R Notebook

Code

Name: Klysman Vieira

Id: 230353014

EDA Singapore Accommodations

Introduction:

During exploratory data analysis (EDA), the dataframe that has information about accommodation in Singapore was worked on. Some dataframe columns contain the following information:

- `neighborhood_group` : Indicates the region of the neighborhood where the accommodation is located.
- `name` : Name of the accommodation.
- `host_id` : ID of the host responsible for the accommodation.
- `host_name` : Host name.
- `neighborhood` : Name of the neighborhood where the accommodation is located.
- `latitude` and `longitude` : Geographic coordinates of the accommodation.
- `number_of_reviews` : Number of reviews received by the accommodation.
- `room_type` : Type of room available in the accommodation.
- `price` : Price of the accommodation.
- `availability_365` : Number of days the accommodation is available to book over the course of a year.

During exploratory data analysis, we explore and summarize the main characteristics and patterns present in these data.

1. Descriptive analysis: We perform a statistical description of the variables, such as mean, median, standard deviation, minimum and maximum. This helps us understand the distribution of data and identify potential discrepancies or outliers.
2. Handling missing or inconsistent data: We identify and handle missing or inconsistent data, such as nulls or outliers. This may involve deleting records with missing data or filling in these values with appropriate techniques.
3. Data visualization: We use graphs and visualizations to represent data in a more understandable way. For example, we can create bar charts to show the distribution of room types or a map to visualize the location of accommodations.

Insights

- We look for interesting patterns and insights in the data that can be useful for decision-making or answering specific questions. For example, we may discover that certain neighborhoods have a greater availability of accommodation throughout the year or that certain room types have a direct relationship with price.

- There are relationships between the variables present in the dataframe. For example, we can analyze whether the price of accommodation varies according to the region of the neighborhood or whether there is any correlation between the number of reviews and availability throughout the year.

Loading the raw dataframe

Hide

```
library(readxl)

file_path <- "G:/My Drive/IPS/Mestrado/UCs/Aprendizage supervisionada/IPS-ESCE SP ML/Session 3/Data/Singapore Airbnb - Raw.xlsx"

data <- read_excel(file_path)

num_rows <- dim(data)[1]
num_cols <- dim(data)[2]

column_names <- colnames(data)

cat("Number of rows:", num_rows, "\n")
```

```
Number of rows: 7923
```

Hide

```
cat("Number of columns:", num_cols, "\n")
```

```
Number of columns: 16
```

Hide

```
cat("Column names:", paste(column_names, collapse = ", "), "\n")
```

```
Column names: id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365
```

Checking for missing data

Hide

```
missing_values <- colSums(is.na(data))

print("Number of missing values in each column:")
```

```
[1] "Number of missing values in each column:"
```

Hide

```
print(missing_values)
```

```
                            id                           name 
                             0                              5 
                       host_id                      host_name 
                            13                              0 
          neighbourhood_group                  neighbourhood 
                             0                              0 
                      latitude                      longitude 
                            20                              9 
                     room_type                          price 
                             0                              0 
                minimum_nights                number_of_reviews 
                             0                              4 
                   last_review              reviews_per_month 
                          2773                           2773 
calculated_host_listings_count               availability_365 
                             0                              0 
```

Display the first rows of data

Hide

```
head(data)
```

id <dbl>	name <chr>	host_id <dbl>	host_na… <chr>	neighbourhood_group <chr>	nei <ch
49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Wo
50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Buk
56334	COZICOMFORT	266763	Francesca	North Region	Wo
71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tan
71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tan
71903	Room 2-near Airport & EXPO	367042	Belinda	East Region	Tan

