ML bpalazzo_4

Brandon Palazzo

10/24/2020

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.3

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ggplot2)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
uni_data <- read.csv("Universities.csv")
```

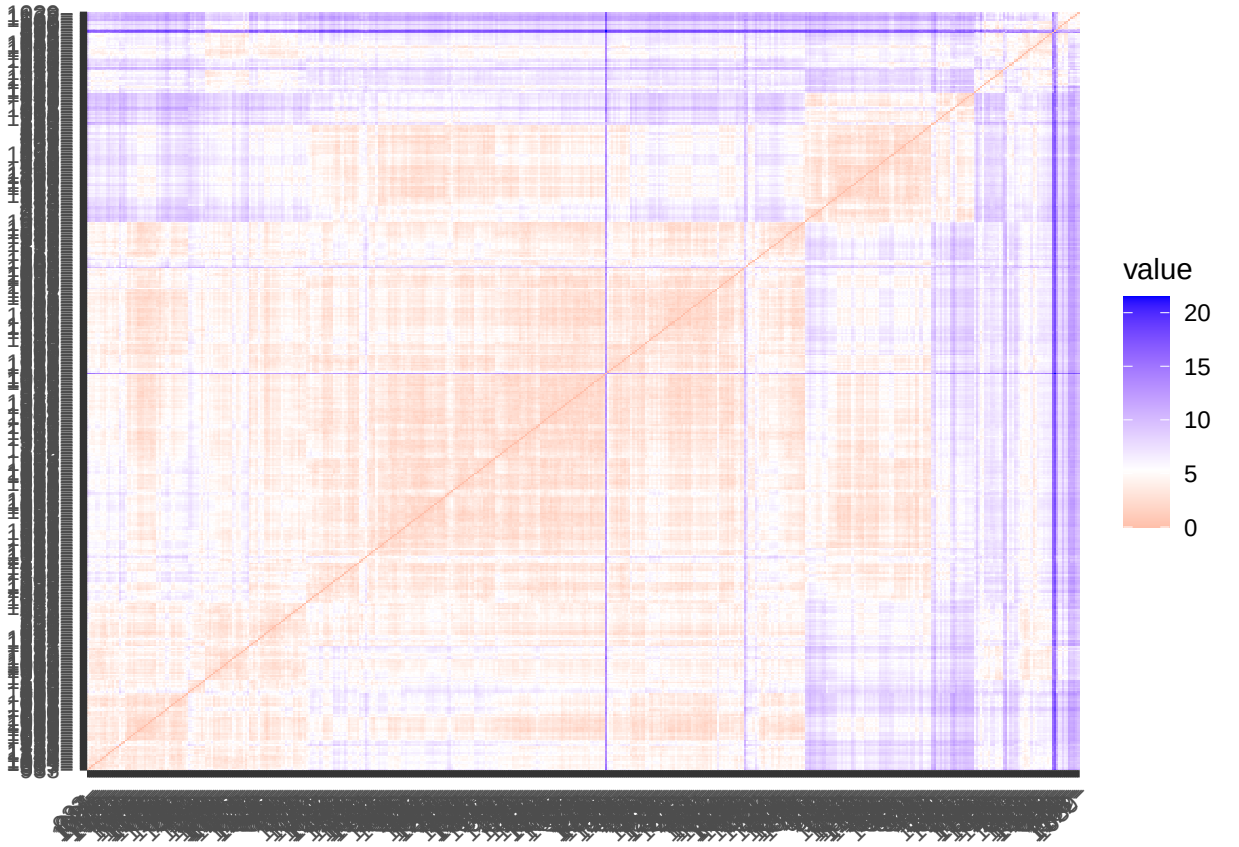
Clean the data

```
  uni_data <- na.omit(uni_data)
```

Normalize the data

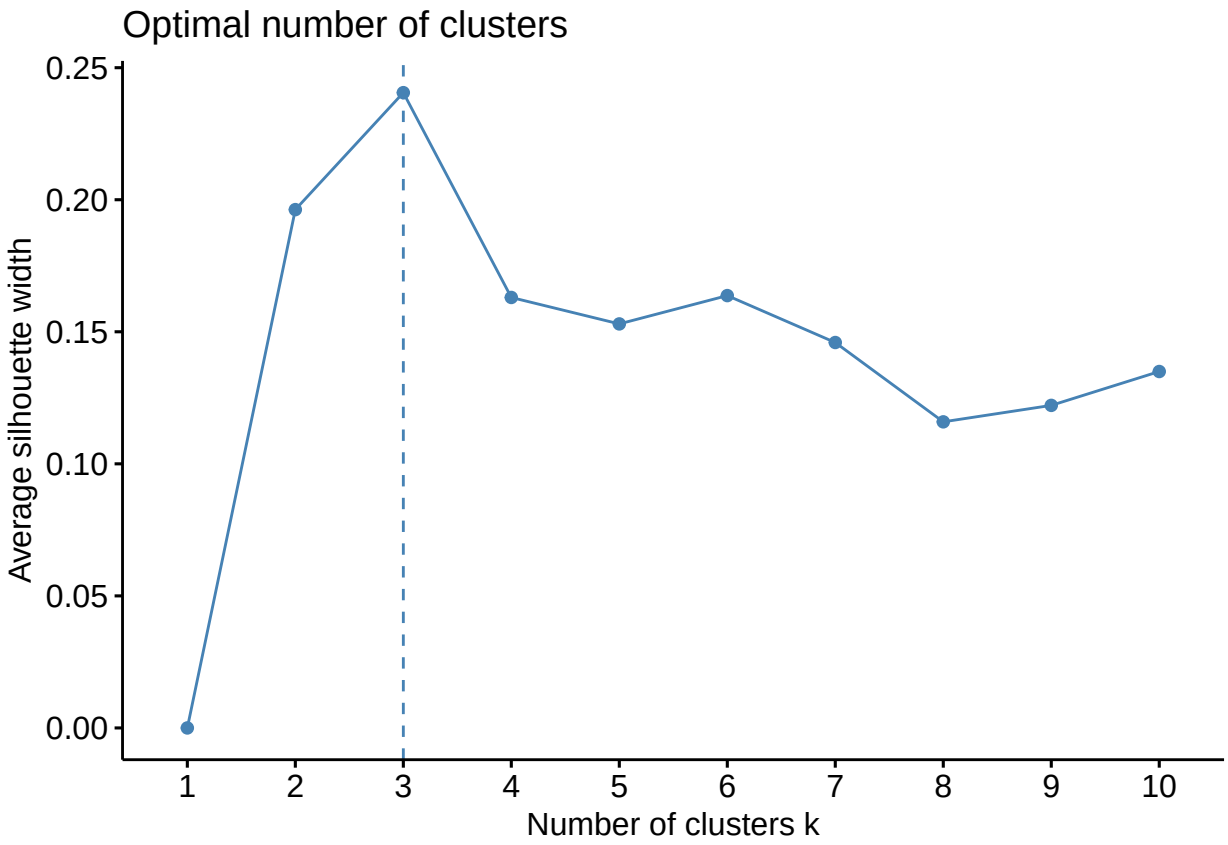
```r
uni_data_num <- uni_data[, c(-1,-2,-3)]


