Modern Data Mining, Case2 Ying Dai

Group Member Ying Dai

Due: 11:59PM, February 7, 2021

Contents

0.1

0.1.1 Case study 2 women in science

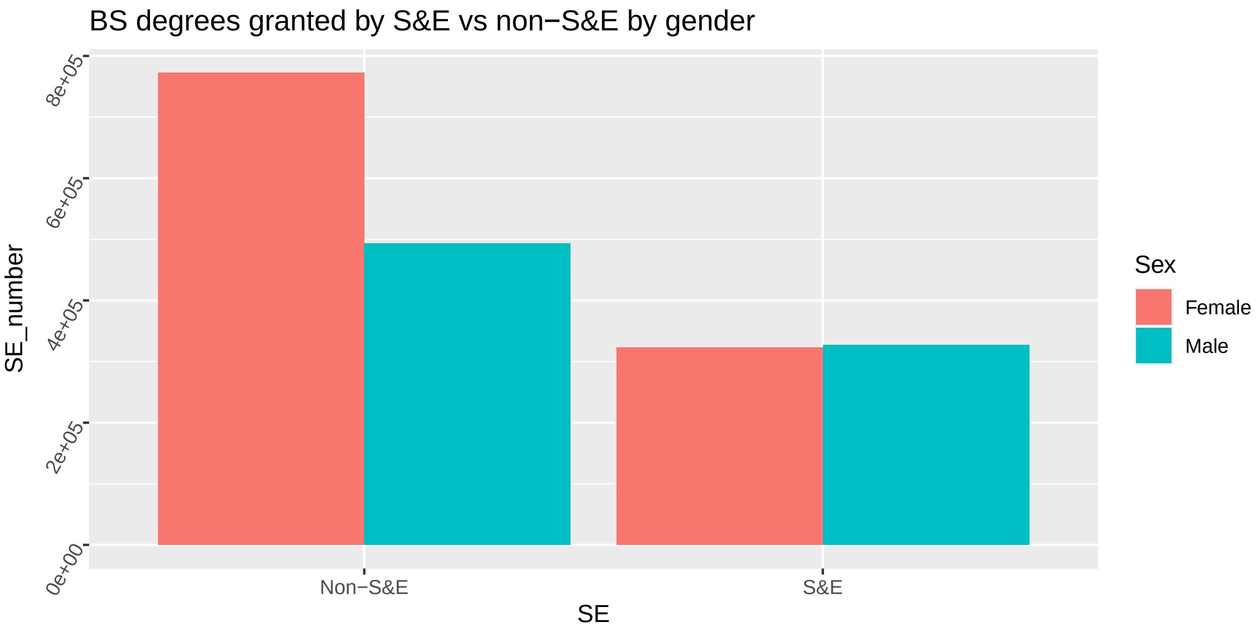
0.2 3.1 Data preparation

```
# 1. Understand and clean the data
# i. read the data
setwd("/Users/yingdai/Desktop/UPenn/Plan of study/STAT571/Homework/Home work 1/data")
Casestudy2 <- read_excel("WomenData_06_16.xlsx")

# ii. rename variables and set natures properly
Casestudy2 %<>%rename(Field = "Field and sex",
          Number = "Degrees Awarded") %>%
   mutate(Field = as.factor(Field),
          Degree = as.factor(Degree),
          Sex = as.factor(Sex),
          Year = as.factor(Year))

str(Casestudy2)

# iii. look for missing values & 2. Brief description of the dataset
skimr::skim(Casestudy2)
```

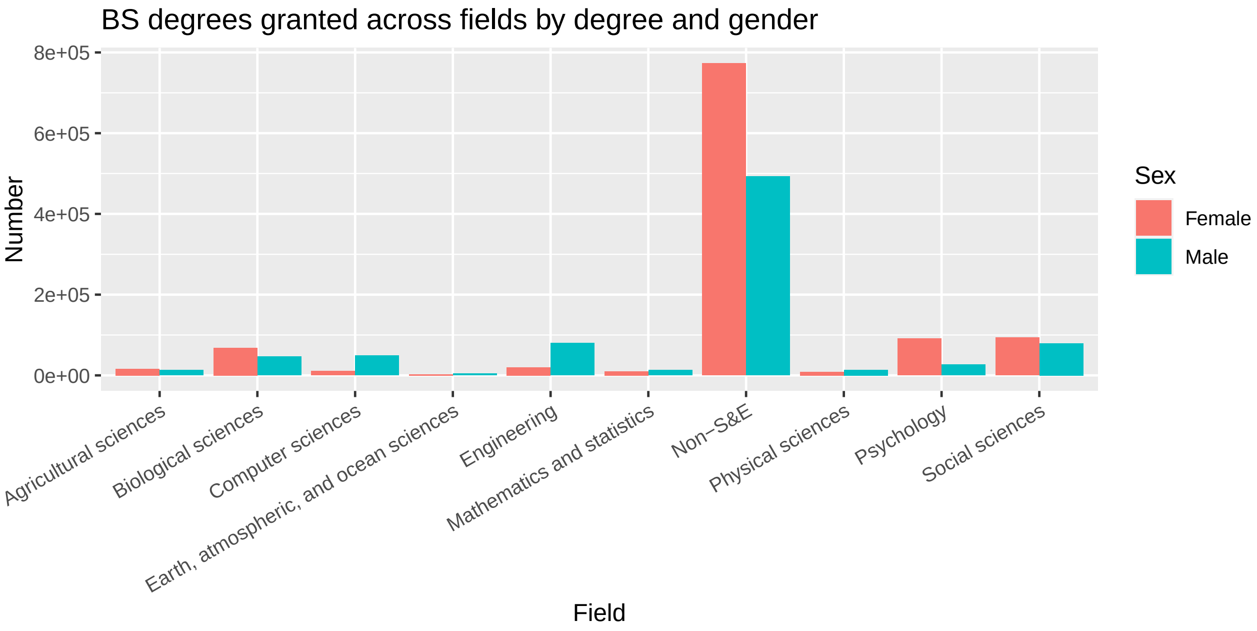
Brief description of the dataset: There is no missing values in this dataset. There are ten fields, three types of degrees, and 11 year's statistics in this dataset.

