Modern Data Mining, HW 1

Alyssa Frantz Angela Chen Lori Sun

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

Contents

1 Overview **2**

1.1 Objectives . . . 2

1.2 Instructions . . . 2

1.3 Review materials . . . 3

2 Case study 1: Audience Size **4**

2.1 Data preparation . . . 4

2.2 Sample properties . . . 10

2.3 Final estimate . . . 12

2.4 New task . . . 13

3 Case study 2: Women in Science **14**

3.1 Data preparation . . . 14

3.2 BS degrees in 2015 . . . 14

3.3 EDA bringing type of degree, field and gender in 2015 . . . 15

3.4 EDA bring all variables . . . 17

3.5 Women in Data Science . . . 19

3.6 Final brief report . . . 20

4 Case study 3: Major League Baseball **21**

4.1 EDA: Relationship between payroll changes and performance . . . 21

4.2 Exploratory questions . . . 21

4.3 Do log increases in payroll imply better performance? . . . 22

4.4 Comparison . . . 24

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".

In this step, we cleaned the data to include only the variables specficied and renamed them accordingly.

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.

We identified and addressed the incomplete and missing values in this dataset as described below.

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.)

We examined 7 of the 35 variables in this dataset: `age`, `gender`, `income`, `education`, `sirius` (whether they listened to Sirius radio), `wharton` (whether they listened to the Wharton Show), and `worktime`. There were initially 1764 observations collected. However, we cleaned the data by omitting observations that had missing or inappropriate data for any of the 7 variables we examined (for example, entries that were left as "select one"). We also converted age into a numerical variable and the remaining 6 categorical variables into factor variables. As a result, we were left with 1725 observations. A summary of the results for each variable can be found below.

Age: Overall

- Mean = 30.3
- Standard deviation = 9.87
- Min = 18
- Max = 76

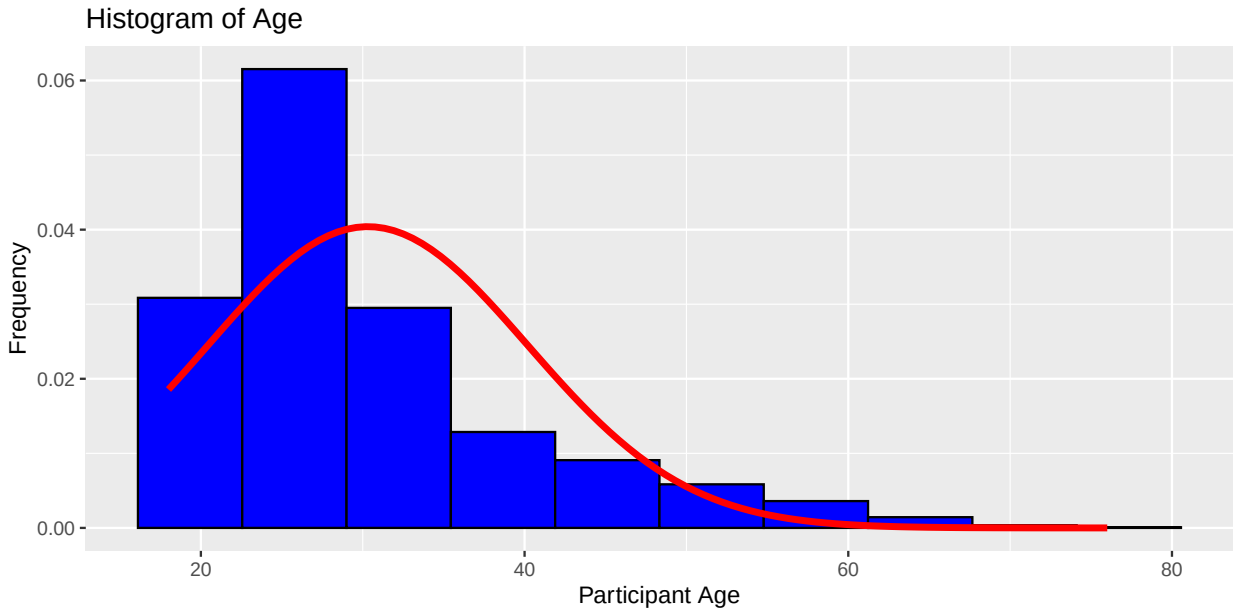
Age: Those who listen to the Wharton show

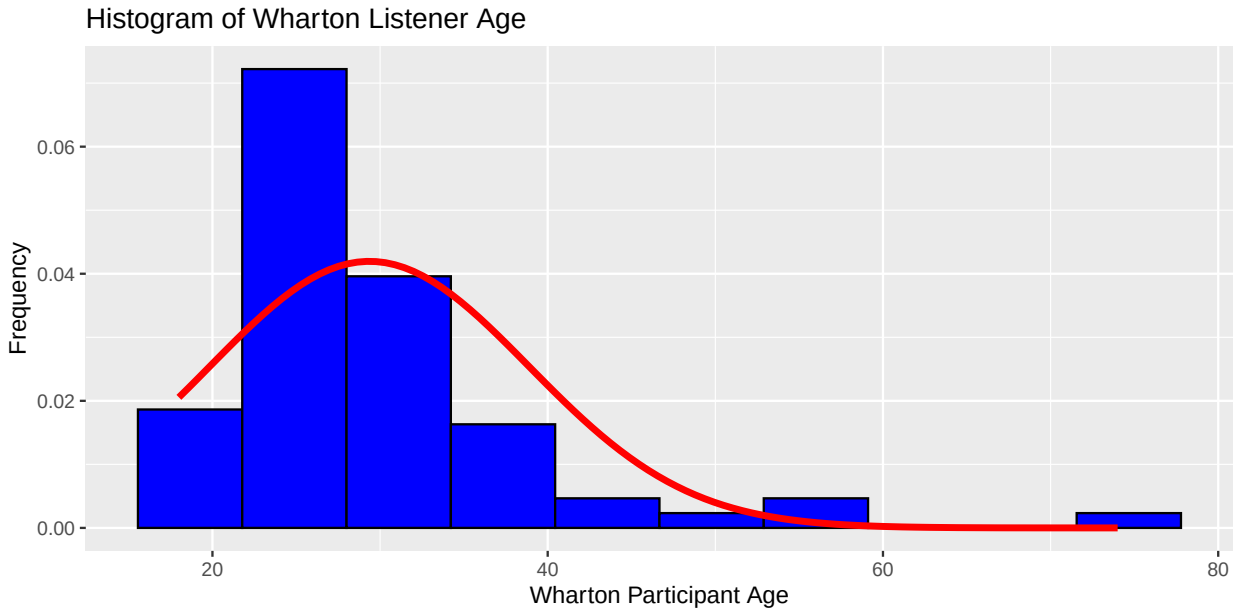
- Mean = 29.4
- Standard deviation = 9.51
- Min = 18
- Max = 74

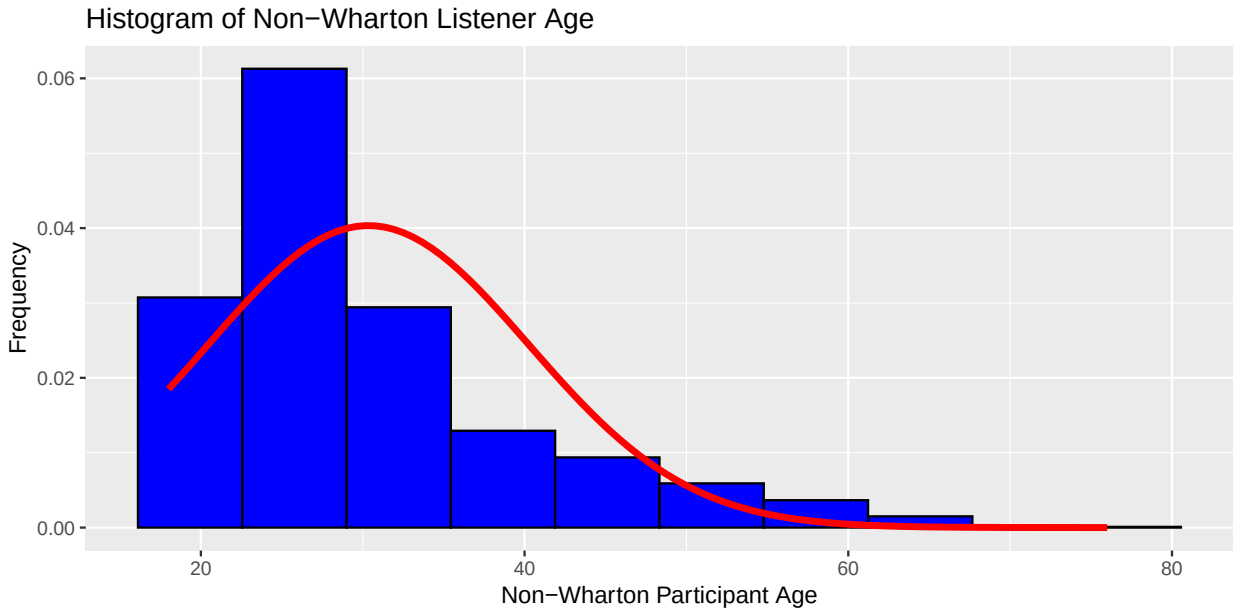
Age: Those who don't listen to the Wharton show

- Mean = 30.4
- Standard deviation = 9.92
- Min = 18
- Max = 76

In the first histogram below, the distribution of age for all survey respondents can be seen. The following 2 histograms display the age distributions for respondents who listen to the Wharton show and those who don't, respectively. Looking at the these two histograms demonstrates that there isn't a significant difference in age between these two populations.







