mbruner3_4

Mark Bruner

10/16/2020

```
library(tidyverse)
```

```
## -- 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

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

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(factoextra)
```

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

```
library(moments)
set.seed(15)
```

Part 1: Preparing & Getting to Know Our Data

```
univ <- read_csv("/Users/markbruner/Google Drive/MSBA/Machine Learning/mbruner3/ML_mbruner3/Assignment
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   ‘College Name‘ = col_character(),
##   State = col_character()
## )
## i Use ‘spec()‘ for the full column specifications.
```

```
str(univ)
```

```
## tibble [1,302 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ College Name            : chr [1:1302] "Alaska Pacific University" "University of Alaska at Fairba
##  $ State                   : chr [1:1302] "AK" "AK" "AK" "AK" ...
##  $ Public (1)/ Private (2) : num [1:1302] 2 1 1 1 1 2 1 1 1 2 ...
##  $ # appli. rec'd          : num [1:1302] 193 1852 146 2065 2817 ...
##  $ # appl. accepted        : num [1:1302] 146 1427 117 1598 1920 ...
##  $ # new stud. enrolled    : num [1:1302] 55 928 89 1162 984 ...
##  $ % new stud. from top 10%: num [1:1302] 16 NA 4 NA NA NA 18 NA 25 67 ...
##  $ % new stud. from top 25%: num [1:1302] 44 NA 24 NA NA 27 78 NA 57 88 ...
##  $ # FT undergrad          : num [1:1302] 249 3885 492 6209 3958 ...
##  $ # PT undergrad          : num [1:1302] 869 4519 1849 10537 305 ...
##  $ in-state tuition        : num [1:1302] 7560 1742 1742 1742 1700 ...
##  $ out-of-state tuition    : num [1:1302] 7560 5226 5226 5226 3400 ...
##  $ room                    : num [1:1302] 1620 1800 2514 2600 1108 ...
##  $ board                   : num [1:1302] 2500 1790 2250 2520 1442 ...
##  $ add. fees               : num [1:1302] 130 155 34 114 155 300 124 84 NA 120 ...
##  $ estim. book costs       : num [1:1302] 800 650 500 580 500 350 300 500 600 400 ...
##  $ estim. personal $       : num [1:1302] 1500 2304 1162 1260 850 ...
##  $ % fac. w/PHD            : num [1:1302] 76 67 39 48 53 52 72 48 85 74 ...
##  $ stud./fac. ratio        : num [1:1302] 11.9 10 9.5 13.7 14.3 32.8 18.9 18.7 16.7 14 ...
##  $ Graduation rate         : num [1:1302] 15 NA 39 NA 40 55 51 15 69 72 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ‘College Name‘ = col_character(),
##   ..   State = col_character(),
##   ..   ‘Public (1)/ Private (2)‘ = col_double(),
##   ..   ‘# appli. rec'd‘ = col_double(),
##   ..   ‘# appl. accepted‘ = col_double(),
##   ..   ‘# new stud. enrolled‘ = col_double(),
##   ..   ‘% new stud. from top 10%‘ = col_double(),
##   ..   ‘% new stud. from top 25%‘ = col_double(),
##   ..   ‘# FT undergrad‘ = col_double(),
##   ..   ‘# PT undergrad‘ = col_double(),
##   ..   ‘in-state tuition‘ = col_double(),
##   ..   ‘out-of-state tuition‘ = col_double(),
##   ..   room = col_double(),
##   ..   board = col_double(),
##   ..   ‘add. fees‘ = col_double(),
##   ..   ‘estim. book costs‘ = col_double(),
##   ..   ‘estim. personal $‘ = col_double(),
##   ..   ‘% fac. w/PHD‘ = col_double(),
##   ..   ‘stud./fac. ratio‘ = col_double(),
##   ..   ‘Graduation rate‘ = col_double()
##   .. )
```

```
head(univ)
```

```
## # A tibble: 6 x 20
##   ‘College Name‘ State ‘Public (1)/ Pr~ ‘# appli. rec’d‘ ‘# appl. accept~
##   <chr>          <chr>            <dbl>           <dbl>           <dbl>
## 1 Alaska Pacifi~ AK                   2             193             146
## 2 University of~ AK                   1            1852            1427
## 3 University of~ AK                   1             146             117
## 4 University of~ AK                   1            2065            1598
## 5 Alabama Agri.~ AL                   1            2817            1920
## 6 Faulkner Univ~ AL                   2             345             320
## # ... with 15 more variables: ‘# new stud. enrolled‘ <dbl>, ‘% new stud. from
## #   top 10%‘ <dbl>, ‘% new stud. from top 25%‘ <dbl>, ‘# FT undergrad‘ <dbl>,
## #   ‘# PT undergrad‘ <dbl>, ‘in-state tuition‘ <dbl>, ‘out-of-state
## #   tuition‘ <dbl>, room <dbl>, board <dbl>, ‘add. fees‘ <dbl>, ‘estim. book
## #   costs‘ <dbl>, ‘estim. personal $‘ <dbl>, ‘% fac. w/PHD‘ <dbl>, ‘stud./fac.
## #   ratio‘ <dbl>, ‘Graduation rate‘ <dbl>
```

```
tail(univ)
```

```
## # A tibble: 6 x 20
##   ‘College Name‘ State ‘Public (1)/ Pr~ ‘# appli. rec’d‘ ‘# appl. accept~
##   <chr>          <chr>            <dbl>           <dbl>           <dbl>
## 1 West Virginia~ WV                   1            1594            1572
## 2 West Virginia~ WV                   1            1869              NA
## 3 West Virginia~ WV                   1            9630            7801
## 4 West Virginia~ WV                   2            1566            1400
## 5 Wheeling Jesu~ WV                   2             903             755
## 6 University of~ WY                   1            2029            1516
## # ... with 15 more variables: ‘# new stud. enrolled‘ <dbl>, ‘% new stud. from
## #   top 10%‘ <dbl>, ‘% new stud. from top 25%‘ <dbl>, ‘# FT undergrad‘ <dbl>,
## #   ‘# PT undergrad‘ <dbl>, ‘in-state tuition‘ <dbl>, ‘out-of-state
## #   tuition‘ <dbl>, room <dbl>, board <dbl>, ‘add. fees‘ <dbl>, ‘estim. book
## #   costs‘ <dbl>, ‘estim. personal $‘ <dbl>, ‘% fac. w/PHD‘ <dbl>, ‘stud./fac.
## #   ratio‘ <dbl>, ‘Graduation rate‘ <dbl>
```

Looking at the data, some initial observations are that there is a wide spread for applications received, applied, and new students. I would think there is a independent/dependent relationship between the applications received and applications accepted/new students enrolled. There is also wide spread for in-state, out-of-state tuition, and PHD. These are areas to look at closer which I will do later after we clean up the dataset.

First thing I want to do is rename the column names to make them easier to use. I will also separate the missing data from the complete cases just in case.

```r
univ %>%
  rename(college_name  = ‘College Name‘, # renaming columns.
         state = State,
         public1_private2 =‘Public (1)/ Private (2)‘,
         appli_recd = "# appli. rec’d",
         appli_accepted = ‘# appl. accepted‘,
         new_stud = "# new stud. enrolled",
         new_stud_10 = "% new stud. from top 10%",
         new_stud_25 = "% new stud. from top 25%",
         ft_undergrad = "# FT undergrad",
         pt_undergrad = "# PT undergrad",
         in_state = "in-state tuition",
         out_state = ‘out-of-state tuition‘,
         add_fees = ‘add. fees‘,
         book_costs = ‘estim. book costs‘,
         personal_costs = ‘estim. personal $‘,
         perc_PHD = ‘% fac. w/PHD‘,
         stud_fac_ratio = ‘stud./fac. ratio‘,
         grad_rate = ‘Graduation rate‘
  ) -> univ

univ_missing <- univ[!complete.cases(univ), ]
univ_complete <- univ[complete.cases(univ), ]

colMeans(is.na(univ_complete)) # shows that we have successfully removed NA’s from the dataset.
```

```
##     college_name            state public1_private2       appli_recd
##                0                0                0                0
##   appli_accepted         new_stud      new_stud_10      new_stud_25
##                0                0                0                0
##     ft_undergrad     pt_undergrad         in_state        out_state
##                0                0                0                0
##             room            board         add_fees       book_costs
##                0                0                0                0
##   personal_costs         perc_PHD   stud_fac_ratio        grad_rate
##                0                0                0                0
```

It looks like columns 4:18, 20 are integer values but are labeled as double. I want to make sure that they are actually integers.

```r
all(univ_complete[, c(4:18, 20)] == round(univ_complete[, c(4:18, 20)]))
```

```
## [1] TRUE
```

All values are integers in columns 4 to 18 and 20 as the logic returned value TRUE meaning that none of the values in the columns have decimal places therefore they are integers.

Creating Complete Cases DF

Separated rows with NA’s from rows with no NA’s.

```
univ_complete[, c(4:18, 20)] <- sapply(univ_complete[, c(4:18, 20)], as.integer) # changed column types

univ_complete$public1_private2 <- as.factor(univ_complete$public1_private2) # Also, need to make public

str(univ_complete)
```

```
## tibble [471 x 20] (S3: tbl_df/tbl/data.frame)
##  $ college_name    : chr [1:471] "Alaska Pacific University" "University of Alaska Southeast" "Birmi
##  $ state           : chr [1:471] "AK" "AK" "AL" "AL" ...
##  $ public1_private2: Factor w/ 2 levels "1","2": 2 1 2 2 2 1 2 2 2 2 ...
##  $ appli_recd      : int [1:471] 193 146 805 608 4414 1797 708 823 605 1721 ...
##  $ appli_accepted  : int [1:471] 146 117 588 520 1500 1260 334 721 405 1068 ...
##  $ new_stud        : int [1:471] 55 89 287 127 335 938 166 274 284 806 ...
##  $ new_stud_10     : int [1:471] 16 4 67 26 30 24 46 52 24 35 ...
##  $ new_stud_25     : int [1:471] 44 24 88 47 60 35 74 87 53 75 ...
##  $ ft_undergrad    : int [1:471] 249 492 1376 538 908 6960 530 954 961 3128 ...
##  $ pt_undergrad    : int [1:471] 869 1849 207 126 119 4698 182 6 99 213 ...
##  $ in_state        : int [1:471] 7560 1742 11660 8080 5666 2220 8644 8800 6398 5504 ...
##  $ out_state       : int [1:471] 7560 5226 11660 8080 5666 4440 8644 8800 6398 5504 ...
##  $ room            : int [1:471] 1620 2514 2050 1380 1424 1935 2382 1935 1450 1650 ...
##  $ board           : int [1:471] 2500 2250 2430 2540 1540 3240 1540 1260 2222 1878 ...
##  $ add_fees        : int [1:471] 130 34 120 100 418 291 120 325 148 1016 ...
##  $ book_costs      : int [1:471] 800 500 400 500 1000 750 500 500 400 700 ...
##  $ personal_costs  : int [1:471] 1500 1162 900 1100 1400 2200 800 1200 1350 910 ...
##  $ perc_PHD        : int [1:471] 76 39 74 63 56 96 79 82 68 71 ...
##  $ stud_fac_ratio  : num [1:471] 11.9 9.5 14 11.4 15.5 6.7 12.6 13.1 13.3 17.7 ...
##  $ grad_rate       : int [1:471] 15 39 72 44 46 33 54 63 75 73 ...
```

Separating Continuous & Categorical Variables

```
univ_continuous <- as.data.frame(univ_complete[, c(4:20)])
```

Exploratory Data Analysis

UNIVARIATE EXPLORATION Summary Statistics

```
summary(univ_complete)
```

```
##  college_name          state           public1_private2   appli_recd
##  Length:471         Length:471         1:128            Min.   :   77
##  Class :character   Class :character   2:343            1st Qu.:  802
##  Mode  :character   Mode  :character                    Median : 1646
##                                                         Mean   : 3147
##                                                         3rd Qu.: 3862
##                                                         Max.   :48094
##  appli_accepted       new_stud       new_stud_10     new_stud_25
##  Min.   :   61.0   Min.   :  27.0   Min.   : 1.00   Min.   :  9.00
##  1st Qu.:  635.5   1st Qu.: 264.0   1st Qu.:15.00   1st Qu.: 40.00
##  Median : 1227.0   Median : 443.0   Median :23.00   Median : 54.00
##  Mean   : 2063.0   Mean   : 780.7   Mean   :28.01   Mean   : 55.65
##  3rd Qu.: 2456.0   3rd Qu.: 896.5   3rd Qu.:36.00   3rd Qu.: 69.00
##  Max.   :26330.0   Max.   :6392.0   Max.   :96.00   Max.   :100.00
##   ft_undergrad    pt_undergrad        in_state       out_state
##  Min.   :  249   Min.   :    1.0   Min.   :  608   Min.   : 1044
##  1st Qu.: 1018   1st Qu.:   81.5   1st Qu.: 3650   1st Qu.: 7290
##  Median : 1715   Median :  299.0   Median : 9858   Median :10100
##  Mean   : 3563   Mean   :  797.5   Mean   : 9407   Mean   :10575
##  3rd Qu.: 4056   3rd Qu.:  869.0   3rd Qu.:13246   3rd Qu.:13286
##  Max.   :31643   Max.   :21836.0   Max.   :20100   Max.   :20100
##       room          board          add_fees         book_costs     personal_costs
##  Min.   : 640   Min.   : 531   Min.   :  10.0   Min.   :  90.0   Min.   : 250
##  1st Qu.:1740   1st Qu.:1750   1st Qu.: 137.5   1st Qu.: 500.0   1st Qu.: 850
##  Median :2090   Median :2082   Median : 280.0   Median : 500.0   Median :1200
##  Mean   :2221   Mean   :2122   Mean   : 379.0   Mean   : 548.8   Mean   :1312
##  3rd Qu.:2663   3rd Qu.:2420   3rd Qu.: 486.0   3rd Qu.: 600.0   3rd Qu.:1600
##  Max.   :4816   Max.   :4541   Max.   :3247.0   Max.   :2340.0   Max.   :6800
##     perc_PHD      stud_fac_ratio    grad_rate
##  Min.   :  8.00   Min.   : 2.90   Min.   : 15.00
##  1st Qu.: 63.00   1st Qu.:11.30   1st Qu.: 53.00
##  Median : 76.00   Median :13.40   Median : 66.00
##  Mean   : 73.21   Mean   :13.96   Mean   : 65.56
##  3rd Qu.: 87.00   3rd Qu.:16.45   3rd Qu.: 79.00
##  Max.   :103.00   Max.   :28.80   Max.   :118.00
```

