Airbnb Project

Helen Zhou

11/26/2017

Midterm Project_Airbnb

Introduction

This project is designed to display data analysis based on Inside Airbnb data set from http://insideairbnb.com/get-the-data.html. Datas contain two cities: Boston and New York City. For each city, I used property listing data.

The project is focused on the difference of properties between two cities. Taking from there, we can see some potential market strategies to promote offers from each city, from company's perspective; or we can see the bargain strategies when we as guests to choose stayover places when we visit the cities.

The major procedure can be summarized as below: 1. Explorary Data Analysis and Visualization 2. Multi-level linear model on price prediction 3. Model Checking 4. Appendix

Due to time limit, the project is only able to explore some aspects. The analysis has limitations on such as safety control, walk score and social interaction between landlords and guests. Feel free to contact me if you have any suggestions!

Data Overview

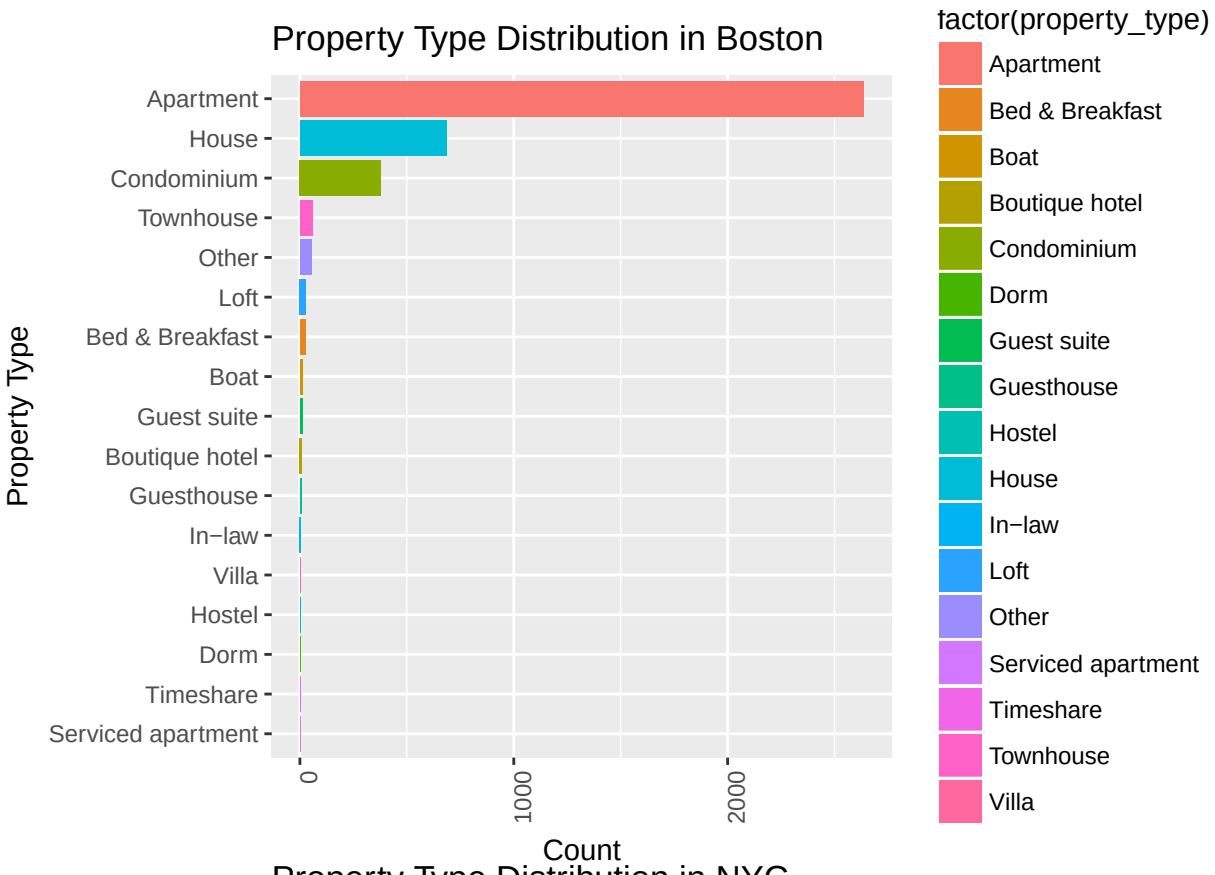
After data cleaning, Boston has 3925 properties on the list and New York City has 33938 properties on the list. The total 36 variables contain property information, host information and hosting policies.

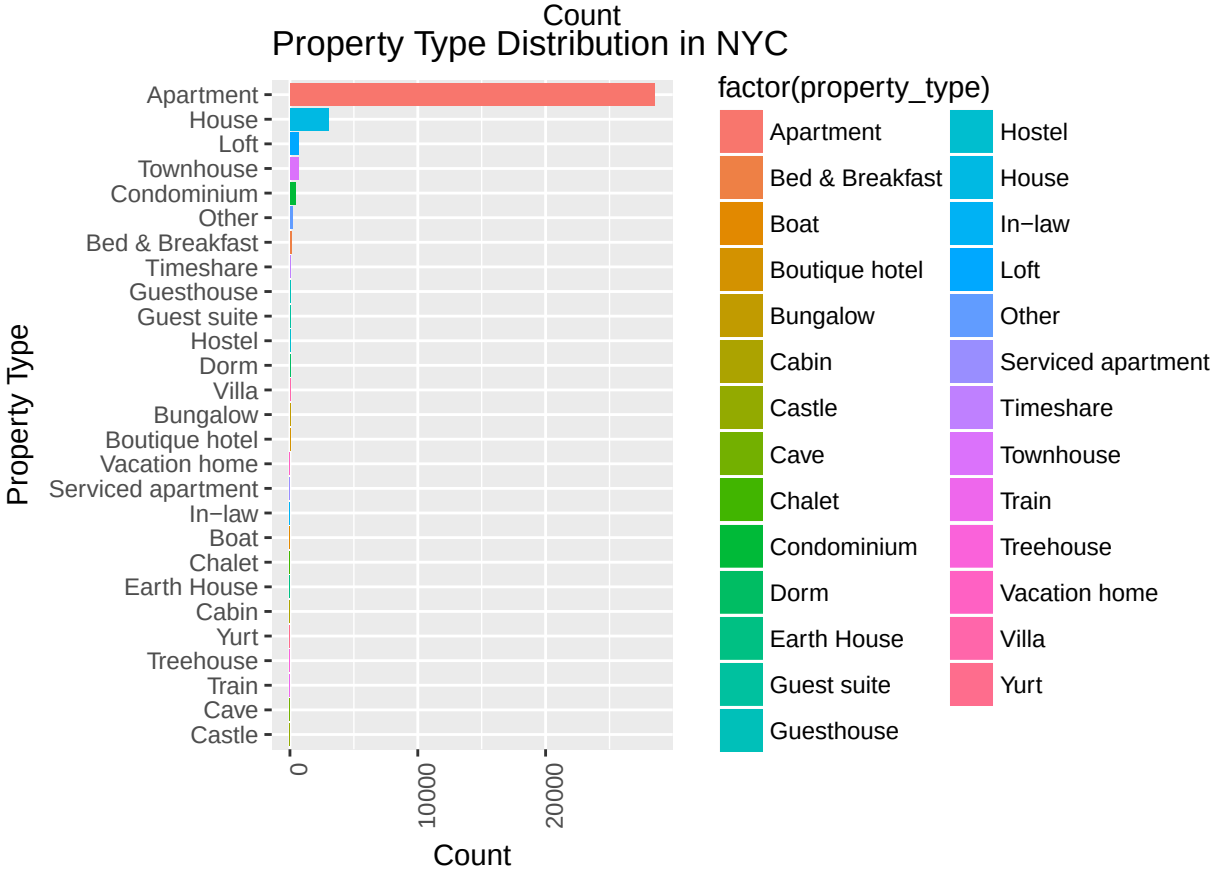
For the outcome variable, I define the price by creating new variable, the price per person, which equals sum of price and cleaning fee divided by guests_included.

