Rough Draft

Tyler Bontrager

2022-11-02

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# IMPORTING DATASETS
tuition_cost <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
```

```
## Rows: 2973 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): name, state, state_code, type, degree_length
## dbl (5): room_and_board, in_state_tuition, in_state_total, out_of_state_tuit...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tc = tuition_cost

tuition_income <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

```
## Rows: 209012 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): name, state, campus, income_lvl
## dbl (3): total_price, year, net_cost
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ti = tuition_income

salary_potential <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste
```

```
## Rows: 935 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): name, state_name
## dbl (5): rank, early_career_pay, mid_career_pay, make_world_better_percent, ...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sp = salary_potential

historical_tuition <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mas
```

```
## Rows: 270 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): type, year, tuition_type
## dbl (1): tuition_cost
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ht = historical_tuition

diversity_school <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste
```

```
## Rows: 50655 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): name, state, category
## dbl (2): total_enrollment, enrollment
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ds = diversity_school
```

```
tcFactored = tc %>%
  mutate(degFactor = as.factor(degree_length))
tcFactored
```

```
## # A tibble: 2,973 x 11
##    name      state state~1 type  degre~2 room_~3 in_st~4 in_st~5 out_o~6 out_o~7
##    <chr>     <chr> <chr>   <chr> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Aaniiih ~ Mont~ MT      Publ~ 2 Year       NA    2380    2380    2380    2380
##  2 Abilene ~ Texas TX      Priv~ 4 Year    10350   34850   45200   34850   45200
##  3 Abraham ~ Geor~ GA      Publ~ 2 Year     8474    4128   12602   12550   21024
##  4 Academy ~ Minn~ MN      For ~ 2 Year       NA   17661   17661   17661   17661
##  5 Academy ~ Cali~ CA      For ~ 4 Year    16648   27810   44458   27810   44458
##  6 Adams St~ Colo~ CO      Publ~ 4 Year     8782    9440   18222   20456   29238
##  7 Adelphi ~ New ~ NY      Priv~ 4 Year    16030   38660   54690   38660   54690
##  8 Adironda~ New ~ NY      Publ~ 2 Year    11660    5375   17035    9935   21595
##  9 Adrian C~ Mich~ MI      Priv~ 4 Year    11318   37087   48405   37087   48405
## 10 Advanced~ Virg~ VA      For ~ 2 Year       NA   13680   13680   13680   13680
## # ... with 2,963 more rows, 1 more variable: degFactor <fct>, and abbreviated
## #   variable names 1: state_code, 2: degree_length, 3: room_and_board,
## #   4: in_state_tuition, 5: in_state_total, 6: out_of_state_tuition,
## #   7: out_of_state_total
```

```
gatheredtc = tcFactored %>%
  filter(type!='Other') %>%
  gather(key="in_out", value="totalCost",c(in_state_total,out_of_state_total))
```

```
gatheredtc
```

```
## # A tibble: 5,944 x 11
##    name       state state~1 type  degre~2 room_~3 in_st~4 out_o~5 degFa~6 in_out
##    <chr>      <chr> <chr>   <chr> <chr>     <dbl>   <dbl>   <dbl> <fct>   <chr>
##  1 Aaniiih N~ Mont~ MT      Publ~ 2 Year       NA    2380    2380 2 Year  in_st~
##  2 Abilene C~ Texas TX      Priv~ 4 Year    10350   34850   34850 4 Year  in_st~
##  3 Abraham B~ Geor~ GA      Publ~ 2 Year     8474    4128   12550 2 Year  in_st~
##  4 Academy C~ Minn~ MN      For ~ 2 Year       NA   17661   17661 2 Year  in_st~
##  5 Academy o~ Cali~ CA      For ~ 4 Year    16648   27810   27810 4 Year  in_st~
##  6 Adams Sta~ Colo~ CO      Publ~ 4 Year     8782    9440   20456 4 Year  in_st~
##  7 Adelphi U~ New ~ NY      Priv~ 4 Year    16030   38660   38660 4 Year  in_st~
##  8 Adirondac~ New ~ NY      Publ~ 2 Year    11660    5375    9935 2 Year  in_st~
##  9 Adrian Co~ Mich~ MI      Priv~ 4 Year    11318   37087   37087 4 Year  in_st~
## 10 Advanced ~ Virg~ VA      For ~ 2 Year       NA   13680   13680 2 Year  in_st~
## # ... with 5,934 more rows, 1 more variable: totalCost <dbl>, and abbreviated
## #   variable names 1: state_code, 2: degree_length, 3: room_and_board,
## #   4: in_state_tuition, 5: out_of_state_tuition, 6: degFactor
```

```
str(tcFactored)
```

```
## tibble [2,973 x 11] (S3: tbl_df/tbl/data.frame)
##  $ name                : chr [1:2973] "Aaniiih Nakoda College" "Abilene Christian University" "Abraha
##  $ state               : chr [1:2973] "Montana" "Texas" "Georgia" "Minnesota" ...
##  $ state_code          : chr [1:2973] "MT" "TX" "GA" "MN" ...
##  $ type                : chr [1:2973] "Public" "Private" "Public" "For Profit" ...
##  $ degree_length       : chr [1:2973] "2 Year" "4 Year" "2 Year" "2 Year" ...
##  $ room_and_board      : num [1:2973] NA 10350 8474 NA 16648 ...
##  $ in_state_tuition    : num [1:2973] 2380 34850 4128 17661 27810 ...
##  $ in_state_total      : num [1:2973] 2380 45200 12602 17661 44458 ...
##  $ out_of_state_tuition: num [1:2973] 2380 34850 12550 17661 27810 ...
##  $ out_of_state_total  : num [1:2973] 2380 45200 21024 17661 44458 ...
##  $ degFactor           : Factor w/ 3 levels "2 Year","4 Year",..: 1 2 1 1 2 2 2 1 2 1 ...
```

```
head(tcFactored)
```

```
## # A tibble: 6 x 11
##   name       state state~1 type  degre~2 room_~3 in_st~4 in_st~5 out_o~6 out_o~7
##   <chr>      <chr> <chr>   <chr> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Aaniiih N~ Mont~ MT      Publ~ 2 Year       NA    2380    2380    2380    2380
## 2 Abilene C~ Texas TX      Priv~ 4 Year    10350   34850   45200   34850   45200
## 3 Abraham B~ Geor~ GA      Publ~ 2 Year     8474    4128   12602   12550   21024
## 4 Academy C~ Minn~ MN      For ~ 2 Year       NA   17661   17661   17661   17661
## 5 Academy o~ Cali~ CA      For ~ 4 Year    16648   27810   44458   27810   44458
## 6 Adams Sta~ Colo~ CO      Publ~ 4 Year     8782    9440   18222   20456   29238
## # ... with 1 more variable: degFactor <fct>, and abbreviated variable names
## #   1: state_code, 2: degree_length, 3: room_and_board, 4: in_state_tuition,
## #   5: in_state_total, 6: out_of_state_tuition, 7: out_of_state_total
```

```
# Time to explore the data!
table(gatheredtc$state_code,gatheredtc$degFactor)
```

```
##
##      2 Year 4 Year Other
##   AK      2     10     0
##   AL     42     66     0
```

```
##    AR     48     44     0
##    AS      2      0     0
##    AZ     46     22     0
##    CA    238    270     0
##    CO     36     40     0
##    CT     28     44     0
##    DC      0     16     0
##    DE      8     10     0
##    FL     66    110     0
##    GA     58    100     0
##    GU      2      0     0
##    HI     16     12     0
##    IA     36     68     0
##    ID      8     18     0
##    IL    104    146     0
##    IN     36     88     0
##    KS     50     54     0
##    KY     30     58     0
##    LA     16     52     0
##    MA     42    144     0
##    MD     32     58     0
##    ME     18     36     0
##    MI     60     96     0
##    MN     66     76     0
##    MO     46    100     0
##    MS     30     34     0
##    MT     22     22     0
##    NC    118    116     0
##    ND     18     18     0
##    NE     20     46     0
##    NH     14     28     0
##    NJ     42     66     0
##    NM     28     20     0
##    NV      8     12     0
##    NY    116    326     0
##    OH     94    160     0
##    OK     30     50     0
##    OR     30     50     0
##    PA     62    258     0
##    PR     12     70     0
##    RI      2     20     0
##    SC     46     68     0
##    SD     10     26     0
##    TN     34     90     0
##    TX    134    164     0
##    UT      8     20     0
##    VA     60     98     0
##    VI      0      2     0
##    VT      6     32     0
##    WA     66     54     0
##    WI     62     72     0
##    WV     18     42     0
##    WY     14      2     0
```

```
bystate = gatheredtc %>%
  group_by(state) %>%
  mutate(freq = n()) %>%
  summarize(numSchools = sum(freq)) %>%
  mutate(prop=numSchools/sum(numSchools)) %>%
  arrange(desc(prop))
bystate
```

```
## # A tibble: 51 x 3
##    state          numSchools   prop
##    <chr>               <int>  <dbl>
##  1 California         258064 0.211
##  2 New York           195364 0.159
##  3 Pennsylvania       102400 0.0835
##  4 Texas               88804 0.0724
##  5 Ohio                64516 0.0526
##  6 Illinois            62500 0.0510
##  7 North Carolina      54756 0.0447
##  8 Massachusetts       34596 0.0282
##  9 Florida             30976 0.0253
## 10 Georgia             24964 0.0204
## # ... with 41 more rows
```

This table shows that California has a very high number of schools proportional to other states in the country.

```
prop.table(table(gatheredtc$degFactor))
```

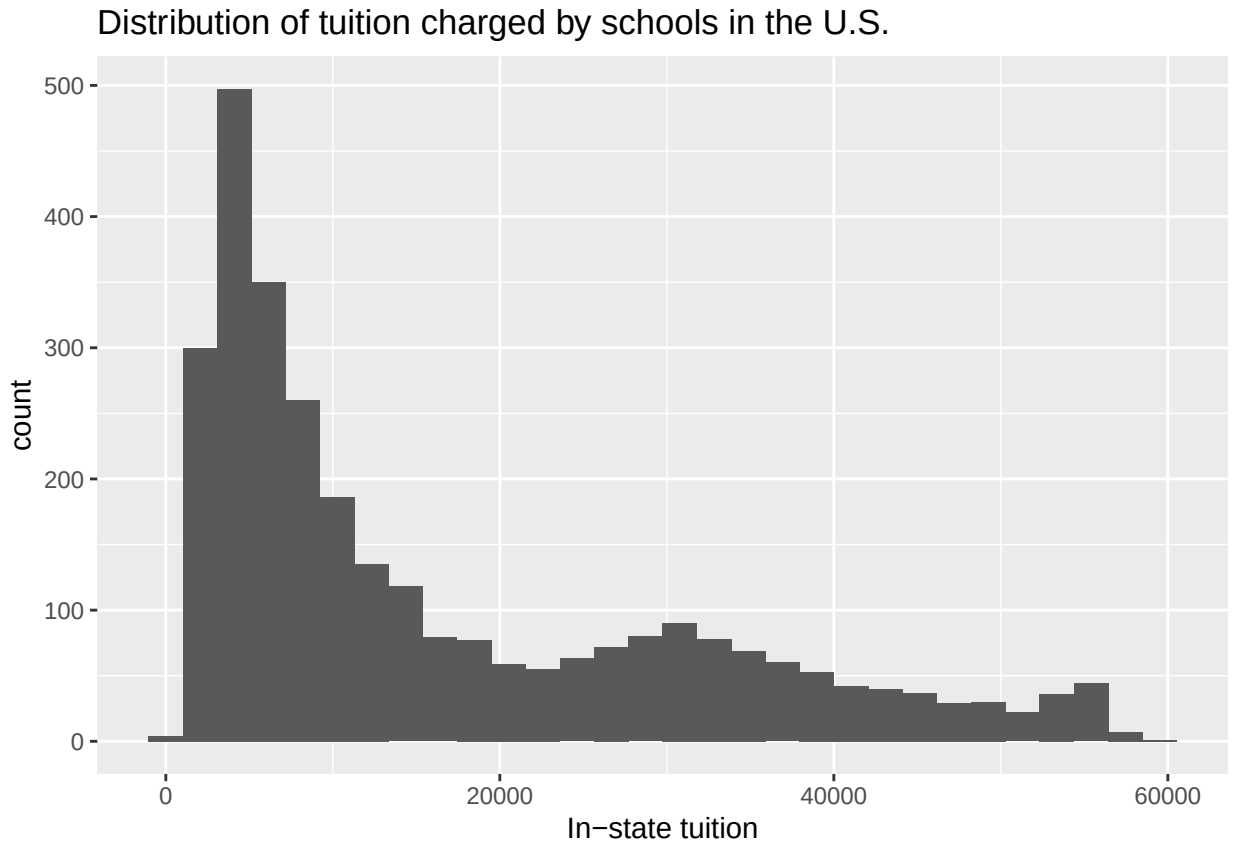
```
##
##    2 Year    4 Year     Other
## 0.3768506 0.6231494 0.0000000
```

```
table(gatheredtc$state)
```

```
##
##        Alabama         Alaska        Arizona       Arkansas     California
##            108             12             68             92            508
##       Colorado    Connecticut       Delaware        Florida        Georgia
##             76             72             18            176            158
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##             28             26            250            124            104
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##            104             88             68             54             90
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##            186            156            142             64            146
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##             44             66             20             42            108
##     New Mexico       New York North Carolina   North Dakota           Ohio
##             48            442            234             36            254
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##             80             80            320             22            114
##   South Dakota      Tennessee          Texas           Utah        Vermont
##             36            124            298             28             38
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##            158            120             60            134             16
```

