QEA Project 1: Smart Airbnb Price Predictor

Lilo Heinrich
Jackie Zeng

February 2020

1 Summary

This report explores the application of linear regression to Airbnb listings data, specifically to predict the price per night of a property listed. Our algorithm models Airbnb price per night as a linear combination of other relevanthttps://www.overleaf.com/project/5e57cf62d5061800019a1527 factors from each listing. The potential for harm of this algorithm is that it may be able to predict which properties are most profitable on Airbnb, allowing people to take advantage of the system and maximize profit by buying up and renting these types of properties. We have found that the location of a property is very significant in determining its price and that the highest ranked cities are Hong Kong, Compenhagen, and Malibu. The greater implication of our findings is that Airbnb may impact the housing market or hotel industry in these high-demand locations. We can take this project further by focusing solely on location data, comparing whether Airbnb rental, long-term rental, or hotel rates are most profitable in different locations, and by how much.

2 Introduction

2.1 Linear Regression

We are using linear regression, a linear approach to modeling the relationship between a dependent variable and one or more independent variables. It is a commonly used type of predictive analysis, used for example by Business Analysts to generate insights on consumer behaviour, understand business and factors influencing profitability, or optimize pricing and promotions on sales of a product.

Linear regression is based on five assumptions: [2]

1) The independent and dependent variables have a linear relationship
2) The data is normally distributed
3) There none/little collinearity between the independent variables
4) No auto-correlation- values cannot be predicted based on preceding values in the series
5) Homoscedasticity- error term is the same across all values of the independent variables

This model can also be misapplied in several ways, such as by relating unrelated variables, leading to spurious correlations, misinterpreting the correlation between variables to always mean causation, or by applying this model to data which does not meet its' assumptions as described above well enough to produce a meaningful result.

2.2 Airbnb

We have chosen to analyze the Airbnb rental market. Airbnb is an online marketplace connecting people who want to rent out their homes with people who are looking for accommodations. It acts as a broker, receiving commissions from each booking. The company has grown since 2008 to serve over 150 million users and host more than five million property listings in over 191 countries. [1]

2.3 Algorithm

We attempted to linearly relate attributes of Airbnb listings to their price in order to understand what factors make a property valuable.

In order to apply linear regression to our data, we first cleaned up our data by removing listings with less than 100 reviews or with incomplete information and narrowed our focus to 20 columns of data including both numerical and categorical data. Next, we used one hot encoding to numerically quantify the presence of different attributes in each of the categorical data columns.

We separated out the price data into one vector and the rest of the data into a matrix and ran linear regression to relate the two, taking linear trends through multidimensional space to figure out which result in the closest approximation of the dependent variable. From this, we made a predicted price vector and compared it to the original using correlation as a metric of how accurate our liner model is, and plotted it on a graph to see the distribution and identify outliers. Lastly, we interpreted the meaning of the attributes weighted the highest, which we found to be either big cities, high-tourism locations, or both.

2.4 Ethical Implications

The potential for harm of our algorithm is that, if given updated data through time, it may be able to detect which locations/cities are most valuable on Airbnb and allow the user to make strategic decisions in buying and listing properties, as well as predicting what a good price is to list a property in that area. This type of predictive algorithm, although simple, could allow landowners and businesspeople to take advantage of the system to maximize their profits due to the the minimal regulations that Airbnb imposes.

2.5 Questions

Question: What factors most affect the price of Airbnb properties?
Sub-question: How do we interpret the results of applying linear regression to a dataset?

3 Methods

Our goal is to determine the listing attributes that influence Airbnb pricing the most. To do this, we used linear regression to predict pricing. We hope to determine the importance of factors such as location, and trustworthiness of host.

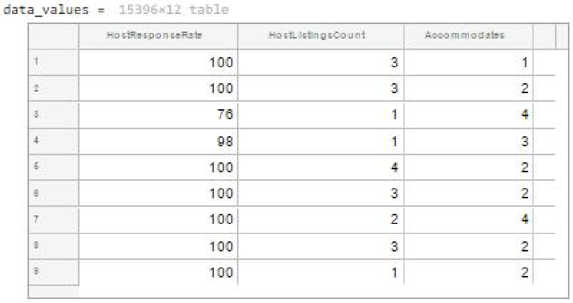
3.1 Data Set

We got our data set from the website OpenDataSoft, which had compiled all of the AirBnb listing data from InsideAirbnb, an independent, non-commercial set of tools and data intended to allow one to explore how Airbnb is really being used in cities around the world.

