Multivariate

Chandan Kumar Pandey

2022-10-23

Data

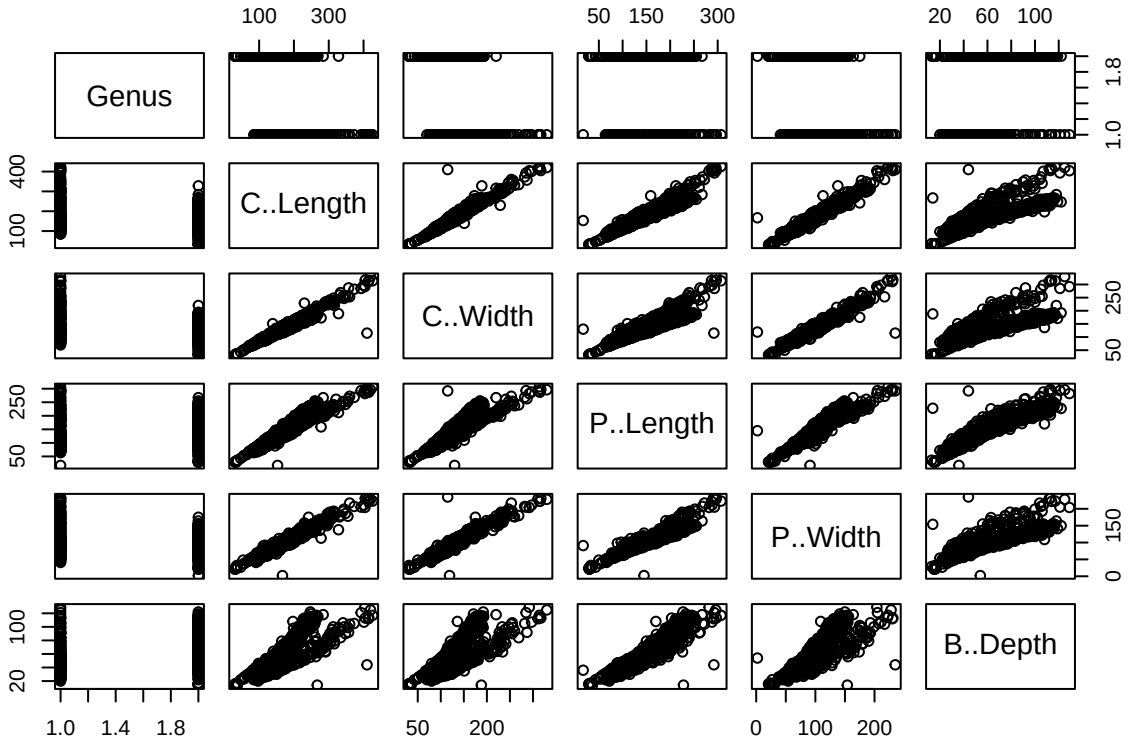
```
Turtle_data <- read.csv("data/Morphometric  freshwater turtles in  Texas.csv")
str(Turtle_data)
```

```
## 'data.frame':    1473 obs. of  13 variables:
##  $ Region         : chr  "Central" "Central" "Central" "Central" ...
##  $ Water.Body.Type: chr  "Terrestrial" "Terrestrial" "Terrestrial" "Terrestrial" ...
##  $ Capture.Method : chr  "Hand" "Hand" "Hand" "Hand" ...
##  $ Genus          : chr  "Trachemys" "Trachemys" "Trachemys" "Trachemys" ...
##  $ Species        : chr  "scripta" "scripta" "scripta" "scripta" ...
##  $ Subspecies     : chr  "elegans" "elegans" "elegans" "elegans" ...
##  $ Sex            : chr  "J" "J" "J" "J" ...
##  $ C..Length      : int  30 36 32 32 33 33 35 37 37 35 ...
##  $ C..Width       : int  31 31 33 33 32 34 36 35 35 36 ...
##  $ P..Length      : int  30 33 29 30 29 31 32 33 36 33 ...
##  $ P..Width       : num  21 32 24 21 24 27 27 24 29 24 ...
##  $ B..Depth       : int  15 16 15 16 16 15 16 17 13 16 ...
##  $ Weight         : int  6 6 7 7 7 8 8 9 9 10 ...
```

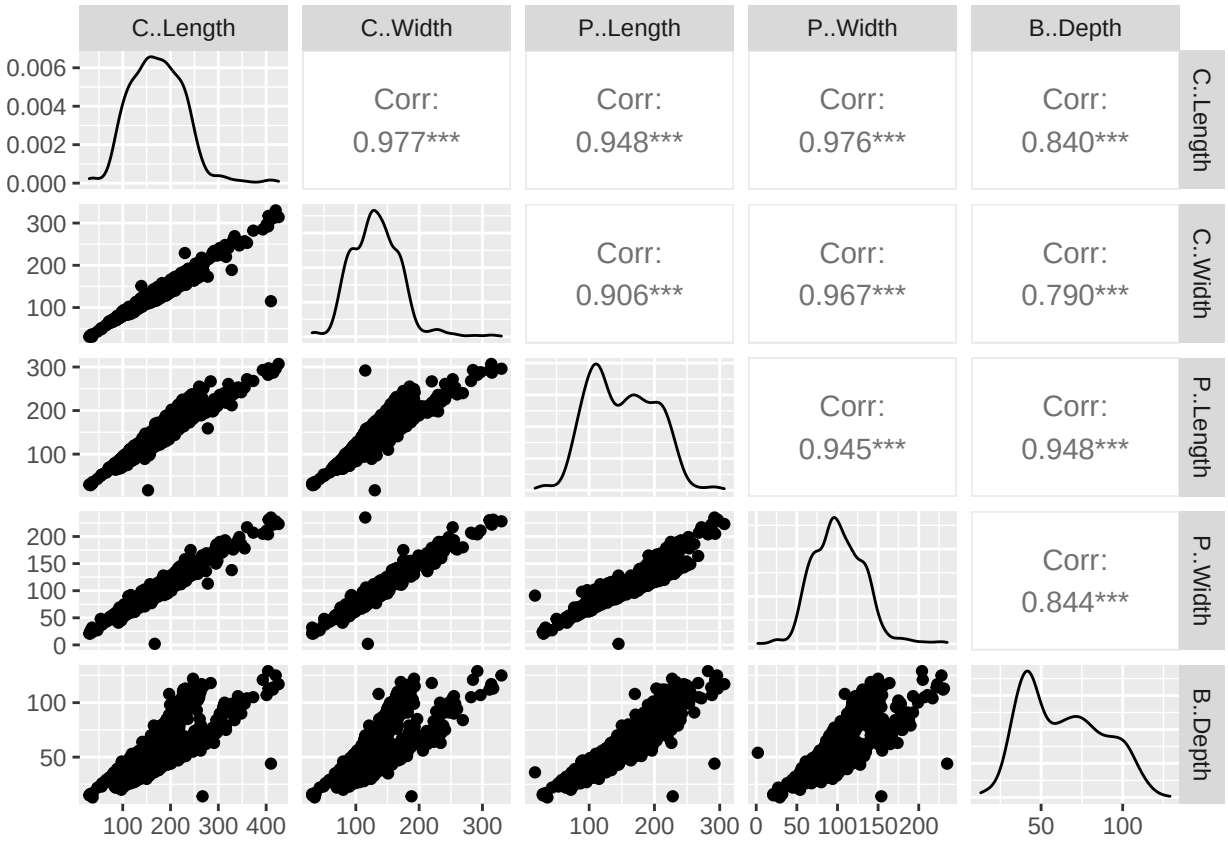
```
PCA_dataset <- Turtle_data[,8:12]
PCA_dataset <- PCA_dataset%>%mutate(index = 1:nrow(PCA_dataset))
PCA_dataset <- na.omit(PCA_dataset)
cor(PCA_dataset[,-7])
```

```
##           C..Length  C..Width P..Length  P..Width  B..Depth     index
## C..Length 1.0000000 0.9767872 0.9478985 0.9761307 0.8397600 0.8880946
## C..Width  0.9767872 1.0000000 0.9063626 0.9671105 0.7903971 0.8525386
## P..Length 0.9478985 0.9063626 1.0000000 0.9454771 0.9479211 0.9260595
## P..Width  0.9761307 0.9671105 0.9454771 1.0000000 0.8439385 0.8852823
## B..Depth  0.8397600 0.7903971 0.9479211 0.8439385 1.0000000 0.8863666
## index     0.8880946 0.8525386 0.9260595 0.8852823 0.8863666 1.0000000
```

