STAT 630: Homework 2

Due: September 16th, 2023 at 11:59pm

Exploratory Data Analysis: The overarching goal of this homework is to explore whether there is any evidence suggestive of discrimination by sex in the employment of the faculty at a single university (University of Washington). To this end, salary data (available on Canvas) was obtained on all faculty members employed by the University during the 1995 academic year. You have been asked to provide an analysis of 1995 salaries with the primary goal of determining whether or not gender discrimination exists with respect to pay. Along with the 1995 salary the following additional variables were also collected:

Variable	Description
id	The anonymous identification number for the faculty member sex
sex	Sex of the faculty member (coded as M or F)
degree	The highest degree obtained by the faculty member (PhD, Professional, Other)
field	Field of research during 1995 (Arts, Professional, Other)
year_degree	Year highest degree attained
start_year	Year starting employment at the university
rank	Faculty rank as of 1995 (Assistant, Associate, Full)
admin	Does faculty member hold an administrative position as of 1995? (0 = No, 1 = Yes)
salary	1995 salary in US dollars

```
install.packages("tidyverse")
```

```
## Installing package into '/home/ly/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
## also installing the dependencies 'systemfonts', 'textshaping', 'ragg', 'rvest', 'xml2'
```

```
## Warning in install.packages("tidyverse"): installation of package 'systemfonts'
## had non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'xml2' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'textshaping'
## had non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'rvest' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'ragg' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'tidyverse'
## had non-zero exit status
```

```
salary <- read.csv("~/Documents/Personal Docs_East Bay/STAT 630/salary.csv")
#install.packages("explore")
library(explore)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(dplyr)
library(qwraps2)
summary(salary)
```

```
##        id          sex         deg        year_degree      field
##  Min.   :   1.0   F: 409   Other: 144   Min.   :48.00   Arts : 220
##  1st Qu.: 456.0   M:1188   PhD  :1350   1st Qu.:69.00   Other:1067
##  Median : 894.0            Prof : 103   Median :76.00   Prof : 310
##  Mean   : 887.8                         Mean   :76.09
##  3rd Qu.:1318.0                         3rd Qu.:84.00
##  Max.   :1770.0                         Max.   :96.00
##    start_year        rank         admin            salary
##  Min.   :48.00   Assist:315   Min.   :0.0000   Min.   : 3042
##  1st Qu.:73.00   Assoc :437   1st Qu.:0.0000   1st Qu.: 4743
##  Median :83.00   Full  :845   Median :0.0000   Median : 5962
##  Mean   :81.12                Mean   :0.1058   Mean   : 6390
##  3rd Qu.:90.00                3rd Qu.:0.0000   3rd Qu.: 7602
##  Max.   :95.00                Max.   :1.0000   Max.   :14464
```

```
glimpse(salary)
```

```
## Rows: 1,597
## Columns: 9
## $ id          <int> 1, 2, 4, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 2~
## $ sex         <fct> F, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, F, M, M, M~
## $ deg         <fct> Other, Other, PhD, PhD, PhD, PhD, PhD, PhD, PhD, PhD, PhD,~
## $ year_degree <int> 92, 91, 96, 66, 70, 75, 82, 68, 64, 68, 79, 72, 73, 72, 67~
## $ field       <fct> Other, Other, Other, Other, Other, Other, Other, Arts, Oth~
## $ start_year  <int> 95, 94, 95, 91, 71, 95, 87, 80, 64, 69, 92, 91, 79, 72, 69~
## $ rank        <fct> Assist, Assist, Assist, Full, Assoc, Assist, Assoc, Full, ~
## $ admin       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ salary      <dbl> 6684.000, 4881.000, 4231.000, 12184.000, 4604.000, 4047.67~
```

```
#View(salary)
```

