Hello R!

NDH802 student

Quick recall

For your own good, you are strongly encouraged to go through the highly recommended reads on Canvas. As a quick recap, in RMarkdown, there are generally four components.

- **The YAML**: where you set the general format of your document (e.g., title, author, output). Within the scope of this course, you'll be working with html and pdf as document outputs. Now, try make it your report by changing "NDH802 student" in line 3 to "your name" (make sure you put "your name" in quotation marks).
- **The text**: where you write generally everything that is not the other three components, like what you're reading. You can format the text, e.g., *italics* and *italics* , **bold** and **bold**, superscript^2, ~~strikethrough~~. Unfortunately, unlike Word, what you type is not what you'll see in the final documents. You need to Knit it first. (More on this later)
- **The code chunks**: where the magic happens in RMarkdown. The first code chunk in this file is line 7 to 13 to install a package you need to translate a RMarkdown file to a pdf file. We will learn more about packages in Live session 2.
- **Inline code**: you can use inline code to refer to some values and it changes when your values change, for example 2021-03-09. Notice what you see in the RMarkdown file, and what you'll see in the final document. In this case, `Sys.Date()` prints the date you run the code, i.e., if you run this file in different dates, it prints out different values. As another example, notice how I integrate the `page` result in the last section. This can come in handy when you write up your assignments or other reports.

About the assignments

You will be asked to use RMarkdown to write up your assignments. Because:

- you can write your thoughts in words, code, print out the results, visualize your data, all in one place
- you don't have to copy your results from Excel and paste them into Word. Why do you do manual labor when you can automate things?
- you don't have to worry about formatting your document. RMarkdown will take care of that, so you can focus on "data analytics", which is why you are here
- I really hope it helps developing a new habit and forever enhance your workflow.

This can also happen. When your code cannot run or it produce errors because of, e.g., what I consider a typo, but I can see your (correct) thought process in there, you will get full score. It's because (1) your understanding is what matters most and (2) I am nice, in general.

Now the fun begins

Creating code chunk

Below, inside `{r}`, the shady zone, is the code chunk

Put your cursor at the end of this line and try Ctrl + Alt + I (OS X: Cmd + Option + I). What do you see?

Little tips: if you're a shortcut fan (like I am), try `Tools -> Keyboard Shortcuts Help` to learn what works in **your** computer. You can even modify/create your own shortcuts. For now, I suggest you go with the currently available ones.

R the calculator

Inside the chunk, type 2+2, then Ctrl(Windows)/Cmd(Mac) + Enter. What do you see?
Now let's create another chunk. You already know how right?

Now, type in x = 2+2, then Ctrl(Windows)/Cmd(Mac) + Enter. What do you see? No, you're not mistaken, nothing new appears this time.

Now, start a new line (press Enter) in the code chunk above, type x and press Ctrl(Windows)/Cmd(Mac) + Enter. What do you see? Also, observe the change(s) in the Global Environment on your top right.

You're starting to get a hang of it, right? When you press Ctrl(Windows)/Cmd(Mac) + Enter, you tell R to run the code on the same row with your cursor. I suggest you do it everytime you type in new code, make sure it works before you write the new ones.

Your turn

Try to use R to compute what you have learnt in the first lecture(s). Nothing fancy (yet), just use it as your calculator. For example, assume Huong gets 9 pt on Assignment 1, 10 pt on Assignment 2, and 7 pt on Assignment 3. The mean of Huong's three assignment is:

```
my_n = 3
my_mean = (9+10+7)/my_n
```

Easy peasy right? Now I'll compute the variance by:

```
my_variance = ((9-my_mean)^2 + (10-my_mean)^2 + (7-my_mean)^2)/(my_n-1)
```

Here, I just type in the exact formula in Canvas. The mean of my assignments points is 8.6666667. The variance of my assignments points is 2.3333333.

Your turn. Compute the **mean** and **variance** of *the number of likes* you get from the most 5 recent Intagram/Facebook posts (or all of your posts if you have less than 5).

```
your_n = 5
your_mean = 1
your_variance = 1
```

The mean of your Instagram/Facebook likes is 1. The variance of your Instagram/Facebook likes is 1.

If you are like, *"Why do I have to do this while it's equally easy to do it with a calculator?"*. You're absolutely right. In the next session, we will learn how to do it the more time- and energy-efficient way.

