MA678 midterm project

Zike Tang

2021/11/9

Abstract

This project focuses on the listings information on Airbnb, trying to figure out the relationship between the price of listings and other information of hosts and listings. In order to get a understanding of the data sets, I tidied the data, drew various plots to describe it both numerically and graphically and made some conclusions on the distribution of review scores, types of listings on the Boston map and etc. As the the price of one listing is related to its nearby listings, I smoothed it on the spatial level With spatial kriging method to get a better continuity of listings price for further modeling. Multilevel models are created including predictors that have different trends among tracts and others related to the price of listings. Conclusion is that listings with licenses permission, owed by hosts with good services quality have higher price on average when being compared with different tracts in Boston.

Introduction

Airbnb is an American technological company that provides an online marketplace for lodging worldwide, mainly homestay for vacations rentals, and tourism activities. Listings price varies from listing to listing and it is one of the main factors that customers will consider when they decide whether or not to rent a house on Airbnb. Factors that may influence the price of listings involve both from the information of the listings such as the location, the room type and the furniture conditions, and the information about hosts from response efficiency to whether or not they are super-hosts and etc. This project aims to figure out how these factors interact with the listing price.

Plus, common descriptive statistic of price of listings does not incorporate its spatial characteristics. In this case, the price of listing of one tract partly influences the price of listing of tracts nearby. Smoothed listing price will be obtained using kriging which will be explained after.

When fitting the models, there are different trends for Boston tracts when considering the relationships between different predictors, which leads to the application of multilevel modeling. In this case, both the variances between different listings and different tracts are considered.

Method

EDA

Two main data sets were used in this project: one contains the listings and hosts information and another involves the listing reviews information from the content of reviews and the date. Data is downloaded from Airbnb website and it was compiled in October 2021. The the spatial information of Boston tracts of 2020 version is also used for presenting distribution on Boston maps. This is obtained from boston government. Table 1 clarifies the name and meanings of predictors that are included in the modeling and details are shown in the model part. Figure 1 shows the distribution of price of listings and its ranks for Boston neighborhoods. West end has the highest listings price on average and most other Boston neighborhoods has large range of listings price. Listings price get highest in the central downtown part where locate most luxury shops and have more people living.

Predictors names	Attribute	Explanation
price	NUMBER	The price of listings
host_response_rate	NUMBER	Proportion of request hosts giving responses
host_response_time	FACTOR	Time taken to give response to customers
host_is_superhost	FACTOR	Whether host is given a superhost title
room_type	FACTOR	The types of listing
license_ornot	FACTOR	Whether the listing have license permitted
GEOID20	FACTOR	Boston tracts

