Universities Clustering

Mark Pirogowicz

October 24, 2019

library(factoextra)

## Loading required package: ggplot2

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------------------------------------------------------ tidyverse 1.2.1 --

## v tibble 2.1.1 v purrr 0.2.4  
## v tidyr 1.0.0 v dplyr 0.8.3  
## v readr 1.1.1 v stringr 1.3.1  
## v tibble 2.1.1 v forcats 0.3.0

## -- Conflicts --------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x psych::%+%() masks ggplot2::%+%()  
## x psych::alpha() masks ggplot2::alpha()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(usmap)

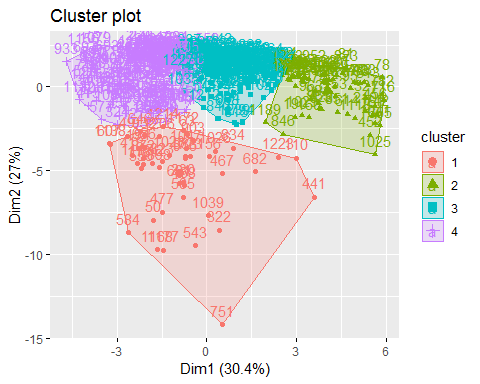
Lets pull in the data and do some dropping of NA rows and some scaling.

#Pull in data  
fullData <- read.csv("C:/Users/mpirogow/Documents/Perm5/06 - School/ML/Clustering/Universities.csv")  
  
#remove any columns that hanve NAs  
dataWithoutNA <- na.omit(fullData)  
dataWithoutNA$Public..1...Private..2. <- as.factor(dataWithoutNA$Public..1...Private..2.)  
  
#Lets scale our variables  
finalCleanedData <- as.data.frame(scale(dataWithoutNA[4:20]))  
finalCleanedData <- cbind(dataWithoutNA[1:3],finalCleanedData)  
  
#One final check to make sure data is scaled  
summary(finalCleanedData)

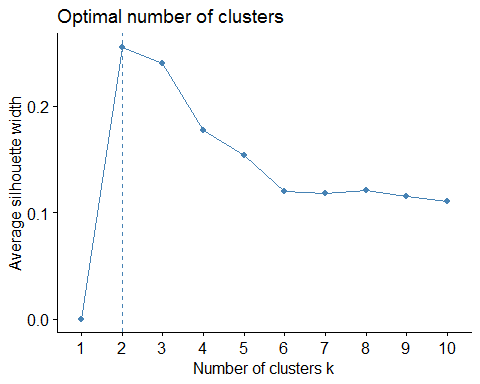
## College.Name State Public..1...Private..2.  
## Trinity College : 4 PA : 42 1:128   
## Augustana College : 2 NY : 38 2:343   
## Monmouth College : 2 OH : 24   
## University of St. Thomas: 2 NC : 23   
## Westminster College : 2 MA : 22   
## Adams State College : 1 TX : 20   
## (Other) :458 (Other):302   
## X..appli..rec.d X..appl..accepted X..new.stud..enrolled  
## Min. :-0.7538 Min. :-0.7996 Min. :-0.8232   
## 1st Qu.:-0.5758 1st Qu.:-0.5701 1st Qu.:-0.5643   
## Median :-0.3686 Median :-0.3339 Median :-0.3688   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.1755 3rd Qu.: 0.1570 3rd Qu.: 0.1265   
## Max. :11.0349 Max. : 9.6923 Max. : 6.1283   
##   
## X..new.stud..from.top.10. X..new.stud..from.top.25. X..FT.undergrad   
## Min. :-1.4618 Min. :-2.29537 Min. :-0.7097   
## 1st Qu.:-0.7042 1st Qu.:-0.77010 1st Qu.:-0.5450   
## Median :-0.2713 Median :-0.08127 Median :-0.3958   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.4322 3rd Qu.: 0.65676 3rd Qu.: 0.1055   
## Max. : 3.6791 Max. : 2.18202 Max. : 6.0139   
##   
## X..PT.undergrad in.state.tuition out.of.state.tuition  
## Min. :-0.51524 Min. :-1.59488 Min. :-2.2105   
## 1st Qu.:-0.46316 1st Qu.:-1.04338 1st Qu.:-0.7619   
## Median :-0.32246 Median : 0.08182 Median :-0.1102   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.04628 3rd Qu.: 0.69594 3rd Qu.: 0.6287   
## Max. :13.61017 Max. : 1.93833 Max. : 2.2091   
##   
## room board add..fees estim..book.costs  
## Min. :-2.2170 Min. :-2.80658 Min. :-1.0370 Min. :-2.8114   
## 1st Qu.:-0.6746 1st Qu.:-0.65614 1st Qu.:-0.6787 1st Qu.:-0.2989   
## Median :-0.1838 Median :-0.07046 Median :-0.2783 Median :-0.2989   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.6196 3rd Qu.: 0.52581 3rd Qu.: 0.3006 3rd Qu.: 0.3139   
## Max. : 3.6384 Max. : 4.26746 Max. : 8.0594 Max. :10.9766   
##   
## estim..personal.. X..fac..w.PHD stud..fac..ratio Graduation.rate   
## Min. :-1.5574 Min. :-3.9127 Min. :-2.8374 Min. :-2.7863   
## 1st Qu.:-0.6775 1st Qu.:-0.6125 1st Qu.:-0.6829 1st Qu.:-0.6923   
## Median :-0.1642 Median : 0.1675 Median :-0.1443 Median : 0.0241   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.4225 3rd Qu.: 0.8276 3rd Qu.: 0.6380 3rd Qu.: 0.7405   
## Max. : 8.0488 Max. : 1.7876 Max. : 3.8056 Max. : 2.8896   
##

