R Project Airbnb

Shakshi Kumari Agrawal

2024-07-07

R Markdown

```
# Data Importing
listings <- read.csv("C:/Users/ishak/Downloads/listing_r.csv")
View(listings)
```

##Codes with outputs

```
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("tidyr")
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library("lubridate")
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
##
## Attaching package: 'lubridate'
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

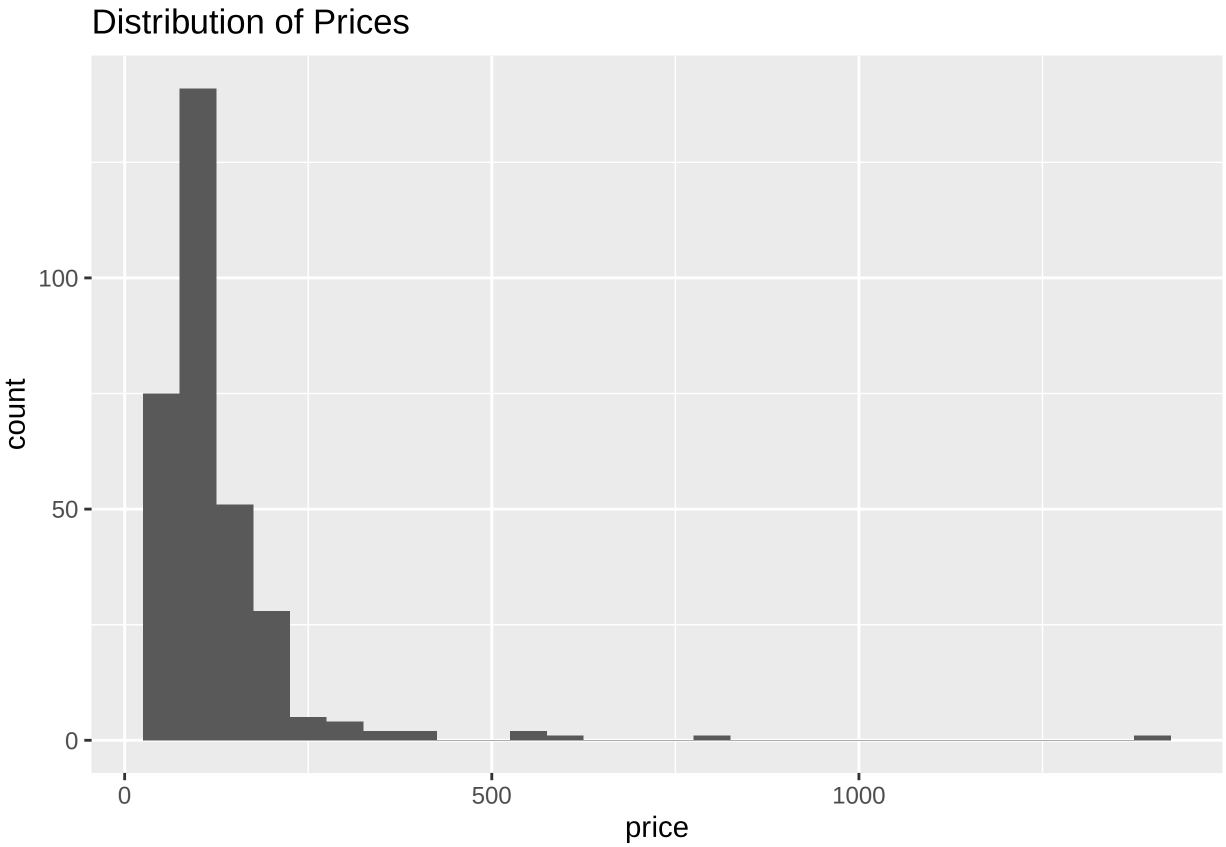
```
colSums(is.na(listings)) # N/A are in -> price, reviews_per_month
```