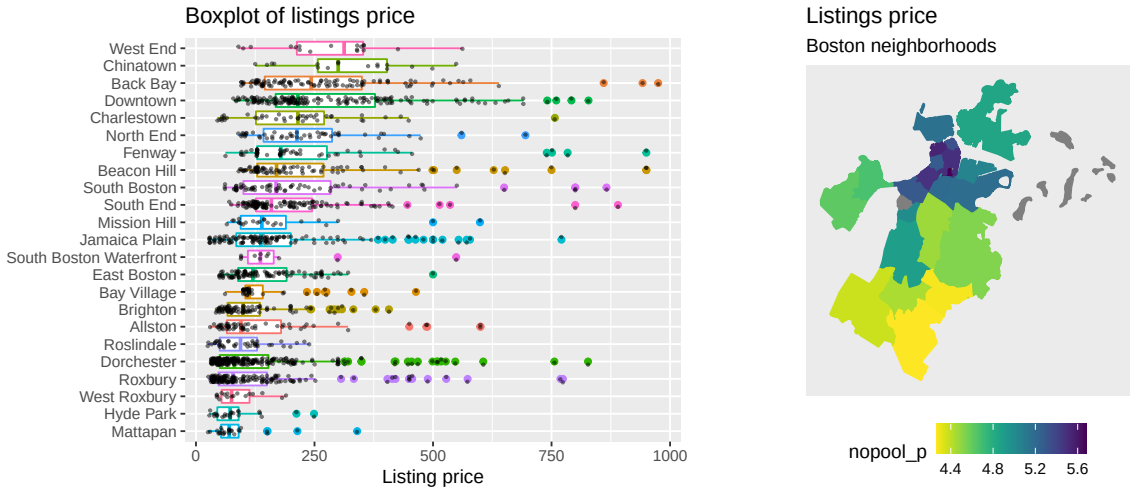


Figure 1: Boxplot and map plot of listings price

Spatial kriging and variogram

Kriging is a process of spatial interpolation given a set of observations while variogram describe of the spatial continuity of the data. Figure 2 shows the spatial relationship of listing price in logarithmic form under the spherical assumptions which is a method commonly being used in testing continuity for spatial data. When distance goes farther, the variogram becomes larger and levels out at a distance point called range. In this project, the listing prices of two nearby tracts are dependent and the prices of listing are nearly independent for those tracts far away from each other. Therefore, spatial distribution of listings does have an impact in further prediction and modeling. Figure 3 shows the comparison before and after kriging on map. The variances of listing price get smaller after kriging.

