Homework 3 - Ali Lakdawala, Kunle Lawal, Anu Rana, Mihir Tulpule

Introduction to Analytical Modeling -

Professor Sokol & Nirmit Chetwani

Question 4.2

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.

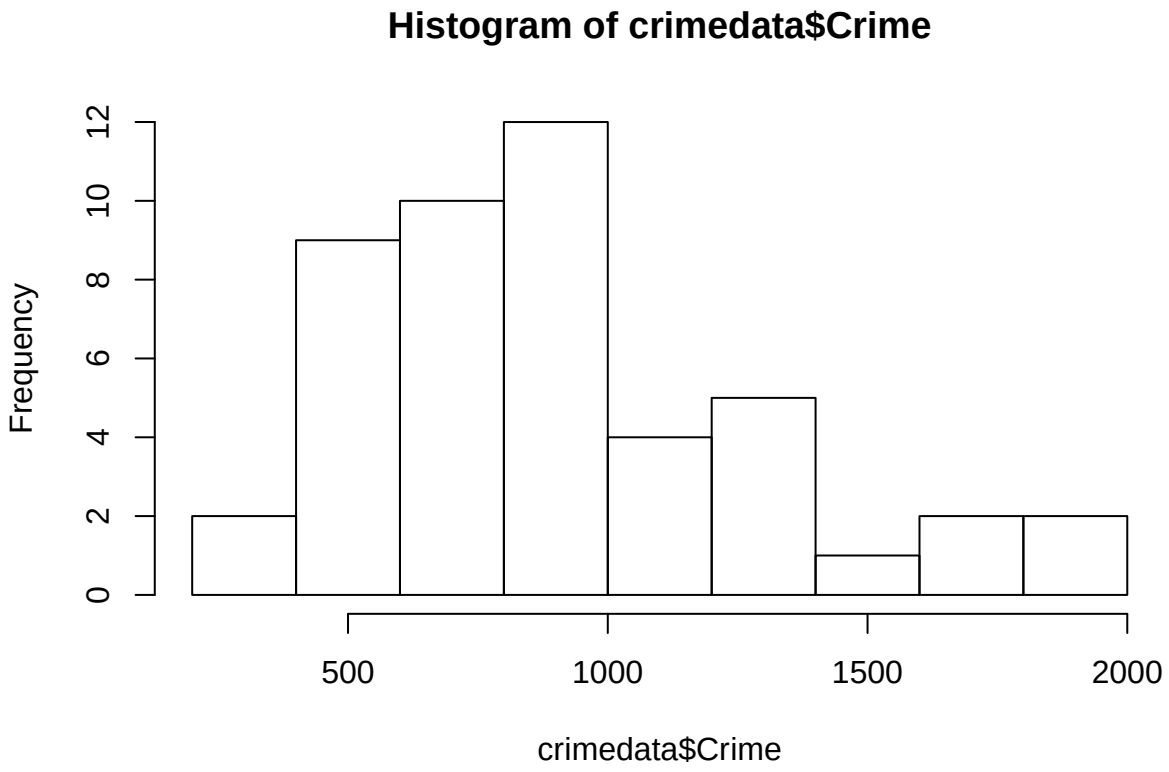
```
rm(list = ls())
setwd("/Users/alimujtaba/Google Drive/isye6501modelling/isye6501homeworks/hw3")
crimedata = read.table("uscrime.txt", header = TRUE)
#Exploring the Data
head(crimedata)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
##       Prob    Time Crime
## 1 0.084602 26.2011   791
## 2 0.029599 25.2999  1635
## 3 0.083401 24.3006   578
## 4 0.015801 29.9012  1969
## 5 0.041399 21.2998  1234
## 6 0.034201 20.9995   682
```

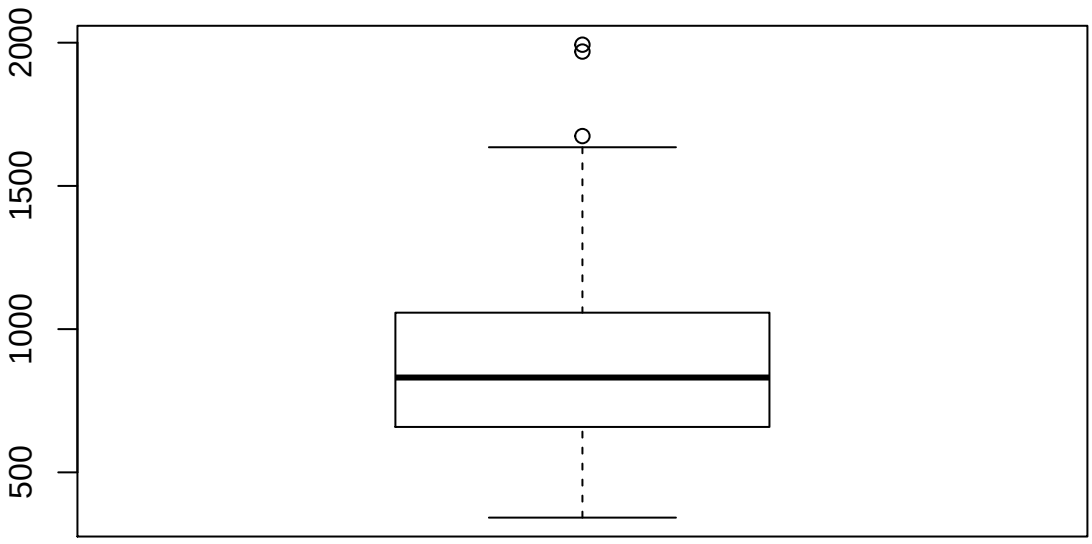
```
summary(crimedata$Crime)
```

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

```
#Visualzing to notice any outliers
hist(crimedata$Crime)
```




```
boxplot(crimedata$Crime)
```




```
# Our boxplot indicates outliers.
table(crimedata$Crime)
```

```
##
##  342  373  439  455  508  511  523  539  542  566  578  653  664  682  696
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  705  742  750  754  791  798  823  826  831  849  856  880  923  929  946
##    1    1    1    1    1    1    1    1    1    2    1    1    1    1    1
##  963  968 1030 1043 1072 1151 1216 1225 1234 1272 1555 1635 1674 1969 1993
##    1    1    1    1    1    1    2    1    1    1    1    1    1    1    1
```

Using the grubbs.test function:

```
require(outliers)
```

```
## Loading required package: outliers
```

```
help("grubbs.test")
#Test is based by calculating score of this outlier G (outlier minus mean and divided by sd)
# Alternative method is calculating ratio of variances of two datasets - full dataset and dataset witho
#The obtained value called U
x = crimedata$Crime
grubbs.test(x) #Defaults with type = 10
```

```
##
##  Grubbs test for one outlier
##
## data:  x
## G = 2.81290, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

