Project Proposal STAT501

2023-05-05

Names: Kien Le, Tony Nguyen

Working title: Airbnb Listing Dataset Statistical Analysis

Project Description:

Guiding question: What are the most critical features that affect the price of Airbnb listings in New York City, and how accurately can we predict the price using regression models?

Domain: Housing Price. The goal is to predict a continuous target variable (in this case, the price) based on a set of input variables (the other columns in the dataset). We will split the dataset into a training set and a testing set, where the training set is used to train the model, and the testing set is used to evaluate its performance. Once we have built and evaluated the model, we can use it to predict the price of new Airbnb listings in the future based on their features. These are our initial thoughts and they might change later on

The dataset we found on the website is through this link: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data and it is called the "New York City Airbnb Open Data".

Outline

1. Data cleaning and Pre-processing
2. Regression Analysis: used to identify the factors that affect the Airbnb rental prices in NYC

Introduction

```
data <- read.csv("airbnb_data.csv")
head(data)
```

```
##     id                                              name host_id   host_name
## 1 2539                Clean & quiet apt home by the park    2787        John
## 2 2595                             Skylit Midtown Castle    2845    Jennifer
## 3 3647               THE VILLAGE OF HARLEM....NEW YORK !    4632   Elisabeth
## 4 3831                    Cozy Entire Floor of Brownstone    4869 LisaRoxanne
## 5 5022 Entire Apt: Spacious Studio/Loft by central park    7192       Laura
## 6 5099         Large Cozy 1 BR Apartment In Midtown East    7322       Chris
##   neighbourhood_group neighbourhood latitude longitude       room_type price
## 1            Brooklyn    Kensington 40.64749 -73.97237    Private room   149
## 2           Manhattan       Midtown 40.75362 -73.98377 Entire home/apt   225
## 3           Manhattan        Harlem 40.80902 -73.94190    Private room   150
## 4            Brooklyn  Clinton Hill 40.68514 -73.95976 Entire home/apt    89
## 5           Manhattan   East Harlem 40.79851 -73.94399 Entire home/apt    80
```

```
## 6           Manhattan   Murray Hill 40.74767 -73.97500 Entire home/apt   200
##   minimum_nights number_of_reviews last_review reviews_per_month
## 1              1                 9  2018-10-19              0.21
## 2              1                45  2019-05-21              0.38
## 3              3                 0                            NA
## 4              1               270  2019-07-05              4.64
## 5             10                 9  2018-11-19              0.10
## 6              3                74  2019-06-22              0.59
##   calculated_host_listings_count availability_365
## 1                              6              365
## 2                              2              355
## 3                              1              365
## 4                              1              194
## 5                              1                0
## 6                              1              129
```

Data Cleaning and Pre-processing

```
# checking duplicated values
sum(duplicated(data))
```

```
## [1] 0
```

There is no duplicated value in our dataset

```
## [1] 10052
```

There are about 10k missing values in our dataset

```
# drop uneccessary columns which does not generate useful insights
data <- data [, -c(1,2,3,4,6,13)]
head(data)
```

```
##   neighbourhood_group latitude longitude       room_type price minimum_nights
## 1            Brooklyn 40.64749 -73.97237    Private room   149              1
## 2           Manhattan 40.75362 -73.98377 Entire home/apt   225              1
## 3           Manhattan 40.80902 -73.94190    Private room   150              3
## 4            Brooklyn 40.68514 -73.95976 Entire home/apt    89              1
## 5           Manhattan 40.79851 -73.94399 Entire home/apt    80             10
## 6           Manhattan 40.74767 -73.97500 Entire home/apt   200              3
##   number_of_reviews reviews_per_month calculated_host_listings_count
## 1                 9              0.21                              6
## 2                45              0.38                              2
## 3                 0                NA                              1
## 4               270              4.64                              1
## 5                 9              0.10                              1
## 6                74              0.59                              1
##   availability_365
## 1              365
## 2              355
```

```
## 3              365
## 4              194
## 5                0
## 6              129
```

```
# Check how many levels of each columns
unique(data$room_type)
```

```
## [1] "Private room"    "Entire home/apt" "Shared room"
```

```
unique(data$neighbourhood_group)
```

```
## [1] "Brooklyn"      "Manhattan"     "Queens"        "Staten Island"
## [5] "Bronx"
```

