STAT 630: Homework 2

Brandon Keck

Due: September 16th, 2024 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

```
## Warning: package 'readr' was built under R version 4.2.3
```

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

```
head(salary)
```

```
## # A tibble: 6 x 9
##      id sex   deg   year_degree field start_year rank   admin salary
##   <dbl> <chr> <chr>       <dbl> <chr>      <dbl> <chr>  <dbl>  <dbl>
## 1     1 F     Other          92 Other         95 Assist     0  6684
## 2     2 M     Other          91 Other         94 Assist     0  4881
## 3     4 M     PhD            96 Other         95 Assist     0  4231
## 4     6 M     PhD            66 Other         91 Full       0 12184
## 5     7 M     PhD            70 Other         71 Assoc      0  4604
## 6     8 M     PhD            75 Other         95 Assist     0  4048.
```

```
library(readr)
library(here)

# Coerce the specified variables to factors
salary$sex <- factor(salary$sex) # sex coerced to factor variable
```

```
salary$degree <- factor(salary$deg) # degree coerced to factor variable
salary$field <- factor(salary$field) # field of discipline coerced to factor variable
salary$rank <- factor(salary$rank) # rank of faculty coerced to factor variable
salary$admin <- factor(salary$admin) # administrator position coerced to factor variable
```

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).

```
#install.packages("dplyr") # Make sure to comment out once ran
library(dplyr) # This calls the dplyr library
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
salary <- salary %>%
  mutate(years_uni = 95 - (start_year))
# using the mutate function which is a part of dplyer package.
# mutate creates a new column which in this case is called years_uni
```

3. Using `gtsummary()` create a table of descriptive statistics for each variable in the dataset, stratified by `sex`.

```
#install.packages("gtsummary") # Make sure to comment out once ran
library(gtsummary) # This calls the gysummary library
```

```
salary %>%
  tbl_summary(by = sex,
              digits = list(
                all_continuous() ~ c(2,2)
              ),
              statistic = list(
                all_continuous() ~ "{mean} ({sd})"
               ))
```

Characteristic	**F** N = 409[1]	**M** N = 1,188[1]
id	919.83 (491.30)	876.76 (511.66)
deg		
Other	56 (14%)	88 (7.4%)
PhD	334 (82%)	1,016 (86%)
Prof	19 (4.6%)	84 (7.1%)
year_degree	81.11 (8.70)	74.37 (9.64)
field		
Arts	80 (20%)	140 (12%)
Other	287 (70%)	780 (66%)
Prof	42 (10%)	268 (23%)
start_year	85.47 (8.02)	79.62 (10.17)

rank		
Assist	145 (35%)	170 (14%)
Assoc	138 (34%)	299 (25%)
Full	126 (31%)	719 (61%)
admin		
0	377 (92%)	1,051 (88%)
1	32 (7.8%)	137 (12%)
salary	5,396.91 (1,481.22)	6,731.64 (2,089.76)
degree		
Other	56 (14%)	88 (7.4%)
PhD	334 (82%)	1,016 (86%)
Prof	19 (4.6%)	84 (7.1%)
years_uni	9.53 (8.02)	15.38 (10.17)

[1] Mean (SD); n (%)

```
# Used from lectures notes with the help of Dr. Moore
```

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

Yes there appears to be a sex discrimination at the University of Washington based on the descriptive statistics above. On average, male faculty earn $6,731.64 while female faculty earn $5,396.91. Also male faculty members hold more senior positions such as Full professors at 61% for males while females hold Full professors at 31%. Additionally female faculty also seem to hold lower positions such as associate and assistant professors compared to their male counterparts.

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 that level of degree and years of experience are the top two confounding variables between the relationship of sex and salary. I believe level of education (Degree) of other, PhD, and professor is closely tied to salary because higher levels of education generally earns more money. Because of this there are males that hold higher levels of degrees that are earning more money than their female counter parts. While years of experience is another confounding variable that generally earns more money as well. The data shows that on average males have more experience at the university level at 15.38 years compared to 9.53 years for females. I myself have witnessed this as I am attempting to enter the workplace and how the industry wants 10+ years of experience and a PhD.

