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.

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.

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.)

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

datasurvey <- datasurvey %>% rename(age = Answer.Age, gender = Answer.Gender, education = Answer.Educat

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

datasurvey <- mutate(datasurvey, age = as.numeric(age))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
datasurvey <- filter(datasurvey, age >= 10 & age <= 110 & worktime >= 7 & !is.na(age))
datasurvey <- datasurvey[-which(datasurvey$income == ""), ]
datasurvey <- datasurvey[-which(datasurvey$gender == ""), ]


summary(datasurvey)
```

```
##       age          gender           education           income
##  Min.   :18.0   Length:1747        Length:1747        Length:1747
##  1st Qu.:23.0   Class :character   Class :character   Class :character
##  Median :28.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :30.4
##  3rd Qu.:34.0
##  Max.   :76.0
##     sirius            wharton            worktime
##  Length:1747        Length:1747        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
```

2.2 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.

2.3 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.

2.4 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.

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?

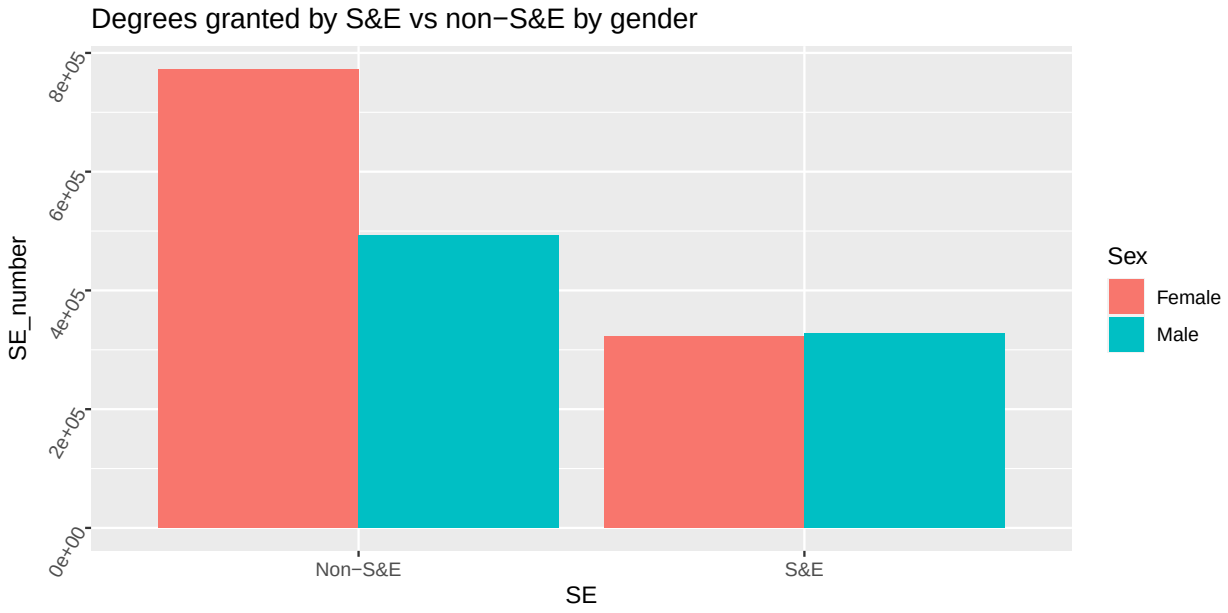
```
length(unique(womendata$Year))
```