```
##                             id                           name
##                              0                              0
##                        host_id                      host_name
##                              0                              0
##                  neighbourhood                       latitude
##                              0                              0
##                      longitude                      room_type
##                              0                              0
##                          price                 minimum_nights
##                             50                              0
##              number_of_reviews                    last_review
##                              0                              0
##              reviews_per_month calculated_host_listings_count
##                             45                              0
##               availability_365          number_of_reviews_ltm
##                              0                              0
```

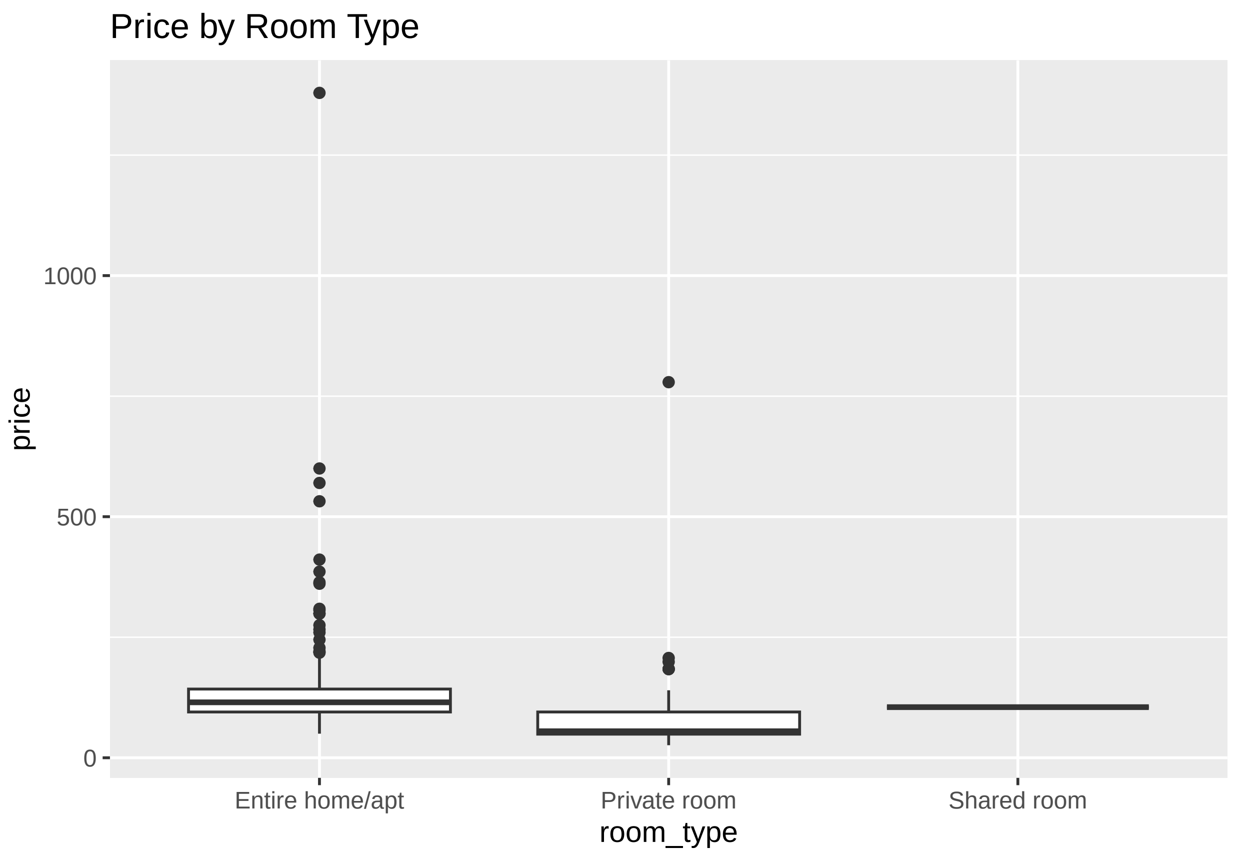
```
# Replace NA in reviews_per_month with 0
listings <- mutate(listings,reviews_per_month = ifelse(is.na(reviews_per_month), 0, reviews_per_month))

# Converting last_review as date format
listings <- mutate(listings,last_review = parse_date_time(last_review, orders = c("Ymd", "Y-m-d", "dmy"

# Drop rows with missing values in other crucial columns
listings <- drop_na(listings,price, minimum_nights, room_type, latitude, longitude,last_review)

# Transforming data
listings <- mutate(listings,room_type = as.factor(room_type))
listings <- mutate(listings,reviews_per_year = reviews_per_month * 12)

