***DMPM Assignment 5***

Name: Rushikesh Jyoti

Division: A

Roll no: 27

SRN: 201901139

Question: build a Linear Regression Model

Code

library(tidyverse)

library(ggplot2)

library(dplyr)

library(Metrics)

library(caret)

library(scales)

library(caTools)

library(corrplot)

dataset <- read.csv("AB\_NYC\_2019.csv")

summary(dataset)

# Find NA Values

print(colSums(is.na(dataset)))

# Fill 0 into NA

dataset$reviews\_per\_month[is.na(dataset$reviews\_per\_month) == TRUE] <- 0

data\_new <- tidyr::separate(dataset, last\_review, c("Year", "Month", "Day"), sep = "-")

data\_new$Year[is.na(data\_new$Year) == TRUE] <- 0

data\_new$Month[is.na(data\_new$Month) == TRUE] <- 0

data\_new$Day[is.na(data\_new$Day) == TRUE] <- 0

data\_new$neighbourhood\_group <- as.factor(data\_new$neighbourhood\_group)

data\_new$room\_type <- as.factor(data\_new$room\_type)

data\_new$Year <- as.integer(data\_new$Year)

data\_new$Month <- as.integer(data\_new$Month)

data\_new$Day <- as.integer(data\_new$Day)

head(data\_new)

print(colSums(is.na(data\_new)))

data\_new <- na.omit(data\_new)

summary(data\_new)

# Correlation

correlation <- cor(data\_new[, sapply(data\_new, is.numeric)])

corrplot(cor(data\_new[, sapply(data\_new, is.numeric)]))

# Plotting the graphs

ggplot(data = data\_new, mapping = aes(neighbourhood\_group, fill = room\_type)) +

geom\_bar(colour = "Black", position = position\_dodge())

price\_roomtype <- data\_new %>%

group\_by(neighbourhood\_group, room\_type) %>%

summarise(Mean\_Price = mean(price))

ggplot(price\_roomtype, aes(x = reorder(neighbourhood\_group, -Mean\_Price), y = Mean\_Price, fill = room\_type)) +

geom\_bar(stat = "identity", colour = "black", position = position\_dodge())

ggplot(data\_new, aes(y = price, x = minimum\_nights, color = neighbourhood\_group)) +

geom\_jitter()

ggplot(data = data\_new, mapping = aes(number\_of\_reviews, price)) +

geom\_point() +

facet\_wrap(data\_new$room\_type)

ggplot(data = data\_new, mapping = aes(availability\_365, price)) +

geom\_point()

ggplot(data = data\_new, mapping = aes(neighbourhood\_group, availability\_365)) +

geom\_boxplot()

# Building the model

model <- lm(price ~ host\_id + neighbourhood\_group + latitude + longitude + room\_type + minimum\_nights + number\_of\_reviews +

Year + calculated\_host\_listings\_count + availability\_365, data = data\_new)

print(model)

print(summary(model))

pred1 <- predict(model)

resd1 <- residuals(model)

x <- cbind(data\_new$price, pred1)

x <- data.matrix(x)

x <- rescale(x)

x <- as.data.frame(x)

mae <- MAE(x$V1, x$pred1)

mse <- mse(x$V1, x$pred1)

rmse <- RMSE(x$V1, x$pred1)

r2 <- R2(x$V1, x$pred1)

cat("\nMAE:", mae, "\n\nMSE:", mse, "\n\nRMSE:", rmse, "\n\nR-squared:", r2, "\n\n")

