week2 homework

Chen Yi-Ju(Ernie)

2020-05-28

Question 4.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

A clustering model can be seen in the case of industry clusters, which means that certain kind of factories/businesses may be more likely to choose to locate at a certain country or city. For example, toy factories tend to be located in places with cheaper wages. Possible factors should include: labor wages, land prices , education level , tax levels

Question 4.2

Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

The iris data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower.

Answer :I ended up with the conclusion of k = 3 has the highest accuracy of 95%.

Process

Loading data and libraries

```
#Loading data and library
setwd("D:/ernie/self-study/GTxMicroMasters/Introduction to Analytics Modeling/week2/homework")
iris <- read.table("iris.txt")
library(tidyverse)
library(factoextra)
library(cluster)
```

First of all, a quick overview of the data

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

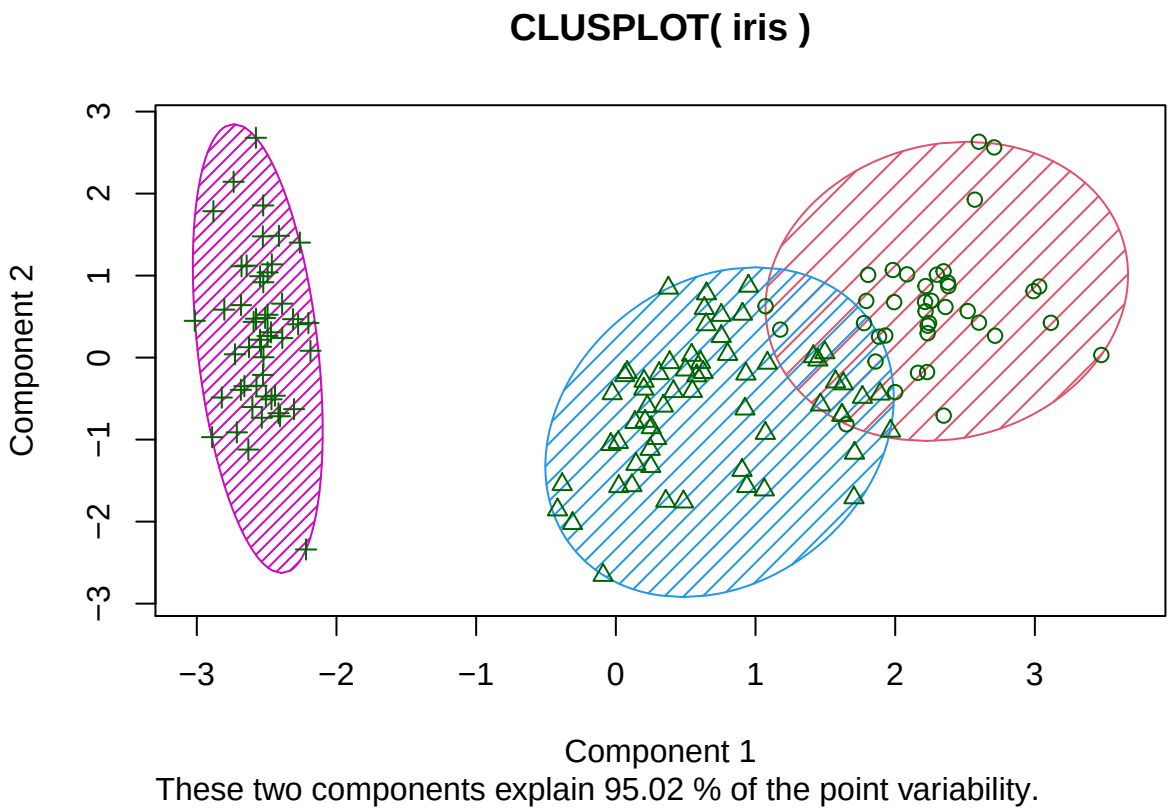
Trying out k at 3 for a model

```
set.seed(101)
irisCluster <- kmeans(iris[,1:4], 3, nstart = 20)
comparison <-table(irisCluster$cluster, iris$Species)
comparison
```

```
##
##     setosa versicolor virginica
##   1      0          2        36
##   2      0         48        14
##   3     50          0         0
```

We can see that the accuracy is quite high, with only a slight mix between versicolor and virginica.

```
clusplot(iris, irisCluster$cluster , color = T, shade = T, labels = 0,lines = 0)
```



But to make sure of which k to use, we try out the elbow method:

```
test_mat<- data.frame(matrix(nrow = 15, ncol = 2,0))
for (k in 1:15){
  irisCluster_test <- kmeans(iris[,1:4], k, nstart = 20)
  j <- irisCluster_test$tot.withinss
  test_mat[k,1] <- k
  test_mat[k,2] <- j
}
colnames(test_mat) <- c("k" , "Total")# Total within-cluster sum of squares
```