```
grubbs.test(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
```

```
# For the two tail test we cannot reject the null hypothesis
```

We find with confidence that 1993 is an outlier using the grubbs.test function.

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?

Example: Heart beats per minute when working out in the gym

Max_heart_rate: (220bpm - age)

Optimal heart rate = less than 0.85(max_heart_rate)

Critical value ->

The rate of change of your heart beats within time range. For example when you're working out, you don't want your heart rate to change to fast because it could lead to cardiac arrest.

Threshold ->

Max threshold: your heart rate is too high/overworking and you're burning muscle leading to muscle deterioration. Min threshold: your heart rate is too low and you are not working out hard enough.

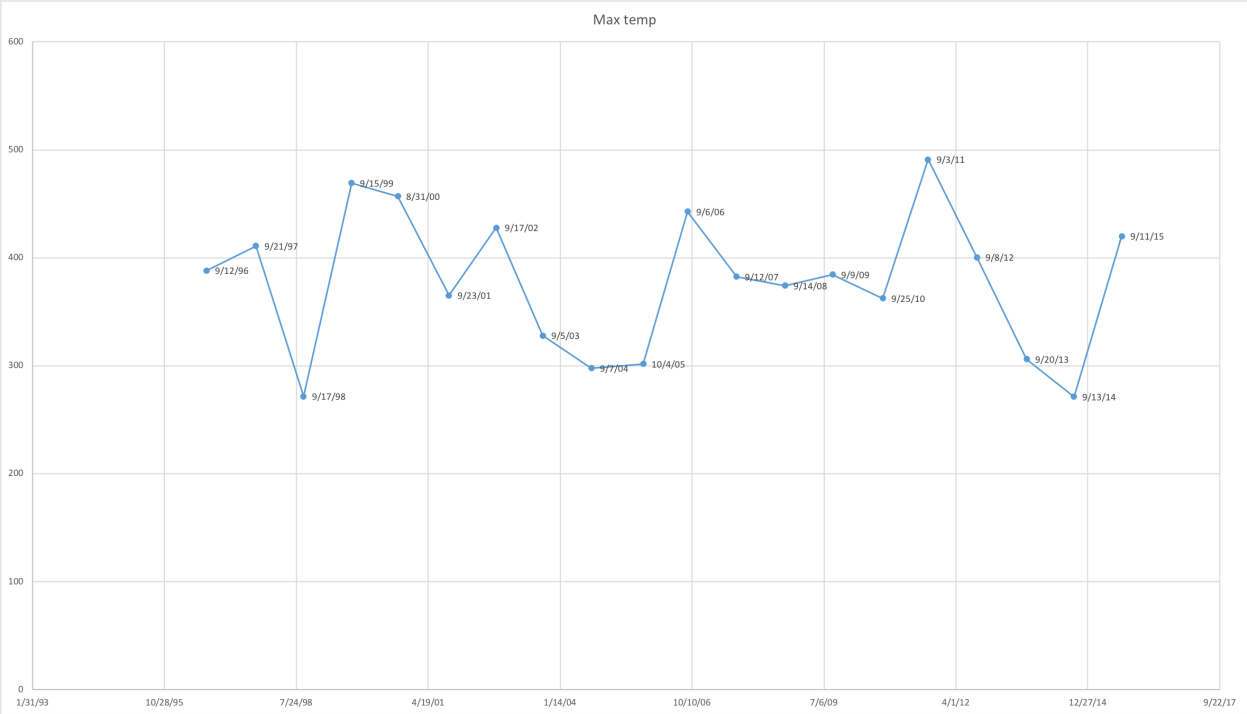
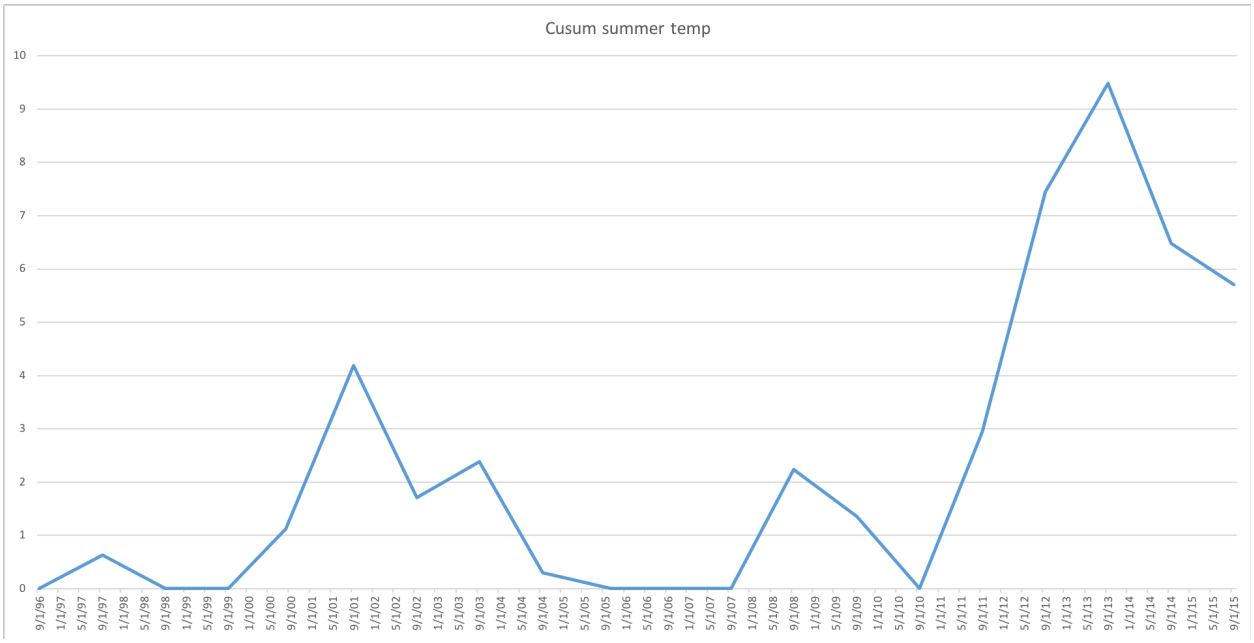


Figure 1: The last day of summer for each year

6.2.2

(a) Using the CUSUM method we found the date with the maximum CUSUM temperature for a particular year. Please reference the attached Excel sheet 2 for the calculation.

(b) Again using the CUSUM method we calculated the average summer temprature for each year. Using this data, we found that overall trend of increasing temperatures in ATL. We noticed a significant spike in average temperature beginning in 2010. Please reference the attached Excel sheet. ("q6.2.txt") for the calculations.



It is also important to determine with mean is a good estimate of this data or whether there are outliers biasing the data. We can determine this easily through boxplots.

