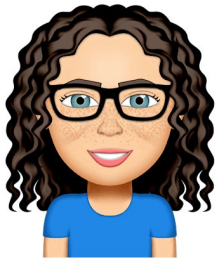
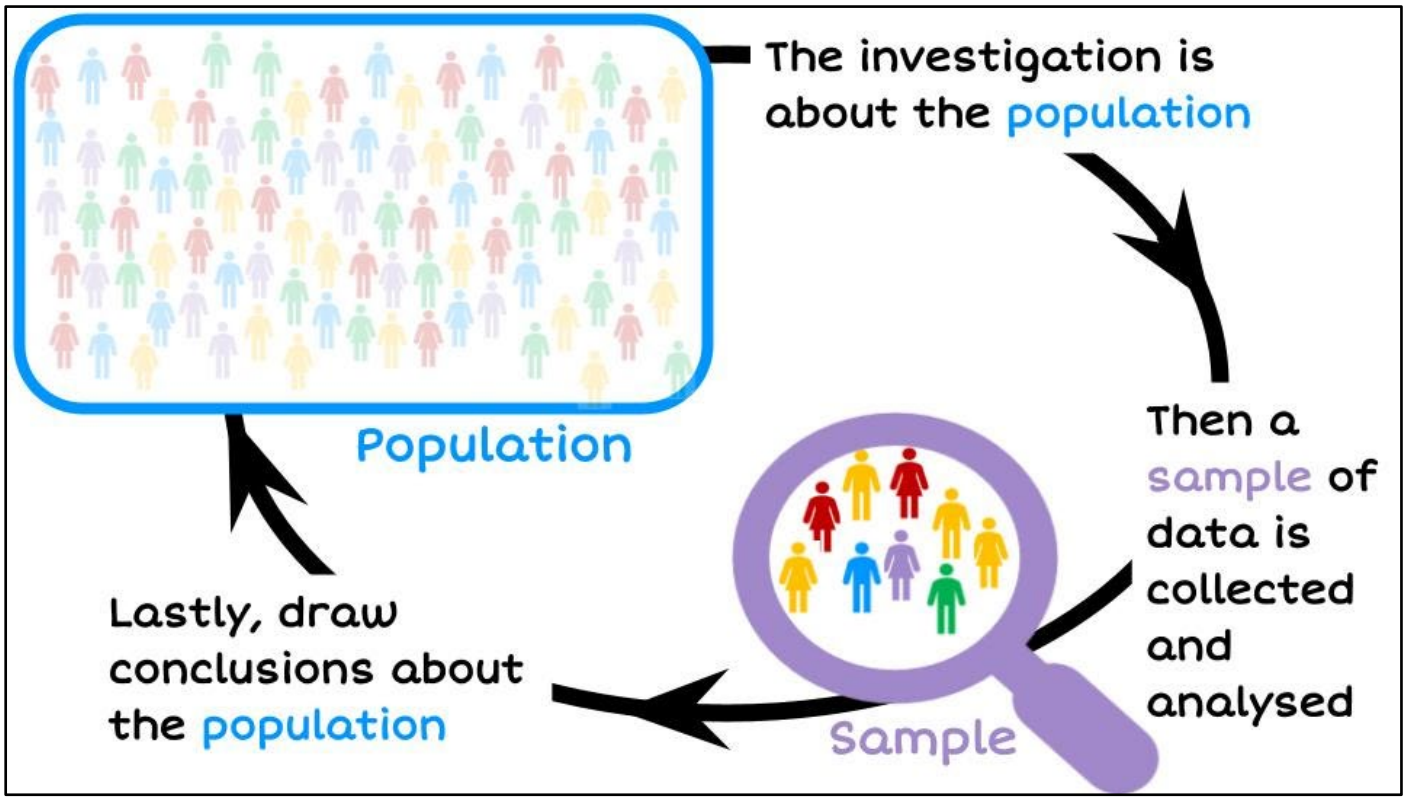
Level 2 Inference Workbook

Name:

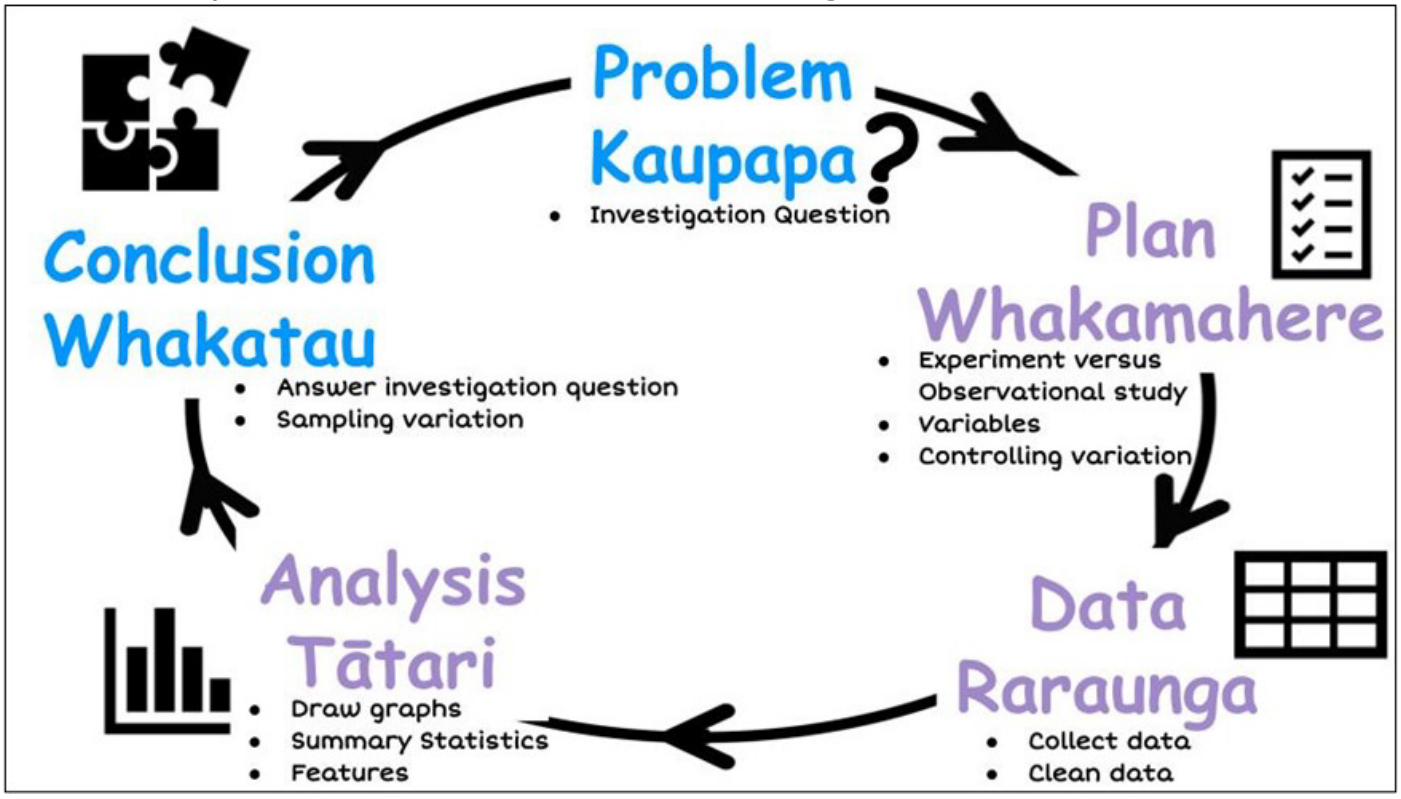


By Liz Sneddon

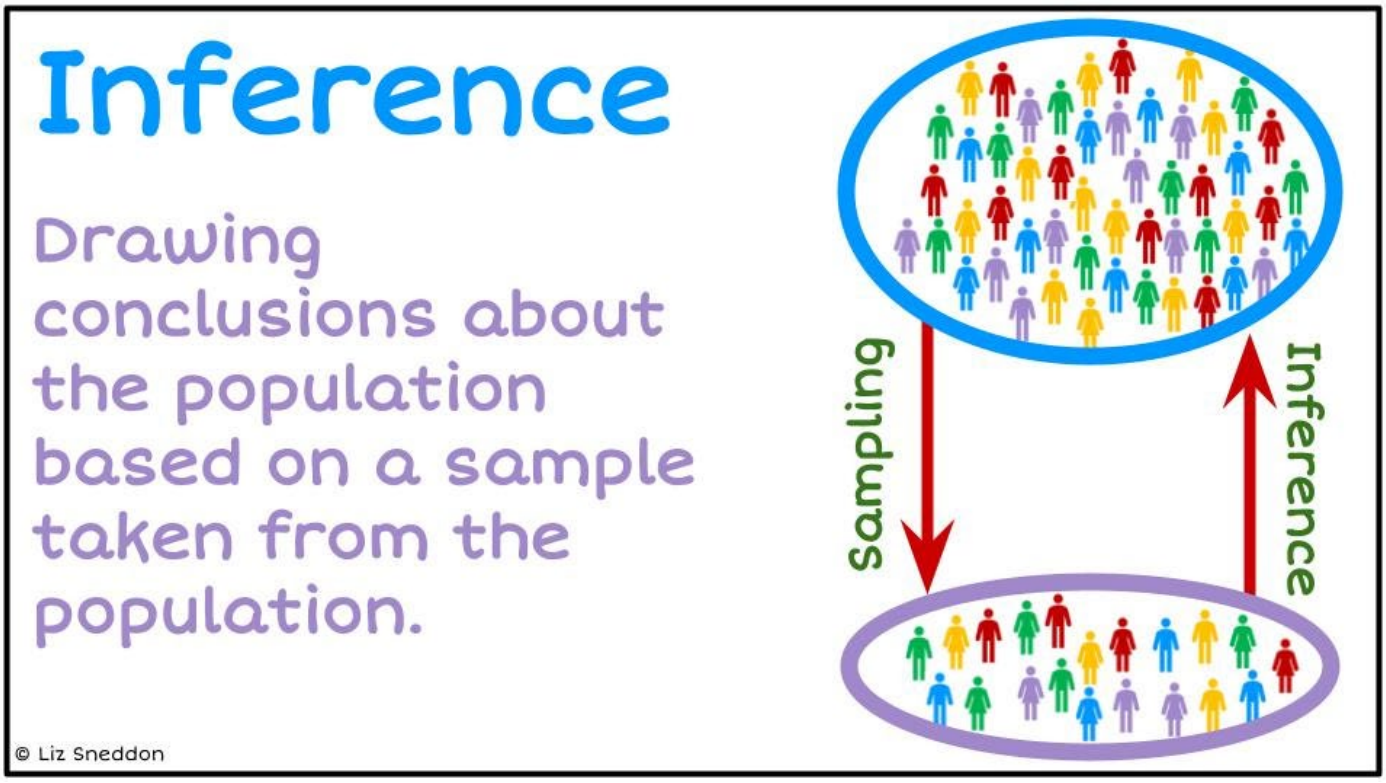
Investigative Process



The PPDAC cycle is the core of all statistical investigations.

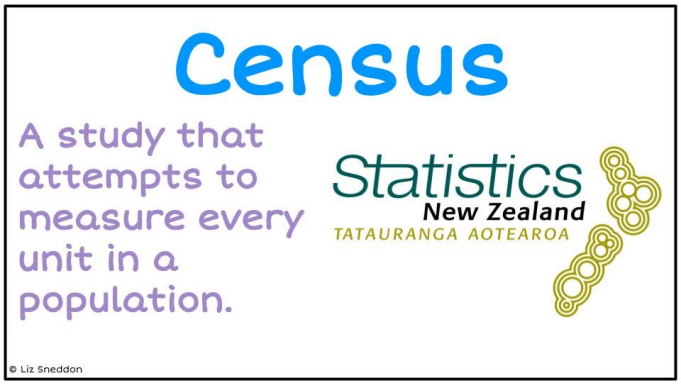


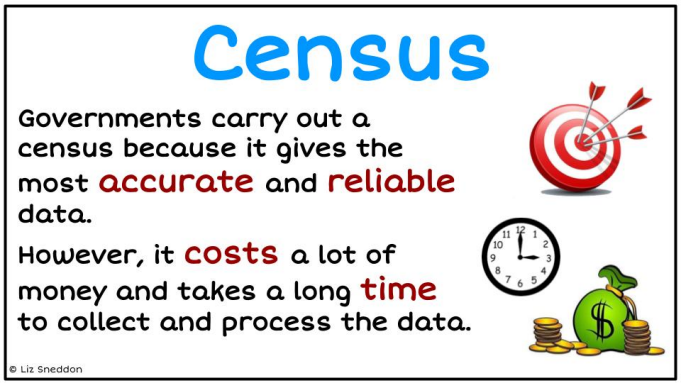
At the end of this investigation, you will need to make an Inference.



Populations and samples

We start with an investigation question about a population. We often have a hypothesis or prediction of what we expect to find.





The government uses this information to help it decide things like:

- Where to build new schools (if there are a lot of young children in one area, they will need a school)
- How many hospitals do we need?
- Roading decisions (e.g., motorways and the harbour bridge),
- Etc.

Exercise 1:

1) Describe the population at your school. Think about ages, genders, ethnicities, etc.

2) Why do we usually take a sample rather than a census?

3) What does the government use the data collected from the census for? Explain.

4) A beverage company wanted to see if people in the United States liked their new logo. Which choice best represents a **population**?

A. A selection of logo artists.

B. Every person in the United States.

C. A selection of shoppers from different states.

D. 3,800 children aged 5 - 15

5) A musician wanted to see what people who bought his last album thought about the songs. Which choice best represents a **sample**?

A. Every person who bought the album.

B. A selection of people who didn't want to buy the album.

C. 250 girls who bought the album.

D. A selection of 3,294 people who bought the album.

6) A gaming website wanted to find out which console its visitors owned. Which choice best represents a **population**?

A. Visitors to the 3DS section.

B. All of the website visitors.

C. Visitors to the PS4 section.

D. Visitors who are on the website for more than 5 minutes.

7) Before a nationwide election, a polling place was trying to see who would win. Which choice best represents a **sample**?

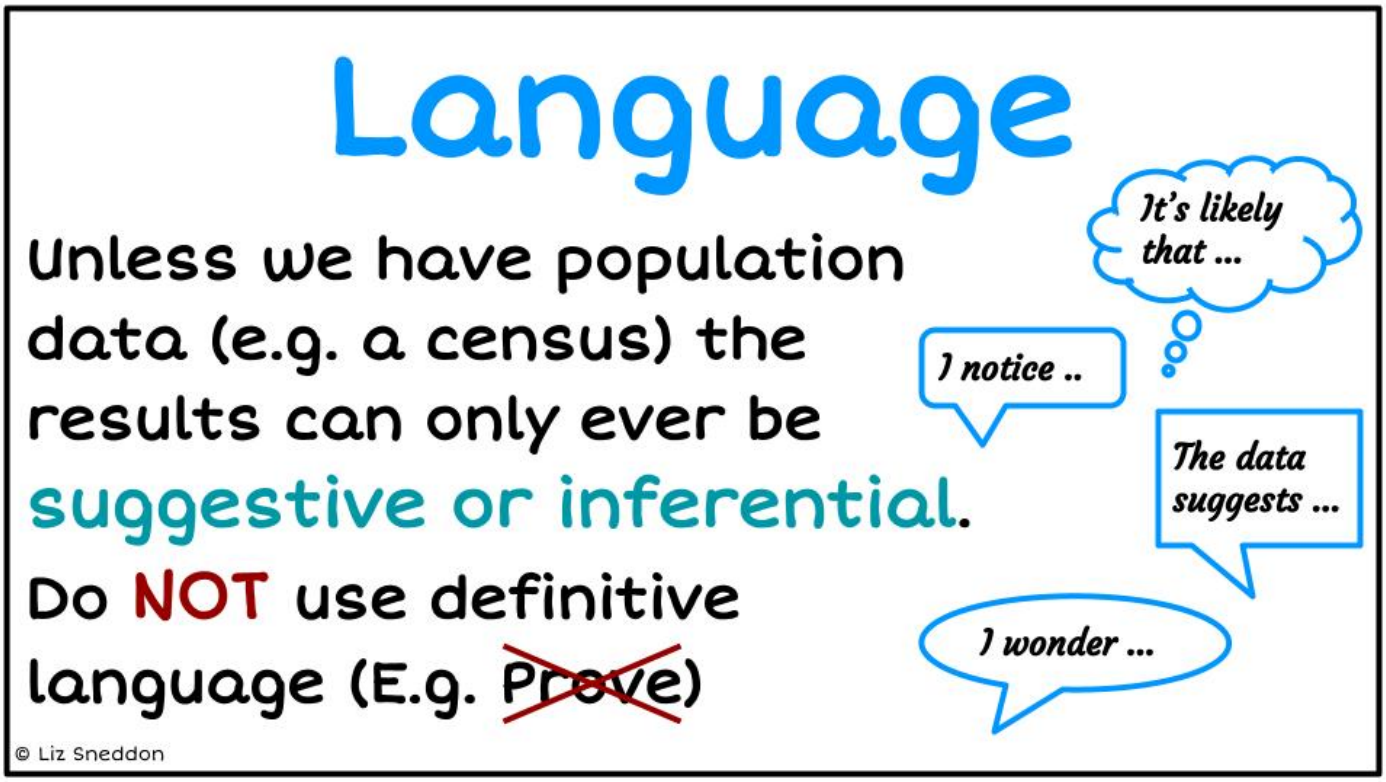
A. A selection of voters over age 50.

B. A selection of male voters.

C. A selection of voters of different ages.

D. All voters

Language



Exercise 2:

Brainstorm ideas for more words that are suggestive versus definitive.

Suggestive words	Definitive words
Maybe	Prove
Suggest	Definitely
Likely	

Average

The NCEA statistics glossary[1] defines the average:

> A term used in two different ways.
>
> When used generally, an average is a number that is representative or typical of the centre of a set of numerical values. In this sense, the number used could be the *mean* or the *median*. Sometimes the mode is used. This use of average has the same meaning as *measure of centre*.
>
> When used precisely, the average is the number obtained by adding all values in a set of numerical values and then dividing this total by the number of values. This use of average has the same meaning as *mean*.

As you can see, there are two different definitions of the word average. Therefore, in order to avoid confusion, **please do not use the word average**. Instead use the specific terminology of **mean, median or mode** as appropriate.

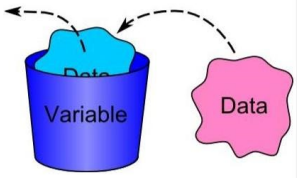
[1] https://seniorsecondary.tki.org.nz/Mathematics-and-statistics/Glossary/Glossary-page-A

Variables & Data

Let's think about the data we've collected and how we structure it to be able to decide what kinds of analysis and graphs we can do.

Variable

A variable is any characteristic, number, or quantity that can be measured or counted



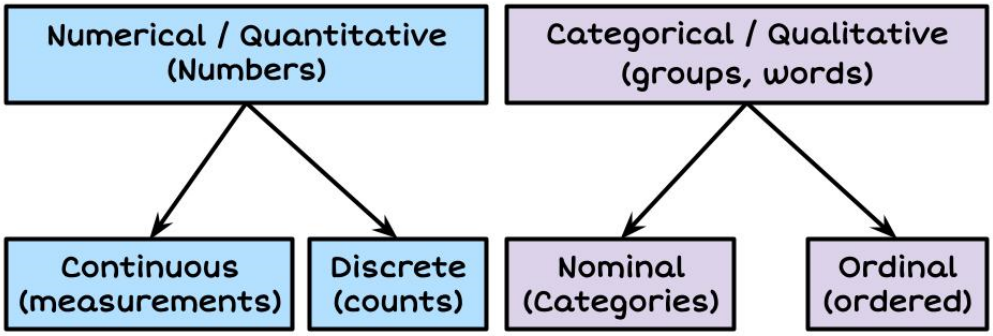
Data

A collection of facts, numbers, or information; the individual values of which are often the results of an experiment or observations.

Organising data

	Variable 1	Variable 2	Variable 3
Object 1			
Object 2			
Object 3			
...			

Data types



Numerical data

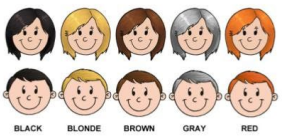
Numerical (numerical) data is data described by numbers.

