Pokemon-K-means.R

danielpeslherbe

2020-08-04

```
##Input data
Master <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/Pokemon KMeans/pokedex_(Update_05.20

##filter unimportant data
Df <- Master[,-c((2:6),8)]

##build dataset for k means clustering only
dataset <- Df[,-c((1:5),(8:11),(19:45))]

##install.packages("VIM")
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

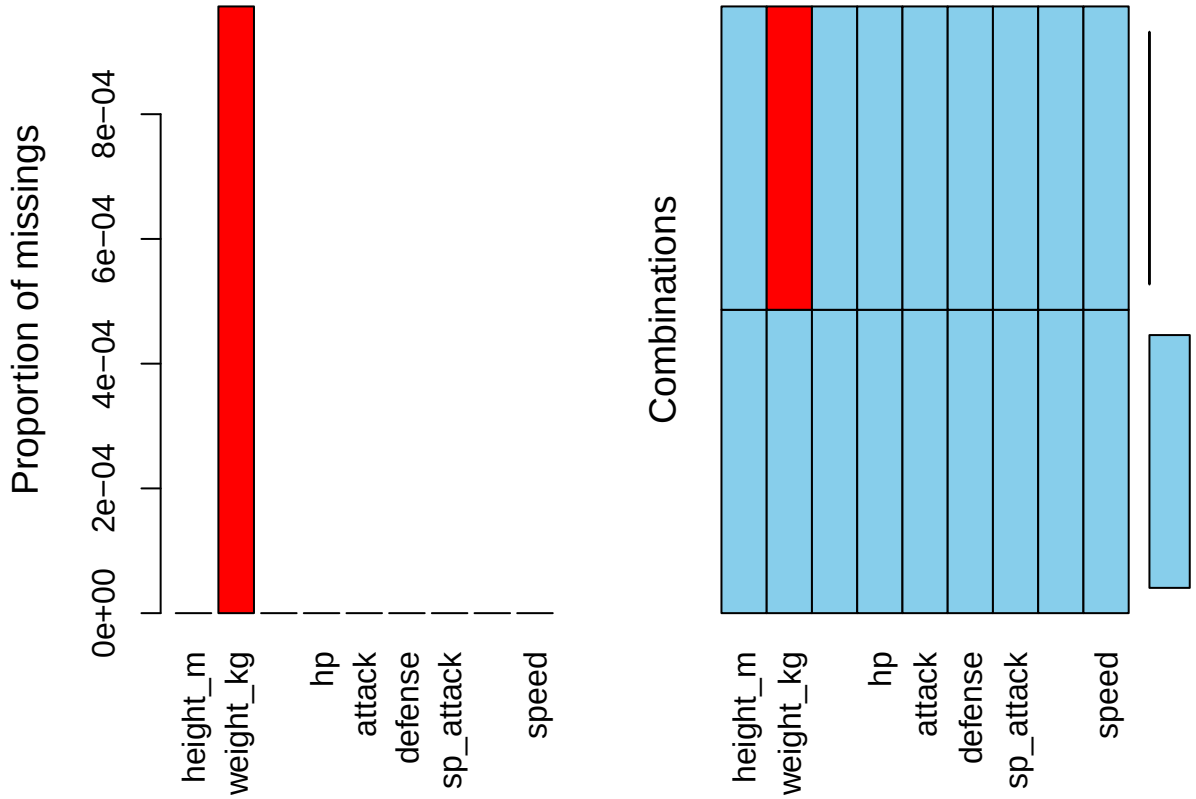
```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

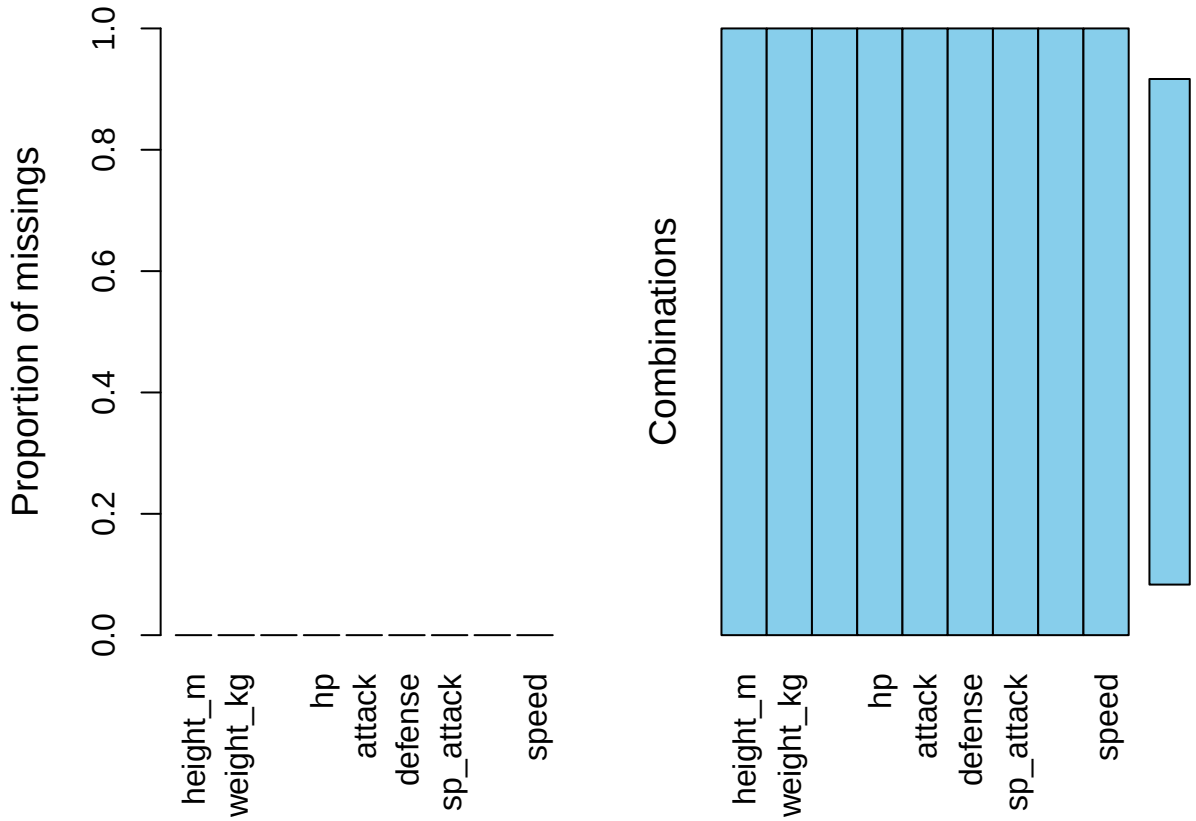
```
##install.packages("cluster")
library(cluster)
##install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.84 loaded
```

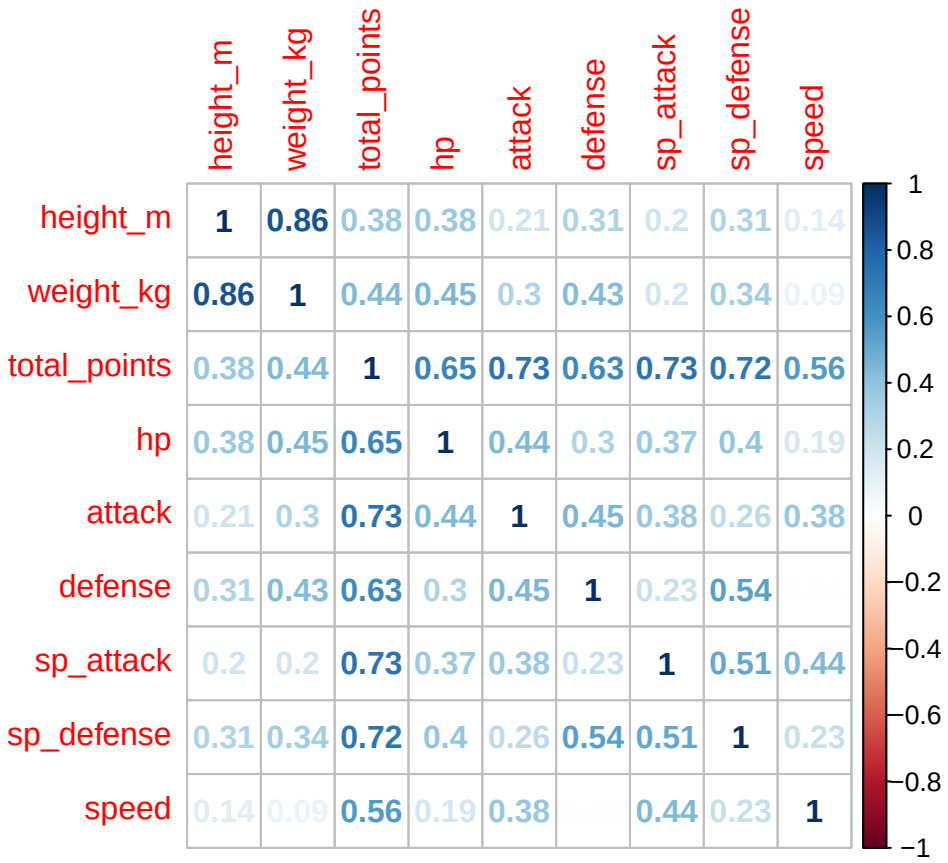
```
aggr(dataset)
```



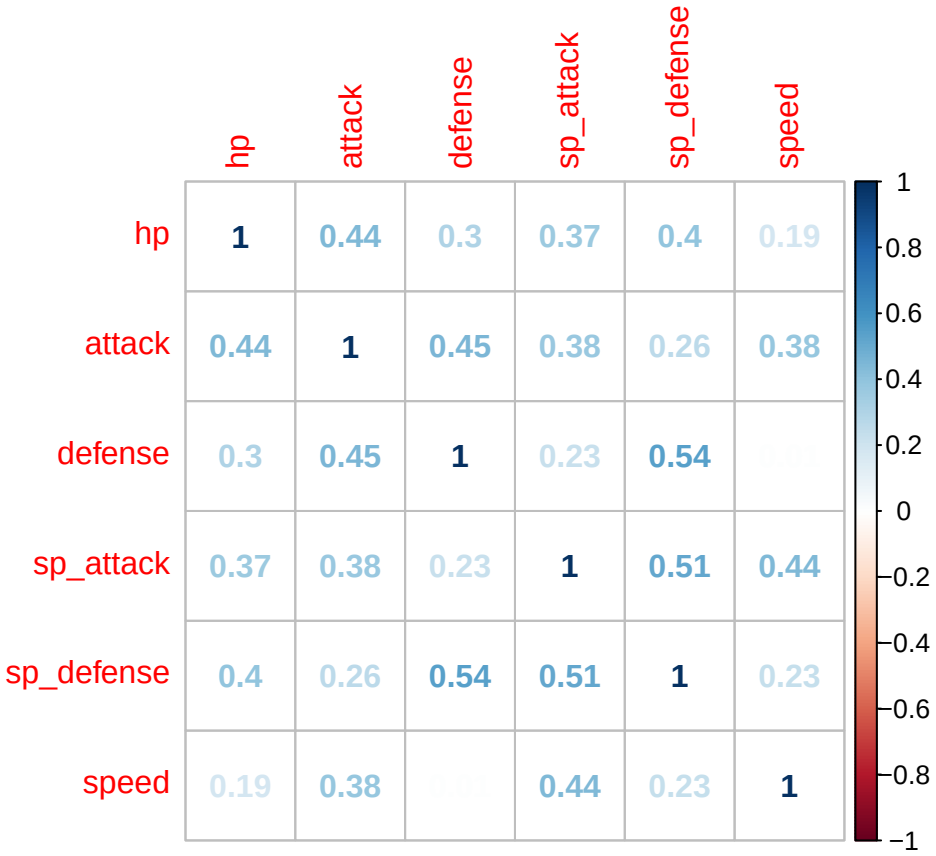

```
##Note that no weight is recorded for observation 1028 - or
##Eternamax Eternatus - thus we will discard it from the dataset
dataset[1028,2] <- dataset[1028,1]*dataset[1027,2]/dataset[1027,1]
aggr(dataset)
```