```
dat <- read.csv("cusum_excel.csv", header = TRUE)

head(dat)
```

```
##      DAY  st.1996   st.1997   st.1998    st.1999  st.2000   st.2001
## 1         0.00000  0.000000  0.000000  0.0000000  0.00000  0.000000
## 2 7/1/18 14.28455  4.325203  6.739837  0.6422764  4.96748  2.447154
## 3 7/2/18 27.56911 12.650407 10.479675  0.0000000 11.93496  7.894309
## 4 7/3/18 40.85366 23.975610 17.219512  3.6422764 20.90244 13.341463
## 5 7/4/18 47.13821 33.300813 23.959350  8.2845528 31.86992 15.788618
## 6 7/5/18 52.42276 35.626016 30.699187 14.9268293 43.83740 20.235772
##     st.2002   st.2003    st.2004   st.2005  st.2006  st.2007   st.2008
## 1  0.000000  0.000000  0.0000000  0.000000  0.00000 0.000000  0.000000
## 2  6.414634  0.000000  0.2357724  7.642276  9.95122 9.601626  2.487805
## 3 12.829268  0.000000  0.0000000 13.284553 19.90244 9.203252  6.975610
## 4 16.243902  5.520325  4.2357724 15.926829 29.85366 5.804878 15.463415
## 5 21.658537 10.040650 10.4715447 18.569106 37.80488 6.406504 22.951220
## 6 31.073171  8.560976 18.7073171 24.211382 44.75610 9.008130 28.439024
##    st.2009   st.2010   st.2011  st.2012   st.2013   st.2014  st.2015
## 1  0.00000 0.0000000  0.000000  0.00000 0.0000000  0.000000 0.000000
## 2 14.00813 0.0000000  6.723577 20.34959 0.3333333  6.056911 1.699187
## 3 23.01626 0.0000000 15.447154 28.69919 3.6666667 15.113821 5.398374
## 4 31.02439 0.0000000 25.170732 43.04878 0.0000000 18.170732 1.097561
## 5 41.03252 0.0000000 31.894309 56.39837 0.0000000 18.227642 2.796748
## 6 40.04065 0.7886179 36.617886 71.74797 1.3333333 20.284553 3.495935
##    DAY.1 Year  X Max.Temp.Date Max.temp X.1 X.2 X.3 X.4
## 1        1996 NA       9/12/96 388.0569  NA  75  76 A76
## 2 7/1/18 1997 NA       9/21/97 410.9919  NA  NA  NA
## 3 7/2/18 1998 NA       9/17/98 271.4472  NA  NA  NA
## 4 7/3/18 1999 NA       9/15/99 469.1707  NA  NA  NA
```

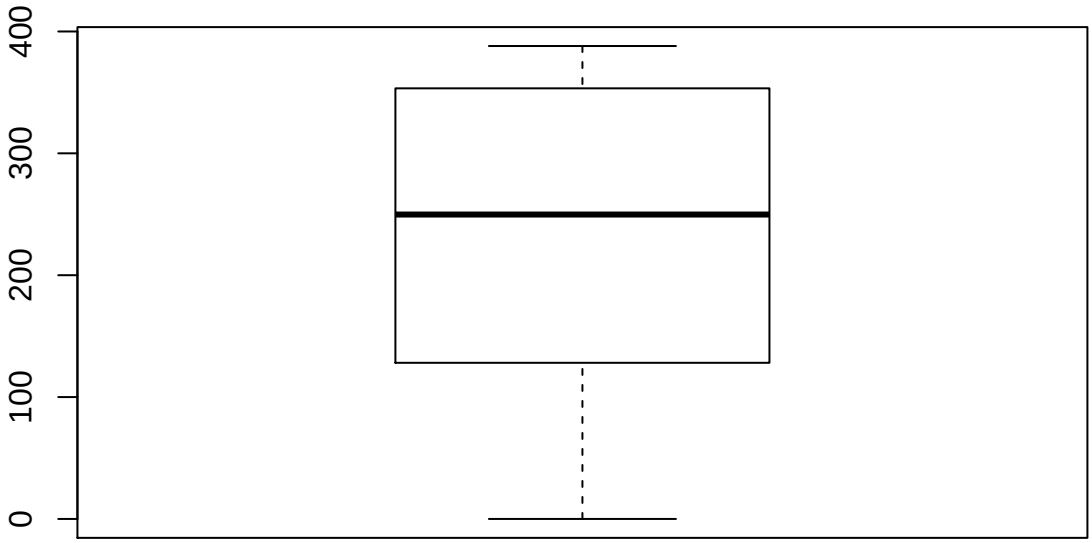
```
## 5 7/4/18 2000 NA       8/31/00 456.9837  NA  NA  NA
## 6 7/5/18 2001 NA       9/23/01 365.0081  NA  NA  NA
```

```
library(ggplot2)

summary(dat$st.1996)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   128.6   249.8   230.9   353.3   388.1
```

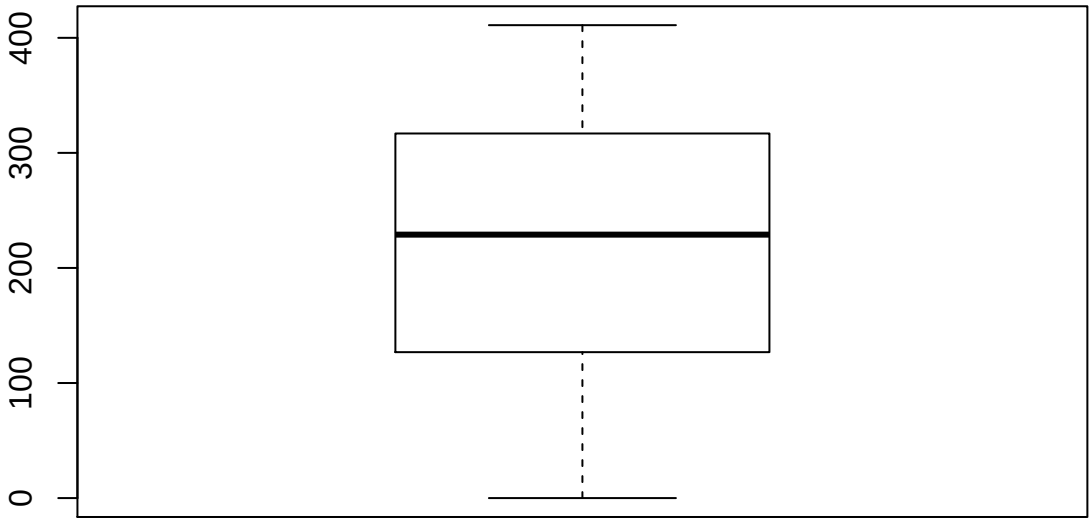
```
boxplot(dat$st.1996)
```




```
summary(dat$st.1997)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   128.1   228.9   219.8   316.8   411.0
```

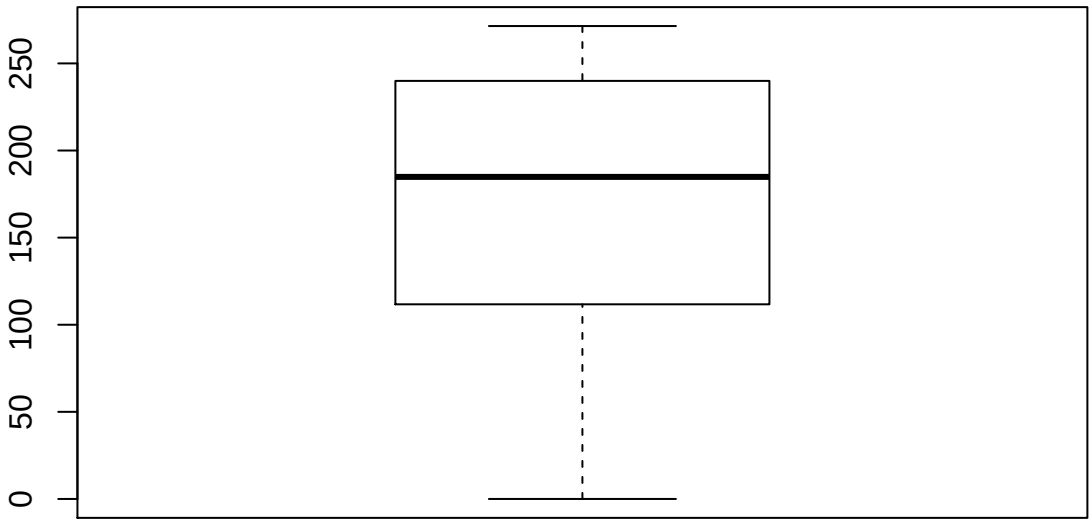
```
boxplot(dat$st.1997)
```




