NDH802 R Application - Seminar #1

Huong Nguyen

Working with data in R

Load your data

If you are familiar with Excel, this step is like opening a file. You can type in

```
salaries = read.csv("https://tinyurl.com/NHD802salaries") #data stored on cloud
```

or click on Import Data set on the top right of your screen.

Explore your data

These commands give you a quick overview of your data. I recommend you to try and run them in Rstudio, but NOT to knit them. It's gonna be a bit messy. To mask a command, we make them a "comment" by placing # in front of the command. Things after # are not code and won't be executed.

```
#View(salaries)
head(salaries)
tail(salaries)
summary(salaries)
#rename a column
```

Select columns from a data frame

There are numerous way to select a column. The three most basic way are:

```
#here you refer to the column by the columns names
salaries$sal_expected
salaries[,"sal_expected"]

#here you refer to the column by the column position from the left ie 1 is the first column from the le
salaries[,2]
```

When to use which?
It's all about your personal preference. I introduce several ways to give you options, **not** to confuse you. You can totally explore, see which you're more comfortable with and stick with it. This is also applicable for most of other commands.

To select more than 1 column:

```
#here you refer to the column by the columns names
salaries[,c("sal_expected", "fairpay", "program")]

#here you refer to the column by the column position from the left
salaries[,c(1,3,10)]
```

Select rows from a data frame

There are numerous way to select a row. The most basic way (with base R) are:

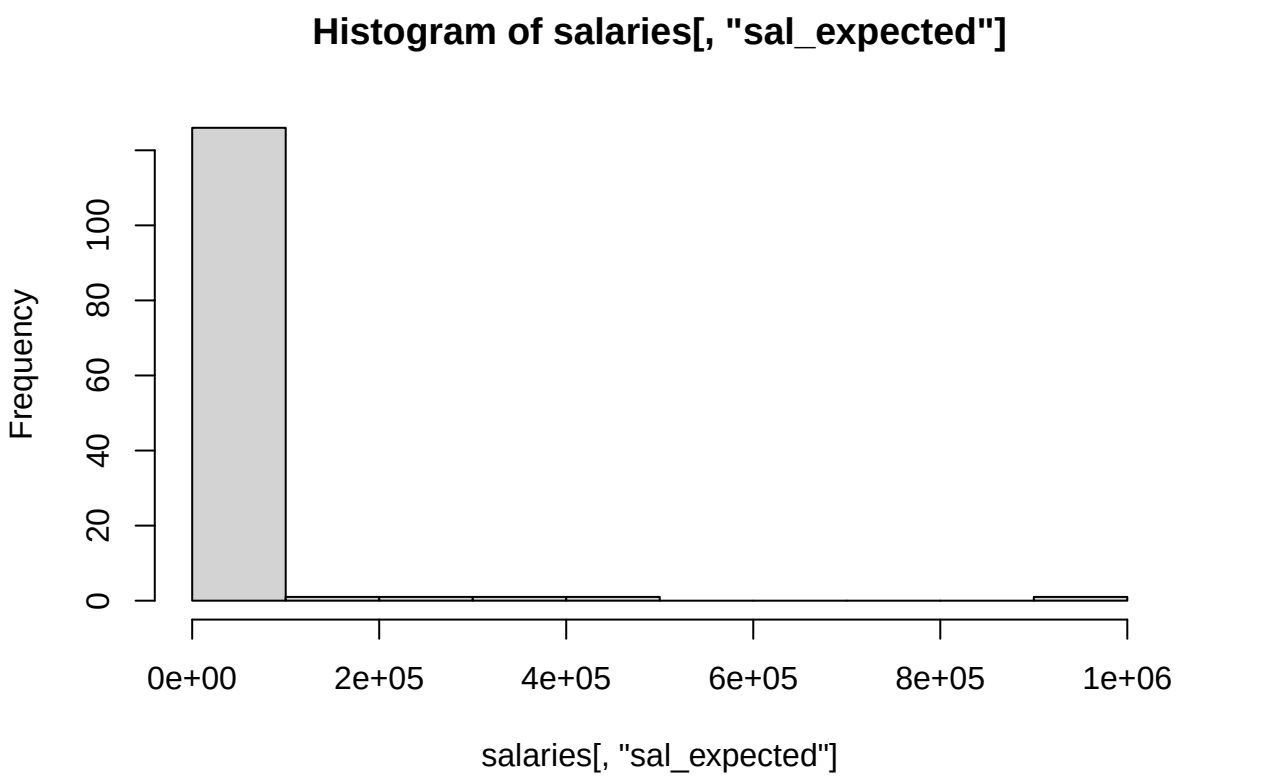
```
salaries[salaries$program == "RM",]
salaries[salaries$whatisyourgender1isfemale == 1,]
```