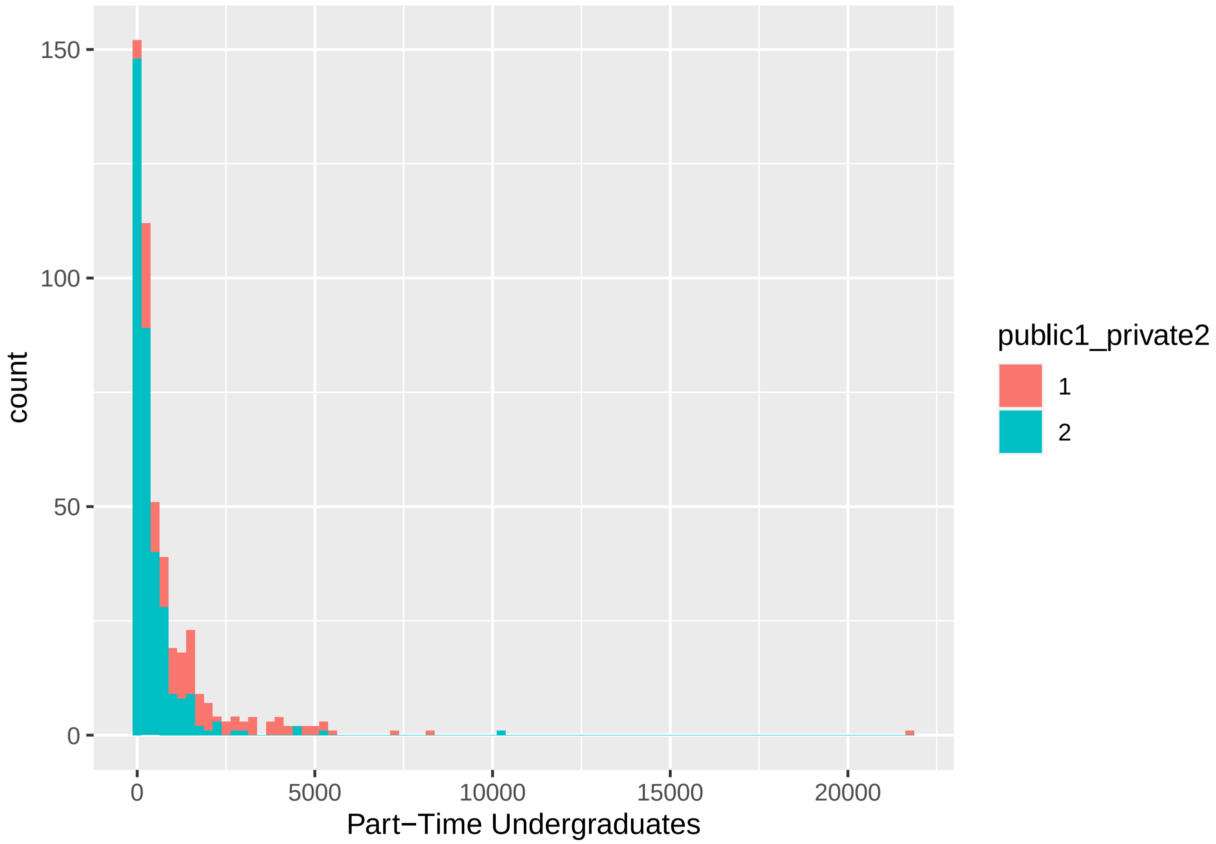
The range is large for applications received, applications accepted, and number of new students. It looks like much of the variables in this set skews positive as the means are larger than the medians. **What is the skewness of this data?**

```r
skewness(univ_complete[, 4:20])
```

```
##     appli_recd appli_accepted       new_stud    new_stud_10    new_stud_25
##     4.13469362     3.64507189     2.77962759     1.31925254     0.24053109
##   ft_undergrad   pt_undergrad       in_state      out_state           room
##     2.82133450     6.89077330     0.08709405     0.44285575     0.68618731
##          board       add_fees     book_costs personal_costs       perc_PHD
##     0.43480660     2.61701729     3.83857078     2.01557652    -0.76988349
## stud_fac_ratio      grad_rate
##     0.44264039    -0.12419865
```

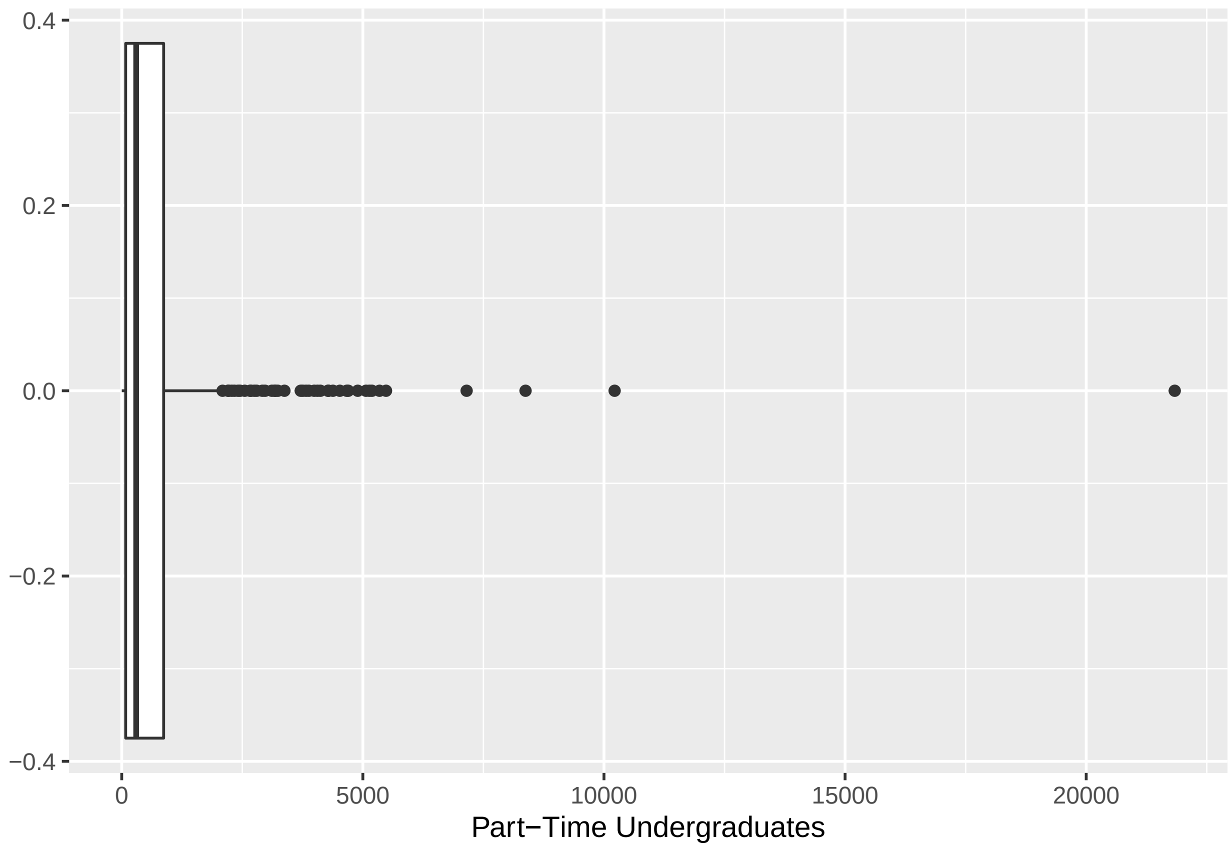
Most of the variables skew positive. Part-time undergrad is highly skewed positive, why? I also thought in-state tuition/out-of-state tuition would skew more positive. Going to look into those more through visualizing those variables as well.

```r
ggplot(data = univ_complete) +
  geom_histogram(mapping = aes(x = pt_undergrad, fill = public1_private2), binwidth = 250) +
  xlab("Part-Time Undergraduates")
```



It appears to skew more positive due to most of the data is between 0 and 2500, then there are some outliers that have more than 6,000 PT undergrads.

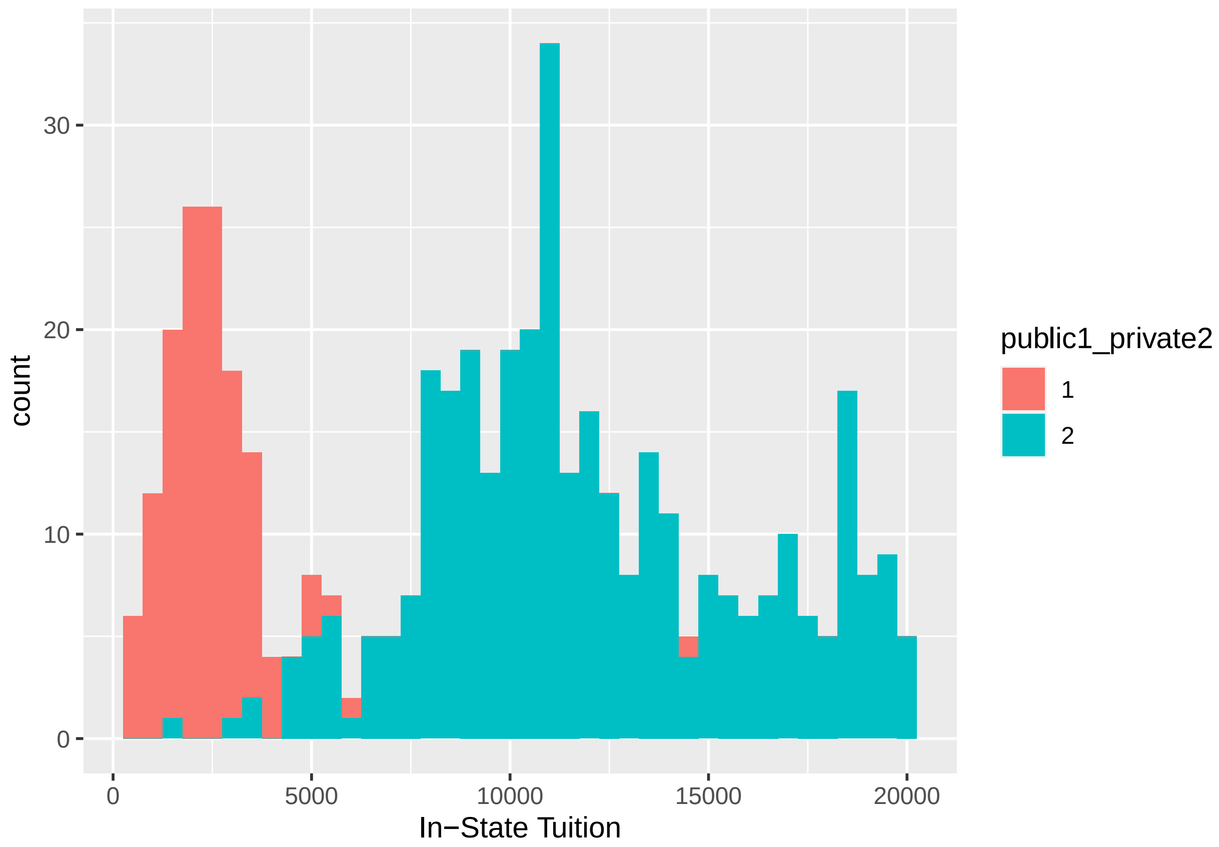
```
ggplot(data = univ_complete) +
  geom_boxplot(mapping = aes(x = pt_undergrad)) +
  xlab("Part-Time Undergraduates") # to better show the outliers.
```



