Exploring Real Data: A look at AirBnB

Amanda Francis and Eric Sullivan

## Airbnb

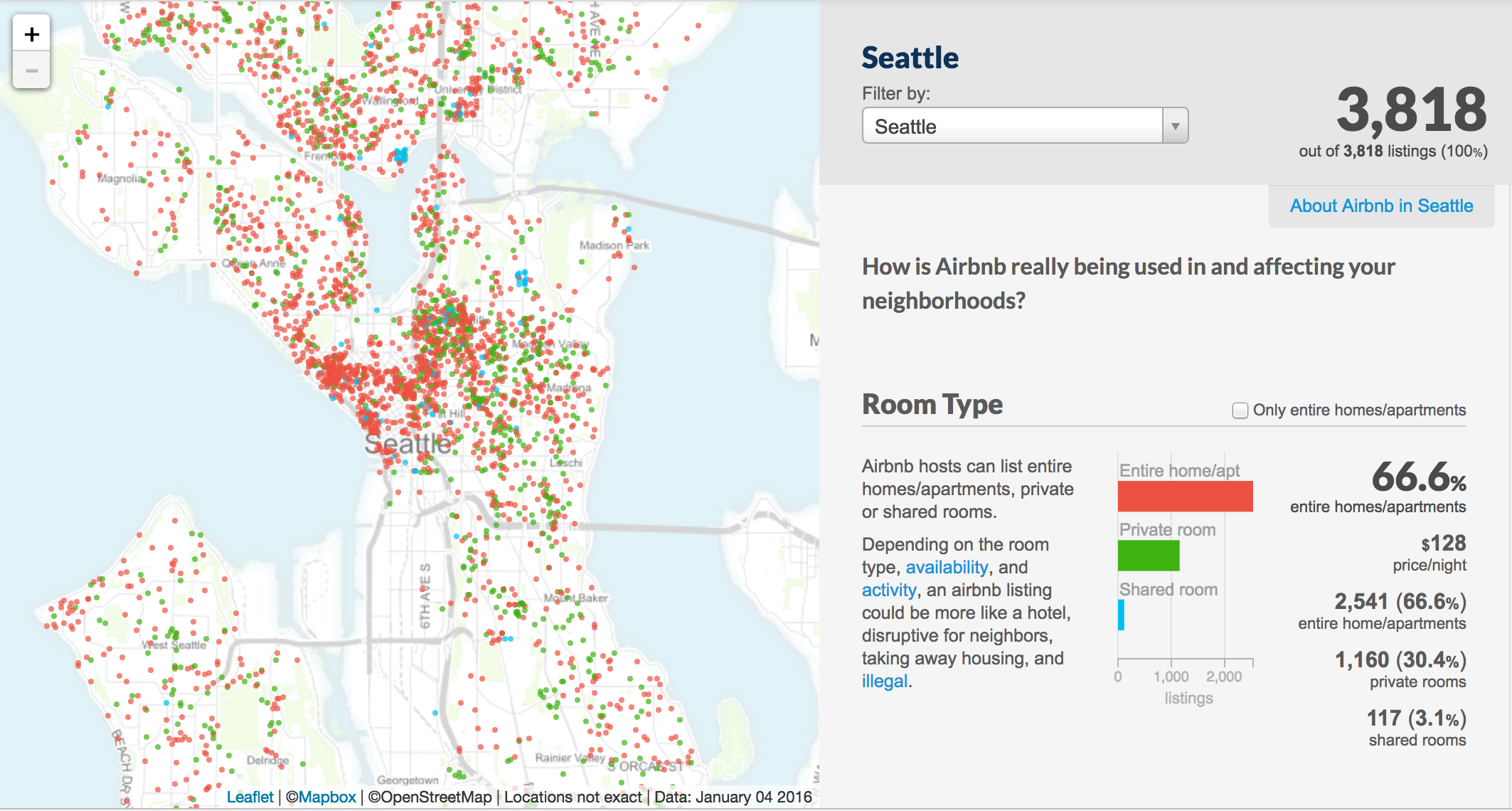
We have all likely experience the traveler’s dilemma: I want to travel cheaply so that my trip can last longer, but the cost of lodging is prohibitively high. In 2008 the San Francisco based company Airbnb changed the nature of finding lodging by creating “ … a trusted community marketplace for people to list, discover, and book unique accommodations around the world”. In their words: “whether an apartment for a night, a castle for a week, or a villa for a month, Airbnb connects people to unique travel experiences, at any price point, in more than 65,000 cities and 191 countries” (www.airbnb.com).

