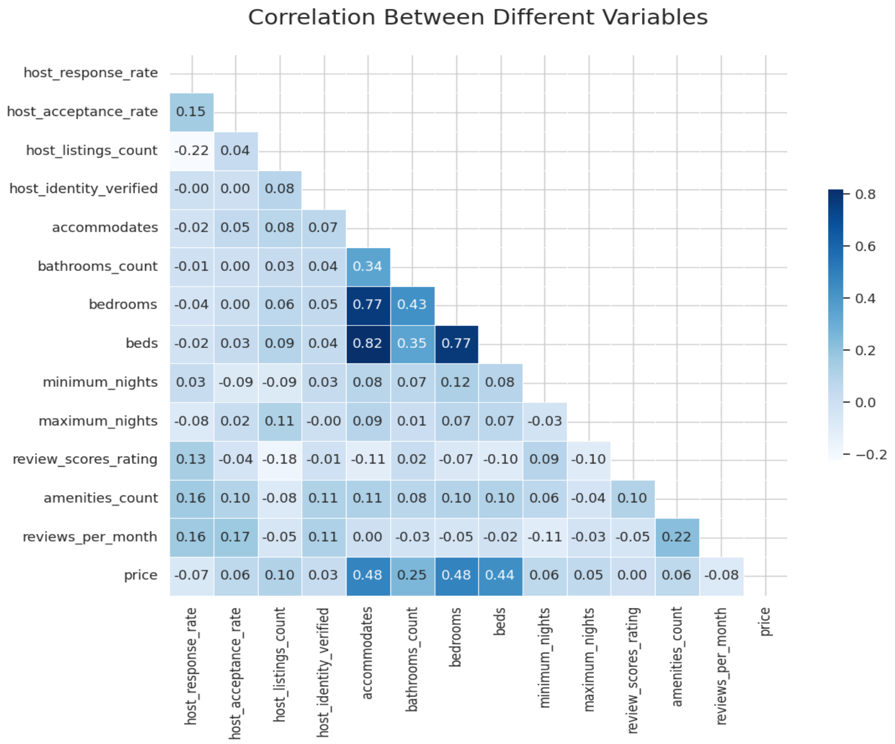

Figure 3: Correlation Analysis

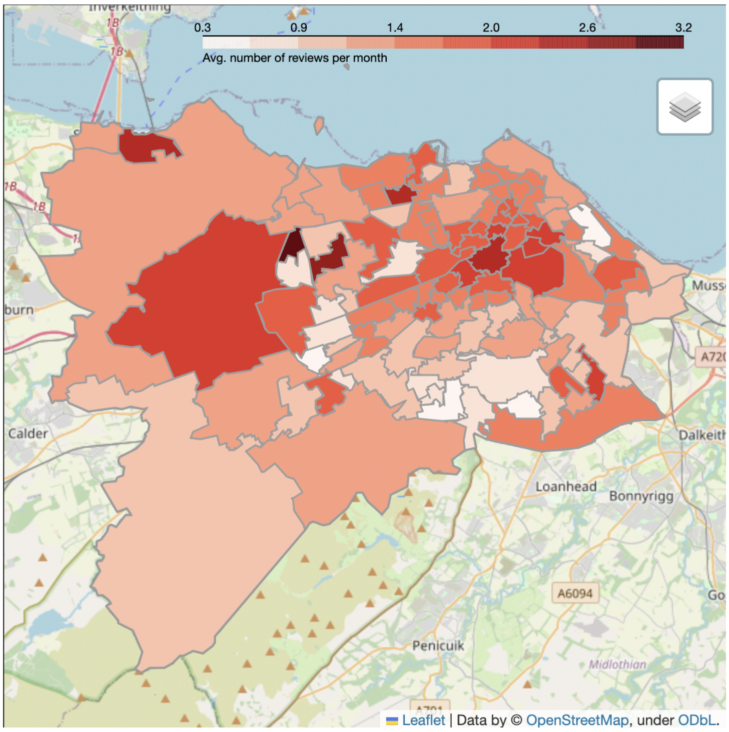

Figure 4: Average number of reviews in each neighbourhoods

Table 1: Feature Importance Analysis: Top 11 Features and their Coefficients for Price and Reviews per Month

