hw1

Liming Ning

2022/1/16

Contents

Case Study 1: Audience Size

Data Preparation

cleaning

```
# selection
talkdata = fread("data/Survey_results_final.csv", encoding = "UTF-8")
talkdata.selected =
  talkdata[,.(age = Answer.Age, # Note: some answers in Age are not numerics. Therefore
              gender = Answer.Gender,
              education = Answer.Education,
              income = Answer.HouseHoldIncome,
              sirius = `Answer.Sirius Radio`,
```

```r
              wharton = `Answer.Wharton Radio`,
              worktime = WorkTimeInSeconds)]
talkdata$Reward[1:10] # question: 5 cents or 10 cents?
```

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

```r
# detect suspect observations
## age
talkdata.selected[!age %in% 10:100] # automatic type coersion when matching
```

```
##             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
```

```r
talkdata.selected[age == "Eighteen (18)", age := "18"] # imputation
talkdata.selected[age == "27`", age := "27"] # imputation
talkdata.selected = talkdata.selected[age %in% 10:100] # delete NAs
talkdata.selected[,age := as.numeric(age)]

## gender
unique(talkdata.selected$gender)
```

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

```r
talkdata.selected[!gender %in% c("Male","Female")]
```

```
##    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
```

```r
talkdata.selected = talkdata.selected[gender != ""] # delete blanks

## education
unique(talkdata.selected$education)
```

```
## [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"
```

```r
talkdata.selected = talkdata.selected[!education %in% c("Other","select one")] # delete

## income
unique(talkdata.selected$income)
```

```
## [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] ""
```

```r
talkdata.selected = talkdata.selected[income != ""] # delete blanks

## sirius
unique(talkdata.selected$sirius)
```

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

```r
talkdata.selected = talkdata.selected[sirius != ""] # delete blanks

## wharton
unique(talkdata.selected$wharton)
```

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

```
talkdata.selected = talkdata.selected[wharton != ""] # delete blanks
talkdata.selected[sirius == "No" & wharton == "Yes"] # These two are weird. Delete
```

```
##    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
```

```
talkdata.selected = talkdata.selected[!(sirius == "No" & wharton == "Yes")]

fwrite(talkdata.selected,"data/talkdata_cleaned.csv",row.names = F)
rm(list = ls())
## worktime: automatically recorded.
# possible improvements: use dplyr. get summary stats in one go and make imputation/dr
# alternatives: do not delete some obs which seems to be valid expect for some missing
```

summary stats

```
# age and worktime, integer
talkdata.selected = fread("data/talkdata_cleaned.csv",encoding = "UTF-8")
age.stat = talkdata.selected %>%
  summarise(mean = mean(age),min = min(age),median = median(age),max = max(age),"std. de
worktime.stat = talkdata.selected %>%
  summarise(mean = mean(worktime),min = min(age),median = median(worktime),max = max(age
cont.var.stat = rbind(age.stat,worktime.stat)
rownames(cont.var.stat) = c("age","worktime")
talkdata.size = nrow(talkdata.selected)
kbl(cont.var.stat, caption = "Summary Statistics for Non-categorical Variables", digits
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = paste("The table reports the summary statistics for non-categorica
            threeparttable = T)
```

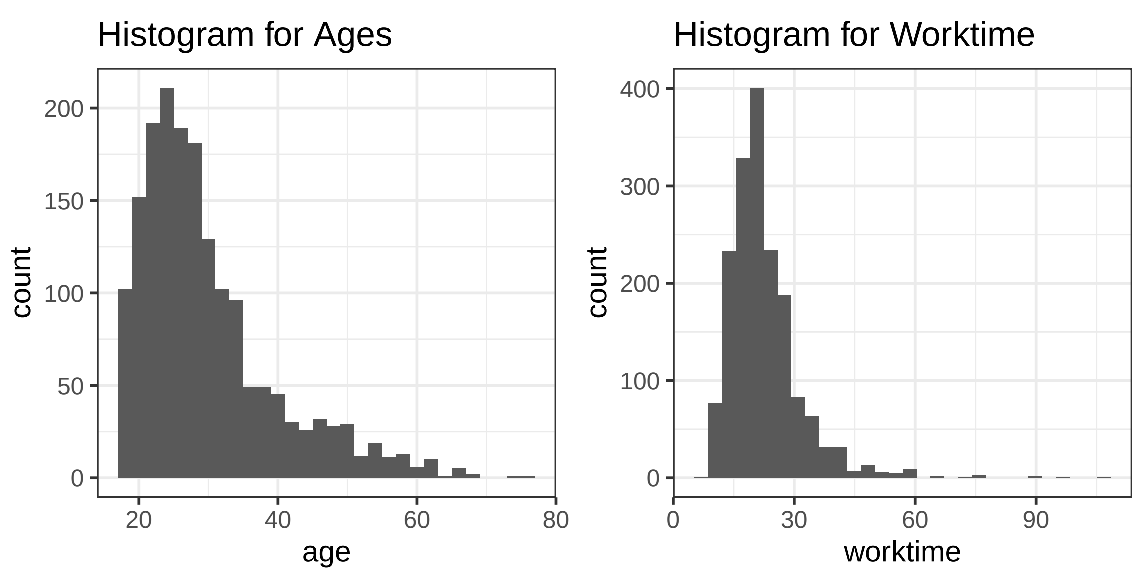
Table 1: Summary Statistics for Non-categorical Variables

	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.

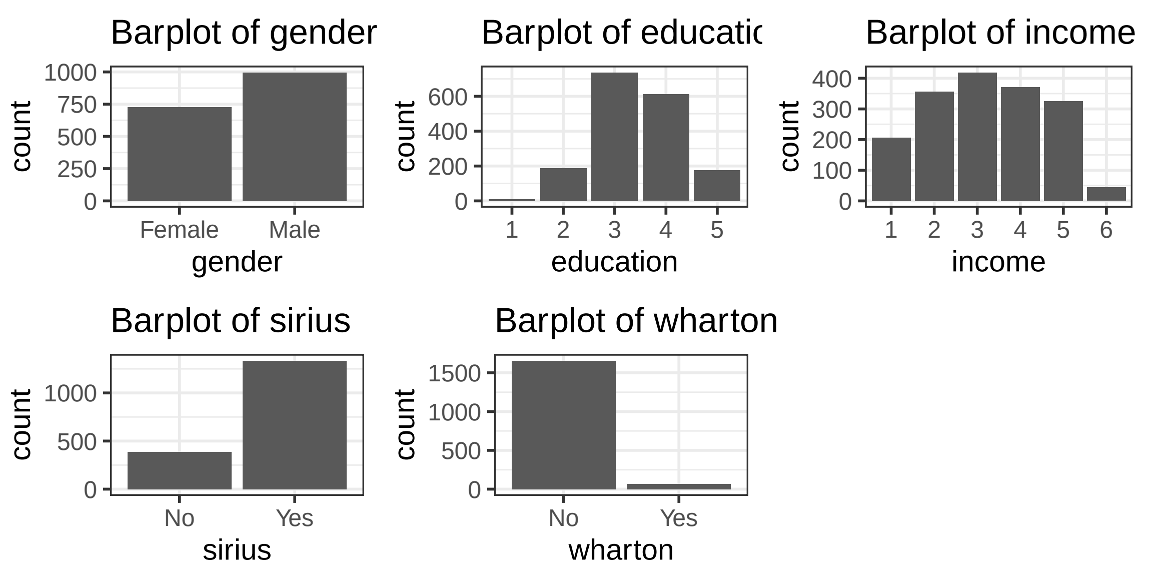
```
age.hist = ggplot(talkdata.selected,aes(x=age))+
  geom_histogram()+
  theme_bw()+
  labs(title = "Histogram for Ages")
worktime.hist = ggplot(talkdata.selected,aes(x=worktime))+
  geom_histogram()+
  theme_bw()+
  labs(title = "Histogram for Worktime")
plot_grid(age.hist,worktime.hist,nrow = 1) # right-skewed
```

