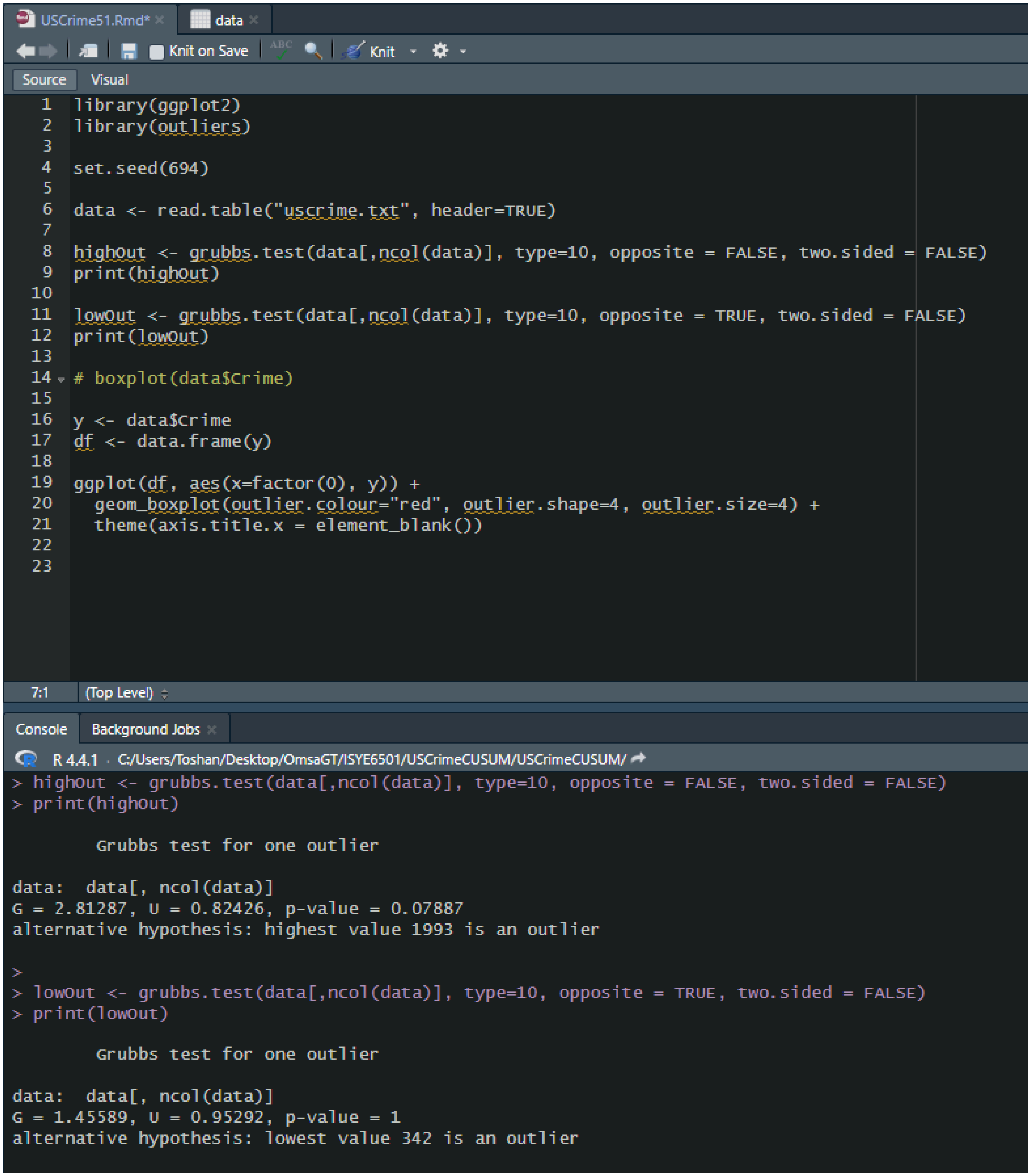
Question 5.1

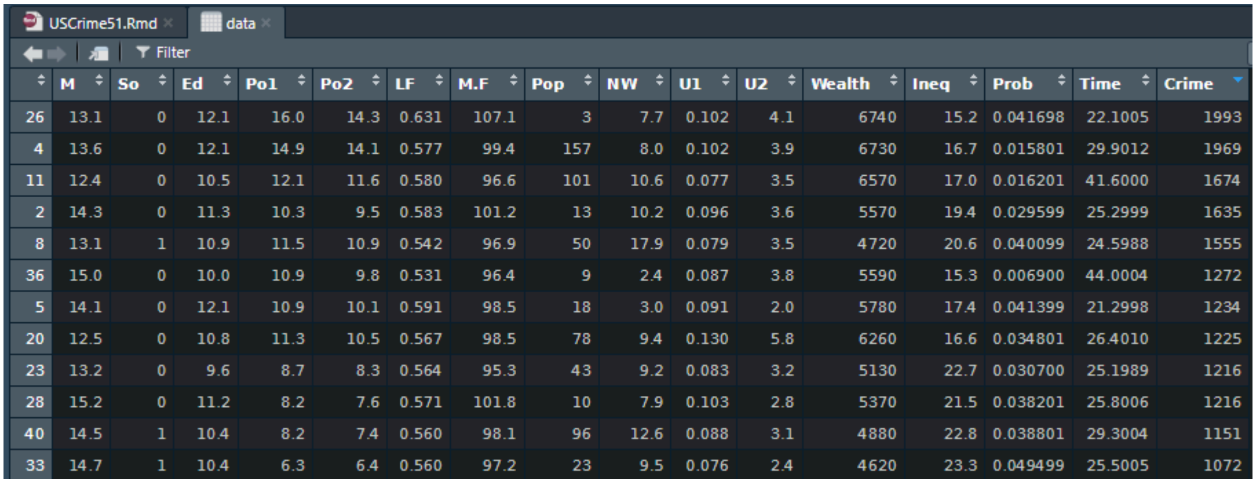
Given the US Crime dataset, my task was to determine where there are any outliers present in the Crime column using the outliers package in R. Figure 1 details the exact code used to produce the outliers based on the data. The value of 1993 is an outlier present in the data. The p-value of this data point is 0.07887, which is statistically significant in proving its case an outlier. The p-value dictates that this result would be produced in about 7% of trials. Furthermore, despite generating a lowest outlier, the p-value of that point is exactly 1. Meaning that it should not be interpreted as an outlier. These claims are further substantiated by the boxplot displayed in Figure 3. Figure 3 is a graphical representation of ggplot2's analysis of the outliers in the dataset, and as seen in the figure, it is clear that the outliers are on the high end of the dataset rather than the low end. Figure 2 gives substance to this conclusion as you can see how much higher the values are of the high end of the dataset.



```r
library(ggplot2)
library(outliers)

set.seed(694)