ggplot(data\_new, aes(y = pred1, x = data\_new$price)) +

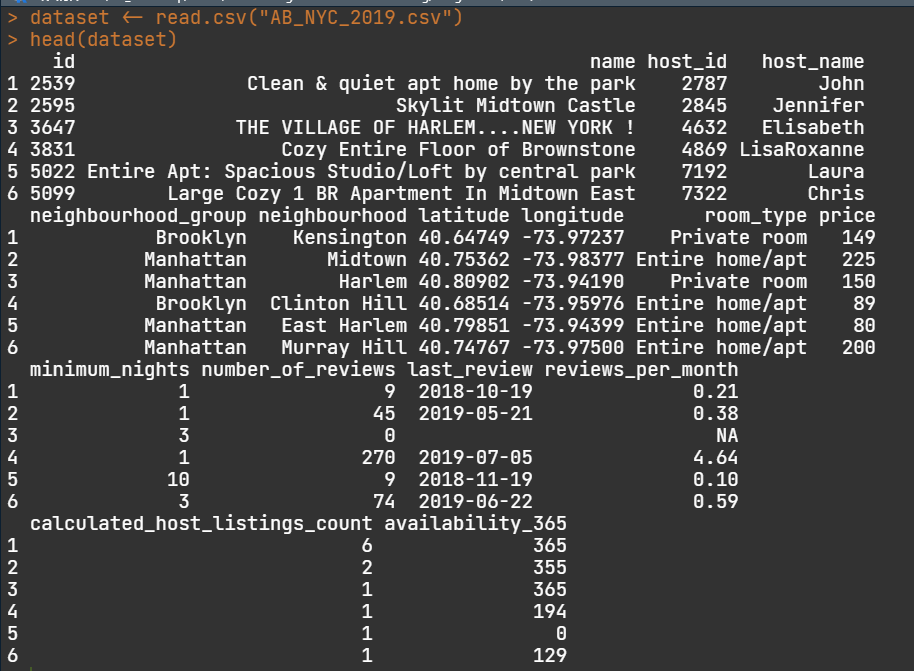
geom\_point() +

geom\_abline(intercept = 0, slope = 1, colour = "Red") +

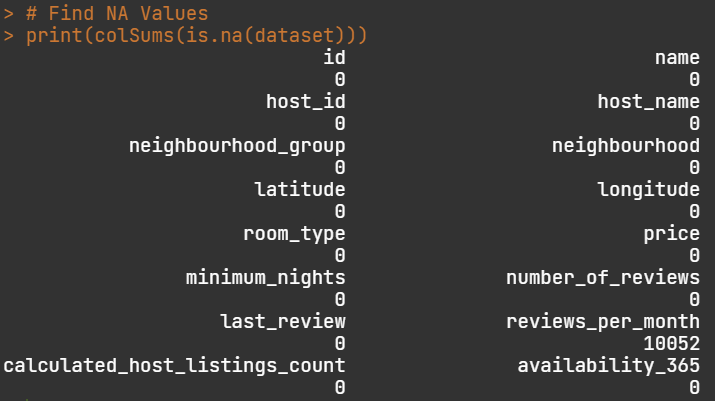
labs(y = "Predicted Values", x = "Actual Values", title = "Predicted vs. Actual Values")