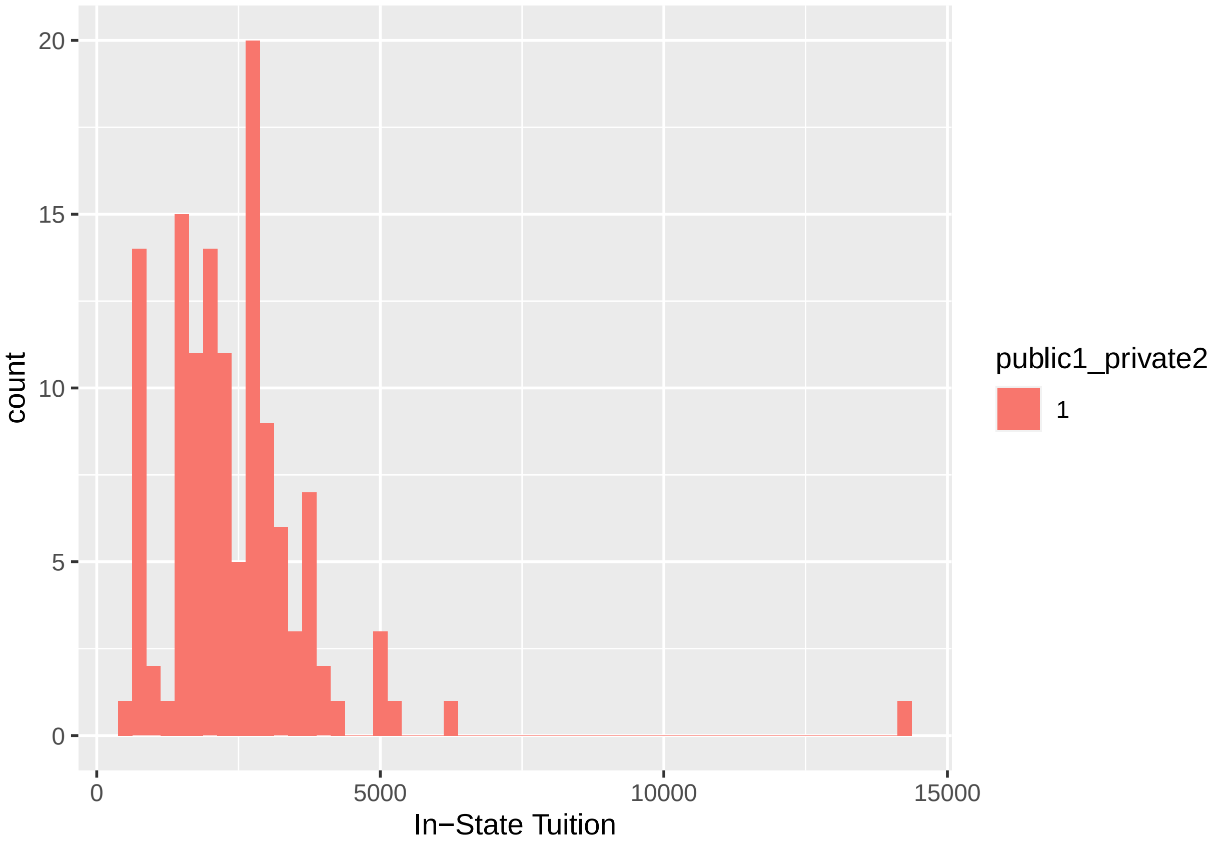
```
univ_complete %>%
  filter(pt_undergrad > 6000) # the four schools who have greater than 6,000 PT undergrads.
```

```
## # A tibble: 4 x 20
##   college_name state public1_private2 appli_recd appli_accepted new_stud
##   <chr>        <chr> <fct>                 <int>          <int>    <int>
## 1 University ~ FL    1                      6986           2959     1918
## 2 Northeaster~ MA    2                     11901           8492     2517
## 3 University ~ MN    1                     11054           6397     3524
## 4 University ~ UT    1                      5095           4491     2400
## # ... with 14 more variables: new_stud_10 <int>, new_stud_25 <int>,
## #   ft_undergrad <int>, pt_undergrad <int>, in_state <int>, out_state <int>,
## #   room <int>, board <int>, add_fees <int>, book_costs <int>,
## #   personal_costs <int>, perc_PHD <int>, stud_fac_ratio <dbl>, grad_rate <int>
```

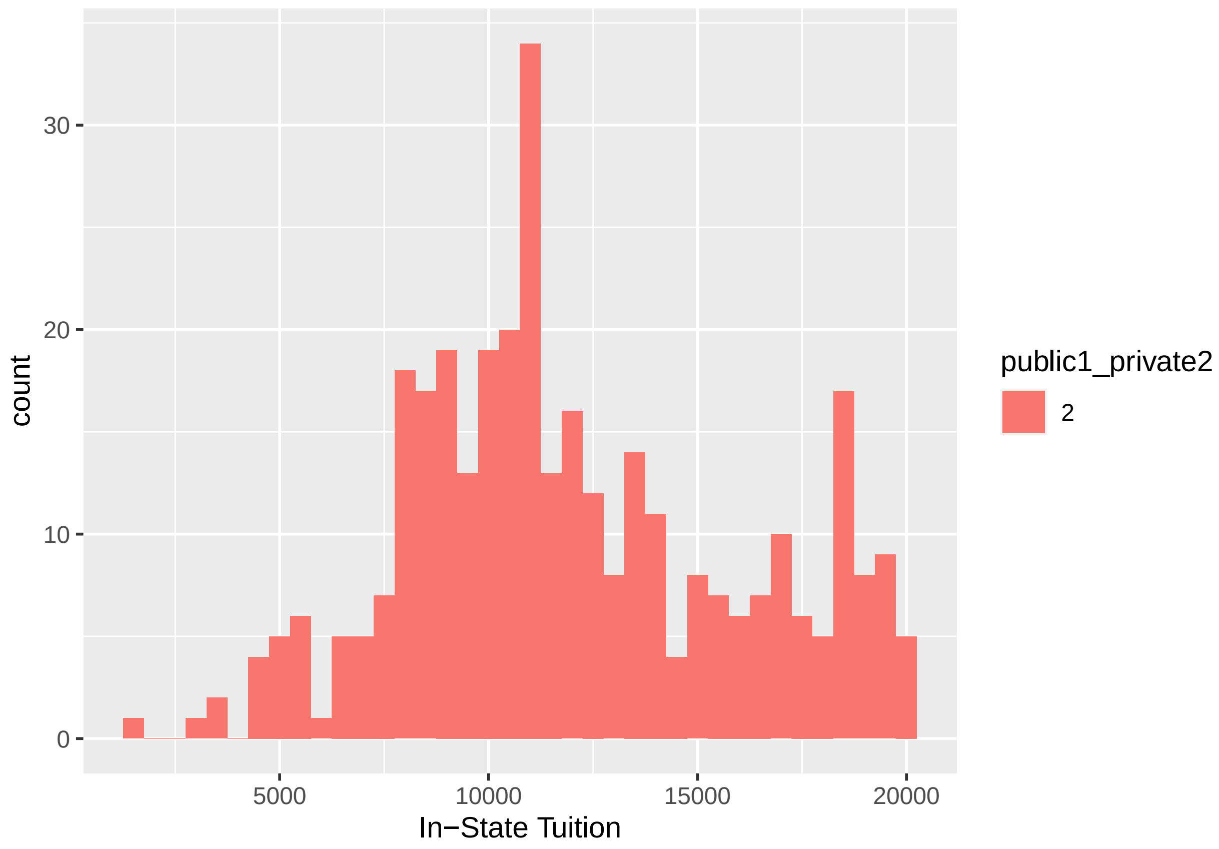
```
ggplot(data = univ_complete) +
  geom_histogram(mapping = aes(x = in_state, fill = public1_private2), binwidth = 500) +
  xlab("In-State Tuition") # going to separate the public and private schools to better see the data.
```



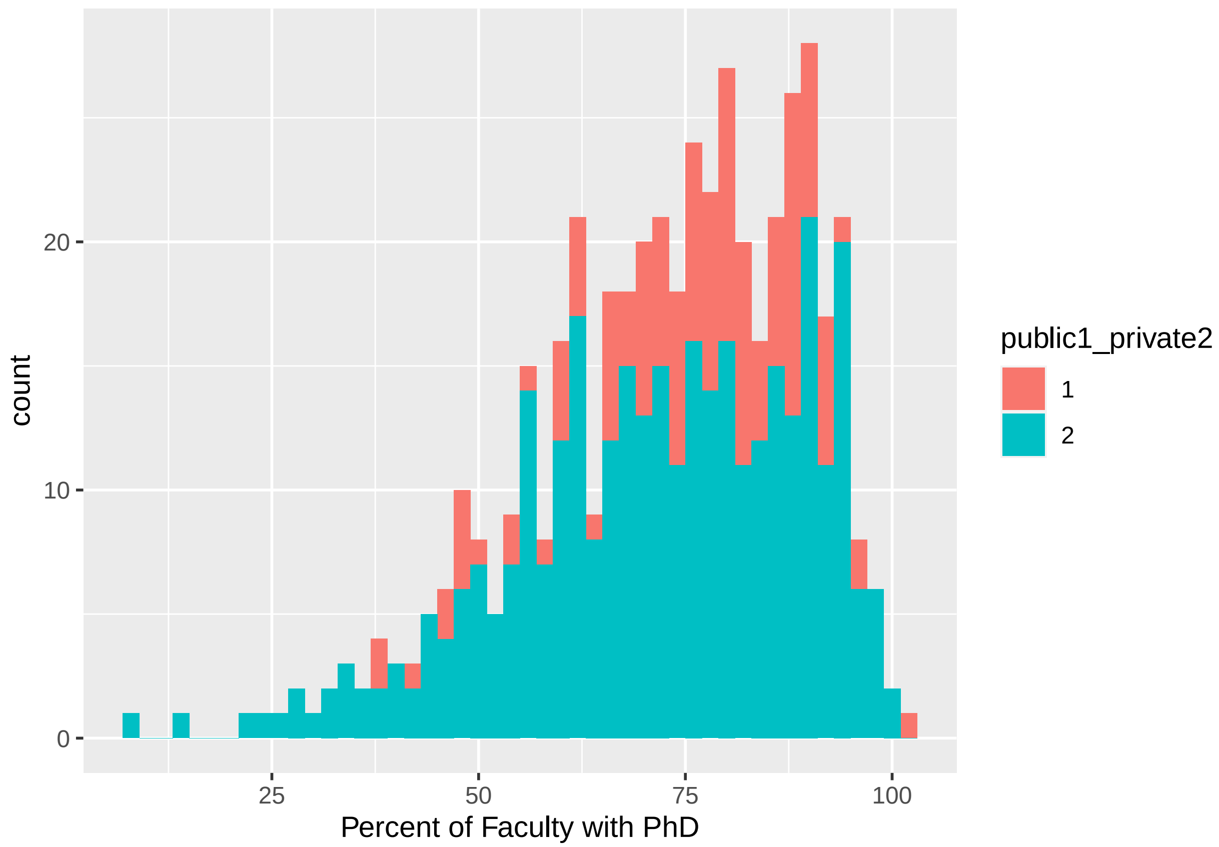
```
univ_complete %>%
  filter(public1_private2 == 1) %>%
  ggplot() +
      geom_histogram(mapping = aes(x = in_state, fill = public1_private2), binwidth = 250) +
  xlab("In-State Tuition") # skews more positive.
```



```
univ_complete %>%
  filter(public1_private2 == 2) %>%
  ggplot() +
      geom_histogram(mapping = aes(x = in_state, fill = public1_private2), binwidth = 500) +
  xlab("In-State Tuition") # more normally distributed.
```



```
ggplot(data = univ_complete) +
  geom_histogram(mapping = aes(x = perc_PHD, fill = public1_private2), binwidth = 2) +
  xlab("Percent of Faculty with PhD")
```



