

**CMSE11432: Principles of Data Analytics 22/23**

**FORM FOR THE SUBMISSION OF ASSESSED GROUP COURSEWORK**

**Title :** **Understanding Prices in Edinburgh’s Airbnb Market**

**Team :** **Group 6**

**Word Count :** **1,496**

We understand that all marks are provisional until ratified by the Faculty Examination Board.

**CMSE11432 Principles of Data Analytics**

**Team Contract**

**Team: Group 6**

**How we will work together**

* Every team member has the right to speak their mind and respect every opinion.
* We will create a cooperative environment for the whole group.
* We would balance the responsibility and help each other in all ways.

**How we will resolve conflicts**

* All decisions will be taken by vote, and every person’s opinion will be considered in the highest possible manner.
* Anyone’s particular circumstances will be taken into consideration as long as they have been previously communicated transparently.
* No blame game to be played; it is on everyone’s shoulders to hold the good and bad.

**Contract agreed by all team members**

In addition to the principles, policies, and processes above, we, the undersigned, also:

1. recognise that individual team members *may* receive different marks from each other based on their level of contribution to the coursework.

2. will notify CO **immediately** if a team member is not contributing as needed or if there are significant unresolved conflicts within the team that threaten our project.

3. will document any cases of significant conflict and its resolution, clearly in writing for the benefit of transparency and as a record for team members and CO.

4. recognise that CO will not intervene to resolve conflict unless we notify them at least two weeks before the submission deadline if we wish them to intervene. This encourages us to raise issues early. CO will of course act in the event of any problems under the university’s code of conduct or well-being policies.

**Signed Agreement**

| **Team Member Name** | **Date Signed & Agreed** |
| --- | --- |
| Ketan Mittal | Ketan Mittal - 23 September 2022 |
| Javiera Mora | Jay Mora - 23 September 2022 |
| Markus Mowatt-Larssen | Markus Mowatt-Larssen - 23 September 2022 |
| Ranty Putri | Ranty Putri - 23 September 2022 |
| Xitao Qiao | Xitao Qiao - 23 September 2022 |

**CMSE11432 Principles of Data Analytics**

Business Analysis Report: Understanding Prices in Edinburgh’s Airbnb Market

**Team: 6**

# Background, Significance, and Research Questions

According to YouGov’s research in Q3 2022, Edinburgh is the #1 popular city in the UK and its tourism industry held a value of 1.87 billion pounds in 2019 (City of Edinburgh Council, 2021. Page 16). In response to the ever-growing demand for affordable accommodation, Airbnb’s presence has consolidated a relevant portion of the market, with the listings increasing by 54% from July 2016 to September 2017 (Rae, 2017. Page 1).

The Airbnb market has changed significantly after the pandemic. Literature states that “the proportion of Capitalist[[1]](#footnote-0) hosts will decline, as they will tend to move to the long-term rentals market, and the proportion of Befriender[[2]](#footnote-1) and Ethicist[[3]](#footnote-2) hosts will increase, moving Airbnb back towards its original ethos of space sharing among ordinary citizens”. (Dolnicar and Zare, 2020. Page 2)

This report aims to discover the property attributes' role in setting the listing price, which would translate both as the value clients assign to each attribute; and the pricing decision from the investor's point of view.

