BIXI Project - Part 1

Ha Dang Vu

2025-09-13

1. Data

```
bixi <- read.csv("bixi_part_1.csv")
head(bixi)
```

```
##                                           station        arrondissement
## 1                              de Bordeaux / Rachel Le Plateau-Mont-Royal
## 2      Square St-Louis (du Square St-Louis / Laval) Le Plateau-Mont-Royal
## 3                         de Brébeuf / du Mont-Royal Le Plateau-Mont-Royal
## 4                             Marie-Anne / Papineau Le Plateau-Mont-Royal
## 5 Parc Hilda-Ramacière (de Bullion / Prince-Arthur) Le Plateau-Mont-Royal
## 6                              des Pins / Hutchison Le Plateau-Mont-Royal
##        lat      long       dur mm jj temp precip
## 1 45.53208 -73.56770 12.900850  9  6 19.5    0.0
## 2 45.51609 -73.57013 15.406917  9  6 15.9   16.4
## 3 45.52922 -73.57785  3.790017  7  2 25.9    0.0
## 4 45.53177 -73.57241  7.313450  5  2 17.7    0.3
## 5 45.51562 -73.57213 12.470333  6  6 21.0    0.0
## 6 45.51080 -73.57822 15.558917  5  3 10.2    0.4
```

2. Adding "weekend" variable

To address whether weekend BIXI trips tend to be longer, we created a new variable (weekend) that classifies each trip as either a weekday (0) or a weekend (1) based on the day of the week (jj). This simplifies the comparison into two groups directly aligned with the business question.

```
# Create weekend variable: 1 = weekend, 0 = weekday
bixi$weekend <- ifelse(bixi$jj %in% c(6, 7), 1, 0)

head(bixi)
```

```
##                                           station        arrondissement
## 1                              de Bordeaux / Rachel Le Plateau-Mont-Royal
## 2      Square St-Louis (du Square St-Louis / Laval) Le Plateau-Mont-Royal
## 3                         de Brébeuf / du Mont-Royal Le Plateau-Mont-Royal
## 4                             Marie-Anne / Papineau Le Plateau-Mont-Royal
## 5 Parc Hilda-Ramacière (de Bullion / Prince-Arthur) Le Plateau-Mont-Royal
## 6                              des Pins / Hutchison Le Plateau-Mont-Royal
##        lat      long       dur mm jj temp precip weekend
## 1 45.53208 -73.56770 12.900850  9  6 19.5    0.0       1
## 2 45.51609 -73.57013 15.406917  9  6 15.9   16.4       1
## 3 45.52922 -73.57785  3.790017  7  2 25.9    0.0       0
```

```
## 4 45.53177 -73.57241  7.313450  5  2 17.7    0.3       0
## 5 45.51562 -73.57213 12.470333  6  6 21.0    0.0       1
## 6 45.51080 -73.57822 15.558917  5  3 10.2    0.4       0
```

