Predicting Airbnb Prices in NYC

Bryan Phillips

1/10/2020

Overview

For this project, I have decided to utilize a large open-source dataset of Airbnb's in New York City in order to create a price prediction model using machine learning techniques from throughout this course. The main technique I will utilize are matrix factorization and regularization.

The dataset contains around 50,000 unique observation on individual Airbnb locations and their price point for 2019. This dataset is available at Kaggle.com at the link here: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

The dataset contains 16 variables summarized below, but the main information I will use for this analysis are Neighbourhood, Room Type, and Price.

```
##        id                                          name
##  Min.   :    2539   Hillside Hotel                   :   18
##  1st Qu.: 9471945   Home away from home              :   17
##  Median :19677284                                    :   16
##  Mean   :19017143   New york Multi-unit building     :   16
##  3rd Qu.:29152178   Brooklyn Apartment               :   12
##  Max.   :36487245   Loft Suite @ The Box House Hotel:   11
##                     (Other)                          :48805
##     host_id                 host_name        neighbourhood_group
##  Min.   :      2438   Michael     :  417   Bronx        : 1091
##  1st Qu.:   7822033   David       :  403   Brooklyn     :20104
##  Median : 30793816   Sonder (NYC):  327   Manhattan    :21661
##  Mean   : 67620011   John        :  294   Queens       : 5666
##  3rd Qu.:107434423   Alex        :  279   Staten Island:  373
##  Max.   :274321313   Blueground  :  232
##                       (Other)     :46943
##             neighbourhood      latitude       longitude
##  Williamsburg      : 3920   Min.   :40.50   Min.   :-74.24
##  Bedford-Stuyvesant: 3714   1st Qu.:40.69   1st Qu.:-73.98
##  Harlem            : 2658   Median :40.72   Median :-73.96
##  Bushwick          : 2465   Mean   :40.73   Mean   :-73.95
##  Upper West Side   : 1971   3rd Qu.:40.76   3rd Qu.:-73.94
##  Hell's Kitchen    : 1958   Max.   :40.91   Max.   :-73.71
##  (Other)           :32209
##             room_type         price          minimum_nights
##  Entire home/apt:25409   Min.   :    0.0   Min.   :   1.00
##  Private room   :22326   1st Qu.:   69.0   1st Qu.:   1.00
##  Shared room    : 1160   Median :  106.0   Median :   3.00
##                          Mean   :  152.7   Mean   :   7.03
##                          3rd Qu.:  175.0   3rd Qu.:   5.00
##                          Max.   :10000.0   Max.   :1250.00
##
##  number_of_reviews  last_review    reviews_per_month
##  Min.   :  0.00            :10052   Min.   : 0.010
##  1st Qu.:  1.00    6/23/19: 1413   1st Qu.: 0.190
##  Median :  5.00    7/1/19 : 1359   Median : 0.720
```

