Lab 1B

The following dataset, titled "Rain in Australia," was obtained via Kaggle.com via author Joe Young et. al.

This dataset contains daily weather observations from numerous Australian weather stations during the year 2016. The dataset measured variables: Date, Location, MinTemp (C), MaxTemp (C), Rainfall (mm), Evaporation (mm per 24 hours), Sunshine (hours per day), WindGustDir (strongest gust in 24 hours), WindGustSpeed (km/h), WindDir9am (direction at 9 AM), ..., RainToday (Yes/No) and RainTomorrow (Yes/No).

```
rain = read.csv("weatherAUS.csv", sep=",", header=TRUE)
rain = na.omit(rain)
head(rain, 4)
```

```
##        Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir
## 1467 1/3/16    Moree    18.5    19.7     22.4        30.4      0.0          ESE
## 1468 1/4/16    Moree    17.6    28.2     20.6         2.2      3.5            E
## 1469 1/5/16    Moree    19.7    27.3      0.2         4.2      3.2            E
## 1470 1/6/16    Moree    17.0    29.8      0.6         4.8     12.0          SSW
##      WindGustSpeed WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am
## 1467            39        ENE          E           19           17          92
## 1468            33          E         SW           26           11          77
## 1469            54         SE          N           11           17          78
## 1470            50         SW         SW           20           28          78
##      Humidity3pm Pressure9am Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm
## 1467          91      1010.1      1010.3        8        8    18.9    18.6
## 1468          54      1012.4      1009.5        8        7    19.7    25.7
## 1469          71      1010.3      1006.6        7        7    23.2    25.7
## 1470          44      1008.6      1007.0        5        2    21.1    27.8
##      RainToday RainTomorrow
## 1467       Yes          Yes
## 1468       Yes           No
## 1469        No           No
## 1470        No           No
```

Question 1

Using 1000 randomly sampled observations from the dataset `rain`, plot a scatter plot of `MaxTempF` (x) vs. `Humidity3PM` (y). `MaxTempF` should be the `MaxTemp` variable converted to degrees F. Color the point on the scatterplot blue if it rained that day and red if it did not rain that day. The plot should have x- and y-axis labels in addition to a title. After the plot, include a discussion on observations from the scatterplot.

Hints: - To sample from `rain`, use `sample_n()` from package `dplyr`. - To create MaxTempF, use conversion (9/5)C + 32 to change C to F. - Create a second new column named Colors that assigns a color based on RainToday variable.

