hw1

Liming Ning

2022/1/16

Contents

Case Study 1: Audience Size

Data Preparation

cleaning

```
##  [1] "$0.05" "$0.05" "$0.05" "$0.05" "$0.05" "$0.05" "$0.05" "$0.05" "$0.05"
## [10] "$0.05"
```

```
##              age gender                                     education
## 1:                 Male                                    select one
## 2:           223   Male            High school graduate (or equivalent)
## 3:        female Female Some college, no diploma; or Associate's degree
## 4: Eighteen (18)   Male            High school graduate (or equivalent)
```

```
## 5:             4   Male       Bachelor’s degree or other 4-year degree
## 6:           27`   Male Some college, no diploma; or Associate’s degree
##              income sirius wharton worktime
## 1:                                        5
## 2: $30,000 - $50,000     No      No       11
## 3:    Above $150,000    Yes      No       21
## 4: $30,000 - $50,000    Yes      No       29
## 5: $50,000 - $75,000    Yes      No       22
## 6: Less than $15,000     No      No       20

## [1] "Female" "Male"   ""

##    age gender                                         education            income
## 1:  47                        Graduate or professional degree $30,000 - $50,000
## 2:  47                        Graduate or professional degree $50,000 - $75,000
## 3:  29        Some college, no diploma; or Associate’s degree $15,000 - $30,000
## 4:  31                        Graduate or professional degree $30,000 - $50,000
## 5:  25        Some college, no diploma; or Associate’s degree Less than $15,000
## 6:  67        Some college, no diploma; or Associate’s degree $50,000 - $75,000
##    sirius wharton worktime
## 1:    Yes      No       54
## 2:    Yes      No       15
## 3:    Yes      No       19
## 4:     No      No       15
## 5:    Yes      No       19
## 6:     No      No       32

## [1] "Some college, no diploma; or Associate’s degree"
## [2] "Graduate or professional degree"
## [3] "Bachelor’s degree or other 4-year degree"
## [4] "High school graduate (or equivalent)"
## [5] "Less than 12 years; no high school diploma"
## [6] "select one"
## [7] "Other"

## [1] "$30,000 - $50,000"  "$15,000 - $30,000"  "$50,000 - $75,000"
## [4] "Above $150,000"     "Less than $15,000"  "$75,000 - $150,000"
## [7] ""

## [1] "No"  "Yes" ""

## [1] "No"  "Yes" ""

##    age gender                            education            income sirius
## 1:  25   Male High school graduate (or equivalent) $15,000 - $30,000     No
## 2:  26   Male High school graduate (or equivalent) $15,000 - $30,000     No
##    wharton worktime
## 1:     Yes       20
## 2:     Yes       25
```

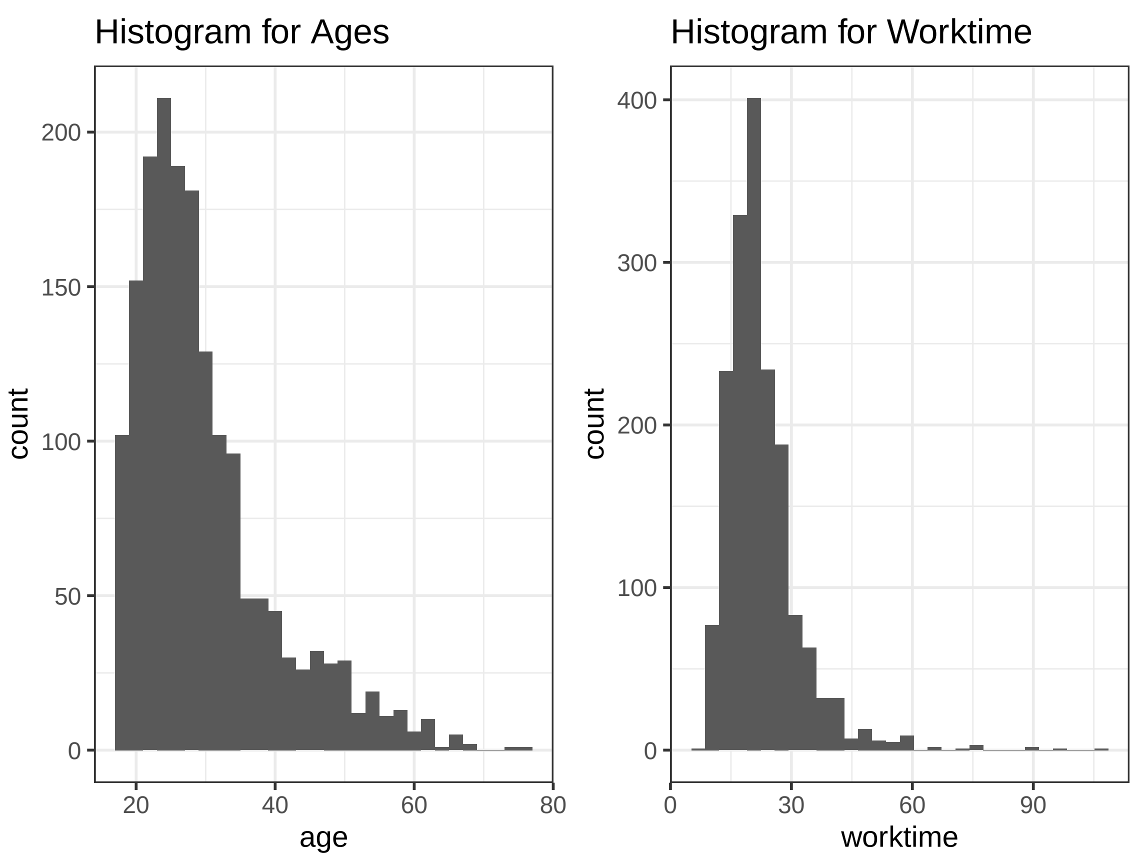
summary stats

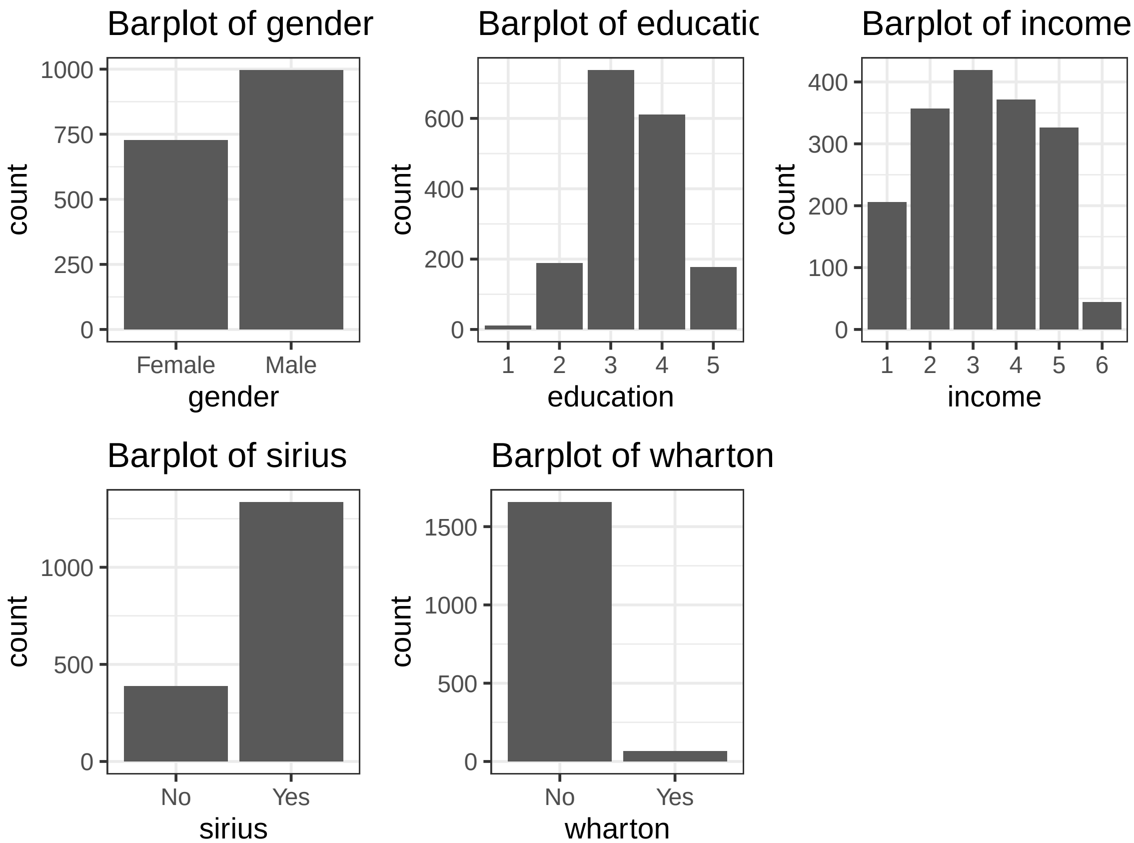
Table 1: Summary Statistics for Non-categorical Variables

	mean	min	median	max	std. dev.
age	30.29	18	28	76	9.84
worktime	22.49	18	21	76	9.30

Note:
The table reports the summary statistics for non-categorical variables in the talkshow data. The valid sample size is 1723.

