Math 218: Final Report

FYP

Javier Merino and Mauricio Moreno

December 8 2022

1.Introduction

1.1 Data

The dataset we are using is Airbnb listings data from Inside Airbnb with variables on info that is shared in a room listing; detailed description of listings dataset below. The data was collected via publicly available data from Airbnb itself. One limitation so far is that the date represented by the data for listing prices is from a single day so we do not have time-series data for analysis. All the data collected was scraped on Sept. 22, 2022, so this is hopefully a good snapshot as to the trends we will explore between listing/host factors and listing price. Another limitation that we are noticing is that cleaning fees are not included in the data which many time is a determining factor in booking rates.

Data Source: Get the Data. (2022). Insideairbnb.com. http://insideairbnb.com/get-the-data

1.2 Response variable of interest

Price is our response variable of interest from the listing data. In this case, price is measured as how many pesos ($MXN) per night per listing. Because our analysis focuses more on the point of view from an Airbnb host, we are interested in what factors are associated with price increase/decrease in a particular market (Mexico City).

1.3 Research Question

"What are the best predictors of price for Airbnb listings in Mexico City?"

1.4 Supervised Learning Methods

```
-Best Subset Selection
-Forward Stepwise Selection
-Backwards Stepwise Selection
-Lasso
```

1.5 Packages

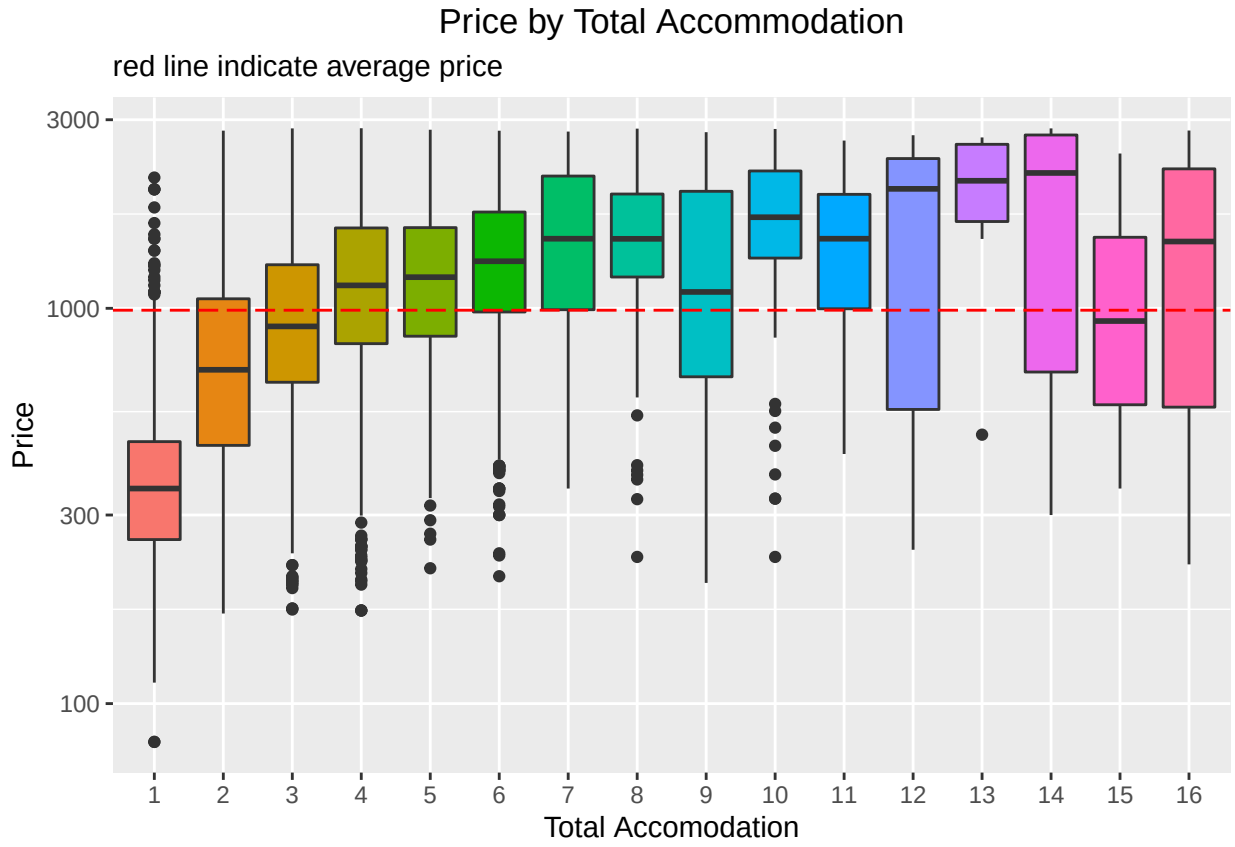
library(tidyverse) library(glmnet) library(fastDummies) library(geojsonio) library(rjson) library(leaflet) library(remotes) library(readr) library(sf) library(leaflet.extras) library(dbscan) library(geosphere) library(magrittr) library(rgdal)

