

Regression

Tyler Bontrager, Ganesh Singh

2022-11-18

```
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()
library(ggplot2)
?ggplot2
```

```
# IMPORTING DATASETS
```

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

```
## 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/mast
```

```
## 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/master/data/2020/2020-01-13/historical_tuition.csv')

## 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/master/data/2020/2020-01-13/diversity_school.csv')

## 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

str(tc)

## spec_tbl_df [2,973 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ name : chr [1:2973] "Aaniiih Nakoda College" "Abilene Christian University" "Abraham Lincoln University" ...
## $ 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 ...
## - attr(*, "spec")=
## .. cols(
## .. name = col_character(),
## .. state = col_character(),
## .. state_code = col_character(),
## .. type = col_character(),
## .. degree_length = col_character(),
## .. room_and_board = col_double(),
## .. in_state_tuition = col_double(),

```

```

## .. in_state_total = col_double(),
## .. out_of_state_tuition = col_double(),
## .. out_of_state_total = col_double()
## .. )
## - attr(*, "problems")=<externalptr>

str(ti)

## spec_tbl_df [209,012 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ name      : chr [1:209012] "Piedmont International University" "Piedmont International University"
## $ state      : chr [1:209012] "NC" "NC" "NC" "NC" ...
## $ total_price: num [1:209012] 20174 20174 20174 20174 20514 ...
## $ year       : num [1:209012] 2016 2016 2016 2016 2017 ...
## $ campus     : chr [1:209012] "On Campus" "On Campus" "On Campus" "On Campus" ...
## $ net_cost   : num [1:209012] 11475 11451 16229 15592 11668 ...
## $ income_lvl : chr [1:209012] "0 to 30,000" "30,001 to 48,000" "48_001 to 75,000" "75,001 to 110,000" ...
## - attr(*, "spec")=
## .. cols(
## ..   name = col_character(),
## ..   state = col_character(),
## ..   total_price = col_double(),
## ..   year = col_double(),
## ..   campus = col_character(),
## ..   net_cost = col_double(),
## ..   income_lvl = col_character()
## .. )
## - attr(*, "problems")=<externalptr>

str(sp)

## spec_tbl_df [935 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ rank      : num [1:935] 1 2 3 4 5 6 7 8 9 10 ...
## $ name      : chr [1:935] "Auburn University" "University of Alabama in Huntsville" ...
## $ state_name : chr [1:935] "Alabama" "Alabama" "Alabama" "Alabama" ...
## $ early_career_pay : num [1:935] 54400 57500 52300 54500 48400 46600 49100 48600 47700 48700 ...
## $ mid_career_pay : num [1:935] 104500 103900 97400 93500 90500 ...
## $ make_world_better_percent: num [1:935] 51 59 50 61 52 53 48 57 56 58 ...
## $ stem_percent : num [1:935] 31 45 15 30 3 12 27 17 17 20 ...
## - attr(*, "spec")=
## .. cols(
## ..   rank = col_double(),
## ..   name = col_character(),
## ..   state_name = col_character(),
## ..   early_career_pay = col_double(),
## ..   mid_career_pay = col_double(),
## ..   make_world_better_percent = col_double(),
## ..   stem_percent = col_double()
## .. )
## - attr(*, "problems")=<externalptr>

str(ht)

## spec_tbl_df [270 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ type      : chr [1:270] "All Institutions" "All Institutions" "All Institutions" ...
## $ year      : chr [1:270] "1985-86" "1985-86" "1985-86" "1985-86" ...
## $ tuition_type: chr [1:270] "All Constant" "4 Year Constant" "2 Year Constant" "All Current" ...

```

```
## $ tuition_cost: num [1:270] 10893 12274 7508 4885 5504 ...
## - attr(*, "spec")=
## .. cols(
## ..   type = col_character(),
## ..   year = col_character(),
## ..   tuition_type = col_character(),
## ..   tuition_cost = col_double()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(ds)
```

```
## spec_tbl_df [50,655 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ name          : chr [1:50655] "University of Phoenix-Arizona" "University of Phoenix-Arizona" "
## $ total_enrollment: num [1:50655] 195059 195059 195059 195059 195059 ...
## $ state          : chr [1:50655] "Arizona" "Arizona" "Arizona" "Arizona" ...
## $ category       : chr [1:50655] "Women" "American Indian / Alaska Native" "Asian" "Black" ...
## $ enrollment     : num [1:50655] 134722 876 1959 31455 13984 ...
## - attr(*, "spec")=
## .. cols(
## ..   name = col_character(),
## ..   total_enrollment = col_double(),
## ..   state = col_character(),
## ..   category = col_character(),
## ..   enrollment = col_double()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
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
```

```
tcFacJoinSp = tcFactored %>%
  inner_join(sp, by=c("name"="name"))
tcFacJoinSp
```

```
## # A tibble: 728 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 718 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
```