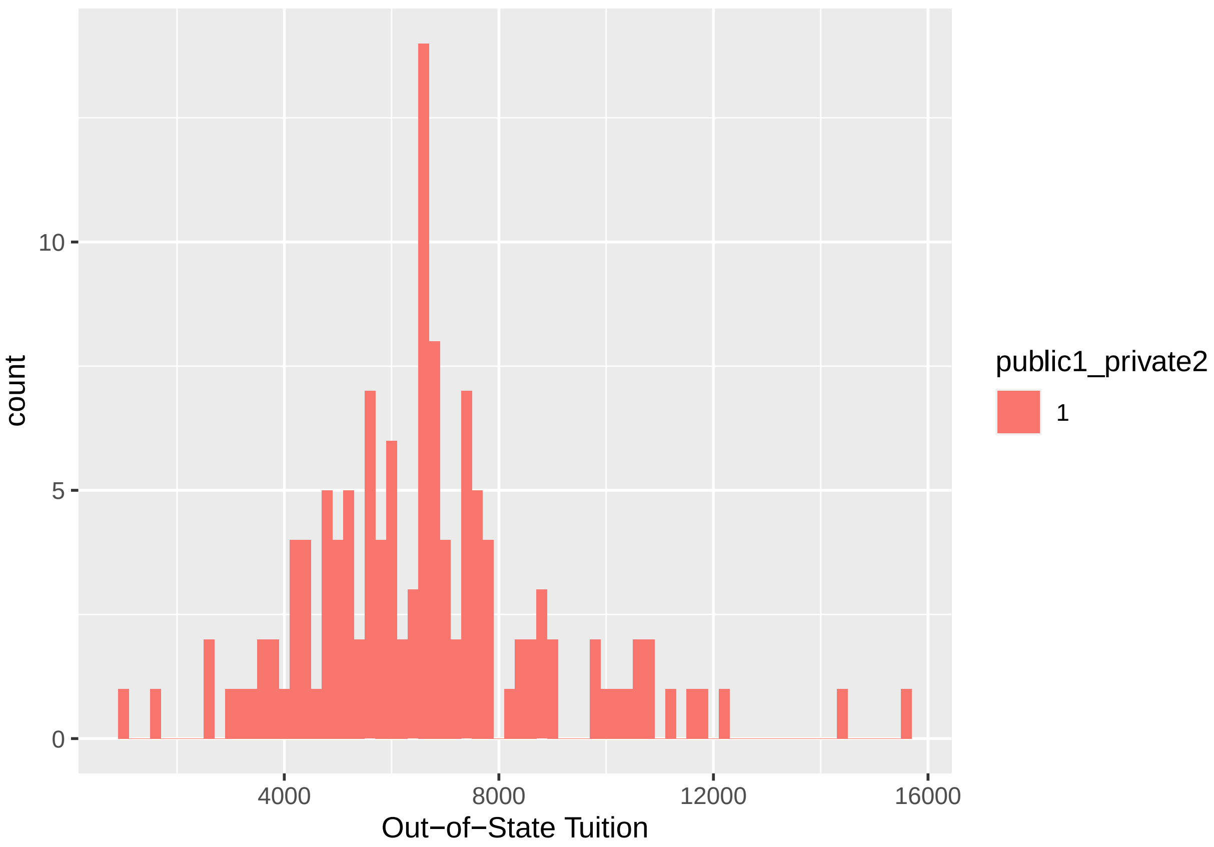
Similar skew for both public and private. Probably because some universities that are more research focused have the higher portion of PHD faculty where more liberal arts focused/religious universities you would expect to not have as high percent of PHD faculty. An odd occurrence appeared where there may exist universities who have greater than 100% PHD faculty…seems strange.

```
univ_complete %>%
  filter(perc_PHD > 100) # the school with 103% PHD faculty.
```

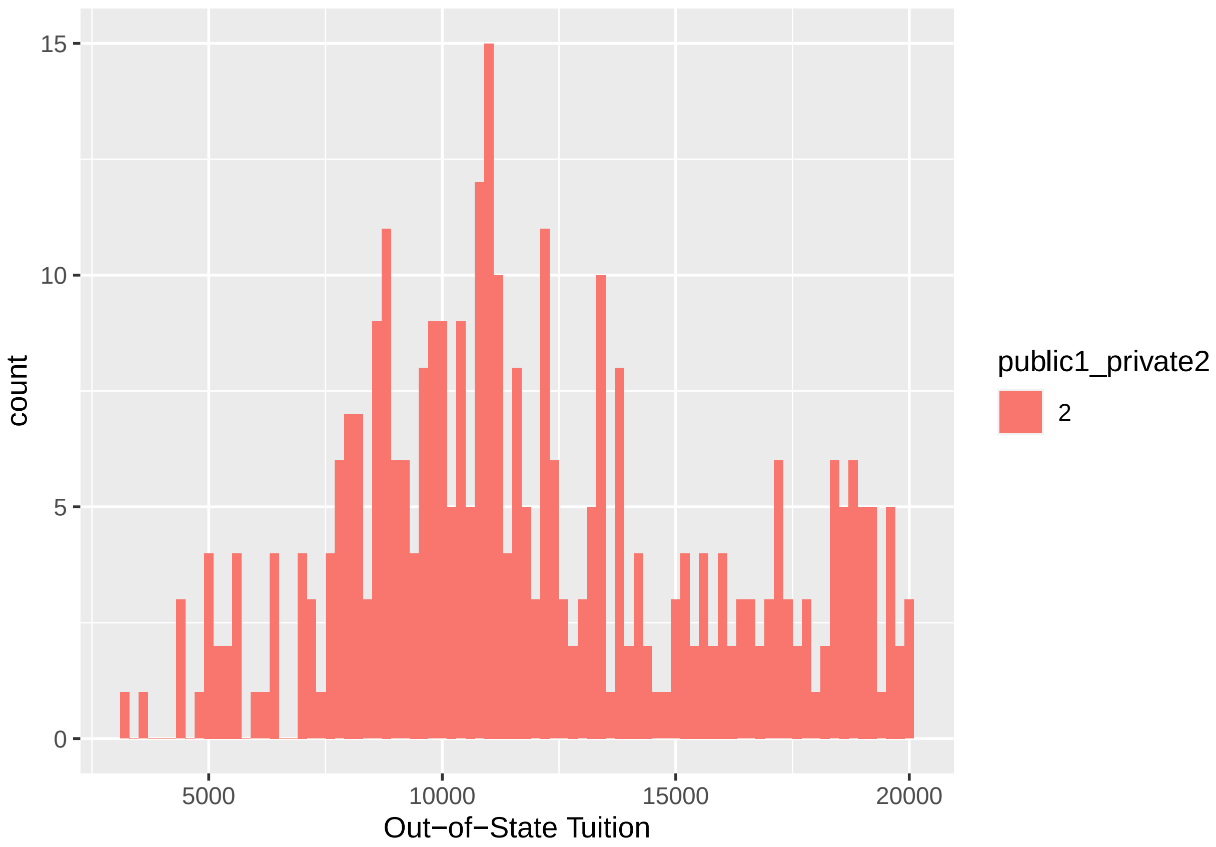
```
## # A tibble: 1 x 20
##   college_name state public1_private2 appli_recd appli_accepted new_stud
##   <chr>        <chr> <fct>                 <int>          <int>    <int>
## 1 Texas A&M U~ TX    1                       529            481      243
## # ... with 14 more variables: new_stud_10 <int>, new_stud_25 <int>,
## #   ft_undergrad <int>, pt_undergrad <int>, in_state <int>, out_state <int>,
## #   room <int>, board <int>, add_fees <int>, book_costs <int>,
## #   personal_costs <int>, perc_PHD <int>, stud_fac_ratio <dbl>, grad_rate <int>
```

There must be a reporting error here. I am not going to remove it since this single mistake shouldn't impact the overall PhD variable significantly and the other information from this university will probably be more helpful than that single mistake.

```
univ_complete %>%
  filter(public1_private2 == 1) %>%
  ggplot() +
      geom_histogram(mapping = aes(x = out_state, fill = public1_private2), binwidth = 200) +
  xlab("Out-of-State Tuition")
```



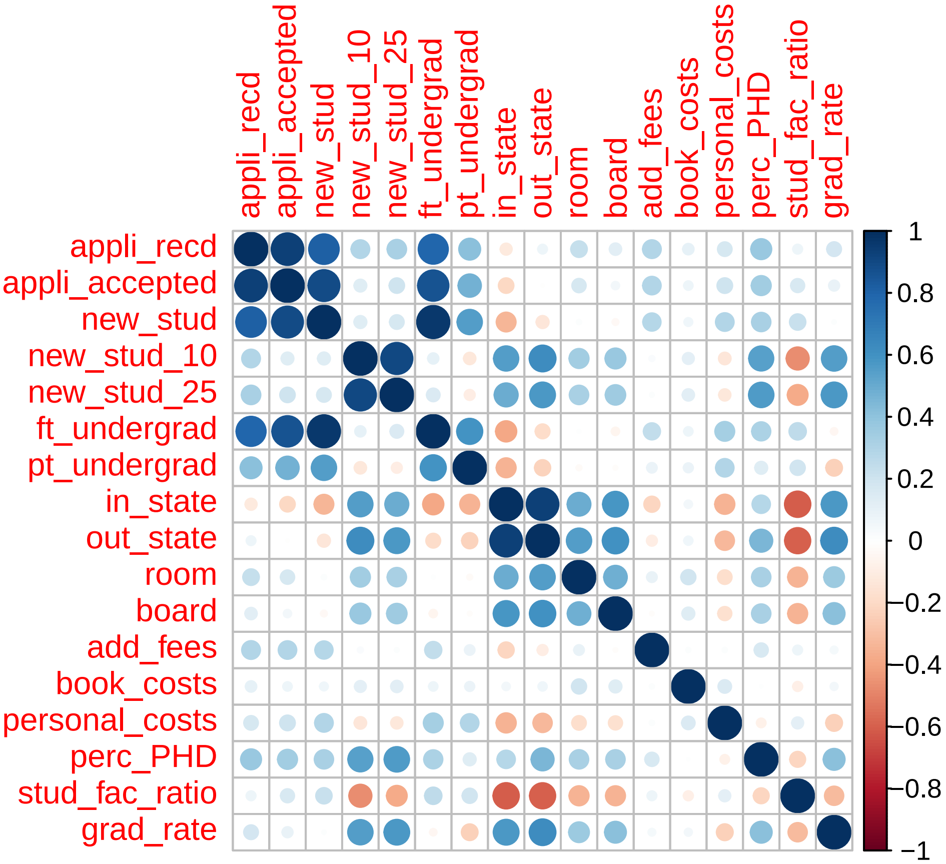
```
univ_complete %>%
  filter(public1_private2 == 2) %>%
  ggplot() +
      geom_histogram(mapping = aes(x = out_state, fill = public1_private2), binwidth = 200) +
  xlab("Out-of-State Tuition")
```



For out-of-state tuition, both public and private universities have a more normal distribution.

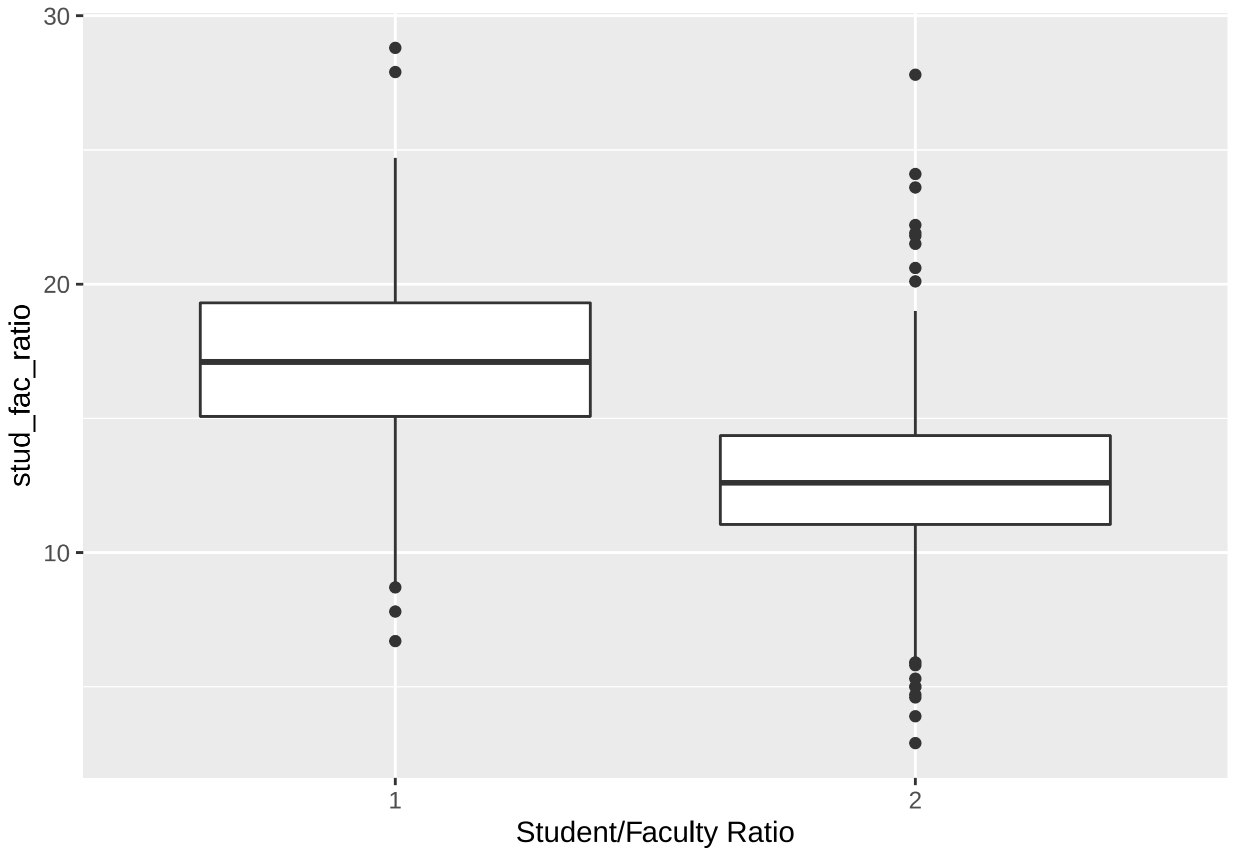
Bivariate Variable Analysis

```
m <- cor(univ_continuous)
corrplot(m, method = "circle")
```