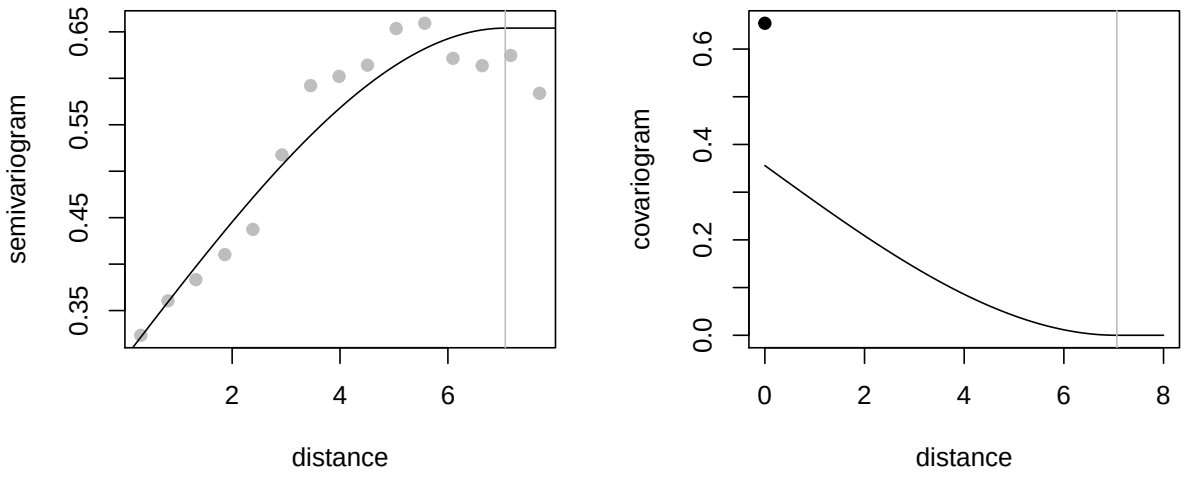


Figure 2: Variogram

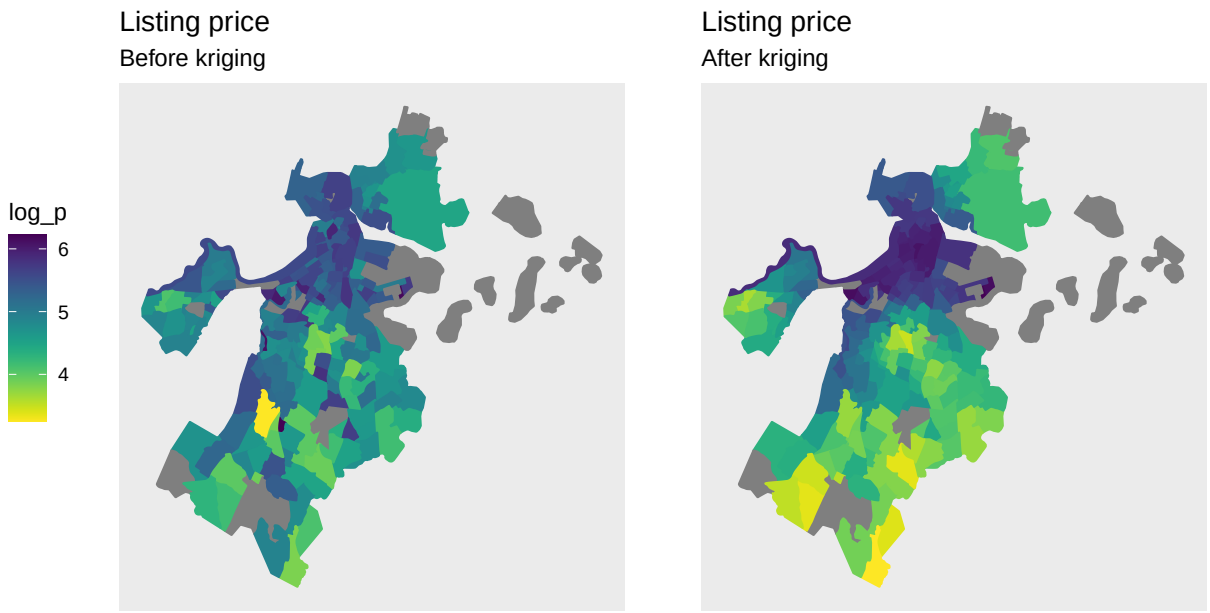


Figure 3: Kriging comparison

Models Fitting

At the beginning, two models are built: one only includes the various intercept for tracts and another includes all the related predictors, such as whether the hosts have profile picture on website, whether their identities are verified, the review scores of listings and the number of total listings owned by each host which is not listed in the above table.

They are excluded because they are not significant. The picture and identity of hosts might not influence customer's decision on renting house. We may consider that the listings with good comments have higher price. The difference of review scores is small among different tracts in this case, which means listings are nice in Boston overall. Moreover, Customers usually do not care about the number of listings owed by the hosts.

After drawing the relationships of other predictors among different tracts(shown in appendix), host response rate and response time are included to be analyzed their various slope effect. Intercepts among tracts are not considered because there is little difference after analysis. Room types, whether host is a superhost and whether the listing has license permitted are kept in simple form in the model. The listings price is taken the logarithmic form to better present the result.

Finally, the final model is defined as below:

```
lmer(log(price)~
       room_type+
       host_is_superhost+
       license_ornot+
       (host_response_rate + host_response_time -1| GEOID20)
```

Result and assessment

The hotel room and the license are positive with listing price while the other predictors are negative with listing price. For the fixed coefficients, intercept, room types, especially the private and shared types, and whether the listing has a license are significant. Whether host is superhost is not that significant comparatively but still has relationship with the listing price. Results are shown in table 2.

Predictors names	Estimate	Std. Error	t.value
intercept	4.90130	0.03431	143.324735
room_type Private room	-0.84347	0.02754	-30.711543
room_type Hotel room	0.19068	0.13268	1.441391
room_type Shared room	-1.38750	0.32233	-4.313799
host_is_superhost	-0.08684	0.02672	-3.260852
license_ornot1	0.46292	0.02976	15.583953

Figure 4 shows the normal quantile for the random coefficients. Simulated Values that have a significant confidence interval which does not overlap zero are highlighted in black. Although the normal quantile for host response rate is not significant, the model still keeps the host response rate since it makes the whole model predict well.

