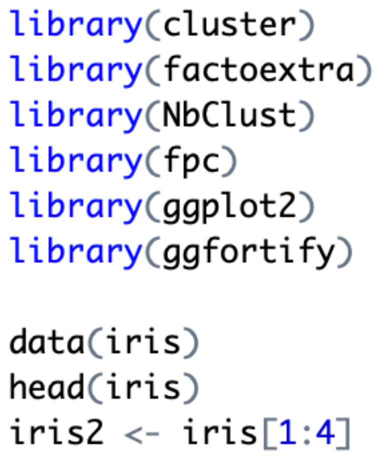
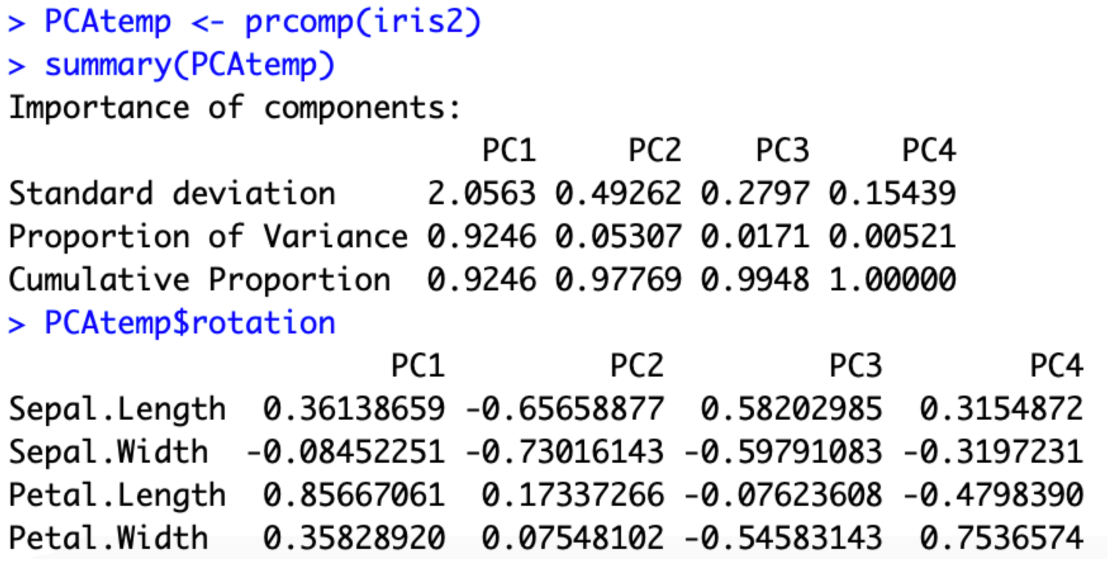


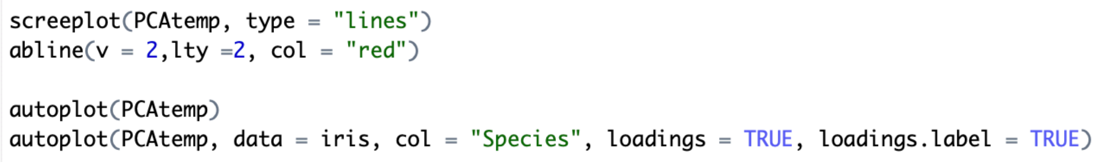
```
library(cluster)
library(factoextra)
library(NbClust)
library(fpc)
library(ggplot2)
library(ggfortify)

data(iris)
head(iris)
iris2 <- iris[1:4]
```

