Thomas Huelsnitz
ISYE 6501
Week 2 Homework

4.1

In my current role I work for a global engineering conglomerate that deals with heavy machinery. Our customers use our machinery in different ways depending on what type of industry they are in or, within an industry, the dynamics of their local market. We call the way the customer uses their machine their "operating profile." A customer's operating profile affects what replacement parts and/or upgrades the customer will be interested in. The customer does not generally tell us up front what additional products they are interested in. It is up to us to look at their operational data and determine what group (type of operating profile) they belong to and then target our sales accordingly.

The operating profile of the machine is determined by several factors. These are what I would use as factors for a clustering model. A few of these are:

1. The ratio of hours that the machine is running in a year to total hours in a year.
2. The ratio of hours that the machine is running at full capacity in a year to total hours in a year.
3. The amount of times that the machine experiences a start-up (going from off to on) in a year.

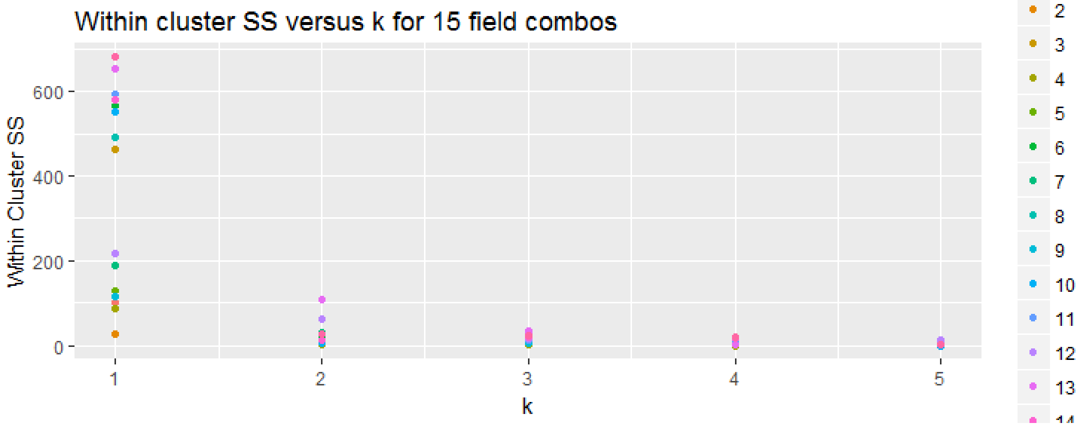
4.2

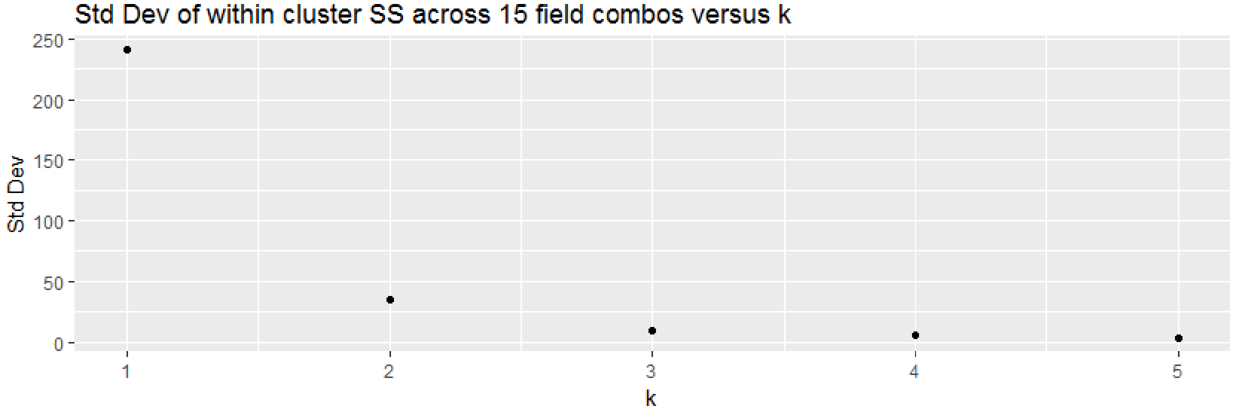
The code for this is in **HW02-4.2.R**

For this exercise, it is important to pretend that we do not know the correct labeling of the points ahead of time. Let's see if we can use a brute force approach using different transformations of input data and kmeans k values to get an idea of how many clusters are in our data set.

