Modern Data Mining, HW 1

Sarah Hayward Annie Vo Jessica Brown

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

Contents

1 Overview

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

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

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

1.1 Objectives

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

1.2 Instructions

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

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

1.3 Review materials

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

2 Case study 1: Audience Size

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

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

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

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

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

```
survey_data = read_csv('Survey_results_final.csv')
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
## Rows: 1763 Columns: 35
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (25): HITId, HITTypeId, Title, Description, Keywords, Reward, CreationTi...
## dbl  (4): MaxAssignments, AssignmentDurationInSeconds, AutoApprovalDelayInSe...
## lgl  (6): NumberOfSimilarHITs, LifetimeInSeconds, RejectionTime, RequesterFe...
```

```
##
## 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.
```

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".

```
#get selected columns and change datatypes of numeric columnss
cleaned_survey_data = survey_data %>% select("Answer.Age", "Answer.Gender", "Answer.Education", "Answer
```

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

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 over thinking. Keep it simple.

```
#get unique values for each column
lapply(cleaned_survey_data, unique)
```

Answer: We found that there were 37 rows with at least one missing value or unspecified answer (including those with both). For most columns, this correlated to an NA value, except for education there are "select one" responses which likely indicate a drop down option where the respondent did not answer the question. For the gender question in particular, it could be possible that the NAs were people who don't identify as male or female, something the survey did not account for; however, we also want to avoid making this assumption as it's very possible they just didn't answer the question. For the numeric columns, the NAs may also include entries which contained characters which would be turned into an NA value when we made the column numeric. To be standardized, for this small sample of surveys not completely filled out, we opted to remove them to preserve the integrity of the data.

```
print(nrow(cleaned_survey_data[xor(rowSums(is.na(cleaned_survey_data)) > 0, cleaned_survey_data$educati
```

```
#drop missing values
final_survey_data = cleaned_survey_data[rowSums(is.na(cleaned_survey_data)) == 0 & cleaned_survey_data$
```

From the initial data, we can also see that there are a few questionable/wrong age values, namely 4 and 223 which were likely filled in incorrectly. We can see that these values consist of only 2 rows, one for each value, after we removed the missing data. The data also affirms that the row for the 4-year-old is incorrect given that there's no way a toddler could complete a bachelor's or 4-year degree.

```
final_survey_data %>% filter(age == 4 | age == 223)
```

We could reasonably assume that the 223 person meant 22 or 23 and that the 4-year-old is likely in their 40s or some age that ends in 4. Given the uncertainty and the very low number of rows, we will also drop these rows.

```
final_survey_data = final_survey_data %>% filter(age != 4 & age != 223)
```

Overall, we ended up removing only 39 rows from our dataset. Our overall sample is big enough that we still have enough datapoints to work with after removing these points, but it's also small enough that incorrect estimations for the missing or wrongly entered questions could make a tangible impact on our results.

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.)

Report: We can first determine the sample size of our data and see that we are considering 1,725 survey results.

```
cat("Sample size: ", nrow(final_survey_data))
```

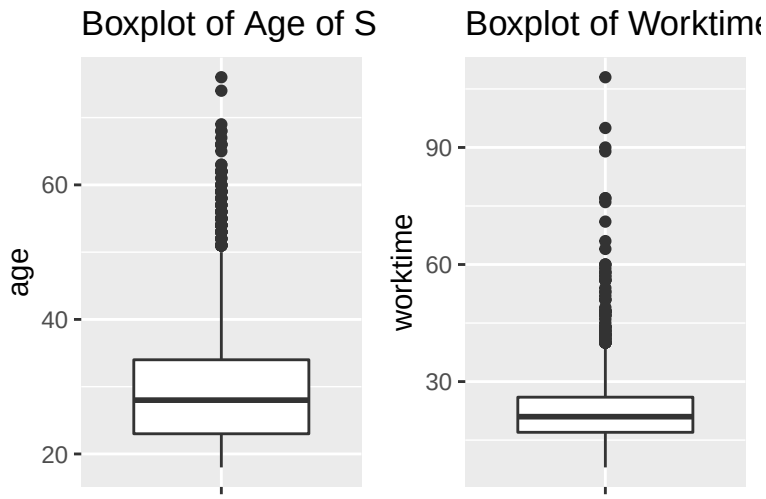
We can then delve into the other summary statistics of the data. While most of our columns are categorical data, we can get distribution information about age and worktime, such as the average age of the respondents which is 30.3 and the average worktime being 22.5. We can also see the ranges we're dealing with, with the youngest repondant being only 18 years-old and the oldest is 76. The time used to complete the survey ranges from just 8 seconds to 108 seconds.

```
summary(final_survey_data)
```

We can now make graphical representations of each of our columns to demonstrate the demographics and distribution of the responses.

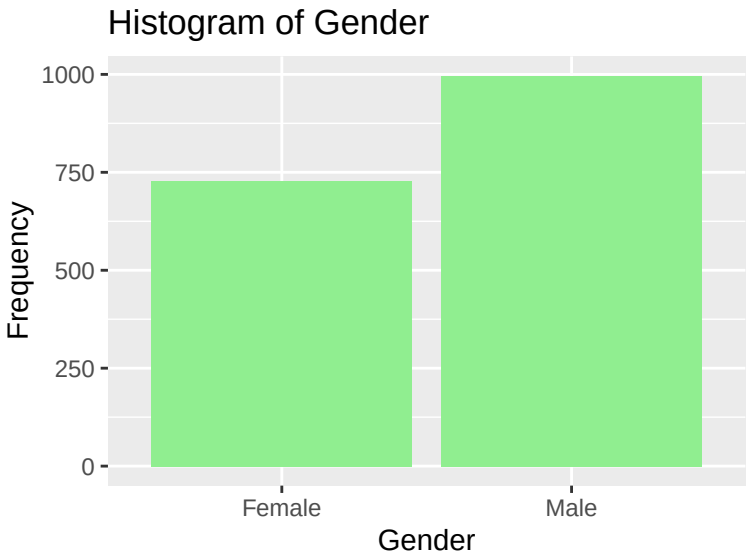
For age and worktime, it makes sense to make boxplots to visualize the distributions of the numerical columns given in the summary.

```
box_p1 <- ggplot(final_survey_data) +
  geom_boxplot(aes(x = "", y = age)) +
  labs(title = "Boxplot of Age of Survey Respondents", x = "")

box_p2 <- ggplot(final_survey_data) +
  geom_boxplot(aes(x = "", y = worktime)) +
  labs(title = "Boxplot of Worktime of Survey Repondents", x = "")

grid.arrange(box_p1, box_p2, ncol = 2)
```



Unlike age and worktime, gender is a categorical variable and thus is better represented using a histogram instead of a boxplot. We can see that there are almost 33% more men which took the survey than women, potentially meaning there are more male survey workers (although it's far from conclusive).

```
ggplot(final_survey_data) +
  geom_bar(aes(x = gender), fill = "light green") +
  labs( title = "Histogram of Gender", x = "Gender" , y = "Frequency")
```



For education and income, both discrete categorical variables, it makes sense to also create histograms to evaluate the distribution of the data. We can see that our data most heavily consists of 4-year college graduates and (likely) many current college students as well as those who did not finish college or those with an associate's degree. Our data is more evenly distributed by income group, with the least amount of respondents in the highest income level ($> \$150{,}000$), which makes sense given they are the population least incentivized by a job that pays 10 cents per survey.

```r
hist_p1 <- ggplot(final_survey_data) +
  geom_bar(aes(x = education), fill = "blue") +
  labs( title = "Histogram of Education", x = "Education Level" , y = "Frequency") +
  scale_x_discrete(labels = function(x) str_wrap(x, width = 10))

hist_p2 <- ggplot(final_survey_data) +
  geom_bar(aes(x = income), fill = "light blue") +
  labs( title = "Histogram of Income", x = "Income" , y = "Frequency")

grid.arrange(hist_p1, hist_p2, nrow = 2)
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>
```