PCA

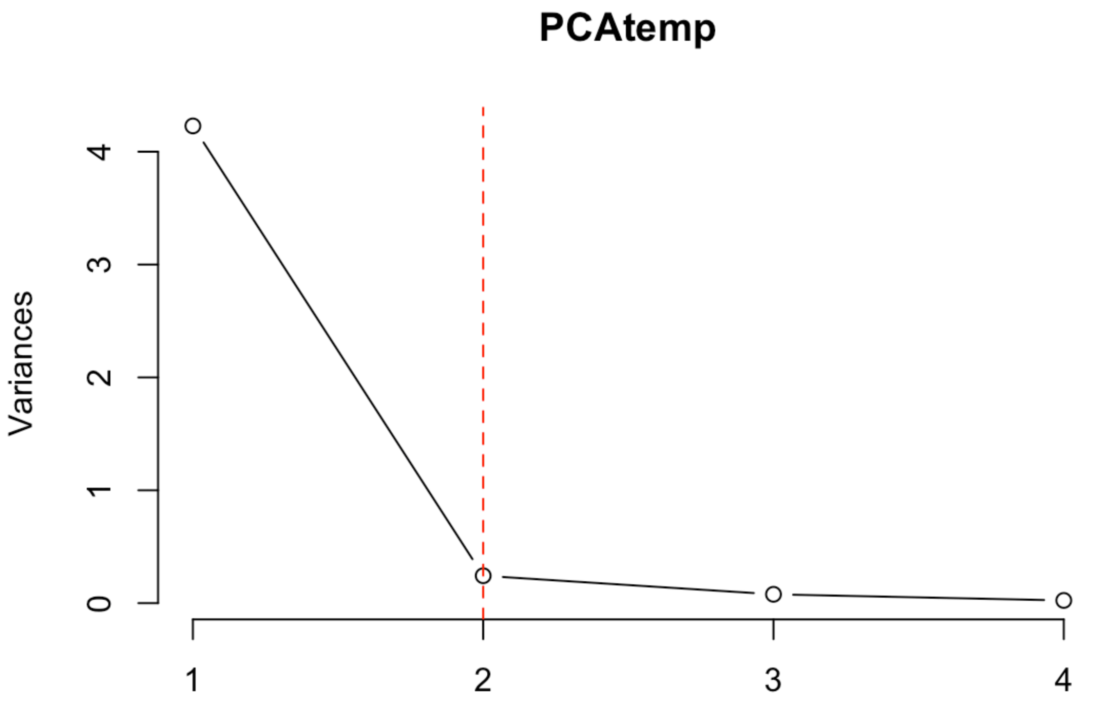


```
> PCAtemp <- prcomp(iris2)
> summary(PCAtemp)
Importance of components:
                          PC1     PC2    PC3     PC4
Standard deviation     2.0563 0.49262 0.2797 0.15439
Proportion of Variance 0.9246 0.05307 0.0171 0.00521
Cumulative Proportion  0.9246 0.97769 0.9948 1.00000
> PCAtemp$rotation
                     PC1         PC2         PC3        PC4
Sepal.Length  0.36138659 -0.65658877  0.58202985  0.3154872
Sepal.Width  -0.08452251 -0.73016143 -0.59791083 -0.3197231
Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798390
Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536574
```

