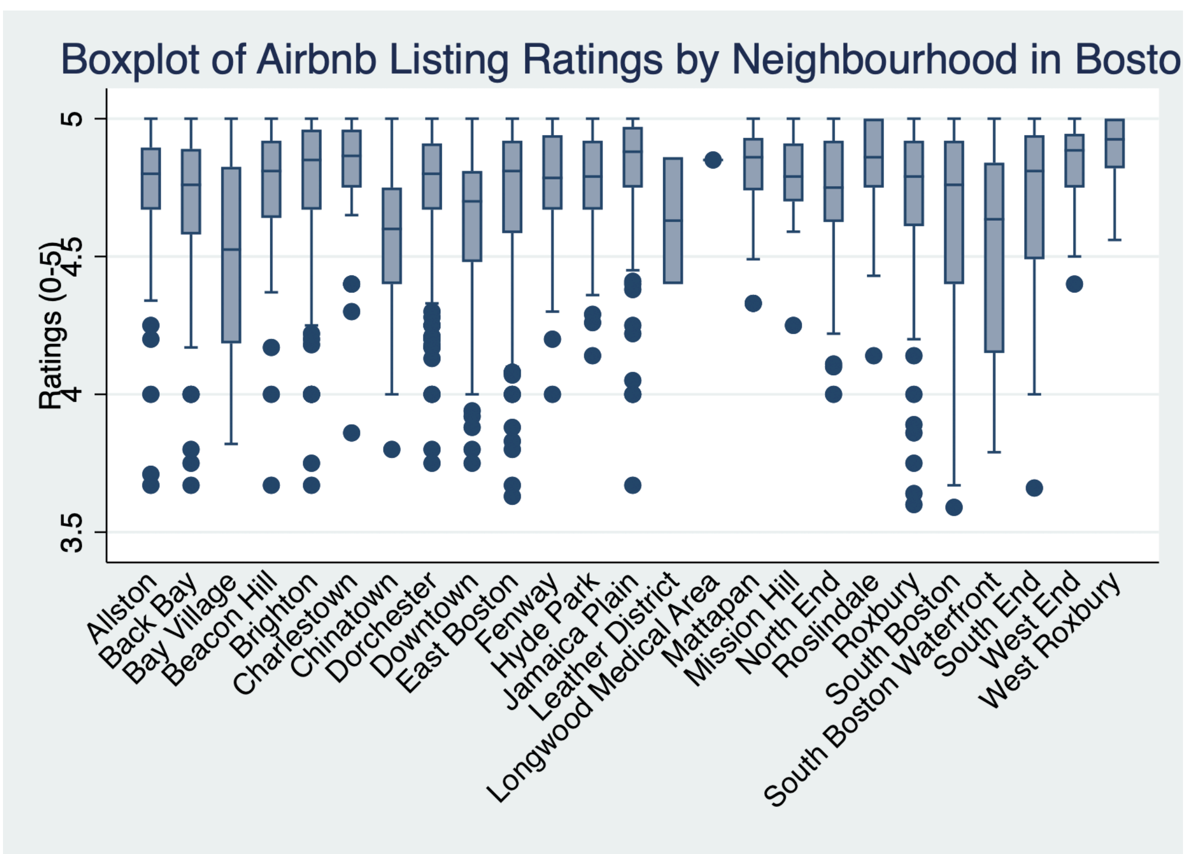
ECON 255 | Data brief # 3
Alessa Somer
7 December 2023