Let's take a closer look

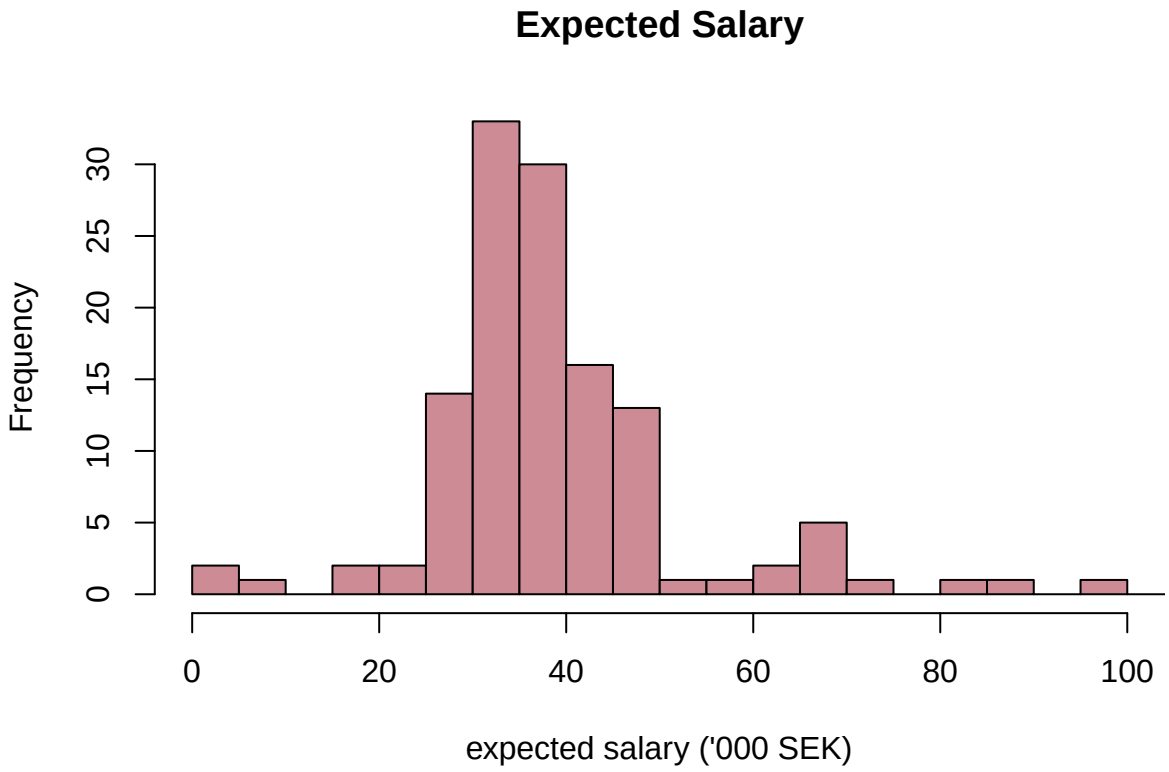
Assume you want to explore the salary expectation from the students, which is the variable `sal_expected`. Two of my most frequent used methods are histogram and box plot.

Let's first try plotting a **histogram**. If you're like "a histo what?", take a quick look here, then come back.

