HW3

Kelly "Scott" Sims

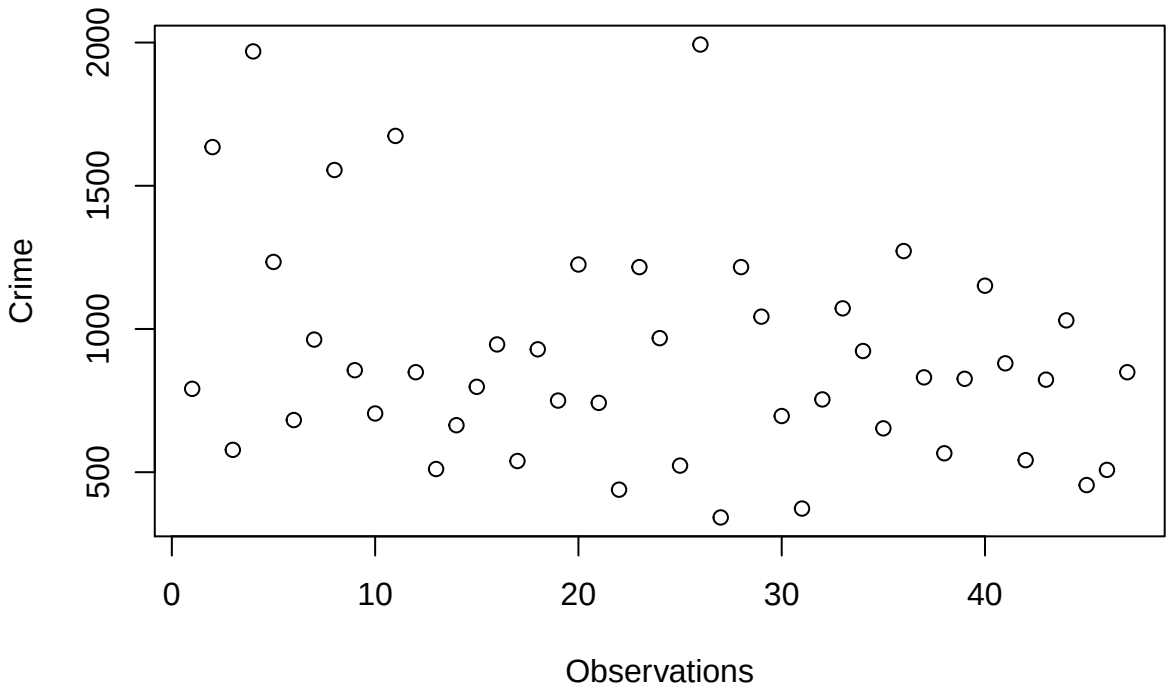
January 29, 2019

QUESTION 5.1

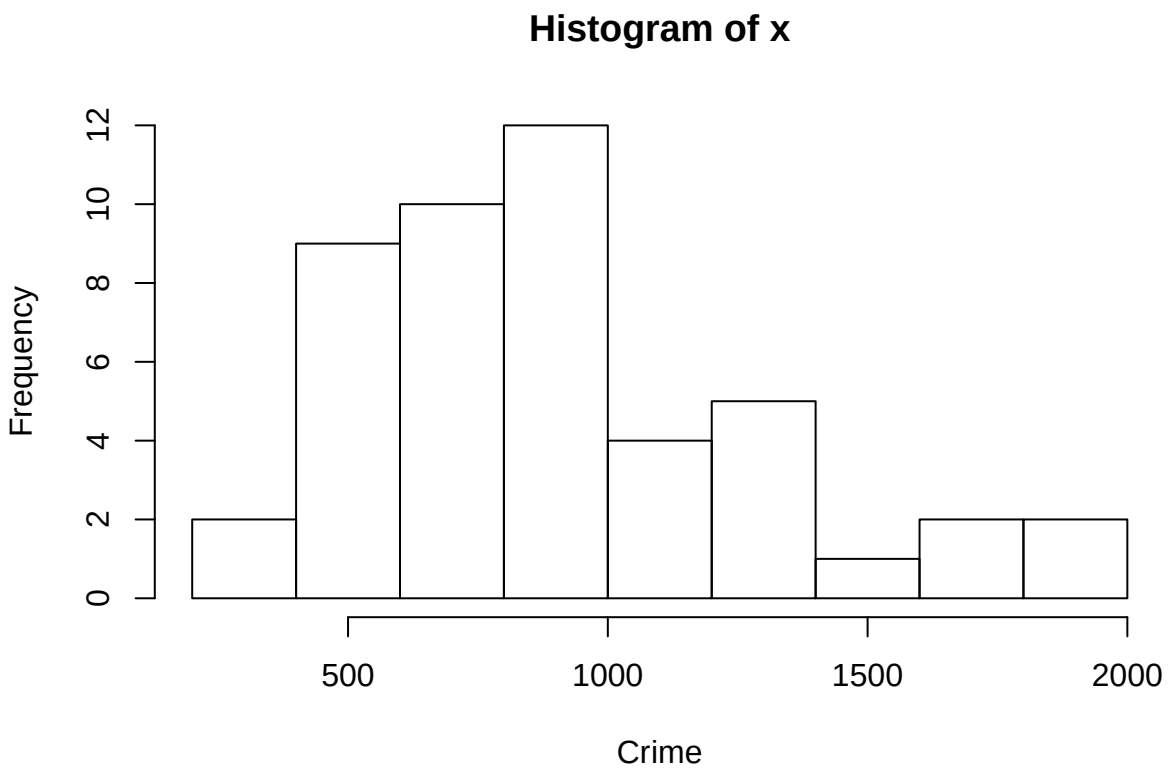
Using crime data from the file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

Let's first explore our data before finding any outliers

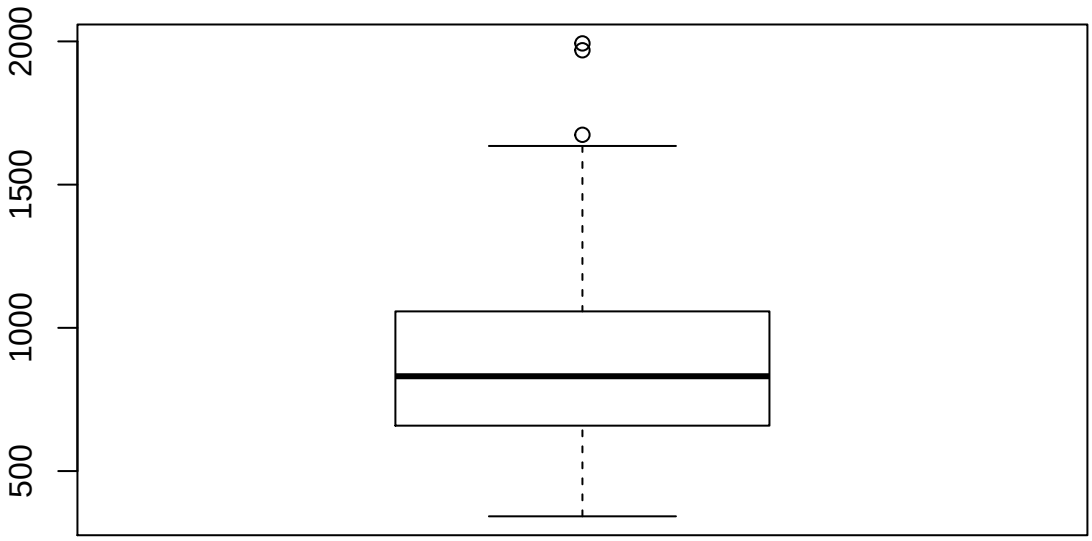
```
library(outliers)

data = read.table('uscrime.txt', header = TRUE)
x = data$Crime

plot(x = seq(1,length(x)), y=x, xlab = "Observations", ylab='Crime')
```




```
hist(x = x, xlab = 'Crime')
```




```
boxplot(x = x, xlab = 'Crime')
```




```
grubbs.test(x = x , type = 11)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  x
```

```
## G = 4.26880, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

*Conclusion

We can see from the plots above that the data is far from normally distributed. There is a right skew to the histogram with an upper bound outlier value of 1993

Question 6.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

Due to the randomness and volatility in the stock market, a change detection model would be helpful for identify a bullish or bearish trend for a specific stock. One of the many metrics tracked of stocks are their day to day, week to week, and month to month percent change. Using historical data/trends, critical values and threshold can be chosen from previous observations. For instances a bearish trend can be signaled from a certain amount of consecutive days of a price decrease in conjuction with a drastic percent change as the threshold instead of a specific price. E.g. if a previous bear trend occured during three days of downturn followed by a 20% drop, then that would be the baseline for detecting another bearish trend in the future.

Question 6.2

Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year.

Let's look at 1996 only first and try to define an algorithm to apply on the whole data set