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




```
# categorical variables
## mapping
keywords.edu = data.table(education = unique(talkdata.selected$education))
keywords.edu[,order := c(3,5,4,2,1)]
```

```
keywords.edu = keywords.edu[order(order)]
for (i in 1:nrow(keywords.edu)) {
  talkdata.selected[education==keywords.edu$education[i],education:=keywords.edu$order[i
} # for education

keywords.income = data.table(income = unique(talkdata.selected$income))
keywords.income[,order := c(3,2,4,6,1,5)]
keywords.income = keywords.income[order(order)]
for (i in 1:nrow(keywords.income)) {
  talkdata.selected[income==keywords.income$income[i],income:=keywords.income$order[i]]
} # for income

## plots
get.bar = function(data,varname,x.label = varname,y.label = "count",...){

  ggplot(data,aes(x=eval(parse(text = varname))))+
    geom_bar(...)+
    xlab(x.label)+
    theme_bw()+
    labs(title = paste("Barplot of ",varname,sep = ""))
}

bar.list = list()
key.catevar = c("gender","education","income","sirius","wharton")
for (i in 1:length(key.catevar)) {
  bar.list[[i]] = get.bar(talkdata.selected,key.catevar[i])
}
plot_grid(plotlist = bar.list,nrow = 2)
```



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

```
# what's the dist? to be found
# multivariate t-stats
```

Final Estimates

```
p = sum(talkdata.selected$wharton=="Yes")/sum(talkdata.selected$sirius=="Yes")
sigma = sqrt(p*(1-p)/sum(talkdata.selected$sirius=="Yes")) # use clt
paste("95% CI: [",round(p-1.96*sigma,digits = 3),", ",round(p+1.96*sigma,digits = 3),"].
```

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

New Task

```
# why not contact Sirius for data?
```

```
rm(list = ls())
```

Case Study 2: Women in Science

Data Preparation

```
degreedata = read_excel("Data/WomenData_06_16.xlsx")
degreedata = data.table(degreedata)
setnames(degreedata,1:5,c("field","degree","sex","year","number"))

# cleaning & summary
which(rowSums(is.na(degreedata))==1) # no NA
```

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

```r
unique(degreedata$field)
```

```
##  [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"
```

```r
unique(degreedata$degree)
```

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

```r
unique(degreedata$sex) # no a third sex
```

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

```r
unique(degreedata$year) # conform to the data description
```

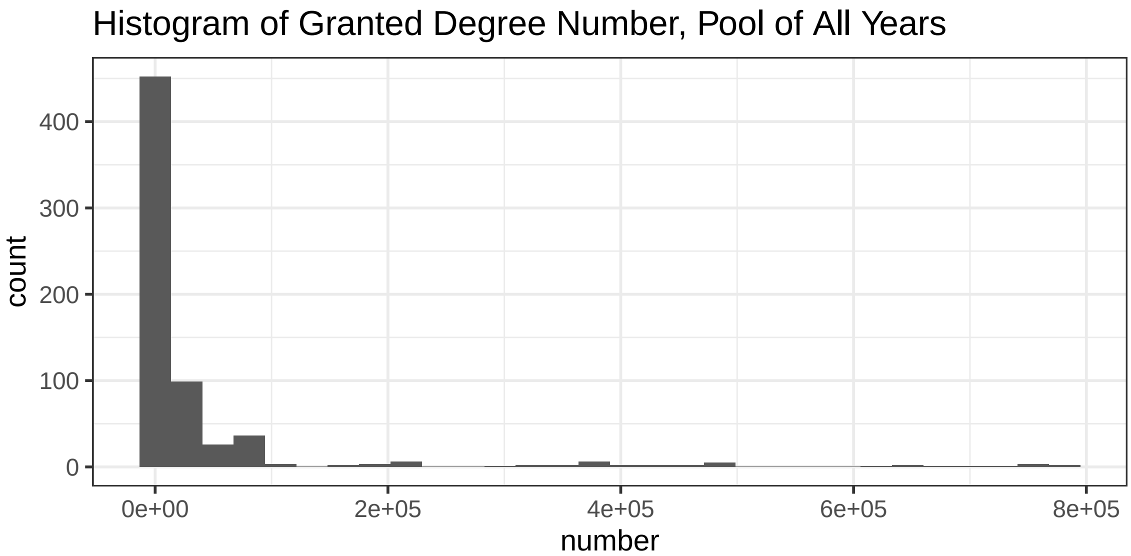
```
##  [1] 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
```

```r
summary(degreedata$number)
```

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

```r
ggplot(degreedata,aes(x=number))+
  geom_histogram()+
  theme_bw()+
  labs(title = "Histogram of Granted Degree Number, Pool of All Years") # outliers
```