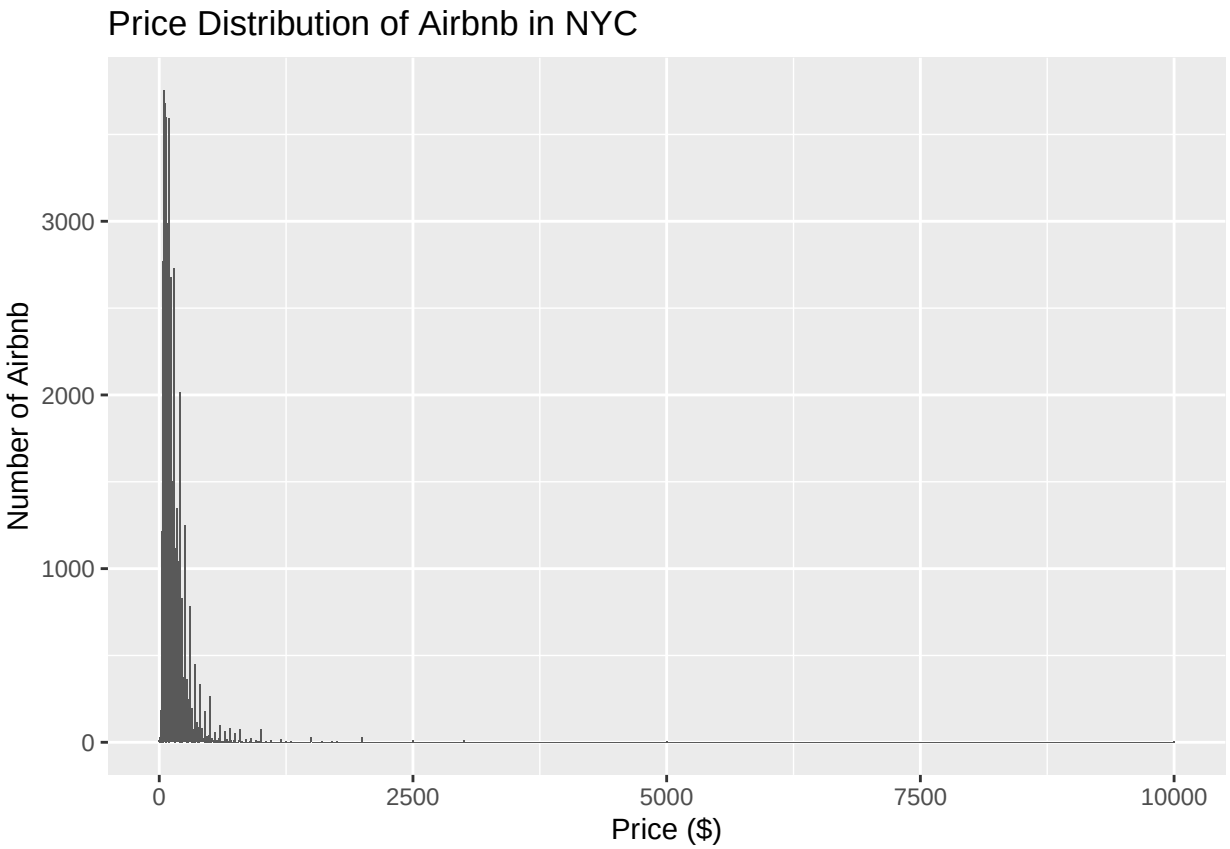
```
# Making these columns factors
data$neighbourhood_group <- factor(data$neighbourhood_group)
data$room_type <- factor(data$room_type)
```

```
# Fill out missing values
data$reviews_per_month <- ifelse(is.na(data$reviews_per_month), 0, data$reviews_per_month)
sum(is.na(data))
```