$$Price\,per\,person = \frac{(Price + Cleaning\,fee)}{Guests\,included}$$

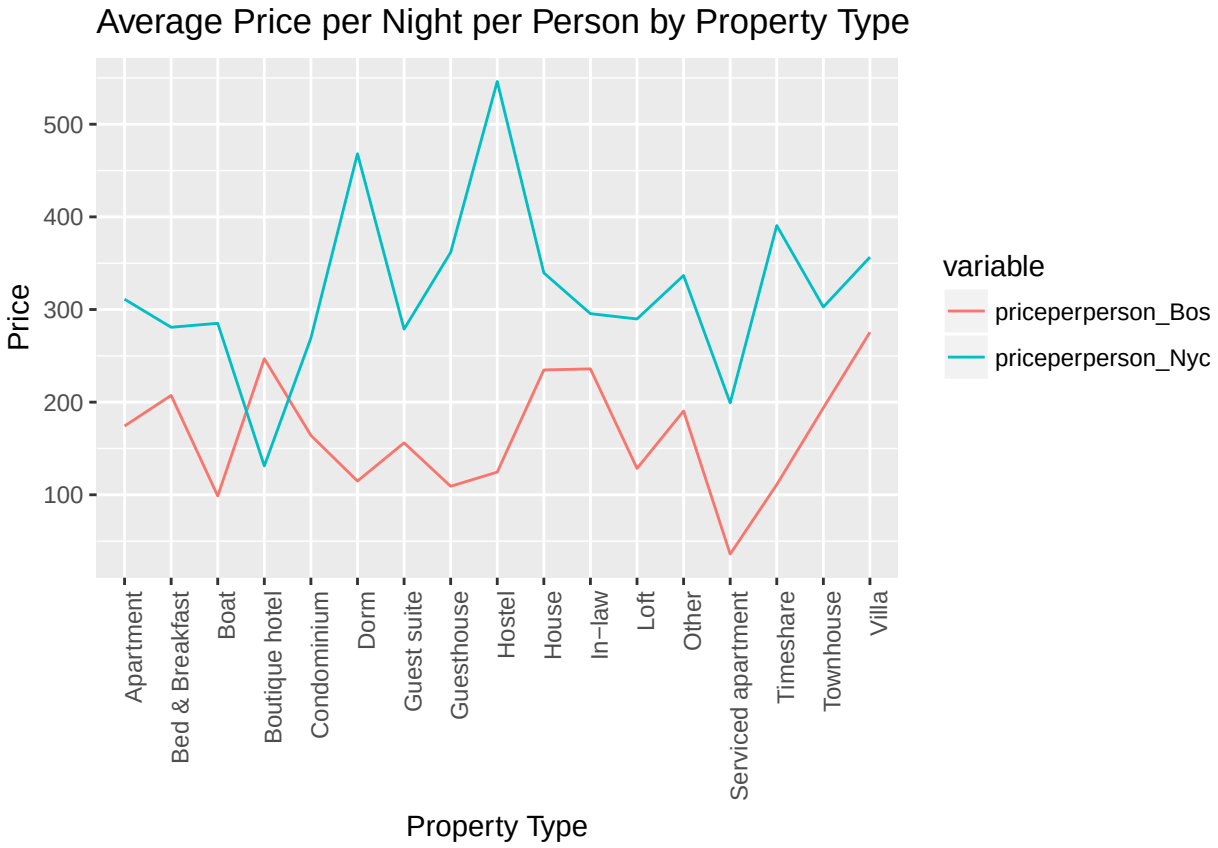
Explorary Data Analysis and Visualization

To find out people's experience with Airbnb in one city, we first need to know what are the options there. Things showed below are somethings I think are important to take a look at.

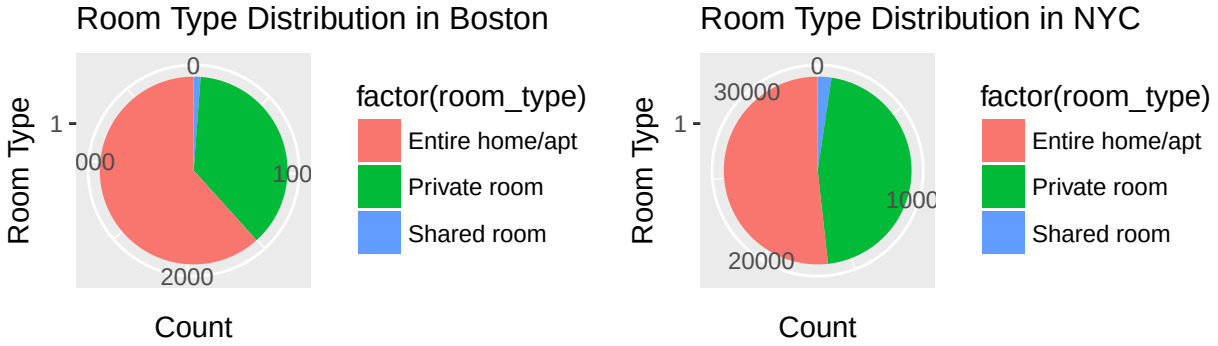

Property Type Distribution in Boston
Property Type
Apartment
House
Condominium
Townhouse
Other
Loft
Bed & Breakfast
Boat
Guest suite
Boutique hotel
Guesthouse
In−law
Villa
Hostel
Dorm
Timeshare
Serviced apartment
0
1000
2000
Count
factor(property_type)
Apartment
Bed & Breakfast
Boat
Boutique hotel
Condominium
Dorm
Guest suite
Guesthouse
Hostel
House
In−law
Loft
Other
Serviced apartment
Timeshare
Townhouse
Villa