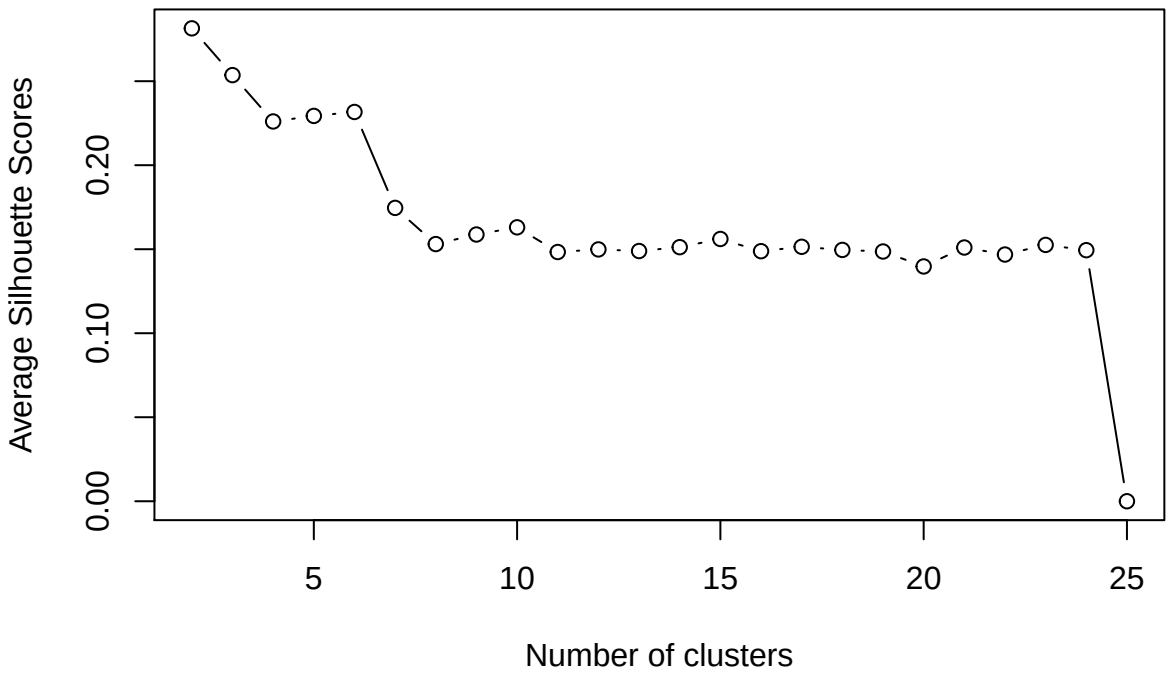

```
##Let us look at the correlation matrix for dataset
corrmatrix <- cor(dataset)
corrplot(corrmatrix, method = 'number')
```




```
##Note since height and weight are heavily correlated
##we will only used height (since weight is correlated and
##is also simulated for observation 1028)
##Note that we also exclude total points (since it is correlated to attack, defense, etc..)

dataset <- dataset[,-c(1,2,3)]

##Let us look at the correlation matrix for dataset
corrmatrix <- cor(dataset)
corrplot(corrmatrix, method = 'number')
```




```
##Scaling the dataset
dataset <- scale(dataset, center = TRUE, scale = TRUE)

set.seed(5)
##let us define the optimal k clusters through silhouette method
silhouettescore <- function(k){
  km <- kmeans(dataset, centers = k)
  ss <- silhouette(km$cluster, dist(dataset))
  mean(ss[,3])
}
k <- 2:25
avgss <- c(rep(0,24))
for (i in min(k):length(k)) {
  avgss[i-1] <- silhouettescore(i)
}
plot(k, type = 'b', avgss, xlab = 'Number of clusters', ylab = 'Average Silhouette Scores')
```




```
optk <- which.max(avgss)+1
optk
```

```
## [1] 2
```

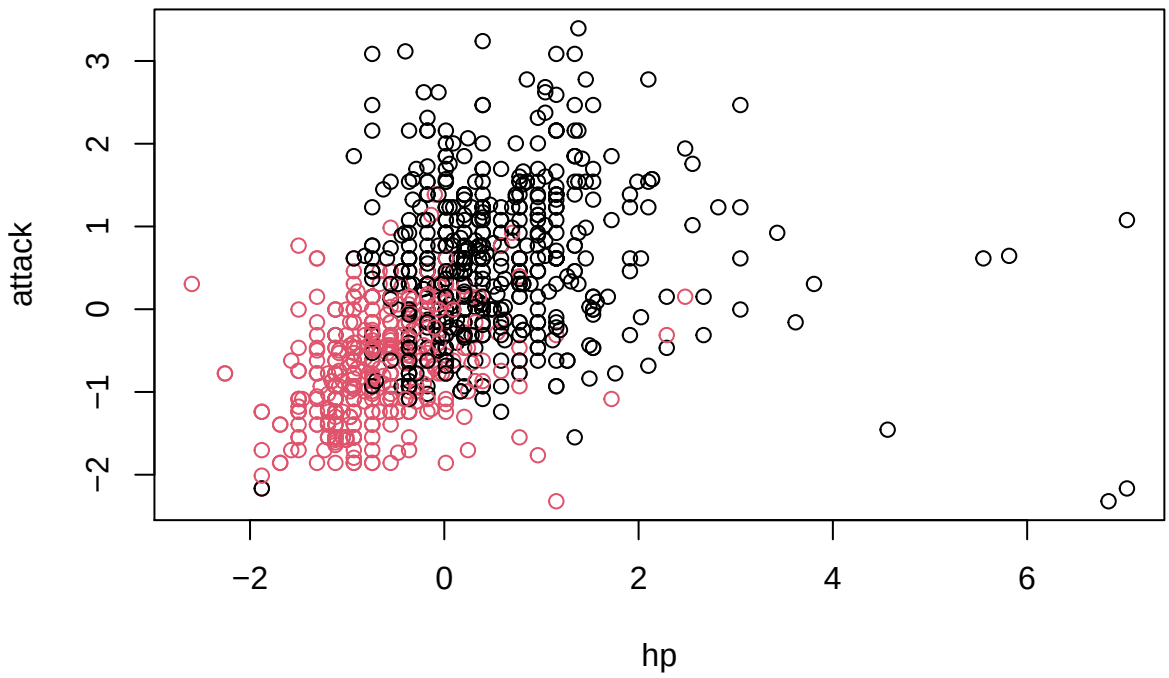
```
kmeansmodel <- kmeans(dataset, optk)
kmeansmodel$tot.withinss
```

