Modern Data Mining, HW 1

Claes Sjoborg Mark Haghani Kevin Yu

Due: 11:59PM, Jan. 29th, 2023

Contents

```
knitr::opts_chunk$set(echo = FALSE, results = "hide", fig.width=8, fig.height=4)
options(scipen = 0, digits = 3)  # controls base R output
# check if you have ISLR package, if not, install it
if(!require('pacman')) {install.packages('pacman')}
```

```
## Loading required package: pacman
```

```
pacman::p_load(ISLR, readxl, tidyverse, magrittr, dplyr, ggplot2)
```

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".

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.

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.

First, checking to see if there are any NA entries in the dataset

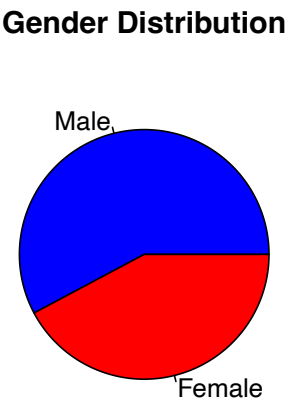
Viewing the data set, we can see that there are some missing/incorrectly input values. My approach was to drop columns that had missing or wrong entries values but if some neeeded their type changed then I hard coded this. These ID's were dropped because they have no/completely wrong inputs: 479, 1643, 663, 1018, 1261, 1251, 1386, 886, 100, 963 764, 784, 479, 497 261, 578, 559, 1481, 856, 1334 There were several that did not choose their level of education so they were also dropped: 377, 179, 1392, 304, 1330, 1294, 605, 379, 1060, 212, 580, 739, 1127, 853, 1040, 1702, 1218

1008: We changed eighteen to 18

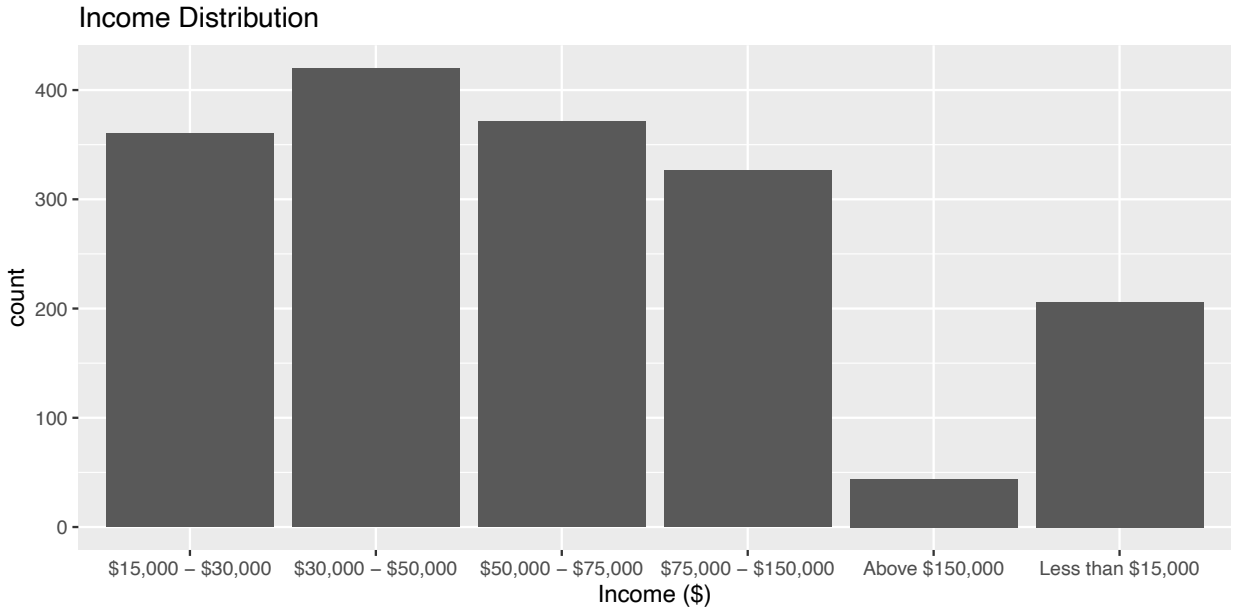
3. Brief summary

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.)

The majority of respondants are in the younger years (>30) which would make sense they would perform a survey for less money as perhaps they are students or young people looking for a quick and easy way to make money.