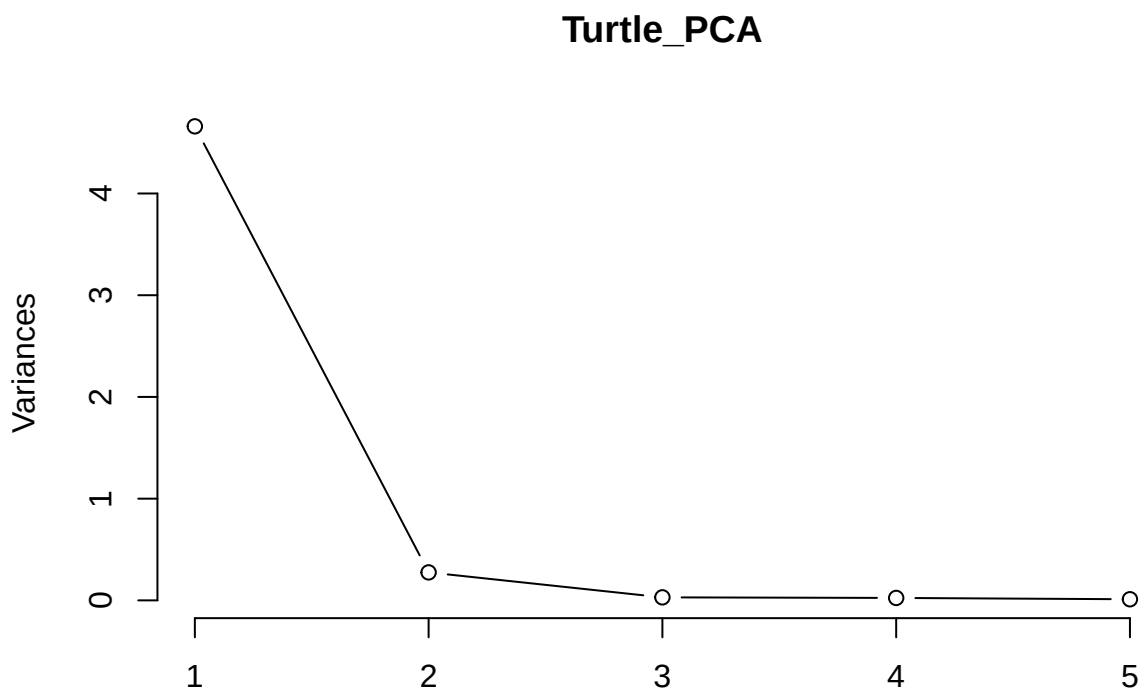
```
plot(Turtle_data[,c(4,8,9,10,11,12)])
```



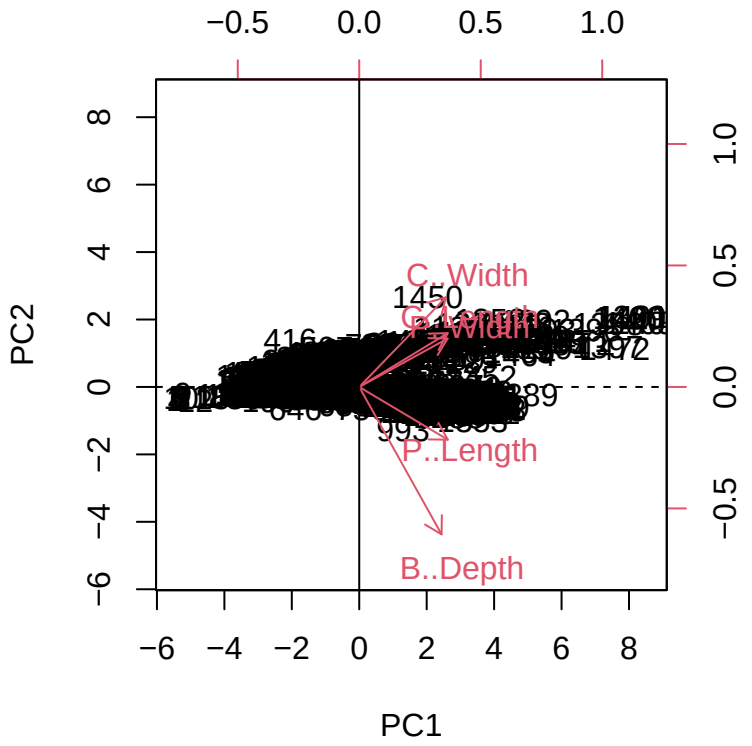

```
ggpairs(PCA_dataset[,-6])
```




```
## PCA analysis
Turtle_PCA <- prcomp(PCA_dataset[,-6],scale=T)
plot(Turtle_PCA, type="l")
```




```
biplot(Turtle_PCA,scale = 0)
abline(a=0,b=0,lty=2)
abline(v = 0)
```




```
summary(Turtle_PCA)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5
## Standard deviation     2.159 0.52447 0.17250 0.15460 0.10573
## Proportion of Variance 0.932 0.05501 0.00595 0.00478 0.00224
## Cumulative Proportion  0.932 0.98703 0.99298 0.99776 1.00000
```