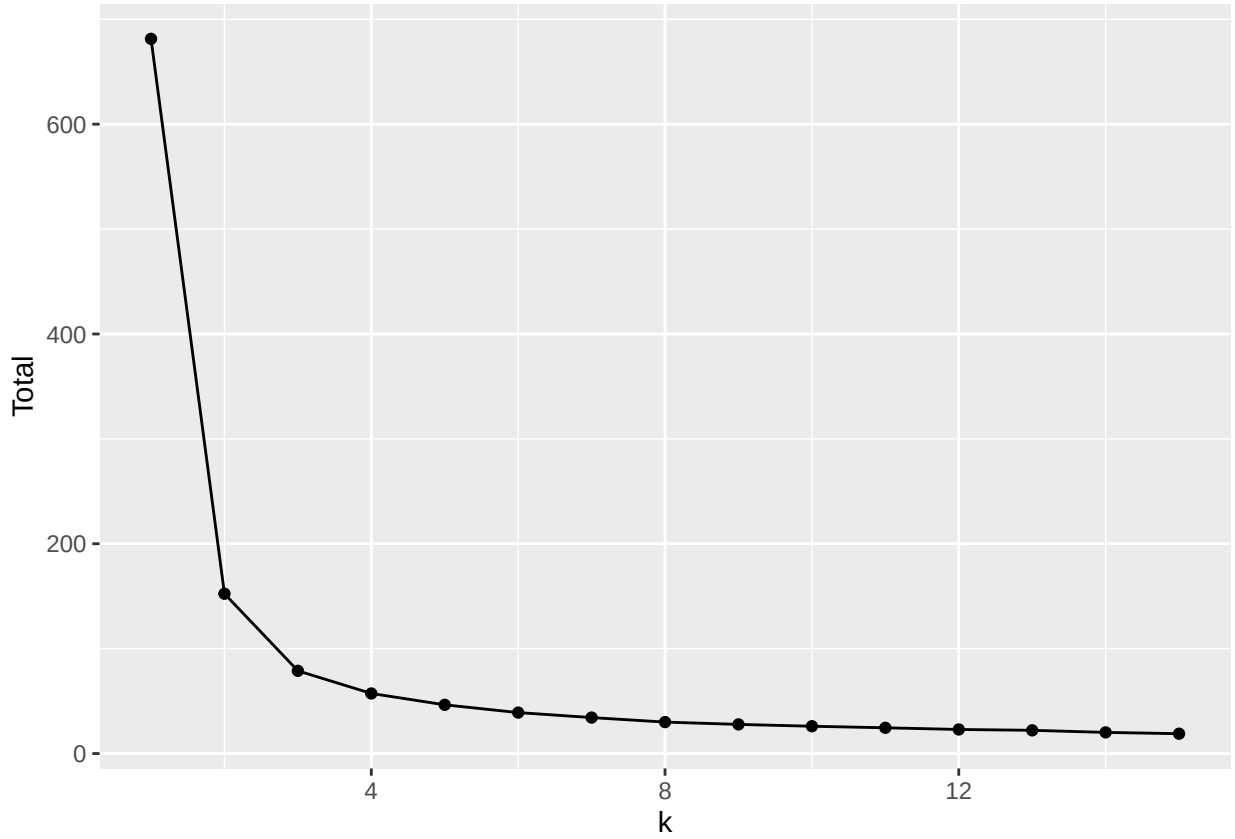
The results, as in a matrix

```
test_mat
```

```
##     k     Total
## 1   1 681.37060
## 2   2 152.34795
## 3   3  78.85144
## 4   4  57.26562
## 5   5  46.44618
## 6   6  39.03999
## 7   7  34.29823
## 8   8  29.98894
## 9   9  27.78609
## 10 10  26.03014
## 11 11  24.54635
## 12 12  22.94478
## 13 13  22.12870
## 14 14  20.16033
## 15 15  18.92890
```

Graphically represented: We can see that k = 3 is indeed an appropriate choice, due to the great decrease in error and the steadiness onwards.