```
## [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.
```

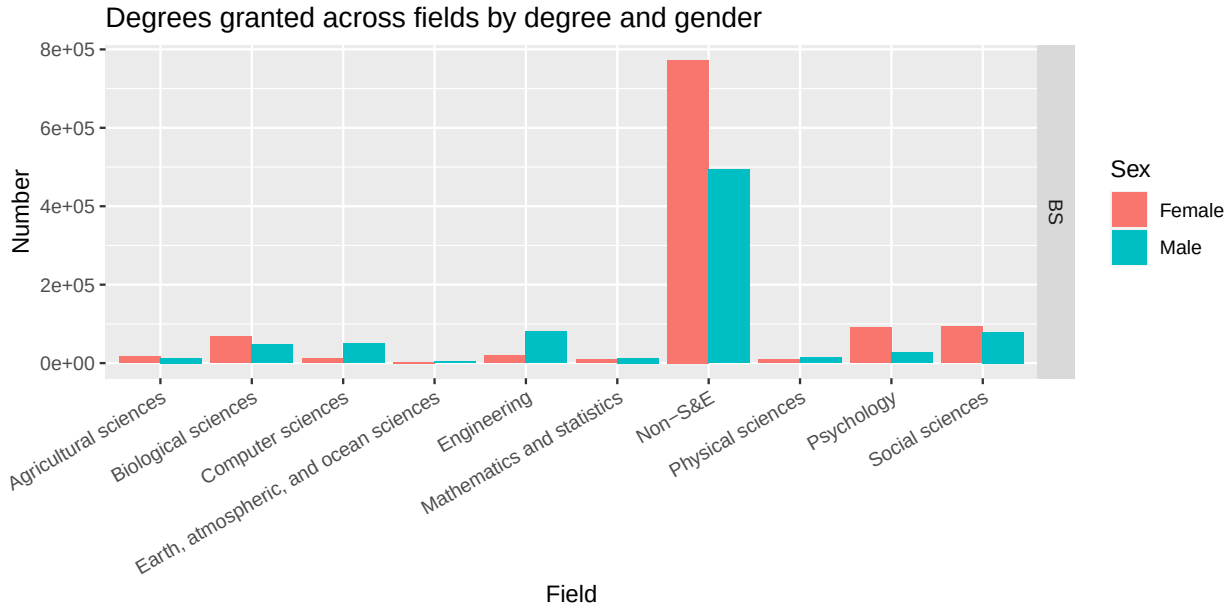
First, it is important to note that there are more females than males in this dataset (2015, BS).

```
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.

