STAT 630: Homework 2

David Teng

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

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

```
sex <- as.factor(salary$sex)
degree <- as.factor(salary$deg)
field <- as.factor(salary$field)
rank <- as.factor(salary$rank)
admin <- as.factor(salary$admin)
```

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).

```
year_uni <- 95 - salary$start_year
```

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

```
library(tidyr)
library(gtsummary)
salary %>% tbl_summary(by = sex)
```

Characteristic	**F** N = 409[1]	**M** N = 1,188[1]
id	925 (504, 1,359)	884 (436, 1,310)
deg		

Other	56 (14%)	88 (7.4%)
PhD	334 (82%)	1,016 (86%)
Prof	19 (4.6%)	84 (7.1%)
year_degree	82 (74, 89)	73 (67, 82)
field		
Arts	80 (20%)	140 (12%)
Other	287 (70%)	780 (66%)
Prof	42 (10%)	268 (23%)
start_year	88 (80, 92)	80 (71, 89)
rank		
Assist	145 (35%)	170 (14%)
Assoc	138 (34%)	299 (25%)
Full	126 (31%)	719 (61%)
admin	32 (7.8%)	137 (12%)
salary	5,016 (4,292, 6,135)	6,313 (5,088, 7,935)

[1] Median (Q1, Q3); n (%)

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

Yes, it does seem like sex discrimination exists at the University. Men tend to hold higher academic ranks, earn more, and have more administrative roles compared to women.

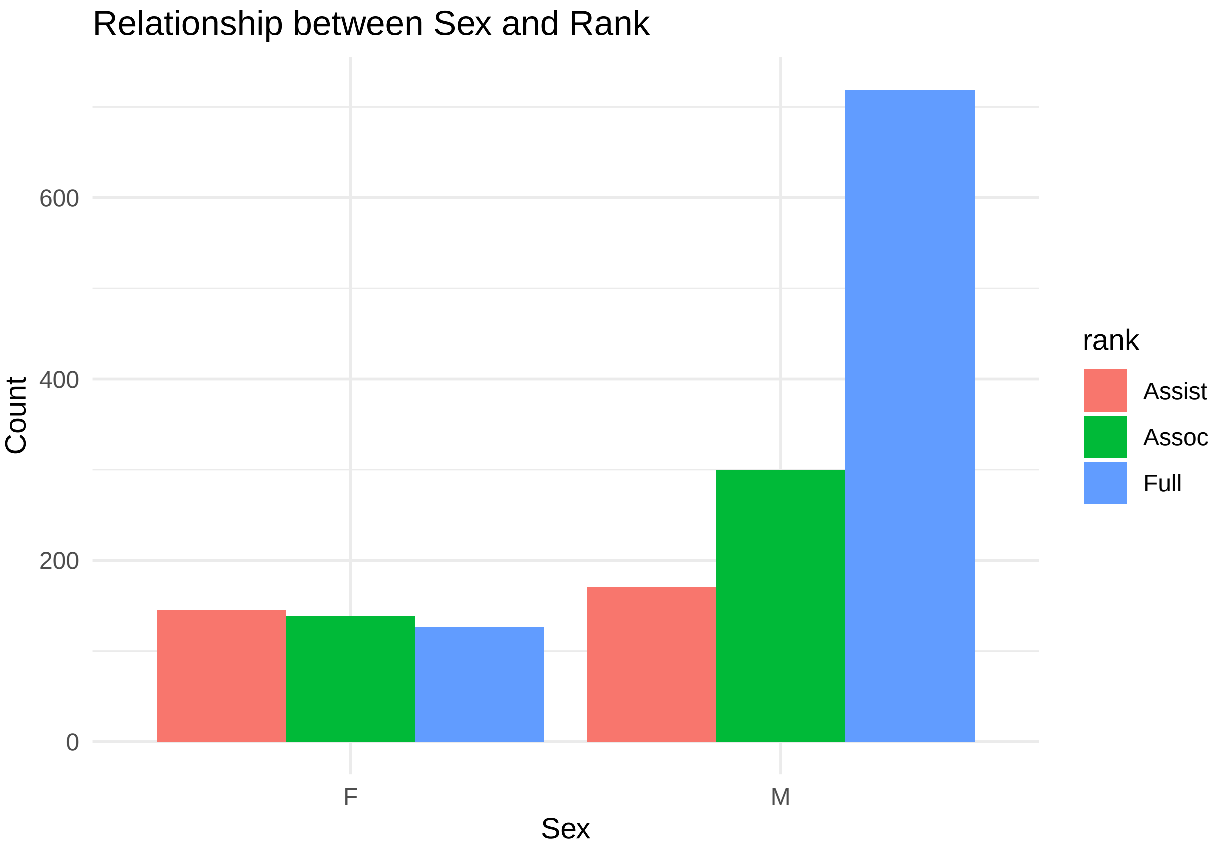
5. Choose what you believe to be the top two confounding variables in the relationship between `sex` and `salary`. Explain how each confounding variable is related to both `sex` and `salary`.