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
```

sp

```
## # A tibble: 935 x 7
##   rank name      state-1 early~2 mid_c~3 make_~4 stem_~5
##   <dbl> <chr>      <chr>      <dbl>      <dbl>      <dbl>      <dbl>
## 1     1 Auburn University Alabama  54400  104500      51      31
## 2     2 University of Alabama in Hunts~ Alabama  57500  103900      59      45
## 3     3 The University of Alabama Alabama  52300   97400      50      15
## 4     4 Tuskegee University Alabama  54500   93500      61      30
## 5     5 Samford University Alabama  48400   90500      52       3
## 6     6 Spring Hill College Alabama  46600   89100      53      12
## 7     7 Birmingham Southern College Alabama  49100   88300      48      27
## 8     8 University of Alabama at Birmi~ Alabama  48600   87200      57      17
## 9     9 University of South Alabama Alabama  47700   86400      56      17
## 10    10 Alabama A&M University Alabama  48700   83500      58      20
## # ... with 925 more rows, and abbreviated variable names 1: state_name,
## #   2: early_career_pay, 3: mid_career_pay, 4: make_world_better_percent,
## #   5: stem_percent
```

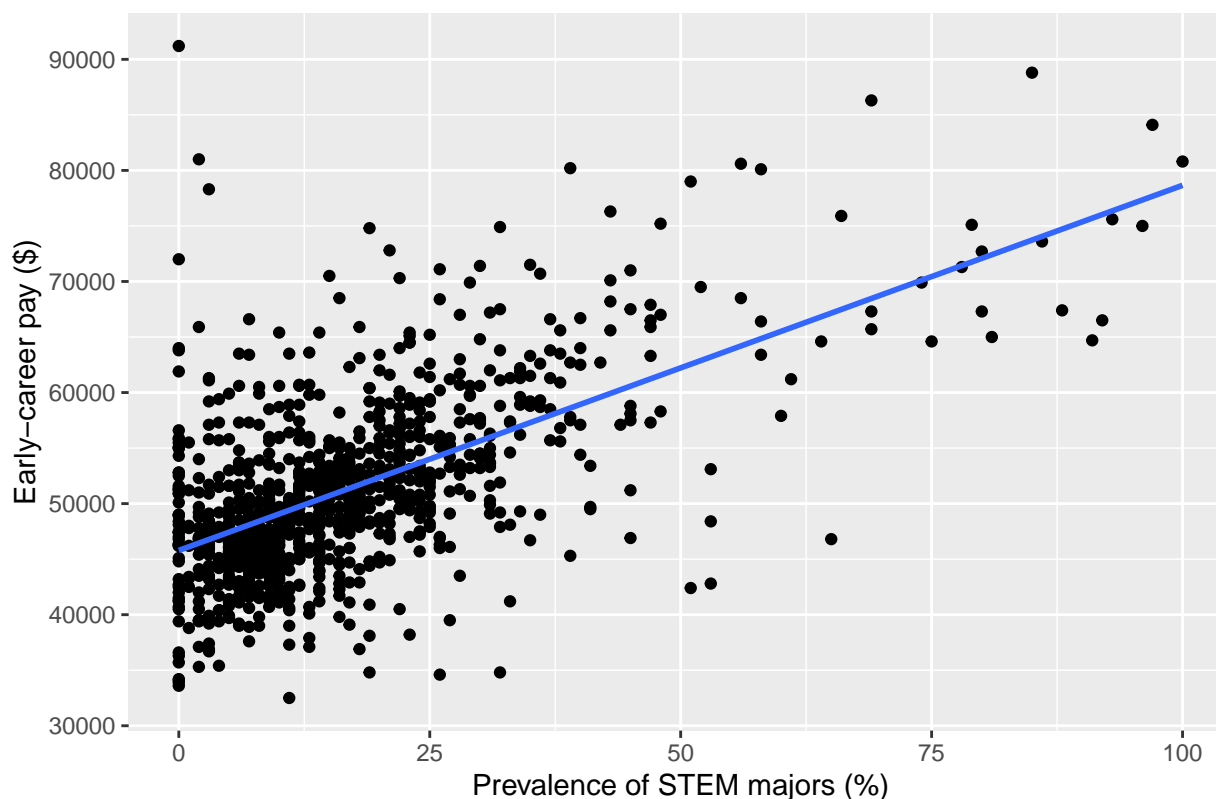
```
tcFacJoinSp
```