`price`		`reviews_per_month`	
Feature	*Coefficient*	*Feature*	*Coefficient*
Accommodates	0.279	Within an hour	0.495
Entire home or apartment	0.098	Host is Superhost	0.282
Beds	-0.091	Amenities Count	0.191
Shared Room	-0.090	Review Scores Value	0.180
Private Room	-0.086	Old Town, Princes Street and Leith Street	0.179
Review Scores Value	-0.083	Host Acceptance Rate	0.166
Reviews Per Month	-0.078	Calendar Host Listings Count	-0.135
Old Town, Princes Street and Leith Street	0.073	Private Room in Condo	0.123
bedrooms	0.072	bedrooms	-0.112
Host Listings Count	0.069	Tollcross	0.105
Host Response Rate	-0.062	Within a few hours	0.095

- Some of the most expensive neighbourhood, such as The Calders and Oxgangs has unreasonably high value of average price. We looked into it and found that it is caused by fake listings in the area. The average price of neighbourhood with more listings is closer to the actual price level, since outliers have less impact on them. Some neighbourhoods with the actual expensive price level are New Town West, Deans Village and Hillside and Calton Hill.

- The graph for the average reviews per month does a better job. It reflects similar to the actual popularity level of airbnb listings in the area. Some of the most popular areas are East Craigs North, Corstorphine North, Queensferry East and Old Town, Princes Street and Leith Street.

- The Old Town, Prince Street and Leith Street is both the most popular and expensive neighbourhood according to the model, not considering other features of the listings.

Multivariate analysis, and Stat and ML methods used in the study

We performed multivariate analysis to identify more complex relationships and patterns in the data. We implemented ridge regression, which is linear regression with L2 regularization, to predict the price per night and reviews per month based on a number of different features.

Ridge regression loss function:

$$Loss = Error(Y - \widehat{Y}) + \lambda \sum_{1}^{n} w_i^2 \tag{1}$$

Before fitting the model, we applied standardization to the feature dataset, since the linear regression perform better on standardized data. We spotted high skewness in price feature, which would have bad influence on the model. To address this problem, we implemented log transformation and fitting the model with the new logprice feature instead.

The models went through 5-fold cross-validation, achieved mean R^2 score for 0.49 and 0.22 for price and popularity respectively on the testing set. We represent the importance of some of the most important features as a table.

Feature Importance Table 1 :

- This table 1 presents the top 11 features that are most strongly associated with both the price and the number of reviews per month for the given Airbnb listings.

- The features are ranked based on their absolute coefficients, which indicate the strength and direction of their relationship with the target variables.
- Coefficients have either a positive or negative sign, indicating the direction of association with the target variables.

Interpretation:

- The results suggest that the number of people accommodated, the type of room (e.g., private or shared), and the location of the listing are among the most important factors that influence both the price and the popularity of Airbnb listings in this area.
- Acommodates is the dominating feature for price, and its correlated features do not have as high the coefficient, this is because the regularization prevent them being to large.
- Listings that offer their services at a lower price are more loved by customers and hence get more reviews.
- The Within an hour feature has the highest correlation with reviews per month, which implies it is crucial for the hosts to reply fast to make the listing popular.

Exploring extension: We also applied the model related to price to give predictions. Comparing the predicted price with the actual price to identify anything weird, to address the extension problem of identifying counterfeit listings. It works, the result shows the listings with price underestimated by the model. for example, those with difference in logprice greater than 2, are mostly unbookable for the whole coming year by checking the website or contact host. By looking into the patterns of these listings, we found that they share some common characteristics. For instance, high number of minimum nights, 0 availability for the next 365 days and distant last review date. We can look into these traits to identify fake listings from genuine ones.