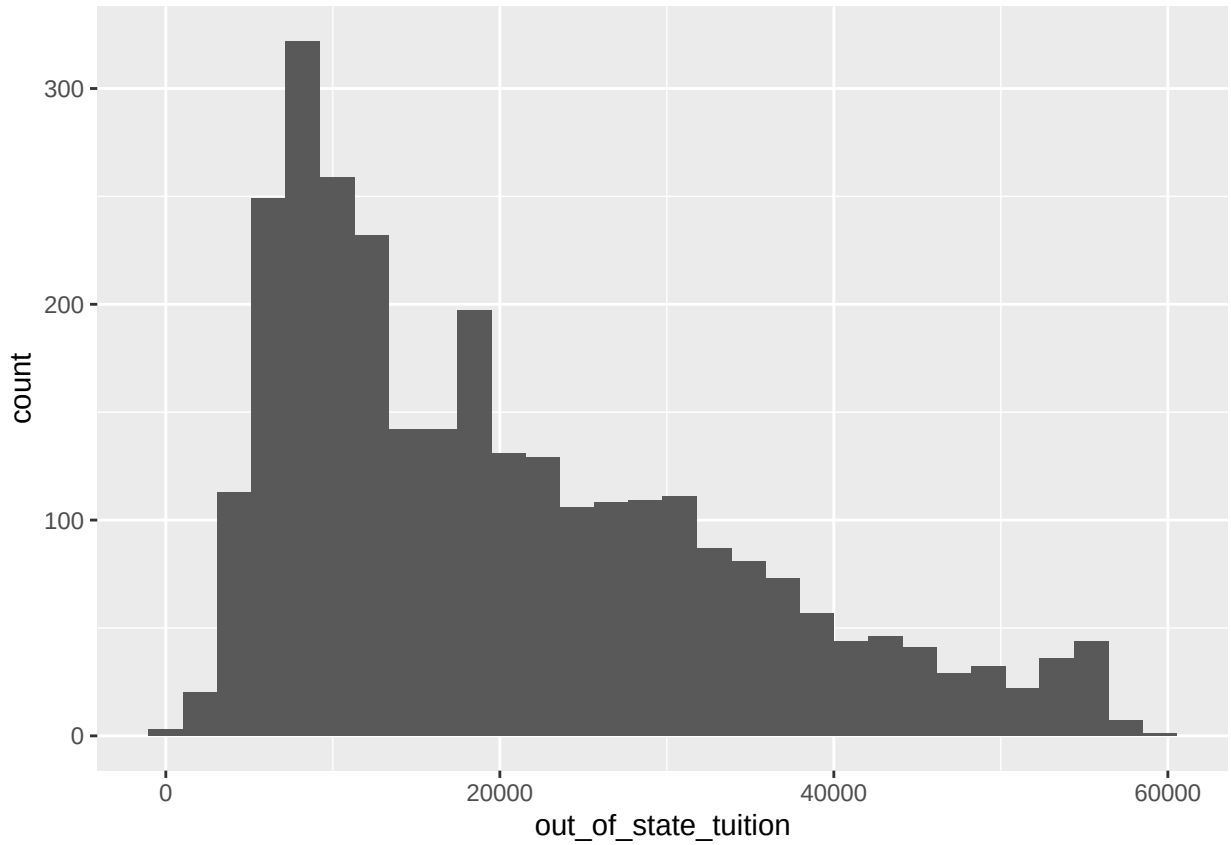
```
ggplot(tcFactored, aes(x=in_state_tuition)) + geom_histogram() +
  ggtitle("Distribution of tuition charged by schools in the U.S.")+
  xlab("In-state tuition")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



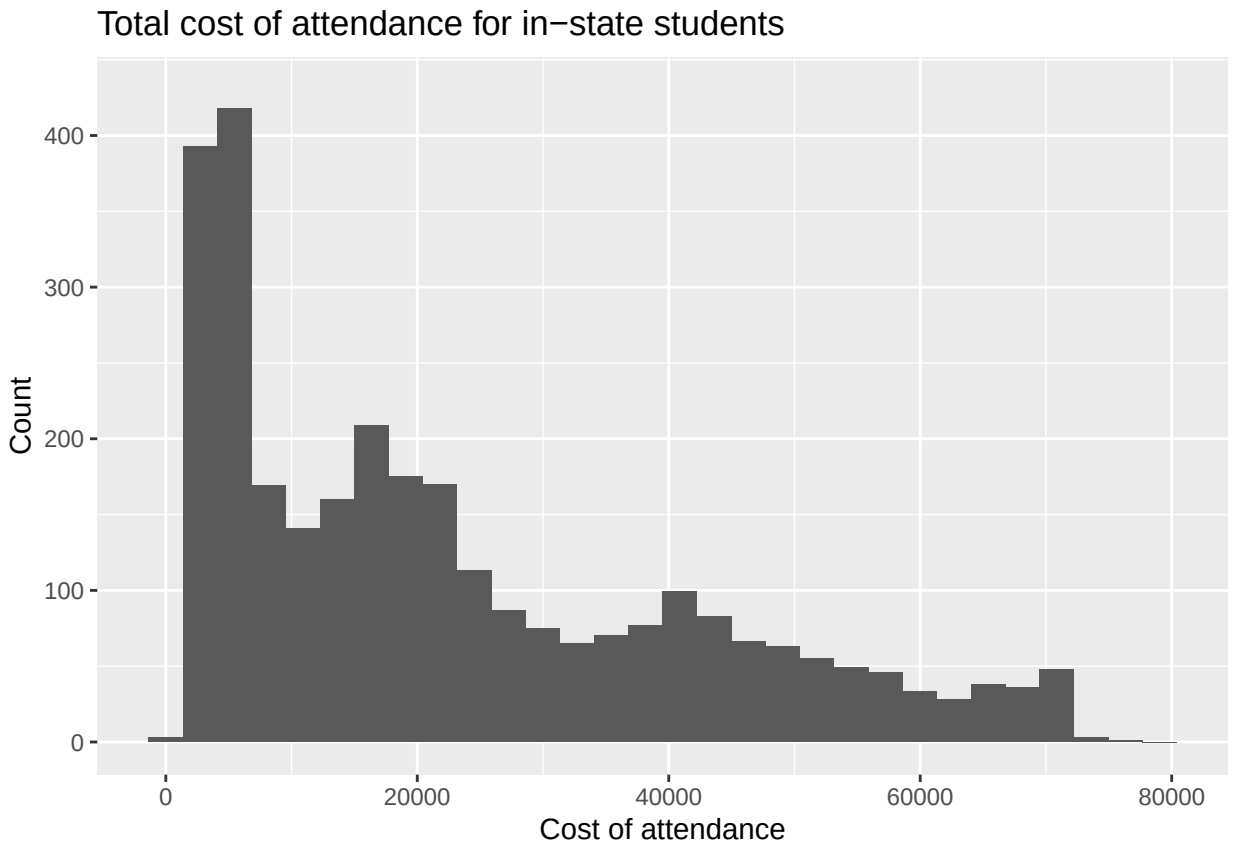

```
ggplot(tcFactored, aes(x=out_of_state_tuition))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



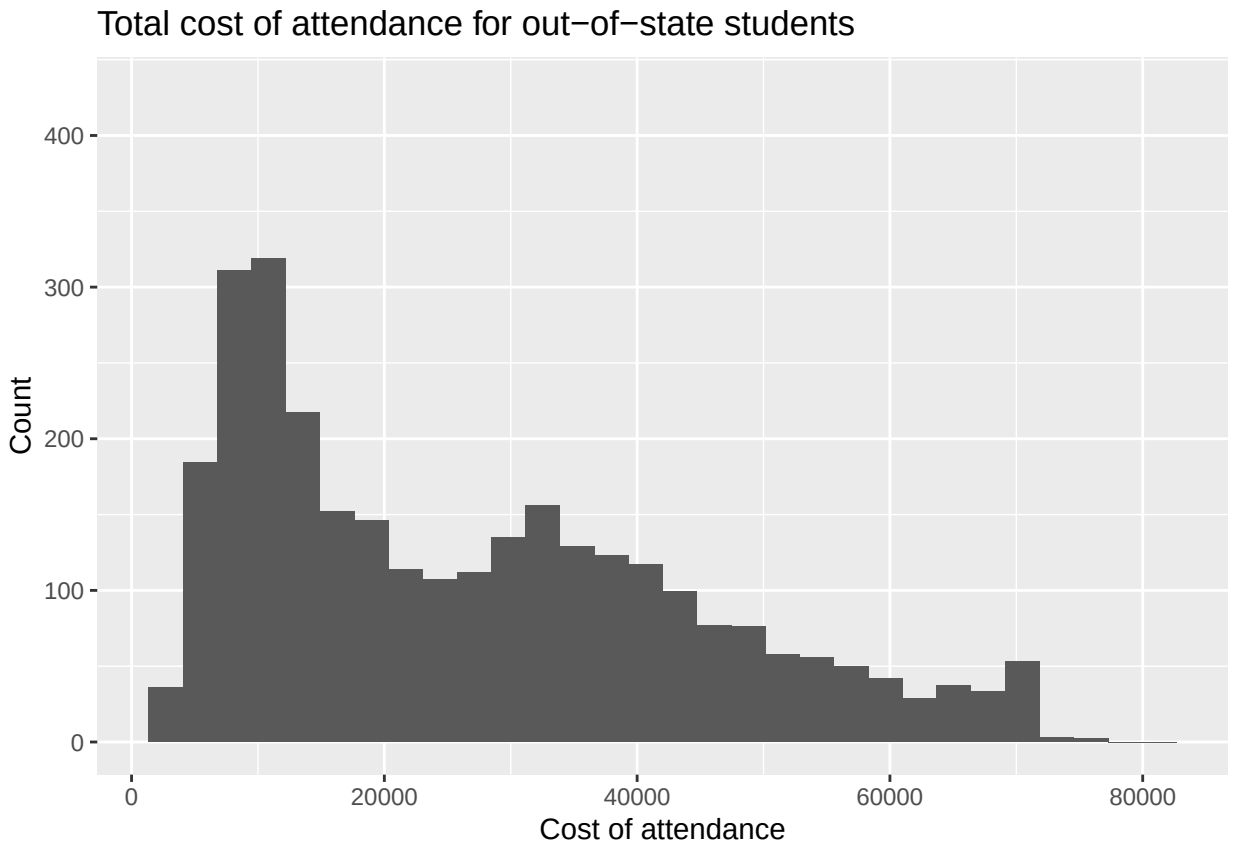

```
ggplot(tcFactored, aes(x=in_state_total))+geom_histogram()+expand_limits(x=80000,y=430) +
  ggtitle("Total cost of attendance for in-state students")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") # for the y axis label
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



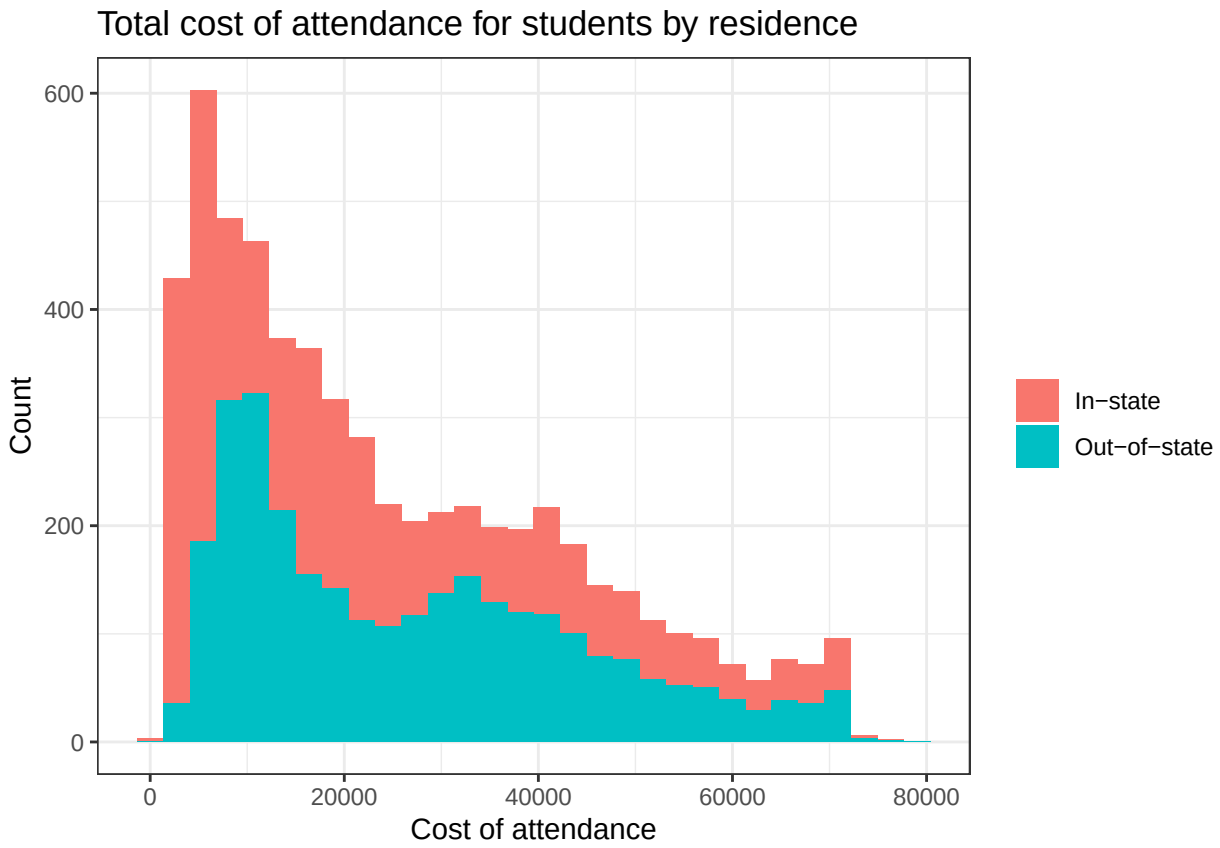