I believe rank and field are the top two confounding variables in the relationship between sex and salary. For rank,I think different rank will have difference salary. Only 31% females get a full time job, males has 61% chance to get a full time job. Additionally, 69% of females are in assistant or associate positions, while only 39% of males are in these positions. As full-time positions get higher salaries, rank is a significant confounding variable.

For field,I think different field will have difference salary. Only 10% of fernales are in professional field, males are 23%. There are 20% of females in art field compare to males are 12%. There are only a tiny difference between others field.Female(70%), male (66%).Due to professional field will get higher salaries, field is a signficant confounding variable

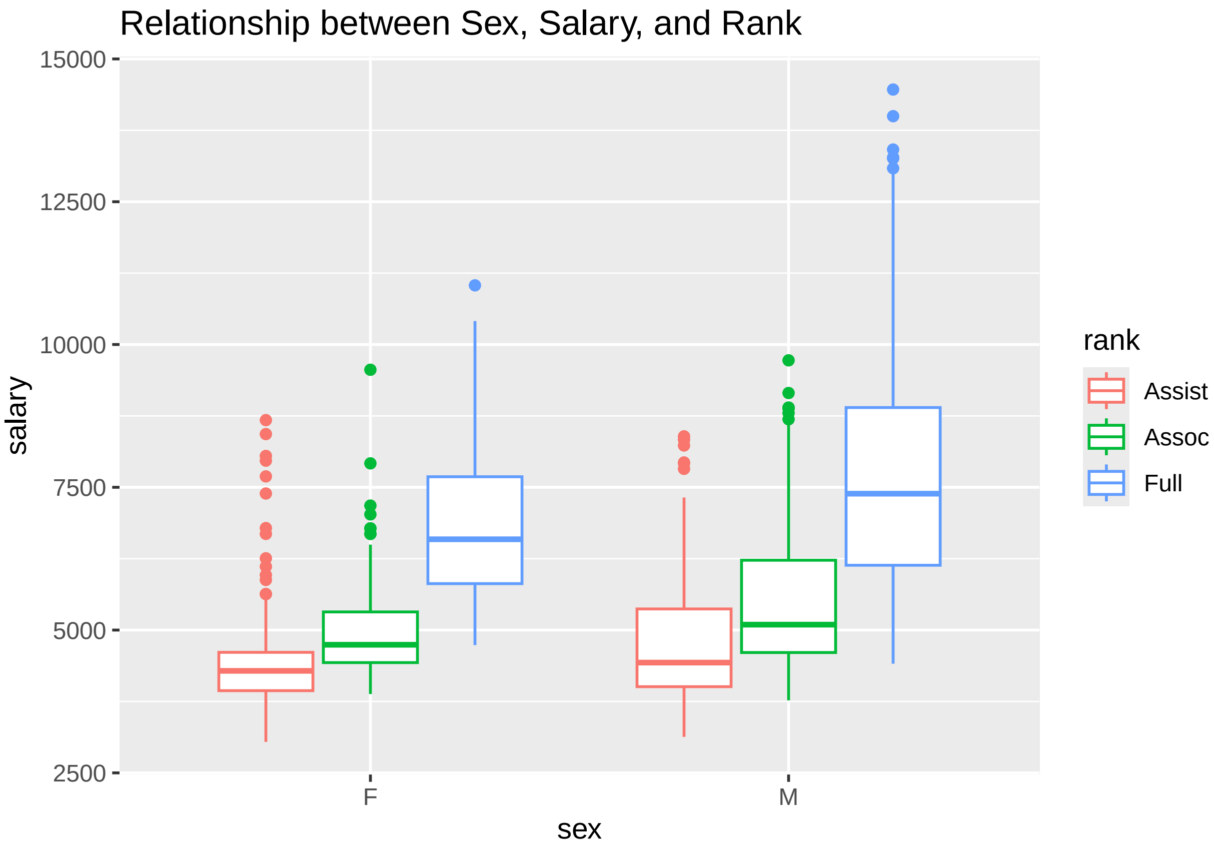
6. Using the R package of your choice, plot the relationship between `sex` and `rank`.

```r
library(ggplot2)
ggplot(salary, aes(x = sex, fill = rank)) +
  geom_bar(position = "dodge")+
  labs(title = "Relationship between Sex and Rank",
       x = "Sex",
       y = "Count") +
  theme_minimal()
```



