Exploratory Data analysis using R

Dimitris Zerkelidis

STUDENT ID : 03400049

EMAIL : dimzerkes@gmail.com

Dataset: phD salaries 2008-9

Reading data and descriptive statistics

This part of the code, is used to read the data from the .csv file that we are given for the phD salaries. After reading the file,then we are printing the first rows to see how our data looks.

```
  #reading csv salaries
  df<- read.table(path)
  head(df)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 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
```

Now we need to ensure that the variables "rank", "discipline" and "sex" are type of factors. This, is going to help us in plotting later and also give us a baseline that is useful when comparing with different levels of factors. The other columns are type of integers. This can be done by the command below which will describe us the variables and their type of class.

```
str(df)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AssocProf","AsstProf",..: 3 3 2 3 3 1 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

The above results can be interpreted as follows:

1. Indeed, rank , discipline and sex are classes of factors.
2. rank column has levels, AssocProf , AsstProf, Prof with baseline AssocProf.
3. discipline column has levels A , B with baseline A.
4. Finally, sex has levels Female, and Male with baseline Female.
5. Salary is an integer
6. years.since.phd & yrs.service is also an integer
7. we have 397 observations

It is useful before trying to discover some "hidden" patterns of our data by plotting, to first see some of the basic descriptive statistic numbers that is being provided by the function summary.

```
#some statistics for our DF
summary(df)
```

```
##         rank     discipline yrs.since.phd    yrs.service        sex
##  AssocProf: 64   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
##  AsstProf : 67   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
```

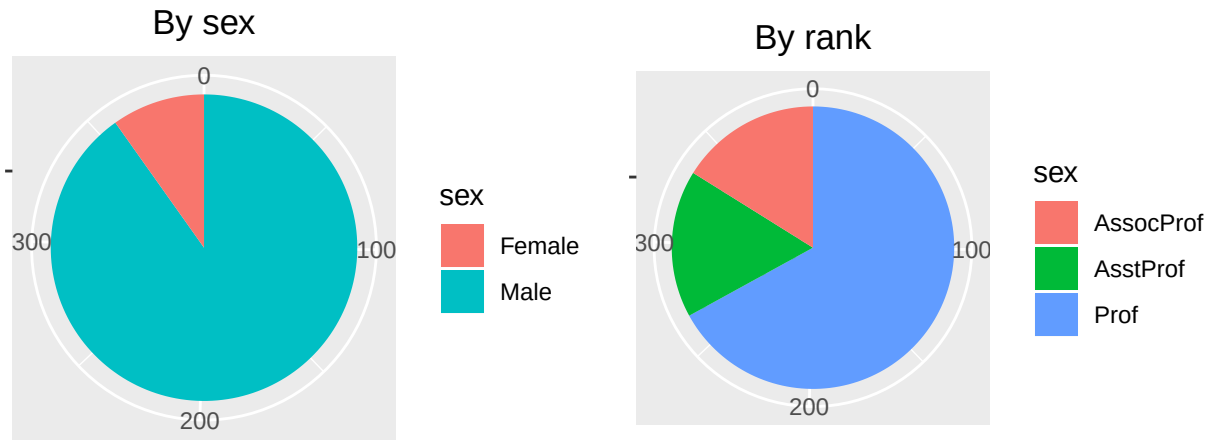
It is worth noticing that, from our 397 entries of our dataset , only 39 are about females and the other 358 are males. Also , more than half of the entries are about professors with discipline being almost balanced for both applied and theoritical departments.

In regards to our integer variables we can see their quantile values.

Below we can see the pie charts so that we can observe graphically what the numbers described us before.

```
pie1 <- ggplot(df, aes(x = "", fill = sex)) +
geom_bar(width = 1) +
theme(axis.line = element_blank(),
plot.title = element_text(hjust=0.5)) +
labs(fill="sex",
x=NULL,
y=NULL,
title="By sex")


pie2 <- ggplot(df, aes(x = "", fill = rank)) +
geom_bar(width = 1) +
theme(axis.line = element_blank(),
plot.title = element_text(hjust=0.5)) +
labs(fill="sex",
x=NULL,
y=NULL,
title="By rank")



