Modern Data Mining, HW 1

Rob Kuan Bethany Hsaio Jose Cervantez

Due: 11:59PM, Feb 4th, 2024

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 HTML or pdf 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 Basic R Tutorial
- 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

1. 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".

Answer:

```
d0 <- read.csv("data/Survey_results_final.csv", header = T) |>
  select(Answer.Age, Answer.Gender, Answer.Education, Answer.HouseHoldIncome, Answer.Sirius.Radio, Answw
  rename(
    age = Answer.Age,
    gender = Answer.Gender,
    education = Answer.Education,
    income = Answer.HouseHoldIncome,
    sirius = Answer.Sirius.Radio,
    wharton = Answer.Wharton.Radio,
    worktime = WorkTimeInSeconds
  )

names(d0)
```

```
## [1] "age"       "gender"    "education" "income"    "sirius"    "wharton"
```

```
## [7] "worktime"
```

2. 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.

Code Cleaning:

```
## assign NA for empty strings
d0 <- d0 %>%
  mutate(across(where(is.character), ~ na_if(., "")))


## age
# table(d0$age)

d0 <- d0 |>
  mutate(age = case_when(
    age == "27`" ~ "27", # Change '27`' to '27'
    age == "Eighteen (18)" ~ "18", # Change 'Eighteen (18)' to '18'
    TRUE ~ age # Set all other values to age
  )) |>
  mutate(
    age = as.numeric(age), # Convert age to numeric
    age = case_when(
      age > 100 | age < 18 ~ NA_real_, # Replace age > 100 with NA
      TRUE ~ age # Keep all other ages as they are
    )
  )


## gender
# table(d0$gender) # good


## education
# table(d0$education)
d0 <- d0 |>
  mutate(education = case_when(
    education %in% c("Other", "select one") ~ NA_character_,
    TRUE ~ education
  ))

## income
# table(d0$income) # good


## sirius
# table(d0$sirius) # good

## wharton
# table(d0$wharton) # good
```

```
## worktime
# table(d0$worktime) # good
```

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.

Reasoning:

- For each variable in our dataset, I executed a 'table' command to visually expect the frequencies of the data. I also used personal judgment to assess whether the item was either missing, entered incorrectly, or unable to tell the difference. In the cases that I was able to quickly identify that the value was entered in an incorrect format, I adjusted the format correctly. For example, changing 'Eighteen (18)' to '18.' For items that were entered incorrectly but I couldn't determine a reasonable alternative or for missing data, I assumed NA for those values.

3. Brief summary

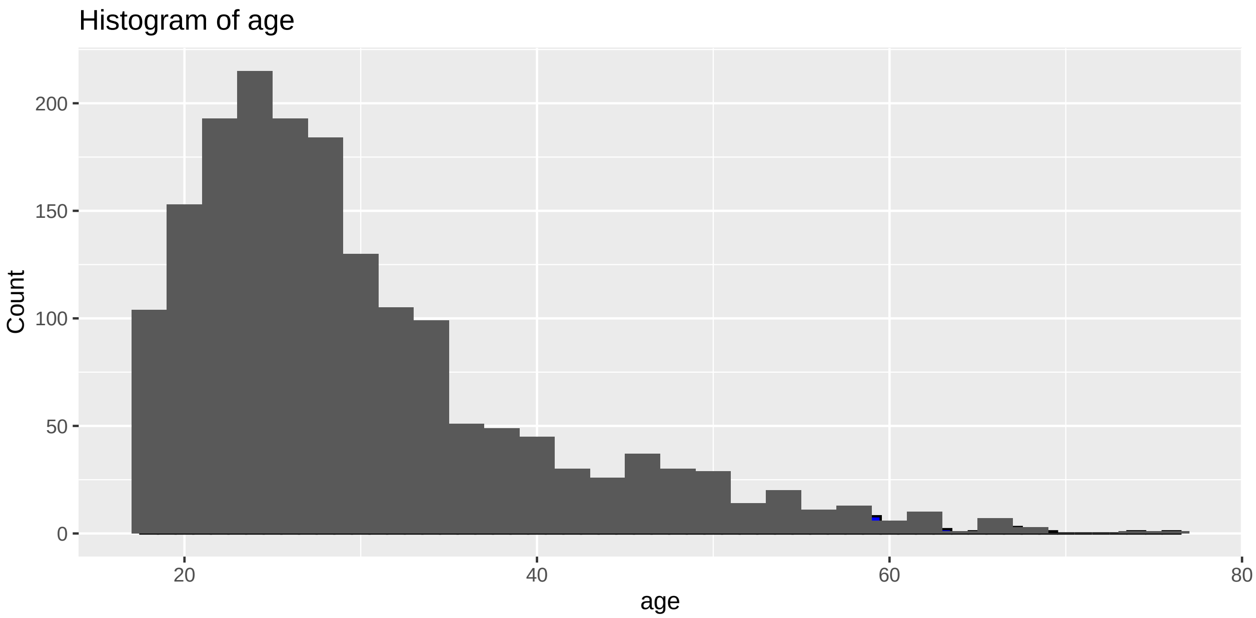
Write a brief report to summarize all the variables collected. Include both summary statistics (including sample size) and graphical displays such as histograms or bar charts where appropriate. Comment on what you have found from this sample. (For example - it's very interesting to think about why would 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.)

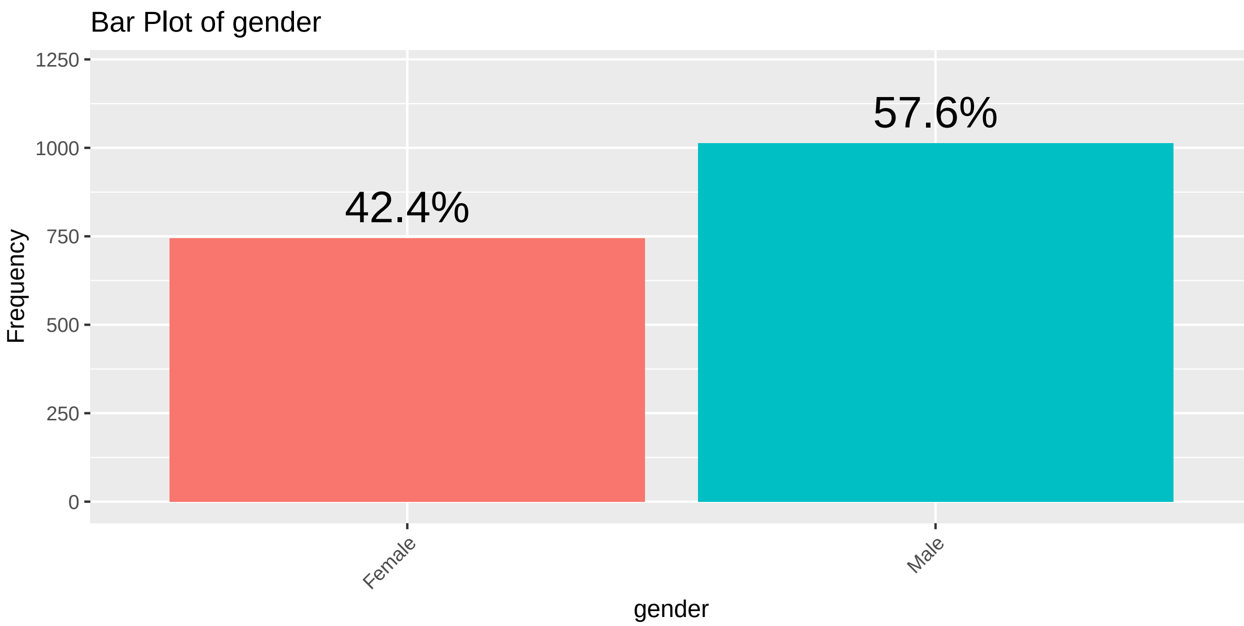
Code for Summary Stats:

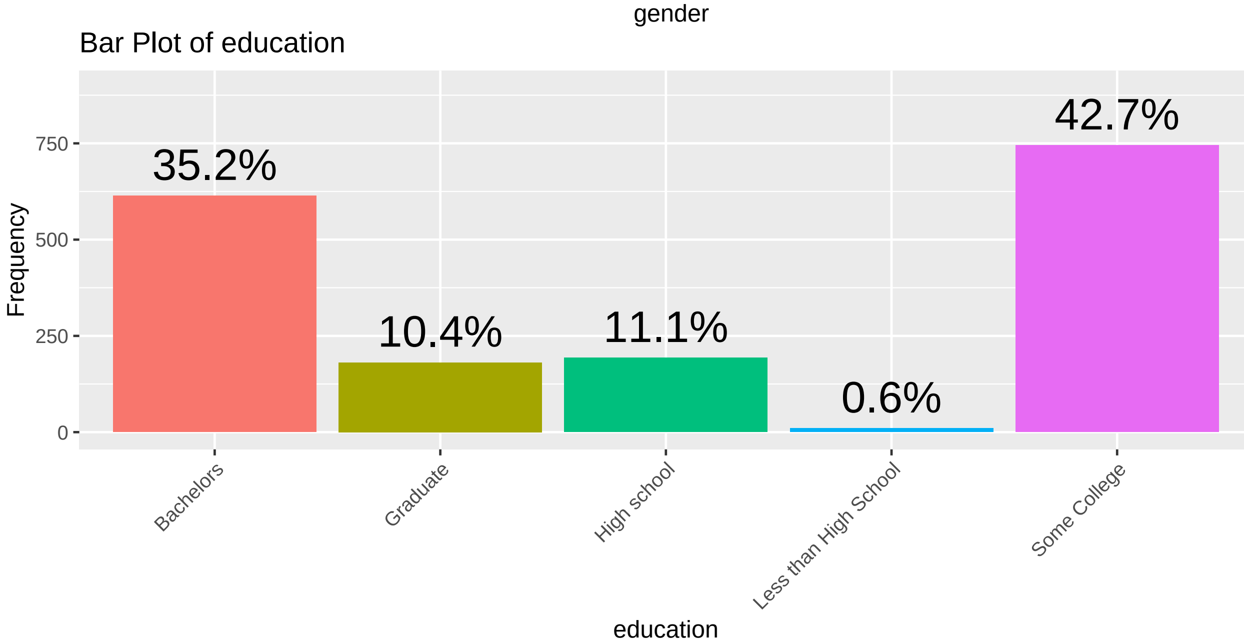
```
summarize_data <- function(df) {
  for (column_name in names(df)) {
    data <- df[[column_name]]

    cat("\n\nSummary for column:", column_name, "\n")

    if (is.numeric(data)) {
      # Numeric summary
      summary_stats <- summary(data)
      print(summary_stats)

      # Histogram using ggplot2
      p <- ggplot(df, aes(x = .data[[column_name]])) +
        geom_histogram(binwidth = 1, fill = "blue", color = "black") +
        labs(title = paste("Histogram of", column_name), x = column_name, y = "Count") +
        theme(legend.position = "none")

      # Handling NA values explicitly
      if (any(is.na(data))) {
        p <- p + geom_histogram(data = df[!is.na(df[[column_name]]), ], aes(x = .data[[column_name]]))
      }
    } else if (is.factor(data) || is.character(data)) {
      # Categorical summary
      cat_table <- data.frame(table(data))
      total <- sum(cat_table$Freq)
      cat_table$Perc <- (cat_table$Freq / total) * 100
      # Find the maximum frequency to adjust ylim
      max_freq <- max(cat_table$Freq)

      # Bar Plot using ggplot2
      p <- ggplot(cat_table, aes(x = data, y = Freq, fill = data)) +
```

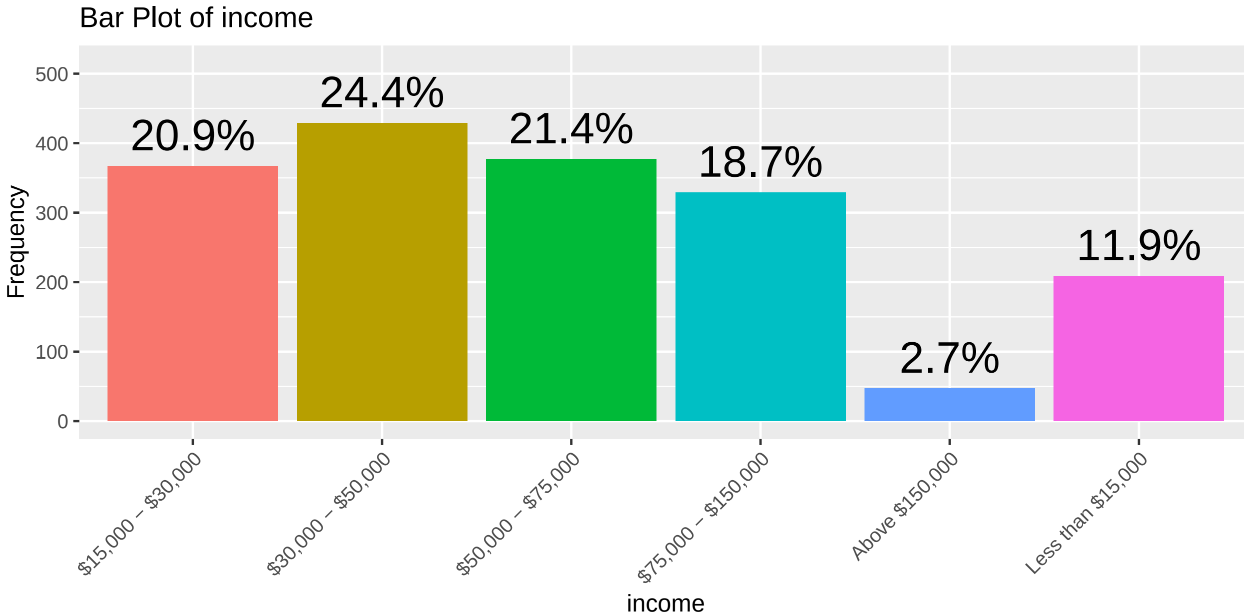
```
        geom_bar(stat = "identity") +
        geom_text(aes(label = sprintf("%.1f%%", Perc)), vjust = -0.5, size = 7, position = position_sta
        labs(title = paste("Bar Plot of", column_name), x = column_name, y = "Frequency") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        ylim(0, max_freq * 1.2) + # Increase ylim by 20%
        theme(legend.position = "none")
    }
    print(p)
  }
}
# Define the mapping from old to new names so it can fit in the graph
education_mapping <- c(
  "Bachelor’s degree or other 4-year degree" = "Bachelors",
  "Some college, no diploma; or Associate’s degree" = "Some College",
  "Graduate or professional degree" = "Graduate",
  "High school graduate (or equivalent)" = "High school",
  "Less than 12 years; no high school diploma" = "Less than High School"
)

# Replace the values in the dataframe
d0$education <- factor(education_mapping[d0$education])

summarize_data(d0)
```