6 rows | 1-6 of 16 columns

Statistical data about the dataframe

Hide

```
summary(data)
```

```
      id                 name              host_id        
 Min.   :   49091   Length:7923        Min.   :    23666  
 1st Qu.:15824582   Class :character   1st Qu.: 23129381  
 Median :24707713   Mode  :character   Median : 63448912  
 Mean   :23404133                      Mean   : 91313156  
 3rd Qu.:32365800                      3rd Qu.:155656938  
 Max.   :38112762                      Max.   :288567551  
                                       NA's   :13         
  host_name         neighbourhood_group neighbourhood      
 Length:7923        Length:7923         Length:7923        
 Class :character   Class :character    Class :character   
 Mode  :character   Mode  :character    Mode  :character   
                                                           
                                                           
                                                           
                                                           
    latitude       longitude      room_type        
 Min.   :1.244   Min.   :103.6   Length:7923        
 1st Qu.:1.296   1st Qu.:103.8   Class :character   
 Median :1.311   Median :103.8   Mode  :character   
 Mean   :1.314   Mean   :103.8                      
 3rd Qu.:1.322   3rd Qu.:103.9                      
 Max.   :1.455   Max.   :104.0                      
 NA's   :20      NA's   :9                          
     price         minimum_nights    number_of_reviews
 Min.   :    0.0   Min.   :   1.00   Min.   :  0.00    
 1st Qu.:   65.0   1st Qu.:   1.00   1st Qu.:  0.00    
 Median :  124.0   Median :   3.00   Median :  2.00    
 Mean   :  192.3   Mean   :  17.53   Mean   : 12.73    
 3rd Qu.:  199.0   3rd Qu.:  10.00   3rd Qu.: 10.00    
 Max.   :65000.0   Max.   :1000.00   Max.   :323.00    
                                     NA's   :4         
  last_review                     reviews_per_month
 Min.   :2013-10-21 00:00:00.00   Min.   : 0.010   
 1st Qu.:2018-11-21 00:00:00.00   1st Qu.: 0.180   
 Median :2019-06-27 00:00:00.00   Median : 0.550   
 Mean   :2019-01-11 17:33:01.04   Mean   : 1.044   
 3rd Qu.:2019-08-07 00:00:00.00   3rd Qu.: 1.370   
 Max.   :2019-08-27 00:00:00.00   Max.   :13.000   
 NA's   :2773                     NA's   :2773     
 calculated_host_listings_count availability_365
 Min.   :  1.00                 Min.   :  0.0   
 1st Qu.:  2.00                 1st Qu.: 54.0   
 Median :  9.00                 Median :260.0   
 Mean   : 40.55                 Mean   :208.7   
 3rd Qu.: 48.00                 3rd Qu.:355.0   
 Max.   :274.00                 Max.   :365.0   
```

Data cleaning

Hide

```
#Remover colunas desnecessárias
data <- data[, c("neighbourhood_group", "name", "host_id", "host_name", "neighbourhood", "latitu
de", "number_of_reviews", "longitude", "room_type", "price", "availability_365")]
```

Filter only complete rows (no null values)

Hide

```
data_clean <- data[complete.cases(data), ]
```

Identify and remove duplicate elements

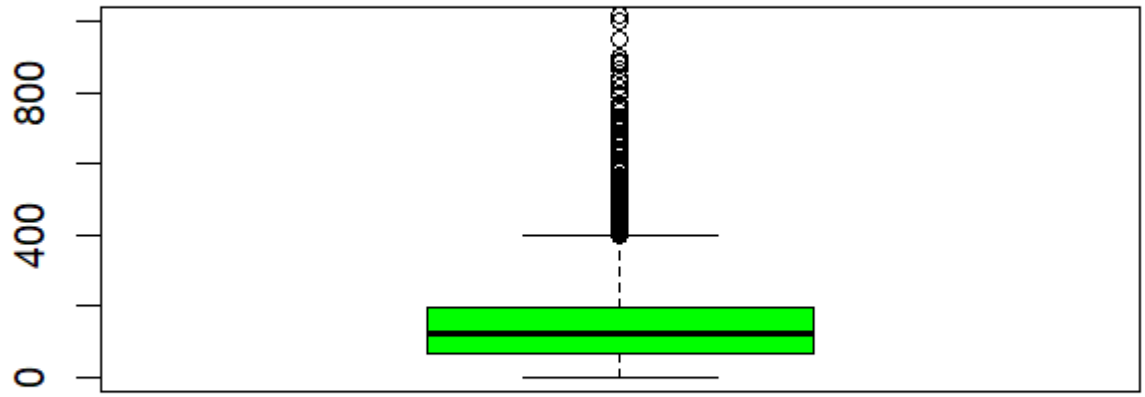
Hide

```
data_unique <- subset(data_clean, !duplicated(data_clean))
```

Boxplot to understand the distribution of data depending on the price column

Hide

```
# Seleciona as colunas desejadas para criar os boxplots
boxplot(data_unique$price, col = "green", ylim = c(0, 1000))
```



