Project title

Report

```r
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   1.0.0
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.5.0
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()
```

```r
library(tidymodels)
```

```
-- Attaching packages -------------------------------------- tidymodels 1.0.0 --
v broom        1.0.2     v rsample      1.1.1
v dials        1.1.0     v tune         1.0.1
v infer        1.0.4     v workflows    1.1.2
v modeldata    1.0.1     v workflowsets 1.0.0
v parsnip      1.0.3     v yardstick    1.1.0
v recipes      1.0.3
-- Conflicts ----------------------------------------- tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Search for functions across packages at https://www.tidymodels.org/find/
```

```r
EADA_2021 = read.csv("data/Copy of EADA_2021.csv")


eada <- EADA_2021[,-27:-4274]
eada <- eada |>
  mutate(ratio_mw_sa = STUAID_MEN_RATIO/STUAID_WOMEN_RATIO)
eada <- eada |>
  mutate(ratio_mw_ep = RECRUITEXP_MEN/RECRUITEXP_WOMEN)
eada <- eada |>
  filter(ratio_mw_sa != "NaN",
         ratio_mw_ep != "NaN",
         ratio_mw_sa != "Inf",
         ratio_mw_ep != "Inf")
glimpse(eada)
```

```
Rows: 1,080
Columns: 28
$ unitid             <int> 100654, 100663, 100706, 100724, 100751, 100830, 10~
$ institution_name   <chr> "Alabama A & M University", "University of Alabama~
$ addr1_txt          <chr> "4900 Meridian Street", "Administration Bldg Suite~
$ addr2_txt          <chr> "", "", "", "", "", "", "", "", "", "", "", "", ""~
$ city_txt           <chr> "Normal", "Birmingham", "Huntsville", "Montgomery"~
$ state_cd           <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "A~
$ zip_text           <int> 35762, 352940110, 35899, 361040271, 354870100, 361~
$ ClassificationCode <int> 2, 1, 5, 2, 1, 5, 1, 12, 9, 12, 12, 2, 12, 12, 4, ~
$ classification_name <chr> "NCAA Division I-FCS", "NCAA Division I-FBS", "NCA~
$ ClassificationOther <chr> "", "", "", "", "", "", "", "", "", "", "", "", ""~
$ EFMaleCount        <int> 1849, 3965, 3965, 1171, 12336, 1103, 11348, 1116, ~
$ EFFemaleCount      <int> 2814, 6387, 2753, 2118, 15387, 2270, 11060, 1625, ~
$ EFTotalCount       <int> 4663, 10352, 6718, 3289, 27723, 3373, 22408, 2741,~
$ sector_cd          <int> 1, 1, 1, 1, 1, 1, 1, 4, 2, 4, 4, 1, 4, 4, 1, 4, 4,~
$ sector_name        <chr> "Public, 4-year or above", "Public, 4-year or abov~
$ STUDENTAID_MEN     <int> 2120835, 4337701, 1596000, 2536436, 9183743, 99818~
$ STUDENTAID_WOMEN   <int> 1649771, 3327452, 1378163, 1795545, 7986599, 13023~
$ STUDENTAID_COED    <int> NA, 91511, NA, NA, NA, NA, NA, NA, NA, NA, NA, 727~
$ STUDENTAID_TOTAL   <int> 3770606, 7756664, 2974163, 4331981, 17170342, 2300~
$ STUAID_MEN_RATIO   <int> 56, 56, 54, 59, 53, 43, 54, 50, 66, 42, 64, 57, 53~
$ STUAID_WOMEN_RATIO <int> 44, 43, 46, 41, 47, 57, 46, 50, 34, 58, 36, 41, 47~
$ STUAID_COED_RATIO  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 5, 0, 0,~
$ RECRUITEXP_MEN     <int> 51495, 39209, 16446, 44407, 637794, 7054, 166893, ~
$ RECRUITEXP_WOMEN   <int> 42965, 16144, 12261, 34022, 96605, 11420, 117118, ~
$ RECRUITEXP_COED    <int> NA, 0, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, ~
```

```
$ RECRUITEXP_TOTAL    <int> 94460, 55353, 28707, 78429, 734399, 18474, 284011,~
$ ratio_mw_sa         <dbl> 1.2727273, 1.3023256, 1.1739130, 1.4390244, 1.1276~
$ ratio_mw_ep         <dbl> 1.1985337, 2.4287042, 1.3413262, 1.3052437, 6.6020~
```

Introduction

Context

Duke University is intrinsically connected to its sports program. Most famously known for their Men's Division 1 Basketball and Football teams, as of 2020, Duke makes an estimated $33,382,946 and $39,669,829 in revenue, respectively (Laster, Jordan, et al. "The Haves and the Have-Nots: Inequities Amongst Duke University's Sports Teams." The Haves and the Have-Nots: Inequities Amongst Duke University's Sports Teams | Black In Blue, 2020, https://blackinblue.trinity.duke.edu/haves-and-have-nots-inequities-amongst-duke-universitys-sports-teams.). Furthermore, Duke spends approximately $1,329,949 per Men's basketball player and $217,047 per football player (Ibid.). Duke's 2020 Equity in Athletics Disclosure Act (EADA) revealed that Duke spent a total of $57,114,331 on men's sports teams and a total of $25,472,405 on women's sports teams (Ibid.) This figure includes recruitment cost, amongst other things. Nationwide, there is a recorded inequity between NCAA sports (Meredith, L. (2017, June 21). NCAA Title IX Report: Spending up, gender gaps remain. The Seattle Times. https://www.seattletimes.com/sports/ncaa-title-ix-report-spending-up-gender-gaps-remain/). As Duke students, we were curious: is this the status quo in collegiate athletics?