**Output**

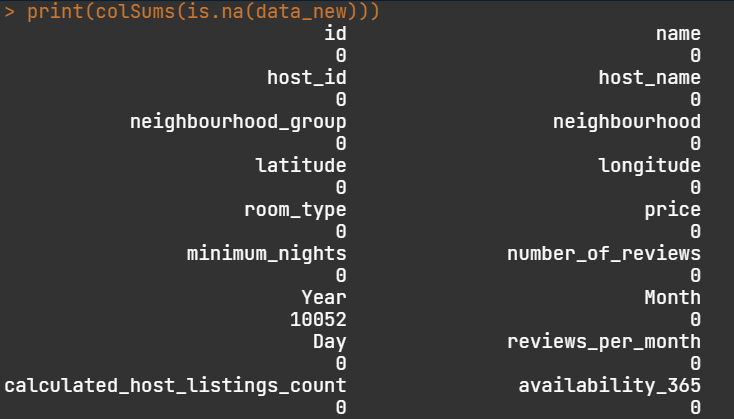
Taking a look at dataset



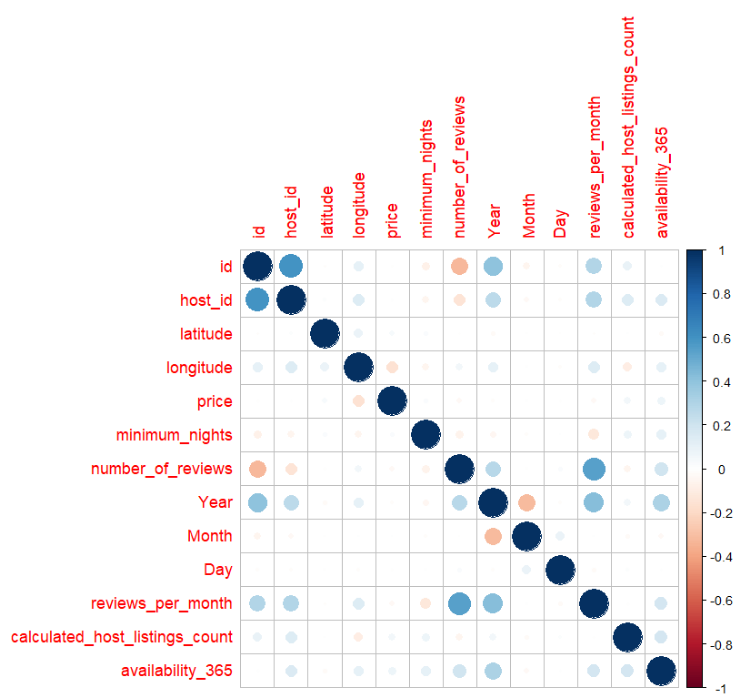
NA values in variables



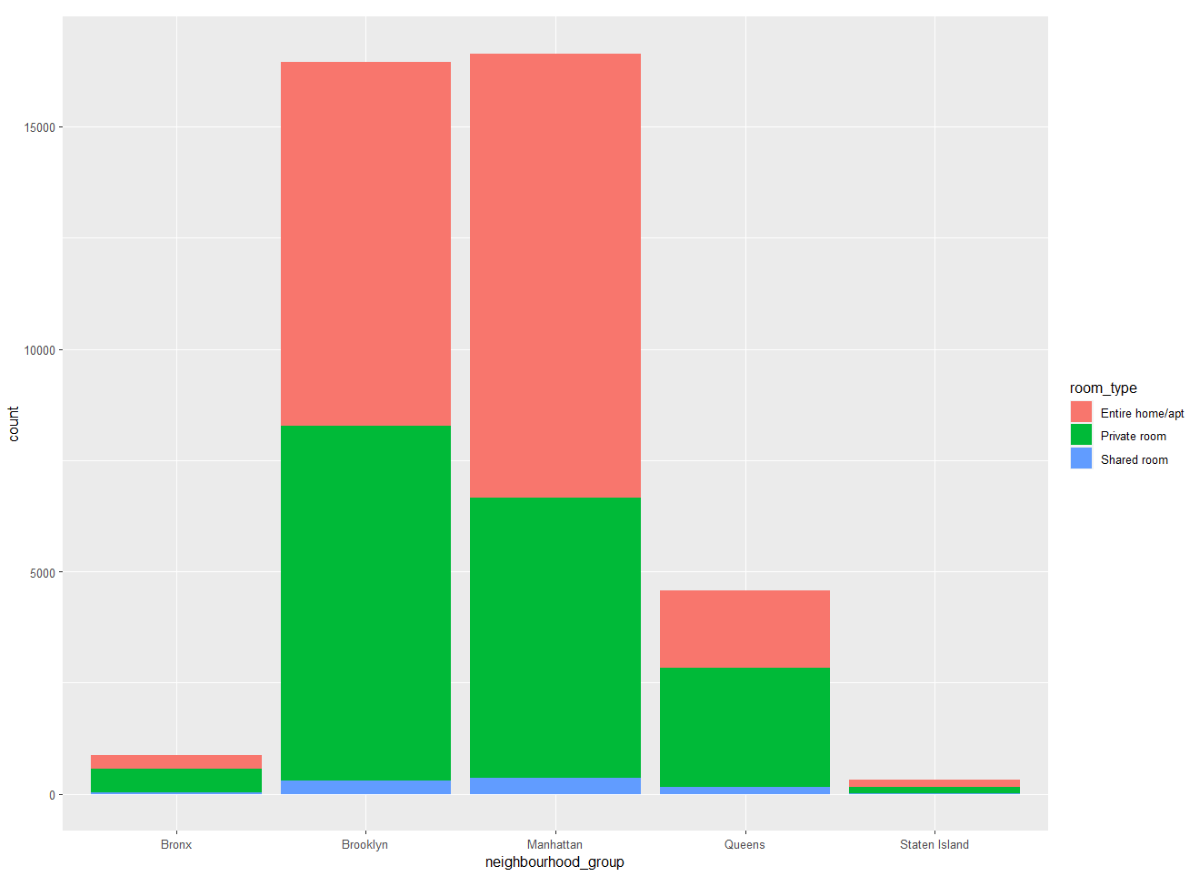
After omitting NA values



