University Prediction with K-Means Clustering

Mark Bruner

10/16/2020

```r
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()
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

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

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(factoextra)
```

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

```r
library(moments)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
set.seed(15)
```

Part 1: Cleaning Data

After viewing the structure of the data, **I changed the column type of applications accepted, received, enrolled, PT/FT undergrads to integers since they are all counts of students. Also, I converted the public/private school column to a factor.**

Additionally, I created a variable “acceptance rate” because it shows the “selectiveness” of a university and removed the accepted application column. The reason I did this was because the columns applications received and accepted doesn’t easily show us the “selectiveness” of a university. I kept the applications received column so we could easily get the accepted column back if needed. I also choose to keep the applications received column over the accepted column because the number of accepted students is dependent on applications received. Applications received will also help us determine large schools from smaller schools.

```r
univ <- read_csv("/Users/markbruner/Google Drive/MSBA/Machine Learning/mbruner3/ML_mbruner3/Assignment 4
```

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

```r
str(univ) # showing intitial structure of the data before the changes.
```

```
## 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) # head and tail of data shows if the data seems normal.
```

```
## # 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>
```

```
univ %>% # renamed columns to make them easier to work with.
  rename(
         college_name  = ‘College Name‘,
         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
```

Changing Variable Type (Integers and Factors)

```
univ[, c(4:6, 9, 10)] <- sapply(univ[, c(4:6, 9, 10)], as.integer) # changed "counts" columns to intege
univ$public1_private2 <- as.factor(univ$public1_private2) # changed public/private to a factor type.
str(univ) # shows that these changes were made accurately.
```

```
## 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 Fairbanks" "Un
##  $ state           : chr [1:1302] "AK" "AK" "AK" "AK" ...
##  $ public1_private2: Factor w/ 2 levels "1","2": 2 1 1 1 1 2 1 1 1 2 ...
##  $ appli_recd      : int [1:1302] 193 1852 146 2065 2817 345 1351 4639 7548 805 ...
##  $ appli_accepted  : int [1:1302] 146 1427 117 1598 1920 320 892 3272 6791 588 ...
##  $ new_stud        : int [1:1302] 55 928 89 1162 984 179 570 1278 3070 287 ...
##  $ new_stud_10     : num [1:1302] 16 NA 4 NA NA NA 18 NA 25 67 ...
##  $ new_stud_25     : num [1:1302] 44 NA 24 NA NA 27 78 NA 57 88 ...
##  $ ft_undergrad    : int [1:1302] 249 3885 492 6209 3958 1367 2385 4051 16262 1376 ...
##  $ pt_undergrad    : int [1:1302] 869 4519 1849 10537 305 578 331 405 1716 207 ...
##  $ in_state        : num [1:1302] 7560 1742 1742 1742 1700 ...
##  $ out_state       : 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 ...
##  $ book_costs      : num [1:1302] 800 650 500 580 500 350 300 500 600 400 ...
##  $ personal_costs  : num [1:1302] 1500 2304 1162 1260 850 ...
##  $ perc_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 ...
##  $ grad_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()
##   .. )
```

Acceptance Rate

```
univ %>%
  mutate(accept_rate =  appli_accepted/appli_recd*100) %>%
  relocate(college_name, state, public1_private2, appli_accepted, accept_rate, appli_recd, new_stud) ->
```

Separating Continuous & Categorical Variables

```
univ_continuous <- as.data.frame(univ[, c(4:21)])
```

Exploratory Data Analysis

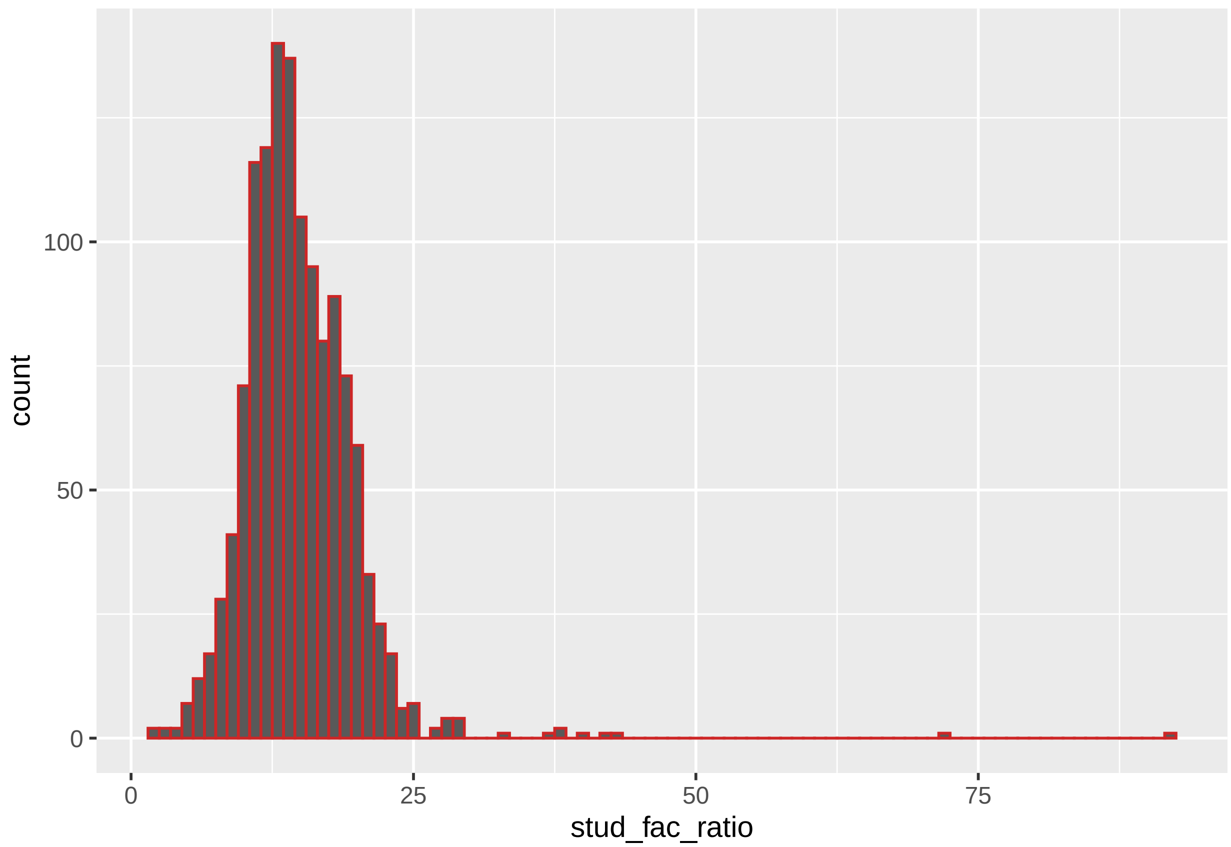
UNIVARIATE EXPLORATION Summary Statistics

```
summary(univ_continuous)
```

```
##  appli_accepted     accept_rate        appli_recd         new_stud
##  Min.   :   35.0   Min.   : 9.139   Min.   :   35.0   Min.   :  18.0
##  1st Qu.:  554.5   1st Qu.: 68.122   1st Qu.:  695.8   1st Qu.: 236.0
##  Median : 1095.0   Median : 78.261   Median : 1470.0   Median : 447.0
##  Mean   : 1870.7   Mean   : 75.479   Mean   : 2752.1   Mean   : 778.9
##  3rd Qu.: 2303.0   3rd Qu.: 86.087   3rd Qu.: 3314.2   3rd Qu.: 984.0
##  Max.   :26330.0   Max.   :100.000   Max.   :48094.0   Max.   :7425.0
##  NA’s   :11        NA’s   :13        NA’s   :10        NA’s   :5
##   new_stud_10     new_stud_25      ft_undergrad    pt_undergrad
##  Min.   : 1.00   Min.   :  6.00   Min.   :   59   Min.   :    1.0
##  1st Qu.:13.00   1st Qu.: 36.75   1st Qu.:  966   1st Qu.:  131.2
##  Median :21.00   Median : 50.00   Median : 1812   Median :  472.0
##  Mean   :25.67   Mean   : 52.35   Mean   : 3693   Mean   : 1081.5
##  3rd Qu.:32.00   3rd Qu.: 66.00   3rd Qu.: 4540   3rd Qu.: 1313.0
##  Max.   :98.00   Max.   :100.00   Max.   :31643   Max.   :21836.0
##  NA’s   :235     NA’s   :202      NA’s   :3       NA’s   :32
##     in_state       out_state         room          board         add_fees
##  Min.   :  480   Min.   : 1044   Min.   : 500   Min.   : 531   Min.   :   9.0
##  1st Qu.: 2580   1st Qu.: 6111   1st Qu.:1710   1st Qu.:1619   1st Qu.: 130.0
##  Median : 8050   Median : 8670   Median :2200   Median :1980   Median : 264.5
##  Mean   : 7897   Mean   : 9277   Mean   :2515   Mean   :2061   Mean   : 392.0
##  3rd Qu.:11600   3rd Qu.:11659   3rd Qu.:3040   3rd Qu.:2402   3rd Qu.: 480.0
##  Max.   :25750   Max.   :25750   Max.   :7400   Max.   :6250   Max.   :4374.0
##  NA’s   :30      NA’s   :20      NA’s   :321    NA’s   :498    NA’s   :274
##    book_costs    personal_costs    perc_PHD      stud_fac_ratio
##  Min.   :  90   Min.   :  75   Min.   :  8.00   Min.   : 2.30
##  1st Qu.: 480   1st Qu.: 900   1st Qu.: 57.00   1st Qu.:11.80
##  Median : 502   Median :1250   Median : 71.00   Median :14.30
##  Mean   : 550   Mean   :1389   Mean   : 68.65   Mean   :14.86
##  3rd Qu.: 600   3rd Qu.:1794   3rd Qu.: 82.00   3rd Qu.:17.60
##  Max.   :2340   Max.   :6900   Max.   :105.00   Max.   :91.80
##  NA’s   :48     NA’s   :181    NA’s   :32       NA’s   :2
##    grad_rate
##  Min.   :  8.00
##  1st Qu.: 47.00
##  Median : 60.00
##  Mean   : 60.41
##  3rd Qu.: 74.00
##  Max.   :118.00
##  NA’s   :98
```