R the data manager

Load your data

Now let's create another chunk, copy and paste this `read.csv("https://raw.githubusercontent.com/lanhuongnguyen276,` (only the text inside and without) into your chunk, then Ctrl(Windows)/Cmd(Mac) + Enter. What do you see?

```
read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Preps/Before_Live_1/salaries
```

```
##    sal_expected sal_expected_other fairpay idealpay idealprod masters worktime
## 1         33000              31000   30000    35000         8       2       60
## 2         32000              32500   35000    34000         7       1       70
## 3         40000              40000   35000    50000         7       3       70
## 4         50000              40000   60000    70000         8       1       80
## 5         40000              35000   40000    40000         8       3      100
## 6         65000              50000   60000    65000        NA       1       70
## 7         50000              30000   35000    35000         7       3       45
## 8         50000              50000   50000    50000         9       3       60
## 9         50000             600000   40000    60000         7       1       40
## 10        37000              40000   33000    42000         6       2       60
## 11        33000              33000   33000    33000         8       1       50
## 12        50000              40000   40000    50000         7       3       55
## 13        35000              32000   35000    35000         9       3       60
## 14        30000              35000   30000    30000         8       1       60
## 15        38000              38000   40000    55000         7       2       50
## 16        36000              32000   32000    40000        NA       2       60
## 17        40000              32000   30000    45000         8       3       50
## 18        35000              30000   30000    35000         8       3       40
## 19        38000              36000   35000    45000         7       3       50
## 20        45000              40000   40000    60000         7       2       60
## 21        72000              65000   60000    75000         8       1       70
## 22        35000              30000   32000    35000         8       2       50
## 23       180000             350000  350000   350000         5       3       75
## 24        20000              20000   20000    25000         9       3       55
## 25        54000              50000   50000    55000         8       2       55
## 26        40000              40000   40000    45000         8       3       40
## 27        35000              35000   32000    34000         8       1       70
## 28        40000              40000   20000   200000         8       1       70
## 29        45000              35000   40000    38000         8       2       58
## 30        32000              28000   30000    35000         9       3       40
## 31        44000              35000   35000   500000        10       3       50
## 32        32000              31000   30000    40000         8       3       45
## 33        45000              33000   33000    50000         9       1       70
## 34        33000              33000   32000    36000         5       2       50
## 35        36000              36000   36000    38000         8       2       55
## 36        37000              35000   35000    40000         7       2       55
## 37        30000              30000   30000    32500         8       1       50
## 38        40000              38000   40000    45000         9       1       50
## 39        40000              28000   35000    40000         6       3       70
## 40        32000              32000   32000    32000         9       2       50
## 41        85000              70000   70000    75000         6       2       50
## 42        50000              40000   45000    55000         9       3       65
## 43        45000              40000   40000    50000         8       2       60
## 44        33000              30000   60000    70000         8       2       50
## 45        35000              35000   35000    40000         8       2       50
## 46        35000              35000   40000    40000         8       1       60
## 47        45000              37000   40000    50000         7       2      100
## 48        30000              30000   30000    30000         7       1       55
## 49        45000              35000   34000       NA         6       2       60
## 50        70000              43000   50000    70000         9       1       60
```

