Modern Data Mining, HW 1

Kevin Sun William Walsh Hanson Wang

Due: 11:59PM, Jan. 30th, 2021

Contents

1 Overview

This is a fast-paced course that covers a lot of material. There will be a large amount of references. You may need to do your own research to fill in the gaps in between lectures and homework/projects. It is impossible to learn data science without getting your hands dirty. Please budget your time evenly. Last-minute work ethic will not work for this course.

Homework in this course is different from your usual homework assignment as a typical student. Most of the time, they are built over real case studies. While you will be applying methods covered in lectures, you will also find that extra teaching materials appear here. The focus will be always on the goals of the study, the usefulness of the data gathered, and the limitations in any conclusions you may draw. Always try to challenge your data analysis in a critical way. Frequently, there are no unique solutions.

Case studies in each homework can be listed as your data science projects (e.g. on your CV) where you see fit.

1.1 Objectives

- Get familiar with `R-studio` and `RMarkdown`
- Hands-on R
- Learn data science essentials
  - gather data
  - clean data
  - summarize data
  - display data
  - conclusion
- Packages
  - `dplyr`
  - `ggplot`

1.2 Instructions

- **Homework assignments can be done in a group consisting of up to three members**. Please find your group members as soon as possible and register your group on our Canvas site.
- **All work submitted should be completed in the R Markdown format.** You can find a cheat sheet for R Markdown here. For those who have never used it before, we urge you to start this homework as soon as possible.
- **Submit the following files, one submission for each group:** (1) Rmd file, (2) a compiled PDF or HTML version, and (3) all necessary data files if different from our source data. You may directly edit this .rmd file to add your answers. If you intend to work on the problems separately within your group, compile your answers into one Rmd file before submitting. We encourage that you at least attempt each problem by yourself before working with your teammates. Additionally, ensure that you can 'knit' or compile your Rmd file. It is also likely that you need to configure Rstudio to properly convert files to PDF. **These instructions** might be helpful.
- In general, be as concise as possible while giving a fully complete answer to each question. All necessary datasets are available in this homework folder on Canvas. Make sure to document your code with comments (written on separate lines in a code chunk using a hashtag `#` before the comment) so the teaching fellows can follow along. R Markdown is particularly useful because it follows a 'stream of consciousness' approach: as you write code in a code chunk, make sure to explain what you are doing outside of the chunk.

- A few good or solicited submissions will be used as sample solutions. When those are released, make sure to compare your answers and understand the solutions.

1.3 Review materials

- Study Advanced R Tutorial (to include `dplyr` and `ggplot`)
- Study lecture 1: Data Acquisition and EDA

2 Case study 1: Audience Size

How successful is the Wharton Talk Show Business Radio Powered by the Wharton School

Background: Have you ever listened to SiriusXM? Do you know there is a **Talk Show** run by Wharton professors in Sirius Radio? Wharton launched a talk show called Business Radio Powered by the Wharton School through the Sirius Radio station in January of 2014. Within a short period of time the general reaction seemed to be overwhelmingly positive. To find out the audience size for the show, we designed a survey and collected a data set via MTURK in May of 2014. Our goal was to **estimate the audience size**. There were 51.6 million Sirius Radio listeners then. One approach is to estimate the proportion of the Wharton listeners to that of the Sirius listeners, p, so that we will come up with an audience size estimate of approximately 51.6 million times p.

To do so, we launched a survey via Amazon Mechanical Turk (MTurk) on May 24, 2014 at an offered price of $0.10 for each answered survey. We set it to be run for 6 days with a target maximum sample size of 2000 as our goal. Most of the observations came in within the first two days. The main questions of interest are "Have you ever listened to Sirius Radio" and "Have you ever listened to Sirius Business Radio by Wharton?". A few demographic features used as control variables were also collected; these include Gender, Age and Household Income.

We requested that only people in United States answer the questions. Each person can only fill in the questionnaire once to avoid duplicates. Aside from these restrictions, we opened the survey to everyone in MTurk with a hope that the sample would be more randomly chosen.

The raw data is stored as `Survey_results_final.csv` on Canvas.

```
library(dplyr)
library(ggplot2)
library(dplyr)
library(grid)
library(ggplot2)
library(lattice)
```

2.1 Data preparation

i. We need to clean and select only the variables of interest.

Select only the variables Age, Gender, Education Level, Household Income in 2013, Sirius Listener?, Wharton Listener? and Time used to finish the survey.

Change the variable names to be "age", "gender", "education", "income", "sirius", "wharton", "worktime".

ii. Handle missing/wrongly filled values of the selected variables

As in real world data with user input, the data is incomplete, with missing values, and has incorrect responses. There is no general rule for dealing with these problems beyond "use common sense." In whatever case, explain what the problems were and how you addressed them. Be sure to explain your rationale for your chosen methods of handling issues with the data. Do not use Excel for this, however tempting it might be.

Tip: Reflect on the reasons for which data could be wrong or missing. How would you address each case? For this homework, if you are trying to predict missing values with regression, you are definitely overthinking. Keep it simple.

```
rawdatasurvey <- read.csv("data/Survey_results_final.csv", header=T, stringsAsFactors = FALSE)
keptvars <- c("Answer.Age", "Answer.Gender", "Answer.Education",
              "Answer.HouseHoldIncome", "Answer.Sirius.Radio",
              "Answer.Wharton.Radio","WorkTimeInSeconds")
datasurvey <- rawdatasurvey[keptvars]

datasurvey <- datasurvey %>% rename(age = Answer.Age, gender = Answer.Gender,
                                    education = Answer.Education, income = Answer.HouseHoldIncome,
                                    sirius = Answer.Sirius.Radio, wharton = Answer.Wharton.Radio,
                                    worktime = WorkTimeInSeconds)

# remove people too young or old, non numeric ages, too fast submissions low-quality, remove no income,

datasurvey <- mutate(datasurvey, age = as.numeric(age))
datasurvey <- filter(datasurvey, age >= 10 & age <= 110 & worktime >= 7 & !is.na(age))
datasurvey <- datasurvey[-which(datasurvey$income == ""), ]
datasurvey <- datasurvey[-which(datasurvey$gender == ""), ]
datasurvey <- datasurvey[-which(datasurvey$sirius == ""), ]
datasurvey <- datasurvey[-which(datasurvey$wharton == ""), ]
datasurvey <- datasurvey[-which(datasurvey$education == "select one"), ]
datasurvey <- datasurvey[-which(datasurvey$wharton == "Yes" & datasurvey$sirius == "No"), ]
```

2.2 Data preperation RESPONSE

To clean the data, people too old (>110) and too young (<10) were removed. Furthermore, people without a numeric age were removed. Samples that were too quick (<7seconds) were also removed due to the likely low-quality of their responses. Incomplete entries in the other columns were removed, in addition to people that claimed to listen to the Wharton station without listening to SiriusXM.

```
summary(datasurvey)
```

```
##       age           gender            education            income         
##  Min.   :18.0   Length:1723        Length:1723        Length:1723       
##  1st Qu.:23.0   Class :character   Class :character   Class :character  
##  Median :28.0   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :30.3                                                           
##  3rd Qu.:34.0                                                           
##  Max.   :76.0                                                           
##     sirius            wharton             worktime    
##  Length:1723        Length:1723        Min.   :  8.0  
##  Class :character   Class :character   1st Qu.: 17.0  
##  Mode  :character   Mode  :character   Median : 21.0  
##                                        Mean   : 22.5  
##                                        3rd Qu.: 26.0  
##                                        Max.   :108.0  
```

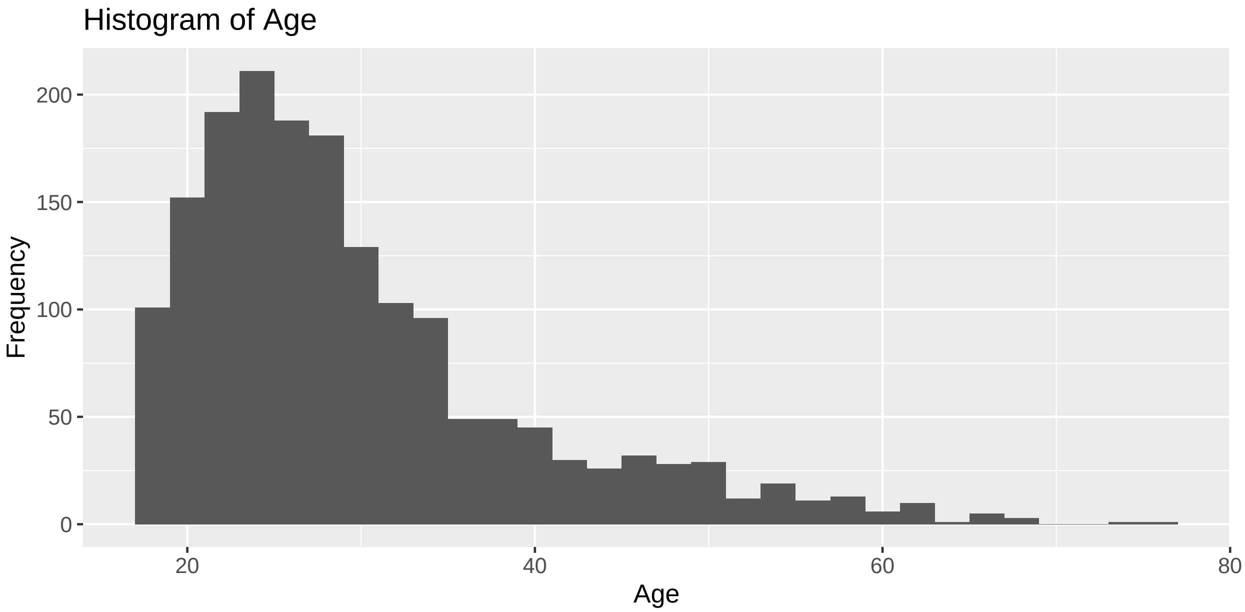
iii. Brief summary

Write a brief report to summarize all the variables collected. Include both summary statistics (including sample size) and graphical displays such as histograms or bar charts where appropriate. Comment on what you have found from this sample. (For example - it's very interesting to think about why would one work for a job that pays only 10cents/each survey? Who are those survey workers? The answer may be interesting even if it may not directly relate to our goal.)