0.3 3.2 BS degrees in 2015

```
# BS degrees in 2015
# 1. Plot of Male vs female in SE and Non-SE field
Casestudy2 %>%
  filter(Year == 2015, Degree == "BS") %>%
  mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(SE, Sex) %>%
  summarise(SE_number = sum(Number)) %>%
  ggplot(aes(x = SE, y = SE_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.y = element_text(angle = 60)) +
  ggtitle("BS degrees granted by S&E vs non-S&E by gender")
```



```
# 2. Plot of male vs female in each field
Casestudy2 %>%
  filter(Year == 2015, Degree == "BS") %>%
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("BS degrees granted across fields by degree and gender")
```

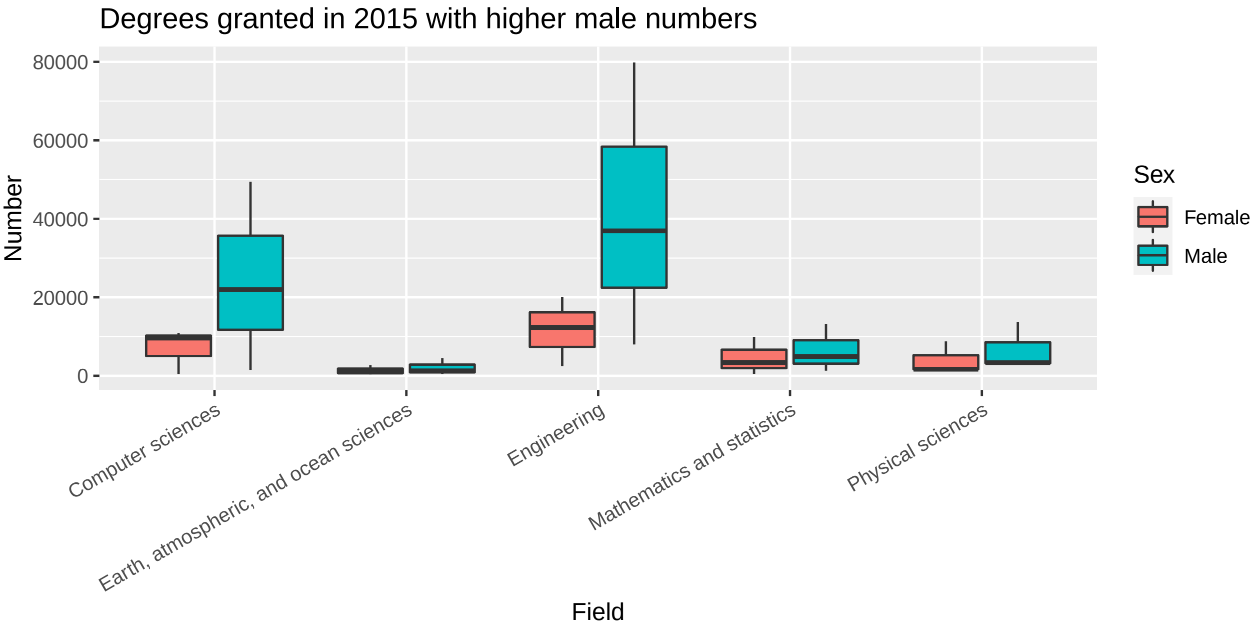


```
# 3. Table summary of number of male and female in each field in year 2015
Casestudy2 %>%
  filter(Year == 2015, Degree == "BS") %>%
  group_by(Field, Sex) %>%
  pivot_wider(names_from = Sex, values_from = "Number") %>%
  summarise(Male = sum(Male),
            Female = sum(Female))
```

From the above first plot we can see that in the year 2015, the total number of males who own a BS degree in science and engineering (SE) field is slightly higher than the number of women. However, when looking at the total number of BS degrees in each field (show in the second plot and the above table), not all fields have a higher male number. Specifically, five SE fields have a higher male total number, while the remaining four SE fields have a higher female total number.

