First Exercise

Lorenzo Meninato

1. Problem Statement: Do women earn less than men at academic institutions? If so, what factors account for the differences in pay.

2. Data:

The Salaries dataset from the “car” package contains 2008-09 academic salary data for professors (assistant, associate, and full professors) for a single college in the US. Data was collected as part of the college’s efforts to monitor salary differences between male and female professors.

3. Analysis:

My analysis will primarily be Level 1 and Level 2. I think there will not be enough data to causally infer any relationships between variable, and further, I would not have the subject matter knowledge to adequately perform that type of causal inference. Still, I believe that from observing the summary statistics and univariate statistsics/plots (Level 1 analysis) I might be able to find appropriate models (or at least obtain useful informations from the wrong models we might use) for the statistical relationships between variable (Level 2 analysis).

4. Univariate statistics and plots:

```
library(car)
library(ggplot2)

summary(Salaries)
```

```
##         rank     discipline yrs.since.phd    yrs.service        sex
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358
##  Prof     :266              Median :21.00   Median :16.00
##                             Mean   :22.31   Mean   :17.61
##                             3rd Qu.:32.00   3rd Qu.:27.00
##                             Max.   :56.00   Max.   :60.00
##      salary
##  Min.   : 57800
##  1st Qu.: 91000
##  Median :107300
##  Mean   :113706
##  3rd Qu.:134185
##  Max.   :231545
```

```
table(Salaries$sex, Salaries$rank)
```

```
##
##           AsstProf AssocProf Prof
##   Female        11        10   18
##   Male          56        54  248
```

There are a few concerns with the data. For starters, there is limited data on female professor salary (only 39 observations). This could be difficult to work with, especially since the main problem we are trying to resolve is the difference in salaries between females and males, ceteris paribus. Further, most professors are of the highest rank (266/397 observations), so if we group the sexes by rank and discipline, some group would have far too few observations. I also thought it would likely be that having both “years since phd” and “years of service” would be redundant and could lead to multicollinearity in later regressions. Statistics confirm this later. At first glance, the subdivision of professors into applied and theoretical fields does not appear to be particularly useful, since it would likely be more fruitful to divide professors by what subject matter they teach, rather than if it is a theoretical or applied field. We would expect business professors to earn more than english professors, if outside market forces would pay the former more than the latter.

1: Average years of service by rank

rank	yrs.service
AsstProf	2.373134
AssocProf	11.953125
Prof	22.815790

2: Median salaries by sex and rank

sex	rank	Salaries$salary
Female	AsstProf	77000.0
Male	AsstProf	80182.0
Female	AssocProf	90556.5
Male	AssocProf	95626.5
Female	Prof	120257.5
Male	Prof	123996.0

3: Average years since PhD by rank

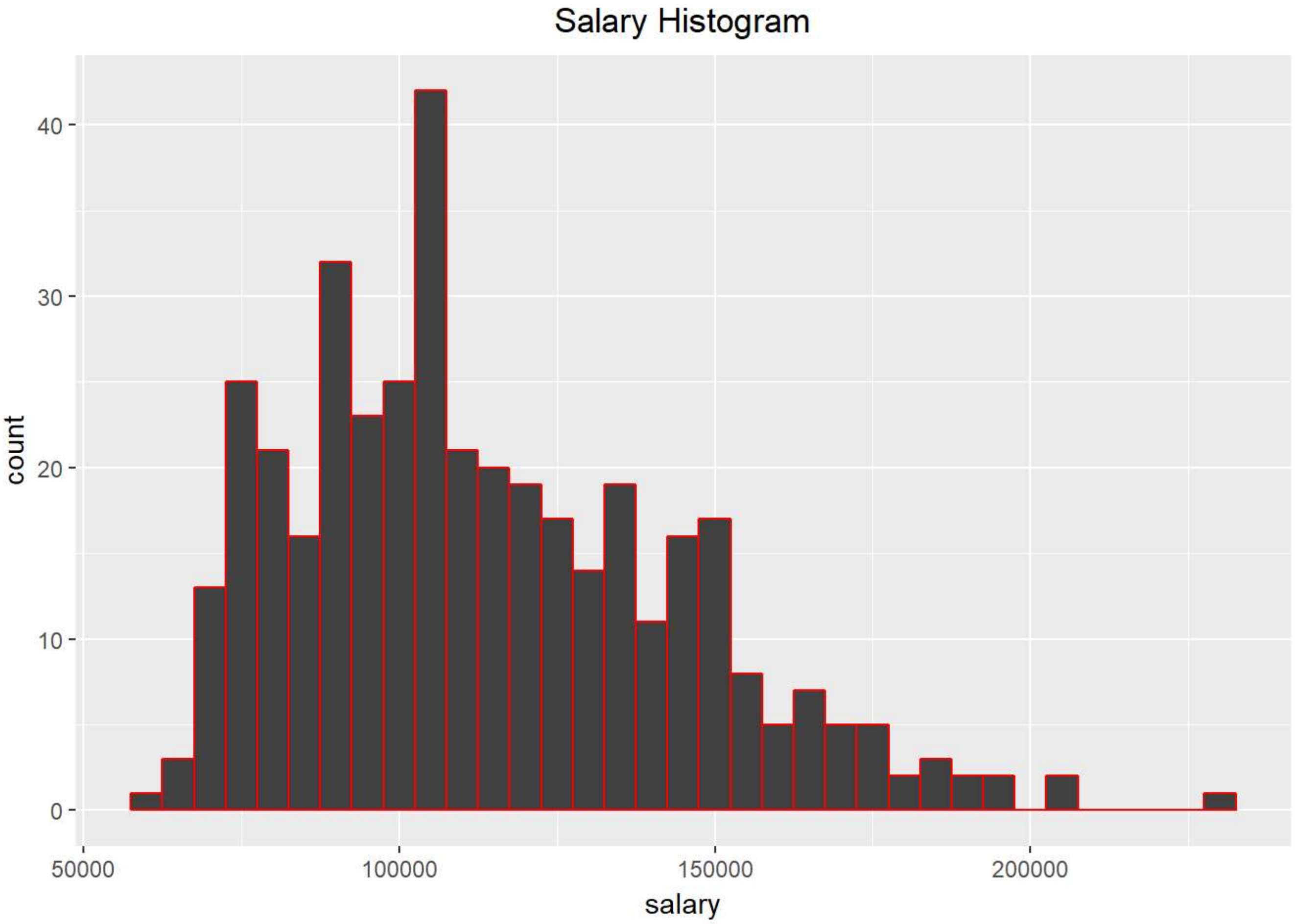
rank	yrs.since.phd
AsstProf	5.104478
AssocProf	15.453125
Prof	28.300752