Education: Overall

- Bachelor's degree or other 4-year degree: 611
- Graduate or professional degree: 177
- High school graduate (or equivalent): 189
- Less than 12 years; no high school diploma: 10
- Other: 2
- Some college, no diploma; or Associate's degree: 736

Education: Those who listen to the Wharton show

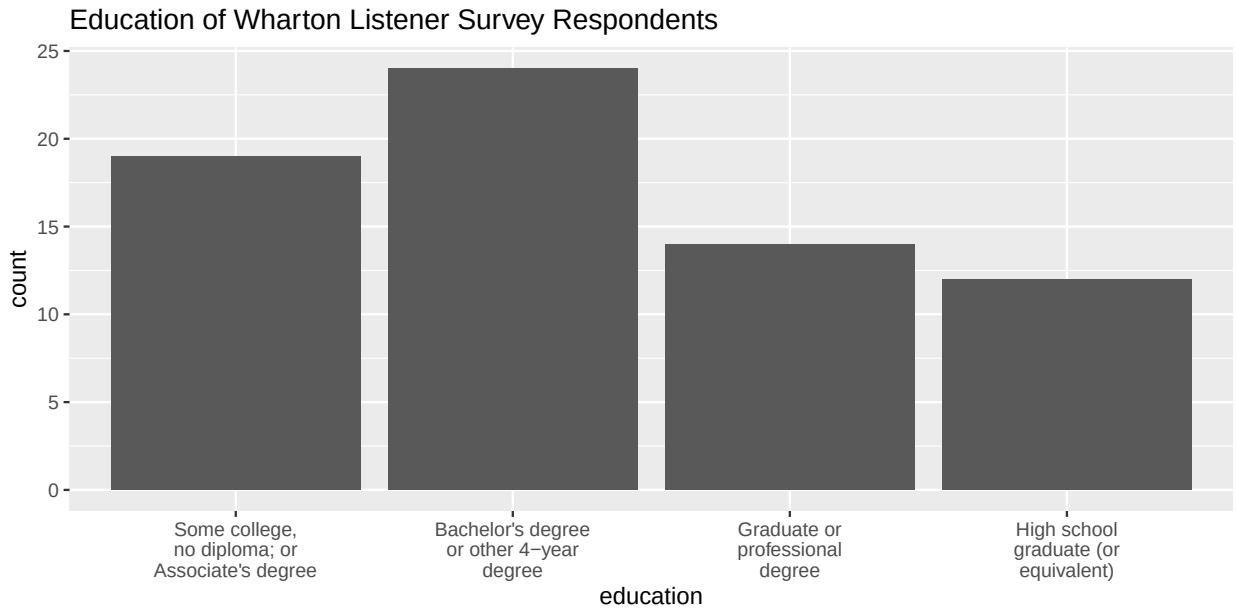
- Bachelor's degree or other 4-year degree: 24 (35%)
- Graduate or professional degree: 14 (20%)
- High school graduate (or equivalent): 12 (17%)

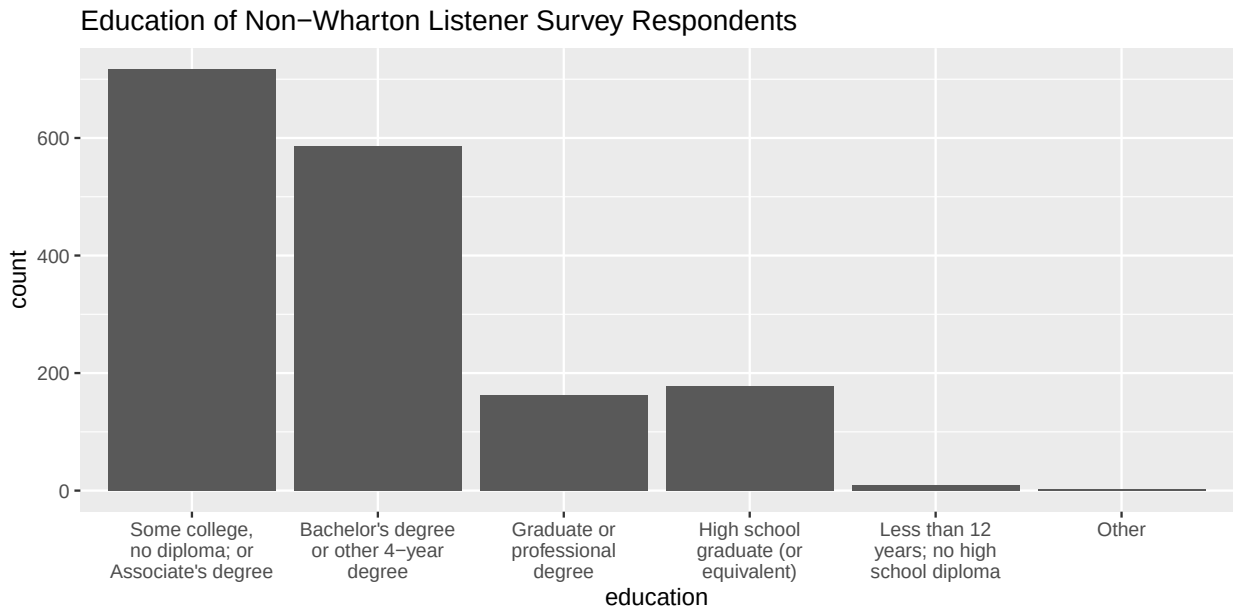
- Less than 12 years; no high school diploma: 0
- Other: 0
- Some college, no diploma; or Associate's degree: 19 (28%)

Education: Those who don't listen to the Wharton show

- Bachelor's degree or other 4-year degree: 586 (35%)
- Graduate or professional degree: 160 (10%)
- High school graduate (or equivalent): 175 (11%)
- Less than 12 years; no high school diploma: 10 (0.8%)
- Other: 2 (0.2%)
- Some college, no diploma; or Associate's degree: 709 (43%)

The plots of the education levels below demonstrate a couple things. First, there were more Non-Wharton listeners among the respondents overall (absolute count was higher). Additionally, listeners to the Wharton show appear to have greater proportions of Bachelor's and graduate degrees, while proportionally more non-Wharton show listeners had some college, no diploma, or an Associate's degree.





Income: Overall

- Less than 15,000: 205
- 15,000 - 30,000: 360
- 30,000 - 50,000: 419
- 50,000 - 75,000: 371
- 75,000 - 150,000: 326
- Above 150,000: 44

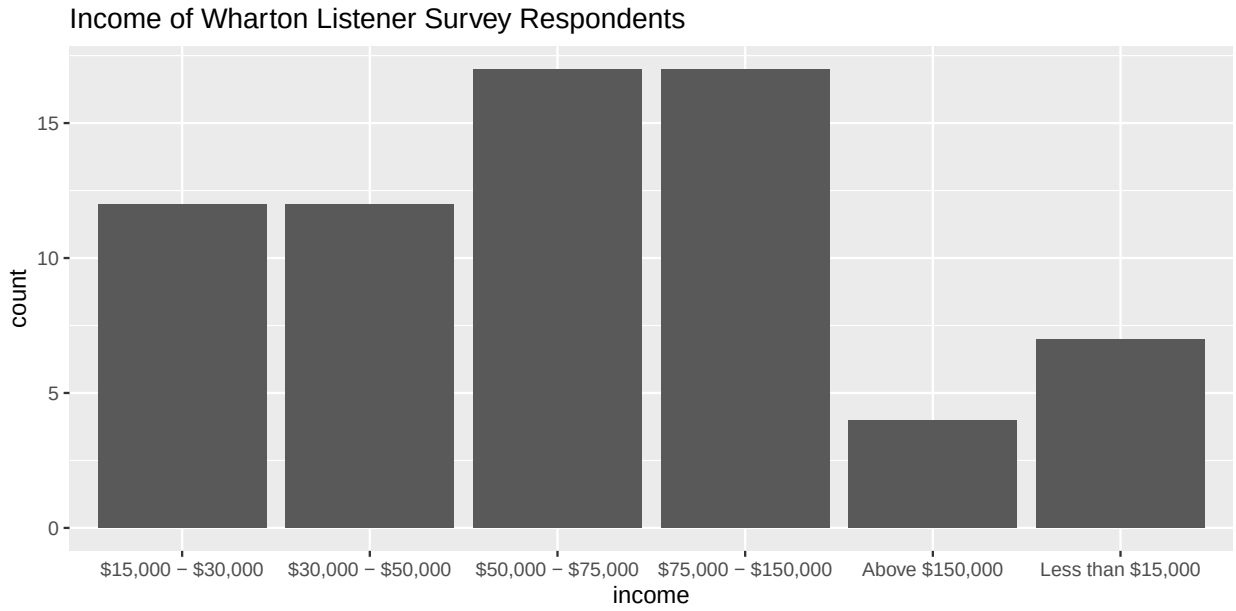
Income: Those who listen to the Wharton show