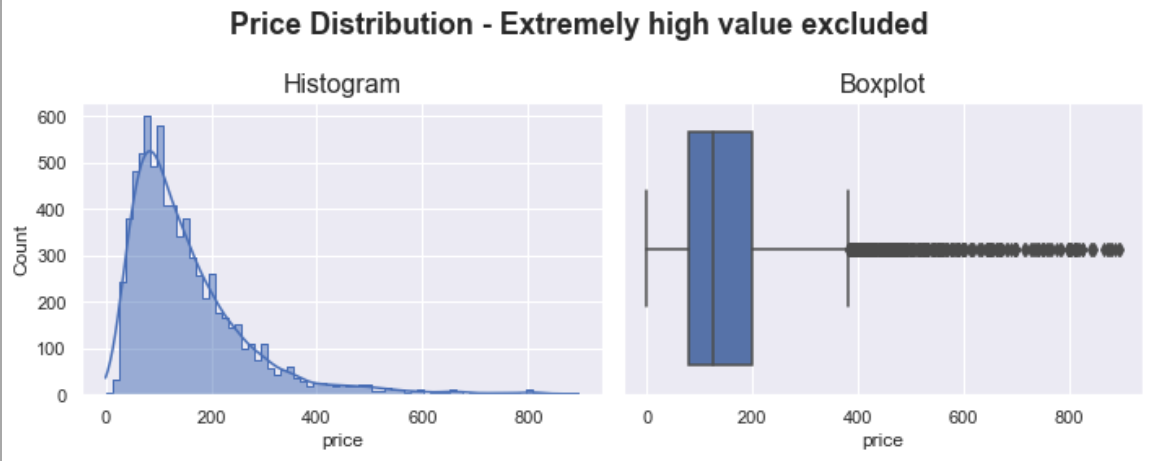
# Data Presentation

This analysis uses a real Edinburgh Airbnb dataset generated from the Inside Airbnb[[4]](#footnote-3) website to evaluate the factors involved in the pricing of an Airbnb property. The dataset consists of three files: calendar, listing, and reviews[[5]](#footnote-4). Several variables used in this analysis are prices, date, minimum nights stayed, maximum nights stayed, last reviewed, average rating, and location.

Table 1. Descriptive Analysis from The “Calendar” Data

|  | **price** | **adjusted\_price** | **minimum\_nights** | **maximum\_nights** | **year** | **month** | **day** | **dayOfWeek** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 316 | 2853576 | 2853576 | 2853576 | 2853576 | 2853576 | 2853576 | 2853576 |
| **mean** | 143.3573608 | 203.9507103 | 5.027388442 | 696.4944666 | 2022.699267 | 6.526032599 | 15.72054222 | 2.995154501 |
| **std** | 103.5754129 | 435.2417156 | 27.18914708 | 492.1377317 | 0.45857694 | 3.447850168 | 8.796240182 | 1.999515812 |
| **min** | 20 | 0 | 1 | 1 | 2022 | 1 | 1 | 0 |
| **25%** | 75 | 77 | 2 | 60 | 2022 | 4 | 8 | 1 |
| **50%** | 119 | 120 | 2 | 1125 | 2023 | 7 | 16 | 3 |
| **75%** | 176.0445205 | 200 | 3 | 1125 | 2023 | 10 | 23 | 5 |
| **max** | 813.3534247 | 20551 | 1000 | 1125 | 2023 | 12 | 31 | 6 |

After preprocessing the data, it could be inferred that 75% of the prices lie under the price of $176 and the median, or percentile 50, roughly comes to $119 per night. The data scope is from September 2022 to September 2023.

Figure 1. Price Distribution (Extremely High Values Excluded)

# Methodology

In descriptive analytics, one of the main concepts to understand is a distribution. They describe the frequency or probability of the occurrence of a certain outcome for a given variable. Distributions are best represented in visualisations such as histograms and boxplots. These display the shapes of a distribution and allow insight such as if the distribution is normal. Other key terms include the mean, or average value of a dataset, and the variance, which describes the typical distance between data points and the mean. Finally, the concept of independence is important as it denotes that a variable is not in correlation with other variables. In the case of the dependent variable, this analysis seeks to understand the nature of its dependence on the independent variables.

Various methods of data preprocessing and analysis are used to gain insight on this dataset. The first step is to preprocess the data to address factors that would otherwise ruin the outcome of the analysis. One vital method is to remove certain rows that contained null values which could not be imputed in any way because of their independence from other events. This was the method used to clean the Calendar data because the date could not be averaged or imputed to provide an improvised datum. For other columns, such as ratings of a host, events which had ratings could be averaged for certain hosts to fill in events with missing ratings for that host. These methods improved the cleanliness of the overall data by removing any datum which showed up as null.

The next method of preprocessing was to remove outliers which would skew the results of the analysis. Outliers are values which are considered extreme in comparison to other values in a distribution. In the case of this data, the Airbnb price for certain listings far exceeded the median values and would therefore compromise the results of a linear regression. By viewing histograms and boxplots, it was determined that the top 3% of listings by price could be removed. After completing these steps, the data was fully processed and ready for analysis.