```
## # A tibble: 728 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 718 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 ($)")
```

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

Early career pay against prevalence of STEM majors in a school



```
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
## -21130  -4042   -636    3052   34769
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45584.50     353.63   128.91  <2e-16 ***
## stem_percent    323.30       15.45    20.93  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6533 on 726 degrees of freedom
## Multiple R-squared:  0.3763, Adjusted R-squared:  0.3754
## F-statistic:  438 on 1 and 726 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
```

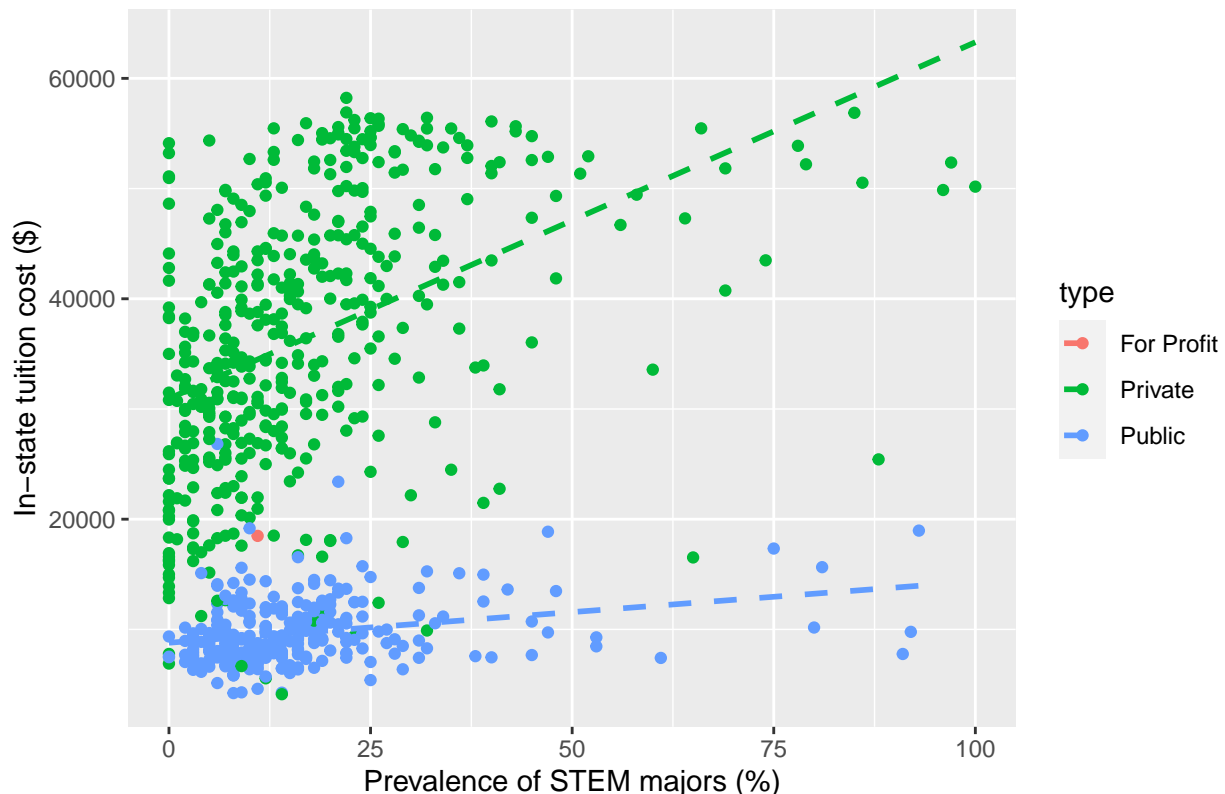
Moderately strong positive association between prevalence of STEM majors at a school and general alumni early-career earnings.

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,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 ($)")
```

