Homework 2 - ISYE6501-OA

5/27/2018

Objectives

The focus of Week2 was Classification, Data Preparation, Outliers, and Change Detection (with specific application to CUSUM). This homework tests several of these aspects.

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.

I want to understand my family's spending habits in order to better balance our budget as well as save for the future. Data can be extracted in csv format from every bank or credit card company's online portal. There could be several attributes/predictors that play into the spending habits

1. **day of week** There could be potentially seven clusters which show the cumulative spending for that day.
2. **category** These categories are spread across: Utilities, Groceries, Home, Automobiles, Discretionary Spending (such as eating out), Retail (clothes and other purchases)
3. **times of year** There may be some times more than others (Xmas, Memorial day, birthdays) where my family and I have more propensity to spend.
4. **location** Do we spend more in one location vs another. This may create too many clusters but worth exploring
5. **person** Who is doing the spending. This may be a predictor , or may be a lens to look at all the other predictors above. For instance, I would run 1 through 4 for myself, and then a separate (facet) view of 1-4 for my wife and so on, so separate spending patterns and look for insights.

Question 4.2

Q4.2 Use the R function kmeans to cluster the points as well as possible [using the iris data]. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

For this question, let's understand the data first. In 1936, a biologist named Ronald Fisher measured 50 data sets across three different species of the iris flower.

Here's a sample of the iris data set:

```
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
```

Some important points before beginning:

- this data set is essentially unsupervised data. We dont necessarily know the grouping. However, the biologist Ron Fisher made this data unique by adding the labels (aka Species) to each data set.
- in real life we would not have this information available
- and given a new measurement of sepal or petal length/width we would have to use the clustering process to identify which cluster that new value would belong to.
- however, since we know that these measurements are grouped into 3 species, we basically already know that the optimum clusters are 3.
- however, we shall still evaluate how the kmeans() algorithm performs for other values of k, and *objectively* attempt to come to the most optimum cluster value
- in short, **we will have selective amnesia during this exercise that k really is 3 , and try to figure it out by analysis!)**

Let's disect this data.. We have four predictors: *(the last column, Species, is the manual categorization done by the biologist, its not a predictor)*

```
colnames(irisTable)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
## [5] "Species"
```

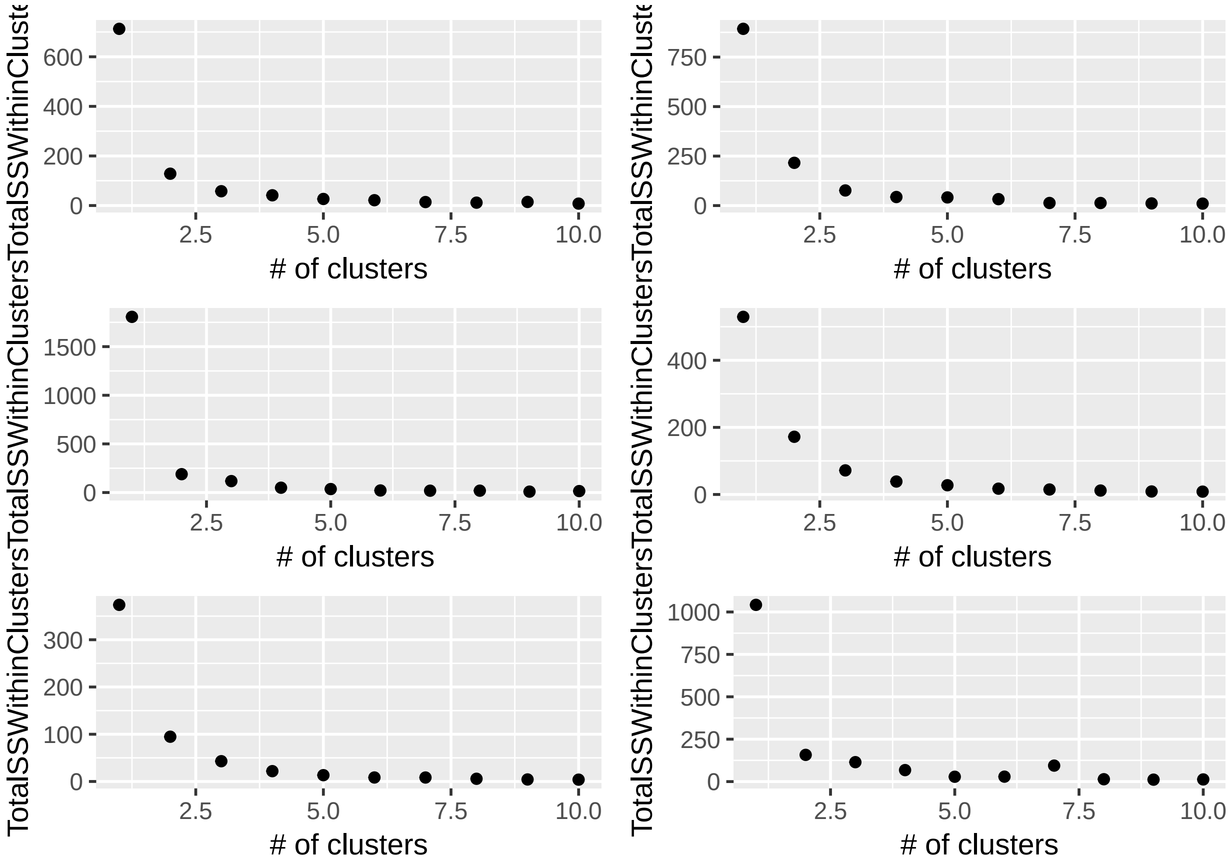
We have 4 predictors that can be combined in groups of 2 (i.e. tuple length is 2) This means we have C(4,2) = 6 of combinations we can use in 2-dimensional clustering, using the combinations formula.. n!/[(n-k)!*k!]