```
temp = read.table('temps.txt', header = TRUE)

x = temp$X1996
avg = mean(x)
s = x[1]
C = 1
threshold = -30
for(i in seq(from = 2, to = length(x))){
  s = s + (x[i] - avg + C)

  s = min(0, s)

  if (s < threshold){
    day_change = temp$DAY[i]
    day_index = i
    break
  }


}

print(as.character(day_change))
```

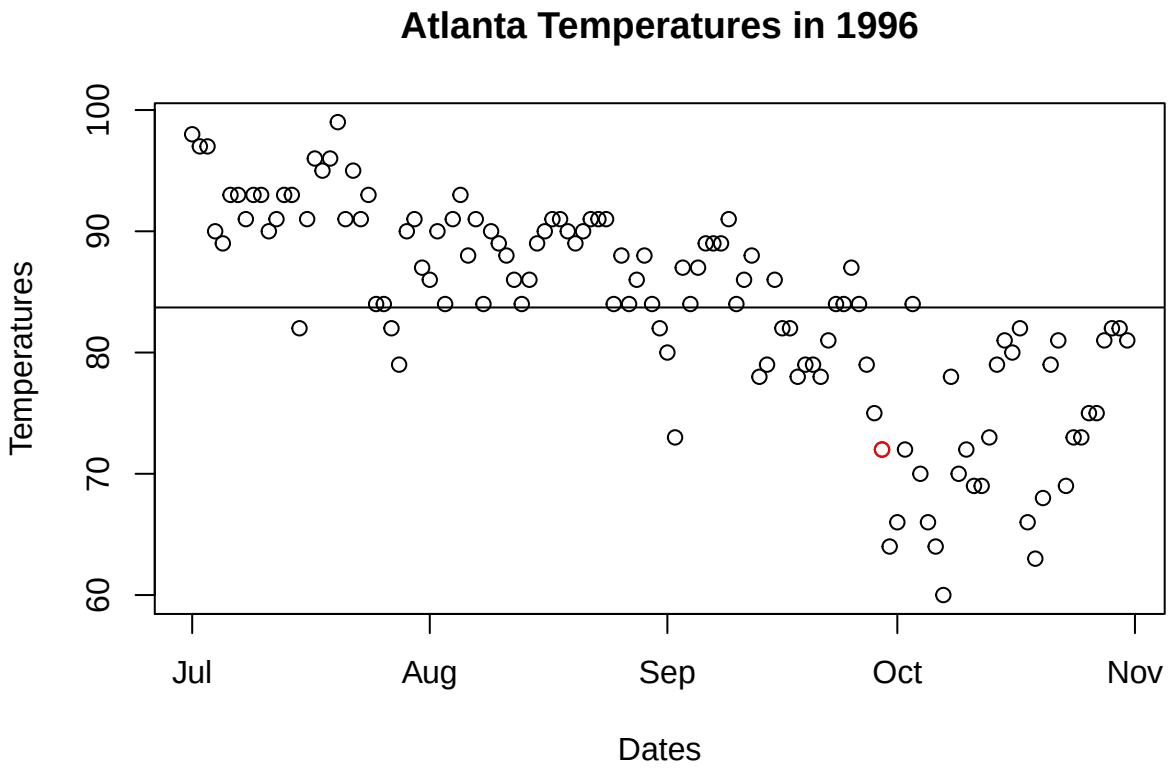
```
## [1] "29-Sep"
```

```
print(day_index)
```

```
## [1] 91
```

From the algorithm above, it appears that September 29th is the official end of Summer for 1996. It appears that there were 91 days of Summer given that the index of Sept. 29th is 91. Let's plot the data to see if this date does appear to be the official end of Summer

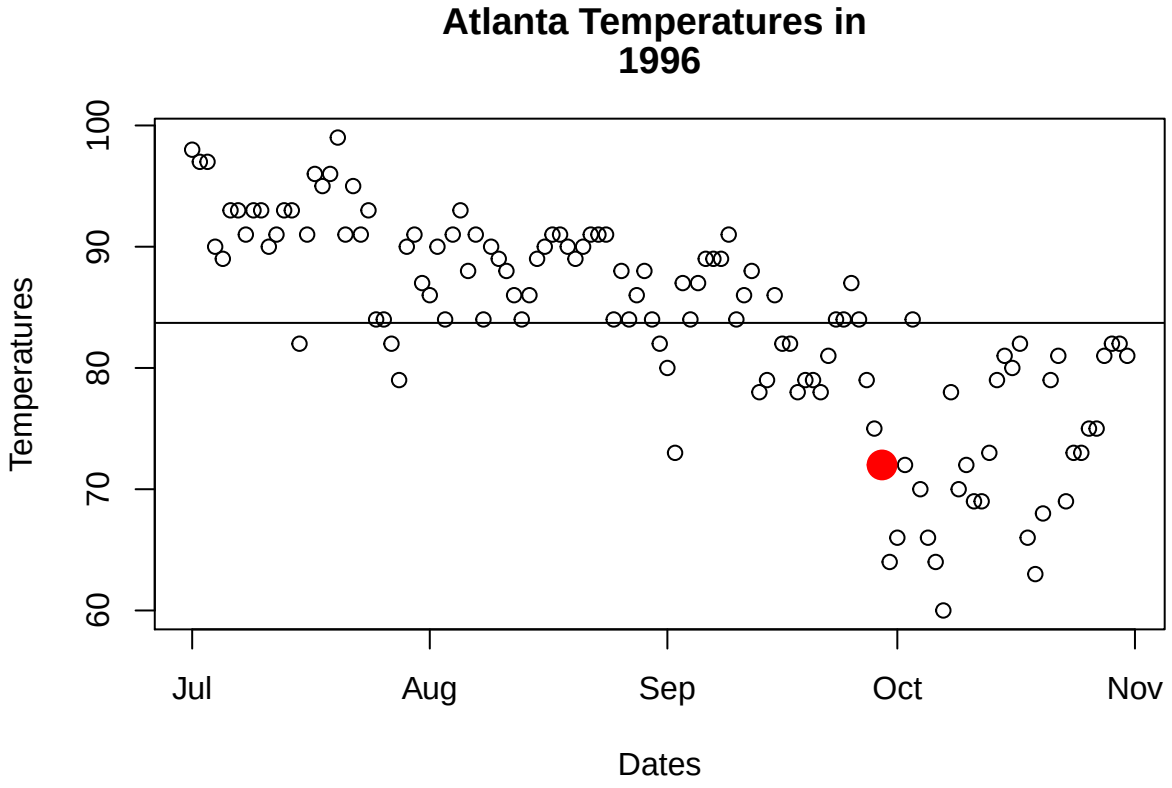
```
xp = as.Date(strptime(temp$DAY, format = '%d-%b'))
plot(x = xp, y = temp$X1996, xlab = 'Dates', ylab='Temperatures', main='Atlanta Temperatures in 1996')
points(x = xp[91], y= temp$X1996[91], col = 'red')
abline(h = avg)
```



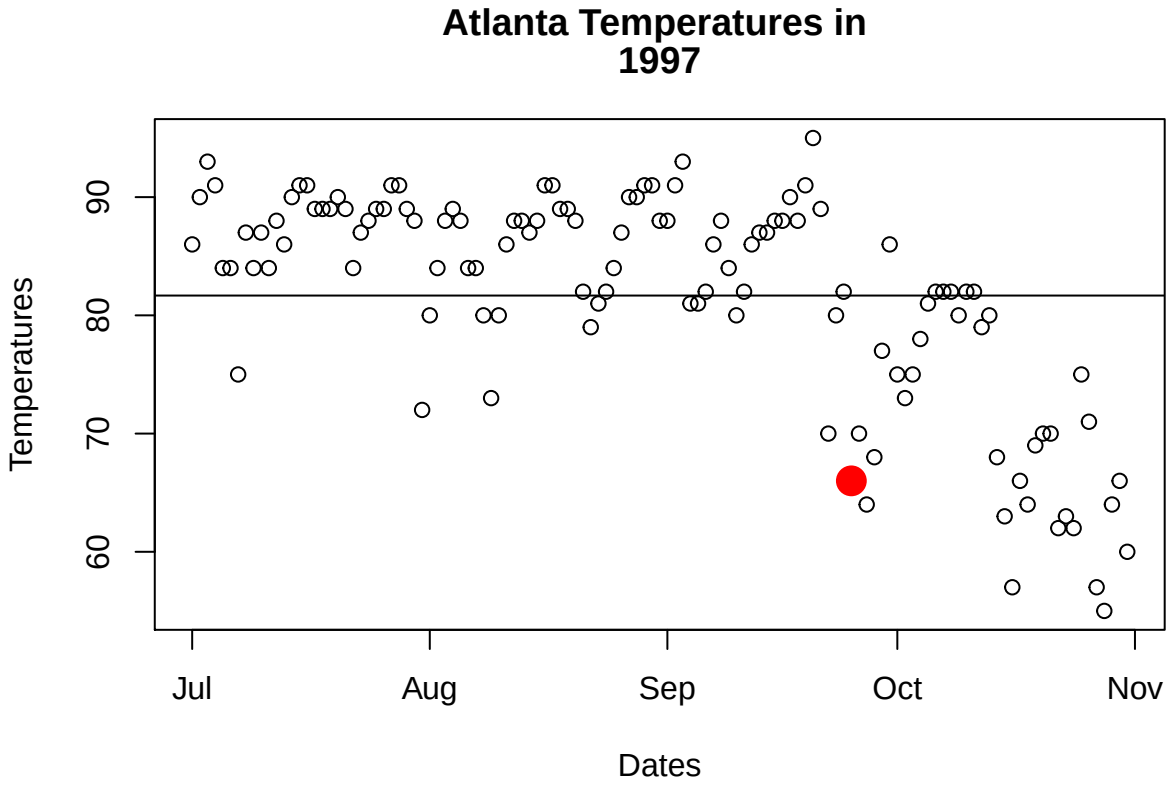
the date mentioned above appears to be pretty reasonable. The weather did appear to cool off a little before this date, but then it raised above the City's average temperature (straight line) before cooling off again for good where it never made it above the Average temperature threshold line. Let's apply this algorithm to all the dates in the data set. We will also be saving the index locations of the dates where summer officially ended to use in the last question. That variable will be called "weather_change_index"