1. Sample Size = 1723

```
p1 <- ggplot(datasurvey) +
geom_histogram(aes(x = age), count = 5) +
labs( title = "Histogram of Age", x = "Age" , y = "Frequency", fill= "blue")

p1 + theme(text = element_text(size = 12))
```



```
incomedf <- as.data.frame(table(datasurvey$income))
incomedf <- incomedf[c(6, 1:5), ] # new order


incomedf <- as.data.frame(table(datasurvey$income))
incomedf <- incomedf %>% rename(income = Var1, count = Freq)
incomedf <- incomedf[c(6, 1:5), ] # new order
incomedf <- incomedf %>% mutate(percent = count / sum(incomedf$count))
incomedf
```

```
##                income count percent
## 6  Less than $15,000   205  0.1190
## 1  $15,000 - $30,000   358  0.2078
## 2  $30,000 - $50,000   419  0.2432
## 3  $50,000 - $75,000   371  0.2153
## 4 $75,000 - $150,000   326  0.1892
## 5     Above $150,000    44  0.0255
```

```
genderdf <- as.data.frame(table(datasurvey$gender))
genderdf <- genderdf %>% rename(gender = Var1, count = Freq)
genderdf <- genderdf %>% mutate(percent = count / sum(genderdf$count))
genderdf
```

```
##   gender count percent
## 1 Female   729   0.423
## 2   Male   994   0.577
```

```
medianage <- median(datasurvey$age)
medianage
```

```
## [1] 28
```

```
edudf <- as.data.frame(table(datasurvey$education))
edudf <- edudf %>% rename(education = Var1, count = Freq)
edudf <- edudf %>% mutate(percent = count / sum(genderdf$count))
edudf
```

```
##                                          education count percent
## 1        Bachelor’s degree or other 4-year degree   611 0.35461
## 2                 Graduate or professional degree   177 0.10273
## 3            High school graduate (or equivalent)   187 0.10853
## 4      Less than 12 years; no high school diploma    10 0.00580
## 5                                            Other     2 0.00116
## 6 Some college, no diploma; or Associate’s degree   736 0.42716
```

```
genderratedf <-datasurvey %>%
  group_by(gender) %>%
  summarise(
    count = sum(age != 0),
    sirius_count = sum(sirius == 'Yes'),
    wharton_count = sum(sirius == 'Yes' & wharton == 'Yes'))

incomeratedf <-datasurvey %>%
  group_by(income) %>%
  summarise(
    count = sum(age != 0),
    sirius_count = sum(sirius == 'Yes'),
    wharton_count = sum(sirius == 'Yes' & wharton == 'Yes'))

ageratedf <-datasurvey %>%
  group_by(gr=cut(age, breaks= seq(0, 100, by = 10)) ) %>%
     summarise(count= n(), sirius_count = sum(sirius == 'Yes'),
    wharton_count = sum(sirius == 'Yes' & wharton == 'Yes'))%>%
     arrange(as.numeric(gr))

genderratedf <- genderratedf %>% mutate(siriuspct = sirius_count / count)
genderratedf <- genderratedf %>% mutate(whartonshare = wharton_count / sirius_count)
genderratedf
```

```
## # A tibble: 2 x 6
##   gender count sirius_count wharton_count siriuspct whartonshare
##   <chr>  <int>        <int>         <int>     <dbl>        <dbl>
## 1 Female   729          555            19     0.761       0.0342
## 2 Male     994          781            48     0.786       0.0615
```

```
incomeratedf <- incomeratedf %>% mutate(siriuspct = sirius_count / count)
incomeratedf <- incomeratedf[c(6, 1:5), ] # new order
incomeratedf <- incomeratedf %>% mutate(whartonshare = wharton_count / sirius_count)
incomeratedf
```

```
## # A tibble: 6 x 6
##   income              count sirius_count wharton_count siriuspct whartonshare
##   <chr>               <int>        <int>         <int>     <dbl>        <dbl>
## 1 Less than $15,000     205          145             7     0.707       0.0483
## 2 $15,000 - $30,000     358          271            10     0.757       0.0369
## 3 $30,000 - $50,000     419          323            12     0.771       0.0372
## 4 $50,000 - $75,000     371          296            17     0.798       0.0574
## 5 $75,000 - $150,000    326          269            17     0.825       0.0632
## 6 Above $150,000         44           32             4     0.727       0.125
```

```
ageratedf <- ageratedf %>% mutate(siriuspct = sirius_count / count)
ageratedf <- ageratedf %>% mutate(whartonshare = wharton_count / sirius_count)
ageratedf
```

```
## # A tibble: 7 x 6
##   gr      count sirius_count wharton_count siriuspct whartonshare
##   <fct>   <int>        <int>         <int>     <dbl>        <dbl>
## 1 (10,20]   158          115             4     0.728       0.0348
## 2 (20,30]   923          718            42     0.778       0.0585
## 3 (30,40]   391          324            15     0.829       0.0463
## 4 (40,50]   158          119             2     0.753       0.0168
## 5 (50,60]    70           45             3     0.643       0.0667
## 6 (60,70]    21           14             0     0.667       0
## 7 (70,80]     2            1             1     0.5         1
```

```
#sirius reaches femals and males equally if similar age distributions
#looks like way more men listen to wharton than females
whartonspct <- sum(incomeratedf$wharton_count) / sum(incomeratedf$count)
whartonspct
```

```
## [1] 0.0389
```

```
whartonpctofsirius <- sum(incomeratedf$wharton_count) / sum(incomeratedf$sirius_count)
whartonpctofsirius
```

```
## [1] 0.0501
```

2.3 Sample properties

The population from which the sample is drawn determines where the results of our analysis can be applied or generalized. We include some basic demographic information for the purpose of identifying sample bias, if any exists. Combine our data and the general population distribution in age, gender and income to try to characterize our sample on hand.

i. Does this sample appear to be a random sample from the general population of the USA?
ii. Does this sample appear to be a random sample from the MTURK population?

Note: You can not provide evidence by simply looking at our data here. For example, you need to find distribution of education in our age group in US to see if the two groups match in distribution. You may need to gather some background information about the MTURK population to have a slight sense if this particular sample seem to a random sample from there… Please do not spend too much time gathering evidence.

2.4 (2.2) Sample properties RESPONSE

From visual inspection of the histogram of survey respondent ages, the sample from this survey is skewed towards younger people. The median age of people in America in 2014 was 37.4 years [1], while the median age of survey respondents was only 28, nearly ten years younger. Thus, this survey population does not represent a random sample of the general US population with respect to age. From inspection of the gender breakdown of this population, males are heavilly over represented. According to the US Census Bureau, the general US population is only 49.2% male– in this survey, respondents were over 57% male. Due to an over representation of men, this population is not a random sample of the US population.[2] Furthermore, this sample skews to lower-income people. According to the US Census Bureau, more than 10% of American Households had incomes above 157,500– less than 3% of sample data came from households with income above 150,000.[3] Thus, the data set is not a random sample of American Households and skews towards households with less income.

According to CloudResearch, 57.5% of participants in MTURK are female, in our dataset 57% of survey responses came from men. Clearly, our data set is not a random sample of the Amazom MTURK population. [5] Income wise, this data set appears to be close to a random sample of the MTRUK data set income wise, with the largest discrepancy coming in the 150k+ income bucket, MTRURK is about 5%, this data is around 3%. Overall, income appears to be the same across MTRUK and the sample. The age distribution appears to match the Amazon MTURK population closer than the overall US population, with the majority of respondents under 40. The median age of MTURK as a whole is a bit higher than 28, our median, according to a graph on CloudResearch [5]. Overall, the sample appears to be a bit younger and more male than MTRUK as a whole.

Sources: 1. https://www.census.gov/data/developers/data-sets/acs-5year.html 2. https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf 3. https://www.census.gov/library/visualizations/2015/demo/distribution-of-household-income--2014.html 5. https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/

2.5 Final estimate

Give a final estimate of the Wharton audience size in January 2014. Assume that the sample is a random sample of the MTURK population, and that the proportion of Wharton listeners vs. Sirius listeners in the general population is the same as that in the MTURK population. Write a brief executive summary to summarize your findings and how you came to that conclusion.