```
## 51       35000        33000   33000   37000         7        2        60
## 52       50000        45000   45000   50000         8        3        60
## 53       30000        30000   28000   32000         7        3        50
## 54       35000        35000   35000   37000         5        1        60
## 55       37000        34000   35000   38000         8        3        60
## 56       27000        30000   27000   35000         4        2        40
## 57           0        40000   40000   50000         8        3        60
## 58     1000000        20000   40000 1000000         8        2        60
## 59       50000        45000   35000   55000         8        2        55
## 60       50000        45000   50000   50000         7        3       110
## 61       34000        32000   30000   35000         7        1        60
## 62       30000        30000   30000   35000         6        3        50
## 63       28000        45000   35000   28000         6        2        40
## 64       32000        32000   30000   36000         7        3        45
## 65       47000        36000   42000   50000         6        2        60
## 66       35000        31000   31000   35000         9        3        40
## 67          NA           NA      NA      NA         7        2        NA
## 68       31000        36000   29000   36000         6        2        55
## 69       50000        35000   50000   70000         8        1        80
## 70       40000        40000   35000   60000        NA       NA        NA
## 71       35000        37000   40000   35000        NA       NA        NA
## 72       30000        30000   28000   30000         6        1        50
## 73      500000       100000   20000  500000        NA       NA        NA
## 74       36500        34000   38000   40000         9        2        50
## 75       30000        30000   30000   30000         8        3        32
## 76       40000        40000   47000   45000         8        2        70
## 77       40000        35000   35000   40000         8        2        60
## 78       20000        25000   20000   25000         7        1        40
## 79       30000        35000   30000   35000         8        3        50
## 80       32000        32000   32000   35000         8        3        45
## 81       40000        35000   40000   45000         6        1        55
## 82       50000        50000   50000   50000         8        3        80
## 83           0        35000       0       0         7        2        55
## 84       33000        31000   30000   36000         6        3        60
## 85       35000        25000   22500   35000         6        3       105
## 86      288000       288000  288000  288000         9        1        65
## 87       35000        40000   30000   45000         7        2        80
## 88       47000        42000   39000   50000         7        2        60
## 89          NA           NA      NA      NA         7        2        NA
## 90       35000        35000   35000   40000         8        3        70
## 91       40000        35000   35000   45000         7        1        60
## 92       42000        42000   42000   45000         9        2        60
## 93       86000        60000   70000  150000         9        1        60
## 94       40000        36000   35000   40000        NA       NA        NA
## 95       45000        45000   42500   45000         6        1        70
## 96       60000        60000   60000   60000         9        1        25
## 97       70000        50000   50000   60000         7        1        50
## 98       45000        50000   40000   55000         7        2         9
## 99       40000        34500   40000   40000         6        3        45
## 100     400000       350000  350000  400000         8        2        60
## 101      35000        30000   34000   33000         6        1        20
## 102      32000        32000   31000   35000         7        2        45
## 103       8000         6000    8000    6000         9        1        40
## 104      45000        45000   38000   50000         8        2        45
```

