Numeric Summaries

Tyler Bontrager, Ganesh Singh

2022-11-02

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      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
```

```
str(tcFactored)
```

```
## tibble [2,973 x 11] (S3: tbl_df/tbl/data.frame)
```

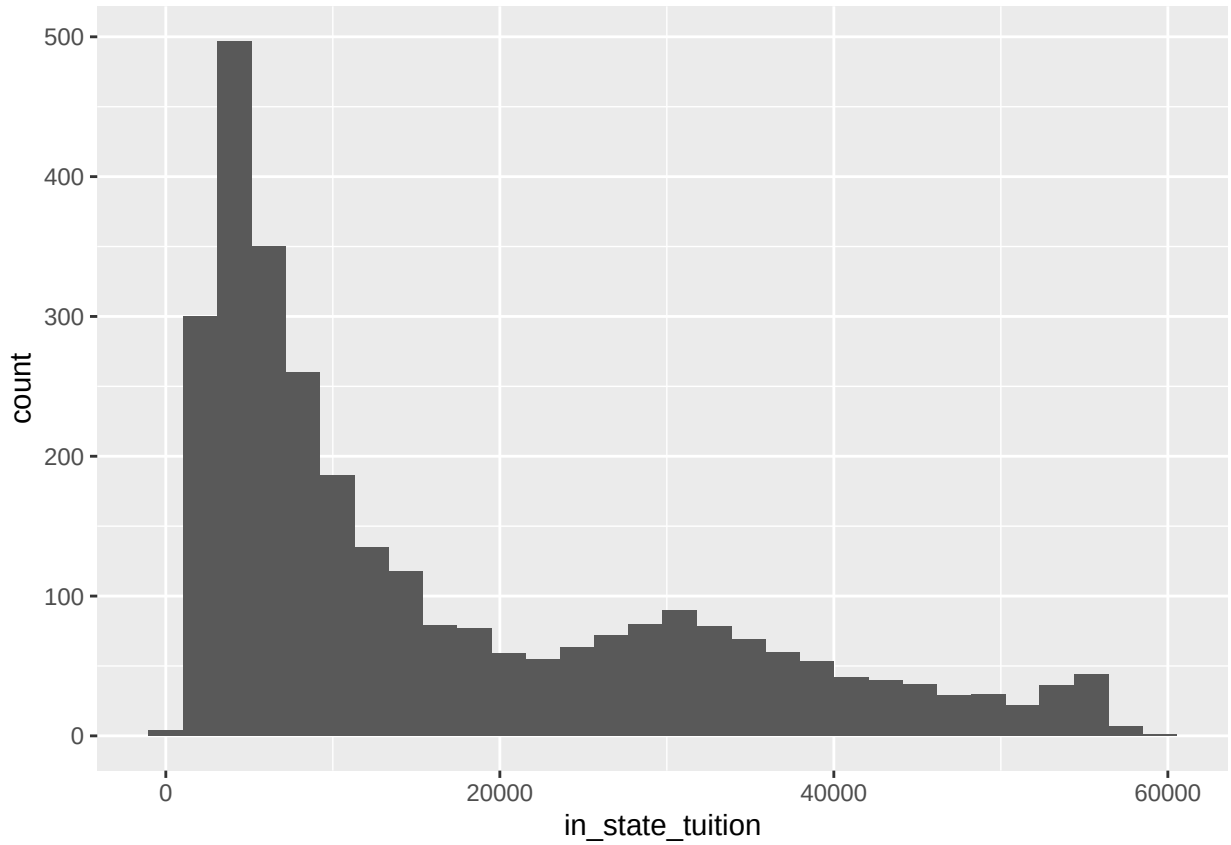
```
##  $ name                : chr [1:2973] "Aaniiih Nakoda College" "Abilene Christian