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



BS degrees in 2015

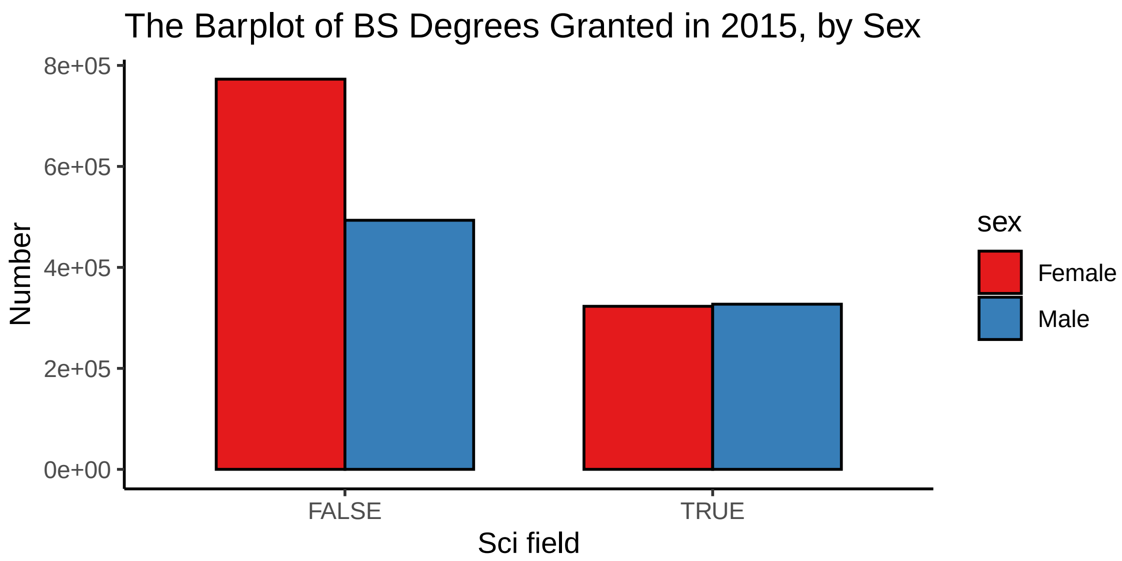
```
degreedata[,sci := field != "Non-S&E"]

degreedata.bs.2015 = degreedata[degree == "BS" & year == 2015]
summary.bs.2015 =
  degreedata.bs.2015[,.(number = sum(number)),by = .(sex,sci)]
summary.bs.2015[, percent := number/sum(number),by = sci]
summary.bs.2015.wide = data.frame(acast(summary.bs.2015,sci~sex,value.var = "number"))
rownames(summary.bs.2015.wide) = c("Non-sci","Sci")
summary.bs.2015.wide = mutate(summary.bs.2015.wide,Female.per =
                                paste(sprintf('%0.1f',round(summary.bs.2015$percent[c(2,

# summary table
kbl(summary.bs.2015.wide, caption = "Summary Statistics for BS Degrees Granted in 2015 b
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = "The table reports the summary statistics for the amount of BS deg
            threeparttable = T)
```

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.

```r
# bar plot
ggplot(summary.bs.2015,aes(x=sci,weight=number,fill=sex))+
  geom_bar(color = "black", width = .7,position = 'dodge')+
  theme_classic()+
  ylab('Number')+
  xlab('Sci field')+
  scale_fill_brewer(palette = "Set1")+
  labs(title = "The Barplot of BS Degrees Granted in 2015, by Sex")
```



Bring in Type of Degree

```r
summary.2015.sex.degree = degreedata[year==2015,
                                   .(number = sum(number)),
                                   by = .(sex,degree)]

# table
summary.2015.sex.degree.wide = data.frame(acast(summary.2015.sex.degree,degree~sex))
```

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

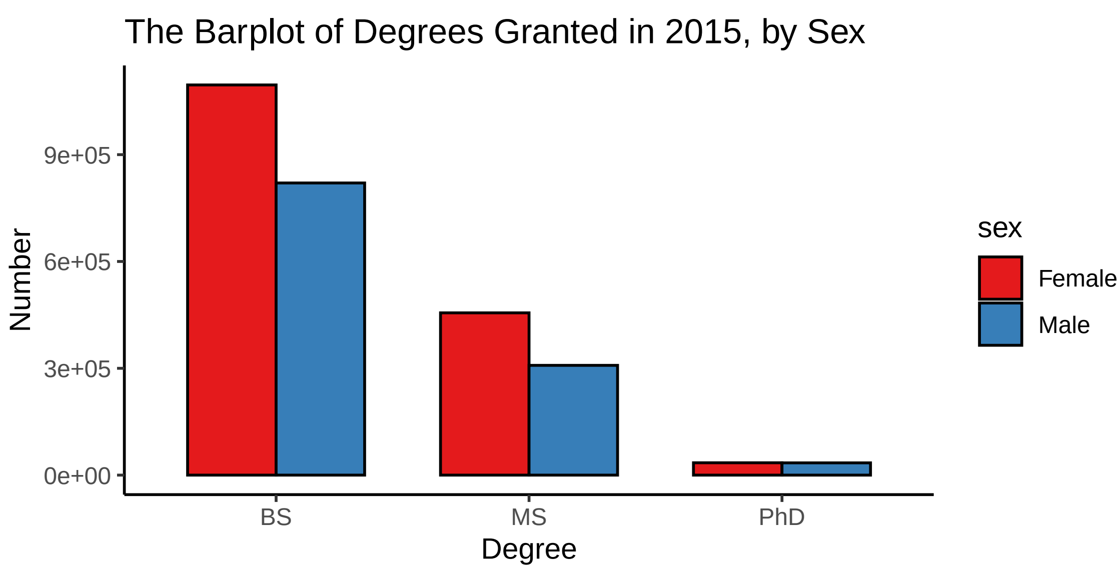
```r
summary.2015.sex.degree.wide = mutate(summary.2015.sex.degree.wide,
                                      Female.per = paste(sprintf('%0.1f',round(Female/(F
kbl(summary.2015.sex.degree.wide, caption = "Summary Statistics for Degrees Granted in 2
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = "The table reports the summary statistics for the amount of degree
            threeparttable = T)
```

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.

```
# barplot
ggplot(summary.2015.sex.degree,aes(x=degree,weight=number,fill=sex))+
  geom_bar(color = "black", width = .7,position = 'dodge')+
  theme_classic()+
  ylab('Number')+
  xlab('Degree')+
  scale_fill_brewer(palette = "Set1")+
  labs(title = "The Barplot of Degrees Granted in 2015, by Sex")
```



Bring All Variables

```
summary.sex.time = degreedata[,
                              .(number = sum(number)),
                              by = .(sex,year)]
summary.sex.time[,percent := number/sum(number),by = year]
```

```r
# table
summary.sex.time.wide = data.frame(acast(summary.sex.time,year~sex,value.var = "number")
summary.sex.time.wide = mutate(summary.sex.time.wide,
                                     Female.per = paste(sprintf('%0.1f',round(Female/(F
kbl(summary.sex.time.wide, caption = "Summary Statistics for Degrees Granted by Sex, 200
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = "The table reports the summary statistics for the amount of degree
            threeparttable = T)
```