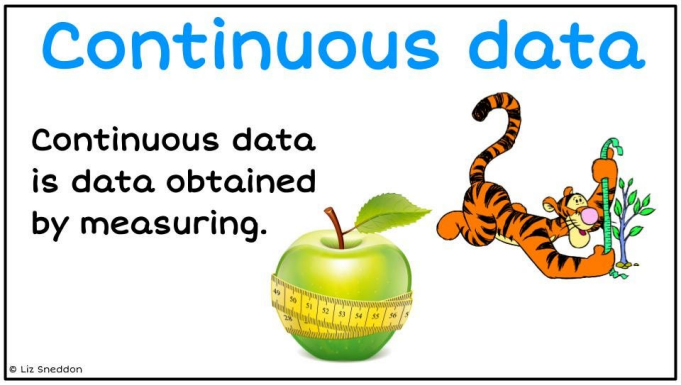
E.g., height, age, number of apples, weight.

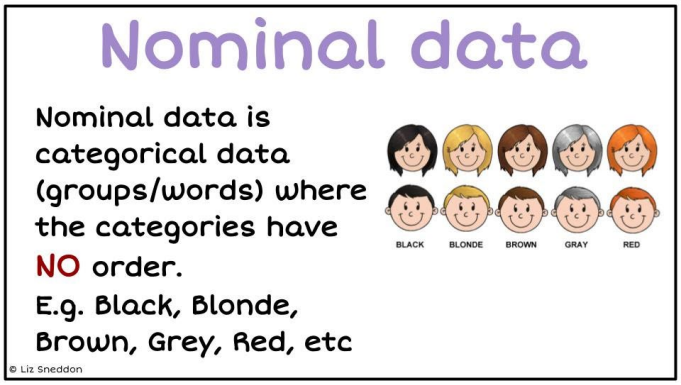
Categorical data

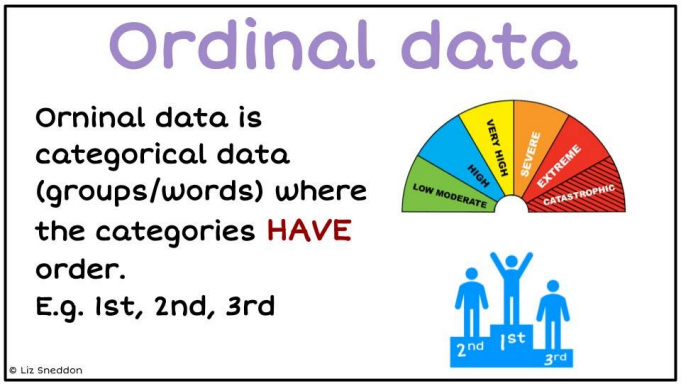
Categorical data is data that cannot be described by numbers. The data will be groups of words.

E.g., gender, ethnicity, apple variety.









Example:

Here is a spreadsheet:

	A	B	C	D	E	F
1	**Age**	**Gender**	**PE?**	**Wall sit time (seconds)**	**Leg up Wall sit time (seconds)**	**Height (cm)**
2	15	F	no	185	15	168
3	15	M	no	70	31	187
4	15	M	no	71	40	170
5	15	F	no	91.76	12.18	156.5
6	15	F	no	74	21	147

Each **row** is a set of **data** belonging to a student.

Each column is a **variable**.

Variable	Data Type
Age	Numeric (Discrete)
Gender	Categorical
Whether they take PE this year	Categorical
Wall sit time	Numeric (Continuous)
Leg up wall sit time	Numeric (Continuous)
Height	Numeric (Continuous)

Exercise 3:

For each of the datasets below, name the numeric and categorical variables.

1) The Bungee spreadsheet:

	A	B	C	D	E	F
1	**Elastic type**	**Length (cm)**	**Number of marbles**	**Stretched length (cm)**	**Weight (grams)**	**Did it hit obstacle(s)?**
8	Narrow	50	20	95	112	No
9	Narrow	50	18	79	112	No
10	Wide	27	20	30	107	No
11	Wide	39	20	60	119	No
12	Wide	45	19	70	106.4	Yes

Variable	Data Type
Elastic Type	
Length	
Number of marbles	
Stretched length	
Weight	
Obstacles	

2) The Marathon dataset:

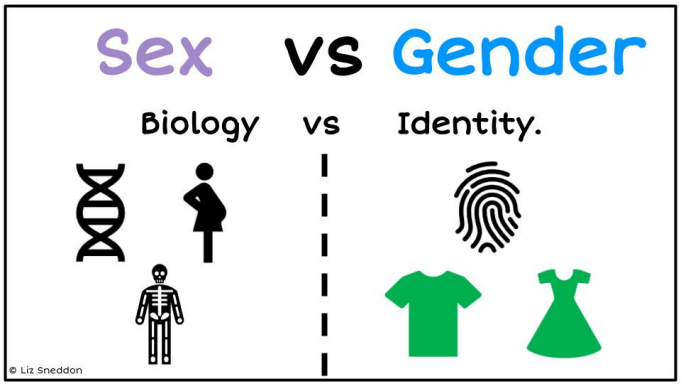
Variable	Description
Minutes	How many minutes they completed the marathon in
Gender	Male (M) or Female (F)
AgeGroup	Younger (under 40) or older (over 40)
Stridelength	The persons average stride length over the marathon in cm.

Variable	Data Type
Minutes	
Gender	
AgeGroup	
Stridelength	

Gender versus Sex for data collection

Statistics New Zealand (Tatauranga Aotearoa) is the government organisation that runs the Census (every 5 years). In the 2023 Census, they changed the questionnaire to include questions on both sex and gender, in order to collect data that represents the diversity of Aotearoa New Zealand, along with more accurate and detailed information across population groups.

Here are updated definitions that Statistics NZ has recently released: [2]



gender: Refers to a person's social and personal identity as male, female, or another gender such as non-binary. Gender may include the gender that a person internally feels ('gender identity'), and/or the gender a person publicly expresses ('gender expression') in their daily life. A person's current gender may differ from the sex recorded at their birth and may differ from what is indicated on their current legal documents. A person's gender may change over time. Some people may not identify with any gender.

sex at birth: Refers to the sex recorded at a person's birth (for example, recorded on their original birth certificate).

Question example:

Sex at birth	Gender
What was your sex at birth? (for example, what was recorded on your birth certificate) □ male □ female	**What is your gender?** □ male □ female □ another gender Please state: _____

For older datasets and reports, we need to remember that worldwide, there are limited and inconsistent practices when collecting sex and gender data in ways that reflect the diversity of the population.

This means that many of the existing datasets that we will use, have used the terms "gender" or "sex" interchangeably, and we need to take care when using these, as respondents are more likely to have recorded their gender, or may have recorded their sex at birth due to concerns about discrimination and society views.

[2] https://www.stats.govt.nz/consultations/sex-and-gender-identity-statistical-standards-consultation#summary

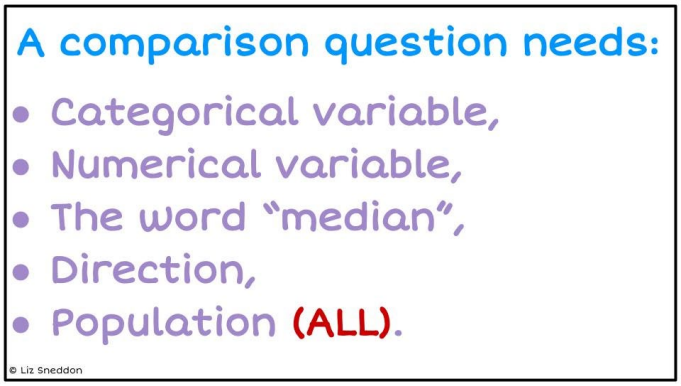
Exercise 4:

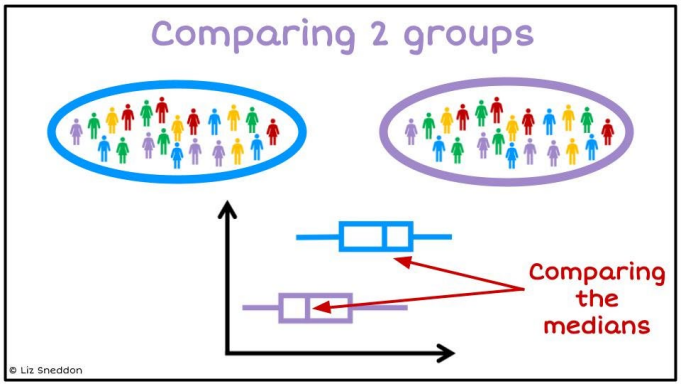
1) Explain why it is important for questionnaires, data analysis and reports to distinguish between gender and sex.

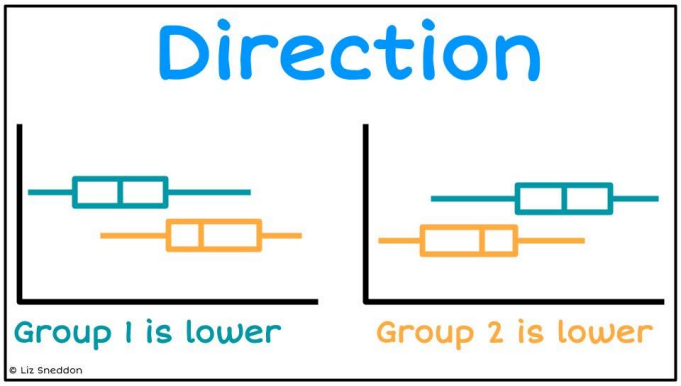
2) For existing datasets where data on male/female has been recorded, what are some of the issues in analysing and drawing conclusions that we need to be aware of?

Problem / Kaupapa

The first step in the investigation is to pose an investigation question. These are the things the investigation question needs to have.







Notes:

- You want to compare the centre of one group with the centre of another group. This means you are not expecting that **ALL** the items/people in one group will be bigger or smaller than **ALL** the items/people in the second group. However, you are wanting to compare if **MANY** of the items/people in one group are bigger or smaller than **MANY** of the items/people in the second group. This is why we are comparing the medians, as they are a measure of the centres of both groups.
- The population refers to **EVERYONE** in a group, this is why we suggest you use the word **ALL** when referring to the population. But be careful not to say **ALL people in one group** being bigger or smaller than **ALL people in the second group**, as we are comparing the central tendency.
- Be careful not to refer to "a" person/item, we are comparing groups of people/items.
- When you refer to the median, it is always the **median of the Numeric variable**, NOT the median of the categorical variable. E.g., **median height**, NOT median boy or median girl.

Example:

I wonder if the **median** travel time (minutes) for students who catch the bus is longer than the **median** travel time (minutes) for students who walk, for **ALL** high school students in NZ, for data from Census at School in 2015.

Exercise 5:

1) Identify from the example above:

Categorical variable	
Two groups being compared	
Numerical variable and units	
Direction	
Population definition	

2) The questions below all are missing one or more requirements. Identify each of the requirements, and which parts are missing. Then rewrite the question.

a) Are 8-year-old boys generally taller (cm) than 8-year-old girls in NZ?

Categorical variable	
Numerical variable and units	
Median	Yes / No
Direction	
Population definition	
Rewritten question:	

b) Do all 18-year-old males tend to have a longer right foot than all 18-year-old females in NZ?

Categorical variable	
Numerical variable and units	
Median	Yes / No
Direction	
Population definition	
Rewritten question:	

c) I wonder if there is a difference between the average school bag weight for girls and boys?

Categorical variable	
Numerical variable and units	
Median	Yes / No
Direction	
Population definition	
Rewritten question:	

d) How does the number of text messages a teenage girl sends daily compare with the number of text messages a teenage boy sends daily, in Auckland?

Categorical variable	
Numerical variable and units	
Median	Yes / No
Direction	
Population definition	
Rewritten question:	

3) **Sample:** the dataset is a random sample of 200 marathon runners in Aotearoa New Zealand.

Population: **ALL** marathon runners in Aotearoa New Zealand.

Variable	Description
Minutes	How many minutes they completed the marathon in
Gender	Male (M) or Female (F)
AgeGroup	Younger (under 40) or older (over 40)
StridelengthCM	The person's average stride length over the marathon in cm.

There are four possible comparison questions that you can write from the dataset above. Write all four questions below.

Investigation 1:

Categorical variable	
Numerical variable & units	
Investigation question:	

Investigation 2:

Categorical variable	
Numerical variable & units	
Investigation question:	

Investigation 3:

Categorical variable	
Numerical variable & units	
Investigation question:	

Investigation 4:

Categorical variable	
Numerical variable & units	
Investigation question:	

4) **Sample:** the dataset is a random sample of top rugby players.

Population: ALL top rugby players on the website: http://www.rugby-sidestep-central.com/.

Variable	Description
Country	New Zealand or South Africa
Position	Forward or Back
Weight	The weight of the player in kilograms (kg)
Height	The height of the player in meters (m)

There are four possible comparison questions that you can write from the dataset above. Write all four questions below.

Investigation 1:

Categorical variable	
Numerical variable & units	
Investigation question:	

Investigation 2:

Categorical variable	
Numerical variable & units	
Investigation question:	

Investigation 3:

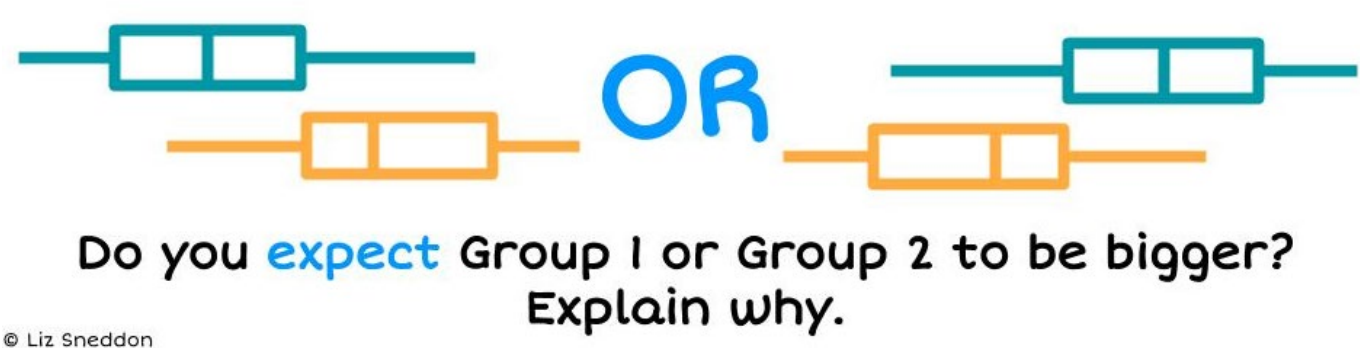
Categorical variable	
Numerical variable & units	
Investigation question:	

Investigation 4:

Categorical variable	
Numerical variable & units	
Investigation question:	

Hypothesis

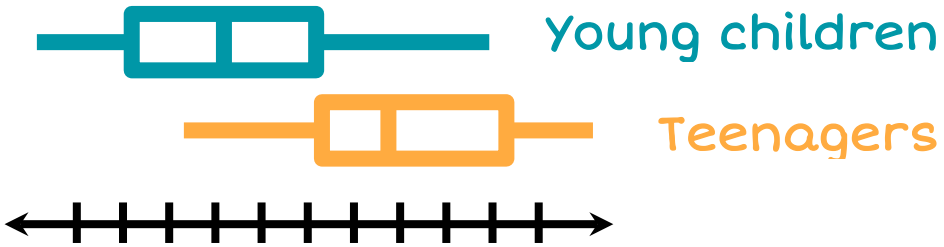
In the **Problem** section, you want to add a hypothesis about what you might see **BEFORE** you explore the data. You then want to use **research** to support this hypothesis and explain **WHY** there may be a difference between the two groups that you are comparing. Later in both your **Analysis** and **Conclusion** sections you can refer to this research and compare if the data agrees with your hypothesis or not.



Example:

Problem:
I wonder if the median weight of teenagers is heavier than the median weight of young children, for all children in NZ.

Hypothesis:
I expect that the median weight of teenagers would be heavier than young children, as they tend to be taller due to growth spurts.



Research from Kids Health[3] suggests that a major growth spurt usually happens for boys between 10 and 15 years of age, and for girls between 8 and 13 years old. As a child has a growth spurt, they grow taller. As they grow taller the weight of their bones, muscles and organs gets heavier. Therefore, suggesting that teenagers would be heavier than young children.

Now I will investigate the dataset to see if the data supports this hypothesis.

(Notice that I added a footnote with the website that I got information from).

[3] https://kidshealth.org/en/parents/childs-growth.html

Research (for Excellence)

For Excellence, you need to include research to support your hypothesis.

Research

Search for information explaining **WHY** there would be this pattern

© Liz Sneddon

Why reference?

To credit the author and avoid plagiarism

To allow readers to find the original source of your information.

Allows readers to assess the validity and reliability of the evidence

To show your argument is supported by evidence

© Liz Sneddon

Plagiarism

the practice of taking someone else's work or ideas and passing them off as one's own.[1]

1. Oxford dictionary

© Liz Sneddon

Avoiding plagiarism

Paraphrase/Reword
(Writing the ideas in your own words, acknowledging where they came from.)

Record & reference all sources.

Use quotation marks

Summarise

© Liz Sneddon

Citing sources

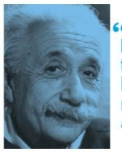
If the ideas are **general knowledge**, or ideas **you** came up with on your own, you **don't** need to cite it.

For everything else, you must cite it.

© Liz Sneddon

Quotes

If you want to **copy exactly** what someone else has written, put **quotation marks** around it, and **tell us who wrote it.**



"Education is not the learning of facts, but the training of the mind to think."
ALBERT EINSTEIN

© Liz Sneddon

Summarise

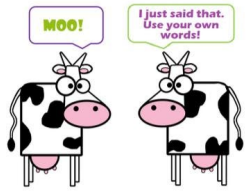
Use your **own words**, focus on the **main ideas only**, and write a **short** sentence or paragraph.

© Liz Sneddon

Paraphrase

Rewriting the information expressed by someone else, **without changing the meaning.**

Source text	Paraphrase
The need for investors to earn a commercial return may put upward pressure on prices	The need for profit is likely to *push up* prices

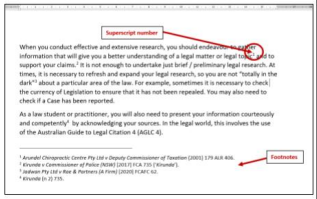


© Liz Sneddon

List of sources

Include a list of your sources.

This could be just a reference list at the end, and/or footnotes.



© Liz Sneddon

Credible sources

Purpose - what is the purpose of the article? To inform, teach, sell, entertain or persuade? Is it factual & unbiased or an opinion?

Author - are they qualified to write on this topic?

Timeliness - is the information recent?

Accuracy - is the evidence trustworthy, reliable and verified?

Relevant - is the information important to explaining WHY?

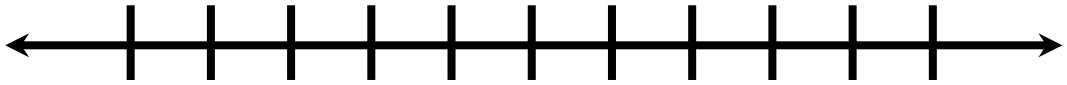
© Liz Sneddon

Exercise 6:

Make a hypothesis and support with research for the situations below.

1) **Problem:**
I wonder if the median time that teenagers spend on phone calls each day is more than the median time that adults (18 years and older) spend on their own phone calls each day, for all teenagers in NZ.

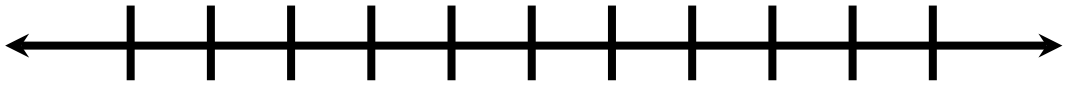
Hypothesis:
Draw the graph you expect to see.



Use research to explain why you expect to find this.

2) **Problem:**
I wonder if the median lung capacity of teenagers who vape is smaller than the median lung capacity of teenagers who don't vape, for ALL teenagers in NZ.

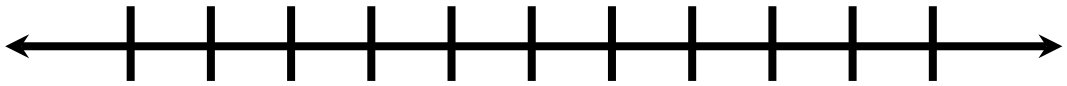
Hypothesis:
Draw the graph you expect to see.