plot_grid(pie1 + coord_polar(theta = "y", start=0), pie2 + coord_polar(theta = "y", start=0), labels =
```

A **B**



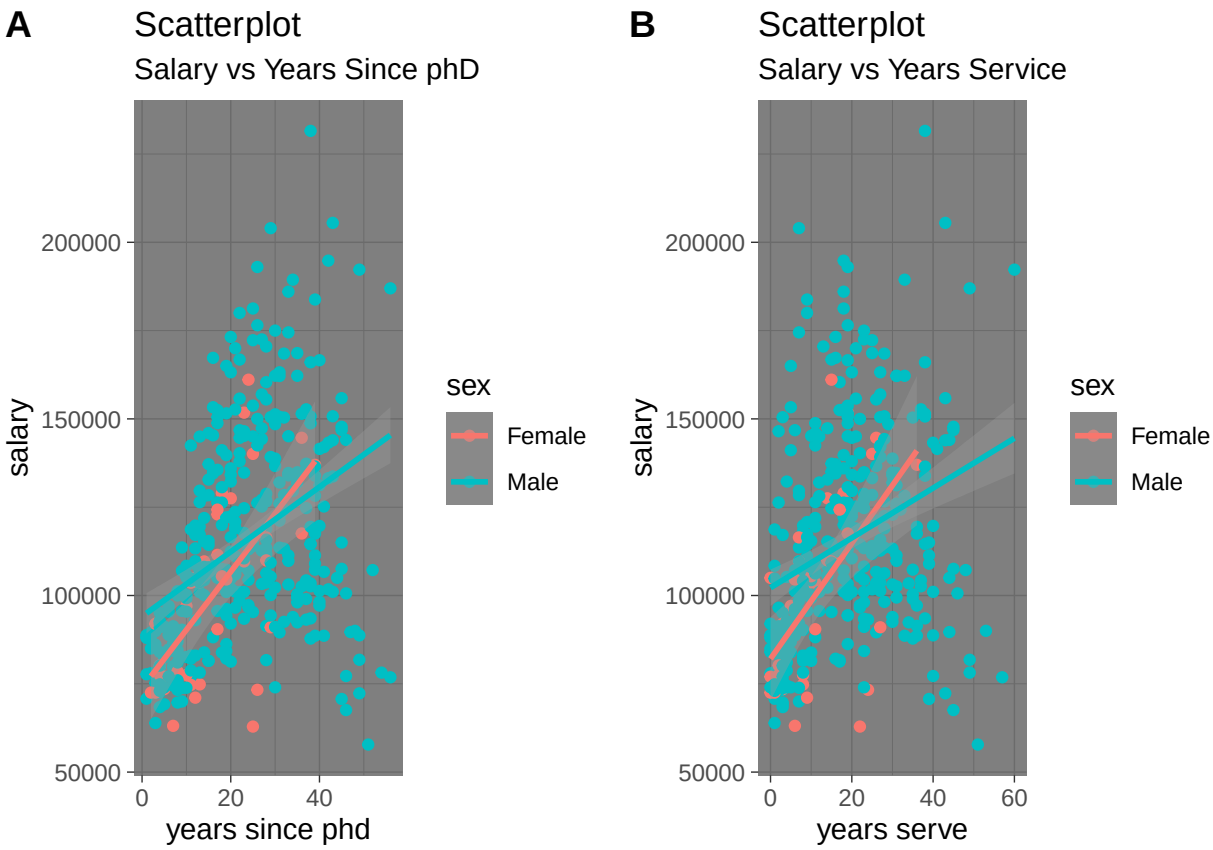
Correlation plots

Salary ~ Years by sex

In this section, we are going to plot years since phd and years service with salary. Also we are grouping each point by gender, so we can see the interaction of points that belong to male and to female in the salary as years are increasing.

The left plot is with years since phd, while the second plot is with years servicing. Both plots have a fitted line for each gender with the linear model method. The lines in our plots tell us that in both genders the bigger amount of years(either phd or service) they are in the field, the more money they get as a salary.

```
gg <- ggplot(df, aes(x = yrs.since.phd, y = salary)) +
geom_point(aes(col=sex)) + labs(subtitle="Salary vs Years Since phD",
y="salary",
x="years since phd",
title="Scatterplot") + geom_smooth(aes(group=sex,col=sex),method="lm") + theme_dark()

gg2 <- ggplot(df, aes(x = yrs.service, y = salary)) +
geom_point(aes(col=sex)) + labs(subtitle="Salary vs Years Service",
y="salary",
x="years serve",
title="Scatterplot")  + geom_smooth(aes(group=sex,col=sex),method="lm") + theme_dark()

plot_grid(gg, gg2, labels = "AUTO")
```

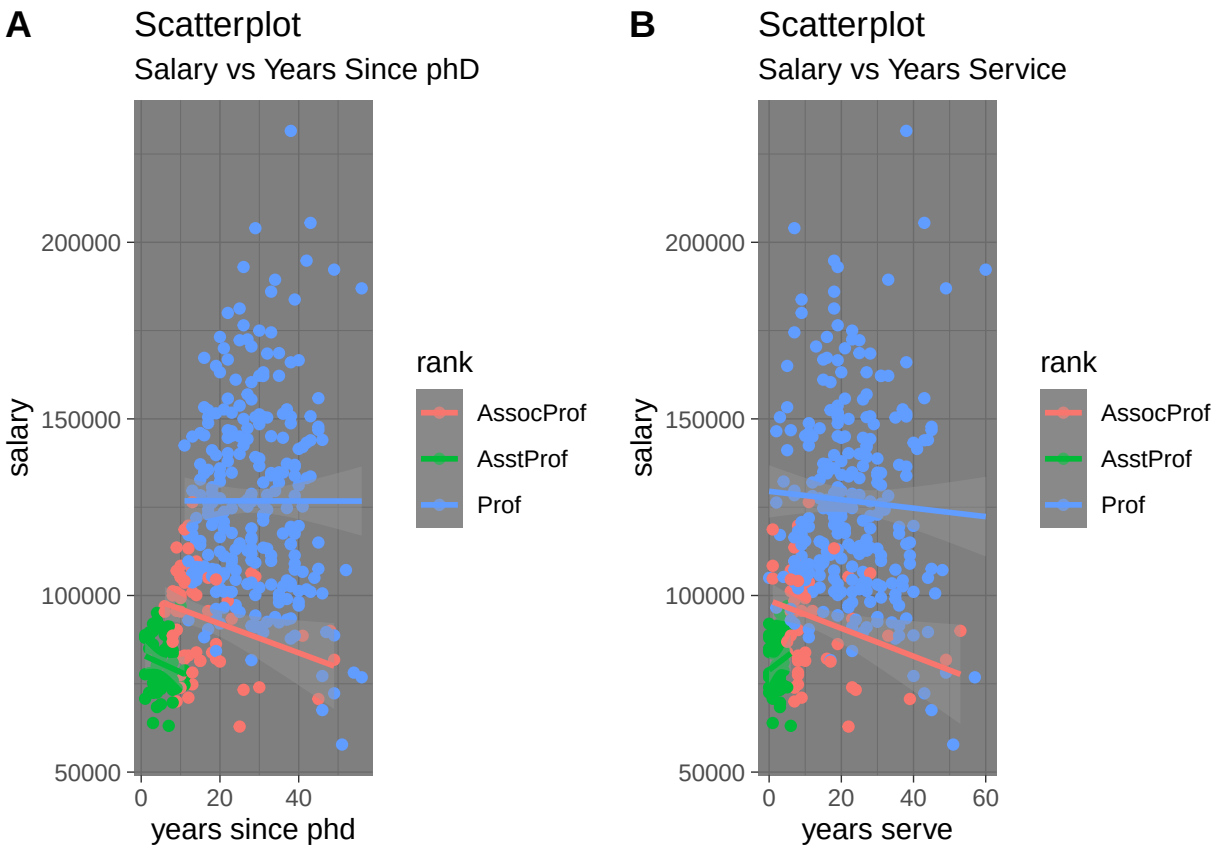


Salary ~ Years by rank

We can create the same scatterplot except now we can assign a color to each point by ranking only. In that case, we will see how the salary is distributed between the ranks.

Here we can notice something strange. The fitted line from the linear model we used has a negative slope. Thus, we can see that as years pass the salary increases but until one point. After that, the salary is being decreased. Maybe, this fact means that the oldest people of the faculty start to get lower salary than before while the younger in age get more.