```
ggplot(tcFactored, aes(x=out_of_state_total))+geom_histogram()+expand_limits(x=80000,y=430) +
  ggtitle("Total cost of attendance for out-of-state students")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") # for the y axis label
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(gatheredtc, aes(x=totalCost,fill=in_out))+geom_histogram()+expand_limits(x=80000,y=430) +
  ggtitle("Total cost of attendance for students by residence")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") + # for the y axis label
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_fill_discrete(name = "Student Residence", labels = c("In-state", "Out-of-state"))
```

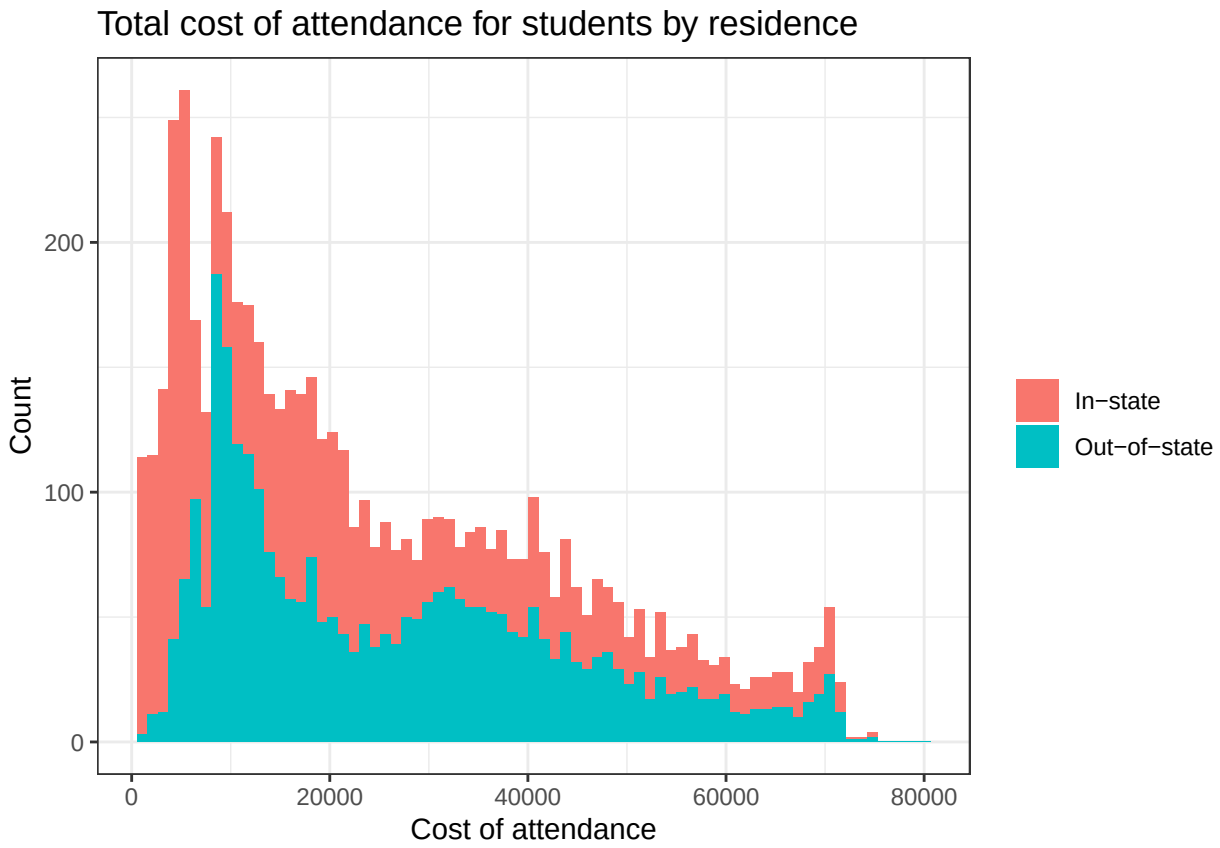
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#ggtitle(label) # for the main title
#xlab(label) # for the x axis label
#ylab(label) # for the y axis label
#labs(...) # for the main title, axis labels and legend titles
```

As the above plots show, it's clear that the distributions are skewed to the right which means that expensive schools are generally less common. It's interesting to see that both of these seem to have similar shapes, and a hint of evidence for a slight bimodal distribution.

```
ggplot(gatheredtc, aes(x=totalCost,fill=in_out))+geom_histogram(bins=75)+expand_limits(x=80000) +
  ggtitle("Total cost of attendance for students by residence")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") + # for the y axis label
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_fill_discrete(name = "Student Residence", labels = c("In-state", "Out-of-state"))
```



Upon further inspection by increasing the bin number, the shape becomes more distinct. The second mode is mostly just a bump for the out-of-state group, but something interesting appears in the in-state group! Is there a cause for this disruption?

```
tcInStateSummr = tcFactored %>%
  group_by(degFactor) %>%
  summarize(median(in_state_total))

tcOutStateSummr = tcFactored %>%
  group_by(degFactor) %>%
  summarize(median(out_of_state_total))

tcInStateSummr
```

```
## # A tibble: 3 x 2
##   degFactor `median(in_state_total)`
##   <fct>                        <dbl>
## 1 2 Year                       4972.
## 2 4 Year                      28287
## 3 Other                        8448
```

```
tcOutStateSummr
```

```
## # A tibble: 3 x 2
##   degFactor `median(out_of_state_total)`
##   <fct>                             <dbl>
## 1 2 Year                            10291
## 2 4 Year                            34888
## 3 Other                             14640
```

This is a simple calculation of the median for 2-year and 4-year schools for total cost to out-of-state students.

```
tcFours = tcFactored %>%
  filter(degFactor=="4 Year")
tcFours
```

```
## # A tibble: 1,852 x 11
##    name      state state~1 type  degre~2 room_~3 in_st~4 in_st~5 out_o~6 out_o~7
##    <chr>     <chr> <chr>   <chr> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Abilene ~ Texas TX      Priv~ 4 Year    10350   34850   45200   34850   45200
##  2 Academy ~ Cali~ CA      For ~ 4 Year    16648   27810   44458   27810   44458
##  3 Adams St~ Colo~ CO      Publ~ 4 Year     8782    9440   18222   20456   29238
##  4 Adelphi ~ New ~ NY      Priv~ 4 Year    16030   38660   54690   38660   54690
##  5 Adrian C~ Mich~ MI      Priv~ 4 Year    11318   37087   48405   37087   48405
##  6 Adventis~ Flor~ FL      Priv~ 4 Year     4200   15150   19350   15150   19350
##  7 Agnes Sc~ Geor~ GA      Priv~ 4 Year    12330   41160   53490   41160   53490
##  8 Alabama ~ Alab~ AL      Publ~ 4 Year     8379    9698   18077   17918   26297
##  9 Alabama ~ Alab~ AL      Publ~ 4 Year     5422   11068   16490   19396   24818
## 10 Alaska B~ Alas~ AK      Priv~ 4 Year     5700    9300   15000    9300   15000
## # ... with 1,842 more rows, 1 more variable: degFactor <fct>, and abbreviated
## #   variable names 1: state_code, 2: degree_length, 3: room_and_board,
## #   4: in_state_tuition, 5: in_state_total, 6: out_of_state_tuition,
## #   7: out_of_state_total
```

```
tcTwos = tcFactored %>%
  filter(degFactor=="2 Year")
```

```
tc4YOOS_Summary = tcFours%>%
  summarise(count_4YOOS=n(),
            min=min(tcFours$out_of_state_total, na.rm=TRUE),
            Q1=quantile(tcFours$out_of_state_total, prob=0.25,na.rm=TRUE),
            med=median(tcFours$out_of_state_total, na.rm=TRUE), #or quantile(AQI,prob=0.5,na.rm=TRUE)
            Q3=quantile(tcFours$out_of_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcFours$out_of_state_total, na.rm=TRUE))

tc4YIS_Summary = tcFours%>%
  summarise(count_4YIS=n(),
            min=min(tcFours$in_state_total, na.rm=TRUE),
            Q1=quantile(tcFours$in_state_total, prob=0.25,na.rm=TRUE),
            med=median(tcFours$in_state_total, na.rm=TRUE), #or quantile(AQI,prob=0.5,na.rm=TRUE)
            Q3=quantile(tcFours$in_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcFours$in_state_total, na.rm=TRUE))

tc2YOOS_Summary = tcTwos%>%
  summarise(count_2YOOS=n(),
            min=min(tcTwos$out_of_state_total, na.rm=TRUE),
            Q1=quantile(tcTwos$out_of_state_total, prob=0.25,na.rm=TRUE),
            med=median(tcTwos$out_of_state_total, na.rm=TRUE), #or quantile(AQI,prob=0.5,na.rm=TRUE)
            Q3=quantile(tcTwos$out_of_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcTwos$out_of_state_total, na.rm=TRUE))

tc2YIS_Summary = tcTwos%>%
  summarise(count_2YIS=n(),
            min=min(tcTwos$in_state_total, na.rm=TRUE),
            Q1=quantile(tcTwos$in_state_total, prob=0.25,na.rm=TRUE),
```

```
            med=median(tcTwos$in_state_total, na.rm=TRUE), #or quantile(AQI,prob=0.5,na.rm=TRUE)
            Q3=quantile(tcTwos$in_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcTwos$in_state_total, na.rm=TRUE))

tc4YOOS_Summary
```

```
## # A tibble: 1 x 6
##   count_4YOOS   min    Q1   med    Q3   max
##         <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1        1852  1430 24951 34888 46670 75003
```

```
tc4YIS_Summary
```

```
## # A tibble: 1 x 6
##   count_4YIS   min    Q1   med     Q3   max
##        <int> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1       1852  1430 18199 28287 44846. 75003
```

```
tc2YOOS_Summary
```

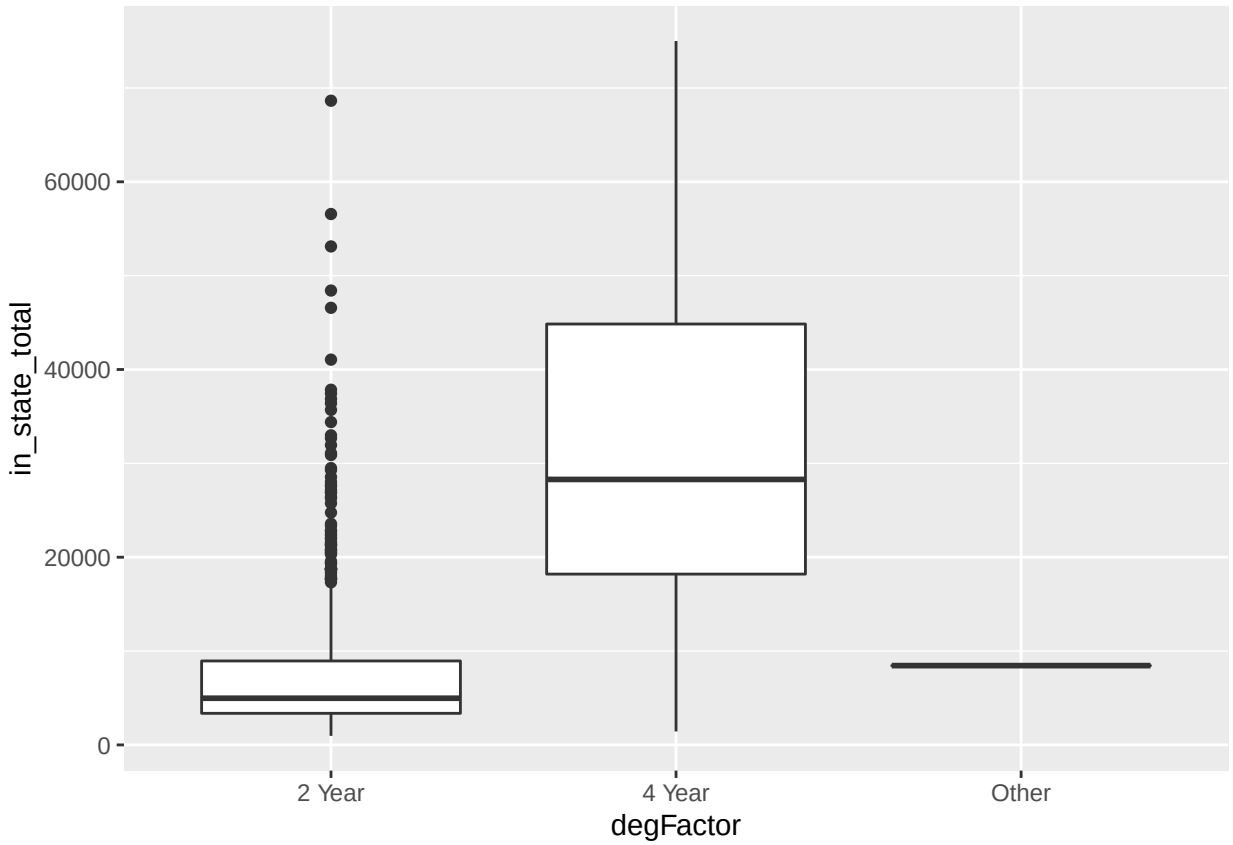
```
## # A tibble: 1 x 6
##   count_2YOOS   min    Q1   med    Q3   max
##         <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1        1120  1376 8196. 10291 13598 68640
```

```
tc2YIS_Summary
```