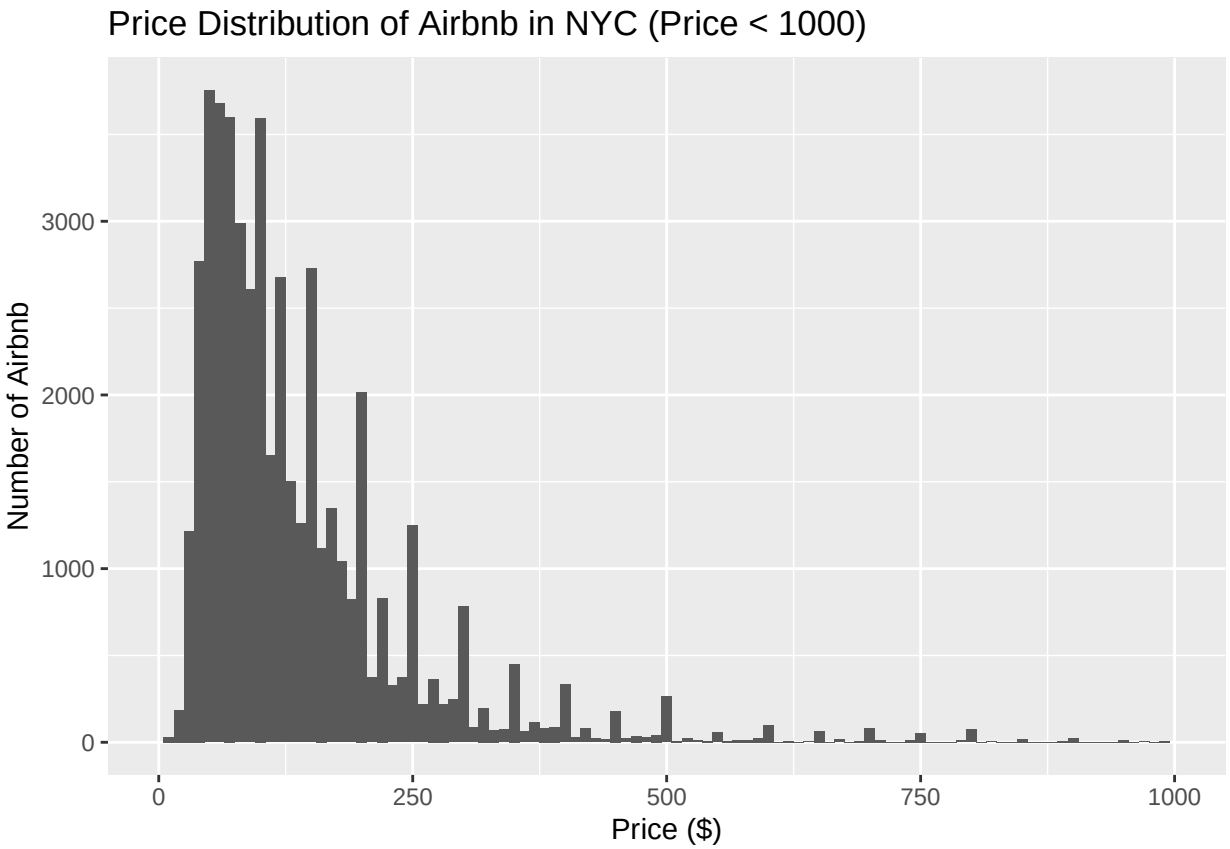
```
## [1] 0
```

Detecting any outliers

```
ggplot(data = data, aes(x = price)) +
    geom_histogram(binwidth = 10) +
    labs(x = "Price ($)", y = "Number of Airbnb", title = "Price Distribution of Airbnb in NYC")
```