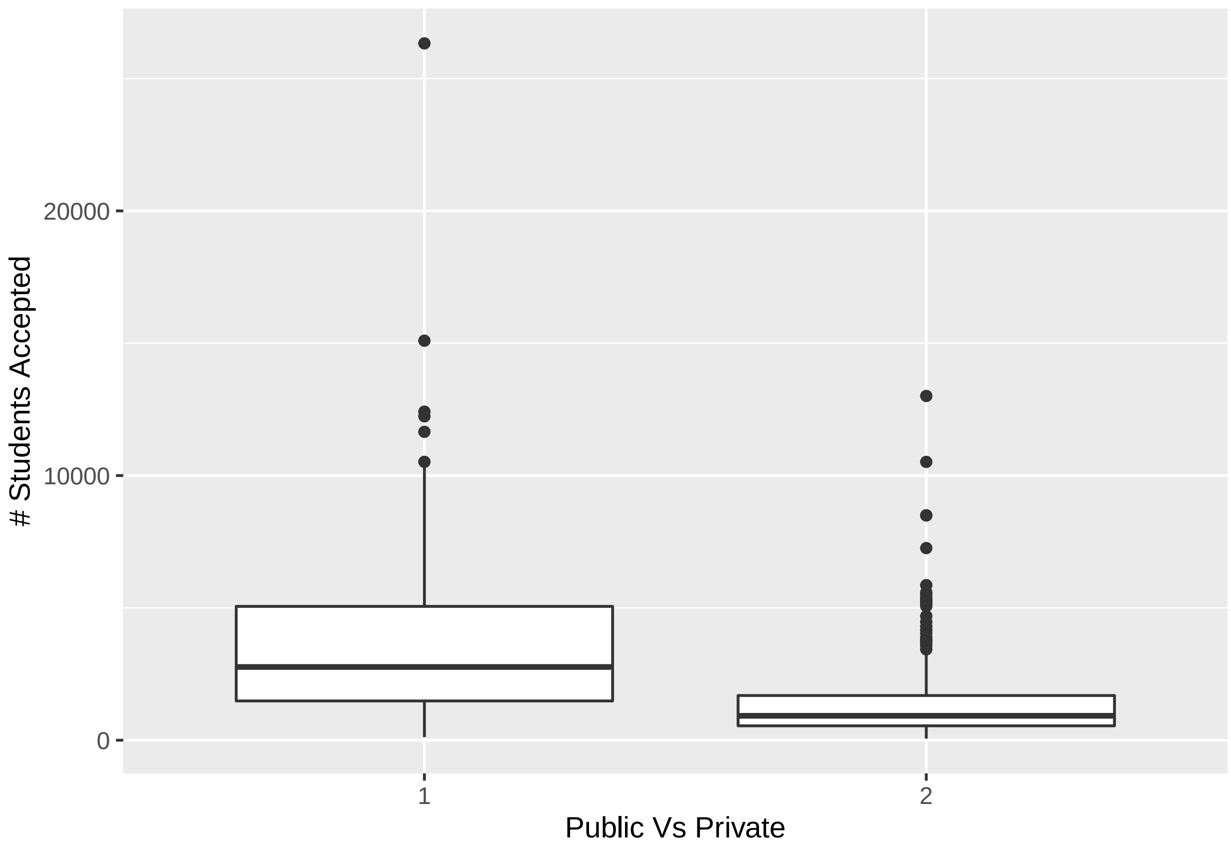
There are stronger correlations between: applications received, accepted, new students, Full-time undergrad; in-state tuition, out-of-state tuition, new students from both top 10 and 25%, room, board, student/faculty ratio, graduation rate; student/faculty ration and percent of faculty with PHD.

```
ggplot(data = univ_complete) +
  geom_boxplot(mapping = aes(x = public1_private2, y = stud_fac_ratio))+
  xlab("Student/Faculty Ratio")
```

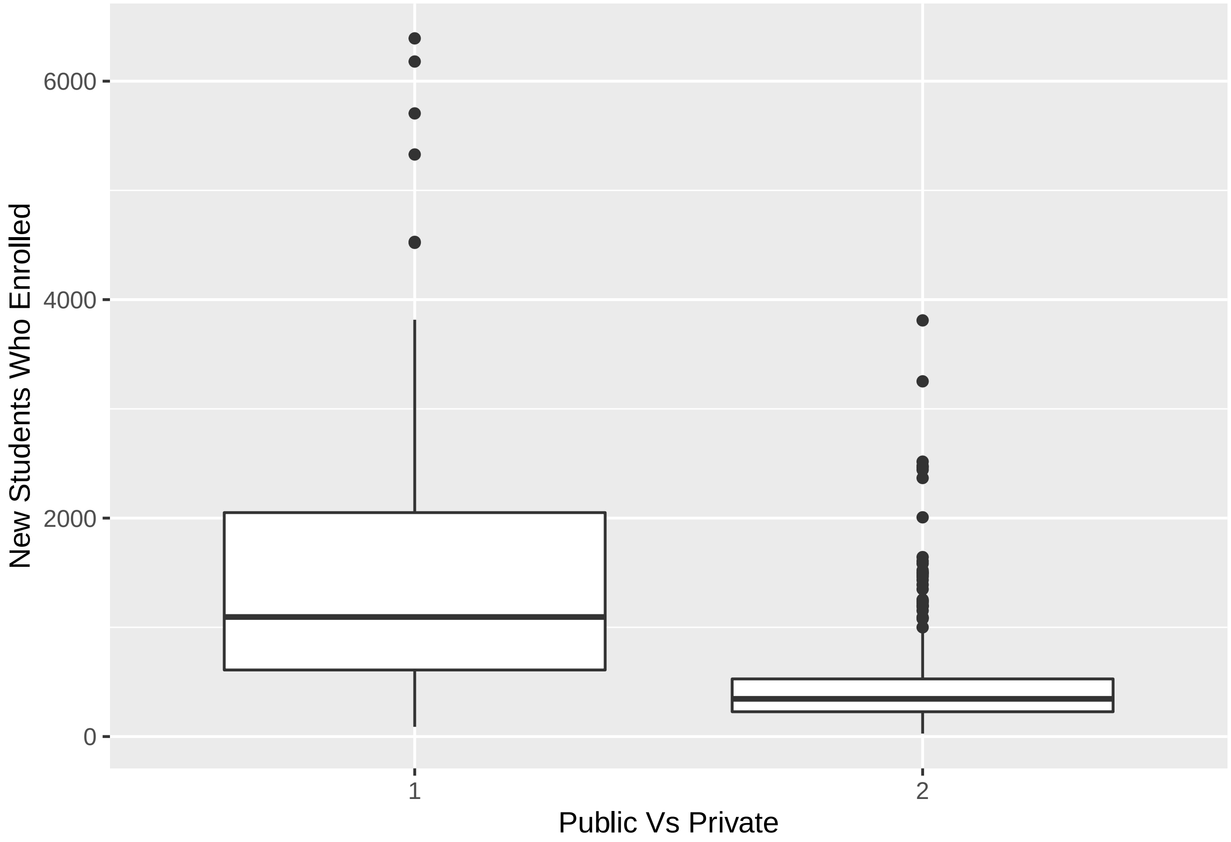


You would expect private schools to have a lower student/faculty ratio so nothing unusual shown above.

```
ggplot(data = univ_complete) +
  geom_boxplot(mapping = aes(x = public1_private2, y = appli_accepted)) +
  xlab("Public Vs Private") +
  ylab("# Students Accepted")
```



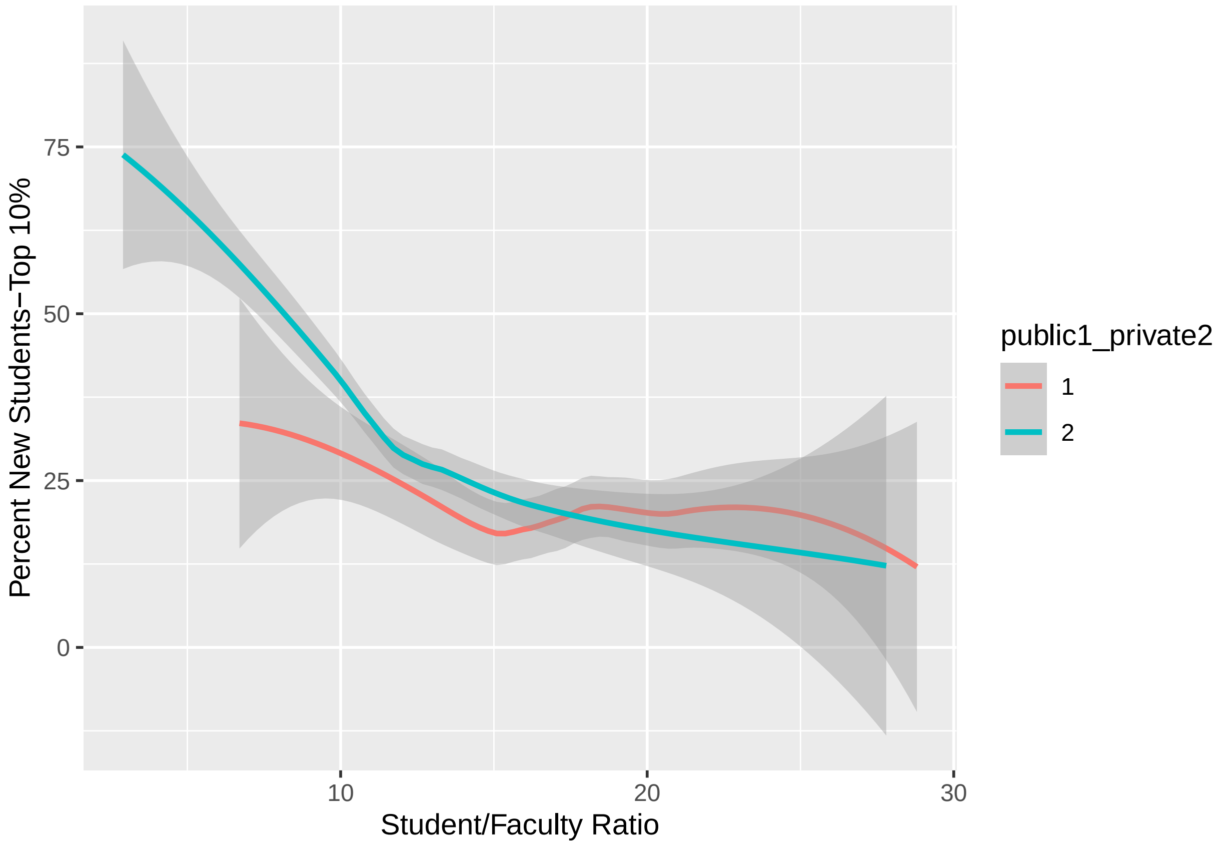
```r
ggplot(data = univ_complete) +
  geom_boxplot(mapping = aes(x = public1_private2, y = new_stud)) +
  xlab("Public Vs Private") +
  ylab("New Students Who Enrolled")
```



Private schools from enrollment and accepted boxplots looks almost identical outside of the change in the y-axis.

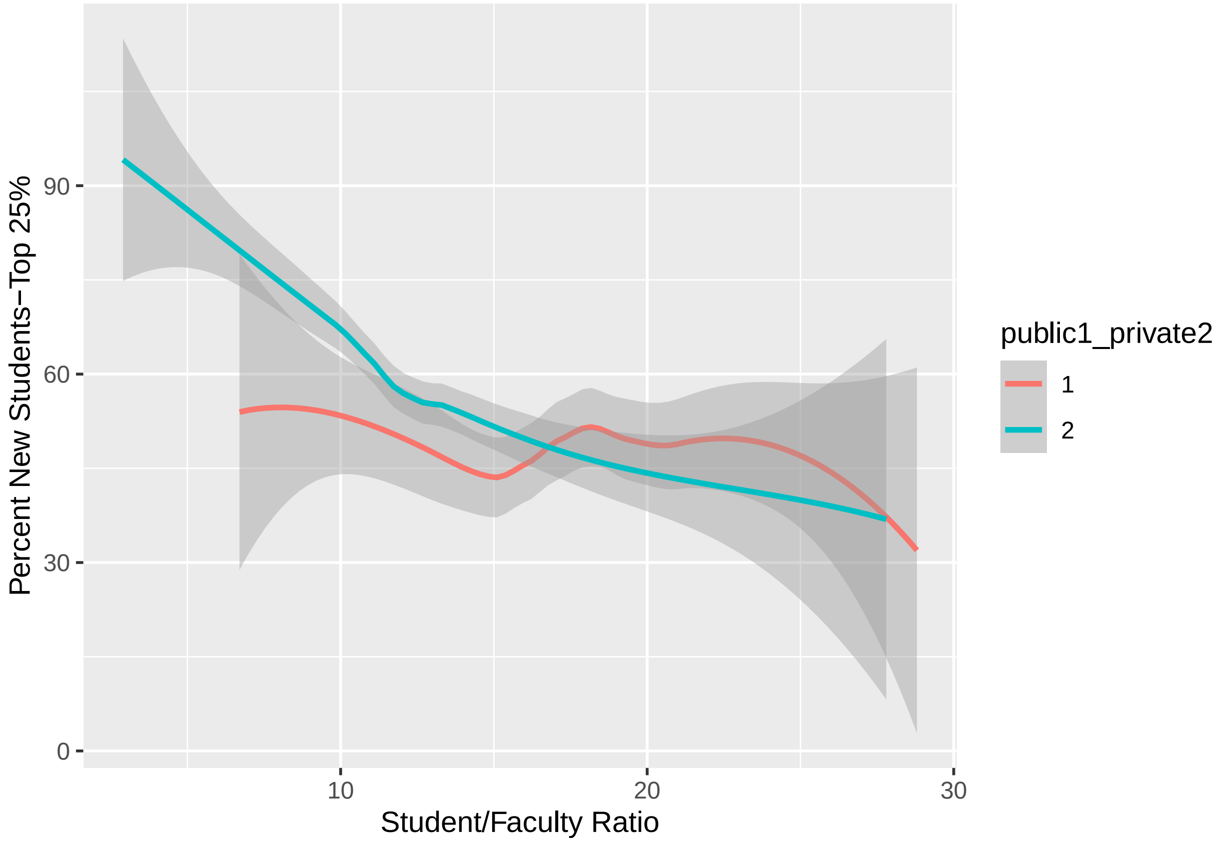
```
ggplot(data = univ_complete) +
  geom_smooth(mapping = aes(x = stud_fac_ratio, y = new_stud_10, color = public1_private2), se = TRUE )
  xlab("Student/Faculty Ratio") +
  ylab("Percent New Students-Top 10%")
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



