Homework 2. Clustering Practice (80 Points)

Zeming Zhang

2023-03-8

```
#install the package
#install.packages("devtools")
#install.packages("factoextra")
#install_github("vqv/ggbiplot")
#install.packages("fpc")
#install.packages('psych')
options(warn = -1)
```

Part 1. USArrests Dataset and Hierarchical Clustering (20 Points)

Consider the "USArrests" data. It is a built-in dataset you may directly get in RStudio. Perform hierarchical clustering on the observations (states) and answer the following questions.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

The output shows that the dataset contains 50 observations (states) and 4 variables (columns) which are the rate of crimes per 100,000 residents:

```
str(USArrests)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

Q1.1. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. (5 points)

```
# Compute Euclidean distance matrix
dist_mat <- dist(USArrests, method = "euclidean")

head(as.matrix(dist_mat)[1:50, 1:50],25)
```

```
##                Alabama    Alaska   Arizona  Arkansas California  Colorado
## Alabama        0.00000  37.17701  63.00833  46.92814   55.52477  41.93256
## Alaska        37.17701   0.00000  46.59249  77.19741   45.10222  66.47594
```

```
## Arizona        63.00833  46.59249   0.00000 108.85192   23.19418  90.35115
## Arkansas       46.92814  77.19741 108.85192   0.00000   97.58202  36.73486
## California     55.52477  45.10222  23.19418  97.58202    0.00000  73.19713
## Colorado       41.93256  66.47594  90.35115  36.73486   73.19713   0.00000
## Connecticut   128.20694 159.40656 185.15953  85.02829  169.27711  98.08119
## Delaware       16.80625  45.18296  58.61638  53.01038   49.29148  41.47783
## Florida       102.00162  79.97450  41.65453 148.73574   60.98073 131.40582
## Georgia        25.84183  57.03026  86.03796  25.58613   73.99730  25.09303
## Hawaii        191.80305 221.19354 248.26897 147.77598  231.07109 159.17918
## Idaho         116.76198 146.48498 176.81767  70.58704  162.61279  90.88641
## Illinois       28.45488  42.91165  45.69781  67.77027   32.71880  47.66907
## Indiana       123.34521 152.80409 181.89780  78.47809  166.22996  93.61506
## Iowa          180.61010 209.98352 239.99146 134.59495  224.63466 152.07975
## Kansas        121.51987 151.48020 180.02891  76.75344  164.51675  92.17972
## Kentucky      127.28417 156.61204 187.69030  81.09285  173.20791 101.02475
## Louisiana      15.45445  32.34888  48.49464  61.54551   41.63556  49.97499
## Maine         154.14529 183.89753 214.32741 107.85073  199.93111 127.90016
## Maryland       64.99362  44.83949  15.01599 111.64291   36.34735  97.30041
## Massachusetts  91.64851 123.25421 145.87591  54.18118  129.52471  59.90000
## Michigan       28.48543  28.85775  39.87242  71.10028   27.74635  51.45483
## Minnesota     164.65096 194.25357 223.08826 119.32464  207.22254 134.76454
## Mississippi    27.39014  28.63512  52.70873  69.68536   55.68357  68.66440
## Missouri       59.78829  89.30672 116.46738  24.89438  100.98891  29.17979
##               Connecticut  Delaware   Florida   Georgia    Hawaii     Idaho
## Alabama         128.20694  16.80625 102.00162  25.84183 191.80305 116.76198
## Alaska          159.40656  45.18296  79.97450  57.03026 221.19354 146.48498
## Arizona         185.15953  58.61638  41.65453  86.03796 248.26897 176.81767
## Arkansas         85.02829  53.01038 148.73574  25.58613 147.77598  70.58704
## California      169.27711  49.29148  60.98073  73.99730 231.07109 162.61279
## Colorado         98.08119  41.47783 131.40582  25.09303 159.17918  90.88641
## Connecticut       0.00000 128.21018 226.30300 104.42653  64.95237  25.28043
## Delaware        128.21018   0.00000  99.10832  33.24530 192.36611 119.42131
## Florida         226.30300  99.10832   0.00000 125.76649 289.42857 217.66518
## Georgia         104.42653  33.24530 125.76649   0.00000 167.12800  93.11606
## Hawaii           64.95237 192.36611 289.42857 167.12800   0.00000  79.75143
## Idaho            25.28043 119.42131 217.66518  93.11606  79.75143   0.00000
## Illinois        139.90647  18.15186  86.55871  45.00267 203.09961 132.81145
## Indiana          16.31625 125.31053 222.92387  98.77287  69.40641  15.40779
## Iowa             57.59557 182.70999 281.01352 156.44581  29.40782  64.13712
## Kansas           14.17392 123.16594 221.08272  97.17407  71.10084  13.96424
## Kentucky         26.34388 130.59743 228.33276 103.04145  70.45970  13.40970
## Louisiana       140.39808  16.97675  87.67035  38.69057 203.97061 130.43328
## Maine            37.64744 156.66665 255.15231 130.47256  50.56679  37.67240
## Maryland        191.16195  63.57798  37.78386  89.50536 254.68757 181.18954
## Massachusetts    40.16528  89.95832 187.04374  68.76227 103.09714  42.53998
## Michigan        147.26656  26.53168  80.35627  47.39810 209.83386 138.39097
## Minnesota        39.74670 166.14166 264.22583 140.32783  31.62040  49.48232
## Mississippi     153.26396  36.47917  85.39046  51.35543 216.83231 140.04164
## Missouri         70.69583  61.37891 157.49175  35.57134 132.93115  62.10443
##                Illinois    Indiana      Iowa     Kansas  Kentucky Louisiana
## Alabama        28.45488 123.345207 180.61010 121.519875 127.28417  15.45445
## Alaska         42.91165 152.804090 209.98352 151.480197 156.61204  32.34888
## Arizona        45.69781 181.897801 239.99146 180.028914 187.69030  48.49464
## Arkansas       67.77027  78.478086 134.59495  76.753436  81.09285  61.54551
```

```
## California     32.71880 166.229961 224.63466 164.516747 173.20791  41.63556
## Colorado       47.66907  93.615063 152.07975  92.179716 101.02475  49.97499
## Connecticut   139.90647  16.316250  57.59557  14.173920  26.34388 140.39808
## Delaware       18.15186 125.310534 182.70999 123.165945 130.59743  16.97675
## Florida        86.55871 222.923866 281.01352 221.082722 228.33276  87.67035
## Georgia        45.00267  98.772871 156.44581  97.174071 103.04145  38.69057
## Hawaii        203.09961  69.406412  29.40782  71.100844  70.45970 203.97061
## Idaho         132.81145  15.407790  64.13712  13.964240  13.40970 130.43328
## Illinois        0.00000 137.256111 195.32929 135.278823 143.59937  17.81123
## Indiana       137.25611   0.000000  58.58404   3.929377  14.60616 136.25594
## Iowa          195.32929  58.584042   0.00000  60.177487  53.99306 193.96662
## Kansas        135.27882   3.929377  60.17749   0.000000  15.76642 134.39494
## Kentucky      143.59937  14.606163  53.99306  15.766420   0.00000 140.93722
## Louisiana      17.81123 136.255936 193.96662 134.394940 140.93722   0.00000
## Maine         170.03332  36.003472  27.87938  36.989863  28.40792 167.82506
## Maryland       53.59338 187.179192 244.93072 185.337881 191.93960  51.47980
## Massachusetts 100.49522  41.544314  97.27713  39.018585  52.12571 102.55150
## Michigan       15.59166 143.065789 201.38135 141.398798 148.85967  16.65233
## Minnesota     178.21364  41.706834  18.71390  43.237715  40.19900 177.60512
## Mississippi    41.24439 147.822258 203.97267 146.023354 150.35159  24.70830
## Missouri       72.31597  65.613108 124.03568  64.015936  72.29869  71.65166
##                   Maine  Maryland Massachusetts  Michigan Minnesota Mississippi
## Alabama       154.14529  64.99362      91.64851  28.48543 164.65096    27.39014
## Alaska        183.89753  44.83949     123.25421  28.85775 194.25357    28.63512
## Arizona       214.32741  15.01599     145.87591  39.87242 223.08826    52.70873
## Arkansas      107.85073 111.64291      54.18118  71.10028 119.32464    69.68536
## California    199.93111  36.34735     129.52471  27.74635 207.22254    55.68357
## Colorado      127.90016  97.30041      59.90000  51.45483 134.76454    68.66440
## Connecticut    37.64744 191.16195      40.16528 147.26656  39.74670   153.26396
## Delaware      156.66665  63.57798      89.95832  26.53168 166.14166    36.47917
## Florida       255.15231  37.78386     187.04374  80.35627 264.22583    85.39046
## Georgia       130.47256  89.50536      68.76227  47.39810 140.32783    51.35543
## Hawaii         50.56679 254.68757     103.09714 209.83386  31.62040   216.83231
## Idaho          37.67240 181.18954      42.53998 138.39097  49.48232   140.04164
## Illinois      170.03332  53.59338     100.49522  15.59166 178.21364    41.24439
## Indiana        36.00347 187.17919      41.54431 143.06579  41.70683   147.82226
## Iowa           27.87938 244.93072      97.27713 201.38135  18.71390   203.97267
## Kansas         36.98986 185.33788      39.01859 141.39880  43.23772   146.02335
## Kentucky       28.40792 191.93960      52.12571 148.85967  40.19900   150.35159
## Louisiana     167.82506  51.47980     102.55150  16.65233 177.60512    24.70830
## Maine           0.00000 218.69989      74.76323 175.94968  19.91909   176.93923
## Maryland      218.69989   0.00000     152.65929  46.12949 228.52871    48.45132
## Massachusetts  74.76323 152.65929       0.00000 108.48839  79.34009   117.97682
## Michigan      175.94968  46.12949     108.48839   0.00000 184.52480    35.44009
## Minnesota      19.91909 228.52871      79.34009 184.52480   0.00000   188.77871
## Mississippi   176.93923  48.45132     117.97682  35.44009 188.77871     0.00000
## Missouri       99.24601 122.05921      35.05382  77.47400 107.09146    86.08496
##                Missouri    Montana  Nebraska    Nevada New Hampshire New Jersey
## Alabama        59.78829 127.392621 134.43697  37.43047    179.736196   83.24302
## Alaska         89.30672 156.673578 164.11426  34.88682    209.254415  114.73557
## Arizona       116.46738 187.540849 193.42360  44.79743    239.255616  135.85040
## Arkansas       24.89438  81.163107  88.97893  74.28869    133.678308   49.84426
## California    100.98891 172.996069 178.10081  26.74696    224.055395  119.04117
## Colorado       29.17979 100.751675 105.66835  48.83421    151.589182   50.42083
```

