12 Final Project part4

Rama Danda

2/26/2020

Summarize the problem statement you addressed. *Identify the students that would complete the courses in 100% of allocated time and 150% time completed with given gender and race analysis factors*

```
inst_grads <- read.csv("cc_institution_grads.csv", header=TRUE)
str(inst_grads)
```

```
## 'data.frame':    1302102 obs. of  10 variables:
##  $ unitid       : int  100760 100760 100760 100760 100760 100760 100760 10
0760 100760 100760 ...
##  $ year         : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 .
..
##  $ gender       : Factor w/ 3 levels "B","F","M": 1 3 2 1 3 2 1 3 2 1 ...
##  $ race         : Factor w/ 6 levels "A","Ai","B","H",..: 6 6 6 5 5 5 3 3
3 4 ...
##  $ cohort       : Factor w/ 3 levels "2y all","4y bach",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ grad_cohort  : Factor w/ 4181 levels "0","1","10","100",..: 3162 963 17
87 2579 704 974 4052 1243 3776 1 ...
##  $ grad_100     : Factor w/ 2314 levels "0","1","10","100",..: 2015 2314 2
314 2314 2314 2314 2314 2314 2314 2314 ...
##  $ grad_150     : Factor w/ 3110 levels "0","1","10","100",..: 60 2247 271
3 2954 2073 2506 877 2484 339 1 ...
##  $ grad_100_rate: Factor w/ 991 levels "0.0","0.1","0.2",..: 86 991 991 99
1 991 991 991 991 991 991 ...
##  $ grad_150_rate: Factor w/ 1002 levels "0.0","0.1","0.2",..: 167 148 181
179 148 206 134 170 113 1002 ...
```

```
head(inst_grads)
```

```
##   unitid year gender race cohort grad_cohort grad_100 grad_150 grad_100_ra
te
## 1 100760 2011      B    X 2y all         446       73      105          16
.4
## 2 100760 2011      M    X 2y all         185     NULL       40          NU
LL
## 3 100760 2011      F    X 2y all         261     NULL       65          NU
LL
## 4 100760 2011      B    W 2y all         348     NULL       86          NU
LL
## 5 100760 2011      M    W 2y all         162     NULL       35          NU
LL
```

```
## 6 100760 2011      F    W 2y all         186     NULL       51          NU
LL
##   grad_150_rate
## 1          23.5
## 2          21.6
## 3          24.9
## 4          24.7
## 5          21.6
## 6          27.4
```

```
colSums(inst_grads=='NULL')
```

```
##        unitid          year        gender          race        cohort
##             0             0             0             0             0
##   grad_cohort      grad_100      grad_150 grad_100_rate grad_150_rate
##        412722        892033        412722        970041        607233
```

```
inst_grads[inst_grads=='NULL'] <- NA

inst_grads_clean <- data.frame(na.omit((inst_grads)))
colSums(inst_grads_clean=='NULL')
```

```
##        unitid          year        gender          race        cohort
##             0             0             0             0             0
##   grad_cohort      grad_100      grad_150 grad_100_rate grad_150_rate
##             0             0             0             0             0
```

```
sum(is.na(inst_grads_clean))
```

```
## [1] 0
```

```
str(inst_grads_clean)
```

```
## 'data.frame':    332061 obs. of  10 variables:
##  $ unitid       : int  100760 100760 100760 101028 101028 101028 101143 10
1143 101143 101161 ...
##  $ year         : int  2011 2012 2013 2011 2012 2013 2011 2012 2013 2011 .
..
##  $ gender       : Factor w/ 3 levels "B","F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ race         : Factor w/ 6 levels "A","Ai","B","H",..: 6 6 6 6 6 6 6 6
6 6 ...
##  $ cohort       : Factor w/ 3 levels "2y all","4y bach",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ grad_cohort  : Factor w/ 4181 levels "0","1","10","100",..: 3162 3589 3
589 1787 1991 1941 3218 3492 3407 4027 ...
##  $ grad_100     : Factor w/ 2314 levels "0","1","10","100",..: 2015 1630 1
710 1238 1645 944 287 37 1672 200 ...
##  $ grad_150     : Factor w/ 3110 levels "0","1","10","100",..: 60 2965 255
9 2295 2272 2140 817 251 2735 328 ...
##  $ grad_100_rate: Factor w/ 991 levels "0.0","0.1","0.2",..: 86 569 679 89
8 68 674 205 108 785 58 ...
```

```
##  $ grad_150_rate: Factor w/ 1002 levels "0.0","0.1","0.2",..: 167 68 893 8
3 68 56 319 153 51 71 ...
```

```
head(inst_grads_clean)
```

```
##    unitid year gender race cohort grad_cohort grad_100 grad_150 grad_100_r
ate
## 1  100760 2011      B    X 2y all         446       73      105          1
6.4
## 19 100760 2012      B    X 2y all         594       40       87
6.7
## 37 100760 2013      B    X 2y all         594       46       54
7.7
## 55 101028 2011      B    X 2y all         261       25       42
9.6
## 73 101028 2012      B    X 2y all         281       41       41          1
4.6
## 91 101028 2013      B    X 2y all         276       20       37
7.2
##    grad_150_rate
## 1           23.5
## 19          14.6
## 37           9.1
## 55          16.1
## 73          14.6
## 91          13.4
```

```
## convert factors dataset of grad_100 to numerical
inst_grads_clean$grad100 <- as.numeric(levels(inst_grads_clean$grad_100))[ins
t_grads_clean$grad_100]
```

```
## Warning: NAs introduced by coercion
```

```
## convert factors dataset of grad_150 to numerical
inst_grads_clean$grad150 <- as.numeric(levels(inst_grads_clean$grad_150))[ins
t_grads_clean$grad_150]
```

```
## Warning: NAs introduced by coercion
```

```
#summarize the data frame to get grad_150 group by year
inst_grads_clean_sum <- inst_grads_clean %>%
  group_by(year,gender,race,unitid) %>%
  summarise(grad_150= sum(grad150),grad_100=sum(grad100))
```

```
head(inst_grads_clean_sum)
```