```
ggplot(data = univ_complete) +
  geom_smooth(mapping = aes(x = stud_fac_ratio, y = new_stud_25, color = public1_private2), se = TRUE )
  xlab("Student/Faculty Ratio") +
  ylab("Percent New Students-Top 25%")
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



This line plot illustrates an interesting trend, universities who have a lower student/faculty ratio also attract a higher percentage of students from the top 10% and 25% of their graduating class. Noticeable higher percentages of students from 10 & 25% in Private schools. Probably Ivy league schools.

```
univ_complete %>%
  filter(new_stud_10 > .45 & stud_fac_ratio <=10) %>%
  group_by(public1_private2) %>%
  select(new_stud_10, new_stud_25, perc_PHD, stud_fac_ratio) %>%
  summarise(n = n(), across (1:4, mean))
```

```
## Adding missing grouping variables: ‘public1_private2‘
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 2 x 6
##   public1_private2     n new_stud_10 new_stud_25 perc_PHD stud_fac_ratio
##   <fct>            <int>       <dbl>       <dbl>    <dbl>          <dbl>
## 1 1                    8        32.5        56.4     82             8.88
## 2 2                   47        52.1        75.0     84.0           7.85
```

It seems like there is may be a relationship between small class sizes, PHD faculty, and students from the top 10 and 25% of their class.

Part 2: K-means Clustering

Normalize Continuous Dataset

```
norm <- preProcess(univ_continuous, method = c("scale", "center"))
univ_continuous <- predict(norm, univ_continuous)
head(univ_continuous)
```

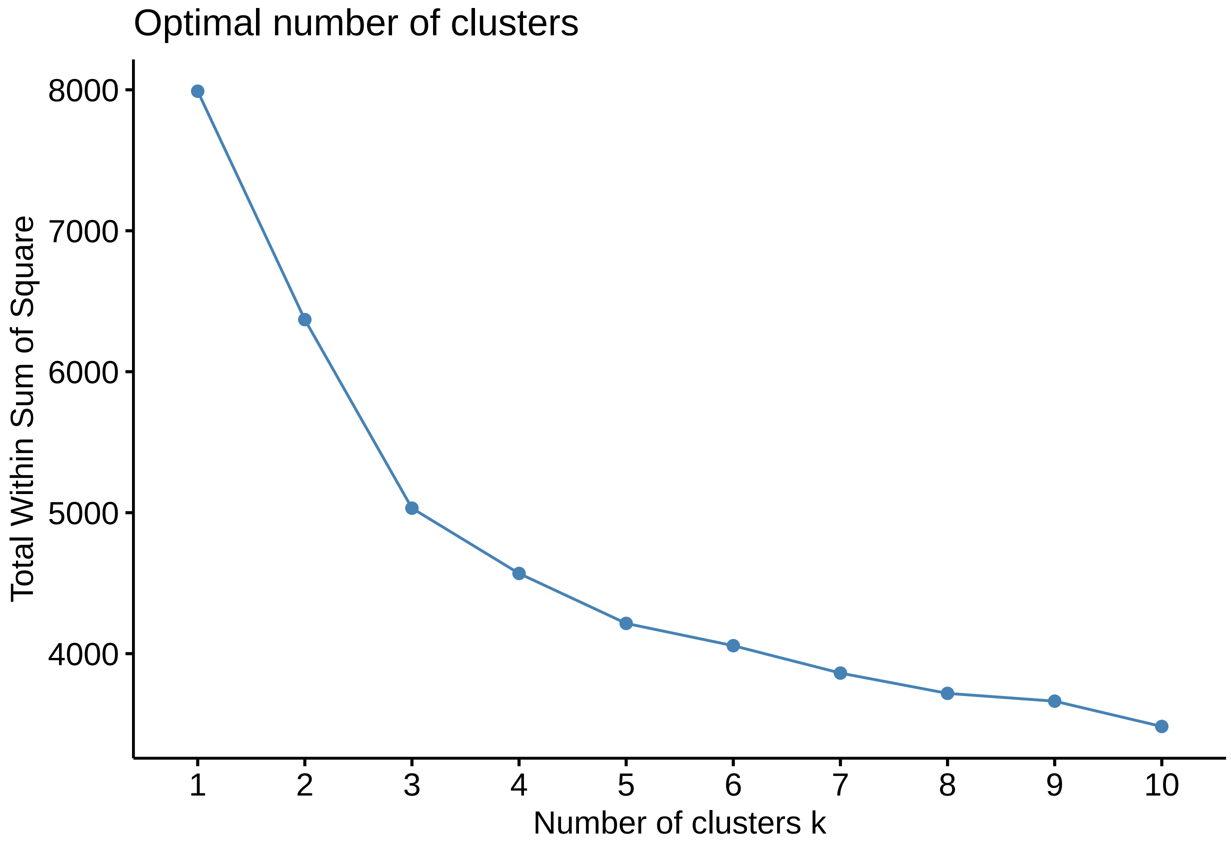
```
##   appli_recd appli_accepted   new_stud new_stud_10 new_stud_25 ft_undergrad
## 1 -0.7253139     -0.7656329 -0.7925715  -0.6500683  -0.5732933   -0.7097404
## 2 -0.7368529     -0.7772155 -0.7554388  -1.2994472  -1.5573355   -0.6576975
## 3 -0.5750612     -0.5890979 -0.5391950   2.1097921   1.5915994   -0.4683728
## 4 -0.6234268     -0.6162571 -0.7139374  -0.1089192  -0.4256870   -0.6478457
## 5  0.3109878     -0.2248447 -0.4867723   0.1075405   0.2139404   -0.5686035
## 6 -0.3315143     -0.3207008  0.1717883  -0.2171490  -1.0161123    0.7275427
##   pt_undergrad   in_state  out_state       room      board   add_fees
## 1    0.0462840 -0.3347297 -0.6993021 -0.8428467  0.6669350 -0.6997824
## 2    0.6802614 -1.3893276 -1.2406234  0.4106795  0.2259098 -0.9695550
## 3   -0.3819742  0.4084555  0.2516051 -0.2399203  0.5434479 -0.7278837
## 4   -0.4343744 -0.2404720 -0.5786992 -1.1793639  0.7374990 -0.7840863
## 5   -0.4389028 -0.6780450 -1.1385749 -1.1176691 -1.0266018  0.1095355
## 6    2.5233243 -1.3026831 -1.4229193 -0.4011680  1.9723696 -0.2473512
##   book_costs personal_costs    perc_PHD stud_fac_ratio  grad_rate
## 1  1.5394532      0.2758088  0.16752615   -0.529035524 -2.7862940
## 2 -0.2989446     -0.2199034 -2.05260943   -1.144600894 -1.4637549
## 3 -0.9117438     -0.6041537  0.04751883    0.009584174  0.3547362
## 4 -0.2989446     -0.3108329 -0.61252148   -0.657278310 -1.1882260
## 5  2.7650518      0.1291484 -1.03254714    0.394312530 -1.0780144
## 6  1.2330536      1.3024317  1.36759945   -1.862760492 -1.7943897
```

```
summary(univ_continuous)
```

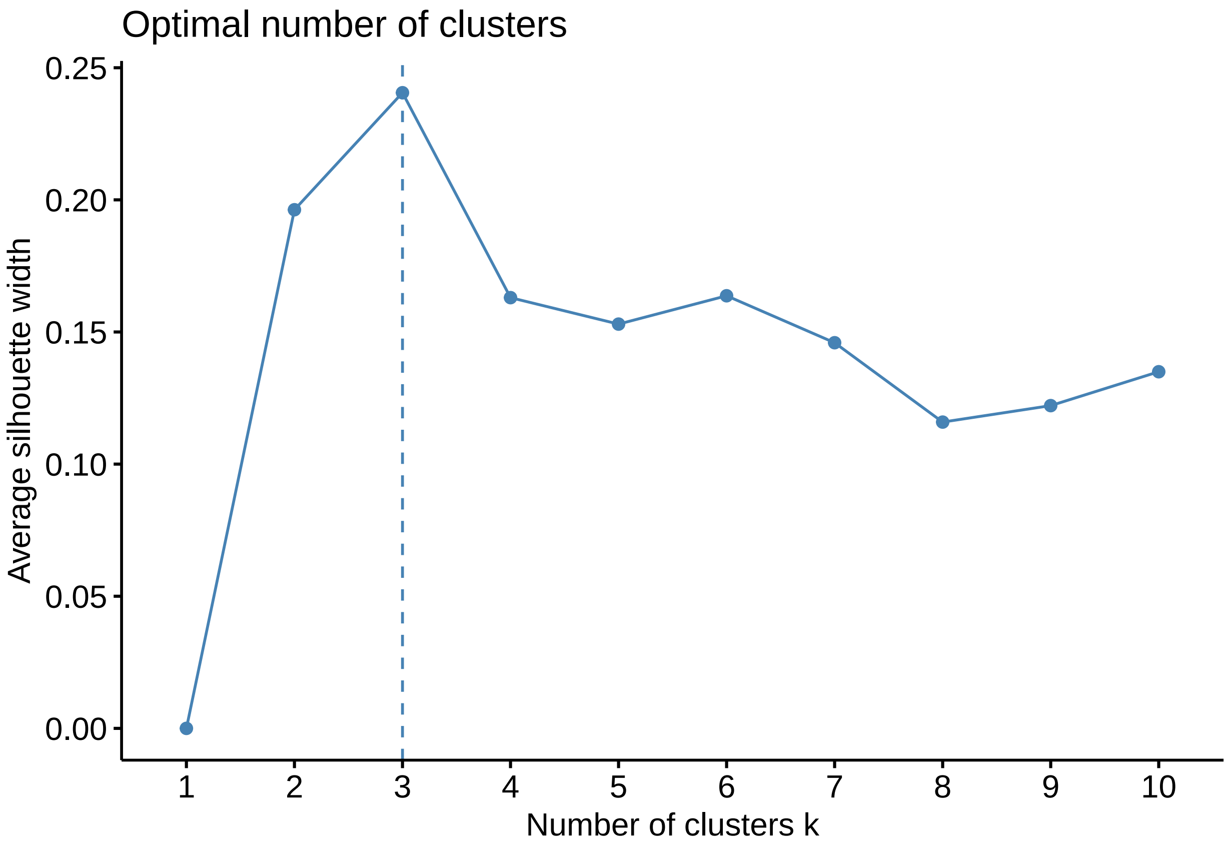
```
##    appli_recd      appli_accepted      new_stud        new_stud_10
##  Min.   :-0.7538   Min.   :-0.7996   Min.   :-0.8232   Min.   :-1.4618
##  1st Qu.:-0.5758   1st Qu.:-0.5701   1st Qu.:-0.5643   1st Qu.:-0.7042
##  Median :-0.3686   Median :-0.3339   Median :-0.3688   Median :-0.2713
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.1755   3rd Qu.: 0.1570   3rd Qu.: 0.1265   3rd Qu.: 0.4322
##  Max.   :11.0349   Max.   : 9.6923   Max.   : 6.1283   Max.   : 3.6791
##   new_stud_25        ft_undergrad      pt_undergrad        in_state
##  Min.   :-2.29537   Min.   :-0.7097   Min.   :-0.51524   Min.   :-1.59488
##  1st Qu.:-0.77010   1st Qu.:-0.5450   1st Qu.:-0.46316   1st Qu.:-1.04338
##  Median :-0.08127   Median :-0.3958   Median :-0.32246   Median : 0.08182
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.65676   3rd Qu.: 0.1055   3rd Qu.: 0.04628   3rd Qu.: 0.69594
##  Max.   : 2.18202   Max.   : 6.0139   Max.   :13.61017   Max.   : 1.93833
##    out_state            room             board              add_fees
##  Min.   :-2.2105   Min.   :-2.2170   Min.   :-2.80658   Min.   :-1.0370
##  1st Qu.:-0.7619   1st Qu.:-0.6746   1st Qu.:-0.65614   1st Qu.:-0.6787
##  Median :-0.1102   Median :-0.1838   Median :-0.07046   Median :-0.2783
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.6287   3rd Qu.: 0.6196   3rd Qu.: 0.52581   3rd Qu.: 0.3006
##  Max.   : 2.2091   Max.   : 3.6384   Max.   : 4.26746   Max.   : 8.0594
##    book_costs      personal_costs       perc_PHD        stud_fac_ratio
##  Min.   :-2.8114   Min.   :-1.5574   Min.   :-3.9127   Min.   :-2.8374
##  1st Qu.:-0.2989   1st Qu.:-0.6775   1st Qu.:-0.6125   1st Qu.:-0.6829
##  Median :-0.2989   Median :-0.1642   Median : 0.1675   Median :-0.1443
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.3139   3rd Qu.: 0.4225   3rd Qu.: 0.8276   3rd Qu.: 0.6380
##  Max.   :10.9766   Max.   : 8.0488   Max.   : 1.7876   Max.   : 3.8056
##    grad_rate
##  Min.   :-2.7863
##  1st Qu.:-0.6923
##  Median : 0.0241
##  Mean   : 0.0000
##  3rd Qu.: 0.7405
##  Max.   : 2.8896
```