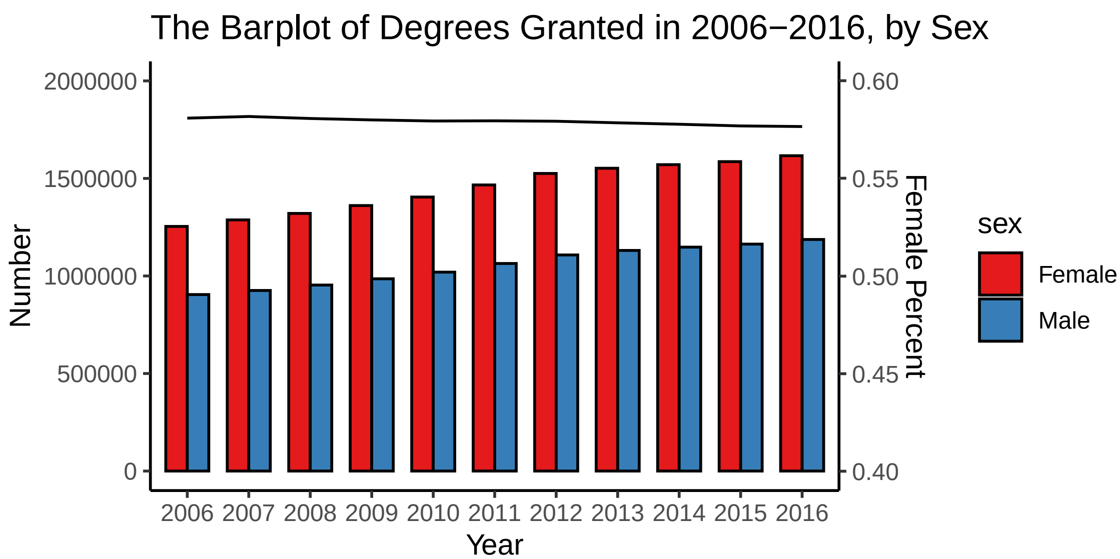
Table 4: Summary Statistics for Degrees Granted by Sex, 2006-2016

	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.

```r
# barplot
ggplot()+
  geom_bar(data = summary.sex.time,aes(x=as.factor(year),weight=number,fill=sex),
           color = "black", width = .7,position = 'dodge')+
  theme_classic()+
  ylab('Number')+
  scale_y_continuous(limits = c(0,2000000),
                     sec.axis = sec_axis(~./10000000+0.4, name = "Female Percent"))+
  geom_line(data = summary.sex.time[sex=="Female"],
            aes(x=as.factor(year),y=10000000*(percent-0.4),group=1))+
  xlab('Year')+
```

```r
  scale_fill_brewer(palette = "Set1")+
  labs(title = "The Barplot of Degrees Granted in 2006-2016, by Sex")
```



Focus on Data Science

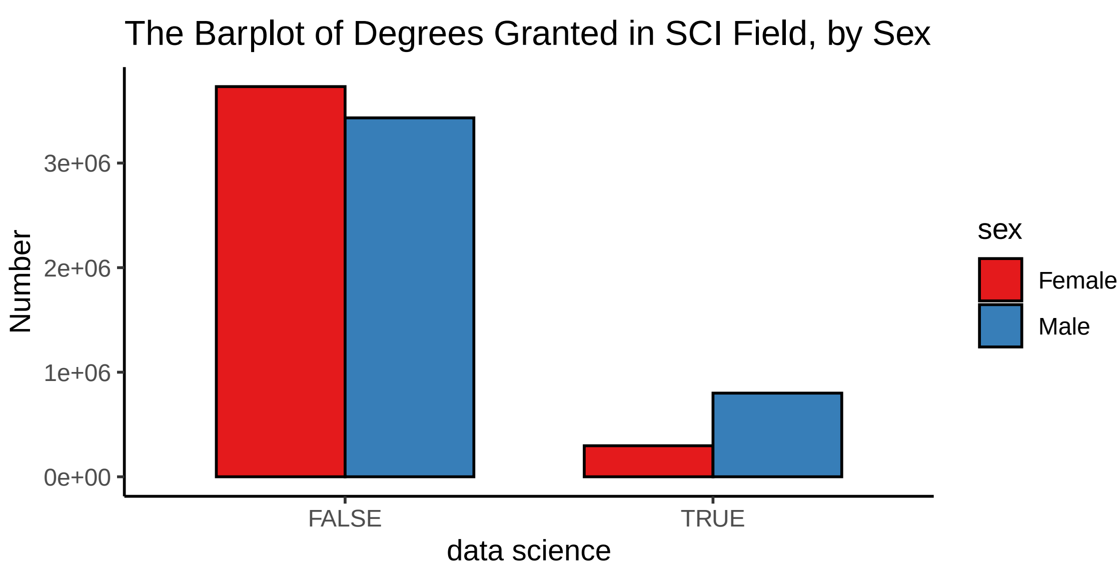
```r
degreedata[, datasci := field %in% c("Computer sciences","Mathematics and statistics")]
summary.ds = degreedata[,
                              .(number = sum(number)),
                              by = .(sci,datasci,sex,year)]
summary.ds.timeagg = summary.ds[,.(number = sum(number)),by = .(sci,datasci,sex)]
summary.ds.timeagg[, percent := number/sum(number),by = .(datasci,sci)]

# sci vs non-sci: we have conclusions before. We want to know whether it's more severe
summary.ds.timeagg.wide = data.frame(acast(summary.ds.timeagg[sci==T],datasci~sex,value.
rownames(summary.ds.timeagg.wide) = c("non-data sci"," data sci")
summary.ds.timeagg.wide = mutate(summary.ds.timeagg.wide,Female.per =
                                 paste(sprintf('%0.1f',round(summary.ds.timeagg$percent[1

# summary table
kbl(summary.ds.timeagg.wide, caption = "Summary Statistics for Degrees Granted in SCI Fi
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = "The table reports the summary statistics for the amount of sci de
            threeparttable = T)
```

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.

```
# barplot
ggplot(summary.ds.timeagg[sci==T],aes(x=datasci,weight=number,fill=sex))+
  geom_bar(color = "black", width = .7,position = 'dodge')+
  theme_classic()+
  ylab('Number')+
  xlab('data science')+
  scale_fill_brewer(palette = "Set1")+
  labs(title = "The Barplot of Degrees Granted in SCI Field, by Sex")
```



By Degree

```
##
summary.ds.bs = degreedata[sci==TRUE&degree=="BS",.(
  number = sum(number)
), by = .(datasci,sex)]
```

```r
summary.ds.bs[, percent := number/sum(number),by = .(datasci)]
summary.ds.ms = degreedata[sci==TRUE&degree=="MS",.(
  number = sum(number)
), by = .(datasci,sex)]
summary.ds.ms[, percent := number/sum(number),by = .(datasci)]
summary.ds.phd = degreedata[sci==TRUE&degree=="PhD",.(
  number = sum(number)
), by = .(datasci,sex)]
summary.ds.phd[, percent := number/sum(number),by = .(datasci)]

# sci vs non-sci: we have conclusions before. We want to know whether it's more severe
summary.ds.bs.wide = data.frame(acast(summary.ds.bs,datasci~sex,value.var = "number"))
rownames(summary.ds.bs.wide) = c("non-data sci"," data sci")
summary.ds.bs.wide = mutate(summary.ds.bs.wide,Female.per =
                                paste(sprintf('%0.1f',round(summary.ds.bs$percent[1:2]*1
summary.ds.ms.wide = data.frame(acast(summary.ds.ms,datasci~sex,value.var = "number"))
rownames(summary.ds.ms.wide) = c("non-data sci"," data sci")
summary.ds.ms.wide = mutate(summary.ds.ms.wide,Female.per =
                                paste(sprintf('%0.1f',round(summary.ds.ms$percent[1:2]*1
summary.ds.phd.wide = data.frame(acast(summary.ds.phd,datasci~sex,value.var = "number"))
rownames(summary.ds.phd.wide) = c("non-data sci"," data sci")
summary.ds.phd.wide = mutate(summary.ds.phd.wide,Female.per =
                                paste(sprintf('%0.1f',round(summary.ds.phd$percent[1:2]*

# summary table
kbl(summary.ds.bs.wide, caption = "Summary Statistics for BS Degrees Granted in SCI Fiel
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = "The table reports the summary statistics for the amount of BS sci
            threeparttable = T)
```