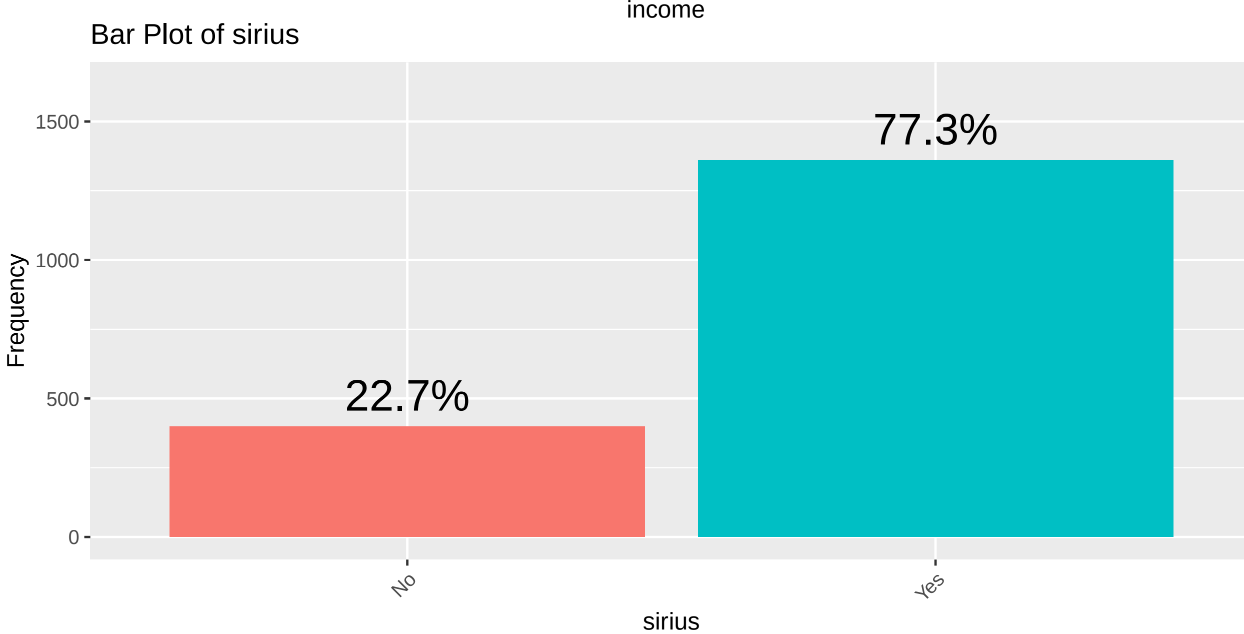
```
## Warning: Removed 4 rows containing non-finite values (`stat_bin()`).

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

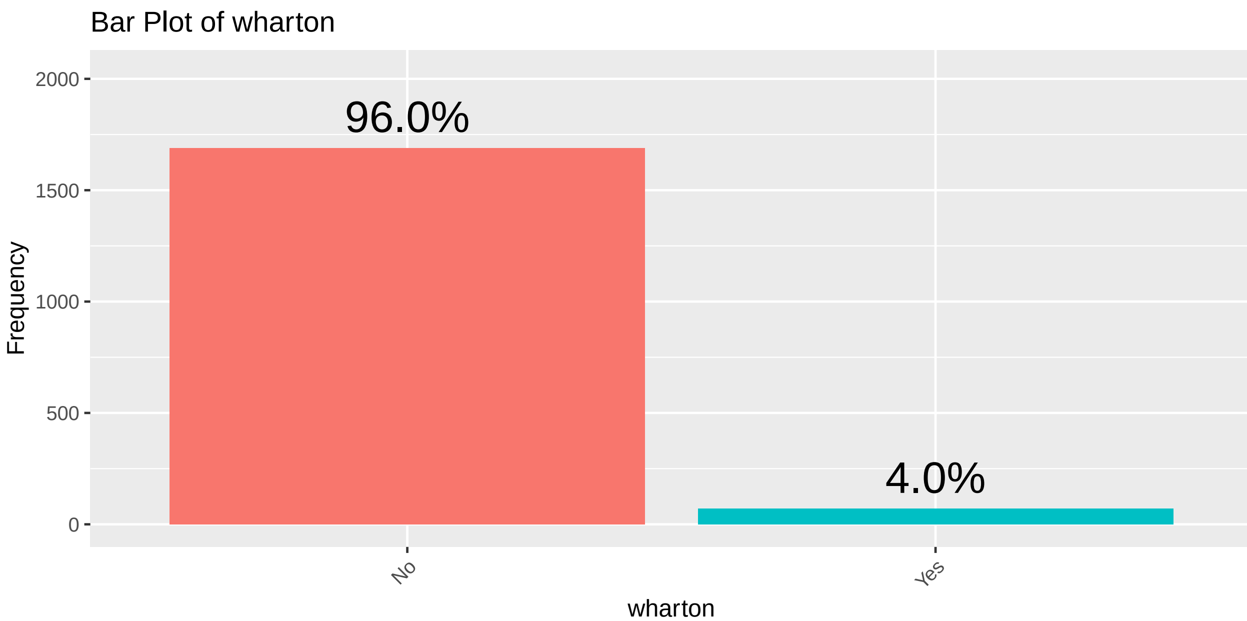


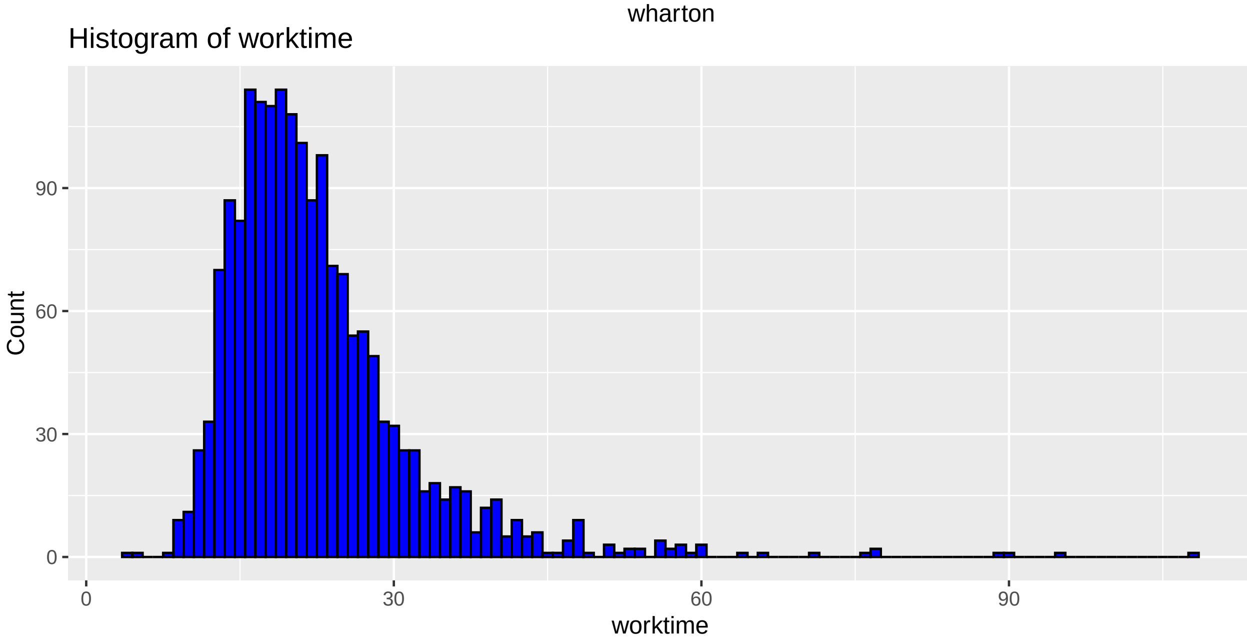

Bar Plot of gender
Frequency
1250
1000
750
500
250
0
42.4%
57.6%
Female
Male
gender


Bar Plot of education
Frequency
750
500
250
0
35.2%
10.4%
11.1%
0.6%
42.7%
Bachelors
Graduate
High school
Less than High School
Some College
education


Bar Plot of income
Frequency
500
400
300
200
100
0
20.9%
24.4%
21.4%
18.7%
2.7%
11.9%
$15,000 – $30,000
$30,000 – $50,000
$50,000 – $75,000
$75,000 – $150,000
Above $150,000
Less than $15,000
income


Bar Plot of sirius
Frequency
1500
1000
500
0
22.7%
77.3%
No
Yes
sirius





Summary Data Report:

- **Sample Size:** The survey collected responses from 1,764 participants.
- **Age Distribution:** The ages of participants range from 18 to 76 years, with a mean age of approximately 30.4 years.
- **Work Time:** The time taken to complete the survey ranged from 4 to 108 seconds, with an average time of approximately 22.5 seconds. This quick completion time reflects the survey's simplicity or possibly the respondents' familiarity with such tasks.
- **Gender Distribution:** The distribution between genders was relatively balanced, with men representing approximately 57.6% and women 42.4% of the sample.
- **Sirius Radio Listenership:** A significant majority, about 77.3%, reported having listened to Sirius Radio, indicating a broad engagement among the survey participants.
- **Wharton Radio Listenership:** Only a small fraction, about 4%, of respondents reported having

listened to the Wharton Radio on SiriusXM, suggesting that while SiriusXM's reach is extensive, the specific audience for the Wharton show is relatively small.

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.

1. Does this sample appear to be a random sample from the general population of the USA? Why it is crucial to have randomness here?

- The sample does not appear to be a random. Our current data says that 75% of the population is within 23-34 years of age. However, according to the Census, only 20% of the population are between the ages of 23-34.[1] Furthermore, the educational attainment is slightly skewed towards some college and bachelor's degree in our sample. According to the Census, it shows that approximately 15% of the population had completed some college, however, our sample shows that 42.7% of the sample hand completed some college. Additionally, according to the Census, it shows that the 23% of the population had completed a bachelor's degree, whereas our sample suggests that approximately 35.2% had completed it.[2]

2. Does this sample appear to be a random sample from the MTURK population?

- The sample does appear to be a random sample from the MTurk population. Namely, the sample skews younger than the general US population, which is consistent with both the MTurk population and our sample.[3] Additionally, our sample is more concentrated around the middle income categories (e.g., 20k - 75k), which is consistent with the MTurk sample, but lower income on average than the US population.[4]

2.3 Final estimate

Give a final estimate of the Wharton audience size by May of 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.

Goal of the Study:

The primary goal of this study was to estimate the audience size for the Wharton School's talk show on Sirius Radio, "Business Radio Powered by the Wharton School," as of May 2014. This estimation aimed to understand the show's reach among Sirius Radio listeners, leveraging a survey conducted via Amazon Mechanical Turk (MTurk).

Method Used:

Data were gathered through a survey launched on MTurk on May 24, 2014, targeting a maximum sample size of 2000 respondents over six days. The survey inquired about listenership to Sirius Radio and the Wharton Business Radio specifically, alongside demographic variables such as gender, age, and household income. The estimation method involved calculating the proportion of Wharton listeners to Sirius listeners within the MTurk sample and applying this proportion to the known total of 51.6 million Sirius Radio listeners to estimate the Wharton show's audience size.

Findings:

[1] https://www.census.gov/data/tables/2022/demo/age-and-sex/2022-age-sex-composition.html
[2] https://www.census.gov/newsroom/press-releases/2023/educational-attainment-data.html