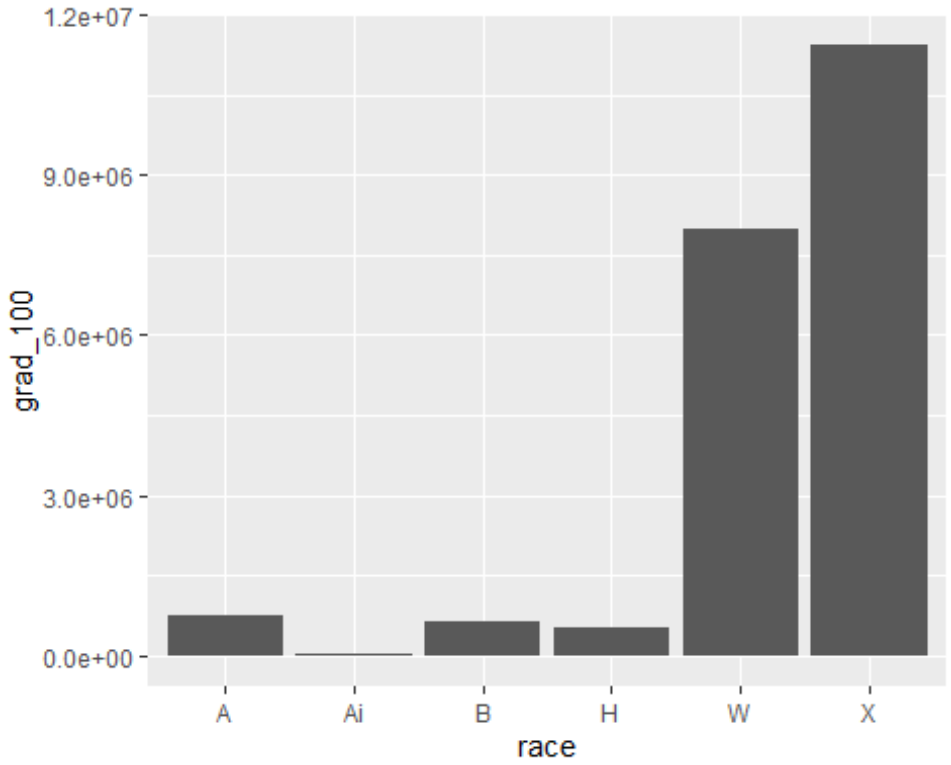
```
## # A tibble: 6 x 6
## # Groups:   year, gender, race [1]
##    year gender race  unitid grad_150 grad_100
##   <int> <fct>  <fct>  <int>    <dbl>    <dbl>
## 1  2002 B      A     100654        1        1
## 2  2002 B      A     100663       24       17
```

```
## 3  2002 B      A     100706        8        2
## 4  2002 B      A     100724        0        0
## 5  2002 B      A     100751       11        3
## 6  2002 B      A     100830        3        1

#str(inst_grads_clean_sum)
```