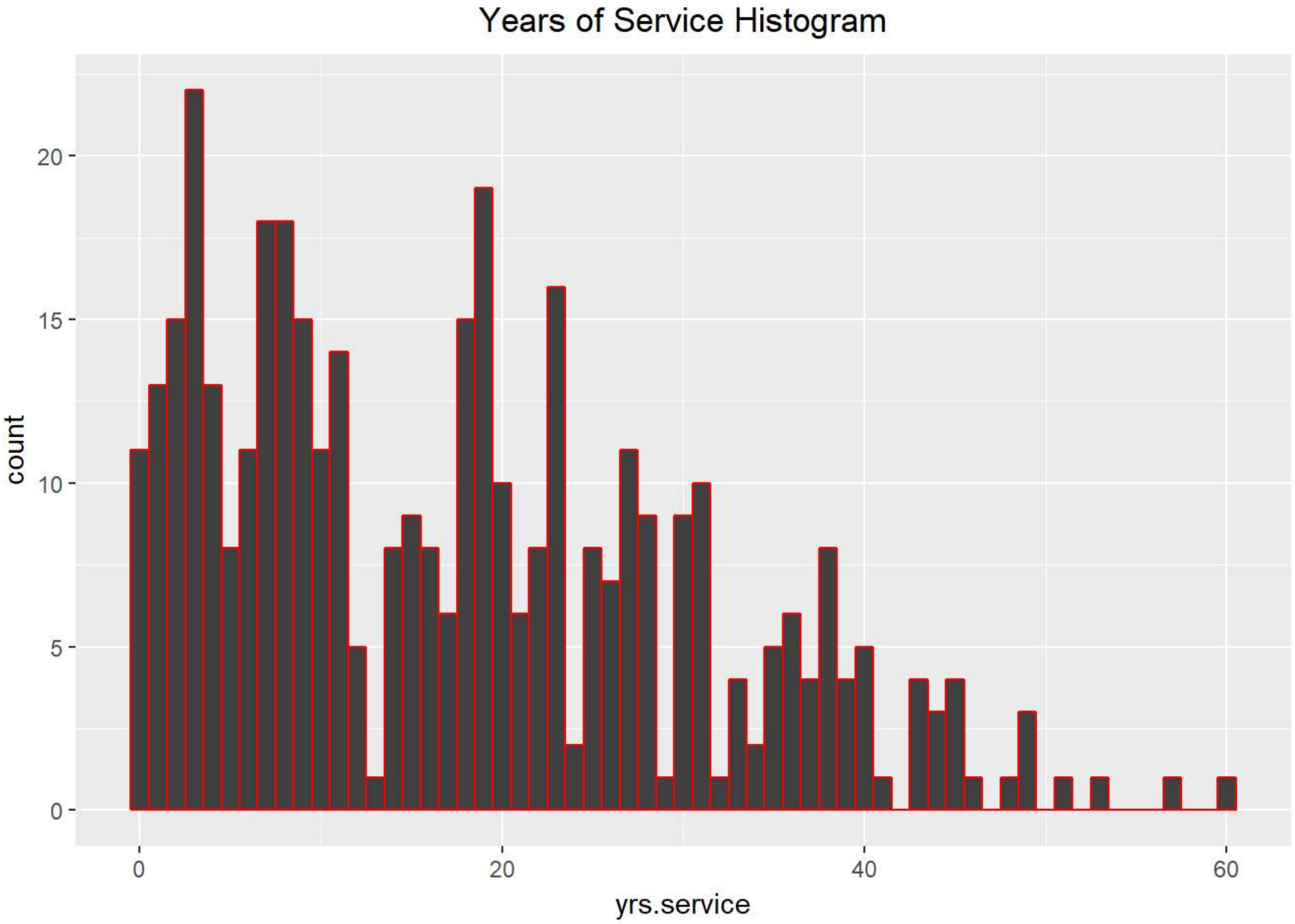
4: Median salary by sex and discipline

discipline	sex	salary
A	Female	78000
B	Female	105450

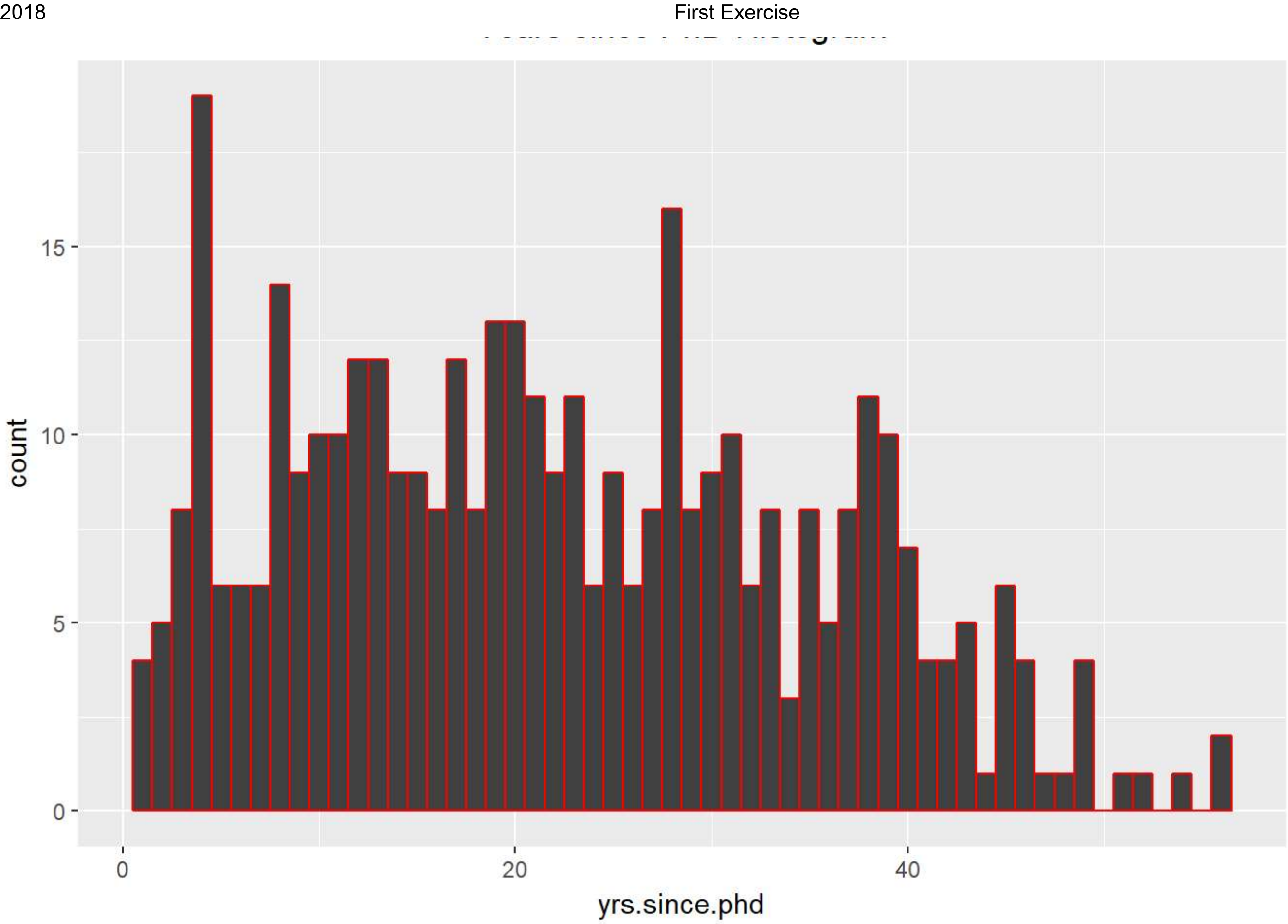
discipline	sex	salary
A	Male	105260
B	Male	113600

The first table is not surprising. The higher the rank of the professor, the more years they have been working at the college. The second table might confirm the college administrators suspicions that female professors were underpaid relative to their male peers. At every rank, female professors have a lower median salary than their male colleagues. However, this difference might be accounted for in differences in experience (“years since phd”) or discipline. Further, the fourth table suggests that applied disciplines tend to pay more than theoretical differences, by a median difference of about $9000. When subdividing the disciplines by sex, the median salary for females in theoretical fields is nearly $30000 less than in applied fields ($8000 difference). But again, there is no reason to believe this is statistical evidence of gender pay discrimination.