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.

```
kbl(summary.ds.ms.wide, caption = "Summary Statistics for MS Degrees Granted in SCI Fiel
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = "The table reports the summary statistics for the amount of MS sci
            threeparttable = T)
```

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.

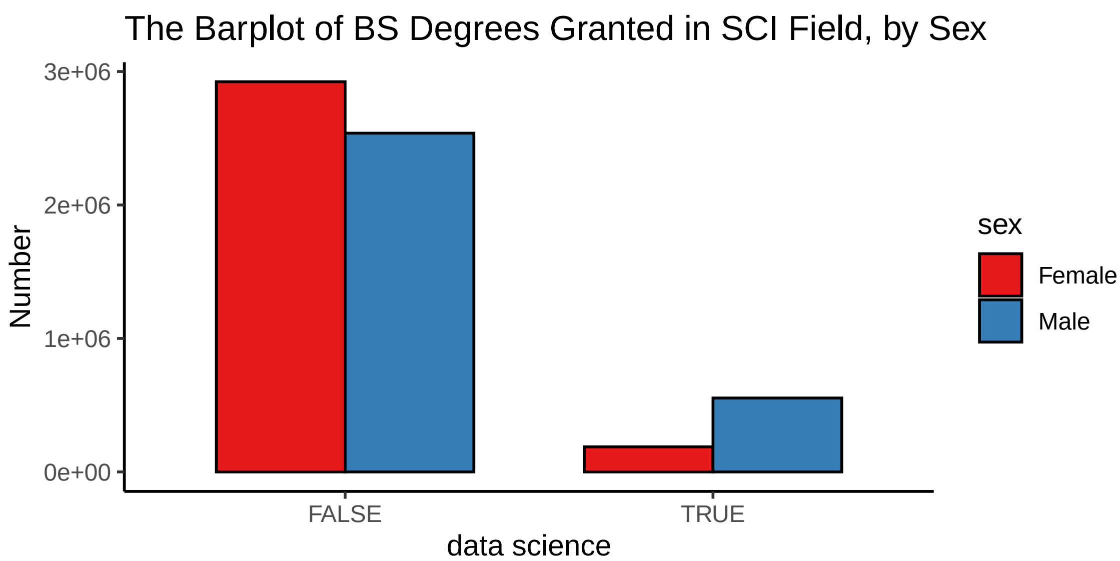
```
kbl(summary.ds.phd.wide, caption = "Summary Statistics for PhD Degrees Granted in SCI Fi
  kable_styling(latex_options = c("HOLD_position"))%>%
   footnote(general = "The table reports the summary statistics for the amount of PhD sc
            threeparttable = T)
```

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

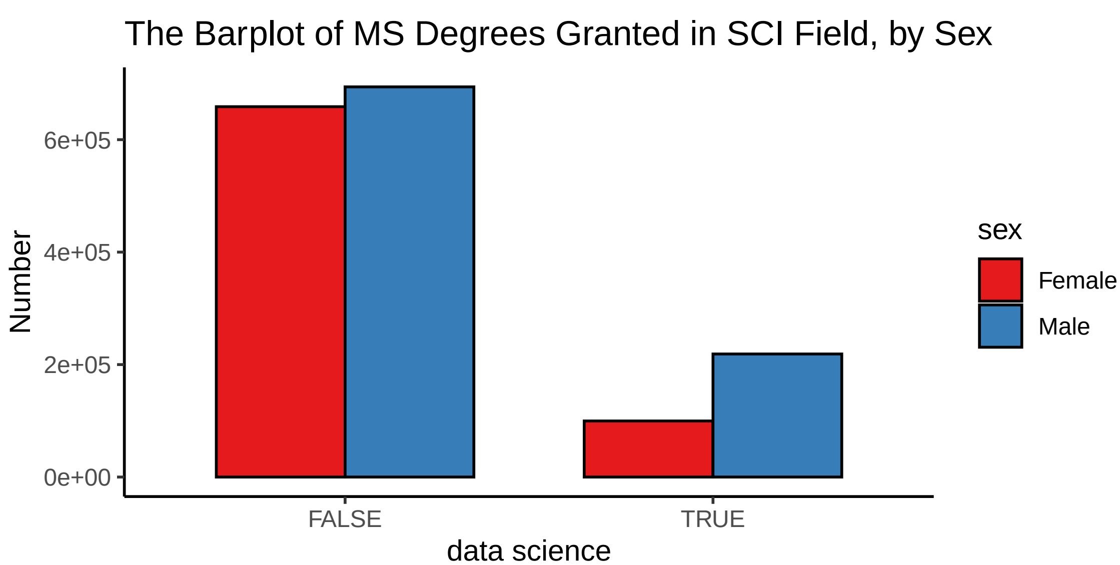
	Female	Male	Female.per
non-data sci	148934	200583	42.6%
data sci	9140	27337	25.1%

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.

```
# barplot
ggplot(summary.ds.bs,aes(x=datasci,weight=number,fill=sex))+
  geom_bar(color = "black", width = .7,position = 'dodge')+
  theme_classic()+
  ylab('Number')+
  xlab('data science')+
  scale_fill_brewer(palette = "Set1")+
  labs(title = "The Barplot of BS Degrees Granted in SCI Field, by Sex")
```

