HW_Clustering

경제학과 2020110210 공소연

2022-11-12

Q1

```
setwd("C:/Users/sy/Documents/bda")
data <- read.table("wine.csv", header=T, sep=",")

## 1
str(data)
```

```
## 'data.frame':    178 obs. of  14 variables:
##  $ Class          : chr  "C" "C" "C" "C" ...
##  $ Alcohol        : num  14.2 13.2 13.2 14.4 13.2 ...
##  $ Malic          : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
##  $ Ash            : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
##  $ Alcalinity     : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
##  $ Magnesium      : int  127 100 101 113 118 112 96 121 97 98 ...
##  $ Phenols        : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
##  $ Flavanoids     : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
##  $ Nonflavanoid   : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
##  $ Proanthocyanins: num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
##  $ Intensity      : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
##  $ Hue            : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
##  $ OD280          : num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
##  $ Proline        : int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
```

```
head(data)
```

```
##   Class Alcohol Malic  Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoid
## 1     C   14.23  1.71 2.43       15.6       127    2.80       3.06          0.28
## 2     C   13.20  1.78 2.14       11.2       100    2.65       2.76          0.26
## 3     C   13.16  2.36 2.67       18.6       101    2.80       3.24          0.30
## 4     C   14.37  1.95 2.50       16.8       113    3.85       3.49          0.24
## 5     C   13.24  2.59 2.87       21.0       118    2.80       2.69          0.39
## 6     C   14.20  1.76 2.45       15.2       112    3.27       3.39          0.34
##   Proanthocyanins Intensity  Hue OD280 Proline
## 1            2.29      5.64 1.04  3.92    1065
## 2            1.28      4.38 1.05  3.40    1050
## 3            2.81      5.68 1.03  3.17    1185
## 4            2.18      7.80 0.86  3.45    1480
## 5            1.82      4.32 1.04  2.93     735
## 6            1.97      6.75 1.05  2.85    1450
```

```
y <- data[,1]
x <- data[,-1]
```

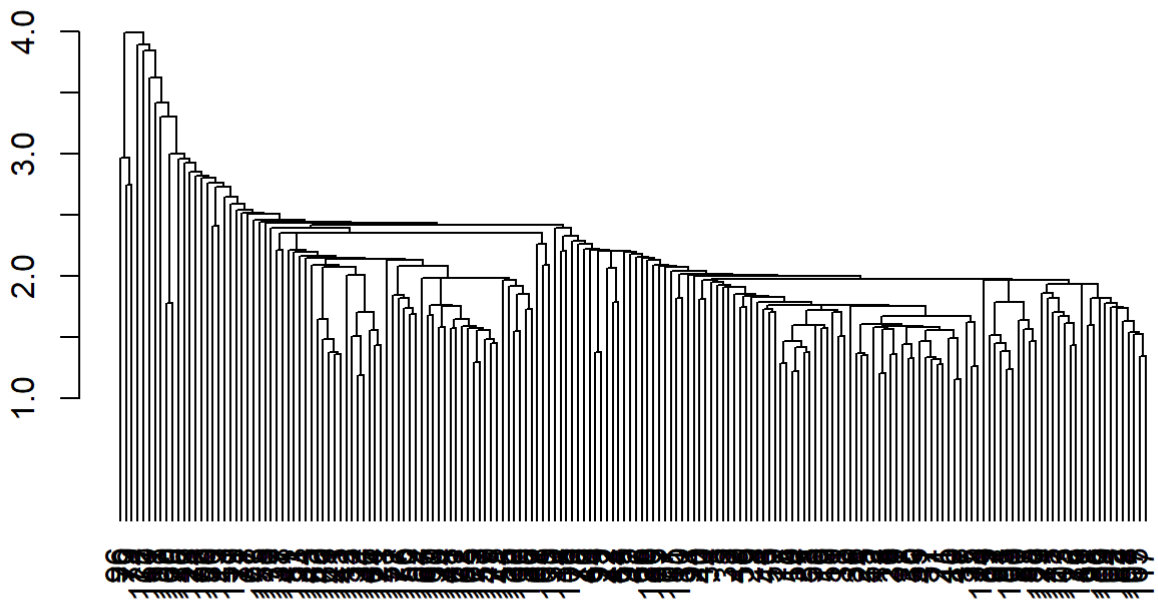
Q2

```
x.norm <- data.frame(sapply(x, scale))
head(x.norm)
```