1. Coerce `sex`, `degree`, `field`, `rank`, and `admin` to factors.

```
salary <- read.csv("~/Documents/Personal Docs_East Bay/STAT 630/salary.csv")
salary$sex_f <- factor(salary$sex)
levels(salary$sex_f)
```

```
## [1] "F" "M"
```

```
salary$sex_f <- as.numeric(salary$sex_f)

salary$deg_f <- factor(salary$deg)
```

```
levels(salary$deg_f)
```

```
## [1] "Other" "PhD"   "Prof"
```

```
salary$deg_f <- as.numeric(salary$deg_f)


salary$field_f <- factor(salary$field)
levels(salary$field_f)
```

```
## [1] "Arts"  "Other" "Prof"
```

```
salary$field_f <- as.numeric(salary$field_f)

salary$rank_f <- factor(salary$rank)
levels(salary$rank_f)
```

```
## [1] "Assist" "Assoc"  "Full"
```

```
salary$rank_f <- as.numeric(salary$rank_f)

salary$admin_f <- factor(salary$admin,
                         labels = c("No", "Yes"))
```

2. Make a new column called `years_uni` and calculate the number of years the instructor has been teaching at the University (note that start year is recorded using only the last two digits of the year, e.g., 95 rather than 1995).

```
salary <- salary %>%
  mutate(years_uni = 1995-(start_year + 1900))
```

3. Create a table of descriptive statistics for each variable in the dataset, stratified by `sex`. Use the `kable()` function to create a publication-quality table (use the table in HW 1 as a guide!). Use the packages in the `tidyverse` (like `dplyr`) and include all of the code used.

```
# Code used to create table
#install.packages("tidyverse")
#library(tidyverse) # Load the tidyverse packages
#tidyverse_packages() # View packages included in the tidyverse
salary %>%
  group_by(sex) %>%
  summarise_if(is.numeric, mean_sd,na_rm = TRUE,denote_sd = "paren")
```

```
## # A tibble: 2 x 11
##   sex   id        year_degree start_year admin salary sex_f deg_f field_f rank_f
##   <fct> <chr>     <chr>       <chr>      <chr> <chr>  <chr> <chr> <chr>   <chr>
## 1 F     919.83 (~ 81.11 (8.7~ 85.47 (8.~ 0.08~ 5,396~ 1.00~ 1.91~ 1.91 (~ 1.95 ~
## 2 M     876.76 (~ 74.37 (9.6~ 79.62 (10~ 0.12~ 6,731~ 2.00~ 2.00~ 2.11 (~ 2.46 ~
## # i 1 more variable: years_uni <chr>
```

```
tab <- matrix(c(919.83, 491.30, 876.76, 511.66, 81.11, 8.70, 74.37, 9.64, 85.47, 8.02, 79.62,
                10.17, 0.08, 0.27, 0.12, 0.32,5396.91, 1481.22, 6731.64, 2089.76),nrow = 2,byrow = TRUE)

knitr::kable(tab)
```

919.83	491.30	876.76	511.66	81.11	8.70	74.37	9.64	85.47	8.02
79.62	10.17	0.08	0.27	0.12	0.32	5396.91	1481.22	6731.64	2089.76

```
salary %>%
  group_by(sex) %>%
  count_pct(deg)
```

```
## # A tibble: 6 x 5
## # Groups:   sex [2]
##   sex   deg       n total   pct
##   <fct> <fct> <int> <int> <dbl>
## 1 F     Other    56   409 13.7
## 2 F     PhD     334   409 81.7
## 3 F     Prof     19   409  4.65
## 4 M     Other    88  1188  7.41
## 5 M     PhD    1016  1188 85.5
## 6 M     Prof     84  1188  7.07
```

```
install.packages("skimr")
```

```
## Installing package into '/home/ly/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
library(skimr)
```

4. Based on the table you created above, does there appear to be sex discrimination at the University? Explain in 2-3 sentences.

It does seem there may be evidence of sex discrimination at the University. The mean salary for males (6,731.64) is notably higher than that of females (5,396.91), indicating a significant disparity. Additionally, the higher standard deviation for males (2,089.76) compared to females (1,481.22) suggests greater variability in male salaries, potentially indicating that there are more extreme cases of overcompensation for males.

