# Project 1

Shannon Owings

3/4/2020

## 0. Introduction

The datasets that I have chosen for this project are entitled "SchoolRankings.csv" and "salaries-by-college-type.csv". The "SchoolRankings"" dataset describes America's top 150 universities in 2019, pulled from Niche.com. This dataset gives variables such as the name of the school, acceptance rate, location, average cost after financial aid, and the 25th to 75th percentile score on SAT for accepted students. The other dataset, "salaries-by-college-type", describes the salaries of five different types of schools: Engineering, State, Liberal Arts, Party, and Ivy League. This data was obtained from the Wall Street Journal. The variables for this dataset include school name, school type, starting median salary, mid-career median salary, mid-career 10th percentile salary, mid-career 25th percentile salary, mid-career 75th percentile salary, and mid-career 90th percentile salary. I acquired these datasets through the Kaggle website by searching the word "college". The subject of college and these two datasets in particular are interesting to me because I am a college student and it is amusing to compare my college to 50 of America's top institutions. It is also interesting to compare the types of schools in terms of acceptance rate, cost, salaries, and more. I liked these datasets because I am very interested in this information currently and can personally relate to the data as a college student. I expect to see potential associations between acceptance rate and price as well as starting median salary and mid-career median salary. I believe that as acceptance rate decreases, cost of attendance will increase. Likewise, I think that as starting median salary increases, mid-career median salary will increase. It will be interesting to see if there are associations between these variables.

## 1. Tidying: Rearranging Wide/Long

```
library(tidyverse)
library(readr)
SchoolRankings <- read_csv("/stor/home/smo884/Project 1/SchoolRankings.csv")
salaries_by_college_type <- read_csv("/stor/home/smo884/Project 1/salaries-by-college-type.csv")
glimpse(SchoolRankings)
```

```
## Observations: 150
## Variables: 5
## $ Institution <chr> "Massachusetts Institute of Technology", "Stanford Univ...
## $ `AR%`       <dbl> 7, 5, 5, 7, 6, 9, 7, 10, 8, 8, 16, 16, 19, 9, 8, 8, 14,...
## $ Location    <chr> "Cambridge, MA", "Stanford, CA", "Cambridge, MA", "New ...
## $ `Price$`    <dbl> 22230, 16562, 17030, 18053, 16302, 24539, 22824, 22011,...
## $ SAT_Range   <chr> "1490-1570", "1390-1540", "1460-1590", "1460-1580", "14...
```

```
glimpse(salaries_by_college_type)
```

```
## Observations: 269
## Variables: 8
```

```
## $ Institution                       <chr> "Massachusetts Institute of Techn...
## $ School_Type                       <chr> "Engineering", "Engineering", "En...
## $ Starting_Median_Salary            <dbl> 72200, 75500, 71800, 62400, 62200...
## $ `Mid-Career_Median_Salary`        <dbl> 126000, 123000, 122000, 114000, 1...
## $ `Mid-Career_10_Percentile_Salary` <dbl> 76800, NA, NA, 66800, NA, 80000, ...
## $ `Mid-Career_25_Percentile_Salary` <dbl> 99200, 104000, 96000, 94300, 8020...
## $ `Mid-Career_75_Percentile_Salary` <dbl> 168000, 161000, 180000, 143000, 1...
## $ `Mid-Career_90_Percentile_Salary` <dbl> 220000, NA, NA, 190000, NA, 18000...
```

```
untidysalaries <- salaries_by_college_type %>% pivot_wider(names_from = "School_Type",
    values_from = "Institution")
glimpse(untidysalaries)
```

```
## Observations: 249
## Variables: 11
## $ Starting_Median_Salary            <dbl> 72200, 75500, 71800, 62400, 62200...
## $ `Mid-Career_Median_Salary`        <dbl> 126000, 123000, 122000, 114000, 1...
## $ `Mid-Career_10_Percentile_Salary` <dbl> 76800, NA, NA, 66800, NA, 80000, ...
## $ `Mid-Career_25_Percentile_Salary` <dbl> 99200, 104000, 96000, 94300, 8020...
## $ `Mid-Career_75_Percentile_Salary` <dbl> 168000, 161000, 180000, 143000, 1...
## $ `Mid-Career_90_Percentile_Salary` <dbl> 220000, NA, NA, 190000, NA, 18000...
## $ Engineering                       <chr> "Massachusetts Institute of Techn...
## $ Party                             <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Liberal Arts`                    <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Ivy League`                      <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ State                             <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
```

```
tidysalaries <- untidysalaries %>% pivot_longer(7:11, names_to = "School_Type",
    values_to = "Institution") %>% na.omit
glimpse(tidysalaries)
```

```
## Observations: 231
## Variables: 8
## $ Starting_Median_Salary            <dbl> 72200, 62400, 61000, 61800, 61100...
## $ `Mid-Career_Median_Salary`        <dbl> 126000, 114000, 114000, 111000, 1...
## $ `Mid-Career_10_Percentile_Salary` <dbl> 76800, 66800, 80000, 63300, 71600...
## $ `Mid-Career_25_Percentile_Salary` <dbl> 99200, 94300, 91200, 80100, 85500...
## $ `Mid-Career_75_Percentile_Salary` <dbl> 168000, 143000, 137000, 150000, 1...
## $ `Mid-Career_90_Percentile_Salary` <dbl> 220000, 190000, 180000, 209000, 1...
## $ School_Type                       <chr> "Engineering", "Engineering", "En...
## $ Institution                       <chr> "Massachusetts Institute of Techn...
```

To begin, I imported and glimpsed at my data. I noticed that besides a handful of NAs, my data was pretty tidy. Because of this, I decided to untidy my salaries dataset by using pivot_wider and naming this new dataset "untidysalaries". This changed my data from long to wide by removing rows and adding columns. A separate column for each school type (Engineering, Party, Liberal Arts, Ivy League, and State) was made and several new NAs were introduced. I then used pivot_longer to undo pivot_wider. This action removed the newly made columns and created the original rows again, with the five different school types falling under the column of "School Type" rather than each having a column of their own. This dataset was named "tidysalaries". These actions were completed in order to demonstrate my tidying skills.

## 2. Joining/Merging

```
library(dplyr)
fulldata <- inner_join(salaries_by_college_type, SchoolRankings)
head(fulldata)
```

```
## # A tibble: 6 x 12
##   Institution School_Type Starting_Median~ `Mid-Career_Med~ `Mid-Career_10_~
##   <chr>       <chr>                  <dbl>            <dbl>            <dbl>
## 1 Harvey Mud~ Engineering            71800           122000               NA
## 2 Georgia In~ Engineering            58300           106000            67200
## 3 Colorado S~ Engineering            58100           106000            62200
## 4 Stevens In~ Engineering            60600           105000            68700
## 5 Bucknell U~ Liberal Ar~            54100           110000            62800
## 6 Colgate Un~ Liberal Ar~            52800           108000            60000
## # ... with 7 more variables: `Mid-Career_25_Percentile_Salary` <dbl>,
## #   `Mid-Career_75_Percentile_Salary` <dbl>,
## #   `Mid-Career_90_Percentile_Salary` <dbl>, `AR%` <dbl>, Location <chr>,
## #   `Price$` <dbl>, SAT_Range <chr>
```

I performed an inner_join because I only wanted to look at the colleges that were in both datasets. I did this in order to keep my dataset tidy and leave all unmatched rows out. By performing this join, 100 observations were dropped from the "SchoolRankings" dataset and 219 observations were dropped from the "salaries_by_college_type" dataset. The datasets were joined by Institution, so any institution that was in one dataset but not the other was dropped. This left 50 institutions that were common to both datasets and a total of 12 columns. Because each dataset was reduced significantly, there could be potential problems with the remaining 50 institutions. It is no longer as representative of the original data. There are 30 Liberal Arts schools, 8 Ivy Leagues, 8 State schools, and 4 Engineering schools. This could pose potential problems with summary statistics or other analyses because some schools are so underrepresented.