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





Notes: We map the education and income levels into integers for better exhibition. For education, 1 means *less than 12 years; no high school diploma*, 2 means *High school graduate (or equivalent)*, 3 means *Some college, no diploma; or Associate's degree*, 4 means *Bachelor's degree or other 4-year degree*, 5 means *Graduate or professional degree*. For income, 1 means *Less than $15,000*, 2 means *$15,000 - $30,000*, 3 means *$30,000 - $50,000*, 4 means *$50,000 - $75,000*, 5 means *$75,000 - $150,000*, 6 means *Above $150,000*.

Sample Properties

Final Estimates

```
## [1] "95% CI: [0.038, 0.062]."
```

New Task

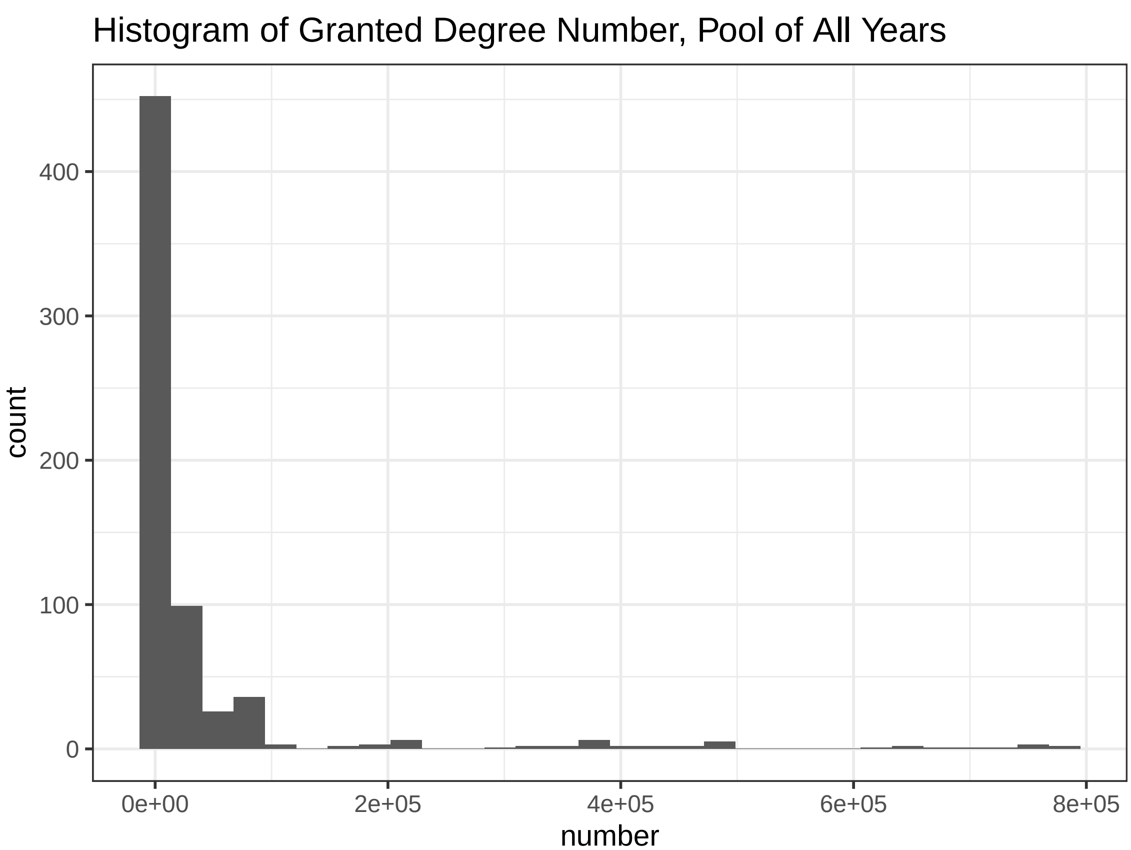
Case Study 2: Women in Science

Data Preparation

```
## integer(0)
```

```
##  [1] "Agricultural sciences"
##  [2] "Biological sciences"
```

```
##  [3] "Computer sciences"
##  [4] "Earth, atmospheric, and ocean sciences"
##  [5] "Mathematics and statistics"
##  [6] "Physical sciences"
##  [7] "Psychology"
##  [8] "Social sciences"
##  [9] "Engineering"
## [10] "Non-S&E"

## [1] "BS"  "MS"  "PhD"

## [1] "Female" "Male"

##  [1] 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     218    2118    6020   41717   18127  781474

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

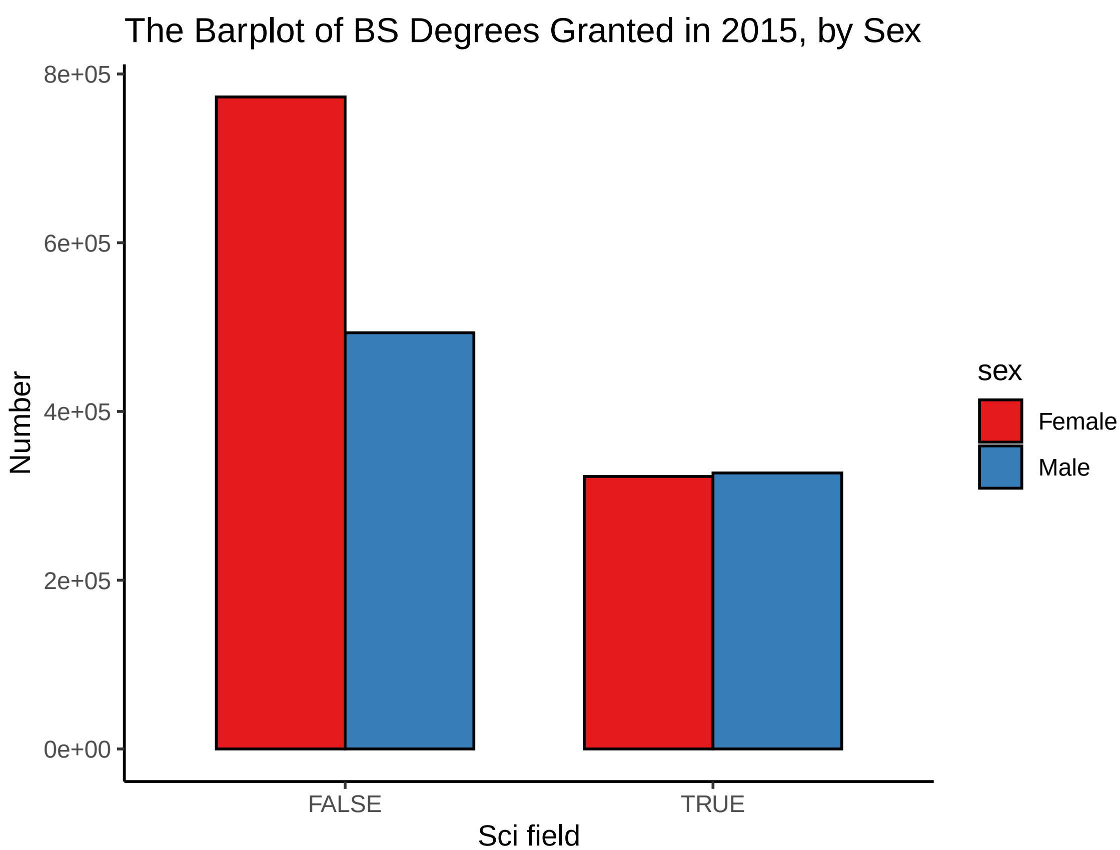


BS degrees in 2015

Table 2: Summary Statistics for BS Degrees Granted in 2015 by Sex

	Female	Male	Female.per
Non-sci	772768	493304	61.0%
Sci	322935	327122	49.7%

Note:
The table reports the summary statistics for the amount of BS degrees granted in 2015 by sex in the US degree data.



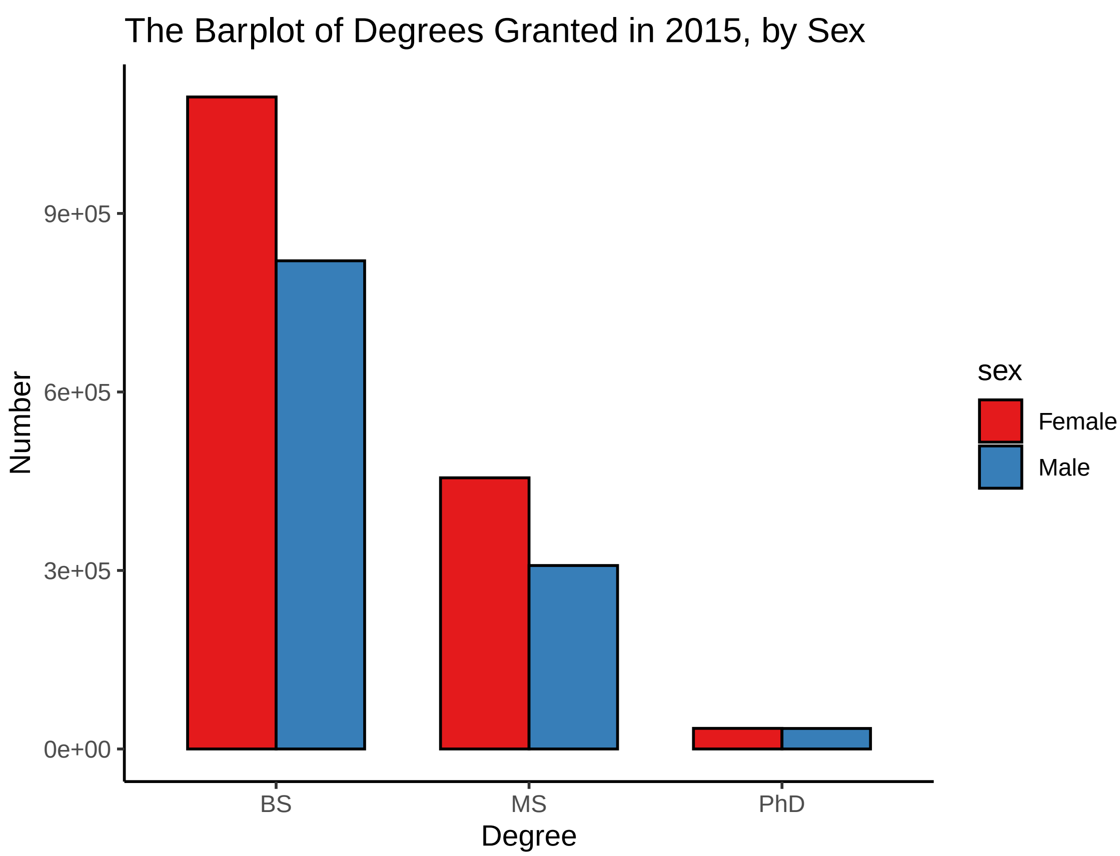
Bring in Type of Degree