As Airbnb has grown in popularity a wealth of data has been accumulated about many aspects of the Airbnb experience. One particular independent website, insideairbnb.com, has developed “ … a set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.” Insideairbnb has gathered public data from Airbnb sites around the world and as mathematicians and statisticians this provides us with a treasure trove of questions to explore! They have data on cities like Paris, Venice, Amsterdam, Boston, Chicago, Seattle, and many more. Their data sets include factors such as average reviews, locations, cleaning costs, rent, and much more.

What follows are a collection of explorations that you can try out along with us. After each exploration we’ll give some of our own thoughts,discussion, and visualization. We encourage the reader to head over to insideairbnb.com, find some interesting data, and start exploring.

### Exploration #1:

Go to insideairbnb.com now and find a data set for the city of your choice. Spend a few minutes exploring the data set and propose several questions. When you press “Get the Data” on insideairbnb.com and find your city you will see a link “see data visually here”. Spend some time exploring their visual aids.



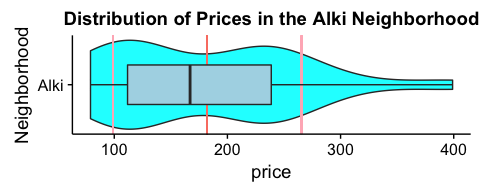
We are located somewhat near Seattle, so for the purposes of discussion we’ll use the Seattle data. You are welcome to use whichever data set you find most interesting. The graphical visualization tool allows you to see several descriptive statistics for your chosen city, animate through the frequency and location of reviews, see top rated hosts, and much more.

## Initial Exploratory Data Analysis

Let’s explore a some questions related to your particular city (These explorations are meant to increase in difficulty as you work through them).

### Exploration #2:

You’re trying to budget your travel money. In what range of daily prices would you expect the average daily price for a private room at an Airbnb in a particular neighborhood to land? **This data isn't normally distributed, so we need to be careful giving confidence intervals for single instances...**

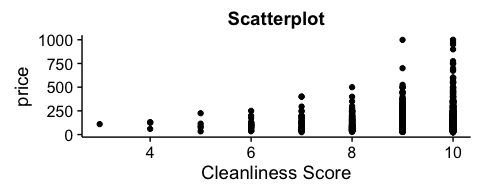
**Discussion:** There is a beautiful beach on the Puget Sound in the Alki neighborhood of Seattle. Let's filter the data to explore the prices from the Alki beach area. We'll start with some basic descriptive statistics: the average rental price there is $170 but the standard deviation is $97! The question in this exploration is really asking for a range of prices, and based on the rather large standard deviation, the estimation range will have a rather large range as well. If this data were normally distributed we would expect *most* of the rentals (~68% of them) to be withing one standard deviation of the mean: between $73 and $267. The large variance in the prices leads us to believe that there is more going on here. Let's create a data visualization to investigate further: 

The plot above is called a violin plot. The curved shape tells us where 'clusters' of the data are located, while the boxplot tells us where the median and quartiles are located. We observe that this data does *not* appear to be normally distributed. In fact, a relatively low concentration of the rentals in the Alki neighborhood are near the median price; many are between $40 and $80 away from the 'center', in other words, this is a bimodal distribution of data. We could investigate the types of rentals in this data, or perhaps, their distance from the coastline to figure out what is creating this interesting shape. In addition, if we want to use prices in this neighborhood as part of some further analysis, we might consider performing some kind of variable transformation. Since the prices here are all greater than zero, bimodal, and fairly "right skewed"", we notice that the distribution is not 'nearly normal' like we have come to expect! A square root or log transformation might prove useful for further analysis.