2.EDA

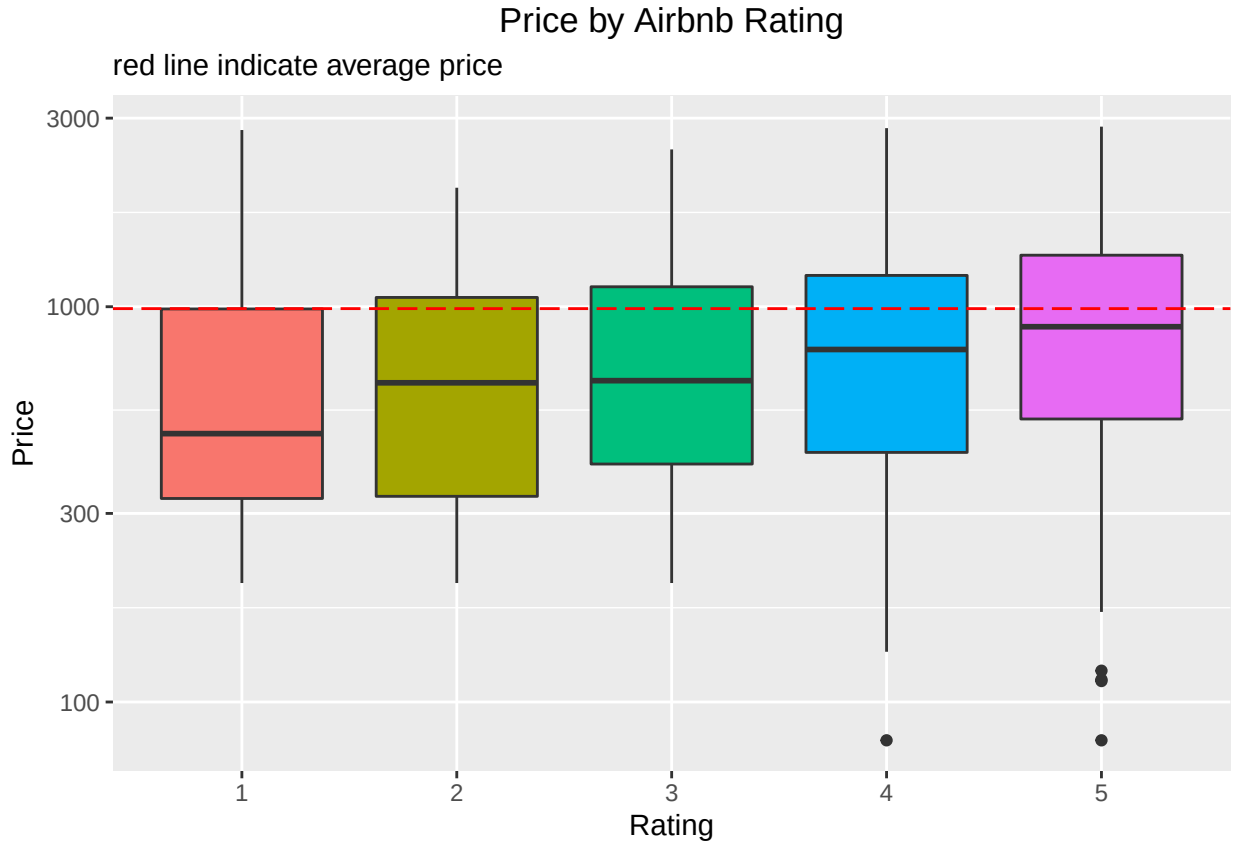
2.1 Box plots

Boxplots were generated for each of three variables that we thought would be important - the type of room, host rating, and total accommodation size. All boxplots revealed expected trends: increased accommodation -> increased listing price, increased host rating -> increased listing price, more private listings -> increased listing price.

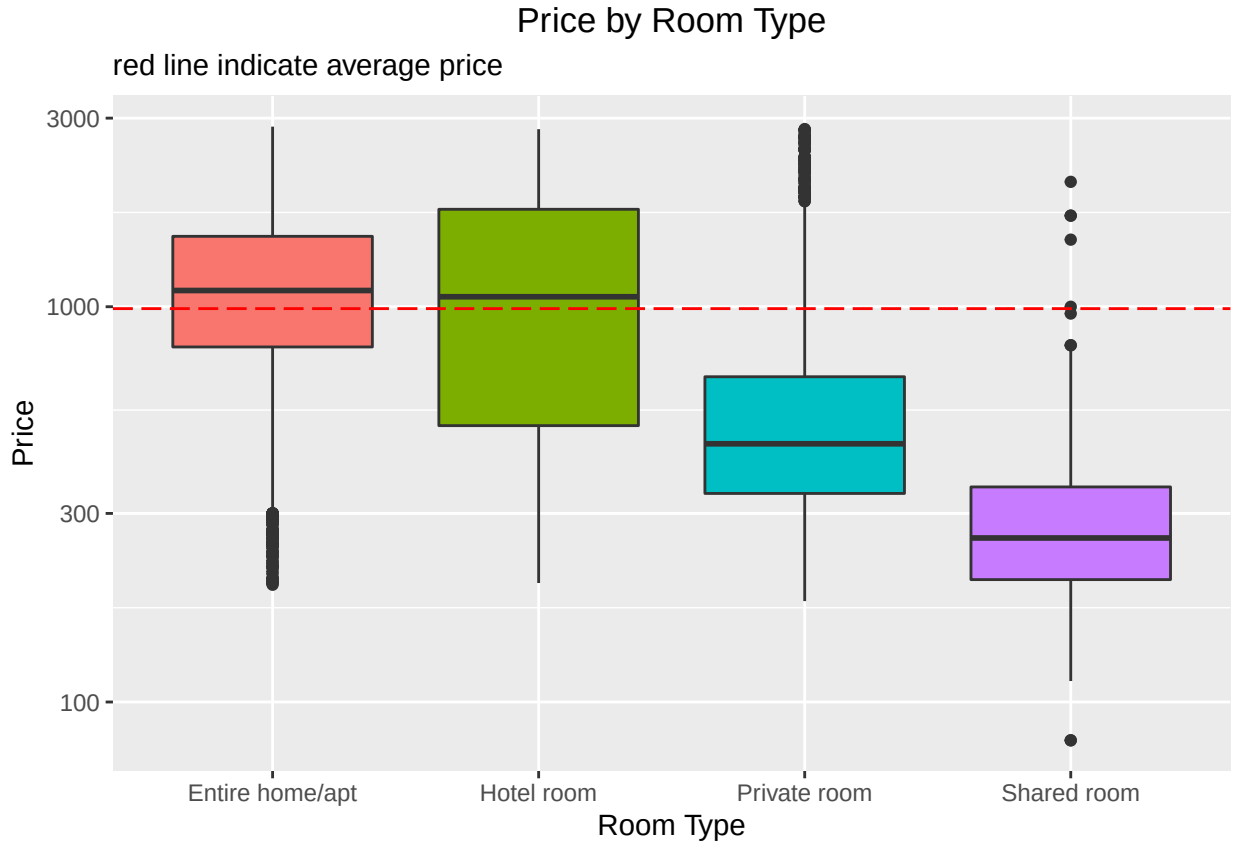
2.1.1 Accommodates In this series of boxplots we are noticing that as total accommodation increases for listings, so do the max and median prices. We also notice larger spreads middle 50% of prices per accommodation size for those that accommodate for 9+ guests.



2.1.2 Rating In this series of boxplots we are noticing that as host rating increases, so do the the max and median prices for their listings. The spread of the middle 50% of prices per listing remain consistent regardless of rating.