1. Introduction

While Airbnbs pose a direct threat to the hotel industry, Airbnbs are far less standardized than hotels and have more variety in the experience they offer. While Airbnb can offer unique and amazing experiences and places to stay for optimal prices, getting scammed through Airbnb listings is certainly not uncommon. Therefore, like most products, customers look to ratings to inform their decisions. Furthermore, ratings of Airbnbs are one of the few metrics and methods to gauge the reliability and value of an Airbnb as a customer. On the other hand, more and more people are using and managing Airbnb listings as a source of income. In fact, "hosts in the US earned $22 billion in supplemental income last year" according to AirBnb (2023). In the case of some people, they invest and buy properties for the sole purpose of listing the property as an Airbnb. Therefore, if hosts can predict what factors lead to higher ratings for their listings, they may be able to make better-guided choices in regards to their listings. Therefore this analysis focuses on the relationship between rating as the response to potential predictor variables. Specifically, this analysis focuses on the potential neighbourhood location of the listing, the number of reviews the listing has, the reviews per month the listing receives, and the gender of the listing hostname, with the rating of the listing. Often times, customers book Airbnbs solely for its location. Therefore, it is worthwhile to explore the relationship neighbourhood of the listing has on the listing's rating.

Figure 1. Boxplot of Ratings for Neighbourhood in Boston



Note *Figure 1 displays box charts for each of the neighbourhoods in Boston from the Inside Airbnb data. On the x-axis, we see the different neighbourhoods in which there are Airbnb listings in Boston. On the Y axis, we have the average rating for Airbnb listing. Based on the boxplots, we can see that there is variation in the ranges and median ratings for various neighbourhoods in comparison to each other. Specifically, we can see that Bay Village has a lower median rating and a larger range of ratings compared to most of the other neighbourhoods. The variance in ratings among neighbourhoods prompts investigation.*
Data source*: Inside Airbnb Boston and Airbnb*

For this analysis, I will be using Inside Airbnb's Boston dataset. Inside Airbnb is a mission-driven project that provides data and advocacy about Airbnb's impact on residential communities. Inside Airbnb has verified and cleaned public data published by Airbnb. Therefore, the data used in this analysis is public, readily accessible for verification, and reliable. Specifically, I use a cleaned version of InsidesAirbnb's Boston data called "Clean AirBnb Listings." This data set contains variables for each listing such as rating, name, number of bedrooms, number of beds, number of bathrooms, neighbourhood, price, reviews per month, etc. However, this analysis will focus on the relationship between rating and neighbourhood, number of reviews, reviews per month, and gender of host specifically.

2. Data

The dataset used in this analysis is from InsideAirbnb's Boston dataset. The unit of observation for this data set is Airbnb listing. The main variables of focus for this analysis are Airbnb listing rating, neighbourhood, number of ratings, reviews per month, and gender of host of the listing. We must include the number of ratings the listing has in our model, as there is potential for customers to be biased about their rating on Airbnb based on previous ratings. It is also important to include the number of reviews per month the listing receives, as this could be an indicator of how many different people stay in the listing per month because to rate the Airbnb listing, you must have stayed at the Airbnb.

Only simple preprocessing of the data and transformations of the variables were needed for this regression. First, we converted the response variable, rating, from a string into a long. Second, we encoded the neighbourhood variable as a categorical variable, where each neighbourhood represents a different category. After conducting univariate analysis, it was revealed that there are varying frequencies of listings from each neighbourhood included in this dataset. For example, the neighbourhood of Dorchester has 310 observations in the dataset, while the neighbourhood of Longwood Medical Area only had one observation. Thus, to ensure that our model is stable, we decided to only include neighbourhoods which have a similar number of observations. Specifically, we only included neighbourhoods that had between 3-6% of the proportion of total observations in the original dataset. Our subset of neighbourhoods for which we analyzed are Allston, Back Bay, Beacon Hill, Brighton, Charlestown, Fenway, Roslindale, South Boston, and South End.

While, the primary focus of this analysis is to study the relationship between rating of Airbnb listings and neighbourhood location, it could be interesting to study the relationship between the rating of the listing and gender of the listing host. In order to preprocess the data to get the gender of the listing host, we used a Python package called "gender-guesser." The python package takes each name and returns its guess of what the gender is of the person with the name. The python package is reliable, having been trained on over 40,000 first names with over 98% accuracy. However, this package was created in 2016 and therefore has the potential for hazards and inaccuracy. In addition, there certainly is no absolute gender for a name, and in fact the typical gender of a name varies by culture, and therefore these results should only be relied on lightly. To gather the data, we iterated through the "host_name" of each listing and generated the gender of each host. The algorithm has the potential outputs: "male," "female," "mostly female", "mostly male", "androgynous", and "unknown." After the algorithm outputted the results, I manually went through the first 100 names and gave the sample a 100% accuracy based on what my guess of the gender of the host would be. Lastly, I collapsed the categories for simplicity to be "male and mostly male" to become "male", "female and mostly female" to become "female", and "unknown and androgynous" to become "ambiguous."