## 3. Wrangling

```
fulldata %>% filter(`Mid-Career_10_Percentile_Salary` != "N/A") %>%
    filter(`Mid-Career_90_Percentile_Salary` != "N/A")
```

```
## # A tibble: 25 x 12
##    Institution School_Type Starting_Median~ `Mid-Career_Med~ `Mid-Career_10_~
##    <chr>       <chr>                  <dbl>            <dbl>            <dbl>
##  1 Georgia In~ Engineering            58300           106000            67200
##  2 Colorado S~ Engineering            58100           106000            62200
##  3 Stevens In~ Engineering            60600           105000            68700
##  4 Bucknell U~ Liberal Ar~            54100           110000            62800
##  5 Colgate Un~ Liberal Ar~            52800           108000            60000
##  6 Lafayette ~ Liberal Ar~            53900           107000            70600
##  7 University~ Liberal Ar~            48600            94600            44500
##  8 St. Olaf C~ Liberal Ar~            45300            86200            41300
##  9 Smith Coll~ Liberal Ar~            44000            83900            45100
## 10 Dartmouth ~ Ivy League             58000           134000            63100
## # ... with 15 more rows, and 7 more variables:
## #   `Mid-Career_25_Percentile_Salary` <dbl>,
## #   `Mid-Career_75_Percentile_Salary` <dbl>,
## #   `Mid-Career_90_Percentile_Salary` <dbl>, `AR%` <dbl>, Location <chr>,
```

```
## #   `Price$` <dbl>, SAT_Range <chr>
```

```
fulldata %>% summarize(mean(Starting_Median_Salary)) %>% glimpse()
```

```
## Observations: 1
## Variables: 1
## $ `mean(Starting_Median_Salary)` <dbl> 50358
```

```
fulldata %>% summarize(mean(`Mid-Career_Median_Salary`)) %>%
    glimpse()
```

```
## Observations: 1
## Variables: 1
## $ `mean(\`Mid-Career_Median_Salary\`)` <dbl> 97900
```

```
fulldata %>% summarize(mean(`AR%`)) %>% glimpse()
```

```
## Observations: 1
## Variables: 1
## $ `mean(\`AR%\`)` <dbl> 33.06
```

```
fulldata %>% summarize(mean(`Price$`)) %>% glimpse()
```

```
## Observations: 1
## Variables: 1
## $ `mean(\`Price$\`)` <dbl> 24868.82
```

```
fulldata %>% summarize(median(`AR%`)) %>% glimpse()
```

```
## Observations: 1
## Variables: 1
## $ `median(\`AR%\`)` <dbl> 28.5
```

```
highAR <- fulldata %>% filter(`AR%` > 33.06) %>% glimpse()
```

```
## Observations: 20
## Variables: 12
## $ Institution                       <chr> "Colorado School of Mines", "Stev...
## $ School_Type                       <chr> "Engineering", "Engineering", "Li...
## $ Starting_Median_Salary            <dbl> 58100, 60600, 50200, 51900, 42400...
## $ `Mid-Career_Median_Salary`        <dbl> 106000, 105000, 106000, 105000, 9...
## $ `Mid-Career_10_Percentile_Salary` <dbl> 62200, 68700, NA, NA, NA, 41300, ...
## $ `Mid-Career_25_Percentile_Salary` <dbl> 87900, 81900, 65600, 54800, 57100...
## $ `Mid-Career_75_Percentile_Salary` <dbl> 142000, 138000, 143000, 157000, 1...
## $ `Mid-Career_90_Percentile_Salary` <dbl> 201000, 185000, NA, NA, NA, 18500...
## $ `AR%`                             <dbl> 56, 44, 40, 42, 51, 43, 46, 37, 3...
## $ Location                          <chr> "Golden, CO", "Hoboken, NJ", "Wor...
## $ `Price$`                          <dbl> 25472, 38469, 27005, 33738, 29784...
## $ SAT_Range                         <chr> "1310-1450", "1320-1470", "1230-1...
```

```
lowAR <- fulldata %>% filter(`AR%` < 33.06) %>% glimpse()
```

```
## Observations: 30
## Variables: 12
## $ Institution                       <chr> "Harvey Mudd College", "Georgia I...
## $ School_Type                       <chr> "Engineering", "Engineering", "Li...
## $ Starting_Median_Salary            <dbl> 71800, 58300, 54100, 52800, 54500...
## $ `Mid-Career_Median_Salary`        <dbl> 122000, 106000, 110000, 108000, 1...
## $ `Mid-Career_10_Percentile_Salary` <dbl> NA, 67200, 62800, 60000, NA, 7060...
```

```
## $ `Mid-Career_25_Percentile_Salary` <dbl> 96000, 85200, 80600, 76700, 84900...
## $ `Mid-Career_75_Percentile_Salary` <dbl> 180000, 137000, 156000, 167000, 1...
## $ `Mid-Career_90_Percentile_Salary` <dbl> NA, 183000, 251000, 265000, NA, 2...
## $ `AR%`                             <dbl> 15, 23, 31, 28, 13, 31, 14, 11, 2...
## $ Location                          <chr> "Claremont, CA", "Atlanta, GA", "...
## $ `Price$`                          <dbl> 38135, 15873, 37817, 22182, 19519...
## $ SAT_Range                         <chr> "1470-1570", "1090-1520", "1250-1...
```

```
fulldata %>% summarize(max(`Mid-Career_Median_Salary`)) %>% glimpse()
```

```
## Observations: 1
## Variables: 1
## $ `max(\`Mid-Career_Median_Salary\`)` <dbl> 134000
```

```
fulldata %>% summarize(max(`AR%`)) %>% glimpse()
```

```
## Observations: 1
## Variables: 1
## $ `max(\`AR%\`)` <dbl> 89
```

```
fulldata %>% filter(`Mid-Career_Median_Salary` == 134000) %>%
    glimpse()
```

```
## Observations: 1
## Variables: 12
## $ Institution                       <chr> "Dartmouth College"
## $ School_Type                       <chr> "Ivy League"
## $ Starting_Median_Salary            <dbl> 58000
## $ `Mid-Career_Median_Salary`        <dbl> 134000
## $ `Mid-Career_10_Percentile_Salary` <dbl> 63100
## $ `Mid-Career_25_Percentile_Salary` <dbl> 90200
## $ `Mid-Career_75_Percentile_Salary` <dbl> 234000
## $ `Mid-Career_90_Percentile_Salary` <dbl> 321000
## $ `AR%`                             <dbl> 10
## $ Location                          <chr> "Hanover, NH"
## $ `Price$`                          <dbl> 22303
## $ SAT_Range                         <chr> "1430-1560"
```

```
fulldata %>% filter(`AR%` == 89) %>% glimpse
```

```
## Observations: 1
## Variables: 12
## $ Institution                       <chr> "Iowa State University"
## $ School_Type                       <chr> "State"
## $ Starting_Median_Salary            <dbl> 45400
## $ `Mid-Career_Median_Salary`        <dbl> 84600
## $ `Mid-Career_10_Percentile_Salary` <dbl> 44400
## $ `Mid-Career_25_Percentile_Salary` <dbl> 60000
## $ `Mid-Career_75_Percentile_Salary` <dbl> 109000
## $ `Mid-Career_90_Percentile_Salary` <dbl> 147000
## $ `AR%`                             <dbl> 89
## $ Location                          <chr> "Ames, IA"
## $ `Price$`                          <dbl> 13949
## $ SAT_Range                         <chr> "1160-1410"
```

```
fulldata %>% summarize_all(function(x) sum(is.na(x)))
```

```
## # A tibble: 1 x 12
```

```
##   Institution School_Type Starting_Median~ `Mid-Career_Med~ `Mid-Career_10_~
##         <int>       <int>            <int>            <int>            <int>
## 1           0           0                0                0               25
## # ... with 7 more variables: `Mid-Career_25_Percentile_Salary` <int>,
## #   `Mid-Career_75_Percentile_Salary` <int>,
## #   `Mid-Career_90_Percentile_Salary` <int>, `AR%` <int>, Location <int>,
## #   `Price$` <int>, SAT_Range <int>
```

```
fulldata %>% group_by(Location) %>% group_by(School_Type) %>%
    summarize(sdprice = sd(`Price$`)) %>% glimpse()
```

```
## Observations: 4
## Variables: 2
## $ School_Type <chr> "Engineering", "Ivy League", "Liberal Arts", "State"
## $ sdprice     <dbl> 10907.573, 5099.009, 5873.948, 3577.301
```

```
byschooltype <- fulldata %>% dplyr::select(-c("Mid-Career_10_Percentile_Salary",
    "Mid-Career_25_Percentile_Salary", "Mid-Career_75_Percentile_Salary",
    "Mid-Career_90_Percentile_Salary")) %>% group_by(School_Type) %>%
    mutate(Salary_Rank = dense_rank(desc(`Mid-Career_Median_Salary`))) %>%
    arrange(School_Type)
newAR <- byschooltype %>% mutate(AR_cat = case_when(`AR%` > 33 ~
    "high", `AR%` == 33 ~ "high", `AR%` < 33 ~ "low"))
newAR %>% group_by(School_Type) %>% group_by(AR_cat) %>% glimpse()
```

```
## Observations: 50
## Variables: 10
## Groups: AR_cat [2]
## $ Institution                <chr> "Harvey Mudd College", "Georgia Institut...
## $ School_Type                <chr> "Engineering", "Engineering", "Engineeri...
## $ Starting_Median_Salary     <dbl> 71800, 58300, 58100, 60600, 58000, 66500...
## $ `Mid-Career_Median_Salary` <dbl> 122000, 106000, 106000, 105000, 134000, ...
## $ `AR%`                      <dbl> 15, 23, 56, 44, 10, 6, 7, 5, 9, 13, 8, 7...
## $ Location                   <chr> "Claremont, CA", "Atlanta, GA", "Golden,...
## $ `Price$`                   <dbl> 38135, 15873, 25472, 38469, 22303, 16302...
## $ SAT_Range                  <chr> "1470-1570", "1090-1520", "1310-1450", "...
## $ Salary_Rank                <int> 1, 2, 2, 3, 1, 2, 3, 4, 5, 6, 7, 8, 1, 2...
## $ AR_cat                     <chr> "low", "low", "high", "high", "low", "lo...
```

```
byschooltype %>% arrange(Salary_Rank) %>% glimpse()
```

```
## Observations: 50
## Variables: 9
## Groups: School_Type [4]
## $ Institution                <chr> "Harvey Mudd College", "Dartmouth Colleg...
## $ School_Type                <chr> "Engineering", "Ivy League", "Liberal Ar...
## $ Starting_Median_Salary     <dbl> 71800, 58000, 54100, 49700, 58300, 58100...
## $ `Mid-Career_Median_Salary` <dbl> 122000, 134000, 110000, 96100, 106000, 1...
## $ `AR%`                      <dbl> 15, 10, 31, 71, 23, 56, 6, 28, 57, 44, 7...
## $ Location                   <chr> "Claremont, CA", "Hanover, NH", "Lewisbu...
## $ `Price$`                   <dbl> 38135, 22303, 37817, 19554, 15873, 25472...
## $ SAT_Range                  <chr> "1470-1570", "1430-1560", "1250-1420", "...
## $ Salary_Rank                <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3...
```

```
byschooltype %>% summarize_all(n_distinct) %>% glimpse()
```

```
## Observations: 4
```

```
## Variables: 9
## $ School_Type                <chr> "Engineering", "Ivy League", "Liberal Ar...
## $ Institution                <int> 4, 8, 30, 8
## $ Starting_Median_Salary     <int> 4, 8, 29, 7
## $ `Mid-Career_Median_Salary` <int> 3, 8, 24, 8
## $ `AR%`                      <int> 4, 7, 27, 7
## $ Location                   <int> 4, 8, 29, 8
## $ `Price$`                   <int> 4, 8, 30, 8
## $ SAT_Range                  <int> 4, 8, 26, 7
## $ Salary_Rank                <int> 3, 8, 24, 8
```

```
summary(byschooltype)
```

```
##  Institution        School_Type        Starting_Median_Salary
##  Length:50          Length:50          Min.   :40500
##  Class :character   Class :character   1st Qu.:44850
##  Mode  :character   Mode  :character   Median :48500
##                                        Mean   :50358
##                                        3rd Qu.:54400
##                                        Max.   :71800
##  Mid-Career_Median_Salary      AR%          Location            Price$
##  Min.   : 74600           Min.   : 5.00   Length:50          Min.   :12117
##  1st Qu.: 84525           1st Qu.:15.00   Class :character   1st Qu.:20050
##  Median : 96300           Median :28.50   Mode  :character   Median :24105
##  Mean   : 97900           Mean   :33.06                      Mean   :24869
##  3rd Qu.:107000           3rd Qu.:45.50                      3rd Qu.:29204
##  Max.   :134000           Max.   :89.00                      Max.   :39794
##   SAT_Range          Salary_Rank
##  Length:50          Min.   : 1.00
##  Class :character   1st Qu.: 3.00
##  Mode  :character   Median : 6.50
##                     Mean   : 8.72
##                     3rd Qu.:13.75
##                     Max.   :24.00
```

```
byschooltype %>% summarize_if(is.numeric, sd, na.rm = T) %>%
    glimpse()
```

```
## Observations: 4
## Variables: 6
## $ School_Type                <chr> "Engineering", "Ivy League", "Liberal Ar...
## $ Starting_Median_Salary     <dbl> 6499.744, 3218.584, 4020.785, 2858.790
## $ `Mid-Career_Median_Salary` <dbl> 8180.261, 10412.047, 11162.591, 4860.335
## $ `AR%`                      <dbl> 18.841444, 2.531939, 13.749357, 13.819629
## $ `Price$`                   <dbl> 10907.573, 5099.009, 5873.948, 3577.301
## $ Salary_Rank                <dbl> 0.8164966, 2.4494897, 6.9467598, 2.4494897
```

```
summarystats <- byschooltype %>% summarize(mean_starting_salary = mean(Starting_Median_Salary),
    sd_starting_salary = sd(Starting_Median_Salary), cor_salaries = cor(Starting_Median_Salary,
        `Mid-Career_Median_Salary`), min_starting_salary = min(Starting_Median_Salary),
    max_starting_salary = max(Starting_Median_Salary))
summarystats2 <- byschooltype %>% summarize(meanAR = mean(`AR%`),
    sdAR = sd(`AR%`), minAR = min(`AR%`), maxAR = max(`AR%`),
    cor_ARPrice = cor(`AR%`, `Price$`))
summarystats3 <- byschooltype %>% summarize(meanprice = mean(`Price$`),
    sdprice = sd(`Price$`), minprice = min(`Price$`), maxprice = max(`Price$`))
```