Discussion

Compared to the price of the entire home or apartment rooms, the price of shared rooms are lower on average. Hotels have the highest price on overage and it makes sense because hotels are often rent for shorter periods

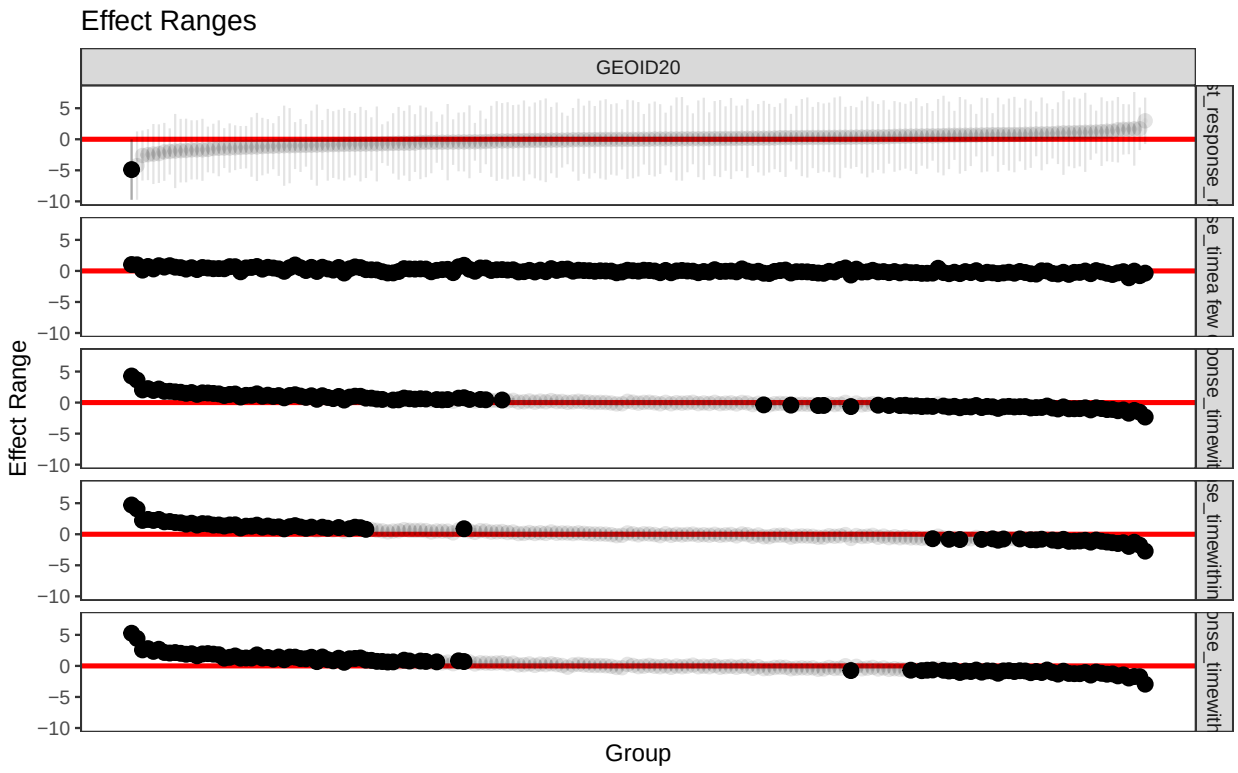


Figure 4: Normal quantile for the random coefficients

compared to other types of rooms. Services such as morning call, breakfast and amusement park in the hotels which are not usually offered by other types of room also add value to the price.

Another interesting point is that if the host is a superhost, the price of listings is a little lower than those price of houses owed by normal hosts. I suppose that hosts get satisfied with the superhost identification, which might encourage hosts to provide discounts or better services for customers. It is not surprised that listings with license permitted have higher price on average. Listings with permissions will guarantee services and refunds, which increases the underlying operation costs.

