Goal: Use the canonical iris dataset to explain how clustering can be used to group uncategorized data (In this case unknown iris species with known: sepal length, sepal width, petal lenth and petal width)

```
data("iris")

# A peek at the data
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

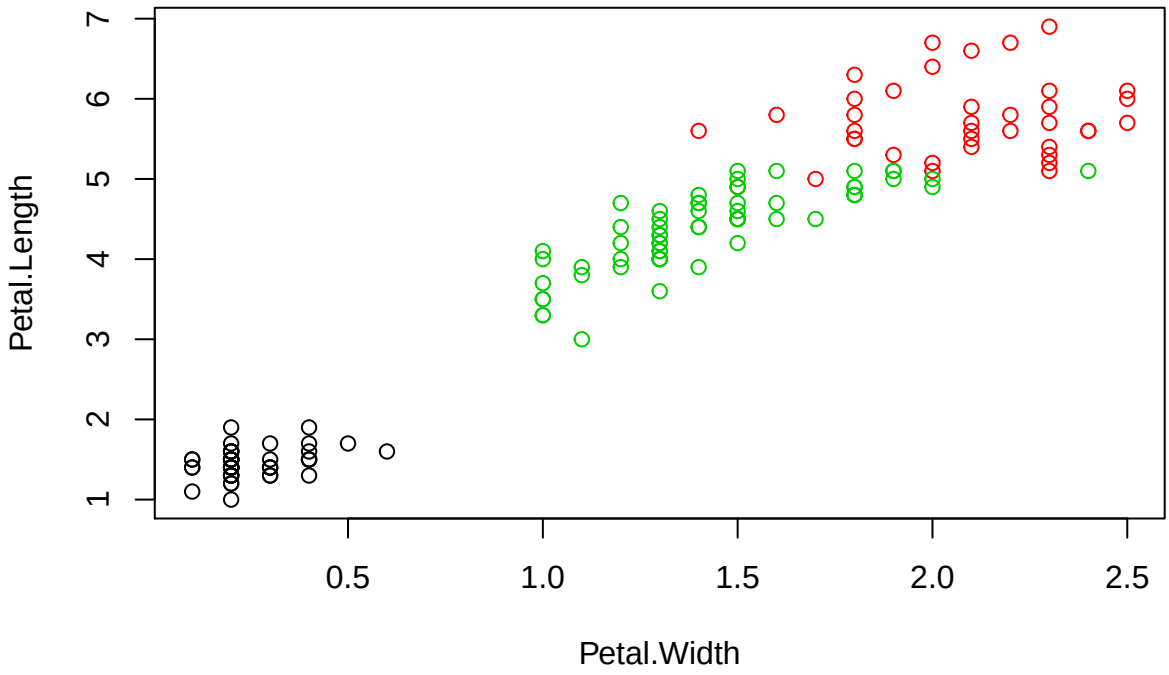
```
# Set random seed.
set.seed(1)

