k-means Clustering of the Old Faithful Dataset

Jon Kinsey

Wed Dec 3 17:40:12 2014

```
# Old Faithful Geyser Data
# Waiting time between eruptions and the duration of the eruption for the
# Old Faithful geyser in Yellowstone National Park, Wyoming, USA.
# The dataset contains 272 observations on 2 variables.
# [,1]  eruptions   numeric Eruption time in mins
# [,2]  waiting numeric Waiting time to next eruption (in mins)
#
names(faithful)
```

```
## [1] "eruptions" "waiting"
```

```
str(faithful) # structure
```

```
## 'data.frame':    272 obs. of  2 variables:
##  $ eruptions: num  3.6 1.8 3.33 2.28 4.53 ...
##  $ waiting  : num  79 54 74 62 85 55 88 85 51 85 ...
```

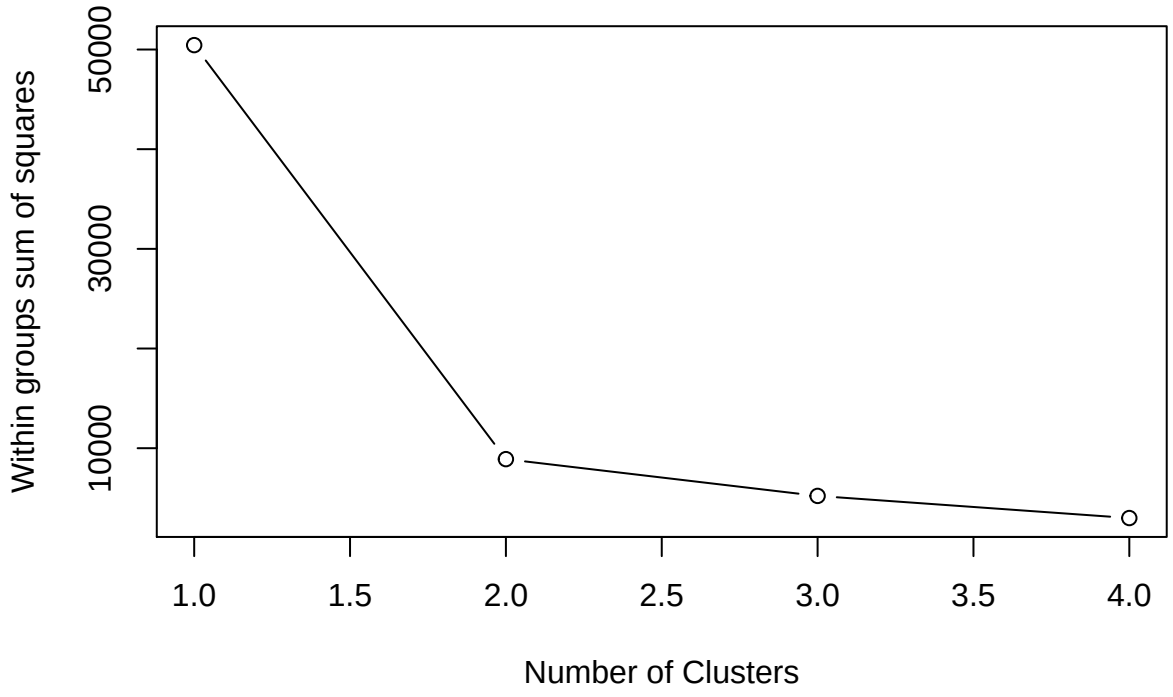
```
attributes(faithful) # attributes
```

```
## $names
## [1] "eruptions" "waiting"
##
## $row.names
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35]  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69]  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102
## [103] 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
## [120] 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136
## [137] 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153
## [154] 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170
## [171] 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187
## [188] 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204
## [205] 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221
## [222] 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238
## [239] 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255
## [256] 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272
##
## $class
## [1] "data.frame"
```

```
summary(faithful)
```

```
##    eruptions       waiting
##  Min.   :1.60   Min.   :43.0
##  1st Qu.:2.16   1st Qu.:58.0
##  Median :4.00   Median :76.0
##  Mean   :3.49   Mean   :70.9
##  3rd Qu.:4.45   3rd Qu.:82.0
##  Max.   :5.10   Max.   :96.0
```

```
#
ssPlot <- function(data, maxCluster = 4) {
  # Initialize within sum of squares
  SSw <- (nrow(data) - 1) * sum(apply(data, 2, var))
  SSw <- vector()
  for (i in 1:maxCluster) {
    SSw[i] <- sum(kmeans(data, centers = i)$withinss)
  }
  plot(1:maxCluster, SSw, type = "b", xlab = "Number of Clusters", ylab = "Within groups sum of squares")
}
ssPlot(faithful)
```




```
#
(kf <- kmeans(faithful, 2))
```

```