```
summary(dat$st.1998)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   112.9   184.9   166.3   239.9   271.4
```

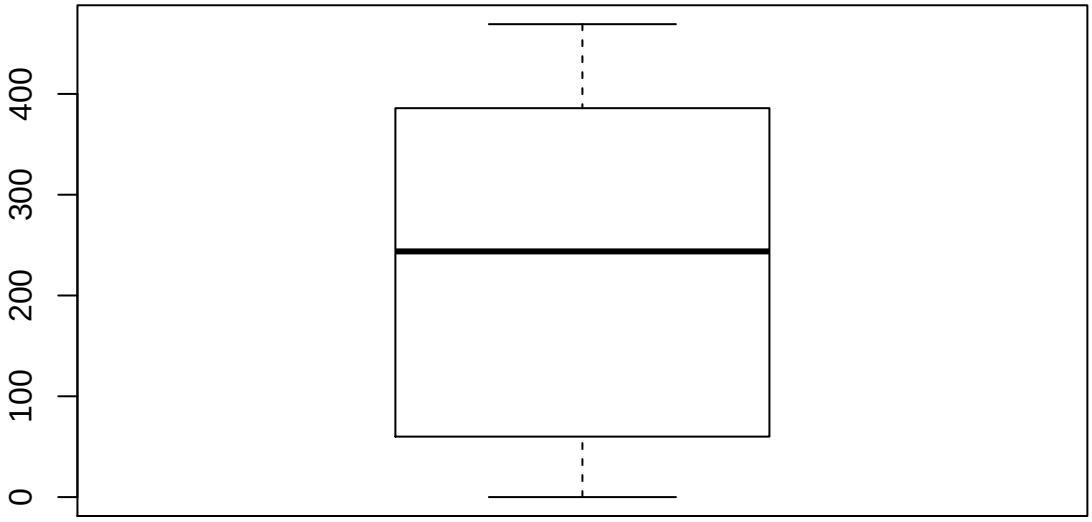
```r
boxplot(dat$st.1998)
```




```r
summary(dat$st.1999)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   60.93  243.74  234.62  384.33  469.17
```

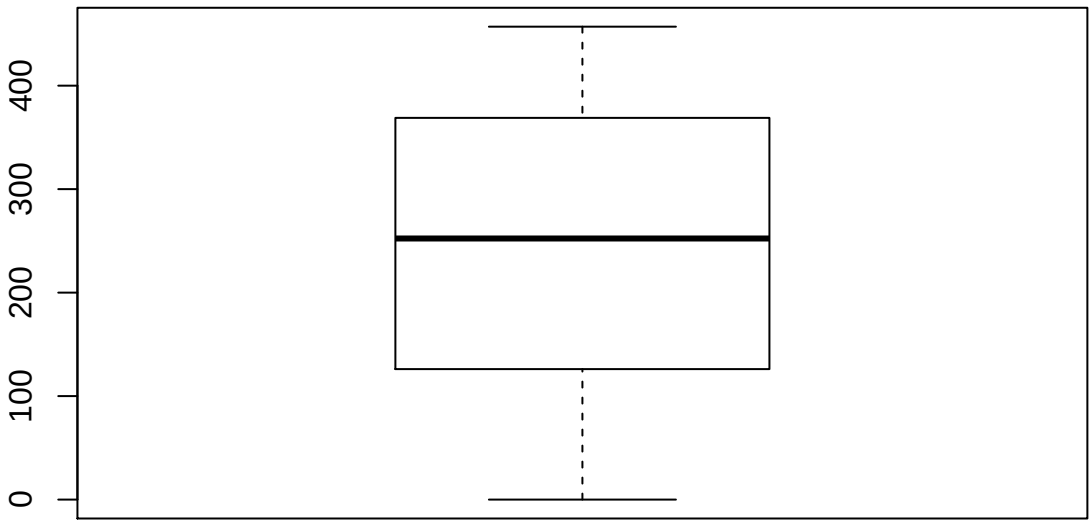
```r
boxplot(dat$st.1999)
```




```r
summary(dat$st.2000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   127.8   252.3   246.9   368.1   457.0
```

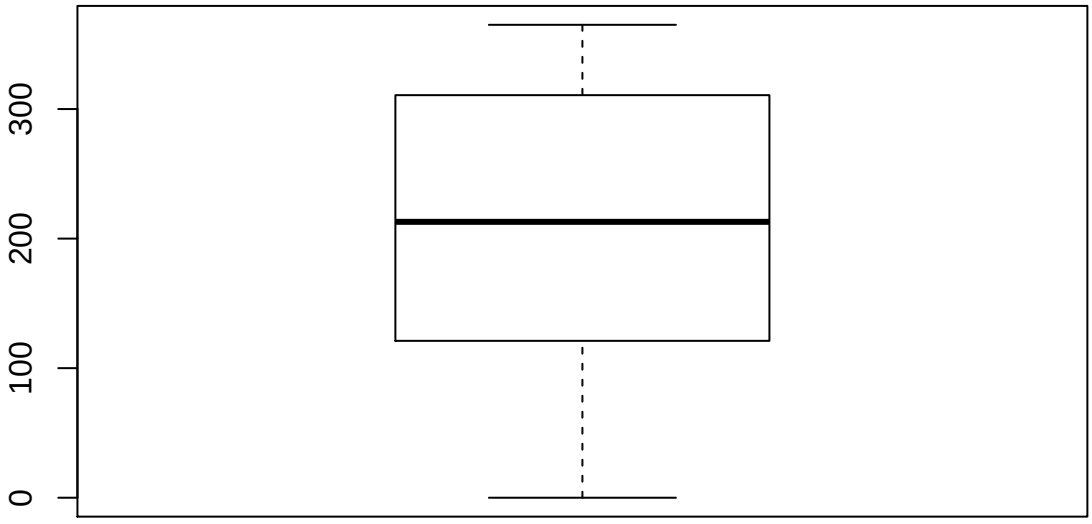
```r
boxplot(dat$st.2000)
```