Table 1: Summary Stats for School Types

| School_Type | mean_starting_salary | sd_starting_salary | cor_salaries | min_starting_salary | max_starting_sala |
|---|---|---|---|---|---|
| **Engineering** | | | | | |
| Engineering | 62200.00 | 6499.744 | 0.9729871 | 58100 | 718 |
| **Ivy League** | | | | | |
| Ivy League | 60475.00 | 3218.584 | 0.4170142 | 56200 | 665 |
| **Liberal Arts** | | | | | |
| Liberal Arts | 47053.33 | 4020.785 | 0.8410719 | 40500 | 545 |
| **State** | | | | | |
| State | 46712.50 | 2858.790 | 0.8534101 | 42800 | 514 |

```
newsummary <- full_join(summarystats, summarystats2)
totalnewsummary <- full_join(newsummary, summarystats3) %>% glimpse()
```

```
## Observations: 4
## Variables: 15
## $ School_Type          <chr> "Engineering", "Ivy League", "Liberal Arts", "...
## $ mean_starting_salary <dbl> 62200.00, 60475.00, 47053.33, 46712.50
## $ sd_starting_salary   <dbl> 6499.744, 3218.584, 4020.785, 2858.790
## $ cor_salaries         <dbl> 0.9729871, 0.4170142, 0.8410719, 0.8534101
## $ min_starting_salary  <dbl> 58100, 56200, 40500, 42800
## $ max_starting_salary  <dbl> 71800, 66500, 54500, 51400
## $ meanAR               <dbl> 34.500, 8.125, 29.700, 69.875
## $ sdAR                 <dbl> 18.841444, 2.531939, 13.749357, 13.819629
## $ minAR                <dbl> 15, 5, 8, 47
## $ maxAR                <dbl> 56, 13, 68, 89
## $ cor_ARPrice          <dbl> -0.0211964, 0.8802455, 0.6650389, 0.1546587
## $ meanprice            <dbl> 29487.25, 22268.88, 26984.97, 17224.00
## $ sdprice              <dbl> 10907.573, 5099.009, 5873.948, 3577.301
## $ minprice             <dbl> 15873, 16302, 18427, 12117
## $ maxprice             <dbl> 38469, 31449, 39794, 22613
```

```
install.packages("kableExtra")
library(knitr)
library(kableExtra)
kable(totalnewsummary[1:4, 1:15], caption = "Summary Stats for School Types") %>%
    kable_styling("striped", full_width = F) %>% pack_rows("Engineering",
    1, 1) %>% pack_rows("Liberal Arts", 3, 3) %>% pack_rows("Ivy League",
    2, 2) %>% pack_rows("State", 4, 4)
```

```
df <- fulldata %>% na.omit %>% select_if(is.numeric)
cor(df)
```

```
##                                Starting_Median_Salary Mid-Career_Median_Salary
## Starting_Median_Salary                      1.0000000                0.9017287
## Mid-Career_Median_Salary                    0.9017287                1.0000000
## Mid-Career_10_Percentile_Salary             0.7837253                0.7482854
## Mid-Career_25_Percentile_Salary             0.9110233                0.9031567
## Mid-Career_75_Percentile_Salary             0.7520029                0.9417837
## Mid-Career_90_Percentile_Salary             0.7016844                0.9032221
## AR%                                        -0.7721737               -0.8133105
## Price$                                      0.1898464                0.1475821
```

```
##                                Mid-Career_10_Percentile_Salary
## Starting_Median_Salary                               0.7837253
## Mid-Career_Median_Salary                             0.7482854
## Mid-Career_10_Percentile_Salary                      1.0000000
## Mid-Career_25_Percentile_Salary                      0.8885681
## Mid-Career_75_Percentile_Salary                      0.5613020
## Mid-Career_90_Percentile_Salary                      0.4938375
## AR%                                                 -0.5482747
## Price$                                               0.3275366
##                                Mid-Career_25_Percentile_Salary
## Starting_Median_Salary                               0.9110233
## Mid-Career_Median_Salary                             0.9031567
## Mid-Career_10_Percentile_Salary                      0.8885681
## Mid-Career_25_Percentile_Salary                      1.0000000
## Mid-Career_75_Percentile_Salary                      0.7434955
## Mid-Career_90_Percentile_Salary                      0.6766279
## AR%                                                 -0.6468835
## Price$                                               0.1745009
##                                Mid-Career_75_Percentile_Salary
## Starting_Median_Salary                               0.7520029
## Mid-Career_Median_Salary                             0.9417837
## Mid-Career_10_Percentile_Salary                      0.5613020
## Mid-Career_25_Percentile_Salary                      0.7434955
## Mid-Career_75_Percentile_Salary                      1.0000000
## Mid-Career_90_Percentile_Salary                      0.9590237
## AR%                                                 -0.8180790
## Price$                                               0.1288950
##                                Mid-Career_90_Percentile_Salary        AR%
## Starting_Median_Salary                               0.7016844 -0.7721737
## Mid-Career_Median_Salary                             0.9032221 -0.8133105
## Mid-Career_10_Percentile_Salary                      0.4938375 -0.5482747
## Mid-Career_25_Percentile_Salary                      0.6766279 -0.6468835
## Mid-Career_75_Percentile_Salary                      0.9590237 -0.8180790
## Mid-Career_90_Percentile_Salary                      1.0000000 -0.8176629
## AR%                                                 -0.8176629  1.0000000
## Price$                                               0.1369273 -0.2390450
##                                   Price$
## Starting_Median_Salary          0.1898464
## Mid-Career_Median_Salary        0.1475821
## Mid-Career_10_Percentile_Salary 0.3275366
## Mid-Career_25_Percentile_Salary 0.1745009
## Mid-Career_75_Percentile_Salary 0.1288950
## Mid-Career_90_Percentile_Salary 0.1369273
## AR%                            -0.2390450
## Price$                          1.0000000
```

I generated several summary statistics. First, I filtered out the NAs from the Mid-Career 10th Percentile Salary column as well as from the Mid-Career 90th Percentile Salary column. I did this because these were the only two columns with any NAs and by removing them, the dataset was cut in half. I then began finding the means of several numeric variables of the full dataset. I found the mean starting median salary of all institutions along with the mean mid-career median salary, acceptance rate, and cost of attendance. It was interesting to see that the mean starting median salary of 50 of America's top institutions was only 50,358 dollars. The mean mid-career median salary of these institutions, however, nearly doubled to 97,900 dollars. I also discovered that the mean acceptance rate of these schools was 33.06% and the mean price

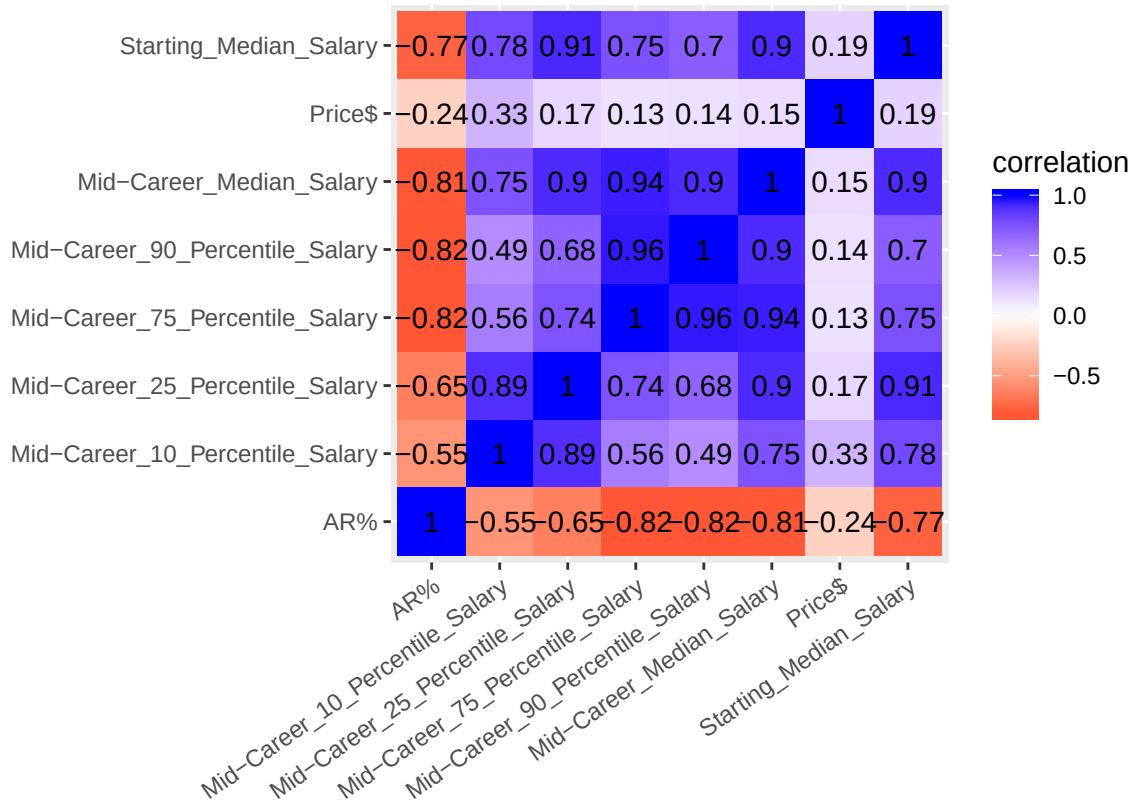
was 24,868.82 dollars. These numbers make sense because these top institutions have lower acceptance rates than other schools in the country. Next, I found the median acceptance rate, which was lower than the mean acceptance rate by 4.56%, and created datasets based on this mean acceptance rate. Any institution above the mean acceptance rate was grouped in a dataset called, "highAR". Likewise, any institution below the mean acceptance rate was grouped in a dataset called, "lowAR." There were 20 schools in the highAR dataset and 30 schools in the lowAR dataset. After finding the maximum mid-career median salary and acceptance rate, which were 134,000 dollars and 89% and belonged to Dartmouth College and Iowa State University, respectively, I calculated the total number of NAs in each column.