The Dataset

The dataset we use to explore this topic comes from the U.S. Department of Education's Equity in Athletics Disclosure Act, which requires many schools to provide information regarding financial expenditure and revenue generation of their athletics departments. The dataset spans 4274 variables, including school names, locations, and cost breakdowns. There are 1844 observations in this dataset; each observation is a different school. The most recently published dataset is from 2021, which is the dataset we use.

Explanation of Topic

Our topic will specifically observe and investigate how much, on average, universities spend on recruitment. We will create a model that predicts this figure, based on the current a university's expenditure on athletic scholarships for student athletes. We will break down a number of variables within the data, looking at whether the institution is public versus private, has a football program or not, and what athletic division it falls under (I, II, or III). Our research question is: To what extent can college athletics' recruitment funds be predicted by the level of athletic scholarships provided to student athletes? We hypothesize that the ratio will, to a high degree, predict the recruitment funding — as schools who allocate more money into students' financial aid will be willing to designate more money for their recruitment.

Ethical Issues

There are very few ethical issues anticipated with this investigation. However, it is important to note that source biases are an likely phenomenon. We must be careful with many self-reported information from university sites, and assuring that the data published can be verified.

Methodology

Research question: To what extent can college athletics' recruitment funds be predicted by the level of athletic scholarships provided to student athletes?