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

Cost of tuition for in-state students against prevalence of STEM majors

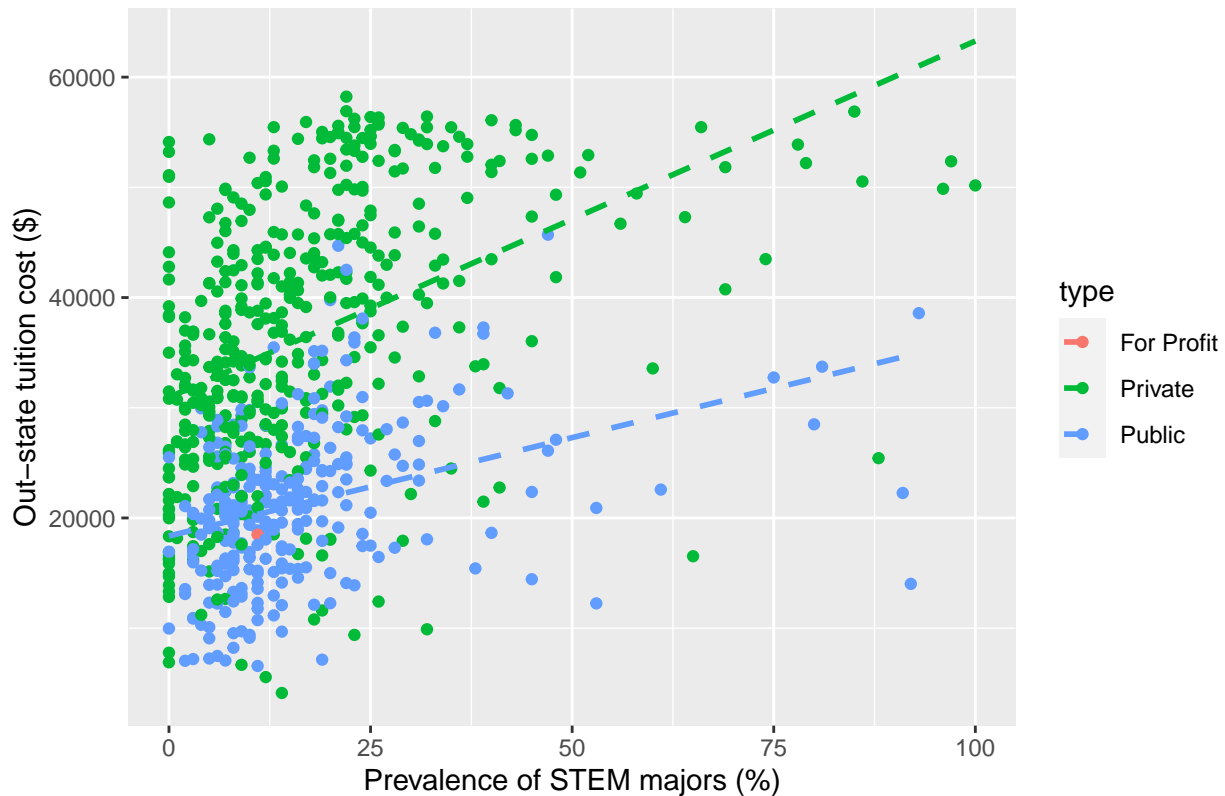


(This graph doesn't show what was seen first for your convenience, but I will be sure to remove the color for the presentation)

```
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 ($)")
```

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


Cost of tuition for out-state students against prevalence of STEM majors



```
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
## -37716 -15270   1182  13232  32159
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21958.94     854.44   25.700 < 2e-16 ***
## stem_percent    258.53      37.33    6.926 9.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15790 on 726 degrees of freedom
## Multiple R-squared:  0.06198,    Adjusted R-squared:  0.06069
## F-statistic: 47.97 on 1 and 726 DF,  p-value: 9.539e-12
```

```
summary(OSSPmodel)
```

```
##
## Call:
```

```
## lm(formula = out_of_state_tuition ~ stem_percent, data = tcFacJoinSp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -38402  -8987  -1332   9183  28283
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25835.03     656.62   39.34  <2e-16 ***
## stem_percent   289.00       28.69   10.07  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12130 on 726 degrees of freedom
## Multiple R-squared:  0.1227, Adjusted R-squared:  0.1215
## F-statistic: 101.5 on 1 and 726 DF,  p-value: < 2.2e-16
```