```
## Using number as value column: use value.var to override.
```

Table 3: Summary Statistics for Degrees Granted in 2015 by Sex

	Female	Male	Female.per
BS	1095703	820426	57.2%
MS	455697	308283	59.6%
PhD	34660	34455	50.1%

Note:
The table reports the summary statistics for the amount of degrees granted in 2015 by sex in the US degree data.

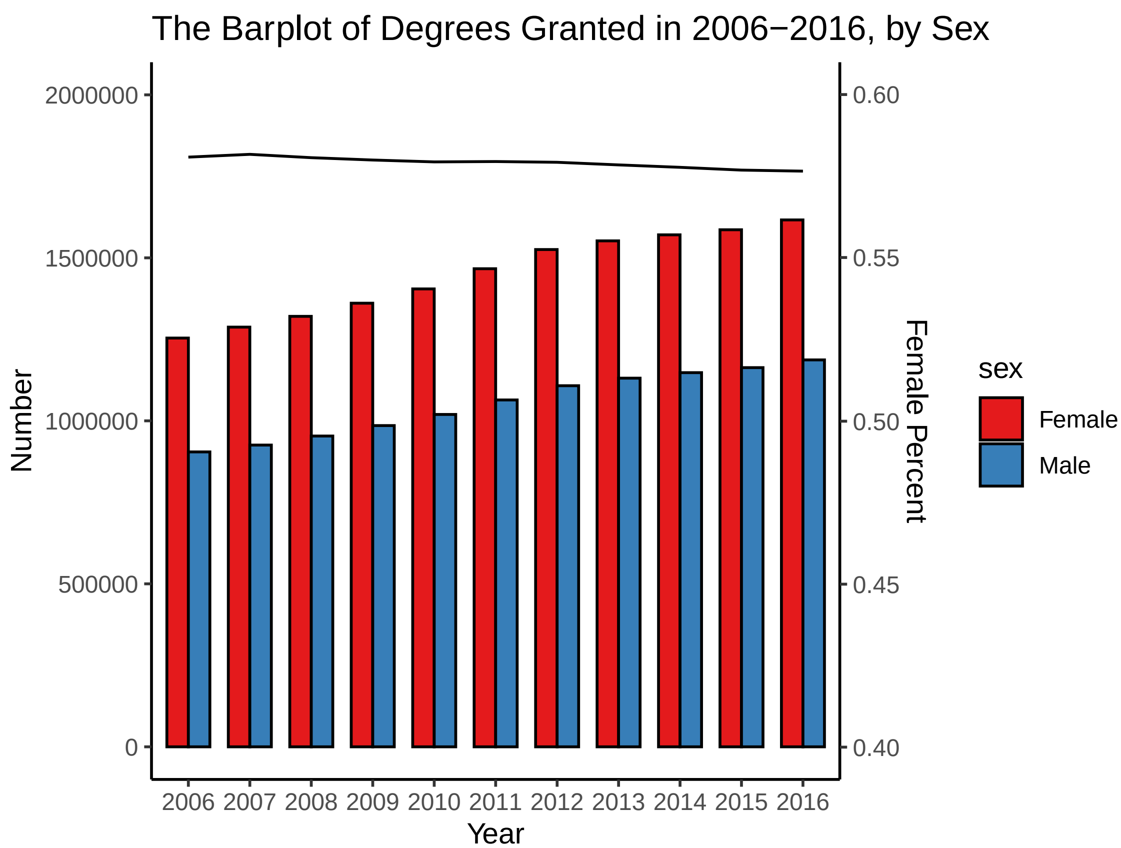


Bring All Variables

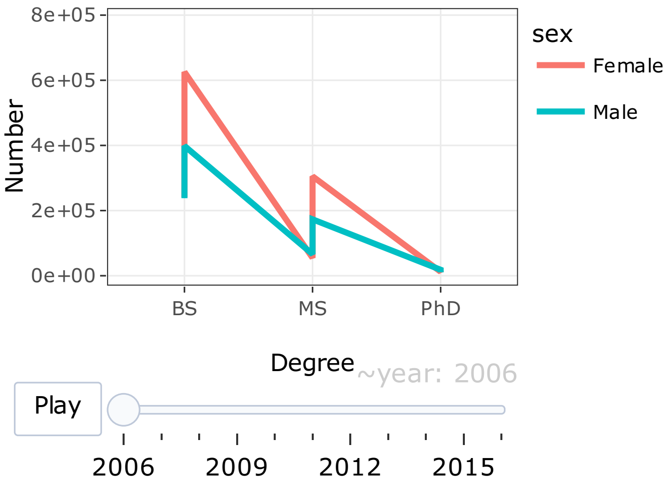
Table 4: Summary Statistics for Degrees Granted by Sex, 2006-2016

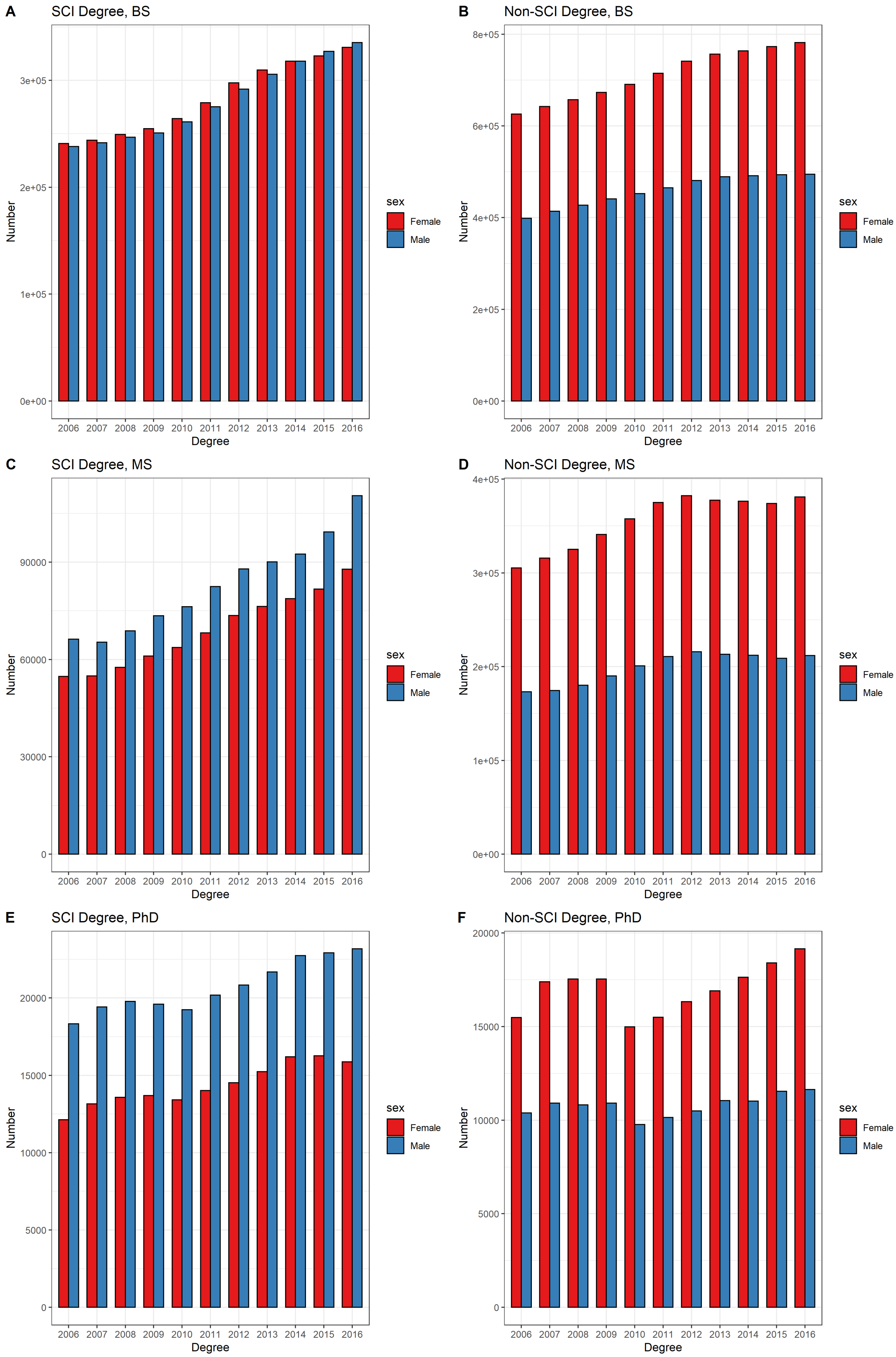
	Female	Male	Female.per
2006	1253917	904679	58.1%
2007	1287439	925621	58.2%
2008	1320480	953360	58.1%
2009	1360820	985411	58.0%
2010	1404646	1019514	57.9%
2011	1466539	1063992	58.0%
2012	1525402	1107721	57.9%
2013	1552075	1130821	57.9%
2014	1570559	1147769	57.8%
2015	1586060	1163164	57.7%
2016	1616307	1186906	57.7%

Note:
The table reports the summary statistics for the amount of degrees granted within 2006-2016 by sex in the US degree data.



The Barplot of Degrees Granted in 2006−2016, by Sex
Number
Female Percent
Year
2000000
1500000
1000000
500000
0
0.60
0.55
0.50
0.45
0.40
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
sex
Female
Male


8e+05
6e+05
4e+05
2e+05
0e+00
Number
BS
MS
PhD
Degree
sex
Female
Male
~year: 2006
Play
2006
2009
2012
2015

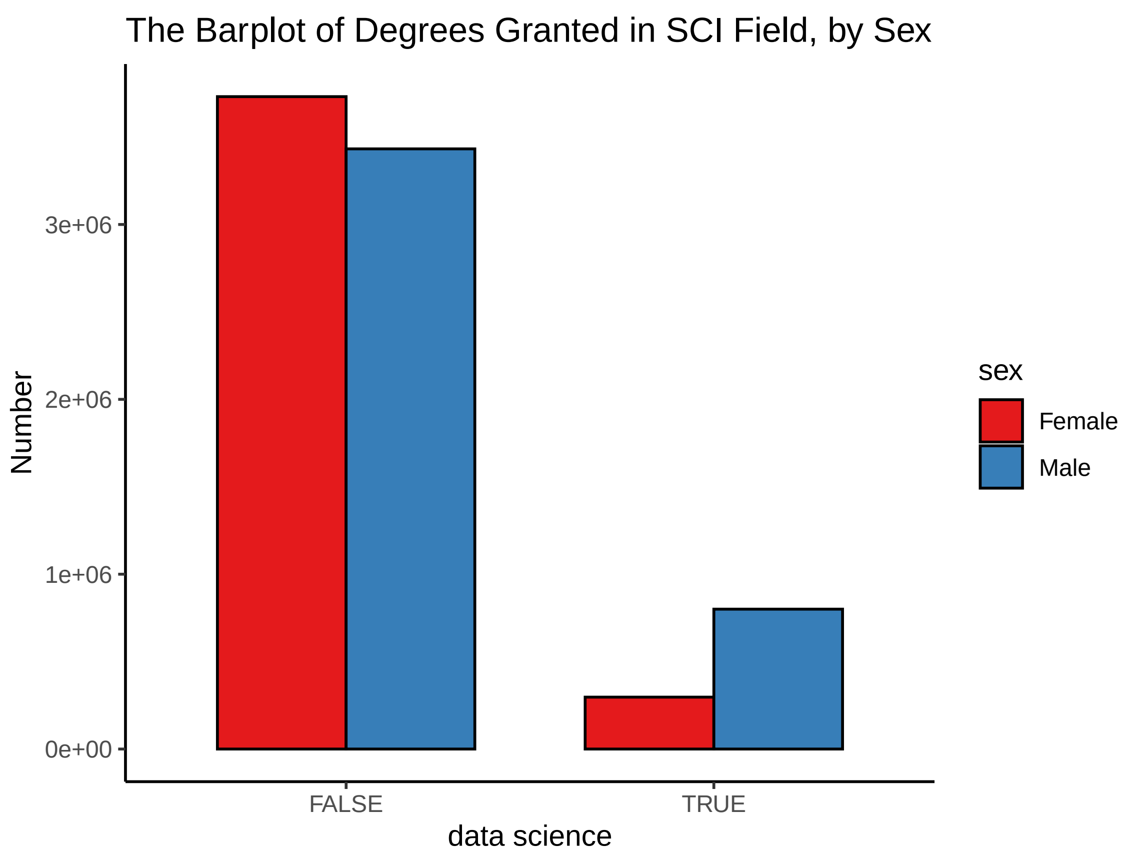

A
SCI Degree, BS
B
Non-SCI Degree, BS
C
SCI Degree, MS
D
Non-SCI Degree, MS
E
SCI Degree, PhD
F
Non-SCI Degree, PhD
Number
Degree
sex
Female
Male
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

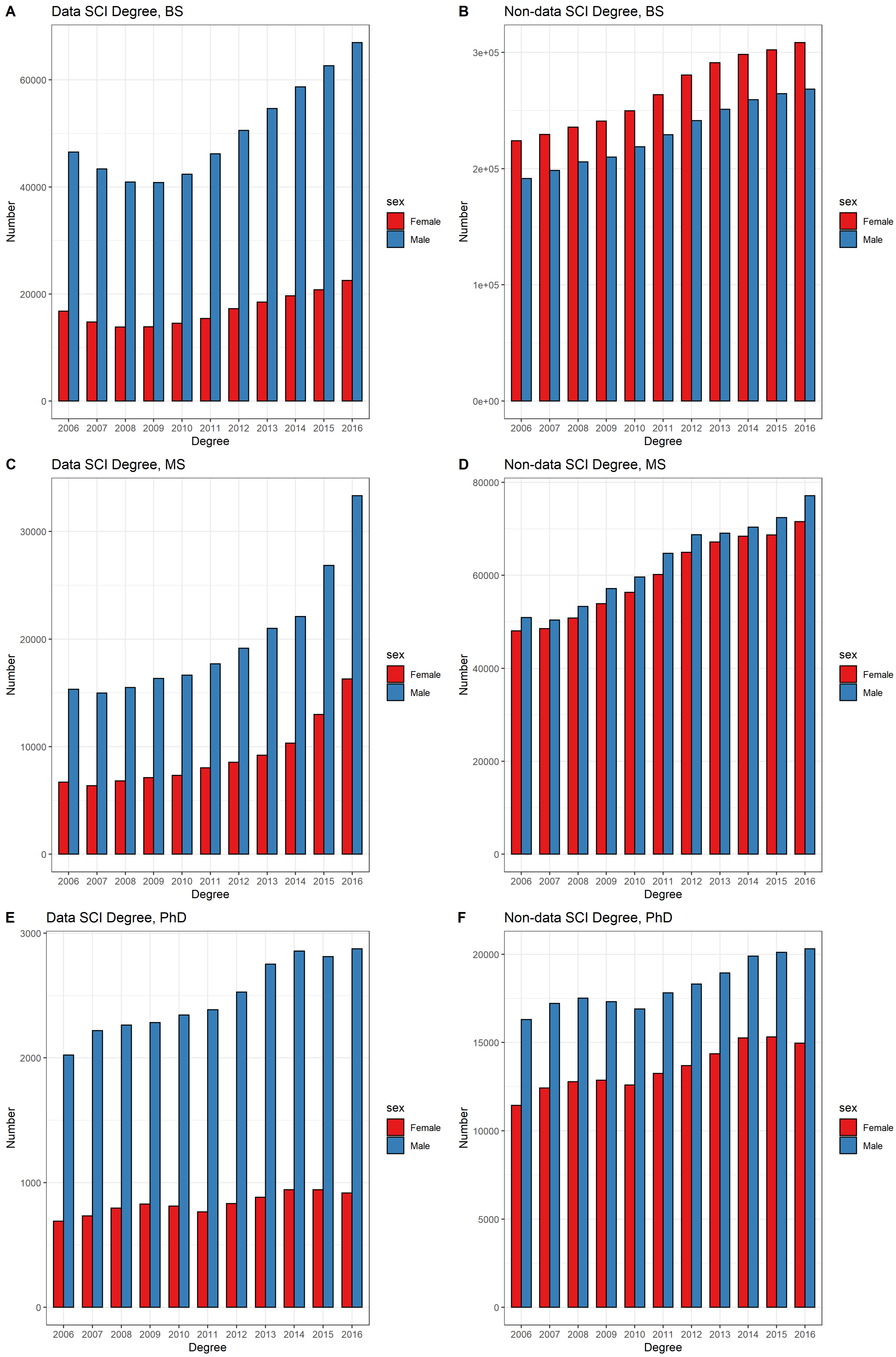
Focus on Data Science

Table 5: Summary Statistics for Degrees Granted in SCI Field by Sex

	Female	Male	Female.per
non-data sci	3731029	3432349	52.1%
data sci	296891	799889	27.1%

Note:
The table reports the summary statistics for the amount of sci degrees granted over the sample period, separated by sex and data science or not.




A
Data SCI Degree, BS
Number
Degree
sex
Female
Male
B
Non-data SCI Degree, BS
C
Data SCI Degree, MS
D
Non-data SCI Degree, MS
E
Data SCI Degree, PhD
F
Non-data SCI Degree, PhD
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

By Degree

Table 6: Summary Statistics for BS Degrees Granted in SCI Field by Sex

	Female	Male	Female.per
non-data sci	2923482	2537905	53.5%
data sci	188047	553709	25.4%

Note:
The table reports the summary statistics for the amount of BS sci degrees granted over the sample period, separated by sex and data science or not.

Table 7: Summary Statistics for MS Degrees Granted in SCI Field by Sex

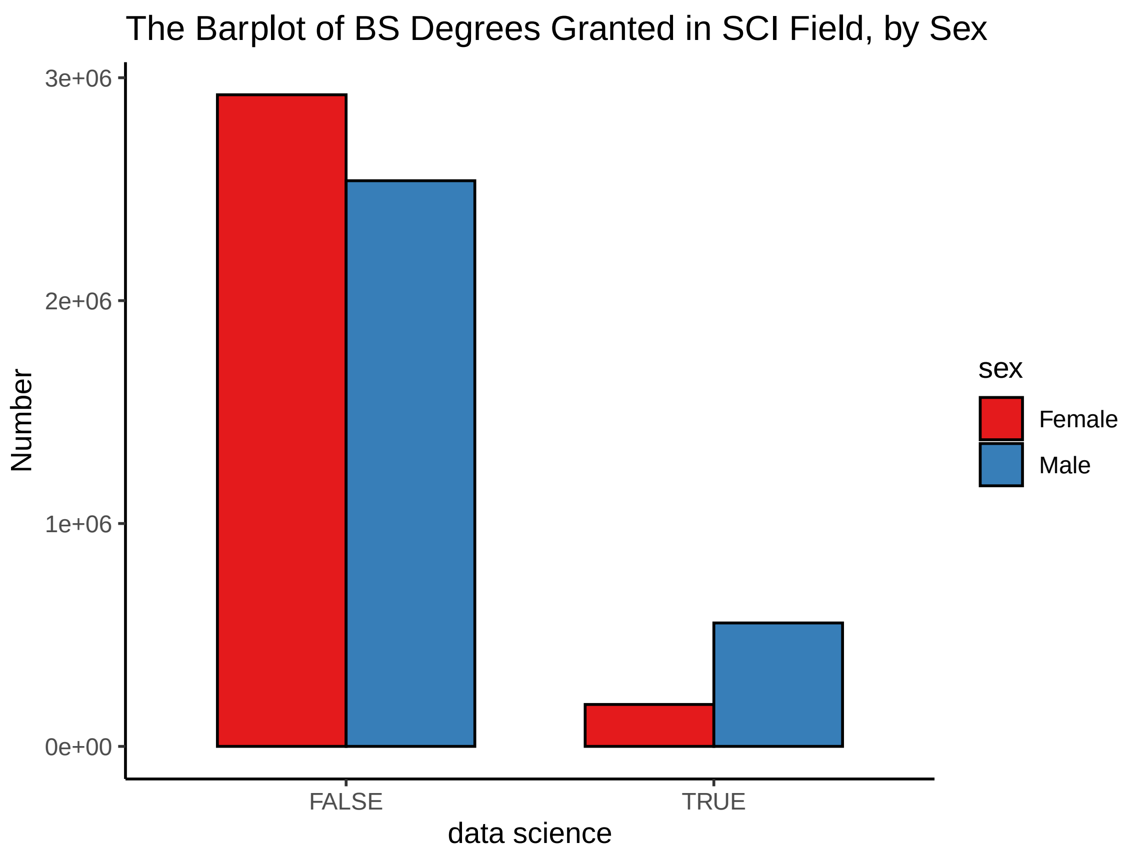
	Female	Male	Female.per
non-data sci	658613	693861	48.7%
data sci	99704	218843	31.3%

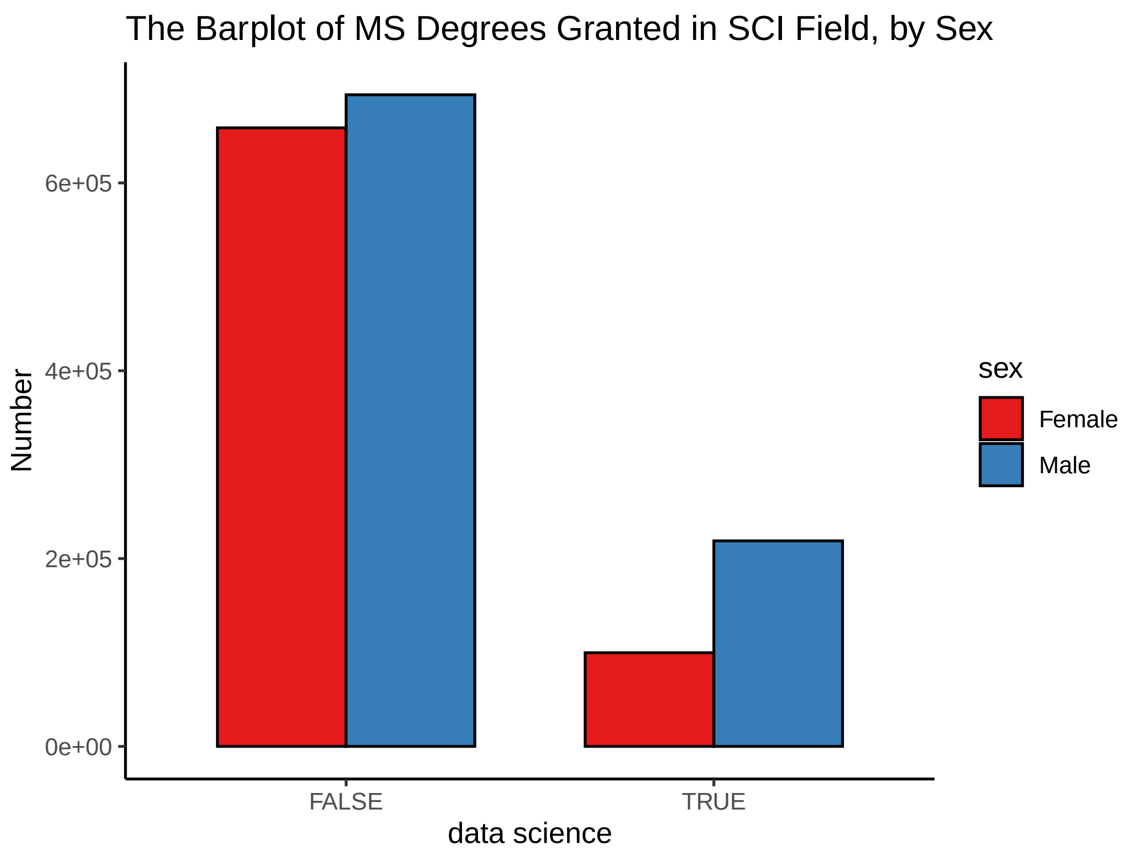
Note:
The table reports the summary statistics for the amount of MS sci degrees granted over the sample period, separated by sex and data science or not.

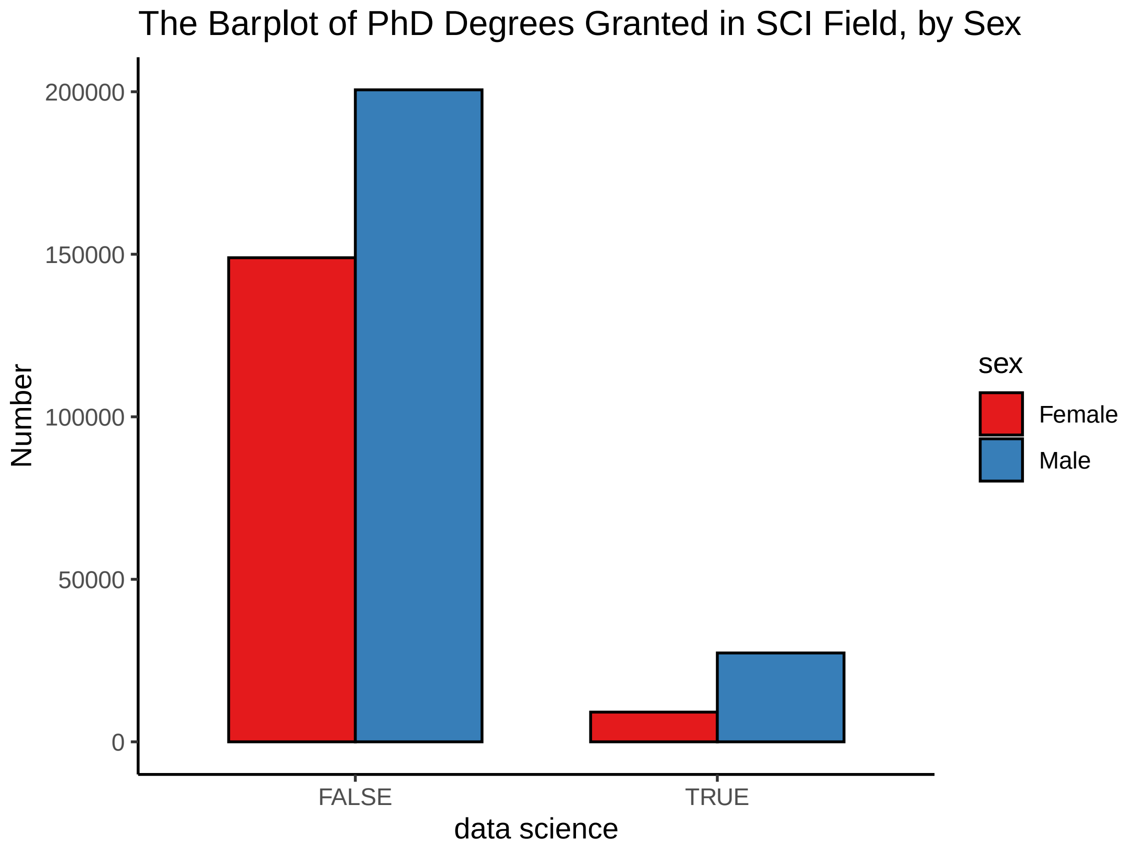
Table 8: Summary Statistics for PhD Degrees Granted in SCI Field by Sex

	Female	Male	Female.per
non-data sci	148934	200583	42.6%
data sci	9140	27337	25.1%

Note:
The table reports the summary statistics for the amount of PhD sci degrees granted over the sample period, separated by sex and data science or not.


The Barplot of BS Degrees Granted in SCI Field, by Sex
3e+06
2e+06
1e+06
0e+00
Number
FALSE
TRUE
data science
sex
Female
Male


The Barplot of MS Degrees Granted in SCI Field, by Sex
Number
6e+05
4e+05
2e+05
0e+00
FALSE
TRUE
data science
sex
Female
Male



##

Final Report

Appendix

Case Study 3: Major League Baseball

Data Preparation

The log difference is more appropriate in this setup because it measures the proportional (relative) change in the payroll. The base payrolls in all teams are not the same, so a same increase in absolute amount may incentivize players differently in different teams; the incentive may be bigger in teams with a smaller payroll, but smaller in teams with a larger payroll. The relative changes measured by the difference of logarithm of payroll can alleviate this problem.

Exploratory Questions

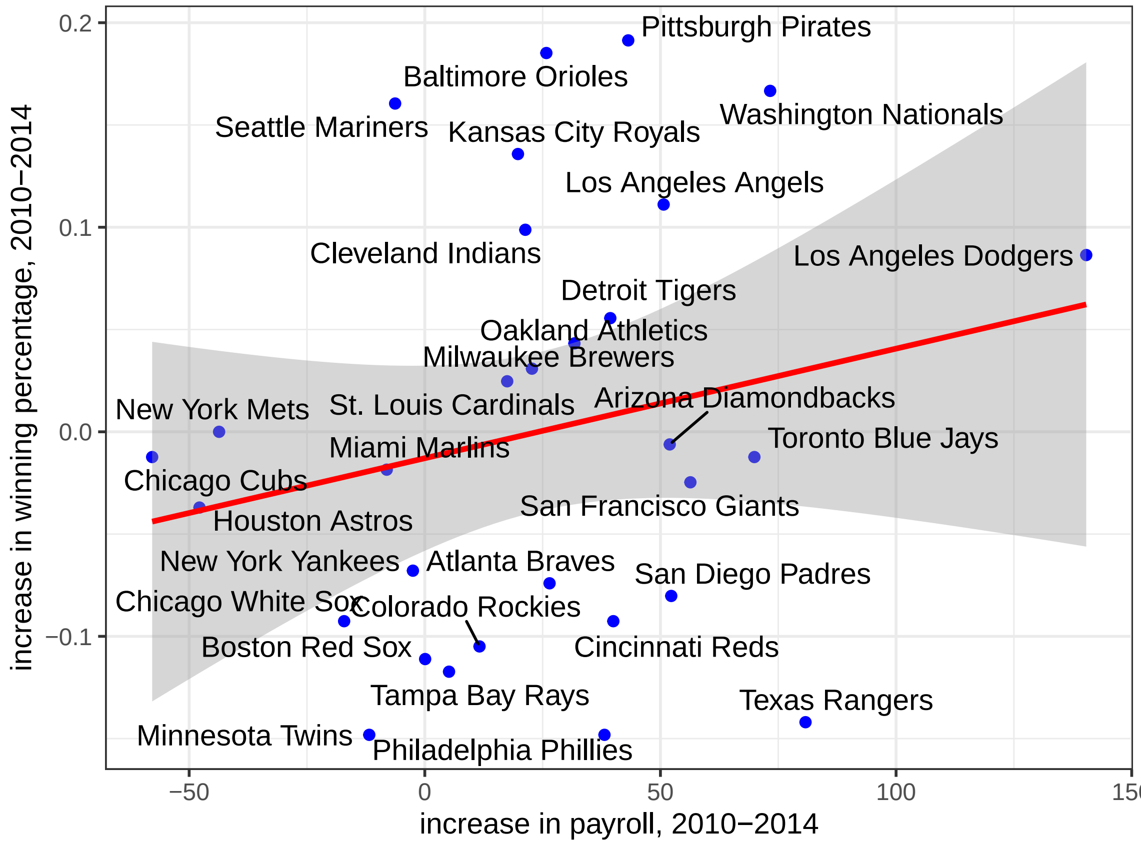
```
##                    team
## 1:  Los Angeles Dodgers
## 2:   Pittsburgh Pirates
## 3:     San Diego Padres
## 4:        Texas Rangers
```

