```
library(tidyverse)
```

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

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

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

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

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

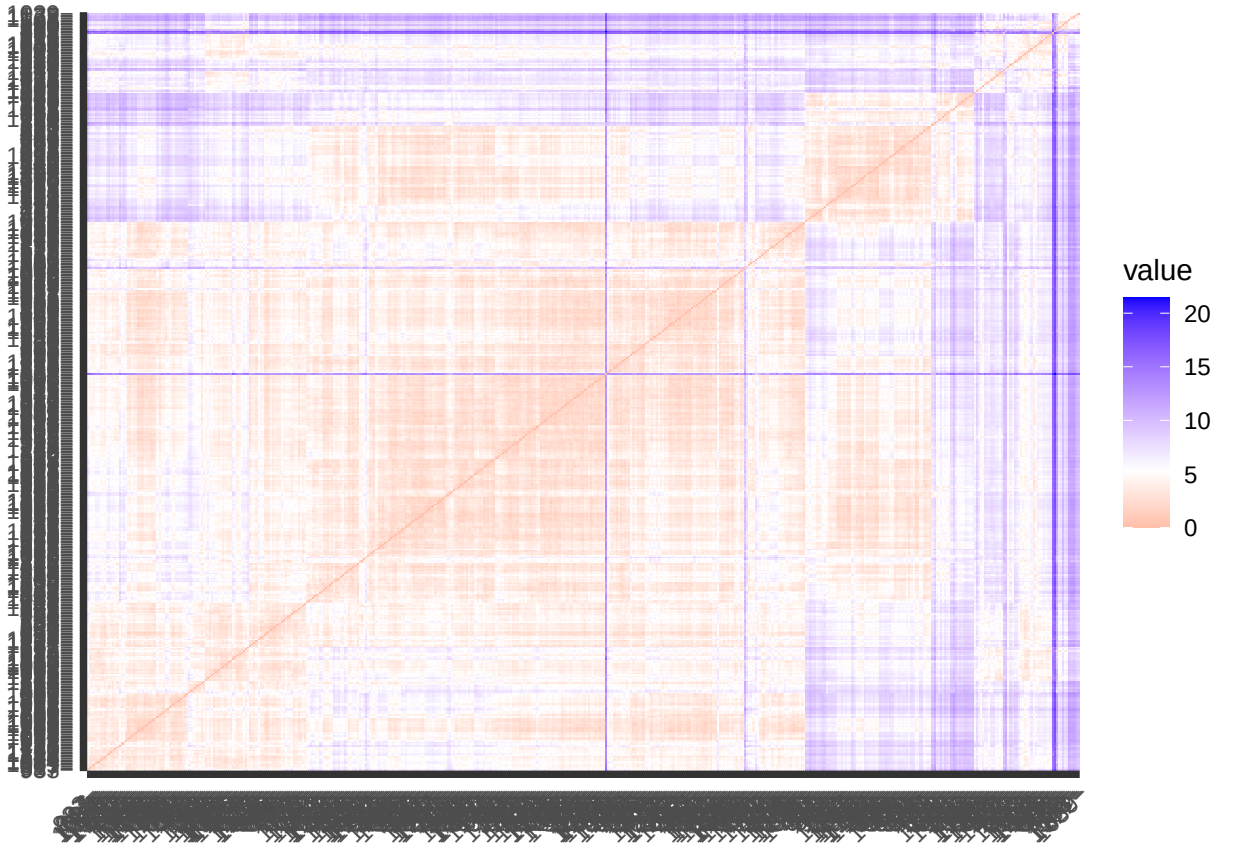
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ISLR)
Universities <- read.csv("Universities.csv")
summary(Universities)
```

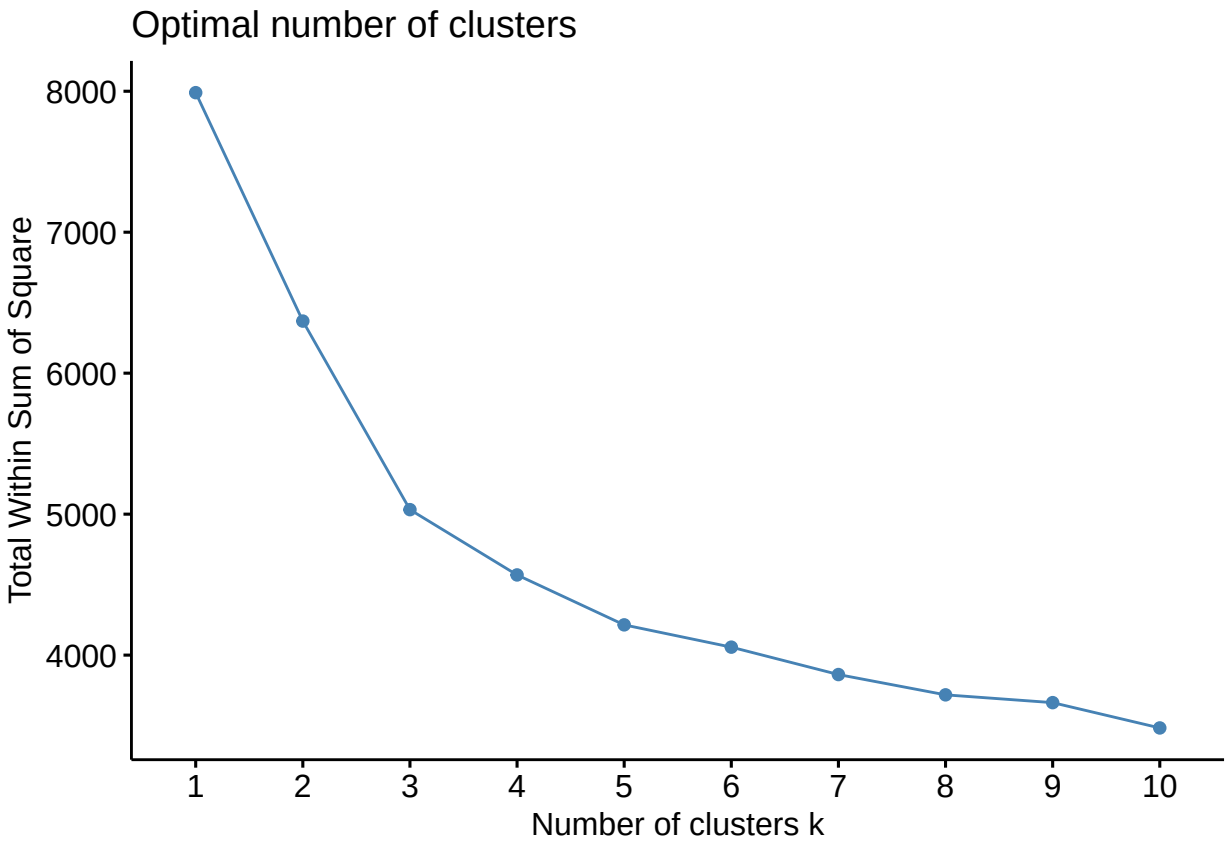
```
##  College.Name          State           Public..1...Private..2.
##  Length:1302        Length:1302        Min.   :1.000
##  Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :2.000
##                                        Mean   :1.639
##                                        3rd Qu.:2.000
##                                        Max.   :2.000
##
##  X..appli..rec.d   X..appl..accepted X..new.stud..enrolled
##  Min.   :   35.0   Min.   :   35.0   Min.   :  18.0
##  1st Qu.:  695.8   1st Qu.:  554.5   1st Qu.: 236.0
##  Median : 1470.0   Median : 1095.0   Median : 447.0
##  Mean   : 2752.1   Mean   : 1870.7   Mean   : 778.9
##  3rd Qu.: 3314.2   3rd Qu.: 2303.0   3rd Qu.: 984.0
##  Max.   :48094.0   Max.   :26330.0   Max.   :7425.0
##  NA's   :10        NA's   :11        NA's   :5
##  X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
##  Min.   : 1.00             Min.   :  6.00            Min.   :   59
##  1st Qu.:13.00             1st Qu.: 36.75            1st Qu.:  966
```

