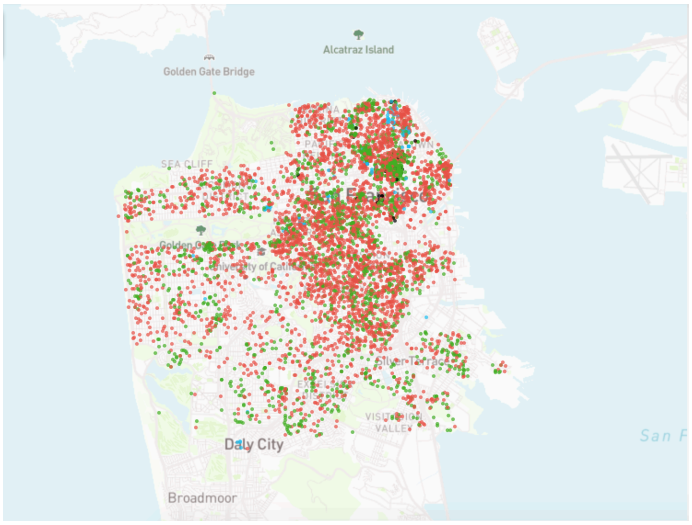
1 Introduction

What is the effect of location and environmental features on Airbnb rental prices? Airbnb is an online platform where people can list and book their accommodations all around the world. This concept proposed by this site meets a real demand: it is to link individuals who want to rent their homes to supplement their ends with travelers seeking accommodation at lower prices. It was founded by Nathan Blecharczyk, Brian Chesky and Joe Gebbia in 2008 in California and as a start-up in Silicon Valley it has attracted over 4 million users in 192 countries worldwide over the years. There has been extensive research done investigating the effect of environmental and location factors on housing, hotels, and rental building listings. Unlike hotels and rental units, Airbnbs are directly priced by the host and therefore are evaluated for price through a different process than the traditional cumulative hotel service or rental building. As a shared stakeholder economic system, although there has been clear evidence that Airbnb creates more economic opportunities, a decade of systematic literature review has shown an ever increasing conflicted host and booking provider relationship and mixed results meeting proportional thresholds for value and pricings of listings. [1] As part of Airbnb's booking system, it tacs on a 3% charge on the side of the host's earnings and a 14% charge on the side of the guest's purchase which are both incorporated into the final price listing at the outset. In this context, we wanted to assess the pricing trend followed by an aggregated, decentralized rental service like Airbnb, its variable effects and analyze any anomalies unique to listings that affect Airbnb rental prices. By using machine learning methods and setting this as a multivariate regression, we aim to see what the predictors for price are within a specific region of listings.

2 Data

Our data comes from Inside Airbnb which is described as " a mission-driven activist project with the objective to: Provide data that quantifies the impact of short-term rentals on housing and

residential communities; and also provides a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals" [2]. Inside Airbnb is a repository of data on Airbnb listings in various cities and countries that is scraped regularly using open source web scraping tools to collect data from the Airbnb platform API on a monthly basis.



In this study, we specifically look into Inside Airbnb data in San Francisco, California compiled on December 4, 2021, to predict the price of new Airbnb properties in that area. California is known for its highly competitive housing market and this is exemplified no better than San Francisco where if we are able to predict Airbnb housing prices we may be able to suggest giving host and user alike a more informed choice in ever increasing related housing and hotel markets. Our San Francisco detailed listings file provided data with 74 covariates mostly concerning host, unit, and review characteristics for each listing. Based on website summary statistics, there were a total of 6,413 listings, of which 63.7% were entire homes or apartments(red) and 33.2% were private rooms(green), with an average price of $246 dollars per listing, and 131 nights of booking in a given calendar year.[2] Additionally, about 61.3% were multi-listings belonging to hosts with multiple listings in comparison to 38.7% which were only single listings.[2] From this preliminary outlook, we can surmise that this has been a popular region for Airbnb as a primary source of income for hosts as the proportion of multi-unit hosts is higher and there are more fully converted units as bookable in comparison to private rooms or

hybrid units hence hosts own more full housing units and are more willing to list multiple units. In addition to this, we directly observe from the geoplot that location is an indicator for the concentration of listings as neighborhoods that are generally farther inward and northeast have more concentrated listings in their proximity in contrast to listings that are more west or south.