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

```r
ggplot(salary, aes(x = sex, y = salary, color = rank)) +
  geom_boxplot()+
  labs(title = "Relationship between Sex, Salary, and Rank")
```

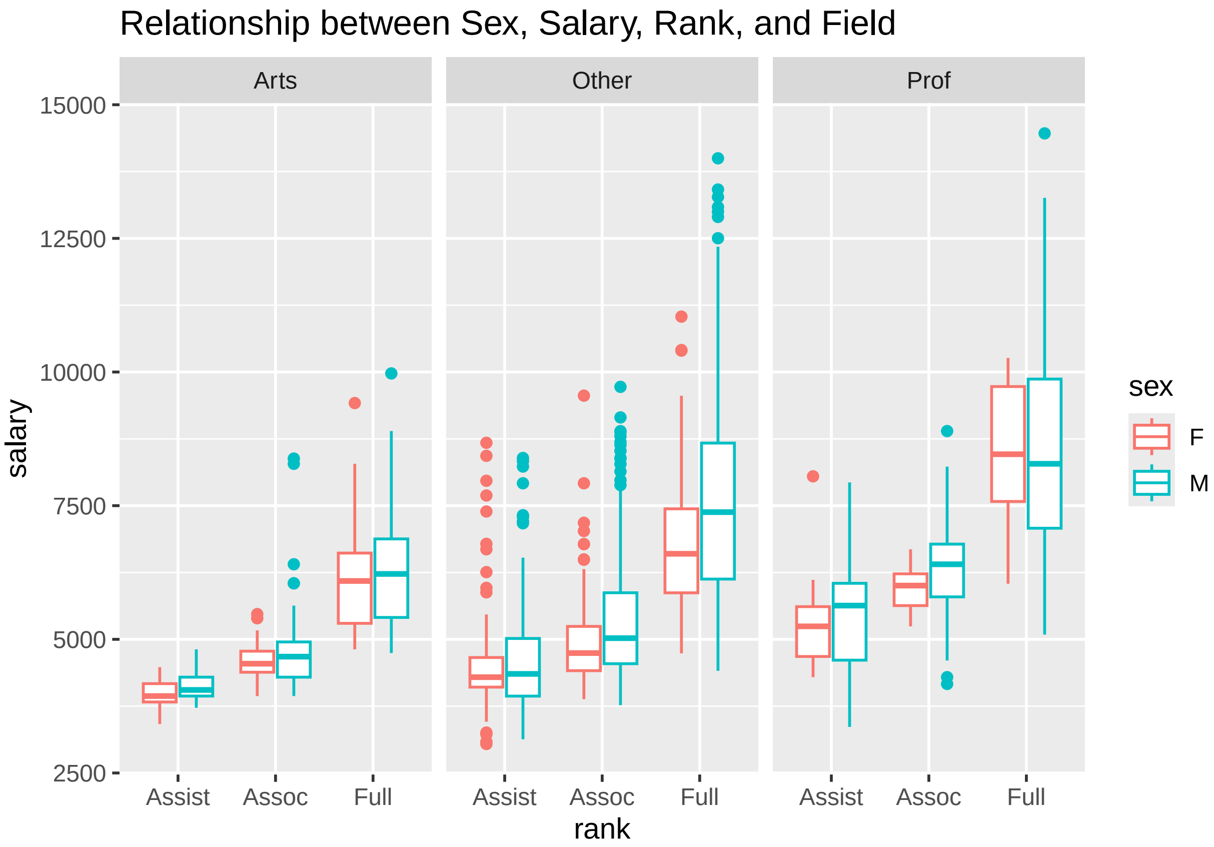


8. Comment on how the relationship between `sex` and `salary` changes for different values of your confounding variable in 1-2 sentences.