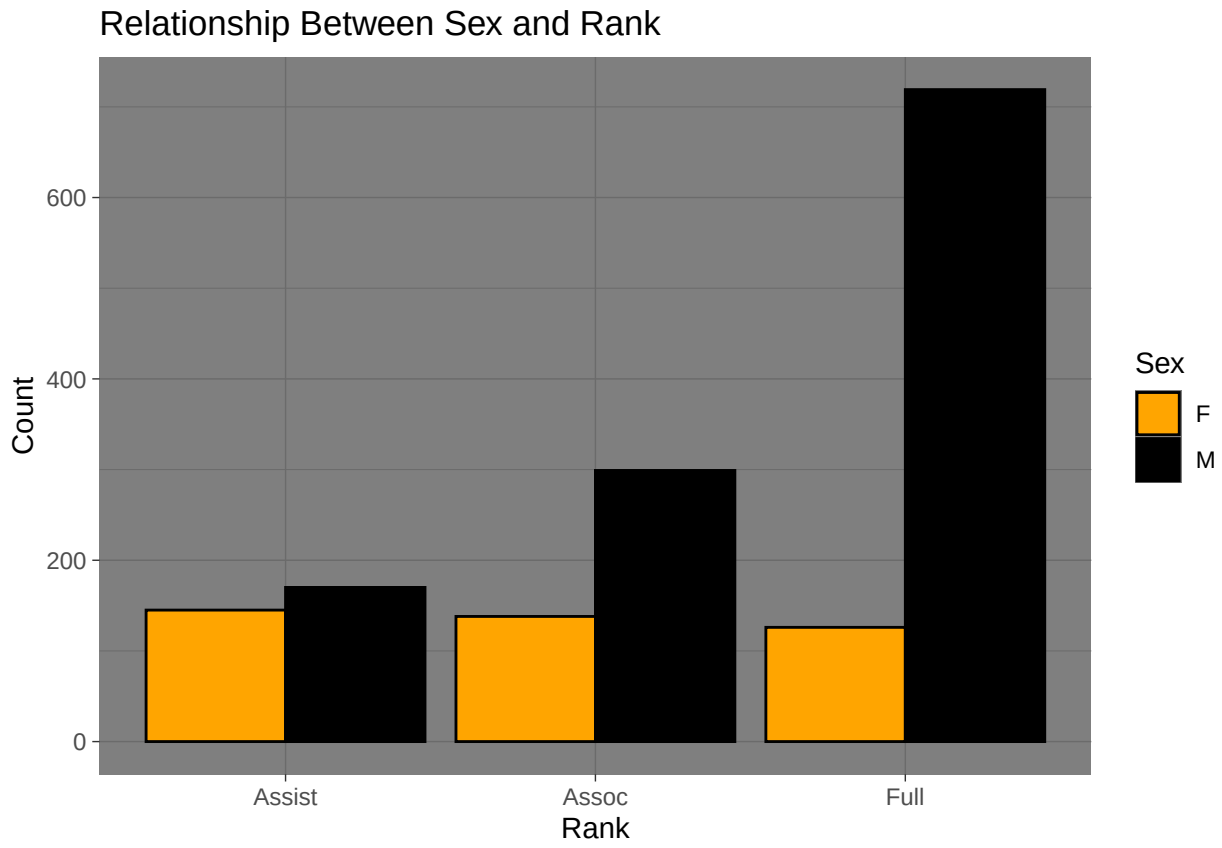
6. Using the R package of your choice, plot the relationship between `sex` and `rank`.

```
#install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
# ggplot2 is my favorite package and one I'm most familiar with
```