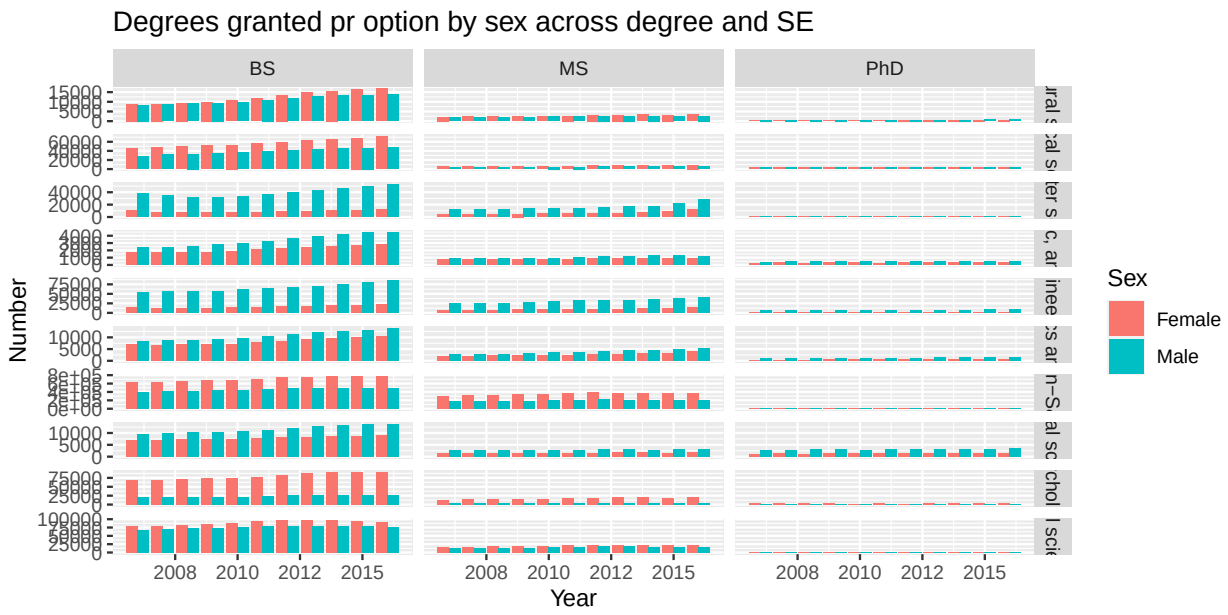
```
womendata %>%
  filter(Year == 2007) %>%
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Degree~., scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across fields by degree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

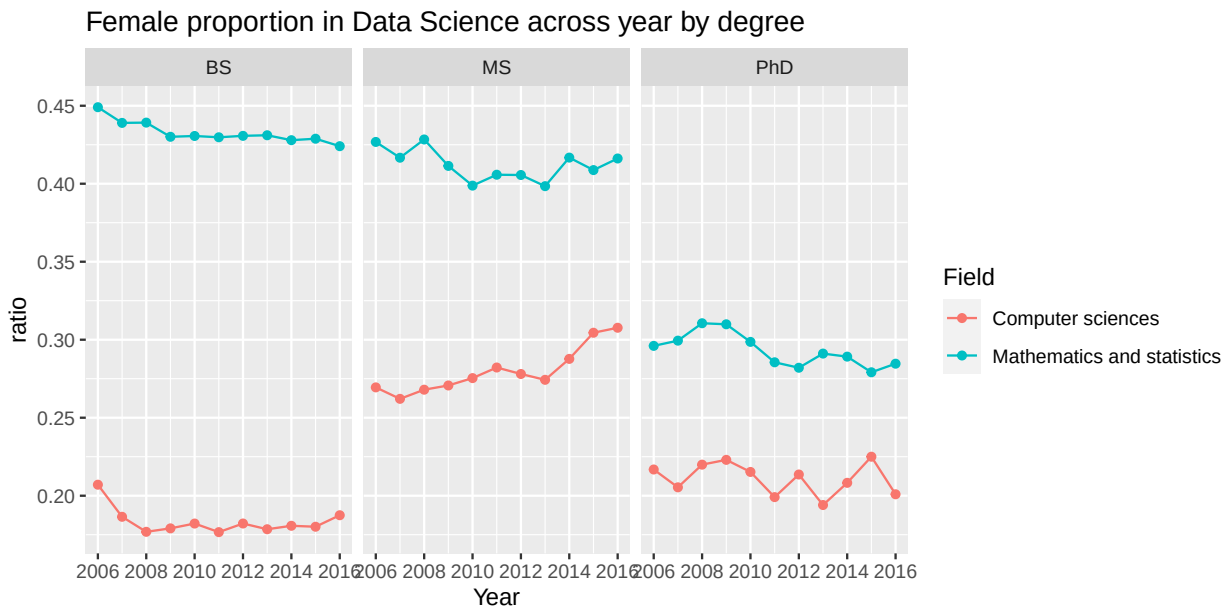
Here, we have graphed visually and numerically (by extracting the ggplot data) the number of degrees across gender (bar color), field, and time (x-axis). First, we discuss the changes over time. The number of degrees generally has a slight increase over time, which is more prominent at the BS level. 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.

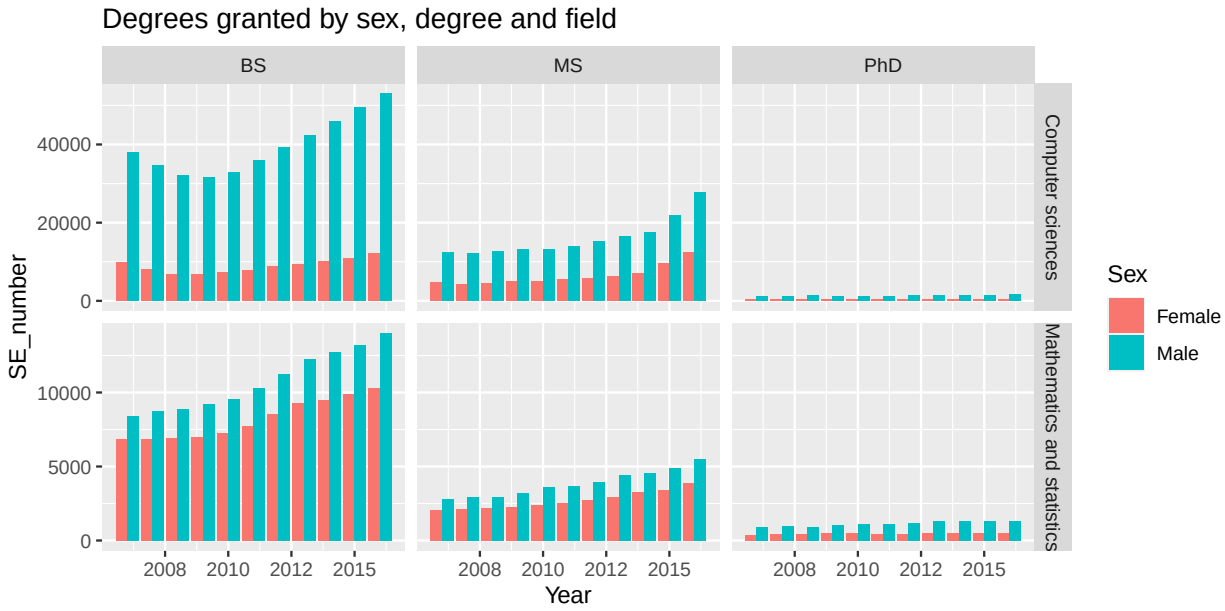
3.5 Women in Data Science

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)

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

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

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

4.2 Exploratory questions

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

ii. Between 2010 and 2014, inclusive, which team(s) "improved" the most? That is, had the biggest 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.

4.4 Comparison

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