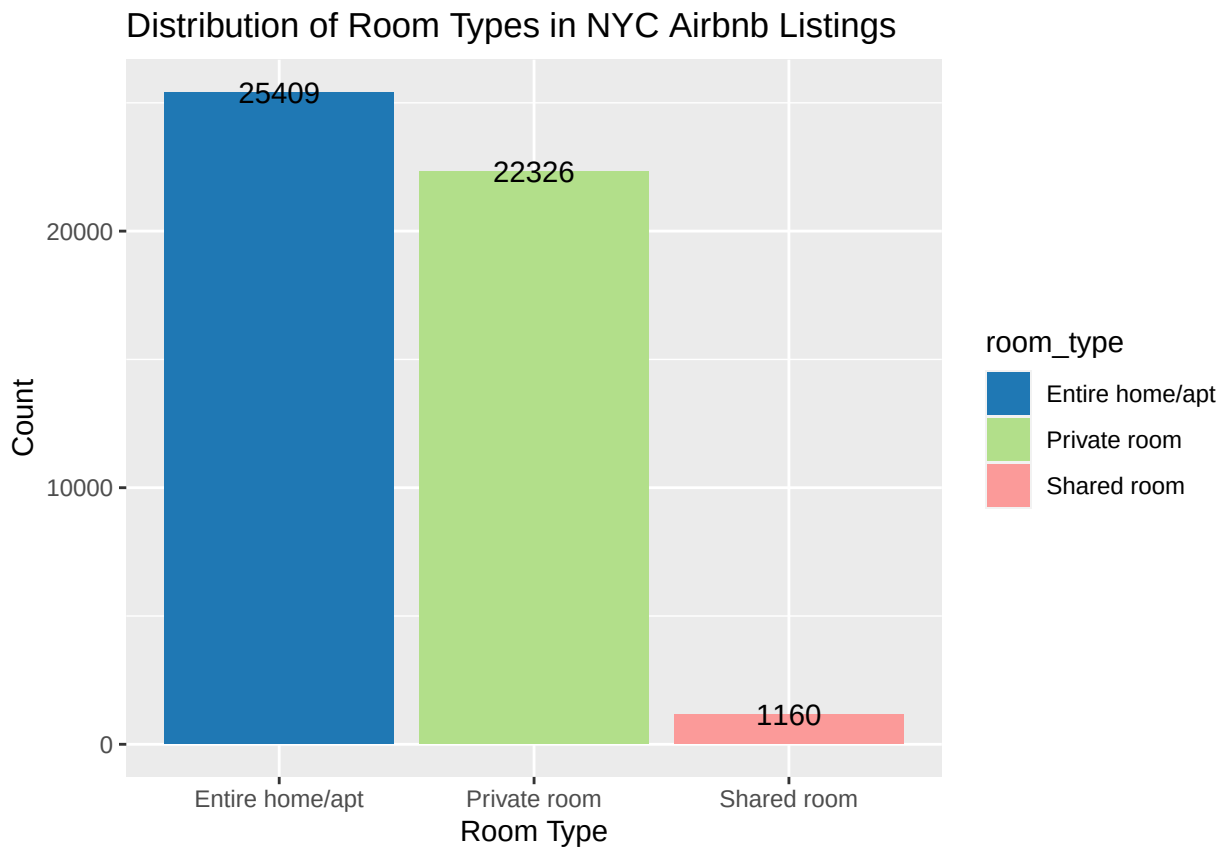



```
ggplot(data = data, aes(x = price)) +
    geom_histogram(binwidth = 10) +
    labs(x = "Price ($)", y = "Number of Airbnb", title = "Price Distribution of Airbnb in NYC (Price <
    xlim(0, 1000)
```