The initial method of analysis was to develop descriptive summaries and visualisations of the data. Python has many built-in functions which enable descriptive analysis to be completed simply. Histograms were created to determine the shape of the distributions for different variables. Trends were also determined from the descriptive analysis such as the variation in price across different times of the year or which locations have the highest pricing.

The final method of analysis was to create linear regressions of the data. Regressions are a method of modelling which seek to form a mathematical relationship between the independent variables and a dependent variable. Linear regressions are those which produce models that are strictly linear in nature (Newbold, 2013. Page 489). In the case of this data, the dependent variable was the price which was modelled by creating a regression with the independent variables such as average rating and location. A multiple linear regression was created which utilised all the independent variables as predictors of the price.

Prior to the regression, it is important to confirm that there is only mild autocorrelation between the variables using the Durbin-Watson statistic. When significant autocorrelation is present, regression results are considered to be without merit because it implies that a variable is correlated with itself rather than other variables. Using Python functions, this can be calculated easily and compared with a benchmark value of 2, which indicates no autocorrelation in the data.

Once the linear regression was computed, the root mean squared error, or RMSE, was calculated by taking the root of the sum of the squared errors. Errors in statistics refer to the difference in value between a model and observed value. The RMSE is key for determining the quality of the model in relation to the observed price points (Newbold, 2013. Page 491). The R-squared value was also calculated. Known as the correlation of determination, this value gives us the strength of the correlation between the dependent and independent variables. Since the regression uses multiple variables, an adjusted R-squared equation is necessary.

# Analysis and Discussion

## 4.1 Understanding Airbnb Properties in Edinburgh

This analysis starts with understanding the Airbnb properties across Edinburgh. Over a 12 month horizon, the summer and spring seasons are the most expensive while November is the cheapest of all. In a future report, the effects of weather on these prices could easily be explored.