3. Exploratory Data Analysis (EDA)

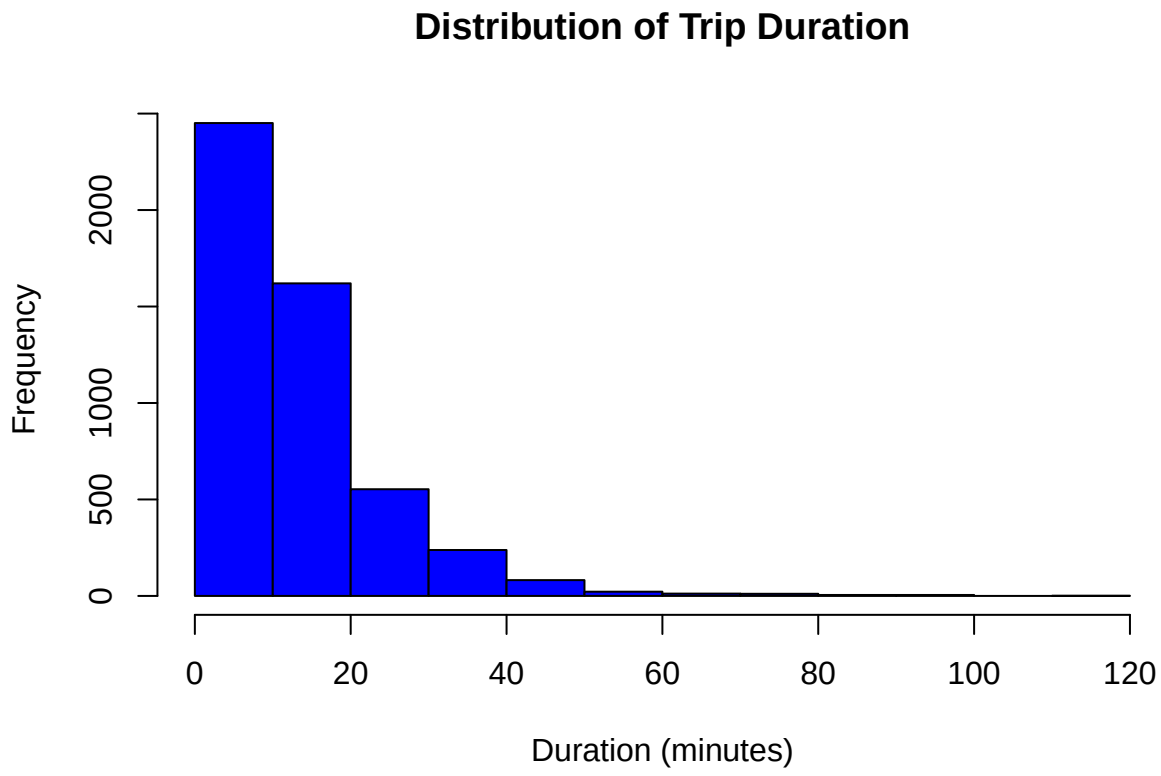
3.1 Trip duration (dur)

```
# Basic summary statistics
summary(bixi$dur)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.022   5.991  10.221  13.221  16.717 112.316
```

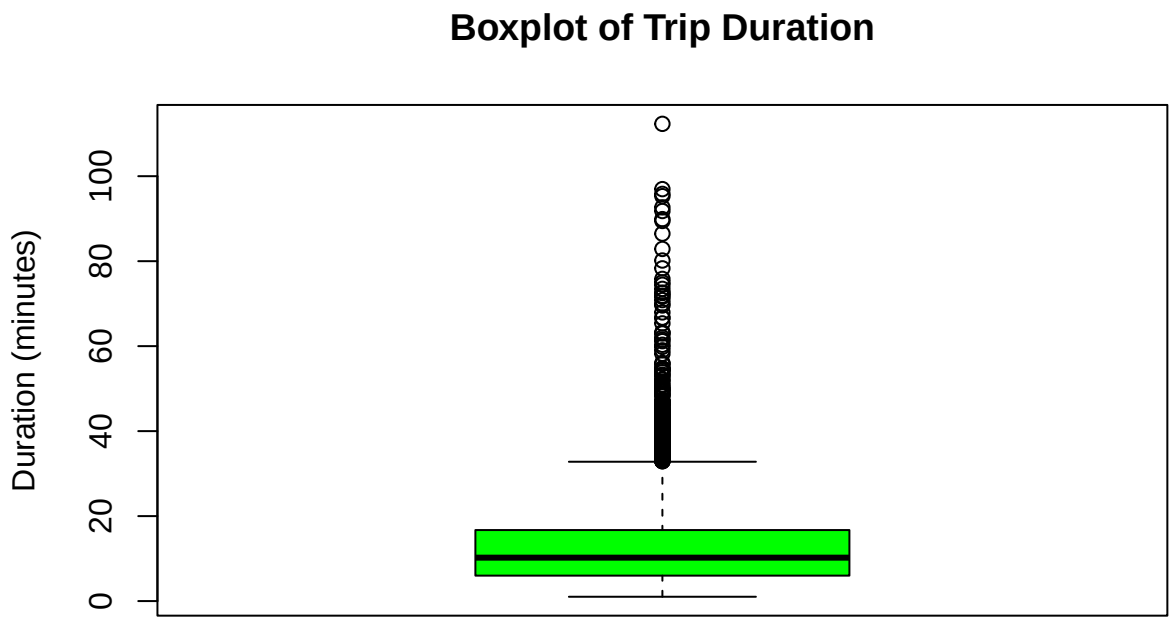
First, we obtain the minimum trip duration is 1.022 min while the maximum is 112.316 min, with mean of 13.221 min and median of 10.221 min.

```
# Histogram of trip duration
hist(bixi$dur,
     main = "Distribution of Trip Duration",
     xlab = "Duration (minutes)",
     col = "blue",
     border = "black")
```