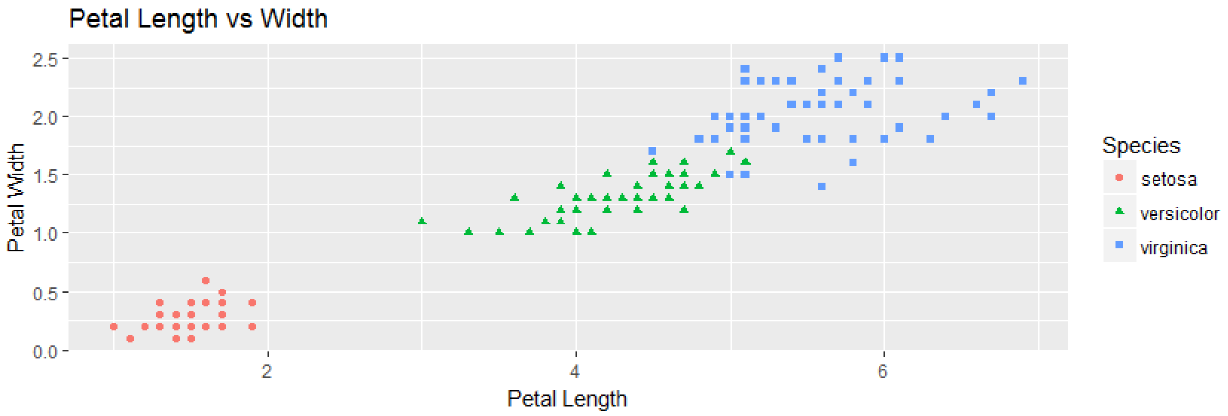
Here we iterate over each of the 15 different possible combinations of the 4 input fields and for each of those combinations we iterate for the k values 1 through 5. A kmeans model is trained for each of these unique combinations. This results in 15*5=75 different models. The performance of these models is recorded in the variable "results." "colsetlist" is a length-75 list where each item is the set of columns used in the corresponding entry in "results." "cluster" is a length-75 list where each item is a vector indicating which cluster each of the data points is assigned to for the corresponding row in "results."

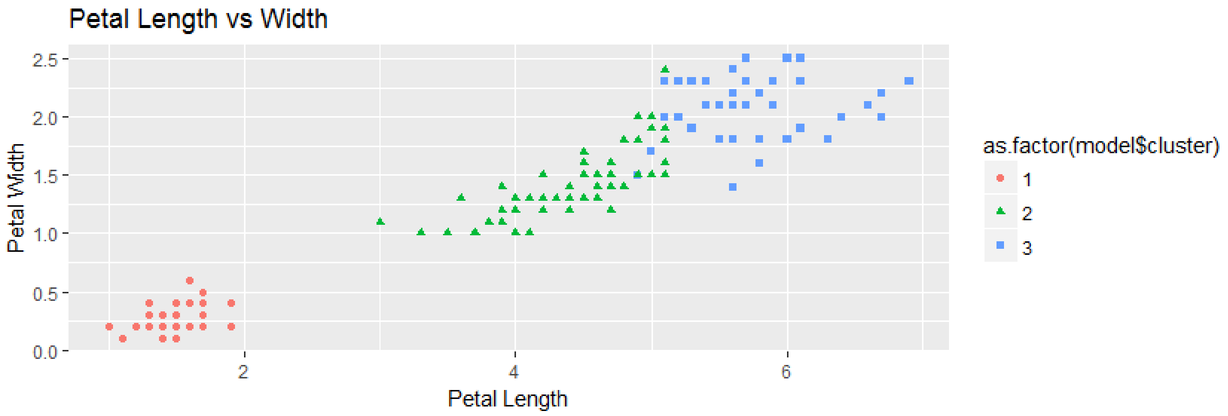
Here we plot some of the data from the "results" dataframe. First is a plot of the within cluster sum of squares for each unique combination of k-value and column combination. If we group this data by each k value and compute the variance we get the bottom plot which has the characteristic elbow at k=3. This indicates that k=3 is a good choice for our model.