Mean is 0 so all the data is now normalized.

```
fviz_nbclust(univ_continuous, kmeans, method = "wss")
```



```
fviz_nbclust(univ_continuous, kmeans, method = "silhouette")
```

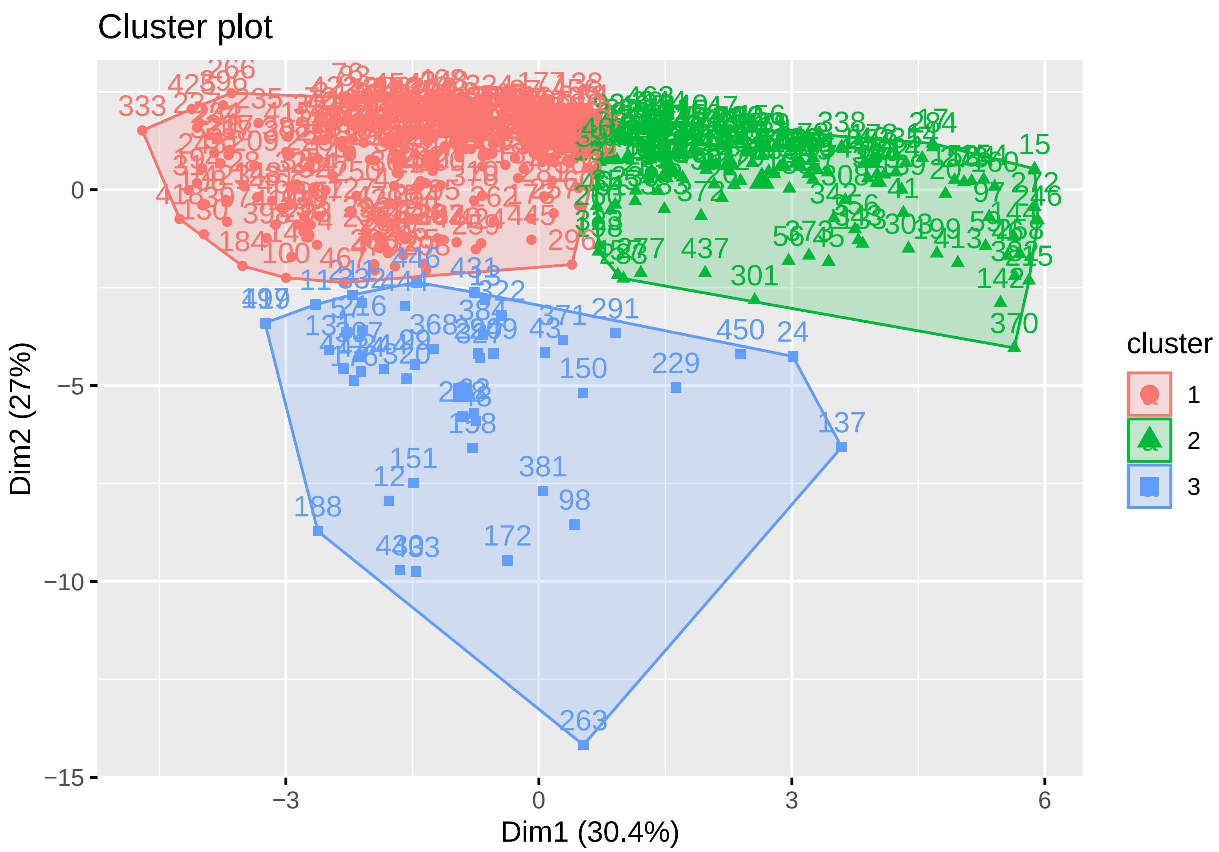


3 clusters would seem to me to be reasonable since you have smaller private and state schools, larger state schools, and ivy league schools. Optimal k would be 3 due to the "elbow" of the curve being at that point.

K-means for k = 3 Analysis

```
univ_3kmeans <- kmeans(univ_continuous, centers = 3, nstart = 25)
```

```
fviz_cluster(univ_3kmeans, data = univ_continuous)
```



### Finding Cluster Distances

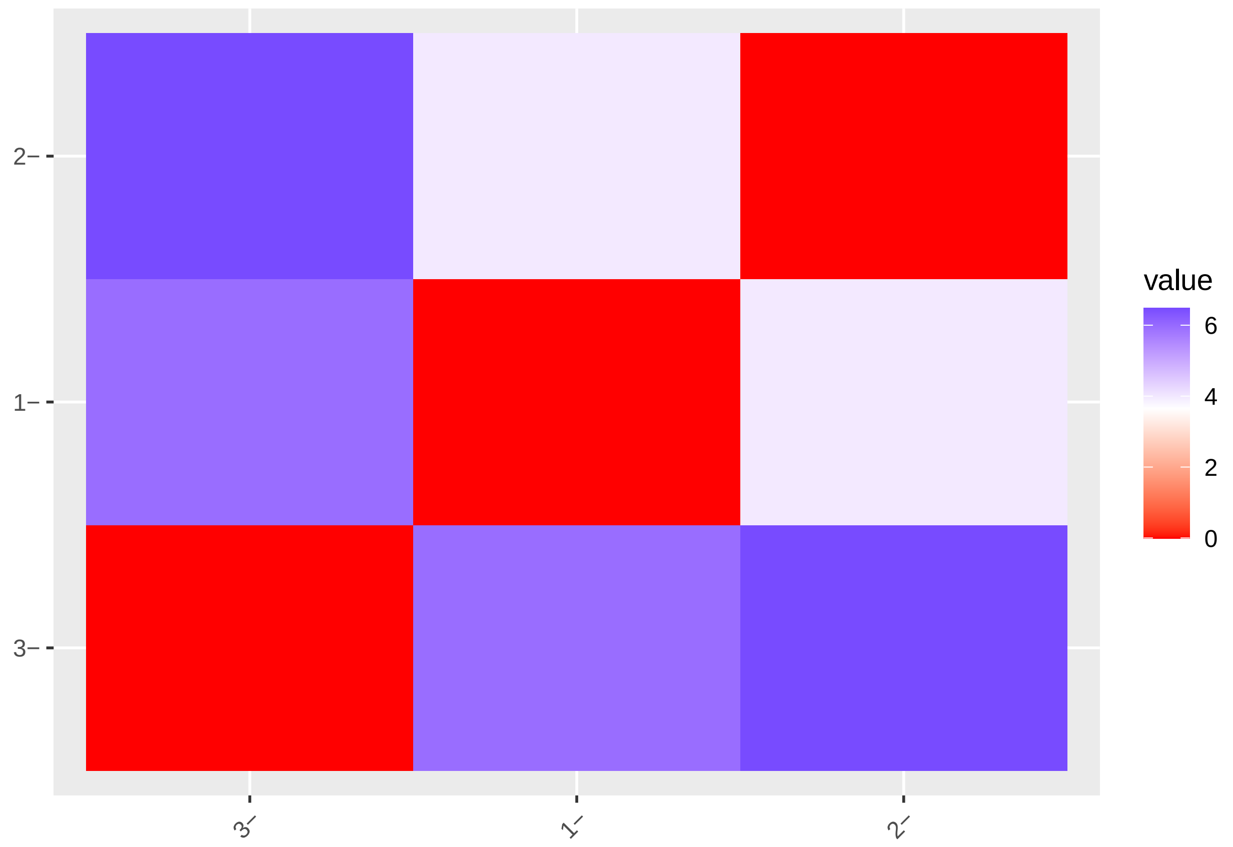
```
get_dist(univ_3kmeans$centers) -> dist_3kmeans
dist_3kmeans
```

```
##          1        2
## 2 3.983054
## 3 5.959276 6.478500
```

```
mean(dist_3kmeans)
```

```
## [1] 5.47361
```

```
fviz_dist(dist_3kmeans)
```

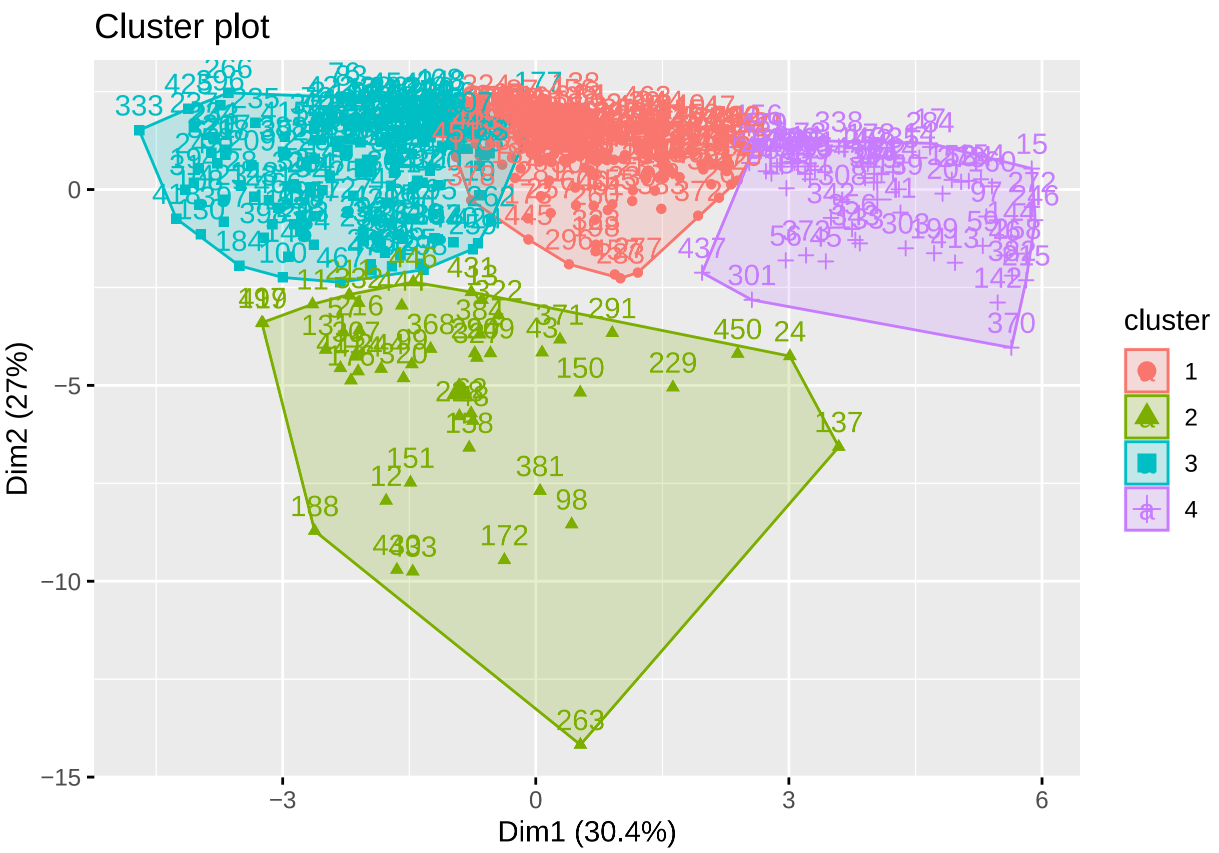


K-means for k = 4 Analysis

Going to test both k values to see if 4 could be better. I suspect it probably will not.

```
univ_4kmeans <- kmeans(univ_continuous, centers = 4, nstart = 25)
```

```
fviz_cluster(univ_4kmeans, data = univ_continuous)
```