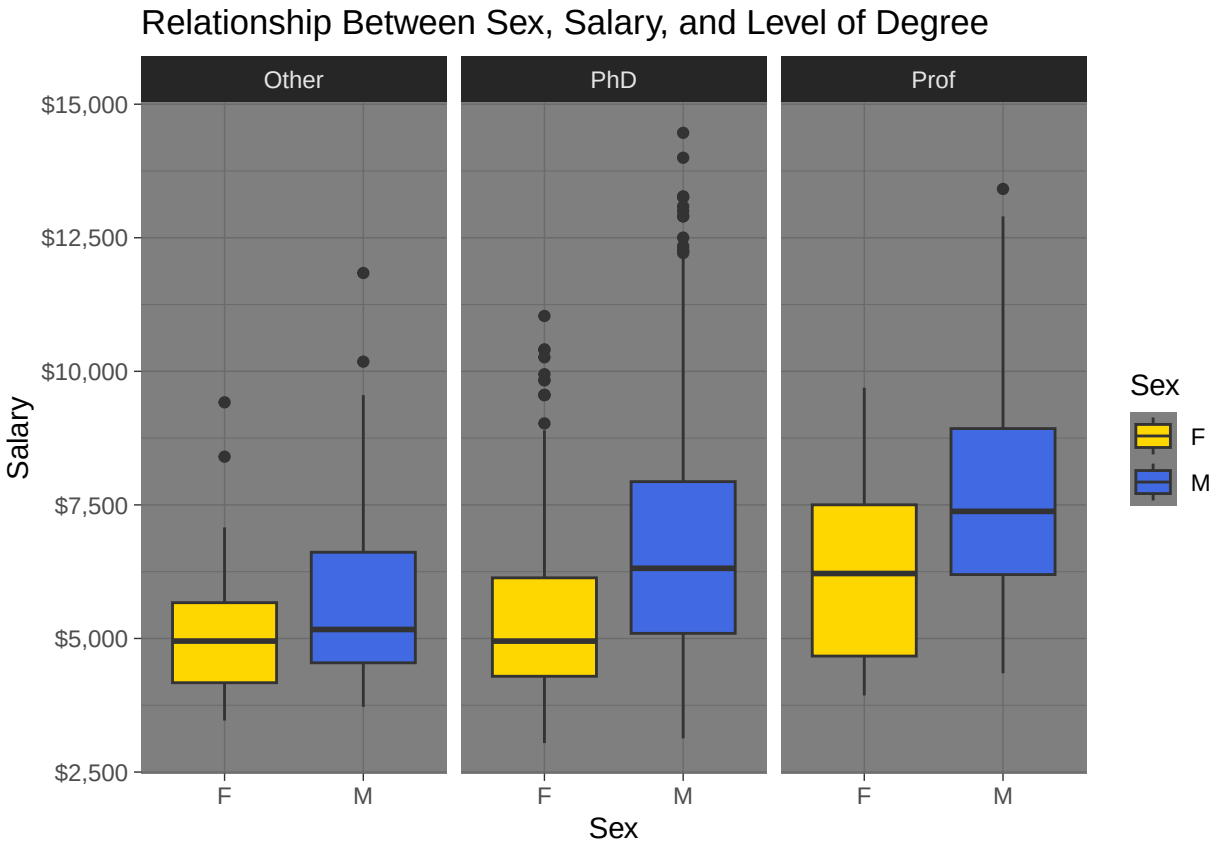
```
ggplot(data = salary, aes(x = rank, fill = sex)) +
  geom_bar(position = "dodge", color = "black") +
  # position = "dodge must be specified to have the bars
  #one beside each other. Cited:https://r-graph-gallery.com/48-grouped-barplot-with-ggplot2
  labs(title = "Relationship Between Sex and Rank",
       x = "Rank",
       y = "Count",
       fill = "Sex") +
  scale_fill_manual(values = c("M" = "black", "F" = "orange")) +
  # Feeling orange and black today as we approach Halloween but
  # more importantly GO OSU BEAVERS!!
  theme_dark() # theme_dark is my favorite
```