Remove lines (outliers) where price >= 500

Hide

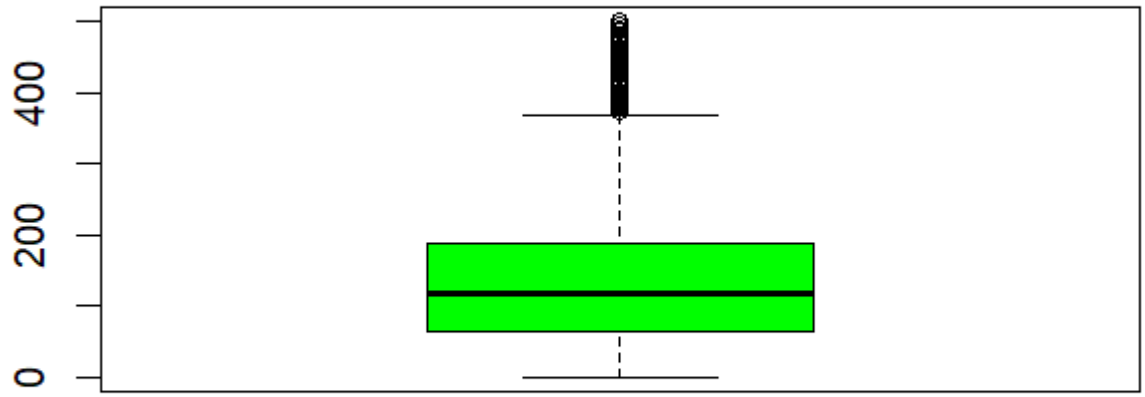
```
data_filtered_500 <- subset(data_unique, price < 500)
```

Graphics

Create a boxplot for the price column

Hide

```
boxplot_output <- boxplot(data_filtered_500$price, col = "green", ylim = c(0, 500))
```



Hide

```
# Acessa os resultados matemáticos do boxplot
min_value <- boxplot_output$stats[1]
q1 <- boxplot_output$stats[2]
median <- boxplot_output$stats[3]
q3 <- boxplot_output$stats[4]
max_value <- boxplot_output$stats[5]

# Imprime os resultados
cat("Valor mínimo:", min_value, "\n")
```

```
Valor mínimo: 0
```

Hide

```
cat("Primeiro quartil:", q1, "\n")
```

```
Primeiro quartil: 65
```

Hide

```
cat("Mediana:", median, "\n")
```

```
Mediana: 119
```

Hide

```
cat("Terceiro quartil:", q3, "\n")
```

```
Terceiro quartil: 187
```

Hide

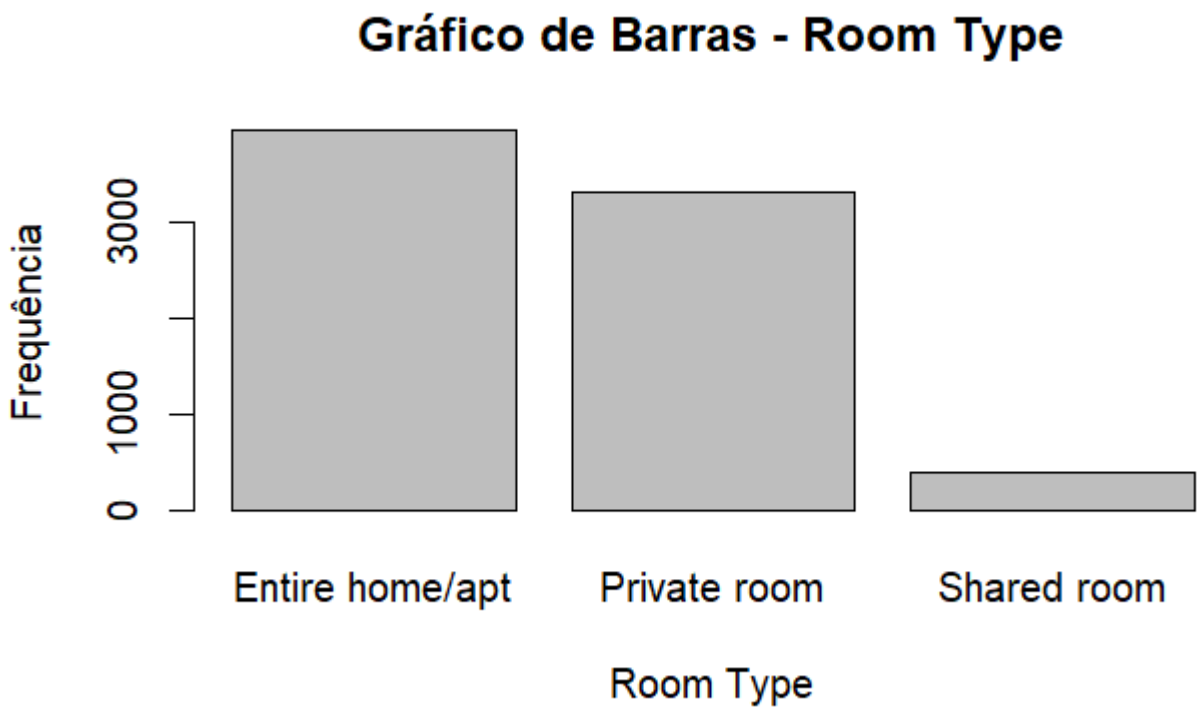
```
cat("Valor máximo:", max_value, "\n")
```

```
Valor máximo: 369
```