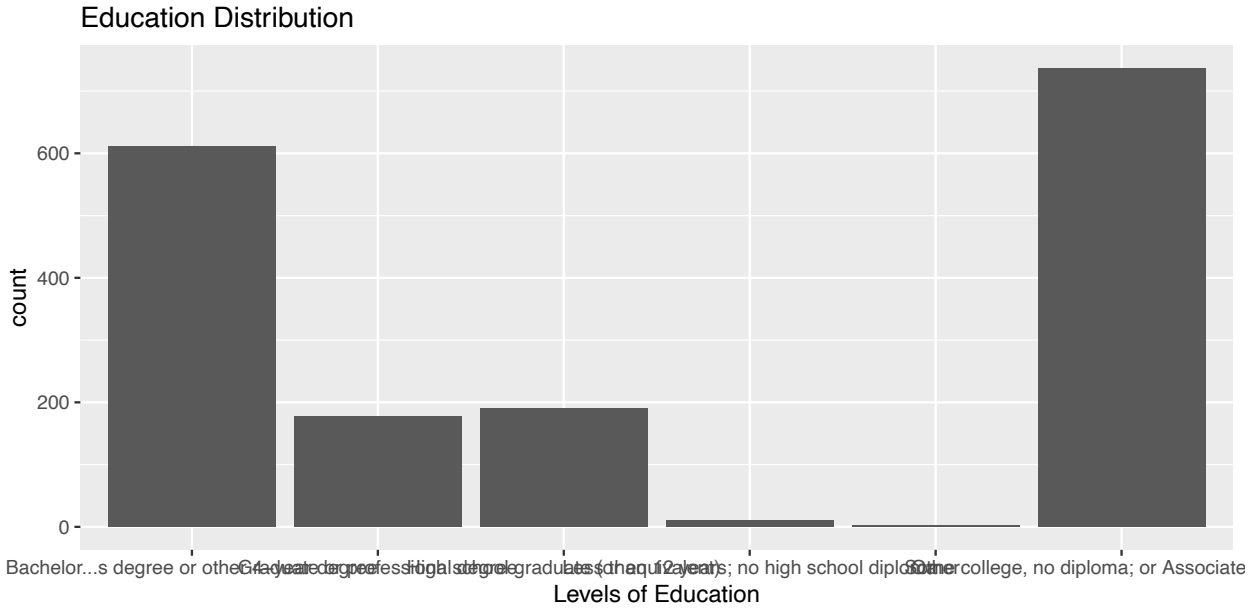


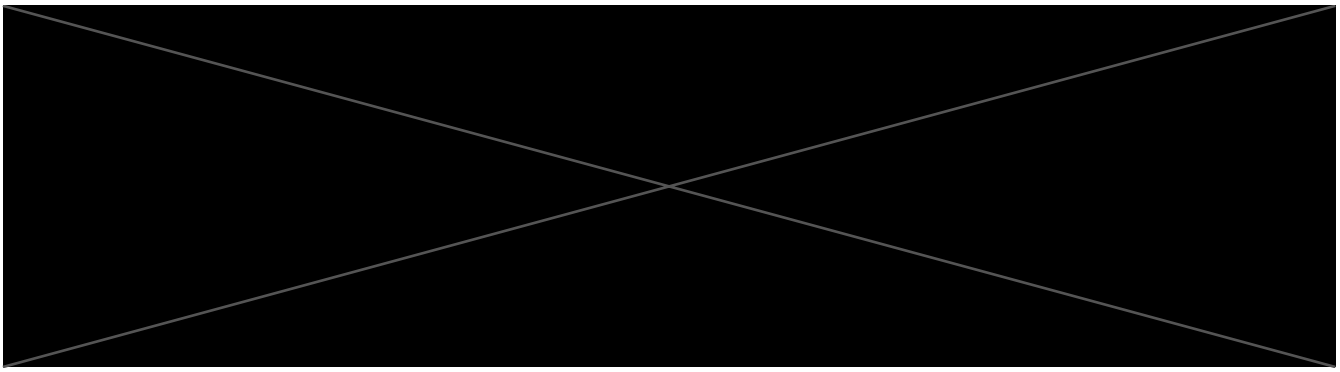
Majority of the survey was filled out by Males (57.6%) vs Females (42.4%).

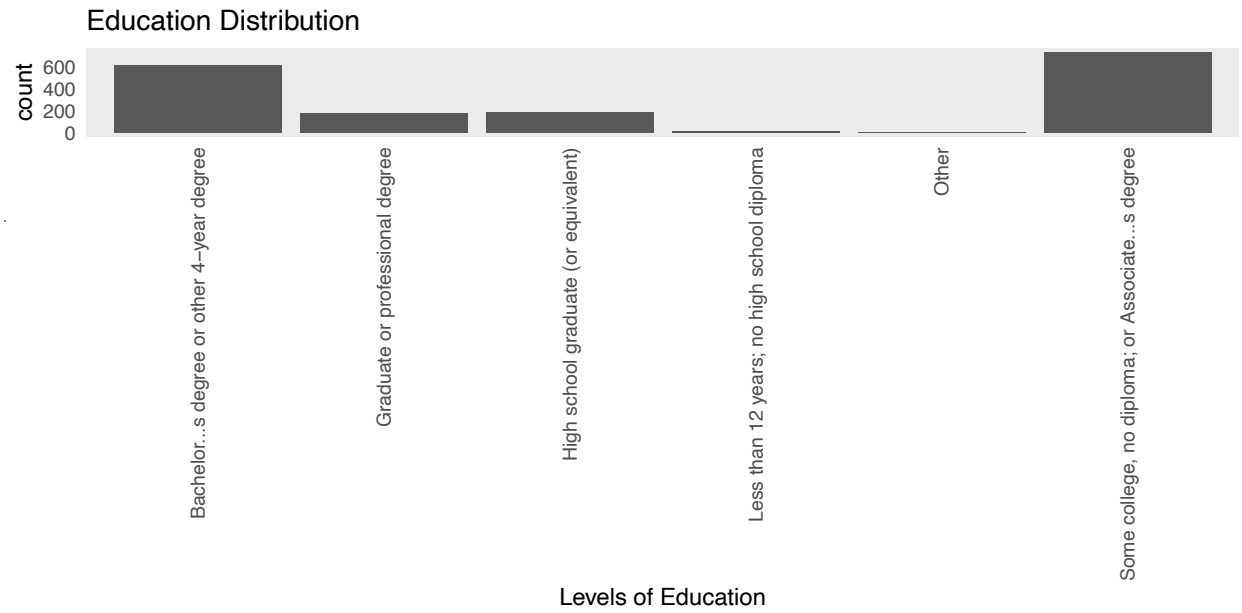


The majority of respondents have an income of >$75K. The largest group is $30-$50K. The majority of respondents being of a lower income bracket would also support the initial hypothesis of younger people wanting to make quick money.

Link between income and Age Gender Distribution show income variation with Age


Education Distribution
600
400
200
0
count
Bachelor...s degree or other 4-year degree
Graduate degree or professional degree
High school graduate
Less than 12 years
Some college, no diploma; or Associate
Levels of Education





The largest group has had some college education or Bachelors which suggests that most of the respondents are students (between ages 18-22).