uni_data_num <- scale(uni_data_num)
distance <- get_dist(uni_data_num)
fviz_dist(distance)
```



Determining the optimal k

```r
fviz_nbclust(uni_data_num, kmeans, method = "silhouette")
```



Before finding the optimal k, I predict that a reasonable amount of clusters should be anywhere from 2-4. After performing the silhouette method, we have determined the optimal k to be 3. Three clusters should be a perfect match to properly fit the data without overfitting.

K Means Algorithm

```
k3 <- kmeans(uni_data_num, centers = 3, nstart = 25)



k3$centers
```

```
##   X..appli..rec.d X..appl..accepted X..new.stud..enrolled
## 1     -0.35953828       -0.34918455            -0.3171053
## 2      0.05140256       -0.04367128            -0.1683551
## 3      1.98179657        2.22992267             2.4447222
##   X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
## 1                -0.5020886               -0.5128195      -0.2952142
## 2                 0.8795798                0.8620961      -0.2324464
## 3                 0.1334215                0.2545856       2.5228452
##   X..PT.undergrad in.state.tuition out.of.state.tuition       room      board
## 1      -0.1217682       -0.4036544           -0.5263964 -0.3588740 -0.3938990
## 2      -0.3130216        1.0620416            1.1158839  0.6698444  0.7756859
## 3       1.7486849       -1.0500277           -0.4918168 -0.0388330 -0.1745795
##     add..fees estim..book.costs estim..personal.. X..fac..w.PHD
## 1 -0.05832646       -0.06621454        0.05935933    -0.5322257
```

```
## 2 -0.04496556        0.07122705       -0.39665857     0.7659627
## 3  0.49531762        0.16358567        0.93858632     0.6840794
##   stud..fac..ratio Graduation.rate
## 1        0.2810858      -0.4171456
## 2       -0.7036167       0.8426062
## 3        0.6139980      -0.2538234
```

```
k3$size
```

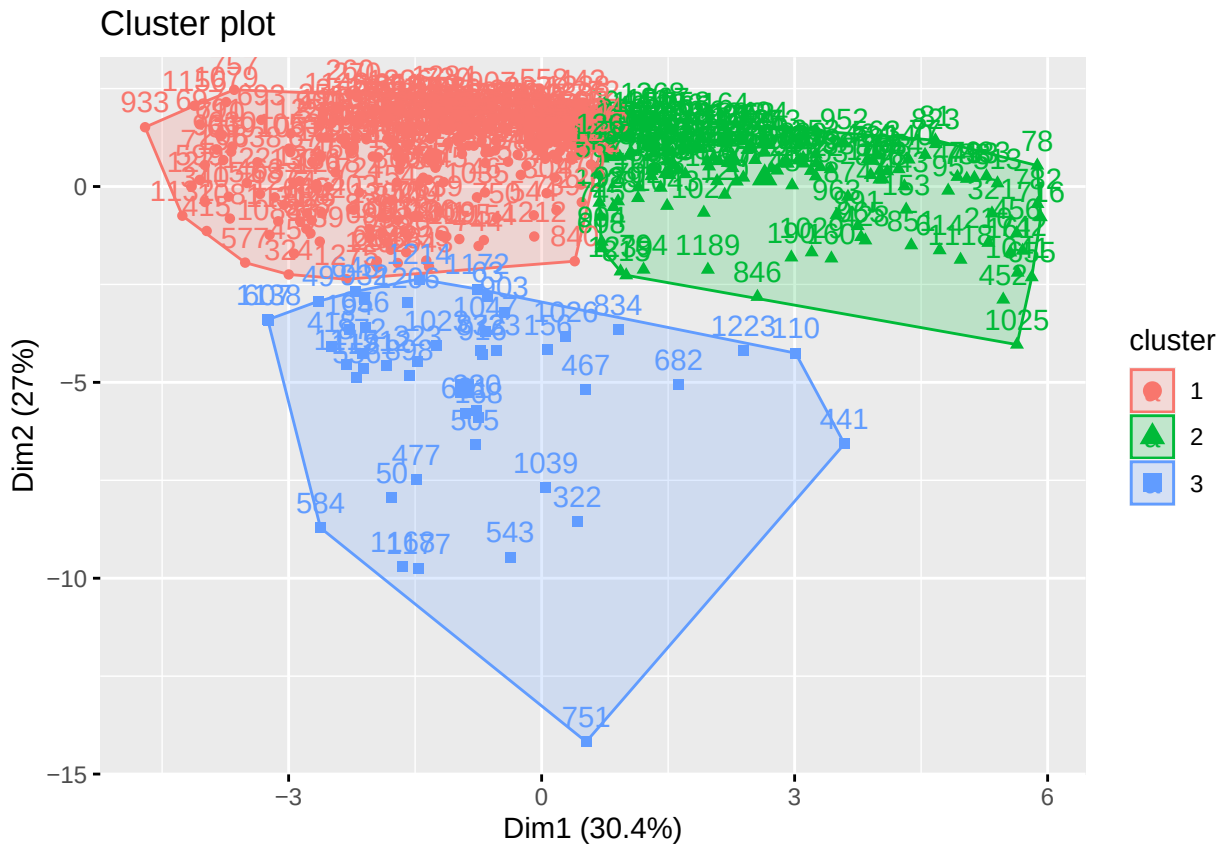
```
## [1] 275 150  46
```