Take a closer look at SE fields with higher male numbers

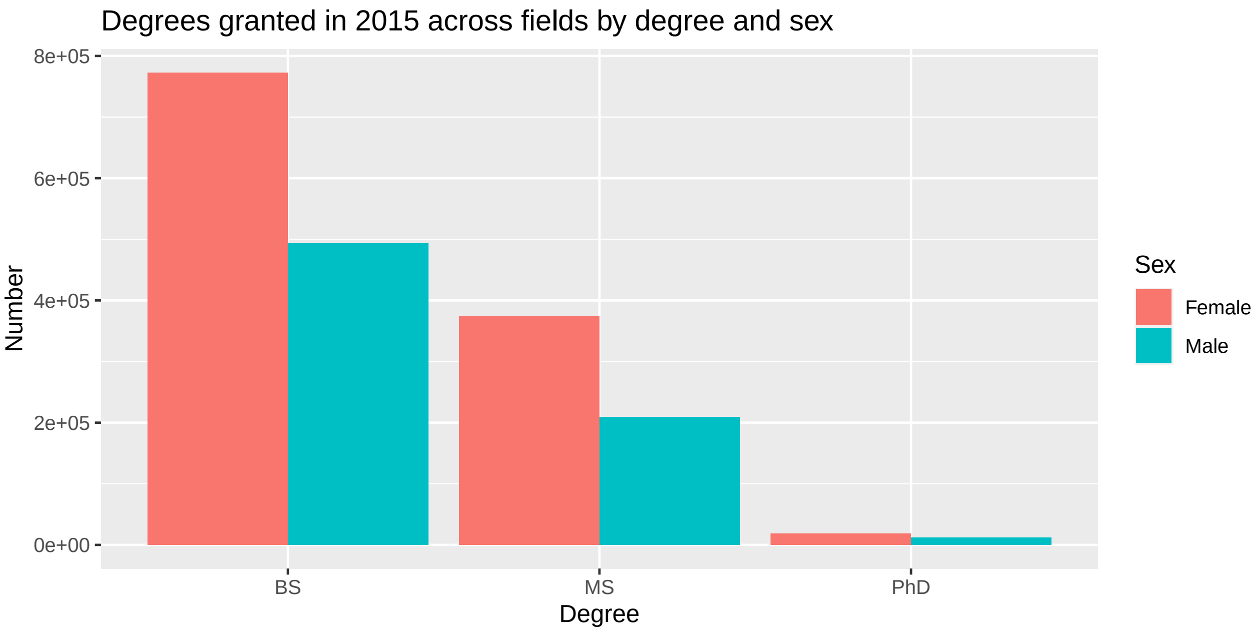
```
Casestudy2 %>% filter(Year == 2015, Field %in%
                        c("Computer sciences", "Mathematics and statistics", "Earth, atmospheric, and oc
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_boxplot() +
  ylab("Number") +
  ggtitle("Degrees granted in 2015 with higher male numbers") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  theme()
```



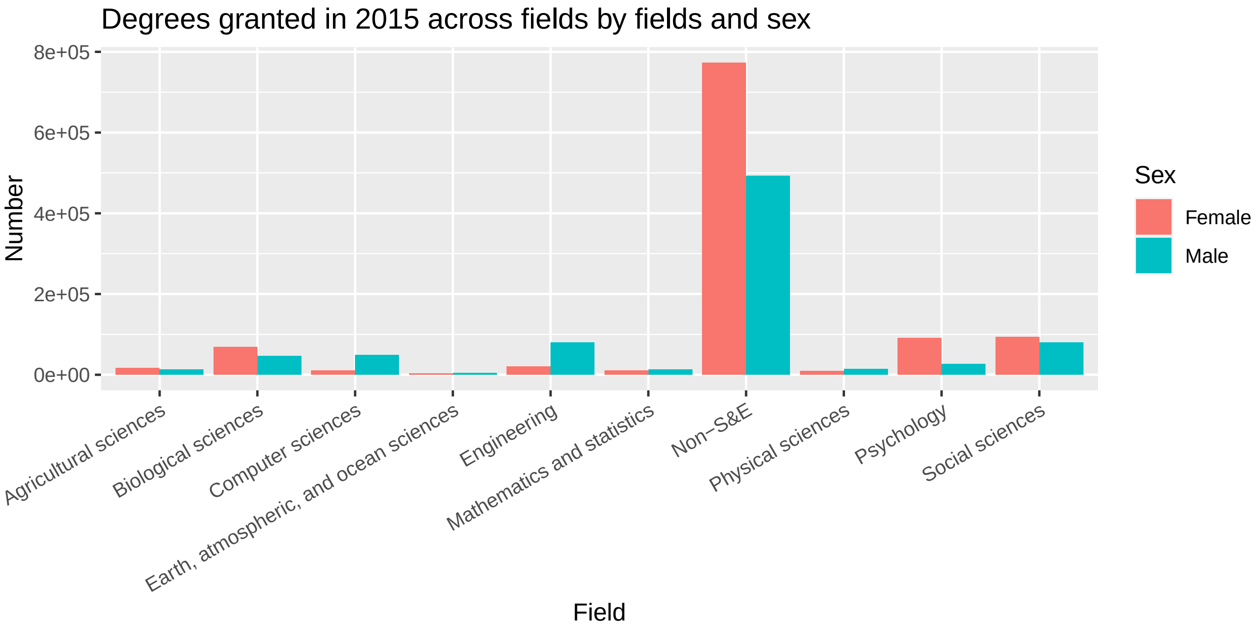
The above boxplot shows that among SE fields with a higher male number in the year 2015, engineering field has the highest difference between male and female number, followed by computer science.