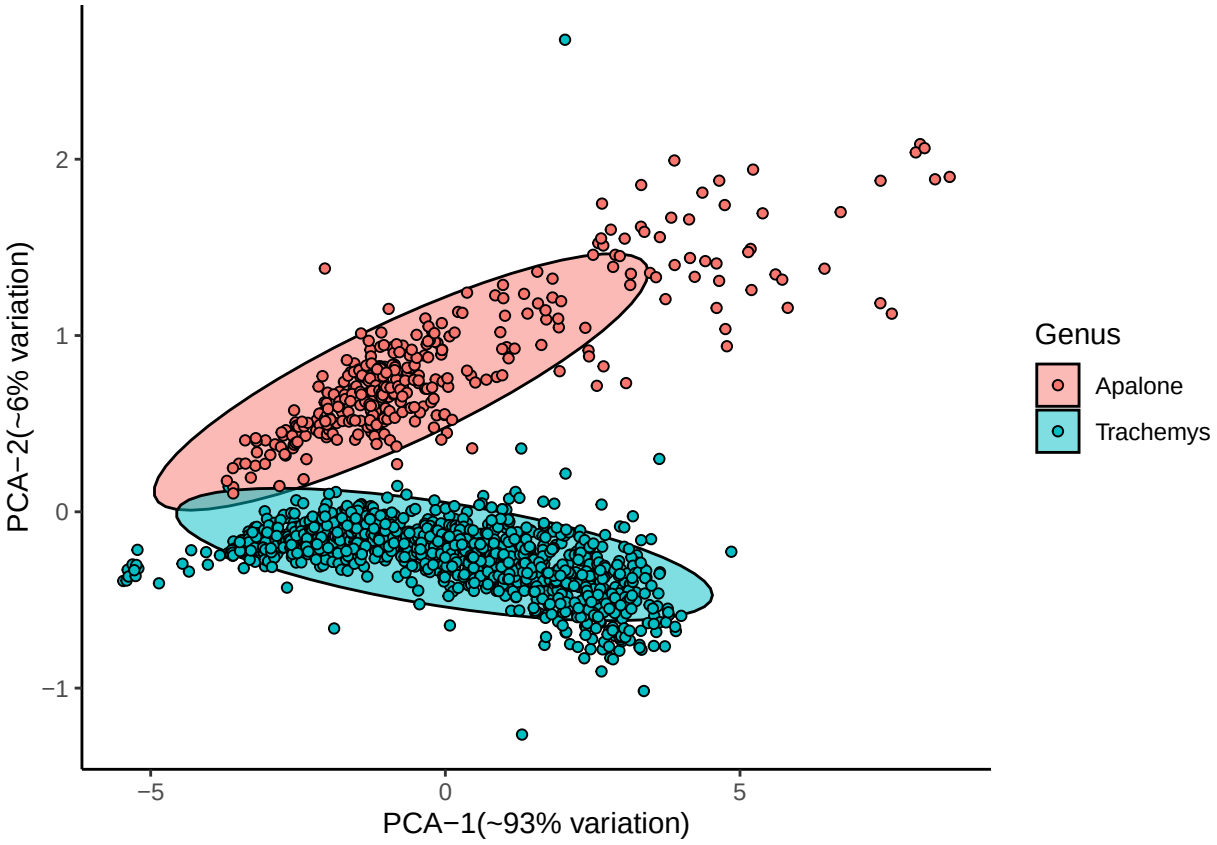
```
Turtle_PCA
```

```
## Standard deviations (1, .., p=5):
## [1] 2.1587266 0.5244658 0.1725018 0.1545961 0.1057273
##
## Rotation (n x k) = (5 x 5):
##                 PC1        PC2         PC3        PC4        PC5
## C..Length 0.4554428  0.2766761 -0.05138604  0.5783427  0.6155497
## C..Width  0.4459680  0.4611456 -0.63033917 -0.2363937 -0.3677605
## P..Length 0.4555458 -0.2711538  0.31556271  0.4704876 -0.6308845
## P..Width  0.4546558  0.2489082  0.63811911 -0.5550914  0.1265321
## B..Depth  0.4236161 -0.7584955 -0.30537874 -0.2831126  0.2680021
```

```
PCA_dataset <-cbind(PCA_dataset,Turtle_PCA$x[,1:2])
PCA_dataset <- cbind(Turtle_data[PCA_dataset$index,][,1:7],PCA_dataset)
```

Let plot the PCA data

```
ggplot(data = PCA_dataset,aes(x=PC1,y=PC2,col=Genus,fill=Genus))+
  stat_ellipse(geom = "polygon",col="black",alpha=0.5)+
  geom_point(shape=21,col="black")+
  labs(x="PCA-1(~93% variation)",y="PCA-2(~6% variation)")+
  theme_classic()
```



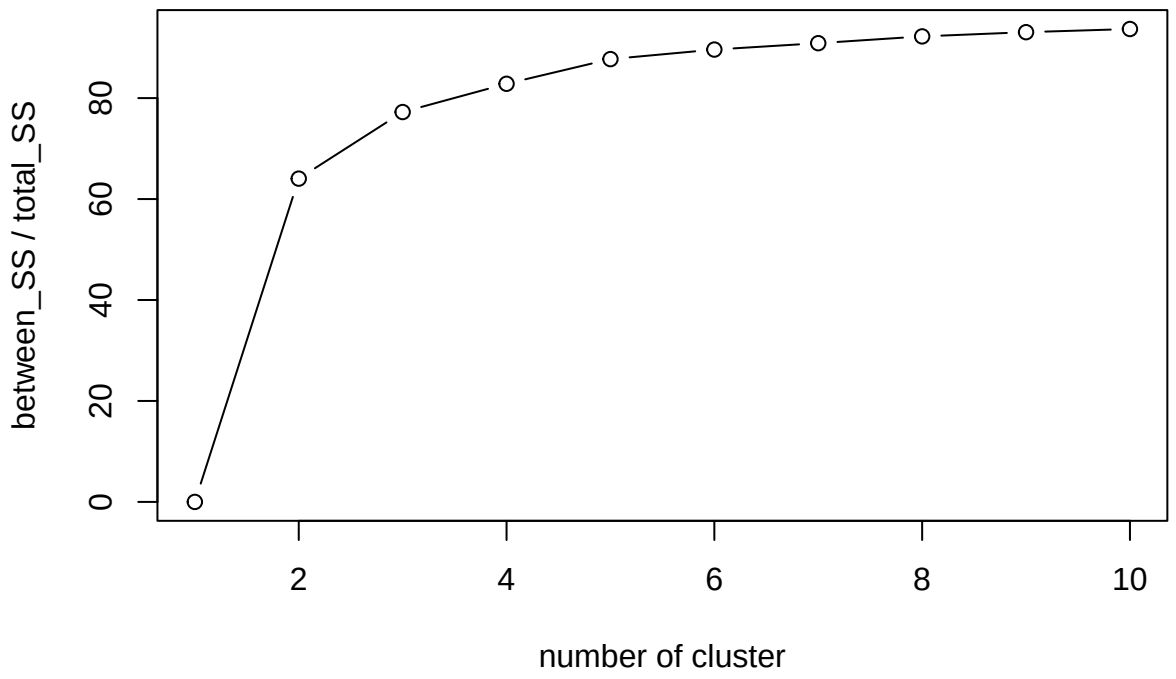
Cluster analysis.

```
cluster_data <- PCA_dataset[,8:12]
## col 14 in PCA dataset is index.
cluster_data <- scale(cluster_data)
summary(cluster_data)
```