```
k3$cluster
```

```
##    1    3   10   12   22   26   32   38   39   46   49   50   63   77   78   79
##    1    1    2    1    1    1    1    1    1    1    3    3    3    2    2    2
##   81   90   92   95   96   97  108  110  112  120  121  122  123  126  127  130
##    2    2    1    2    1    2    2    3    2    2    1    1    2    1    1    1
##  134  139  140  146  148  149  151  152  153  154  156  158  160  161  164  168
##    1    1    2    1    2    2    2    2    2    1    3    2    2    2    2    3
##  169  172  174  176  181  186  188  190  194  208  210  220  228  230  235  236
##    1    1    2    1    1    2    2    2    3    1    2    1    2    3    1    1
##  239  244  245  246  247  248  250  251  252  258  259  260  262  263  264  265
##    1    1    1    1    2    1    1    1    2    1    1    1    1    1    1    1
##  268  269  270  272  275  277  287  294  297  298  302  304  312  317  319  320
##    1    1    1    1    1    2    1    2    2    2    1    2    1    1    1    1
##  321  322  323  324  326  329  330  331  332  333  336  340  349  352  359  360
##    2    3    3    1    1    2    2    2    1    1    1    1    1    1    1    2
##  362  365  366  368  369  370  376  377  382  387  390  395  398  400  403  405
##    2    2    2    1    1    1    1    1    1    1    1    2    1    1    1    2
##  412  415  418  421  428  433  436  438  441  442  443  444  451  452  454  456
##    1    1    3    2    2    2    2    2    3    1    2    2    1    2    1    2
##  457  460  462  463  464  467  477  479  481  483  484  494  500  505  511  513
##    1    1    1    1    2    3    3    2    2    2    2    2    1    3    1    2
##  514  515  520  522  525  526  527  528  537  538  541  543  544  551  552  556
##    2    1    1    1    1    2    2    1    2    2    1    3    1    1    1    3
##  558  563  564  565  567  571  572  577  578  579  582  584  587  589  591  595
##    1    2    1    1    2    2    2    1    1    1    1    3    1    1    1    1
##  598  599  605  606  607  608  614  615  616  617  618  626  627  629  630  633
##    1    1    1    1    3    2    2    1    1    1    1    1    1    1    1    1
##  638  642  646  647  651  654  655  656  658  659  660  662  666  667  670  673
##    1    1    3    1    1    2    2    3    1    1    1    1    1    1    1    1
##  674  676  677  681  682  684  687  688  689  691  693  696  697  698  702  703
##    1    1    2    3    3    1    1    1    1    1    1    1    1    1    2    1
##  704  705  710  712  713  716  717  720  723  724  726  732  733  736  737  738
##    1    1    1    3    1    2    1    1    3    1    1    2    1    1    1    1
##  739  742  744  745  746  750  751  754  756  757  769  771  772  777  778  782
##    1    2    1    2    1    1    3    1    1    1    2    1    1    2    2    2
##  783  789  792  793  794  795  801  803  804  814  815  823  824  825  826  828
##    2    1    2    2    2    1    2    2    1    1    2    2    2    2    1    1
##  831  833  834  836  837  838  839  840  841  843  844  845  846  848  851  860
##    2    3    3    1    1    1    1    1    1    1    1    1    2    2    2    1
##  869  870  872  874  875  878  879  882  885  889  891  892  893  894  896  898
##    1    1    3    2    1    1    1    2    1    2    1    2    1    1    2    3
```

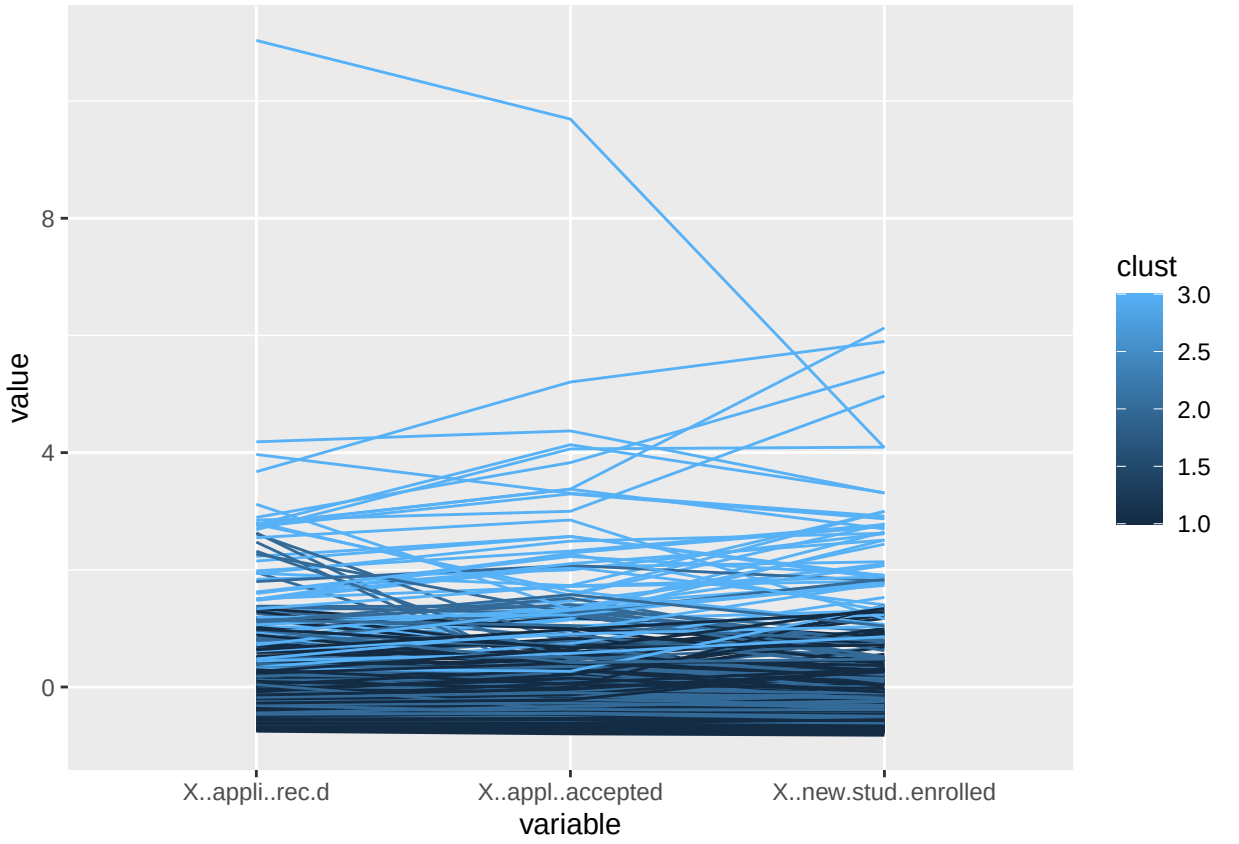
```
##  901  903  904  907  911  912  916  917  928  929  931  932  933  939  943  945
##    1    3    2    1    2    1    3    1    1    1    1    3    1    1    1    2
##  950  952  955  958  959  963  965  967  969  971  974  975  977  978  979  986
##    2    2    2    2    2    2    2    2    2    2    1    2    1    1    2    2
##  987  988  989  991  992  994  996  997 1001 1009 1010 1014 1017 1020 1021 1023
##    1    2    2    2    1    1    2    2    1    1    1    1    1    1    2    3
## 1024 1025 1026 1027 1029 1030 1031 1032 1033 1035 1036 1037 1039 1041 1043 1047
##    1    2    3    2    2    2    1    1    1    1    1    1    3    2    2    3
## 1048 1051 1052 1053 1055 1059 1060 1061 1064 1065 1075 1079 1081 1084 1087 1089
##    1    1    1    1    1    1    2    1    1    2    1    1    1    1    1    1
## 1090 1095 1096 1098 1101 1102 1105 1107 1110 1111 1115 1117 1118 1121 1125 1127
##    1    1    1    1    1    1    1    2    1    1    3    2    2    1    1    1
## 1131 1132 1138 1139 1143 1146 1152 1154 1156 1158 1163 1164 1166 1168 1172 1176
##    1    1    3    1    1    1    1    1    1    1    2    2    1    3    3    1
## 1177 1181 1185 1188 1189 1192 1194 1195 1196 1198 1204 1206 1212 1214 1218 1221
##    3    3    1    1    2    1    2    1    2    1    1    3    1    3    1    1
## 1222 1223 1227 1231 1232 1236 1237 1238 1239 1246 1253 1257 1258 1262 1268 1269
##    2    3    1    1    1    1    2    1    2    2    2    2    1    2    2    2
## 1273 1274 1275 1284 1285 1292 1302
##    1    1    1    1    1    1    1
```