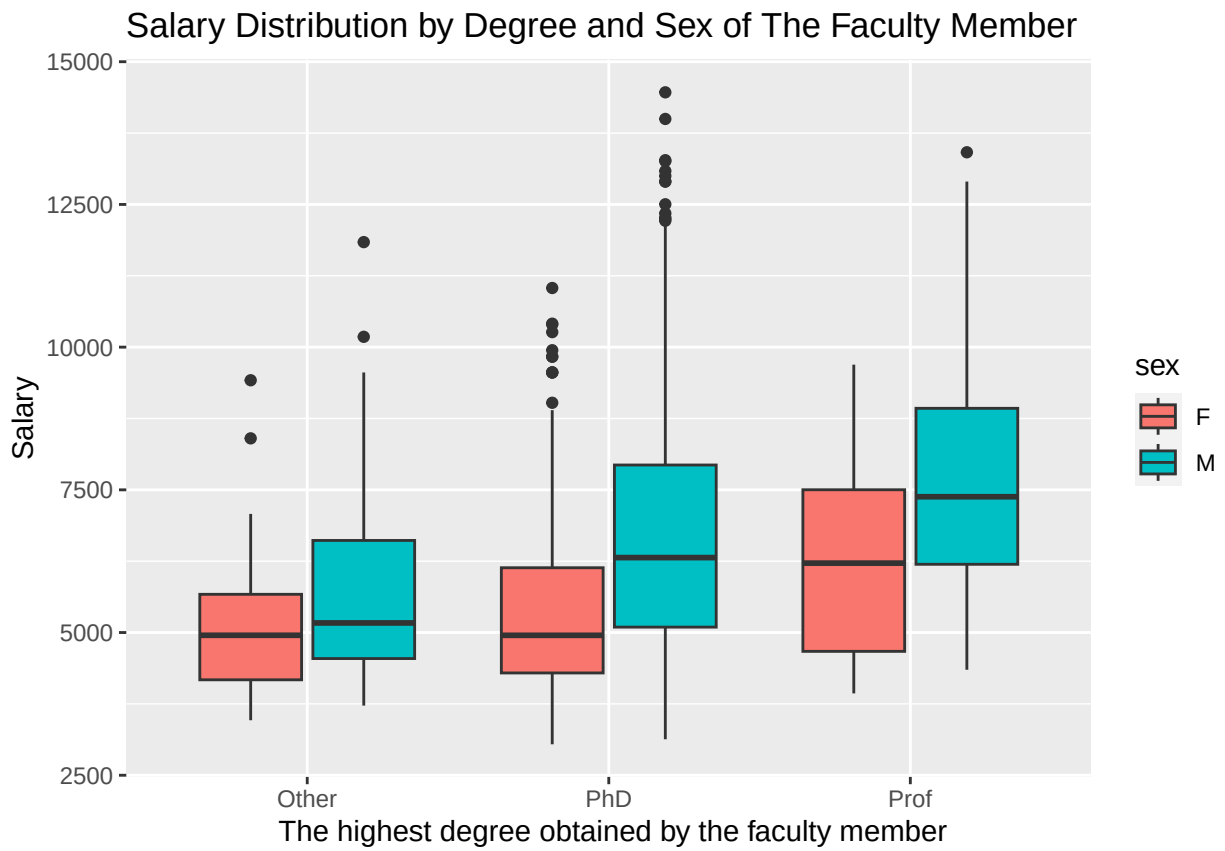
5. Choose what you believe to be the top two confounding variables in the relationship between sex and salary. Explain how each confounding variable is related to both sex and salary.

I believe Field and Degree are the top two confounding variables in the relationship between sex and salary. As we can see these two graphs plotted below.