Use research to explain why you expect to find this.

3) **Problem:**
I wonder if the median house price for a 2-bedroom home tends to sell for higher prices than the median house price of 4-bedroom homes, for ALL homes in NZ.

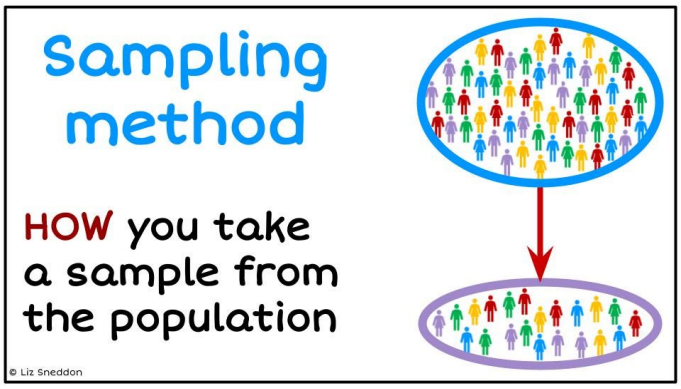
Hypothesis:
Draw the graph you expect to see.

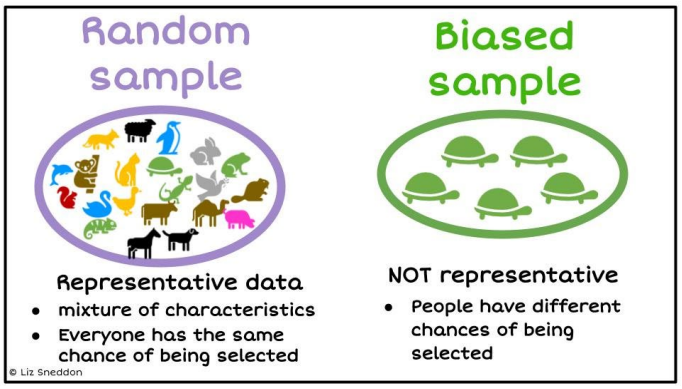


Use research to explain why you expect to find this.

Plan / Whakamahere

One of the most important components of being able to form a conclusion is that the people (or objects) that the data is collected from are **randomly selected**, so that the **sample data is representative of the population**.



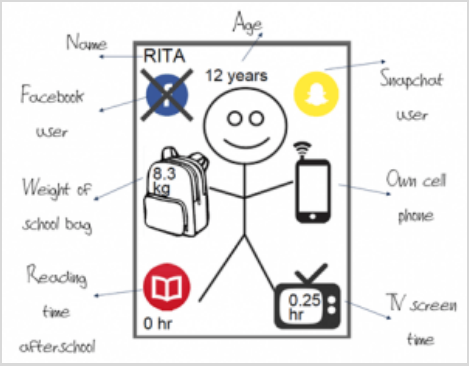


Example

If I do a questionnaire with **only blue-eyed** students, then I have a **biased sample**. This means I do not have any information about people with other coloured eyes (E.g., brown, green, grey, etc.), so my data does not represent the population of all people, only the people with blue eyes.

Exercise 7:

1) Collect a bag of data[4] from your teacher. Take a random sample of 10 students. Record your data below.

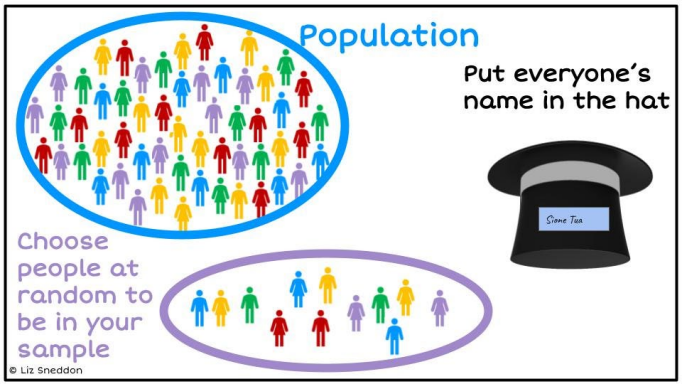
 **Name**	**Age (years)**	**Do you have Facebook? Yes / No**	**Do you have Snapchat? Yes / No**	**School bag weight (kg)**	**Do you have a Cell phone? Yes / No**	**Reading time yesterday (hours)**	**TV time yesterday (hours)**

[4] Created by Anna Ferguson: http://teaching.statistics-is-awesome.org/census-at-school-stick-people-data-cards/

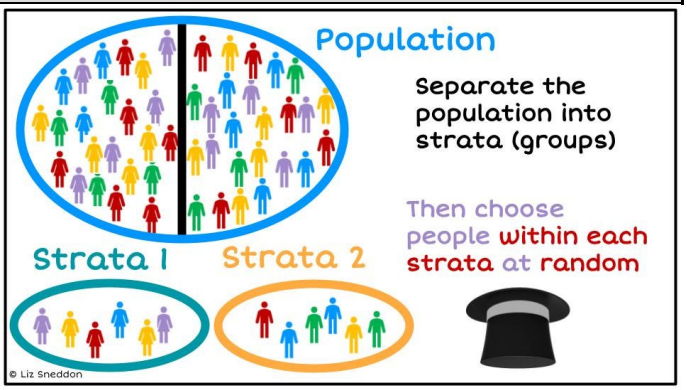
Sampling methods

These are the different random sampling methods:

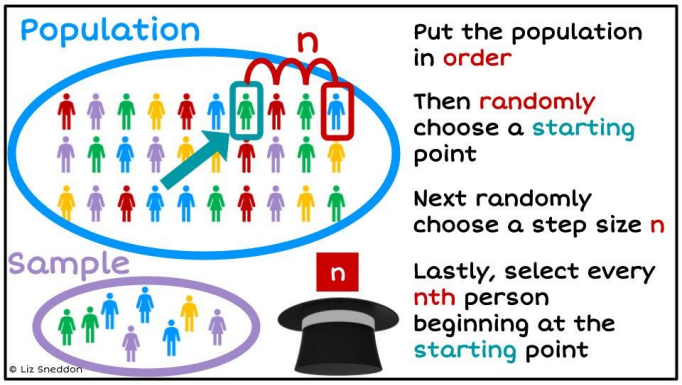
Simple random



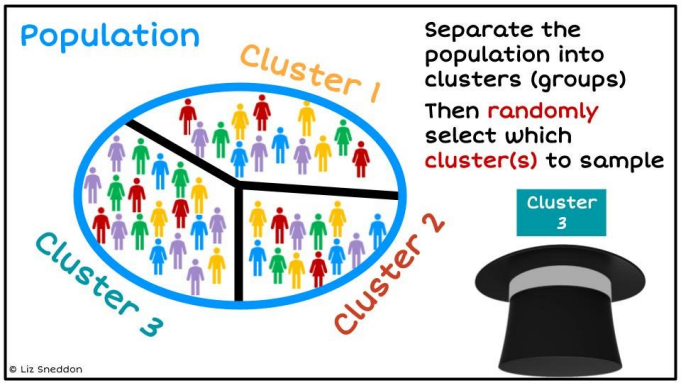
Stratified



Systematic

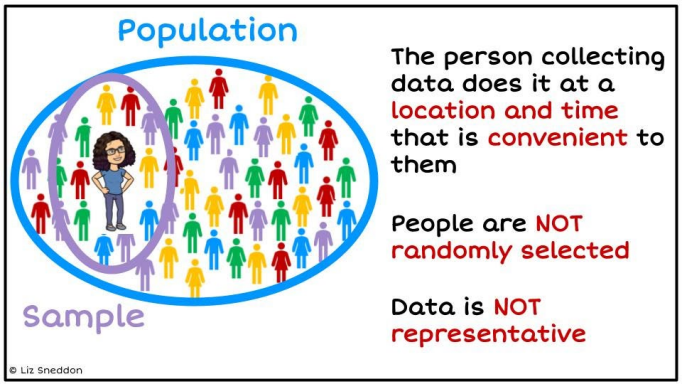


Cluster

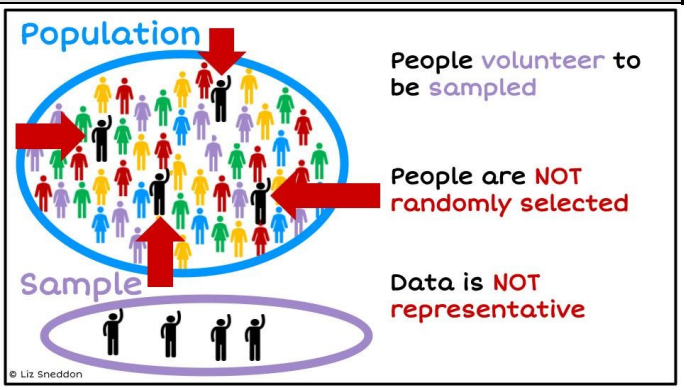


These are some biased sampling methods:

Convenience



Self-selected



Exercise 8:

1) Below are a number of descriptions of sampling methods. Decide which sampling method has been described and write the name below the description.

	Situation	Sampling method	Random or biased
a)	Mrs Sneddon prints out her Year 12 class list. She randomly starts at the 10th person on the list, and then chooses every 4th person after that.		
b)	At assembly, the principal asks students who want to take part in the blood drive later this year to stand up.		
c)	Mrs Sneddon needs help setting up classrooms for an assessment and asks five students seated nearby to help.		
d)	Mrs Sneddon separates her class into boys and girls, and randomly choose 3 girls and 3 boys to participate in an experiment.		
e)	Mrs Sneddon was doing a survey of students' opinions on the canteen, randomly chose Elliot quad and she asked all the students in the quad their opinion.		
f)	Mrs Sneddon randomly chooses a student from her class to collect messages from the office.		

2) Your teacher gives you a list of students in your class. Describe how you could take a simple random sample.

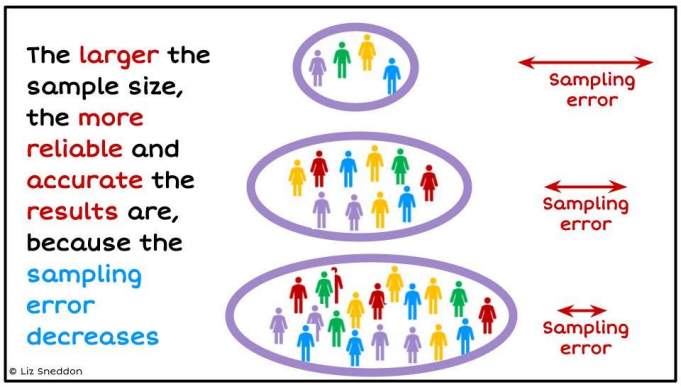
3) Describe how you could take a stratified random sample from your Year 12 class, separating by gender.

Sample size

We want to take a big enough **sample size**, so that the results are **reliable and precise** enough to represent the population.

The more data we have, the greater the precision of our results, and the lower the variation.

With a small sample size, it is much harder to find differences. With a larger sample size, you can find differences more easily.



Exercise 9:

1) Circle the words that complete the sentences below.

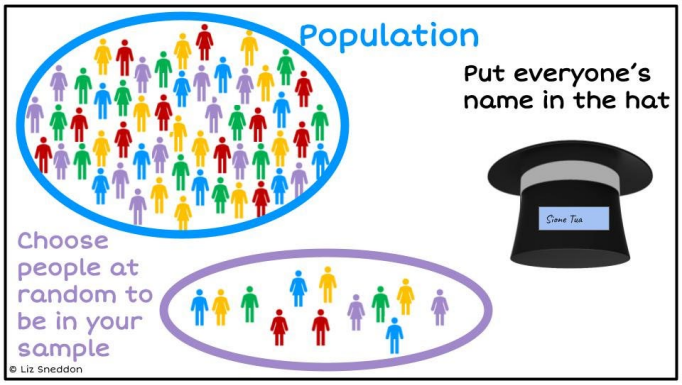
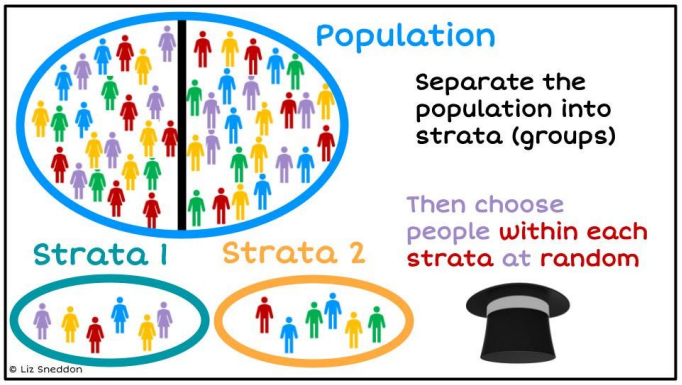
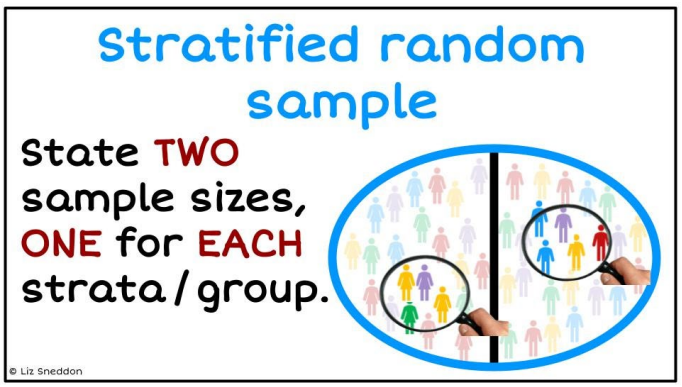
a) Smaller sample sizes take a **shorter / longer** time to collect data, but results are **more / less** precise.

b) Larger sample sizes take a **shorter / longer** time to collect data and results are **more / less** precise.

2) Mrs Sneddon is going to survey 35 girls and 40 boys to investigate their use of iPads at home. Explain why it is ok for the sample sizes to be different, and how these sample sizes affect the variation.

Putting random sampling altogether

Depending on which sampling method is used, will determine how we state the sample size. We will focus on two sampling methods, both random sampling methods.

Simple random	Stratified
	
	

Example 1:

When investigating the difference in travel time for high school students between catching the bus and walking, I will take a **simple random sample** of **100 students**.

Example 2:

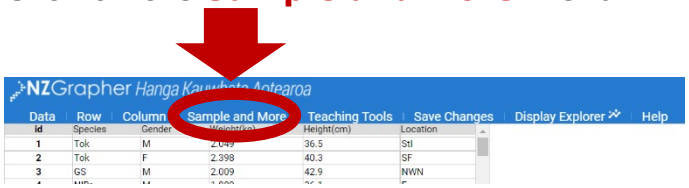
When investigating the difference in travel time for high school students between catching the bus and walking, I will take a **stratified random sample** of **100 students who catch the bus**, and **100 students who walk to school**. This gives a **total of 200 students** in my sample.

Taking a sample with NZGrapher

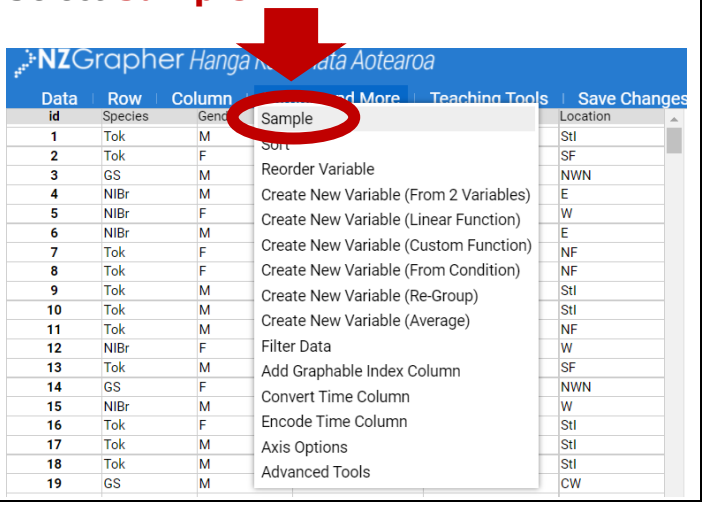
NZGrapher has two sampling methods that you can use: Simple random sampling, and Stratified sampling.

Simple random sample:

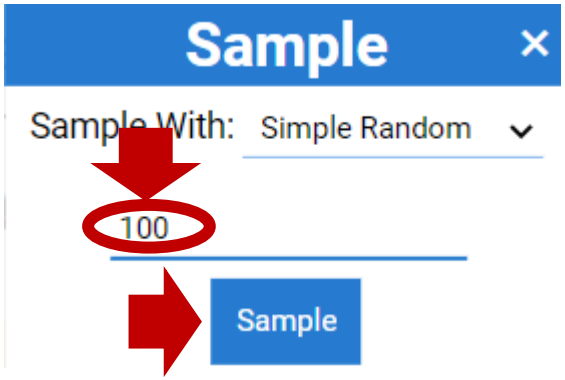
Step 1:
Click on the **Sample and More** menu.



Step 2:
Select **Sample**



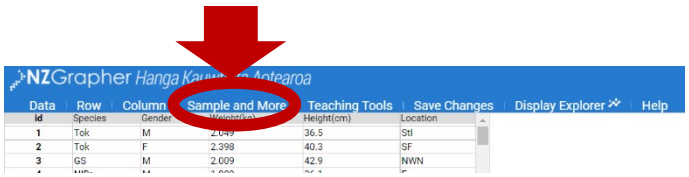
Step 3:
The default option is to use Simple Random sampling. Enter the sample size you want and press the **Sample** button.



Stratified random sample:

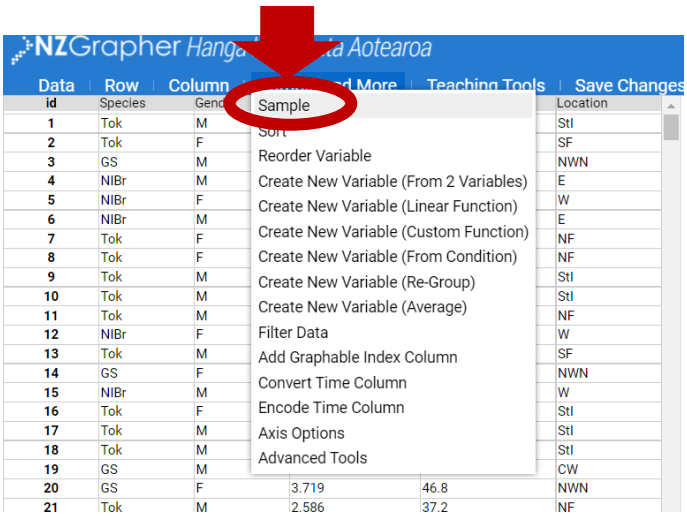
Step 1:

Click on the **Sample and More** menu.



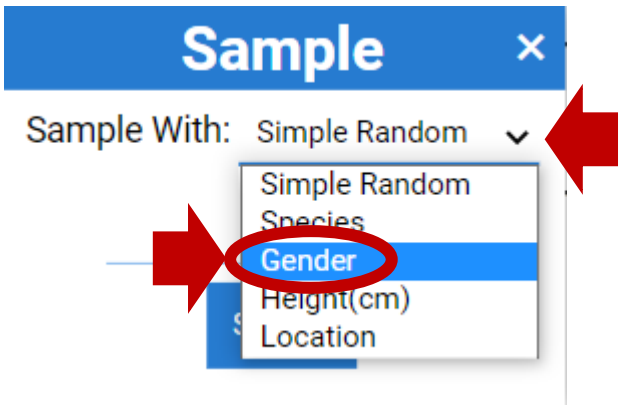
Step 2:

Select **Sample**



Step 3:

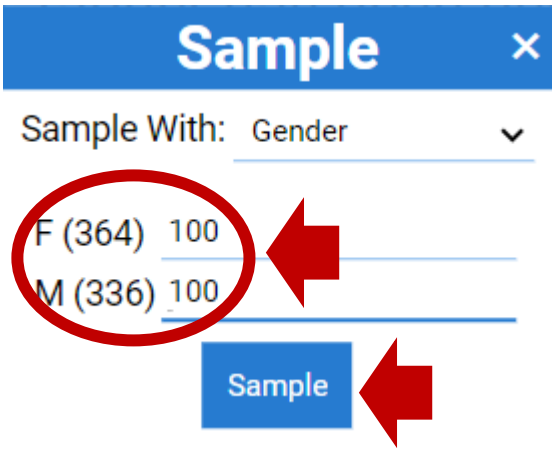
Choose the variable you want to stratify with (this will be the **categorical variable** you chose for your investigation question).