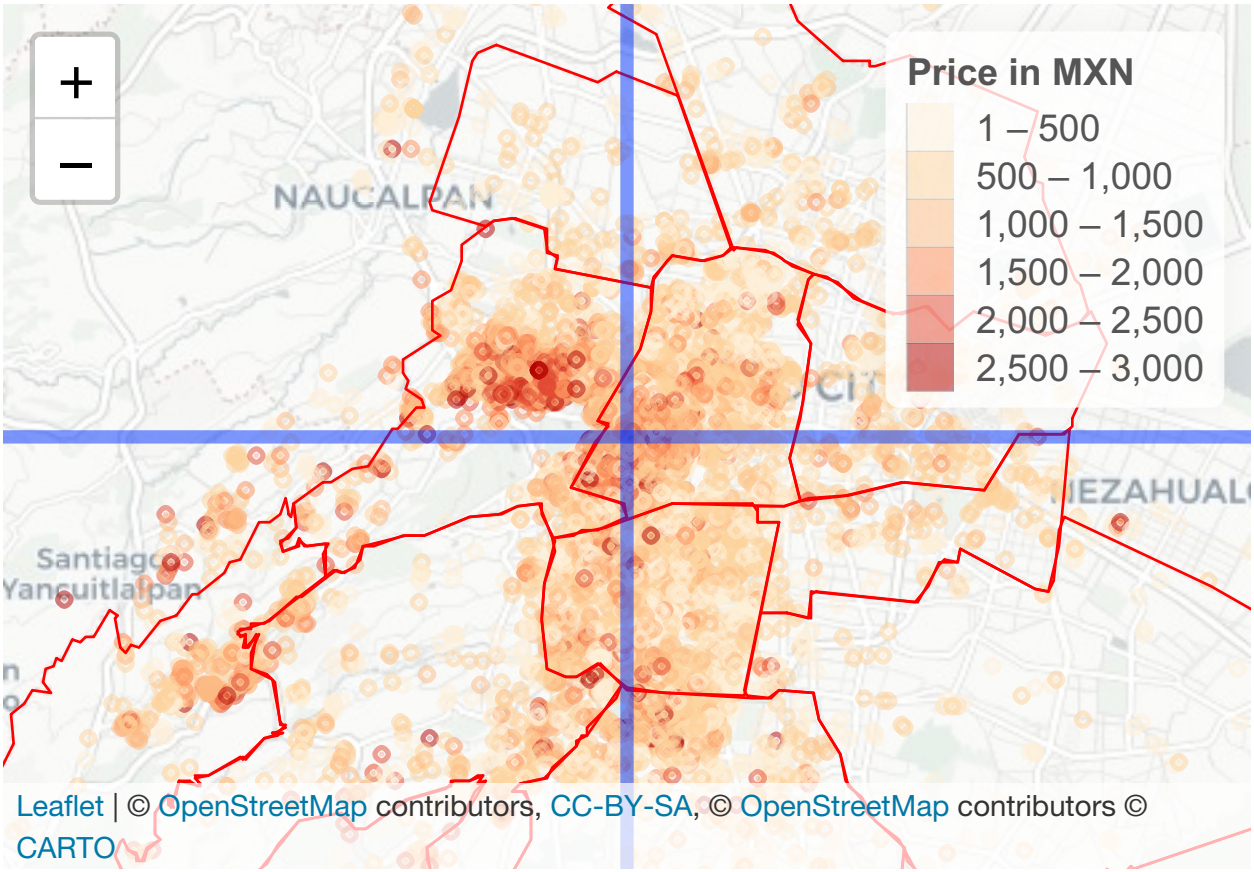


2.1.3 Room Type In this series of boxplots we are noticing that more private listings are more expensive than shared rooms. The spread of middle 50% listing prices for Hotel rooms is much greater than the other three room-type listings.



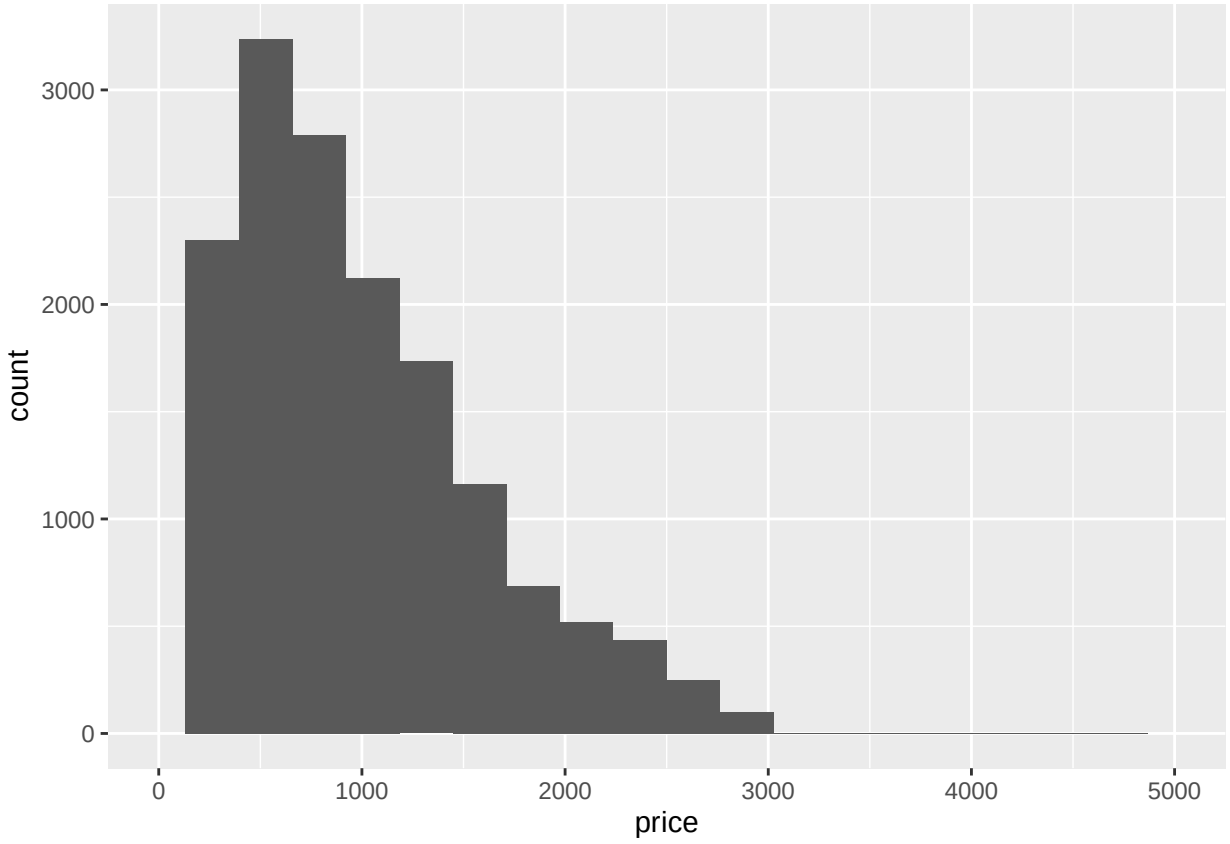
2.2 Heat map leaflet price

A leaflet map was generated, displaying all the Airbnb listings in the map of Mexico city to explore geographical influence in price. The solid blue lines are the median longitude and latitude to show where listings are clustered about.