There are potential hazards for this dataset. First, it is possible that some of these listings are not currently up to date. Additionally, there is the possibility for confounding factors that relate to our response variable of listing ratings that are not represented in the data, and subsequently the regression models. For example, there is the possibility that price could have an impact on the rating of an Airbnb listing. However, in this dataset, price and rating for listings do not have a linear relationship, and therefore we can't include price in our regression model.

3. Results

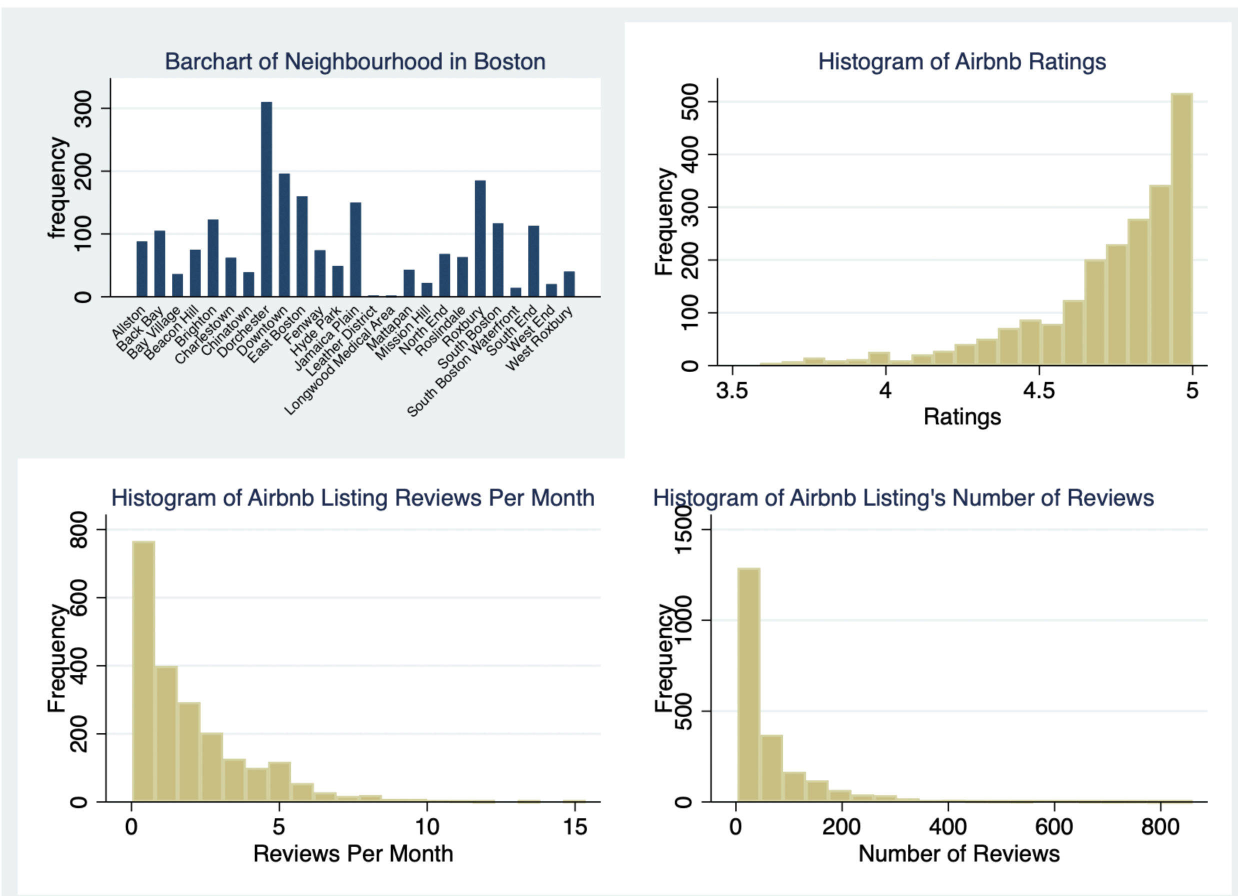
Four OLS regressions were run to investigate the relationship between Airbnb ratings and its listing information. OLS creates easily interpreted results. The first regression ran rating on neighbourhood to look at the relationship between the two variables before adjusting for other possible predictor variables. The second regression ran rating on neighbourhood and reviews per month. The third regression ran rating on neighbourhood, reviews per month, and total number of reviews. The fourth regression ran rating on neighbourhood, reviews per month, total number of reviews, and gender of the listing host. Additionally, since the gender of the hostname for each listing is only a prediction, we included them only in the full model, while still running regressions on the reduced focus variables. OLS was chosen as the model because of its high interpretability. In addition, OLS is a great way to measure continuous response variables. The main purpose of this paper is to study the relationship between neighbourhoods and ratings. Therefore the initial model only includes neighbourhoods as the predictor variable. In additional models, we include predictor variables of the number of reviews and reviews per month each listing has to ensure that we are not excluding important confounding variables from the model. However, based on the scatterplots in Figure 3, we must approach these model with caution, since the relationship between rating and number of reviews per month and the relationship between rating and total number of reviews is not exactly linear. Based on the scatterplot, we can see that there is a positive association, but due to the high frequency of ratings with high values and low x values, the linear relationship is broken.