The histogram shows that it is positively skewed (right-skewed). Moreover, most BIXI trips are short in duration, with the most from lower values (around 10 min).

```r
# Boxplot of trip duration
boxplot(bixi$dur,
        main = "Boxplot of Trip Duration",
        ylab = "Duration (minutes)",
        col = "green")
```

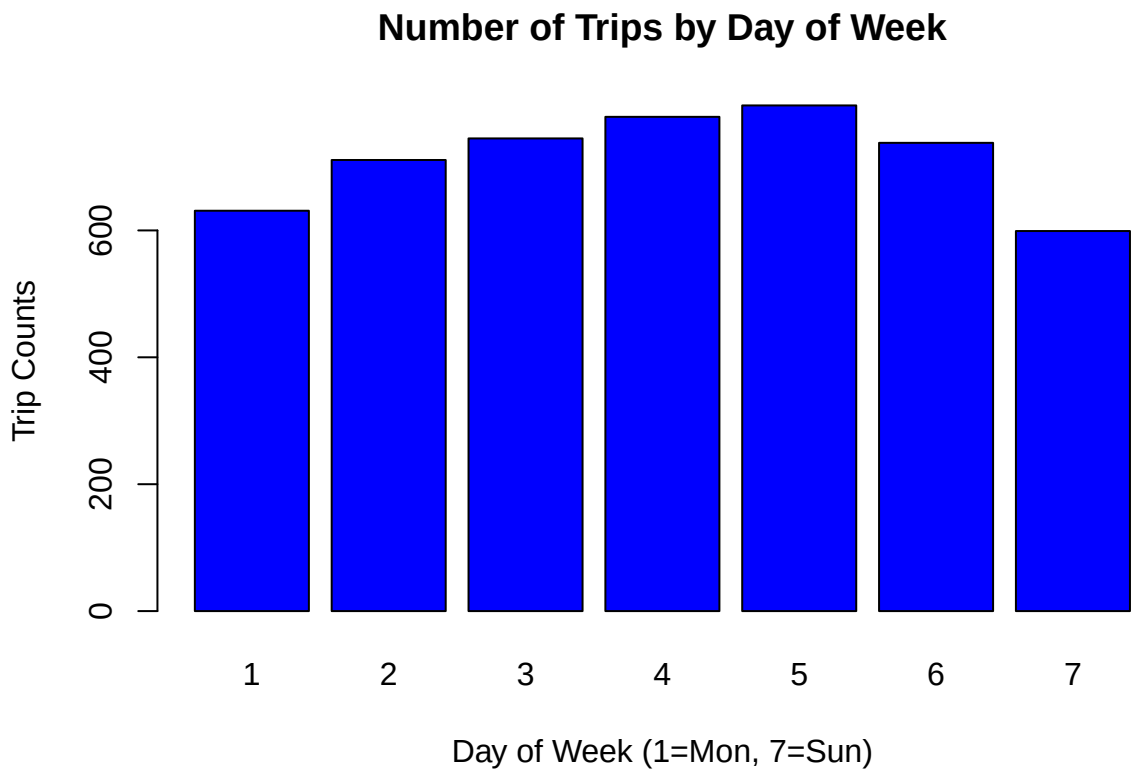


The boxplot indicates that the median is closer to the lower quantile than the upper quantile, and there are many outliers above the maximum - suggests that a lot of trip duration are longer than usual.

From a business perspective, this suggests that while most riders use BIXI for short commutes, a smaller numbers rent it for more long leisure trips.