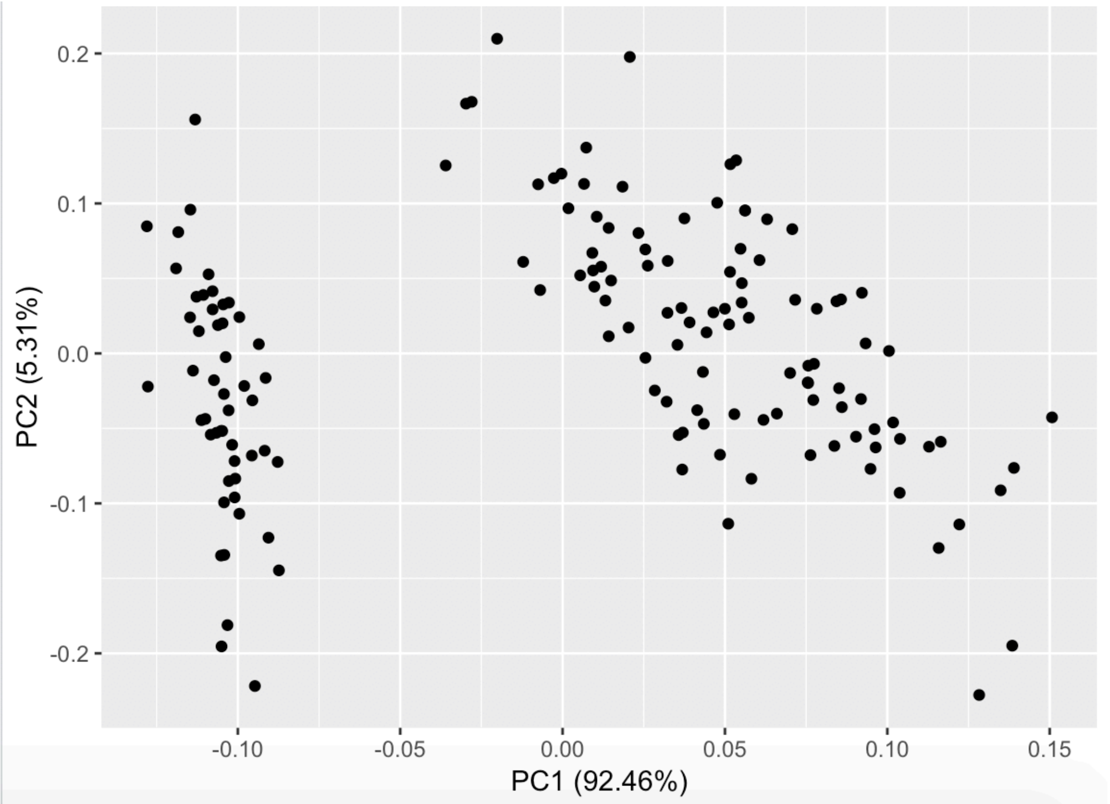
The summary shows that the first two Principal Components account for most of the variation in all four components. "Petal. Length" has a large weight in principal component 1, while "Sepal. Width" has a small weight. The second principal component is mostly "Sepal. Width" and "Sepal. Length" with small weightings on the other variables.

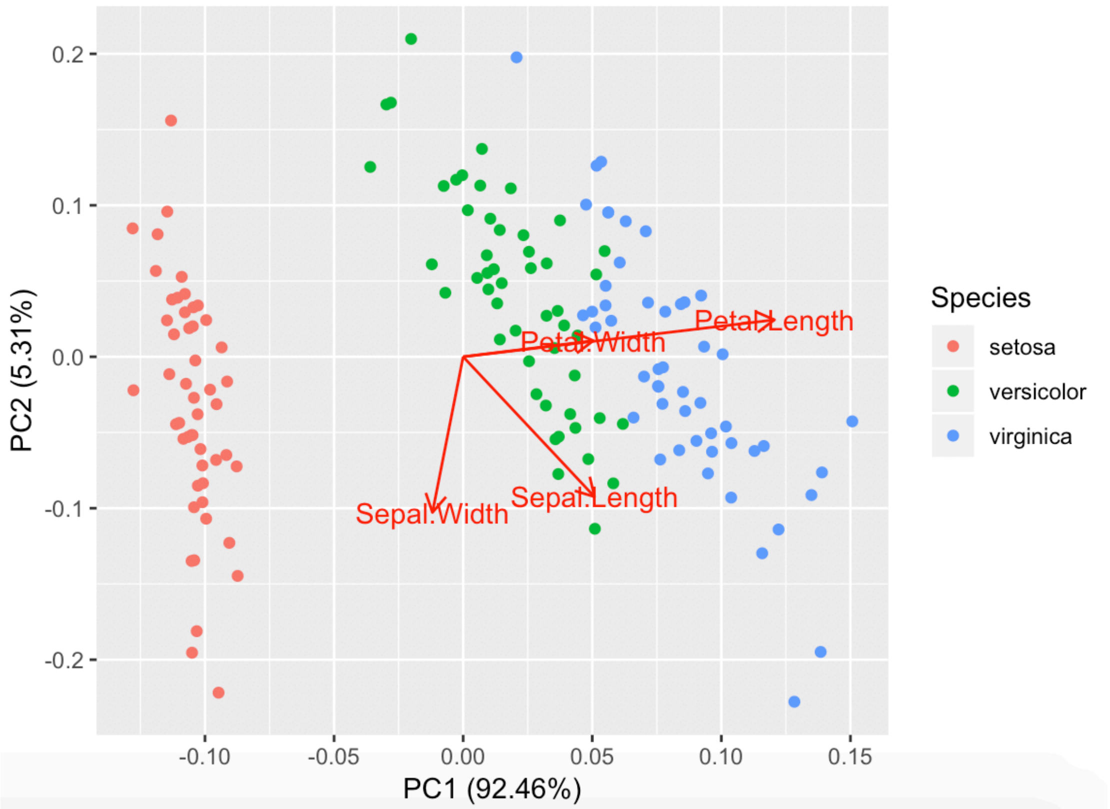


```
screeplot(PCAtemp, type = "lines")
abline(v = 2,lty =2, col = "red")

autoplot(PCAtemp)
autoplot(PCAtemp, data = iris, col = "Species", loadings = TRUE, loadings.label = TRUE)
```



