# STA130 Week 1 Tutorial

Kaixuan (Bryan) Hu

September 12, 2019

# Introduction

- Tutorial: 20% of the final grade.
- Attendance is mandatory.
- 1 point for attendance, 1 point for practice problem, and 4 points for writing/presentation exercise.
- 2nd week will be a half tutorial, the second half is reserved for the mentorship program which is worth 3% of the final grade.

- Have questions completed and submitted to Quercus, no emailed homework will be accepted.
- Tutorial is NOT the place for troubleshooting R code. You should be prepared to discuss your results during tutorial. Instead, go to OH or post questions to the discussion board ahead of tutorial.
- Participate in group work and class discussions
- You will get practice for writing/ oral presentation skills during the tutorial
- Show up on time, tutorial starts at 2:10PM. If they need to miss a tutorial or leave early for a test, you should let me know ahead of time.
- Tutorial is a safe and friendly environment to practice your communication skills (critical skills for the future careers, in statistics or otherwise)

# Introduce yourself!

- Tell us about your name, your preferred name and what do you want to study in UofT?
- What do you like to do in your spare time? (Sports, movie, music, game, etc.)
- One fun fact about yourself.
- Or anything else you would like to share.

# Outline for today

- Learn how to describe data visualizations aided by the vocabulary list.
- Later you will be assigned to small groups for group discussion regarding question 2 from the homework