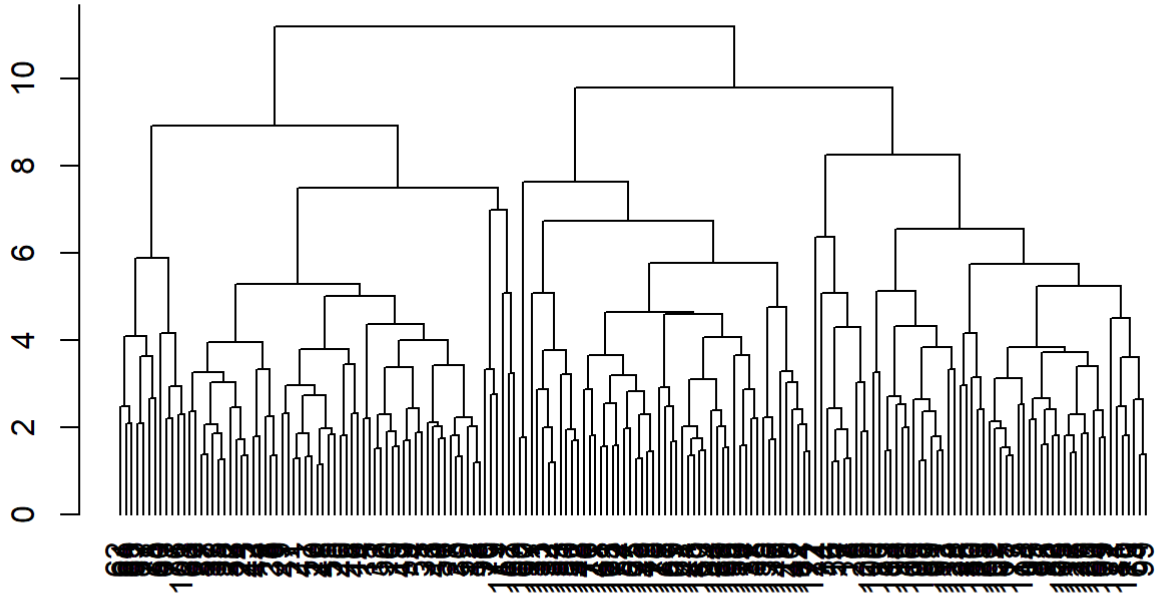
```
##     Alcohol       Malic        Ash Alcalinity  Magnesium   Phenols Flavanoids
## 1 1.5143408 -0.56066822  0.2313998 -1.1663032 1.90852151 0.8067217  1.0319081
## 2 0.2455968 -0.49800856 -0.8256672 -2.4838405 0.01809398 0.5670481  0.7315653
## 3 0.1963252  0.02117152  1.1062139 -0.2679823 0.08810981 0.8067217  1.2121137
## 4 1.6867914 -0.34583508  0.4865539 -0.8069748 0.92829983 2.4844372  1.4623994
## 5 0.2948684  0.22705328  1.8352256  0.4506745 1.27837900 0.8067217  0.6614853
## 6 1.4773871 -0.51591132  0.3043010 -1.2860793 0.85828399 1.5576991  1.3622851
##   Nonflavanoid Proanthocyanins  Intensity        Hue     OD280     Proline
## 1   -0.6577078       1.2214385  0.2510088  0.3610679 1.8427215  1.01015939
## 2   -0.8184106      -0.5431887 -0.2924962  0.4048188 1.1103172  0.96252635
## 3   -0.4970050       2.1299594  0.2682629  0.3173170 0.7863692  1.39122370
## 4   -0.9791134       1.0292513  1.1827317 -0.4264485 1.1807407  2.32800680
## 5    0.2261576       0.4002753 -0.3183774  0.3610679 0.4483365 -0.03776747
## 6   -0.1755994       0.6623487  0.7298108  0.4048188 0.3356589  2.23274072
```