After computing summary statistics for the full dataset, I decided to use the group_by function. First, I grouped by the two categorial variables of "School Type" and "Location". I then created a new dataset where all of the salary columns were removed except for "Starting Median Salary" and "Mid-Career Median Salary", which I believed to be the two most useful salary columns. I then grouped this dataset by "School-Type" and created a column via the mutate function called "Salary Rank", which ranked each institution within its respective school type in terms of Mid-Career Median Salary. Harvey Mudd College was the highest ranked Engineering school, Dartmouth College was the highest ranked Ivy League school, Bucknell University was the highest ranked Liberal Arts school, and Texas A&M University was the highest ranked State school. I then created several other summary statistics for this data grouped by "School Type" including mean, standard deviation, minimum, and maximum of starting median salary, acceptance rate, and price. I also found the correlation between starting median salary and mid-career median salary as well as acceptance rate and price. All of these statistics were joined together and displayed in a table. It was interesting to see that the mean starting salary was higher for the engineering and Ivy League groups and lower for the Liberal Arts and State groups compared to the overall mean. Conversely, the mean acceptance rates for the Ivy League and Liberal Arts schools were lower than the overall mean whereas the Enginerring and State school mean acceptance rates were higher. I also found it fascinating that a Liberal Arts school had the maximum cost of attendance out of all 50 schools. As I predicted in the beginning, there was a positive correlation between mean starting median salary and mean mid-career median salary; however, there did not seem to be a correlation consistent across all four groups between acceptance rate and cost of attendance. Finally, I created a correlation matrix with my numeric variables.

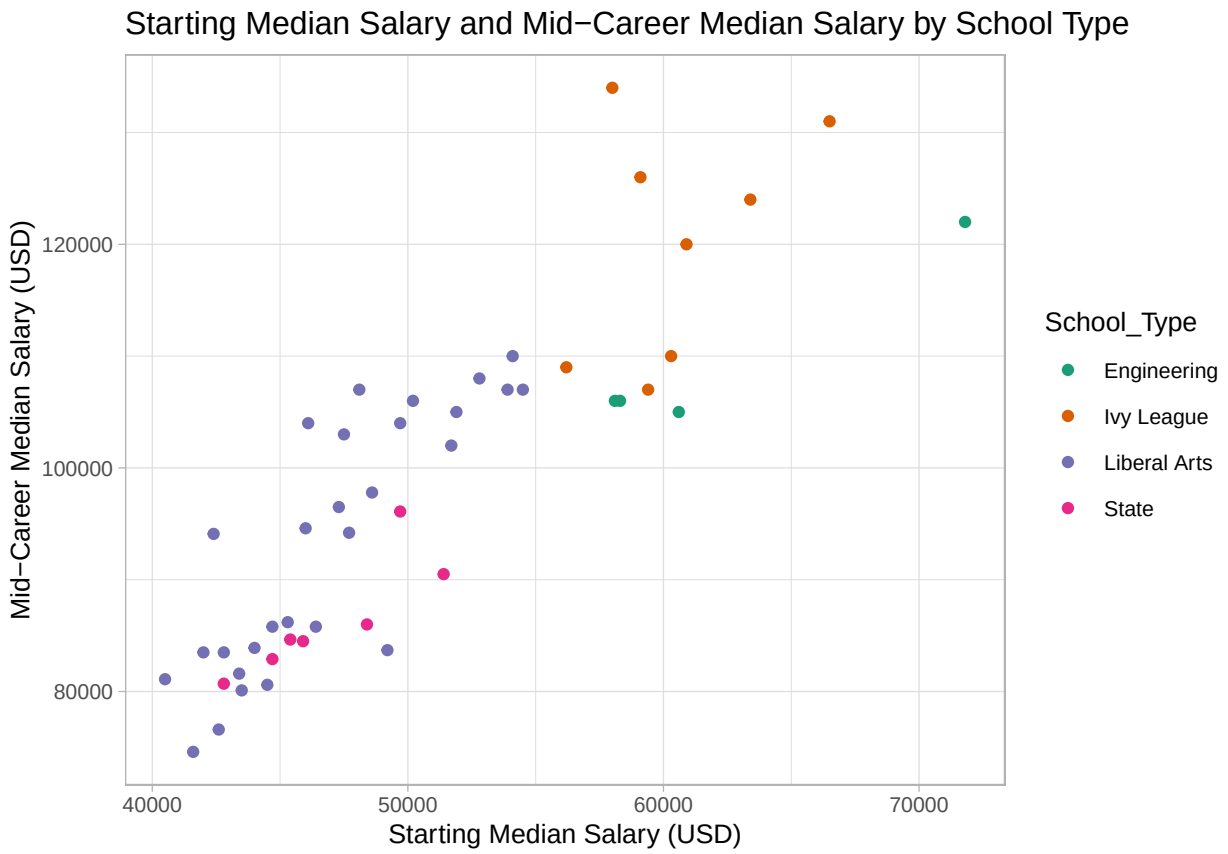
## 4. Visualizing

```
library(ggplot2)

tidycor <- cor(df) %>% as.data.frame %>% rownames_to_column %>%
    pivot_longer(-1, names_to = "name", values_to = "correlation")
tidycor %>% ggplot(aes(rowname, name, fill = correlation)) +
    geom_tile() + scale_fill_gradient2(low = "red", mid = "white",
    high = "blue") + geom_text(aes(label = round(correlation,
    2)), color = "black", size = 4) + coord_fixed() + xlab("") +
    ylab("") + theme(axis.text.x = element_text(angle = 35, hjust = 1))
```

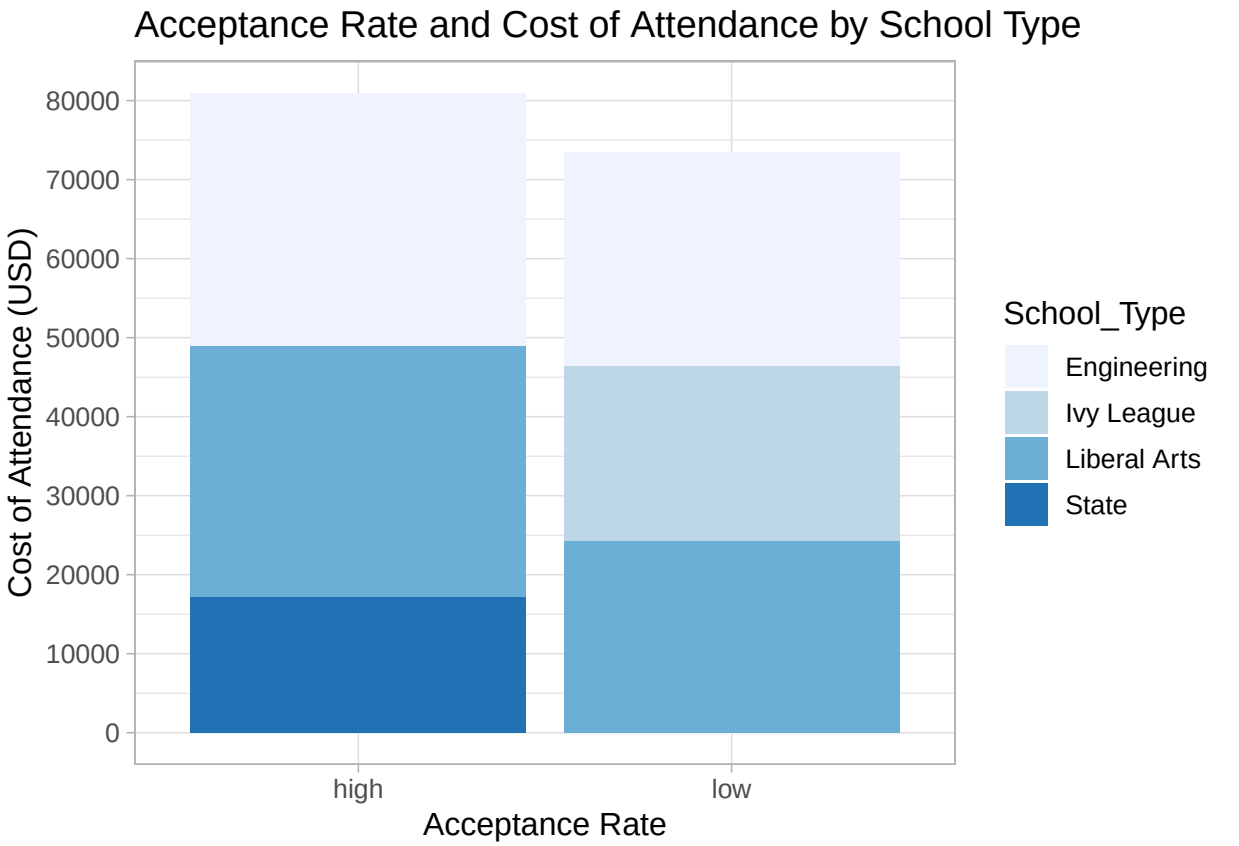



```
ggplot(fulldata, aes(Starting_Median_Salary, `Mid-Career_Median_Salary`)) +
    geom_point(aes(y = `Mid-Career_Median_Salary`, color = School_Type),
        stat = "summary", fun.y = "mean") + ggtitle("Starting Median Salary and Mid-Career Median Salary
    scale_y_continuous(name = "Mid-Career Median Salary (USD)") +
    xlab("Starting Median Salary (USD)") + theme_light(base_size = 10) +
    scale_color_brewer(palette = "Dark2")
```




```
newAR <- byschooltype %>% mutate(AR_cat = case_when(`AR%` > 33 ~
    "high", `AR%` == 33 ~ "high", `AR%` < 33 ~ "low"))

ggplot(newAR, aes(AR_cat, `Price$`)) + geom_bar(aes(y = `Price$`,
    fill = School_Type), stat = "summary", fun.y = "mean") +
    ggtitle("Acceptance Rate and Cost of Attendance by School Type") +
    xlab("Acceptance Rate") + ylab("Cost of Attendance (USD)") +
    scale_fill_brewer() + theme_light(base_size = 12) + scale_y_continuous(breaks = seq(0,
    80000, 10000))
```




In the correlation matrix, a correlation is given between every numeric variable. The strongest correlation is 0.96 between mid-career 90th percentile salary and mid-career 75th percentile salary. The weakest correlation is 0.13 between cost of attendance and mid-career 75th percentile salary. In the second plot, the correlation between starting median salary and mid-career median salary by school type is explored. As starting median salary increases, so does mid-career median salary, suggesting that there is a strong positive correlation between these two variables. The points are colored by school type and it is interesting to note that Ivy Leagues and Engineering schools have the highest salaries and Liberal Arts schools and State schools have lower salaries. This indicates that schools that have higher starting median salaries are also likely to have higher mid-career median salaries, although a few Liberal Arts schools surpassed Ivy League and Engineering schools in mid-career median salaries despite having a lower starting median salary.

In the second plot, the relationship between accceptance rate and cost of attendance was explored based on school type. To begin, I mutated acceptance rate from a numeric variable to a categorical variable by recoding all acceptance rates above the mean to “high”, indicating that these schools had a high acceptance rate, and recoded all acceptance rates below or equal to the mean to “low”, indicating that these schools had low acceptance rates. I also changed the number of tick marks on the y-axis to include more cost of attendance prices. After completing these actions, I noticed a few relationships in this plot. To begin with, no State schools were classified as having a low acceptance rate and no Ivy Leagues were classified as having a high acceptance rate, but the other two school types fell into both categories. It is also interesting that State schools appear to have the lowest cost of attendance while Engineering and Liberal Arts schools seem to have similar cost of attendance prices. Ivy Leagues appear to cost less than Engineering and Liberal Arts schools with low acceptance rates, which was surprising.

