Exploratory Data Analysis in R

Maria Kaiktzoglou

STUDENT ID : 03400052

EMAIL : makaiktzoglou@gmail.com

Dataset: phD salaries 2008-9

Reading the data, basic checks and first plots

We read the data into R and save the results to a data frame

```
data <- read.table('/home/maria/Desktop/Desktop/Salaries.csv', header = T)
head(data, 5)
```

```
##       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
```

We check that the data types are correct. They turn out to be. We also check whether there are any missing values. It turns out that there aren't any.

```
data_frame(names(data), sapply(data, class))
```

```
## # A tibble: 6 x 2
##   `names(data)` `sapply(data, class)`
##   <chr>         <chr>
## 1 rank          factor
## 2 discipline    factor
## 3 yrs.since.phd integer
## 4 yrs.service   integer
## 5 sex           factor
## 6 salary        integer
```

```
cat('number of missing data / null values: ', sum(is.na(data)) + sum(is.null(data)))
```

```
## number of missing data / null values:  0
```

Distribution of sex, rank and discipline in the sample

In the beginning it is useful to get some insight into the sample. Namely, how the 'sex', 'rank' and 'discipline' are distributed among the participants.

In the table below we see that the vast majority of the participants consists of males (~90%)

```
sex_count <- dplyr::count(data, sex)
sex_count['percent %']<- round(100*sex_count$n/sum(sex_count$n), 1)
names(sex_count)[2]<- 'frequency'
sex_count
```

```
## # A tibble: 2 x 3
##   sex    frequency `percent %`
##   <fct>      <int>       <dbl>
## 1 Female        39         9.8
## 2 Male         358        90.2
```

Approximately 2/3 of the participants are professors, 1/6 are Associate Professors and 1/6 Assistant Professors.

```
df1 <- as.data.frame(table(data$rank))
piepercent<- round(100*df1$Freq/sum(df1$Freq), 1)  # percentages
df1['percent %']<-piepercent
colnames(df1) <- c("rank", "frequency", 'percent %')
df1
```

```
##        rank frequency percent %
## 1 AssocProf        64      16.1
## 2  AsstProf        67      16.9
## 3      Prof       266      67.0
```

```
pie <- ggplot(df1, aes(x = "", y=frequency, fill = factor(rank))) + geom_bar(width = 1, stat = "identit
  )
pie + coord_polar(theta = 'y')
```

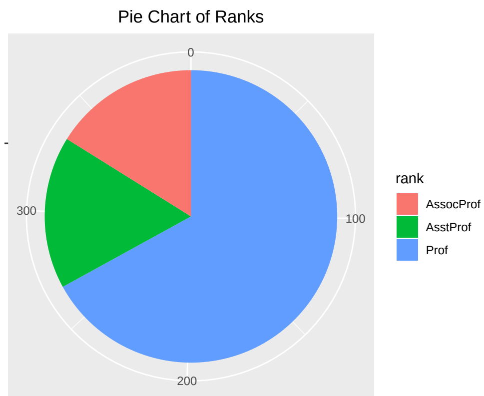

Figure 1

Next, we see that there are slightly more participants who come from applied departments compared to theoretical departments. It can be seen that it is males who account for this difference.

```
disc_count <- dplyr::count(data, discipline)
disc_count['percent %']<- round(100*disc_count$n/sum(disc_count$n), 1)
names(disc_count)[2]<- 'frequency'
disc_count
```

```
## # A tibble: 2 x 3
##   discipline frequency `percent %`
##   <fct>          <int>       <dbl>
## 1 A                181        45.6
## 2 B                216        54.4
```

```
ggplot(data, aes(x = discipline, fill=sex)) + geom_bar(position='dodge')+labs(title= "Disciplines Barcha
```