Q3

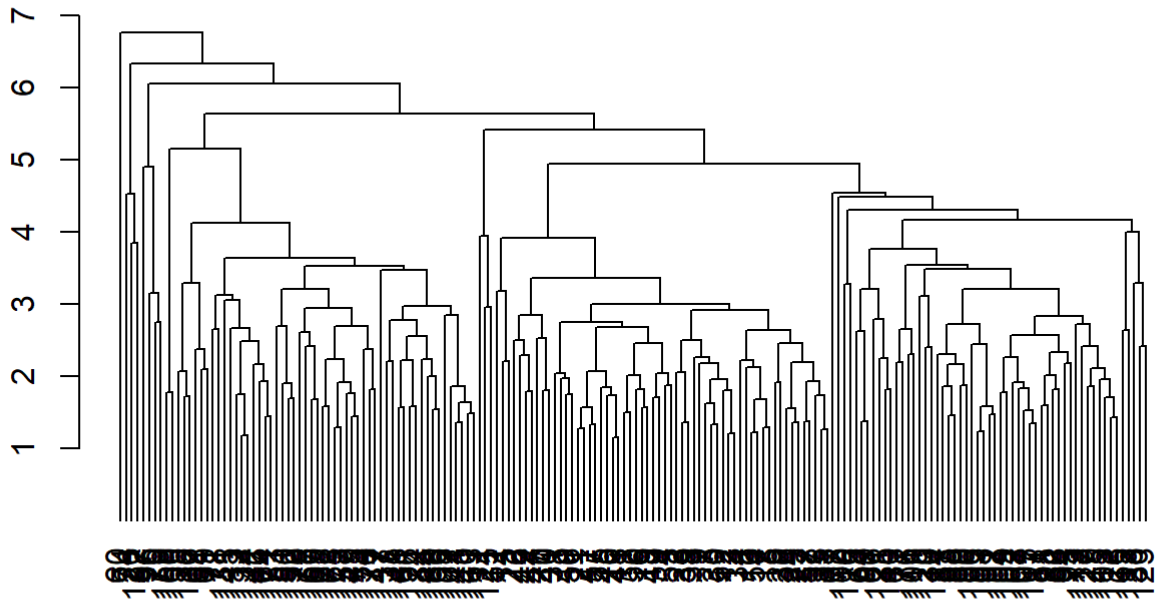
```
d.norm <- dist(x.norm, method="euclidean")
hc1_single <- hclust(d.norm, method="single")
plot(hc1_single, hang=-1, ann=FALSE)
```



```
hc2_complete <- hclust(d.norm, method="complete")
plot(hc2_complete, hang=-1, ann=FALSE)
```




```
hc3_average <- hclust(d.norm, method="average")
plot(hc3_average, hang=-1, ann=FALSE)
```