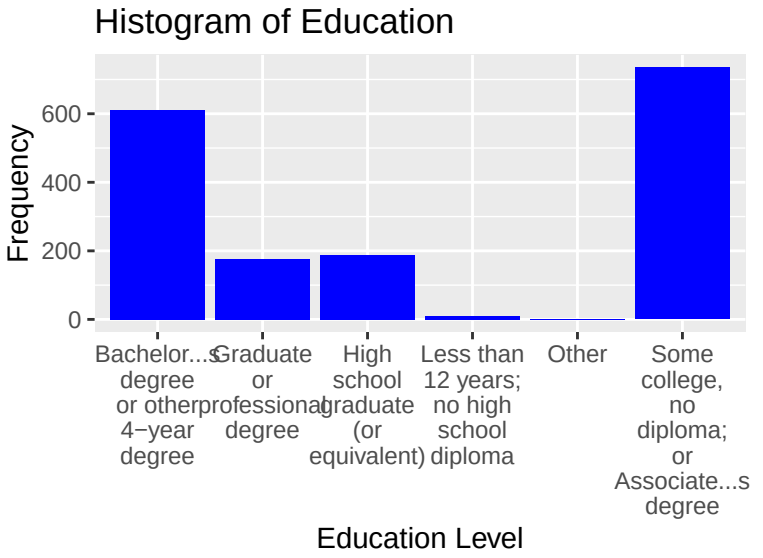
```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>
```

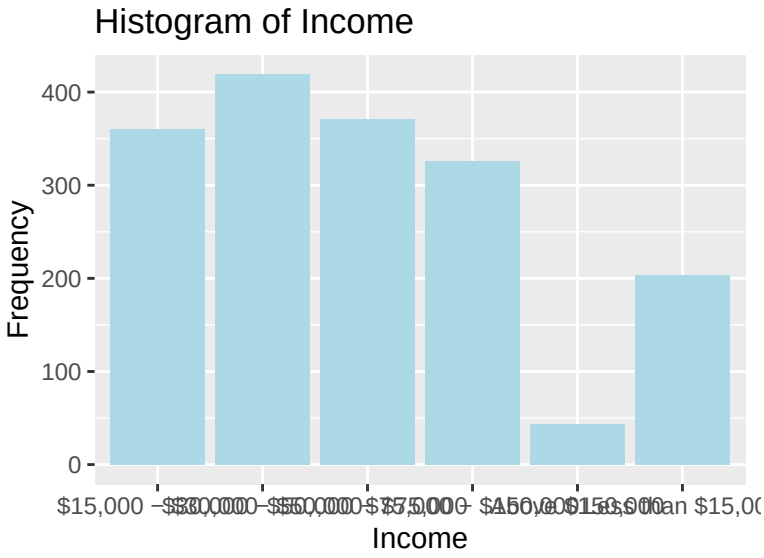
```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Bachelor's' in 'mbcsToSbcs': dot substituted for <99>
```

```
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Associate's' in 'mbcsToSbcs': dot substituted for <99>
```

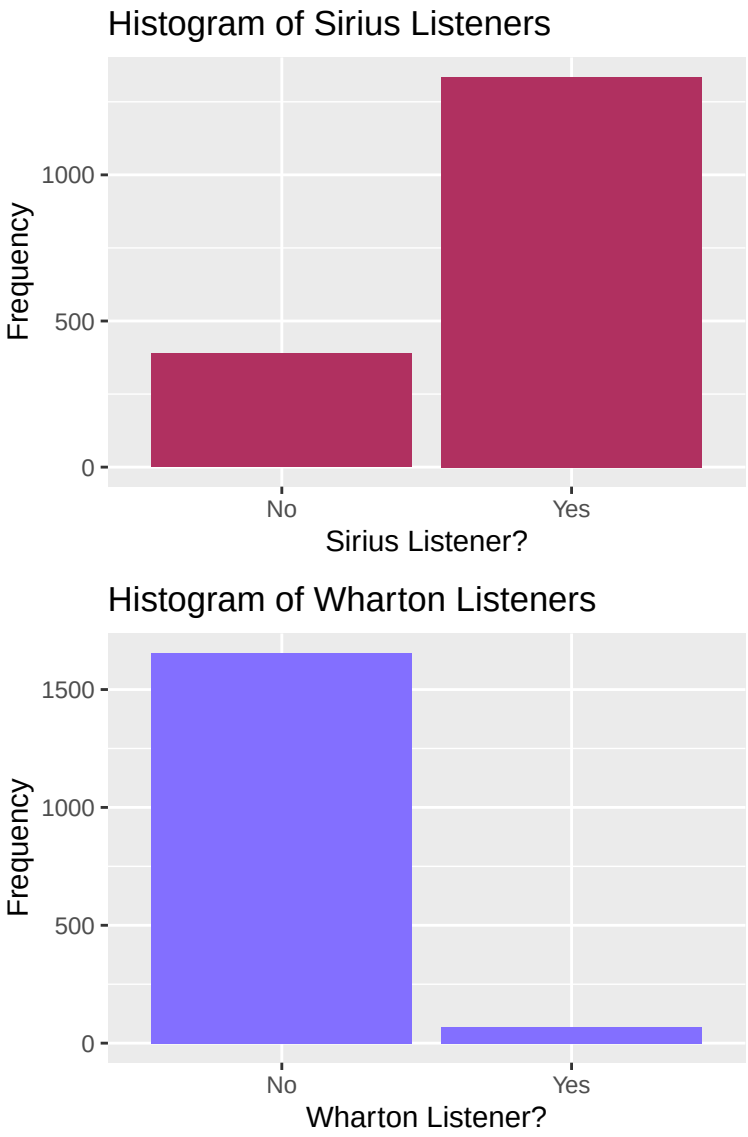




Lastly, we can represent the listening habits of histograms as well given that they're also discrete categorical variables. From the graphs, it's evident that although there are plenty of Sirius listeners in the survey workers, there are few people which have listen Sirius Business Radio by Wharton.

```
hist_p3 <- ggplot(final_survey_data) +
  geom_bar(aes(x = sirius), fill = "maroon") +
  labs( title = "Histogram of Sirius Listeners", x = "Sirius Listener?" , y = "Frequency")
```

```
hist_p4 <- ggplot(final_survey_data) +
  geom_bar(aes(x = wharton), fill = "slateblue1") +
  labs( title = "Histogram of Wharton Listeners", x = "Wharton Listener?" , y = "Frequency")

grid.arrange(hist_p3, hist_p4, nrow = 2)
```



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?

In terms of gender, this sample does not appear to be a random sample from the general population of the US given the disproportionate number of males compared to females given that in the US there are actually more women than men (161.7 million vs 156.7 million in 2014). The income distributions are more realistic of the US population given that the median income levels in our age group are around $30,000-50,000 (https://dqydj.com/average-median-top-income-by-age-percentiles/) which is not only the mode of our distribution, but the ranges around it are also highly populated. When comparing our education data to the distribution in the US (https://www.statista.com/statistics/785618/educational-attainment-by-age-group-us/), we can see that our sample is more educated given the larger percentage of those with at least a bachelor's degree and disproportionally small number of people with less than a high school diploma (<1% instead of around 10%) and those with only a high school diploma (11% instead of around 26-33%).

```
#get distribution of education data
final_survey_data %>% count(education) %>% mutate(n, n / sum(n))
```

This sample does not seem to be a representative random sample of the general population, as the sample is skewed younger and within that age group, it is more male heavy and more educated.

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

When considering the general MTURK population, we again see that our sample has a disproportionate number of men instead of women given that across each age group, there are actually more female participants on MTURK. However, it is important to note that the older MTURK workers are most heavily concentrated with females and our data consists of mostly younger people. This sample does seem to be representative age-wise of the relatively young MTURK population, as most data entries are for people under 40. As pointed out on the website from which we pull this MTURK data (https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/), the distribution of income is similar to that of the US population, which we have already noted is similarly represented in our sample. For education, it does seem like our data is similar to the distribution of education levels among US MTURK workers (https://crowdsourcing-class.org/readings/downloads/platform/demographics-of-mturk.pdf). It does seem like this sample is a representative random sample from the MTURK population outside of the gender distribution.

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

2.3 Final estimate

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

To be specific, you should include:

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

Before writing the executive summary, we need to first estimate the Wharton audience using the proportion p of Wharton Sirius listeners from our dataset.

```r
#get proportion p of wharton listener out of the sirius listeners
prop_table = final_survey_data %>% filter(sirius == "Yes") %>% count(sirius, wharton) %>% mutate(n, p =
prop_table
```

```r
cat("Our estimation of the audiance size is: ", 51600000 * prop_table$p[1])
```

Executive summary: We conducted a study to measure the success of the Wharton Talk Show on Sirius Radio, that is, estimate the number of listeners. We got our data from a survey we launched on Amazon Mechanical Turk (MTURK) in May 2014 which included basic demographic questions and asked repondents whether they were a Sirius Radio listener and if they have listened to Sirius Business Radio by Wharton. We ended up with a random sample of 1,725 responses from the MTURK population, a population which we assumed shares the same proportion of Wharton listeners vs. Sirius listeners in the general population. Therefore, to estimate the audience size for the Wharton Talk Show, we calculated the proportion (p) of Sirius listeners which have listened to the Wharton show and got about 5%. We then multiplied that p value by 51.6 million, overall number of Sirius listeners to get an estimated approximate audience size of 2,587,725 people. Our study is not without limitations however, given that it is very dependent on our assumption about the consistency of our sample compared to the population of Sirius listeners. There is no evidence to support this and our sample is shown to be somewhat differently distributed than the US population of the same age range. We are also making inferences on a population of 51.6 million listeners based on less than 2000 people, which is a large claim.