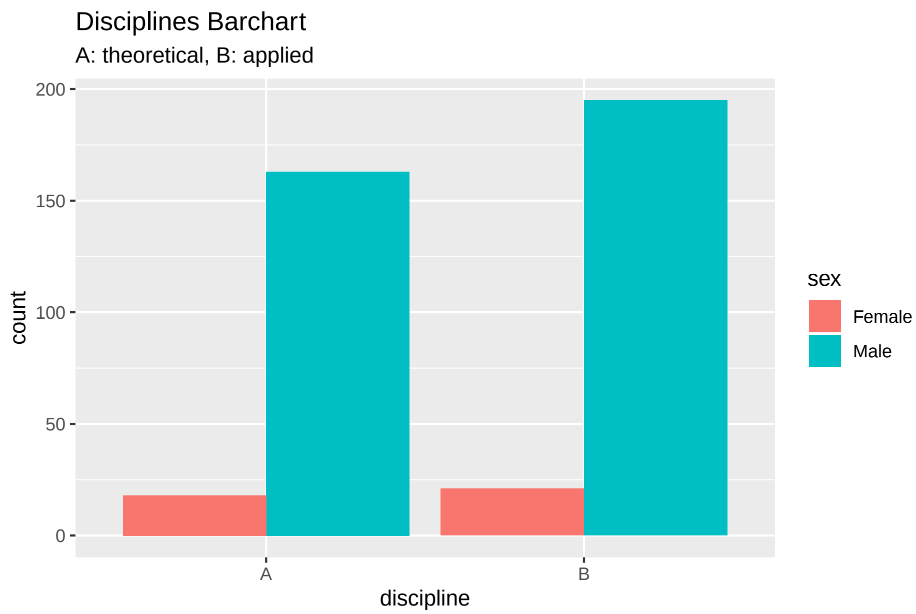

Figure 2

```
rank_count <- dplyr::count(data,rank)
rank_count['overall percent']<- round(100*rank_count$n/sum(rank_count$n), 1)
names(rank_count)[2]<- 'frequency'
tmp1<-dplyr::count(data, rank,sex)[dplyr::count(data, rank,sex)$sex=='Male',][3]
rank_count['males percent']<-round(100*tmp1/sum(rank_count['frequency']),1)
tmp1<-dplyr::count(data, rank,sex)[dplyr::count(data, rank,sex)$sex=='Female',][3]
rank_count['females percent']<-round(100*tmp1/sum(rank_count['frequency']),1)
rank_count[c(1,3,4,5)]
```

```
## # A tibble: 3 x 4
##   rank      `overall percent` `males percent` `females percent`
##   <fct>                 <dbl>           <dbl>             <dbl>
## 1 AssocProf              16.1            13.6               2.5
## 2 AsstProf               16.9            14.1               2.8
## 3 Prof                   67              62.5               4.5
```

```
ggplot(data, aes(x = rank, fill=sex)) + geom_bar(position='fill')+labs(title = "Proportion of males/fema
```

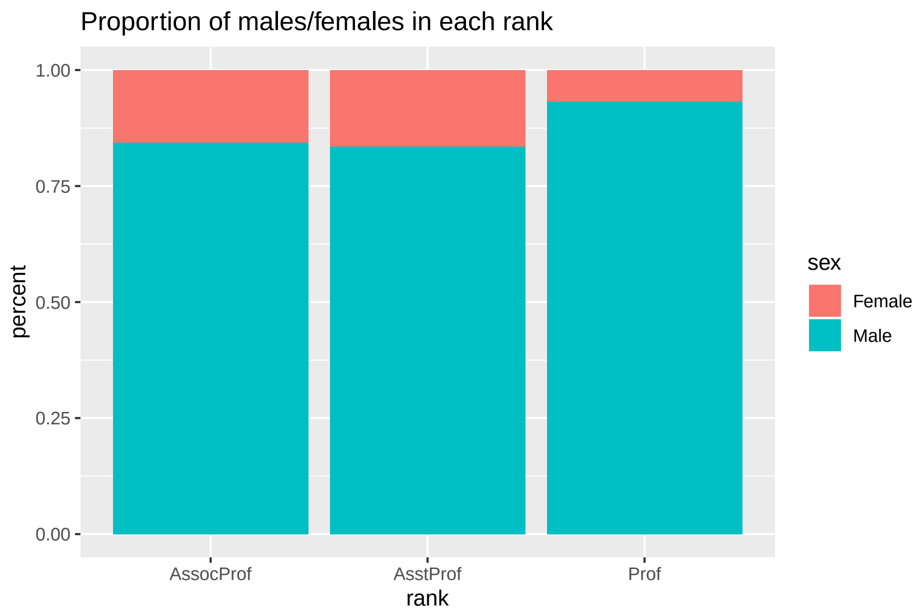

Figure 3

An important outcome from the above is that the 'Professors' rank is mainly occupied by males. We remind ourselves that females take up 10% of the sample, and while they occupy 16% and 17% of the 'Associate Professors' and 'Assistant Professors' respectively, they occupy merely 6% of the 'Professors' positions.

Salary plots with respect to rank

It is important now to examine how salary relates to rank. The following boxplots shed some light to this question as we can see that the median as well as the range increase as the rank increases (e.g. the higer the rank the higher the median salary and the range of salary). It is worth noticing that there exist outliers in the 'Professors' salary.

```
g <- ggplot(data, aes(rank, salary))
g + geom_boxplot(varwidth=T, fill = 'darkseagreen3') + theme(axis.text.x = element_text(angle=65, vjust=
```