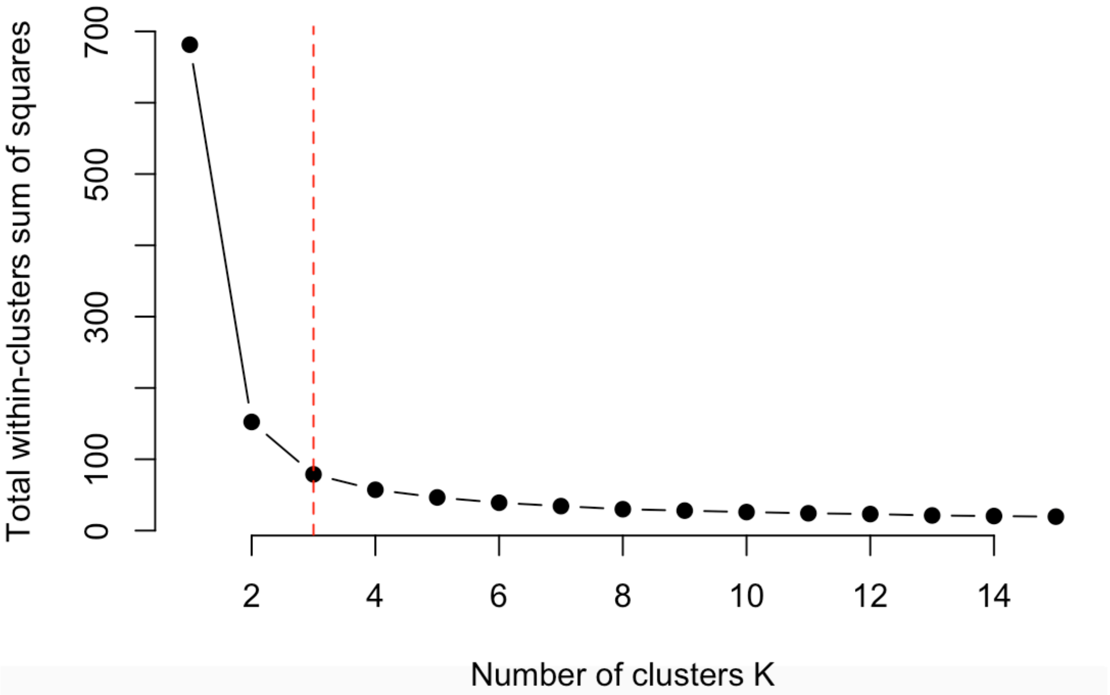
As can be seen from the figure, after the second principal component, the line changes tend to be stable, so the first two principal components can be selected for analysis.