# Chop up iris in my_iris and species
my_iris <- iris[-5]  ##remove the species column (total=3)
species <- iris$Species

# Perform k-means clustering on my_iris: kmeans_iris
kmeans_iris <- kmeans(my_iris, 3)

# Compare the actual Species to the clustering using table()
table(species, kmeans_iris$cluster)
```

```
##
## species       1  2  3
##   setosa     50  0  0
##   versicolor  0  2 48
##   virginica   0 36 14
```

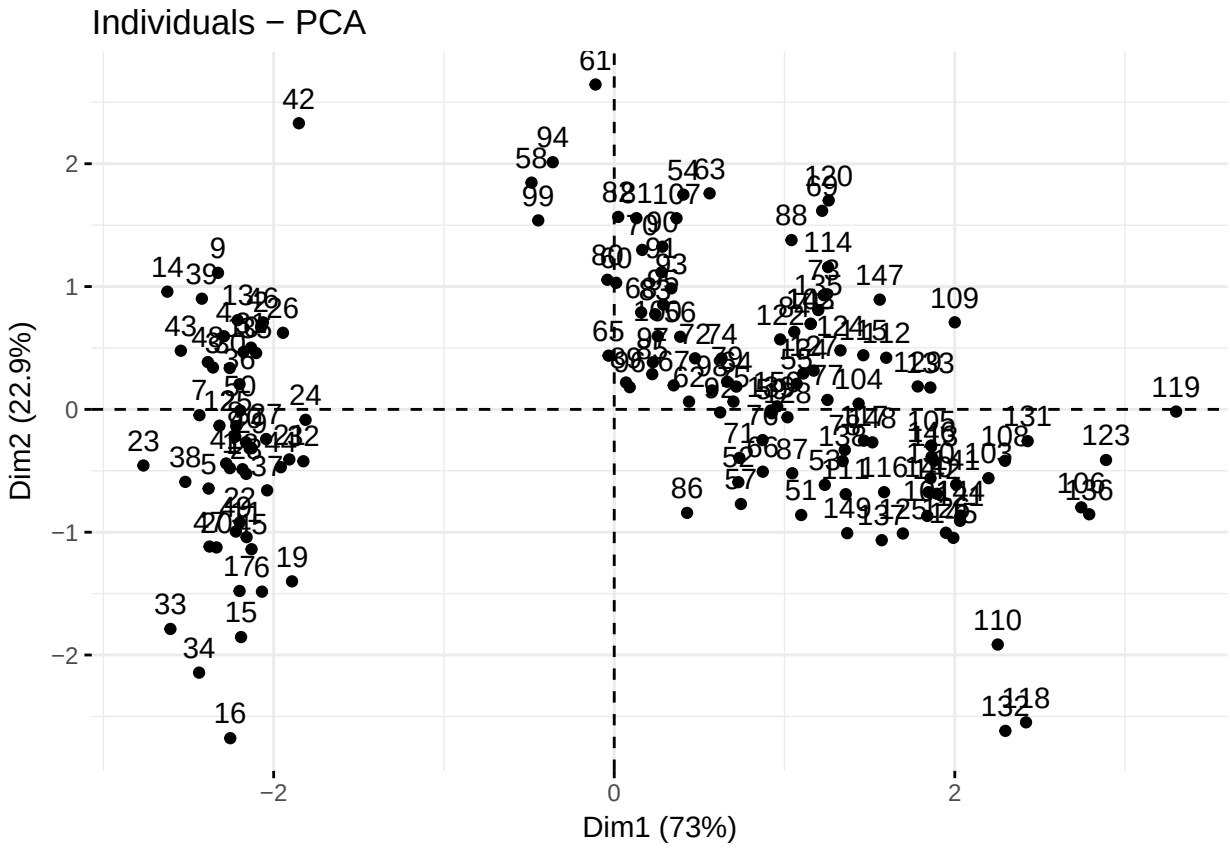
```
# Plot Petal.Width against Petal.Length, coloring by cluster
plot(Petal.Length ~ Petal.Width, data = my_iris, col = kmeans_iris$cluster)