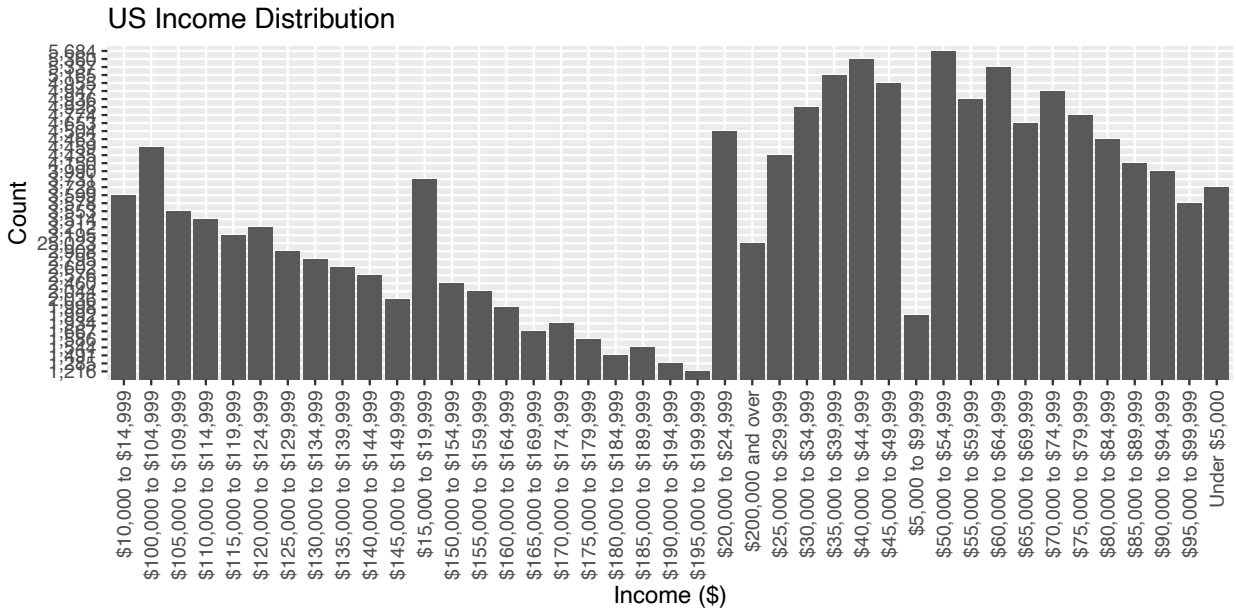
2.2 Sample properties

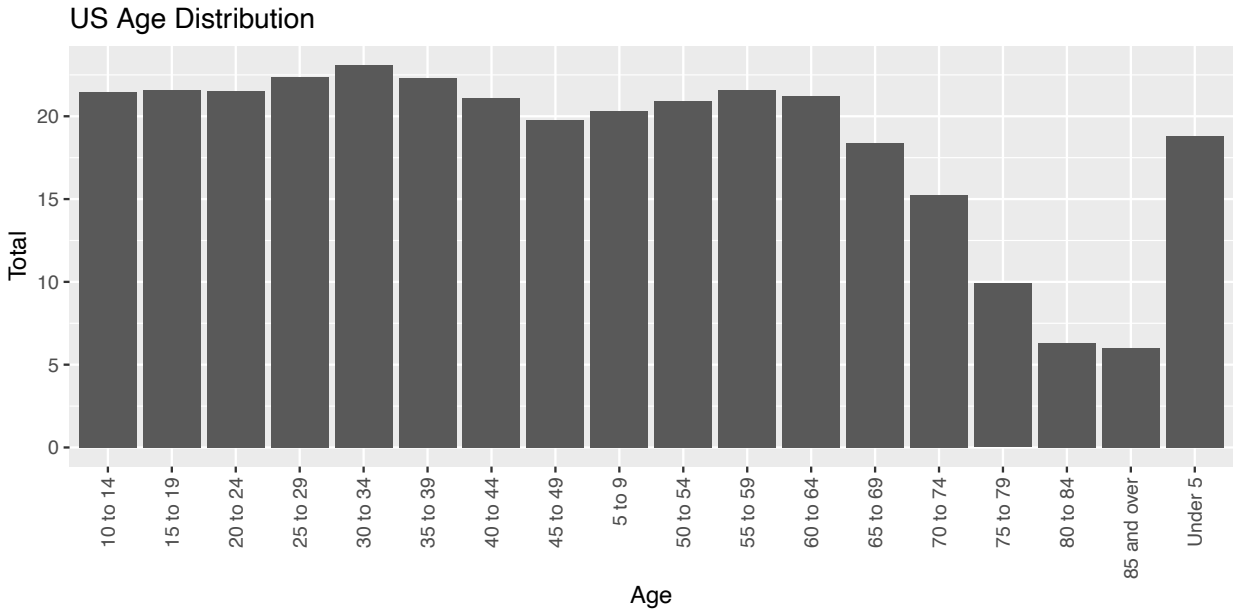
The population from which the sample is drawn determines where the results of our analysis can be applied or generalized. We include some basic demographic information for the purpose of identifying sample bias, if any exists. Combine our data and the general population distribution in age, gender and 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? It does not. The general US population has different income, Age, and levels of education (explored in plots below)
2. 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.

Our sample is very similar to a random sample from MTURK's population (in Age distribution and income levels). (https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/)





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.

Goal: The goal of the study is to use the proportion of Sirius Radio listeners who also specifically listen to the Wharton Radio show and use this figure to estimate the audience size specifically for January 2014.

Method Used: A survey of 1743 participants were used from the MTURK platform. The data that was gathered included: age, gender, income, education level, and if they listen wharton & sirius radio.

Findings: p value is 0.0491 (4.91%) of sirius listeners listen to the wharton radio. Therefore if we assume there are 51.6 million sirius lsiteners, wharton's audience size would be 2,533,560 listeners. This would be for the total year so if we wanted to do a single month we could collect more specific data for January to see if there are seasonal changes. Alternatively, a more crude approach could be dividing the number by 12 (months) and seeing the audience size for one month.

Limitations of the study. Relatively small sample size with not all pariticipants being Sirius XM listeners. Also, the MTURK population might not be representative of the US population as a whole. There may be significant demographic differences.

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.

Method: An online survey will be used to estimate the audience size of the Wharton Business Radio Show. The survey will be conducted using SurveyMonkey, a popular online survey tool. Instead of going through Sirius XM, we will be finding out the total population that listens to wharton radio directly.

Data: Stratified sampling will ensure we can accurately represent the population. subcategories will be made based on age, education level, gender, income, and if they listen to Wharton Radio. This demographic information will be collected in the beginning of the survey. The sample size will be around 2000 people.

Distribution: Through social media (Facebook) and apps designed for survey collection. To cater to the elder demographic, letters will also be sent as well as emails.

Budget: Our budget is 1000 dollars. If we have 2000 people we can offer 40 cents per survey we use (which leaves us $200). SurveyMonkey's flat fee is $39/month. We will advertise for 1 month which gives us enough time to collect the information we need.

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`)

c). Set the variable natures properly.

d). Any missing values?

There are no missing values in df

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

a). How many fields are there in this data?

There are 10 fields in this data set

b). What are the degree types?