```
##  Median :21.00             Median : 50.00            Median : 1812
##  Mean   :25.67             Mean   : 52.35            Mean   : 3693
##  3rd Qu.:32.00             3rd Qu.: 66.00            3rd Qu.: 4540
##  Max.   :98.00             Max.   :100.00            Max.   :31643
##  NA's   :235               NA's   :202               NA's   :3
##  X..PT.undergrad   in.state.tuition out.of.state.tuition      room
##  Min.   :    1.0   Min.   :  480    Min.   : 1044        Min.   : 500
##  1st Qu.:  131.2   1st Qu.: 2580    1st Qu.: 6111        1st Qu.:1710
##  Median :  472.0   Median : 8050    Median : 8670        Median :2200
##  Mean   : 1081.5   Mean   : 7897    Mean   : 9277        Mean   :2515
##  3rd Qu.: 1313.0   3rd Qu.:11600    3rd Qu.:11659        3rd Qu.:3040
##  Max.   :21836.0   Max.   :25750    Max.   :25750        Max.   :7400
##  NA's   :32        NA's   :30       NA's   :20           NA's   :321
##      board        add..fees      estim..book.costs estim..personal..
##  Min.   : 531   Min.   :   9.0   Min.   :  90      Min.   :  75
##  1st Qu.:1619   1st Qu.: 130.0   1st Qu.: 480      1st Qu.: 900
##  Median :1980   Median : 264.5   Median : 502      Median :1250
##  Mean   :2061   Mean   : 392.0   Mean   : 550      Mean   :1389
##  3rd Qu.:2402   3rd Qu.: 480.0   3rd Qu.: 600      3rd Qu.:1794
##  Max.   :6250   Max.   :4374.0   Max.   :2340      Max.   :6900
##  NA's   :498    NA's   :274      NA's   :48        NA's   :181
##  X..fac..w.PHD    stud..fac..ratio Graduation.rate
##  Min.   :  8.00   Min.   : 2.30    Min.   :  8.00
##  1st Qu.: 57.00   1st Qu.:11.80    1st Qu.: 47.00
##  Median : 71.00   Median :14.30    Median : 60.00
##  Mean   : 68.65   Mean   :14.86    Mean   : 60.41
##  3rd Qu.: 82.00   3rd Qu.:17.60    3rd Qu.: 74.00
##  Max.   :105.00   Max.   :91.80    Max.   :118.00
##  NA's   :32       NA's   :2        NA's   :98
```

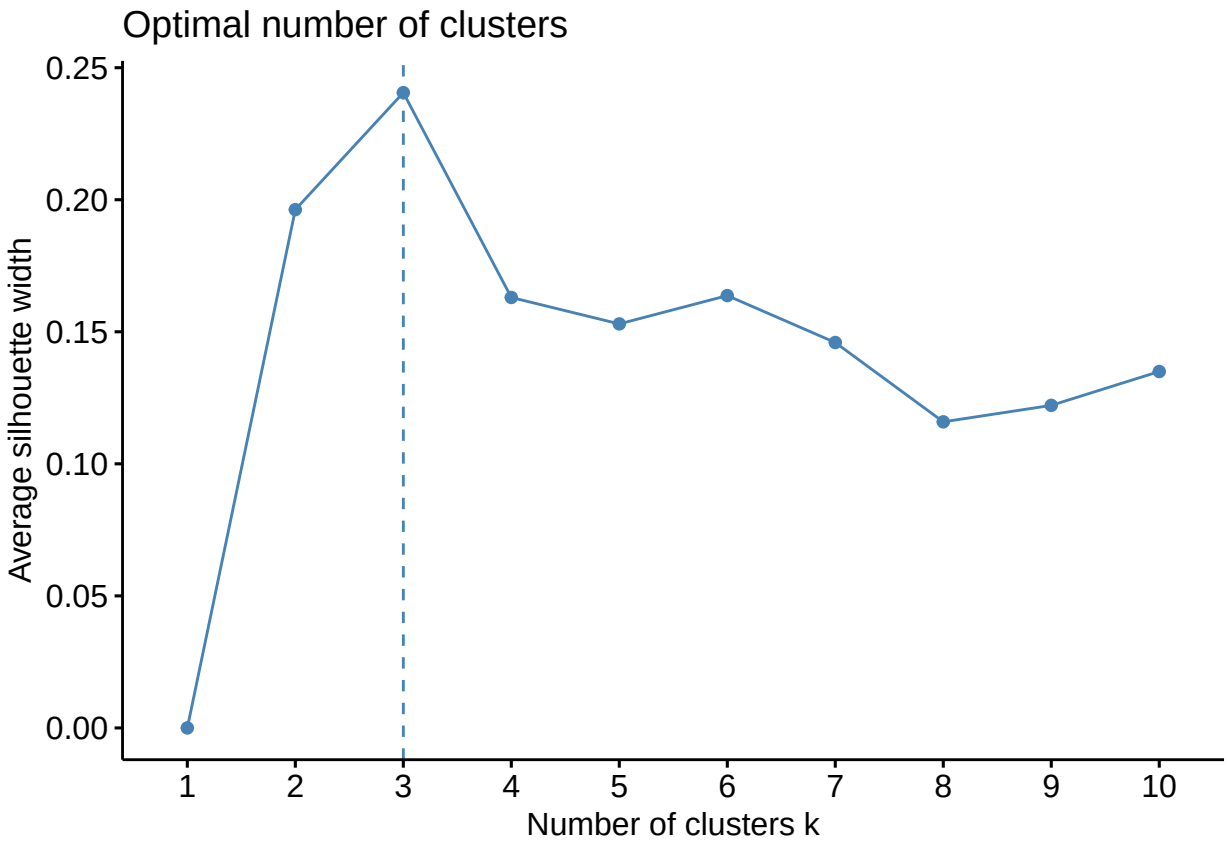
```r
# Part 1. Remove all rows with N/A's
Univ_Cleaned <-Universities[complete.cases(Universities),] # Removes all cases with N/A
Univ_Cleaned$Public..1...Private..2. <- ifelse(Univ_Cleaned$Public..1...Private..2. == 1, "Public", "Pr

