425 Final Project

*Data analysis and visualization of Boston AirBnb data*

* Group:
  + Gopalakrishnan Shanker (NetID: gs10);
  + Manasa Hariharan (NetID: manasah2);
  + Marcos Feole (NetID: mfeole2).

1. Introduction

In the present work we performed a data analysis and visualization of different traits present in an AirBnb dataset from the Boston city[[1]](#footnote-0). Airbnb is an online lodging rental marketplace, where people can rent out their homes, rooms and beds to guests. Airbnb does not own any of the rooms on its website, it is merely a broker that facilitates the interaction between the host and the guest and receives a commission fee from both of them. The cost of lodging is chosen by the host.

This data contains information about the various listings on Airbnb in Boston, MA, and the various reviews that the listings have received. We want to look at the various factors and trends present in the Boston listings on Airbnb. We want to build a story on how the words used by the hosts in their descriptions of their listings has affected their final reviews, and how these have interacted with the prices that the guests are willing to pay.

1. Exploring the data

The columns present in the data provide us with the following information:

* Host information: host ID, host response time, total listings owned, host review scores.
* Listing information: number of bedrooms, bathrooms, amenities, size, price, neighbourhood, house rules, location, text descriptions of the house/apartment.
* Text reviews given by each guest.

While exploring the data we found some features (or columns) that needed a little amount of preprocessing and cleaning. To initially and partially clean the datasets we developed a script in R (“preproc.R”), where we turned the NA values in “beds” and “bathrooms” for their median value respectively, since there were only a few rows with NA values present. We also cleaned the “price” and “host response rate” variables, by turning it from character to numeric, after cleaning for non-numeric characters (like the “$” and the “%” symbols). We also created a column called “host\_period”, which is the number of days a host has been logged into AirBnb as host (we subtracted the “host\_since” date from the current date). We also realized that there were a significantly large amount of rows that had no guest scores, nor guest reviews (approx. 20% of the data). We decided not to do anything with these observations, but used a subset of the complete data when working with these features.

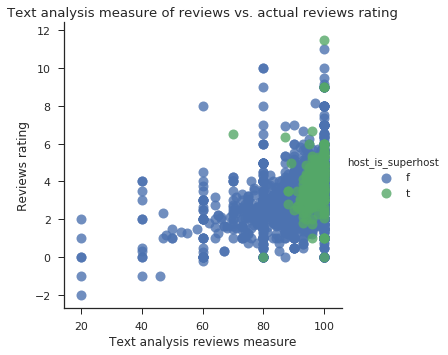
Finally, from the original 95 columns present in the listings data, we removed 54 of them because they mostly were unuseful for our analysis, and worked with the remaining 41 (we then created more variables from these columns and also the two other data files: reviews and calendar).

1. Text analysis of reviews, descriptions and comments

We performed some basic and explainable text analysis to the guest reviews, host descriptions of listings and host descriptions of the listings’ neighborhood, in order to come up with some relation between the text comments and the different features of the data.

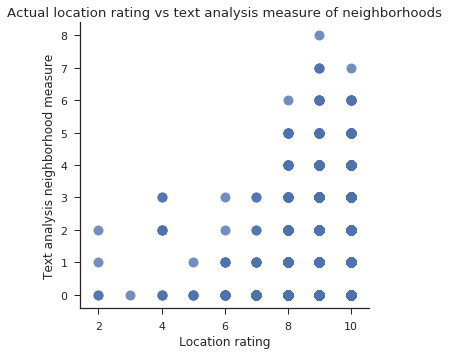
1. First we parsed every guest review (see Python code attached: reviews\_analysis), and removed bad characters such as points, commas, apostrophes, etc.
2. Then we separated each review into words and built a data structure that contained the frequency of each word in all reviews.
3. With that, we read the most frequent words (almost 1,000 of them), and manually categorized them as positive, negative or neutral.
4. After that we created three measures for each review by counting the number of positive, negative, and the difference (positive minus negative) words in each of them. Then we averaged this measure for each listing (in general each listing has more than one review). The columns created then for each listing were: “pos\_review”, “neg\_review”, and “dif\_review”.
5. The same process was repeated for the host’s neighborhood descriptions, and the host listings descriptions, with two differences: 1. there were no negative words found in these descriptions, so only one measured was built for each (“descr\_measure” and “neigh\_measure”), and 2. there was only one description per listing, so no average was taken.