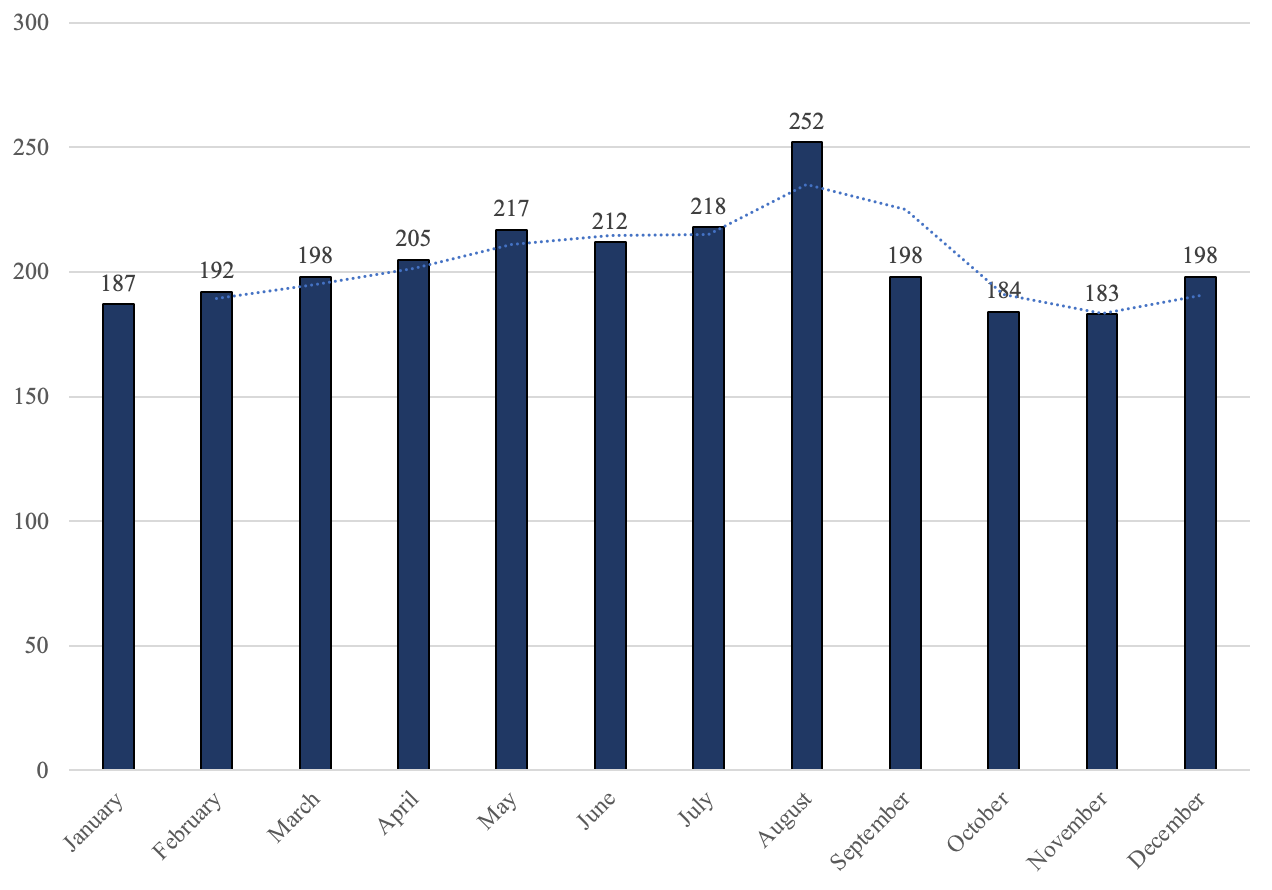


Figure 2. Average Price per Month for Airbnb Properties in Edinburgh

In addition, it can also be seen from Figure 3 below that the Airbnb business in Edinburgh has excellent potential as the booking rate is above 50% for Edinburgh properties through September 2023 with the popular summer tourist season reaching even higher.

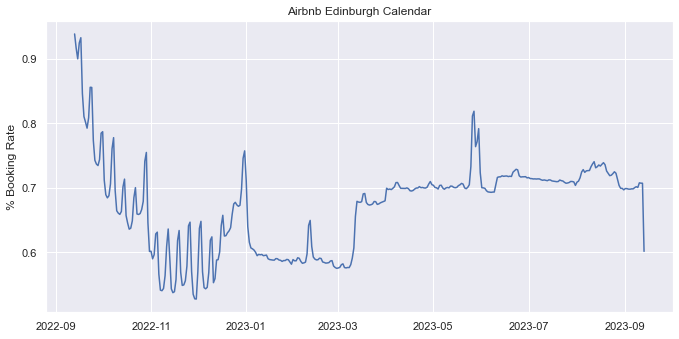


Figure 3. Booking Rate of Airbnb Properties in Edinburgh, September 2022 - September 2023

## 4.2 Factors Affecting Listing Price

According to the multiple linear regression results, there are significant relationships between several attributes and Airbnb property prices. Prior to modelling the regression, it was necessary to measure autocorrelation using the Durbin-Watson test statistic. The score is 1.802, which shows a slight positive autocorrelation but is still within the acceptable range. The regression analysis found a 0.34% increase in price for every 1% increase in the number of accommodates (guests) the property can hold and a 0.13% increase in price for every 1% increase in the number of houses rented/listed by hosts. To validate the results, it was calculated that the adjusted R2 value is approximately 0.45, meaning the model is a moderate fit but does not fully capture the reality. The RMSE is around 0.7 which is within an acceptable range.

Using another approach, looking at the Airbnb Properties heatmap (see Figure 4 below), it can be inferred that most Airbnb properties in Edinburgh are located in the city centre and the northeastern part of the city. The heatmap also indicates that prices are higher in the city centre, indicating that location plays a key role in determining Airbnb property prices.

# 

Figure 4. Airbnb Properties Heatmap in Edinburgh

# Conclusion

Results obtained on this analysis can prove useful for Airbnb property investors and their decision making. Firstly, the relationship between the number of guests and the price is strong enough to suggest that if property investors are able to increase the number of beds in a property, they can theoretically increase the price for the same space and still be competitive with other properties with similar numbers of guests allowed.

Secondly, the correlation between number of properties owned and price suggests that investors who seek out many properties for Airbnb listings and maintain them are able to list at a greater price. While the exact cause of this correlation is not to be explored in this report, it is surmisable that these investors have more money to spend on their properties and are successful in delivering high quality experiences for customers because of this.