In-state Slope: 258.53 Y-intercept: 21958.94

Out-of-state Slope: 289.00 Y-intercept: 25835.03

Notes for story-telling:

Hypothesis: stem_percent high => stem equipment high; cost++?

room and board not considered because r&b charged by a non-academic department, and stem is academic
in-state plot revealed cluster out-of-state students do not generally enjoy benefits of tuition discounts at public universities

Hypothesis: public => less burden on student due to public/instate status => generally less tuition, but still guessing stem equipment drives up cost.

```
cor(tcFacJoinSp$in_state_total,tcFacJoinSp$stem_percent)
```

```
## [1] 0.2781892
```

```
#duh, cor doesn't care about the clusters.
```

```
tcFacJoinSpPUB = tcFacJoinSp %>%
  filter(type=="Public")
tcFacJoinSpPUB
```

```
## # A tibble: 276 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 Alabama ~ Alab~ AL      Publ~ 4 Year      5422    11068     16490     19396     24818
## 3 Alcorn S~ Miss~ MS      Publ~ 4 Year      9608     7144     16752      7144     16752
## 4 Appalach~ Nort~ NC      Publ~ 4 Year      8304     7214     15518     22021     30325
## 5 Arkansas~ Arka~ AR      Publ~ 4 Year      7870     9068     16938     15848     23718
## 6 Armstron~ Geor~ GA      Publ~ 4 Year     11385     6384     17769     19866     31251
## 7 Athens S~ Alab~ AL      Publ~ 4 Year        NA     6810     6810     12870     12870
## 8 Auburn U~ Alab~ AL      Publ~ 4 Year     13332    11276     24608     30524     43856
## 9 Auburn U~ Alab~ AL      Publ~ 4 Year      6980    10288     17268     22048     29028
## 10 Augusta ~ Geor~ GA      Publ~ 4 Year      9640    10758     20398     29796     39436
## # ... with 266 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

tcFacJoinSpPRIV = tcFacJoinSp %>%
  filter(type=="Private")
tcFacJoinSpPRIV

## # A tibble: 451 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 Adventis~ Flor~ FL      Priv~ 4 Year    4200   15150   19350   15150   19350
## 2 Agnes Sc~ Geor~ GA      Priv~ 4 Year   12330   41160   53490   41160   53490
## 3 Alaska P~ Alas~ AK      Priv~ 4 Year    7300   20830   28130   20830   28130
## 4 Albany C~ New ~ NY      Priv~ 4 Year   10920   35105   46025   35105   46025
## 5 Albertus~ Conn~ CT      Priv~ 4 Year   13200   32060   45260   32060   45260
## 6 Albion C~ Mich~ MI      Priv~ 4 Year   12380   45775   58155   45775   58155
## 7 Allen Co~ Iowa  IA      Priv~ 4 Year    7282   19970   27252   19970   27252
## 8 Alma Col~ Mich~ MI      Priv~ 4 Year   10998   40258   51256   40258   51256
## 9 Amberton~ Texas TX      Priv~ 4 Year     NA   12840   12840   12840   12840
## 10 Amherst ~ Mass~ MA      Priv~ 4 Year   14740   56426   71166   56426   71166
## # ... with 441 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

cor(tcFacJoinSpPRIV$in_state_tuition,tcFacJoinSpPRIV$stem_percent)

## [1] 0.4328422

cor(tcFacJoinSpPUB$in_state_tuition,tcFacJoinSpPUB$stem_percent)

## [1] 0.2654226

cor(tcFacJoinSpPRIV$out_of_state_tuition,tcFacJoinSpPRIV$stem_percent)

## [1] 0.4328422

cor(tcFacJoinSpPUB$out_of_state_tuition,tcFacJoinSpPUB$stem_percent)

## [1] 0.3564587
```

Moderately weak positive association between tuition cost and prevalence of stem majors at a school. But only MODERATELY! There still is some good correlation here. Other ideas?

Observation notes:

Correlations for both seem to be generally higher than that of the one taken before splitting.

Question: what is a stronger explanation for expensive schools? Are there other factors beyond funding STEM departments?

This dataset does not contain readily-usable data on ivy-league status or reputation. What if schools use their alumni's early career salary as a selling point to justify higher prices?