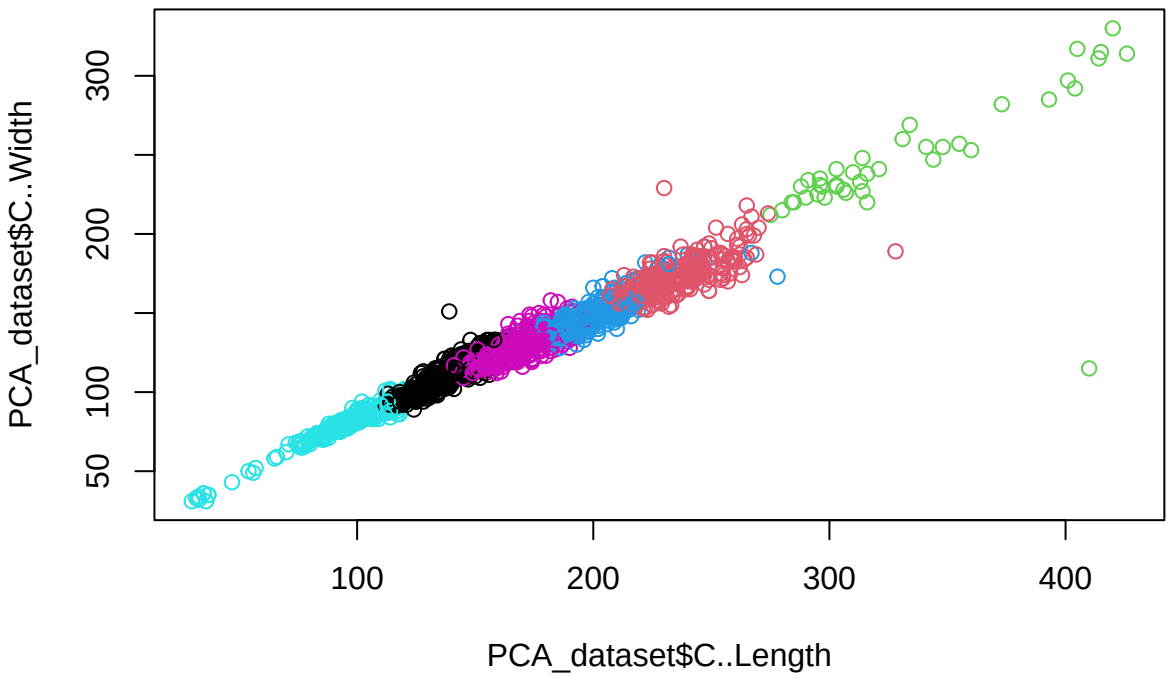
```
##    C..Length          C..Width          P..Length          P..Width
##  Min.   :-2.53734   Min.   :-2.62196   Min.   :-2.68334   Min.   :-3.23836
##  1st Qu.:-0.74389   1st Qu.:-0.72736   1st Qu.:-0.81199   1st Qu.:-0.73606
##  Median :-0.04249   Median :-0.03959   Median :-0.04735   Median :-0.05362
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.67667   3rd Qu.: 0.59626   3rd Qu.: 0.79777   3rd Qu.: 0.66132
##  Max.   : 4.49441   Max.   : 5.13812   Max.   : 3.15205   Max.   : 4.33353
##     B..Depth
##  Min.   :-2.0789
```

```
##  1st Qu.:-0.8882
##  Median :-0.1081
##  Mean   : 0.0000
##  3rd Qu.: 0.7952
##  Max.   : 2.6839
```

```
klu <- list()
for(i in 1:10){
  klu[[i]] <- kmeans(cluster_data,i)
}
ss_value <- list()
for(j in 1:length(klu)){
  ss_value [[j]]<-(klu[[j]]$betweenss/klu[[j]]$totss)*100
}
plot(x=1:length(ss_value),y=ss_value,type = "b",xlab = "number of cluster",
     ylab = "between_SS / total_SS")
```




```
plot(PCA_dataset$C..Length,PCA_dataset$C..Width,col=klu[[6]]$cluster)
```

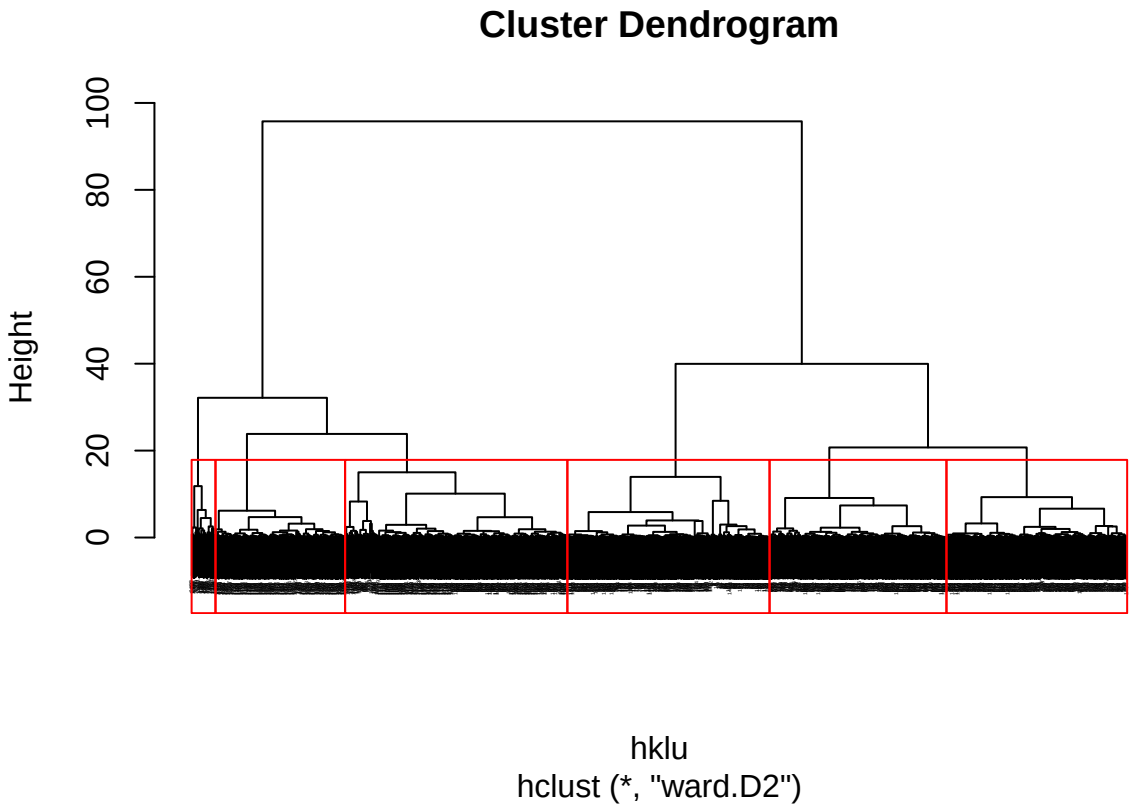


hiarchiel cluster analysis,

```
hklu <- dist(cluster_data)
fit_hklu <- hclust(hklu,method = "ward.D2")
plot(fit_hklu,cex=0.2)
sub_group <- cutree(fit_hklu,2)
table(sub_group)
```

```
## sub_group
##   1   2
## 864 580
```

```
plot(fit_hklu,cex=0.2)
rect.hclust(fit_hklu,k = 6,border = "red")
```