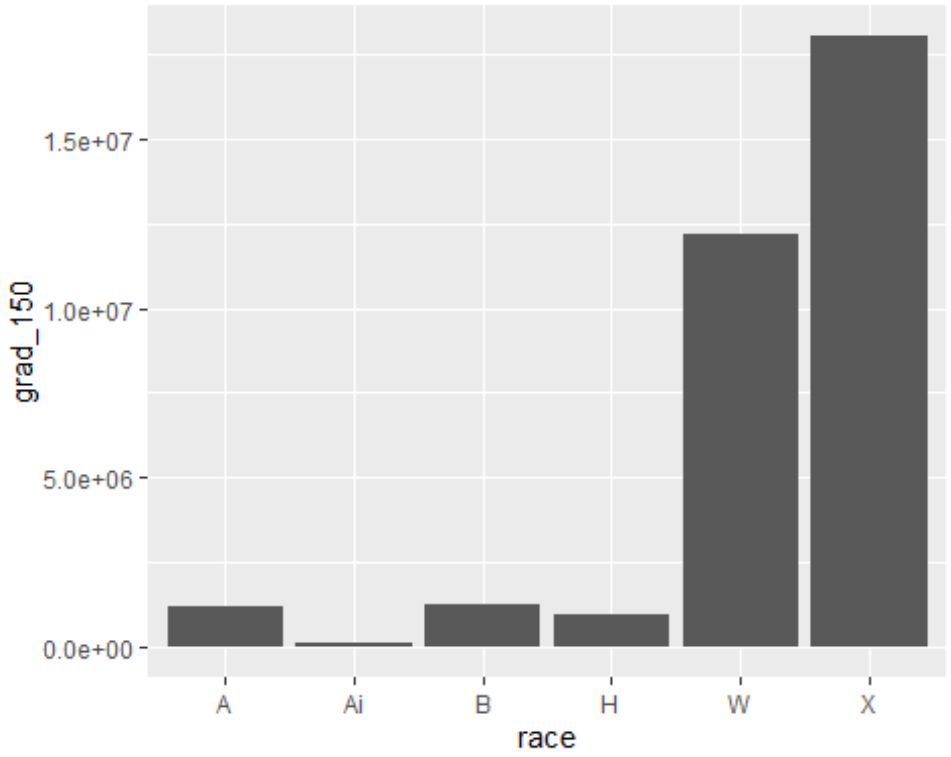
Summarize how you addressed this problem statement

Uncovered the trend of particular race has high percentage of 100 completion vs other genders in 150% range

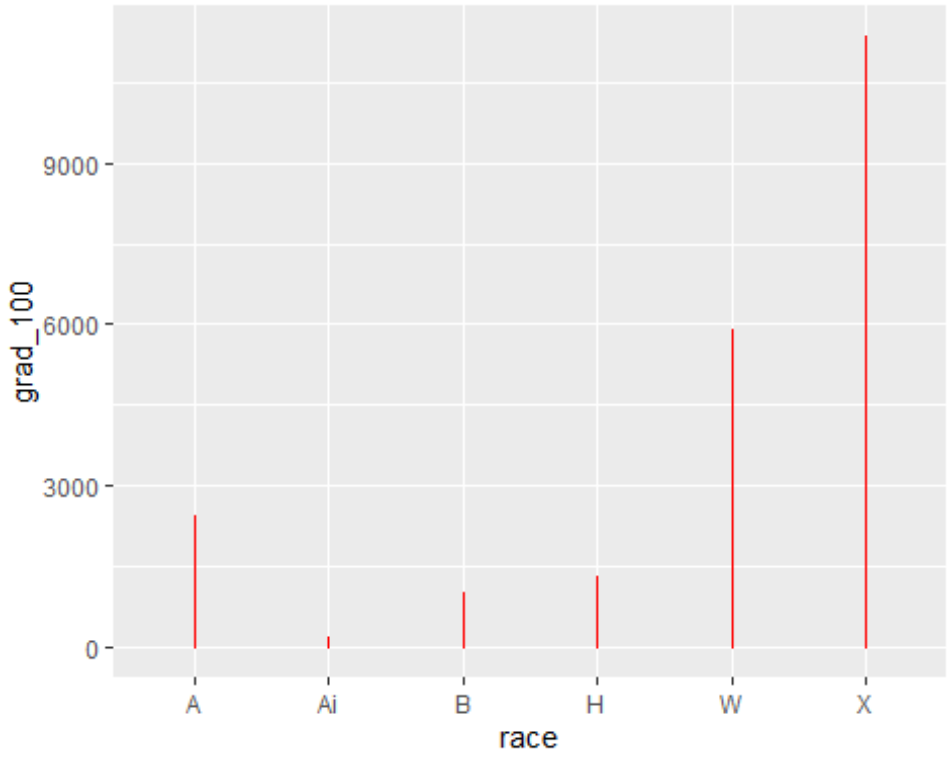
```
 ggplot(inst_grads_clean_sum,aes(x=race,y=grad_100),col(year,as.factor = TRUE
))+
    geom_bar(stat = "identity")
```




```
    ggplot(inst_grads_clean_sum,aes(x=race,y=grad_150),col(year,as.factor = T
RUE))+
    geom_bar(stat = "identity")
```