```
## Connecticut    70.69583  24.746313  17.86505 146.55108     57.043843   51.19668
## Delaware       61.37891 130.393136 136.37833  35.05325    181.854695   80.87799
## Florida       157.49175 228.327856 234.46401  84.25586    280.247480  176.89717
## Georgia        35.57134 103.302081 110.19573  50.56758    155.665603   60.77829
## Hawaii        132.93115  69.885120  59.93071 207.73360     31.220666  113.18732
## Idaho          62.10443  11.764353  19.90427 138.76743     63.208702   52.82234
## Illinois       72.31597 143.447273 148.80679  22.36605    194.607657   90.39934
## Indiana        65.61311  13.512957  12.59603 142.22166     58.096988   51.93149
## Iowa          124.03568  53.529898  46.60955 200.73886      2.291288  108.24181
## Kansas         64.01594  14.406943  13.78913 140.77088     59.594127   49.67494
## Kentucky       72.29869   3.834058  13.34916 148.92394     53.141321   62.29398
## Louisiana      71.65166 141.035457 147.58286  28.47244    193.137723   93.29823
## Maine          99.24601  27.733914  23.71771 176.13134     26.530925   85.84340
## Maryland      122.05921 191.924595 198.50866  53.21701    244.109668  143.04269
## Massachusetts  35.05382  51.250073  52.32638 107.55431     96.729158   11.45644
## Michigan       77.47400 148.808266 154.78953  13.29737    200.707150   98.63458
## Minnesota     107.09146  39.384515  30.34996 183.52782     18.828701   90.19590
## Mississippi    86.08496 150.610425 158.46956  47.62793    202.982167  110.01627
## Missouri        0.00000  72.098821  77.45308  76.96805    123.427307   28.51175
##               New Mexico   New York North Carolina North Dakota      Ohio
## Alabama         51.64349  33.710829      101.96102    192.41614 117.38761
## Alaska          33.52193  43.182983       79.37607    221.37859 147.37334
## Arizona         13.89604  40.853519       57.61961    252.80819 174.33818
## Arkansas        97.93120  73.762118      147.18424    145.85554  74.36975
## California      24.49510  26.900929       80.33212    238.21446 157.99851
## Colorado        81.73622  52.278102      138.97759    165.75093  85.81754
## Connecticut    176.58032 145.268166      229.50401     73.03896  15.03629
## Delaware        50.08932  24.189461      102.86156    195.27227 118.17919
## Florida         51.14724  81.542198       38.52791    293.62275 215.46661
## Georgia         75.17772  50.643657      127.33597    168.61142  92.88364
## Hawaii         239.72655 208.186095      293.60024     41.33594  74.46771
## Idaho          166.96961 138.541907      217.44372     75.99901  22.69207
## Illinois        39.13579   6.236986       96.21419    208.58583 129.31114
## Indiana        172.48145 142.699755      225.01922     72.75747  12.21352
## Iowa           230.49356 200.856292      281.50432     17.54879  67.43901
## Kansas         170.71605 140.757309      223.10896     74.33391  10.92016
## Kentucky       177.63032 149.261515      228.13139     65.72830  26.11073
## Louisiana       37.76255  21.417283       90.70816    206.24056 129.56948
## Maine          204.55312 175.732439      254.43997     38.66445  46.44351
## Maryland        15.89025  49.798896       44.64056    257.06906 180.33569
## Massachusetts  137.91171 105.673696      192.40062    112.20945  31.23171
## Michigan        30.42187  15.066519       89.03263    214.24409 135.78192
## Minnesota      213.90776 183.630063      266.03295     35.69832  49.48141
## Mississippi     39.98862  43.531598       78.07439    214.76995 142.75129
## Missouri       107.09795  77.722712      161.45715    137.36466  58.63557
##                Oklahoma    Oregon Pennsylvania Rhode Island South Carolina
## Alabama         85.84870  78.38686   131.085087     70.33811       44.18292
## Alaska         116.42942 106.93012   161.600897    103.90380       27.55649
## Arizona        143.93141 135.67288   188.866222    122.41887       36.89092
## Arkansas        43.01267  36.89512    86.990862     42.18531       89.24887
## California     128.77935 120.03958   172.999364    107.21311       47.06134
## Colorado        57.09974  47.36412   101.039596     43.87949       82.64194
## Connecticut     43.03603  53.24284     8.027453     64.83710      172.20677
## Delaware        87.19593  80.30722   132.003674     66.20801       48.72515
```

```
## Florida       184.98392 176.81066   229.949581    163.31246       65.18712
## Georgia        61.75986  54.05090   106.822376     50.99265       69.19458
## Hawaii        106.07417 114.49004    61.237978    128.62822      235.80098
## Idaho          34.73672  43.85544    23.112334     63.56453      159.76586
## Illinois       99.29471  91.72971   143.769329     77.04862       46.29525
## Indiana        38.13607  46.84208    11.662761     66.18708      167.03021
## Iowa           96.13038 105.07278    52.485903    121.79672      223.79517
## Kansas         36.11592  45.45217    11.256109     63.42594      165.25983
## Kentucky       45.20288  54.00963    20.555291     74.52308      170.22494
## Louisiana      98.43922  90.89367   143.600487     80.02006       35.00129
## Maine          71.28878  80.63554    32.218783     97.87206      196.95781
## Maryland      149.28138 141.15314   194.556958    129.30066       28.97758
## Massachusetts  17.65021  24.35672    44.984108     26.34388      135.67402
## Michigan      105.40522  96.69788   150.488139     86.74059       37.63044
## Minnesota      79.28569  88.21678    34.712534    104.35061      208.24901
## Mississippi   111.07952 103.93883   155.864942     96.50249       21.16719
## Missouri       28.39014  19.69822    73.295157     27.06234      103.66605
##               South Dakota Tennessee     Texas      Utah   Vermont  Virginia
## Alabama         151.089113  48.34760  41.56609 118.50270 190.37069  80.29533
## Alaska          179.948131  77.88453  72.36221 148.27609 218.29047 110.64669
## Arizona         211.751576 108.25812  93.27599 174.25734 251.48926 139.42471
## Arkansas        104.455206  12.61428  32.74462  76.43900 143.52857  36.42156
## California      197.524378  94.72766  77.38023 157.49263 237.43546 124.82091
## Colorado        125.302115  28.00589  14.50103  85.62552 165.04769  53.41685
## Connecticut      40.039231  82.19276  92.65916  15.75595  76.61736  49.30720
## Delaware        154.422829  53.34323  39.66522 118.51456 194.25460  82.67872
## Florida         252.438844 148.59287 134.17992 215.53385 292.02008 180.28602
## Georgia         127.294776  23.42755  22.85126  94.29236 166.72492  56.02874
## Hawaii           55.686713 144.38594 155.29601  74.13973  51.91926 111.85030
## Idaho            35.219313  70.16160  86.41007  27.42353  75.34693  38.13214
## Illinois        167.874953  65.67534  48.17198 129.24028 207.92566  95.20242
## Indiana          34.753417  75.70872  89.55166  17.13505  73.72272  43.06716
## Iowa             32.385336 133.38801 147.87119  68.99681  26.24900 100.81691
## Kansas           36.247483  74.22297  87.71055  15.90126  75.53595  41.27396
## Kentucky         25.001200  80.09126  96.65216  31.47713  64.83255  48.48505
## Louisiana       165.023998  61.61923  50.18147 130.33162 204.57820  93.31592
## Maine             8.537564 107.59656 123.25035  49.38846  39.96961  75.36823
## Maryland        215.780560 112.30503  99.88619 180.71696 255.12226 144.25758
## Massachusetts    74.710173  48.85489  53.68920  30.18278 114.19654  23.85728
## Michigan        173.113200  69.15526  55.17717 135.97445 212.79619 100.70909
## Minnesota        25.349951 117.29983 130.57320  50.63842  41.78445  84.45283
## Mississippi     173.492882  73.28335  68.86305 144.28115 211.87973 105.07483
## Missouri         96.711943  15.50258  25.49471  59.37786 136.67202  24.27962
##               Washington West Virginia  Wisconsin   Wyoming
## Alabama         92.82047     156.79241 183.775733  75.50709
## Alaska         122.14700     185.64086 213.575397 106.74010
## Arizona        149.29786     218.00608 242.312381 135.38039
## Arkansas        51.20478     110.07111 138.344245  30.98726
## California     133.10657     204.25371 226.457502 121.72034
## Colorado        60.64206     132.36011 154.115217  52.03672
## Connecticut     38.33406      47.89572  58.056696  54.06015
## Delaware        93.60433     160.56242 185.194195  77.93491
## Florida        190.55563     258.46054 283.423799 176.11261
## Georgia         68.59096     133.22965 159.511880  52.11909
```

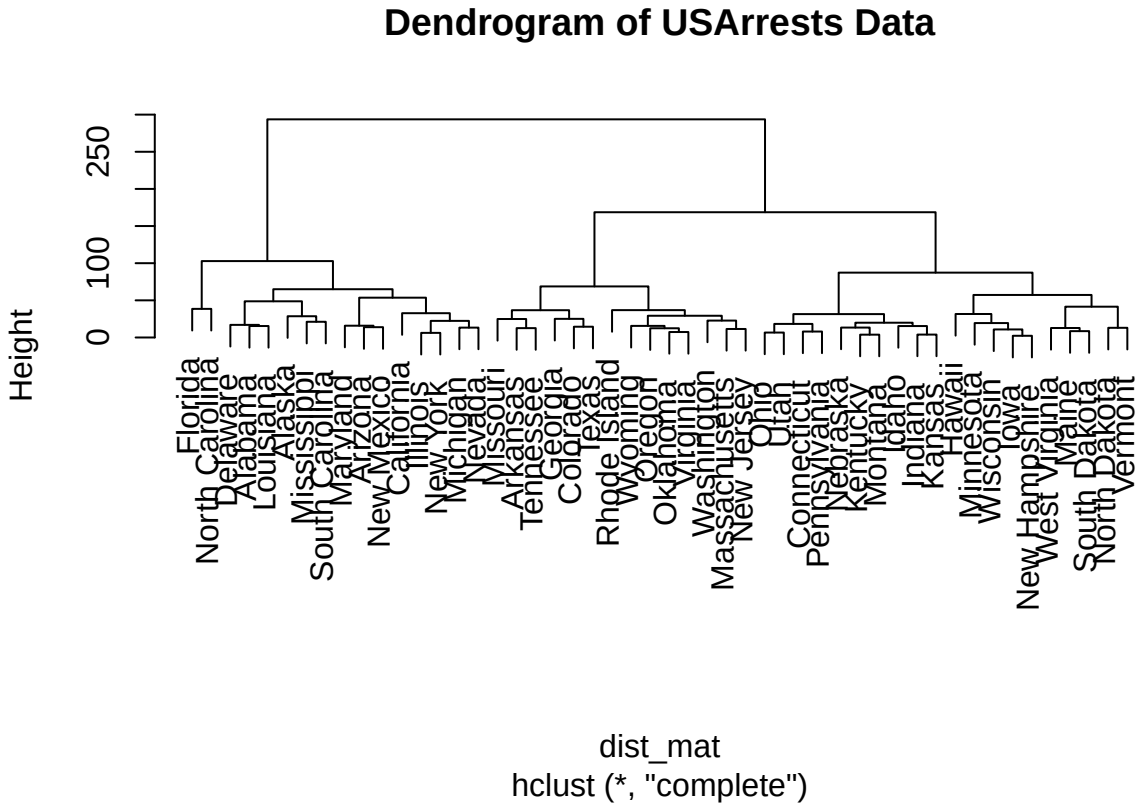
```
## Hawaii          99.69298      57.27102  20.824265 117.37721
## Idaho           33.64461      42.18554  68.151009  41.67253
## Illinois       104.69862     174.35074 197.332410  91.41400
## Indiana         33.54519      42.88520  61.042608  48.56254
## Iowa            91.66379      31.06847   9.508417 105.23141
## Kansas          31.94120      44.28070  62.509199  46.45858
## Kentucky        43.21458      31.90611  58.417977  52.69630
## Louisiana      104.93312     170.91957 196.748062  88.86799
## Maine           68.33864      12.77537  33.678628  79.04385
## Maryland       155.29601     221.62719 247.739157 139.78230
## Massachusetts   16.06767      82.40564  98.033107  27.84331
## Michigan       110.66083     179.46476 203.835080  97.16141
## Minnesota       74.21172      29.16093  19.437592  89.29894
## Mississippi    118.60110     178.54411 207.706379  99.74337
## Missouri        33.57082     103.62480 126.430692  23.50745
```

Q1.2. Cut the dendrogram at a height that results in three distinct clusters. Interpret the clusters. Which states belong to which clusters? (5 points)

```r
# Perform hierarchical clustering
hc <- hclust(dist_mat, method = "complete")

# Cut dendrogram into 3 clusters
clusters <- cutree(hc, k = 3)

# Count number of states in each cluster
clusters
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas          Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

```r
# Plot dendrogram
plot(hc, main = "Dendrogram of USArrests Data")
```



Q1.3 Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. Obtain three clusters. Which states belong to which clusters?(5 points)

```
# Scale the variables to have standard deviation one
scaled_data <- scale(USArrests)

head(scaled_data, 50)
```

```
##                   Murder     Assault    UrbanPop         Rape
## Alabama         1.24256408  0.78283935 -0.52090661 -0.003416473
## Alaska          0.50786248  1.10682252 -1.21176419  2.484202941
## Arizona         0.07163341  1.47880321  0.99898006  1.042878388
## Arkansas        0.23234938  0.23086801 -1.07359268 -0.184916602
## California      0.27826823  1.26281442  1.75892340  2.067820292
## Colorado        0.02571456  0.39885929  0.86080854  1.864967207
## Connecticut    -1.03041900 -0.72908214  0.79172279 -1.081740768
## Delaware       -0.43347395  0.80683810  0.44629400 -0.579946294
## Florida         1.74767144  1.97077766  0.99898006  1.138966691
## Georgia         2.20685994  0.48285493 -0.38273510  0.487701523
## Hawaii         -0.57123050 -1.49704226  1.20623733 -0.110181255
## Idaho          -1.19113497 -0.60908837 -0.79724965 -0.750769945
## Illinois        0.59970018  0.93883125  1.20623733  0.295524916
## Indiana        -0.13500142 -0.69308401 -0.03730631 -0.024769429
## Iowa           -1.28297267 -1.37704849 -0.58999237 -1.060387812
## Kansas         -0.41051452 -0.66908525  0.03177945 -0.345063775
## Kentucky        0.43898421 -0.74108152 -0.93542116 -0.526563903
## Louisiana       1.74767144  0.93883125  0.03177945  0.103348309
## Maine          -1.30593210 -1.05306531 -1.00450692 -1.434064548
```