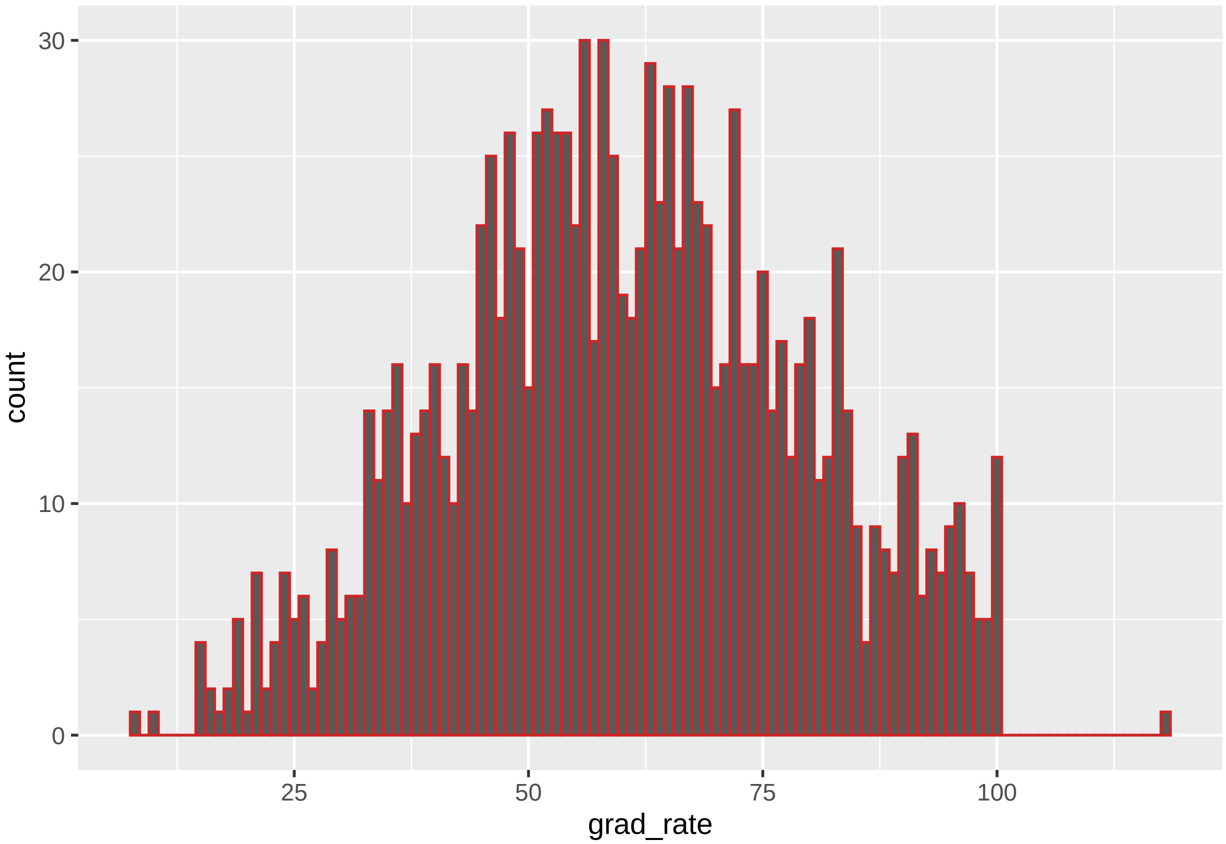
```r
univ_continuous %>%
  ggplot(mapping = aes(x= stud_fac_ratio)) +
  geom_histogram(color = "firebrick3", binwidth = 1)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



```
univ_continuous %>%
  ggplot(mapping = aes(x= grad_rate)) +
  geom_histogram(color = "firebrick3", binwidth = 1)
```