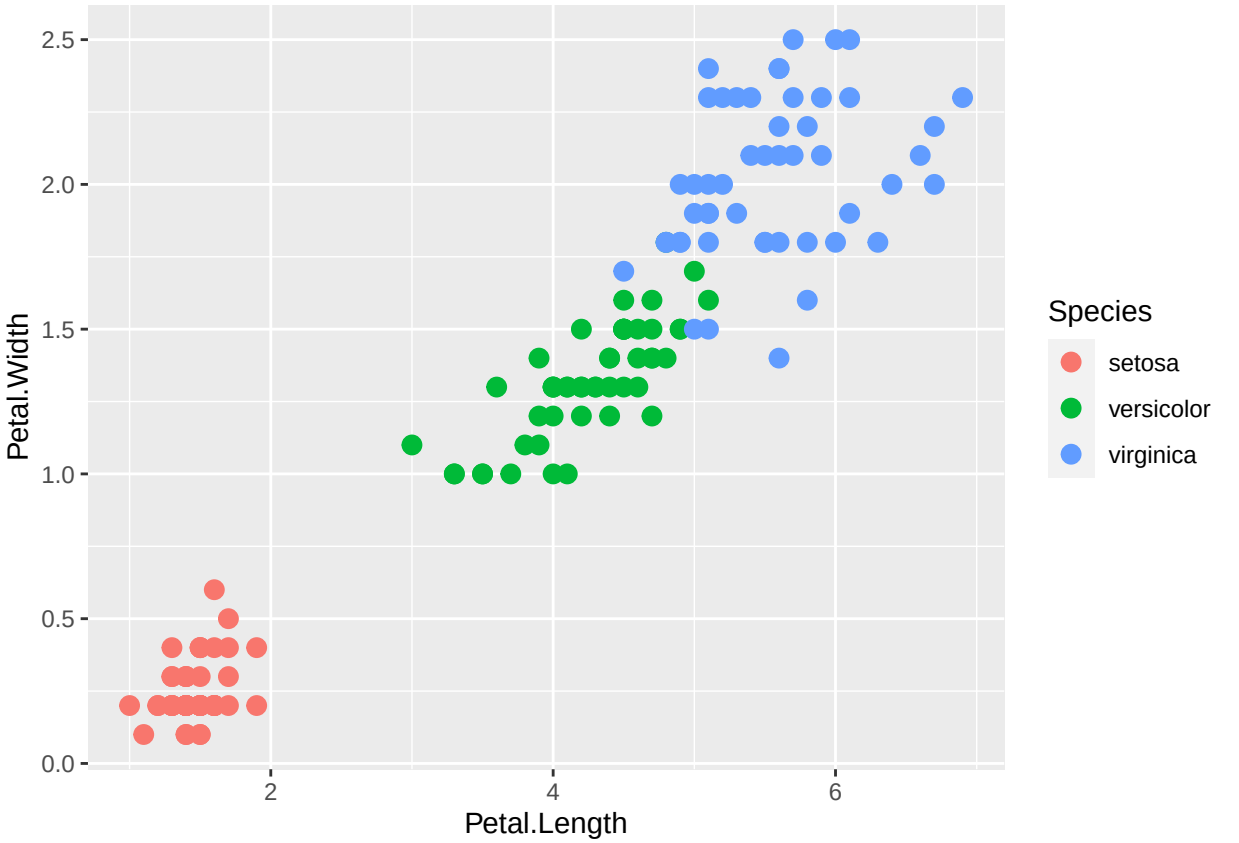
```
ggplot(test_mat , aes(x = k , y = Total ))+
  geom_point()+
  geom_line()
```



Lastly, in this case k = 3 is quite obvious to begin with since we already know how many species are there

for us to categorize, we can see it even more clearly from this plot.

```
#Visualization

pl <- ggplot(iris, aes(Petal.Length ,Petal.Width, color = Species))+
  geom_point(size = 3)
print(pl)
```



In coclusion, k = 3 is the most accurate model with 95% accuracy

Question 5.1

##Using crime data from the file uscrime.txt ,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. ### Answer: Yes and No. The reason for saying that is that it depends on how you define outliers as to how far it is. It is a "weak outlier" , if one defines p-value to be less than 0.1, but not so if it is put under the stricter definition of p-value = 0.05. According to the plots and tests, there are outliers far enough well out of 1.5 SD of the data.However, even with the largest of the outliers, they are not far enough (to large of a p-value:0.07) to be strictly statistically

Process

Loading data

```
library(outliers)
crime <- read.table("uscrime.txt" , header = T)
crimePerHundred <-crime$Crime
```