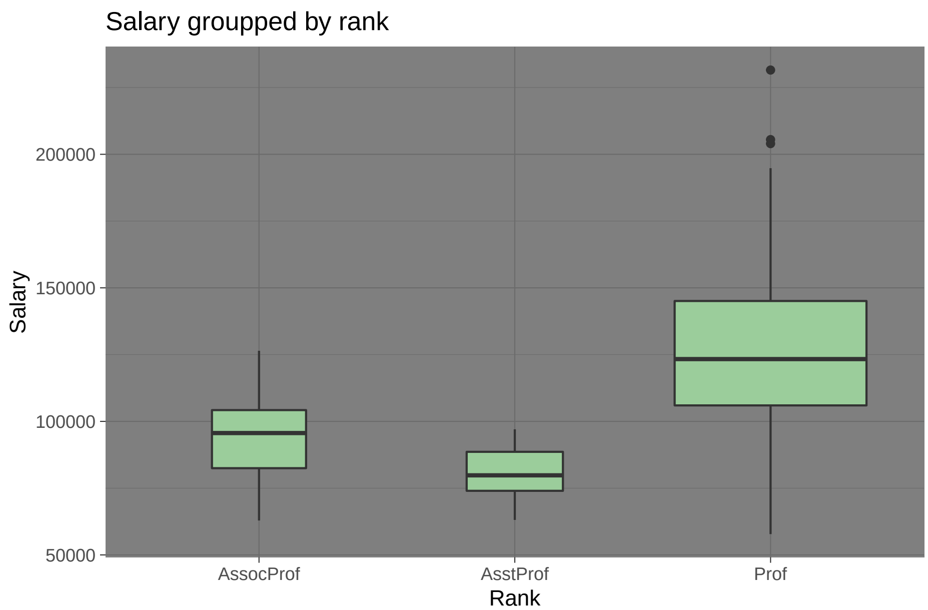

Figure 4

We can also view in the table and the lollipop chart below the average salaries of each rank, split by sex.

```
salary_data <- aggregate(data$salary, by=list(data$rank, data$sex),FUN=mean) # aggregate/group
colnames(salary_data) <- c("rank", 'sex',"salary") # change column names
salary_data <- salary_data[order(salary_data$salary), ] #
salary_data
```

```
##        rank    sex    salary
## 2  AsstProf Female  78049.91
## 5  AsstProf   Male  81311.46
## 1 AssocProf Female  88512.80
## 4 AssocProf   Male  94869.70
## 3      Prof Female 121967.61
## 6      Prof   Male 127120.82
```

```
theme_set(theme_bw())
ggplot(salary_data, aes(x=rank, y=salary, color=sex)) + geom_point(size=3) + geom_segment(aes(x=rank, x
```

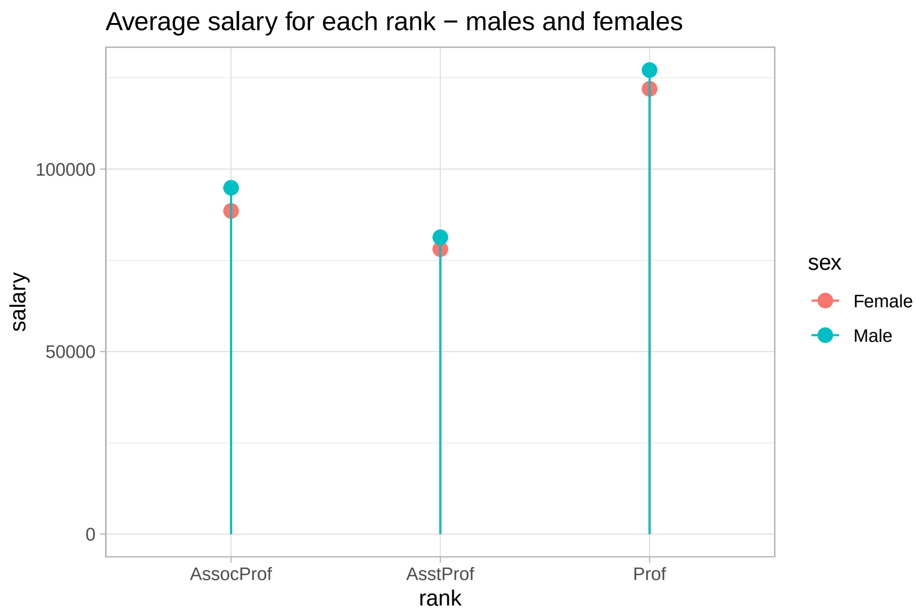

Figure 5

We can detect small yet clear differences in the average salary between males and females for every rank. We come back to boxplots, this time creating a distinct boxplot for every sex and every rank that will contain more information about the distribution of the observations (quantile range and outliers).

```
g <- ggplot(data, aes(rank, salary))
g + geom_boxplot(aes(fill=sex)) +
theme(axis.text.x = element_text(angle=65, vjust=0.6))+ labs(title="Boxplots for salary groupped by ranl
```

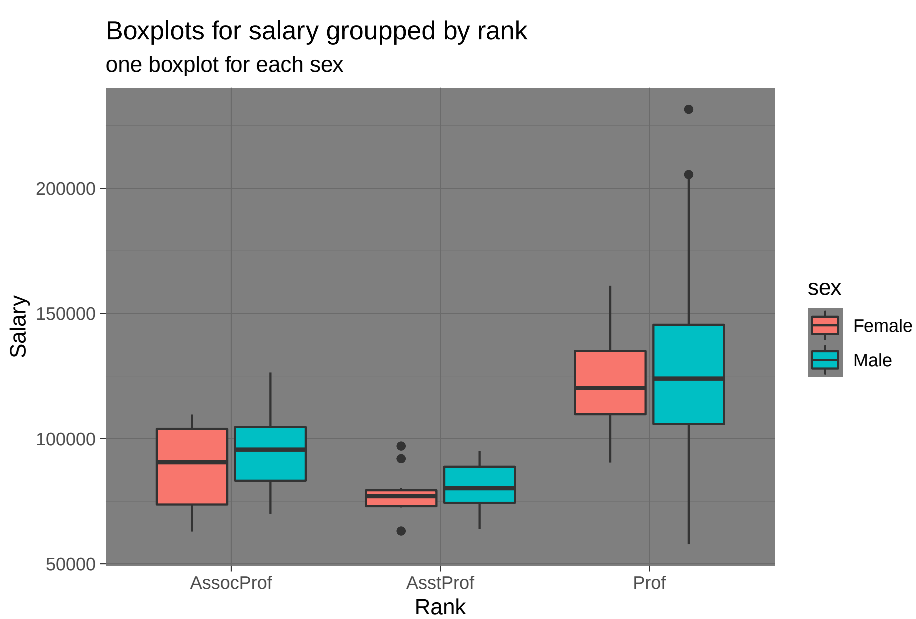

Figure 6

Comments:

1. In all three ranks the median salary of males is higher than this of females

2. The range of salaries in Associate Professors is bigger for females and skewed to lower salary values. The range of salary in Assistant Professors is bigger for males and skewed to higher salary values. The respective range in Professors is also bigger for males.

3. There are outlying values -both high and low- in female Assistant Professors' salary. There are high outlying values in male Professors' salary.

Boxplots are not informative when it comes to how observations are distributed within the quantiles; to get some insight we use diverging bars. We constructed four pairs of diverging bars; the first is about salary regardless the rank, and the other three are for a specific rank. We normalized the salary so that the mean is 0, as it can be seen. The green area at the right of the mean represents the number of males/females whose salary is above the mean, while the red area at the left of the mean represents the number of males/females whose salary is below the mean.

```
# for professors
data$salary_z <- round((data$salary - mean(data$salary))/sd(data$salary), 2) # normal. salary
data$salary_type <- ifelse(data$salary_z < 0, "below", "above") # above / below average
data <- data[order(data$salary_z), ]

ggplot(data, aes(x=sex, y=salary_z, label=salary_z)) +
geom_bar(stat='identity', aes(fill=salary_type), width=.5) + scale_fill_manual(name="USD",
labels = c("Above Average", "Below Average"), values = c("above"="#00ba38", "below"="#f8766d")) + labs(
```

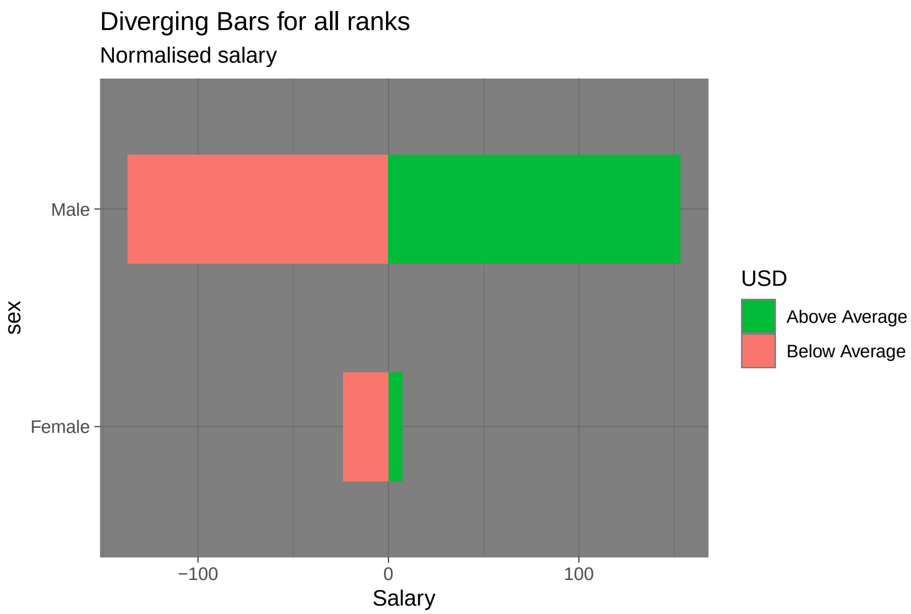

Figure 7

```
data1<-data[data$rank=='Prof',]
data1$salary_z <- round((data1$salary - mean(data1$salary))/sd(data1$salary), 2)
data1$salary_type <- ifelse(data1$salary_z < 0, "below", "above")
data1 <- data1[order(data1$salary_z), ]

ggplot(data1, aes(x=sex, y=salary_z, label=salary_z)) + geom_bar(stat='identity', aes(fill=salary_type)
```