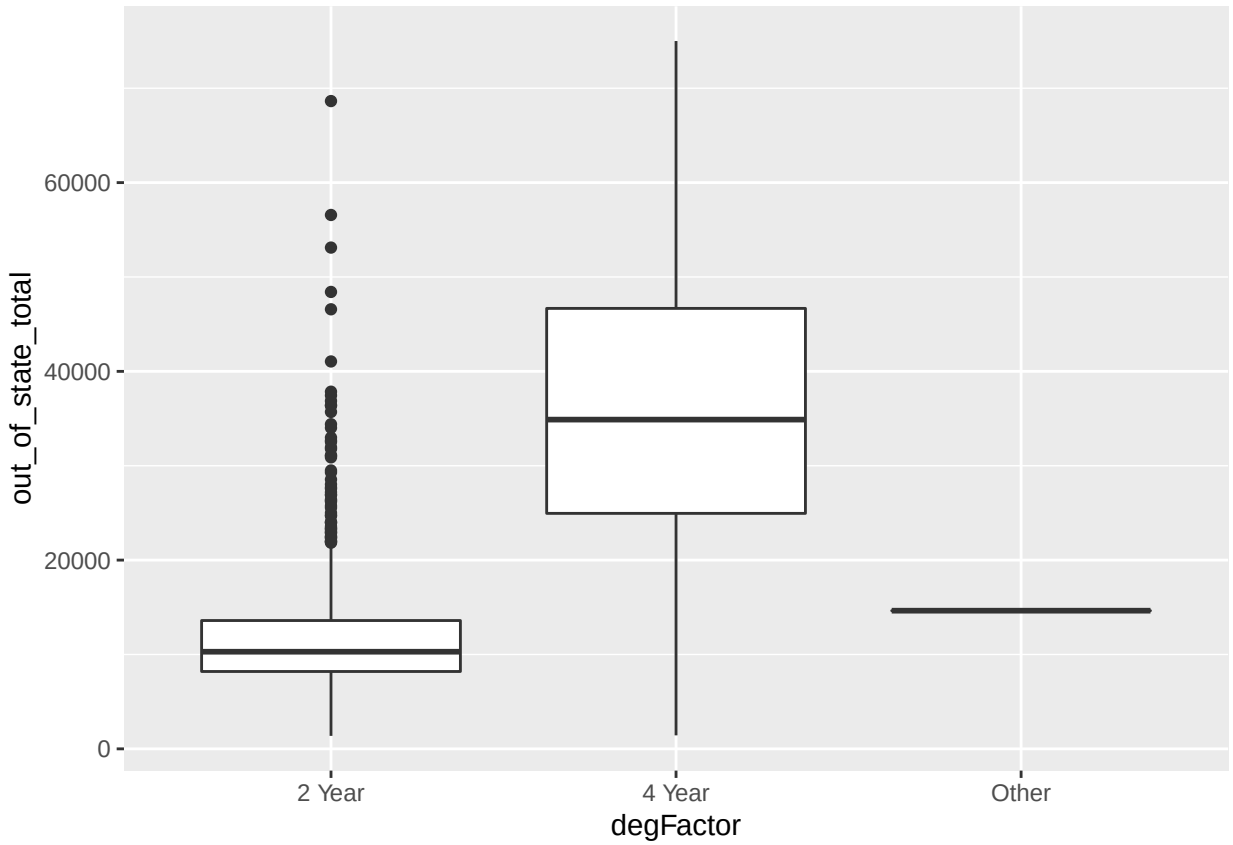
```
## # A tibble: 1 x 6
##   count_2YIS   min    Q1   med    Q3   max
##        <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1       1120   962 3364. 4972.  8946 68640
```

These are the 5-number summaries for each of the categorical variables of interest.

```
ggplot(tcFactored, aes(x = degFactor, y = in_state_total)) +  # ggplot function
  geom_boxplot()
```



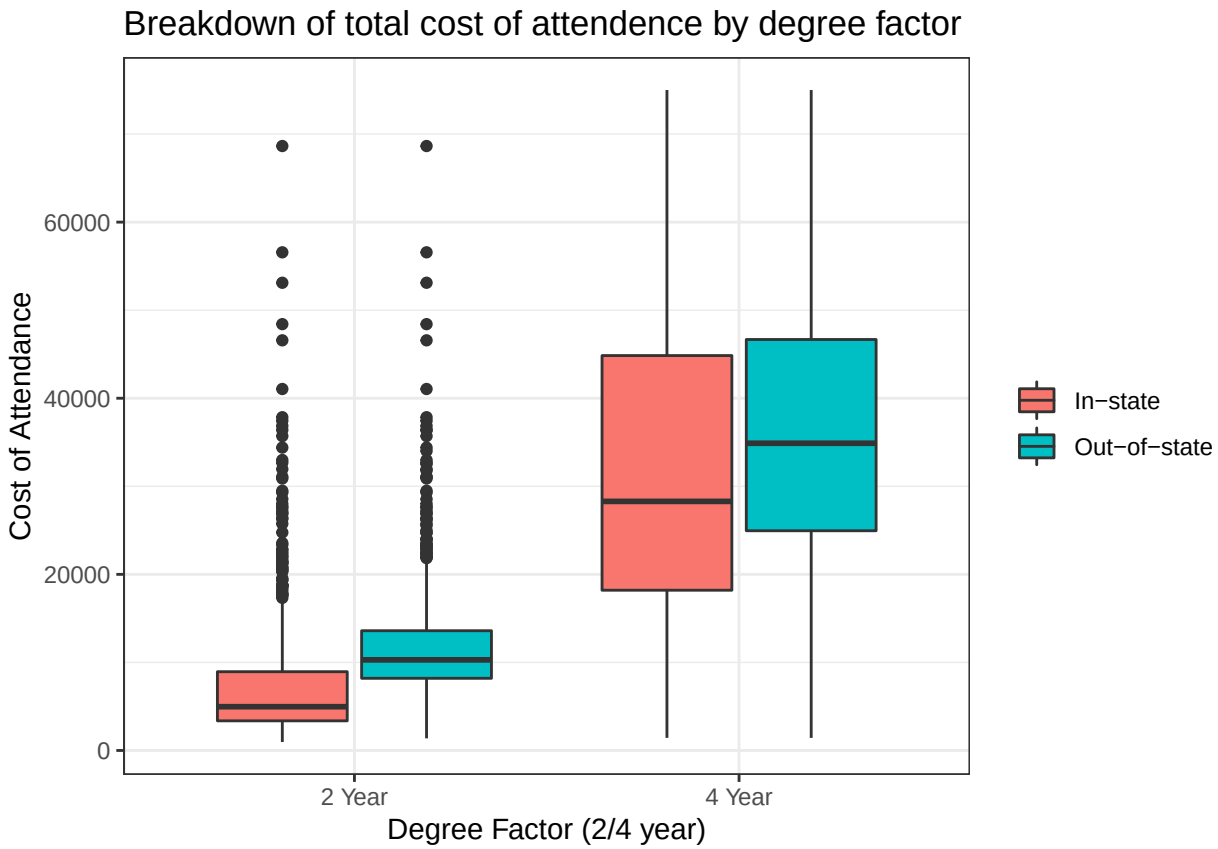

```
ggplot(tcFactored, aes(x = degFactor, y = out_of_state_total)) +  # ggplot function
  geom_boxplot()
```



These box plots (couldn't figure out how to make an overlayed boxplot with both in/out of state variables) show a clear difference in the general cost between 2-year and 4-year institutions, and that out-of-state students generally pay more.

```
#ggplot(tcFactored, aes(x=tcInStateSummr$degFactor, fill=tcInStateSummr$in_state_total)) +
#  geom_histogram( color="#e9ecef", alpha=0.6, position = 'identity') +
#  scale_fill_manual(values=c("#69b3a2", "#404080"))
```

```
ggplot(gatheredtc, aes(x = degFactor, y = totalCost, fill=in_out)) +  # ggplot function
  geom_boxplot()+
  ggtitle("Breakdown of total cost of attendence by degree factor")+ # for the main title
  xlab("Degree Factor (2/4 year)")+ # for the x axis label
  ylab("Cost of Attendance")+ # for the y axis label
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_fill_discrete(name = "Student Residence", labels = c("In-state", "Out-of-state"))
```

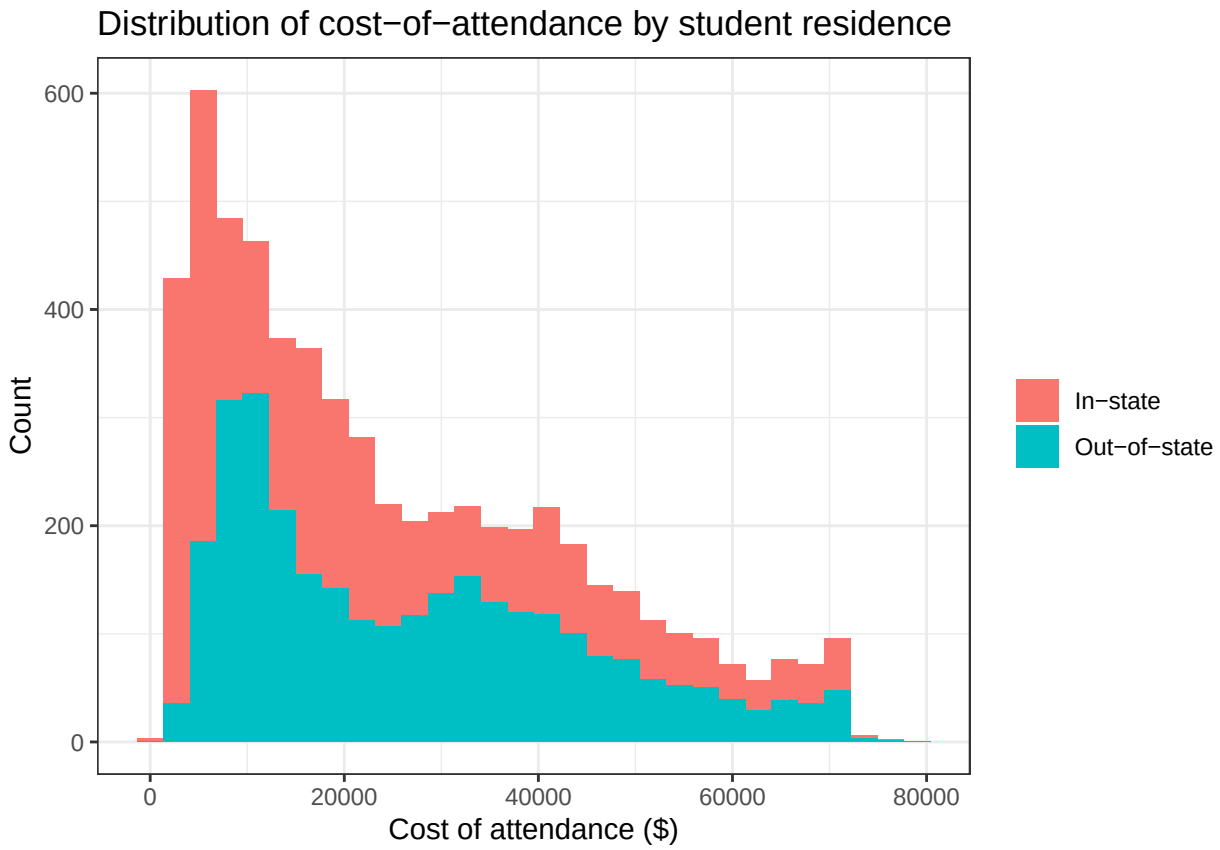


There is clearly a difference here between how much students should expect to pay given their residency status, but it isn't as absurdly significant as we were anticipating given that we hear from high school guidence counselors, specifically about 4-year institutions. Therefore, we should look for another potential explanation for the contribution to higher costs of attendance for some students.