To be specific, you should include:

1. Goal of the study
2. Method used: data gathering, estimation methods
3. Findings

4. Limitations of the study.

```
numhouseholds <- 123.23 #statistia 2014

# https://www.statista.com/statistics/183635/number-of-households-in-the-us/

avghhsize <- 2.54

# https://www.statista.com/statistics/183648/average-size-of-households-in-the-us/

incomeratedf <- incomeratedf %>% mutate(hhcount = c(15657,
20007,
22579,
21227,
31034,
14081
))
# https://en.wikipedia.org/wiki/Household_income_in_the_United_States#Distribution_of_household_income_

incomeratedf <- incomeratedf %>% mutate(siriushhs = siriuspct*hhcount)
```

2.6 (2.3) Final estimate Response

1. The Goal of this study is to estimate the Wharton audience size in the United States by using the sample data from the MTURK population. This study operates under two critical limiting assumptions- the sample is a random sample of the MTURK population, and that the proportion of Wharton listeners vs. Sirius listeners in the general population is the same as that in the MTURK population.Thus, the goal of this study is to estimate the audience size of Sirius XM in the United States using the MTURK data, and apply the same proportion of wharton listeners of SiriusXM listeners as MTURK.

2. The method this study uses to estimate the audience size of the Wharton station, through estimating Sirius XM's audience size, will use different penetration rates across household income buckets. SiriusXM penetration varied along both age and income ranges, but due to limited samples from older age ranges (less than 100 sample points for people over 50 years old), income-varying penetration rates were selected for estimating SiriusXM's reach. Having far more data points for different income buckets representing the vast majority of the US Population (only around 10% of the population was in households with income greater than 150k). Using data from the US Census Bureau, the number of households in each income bracket can be found. Using this methodology, a total number of SiriusXM listening households is found. Combining this with the average household size in 2014 and the penetration of Wharton compared to SiriusXM found in the MTRURK data, a total audience estimation for the Wharton channel is estimated.

3. From this data set and the above explained methodology, the Wharton channel's audiance is estimated to be 12.3 million people, or around 3.8% of the 2014 US Population.

```
whartonaudiance <- sum(incomeratedf$siriushhs) * avghhsize * 1000 * whartonpctofsirius
whartonaudiance
```

```
## [1] 12280673
```

```
estpct <- whartonaudiance / (318400*1000)
estpct
```

```
## [1] 0.0386
```

4. The greatest limitation to this study is the "income wharton listeining propensity homogeneity" introduced by the limiting assumption in the problem. When we remove this limiting assumption, the audiance size increases by around 50% because wealthier households have a higher propensity to listen to the Wharton channel. Further limitations include a constant household size across income brackets and not considering the impact of age.

2.7 New task

Now suppose you are asked to design a study to estimate the audience size of Wharton Business Radio Show as of today: You are given a budget of $1000. You need to present your findings in two months.

Write a proposal for this study which includes:

1. Method proposed to estimate the audience size.
2. What data should be collected and where it should be sourced from. Please fill in the google form to list your platform where surveys will be launched and collected HERE

A good proposal will give an accurate estimation with the least amount of money used.

1. Google Display Network ads paid per-click on general population websites (news) or a balanced approach to gender segments (sports, cooking recipe sites) and second order organic referrals to the form.

2. To estimate the Wharton audience size, we must reach SiriusXM listeners. Doing this direct, though ads on SiriusXM stations is prohibitively expensive– SiriusXM requires a minimum monthly spend of 10,000. 1 Alternatively, and cheaper, we could run ads drawing only SiriusXM listeners through Google Banner Ads. Running an ad asking website viewers to "vote on new SiriusXM Channels," paying on a cost-per-click basis, would ensure most people who click the ad listen to SiriusXM. On top of paid inbound traffic, the top of the linked form would include an easy to share feature so all users, especially mobile users, could forward the link to their friends that listen to SiriusXM. This would create a second, free, organic traffic source. People would be motivated to share the survey under the belief that this directly influences future SiriusXM channels. In the survey, we would first ask for the desired information, and sneak in "do you listen to the Wharton channel" in the middle. The cost-per-click on the google display network is under 1 dollar for most industries, so we would likely generate around 1000 clicks to the survey. From there a portion would fill it out, and a smaller portion would share with friends. With enough organic referrals, under the guise that respondents are impacting future channel lineups, the survey would likely generate over 1000 data points of people interested in changing SiriusXM's lineup, a population likely resembling the underlying SiriusXM listening base.

Perhaps the largest challenge in estimating Wharton's reach is collecting enough data points for the upper income buckets since these were observed to be more likely to listen to the channel. By promising a non-monetary benefit that they would value, like influencing new channels, this could possibly reach these wealthier demographics. Organic referrals within these demographics would strengthen this method's reach of richer individuals.

CPC for Google Ads: https://www.webfx.com/blog/marketing/much-cost-advertise-google-adwords/

3 Case study 2: Women in Science

Are women underrepresented in science in general? How does gender relate to the type of educational degree pursued? Does the number of higher degrees increase over the years? In an attempt to answer these questions, we assembled a data set (WomenData_06_16.xlsx) from NSF about various degrees granted in the U.S. from 2006 to 2016. It contains the following variables: Field (Non-science-engineering (Non-S&E) and sciences (Computer sciences, Mathematics and statistics, etc.)), Degree (BS, MS, PhD), Sex (M, F), Number of degrees granted, and Year.

Our goal is to answer the above questions only through EDA (Exploratory Data Analyses) without formal testing. We have provided sample R-codes in the appendix to help you if needed.

3.1 Data preparation

1. Understand and clean the data

Notice the data came in as an Excel file. We need to use the package readxl and the function read_excel() to read the data WomenData_06_16.xlsx into R.

i. Read the data into R.
ii. Clean the names of each variables. (Change variable names to Field,Degree, Sex, Year and Number)
iii. Set the variable natures properly.
iv. Any missing values?

There are no missing values.

```
womendata <- read_excel("data/WomenData_06_16.xlsx")
womendata <- womendata %>% rename(Field = "Field and sex",
                                  Degree = "Degree", Sex = "Sex",
                                  Year = "Year",
                                  Number = "Degrees Awarded")
womendata %<>%
  mutate(Field = as.factor(Field), Degree = as.factor(Degree), Sex = as.factor(Sex))

which(is.na(womendata$Field))
```

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

```
which(is.na(womendata$Degree))
```

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

```
which(is.na(womendata$Sex))
```

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

```
which(is.na(womendata$Year))
```

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

```
which(is.na(womendata$Number))
```

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

```
summary(womendata)
```

```
##                                     Field     Degree        Sex
##  Agricultural sciences                 : 66   BS :220   Female:330
##  Biological sciences                   : 66   MS :220   Male  :330
##  Computer sciences                     : 66   PhD:220
##  Earth, atmospheric, and ocean sciences: 66
##  Engineering                           : 66
##  Mathematics and statistics            : 66
##  (Other)                               :264
##       Year          Number
##  Min.   :2006   Min.   :   218
##  1st Qu.:2008   1st Qu.:  2118
##  Median :2011   Median :  6020
##  Mean   :2011   Mean   : 41717
##  3rd Qu.:2014   3rd Qu.: 18127
##  Max.   :2016   Max.   :781474
##
```

2. Write a summary describing the data set provided here.

i. How many fields are there in this data?

```
length(unique(womendata$Field))
```

```
## [1] 10
```

ii. What are the degree types?

```
unique(womendata$Degree)
```

```
## [1] BS  MS  PhD
## Levels: BS MS PhD
```

iii. How many year's statistics are being reported here?

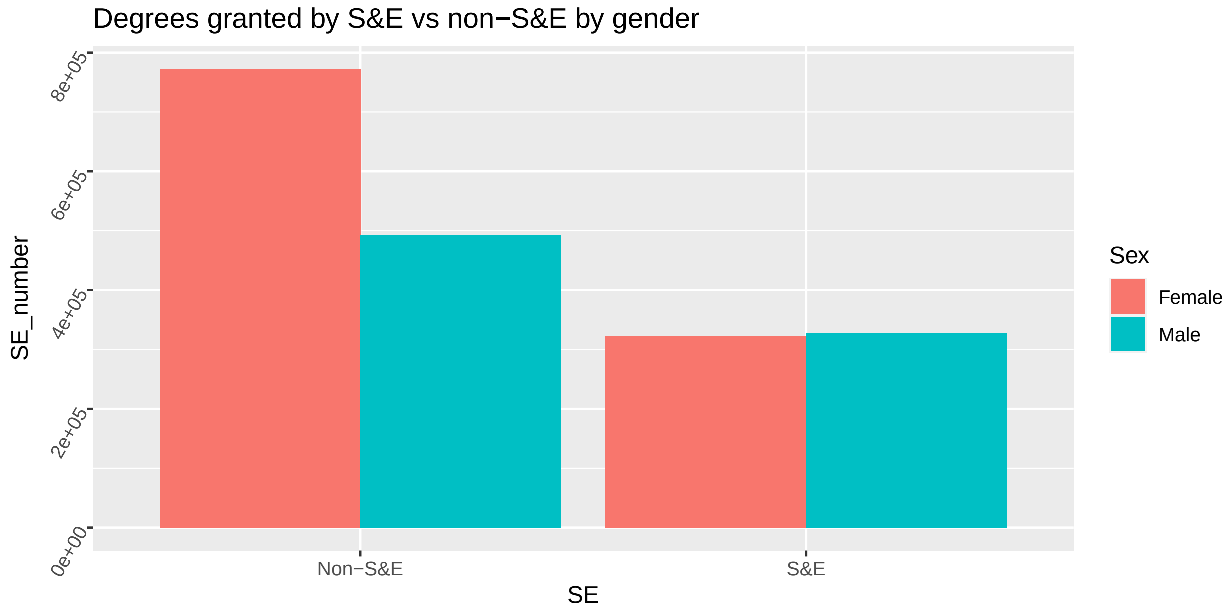
```
length(unique(womendata$Year))
```

```
## [1] 11
```