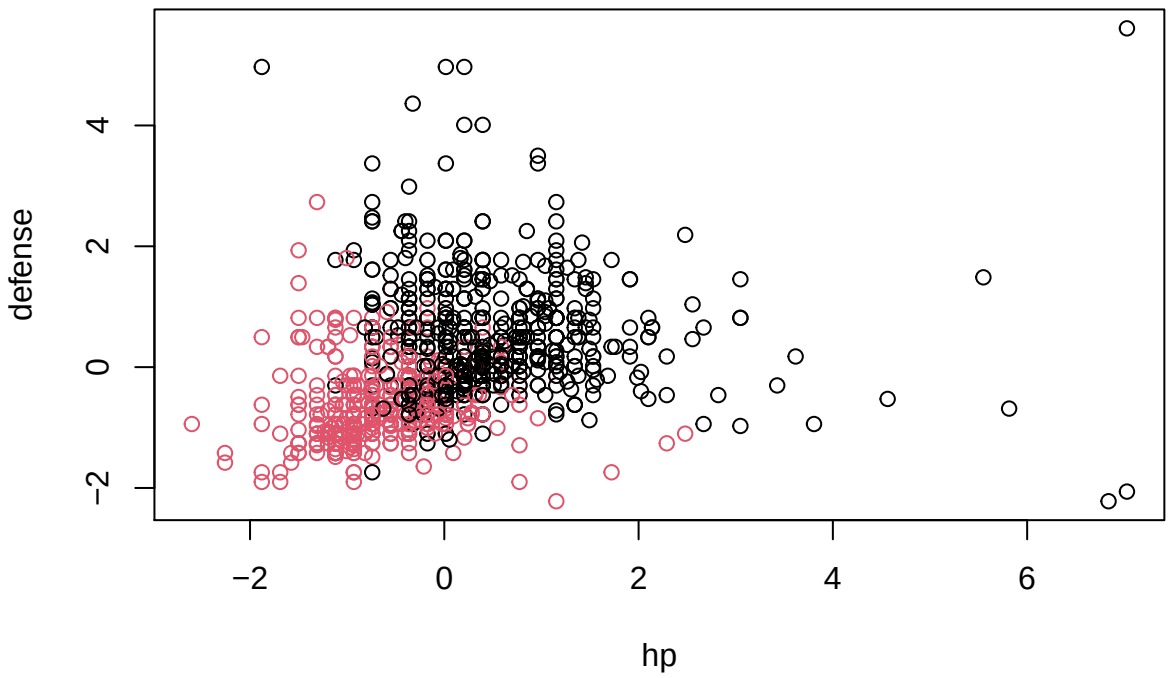
```
## [1] 4243.902
```

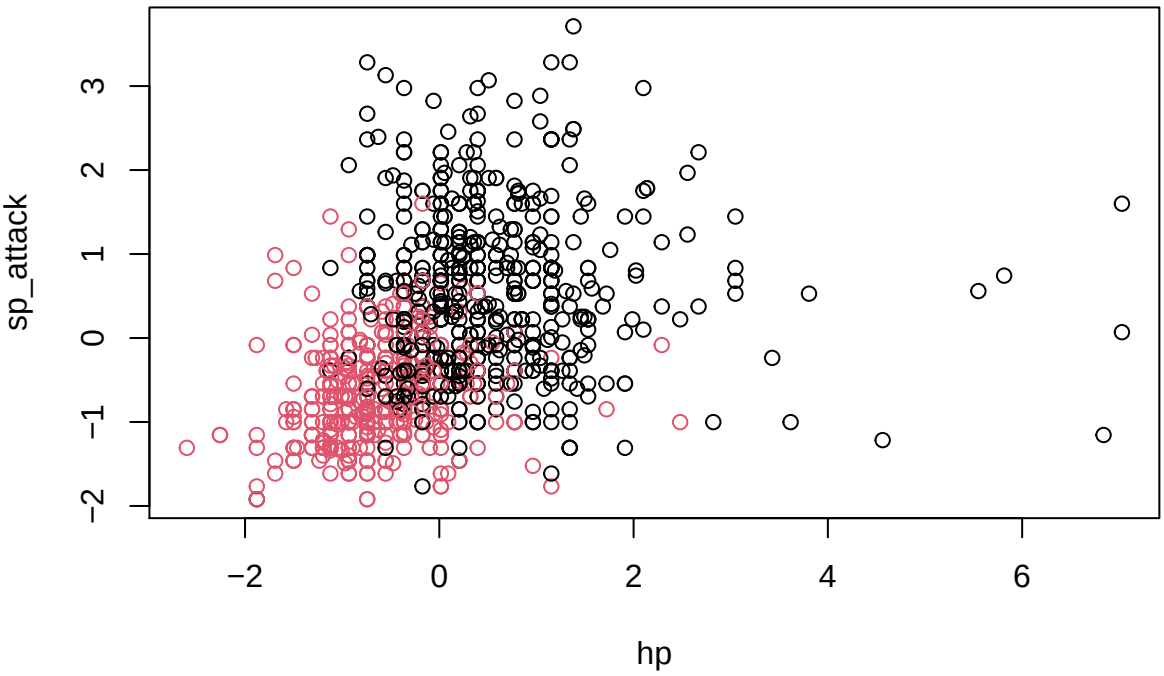
```
kmeansmodel$size
```

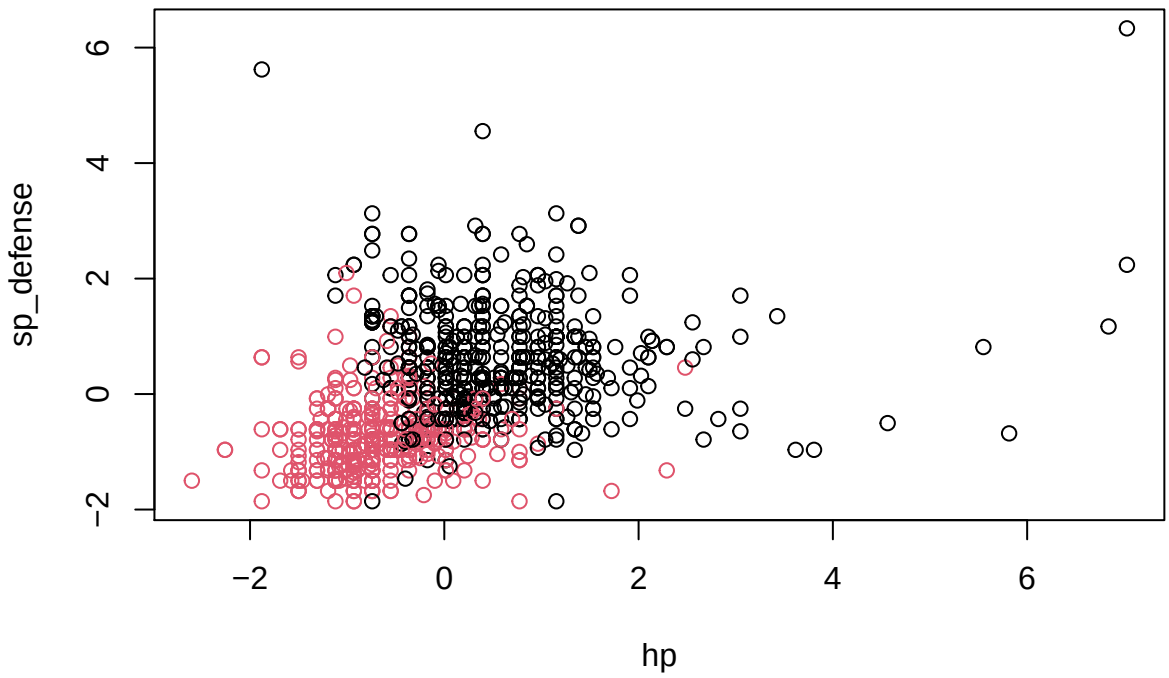
```
## [1] 574 454
```

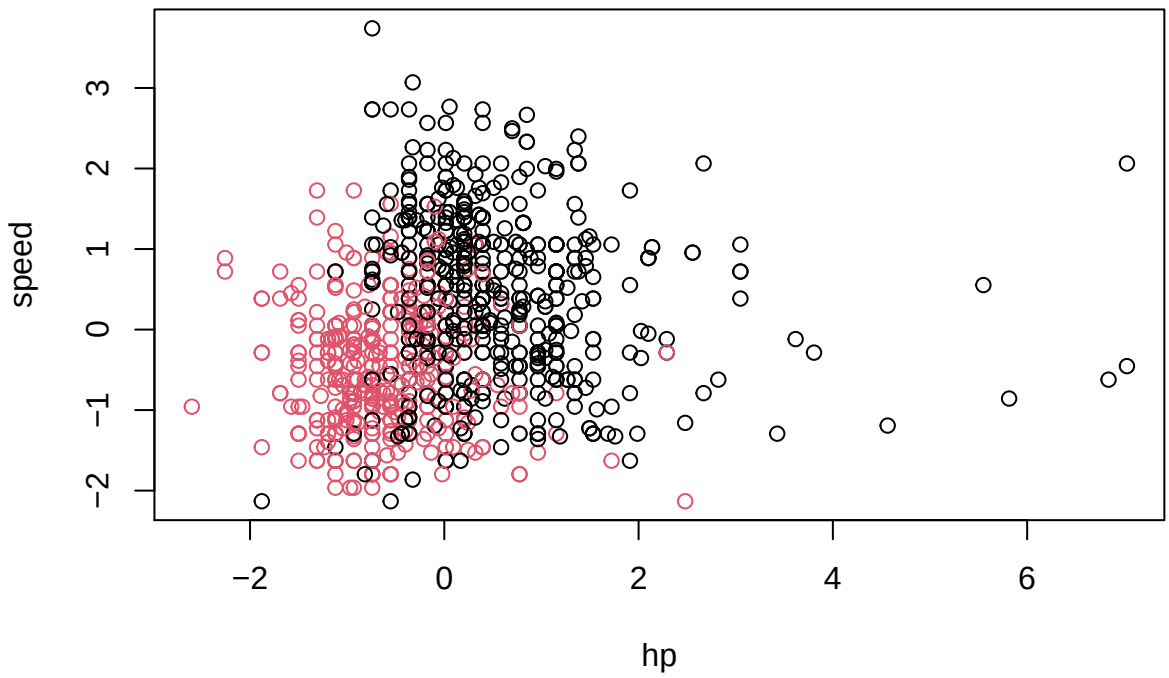
```
dataset <- as.data.frame(cbind(dataset, kmeansmodel$cluster))

for (i in 1:6) {
  for (j in 2:6) {
    if(i < j){
      plot(dataset[,i], dataset[,j], col = dataset[,7], xlab = names(dataset[i]), ylab = names(dataset[
    }
  }
}
```