## Dimensionality Reduction

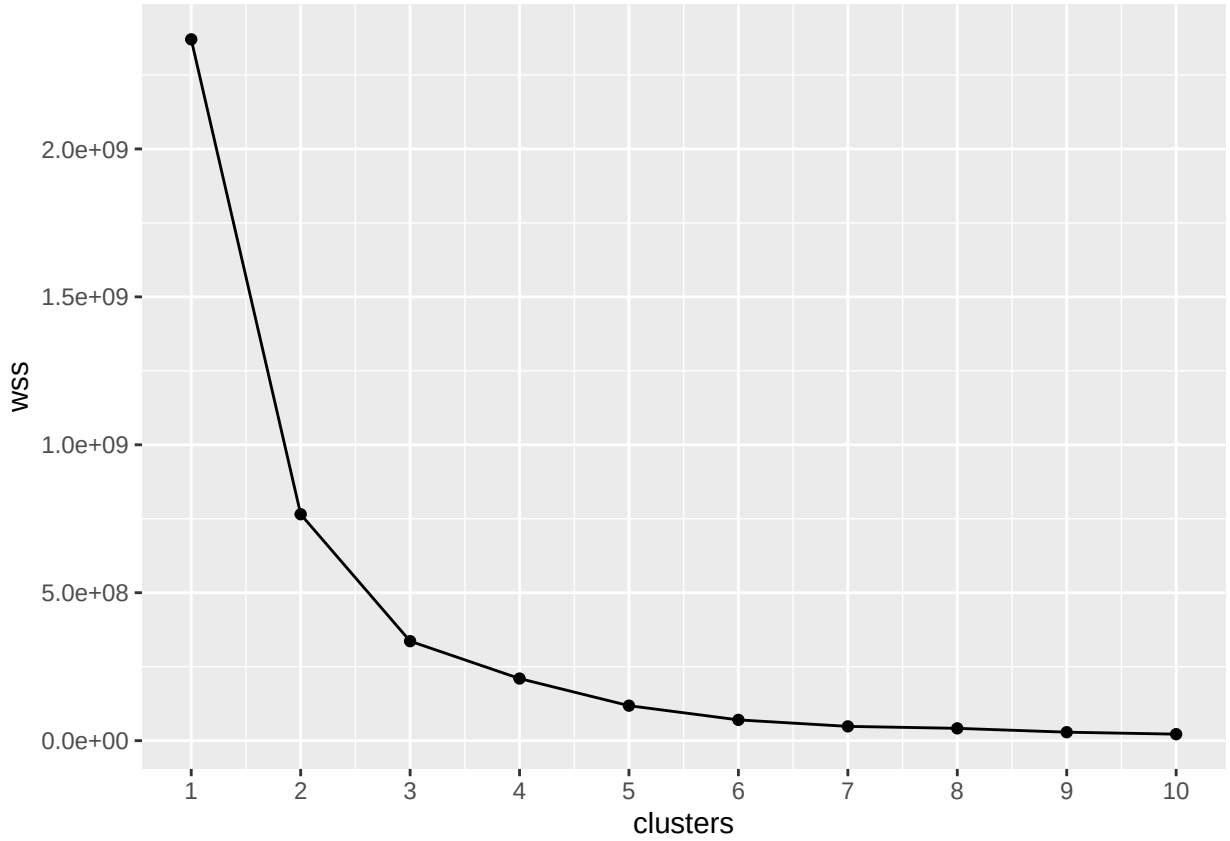
```
library("tidyverse")
install.packages("GGally")
library("GGally")
library(cluster)
library(dplyr)
library(ggplot2)

# K-means Clustering

fulldata2 <- fulldata %>% dplyr::select(-Institution, -School_Type,
    -Location, -`Mid-Career_10_Percentile_Salary`, -`Mid-Career_90_Percentile_Salary`)
fulldata2
```

```
## # A tibble: 50 x 7
##    Starting_Median~ `Mid-Career_Med~ `Mid-Career_25_~ `Mid-Career_75_~ `AR%`
##               <dbl>            <dbl>            <dbl>            <dbl> <dbl>
##  1            71800           122000            96000           180000    15
##  2            58300           106000            85200           137000    23
##  3            58100           106000            87900           142000    56
##  4            60600           105000            81900           138000    44
##  5            54100           110000            80600           156000    31
##  6            52800           108000            76700           167000    28
##  7            54500           107000            84900           162000    13
##  8            53900           107000            79300           144000    31
##  9            48100           107000            74600           146000    14
## 10            50200           106000            65600           143000    40
## # ... with 40 more rows, and 2 more variables: `Price$` <dbl>, SAT_Range <chr>
```

```
wss <- vector()
for (i in 1:10) {
    temp <- fulldata2 %>% select(`AR%`, `Price$`) %>% kmeans(i)
    wss[i] <- temp$tot.withinss
}
ggplot() + geom_point(aes(x = 1:10, y = wss)) + geom_path(aes(x = 1:10,
    y = wss)) + xlab("clusters") + scale_x_continuous(breaks = 1:10)
```




```
cluster1 <- fulldata2 %>% dplyr::select(`AR%`, `Price$`)
cluster2 <- fulldata2 %>% dplyr::select(Starting_Median_Salary,
    `Mid-Career_Median_Salary`)
cluster3 <- fulldata2 %>% dplyr::select(`Price$`, `Mid-Career_Median_Salary`)

kmeans1 <- cluster1 %>% scale %>% kmeans(2)
kmeans1
```

```
## K-means clustering with 2 clusters of sizes 20, 30
##
## Cluster means:
##          AR%     Price$
## 1  0.2994486  0.9656736
## 2 -0.1996324 -0.6437824
##
## Clustering vector:
##  [1] 1 2 1 1 1 2 2 2 2 1 1 2 1 1 2 2 2 2 2 2 1 1 2 1 1 2 2 1 1 1 1 1 1 2 2 2 2 2
## [39] 2 1 2 2 2 2 2 2 1 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 21.31712 42.60972
##  (between_SS / total_SS =  34.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans1$size
```

```
## [1] 20 30
```

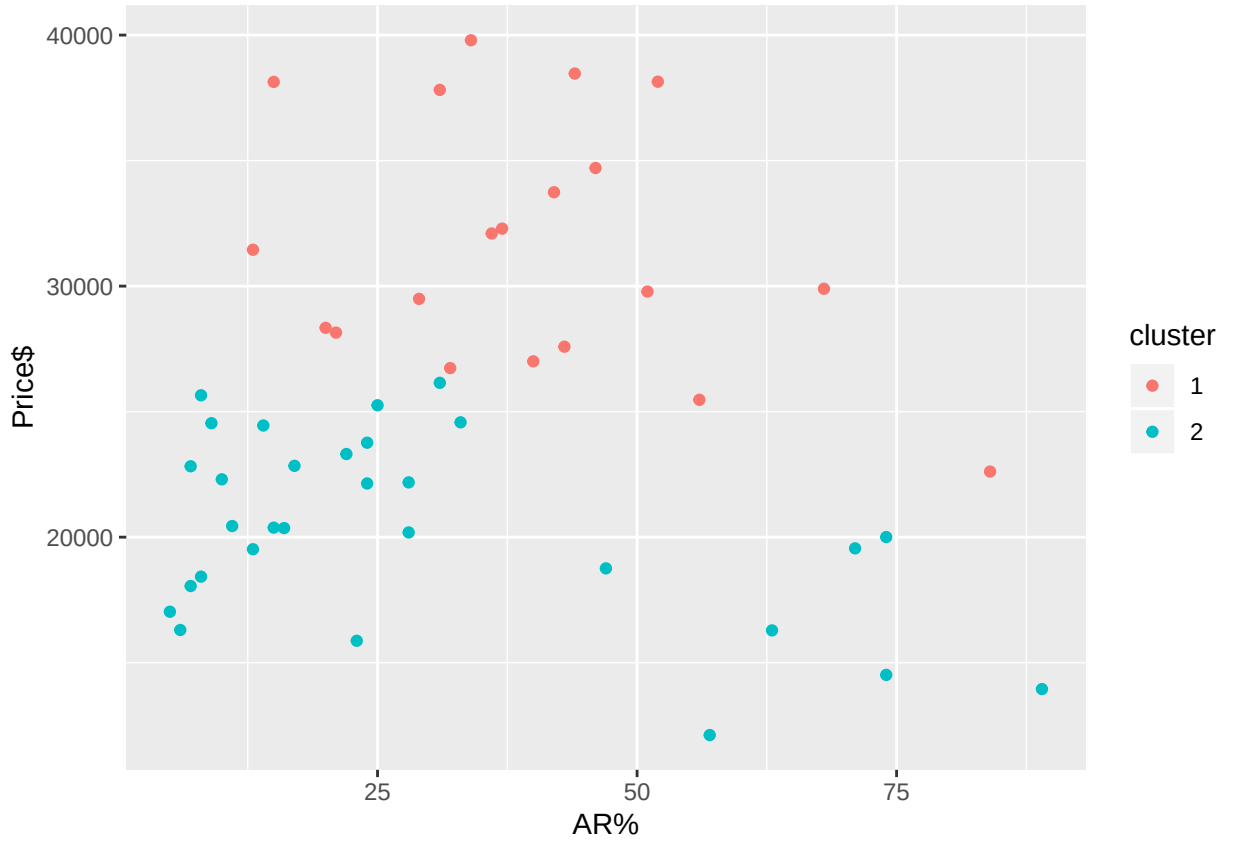
```
kmeans1$center
```

```
##          AR%     Price$
## 1  0.2994486  0.9656736
## 2 -0.1996324 -0.6437824
```