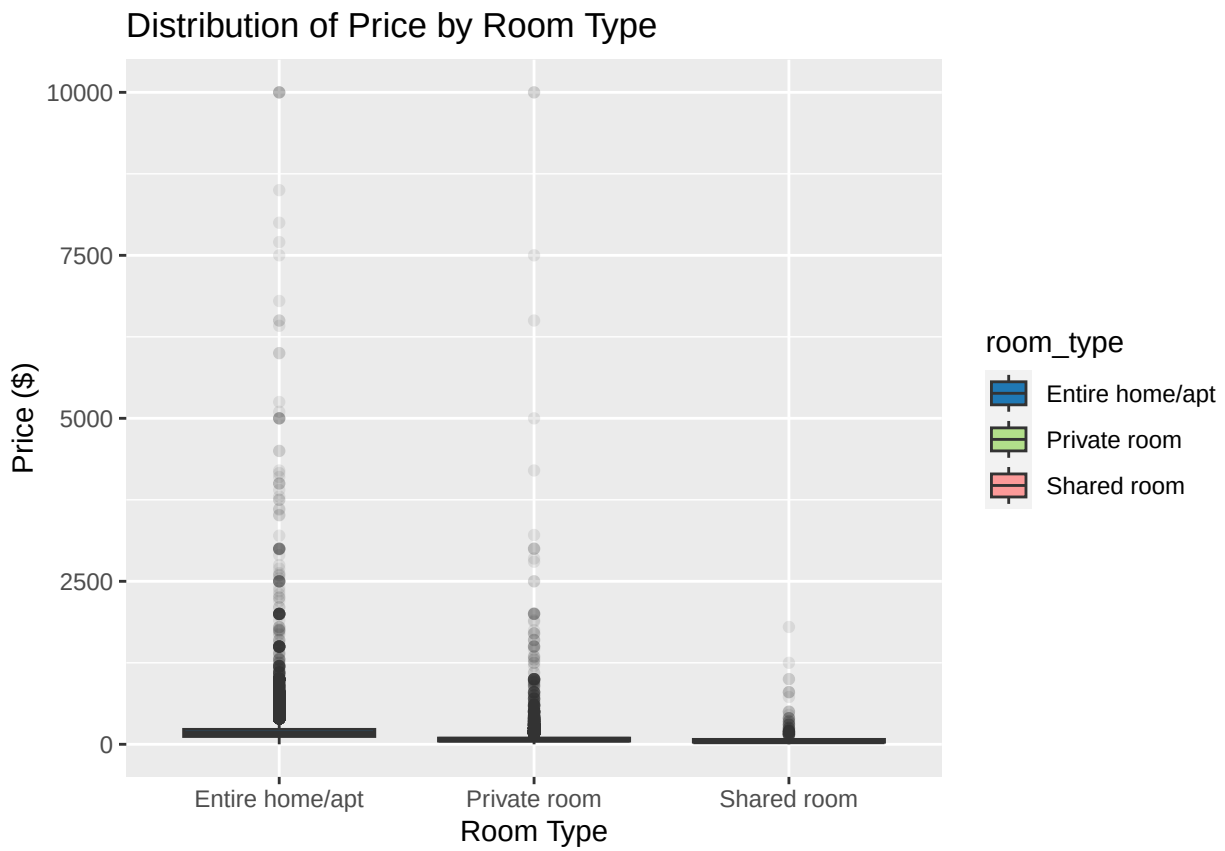
> The price range in NYC for airbnb is from 20 to more than 1000 per night, with the peak of about 100 dollars per night. I set the range of price to smaller than 1000 because the dataset is right-skewed so by doing that, it is easier to see the distribution.

```
#plot distribution of Room Types in NYC Airbnb Listings
ggplot(data = data, aes(x = room_type, fill = room_type)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label=..count..)) +
  labs(x = "Room Type", y = "Count", title = "Distribution of Room Types in NYC Airbnb Listings") +
  scale_fill_manual(values = c("#1f78b4", "#b2df8a", "#fb9a99"))
```

