BDA Prac 1 K-Means algorithm

Code

Hide

```
install.packages("plyr")
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please downlo
ad and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/User/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/plyr_1.8.8.zip'
Content type 'application/zip' length 1153074 bytes (1.1 MB)
downloaded 1.1 MB
```

```
package ‘plyr’ successfully unpacked and MD5 sums checked
Warning in install.packages :
  cannot remove prior installation of package ‘plyr’
Warning in install.packages :
  problem copying C:\Users\User\AppData\Local\R\win-library\4.2\00LOCK\plyr\libs\x64\plyr.dll
to C:\Users\User\AppData\Local\R\win-library\4.2\plyr\libs\x64\plyr.dll: Permission denied
Warning in install.packages :
  restored ‘plyr’

The downloaded binary packages are in
    C:\Users\User\AppData\Local\Temp\RtmpiwDCp8\downloaded_packages
```

Hide

```
install.packages("ggplot2")
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please downlo
ad and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/User/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggplot2_3.4.1.zip'
Content type 'application/zip' length 4226768 bytes (4.0 MB)
downloaded 4.0 MB
```

```
package ‘ggplot2’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\User\AppData\Local\Temp\RtmpiwDCp8\downloaded_packages
```

Hide

```
install.packages("cluster")
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please downlo
ad and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/User/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/cluster_2.1.4.zip'
Content type 'application/zip' length 585941 bytes (572 KB)
downloaded 572 KB
```

```
package ‘cluster’ successfully unpacked and MD5 sums checked
Warning in install.packages :
  cannot remove prior installation of package ‘cluster’
Warning in install.packages :
  problem copying C:\Users\User\AppData\Local\R\win-library\4.2\00LOCK\cluster\libs\x64\clust
er.dll to C:\Users\User\AppData\Local\R\win-library\4.2\cluster\libs\x64\cluster.dll: Permiss
ion denied
Warning in install.packages :
  restored ‘cluster’

The downloaded binary packages are in
    C:\Users\User\AppData\Local\Temp\RtmpiwDCp8\downloaded_packages
```

Hide

```
install.packages("lattice")
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please downlo
ad and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/User/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/lattice_0.20-45.zip'
Content type 'application/zip' length 1171761 bytes (1.1 MB)
downloaded 1.1 MB
```

```
package ‘lattice’ successfully unpacked and MD5 sums checked
Warning in install.packages :
  cannot remove prior installation of package ‘lattice’
Warning in install.packages :
  problem copying C:\Users\User\AppData\Local\R\win-library\4.2\00LOCK\lattice\libs\x64\latti
ce.dll to C:\Users\User\AppData\Local\R\win-library\4.2\lattice\libs\x64\lattice.dll: Permiss
ion denied
Warning in install.packages :
  restored ‘lattice’

The downloaded binary packages are in
    C:\Users\User\AppData\Local\Temp\RtmpiwDCp8\downloaded_packages
```

Hide

```
install.packages("grid")
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please downlo
ad and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/User/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
```

```
Warning in install.packages :
  package ‘grid’ is a base package, and should not be updated
```

Hide

```
install.packages("gridExtra")
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please downlo
ad and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/User/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/gridExtra_2.3.zip'
Content type 'application/zip' length 1109591 bytes (1.1 MB)
downloaded 1.1 MB
```

```
package ‘gridExtra’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\User\AppData\Local\Temp\RtmpiwDCp8\downloaded_packages
```

Hide

```
library(plyr)
library(ggplot2)
library(cluster)
library(lattice)
library(grid)
library(gridExtra)
```

Hide

```
grade_input=as.data.frame(read.csv("F:/GitHub/Practical_BscIT_MscIT_Ninad/MscIT/Semester 2/Bi
gDataAnalytics/Dataset/grades_km_input.csv"))
kmdata_orig=as.matrix(grade_input[, c ("Student","English","Math","Science")])
kmdata=kmdata_orig[,2:4]
kmdata[1:10,]
```