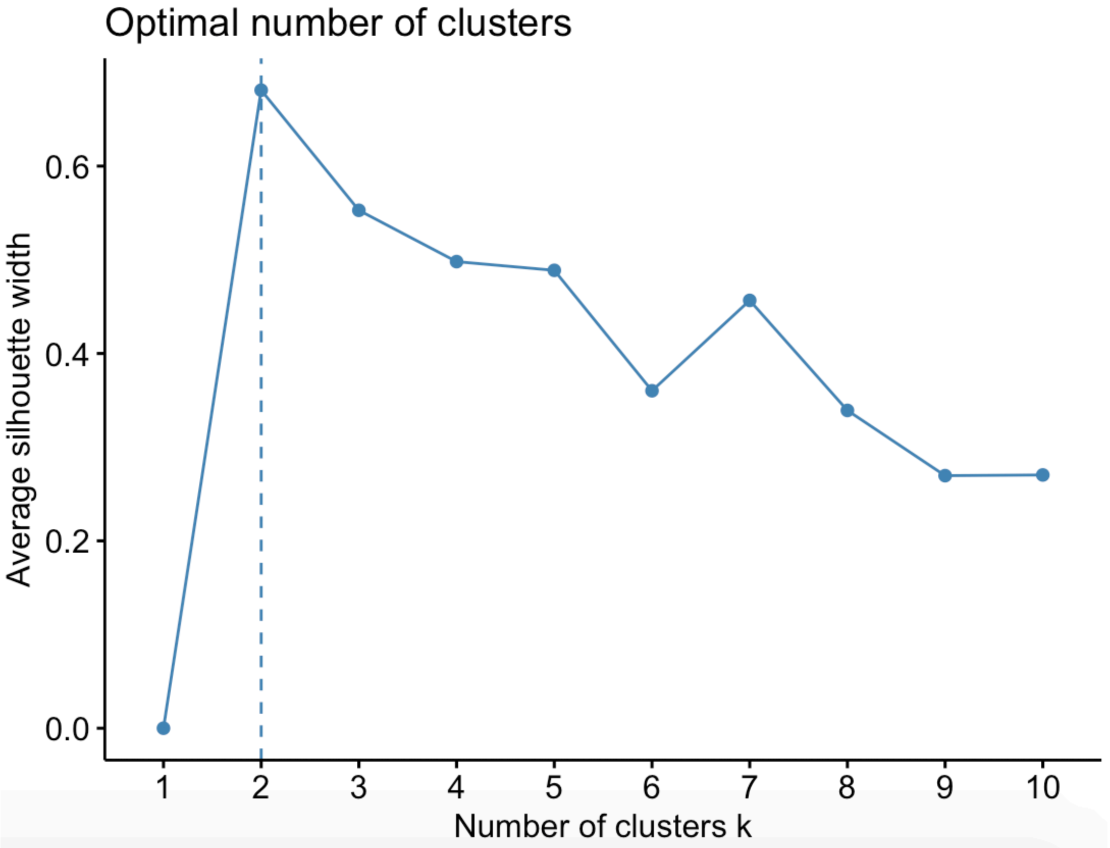


According to the graph, the first two principal components can be used to roughly distinguish the three species of Iris.

```
# K-means

k.max <- 15
wss <- sapply(1:k.max, function(k){kmeans(iris2, k, nstart = 10)$tot.withinss})
plot(1:k.max,wss,type = "b",pch = 19, frame = FALSE,
     xlab = "Number of clusters K",ylab = "Total within-clusters sum of squares")
abline(v = 3,lty =2, col = "red")

fviz_nbclust(iris2, kmeans, method = c("silhouette"))

iris.kmeans <- kmeans(iris2,3)
autoplot(iris.kmeans, data = iris2, frame = TRUE, frame.type = "norm")

dis <- dist(iris2)^2
sil <- silhouette(iris.kmeans$cluster, dis)
plot(sil)

plot(iris$Species,iris.kmeans$cluster)
```