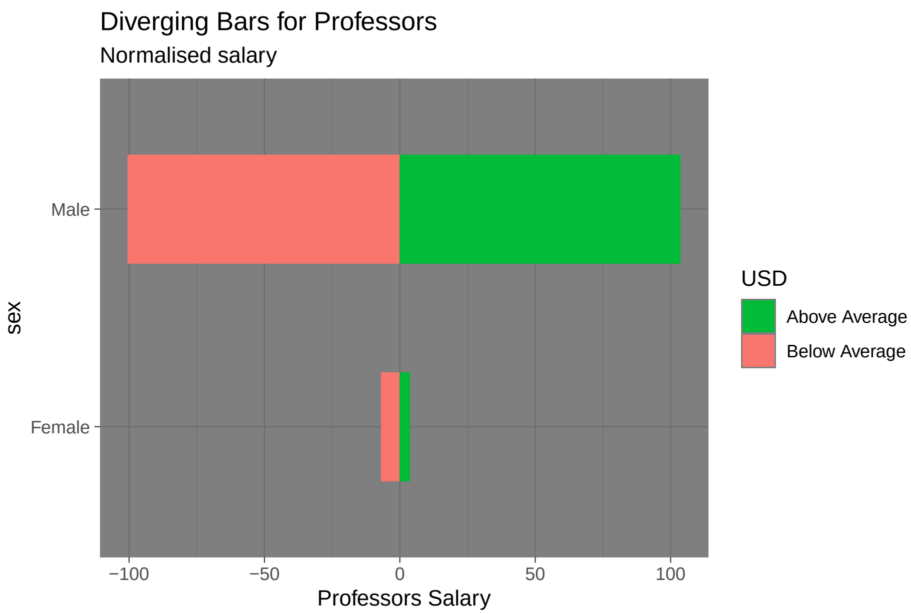

Figure 8

```r
# For Associate Professors
data1<-data[data$rank=='AssocProf',]
data1$salary_z <- round((data1$salary -
mean(data1$salary))/sd(data1$salary), 2)
data1$salary_type <- ifelse(data1$salary_z < 0, "below", "above")
data1 <- data1[order(data1$salary_z), ]

ggplot(data1, aes(x=sex, y=salary_z, label=salary_z)) + geom_bar(stat='identity', aes(fill=salary_type)
```

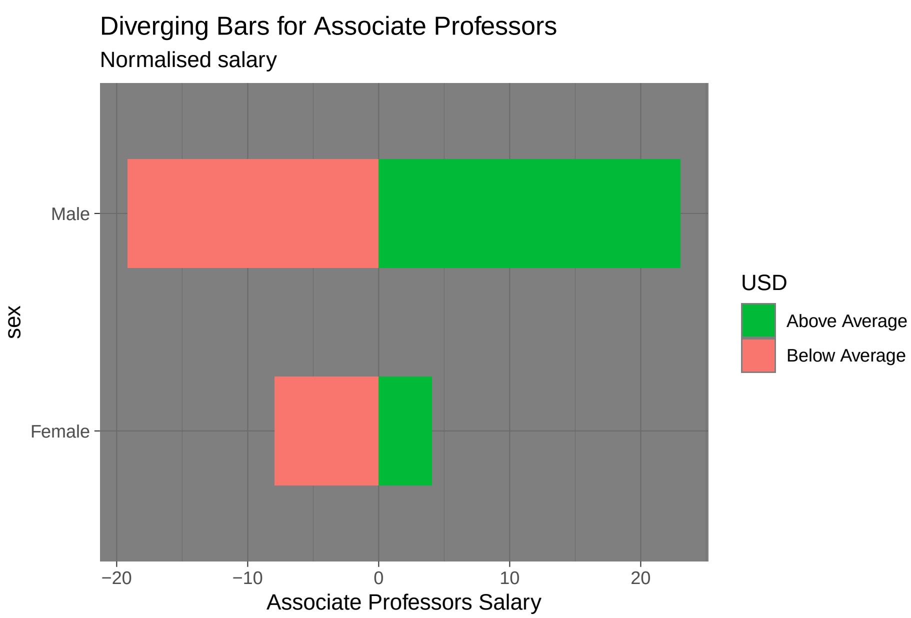

Figure 9

```r
# For Assistant Professors
data1<-data[data$rank=='AsstProf',]
data1$salary_z <- round((data1$salary - mean(data1$salary))/sd(data1$salary), 2)
data1$salary_type <- ifelse(data1$salary_z < 0, "below", "above")
data1 <- data1[order(data1$salary_z), ]

ggplot(data1, aes(x=sex, y=salary_z, label=salary_z)) + geom_bar(stat='identity', aes(fill=salary_type)
```

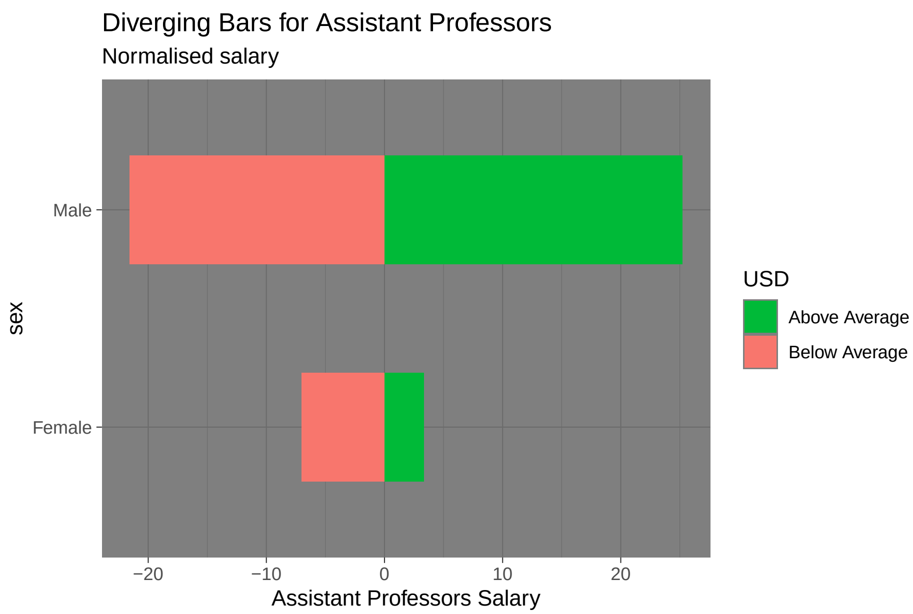

Figure 11

In all cases the green area (above mean) is bigger than the red area (below mean) for the males and vice versa for the females. This means that the average salary is lifted by males' salaries that tend to be higher than those of females.