```
summary(dat$st.2001)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   121.5   213.0   204.5   309.6   365.0
```

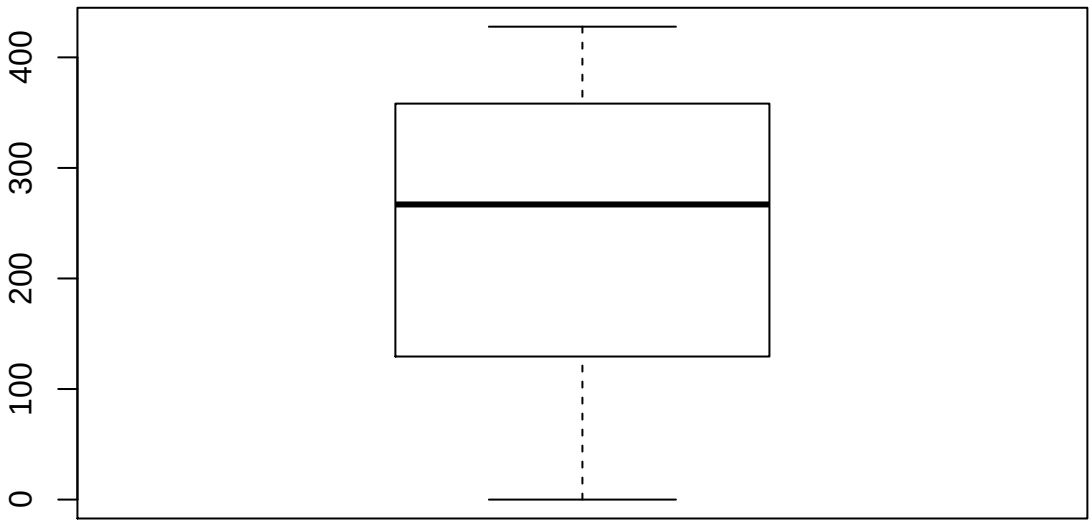
```
boxplot(dat$st.2001)
```




```
summary(dat$st.2002)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   131.3   267.0   244.8   357.9   427.8
```

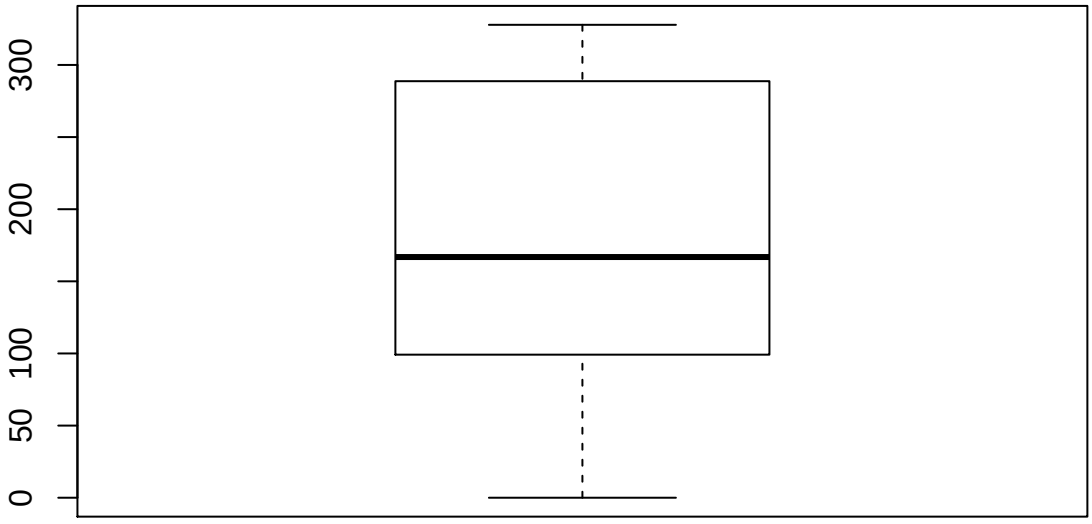
```
boxplot(dat$st.2002)
```




```
summary(dat$st.2003)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   99.29  166.85  177.67  288.53  327.82
```

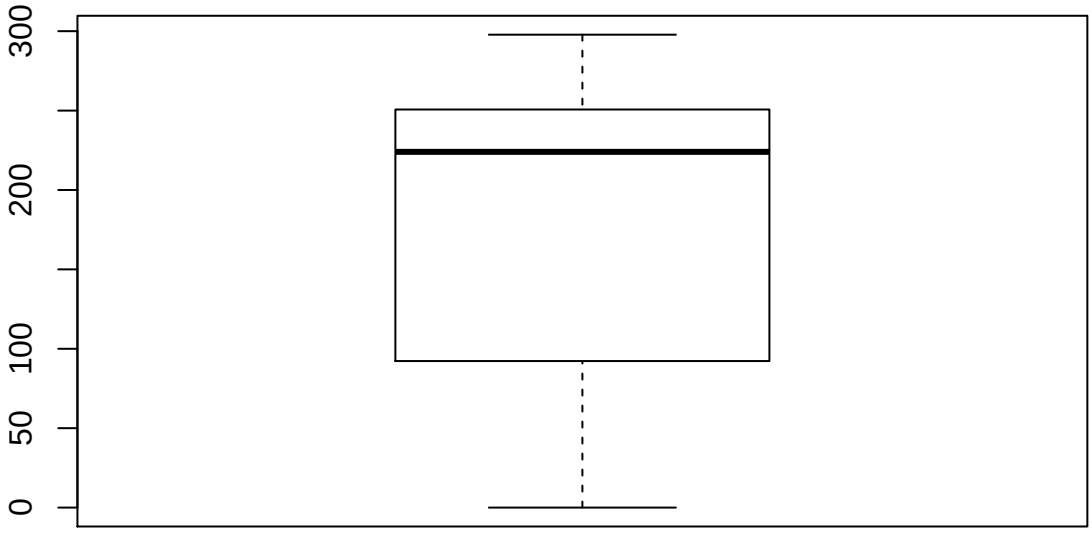
```
boxplot(dat$st.2003)
```




```
summary(dat$st.2004)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   92.75  224.02  177.04  250.63  297.80
```

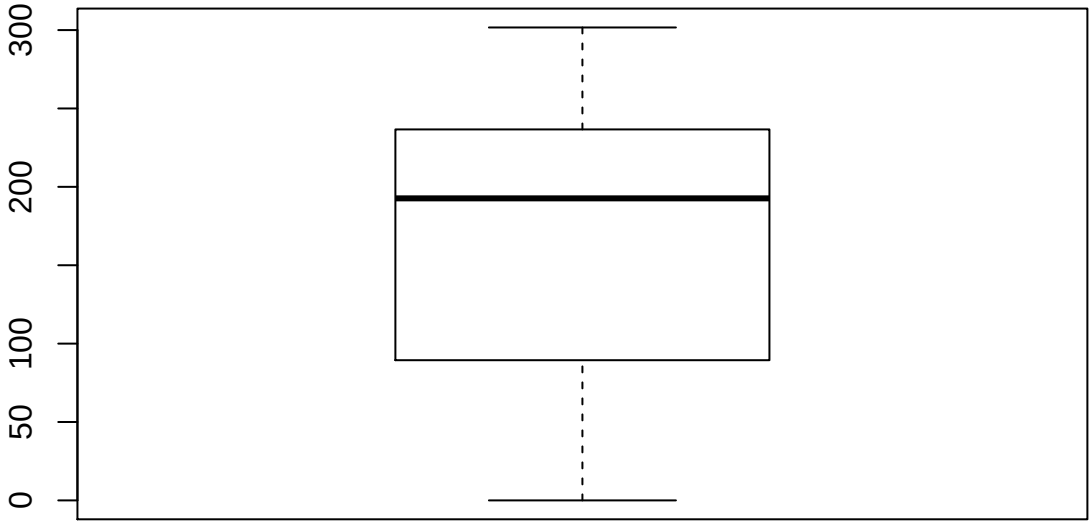
```
boxplot(dat$st.2004)
```