Property Type Distribution in NYC
Property Type
Apartment
House
Loft
Townhouse
Condominium
Other
Bed & Breakfast
Timeshare
Guesthouse
Guest suite
Hostel
Dorm
Villa
Bungalow
Boutique hotel
Vacation home
Serviced apartment
In−law
Boat
Chalet
Earth House
Cabin
Yurt
Treehouse
Train
Cave
Castle
0
10000
20000
Count
factor(property_type)
Apartment
Bed & Breakfast
Boat
Boutique hotel
Bungalow
Cabin
Castle
Cave
Chalet
Condominium
Dorm
Earth House
Guest suite
Guesthouse
Hostel
House
In−law
Loft
Other
Serviced apartment
Timeshare
Townhouse
Train
Treehouse
Vacation home
Villa
Yurt

As we can see, apartment style buildings take the majority part of the properties in Boston and New York. There are variety here and there with small count relatively. Overall, I think properties provided in two cities are similar. Then, how about the price related to different properties?



Isn't it interesting that different property style leads to very different price variety in two cities? I will treat property tyle as group effect in the price prediction later.



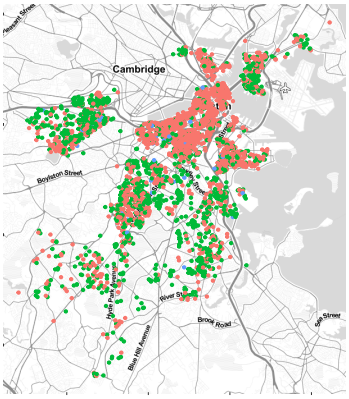
From the charts, we can see that Boston has more Entire home/apt property percentage than NYC has, and less percentage of private rooms as NYC has. Again, what are the price distribution related to this?

Table 1: Average Price by Room Type

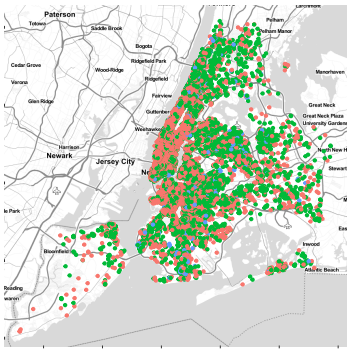
Room Type	Average Price(Boston)	Average Price(New York)
Entire home/apt	134.8235	206.0178
Private room	262.7514	428.8366
Shared room	279.5426	403.5122

From the table, though New York on average has higher price than Boston does, the change rates among different room types are very similar.

Property Distribution in Boston



Property Distribution in NYC

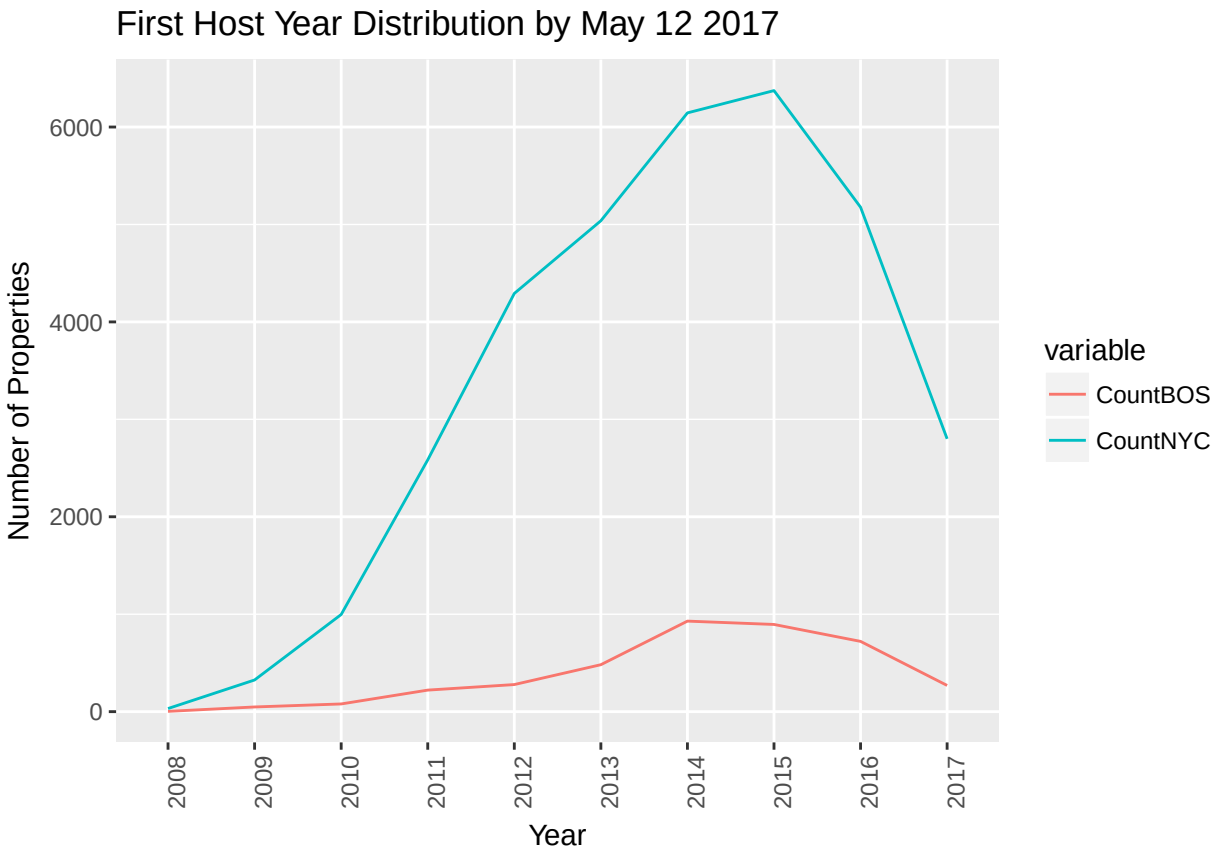


Some find-outs:

1. Different room types distributed in Boston is less blended as they are in New York.
2. New York is densely placed by airbnb properties almost all the places. While Boston is a little scatterd.
3. Downtown/Touristy areas are the densest part.