### Exploration #3:

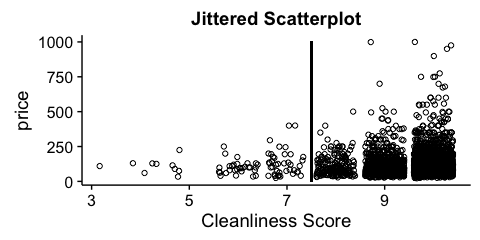
Do you wonder if hosts with cleaner rentals charge more? How about a more specific question: is there a statistical difference in daily price between the rentals that are rated most cleanly versus the rentals that are rated least cleanly?

**Discussion:** The first order of business in this explanation is to decide what *most cleanly* and *least cleanly* mean to us. The cleanliness rating scale goes from 0 to 10, but we shouldn't assume that cleanliness ratings are evenly distributed among these numbers, or that the numbers 0-10 represent a nice linear progression of cleanliness. In fact, many travelers might feel guilty giving a low cleanliness rating, so we may expect the ratings to be artificially inflated. Having learned from our last exploration, this time let's look at a visualization of the data first. How about a scatterplot of cleanliness rating versus price?

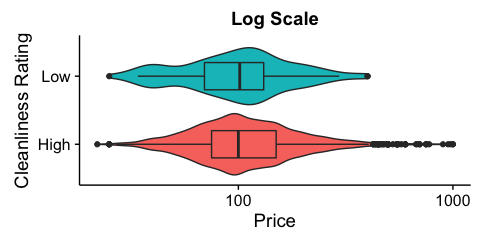


One problem with this graphic is that the dots we've plotted are likely hiding information from us! Some of the dots might represent only one observation, while some may have many rental observations which happen to have the same cleanliness review score and price as many others.

One option is to add a ‘jitter’ to our graphic, so that we can see all of our points (or at least get a sense about where our clusters are.



Here we notice that the sample sizes are wildly different: 80 rentals with lower cleanliness ratings and 3,165 with higher ratings. Perhaps people are just too nice in their ratings or perhaps the Airbnb properties in Seattle are clean! It *looks* like the rentals rated higher for cleanliness get better prices, perhaps with a few outliers. Our scatterplot dots are still so clustered that it makes it hard for us to see what's going on. Let's try one more data visualization. In the last exploration we looked at a violin plot, and it seemed useful; this time let's look at a side-by-side violin plot. In this case, we've added a log-scale to the price variable, to make the shapes easier to see.

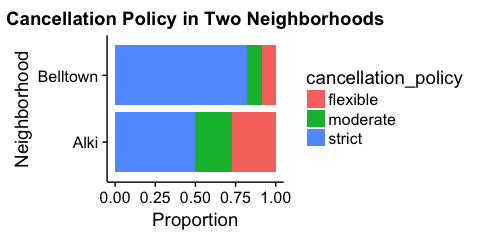


We see that actually, the bulk of the rental prices for *both* High and Low ratings fall below $250, but the high rentals have some 'outlier' observations with much higher prices. On the whole, though, the average rental in either category doesn't appear to be much different. If we conduct a statistical -test on two means, we come to roughly the same conclusion (with a p-value of about 0.26), so there doesn't appear to be strong evidence of a difference in price between the two groups.

### Exploration #4:

I want to be safe and allow myself an out in case my travel plans fall through, but I still want to have several Airbnb rentals to choose from. Is there a difference between the proportion of rentals in one neighborhood that have a strict cancellation policy as compared to the rentals in another neighborhood?

**Discussion:** Let’s say that we want to stay on the waterfront in Seattle so we’ll compare Alki beach and Belltown; both of which are on the water.

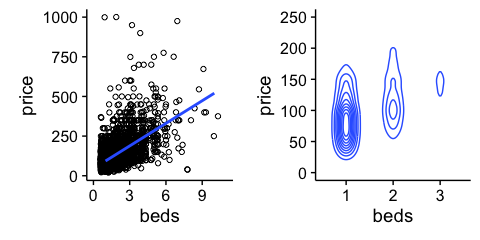


The visualization above shows us the distribution of cancellation policies in the two neighborhoods, and a statistical test on the difference of the proportions of strict policies in each neighborhood gives us a p-value around .0004, and we conclude that there is evidence to suggest that rentals in Belltown have a much stricter cancellation policy than those in Alki. The safe bet seems to be Alki, but if we want to spend our social time in downtown we now need to consider the transportation costs.