In cleaning our data in order to run initial regressions and the lasso analysis with cross validation, first all the variables relating to web scraping ids, urls and descriptions were removed. Personal descriptions, names and id's for the hosts were also removed as we were only concerned with the listings themselves and also did not have the means to conduct textual analysis to see any effects. Two variables, amenities and verification, which were columns of the list of tags for amenities and host verifications for each observation, were also parsed and extracted. Each verification was then added as an individual column of 0 or 1 of the listings containing the host verification and the original verification column was removed. For amenities, since there about 890 amenity tags, some repeating similar features and others self described and highly specific and unique amenities, we collected the 54 most common amenities from a 2020 Inside Airbnb analysis and created additional column variables based on each amenity in the same manner.[3] We also converted all true, false variables Repeated variables and empty columns possibly from incomplete web scraping were also removed. Lastly, all NA's were omitted. In summary, this reduced our total number of observations from 6,413 to 3,192 with 50 of the original variables retained and 54 individual amenities and 16 individual verification variables added to make a total of 120 potential covariates. From analysis, the wide majority of observations that were omitted from NA's came from lack of review variables data. This could mean that those listings were either never booked or brand new in the region to where they did not have review variable data. Given that these are new or unused listings, it is assumed that their

exclusion will not take away much from the data. With this cleaned data set, we are limited in our inference and conclusions by the NA's which were removed and other variables which remove a portion of representative observations within the population of listings provided our data. Additionally, this set of Airbnb observations are only representative of the listings that were scrapped on December 4, 2021 from the Airbnb platform through Inside Airbnb's algorithm. This does not necessarily represent all Airbnb units in San Francisco, only those picked up by the scrape and from that only the subset that contained sufficient review data and no NA's. Hence we are limited by these concentric modifications but given a large number of observable listings and spread throughout the city, we assume this subset to be adequate for representative analysis. Based on all these preliminary findings, the major variables which we selected for the analysis were the factor variables host_neighbourhood, bathrooms_text, and property type and bedrooms and min_nights_avg_atm. These variables all have implied causative associations as there is a physical necessity of cost for a unit based on the number of beds, baths, property type, local neighborhood and the average minimum booking in the last calendar year to set prices. Analyzing each variable with boxplots we found that on average there were about 60 listings per neighborhood with an even percentile range, there were less than 2 bedrooms for the wide majority of listings with only 5 outliers that had more than 4 bedrooms, the minimum average nights bookable per year were less than a 100 at about 30 nights with a handful of outliers, and property types were evenly distributed with a median value a little greater than 10 listings per the 44 different properties offered. Bathrooms were textually described and self entered so there was greater range and only a small range from 5 to under ten where most bathroom categories fit.

From these findings, we can see that listings were distributed fairly through each neighborhood and each property type and that on average bookings were usually within 30 day intervals and the wide majority of listings had 2 or less beds.

In further analysis, we ran a regression on reviews scores listing rating in each neighborhood given price as some interesting correlations we wanted to discover were how review scores listing rating correlate with price in each neighborhood. On average, there was a positive effect of 58 dollar increase for 1 point increase on review scores listing rating but the graphs differed depending on each neighborhood and for some there was a general pattern of a negative correlative relationship between price and review score listing rating. This could possibly indicate that the perceived value as seen through the rating of the listing was relative and irrespective of price so in indifference to the extravagance of a listing, a guest could have scored a lower cost listing, higher for it's lower price as a result of indifference to the additional qualities of the listing which can increase its price. Furthermore, we ran a regression on the total listings of the host of a given listing and the years the host has been active as an experienced and multi-listing host may have a different process of pricing than other hosts. In this, again we saw an increase in 4.9 and 7.9 dollars for a 1 unit increase in years the host was active and total listings owned by the listing's host. Again, among the panel regression graphs per property type we see many of the graphs are flat and subsets of them are radically high or low. In constructing our lasso model for predicting price, this is a decently clear indicator that covariates relating to host characteristics and guest reviews substantially differ based on the physical and environmental characteristics of the listing and that creating interactions from all variables on the physical and environmental characteristics like neighborhood and property type would serve as a better predictor of price while utilizing the data relating to host characteristics and guest reviews.