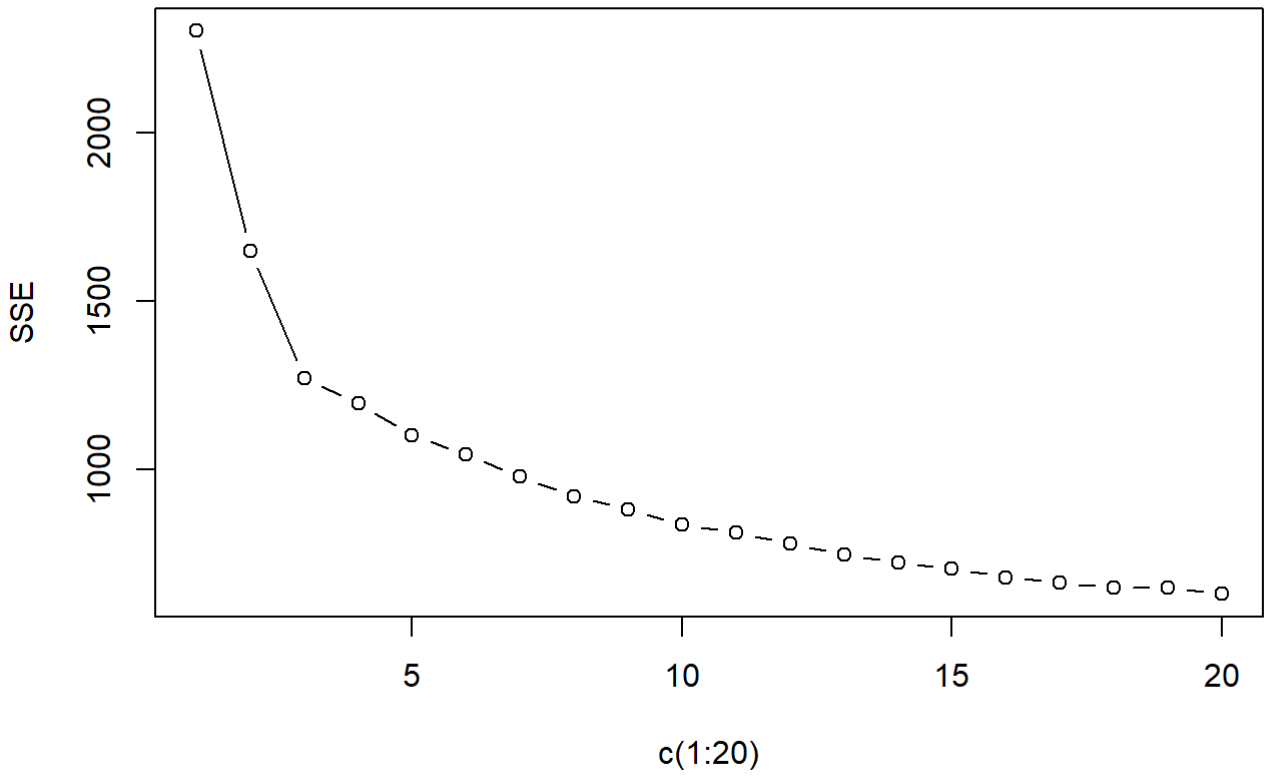


Q4

```
# 1) SSE
SSE <- c()

for (i in 1:20){
  set.seed(1)
  kmeans_cluster <- kmeans(x.norm, i)
  SSE[i] <- kmeans_cluster$tot.withinss
}

plot(c(1:20), SSE, type="b")
```