University" "Abraha
##  $ state               : chr [1:2973] "Montana" "Texas" "Georgia" "Minnesota" ...
##  $ state_code          : chr [1:2973] "MT" "TX" "GA" "MN" ...
##  $ type                : chr [1:2973] "Public" "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
```

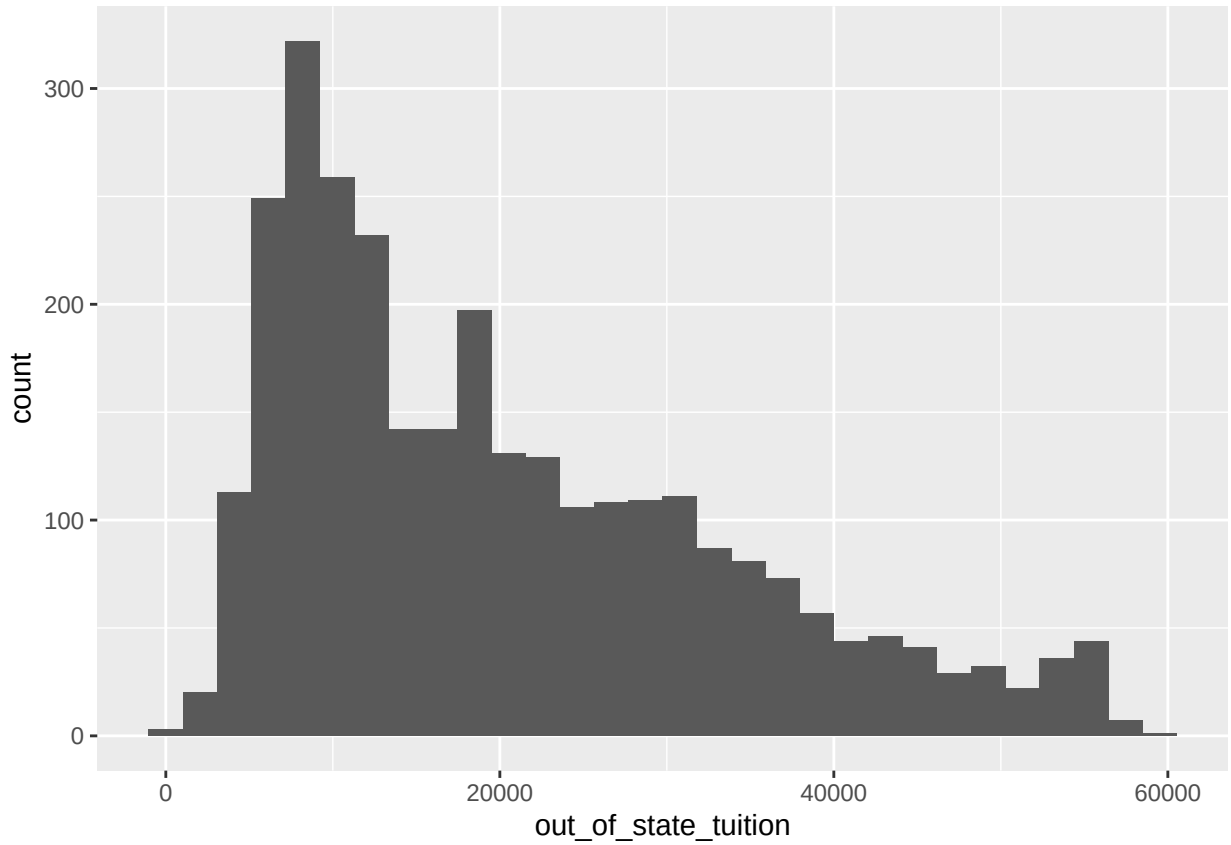
```
ggplot(tcFactored, aes(x=in_state_tuition)) + geom_histogram()
```