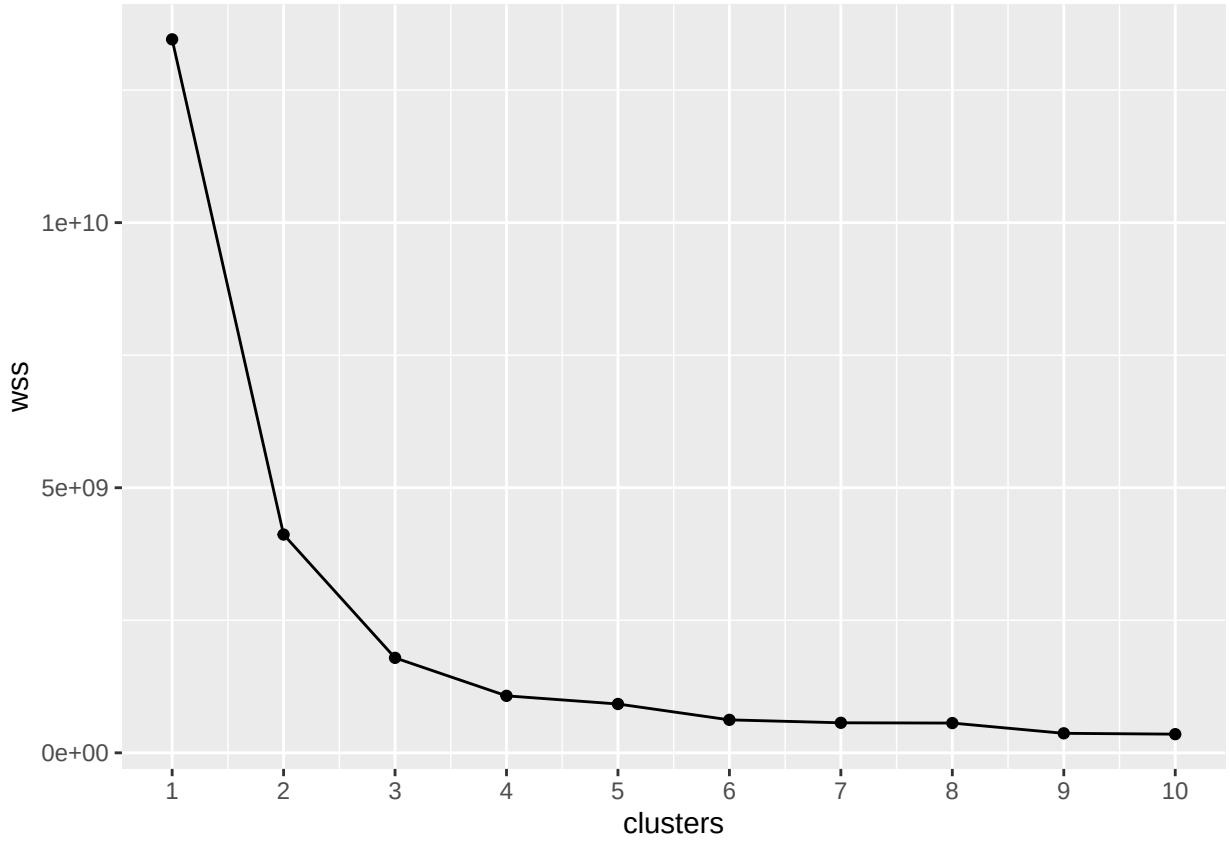
```
kmeans1$cluster
```

```
##  [1] 1 2 1 1 1 2 2 2 2 1 1 2 1 1 2 2 2 2 2 2 1 1 2 1 1 2 2 1 1 1 1 1 1 2 2 2 2 2
## [39] 2 1 2 2 2 2 2 1 2 2 2 2
```

```
kmeansclust <- fulldata2 %>% mutate(cluster = as.factor(kmeans1$cluster))
kmeansclust %>% ggplot(aes(`AR%`, `Price$`, color = cluster)) +
    geom_point()
```




```
wss2 <- vector()
for (i in 1:10) {
    temp <- fulldata2 %>% select(Starting_Median_Salary, `Mid-Career_Median_Salary`) %>%
        kmeans(i)
    wss[i] <- temp$tot.withinss
}
ggplot() + geom_point(aes(x = 1:10, y = wss)) + geom_path(aes(x = 1:10,
    y = wss)) + xlab("clusters") + scale_x_continuous(breaks = 1:10)
```




```
kmeans2 <- cluster2 %>% scale %>% kmeans(2)
kmeans2
```

```
## K-means clustering with 2 clusters of sizes 31, 19
##
## Cluster means:
##   Starting_Median_Salary Mid-Career_Median_Salary
## 1             -0.6493326                -0.615673
## 2              1.0594375                 1.004519
##
## Clustering vector:
##  [1] 2 2 2 2 2 2 2 2 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2
## [39] 2 2 2 2 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 14.64982 18.03105
##  (between_SS / total_SS =  66.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans2$size
```

```
## [1] 31 19
```

```
kmeans2$center
```

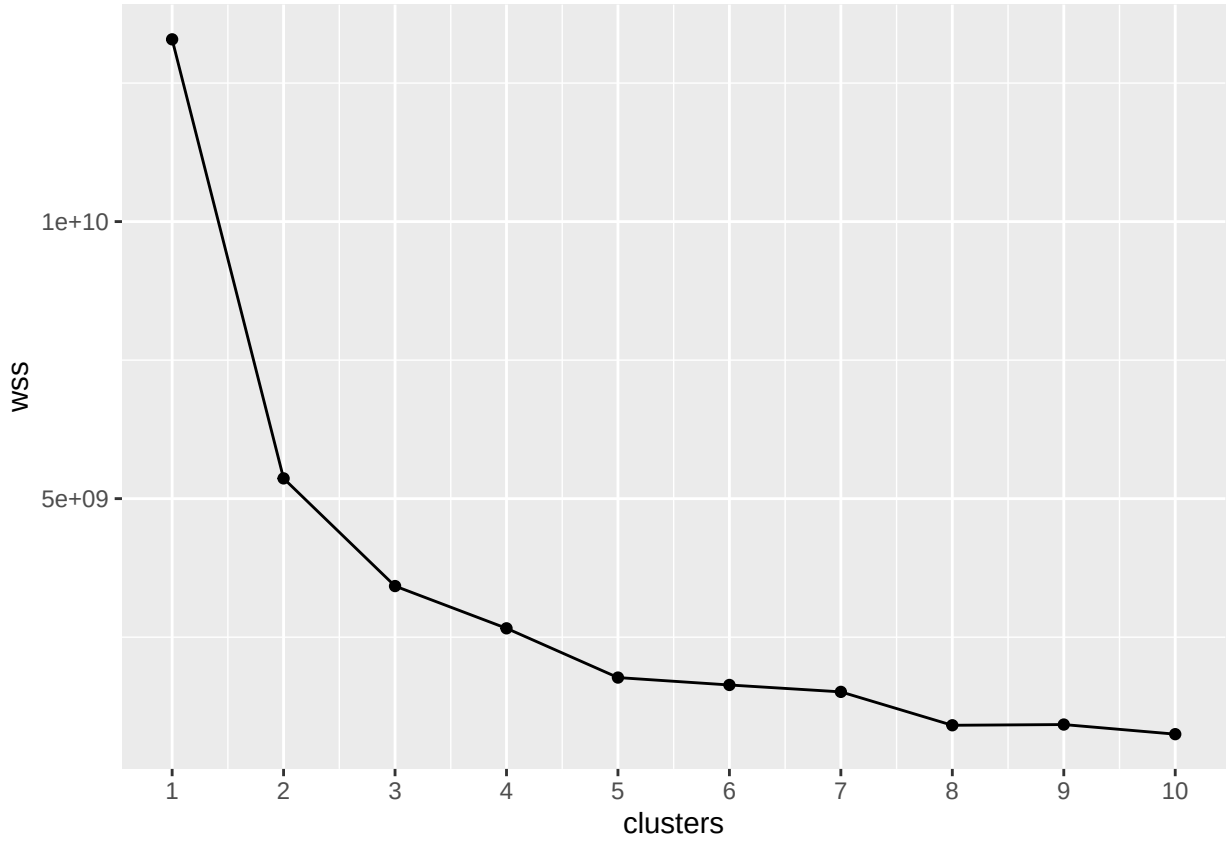
```
##   Starting_Median_Salary Mid-Career_Median_Salary
## 1              -0.6493326                -0.615673
## 2               1.0594375                 1.004519
```

```
kmeans2$cluster
```

```
##  [1] 2 2 2 2 2 2 2 2 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2
## [39] 2 2 2 2 1 1 1 1 1 1 1 1
```

```
kmeansclust2 <- fulldata2 %>% mutate(cluster = as.factor(kmeans1$cluster))
kmeansclust2 %>% ggplot(aes(Starting_Median_Salary, `Mid-Career_Median_Salary`,
    color = cluster)) + geom_point()
```


```
wss3 <- vector()
for (i in 1:10) {
    temp <- fulldata2 %>% select(`Price$`, `Mid-Career_Median_Salary`) %>%
        kmeans(i)
    wss[i] <- temp$tot.withinss
}
ggplot() + geom_point(aes(x = 1:10, y = wss)) + geom_path(aes(x = 1:10,
    y = wss)) + xlab("clusters") + scale_x_continuous(breaks = 1:10)
```




```
kmeans3 <- cluster3 %>% scale %>% kmeans(2)
kmeans3
```

```
## K-means clustering with 2 clusters of sizes 32, 18
##
## Cluster means:
##       Price$ Mid-Career_Median_Salary
## 1 -0.6031041                0.1990902
## 2  1.0721850               -0.3539381
##
## Clustering vector:
##  [1] 2 1 1 2 2 1 1 1 1 1 2 1 2 2 1 1 1 1 1 1 2 2 1 2 2 1 1 2 2 2 2 2 2 2 2 1 1 1 1
## [39] 1 2 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 39.51511 22.62965
##  (between_SS / total_SS =  36.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans3$size
```

```
## [1] 32 18
```

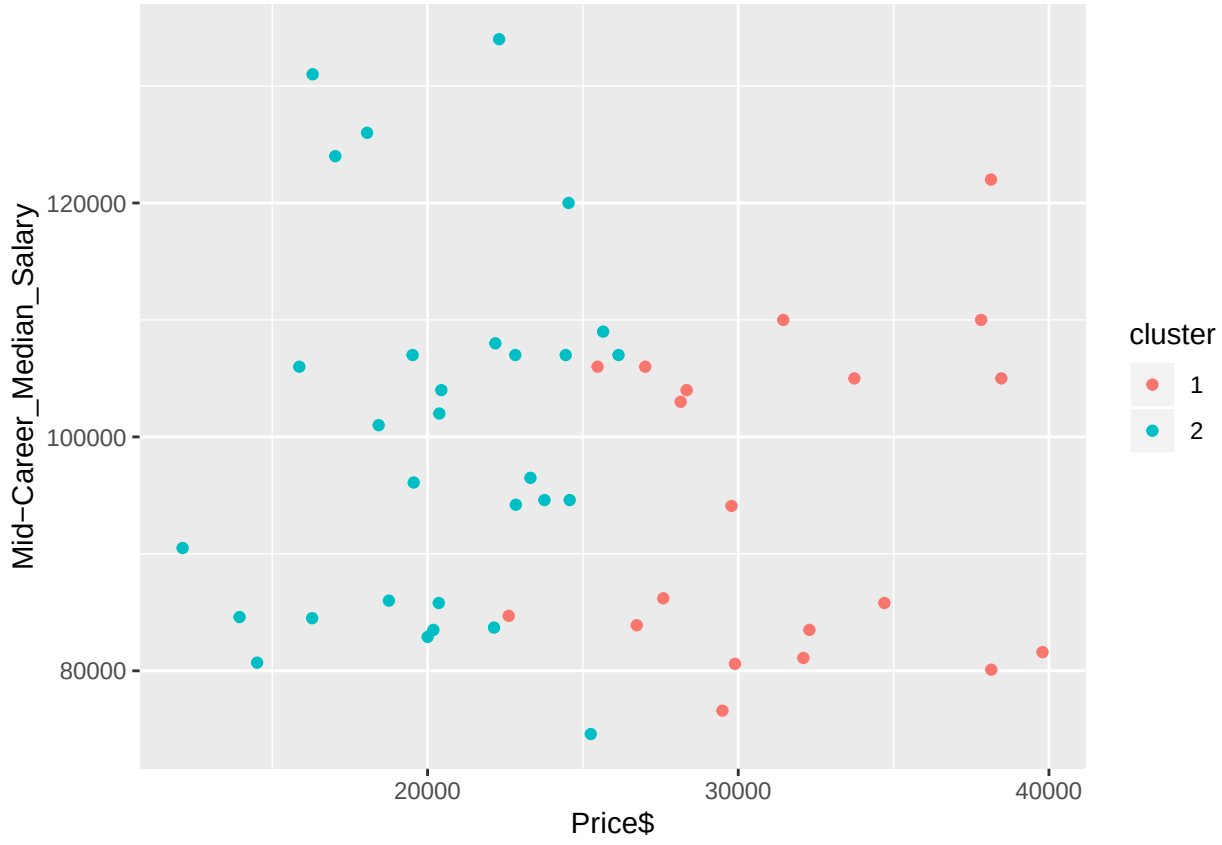
```
kmeans3$center
```

```
##       Price$ Mid-Career_Median_Salary
## 1 -0.6031041                0.1990902
## 2  1.0721850               -0.3539381
```