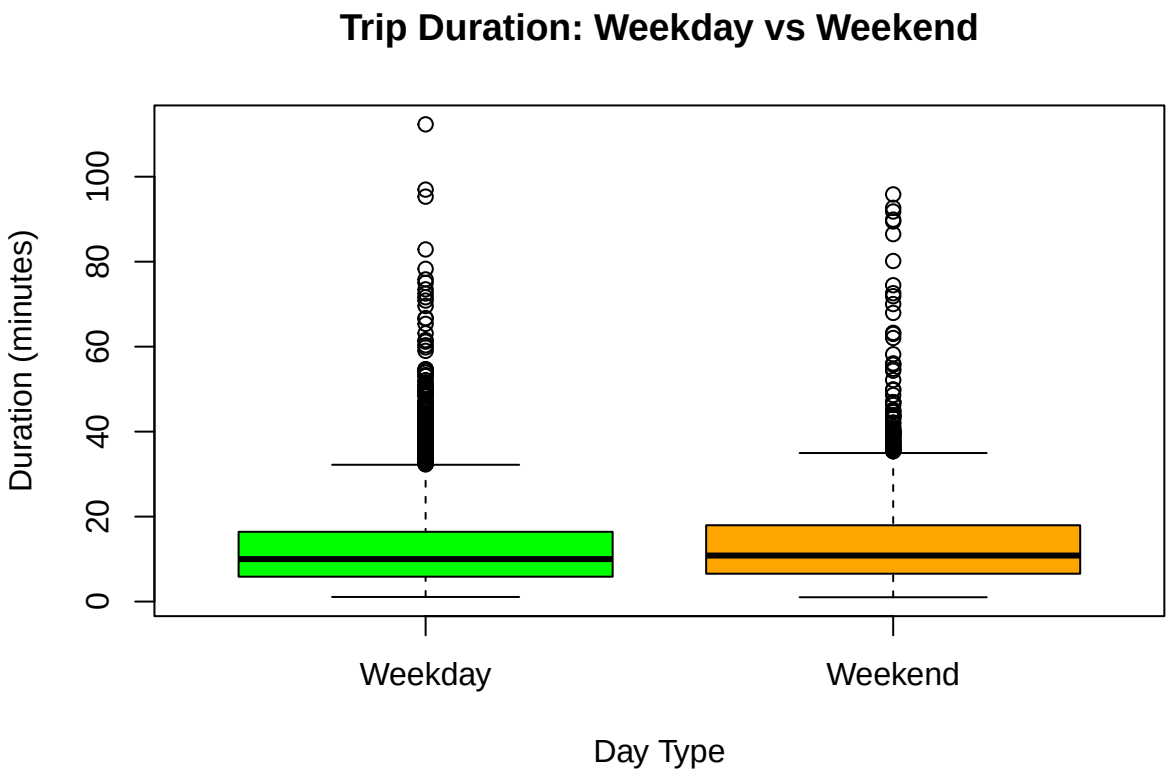
3.2 Day of the week (jj) and Weekend (weekend)

```r
# Bar plot: number of trips by day of week
barplot(table(bixi$jj),
        main = "Number of Trips by Day of Week",
        xlab = "Day of Week (1=Mon, 7=Sun)",
        ylab = "Trip Counts",
        col = "blue")
```



Trips are most popular on Fridays, following by Thursdays and Wednesdays. The lowest count is on Sundays, suggesting bike is primarily used for weekday commuting but still observes a significant numbers for leisure usage on weekends.

```
# Boxplot: duration by weekend (0=Weekday, 1=Weekend)
boxplot(dur ~ weekend, data = bixi,
        names = c("Weekday", "Weekend"),
        main = "Trip Duration: Weekday vs Weekend",
        xlab = "Day Type",
        ylab = "Duration (minutes)",
        col = c("green", "orange"))
```



From this comparing boxplots, we see that weekend trips have a slightly higher median and a higher upper quantile. While the weekend distribution shows a longer right tail, with more trip durations around 100 min area, but the overall difference between weekday and weekend distributions is modest.

```
# Mean trip duration by group
tapply(bixi$dur, bixi$weekend, mean)
```

```
##        0        1
## 12.80640 14.35699
```

```
# Median trip duration by group
tapply(bixi$dur, bixi$weekend, median)
```

```
##        0        1
## 10.00143 10.83205
```

- Mean duration: Weekdays = 12.8 min, Weekends = 14.4 min
- Median duration: Weekdays = 10.0 min, Weekends = 10.8 min

This supports the idea that weekend trips tend to be longer on average, though the difference is not dramatic.