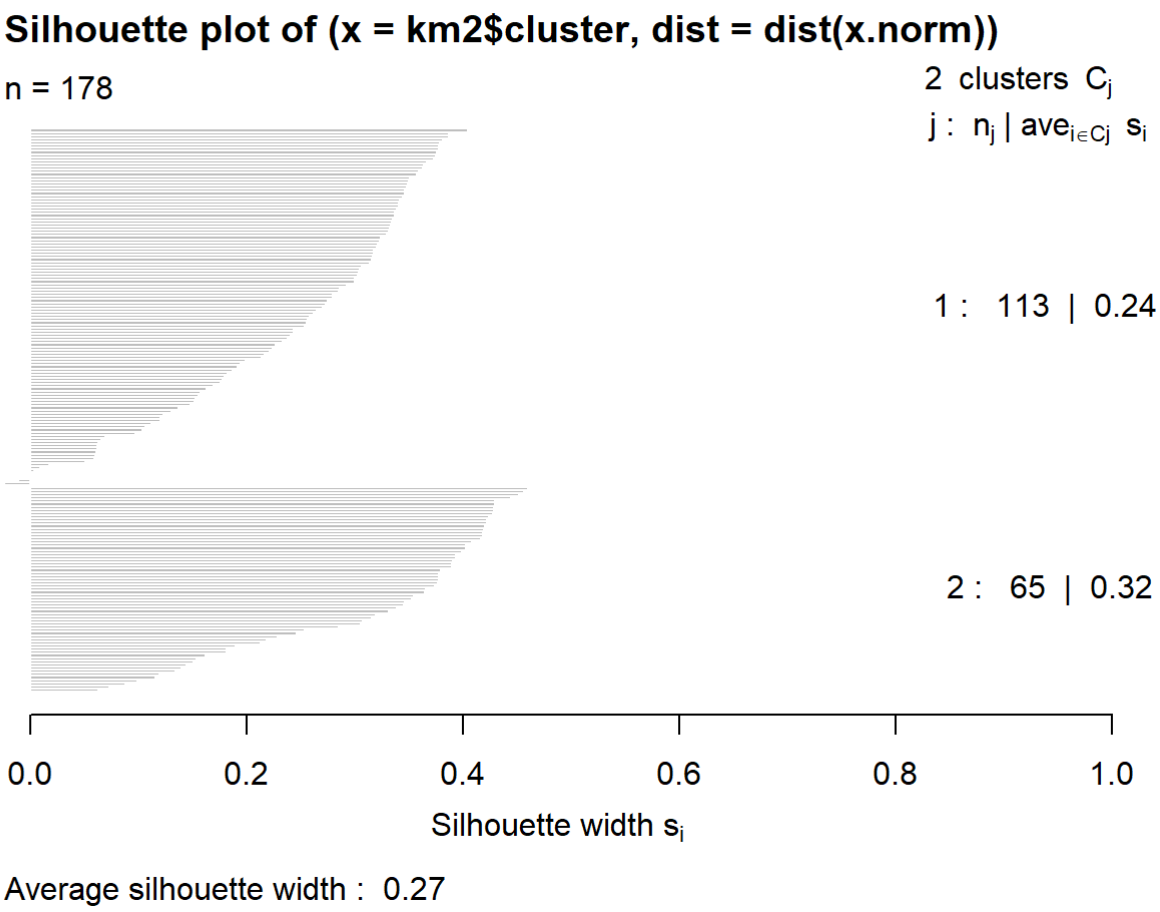


k=3

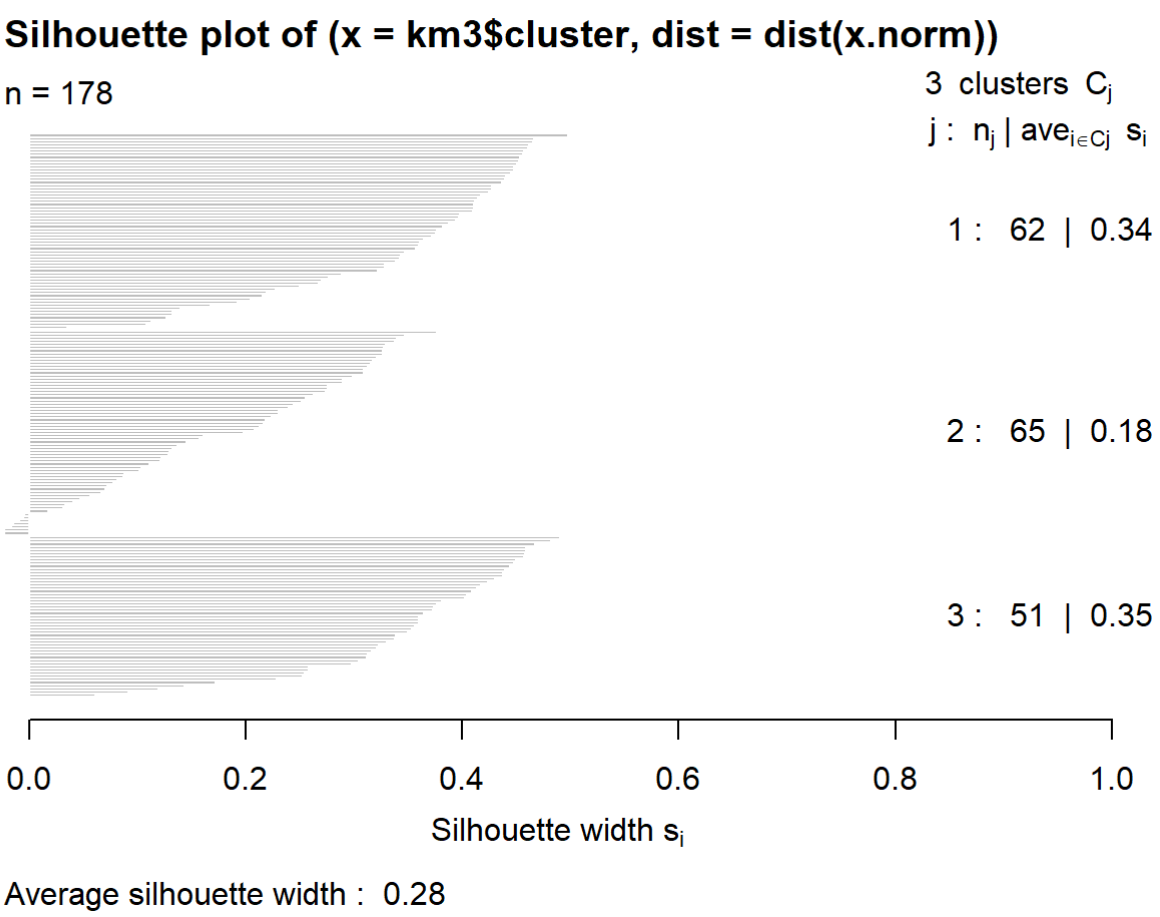
```
# 2) Silhouette
library("cluster")
```

```
## Warning: 패키지 'cluster'는 R 버전 4.2.2에서 작성되었습니다
```

```
km2 <- kmeans(x.norm, 2)
sil2 <- silhouette(km2$cluster, dist(x.norm))
plot(sil2)
```