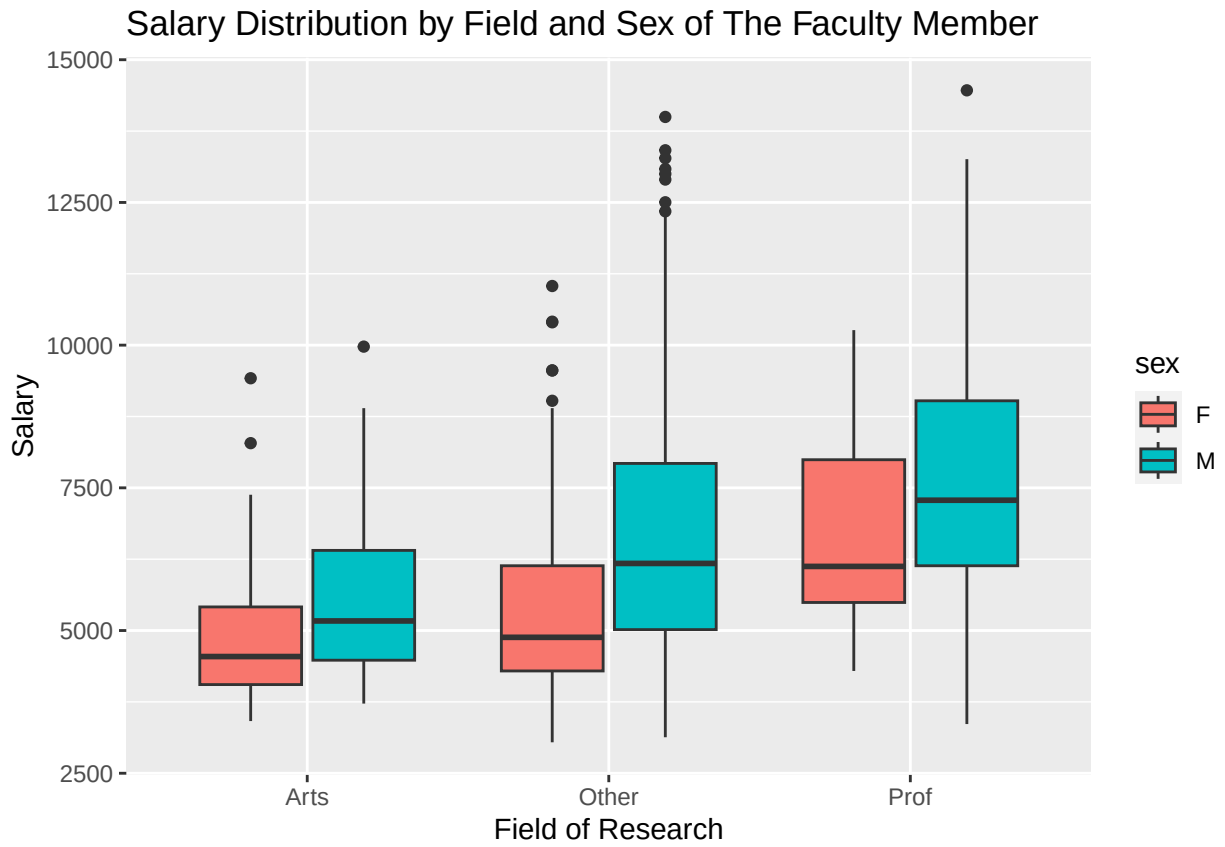
For field of conducted research, we can see that it also changes according to gender and salary as well. The salary of Professional faculty members is the highest. Below Professional's salary is Other and Arts. The mean and standard deviation of male faculty members are also higher than female in each field.