In Boston, area related words are most commonly mentioned, such as “near”, “downtown”,“fenway” and “boston”, which implies that hosts tend to use location convenience and popularity to attract guests. Descriptive words used are “cozy”,“spacious” and “luxury”. In New York, room types are most commonly mentioned, like “private (room)”, “studio” and “apartment”, which implies hosts emphasize on the room types to attract guests. Descriptive words used are “cozy”,“spacious” and “sunny”.

Next, I wonder if hosts expand in differently ratios in both cities so I make two tables below. at the first host year of the hosts



We can see they both have similar trends of peek of increasing/decreasing, however, the slope varies. New York grows much faster than Boston does.

How about super-host influence? Super host is defined for hosts who have 5 star views, high response rate and large number of hosting experiences. In general, it's a title of the hosts who are commited to their hostings. I am curious to see its influence on the price, which also indirectly implying hosting experience influence on the price.

Table 2: Super Host Percentage and Average Price in Boston

Super Host?	Percentage	Average Price
f	0.77	182.9886
t	0.23	187.5686

Table 3: Super Host Percentage and Average Price in New York

Super Host?	Percentage	Average Price
f	0.86	314.1452
t	0.14	304.7707

Boston has higher rate of Super host than New York does. However, both cities don't have big difference of price in terms of super host or not, moreover, Super host leads to slightly higher price in Boston but lower price in New York. As a conclusion, I think price influence is decided by property information (room type, property type etc) but not host experience.

Model and model checking

I fit a multi-level model for each city. Here, I treat intercept as random

$$y_i^{price} = \mu + \gamma_{ji} + \delta_{ki} + e_i$$

$$\gamma_i \sim N(0, \sigma_\gamma^2)$$
$$\delta_k \sim N(0, \sigma_\delta^2)$$

```
# Multi-level model
# Boston
m1 <- lmer(formula = priceperperson ~ 1 + (1 | room_type) + (1 | property_type) + (1 | neighbourhood), d
# New York
m2 <- lmer(formula = priceperperson ~ 1 + (1 | room_type) + (1 | property_type) + (1 | neighbourhood), d
summary(m1) # summary of Boston Model
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: priceperperson ~ 1 + (1 | room_type) + (1 | property_type) +
##     (1 | neighbourhood)
##    Data: bos_list
##
## REML criterion at convergence: 47914.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5265 -0.7125 -0.0376  0.6674  3.3289
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  neighbourhood (Intercept)    56.01   7.484
##  property_type (Intercept)   808.68  28.437
##  room_type     (Intercept)  6105.28  78.136
##  Residual                  11621.78 107.804
## Number of obs: 3925, groups:
## neighbourhood, 30; property_type, 17; room_type, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   204.12      46.48   4.392
```

```
summary(m2) # summary of New York Model
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: priceperperson ~ 1 + (1 | room_type) + (1 | property_type) +
##     (1 | neighbourhood)
##    Data: nyc_list
##
## REML criterion at convergence: 448840.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5631 -0.7085 -0.0855  0.6568  3.3420
##
```

```
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  neighbourhood (Intercept)   879.7   29.66
##  property_type (Intercept)  4515.5   67.20
##  room_type     (Intercept) 13895.9  117.88
##  Residual                  32215.6  179.49
## Number of obs: 33938, groups:
## neighbourhood, 200; property_type, 27; room_type, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   341.41      70.17   4.866
```

```
#m1
sig2roomBOS = as.vector(VarCorr(m1)$room_type)
sig2propertyBOS = as.vector(VarCorr(m1)$property_type)
sig2neighborBOS = as.vector(VarCorr(m1)$neighbourhood )
sig2roomBOS
```

```
## [1] 6105.283
```

```
sig2propertyBOS
```

```
## [1] 808.6806
```

```
sig2neighborBOS
```

```
## [1] 56.01093
```

```
#m2
sig2roomNYC = as.vector(VarCorr(m2)$room_type)
sig2propertyNYC = as.vector(VarCorr(m2)$property_type)
sig2neighborNYC = as.vector(VarCorr(m2)$neighbourhood )
sig2roomNYC
```

```
## [1] 13895.95
```

```
sig2propertyNYC
```

```
## [1] 4515.502
```

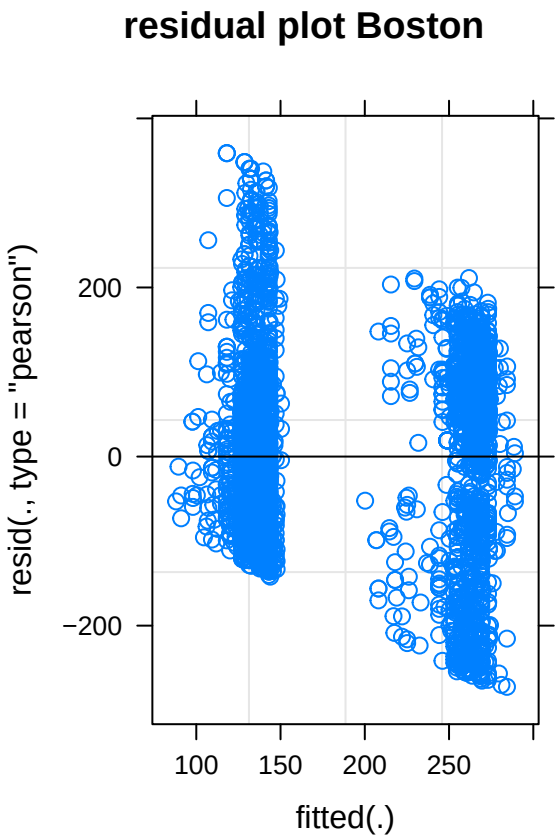
```
sig2neighborNYC
```

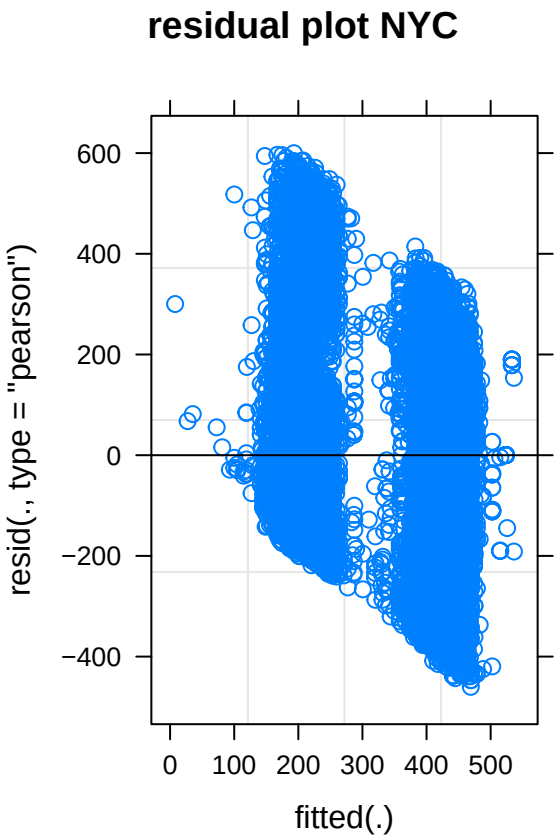
```
## [1] 879.6626
```

The first 1 is the fixed effect. The term (1 | variable) means that there is a random effect for each site and this effect is nested within the intercept (the whole model).