Lets remove unneeded variables, figure out the optimal k, and run k means. Lets also break out the summary statistics by cluster

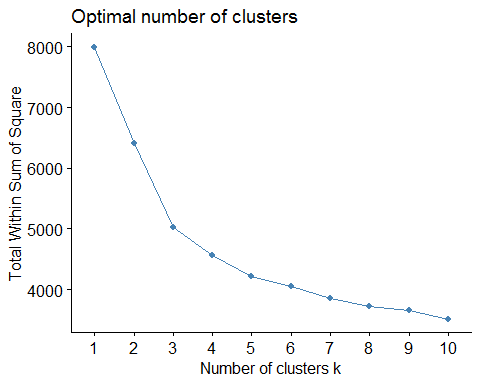
#ability to repeat  
set.seed(123)  
  
#remove the variables we are not clusting by  
data <- finalCleanedData[4:20]  
  
#create our cluster with 4 as the k (found below)  
k4 <- kmeans(x = data, centers = 4, nstart = 25)  
  
#lets toss the cluster number on the dataset in order to break out summary statistics later  
dataWithCluster <- cbind(finalCleanedData,k4$cluster)  
  
#lets visualize the clusters  
fviz\_cluster(object = k4, data = data)



#2 clusters are optimal  
fviz\_nbclust(data, kmeans, method = "silhouette")



#four are optimal here. I chose 4 based on data apperance  
fviz\_nbclust(data, kmeans, method = "wss")



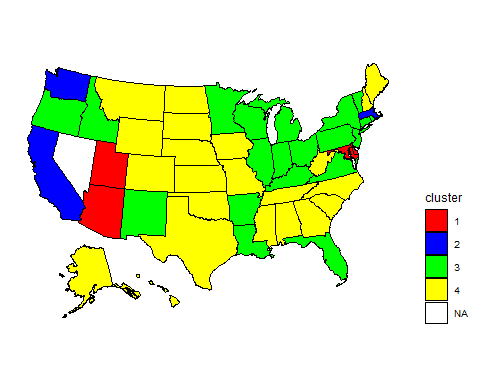
#summary statistics broken down by cluster (psych package)  
describeBy(dataWithCluster, dataWithCluster$`k4$cluster`)