These are:

1. Sepal.Length plotted against Sepal.Width
2. Sepal.Length plotted against Petal.Length
3. Sepal.Length plotted against Petal.Width
4. Sepal.Width plotted against Petal.Length
5. Sepal.Width plotted against Petal.Width
6. Petal.Length plotted against Petal.Width

Against each of these combinations, we run cluster k value of 1 through 10 to see the outcome.. Specifically, we plot the number of clusters against the total sum of squares within each cluster.

The argument being that the sum of squares within a cluster needs to be minimized by choosing the appropriate number of clusters.

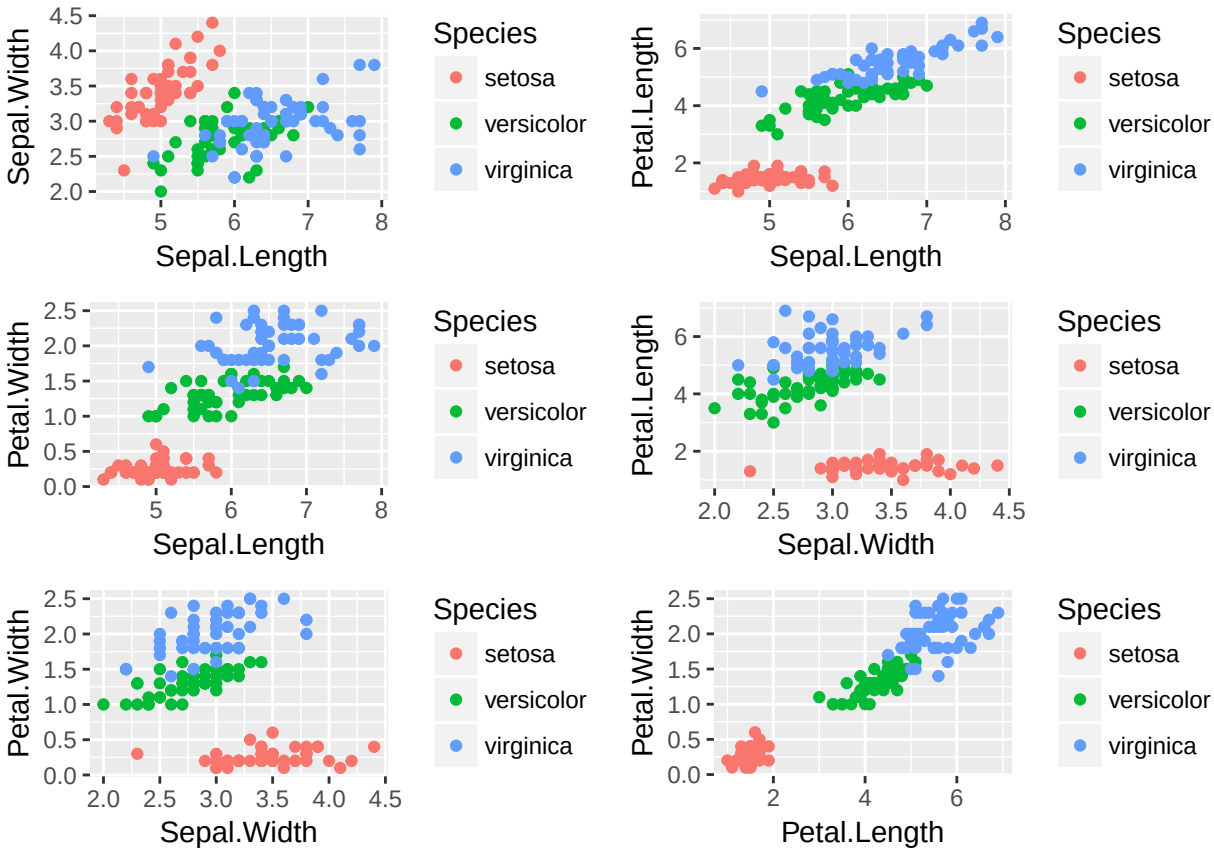


Summary

As can be seen from the above, there is diminishing improvement of the total sum of squares within each cluster after k = 3 . Implies that the optimum number of clusters witnessed within the data set is 3, as that is where the "kink" in the elbow curve lies.