- Private vs. public
- Football vs. no football
- Division

We will answer our research question by first creating subsets of our data by the factors listed above, and removing any outliers using the interquartile range rule. Then we create regression models comparing athletic scholarships to recruitment funds, and then calculate the AIC for each model to determine the effectiveness of each predictor. We chose to use regression models because our research question looks at the extent to which one variable can predict another. We chose to use AIC calculations because we are comparing the effectiveness of prediction based on which variables are included in the model, and AIC is an effective calculation for comparing models that is not simply dependent on the number of predictors.

```
eada_NCAA <- eada |>
  filter(classification_name == "NCAA Division I-FCS" |
        classification_name == "NCAA Division I-FBS" |
        classification_name == "NCAA Division I without football" |
        classification_name == "NCAA Division II with football" |
        classification_name == "NCAA Division II without football" |
        classification_name == "NCAA Division III with football" |
        classification_name == "NCAA Division III without football")


Q1_NCAA_ep <- quantile(eada_NCAA$ratio_mw_ep, .25)
Q1_NCAA_sa <- quantile(eada_NCAA$ratio_mw_sa, .25)

Q3_NCAA_ep <- quantile(eada_NCAA$ratio_mw_ep, .75)
Q3_NCAA_sa <- quantile(eada_NCAA$ratio_mw_sa, .75)

IQR_NCAA_ep <- IQR(eada_NCAA$ratio_mw_ep)
IQR_NCAA_sa <- IQR(eada_NCAA$ratio_mw_sa)

eada_NCAA <- subset(eada_NCAA, eada_NCAA$ratio_mw_ep> (Q1_NCAA_ep - 1.5*IQR_NCAA_ep) & ead
```

```
eada_NCAA <- subset(eada_NCAA, eada_NCAA$ratio_mw_sa> (Q1_NCAA_sa - 1.5*IQR_NCAA_sa) & ead
```

```
eada_private <- eada_NCAA |>
  filter(sector_name == "Private nonprofit, 4-year or above" |
         sector_name == "Private for-profit, 4-year or above")



Q1_priv_ep <- quantile(eada_private$ratio_mw_ep, .25)
Q1_priv_sa <- quantile(eada_private$ratio_mw_sa, .25)

Q3_priv_ep <- quantile(eada_private$ratio_mw_ep, .75)
Q3_priv_sa <- quantile(eada_private$ratio_mw_sa, .75)

IQR_priv_ep <- IQR(eada_private$ratio_mw_ep)
IQR_priv_sa <- IQR(eada_private$ratio_mw_sa)

eada_private <- subset(eada_private, eada_private$ratio_mw_ep> (Q1_priv_ep - 1.5*IQR_priv_

eada_private <- subset(eada_private, eada_private$ratio_mw_sa> (Q1_priv_sa - 1.5*IQR_priv_
```

```
eada_public <- eada_NCAA |>
  filter(sector_name == "Public, 4-year or above" |
         sector_name == "Public, 2-year")

Q1_public_ep <- quantile(eada_public$ratio_mw_ep, .25)
Q1_public_sa <- quantile(eada_public$ratio_mw_sa, .25)

Q3_public_ep <- quantile(eada_public$ratio_mw_ep, .75)
Q3_public_sa <- quantile(eada_public$ratio_mw_sa, .75)

IQR_public_ep <- IQR(eada_public$ratio_mw_ep)
IQR_public_sa <- IQR(eada_public$ratio_mw_sa)

eada_public <- subset(eada_public, eada_public$ratio_mw_ep> (Q1_public_ep - 1.5*IQR_public

eada_public <- subset(eada_public, eada_public$ratio_mw_sa> (Q1_public_sa - 1.5*IQR_public
```

```
eada_nofootball <- eada |>
  filter(classification_name == "NCAA Division I without football" |
```

```
        classification_name == "NCAA Division II without football" |
        classification_name == "NCAA Division III without football")

Q1_nofootball_ep <- quantile(eada_nofootball$ratio_mw_ep, .25)
Q1_nofootball_sa <- quantile(eada_nofootball$ratio_mw_sa, .25)

Q3_nofootball_ep <- quantile(eada_nofootball$ratio_mw_ep, .75)
Q3_nofootball_sa <- quantile(eada_nofootball$ratio_mw_sa, .75)

IQR_nofootball_ep <- IQR(eada_nofootball$ratio_mw_ep)
IQR_nofootball_sa <- IQR(eada_nofootball$ratio_mw_sa)

eada_nofootball <- subset(eada_nofootball, eada_nofootball$ratio_mw_ep> (Q1_nofootball_ep

eada_nofootball <- subset(eada_nofootball, eada_nofootball$ratio_mw_sa> (Q1_nofootball_sa
```

```
eada_football <- eada |>
  filter(classification_name != "NCAA Division I without football" |
         classification_name != "NCAA Division II without football" |
         classification_name != "NCAA Division III without football")

Q1_football_ep <- quantile(eada_football$ratio_mw_ep, .25)
Q1_football_sa <- quantile(eada_football$ratio_mw_sa, .25)

Q3_football_ep <- quantile(eada_football$ratio_mw_ep, .75)
Q3_football_sa <- quantile(eada_football$ratio_mw_sa, .75)

IQR_football_ep <- IQR(eada_football$ratio_mw_ep)
IQR_football_sa <- IQR(eada_football$ratio_mw_sa)

eada_football <- subset(eada_football, eada_football$ratio_mw_ep> (Q1_football_ep - 1.5*IQ

eada_football <- subset(eada_football, eada_football$ratio_mw_sa> (Q1_football_sa - 1.5*IQ
```

```
eada_divisionI <- eada |>
  filter(classification_name == "NCAA Division I-FCS" |
         classification_name == "NCAA Division I-FBS" |
         classification_name == "NCAA Division I without football")
```

```r
Q1_divisionI_ep <- quantile(eada_divisionI$ratio_mw_ep, .25)
Q1_divisionI_sa <- quantile(eada_divisionI$ratio_mw_sa, .25)

Q3_divisionI_ep <- quantile(eada_divisionI$ratio_mw_ep, .75)
Q3_divisionI_sa <- quantile(eada_divisionI$ratio_mw_sa, .75)

IQR_divisionI_ep <- IQR(eada_divisionI$ratio_mw_ep)
IQR_divisionI_sa <- IQR(eada_divisionI$ratio_mw_sa)

eada_divisionI <- subset(eada_divisionI, eada_divisionI$ratio_mw_ep> (Q1_divisionI_ep - 1.

eada_divisionI <- subset(eada_divisionI, eada_divisionI$ratio_mw_sa> (Q1_divisionI_sa - 1.
```