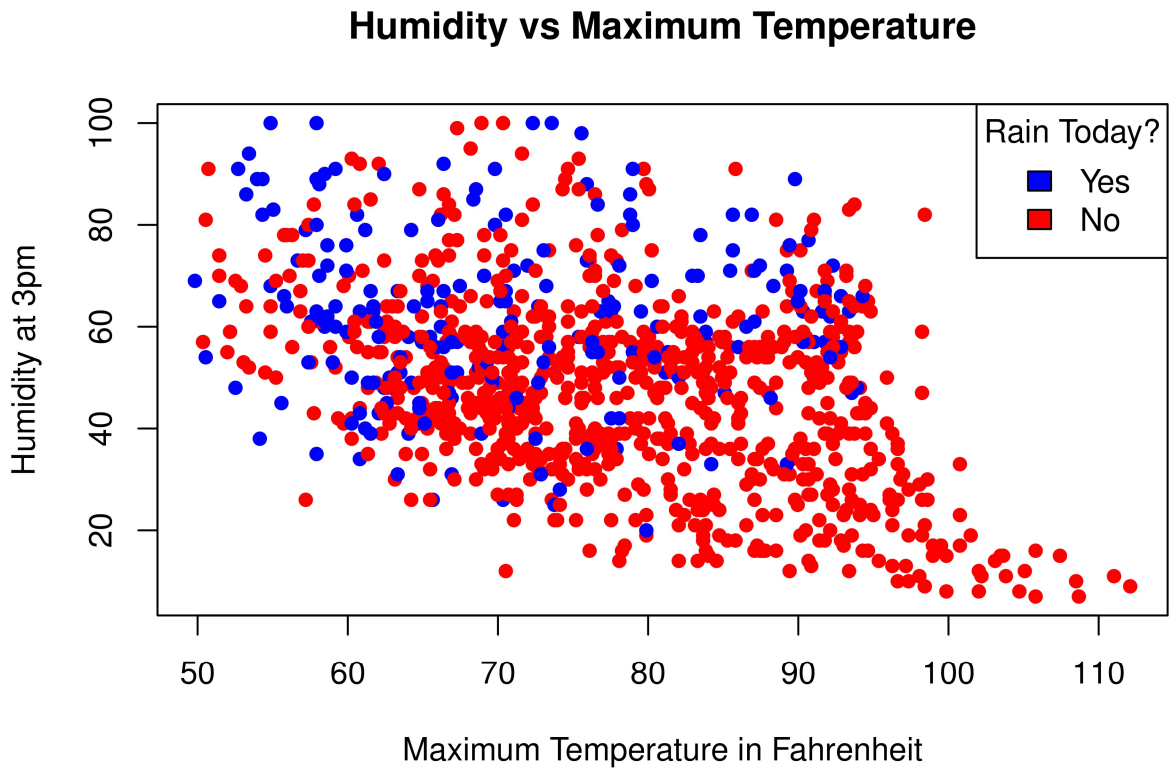
```
library(dplyr)
```

```
##
## 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
```

```
rainSample = sample_n(rain,1000)
MaxTempF = ((9 * rainSample$MaxTemp )/5) + 32
Colors= ifelse(rainSample$RainToday == "No", "red", "blue")#red = no, blue = yes
plot(MaxTempF,rainSample$Humidity3pm, pch = 16,
     xlab = "Maximum Temperature in Fahrenheit", ylab = "Humidity at 3pm",
     main = "Humidity vs Maximum Temperature", col = Colors)
legend("topright", legend = c("Yes","No"), fill = c("blue", "red"),
       title = "Rain Today?")
```



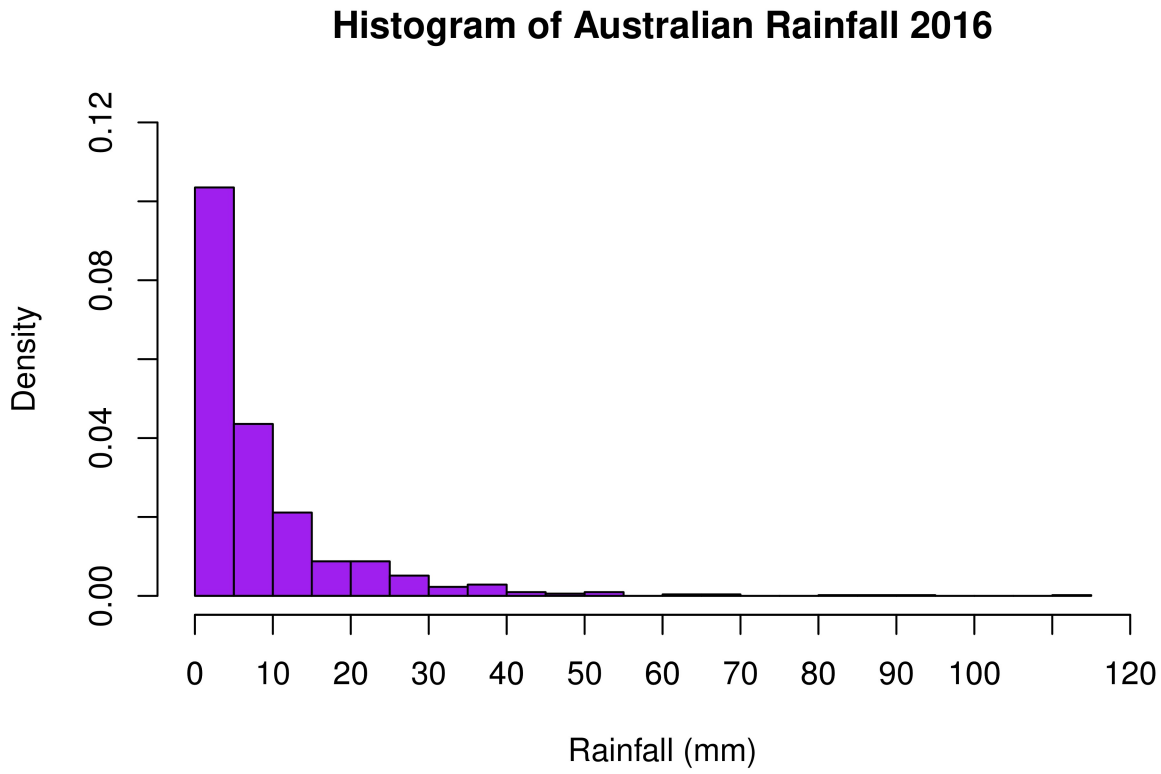
The negative slope on the scatter plot shows an inverse relationship between the Maximum Temperature and Humidity at 3pm with potential outliers. This means that the higher the max temperature the lower the humidity level. It also shows that at higher max temperature (or at lower humidity level) it is less likely to rain.