3.3 Weather condition (temp, prec)

```
summary(bixi$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.20   15.70   20.10   18.95   22.40   28.70
```

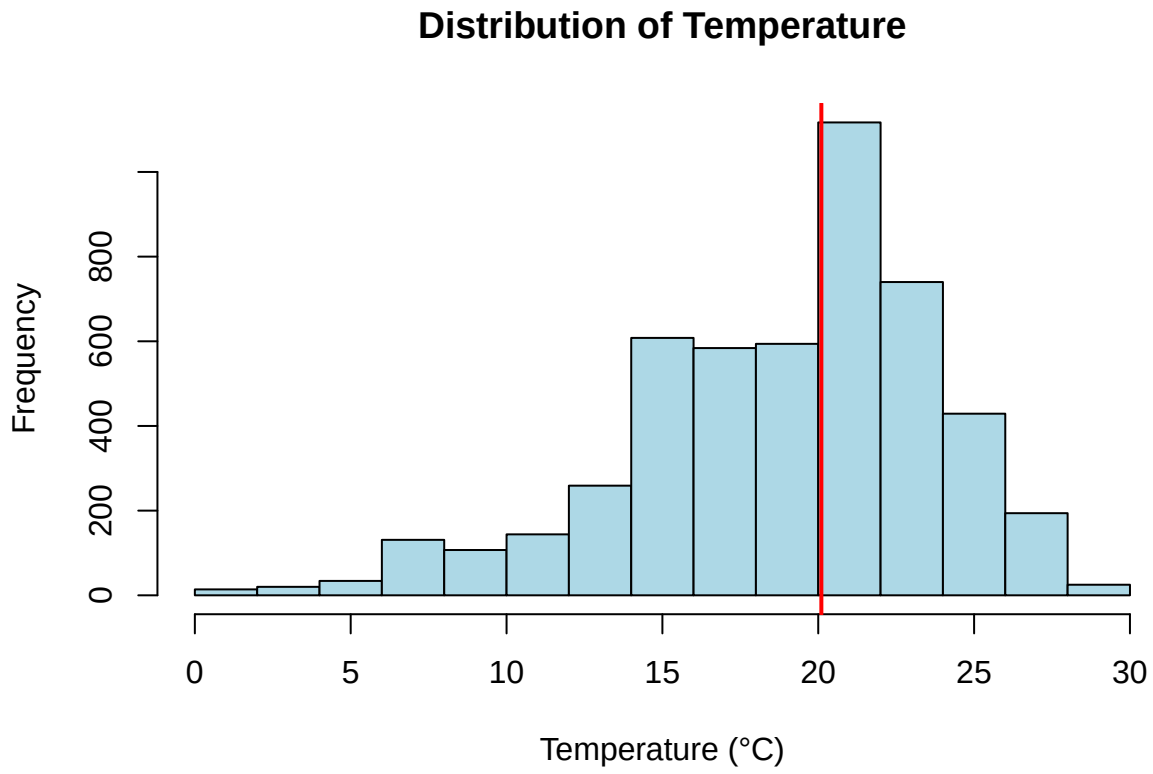
```
summary(bixi$prec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.300   2.285   1.700 150.200
```

The summary shows that most trips happen during the cool condition (15 to 22). The minimum number 1.20 suggests that it might occur in early-season or outliers, while maximum temperature 28.7 is warm but reasonable form summer or fall