[3] https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/
[4] https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/

After cleaning the dataset, it revealed that out of 1358 respondents who have listened to Sirius Radio, 68 reported listening to the Wharton Business Radio. This results in a proportion of approximately 4.98% of Sirius listeners also tuning into the Wharton show. By applying this proportion to the total Sirius audience, we estimate the Wharton Business Radio's audience size to be around 2,570,549 listeners as of May 2014.

Total Sirius Listeners: 1,365
Wharton Listeners Count: 68
Proportion Wharton: 0.0498
Estimated Wharton Audience: 2,570,549

No Yes 1690 70

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. (Can we use ChatGPT to get us a rough estimate?)

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.

Proposal:

To estimate the audience size of Wharton Business Radio Show (as of today). We will launch a study on Prolific Academic. According to the website, it costs approximately $8/hr for each participant. Our survey should take 3-minutes, which means that it will cost $0.40 per participant. We will recruit approximately 2,500 participants, for a total cost of $1,000. We will use the following questions to estimate the Wharton Radio Show's audience size:

1. Do you listen to podcasts? (Yes/No)
2. What types of podcasts do you listen to?
   - Business
   - News
   - Entertainment
   - Educational
   - Sports
   - Other
3. Please select the podcasts that you actively listen to:
   - Wharton Business Radio
   - The Joe Rogan Show
   - The Daily
   - Arm Chair Expert
   - Huberman Lab
   - Lex Fridman Podcast

- Other [text entry]

After we ask these questions, we will calculating the proportion of Wharton listeners to total podcast listeners within the Prolific sample and applying this proportion to the known estimate of total 100 million podcast listeners[5] to estimate the Wharton show's audience size. We use the total podcast estimate because the Wharton audience is influenced by various sources (e.g., Apple podcasts, Spotify, etc.), and not just a single source (e.g., SiriusXM.)

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.

a). Read the data into R.

b). Clean the names of each variables. (Change variable names to `Field`,`Degree`, `Sex`, `Year` and `Number`)

I inspected the variable names by looking at head(wsci). Only "Field and sex" and "Degrees Awarded" are not the correct names, so I rename those.

c). Set the variable natures properly. I inspected the variable natures using str(wsci). Field, Degree, and Sex are strings, which is appropriate for their data. Year and Number are numeric, which are also appropriate.

d). Any missing values? I checked for missing values by inspecting the summary of wsci, the table of values for each variable, and calculating the sum of all NA values. This reveals that there are no missing values. All columns have the same length, and there are no null values in any of the columns.

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

a). How many fields are there in this data? I found the number of fields by looking at the length of the list of unique values in Field. There are 10 unique fields.

b). What are the degree types? I got the degree types by looking at the unique values in Degree. They are BS, MS, PhD.

c). How many year's statistics are being reported here? I found this by looking at the length of the list of unique values in Year. There are 11 years of statistics in this dataset.

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.

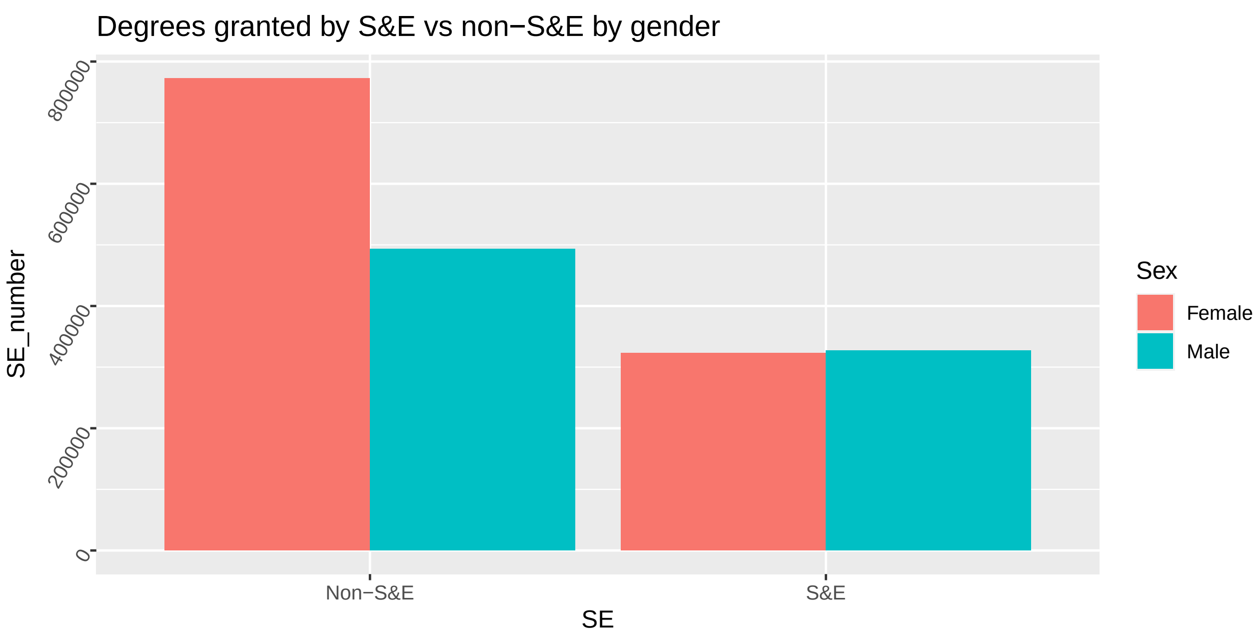
[5] https://explodingtopics.com/blog/podcast-listeners

To examine this, I filtered the data to include just data on BS degrees from 2015. Then, I looked at summary statistics of the number of degrees earned by men and women and plotted the number of S&E and non-S&E degrees earned by men and women.

There is not evidence that there are more males in science-related fields. In 2015, 327122 males earned S&E degrees and 493304 males earned non-S&E degrees, indicating that there are fewer males in science-related fields. Additionally, there is not evidence that there are more men than women obtaining S&E BS degrees. In 2015, 322935 women and 327122 earned S&E degrees, showing that roughly the same number of females as males obtained S&E degrees.