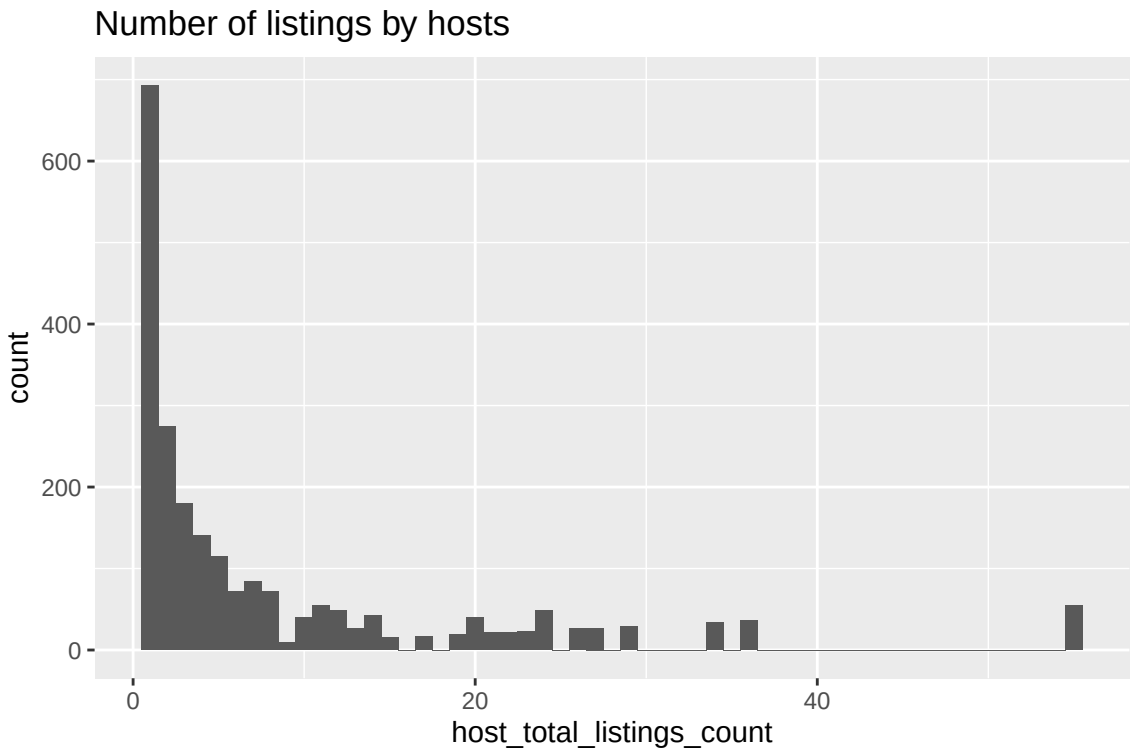
There are four types of host response time in this case: within an hour, within a few hours, within a day and within a few days or more. When looking at the random effects, variance of long response is small comparatively, which means there are little differences of price change among tracts for those listings with hosts giving slow responses. This can be explained that for listings with hosts who do not give timely response, their price change behaves similarly and contact services of hosts do mainly influence their listings price. Therefore, it is suggested that hosts should pay more attention to give instant feedback for customers or they will lose them in a short period.

In conclusion, type of listings, hosts information and their services have relationship with the price of listings. Hosts are advised to improve their services quality to better serve the customers.