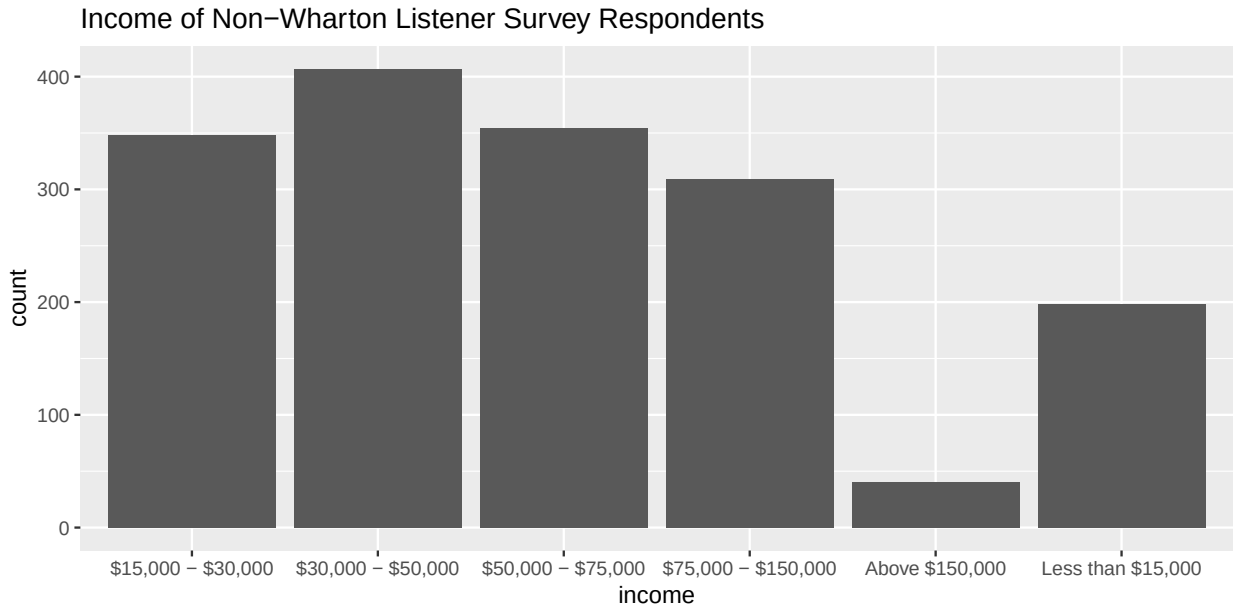
- Less than 15,000: 7 (10%)
- 15,000 - 30,000: 12 (17%)
- 30,000 - 50,000: 12 (17)
- 50,000 - 75,000: 17 (25%)
- 75,000 - 150,000: 17 (25%)
- Above 150,000: 4 (6%)

Income: Those who don't listen to the Wharton show

- Less than 15,000: 197 (12%)
- 15,000 - 30,000: 343 (21%)
- 30,000 - 50,000: 405 (25)
- 50,000 - 75,000: 352 (21%)
- 75,000 - 150,000: 305 (18%)
- Above 150,000: 40 (3%)

Again, these data are shown below in histograms for comparison across the two populations. The plots of the income range distributions below demonstrate that there appears to be an overall higher income amongst those who answered that they listen to the Wharton Show.





Gender: Overall

- Female: 729 (42%)
- Male: 996 (58%)

Gender: Those who listen to the Wharton show

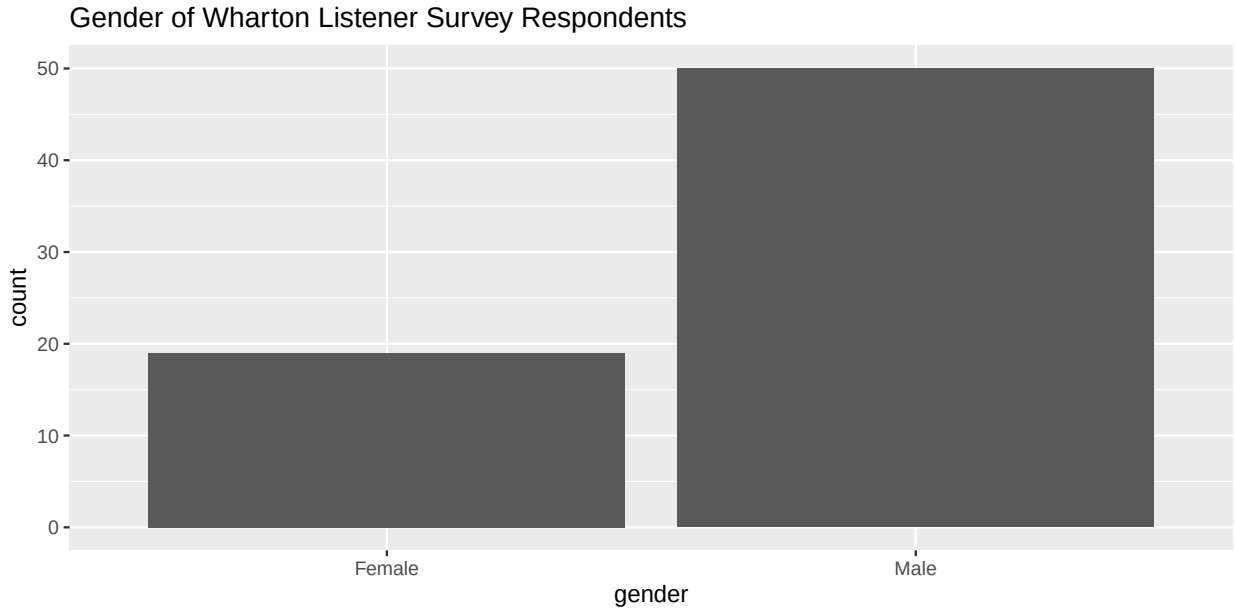
- Female: 19 (28%)
- Male: 50 (72%)

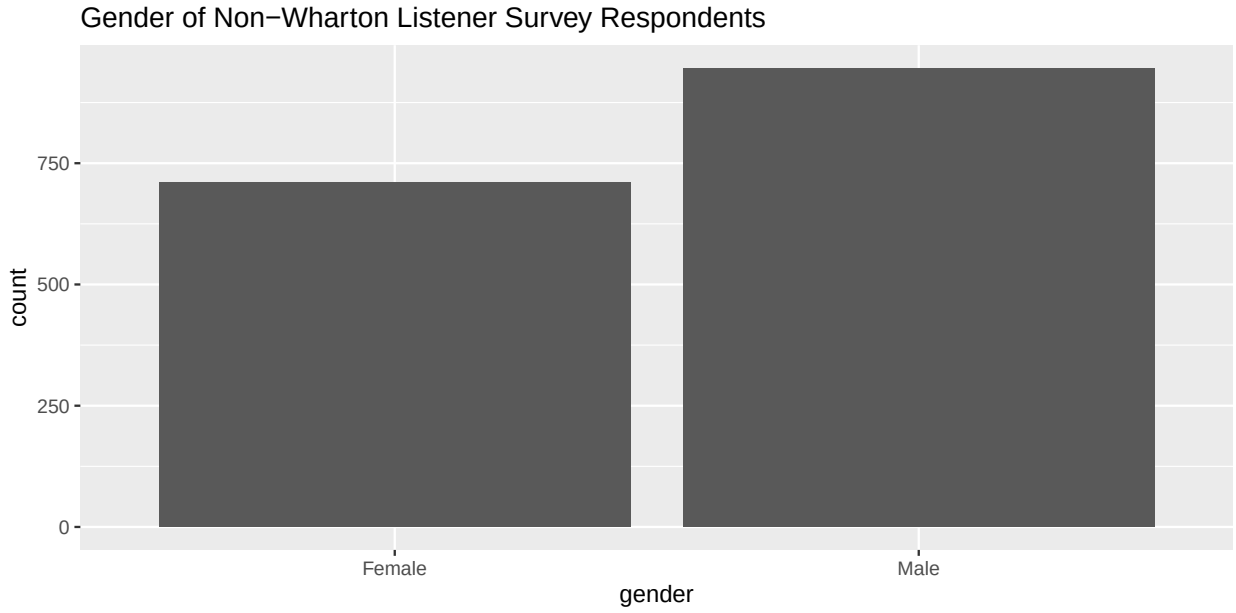
Gender: Those who don't listen to the Wharton show

- Female: 710 (43%)

- Male: 946 (57%)

The distribution of genders for Wharton show listeners reveal that there is a significantly greater proportion of male listeners (70%). In the total sample size and the Non-Wharton show listeners, there is closer to a 50:50 split between males and females.