3.1. Descriptive Statistics

Figure 2 displays the univariate distributions of the variables in our model. Based on the bar chart of the neighbourhoods for Airbnb listings in Boston in Figure 2, we can see the varied frequencies of neighbourhood observations. For example, Dorchester has a high frequency of 305, while

Longwood Medical Area has a frequency of 1. The histogram of the Ratings of Airbnb listings in Boston shows that, on average, the ratings are relatively high. The distribution of the listing ratings is left-skewed and unimodal. With the outlier removed, the mean of the listing ratings is 4.7. This left skewness is logical, as listings with low ratings are less likely to get booked, and thus would have a hard time staying on the market. Therefore, it makes sense that our data does not capture many listings with low ratings. The distribution of the number of reviews each Airbnb receives is right skewed and unimodal with a peak at around 1 review. The mean number of reviews each Airbnb receives per month is 2.02. Additionally, the histogram reveals that there are outliers in the upper range. The distribution of the total number of reviews for each listing is left skewed and unimodal with a peak at around 20 reviews. The mean number of total reviews for each listing is 68.7. Additionally, the histogram and numerical summary reveal that the total number of reviews contains outliers in the upper range.

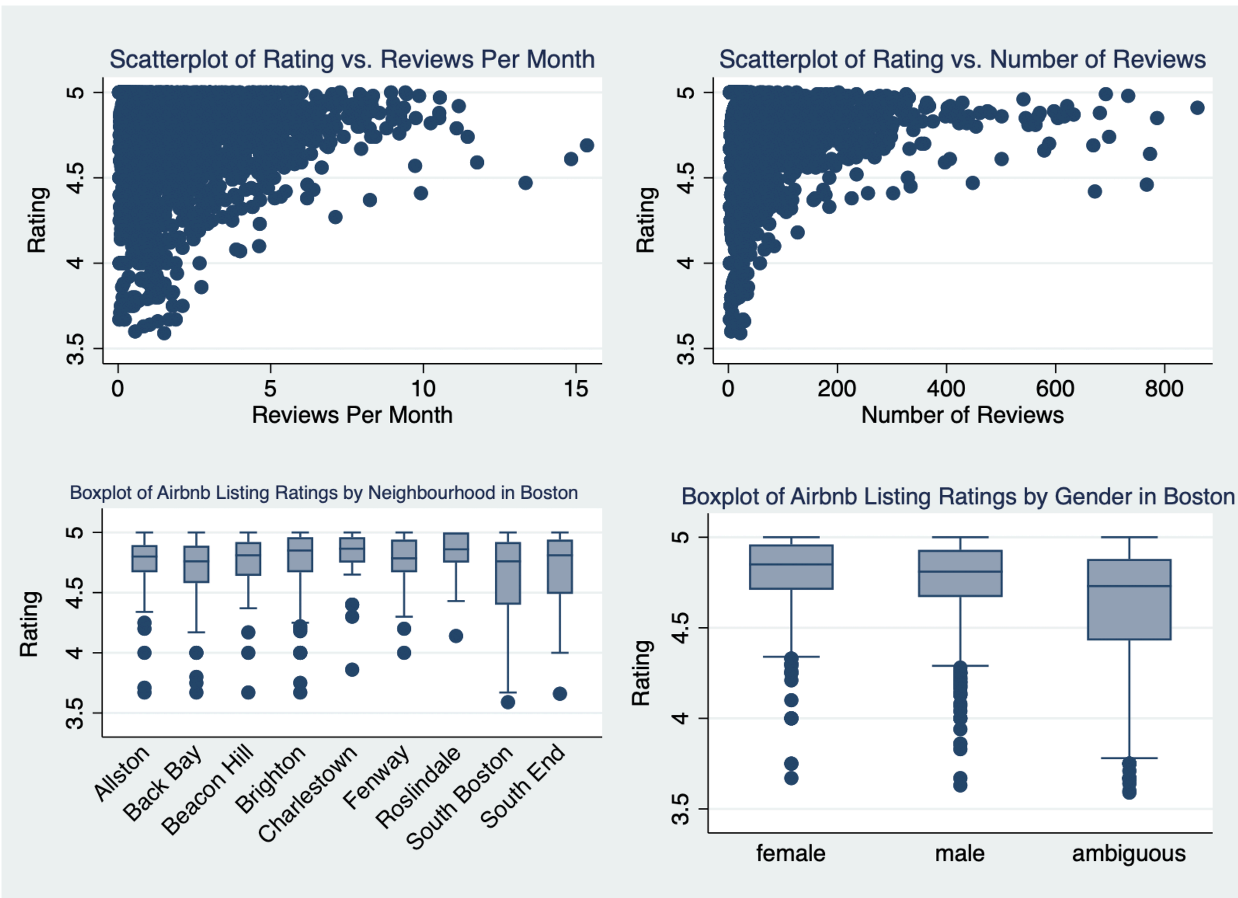
Figure 2. Univariate Distributions of Neighbourhoods, Ratings, Number of Reviews, and Reviews per Month



***Note**: Figure 2 shows the univariate summaries of the two main focus variables in this analysis. Upper Left: Barchart of the Neighbourhood Listings in Boston. The x-axis shows the different neighbourhoods in Boston that have Airbnb listings. On the y-axis are the frequencies of each neighbourhood listing. The Upper Right: Histogram of the Ratings of Airbnb listings in Boston from the InsideAirbnb data. On the x-axis are the ratings of each listing on*

a scale from 0-5. The y-axis is the frequency of each rating. Note: One outlier for a listing with a rating of 0 was removed in this histogram to show the ratings' distribution. Lower Left: Histogram of the number of reviews each Airbnb listing receives per month. The x-axis shows the reviews per month the Airbnb listing receives, and the y-axis shows the frequency of the reviews per month. Lower Right: Histogram of the number of reviews an Airbnb listing has received in total. The x-axis shows the number of reviews the Airbnb listing has in total, and the y-axis shows the frequency of the number of reviews.
Data source*: Inside Airbnb Boston and Airbnb*

Figure 3. Bivariate Summaries of Rating and Reviews Per Month, Number of Reviews, Neighbourhood, and Gender