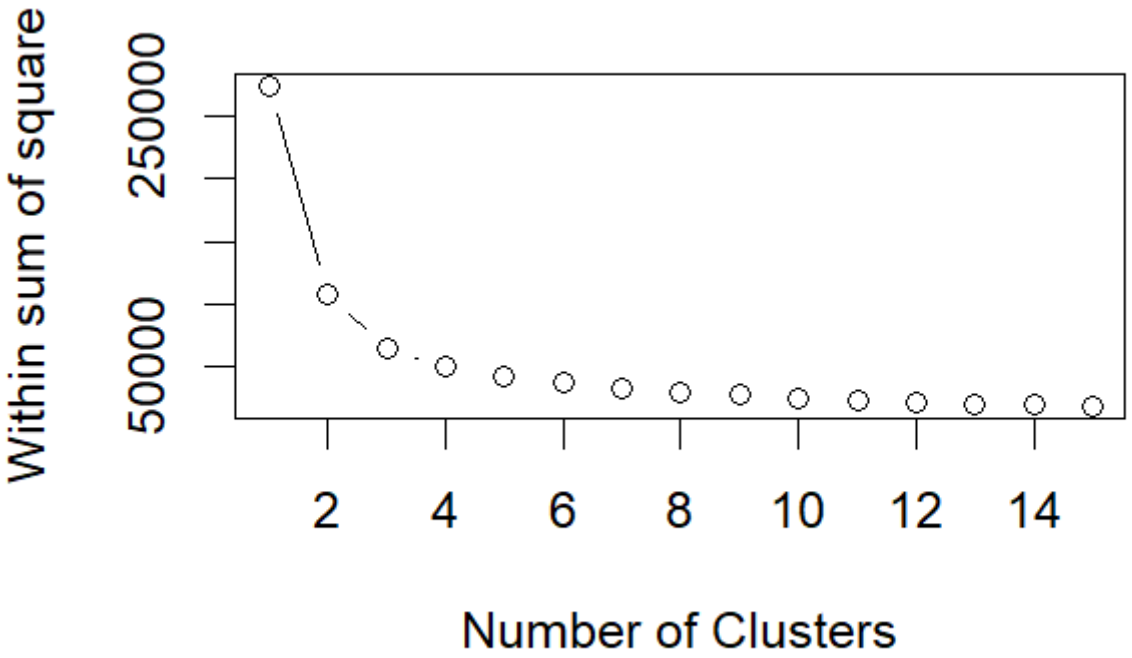
```
      English Math Science
 [1,]      99   96      97
 [2,]      99   96      97
 [3,]      98   97      97
 [4,]      95  100      95
 [5,]      95   96      96
 [6,]      96   97      96
 [7,]     100   96      97
 [8,]      95   98      98
 [9,]      98   96      96
[10,]      99   99      95
```

Hide

```
# the k-means algorithm is used to identify clusters for k = 1, 2, .. . , 15. For each value
of k, the WSS is calculated.
wss=numeric(15)
```

Hide

```
# the option n start=25 specifies that the k-means algorithm will be repeated 25 times, each
starting with k random initial centroids
for(k in 1:15)wss[k]=sum(kmeans(kmdata,centers=k,nstart=25)$withinss)
plot(1:15,wss,type="b",xlab="Number of Clusters",ylab="Within sum of square")
```



Hide

```
#As can be seen, the WSS is greatly reduced when k increases from one to two. Another substan
tial reduction in WSS occurs at k = 3. However, the improvement in WSS is fairly linear fork
> 3.
km = kmeans(kmdata,3,nstart=25)
km
```

```
K-means clustering with 3 clusters of sizes 158, 218, 244

Cluster means:
   English     Math  Science
1 97.21519 93.37342 94.86076
2 73.22018 64.62844 65.84862
3 85.84426 79.68033 81.50820

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [31] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [61] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [91] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[121] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[151] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 1 1 3
[181] 3 1 3 3 3 1 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[211] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[241] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[271] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[301] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[331] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[361] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 3 2 3 2 3 3 3
[391] 2 2 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[421] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2
[451] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 3 2 2 2 2 2 2 2 3 2
[481] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[511] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[541] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[571] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[601] 3 3 2 2 3 3 3 3 1 1 3 3 3 2 2 3 2 3 3 3

Within cluster sum of squares by cluster:
[1]  6692.589 34806.339 22984.131
 (between_SS / total_SS =  76.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Hide

```
c( wss[3] , sum(km$withinss))
```

```
[1] 64483.06 64483.06
```

Hide

```
df=as.data.frame(kmdata_orig[,2:4])
df$cluster=factor(km$cluster)
centers=as.data.frame(km$centers)
```

Hide

```
g1=ggplot(data=df, aes(x=English, y=Math, color=cluster )) + geom_point() + theme(legend.posi
tion="right") + geom_point(data=centers,aes(x=English,y=Math, color=as.factor(c(1,2,3))),size
=10, alpha=.3, show.legend =FALSE)

g2=ggplot(data=df, aes(x=English, y=Science, color=cluster )) + geom_point () +geom_point(dat
a=centers,aes(x=English,y=Science, color=as.factor(c(1,2,3))),size=10, alpha=.3, show.legend=
FALSE)

g3 = ggplot(data=df, aes(x=Math, y=Science, color=cluster )) + geom_point () + geom_point(dat
a=centers,aes(x=Math,y=Science, color=as.factor(c(1,2,3))),size=10, alpha=.3, show.legend=FAL
SE)
tmp=ggplot_gtable(ggplot_build(g1))
```

Hide

```
grid.arrange(arrangeGrob(g1 + theme(legend.position="none"),g2 + theme(legend.position="non
e"),g3 + theme(legend.position="none"),top ="High School Student Cluster Analysis" ,ncol=1))
```