BOSTON UNIVERSITY

Airbnb Price Analysis Report for neighbourhoods in Boston and Hawaii

Jing Wu

December 12, 2022

This report aims to analyze the factors which may affect the price of houses or rooms in Airbnb, and conduct multilevel analysis towards the price in different neighbourhoods of Hawaii and Boston.

1 Instruction

1.1 Background

Airbnb is a platform which provides the service of renting rooms and houses for travellers. The places available for rent on the platform have various properties, like the number of rooms and room types, they are also located in different neighbourhoods of the city.

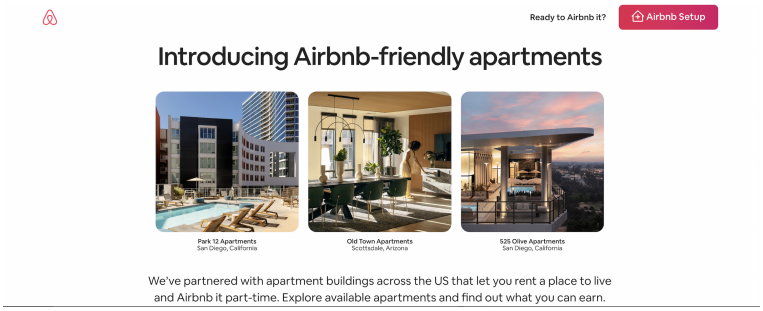

Figure 1.1: Airbnb

1.2 Data resource

The Airbnb data and geojson file, which contains the shape file of different neighbourhoods of the cities, is downloaded from the website Inside Airbnb. Moreover, the report also

contains Hawaii's annual rainfall data, which is downloaded from the website Rainfall Atlas of Hawaii.

Airbnb: `http://insideairbnb.com/get-the-data/`

Rainfall Atlas of Hawaii: `http://rainfall.geography.hawaii.edu/downloads.html`

1.3 Technique

1.3.1 Data wrangling

The Boston data in Airbnb contains data from greater Boston, while the shape file only contains Boston(without some neighbourhoods like Brookline and Cambridge). The data used was filtered with the shape file.

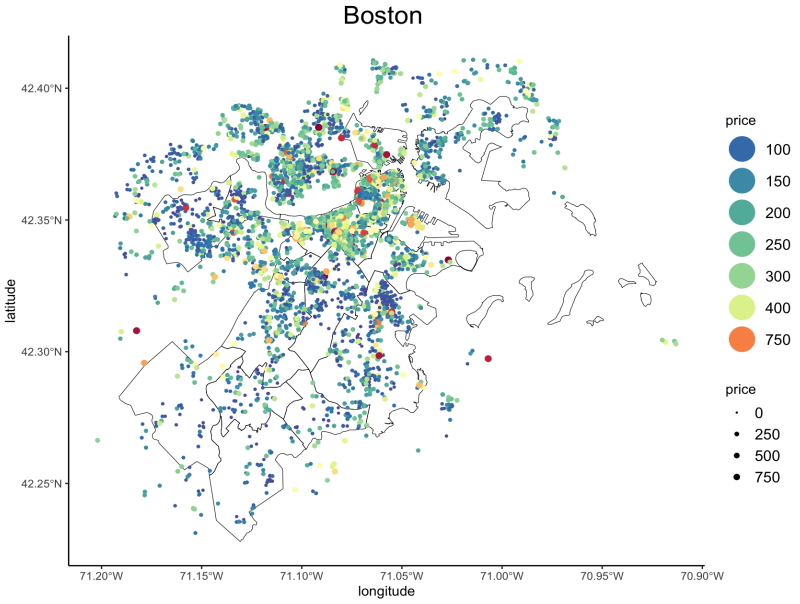

Figure 1.2: Raw Boston data

In addition, the *bathrooms text* in the raw data contains the number of bathrooms and the type of bathroom, and the *bathrooms* is invalid, so the *bathrooms text* variables were cleaned and separated to *bathroom type* and *bathroom number* before analysis.