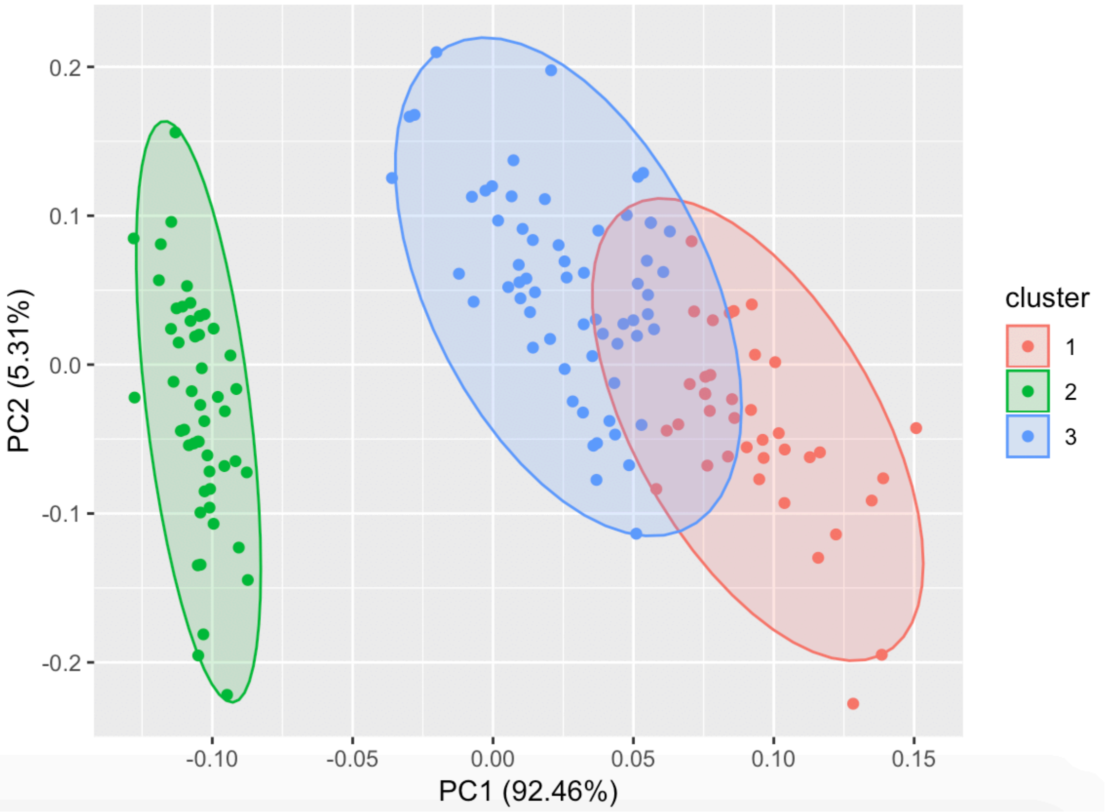


The leaving off at the number of clusters being 3 indicates this would be a good choice for the number of clusters.