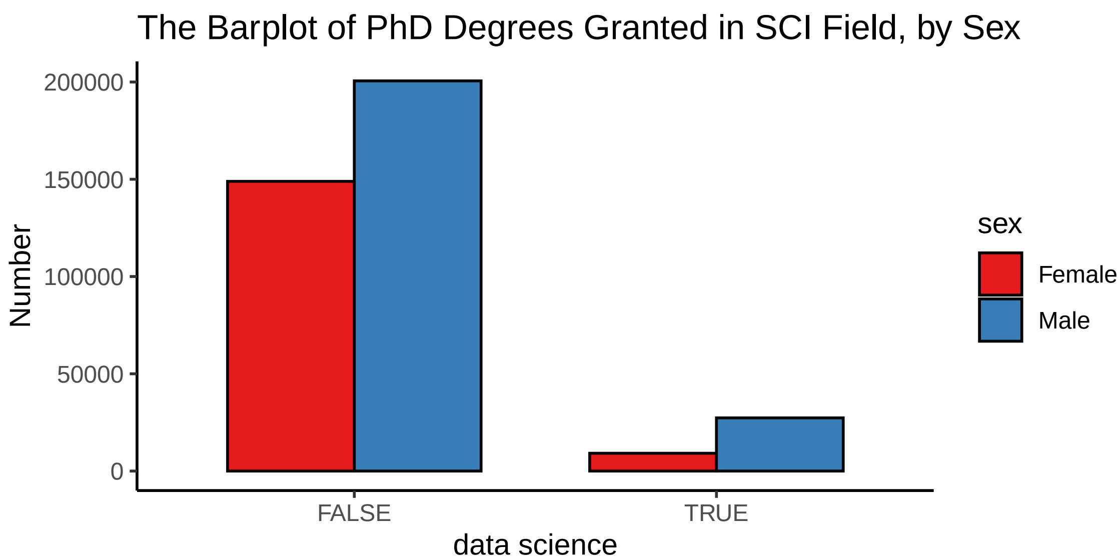



```
ggplot(summary.ds.ms,aes(x=datasci,weight=number,fill=sex))+
  geom_bar(color = "black", width = .7,position = 'dodge')+
  theme_classic()+
  ylab('Number')+
  xlab('data science')+
  scale_fill_brewer(palette = "Set1")+
  labs(title = "The Barplot of MS Degrees Granted in SCI Field, by Sex")
```




```
ggplot(summary.ds.phd,aes(x=datasci,weight=number,fill=sex))+
  geom_bar(color = "black", width = .7,position = 'dodge')+
  theme_classic()+
  ylab('Number')+
  xlab('data science')+
```

```
  scale_fill_brewer(palette = "Set1")+
  labs(title = "The Barplot of PhD Degrees Granted in SCI Field, by Sex")
```



Final Report

Appendix

Case Study 3: Major League Baseball

Data Preparation

```
rm(list = ls())
paydata.wide = fread("data/MLPayData_Total.csv", encoding = "UTF-8")
setnames(paydata.wide,"Team.name.2014","team")

paydata.long = fread("data/baseball.csv", encoding = "UTF-8")

# test conversion
# paydata.wide[,.(p1998:p2014)] # do not try to use .() and : simultaneously. otherwis

# pay = paydata.wide[,team,p1998:p2014] %>%
#   pivot_longer(cols = p1998:p2014,
#                names_prefix = "p",
#                names_to = "year",
#                values_to = "payroll")
# win.num = paydata.wide[,team,X1998:X2014] %>%
#   pivot_longer(cols = X1998:X2014,
#                names_prefix = "X",
```

```r
#                 names_to = "year",
#                 values_to = "win.num")
# win.per = paydata.wide[,team,X1998.pct:X2014.pct] %>%
#   pivot_longer(cols = X1998.pct:X2014.pct,
#                 names_prefix = "X",
#                 names_to = "year",
#                 values_to = "win.pct") %>%
#   mutate(year = substr(year,1,4))
# test.long = pay %>%
#   right_join(win.num, by = c("year","team")) %>%
#   right_join(win.per, by = c("year","team"))

# create increments
paydata.long[order(team,year),log.pay := log(payroll)]
paydata.long[,pay.diff := c(NA,diff(payroll)),by = team]
paydata.long[,log.pay.diff := c(NA,diff(log.pay)),by = team]
paydata.new = paydata.long[,.(team,year,diff_log=log.pay.diff,win_pct)]
```

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

```r
increase.data = paydata.new[year %in% 2010:2014,.(
  pay.increase = sum(diff_log[-1]), # from 2010 to 2014, increase in log(payroll)
  win.pct.increase = win_pct[length(win_pct)]-win_pct[1]
), by = team]
increase.data %>%
  filter(rank(-pay.increase)<6) %>%
  select(team) # top 5 teams in payroll increase
```

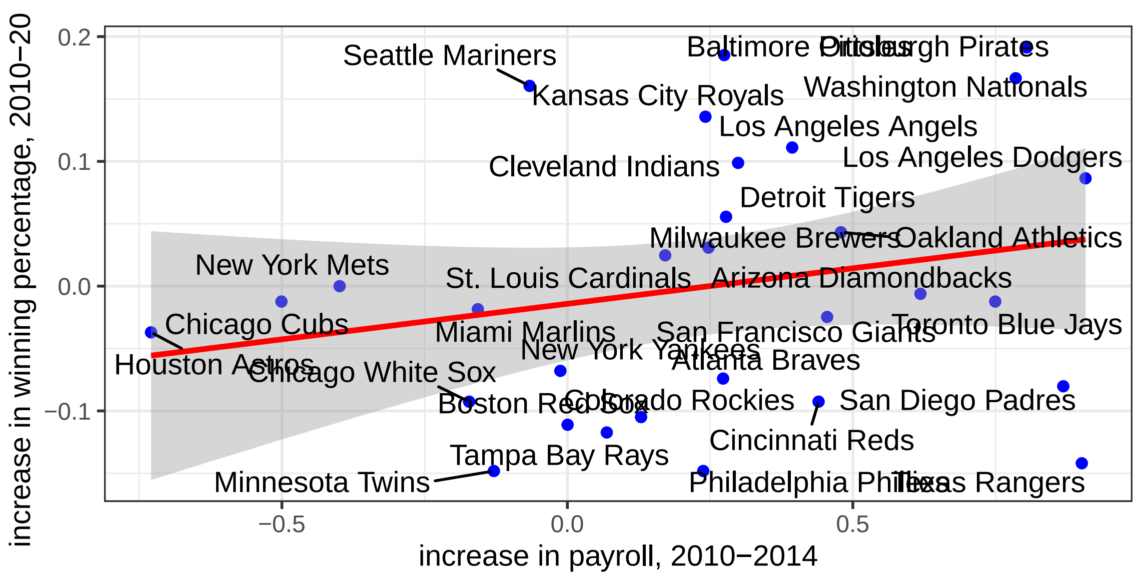
```
##                     team
## 1:   Los Angeles Dodgers
## 2:    Pittsburgh Pirates
## 3:      San Diego Padres
## 4:         Texas Rangers
## 5: Washington Nationals
```

```r
increase.data %>%
  filter(rank(-win.pct.increase)<6) %>%
  select(team) # top 5 teams in winning percentage increase
```