BS, MS, PhD

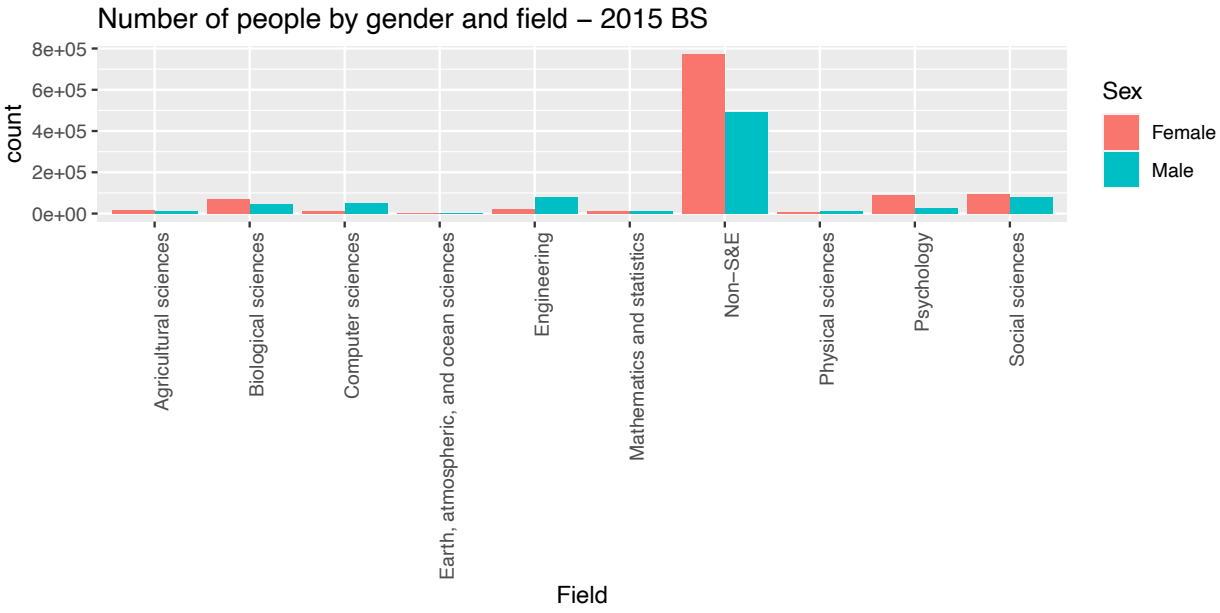
c). How many year's statistics are being reported here?

There are 11 years being reported

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.

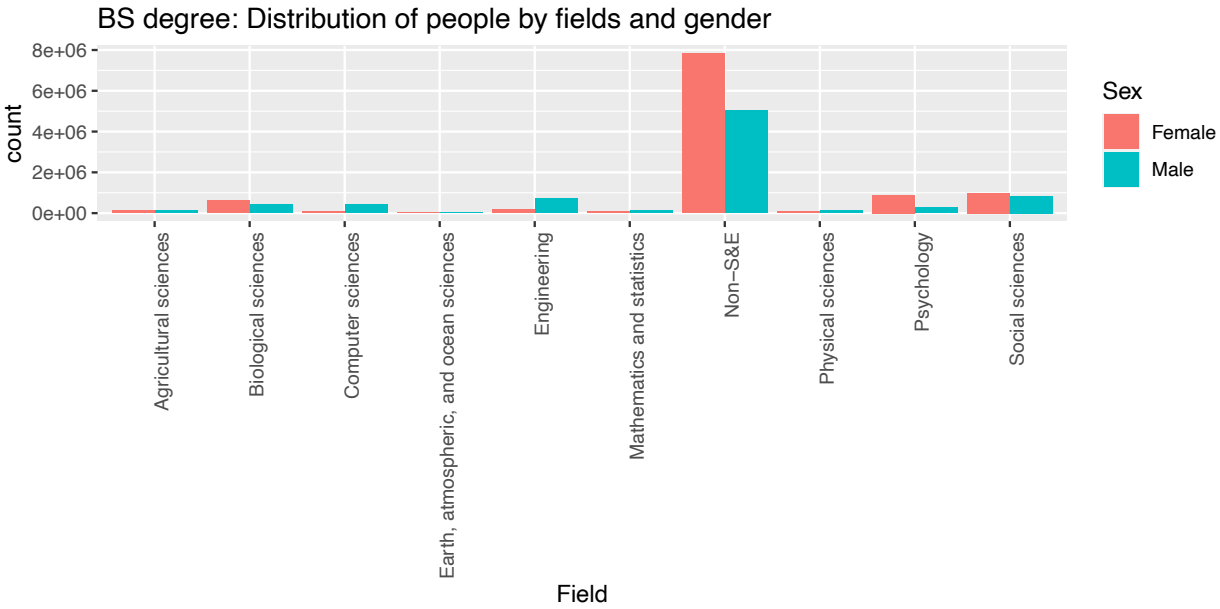
In 2015, there are a similar number of men and women getting a BS in Science Related fields. However, there are many more women BS degrees than men BS degrees for Non Science Related fields. The summary statistics are that in Non-S&E there are: (Females = 772768), (Males = 493304). In S&E there are: (Females = 322935), (Males = 327122).

