New York Airbnb Analysis

Simu Huang

2020/12/8

Abstract

In this report, we use the "lm" and "lmer" model to determine if it is possible to predict the price of rentals on Airbnb in New York area, based on the location, room type and other related factors, and choose a more appropriate model from these two models to predict prices.

Introduction

For people looking for accommodation, Airbnb is a good choice because it offers a variety of options and is easy to use during travel. Therefore, for travelers, they want to know exactly what will affect the rent and how the rent will change in the future in order to better plan their travel plans.

In this research, we use the dataset containing Airbnb listings in New York State in 2020, including data on prices, rental attributes and locations, to discover whether it is possible to find out the key factors affecting prices and predict rental prices.

Methods

Data cleaning and selection

Since most of the available rooms in our data are located in the Manhattan and Brooklyn boroughs, we focus on the price of these two boroughs separately. And select all the data for 2020.

EDA

The figure in the appendix respectively show the price distribution of all available rooms on Airbnb in New York in 2020, the prices of different room types, and the relationship between monthly reviews and prices. We can roughly see that most rooms in Airbub's New York area cost between 50 and 100. Hotel rooms are usually more expensive, and the price range is larger than the other three types. Among the four types of rental housing, shared housing is cheaper. At the same time, there is no obvious linear relationship between the number of reviews and the room price, but we can still see that cheaper rooms have lower frequency of reviews for more expensive rooms.

From the figure, We can see that most of the rooms in Airbnb in New York area locate in the region between longitude -74 to -73.9 and latitude 40.6 to 40.8. And most of the rooms are entire home/apartment and private room.

Method

Modeling

The variables we use here are the `room_type`, `longitude`, `latitude`, `minimum_nights` (Minimum number of days per booking), `calculated_host_listings_count` (The number of properties owned by the host) and `price` (the room price in dollars).

First, we use the linear model here to find that, if there is a simple linear relationship between prices and other variables. In addition to location and type of the room, the price would go down as the number of minimum-stay increases.

The differences between room types can also affect prices so in this step we use the multilevel linear model here to find whether it works.

Validation

First we check the residual plots of the linear regression model. We can find that most of the points in the residual plot do not follow a straight line and are obviously not evenly distributed. In the residual plot of the linear regression model for Brooklyn, most of the points are concentrated around the fitted value 4.0 and 4.5 to 5.0. For Manhattan, the points in the residual plot are clustered between the 4.0 and 5.0. Therefore, it means that the simple linear model does not work here.