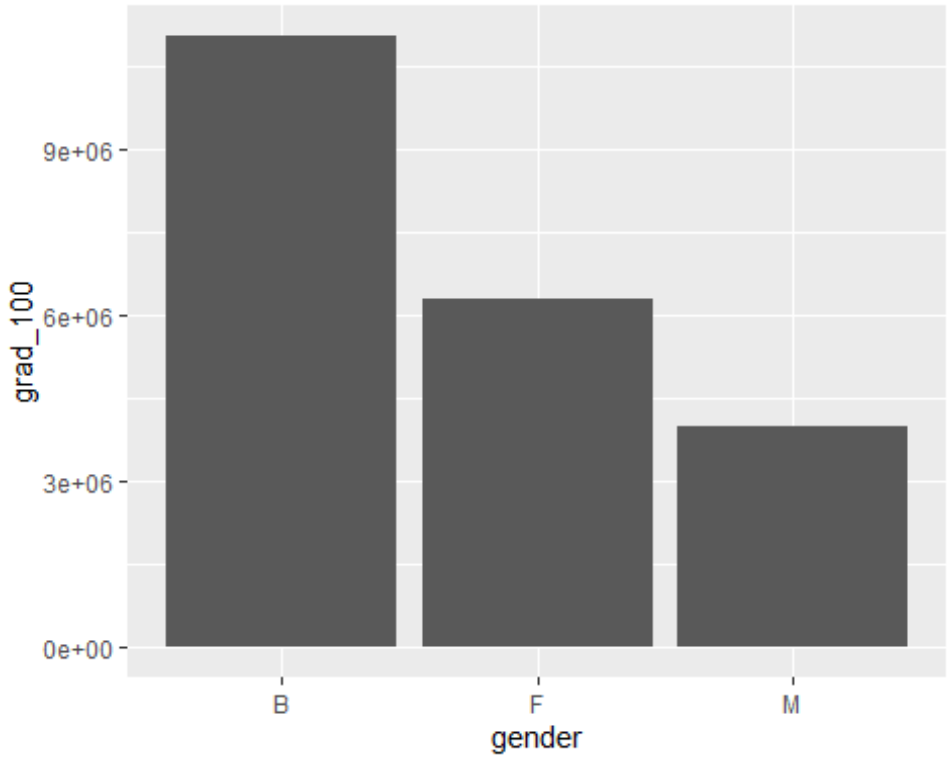

```
      p= ggplot()+
  geom_line(data=inst_grads_clean_sum, aes(x=race,y=grad_100),color = "blue")
+
    geom_line(data=inst_grads_clean_sum,aes(x=race,y=grad_150),color = "red")
  print(p)
```