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.

Answer: For the new survey, we immediately wanted to account for the fact that Wharton Business Radio is not exclusive to SiriusXM anymore, given that clips are available for listening or download online and it's also available on platforms like Apple Music. For this reason, instead of estimating the proportion p of Sirius listeners that listen to Wharton Business Radio, we want to expand to estimate the proportion of the general US population which have listened to the show. To collect this random sampling of US citizens, we would like to use the platform which brings together a wide variety of characters: Twitter. We would like to create a free survey on Qualtrics that we would link in the Twitter promoted tweet which we would limit geographically to the US and age-wise to those above the age of 16. The survey would have the same questions as the original study from 2014, besides the question on whether the person is a Sirius listener and instead include a drop-down question on where they listened to the show. While this won't help with our prediction, it would support our hypothesis that Wharton Business Radio listeners are not limited to SiriusXM subscribers and improve on the original study.

Twitter gave an estimate that there would be a cost of $0.50 per interaction, and aiming for 2,000 responses, the total cost of data collection would then be $1000. We would also only leave the ad and survey open for 1 month to collect the data and provide an additional month for us to analyze and put together our findings.

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.

```
#install.packages("readxl")
library("readxl")
```

i. Read the data into R.

```
degree_data = read_excel('WomenData_06_16.xlsx')
```

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

```
# Most variable names are already clean
degree_data <- degree_data %>% rename(Field = `Field and sex`, Number = `Degrees Awarded`)
```

iii. Set the variable natures properly.

```
str(degree_data)
summary(degree_data)
# Changing Degree and Sex to factors for data consistency
degree_data[,c('Degree','Sex')] <- lapply(degree_data[,c('Degree','Sex')], as.factor)
# Chaning Year and Number to integers for data consistency
degree_data[,c('Year','Number')] <- lapply(degree_data[,c('Year','Number')], as.integer)
str(degree_data)
summary(degree_data)
```

iv. Any missing values?

```
sum(is.na(degree_data))
```

Answer: There are no missing values.

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

i. How many fields are there in this data?
ii. What are the degree types?
iii. How many year's statistics are being reported here?

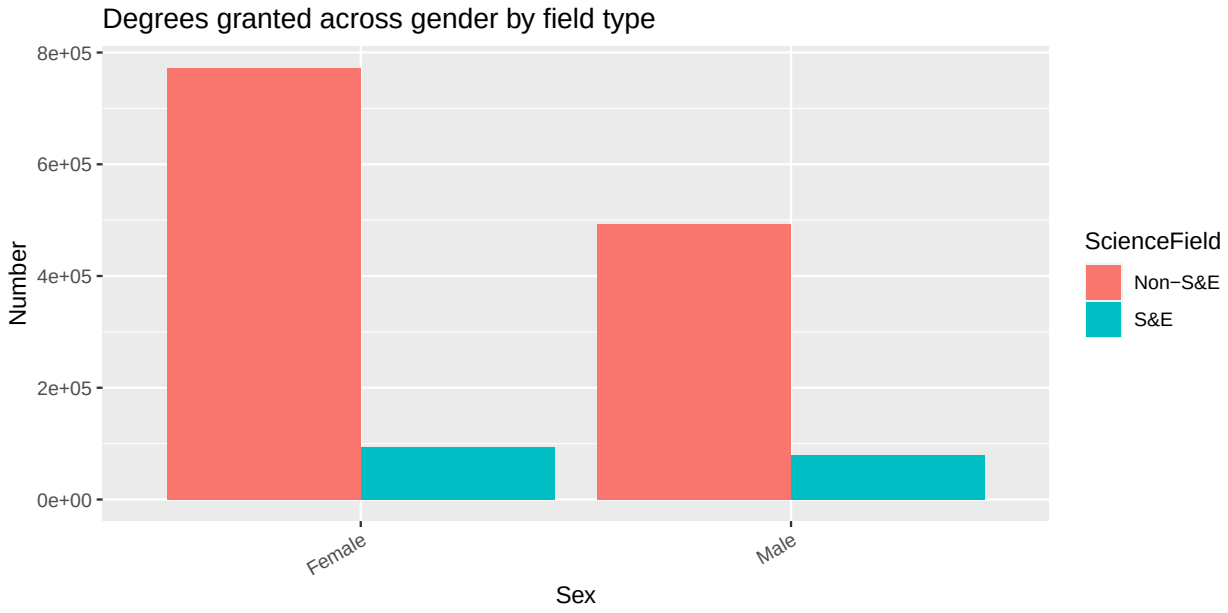
```
summary(degree_data)
knitr::kable(degree_data)
length(unique(degree_data$Field))
```

Answer: In this data, there are 660 different entries and 5 different variables: Field, Degree, Sex, Year, and Number. Within the Field variable, there are 10 distinct fields. The different degree types are BS, MS, and PhD. There are 11 years of statistics reported with the data spanning from 2006-2016. The minimum number of degrees awarded for a single degree type and for a particular sex in a year is 218, while the max is 781474.

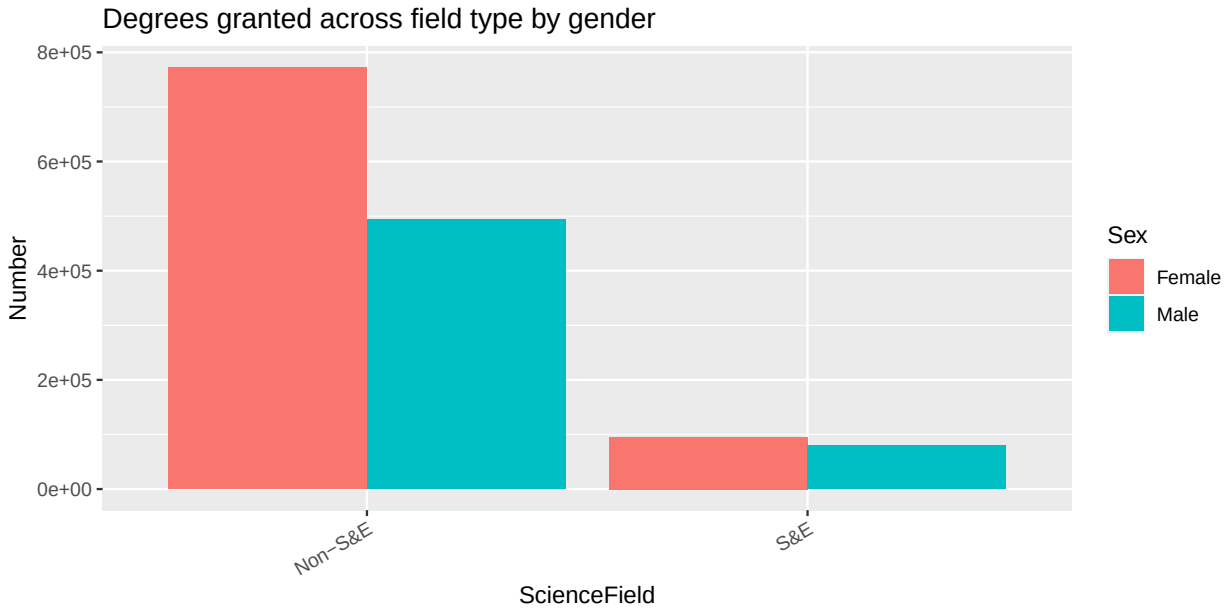
3.2 BS degrees in 2015

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.

```
# Create another variable that codes the field into S&E or Non-S&E
degree_data %<>% mutate(ScienceField = ifelse(Field != "Non-S&E" , "S&E", "Non-S&E"))

sum(degree_data[which(degree_data$Degree == 'BS' & degree_data$Year == 2015 & degree_data$Sex == "Male"

sum(degree_data[which(degree_data$Degree == 'BS' & degree_data$Year == 2015 & degree_data$Sex == "Male"

degree_data %>%
  # Only take data that are BS degrees in 2015
  filter(Degree == 'BS' & Year == 2015) %>%
  ggplot(aes(x = Sex, y = Number, fill = ScienceField)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across gender by field type")
```




```
degree_data %>%
  # Only take data that are BS degrees in 2015
  filter(Degree == 'BS' & Year == 2015) %>%
  ggplot(aes(x = ScienceField, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across field type by gender")
```