3.2 BS degrees in 2015

Is there evidence that more males are in science-related fields vs `Non-S&E`? Provide summary statistics and a plot which shows the number of people by gender and by field. Write a brief summary to describe your findings.

```
womendataBS2015 <- filter(womendata, Degree == "BS" & Year == 2015)
womendataBS2015 %>%
  mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(SE, Sex) %>%
  summarise(SE_number = sum(Number)) %>%
  ggplot(aes(x = SE, y = SE_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.y = element_text(angle = 60)) +
  ggtitle("Degrees granted by S&E vs non-S&E by gender")
```



```
womendataBS2015 %>%
  group_by(Sex) %>%
  summarise(deg = mean(Number))
```

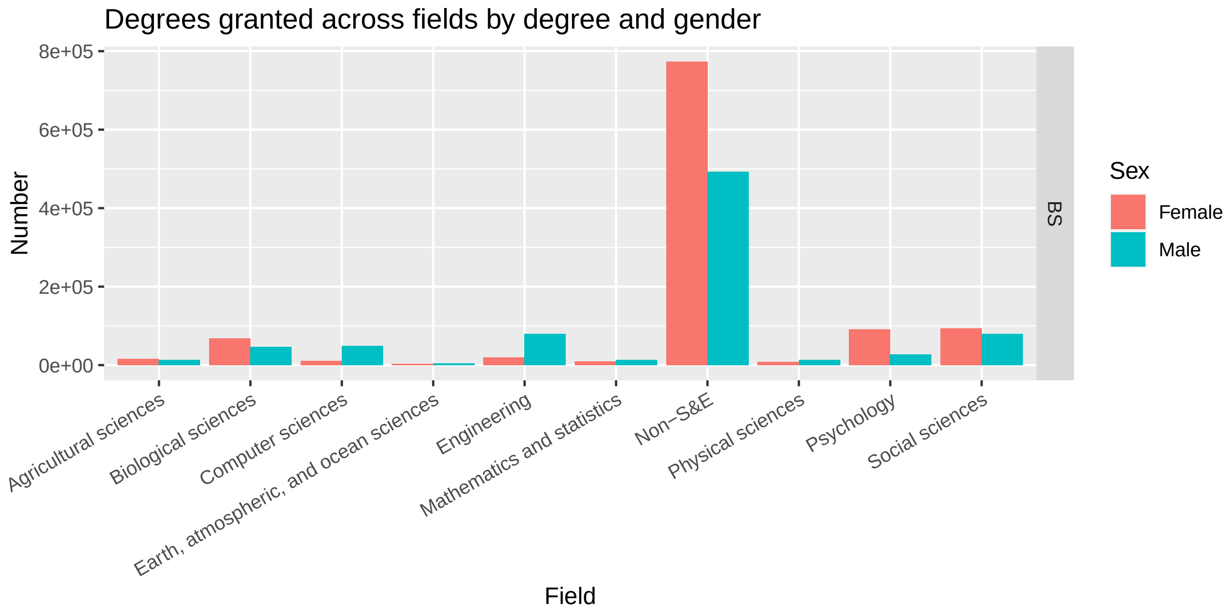
```
## # A tibble: 2 x 2
##   Sex       deg
##   <fct>   <dbl>
## 1 Female 109570.
## 2 Male    82043.
```

First, it is important to note that there are more females than males in this dataset (2015, BS).

```
womendataBS2015 %>%  # to get the average number of ppl by gender
  group_by(Field, Sex) %>%
  summarise(deg = mean(Number))
```

```
## # A tibble: 20 x 3
## # Groups:   Field [10]
##    Field                                  Sex       deg
##    <fct>                                  <fct>   <dbl>
##  1 Agricultural sciences                  Female  16234
##  2 Agricultural sciences                      Male    13226
##  3 Biological sciences                        Female  68570
##  4 Biological sciences                        Male    46554
##  5 Computer sciences                          Female  10863
##  6 Computer sciences                          Male    49446
##  7 Earth, atmospheric, and ocean sciences Female   2701
##  8 Earth, atmospheric, and ocean sciences Male     4454
##  9 Engineering                                Female  20057
## 10 Engineering                                Male    79849
## 11 Mathematics and statistics                 Female   9922
## 12 Mathematics and statistics                 Male    13214
## 13 Non-S&E                                    Female 772768
## 14 Non-S&E                                    Male   493304
## 15 Physical sciences                          Female   8765
## 16 Physical sciences                          Male    13716
## 17 Psychology                                 Female  91688
## 18 Psychology                                 Male    27080
## 19 Social sciences                            Female  94135
## 20 Social sciences                            Male    79583
```

```
womendataBS2015 %>%
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Degree~., scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across fields by degree and gender")
```



We plot the absolute number of people by gender and by field. As expected by the fact that there are more women in the dataset, there are more females in non-S&E fields than men. However, there are more men in select fields such as Math & Statistics, Physical Sciences, Earch/Atmospheric/Ocean Sciences, Computer Sciences, and Engineering.

```
womendata %>%
  filter(Degree == "BS" & Year == 2015)  %>%
  mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(SE, Sex, Year) %>%
  summarise(SE_number = sum(Number)) %>%
  group_by(SE, Year) %>%
  mutate(ratio = SE_number / sum(SE_number)) %>%
  filter(Sex == "Female")
```

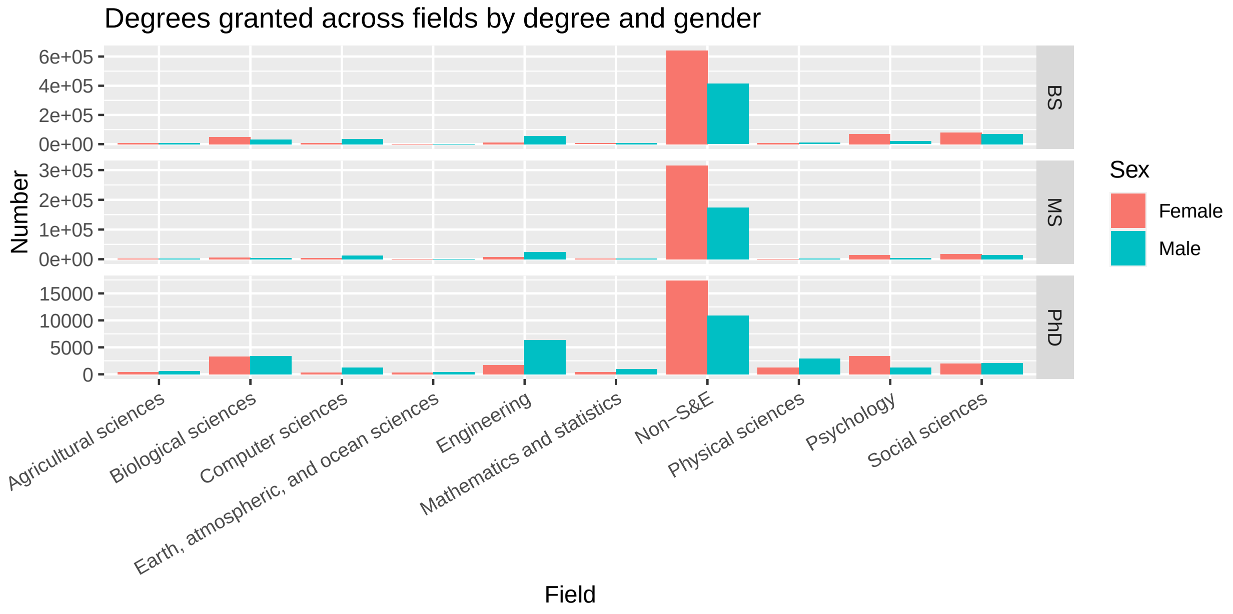
```
## # A tibble: 2 x 5
## # Groups:   SE, Year [2]
##   SE      Sex      Year SE_number ratio
##   <chr>   <fct>   <dbl>     <dbl> <dbl>
## 1 Non-S&E Female   2015    772768 0.610
## 2 S&E     Female   2015    322935 0.497
```

Charting the ratio of females in Non-S&E and S&E fields, we can see that there is not sufficient evidence of more males in either degree type. In fact, there are more females in non-S&E fields, and an approximately equal gender ratio in S&E fields.