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

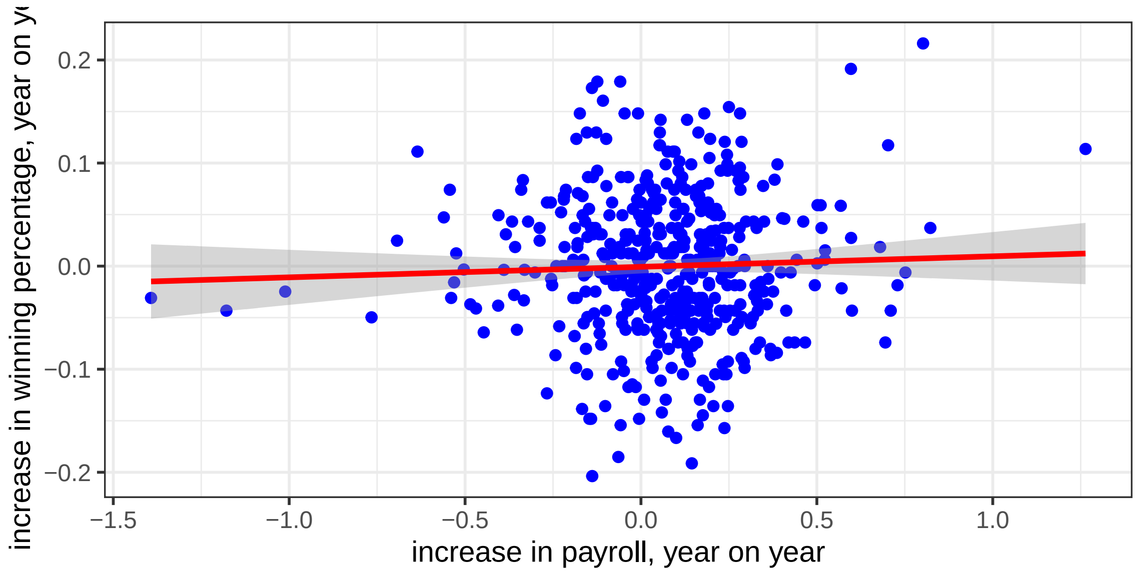
prediction

```r
# 2010-2014
ggplot(increase.data,aes(x=pay.increase,y=win.pct.increase))+
  geom_point(color="blue",size=1.5)+
  geom_smooth(method = "lm",formula = y~x, color="red")+
  geom_text_repel(aes(label = team))+
  xlab("increase in payroll, 2010-2014")+
  ylab("increase in winning percentage, 2010-2014")+
  theme_bw()+
  labs("Relationship between Payroll Increase and Performance Increase, 2010-2014")
```




```r
# all years, year-on-year basis
paydata.long[,diff_win_pct := c(NA,diff(win_pct)), by = team]

ggplot(paydata.long,aes(x=log.pay.diff,y=diff_win_pct))+
```

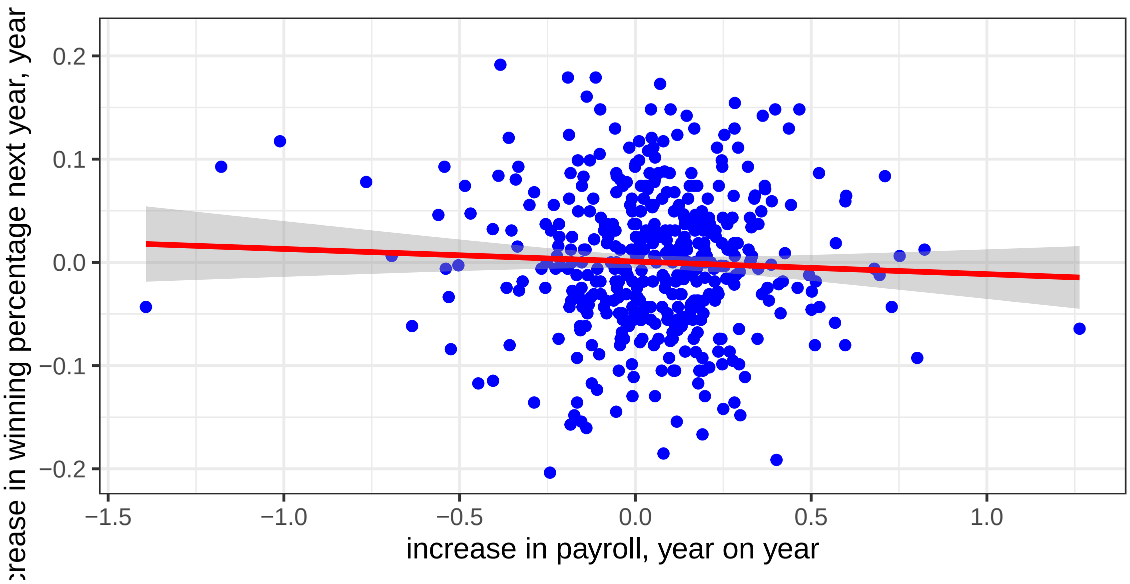
```r
  geom_point(color="blue",size=1.5)+
  geom_smooth(method = "lm", formula = y~x, color="red")+
  xlab("increase in payroll, year on year")+
  ylab("increase in winning percentage, year on year")+
  theme_bw()+
  labs("Relationship between Payroll Increase and Performance Increase, All Years")
```




```r
summary(lm(diff_win_pct~log.pay.diff,data = paydata.long)) # not significant
```

```
##
## 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
```

```r
# how about next year?
paydata.long[,diff_win_pct_next := c(diff(win_pct),NA), by = team]

ggplot(paydata.long,aes(x=log.pay.diff,y=diff_win_pct_next))+
  geom_point(color="blue",size=1.5)+
  geom_smooth(method = "lm", formula = y~x, color="red")+
  xlab("increase in payroll, year on year")+
  ylab("increase in winning percentage next year, year on year")+
  theme_bw()+
  labs("Relationship between Payroll Increase and Performance Increase Next Year, All Ye
```

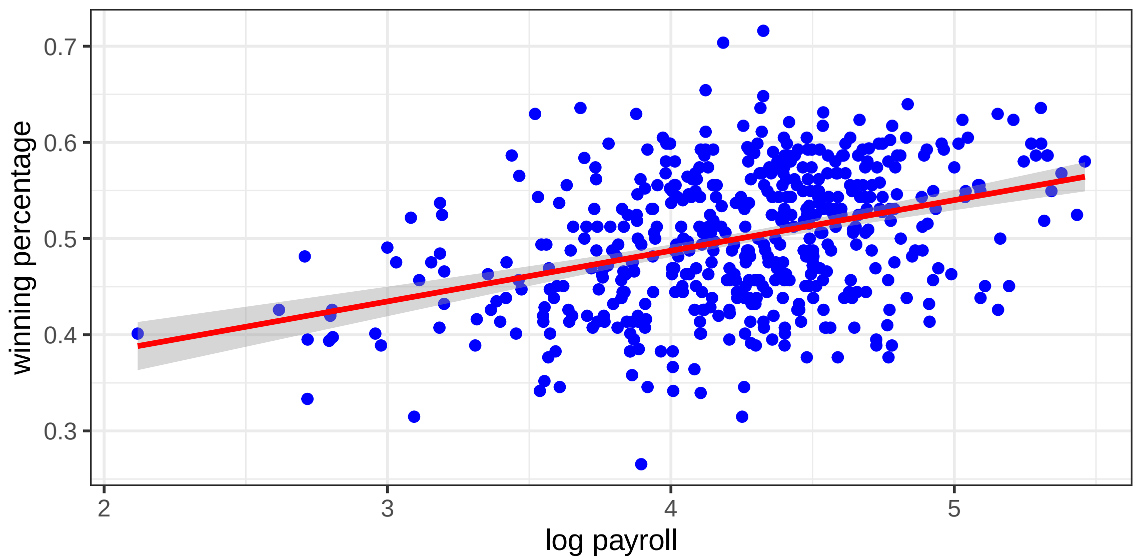



```r
summary(lm(diff_win_pct_next~log.pay.diff,data = paydata.long)) # negative and not sign
```

```
##
## 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
```

```
# payroll to winning percentage
ggplot(paydata.long,aes(x=log.pay,y=win_pct))+
  geom_point(color="blue",size=1.5)+
  geom_smooth(method = "lm", formula = y~x, color="red")+
  xlab("log payroll")+
  ylab("winning percentage")+
  theme_bw()+
  labs("Relationship between Payroll and Performance, All Years")
```