# Exploratory Data Analysis
summary(listings)
```

```
##        id                name              host_id           host_name
##  Min.   :1.489e+06   Length:313         Min.   :   649068   Length:313
##  1st Qu.:4.733e+07   Class :character   1st Qu.: 47625981   Class :character
##  Median :6.686e+17   Mode  :character   Median :172539578   Mode  :character
##  Mean   :5.052e+17                      Mean   :215274022
##  3rd Qu.:9.022e+17                      3rd Qu.:382970529
##  Max.   :1.147e+18                      Max.   :552465537
##  neighbourhood         latitude       longitude                room_type
##  Length:313         Min.   :42.63   Min.   :-73.83   Entire home/apt:231
##  Class  :character  1st Qu.:42.65   1st Qu.:-73.79   Private room   : 81
##  Mode   :character  Median :42.66   Median :-73.77   Shared room    :  1
##                     Mean   :42.66   Mean   :-73.78
##                     3rd Qu.:42.66   3rd Qu.:-73.76
##                     Max.   :42.69   Max.   :-73.74
##      price        minimum_nights   number_of_reviews
##  Min.   :  26.0   Min.   : 1.000   Min.   :  1.00
##  1st Qu.:  79.0   1st Qu.: 1.000   1st Qu.:  8.00
##  Median : 105.0   Median : 1.000   Median : 29.00
##  Mean   : 125.3   Mean   : 3.489   Mean   : 70.32
##  3rd Qu.: 136.0   3rd Qu.: 2.000   3rd Qu.: 84.00
##  Max.   :1379.0   Max.   :30.000   Max.   :795.00
##   last_review                     reviews_per_month
##  Min.   :2018-07-27 00:00:00.00   Min.   : 0.040
##  1st Qu.:2024-02-19 00:00:00.00   1st Qu.: 0.690
##  Median :2024-04-14 00:00:00.00   Median : 1.720
##  Mean   :2024-02-05 05:17:26.65   Mean   : 2.283
##  3rd Qu.:2024-04-26 00:00:00.00   3rd Qu.: 3.160
##  Max.   :2024-05-06 00:00:00.00   Max.   :11.050
##  calculated_host_listings_count availability_365 number_of_reviews_ltm
##  Min.   : 1.00                  Min.   :  3.0    Min.   :  0.00
##  1st Qu.: 1.00                  1st Qu.:126.0    1st Qu.:  4.00
##  Median : 3.00                  Median :243.0    Median : 14.00
##  Mean   : 5.54                  Mean   :220.4    Mean   : 21.21
##  3rd Qu.: 9.00                  3rd Qu.:318.0    3rd Qu.: 29.00
##  Max.   :22.00                  Max.   :365.0    Max.   :131.00
##  reviews_per_year
##  Min.   :  0.48
##  1st Qu.:  8.28
##  Median : 20.64
##  Mean   : 27.40
##  3rd Qu.: 37.92
##  Max.   :132.60
```