```
fviz_cluster(k3, data = uni_data_num)
```

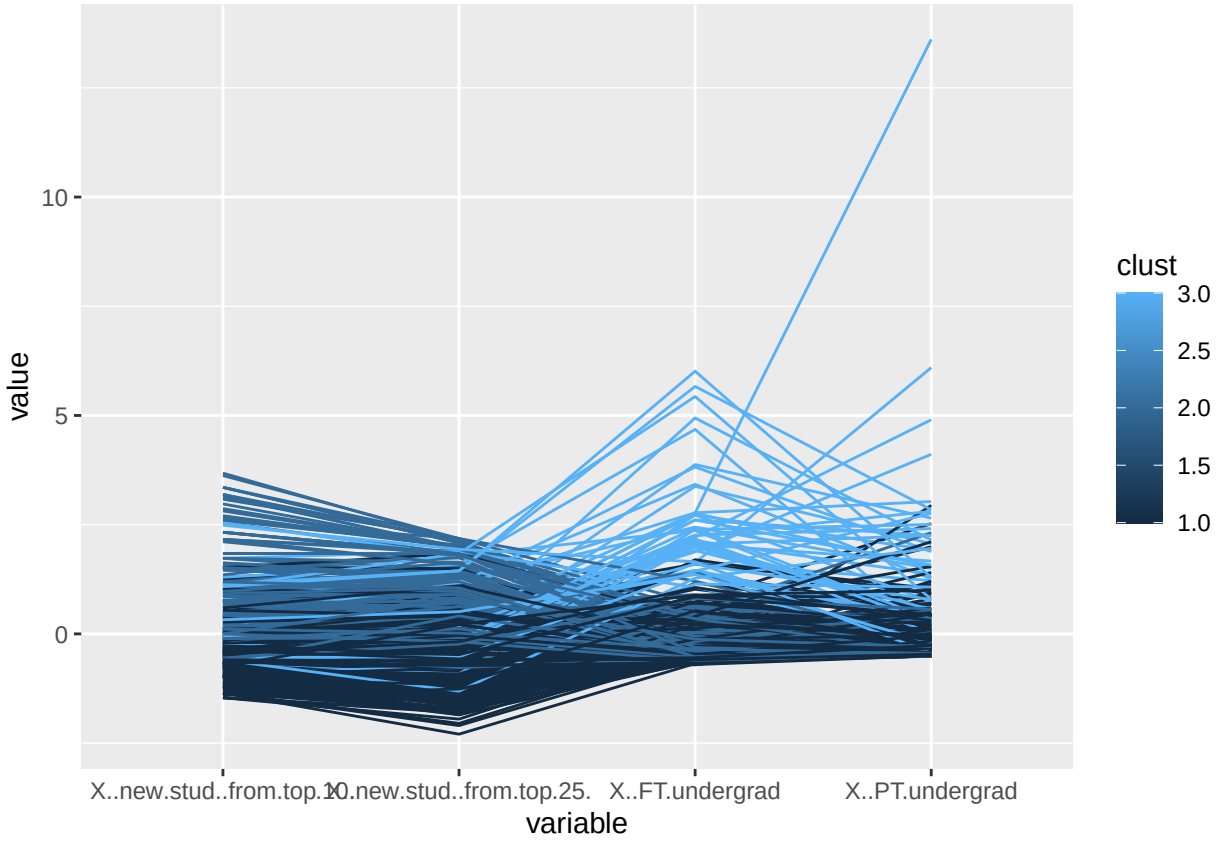