Question 2

Part A

Plot a histogram of `Rainfall`, including all values from `rain` dataset given that it had rained that day. Display density on the y-axis. Label the axes and give the graph a title. Colorize the bars. Include a discussion after the plot.

```
hist(x = rain$Rainfall[rain$RainToday == "Yes"], freq = F,xlim = c(0, 120),
     ylim = c(0,0.12), xlab = "Rainfall (mm)",
     main = "Histogram of Australian Rainfall 2016", col = "purple",
     breaks = 20)
axis(1, at = seq(0,120,10))
```



From the histogram, the data is right skewed which means the mean is greater than the median. The median can be estimated to about 5mm. According to the histogram, majority of rainy days had a rainfall between 0 to 5mm of rain.

Part B

Find the mean and median of the `Rainfall` in mm for days that it rained in Australia. Do these numbers correspond with the image in part A?

```
rainfallMean = mean(rain$Rainfall[rain$RainToday == "Yes"])
sprintf('Mean of Rainfall for days it rained: %fmm', rainfallMean)
```

```
## [1] "Mean of Rainfall for days it rained: 8.779924mm"
```

```
rainfallMedian = median(rain$Rainfall[rain$RainToday == "Yes"])
sprintf('Median of Rainfall for days it rained: %.2fmm', rainfallMedian)
```

```
## [1] "Median of Rainfall for days it rained: 4.80mm"
```