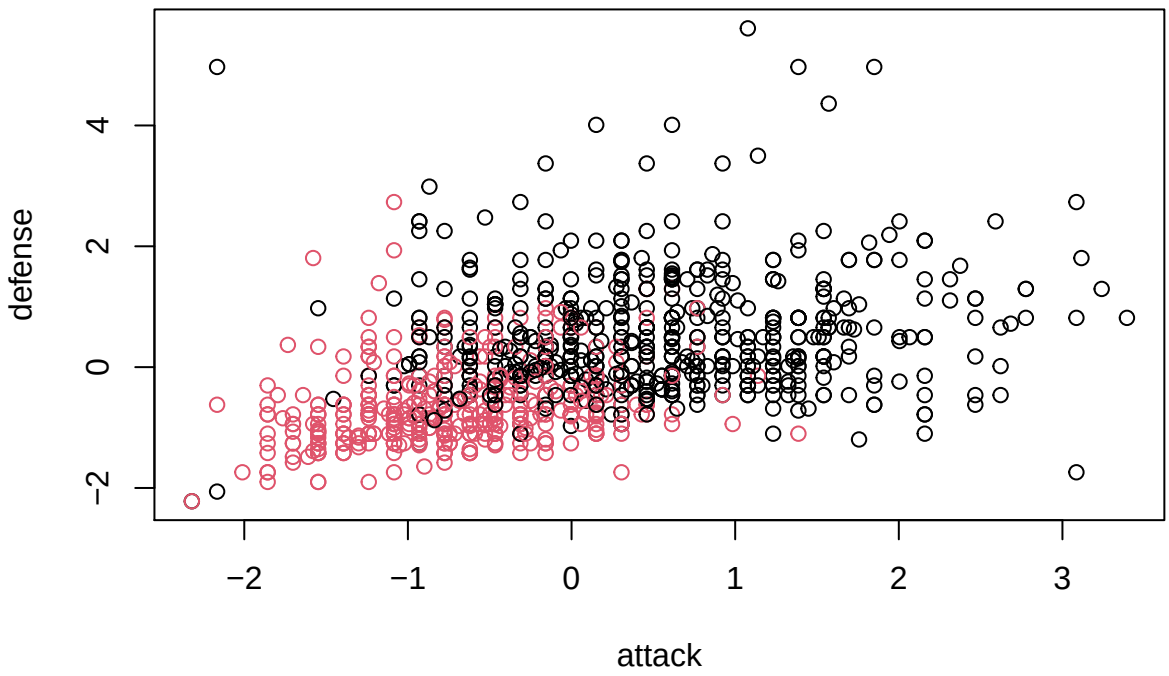

attack
hp
−2
−1
0
1
2
3
−2
0
2
4
6

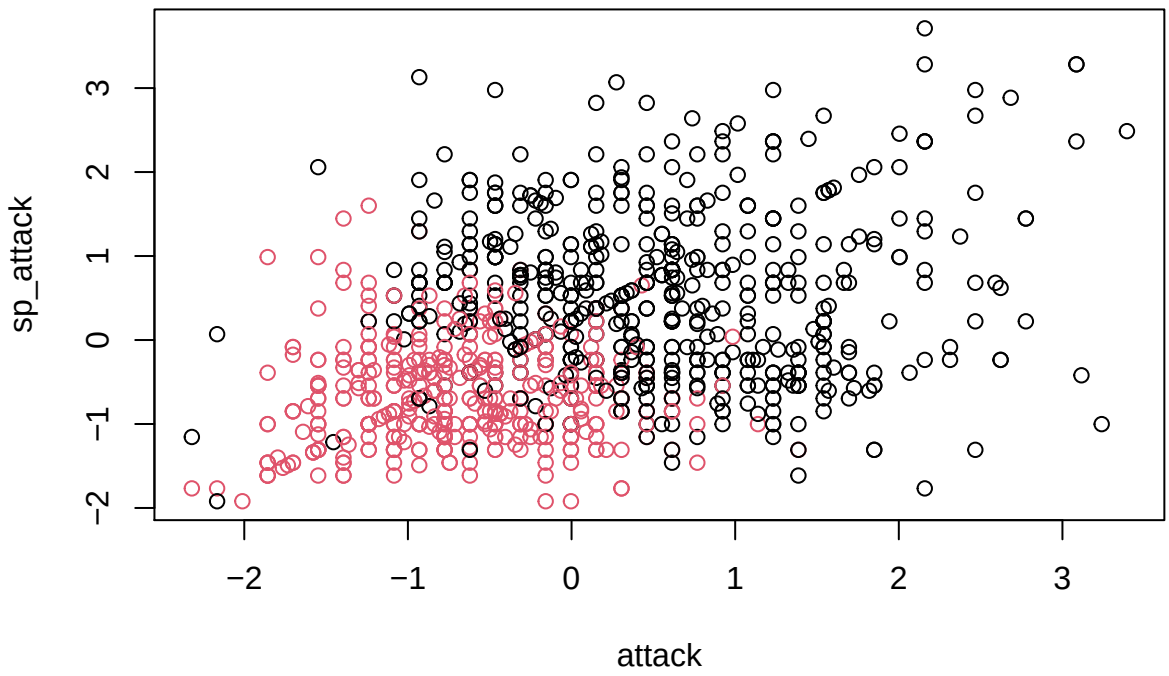

defense
hp
−2
0
2
4
6
−2
0
2
4

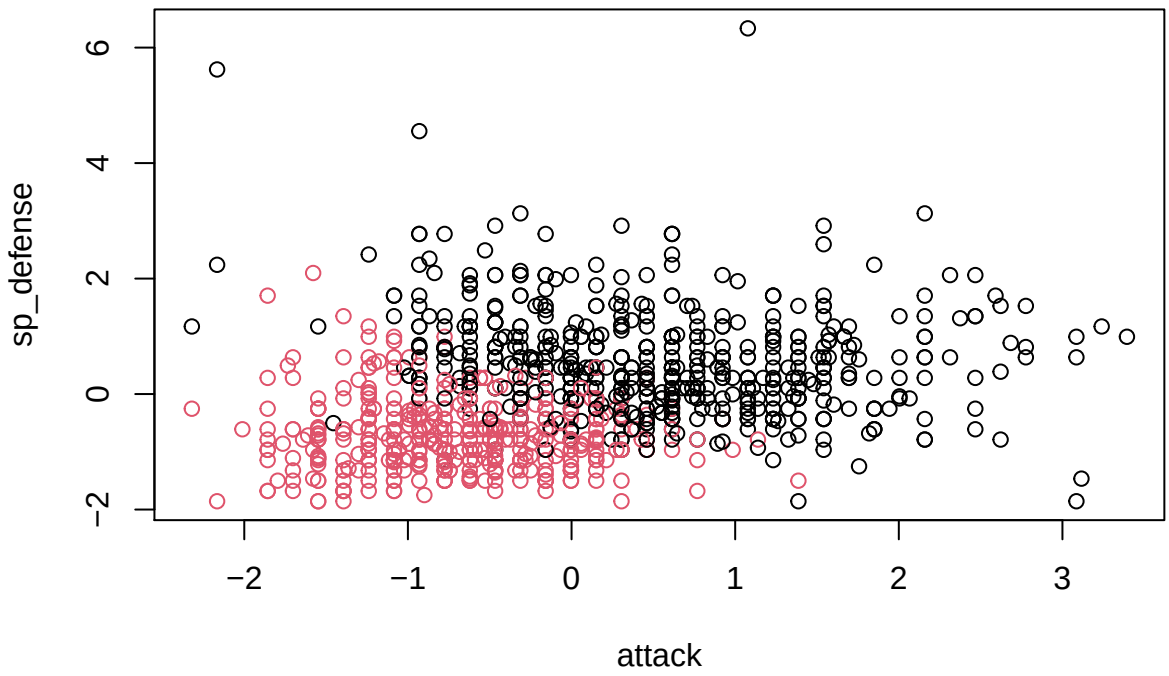

sp_attack
hp
−2
−1
0
1
2
3
−2
0
2
4
6

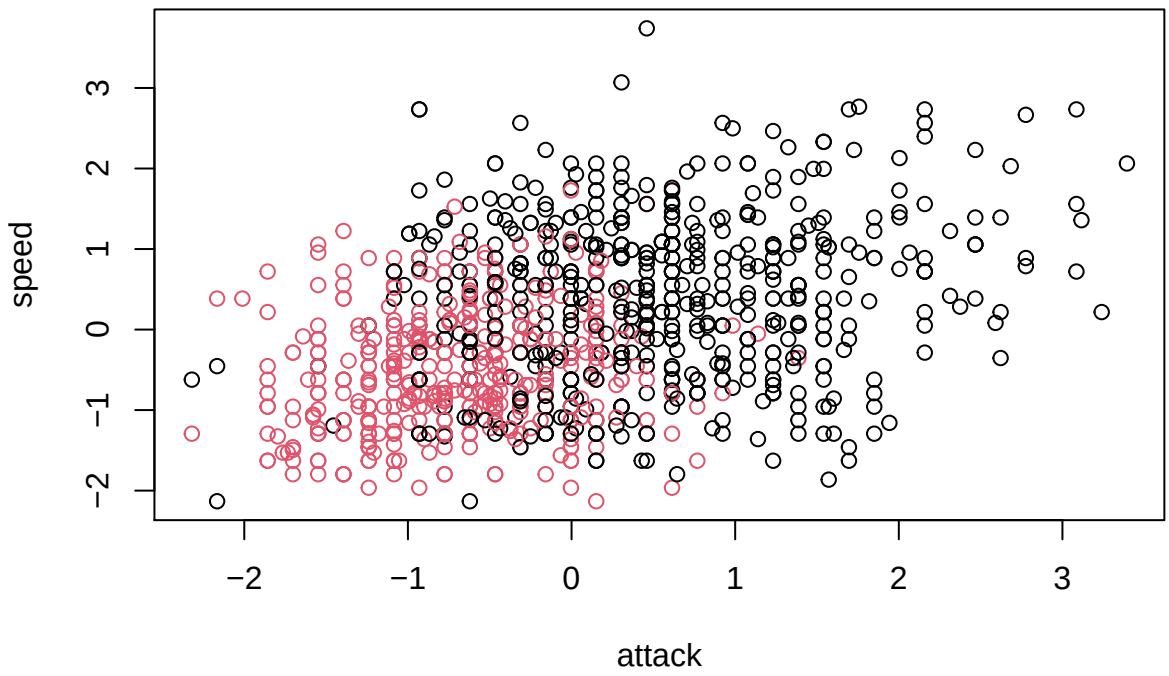

sp_defense
hp
−2
0
2
4
6

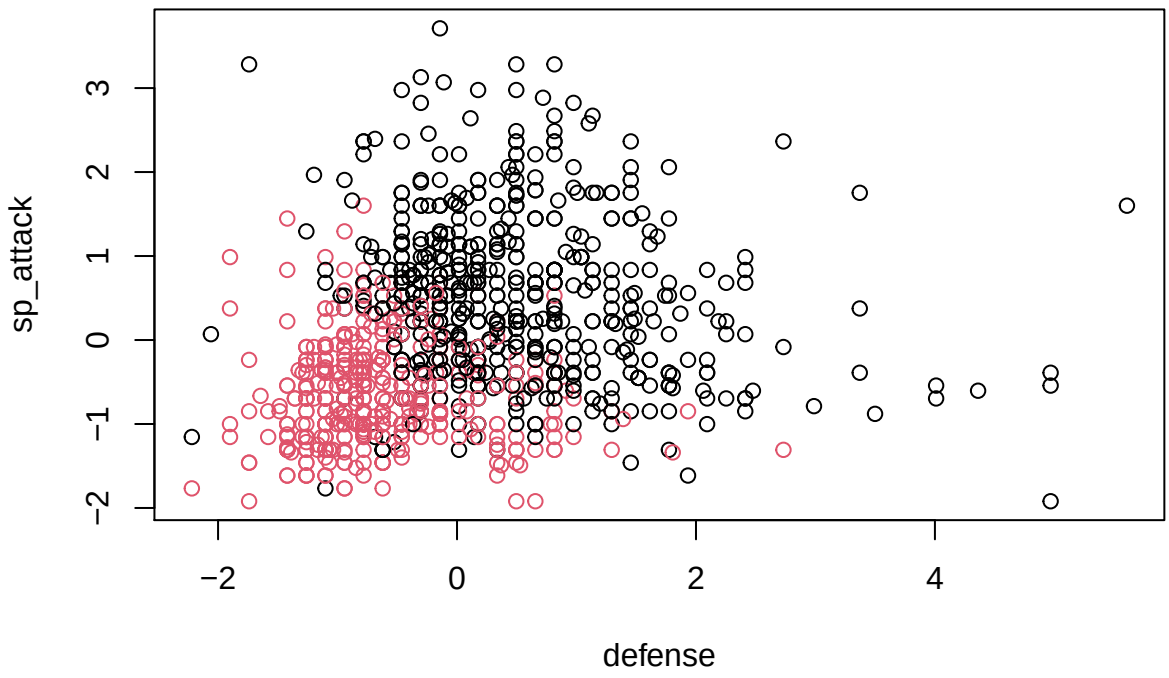

speed
hp
−2
−1
0
1
2
3
−2
0
2
4
6

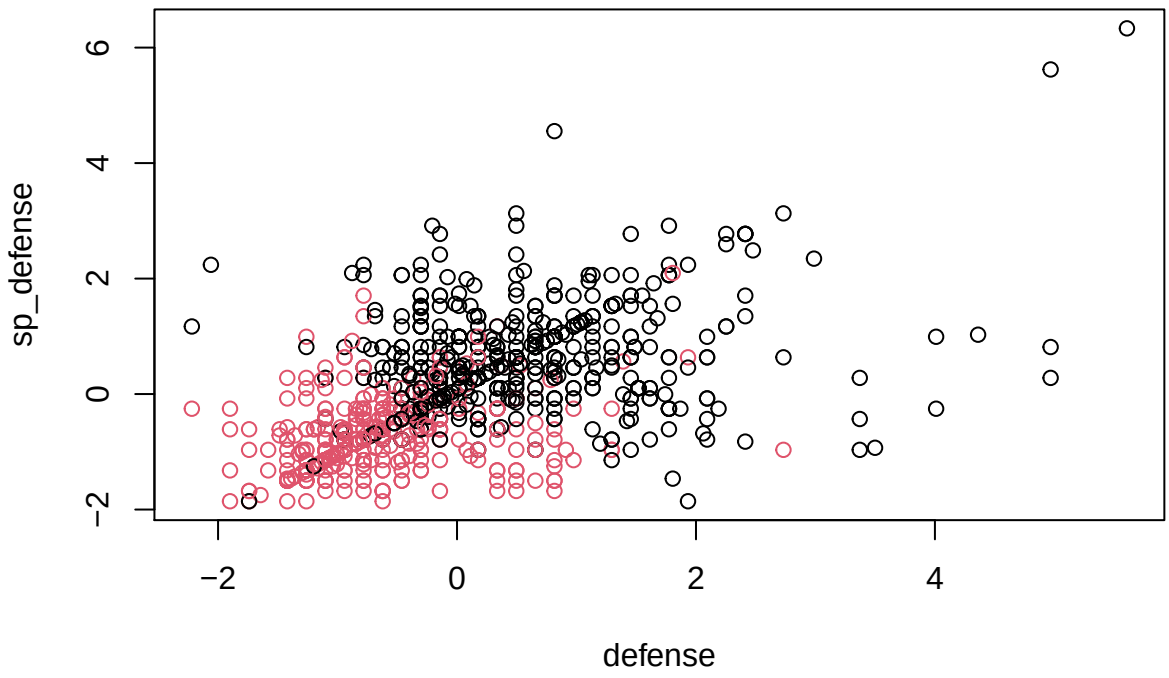

defense
attack
−2
0
2
4
−2
−1
0
1
2
3

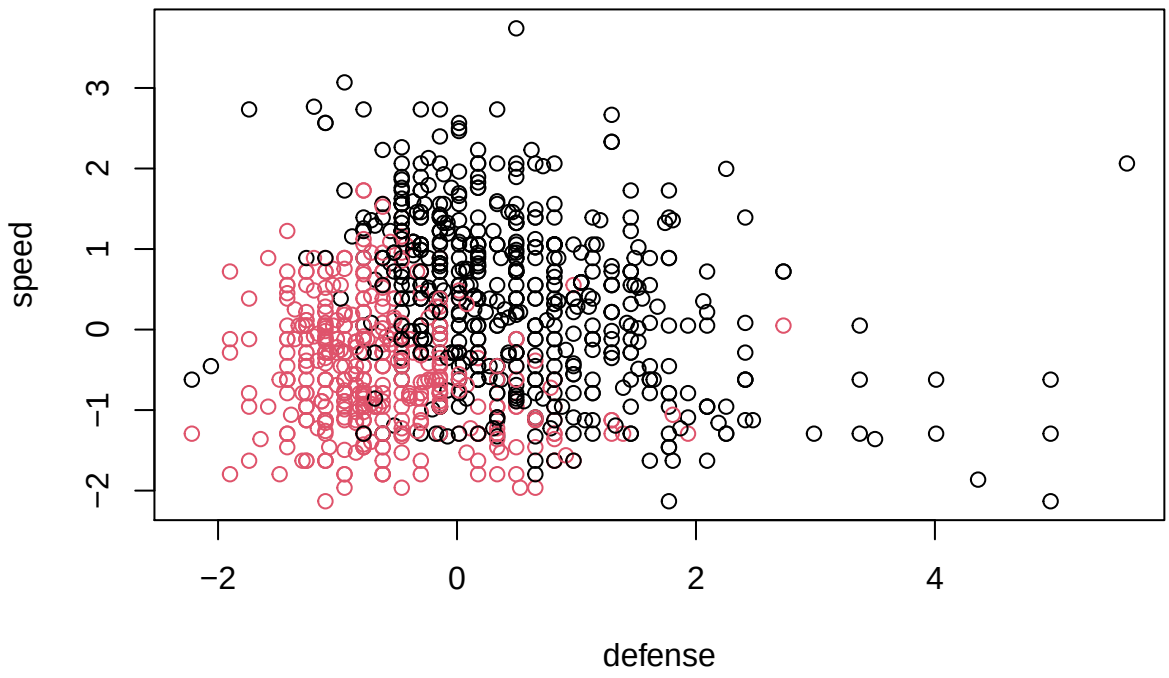

sp_attack
attack
−2
−1
0
1
2
3

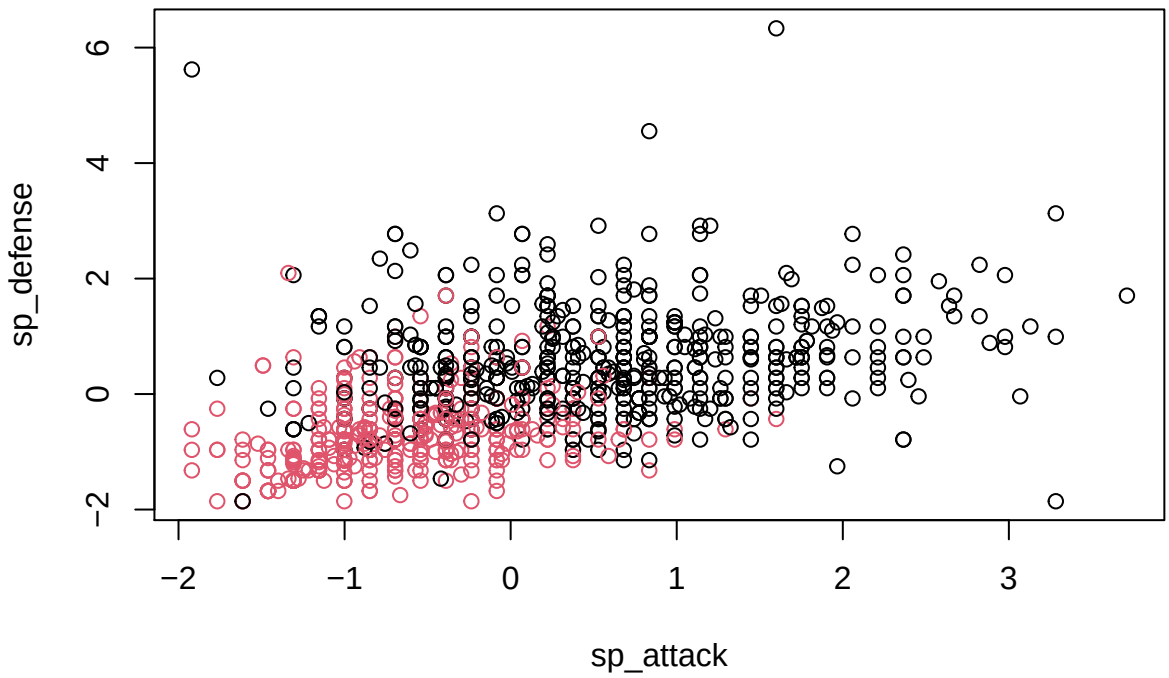

sp_defense
attack
−2
0
2
4
6
−2
−1
0
1
2
3

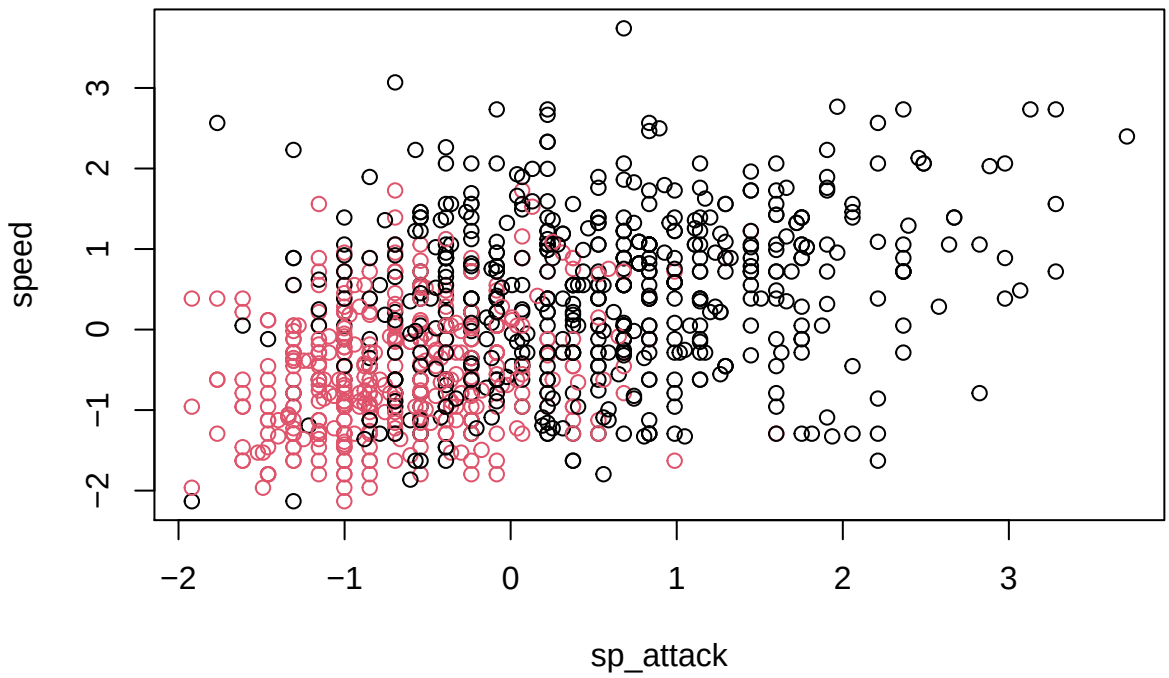

speed
attack
−2
−1
0
1
2
3

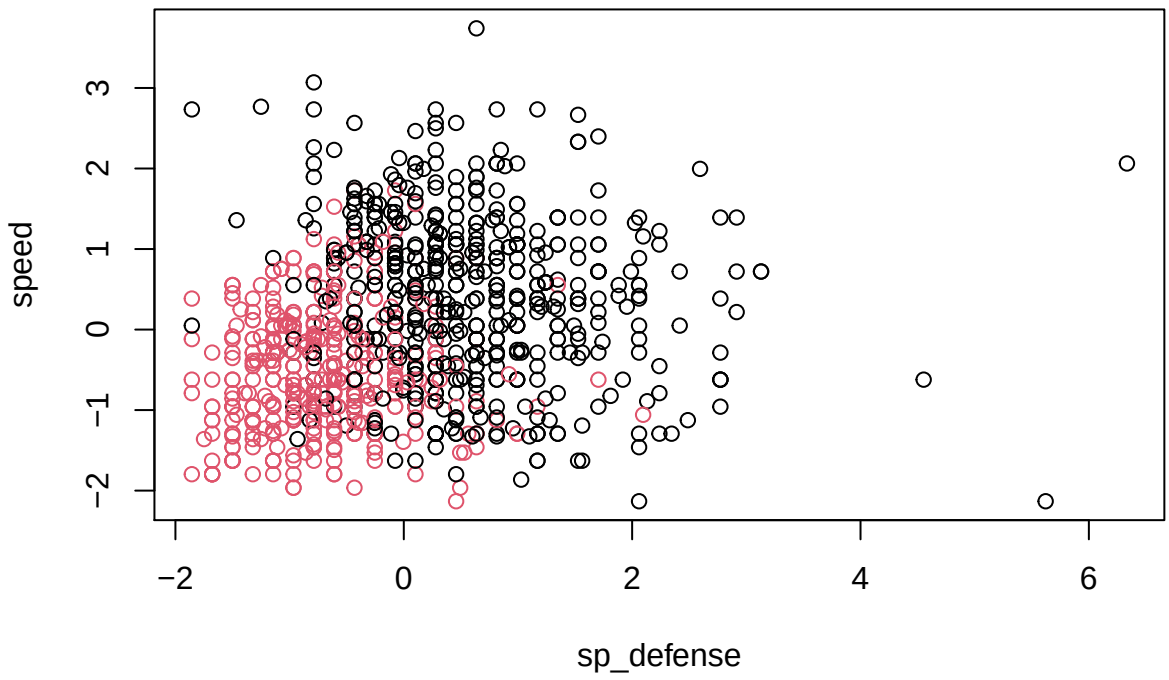

sp_attack
3
2
1
0
−1
−2
−2
0
2
4
defense


sp_defense
6
4
2
0
−2
−2
0
2
4
defense


speed
defense
−2
−1
0
1
2
3
−2
0
2
4


sp_defense
6
4
2
0
−2
−2
−1
0
1
2
3
sp_attack


speed
sp_attack
−2
−1
0
1
2
3

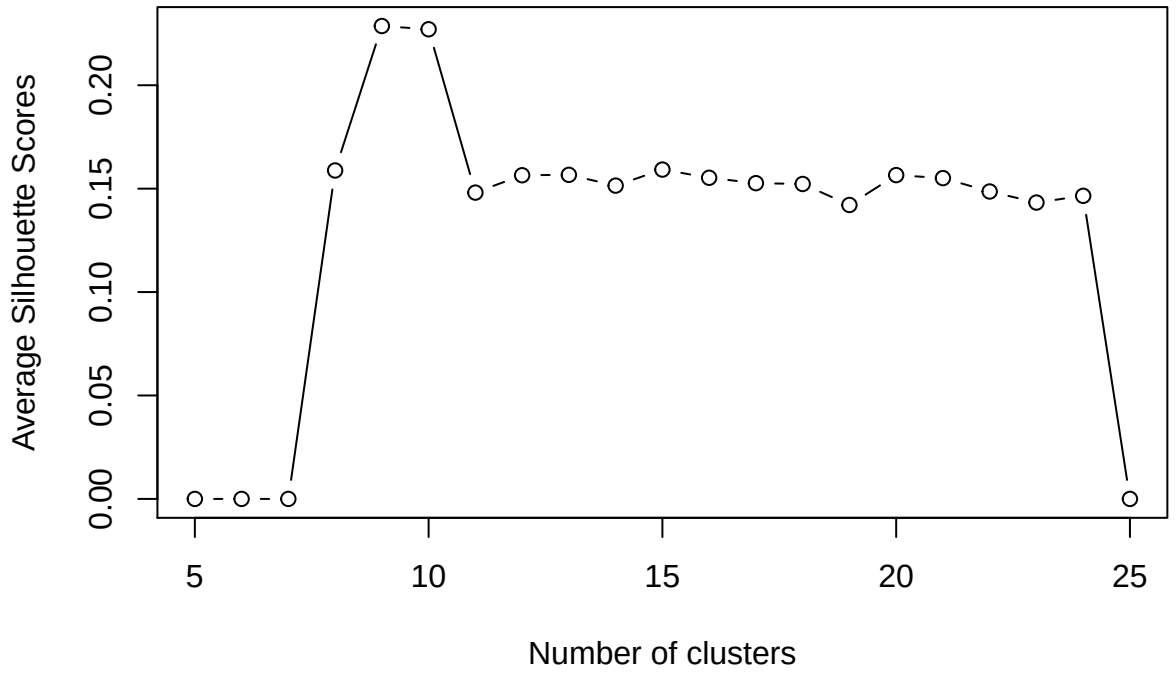



```
kmeansmodel$centers
```

```
##           hp     attack    defense  sp_attack sp_defense      speed
## 1  0.4952192  0.5300545  0.4840779  0.5136510  0.5449158  0.3930374
## 2 -0.6261141 -0.6701571 -0.6120280 -0.6494177 -0.6889465 -0.4969239
```

```