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?

US population statistics, obtained from the US census:

- Average Age: 38.1 (2019)
- Average Income: 31,133 USD (2019)
- Percent of Women/Men:50.8%, 49.2% (2019)
- Average Education: HS or higher if 25+ (88%), Bachelor's or higher if 25+ (32.1%)

Data from the survey sample:

- Average Age: 30.3
- Average Income: About 25% of the survey respondents fell into the "50,000 - 75,000" income range
- Percent of Women/Men: 42%/58%
- Average Education: HS or higher if 25+ (99%), Bachelor's or higher if 25+ (89%)

Overall: This sample does not appear to be a random sample from the general population of the US. In most of the categories, the averages for the variables of interest vary significantly. For example, our sample was on average 8 years younger than the overall US population and appears to make substantially more than the national average income. Furthermore, our sample has a much larger proportion of men, which further indicates that it is not a random sample from the general population. Lastly, with the examination of education data from our sample, when comparing to survey participants over 25 years old, nearly every participant had a high school degree or higher, and almost three times the percent of people in our sample had a bachelor's or higher.

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

MTURK population statistics, obtained from cloudresearch:

- Average Age: "37% of people on MTurk are in their 30's, another 17% are in their 40's, and roughly 11% are in their 50's."
- Average Income: Research compares average annual incomes of MTURK againt the US population and reveals that the inequal distribution of wealth is much stronger with the US population. For MTURK, the annual income ranges with the largest percentages are 100-150k, 20-30k, and 50-60k at (11.97%, 11.67% and 11.22% respectively).
- Percent of Women/Men: 57%/43%
- Average Education: According to an NYU Stern study, the (self-declared) educational level of the workers is generally higher than the general US and Indian population. This Stern study noted that about 50% of the Indian population have a Bachelor's.

Data from survey sample:

- Average Age: 30.3
- Average Income: About 25% of the survey respondents fell into the "$50,000 - $75,000" income range
- Percent of Women/Men: 42%/58%
- Average Education: HS or higher if 25+ (99%), Bachelor's or higher if 25+ (89%).
  - Based on this information, the average education level for people in MTURK and this sample size were both generally above the US average. The MTURK education level is also above the India average ($>$ 50% have Bachelor's in India compared to 89% with at least a Bachelor's in this survey).

– Given the lack of specific information found, it's difficult to draw conclusions about the accuracy and representation of this sample, but both our sample and the MTURK population generally skew to to be more educated than the general population.

Overall: The sample doesn't appear to be a random sample of MTURK given the difference in gender ratio. However, the average age, education, and income seem to align more than the comparison of data for this sample size and the US population.

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.

Goal of the study

The goal of this study was to estimate the audience size of the 2014 Sirius Radio show, "Business Radio Powered by the Wharton School" through a survey conducted through MTURK. Given the total number of Sirius Radio listeners, we aimed to estimate the proportion of listeners who specifically listen to the Wharton talk show.

Methods

1. Data collection/preparation: We first selected our variables of interest. These were: `Age`, `Gender`, `Education Level`, `Household Income in 2013`, `Sirius Listener`, and `Wharton Listener`. The data was then subsetted to remove missing or inappropriate data. If any of the observations which contained a blank response or default response ("select one"), the entire observation was removed. The age variable was also transformed into a numeric format, and the others were converted to factor variables.

2. Estimation method: To estimate the total number of listeners, we take the proportion ($p = .0504$) of Sirius listeners who listen to the Wharton show, and multiply that by the total number of Sirius listeners given above (51.6 million).

Findings

We find that: 51.6 million total listeners x .0504 = **2.6 million** people listening to "Business Radio Powered by the Wharton School."

Other findings related to the data include that those who listen to the Wharton radio appear to have a greater proportion of individuals with bachelor's and graduate degrees. The bar plot of their incomes also indicates that those who answered "Yes" to listening to Wharton radio generally are in a higher income range. Furthermore, there is a higher ratio of men to women amongst listeners of the Wharton show in comparison to the rest of the sample size.

Limitations of the study

The primary limitation of this study is the representativeness of the sample size. After cleaning the data, the analysis above indicates that the sample size is not an accurate representation of the US population. In particular, the gender distribution was highly inaccurate in comparison to both the US population and the general MTURK distribution. Furthermore, the proportion of Wharton listeners was very small, so it was difficult to make conclusions between the people in the sample who did and didn't listen to the Wharton radio in particular. Since there were only 69 iindividuals who said "Yes", generalizations made about the income and education range of Wharton show listeners may not be accurate.

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.

Method proposed to estimate the audience size:

If our objective is to estimate the audience size of Wharton Business Radio Show, we could do so by taking a sample directly from Sirius Radio listeners and find the proportion of Wharton Business Show listeners by directly examining the data from Sirius. With the 1,000 USD budget given to conduct the survey, we would announce on Sirius Radio to its listeners that completing the survey will automatically enter them into a competition for a 50USD prize (hoping that a larger prize may incentivize a more random sample) and that 10 winners will be selected (for a total of 500USD). Listeners would be directed to the Sirius website to access the survey. The remaining 500USD would be allocated to create the survey, have it announced on Sirius Radio during times where there may be the most listeners, and display an ad to the survey on the Sirius website.

Data to be collected:

- Age
- Gender
- Ethnicity
- Income
- Education
- Do you listen to Wharton Radio?

Source of data:

The survey would be announced directly on Sirius Radio 3x a day (every 8 hours) for one month to aim to capture the most responses. An ad would also be displayed on the Sirius website with a link to the survey form.

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.

3.1 Data preparation

1. Understand and clean the data

Because the data came as an Excel file, we used the package `readxl` and the function `read_excel()` to read the data `WomenData_06_16.xlsx` into R.

In the process of cleaning this dataset, we renamed the `Field and sex`, `Degree`, `Sex`, `Year` and `Degrees Awarded` variables to `field`, `degree`, `sex`, `year`, and `number`, respectively. We also set the `field`, `degree`, and `sex` variables as factors to accommodate their nature as categorical variables. There do not appear to be any missing values in this dataset.

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

In this dataset, we have 660 total observations (330 female, 330 male). Years of data range from 2006 to 2016, and degree types are split evenly among BS, MS and PhD degrees (220 of each). Furthermore, the data includes 66 observations of each of the following 10 fields:

- Agricultural sciences
- Biological sciences
- Computer sciences
- Earth, atmospheric, and ocean sciences
- Engineering
- Mathematics and statistics
- Non-S&E (non-science-engineering)
- Physical sciences
- Psychology
- Social sciences

In the following sections, we dive deeper in our exploration of the data.

3.2 BS degrees in 2015

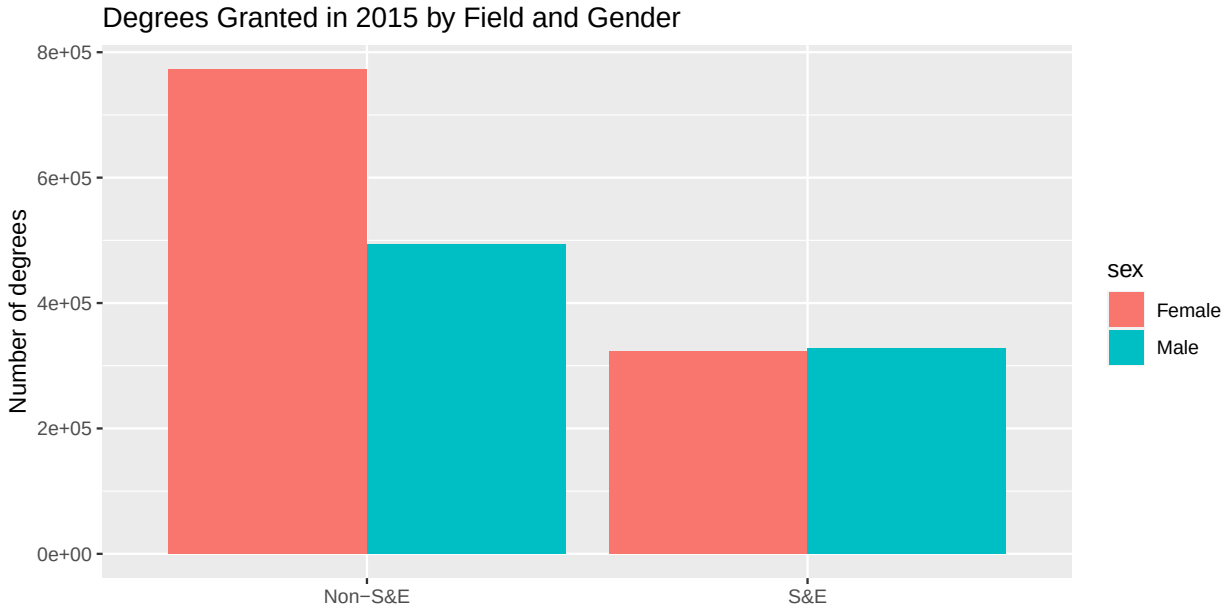
Is there evidence that more males are in science-related fields vs `Non-S&E`? Provide summary statistics and a plot which shows the number of people by gender and by field. Write a brief summary to describe your findings.