```
## `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`.
```

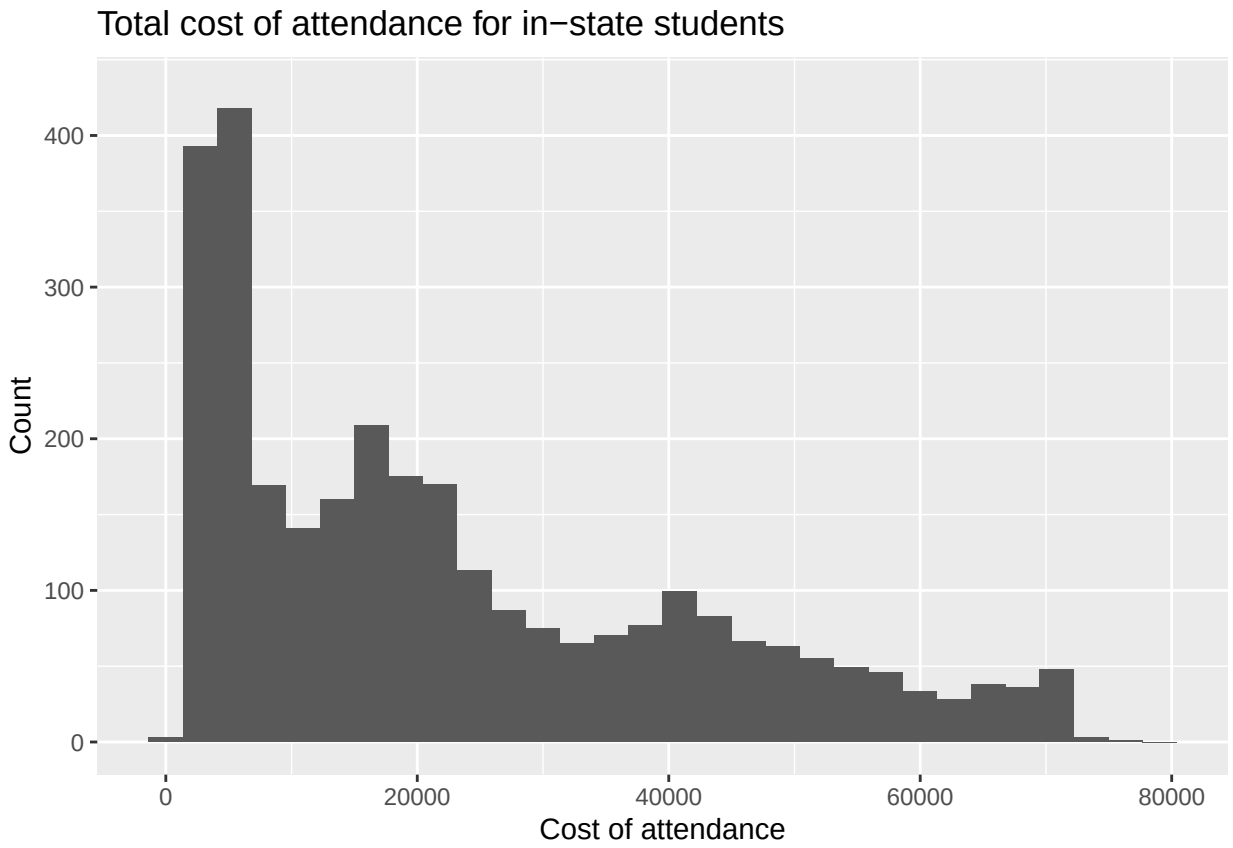