Some post-analysis confirmation:

As mentioned earlier, Ronald Fisher already provided us the answer (i.e. number of clusters) within his data. Knowing that, and working backwards, if we plot the three clusters (categorized by "Species" using ggplot2, we see the following:



The clean demarcation in almost all cases of shows that indeed the number of clusters is 3.

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.

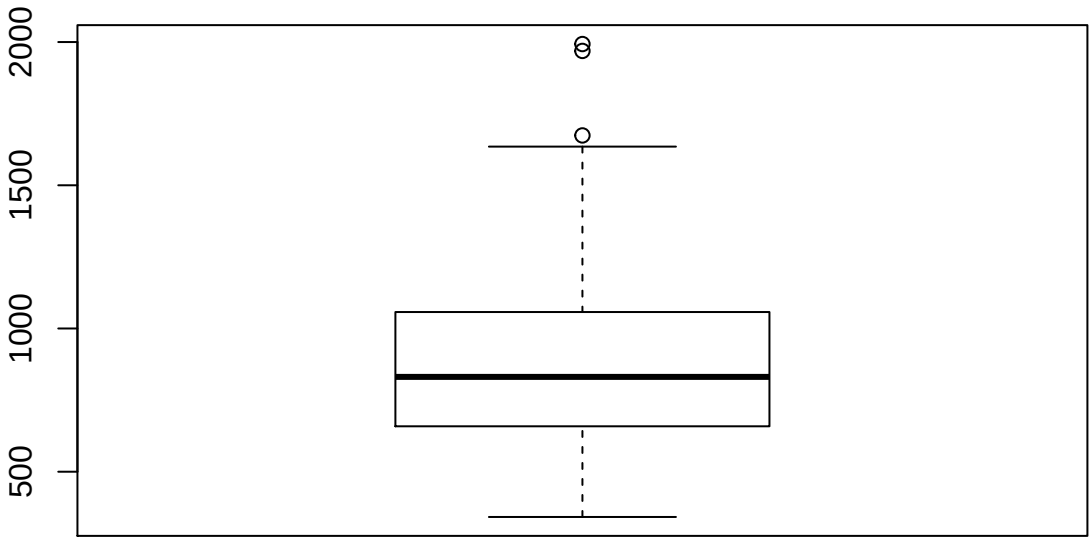
First, we pull down the data and load it...

```
dataFile5_1 <- "uscrime.txt"
if (!file.exists(dataFile5_1)) {
  crimeDataURL <- paste0(c("http://www.statsci.org/data/general/uscrime.txt"))
  download.file(crimeDataURL, dataFile5_1) }

crimeDataTable <- read.table(dataFile5_1, header = TRUE )
```