```
## 5: Washington Nationals
```

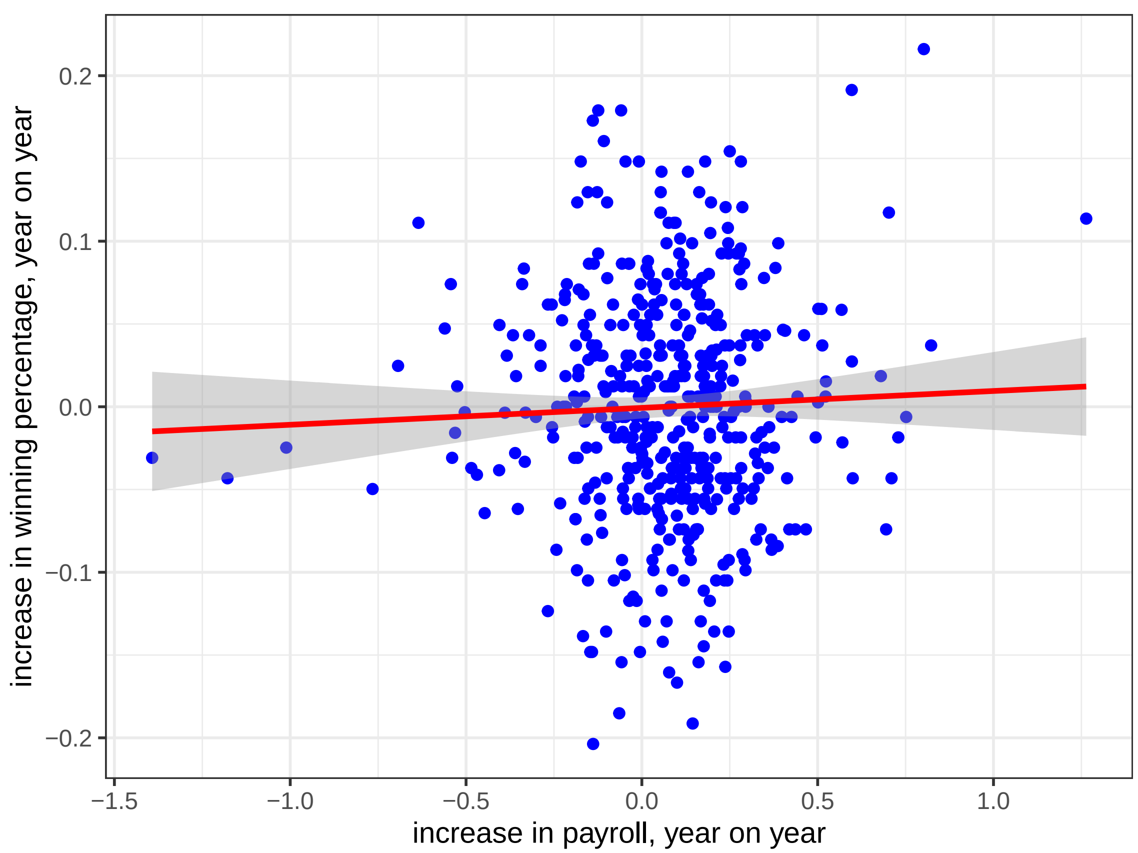
```
##                    team
## 1:  Los Angeles Dodgers
## 2: San Francisco Giants
## 3:        Texas Rangers
## 4:    Toronto Blue Jays
## 5: Washington Nationals
```