3 Methodology

In predicting prices for Airbnbs in San Francisco based on our covariates we will be using a lasso multivariate regression with 120 variables interacting on neighborhood and property type. This regression will be a regularized linear regression model that penalizes the L1 norm of weights that pushes the weights to zero, thereby constraining the model to be simpler over numerous potential coefficients of varying complexity. Additionally, we will use cross fold validation with 10 folds to verify the model and evaluate it with an in sample and out of sample coefficient of determination which will show the amount of accounted for deviations within the data by the model. In using this model, we are assuming a parametric characterization of the data which means that each covariate in totality serves as a scaled linear combination to produce a predicted output. Underlying this, we are also assuming that the data takes a linear relationship. Additionally, in order for the model to be representative there is also a need to assume that the data which it is trained on will be representative of the population, randomly selected and low in bias. An advantage of this model is a reduction of the effect of collinearity among multiple variables in the model, especially among multi-correlating variables which we assumed would be present specifically in the review scores which are reflective or preexisting characteristics of the listings.

As a result of the implied causative relationship between the physical characteristics of a listed Airbnb unit and its price, there is plausibility to the parametric and linear application of the model onto the data as there is direct evidence of an increase in one leading to an increase in another. In terms of other covariates, the level of plausibility is unclear as there is no evidence to substantiate a parametric and linear application onto the other factors which are being supported to predict the price of the listing. In addition, concerning the modification and acquisition of the

data there it is also not necessarily a conclusive evaluation of the region and its units. We can say, however, that our findings are plausible to the extent of the particular modification and subsetting of data which we have established, irrespective of any other external application.

4 Main Results

We ran two lasso multivariate regressions with cross fold validation: one on all variables with no interactions and the other on all variables with interactions on neighborhood and property type. On the first lasso, latitude, property type-private room in resort, bathroom_text-5 baths, and El Granada neighborhood were the highest increasing covariates at 1021, 290, 201 and 193 dollar increases with a unit increase in each, and Newport beach, Kailua/Kona, Diamond Heights, and Santa Cruz neighborhoods as the highest decreasing covariates at 104, 103, 100 and 74 dollar decrease for a unit increase in each. For this model our in sample R^2 was about 0.615 and our out of sample R^2 was about 0.661. In our second improved model which created interactions with neighborhood and property type with all other variables, we found that our highest increasing covariates were property_type-private room in resort, property_type-entire place, latitude, property_type-room in serviced apartment and property_type-floor at 625, 341, 85, 83 and 58 dollar increases in price for a unit increase in each and property_type-Entire residential home:bathrooms_text-1 bath, bathrooms_text-1 bath, property_typePrivate room in residential home:bathrooms_text1 shared bath, and property_type-Entire rental unit:bathrooms_text1 bath as the highest decreasing covariates at about 3, 2, 1, and 0.14 dollar decrease for a unit increase in each. For this improved model our in sample R^2 was 0.982 and out of sample R^2 was 0.983. Based on these betas, there are definitely some outlier factor variables which we did not account for within property types as the drastic increases in price by well over the average listing price is a result of these single factor levels on property. We have determined in our second model with

the given interaction that a lower number of bathrooms can decently predict a lower cost irrespective of the neighborhood or property type. In both regressions, latitude remained among the top increasing price covariates which is a result which we expected and confirmed our initial inferences that prices for listings would increase the further north that a listing was located in the city.

5 Further Discussion and Conclusion

Given the numerous variables which were provided in the data, we may have been overly ambitious with including the majority of them in our model and adding extra variables for individual amenities and host verifications which did not produce any unique impact on predicting prices. In future analysis, we might consider using these amenities frequencies to predict the review value score which a guest would give. Furthermore, with added survey data on whether or not a potential guest would book a listing based on these factors could be added to the data set to predict the likelihood of an Airbnb listing being booked given these factors. We found the most difficult part of the process to be cleaning and extracting the data so that the algorithms would function properly.

References:

1.) Hati, Sri Rahayu Hijrah, et al. “A Decade of Systematic Literature Review on Airbnb: The Sharing Economy from a Multiple Stakeholder Perspective.” Heliyon, vol. 7, no. 10, Oct. 2021, p. e08222. DOI.org (Crossref), https://doi.org/10.1016/j.heliyon.2021.e08222.

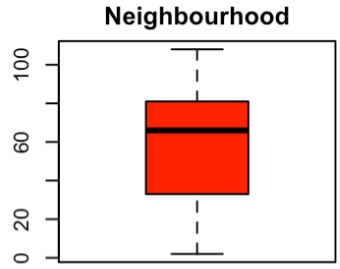
2.) “Inside Airbnb. Adding Data to the Debate.” Inside Airbnb, http://insideairbnb.com. Accessed 18 Mar. 2022.