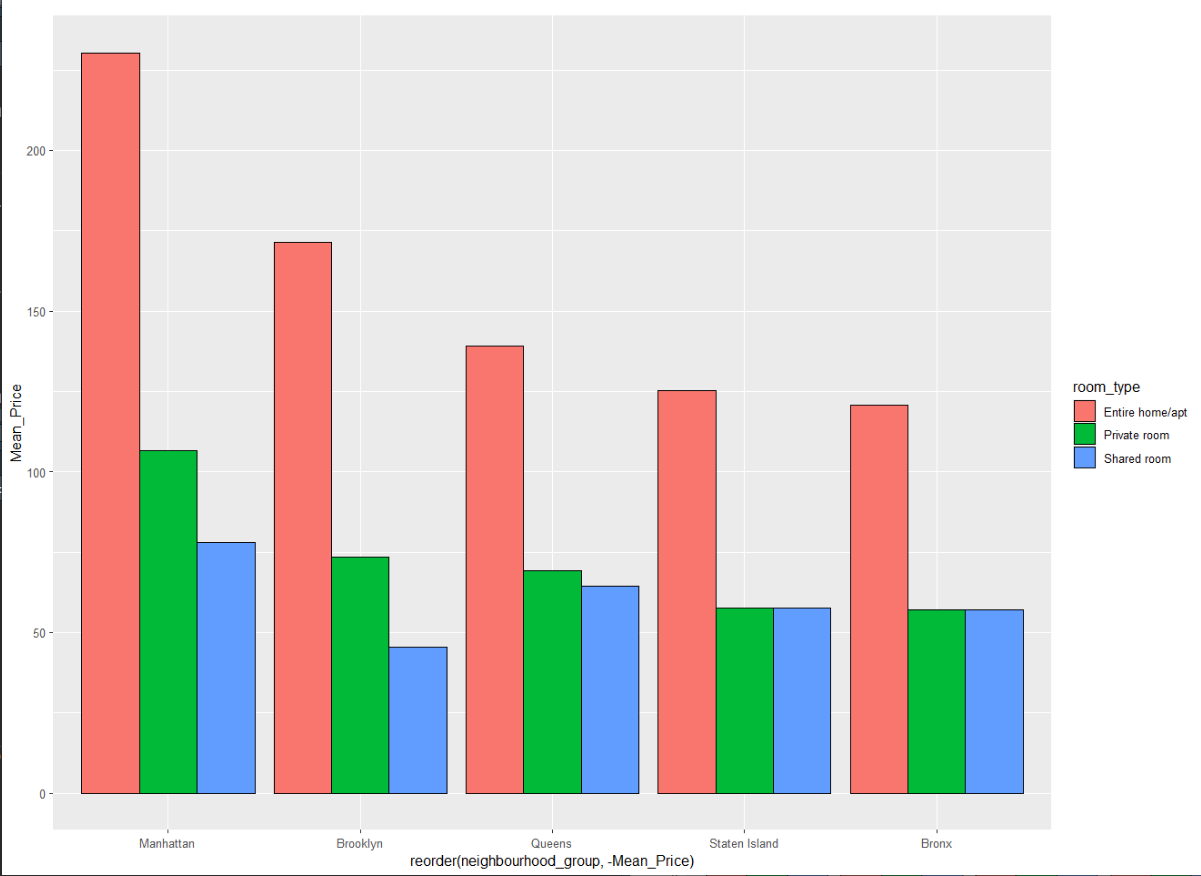
Correlation Matrix



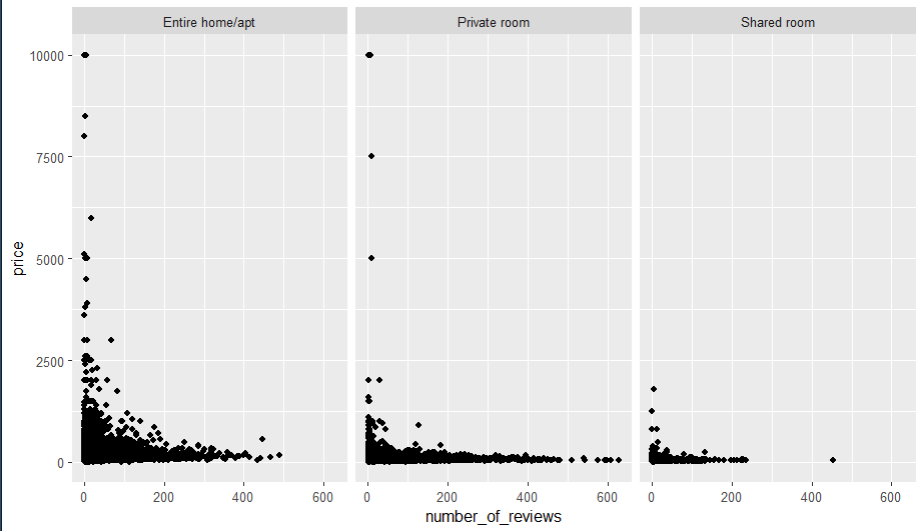
Count of each neighbourhood group



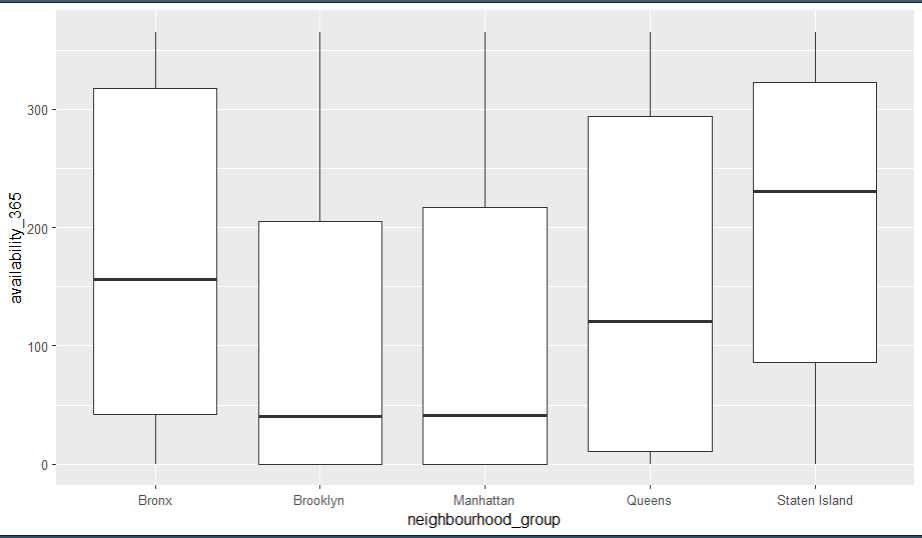