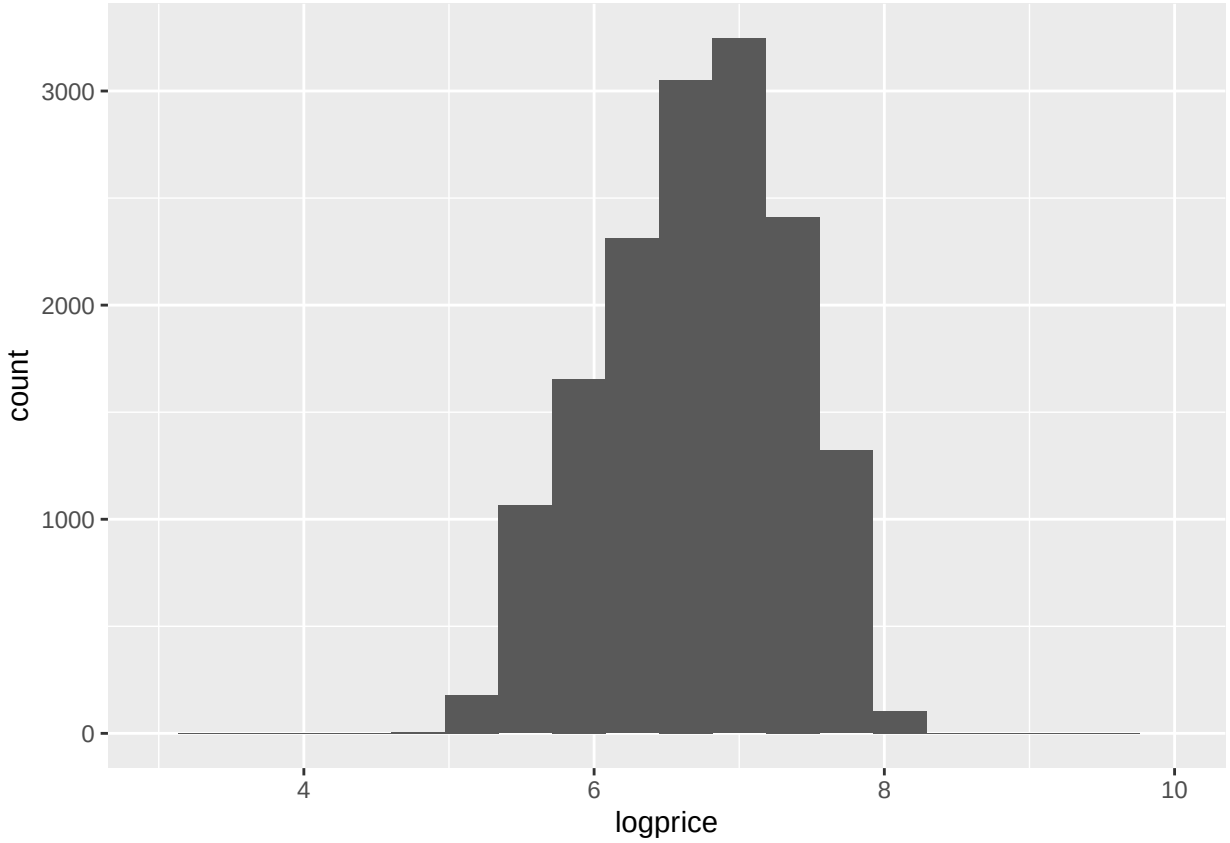


2.3 Distribution of Price

2.3.1 Right Skewed The following figure shows the distribution of price.



2.3.2 Log Normal Distribution Upon taking the the log of the price variable, we were able to get a normal distribution of the price responses.



3 Methodology

3.1 Statistical learning methods

For all models, we are fitting on the response variable of log(price) of listings.

For Stepwise and Best Subset models:

We decided to use the calculated BIC to determine the best number of variables (predictors) to include for stepwise forward, stepwise backward, and best subset models. We made the decision to use BIC instead of lowest AIC or CP since those two methods yielded inclusion of the majority of total predictors whereas BIC selected for for only 13-16 of the 25 total predictors.

For Lasso: Lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. The variable selection was determined by our optimal lambda value. Optimal lambda was obtained by performing cross-validation (10 folds, determined by examples executed in hw/lab models) on our data, and selecting the lambda that corresponded with lowest MSE.

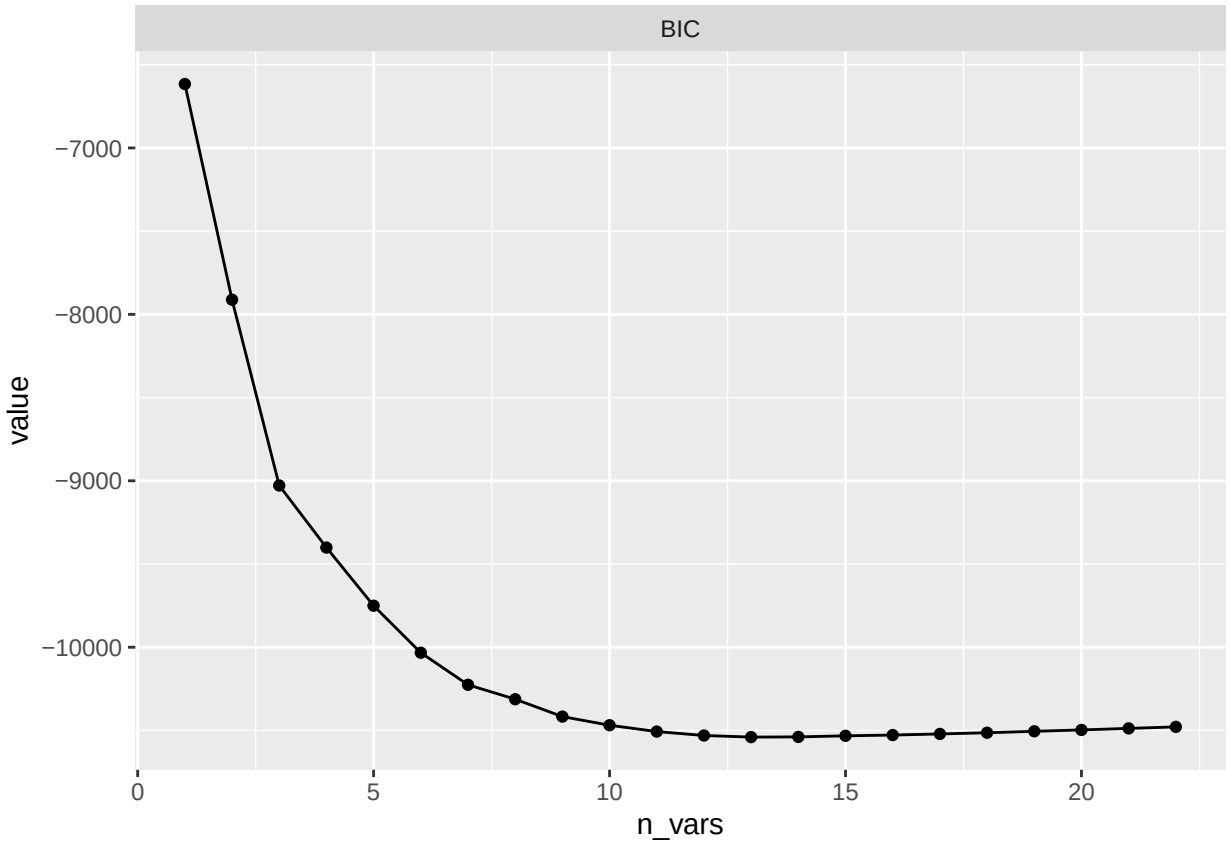
We used scaled data for all models since Lasso requires it and we wanted to keep as many factors consistent across the modelling for better interpretation of results. Ridge regression was not considered because it does not perform variable selection and we made an assumption that not all parameters were influential on the response variable of log-price. Results' table will be in results section.