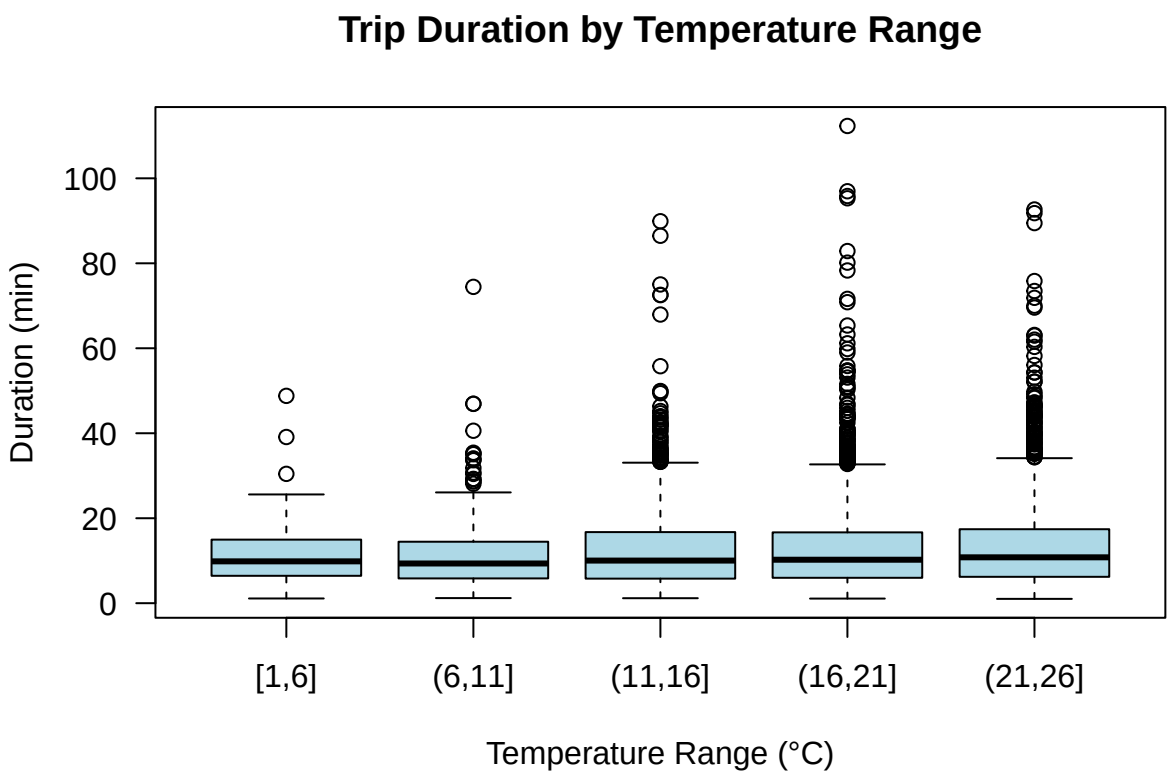
Moreover, most days have just little to no rain since the median is 0.300. There are some extreme values (up to 150mm of rainfall) are present - they can be outliers or some heavy rain days.

```r
# Histogram of temperature
hist(bixi$temp,
     main = "Distribution of Temperature",
     xlab = "Temperature (°C)",
     col = "lightblue",
     border = "black")

# Add a vertical line at the median
abline(v = median(bixi$temp, na.rm = TRUE), col = "red", lwd = 2)
```

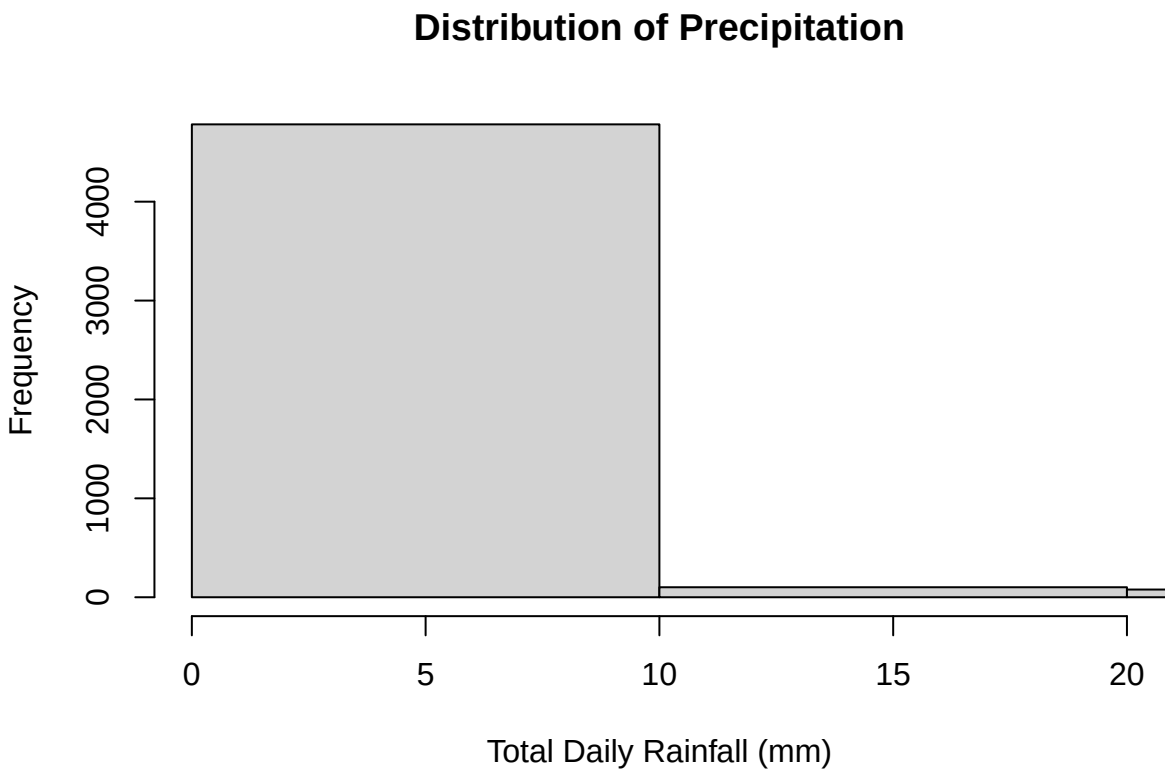



```r
# Bin temperature into 5°C ranges
bixi$temp_bin <- cut(bixi$temp,
                     breaks = seq(floor(min(bixi$temp)),
                                  ceiling(max(bixi$temp)), by = 5),
                     include.lowest = TRUE)

# Boxplot of duration by temperature bins
boxplot(dur ~ temp_bin, data = bixi,
```