```
## Warning: Removed 98 rows containing non-finite values (stat_bin).
```



The range is large for applications received, enrolled new students, and pt/ft students. Percent of faculty with a PHD has a max of 105%. Most of the data skews positive as they have medians less than their means except for perc_PHD and acceptance rate which skews negative. Also, stud_fac_ratio & grad_rate have a fairly close mean and median which means they follow a fairly normal distribution. (Shown above)

You would expect that most of the data would skew negative since columns are mostly counts or costs. You would assume that the lower counts would occur more frequently and the higher counts to occur less frequently creating a positive skew of the data. Same is true for costs.

NOTE: After creating two df with and without outliers, the clustering model showed more overlap and less distance between clusters in the removed outlier df compared to the df with the outliers included. For this reason I decided to keep the outliers in the dataset since they seemed to help create more defined clusters which is what we want.

Part 2: K-means Clustering

Normalize Continuous Dataset

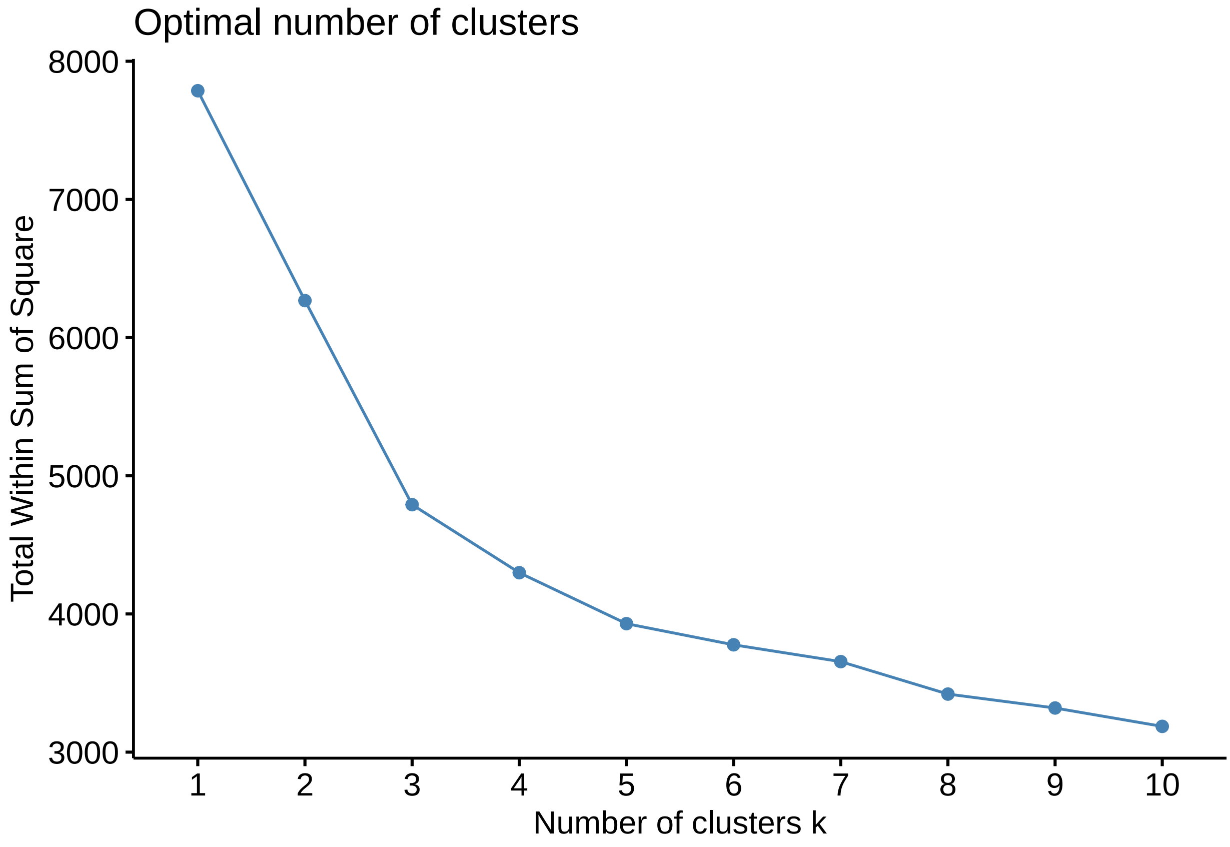
```
univ_complete <- univ_continuous[complete.cases(univ), ]
univ_complete_orig <- univ[complete.cases(univ), ] # Keeping the original data separate to combine clus


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

NOTE: I decided to keep Tufts University in the normalization of the dataset. The reason is that the column we will be predicting for will be missing. Also, it has other information that I think is more valuable to keep in rather than to remove.