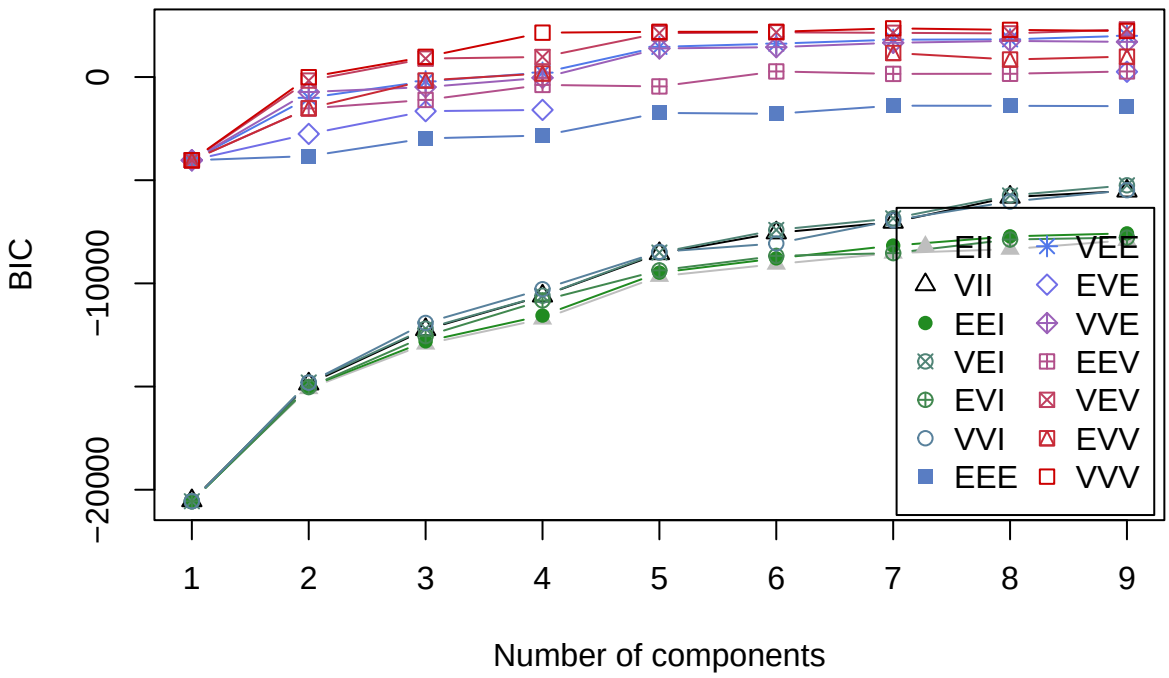

```
 ##model based cluster
library(mclust)
```

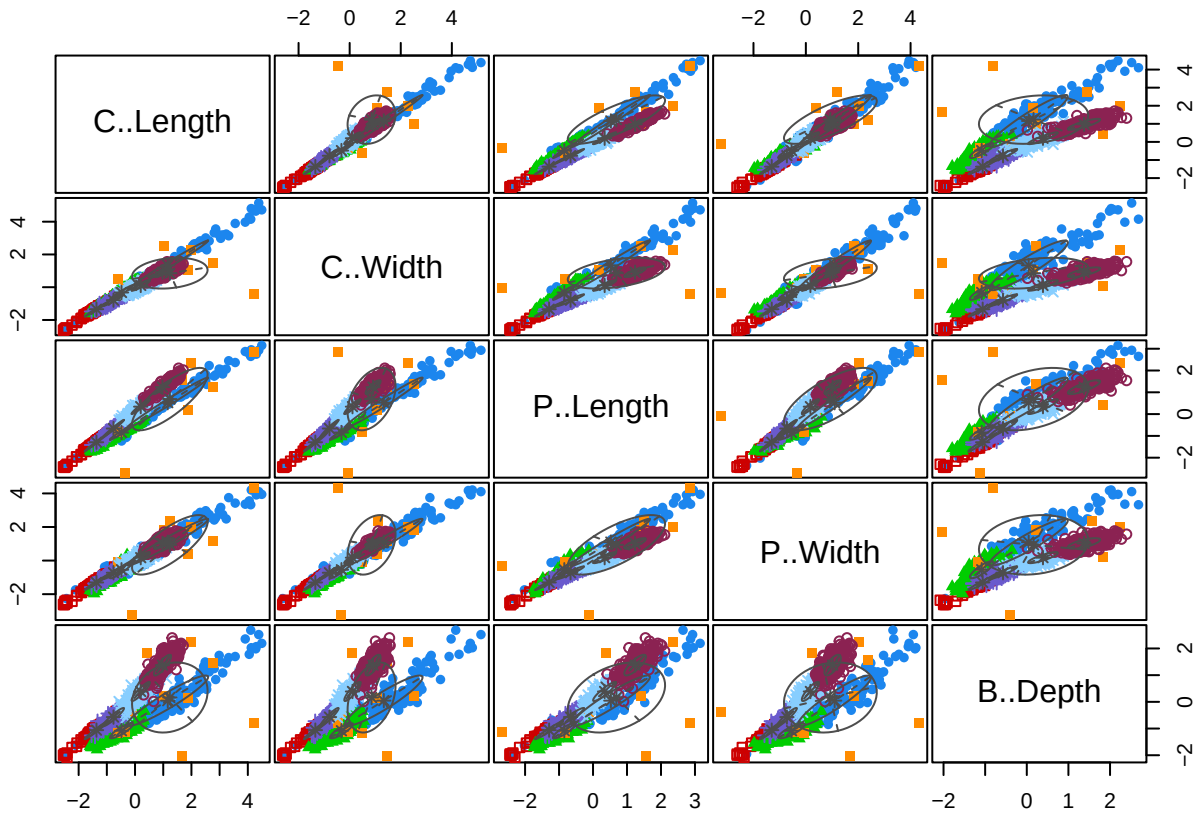
```
## Package 'mclust' version 5.4.10
## Type 'citation("mclust")' for citing this R package in publications.
```