```
clust <- k3$cluster
uni_data_num_clus <- cbind(uni_data_num, clust)
```

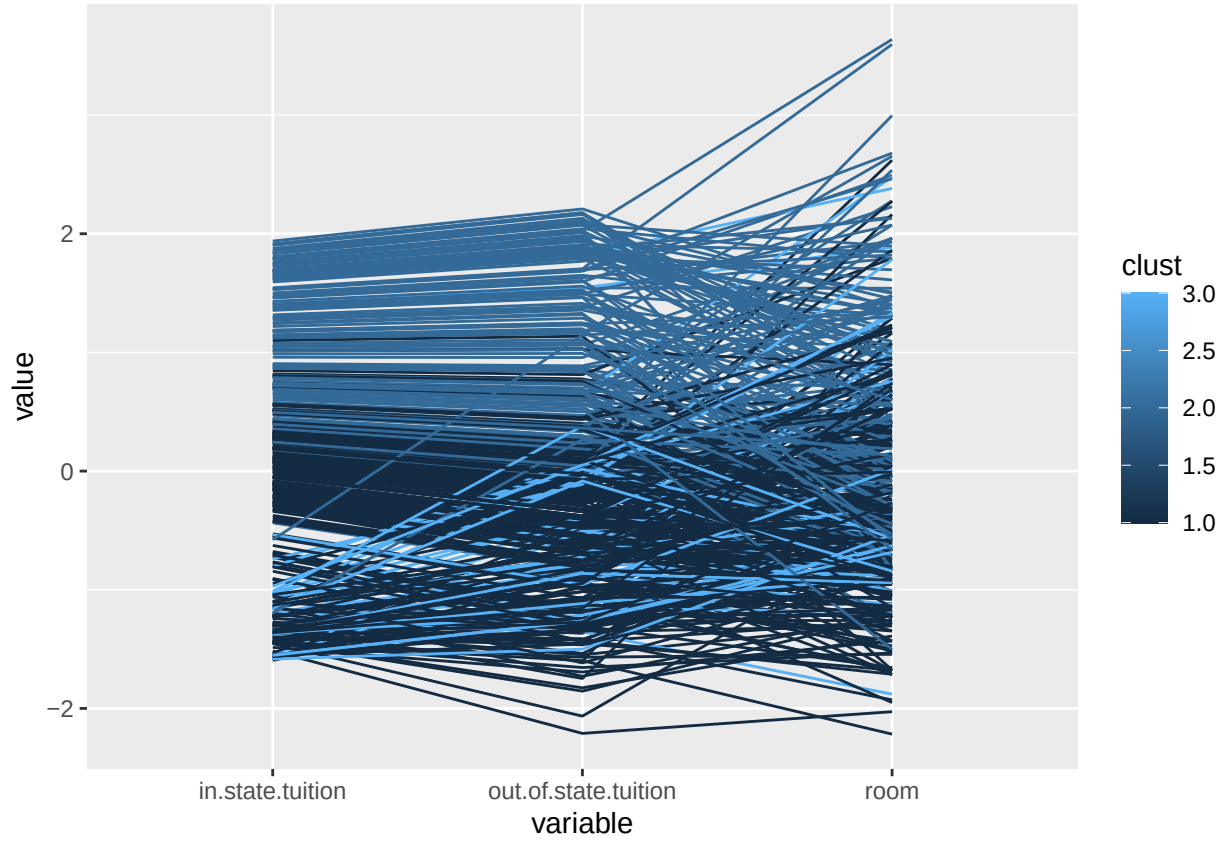
```
ggparcoord(uni_data_num_clus, columns = 1:3, groupColumn = 18)
```



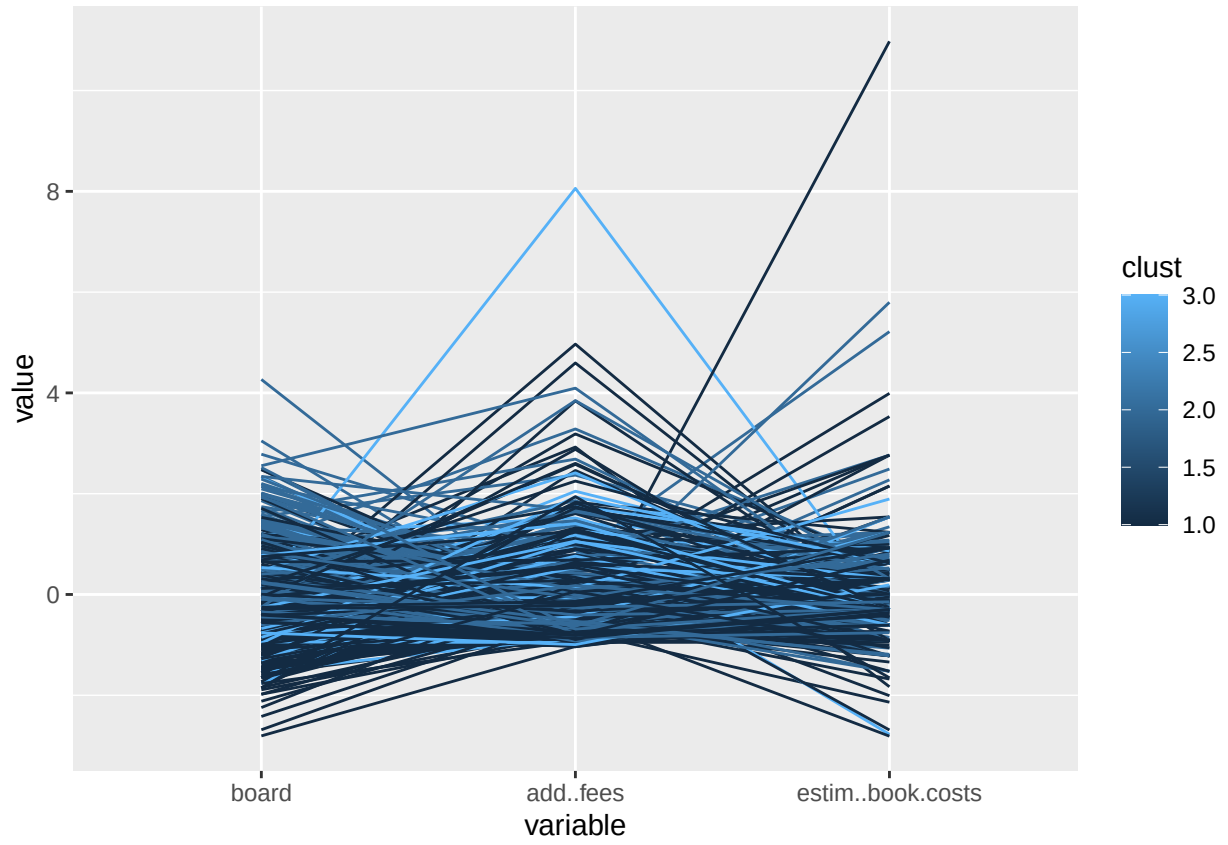