```
xp = as.Date(strptime(temp$DAY, format = '%d-%b'))
weather_change_index = rep(0, 20)
for (i in seq(2,21)){

  x = temp[,i]
  avg = mean(x)
  C = 3
  threshold = -18
  s = x[1] - avg + C
```

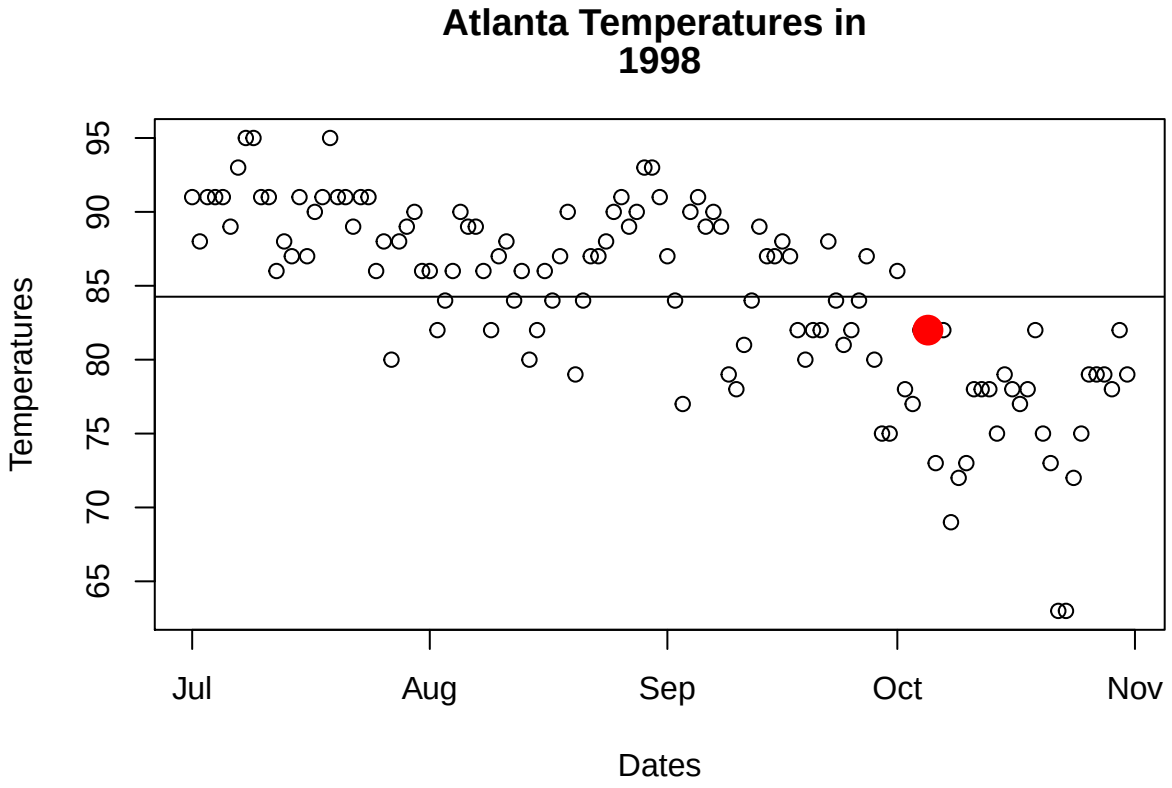
```
  for(j in seq(from = 2, to = length(x))){
    s = s + (x[j] - avg + C)
    s = min(0, s)

    if (s < threshold){
      day_change = temp$DAY[j-1]
      day_index = j-1
      weather_change_index[i-1] = day_index
      break
    }


  }

  plot(x = xp, y = x, xlab = 'Dates', ylab='Temperatures', main=c('Atlanta Temperatures in ', as.charac
  points(x = xp[day_index], y= x[day_index], col = 'red', pch=19, cex = 2)
  abline(h = avg)
  print(as.character(day_change))