4

```
## 105      30000          30000 25000  45000           9        1       45
## 106      40000          35000 38000  42000          NA       NA       NA
## 107      45000          60000 45000  50000           8        1       50
## 108     100000          40000 80000 100000           9        1       50
## 109      36000          35000 40000  45000           6        2       40
## 110      30000          30000 30000  35000           9        1       35
## 111      35000          30000 30000  35000           7        1       40
## 112      25000          25000 25000  30000           8        1       50
## 113      40000          35000 35000  40000           9        3       50
## 114      34000          34000 34000  36000           9        2       40
## 115      70000          35000 35000  40000          10        3       50
## 116      45000          30000 30000  35000           9        2       50
## 117      40000          37000 45000  48000          NA       NA       NA
## 118      45000          45000 50000  50000           9        3       40
## 119      32000          32000 35000  37000           9        3       40
## 120      40000          29000 29000  40000           8        3       60
## 121      45000          40000 40000  47000           8        1       40
## 122      30000          25000 30000  30000           6        2       45
## 123      45000          45000 40000  45000           8        1       40
## 124      30000          30000 30000  30000           2        3       60
## 125      25000          25000 25000  35000           8        2       10
## 126      35000          37000 35000  36000           6        3       40
## 127      35000          33000 35000  35000           7        3       40
## 128      70000          40000 60000  50000          NA       NA       NA
## 129      40000          35000 30000  35000           8        2       50
## 130      40000          35000 40000  60000          NA       NA       NA
## 131      70000          55000 55000  70000           8        1       40
## 132      65000          60000 65000  65000          10        1       40
## 133      40000          28000 37000  50000           8        1       90
##     howmanymon whatisyourgender1isfemale program
## 1          6.0                         2      BE
## 2         10.0                         2      BE
## 3         12.0                         2      BE
## 4          0.0                         2      BE
## 5          0.0                         2      BE
## 6          6.0                         2      BE
## 7          2.0                         1      BE
## 8         12.0                         2      BE
## 9          6.0                         2      BE
## 10         6.0                         2      BE
## 11         8.0                         2      BE
## 12         0.0                         2      BE
## 13         6.0                         2      BE
## 14         2.0                         2      BE
## 15         4.0                         1      BE
## 16         0.0                         2      BE
## 17        12.0                         2      BE
## 18         3.0                         2      BE
## 19        12.0                         2      BE
## 20         2.0                         2      BE
## 21         1.0                         2      BE
## 22         0.0                         1      BE
## 23        18.0                         2      BE
## 24         0.0                         2      BE
```

```
## 25      12.0              2      BE
## 26       6.0              2      BE
## 27      12.0              2      BE
## 28       8.0              1      BE
## 29       2.0              2      BE
## 30       0.0              2      BE
## 31       6.0              2      BE
## 32      40.0              1      BE
## 33       2.0              2      BE
## 34      10.0              2      BE
## 35       4.0              1      BE
## 36       4.0              2      BE
## 37       0.0              2      BE
## 38      18.0              2      BE
## 39       3.0              2      BE
## 40       3.0              2      BE
## 41       9.0              2      BE
## 42       4.0              2      BE
## 43       4.0              2      BE
## 44      18.0              2      BE
## 45       3.0              2      BE
## 46       3.0              1      BE
## 47       7.0              2      BE
## 48       1.5              2      BE
## 49       6.0              2      BE
## 50      15.0              1      BE
## 51       6.0              2      BE
## 52       4.0              2      BE
## 53      24.0              1      BE
## 54       3.5              1      BE
## 55       3.0              2      BE
## 56       8.0              1      BE
## 57      20.0              2      BE
## 58       0.0              2      BE
## 59      12.0              2      BE
## 60       0.0              1      BE
## 61       1.5              1      BE
## 62       3.0              2      BE
## 63       3.0              2      BE
## 64      24.0              1      BE
## 65       4.0              2      BE
## 66      20.0              1      BE
## 67        NA              2      BE
## 68       4.0              2      BE
## 69       6.0              2      BE
## 70        NA             NA      BE
## 71        NA             NA      BE
## 72       1.0              2      BE
## 73        NA             NA      BE
## 74       4.0              2      BE
## 75      24.0              1      BE
## 76      20.0              1      BE
## 77       3.0              2      BE
## 78       0.0              1      BE
```

```
## 79      12.0               1     BE
## 80      18.0               2     BE
## 81       8.0               2     BE
## 82      24.0               1     BE
## 83      36.0               1     BE
## 84      12.0               2     BE
## 85       1.0               2     BE
## 86      12.0               1     BE
## 87      12.0               2     BE
## 88      12.0               2     BE
## 89        NA               2     BE
## 90       0.0               1     BE
## 91       5.0               2     BE
## 92       6.0               2     BE
## 93       2.0               2     BE
## 94        NA              NA     BE
## 95       7.0               2     BE
## 96      24.0               1     RM
## 97       0.0               2     RM
## 98      16.0               2     RM
## 99       8.0               2     RM
## 100      0.0               2     RM
## 101      6.0               1     RM
## 102      9.0               1     RM
## 103      0.0               2     RM
## 104     24.0               1     RM
## 105     12.0               1     RM
## 106       NA              NA     RM
## 107     18.0               1     RM
## 108      0.0               2     RM
## 109     24.0               1     RM
## 110     12.0               1     RM
## 111      2.0               1     RM
## 112      0.0               1     RM
## 113      2.0               1     RM
## 114     18.0               1     RM
## 115      0.0               2     RM
## 116     18.0               1     RM
## 117       NA              NA     RM
## 118     36.0               1     RM
## 119     10.0               2     RM
## 120      6.0               2     RM
## 121     15.0               1     RM
## 122      3.0               1     RM
## 123      2.0               2     RM
## 124     30.0               1     RM
## 125      8.0               1     RM
## 126     14.0               1     RM
## 127      6.0               2     RM
## 128       NA              NA     RM
## 129     12.0               1     RM
## 130       NA              NA     RM
## 131      6.0               2     RM
## 132     24.0               1     RM
```

```
## 133          6.0                           2       RM
```