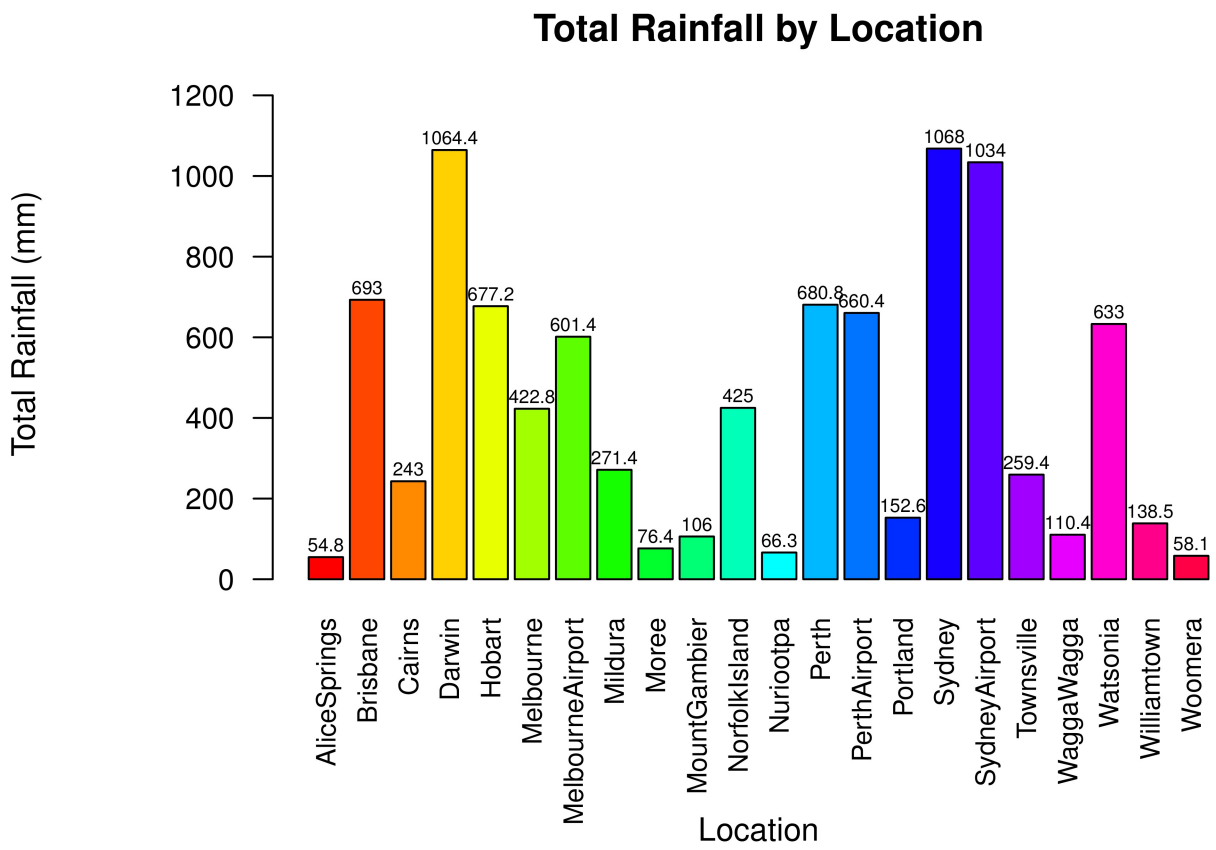
In right skewed histograms, the mean is greater than the median. The histogram in part A is right skewed and the result corresponds because the mean (approx. 8.78) is greater than the median (4.8). The median is also close to the estimate in part A.

Question 3

Part A

Plot a bar graph representing *total* `Rainfall` in 2016 by `Location`. Make sure to label the x- and y- axes, colorize the bars, include a title, and include the appropriate response labeled above each bar. Utilize all observations from dataset `rain`. Include a discussion after the plot.

```
cities = tapply(X = rain$Rainfall, INDEX = rain$Location, FUN = sum)
par(mar = c(7,7,2,0) + 0.1, mgp = c(6,1,0))
citiesPlot <- barplot(cities, ylim = c(0, max(cities)+ 200), xlab = "Location",
                      ylab = "Total Rainfall (mm)",
                      las = 2, cex.names = 0.8, cex.axis = 0.9,
                      col = rainbow(22), main = "Total Rainfall by Location")
text(citiesPlot, y = cities+30, labels = as.character(cities), cex = 0.6)
```