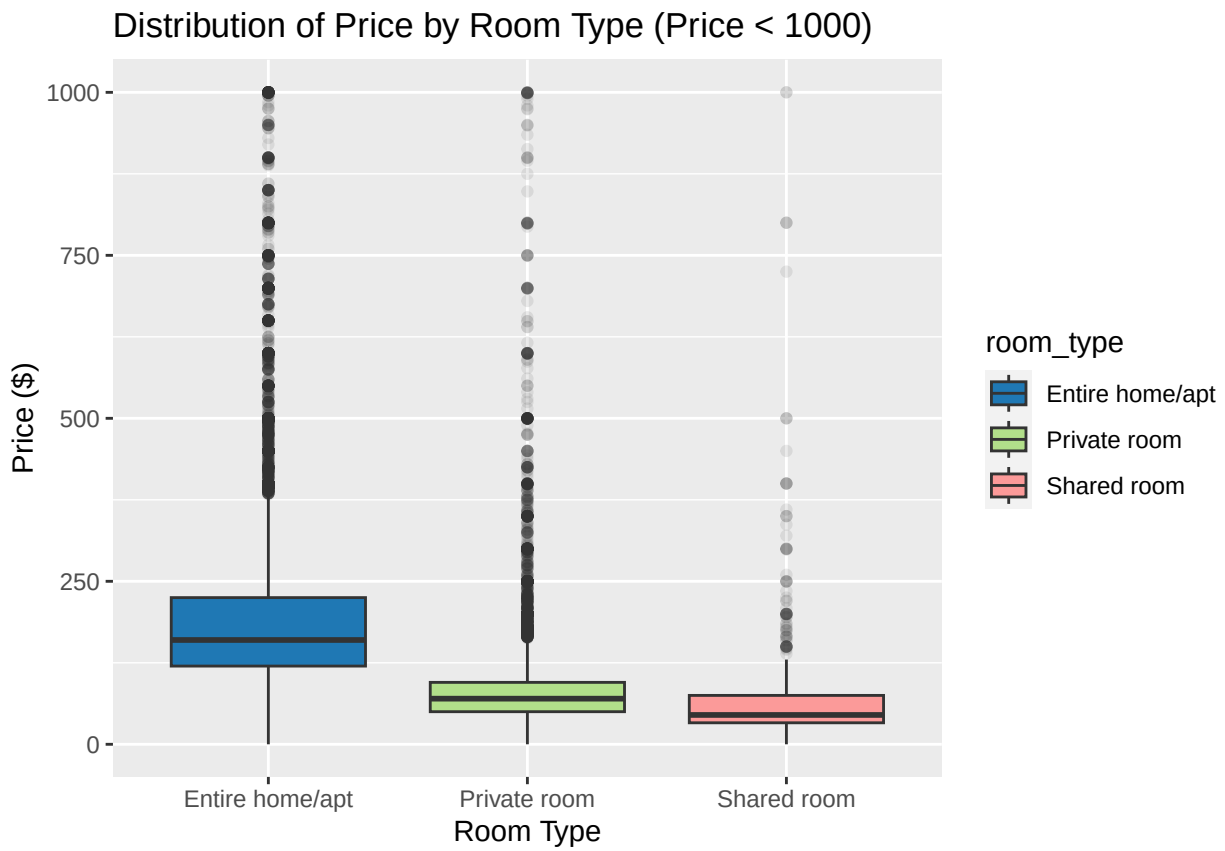


> In this dataset, most of the room type are "entire home/apt" and "private room".

```
#plot distribution of Room Types in NYC Airbnb Listings
ggplot(data = data, aes(x=room_type, y=price, fill=room_type)) +
  geom_boxplot(outlier.alpha=0.1) +
  labs(title="Distribution of Price by Room Type", x="Room Type", y="Price ($)") +
  scale_fill_manual(values=c("#1f78b4", "#b2df8a", "#fb9a99"))
```




```
#plot distribution of Room Types in NYC Airbnb Listings with price smaller than $1000
ggplot(data = data, aes(x=room_type, y=price, fill=room_type)) +
  geom_boxplot(outlier.alpha=0.1) +
  labs(title="Distribution of Price by Room Type (Price < 1000)", x="Room Type", y="Price ($)") +
  scale_fill_manual(values=c("#1f78b4", "#b2df8a", "#fb9a99")) +
  ylim(0, 1000)
```