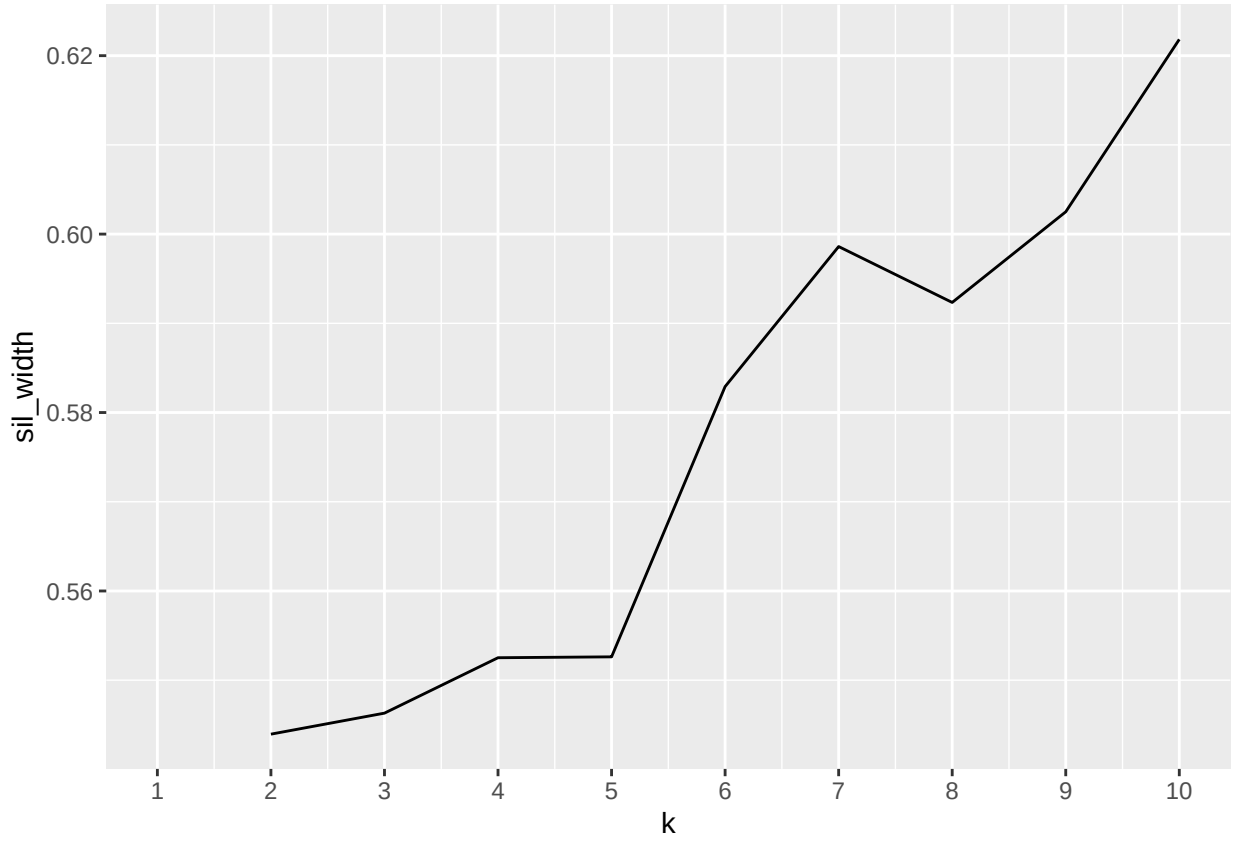
```
kmeans3$cluster
```

```
##  [1] 2 1 1 2 2 1 1 1 1 1 2 1 2 2 1 1 1 1 1 1 2 2 1 2 2 1 1 2 2 2 2 2 2 2 1 1 1 1
## [39] 1 2 1 1 1 1 1 1 1 1 1 1
```

```
kmeansclust3 <- fulldata2 %>% mutate(cluster = as.factor(kmeans1$cluster))
kmeansclust3 %>% ggplot(aes(`Price$`, `Mid-Career_Median_Salary`,
    color = cluster)) + geom_point()
```



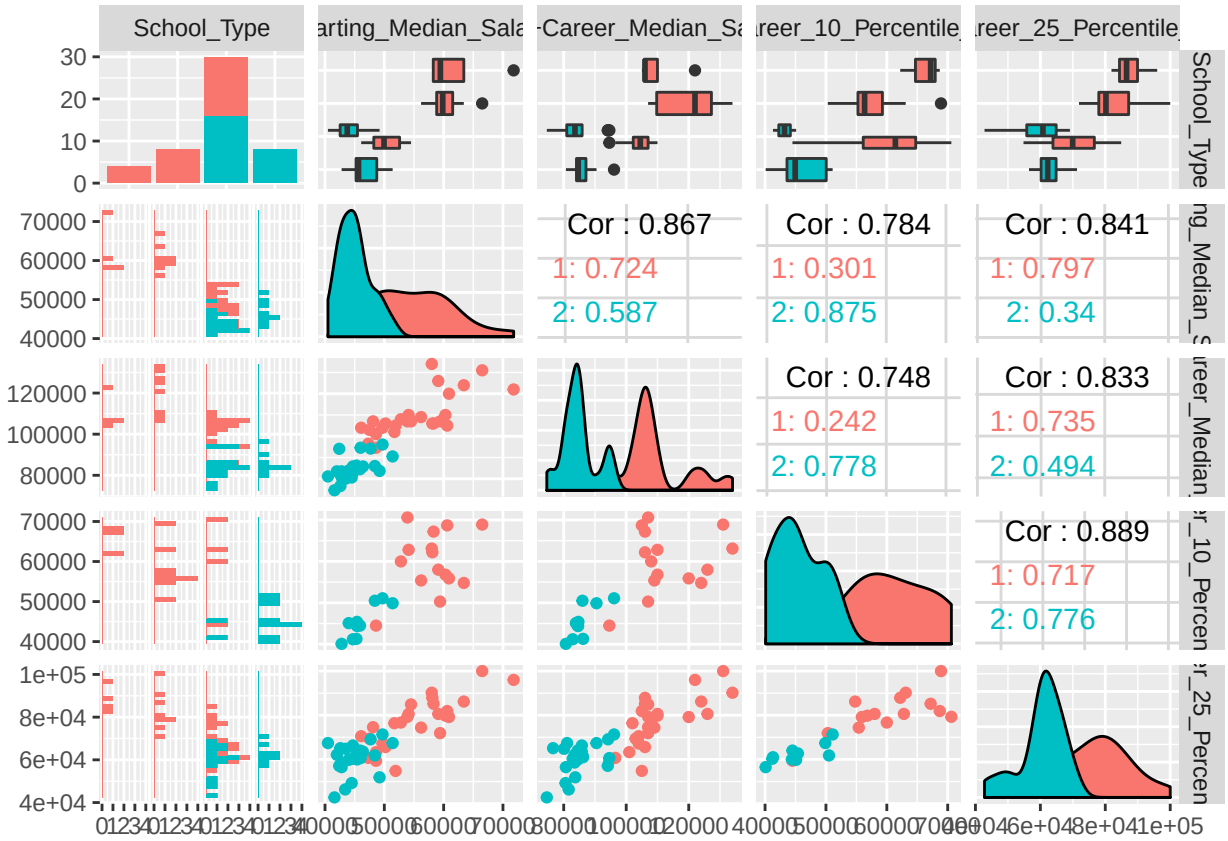

```
# PAM
pam_dat <- fulldata2 %>% select(`AR%`, `Price$`)
sil_width <- vector()
for (i in 2:10) {
    pam_fit <- pam(pam_dat, k = i)
    sil_width[i] <- pam_fit$silinfo$avg.width
}
ggplot() + geom_line(aes(x = 1:10, y = sil_width)) + scale_x_continuous(name = "k",
    breaks = 1:10)
```



