WK3_HW3

Anon Skywalker

9/7/2020

Quesion 5.1

Using the uscrime data set, I specified the last column involving crimes per 100,000 people and ran it through the grubbs.test function. This function provides an easy way to see outliers within a given column of data. For this data set, I used type = 11 in the function to get both the lowest outlier and the highest outlier from the mean.

The lowest: 342 The highest: 1993

Now that we have identified the outliers, it is important to asses the importance of them as they relate to our data set. By sorting the last column, we can see that the crime rates are all within a decent range. There are no days where the value is so far from the mean that it should be considered an anomally and the same is to be said for the highest crime rate. *For the uscrime data, I would not remove either outlier for this reason.*

```
#Lowest value from the mean: 342
#highest value: 1993
#TYPE 10 = one outlier above the mean
#G = 1.45589, U = 0.95292, p-value < 2.2e-16
#type 11: searches for 2 outliers
#G = 4.26877, U = 0.78103, p-value = 1

grubbs.test(crime.data$Crime, type = 11)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  crime.data$Crime
## G = 4.26877, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

```
sort(crime.data$Crime)
```

```
##  [1]  342  373  439  455  508  511  523  539  542  566  578  653  664  682  696
## [16]  705  742  750  754  791  798  823  826  831  849  849  856  880  923  929
## [31]  946  963  968 1030 1043 1072 1151 1216 1216 1225 1234 1272 1555 1635 1674
## [46] 1969 1993
```

Question 6.1

As a clinical data analyst, a great detection model that we could use within the healthcare setting would be a cusum that looks at our number of tickets in the queue. As the team completes tickets the value will decrease and indicate that workloads are under control. As tickets increase, the need may arise for additional help to be called upon via contract positions or additional fte positions.

A good baseline would be around 40 tickets during the summer season, as that is what we usually gravitate around

Question 6.2 Part 1

For the final set of questions, we utilize temperature data from georgia through the summers of years 1995 to 2015. By using a manual cusum approach in R due to my severe lack of excel skills, we can find the unofficial summer end for each year and then plot those days, as well as other summary information, to find out whether or not georgia is subject to global warming (which we know is true…).

The first portion of my code involves the Thresh and C varibles that can be altered to modify the sensitivity of the cusum model. My setup for cusum involves using the average temperature per year of the data set as a baseline stored as list variable “mu”. This is then used in the cusum equation to find decreases in temperature each year.

More code explanation: using a couple of loops, I was able to make sure that the equation would be used for each year (for loop that increments y) and for each day in that year (for loop that increments i). Should the value of a summer day be negative after running through the equation, that value is set to 0 via the if statement that executes for the first day of summer each year and then again in the next if statement within the while loop.

Data for part 1: Below is a read-out of the day summer ended for each year 1995-2015 and their corresponding temps on that day.

1. 1-Oct 66
2. 27-Sep 64
3. 9-Oct 72
4. 30-Sep 71
5. 18-Sep 73
6. 29-Sep 71
7. 29-Sep 73
8. 2-Oct 68
9. 13-Oct 64
10. 12-Oct 74
11. 13-Oct 62
12. 12-Oct 72
13. 19-Oct 65
14. 6-Oct 71
15. 2-Oct 78
16. 1-Oct 65
17. 7-Oct 68
18. 19-Oct 63
19. 4-Oct 65
20. 27-Sep 71

    These were the values I got after tuning my cusum to better detect temperatures that I would deem as cold enough to warrant summer coming to an end.

```
  Thresh <- 40
  C <- 3
  #stores avg temp per year for use as a baseline (mu)
  mu <- rep(NA, ncol(temp.data) - 1)
  #stores each last day of summer each year
  summer.end <- rep(NA, ncol(temp.data) - 1)
  #stores the last temp of summer each year
  temp.end <- rep(NA, ncol(temp.data) - 1)
  #for calculating and later plotting of the length of summer
  summer.length <- rep(NA, ncol(temp.data) - 1)
  #For temp storage in for/while loop
  S <- rep(NA, nrow(temp.data))

#for each year except the day column!
for (y in 2:ncol(temp.data))
{
  i<- 1
  #avg temp for each year
  mu[y - 1] <- mean(temp.data[,y])