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

```
ggplot(data = salary, aes(x = sex, y = salary, fill = sex)) +
  geom_boxplot() + # Used deg as confounding variable.
  # Tried using years_uni and it turned out to be a mess.
  facet_wrap(~deg) + # Need facet wrap to separate each box-plot by degree type
  scale_y_continuous(labels = scales::dollar_format()) +
  # modify the scale of y axis. the dollar_formats numbers
  # as currency as part of the ggplot2 package
```

```
  labs(title = "Relationship Between Sex, Salary, and Level of Degree",
       x = "Sex",
       y = "Salary",
       fill = "Sex") +
  scale_fill_manual(values = c("M" = "royalblue", "F" = "gold")) + # Go Bears!!
  theme_dark() # theme_dark is favorite
```



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

From the box-plot, we can see that sex and salary changes across the different levels of degree of faculty. Both for male and female the median for the "Other" level is about the same. However, as we move to the "PhD" and "Prof" levels we witness a clear disparity between males and females. This is more pronounced at the PhD level where disparity between male and female salaries is the largest.

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

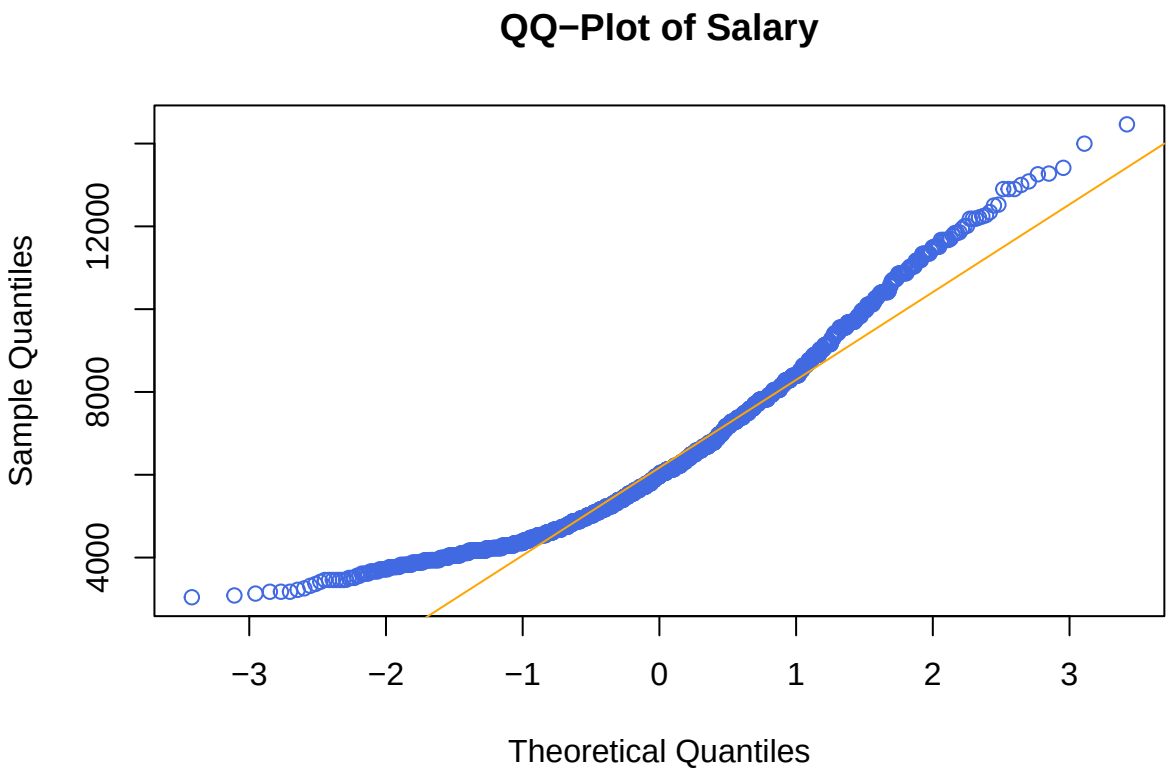
```
ggplot(data = salary, aes(x = years_uni, y = salary, color = sex, shape = deg)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col ="darkred") +
  # lm adds the linear line to our plot. Was trying for a
  # maroon color but this is close enough
  scale_y_continuous(labels = scales::dollar_format()) + # modify the scale of y axis.
  #the dollar_formats numbers as currency as part of the ggplot2 package
```

```
  labs(title = "Relationship Between Sex, Salary, Years at University, and Level of Degree",
       x = "Years at University",
       y = "Salary",
       color = "Sex") +
  scale_color_manual(values = c("M" = "royalblue", "F" = "lightpink")) +
  theme_dark()
```


```
# the confounding variables used are years_uni and degree.
# This is to see the relationship between salary, sex, Years at University and Level of Degree
# had to look up how to fill the shapes color because
# scale_fill_manual is used for differentiating sex. # Cited:https://ggplot2.tidyverse
#.org/reference/scale_manual.html
```

