STAT 8030 Mini Project 1

Based on the study conducted by a researcher at a large public research university, we have to find out how salaries are different for different professors based on their gender, seniority, discipline, years since their PhD, and years since their service at the university. So, after importing the "Salaries.csv" file, we can see that there are 397 rows and 6 columns. These 6 columns are variables/factors.

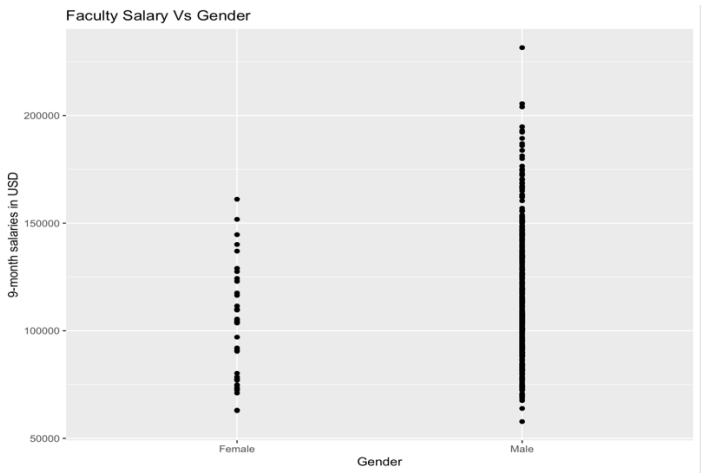
rank	rank of professor: assistant (AsstProf), associate (AssocProf), or full (Prof)
discipline	A=science and mathematics; B=engineering
years.since.phd	Number of years since PhD
years.service	Number of years employed at the University
gender	gender of the faculty member
salary	9-month salary in USD

This is what the dataset looks like in R studio after I've have imported it in my R Studio

	rank	discipline	yrs.since.phd	yrs.service	gender	salary
	<chr>	<chr>	<dbl>	<dbl>	<chr>	<dbl>
1	Prof	B	19	18	Male	139750
2	Prof	B	20	16	Male	173200
3	AsstProf	B	4	3	Male	79750
4	Prof	B	45	39	Male	115000
5	Prof	B	40	41	Male	141500
6	AssocProf	B	6	6	Male	97000
7	Prof	B	30	23	Male	175000
8	Prof	B	45	45	Male	147765
9	Prof	B	21	20	Male	119250
10	Prof	B	18	18	Female	129000

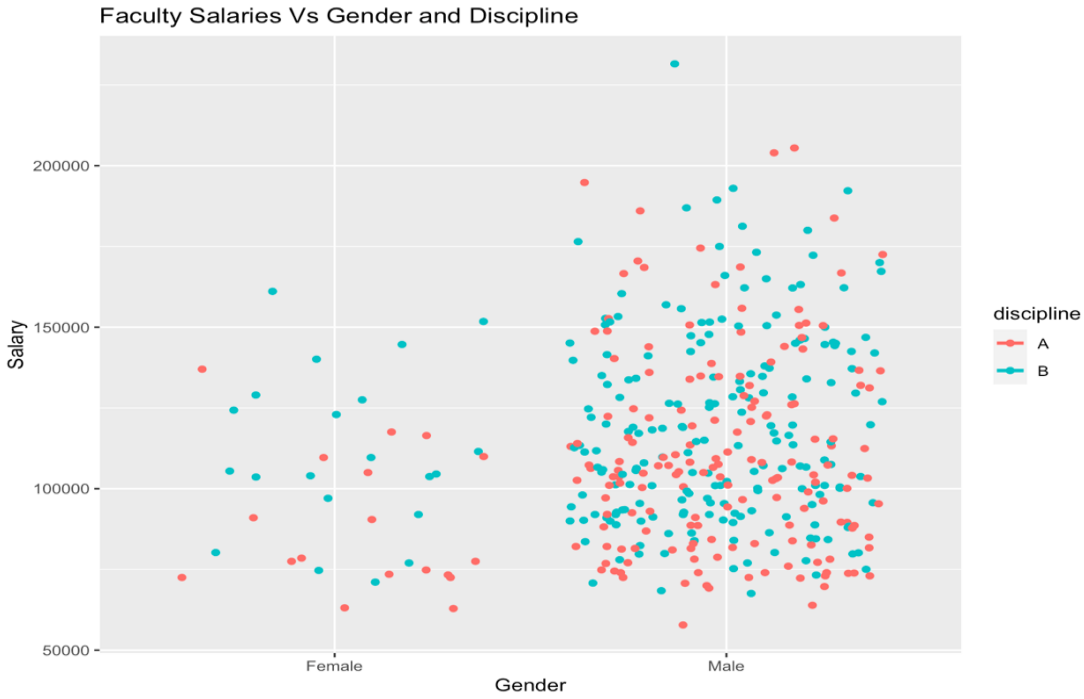
… with 387 more rows

When concerned to the key questions, I've evaluated the mean of Salaries based on gender first to get an idea of the differences in salaries before proceeding for other factors like "Seniority" and "Discipline". Please check the below screenshot to see that Male faculty earn more than female faculty. And I observed that the number of male faculty in the data are more than the number of female faculty. This could impact the data in some way or the other if we are interpreting the differences in salaries based on gender.
The mean salary of female faculty was found to be 101002.4 and the mean salary of male faculty was found to be 115090.4