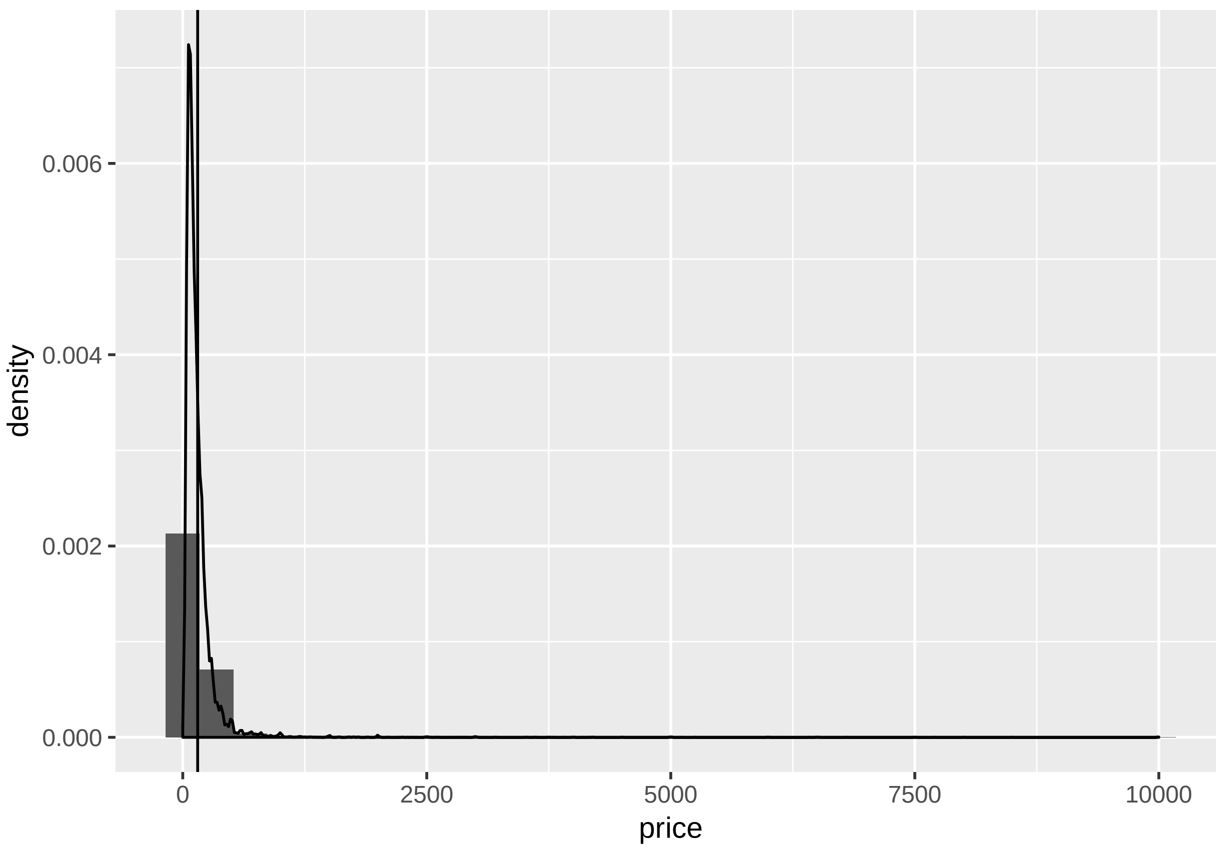
```
##  Mean   : 23.27    6/30/19: 1341   Mean   : 1.373
##  3rd Qu.: 24.00    6/24/19:  875   3rd Qu.: 2.020
##  Max.   :629.00    7/7/19 :  718   Max.   :58.500
##                    (Other):33137   NA's   :10052
##  calculated_host_listings_count availability_365
##  Min.   :  1.000               Min.   :  0.0
##  1st Qu.:  1.000               1st Qu.:  0.0
##  Median :  1.000               Median : 45.0
##  Mean   :  7.144               Mean   :112.8
##  3rd Qu.:  2.000               3rd Qu.:227.0
##  Max.   :327.000               Max.   :365.0
##
```

The way in which I will be evaluating the overall performance of the model will be to calculate Root Mean Squared Error (RMSE) of each model.

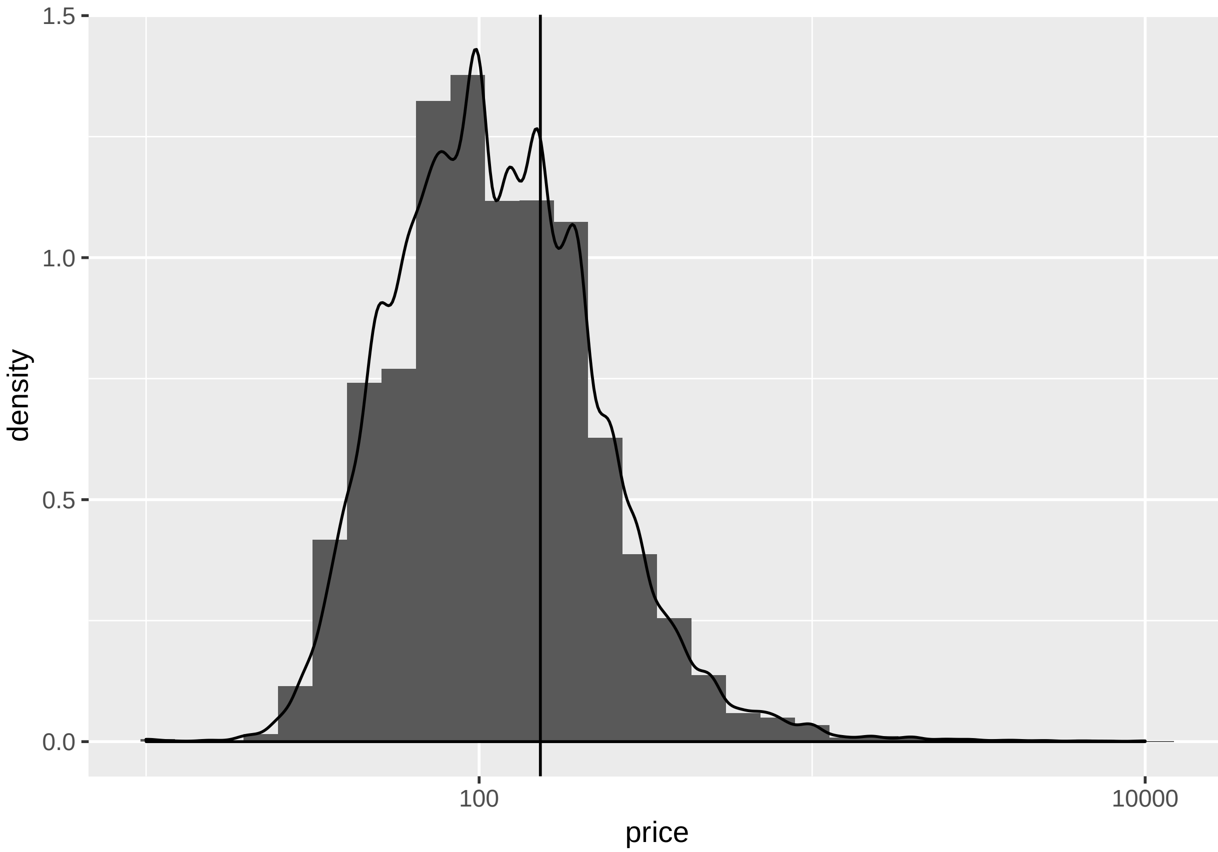
All of the code for this analysis is available on my github linked here: https://github.com/bkphillips/NYC_Airbnb_Price_Predict