9. Create a QQ-plot of salary. Use `qqline()` to ad a reference line to the plot. Do the points on the QQ-plot fall on the straight line? Comment on any deviations in the data from the normal distribution.

```
qqnorm(salary$salary, main = "QQ-Plot of Salary", col = "royalblue")
qqline(salary$salary, col = "orange")
```



From the QQ-Plot of salary we can see that both the lower and upper tails shows deviations. These occur at the points of below -2 and above 2 on the plot.From the QQ-Plot we observe that there is some skewness in the salary distribution particularly in the upper portion. The points that do follow the line suggests that mid-range salaries follow a more normal distribution than that of lower and higher paying salaries.

10. Calculate the proportion of salaries in the dataset that are greater than $8,000.

```
high_salary <- sum(salary$salary > 8000) / nrow(salary)
high_salary
```

```
## [1] 0.2028804
```

The proportion of salaries in the dataset that are greater than $8,000 is 0.2028804 or 20.2%.

11. Let us assume that `salary` is normally distributed regardless of what you found in Question 9. With the mean and standard deviation of `salary`, calculate the probability that a randomly selected salary is greater than $8,000 using the `pnorm()` function. Comment on how this answer compares to the proportion you found in Question 10. Why are are they similar or different?

```
mean_salary <- mean(salary$salary) # Mean salary
sd_salary <- sd(salary$salary) # SD Salary

pnorm(8000, mean_salary, sd_salary, lower.tail = FALSE) # Using the pnorm() function
```

```
## [1] 0.2146003
```

The calculated probability of 21% (0.2146003) and the proportion of salaries > 8,000 is 20% (0.2028804) which is close but just different enough.What we can make of this information is that the similarities of the closeness of the values suggests that the data doesn't deviate from normality when comparing values of 8,000.

12.Was there anything you found difficult with this homework? What topics (if any) do you feel you still need more work on?

The parts that I found difficult were creating the table of descriptive statistics. I was trying to figure out how to create the table but I was confused about why we would remove the columns argument. After talking with Dr. Moore about it she explained that since we want to summarize all the variables in the data set we do not need the columns argument for specific columns. The other section that I found difficult was the challenge question. I always like to fill my plots with different colors. I found that the colors that were being used was difficult for me to distinguish which variable was which. But now with the darkred color I can see the linear relationship between the variables.

13) Give yourself a rating for this assignment using the EMRN rubric.

E - Excellent

M - Meeting expectations

R - Revision needed

N - Not assessible (mostly blank or did not complete)

I believe that I earned an E for excellent on this assignment. I definitely went over board with my comments on this project and sited all of the sites that I used that I didn't inherently know off the top of my head. I believe that I answered each question to the fullest and did the best job that I could :).