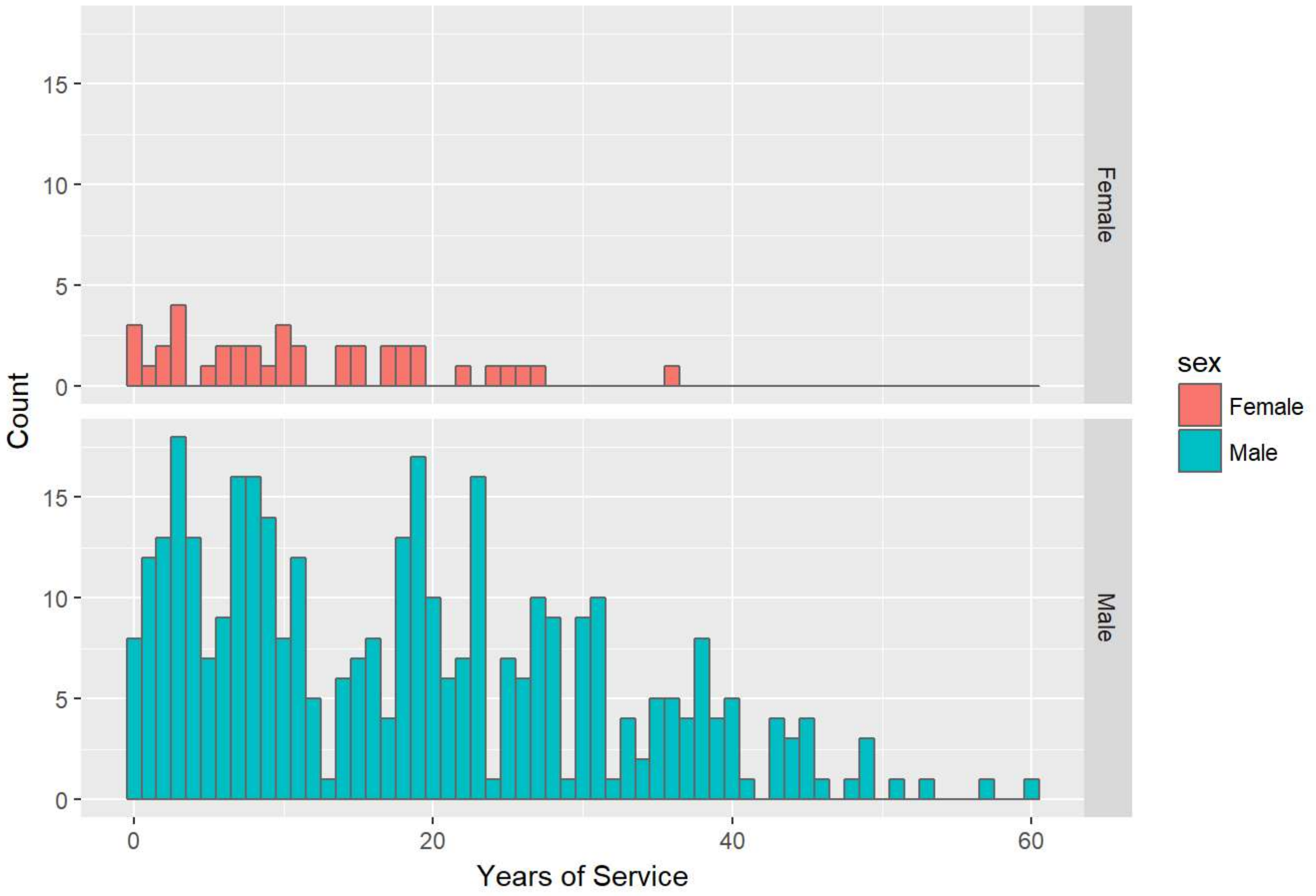
Years since PhD Histogram



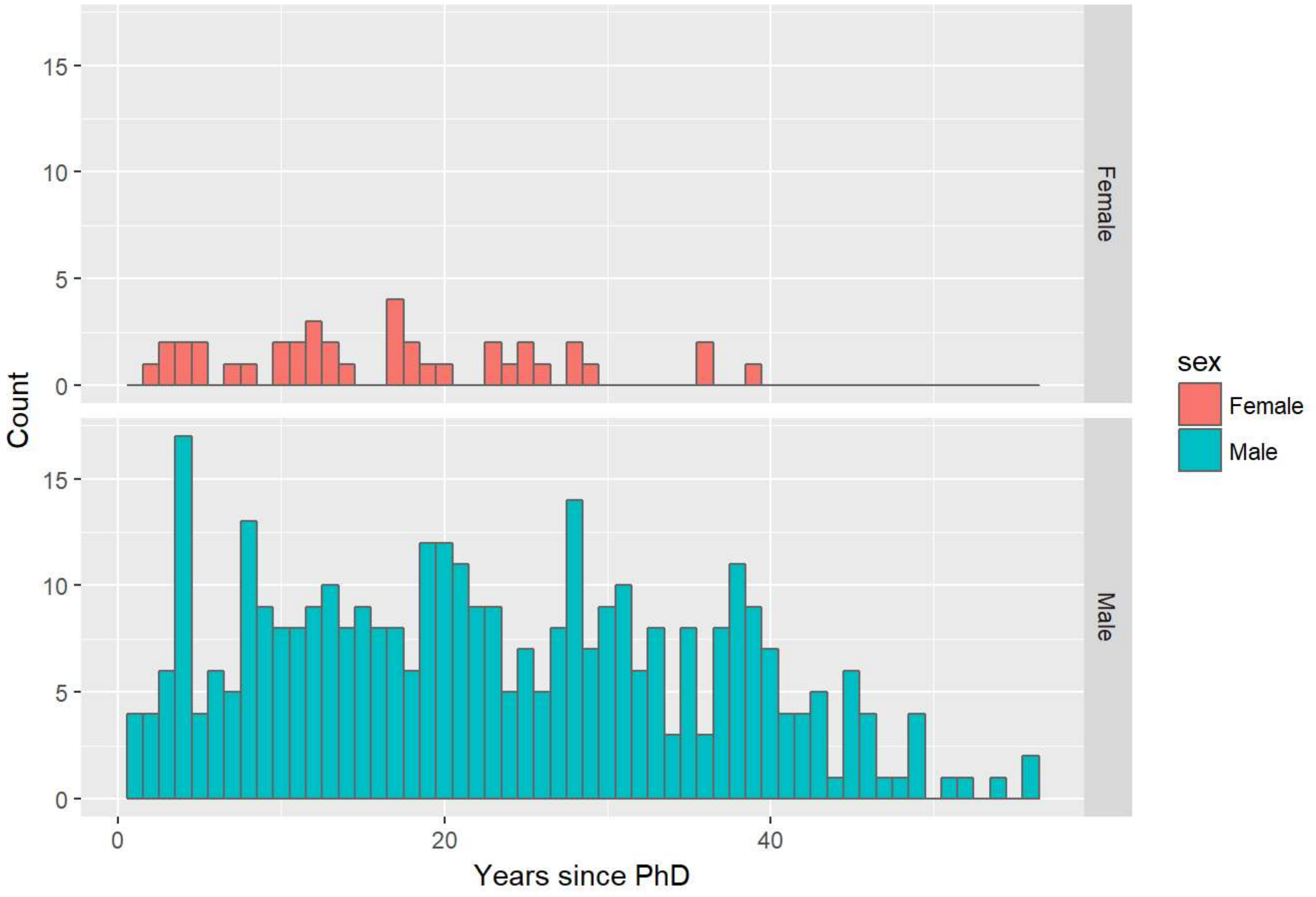
The salary distrbution is positively skewed. There seems to be certain clusters of salaries around certain pricepoints ($75k, $90k and $105k). If there is a bias for administrators to pay amounts near certain round numbers, this could effect regression results later. The “years since phd” data is a lot less right skewed than the years of service data. This could have been caused by the college hiring professors with prior experience.