##   
## Descriptive statistics by group   
## group: 1  
## vars n mean sd median trimmed mad  
## College.Name\* 1 46 836.89 344.11 1013.50 881.53 181.62  
## State\* 2 46 28.15 13.20 29.00 28.79 14.83  
## Public..1...Private..2.\* 3 46 1.11 0.31 1.00 1.03 0.00  
## X..appli..rec.d 4 46 1.98 1.69 1.61 1.76 1.31  
## X..appl..accepted 5 46 2.23 1.59 1.86 2.03 1.04  
## X..new.stud..enrolled 6 46 2.44 1.26 2.12 2.27 1.03  
## X..new.stud..from.top.10. 7 46 0.13 0.83 -0.14 0.03 0.68  
## X..new.stud..from.top.25. 8 46 0.25 0.84 0.07 0.22 0.77  
## X..FT.undergrad 9 46 2.52 1.20 2.19 2.35 0.80  
## X..PT.undergrad 10 46 1.75 2.24 1.33 1.38 1.38  
## in.state.tuition 11 46 -1.05 0.67 -1.20 -1.19 0.25  
## out.of.state.tuition 12 46 -0.49 0.69 -0.71 -0.57 0.52  
## room 13 46 -0.04 1.01 -0.49 -0.13 0.66  
## board 14 46 -0.17 0.80 -0.28 -0.22 0.69  
## add..fees 15 46 0.50 1.46 0.21 0.32 1.16  
## estim..book.costs 16 46 0.16 0.73 0.25 0.17 0.80  
## estim..personal.. 17 46 0.94 1.09 0.72 0.89 1.08  
## X..fac..w.PHD 18 46 0.68 0.37 0.77 0.71 0.36  
## stud..fac..ratio 19 46 0.61 1.07 0.88 0.64 0.93  
## Graduation.rate 20 46 -0.25 0.80 -0.14 -0.25 0.82  
## k4$cluster 21 46 1.00 0.00 1.00 1.00 0.00  
## min max range skew kurtosis se  
## College.Name\* 32.00 1226.00 1194.00 -1.09 -0.06 50.74  
## State\* 4.00 46.00 42.00 -0.36 -1.11 1.95  
## Public..1...Private..2.\* 1.00 2.00 1.00 2.43 4.01 0.05  
## X..appli..rec.d 0.31 11.03 10.72 3.31 15.10 0.25  
## X..appl..accepted 0.27 9.69 9.42 2.33 8.09 0.24  
## X..new.stud..enrolled 0.86 6.13 5.27 1.26 1.21 0.19  
## X..new.stud..from.top.10. -0.87 2.54 3.41 1.17 0.88 0.12  
## X..new.stud..from.top.25. -1.31 1.94 3.25 0.43 -0.75 0.12  
## X..FT.undergrad 1.07 6.01 4.95 1.40 1.29 0.18  
## X..PT.undergrad -0.44 13.61 14.05 3.35 14.58 0.33  
## in.state.tuition -1.58 1.63 3.22 2.85 7.97 0.10  
## out.of.state.tuition -1.50 1.82 3.32 1.33 2.13 0.10  
## room -1.88 2.48 4.36 0.82 -0.19 0.15  
## board -1.77 2.32 4.08 0.71 0.69 0.12  
## add..fees -1.01 8.06 9.07 2.97 13.02 0.22  
## estim..book.costs -2.77 1.89 4.67 -0.92 4.12 0.11  
## estim..personal.. -1.04 3.40 4.44 0.46 -0.67 0.16  
## X..fac..w.PHD -0.19 1.19 1.38 -0.66 -0.47 0.06  
## stud..fac..ratio -1.58 2.75 4.33 -0.26 -0.79 0.16  
## Graduation.rate -1.74 1.62 3.36 -0.02 -0.75 0.12  
## k4$cluster 1.00 1.00 0.00 NaN NaN 0.00  
## --------------------------------------------------------   
## group: 2  
## vars n mean sd median trimmed mad  
## College.Name\* 1 67 589.10 390.22 495.00 575.84 437.37  
## State\* 2 67 26.25 13.04 28.00 26.53 13.34  
## Public..1...Private..2.\* 3 67 1.99 0.12 2.00 2.00 0.00  
## X..appli..rec.d 4 67 0.44 0.84 0.11 0.33 0.58  
## X..appl..accepted 5 67 0.16 0.61 -0.04 0.09 0.43  
## X..new.stud..enrolled 6 67 0.00 0.55 -0.22 -0.06 0.30  
## X..new.stud..from.top.10. 7 67 1.65 1.02 1.51 1.64 1.12  
## X..new.stud..from.top.25. 8 67 1.43 0.58 1.49 1.48 0.66  
## X..FT.undergrad 9 67 -0.11 0.45 -0.31 -0.17 0.25  
## X..PT.undergrad 10 67 -0.38 0.22 -0.47 -0.44 0.05  
## in.state.tuition 11 67 1.50 0.45 1.64 1.57 0.25  
## out.of.state.tuition 12 67 1.68 0.42 1.82 1.73 0.33  
## room 13 67 1.19 0.87 1.09 1.18 0.73  
## board 14 67 0.99 0.90 0.93 0.95 0.82  
## add..fees 15 67 0.08 1.07 -0.15 -0.13 0.75  
## estim..book.costs 16 67 0.31 1.10 0.06 0.17 0.53  
## estim..personal.. 