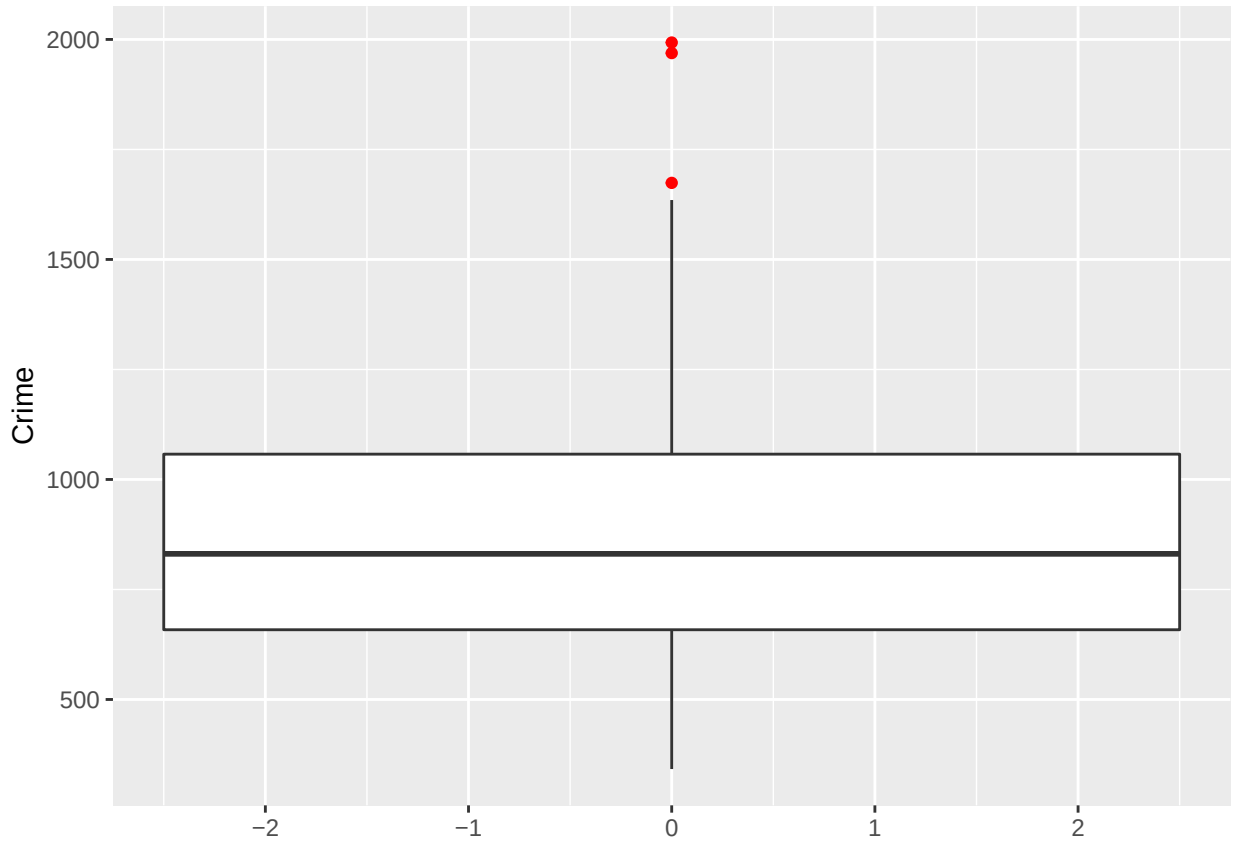
Quick Overview

```
summary(crimePerHundred)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   342.0   658.5   831.0   905.1  1057.5  1993.0
```

Box-plot show that there are some outliers, outside 1.5SDs of the data.

```
ggplot(crime, aes(y = Crime)) +
  geom_boxplot(outlier.color = "red" ,width = 5)
```