```
gg3 <- ggplot(df, aes(x = yrs.since.phd, y = salary)) +
geom_point(aes(col=rank)) + labs(subtitle="Salary vs Years Since phD",
y="salary",
x="years since phd",
title="Scatterplot") + geom_smooth(aes(group=rank,col=rank),method="lm") + theme_dark()

gg4 <- ggplot(df, aes(x = yrs.service, y = salary)) +
geom_point(aes(col=rank)) + labs(subtitle="Salary vs Years Service",
y="salary",
x="years serve",
title="Scatterplot")  + geom_smooth(aes(group=rank,col=rank),method="lm") + theme_dark()

plot_grid(gg3, gg4, labels = "AUTO")
```



Deviations

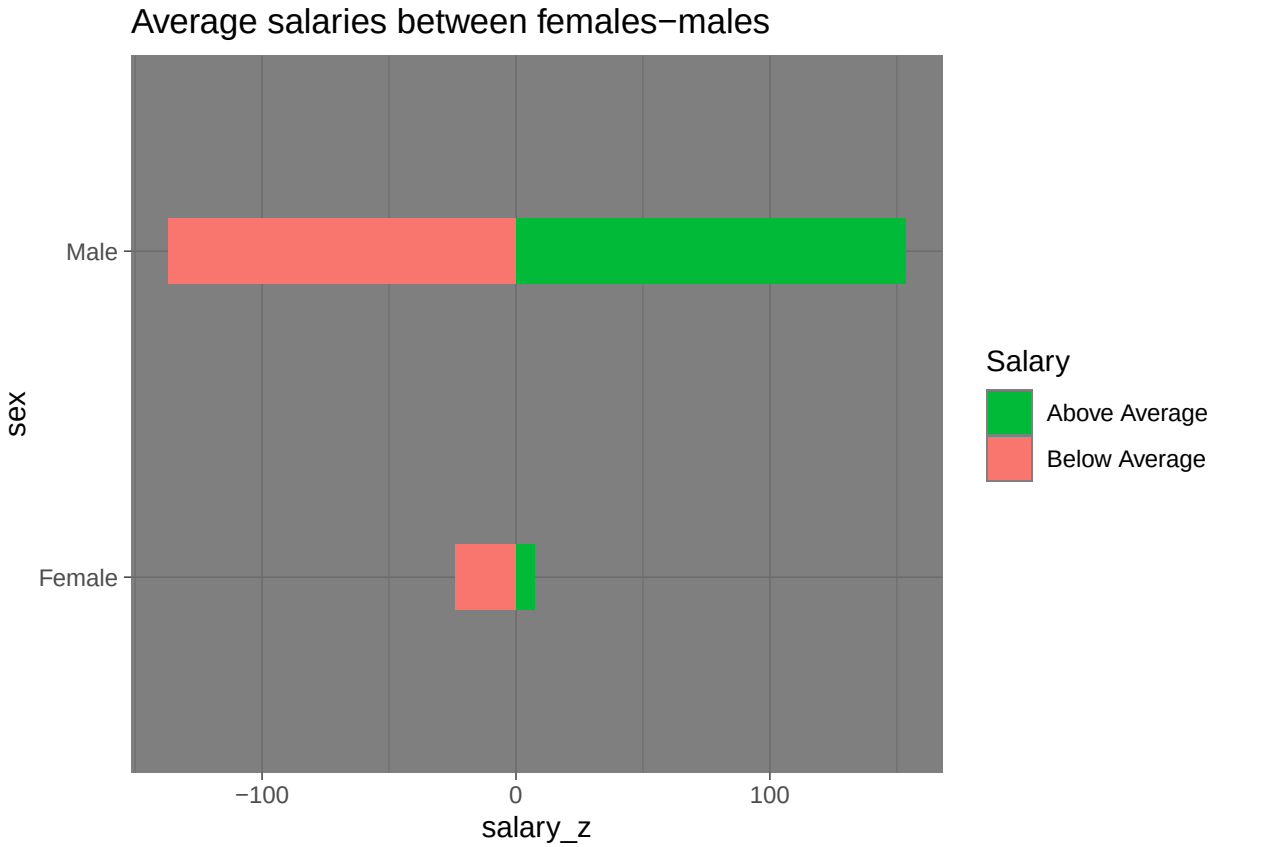
By the deviation charts we would like to observe the difference between the two genders in earning the avg. salary and also inspect the difference in avg. salary and rankings (prof, assocProf, assistProf).

AVG_SALARY VS GENDER

Below we can see that more women have salaries below average than above whilst in men it is balanced.