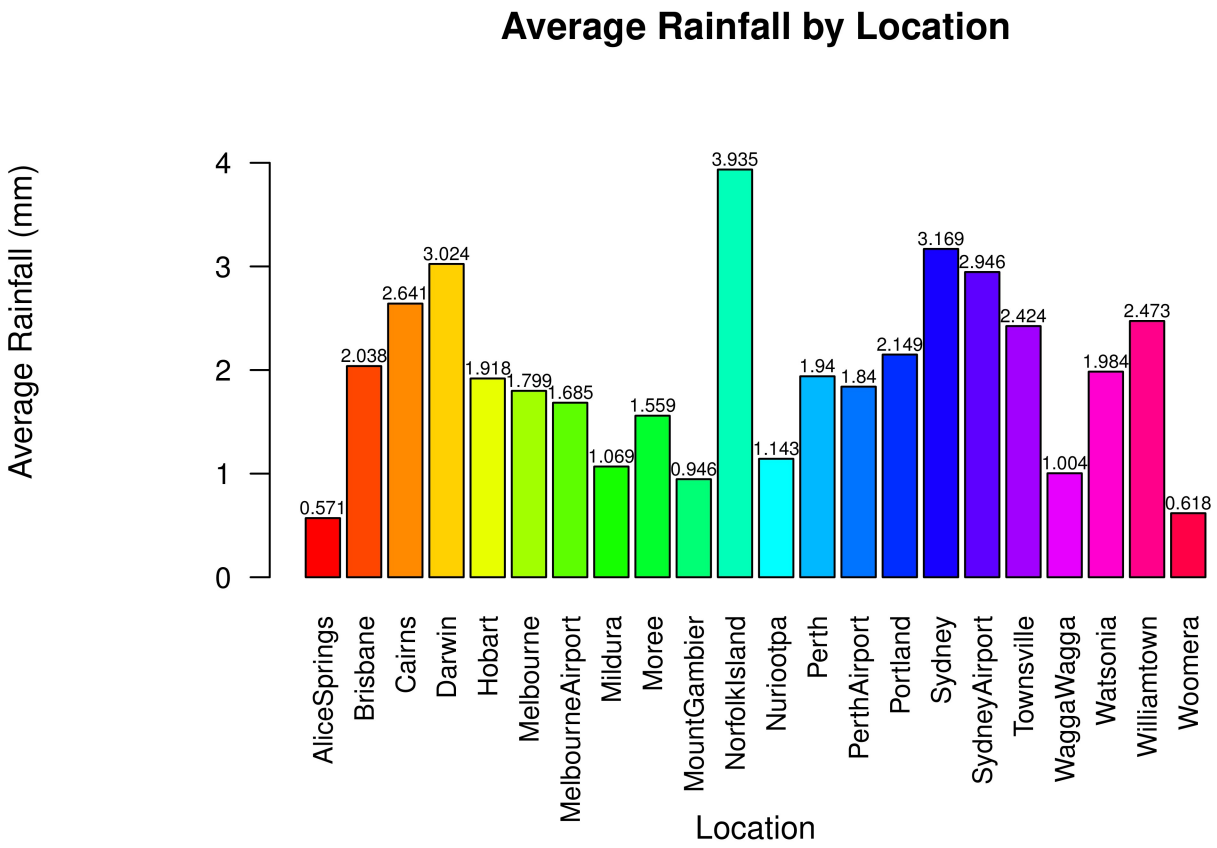


This bar graph shows the total rainfall by location. Sydney has the highest total rainfall, while AliceSprings has the least total rainfall. (A legend was not included because it would be repetitive)

Part B

Plot a bar graph representing *average* `Rainfall` per day based on `Location`. Make sure to label the x- and y- axes, colorize the bars, include a title, and include the appropriate response labeled above each bar. Utilize all observations from dataset `rain`. Include a discussion after the plot.

```
citiesM = tapply(X = rain$Rainfall, INDEX = rain$Location, FUN = mean)
par(mar = c(7,7,2,0) + 0.1, mgp = c(6,1,0))
citiesMPlot <- barplot(citiesM, ylim = c(0, max(citiesM)+1), xlab = "Location",
                       ylab = "Average Rainfall (mm)", las = 2, cex.names = 0.8,
                       cex.axis = 0.9, col = rainbow(22),
                       main = "Average Rainfall by Location")
text(citiesMPlot, y = citiesM+0.1, labels = as.character(round(citiesM,3)),
     cex = 0.6)
```