### Exploration #5:

Now I'm considering inviting some friends on my trip, but first I want to know: For each additional bed, how much more should I expect a rental to cost?

**Discussion:** Let's start with a visualization of our data points in a (jittered) scatterplot and a contour plot showing which combinations of number of beds and rental price occur most frequently.



A simple linear regression tells us that the best-fit line for our data is approximately , so, I should expect to pay about $93 for a rental with one bed and $47 more per additional bed. As an additional exploration, you might try find a 95% confidence interval for the true slope (marginal price) of this regression line. What about a 95% confidence interval for the average price of rentals with 3 beds? What about a predicted interval for the *actual* price of a rental with 3 beds? (How are these questions different from each other?)

## Moving Beyond the Basics

In this section we pose more questions and leave the investigation open to you. If you would like to see how we answered these questions wth the Seattle data, you can visit our workfile at mathquest.carroll.edu. Several of these exercises suggest more advanced techniques, such as multiple regression, logistic regression, or machine learning techniques to answer.

### Exploration #6:

What variables in this data set do you think would be the most useful in predicting the price of a particular rental? Try accommodates, beds, bathrooms, guests\_included, review\_scores\_rating, and various neighborhoods, and see if you can build a model (maybe a multiple regression, regression tree, or artificial neural network) that is relatively good at predicting prices. A good idea would be to split your data into a training and test sets so that you can test your predictions on data that your model has not seen. Which variables turn out to be the most important? The least inportant? Does this surprise you? Are there any variables that amount to just 'noise' in this analysis? How reliable do you think your model is? Do you think this same model would work in other regions of the world? Why or why not? Do you think it would be necessary to pre-process that data in any way? For example, do you think that it would be necessary to do a principal component analysis on your data first in order to reduce the dimension of the data set? Do you think that the variables should be transformed in any way? Do you think any additional features should be included in your data (e.g. the square of a feature, the product of two features, etc)?

### Exploration #7:

Can you use the data to predict whether a rental is an "Entire home" or a "Private room"" based on the other characteristics of that rental? What variables in this data set do you think would be the most useful to you? Try using the same variables as in the last exploration (including price) to build a model? Which variables are the most important? The least important? Does this surprise you? Are there any variables that amount to just 'noise' in this analysis? How reliable do you think your model is? **Do you think this same model would work in other regions of the world? Why or why not? Once you have a model working on your chosen city, try your model on a different city and see how it performs.** It would also be wise in this case to split your data into a training set and a test set so that you can measure the performance of your model.

### Exploration #8:

Can you use the data to predict which neighborhood a rental is in based on other characteristics in the data? Pick three of four neighborhoods which contain "many" rentals and keep only data from these neighborhoods. Can you build a model (maybe hierarchical clustering, k-means, k-nearest neighbors, support vector classifiers, or something else) which can predict which neighborhood you're in, based on certain *important* variables? Which variables do you think you should use? Do you think using latitude and/or longitude is like cheating? How well does your model work?

## Conclusions

The Insideairbnb data set is robust enough that anyone with statistical training, no matter the sophistication, can ask meaningful and challenging questions. In recent times the field of data science, along with the associated mathematical tools, has become a successful and popular way to analyze data sets of this type. If you want to learn morea bout exploring and using data for analysis and predictions there are starter courses on the web (e.g., Coursera and DataCamp) and many universities and colleges around the world are implementing Data Science programs. Most importantly, as you may have found from the last three exercises, the level of mathematical sophistication associated with data analysis can be quite high, so if you are a student reading this article and you find the ideas interesting then we recommend taking courses in statistics, computer science and computational mathematics.

There are many data sets like Insiderairbnb available for free on the internet. For example, the University of California, Irnive, hosts a collection of machine learning data sets that are at least as rich and interesting (archive.ics.uci.edu/ml/). Ohter sites to look at when seeking include data.gov, kaggle.com, quandl.com, gapminder.org, flowingdata.com, and many others. The world of data and data analysis is growing in importance socially and mathematicians and statisticians are uniquely positioned with their training to approach these questions with proper skills and mindsets.