##From this information, we infer that cluster 2 pokemon can be categorized on the weaker side
##while cluster 1 pokemon are stronger

results <- as.data.frame(cbind(Master[,c(2,3,6,7,10,11,15,16,17,18)], dataset[,7]))
results[,11] <- as.numeric(results[,11])
for (i in 1:dim(results)[1]) {
  if((results[i,11] == 2)) {
    results[i,11] <- 'weak'
  }
  if((results[i,11] == 1)){
    results[i,11] <- 'strong'
  }
}
colnames(results) <- c("Pokedex Number", "Name", "Generation", "Status", "1st Type", "2nd Type",
                       "1st Ability", "2nd Ability", "Hidden Ability", "Total Points", "Cluster")

write.csv(results, file = "~/Library/Mobile Documents/com~apple~CloudDocs/Pokemon KMeans/Results.csv")



set.seed(5)
```

```r
##let us define the optimal k clusters through silhouette method
silhouettescore <- function(k){
  km <- kmeans(dataset[,-7], centers = k)
  ss <- silhouette(km$cluster, dist(dataset[,-7]))
  mean(ss[,3])
}
k <- 5:25
avgss <- c(rep(0,21))
for (i in min(k):length(k)) {
  avgss[i-1] <- silhouettescore(i)
}
plot(k, type = 'b', avgss, xlab = 'Number of clusters', ylab = 'Average Silhouette Scores')
```




```r
optk <- which.max(avgss)+1
optk
```

```
## [1] 6
```

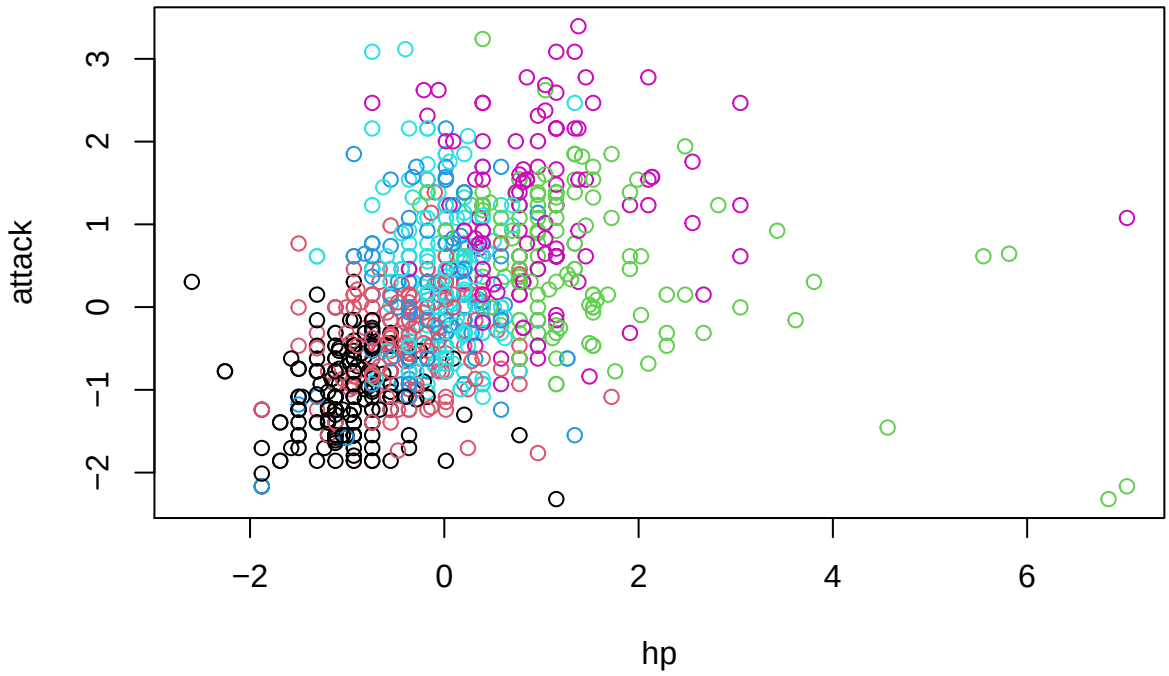
```r
kmeansmodel <- kmeans(dataset, optk)
kmeansmodel$tot.withinss
```