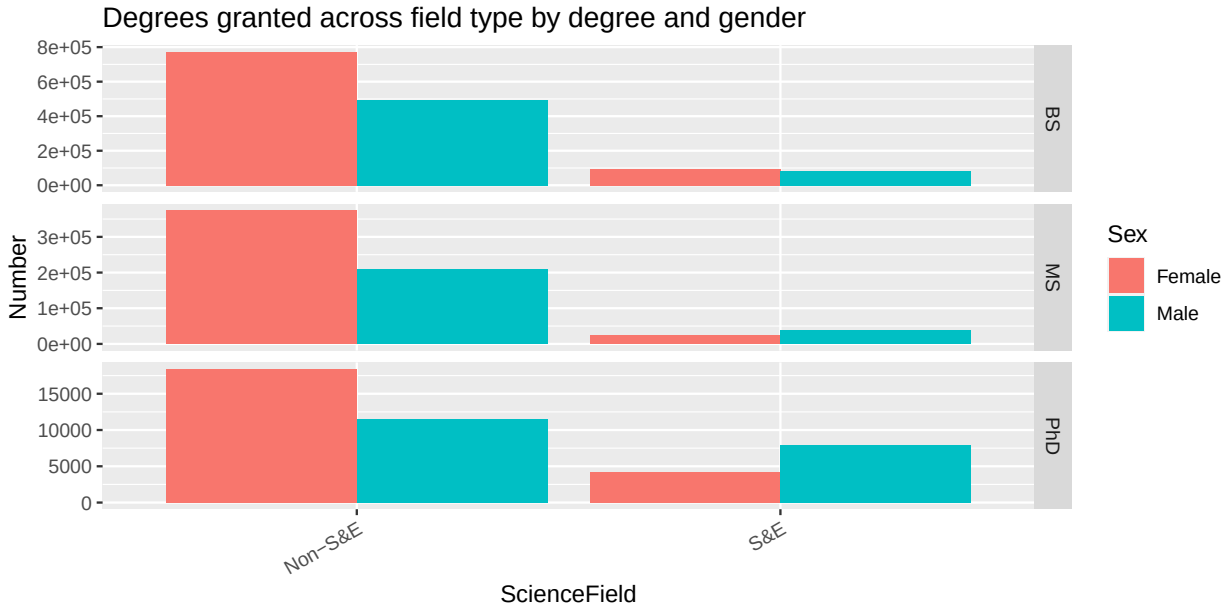
Answer: Only looking at the histogram side for male (not comparing to the female side), there is no evidence that there are more males in science-related fields than males in non-S&E. In fact, there are around 1.5 times as many males in non-S&E than there are in science-related fields. There is a similar trend when looking at the female side. However, there is a greater difference of females in Non-S&E than science-related. This

greater difference is largely due to there being more female in Non-S&E fields than there are male since the number of female and male in science-related fields is roughly equal. Comparatively to the ratio of female in science-related fields vs Non-S&E, there are more male in science-related fields vs Non-S&E.

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

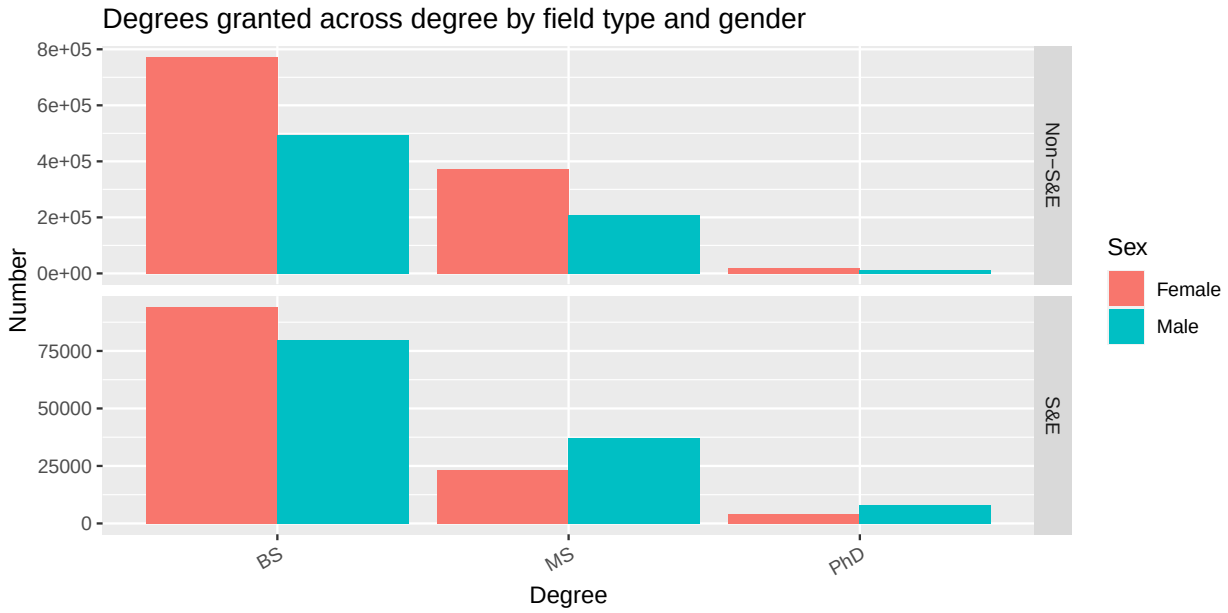
Describe the number of people by type of degree, field, and gender. Do you see any evidence of gender effects over different types of degrees? Again, provide graphs to summarize your findings.

```
degree_data %>%
  # Filter to only 2015
  filter(Year == 2015) %>%
  ggplot(aes(x = ScienceField, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Degree~., scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across field type by degree and gender")
```




```
degree_data %>%
  # Filter to only 2015
  filter(Year == 2015) %>%
  ggplot(aes(x = Sex, y = Number, fill = ScienceField)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Degree~., scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across gender by degree and field type")
```


```
degree_data %>%
  # Filter to only 2015
  filter(Year == 2015) %>%
  ggplot(aes(x = Degree, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(ScienceField~., scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across degree by field type and gender")
```



Answer: Within the Non-S&E fields, there is consistently more degrees awarded to females than males (about double as much) across all types of degrees so there is evidence of a general gender effect over all Non-S&E field degrees that more are awarded to females but no evidence that the gender effect changes based on degree type. However, with science-related fields, the gap between degrees awarded to females and males

widens the higher the degree is. For the BS degree, there is roughly an equal amount of degrees awarded to each sex, with a little more to female than male. However, for the MS degree, there is about double as many degrees awarded to male than female. This difference is replicated in the PhD degree. Therefore, because of the difference of degrees awarded to each sex in MS and PhD degrees but lack of difference in BS degree, there is evidence of gender effects over the MS and PhD degrees that more degrees are awarded to males (about twice as much).

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?

```
degree_data %>%
  group_by(ScienceField, Sex) %>%
  summarise(ScienceField_number = sum(Number)) %>%
  group_by(ScienceField) %>%
  mutate(ratio = ScienceField_number / sum(ScienceField_number))
```

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

```
degree_data %>%
  group_by(ScienceField, Sex, Degree) %>%
  summarise(ScienceField_number = sum(Number)) %>%
  group_by(ScienceField, Degree) %>%
  mutate(ratio = ScienceField_number / sum(ScienceField_number))
```

```
## `summarise()` has grouped output by 'ScienceField', 'Sex'. You can override using the `.groups` argu
```

```
degree_data %>%
  group_by(ScienceField, Sex, Year, Degree) %>%
  summarise(ScienceField_number = sum(Number)) %>%
  ggplot(aes(x = Year, y = ScienceField_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(ScienceField~Degree, scales = "free_y") +
  ggtitle("Degrees granted across time by sex, degree and SE")
```