}
```



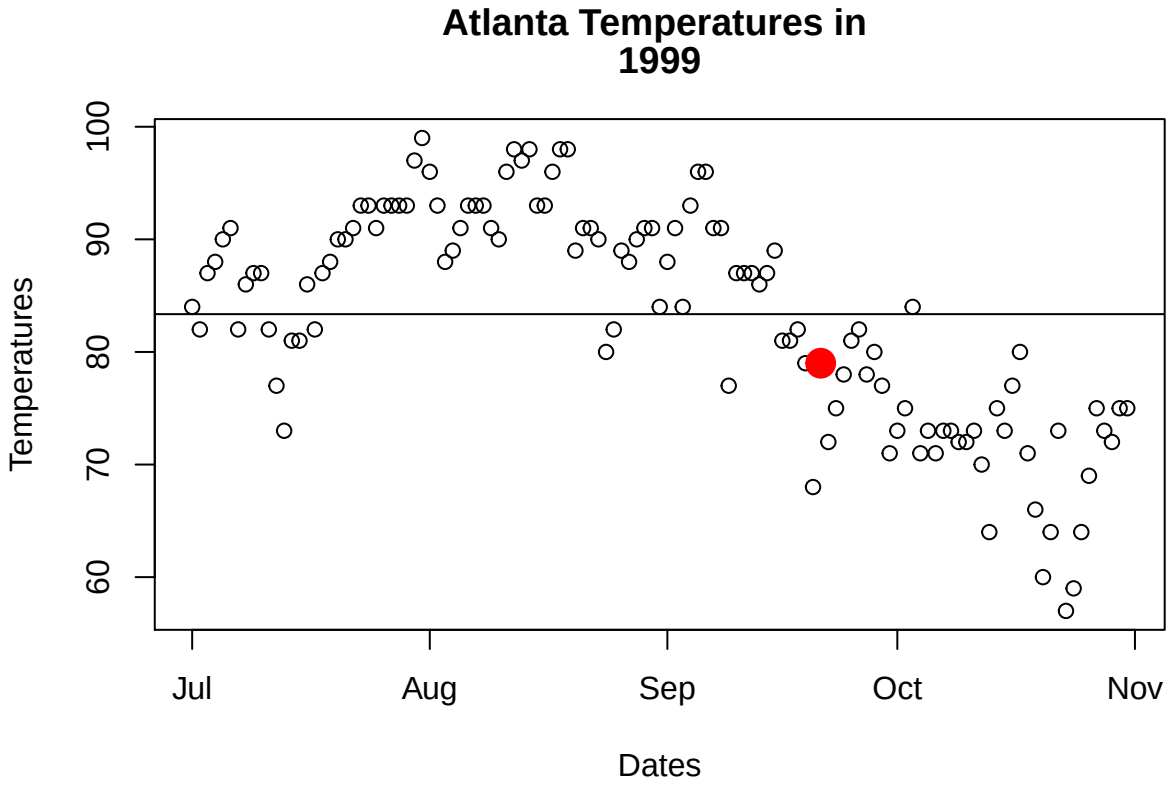

```
## [1] "29-Sep"
```



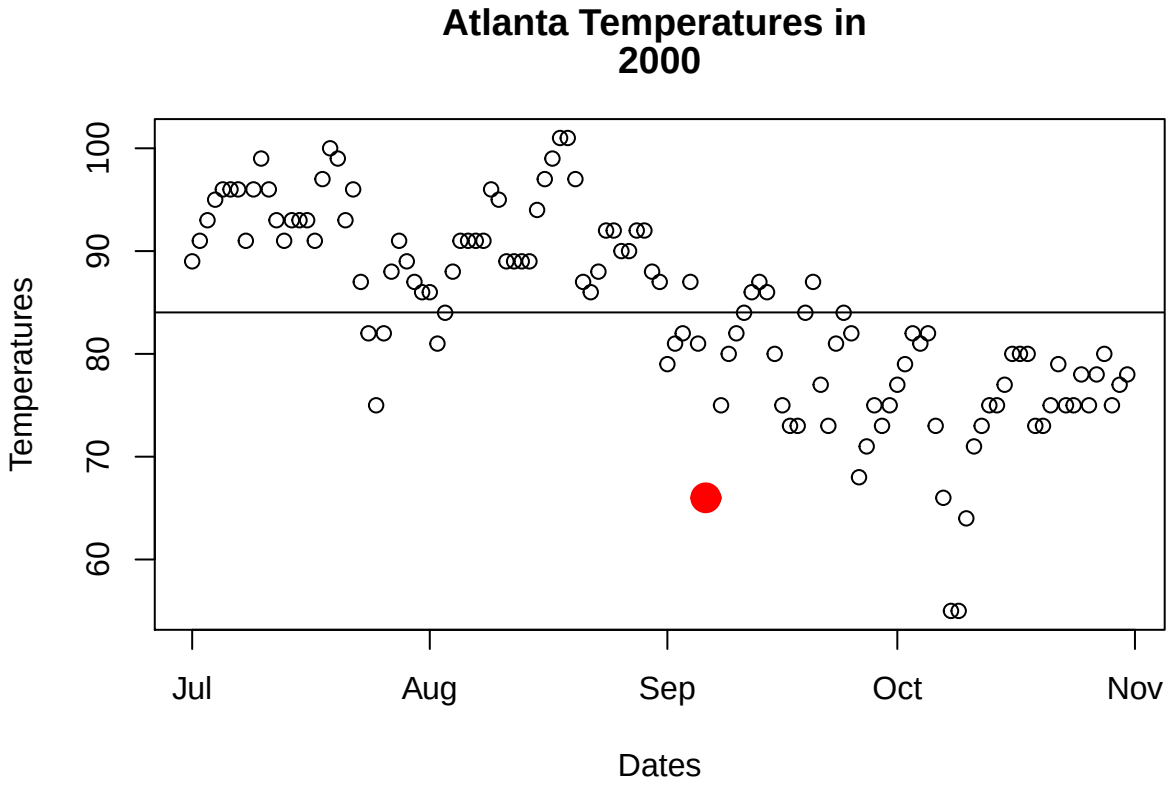

```
## [1] "25-Sep"
```



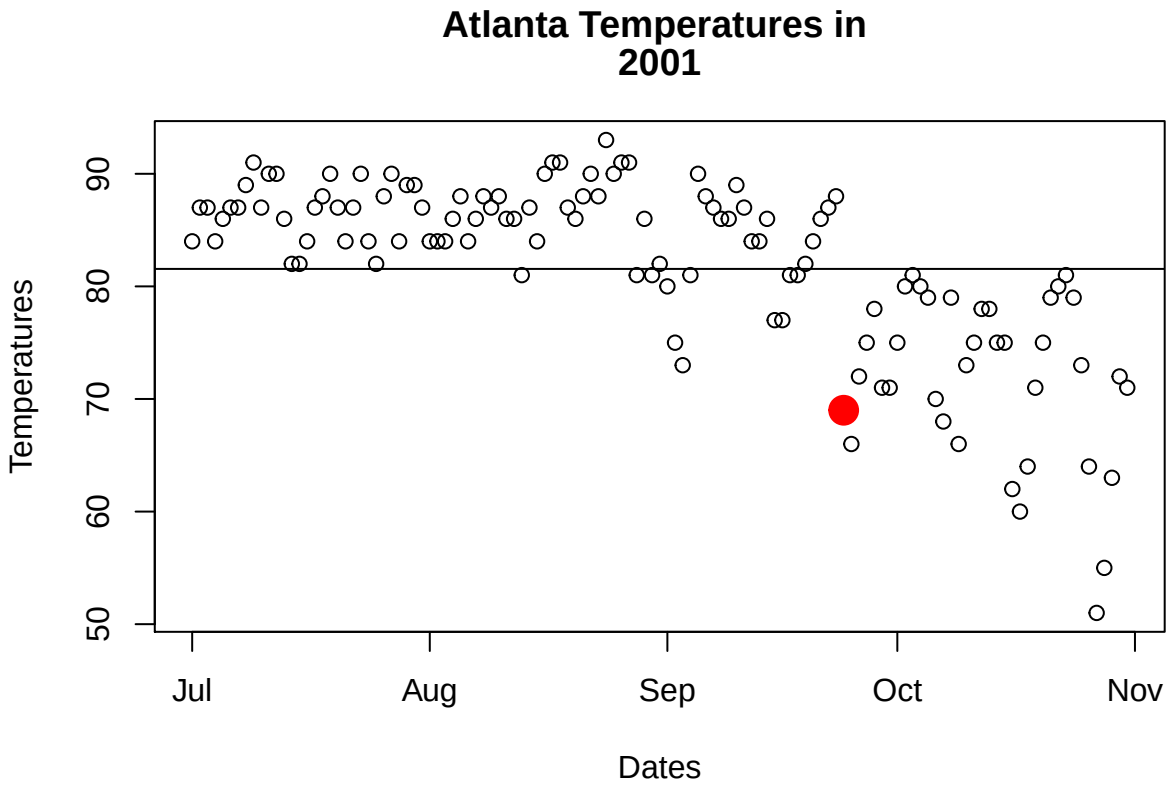

```
## [1] "5-Oct"
```



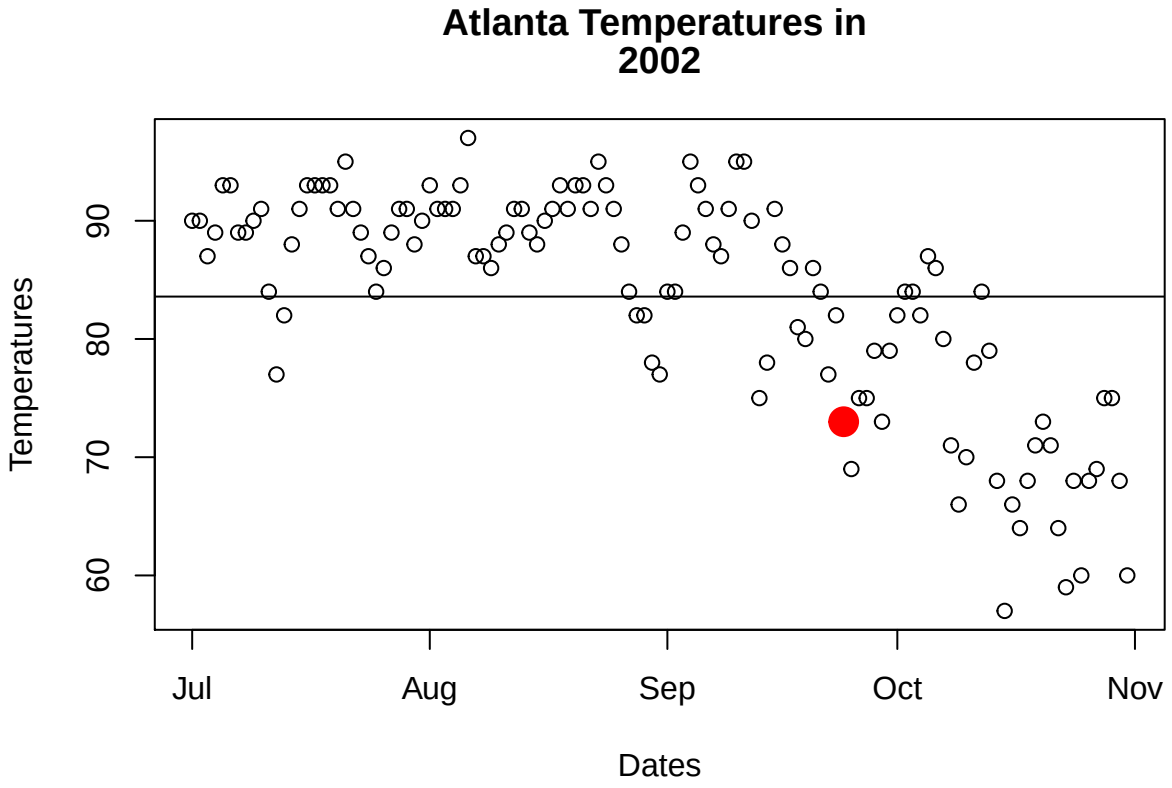

```
## [1] "21-Sep"
```



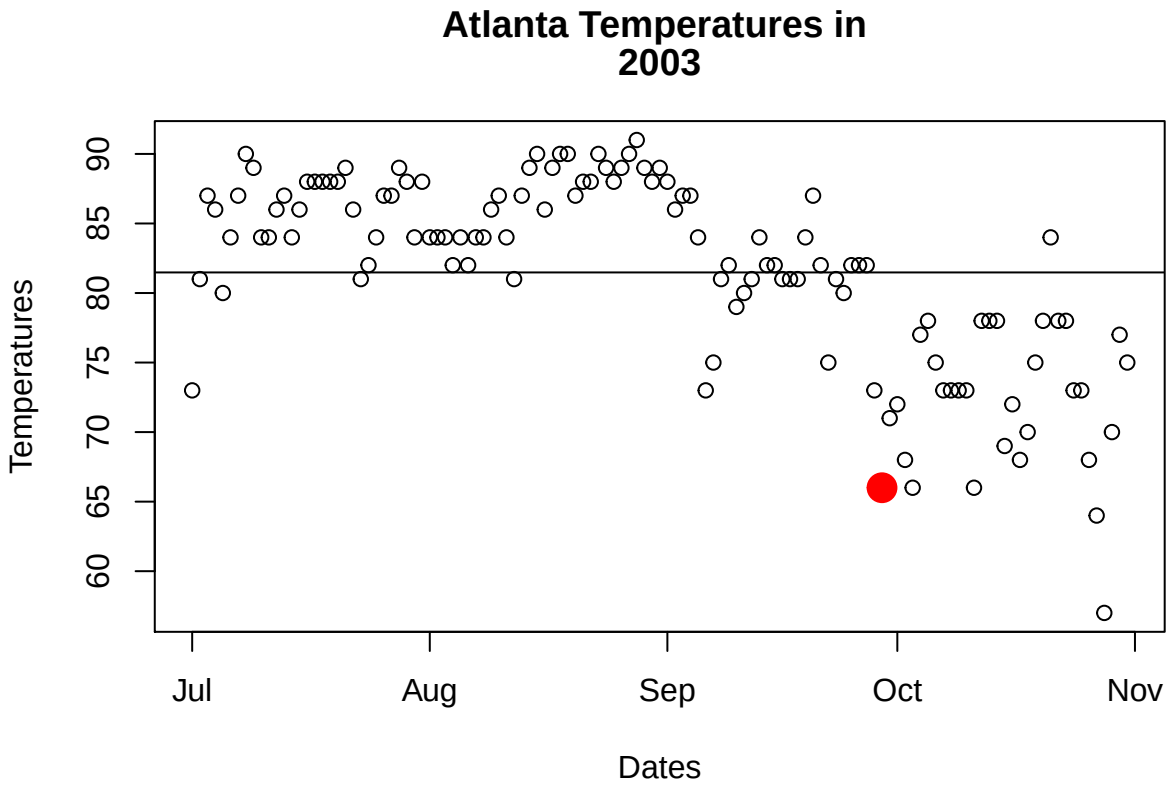

```
