# input data
import state_income.csv

#Sort data
tmp = sort(state_income$V2)

# create 4 clusters
kmeans=kmeans(tmp,4,15)

#Visualize the cluster centers
points(kmeans$centers, col = 1:4, pch=20)
# create another “dimension”
off = rnorm(51, mean=0, sd = 10)

# look at data
off

# scale data
off = off * 200
# visualize as xy
plot(off ~ state_income$V2)

# convert to matrix
cbind(state_income$V2,off)

# create 3 clusters via k-means
kmeans=kmeans(cbind(state_income$V2,off),3,15)
# exam kmeans object

kmeans

# note: clustering statistics

- cluster sizes
- cluster means
- clustering vector
- within cluster sum of squares
# Visualize clusters
plot(state_income$V2, off, col=kmeans$cluster)

# Visualize the cluster centers
points(kmeans$centers, col = 1:3, pch=20)
Is 3 the best number of clusters?

# explore different number of clusters

withinSumSqrs = numeric(20)

for (k in 1:20) withinSumSqrs[k] = sum(kmeans(cbind(state_income$V2,off),centers=k)$withinss)

# Visualize within cluster sum of square

plot(1:20, withinSumSqrs, type="b", xlab="# Clusters", ylab="Within sum of square")