1.3.2 Data set combining

Hawaii's annual rainfall data is downloaded as the ASCII Grid file. The central idea of matching the latitude and longitude of the price data and grid data is to find the grid where the house is located and take corresponding grid data.

1.4 Method used

The main methods are exploratory data analysis and multilevel analysis.

2 analysis towards Boston

2.1 EDA

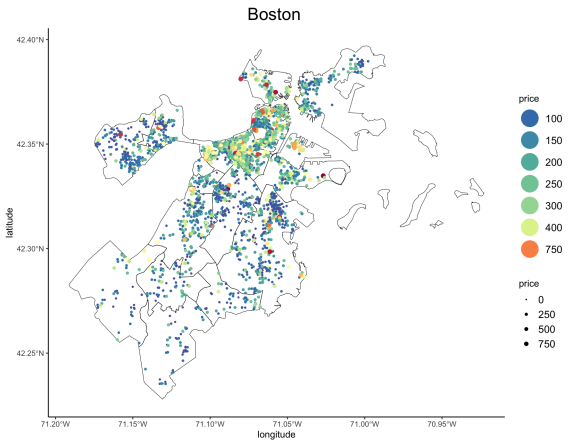

Figure 2.1: Boston Price Data

From the plot, some points are centred in specific neighbourhoods, while others are approximately randomly distributed. In addition, most of the points with high price value tend to be located in specific neighbourhoods.

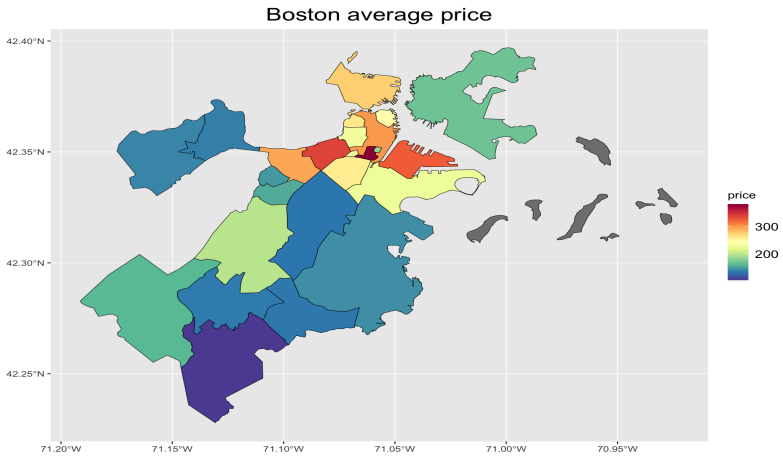

Figure 2.2: Boston Average price Data

It is very clear that the average price in some neighbourhoods(Back Bay, Chinatown) are higher than others.

variables	value
(Intercept)	53.15
room typeHotel room	153.10
room typePrivate room	−63.35
room typeShared room	−20.68
accommodates	17.86
bathroom numbers	64.96
bathroom typeHalf-bath	−90.92
bathroom typeprivate	41.94
bathroom typeshared	−2.92
bedrooms	20.05
beds	8.32

Table 2.1: Coefficient for the fixed estimates

	group	
1	Chinatown	153.87
2	Back Bay	129.19
3	South Boston Waterfront	95.20
4	Downtown	85.06
5	Fenway	74.40
6	West End	43.38
7	Bay Village	39.36
8	North End	27.13
9	South End	27.07
10	Charlestown	11.18

Table 2.2: top 10 values of random estimates

2.2 Multilevel analysis

2.2.1 Multilevel linear model

We construct the multilevel model with the variables(room type, accommodates time, bathroom numbers, bathroom type, bed numbers, and the random intercept with the neighbourhoods). Table 2.1 show us the coefficients of the fitted model.

From the table we can find out the hotel room will have higher price than private room and shared room in average, and 1 bedrooms increase will lead to 20.05 higher price in average. Table 2.2 is about the neighbourhoods with high baseline. In addition, the houses with private bathroom will have higher prices than that with shared bathrooms or half-bathrooms in average.