```r
# Visualizations
ggplot(listings, aes(x = price)) +
  geom_histogram(binwidth = 50) +
  ggtitle("Distribution of Prices")
```



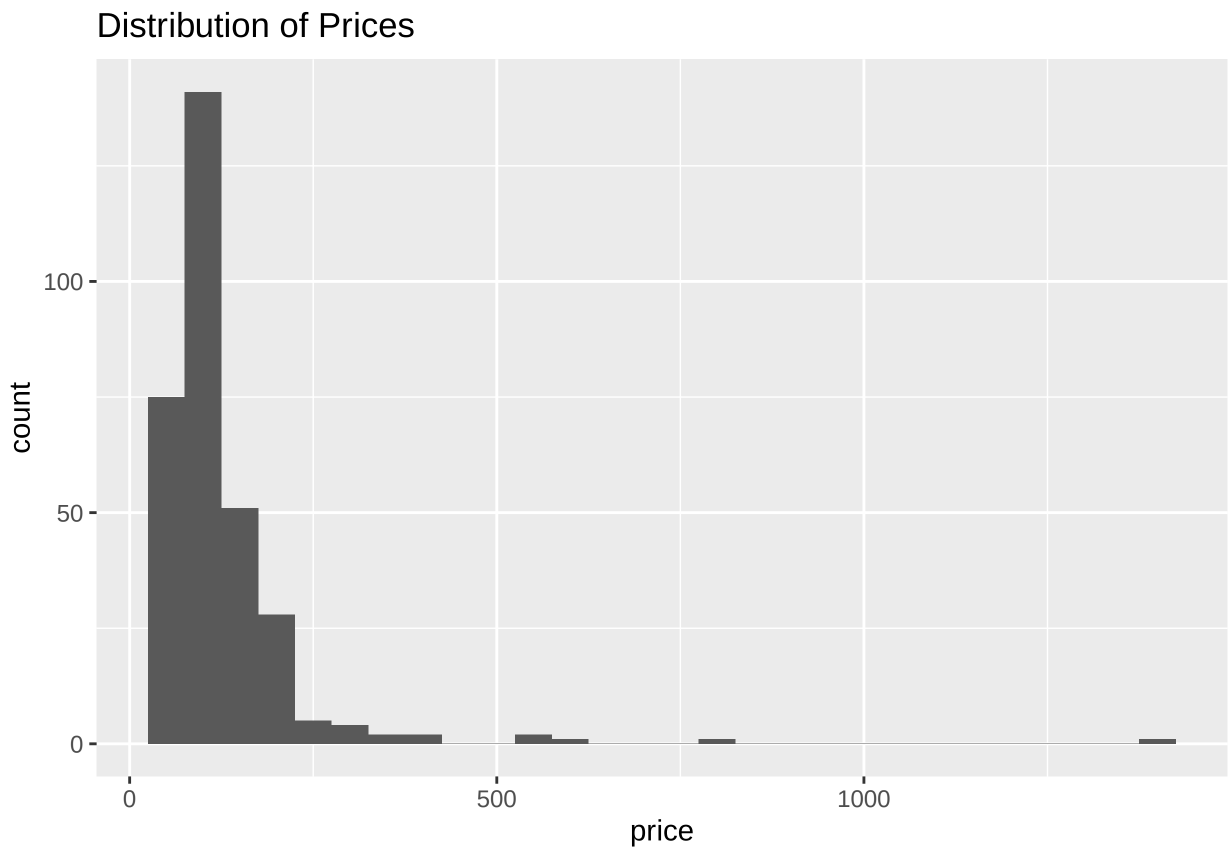
```
ggplot(listings, aes(x = room_type, y = price)) +
  geom_boxplot() +
  ggtitle("Price by Room Type")
```



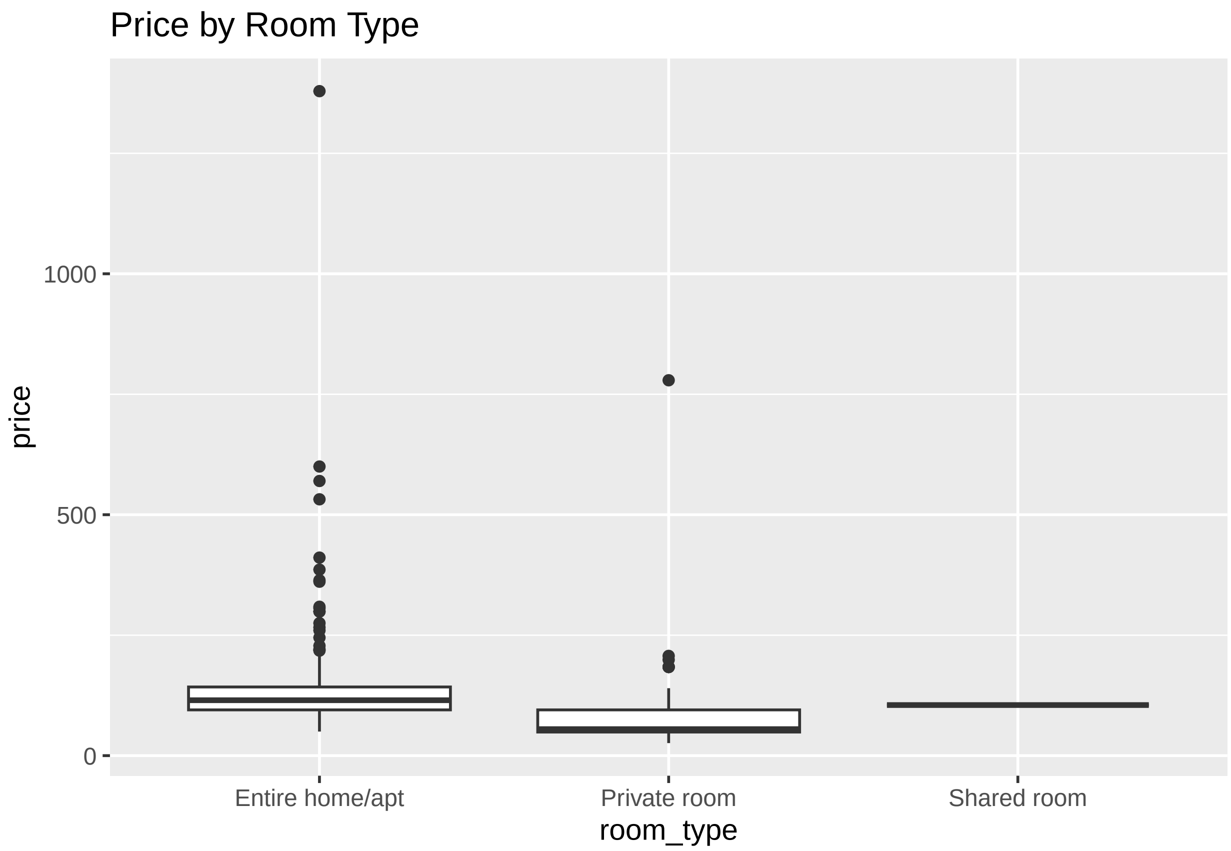
```
# Exploratory Data Analysis
summary(listings)
```