```
df$salary_type <- rownames(df)
df$salary_z <- round((df$salary - mean(df$salary)) / sd(df$salary),2)

df$salary_type <- ifelse(df$salary_z < 0, "below","above")

df <- df[order(df$salary_z),] #sort

## convert to factor to retain sorted order in plot
df$salary_type <- factor(df$salary_type, levels=rev(unique(df$salary_type)),ordered=TRUE)

#diverging barcharts
ggplot(df, aes(x=sex, y=salary_z, label=salary_z)) +
geom_bar(stat='identity', aes(fill=salary_type), width=.2) +
```

```
scale_fill_manual(name="Salary",
labels = c("Above Average", "Below Average"),
values = c("above"="#00ba38", "below"="#f8766d")) +
labs(title= "Average salaries between females-males") +
coord_flip() + theme_dark()
```

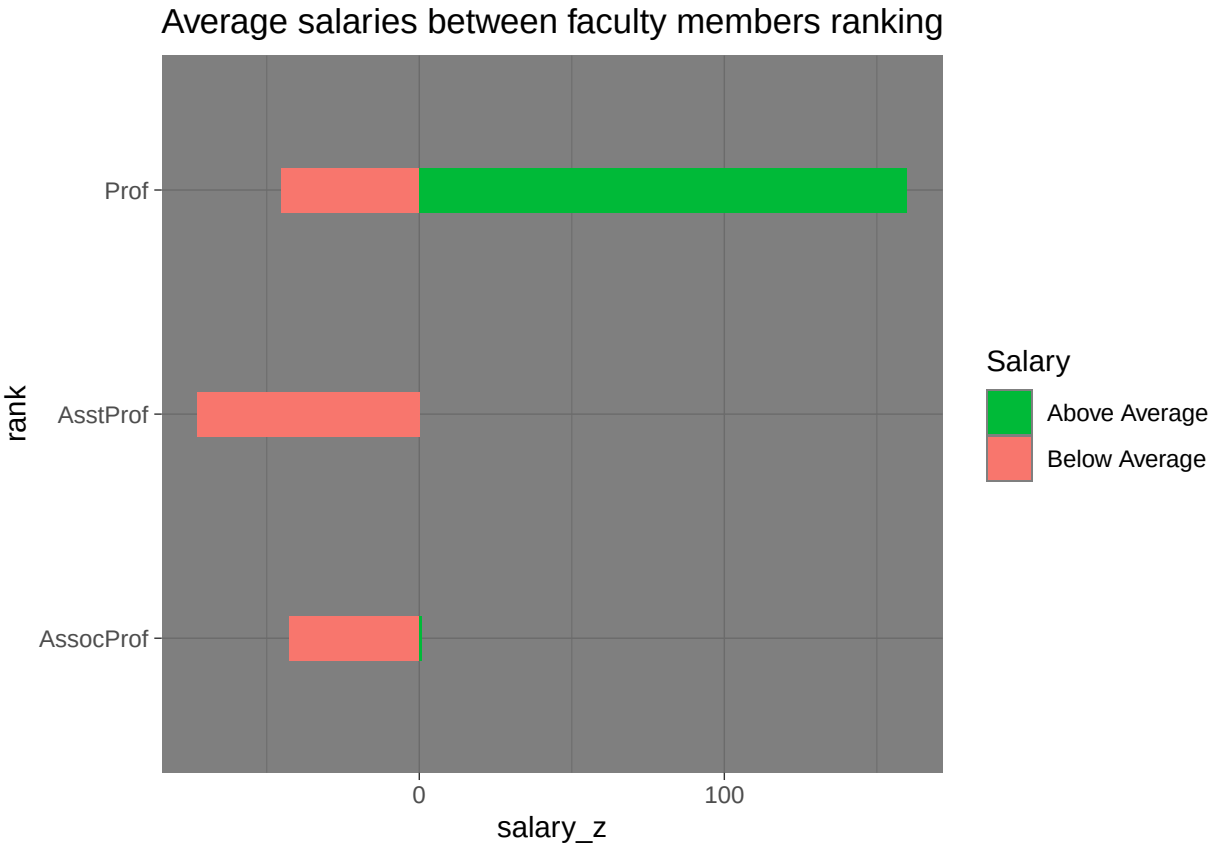


AVG_SALARY VS RANK

It is worth noticing in this graph that only professors are earning above average salary and a really small amount of associate professors. The assistant professors, are in the entire, below average salary. In conclusion, maybe females were mostly under average salary because not a lot of them belong to professor's rank.