3.3 EDA bringing type of degree, field and gender in 2015

Describe the number of people by type of degree, field, and gender. Do you see any evidence of gender effects over different types of degrees? Again, provide graphs to summarize your findings.

```
womendata %>%
  filter(Year == 2007) %>%
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Degree~., scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across fields by degree and gender")
```

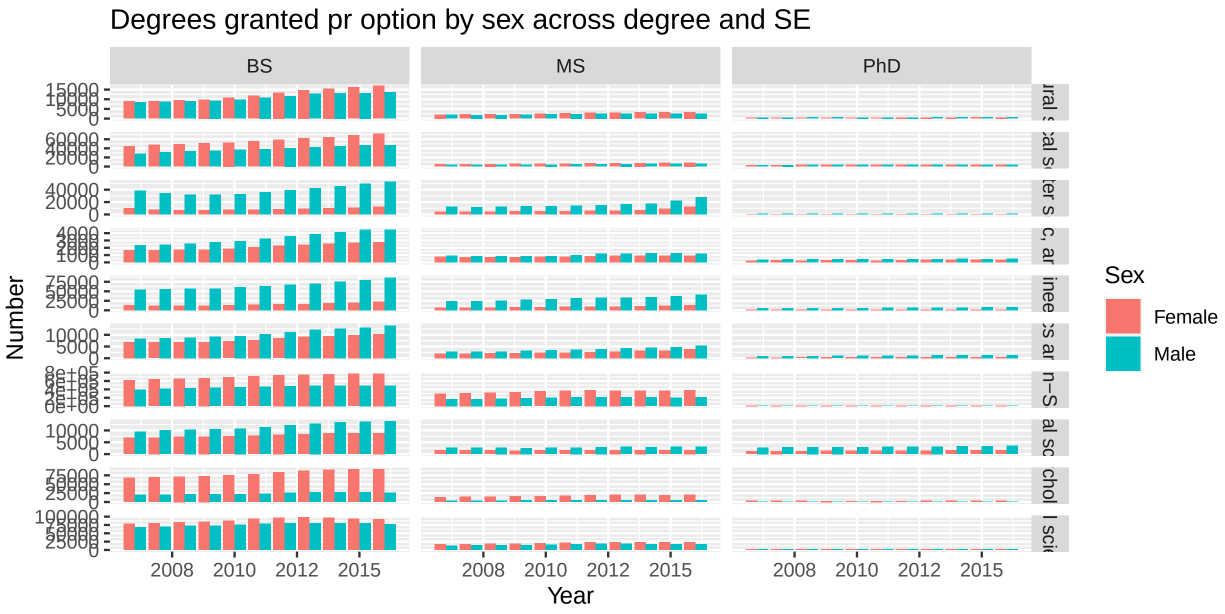


Across all 3 degrees, we see more females than males in non-S&E fields. Some noticeable differences across degrees is that there are more males in Social Sciences at the PHD level, but more females in Social Sciences at the MS and BS levels. Similarly, the discrepancy (more males than females) is more noticeable at the PHD level in Engineering and Computer Sciences fields.

3.4 EDA bring all variables

In this last portion of the EDA, we ask you to provide evidence numerically and graphically: Do the number of degrees change by gender, field, and time?

```
plot3.4 <- ggplot(womendata, aes(x = Year, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Field~Degree, scales = "free_y") +
  ggtitle("Degrees granted pr option by sex across degree and SE")
plot3.4
```



```
print_output(ggplot_build(plot3.4)$data[[1]])
```

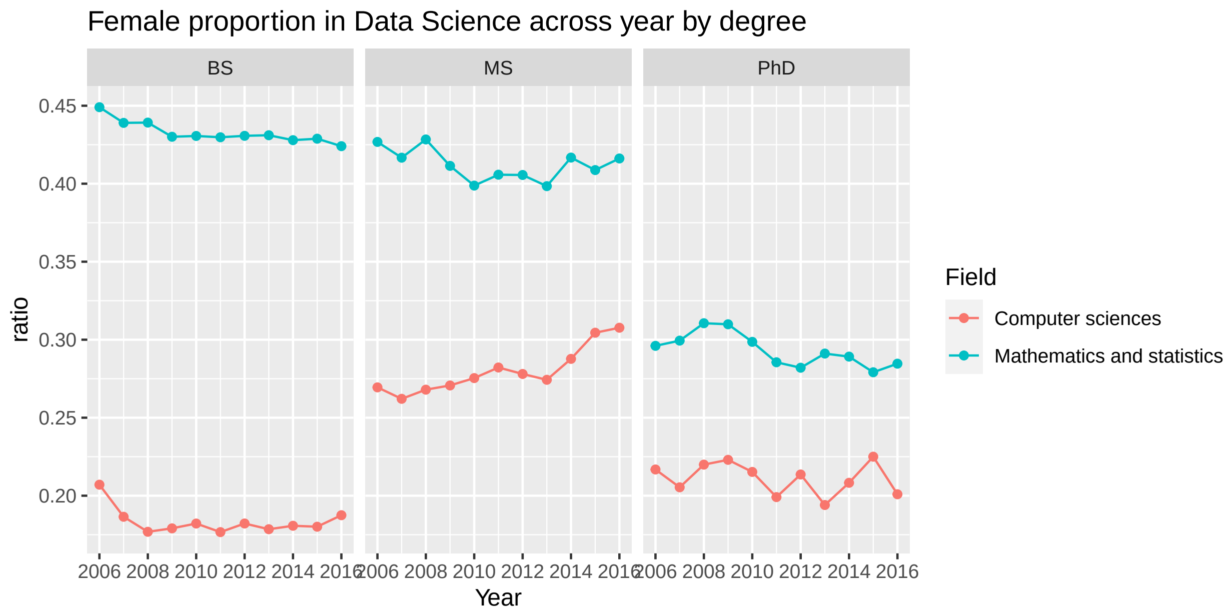
```
         fill    x     y PANEL group flipped_aes ymin   ymax xmin xmax colour
   1  #F8766D 2006  8909     1     1       FALSE    0   8909 2006 2006     NA
   2  #F8766D 2007  8915     1     1       FALSE    0   8915 2007 2007     NA
   3  #F8766D 2008  9457     1     1       FALSE    0   9457 2008 2008     NA
   4  #F8766D 2009  9818     1     1       FALSE    0   9818 2009 2009     NA
   5  #F8766D 2010 10709     1     1       FALSE    0  10709 2010 2010     NA
   6  #F8766D 2011 11855     1     1       FALSE    0  11855 2011 2011     NA
   7  #F8766D 2012 13444     1     1       FALSE    0  13444 2012 2012     NA
   8  #F8766D 2013 14826     1     1       FALSE    0  14826 2013 2013     NA
   9  #F8766D 2014 15525     1     1       FALSE    0  15525 2014 2014     NA
   10 #F8766D 2015 16234     1     1       FALSE    0  16234 2015 2015     NA
   11 #F8766D 2016 16934     1     1       FALSE    0  16934 2016 2016     NA
   12 #00BFC4 2006  8398     1     2       FALSE    0   8398 2006 2006     NA
   13 #00BFC4 2007  8781     1     2       FALSE    0   8781 2007 2007     NA
   14 #00BFC4 2008  9017     1     2       FALSE    0   9017 2008 2008     NA
```

Here, we have graphed visually and numerically (by extracting the ggplot data) the number of degrees across gender (bar color), field, and time (x-axis). First, we discuss the changes over time. The number of degrees generally has a slight increase over time, which is more prominent at the BS level. Across genders, males still have a larger number of total degrees, and this dispartiy is most prominent in Engineering and Computer Science fields. This disparity has not been alleviated over the time, despite growing number of degrees in both genders. The fields with largely more females are still Psychology, Non-S&E, etc., which further supprots the lack of representation of women in STEM fields.