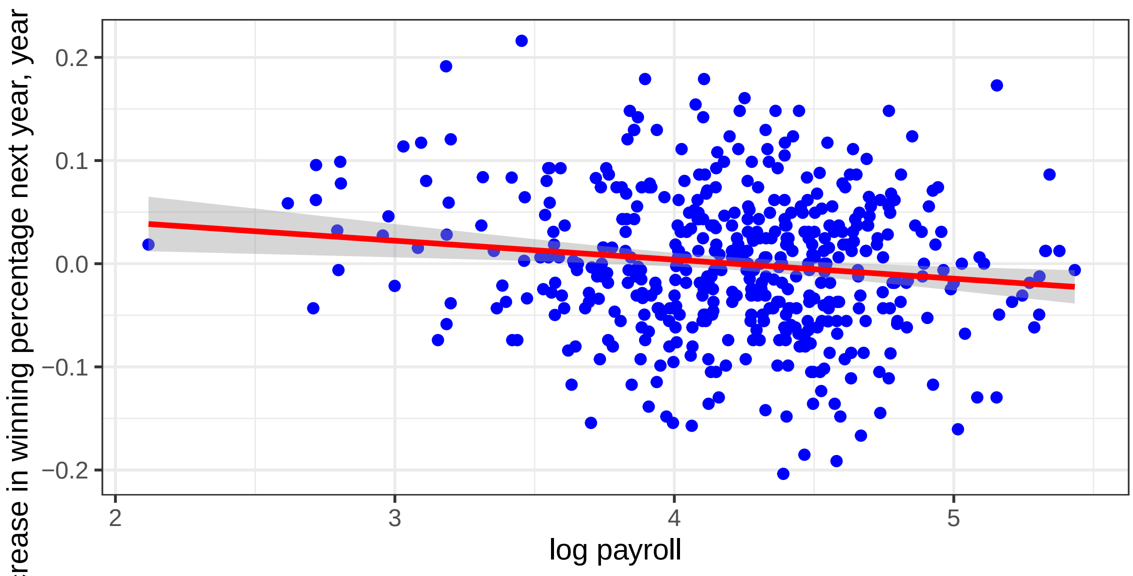



```
summary(lm(win_pct~log.pay,data = paydata.long)) # good prediction
```

```
##
## 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
```

```
# payroll to increase in winning percentage next year
ggplot(paydata.long,aes(x=log.pay,y=diff_win_pct_next))+
  geom_point(color="blue",size=1.5)+
  geom_smooth(method = "lm", formula = y~x, color="red")+
  xlab("log payroll")+
  ylab("increase in winning percentage next year, year on year")+
  theme_bw()+
  labs("Relationship between Payroll and Performance Increase Next Year, All Years")
```

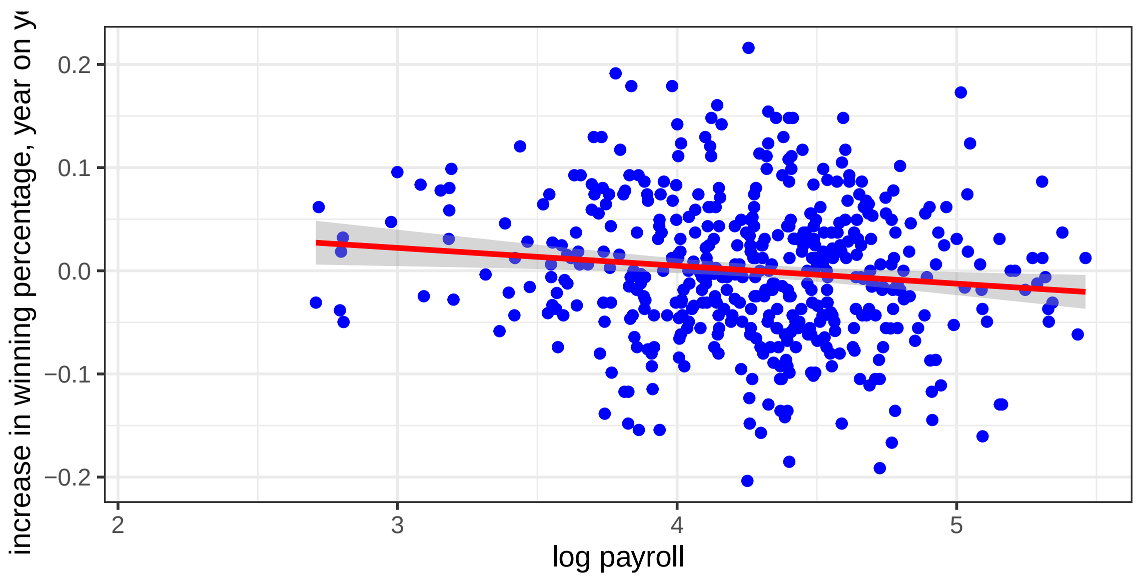



```
summary(lm(diff_win_pct_next~log.pay,data = paydata.long)) # not bad prediction? can be
```

```
##
## 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
```

```
# pay roll to increase in winning percentage current year
ggplot(paydata.long,aes(x=log.pay,y=diff_win_pct))+
  geom_point(color="blue",size=1.5)+
  geom_smooth(method = "lm", formula = y~x, color="red")+
  xlab("log payroll")+
  ylab("increase in winning percentage, year on year")+
  theme_bw()+
  labs("Relationship between Payroll and Performance Increase Next Year, All Years")
```




```
summary(lm(diff_win_pct~log.pay,data = paydata.long)) # not bad prediction? can be spur
```

```
##
## 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
```

```
# The last two prediction power should be taken cautiously. It may stem from some risi
```

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.

```
save.image("hw1.RData")
```