```
df$rank_type <- rownames(df)
df$salary_z <- round((df$salary - mean(df$salary)) / sd(df$salary),2)

df$rank_type <- ifelse(df$salary_z < 0, "below","above")

df <- df[order(df$salary_z),] #sort

## convert to factor to retain sorted order in plot
df$rank_type <- factor(df$rank_type, levels=rev(unique(df$rank_type)),ordered=TRUE)

#diverging barcharts
ggplot(df, aes(x=rank, y=salary_z, label=salary_z)) +
geom_bar(stat='identity', aes(fill=rank_type), width=.2) +
```

```
scale_fill_manual(name="Salary",
labels = c("Above Average", "Below Average"),
values = c("above"="#00ba38", "below"="#f8766d")) +
labs(title= "Average salaries between faculty members ranking") +
coord_flip() + theme_dark()
```



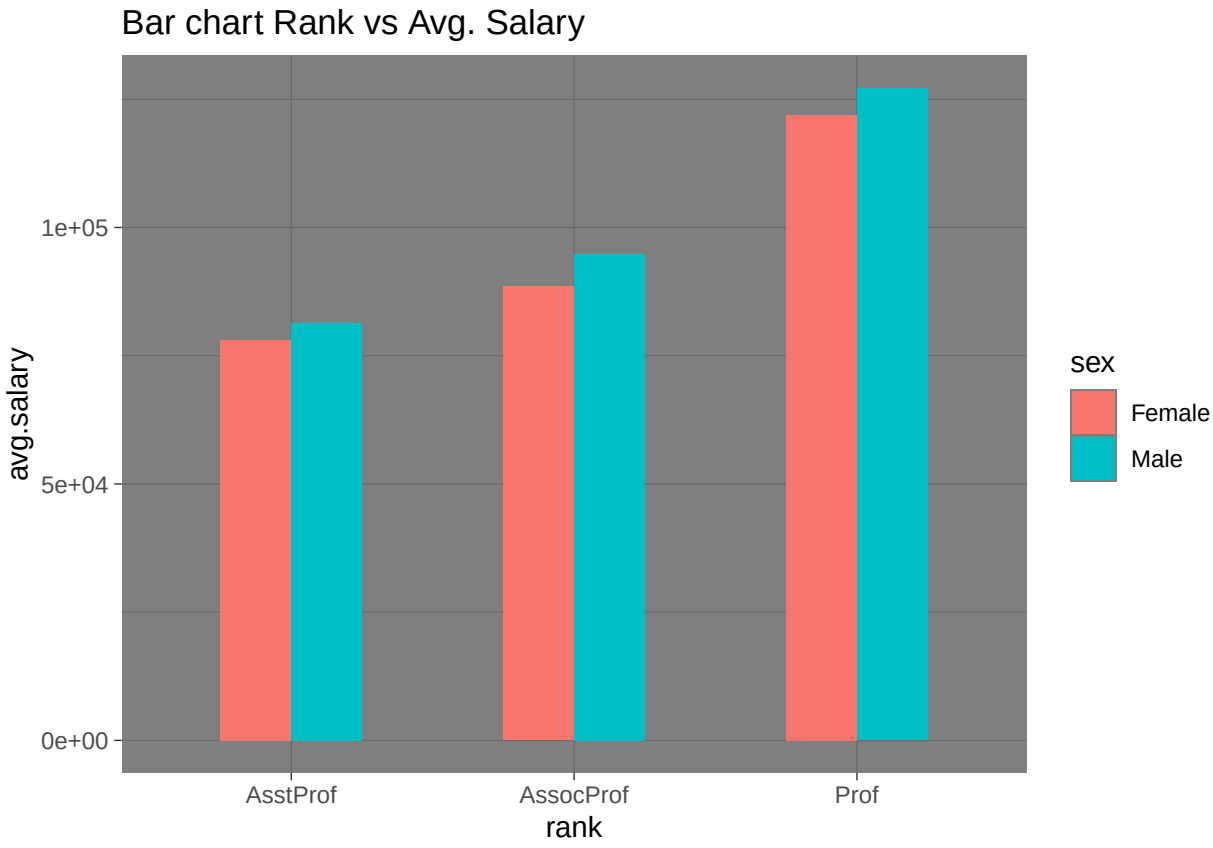
RANKING PLOTS ~ ORDERED BAR CHARTS

AVG SALARY ~ RANK

In this graph we are examining the relationship between the amount of average salary and the ranking of faculty members distinguished by sex. It is obvious that the higher ranking you are the more salary you earn. We can also notice that between sexes in all rankings, males earn more money in average than females.

```
avg_slry <- aggregate(df$salary, by=list(df$rank,df$sex),
FUN=mean) # aggregate