# The following code prepares the data for analysis
Univ_data <- Univ_Cleaned[,c(4:20)] # creates new dataset without categorical
scaleduniv <- scale(Univ_data) # scales new dataset
distance <- get_dist(scaleduniv)
fviz_dist(distance) # looks at distance
```




```
# Part 2: K-means
# The first part of code determines how many clusters to use
fviz_nbclust(scaleduniv, kmeans, method = "wss") #wss method
```



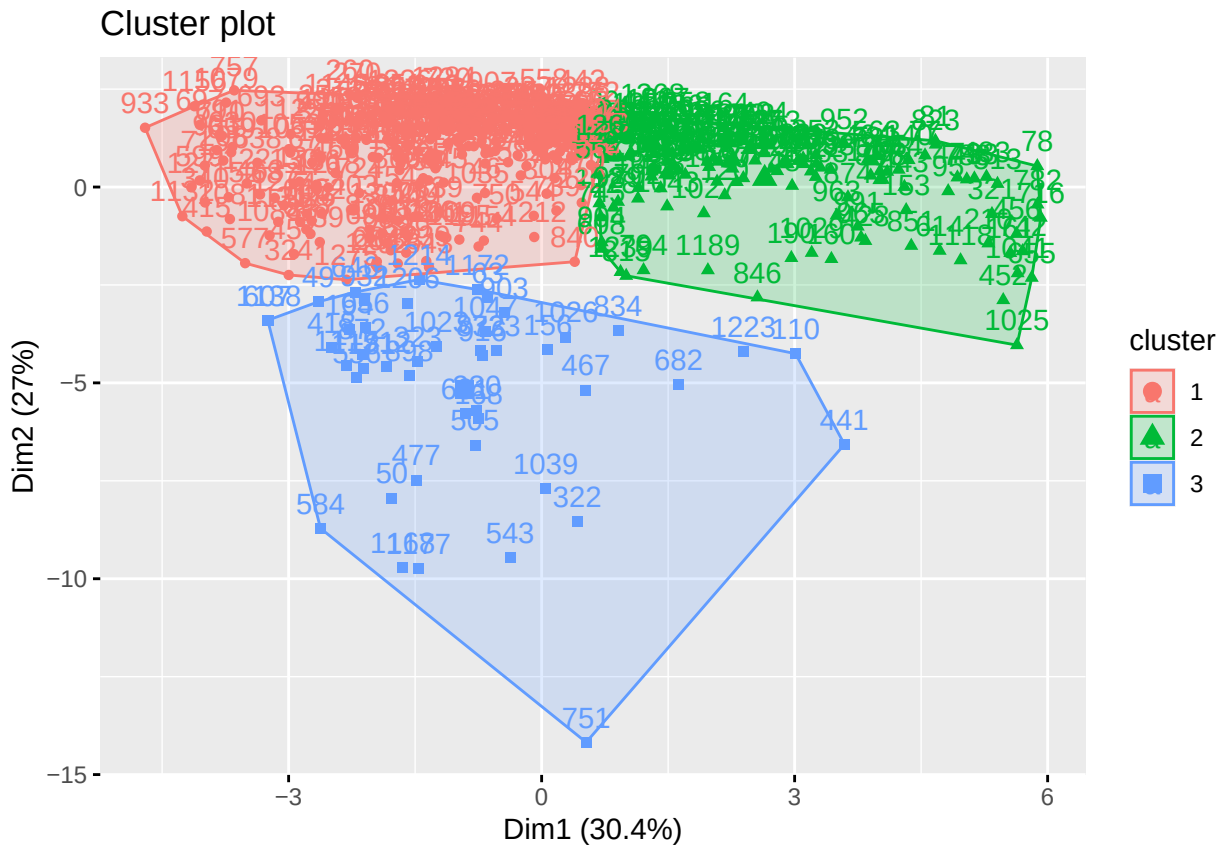

```
fviz_nbclust(scaleduniv, kmeans, method = "silhouette") #silhouette method
```




```
# The WSS suggests that 3 or 4 clusters would be best. The silhouette method suggests 3 would be best.

k3<-kmeans(scaleduniv, centers = 3, nstart = 25) #kmeans formula
str(k3)
```

```
## List of 9
##  $ cluster     : Named int [1:471] 1 1 2 1 1 1 1 1 1 1 ...
##   ..- attr(*, "names")= chr [1:471] "1" "3" "10" "12" ...
##  $ centers     : num [1:3, 1:17] -0.3595 0.0514 1.9818 -0.3492 -0.0437 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "1" "2" "3"
##   .. ..$ : chr [1:17] "X..appli..rec.d" "X..appl..accepted" "X..new.stud..enrolled" "X..new.stud..fr
##  $ totss       : num 7990
##  $ withinss    : num [1:3] 2562 1425 1045
##  $ tot.withinss: num 5032
##  $ betweenss   : num 2958
##  $ size        : int [1:3] 275 150 46
##  $ iter        : int 3
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
fviz_cluster(k3, data = scaleduniv) # displays the clusters
```




```
# Part 3. Compare the summary stats for each cluster in k3
summary(k3)
```

```
##              Length Class  Mode
## cluster      471    -none- numeric
## centers       51    -none- numeric
## totss          1    -none- numeric
## withinss       3    -none- numeric
## tot.withinss   1    -none- numeric
## betweenss      1    -none- numeric
## size           3    -none- numeric
## iter           1    -none- numeric
## ifault         1    -none- numeric
```