Plotting box-plot:

```
boxplot(crimeDataTable$Crime)
```



Since all the outliers on one side (i.e one tail, we just use one tailed grubb test)

Running Grubs test with one-tailed (opposite = false)

```
grubbs.test(crimeDataTable$Crime, type = 10, opposite = FALSE)
```

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

- Also the p value of 0.07887 (is very close to less than 0.05), so we can reject the null hypothesis given we're playing with 90% confidence intervals.
- And agree with the alternative hypothesis is accurate so 1993 is indeed an outlier.

the highest value of 1993 is an outlier. Let's not just take grubbs word for it; let's plot this data too!

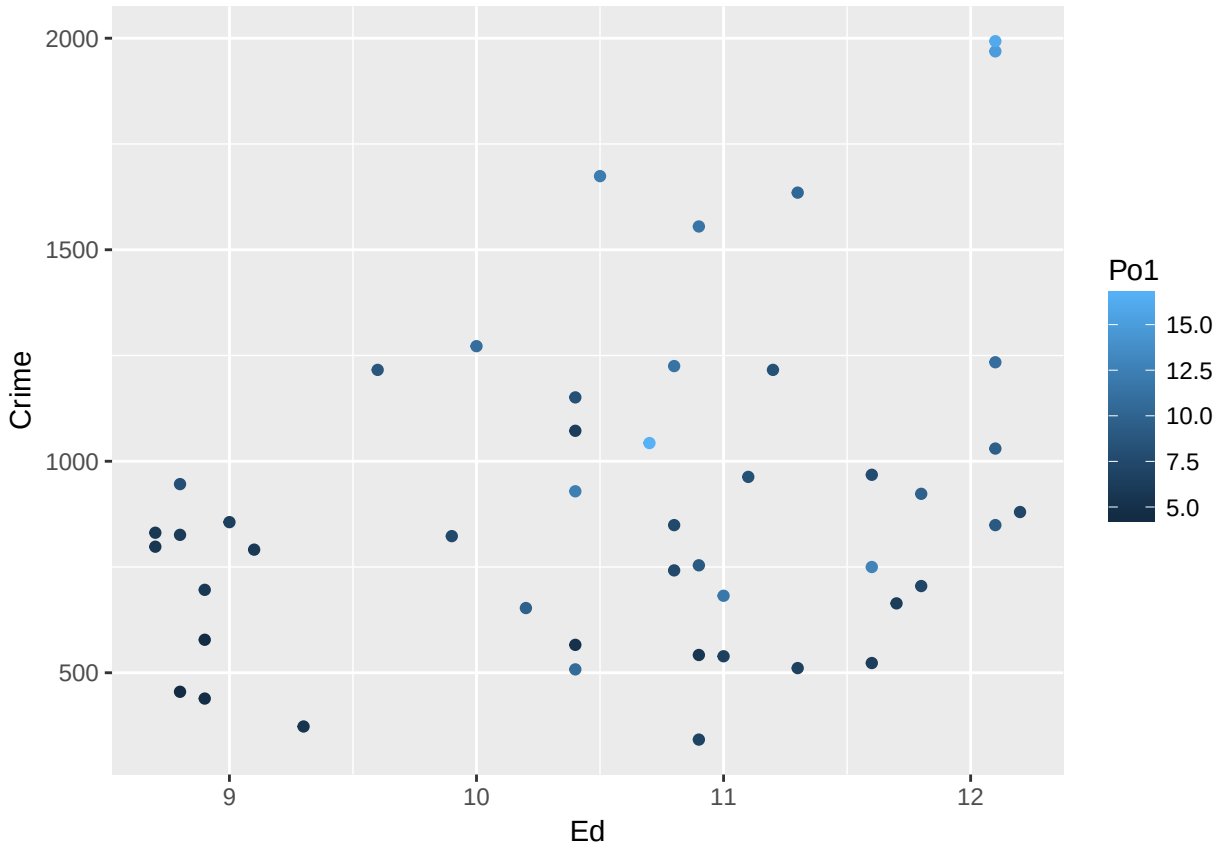
First, we need to understand the various attributes here !(http://www.statsci.org/data/general/uscrime.html)