Step 4:

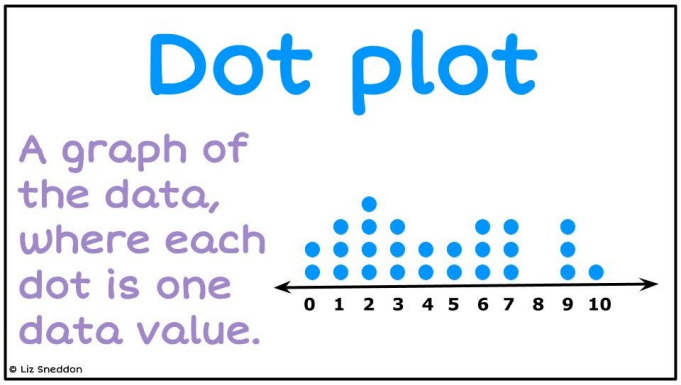
You will then see the groups for the categorical variable (E.g., Male and Female). Beside each group name is the sample size for that group. (E.g., F(364)).

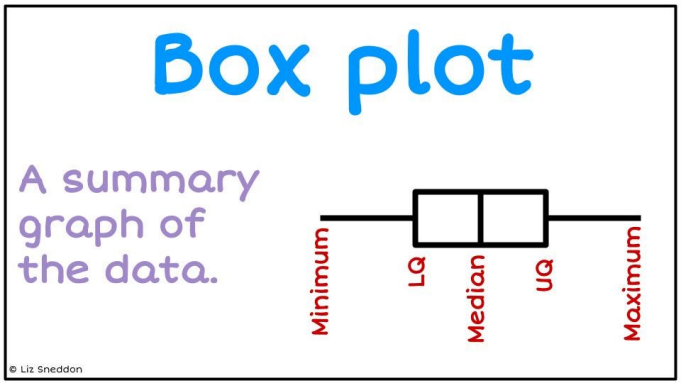
Enter a sample size for **BOTH** groups, then press the **Sample** button.

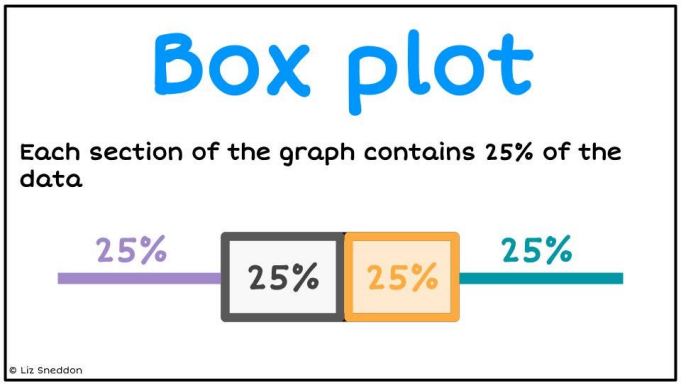


Data / Raraunga

Once you have selected the data using a random sampling method, the next step is to process the data, drawing graphs and getting summary statistics.



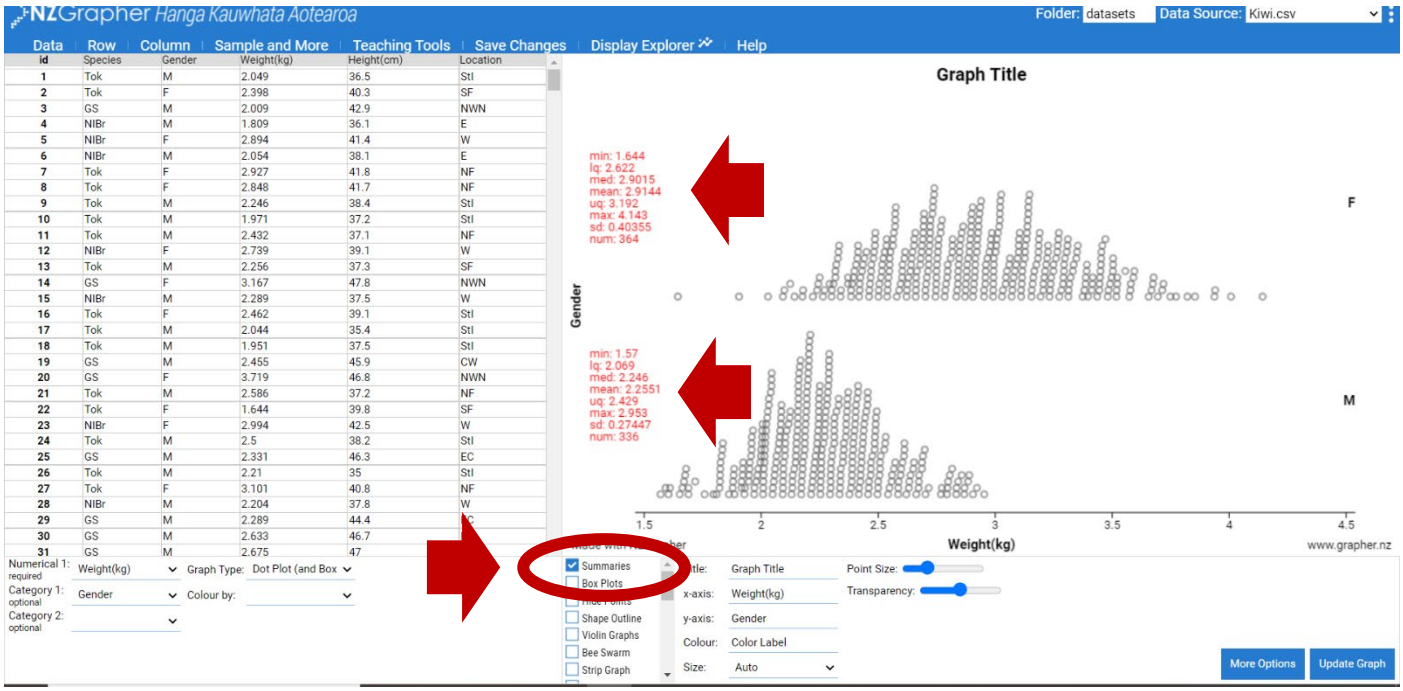




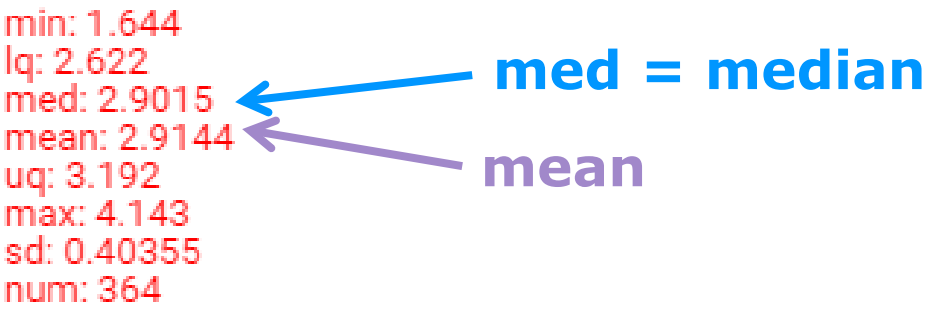
Graphs with NZGrapher

Along with adding the summary statistics, there are several additional options that you need to select when making a Dot plot and Boxplot for Year 12.

1) Add summary statistics to your graph.

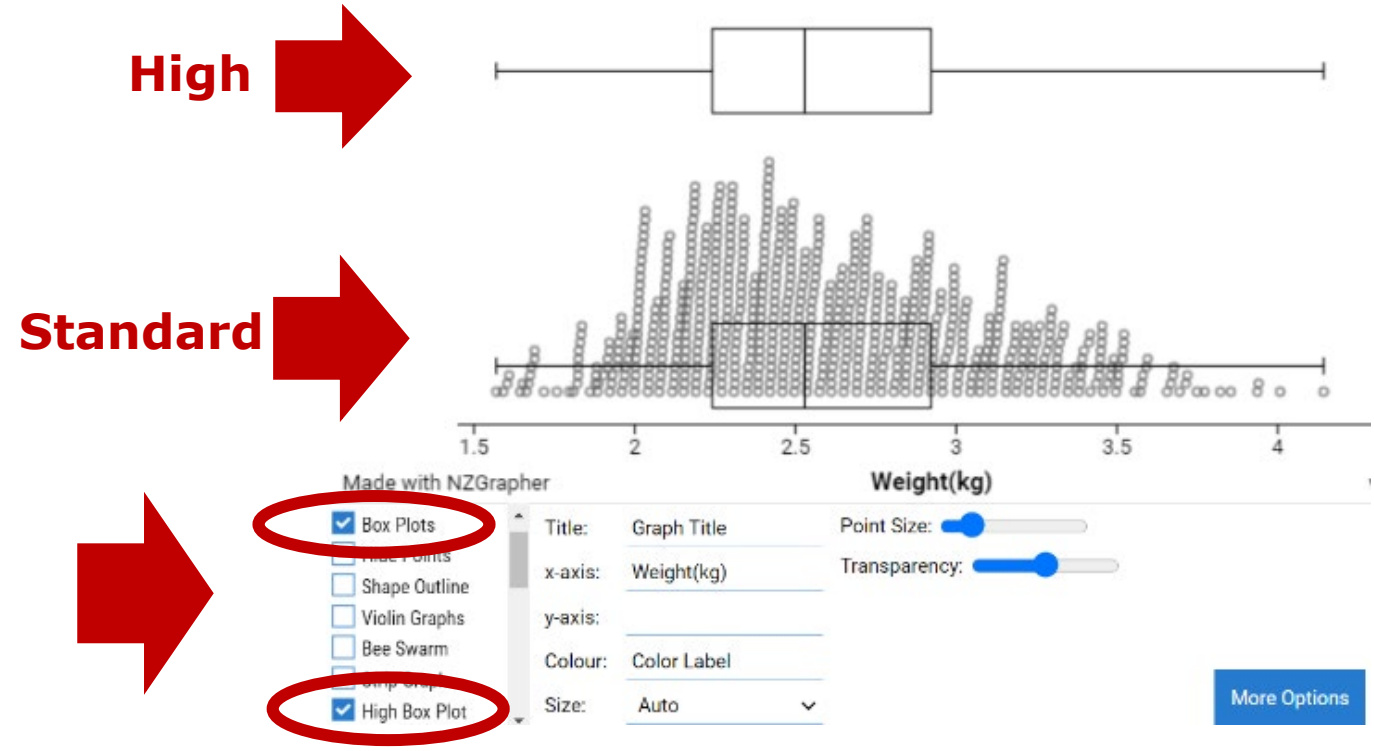


Be careful that you don't mix up the **median** and the **mean** values.

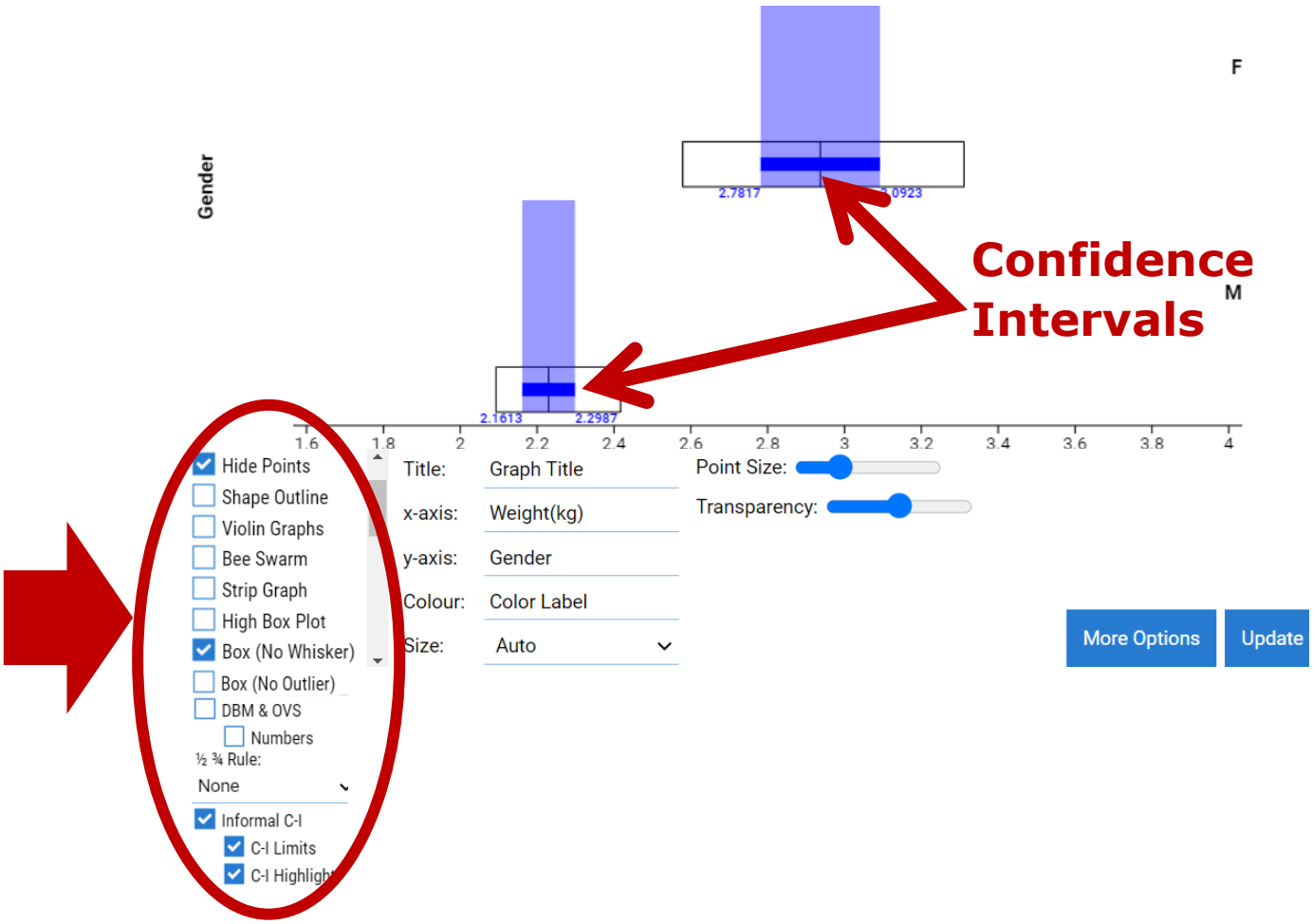


2) Add titles and axis labels to your graph. Press the Update Graph button when you have finished.

3) There are several options for box plots, one that sits **on top of** the data, and one that sits **above** the data (tick the High Box plot option).



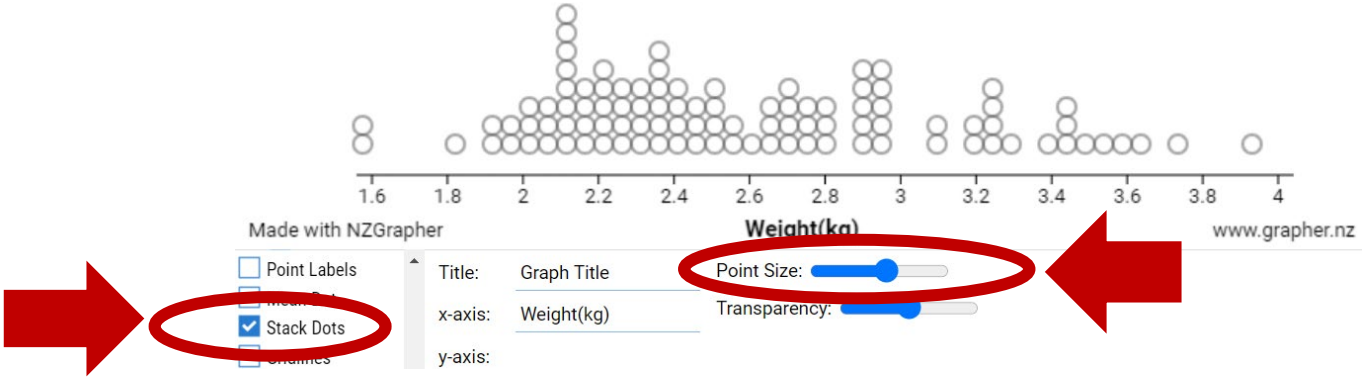
4) You need to add the Confidence Interval, limits and highlights to your graph. In order to see these limits more clearly, there are several options I select: Hide points, Box (No Whisker), Informal C-I, C-I limits, C-I Highlight.



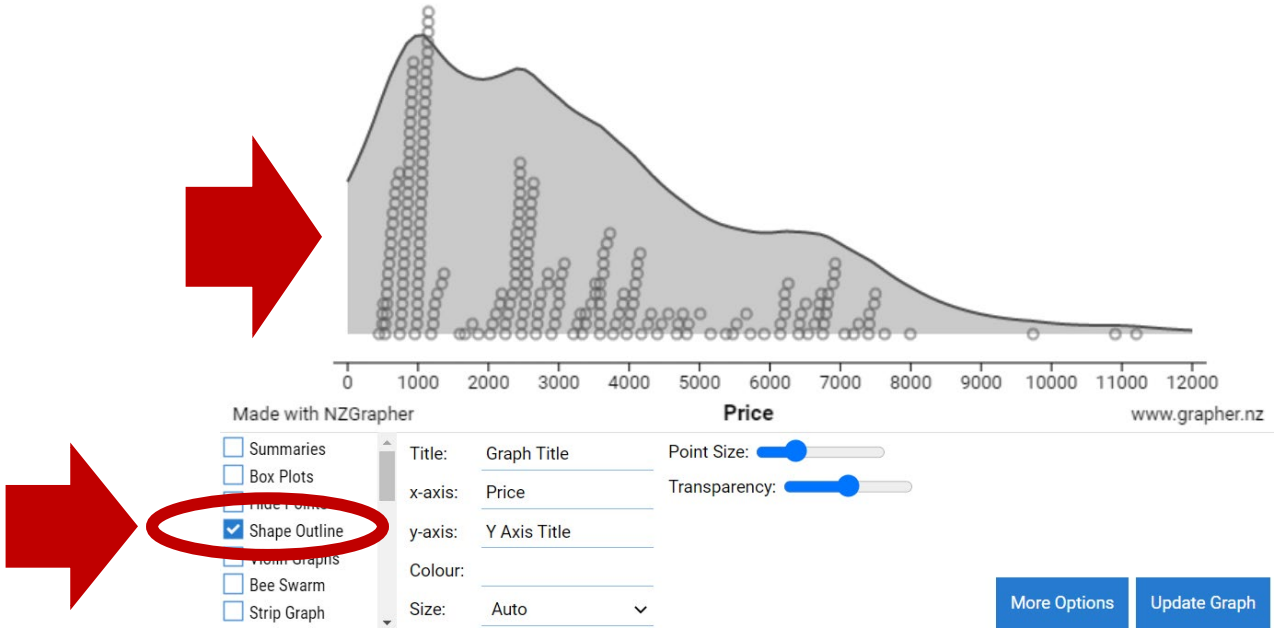
NZGrapher Tips

Here are some additional tips which I find useful.

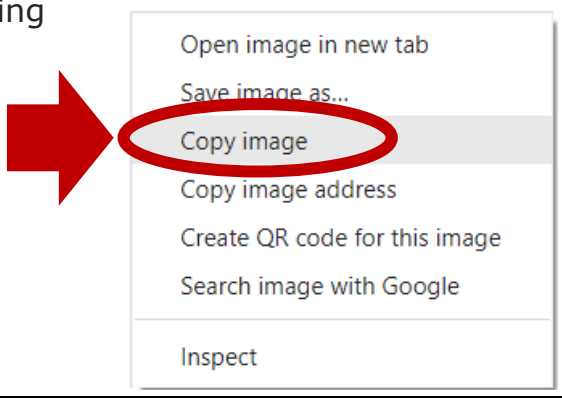
Tip 1: You can tidy up the way the dots look, by selecting **Stack Dots**, and you can also change the **point size** to better see the shape of the data.



Tip 2: There are some new graphs that you can change / add. I like the Shape outline as the overall shape can be seen more easily.