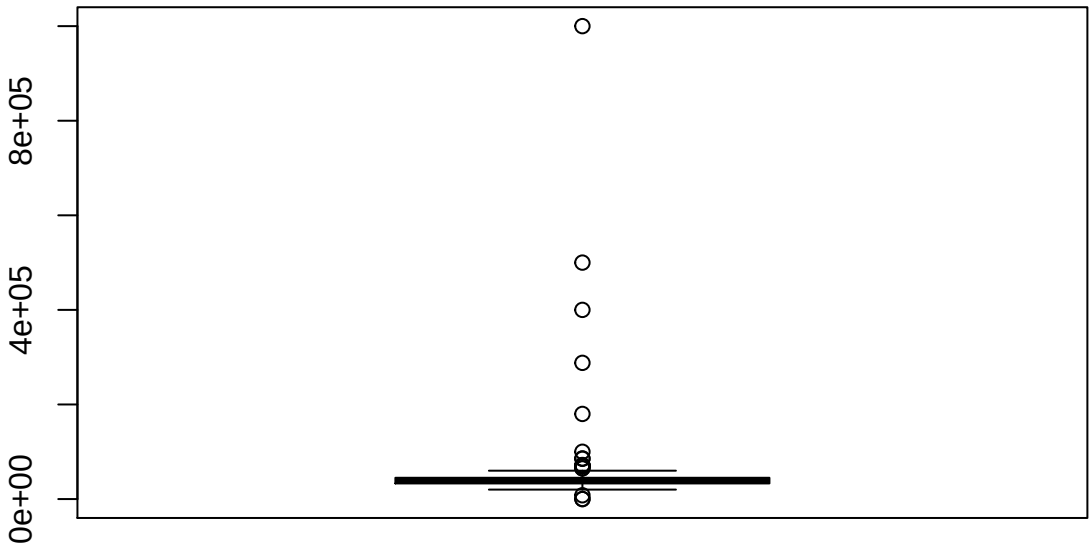
```
hist(salaries[,"sal_expected"]) #this in the assignment will get 0.5pt
```



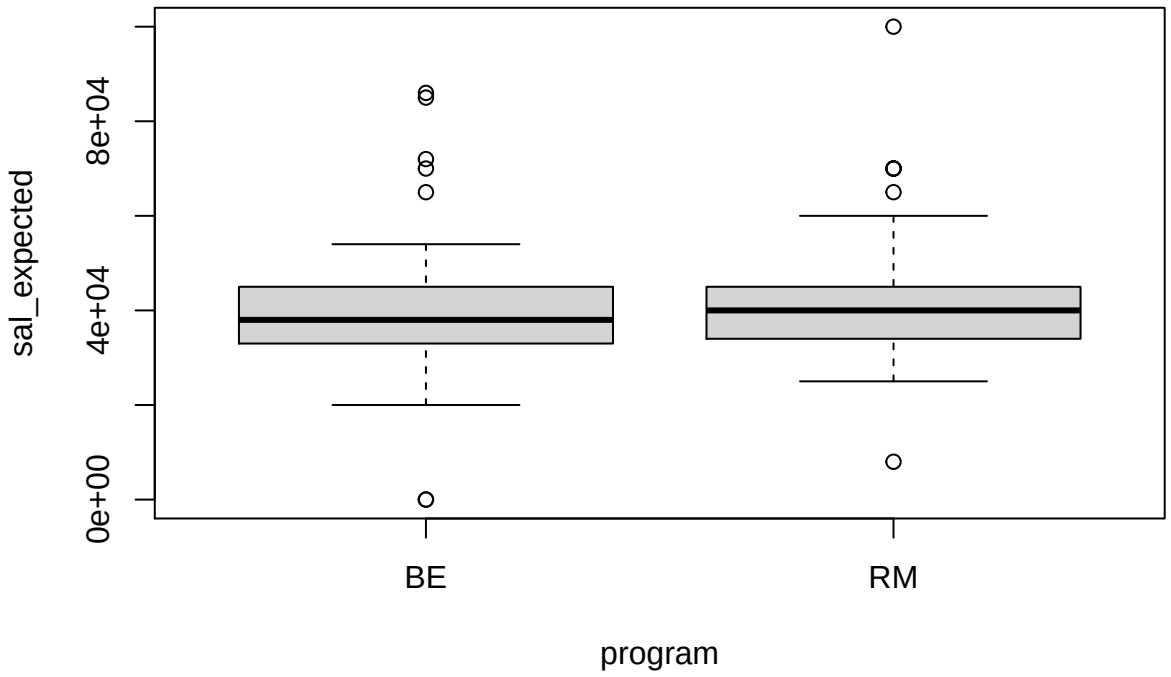
This looks neither aesthetically appealing nor informative, right? How do you think we can make it better?