Let us check if our measure of “positive minus negative” reviews correlates with the actual score given by customers:



Indeed, we can see that our text analysis for reviews correlates with the actual review rating; this is a first validation of our text analysis model. Also, we can see our first interesting feature coming out from the data: all hosts marked as “superhosts” are packed in a high review measure section, as defined by our text analysis model.

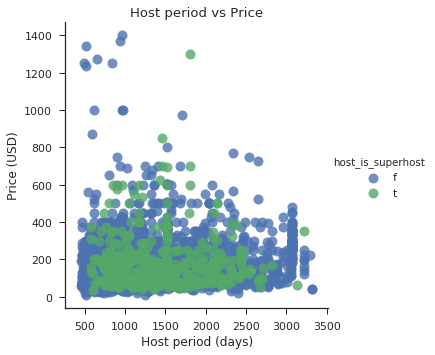
Let us use the same text analysis model to analyze host’s neighborhood descriptions. The following graph shows how it correlates with the actual “location rating” given by the guests:



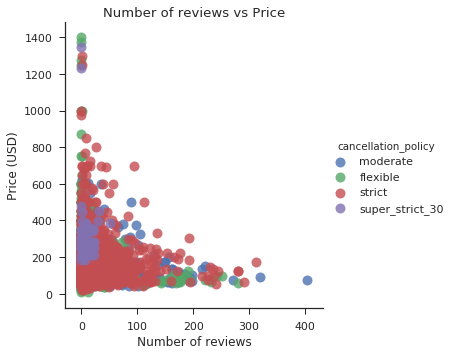
We can see that low location rating corresponds to a low measure in the neighborhood description, and also that “good” descriptions correspond to higher ratings. But, there are many cases in which a “bad” neighborhood descriptions correspond to high location rating, that means that if the location is good, customers put a good rating independently of the host’s neighborhood description.

1. Patterns found in the data: a storytelling

We searched for different patterns in the data by building simple models, and also analyzing the visual relationship between relevant features. We built scatter plots relating many different variables to each other (see the Python code attached: plots). This types of graphs were very enlightening in terms of relationships between different features of the data.



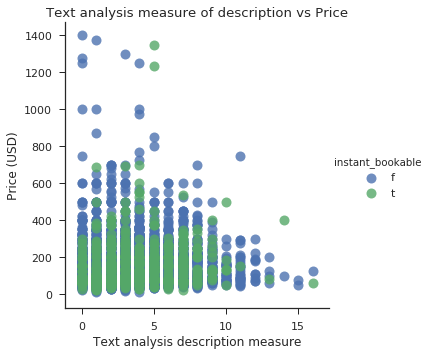
For example, we can see an interesting pattern when we analyze the pattern in the previous graph between ‘host\_period’ and ‘price’, colored and labeled by the variable “superhost” variables, defined by AirBnb. First of all, hosts marked as ‘superhosts’ by AirBnb don’t have listings with excessively high prices (except one) (we can draw this line around 900 USD). Also, we can clearly see in this graph how listings with high prices correspond only to hosts with less days registered in AirBnb. This might give us a hindsight that some of the newer hosts are still learning to adapt to how AirBnb works and how to get customers. Let’s see if more evidence supports our hypothesis:



In the last figure we graphed the number of reviews for each listing against the price. We can see again that the listings that have higher prices correspond to a very low number of reviews, and that could also mean less opportunities to learn from customers reviews (i.e. less experience in AirBnb). This supports our hypothesis that less experienced hosts tend to put higher prices to their apartment offers. Also, a high number of reviews only appear for listings with very low prices.