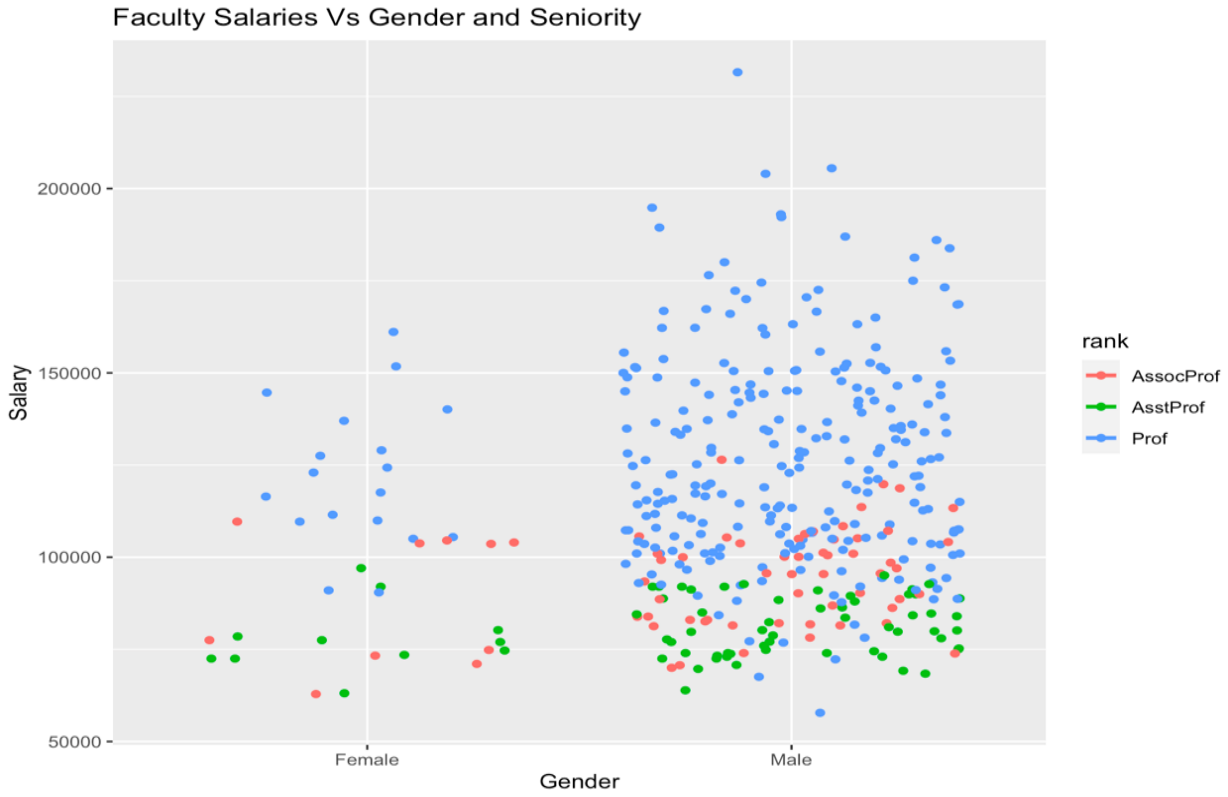
	Group.1	x
1	Female	101002.4
2	Male	115090.4

Let's compute similarly for Seniority and Discipline.
The mean salary of faculty of Discipline A is 108548.4 (Science and Mathematics) and the mean salary of faculty of Discipline B is 118028.7 (Engineering)
It means faculty belonging to Engineering Discipline tend to earn more than faculty belonging to Science and Mathematics field.
The mean salary of Professors is 126772.11, Associate Professors is 93876.44, and for Assistant Professors is 80775.99: So, it means Professors tend to earn more than any other faculty.
Now, after we have explicitly computed salaries based on discipline and rank/seniority, let's plot a graph for Salary of faculty Vs gender and discipline to see how much impact gender can make on the salaries of faculty based on discipline. Please check the screenshot below:



We can see that male faculty from discipline A, B are earning more than female from discipline A,B

Now, let's plot a graph for Salary of faculty Vs gender and rank/seniority to see how much impact gender can make on the salaries of faculty based on their rank. Please check the screenshot below:



We can see that even here, male professors are earning more than female professors and also, we can see there is a huge difference in the number of faculty when it comes to gender.