data <- read.table("uscrime.txt", header=TRUE)

highOut <- grubbs.test(data[,ncol(data)], type=10, opposite = FALSE, two.sided = FALSE)
print(highOut)

lowOut <- grubbs.test(data[,ncol(data)], type=10, opposite = TRUE, two.sided = FALSE)
print(lowOut)

# boxplot(data$Crime)

y <- data$Crime
df <- data.frame(y)

ggplot(df, aes(x=factor(0), y)) +
  geom_boxplot(outlier.colour="red", outlier.shape=4, outlier.size=4) +
  theme(axis.title.x = element_blank())


```

```
> highOut <- grubbs.test(data[,ncol(data)], type=10, opposite = FALSE, two.sided = FALSE)
> print(highOut)

        Grubbs test for one outlier

data:  data[, ncol(data)]
G = 2.81287, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier

>
> lowOut <- grubbs.test(data[,ncol(data)], type=10, opposite = TRUE, two.sided = FALSE)
> print(lowOut)

        Grubbs test for one outlier

data:  data[, ncol(data)]
G = 1.45589, U = 0.95292, p-value = 1
alternative hypothesis: lowest value 342 is an outlier
```

Figure 1. This figure denotes the code used to determine the highest and lowest outlier from the US Crime dataset, as well as the code for generating the box plot seen in Figure 3.