# Principle component analysis
matrix <- scale(my_iris)  ##Scaling needed (new.x=(x-x.bar)/sd)
pr.out <- prcomp(matrix)

## library('devtools') install_github('kassambara/factoextra')
library("factoextra")
```

```
## Loading required package: ggplot2
```

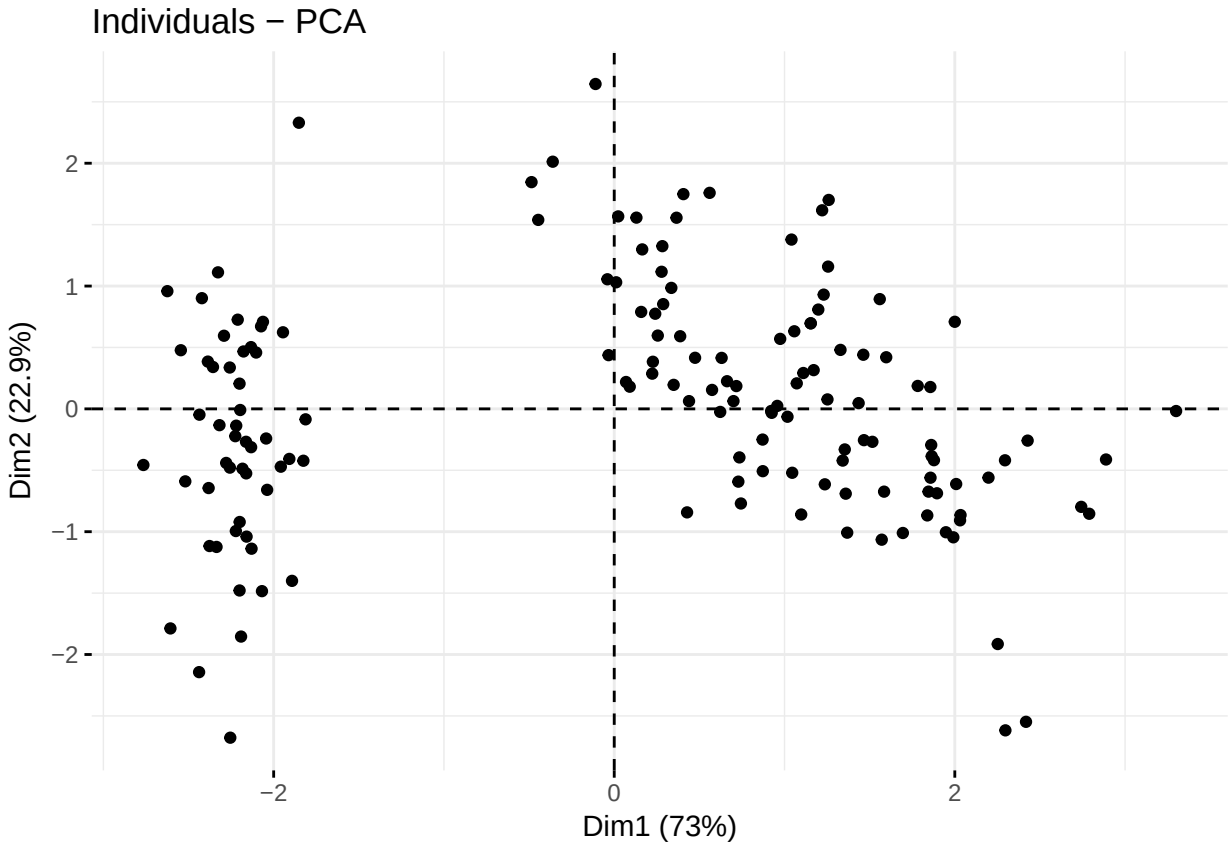
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```



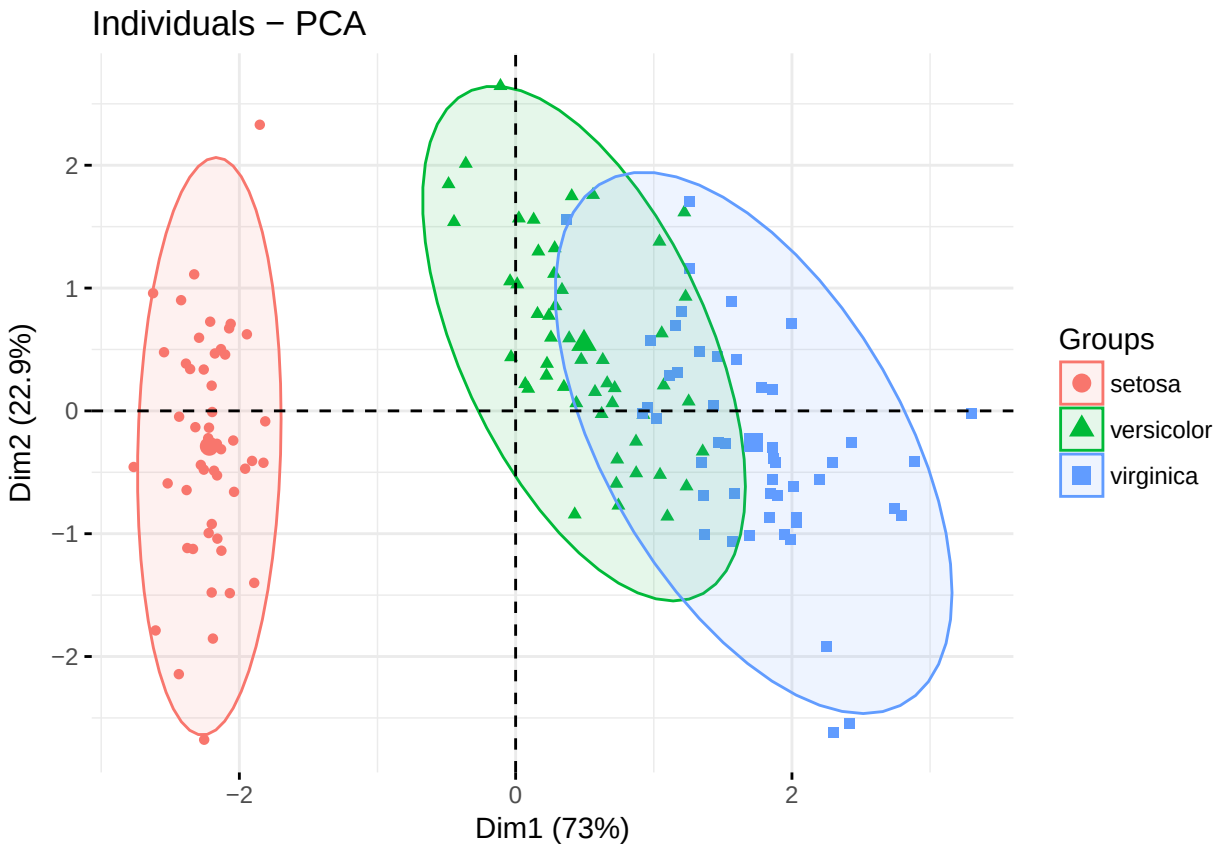

```
## The default
fviz_pca_ind(pr.out)
```




```
## Use points only
fviz_pca_ind(pr.out, geom = "point")
```




```
## Color individuals by groups and circle them
p <- fviz_pca_ind(pr.out, label = "none", habillage = iris$Species, addEllipses = TRUE,
    ellipse.level = 0.95)
print(p)
```


Individuals − PCA
Dim2 (22.9%)
Dim1 (73%)
2
1
0
−1
−2
−2
0
2
Groups
setosa
versicolor
virginica