colnames(avg_slry) <- c("rank","sex","avg.salary") # change column names

avg_slry<- avg_slry[order(avg_slry$avg.salary), ] # sort by avg. salary

# to retain the order in plot.
avg_slry$rank <- factor(avg_slry$rank, levels = list("AsstProf","AssocProf","Prof"))

# Draw plot
```

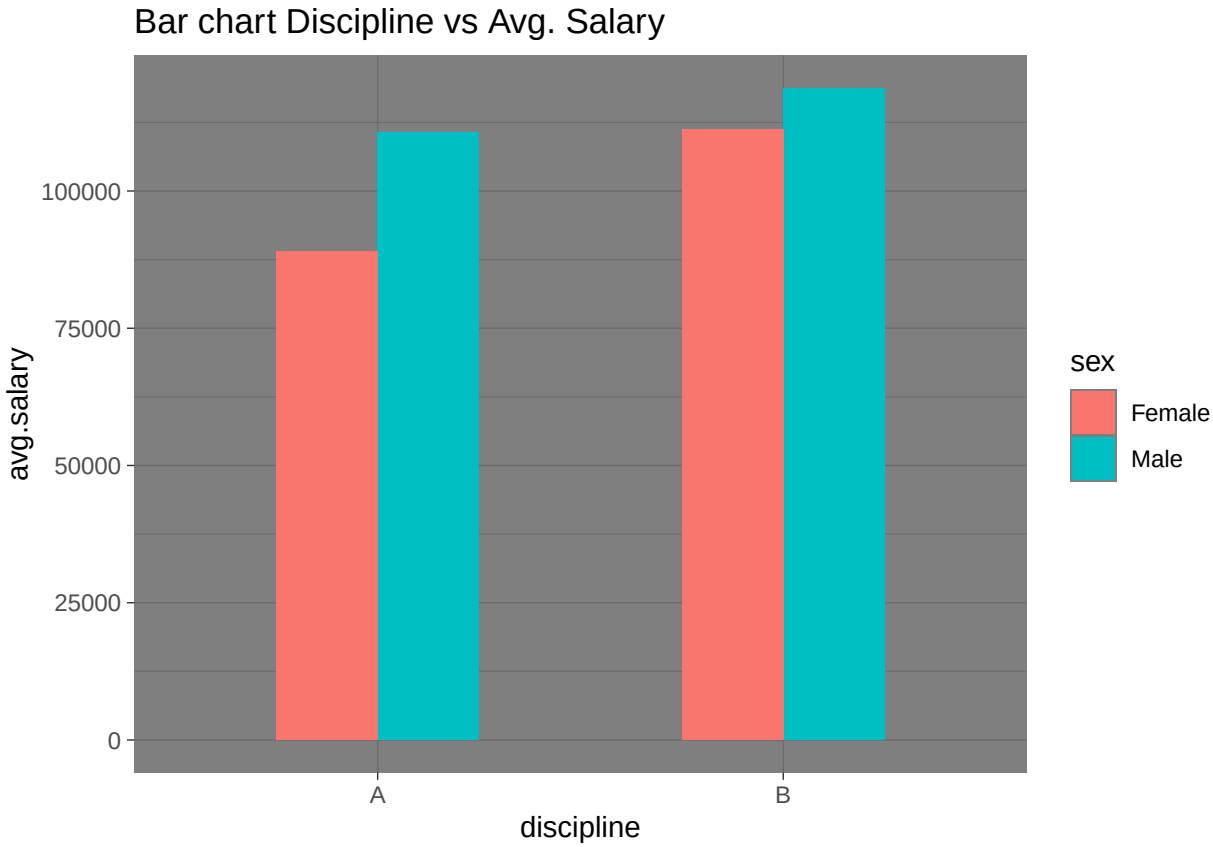
```r
ggplot(avg_slry, aes(x=rank, y=avg.salary)) +
geom_bar(position="dodge",stat="identity",width=.5,aes(fill=sex)) +
labs(title="Bar chart Rank vs Avg. Salary") +
theme(axis.text.x = element_text(angle=65, vjust=0.6)) + theme_dark()
```



AVG SALARY ~ DISCIPLINE

Here, it is visible that applied departments earn in average more money than the theoritical ones. Again, we notice that men earn more than women independently of their discipline.

```r
avg_slry <- aggregate(df$salary, by=list(df$discipline,df$sex),FUN=mean) # aggregate

colnames(avg_slry) <- c("discipline","sex","avg.salary") # change column names

avg_slry<- avg_slry[order(avg_slry$avg.salary), ] # sort by avg. salary

# Draw plot
ggplot(avg_slry, aes(x=discipline, y=avg.salary)) +
geom_bar(position="dodge",stat="identity",width=.5,aes(fill=sex)) +
labs(title="Bar chart Discipline vs Avg. Salary") +
theme(axis.text.x = element_text(angle=0, vjust=0.6)) + theme_dark()
```



BARPLOTS

Count plot rank VS sex and discipline VS sex

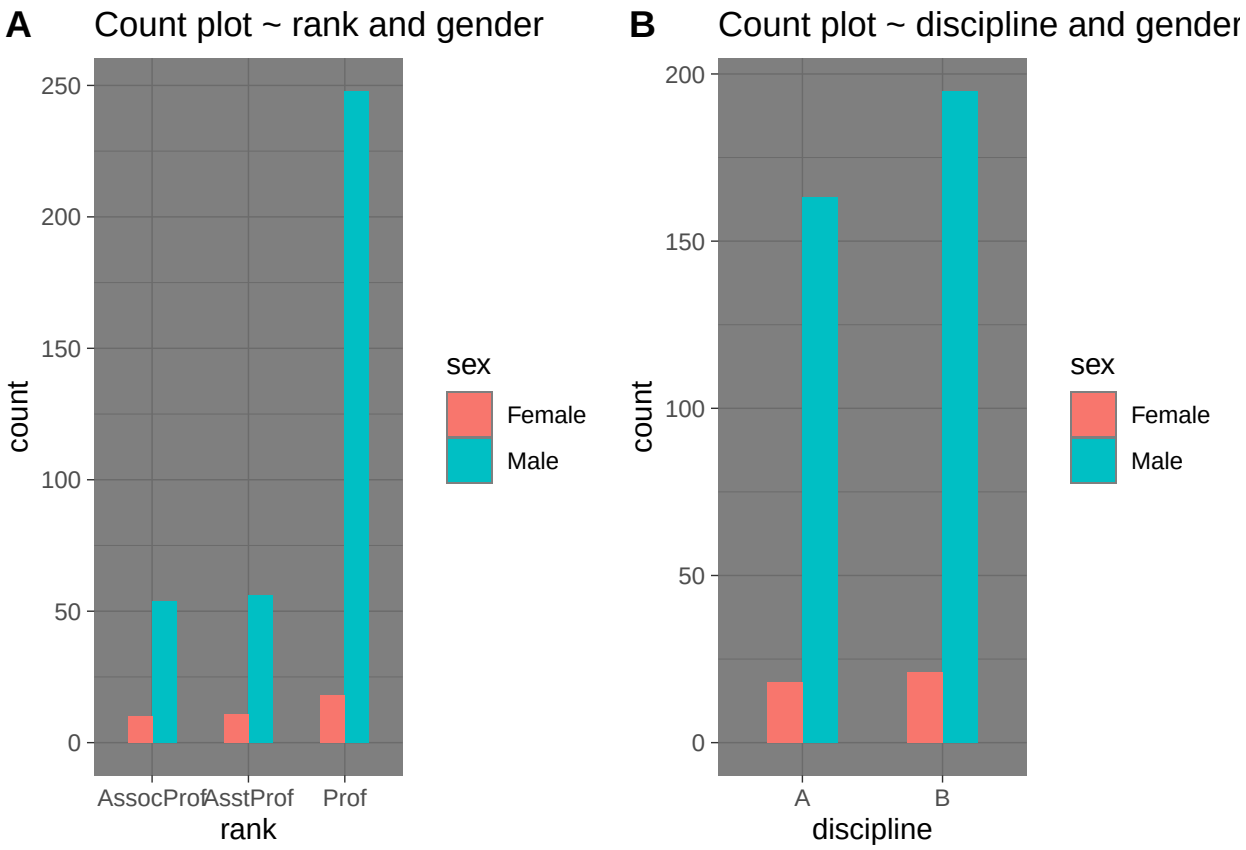
rank vs sex

This specific plot just shows the distribution of females and males on rankings. Again we notice how men earn more in every ranking. Also, the number of professors is larger than the other rankings in both sexes.

discipline vs sex

The next plot can help us observe that both males and females tend more in the applied departments than the theoritical ones. Also we can notice the difference between the number of males and females in each discipline which was mentioned in the beginning.