Now we will fit a model with k=3 to the full dataset and do a visual comparison of how accurately it labels the data:



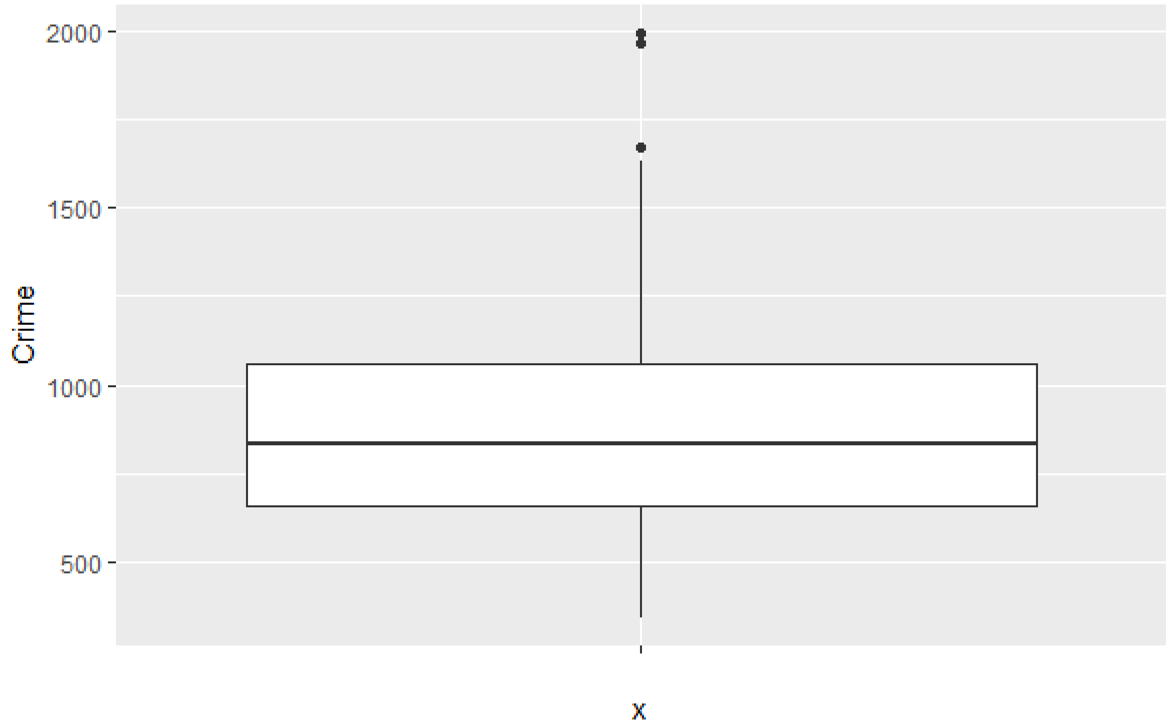


The model shown above has an accuracy of 89.3%

<u>5.1</u>

The code for this is in **HW02-5.1.R**

Here we start by plotting the data on a box and whisker plot:



Running grubbs.test on this data set gives following:

```
	Grubbs test for one outlier

data:  crime[, "Crime"]
G = 2.81290, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier
```

According to this, the null hypothesis is that the highest value is *not* an outlier and the p-value for this hypothesis is .07887. This is a ‘low’ p-value but the choice of what constitutes a good p-value depends on context. I cannot look at this data set and just arbitrarily choose a p-value threshold for determining whether to accept or reject the null hypothesis. In some situations, it might be tempting to throw this data point away. For instance, if these were sensor measurements of some physical system, I would probably discard this point. In this instance however, I do not think that this point should be discarded. This data set is widely used and came about as part of an in-depth investigation into crime statistics in the late 70s. It is unlikely that this potential outlier value is a mistake. On the contrary, it is probably a data point that we are very interested in.