Analysis

First looking at the price information, I notice it contains some pretty large outliers as seen in the density plot below:

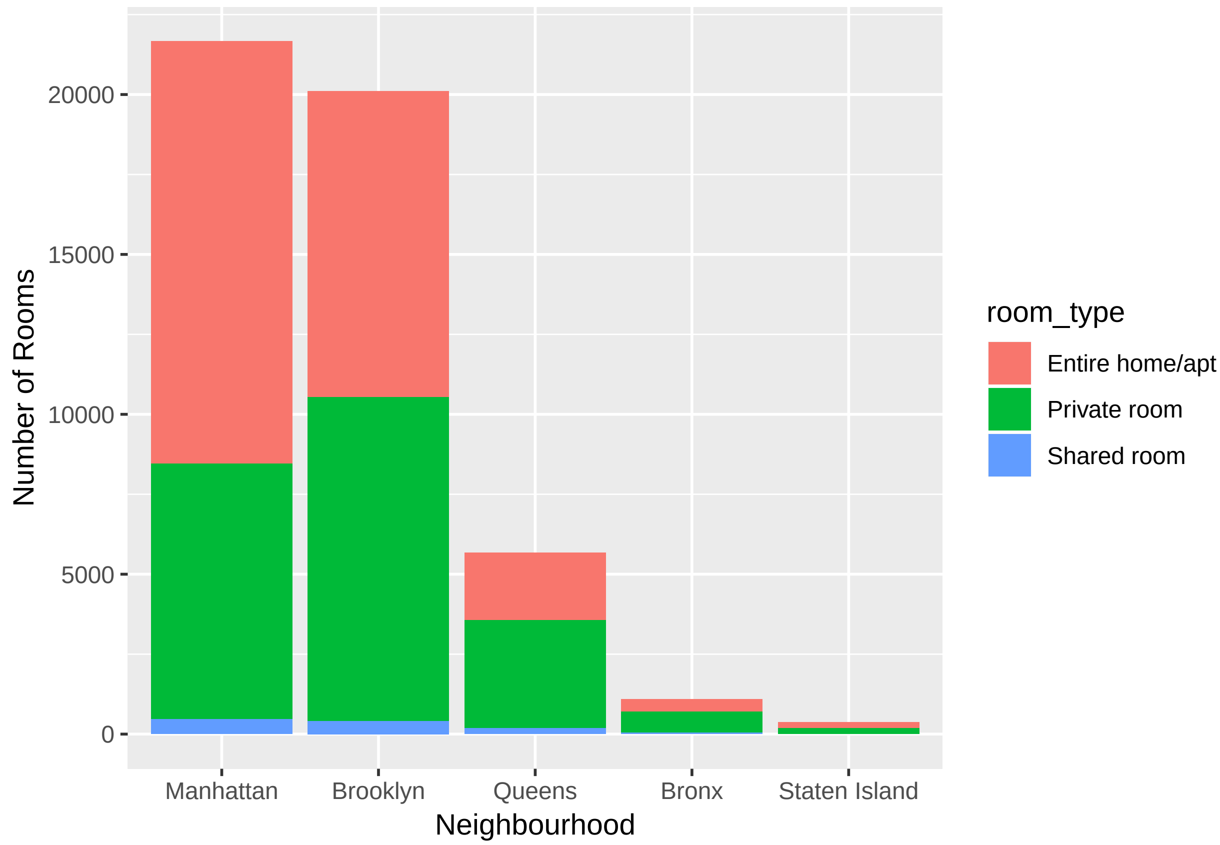


I then decided to look at the log distribution below, where the average price of $152 becomes more apparent. (average shown with vertical line)