From the table, we can find Chinatown, Back Bay and South Boston Waterfront is top 3 neighbourhoods with highest baselines in Boston.

In addition, the standard deviation between groups and residual are 69.41 and 104.24. The deviation between group is quite large, for its percentage with the residual is more than 0.65.

2.2.2 Multilevel logistic model

In addition to the factors that affect housing prices, what will affect the probability of high housing prices is also an interesting question. A multilevel logistic model is built to

q	grp	condval
1	Chinatown	4.30
2	Back Bay	2.27
3	Downtown	1.77
4	South Boston Waterfront	1.69
5	Bay Village	1.67
6	Fenway	1.51
7	North End	1.11
8	South End	0.59
9	Beacon Hill	0.50
10	South Boston	0.44

Table 2.3: top 10 values of random estimates

q	fixef.M1.2.
(Intercept)	−6.41
room typeHotel room	5.42
room typePrivate room	0.365
room typeShared room	3.29
accommodates	0.22
bathroom numbers	1.65
bathroom typeHalf-bath	−9.84
bathroom typeprivate	0.43
bathroom typeshared	−2.26
bedrooms	0.11
beds	0.36

Table 2.4: Fixed random effect with the multilevel logistic regression

determine the questions, with the definition of high price as more than 350 dollars per night.

Table 2.3 shows the random effect of the model, and table 2.4 shows the fixed random effect. For the fixed random effect, we can find that 1 increase in beds will lead to 0.36 increase of the log odds of the possibility of the house to be a high-price house.

The result of random random effect table seems close to that of the the linear model, which means that expensive places are more likely to be located in neighbourhoods with higher average prices. The top 3 neighbourhoods are still China town, Back Bay and South Boston.

3 Analysis towards Hawaii

3.1 EDA

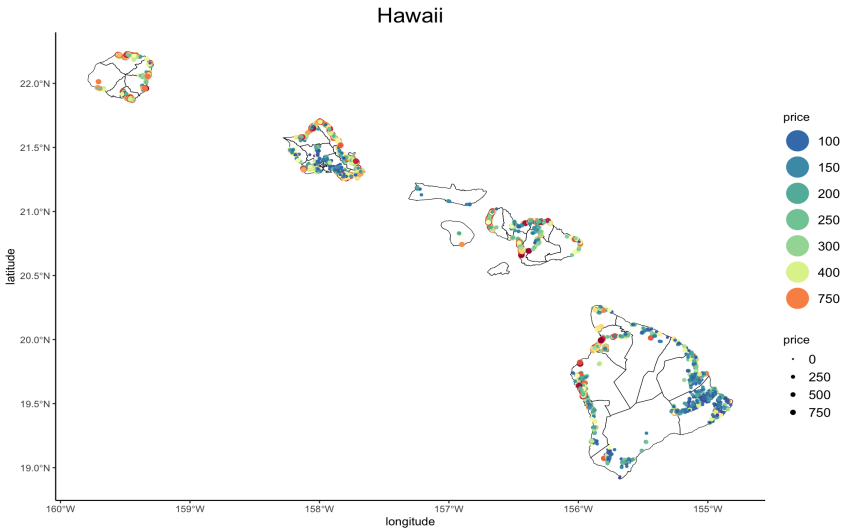

Figure 3.1: Hawaii Price Data

According to figure 3.1, points are distributed in the boundary of the neighbourhoods. It is because of the geographical characteristics of Hawaii. Unlike Boston, Hawaii has many volcanoes, forests, and a much more complex climate.