## K-means clustering with 2 clusters of sizes 100, 172
##
## Cluster means:
##   eruptions waiting
## 1     2.094   54.75
## 2     4.298   80.28
##
## Clustering vector:
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   2   1   2   1   2   1   2   2   1   2   1   2   2   1   2   1   1   2
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   1   2   1   1   2   2   2   2   1   2   2   2   2   2   1   2   2   1
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   1   2   1   2   2   1   2   1   2   2   1   1   2   1   2   2   1   2
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   1   2   2   1   2   2   1   2   1   2   1   2   2   2   1   2   2   1
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##   2   2   1   2   1   2   2   2   2   2   2   1   2   2   2   2   1   2
##  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##   1   2   1   2   1   2   2   2   1   2   1   2   1   2   2   1   2   1
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
##   2   2   2   1   2   2   1   2   1   2   1   2   1   2   2   1   2   2
## 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
##   1   2   1   2   1   2   1   2   1   2   1   2   1   2   2   1   2   2
## 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
##   2   1   2   1   2   1   2   2   1   2   2   2   2   2   1   2   1   2
## 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   1   2   1   2   1   2   1   2   1   1   2   2   2   2   2   1   2   2
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
##   1   2   2   2   1   2   2   1   2   1   2   1   2   2   2   2   2   2
## 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216
##   1   2   1   2   2   1   2   1   2   2   1   2   2   2   1   2   1   2
## 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
##   1   2   1   2   1   2   1   2   2   2   2   2   2   2   2   1   2   1
## 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252
##   2   1   1   2   2   1   2   1   2   1   2   2   1   2   1   2   1   2
## 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
##   2   2   2   2   2   2   1   2   2   2   1   2   1   1   2   2   1   2
## 271 272
##   1   2
##
## Within cluster sum of squares by cluster:
## [1] 3456 5446
##  (between_SS / total_SS =  82.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
#
# Compare the eruptions label with the clustering result
table(faithful$eruptions, kf$cluster)
```