```
##                    team
## 1:    Baltimore Orioles
## 2:   Kansas City Royals
## 3:   Pittsburgh Pirates
## 4:     Seattle Mariners
## 5: Washington Nationals
```

prediction

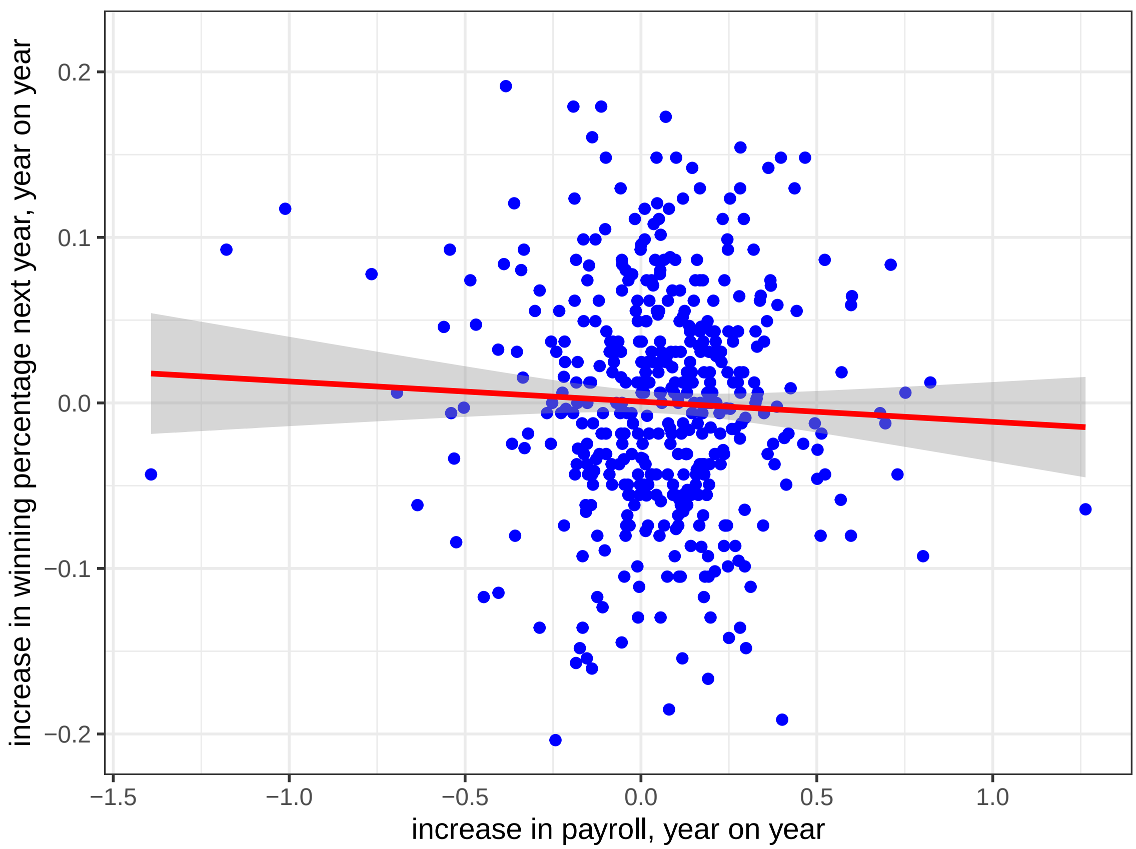


Pittsburgh Pirates
Baltimore Orioles
Seattle Mariners
Kansas City Royals
Washington Nationals
Los Angeles Angels
Cleveland Indians
Los Angeles Dodgers
Detroit Tigers
Oakland Athletics
Milwaukee Brewers
New York Mets
St. Louis Cardinals
Arizona Diamondbacks
Miami Marlins
Toronto Blue Jays
Chicago Cubs
San Francisco Giants
Houston Astros
New York Yankees
Atlanta Braves
San Diego Padres
Chicago White Sox
Colorado Rockies
Boston Red Sox
Cincinnati Reds
Tampa Bay Rays
Texas Rangers
Minnesota Twins
Philadelphia Phillies
increase in winning percentage, 2010−2014
increase in payroll, 2010−2014



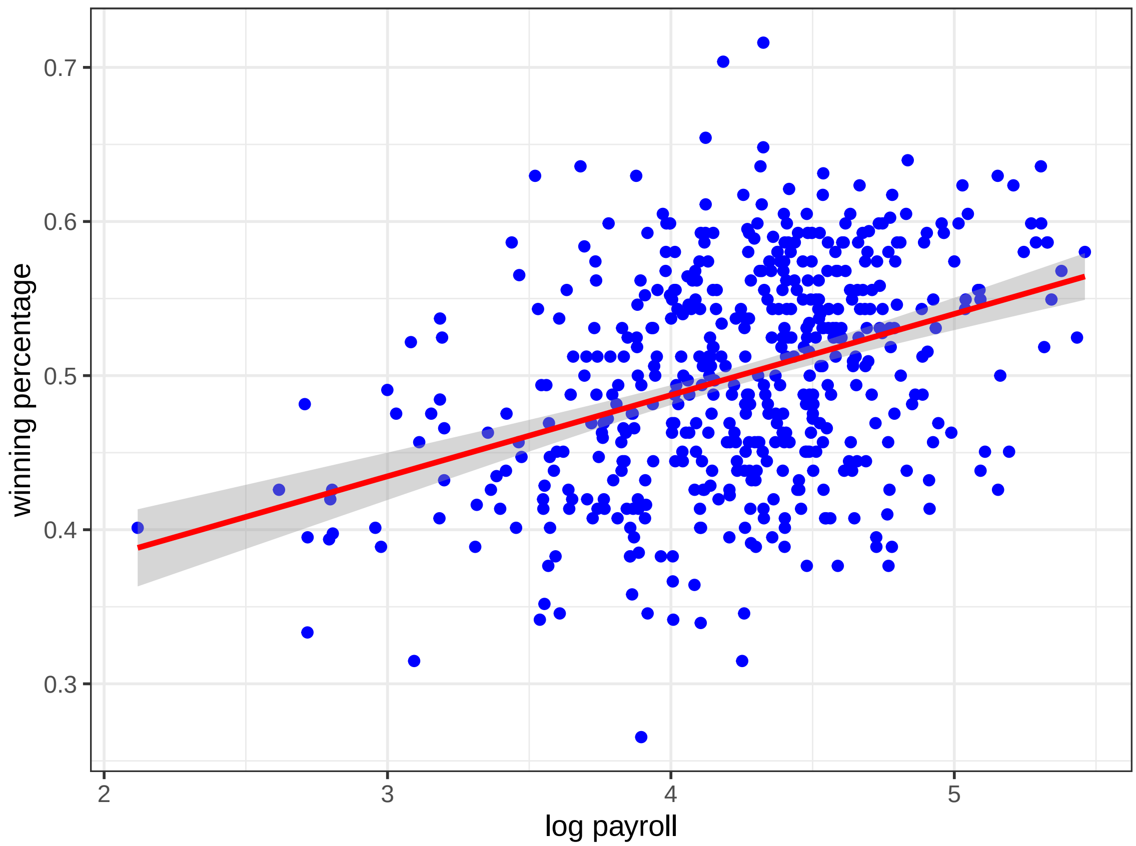
increase in winning percentage, year on year
increase in payroll, year on year