```
##        id                 name              host_id           host_name
##  Min.   :1.489e+06   Length:313         Min.   :   649068   Length:313
##  1st Qu.:4.733e+07   Class :character   1st Qu.: 47625981   Class :character
##  Median :6.686e+17   Mode  :character   Median :172539578   Mode  :character
##  Mean   :5.052e+17                      Mean   :215274022
##  3rd Qu.:9.022e+17                      3rd Qu.:382970529
##  Max.   :1.147e+18                      Max.   :552465537
##  neighbourhood         latitude       longitude                room_type
##  Length:313         Min.   :42.63   Min.   :-73.83   Entire home/apt:231
##  Class :character   1st Qu.:42.65   1st Qu.:-73.79   Private room   : 81
##  Mode  :character   Median :42.66   Median :-73.77   Shared room    :  1
##                     Mean   :42.66   Mean   :-73.78
##                     3rd Qu.:42.66   3rd Qu.:-73.76
##                     Max.   :42.69   Max.   :-73.74
##      price        minimum_nights   number_of_reviews
##  Min.   :  26.0   Min.   : 1.000   Min.   :  1.00
##  1st Qu.:  79.0   1st Qu.: 1.000   1st Qu.:  8.00
##  Median : 105.0   Median : 1.000   Median : 29.00
##  Mean   : 125.3   Mean   : 3.489   Mean   : 70.32
##  3rd Qu.: 136.0   3rd Qu.: 2.000   3rd Qu.: 84.00
##  Max.   :1379.0   Max.   :30.000   Max.   :795.00
##   last_review                     reviews_per_month
##  Min.   :2018-07-27 00:00:00.00   Min.   : 0.040
##  1st Qu.:2024-02-19 00:00:00.00   1st Qu.: 0.690
##  Median :2024-04-14 00:00:00.00   Median : 1.720
##  Mean   :2024-02-05 05:17:26.65   Mean   : 2.283
##  3rd Qu.:2024-04-26 00:00:00.00   3rd Qu.: 3.160
##  Max.   :2024-05-06 00:00:00.00   Max.   :11.050
##  calculated_host_listings_count availability_365 number_of_reviews_ltm
##  Min.   : 1.00                  Min.   :  3.0    Min.   :  0.00
##  1st Qu.: 1.00                  1st Qu.:126.0    1st Qu.:  4.00
##  Median : 3.00                  Median :243.0    Median : 14.00
##  Mean   : 5.54                  Mean   :220.4    Mean   : 21.21
##  3rd Qu.: 9.00                  3rd Qu.:318.0    3rd Qu.: 29.00
##  Max.   :22.00                  Max.   :365.0    Max.   :131.00
##  reviews_per_year
##  Min.   :  0.48
##  1st Qu.:  8.28
##  Median : 20.64
##  Mean   : 27.40
##  3rd Qu.: 37.92
##  Max.   :132.60
```

```r
# Visualizations
ggplot(listings, aes(x = price)) +
  geom_histogram(binwidth = 50) +
  ggtitle("Distribution of Prices")
```