0.4 3.3 Describe number of people by type of degree, field, and sex in 2015

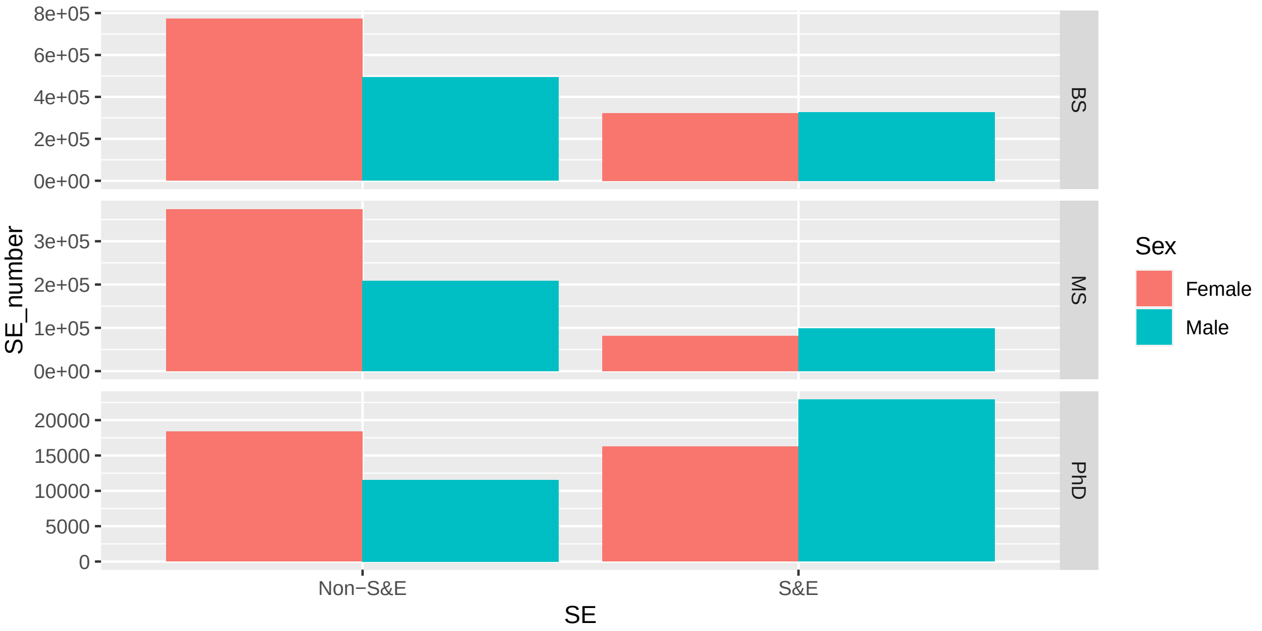
```
# by degree and sex
Casestudy2 %>%filter(Year == 2015) %>%
  ggplot(aes(x = Degree, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme() +
  ggtitle("Degrees granted in 2015 across fields by degree and sex")
```



```
#  by fields and sex
Casestudy2 %>%filter(Year == 2015) %>%
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge")+
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted in 2015 across fields by fields and sex")
```



```
# by SE vs Non-SE and sex
Casestudy2 %>%filter(Year == 2015) %>%
  mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(Degree, SE, Sex) %>%
  summarise(SE_number = sum(Number)) %>%
  ggplot(aes(x = SE, y = SE_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(Degree~., scales = "free_y")
```



```
  theme(axis.text.y = element_text(angle = 60)) +
  ggtitle("Degrees granted in 2015 by S&E vs non-S&E by sex")
```

The above three plots show that gender has an effect on different types of degrees. Specifically, more women had a degree in non-SE fields, while more men had a degree in SE fields across BS, MS, and PhD degree.

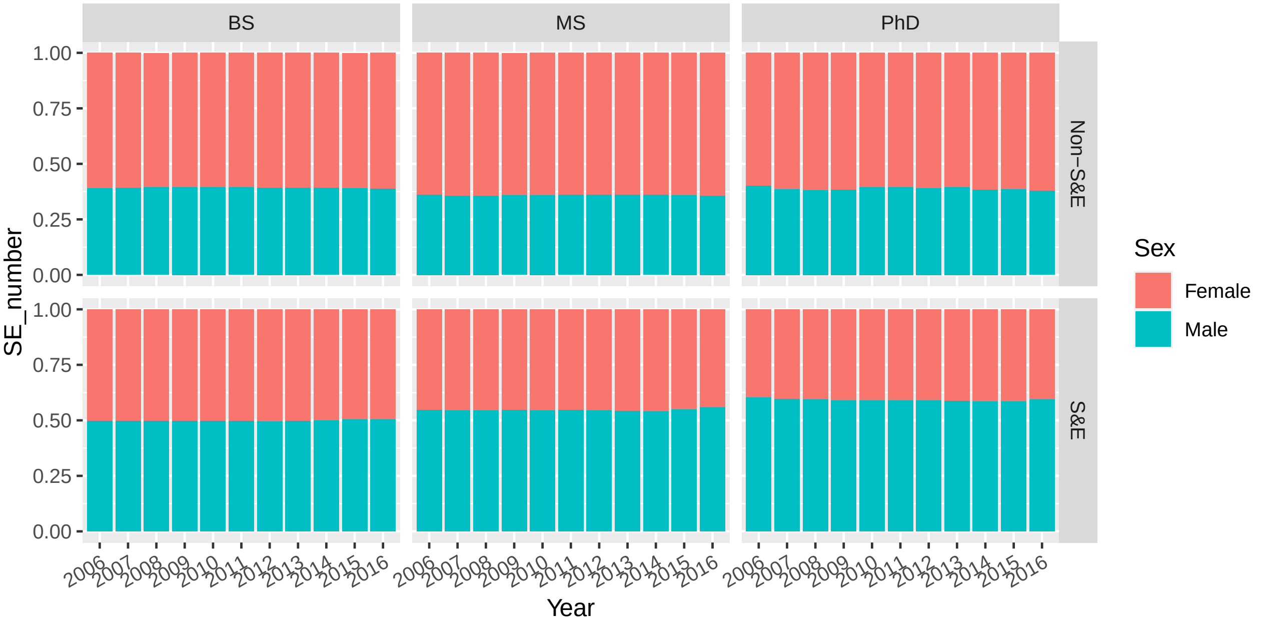
0.5 3.4 Bring all variables

```
# Do the number of degrees change by gender, field, and time?
# Table summary by degrees and gender
x <- Casestudy2 %>%