17 67 -0.49 0.55 -0.53 -0.52 0.54  
## X..fac..w.PHD 18 67 1.05 0.38 1.13 1.08 0.27  
## stud..fac..ratio 19 67 -1.12 0.65 -0.97 -1.08 0.49  
## Graduation.rate 20 67 1.12 0.54 1.24 1.15 0.49  
## k4$cluster 21 67 2.00 0.00 2.00 2.00 0.00  
## min max range skew kurtosis se  
## College.Name\* 16.00 1266.00 1250.00 0.30 -1.34 47.67  
## State\* 5.00 48.00 43.00 -0.29 -1.26 1.59  
## Public..1...Private..2.\* 1.00 2.00 1.00 -7.82 60.09 0.01  
## X..appli..rec.d -0.56 2.63 3.19 1.08 0.25 0.10  
## X..appl..accepted -0.60 2.08 2.67 1.08 0.33 0.07  
## X..new.stud..enrolled -0.70 1.84 2.54 1.30 1.50 0.07  
## X..new.stud..from.top.10. -0.16 3.68 3.84 0.22 -0.96 0.12  
## X..new.stud..from.top.25. -0.13 2.18 2.31 -0.72 -0.28 0.07  
## X..FT.undergrad -0.64 1.41 2.05 1.29 1.24 0.05  
## X..PT.undergrad -0.52 0.48 0.99 2.59 6.13 0.03  
## in.state.tuition -1.20 1.94 3.13 -3.50 17.61 0.05  
## out.of.state.tuition 0.27 2.21 1.94 -1.30 1.58 0.05  
## room -0.84 3.64 4.48 0.30 0.59 0.11  
## board -0.62 4.27 4.89 0.77 1.09 0.11  
## add..fees -0.87 4.09 4.96 2.06 4.39 0.13  
## estim..book.costs -1.22 5.80 7.02 3.07 12.24 0.13  
## estim..personal.. -1.48 0.99 2.47 0.54 -0.12 0.07  
## X..fac..w.PHD -0.01 1.61 1.62 -0.74 -0.22 0.05  
## stud..fac..ratio -2.84 0.32 3.15 -0.60 -0.12 0.08  
## Graduation.rate -0.36 1.90 2.26 -0.57 -0.49 0.07  
## k4$cluster 2.00 2.00 0.00 NaN NaN 0.00  
## --------------------------------------------------------   
## group: 3  
## vars n mean sd median trimmed mad  
## College.Name\* 1 183 578.72 362.08 539.00 566.33 412.16  
## State\* 2 183 28.57 13.18 33.00 29.08 14.83  
## Public..1...Private..2.\* 3 183 1.96 0.21 2.00 2.00 0.00  
## X..appli..rec.d 4 183 -0.37 0.35 -0.47 -0.43 0.19  
## X..appl..accepted 5 183 -0.33 0.39 -0.45 -0.40 0.24  
## X..new.stud..enrolled 6 183 -0.40 0.30 -0.48 -0.44 0.19  
## X..new.stud..from.top.10. 7 183 0.01 0.64 -0.05 -0.03 0.64  
## X..new.stud..from.top.25. 8 183 0.11 0.73 0.12 0.12 0.73  
## X..FT.undergrad 9 183 -0.40 0.25 -0.48 -0.45 0.14  
## X..PT.undergrad 10 183 -0.26 0.34 -0.38 -0.32 0.17  
## in.state.tuition 11 183 0.41 0.51 0.39 0.42 0.40  
## out.of.state.tuition 12 183 0.30 0.56 0.24 0.28 0.52  
## room 13 183 0.08 0.85 -0.04 0.01 0.73  
## board 14 183 0.33 0.81 0.18 0.28 0.70  
## add..fees 15 183 -0.19 0.74 -0.43 -0.31 0.48  
## estim..book.costs 16 183 -0.16 0.75 -0.30 -0.23 0.68  
## estim..personal.. 17 183 -0.30 0.92 -0.46 -0.41 0.54  
## X..fac..w.PHD 18 183 0.08 0.78 0.17 0.14 0.80  
## stud..fac..ratio 19 183 -0.18 0.71 -0.22 -0.24 0.53  
## Graduation.rate 20 183 0.40 0.71 0.35 0.38 0.65  
## k4$cluster 21 183 3.00 0.00 3.00 3.00 0.00  
## min max range skew kurtosis se  
## College.Name\* 4.00 1273.00 1269.00 0.25 -1.05 26.77  
## State\* 2.00 49.00 47.00 -0.31 -1.18 0.97  
## Public..1...Private..2.