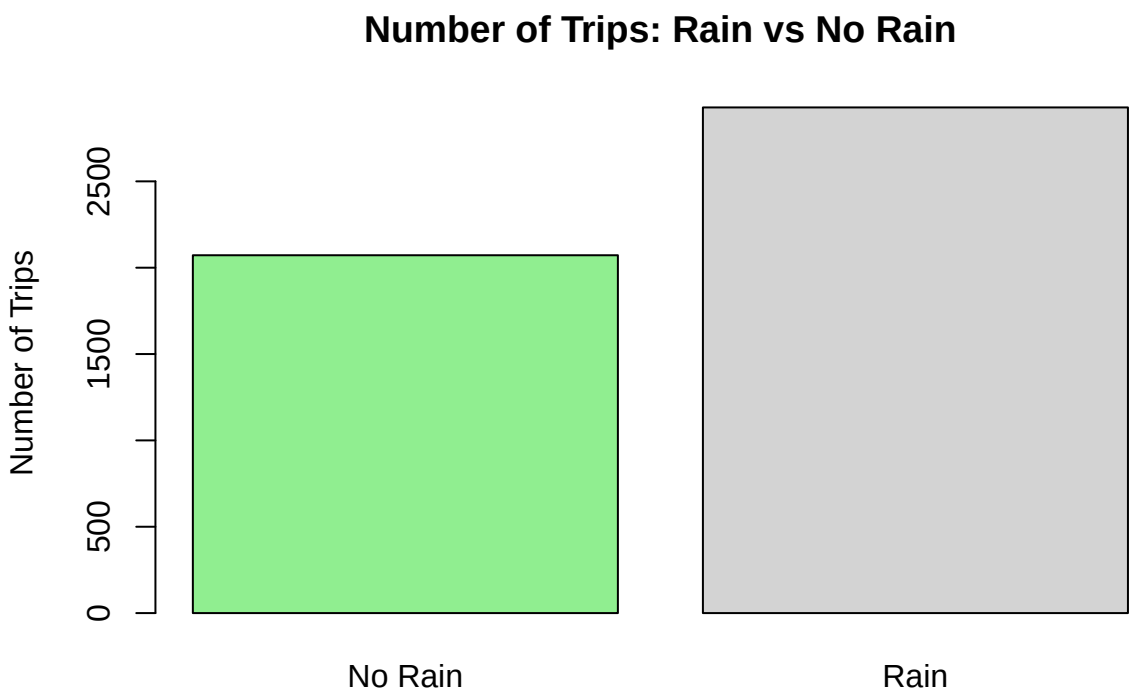
```
        main = "Trip Duration by Temperature Range",
        xlab = "Temperature Range (°C)",
        ylab = "Duration (min)",
        col = "lightblue", las = 1)
```