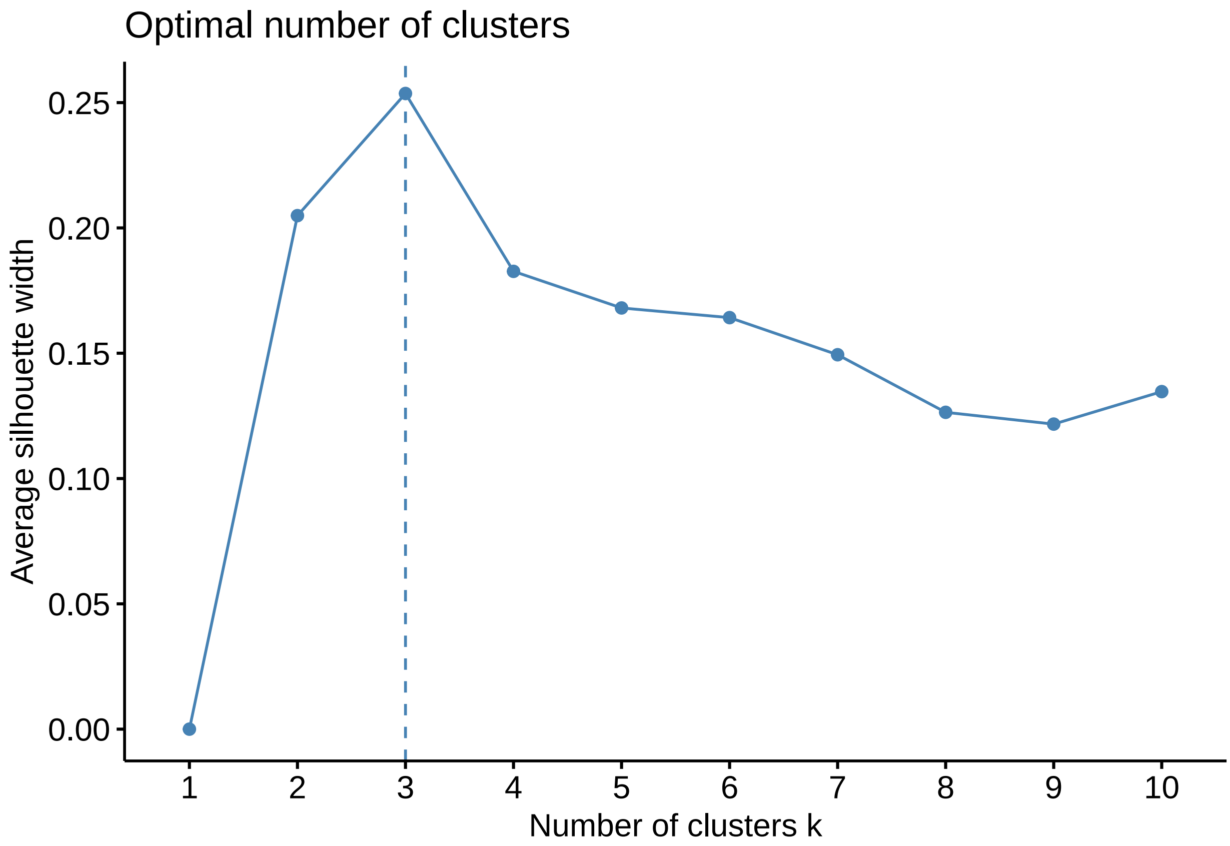
```
colMeans(is.na(univ_complete))
```

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

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



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

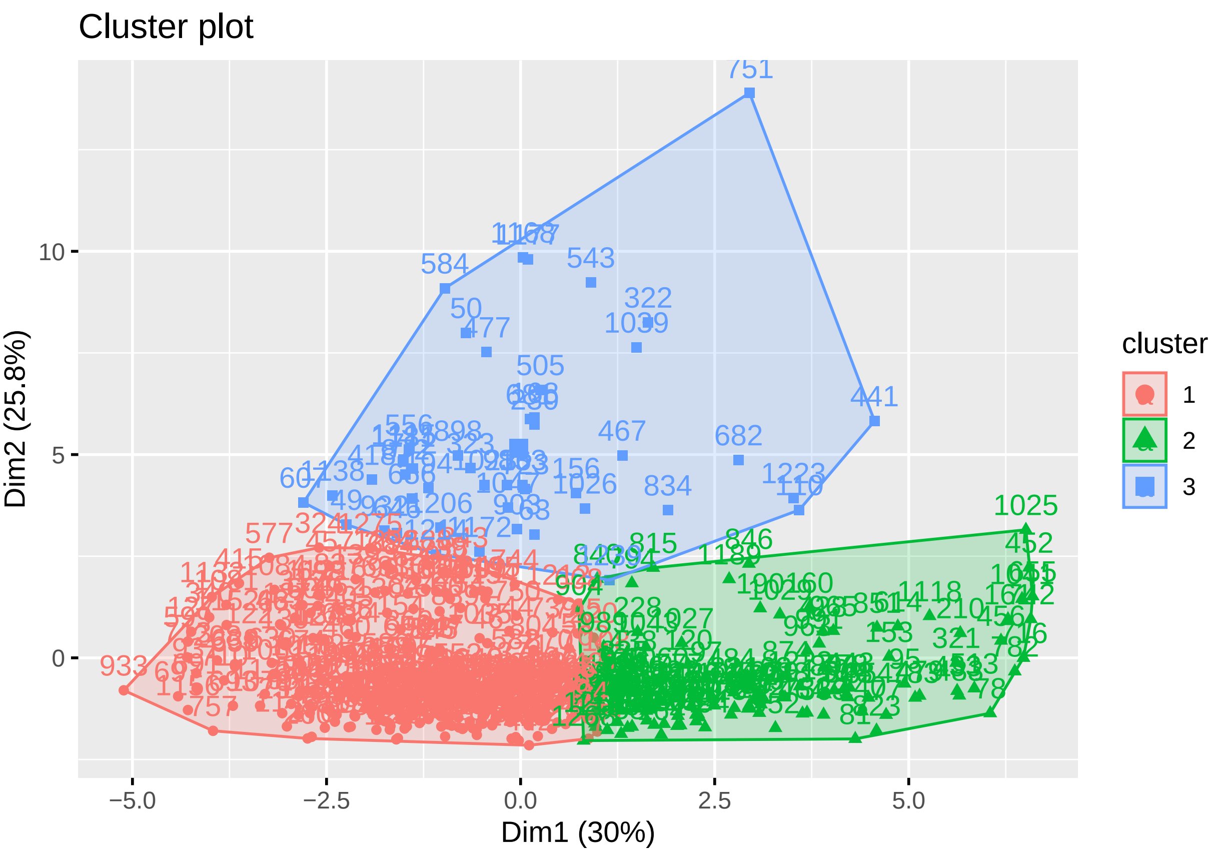


3 clusters would seem to me to be reasonable since, from my 15 years of working in a higher ed setting, you basically have 3 types of universities: 1) smaller private and state schools, 2) larger state schools, and 3) ivy league schools. Also, optimal k would be 3 due to the "elbow" of the curve being at that point and using the information from the silhouettte method.

K-means for k = 3 Analysis

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

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



Combine Cluster labels to the unnormalized dataset.

The reason I am doing this is to help include observations of the categorical variables and to also see trends in the clusters better.

```
univ_complete_orig <- cbind(univ_complete_orig, 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_accepted accept_rate appli_recd   new_stud new_stud_10 new_stud_25
## 1    -0.29920344   0.1533749 -0.3071808 -0.3268645  -0.3506859  -0.3104094
## 2     0.06975978  -0.6753957  0.2509845 -0.1551384   1.1902598   1.1542915
## 3     2.55039277  -0.2735606  2.3692062  2.5017239   0.2479596   0.3959052
##   ft_undergrad pt_undergrad    in_state  out_state       room       board
## 1   -0.3347932   -0.2880181 -0.06100492 -0.1614048 -0.4455185 -0.17740955
## 2   -0.2583164   -0.4817066  1.46985042  1.5778091  0.1876129  0.77317307
## 3    2.5259406    1.4233771 -0.79058727 -0.1611344 -0.2615128 -0.06524493
##      add_fees  book_costs personal_costs   perc_PHD stud_fac_ratio   grad_rate
## 1 -0.08839121 -0.07275290    -0.04593447 -0.1793118    -0.01022405 -0.08584503
## 2 -0.02327906  0.08965159    -0.57099334  1.0265918    -0.70846739  1.18501128
## 3  0.34270731  0.14280660     0.75905738  0.8983642     0.26229756  0.04950538
```

Cluster Labels to Normalize Dataset

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

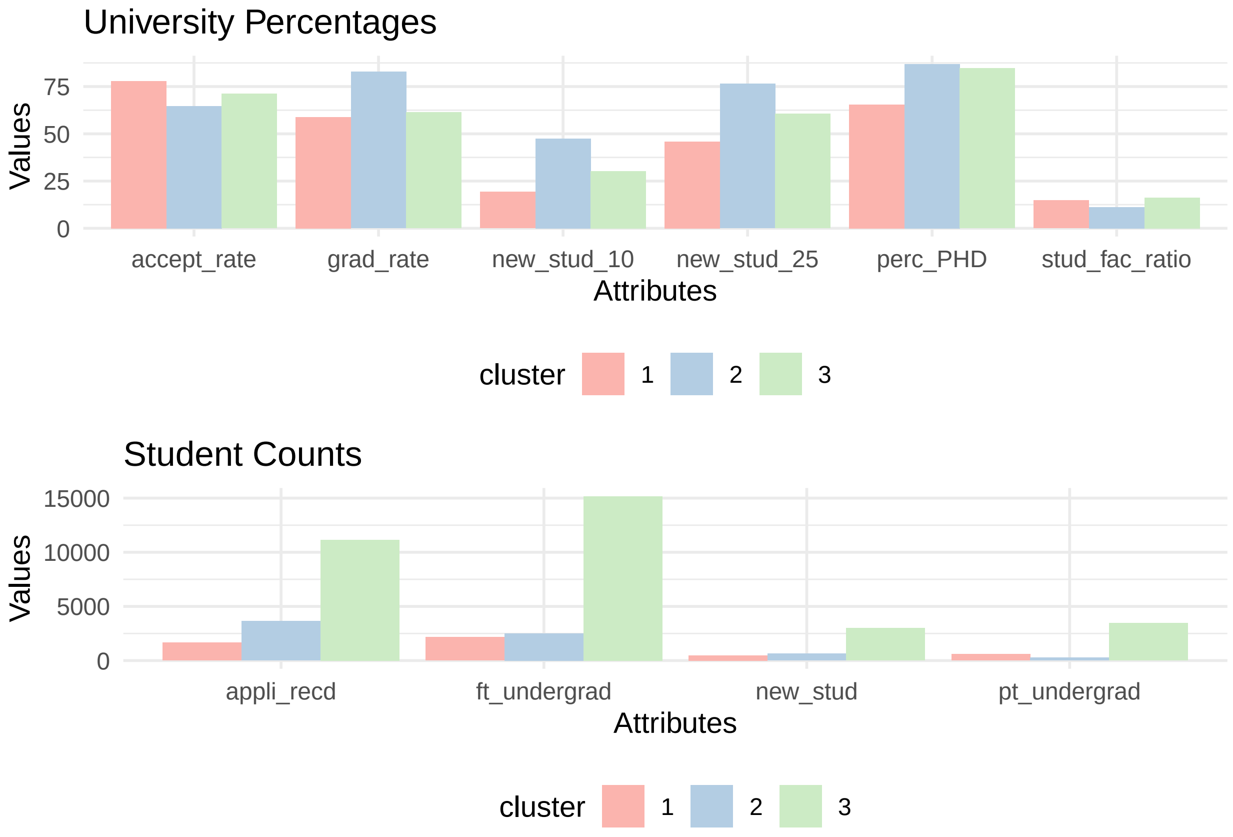
Comparing Clusters Graphically

```
univ_complete_orig %>%
  group_by(cluster) %>%
  summarise(across(5:21, mean)) -> univ_key # created a df of the means of each cluster unnormalized.
```