  pivot_wider(names_from = Sex, values_from = "Number") %>%
  group_by(Degree) %>%
  summarize(Male = sum(Male),
            Female = sum(Female))

x <- x %>%
  mutate(More = x$Male > x$Female)

x$More <- as.character(x$More)
x$More <- str_replace(x$More, ’FALSE’, ’Female’)
x$More <- str_replace(x$More, ’TRUE’,’Male’)
x$More <- as.factor(x$More)

formattable(x, align = c("l"),
            list(‘Degree‘ = formatter("span", style = ~ style(color = "grey", font.weight = "bold")),
                 ‘Male‘ = color_tile("white", "green"),
                 ‘Female‘ = color_tile("white","red"),
                 ‘More‘ = color_tile("green","red")))
```

```
# Table summary by field and gender
x <- Casestudy2 %>%
  pivot_wider(names_from = Sex, values_from = "Number") %>%
  group_by(Field) %>%
  summarize(Male = sum(Male),
            Female = sum(Female))

x <- x %>%
  mutate(More = x$Male > x$Female)

x$More <- as.character(x$More)
x$More <- str_replace(x$More, ’FALSE’, ’Female’)
x$More <- str_replace(x$More, ’TRUE’,’Male’)
x$More <- as.factor(x$More)

formattable(x, align = c("l"),
            list(‘Field‘ = formatter("span", style = ~ style(color = "grey", font.weight = "bold")),
                 ‘Male‘ = color_tile("white", "green"),
                 ‘Female‘ = color_tile("white","red"),
                 ‘More‘ = color_tile("green","red")))
```