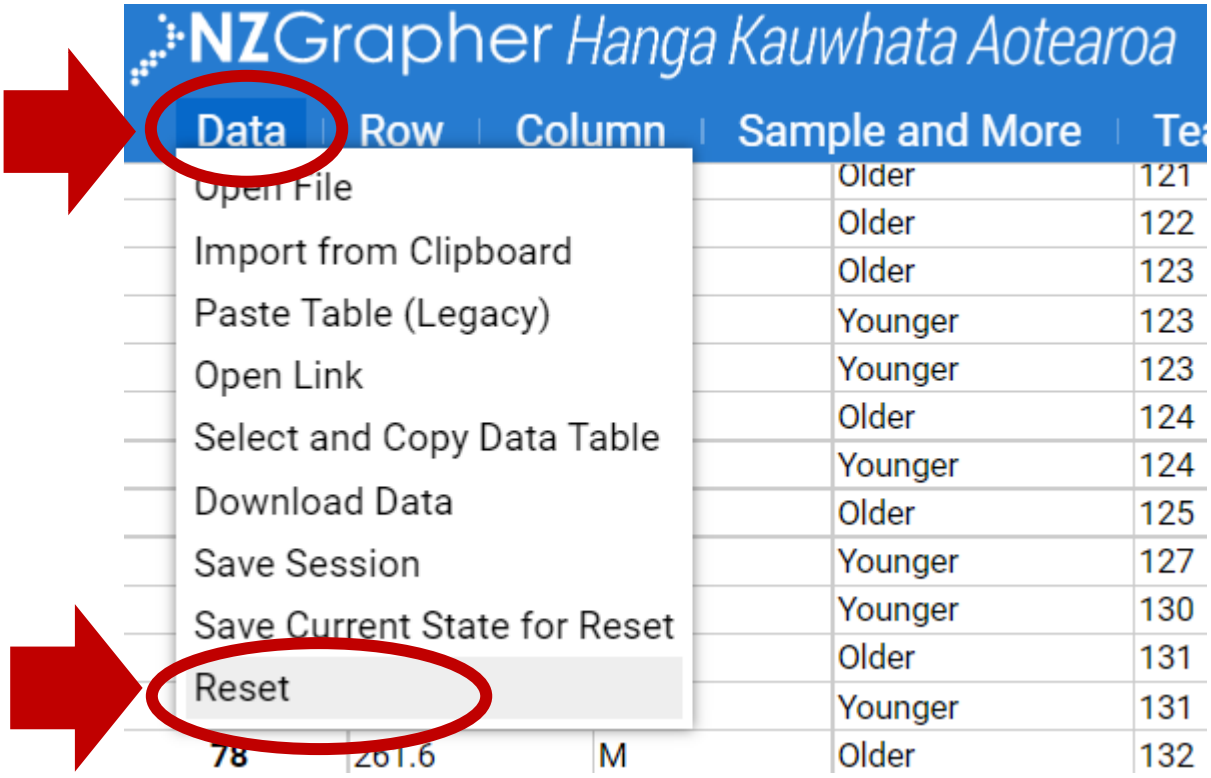


Tip 3: You can copy each graph by right clicking over the image, and then selecting **Copy Image**.



Tip 4: You can select the **Box plot (no Outlier)** which will draw the box plot **EXCLUDING** the outliers in the tail. This allows you to easily identify the outliers.

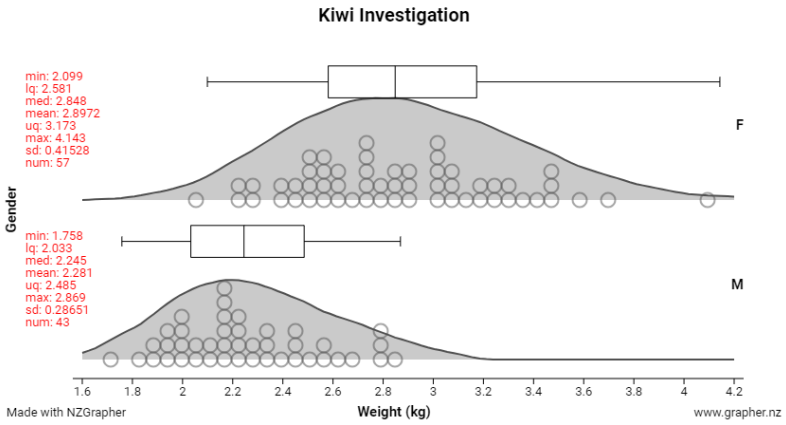
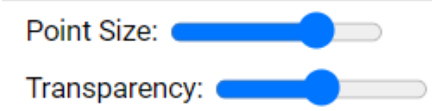
Tip 5: If **you** make a mistake when you take a random sample, then go to the Data menu and select the **Reset** button. This will reload the full dataset and you can start again.



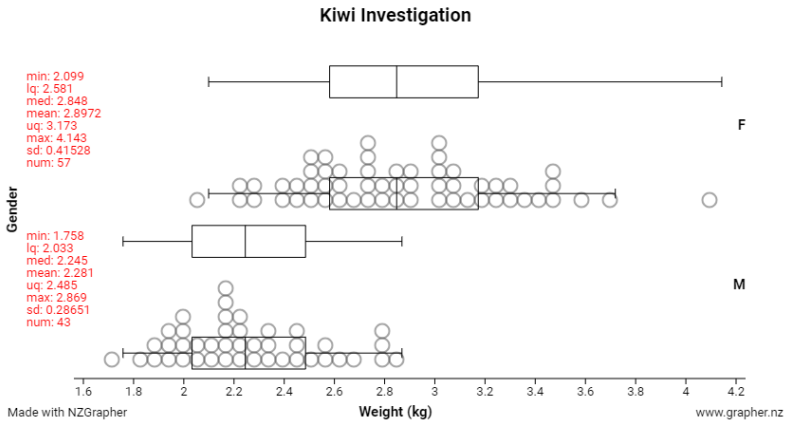
Putting NZGrapher together

We need to take a random sample, then using these selections and tips, these are the three graphs I usually produce to help me analyse and draw conclusions using the data.

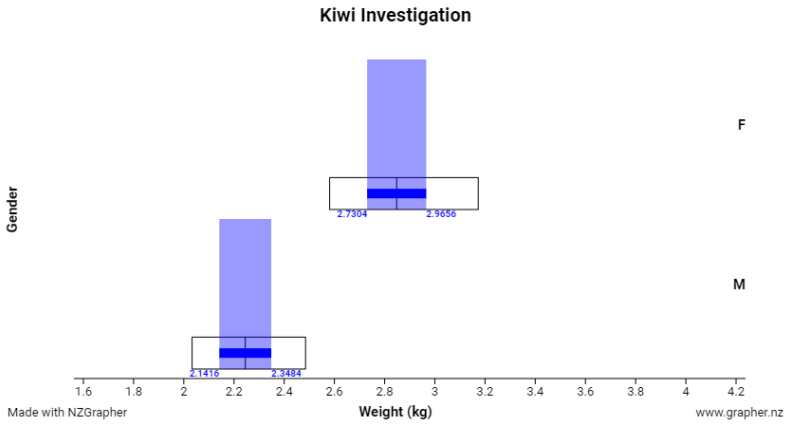
Graph 1: Analysis of features

Select the following options:	Graph
☑ Summaries ☐ Box Plots ☑ High Box Plot ☐ Hide Points ☑ Shape Outline ☐ Violin Graphs ☐ Bee Swarm ☐ Strip Graph ☐ Box (No Whisker) ☐ Box (No Outlier) ☐ DBM & OVS ☐ Numbers ½ ¾ Rule: None ☐ Informal C-I ☐ C-I Limits ☐ C-I Highlight ☐ Point Labels ☐ Mean Dot ☑ Stack Dots ☐ Gridlines	I took a simple random sample of 100 for the Kiwi dataset before making this graph.  I also changed the point size: 

Graph 2: Outliers

Select the following options:	Graph
☑ Summaries ☐ Box Plots ☑ High Box Plot ☐ Hide Points ☐ Shape Outline ☐ Violin Graphs ☐ Bee Swarm ☐ Strip Graph ☐ Box (No Whisker) ☑ Box (No Outlier) ☐ DBM & OVS ☐ Numbers ½ ¾ Rule: None ☐ Informal C-I ☐ C-I Limits ☐ C-I Highlight ☐ Point Labels ☐ Mean Dot ☑ Stack Dots ☐ Gridlines	

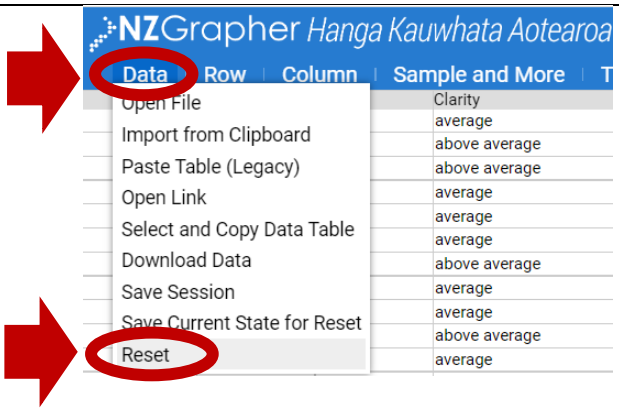
Graph 3: Conclusion

Select the following options:	Graph
☐ Summaries ☐ Box Plots ☐ High Box Plot ☑ Hide Points ☐ Shape Outline ☐ Violin Graphs ☐ Bee Swarm ☐ Strip Graph ☑ Box (No Whisker) ☐ Box (No Outlier) ☐ DBM & OVS ☐ Numbers ½ ¾ Rule: None ☑ Informal C-I ☑ C-I Limits ☑ C-I Highlight ☐ Point Labels ☐ Mean Dot ☑ Stack Dots ☐ Gridlines	

Exercise 10:

1) Go to NZGrapher and select the **Marathon dataset**. You need to generate dot plots for the following variables, using the sampling methods and sample sizes stated below.

Create both a dot plot with summary statistics, a box plot and the shape outline, along with a second graph showing the boxes and confidence intervals for each scenario.

You will need to reset the dataset after each question below so that you can do the next question. To do this, go to the **Data** menu on the top left of the page, and select **Reset**.	

a)

Variables	Minutes, Gender
Sampling method	Simple random sample
Sample size	100

b)

Variables	Stridelength, Agegroup
Sampling method	Stratified random sample
Sample size	50 in each group

c)

Variables	Minutes, Agegroup
Sampling method	Stratified random sample
Sample size	80 older people, 90 younger people

d)

Variables	Stridelength, Gender
Sampling method	Simple random sample
Sample size	150

2) The dataset described below is a sample of rugby data taken from http://www.rugby-sidestep-central.com/

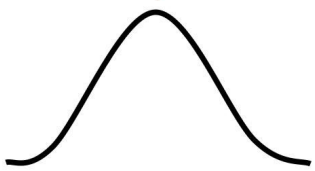
Variable	Description
Country	New Zealand or South Africa
Position	Forward or Back
Weight	The weight of the player in kilograms (kg)
Height	The height of the player in meters (m)

a) Choose 2 variables from the dataset above and describe how you would take a **stratified random sample** using NZGrapher, including stating the sample size(s).

b) Go to NZGrapher and carry out the stratified random sample you described. Create appropriate graphs.

Measures of centre

Centre



- Mean
- Median
- Mode

© Liz Sneddon

$$\text{Mean} = \frac{\text{Sum of data}}{\text{Sample size}}$$

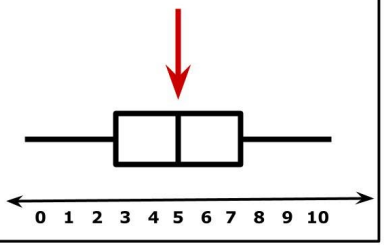
Median = middle value

Mode = most frequent value

© Liz Sneddon

Median

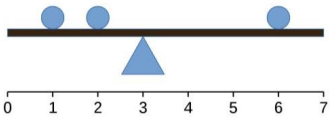
A value where **50%** of the data lies **above** & **below** it.



© Liz Sneddon

Mean

The **average** of a set of values.

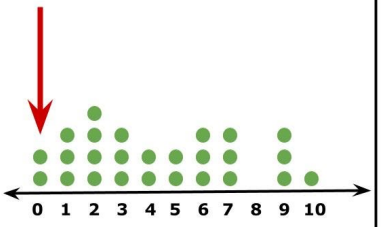


© Liz Sneddon

Measures of spread

Minimum

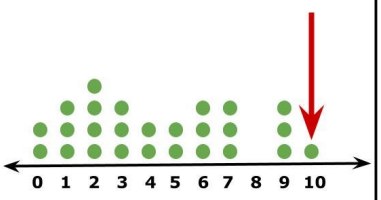
The **smallest** data value.



© Liz Sneddon

Maximum

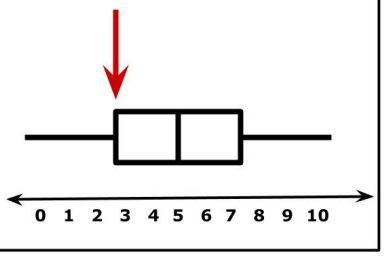
The **largest** data value.



© Liz Sneddon

Lower Quartile (LQ)

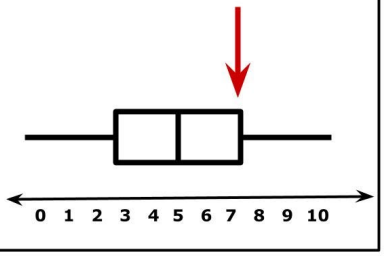
A value where **25%** of the data lies **below** it.



© Liz Sneddon

Upper Quartile (UQ)

A value where **25%** of the data lies **above** it.

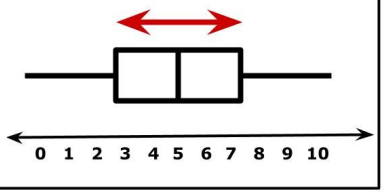


© Liz Sneddon

Interquartile range

IQR = UQ - LQ

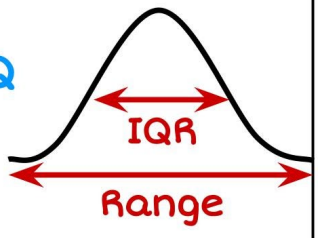
The spread of the **middle 50%** of the data.



© Liz Sneddon

Range = Max - Min

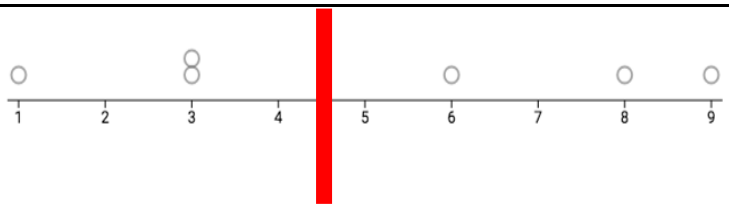
IQR = UQ - LQ



© Liz Sneddon

Example:

Estimate the center, and find the mean, median and mode. Then calculate the IQR.



Data	9, 3, 1, 8, 3, 6
	Put the numbers in order: 1, 3, 3, 6, 8, 9
Minimum	1
Maximum	9
Median	Find the number(s) in the middle: 1, 3, 3, 6, 8, 9 Find the median = $\frac{3+6}{2}$ = 4.5
Mean	$\frac{9+3+1+8+3+6}{6}$ = 5
Mode	3
LQ	Take the numbers **below** the median: 1, 3, 3 Find the middle of these numbers LQ = 3
UQ	Take the numbers **above** the median: 6, 8, 9 Find the middle of these numbers UQ = 8
IQR	UQ – LQ = 8 – 3 = 5

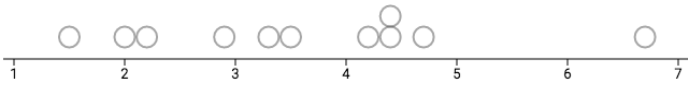
Exercise 11:

Calculate the summary statistics from the graph, including units.

1)

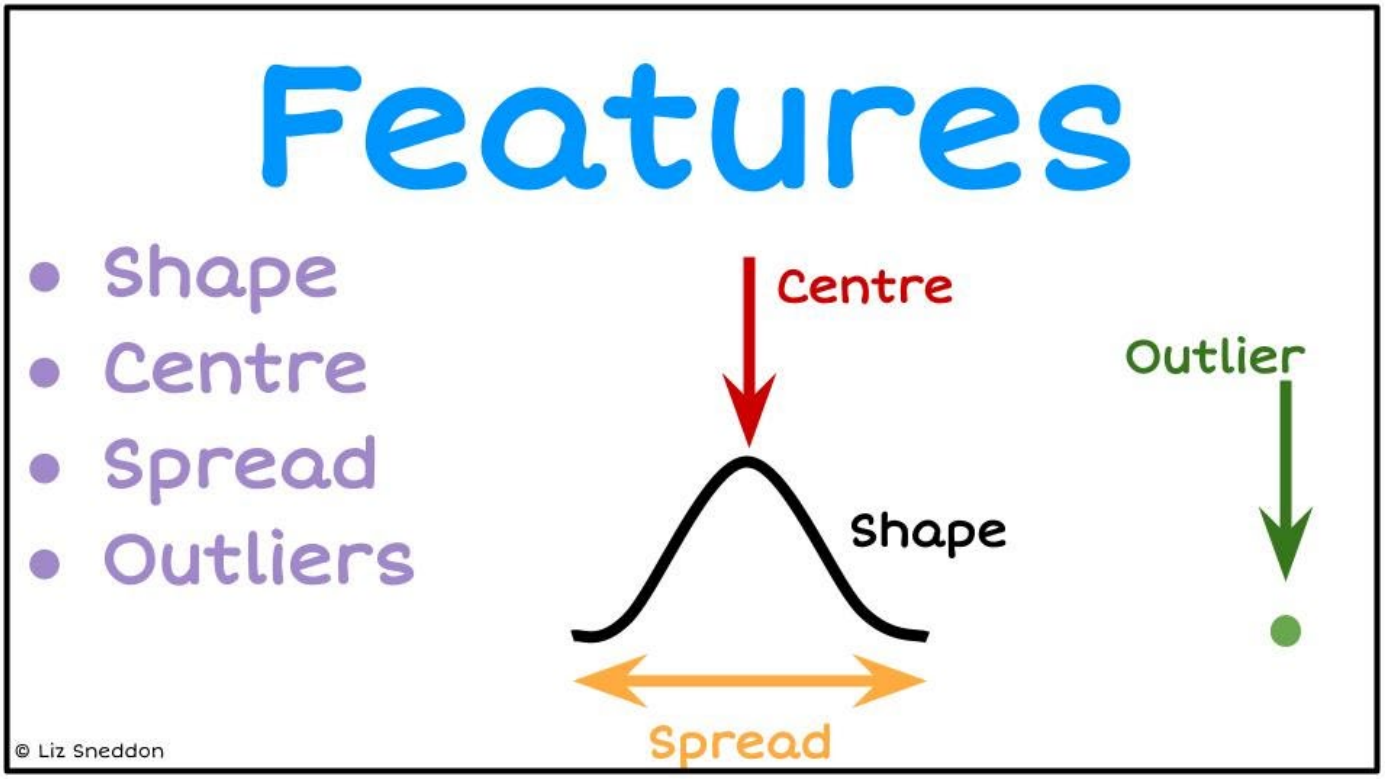
Data	4, 6, 3, 8, 2, 4, 9
Ordered data	
Minimum	
Maximum	
Median	
Mean	
LQ	
UQ	
IQR	

2)



Data	4.4 4.7 3.5 2.2 4.2 6.7 2.9 4.4 1.5 2.0 3.3
Ordered data	
Minimum	
Maximum	
Median	
Mean	
LQ	
UQ	
IQR	

Analysis / Tātari

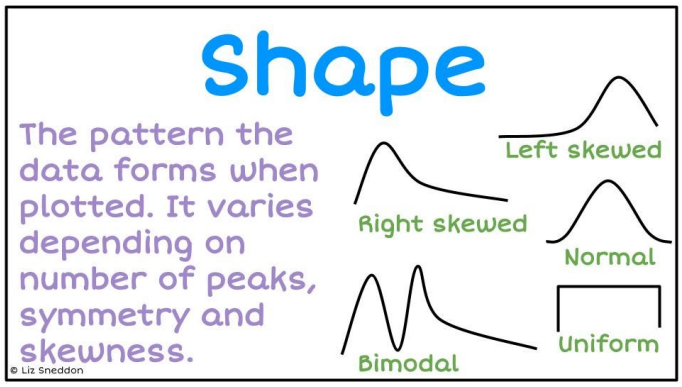


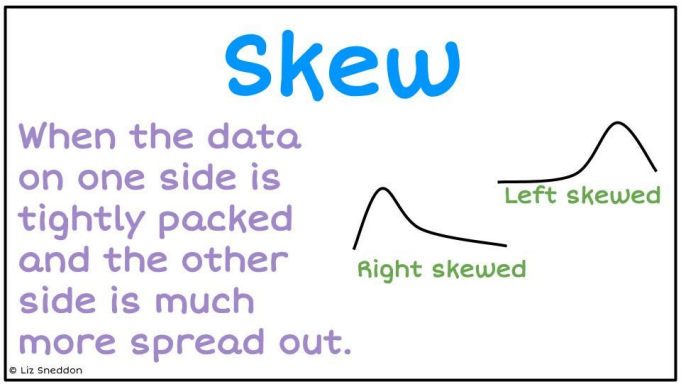
Remember to use **suggestive language** when you analyse the features and patterns in your **sample data**.

