How much should you charge someone to live in your house? Or how much would you pay to live in someone else’s house? Would you pay more or less for a planned vacation or for a spur-of-the-moment getaway?

We are trying to figure out the factors that influence Boston Airbnb price. Our variables include:

* Host Features(Hosting Year, If Host respond message fast, if Host is “Airbnb Super Host”…)
* Property Features(Location, Number of rooms, Room type, Number of beds, Amenities…)
* Service Features(Security Deposit, Cleaning Fee, Extra People Price…)
* Reviews(Number of Review, Average Review Score, Latest Review time…)

To Start:  
As we start our regression model, we reviewed our datasets and went several extra data cleaning steps for better regression expression.  
1) Deleted variables such as “Host Name” and “Host ID” since it won’t influence the price, there is no way that host “Bills” can reasonably charge more can host “Johns”.

2) Turn listed Amenities Text to Amenities Count. There are 50+ different amenities type on Airbnb, however we noticed that actually almost all hosts provide those essential types such as wifi and hot tubs. So we just use the count of different Amenities.

3) Removed Zipcode instead using “Neighborhood” Neighborhood are more accurate and fit consumer’s behavior better.

4) Removed some other overlapped meaningless variables and reserve only one such as “30 days avability”,”60 days avability”, and “90 days avability”

Part 0. Variable Selection

We ran a regression for all variables and applied caret package to test the importance of variables with a 5-fold cross-validation, here is the result.

#<Code Here>

We can see that Based on our experience on Airbnb rent and our observation on the dataset, number of bathrooms, accommodates(How many people can accommodate), location(Neighborhoods) and Room Types are most influential factors. Based on this brief impression, we picked XXX,XXX,XXX,XXX,XXX as our variables.

Part 1. Descriptive Statistics

Use descriptive statistics, plots, and visualizations to do the following:

1. Show descriptive statistics for relevant and important variables.

* box-and-whisker plots for relevant and important variables.
* the minimum, maximum, and average (mean, median, mode) and standard deviation /variance of important variables.

2. Create a scatterplot among the variables to find potentially linear or curvilinear relationships.

That should help you identify both a target variable and candidate predictor variables.

3. Choose a target variable and justify that choice.

Firstly, we analyzed our predictor: average\_price we can see the distribution of the price skewed leftward, most Airbnb have price range from $50-$250, mean price of all Boston-Area Airbnb is $209.17, but there is also a fat tail at around $500, those might be whole-rent Luxury apartment/house

#<Code Here>

Then we examined other variables and their relationship with price. At the beginning ,we box-plotted several must-include variables and described their relationship with our price.

1. Bed type and bedroom type

Most Airbnb provide real bed or airbed given their average price are higher than 125. Airbed, futon and pull-out sofa price is slightly lower bed types do influenced price. Most Airbnb provide entire home/apt for better privacy. Privacy is valuable, as a result, Entire home/apt also more expensive than Private Room and Shared Room.

#<Code Here>

1. Number of Bedrooms and number of beds

Despite the fact that 5-bedroom and 6-bedroom listings’ price have very wide distribution, we can see bedrooms amount pushes up the price. Same as bedroom number, bed number also have a relatively positive relationship with price. We believe there might be strong correlation between these two variables, which we will discuss later.

#<Code Here>

1. Safety Issue: host\_has\_profile\_pic & host identity verified

Seems like safety are not really a big deal for Airbnb pricing, hosts whose profile picture uploaded or identity verified can’t bid on higher price. We need further t-test to check if their average price are different.

#<Code Here>

1. host listings count

We assume that a hosts with many Airbnb properties are more likely adjust price, from the distribution we can see that actually 50% hosts have less than 4 listings. The price would climb slightly higher when more listings hosted by same person, this situation become especially obvious in the latter 50% hosts with more than 4 listings.

#<Code Here>

1. review\_scores\_value

People seems really satisfied about Boston’s Airbnb, overall average review score is 9.26 with only 0.85 std. When we look at the box plot, we can realize those expensive listings tend to have 9+ review. But in fact 9+ reviewed Airbnbs’ price scattered from lowest to highest, and the average price of 10s are even lower than others. Maybe low price relaxed people’s requirement.

#<Code Here>

1. review\_per\_month

Every Airbnb listing gets around 2 reviews every month, while the most frequently reviewed Airbnb can get 13+. The frequency of people reviewing Airbnb seems have negative relationship with the price, low price Airbnb get higher review, maybe because low-priced Airbnbs are booked more frequently.

#<Code Here>

1. Cleaning\_fee & Security Deposit

Despite the fact that many hosts required 0 cleaning fees, the average cleaning fee for a listing is $72, which is slightly overpriced in our view. Cleaning fee goes up with the price, obviously luxury room required better maintenance.

#<Code Here>

1. host\_response\_rate

Seems like host with higher response rate also can not require higher price, their effort on actively marketing their listings makes them undervalued their properties.

#<Code Here>

After all the variable visualization and logic analysis , we have chosen variable below as our target variables for further regression modelling:

(麻烦Modelling的同学填一下)