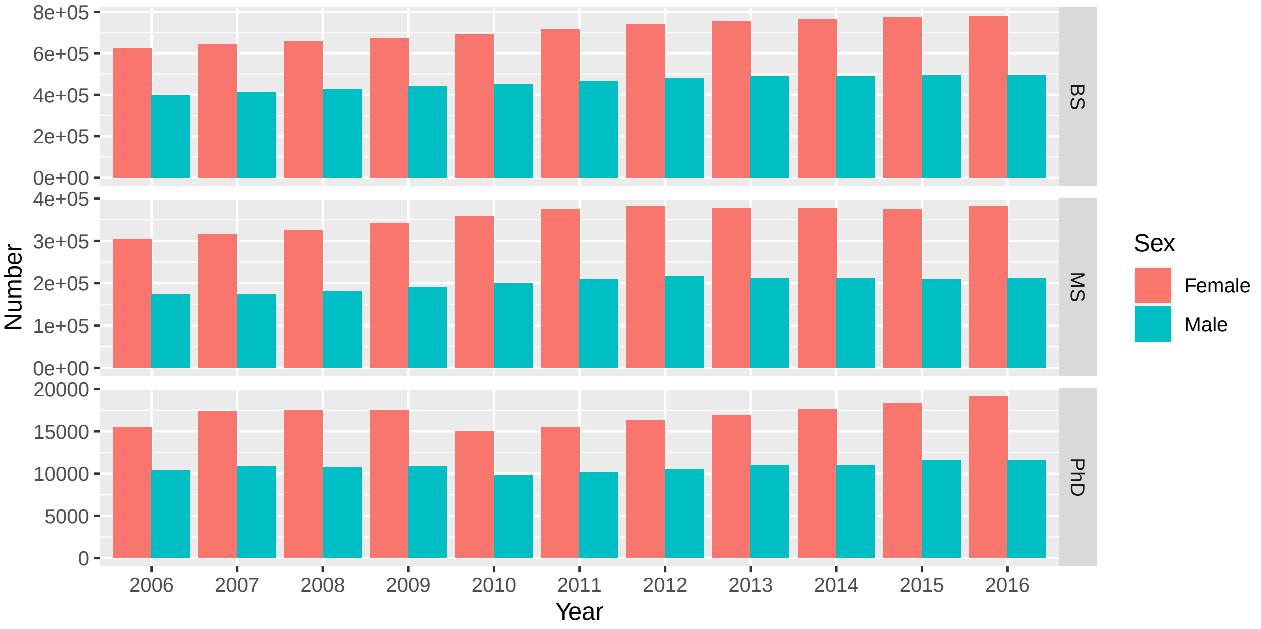
```
# Table summary by year and gender
x <- Casestudy2 %>%
  pivot_wider(names_from = Sex, values_from = "Number") %>%
  group_by(Year) %>%
  summarize(Male = sum(Male),
            Female = sum(Female))

x <- x %>%
  mutate(More = x$Male > x$Female)

x$More <- as.character(x$More)
x$More <- str_replace(x$More, ’FALSE’, ’Female’)
x$More <- str_replace(x$More, ’TRUE’,’Male’)
x$More <- as.factor(x$More)

formattable(x, align = c("l"),
            list(‘Year‘ = formatter("span", style = ~ style(color = "grey", font.weight = "bold")),
                 ‘Male‘ = color_tile("white", "green"),
                 ‘Female‘ = color_tile("white","red"),
                 ‘More‘ = color_tile("green","red")))
```

```
# plot type of degrees by sex and SE/non-SE across years
Casestudy2 %>%
  mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(SE, Sex, Year, Degree) %>%
  summarise(SE_number = sum(Number)) %>%
  ggplot(aes(x = Year, y = SE_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "fill") +
  facet_grid(SE~Degree, scales = "free_y") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```

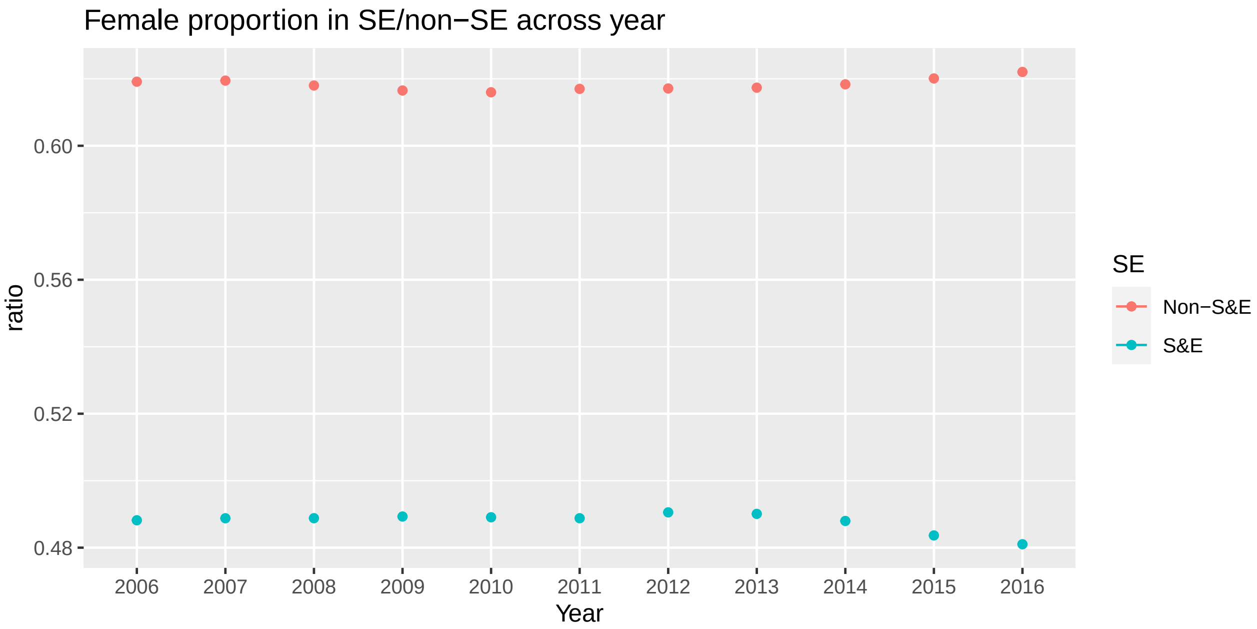