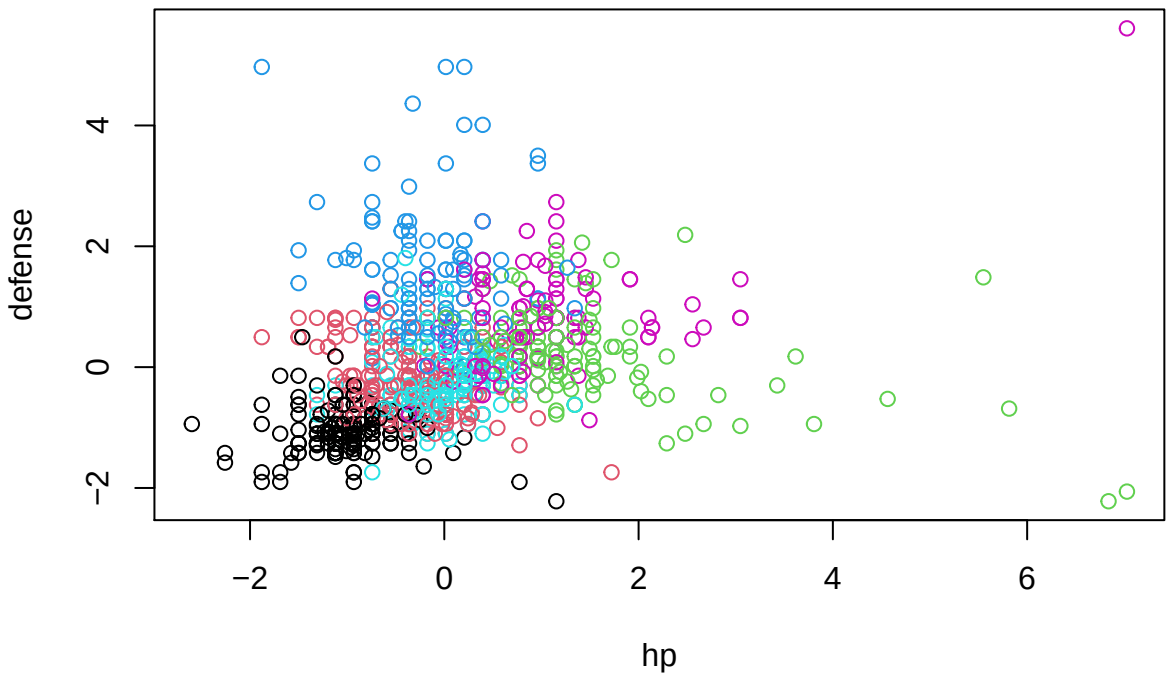
```
## [1] 2919.899
```

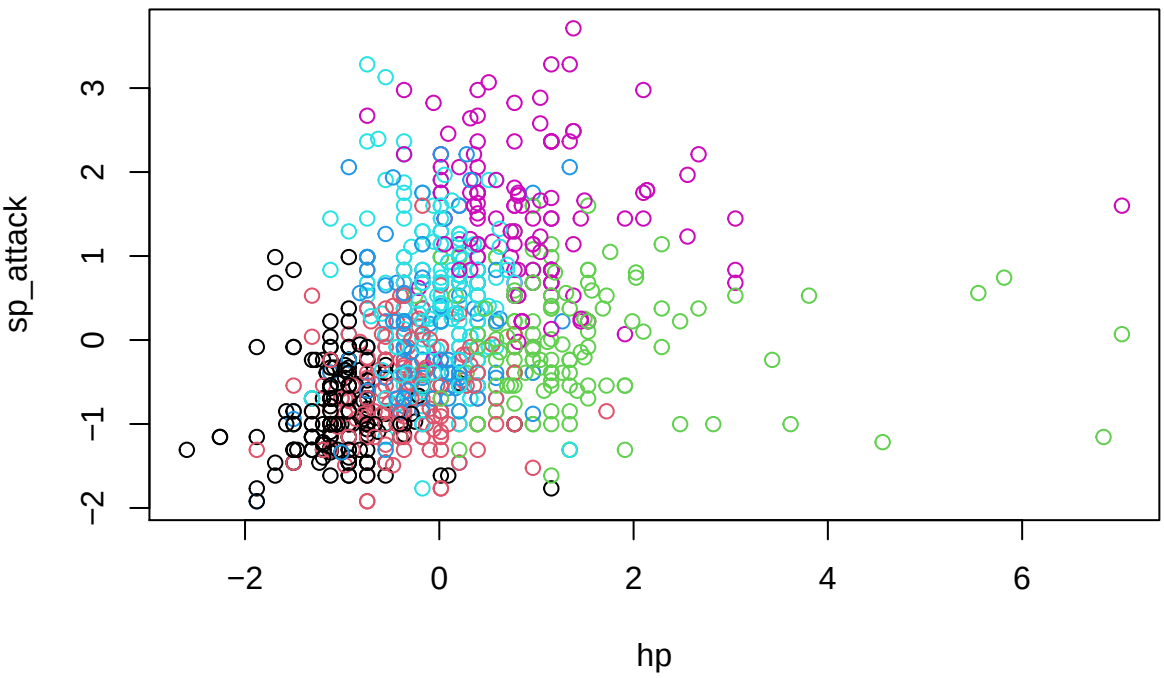
```r
kmeansmodel$size
```

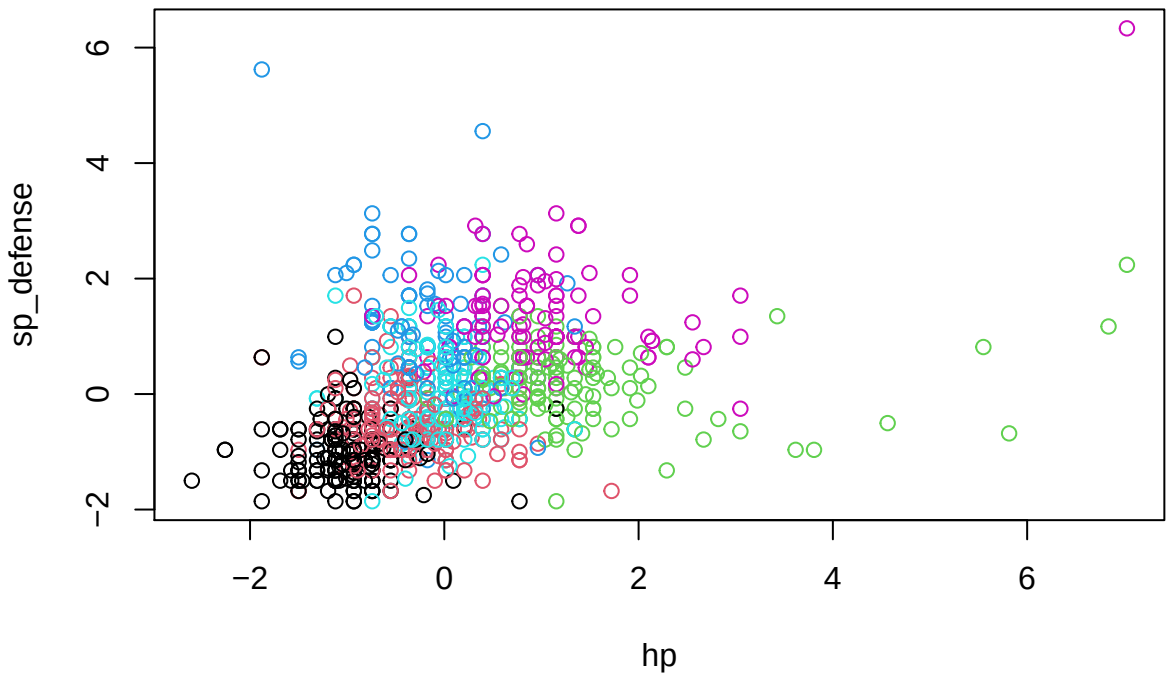
```
## [1] 194 223 142 127 213 129
```

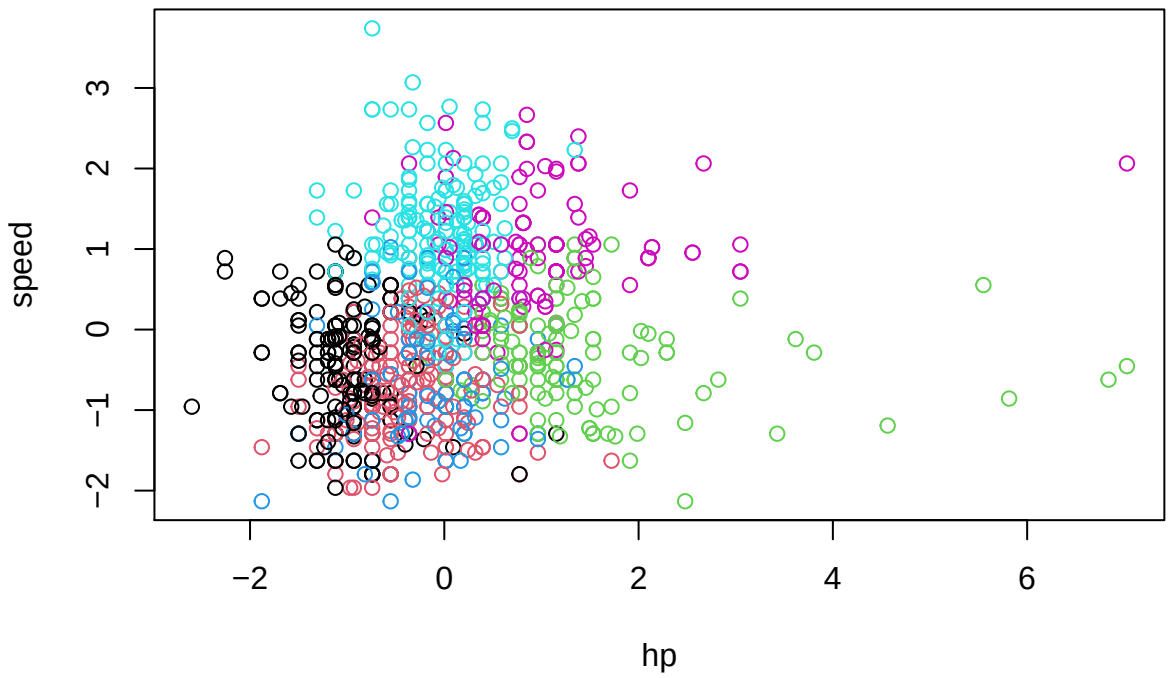
```r
dataset <- as.data.frame(cbind(dataset, kmeansmodel$cluster))

for (i in 1:6) {
  for (j in 2:6) {
    if(i < j){
      plot(dataset[,i], dataset[,j], col = dataset[,8], xlab = names(dataset[i]), ylab = names(dataset[
    }
  }
}
```