The salary gap between men and women changes with rank. Men generally earn more than women at each rank, but the gap is larger in higher positions.

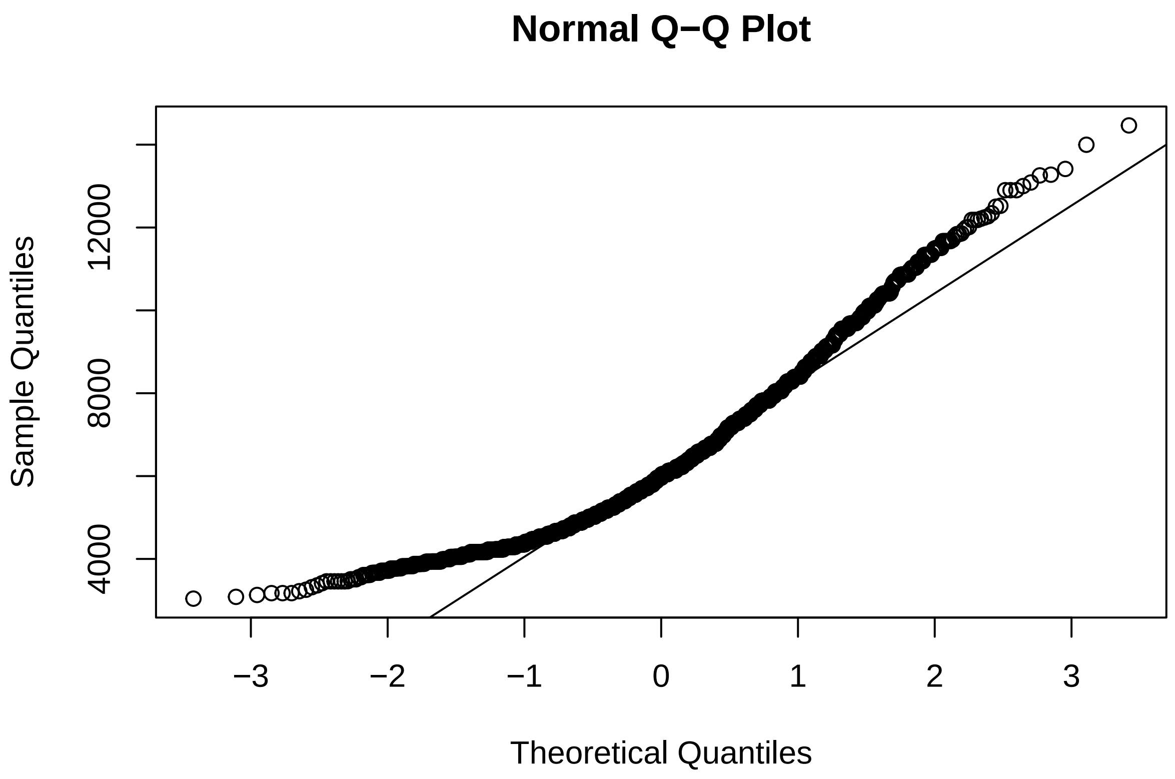
Challenge question: Visualize the relationship between `sex`, `salary`, and both of your confounding variables in a single plot.

```r
ggplot(salary, aes(x = rank, y = salary, color = sex)) +
  geom_boxplot() +
  facet_wrap(~ field)+
labs(title = "Relationship between Sex, Salary, Rank, and Field")
```



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$salary)
qqline(salary$salary)
```



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

```
proportion <- mean(salary$salary > 8000)
proportion
```

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

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$salary, na.rm = TRUE)
sd_salary <- sd(salary$salary, na.rm = TRUE)

probability <- 1 - pnorm(8000, mean = mean_salary, sd = sd_salary)
probability
```

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

The probability and proportion are different. The assumption of normality may not be accurate.

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

I have been working with data and making charts, but figuring out how different variables affect each other is still a bit tricky for me. I have been putting in my best effort and a lot of time on Homework Two. Thank Wendy for creating a lot of questions for us so that we can learn much from them. I will keep trying my best for more homework in the future.

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

E - Excellent