We will go through each of these features separately, before putting it altogether.

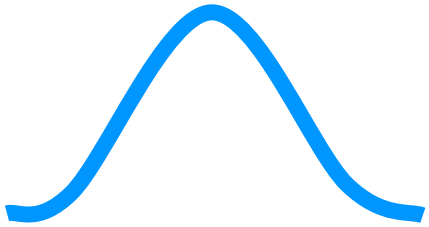
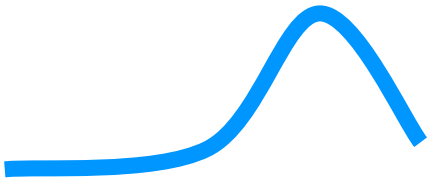
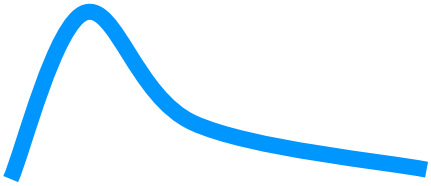
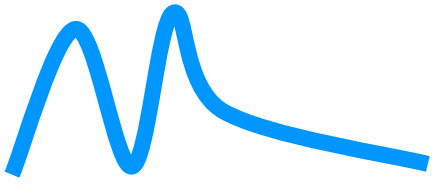
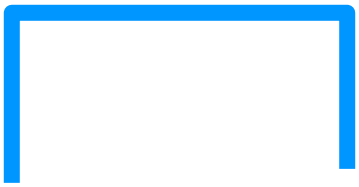
Shape

When we analyse the shape, we are taking a step back and thinking about what shape the distribution is that this sample has come from. It's like standing at the back of the classroom and seeing the general pattern of the data on the board.



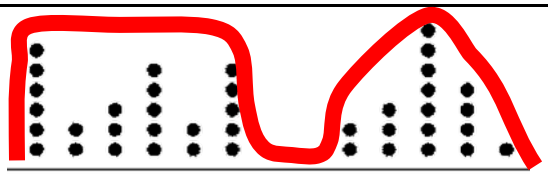


Identifying shapes:

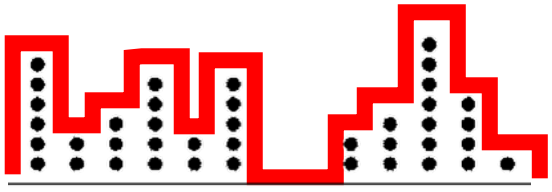
		
Normal distribution	**Left skewed**	**Right Skewed**
		
Bimodal	**Uniform**	

Example:

Draw shapes like this:



Not like this:

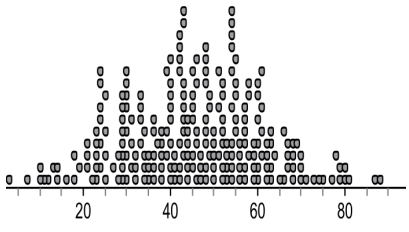


The data has 2 peaks, so looks approximately bimodal in shape.

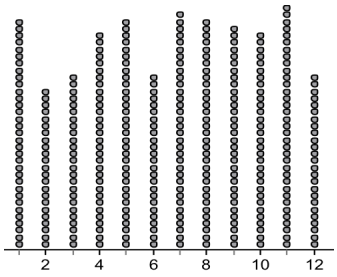
Exercise 12:[5]

Sketch over the top of each graph and then state what shape it most closely matches.

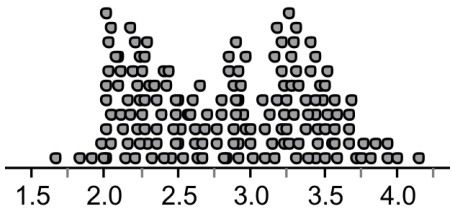
1.



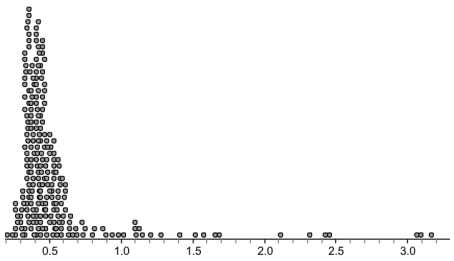
2.



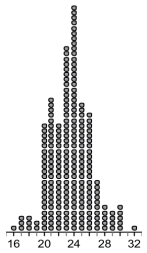
3.



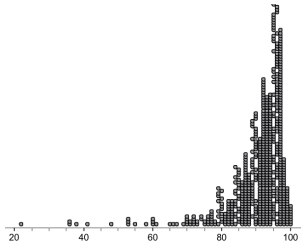
4.



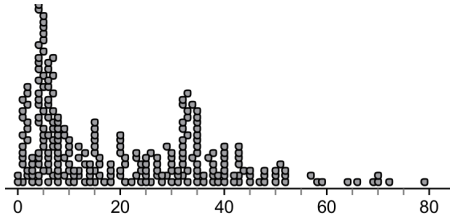
5.



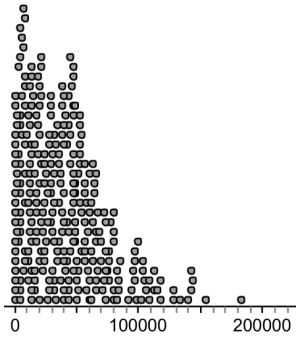
6.



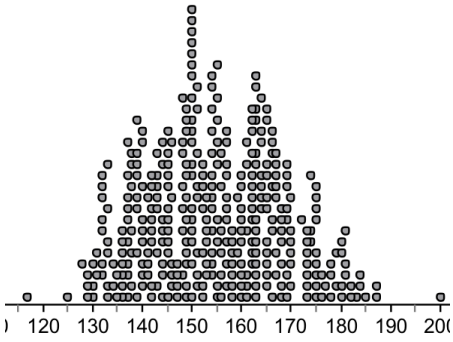
7.



8.

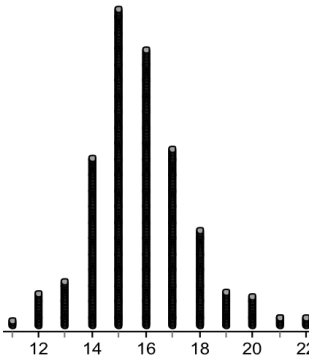


9.

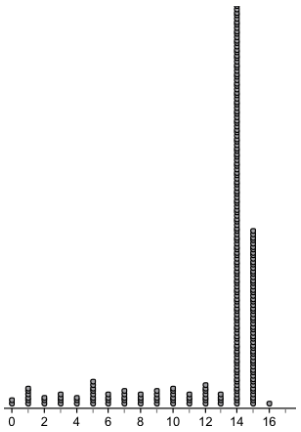


[5] Thanks to Dr Pip Arnold for the graphs.

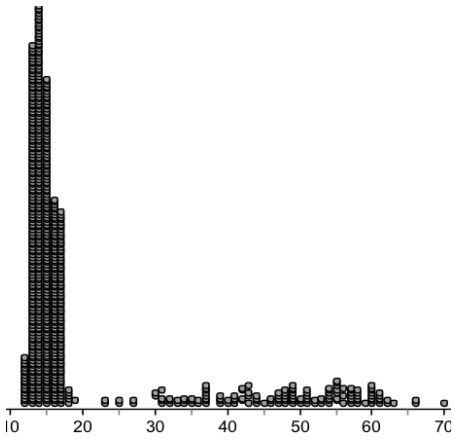
10.



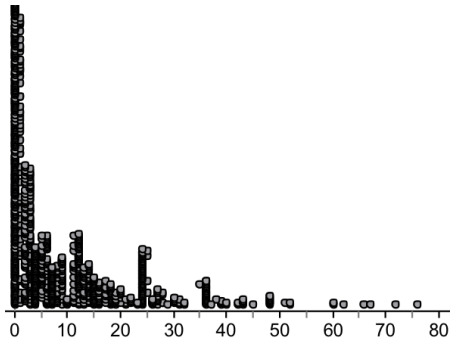
11.



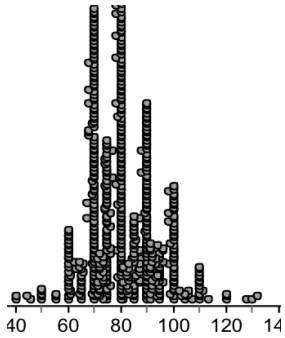
12.



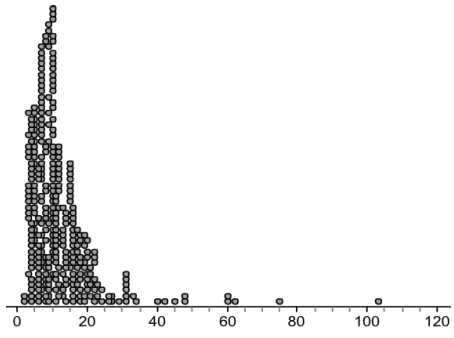
13.



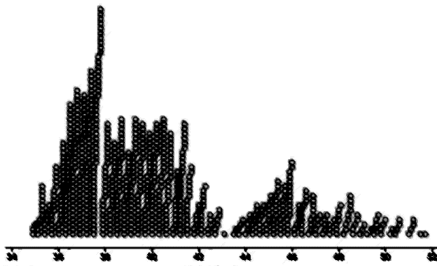
14.



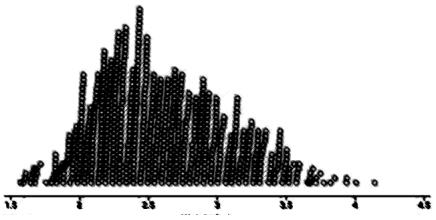
15.



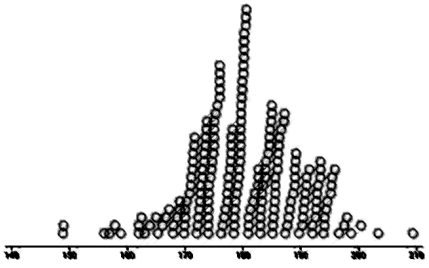
16.

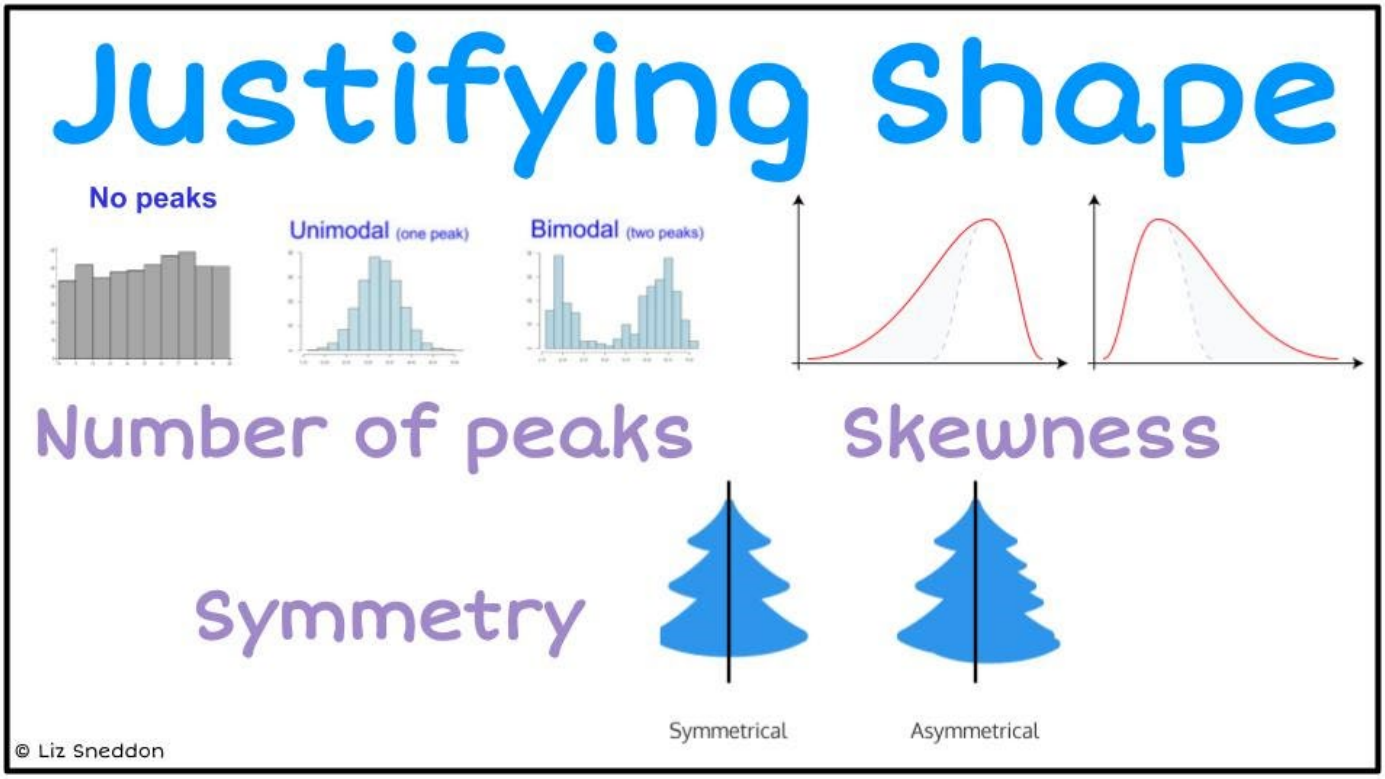


17.

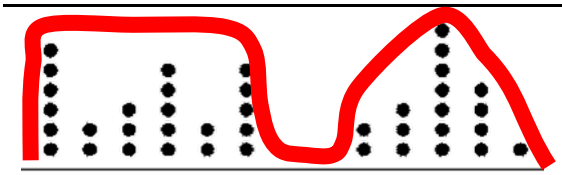


18.



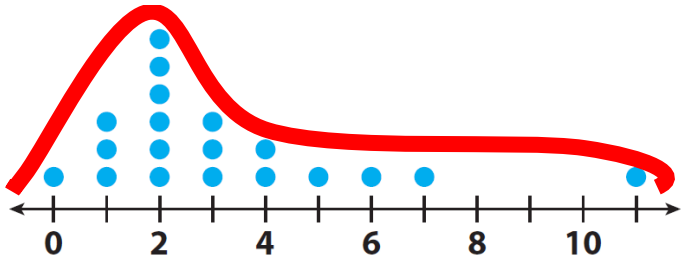


Examples:

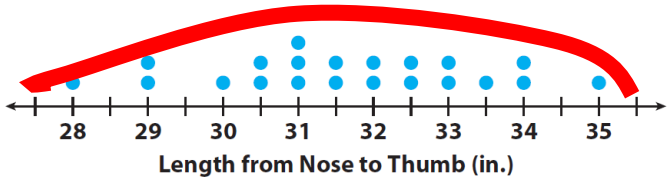


There are two hills / peaks, the data is not symmetric in shape and the tails on each side are different.

Overall, the shape of the data appears to be bimodal.



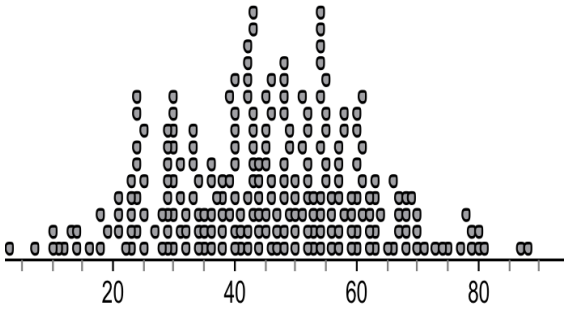
The shape of the data looks right skewed, as there is a longer tail on the right, with one hill / peak. It is not symmetric.



The shape of the data looks approximately normal, as there is a symmetric bell-shaped curve with one hill / peak and the tails on both sides are similar.

Exercise 13:

For each graph, select the correct symmetry, peaks, and tail description.



Symmetric (circle one): Yes / No

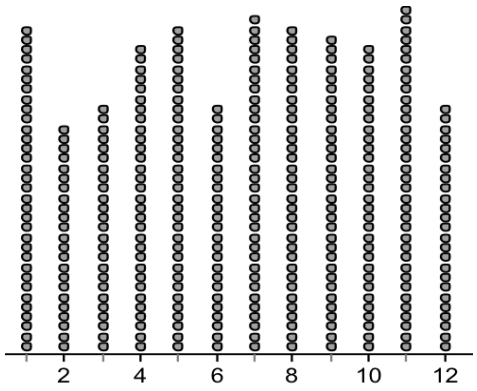
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

☐ Left hand tail longer

☐ Right hand tail longer



Symmetric (circle one): Yes / No

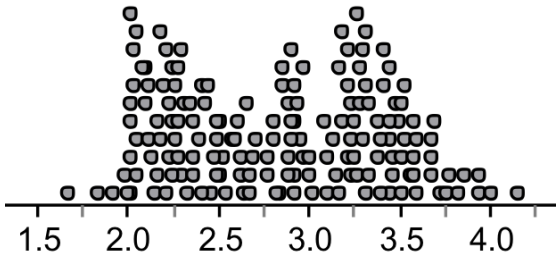
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

☐ Left hand tail longer

☐ Right hand tail longer



Symmetric (circle one): Yes / No

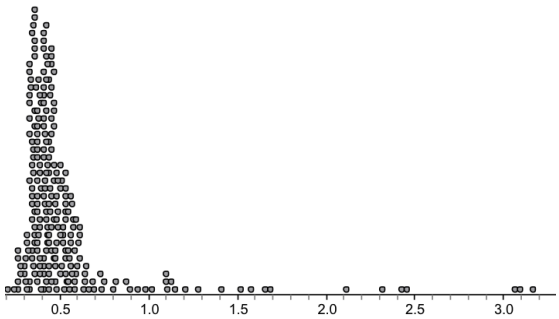
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

☐ Left hand tail longer

☐ Right hand tail longer



Symmetric (circle one): Yes / No

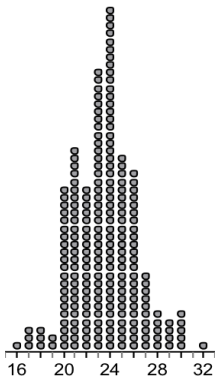
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

☐ Left hand tail longer

☐ Right hand tail longer



Symmetric (circle one): Yes / No

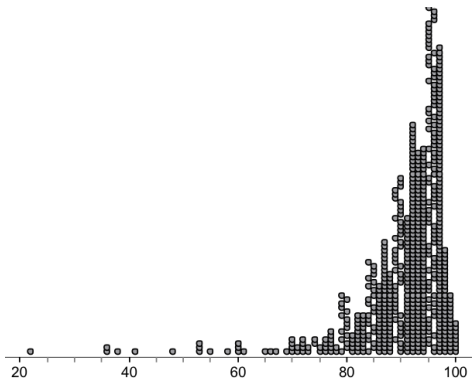
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

☐ Left hand tail longer

☐ Right hand tail longer



Symmetric (circle one): Yes / No

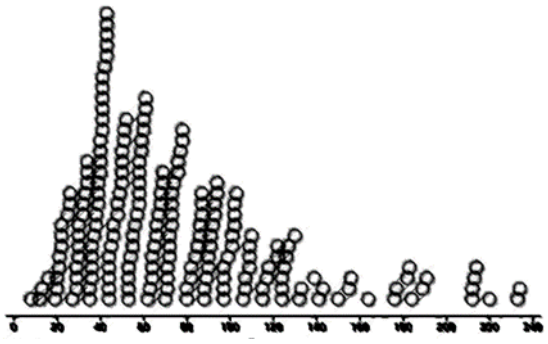
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