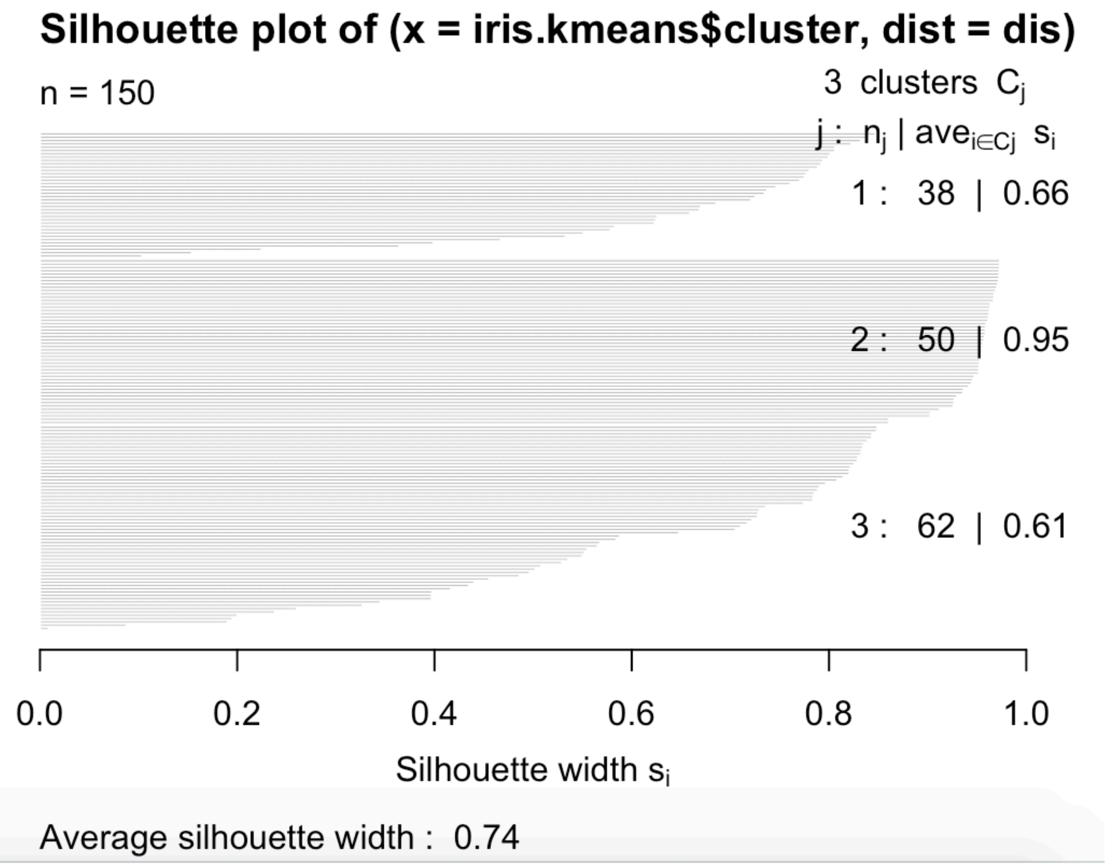


The average silohouette is maximum for K = 2.

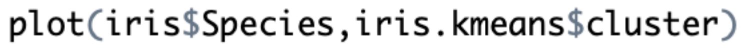
According to the above two methods, it is considered feasible to take k as 2 or 3. However, we have 3 labels, and it is finally decided that 3 clusters is appropriate.



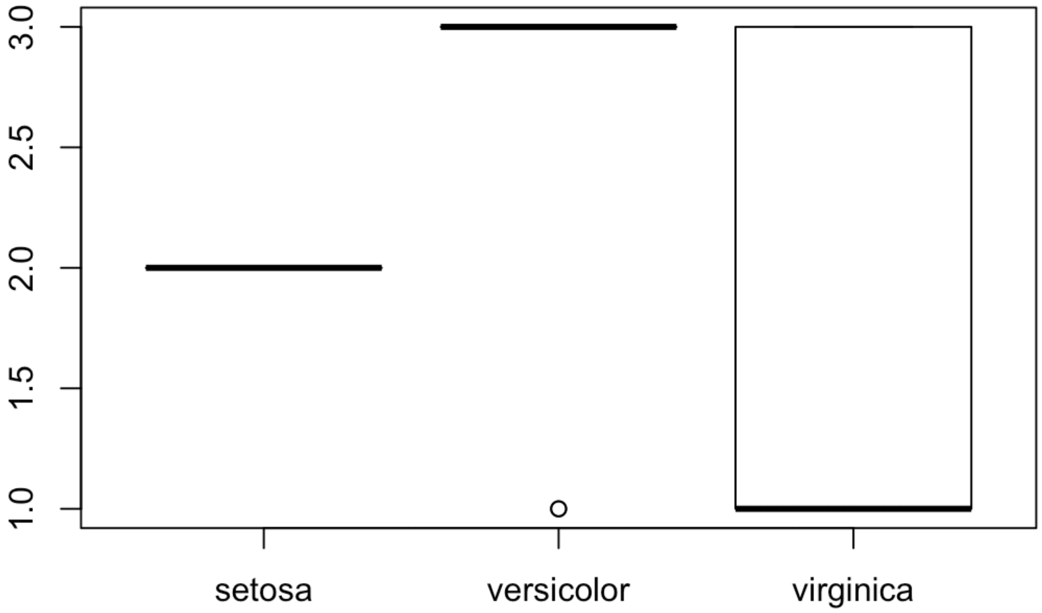
As shown in the figure, k-means has a good discrimination degree for cluster 1 and a poor discrimination degree for cluster 2 and 3.



There are no negative silohouette values and most of the species in each cluster have a moderately silohouette value.



```
plot(iris$Species,iris.kmeans$cluster)
```



It can be seen from the figure that k-means has a high identification degree for species setosa and versicolor, and a poor identification ability for species virginica.