```
gatheredtc = tcFactored %>%
  gather(key="in_out", value="totalCost",c(in_state_total,out_of_state_total))
gatheredtc
```

```
## # A tibble: 5,946 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 5,936 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
```

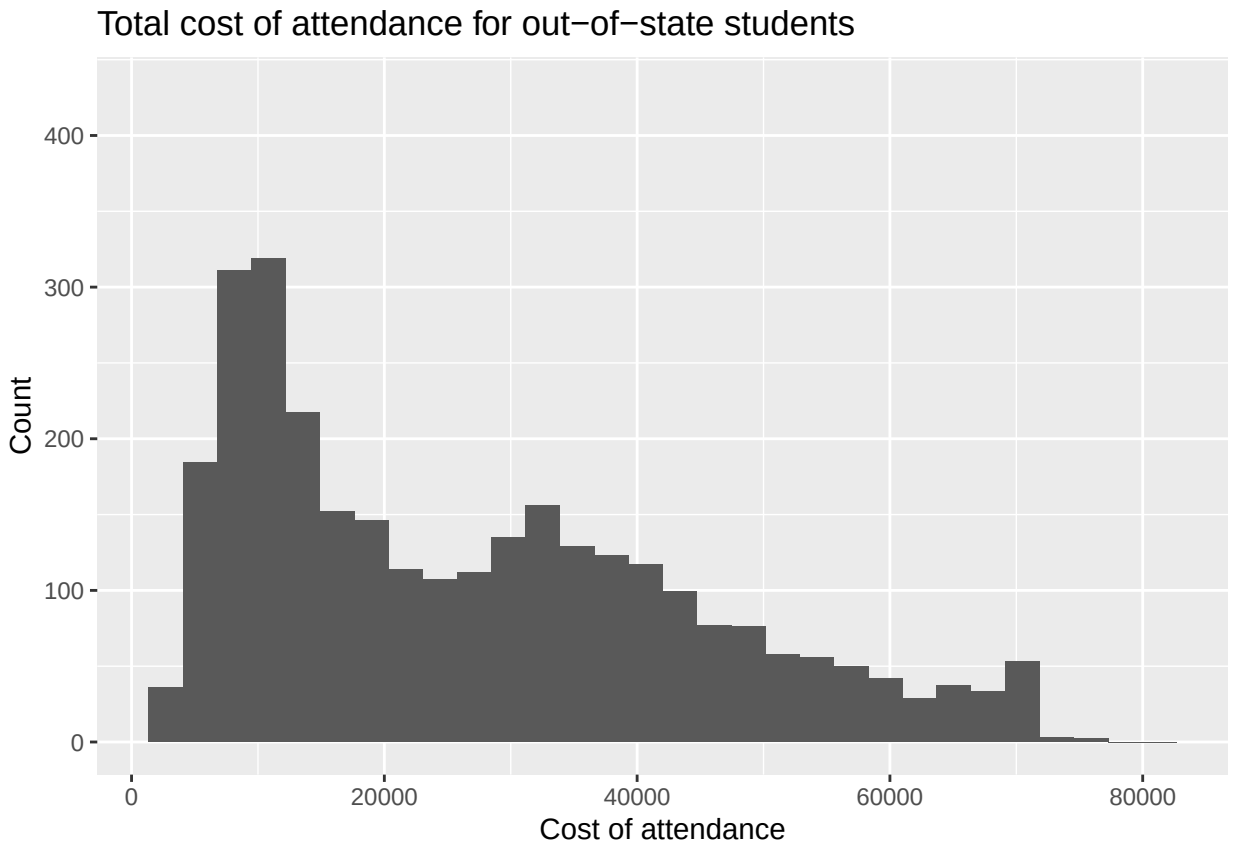
```
ggplot(tcFactored, aes(x=in_state_total))+geom_histogram()+expand_limits(x=80000,y=430) +
  ggtitle("Total cost of attendance for in-state students")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") # for the y axis label
```

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



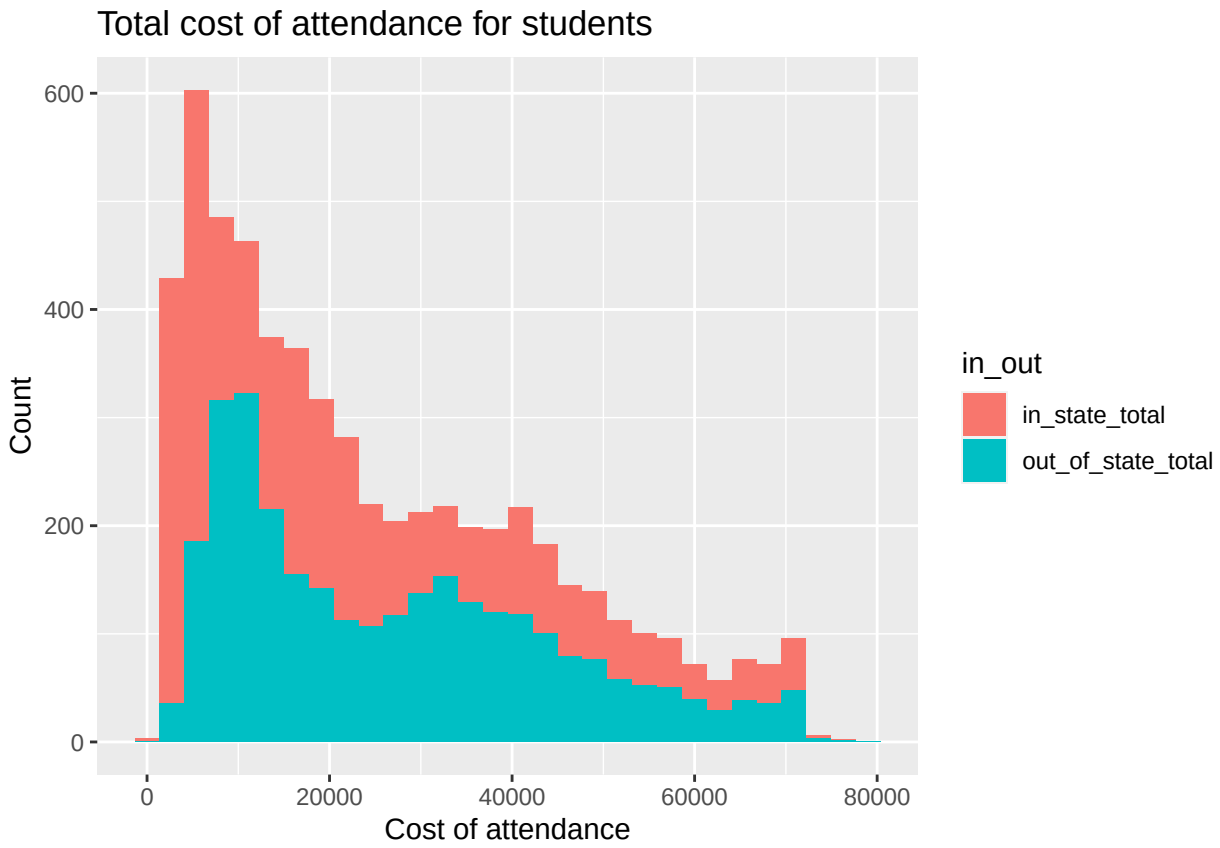