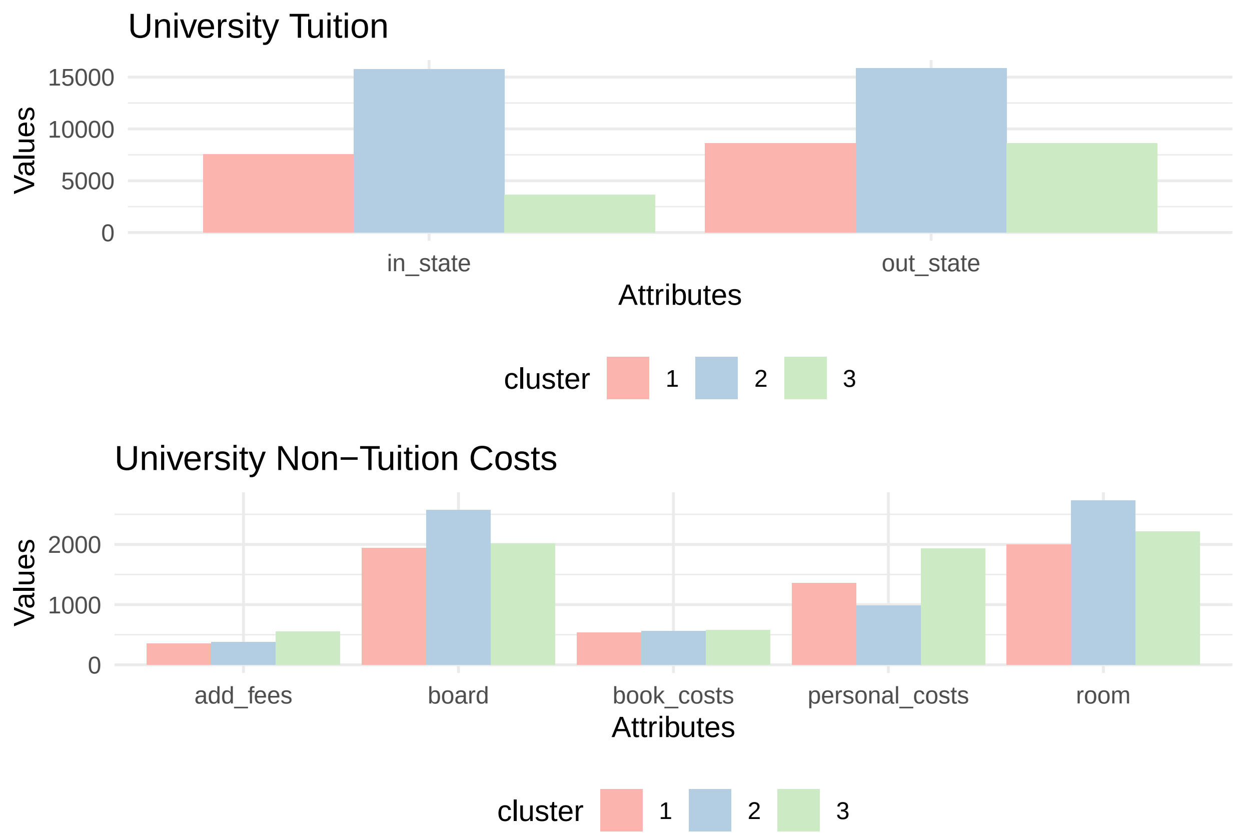
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
univ_key$cluster <- as.factor(univ_key$cluster) # made the cluster column a factor.

univ_key %>%  # rearranged the columns to organize the columns in more logically way. Groupings as foll
  relocate(cluster, stud_fac_ratio, accept_rate, new_stud_10, new_stud_25, perc_PHD, grad_rate) -> univ_

univ_key1 <- univ_key[, c(1:7)] # percentage columns
univ_key2 <- univ_key[, c(1, 8:11)] # count columns
univ_key3 <- univ_key[, c(1, 12:13)] # tuition columns
univ_key4 <- univ_key[, c(1, 14:18)] # costs columns

# reoganizing each key df into a "key", "value" column to be able to represent the data easier graphica
univ_key1 %>%
  gather(key = "key", value = "value", -cluster) -> univ_key1
univ_key2 %>%
  gather(key = "key", value = "value", -cluster) -> univ_key2
univ_key3 %>%
  gather(key = "key", value = "value", -cluster) -> univ_key3
univ_key4 %>%
  gather(key = "key", value = "value", -cluster) -> univ_key4

ggplot(univ_key1) +
 aes(x = key, fill = cluster, weight = value) +
 geom_bar(position = "dodge") +
 scale_fill_brewer(palette = "Pastel1") +
 labs(x = "Attributes", y = "Values", title = "University Percentages") +
 theme_minimal() +
 theme(legend.position = "bottom") -> p1

ggplot(univ_key2) +
 aes(x = key, fill = cluster, weight = value) +
 geom_bar(position = "dodge") +
 scale_fill_brewer(palette = "Pastel1") +
 labs(x = "Attributes", y = "Values", title = "Student Counts") +
```

```r
 theme_minimal() +
 theme(legend.position = "bottom") -> p2

ggplot(univ_key3) +
 aes(x = key, fill = cluster, weight = value) +
 geom_bar(position = "dodge") +
 scale_fill_brewer(palette = "Pastel1") +
 labs(x = "Attributes", y = "Values", title = "University Tuition") +
 theme_minimal() +
 theme(legend.position = "bottom") -> p3

ggplot(univ_key4) +
 aes(x = key, fill = cluster, weight = value) +
 geom_bar(position = "dodge") +
 scale_fill_brewer(palette = "Pastel1") +
 labs(x = "Attributes", y = "Values", title = "University Non-Tuition Costs") +
 theme_minimal() +
 theme(legend.position = "bottom") -> p4