Testing using the grubbs test: It gives us the furthest outlier. But the p-value is slightly too large.

```
grubbs.test(crimePerHundred)
```

```
##
##  Grubbs test for one outlier
##
## data:  crimePerHundred
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

Conclusion: There are ouliers under loosely defined conditions.

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?

In my country (Taiwan), we have big downpours of rain during monsoon season and typhoon season with the extreme of more than 1000centimeters of rain per 24 hours. Therefore, the water level of rivers and stream would be my choice of applying a Change Detection. In terms of the C value and threshold. Because in case of an breach of embankment, floods may cause extremely big damage to both property and lives. Therefore, an optimal choice would be a small C and an relatively high threshold as you can never be too careful.

Question 6.2

1. 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.

Answer: October-2 is my answer.

Process

Loading Data

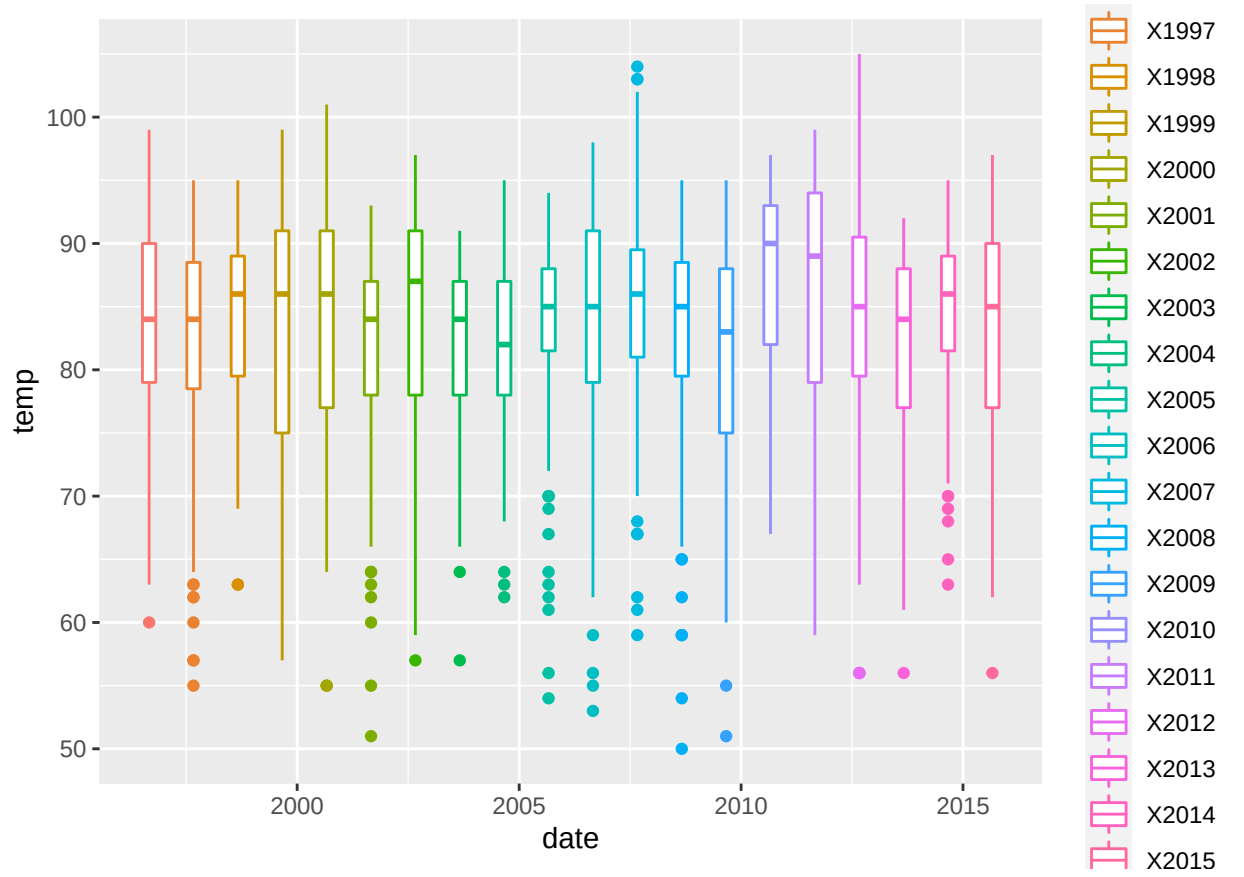
```
library(qcc)
library(dplyr)
library(tidyr)
library(lubridate)
#loading and tidyiing data
weather <- data.frame(read.table("temps.txt" , header = T))%>%
  gather(year,temp,-DAY)%>%
  mutate(year = as.factor(year),
         date = paste(DAY,year,sep = "-"))%>%
  mutate(date = dmy(date),
         month = month(date),
         day = day(date))
summary(weather)
```

```
##       DAY                 year           temp             date
##  Length:2460        X1996  : 123   Min.   : 50.00   Min.   :1996-07-01
##  Class :character   X1997  : 123   1st Qu.: 79.00   1st Qu.:2001-05-01
##  Mode  :character   X1998  : 123   Median : 85.00   Median :2006-03-01
##                     X1999  : 123   Mean   : 83.34   Mean   :2006-03-01
##                     X2000  : 123   3rd Qu.: 90.00   3rd Qu.:2010-12-30
##                     X2001  : 123   Max.   :105.00   Max.   :2015-10-31
##                     (Other):1722
##      month             day
##  Min.   : 7.000   Min.   : 1.00
##  1st Qu.: 7.000   1st Qu.: 8.00
##  Median : 8.000   Median :16.00
##  Mean   : 8.496   Mean   :15.88
##  3rd Qu.:10.000   3rd Qu.:24.00
##  Max.   :10.000   Max.   :31.00
##
```

EDA: box plots of temperture by year

```
#average temperture per year
avg_by_year <-weather %>%
  group_by(year)%>%
  summarise(avg_year = mean(temp))
 pl2 <-ggplot(weather, aes(x = date, y = temp , color = year))+
   geom_boxplot()
```