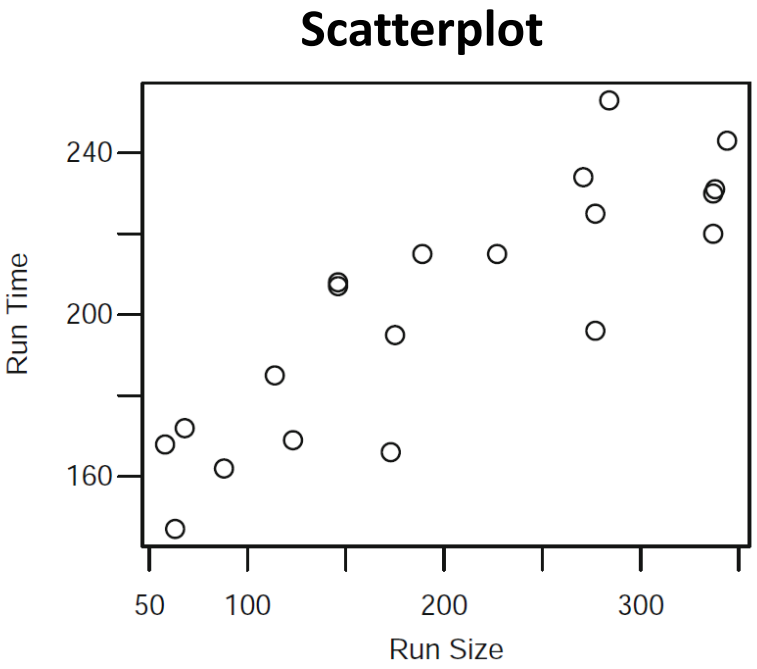
# Data Visualization

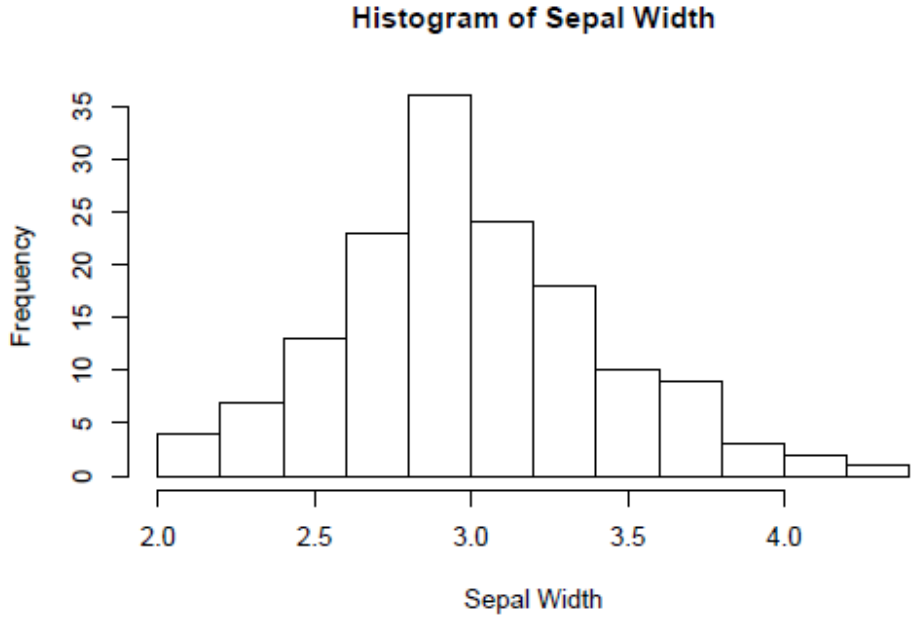
- What a visualization does and how to describe a visualization?

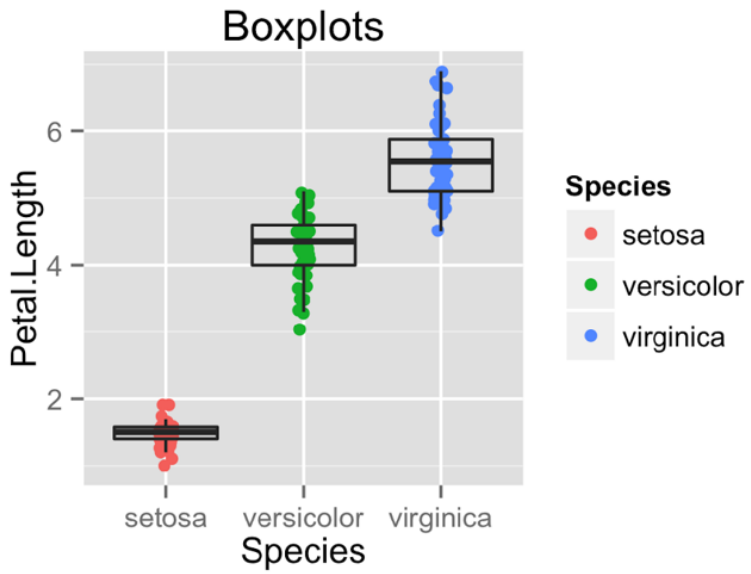
Basically, Exploring a dataset using graphs and come up with a story to describe the dataset.

- What are the most effective types of graphs to summarize information in categorical or quantitative (numerical) variables?

Typically, **histogram**, **scatterplot, Boxplots** are the most commonly used graphs.




How do you describe them?

# Key things to look at:

- Look for **distribution**

What does the distribution tell you about for each types of data?

e.g. Normally distributed;  Symmetric;  Skewed right;  Skewed left;
Bimodal (i.e. double peak) distributed;  Multimodal (i.e. multiple peak) distributed.

Use Statistics terminology! (Vocabularies when describing distributions of variables or relationships between two variables)

- Look for **relationships between two variables (e.g. trend)**

e.g. Linear (positive or negative) / non-linear relationship;

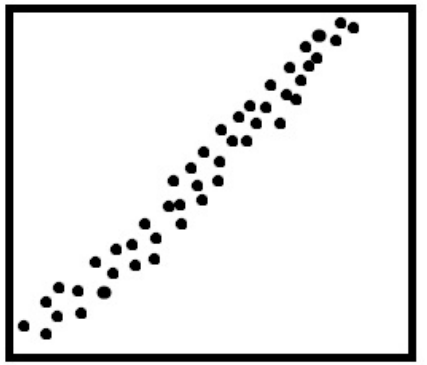
- Boxplots, histograms:

1. Where it is centred (towards the left, right, middle)
2. How much spread? (relative to what?)
3. The tails relative to a normal distribution (fat-tailed or heavy-tailed & thin-tailed)
4. Modes (i.e. peak): where, how many? (unimodal, bimodal, multimodal, uniform)
5. Symmetric; Skewness (left-skewed, right-skewed)
6. Outliers; Extreme values
7. Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

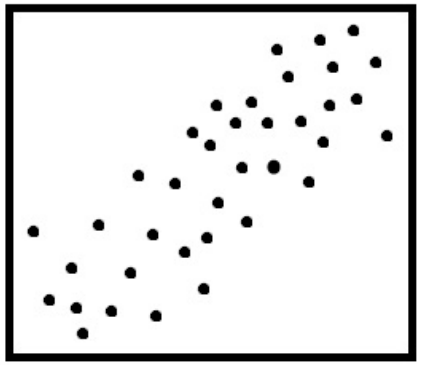
- Scatterplots:

Used to compare the relationship between two variables

1. Strong / weak relationship
2. Linear (positive or negative) / non-linear relationship
3. Outliers (deviation from what?)
4. Any visible clusters forming.



strong positive linear association



weak positive linear association

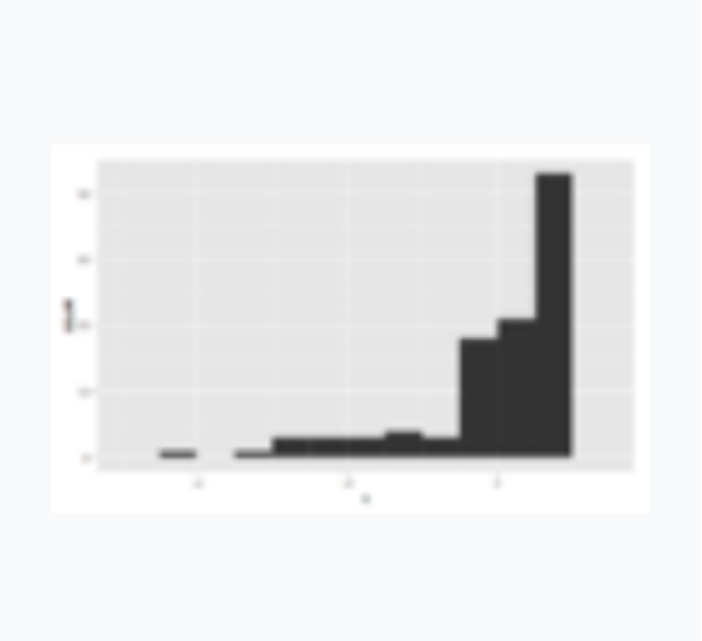
One of the **objective** of this course is to get you familiar with the statistical language, so it is helpful to understand and memorize the word or phrase shown above.

# Try it by yourself

- Describe the following graphs using one or more correct (and precise) vocabularies.

Symmetric or unimodal

Skewed right



Skewed left

Bimodal

multimodal

Symmetric

- Some good questions from the office hour
- What does column, variables, rows, observation refers to?

```
### (b) Use the `glimpse()` function to view properties of the `oly12` dataset.
How many observations does it include? How many variables are measured for each
observation? How many rows and columns does the `oly12` data frame have?