5. Bivariate analysis:

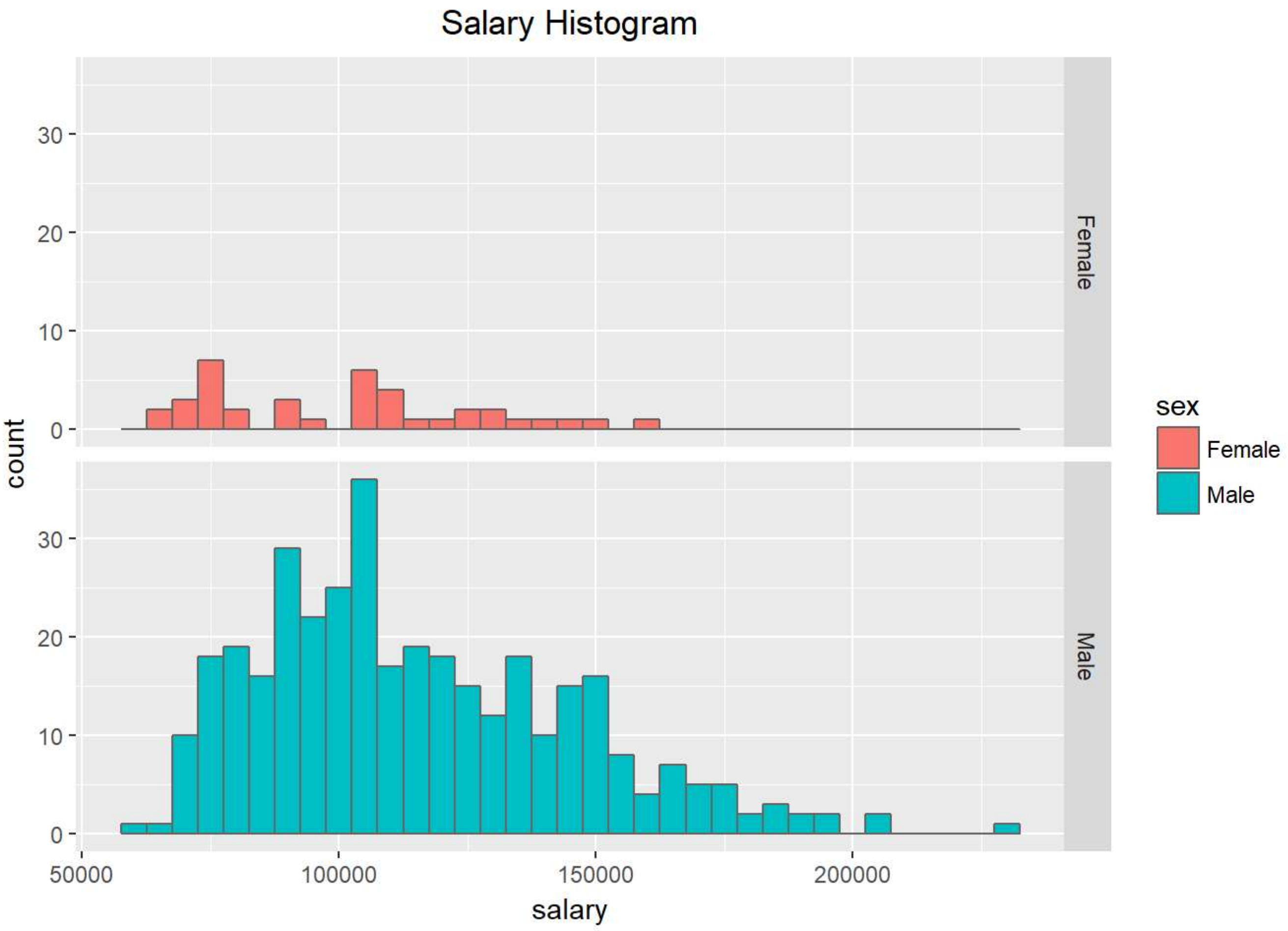
Years of service for male and female professors



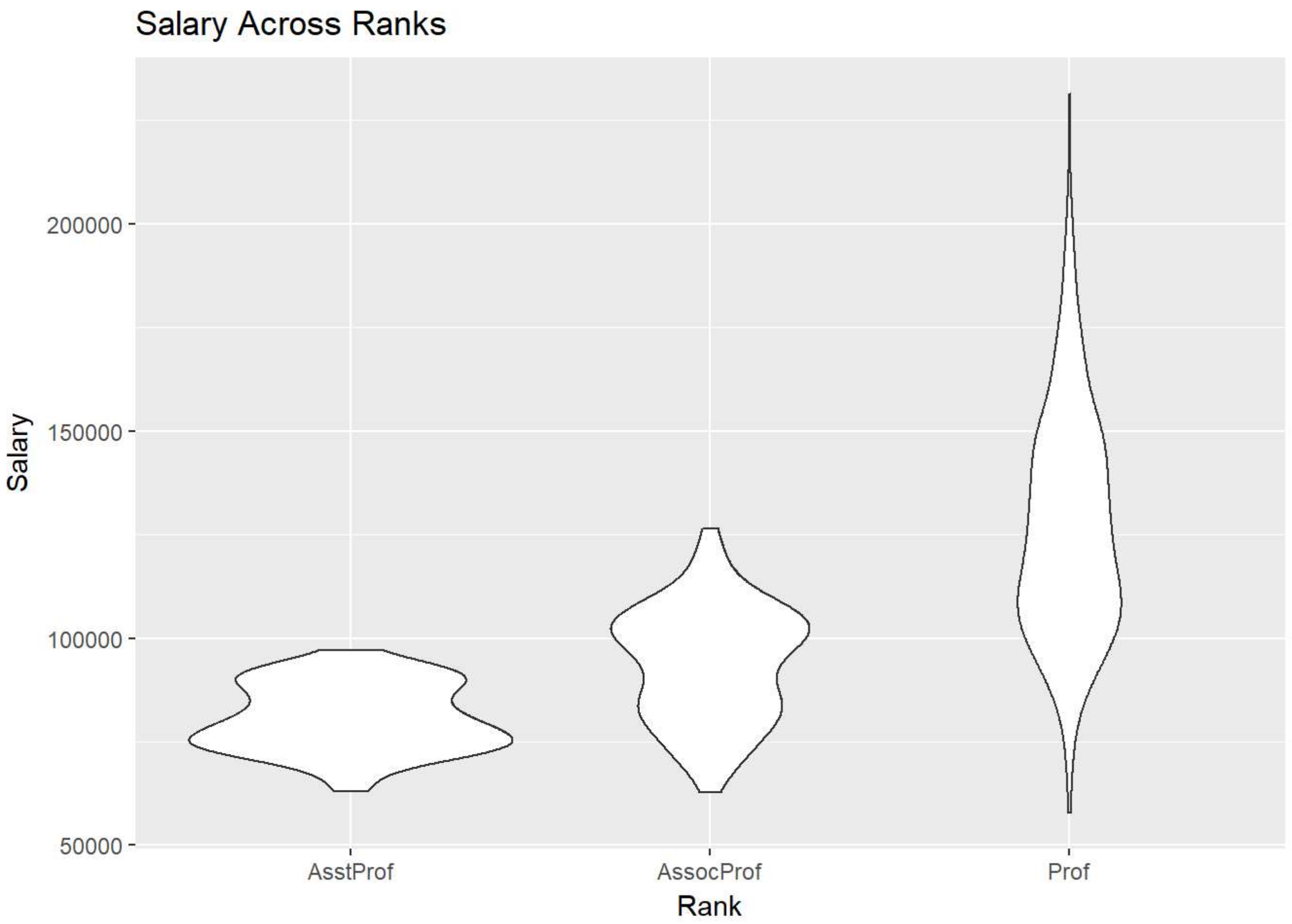
Years since PhD for male and female professors