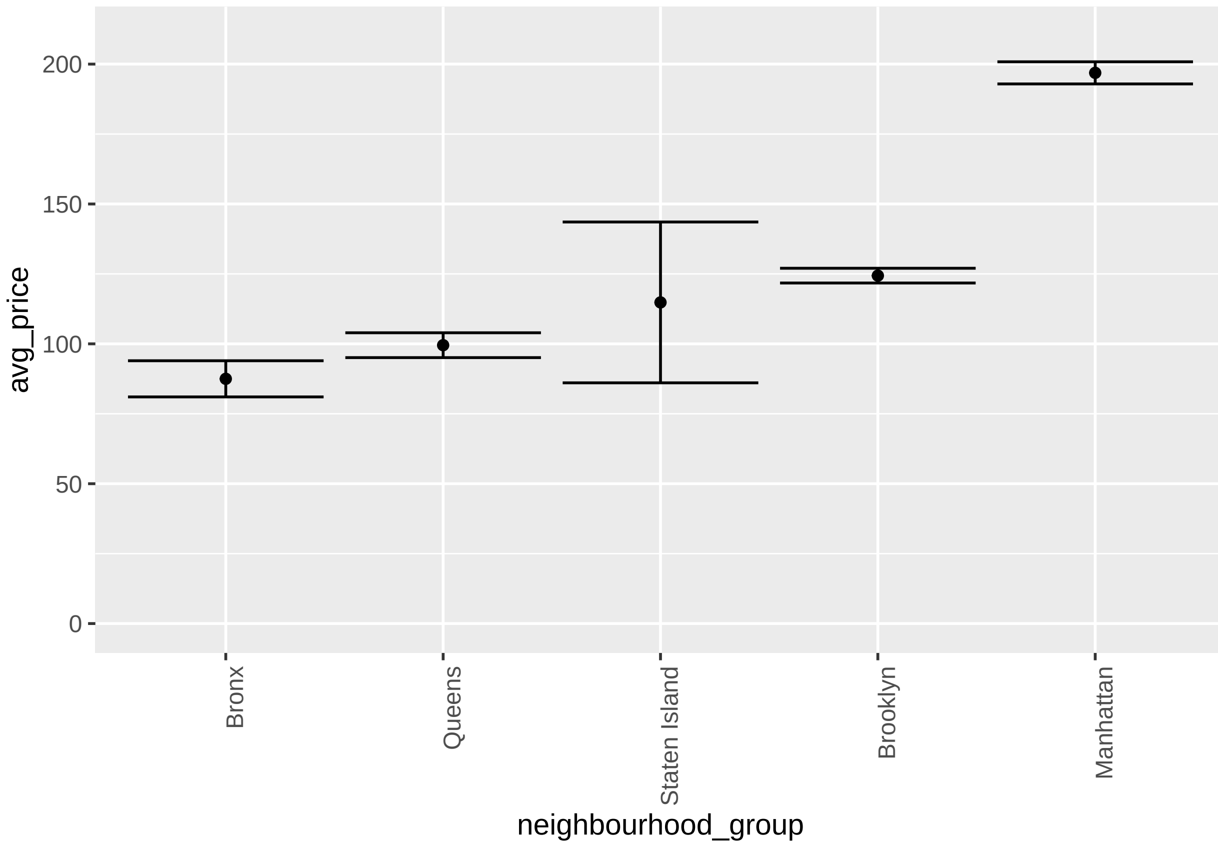


The other key descriptive variable are Neighbourhood Group, Neighborhood, and Room type.