Now let's try **box plot**. Similarly, if you need a quick understanding of box plot, take a look here, then come back. „a'



This looks nothing like the video. Why do you think it is and can you help me fix it?



Means and standard deviation (SD)

Calculate the mean and SD of a variable

Intuitively, you would type

```
mean(salaries[,"sal_expected"])
```

```
## [1] NA
```

```
sd(salaries[,"sal_expected"])
```

```
## [1] NA
```

but it doesn't work. why? Because we have NA in our data. To ask R to ignore the NA values,

```
mean(salaries[,"sal_expected"], na.rm = TRUE)
```

```
## [1] 56980.92
```

```
sd(salaries[,"sal_expected"], na.rm = TRUE)
```

```
## [1] 101193.9
```

Can I calculate the mean of salary expected of the RM, and the mean of salary expected of the BE students separately?

```
mean(salaries[salaries$program == "BE","sal_expected"], na.rm = TRUE)
```

```
## [1] 58844.09
```

```
mean(salaries[salaries$program == "RM","sal_expected"], na.rm = TRUE)
```

```
## [1] 52421.05
```

Can I calculate the mean of salary expected of the female RM student?

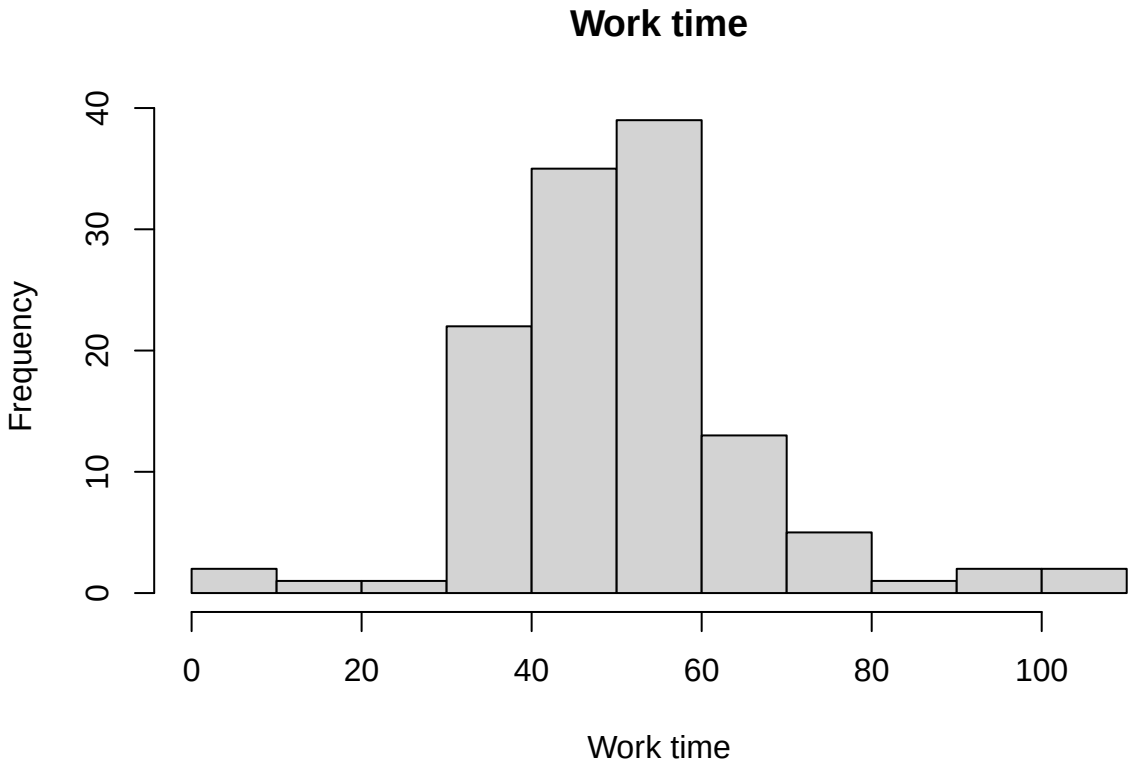
```
mean(salaries[salaries$program == "BE" & salaries$whatisyourgender1isfemale == "1", "sal_expected"], na
```

```
## [1] 46608.7
```