Note*: Upper Left: Scatterplot of Rating vs. Reviews per Month of Airbnb Listings in Boston. On the x-axis are the reviews per month for a listing in Boston, and on the y-axis are the ratings. Based on the scatter plot, we can see a positive correlation between ratings and reviews per month, although it is not approximately linear. Upper Left: Scatterplot of Rating vs. Number of Reviews of Airbnb Listings in Boston. On the x-axis are the total number of reviews for a listing in Boston, and on the y-axis are the ratings. Based on the scatter plot, we can see a positive correlation between ratings and total listing reviews, although it is not approximately linear. Lower Left: Boxplot of Ratings by Neighbourhood of Listings in Boston. On the x-axis are the neighbourhoods for listings in Boston, and on the y-axis are the ratings. Based on the boxplots, we can see slight variations in the median ratings among different*

neighbourhood listings. Additionally, we can see different ranges and interquartile ranges of ratings for different neighborhood listings Lower Right: Boxplot of Ratings by Gender of Hostname in Boston. On the x-axis are the three gender categories represented in the data (female, male, and ambiguous), and on the y-axis are the ratings. Based on the boxplots, we can see that listings with evidently female host names have a slightly higher median rating. In addition, listings with either female or male host names have higher median ratings than listings with gender-ambiguous host names.
Data source*: Inside Airbnb Boston and Airbnb*

The relationship between ratings and the neighbourhoods that listings are in is interesting to look at from an economic perspective because this relationship could measure how much visitors enjoy staying in certain neighbourhoods. Managing Airbnb properties has become a source of income for many people. However, in order to have listings on Airbnb, the individual must own or manage property, which are large investments. If people can measure and guage what qualities in properties have the potential to be successful Airbnbs, this can help direct people when investing in properties. The second part of this analysis discusses how the gender of the listing host based off of their name, impacts the rating of the listing. Investigating this relationship is interesting because we have the potential to measure how gender steretypes might influences people's choices in booking Airbnb's. Unlike neighbourhood of the Airbnb listings, which are impossible to change once owning a property, it is an easy process to alter the profile name for your listing on Airbnb. If there are differences among the relationships between ratings and genders, it is worth investigating what gender stereotypes may be influencing these differing relationships. Additionally, Airbnb hosts could be loosing profits from a bad rating resulting from their name posted on their listing. Hosts can easily alter their name to sound more like people's preferred host gender in order to genrate more profits.

3.2. Regression Results

Coefficient estimates, standard errors, and p-values from our final models are shown in Table 1. Based on the first three regressions, we can see that the standard errors are smallest for the model with rating as the response variable, and neighbourhoods and reviews per month as the predictor variables. Thus we should use this model for our interpretations, as it is able to explain the most variability in ratings. In the second regression model, we found that the expected ratings for Airbnb listings in Charlestown are higher than ratings for listings in Allston by 0.0948, after adjusting for the number of reviews per month the listing has; with $p < 0.01$, this coefficient is significantly different from zero. In the second regression model, we found that the expected ratings for Airbnb listings in Roslindale are higher than ratings for listings in Allston by 0.0895, after adjusting for the number of reviews per month the listing has; with $p < 0.01$, this coefficient is significantly different from zero. In the second regression model, we found that the expected ratings for Airbnb listings in South Boston are lower than ratings for listings in Allston by 0.150, after adjusting for the number of reviews per month the listing has; with $p < 0.01$, this coefficient is significantly different from zero. We found that each review increase in the number of reviews per month for an Airbnb listing in Boston is associated with a 0.0235 increase in the listing's rating on average; with $p < 0.001$, this coefficient is significantly different from 0. The regression in column 4 of Table 1 displays the model results of rating on neighbourhood, number of reviews, reviews per month, and gender of the hostname. From this regression model, we found that the neighbourhoods of Roslindale and Boston still had differences in associations than the neighbourhood of Allston with ratings. Additionally, reviews per month still has a significant positive association with ratings, after adjusting for other variables. Based on the regression model in column 4, we found that the expected ratings for listings with either male or ambiguous-sounding hostnames are lower than expected ratings for listings with female-sounding hostnames; with $p < 0.01$, both coefficients are significantly different than zero.