☐ Left hand tail longer

☐ Right hand tail longer



Symmetric (circle one): Yes / No

Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

☐ Left hand tail longer

☐ Right hand tail longer

Shapes with NZGrapher

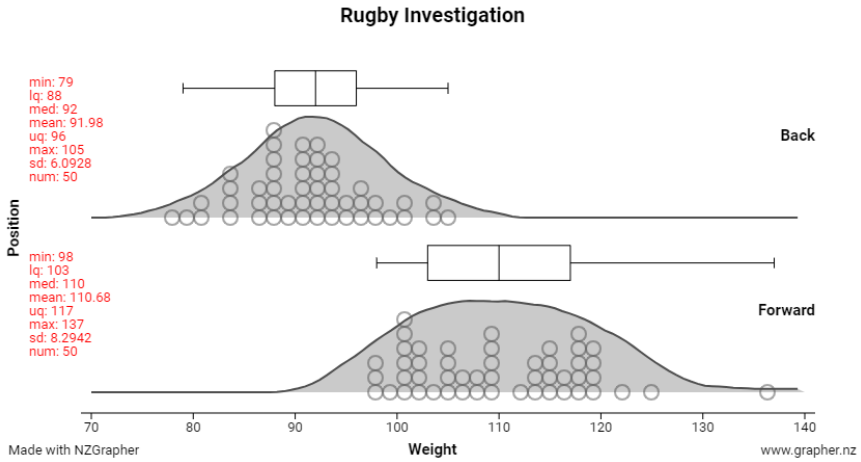
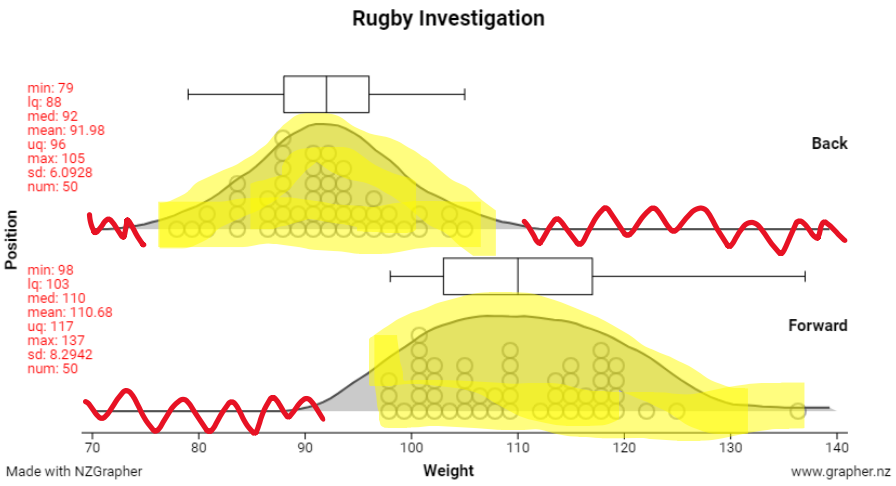
When NZGrapher draws the shape outline over each dataset, it is endeavoring to show what the **shape of the population BEHIND the sample of data** might look like. This is why the curves are smooth, and this is the shape we are trying to identify.

When NZGrapher we need to remember that when we identify the shape, we **ONLY** consider the shape between the **minimum and maximum** values of **EACH group**. (Use the whiskers on the box plot to remind you of these minimums and maximums).

The best way to do this, is to use a highlighter to colour in the data section of the graph to allow you to focus on only that section.

You also need to **IGNORE** any of the tails that extend beyond the data.

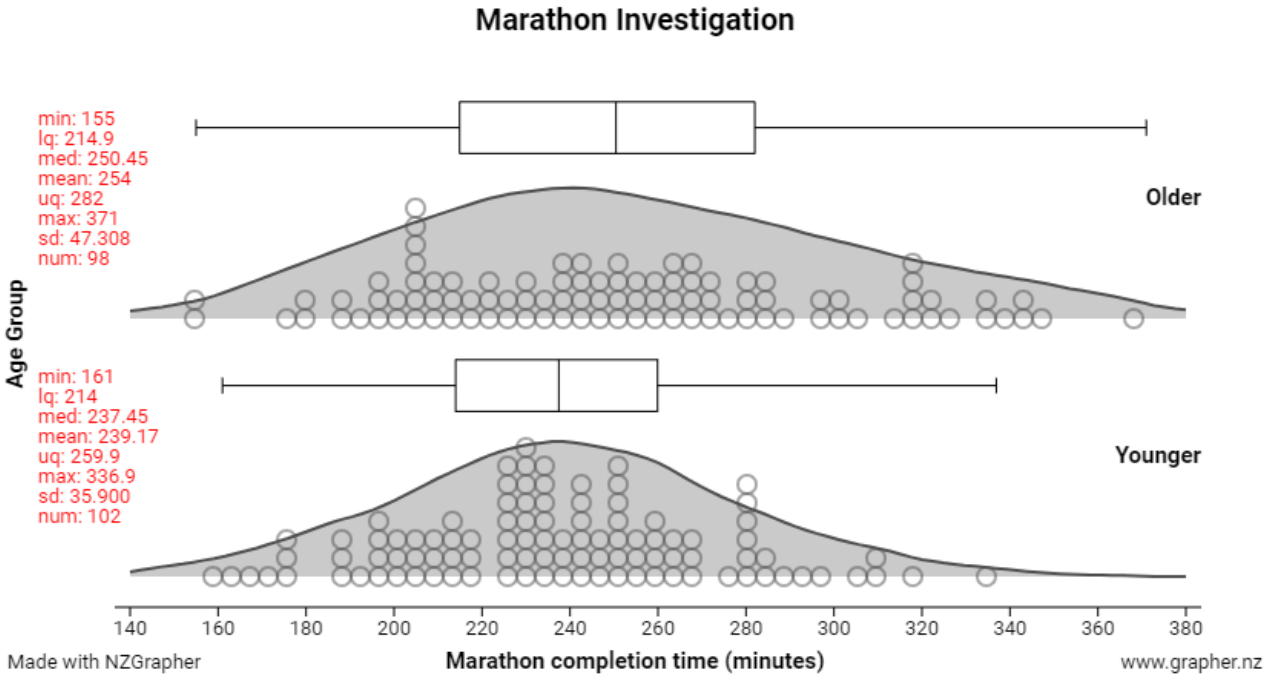
Example:

Step 1: Get the graph from NZGrapher	
Step 2: Highlight the areas between the minimum and maximum of each group, ignoring the tails.	
Step 3: Identify the shape in context, referring to both the numeric and categorical variables. Then justify both shapes.	The shape of weights for the sample of back rugby players is approximately Normal, because the weights are roughly symmetrical, unimodal and follow a bell-shaped curve. The shape of weights for the forwards rugby players is skewed to the right, because the weights are unimodal, asymmetric and more spread out the right-hand side.

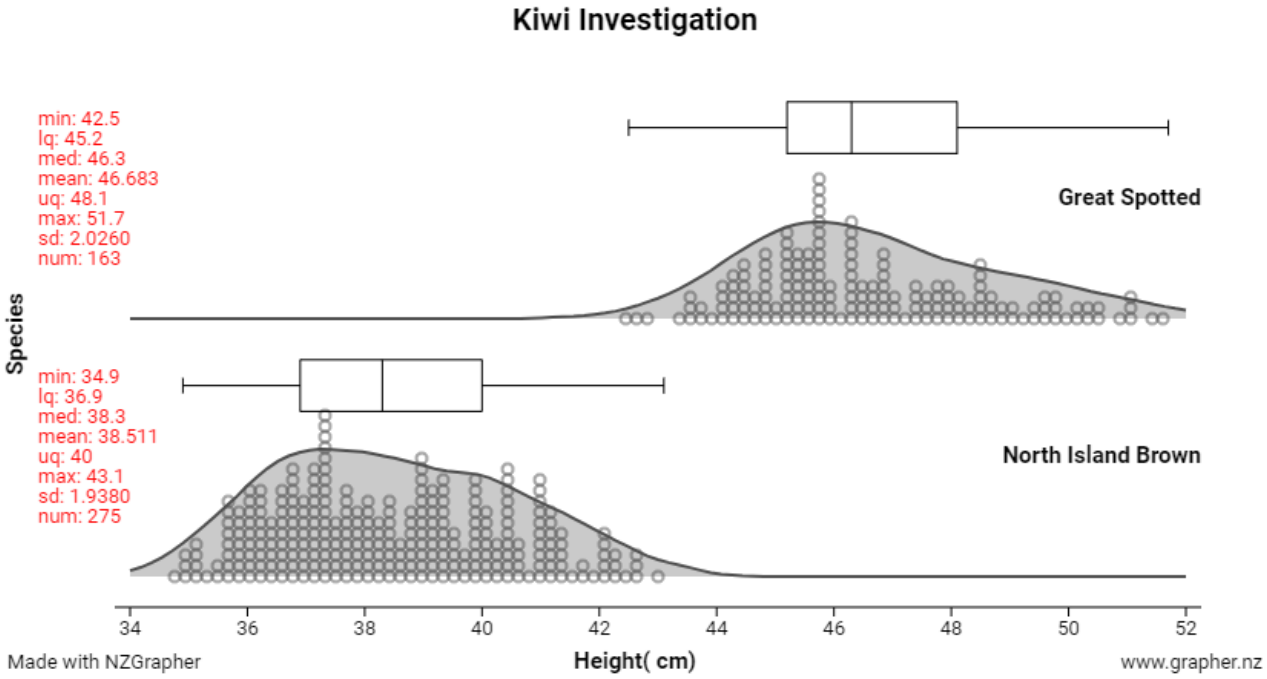
Exercise 14:

Identify and justify the shapes of each of the graphs below, making sure your comments are in context.

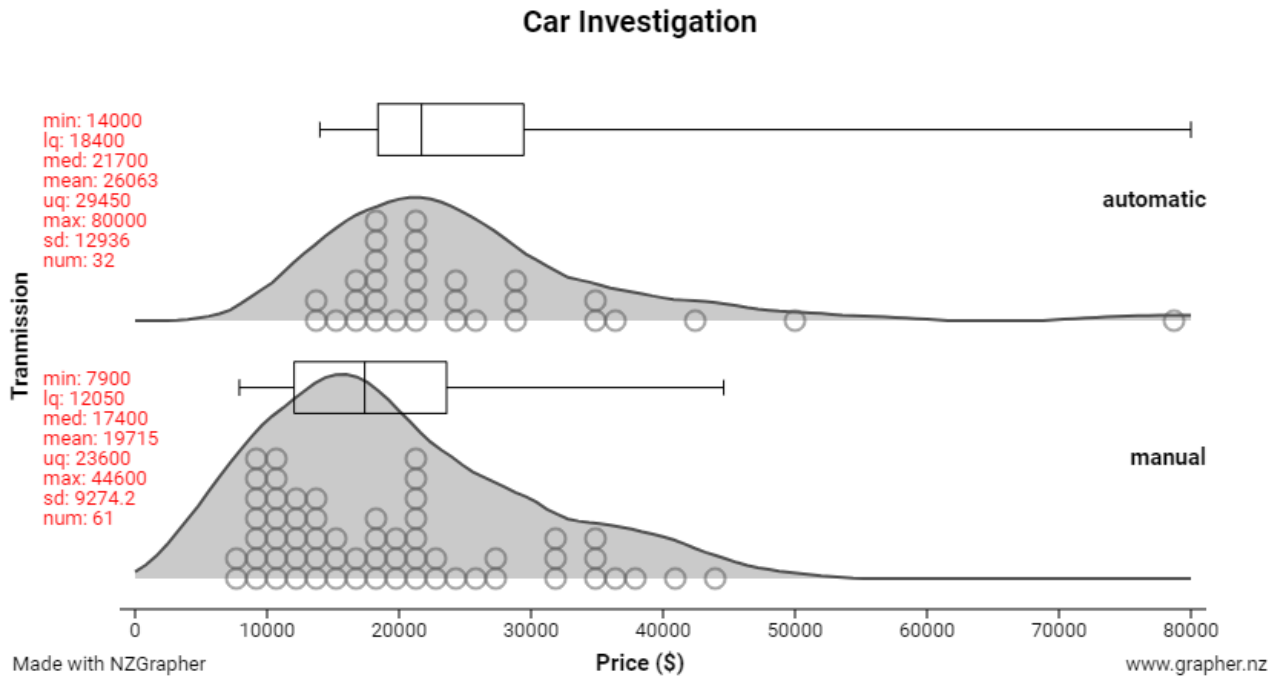
a)



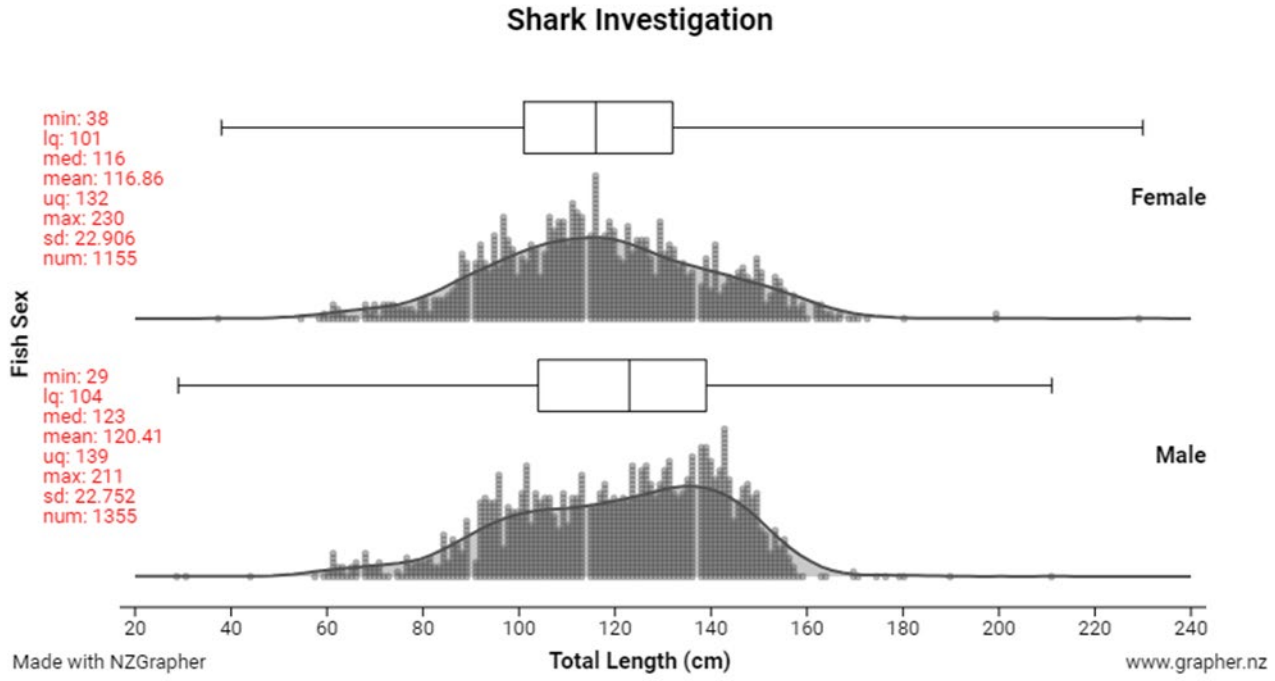
b)


Kiwi Investigation
min: 42.5
lq: 45.2
med: 46.3
mean: 46.683
uq: 48.1
max: 51.7
sd: 2.0260
num: 163
Great Spotted
Species
min: 34.9
lq: 36.9
med: 38.3
mean: 38.511
uq: 40
max: 43.1
sd: 1.9380
num: 275
North Island Brown
34
36
38
40
42
44
46
48
50
52
Made with NZGrapher
Height(cm)
www.grapher.nz

c)


Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Made with NZGrapher
Price ($)
www.grapher.nz

d)

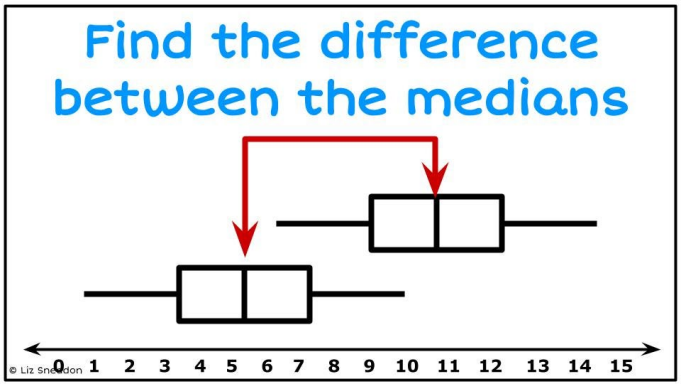

Shark Investigation
min: 38
lq: 101
med: 116
mean: 116.86
uq: 132
max: 230
sd: 22.906
num: 1155
Female
Fish Sex
min: 29
lq: 104
med: 123
mean: 120.41
uq: 139
max: 211
sd: 22.752
num: 1355
Male
20
40
60
80
100
120
140
160
180
200
220
240
Made with NZGrapher
Total Length (cm)
www.grapher.nz

Comparing Centre

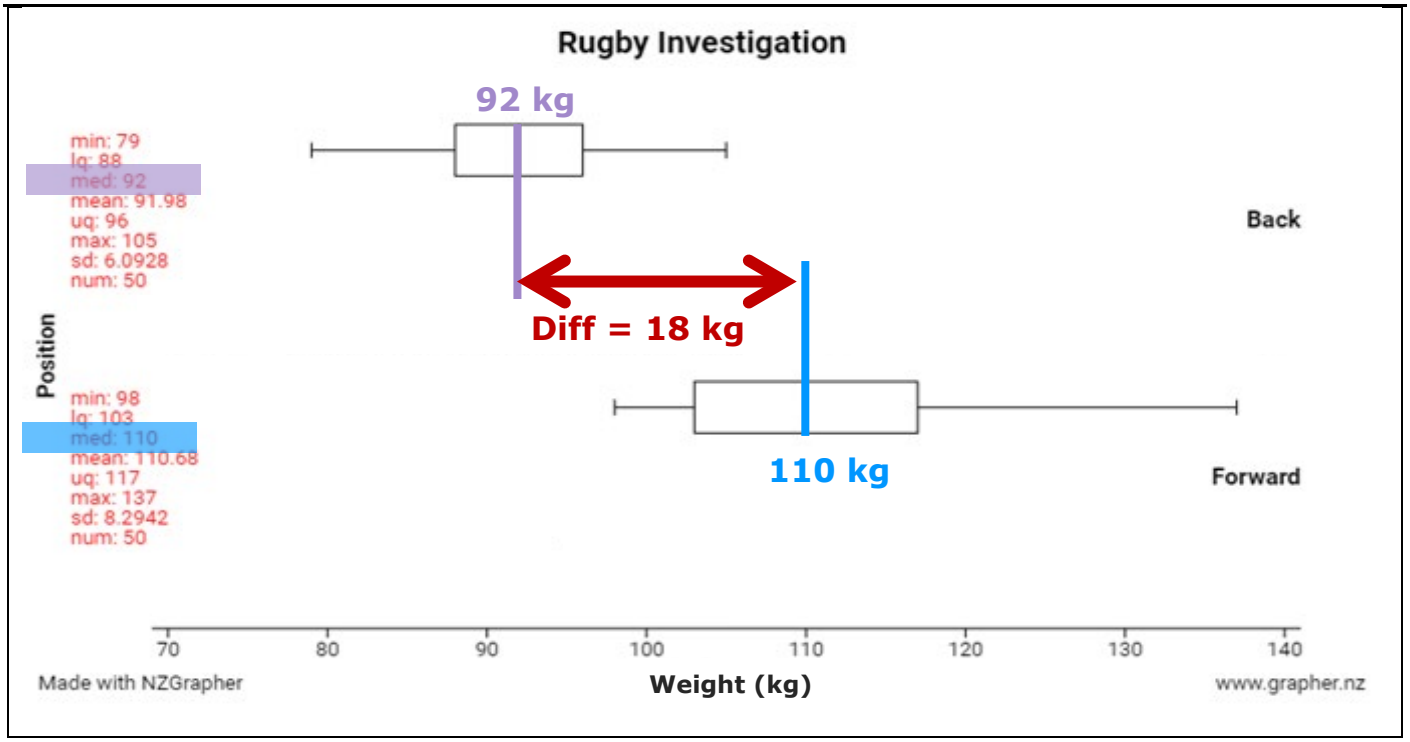
Compare the centres by **finding the medians** of both groups, identifying **which group has a larger median**, and then calculating **HOW** much bigger the median is.