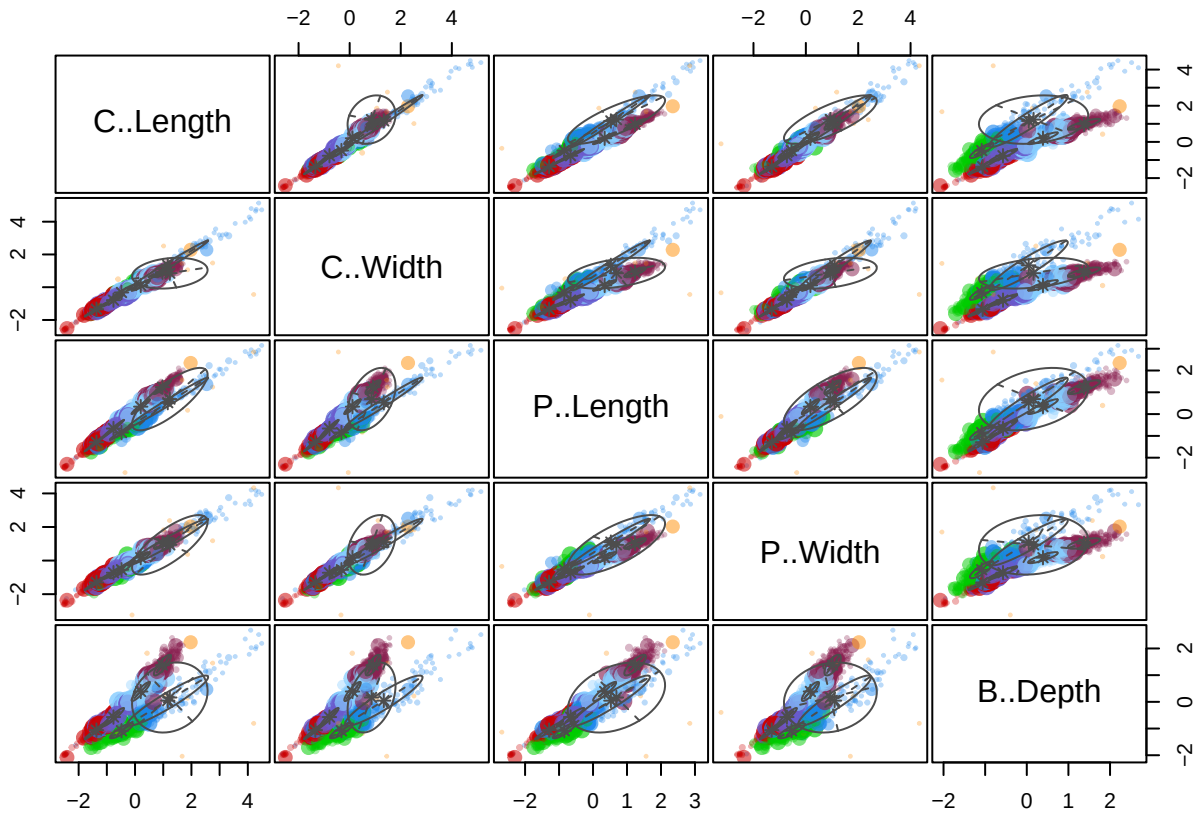
```
fit_mclust  <- Mclust(cluster_data)
fit_mclust
```

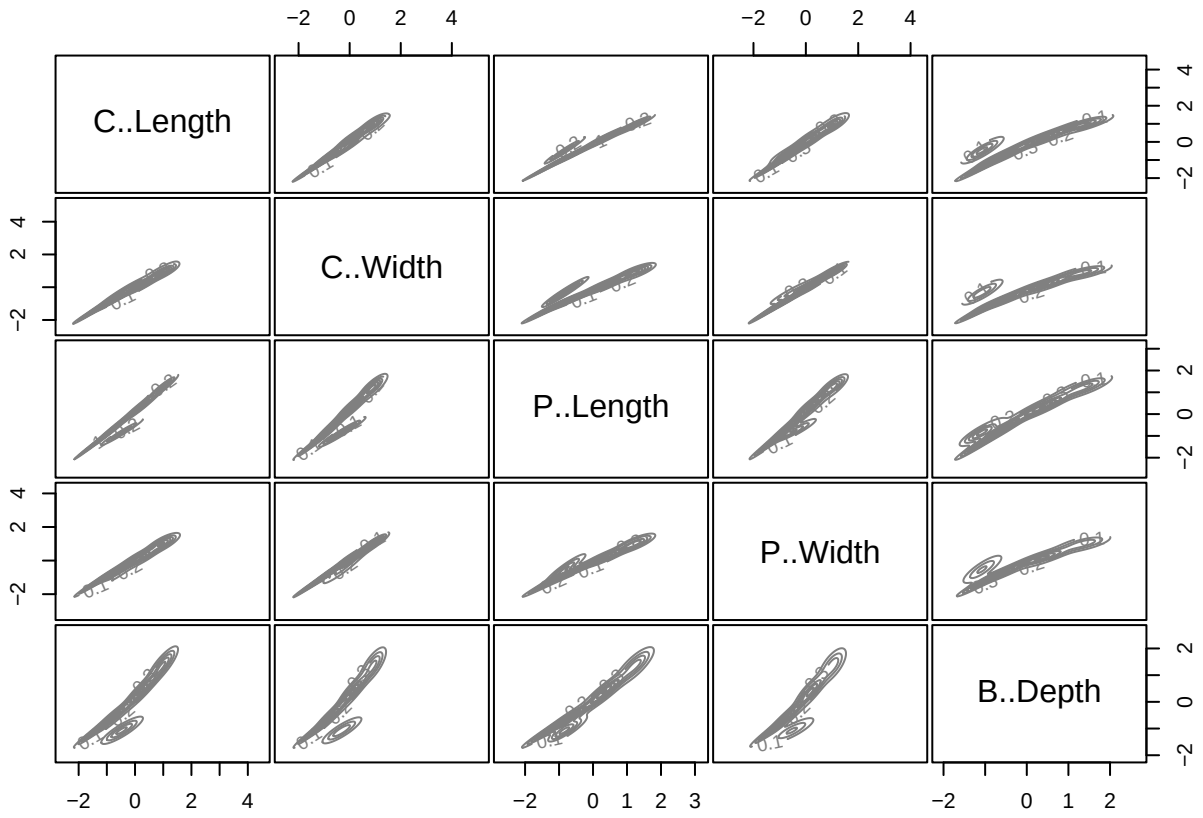
```
## 'Mclust' model object: (VVV,7)
##
## Available components:
##  [1] "call"           "data"           "modelName"      "n"
##  [5] "d"              "G"              "BIC"            "loglik"
##  [9] "df"             "bic"            "icl"            "hypvol"
## [13] "parameters"     "z"              "classification" "uncertainty"
```

```
plot(fit_mclust)
```