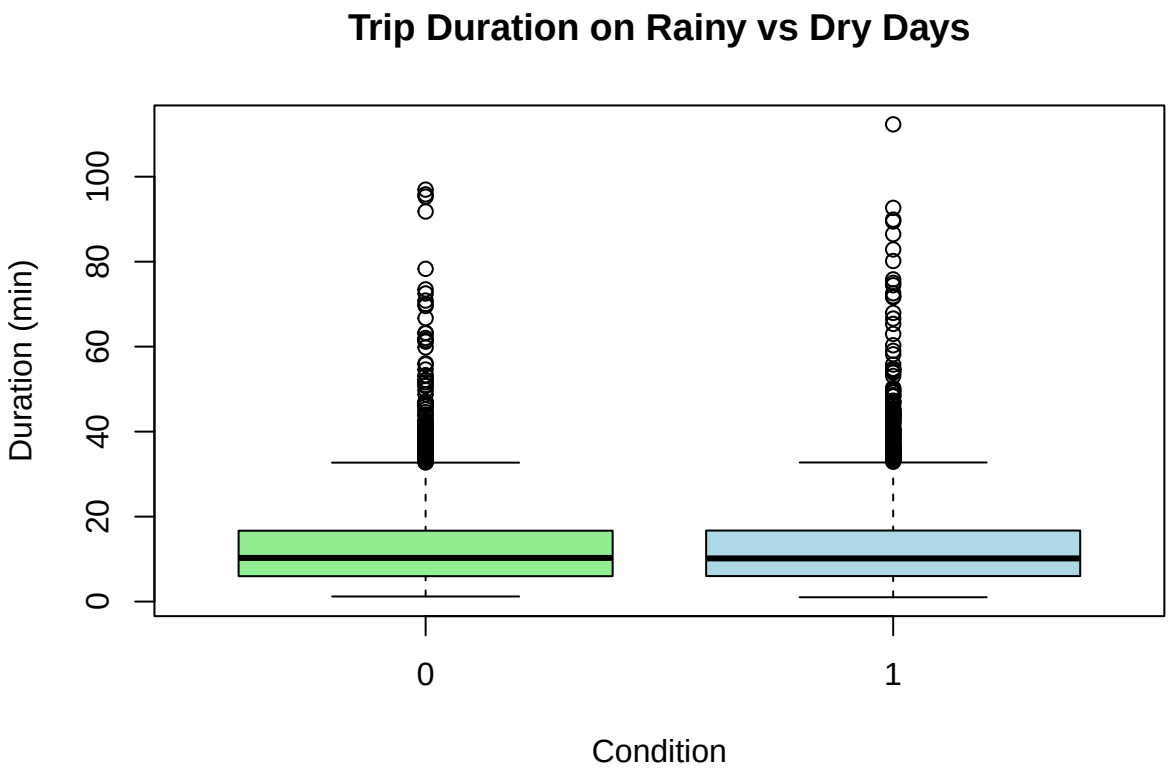

```
# Histogram of precipitation
hist(bixi$prec,
     main = "Distribution of Precipitation",
     xlab = "Total Daily Rainfall (mm)",
     col = "lightgray",
     border = "black",
     xlim = c(0, 20))  # zoom in on most values (ignore extreme 150mm)
```




```r
# Create rain indicator: 1 = rain, 0 = no rain
bixi$rain <- ifelse(bixi$prec > 0, 1, 0)

# Barplot of rainy vs non-rainy days
barplot(table(bixi$rain),
        names.arg = c("No Rain", "Rain"),
        main = "Number of Trips: Rain vs No Rain",
        col = c("lightgreen", "lightgray"),
        ylab = "Number of Trips")
```




```
# Boxplot of duration by rain/no rain
boxplot(dur ~ rain, data = bixi,
        main = "Trip Duration on Rainy vs Dry Days",
        xlab = "Condition",
        ylab = "Duration (min)",
        col = c("lightgreen", "lightblue"))
```


Trip Duration on Rainy vs Dry Days
Duration (min)
0
20
40
60
80
100
0
1
Condition