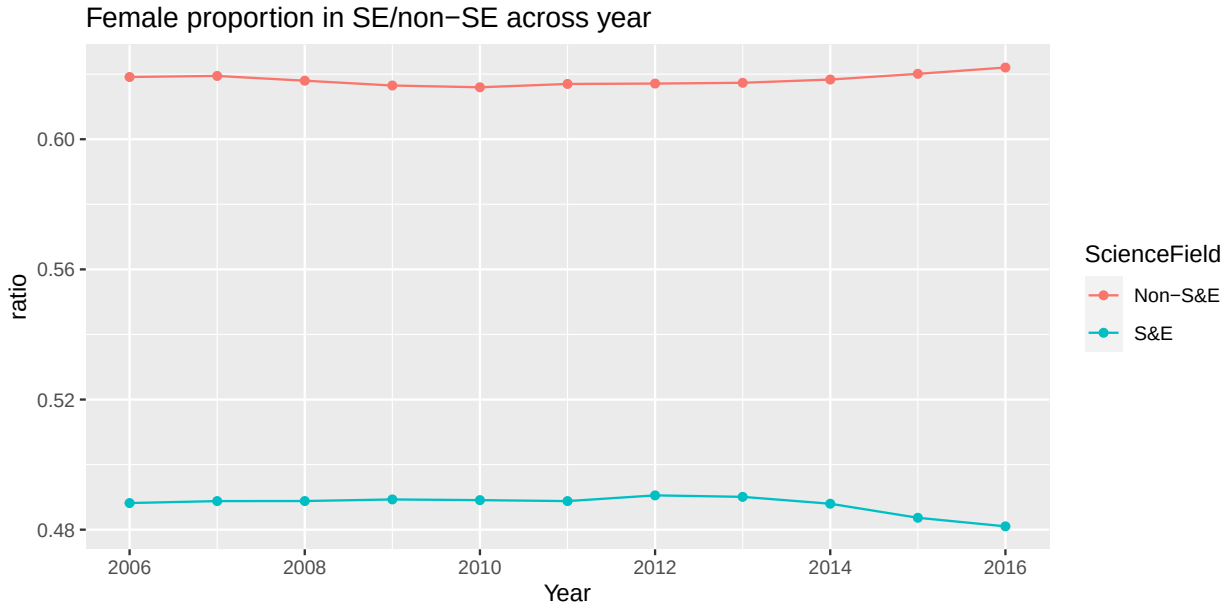
```
## `summarise()` has grouped output by 'ScienceField', 'Sex', 'Year'. You can override using the `.grou
```


```
degree_data %>%
  group_by(ScienceField, Sex, Year, Degree) %>%
  summarise(ScienceField_number = sum(Number)) %>%
  ggplot(aes(x = Year, y = ScienceField_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "fill") +
  facet_grid(ScienceField~Degree, scales = "free_y") +
  ggtitle("Degrees granted proportion by sex across degree and SE")
```

```
## `summarise()` has grouped output by 'ScienceField', 'Sex', 'Year'. You can override using the `.group
```

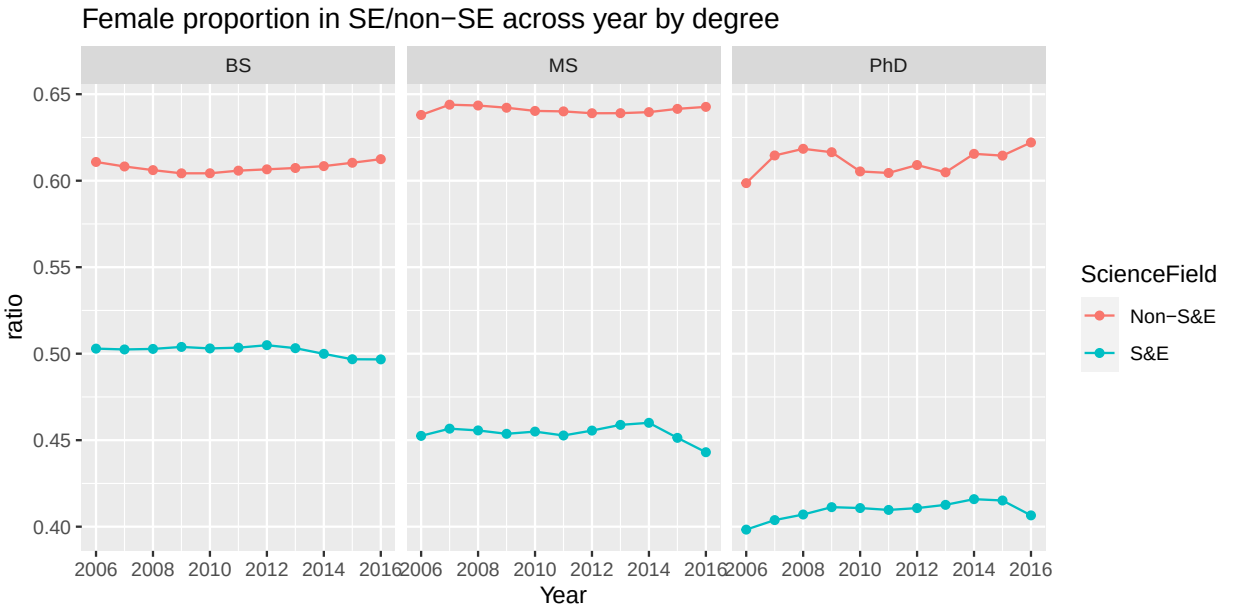
```
degree_data %>%
  group_by(ScienceField, Sex, Year) %>%
  summarise(ScienceField_number = sum(Number)) %>%
  group_by(ScienceField, Year) %>%
  mutate(ratio = ScienceField_number / sum(ScienceField_number)) %>%
  filter(Sex == "Female") %>%
  ggplot(aes(x = Year, y = ratio, color = ScienceField)) +
  geom_point() + geom_line() +
  ggtitle("Female proportion in SE/non-SE across year")
```

```
## `summarise()` has grouped output by 'ScienceField', 'Sex'. You can override using the `.groups` argu
```




```
degree_data %>%
  group_by(ScienceField, Sex, Year, Degree) %>%
  summarise(ScienceField_number = sum(Number)) %>%
  group_by(ScienceField, Year, Degree) %>%
  mutate(ratio = ScienceField_number / sum(ScienceField_number)) %>%
  filter(Sex == "Female") %>%
  ggplot(aes(x = Year, y = ratio, color = ScienceField)) +
  geom_point() + geom_line() +
  facet_grid(~Degree)+
  ggtitle("Female proportion in SE/non-SE across year by degree")
```

```
## `summarise()` has grouped output by 'ScienceField', 'Sex', 'Year'. You can override using the `.grou
```

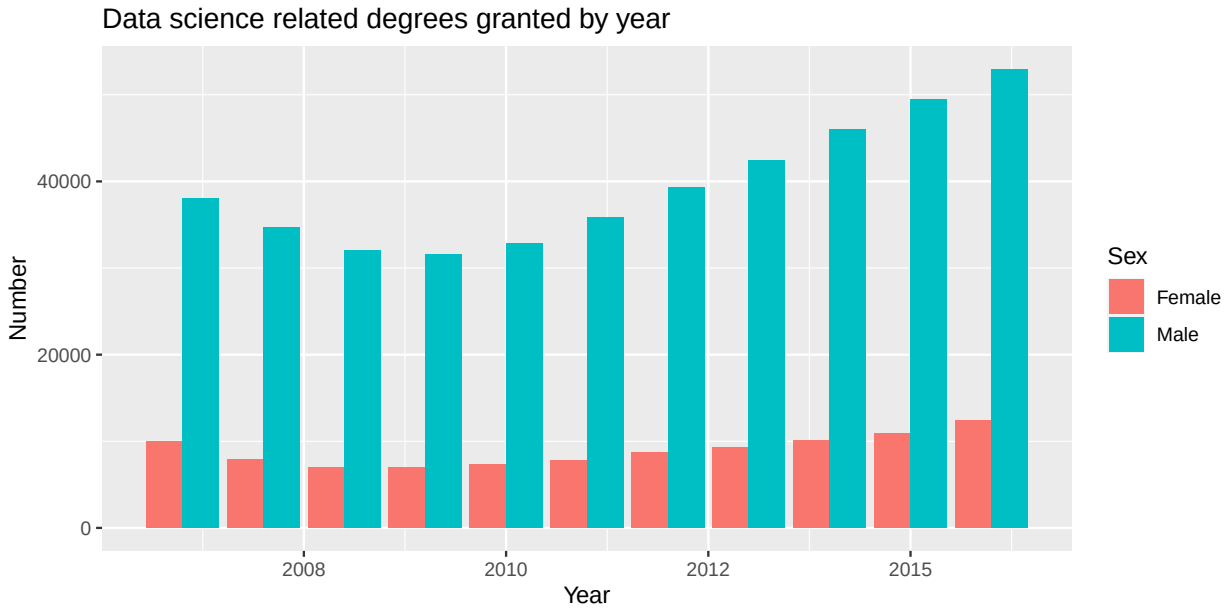


Answer: For Non-S&E fields, there have consistently been about 1.5 as many female as male receiving degrees throughout the years and degree types (with a little less of a difference in PhD than BS). For science-related fields, there have consistently throughout the years been an equal amount of BS degrees awarded, a slightly higher amount of MS degrees awarded for males, and a slightly higher amount than seen in the MS degrees in PhD degrees. When looking at the trend for female proportion across the years for each field type, females make up more than 60% of the proportion in Non-S&E fields throughout all the years, but make up less than 50% of the proportion in science-related fields. In the more recent years (2014-2016), the proportion of females in science-related fields have decreased but have increased for Non-S&E. These trends are more exaggerated when separating the different degree types. Females are least represented throughout the years in science-related fields when looking at PhD degrees (41%), while they are about equally represented in BS degrees (50%) and in between BS and PhD representation for MS degrees (45%). Female proportion in Non-S&E are similar throughout the years for BS and PhD degrees at about 61% (although less consistent for PhD), while proportion in MS degrees is a bit higher at 64%.