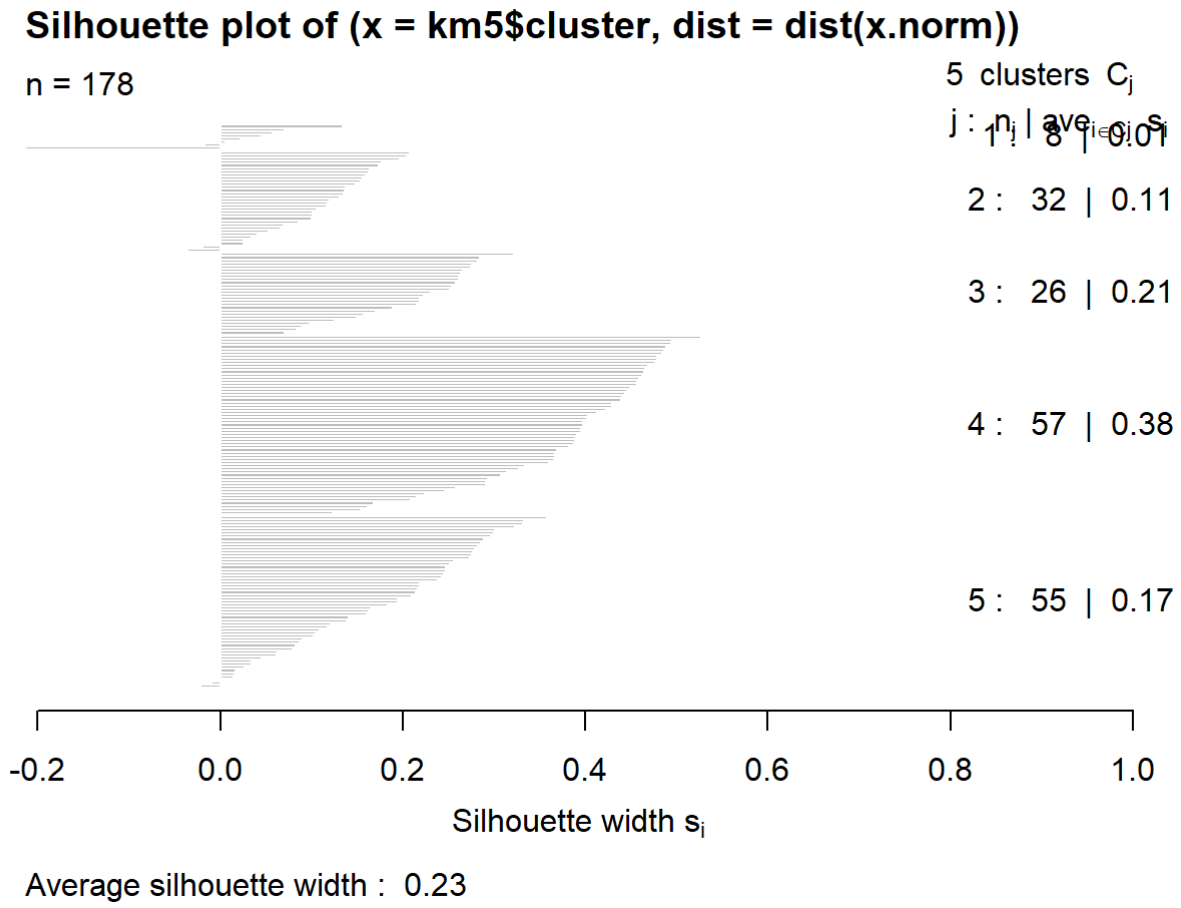



```
km3 <- kmeans(x.norm, 3)
sil3 <- silhouette(km3$cluster, dist(x.norm))
plot(sil3)
```




```
km4 <- kmeans(x.norm, 4)
sil4 <- silhouette(km4$cluster, dist(x.norm))
plot(sil4)
```


```
km5 <- kmeans(x.norm, 5)
sil5 <- silhouette(km5$cluster, dist(x.norm))
plot(sil5)
```