```
g <- ggplot(df, aes(rank)) + geom_bar(position="dodge",aes(fill=sex), width = 0.5) +
theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
labs(title="Count plot ~ rank and gender") + theme_dark()

g2 <- ggplot(df, aes(discipline))  + geom_bar(position="dodge",aes(fill=sex), width = 0.5) +
theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
labs(title="Count plot ~ discipline and gender") + theme_dark()

plot_grid(g, g2, labels = "AUTO")
```



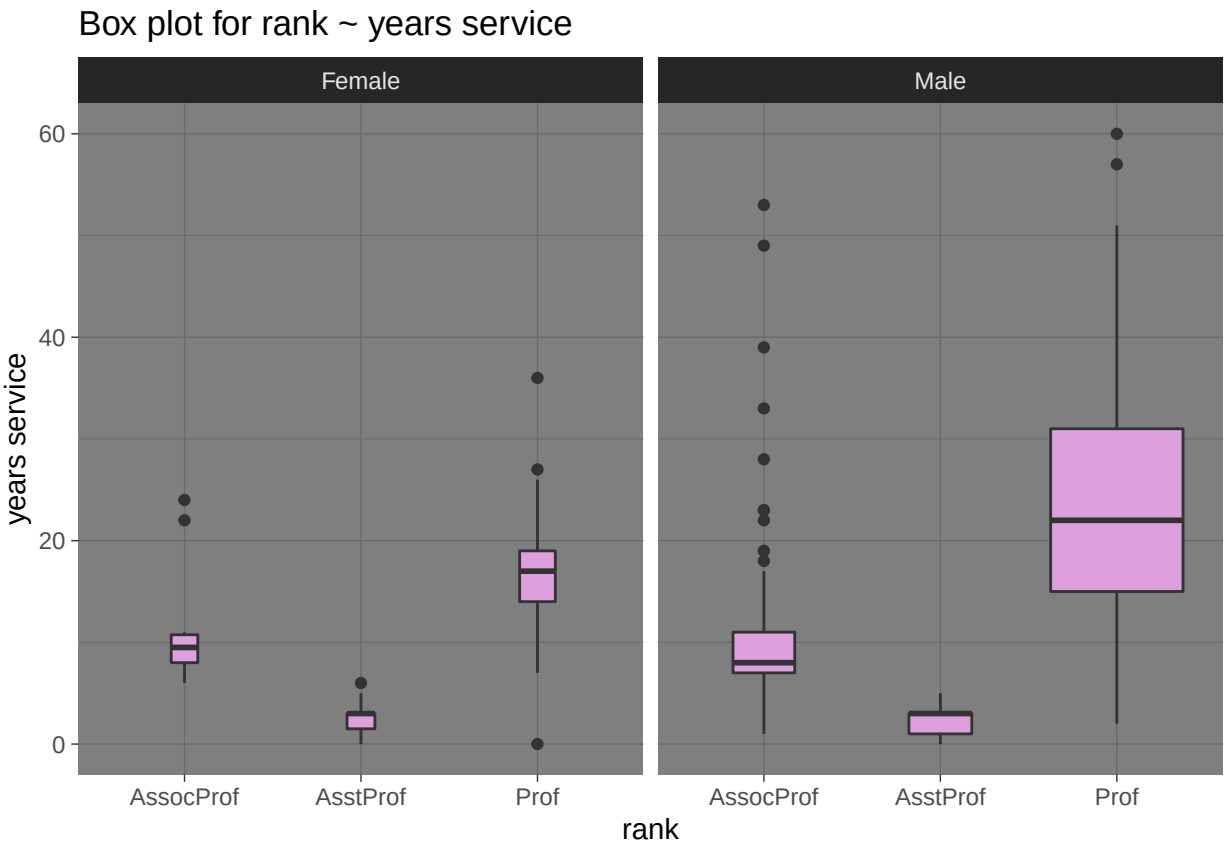
BOXPLOTS

Below we have created 2 grid plots. The first uses years of servicing and the other uses the salary. Both grids show for both genders that the more years you have in either category the more chances you have to be at a higher rank in the faculty and so to increase the possibilities of a higher salary.

boxplot rank~ years service

We can infer by this plot that males have more years than females on the ranking of Professor. In the rank of Associate Professor males have a lot of outliers compared to females while female median is higher than the male one. As for Assistant professors things appear balanced to both genders.

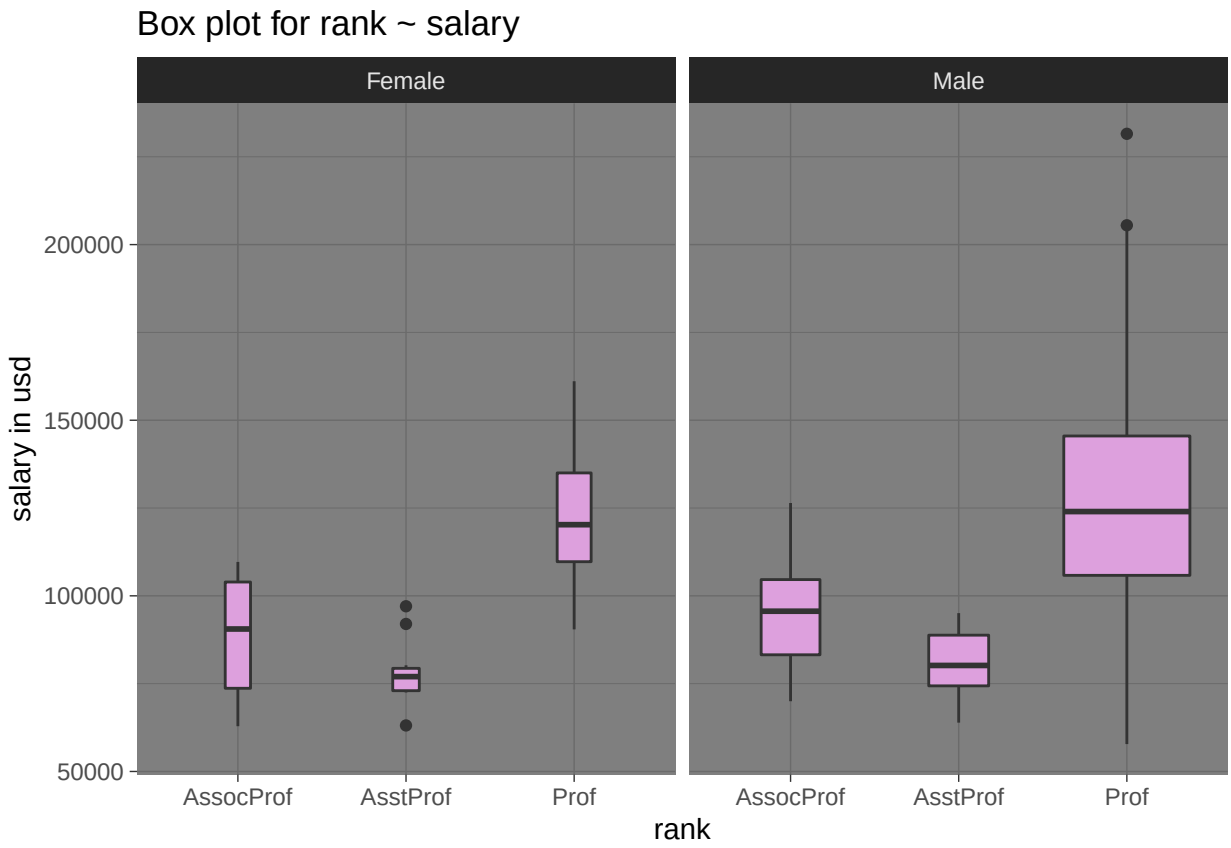
```
g <- ggplot(df, aes(rank, yrs.service))
g + geom_boxplot(varwidth=T, fill="plum") +
labs(title="Box plot for rank ~ years service",
x="rank",
y="years service") + facet_grid(~sex) + theme_dark()
```



boxplot rank~ salary

In this grid, we take account ranking and salary for both sexes. We can notice that the medium values for Females and Males of all rankings are quite close, but men's slightly bigger.

```