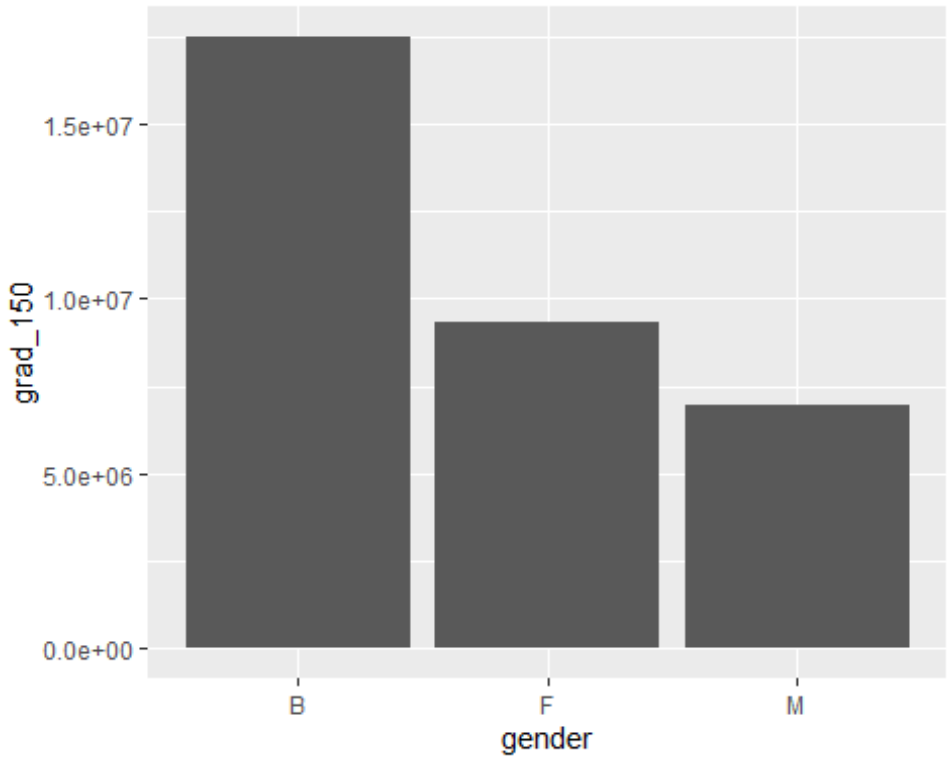
Summarize the interesting insights that your analysis provided.

since the data is in very clustered manner and has to do lot of un earthing of measures combining/summarizing the data and plotting them

```
inst_grads_clean_sum <- inst_grads_clean %>%
  group_by(year,gender,race,unitid) %>%
  summarise(grad_150= sum(grad150),grad_100=sum(grad100))

 ggplot(inst_grads_clean_sum,aes(x=gender,y=grad_100),col(year,as.factor = TR
UE))+
    geom_bar(stat = "identity")
```




```
    ggplot(inst_grads_clean_sum,aes(x=gender,y=grad_150),col(year,as.factor =
TRUE))+
    geom_bar(stat = "identity")
```



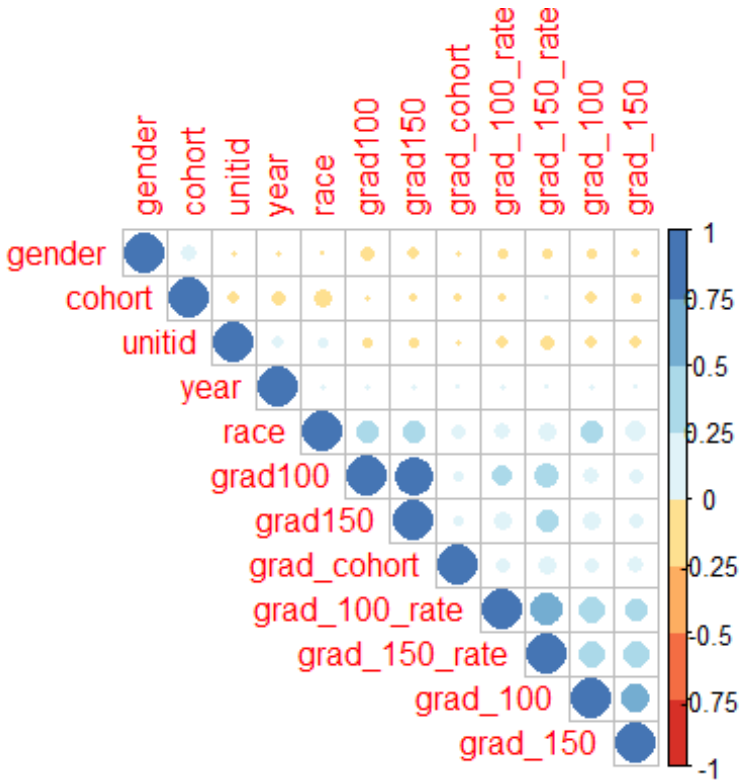
Summarize the implications to the consumer (target audience) of your analysis *Data can only be used for clustering them in to multiple clusters and can identify if a test object can be placed in one of the clusters*

```
#converting factors to numaric

inst_grads_clean$grad_100  <- as.numeric((inst_grads_clean$grad_100  ))
inst_grads_clean$grad_150 <- as.numeric((inst_grads_clean$grad_150 ))
inst_grads_clean$grad_100_rate <- as.numeric((inst_grads_clean$grad_100_rate
))
inst_grads_clean$grad_150_rate <- as.numeric((inst_grads_clean$grad_150_rate
))

inst_grads_clean$gender <- as.numeric((inst_grads_clean$gender ))

inst_grads_clean$race <- as.numeric((inst_grads_clean$race ))

inst_grads_clean$cohort <- as.numeric((inst_grads_clean$cohort ))

inst_grads_clean$grad_cohort <- as.numeric((inst_grads_clean$grad_cohort ))

inst_grads_clean <- data.frame(na.omit((inst_grads_clean)))

sum(is.na(inst_grads_clean))

## [1] 0

head(inst_grads_clean)

##    unitid year gender race cohort grad_cohort grad_100 grad_150 grad_100_r
ate
## 1  100760 2011      1    6      1        3162     2015       60
86
## 19 100760 2012      1    6      1        3589     1630     2965
569
## 37 100760 2013      1    6      1        3589     1710     2559
679
## 55 101028 2011      1    6      1        1787     1238     2295
898
## 73 101028 2012      1    6      1        1991     1645     2272
68
## 91 101028 2013      1    6      1        1941      944     2140
674
##    grad_150_rate grad100 grad150
## 1            167      73     105
## 19            68      40      87
## 37           893      46      54
## 55            83      25      42
## 73            68      41      41
## 91            56      20      37
```

```r
#str(inst_grads_clean)

M <-cor(inst_grads_clean)
M
```