grid.arrange(p1, p2)
```



```r
grid.arrange(p3, p4)
```



Cluster 1: The universities in cluster 1 have:

1. Lower student/faculty ratio
2. **Lowest** percent of faculty with PHD
3. Lower percentage of new students from the top 10/25% of their class
4. **Lowest** graduation rate but high acceptance rate
5. **Lowest** part-time/full-time undergraduates & new students
6. **Lowest** amount of applications from students
7. Tuition is about the same for students in-state as out-of-state

Cluster 2: The universities in cluster 2 have:

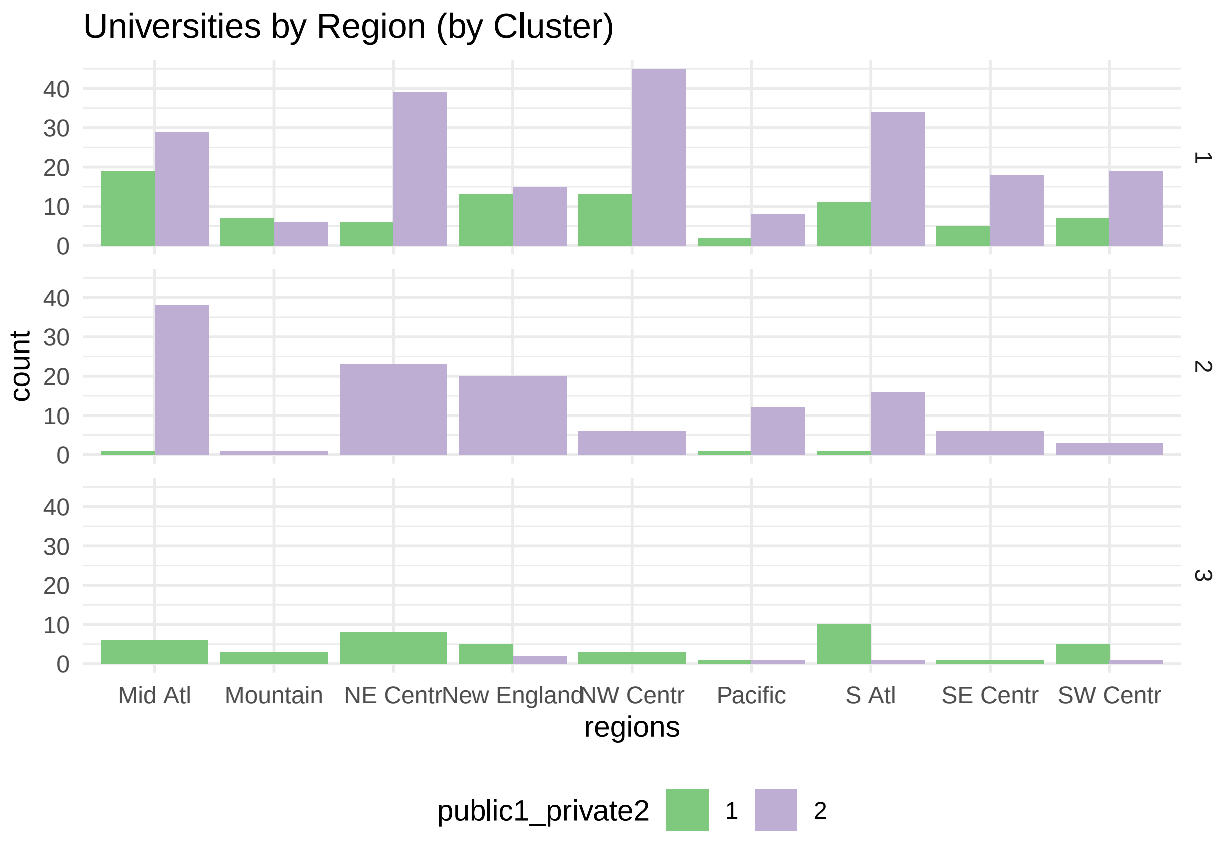
1. **Lowest** student/faculty ratio
2. **Highest** percent of faculty with PhD's
3. **Highest** percent of new students from the top 10 & 25%
4. **Lowest** acceptance rate
5. Students are mostly full-time
6. **Highest** tuition with in-state tuition equal to out-of-state tuition
7. **Highest** room and board

Cluster 3: The universities in cluster 3 have:

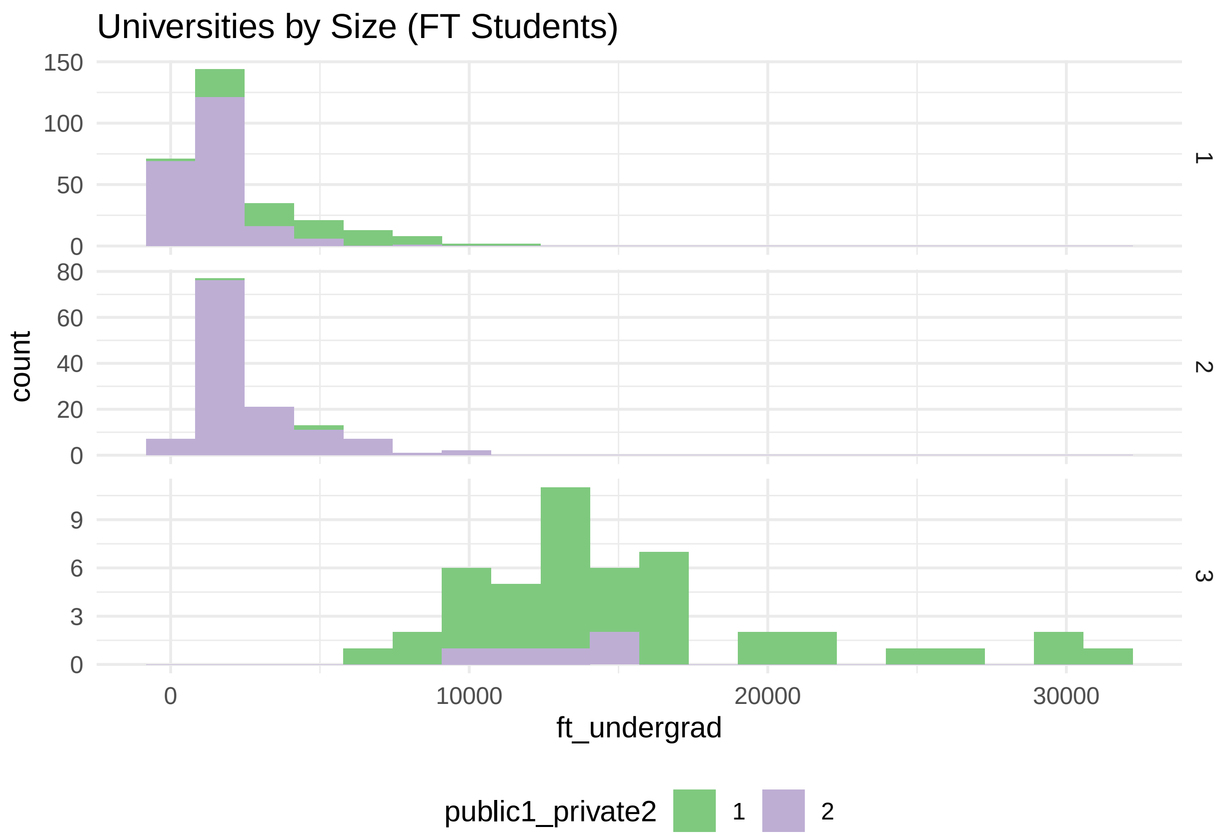
1. **Highest** student/faculty ratio

2. **Highest** applications received
3. **Highest** accepted new students
4. Lower acceptance rate than cluster 1
5. Higher graduation rate than cluster 1
6. Higher new students from the top 10 & 25% than cluster 1
7. **Lowest** in-state tuition but comparable out-of-state tuition to cluster 1

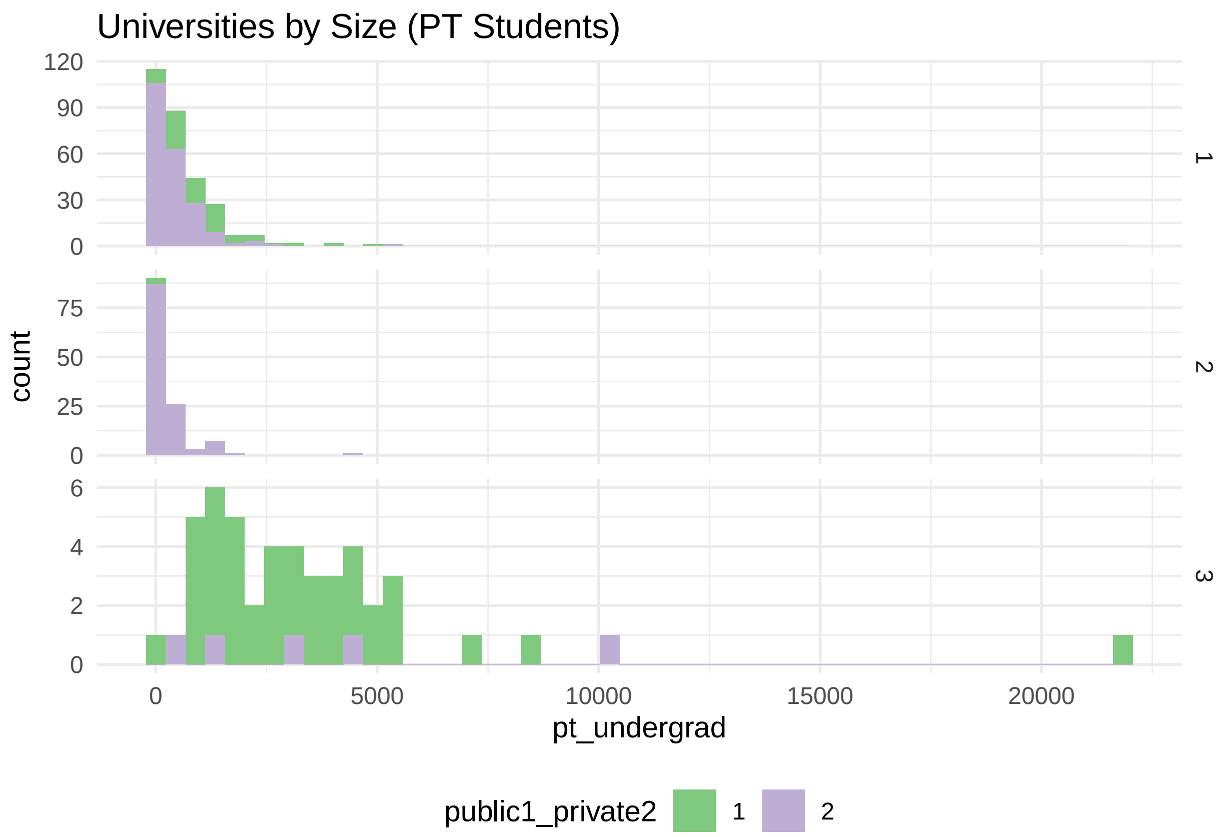
```r
# To better compare locations of universities, I used the US Census Bureau division of regions https://