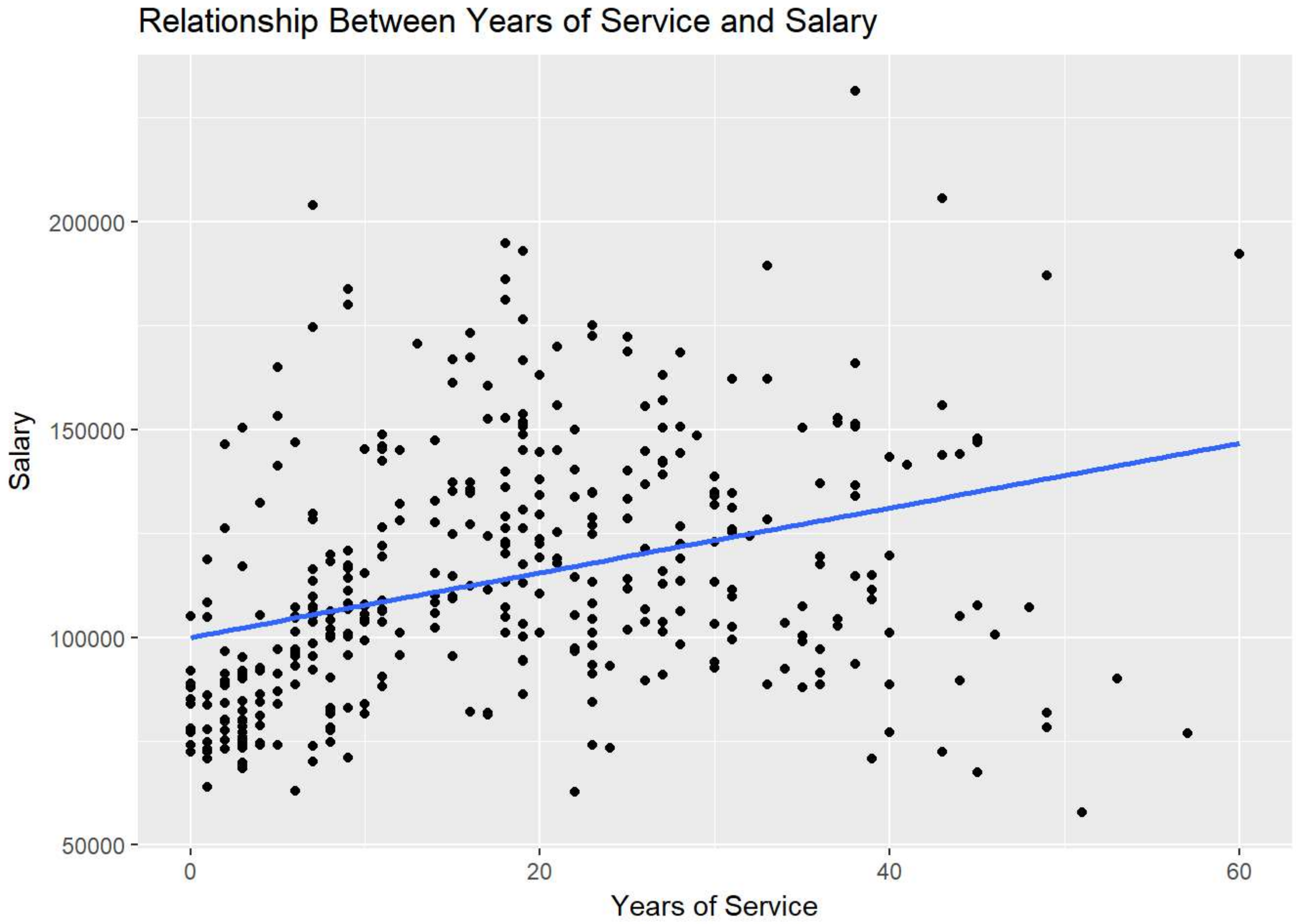
Histograms of “years since phd” and “years of service” visually suggest that the two variables are highly correlated. I think that it would be preferable to exclude “years of service” in a future report, since professors that have been researchers for many years at other schools would have less years of service, but likely could demand high salaries, since they have done much work since their PhD. Further, especially at the tail-end of these distributions, we can see that there is much less data for females beyond 25-30 years of experience. The histogram for female professors (for “years since phd”), is also nearly uniformly distributed, which we would not expect.



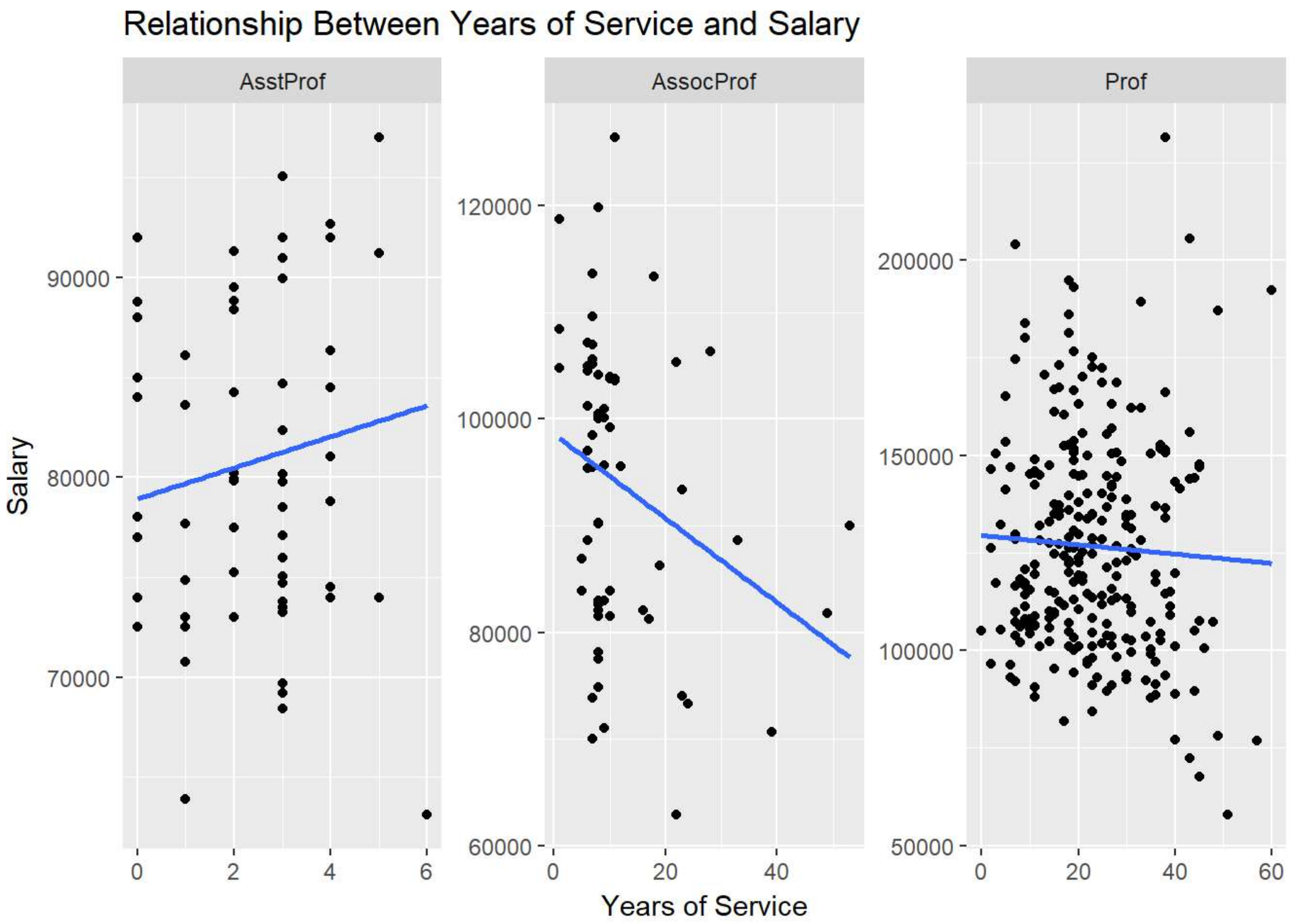
For females there is little data between $75k and $100k, with salaries tending to be clustered around those two boundaries. There could be a reason for this bimodal distribution. We do knot know why this is the case.