6.1

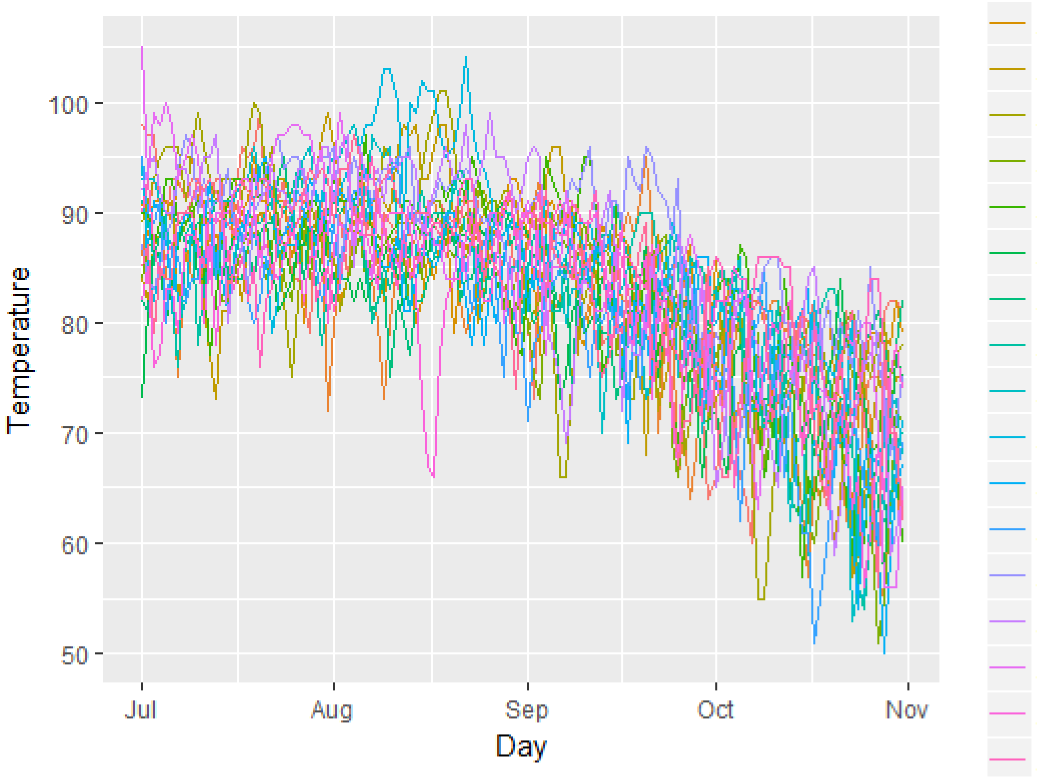
A change detection model would be appropriate for many situations in a hospital. One particular use would be to track a patient’s blood oxygen level (pulse ox). Any amount greater than 93% of maximum saturation is acceptable. As pulse ox trends below 93%, the blood does not have enough oxygen in it for the body to function properly. This can be caused by acute and chronic issues. You would want an algorithm that would trip an alarm rapidly if the pulse ox dropped substantially below 93%. You would want it to also trip an alarm if the pulse ox was only a bit below 93% but had been there for a long period of time. It would also need to be robust against a rapid drop that is only a small amount below 93% (likely due to a sensor error). This behavior would be achieved by a careful choice of a critical value and a threshold.

6.2

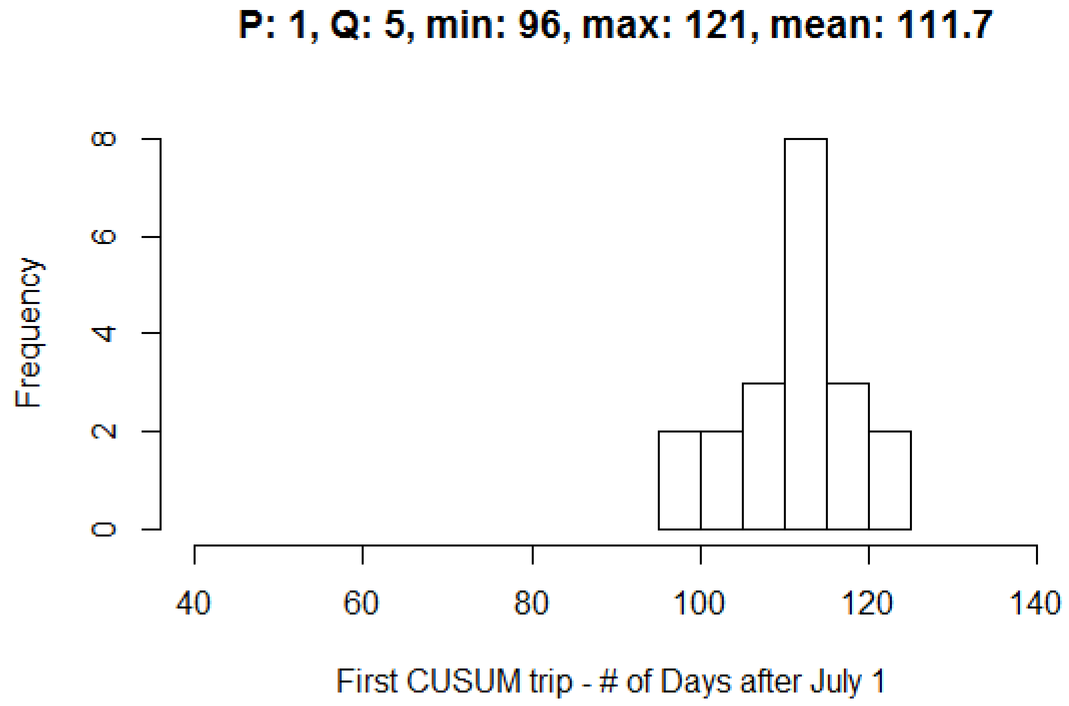
The code for this is in **HW02-6.2.R**

1)