  #for the first day of summer per year
  S[1] <- (mu[y-1] - temp.data[1,y] - C)

    if(S[1] < 0)
    {
      S[1] = 0
    }

  while (S[i] < Thresh)

  {
    i = i + 1
    S[i] <- S[i-1] + (mu[y-1] - temp.data[i,y] - C)

    if(S[i] < 0)
    {
      S[i] = 0
    }

  }
  #temp at last day for current year
  temp.end[y-1] <- temp.data[i,y] #for q1

  #stores the last day of summer for current year
  summer.end[y-1] <- toString(temp.data$DAY[i])
  summer.length[y-1] <- i

  compare.df <- data.frame(summer.end, temp.end)
}
 summer.end
```

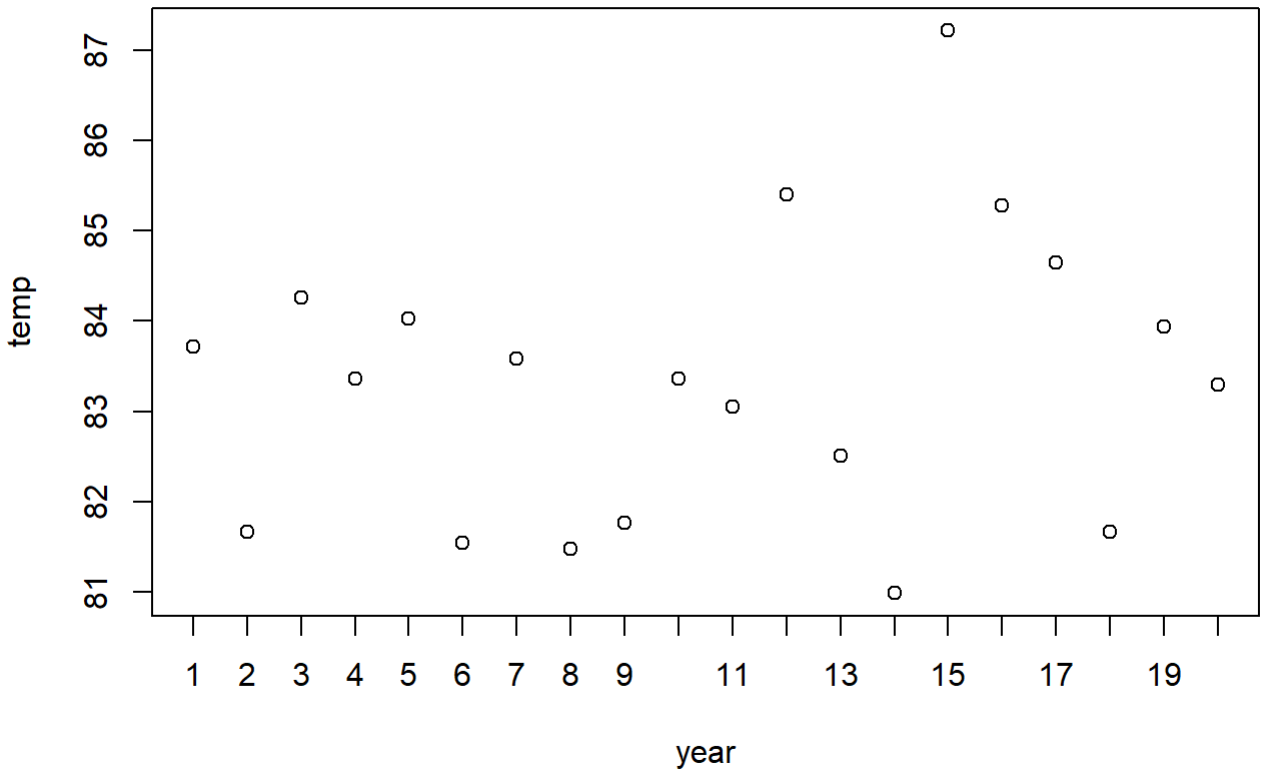
```
##  [1] "1-Oct"  "27-Sep" "9-Oct"  "30-Sep" "18-Sep" "29-Sep" "29-Sep" "2-Oct"
##  [9] "13-Oct" "12-Oct" "13-Oct" "12-Oct" "19-Oct" "6-Oct"  "2-Oct"  "1-Oct"
## [17] "7-Oct"  "19-Oct" "4-Oct"  "27-Sep"
```