For degree obtained by the faculty member, it also differ according to salary and gender. The mean and standard deviation of male faculty member's salary are also higher than female's in each type of degree.

```
ggplot(salary, aes(x = deg, y = salary, fill = sex)) +
  geom_boxplot()+
  labs(title = "Salary Distribution by Degree and Sex of The Faculty Member",
       x = " The highest degree obtained by the faculty member",
       y = "Salary",
       )
```

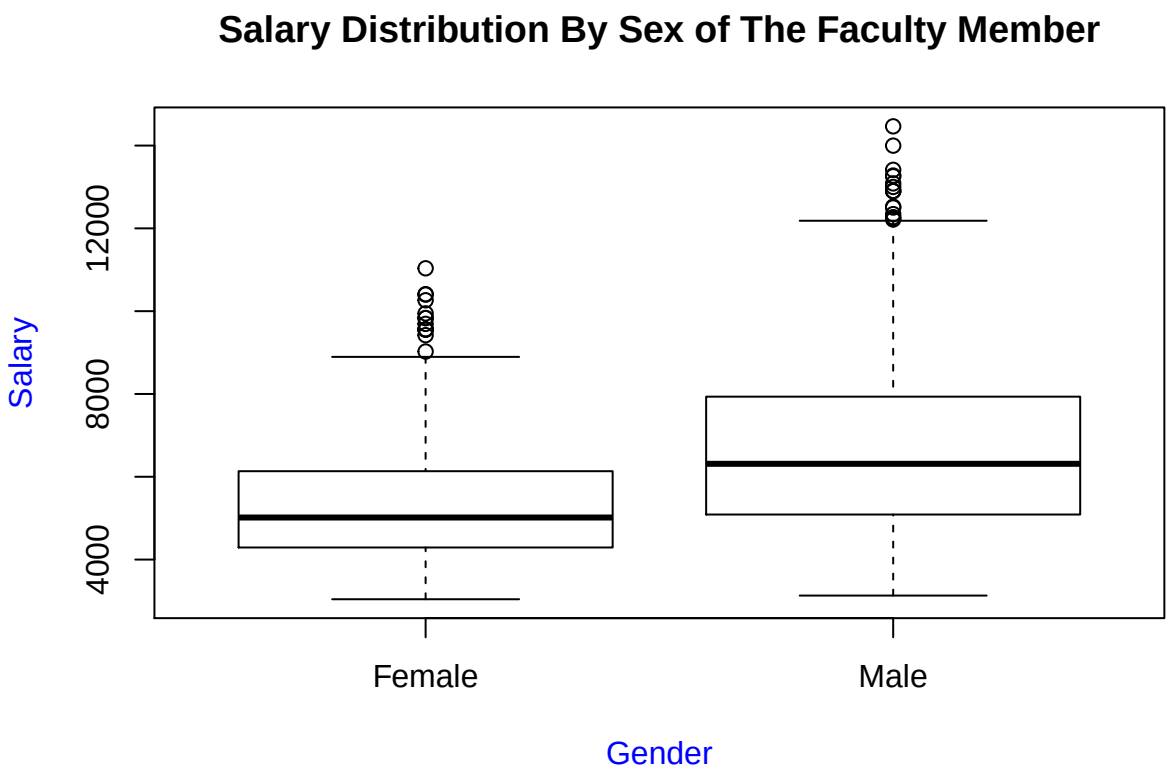



```
ggplot(salary, aes(x = field, y = salary, fill = sex)) +
  geom_boxplot()+
  labs(title = "Salary Distribution by Field and Sex of The Faculty Member",
       x = "Field of Research",
       y = "Salary",
       )
```