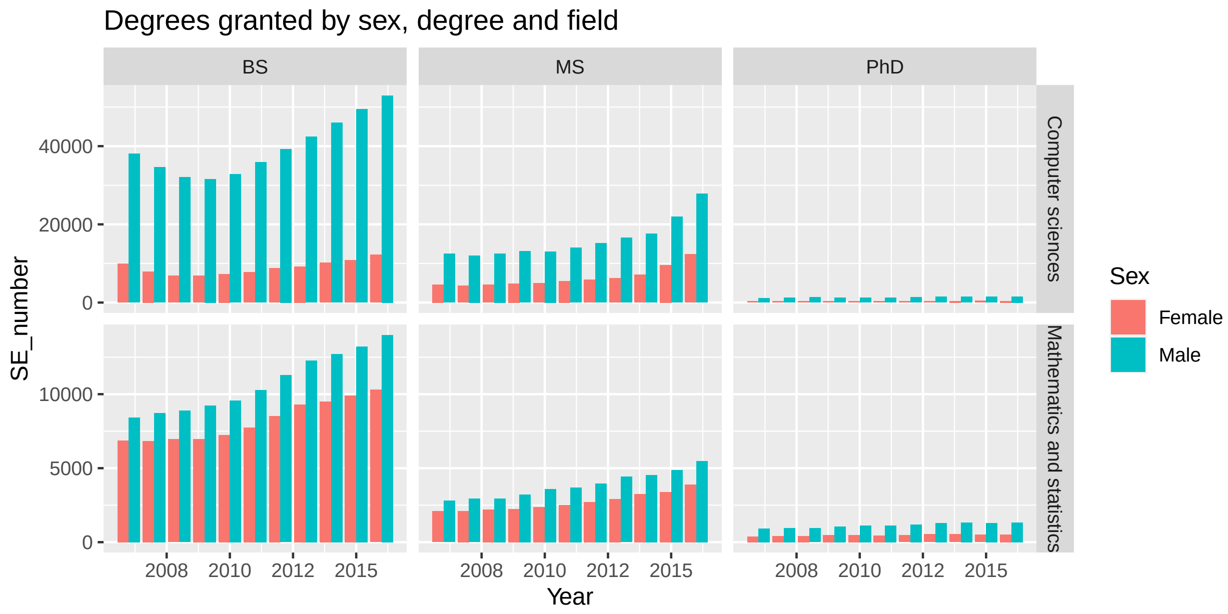
3.5 Women in Data Science

Finally, is there evidence showing that women are underrepresented in data science? Data science is an interdisciplinary field of computer science, math, and statistics. You may include year and/or degree.

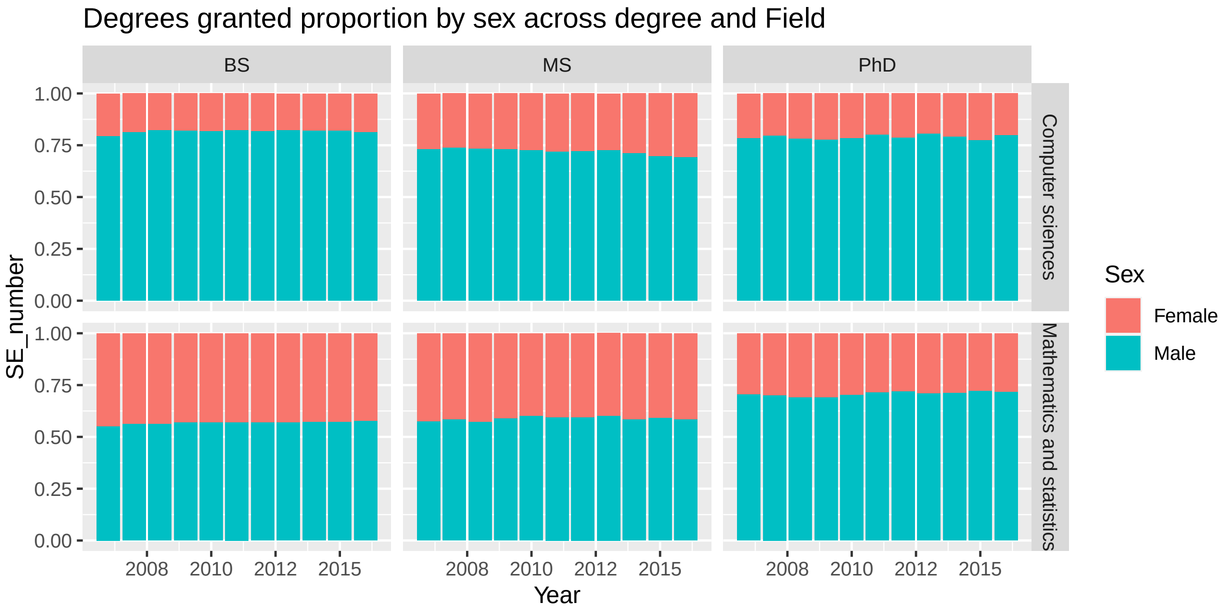
```
womendata %>%
  filter(Field == "Computer sciences" | Field == "Mathematics and statistics") %>%
  # mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(Field, Sex, Year, Degree) %>%
  summarise(SE_number = sum(Number)) %>%
  group_by(Field, Year, Degree) %>%
  mutate(ratio = SE_number / sum(SE_number)) %>%
  filter(Sex == "Female") %>%
  ggplot(aes(x = Year, y = ratio, color = Field)) +
  geom_point() + geom_line() +
  facet_grid(~Degree)+
  ggtitle("Female proportion in Data Science across year by degree")
```



```
womendata %>%
  filter(Field == "Computer sciences" | Field == "Mathematics and statistics") %>%
  # mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(Field, Sex, Year, Degree) %>%
  summarise(SE_number = sum(Number)) %>%
  ggplot(aes(x = Year, y = SE_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Field~Degree, scales = "free_y") +
  ggtitle("Degrees granted by sex, degree and field")
```



```
womendata %>%
  filter(Field == "Computer sciences" | Field == "Mathematics and statistics") %>%
  # mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(Field, Sex, Year, Degree) %>%
  summarise(SE_number = sum(Number)) %>%
  ggplot(aes(x = Year, y = SE_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "fill") +
  facet_grid(Field~Degree, scales = "free_y") +
  ggtitle("Degrees granted proportion by sex across degree and Field")
```



From the line graph of the female proportion in data science across year by degree, we see that in all 3 degrees (BS, MS, PHD), and both subfields of data science (CS, Math and Stats), we see the female proportion is significantly lower than males. This difference is especially pronounced in the Bachelors degree, without more than double the men than women. Furthermore, this problem is not fixed over time, except potentially in the Masters degree, with increasing female proportion in specifiically Computer Science. The bar charts support this claim, but give absolute numbers rather than the proportion.

3.6 Final brief report

Summarize your findings focusing on answering the questions regarding if we see consistent patterns that more males pursue science-related fields. Any concerns with the data set? How could we improve on the study?

As mentioned in the first part of this case problem, one concern of the dataset is that there is a class imbalance. Specificially, there are more females in the datset, and more datapoints in the non-S&E, computer sciences, and engineering fields, as these are more popular fields/majors. To improve this class imbalance, undersampling the heavy classes or collecting more datapoints on the lighter classes could be used.

Looking past the large number of famles in the dataset, we first focused our analysis on Bachelors degree in 2015. THere were more women than men in certain non STEM fields. Expanding the analysis to other degrees, we see similar results at the Masters and PHD levels, although there are fewer number of people with those degree levels. Expanding the anlaysis once more time to inlcude all years, we see that more people are receiving degrees at all 3 levels over time. However, the gender difference is not alleviated over time, especiall in Engineering, Computer Science, and Math/Statistics fields.

3.7 Appendix

To help out, we have included some R-codes here as references. You should make your own chunks filled with texts going through each items listed above. Make sure to hide the unnecessary outputs/code etc.

1. Clean data
2. A number of sample analyses

4 Case study 3: Major League Baseball

We would like to explore how payroll affects performance among Major League Baseball teams. The data is prepared in two formats record payroll, winning numbers/percentage by team from 1998 to 2014.

Here are the datasets:

-`MLPayData_Total.csv`: wide format -`baseball.csv`: long format

Feel free to use either dataset to address the problems.

4.1 EDA: Relationship between payroll changes and performance

Payroll may relate to performance among ML Baseball teams. One possible argument is that what affects this year's performance is not this year's payroll, but the amount that payroll increased from last year. Let us look into this through EDA.

Create increment in payroll

i. To describe the increment of payroll in each year there are several possible approaches. Take 2013 as an example:

   - option 1: diff: payroll_2013 - payroll_2012
   - option 2: log diff: log(payroll_2013) - log(payroll_2012)

```
baseball<-read.csv("data/MLPayData_Total.csv")
names(baseball)
```

```
##  [1] "Team.name.2014" "p1998"          "p1999"          "p2000"
##  [5] "p2001"          "p2002"          "p2003"          "p2004"
##  [9] "p2005"          "p2006"          "p2007"          "p2008"
## [13] "p2009"          "p2010"          "p2011"          "p2012"
## [17] "p2013"          "p2014"          "X2014"          "X2013"
## [21] "X2012"          "X2011"          "X2010"          "X2009"
## [25] "X2008"          "X2007"          "X2006"          "X2005"
## [29] "X2004"          "X2003"          "X2002"          "X2001"
## [33] "X2000"          "X1999"          "X1998"          "X2014.pct"
## [37] "X2013.pct"      "X2012.pct"      "X2011.pct"      "X2010.pct"
## [41] "X2009.pct"      "X2008.pct"      "X2007.pct"      "X2006.pct"
## [45] "X2005.pct"      "X2004.pct"      "X2003.pct"      "X2002.pct"
## [49] "X2001.pct"      "X2000.pct"      "X1999.pct"      "X1998.pct"
```

Explain why the log difference is more appropriate in this setup.

