6.4

a. use range to standardize the data

rge<- sapply(crime, function(x) diff(range(x)))

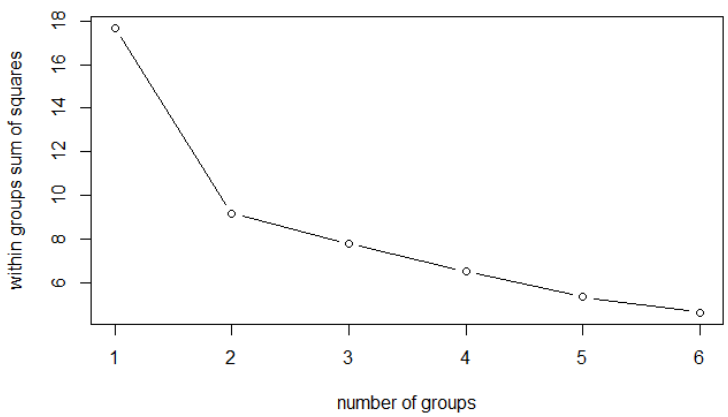
crime_s <- sweep(crime, 2, rge, FUN = "/")

(cluster<-kmeans(crime_s,centers=3))

for (i in 1:6)

wss[i]<-sum(kmeans(crime_snew,centers=i)$withinss)

plot(1:6,wss, type="b", xlab="number of groups",ylab="within groups sum of squares")




```
> rge<- sapply(crime, function(x) diff(range(x)))
> crime_s <- sweep(crime, 2, rge, FUN = "/")
> (cluster<-kmeans(crime_s,centers=3))
K-means clustering with 3 clusters of sizes 15, 13, 23

Cluster means:
     Murder      Rape    Robbery   Assault  Burglary     Theft   Vehicle
1 0.2522222 0.5851609 0.27594824 0.5231656 0.6739651 0.9008955 0.6010654
2 0.3902564 0.8603802 0.34898569 0.6825109 0.9553796 1.2955479 0.6443625
3 0.1508696 0.3738703 0.08020488 0.2609379 0.4782846 0.8438676 0.2407187

Clustering vector:
ME NH VT MA RI CT NY NJ PA OH IN IL MI WI MN IA MO ND SD NE KS DE MD DC VA WV NC SC GA FL KY TN AL MS AR LA OK
 3  3  3  1  1  1  1  1  3  1  3  1  2  3  3  3  1  3  3  3  3  1  1  2  3  3  3  1  1  2  3  1  1  3  3  2  1
TX MT ID WY CO NM AZ UT NV WA OR CA AK HI
 2  3  3  3  2  2  2  3  2  2  2  2  2  3

Within cluster sum of squares by cluster:
[1] 2.131020 3.051491 2.159315
 (between_SS / total_SS =  58.4 %)
```

b. use standard deviation to standardize the data

sd<- sapply(crime, function(x) sd(x))

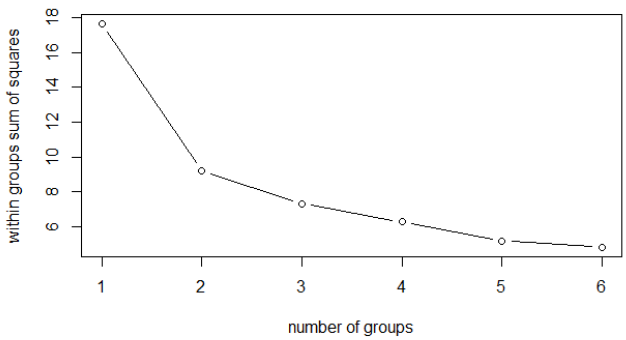
crime_s <- sweep(crime, 2, sd, FUN = "/")

(cluster<-kmeans(crime_snew,centers=3))

for (i in 1:6)

wss[i]<-sum(kmeans(crime_snew,centers=i)$withinss)

plot(1:6,wss, type="b", xlab="number of groups",ylab="within groups sum of squares")




```
K-means clustering with 3 clusters of sizes 13, 23, 15

Cluster means:
     Murder     Rape   Robbery  Assault Burglary    Theft   Vehicle
1 2.4305647 3.607813 1.8915908 2.926259 4.158579 5.116538 2.5241626
2 0.9396341 1.567742 0.4347308 1.118769 2.081879 3.332707 0.9429678
3 1.5708709 2.453742 1.4957094 2.243068 2.933638 3.557928 2.3545550

Clustering vector:
ME NH VT MA RI CT NY NJ PA OH IN IL MI WI MN IA MO ND SD NE KS DE MD DC VA WV NC SC GA FL
 KY TN AL MS AR LA OK TX MT
 2  2  2  3  3  3  3  3  2  3  2  3  1  2  2  2  3  2  2  2  2  3  3  1  2  2  2  3  3  1
  2  3  3  2  2  1  3  1  2
ID WY CO NM AZ UT NV WA OR CA AK HI
 2  2  1  1  1  2  1  1  1  1  1  2

Within cluster sum of squares by cluster:
[1] 71.54982 40.23507 42.24884
 (between_SS / total_SS =  56.0 %)

c. Compare two results.
```

Type of scale	The # of clusters	The # of items in each cluster	between_SS / total_SS
by range	3	13, 23, 15	58.4%
by sd	3	13, 23, 15	56%

If the data scaled by range, the good fit of the three clusters to the original data set is 58.4%. By assigning the samples to 3 clusters rather than 51 clusters achieved a reduction in sums of squares of 58.4%. If the data scaled by standard deviation, the good fit of the three clusters to the original data set is 56.4%. By assigning the samples to 3 clusters rather than 51 clusters achieved a reduction in sums of squares of 56.4%. In general, divide each variable by its **sample range (max – min);** Milligan and Cooper (1988) found that this approach best preserved the clustering structure.