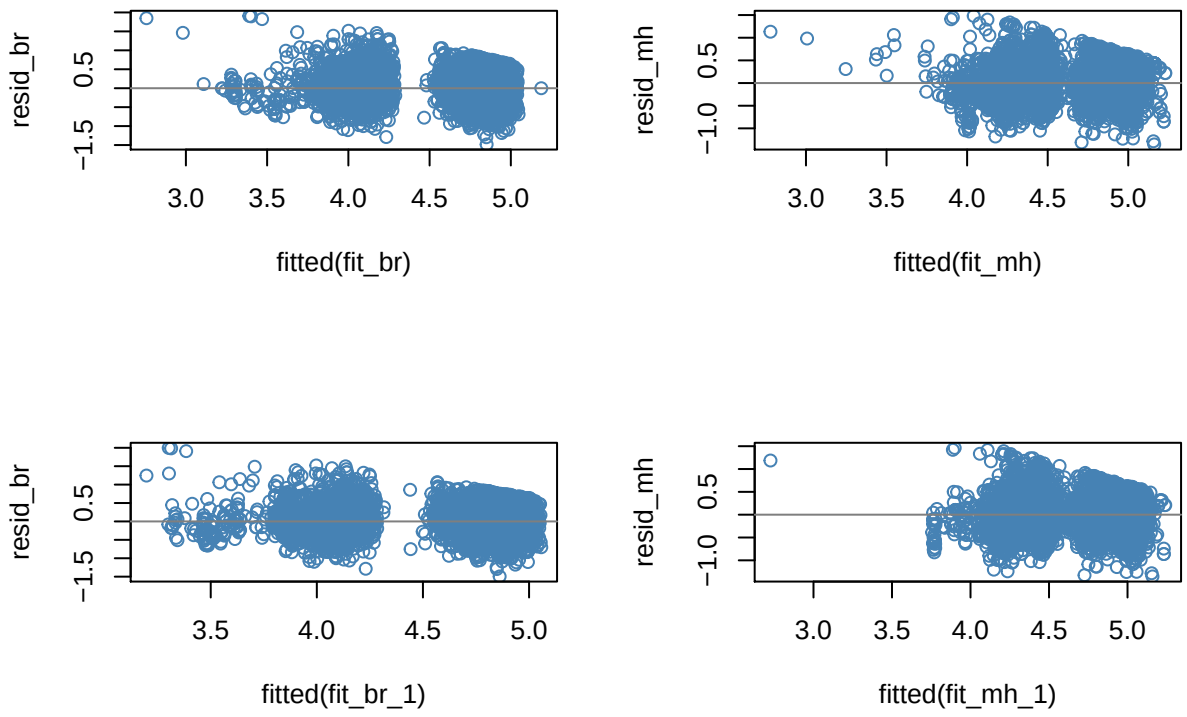


Figure 1: Residual Plot

Then we check the residual plots of the multilevel linear model. Compared to the linear regression result above, although the points in these plots are already more evenly distributed and we can roughly see that

they are close to the 0 line. But these points are clearly divided into many parts, and it also proves that the LMER model here cannot contain all the data perfectly.