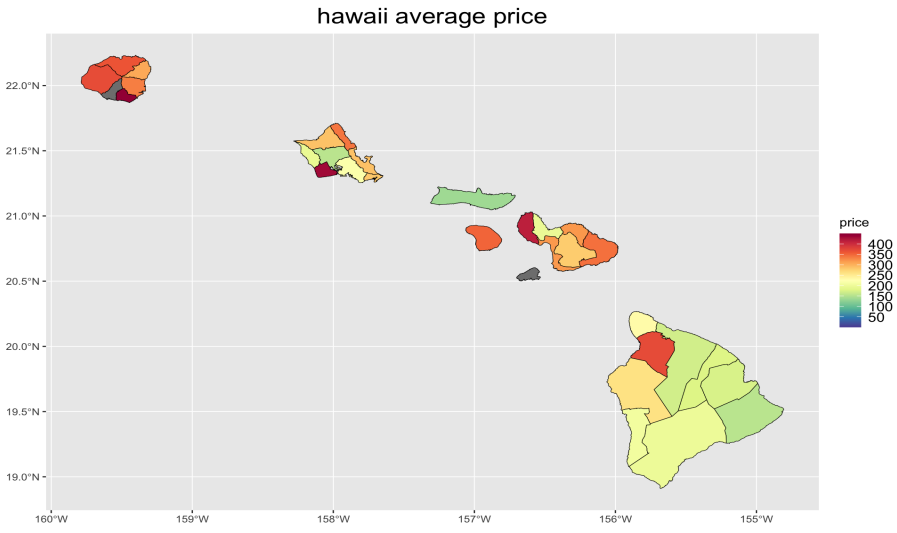

Figure 3.2: Hawaii Average Price Data

Figure 3.2 shows the average price in neighbourhoods of Hawaii. Some neighbourhoods(like Hana and Lahaina) have siginificatnly higher prices than others.