3.2 Data Processing

The Airbnb data set we chose to work with came in the format of a 494955 x 89 table, with information on 494955 listings all over the world and 89 associated attributes. The first steps we took was to remove columns that were difficult to parse or less relevant to our question and listings that had less than 100 reviews. We assume it to be valid to focus on only the higher-reviewed properties because they have more accurate review data and have presumably received more visitors, making the data more reliable. This process dramatically decreased the data set we were working with to 15396 x 20.

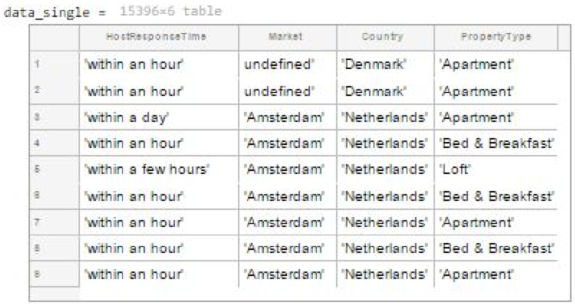
The remaining attributes of listings were then separated into three types: **numerical value variables** (HostResponseRate, HostListingsCount, Accommodates, Bathrooms, Bedrooms, Beds, Price, MinimumNights, MaximumNights, NumberofReviews), **single categorical variables** (HostResponseTime, PropertyType, Country, Market, RoomType, CancellationPolicy), and **multi categorical variables** (Amenities, Features).



data_values = 15396×12 table

	HostResponseRate	HostListingsCount	Accommodates
1	100	3	1
2	100	3	2
3	76	1	4
4	98	1	3
5	100	4	2
6	100	3	2
7	100	2	4
8	100	3	2
9	100	1	2

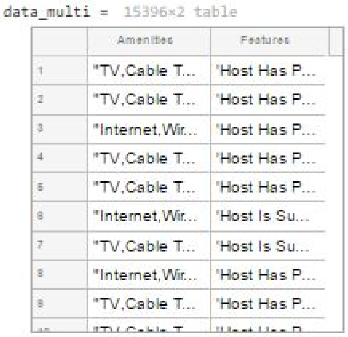
(a) numerical value variables



data_single = 15396×6 table

	HostResponseTime	Market	Country	PropertyType
1	'within an hour'	undefined'	'Denmark'	'Apartment'
2	'within an hour'	undefined'	'Denmark'	'Apartment'
3	'within a day'	'Amsterdam'	'Netherlands'	'Apartment'
4	'within an hour'	'Amsterdam'	'Netherlands'	'Bed & Breakfast'
5	'within a few hours'	'Amsterdam'	'Netherlands'	'Loft'
6	'within an hour'	'Amsterdam'	'Netherlands'	'Bed & Breakfast'
7	'within an hour'	'Amsterdam'	'Netherlands'	'Apartment'
8	'within an hour'	'Amsterdam'	'Netherlands'	'Bed & Breakfast'
9	'within an hour'	'Amsterdam'	'Netherlands'	'Apartment'

(b) single categorical variable



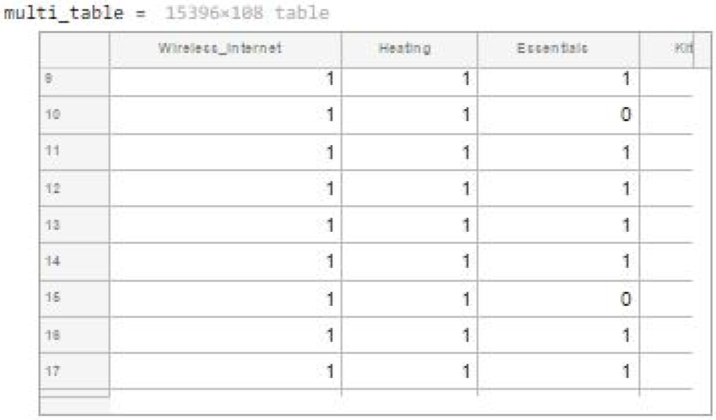
data_multi = 15396×2 table

	Amenities	Features
1	"TV,Cable T...	'Host Has P...
2	"TV,Cable T...	'Host Has P...
3	"Internet,Wir...	'Host Has P...
4	"TV,Cable T...	'Host Has P...
5	"TV,Cable T...	'Host Has P...
6	"Internet,Wir...	'Host Is Su...
7	"TV,Cable T...	'Host Is Su...
8	"Internet,Wir...	'Host Has P...
9	"TV,Cable T...	'Host Has P...