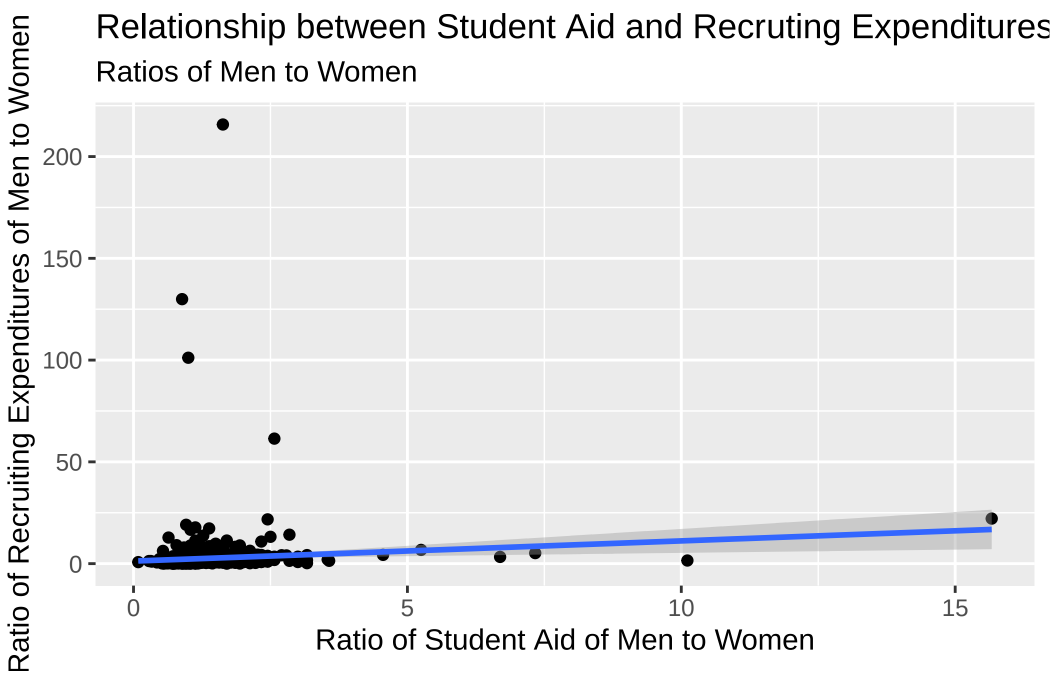
```r
eada_divisionII <- eada |>
  filter(classification_name == "NCAA Division II with football" |
        classification_name == "NCAA Division II without football")

Q1_divisionII_ep <- quantile(eada_divisionII$ratio_mw_ep, .25)
Q1_divisionII_sa <- quantile(eada_divisionII$ratio_mw_sa, .25)

Q3_divisionII_ep <- quantile(eada_divisionII$ratio_mw_ep, .75)
Q3_divisionII_sa <- quantile(eada_divisionII$ratio_mw_sa, .75)

IQR_divisionII_ep <- IQR(eada_divisionII$ratio_mw_ep)
IQR_divisionII_sa <- IQR(eada_divisionII$ratio_mw_sa)

eada_divisionII <- subset(eada_divisionII, eada_divisionII$ratio_mw_ep> (Q1_divisionII_ep

eada_divisionII <- subset(eada_divisionII, eada_divisionII$ratio_mw_sa> (Q1_divisionII_sa
```

```r
eada_divisionIII <- eada |>
  filter(classification_name == "NCAA Division III with football" |
         classification_name == "NCAA Division III without football")

Q1_divisionIII_ep <- quantile(eada_divisionIII$ratio_mw_ep, .25)
Q1_divisionIII_sa <- quantile(eada_divisionIII$ratio_mw_sa, .25)

Q3_divisionIII_ep <- quantile(eada_divisionIII$ratio_mw_ep, .75)
Q3_divisionIII_sa <- quantile(eada_divisionIII$ratio_mw_sa, .75)
```

```r
IQR_divisionIII_ep <- IQR(eada_divisionIII$ratio_mw_ep)
IQR_divisionIII_sa <- IQR(eada_divisionIII$ratio_mw_sa)

eada_divisionIII <- subset(eada_divisionIII, eada_divisionIII$ratio_mw_ep> (Q1_divisionIII

eada_divisionIII <- subset(eada_divisionIII, eada_divisionIII$ratio_mw_sa> (Q1_divisionIII
```

```r
eada |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)
  ) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Ratios of Men to Women")
```

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



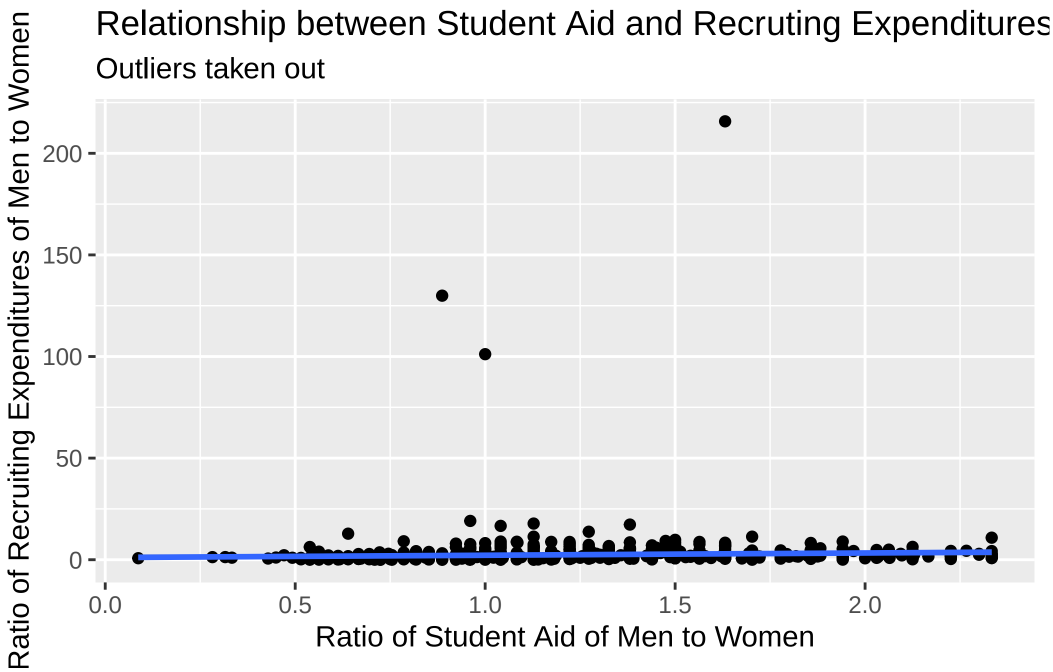
```
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada) |>
  glance() |>
  pull(AIC, r.squared)
```

```
0.00768106755277727
          7723.431
```