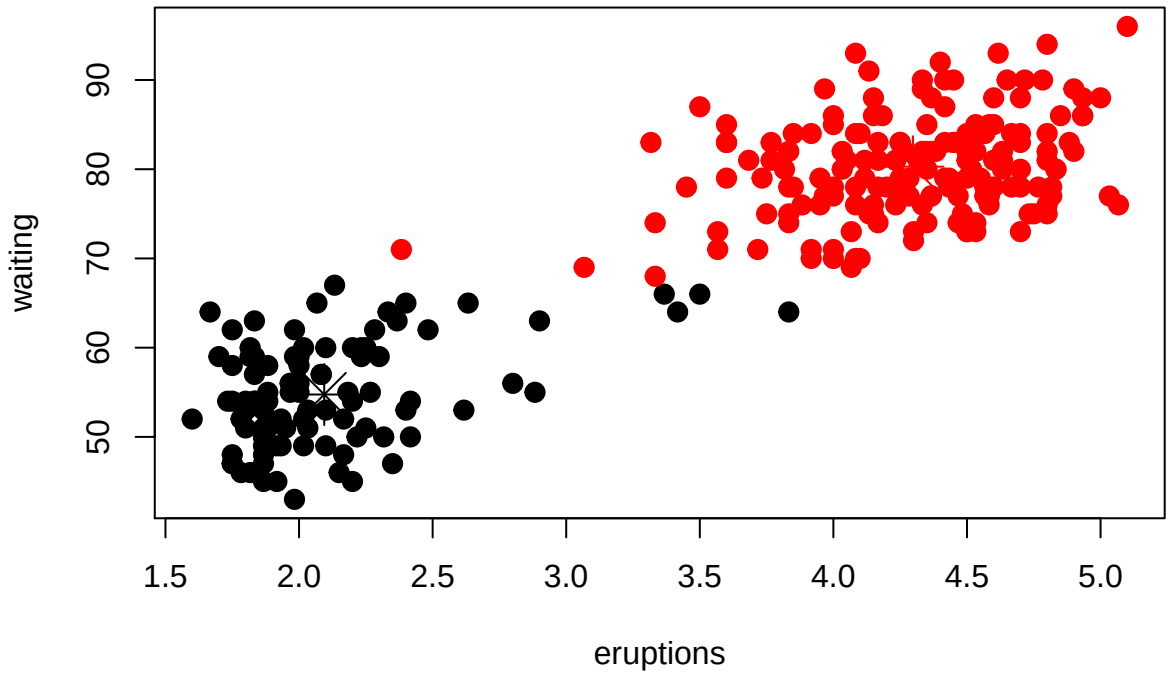
```
##
##         1 2
##   1.6   1 0
##   1.667 1 0
##   1.7   1 0
##   1.733 1 0
##   1.75  6 0
```

```
##   1.783 2 0
##   1.8   4 0
##   1.817 3 0
##   1.833 7 0
##   1.85  2 0
##   1.867 8 0
##   1.883 4 0
##   1.917 2 0
##   1.933 2 0
##   1.95  1 0
##   1.967 3 0
##   1.983 3 0
##   2     4 0
##   2.017 3 0
##   2.033 2 0
##   2.067 1 0
##   2.083 2 0
##   2.1   3 0
##   2.133 1 0
##   2.15  1 0
##   2.167 2 0
##   2.183 1 0
##   2.2   3 0
##   2.217 1 0
##   2.233 2 0
##   2.25  2 0
##   2.267 1 0
##   2.283 1 0
##   2.3   1 0
##   2.317 1 0
##   2.333 1 0
##   2.35  1 0
##   2.367 1 0
##   2.383 0 1
##   2.4   2 0
##   2.417 2 0
##   2.483 1 0
##   2.617 1 0
##   2.633 1 0
##   2.8   1 0
##   2.883 1 0
##   2.9   1 0
##   3.067 0 1
##   3.317 0 1
##   3.333 0 2
##   3.367 1 0
##   3.417 1 0
##   3.45  0 1
##   3.5   1 1
##   3.567 0 2
##   3.6   0 4
##   3.683 0 1
##   3.717 0 1
##   3.733 0 1
```

```
##   3.75  0 1
##   3.767 0 2
##   3.817 0 1
##   3.833 1 4
##   3.85  0 2
##   3.883 0 1
##   3.917 0 3
##   3.95  0 2
##   3.966 0 1
##   3.967 0 1
##   4     0 6
##   4.033 0 2
##   4.05  0 1
##   4.067 0 2
##   4.083 0 5
##   4.1   0 2
##   4.117 0 2
##   4.133 0 2
##   4.15  0 4
##   4.167 0 4
##   4.183 0 1
##   4.2   0 1
##   4.233 0 3
##   4.25  0 4
##   4.267 0 2
##   4.283 0 2
##   4.3   0 2
##   4.317 0 1
##   4.333 0 5
##   4.35  0 4
##   4.366 0 1
##   4.367 0 3
##   4.383 0 1
##   4.4   0 1
##   4.417 0 4
##   4.433 0 2
##   4.45  0 3
##   4.467 0 2
##   4.483 0 1
##   4.5   0 8
##   4.517 0 1
##   4.533 0 5
##   4.55  0 1
##   4.567 0 3
##   4.583 0 4
##   4.6   0 4
##   4.617 0 1
##   4.633 0 3
##   4.65  0 1
##   4.667 0 2
##   4.7   0 6
##   4.716 0 1
##   4.733 0 1
##   4.75  0 1
```

```
##   4.767 0 1
##   4.783 0 1
##   4.8   0 6
##   4.817 0 2
##   4.833 0 2
##   4.85  0 1
##   4.883 0 1
##   4.9   0 2
##   4.933 0 3
##   5     0 1
##   5.033 0 1
##   5.067 0 1
##   5.1   0 1
```

```
#
par(mfrow=c(1,1)) # make sure only 1 plot per page
# plot of waiting time between eruptions and the duration of the eruption
plot(faithful[c("eruptions", "waiting")], col=kf$cluster, pch=20,cex=2)
points(kf$centers[,c("eruptions", "waiting")], col=1:3, pch=8, cex=3)
```




```
#
# cluster centers:
kf$centers
```

```
##   eruptions waiting
## 1     2.094   54.75
## 2     4.298   80.28
```