Finally, the heat map in Figure 4 is a strong visualisation of the value of location. Investing in a property close to the city centre and the attractions that Edinburgh has to offer clearly translates to more returns on your investment, and it is greatly valued by tourists.

# Limitations

One of the limitations of the analysis is that in the neighbourhood data, 29% of null values were found, which limited the project's scope in exploring the insights based on different areas in the city. Another limitation is that the data might indicate that there is strong multicollinearity, meaning that the several independent variables in this model might be correlated, potentially affecting regression model’s efficiency.

# 

# References

**City of Edinburgh Council**. 2021. *”Edinburgh by numbers 2021”. Page 16.* [*https://www.edinburgh.gov.uk/downloads/file/30669/edinburgh-by-numbers-2021*](https://www.edinburgh.gov.uk/downloads/file/30669/edinburgh-by-numbers-2021)*.* accessed on October 23. 2022.

**Dolnicar. S. and Zare. S.**. 2020. *”COVID19 and Airbnb – Disrupting the Disruptor”. Page 2.* [*https://doi.org/10.1016/j.annals.2020.102961*](https://doi.org/10.1016/j.annals.2020.102961)

**Inside Airbnb**. 2022. *”Edinburgh”. CSV databases: “Calendar”, “Listings”, and “Reviews”* accessed on October 11, 2022*.* [*http://insideairbnb.com/edinburgh/*](http://insideairbnb.com/edinburgh/)

**Newbold, P., Carlson, W.L. and Thorne, B.M.,**2013. *Statistics for Business and Economics. Harlow, England: Pearson.*

**Rae. A.** 2017. *”Analysis of Short-Term Lets Data for Edinburgh”. Page 1.* [*https://greens.scot/sites/default/files/AnalysisShortTermLetsDataforEdinburgh\_AlasdairRae\_PDF-FINAL\_0.pdf*](https://greens.scot/sites/default/files/AnalysisShortTermLetsDataforEdinburgh_AlasdairRae_PDF-FINAL_0.pdf)*.* accessed on October 23. 2022.

**YouGov.co.uk**. 2022. *”The Most Popular Cities (Q3 2022)”.* [*https://yougov.co.uk/ratings/travel/popularity/cities/all*](https://yougov.co.uk/ratings/travel/popularity/cities/all)*.* accessed on October 23. 2022.

# 

# 

# 

# 

# Appendix

## Appendix A: Data Structures of Edinburgh Airbnb Data

| **Data File Name** | **Column** | **#Rows** |
| --- | --- | --- |
| Calendar | listing\_id, date, available, price, adjusted\_price, minimum\_nights, maximum\_nights | 986,891 |
| Listing | id, name, host\_id, host\_name, neighbourhood\_group, neighbourhood, latitude, longitude, room\_type, price, minimum\_nights, number\_of\_reviews, last\_review, reviews\_per\_month, calculated\_host\_listings\_count, availability\_365, number\_of\_reviews\_ltm, license | 7,833 |
| Reviews | listing\_id, date | 475,798 |

## 

## 

## Appendix B: OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.454

Model: OLS Adj. R-squared: 0.452

Method: Least Squares F-statistic: 200.2

Date: Mon, 24 Oct 2022 Prob (F-statistic): 0.00

Time: 18:40:53 Log-Likelihood: -7560.4

No. Observations: 6772 AIC: 1.518e+04

Df Residuals: 6743 BIC: 1.538e+04

Df Model: 28

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 2.4e-15 0.009 2.67e-13 1.000 -0.018 0.018