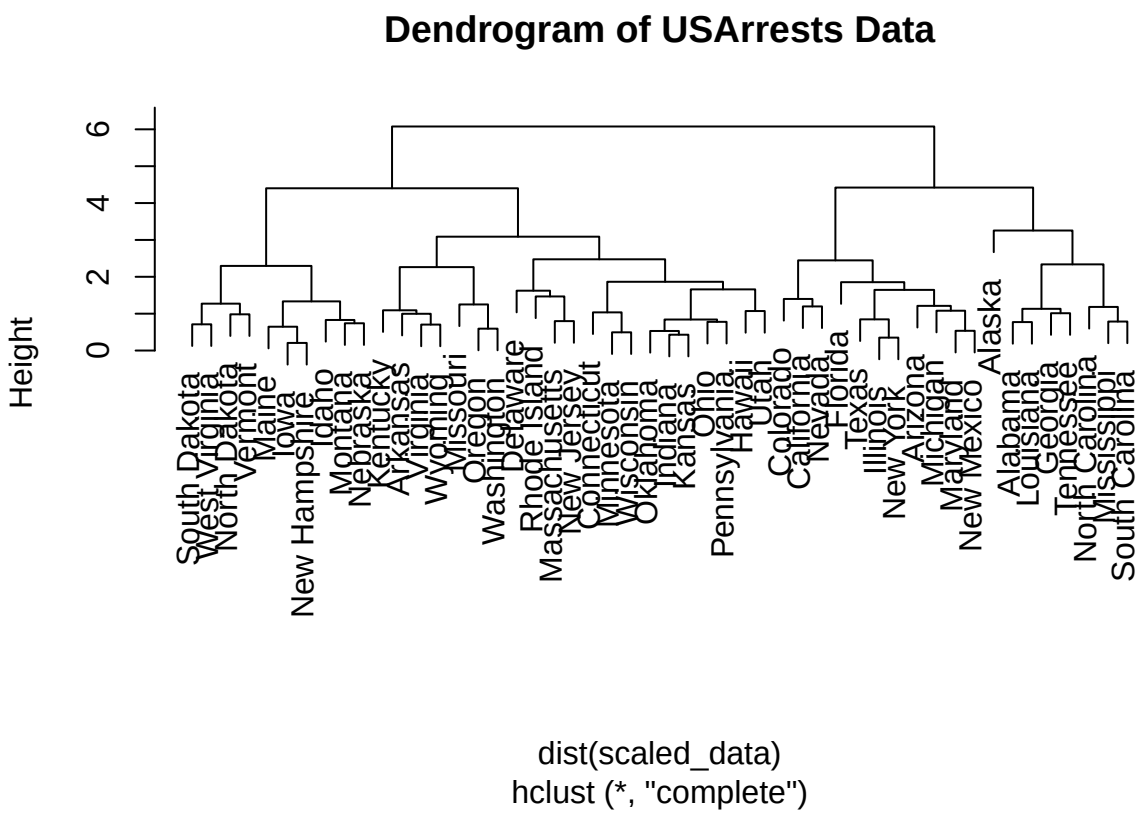
```
## Maryland        0.80633501  1.55079947  0.10086521  0.701231086
## Massachusetts  -0.77786532 -0.26110644  1.34440885 -0.526563903
## Michigan        0.99001041  1.01082751  0.58446551  1.480613993
## Minnesota      -1.16817555 -1.18505846  0.03177945 -0.676034598
## Mississippi     1.90838741  1.05882502 -1.48810723 -0.441152078
## Missouri        0.27826823  0.08687549  0.30812248  0.743936999
## Montana        -0.41051452 -0.74108152 -0.86633540 -0.515887425
## Nebraska       -0.80082475 -0.82507715 -0.24456358 -0.505210947
## Nevada          1.01296983  0.97482938  1.06806582  2.644350114
## New Hampshire  -1.30593210 -1.36504911 -0.65907813 -1.252564419
## New Jersey     -0.08908257 -0.14111267  1.62075188 -0.259651949
## New Mexico      0.82929443  1.37080881  0.30812248  1.160319648
## New York        0.76041616  0.99882813  1.41349461  0.519730957
## North Carolina  1.19664523  1.99477641 -1.41902147 -0.547916860
## North Dakota   -1.60440462 -1.50904164 -1.48810723 -1.487446939
## Ohio           -0.11204199 -0.60908837  0.65355127  0.017936483
## Oklahoma       -0.27275797 -0.23710769  0.16995096 -0.131534211
## Oregon         -0.66306820 -0.14111267  0.10086521  0.861378259
## Pennsylvania   -0.34163624 -0.77707965  0.44629400 -0.676034598
## Rhode Island   -1.00745957  0.03887798  1.48258036 -1.380682157
## South Carolina  1.51807718  1.29881255 -1.21176419  0.135377743
## South Dakota   -0.91562187 -1.01706718 -1.41902147 -0.900240639
## Tennessee       1.24256408  0.20686926 -0.45182086  0.605142783
## Texas           1.12776696  0.36286116  0.99898006  0.455672088
## Utah           -1.05337842 -0.60908837  0.99898006  0.178083656
## Vermont        -1.28297267 -1.47304350 -2.31713632 -1.071064290
## Virginia        0.16347111 -0.17711080 -0.17547783 -0.056798864
## Washington     -0.86970302 -0.30910395  0.51537975  0.530407436
## West Virginia  -0.47939280 -1.07706407 -1.83353601 -1.273917376
## Wisconsin      -1.19113497 -1.41304662  0.03177945 -1.113770203
## Wyoming        -0.22683912 -0.11711392 -0.38273510 -0.601299251
```

```r
# Perform hierarchical clustering with complete linkage and Euclidean distance
hc <- hclust(dist(scaled_data), method = "complete")

# Cut dendrogram into 3 clusters
clusters <- cutree(hc, k = 3)

# Count number of states in each cluster
clusters
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
```

```
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

```
# Plot dendrogram
plot(hc, main = "Dendrogram of USArrests Data")
```



Q1.4 What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer. *(5 points)*

Answer: Scaling the variables in the USArrests dataset before computing inter-observation dissimilarities can have a significant effect on the hierarchical clustering obtained. It helps to avoid biases caused by differences in variable scales and ensures that all variables are given equal weight. Standardization is a suitable method for scaling, as it preserves the original distribution of the variables and avoids distorting the relationships between variables.

Therefore, in my opinion, the variables should be scaled before computing inter-observation dissimilarities. This will help to avoid biases caused by differences in variable scales and ensure that all variables are given equal weight. Standardization is a suitable method for scaling in this case, as it preserves the original distribution of the variables and avoids distorting the relationships between variables.

Part 2. Market Segmentation (60 Points)

An advertisement division of large club store needs to perform customer analysis the store customers in order to create a segmentation for more targeted marketing campaign

You task is to identify similar customers and characterize them (at least some of them). In other word perform clustering and identify customers segmentation.

This data-set is derived from https://www.kaggle.com/imakash3011/customer-personality-analysis

```
Colomns description:
People
  ID: Customer's unique identifier
  Year_Birth: Customer's birth year
  Education: Customer's education level
  Marital_Status: Customer's marital status
  Income: Customer's yearly household income
  Kidhome: Number of children in customer's household
  Teenhome: Number of teenagers in customer's household
  Dt_Customer: Date of customer's enrollment with the company
  Recency: Number of days since customer's last purchase
  Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

  MntWines: Amount spent on wine in last 2 years
  MntFruits: Amount spent on fruits in last 2 years
  MntMeatProducts: Amount spent on meat in last 2 years
  MntFishProducts: Amount spent on fish in last 2 years
  MntSweetProducts: Amount spent on sweets in last 2 years
  MntGoldProds: Amount spent on gold in last 2 years

Place
  NumWebPurchases: Number of purchases made through the company's website
  NumStorePurchases: Number of purchases made directly in stores
```

Assume that data was current on 2014-07-01

Q2.1. Read Dataset and Data Conversion to Proper Data Format *(12 points)*

Read "m_marketing_campaign.csv" using `data.table::fread` command, examine the data.

```
# fread m_marketing_campaign.csv and save it as df (2 points)
df <- data.frame(fread("m_marketing_campaign.csv"))

head(df)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957  Bachelor         Single  58138       0        0  04-09-2012
## 2 2174       1954  Bachelor         Single  46344       1        1  08-03-2014
## 3 4141       1965  Bachelor       Together  71613       0        0  21-08-2013
## 4 6182       1984  Bachelor       Together  26646       1        0  10-02-2014
## 5 5324       1981       PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967    Master       Together  62513       0        1  09-09-2013
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain
```

```
## 1           88               8                4        0
## 2            6               1                2        0
## 3           42               8               10        0
## 4            5               2                4        0
## 5           15               5                6        0
## 6           14               6               10        0
```

```
# Convert Year_Birth to Age (assume that current date is 2014-07-01) (2 points)

df$Age <- 2014 - df$Year_Birth

# Dt_Customer is a date (it is still character), convert it to membership days (i.e. number of days per
# hint: note European date format, use as.Date with proper format argument (2 points)

Dt_Customer <- as.Date(df$Dt_Customer, format = "%d-%m-%Y")
df$MembershipDays <- as.Date("2014-07-01") - Dt_Customer

head(df)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957  Bachelor         Single  58138       0        0  04-09-2012
## 2 2174       1954  Bachelor         Single  46344       1        1  08-03-2014
## 3 4141       1965  Bachelor       Together  71613       0        0  21-08-2013
## 4 6182       1984  Bachelor       Together  26646       1        0  10-02-2014
## 5 5324       1981       PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967    Master       Together  62513       0        1  09-09-2013
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain Age MembershipDays
## 1           88               8                4        0  57       665 days
## 2            6               1                2        0  60       115 days
## 3           42               8               10        0  49       314 days
## 4            5               2                4        0  30       141 days
## 5           15               5                6        0  33       163 days
## 6           14               6               10        0  47       295 days
```

```
# Summarize Education column (use table function) (2 points)
table(df$Education)
```

```
##
##  Associate   Bachelor HighSchool     Master        PhD
##        200       1114         54        363        478
```

```
# Lets treat Education column as ordinal categories and use simple levels for
# distance calculations

# Assuming following order of degrees:
#    HighSchool, Associate, Bachelor, Master, PhD
# factorize Education column (hint: use factor function with above levels)

df$Education <- factor(df$Education, levels = c("High School", "Associate",
```

```
                                                "Bachelor", "Master",
                                                "PhD"))
head(df)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957  Bachelor         Single  58138       0        0  04-09-2012
## 2 2174       1954  Bachelor         Single  46344       1        1  08-03-2014
## 3 4141       1965  Bachelor       Together  71613       0        0  21-08-2013
## 4 6182       1984  Bachelor       Together  26646       1        0  10-02-2014
## 5 5324       1981       PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967    Master       Together  62513       0        1  09-09-2013
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain Age MembershipDays
## 1           88               8                 4        0  57       665 days
## 2            6               1                 2        0  60       115 days
## 3           42               8                10        0  49       314 days
## 4            5               2                 4        0  30       141 days
## 5           15               5                 6        0  33       163 days
## 6           14               6                10        0  47       295 days
```

```
# Summarize Marital_Status column (use table function)
table(df$Marital_Status)
```

```
##
## Divorced  Married   Single Together    Widow
##      232      857      471      573       76
```

```
# Lets convert single Marital_Status categories for 5 separate binary categories
# (2 points)
# Divorced, Married, Single, Together and Widow, the value will be 1 if customer
# is in that category and 0 if customer is not
# hint: use dummyVars from caret package, model.matrix or simple comparison
# (there are only 5 groups)
# Convert Marital_Status to separate binary categories
marital_dummies <- dummyVars(~ Marital_Status, data = df)
df_dummies <- predict(marital_dummies, newdata = df)
df <- cbind(df, df_dummies)
df <- df[complete.cases(df), ]

head(df)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957  Bachelor         Single  58138       0        0  04-09-2012
## 2 2174       1954  Bachelor         Single  46344       1        1  08-03-2014
## 3 4141       1965  Bachelor       Together  71613       0        0  21-08-2013
## 4 6182       1984  Bachelor       Together  26646       1        0  10-02-2014
## 5 5324       1981       PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967    Master       Together  62513       0        1  09-09-2013
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
```