```
k3$size
```

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

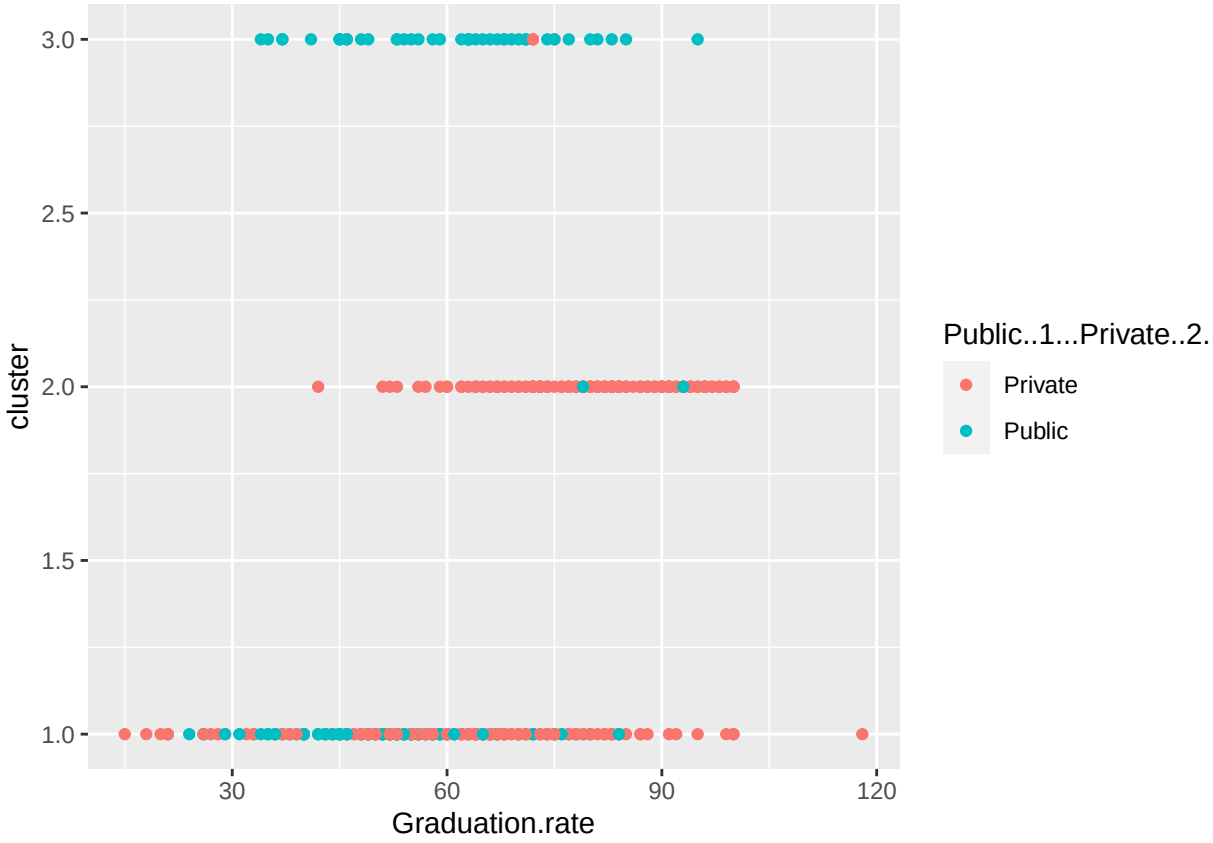
```
# The sizes show that kmeans considers clusters 1 to be much larger than 2, and 2 to be much larger tha
k3$centers
```