Mean price for each neighbourhood

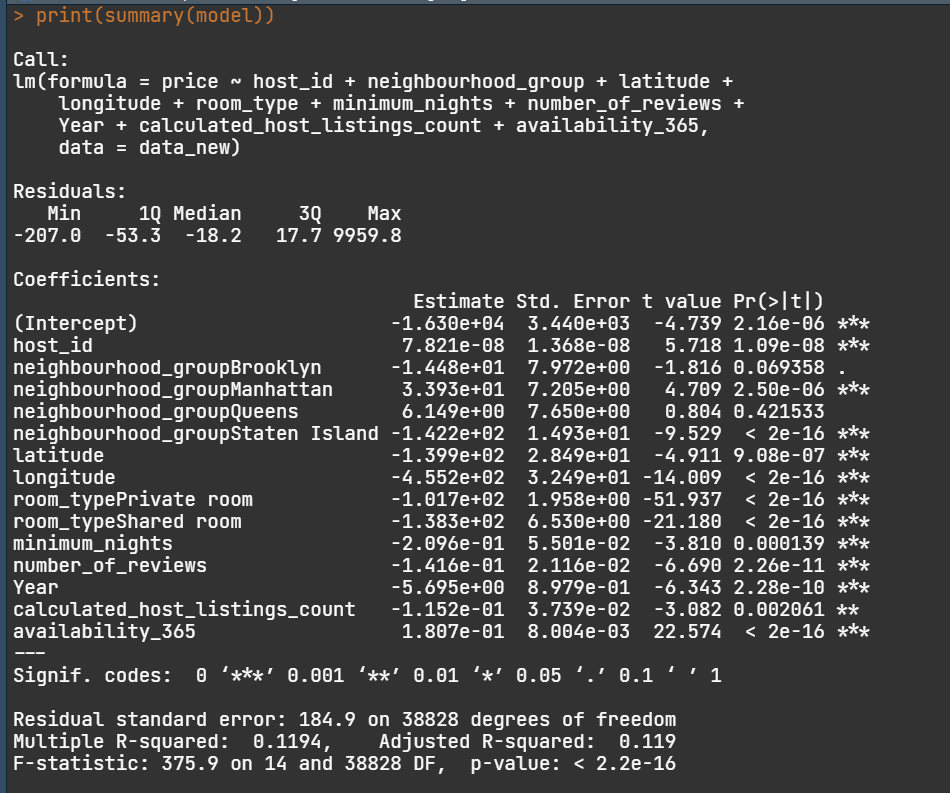


Price of rooms based on reviews grouped by type of room



Box plot of room availability and neighbourhood



Building the model

Metrics