## [1] "6-Sep"
```



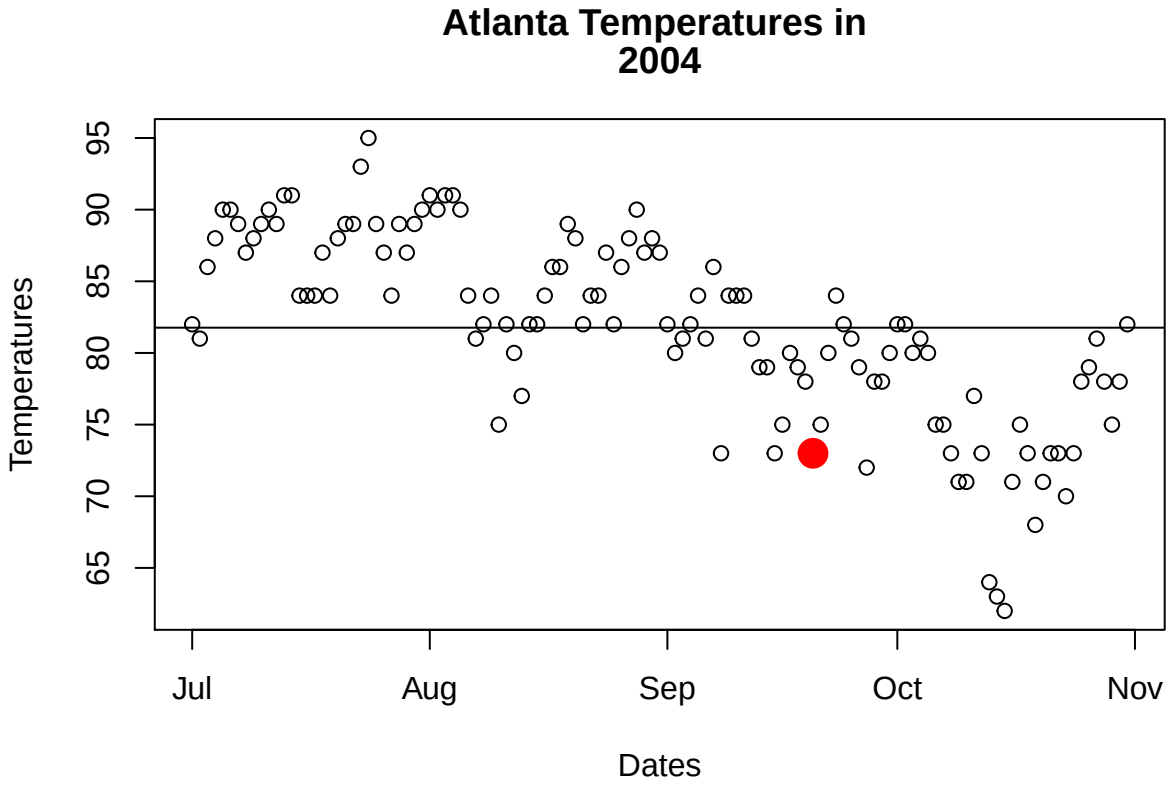

```
## [1] "24-Sep"
```



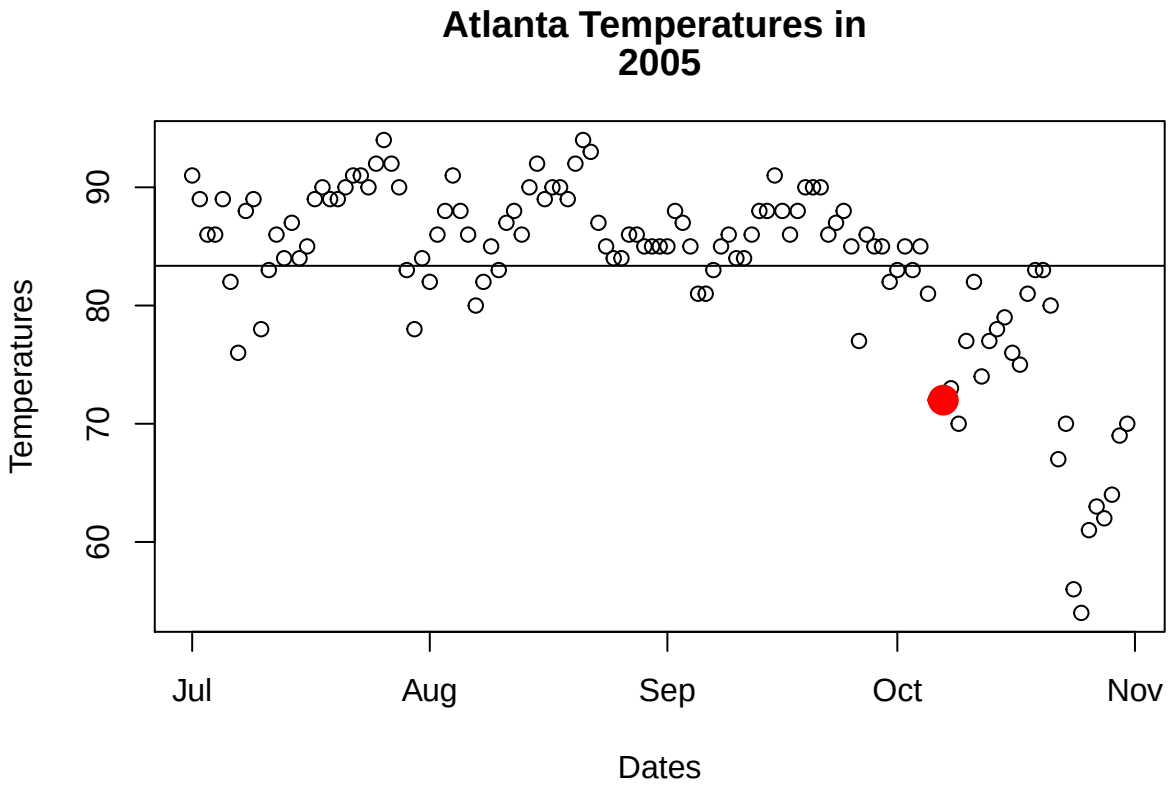

```
## [1] "24-Sep"
```



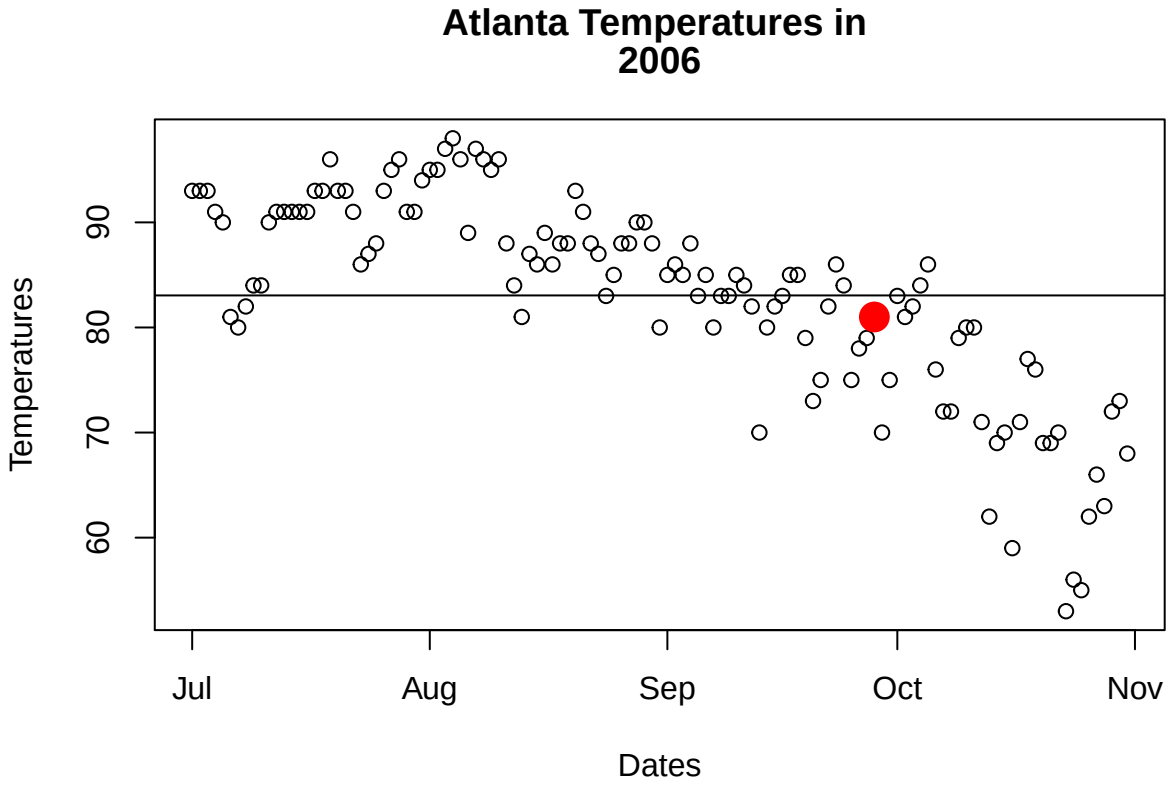

```
## [1] "29-Sep"
```



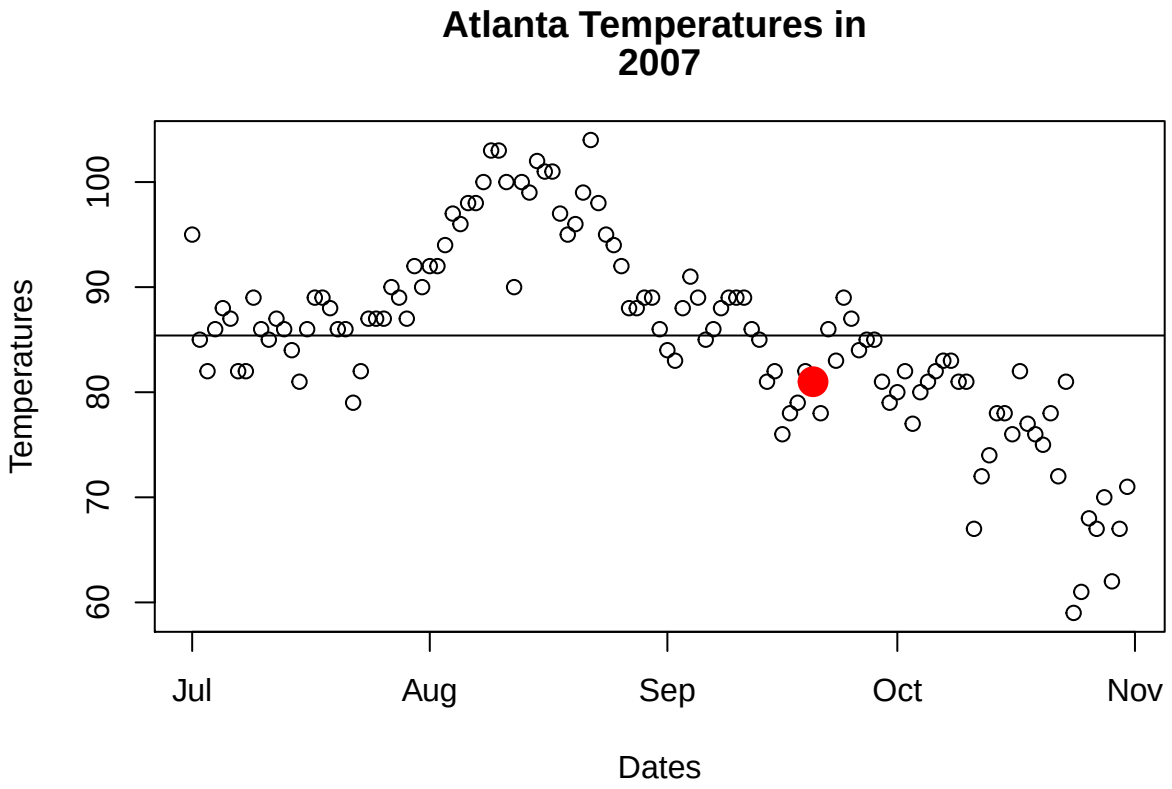

```
## [1] "20-Sep"
```



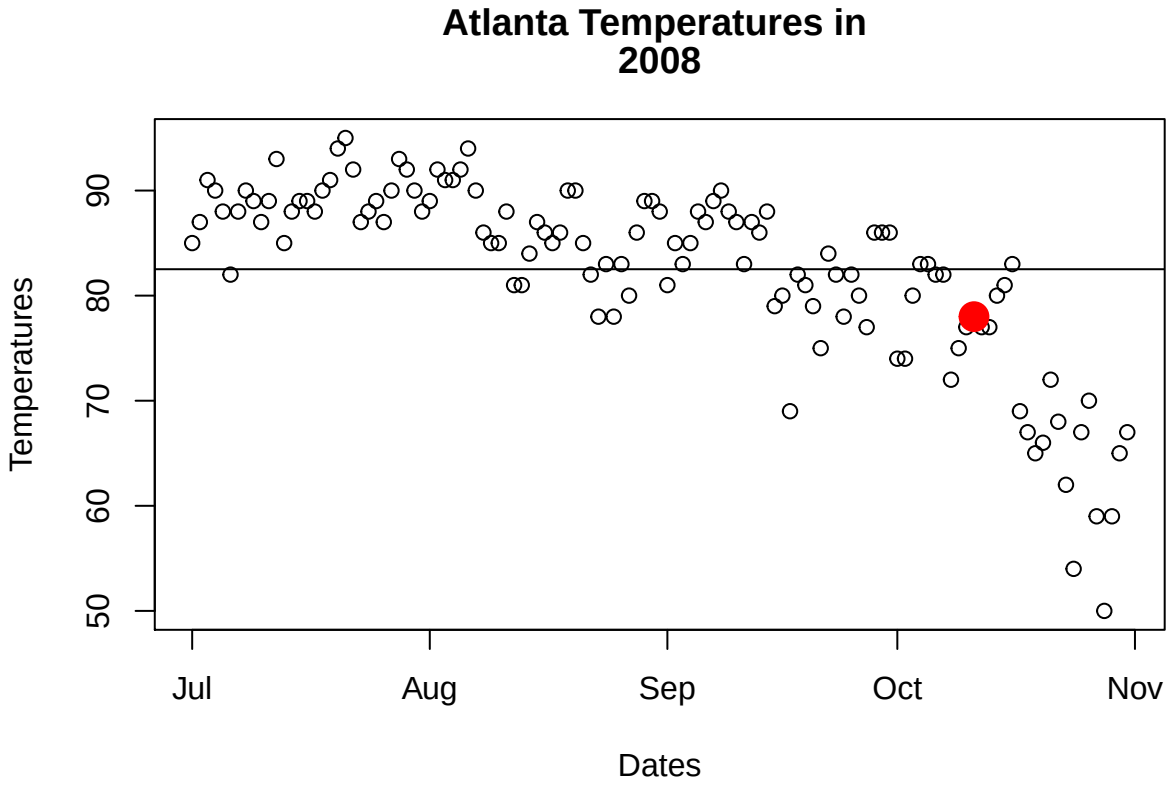