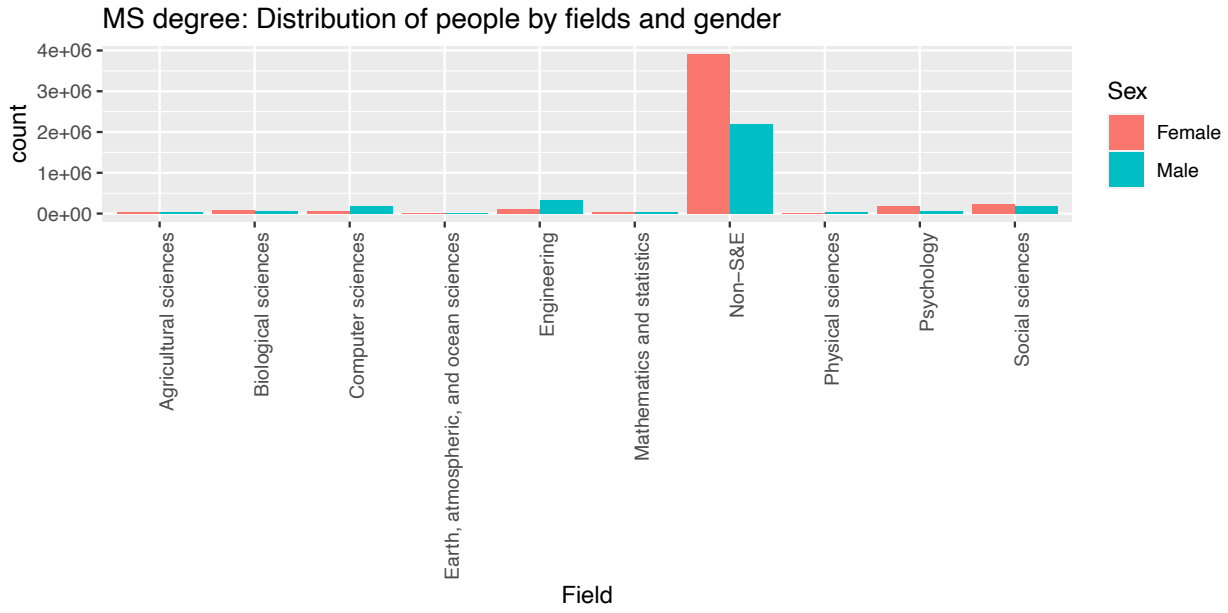


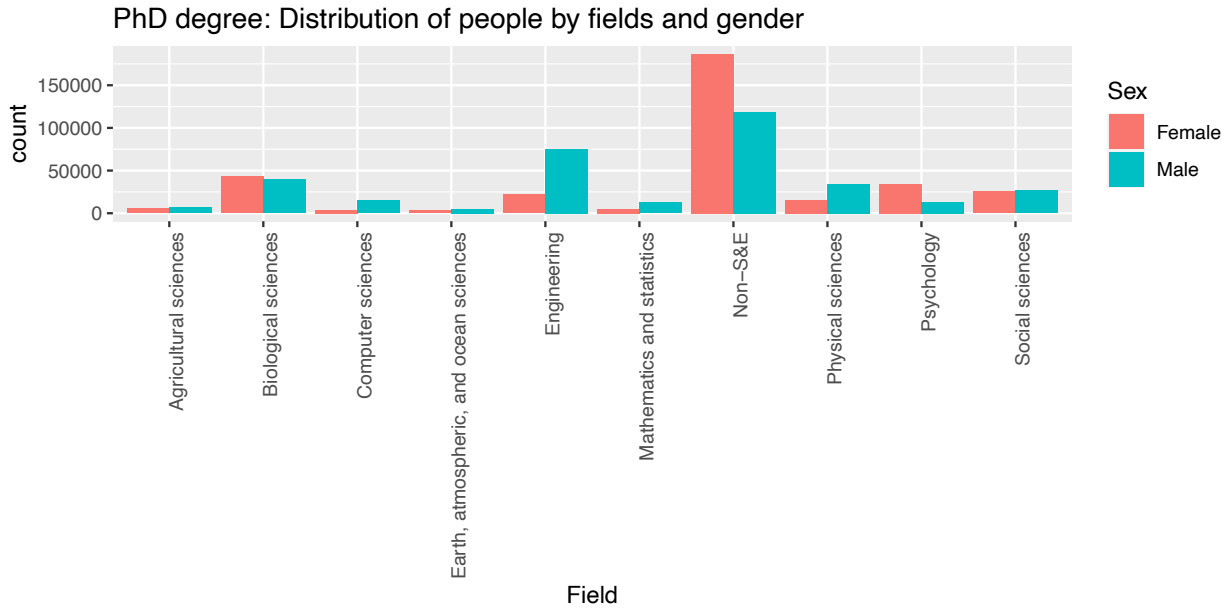
We found that there are more women studying Non S&E subjects and that Engineering subjects tend to have the largest discrepency between men and women outside of Non S&E subjects. We also saw that Psychology had the largest difference in favor of women (more women then men) of all the S&E subjects.

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.