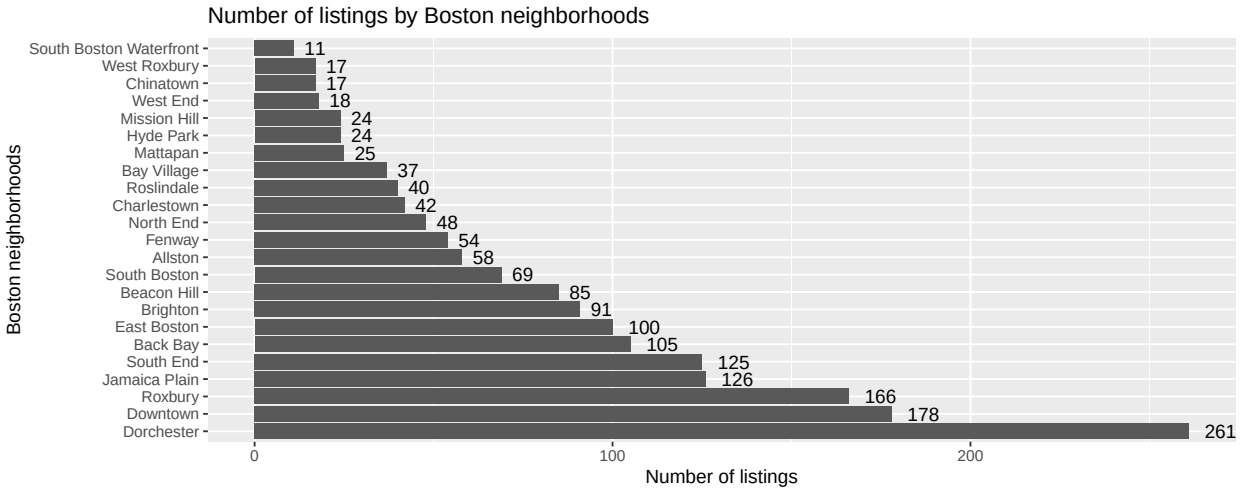
Appendix

Additional plots

- Most hosts have less than five listings.