```
## [1] "7-Oct"
```



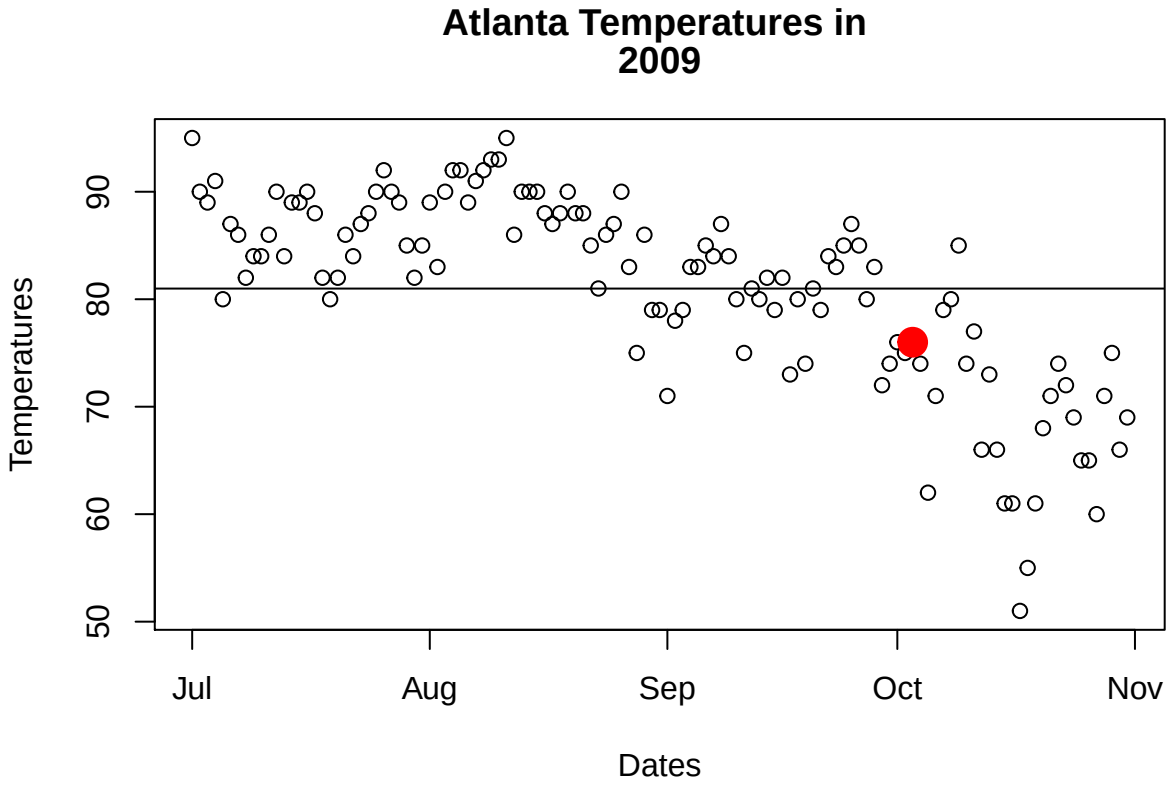

```
## [1] "28-Sep"
```



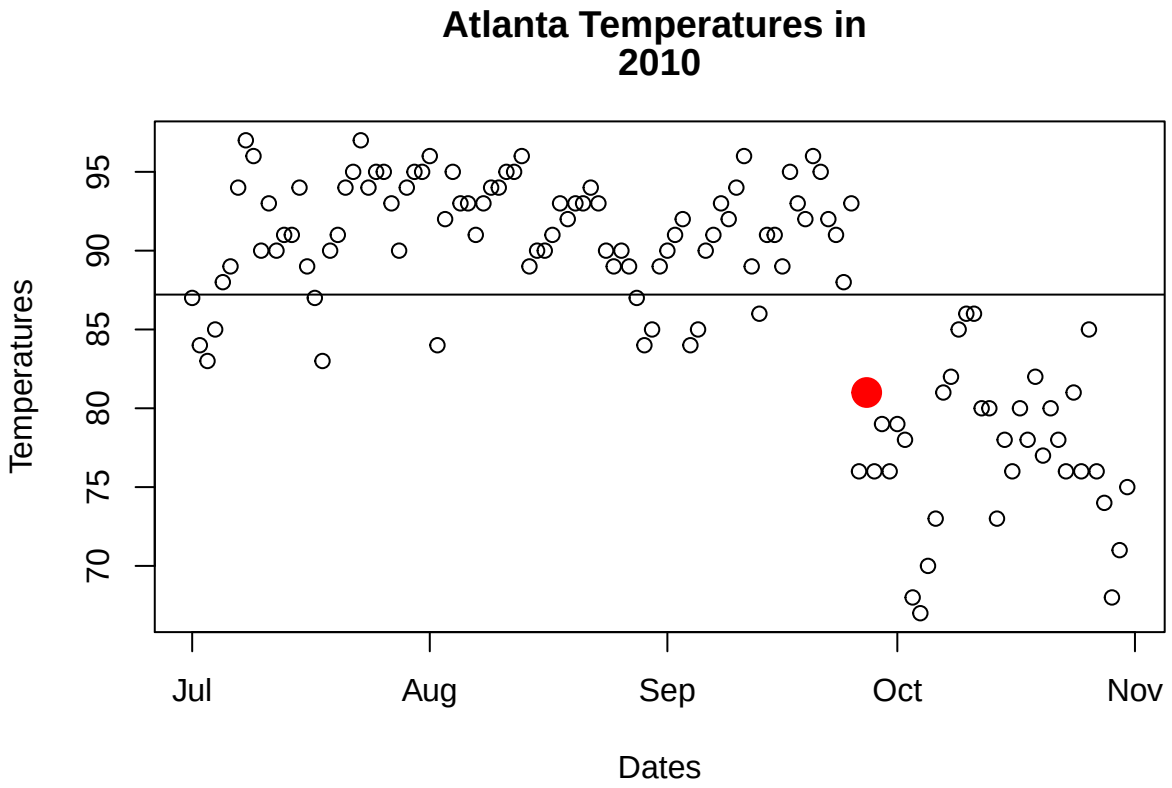

```
## [1] "20-Sep"
```



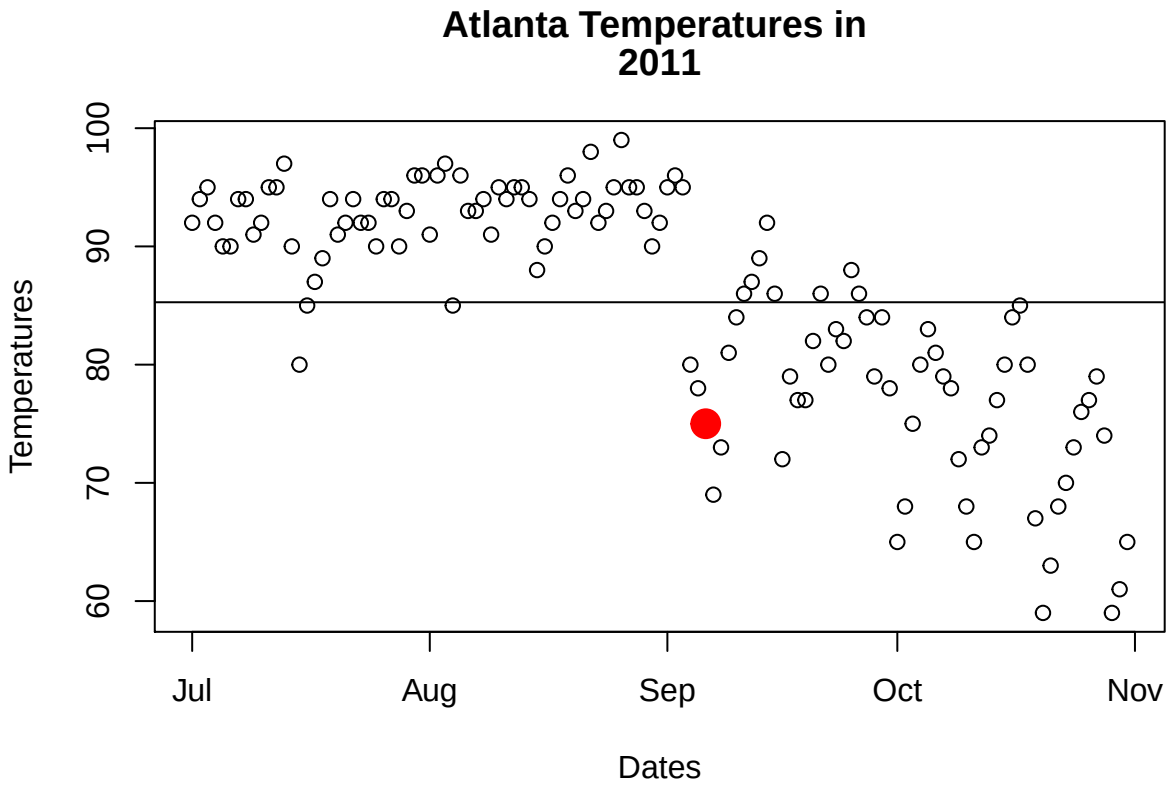

```
## [1] "11-Oct"
```



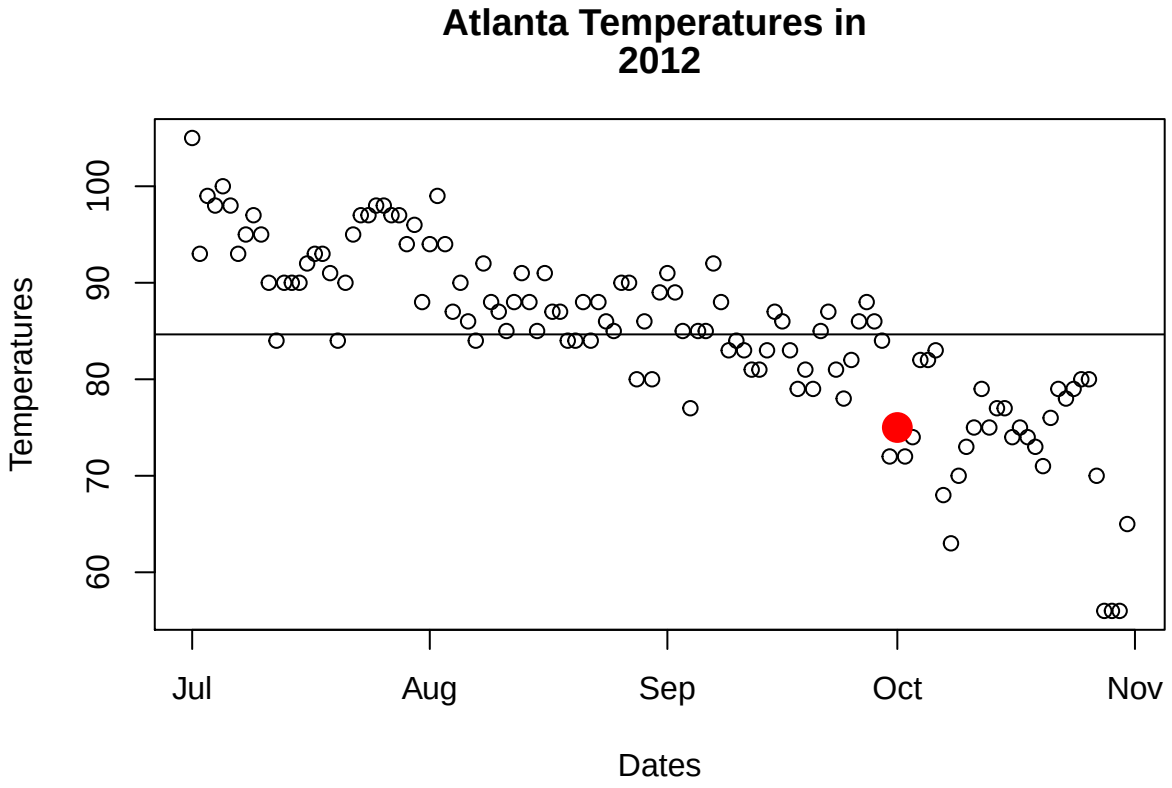

```
## [1] "3-Oct"
```



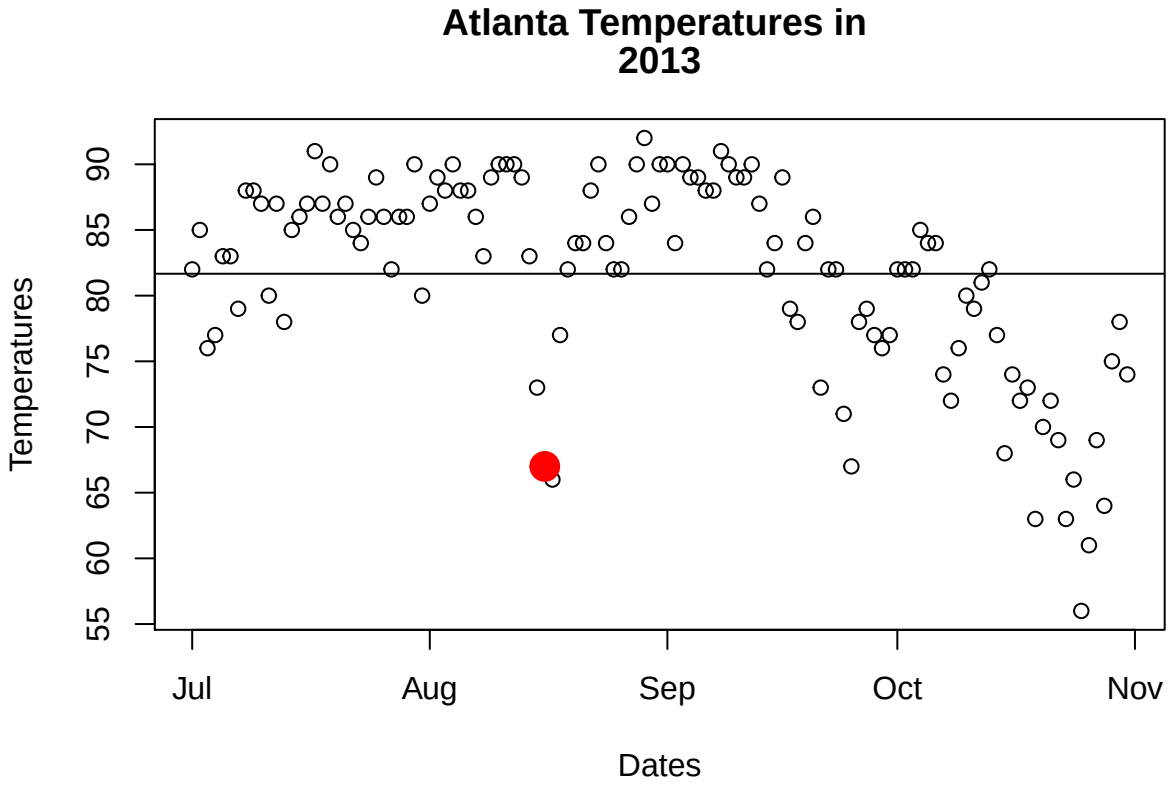

