# Outline for today

- Go over Q1 & some terminology.
- This week, we are having a presentation exercise based on Q2.
- Group Discussion for Q2
- Group presentation preparation
- Group presentation

Each group of 4 will have around 5 to 6 minutes to present what you discuss on Q2

# Some terminology

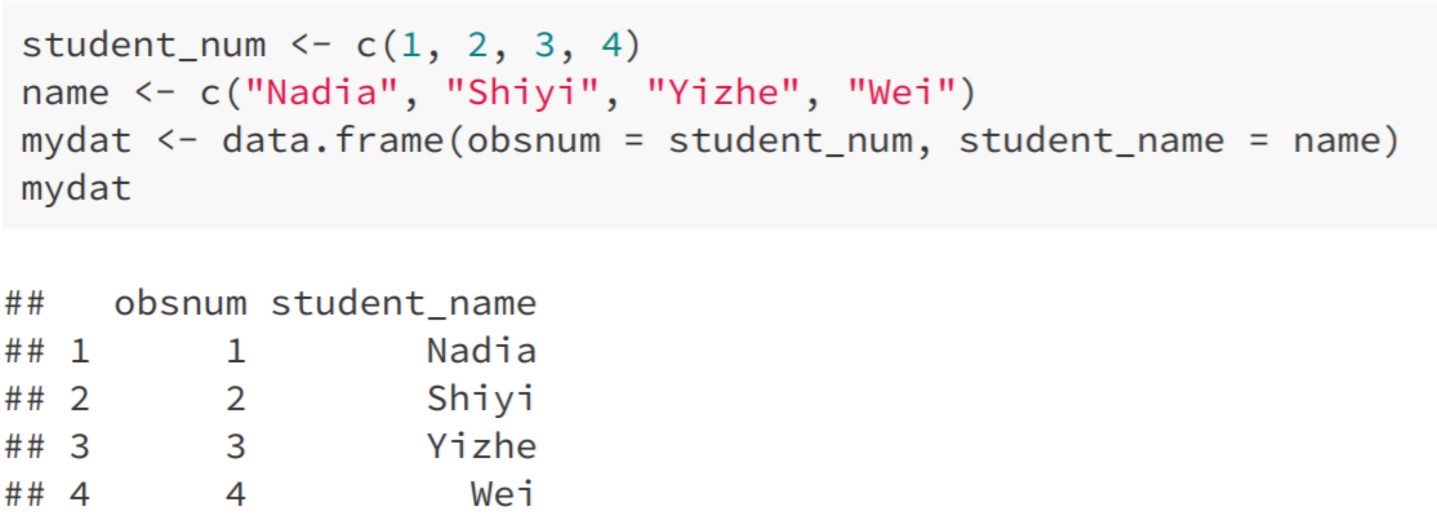
- Data Frame

- Data frames are the two-dimensional version of a list. They are the most useful storage structure for data analysis
- Think of it as a table to store data.
- How do we create a data frame in R?
- By using data.frame() function

- 1D-Example: in Q1
- aves <- replicate(n = 50, expr = ave_years(500))
- dat <- data.frame(aves)

- 2D-Example from slides:
- Student number & name

has 1-1 correspondence.



```r
student_num <- c(1, 2, 3, 4)
name <- c("Nadia", "Shiyi", "Yizhe", "Wei")
mydat <- data.frame(obsnum = student_num, student_name = name)
mydat
```

```
##   obsnum student_name
## 1      1        Nadia
## 2      2        Shiyi
## 3      3        Yizhe
## 4      4          Wei
```

# Variance & Standard deviation

- Important concept in Statistics.
- Relationship: sd = $\sqrt{Var}$
- What do they tell you about the data set?
- They tell you how far the data are spread out from their average value.
- High variance means your data set is **spread**.
- In R, how do we calculate the variance & standard deviation of the data set?
- By using sd() function
- Example: data set: student_height, you do: **sd(student_hight)** & **(sd(student_hight))^2** to calculate sd & var

# Question 1

- What is the shape of the distribution of the average years of study for each year?

The shape of the distribution is symmetric, and centered around aves = 2.5.

Another thing to notice is the spread, the average year of study spreads from 2.40 to 2.60 approximately.

# Group Discussion on question 2

- Focus on (and describe or explain to each other) the following:
- a. Describe what you did to create the variable.
- b. Explain why you did it this way.
- c. Compare graphs or summary statistics on the created variables.

- For example, for question 1, we created a variable year to represent each UofT student, and we want to see what is the distribution of the average year of study for each year.

# Group presentation preparation

- Each group will have around 5 to 6 minutes to present what they discussed on question 2.

# Possible speaking template for group presentation

- Introduce the variables you want to work with.
- Define the problem you want to solve. (e.g. what is the relationship between parents height & their kids')
- Explain how the problem you defined in the previous step can be solved with the variables.
- What conclusion can you draw?
- Use plot and graph to support your conclusion you make from the data set. (explain what graphs do you use and describe them accordingly.)
- Summarize the results.

# Group presentation!