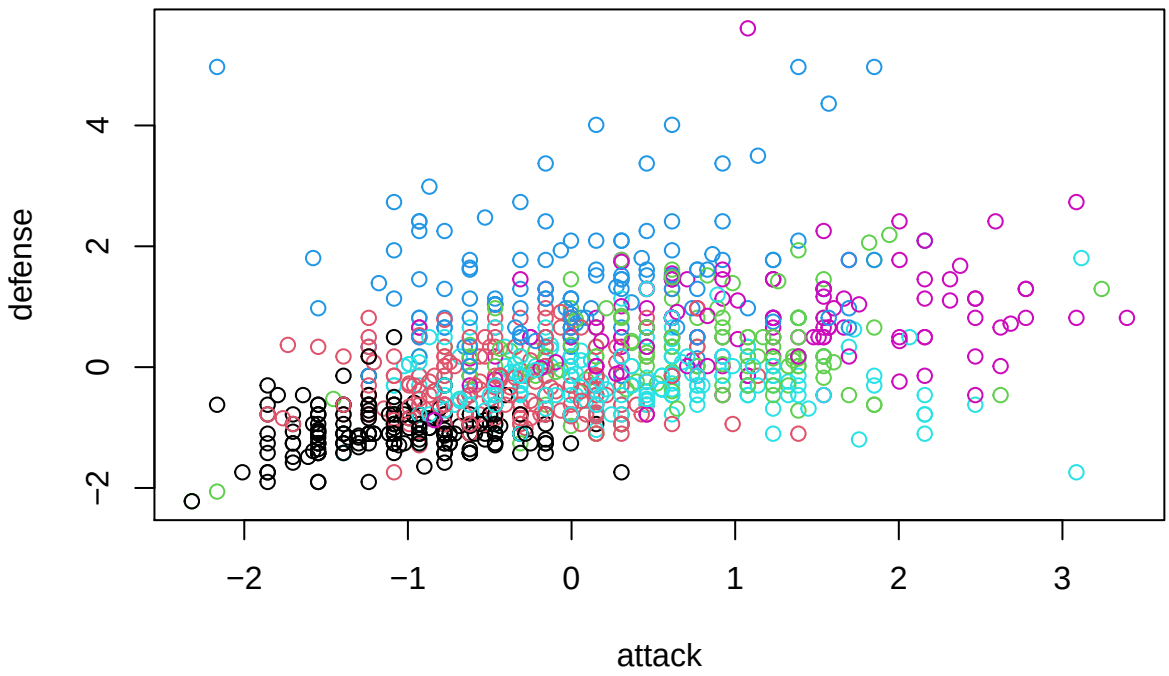


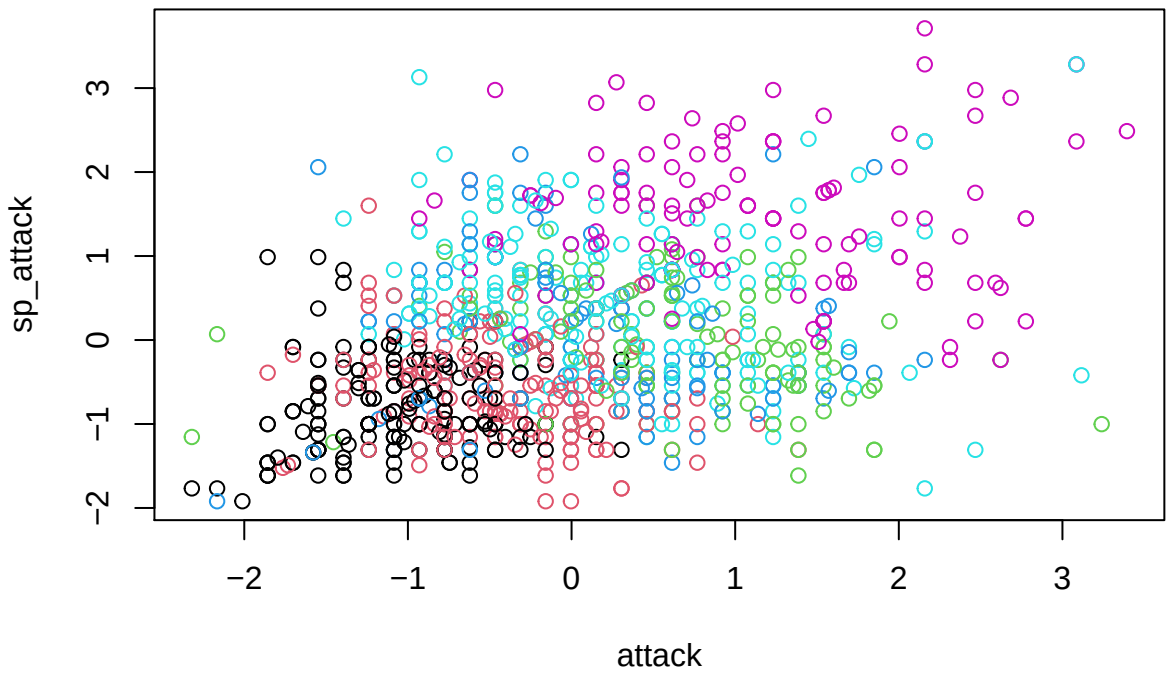

defense
hp
−2
0
2
4
6
−2
0
2
4

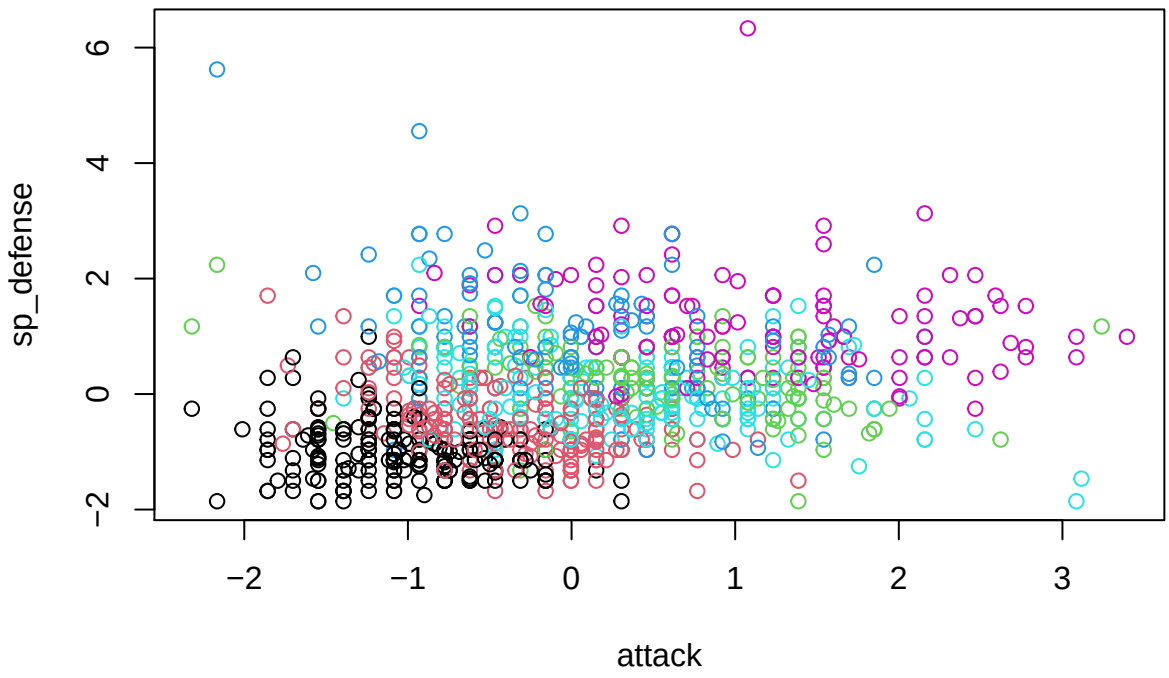

sp_attack
hp
−2
−1
0
1
2
3
−2
0
2
4
6

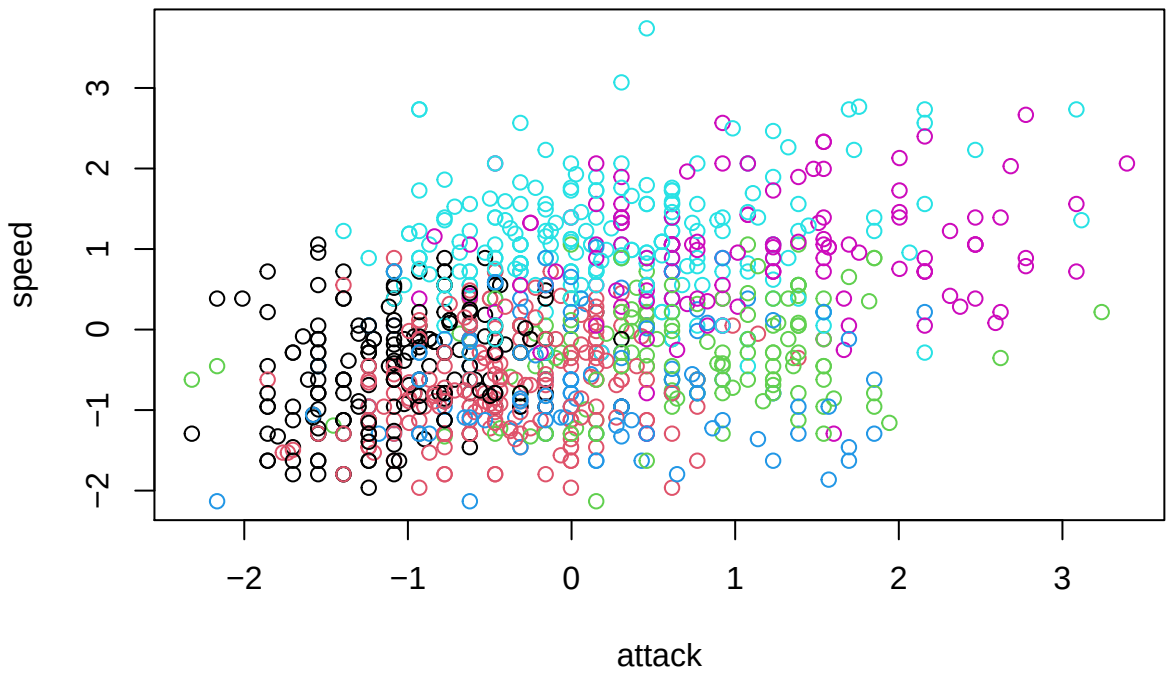

sp_defense
hp
−2
0
2
4
6
−2
0
2
4
6

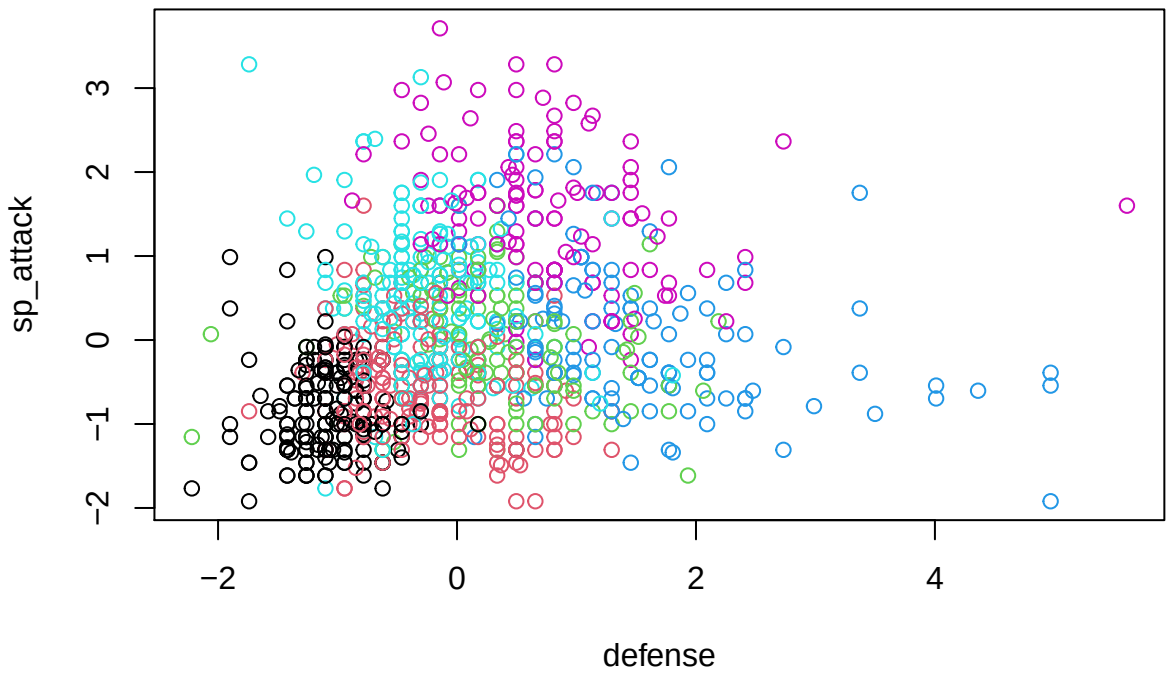

speed
hp
−2
−1
0
1
2
3
−2
0
2
4
6

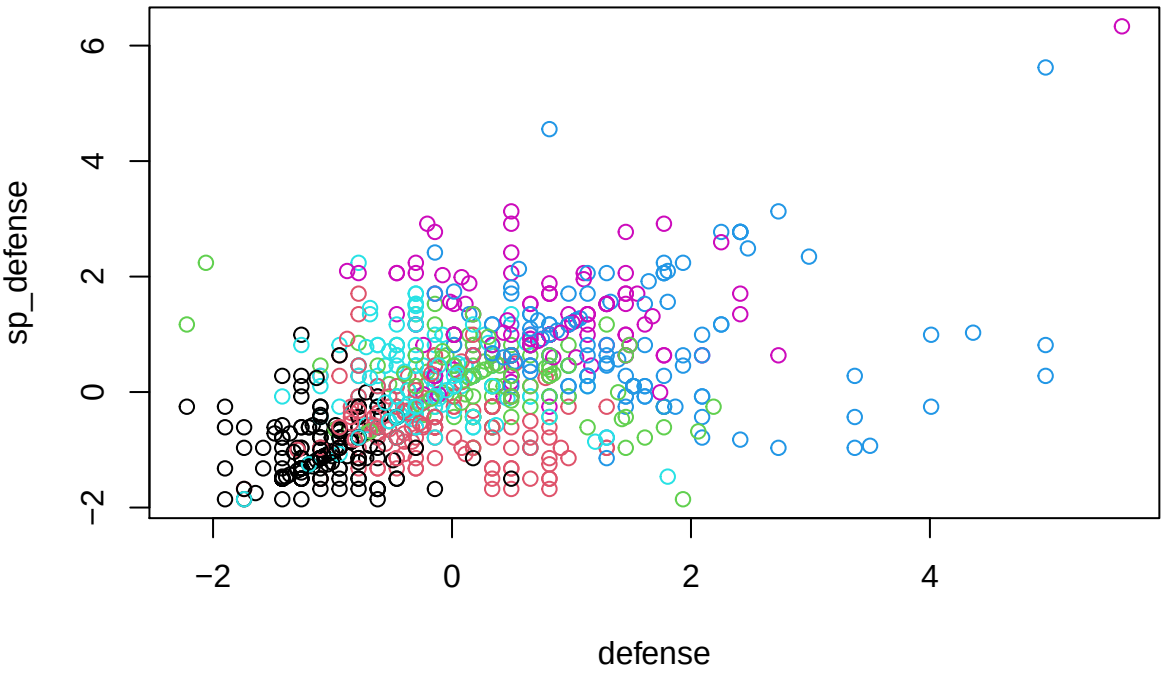

defense
attack
−2
0
2
4
−2
−1
0
1
2
3

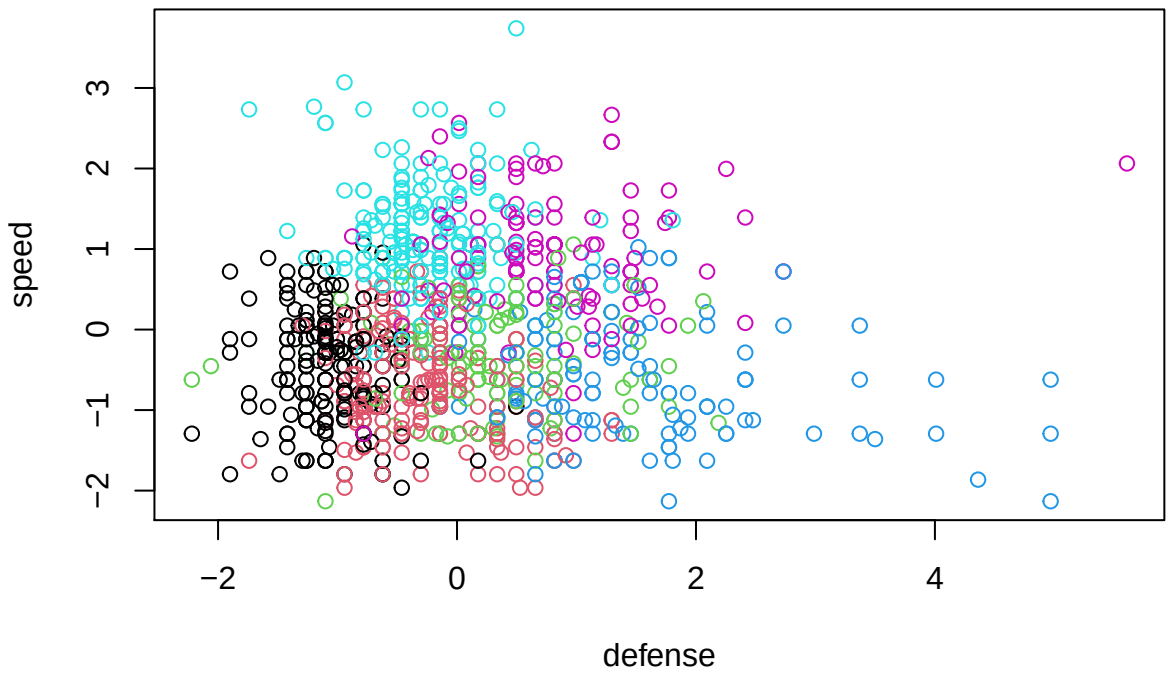

sp_attack
attack
−2
−1
0
1
2
3

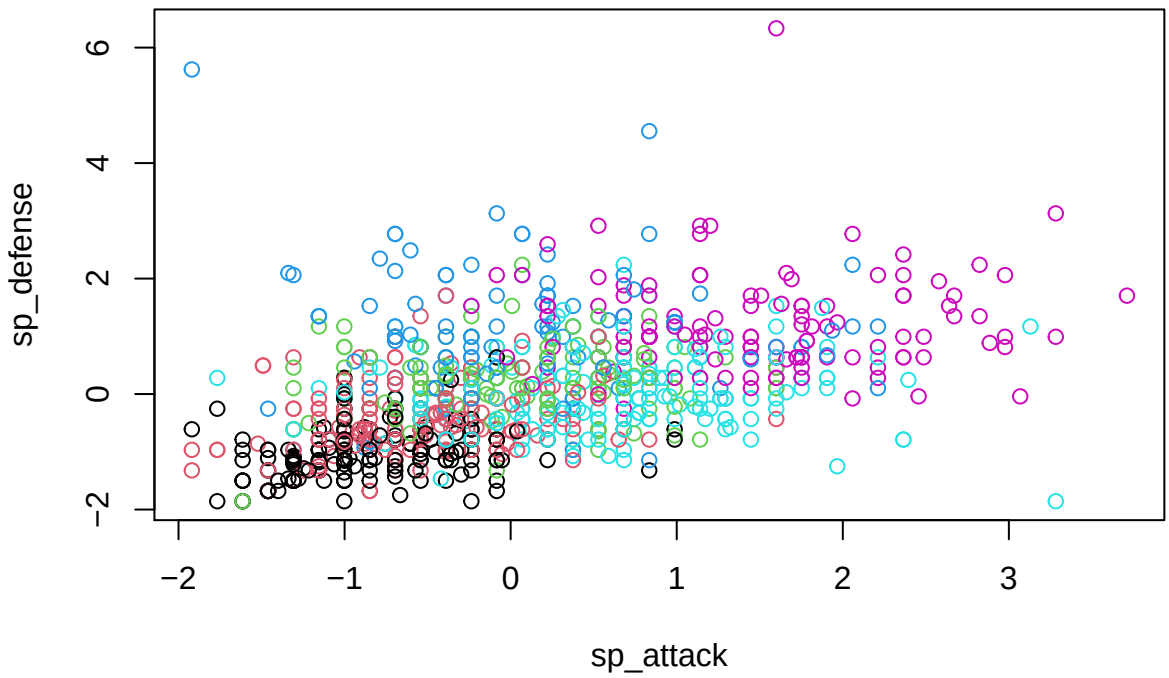

sp_defense
attack
6
4
2
0
−2
−2
−1
0
1
2
3

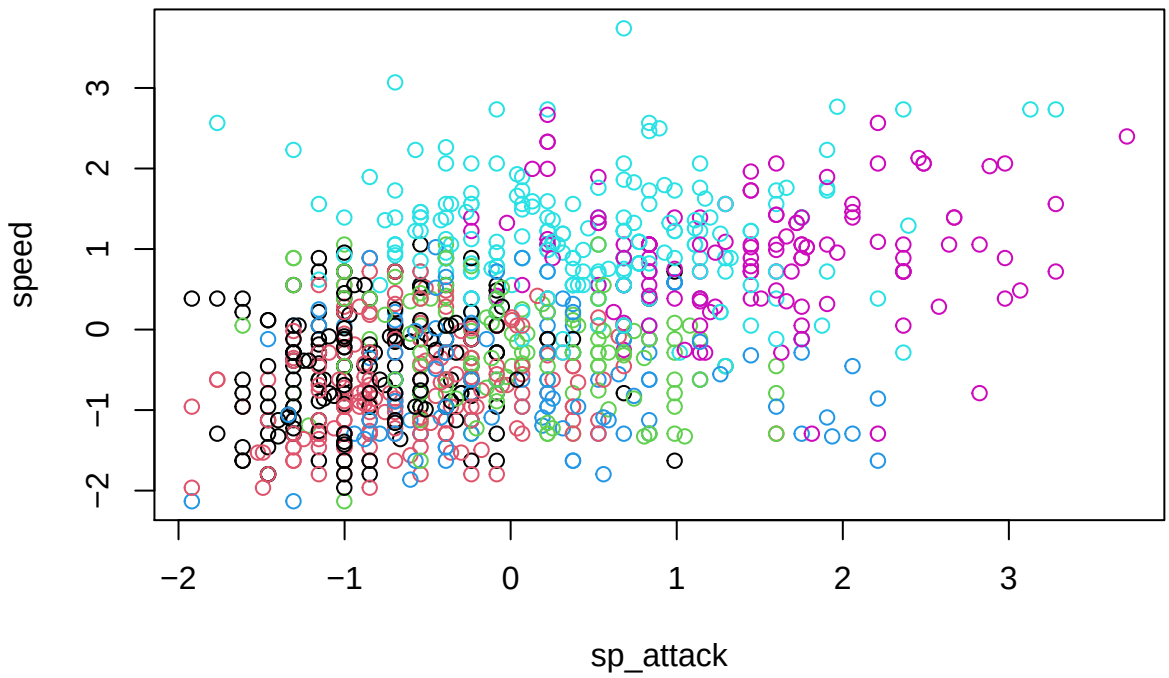

speed
attack
−2
−1
0
1
2
3

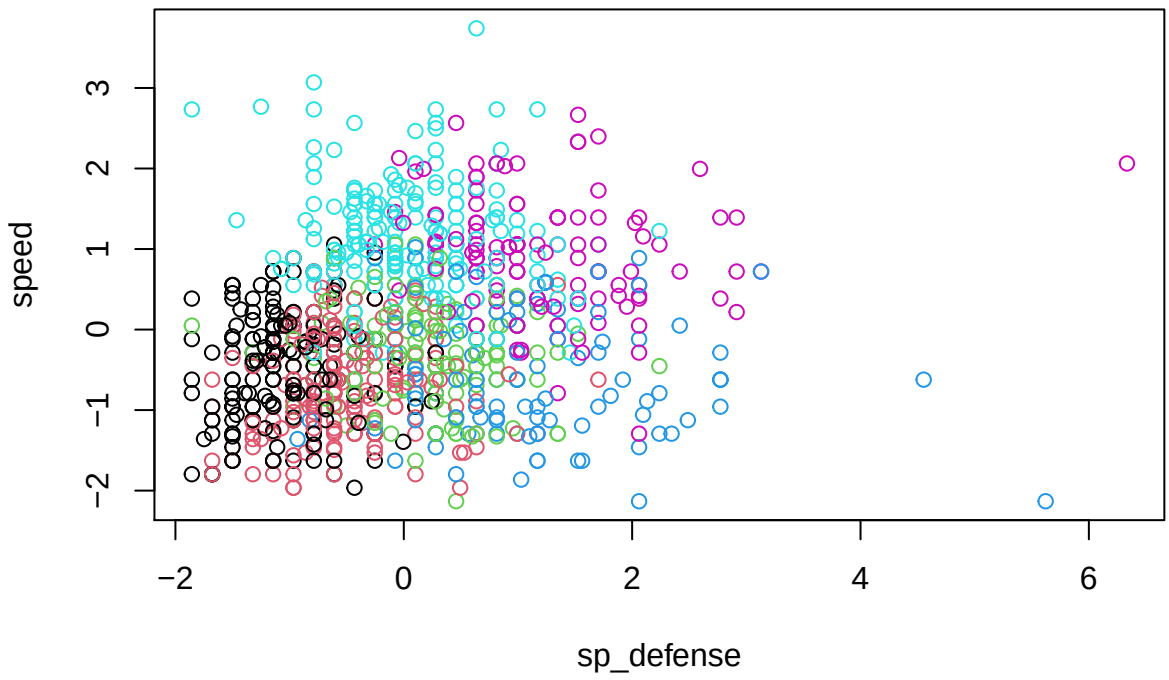

sp_attack
3
2
1
0
−1
−2
−2
0
2
4
defense


sp_defense
6
4
2
0
−2
−2
0
2
4
defense


speed
defense
−2
−1
0
1
2
3
−2
0
2
4


sp_defense
6
4
2
0
−2
−2
−1
0
1
2
3
sp_attack


speed
sp_attack
−2
−1
0
1
2
3




```
kmeansmodel$centers
```

```
##            hp      attack    defense    sp_attack  sp_defense      speed
## 1 -0.99301457 -1.04617375 -1.0547421 -0.826741073 -0.99695878 -0.4721187
## 2 -0.36316077 -0.38754276 -0.2727654 -0.577229611 -0.50481763 -0.6724617
## 3  1.31017450  0.60454396  0.2477921 -0.008192179  0.18887565 -0.3118777
## 4 -0.16812378  0.08094902  1.4814636  0.081438126  0.91156089 -0.5497746
## 5 -0.04680911  