pacific <- c("CA", "OR", "WA", "HI", "AK")
mountain <- c("AZ", "NV", "ID", "MT", "WY", "CO", "NM", "UT")
nw_central <- c("SD", "ND", "NE", "KS", "MO", "IA", "MN")
ne_central <- c("WI", "MI", "IL", "IN", "OH")
sw_central <- c("OK", "TX", "AR", "LA")
se_central <- c("KY", "TN", "MS", "AL")
s_atlantic <- c("GA", "FL", "SC", "NC", "WV", "VA", "MD", "DE", "DC")
mid_atlantic <- c("NY", "PA", "NJ")
new_england <- c("CT", "RI", "MA", "NH", "VT", "ME")

region.list <- list(
  Pacific = pacific,
  Mountain = mountain,
  "NW Centr" = nw_central,
  "NE Centr" = ne_central,
  "SW Centr" = sw_central,
  "SE Centr" = se_central,
  "S Atl" = s_atlantic,
  "Mid Atl" = mid_atlantic,
  "New England" = new_england
  )

# A function to apply region names to the new region column in the df.
univ_complete_orig$regions <- sapply(univ_complete_orig$state,
                 function(x) names(region.list)[grep(x,region.list)])

# Organizing the regions, state, and college name columns together.
univ_complete_orig$cluster <- as.factor(univ_complete_orig$cluster)
univ_complete_orig %>%
  relocate(college_name, state, regions) -> univ_complete_orig

univ_complete_orig$regions <- as.character(univ_complete_orig$regions)
univ_complete_orig$regions <- as.factor(univ_complete_orig$regions)

ggplot(univ_complete_orig) +
 aes(x = regions, fill = public1_private2) +
 geom_bar(position = "dodge") +
 scale_fill_brewer(palette = "Accent") +
 labs(title = "Universities by Region (by Cluster)") +
 theme_minimal() +
 facet_grid(vars(cluster), vars()) +
 theme(legend.position = "bottom")
```



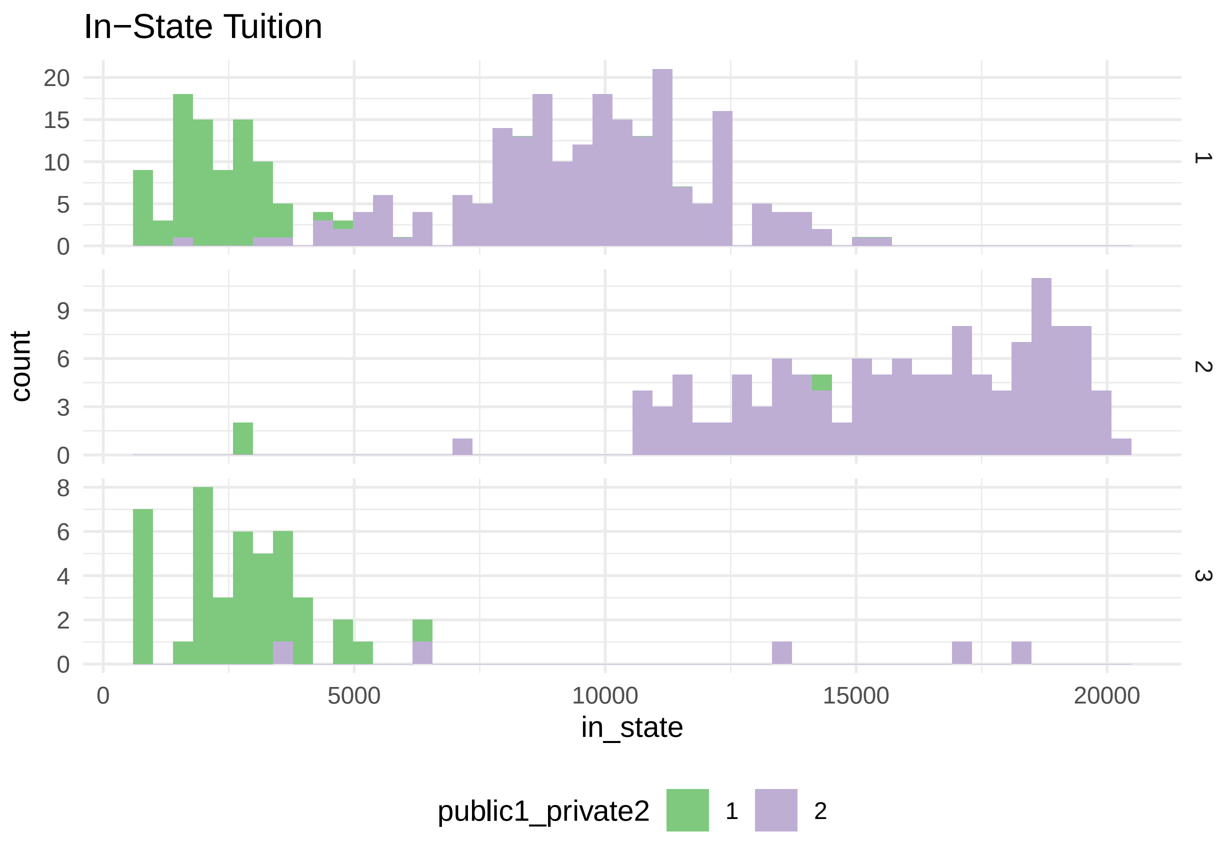
```
ggplot(univ_complete_orig) +
 aes(x = ft_undergrad, fill = public1_private2) +
 geom_histogram(bins = 20) +
 scale_fill_brewer(palette = "Accent") +
 labs(title = "Universities by Size (FT Students)") +
 theme_minimal() +
 facet_grid(vars(cluster), vars(), scales = "free") +
 theme(legend.position = "bottom")
```



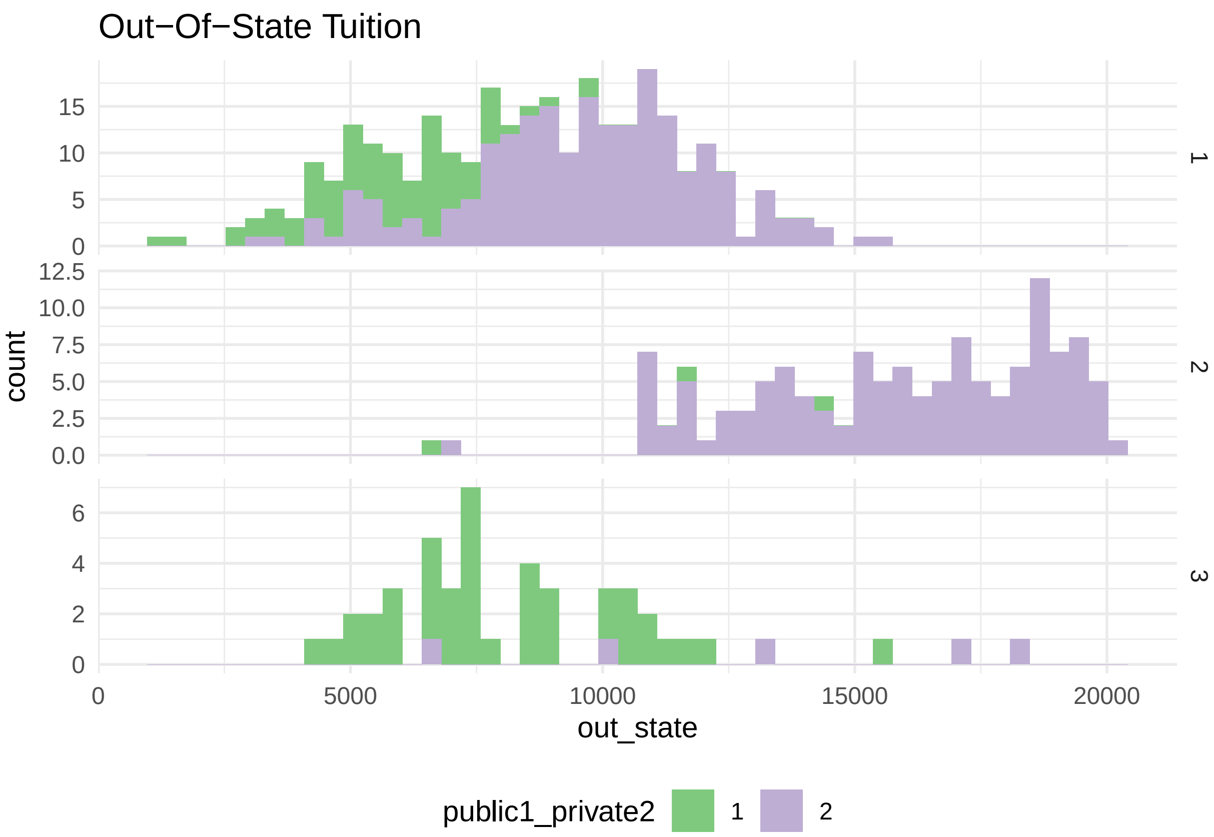
```r
ggplot(univ_complete_orig) +
 aes(x = pt_undergrad, fill = public1_private2) +
 geom_histogram(bins = 50) +
 scale_fill_brewer(palette = "Accent") +
 labs(title = "Universities by Size (PT Students)") +
 theme_minimal() +
 facet_grid(vars(cluster), vars(), scales = "free") +
 theme(legend.position = "bottom")
```