```
## [1] "27-Sep"
```


Atlanta Temperatures in 2011
Temperatures
100
90
80
70
60
Jul
Aug
Sep
Oct
Nov
Dates

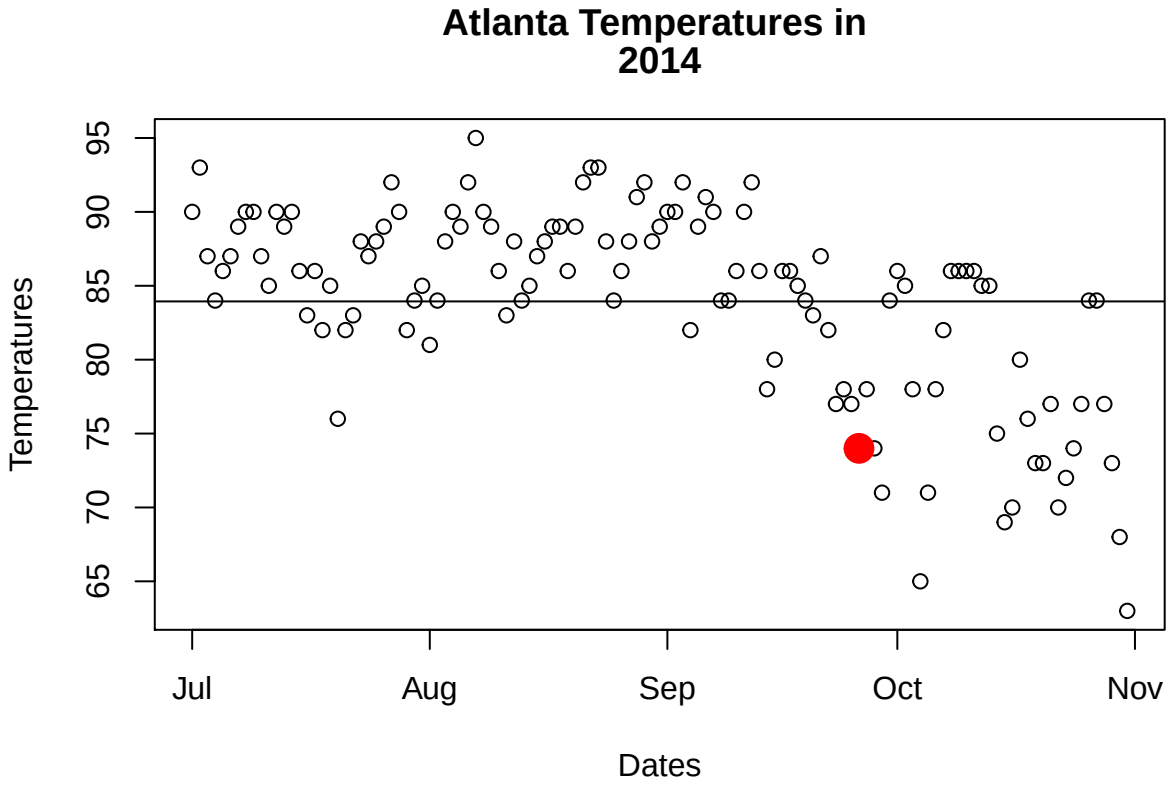
```
## [1] "6-Sep"
```


Atlanta Temperatures in 2012
Temperatures
100
90
80
70
60
Jul
Aug
Sep
Oct
Nov
Dates

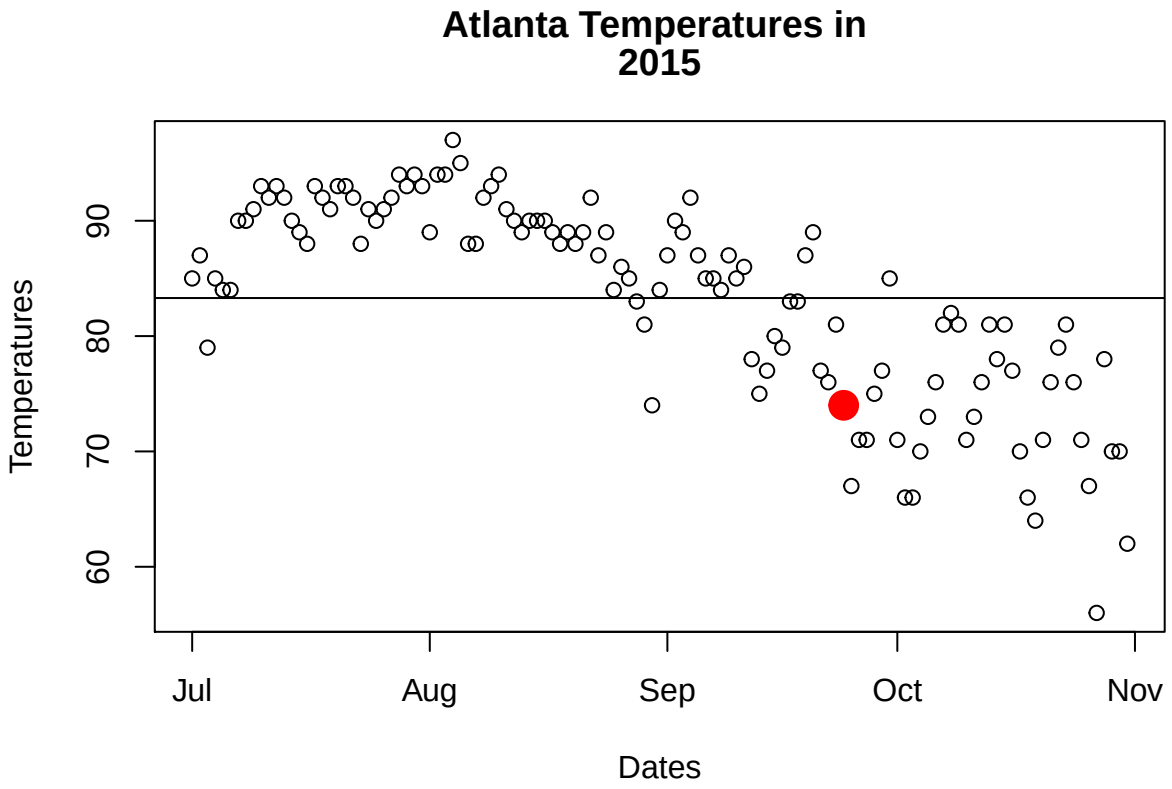
```
## [1] "1-Oct"
```