g <- ggplot(df, aes(rank, salary))
g + geom_boxplot(varwidth=T, fill="plum") +
labs(title="Box plot for rank ~ salary",
x="rank",
y="salary in usd") + facet_grid(~sex) + theme_dark()
```


Box plot for rank ~ salary
Female
Male
200000
150000
100000
50000
salary in usd
AssocProf
AsstProf
Prof
AssocProf
AsstProf
Prof
rank

Conclusions

After our analysis we can conclude some facts. It is obvious, that men earn a higher salary in average than women. Generally, women tend to get below average mostly as we could observe from the deviation charts.Also, its worth to mention that only professors get more than the average salary and a few associate professors. From our scatterplots and boxplots we noticed that the more years phd/Service you have the more the chances to be on a higher rank and thus earn a bigger salary. Females, have less years than males. This might justifies the reason they have lower salary in average. Another point we can make is that the members who belong on applied departments get paid more than the members on the theoritical ones. This applies to both sexes. It is also worth to mention the fact of the negative correlation between years serving/phd and salary, which as we said maybe is justified from the action of paying less the older professors so that the younger can earn more.

Finally, we will close this topic by addressing the existance of a bias towards men to earn more money. It is not possible to prove it since there is no specific pattern. Salary is connected with rank which is connected with years servicing and women have much less years than men in the field.