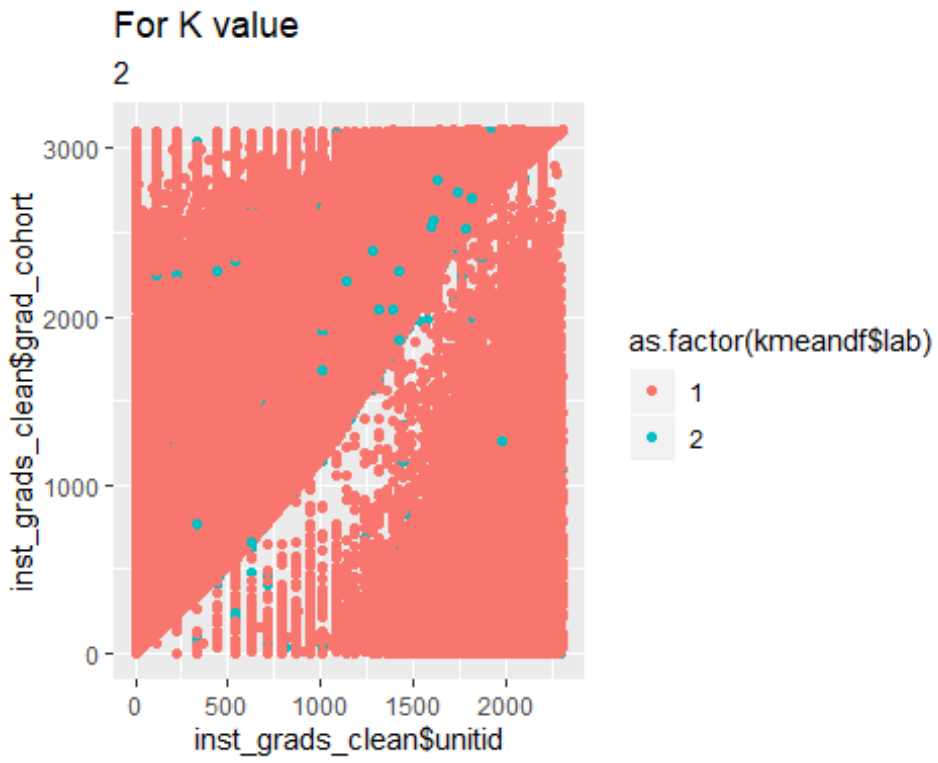
```
##                    unitid         year       gender         race      coho
rt
## unitid         1.00000000  0.105427443 -0.018997796  0.052468992 -0.079466
31
## year           0.10542744  1.000000000 -0.022022934  0.026051942 -0.138531
77
## gender        -0.01899780 -0.022022934  1.000000000 -0.009470576  0.172445
16
## race           0.05246899  0.026051942 -0.009470576  1.000000000 -0.192341
00
## cohort        -0.07946631 -0.138531767  0.172445164 -0.192340996  1.000000
00
## grad_cohort   -0.02368939  0.007744420 -0.026979661  0.119062592 -0.039336
73
## grad_100      -0.10437394  0.017161207 -0.071804877  0.282735072 -0.082856
22
## grad_150      -0.07798409  0.009198045 -0.045722441  0.235376625 -0.055545
59
## grad_100_rate -0.09858721  0.018174723 -0.053233628  0.160824675 -0.040325
87
## grad_150_rate -0.10892559  0.013184901 -0.057083933  0.194868587  0.012886
66
## grad100       -0.05571070  0.023729385 -0.108544035  0.285752084 -0.027845
58
## grad150       -0.06127394  0.016654706 -0.106939471  0.296650389 -0.034045
33
##               grad_cohort    grad_100     grad_150 grad_100_rate grad_150_
rate
## unitid        -0.02368939 -0.10437394 -0.077984088   -0.09858721   -0.1089
2559
## year           0.00774442  0.01716121  0.009198045    0.01817472    0.0131
8490
## gender        -0.02697966 -0.07180488 -0.045722441   -0.05323363   -0.0570
8393
## race           0.11906259  0.28273507  0.235376625    0.16082468    0.1948
6859
## cohort        -0.03933673 -0.08285622 -0.055545588   -0.04032587    0.0128
8666
## grad_cohort    1.00000000  0.12774584  0.173282252    0.13291643    0.2039
3681
## grad_100       0.12774584  1.00000000  0.532928306    0.40597498    0.4369
6648
## grad_150       0.17328225  0.53292831  1.000000000    0.31907669    0.3975
3529
## grad_100_rate  0.13291643  0.40597498  0.319076688    1.00000000    0.6541
```