```
## [1] "16-Aug"
```




```
## [1] "26-Sep"
```




```
## [1] "24-Sep"
```

*__Conclusion__

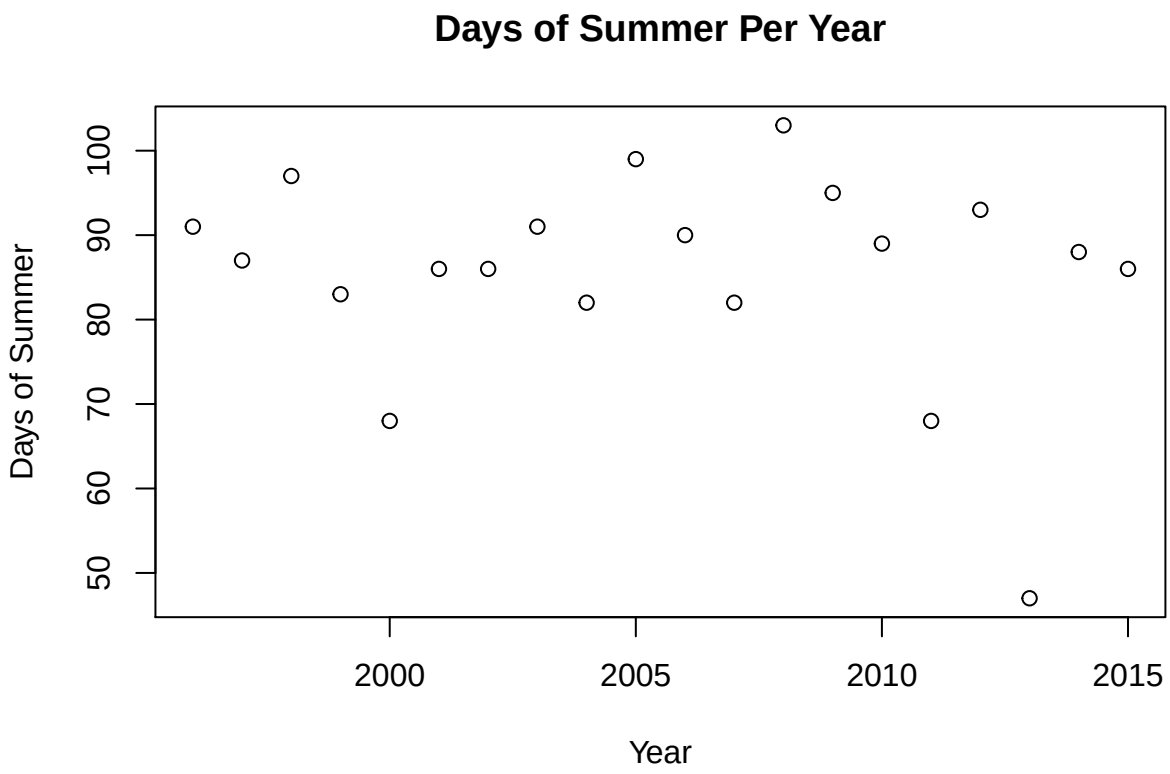
It appears that our algorithm works very well with the exception of 2013. There was an outlier of a week for temperature in August where the temps dropped to nearly 65 degrees before climbing to back over 90 degrees. It was hard to account for this in the cumsum algorithm. As such, the year will just be marked as an outlier for the algorithm.

Question 6.3

Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when)

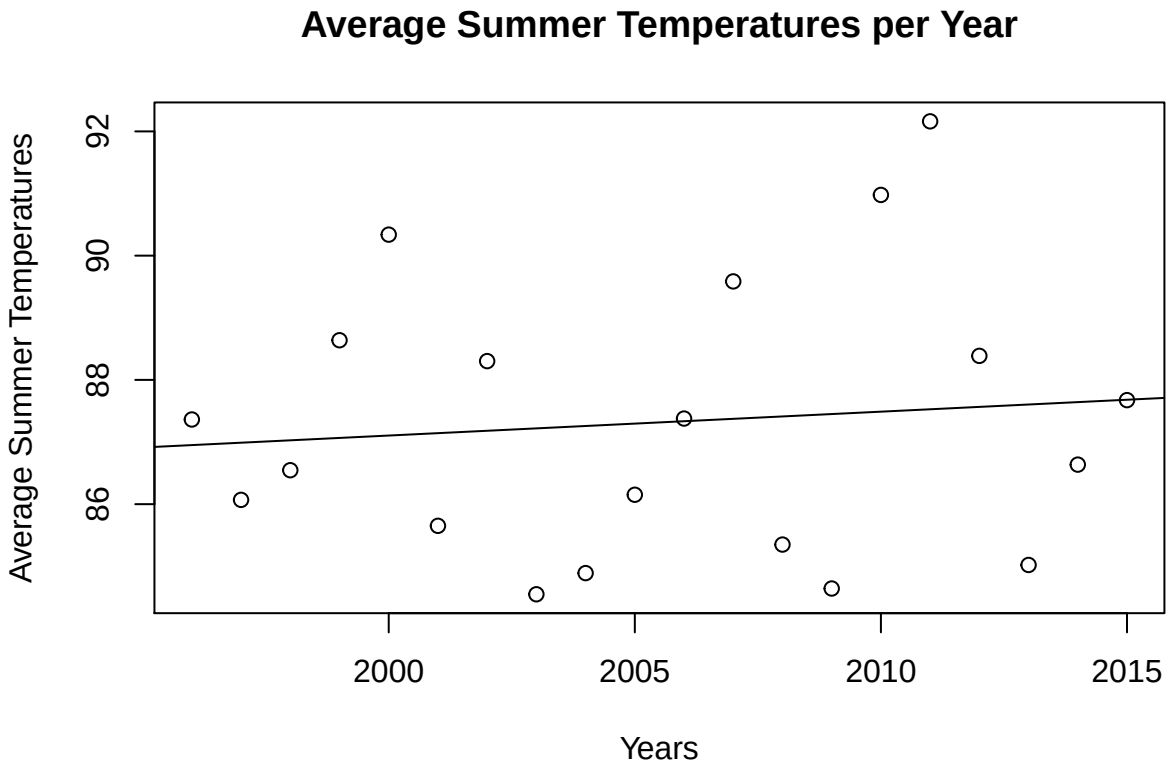
We stored the index locations of the days when summer offically ended during the above CUMSUM alogorithm. This index can also be interpreted as the total number of days of summer. Let's plot this data and see how many days of summer there are for each year

```
plot(x = seq(from = 1996, to = 2015), y = weather_change_index, xlab = 'Year', ylab = 'Days of Summer',
```



We can also use these indices extracted from the CUMSUM algorithm above, to extract only the summer temperatures and then calculate the averages for each year to see if Atlanta's summer climate has gotten warmer.

```
averages = rep(0,20)
for (i in seq(2,21)){
  x = as.matrix(temp[,i])
  x = x[1:weather_change_index[i-1]]
  averages[i-1] = mean(x)

}

years = seq(1996, 2015)
data = c(years, averages)

plot(x = years, y = averages, xlab ='Years', ylab = 'Average Summer Temperatures', main = 'Average Summ
abline(lm(formula = averages~years))
```



*__Conclusion__

After plotting the average temperatues of summer for each year and the data's corresponding regression line, we can see that there is a slight increase in temperature for summer year to year. This is confirmed visually in plotting the data as well as the upward sloping regression line