	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
26	13.1	0	12.1	16.0	14.3	0.631	107.1	3	7.7	0.102	4.1	6740	15.2	0.041698	22.1005	1993
4	13.6	0	12.1	14.9	14.1	0.577	99.4	157	8.0	0.102	3.9	6730	16.7	0.015801	29.9012	1969
11	12.4	0	10.5	12.1	11.6	0.580	96.6	101	10.6	0.077	3.5	6570	17.0	0.016201	41.6000	1674
2	14.3	0	11.3	10.3	9.5	0.583	101.2	13	10.2	0.096	3.6	5570	19.4	0.029599	25.2999	1635
8	13.1	1	10.9	11.5	10.9	0.542	96.9	50	17.9	0.079	3.5	4720	20.6	0.040099	24.5988	1555
36	15.0	0	10.0	10.9	9.8	0.531	96.4	9	2.4	0.087	3.8	5590	15.3	0.006900	44.0004	1272
5	14.1	0	12.1	10.9	10.1	0.591	98.5	18	3.0	0.091	2.0	5780	17.4	0.041399	21.2998	1234
20	12.5	0	10.8	11.3	10.5	0.567	98.5	78	9.4	0.130	5.8	6260	16.6	0.034801	26.4010	1225
23	13.2	0	9.6	8.7	8.3	0.564	95.3	43	9.2	0.083	3.2	5130	22.7	0.030700	25.1989	1216
28	15.2	0	11.2	8.2	7.6	0.571	101.8	10	7.9	0.103	2.8	5370	21.5	0.038201	25.8006	1216
40	14.5	1	10.4	8.2	7.4	0.560	98.1	96	12.6	0.088	3.1	4880	22.8	0.038801	29.3004	1151
33	14.7	1	10.4	6.3	6.4	0.560	97.2	23	9.5	0.076	2.4	4620	23.3	0.049499	25.5005	1072

Figure 2. This figure is the US Crime dataset sorted by greatest numerical value under the column Crime.