Salary plots with respect to years of service/since phd

Let's look at the years of service factor and how it relates to sex and salary. We group the data by years of service and calculate the average salary for each value of this variable (e.g. average salary for 1 year of service, for 2 years of service etc.). This is done for each sex separately as well as for both sexes together.

```r
data_select <- data[(data$yrs.service > 50) & (data$sex=='Male') & (data$salary>150000),]

require(gridExtra)
```

```
## Loading required package: gridExtra

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
gr_ser<-group_by(data, yrs.service, sex)%>%summarise(mean_salary=mean(salary))
gr_ser2<-group_by(data, yrs.service)%>%summarise(mean_salary=mean(salary))

plt1<-ggplot(gr_ser, aes(x = yrs.service, y = mean_salary, color=sex)) + geom_point()+geom_smooth() + t
plt2<-ggplot(gr_ser2, aes(x = yrs.service, y = mean_salary)) + geom_point()+geom_smooth() + theme_dark(
data=data_select,color="red",size=1,expand=0.001) +theme(plot.caption = element_text(hjust=0))
grid.arrange(plt1, plt2, ncol=2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

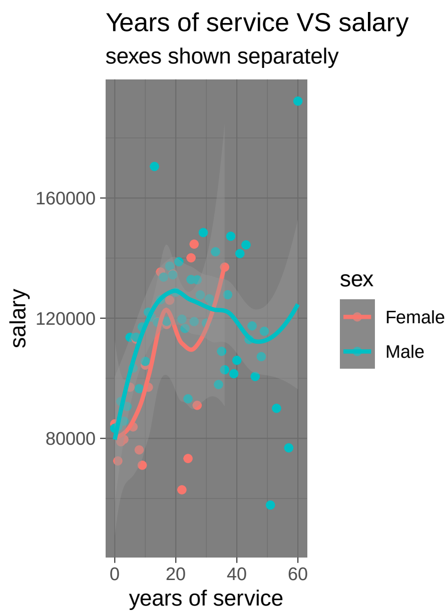

Figure 12

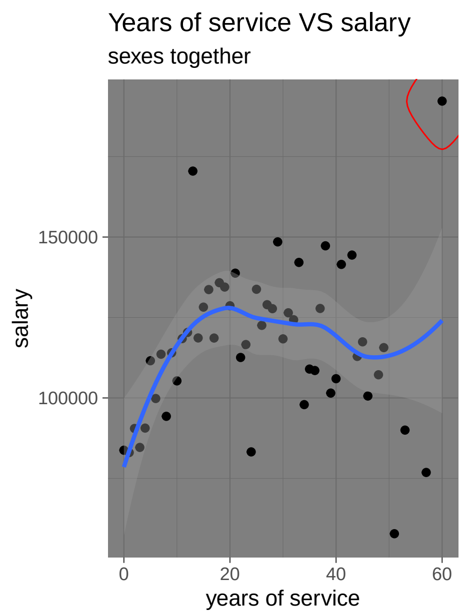

Figure 13

One thing to note is that females have been working in universities far fewer years than males. The trend that appears (both sexes plot) is noteworthy, since we can see it is descending. At the right, the line is lifted by an extreme outlying value (circled). We are doing one more scatterplot with distinct colors for each rank.

```
ggplot(data, aes(x = yrs.service, y = salary, color = rank)) + geom_point() + theme_dark() +geom_smooth
```

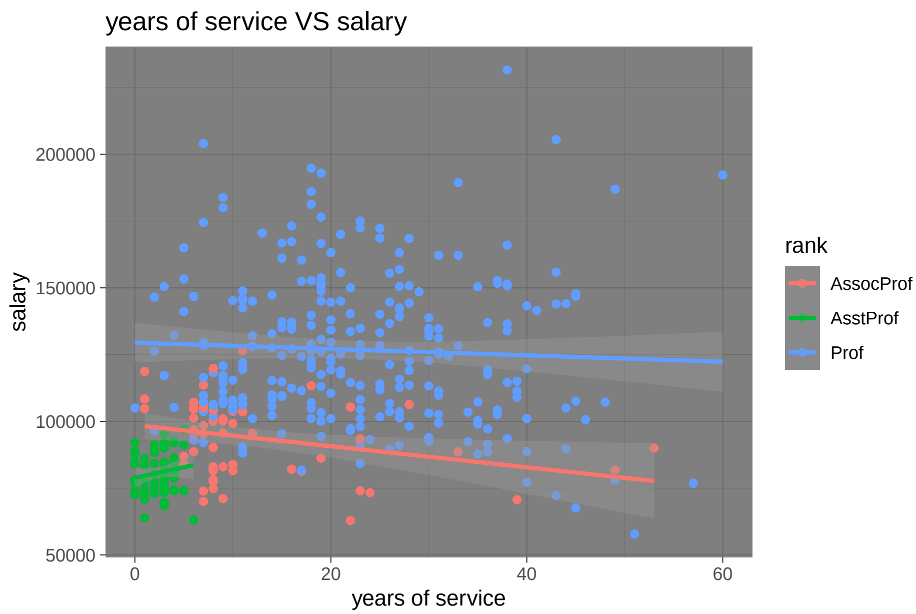

Figure 14

Interpretation: Associate Professors account more for this descending trend. There are some with many years of service but whose salary does not grow along with the years. We also notice that Professors' salary presents a descending trend - and a few high outliers- when the years of service are more than 30. In other words, older professors get less well paid than younger professors.

We continue with the same scatterplots - salary VS years of service/since phd - but now for professors and Associate professors separately. We exclude Assistant Professors due to little data.