```
## `summarise()` has grouped output by 'SE'. You can override using the `.groups`
## argument.
```

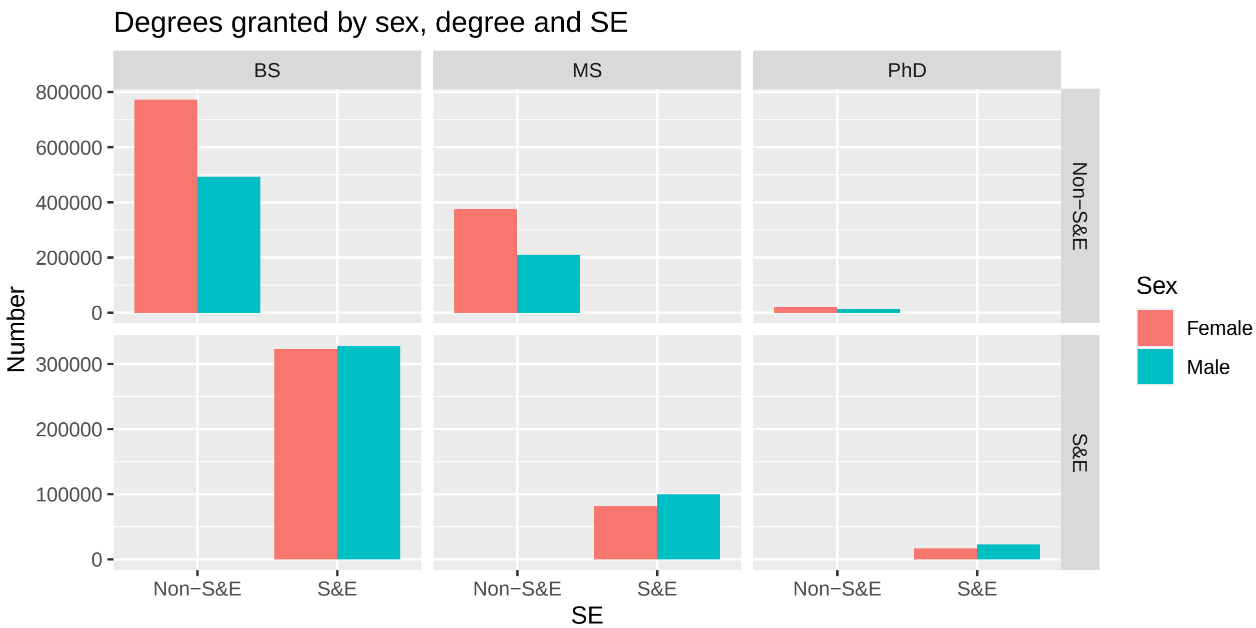
```
## `summarise()` has grouped output by 'SE'. You can override using the `.groups`
## argument.
```



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.

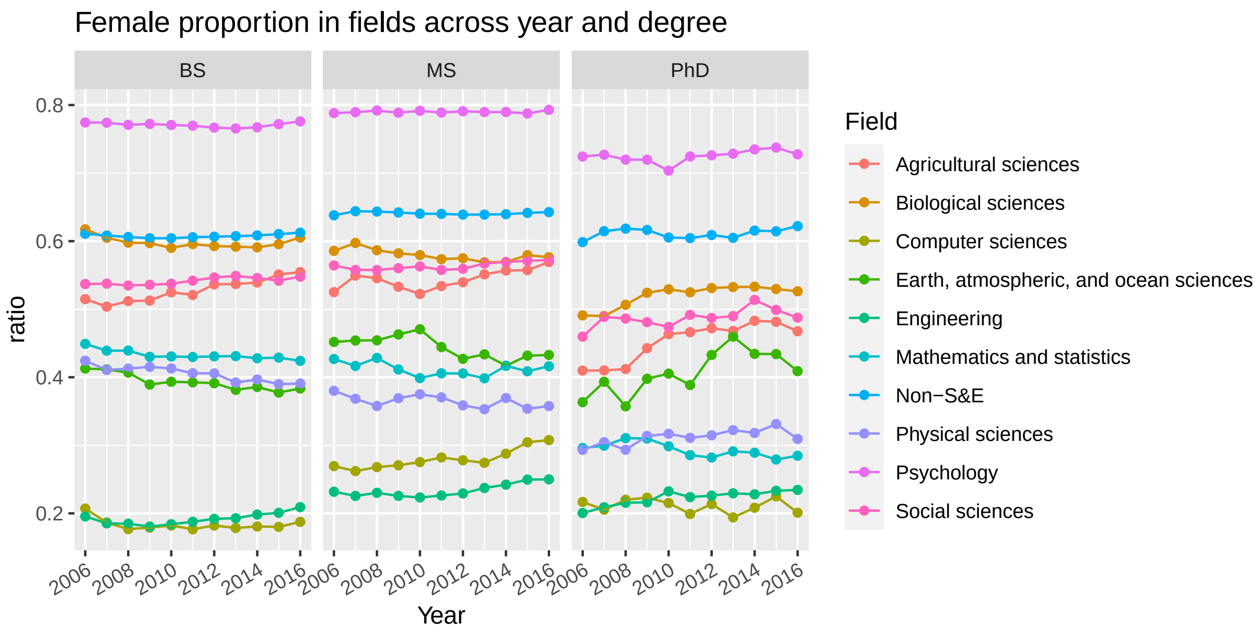
When we include the Degree in our analyses, we see that overall, more women than men pursue degrees at the BS, MS, and PhD levels. When we consider S&E vs non-S&E fields, more men than women earn S&E degrees at the MS and PhD levels. Once we include all fields, we notice larger gender effects for non-S&E, psychology, biological sciences, and engineering fields across all degrees. For non-S&E, psychology, and biological sciences degrees, more women than men pursue these. On the other hand, for engineering degrees, more men than women pursue these.

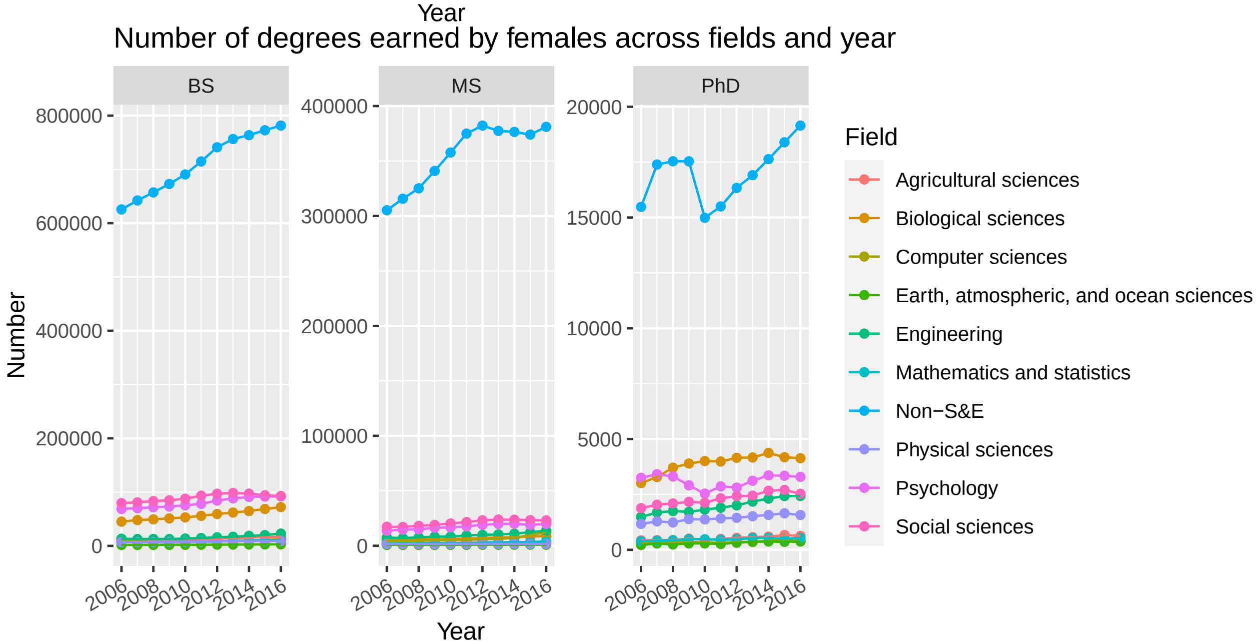


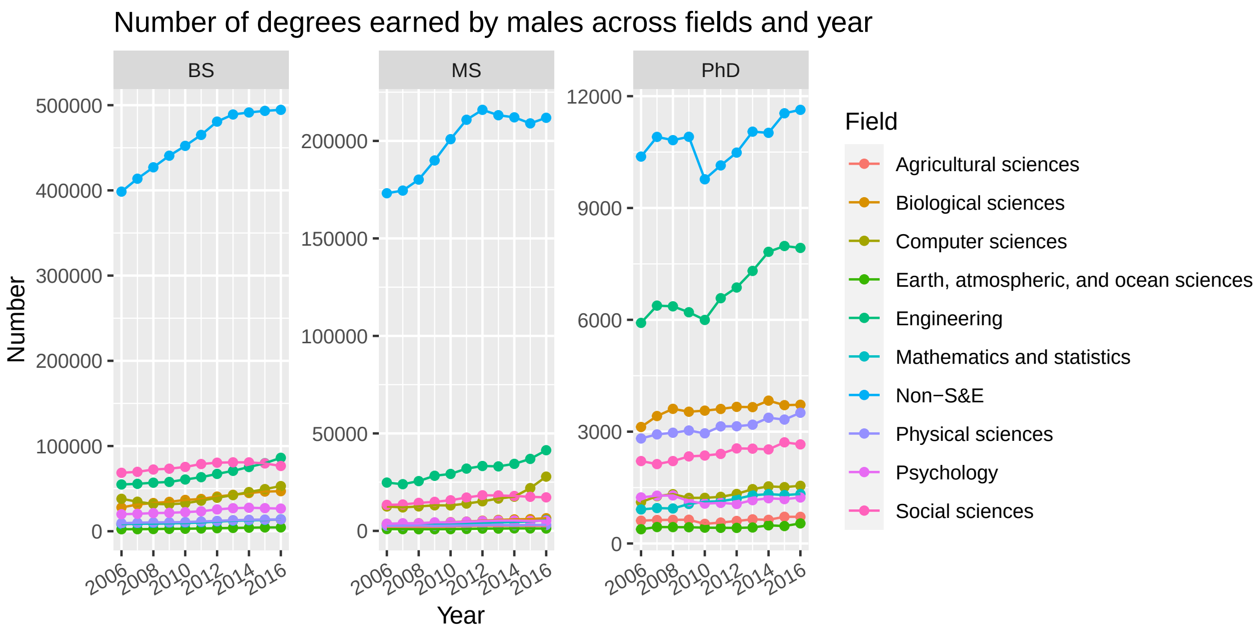
3.4 EDA bring all variables

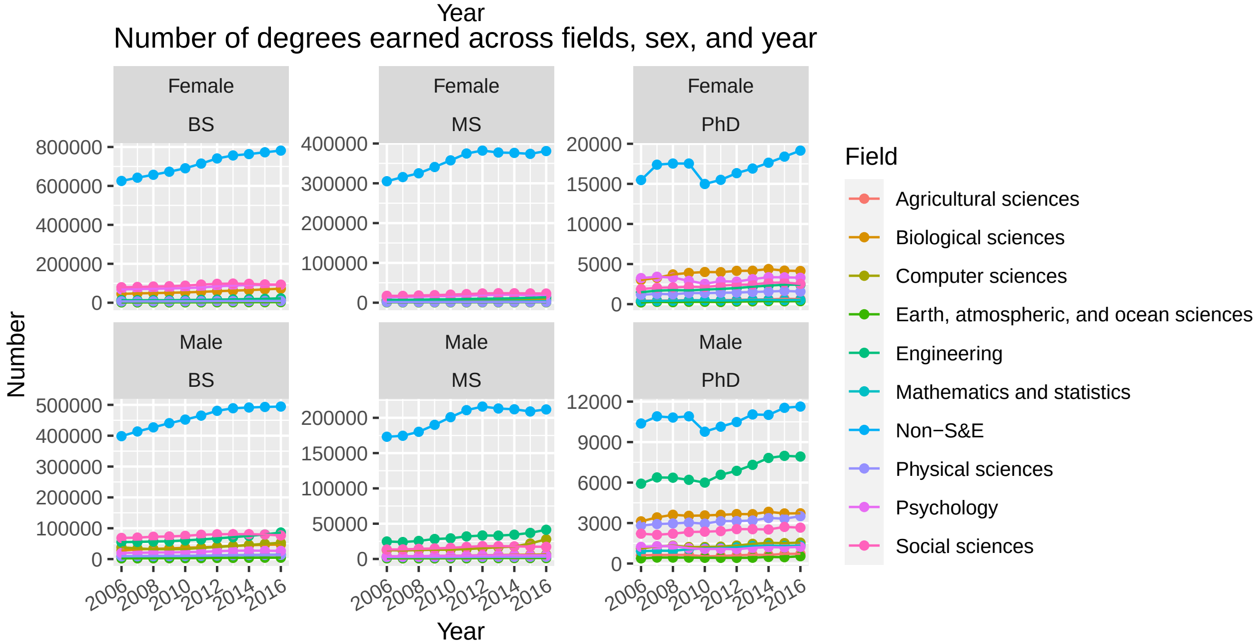
In this last portion of the EDA, we ask you to provide evidence numerically and graphically: Do the number of degrees change by gender, field, and time?

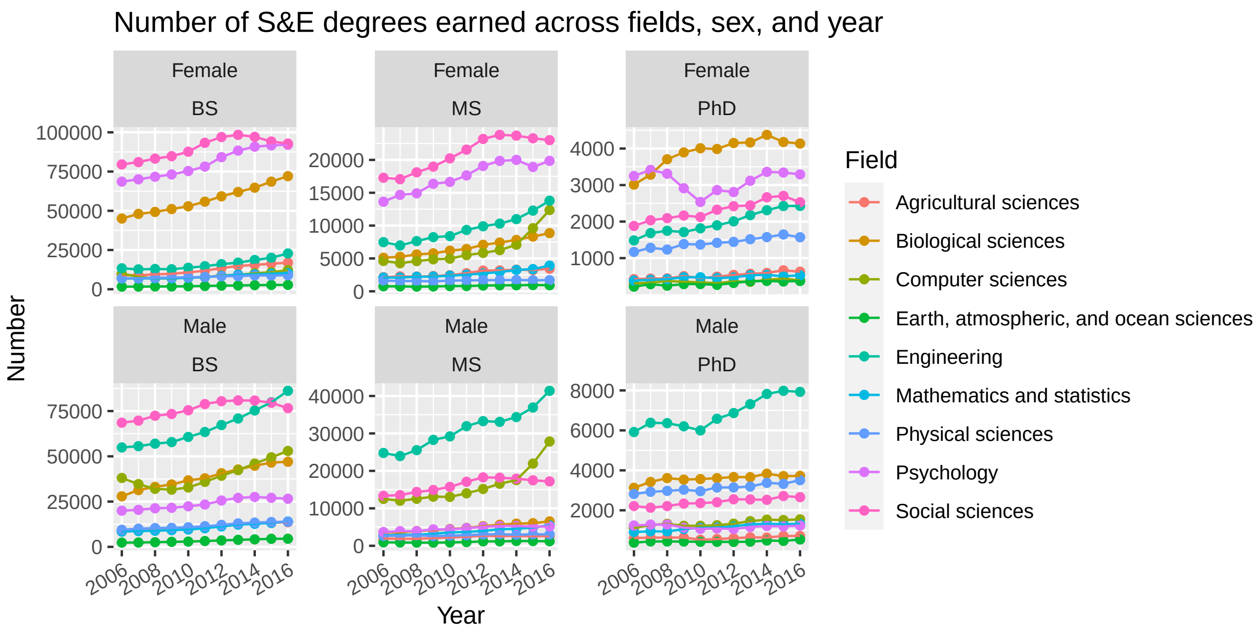
When we display the proportion of females in each degree over time, we see that the relative proportion of females remains fairly stable over time for most fields in all degrees. We can also examine the number of degrees earned by gender. When we only consider females, we see overall increases in the number of degrees earned across all fields and degree types. We see the same trend when we only consider males. Comparing females with males, more women than men pursue degrees across all degree types. For both men and women, non S&E degrees are the most pursued degree. Excluding these from our analysis, we see that consistently more men than women pursue engineering degrees across all degree types.


Female proportion in fields across year and degree
BS
MS
PhD
ratio
0.8
0.6
0.4
0.2
2006
2008