```
ggplot(tcFactored, aes(x=out_of_state_total))+geom_histogram()+expand_limits(x=80000,y=430) +
  ggtitle("Total cost of attendance for out-of-state students")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") # for the y axis label
```

```
## `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")+ # 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`.
```




```
#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, but the in-state and out-of-state costs see a shift in the primary mode. This may be because out-of-state costs are usually more expensive than in-state costs to students.

```
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
```

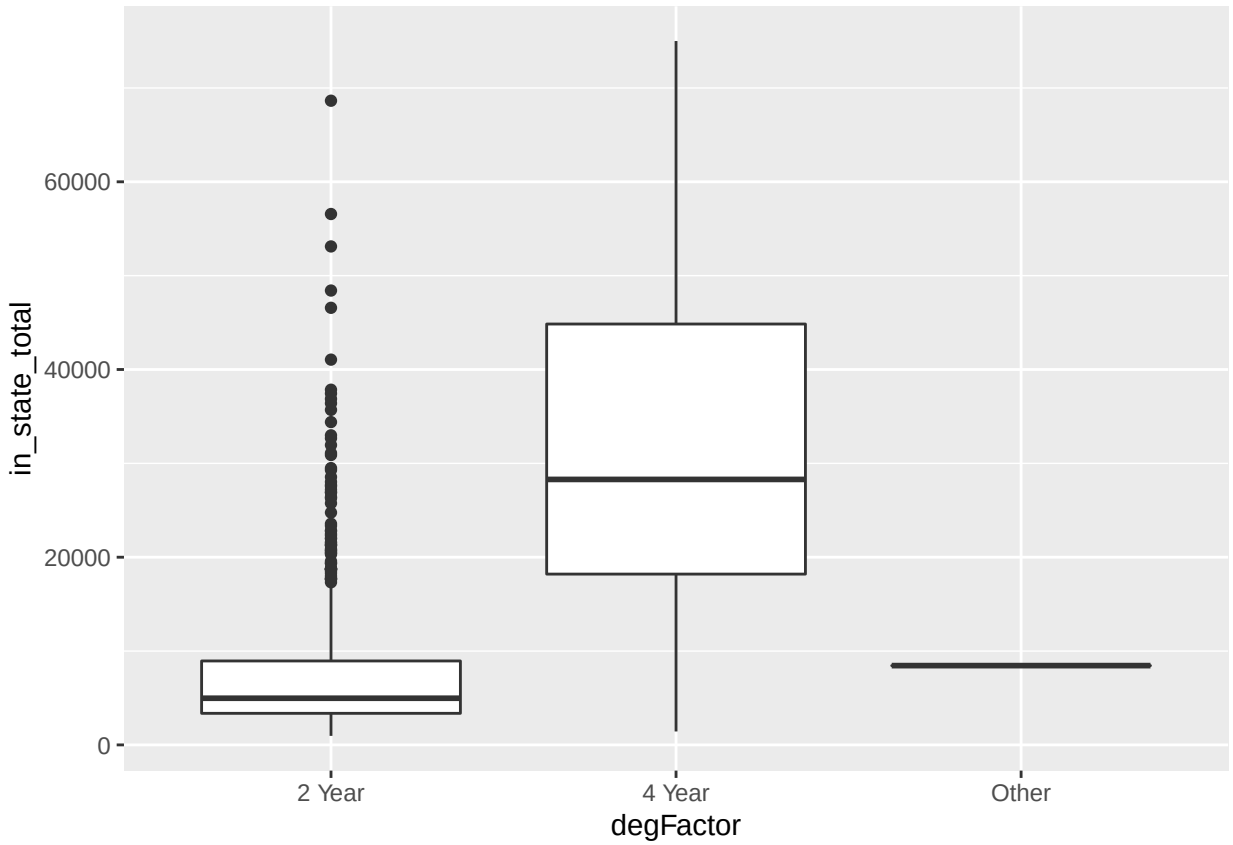
```
## # A tibble: 1 x 6
##   count_2YOOS   min    Q1   med    Q3   max
##         <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1        1120  1376 8196. 10291 13598 68640
```

```
tc2YIS_Summary
```