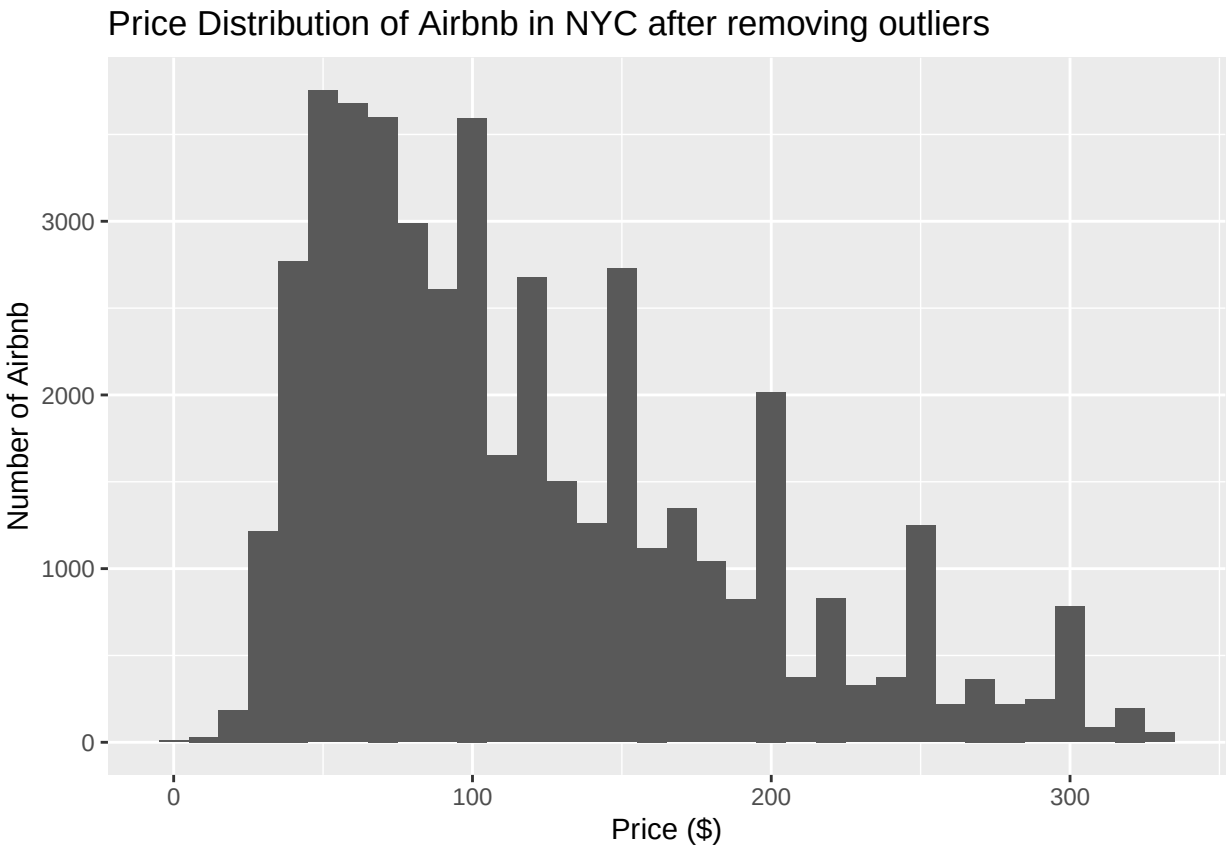


> There are a lot of outliers so I set the limit of y-axis to 1000 to have a better visualization of the dataset. We could see that the Shared room does not have much that is above $2000. However, the other have several that are approximately $10000