```r
ggplot(listings, aes(x = room_type, y = price)) +
  geom_boxplot() +
  ggtitle("Price by Room Type")
```



```r
# Feature Engineering
# Creating a variable which contains distance of a place from the given landmark
landmark_lat <- 40.748817
landmark_lon <- -73.985428

# Define a function to calculate the Haversine distance
haversine_distance <- function(lat1, lon1, lat2, lon2) {
  # Convert degrees to radians
  radians <- pi / 180
  lat1 <- lat1 * radians
  lon1 <- lon1 * radians
  lat2 <- lat2 * radians
  lon2 <- lon2 * radians

  # Haversine formula
  dlat <- lat2 - lat1
  dlon <- lon2 - lon1
  a <- sin(dlat / 2)^2 + cos(lat1) * cos(lat2) * sin(dlon / 2)^2
  c <- 2 * atan2(sqrt(a), sqrt(1 - a))

  # Radius of Earth in kilometers
  R <- 6371
  distance <- R * c
  return(distance)
}

listings <- mutate(listings,distance_from_landmark = haversine_distance(latitude, longitude, landmark_la

# Modeling
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: lattice
```

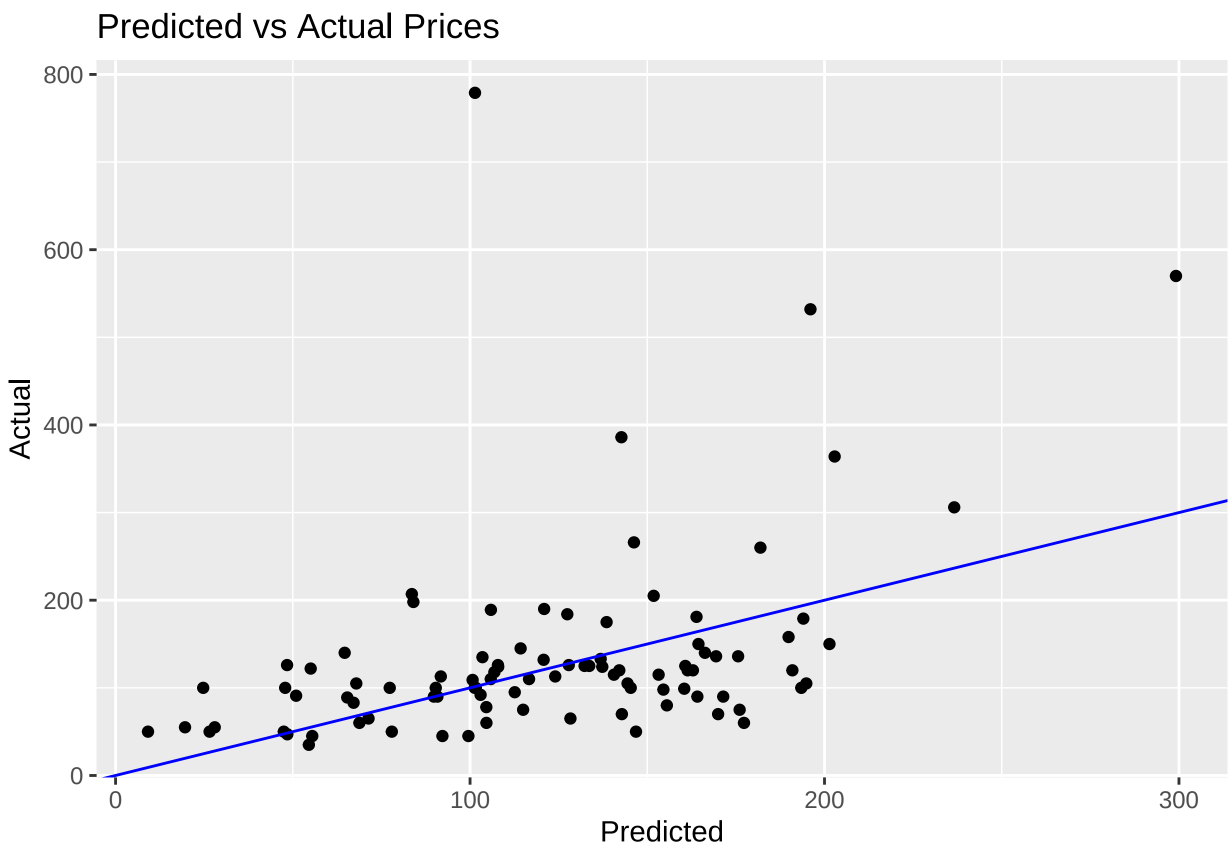
```
set.seed(123)
train_indices <- sample(seq_len(nrow(listings)), size = 0.7 * nrow(listings))
train_data <- listings[train_indices, ]
test_data <- listings[-train_indices, ]