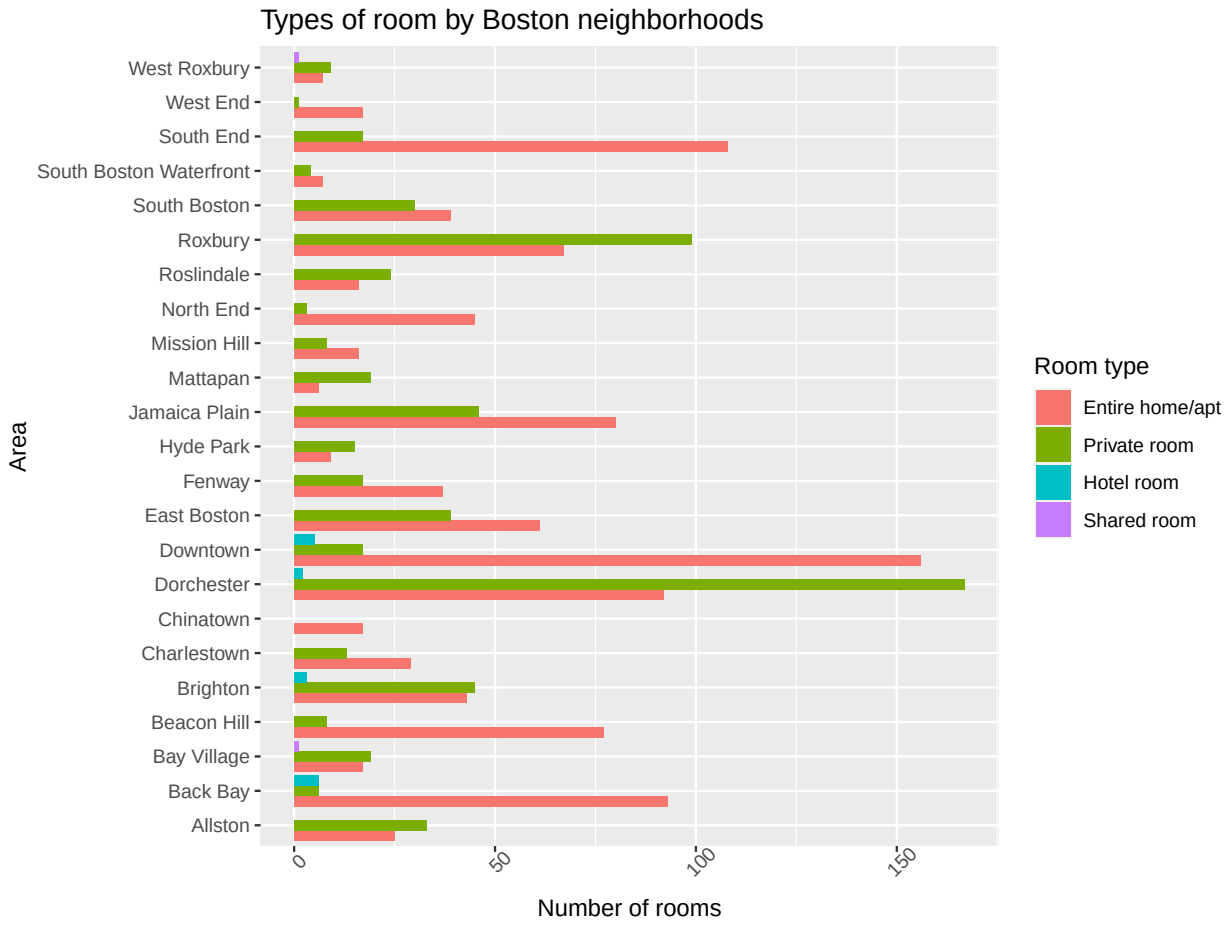


- Dochester, Downtown and Roxbury rank first three of the number of listings. Dochester ranks first because it is a large area having the opportunity to cover more lisitngs. Downtown is the place where most people live in.

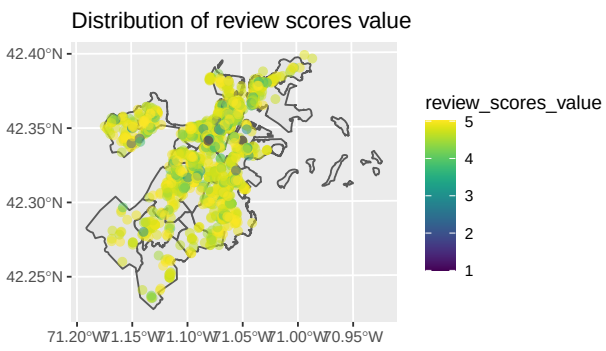


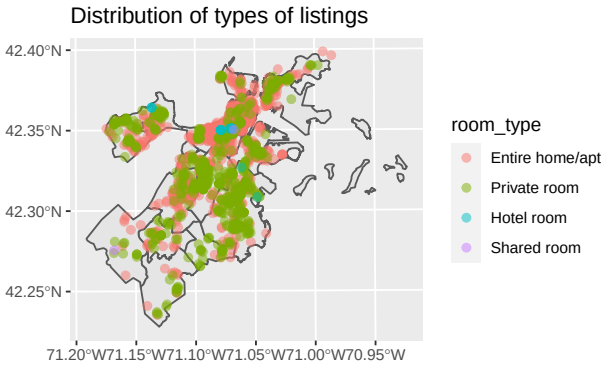
- Almost all the neighborhoods have their peak around and above the mean review scores value.

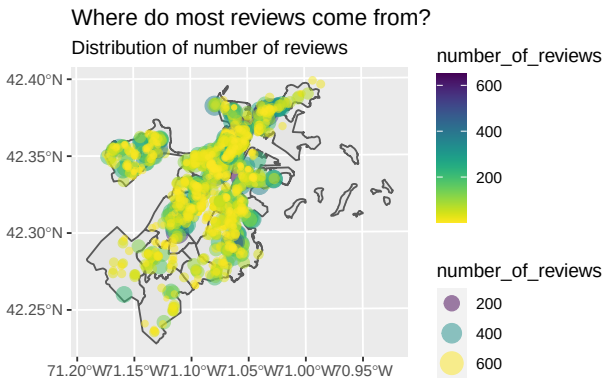
- Entire home or apartments and private rooms are the main types of room averall.
- Downtown has the most number of entire home or apartments and Dorchester has the most private rooms.