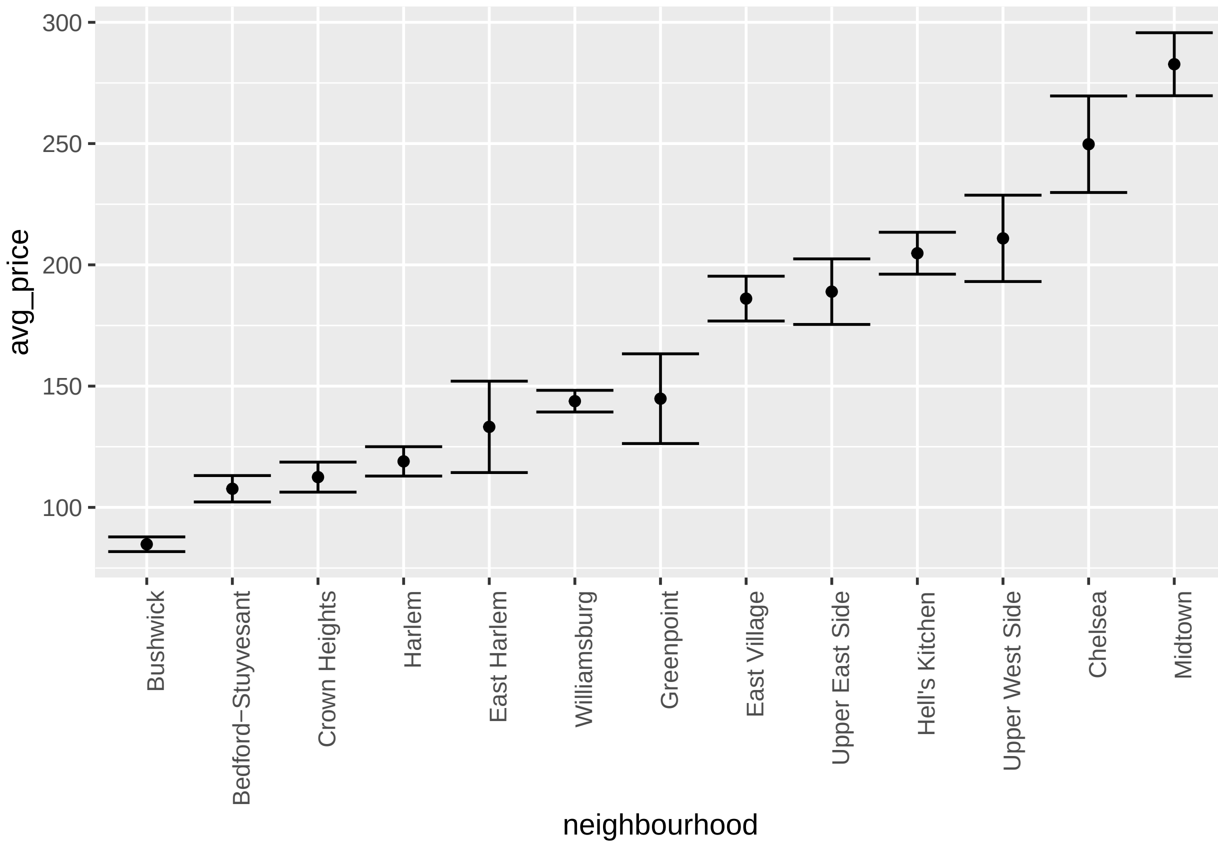
There are 221 unique neighborhoods, so for the purpose of describing the dataset I will mostly show the 5 main groups in which they fall into. There are also 3 main room types: Entire home, private room, or shared room. Below you can see a count of the types of room in the different areas of NYC. You can see the majority of locations are in Manhattan and Brooklyn. They are also mostly entire home/apt. or private rooms.