```
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain Age MembershipDays
## 1           88               8                 4        0  57       665 days
## 2            6               1                 2        0  60       115 days
## 3           42               8                10        0  49       314 days
## 4            5               2                 4        0  30       141 days
## 5           15               5                 6        0  33       163 days
## 6           14               6                10        0  47       295 days
##   Marital_StatusDivorced Marital_StatusMarried Marital_StatusSingle
## 1                      0                     0                    1
## 2                      0                     0                    1
## 3                      0                     0                    0
## 4                      0                     0                    0
## 5                      0                     1                    0
## 6                      0                     0                    0
##   Marital_StatusTogether Marital_StatusWidow
## 1                      0                   0
## 2                      0                   0
## 3                      1                   0
## 4                      1                   0
## 5                      0                   0
## 6                      1                   0
```

```
# lets remove columns which we will no longer use:
# remove ID, Year_Birth, Dt_Customer, Marital_Status
# and save it as df_sel
df_sel <- df[, !(names(df) %in% c("ID", "Year_Birth", "Dt_Customer",
                                  "Marital_Status"))]

# Convert Education to integers
# hint: use as.integer function, if you use factor function earlier
# properly then HighSchool will be 1, Associate will be 2 and so on)
df_sel$Education <- as.integer(df_sel$Education)

head(df_sel)
```

```
##   Education Income Kidhome Teenhome Recency MntWines MntFruits MntMeatProducts
## 1         3  58138       0        0      58      635        88              546
## 2         3  46344       1        1      38       11         1                6
## 3         3  71613       0        0      26      426        49              127
## 4         3  26646       1        0      26       11         4               20
## 5         5  58293       1        0      94      173        43              118
## 6         4  62513       0        1      16      520        42               98
##   MntFishProducts MntSweetProducts MntGoldProds NumWebPurchases
## 1             172               88           88               8
## 2               2                1            6               1
## 3             111               21           42               8
## 4              10                3            5               2
## 5              46               27           15               5
## 6               0               42           14               6
##   NumStorePurchases Complain Age MembershipDays Marital_StatusDivorced
```

```
## 1                 4        0  57       665 days                       0
## 2                 2        0  60       115 days                       0
## 3                10        0  49       314 days                       0
## 4                 4        0  30       141 days                       0
## 5                 6        0  33       163 days                       0
## 6                10        0  47       295 days                       0
##   Marital_StatusMarried Marital_StatusSingle Marital_StatusTogether
## 1                     0                    1                      0
## 2                     0                    1                      0
## 3                     0                    0                      1
## 4                     0                    0                      1
## 5                     1                    0                      0
## 6                     0                    0                      1
##   Marital_StatusWidow
## 1                   0
## 2                   0
## 3                   0
## 4                   0
## 5                   0
## 6                   0
```

```
# lets scale (2 points)
# run scale function on df_sel and save it as df_scale
# that will be our scaled values which we will use for analysis
df_scale <- as.data.frame(scale(df_sel[, sapply(df_sel, is.numeric)]))
df_sel <- cbind(df_scale, df_sel[, !sapply(df_sel, is.numeric)])

head(df_scale)
```

```
##    Education     Income    Kidhome   Teenhome    Recency   MntWines  MntFruits
## 1 -0.5537597  0.2039592 -0.8133003 -0.9456314  0.3070815  0.9522363  1.5267941
## 2 -0.5537597 -0.2682874  1.0472753  0.8886041 -0.3823602 -0.8911466 -0.6414126
## 3 -0.5537597  0.7435152 -0.8133003 -0.9456314 -0.7960253  0.3348212  0.5548393
## 4 -0.5537597 -1.0570201  1.0472753 -0.9456314 -0.7960253 -0.8911466 -0.5666469
## 5  1.5791305  0.2101656  1.0472753 -0.9456314  1.5480766 -0.4125760  0.4053078
## 6  0.5126854  0.3791397 -0.8133003  0.8886041 -1.1407461  0.6125103  0.3803859
##   MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds NumWebPurchases
## 1       1.6602035       2.4381918       1.46115666   0.83825745       1.4096652
## 2      -0.7308919      -0.6566656      -0.63826363  -0.73725893      -1.1446650
## 3      -0.1951094       1.3276841      -0.15563828  -0.04556881       1.4096652
## 4      -0.6689006      -0.5110253      -0.59000110  -0.75647254      -0.7797607
## 5      -0.2349610       0.1443563      -0.01085067  -0.56433640       0.3149522
## 6      -0.3235201      -0.6930757       0.35111834  -0.58355001       0.6798565
##   NumStorePurchases    Complain        Age Marital_StatusDivorced
## 1       -0.57607035 -0.09917716  0.9726996             -0.3464198
## 2       -1.19003257 -0.09917716  1.2245117             -0.3464198
## 3        1.26581631 -0.09917716  0.3012007             -0.3464198
## 4       -0.57607035 -0.09917716 -1.2936091             -0.3464198
## 5        0.03789187 -0.09917716 -1.0417970             -0.3464198
## 6        1.26581631 -0.09917716  0.1333260             -0.3464198
##   Marital_StatusMarried Marital_StatusSingle Marital_StatusTogether
## 1            -0.7967173            1.9378934             -0.5916824
## 2            -0.7967173            1.9378934             -0.5916824
## 3            -0.7967173           -0.5157848              1.6893116
## 4            -0.7967173           -0.5157848              1.6893116
```

```
## 5             1.2545679          -0.5157848             -0.5916824
## 6            -0.7967173          -0.5157848              1.6893116
##   Marital_StatusWidow
## 1          -0.1898446
## 2          -0.1898446
## 3          -0.1898446
## 4          -0.1898446
## 5          -0.1898446
## 6          -0.1898446
```

PCA

Q2.2. Run PCA *(6 points)*

```
set.seed(10)

# Run PCA on df_scale, make biplot and scree plot/percentage variance explained plot
# save as pc_out, we will use pc_out$x[,1] and pc_out$x[,2] later for plotting
df_scale <- df_scale[complete.cases(df_scale), ]
pc_out <- prcomp(df_scale, scale. = TRUE)


# Create biplot using ggbiplot
biplot_gg <- ggbiplot(pc_out, obs.scale = 1, var.scale = 1,
                      groups = df_sel$Response,
                      ellipse = TRUE, circle = FALSE,
                      alpha = 0.5) +
  theme(legend.direction = "horizontal", legend.position = "top") +
  ggtitle("Biplot using ggbiplot")

# Convert biplot to Plotly object
ggplotly(biplot_gg)
```

```
# Create scree plot using ggplot2
screeplot_gg <- ggplot(data.frame(PC = 1:length(pc_out$sdev),
                                  Var = pc_out$sdev^2/sum(pc_out$sdev^2)),
                       aes(x = PC, y = Var)) +
  geom_bar(stat = "identity", fill = "#1f77b4") +
  geom_line(aes(x = PC, y = cumsum(Var)), color = "#2ca02c", size = 1.5) +
  scale_y_continuous(breaks = seq(0, 1, by = 0.1), labels = percent) +
  labs(x = "Principal Component", y = "Variance Explained",
       title = "Scree Plot / Percentage of Variance Explained") +
  theme(plot.title = element_text(hjust = 0.4))

# Convert scree plot to Plotly object
ggplotly(screeplot_gg)
```

Q2.3 Comment on observation (any visible distinct clusters?) *(2 points)*

Based on the given information, it appears that there may be 3 or 4 potential clusters. The data points on the right side of the graph and the directional attributes represented by the arrow lines are separated into 3 distinct groups, which suggests the possibility of 3 clusters. However, there is also a chance that there could be a fourth cluster. Further analysis and exploration of the data would be needed to determine the exact number and nature of the clusters present.

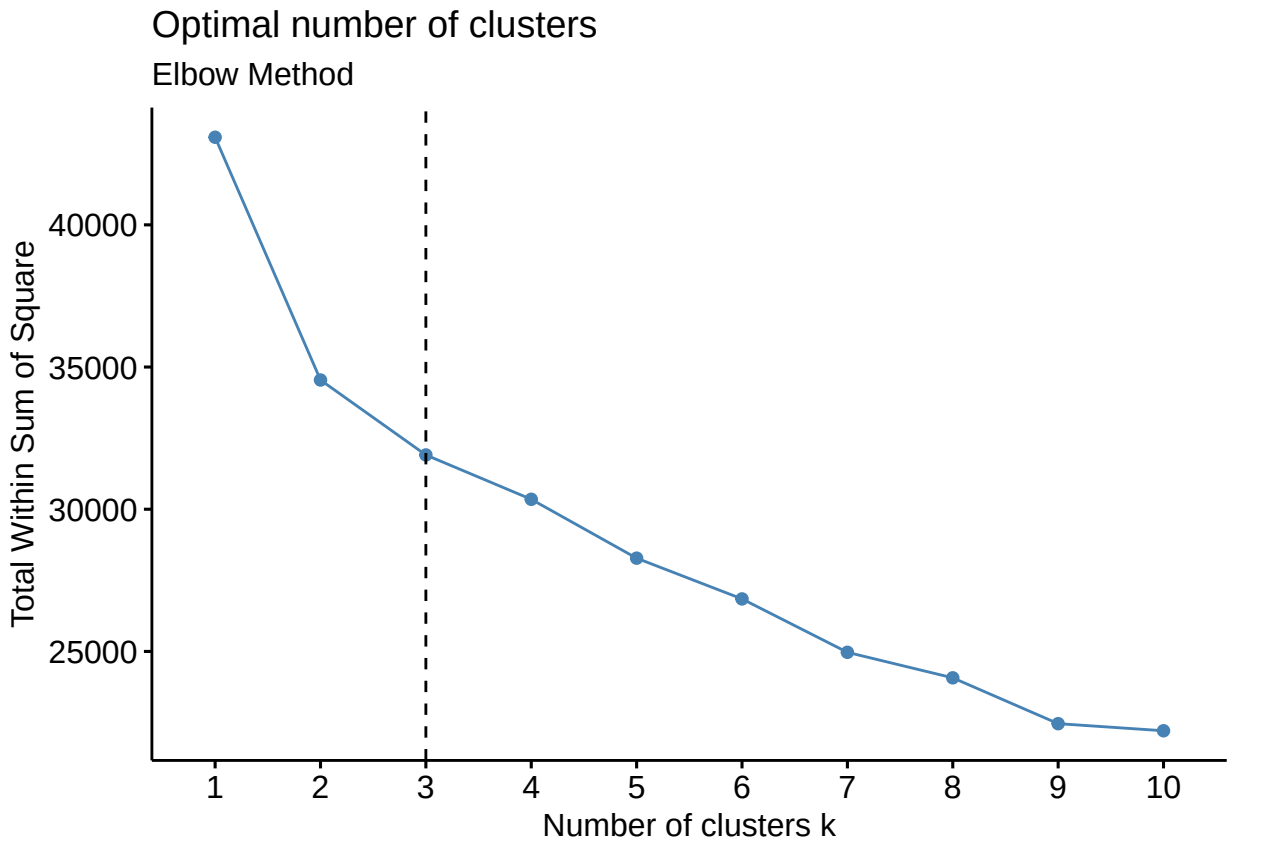
Cluster with K-Means

In questions Q2.4 to Q2.9 use K-Means method for clustering

Selecting Number of Clusters

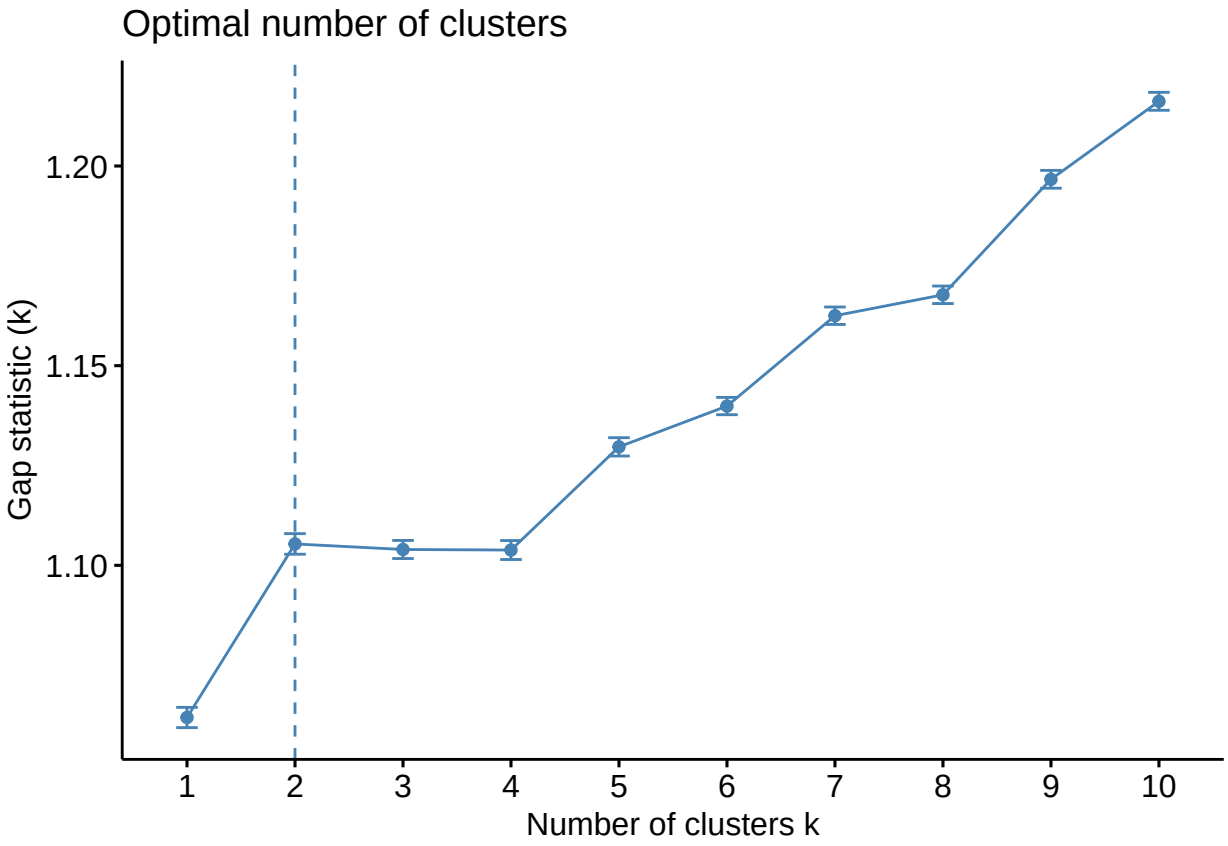
Q2.4 Select optimal number of clusters using elbow method. *(4 points)*