Create a bar chart for the “room_type” column

Hide

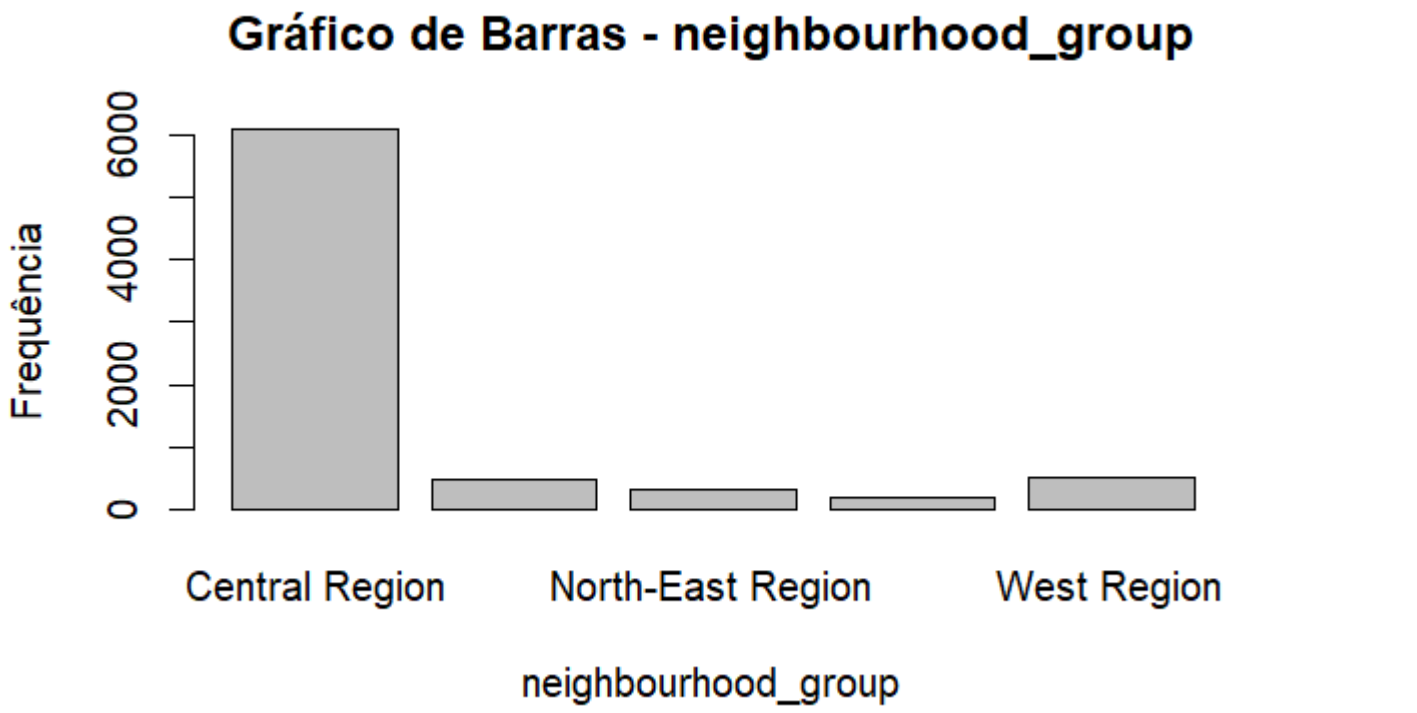
```
barplot(table(data_filtered_500$room_type), main = "Gráfico de Barras - Room Type", xlab = "Room Type", ylab = "Frequência")
```



Creates a bar chart for the “room_type” column. Shows the regions of greatest interest in Singapore

Hide

```
barplot(table(data_filtered_500$neighbourhood_group), main = "Gráfico de Barras - neighbourhood_
group", xlab = "neighbourhood_group", ylab = "Frequência")
```



Hide

```
# Obtém o índice da linha com o maior valor em number_of_reviews
index <- which.max(data_filtered_500$number_of_reviews)

# Obtém a linha completa com o maior valor em number_of_reviews
row_with_max_reviews <- data_filtered_500[index, ]

# Imprime a linha completa
print(row_with_max_reviews)
```

neighbourhood_group <chr>	name <chr>	host_id <dbl>	host_na… <chr>	neighbourhoo <chr>
East Region	Luxuriously Spacious Studio Apt.	7642747	Shirley	Bedok