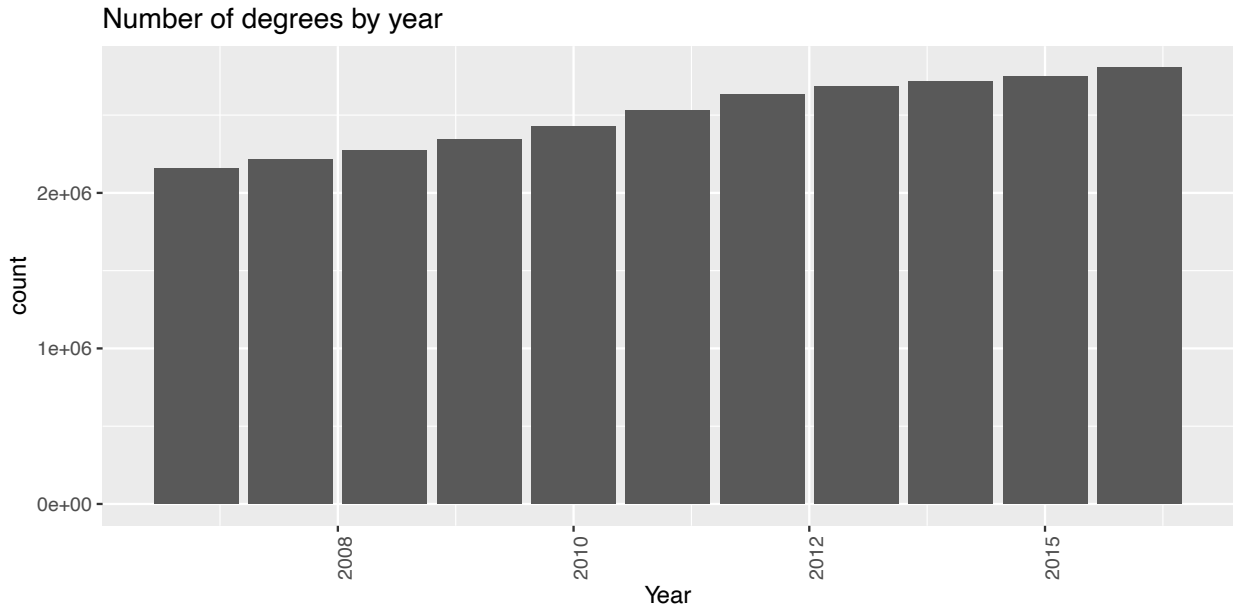


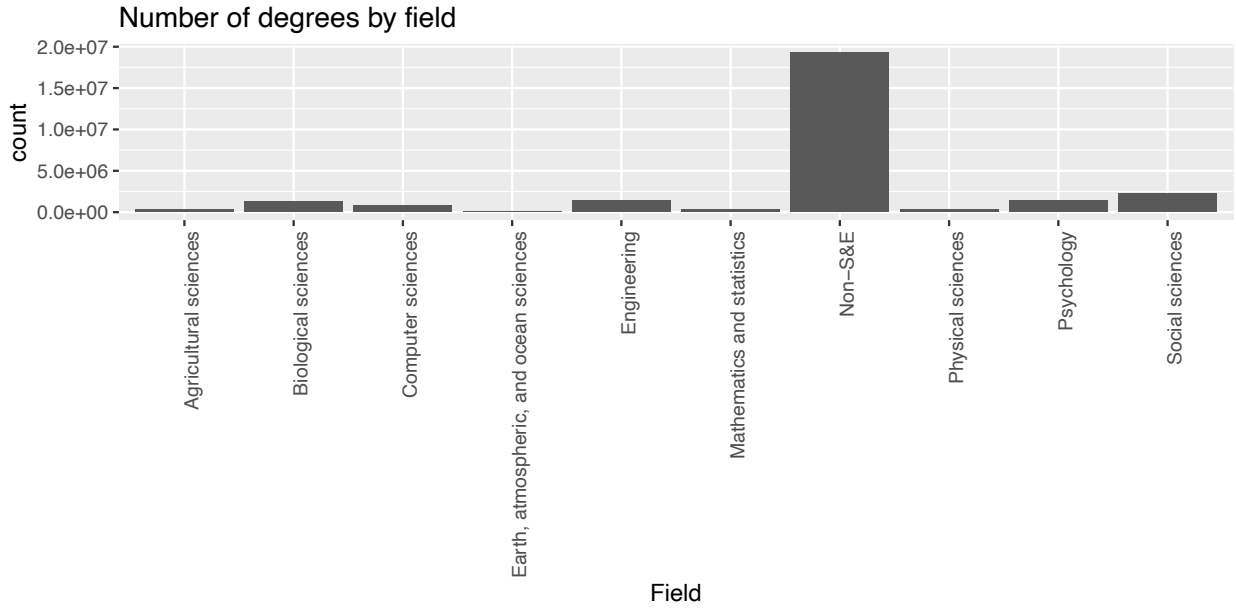


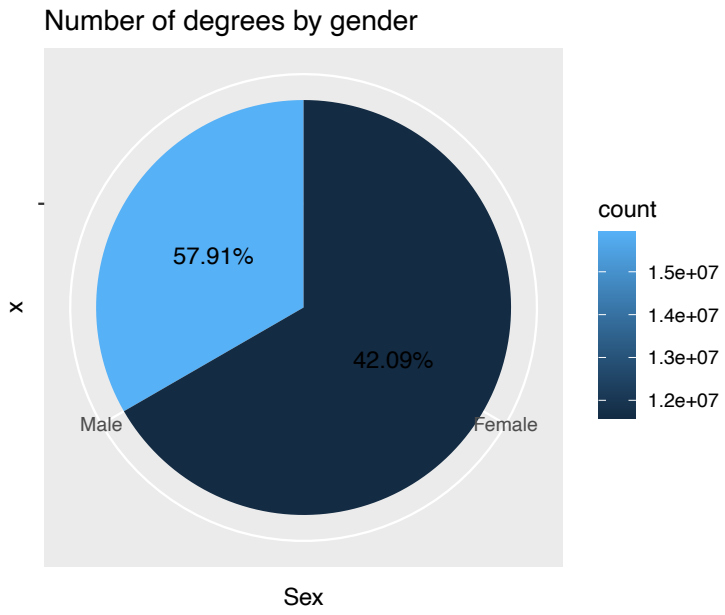
3.4 EDA bring all variables

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?

The number of degrees increases over the years. They spike for Non S&E subjects and they are dominated by a 57.91% women majority.

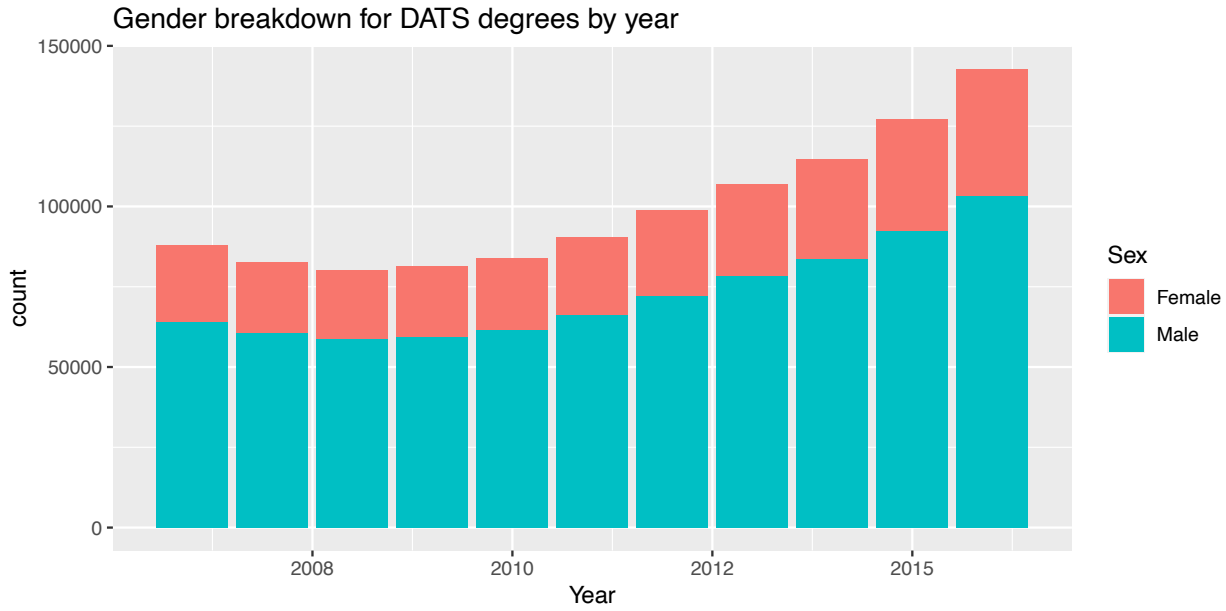

Number of degrees by year
count
2e+06
1e+06
0e+00
2008
2010
2012
2015
Year


Number of degrees by field
count
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Agricultural sciences
Biological sciences
Computer sciences
Earth, atmospheric, and ocean sciences
Engineering
Mathematics and statistics
Non−S&E
Physical sciences
Psychology
Social sciences
Field



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 conclusive evidence that women are underrepresented in data science across many year as shown by the above graph. Every year the number of women studying data science is less than half of men studying data science.