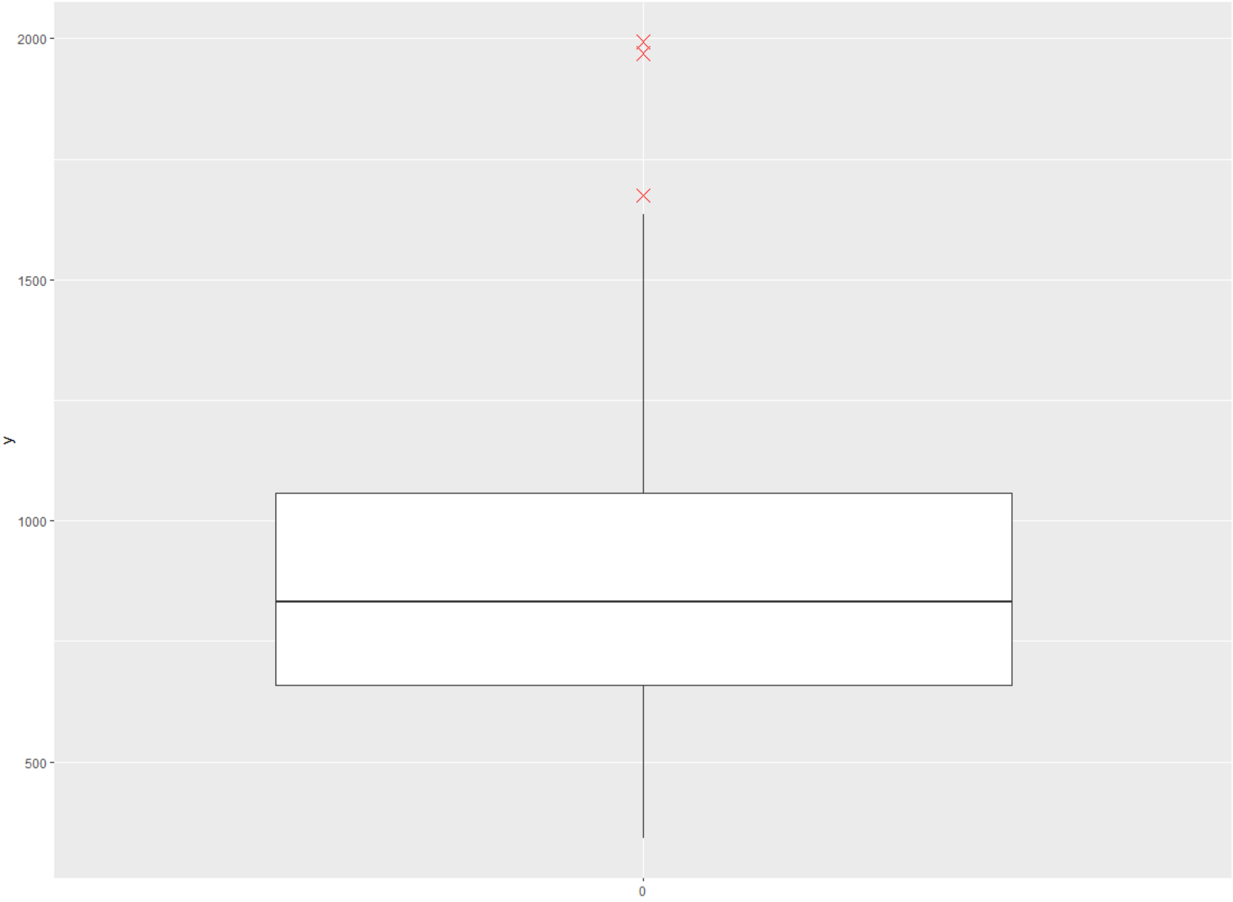


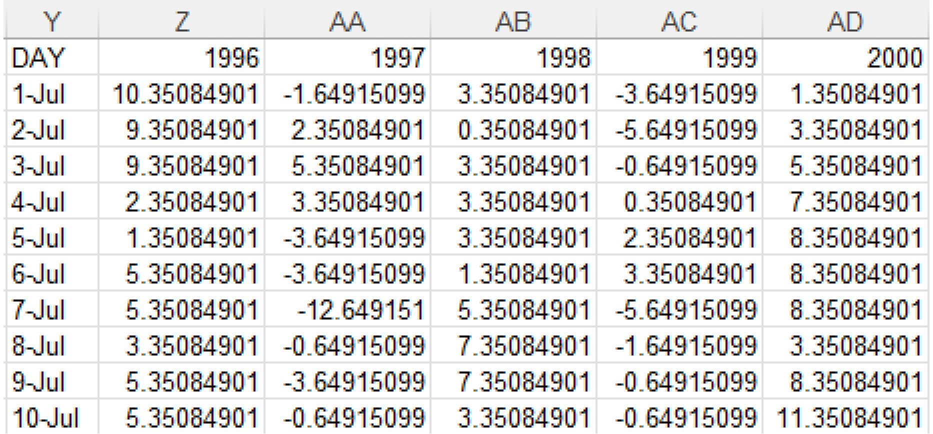
Figure 3. This figure depicts a boxplot of the US Crime dataset, with outliers denoted as a red X.

Question 6.1

The Change Detection model could be useful in many avenues of life, but for me specifically, it would have the greatest effect on the ability to read and react to stock prices. This would allow me to buy and sell at the greatest time to maximize profit. Furthermore, it would allow me to determine if anything has changed with a company I am invested in based on any change above a certain threshold. By this, I mean that if a stock price has begun dropping significantly, I could do some research and determine whether it is the result of bad PR, a malfunction in one of their factories, etc. To determine this threshold, I would set strict critical values and lower thresholds as I do not want to be notified too late that my money has been lost due to a plummeted stock price.

Question 6.2

1. In Figure 5, this graph depicts a significant downward trend around the 25th of September, however it consistently starts violating the lower bound of the threshold around the 13th of September. Given that the Autumn Equinox is typically the 22nd or 23rd of September, using that in conjunction with the perceived data in Figure 5 we can conclude that the end of summer falls right near the 25th of September.
2. In the following figures, I chose to normalize and use the CUSUM approach to model the data. Figure 4 is a small snippet of the final normalized data table using a C-value of ½ the standard deviation of the original dataset (taken from office hours). I utilized the CUSUM equation to normalize the data and set a threshold value of 12. In the beginning, I started with a threshold value of 5 times the standard deviation, but I found that it did not truly model the data in an empirical fashion. Taking into consideration this and what I gained from office hours (selecting T and C values to reduce subjectivity as much as possible), I found that a T value of 12 fit the data very nicely. Based on the final product generated in Figure 5, it can be determined that Atlanta's summer climate did see an increase in the years 2000, 2007, 2011, and 2012. Otherwise, the data was not significant enough to determine that the summer climate did get warmer over time.