5 Discussion and conclusions

Summary of findings

- Generally, the physical characteristics of a listing and the services provided by the host are the main factors that make a good Airbnb listing. Listings with more amenities tend to be more popular, as do those with responsive and accepting hosts.
- The dataset we used to build our models is limited in terms of the information available, which makes it challenging to obtain accurate predictions. Our review model performed poorly with an R^2 value of 0.22, while our price model did better with an R^2 value of 0.49.
- We created a map that displays the average price for each neighborhood, and we observed that The Calders is the most expensive neighborhood. However, we suspect that the high prices in this neighborhood are due to the presence of fake listings. This issue affects several neighborhoods where prices appear to be inflated despite the relatively small number of listings.
- Based on our analysis of reviews per month by neighborhood, we found that East Craigs North, Cortosphire North, Queensferry East, Old Town, Princes Street, and Leith Street are some of the most popular neighborhoods, in that order. We consider these results to be valid.
- We have spotted some characteristics that can be used to identify the authenticity of the listing. Such as 0 availability or unfairly high prices.

Evaluation of own work

Strengths

- Prior to applying any models, the data was cleaned diligently to ensure the highest quality dataset possible. By maximizing the amount of information retained after data cleaning, our model was able to extract the most valuable insights from the dataset.

- During our analysis, we detected the existence of multicollinearity and addressed it through the application of regularisation techniques.

- Upon detecting significant skewness in the price feature (`skew_value` = 6.97), we applied a log transformation to it. This had a significant impact on model fitting, as evidenced by the doubling of the R^2 score in the price model after transformation.

Limitations

- After conducting a thorough analysis of the data, we have found that using the `number_of_reviews` as the sole indicator of a listing's popularity may not always be reliable.

- Automating the methodology developed for detecting counterfeit listings is not a viable option. And, manual verification is not practical given the large size of the dataset.

- Due to the lack of analysis on image data extracted from `picture_url`, and text-based features such as `amenities` and `description`, this study was unable to conduct a comprehensive analysis of the listings dataset.

Comparison with any other related work

- The scope of our study overlaps with this paper [2]. However, our study differs in that we examine a different geographic region and utilize different data sources and analytical methods.

- We take use of the idea of log transformation in the study, since it effectively transformed skewed data to be more centered.

- Our findings are consistent with those reported in that paper in that the physical characteristics of the listing and the services offered by the host are important factors in pricing decisions.

- However, we also identify other factors that were not reported in their study, such as the presence of fake listings in certain neighborhoods.

- Additionally, our study identifies the issue of fake listings and their impact on pricing decisions, which was not addressed in their study.

Improvements and extensions

- By applying the methodology outlined in the extension problem section to our data preprocessing, we can effectively minimize the noise in the dataset. This should result in improved performance of our model and more accurate representation of the actual price levels on the graph of price against different neighborhoods.

- We did not take use of reviews dataset and images from picture_url feature from listings dataset. There are available tools to analyse them, such as pre-trained CNN models for the image process and LLM models for analysing reviews. Such exploration may provide us different insights

References

[1] Daniel Guttentag et al. “Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study”. In: *Journal of Travel Research* 57 (3 2018). ISSN: 15526763. DOI: `10.1177/0047287517696980`.

[2] Daniel A. Guttentag, Sarah L. J. Smith, and Luke R. Potwarka. “The Price is Right? An Examination of the Factors That Drive Pricing Decisions on Airbnb”. In: *Journal of Travel Research* 56.3 (2017), pp. 295–308. DOI: `10.1177/0047287516659386`.

[3] Junfeng Jiao and Shunhua Bai. “An empirical analysis of Airbnb listings in forty American cities”. In: *Cities* 99 (2020). ISSN: 02642751. DOI: `10.1016/j.cities.2020.102618`.

[4] Karen Xie and Zhenxing Mao. “The impacts of quality and quantity attributes of Airbnb hosts on listing performance”. In: *International Journal of Contemporary Hospitality Management* 29 (9 2017). ISSN: 09596119. DOI: `10.1108/IJCHM-07-2016-0345`.