```
##
## Call:
## lm(formula = diff_win_pct ~ log.pay.diff, data = paydata.long)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.201605 -0.044979 -0.001237  0.043731  0.208554
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0006855  0.0032666  -0.210    0.834
## log.pay.diff  0.0102008  0.0123781   0.824    0.410
##
## Residual standard error: 0.06919 on 478 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.001419,   Adjusted R-squared:  -0.0006703
## F-statistic: 0.6791 on 1 and 478 DF,  p-value: 0.4103
```



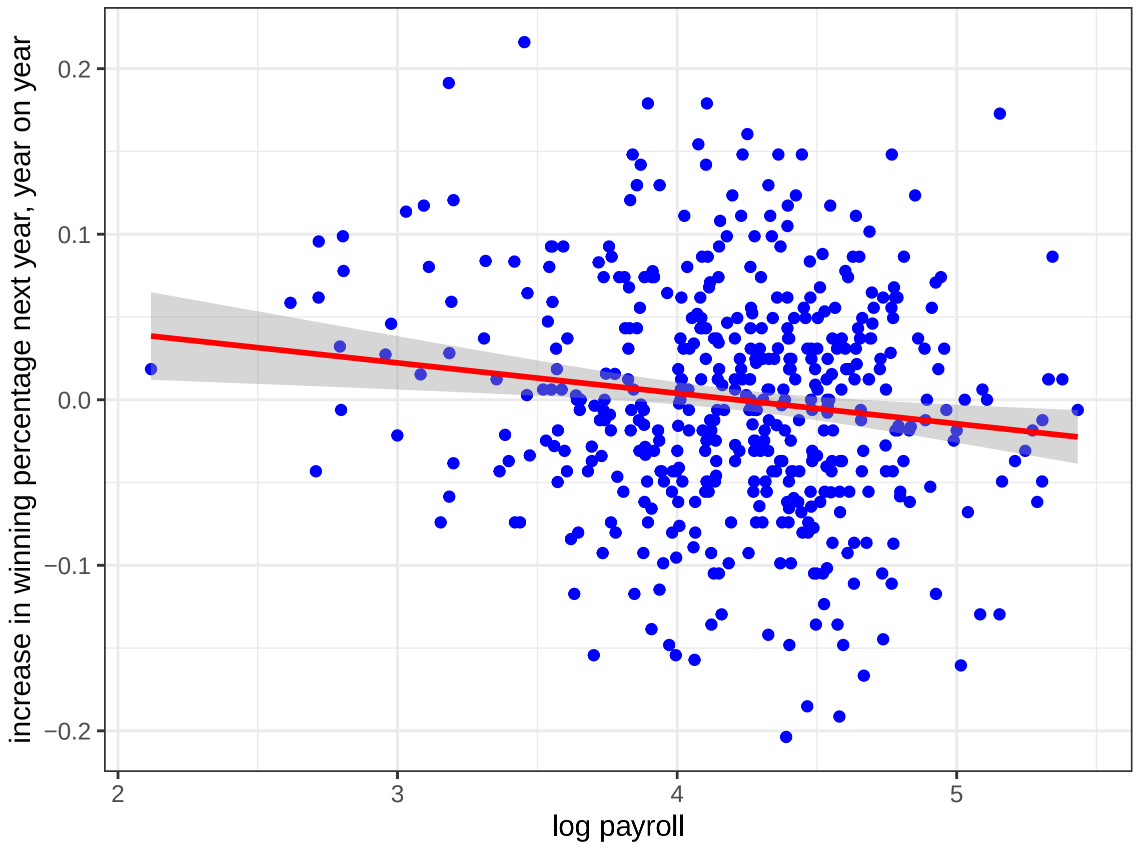