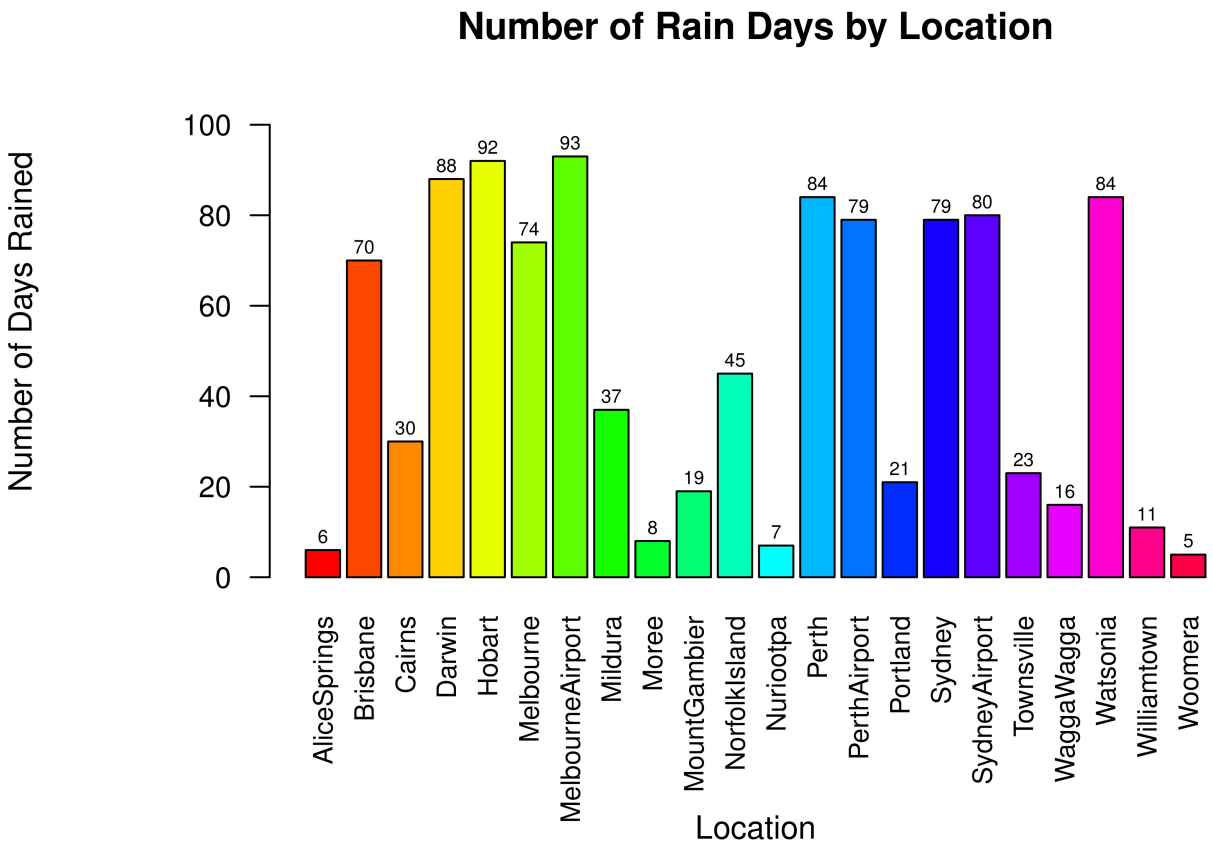


This bar graph shows the average rainfall per day based on location. Norfolk Island has the highest average rainfall per day, while AliceSprings has the least. (A legend was not included because it would be repetitive and the values above the bar have been rounded to 3 decimals)

Part C

Plot a bar graph representing *total* count of days rained in 2016 based on `Location`. Make sure to label the x- and y- axes, colorize the bars, include a title, and include the appropriate response labeled above each bar. Utilize all observations from dataset `rain`. Include a discussion after the plot.

```
citiesC = table(rain$Location[rain$RainToday == "Yes"])
par(mar = c(7,7,2,0) + 0.1, mgp = c(6,1,0))
citiesCPlot <- barplot(citiesC, ylim = c(0, max(citiesC)+20), xlab = "Location",
                       ylab = "Number of Days Rained", las = 2, cex.names = 0.8,
                       cex.axis = 0.9, col = rainbow(22),
                       main = "Number of Rain Days by Location")
text(citiesCPlot, y = citiesC+3, labels = as.character(citiesC), cex = 0.6)
```