To answer this question, we grouped all fields that were not `Non-S&E` into a new variable called `S&E`, which represents the science-related fields. Then, we filtered the data for BS degrees in 2015. Lastly, we identified the number of people by gender and field:

sex	SE	degrees
Female	Non-S&E	772768
Male	Non-S&E	493304
Female	S&E	322935
Male	S&E	327122

We can see from the table above that, within the science-related fields in 2015, there was not a significant gender disparity at the BS level, with about 323K women in science-related fields, and 327K men in science-related fields. However, by contrast, many more women were in non-science-related BS programs in 2015 (773K vs. 493K, respectively).

We represent these numbers in bar graph form as follows:



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.

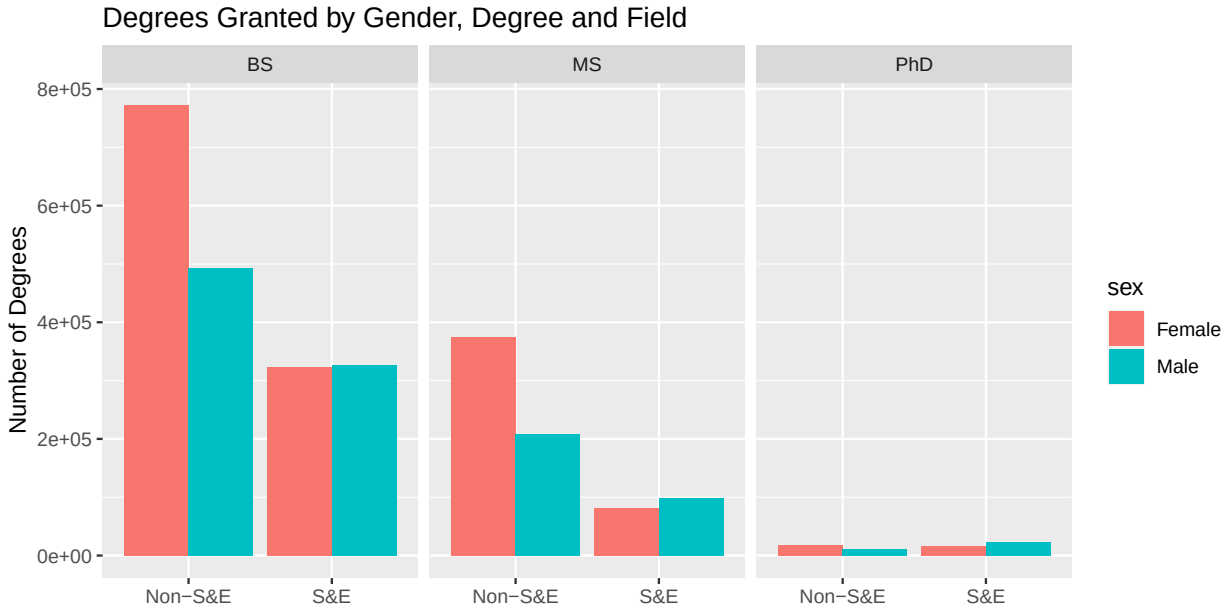
Again, we grouped the non-science-related fields into one category, `S&E`. The table below shows the number of degrees by degree type, field and gender, sorted in descending order:

degree	SE	sex	count
BS	Non-S&E	Female	772768
BS	Non-S&E	Male	493304
MS	Non-S&E	Female	374024
BS	S&E	Male	327122
BS	S&E	Female	322935
MS	Non-S&E	Male	209001
MS	S&E	Male	99282
MS	S&E	Female	81673
PhD	S&E	Male	22914
PhD	Non-S&E	Female	18396

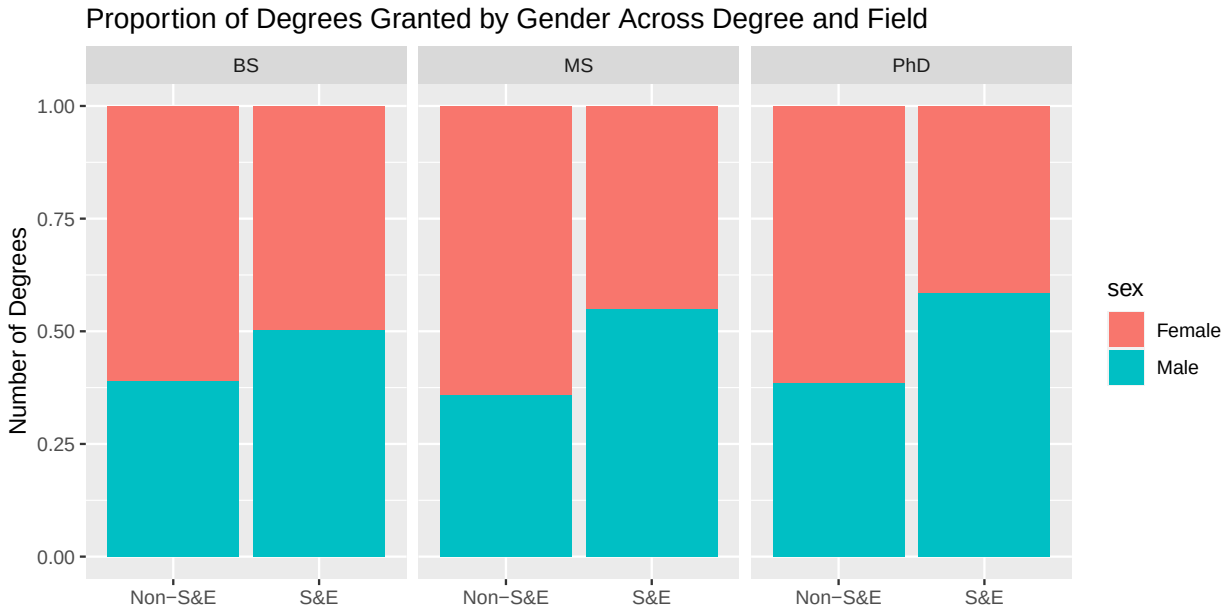
degree	SE	sex	count
PhD	S&E	Female	16264
PhD	Non-S&E	Male	11541

Overall, in 2015 the greatest number of degrees are non-science-related and awarded at the BS level (as previously identified, these number 773K and 493K for females and males, respectively). Interestingly, while fewer MS degrees were awarded than BS degrees overall, more women obtained MS degrees in non-S&E fields (374K) than women or men at the BS level in S&E fields (323K and 327K, respectively). We also note that at each level, more S&E degrees are obtained by men (although not by a large amount). The fewest number of degrees were PhDs awarded to women in S&E fields (16.3K) and men in the non-S&E fields (11.5K).

We graphically represent these findings below:



In the image above, the number of PhDs is too small to meaningfully observe the relationship of degrees awarded to men vs. women. The following presents the same information using stacked bars to more clearly show the proportions of men vs. women across degrees. Note that the proportional pattern is similar across all three degree types, with men having slightly greater representation in science-related fields, and women comprising a greater proportion of degrees in non-science fields:



3.4 EDA bring all variables

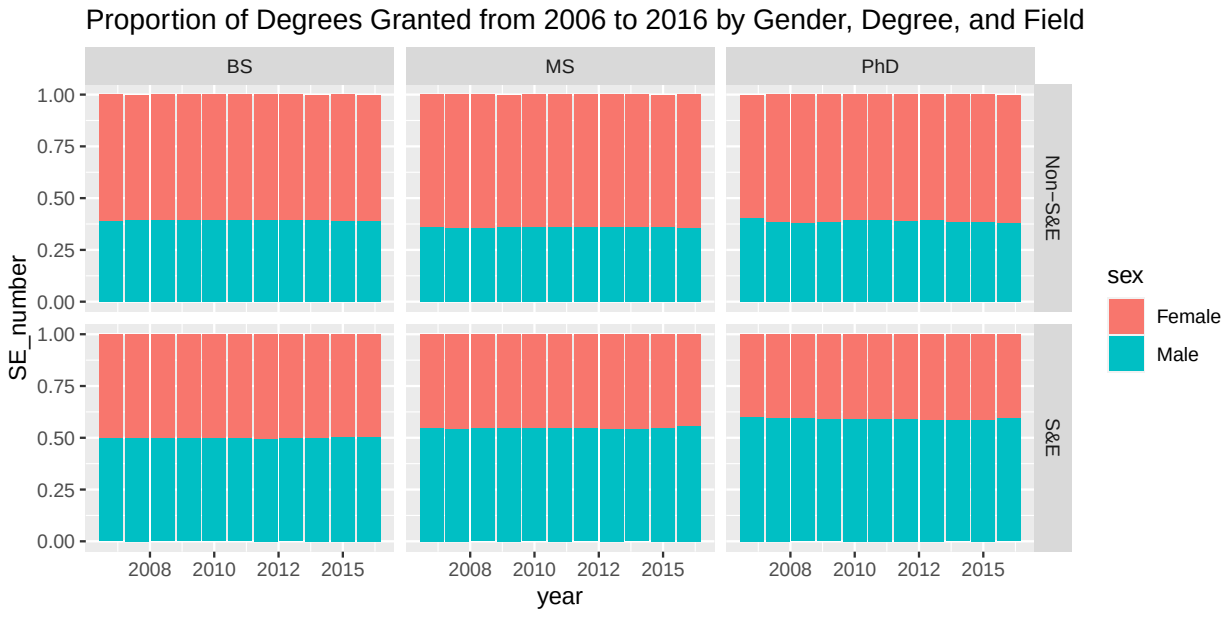
In this last portion of the EDA, we ask you to provide evidence numerically and graphically: Do the number of degrees change by gender, field, and time?