(c) multi categorical variable

Figure 1

To process this data using linear regression, we needed to convert the categorical variables into numerical values that can be stored in a matrix. To do this, we used one hot encoding to parse the top n attributes in a given categorical column into n columns. If the attribute existed in a listing, it was given a value of 1, else, 0. To account for varying numbers of attributes per column, we added a check that will round down to the maximum number of attributes if n is greater.



```
multi_table =  15396×188 table
```

	Wireless_Internet	Heating	Essentials	Kit
9	1	1	1	
10	1	1	0	
11	1	1	1	
12	1	1	1	
13	1	1	1	
14	1	1	1	
15	1	1	0	
16	1	1	1	
17	1	1	1	

Figure 2: Example of one hot encoding of multi-categorical variables

To match the range of the one hot encoded categorical variables, we scaled the numerical value variables to a range of 0 - 1. This ensures that the model's coefficients are not differently scaled due to the numerical values' range being greater.

To create the final attributes matrix, we combined the numerical values variables and the one-hot encoded categorical variables using the top 500 components. The final matrix's dimensions are 15396 x 201. There are now more columns because the one hot encoding has parsed categorical attributes into many columns.

3.3 Linear Regression

To determine the relationship between different attributes and pricing of a particular property for one night's stay, we performed linear regression according to

$$Ax = b \tag{1}$$

where A is a matrix with listings as rows and attributes as columns, x is a column vector that determines how each attribute contributes to the price, and b is a column vector of the prices of listings.

Once we determined x, we found the top attributes that contributed to price and predicted the price of listings using our model by multiplying x with the original attribute matrix. We then found the correlation between the actual prices and predicted prices.

4 Detailed Findings

4.1 Top Attributes

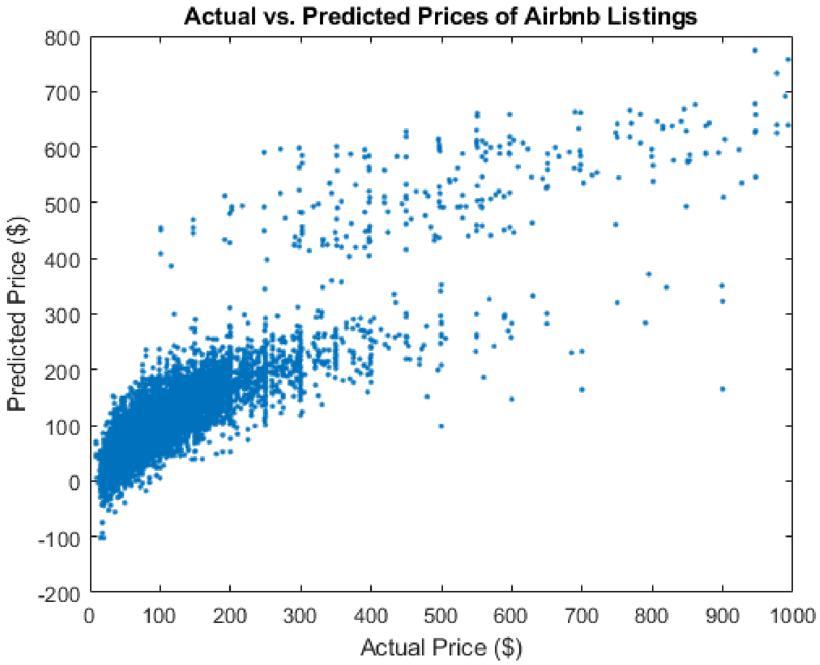
The top 10 attributes that influenced price are

1. H.K	**2. Copenhagen**	**3. Reviews**	**4. Bedrooms**	**5. Accommodates**
0.4514	0.4235	0.2216	0.1878	0.1856
6. Stove	**7. Bathrooms**	**8. Malibu**	**9. Extra Linens**	**10. Venice**
0.0838	0.0825	0.0775	0.0491	0.0448

From this table, we can see that reviews from other Airbnb users and various amenities play a big role in the per-night pricing of Airbnb listings.

However, what we found most interesting is that four out of the top ten attributes that contribute to price are locations - Hong Kong, Copenhagen, Malibu, and Venice. It makes sense that these locations correlate with high Airbnb prices because these are all major cities with a booming tourism industry. Since listings in these major cities will fetch higher short-term Airbnb prices than long-term leases, landowners may be more inclined to rent their properties through Airbnb rather than providing permanent housing for locals. By decreasing the amount of housing available to residents, Airbnb increases housing scarcity.