Total Cost of Attendance by Student Residence and College Type

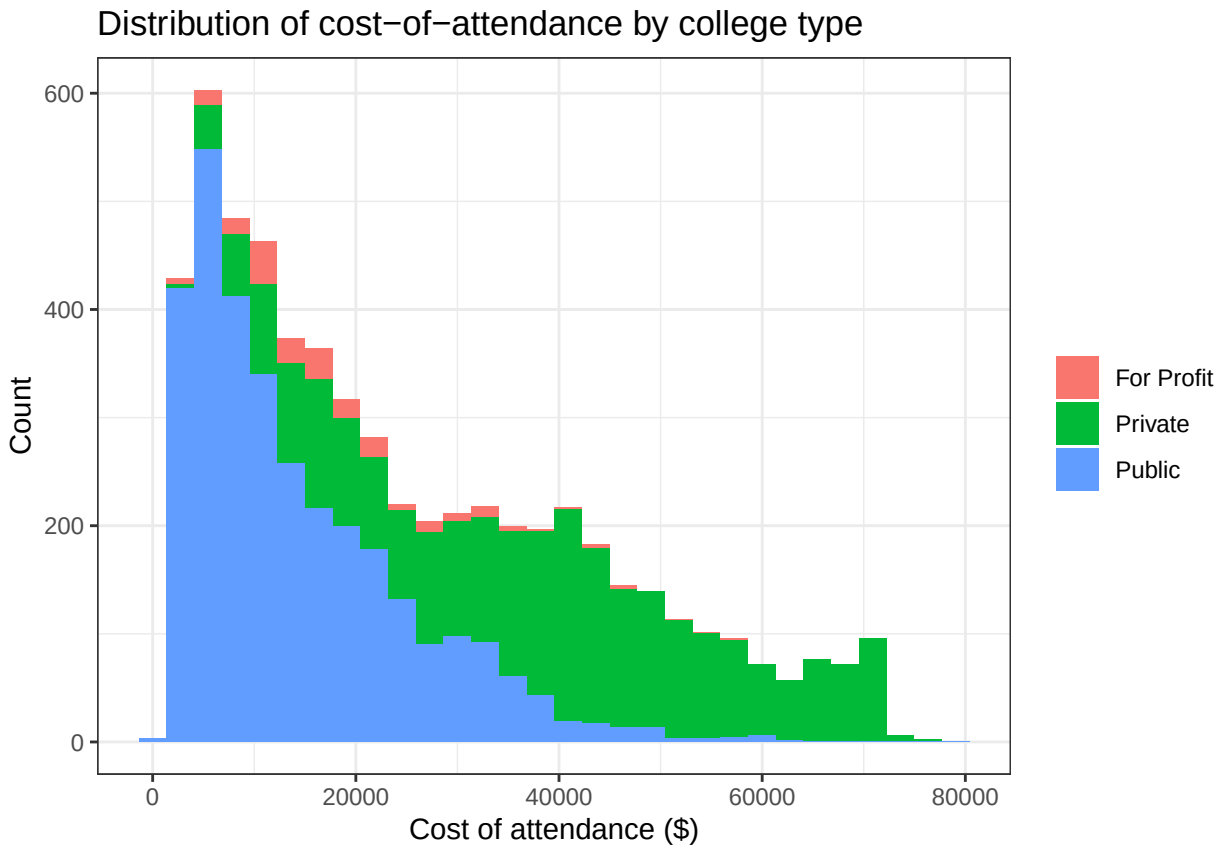
```
ggplot(gatheredtc,aes(x=totalCost, fill=in_out)) +
  geom_histogram() +
  expand_limits(x=80000) +
  ggtitle("Distribution of cost-of-attendance by student residence")+
  xlab("Cost of attendance ($)") +
  ylab("Count") +
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_fill_discrete(labels = c("In-state", "Out-of-state"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

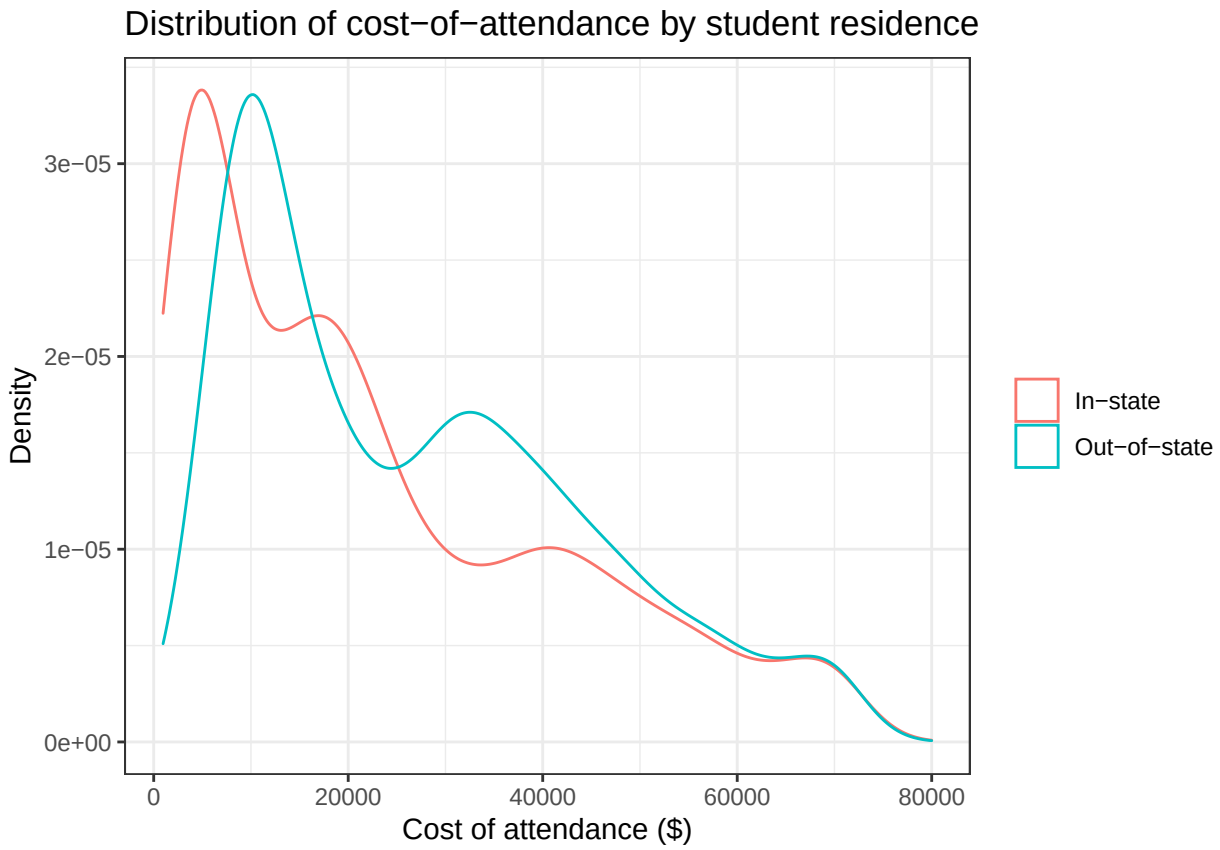



```
ggplot(gatheredtc,aes(x=totalCost, fill=type)) +
  geom_histogram() +
  expand_limits(x=80000) +
  ggtitle("Distribution of cost-of-attendance by college type")+
  xlab("Cost of attendance ($)") +
  ylab("Count") +
  theme_bw()+theme(
    legend.title = element_blank()) + scale_fill_discrete(labels=c('For Profit','Private','Public'))
```

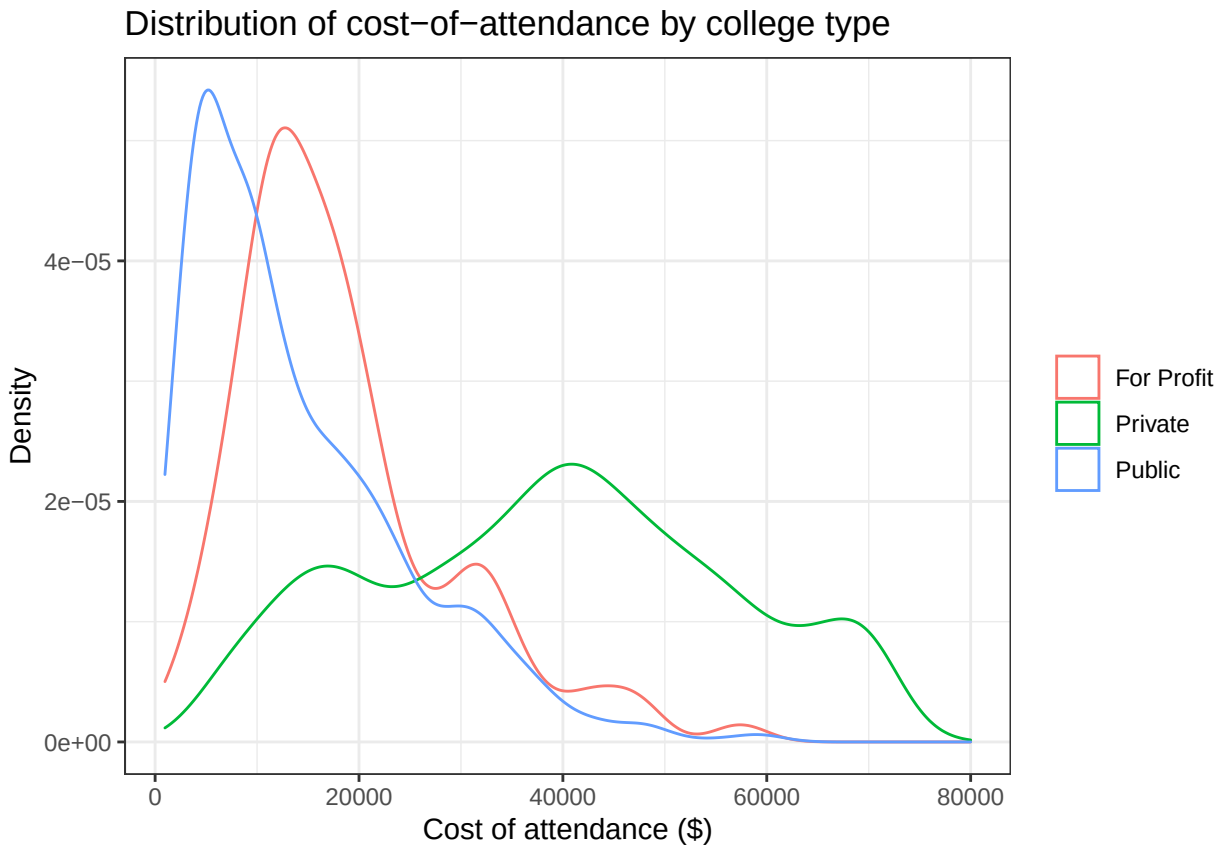
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(gatheredtc,aes(x=totalCost, color=in_out)) +
  geom_density() +
  expand_limits(x=80000) +
  ggtitle("Distribution of cost-of-attendance by student residence")+
  xlab("Cost of attendance ($)") +
  ylab("Density") +
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_color_discrete(labels = c("In-state", "Out-of-state"))
```




```
ggplot(gatheredtc,aes(x=totalCost, color=type)) +
  geom_density() +
  expand_limits(x=80000) +
  ggtitle("Distribution of cost-of-attendance by college type")+
  xlab("Cost of attendance ($)") +
  ylab("Density") +
  theme_bw()+theme(
    legend.title = element_blank()) + scale_color_discrete(labels=c('For Profit','Private','Public'))
```




```
# ggplot(tcFactored, aes(x=in_state_total))+geom_histogram()+expand_limits(x=80000,y=430) +
#   ggtitle("Total cost of attendance for in-state students")+ # for the main title
#   xlab("Cost of attendance")+ # for the x axis label
#   ylab("Count") # for the y axis label
#
#
# ggplot(tcFactored, aes(x=out_of_state_total))+geom_histogram()+expand_limits(x=80000,y=430) +
#   ggtitle("Total cost of attendance for out-of-state students")+ # for the main title
#   xlab("Cost of attendance")+ # for the x axis label
#   ylab("Count") # for the y axis label


#ggtitle(label) # for the main title
#xlab(label) # for the x axis label
#ylab(label) # for the y axis label
#labs(...) # for the main title, axis labels and legend titles
```

This is a density plot that shows how high the cost of attendance is for schools across the country relative to each other given institution type.

```
# Time to explore the data!

# Commenting out ggplot stuff to do dplyr first
#ggplot(ti,aes(x=year, y=total_price)) + geom_point()

#this is the median cost of attendance for instate/outstate

median_IN_COA <- tc %>%
```

```
  filter(degree_length=='4 Year') %>%
  group_by(state_code)%>%
  summarize(median_instate_COA = median(in_state_total))

median_IN_COA
```

```
## # A tibble: 53 x 2
##    state_code median_instate_COA
##    <chr>                   <dbl>
##  1 AK                      17017
##  2 AL                      18646
##  3 AR                      19023
##  4 AZ                      25037
##  5 CA                      30416
##  6 CO                      20976.
##  7 CT                      46455
##  8 DC                      50702.
##  9 DE                      26542
## 10 FL                      23352
## # ... with 43 more rows
```

```
median_OUT_COA <- tc %>%
  filter(degree_length=='4 Year') %>%
  group_by(state_code)%>%
  summarize(median_outstate_COA = median(out_of_state_total))

median_OUT_COA
```

```
## # A tibble: 53 x 2
##    state_code median_outstate_COA
##    <chr>                    <dbl>
##  1 AK                       28604
##  2 AL                       27880
##  3 AR                       23709
##  4 AZ                       37190
##  5 CA                       36103
##  6 CO                       36096
##  7 CT                       48656.
##  8 DC                       50702.
##  9 DE                       30700
## 10 FL                       32000
## # ... with 43 more rows
```

```
#  mutate(mean_instate_COA=mean(in_state_tuition)) %>%
#  mutate(mean_outofstate_COA=mean(out_of_state_tuition))%>%
```

```
#In the following graph, I want to find out whether colleges with higher STEM enrollment tend to cost m

jointisp = ti %>%
  left_join(sp) %>%
  group_by(stem_percent)%>%
  summarize(medianNet=median(net_cost))
```

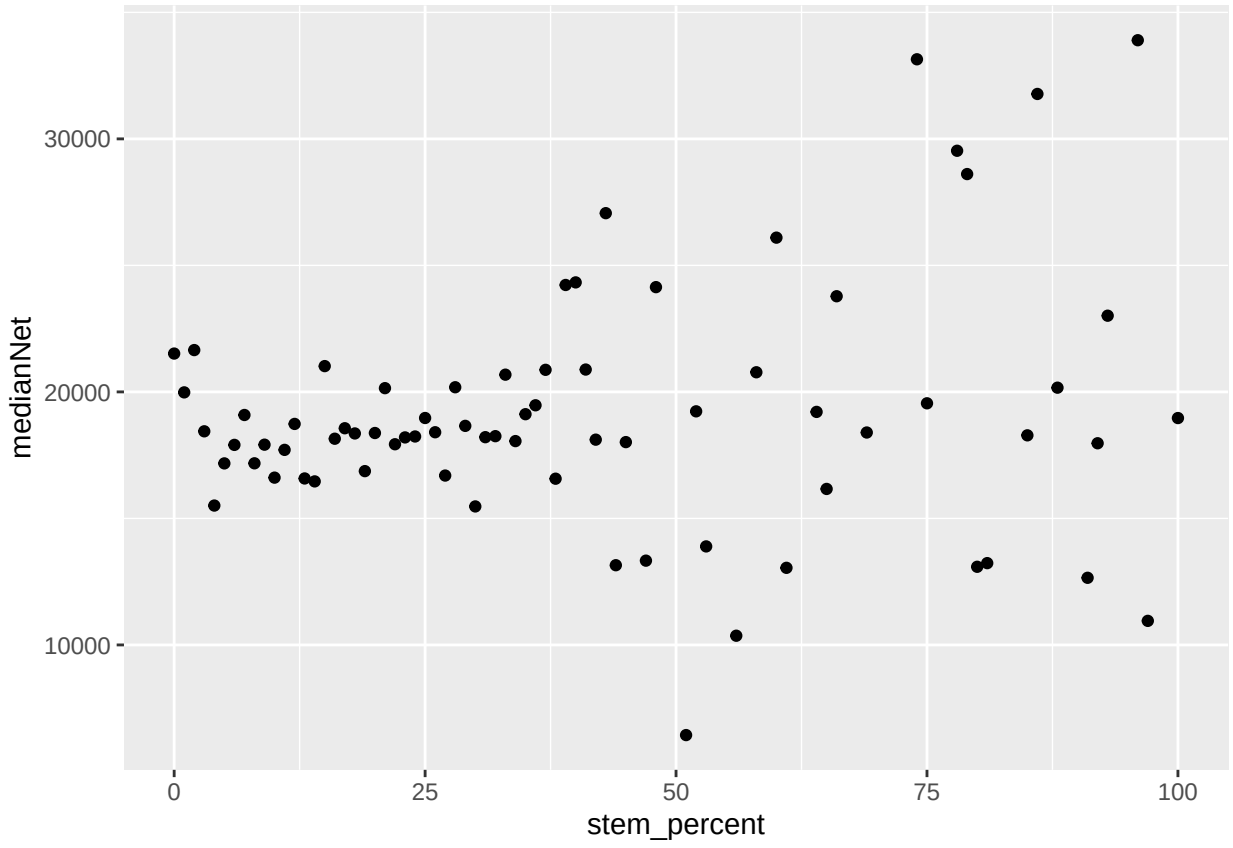
```
## Joining, by = "name"
```

```
jointisp
```

```
## # A tibble: 75 x 2
##    stem_percent medianNet
##           <dbl>     <dbl>
##  1            0    21516.
##  2            1    19981.
##  3            2    21653
##  4            3    18443.
##  5            4    15509
##  6            5    17175
##  7            6    17907
##  8            7    19088
##  9            8    17178
## 10            9    17915
## # ... with 65 more rows
```

```
ggplot(jointisp, aes(stem_percent,medianNet)) + geom_point()
```