```
require(gridExtra)
# For Professors - Service years
gr_srv_prof<-group_by(data[data$rank=='Prof',], yrs.service, sex)%>%summarise(mean_salary_professors=mea
plt1<-ggplot(gr_srv_prof, aes(x = yrs.service, y = mean_salary_professors, color=sex)) + geom_point()+ge
# For Associate Professors - Service years
gr_srv_AssocProf<-group_by(data[data$rank=='AssocProf',], yrs.service, sex)%>%summarise(mean_salary_Asso
plt2<-ggplot(gr_srv_AssocProf, aes(x = yrs.service, y = mean_salary_AssocProfessors, color=sex)) + geom_
grid.arrange(plt1, plt2, ncol=2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

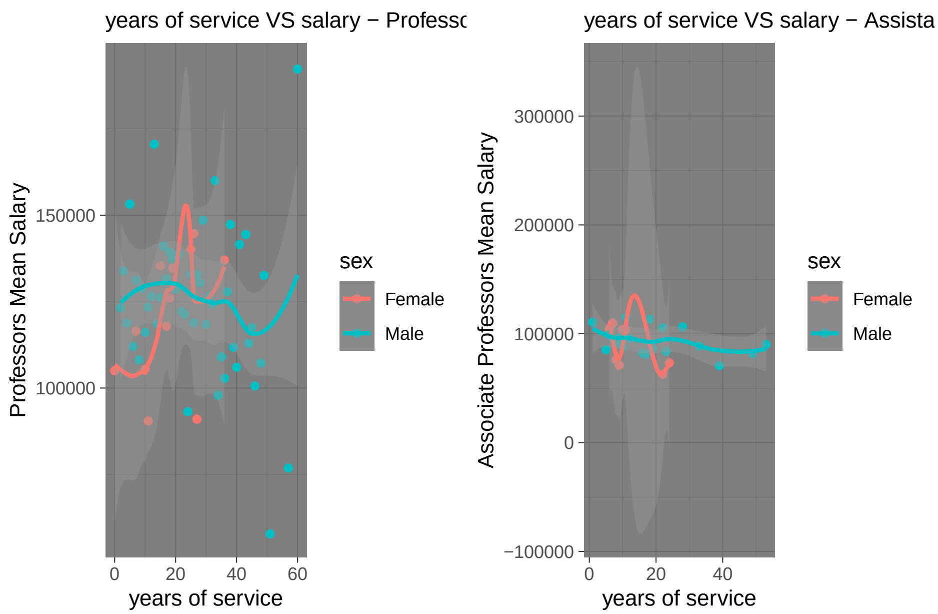

Figure 15

Figure 16

```
# For Professors - phd years
gr_phd_Prof<-group_by(data[data$rank=='Prof',], yrs.since.phd, sex)%>%summarise(mean_salary_Prof=mean(sa
plt3<-ggplot(gr_phd_Prof, aes(x = yrs.since.phd, y = mean_salary_Prof, color=sex)) + geom_point()+geom_s
# For Associate Professors - phd years
gr_phd_AssocProf<-group_by(data[data$rank=='AssocProf',], yrs.since.phd, sex)%>%summarise(mean_salary_As
plt4<-ggplot(gr_phd_AssocProf, aes(x = yrs.since.phd, y = mean_salary_AssocProf, color=sex)) + geom_poin
grid.arrange(plt3, plt4, ncol=2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Figure 17

Figure 18

Regarding the male professors, we see a linear trend whose slope somehow drops after the value 30 in years of service, which comes in accordance with the scatterplot before. We should ignore the uprise in the slope in the end as it is induced by a high outlier. At the same time, female professors seem to have an increase in salary with the passage of years, although the data is not sufficient and there is a non negligible variance.

When it comes to Associate Professors, we notice again a descending trend, which comes in accordance with our previous observations. We would rather not make any inferences for females(not a specific trend, little data)

Lastly, we are plotting histograms of salary for professors and associate professors (not assistant professors due to little data).