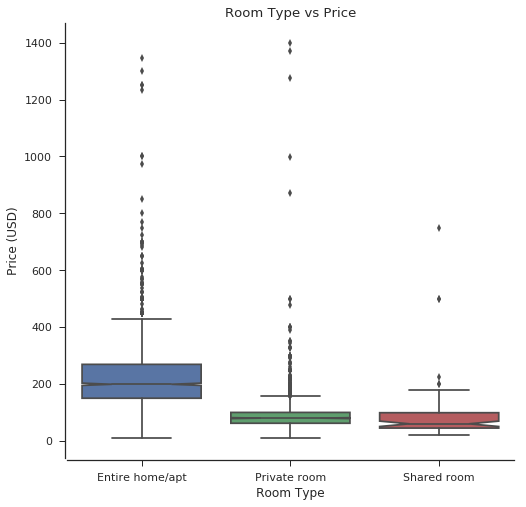
An interesting additional feature of the latter graph is that hosts with a “super strict” cancellation policy are all packed in a region with a very low number of reviews. This also supports our hypothesis; as hosts have more reviews, they learn which of the conditions they put work and which not.

Let us now apply our text analysis (see “Text analysis” section) to the host’s descriptions of their apartments. We find the following:



We see again that all listings with high prices tend to have a low description measure (the opposite is not true). Let’s remember that the description measure is a quantity that describes how many “positive” words are included in your description. Also, higher prices do not offer (except in two cases) instant booking of their apartments. All of these still supports our initial hypothesis: newer hosts who put higher prices are consistently bad at offering their apartments also with bad descriptions.

Let us now analyze a little what does indeed define a listing price. Let see first a graph between room type (a categorical variable) and price:

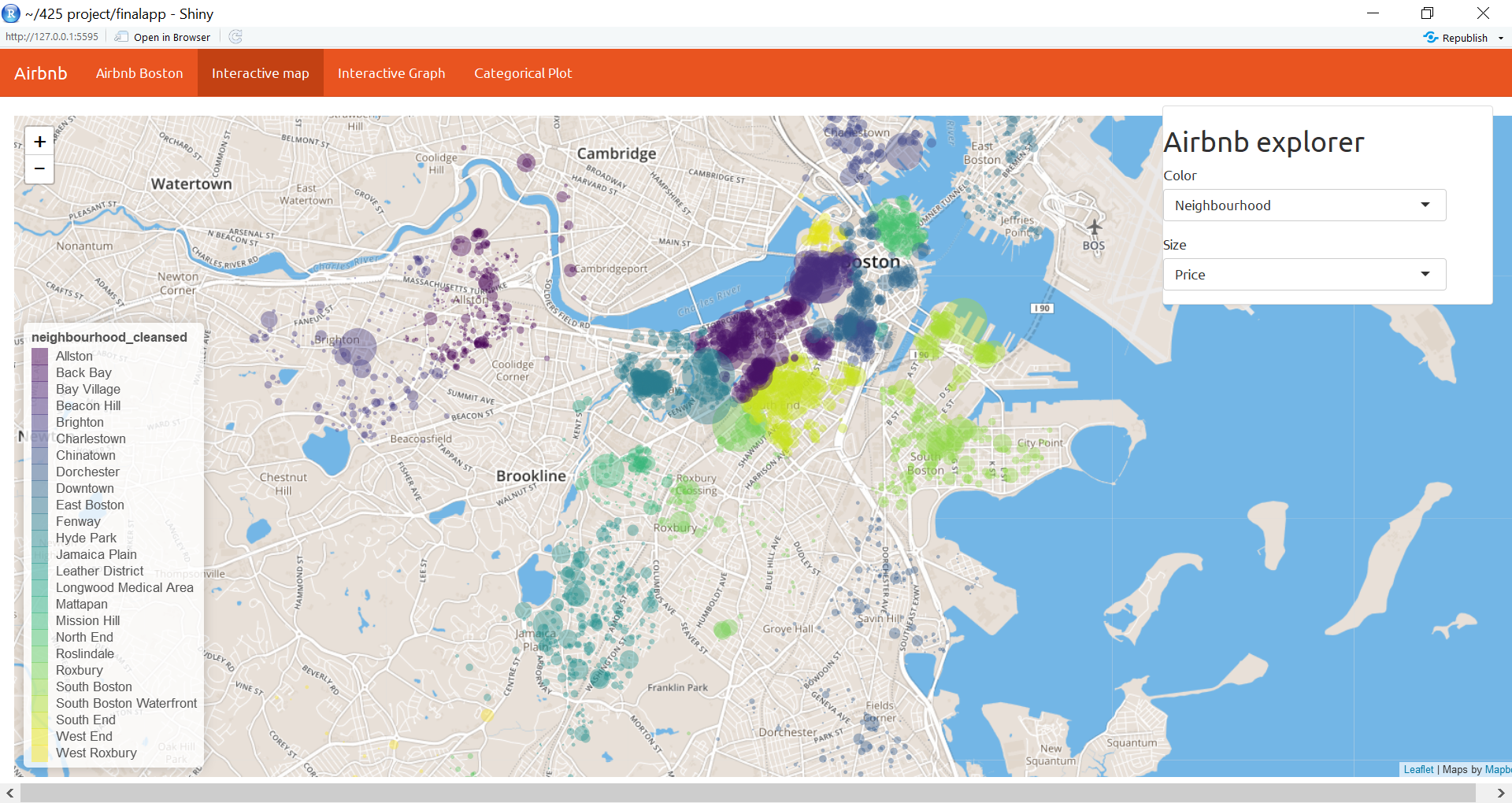


In the last figure we can see how the room type is a clear factor in the price. Not only “entire apartments” have a greater median price (and a lot more of outliers), but also “shared rooms” do not possess any outlier that we would indicate as a “high price”. All high prices correspond to “private rooms” and “entire apartments”, with a median price clearly greater for the latter.