Your turn

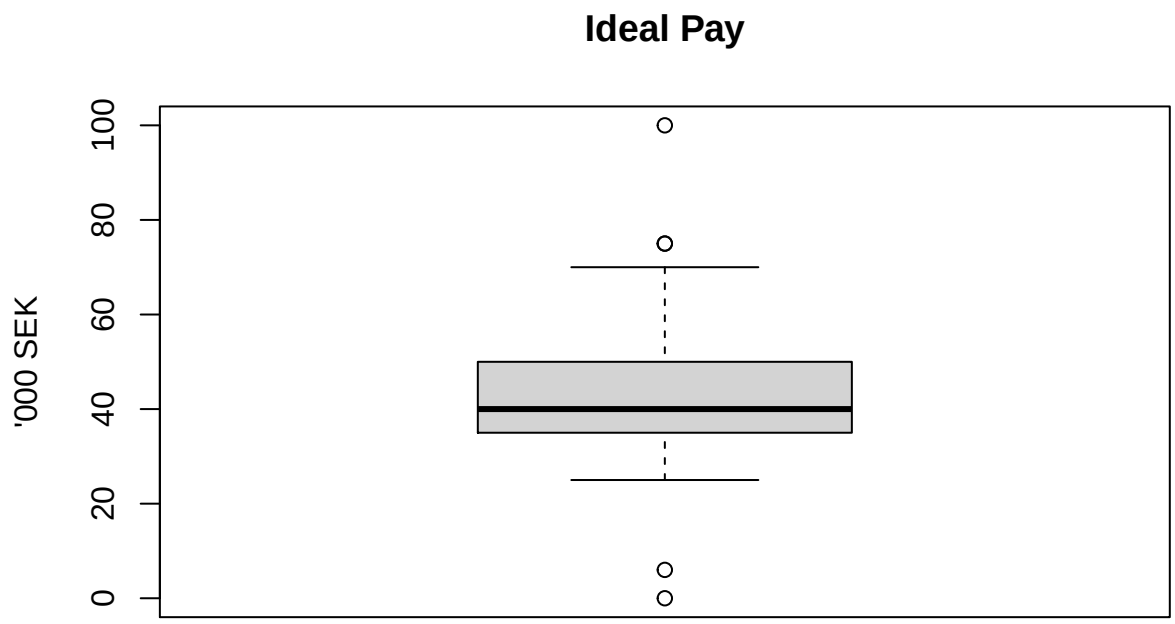
(1) Plot a histogram for `worktime`.

```
hist(salaries$worktime, main = "Work time", xlab = "Work time")
```