However, descriptive summary also needs an inferential analysis to support it. So, for the inferential analysis, we can consider:

- Null Hypothesis: U1-U2 = 0 [The faculty salaries are equal]
- Alternate Hypothesis: U1-U2 is not equal to 0 [The faculty salaries are not equal]
- Significance level to be 0.05

To support the hypothesis, I used "lm" function to fit the linear regression model
Here, in this, I have computed all the variables in the lm function, and got to know that the p-value is 2.2e-16 i,e., 0.00000000000000022
So, since the p-value is less than 0.05, we can reject the Null-Hypothesis.

I've calculated it for each variable that could affect the response variable(salary), still in all the cases, it rejects the Null-Hypothesis.

```
lm(formula = salary ~ ., data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-65248 -13211  -1775  10384  99592

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    78862.8     4990.3  15.803  < 2e-16 ***
rankAsstProf  -12907.6     4145.3  -3.114  0.00198 **
rankProf       32158.4     3540.6   9.083  < 2e-16 ***
disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
yrs.since.phd    535.1      241.0   2.220  0.02698 *
yrs.service     -489.5      211.9  -2.310  0.02143 *
genderMale      4783.5     3858.7   1.240  0.21584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22540 on 390 degrees of freedom
Multiple R-squared:  0.4547,    Adjusted R-squared:  0.4463
F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

Even if we compute just for gender, discipline, and rank, we get the following values

```
Call:
lm(formula = salary ~ rank, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-68972 -16376  -1580  11755 104773

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     93876       2954  31.777  < 2e-16 ***
rankAsstProf   -13100       4131  -3.171  0.00164 **
rankProf        32896       3290   9.997  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23630 on 394 degrees of freedom
Multiple R-squared:  0.3943,    Adjusted R-squared:  0.3912
F-statistic: 128.2 on 2 and 394 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = salary ~ gender, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-57290 -23502  -6828  19710 116455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   101002       4809  21.001  < 2e-16 ***
genderMale     14088       5065   2.782  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30030 on 395 degrees of freedom
Multiple R-squared:  0.01921,   Adjusted R-squared:  0.01673
F-statistic: 7.738 on 1 and 395 DF,  p-value: 0.005667
```

```
Call:
lm(formula = salary ~ discipline, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-50748 -24611  -4429  19138 113516

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   108548       2227  48.751  < 2e-16 ***
disciplineB     9480       3019   3.141  0.00181 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29960 on 395 degrees of freedom
Multiple R-squared:  0.02436,   Adjusted R-squared:  0.02189
F-statistic: 9.863 on 1 and 395 DF,  p-value: 0.001813
```

And also, no other career related variables have any effect on the salary difference between males and females. This could be improved if there were more female professors than male professors. Or maybe if it was somewhat specific.

Considering the collinearity:

We can compute it through VIF. We can automatically find it through loading “car” library and then using this command:

- `vif(lm.Salaries)`
- `cov2cor(vcov(lm.Salaries))`

This is the result we get:

```
> vif(lm.Salaries)
                 GVIF Df GVIF^(1/(2*Df))
rank         2.013193  2        1.191163
discipline   1.064105  1        1.031555
yrs.since.phd 7.518936  1        2.742068
yrs.service  5.923038  1        2.433729
gender       1.030805  1        1.015285
> cov2cor(vcov(lm.Salaries))
              (Intercept) rankAsstProf    rankProf disciplineB yrs.since.phd yrs.service   genderMale
(Intercept)     1.0000000  -0.50834213 -0.20960150 -0.34951896  -0.365469316  0.18441850 -0.620351243
rankAsstProf   -0.5083421   1.00000000  0.40072682  0.03837886   0.179587226 -0.02053019 -0.014799268
rankProf       -0.2096015   0.40072682  1.00000000 -0.04107711  -0.310498496  0.09415452 -0.067834101
disciplineB    -0.3495190   0.03837886 -0.04107711  1.00000000   0.181527164 -0.08873051 -0.030135563
yrs.since.phd  -0.3654693   0.17958723 -0.31049850  0.18152716   1.000000000 -0.85164037  0.001429303
yrs.service     0.1844185  -0.02053019  0.09415452 -0.08873051  -0.851640370  1.00000000 -0.048710271
genderMale     -0.6203512  -0.01479927 -0.06783410 -0.03013556   0.001429303 -0.04871027  1.000000000
```

As we know that VIF level of 10 indicates high collinearity and around 5-10 indicates moderate and below 5 indicates low. So, we can see that rank, discipline, and gender have low collinearity. As these are the variables that could help predict the salaries and having a low collinearity, whereas variables like yrs.since.phd and yrs.service have moderate collinearity, it could hamper the analysis.

After the analysis, we can say that the data was not collected by rigorous sampling method. It was done based on the publicly available data, both the ways of collecting the data i.e. from public database and web was not authenticated. The ratio of female vs male faculty members was also not 1:1. Instead, the data could have been more accurately collected if it was directly given out by the faculty members themselves anonymously with more parameters involving their background such as, their passport issuing country, their visa status, their work authorization in the country etc. The data collected should have been of the ratio 1:1 in terms of both female, male faculty and Mathematics, Science & Engineering departments faculty.