```
set.seed(11)

fviz_nbclust(df_scale, kmeans, method = "wss",
             k.max= 10, nstart= 20, iter.max= 20) +
  geom_vline(xintercept = 3,linetype = 2) + labs(subtitle = "Elbow Method")
```



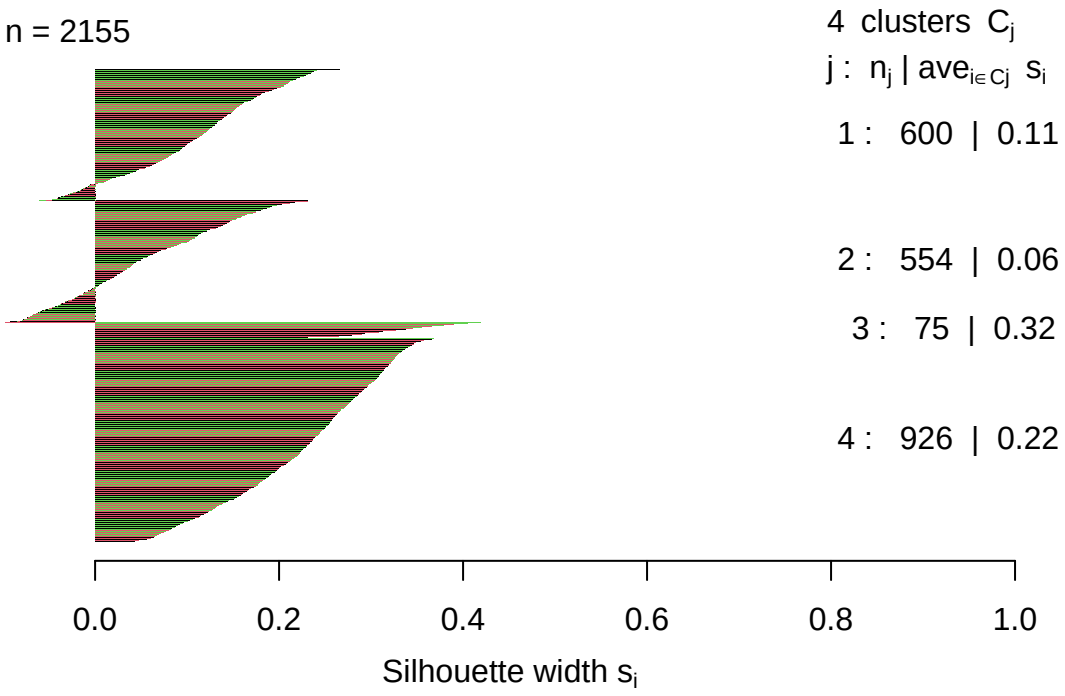
Q2.5 Select optimal number of clusters using Gap Statistic. *(4 points)*

```
set.seed(12)

gap_stat <- clusGap(df_scale, kmeans, K.max = 10, nstart = 25, B = 50)
fviz_gap_stat(gap_stat)
```



Q2.6 Select optimal number of clusters using Silhouette method. *(4 points)*

```
set.seed(13)

kms <- kmeans(df_scale, centers = 4)
plot(silhouette(kms$cluster, daisy(df_scale)), col=1:3, border=NA)
```



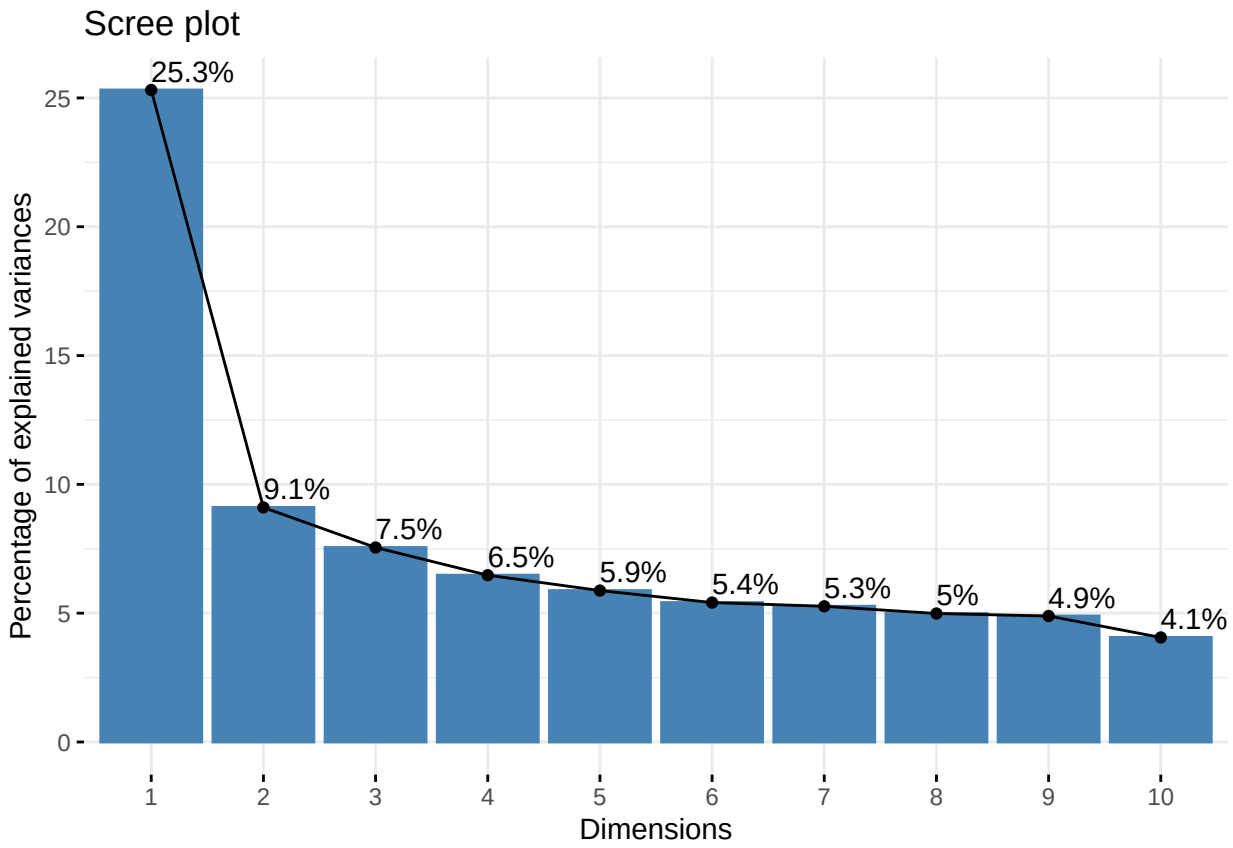
Q2.7 Which k will you choose based on elbow, gap statistics and silhuettes as well as clustering task (market segmentation for advertisement purposes, that is two groups don't provide sufficient benefit over a single groups)?*(4 points)*

Based on the analysis of the cluster gap method, there are two potential options for the number of clusters to choose: 5 and 7. The cluster gap method suggests that 7 clusters would be the next global maximum in terms of the gap statistic, indicating that the data may be best represented by 7 distinct clusters. However, there is also a maximum at 5 clusters on both the cluster gap graph and the elbow chart. Choosing 7 clusters would be a more reliable option as there is strong evidence from the cluster gap graph that 7 clusters are a good fit for the dataset. However, 5 clusters may also be a reasonable choice based on the elbow chart and the cluster gap graph, although it may not capture all the nuances of the data as well as 7 clusters. Ultimately, the decision of how many clusters to use will depend on the specific goals of the analysis and the trade-off between model complexity and fit to the data.

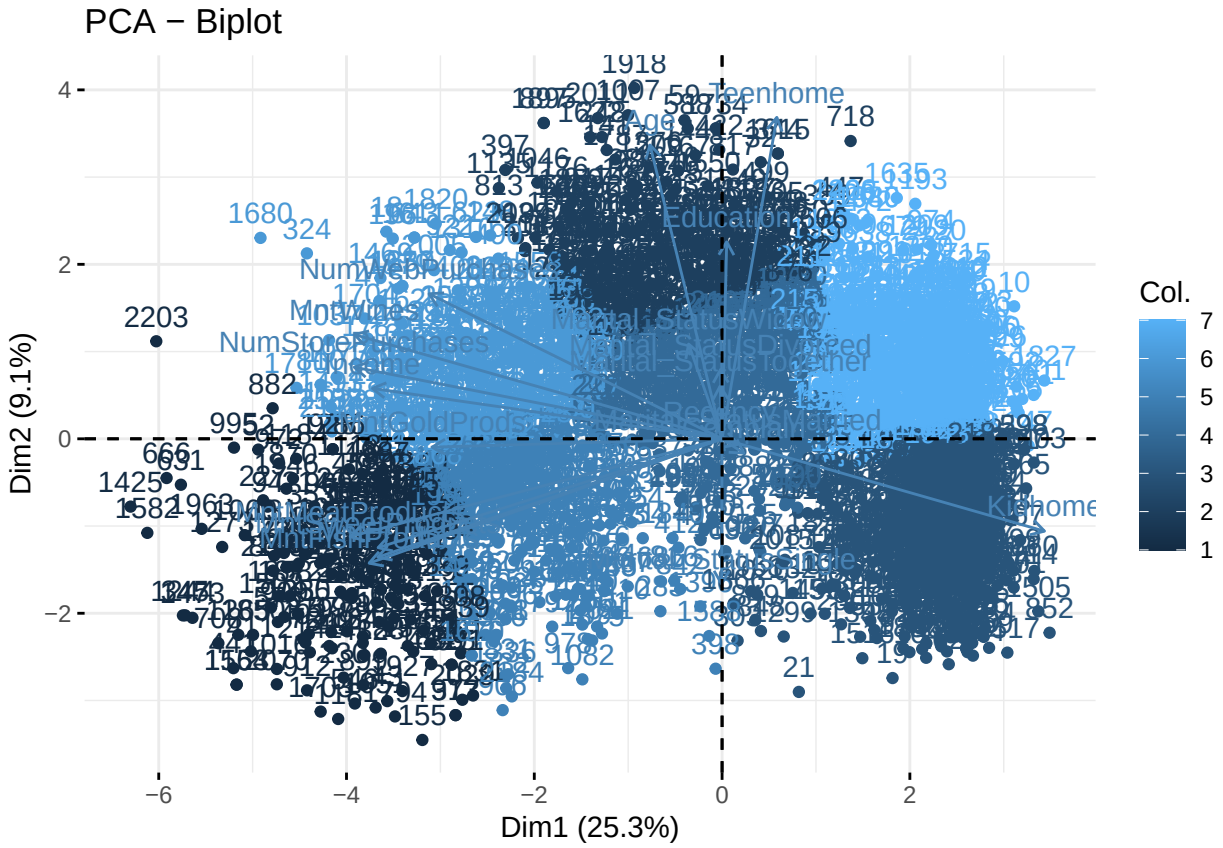
Clusters Visulalization

Q2.8 Make k-Means clusters with selected k_kmeans (store result as km_out). Plot your k_kmeans clusters on biplot (just PC1 vs PC2) by coloring points by their cluster id.*(4 points)*