The estimated variance of Room Type in Boston model is 6105, for property type in Boston is 809, and for neighborhood in Boston is 56.

The estimated variance of Room Type in New York model is 11908, for property type in Boston is 2851, and for neighborhood in Boston is 801.





There are splited patterns in the residual plots.

Table 4: Room Type Fitted Value in Boston

	mean	fitted
Entire home/apt	134.8235	134.8936
Private room	262.7514	262.7032
Shared room	279.5426	277.4255

Table 5: Property Type Fitted Value in Boston

	mean	fitted
Apartment	174.23063	174.1119
Bed & Breakfast	207.28846	217.1552
Boat	98.73397	108.3156
Boutique hotel	246.90000	236.8341
Condominium	164.10794	163.6242
Dorm	114.83333	222.3484
Guest suite	156.06818	177.3600
Guesthouse	109.16667	188.5344
Hostel	124.50000	208.1235
House	234.77928	234.2348
In-law	235.90000	165.0269
Loft	128.43452	133.2851
Other	190.47346	181.8454
Serviced apartment	36.00000	110.6729

	mean	fitted
Timeshare	111.00000	236.8823
Townhouse	193.82026	192.0133
Villa	275.50000	228.0854

Table 6: Neighbourhood Fitted Value in Boston

	mean	fitted
Allston-Brighton	226.68017	223.9270
Back Bay	158.21063	154.5482
Beacon Hill	149.68996	150.5944
Brookline	177.25000	207.6484
Charlestown	174.39446	169.0645
Chelsea	30.33333	127.2623
Chestnut Hill	211.50000	201.6099
Chinatown	124.01307	141.6751
Dorchester	230.97495	229.7301
Downtown	145.11618	147.6771
Downtown Crossing	142.65357	139.7514
East Boston	206.64197	205.8052
Fenway/Kenmore	157.37011	161.8195
Financial District	210.44286	164.5644
Government Center	143.50000	136.8649
Hyde Park	230.91228	226.9589
Jamaica Plain	186.99951	191.6711
Leather District	183.86364	154.0340
Mattapan	246.09658	251.8736
Mission Hill	226.99726	226.3954
North End	136.87216	141.9282
Revere	99.00000	136.5878
Roslindale	232.08570	220.3651
Roxbury	223.35583	216.9020
Somerville	319.50000	263.3350
South Boston	155.03179	162.0207
South End	156.10601	156.7528
Theater District	127.47500	143.8136
West End	165.79808	154.2453
West Roxbury	217.73227	211.9764

```
summary(aov(priceperperson ~ Error(property_type), data=bos_list))
```

```
##
## Error: property_type
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Residuals 16 2555305  159707
##
## Error: Within
##              Df   Sum Sq Mean Sq F value Pr(>F)
## Residuals 3908 58836868   15055
```

```r
summary(aov(priceperperson ~ Error(room_type), data=bos_list))
```

```
##
## Error: room_type
##           Df   Sum Sq Mean Sq F value Pr(>F)
## Residuals  2 15319794 7659897
##
## Error: Within
##              Df   Sum Sq Mean Sq F value Pr(>F)
## Residuals 3922 46072380   11747
```

```r
summary(aov(priceperperson ~ Error(neighbourhood), data=bos_list))
```

```
##
## Error: neighbourhood
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Residuals 29 4981902  171790
##
## Error: Within
##              Df   Sum Sq Mean Sq F value Pr(>F)
## Residuals 3895 56410272   14483
```

```r
summary(aov(priceperperson ~ Error(property_type), data=nyc_list))
```

```
##
## Error: property_type
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Residuals 26 7568644  291102
##
## Error: Within
##               Df    Sum Sq Mean Sq F value Pr(>F)
## Residuals 33911 1.526e+09   44991
```

```r
summary(aov(priceperperson ~ Error(room_type), data=nyc_list))
```

```
##
## Error: room_type
##           Df    Sum Sq   Mean Sq F value Pr(>F)
## Residuals  2 416506849 208253424
##
## Error: Within
##               Df    Sum Sq Mean Sq F value Pr(>F)
## Residuals 33935 1.117e+09   32909
```

```r
summary(aov(priceperperson ~ Error(neighbourhood), data=nyc_list))
```

```
##
## Error: neighbourhood
##            Df   Sum Sq Mean Sq F value Pr(>F)
## Residuals 199 95451988  479658
##
## Error: Within
##               Df    Sum Sq Mean Sq F value Pr(>F)
## Residuals 33738 1.438e+09   42617
```