Question 6.2 Part 2

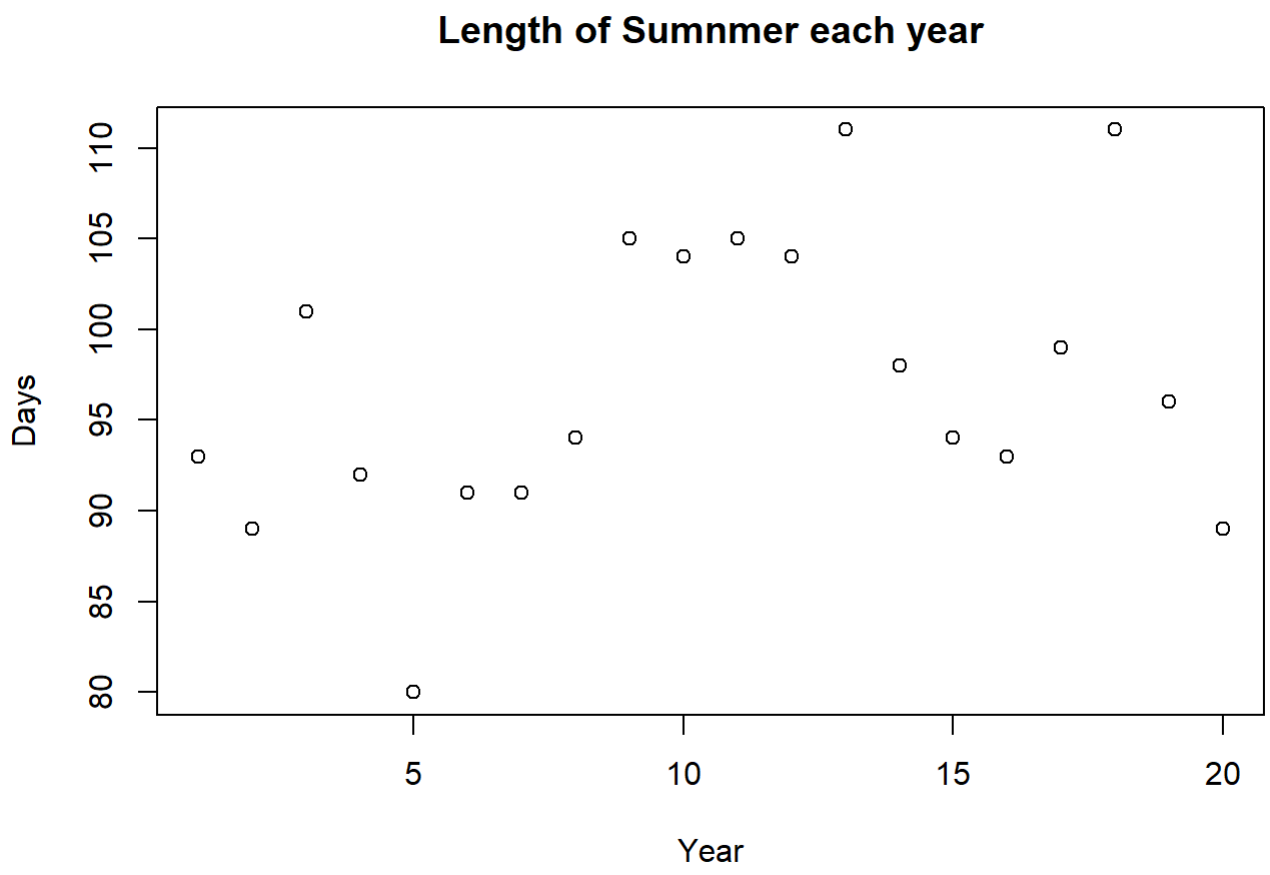
For the second part of the question, we are asked to judge whether or not atlanta is warming up based on the data we got from the cusum model in the previous question. At first, the summer.end data that displayed each day that summer ended on for each year has some evidence of summer ending at a later date each year but becomes diluted quickly towards the end of the collected data.

As another attempt to find out an answer to this question, I gathering the length of each summer in another list defined as summer.length in my code above and plotted the info this provided. By looking at the summer length plot, we can still see the same problem we had before with the day summer