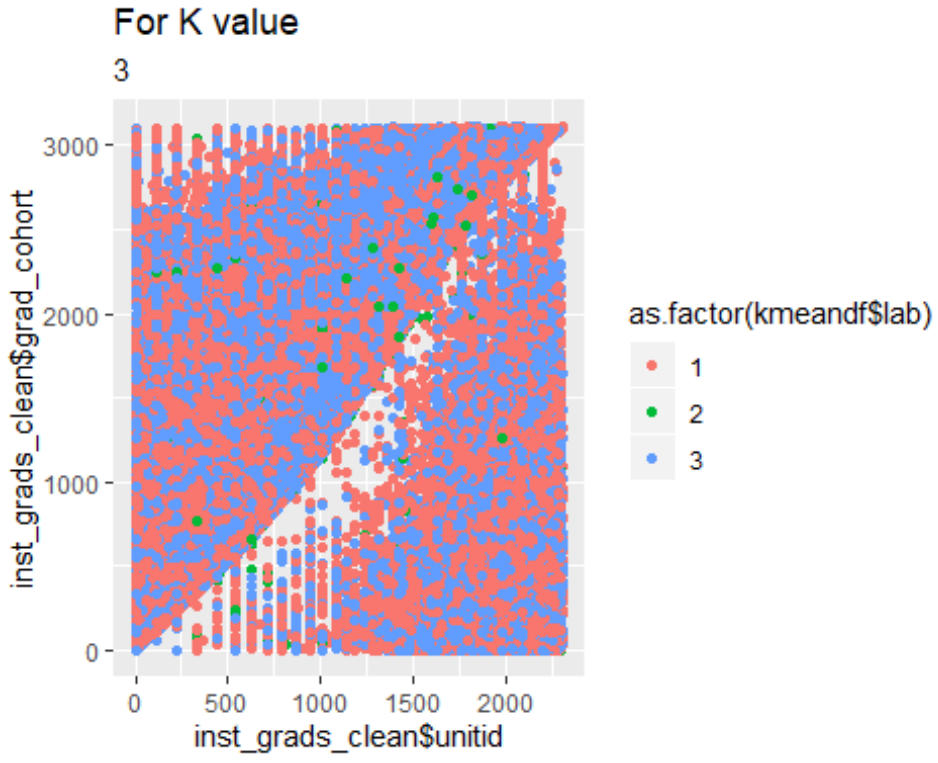
```
3127
## grad_150_rate  0.20393681  0.43696648  0.397535289    0.65413127    1.0000
0000
## grad100        0.05905466  0.18184336  0.125433715    0.25517435    0.3416
2165
## grad150        0.05564294  0.18744603  0.133111858    0.18533299    0.3040
5356
##                   grad100     grad150
## unitid         -0.05571070 -0.06127394
## year            0.02372938  0.01665471
## gender         -0.10854403 -0.10693947
## race            0.28575208  0.29665039
## cohort         -0.02784558 -0.03404533
## grad_cohort     0.05905466  0.05564294
## grad_100        0.18184336  0.18744603
## grad_150        0.12543371  0.13311186
## grad_100_rate   0.25517435  0.18533299
## grad_150_rate   0.34162165  0.30405356
## grad100         1.00000000  0.95750095
## grad150         0.95750095  1.00000000

corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```

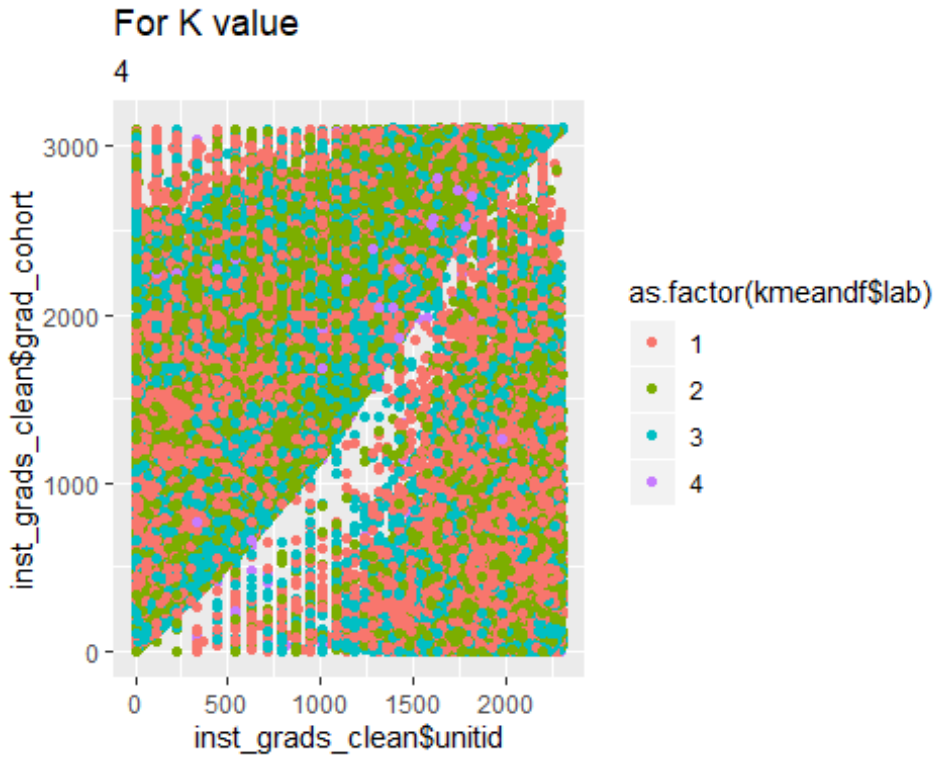