`Variable  Description`

M percentage of males aged 14–24 in total state population So indicator variable for a southern state Ed mean years of schooling of the population aged 25 years or over Po1 per capita expenditure on police protection in 1960 Po2 per capita expenditure on police protection in 1959 LF labour force participation rate of civilian urban males in the age-group 14-24 M.F number of males per 100 females Pop state population in 1960 in hundred thousands NW percentage of nonwhites in the population U1 unemployment rate of urban males 14–24 U2 unemployment rate of urban males 35–39 Wealth wealth: median value of transferable assets or family income Ineq income inequality: percentage of families earning below half the median income Prob probability of imprisonment: ratio of number of commitments to number of offenses Time average time in months served by offenders in state prisons before their first release Crime crime rate: number of offenses per 100,000 population in 1960

This plot below shows the crime rate plotted against the state's education situation (mean years of schooling of populate >25). I colored the points based on the amount of police protection available in that state.

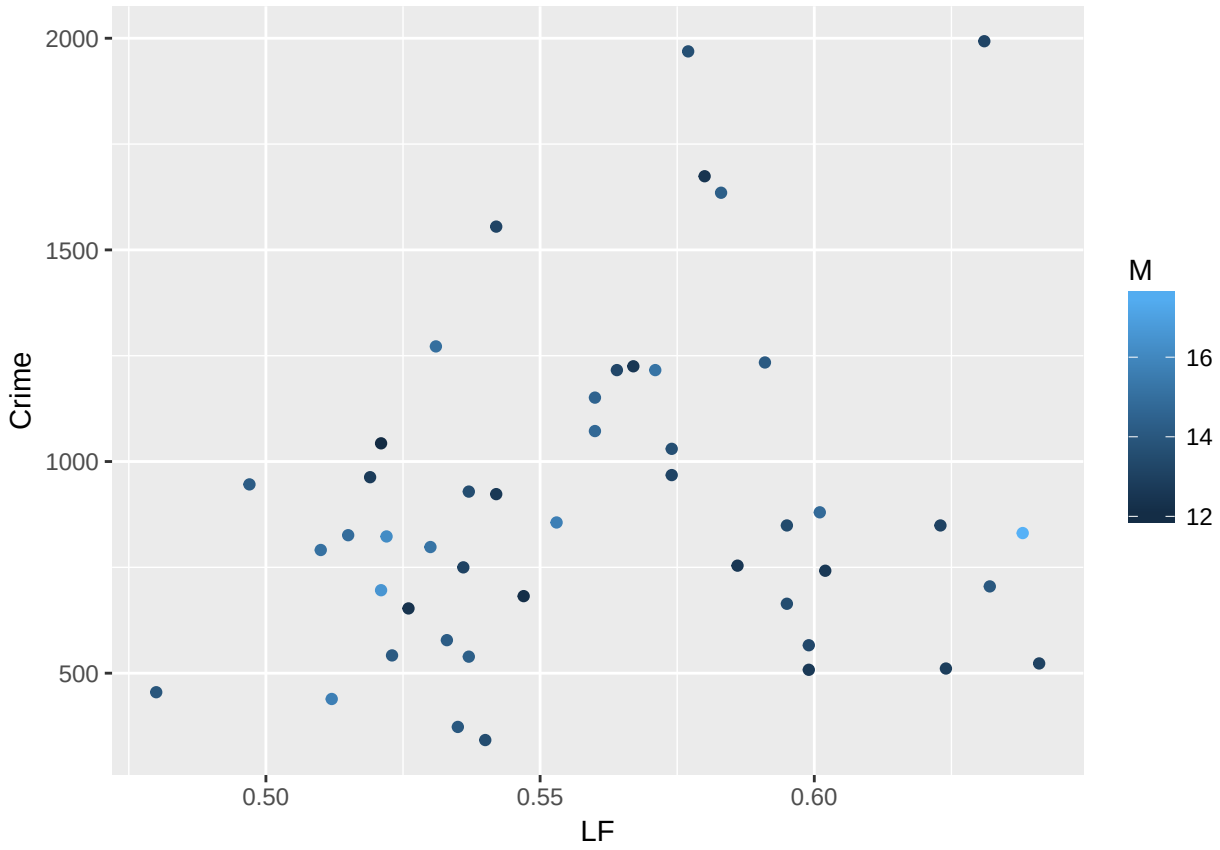
```
ggplot(data = crimeDataTable) +
  geom_point(mapping = aes(Ed, Crime, color = Po1))
```