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?

The percentage of women enrolled in Data Science fields is much less than that of for men. Over the years the distribution of women and men taking Data Science and Non Data Science courses stays consistent whilst both the number of women and men grow. We consistently found that Computer Science and Engineering had the largest differences between men and women outside of Non S&E subjects

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 0.Get Started

To get started, we would like to load the two datasets first.

4.2 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.

The log difference is a more appropriate metric to evaluate the increase in payroll each year for different reasons. Firstly, log difference measures the change in payroll as a percentage change. Option 1, on the other hand, only takes the absolute difference between the current year's payroll and the previous year's payroll. However, since the starting payroll varies for each team, log difference is a better option that takes into account that each team begins at a different payroll level. Additionally, using log difference can correct skewness and outliers.

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

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

4.3 Exploratory questions

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

Based on the results, we can see that the Los Angeles Dodgers, Texas Rangers, San Diego Padres, Pittsburgh Pirates, and Washington Nationals experienced the highest increase in payroll between 2010 and 2014. This information can guide us in our investigation to determine which teams made the most improvement during that time period.

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

Based on the results, the Pittsburgh Pirates, Baltimore Orioles, Seattle Mariners, Washington Nationals, and Kansas City Royals showed the most improvement between 2010 and 2014. Only the Pittsburgh Pirates and Washington Nationals from this list also appeared in the top 5 payroll increase list. These results suggest that an increase in payroll might be related to improved team performance, but performance can also be influenced by factors such as team management, team chemistry, playbooks, etc. Further investigation is needed to determine the extent to which payroll impacts team performance.

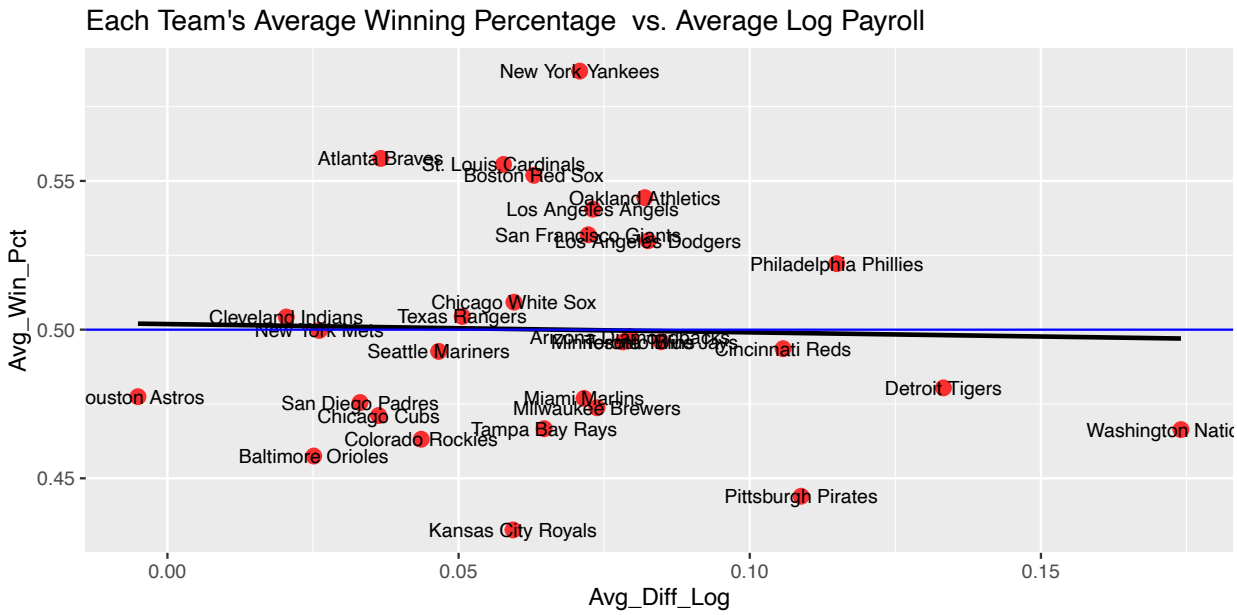
4.4 Do log increases in payroll imply better performance?

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 begin by conducting a correlation test between the winning percentage and the log increase in payroll for each team each year:

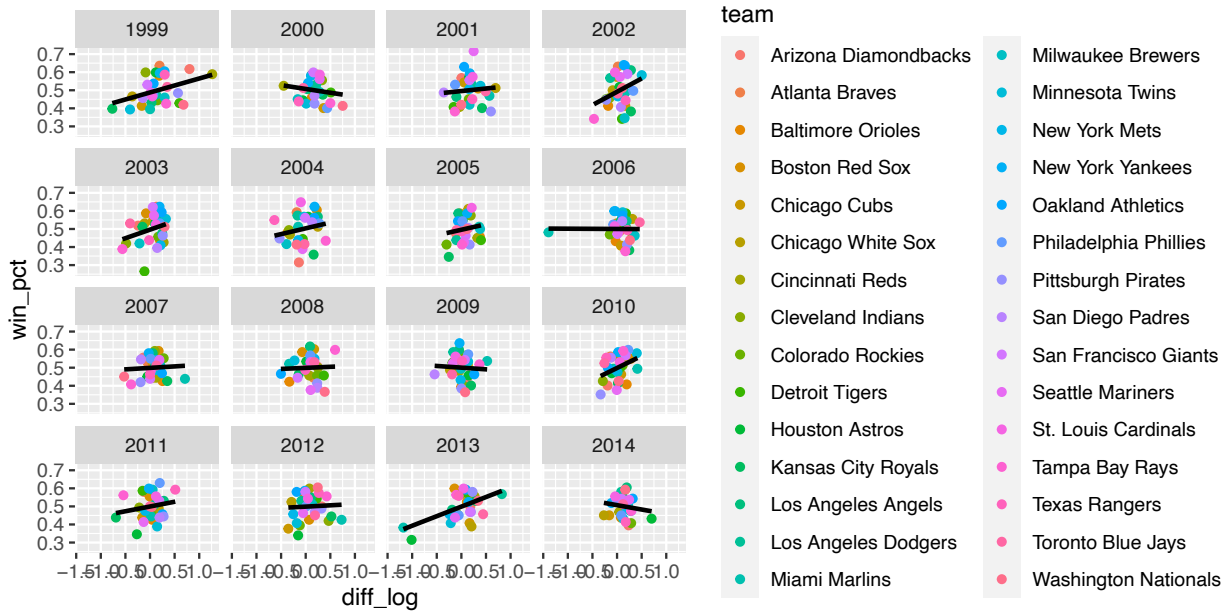
The results indicate that there is a weak positive correlation (0.167) between the log difference in payroll and the winning percentage. The p-value of the Pearson's product-moment correlation test is 2e-04, which means that we fail to reject the null hypothesis that the true correlation is equal to zero. This suggests that there is some relationship between the log difference in payroll and the winning percentage, but it is not very strong. Further investigation is necessary to better understand this relationship.