4 Results

4.1 Best Subset

The following plot shows BIC on the y axis and number of predictors in the x axis. The lowest BIC corresponds to the model with 13 predictors

```
## Reordering variables and trying again:
```




```
##    BIC n_vars
##     13      1
```

The best variables for predicting airbnb listing price according to the best subset model are:

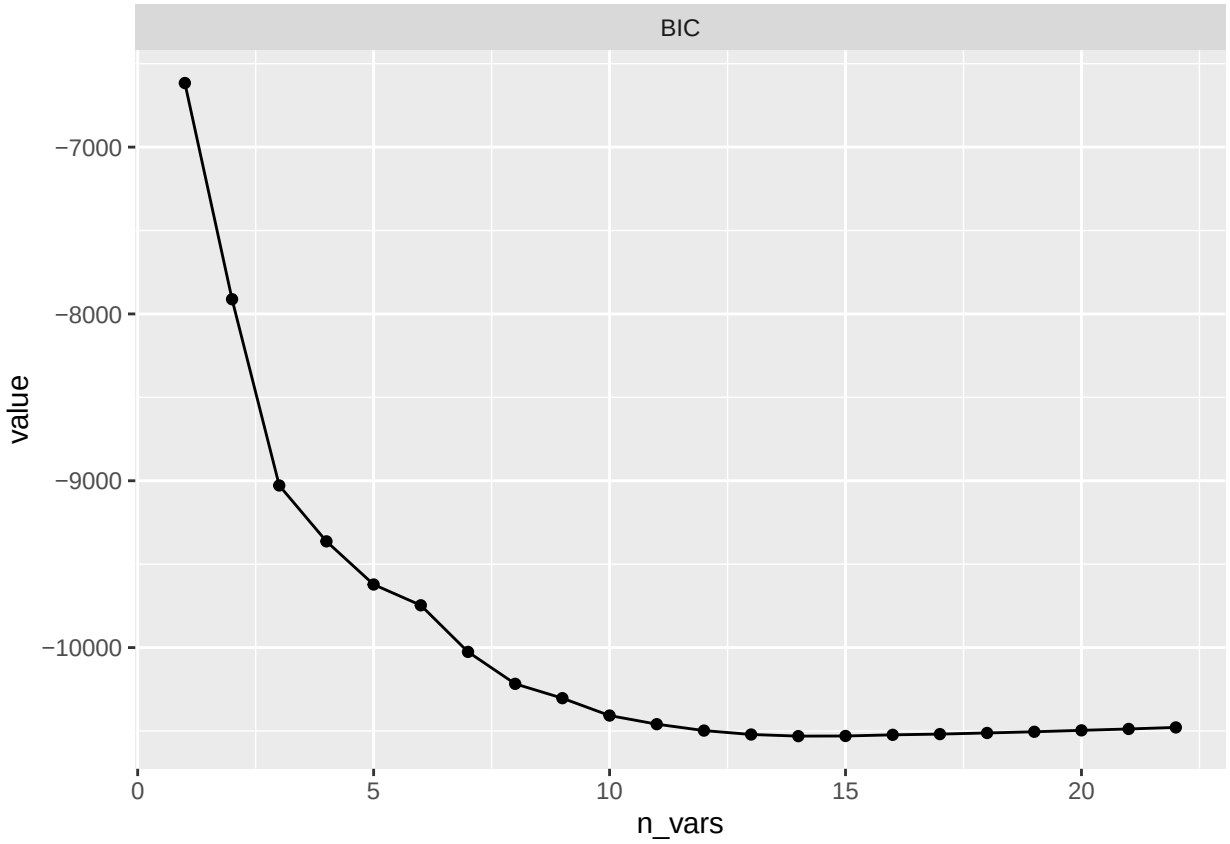
```
##                                      var
## 1                           accommodates
## 2                               bedrooms
## 3             host_total_listings_count
## 4                               latitude
## 5                  review_scores_rating
## 6      host_response_time_within.an.hour
## 7                host_has_profile_pic_t
## 8                number_of_reviews_l30d
## 9                    instant_bookable_t
## 10 host_response_time_within.a.few.hours
## 11               minimum_nights_avg_ntm
## 12                             longitude
## 13                                  beds
```

The number of predictors decreased from 25 to 13

4.2 Backwards selection

The following plot shows BIC on the y axis and number of predictors in the x axis. The lowest BIC corresponds to the model with 14 predictors

```
## Reordering variables and trying again:
```




```
##    BIC n_vars
##     14      1
```

The best variables for predicting airbnb listing price according to the forward selection model are:

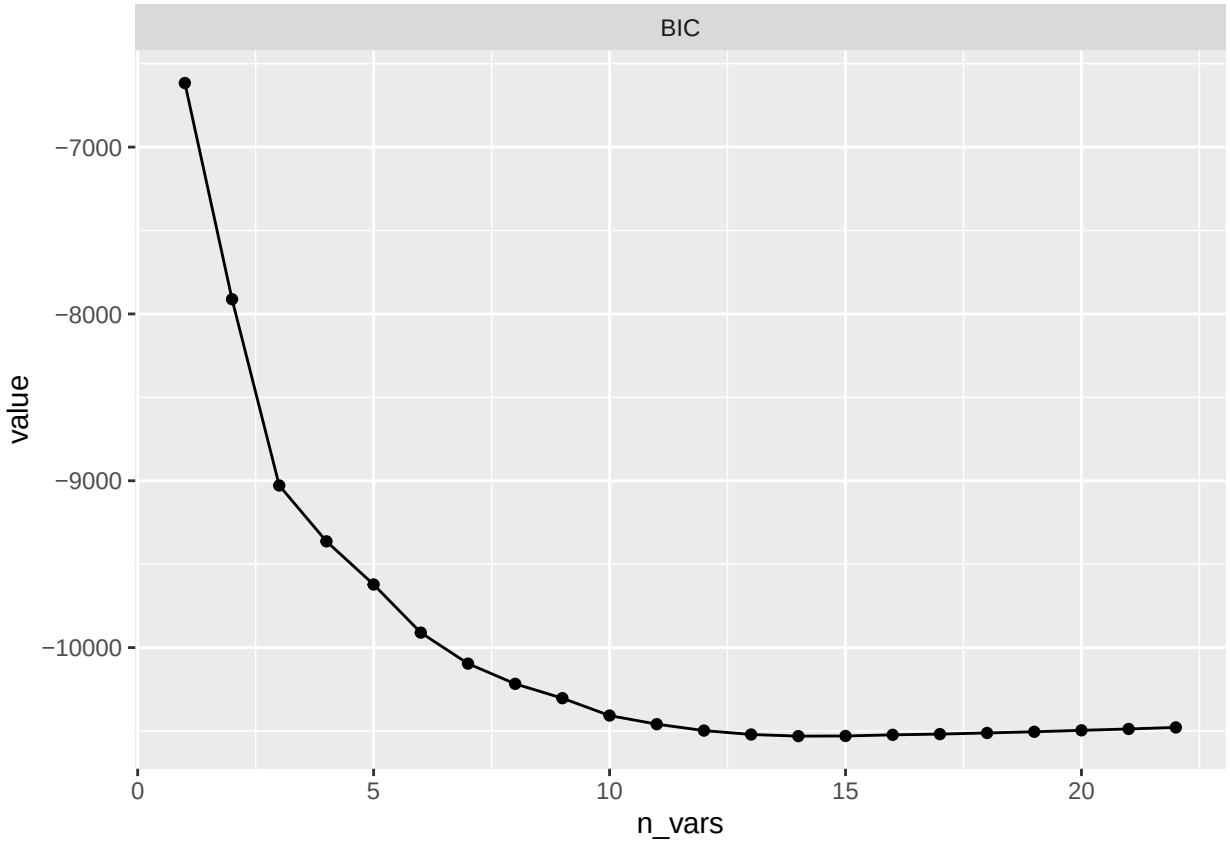
```
##                                    var
## 1                          accommodates
## 2                              bedrooms
## 3              host_total_listings_count
## 4                              latitude
## 5                   review_scores_rating
## 6                     instant_bookable_t
## 7                 host_has_profile_pic_t
## 8                   room_type_Hotel.room
## 9                 number_of_reviews_l30d
## 10                minimum_nights_avg_ntm
## 11 host_response_time_within.a.few.hours
## 12                                  beds
## 13                             longitude
## 14                room_type_Private.room
```

The number of predictors decreased from 25 to 14

4.3 Forward selection

The following plot shows BIC on the y axis and number of predictors in the x axis. The lowest BIC corresponds to the model with 14 predictors

```
## Reordering variables and trying again:
```




```
##    BIC n_vars
##     14      1
```

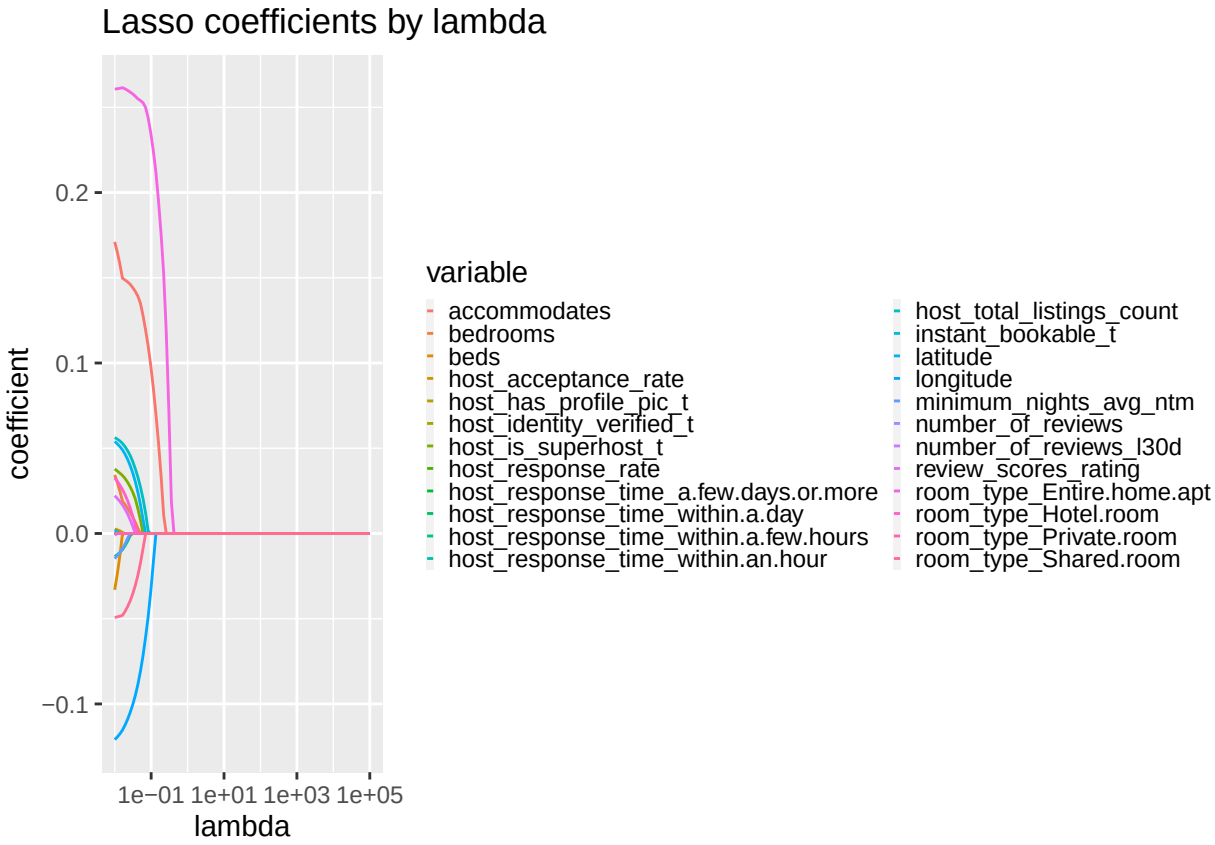
The best variables for predicting airbnb listing price according to forward selection model are:

```
##                                      var
## 1                           accommodates
## 2                               bedrooms
## 3             host_total_listings_count
## 4                               latitude
## 5                  review_scores_rating
## 6     host_response_time_within.an.hour
## 7                  room_type_Hotel.room
## 8                host_has_profile_pic_t
## 9                number_of_reviews_l30d
## 10                   instant_bookable_t
## 11 host_response_time_within.a.few.hours
## 12               minimum_nights_avg_ntm
## 13                            longitude
## 14                                 beds
```