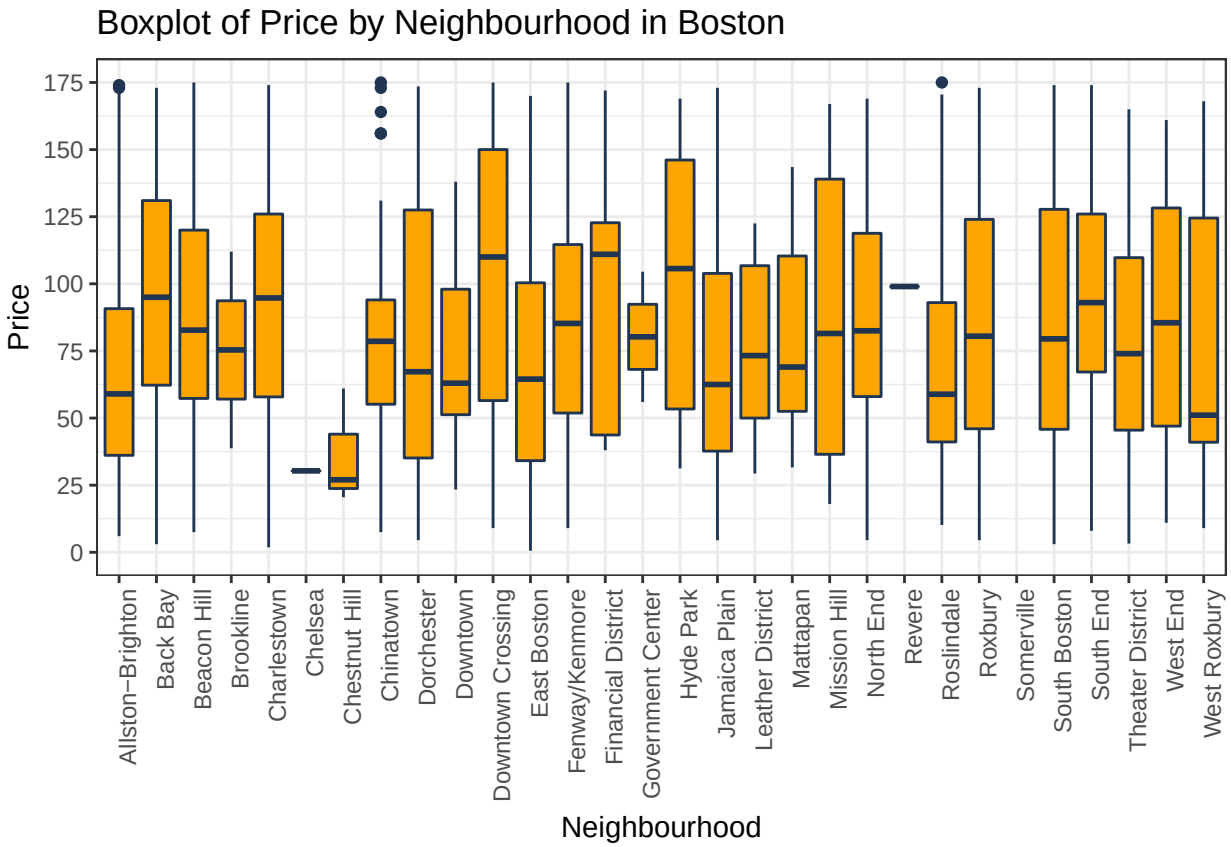
Discusstion and Limitation

In this project I am glad to find out that host experience doesn't have much influence on price prediction, but location, room types and property types do. My model are created based on the listed properties on Boston and New York. In general, the Boston model fits better than the New York one.

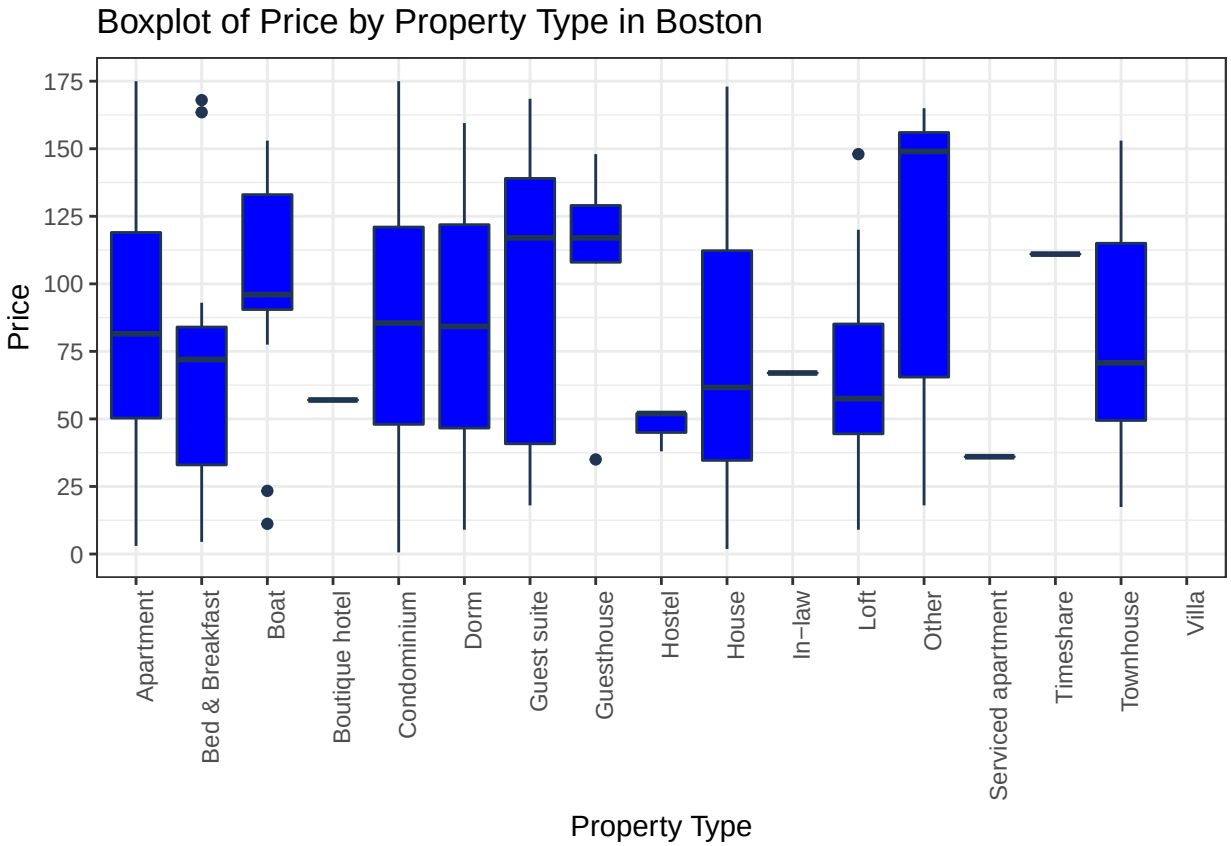
Limitations: In this project, I didn't take safety reason into consideration, though it is usually one of the main concerns when guests are looking for places to stay. Behind the varibales of location, there are convenience, popularity and safety reasons too, which I will need to use other resources such as walk score and tourist map. I also haven't solved the NA values in the New York model, which I hope I will complete an in-depth analysis of multilevel regression interpretion when I have time.