```
kmeandf<-data.frame(inst_grads_clean$unitid,inst_grads_clean$grad_cohort,
                    inst_grads_clean$grad_100,
```

```
                    inst_grads_clean$grad_150)
head(kmeandf)

##   inst_grads_clean.unitid inst_grads_clean.grad_cohort
## 1                  100760                         3162
## 2                  100760                         3589
## 3                  100760                         3589
## 4                  101028                         1787
## 5                  101028                         1991
## 6                  101028                         1941
##   inst_grads_clean.grad_100 inst_grads_clean.grad_150
## 1                      2015                        60
## 2                      1630                      2965
## 3                      1710                      2559
## 4                      1238                      2295
## 5                      1645                      2272
## 6                       944                      2140

#set.seed(20)

#cluserting from k value as 2:12
looop <- 2:4
for (i in looop){
  clusters <- kmeans(kmeandf,i)
  kmeandf$lab <- as.factor(clusters$cluster)
  myplot <- ggplot(data=kmeandf, aes(x=inst_grads_clean$unitid,y=inst_grads_c
lean$grad_cohort))+
    geom_point(aes(x =inst_grads_clean$grad_100, y = inst_grads_clean$grad_15
0, colour = as.factor(kmeandf$lab)),
               data = kmeandf)+
    ggtitle("For K value",i)
  plot(myplot)
}
```


For K value
2
inst_grads_clean$grad_cohort
3000
2000
1000
0
0
500
1000
1500
2000
inst_grads_clean$unitid
as.factor(kmeandf$lab)
1
2


For K value
3
inst_grads_clean$grad_cohort
3000
2000
1000
0
0
500
1000
1500
2000
inst_grads_clean$unitid
as.factor(kmeandf$lab)
1
2
3



Discuss the limitations of your analysis and how you, or someone else, could improve or build on it. *The limitation of the analysis is that a general completion of college with a specific cohorts completing in 100 or 150% is only available and does not explain what has happened to the rest of the population. Data set is limiting and has to be tied to other large datasets of insitutions. Which can be done as an extension*