- Point map of various predictors







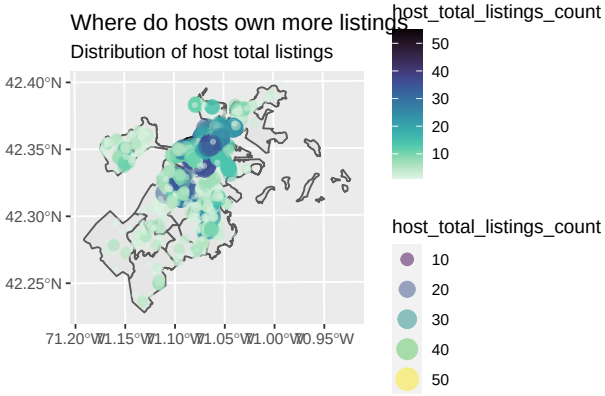


Figure 5: Most positive words in reviews

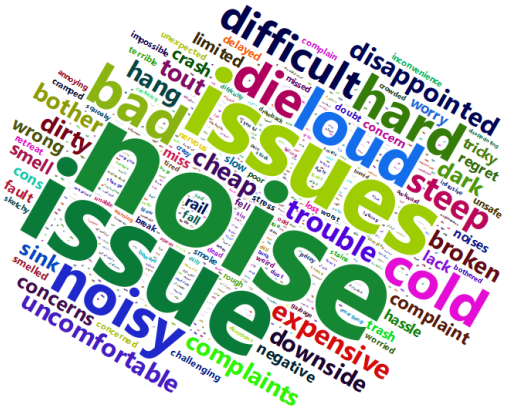
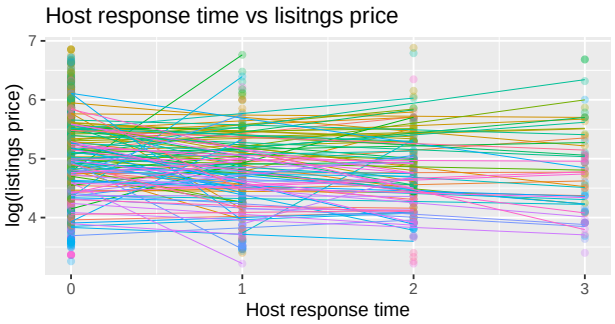
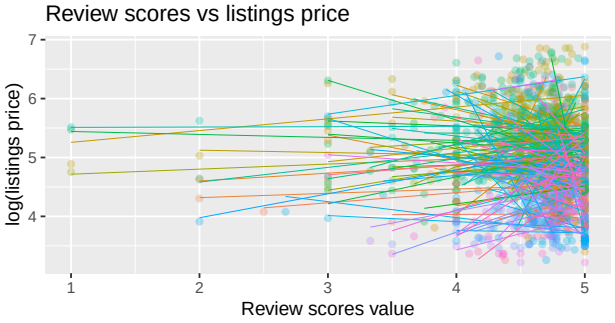
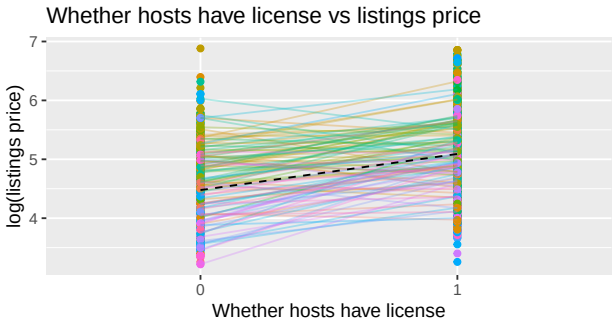


Figure 6: Most negative words in reviews

- Plots between predictors to see if there is a different pattern among different tracts







Citation

- Mapping: https://map-rfun.library.duke.edu/032_thematic_mapping_geom_sf.html
- Text mining: https://www.tidytextmining.com/