```
##   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
```

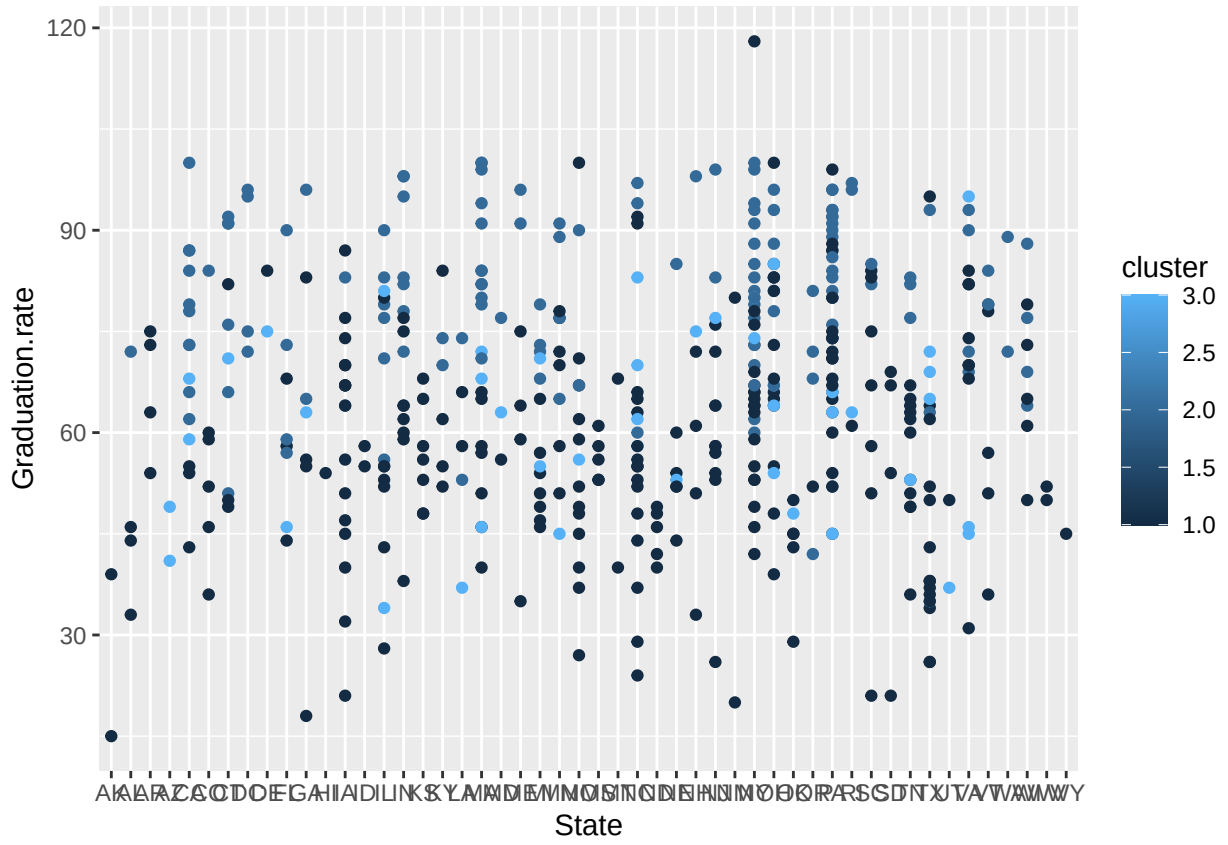
```
## 3      1.98179657        2.22992267             2.4447222
##   X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad
## 1                -0.5020886                -0.5128195      -0.2952142
## 2                 0.8795798                 0.8620961      -0.2324464
## 3                 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
```

```
# Universities in cluster 1 have relatively low graduation rates, relatively low room and board fees, l
# Universities in cluster 2 have relatively higher graduation rates, relatively high room and board fee
# Universities in cluster 3 have relatively low graduation rates, relatively low room and board fees, l
```

```
# Part 4 and 5:
Univ_w_cl <- cbind(Univ_Cleaned, k3$cluster) # Returns the cluster to the original dataset
Univ_w_cl$cluster <- Univ_w_cl$`k3$cluster`

# Uses ggplot to show public and private
ggplot() +
  geom_point(data = Univ_w_cl,
             mapping = aes(x = Graduation.rate,
             y = cluster,
             colour = Public..1...Private..2.))
```




```
# Cluster 1 is mostly private, and Cluster 3 is almost entirely public. Cluster 2 is mixed but is mostly
ggplot() +
  geom_point(data = Univ_w_cl,
             mapping = aes(x = State,
                           y = Graduation.rate,
                           colour = cluster))
```




```
# On close inspection of the data, the small set of cluster 3 seems to spread out by state. Those most

#Based on this data, I would suggest that cluster 1 is mostly private, exclusive universities. Cluster
```

```
# Part 6.
# Tufts is likely cluster 2 based on data. Tufts is a private institution with a low acceptance rate an
# It's missing value is pt.undergrad
# the value for this scaled is -0.3130216
```