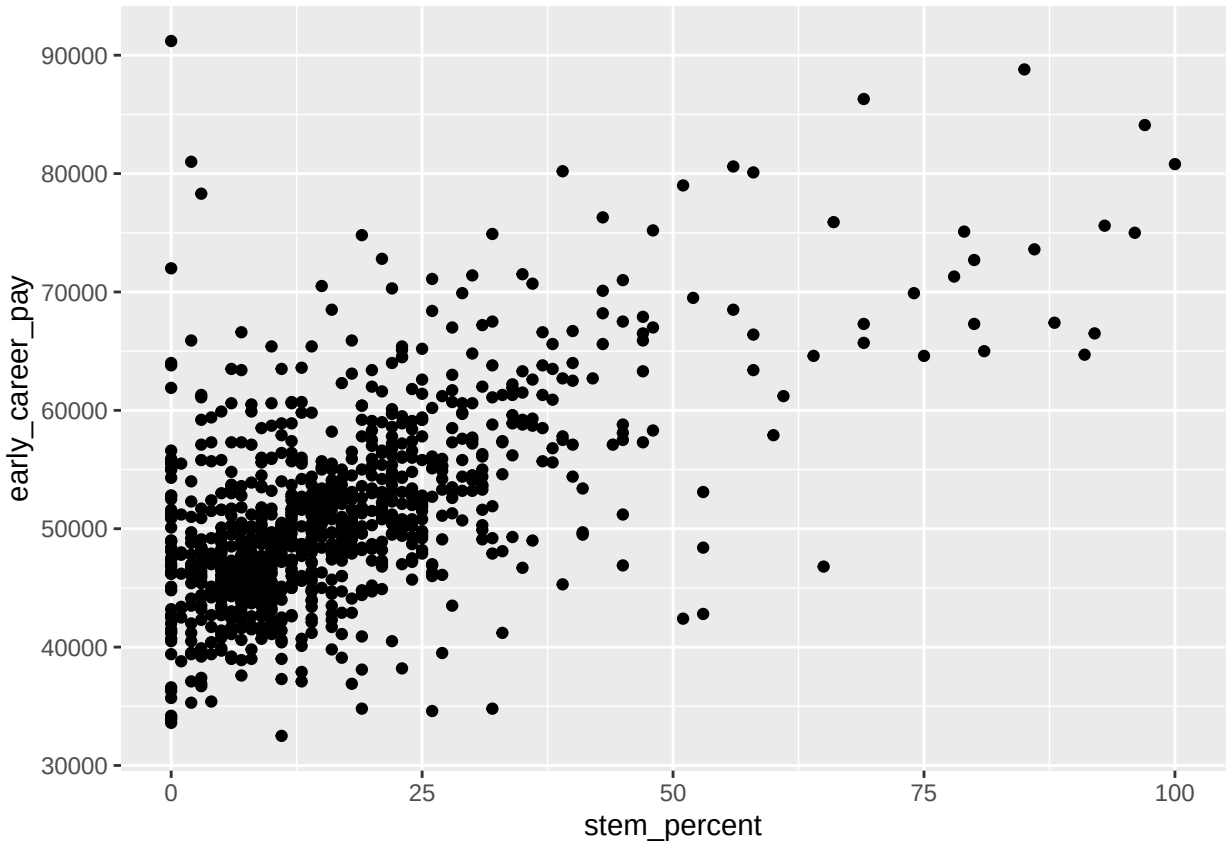
```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
#My conclusion is that there is no association between higher STEM enrollment and median net cost.
```

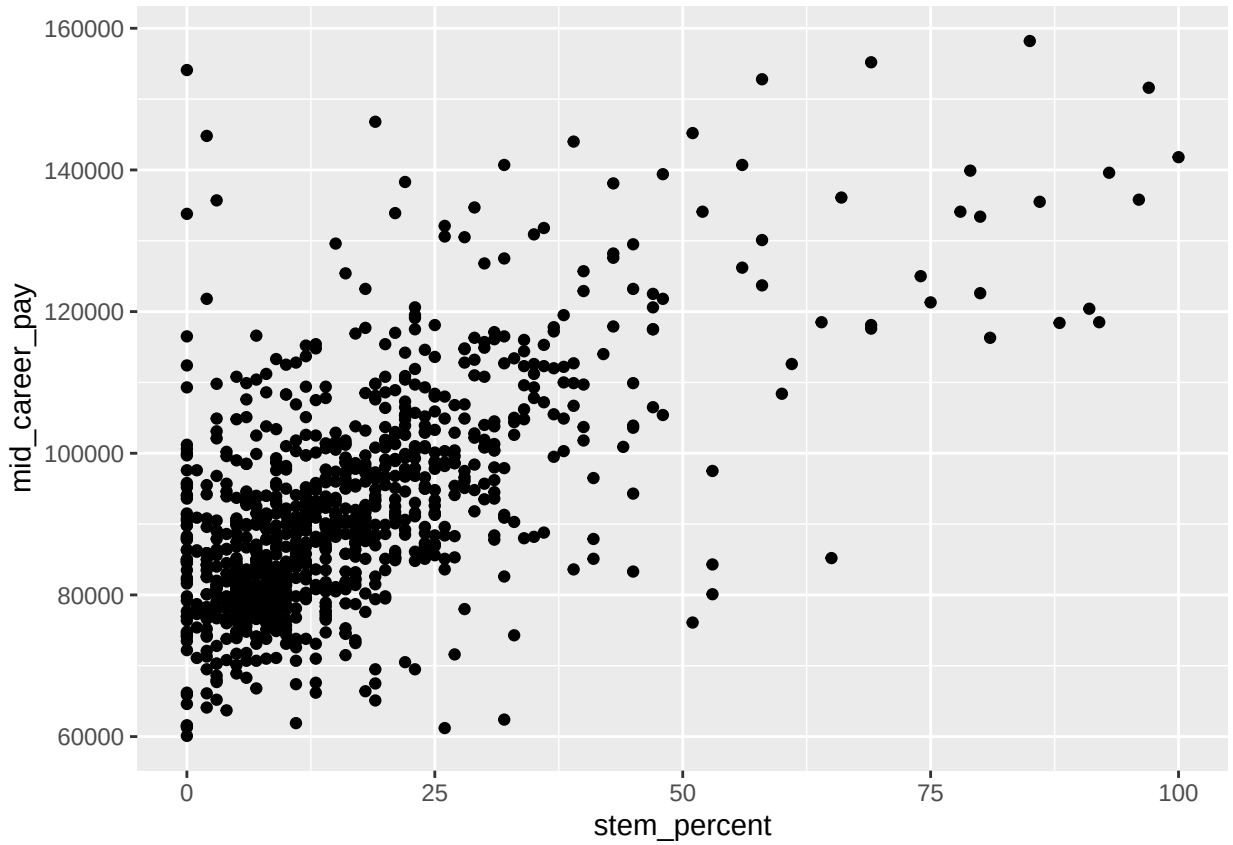
In this graph I want to see whether a higher STEM enrollment has a high association with early career pay. In a later graphic, I will want to see whether the trend keeps for mid-career pay.

```
ggplot(sp, aes(stem_percent,early_career_pay)) + geom_point()
```

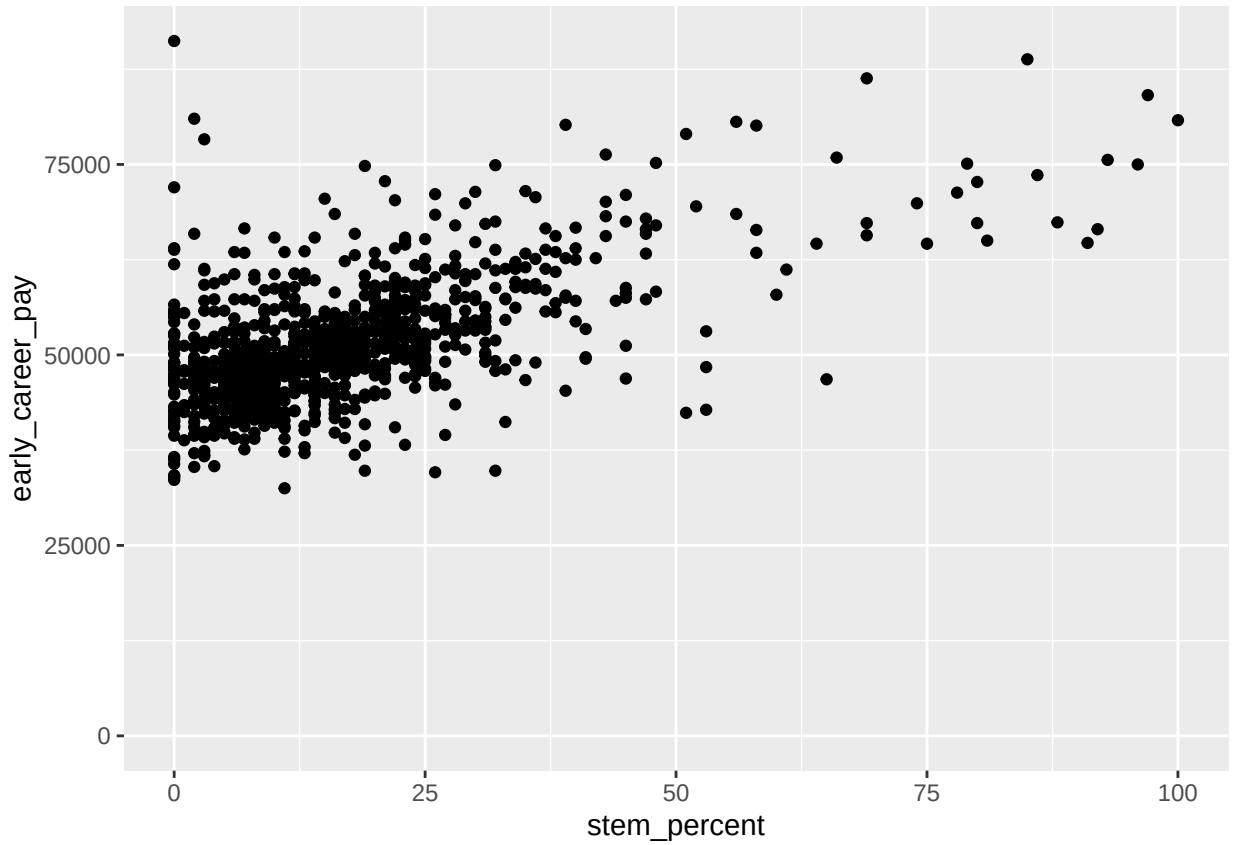


My conclusion is that there seems to be a weak positive correlation between these two variables.