Similar to giving 2+2 a shorter name "x", you can name your data table. Now try paste it to your code chunk

```
salaries = read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Preps/Before_Liv
```

```r
salaries = read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Preps/Before_Liv
```

This is loading your data into R Environment. From now on (until you restart your R session), whenever you type `salaries`, R will understand you refer to the above data.

Explore your data

These commands give you a quick overview of your data. Try each of them, by putting your cursor in each of the three rows and press Ctrl(Windows)/Cmd(Mac) + Enter. What do you see?

```r
head(salaries)
```

```
##   sal_expected sal_expected_other fairpay idealpay idealprod masters worktime
## 1        33000              31000   30000    35000         8       2       60
## 2        32000              32500   35000    34000         7       1       70
## 3        40000              40000   35000    50000         7       3       70
## 4        50000              40000   60000    70000         8       1       80
## 5        40000              35000   40000    40000         8       3      100
## 6        65000              50000   60000    65000        NA       1       70
##   howmanymon whatisyourgender1isfemale program
## 1          6                         2      BE
## 2         10                         2      BE
## 3         12                         2      BE
## 4          0                         2      BE
## 5          0                         2      BE
## 6          6                         2      BE
```

```r
tail(salaries)
```

```
##     sal_expected sal_expected_other fairpay idealpay idealprod masters worktime
## 128        70000              40000   60000    50000        NA      NA       NA
## 129        40000              35000   30000    35000         8       2       50
## 130        40000              35000   40000    60000        NA      NA       NA
## 131        70000              55000   55000    70000         8       1       40
## 132        65000              60000   65000    65000        10       1       40
## 133        40000              28000   37000    50000         8       1       90
##     howmanymon whatisyourgender1isfemale program
## 128         NA                        NA      RM
## 129         12                         1      RM
## 130         NA                        NA      RM
## 131          6                         2      RM
## 132         24                         1      RM
## 133          6                         2      RM
```

```r
summary(salaries)
```

```
##   sal_expected    sal_expected_other    fairpay          idealpay
## Min.   :      0  Min.   :  6000     Min.   :     0  Min.   :      0
## 1st Qu.:  33000  1st Qu.: 31500     1st Qu.: 30000  1st Qu.:  35000
## Median :  40000  Median : 35000     Median : 35000  Median :  40000
## Mean   :  56981  Mean   : 48046     Mean   : 43947  Mean   :  66673
## 3rd Qu.:  45000  3rd Qu.: 40000     3rd Qu.: 40000  3rd Qu.:  50000
## Max.   :1000000  Max.   :600000     Max.   :350000  Max.   :1000000
## NA's   :2        NA's   :2          NA's   :2       NA's   :3
##   idealprod        masters         worktime        howmanymon
## Min.   : 2.000  Min.   :1.000   Min.   :  9.00  Min.   : 0.000
## 1st Qu.: 7.000  1st Qu.:1.000   1st Qu.: 45.00  1st Qu.: 2.000
## Median : 8.000  Median :2.000   Median : 55.00  Median : 6.000
## Mean   : 7.569  Mean   :2.008   Mean   : 54.67  Mean   : 8.654
## 3rd Qu.: 8.000  3rd Qu.:3.000   3rd Qu.: 60.00  3rd Qu.:12.000
## Max.   :10.000  Max.   :3.000   Max.   :110.00  Max.   :40.000
## NA's   :10       NA's   :8       NA's   :10      NA's   :10
## whatisyourgender1isfemale   program
## Min.   :1.000              Length:133
## 1st Qu.:1.000              Class :character
## Median :2.000              Mode  :character
## Mean   :1.648
## 3rd Qu.:2.000
## Max.   :2.000
## NA's   :8
```

Knit your document,

For the assignments, you will submit both the RMarkdown file (so something like this), and a pdf document. First you need to install a sort of software by running the below code chunk, then put a # before them, like `#install.packages("tinytex")` and `#tinytex::install_tinytex()`. By this, we are making them comments, and they won't be run.

```
#install.packages("tinytex")
#tinytex::install_tinytex()
```

Now, click Knit (top left of the RMarkdown screen) or, press Ctrl/Cmd + Shift + K, click "Install"/"Yes" to the windows popping up. Observe the pdf document (appreciate your hard work) then come back here (please!!!).
Note: If you can not knit pdf document, please change it to html_document in line 4, try again and shoot me a quick message.

Did you freeze for a second when you saw 5 pages of data? `salaries` has 365 rows. In the assignments, you will be working with 1,000,000 rows which is appx. 13699 pages. Isn't that crazy? So, when you write your assignments, please

- don't print out the data, or things that are not relevant to the questions. If there is some cool code that you want me to read, but you don't want to run, make them a "comment" (more on this in the live session)
- knit your document frequently, see if it works, if you like how it looks. Just like trying your code, if it doesn't work, you want to know *soon*.

Tips. When you name your Rmd file, you may **not** want to include spaces, dots and commas. For example, this, "NDH802 assignment 1, group 1", is **not** recommended. R is smart and will try to "fill

in the blank", but in my experience, it doesn't work perfectly every time. Therefore, to avoid unnecessary trouble, you should try "NDH802-assignment-group-1", "NDH802_assignment_1_Group_1", or "NDH802Assignment1Group1" instead. On the other hand, you can be more creative with the `title` (the text in line 2).

Phew,

I think that's enough for now. How do you feel?
If you're still here, thank you very much for your effort. It will pay off. If you have questions, please bring them to the class, or start a discussion on Canvas, or shoot me a message, or send me an email. The more you ask at the early stage, the less you do later.

Huong