Outliers Taken Out

```
Q1_no_outs_ep <- quantile(eada$ratio_mw_ep, .25)
Q1_no_outs_sa <- quantile(eada$ratio_mw_sa, .25)

Q3_no_outs_ep <- quantile(eada$ratio_mw_ep, .75)
Q3_no_outs_sa <- quantile(eada$ratio_mw_sa, .75)

IQR_no_outs_ep <- IQR(eada$ratio_mw_ep)
IQR_no_outs_sa <- IQR(eada$ratio_mw_sa)

eada_no_outs <- subset(eada, eada$ratio_mw_ep> (Q1_no_outs_ep - 1.5*IQR_no_outs_ep) & eada

eada_no_outs <- subset(eada, eada$ratio_mw_sa> (Q1_no_outs_sa - 1.5*IQR_no_outs_sa) & eada

eada_no_outs |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)) +
  geom_point() + geom_smooth(method = "lm") +
  labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Outliers taken out")
```

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



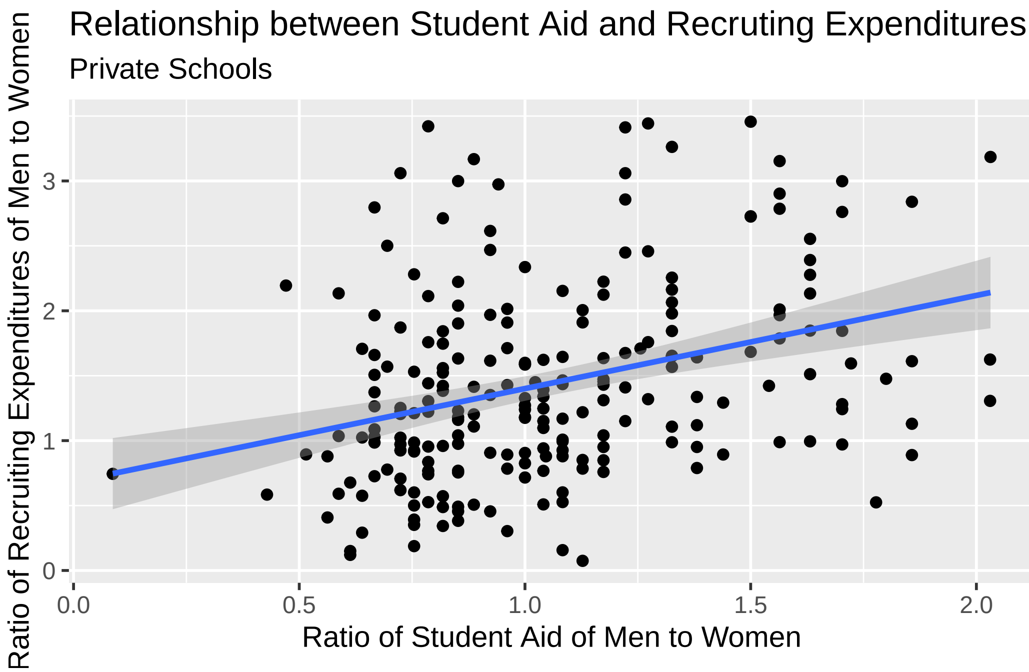

```
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_no_outs) |>
  glance() |>
  pull(AIC, r.squared)
```

```
0.00281077907311822
            7409.35
```

Private Schools

```
eada_private |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)) +
  geom_point() + geom_smooth(method = "lm") +
   labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Private Schools")
```

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




```r
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_private) |>
  glance() |>
  pull(AIC, r.squared)
```

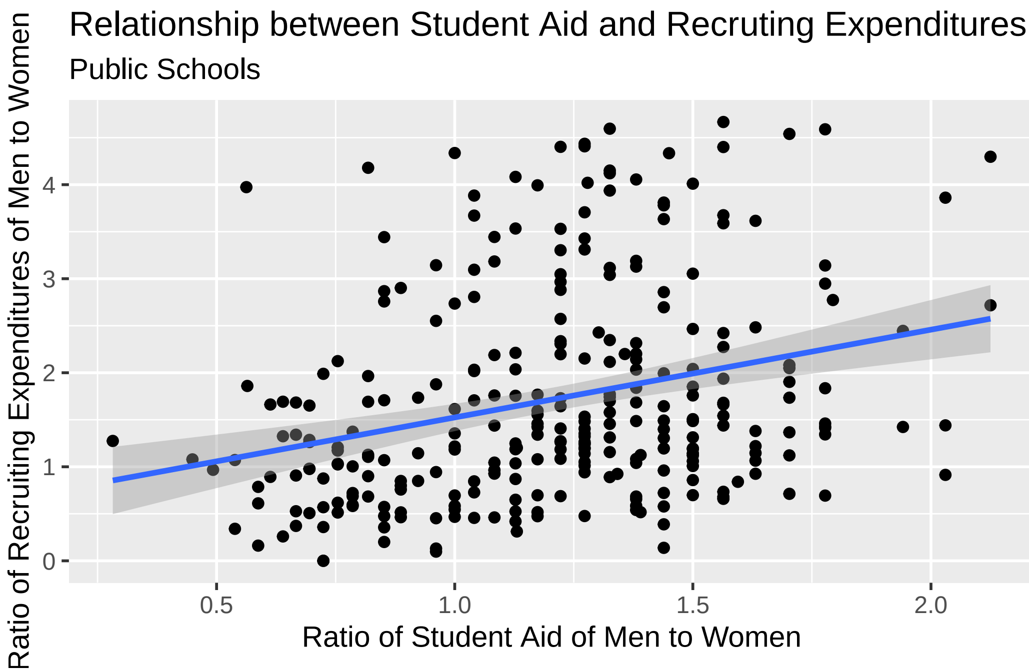
```
0.112223105145674
         490.8492
```