The logarithmic is helpful when the data covers an extensive range of values. Using the logarithms of the values rather than the actual values reduces a wide range to a more manageable size. In the case of this dataset, most baseball teams' total payroll more than quadrupled for 16 years. Without doing the logarithmic transformation, the increases in payroll would seem much more significant in the later years than the beginning years, potentially skewing the results of our regression and PCA analysis results. Therefore, taking the log difference is way more appropriate in this setup.

ii. Create a new variable `diff_log=log(payroll_2013) - log(payroll_2012)`. Hint: use `dplyr::lag()` function.

```
diff_log = log(baseball[3:18]) - log(baseball[2:17])
combined <- cbind(baseball,diff_log)
dim(combined)
```

```
## [1] 30 68
```

```
filtered1 = cbind(combined[1],combined[36:52])
filtered2 = cbind(combined[1],combined[53:68])
filtered3 = cbind(combined[1],combined[2:18])
```

iii. Create a long data table including: team, year, diff_log, win_pct

```
long_filtered1 = filtered1 %>% pivot_longer(!Team.name.2014, names_to =  "year",
                                            names_prefix = "X", values_to = "win_pct")
long_filtered2 = filtered2 %>% pivot_longer(!Team.name.2014, names_to =  "year",
                                            names_prefix = "p", values_to = "diff_log")
olClean <- function(x){
    x <- str_remove(x,".pct")
}
long_filtered3 = filtered3 %>% pivot_longer(!Team.name.2014, names_to =  "year",
                                            names_prefix = "p", values_to = "total_pay")
long_filtered1["year"] = sapply(long_filtered1["year"],olClean)
Final_df = full_join(long_filtered1,long_filtered2,by = c("Team.name.2014","year"))
Final_df = full_join(Final_df,long_filtered3,by = c("Team.name.2014","year"))
names(Final_df)[names(Final_df) == 'Team.name.2014'] <- 'Team'
head(Final_df)
```

```
## # A tibble: 6 x 5
##   Team                 year  win_pct diff_log total_pay
##   <fct>                <chr>   <dbl>    <dbl>     <dbl>
## 1 Arizona Diamondbacks 2014    0.395    0.235     113.
## 2 Arizona Diamondbacks 2013    0.5      0.182      89.1
## 3 Arizona Diamondbacks 2012    0.5      0.326      74.3
## 4 Arizona Diamondbacks 2011    0.580   -0.124      53.6
## 5 Arizona Diamondbacks 2010    0.401   -0.192      60.7
## 6 Arizona Diamondbacks 2009    0.432    0.106      73.6
```

```
summary(Final_df)
```

```
##                     Team         year              win_pct         diff_log
##  Arizona Diamondbacks: 17   Length:510         Min.   :0.265   Min.   :-1.39
##  Atlanta Braves      : 17   Class :character   1st Qu.:0.444   1st Qu.:-0.06
##  Baltimore Orioles   : 17   Mode  :character   Median :0.500   Median : 0.08
##  Boston Red Sox      : 17                      Mean   :0.500   Mean   : 0.07
##  Chicago Cubs        : 17                      3rd Qu.:0.556   3rd Qu.: 0.20
##  Chicago White Sox   : 17                      Max.   :0.716   Max.   : 1.26
##  (Other)             :408                                      NA's   :30
##    total_pay
##  Min.   :  8.3
##  1st Qu.: 51.3
##  Median : 73.3
##  Mean   : 78.1
##  3rd Qu.: 95.0
##  Max.   :235.3
##
```

4.2 Exploratory questions

i. Which five teams had highest increase in their payroll between years 2010 and 2014, inclusive?

```
explore1 <- combined[,c("Team.name.2014","X2010","X2014")]
explore1["payroll"] = explore1["X2010"] - explore1["X2014"]
library(dplyr)
arrange(explore1,payroll)
```

```
##           Team.name.2014 X2010 X2014 payroll
## 1     Pittsburgh Pirates    57    88     -31
## 2      Baltimore Orioles    66    96     -30
## 3   Washington Nationals    69    96     -27
## 4       Seattle Mariners    61    87     -26
## 5     Kansas City Royals    67    89     -22
## 6     Los Angeles Angels    80    98     -18
## 7      Cleveland Indians    69    85     -16
## 8    Los Angeles Dodgers    80    94     -14
## 9         Detroit Tigers    81    90      -9
## 10     Oakland Athletics    81    88      -7
## 11     Milwaukee Brewers    77    82      -5
## 12   St. Louis Cardinals    86    90      -4
## 13         New York Mets    79    79       0
## 14  Arizona Diamondbacks    65    64       1
## 15          Chicago Cubs    75    73       2
## 16     Toronto Blue Jays    85    83       2
## 17         Miami Marlins    80    77       3
## 18  San Francisco Giants    92    88       4
## 19        Houston Astros    76    70       6
## 20      New York Yankees    95    84      11
## 21        Atlanta Braves    91    79      12
## 22      San Diego Padres    90    77      13
## 23     Chicago White Sox    88    73      15
## 24       Cincinnati Reds    91    76      15
## 25      Colorado Rockies    83    66      17
## 26        Boston Red Sox    89    71      18
## 27        Tampa Bay Rays    96    77       19
## 28         Texas Rangers    90    67       23
## 29       Minnesota Twins    94    70       24
## 30 Philadelphia Phillies    97    73       24
```

Pirates, Orioles, Nationals, Mariners, Royals.

ii. Between 2010 and 2014, inclusive, which team(s) “improved” the most? That is, had the biggest percentage gain in wins?

```
explore2 <- combined[,c("Team.name.2014","X2010.pct","X2014.pct")]
explore2["win"] = combined["X2010.pct"] - combined["X2014.pct"]
library(dplyr)
arrange(explore2,win)
```

```
##           Team.name.2014 X2010.pct X2014.pct      win
## 1      Pittsburgh Pirates     0.352     0.543 -0.19136
## 2       Baltimore Orioles     0.407     0.593 -0.18519
## 3    Washington Nationals     0.426     0.593 -0.16667
## 4        Seattle Mariners     0.377     0.537 -0.16049
## 5      Kansas City Royals     0.414     0.549 -0.13580
## 6      Los Angeles Angels     0.494     0.605 -0.11111
## 7       Cleveland Indians     0.426     0.525 -0.09877
## 8     Los Angeles Dodgers     0.494     0.580 -0.08642
## 9          Detroit Tigers     0.500     0.556 -0.05556
## 10      Oakland Athletics     0.500     0.543 -0.04321
## 11      Milwaukee Brewers     0.475     0.506 -0.03086
## 12    St. Louis Cardinals     0.531     0.556 -0.02469
## 13          New York Mets     0.488     0.488  0.00000
## 14   Arizona Diamondbacks     0.401     0.395  0.00617
## 15       Toronto Blue Jays     0.525     0.512  0.01235
## 16           Chicago Cubs     0.463     0.451  0.01235
## 17           Miami Marlins     0.494     0.475  0.01852
## 18    San Francisco Giants     0.568     0.543  0.02469
## 19         Houston Astros     0.469     0.432  0.03704
## 20       New York Yankees     0.586     0.519  0.06790
## 21         Atlanta Braves     0.562     0.488  0.07407
## 22       San Diego Padres     0.556     0.475  0.08025
## 23      Chicago White Sox     0.543     0.451  0.09259
## 24        Cincinnati Reds     0.562     0.469  0.09259
## 25       Colorado Rockies     0.512     0.407  0.10494
## 26         Boston Red Sox     0.549     0.438  0.11111
## 27         Tampa Bay Rays     0.593     0.475  0.11728
## 28          Texas Rangers     0.556     0.414  0.14198
## 29  Philadelphia Phillies     0.599     0.451  0.14815
## 30        Minnesota Twins     0.580     0.432  0.14815
```

Pirates, Orioles, Nationals, Mariners, Royals etc. improved the most with the highest percentage gain in wins.

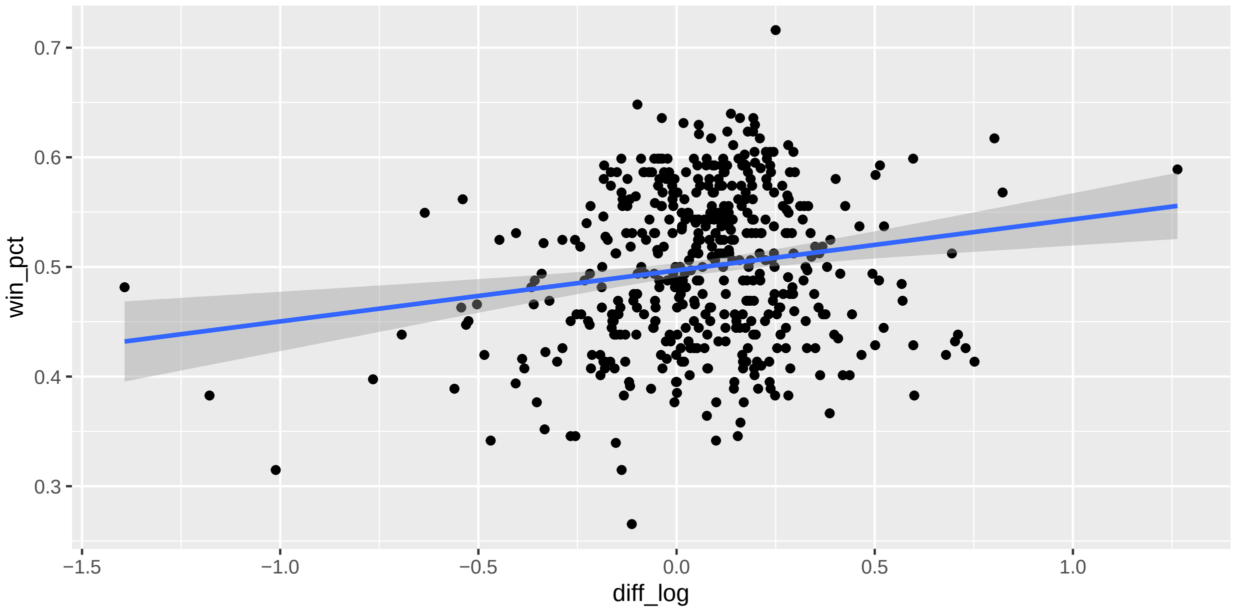
4.3 Do log increases in payroll imply better performance?