# Exclude 'name' and 'host_name' variables from the model
train_data <- select(train_data, -name, -host_name)
test_data <- select(test_data, -name, -host_name)

# Building a regression model
model <- lm(price ~ ., data = train_data)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ ., data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158.14  -41.65   -6.94   22.84 1129.74
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.189e+06  9.229e+06   0.129   0.8976
## id                              8.861e-18  2.312e-17   0.383   0.7019
## host_id                         1.026e-07  4.432e-08   2.315   0.0217 *
## neighbourhoodELEVENTH WARD     -8.476e+01  8.083e+01  -1.049   0.2957
## neighbourhoodFIFTEENTH WARD    -1.566e+02  8.563e+01  -1.829   0.0689 .
## neighbourhoodFIFTH WARD        -1.298e+02  1.017e+02  -1.276   0.2034
## neighbourhoodFIRST WARD         2.431e+01  8.731e+01   0.278   0.7810
## neighbourhoodFOURTEENTH WARD   -8.238e+01  6.073e+01  -1.356   0.1766
## neighbourhoodFOURTH WARD       -2.236e+01  1.276e+02  -0.175   0.8611
## neighbourhoodNINTH WARD        -6.421e+01  6.349e+01  -1.011   0.3132
## neighbourhoodSECOND WARD        1.901e+01  8.668e+01   0.219   0.8266
## neighbourhoodSEVENTH WARD      -2.291e+01  7.278e+01  -0.315   0.7532
## neighbourhoodSIXTH WARD        -3.959e+01  7.530e+01  -0.526   0.5996
## neighbourhoodTENTH WARD        -5.665e+01  6.764e+01  -0.838   0.4033
## neighbourhoodTHIRD WARD        -4.857e+01  9.619e+01  -0.505   0.6142
## neighbourhoodTHIRTEENTH WARD   -1.421e+02  7.394e+01  -1.922   0.0560 .
## neighbourhoodTWELFTH WARD      -1.435e+02  9.550e+01  -1.503   0.1345
## latitude                       -3.694e+04  2.529e+05  -0.146   0.8840
## longitude                      -4.190e+03  1.463e+04  -0.286   0.7749
## room_typePrivate room          -8.152e+01  1.885e+01  -4.324 2.47e-05 ***
## room_typeShared room           -3.695e+01  1.168e+02  -0.316   0.7521
## minimum_nights                 -4.964e-01  1.232e+00  -0.403   0.6875
## number_of_reviews               4.138e-02  1.315e-01   0.315   0.7533
## last_review                    -5.125e-07  5.104e-07  -1.004   0.3166
## reviews_per_month              -9.339e+00  7.881e+00  -1.185   0.2375
## calculated_host_listings_count  2.923e+00  1.413e+00   2.069   0.0399 *
## availability_365                1.066e-01  6.984e-02   1.526   0.1286
## number_of_reviews_ltm           3.528e-01  5.825e-01   0.606   0.5455
## reviews_per_year                       NA         NA      NA       NA
## distance_from_landmark          3.675e+02  2.286e+03   0.161   0.8724
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 102.8 on 190 degrees of freedom
## Multiple R-squared:  0.2082, Adjusted R-squared:  0.09157
## F-statistic: 1.785 on 28 and 190 DF,  p-value: 0.01271
```

```
# 6. Model Evaluation
# Predicting on test set
predictions <- predict(model, newdata = test_data)

# Calculate RMSE
rmse <- sqrt(mean((predictions - test_data$price)^2))
print(paste("RMSE: ", rmse))
```

```
## [1] "RMSE:  101.554272164889"
```

```
# Visualizing model performance
ggplot(data.frame(Predicted = predictions, Actual = test_data$price), aes(x = Predicted, y = Actual)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = 'blue') +
  ggtitle("Predicted vs Actual Prices")
```


Predicted vs Actual Prices
Actual
800
600
400
200
0
Predicted
0
100
200
300