Public Schools

```r
eada_public |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)
  ) + geom_point() + geom_smooth(method = "lm") +
 labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
```

```
        title = "Relationship between Student Aid and Recruting Expenditures",
        subtitle = "Public Schools")
```

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




```
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_public) |>
  glance() |>
  pull(AIC, r.squared)
```

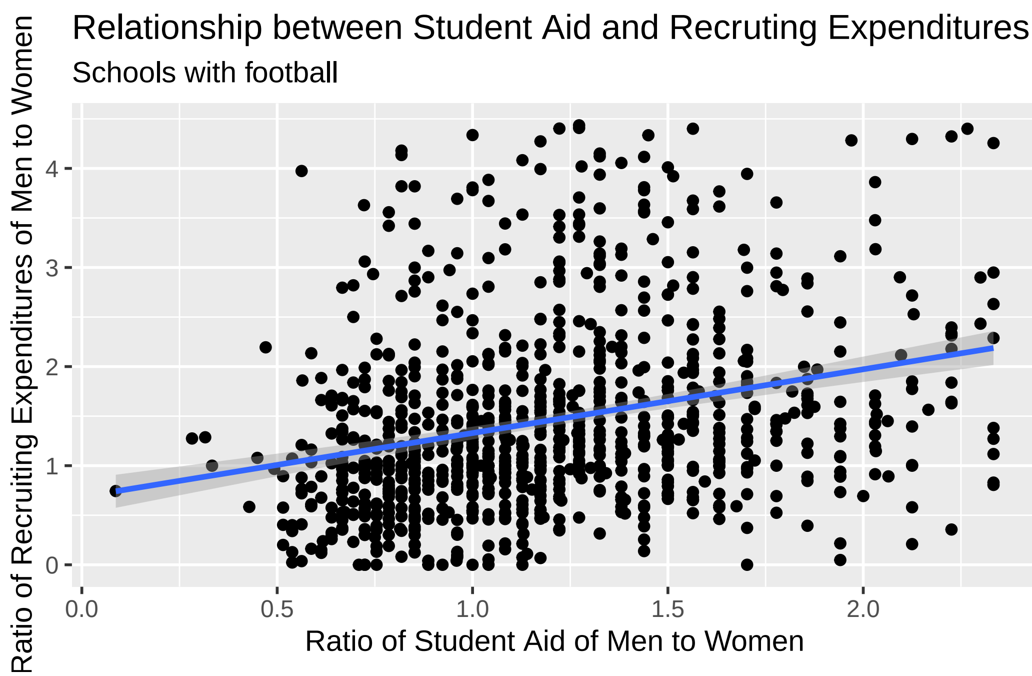
```
0.0818346252647914
          869.6874
```

Football Schools

```
eada_football |>
  ggplot(
    aes(x = ratio_mw_sa,
```

```
        y = ratio_mw_ep)
  ) + geom_point() + geom_smooth(method = "lm") +
   labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Schools with football")
```

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



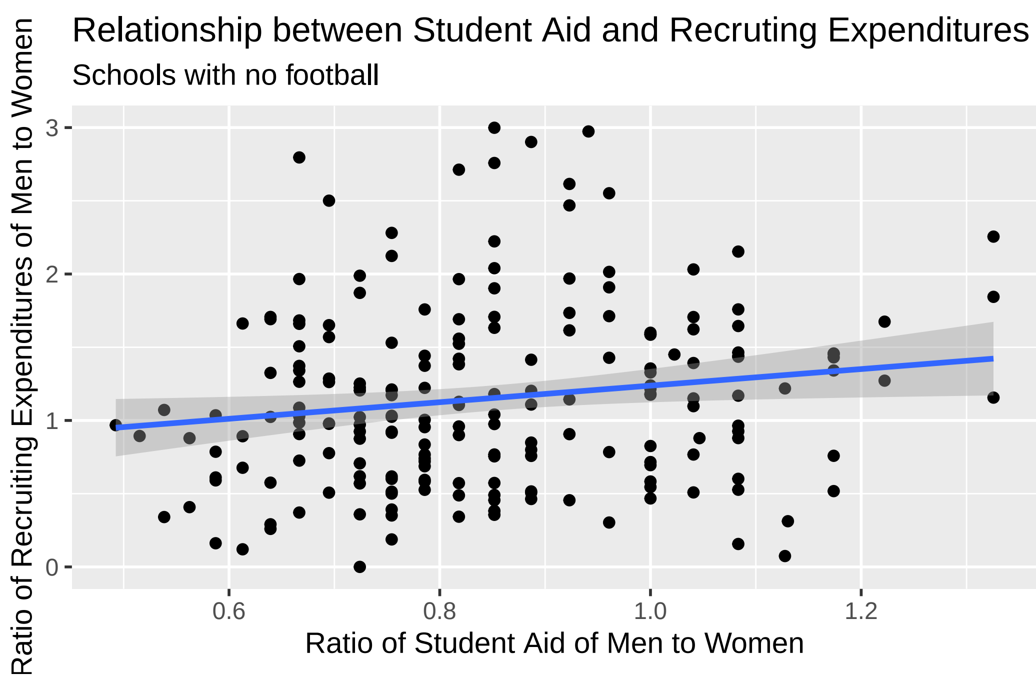

```
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_football) |>
  glance() |>
  pull(AIC, r.squared)
```

```
0.0786090754011517
          2484.094
```