0.24415377 -0.2440637  0.437618752  0.04878462  1.1131722
## 6  0.92175938  1.09495521  0.7294563  1.447424997  1.18608246  0.9190101
##         V7
## 1 2.000000
## 2 2.000000
## 3 1.028169
## 4 1.055118
## 5 1.122066
## 6 1.000000
```

```
##this gives more diverse clusters which will give us indicators into
##pokemon strength depending on multiple more variables (ie tanks with high
##health, speed oriented pokemon, etc...)

results2 <- as.data.frame(cbind(Master[,c(2,3,6,7,10,11,15,16,17,18)], dataset[,8]))
results2[,11] <- as.numeric(results2[,11])
for (i in 1:dim(results)[1]) {
  if((results2[i,11] == 1)) {
    results2[i,11] <- 'small weak'
  }
```

```
  if((results2[i,11] == 2)){
    results2[i,11] <- 'big weak'
  }
  if((results2[i,11] == 3)){
    results2[i,11] <- 'health tanks'
  }
  if((results2[i,11] == 4)){
    results2[i,11] <- 'defensive specialists'
  }
  if((results2[i,11] == 5)){
    results2[i,11] <- 'fast attackers'
  }
  if((results2[i,11] == 6)){
    results2[i,11] <- 'overall strongest'
  }
}
colnames(results2) <- c("Pokedex Number", "Name", "Generation", "Status", "1st Type", "2nd Type",
                        "1st Ability", "2nd Ability", "Hidden Ability", "Total Points", "Cluster")

write.csv(results2, file = "~/Library/Mobile Documents/com~apple~CloudDocs/Pokemon KMeans/Results2.csv")
```