```
  ggtitle("Degrees granted proportion by sex across degree and SE")
```

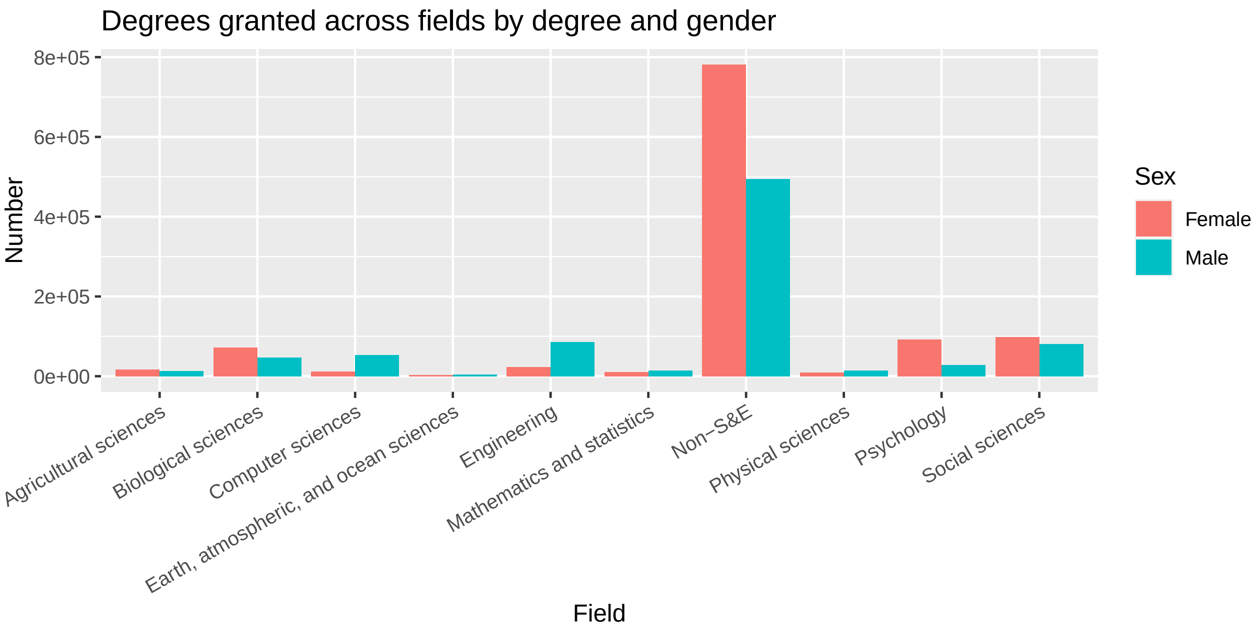
```
# plot degree numbers by year and sex regardless of fields
Casestudy2 %>%
  ggplot(aes(x = Year, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
   facet_grid(Degree~., scales = "free_y")
```



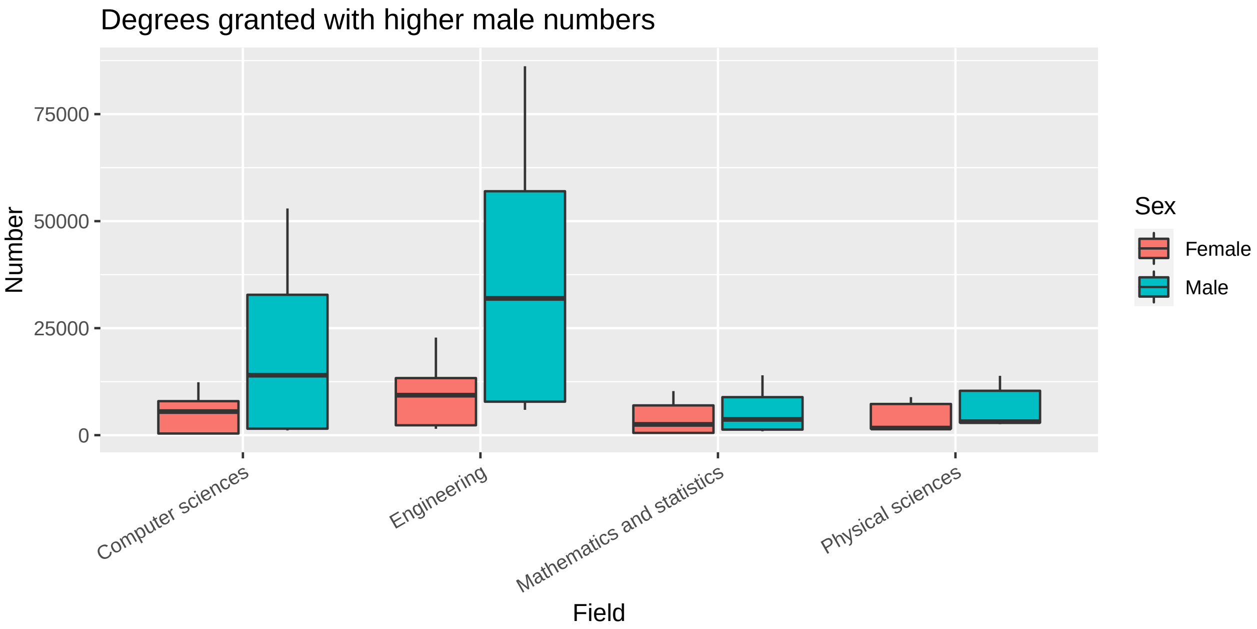
```
  ggtitle("Degrees granted proportion by year and SE")