```
set.seed(14)

# Select k_kmeans based on elbow method
fviz_eig(pc_out, addlabels = TRUE)
```




```r
k_kmeans <- 7

# Perform k-means clustering with proper seed
km_out <- kmeans(pc_out$x[, 1:2], centers = k_kmeans, nstart = 25)

# Add cluster column to original data
df$cluster <- km_out$cluster

# Plot k-means clusters on biplot
fviz_pca_biplot(pc_out, col.ind = km_out$cluster)
```



Q2.9 Do you see any grouping? Comment on you observation. *(2 points)*

Answer Although there are 7 centroids that are colored, it is evident from the grouping of data points that there are only 3 or 4 distinct clusters. Therefore, selecting 7 clusters would result in an excessive number of clusters.

Characterizing Cluster

Q2.10 Perform descriptive statistics analysis on obtained cluster. Based on that does one or more group have a distinct characteristics? *(8 points)* Hint: add cluster column to original df dataframe

```
# Set seed for reproducibility
set.seed(15)

# Add cluster column to original data
df_scale$cluster <- km_out$cluster

# Perform descriptive statistics analysis
head(describeBy(df_scale[, -ncol(df_scale)], group = df_scale$cluster))
```

```
## $`1`
##                vars   n  mean   sd median trimmed  mad   min   max
## Education         1 185 -0.37 0.76  -0.55   -0.42 0.00 -1.62  1.58
## Income            2 185  1.15 1.78   1.07    1.03 0.40 -0.25 24.57
## Kidhome           3 185 -0.78 0.24  -0.81   -0.81 0.00 -0.81  1.05
## Teenhome          4 185 -0.74 0.58  -0.95   -0.91 0.00 -0.95  0.89
## Recency           5 185 -0.02 0.98  -0.07   -0.03 1.18 -1.69  1.72
## MntWines          6 185  0.90 0.84   0.78    0.86 0.92 -0.90  3.19
```

```
## MntFruits                 7 185  1.86 1.32   1.93    1.86 1.63 -0.67  4.29
## MntMeatProducts           8 185  1.72 1.09   1.70    1.74 1.31 -0.68  3.60
## MntFishProducts           9 185  1.87 1.20   2.02    1.89 1.48 -0.69  4.02
## MntSweetProducts         10 185  1.79 1.33   1.82    1.79 1.72 -0.66  4.12
## MntGoldProds             11 185  1.03 1.25   0.76    0.92 1.31 -0.68  3.93
## NumWebPurchases          12 185  0.45 0.76   0.31    0.38 0.54 -0.78  2.50
## NumStorePurchases        13 185  0.84 0.85   0.96    0.83 1.37 -0.88  2.19
## Complain                 14 185 -0.04 0.75  -0.10   -0.10 0.00 -0.10 10.08
## Age                      15 185 -0.39 0.98  -0.45   -0.43 1.00 -2.22  2.06
## Marital_StatusDivorced   16 185 -0.12 0.83  -0.35   -0.35 0.00 -0.35  2.89
## Marital_StatusMarried    17 185 -0.06 0.99  -0.80   -0.14 0.00 -0.80  1.25
## Marital_StatusSingle     18 185  0.19 1.11  -0.52    0.06 0.00 -0.52  1.94
## Marital_StatusTogether   19 185  0.02 1.02  -0.59   -0.10 0.00 -0.59  1.69
## Marital_StatusWidow      20 185 -0.10 0.69  -0.19   -0.19 0.00 -0.19  5.27
##                         range  skew kurtosis   se
## Education                3.20  1.02     1.34 0.06
## Income                  24.82 12.34   159.74 0.13
## Kidhome                  1.86  7.60    56.04 0.02
## Teenhome                 1.83  2.42     3.86 0.04
## Recency                  3.41  0.11    -1.16 0.07
## MntWines                 4.09  0.30    -0.61 0.06
## MntFruits                4.96 -0.06    -1.15 0.10
## MntMeatProducts          4.28 -0.14    -1.03 0.08
## MntFishProducts          4.72 -0.15    -1.13 0.09
## MntSweetProducts         4.78 -0.08    -1.17 0.10
## MntGoldProds             4.61  0.61    -0.71 0.09
## NumWebPurchases          3.28  0.82     0.24 0.06
## NumStorePurchases        3.07  0.03    -1.22 0.06
## Complain                10.18 13.38   178.03 0.06
## Age                      4.28  0.32    -0.56 0.07
## Marital_StatusDivorced   3.23  3.34     9.17 0.06
## Marital_StatusMarried    2.05  0.59    -1.66 0.07
## Marital_StatusSingle     2.45  0.94    -1.13 0.08
## Marital_StatusTogether   2.28  1.03    -0.95 0.07
## Marital_StatusWidow      5.45  7.60    56.04 0.05
##
## $`2`
##                         vars   n  mean   sd median trimmed  mad   min   max
## Education                  1 264  0.69 0.95   0.51    0.75 1.58 -1.62  1.58
## Income                     2 264  0.26 0.43   0.27    0.26 0.39 -1.95  1.62
## Kidhome                    3 264 -0.66 0.52  -0.81   -0.81 0.00 -0.81  1.05
## Teenhome                   4 264  0.99 0.70   0.89    0.90 0.00 -0.95  2.72
## Recency                    5 264  0.07 0.99   0.13    0.09 1.23 -1.69  1.72
## MntWines                   6 264  0.74 0.84   0.61    0.67 0.74 -0.88  3.40
## MntFruits                  7 264 -0.33 0.36  -0.44   -0.40 0.33 -0.67  1.55
## MntMeatProducts            8 264 -0.23 0.37  -0.31   -0.27 0.30 -0.73  1.30
## MntFishProducts            9 264 -0.36 0.44  -0.49   -0.45 0.30 -0.69  2.22
## MntSweetProducts          10 264 -0.34 0.39  -0.49   -0.42 0.25 -0.66  1.08
## MntGoldProds              11 264  0.17 1.03  -0.18   -0.03 0.60 -0.85  5.32
## NumWebPurchases           12 264  0.93 1.01   0.86    0.89 0.81 -1.14  7.61
## NumStorePurchases         13 264  0.58 0.80   0.65    0.56 0.91 -1.80  2.19
## Complain                  14 264 -0.06 0.63  -0.10   -0.10 0.00 -0.10 10.08
## Age                       15 264  0.75 0.70   0.89    0.77 0.81 -0.71  2.15
## Marital_StatusDivorced    16 264  0.24 1.25  -0.35   -0.01 0.00 -0.35  2.89
```

```
## Marital_StatusMarried    17 264 -0.04 0.99  -0.80   -0.11 0.00 -0.80  1.25
## Marital_StatusSingle     18 264 -0.36 0.60  -0.52   -0.52 0.00 -0.52  1.94
## Marital_StatusTogether   19 264  0.06 1.03  -0.59   -0.06 0.00 -0.59  1.69
## Marital_StatusWidow      20 264  0.37 1.66  -0.19   -0.16 0.00 -0.19  5.27
##                        range  skew kurtosis   se
## Education               3.20 -0.40    -1.46 0.06
## Income                  3.56 -0.75     3.80 0.03
## Kidhome                 1.86  3.00     7.01 0.03
## Teenhome                3.67  0.58     3.54 0.04
## Recency                 3.41 -0.13    -1.12 0.06
## MntWines                4.27  0.73     0.21 0.05
## MntFruits               2.22  1.68     3.63 0.02
## MntMeatProducts         2.03  1.26     1.59 0.02
## MntFishProducts         2.91  2.92    12.12 0.03
## MntSweetProducts        1.74  1.57     2.09 0.02
## MntGoldProds            6.17  1.98     4.39 0.06
## NumWebPurchases         8.76  1.81    10.18 0.06
## NumStorePurchases       3.99  0.11    -0.64 0.05
## Complain               10.18 16.06   257.02 0.04
## Age                     2.85 -0.23    -0.95 0.04
## Marital_StatusDivorced  3.23  1.64     0.69 0.08
## Marital_StatusMarried   2.05  0.55    -1.71 0.06
## Marital_StatusSingle    2.45  3.53    10.50 0.04
## Marital_StatusTogether  2.28  0.95    -1.10 0.06
## Marital_StatusWidow     5.45  2.61     4.83 0.10
##
## $`3`
##                        vars   n  mean   sd median trimmed  mad   min   max
## Education                 1 513 -0.37 0.88  -0.55   -0.45 0.00 -1.62  1.58
## Income                    2 513 -0.92 0.41  -0.90   -0.92 0.43 -2.05  0.54
## Kidhome                   3 513  0.86 0.81   1.05    0.93 0.00 -0.81  2.91
## Teenhome                  4 513 -0.75 0.57  -0.95   -0.92 0.00 -0.95  0.89
## Recency                   5 513  0.01 1.00   0.00    0.01 1.33 -1.69  1.72
## MntWines                  6 513 -0.84 0.11  -0.89   -0.87 0.04 -0.92 -0.11
## MntFruits                 7 513 -0.53 0.19  -0.59   -0.57 0.07 -0.67  0.93
## MntMeatProducts           8 513 -0.64 0.35  -0.70   -0.68 0.05 -0.76  6.88
## MntFishProducts           9 513 -0.55 0.19  -0.62   -0.58 0.11 -0.69  0.84
## MntSweetProducts         10 513 -0.53 0.16  -0.59   -0.56 0.11 -0.66  0.69
## MntGoldProds             11 513 -0.55 0.41  -0.66   -0.61 0.20 -0.85  4.18
## NumWebPurchases          12 513 -0.76 0.48  -0.78   -0.83 0.54 -1.51  1.04
## NumStorePurchases        13 513 -0.89 0.28  -0.88   -0.90 0.46 -1.80  0.34
## Complain                 14 513  0.06 1.26  -0.10   -0.10 0.00 -0.10 10.08
## Age                      15 513 -0.80 0.61  -0.79   -0.82 0.62 -2.30  1.56
## Marital_StatusDivorced   16 513 -0.15 0.77  -0.35   -0.35 0.00 -0.35  2.89
## Marital_StatusMarried    17 513  0.07 1.02  -0.80    0.04 0.00 -0.80  1.25
## Marital_StatusSingle     18 513  0.24 1.13  -0.52    0.12 0.00 -0.52  1.94
## Marital_StatusTogether   19 513 -0.12 0.93  -0.59   -0.28 0.00 -0.59  1.69
## Marital_StatusWidow      20 513 -0.19 0.00  -0.19   -0.19 0.00 -0.19 -0.19
##                        range  skew kurtosis   se
## Education               3.20  0.73     0.20 0.04
## Income                  2.59  0.01     0.18 0.02
## Kidhome                 3.72 -0.52     1.70 0.04
## Teenhome                1.83  2.50     4.25 0.03
## Recency                 3.41  0.01    -1.25 0.04
```