ended on. I believe that in order to answer this question with my current cusum model, we would need more years worth of temp data to have a conlusive answer that is easily plotted. However, I would say that my model does show a slight correlation when plotting the length of summers out.

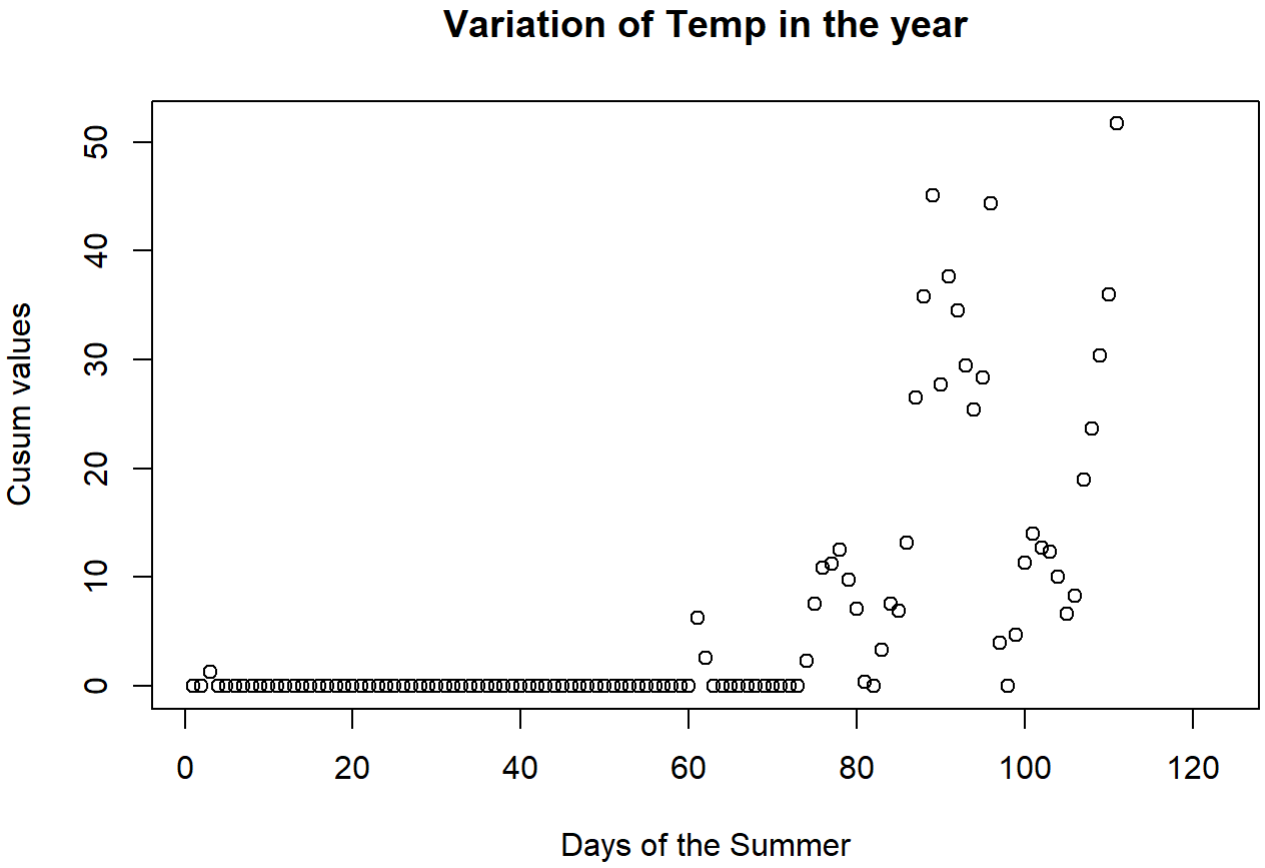
```
x.seq <- c(1:20)
x.lab <- c(1995:2015)
plot(mu, xlab = "year", ylab = "temp", xaxt = "n")
axis(side = 1, at = x.seq)
```



```
plot(summer.length, main = "Length of Sumnmer each year", xlab = "Year", ylab = 'Days')
```




```
plot(S, main = "Variation of Temp in the year" ,xlab = "Days of the Summer", ylab = "Cusum value
s")
```


Variation of Temp in the year
Cusum values
0
10
20
30
40
50
0
20
40
60
80
100
120
Days of the Summer