\* 1.00 2.00 1.00 -4.43 17.69 0.02  
## X..appli..rec.d -0.74 1.34 2.07 2.49 7.53 0.03  
## X..appl..accepted -0.77 1.58 2.35 2.33 6.97 0.03  
## X..new.stud..enrolled -0.76 1.04 1.81 2.00 5.37 0.02  
## X..new.stud..from.top.10. -1.35 2.11 3.46 0.61 0.07 0.05  
## X..new.stud..from.top.25. -1.75 1.84 3.59 -0.15 -0.30 0.05  
## X..FT.undergrad -0.70 0.81 1.51 2.08 5.14 0.02  
## X..PT.undergrad -0.52 2.32 2.83 3.22 17.73 0.02  
## in.state.tuition -1.34 1.64 2.98 -0.57 1.42 0.04  
## out.of.state.tuition -1.30 1.83 3.12 0.18 0.04 0.04  
## room -1.68 3.00 4.67 0.82 0.71 0.06  
## board -1.98 2.78 4.76 0.60 0.54 0.06  
## add..fees -1.04 2.93 3.96 1.76 3.59 0.05  
## estim..book.costs -1.83 2.77 4.60 1.25 2.79 0.06  
## estim..personal.. -1.34 8.05 9.39 4.56 36.01 0.07  
## X..fac..w.PHD -3.07 1.43 4.50 -0.76 0.78 0.06  
## stud..fac..ratio -2.40 3.55 5.95 1.12 4.41 0.05  
## Graduation.rate -1.30 2.89 4.19 0.23 0.05 0.05  
## k4$cluster 3.00 3.00 0.00 NaN NaN 0.00  
## --------------------------------------------------------   
## group: 4  
## vars n mean sd median trimmed mad  
## College.Name\* 1 175 675.63 354.99 681.00 684.33 404.75  
## State\* 2 175 27.91 12.72 28.00 28.32 16.31  
## Public..1...Private..2.\* 3 175 1.55 0.50 2.00 1.57 0.00  
## X..appli..rec.d 4 175 -0.30 0.47 -0.50 -0.39 0.25  
## X..appl..accepted 5 175 -0.30 0.46 -0.51 -0.37 0.27  
## X..new.stud..enrolled 6 175 -0.23 0.53 -0.46 -0.31 0.34  
## X..new.stud..from.top.10. 7 175 -0.68 0.43 -0.70 -0.71 0.40  
## X..new.stud..from.top.25. 8 175 -0.73 0.70 -0.77 -0.76 0.73  
## X..FT.undergrad 9 175 -0.20 0.54 -0.42 -0.29 0.30  
## X..PT.undergrad 10 175 -0.04 0.57 -0.28 -0.16 0.28  
## in.state.tuition 11 175 -0.72 0.62 -0.77 -0.74 0.85  
## out.of.state.tuition 12 175 -0.82 0.50 -0.81 -0.82 0.55  
## room 13 175 -0.53 0.74 -0.57 -0.56 0.62  
## board 14 175 -0.68 0.75 -0.70 -0.69 0.76  
## add..fees 15 175 0.04 1.02 -0.28 -0.13 0.65  
## estim..book.costs 16 175 0.00 1.20 -0.30 -0.09 0.91  
## estim..personal.. 17 175 0.25 0.96 0.17 0.19 0.92  
## X..fac..w.PHD 18 175 -0.67 0.98 -0.61 -0.63 1.07  
## stud..fac..ratio 19 175 0.46 0.94 0.39 0.42 0.95  
## Graduation.rate 20 175 -0.78 0.79 -0.75 -0.78 0.65  
## k4$cluster 21 175 4.00 0.00 4.00 4.00 0.00  
## min max range skew kurtosis se  
## College.Name\* 2.00 1267.00 1265.00 -0.18 -1.02 26.84  
## State\* 1.00 51.00 50.00 -0.19 -0.88 0.96  
## Public..1...Private..2.\* 1.00 2.00 1.00 -0.22 -1.96 0.04  
## X..appli..rec.d -0.75 1.29 2.04 1.43 1.28 0.04  
## X..appl..accepted -0.80 1.18 1.98 1.15 0.32 0.04  
## X..new.stud..enrolled -0.82 1.39 2.21 1.26 0.72 0.04  
## X..new.stud..from.top.10. -1.46 0.59 2.06 0.56 -0.02 0.03  
## X..new.stud..from.top.25. -2.30 1.30 3.59 0.40 -0.15 0.05  
## X..FT.undergrad -0.71 1.70 2.41 1.38 1.19 0.04  
## X..PT.undergrad -0.51 2.94 3.45 2.36 6.92 0.04  
## in.state.tuition -1.59 0.71 2.30 0.16 -1.50 0.05  
## out.of.state.tuition -2.21 0.64 2.85 -0.05 -0.40 0.04  
## room -2.22 2.27 4.49 0.49 0.47 0.06  
## board -2.81 1.97 4.78 0.14 0.51 0.06  
## add..fees -1.03 4.97 6.00 2.03 5.27 0.08  
## estim..book.costs -2.81 10.98 13.79 4.61 38.58 0.09  
## estim..personal.. -1.56 4.24 5.79 0.88 1.47 0.07  
## X..fac..w.PHD -3.91 1.79 5.70 -0.38 0.24 0.07  
## stud..fac..ratio -1.86 3.81 5.67 0.44 0.74 0.07  
## Graduation.rate -2.79 1.90 4.68 0.17 0.75 0.06  
## k4$cluster 4.00 4.00 0.00 NaN NaN 0.00