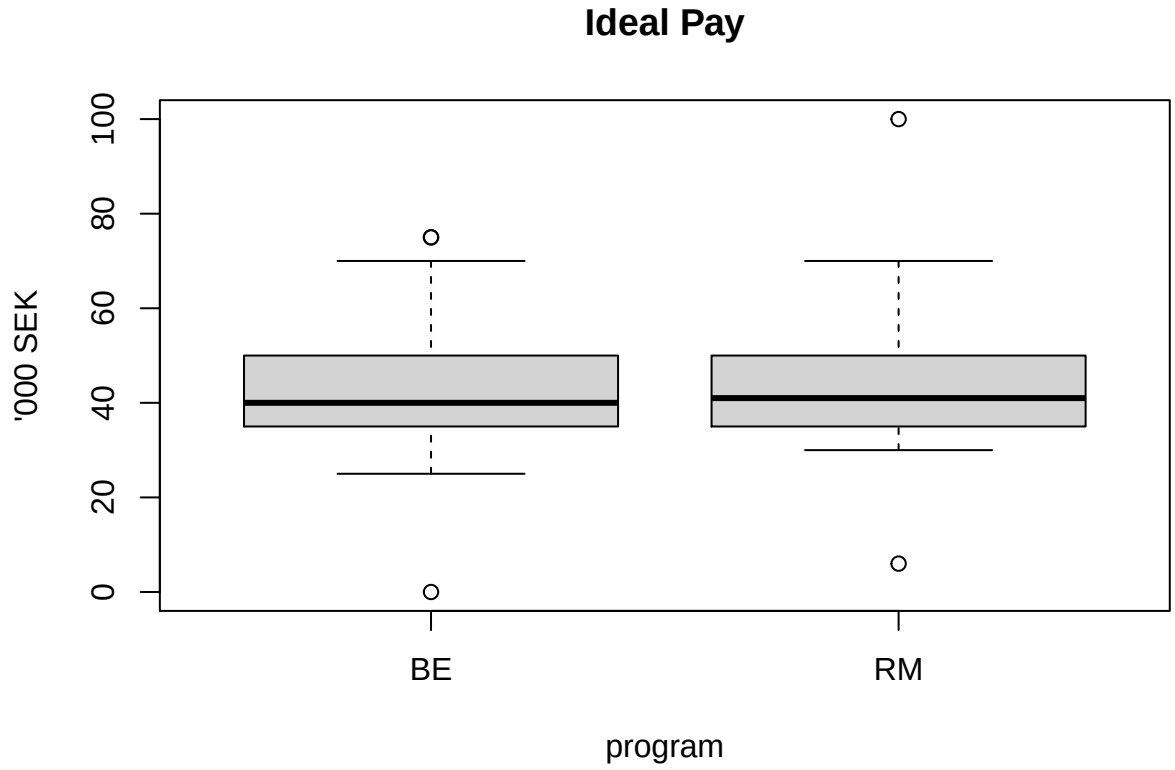
(2) Make a box plot for `idealpay`.

```
boxplot(salaries$idealpay/1000, ylim = c(0, 100), ylab = "'000 SEK", main = "Ideal Pay")
```



(3) Make a box plot for `idealpay`, for RM and BE students separately.

```
boxplot(idealpay/1000 ~ program, data = salaries, ylim = c(0,100),
        main = "Ideal Pay",
        ylab = "'000 SEK")
```



(4) Compute the mean and variance of `fairpay`.

```
mean(salaries$fairpay, na.rm = TRUE)
```

```
## [1] 43946.56
```

```r
var(salaries$fairpay, na.rm = TRUE)
```

```
## [1] 2067262507
```

(5) Compute the mean and variance of `fairpay` of the female students.

```r
mean(salaries[salaries$whatisyourgender1isfemale == 1, "fairpay"], na.rm = TRUE)
```

```
## [1] 40363.64
```

```r
var(salaries[salaries$whatisyourgender1isfemale == 1, "fairpay"], na.rm = TRUE)
```

```
## [1] 1579725159
```

(6) Compute the mean and variance of `fairpay` of the students who are female and belong to RM program.

```r
mean(salaries[salaries$program == "RM" & salaries$whatisyourgender1isfemale == 1, "fairpay"], na.rm
```

```
## [1] 36285.71
```

```r
var(salaries[salaries$program == "RM" & salaries$whatisyourgender1isfemale == 1, "fairpay"], na.rm
```

```
## [1] 117114286
```

Or,

try Assignment 1, Question 1.