1 row | 1-6 of 11 columns

Hide

```
NA
```

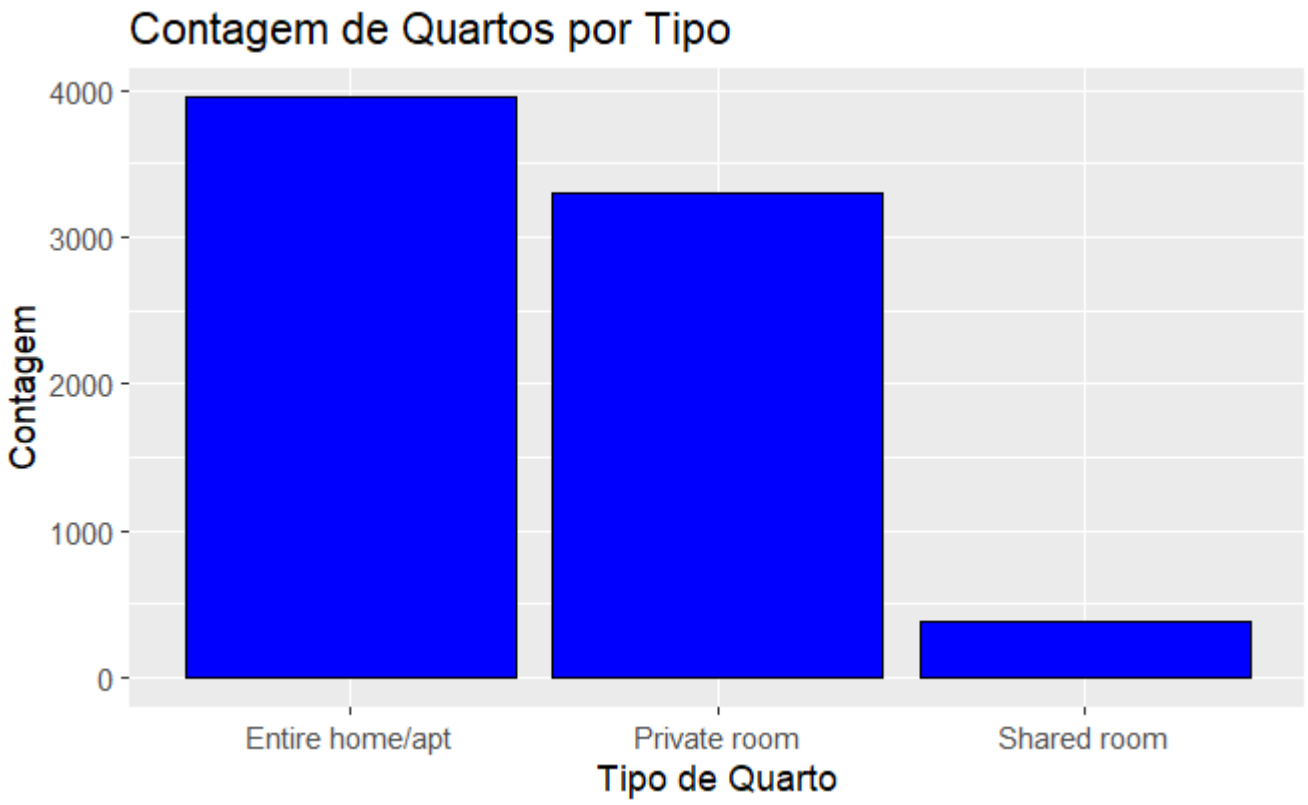
GGPLOT2

Bar graphs

Hide

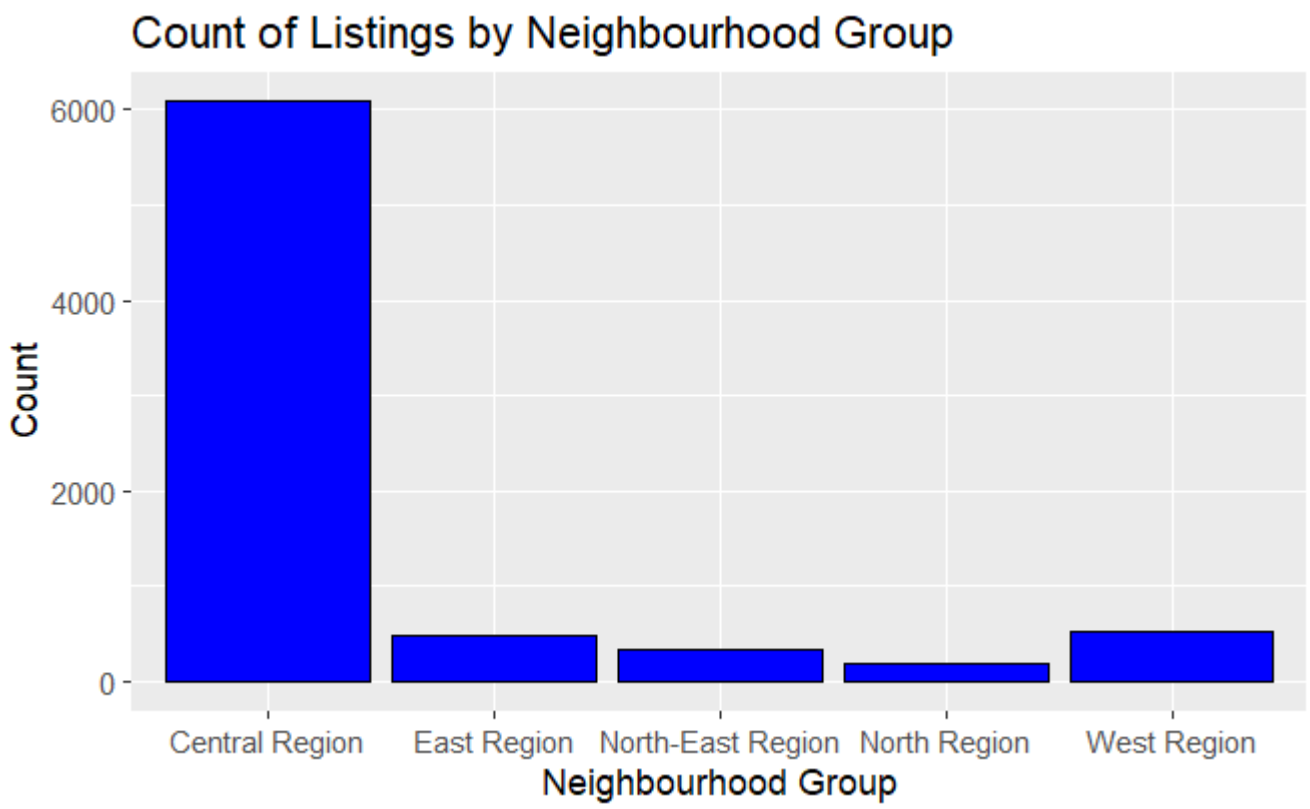
```
library(ggplot2)

# Cria o gráfico de barras usando ggplot2
ggplot(data_filtered_500, aes(x = room_type)) +
  geom_bar(fill = "blue", color = "black") +
  labs(x = "Tipo de Quarto", y = "Contagem") +