Is there evidence to support the hypothesis that higher increases in payroll on the log scale lead to increased performance?

Pick up a few statistics, accompanied with some data visualization, to support your answer.

From the graph below, we see that there is evidence to support the hypothesis that higher increase in payroll on the log scale leads to higher performance. This is supported by a positive coefficient with a t-value of 3.7 that is statistically significant.

```
Final_df %>% ggplot() + aes(x = diff_log, y = win_pct) + geom_point() +
  geom_smooth(method='lm', formula= y~x)
```



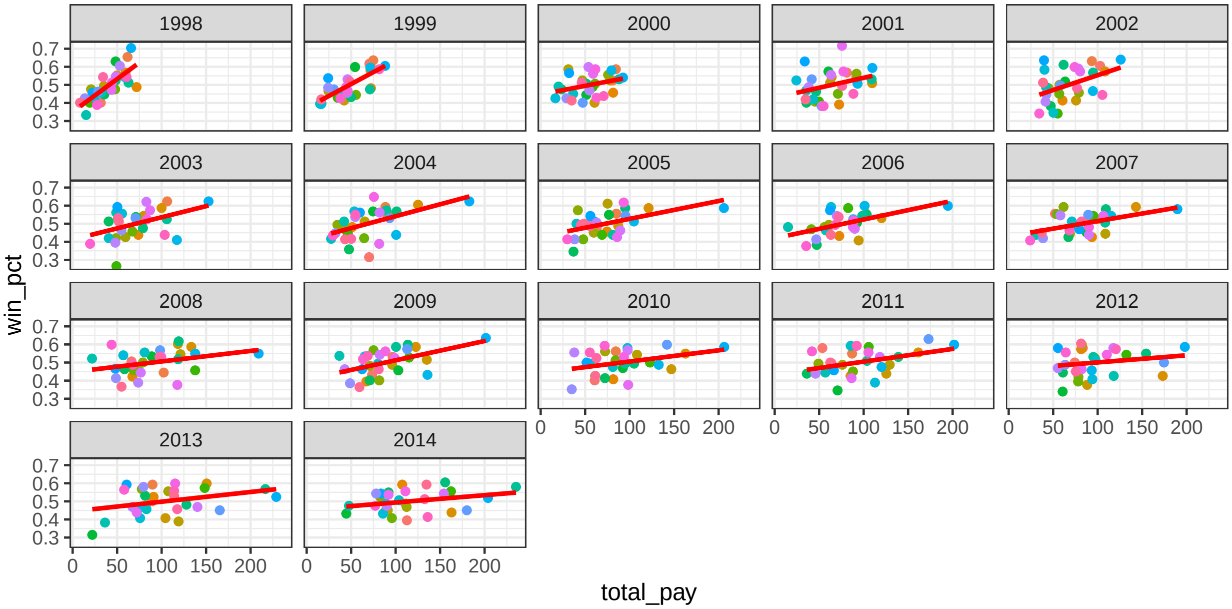
```
#Final_df %>% ggplot() + aes(x = diff_log, y = win_pct) + geom_point()
#ggplot(data = Final_df, x = "diff_log", y = "win_pct")
simple.fit = lm(win_pct ~ diff_log, data = Final_df)
summary(simple.fit)
```

```
##
## Call:
## lm(formula = win_pct ~ diff_log, data = Final_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22615 -0.05395  0.00185  0.05384  0.20756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.49685    0.00332   149.8  < 2e-16 ***
## diff_log     0.04656    0.01257     3.7  0.00024 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0703 on 478 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.0279, Adjusted R-squared:  0.0259
## F-statistic: 13.7 on 1 and 478 DF,  p-value: 0.000236
```

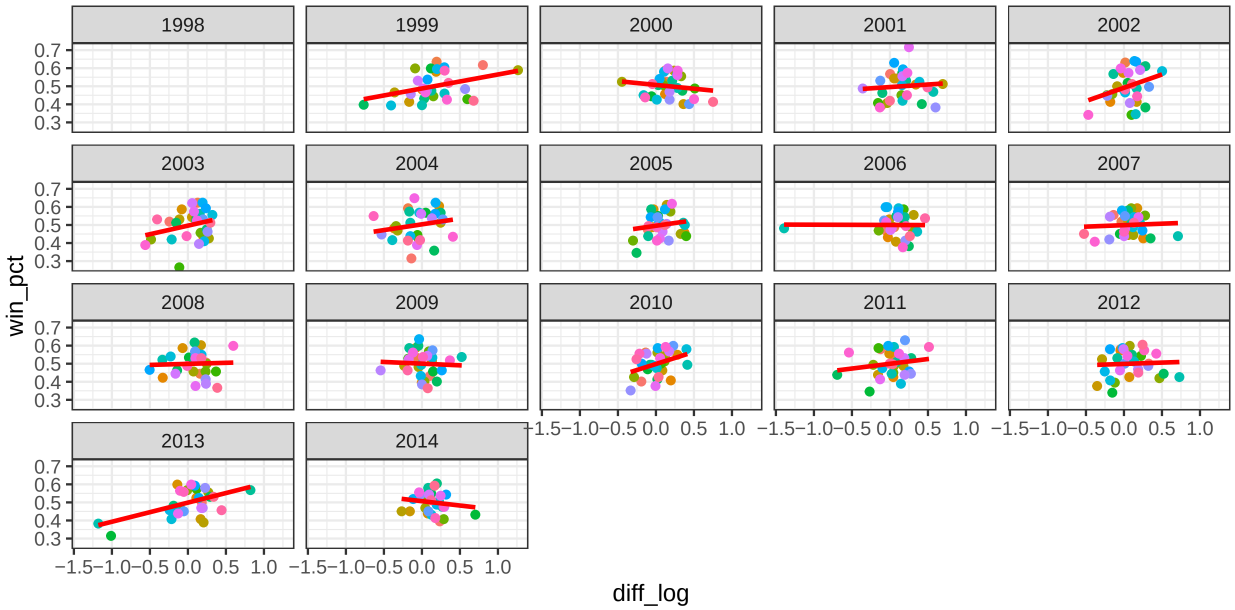
4.4 Comparison

Which set of factors are better explaining performance? Yearly payroll or yearly increase in payroll? What criterion is being used?

```
Final_df %>%
  ggplot(aes(x=total_pay, y=win_pct, group = year, color=Team)) +
  geom_point()+
  geom_smooth(method="lm", formula=y~x, se=F,color = "red")+
  facet_wrap(~year) +
  theme_bw() +
  theme(legend.position = 0)
```



```
Final_df %>%
  ggplot(aes(x=diff_log, y=win_pct, group = year, color=Team)) +
  geom_point()+
  geom_smooth(method="lm", formula=y~x, se=F,color = "red")+
  facet_wrap(~year) +
  theme_bw() +
  theme(legend.position = 0)
```



```
simple.fit = lm(win_pct ~ diff_log, data = Final_df)
summary(simple.fit)
```

```
##
## Call:
## lm(formula = win_pct ~ diff_log, data = Final_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22615 -0.05395  0.00185  0.05384  0.20756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.49685    0.00332   149.8  < 2e-16 ***
## diff_log     0.04656    0.01257     3.7  0.00024 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0703 on 478 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.0279, Adjusted R-squared:  0.0259
## F-statistic: 13.7 on 1 and 478 DF,  p-value: 0.000236
```

```
simple.fit = lm(win_pct ~ total_pay, data = Final_df)
summary(simple.fit)
```

```
##
## Call:
## lm(formula = win_pct ~ total_pay, data = Final_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21559 -0.05102  0.00245  0.05260  0.21767
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.448793   0.006855    65.5  < 2e-16 ***
## total_pay   0.000655   0.000079     8.3  9.6e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0675 on 508 degrees of freedom
## Multiple R-squared:  0.119,  Adjusted R-squared:  0.118
## F-statistic: 68.8 on 1 and 508 DF,  p-value: 9.64e-16
```

Based on the line graphs, there is strong evidenece that yearly pay roll is better at explaining performance, based on the positive correlations throughout the years. This conclusion is supported by the fact the R^2 is higher for total pay than diff log. Furthermore, the t-value of 8.3 and p-value for total_pay is more statistically significant. Meanwhile, there are some years where the regression fit for diff_log and win_pct are negative, for example, 2014.