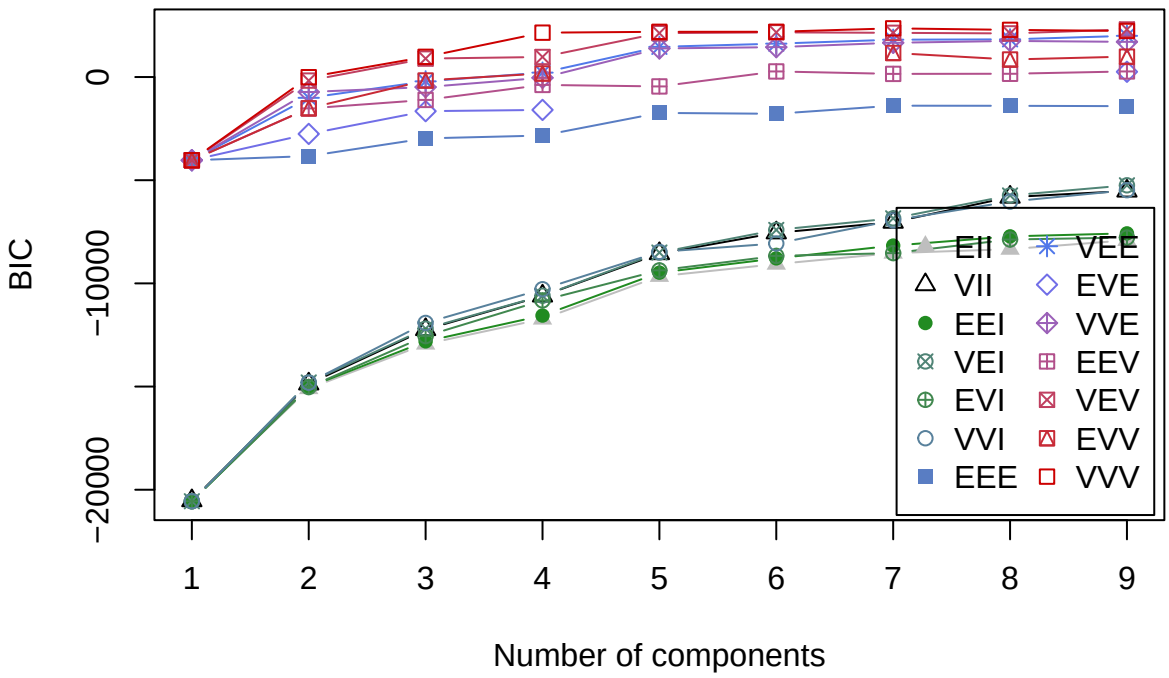

BIC
0
−10000
−20000
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV

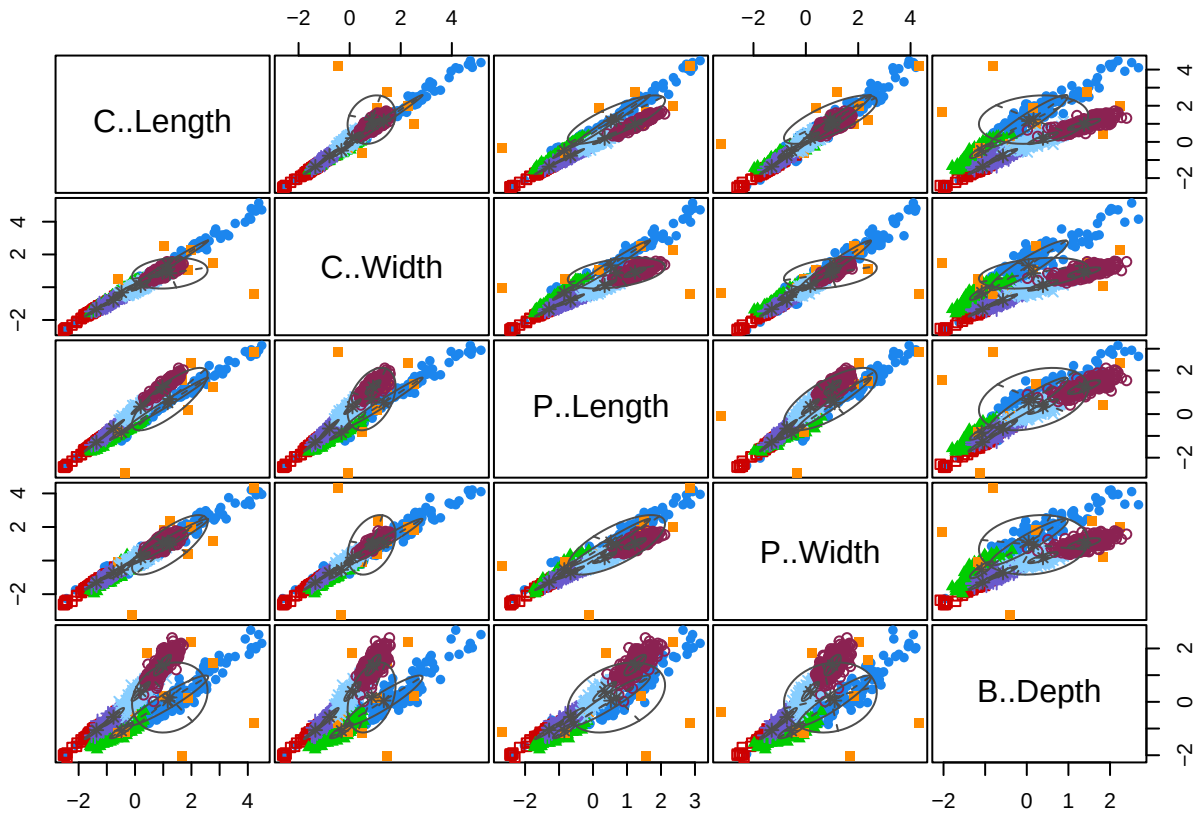

C..Length
C..Width
P..Length
P..Width
B..Depth

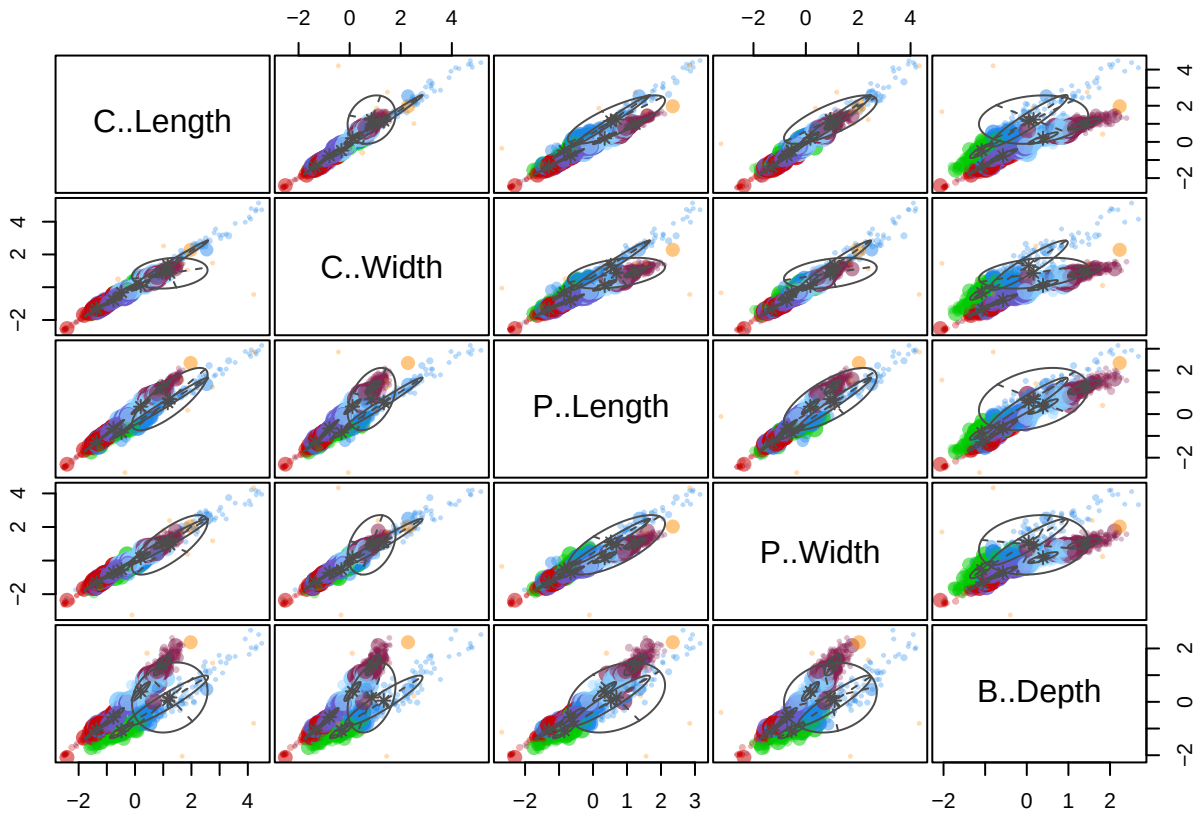

C..Length
C..Width
P..Length
P..Width
B..Depth

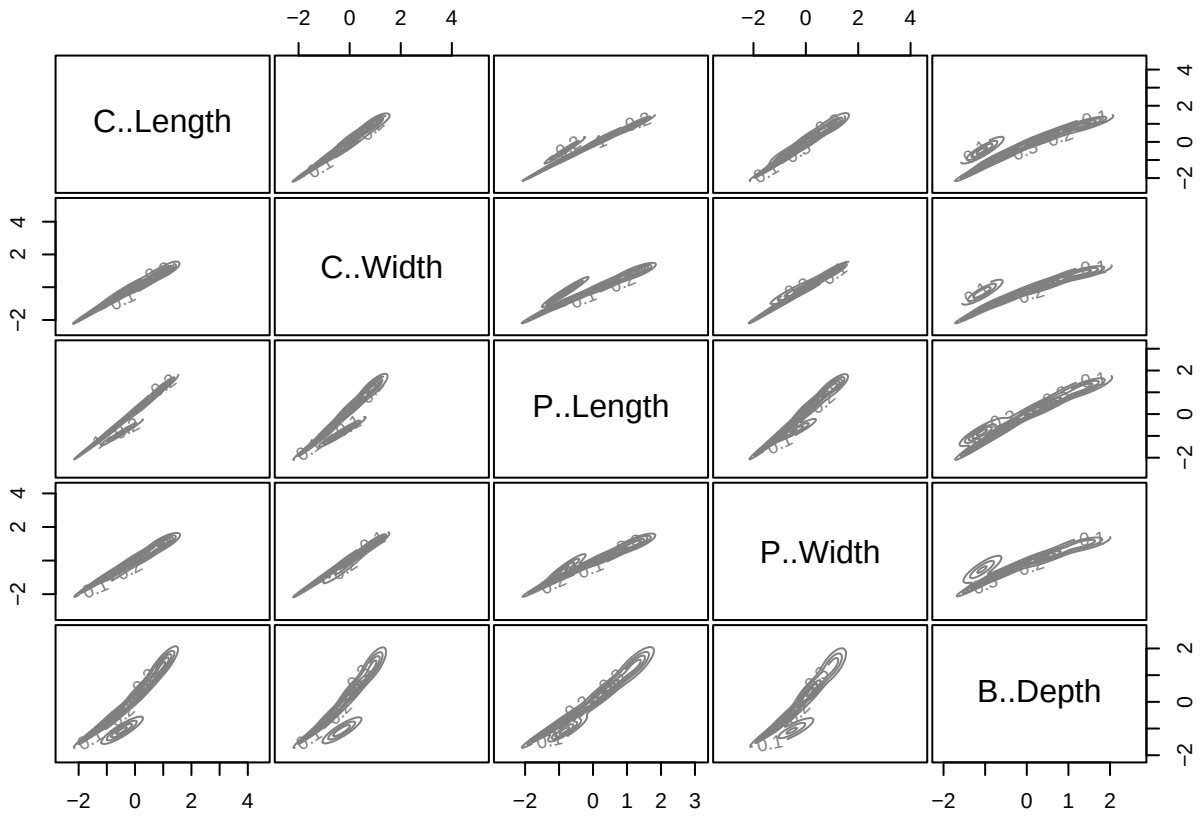



```
plot(fit_mclust)