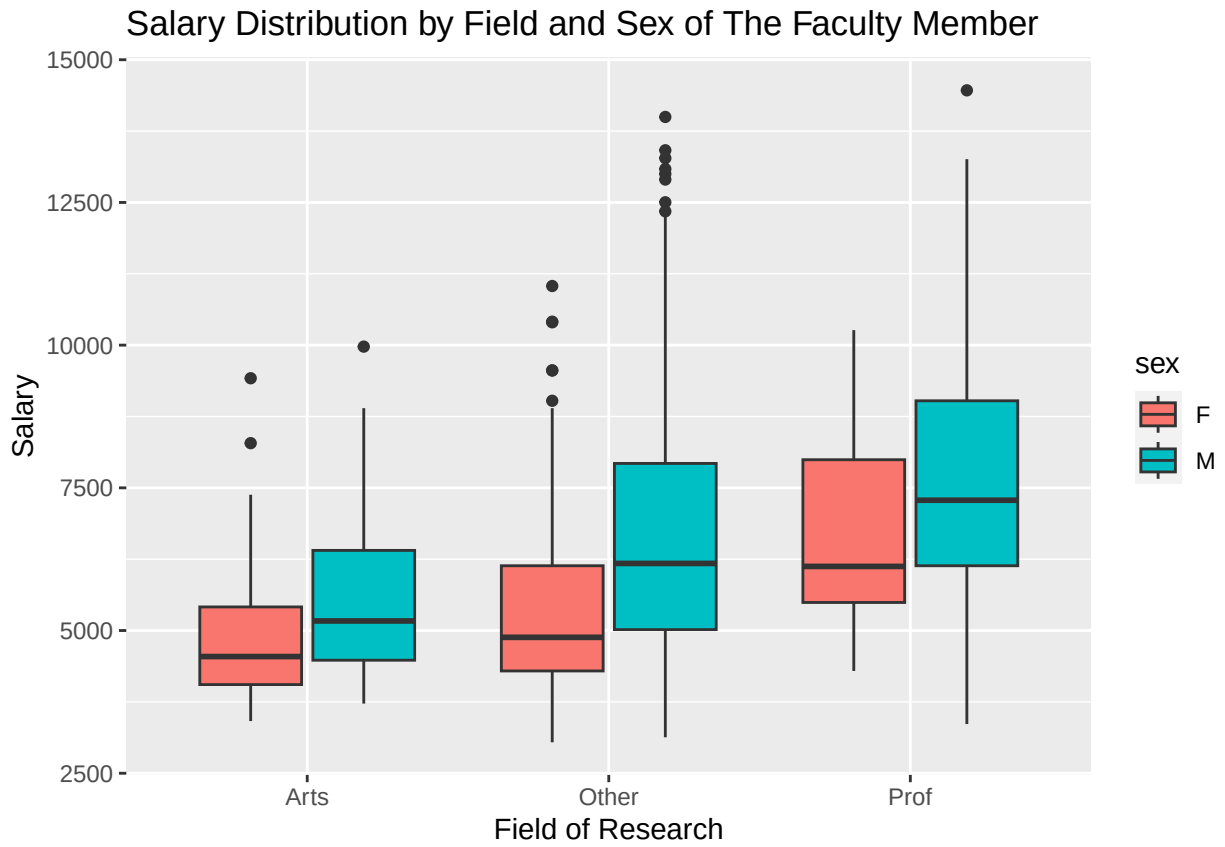
6. Using `base` R, plot the relationship between sex and salary.

```
label = c("Female", "Male")
boxplot(salary~sex, salary, names = label,
        main = "Salary Distribution By Sex of The Faculty Member",
        xlab = "Gender",
        ylab = "Salary",
        col.lab="blue")
```



7. Using `ggplot2`, plot the relationship between sex, salary, and one of your confounding variables.

```
ggplot(salary, aes(x = field, y = salary, fill = sex)) +
  geom_boxplot()+
  labs(title = "Salary Distribution by Field and Sex of The Faculty Member",
       x = "Field of Research",
       y = "Salary",
       )
```