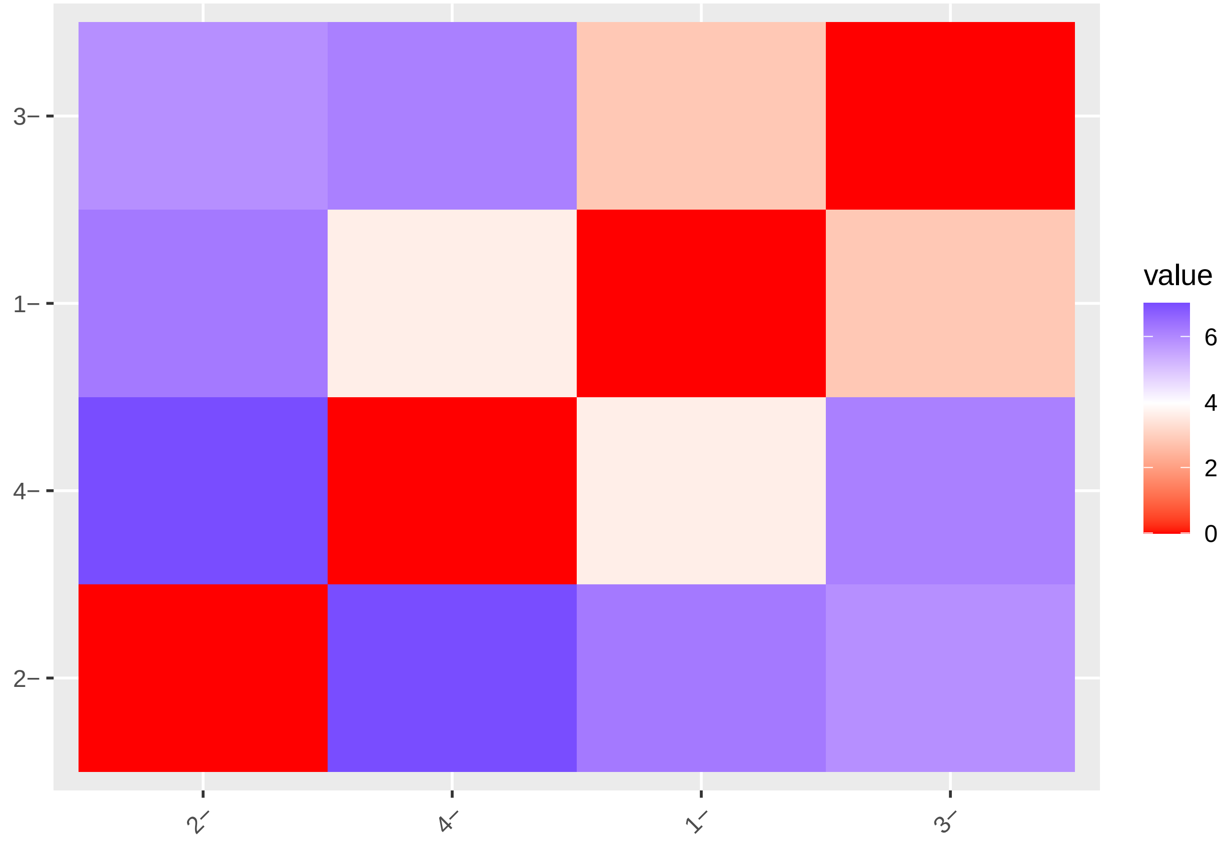
```
get_dist(univ_4kmeans$centers) -> dist_4kmeans
dist_4kmeans
```

```
##          1        2        3
## 2 6.271104
## 3 2.821317 5.877195
## 4 3.613022 7.015434 6.142001
```

```
mean(get_dist(univ_4kmeans$centers))
```

```
## [1] 5.290012
```

```
fviz_dist(dist_4kmeans)
```



In k = 4, clusters 1, 3, & 4 are close together and doesn't have much distance from each other. k = 3 has a higher distance average than k = 4 and also seems each cluster is further apart/better clustering than k = 4.

Combine Cluster labels to the unnormalized dataset.

Doing this to help include observations of the categorical variables and to also see trends in the clusters better.

```
univ_continuous<- cbind(univ_continuous, cluster = univ_3kmeans$cluster)
```

Cluster centers

Creating a df for the centers and will use later for Tufts University.

```
univ_centers <- data.frame(univ_3kmeans$centers)
univ_centers
```

```
##    appli_recd appli_accepted   new_stud new_stud_10 new_stud_25 ft_undergrad
## 1 -0.35953828    -0.34918455 -0.3171053  -0.5020886  -0.5128195   -0.2952142
## 2  0.05140256    -0.04367128 -0.1683551   0.8795798   0.8620961   -0.2324464
## 3  1.98179657     2.22992267  2.4447222   0.1334215   0.2545856    2.5228452
##   pt_undergrad   in_state  out_state       room      board    add_fees
## 1   -0.1217682 -0.4036544 -0.5263964 -0.3588740 -0.3938990 -0.05832646
## 2   -0.3130216  1.0620416  1.1158839  0.6698444  0.7756859 -0.04496556
## 3    1.7486849 -1.0500277 -0.4918168 -0.0388330 -0.1745795  0.49531762
##    book_costs personal_costs   perc_PHD stud_fac_ratio  grad_rate
## 1 -0.06621454     0.05935933 -0.5322257      0.2810858 -0.4171456
## 2  0.07122705    -0.39665857  0.7659627     -0.7036167  0.8426062
## 3  0.16358567     0.93858632  0.6840794      0.6139980 -0.2538234
```

Cluster Labels to Normalize Dataset

```
univ_complete <- cbind(univ_complete, cluster = univ_3kmeans$cluster)
```

Created a Variable "Acceptance rate"

I found this more helpful in comparing the clusters than application received and accepted.

```
univ_complete %>%
  mutate(accept_rate = appli_accepted/appli_recd) -> univ_complete
```

Comparing Clusters

```
univ_complete %>%
  group_by(cluster) %>%
  summarise(across(4:21, mean)) # focused on the mean for each cluster since it also represents the cen
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 19
##   cluster appli_recd appli_accepted new_stud new_stud_10 new_stud_25
##     <int>      <dbl>          <dbl>    <dbl>       <dbl>       <dbl>
## 1       1      1683.          1189.     490.        18.7        45.2
## 2       2      3357.          1954.     627.        44.3        73.2
## 3       3     11219.          7646.    3019.        30.5        60.8
## # ... with 13 more variables: ft_undergrad <dbl>, pt_undergrad <dbl>,
## #   in_state <dbl>, out_state <dbl>, room <dbl>, board <dbl>, add_fees <dbl>,
## #   book_costs <dbl>, personal_costs <dbl>, perc_PHD <dbl>,
## #   stud_fac_ratio <dbl>, grad_rate <dbl>, accept_rate <dbl>
```

Cluster 1: higher acceptance rate, lower graduation rate, lower percent of PHD faculty, lower tuition, lower pt undergrad, lower percent incoming freshmen from the top 10 and 25% of HS graduating class.

Cluster 2: lower acceptance rate, higher graduation rate, lower faculty/student ratio, high percent of PHD, not much variance between in-state/out-of-state tuition, high tuition, high percent incoming freshmen from the top 10 and 25% of HS graduating class.

Cluster 3: closer to average acceptance rate, high student/faculty ratio, high percent of PHD faculty, low in-state tuition, higher pt undergrad, high ft undergrad, closer to average percent incoming freshmen from the top 10 and 25% of HS graduating class, high amount of applications received and accepted.

NOTE: a summary/observations for the next three tables is located below Table 3. Table label at bottom of each table for greater clarity.

```
univ_complete %>% # to see what proportion in each cluster is public/private.
  group_by(cluster) %>%
  count(public1_private2)
```

```
## # A tibble: 6 x 3
## # Groups:   cluster [3]
##   cluster public1_private2     n
##     <int> <fct>            <int>
## 1       1 1                   84
## 2       1 2                  191
## 3       2 1                    3
## 4       2 2                  147
## 5       3 1                   41
## 6       3 2                    5
```

TABLE 1 (above)

```
univ_complete %>% # to see what proportion each cluster is located by state.
  group_by(cluster) %>%
  count(state) %>%
  filter(n>=10)
```

```
## # A tibble: 12 x 3
## # Groups:   cluster [2]
##    cluster state     n
##      <int> <chr> <int>
##  1       1 IA       16
##  2       1 MO       12
##  3       1 NC       16
##  4       1 NY       18
##  5       1 OH       13
##  6       1 PA       19
##  7       1 TN       11
##  8       1 TX       14
##  9       2 CA       10
## 10       2 MA       12
## 11       2 NY       18
## 12       2 PA       20
```

TABLE 2 (above)

```
univ_complete %>%
  group_by(cluster) %>%
  filter(cluster == 3 & public1_private2 == 2)
```

```
## # A tibble: 5 x 22
## # Groups:   cluster [1]
##   college_name state public1_private2 appli_recd appli_accepted new_stud
##   <chr>        <chr> <fct>                 <int>          <int>    <int>
## 1 University ~ CA    2                     12229           8498     2477
## 2 University ~ DE    2                     14446          10516     3252
## 3 Boston Univ~ MA    2                     20192          13007     3810
## 4 Northeaster~ MA    2                     11901           8492     2517
## 5 Baylor Univ~ TX    2                      6075           5349     2367
## # ... with 16 more variables: new_stud_10 <int>, new_stud_25 <int>,
## #   ft_undergrad <int>, pt_undergrad <int>, in_state <int>, out_state <int>,
## #   room <int>, board <int>, add_fees <int>, book_costs <int>,
## #   personal_costs <int>, perc_PHD <int>, stud_fac_ratio <dbl>,
## #   grad_rate <int>, cluster <int>, accept_rate <dbl>
```

TABLE 3 (above)

SUMMARY Cluster 1 is mostly private religious schools, private liberal art schools, and small state schools in the midwest/east regions of the US.

Cluster 2 is Ivy league universities mostly located in the East, New England areas including California.

Cluster 3 are mostly large state schools spread all over the US.

Possible Additional External Information

Other external information that could help to explain these clusters could be financial aid awarded, scholarships awarded, GPA, ethnicity, & socieoeconomic status.

Part 3: Tufts University

1. Separate Tufts information into df.

```
univ_missing %>%
  filter(college_name == "Tufts University") -> tufts
tufts
```

```
## # A tibble: 1 x 20
##   college_name state public1_private2 appli_recd appli_accepted new_stud
##   <chr>        <chr>            <dbl>      <dbl>          <dbl>    <dbl>
## 1 Tufts Unive~ MA                   2       7614           3605     1205
## # ... with 14 more variables: new_stud_10 <dbl>, new_stud_25 <dbl>,
## #   ft_undergrad <dbl>, pt_undergrad <dbl>, in_state <dbl>, out_state <dbl>,
## #   room <dbl>, board <dbl>, add_fees <dbl>, book_costs <dbl>,
## #   personal_costs <dbl>, perc_PHD <dbl>, stud_fac_ratio <dbl>, grad_rate <dbl>
```

2. Normalize Tufts df using the preProcess univ_continuous df normalization.

```
tufts_original <- tufts
tufts_norm <- predict(norm, tufts)
tufts_norm
```

```
## # A tibble: 1 x 20
##   college_name state public1_private2 appli_recd appli_accepted new_stud
##   <chr>        <chr>            <dbl>      <dbl>          <dbl>    <dbl>
## 1 Tufts Unive~ MA                   2       1.10          0.616    0.463
## # ... with 14 more variables: new_stud_10 <dbl>, new_stud_25 <dbl>,
## #   ft_undergrad <dbl>, pt_undergrad <dbl>, in_state <dbl>, out_state <dbl>,
## #   room <dbl>, board <dbl>, add_fees <dbl>, book_costs <dbl>,
## #   personal_costs <dbl>, perc_PHD <dbl>, stud_fac_ratio <dbl>, grad_rate <dbl>
```

Tufts Distance from Cluster Centers

```
tufts_dist <- rbind(univ_centers, tufts_norm[4:20])
get_dist(tufts_dist, method = "euclidean")
```

```
##           1        2        3
## 2  3.983054
## 3  5.959276 6.478500
## 11 6.640413 2.751310 6.905137
```

Tufts is closest to cluster 2, at a distance of 2.75. Tufts University should be included in cluster 2.

```
univ_complete %>%
  filter(cluster == 2) %>%
  summarise(mean(pt_undergrad))
```

```
##   mean(pt_undergrad)
## 1           313.5867
```

This is the value that should be imputed into the PT undergrad column in the Tufts University df.

Imputing Missing Value

```
univ_complete %>%
  filter(cluster == 2) -> c2 # created a new df with only cluster 2 so I could find the mean of the pt_

tufts[is.na(tufts$pt_undergrad), "pt_undergrad"] <- mean(c2$pt_undergrad)
tufts <- rbind(tufts_original, tufts)
tufts
```

```
## # A tibble: 2 x 20
##   college_name state public1_private2 appli_recd appli_accepted new_stud
##   <chr>        <chr>            <dbl>      <dbl>          <dbl>    <dbl>
## 1 Tufts Unive~ MA                   2       7614           3605     1205
## 2 Tufts Unive~ MA                   2       7614           3605     1205
## # ... with 14 more variables: new_stud_10 <dbl>, new_stud_25 <dbl>,
## #   ft_undergrad <dbl>, pt_undergrad <dbl>, in_state <dbl>, out_state <dbl>,
## #   room <dbl>, board <dbl>, add_fees <dbl>, book_costs <dbl>,
## #   personal_costs <dbl>, perc_PHD <dbl>, stud_fac_ratio <dbl>, grad_rate <dbl>
```

```
# showing Tufts information before imputing the value and after imputing the value to show that nothing
```

```
tufts %>%
  select(college_name, pt_undergrad) # shows that I correctly imputed the average for cluster 2 into pt_
```

```
## # A tibble: 2 x 2
##   college_name     pt_undergrad
##   <chr>                   <dbl>
## 1 Tufts University          NA
## 2 Tufts University         314.
```