When adding the time element into the EDA, the datatable grew to 44 rows, so we refrained from including it in the report. While it may be difficult to glean patterns from the numerical data given the number of observations, we arranged the data by gender, field, and year for clarity. In doing so, we see that the number of women in science-related fields did increase from 308K degrees (across all levels) in 2006 to 435K S&E degrees in 2016. The number of women obtaining degrees in non-S&E fields also increased in this same period, from 946K in 2006 to 1.2M in 2016. The number of men receiving degrees in both S&E fields and non-S&E fields also increases over time: in S&E fields, the number increases from 323K in 2006 to 469K in 2016, and in non-S&E fields, these numbers are 582K and 718K, respectively.

These findings are more digestible in graphical form, as shown below:

As discussed above, the total number of degrees increases over time for both men and women. This increase is most apparent for BS degrees. It is also perceptible for MS degrees, but less so for PhD degrees. For BS degrees, the number of females vs. males obtaining degrees in science-related fields is quite similar, whereas clearly more females obtain BS degrees in non-S&E fields. Similar patterns can be seen for MS and PhD degrees. Fewer degrees overall are granted at the MS level, and PhDs are granted the least.

The same information from above is presented in stacked bar graph form to better visualize men vs. women proportions at the MS and PhD degree levels:



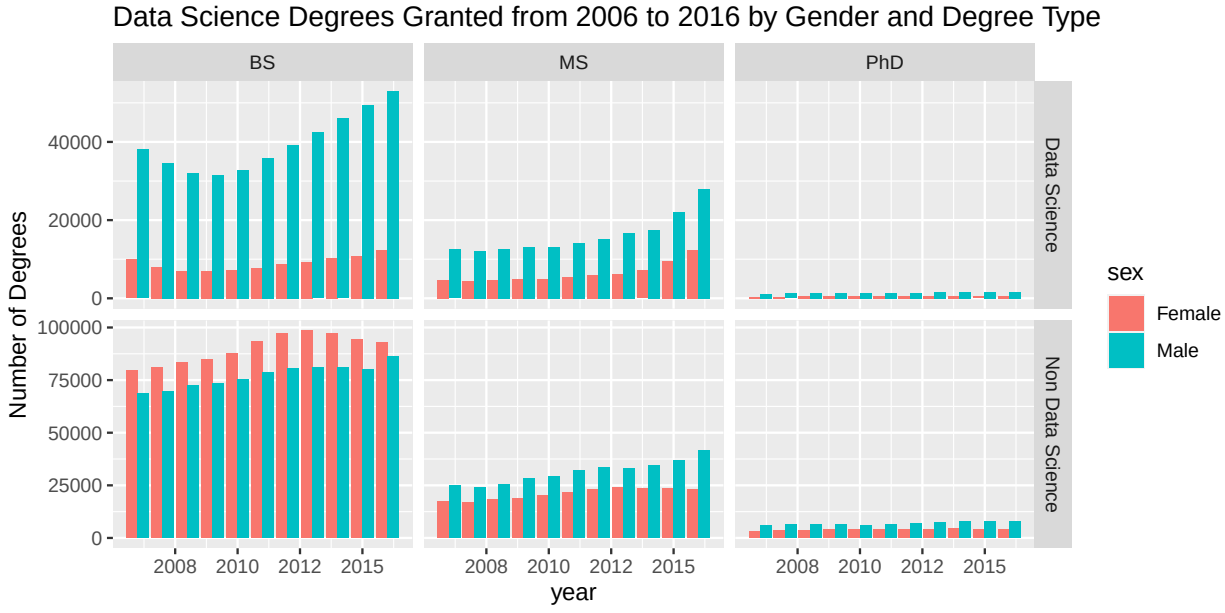
Again, we see here that more females than males are granted non-S&E degrees at all levels. Within fields, the proportion of males vs. females is approxiately constant across degree types and years.

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.

To examine the data for an answer to this question, we created a new variable, `Data Science`, which comprised of the `Computer sciences` and `Mathematics and statistics` fields. We first examined the data by year and gender within the field of data science. In doing so, we do find evidence that women are underrepresented in data science. We see that each year, women are outnumbered by men. Furthermore, while the number of men and women in data science both increase over time, this disparity remains in place. Specifically, in 2006 the numbers of men and women in the field were 63.8K and 24.2K, respectively (164% more men), and these numbers in 2016 were 103K and 39.7K, respectively (160% more men).

Again, we represent these data graphically below, also breaking the data down by degree type:



Seen graphically above, the disparity between numbers of men and women in data science is quite drastic at

all levels and time points. The stacked bars make quite clear the fact that the underrepresentation of women in data science has not changed over time.

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 this dataset, we do consistently see more men pursuing science-related fields than women. However, if we consider "science-related fields" to include all fields included in the data except the Non-S&E category, this difference is quite subtle. For instance, in section 3.2 we see that only slightly more men pursued S&E fields than women, which is actually a pattern that remains consistent across years (see section 3.4). However, when the S&E fields are focused to specifically examine data science, we see that the disparity is far more severe (section 3.5).

A limitation in this analysis is that we grouped all the non-S&E fields into one "science-related" field, S&E. However, in reality, S&E encompasses a broad range of topics of study, from psychology to engineering. As we saw, once we focused on the data science field, underrepresentation of women increased. This relative underrepresentation of women in data science implies that women experience greater representation in other science-related fields, such as psychology or the biological sciences. In future analyses, more information could be gleaned by examining fields of study at a more granular level.

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.

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.

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.

The log difference in payroll is more appropriate than absolute difference in payroll because the log difference controls for differences in base size. For example, a 1-dollar increase in salary is a 10% increase if the base is 10 but is a 1% increase if the base is 100. Essentially, log differences capture relative differences as opposed to absolute differences to make comparisons across groups more appropriate.

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

team	year	diff_log	win_pct
Arizona Diamondbacks	1998	NA	0.401
Arizona Diamondbacks	1999	0.802	0.617
Arizona Diamondbacks	2000	0.139	0.525
Arizona Diamondbacks	2001	0.002	0.568
Arizona Diamondbacks	2002	0.236	0.605
Arizona Diamondbacks	2003	-0.243	0.519

We created a new variable, `diff_log` using `log(payroll) - lag(log(payroll)`. We then created a long table that includes the `team`, `year`, `diff_log`, and `win_pct` variables only. The table has a total of 510 observations; the first 6 are shown in the table above.

4.2 Exploratory questions

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

The five teams with the highest increase in payroll from 2010 to 2014 were the:

1. LA Dodgers
2. Texas Rangers

3. SD Padres
4. Pittsburgh Pirates
5. Washington Nationals

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

The five teams with the highest increase in percentage wins from 2010 to 2014 were the:

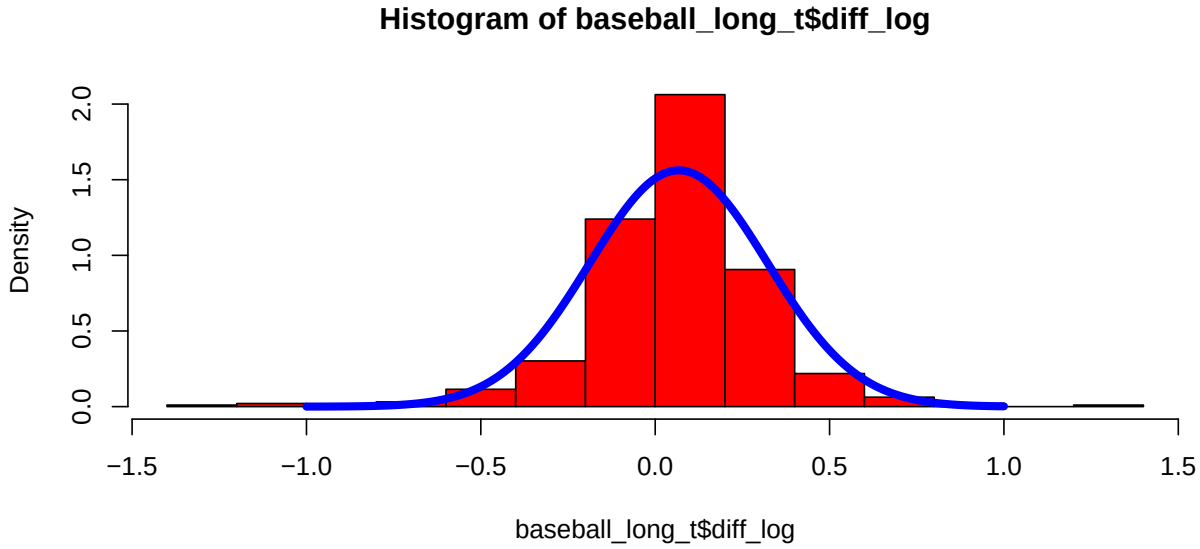
1. Pittsburgh Pirates
2. Baltimore Orioles
3. Seattle Mariners
4. Washington Nationals
5. KC Royals