```
##
## Call:
## lm(formula = diff_win_pct_next ~ log.pay.diff, data = paydata.long)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.207438 -0.045105 -0.000575  0.045353  0.185903
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0007709  0.0033460   0.230    0.818
## log.pay.diff -0.0121974  0.0125591  -0.971    0.332
##
## Residual standard error: 0.069 on 448 degrees of freedom
##   (60 observations deleted due to missingness)
## Multiple R-squared:  0.002101,   Adjusted R-squared:  -0.0001265
## F-statistic: 0.9432 on 1 and 448 DF,  p-value: 0.332
```



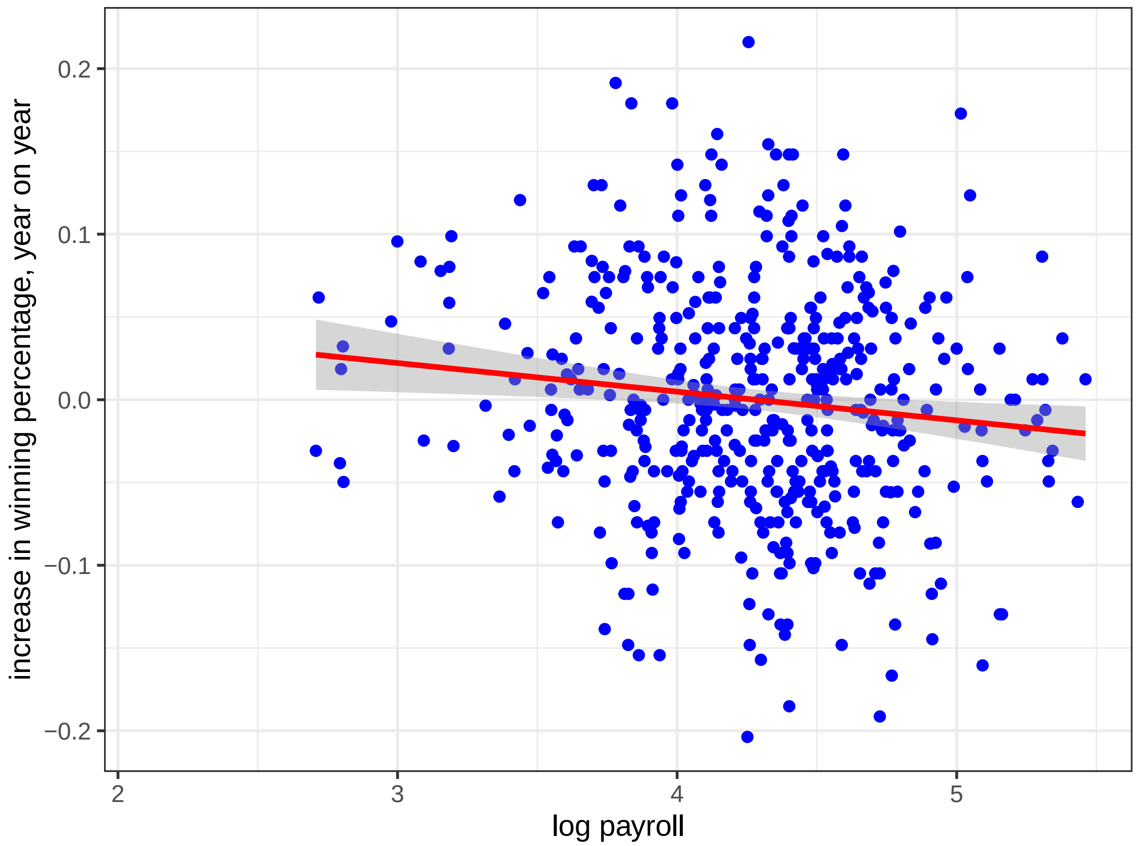

