Week 10 Mini Project Clustering

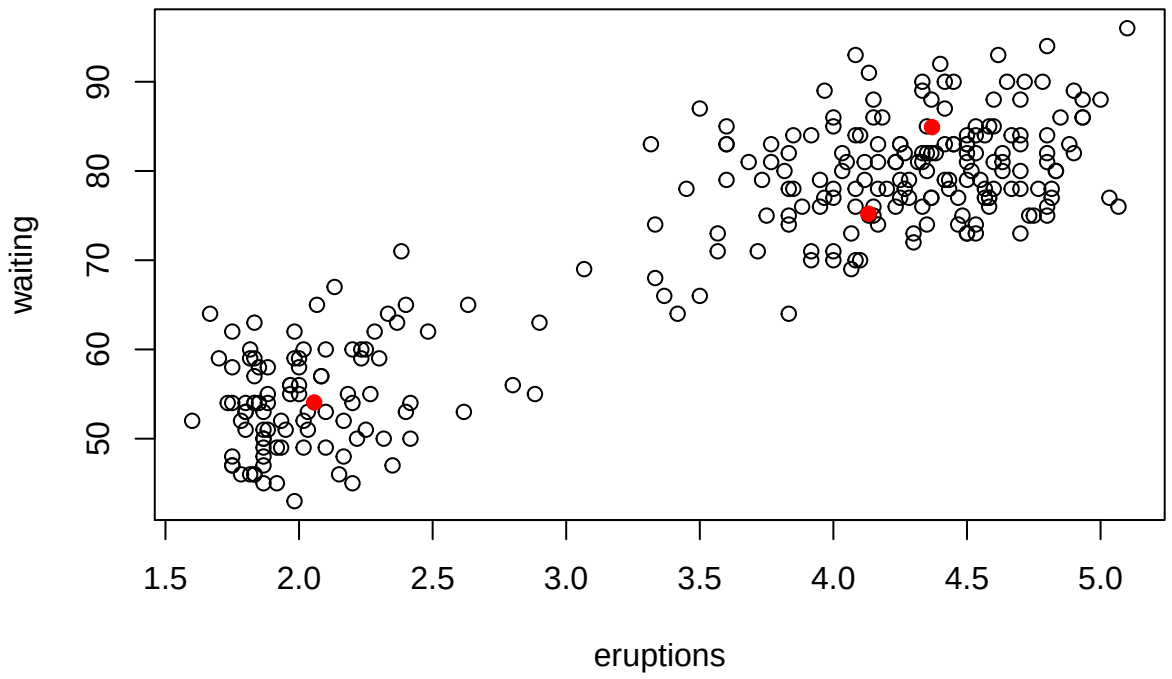
Ben Arancibia

10/31/2015

This project will focus on taking data from the R package cluster.datasets nd performing a cluster analysis on the data. The first thing to do is setup the appropriate environment using the following.

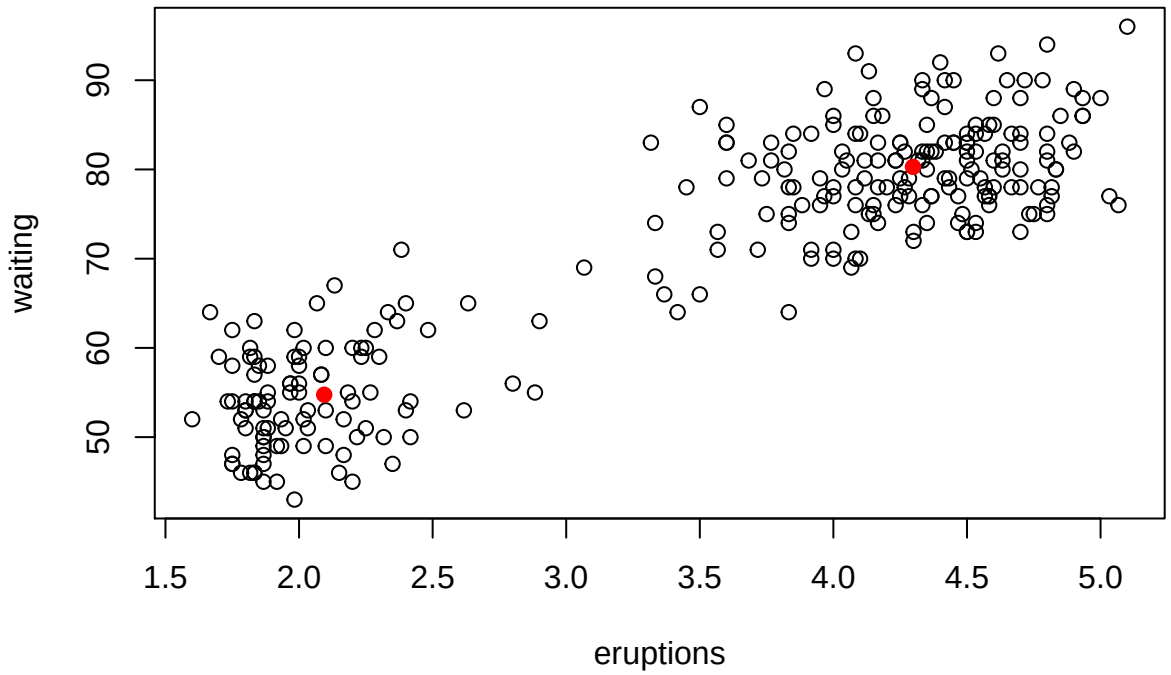
I am going to cluster waiting time between eruptions and the duration of eruptions for the Old Faithful geyser in Yellowstone. This dataset faithful is part of base R.

```
data <- faithful

kmeans.data<- kmeans(data, 3)

plot(data)
points(kmeans.data$centers, pch=19, col="red")
```



Looking at the initial datasets there are two clusters.

```
kmeans.data<- kmeans(data, 2)

plot(data)
points(kmeans.data$centers, pch=19, col="red")
```



Next do this in Spark.

```
df <- createDataFrame(sqlContext, faithful)
df
```

```
## DataFrame[eruptions:double, waiting:double]
```

```
getScaled <- function(data, column, min, max){
    if(max != min) eval(parse(text=paste("data$",column," <- (data$", column, "-",min,")/(",max,"-",min
    return (data)
}

km_scale <- function(data, numericVars){
    scales <- c()
    for(var in numericVars){
        min <- getMin(data, var)
        max <- getMax(data, var)
        data <- getScaled(data, var, min, max)
        scales <- c(scales, eval(parse(text=paste("c('",var, "_min' = min, '",var,"_max' = max)", sep="'
    }
    return(list("data"= data, "scales"=scales))
}

getClusters <- function(data, k){
    data$cluster <- cast(data[[1]]*0, 'integer')
    for(i in 1:k){
        data$temp_cluster <- cast(data[[1]]*0, 'integer')
        for(j in 1:k){
            if(i < j){
                eval(parse(text=paste("data$temp_cluster <-
                        data$temp_cluster + cast(data$dist_",i," <= data$dist_",j,", 'integer')", sep=""
            }else if(j < i){
```

```
                eval(parse(text=paste("data$temp_cluster <-
                       data$temp_cluster + cast(data$dist_",i," < data$dist_",j,", 'integer')", sep="")
            }
        }
        eval(parse(text=paste("data$cluster <- data$cluster + cast(data$temp_cluster == ",(k-1),", 'int
        data <- removeColumns(data, c("temp_cluster"))
    }
    data <- setType(data, c("cluster"), "integer")
    return(data)
}
```