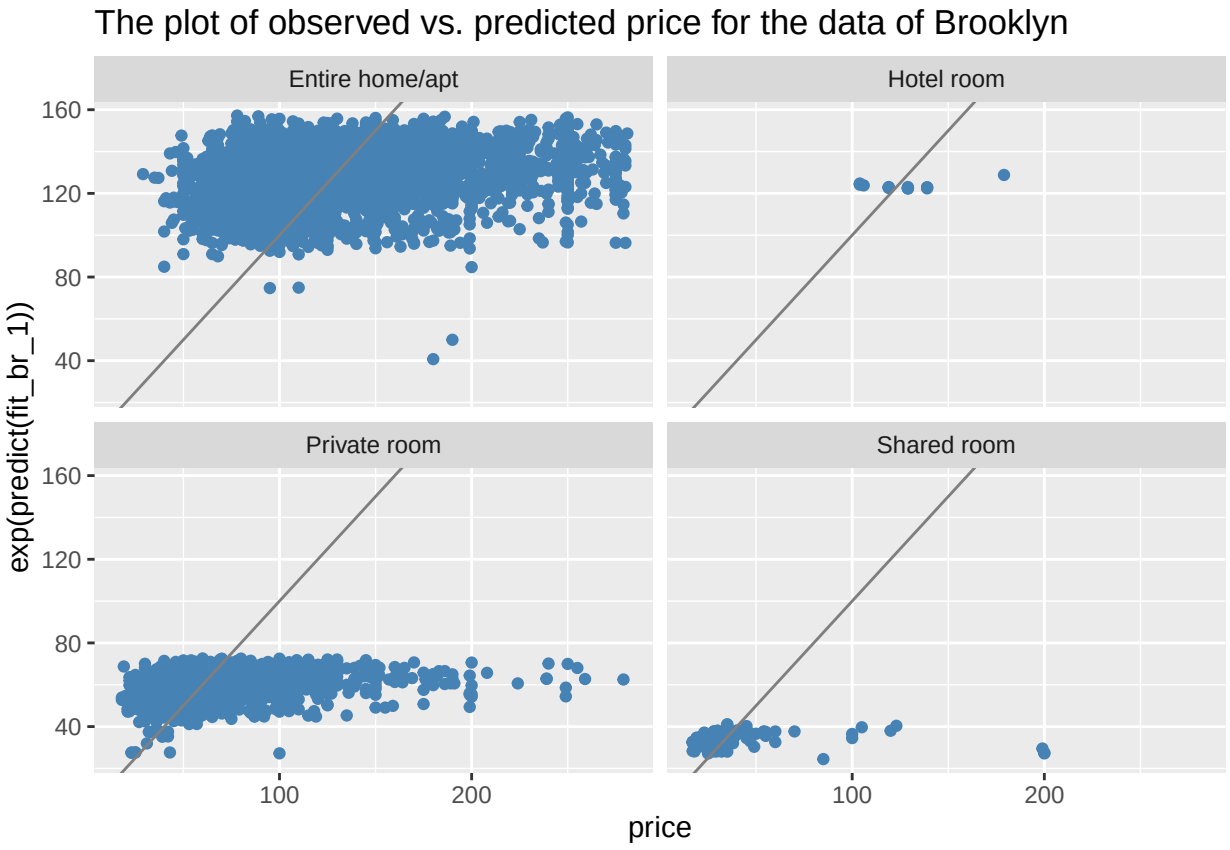
Result

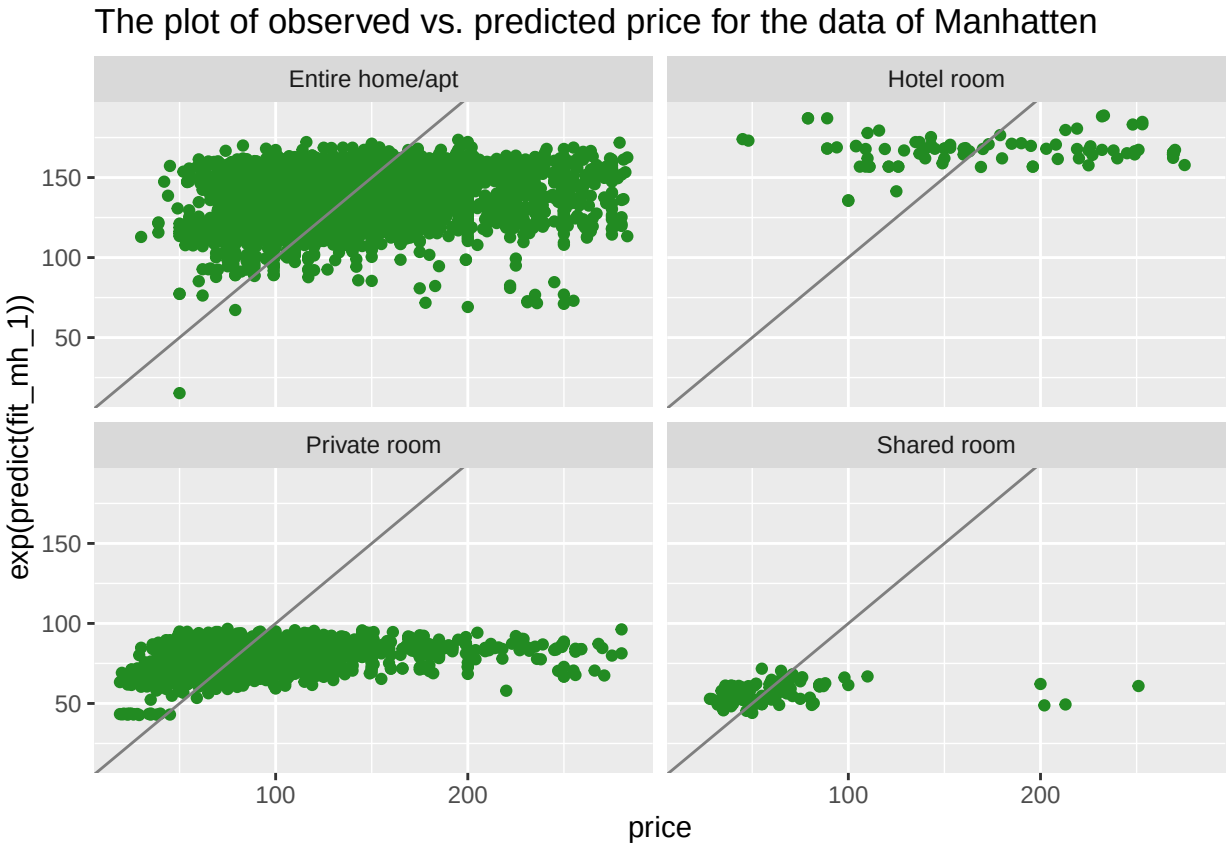
Estimation

In these two boroughs, one the number of minimum nights and the number of property each host own increase, the price will go down. And the availability in a year of the room and the number of times the room information was accessed have a slightly positive effect on the price. In the Brooklyn, for the room in the latitude 40.68 and 73.95, without minimum booking nights and can be booked all year around, the price of the entire home is 131.6 dollars, the price of hotel room is 121.6, the price of private room is 61.6, and the price of 37.8.

Predictions

The figure shows that many points are near a line with slope 1 and intercept 0, which means that this model captures some patterns well. But we can clearly see that the predicted prices are clustered in a restricted range. For example, for the private room, most of the predicted value is in the range 50 to 100. It may represent that we lose some features of the data.