```
ggparcoord(uni_data_num_clus, columns = 4:7, groupColumn = 18)
```



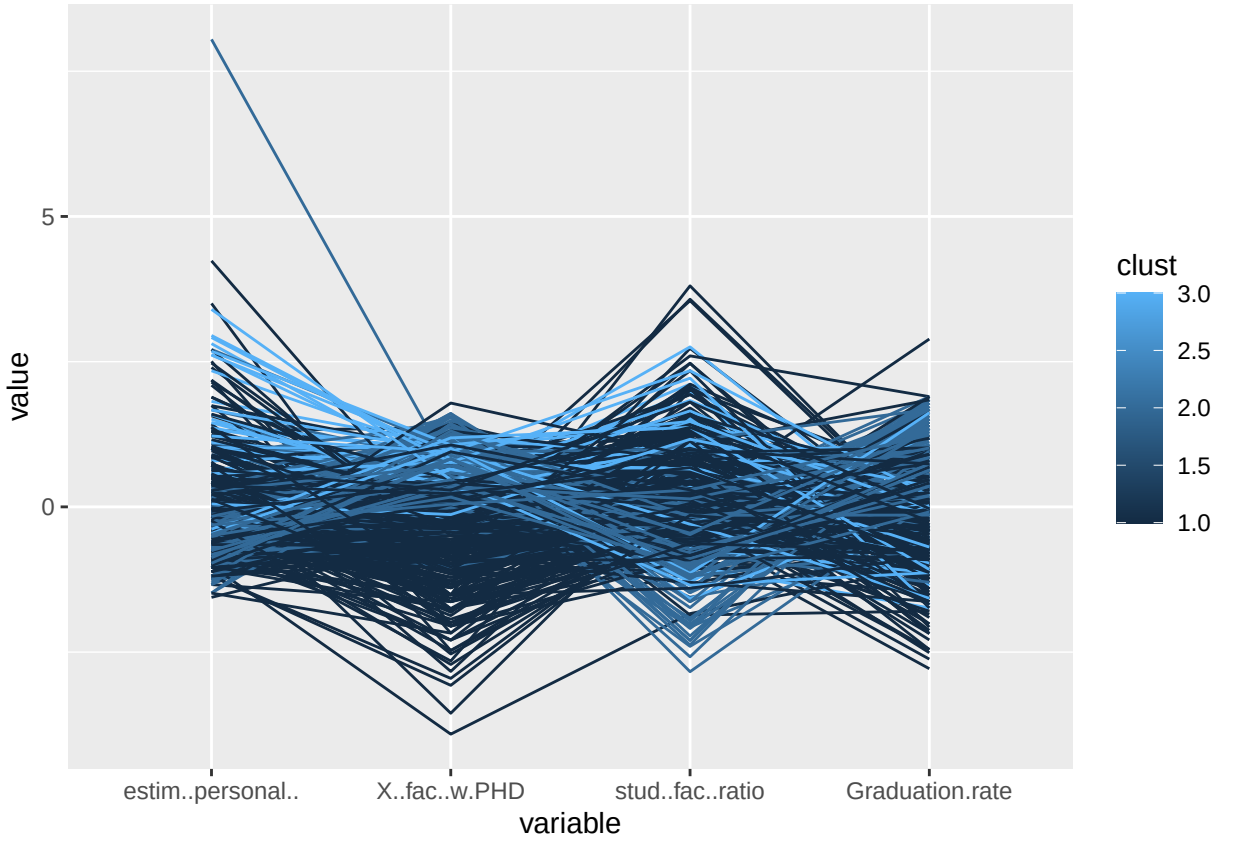