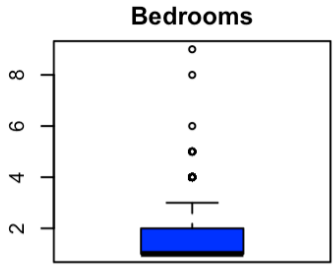
3.) “Amenities Offered in Airbnb Listings.” BnB Facts, 9 July 2020, https://bnbfacts.com/amenities-offered-in-airbnb-listings/.

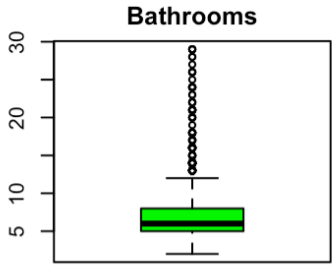
4.) Taddy, Matt. Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions. New York: McGraw-Hill Education, 2019.

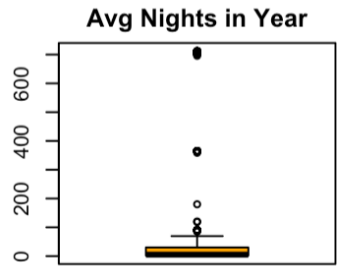
Graphs and Figures (also included in R)

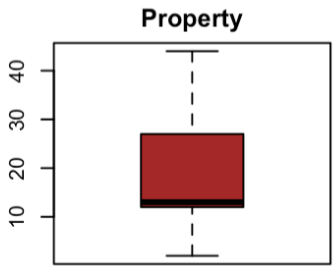
1. Boxplot of Main variables



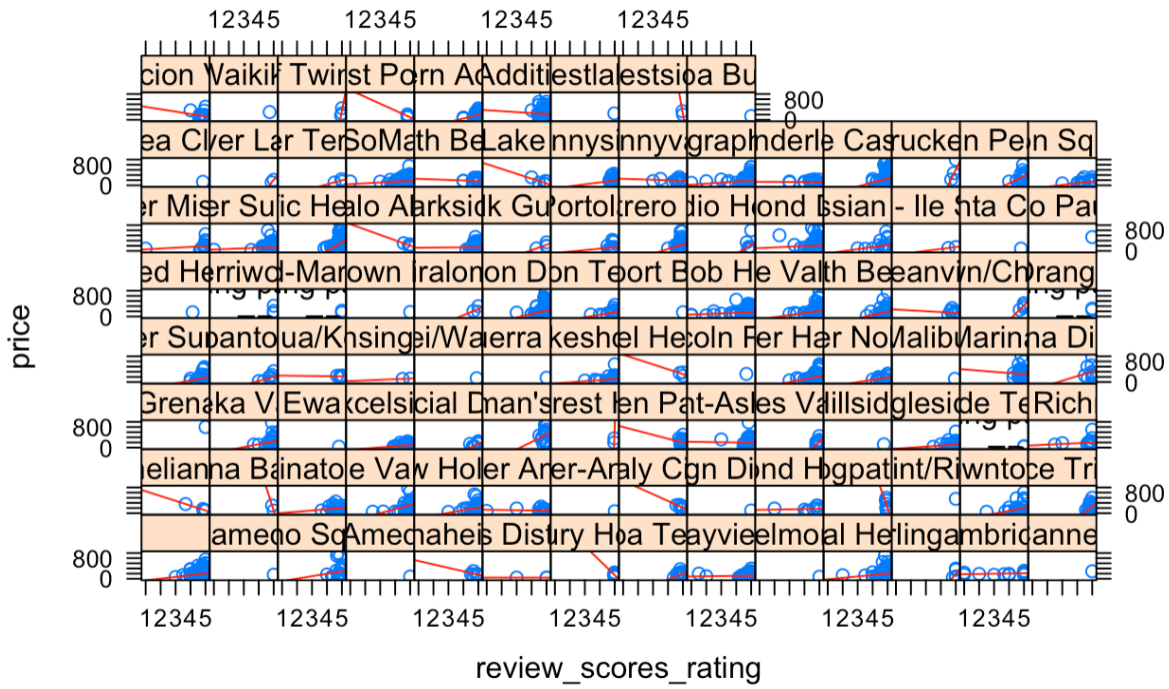




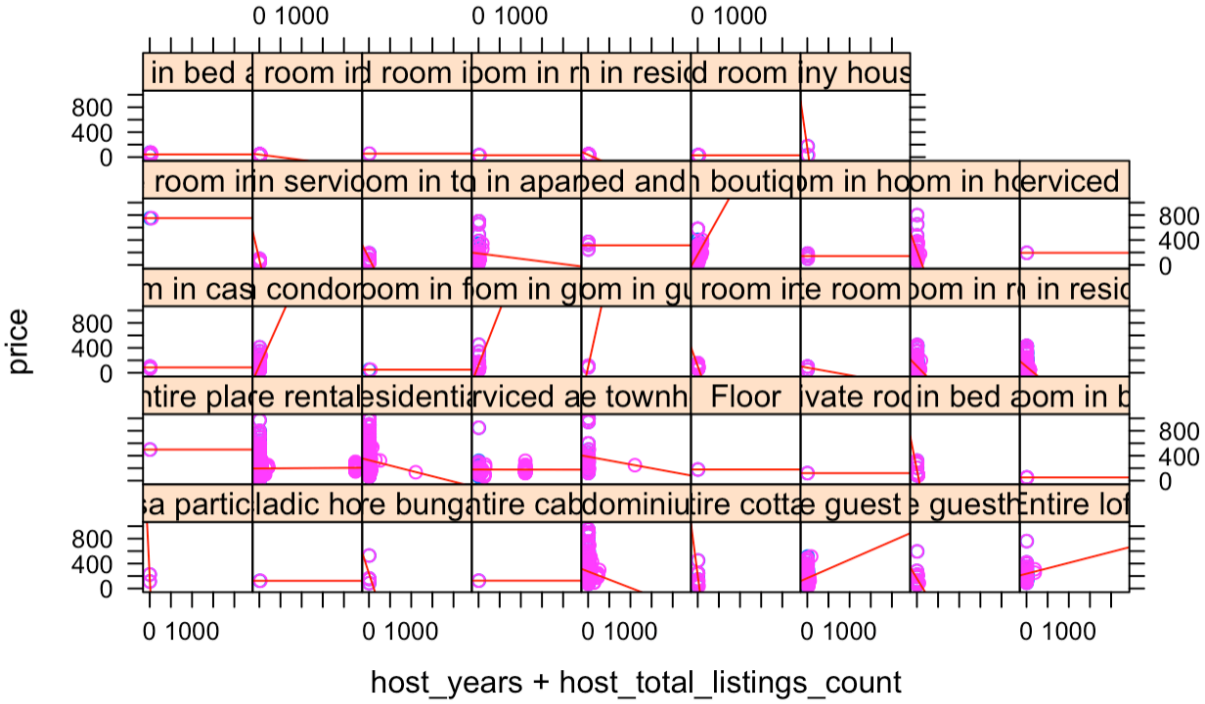




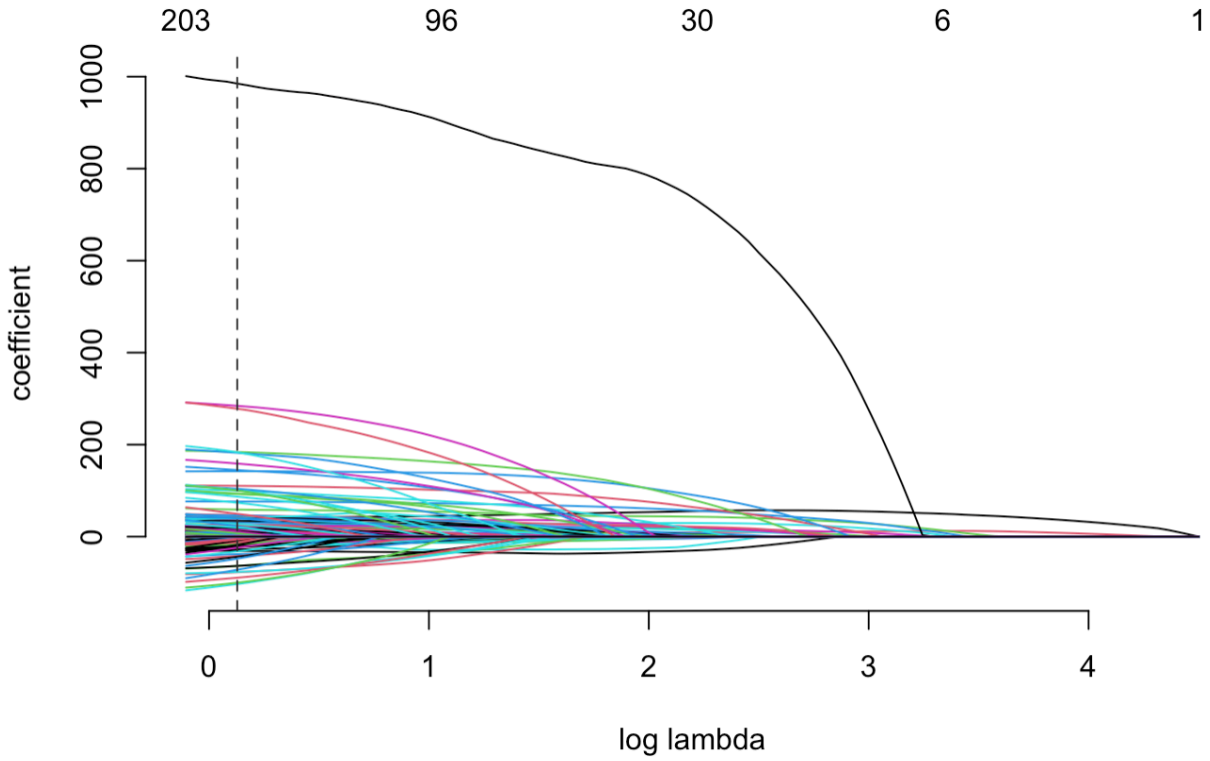
2. Multi-Panel plot of review scores rating and price for different neighborhoods