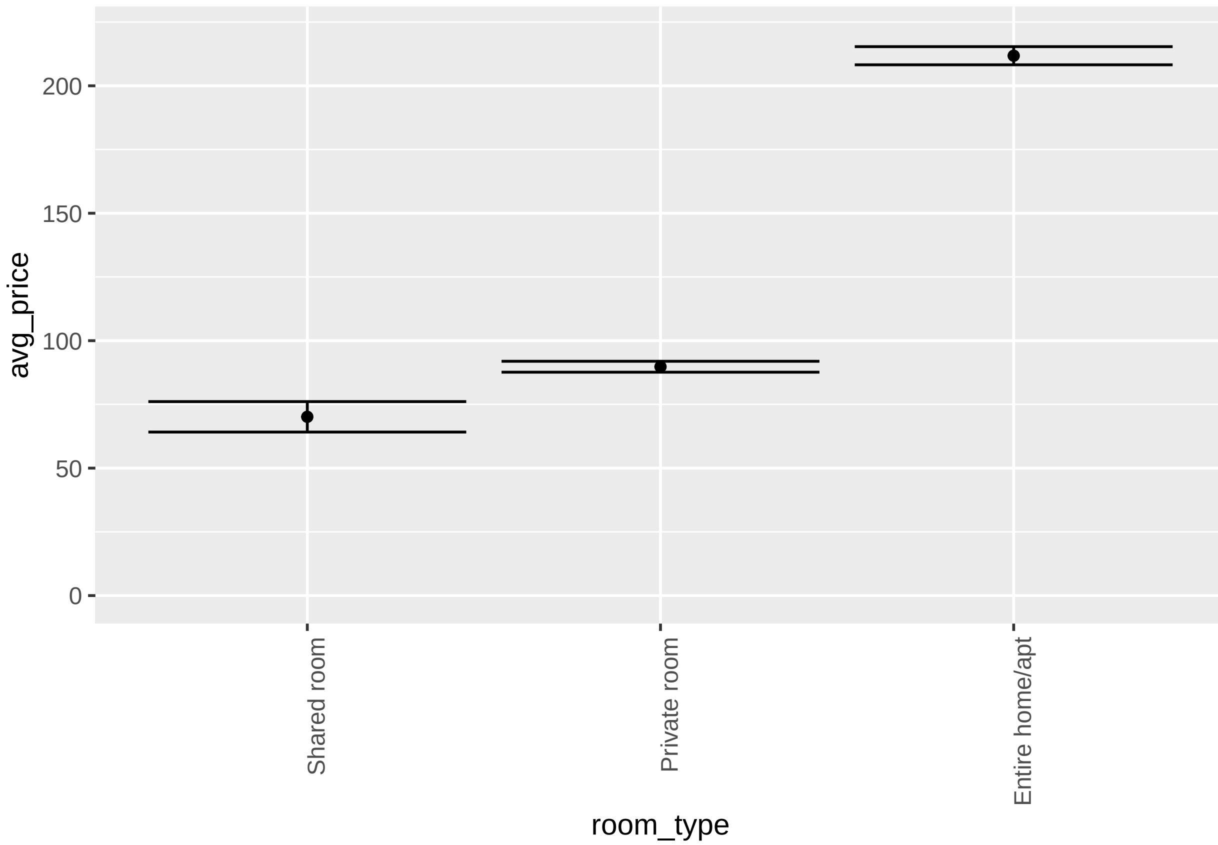
Below is the average price for each neighborhood with Manhattan noticeably more expensive compared to the other neighbourhoods.



Below are the average price of the neighborhoods that have over 1,000 Airbnb locations. Midtown is most expensive and Bushwick is the cheapest.



Below is the variation of average price by room type with Entire home/apt. being drastically more expensive



In order to create my training and testing datasets, I decided to remove locations with a price of $0 or anything above $500 based on the outliers that were seen in the initial density plots. I then partitioned the data into 70% for the training and 30% for the testing set.

Modeling and Results

I then began testing just the average first model on the price data, which gave a RMSE of 85.4. When I added the neighborhood group effects (b__g), this brought it down to 80.6. When testing neighborhood effects (b__n), it had a much better performance of 74.6, so I decided to stay with just b__n. The fourth model then used the room type effects (b__t) which significantly brought down the RMSE to 64.5.

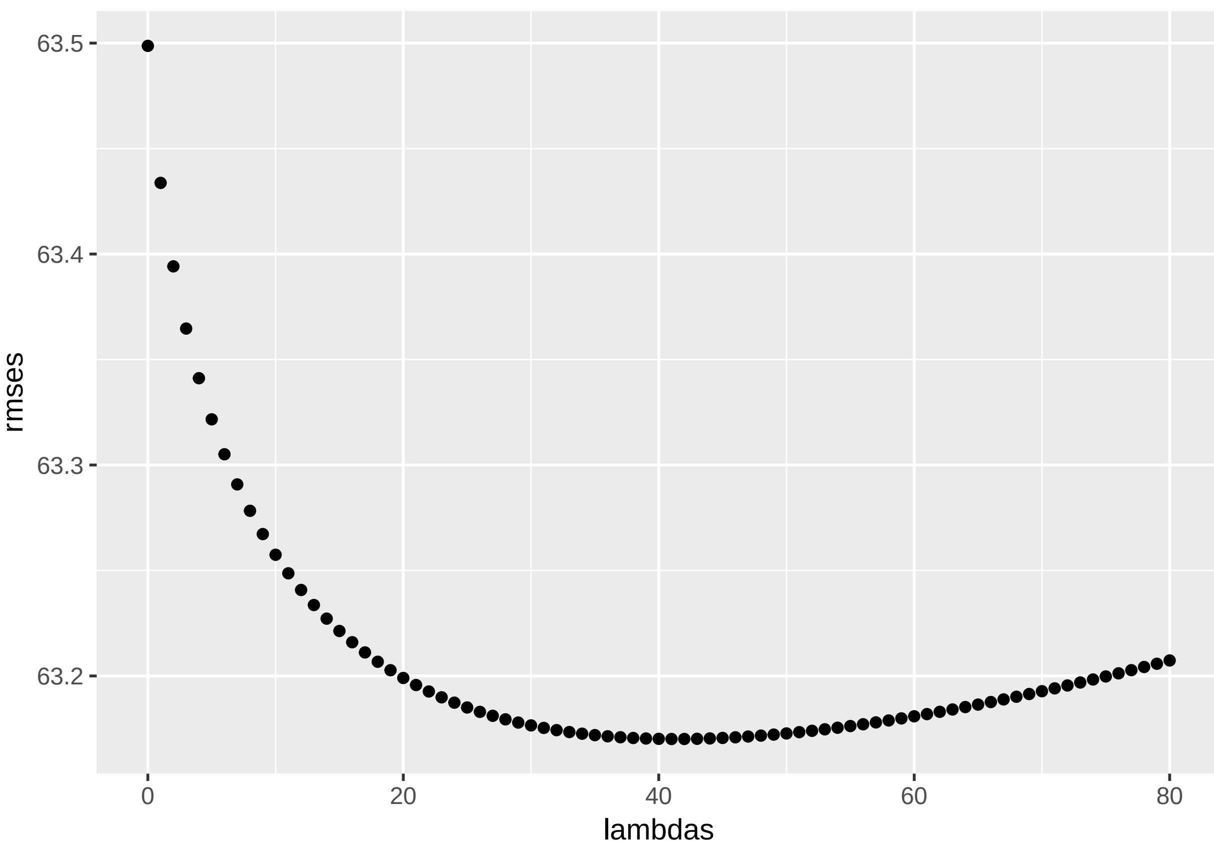
Then I then tried regularlizing the data because I figured that neighbourhoods that had more listings probably have more trustworthy prices that are more accurate. This only brought down my RSME to 64.1.