```{r}
### Type your code here
glimpse(oly12)
```
```

```
Observations: 10,384
Variables: 14
$ Name    <fct> Lamusi A, A G Kruger, Jamale Aarrass, Abdelhak Aatakni, Mar…
$ Country <fct> People's Republic of China, United States of America, Franc…
$ Age     <int> 23, 33, 30, 24, 26, 27, 30, 23, 27, 19, 37, 28, 28, 28, 22,…
$ Height  <dbl> 1.70, 1.93, 1.87, NA, 1.78, 1.82, 1.82, 1.87, 1.90, 1.70, N…
$ Weight  <int> 60, 125, 76, NA, 85, 80, 73, 75, 80, NA, NA, NA, 60, 64, 62…
$ Sex     <fct> M, M, M, M, F, M, F, M, M, M, M, M, F, F, M, F, M, M, M, M,…
$ DOB     <date> 1989-02-06, NA, NA, 1988-09-02, NA, 1984-06-09, NA, 1989-0…
$ PlaceOB <fct> NEIMONGGOL (CHN), Sheldon (USA), BEZONS (FRA), AIN SEBAA (M…
$ Gold    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Silver  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Bronze  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Total   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Sport   <fct> Judo, Athletics, Athletics, Boxing, Athletics, Handball, Ro…
$ Event   <fct> "Men's -60kg", "Men's Hammer Throw", "Men's 1500m", "Men's …
```

- What is the difference between numerical variables and categorical variables?
- When do we use bar plot, histogram, and Scatterplots? (i.e. to describe what type of variables?)
- What are the corresponding R-functions for those plots?
- When using histogram, difference between bins_number and binwidth?
- Tips for doing homework: always check the slides first, the slides contain very similar solution that you only need to modify a few things to get the answer for homework.

# Group Discussion

- Form a groups of 3-4
- FIRST 5 mins: Ice breaker activity. Next, discuss about questions from the homework.
- And explain to each other:
  - What do you notice about the number of bins a histogram has, its shape and precision?
  - In Question 2e, you could have presented both sexes in the same plot or presented them on separate plots? What are some considerations for which presentation you may want to choose (e.g. what are the pros and cons of each one)?
  - If presenting two plots side by side, what are some things to consider to ensure they are comparable and reader-friendly?

# synthesizing information from visualizations

- You should think about the most logical order in which to lead the reader through the visual.
- *Possible writing templates:*
- a. Describe what their graphs are telling us; i.e., what type of relationship(s) are apparent, the x- and y-axis labels should be clear, etc.
- b. Come up with a “story” of your main results, use a few of your graphs. You can make one not originally asked in the question, if it will help and is appropriate; e.g. you are interested in describing what the characteristics of the average Olympic athlete; striking features you observed; compare and contrast features; etc.
- c. Consider the logical order that you ‘tell your story’ – how will you most effectively tell your ‘story’.
- d. Make sure you provide figures to support your ‘story’.

Other things to notice:

- Give some context to the variables you are graphing based on what you know about the dataset (units and types of variables involved should be clear)

Either:

- Give the most striking features of the graphs (contrast or similarity).

-Synthesize these features and make a conclusion based on these features.

Or:

- Make a statement or conclusion based on your impression.
- Explain each of the features of the graphs (contrast or similarity) that support your statement or conclusion.

*Vocabulary list*

**Histograms, boxplots, bar graphs:**

1. Where it is centred (towards the left, right, middle)
2. How much spread? (relative to what?)
3. The tails relative to a normal distribution (fat-tailed or heavy-tailed & thin-tailed)
4. Modes (i.e. peak): where, how many? (unimodal, bimodal, multimodal, uniform)
5. Symmetric; Skewness (left-skewed, right-skewed)
6. Outliers; Extreme values
7. Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

**Scatterplots**:

1. Strong / weak relationship
2. Linear (positive or negative) / non-linear relationship
3. Outliers (deviation from what?)
4. Any visible clusters forming.

# Writing exercise

- Write a short paragraph to describe coherently the graphs you produced from question 2 and structure these graphs to tell a story.
- Use at least 3 graphs from question 2 to support the story.
- HAND IN for evaluation via Quercus by the end of tutorial.
- I would suggest you write in a word processor (e.g. MS Word) to avoid losing any work. i.e. don’t type your work directly in Quercus.