Y	Z	AA	AB	AC	AD
DAY	1996	1997	1998	1999	2000
1-Jul	10.35084901	-1.64915099	3.35084901	-3.64915099	1.35084901
2-Jul	9.35084901	2.35084901	0.35084901	-5.64915099	3.35084901
3-Jul	9.35084901	5.35084901	3.35084901	-0.64915099	5.35084901
4-Jul	2.35084901	3.35084901	3.35084901	0.35084901	7.35084901
5-Jul	1.35084901	-3.64915099	3.35084901	2.35084901	8.35084901
6-Jul	5.35084901	-3.64915099	1.35084901	3.35084901	8.35084901
7-Jul	5.35084901	-12.649151	5.35084901	-5.64915099	8.35084901
8-Jul	3.35084901	-0.64915099	7.35084901	-1.64915099	3.35084901
9-Jul	5.35084901	-3.64915099	7.35084901	-0.64915099	8.35084901
10-Jul	5.35084901	-0.64915099	3.35084901	-0.64915099	11.35084901

Figure 4. This figure is a small snippet of the normalized data using the CUSUM equation (C-value=4.31, T-value=12).

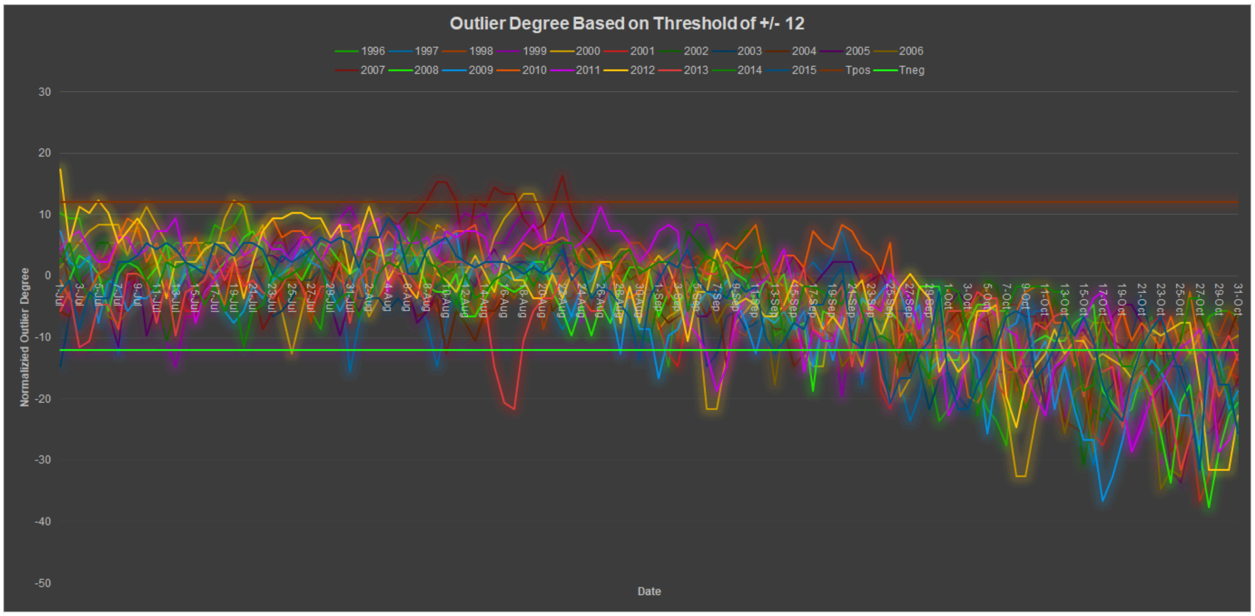


Figure 5. This graph displays exactly how the data changed over time. It also shows the threshold for the upper and lower bound.