Schools with No Football

```r
eada_nofootball |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)
  ) + geom_point() + geom_smooth(method = "lm") +
   labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Schools with no football")
```

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



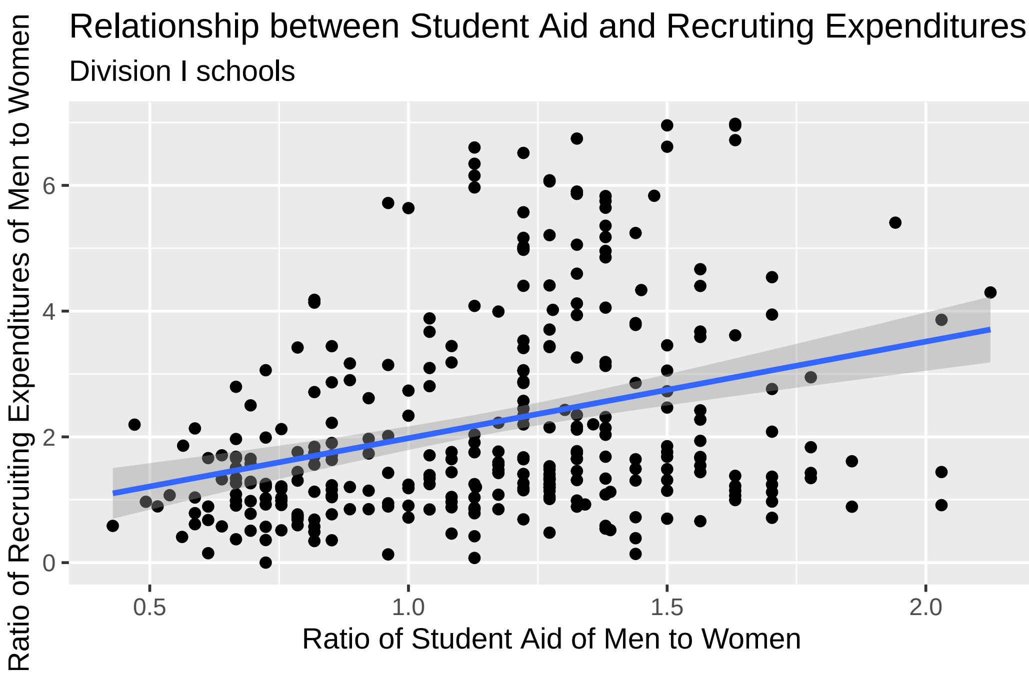

```r
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_nofootball) |>
  glance() |>
  pull(AIC, r.squared)
```

```
0.0251744256907911
          374.6477
```

Division I Schools

```
eada_divisionI |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)
  ) + geom_point() + geom_smooth(method = "lm") +
   labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Division I schools")
```

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




```
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_divisionI) |>
  glance() |>
  pull(AIC, r.squared)
```

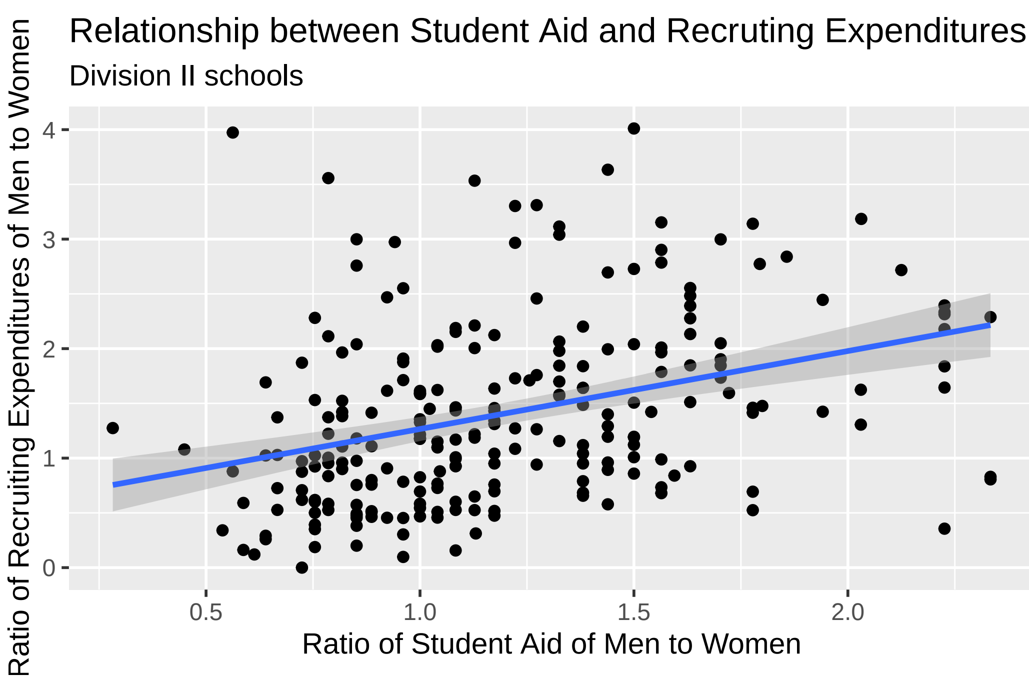
```
0.103162960388741
```

```
    1167.375
```