3.2 Multilevel analysis

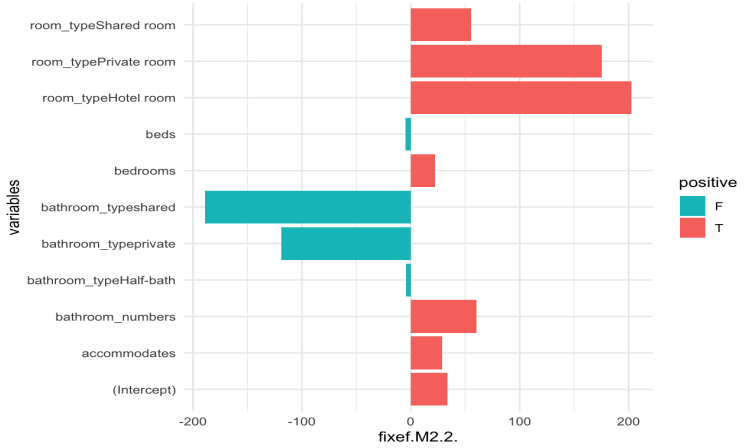

Figure 3.3: Coefficients of Fixed Effect

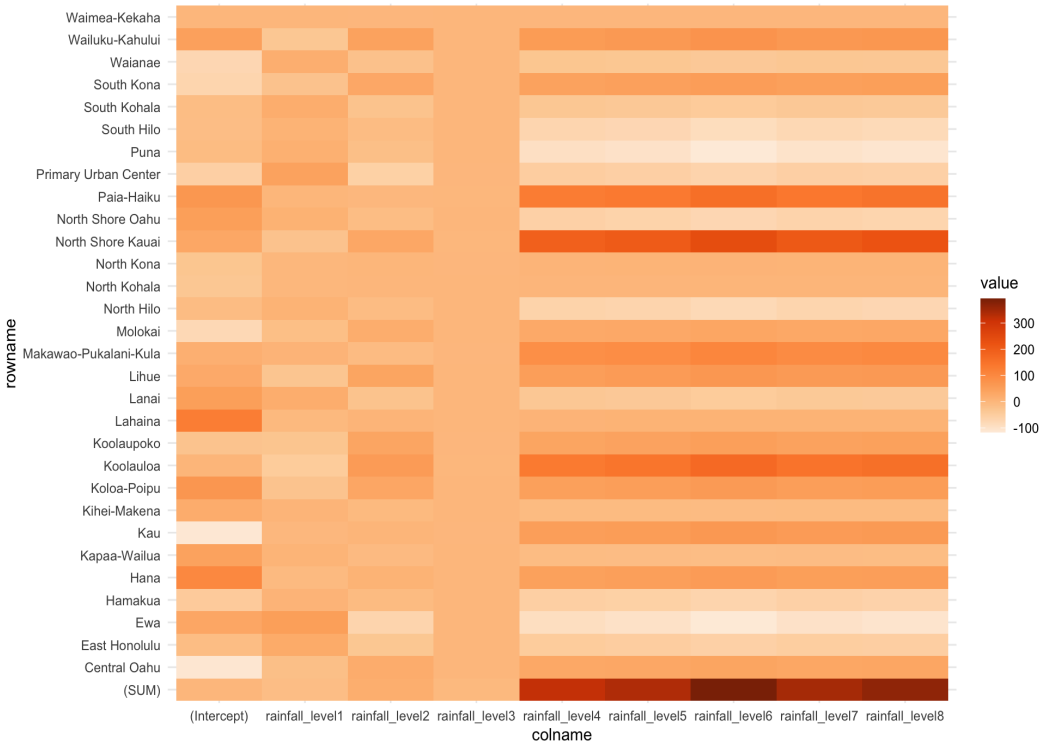

Figure 3.4: Coefficients of Random Effect

To combine the weather data into analysis, we download the annual rainfall data and convert them into eight levels(with the same cutting method of the rainfall amount picture).

We construct the multilevel model with the variables(room type, accommodates time, bathroom numbers, bathroom type, bed numbers, rainfall level with the neighbourhoods, and the random intercept with the neighbourhoods).

Figure 3.3 show us the fixed coefficients of the model. Like Boston, the hotel room still tends to have a higher price than others on average. However, the increasing of beds will lead to decrease of the price in this model in average.

Figure 3.4 show us the random coefficients. In general, houses located in humid locations tend to have higher prices than those located in extremely dry locations. However, in some neighbour hood like East Honolulu and Hana, the houses in dry locations with same conditions in other variables will have highest prices in average.

4 Compare the data in Boston and Hawaii

4.1 EDA

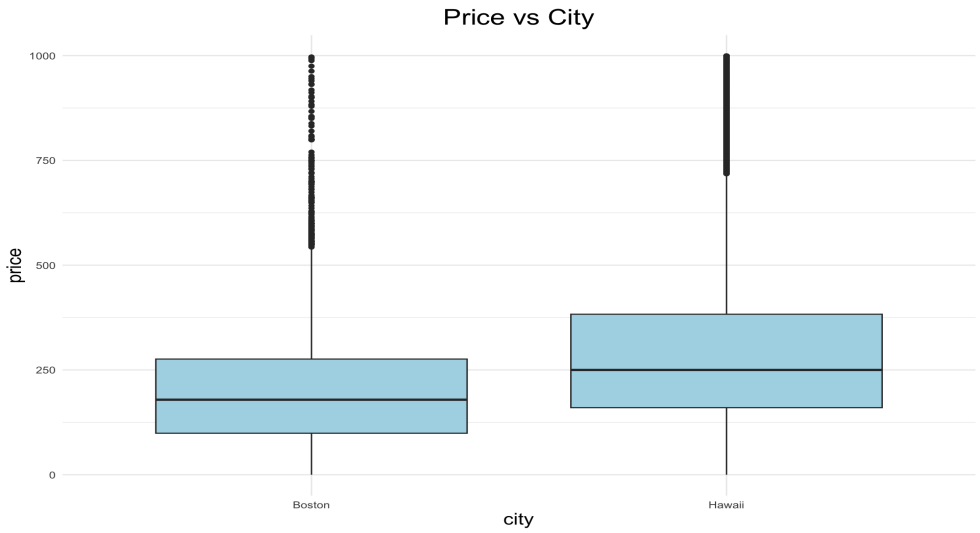

Figure 4.1: Price in differen cities

The box plot shows the distribution of the house prices in two cities. In general, the prices of Hawaii is higher than that in Boston

4.2 Multilevel analysis

The two data sets are combined, and a multilevel linear model with variables((room type, accommodates time, bathroom numbers, bathroom type, bed numbers, rainfall level with the neighbourhoods, and the random intercept with the neighbourhoods) is built with the combined data.