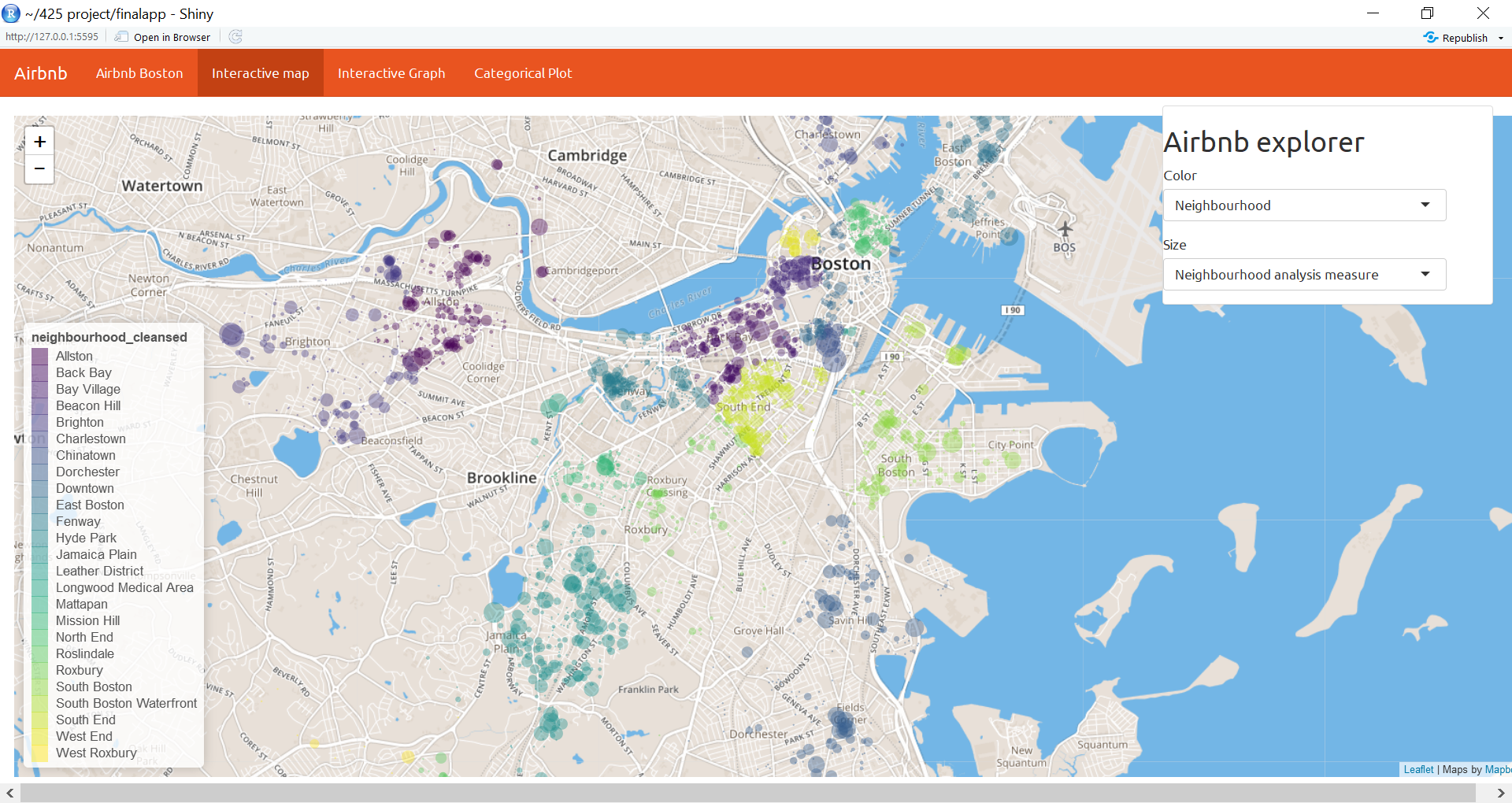
1. Building a user interface application

For the purposes of carry on analyzing this dataset and contain all the plots presented in this report, we built an interactive dashboard using R Shiny, with multiple tabs, each fulfilling different purposes:

* Introduction Tab: explanation about the data.
* Interactive Graphs, both scatter-plots and box-plot, to visualize the different features of the data. User can apply ranges to different variables and select which variables can be visualized in the “x” and “y” axes.
* Interactive Map with visualization of the listings with colors and sizes defined by the user.



* Looking at the price of houses across the city, coloured by neighbourhood, we can immediately spot the expensive neighbourhoods inside the city. The prices of Airbnb listings reduces significantly as we move out of the city towards the suburbs



* We can also represent ‘neighbourhood analysis measure’ using the size of circles. This is a measure of the sentiment of the words used in the description of the neighbourhood by the hosts. Doing so shows an interesting cluster of positively worded descriptions in the center of the city, where the high prices are located too.

1. Final remarks

In the present work we performed a data cleaning and analysis of different features of the data. First we created new variables by performing an explainable text analysis model on the guest reviews and host descriptions. We then found a correlation between these variables and the rest of the data, and also in the data itself. We found that listings with higher prices correspond to newer hosts (low “host\_period”), and also that they have less experience with AirBnb guests (as seen by their number of reviews). We also found that they are not doing a good job in describing their own listings, as defined by our text analysis description measure.

Apart from these, we built an interactive user interface using R Shiny in which all of these and many more features of the data can be explored. It also includes a tab with an interactive map to look for patterns against location and neighborhood.

1. Acknowledgements
2. <https://shiny.rstudio.com/gallery/movie-explorer.html>
3. <https://shiny.rstudio.com/gallery/superzip-example.html>

1. Source: <https://www.kaggle.com/airbnb/boston/data> [↑](#footnote-ref-0)