```

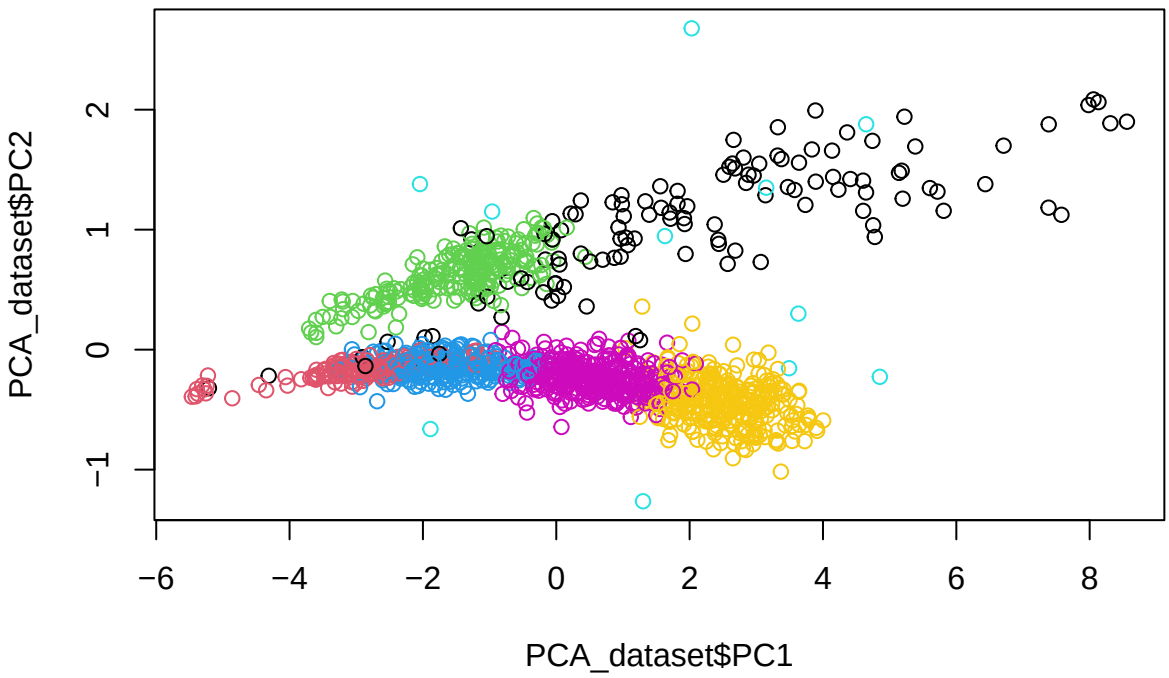

BIC
0
−10000
−20000
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV


C..Length
C..Width
P..Length
P..Width
B..Depth


C..Length
C..Width
P..Length
P..Width
B..Depth




```
plot(PCA_dataset$PC1,PCA_dataset$PC2,col=fit_mclust$classification)
```


PCA_dataset$PC2
PCA_dataset$PC1
−6
−4
−2
0
2
4
6
8
−1
0
1
2