4.2 Actual Prices vs. Predicted Prices

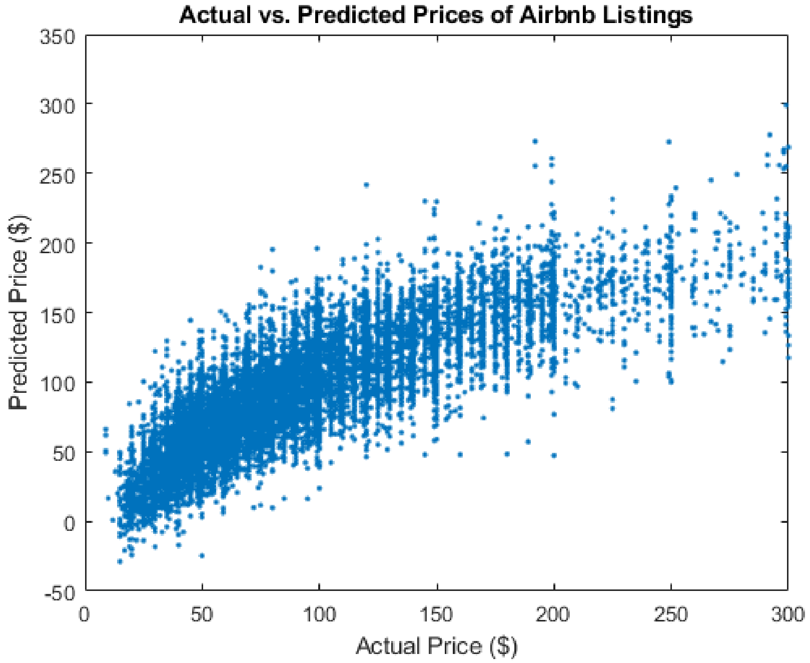


$$Corr = 0.8669 \tag{2}$$

The above graph is a plot of real one-night prices of Airbnbs vs. prices predicted through our algorithm. The correlation coefficient between the actual prices and predicted prices is 0.8362, indicating a strong positive linear relationship and validates our predictions.

The graph shows a cluster in the $0-300 price range. This cluster makes sense because most homes listed for one night are under $300. As the prices get higher, the predictions become inaccurate, seeming to cap out around $300. There is also a trailing group of outliers from around $300-$1000 on the x-axis and $500-$600 on the y-axis, the higher-priced range. We believe this is because we don't have enough data points for those areas - our data isn't normally distributed enough for it to work outside of the range. This means that linear regression may be the wrong model for this data.

4.3 Actual Prices vs. Predicted Prices for $0-300 range



$$Corr = 0.8249 \tag{3}$$

We attempted filtering out these luxury stays and re-running this model for only the $0-300 price range to see how it affects the correlation factor and distribution of the data. We deleted the 390 listings out of 15936 (2.5% of our data) and reran the program, producing the following graph and a 0.8249 correlation factor. This correlation is lower than with the full dataset, but we still think it is more accurate without these outliers. Interestingly, this graph more clearly exhibits a slightly curved shape, indicating that a linear approximation may not be the best model.

5 Recommendations

The key finding of this Airbnb investigation is that location influences pricing of an Airbnb property dramatically. the implications of this are that in popular tourist destinations, landlords can make more money on holiday lets than long term tenants, making it profitable to buy and list multiple properties on AirBnb. There is no limit on how many listings a host can advertise on AirBnb, making this possible. This may exacerbate existing problems with housing scarcity and unaffordability, crowding out locals. For example, New Orleans is already facing this problem. [3] Or, on the flip side, it may take business away from the hotel industry, such as in Hong Kong.

The results obtained from this project show how businesspeople can take advantage of systems such as this to maximize their profit, possibly at the detriment of others, and reflects the current ethical debate around Airbnb.

In terms of technical improvements, in the future, it would be interesting to implement logistic regression instead of linear regression to get a better price prediction. Additionally, it would be interesting to rerun this algorithm with a focus on only location data and compare whether Airbnb rental, long-term rental, or hotel rates are most profitable in different locations, and by how much.

References

[1] Much Needed. Airbnb by the numbers: Usage, demographics, and revenue growth. https://muchneeded.com/airbnb-statistics/.

[2] Statistics Solutions. Assumptions of linear regression. https://www.statisticssolutions.com/assumptions-of-linear-regression/.

[3] The Verge. Activists say airbnb makes new orleans housing shortage worse. https://www.theverge.com/2018/3/28/17172946/airbnb-new-orleans-housing-crisis-gentrification-str.