In the above analysis, a line of best fit was constructed based on the average log payroll increase and the average winning percentage. The results show that while there is a positive relationship between the two variables, it is quite weak. This is demonstrated by outliers such as the Washington Nationals, whose high payroll difference did not correspond with a significant increase in their average winning percentage.

To further explore this relationship, a regression test was performed to assess the impact of the log payroll difference on the average winning percentage. This will provide more information about the relationship between the two variables and allow for a more thorough investigation into the factors that influence the winning percentage of a team.

The results of the regression test further support the conclusion that, despite the weak correlation, there is a significant impact of the log payroll difference on the winning percentage. This makes sense as teams that invest more in acquiring top players are likely to perform better on the field. However, it's important to note that baseball is a team sport and success on the field is determined by many factors beyond just payroll. Further analysis is necessary to fully understand the impact of payroll changes on winning percentage and to identify other factors that may influence the relationship.

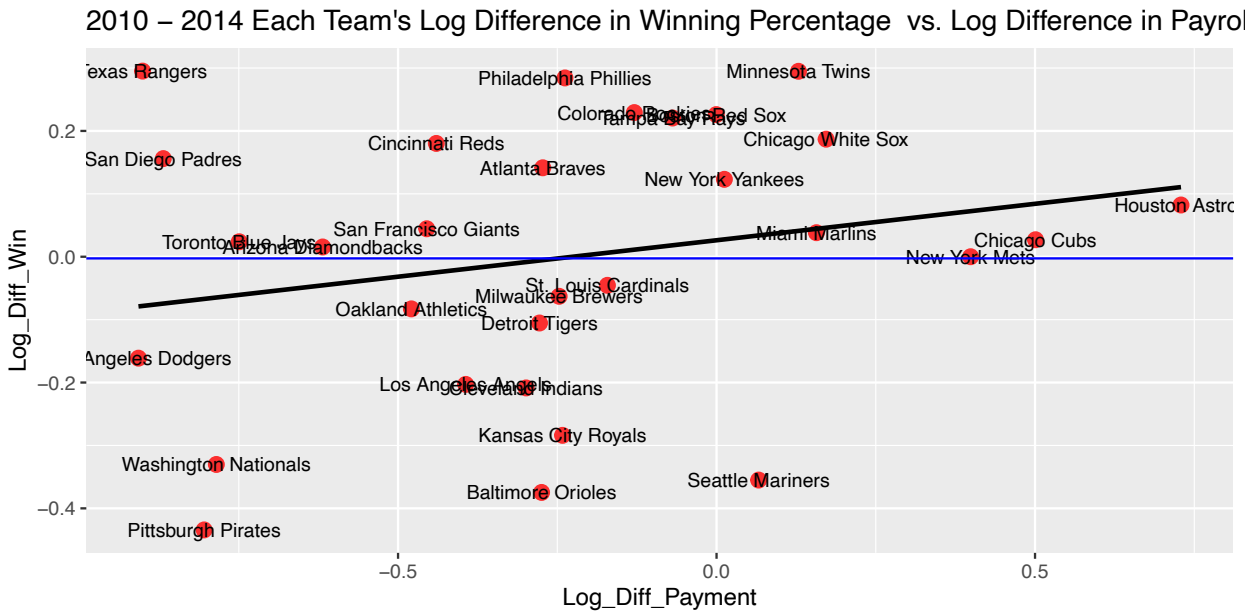


After plotting the relationship between the log difference in payroll and the winning percentage, it becomes evident that the relationship between these two variables changes from year to year. In some years, there is a positive relationship between the log difference in payroll and the winning percentage, while in other years the relationship is negative.

Therefore, we narrow our analysis to the years 2010 to 2014. By conducting a similar analysis on this time frame, we aim to gain a clearer understanding of the impact of the log difference in payroll on the winning percentage during these years.

The correlation between the log difference in payroll and the log difference in winning between the years 2010 and 2014 is 0.227, which is stronger than the overall correlation of 0.167. This indicates that during the years 2010-2014, the impact of payroll on winning percentage was greater compared to the overall relationship.

However, the correlation coefficient of 0.227 represents a weak positive correlation, meaning there is a slight positive relationship between the two variables during this specific time frame. This suggests that payroll may have had a slightly bigger impact on a team's success during the years 2010-2014, though other factors should also be considered.



The plot confirms our previous conclusion that there are multiple factors that can impact a team's performance in addition to an increase in payroll. The plot also highlights the fact that the impact of these factors can vary from year to year.

Going forward, it would be beneficial to delve further into the other factors that influence a team's success. This will allow us to gain a more comprehensive understanding of what drives a team's performance. Additionally, it may be useful to compare the benefits of evaluating a team's performance using either the log of the yearly payroll increase or the yearly payroll amount to determine which method is more effective.

4.5 Comparison

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

Based on the statistical analysis, using both yearly payroll and yearly increase in payroll show that they have a positive impact on team performance and are statistically significant. However, the regression model using yearly increase in payroll has a slightly higher R-squared, indicating it better explains team performance.

Yearly increase in payroll is a better metric to evaluate a team's performance compared to yearly payroll because it takes into account the effects of monetary inflation. Additionally, every team has a different situation and reacts differently to changes in payroll. A team with a lot of rookies would benefit more from hiring star players, top coaches, and so on compared to a championship team. Hence, the yearly increase in payroll is a more accurate metric to evaluate a team's performance.