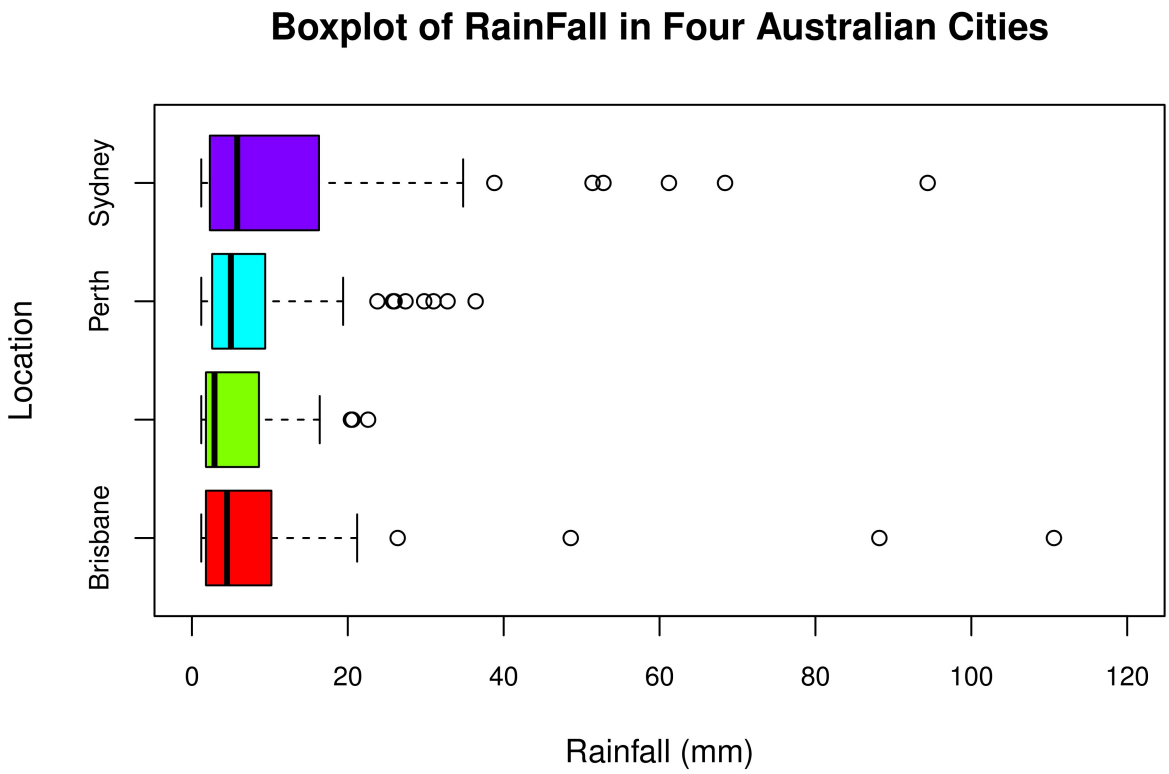


This bar chart shows the total number of days it rained based on location. Melbourne Airport experienced the most days with rain, while Woomera had the least days with rain. (A legend was not included because it would be repetitive)

Question 4

For locations `Brisbane`, `Melbourne`, `Perth`, and `Sydney`, plot side by side boxplots of the variable `Rainfall` for days where it rained. The graph should include a title, x- and y-axis labels, and unique colors for the 4 boxplots. The data should include all observations from the `rain` dataset. Include a discussion after the plot.

```
rainCity=rain[rain$Location %in% c('Brisbane', 'Melbourne', 'Perth', 'Sydney'),]
y = seq(0,120,5)
boxplot(Rainfall ~ Location, data = rainCity[rainCity$RainToday == "Yes", ],
        col = rainbow(4), xlab = "Rainfall (mm)",
        main = "Boxplot of RainFall in Four Australian Cities", horizontal = T,
        cex.axis = 0.8, ylim = c(0, 120))
```



This boxplot summarizes 5 information (median, Q1, Q3, min, max) about the rainfall in 4 different cities while showing outliers. The dark black line running across each box shows the median of the data. The circle unfilled dots show the outliers. The plot shows that the median of any of these four cities is less than 10mm.