4.3 Do log increases in payroll imply better performance?

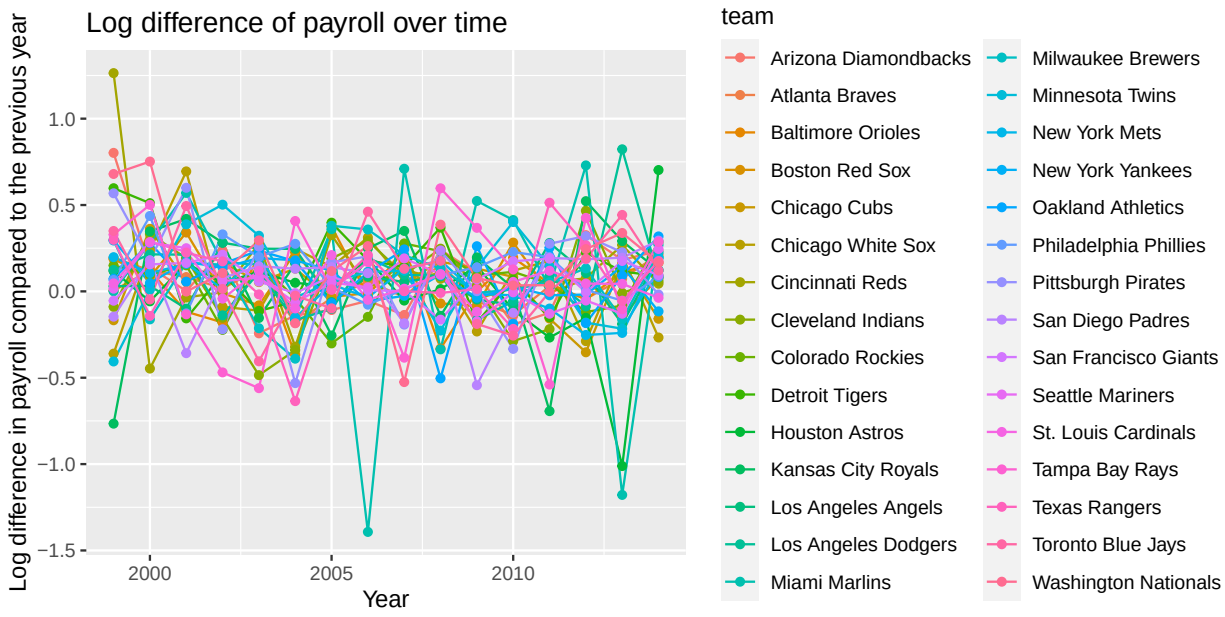
Is there evidence to support the hypothesis that higher increases in payroll on the log scale lead to increased performance? Pick up a few statistics, accompanied with some data visualization, to support your answer.

We felt that the best way to examine this hypothesis was through data visualization. Our analyses of the data are as follows.

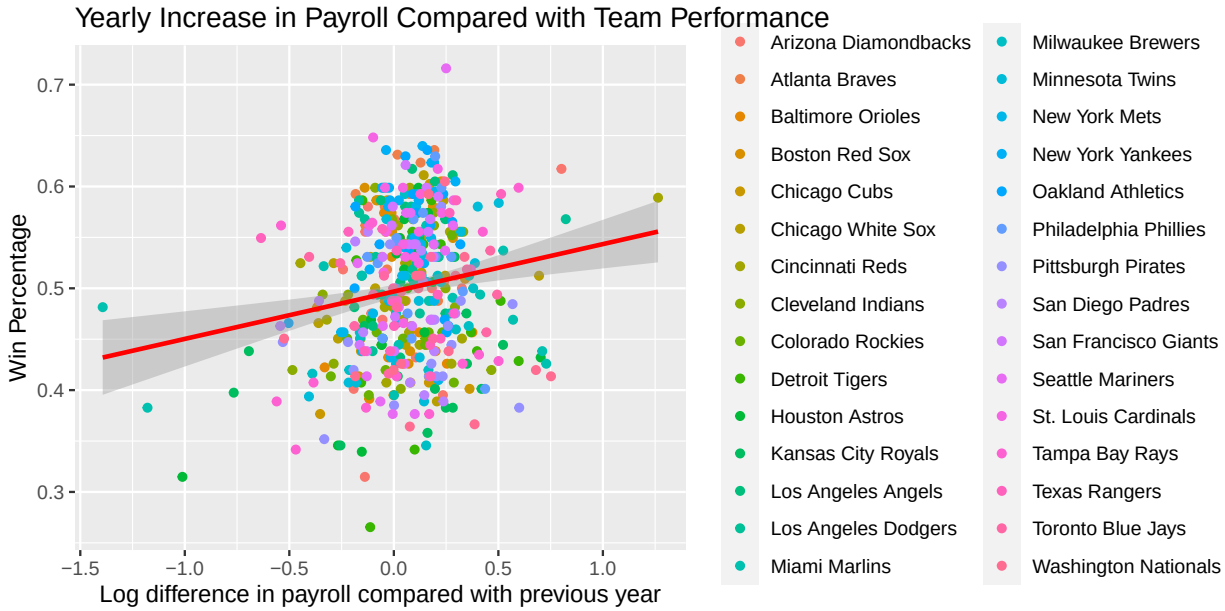
First, we plot the distribution of the log difference in payroll each year against a normal distribution to verify that the distribution of our variable of interest is approximately normal:



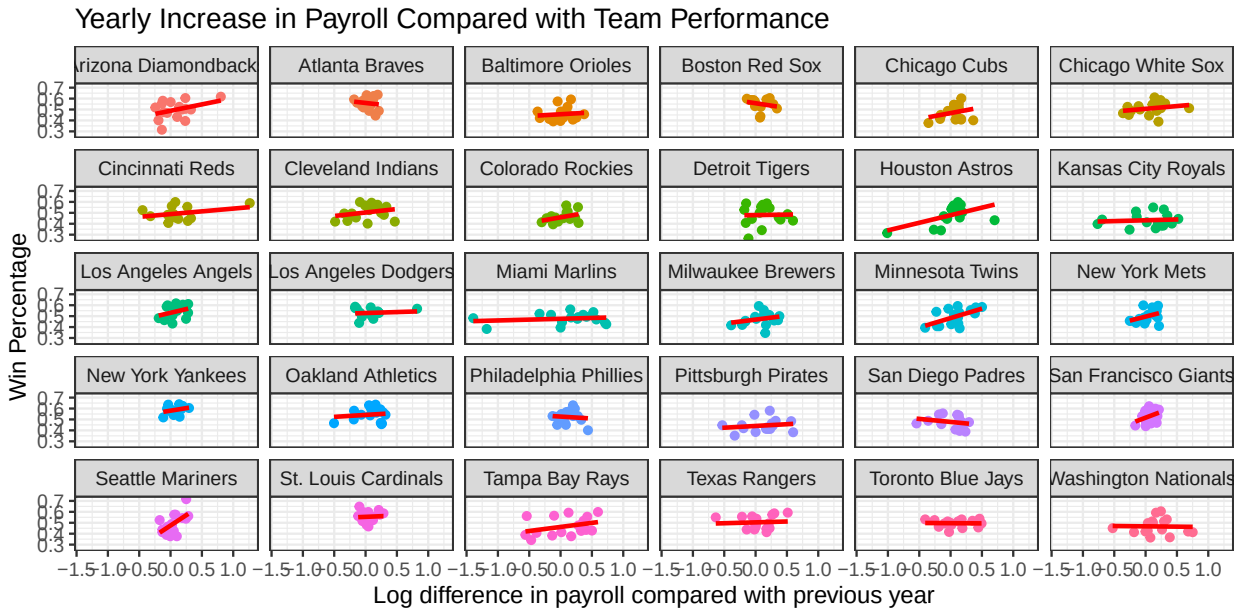
We also plot the log pay difference over time to check for any existing pattern over time. As we see below, there is no clear pattern to pay increases over time. As in, while pay may be increasing over time, the difference year to year is not increasing.