The number of predictors decreased from 25 to 14

4.4 Lasso

The following plot shows the lasso coefficients by lambda.



The value of the best lambda is the following:

```
## [1] 0.001076295
```

The best variables for predicting airbnb listing price according to the lasso model are:

```
##                                     var
## 1              room_type_Entire.home.apt
## 2                           accommodates
## 3              host_total_listings_count
## 4                               latitude
## 5                     host_is_superhost_t
## 6                               bedrooms
## 7                    room_type_Hotel.room
## 8                    review_scores_rating
## 9                    host_acceptance_rate
## 10                     instant_bookable_t
## 11                number_of_reviews_l30d
## 12 host_response_time_within.a.few.hours
## 13                minimum_nights_avg_ntm
## 14                                  beds
## 15                 room_type_Shared.room
## 16                             longitude
```

The number of predictors decreased from 25 to 16

5 Discussion

```
##                                    var coefs_backward_per coefs_forward_per
## 1                          (Intercept)       816.43073418      816.43073418
## 2                         accommodates        26.39973733       41.88210955
## 3                             bedrooms         8.48636399       10.19173592
## 4                                 beds       -12.25562668      -14.77620882
## 5                  host_acceptance_rate                 NA                NA
## 6                host_has_profile_pic_t         0.57968353        0.52105413
## 7                  host_is_superhost_t                 NA                NA
## 8  host_response_time_within.a.few.hours        -2.65012596       -0.95458445
## 9      host_response_time_within.an.hour                 NA        3.72003028
## 10            host_total_listings_count         6.63662396        9.17015810
## 11                   instant_bookable_t         0.64784693       -0.54091527
## 12                            latitude         6.11642386        8.37184412
## 13                           longitude       -12.42928211      -13.73534724
## 14               minimum_nights_avg_ntm        -2.63982573       -1.66988837
## 15               number_of_reviews_l30d        -0.73793419        0.09284889
## 16                 review_scores_rating         4.39979623        5.14775686
## 17            room_type_Entire.home.apt                 NA                NA
## 18                 room_type_Hotel.room        -0.04737278        1.67787731
## 19               room_type_Private.room       -21.79636686                NA
## 20                room_type_Shared.room                 NA                NA
```

```
##                                    var coefs_bestsubset_per coefs_lass_per
## 1                          (Intercept)          816.43073418    816.4307342
## 2                         accommodates           41.78056603     18.6567858
## 3                             bedrooms           10.14872581      3.4924840
## 4                                 beds          -14.70453610     -3.2510639
## 5                  host_acceptance_rate                    NA      0.2641350
## 6                host_has_profile_pic_t            0.53441857             NA
## 7                  host_is_superhost_t                    NA      3.8535599
## 8  host_response_time_within.a.few.hours           -0.92270084     -1.3434529
## 9      host_response_time_within.an.hour            3.74165692             NA
## 10            host_total_listings_count            9.15869005      5.7909863
## 11                   instant_bookable_t           -0.45025116      0.2048331
## 12                            latitude            8.40509212      5.5518076
## 13                           longitude          -13.74949416    -11.3902867
## 14               minimum_nights_avg_ntm           -1.68761674     -1.4646428
## 15               number_of_reviews_l30d            0.04166698     -0.1060459
## 16                 review_scores_rating            5.12396963      2.2422445
## 17            room_type_Entire.home.apt                    NA     29.7897366
## 18                 room_type_Hotel.room                    NA      3.3034267
## 19               room_type_Private.room                    NA             NA
## 20                room_type_Shared.room                    NA     -4.8084392
```

To better understand and compare across the 4 different approaches for modeling variables and their effects on the response variable log-price, we took the variables of interest from each model and joined them in a single data frame by variable name (inclusive). The resulting data frame demonstrates the shared variables of interest for modeling effects on price for Airbnb listings(If NA it means that the specific model did not include that variable). Aside from showcasing conserved variables, we are also able to see at a glance for which variables are magnitude and sign preserved.

We decided to mutate new columns for our results and exponentiate the coefficients, subtract one from that value, and multiply by 100. This is due to literature on how to interpret log transformations. The result is

that we can now interpret changes in standard deviations of one variable causing a percent change in the response variable price.

For example:

The interpretation of the coefficient of the variable accommodates will be the following.

Best Subset

For every increase in 1 standard deviations of the variable accommodates it will cause a 41.78% percent increase in the response variable price.

Forward Selection

For every increase in 1 standard deviations of the variable accommodates it will cause a 41.88% percent increase in the response variable price.

Backward Selection

For every increase in 1 standard deviations of the variable accommodates it will cause a 26.4% percent increase in the response variable price.

Lasso

For every increase in 1 standard deviations of the variable accommodates it will cause a 18.66% percent increase in the response variable price.

The top 11 best predictors, those kept in the 4 models are:

```
##  [1] "accommodates"
##  [2] "bedrooms"
##  [3] "beds"
##  [4] "host_response_time_within.a.few.hours"
##  [5] "host_total_listings_count"
##  [6] "instant_bookable_t"
##  [7] "latitude"
##  [8] "longitude"
##  [9] "minimum_nights_avg_ntm"
## [10] "number_of_reviews_l30d"
## [11] "review_scores_rating"
```

5.1 Advice to Airbnb Hosts and Guest in Mexico City

5.1.1 If you want to increase the price of your listings: Create listings with high carrying capacity (accommodates, beds, bedrooms)

Respond fast, within an hour or few hours

Incentivize your guests to rate your listing

5.1.2 If you want to find a cheaper listing: Holding all other factors constant, cheaper listings are expected to be South-East of downtown Mexico City.