  ggtitle("Contagem de Quartos por Tipo")
```



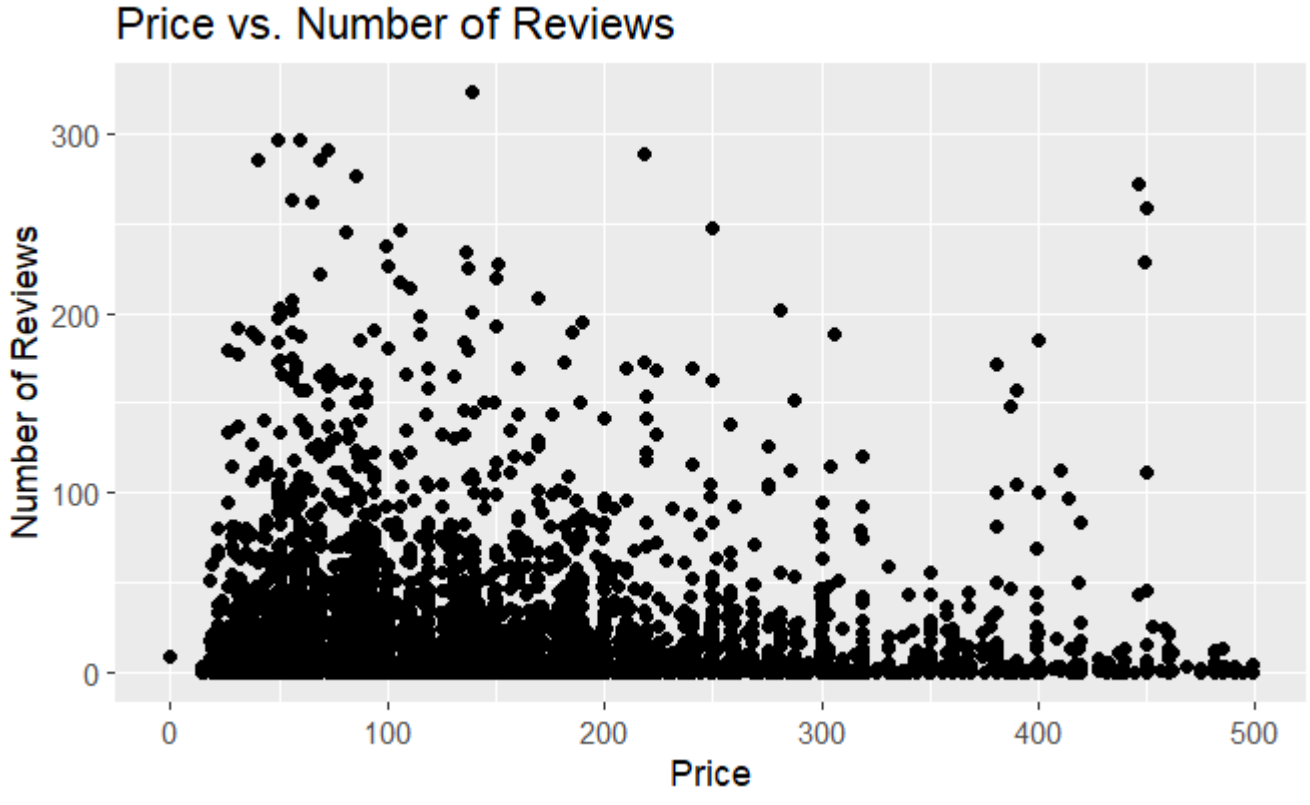
Hide

```
ggplot(data_filtered_500, aes(x = neighbourhood_group)) +
  geom_bar(fill = "blue", color = "black") +
  labs(x = "Neighbourhood Group", y = "Count") +
  ggtitle("Count of Listings by Neighbourhood Group")
```