```
#require(gridExtra)
g1 <-ggplot(data[data$rank=='Prof',], aes(salary, color=sex)) +scale_fill_brewer(palette = "Spectral") +
g1
```

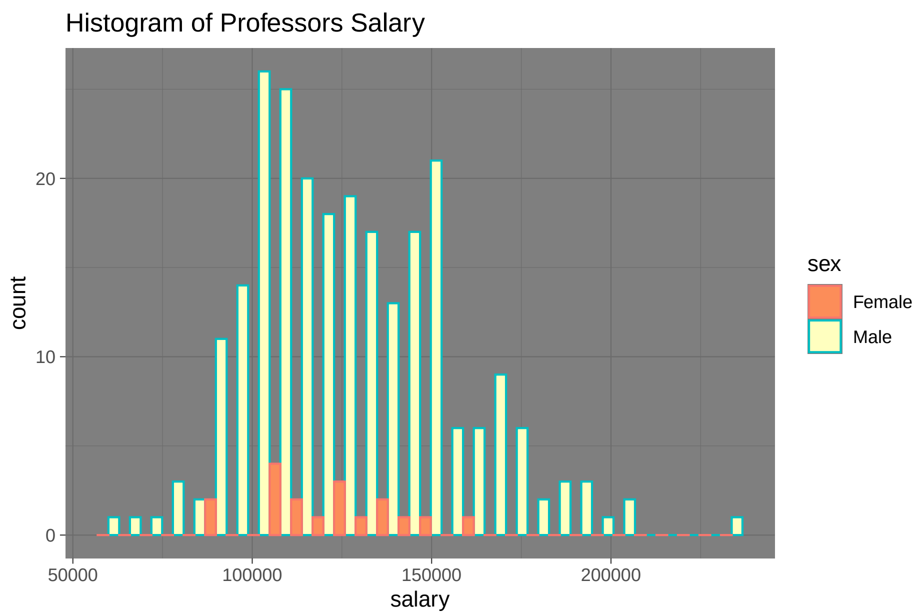

figure 19

```
g2 <- ggplot(data[data$rank=='AssocProf',], aes(salary,color=sex))+ geom_histogram(aes(fill=sex), positi
g2
```

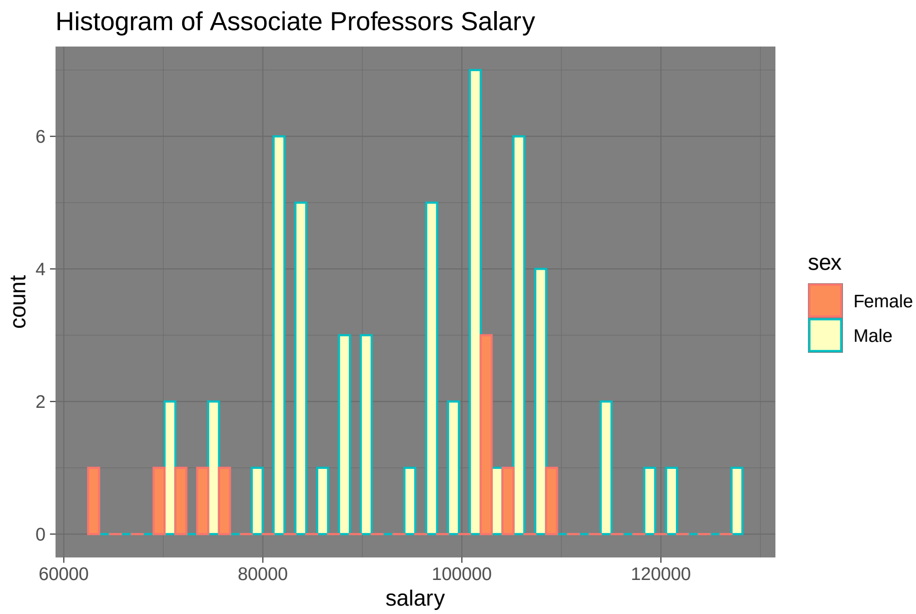

figure 20

```
#grid.arrange(g1, g2, ncol=2)
```

Most female professors receive a salary that ranges between 120,000 and 160,000 USD, with a few receiving 90,000 USD. We could say the distribution tends to have a right skewness. Male professors' salary ranges mostly between 90,000 and 200,000 USD. The distribution could be described as being in between right skewed and bimodal -in a pretty lenient statistic description!-.

Concerning associate professors, females' salary is either in the range 63,000-77,000 USD or 103,000-110,000 USD, a quite significant discrepancy of whose causes we are unaware. At the same time, around 80% of males' salaries range between 80,000 and 108,000 USD, the rest receivning either more or less. There is also significant variance.

These histograms would be better understood if accompaigned with scatterplots, however, due to our constraints in the size of this analysis, we will not do them.

Conclusions

We can reach some final conclusions after this exploratory analysis.

First of all, there is a small yet clear difference between males' and females' salary. Males' average salary is higher. The same holds for the median salary. However, the distribution of the salaries within the quantiles is not uniform at all neither does it follow some particular pattern. Diverging barcharts showed that most females get lower (below average) salaries and other plots showed that there are some females who get high salaries (outliers) which results in an increased mean. Professors get higher salary and there are more male professors than female professors. Furthermore, females have significantly fewer years of service/since phd, something that indicates that females have only recently joined this work field. Although an index of gender discrimination in jobs is the rank/position, something that goes together with the amount of salary, this goes beyond the scope of this analysis, which focuses on differences in the salary. We cannot infer a specific behaviour when it comes to females' salary with respect to their years of service/since phd. For males, there seems to be a slightly negative correlation, e.g. older males are paid less than younger males. Since there is less data from people with many years of service/since phd, hence we choose to be cautious and limited in our comments.