Appendix

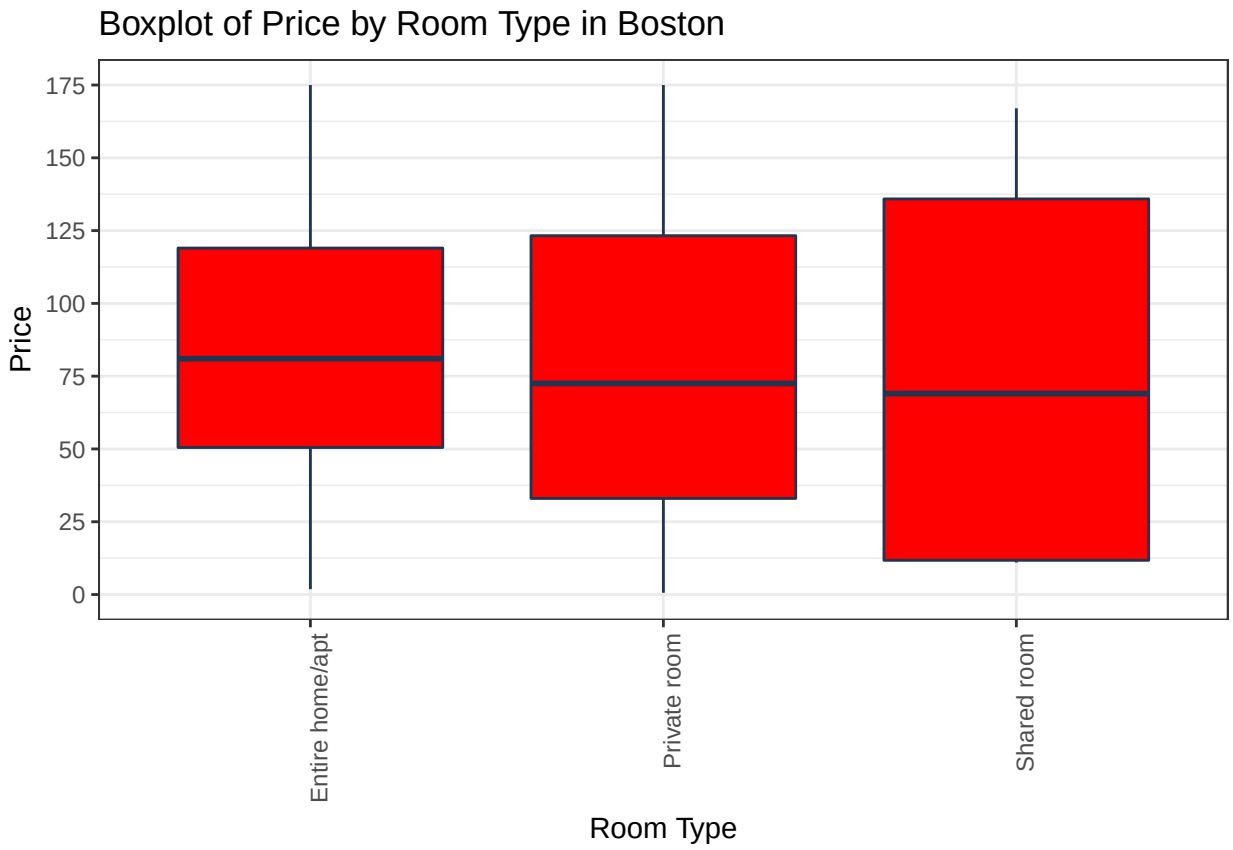
```
## Warning: Removed 1789 rows containing non-finite values (stat_boxplot).
```



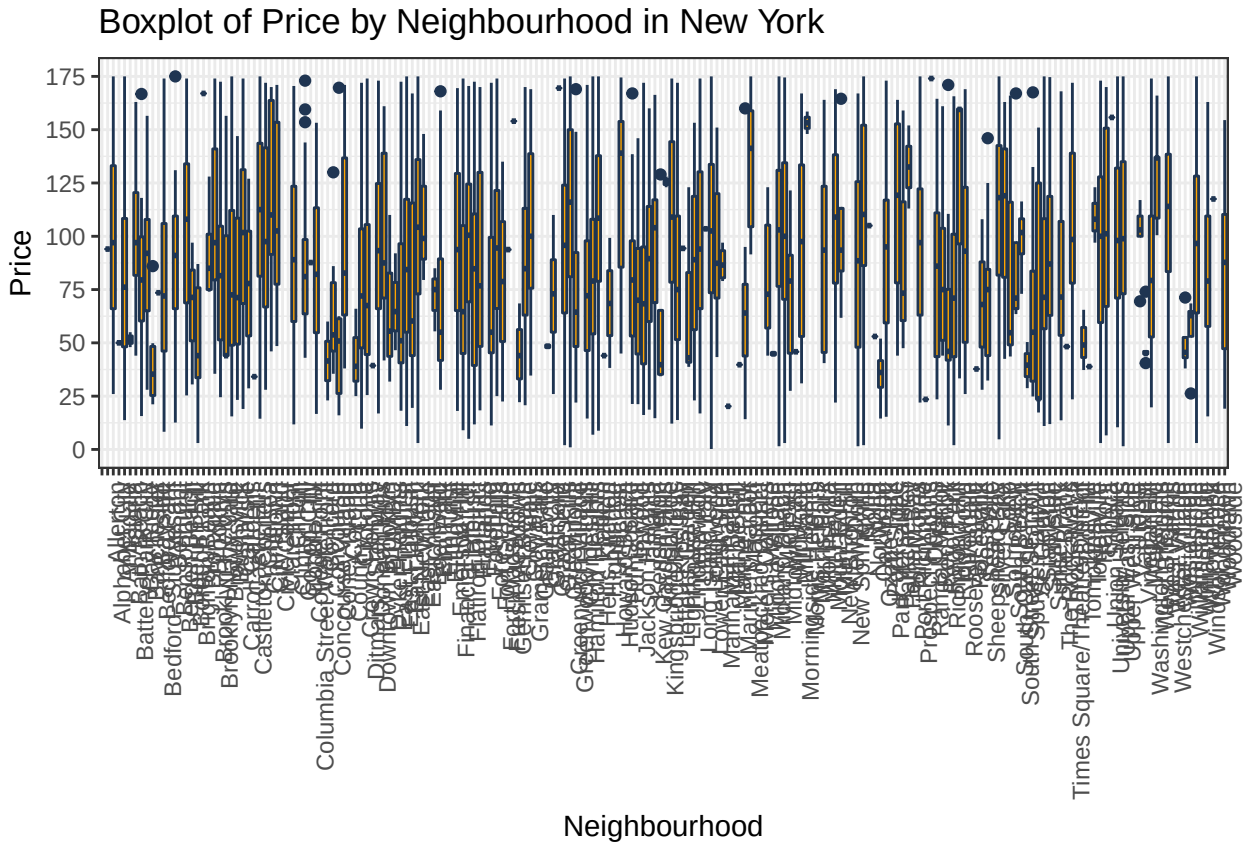

```
## Warning: Removed 1789 rows containing non-finite values (stat_boxplot).
```