```
ggparcoord(uni_data_num_clus, columns = 8:10, groupColumn = 18)
```




```
ggparcoord(uni_data_num_clus, columns = 11:13, groupColumn = 18)
```




```
ggparcoord(uni_data_num_clus, columns = 14:17, groupColumn = 18)
```



Cluster 1: Small, Local, Commonly Private Schools. These schools have very few/low top 10 and 25 students, graduation rate, and faculty with a PHD. These schools also have low amount of applications received/accepted, new students enrolled, full-time/part-time undergrad, in-state/out-state tuition, room, and board. These schools have around avg admin fees, book cost, and personal loan amounts. Lastly, this cluster has a high faculty to student ratio.

Cluster 2: Expensive, Exclusive, Private Schools. These schools have a low amount of students enrolled, full-time/part-time undergrad, personal student loans,and a very low student/faculty ratio. These schools have an avg amount of applications received/accepted, admin fees, and book costs. Lastly, these schools have very high amount of top 10 and 25 students, in-state/out-state tuition, room, board, faculty with PHD's, and graduation rate.

Cluster 3: Big, Inexpensive, Public Schools. These schools have super low in-state tuition, very low out-state tuition, low board, and low graduation rate. These schools have an avg amount of top 10 students and room costs. Lastly, these schools have super high amount of applications received/accepted, students enrolled, full-time/part-time students, personal loan amounts, very high faculty/student ratio, faculty with PHD, high admin fees, book costs, and slightly high top 25 students.

Clusters by State

```
table(uni_data$State, k3$cluster)
```

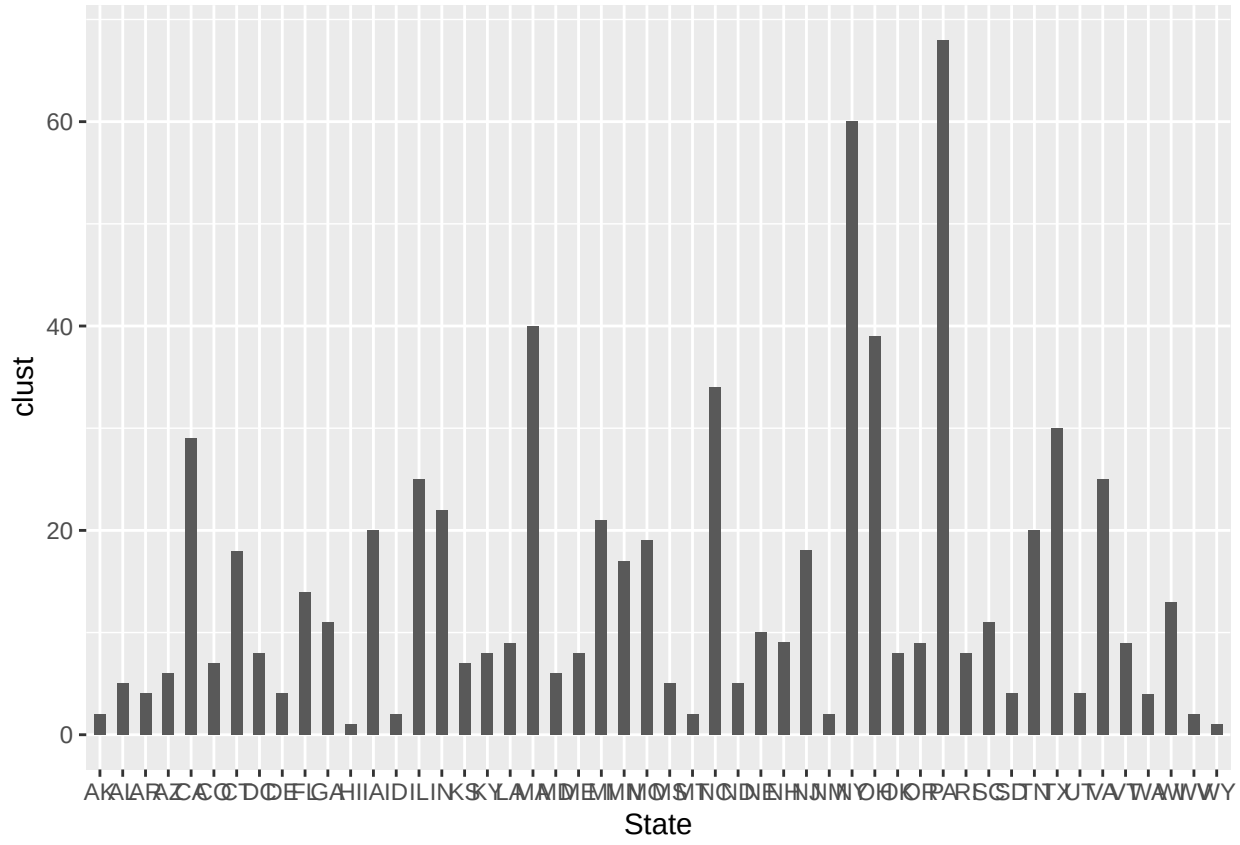
```
##
##       1  2  3
##   AK  2  0  0
##   AL  3  1  0
##   AR  4  0  0
```