```
pl2
```



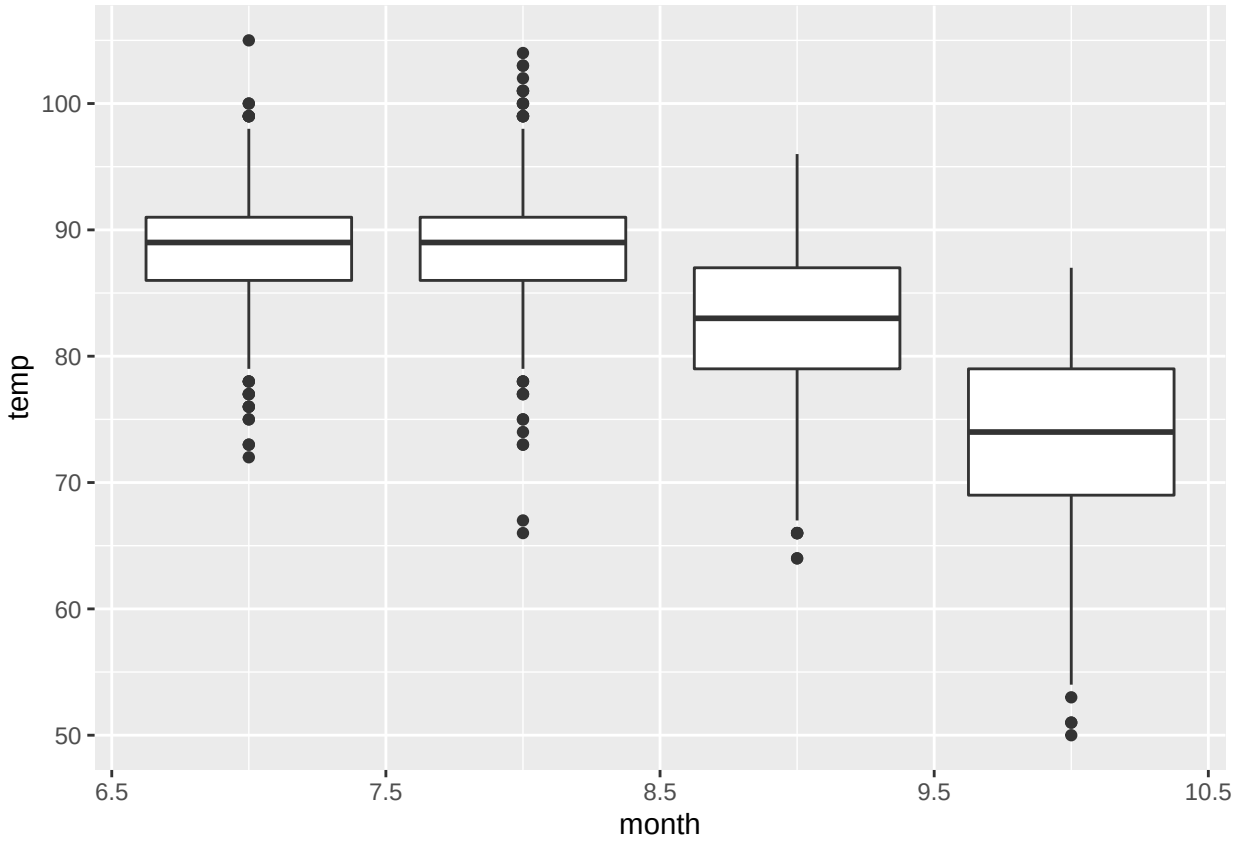
EDA2: boxplot of temperature by month

```
#average temperture per month
avg_by_month <-weather %>%
  group_by(month)%>%
  summarise(avg_month = mean(temp))
avg_by_month
```

```
## # A tibble: 4 x 2
##   month avg_month
##   <dbl>     <dbl>
## 1     7      88.8
## 2     8      88.6
## 3     9      82.7
## 4    10      73.3
```

```
pl3 <-ggplot(weather, aes(x = month, y = temp,group = month))+
  geom_boxplot()
```

```
pl3
```



The Cusum Process

```
#cusum
weather_2 <- data.frame(read.table("temps.txt" , header = T))
day_mean <- rowMeans(weather_2[,-1])
summary(day_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.60   77.78   85.90   83.34   88.78   91.15
```

Setting C as 4, which is about half way between the 1st Quarter and the median

```
# mean - Xi (want to find decrease)
total_mean<- mean(day_mean)
diff <- total_mean - day_mean %>%
  data.frame()
#setting C
C <- 4 #half way between median and 1st quarter
diff <- diff - C
diff<- mutate(diff, date = weather_2$DAY)
colnames(diff)<- c("Xi-m-C" , "date")
```

Deriving the St

```
S <- matrix(nrow = 124,0)