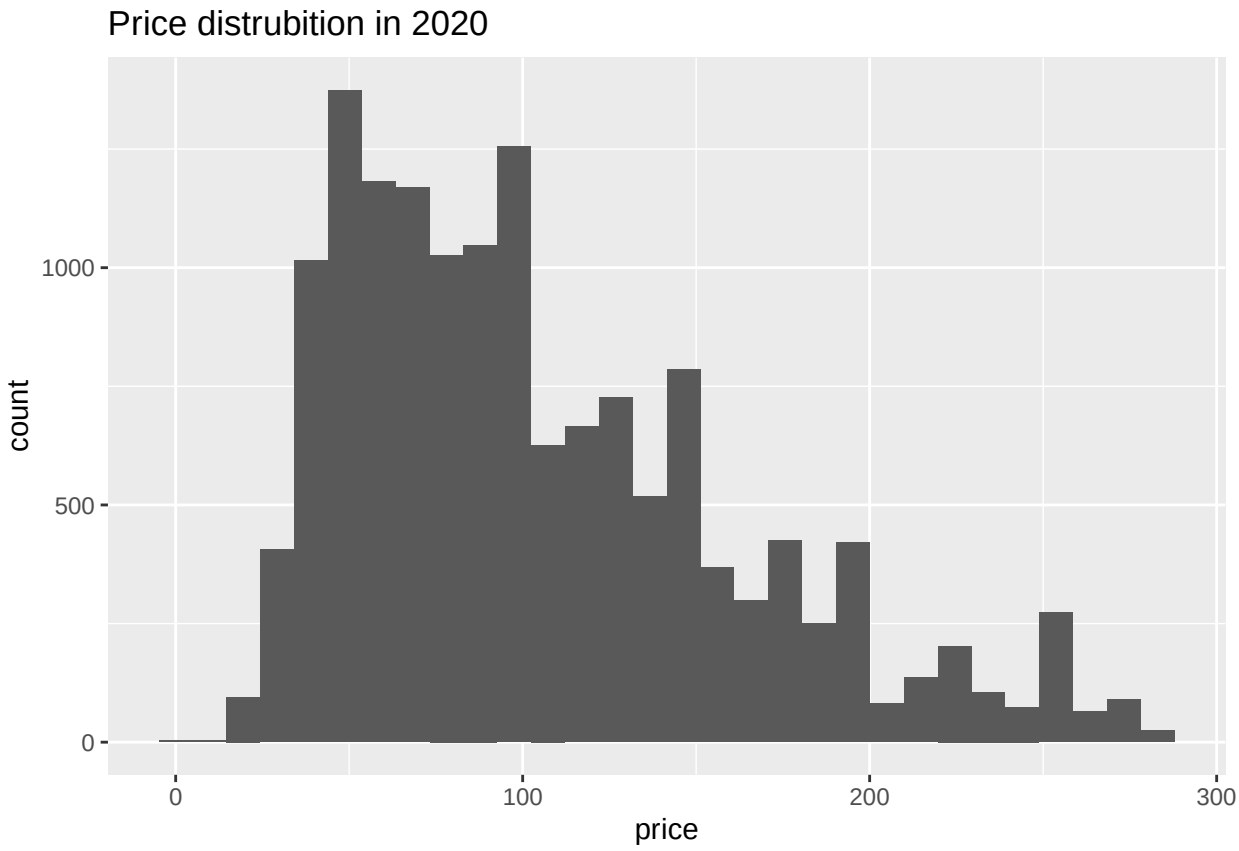


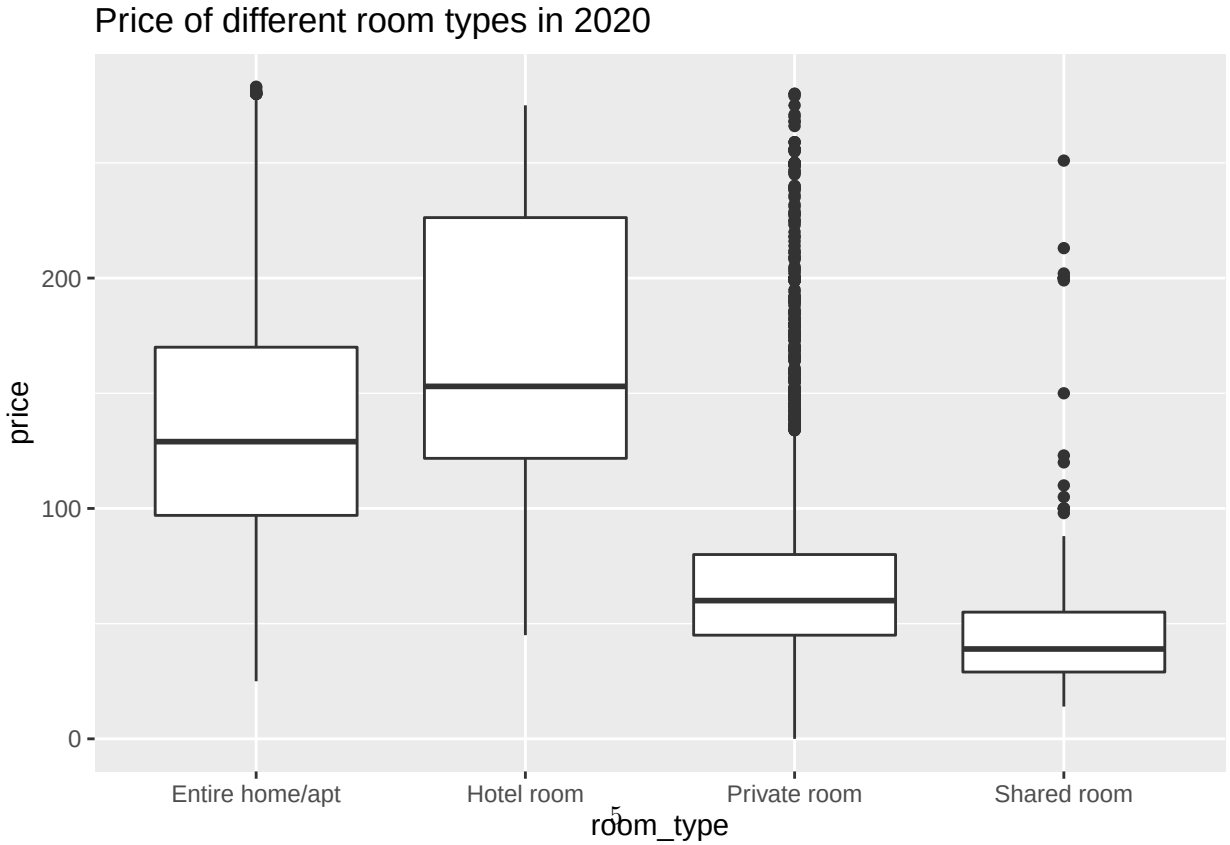
Discussion

From the modeling process above, we can find that the multilevel linear model is a relatively better model here. The type of room has the biggest impact on prices and different boroughs and locations also have an impact on prices. However, we cannot precisely predict the price based on the predictors that we have in the data. Because in the model, we only conclude the change of location, but what we should consider more is the distance between these rooms and key places in that borough, such as the high street, tourist attractions and public transportation. Besides, there may be other factors that affect prices, such as whether the booked time is a working day or a holiday, but this information does not conclude in the data. For future study, we need data with more detailed information.