Starting with a plot of temperature versus day for all years in the data set – this is a bit of a mess.



We want to use CUSUM to determine when the temperature typically switches from summer to fall (unofficial summer end). In order to do this I created a function which implements the CUSUM procedure shown in the lectures. I set it so that the threshold is 5 times the standard deviation of the dataset (P on the graph/ in the function) and the C value shown in the lecture is 1 times the standard deviation (Q on the graph/ in the function). This was applied to each year in the dataset and the first day of surpassing the threshold was recorded. This is shown below in a histogram and a time series plot.

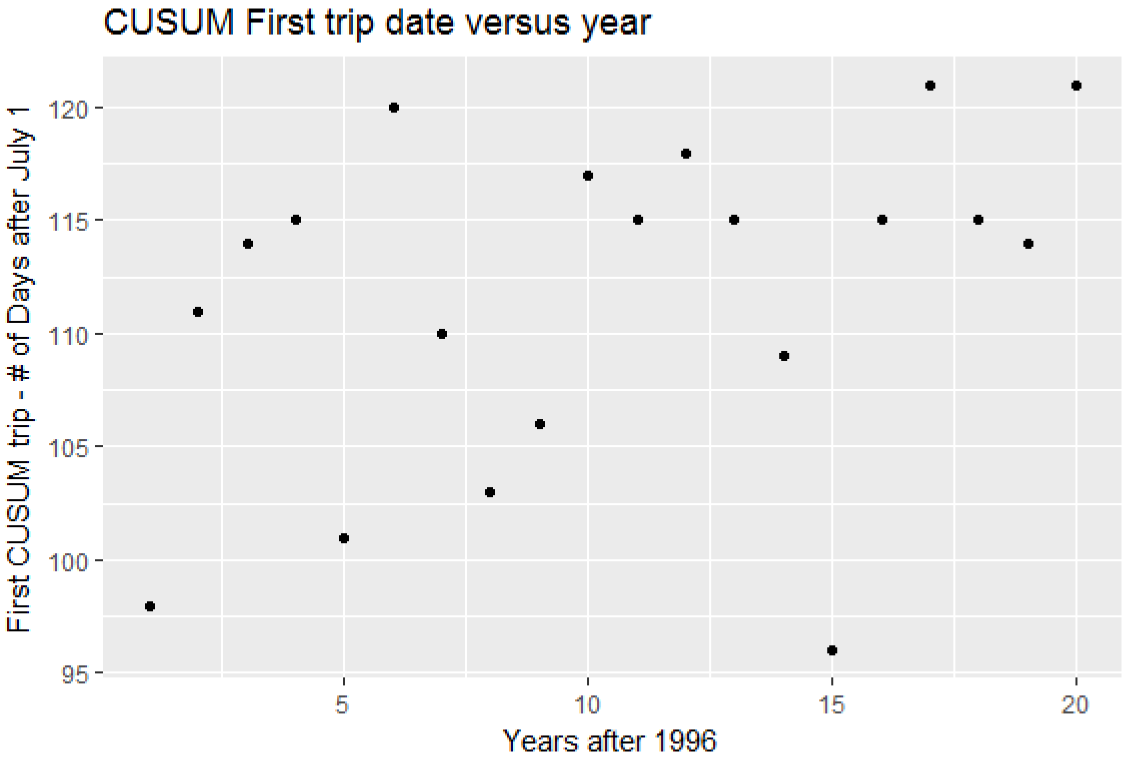


From this it looks like the weather makes a notable shift, on average, 110 days after July 1st. This could be considered the unofficial summer end.

2)

Note from the timeseries plot that this seems to be happening later in the year as the years progress.

Thomas Huelsnitz
ISYE 6501
Week 2 Homework



It seems like Atlanta’s summer climate has gotten warmer in this time since the transition from summer to fall is happening later in the year.