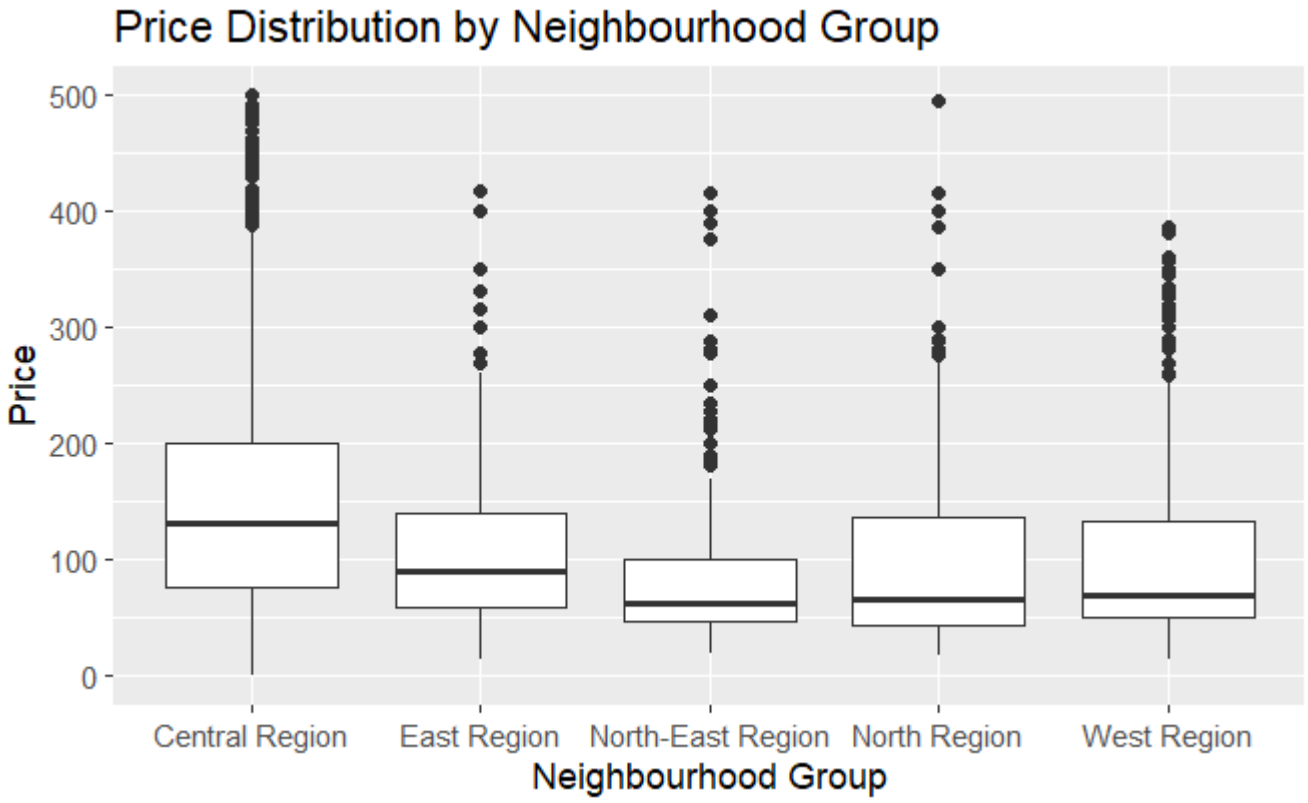
Hide

```
ggplot(data_filtered_500, aes(x = price, y = number_of_reviews)) +
  geom_point() +
  labs(x = "Price", y = "Number of Reviews") +
  ggtitle("Price vs. Number of Reviews")
```