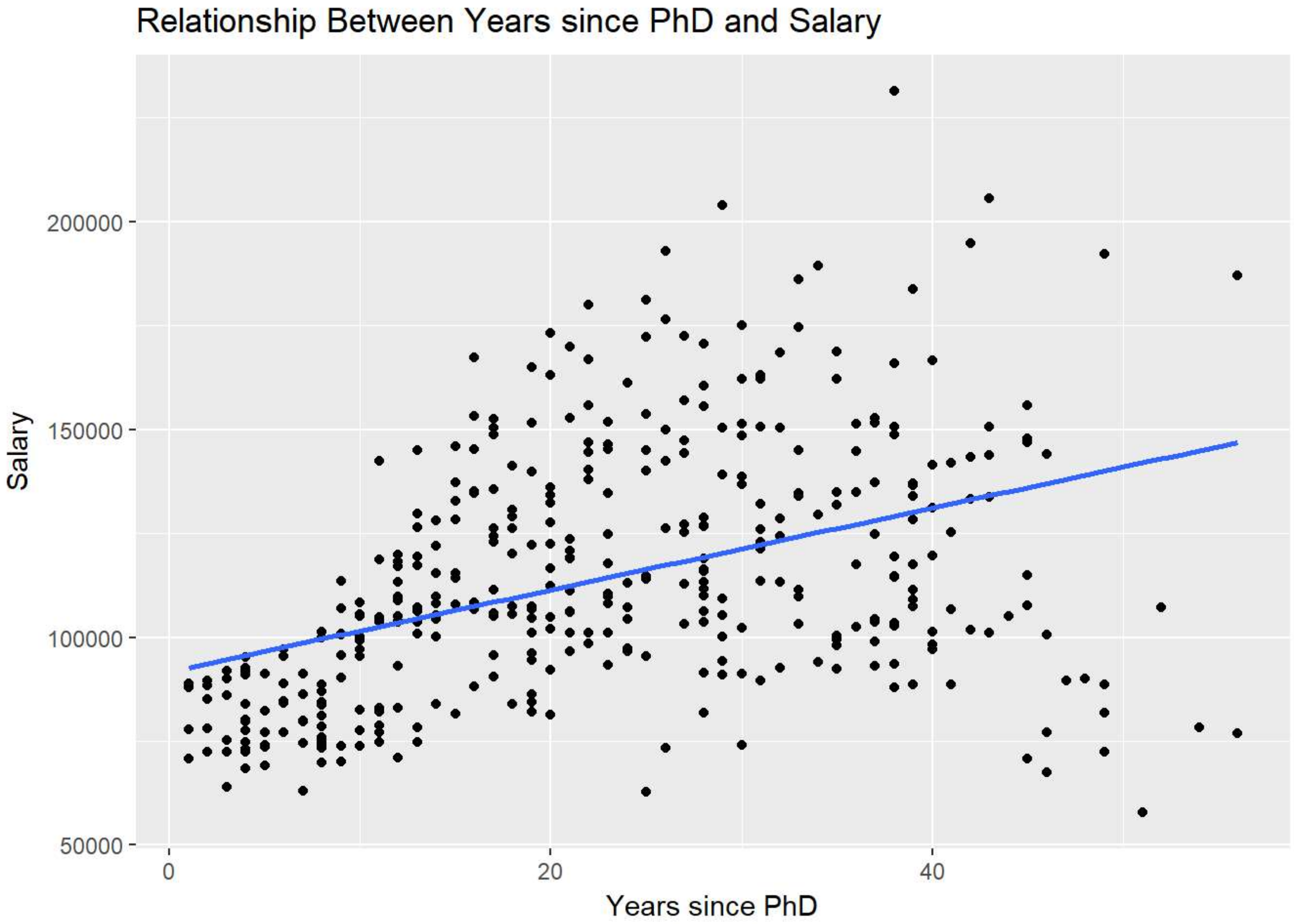
For the violin plot ("Salary across ranks"), note how the distribution of salaries is skewed right for assistant professors and full professors while it is skewed left for associate professors. This might be explained by the fact that assistant professors are more readily promoted to associate professor, and thus earn more, while associate professors might increase their salary over a longer period of time but would not be so easily promoted to full professor.



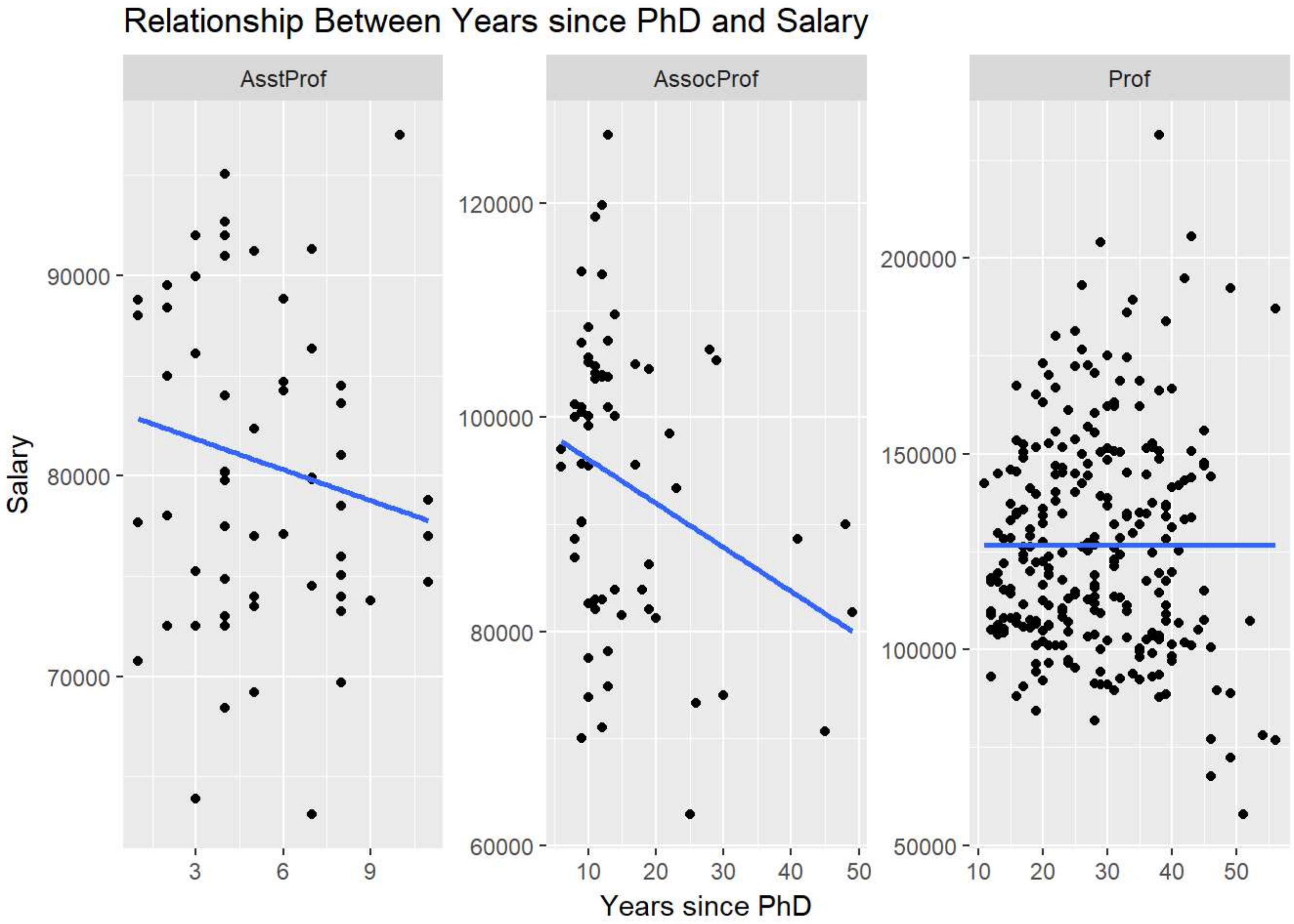
According the scatterplot ("Relationship between years of service and salary"), we see a weak positive correlation between the two variables. How linear their relationship is is unclear.