for (i in 1:123){
```

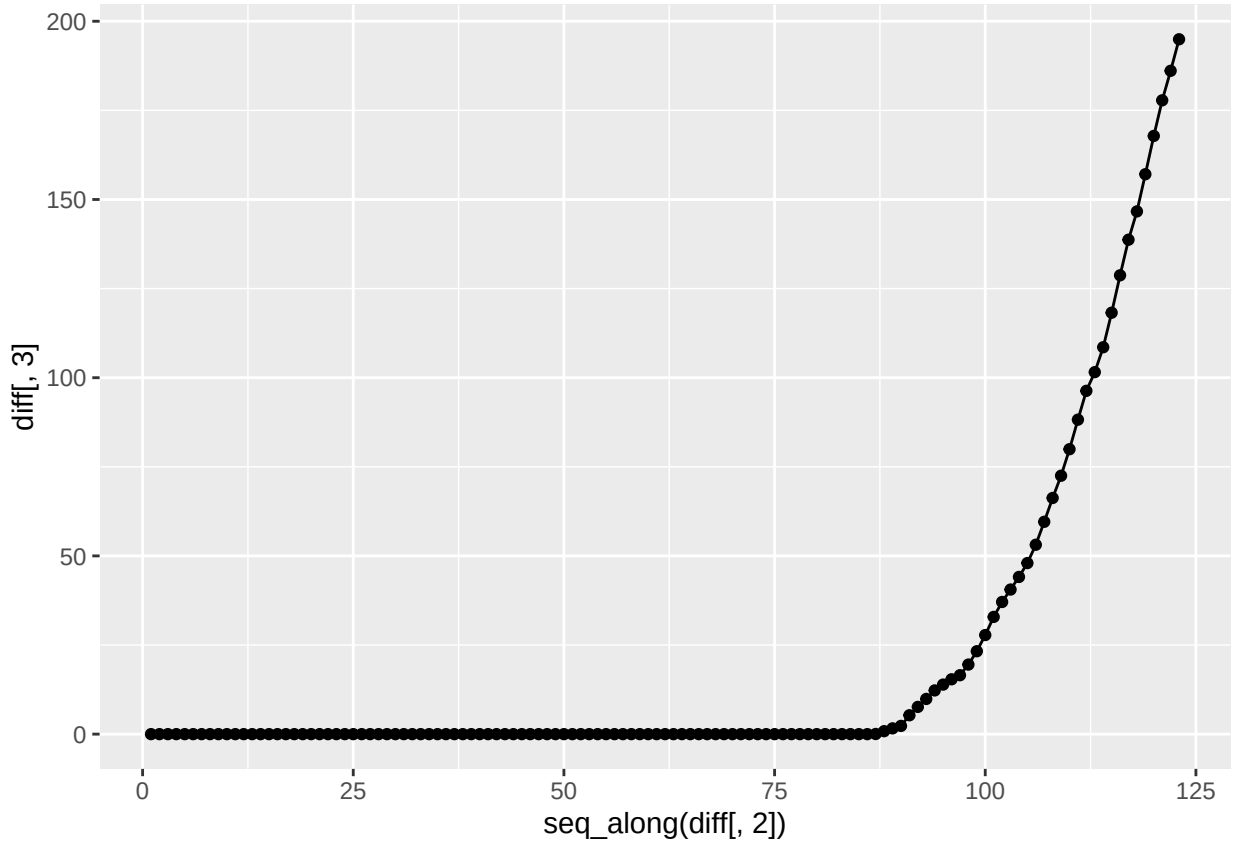
```
  if (i == 1){
    S[i+1,1] <- max(0,S[1,1]+diff[i,1])
  }else{
    S[i+1,1] <- max(0,S[i,1]+diff[i,1])
  }

}
S<- data.frame(S)
diff <- diff%>%
  mutate(S = S[2:124,1])
```

```
head(diff)
```

```
##      Xi-m-C  date S
## 1 -9.510976 1-Jul 0
## 2 -9.010976 2-Jul 0
## 3 -9.060976 3-Jul 0
## 4 -9.010976 4-Jul 0
## 5 -8.910976 5-Jul 0
## 6 -8.510976 6-Jul 0
```

Graphically:

```
pl4 <- ggplot(diff, aes(x= seq_along(diff[,2]), y = diff[,3] ))+
        geom_point()+
        geom_line()
pl4
```



I set the threshold at 10:

```
#threshold set at 10
S2 <-subset(diff,S>10)
S2
```

```
##        Xi-m-C   date         S
## 94   2.389024  2-Oct  12.26220
## 95   1.639024  3-Oct  13.90122
## 96   1.489024  4-Oct  15.39024
## 97   1.139024  5-Oct  16.52927
## 98   2.989024  6-Oct  19.51829
## 99   3.739024  7-Oct  23.25732
## 100  4.539024  8-Oct  27.79634
## 101  5.089024  9-Oct  32.88537
## 102  4.189024 10-Oct  37.07439
## 103  3.489024 11-Oct  40.56341
## 104  3.539024 12-Oct  44.10244
## 105  3.889024 13-Oct  47.99146
## 106  5.139024 14-Oct  53.13049
## 107  6.439024 15-Oct  59.56951
## 108  6.689024 16-Oct  66.25854
## 109  6.239024 17-Oct  72.49756
## 110  7.439024 18-Oct  79.93659
## 111  8.289024 19-Oct  88.22561
## 112  8.089024 20-Oct  96.31463
## 113  5.239024 21-Oct 101.55366
## 114  6.989024 22-Oct 108.54268
## 115  9.689024 23-Oct 118.23171
## 116 10.489024 24-Oct 128.72073
## 117  9.989024 25-Oct 138.70976
## 118  7.939024 26-Oct 146.64878
## 119 10.439024 27-Oct 157.08780
## 120 10.739024 28-Oct 167.82683
## 121  9.989024 29-Oct 177.81585
## 122  8.289024 30-Oct 186.10488
## 123  8.839024 31-Oct 194.94390
```

Conclusion : Oct-2.

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

Answer : Yes, from 2010.

Process:

Loading Data + setting C and threshold