Table 1. Rating Regression Model Results

	(1)	(2)	(3)	(4)
	numeric_rating	numeric_rating	numeric_rating	numeric_rating
Allston	0	0	0	0
	(.)	(.)	(.)	(.)
Back Bay	-0.0487	-0.0577	-0.0476	-0.0120
	(0.0388)	(0.0383)	(0.0387)	(0.0378)
Beacon Hill	0.00152	-0.00550	-0.00567	0.00457
	(0.0394)	(0.0390)	(0.0390)	(0.0362)
Brighton	0.0103	0.0108	0.0153	0.00597
	(0.0372)	(0.0362)	(0.0362)	(0.0345)
Charlestown	0.0854*	0.0948**	0.0882*	0.0526
	(0.0368)	(0.0362)	(0.0366)	(0.0362)
Fenway	0.0335	-0.00221	0.00607	-0.00884
	(0.0358)	(0.0358)	(0.0359)	(0.0353)
Roslindale	0.0901**	0.0895**	0.0981**	0.0769*
	(0.0345)	(0.0336)	(0.0339)	(0.0321)
South Boston	-0.143**	-0.150**	-0.150***	-0.118**
	(0.0466)	(0.0455)	(0.0452)	(0.0406)
South End	-0.0308	-0.0356	-0.0400	-0.0121
	(0.0382)	(0.0376)	(0.0377)	(0.0343)
reviews_per_month		0.0326***	0.0235***	0.0302***
		(0.00504)	(0.00541)	(0.00525)
number_of_reviews			0.000324***	0.000179*
			(0.0000935)	(0.0000902)
female				0
				(.)

male				-0.0568**
				(0.0188)
ambiguous				-0.239***
				(0.0251)
Constant	4.742***	4.691***	4.685***	4.771***
	(0.0273)	(0.0291)	(0.0294)	(0.0277)
N	820	820	820	820

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Note*: Column 1 shows the regression of* Rating *on Neighbourhoods in Boston of Airbnb listings. Column 2 shows the regression of* Rating *on Neighbourhood and the number of reviews Airbnb listings receive per month. Column 3 shows the regression of* Rating *on Neighbourhood, the number of reviews Airbnbs listings receive per month, and total number of reviews the Airbnb listing has. Column 3 shows the regression of* Rating *on Neighbourhood, the number of reviews Airbnbs listings receive per month, total number of reviews the Airbnb listing has, and assumed gender of the host based on name. Note: gender was predicted by Python algorithm "gender-guesser." Underneath each coefficient, is the coefficient's standard error. We can see that coefficients with the smallest standard errors are the coefficients from Regression 2 and 4. Every variable in the regression is included in Table 1.*
Data source*: Inside Airbnb Boston and Airbnb.*

4. Conclusion

It is logical that there are differences in ratings of listings for different neighbourhoods in Boston. Rent prices are often times driven by the "desirability" of the neighborhood of the property. Therefore, it is plausible that Charlestown and Roslindale have higher expected ratings, while South Boston has lower expected ratings. It would be interesting to investigate the differences between these neighbourhoods, such as average income for households in these neighborhoods, and why perhaps there might be a difference in ratings for Airbnb listings. It would also be interesting to investigate the consensus of what might be the most "desirable" neighborhoods in Boston are, and see if these correlate in any way to the positive and negative associations found in this analysis. We also found that there is a positive association between reviews per month and rating. Perhaps, people only feel the desire to make public rating if they have had a positive experience with the Airbnb. Lastly, the analysis shows that listings with host names that sound female are associated with high expected ratings, after adjusting for neighbourhood of the listing, reviews per month of the listing, and number of reviews total for the listing. This is an interesting finding, and has potential for further analysis. It is worthwhile to gather information on what qualities customers look for in an Airbnb host, and why people may think female hosts embody these qualities over other genders. In conclusion, these findings are intriguing and help share the story of what people's current preferences for Airbnbs.

5. References

Airbnb. (2023). Retrieved from https://news.airbnb.com/hosts-in-the-us-earned-22-billion-in-supplemental-income-last-year/

Inside Airbnb (2023). Retrieved from http://insideairbnb.com/

Python (2016). Retrieved from https://pypi.org/project/gender-guesser/