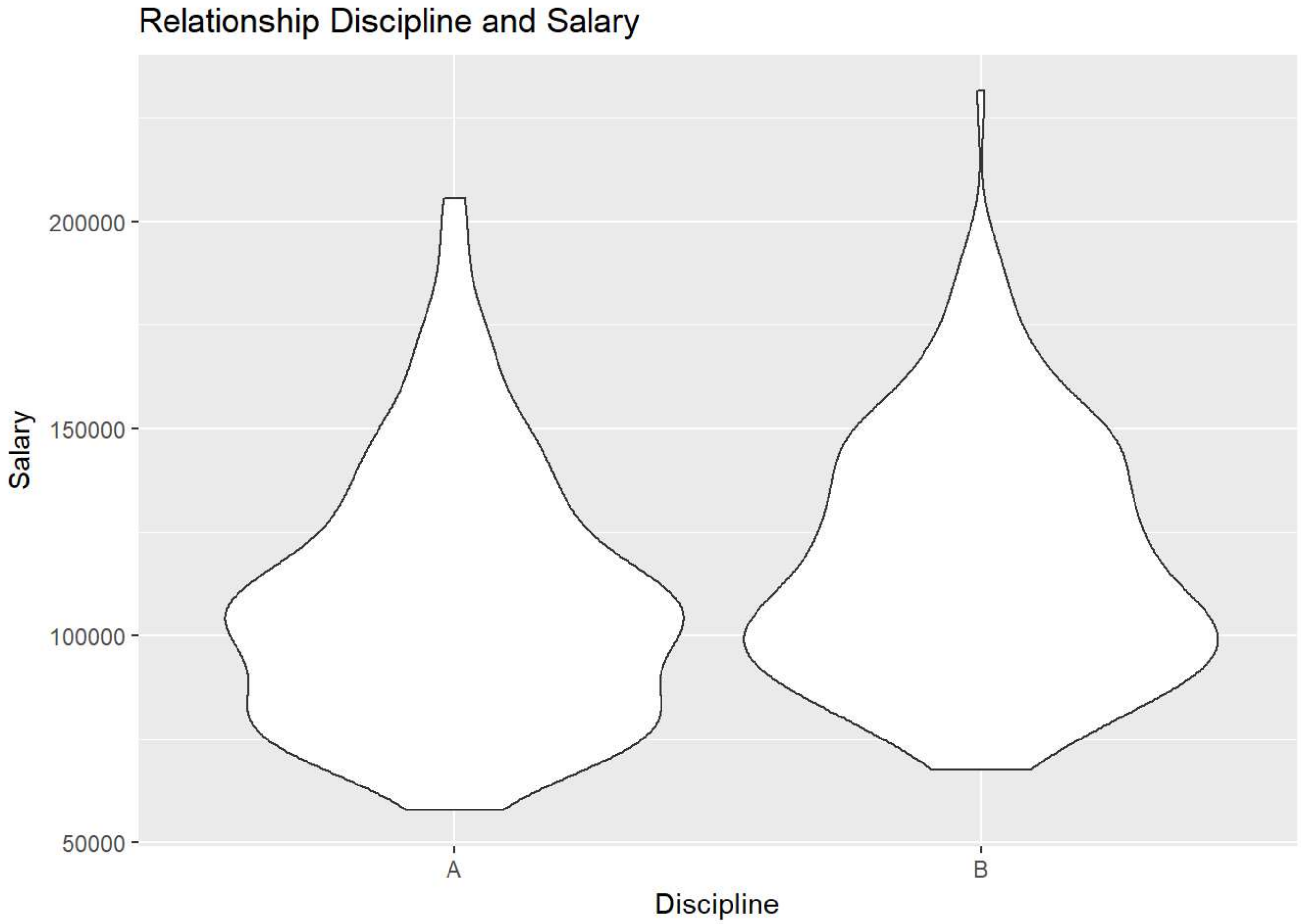
For more expereienced professors, having worked at the college longer does not seem to positively correlate with a higher salary. This might be because professors are not negotiating with other colleges for higher, more competitive salaries if they have worked at the same college for so long. And this cannot be the case for assistant professors that are probably much younger.



Using “years since PhD” we obtain similar results.



Here we see a near-zero or even negative relationship between “years since phd” and “salary”, when salaries are subdivided by rank. This result is fairly surprising. At least for the two lesser ranks, it might be that professors that are not promoted to full professors relatively soon after obtaining their PhD are less desirable, thus are paid less.



The applied discipline appears to have a slightly more right skewed distribution than its theoretical counterpart. This could be because of a few influential observations. For instance, certain fields like statistics and computer science are heavily desired by industry, so professors in those fields could demand a higher salary.

```
salarydata = Salaries[, c(3,4,6)]
cor(salarydata)
```

```
##               yrs.since.phd yrs.service    salary
## yrs.since.phd     1.0000000   0.9096491 0.4192311
## yrs.service       0.9096491   1.0000000 0.3347447
## salary            0.4192311   0.3347447 1.0000000
```

For our quantitative variables, we can see that “years of service” and “years since phd” are highly correlated as expected.

6. Training and test data:

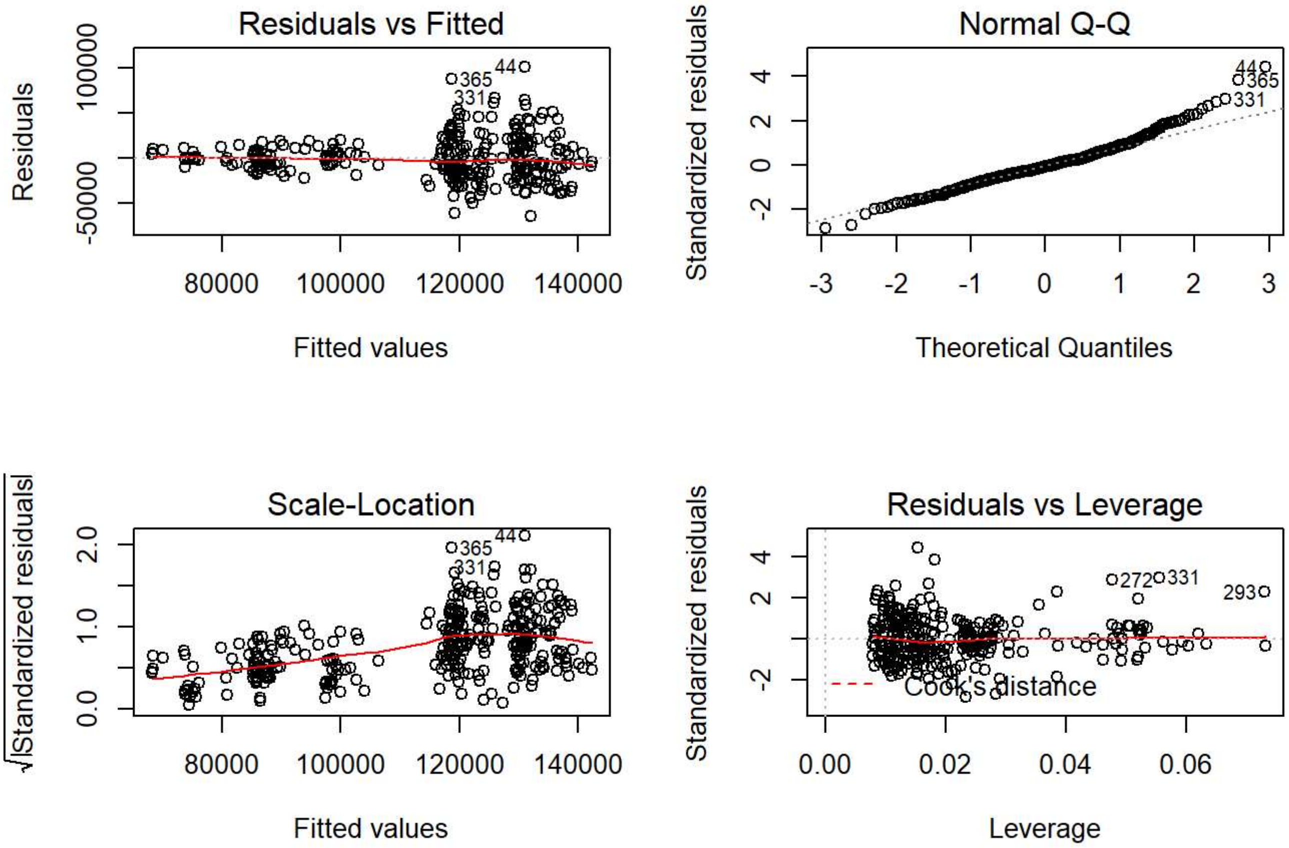
```
Index<-sample(1:397,80, replace=FALSE)
Test <- Salaries[Index,]
Train <- Salaries[-Index,]
```