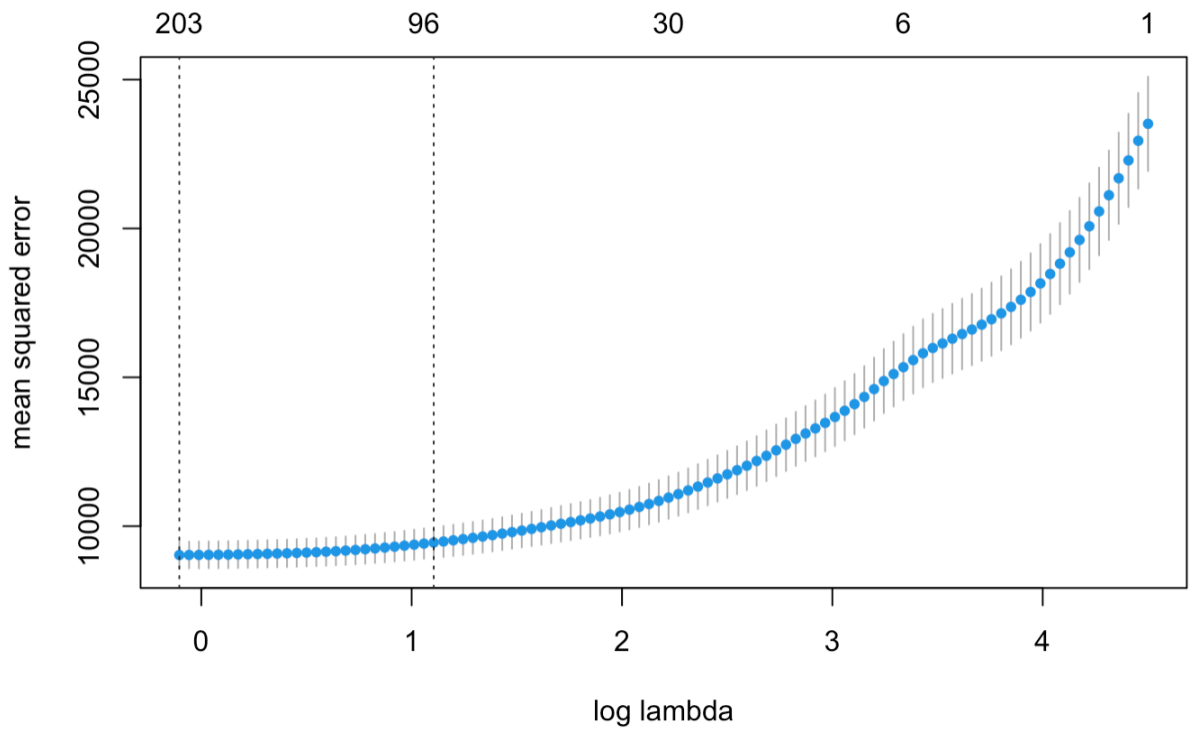
3.Multi-Panel plot of host years+total listings and price for different neighborhoods



4. Coefficient graph for first lasso model



5. Regularization path for 10 fold CV on first model



6. Regularization path on 10 fold CV for interacted model