```
##    AZ  0  0  2
##    CA  3 10  2
##    CO  5  1  0
##    CT  3  6  1
##    DC  0  4  0
##    DE  1  0  1
##    FL  3  4  1
##    GA  4  2  1
##    HI  1  0  0
##    IA 16  2  0
##    ID  2  0  0
##    IL  7  6  2
##    IN  8  7  0
##    KS  7  0  0
##    KY  4  2  0
##    LA  2  2  1
##    MA  7 12  3
##    MD  1  1  1
##    ME  4  2  0
##    MI  7  4  2
##    MN  6  4  1
##    MO 12  2  1
##    MS  5  0  0
##    MT  2  0  0
##    NC 16  3  4
##    ND  5  0  0
##    NE  5  1  1
##    NH  4  1  1
##    NJ  9  3  1
##    NM  2  0  0
##    NY 18 18  2
##    OH 13  7  4
##    OK  5  0  1
##    OR  1  4  0
##    PA 19 20  3
##    RI  1  2  1
##    SC  7  2  0
##    SD  4  0  0
##    TN 11  3  1
##    TX 14  2  4
##    UT  1  0  1
##    VA  8  4  3
##    VT  5  2  0
##    WA  0  2  0
##    WI  5  4  0
##    WV  2  0  0
##    WY  1  0  0
```

```
uni_data_clus <- cbind(uni_data, clust)
```

```
p <- ggplot(data= uni_data_clus, aes(x= State, y=clust)) +
  geom_bar(stat="identity", width = 0.5)
```

```
p
```



After comparing the state and the number of each cluster, the only relationship is that the larger states tend to have more universities, however because of the amount of schools we had to leave off because of missing data, it is hard to draw a concise conclusion.

```
table(uni_data$Public..1...Private..2., k3$cluster)
```

```
##
##       1   2   3
##   1  84   3  41
##   2 191 147   5
```

After comparing whether a school is public or private and the number of clusters in each category, we notice that cluster 2 is for private schools, cluster 3 is for public schools, and cluster one favors private school, but can also be public as well.

Generally speaking, most universities in the United States fall into three different categories. Those categories include big, state schools, small private or local colleges, and fancy, expensive private schools. For instance, our school, Kent State, would fall into cluster 3 because of our large size, inexpensive tuition, and high amounts of personal loans.

Tufts University

```r
tufts_pt <- mean(uni_data$X..PT.undergrad)

tufts_pt
```

```
## [1] 797.4544
```

```r
new_row <- c("Tufts University", "MA",  2,  7614, 3605, 1205,   60, 90, 4598,   797.4544,   19701,  197
uni_data_tufts <- rbind(uni_data, new_row)

uni_data_tufts_num <- uni_data_tufts[, c(-1,-2,-3)]



uni_data_tufts_num$X..appli..rec.d <- as.numeric(uni_data_tufts_num$X..appli..rec.d)
uni_data_tufts_num$X..appl..accepted <- as.numeric(uni_data_tufts_num$X..appl..accepted)
uni_data_tufts_num$X..new.stud..enrolled <- as.numeric(uni_data_tufts_num$X..new.stud..enrolled)
uni_data_tufts_num$X..new.stud..from.top.10. <- as.numeric(uni_data_tufts_num$X..new.stud..from.top.10.)
uni_data_tufts_num$X..new.stud..from.top.25. <- as.numeric(uni_data_tufts_num$X..new.stud..from.top.25.)
uni_data_tufts_num$X..FT.undergrad <- as.numeric(uni_data_tufts_num$X..FT.undergrad)
uni_data_tufts_num$X..PT.undergrad <- as.numeric(uni_data_tufts_num$X..PT.undergrad)
uni_data_tufts_num$in.state.tuition <- as.numeric(uni_data_tufts_num$in.state.tuition)
uni_data_tufts_num$out.of.state.tuition <- as.numeric(uni_data_tufts_num$out.of.state.tuition)
uni_data_tufts_num$room <- as.numeric(uni_data_tufts_num$room)
uni_data_tufts_num$board <- as.numeric(uni_data_tufts_num$board)
uni_data_tufts_num$add..fees <- as.numeric(uni_data_tufts_num$add..fees)
uni_data_tufts_num$estim..book.costs <- as.numeric(uni_data_tufts_num$estim..book.costs)
uni_data_tufts_num$estim..personal.. <- as.numeric(uni_data_tufts_num$estim..personal..)
uni_data_tufts_num$X..fac..w.PHD <- as.numeric(uni_data_tufts_num$X..fac..w.PHD)
uni_data_tufts_num$stud..fac..ratio <- as.numeric(uni_data_tufts_num$stud..fac..ratio)
uni_data_tufts_num$Graduation.rate <- as.numeric(uni_data_tufts_num$Graduation.rate)



uni_data_tufts_scale <- scale(uni_data_tufts_num)



k3_tufts <- kmeans(uni_data_tufts_scale, centers = 3, nstart = 25)

k3_tufts$cluster[472]
```

```
## 472
##   3
```

Our kmeans clustering algorithm classifies Tufts University as cluster 2; small, private, exclusive university.