The statistic of 1993 crimes / 100k population is the far upper right and different from the other states with high crime rate where there is lower per capita expenditure on police protection:

This plot below shows crime rate against LF (labor force participation rate in ages 14-24)

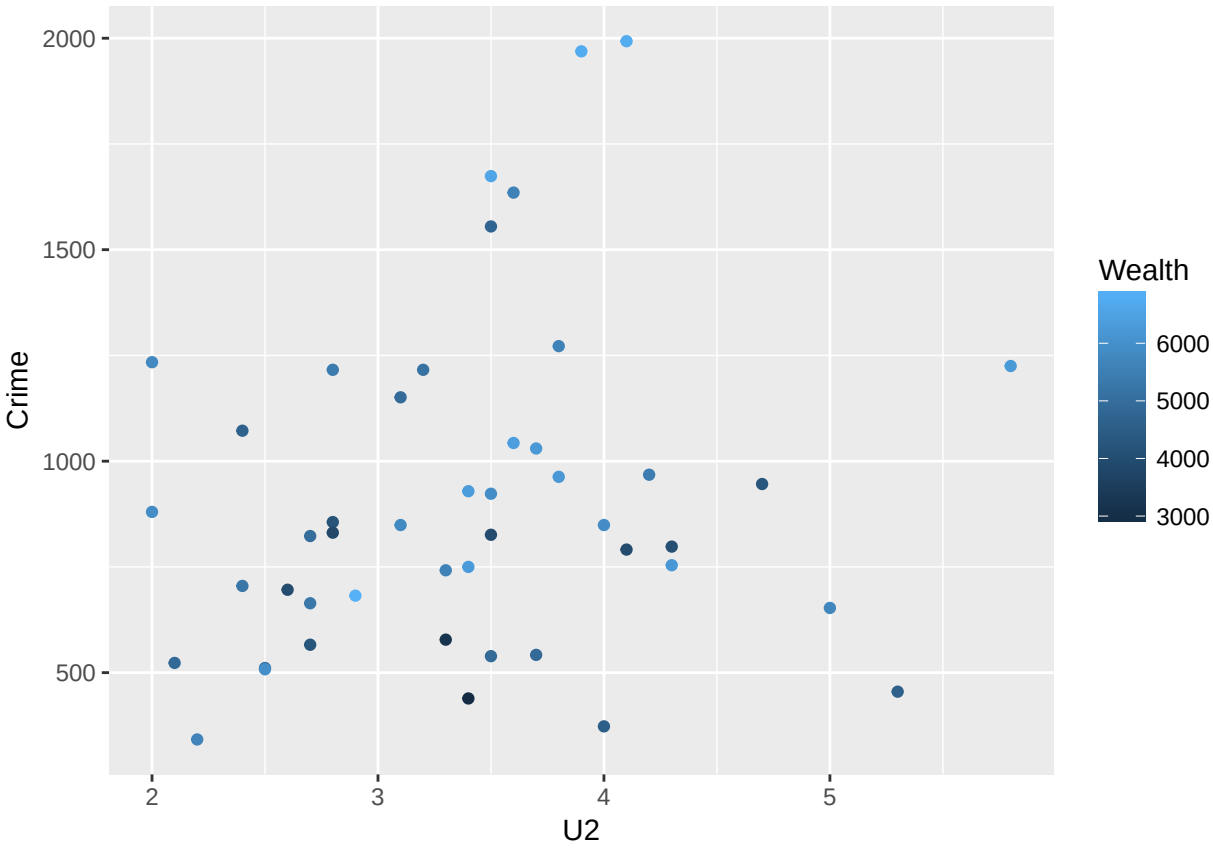
```
ggplot(data = crimeDataTable) +
  geom_point(mapping = aes(LF, Crime, color = M))
```