#Lets map the state the the cluster that most ofter occurs to see if there is something cool happening. We need to convert appreviations to names to map. This is a lookup table  
#We will join to this  
st\_crosswalk <- tibble(state = state.name) %>%  
 bind\_cols(tibble(abb = state.abb)) %>%   
 bind\_rows(tibble(state = "District of Columbia", abb = "DC"))  
  
stateClusterData <- dataWithCluster[,c(2,21)]  
names(stateClusterData)[names(stateClusterData) == 'State'] <- 'abb'  
stateClusterData <- left\_join(stateClusterData, st\_crosswalk, by = "abb")

## Warning: Column `abb` joining factor and character vector, coercing into  
## character vector

mode <- function(codes){  
 which.max(tabulate(codes))  
}  
  
stateClusterData <- stateClusterData %>%   
 group\_by(state) %>%   
 summarise(`k4$cluster` = mode(`k4$cluster`))  
  
stateClusterData$`k4$cluster` <- as.factor(stateClusterData$`k4$cluster`)  
  
#lets map it!  
plot\_usmap(data = stateClusterData, values = "k4$cluster") + scale\_fill\_manual(values=c("red", "blue", "green","yellow"), name="cluster") + theme(legend.position = "right")



Lets look at Tufts and figure out what cluster it would be in if we used only the data we had

#GEt a DF of just Tufts data  
tufts <- fullData %>%  
 filter(College.Name=="Tufts University")   
   
  
#remove NAs  
tufts <- Filter(function(x) !all(is.na(x)), tufts)

To be honest I dont know how to do this part. I dont have scaled data for Tufts because that row was dropped before I scaled for clustering because it contained an NA value. If I scale before dropping NAs then my cluster data will no longer be 0 at the Mean. If I get scaled data for Tufts after the fact it does not line up because the scaling does not match. Hopefully writing this down will show that I put thought into it and I would like to see the solution so I can understand. I made an extra map chart to prove that I want to learn :)

All other answers are written out in the quiz.