7. Linear regression:

```
out1 <- lm(salary ~ ., data = Train)
summary(out1)
```

```
##
## Call:
## lm(formula = salary ~ ., data = Train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64479 -13303  -1736  11613 100568
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    66914.5     5340.0  12.531  < 2e-16 ***
## rankAssocProf  11949.8     4759.0   2.511   0.0125 *
## rankProf       43337.0     4856.3   8.924  < 2e-16 ***
## disciplineB    12672.2     2640.1   4.800 2.47e-06 ***
## yrs.since.phd    588.6      270.5   2.175   0.0303 *
## yrs.service     -521.1      239.1  -2.179   0.0301 *
## sexMale         5488.8     4652.8   1.180   0.2390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22930 on 310 degrees of freedom
## Multiple R-squared:  0.4315, Adjusted R-squared:  0.4205
## F-statistic: 39.21 on 6 and 310 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(out1)
```



```
vif(out1)
```

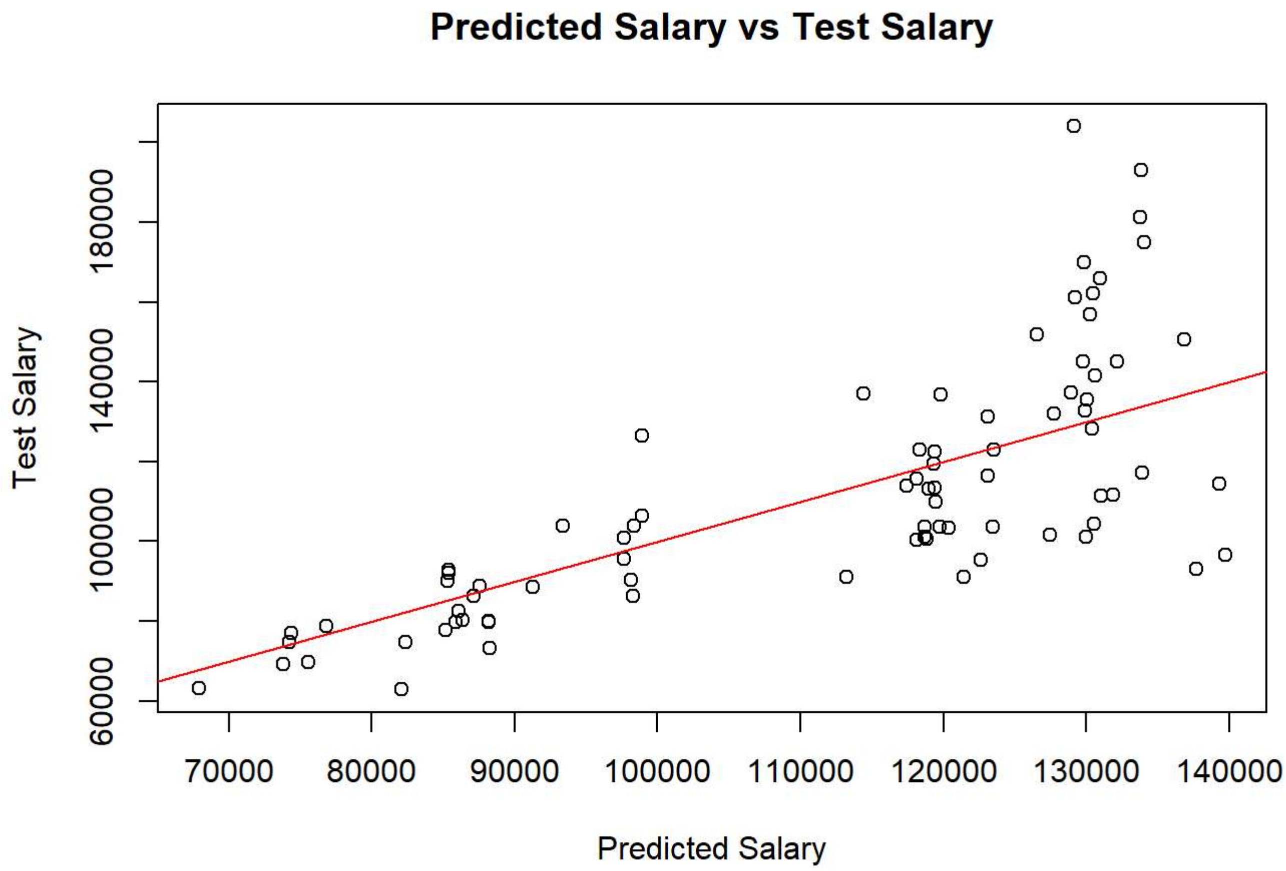
```
##                   GVIF Df GVIF^(1/(2*Df))
## rank          1.943108  2        1.180658
## discipline    1.043768  1        1.021650
## yrs.since.phd 7.549867  1        2.747702
## yrs.service   6.088660  1        2.467521
## sex           1.050774  1        1.025073
```

```
#hccm(out1)
#sqrt(hccm(out1))
```

The variance inflation factors do suggest that “years since phd” and “years of service” are multicollinear, as I suspected earlier. The quality of the fit is not great. Our adjusted R-squared value is 0.45. Examining the plot of residuals versus fitted points we see several outliers and heteroskedasticity. The normal Q-Q plot is decent for many observations, but again, the fit is not great.

Certain categorical variables are ommitted to prevent multicollinearity.

```
Preds1<-predict(out1,newdata=Test)
plot(Preds1, Test$salary, xlab = "Predicted Salary", ylab = "Test Salary", main = "Predicted Sal
ary vs Test Salary")
abline(0,1,col="red")
```



Here I applied the linear regression model from the training data onto the test data. We can see that the fit is not great. The same problem of heteroskedasticity remains.

8. Regression Results:

At a level 2 level, from our regression results we can infer that if a professor’s sex is male instead of female, we would expect an increase in salary of $3000 dollars. However, this result is not statistically significant. Without a different regression technique or analysis we could not say otherwise. This is different from what we could have inferred from our Level 1 analysis of the differences in salaries across groups. Further, the results of the regression suggest that only the rank and discipline of a professor matte (statistically, from this model’s perspective).

9. Conclusions:

A preliminary statistical analysis of the Salaries dataset finds that their is no statistical evidence for gender pay discrimination. We can conclude, with some confidence, that professor rank and discipline (applied vs theoretical) do affect professor salary. However, especially with respect to the overarching problem of gender pay discrimination, these results should be taken lightly since there are few observation for female professor salaries, the inclusion of both “years since phd” and “years of service” likely are causing multicollinearity, and it would probably be more fruitful to divide professor disciplines into subject matters, rather than a “theoretical vs applied” framework.