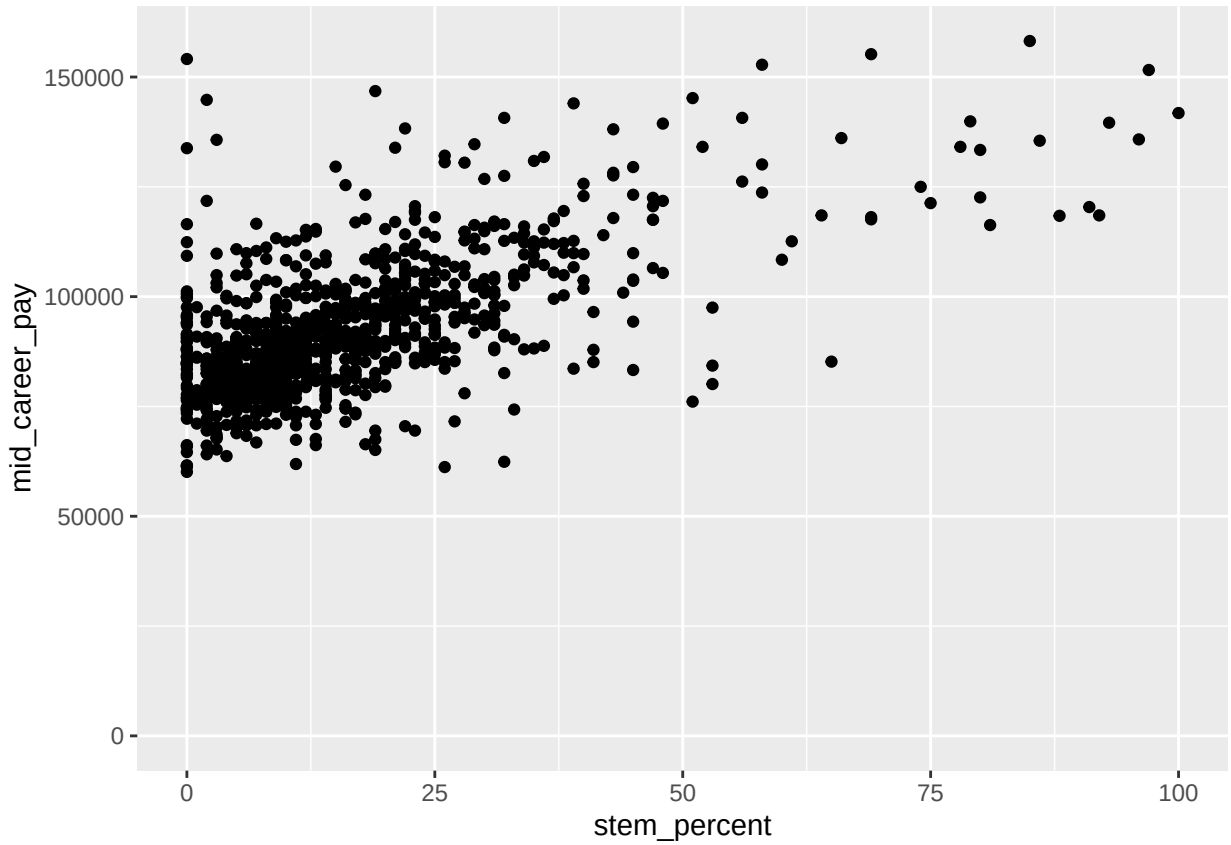
```
#In this graphic, perhaps the trend keeps?

ggplot(sp, aes(stem_percent,mid_career_pay)) + geom_point()
```



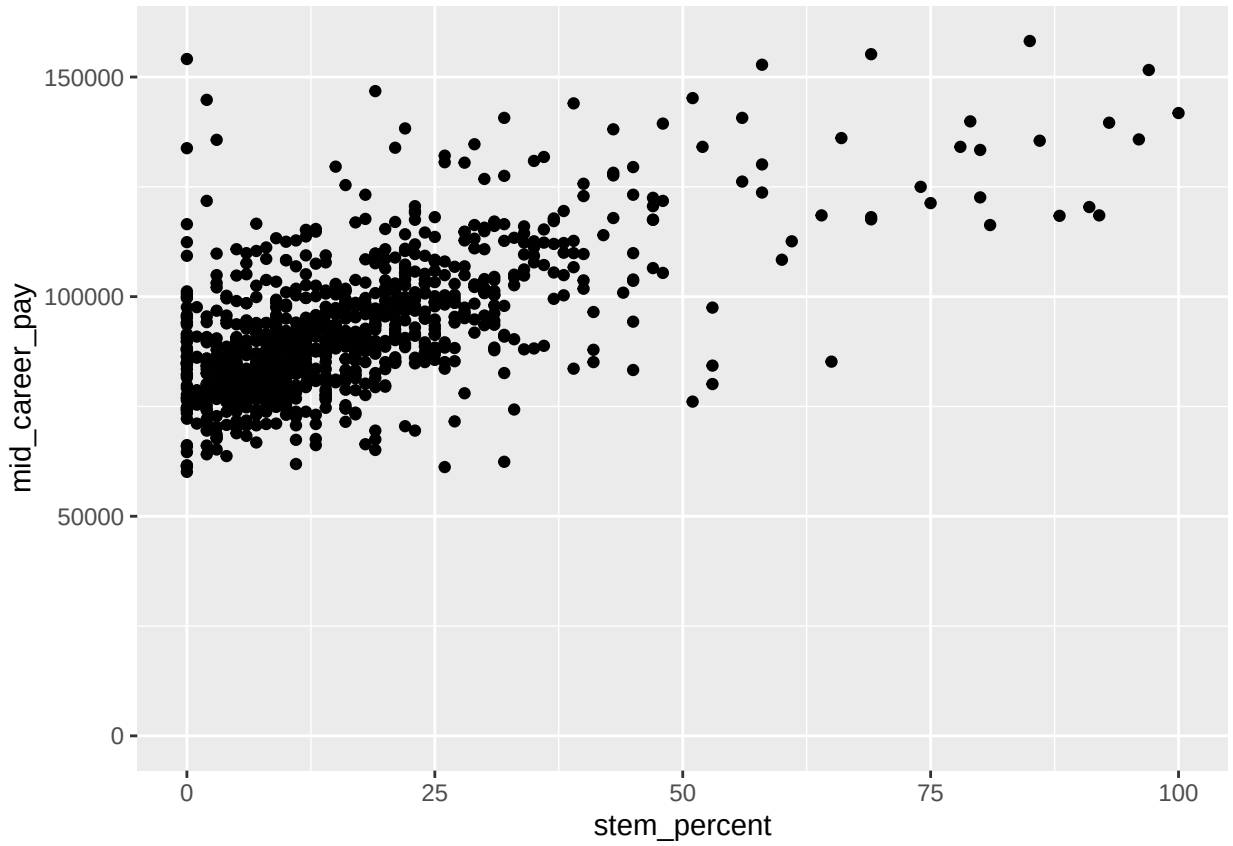

```
#attempting to make the graphics easier to differentiate. I'm not sure what the difference is.

ggplot(sp, aes(x=stem_percent,y=early_career_pay)) + geom_point() + expand_limits(x=0,y=0)
```




```
ggplot(sp, aes(x=stem_percent,y=mid_career_pay)) + geom_point() + expand_limits(x=0,y=0)
```




```
ggplot(sp, aes(x=stem_percent,y=mid_career_pay)) + geom_point() + expand_limits(x=0,y=0)
```




```
income = ti %>%
  group_by(income_lvl) %>%
  summarize(Count=n()) %>%
  mutate(Percent = round((Count/sum(Count)*100))) %>%
  arrange(desc(Count))
income
```

```
## # A tibble: 5 x 3
##   income_lvl        Count Percent
##   <chr>             <int>   <dbl>
## 1 0 to 30,000       44969      22
## 2 30,001 to 48,000  43384      21
## 3 48_001 to 75,000  42600      20
## 4 75,001 to 110,000 40403      19
## 5 Over 110,000      37656      18
```

```
incomeByState = ti %>%
  group_by(income_lvl,state) %>%
  summarize(Count=n()) %>%
  mutate(Percent = round((Count/sum(Count)*100))) %>%
  arrange(desc(Count))
```

```
## `summarise()` has grouped output by 'income_lvl'. You can override using the
## `.groups` argument.
```

```
incomeByState
```

```
## # A tibble: 255 x 4
## # Groups:   income_lvl [5]
```

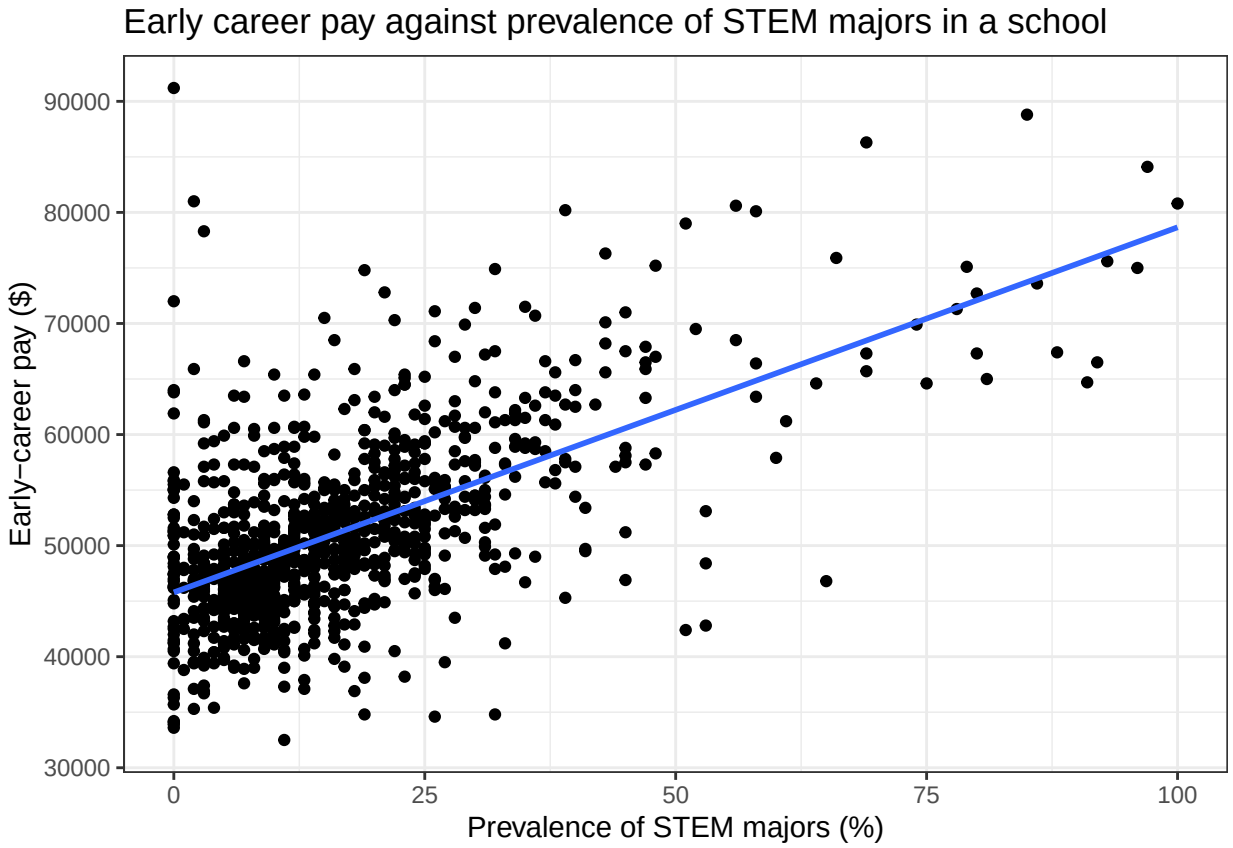
```
##    income_lvl        state Count Percent
##    <chr>             <chr> <int>   <dbl>
##  1 0 to 30,000       NY     3460       8
##  2 30,001 to 48,000  NY     3357       8
##  3 48_001 to 75,000  NY     3333       8
##  4 0 to 30,000       CA     3290       7
##  5 75,001 to 110,000 NY     3125       8
##  6 30,001 to 48,000  CA     2909       7
##  7 Over 110,000      NY     2877       8
##  8 48_001 to 75,000  CA     2841       7
##  9 0 to 30,000       PA     2776       6
## 10 30,001 to 48,000  PA     2736       6
## # ... with 245 more rows
```

```
tcFacJoinSp = tcFactored %>%
  filter(tcFactored$type!='For Profit') %>%
  inner_join(sp, by=c("name"="name"))
tcFacJoinSp
```

```
## # A tibble: 727 x 17
##    name      state state~1 type  degre~2 room_~3 in_st~4 in_st~5 out_o~6 out_o~7
##    <chr>     <chr> <chr>   <chr> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Adams St~ Colo~ CO      Publ~ 4 Year     8782    9440   18222   20456   29238
##  2 Adventis~ Flor~ FL      Priv~ 4 Year     4200   15150   19350   15150   19350
##  3 Agnes Sc~ Geor~ GA      Priv~ 4 Year    12330   41160   53490   41160   53490
##  4 Alabama ~ Alab~ AL      Publ~ 4 Year     5422   11068   16490   19396   24818
##  5 Alaska P~ Alas~ AK      Priv~ 4 Year     7300   20830   28130   20830   28130
##  6 Albany C~ New ~ NY      Priv~ 4 Year    10920   35105   46025   35105   46025
##  7 Albertus~ Conn~ CT      Priv~ 4 Year    13200   32060   45260   32060   45260
##  8 Albion C~ Mich~ MI      Priv~ 4 Year    12380   45775   58155   45775   58155
##  9 Alcorn S~ Miss~ MS      Publ~ 4 Year     9608    7144   16752    7144   16752
## 10 Allen Co~ Iowa  IA      Priv~ 4 Year     7282   19970   27252   19970   27252
## # ... with 717 more rows, 7 more variables: degFactor <fct>, rank <dbl>,
## #   state_name <chr>, early_career_pay <dbl>, mid_career_pay <dbl>,
## #   make_world_better_percent <dbl>, stem_percent <dbl>, and abbreviated
## #   variable names 1: state_code, 2: degree_length, 3: room_and_board,
## #   4: in_state_tuition, 5: in_state_total, 6: out_of_state_tuition,
## #   7: out_of_state_total
```

```
ggplot(sp, aes(stem_percent,early_career_pay)) + geom_point() +
  geom_smooth(method="lm",se=FALSE)+
  ggtitle("Early career pay against prevalence of STEM majors in a school")+
  xlab("Prevalence of STEM majors (%)")+
  ylab("Early-career pay ($)")+
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(sp$stem_percent,sp$early_career_pay)
```

```
## [1] 0.6050609
```

There seems to be a moderately strong correlation between the prevalence of STEM majors and how much graduates tend to

```
ECSPmodel = lm(early_career_pay~stem_percent,data=tcFacJoinSp)
summary(ECSPmodel)
```

```
##
## Call:
## lm(formula = early_career_pay ~ stem_percent, data = tcFacJoinSp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21134  -4041   -624   3056  34761
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45592.77     353.98  128.80   <2e-16 ***
## stem_percent   323.16      15.46   20.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6536 on 725 degrees of freedom
## Multiple R-squared:  0.3762, Adjusted R-squared:  0.3753
## F-statistic: 437.2 on 1 and 725 DF,  p-value: < 2.2e-16
```