```
ann_data<-weather%>%
        group_by(year)%>%
         summarise(year_avg = mean(temp))
year_mean<- mean(ann_data$year_avg)
sd(ann_data$year_avg)
```

```
## [1] 1.582457
```

```
#setting C
C2 <- 1
thresh2 <- 2#total of C and threshold would be approx. 2 Sd. away from mean
```

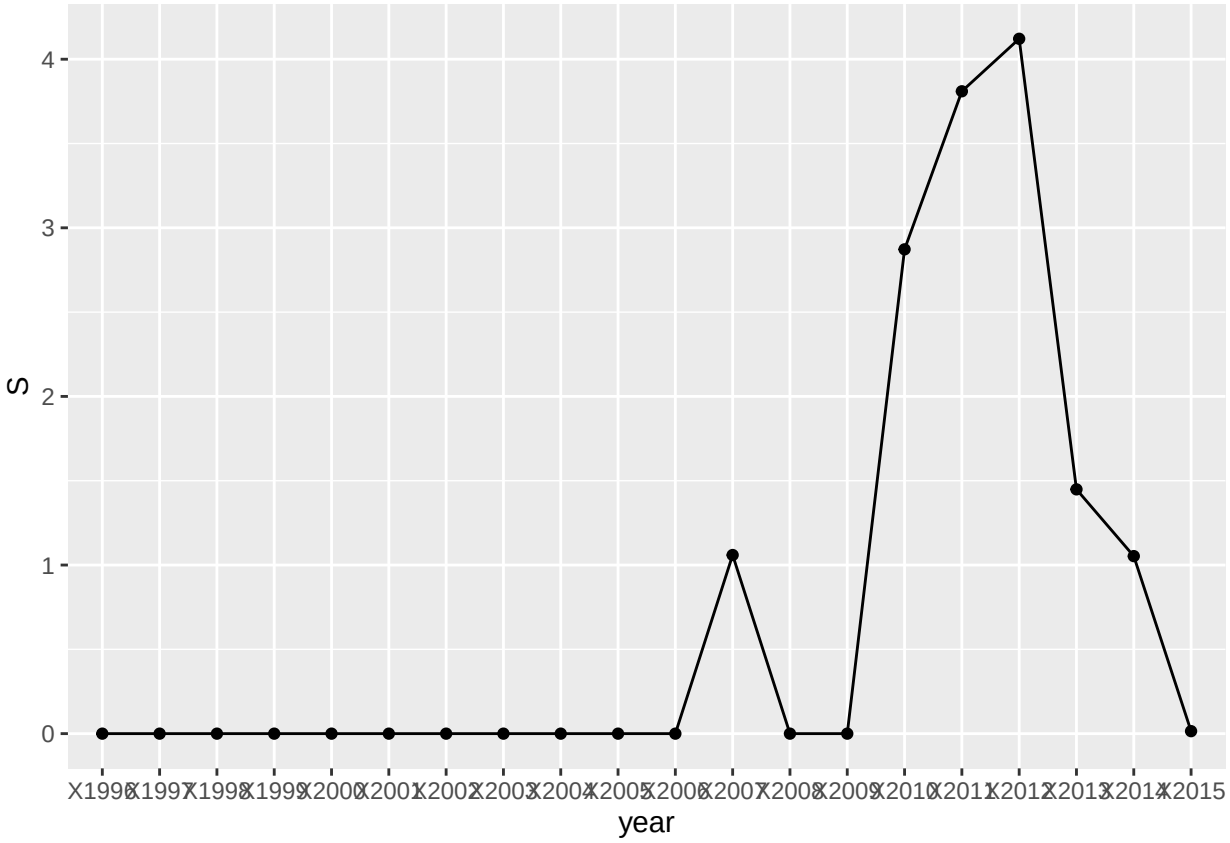
deriving S

```
ann_data <-ann_data %>%
  mutate(diff = year_avg - year_mean-C2)%>%
  data.frame()
S2 <-matrix(nrow = 21,0)
for (i in 1:20){
  if (i == 1){
    S2[i+1,1] <- max(0,S2[1,1]+ann_data[i,3])
  }else{
    S2[i+1,1] <- max(0,S2[i,1]+ann_data[i,3])
  }

}
```

graphically represented:

```
S2 <- data.frame(S2)
ann_data <-ann_data %>%
        mutate(S = S2[2:21,])
pl5 <- ggplot(ann_data, aes(x= year, y = S , group = 1))+
  geom_point()+
  geom_line()
```

```
pl5
```



The earliest year to cross the threshold is 2010.

```
subset(ann_data, S > thresh2)
```

```
##     year year_avg      diff        S
## 15 X2010 87.21138 2.8723577 2.872358
## 16 X2011 85.27642 0.9373984 3.809756
## 17 X2012 84.65041 0.3113821 4.121138
```

Conclusion: The temperature has been rising since 2010