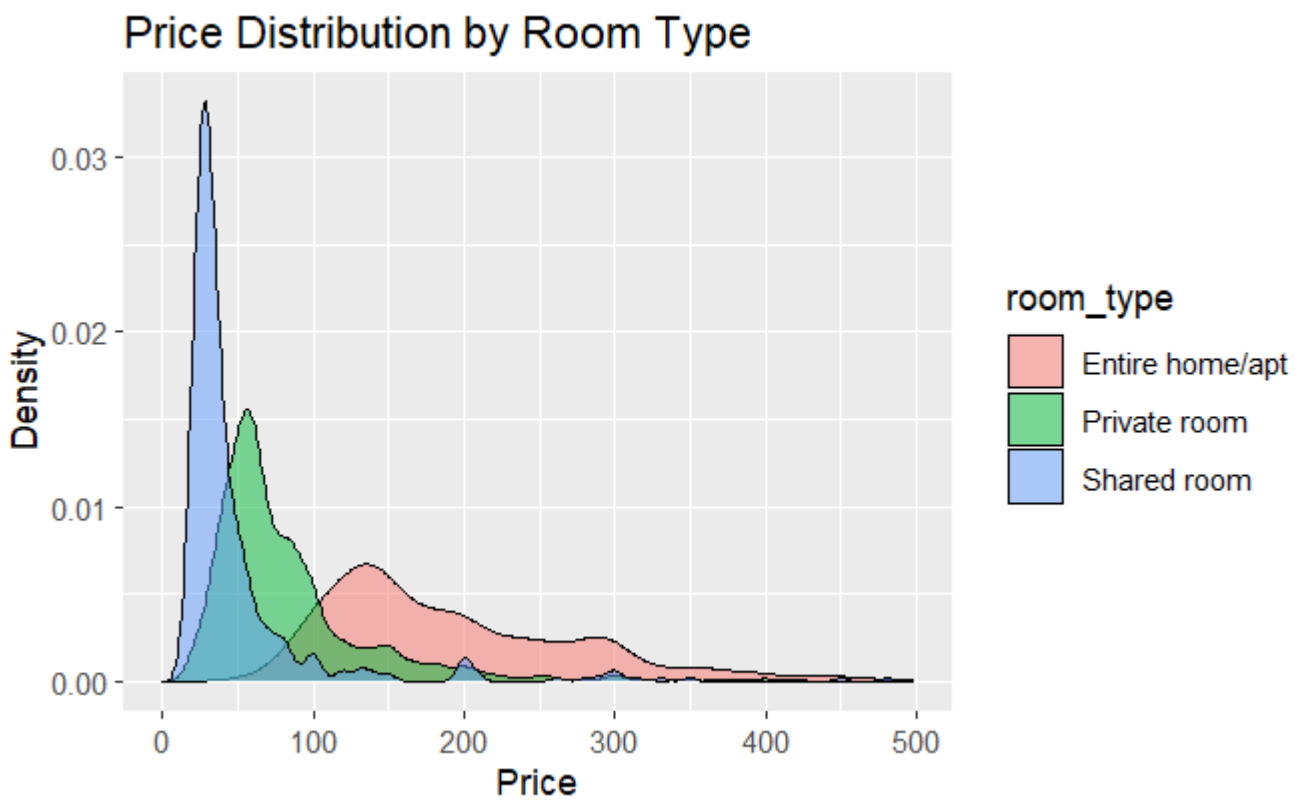
Hide

```
ggplot(data_filtered_500, aes(x = neighbourhood_group, y = price)) +
  geom_boxplot() +
  labs(x = "Neighbourhood Group", y = "Price") +
  ggtitle("Price Distribution by Neighbourhood Group")
```



Hide

```
ggplot(data_filtered_500, aes(x = price, fill = room_type)) +
  geom_density(alpha = 0.5) +
  labs(x = "Price", y = "Density") +
  ggtitle("Price Distribution by Room Type")
```



#salvando o novo dataframe data_filtered_500_price em formato XLSX

Hide

```
library(openxlsx)
new_data <- "G:/My Drive/IPS/Mestrado/UCs/Aprendizage supervisionada/IPS-ESCE SP ML/Session 3/Da
ta/data_filtered_500_price.xlsx"
write.xlsx(data_filtered_500, file = new_data)
```