Again, most of the states with high youth labor force have lower crime (less than 1000/100k population). Yet, this state has the highest in the country! Another clear sign of an outlier.

Finally, this plot below shows (U2) the unemployment rate of urban males (35-39) plotted against the crime rate.

```
ggplot(data = crimeDataTable) +
  geom_point(mapping = aes(U2, Crime, color = Wealth))
```



Again as we can witness, the state with the highest , 1993/100K crime rate in in stark difference than the other states also in the 4% range of unemployment.

So, in short, Grubbs.test is fairly accurate in its estimation

---

Question6.1

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?

I am a co-founder of waada (http://waada.org). The purpose of this non profit is to help folks with mental illness using technology. The reason why CUSUM / Change Detection is so germane to this organization is that I can can use the CUSUM algorithm to detect mood changes. Unless the person has bi-polar depression , where the changes are obvious, depression in people who are prone to it creeps in gradually until its too late for the care giver to make an impact. In this situation, the "slippery slope" hits the depressed person and they stay depressed for weeks or months. Sometimes crude measures like medicine have to be taken to lift them out, but those are mostly artificial and there is no way to measure the exact quantity to be taken by the person to get better since measuring the "extent" of depression is so subjective. Therefore there is almost always a slight overdose of the medicine, which in the long term is severely adverse to the health of the patient, since he or she invariably becomes dependent on that medicine, akin to a drug addict.

The concept I have is as follows. Suppose we are able to take in physiological information (heartbeats through fitbits or iWatch wearables), phone mobility (through its gyroscope), # of calls made, length of calls, we can start creating a pattern around this person. We can also allow this person to directly enter into the phone via an app if they are feeling down or not (taking care to give something back in return, like a calming remedy , a song, breathing techniques etc so we can motivate the person to enter the data).

This data can then be used to create a "mental wellness score", at which point you have mapped the subjective "extent" of depression into objective numbers. After that, applying the CUSUM technique is simple.

Prompt identification of the onset of depression would allow us to engage in proactive measures like involving the caregiver much sooner, or providing special services through the mobile app around improving breathing techniques and a more engaged package of activities. if the onset is pretty severe, healthcare and even emergency services (suicide hotline) could be put on notice.

Specifics around C and T for this use-case

Threshold T - the threshold T is decided upon based on how costly it is to qualify that a change is detected

if the threshold is low, the cost is that the patient may not have fallen ill, and we have the cost to bear of engaging extra healthcare services and the cost of pulling in the caregiver sooner than normal.

if the threshold is too high, we may never find out until its too late. the cost can be catastrophic. This is why I'd tend to stay on the lower range of the threshold.

Dampener C - the value C depends on how costly it is to be too sensitive to change.

A low value of C will raise false alarms, if the mood is oscillating close to the border line. This may be the case if the person is taking drugs/medicine to improve their well being, and spurious random data can also cause the model to be inaccurate. The cost of false alarms is pretty much the same as a low threshold above.

If however the dampener is high, the change may get detected a little later than desired. (e.g. on Day4 when the actual depression started getting established on Day2). The costs over here, though not as catastrophic as a high Threshold value, is still high. namely the cost is that a high C renders the model useless. This is because when a person has been consistently in a depressed state for 1-2 days, its going to get much harder to pull them out of it, and the latter was the whole purpose of this initiative.

Therefore I'd err towards a mid to lower range of C.

Note: I have concatenated the answer to Question 6.2 below this Rmd. It was written in Google Docs and exported to PDF in case you are wondering why the formatting and fonts etc changed so much!