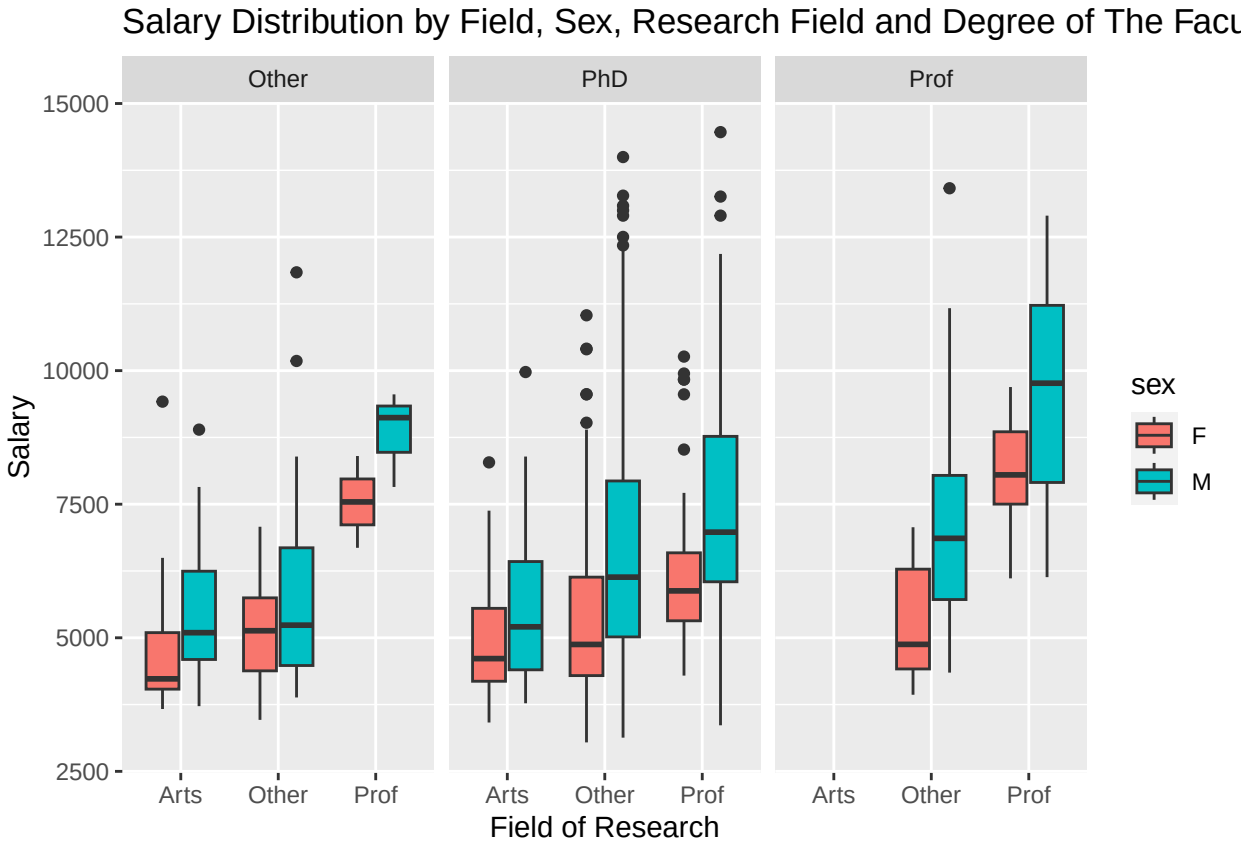


8. Comment on how the relationship between sex and salary changes for different values of your confounding variable in 1-2 sentences.

As shown on the boxplot above, we can see that the salary of professional field of research is highest and has greatest variability compared to Other and Arts. The salary of Art field of research is lowest. So we can see that the salary of the faculty member increases from Arts to Professional research. However, the range and amount of salary of Male staff is always higher than that of female.

Challenge question: Visualize the relationship between sex, salary, and both of your confounding variables in a single plot.

```
ggplot(salary) +
  geom_boxplot(aes(x = field, y = salary, fill = sex)) +
  facet_wrap(~ deg)+
  labs(title = "Salary Distribution by Field, Sex, Research Field and Degree of The Faculty Member",
       x = "Field of Research",
       y = "Salary",)
```


Salary Distribution by Field, Sex, Research Field and Degree of The Facu
Other
PhD
Prof
15000
12500
10000
7500
5000
2500
Salary
Arts
Other
Prof
Field of Research
sex
F
M