```
summary(dat$st.2005)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   89.74  192.67  166.48  236.51  301.66
```

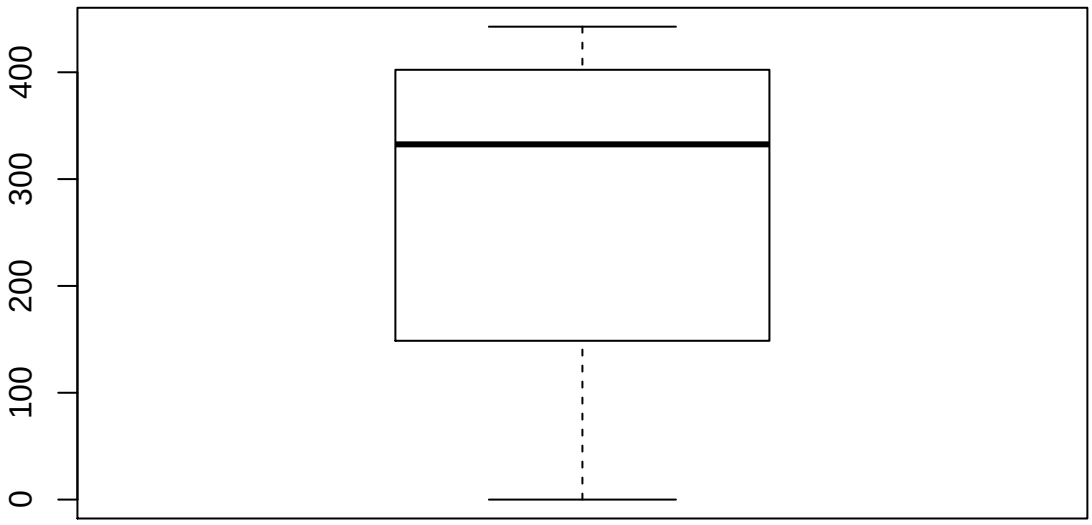
```
boxplot(dat$st.2005)
```




```
summary(dat$st.2006)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   149.0   332.5   273.5   401.8   442.7
```

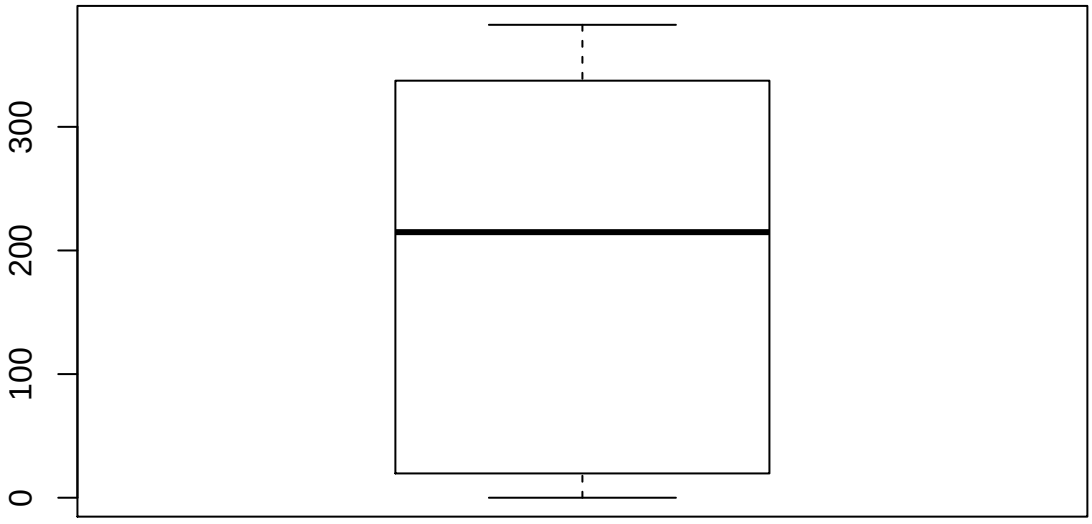
```
boxplot(dat$st.2006)
```




```
summary(dat$st.2007)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   20.05  214.83  193.91  337.25  382.52
```

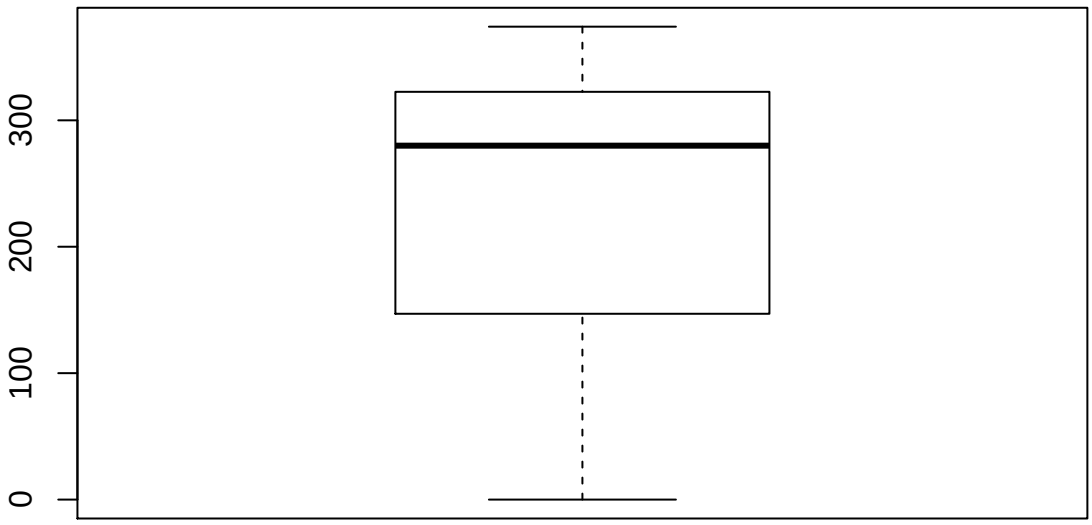
```
boxplot(dat$st.2007)
```




```
summary(dat$st.2008)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   148.1   280.0   234.9   322.4   374.1
```

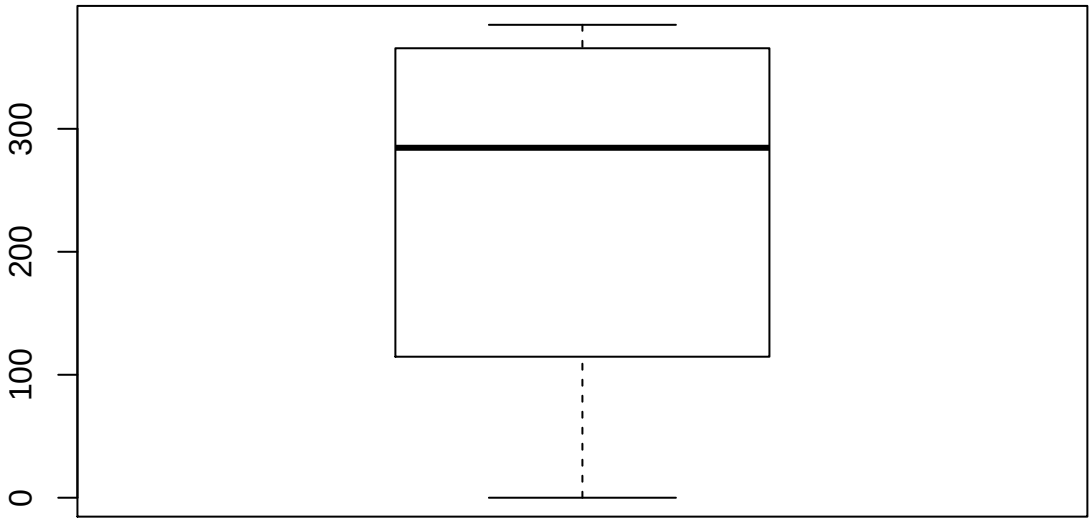
```
boxplot(dat$st.2008)
```