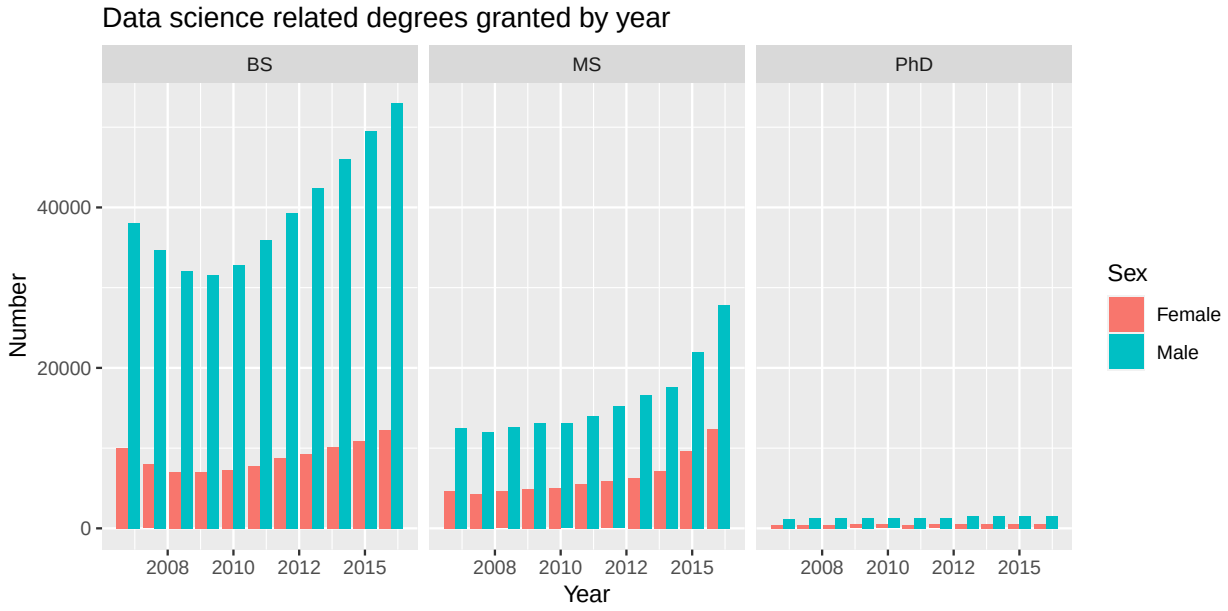
3.5 Women in Data Science

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.

```
# Create another variable that codes the field into DS or Non-DS
degree_data %<>% mutate(DataScience = ifelse(Field != "Mathematics and statistics" & Field != "Computer

degree_data %>%
  filter(DataScience == "DS") %>%
  ggplot(aes(x = Year, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(scales = "free_y") +
  ggtitle("Data science related degrees granted by year")
```




```
degree_data %>%
  filter(DataScience == "DS") %>%
  ggplot(aes(x = Year, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(~Degree, scales = "free_y") +
  ggtitle("Data science related degrees granted by year")
```



Answer: There is evidence of women being underrepresented in data science as there is a large difference in the number of computer science or math and statistics degrees awarded to males compared to females. This difference is about 5 times as much for BS degrees, twice as much for MS degrees, and 4 times as much for PhD degrees. Overall, degrees in computer science or math and statistics are awarded about 4-5 times more to males than females.

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?

Answer: The proportion of females vs males in science-related fields have stayed pretty consistent over time, even when separating my degree type. However, when separating my degree type, the proportion of females vs males in science-related fields change. The higher the degree type, the less females there is proportionately receiving degrees. There is about an equal amount of BS degrees awarded to females as males (50%), then there is less MS degrees proportionately awarded (45%), and even less PhD degrees proportionately awarded (41%). As the degree type increases, there is about a 5% decrease in representation of females proportionately to males. Therefore, there is evidence to suggest that more males pursue science related fields with higher degree types. One concern with the data set is that the report states "Surveys conducted by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation provided a large portion of the data used in this report", we wonder if the surveys were represented of the whole population. We can improve on the study by looking into other ways to be involved in the sciences besides degrees like technical school certificates.

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.

```
baseball <- read_csv("baseball.csv")
```

```
## 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.
```

```
payData <- read_csv("MLPayData_Total.csv")
```

```
## Rows: 30 Columns: 52

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Team.name.2014
## dbl (51): p1998, p1999, p2000, p2001, p2002, p2003, p2004, p2005, p2006, p20...

##
## 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.
```

4.1 EDA: Relationship between payroll changes and performance

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

Create increment in payroll

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

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

```
#let's look at the differences for 2013
payData$p2013 - payData$p2012
log(payData$p2013) - log(payData$p2012)
summary(payData$p2013 - payData$p2012)
summary(log(payData$p2013) - log(payData$p2012))
```

We can see that the values in the first set have a much larger range from -81.7 to 121.5, whereas the second set using the log the values lie closer together and the range is from about -1 to 1.

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

Answer: We want to use the log difference as it allows us to approximate the percent change in the payroll rather than just the difference in pay. By converting the differences to the log scale we can compare the different teams on a more standardized scale. As we saw above, it put all the log payroll differences on a much smaller range from about -1 to 1. This allows us to level our the teams that get more funding to those with less funding.

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

```
#creating the diff_log function
baseball$diff_log <- log(baseball$payroll) - lag(log(baseball$payroll))
#note we set the year 1998 to NA since the row before is a different team, and
#that is the first year that we have data for
baseball$diff_log[baseball$year==1998] <- NA
```

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

```
#creating long table
baseballNew <- baseball[,c("team", "year", "diff_log", "win_pct")]
```

4.2 Exploratory questions

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

```
#payroll differences between 2010 and 2014 (p2014-p2010)
#make a new data set with just 2010 and 2014
tenFourteen <- subset(baseballNew, year == 2010 | year == 2014)
tenFourteen$fourYearDiff <- tenFourteen$diff_log - lag(tenFourteen$diff_log)
i <- seq(1, 60, 2)
#we set odd numbers to NA since we don't need the differences for different teams
tenFourteen$fourYearDiff[i] <- NA
j <- order(tenFourteen$fourYearDiff, decreasing = TRUE)[1:5]
tenFourteen[j,c("team", "fourYearDiff")]
```

Answer: The teams with the highest increase in payroll between 2010 and 2014 inclusive are 1. Houston Astros 0.812
2. Oakland Athletics 0.506
3. Arizona Diamondbacks 0.427
4. San Diego Padres 0.418
5. Texas Rangers 0.393

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

```
#determining biggest percent gain in wins
tenFourteen$pctChange <- tenFourteen$win_pct - lag(tenFourteen$win_pct)
i <- seq(1, 60, 2)
#we set odd numbers to NA since we don't need the differences for different teams
tenFourteen$pctChange[i] <- NA
j <- order(tenFourteen$pctChange, decreasing = TRUE)
tenFourteen[j,c("team", "pctChange")]
```

Answer: The team with the biggest percent change, or that improved the most from 2010 to 2014 was the Pittsburgh Pirates. Their percent change was 0.19136.The next closest team was the Baltimore Orioles with a win percent change of 0.18519 followed by the Washington Nationals at 0.16667.