Below is a table of results of modeling on my training data and testing my final model on my training dataset:

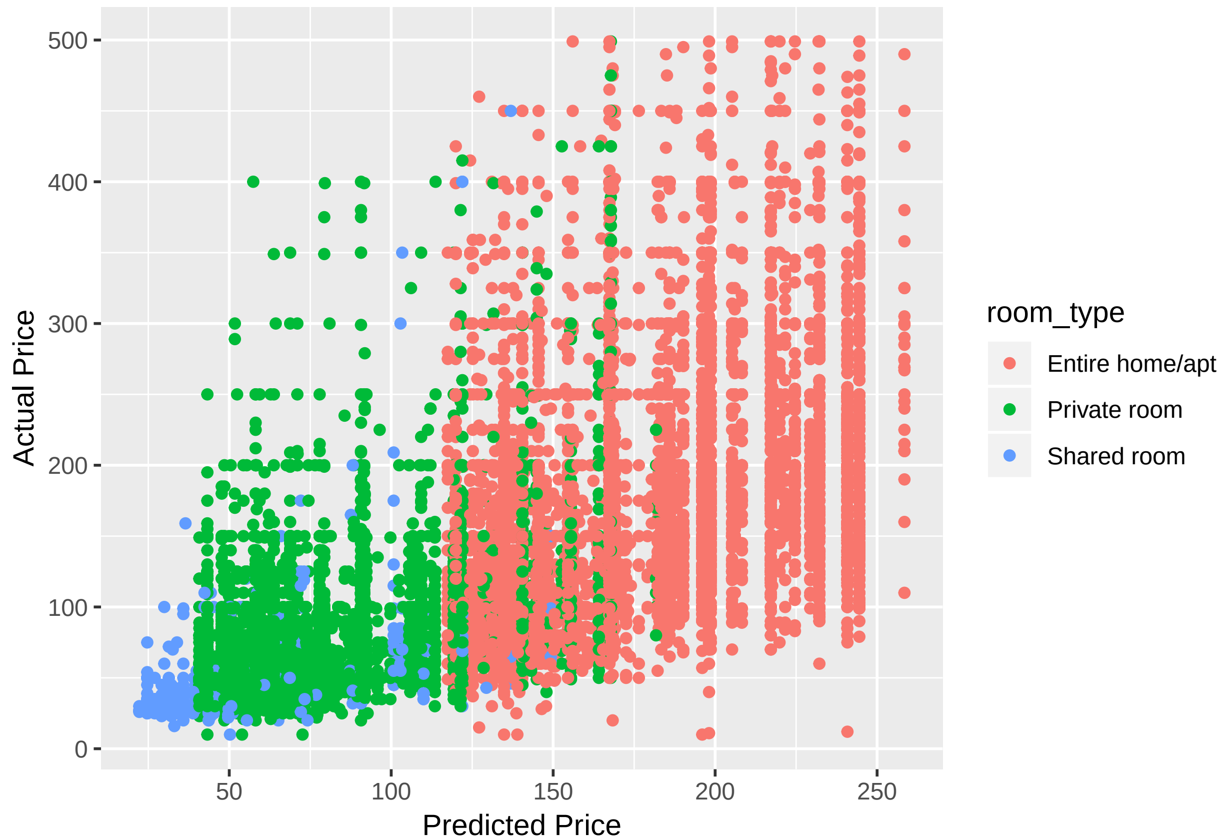
method	RMSE
Just the average	85.41663
Neighbourhood Group Mode	80.63411
Neighbourhood Model	74.57021
Neighbourhood + Room Type Model	64.45839
Regularized Neighbourhood + Room Type Model	64.18415
Testing Final Regularized Neighbourhood + Room Type Model	63.17012