```
summary(dat$st.2009)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   115.4   284.6   240.2   365.0   384.6
```

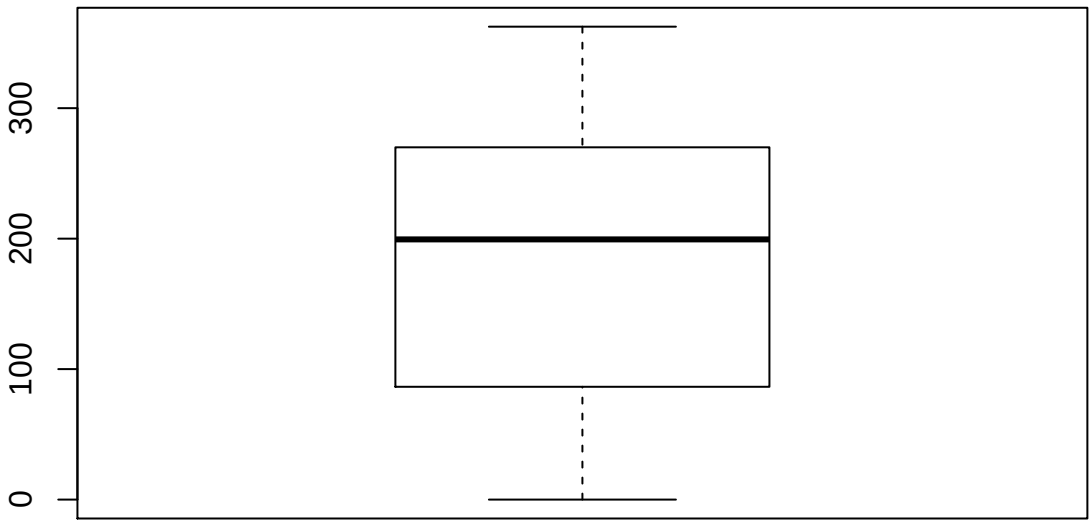
```
boxplot(dat$st.2009)
```




```
summary(dat$st.2010)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   87.61  199.41  183.03  269.84  362.46
```

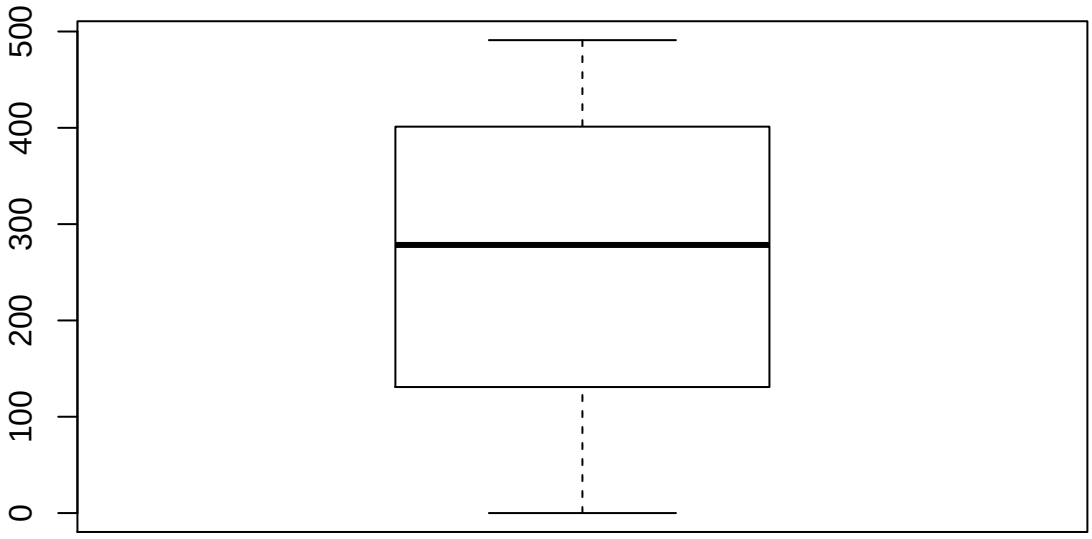
```
boxplot(dat$st.2010)
```




```
summary(dat$st.2011)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   132.2   278.4   265.9   401.1   491.0
```

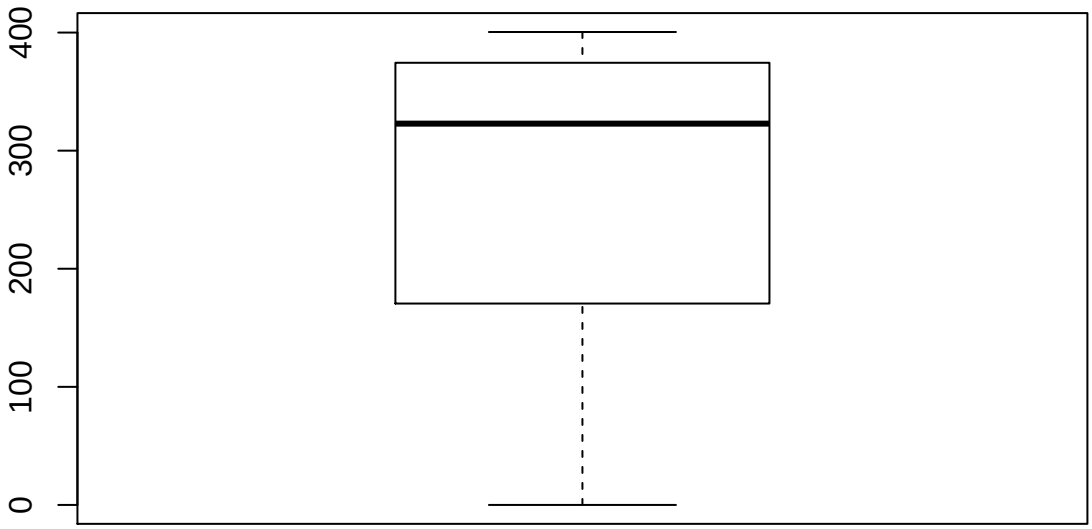
```
boxplot(dat$st.2011)
```