4.3 Do log increases in payroll imply better performance?

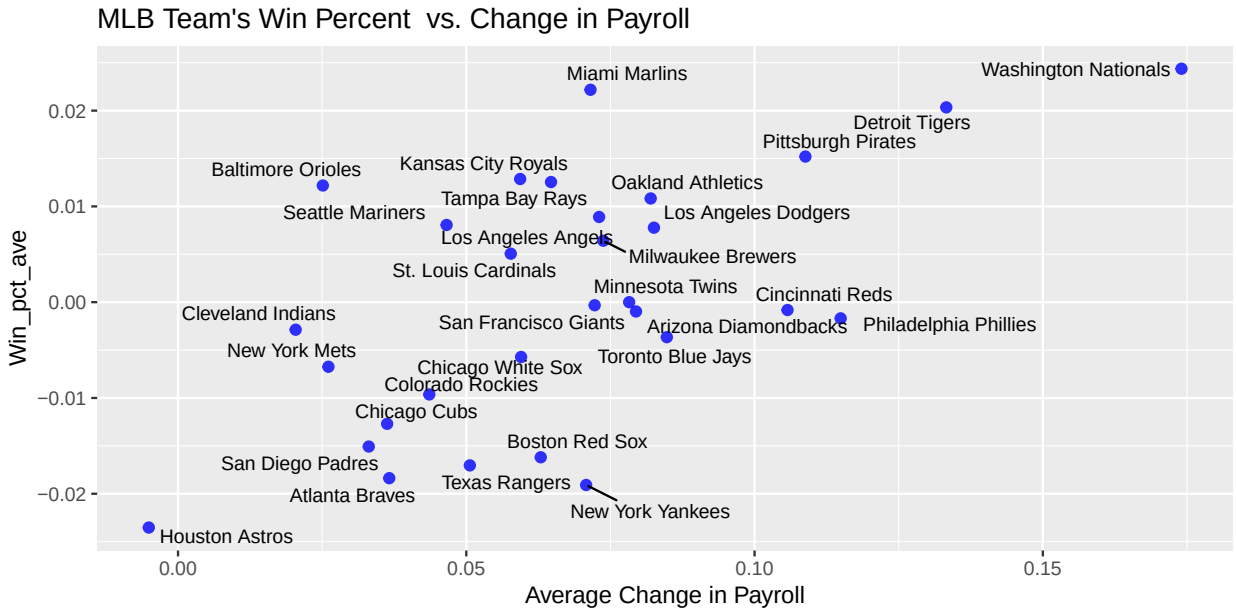
Is there evidence to support the hypothesis that higher increases in payroll on the log scale lead to increased performance? Pick up a few statistics, accompanied with some data visualization, to support your answer.

```
#creating the win_pct_change function
baseball$win_pct_change <- log(baseball$win_pct) - lag(log(baseball$win_pct))
#note we set the year 1998 to NA since the row before is a different team, and
#that is the first year that we have data for
baseball$win_pct_change[baseball$year==1998] <- NA
```

```
# create average change in winning percentage and diff_log for each team
data_agg <-baseball %>%
  group_by(team) %>%
  summarise(
    diff_log_avg = mean(diff_log, na.rm = TRUE),
    win_pct_ave = mean(win_pct_change, na.rm = TRUE))
str(data_agg)
summary(data_agg)
```

Above we created a new data set that found the average difference payroll on the log scale for the team covering all the differences between 1998-2014 as well as the average win percentage change for all of those years. This allows us to look at the overall performance of the teams in just two variables.

```
# now let's look at the averages on a plot and see if there is some relationship
#install.packages("ggrepel")
library(ggrepel)
data_agg %>%
  ggplot(aes(x = diff_log_avg, y = win_pct_ave)) +
  geom_point(color = "blue", size= 2, alpha = .8) +
  geom_text_repel(aes(label = team), size = 3) +
  labs(title = "MLB Team's Win Percent  vs. Change in Payroll",
       x = "Average Change in Payroll",
       y = "Win_pct_ave")
```

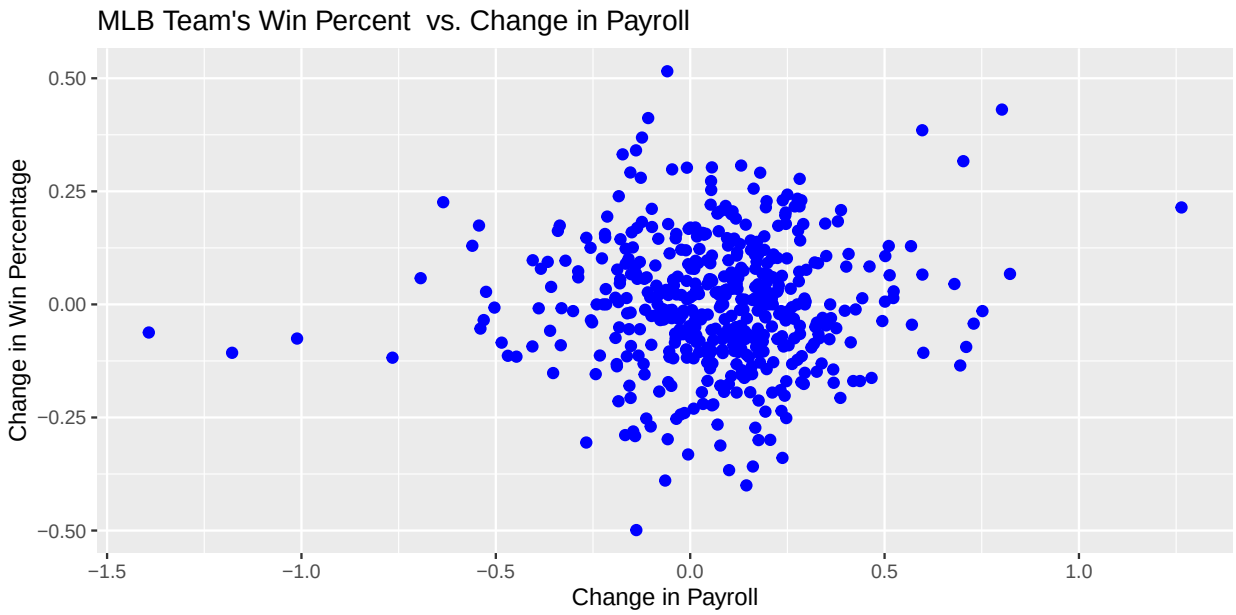



```
cor(data_agg$diff_log_avg, data_agg$win_pct_ave)
```

Answer: Based on this graph of the average change in payroll, (average of diff_log for each team)and the win percent average change over the years 1998-2014, there does appear to be a moderate positive correlation between the variables. The correlation between the variables diff_log_avg and win_pct_avg is 0.592.That means that an increase in the difference in payroll would be correlate to a increase in the average win percentage change for the teams.

However, this is looking at the performance of each team averaged over the years and it could be different if we compare each year to the increased performance and thus have more data points. So below we will look at all the teams and all of the changes between the years.

```
# we remove the NA functions to make it easier to plot the graphs and compute the correlation
baseballSubset <- subset(baseball, !is.na(diff_log) & !is.na(win_pct_change))

baseballSubset %>%
  ggplot(aes(x = diff_log, y = win_pct_change)) +
  geom_point(color = "blue", size= 2) +
  labs(title = "MLB Team's Win Percent  vs. Change in Payroll",
       x = "Change in Payroll",
       y = "Change in Win Percentage")
```



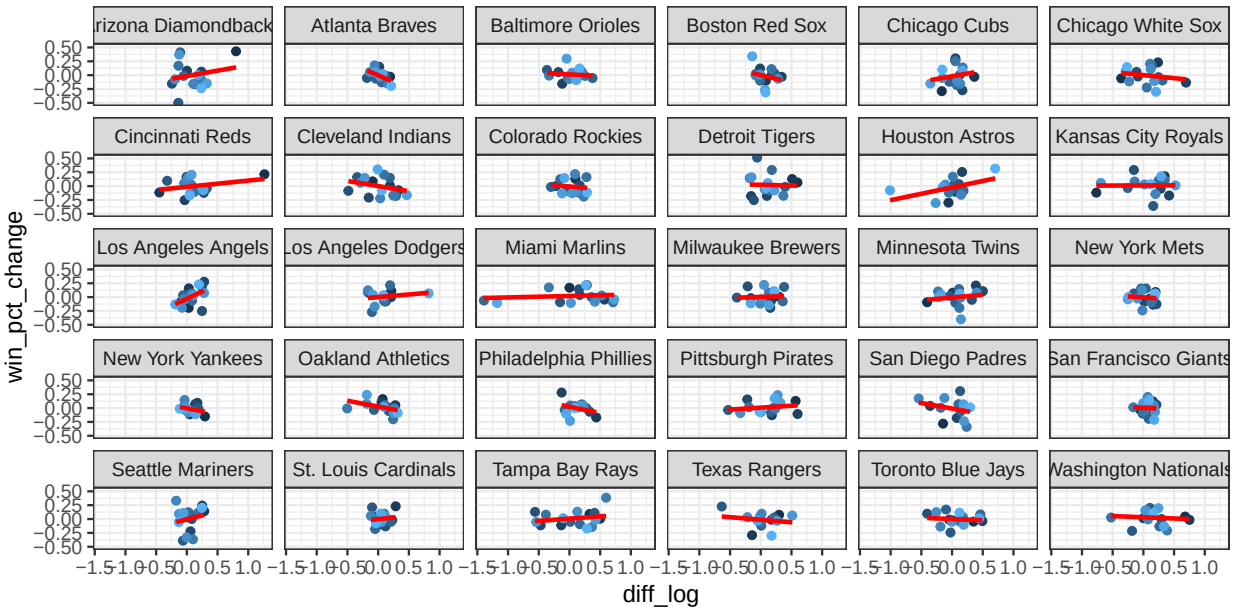

```
cor(baseballSubset$diff_log, baseballSubset$win_pct_change)
```

Answer: Now from looking at this plot, there appears to be essentially not correlation between changes in payroll on the log scale and the change in win percentage for teams. This is further confirmed by a correlation value of 0.0391 which is essentially zero. A correlation value of zero indicates no correlation at all.