```
## MntWines                0.81  2.97    10.57 0.00
## MntFruits               1.60  3.27    14.28 0.01
## MntMeatProducts         7.64 18.97   400.02 0.02
## MntFishProducts         1.53  3.07    13.65 0.01
## MntSweetProducts        1.35  2.57    10.05 0.01
## MntGoldProds            5.03  5.35    46.13 0.02
## NumWebPurchases         2.55  1.25     2.06 0.02
## NumStorePurchases       2.15  0.61     2.84 0.01
## Complain               10.18  7.80    58.90 0.06
## Age                     3.86  0.36     0.25 0.03
## Marital_StatusDivorced  3.23  3.68    11.56 0.03
## Marital_StatusMarried   2.05  0.30    -1.91 0.04
## Marital_StatusSingle    2.45  0.84    -1.30 0.05
## Marital_StatusTogether  2.28  1.43     0.05 0.04
## Marital_StatusWidow     0.00  -Inf      NaN 0.00
##
## $`4`
##                         vars   n  mean   sd median trimmed  mad   min   max
## Education                  1 264 -0.12 0.90  -0.55   -0.16 0.00 -1.62  1.58
## Income                     2 264  0.12 0.63   0.09    0.08 0.44 -1.15  4.19
## Kidhome                    3 264 -0.31 0.83  -0.81   -0.42 0.00 -0.81  1.05
## Teenhome                   4 264  0.28 0.89   0.89    0.34 0.00 -0.95  2.72
## Recency                    5 264 -0.09 0.96  -0.11   -0.10 1.20 -1.69  1.72
## MntWines                   6 264 -0.02 0.68  -0.20   -0.12 0.49 -0.92  2.94
## MntFruits                  7 264 -0.20 0.47  -0.37   -0.28 0.30 -0.67  1.90
## MntMeatProducts            8 264 -0.22 0.43  -0.35   -0.30 0.26 -0.75  2.29
## MntFishProducts            9 264 -0.17 0.50  -0.31   -0.24 0.40 -0.69  2.04
## MntSweetProducts          10 264 -0.20 0.50  -0.32   -0.28 0.36 -0.66  2.86
## MntGoldProds              11 264  0.14 0.87  -0.11    0.01 0.64 -0.85  3.51
## NumWebPurchases           12 264  0.30 0.73   0.31    0.29 0.54 -1.51  2.50
## NumStorePurchases         13 264  0.18 0.65   0.04    0.15 0.46 -1.80  1.88
## Complain                  14 264  0.02 1.08  -0.10   -0.10 0.00 -0.10 10.08
## Age                       15 264  0.07 0.83  -0.03    0.04 0.87 -1.88  2.15
## Marital_StatusDivorced    16 264 -0.15 0.77  -0.35   -0.35 0.00 -0.35  2.89
## Marital_StatusMarried     17 264  0.06 1.01  -0.80    0.02 0.00 -0.80  1.25
## Marital_StatusSingle      18 264  0.06 1.04  -0.52   -0.10 0.00 -0.52  1.94
## Marital_StatusTogether    19 264  0.04 1.02  -0.59   -0.09 0.00 -0.59  1.69
## Marital_StatusWidow       20 264 -0.13 0.58  -0.19   -0.19 0.00 -0.19  5.27
##                         range  skew kurtosis   se
## Education                3.20  0.61    -0.37 0.06
## Income                   5.34  2.88    16.77 0.04
## Kidhome                  1.86  1.04    -0.93 0.05
## Teenhome                 3.67 -0.52    -1.23 0.05
## Recency                  3.41  0.10    -1.14 0.06
## MntWines                 3.86  1.65     3.18 0.04
## MntFruits                2.57  1.50     2.08 0.03
## MntMeatProducts          3.05  2.43     8.26 0.03
## MntFishProducts          2.73  1.48     2.70 0.03
## MntSweetProducts         3.52  2.23     7.62 0.03
## MntGoldProds             4.36  1.37     1.54 0.05
## NumWebPurchases          4.01  0.27    -0.15 0.05
## NumStorePurchases        3.68  0.26     0.26 0.04
## Complain                10.18  9.17    82.36 0.07
## Age                      4.03  0.31    -0.67 0.05
```

```
## Marital_StatusDivorced  3.23  3.66    11.45 0.05
## Marital_StatusMarried   2.05  0.34    -1.89 0.06
## Marital_StatusSingle    2.45  1.24    -0.45 0.06
## Marital_StatusTogether  2.28  0.99    -1.02 0.06
## Marital_StatusWidow     5.45  9.17    82.36 0.04
##
## $`5`
##                        vars   n  mean   sd median trimmed  mad   min   max
## Education                 1 221 -0.30 0.94  -0.55   -0.37 0.00 -1.62  1.58
## Income                    2 221  0.75 0.63   0.76    0.76 0.43 -1.22  4.31
## Kidhome                   3 221 -0.50 0.74  -0.81   -0.68 0.00 -0.81  2.91
## Teenhome                  4 221 -0.77 0.54  -0.95   -0.95 0.00 -0.95  0.89
## Recency                   5 221  0.03 1.02   0.07    0.05 1.28 -1.69  1.72
## MntWines                  6 221  0.50 0.85   0.28    0.42 0.85 -0.92  3.01
## MntFruits                 7 221  0.63 1.03   0.38    0.50 0.78 -0.67  4.17
## MntMeatProducts           8 221  1.01 1.10   0.89    0.88 0.97 -0.52  6.88
## MntFishProducts           9 221  0.82 1.05   0.56    0.71 1.00 -0.69  4.00
## MntSweetProducts         10 221  0.70 1.02   0.47    0.58 0.89 -0.66  4.04
## MntGoldProds             11 221  0.40 0.93   0.17    0.27 0.74 -0.85  3.87
## NumWebPurchases          12 221  0.22 0.81  -0.05    0.15 0.54 -1.51  2.50
## NumStorePurchases        13 221  0.66 0.88   0.65    0.65 0.91 -1.80  2.19
## Complain                 14 221 -0.05 0.68  -0.10   -0.10 0.00 -0.10 10.08
## Age                      15 221 -0.41 1.01  -0.54   -0.44 1.12 -2.22  1.98
## Marital_StatusDivorced   16 221 -0.11 0.84  -0.35   -0.35 0.00 -0.35  2.89
## Marital_StatusMarried    17 221 -0.04 0.99  -0.80   -0.10 0.00 -0.80  1.25
## Marital_StatusSingle     18 221  0.37 1.18  -0.52    0.29 0.00 -0.52  1.94
## Marital_StatusTogether   19 221 -0.16 0.90  -0.59   -0.33 0.00 -0.59  1.69
## Marital_StatusWidow      20 221 -0.17 0.37  -0.19   -0.19 0.00 -0.19  5.27
##                        range  skew kurtosis   se
## Education               3.20  0.75    -0.10 0.06
## Income                  5.54  1.74    11.46 0.04
## Kidhome                 3.72  2.19     3.97 0.05
## Teenhome                1.83  2.74     5.55 0.04
## Recency                 3.41 -0.12    -1.29 0.07
## MntWines                3.93  0.76     0.00 0.06
## MntFruits               4.83  1.10     0.83 0.07
## MntMeatProducts         7.40  1.88     6.73 0.07
## MntFishProducts         4.70  0.85     0.17 0.07
## MntSweetProducts        4.71  0.99     0.59 0.07
## MntGoldProds            4.73  1.30     1.49 0.06
## NumWebPurchases         4.01  0.69     0.09 0.05
## NumStorePurchases       3.99  0.01    -0.68 0.06
## Complain               10.18 14.66   214.03 0.05
## Age                     4.20  0.27    -0.91 0.07
## Marital_StatusDivorced  3.23  3.28     8.78 0.06
## Marital_StatusMarried   2.05  0.53    -1.73 0.07
## Marital_StatusSingle    2.45  0.57    -1.68 0.08
## Marital_StatusTogether  2.28  1.57     0.46 0.06
## Marital_StatusWidow     5.45 14.66   214.03 0.02
##
## $`6`
##                        vars   n  mean   sd median trimmed  mad   min   max
## Education                 1 262  0.17 1.03  -0.55    0.16 1.58 -1.62  1.58
## Income                    2 262  0.79 0.42   0.76    0.79 0.42 -0.80  2.43
```