Below is my final model used on the test set and the plot of the RSME's that were used to fine tune the lambda's for the regularization technique. You can see the optimal lambda for minimized RMSE is around 40:

```r
#Testing the Final Regularized Model of b_n + b_t
lambdas <- seq(0, 80, 1)
rmses <- sapply(lambdas, function(l){
  mu <- mean(train_set$price)
  b_n <- train_set %>%
    group_by(neighbourhood) %>%
    summarize(b_n = sum(price - mu)/(n()+l))
  b_t <- train_set %>%
    left_join(b_n, by="neighbourhood") %>%
    group_by(room_type) %>%
    summarize(b_t = sum(price - b_n - mu)/(n()+l))
  predicted_price <-
    test_set %>%
    left_join(b_n, by = "neighbourhood") %>%
    left_join(b_t, by = "room_type") %>%
    rowwise()  %>%
    mutate(pred = sum( mu, b_n, b_t, na.rm=TRUE)) %>%
    .$pred
  return(RMSE(predicted_price, test_set$price))
})
qplot(lambdas, rmses)
```



I then plotted the predicted prices to see how they compared with the actual prices. The plot is also colored by the room type. It looks as though even though I removed large outliers, the more expensive locations are causing a large amount of the error. It also looks as though some neighbourhoods are also clumping into columns.



Below is a random sample of 20 predicted prices (pred) and actual price (price) with their difference (diff). You can see the majority fall within \$40 of the actual price, but the more expensive locations tend to be farther off target.

```
## Source: local data frame [20 x 6]
## Groups: <by row>
##
## # A tibble: 20 x 6
##    neighbourhood_group neighbourhood      room_type       price  pred     diff
##    <fct>               <fct>              <fct>           <int> <dbl>    <dbl>
##  1 Manhattan           Midtown            Private room      280 168.  -112.
##  2 Manhattan           Tribeca            Entire home/~     358 258.   -99.6
##  3 Manhattan           Gramercy           Entire home/~     300 206.   -93.9
##  4 Brooklyn            Fort Greene        Entire home/~     275 183.   -91.6
##  5 Brooklyn            Greenpoint         Entire home/~     205 168.   -36.6
##  6 Brooklyn            Bedford-Stuyvesa~ Entire home/~     170 135.   -35.2
##  7 Brooklyn            Williamsburg       Entire home/~     190 167.   -22.6
##  8 Brooklyn            Bedford-Stuyvesa~ Private room       65  58.1   -6.86
##  9 Brooklyn            Bedford-Stuyvesa~ Private room       62  58.1   -3.86
## 10 Brooklyn            Bushwick           Private room       45  43.2   -1.75
## 11 Brooklyn            Bedford-Stuyvesa~ Private room       58  58.1    0.144
## 12 Brooklyn            Bedford-Stuyvesa~ Entire home/~     119 135.    15.8
## 13 Bronx               Schuylerville      Private room       60  83.7   23.7
## 14 Manhattan           East Harlem        Private room       50  77.9   27.9
## 15 Manhattan           Chelsea            Entire home/~     180 232.    52.2
## 16 Manhattan           Harlem             Entire home/~      89 145.    56.5
## 17 Queens              Long Island City   Entire home/~      99 156.    57.0
```

```
## 18 Manhattan           Hell's Kitchen    Entire home/~   160 217.    57.2
## 19 Manhattan           Harlem            Entire home/~    77 145.    68.5
## 20 Brooklyn            Prospect Heights  Entire home/~   100 169.    69.1
```

Conclusion

Using matrix factorization of the key descriptive factors of the location, I was able to more acurately predict the price of each airbnb. I was surprised to see that regularization did not improve the prediction of the price by much. The large price outliers are a challenging aspect of this dataset. It would be helpful if there was further information given about each location that could help convey other aspects that lead to a higher or lower price such as the quality of the space, amenities, or walking score. Further information may help predict these outliers. I would also like to add a confidence interval that would likely fall within the majority of the given prices as seen by the random sample where the majority of locations are within $40 of the actual price.