As further evidence, we can also use the graphs from the prior question to see that the top 5 teams with the greater percent increase in payroll between 2010 and 2014 do not line up with the top 5 teams that had the biggest percent change in wins from those years

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

```
## Warning: Removed 30 rows containing non-finite values (stat_smooth).

## Warning: Removed 30 rows containing missing values (geom_point).
```



Answer: From the plots above, we are able to look at the changes in payroll on the log scale and the win percentage change for team. We can see that there is no clear positive correlation for all of the teams Some have a least squares line that is flat, some have positive, and have negative. This means that the correlation for each of the teams can be different, for some teams an increase is payroll on the log scale could lead to an increase of percent change in wins but for other teams it could lead to a decrease. Because of this we can say that an increase in payroll for some individual teams doesn't lead to increased performance. In fact as we saw, when we look at all the data for the teams combined, the correlation and scatter plot indicates no correlation between increased payroll and increased performance. Thus we can conclude that increased payroll does not lead to increased performance.

4.4 Comparison

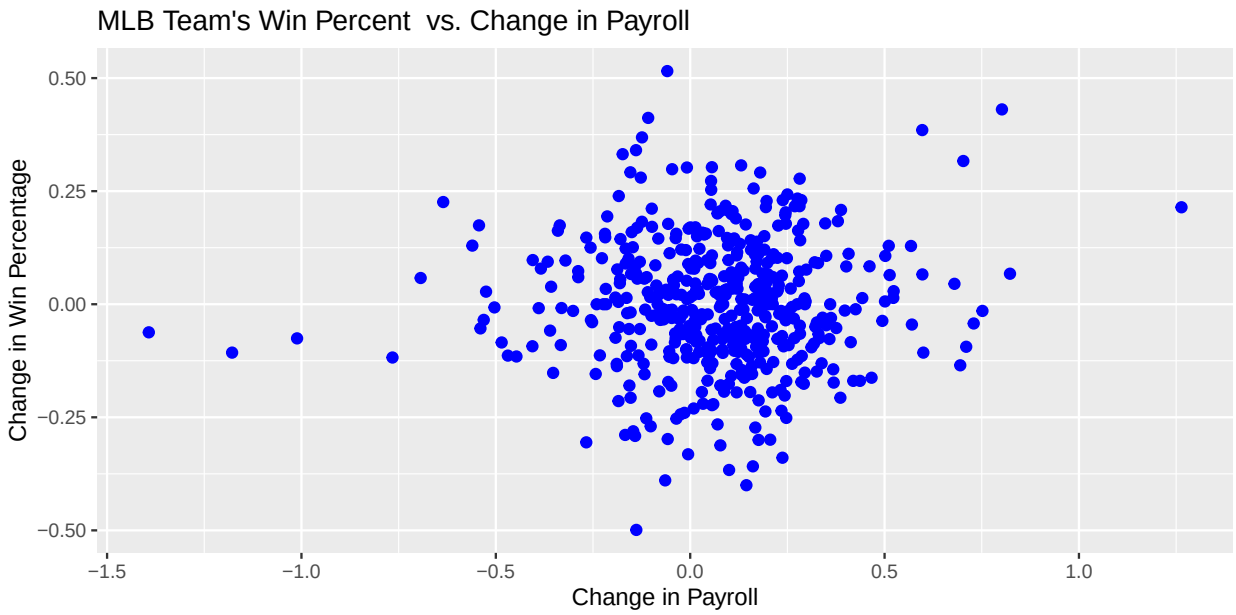
Which set of factors are better explaining performance? Yearly payroll or yearly increase in payroll? What criterion is being used?

Answer: Yearly Payroll

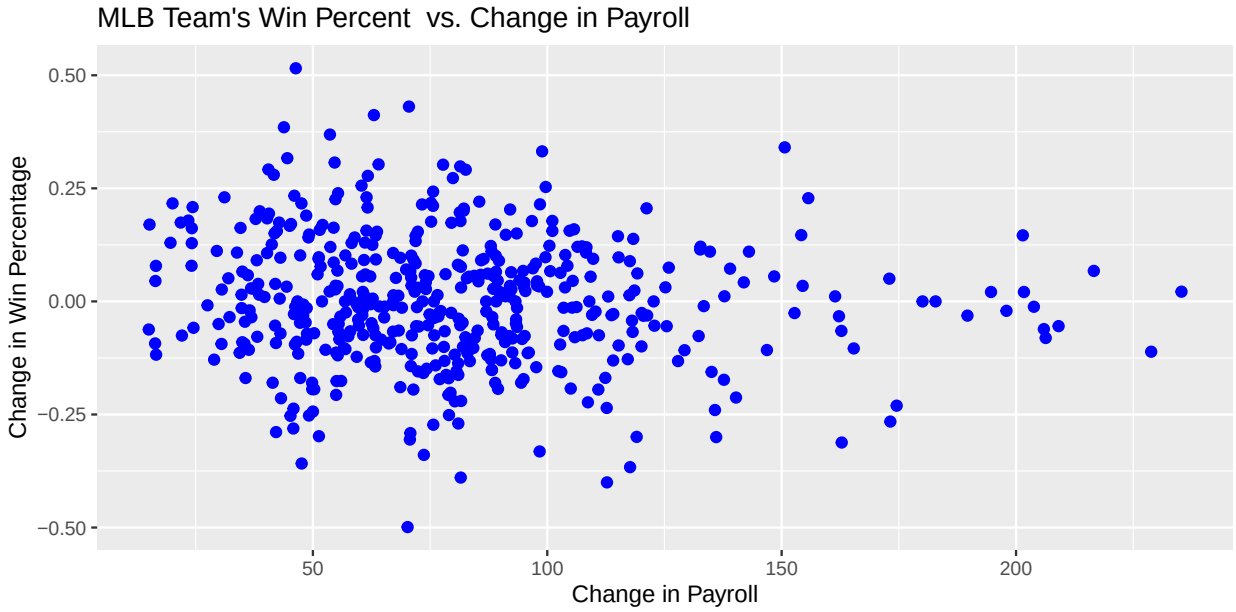
```
cor(baseballSubset$payroll, baseballSubset$win_pct_change)
cor(baseballSubset$diff_log, baseballSubset$win_pct_change)
```

We can see that Yearly payroll has a slightly higher correlation when looking at the absolute value with win percentage than the yearly increase in payroll. We can also look at these plots below to see if there is a visual correlation as well.

```
baseballSubset %>%
  ggplot(aes(x = diff_log, y = win_pct_change)) +
  geom_point(color = "blue", size= 2) +
  labs(title = "MLB Team's Win Percent  vs. Change in Payroll",
       x = "Change in Payroll",
       y = "Change in Win Percentage")
```




```
baseballSubset %>%
  ggplot(aes(x = payroll, y = win_pct_change)) +
  geom_point(color = "blue", size= 2) +
  labs(title = "MLB Team's Win Percent  vs. Change in Payroll",
       x = "Change in Payroll",
       y = "Change in Win Percentage")
```



Although neither plot indicates a strong correlation between the two variable, you can see there is a weak positive correlation between Yearly Payroll and Win percentage change compared to the other scatter plot of change in payroll and win percentage that does not indicate any correlation. Thus although both yearly payroll and change in year payroll on the log scale have low correlations with values of-0.109 and 0.0391 respectively, if I had to choose one to describe performance, it would be Yearly payroll. Notice that the correlation being negative hear means that for the entire MLB, a higher payroll weakly correlated to a

decreased in win percent change between the years.