```
## Kidhome                    3 262 -0.74 0.36  -0.81   -0.81 0.00 -0.81  1.05
## Teenhome                   4 262  0.06 0.98   0.89    0.04 0.00 -0.95  2.72
## Recency                    5 262 -0.01 1.03   0.03   -0.01 1.38 -1.69  1.72
## MntWines                   6 262  1.11 1.00   0.88    1.07 1.03 -0.91  3.49
## MntFruits                  7 262  0.74 1.02   0.53    0.64 1.00 -0.67  4.17
## MntMeatProducts            8 262  0.74 0.81   0.50    0.65 0.75 -0.74  3.38
## MntFishProducts            9 262  0.66 0.94   0.49    0.59 0.97 -0.69  4.00
## MntSweetProducts          10 262  0.73 1.06   0.46    0.62 0.98 -0.66  5.66
## MntGoldProds              11 262  0.72 1.21   0.41    0.60 1.22 -0.85  3.91
## NumWebPurchases           12 262  0.89 0.97   0.68    0.84 1.08 -0.78  8.34
## NumStorePurchases         13 262  1.09 0.87   1.27    1.15 0.91 -1.80  2.19
## Complain                  14 262 -0.02 0.89  -0.10   -0.10 0.00 -0.10 10.08
## Age                       15 262  0.64 0.93   0.72    0.65 1.00 -1.46  5.84
## Marital_StatusDivorced    16 262  0.10 1.11  -0.35   -0.19 0.00 -0.35  2.89
## Marital_StatusMarried     17 262  0.01 1.00  -0.80   -0.04 0.00 -0.80  1.25
## Marital_StatusSingle      18 262 -0.24 0.77  -0.52   -0.48 0.00 -0.52  1.94
## Marital_StatusTogether    19 262  0.05 1.03  -0.59   -0.07 0.00 -0.59  1.69
## Marital_StatusWidow       20 262  0.23 1.45  -0.19   -0.19 0.00 -0.19  5.27
##                         range  skew kurtosis   se
## Education                3.20  0.27    -1.27 0.06
## Income                   3.23 -0.08     1.01 0.03
## Kidhome                  1.86  4.79    21.05 0.02
## Teenhome                 3.67  0.18    -1.20 0.06
## Recency                  3.41  0.01    -1.26 0.06
## MntWines                 4.39  0.40    -0.64 0.06
## MntFruits                4.83  0.94     0.59 0.06
## MntMeatProducts          4.13  0.90     0.17 0.05
## MntFishProducts          4.70  0.70     0.10 0.06
## MntSweetProducts         6.32  1.05     1.36 0.07
## MntGoldProds             4.76  0.71    -0.49 0.07
## NumWebPurchases          9.12  1.84    11.84 0.06
## NumStorePurchases        3.99 -0.51    -0.73 0.05
## Complain                10.18 11.25   125.02 0.05
## Age                      7.30  0.43     2.32 0.06
## Marital_StatusDivorced   3.23  2.09     2.40 0.07
## Marital_StatusMarried    2.05  0.44    -1.82 0.06
## Marital_StatusSingle     2.45  2.47     4.10 0.05
## Marital_StatusTogether   2.28  0.96    -1.08 0.06
## Marital_StatusWidow      5.45  3.17     8.10 0.09
```

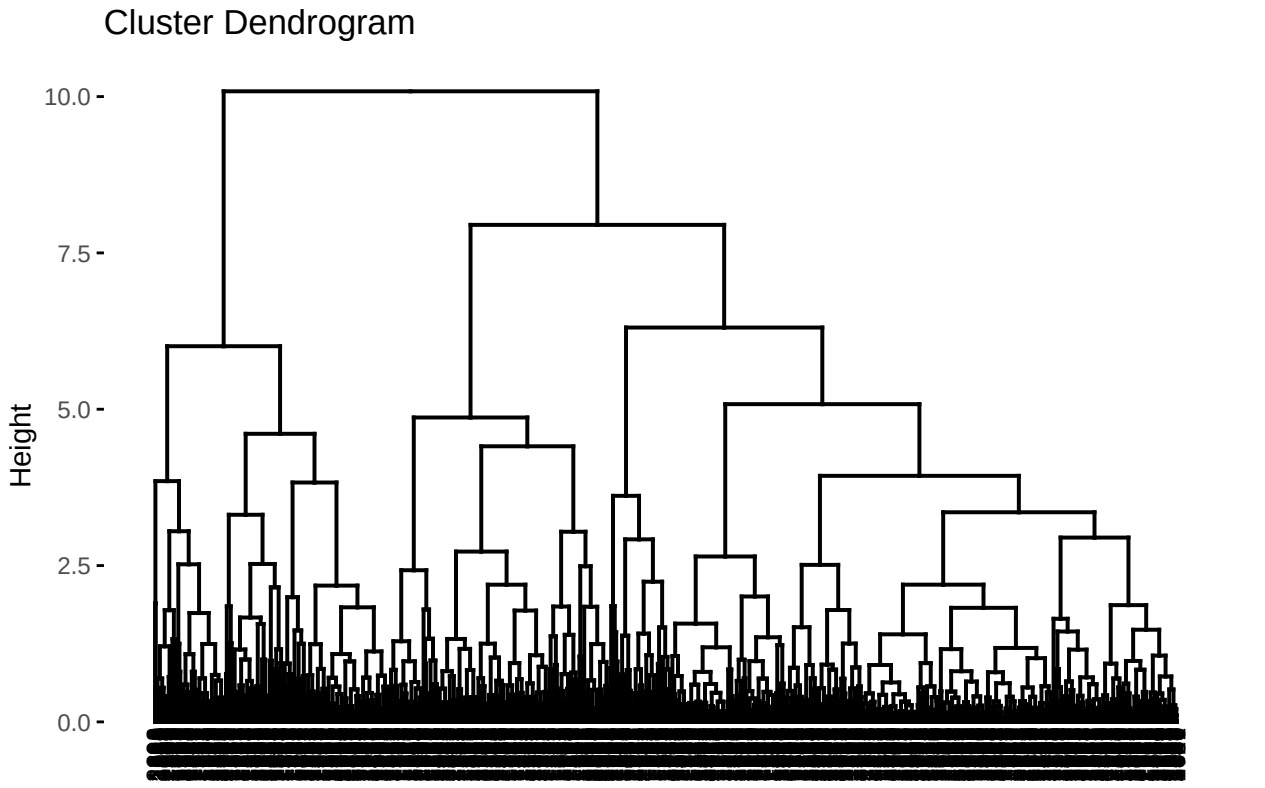
Cluster with Hierarchical Clustering

Q2.11 Perform clustering with Hierarchical method (Do you need to use scaling here?). Try complete, single and average linkage. Plot dendagram, based on it choose linkage and number of clusters, if possible, explain your choice. *(8 points)*

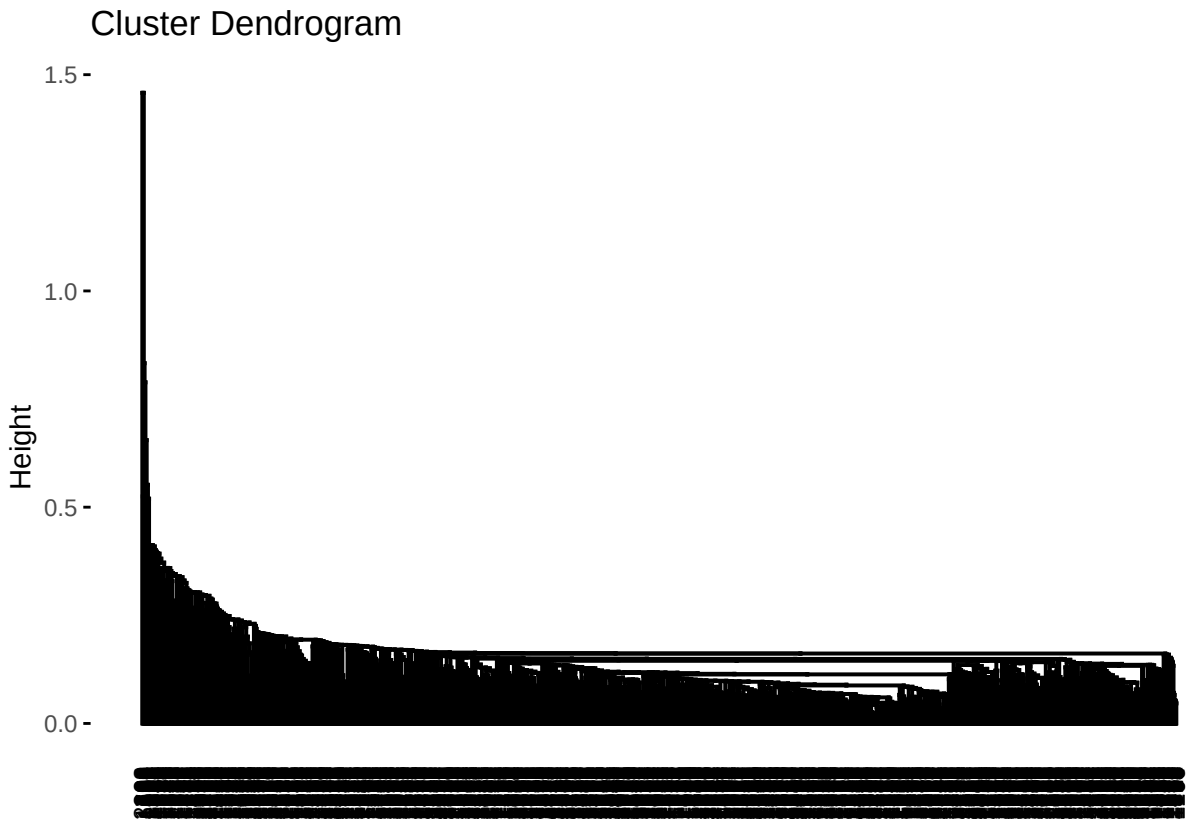
```r
set.seed(16)

# Perform hierarchical clustering with complete linkage
hclust_out_complete <- hclust(dist(pc_out$x[, 1:2]), method = "complete")

# Perform hierarchical clustering with single linkage
hclust_out_single <- hclust(dist(pc_out$x[, 1:2]), method = "single")
```

```r
# Perform hierarchical clustering with average linkage
hclust_out_average <- hclust(dist(pc_out$x[, 1:2]), method = "average")
```

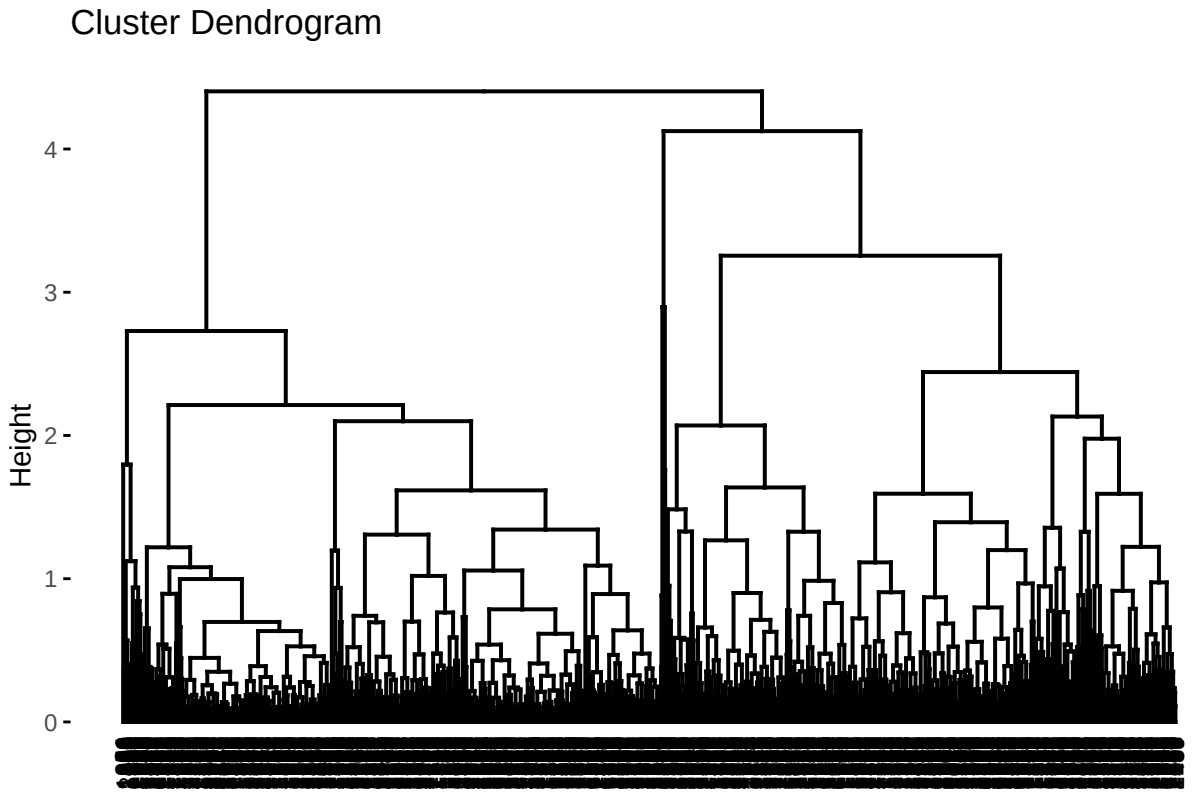
```r
# Plot dendrogram
fviz_dend(hclust_out_complete, cex = 0.6)
```



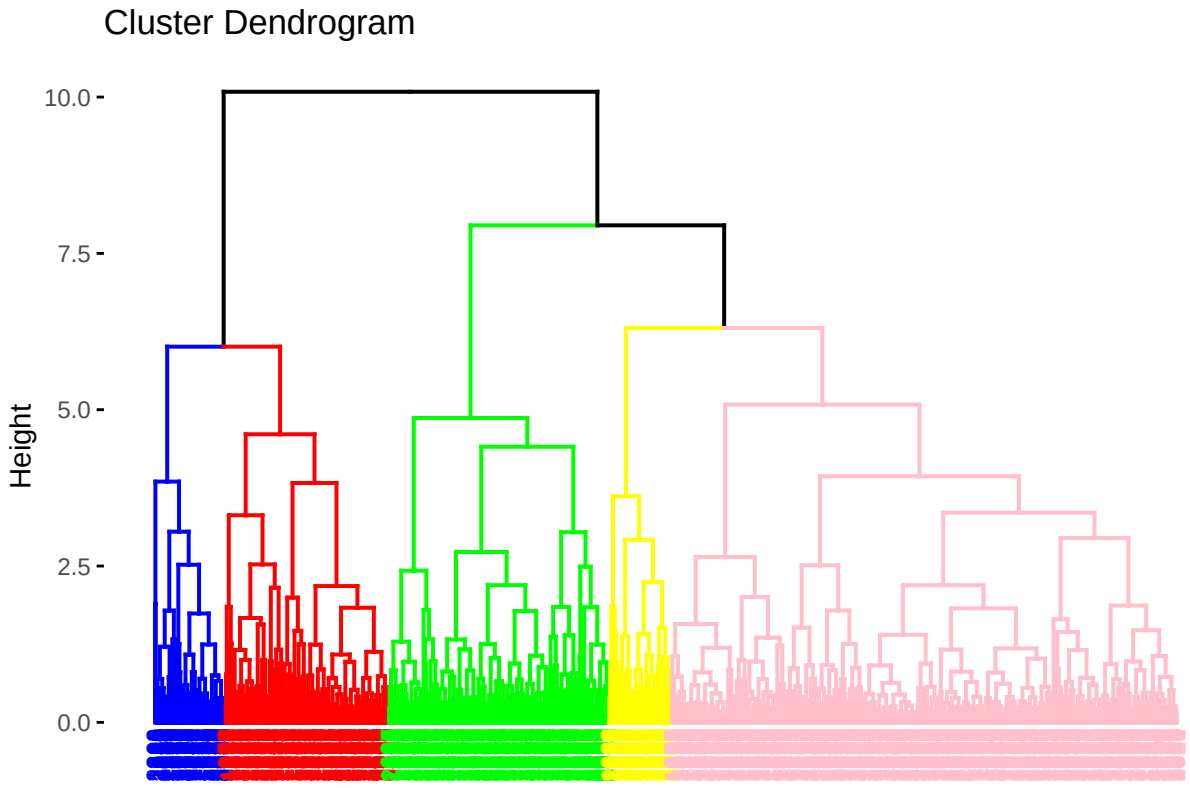

```r
fviz_dend(hclust_out_single, cex = 0.6)
```




```
fviz_dend(hclust_out_average, cex = 0.6)
```



```r
# Choose linkage and number of clusters based on dendrogram
hclust_out <- hclust_out_complete
num_clusters <- 5

# Cut tree to obtain clusters
df$cluster <- cutree(hclust_out, k = num_clusters)

# Plot dendrogram with clusters highlighted
fviz_dend(hclust_out, k = num_clusters, cex = 0.6,
          k_colors = c("blue", "red", "green", "yellow", "pink"))
```



Additional grading criteria:

G3.1 Was all random methods properly seeded? *(2 points)*

Yes, all random methods are properly seeded with seeds 10 to 16.