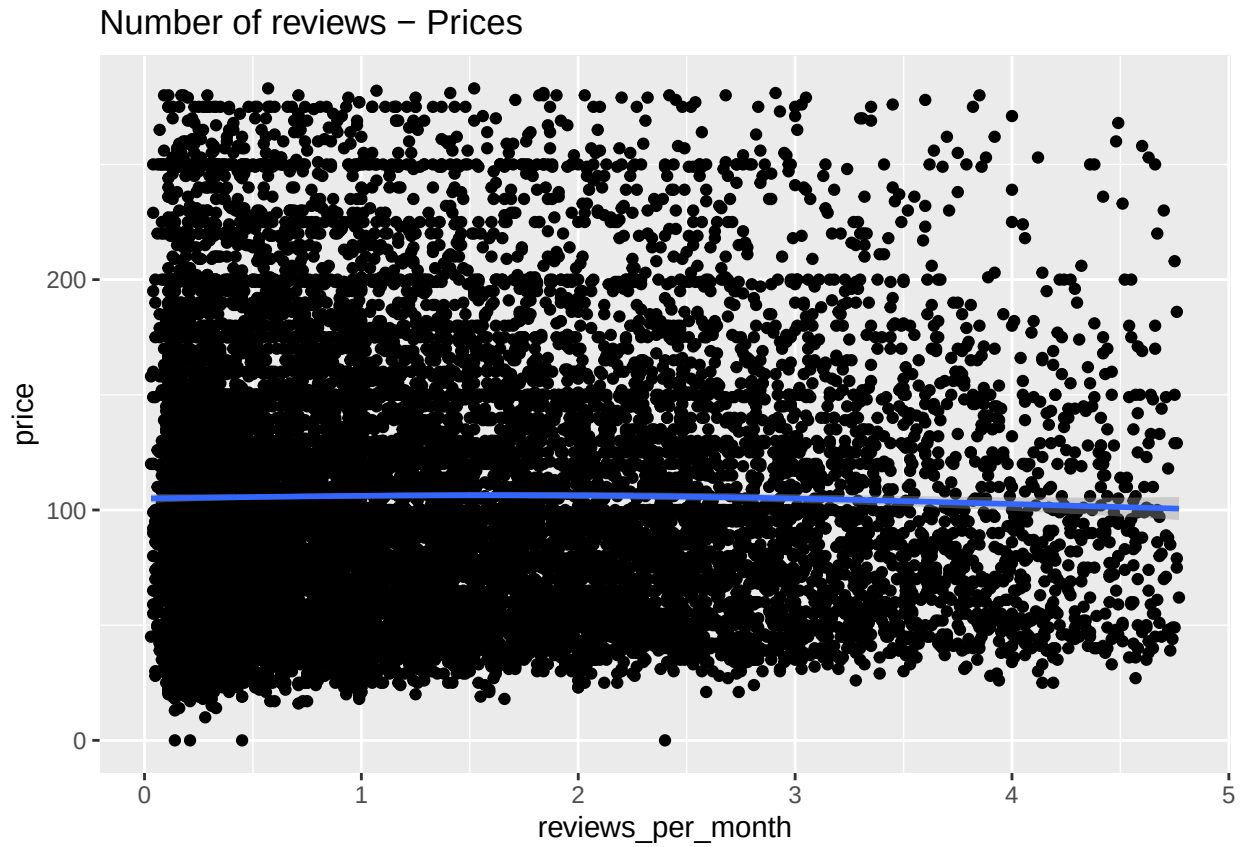
Appendix

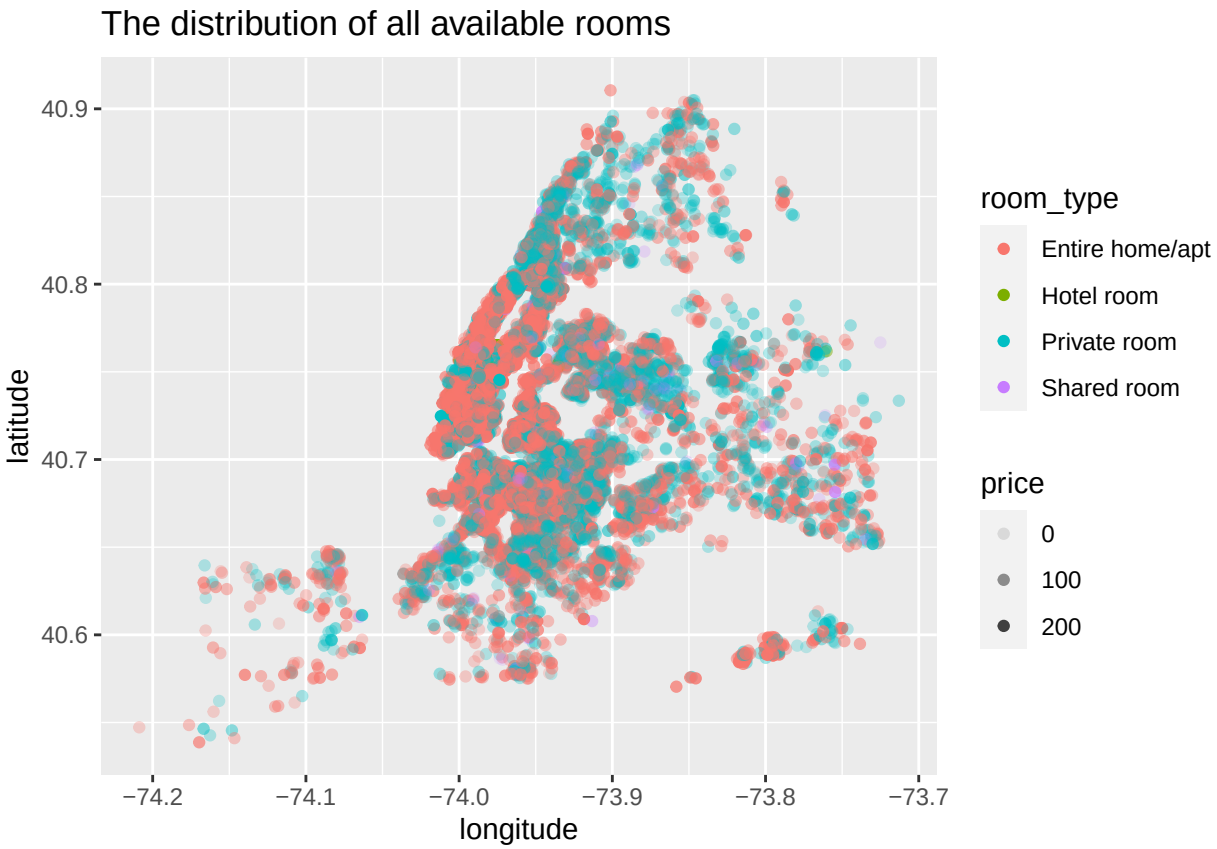
EDA





```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```





Reference

1: [Airbnb - New York]:(http://insideairbnb.com/get-the-data.html)