2010
2012
2014
2016
Year
Field
Agricultural sciences
Biological sciences
Computer sciences
Earth, atmospheric, and ocean sciences
Engineering
Mathematics and statistics
Non−S&E
Physical sciences
Psychology
Social sciences


Number of degrees earned by females across fields and year
BS
MS
PhD
Number
800000
600000
400000
200000
0
400000
300000
200000
100000
20000
15000
10000
5000
2006
2008
2010
2012
2014
2016
Year
Field
Agricultural sciences
Biological sciences
Computer sciences
Earth, atmospheric, and ocean sciences
Engineering
Mathematics and statistics
Non−S&E
Physical sciences
Psychology
Social sciences


Number of degrees earned by males across fields and year
BS
MS
PhD
Number
Year
2006
2008
2010
2012
2014
2016
Field
Agricultural sciences
Biological sciences
Computer sciences
Earth, atmospheric, and ocean sciences
Engineering
Mathematics and statistics
Non−S&E
Physical sciences
Psychology
Social sciences

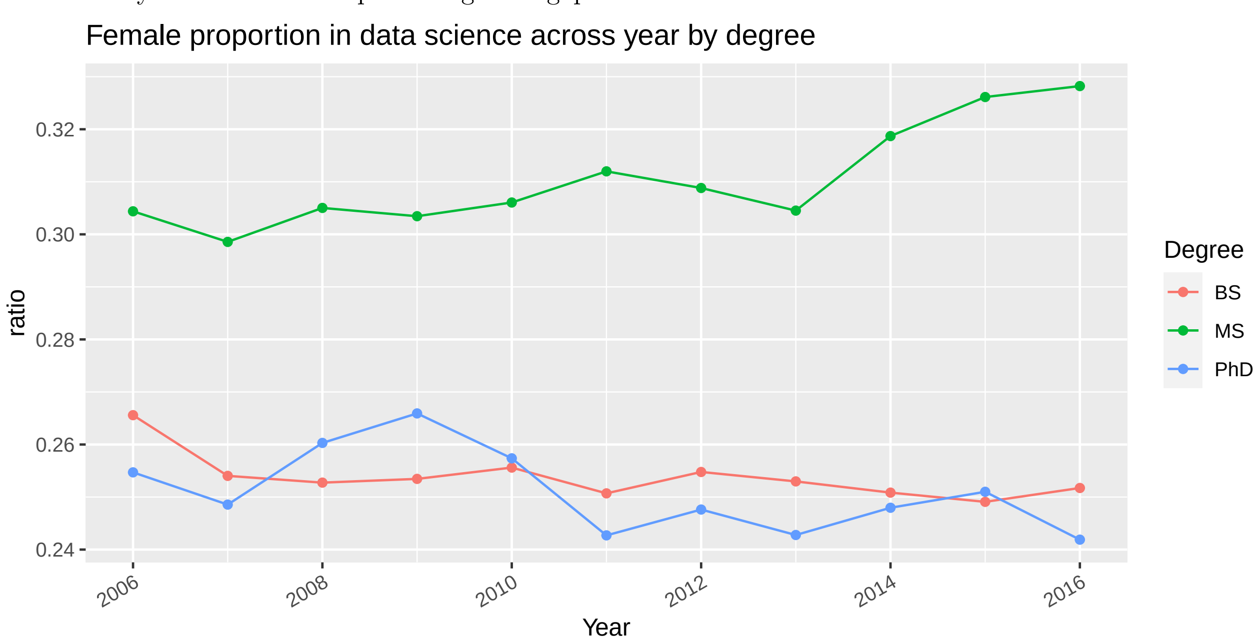

Number of degrees earned across fields, sex, and year
Female
BS
Female
MS
Female
PhD
Male
BS
Male
MS
Male
PhD
Number
Year
Field
Agricultural sciences
Biological sciences
Computer sciences
Earth, atmospheric, and ocean sciences
Engineering
Mathematics and statistics
Non−S&E
Physical sciences
Psychology
Social sciences

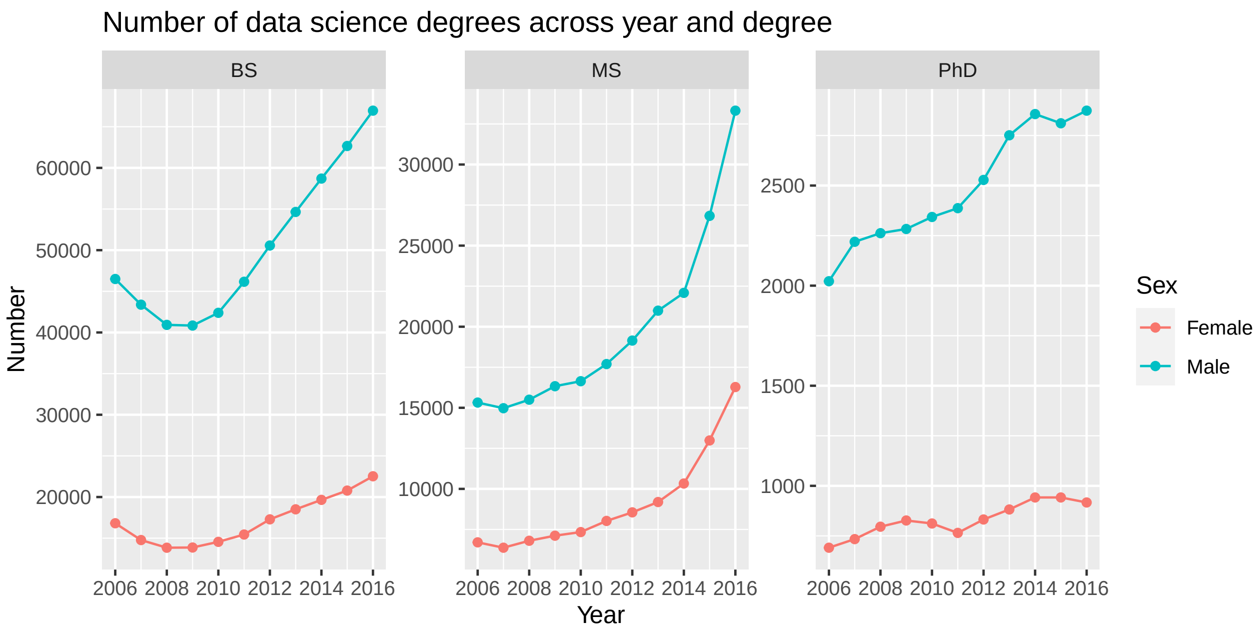


3.5 Women in Data Science

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.

There is evidence that women are underrepresented in data science. For all 3 degrees, less than 1/2 of data science degree pursuers are women. The ratio of women seeking data science BS and PhD degrees has decreased over time, showing that the gender gap has widened. Interestingly, the ratio of women seeking data science MS has increased over time. However, the ratio of female MS degree holders remains around 1/3. When we look at the number of women and men pursuing data science degrees, we see that there has been an increase for both men and women over time. However, for BS, MS, and PhD degrees, the rate of increase for men is higher than that for women. All of this provides supporting evidence that women are underrepresented in data science. With the rate of increase for men outpacing that of women, this suggests that we may not be able to expect the gender gap in data science to decrease in the near future.





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?

In summary, our data analysis shows that in general, more males pursue science-related fields. That is, across time and all three degree types, more men than women pursue science-related degrees. This is especially true for the fields related to engineering, data science, and "harder" sciences. However, when we consider "softer" sciences, such as psychology and social science, mroe women than men pursue these degrees. One exception to this rule is biological sciences, in which more women than men pursue these degrees.

One concern with this dataset is that we can only see the number of degrees earned by women and men across years and fields. It does not give us any information, such as in-major GPAs, of how well women and men did in their degrees. Without this information, there is no way to measure how successful women and men are at completing their degrees, which could also impact their future career outcomes. Additionally, without career outcome information, we cannot conclusively state that science-related fields are male dominated in the workforce. Another concern is that this dataset only provides information up until 2016, which bars us from analyzing the gender difference in science-related fields in more recent years.

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

a). 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.

Answer: In this set up, the log difference is more appropriate because it allows us to get a better understanding of the percentage change of payroll between each year. Since players have different magnitudes in pay, taking the log allows us to base our analysis on the percentage change.

---

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