	group	
1	Chinatown	165.55
2	Back Bay	133.45
3	Lahaina	130.07
4	Hana	122.43
5	Koloa-Poipu	111.22
6	South Boston Waterfront	102.73
7	Paia-Haiku	92.65
8	Ewa	91.87
9	Downtown	83.08
10	Lihue	72.28
11	Fenway	70.95
12	Lanai	51.27
13	North Shore Kauai	47.54
14	North Shore Oahu	44.70
15	Kapaa-Wailua	41.19

Table 4.1: top 15 values of random estimates

Table 4.1 shows the random intercept of the model. Surprisingly, although Hawaii has a higher average, when data is combined, Chinatown and Back Bay, from Boston, still have a higher baseline than any other neighbourhoods in Boston and Hawaii.

5 Discussion

1. The analysis towards Boston could go deeper. The traffic around the place, and the famous attractions, there are a lot of other factors that may affect the prices in the city.

2. The residual plots for the multilevel linear models seem to have some patterns, which means the model still needs to be improved.

3. A 3-level multilevel model was tried to be conducted with combined data, while R failed to compute it. It may be due to the high correlation within the group.

4. Prediction, confidence interval...The analysis can also go deeper statistically.

6 Appendix

6.1 Annual Rainfall in Hawaii

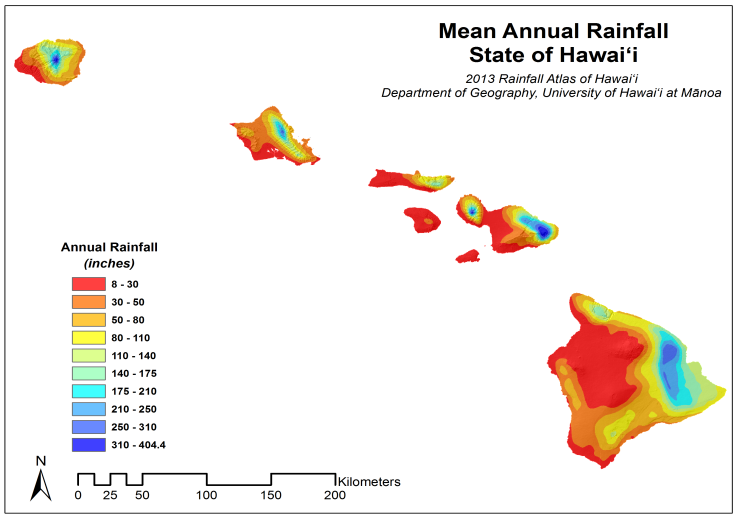

Figure 6.1: Annual Rainfall

Figure 6.1 shows the annual rainfall in Hawaii.
It is downloaded from `http://rainfall.geography.hawaii.edu/downloads.html`.

6.2 Definition of high price

It is determined by the whole data. The price data which is more than 0.85 of other data is defined as high-price data.

6.3 Residual plot for the models

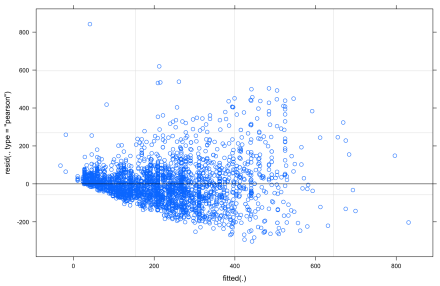

Figure 6.2: multilevel linear model of Boston

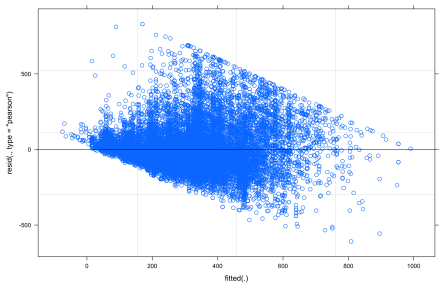

Figure 6.3: multilevel linear model of Hawaii