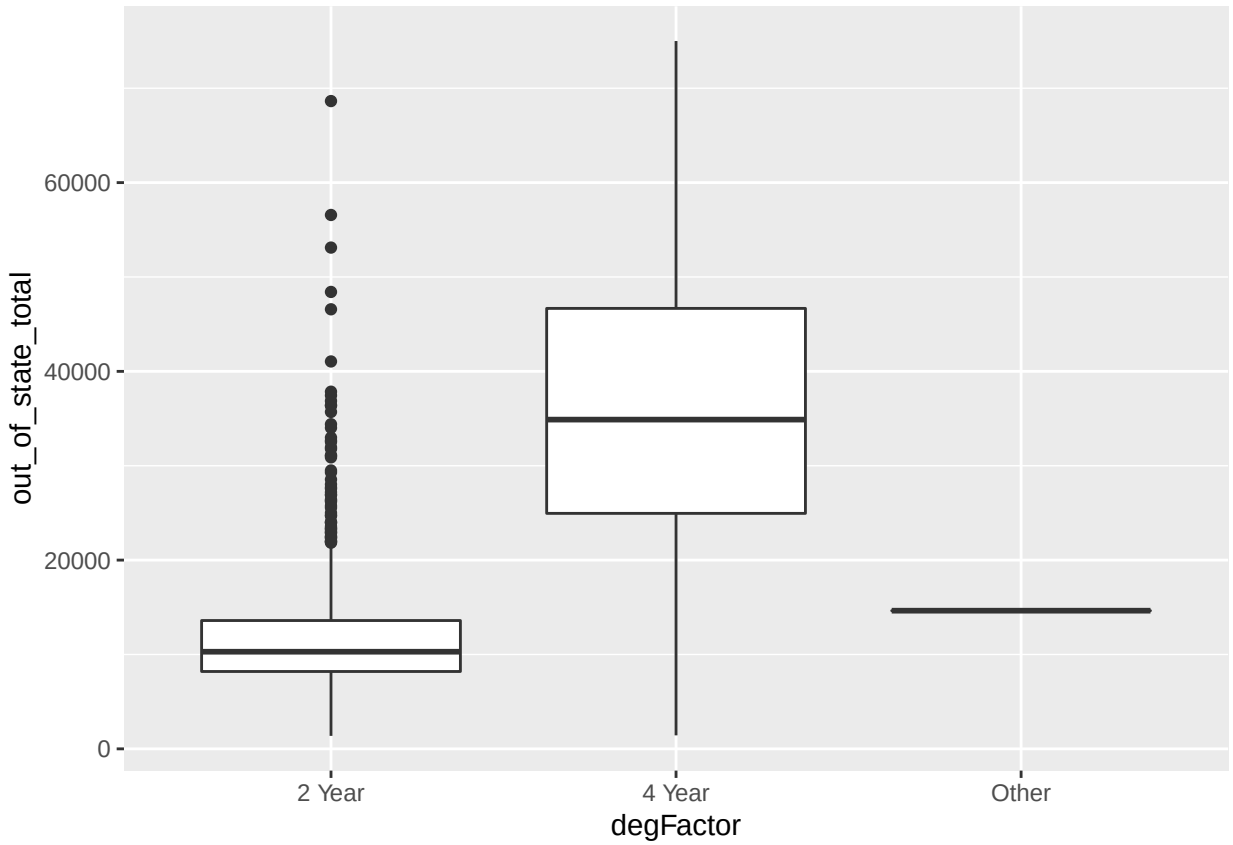
```
## # A tibble: 1 x 6
##   count_2YIS   min    Q1   med    Q3   max
##        <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1       1120   962 3364. 4972.  8946 68640
```