```
## Rows: 510 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): team
## dbl (4): year, payroll, win_num, win_pct
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Answer: Look to .rmd document for the code.

---

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

Answer: Look to .rmd document for the code.

4.2 Exploratory questions

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

team	diff_log_2010_2014	percent_change_2010_2014
Los Angeles Dodgers	0.908	148
Texas Rangers	0.901	146
San Diego Padres	0.869	138
Pittsburgh Pirates	0.804	124
Washington Nationals	0.785	119

Answer: The top five teams are the LA Dogdgers, the Texas Rangers, the San Diego Padres, the Pittsburg Pirates, and the Washington Nationals. (see table above)

---

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

Table 2: Table continues below

team	diff_log_win_pct_2010_2014	percent_change_win_pct_2010_2014
Pittsburgh Pirates	0.434	54.4
Baltimore Orioles	0.375	45.5
Seattle Mariners	0.355	42.6
Washington Nationals	0.33	39.1
Kansas City Royals	0.284	32.8

Answer: The Pittsburgh Pirates improved the most (see table above).

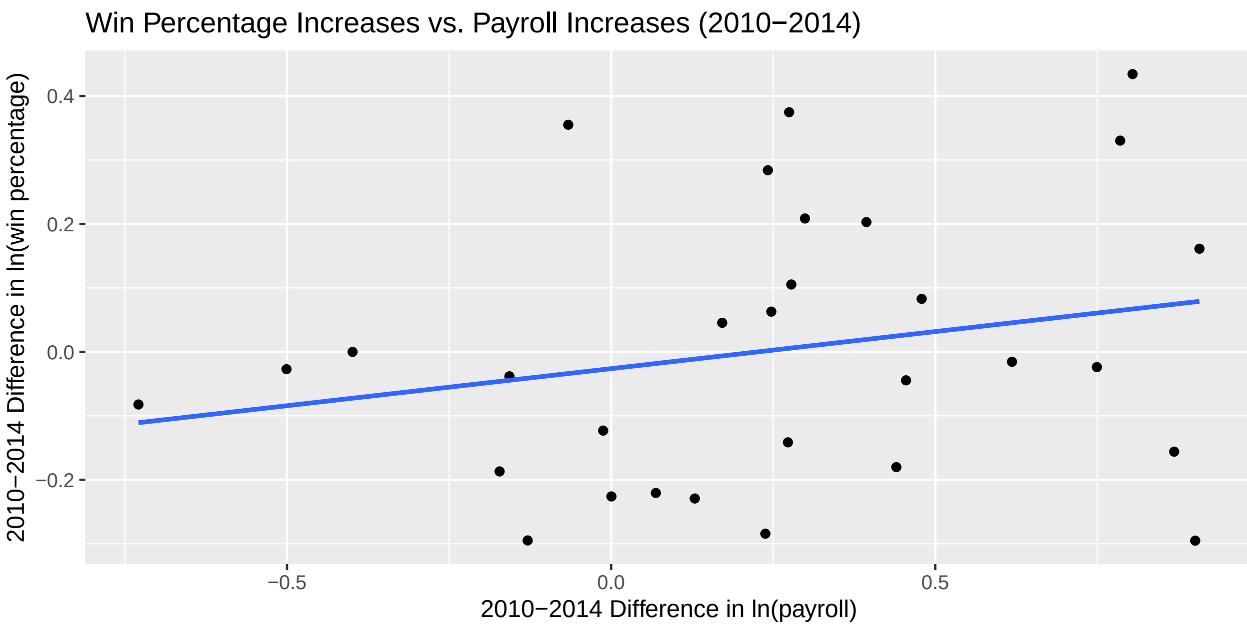
4.3 Do log increases in payroll imply better performance?

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.

Table 4: Regression of difference in ln(win percentage) from 2010-2014 on difference in ln(payroll) from 2010-2014

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0262	0.0452	-0.58	0.567
diff_log_2010_2014	0.116	0.0939	1.23	0.227



Answer: The table of statistics above and the corresponding graph provides some evidence that higher % changes in payroll on the log scale correlate with higher % changes in win percentage. However, it's important to note that even though higher % changes in payroll **predict** higher % changes in win percentage, it doesn't necessarily mean that higher % changes in payroll **cause** higher % changes in win percentage. This is because there may be ommited variables that are correlated with both payroll and win percentage – so we cannot conclusively say that one "leads" to the other.

If we regress % change in win percentage on % change in payroll, we find that although the results directionally affirm our hypothesis, they are not statistically significant. In fact, the r-square for this model is extremely low, which makes it plausible that there there is omitted variables which would bias this estimate. This implies the results are only suggestive, and not conclusive.

4.4 Comparison

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

Answer: This question isn't very clear as its not clear how "performance" is being defined (e.g., win percentage, percentage point improvement in win percentage, percentage increase in win percentage, etc.).

If we assume that "performance" is being defined as win percentage, then we can compare the two models to see which variable better explains the variance in win percentage using the the R-Squared metric. Below, you can find two regressions: one with win percentage regressed on payroll, and another with win percentage regressed on the log difference in payroll. The R-Squared for the first model is 0.12, while the R-Squared for the second model is 0.03. This suggests that the yearly payroll is significantly better at explaining the variance in win percentage than the yearly increase in payroll.

Table 5: Regression of win percentage on yearly increase in payroll

term	estimate	std.error	statistic	p.value
(Intercept)	0.497	0.00332	150	0
diff_log	0.0466	0.0126	3.7	0.000236

Table 6: Regression of win percentage on yearly payroll

term	estimate	std.error	statistic	p.value
(Intercept)	0.449	0.00686	65.5	9.24e-250
payroll	0.000655	7.9e-05	8.3	9.64e-16

We can also look at another criterion to compare models against one another: the Bayesian Information Criterion (BIC). The BIC is a criterion for model selection among a finite set of models, and can provide a rough approximation to the logarithm of the Baye's factor (. The difference in BIC between the two models is 115, which is an incredibly large Bayes factor (approximately equal to $e^{(115/2)}$). This shows convincingly that payroll is better at explaining win percentage than yearly increase in payroll.

Table 7: BIC comparison for both models

payroll_increase_BIC	payroll_BIC	BIC_delta
-1171	-1286	115