```
summary(dat$st.2012)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   170.9   322.9   272.0   374.3   400.5
```

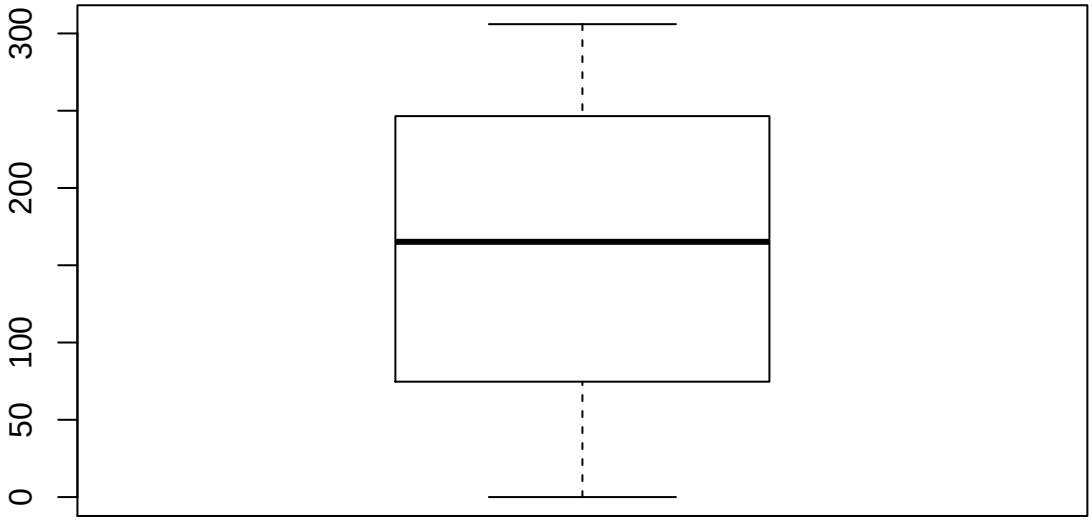
```
boxplot(dat$st.2012)
```




```
summary(dat$st.2013)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   75.17  165.17  158.40  244.92  306.00
```

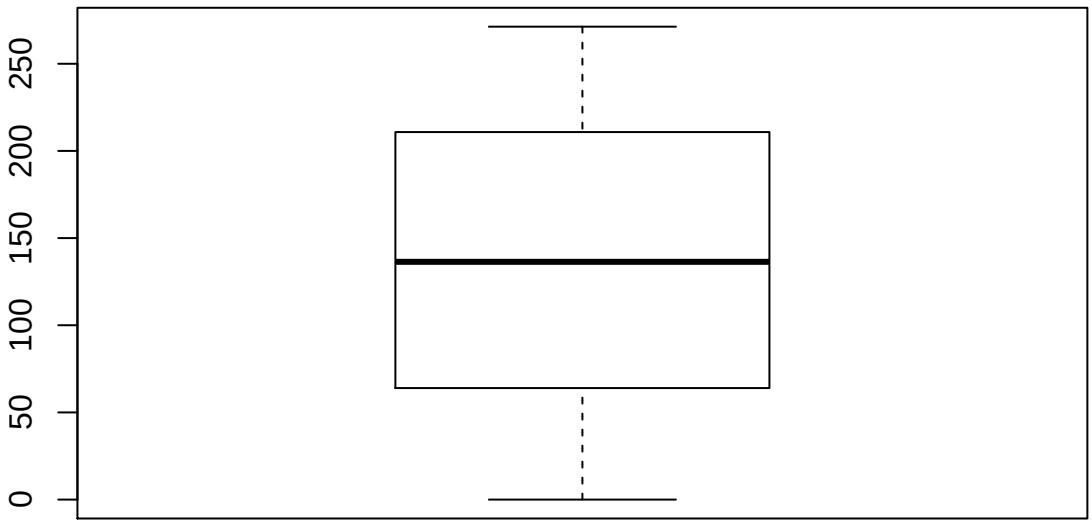
```
boxplot(dat$st.2013)
```




```
summary(dat$st.2014)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   64.02  136.38  140.94  210.56  271.27
```

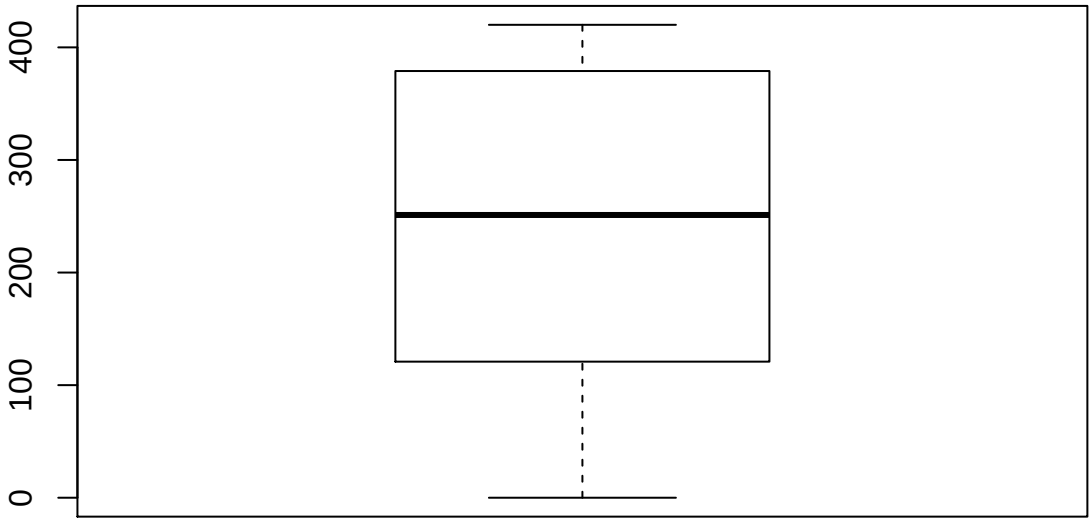
```
boxplot(dat$st.2014)
```




```
summary(dat$st.2015)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   121.9   251.1   239.6   378.8   420.0
```

```
boxplot(dat$st.2015)
```



We find that the 75th percentile could be a good estimate for our CUMSUM calculations.

Howevever, we get a very similar result with 2010 as the main inflection year. Please see attached excel sheet for calculations.

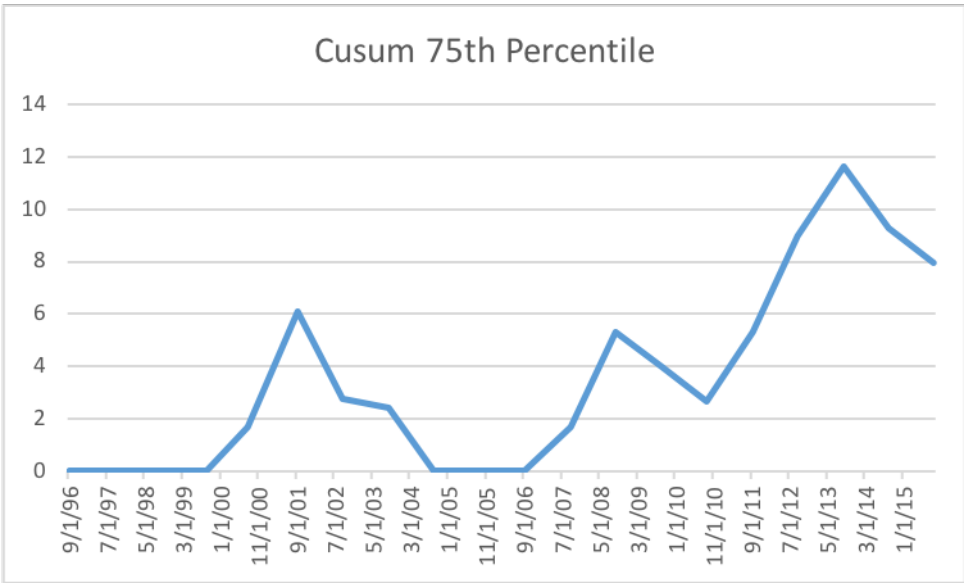


Figure 2: CUSUM using 75th percentile for temperature for each of the years.