k=3일 때 Average silhouette width가 가장 크므로 optimal k는 3

Q5

```
set.seed(1122334455)
km <- kmeans(x.norm, 3)
comp <- data.frame(km$cluster, y)
comp
```

```
##     km.cluster y
## 1            3 C
## 2            3 C
## 3            3 C
## 4            3 C
## 5            3 C
## 6            3 C
## 7            3 C
## 8            3 C
## 9            3 C
## 10           3 C
## 11           3 C
## 12           3 C
## 13           3 C
## 14           3 C
## 15           3 C
## 16           3 C
## 17           3 C
## 18           3 C
## 19           3 C
## 20           3 C
## 21           3 C
## 22           3 C
## 23           3 C
## 24           3 C
## 25           3 C
## 26           3 C
## 27           3 C
## 28           3 C
## 29           3 C
## 30           3 C
## 31           3 C
## 32           3 C
## 33           3 C
## 34           3 C
## 35           3 C
## 36           3 C
## 37           3 C
## 38           3 C
## 39           3 C
## 40           3 C
## 41           3 C
## 42           3 C
## 43           3 C
## 44           3 C
## 45           3 C
## 46           3 C
## 47           3 C
## 48           3 C
## 49           3 C
## 50           3 C
## 51           3 C
## 52           3 C
## 53           3 C
## 54           3 C
```

```
## 55          3 C
## 56          3 C
## 57          3 C
## 58          3 C
## 59          3 C
## 60          2 A
## 61          2 A
## 62          1 A
## 63          2 A
## 64          2 A
## 65          2 A
## 66          2 A
## 67          2 A
## 68          2 A
## 69          2 A
## 70          2 A
## 71          2 A
## 72          2 A
## 73          2 A
## 74          3 A
## 75          2 A
## 76          2 A
## 77          2 A
## 78          2 A
## 79          2 A
## 80          2 A
## 81          2 A
## 82          2 A
## 83          2 A
## 84          1 A
## 85          2 A
## 86          2 A
## 87          2 A
## 88          2 A
## 89          2 A
## 90          2 A
## 91          2 A
## 92          2 A
## 93          2 A
## 94          2 A
## 95          2 A
## 96          3 A
## 97          2 A
## 98          2 A
## 99          2 A
## 100         2 A
## 101         2 A
## 102         2 A
## 103         2 A
## 104         2 A
## 105         2 A
## 106         2 A
## 107         2 A
## 108         2 A
## 109         2 A
## 110         2 A
```

```
## 111          2 A
## 112          2 A
## 113          2 A
## 114          2 A
## 115          2 A
## 116          2 A
## 117          2 A
## 118          2 A
## 119          1 A
## 120          2 A
## 121          2 A
## 122          3 A
## 123          2 A
## 124          2 A
## 125          2 A
## 126          2 A
## 127          2 A
## 128          2 A
## 129          2 A
## 130          2 A
## 131          1 B
## 132          1 B
## 133          1 B
## 134          1 B
## 135          1 B
## 136          1 B
## 137          1 B
## 138          1 B
## 139          1 B
## 140          1 B
## 141          1 B
## 142          1 B
## 143          1 B
## 144          1 B
## 145          1 B
## 146          1 B
## 147          1 B
## 148          1 B
## 149          1 B
## 150          1 B
## 151          1 B
## 152          1 B
## 153          1 B
## 154          1 B
## 155          1 B
## 156          1 B
## 157          1 B
## 158          1 B
## 159          1 B
## 160          1 B
## 161          1 B
## 162          1 B
## 163          1 B
## 164          1 B
## 165          1 B
## 166          1 B
```

```
## 167          1 B
## 168          1 B
## 169          1 B
## 170          1 B
## 171          1 B
## 172          1 B
## 173          1 B
## 174          1 B
## 175          1 B
## 176          1 B
## 177          1 B
## 178          1 B
```

Q6

```
y.i <- NULL
y.i[y=="A"] <- 2
y.i[y=="B"] <- 1
y.i[y=="C"] <- 3

library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
accuracy(km$cluster, y.i)
```

```
##          ME     RMSE        MAE MPE     MAPE
## Test set  0 0.183597 0.03370787   0 1.685393
```