Remove outliers

```
#calculate IQR and quantile to remove outliers
quantile <- quantile(data$price, probs=c(.25, .75))
iqr <- IQR(data$price)
data_no_OL <- data %>% filter(price > (quantile[1] - 1.5*iqr) & price < (quantile[2] + 1.5*iqr))

#Plot the dataset without the outliers
ggplot(data = data_no_OL, aes(x = price)) +
    geom_histogram(binwidth = 10) +
    labs(x = "Price ($)", y = "Number of Airbnb", title = "Price Distribution of Airbnb in NYC after re
```



> This is the dataset after we remove the outliers

Regression Model

```
# Build a linear regression model using the no outliers data
model <- lm(price ~ neighbourhood_group + latitude + longitude + minimum_nights + room_type + number_of_
summary(model)
```

```
##
## Call:
## lm(formula = price ~ neighbourhood_group + latitude + longitude +
##     minimum_nights + room_type + number_of_reviews + availability_365 +
##     calculated_host_listings_count, data = data_no_OL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -175.376  -30.553   -7.956   20.769  268.226
##
## Coefficients:
##                                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                    -1.711e+04  7.009e+02  -24.409  < 2e-16 ***
## neighbourhood_groupBrooklyn    -8.116e+00  1.913e+00   -4.242 2.22e-05 ***
## neighbourhood_groupManhattan    2.222e+01  1.728e+00   12.857  < 2e-16 ***
## neighbourhood_groupQueens       4.958e+00  1.836e+00    2.701  0.00692 **
## neighbourhood_groupStaten Island -7.837e+01  3.640e+00  -21.528  < 2e-16 ***
```

```
## latitude                          -7.523e+01  6.889e+00  -10.921  < 2e-16 ***
## longitude                         -2.748e+02  7.880e+00  -34.872  < 2e-16 ***
## minimum_nights                    -2.134e-01  1.183e-02  -18.034  < 2e-16 ***
## room_typePrivate room             -7.563e+01  4.768e-01 -158.622  < 2e-16 ***
## room_typeShared room              -1.009e+02  1.497e+00  -67.396  < 2e-16 ***
## number_of_reviews                 -5.044e-02  5.203e-03   -9.693  < 2e-16 ***
## availability_365                   5.770e-02  1.900e-03   30.365  < 2e-16 ***
## calculated_host_listings_count     9.309e-02  7.844e-03   11.867  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49 on 45905 degrees of freedom
## Multiple R-squared:  0.4826, Adjusted R-squared:  0.4825
## F-statistic:  3569 on 12 and 45905 DF,  p-value: < 2.2e-16
```

The linear regression model on data without outliers has an R-squared value of 0.4825, which means that the model explains approximately 48.25% of the variability in the target variable (price). Some of the coefficient can be interpreted as follow: Airbnb listings with a private room have an expected price that is approximately $75.63 lower than listings with the reference room type (entire home/apt), holding all other variables constant. Airbnb listings with a shared room have an expected price that is approximately $100.90 lower than listings with the reference room type (entire home/apt), holding all other variables constant.

Model with prediction

```r
# Split the data into a training set and a testing set
set.seed(123)
split <- sample.split(data_no_OL, SplitRatio = 0.8)
train <- subset(data_no_OL, split == TRUE)
test <- subset(data_no_OL, split == FALSE)

# Use the model to predict prices on the testing set
predictions <- predict(model, newdata = test)

# Calculate the root mean squared error (RMSE) to evaluate the performance of the model
rmse <- sqrt(mean((test$price - predictions)^2))
print(rmse)
```

```
## [1] 48.73927
```

First, we create a training set and test set. The training set has 80% of the data and the test set has 20%. THen we Create predictions fr the model. The printed result is the root mean squared error (RMSE) of the model and it is 49.84541, which means that the average difference between the predicted price and the actual price on the testing set is around $49. This indicates that the model is relatively accurate in predicting the prices of Airbnb listings based on the given features.