Based on the coefficients of latitude and longitude

Best Subset

-*Latitude* For every increase in 1 standard deviations of the variable Latitude it will cause a 8.4% percent increase in the response variable price.

-*Longitude* For every increase in 1 standard deviations of the variable longitude it will cause a 13.74% percent increase (increase because longitude is negative, and the coefficient is also negative) in the response variable price.

Backwards selection

-*Latitude* For every increase in 1 standard deviations of the variable Latitude it will cause a 6.1% percent increase in the response variable price.

-*Longitude* For every increase in 1 standard deviations of the variable longitude it will cause a 12.4% percent increase (increase because longitude is negative, and the coefficient is also negative) in the response variable price.

Forward selection

-*Latitude* For every increase in 1 standard deviations of the variable Latitude it will cause a 8.37% percent increase in the response variable price.

-*Longitude* For every increase in 1 standard deviations of the variable longitude it will cause a 13.73% percent increase (increase because longitude is negative, and the coefficient is also negative) in the response variable price.

Lasso

-*Latitude* For every increase in 1 standard deviations of the variable Latitude it will cause a 5.55% percent increase in the response variable price.

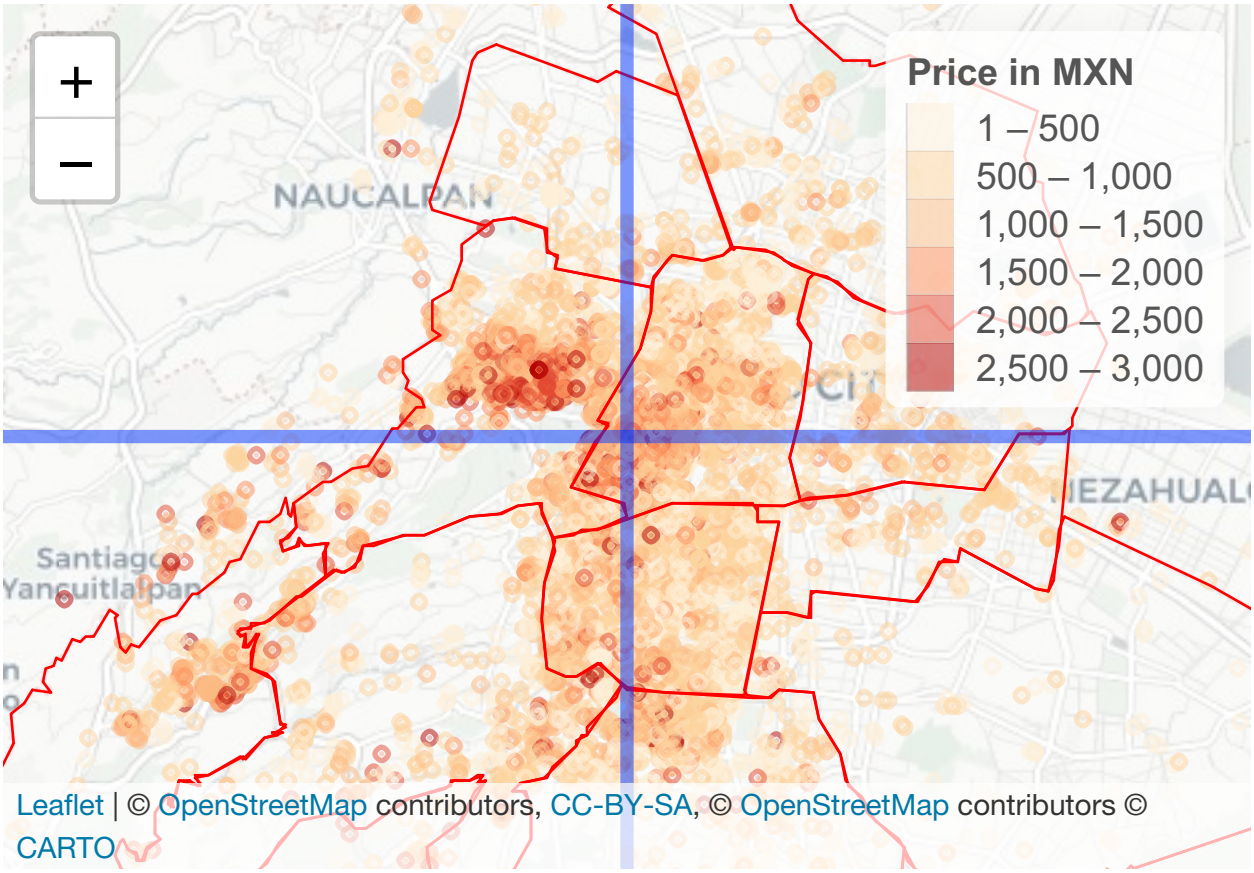
-*Longitude* For every increase in 1 standard deviations of the variable longitude it will cause a 11.39% increase (increase because longitude is negative, and the coefficient is also negative) decrease in the response variable price.

All our 4 models agree that:

Listings with a higher Latitude, more to the north have a higher price.

Listings with a higher Longitude, more to the west have a higher price. (The coefficient is negative because the longitude in our data set has a negative sign)

This is confirmed in our EDA map displaying the price of airbnb listings in mexico city. More expensive airbnbs are in the North West. Cheaper airbnbs are located in the South East part of Mexico City



5.2 Limitations

Some limitations that we experienced were with respect to what data was available, and how we transformed the response variable. As stated in the data section, we did not use time series data but rather used data scraped on a single day in September. Because of that, we could not see if variables had effects over time which is something that we would expect - especially with varying trends in neighborhood preferences. Another limitation of our scope was that we were measuring the response as log-price. Since we made that transformation in order to proceed with the Lasso method, we decided to use that scaled measure for the other three models. Doing so skews and complicates interpretation of the coefficients.

5.3 Areas of further research

For future research on this topic with this data, we would like to start off by first limiting our variables of interest. We could do this by first creating a correlation matrix to determine which factors are more heavily correlated with each other. By conducting this analysis, we could remove certain variables if others explain the same "idea" better (e.g. choosing between beds vs accommodates variables). Another good method that

we may wish to implement is the use of a tree for clustering. By doing so, we could get a quick overlook as to which variables are best for determining clustering between low, medium, and high priced listings. From those, we could determine which variables maybe good to look into, and if a clustering approach may be better than a one that assumes a linear relationship (i.e. what we completed in this analysis). Another area that we may want to consider for further research is using external data on property prices, time series data on tourism, comparing across certain dates of interest (like holidays), etc. and their influence on Airbnb listing prices in Mexico City.