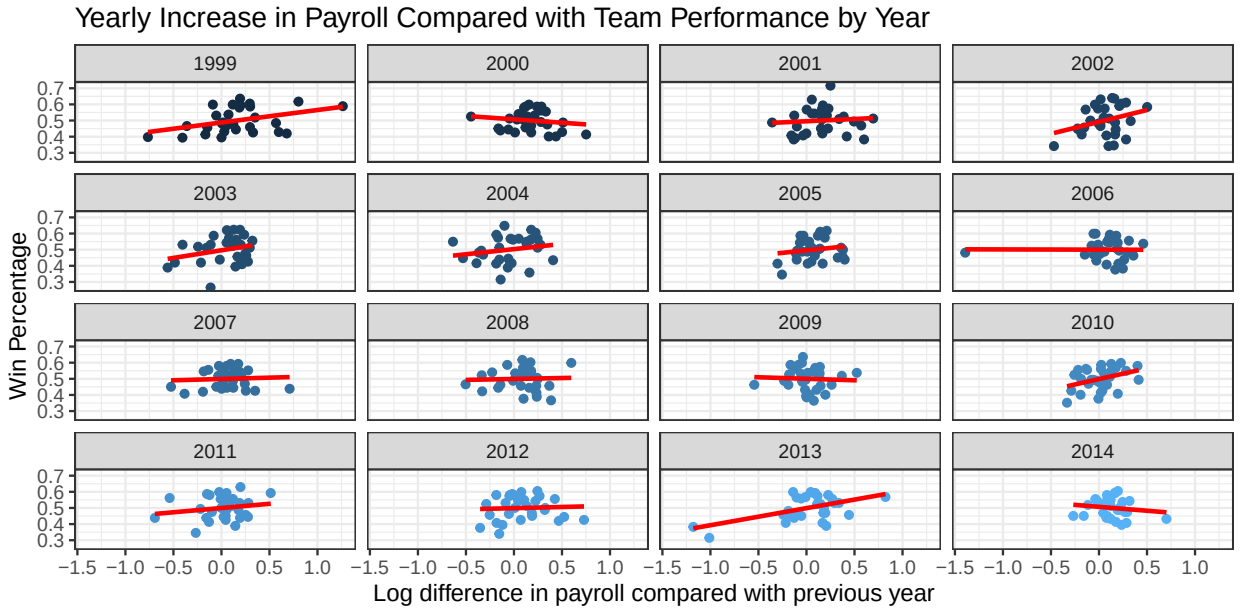
Next, we fit a linear model to our dataset, with the log difference in payroll on the x-axis, and win percentage on the y-axis. We also include a line of best fit; in doing so, we see that there does appear to be a positive relationship between the log difference in payroll and the team's performance. However, the strength of the relationship is questionable, and the correlation of the two variables does not appear to be very strong.



We parse out the data above by team, examining this relationship of log difference in payroll and win percentage as follows:



We also examine this relationship of log difference in payroll and win percentage year by year:

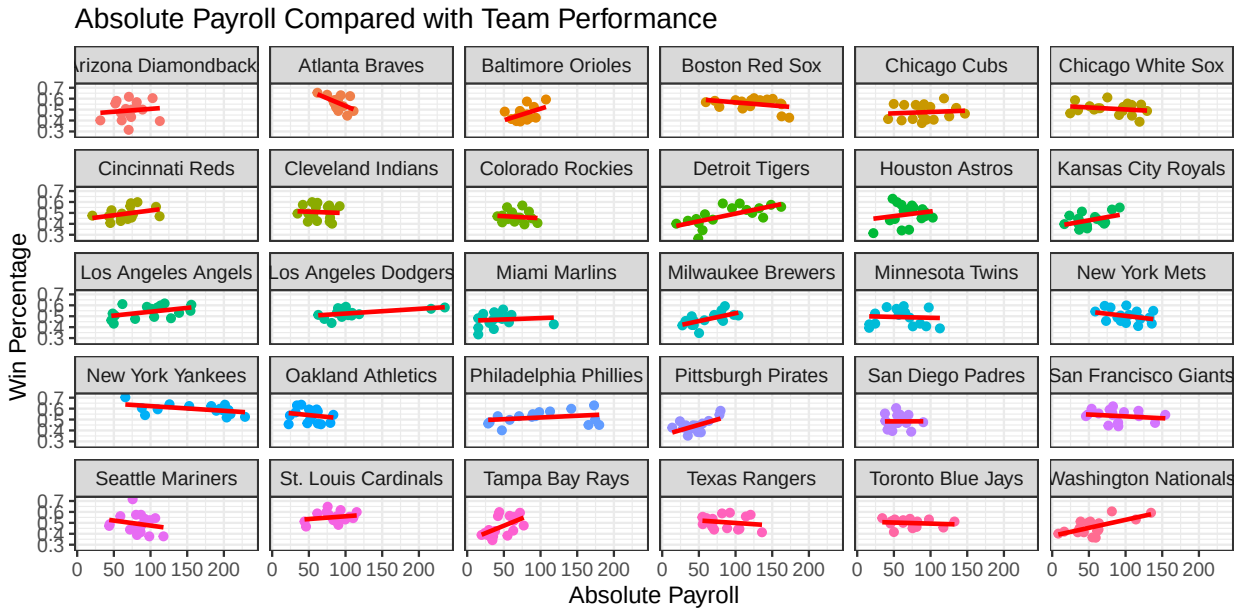


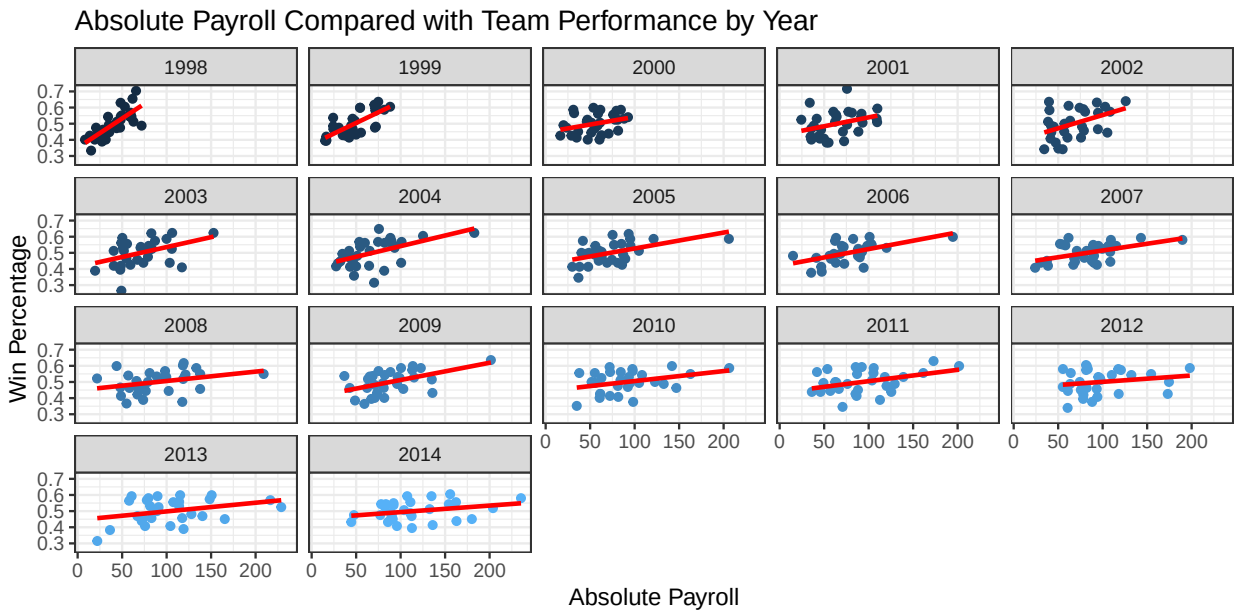
As these visualizations show, the correlation between increase in payroll and win % vary from team to team. Some teams show a very strong positive correlation (such as Arizona, Houston, and Minnesota) while others show very low correlation (e.g., San Diego Padres, Miami Marlins). Thus, there is some evidence to suggest that increases in the log difference in payroll lead to better outcomes on a team-by-team basis. Additionally, the data also show that the log increases in payroll year over year are consistent, so we do not have to factor in the "time effect," and only need to consider how the effect may vary by team.

4.4 Comparison

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

To answer this question, we repeat our analyses above. However, rather than use the log difference, we plot the win percentage of teams against the absolute yearly payroll. Again, we first stratify by team, and then by year:





Comparing these data to our previous plots, yearly *increase* in payroll appears to explain performance better than yearly payroll. As we saw in class, payroll increases over time, and that increase over time created noise in the analysis. However, by using the yearly increase in payroll, we eliminate the noise of the timing effect. As such, this relationship is only affected by the variation among the different teams. Furthermore, when we examine the log increase in payroll correlations with win percentage on a team-by-team basis, there are more teams with positive correlation between the 2 variables when compared to the analogous plot using absolute payroll, where many teams have little or even negative correlation between payroll and performance. As such, this would also sugget that yearly increase in payroll may be better at explaining performance than absolute payroll.