```
##
## Call:
## lm(formula = win_pct ~ log.pay, data = paydata.long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21640 -0.04691  0.00447  0.05019  0.21151
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.276629   0.024943  11.090   <2e-16 ***
## log.pay     0.052682   0.005842   9.018   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06678 on 508 degrees of freedom
## Multiple R-squared:  0.138,  Adjusted R-squared:  0.1363
## F-statistic: 81.32 on 1 and 508 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = diff_win_pct_next ~ log.pay, data = paydata.long)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.200438 -0.046946 -0.000896  0.044609  0.202100
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.077443   0.026526   2.919  0.00367 **
```

```
## log.pay     -0.018385   0.006253  -2.940  0.00344 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06862 on 478 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.01776,    Adjusted R-squared:  0.01571
## F-statistic: 8.643 on 1 and 478 DF,  p-value: 0.003442
```




```
##
## Call:
## lm(formula = diff_win_pct ~ log.pay, data = paydata.long)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.204195 -0.046024 -0.001478  0.044662  0.215630
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.074104   0.028295   2.619  0.00910 **
## log.pay     -0.017315   0.006571  -2.635  0.00868 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06874 on 478 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.01432,    Adjusted R-squared:  0.01226
## F-statistic: 6.944 on 1 and 478 DF,  p-value: 0.008683
```

Overall, current payroll predicts current performance well. As for changes in performance, there is weak evidence that increase in current performance is positively correlated to increase in payroll, but still not very predictive. It is surprising that the increase in performance, no matter current or future, is negatively correlated with the current payroll, to some degree. However, we should be cautious about this conclusion as it may be mainly driven by some rising small teams.

One more thing to note is that correlation does not mean causality.