```
ggplot(univ_complete_orig) +
 aes(x = in_state, fill = public1_private2) +
 geom_histogram(bins = 50) +
 scale_fill_brewer(palette = "Accent") +
 labs(title = "In-State Tuition") +
 theme_minimal() +
 facet_grid(vars(cluster), vars(), scales = "free") +
 theme(legend.position = "bottom")
```



```
ggplot(univ_complete_orig) +
 aes(x = out_state, fill = public1_private2) +
 geom_histogram(bins = 50) +
 scale_fill_brewer(palette = "Accent") +
 labs(title = "Out-Of-State Tuition") +
 theme_minimal() +
 facet_grid(vars(cluster), vars(), scales = "free") +
 theme(legend.position = "bottom")
```



SUMMARY I believe that **Cluster 1** represent public state schools because they are evenly spread across the country. Also, the cost of in-state tuition is significantly lower than the out-of-state tuition. The Private schools in this cluster are located in the North East, North West central, and South Atlantic regions. Due to the size of the private schools, the cost of tuition being higher, and being located mostly in the north/south regions they are probably mostly religious and liberal arts private schools.

I believe that **Cluster 2** are elite or prestigious universitiesbecause they are located mostly in the East North Central, New England, and Middle Atlantic regions with smaller numbers of FT undergraduates. They have a very high percent of PhD faculty, have basically no PT undergraduates, and have a very high tuition cost for both in-state and out-of-state.

I believe **Cluster 3** are mostly large state schools spread fairly evenly accross the country, have a lower in-state tuition, higher percent of PhD faculty, high amount of FT undergraduates, and high PT undergraduates.

Possible Additional External Information

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

Part 3: Tufts University

1. Separate Tufts information into df.

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

```
## # A tibble: 1 x 21
##   college_name state public1_private2 appli_accepted accept_rate appli_recd
##   <chr>        <chr> <fct>                     <int>       <dbl>      <int>
## 1 Tufts Unive~ MA    2                          3605        47.3       7614
## # ... with 15 more variables: new_stud <int>, new_stud_10 <dbl>,
## #   new_stud_25 <dbl>, ft_undergrad <int>, pt_undergrad <int>, 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 21
##   college_name state public1_private2 appli_accepted accept_rate appli_recd
##   <chr>        <chr> <fct>                     <dbl>       <dbl>      <dbl>
## 1 Tufts Unive~ MA    2                         0.771       -1.76       1.37
## # ... with 15 more variables: new_stud <dbl>, 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[, 4:21])
get_dist(tufts_dist, method = "euclidean")
```

```
##           1        2        3
## 2  4.032770
## 3  6.176519 6.625517
## 11 6.608466 2.728890 6.946319
```

Tufts is closest to cluster 2, at a distance of 2.73. Tufts University should be included in cluster 2. This means that the Kmeans algorithm is predicting that Tufts University is a Ivy League school. According to US News & World report it confirms that it is an elite university ranking as #30 in the nation.

Citation: https://www.usnews.com/best-colleges/tufts-university-2219

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

```
##   mean(pt_undergrad)
## 1           276.0156
```

This is the value that should be imputed into the PT undergrad column in the Tufts University df. Meaning that they have an average of 276 PT undergraduates. The 2019-2020 Tufts University had at total of 165 PT undergraduates vs. 5643 FT undergraduates. Although the mean of cluster 2 PT undergrad value isn't "exact" what it does successfully communicate is that this university has a very lower number of PT undergraduates.

Citation https://provost.tufts.edu/institutionalresearch/about-tufts/common-data-set/

Imputing Missing Value

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

tufts$pt_undergrad <- as.double(tufts$pt_undergrad)
tufts[is.na(tufts$pt_undergrad), "pt_undergrad"] <- mean
tufts_demo <- rbind(tufts_original, tufts)
tufts_demo %>%
  select(pt_undergrad)
```

```
## # A tibble: 2 x 1
##   pt_undergrad
##          <dbl>
## 1           NA
## 2         276.
```

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