Slope: 323.30 Y-intercept: 45584.50

```
cor(x=sp$stem_percent,y=sp$early_career_pay)
```

```
## [1] 0.6050609
```

```
cor(x=sp$stem_percent,y=sp$mid_career_pay)
```

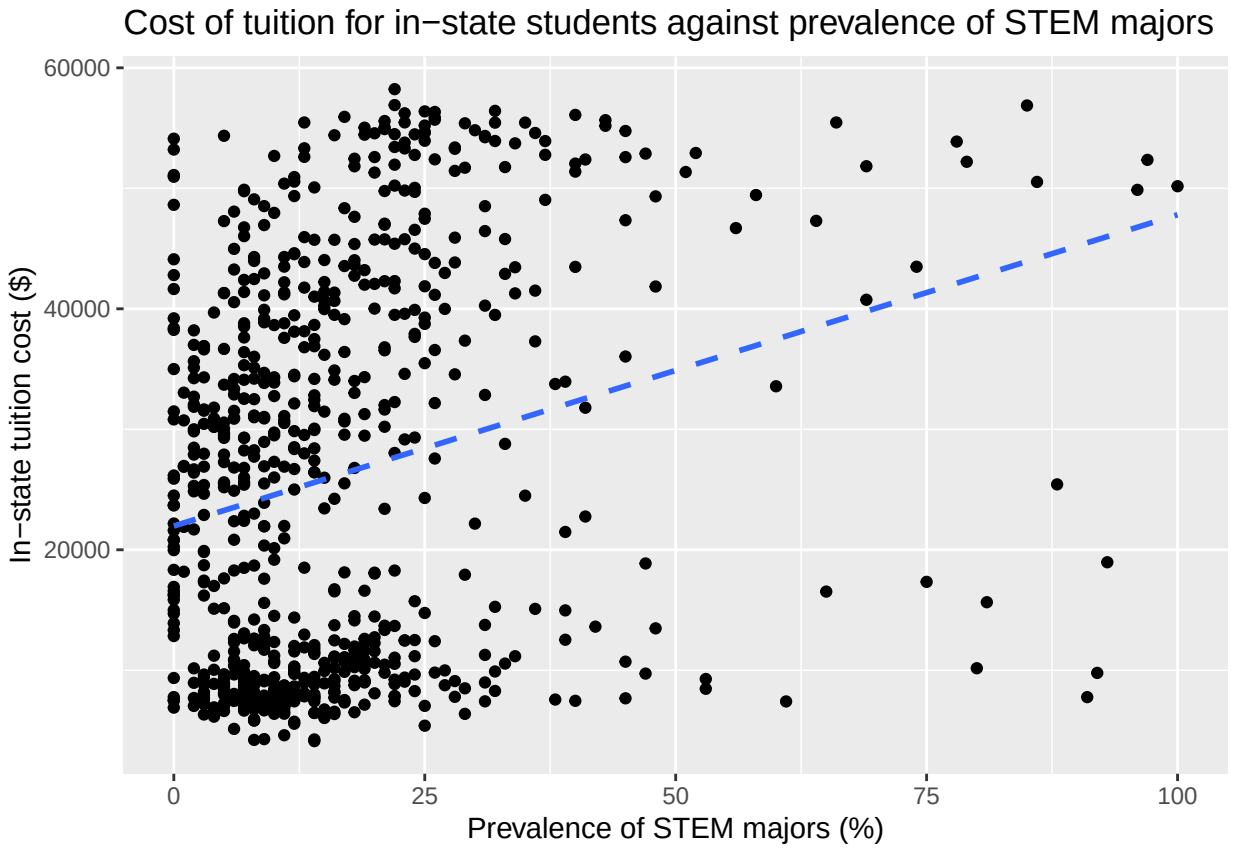
```
## [1] 0.6212143
```

Moderately strong positive association between prevalence of STEM majors at a school and general alumni early-career earnings.

Boring. STEM jobs tend to pay really well. Let's do something fun. Hear me out...

```
ggplot(tcFacJoinSp, aes(x=stem_percent,y=in_state_tuition)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE,lty=2) +
  ggtitle("Cost of tuition for in-state students against prevalence of STEM majors")+
  xlab("Prevalence of STEM majors (%)")+
  ylab("In-state tuition cost ($)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(x=tcFacJoinSp$stem_percent,y=tcFacJoinSp$in_state_tuition)
```

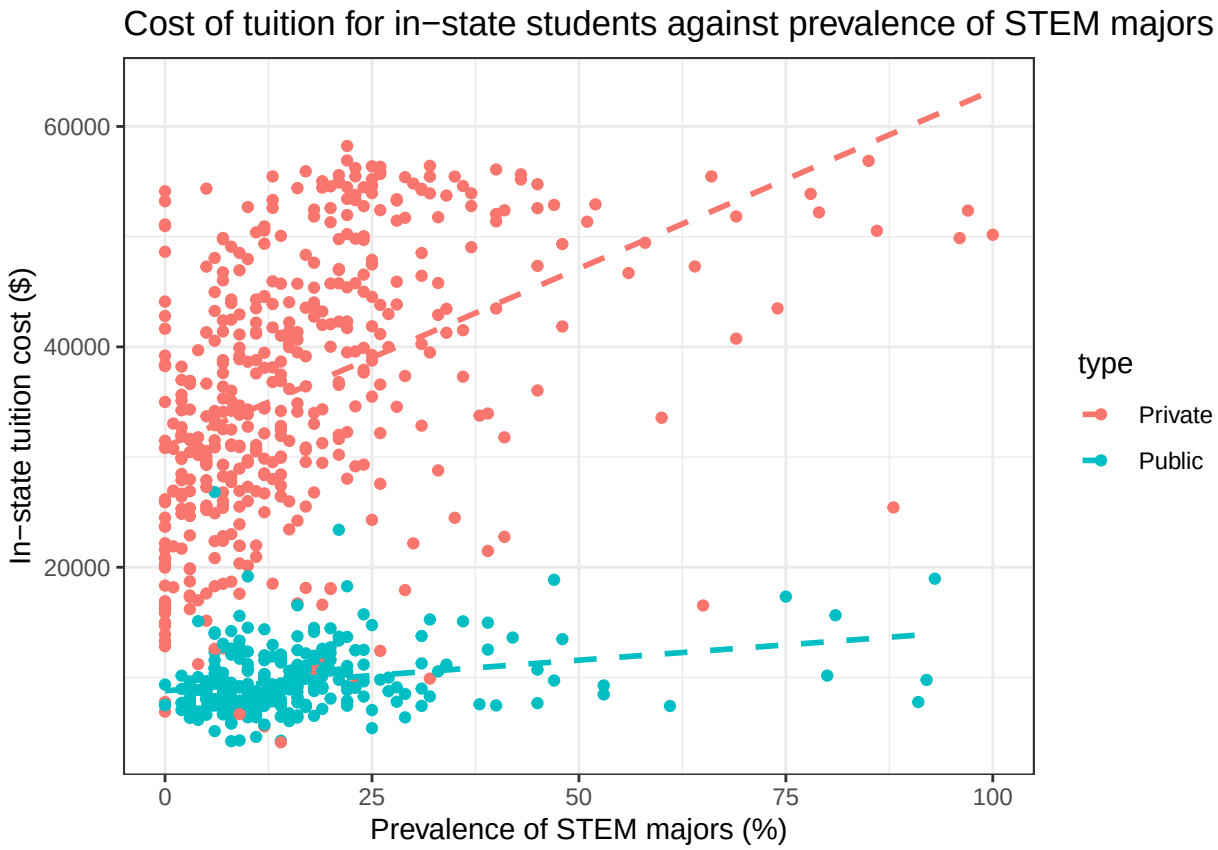
```
## [1] 0.2487816
```

Terrible fit. Couldn't predict the broad side of a barn!

```
ggplot(tcFacJoinSp, aes(x=stem_percent,y=in_state_tuition,color=type)) +
  geom_point() +
```

```
  geom_smooth(method="lm",se=FALSE,lty=2) +
  ggtitle("Cost of tuition for in-state students against prevalence of STEM majors")+
  xlab("Prevalence of STEM majors (%)")+
  ylab("In-state tuition cost ($)")+
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



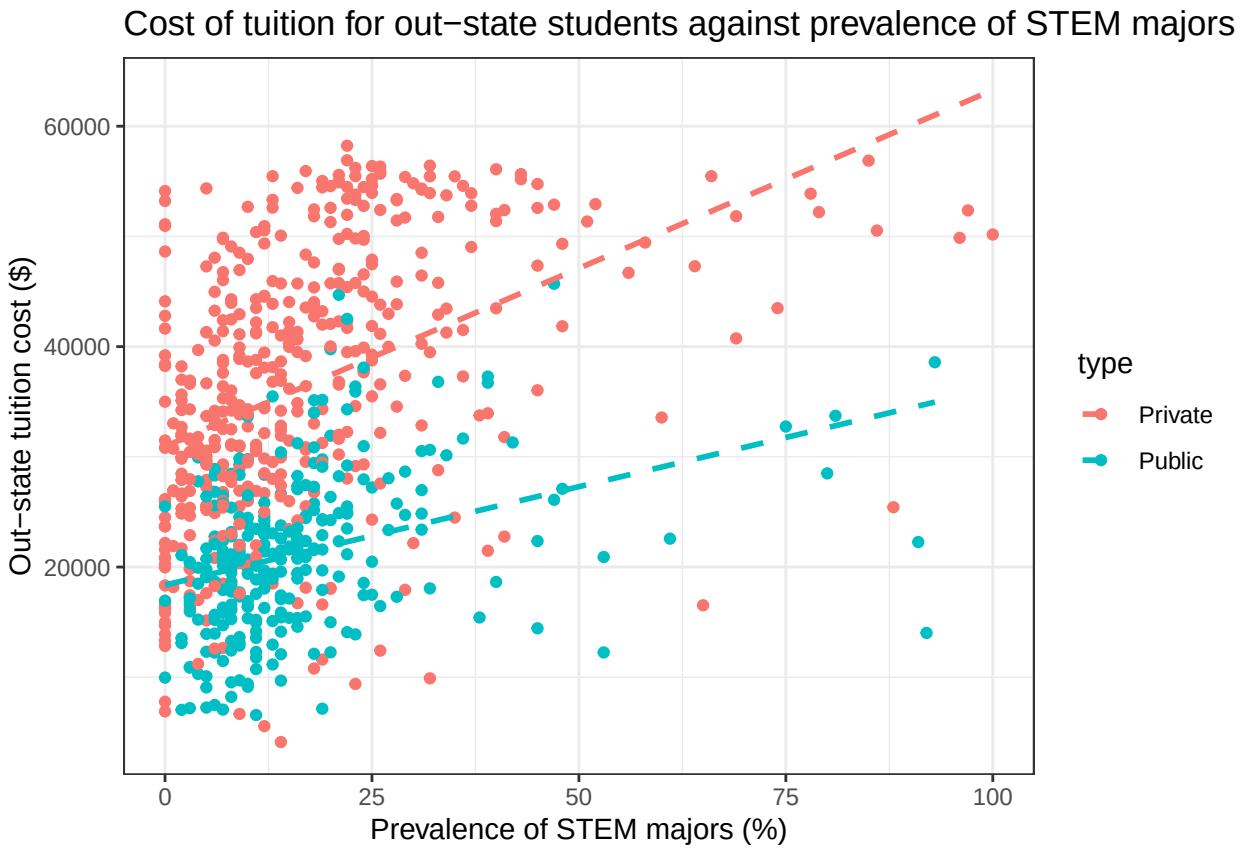

```
correlationsIS <- tcFacJoinSp %>%
  group_by(type) %>%
  summarise(r = cor(stem_percent, in_state_tuition))
correlationsIS
```

```
## # A tibble: 2 x 2
##   type        r
##   <chr>   <dbl>
## 1 Private 0.433
## 2 Public  0.265
```

Moderately weak/moderate correlation between variables given institution type.

```
ggplot(tcFacJoinSp, aes(x=stem_percent,y=out_of_state_tuition,color=type)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE,lty=2) +
  ggtitle("Cost of tuition for out-state students against prevalence of STEM majors")+
  xlab("Prevalence of STEM majors (%)")+
  ylab("Out-state tuition cost ($)")+
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
correlationsOOS <- tcFacJoinSp %>%
  group_by(type) %>%
  summarise(r = cor(stem_percent, out_of_state_tuition))
correlationsOOS
```

```
## # A tibble: 2 x 2
##   type        r
##   <chr>   <dbl>
## 1 Private 0.433
## 2 Public  0.356
```

Correlations show that there is a moderate correlation between the two variables given type.

These two graphs support the idea that tuition tends to be more expensive for schools that see greater proportions of STEM enrollment. Possibly due to the expensive equipment that generally comes with academic amenities like labs and associated equipment.

```
ISSPmodel=lm(in_state_tuition~stem_percent,data=tcFacJoinSp)
OSSPmodel=lm(out_of_state_tuition~stem_percent,data=tcFacJoinSp)

summary(ISSPmodel)
```

```
##
## Call:
## lm(formula = in_state_tuition ~ stem_percent, data = tcFacJoinSp)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -37709 -15280   1204  13279  32147
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21970.98     855.47  25.683  < 2e-16 ***
## stem_percent   258.33      37.35   6.916 1.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15790 on 725 degrees of freedom
## Multiple R-squared:  0.06189,    Adjusted R-squared:  0.0606
## F-statistic: 47.83 on 1 and 725 DF,  p-value: 1.02e-11
```

```
summary(OSSPmodel)
```

```
##
## Call:
## lm(formula = out_of_state_tuition ~ stem_percent, data = tcFacJoinSp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38392  -8930  -1347   9181  28263
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25855.11     657.14   39.34   <2e-16 ***
## stem_percent   288.67      28.69   10.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12130 on 725 degrees of freedom
## Multiple R-squared:  0.1225, Adjusted R-squared:  0.1213
## F-statistic: 101.2 on 1 and 725 DF,  p-value: < 2.2e-16
```