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()
```



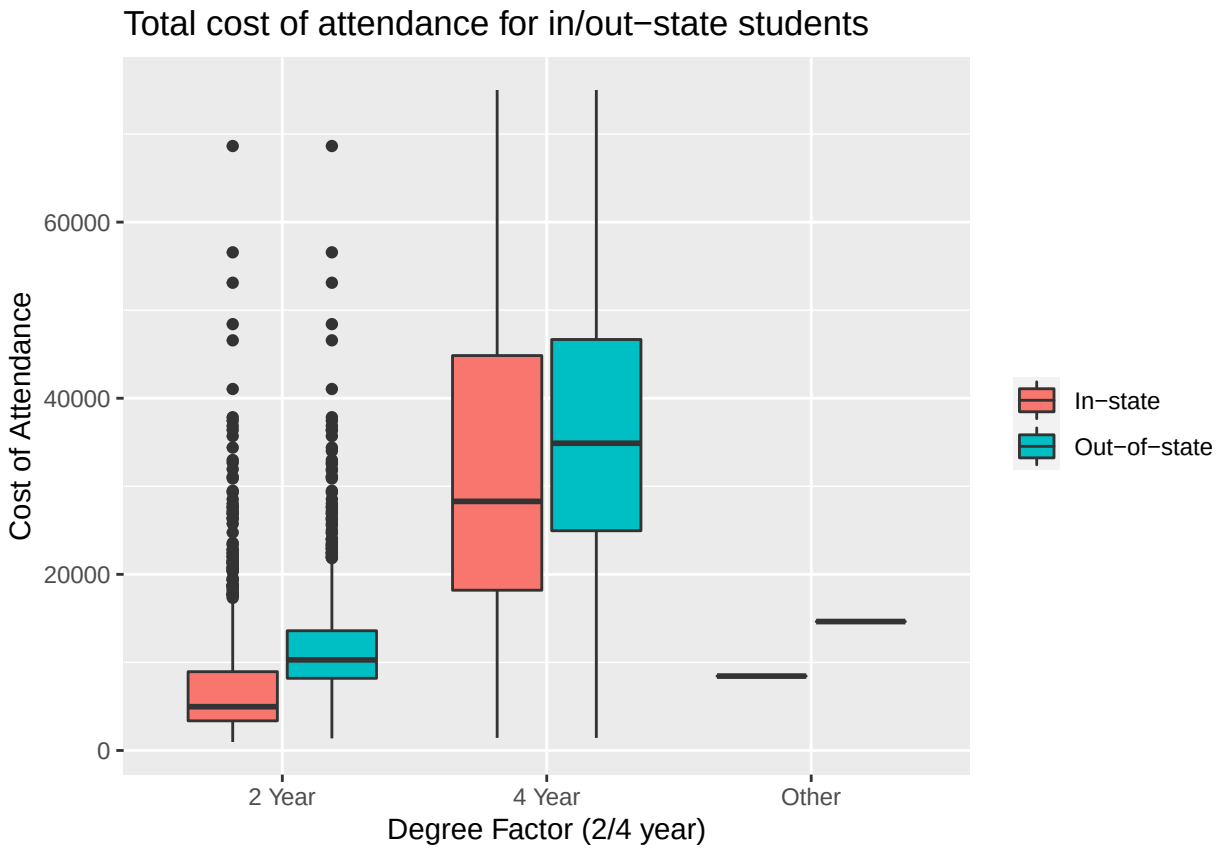

```
ggplot(tcFactored, aes(x = degFactor, y = out_of_state_total)) +  # ggplot function
  geom_boxplot()
```



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("Total cost of attendance for in/out-state students")+ # 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(
    legend.title = element_blank(),
    ) + scale_fill_discrete(name = "Student Residence", labels = c("In-state", "Out-of-state"))
```


Total cost of attendance for in/out−state students
Cost of Attendance
60000
40000
20000
0
2 Year
4 Year
Other
Degree Factor (2/4 year)
In−state
Out−of−state