Division II Schools

```
eada_divisionII |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)
  ) + geom_point() + geom_smooth(method = "lm") +
   labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Division II schools")
```

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




```
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_divisionII) |>
  glance() |>
```

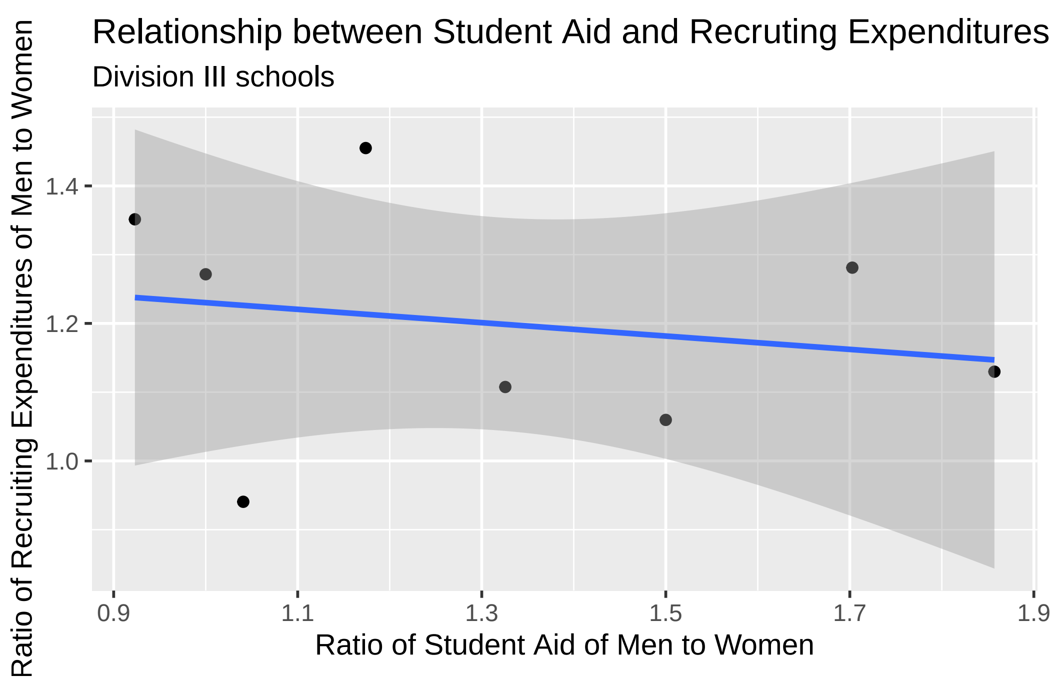
```
  pull(AIC, r.squared)
```

```
0.126985443525168
         554.4589
```

Division III Schools

```
eada_divisionIII |>
  ggplot(
    aes(x = ratio_mw_sa,
        y = ratio_mw_ep)
  ) + geom_point() + geom_smooth(method = "lm") +
   labs(x = "Ratio of Student Aid of Men to Women",
       y = "Ratio of Recruiting Expenditures of Men to Women",
       title = "Relationship between Student Aid and Recruting Expenditures",
       subtitle = "Division III schools")
```

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



```
linear_reg() |>
  set_engine("lm") |>
  fit(ratio_mw_ep ~ ratio_mw_sa, data = eada_divisionIII) |>
  glance() |>
  pull(AIC, r.squared)
```

```
0.0389448055922119
         -1.120288
```

Results

This prediction model was aimed at testing whether the financial allocation universities spent on athletic scholarship could predict the amount they allocated for recruitment. We filtered the data through three different variables: whether the school had a football program or not, whether the school was public or private, and whether the school was Division I, II, or III.

It is important to note the biases that may have effected the data outcome. Division III schools do not allocate any money towards athletic scholarships. Furthermore, financial capabilities widely vary between public and private schools, as well as, how much income the school makes from their athletic programs.

After filtering the data, we ran an AIC metric on all the subsets of the data to compare the different regression models. There was a more accurate trend observed in the private vs. public data-set, meaning that model offers the best fit. Amongst all the graphs, there is a low positive correlation observed. Then we ran an R squared test. This would tell us the significance in the correlation between the two variables: scholarship funds versus recruitment funds. The R squared test revealed that there is a very insignificant correlation between the two chosen variables.