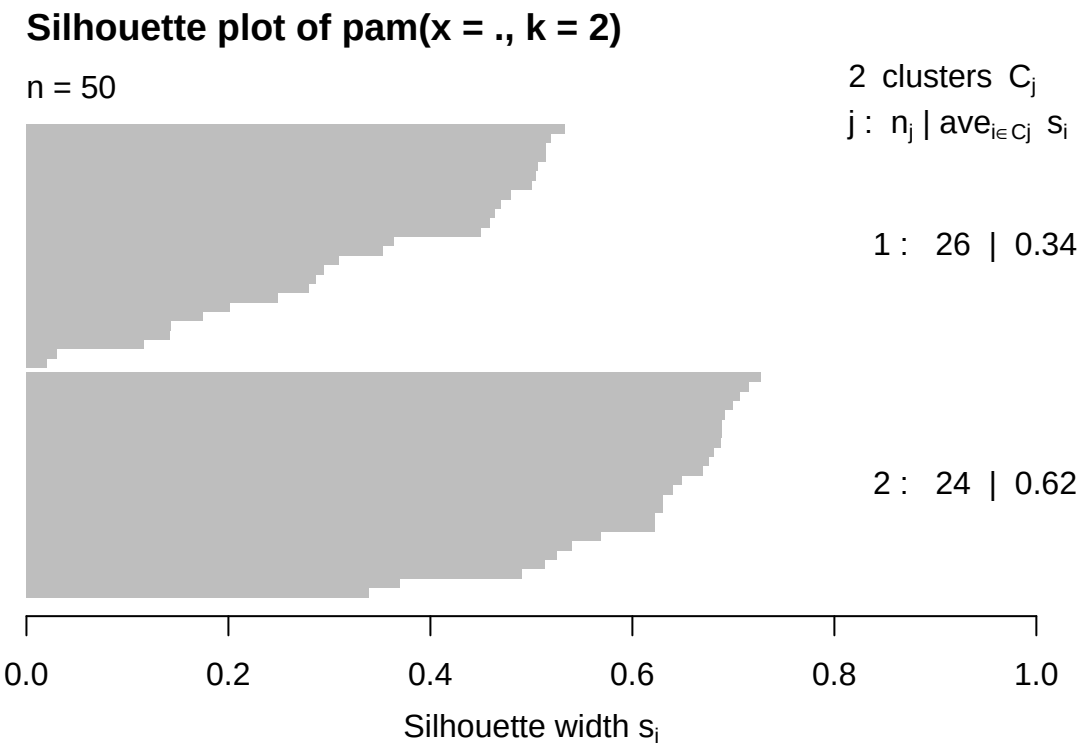

```
pam1 <- fulldata2 %>% pam(k = 2)
pam1
```

```
## Medoids:
##      ID Starting_Median_Salary Mid-Career_Median_Salary
## [1,] 41                  56200                   109000
## [2,] 22                  45300                    86200
##      Mid-Career_25_Percentile_Salary Mid-Career_75_Percentile_Salary AR% Price$
## [1,]                           74400                          159000   8  25651
## [2,]                           61000                          120000  43  27587
##      SAT_Range
## [1,]        NA
## [2,]        NA
## Clustering vector:
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1
## [39] 1 1 1 1 2 2 2 2 2 2 2 2
## Objective function:
##    build     swap
## 24265.74 20807.94
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

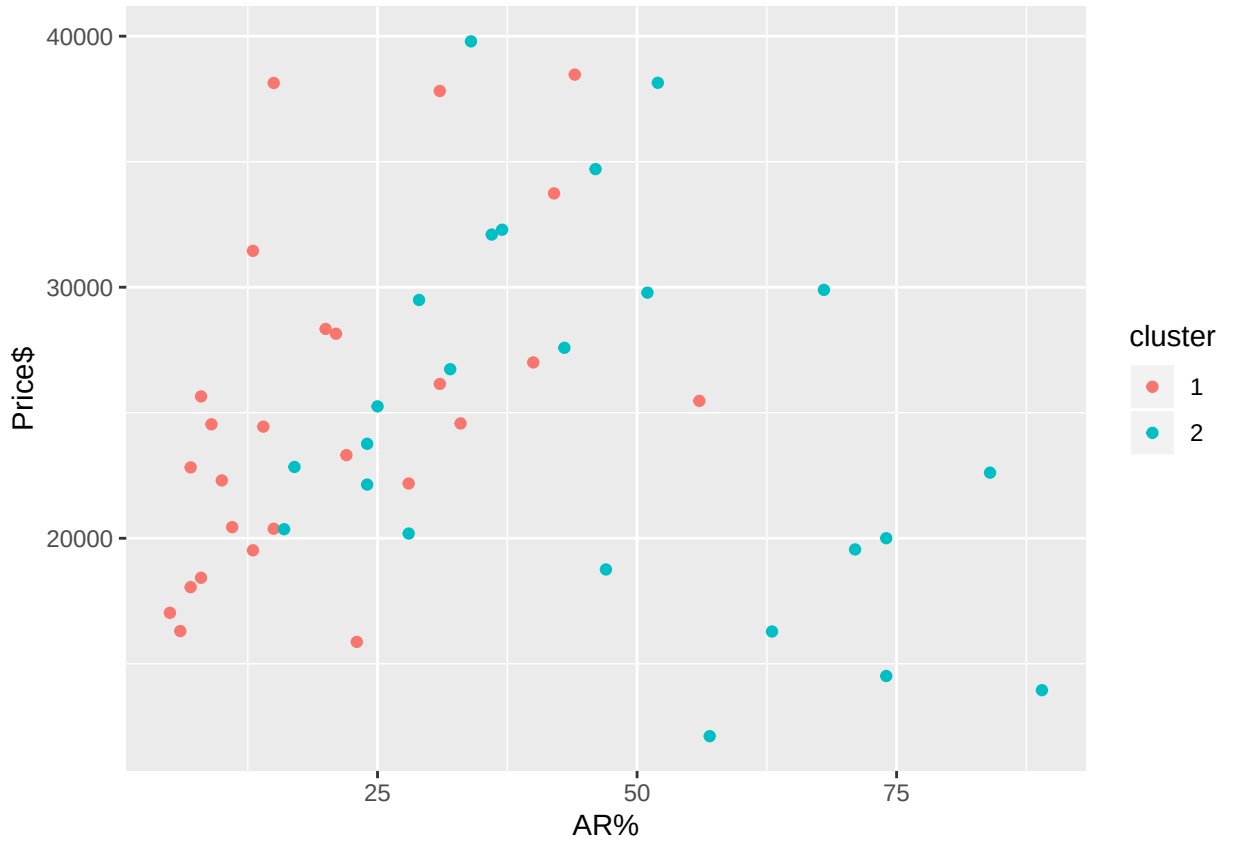
```
fulldata %>% mutate(cluster = as.factor(pam1$clustering)) %>%
    ggpairs(columns = 2:6, aes(color = cluster))
```




```
plot(pam1, which = 2)
```




```
pamclust <- fulldata2 %>% mutate(cluster = as.factor(pam1$clustering))
pamclust %>% ggplot(aes(`AR%`, `Price$`, color = cluster)) +
    geom_point()
```




```
pamclust %>% group_by(cluster) %>% summarize_if(is.numeric, mean,
    na.rm = T)
```

```
## # A tibble: 2 x 7
##   cluster Starting_Median~ `Mid-Career_Med~ `Mid-Career_25_~ `Mid-Career_75_~
##   <fct>              <dbl>            <dbl>            <dbl>            <dbl>
## 1 1                 55292.          109812.           76796.          162154.
## 2 2                 45012.           84996.           60408.          120625
## # ... with 2 more variables: `AR%` <dbl>, `Price$` <dbl>
```

```
fulldata[pam1$id.med, ]
```

```
## # A tibble: 2 x 12
##   Institution School_Type Starting_Median~ `Mid-Career_Med~ `Mid-Career_10_~
##   <chr>       <chr>                 <dbl>            <dbl>            <dbl>
## 1 Brown Univ~ Ivy League            56200           109000            55400
## 2 St. Olaf C~ Liberal Ar~           45300            86200            41300
## # ... with 7 more variables: `Mid-Career_25_Percentile_Salary` <dbl>,
## #   `Mid-Career_75_Percentile_Salary` <dbl>,
## #   `Mid-Career_90_Percentile_Salary` <dbl>, `AR%` <dbl>, Location <chr>,
## #   `Price$` <dbl>, SAT_Range <chr>
```

```
pam1$silinfo$avg.width
```

```
## [1] 0.4731273
```

For this dataset, I performed k-means clustering on my numeric variables. To begin, I created new data where I removed all of the categorical variables and then I plotted WSS against the number of clusters to

determine the best number of clusters to use. The WSS appeared to drop rapidly at 2, and so I decided that 2 clusters was best based on the plot. Next, I made my k-means clusters. The first k-means cluster included the variables of acceptance rate and cost of attendance. I scaled these variables and checked the k-means size, center, and cluster. After assigning each of the observations to the cluster whose center is closest and saving the cluster assignment as a column in the dataset, I graphed this first k-means and colored the data by final cluster assignment. I saw two distinct groups. The first cluster has an overall lower cost of attendance than the second cluster, however both clusters vary in acceptance rate. There does seem to be a general trend, however, that the lower the acceptance rate the higher cost of attendance. The size of the clusters was 30 and 20, with the lower cost of attendance cluster having 30 institutions and the higher cost of attendance cluster having 20 institutions.

Next, I performed a second k-means clustering on the numeric variables of starting median salary and mid-career median salary. I first determined how many clusters were appropriate. I decided to use 2 clusters, with one cluster having 19 institutions and the other cluster having 32 institutions. I then created a ggplot to visualize the clusters. These clusters appeared to be entertwined. These clusters did not map nicely at all to these variables and there is no distinction between the two clusters since they have similar starting median salaries and mid-career median salaries. There is a positive correlation between the two variables, however, in both clusters.

Finally, I performed a third k-means clustering on the numeric variables of cost of attendance and mid-career median salary. I determined how many clusters were appropriate for these numeric variables and once again decided on two clusters. I graphed these clusters via ggplot. The sizes of these clusters are 24 and 26. These clusters are separated once again based on cost of attendance. One cluster has a consistently higher cost of attendance while the other cluster has a consistently lower cost of attendance, however the distinction between mid-career median salary is not as clear. It is clear that the first and third k-means mapped much more nicely and showed the clusters much more distinctly than the second k-means plot. For the PAM method, I chose the number of clusters based on average silhouette width and erred on the side of fewer clusters. I chose 2 clusters to maximize the silhouette width. Again, I assigned observations to the cluster whose center is closest. I saved the cluster assignment and plotted it. I ran PAM and visualized it. Based on this, the mid-career 10th percentile salary and mid-career 25th percentile salary have the strongest correlation. Every correlation is positive. In terms of the average silhouette width, which is 0.43, the structure is considered weak and could be artifical. Next, I found the means for each variable and the final medoids, who were most representative of the cluster. Overall, I processed the data, chose the number of clusters, ran both k-means and PAM cluster analysis, and visualized the clusters. The PAM cluster did not map very nicely to the variables or show the clusters very distinctly. Likewise, the cluster solution was not very good, so I liked the k-means clustering plots better and was able to interpret them better.

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.4 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/libopenblasp-r0.2.20.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
##  [1] cluster_2.0.6    GGally_1.5.0     kableExtra_1.1.0 knitr_1.28
##  [5] forcats_0.4.0    stringr_1.4.0    dplyr_0.8.3      purrr_0.3.3
##  [9] readr_1.3.1      tidyr_1.0.0.9000 tibble_2.1.3     ggplot2_3.2.1
## [13] tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_0.2.5   xfun_0.13          reshape2_1.4.3     haven_2.2.0
##  [5] lattice_0.20-35    colorspace_1.4-1   vctrs_0.2.1        generics_0.0.2
##  [9] viridisLite_0.3.0  htmltools_0.3.6    yaml_2.2.0         utf8_1.1.4
## [13] rlang_0.4.2        pillar_1.4.2       glue_1.3.1         withr_2.1.2
## [17] DBI_1.0.0          RColorBrewer_1.1-2 dbplyr_1.4.2       modelr_0.1.5
## [21] readxl_1.3.1       plyr_1.8.4         lifecycle_0.1.0    munsell_0.5.0
## [25] gtable_0.3.0       cellranger_1.1.0   rvest_0.3.5        evaluate_0.14
## [29] labeling_0.3       fansi_0.4.0        broom_0.5.2        Rcpp_1.0.2
## [33] scales_1.0.0       backports_1.1.4    formatR_1.7        webshot_0.5.2
## [37] jsonlite_1.6       fs_1.3.1           hms_0.5.3          digest_0.6.20
## [41] stringi_1.4.3      grid_3.4.4         cli_1.1.0          tools_3.4.4
## [45] magrittr_1.5       lazyeval_0.2.2     crayon_1.3.4       pkgconfig_2.0.2
## [49] zeallot_0.1.0      xml2_1.2.2         reprex_0.3.0       lubridate_1.7.4
## [53] reshape_0.8.8      assertthat_0.2.1   rmarkdown_2.1      httr_1.4.1
## [57] rstudioapi_0.10    R6_2.4.0           nlme_3.1-131       compiler_3.4.4
```

```
## [1] "2020-05-08 01:03:19 CDT"
```

```
##                                      sysname
##                                      "Linux"
##                                      release
##                            "4.15.0-99-generic"
##                                      version
## "#100-Ubuntu SMP Wed Apr 22 20:32:56 UTC 2020"
##                                     nodename
##                    "educcomp01.ccbb.utexas.edu"
##                                      machine
##                                     "x86_64"
##                                        login
##                                    "unknown"
##                                         user
##                                     "smo884"
##                                effective_user
##                                     "smo884"
```