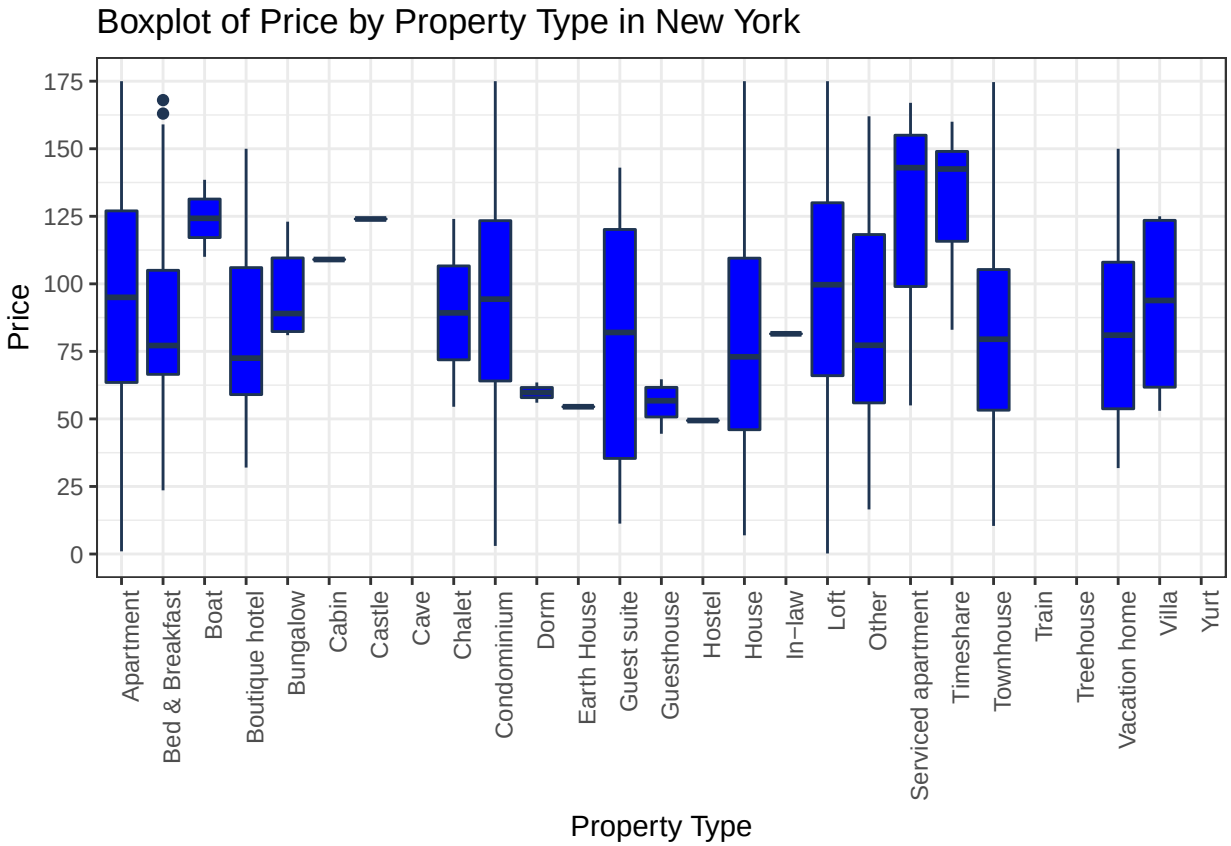

```
## Warning: Removed 1789 rows containing non-finite values (stat_boxplot).
```



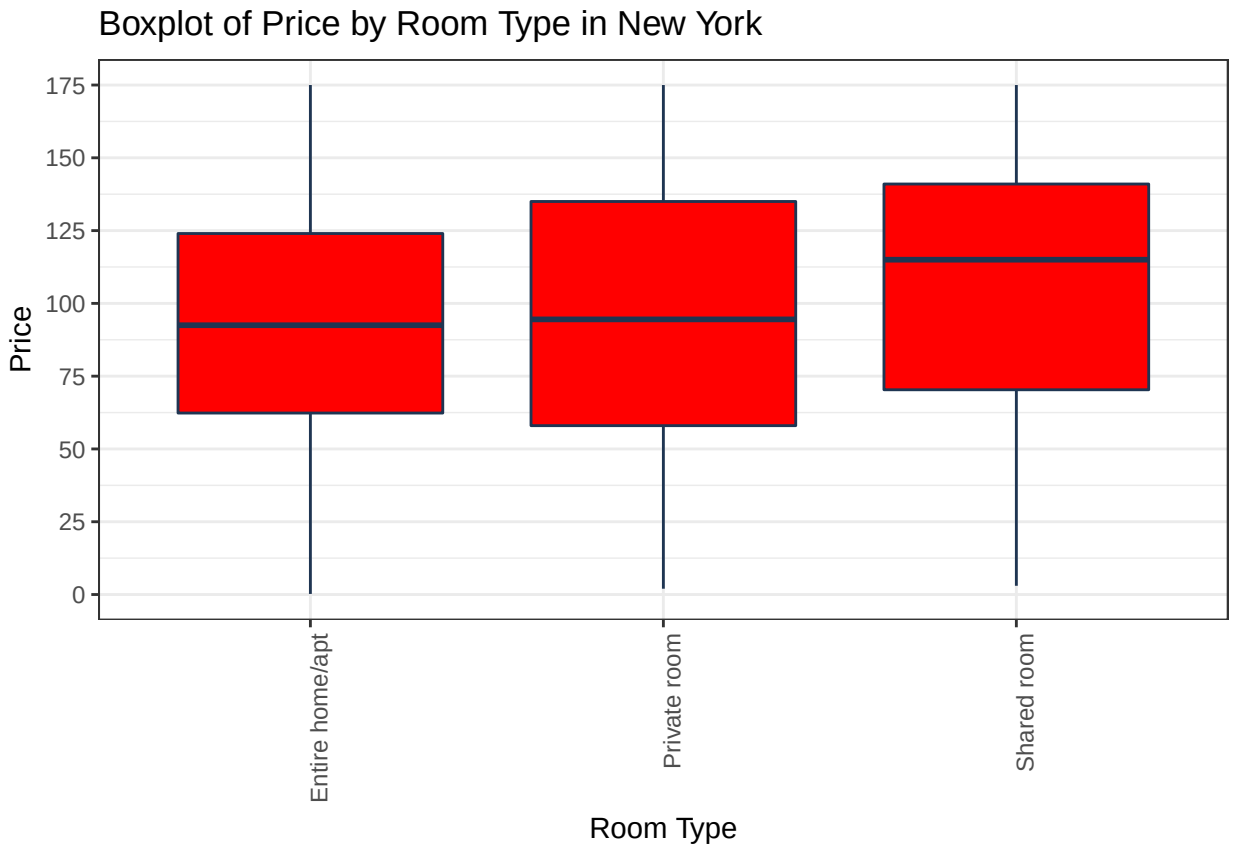

```
## Warning: Removed 21614 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 21614 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 21614 rows containing non-finite values (stat_boxplot).
```


Boxplot of Price by Room Type in New York
Price
175
150
125
100
75
50
25
0
Entire home/apt
Private room
Shared room
Room Type