x1 0.0738 0.012 6.090 0.000 0.050 0.098

x2 -3.906e-16 1.72e-16 -2.271 0.023 -7.28e-16 -5.34e-17

x3 0.0044 0.010 0.425 0.671 -0.016 0.025

x4 0.0184 0.009 2.021 0.043 0.001 0.036

x5 -0.0023 0.009 -0.249 0.803 -0.020 0.016

x6 0.3351 0.011 31.375 0.000 0.314 0.356

x7 0.0310 0.037 0.844 0.399 -0.041 0.103

x8 -0.0253 0.011 -2.283 0.022 -0.047 -0.004

x9 -0.0305 0.035 -0.859 0.390 -0.100 0.039

x10 -0.0384 0.038 -1.002 0.317 -0.114 0.037

x11 -0.0346 0.033 -1.057 0.290 -0.099 0.030

x12 -0.1463 0.039 -3.780 0.000 -0.222 -0.070

x13 -0.0233 0.037 -0.629 0.529 -0.096 0.049

x14 0.1770 0.055 3.241 0.001 0.070 0.284

x15 0.1662 0.025 6.552 0.000 0.116 0.216

x16 0.0144 0.055 0.265 0.791 -0.092 0.121

x17 -0.0285 0.042 -0.677 0.499 -0.111 0.054

x18 0.0914 0.013 7.267 0.000 0.067 0.116

x19 -0.0978 0.014 -7.170 0.000 -0.125 -0.071

x20 0.0653 0.016 3.978 0.000 0.033 0.097

x21 -0.0878 0.014 -6.261 0.000 -0.115 -0.060

x22 0.1346 0.344 0.392 0.695 -0.539 0.809

x23 -0.0577 0.324 -0.178 0.859 -0.693 0.578

x24 0.0383 0.052 0.734 0.463 -0.064 0.140

x25 0.0223 0.018 1.241 0.215 -0.013 0.057

x26 -0.0099 0.015 -0.671 0.502 -0.039 0.019

x27 -0.0547 0.016 -3.409 0.001 -0.086 -0.023

x28 -0.3548 0.012 -28.879 0.000 -0.379 -0.331

x29 -0.1483 0.016 -9.503 0.000 -0.179 -0.118

==============================================================================

Omnibus: 360.177 Durbin-Watson: 1.802

Prob(Omnibus): 0.000 Jarque-Bera (JB): 421.269

Skew: 0.581 Prob(JB): 3.33e-92

Kurtosis: 3.375 Cond. No. 1.32e+16

==============================================================================

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 1.47e-28. This might indicate that there are

strong multicollinearity problems or that the design matrix is singular.

## 

## 

## Appendix C: Explanation of Variables

|  | **variable name** | **note** |
| --- | --- | --- |
| x1 | id | id of the house |
| x2 | scrape\_id | the id of scraped information |
| x3 | host\_id |  |
| x4 | latitude |  |
| x5 | longitude |  |
| x6 | accommodates | the number of accommodates it can hold |
| x7 | minimum\_nights |  |
| x8 | maximum\_nights |  |
| x9 | minimum\_minimum\_nights | the minimum of minimum\_nights |
| x10 | maximum\_minimum\_nights |  |
| x11 | minimum\_maximum\_nights |  |
| x12 | maximum\_maximum\_nights |  |
| x13 | minimum\_nights\_avg\_ntm | the average of minimum\_nights |
| x14 | maximum\_nights\_avg\_ntm |  |
| x15 | availability\_30 | availiability in 30 days |
| x16 | availability\_60 |  |
| x17 | availability\_90 |  |
| x18 | availability\_365 |  |
| x19 | number\_of\_reviews |  |
| x20 | number\_of\_reviews\_ltm | no. of reviews over the last twelve months |
| x21 | number\_of\_reviews\_l30d | no. of reviews over the last 30 days |
| x22 | calculated\_host\_listings\_count | no. of houses rented/listed by hosts |
| x23 | calculated\_host\_listings\_count\_entire\_homes |  |
| x24 | calculated\_host\_listings\_count\_private\_rooms |  |
| x25 | calculated\_host\_listings\_count\_shared\_rooms |  |
| x26 | reviews\_per\_month |  |
| x27 | rt\_Hotel room | room type: hotel room |
| x28 | rt\_Private room |  |
| x29 | rt\_Shared room |  |

1. Capitalist hosts engage in short-term renting to make commercial profits [↑](#footnote-ref-0)
2. Befrienders enjoy the social aspect of hosting [↑](#footnote-ref-1)
3. Ethicist are the true believers in the principle of sharing [↑](#footnote-ref-2)
4. http://insideairbnb.com/ [↑](#footnote-ref-3)
5. For the complete data structure. see [Appendix A](#_og6qskxqip55) [↑](#footnote-ref-4)