# plot by female proportion in SE/non-SE across years
Casestudy2 %>%
  mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(SE, Sex, Year) %>%
  summarise(SE_number = sum(Number)) %>%
  group_by(SE, Year) %>%
  mutate(ratio = SE_number / sum(SE_number)) %>%
  filter(Sex == "Female") %>%
  ggplot(aes(x = Year, y = ratio, color = SE)) +
  geom_point() + geom_line() +
  ggtitle("Female proportion in SE/non-SE across year")
```



```
# Plot of degree numbers by field and sex
Casestudy2 %>%
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Degrees granted across fields by degree and gender")
```



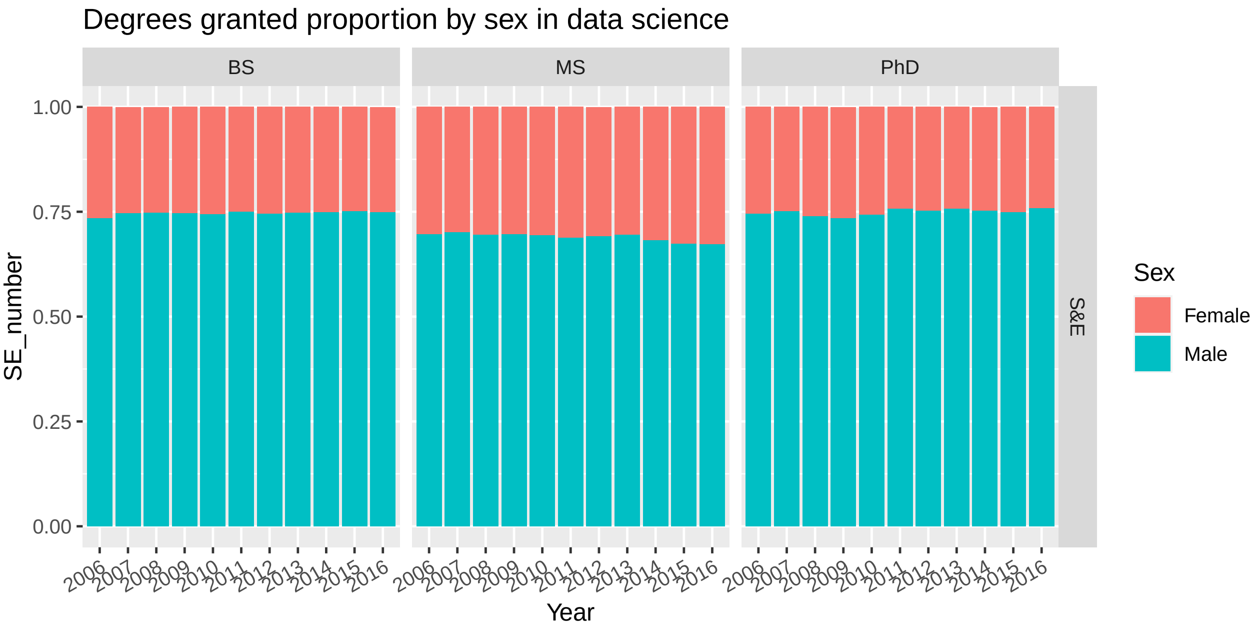
```
Casestudy2 %>% filter(Field %in%
                        c("Computer sciences", "Engineering", "Mathematics and statistics", "Physical s
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_boxplot() +
  ylab("Number") +
  ggtitle("Degrees granted with higher male numbers") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  theme()
```



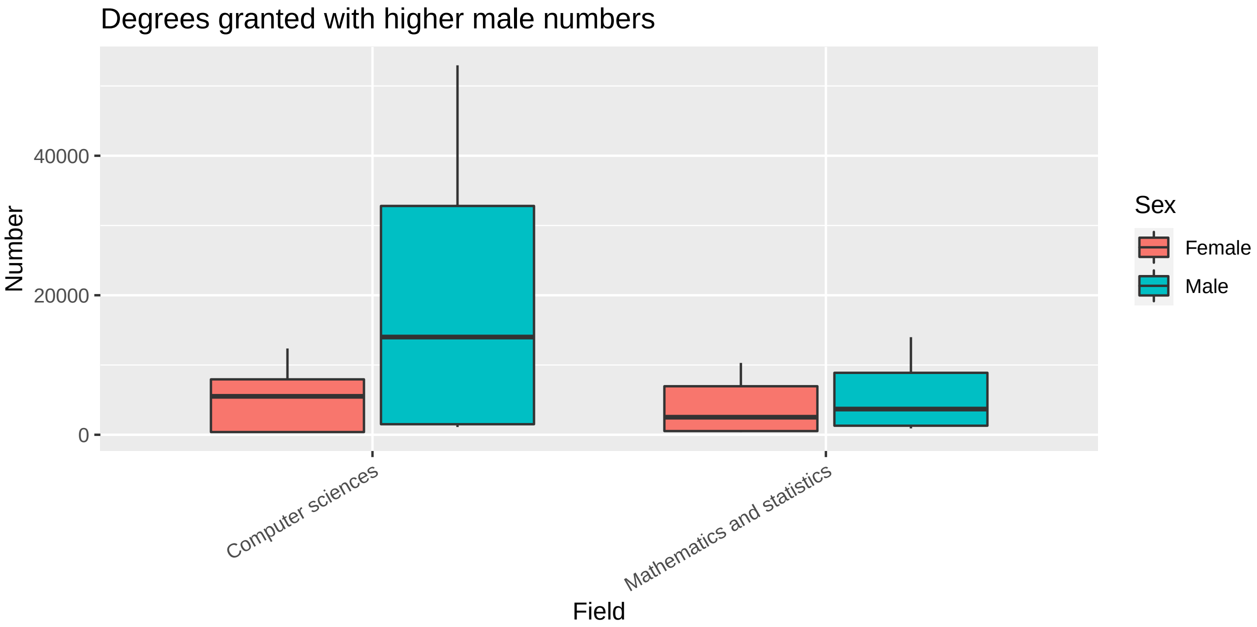
The above tables and plots show that overall, more women got a degree than men across time. However, the proportion of women in non-SE fields is consistently higher than the proportion of women in SE fields across years, suggesting that women tend to enter non-SE fields across time. Further, even within the SE fields, the effect of gender on degree number still exists. Specifically, more women pursued a degree in agriculture, biology, psychology, and social sciences; and engineering and computer science seem to be the two fields with women mostly underrepresented.

0.6 3.3 Women in data science

```
Casestudy2 %>%
  filter(Field %in% c("Computer sciences", "Mathematics and statistics")) %>%
  mutate(SE = ifelse(Field!="Non-S&E" , "S&E", "Non-S&E")) %>%
  group_by(SE, Sex, Year, Degree) %>%
  summarise(SE_number = sum(Number)) %>%
  group_by(Sex, Year, Degree) %>%
  ggplot(aes(x = Year, y = SE_number, fill = Sex)) +
  geom_bar(stat = "identity", position = "fill") +
  facet_grid(SE~Degree, scales = "free_y") +
  ggtitle("Degrees granted proportion by sex in data science") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  theme()
```



```
Casestudy2 %>% filter(Field %in%
                      c("Computer sciences", "Mathematics and statistics")) %>%
  ggplot(aes(x = Field, y = Number, fill = Sex)) +
  geom_boxplot() +
  ylab("Number") +
  ggtitle("Degrees granted with higher male numbers") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  theme()
```



The above plots show that women consistently have lower proportions (i.e. underrepresented) in data science across years. Within data science subgroup, computer science has a higher gender gap.

0.7 3.6 Final conclusion

Based on this dataset, there is a consistently lower proportion of women in science-related fields across years. However, it remains unclear which fields consisted of the non-SE related fields, or why the above nine fields were chosen to represent "Science-related" fields. Other fields, such as chemical sciences and medical and health sciences, among others, belong to science as well. Future studies could improve by asking participants to clarify their degree fields in the data collection process.