For Merit, make sure that you include the median values and units of each group as evidence.

For Excellence, tell the story and connect to research, explaining why one group may (or may not) be bigger or smaller than the second group.



Example:



The median weight for my sample of backs is 92kg. The median weight for my sample of forwards is 110 kg.

Difference = forwards median – backs median

= 110 kg – 92 kg

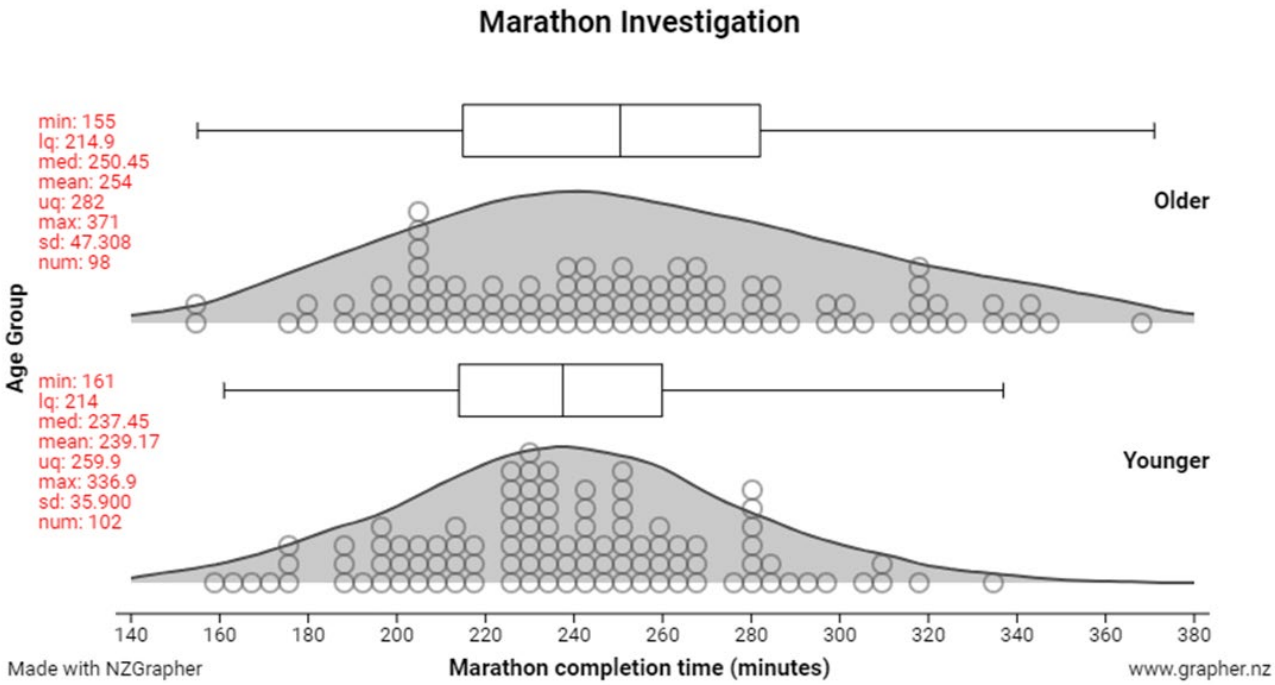
= 18 kg

In my sample, the median weight for forwards rugby players is heavier than backs by 18 kg, as forwards need more muscles and weight to be able to both hold the line and push the line forwards, whereas backs tend to need to run fast, which often is less bulk than rugby forwards.

Exercise 15:

Compare the medians for the graphs below.

1)

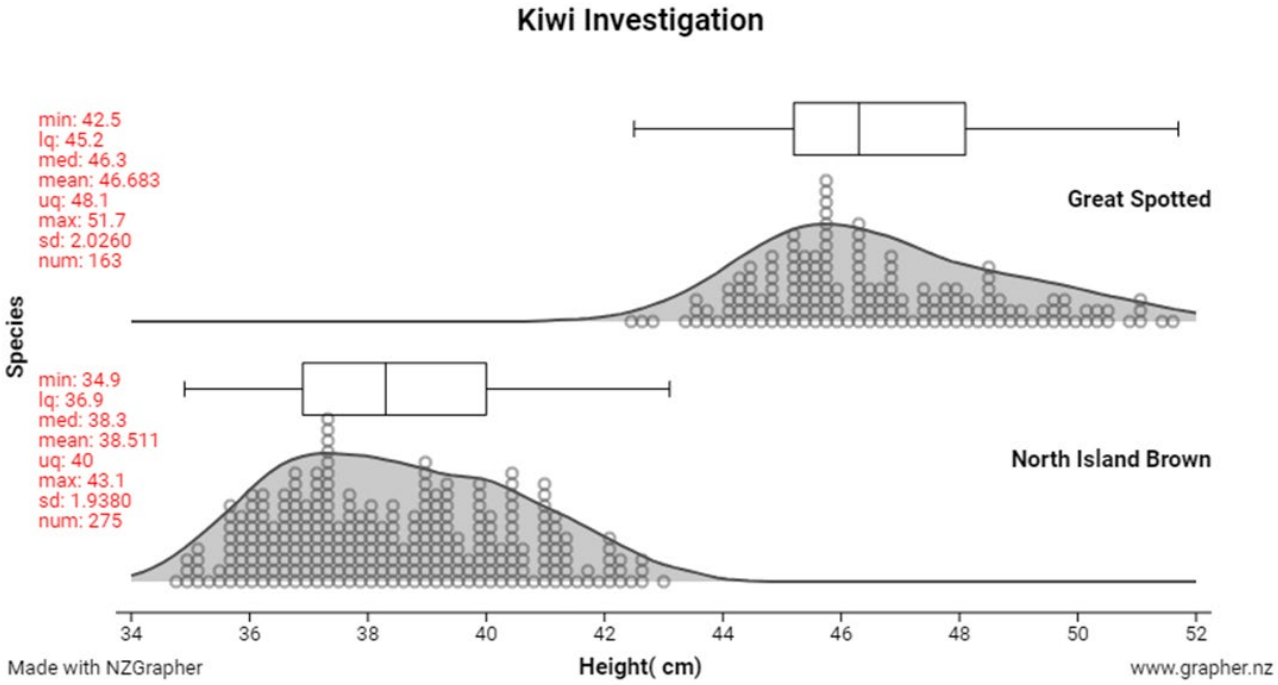


Median marathon time for Older runners	
Median marathon time for Younger runners	
Which group has a longer median time?	Older / Younger
Difference in median times	

In the sample, the median marathon time for older runners (_____ minutes) is _____ than the median marathon time for younger runners (_____ minutes) by _____.

_____ runners may take longer to finish a marathon because …

2)

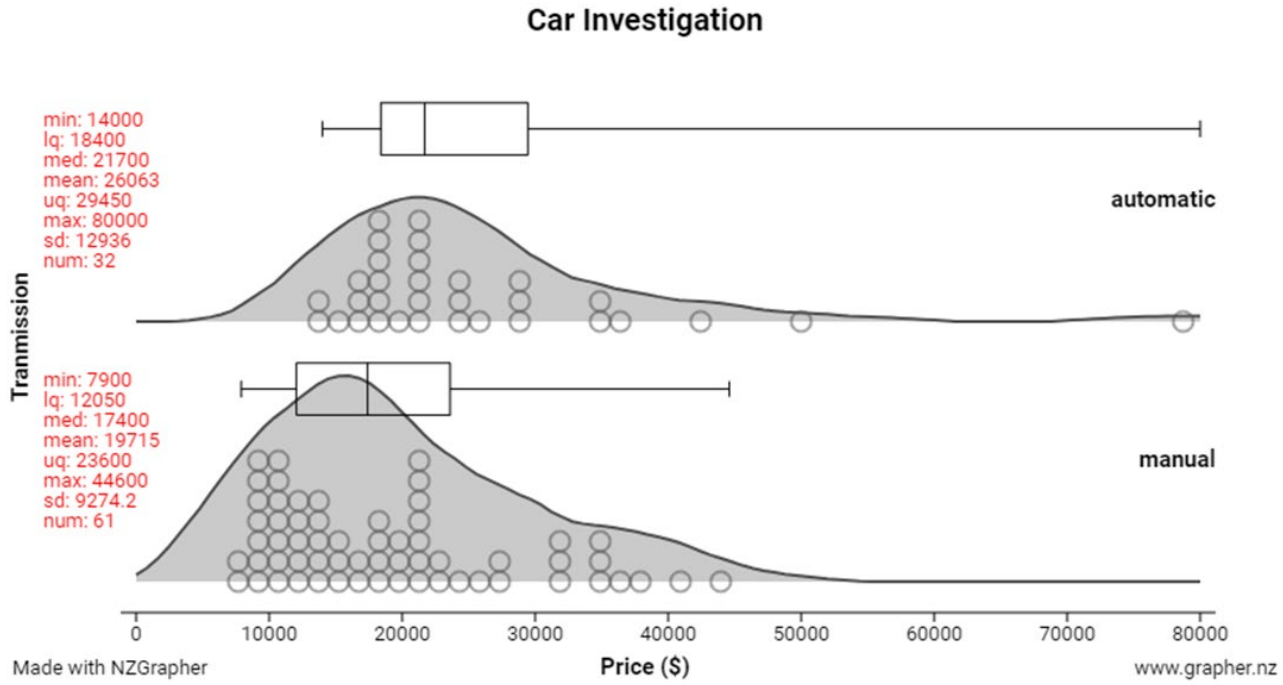


Median height for Great Spotted kiwis	
Median height for North Island Brown kiwis	
Which group has a larger median height?	Great Spotted / North Island Brown
Difference in median height	

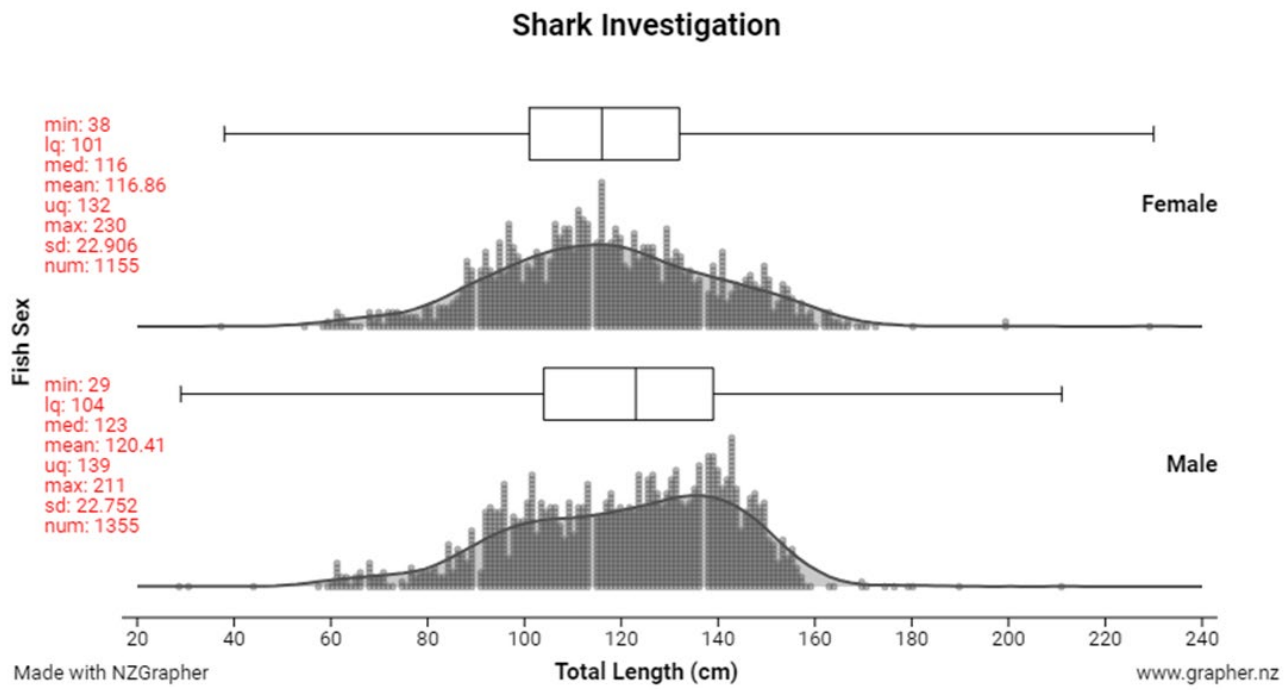
In the sample, the median height of Great Spotted kiwis (_____ cm) is _____ than the median height of North Island Brown kiwis (_____ cm) by _____.

_____ kiwis may be taller because …

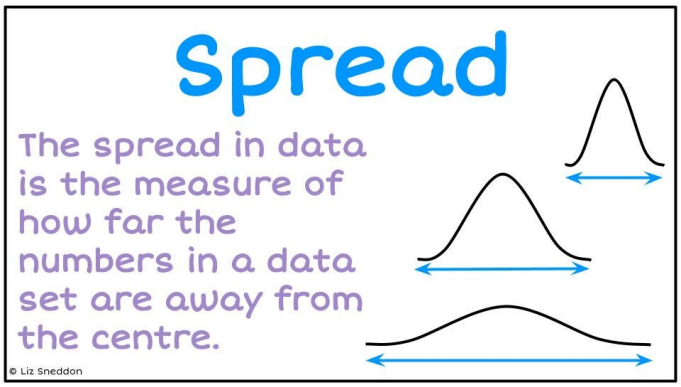
3)

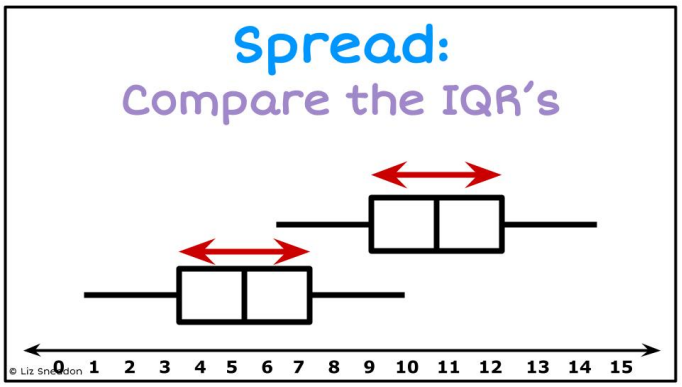

Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Price ($)
Made with NZGrapher
www.grapher.nz

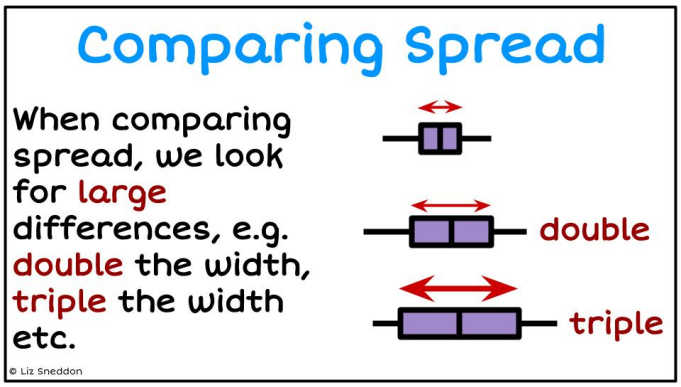
4)

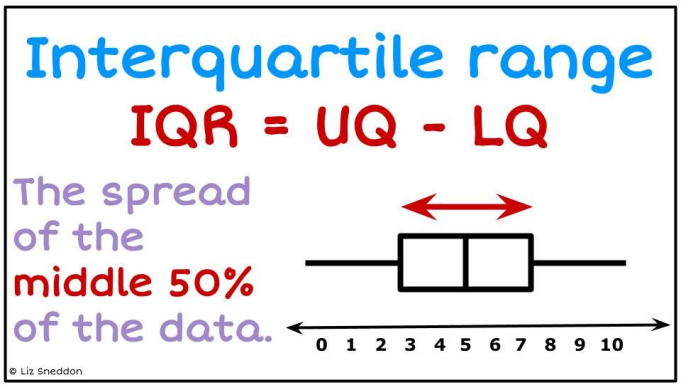

Shark Investigation
min: 38
lq: 101
med: 116
mean: 116.86
uq: 132
max: 230
sd: 22.906
num: 1155
Female
Fish Sex
min: 29
lq: 104
med: 123
mean: 120.41
uq: 139
max: 211
sd: 22.752
num: 1355
Male
20
40
60
80
100
120
140
160
180
200
220
240
Made with NZGrapher
Total Length (cm)
www.grapher.nz

Comparing Spread









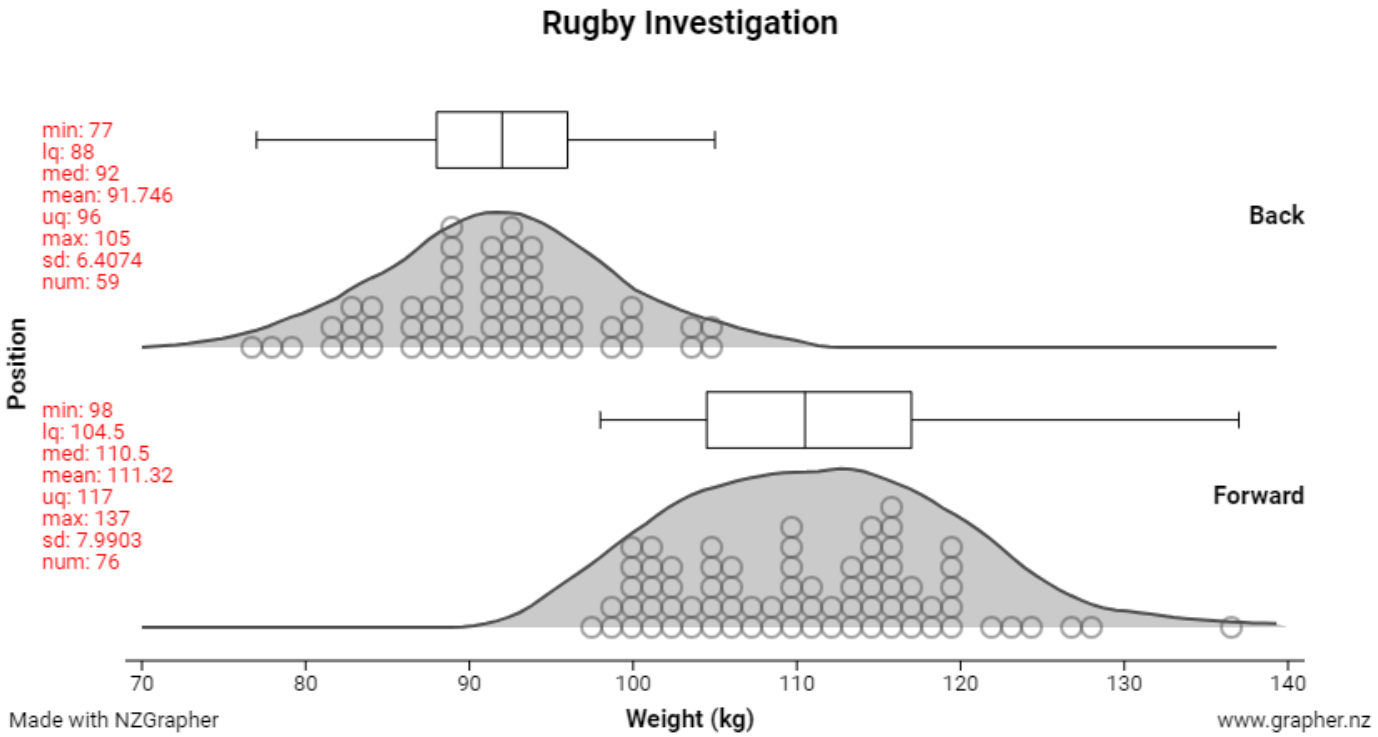
The IQR is the middle 50% of your sample data, and in our boxplot, it is the box.

When comparing the spread, don't calculate the difference between the IQR's, look for whether **one IQR is a multiple of the other**. E.g., double, triple, etc.

For Merit, make sure that you include the IQR values and units of each group as evidence.

For Excellence, you **MAY** be able to tell the story, explaining why one group may (or may not) be more or less spread out than the second group. It is not always easy to explain why one group is more or less spread out, so be careful not to confuse this with comparing the centres and medians.

Example:



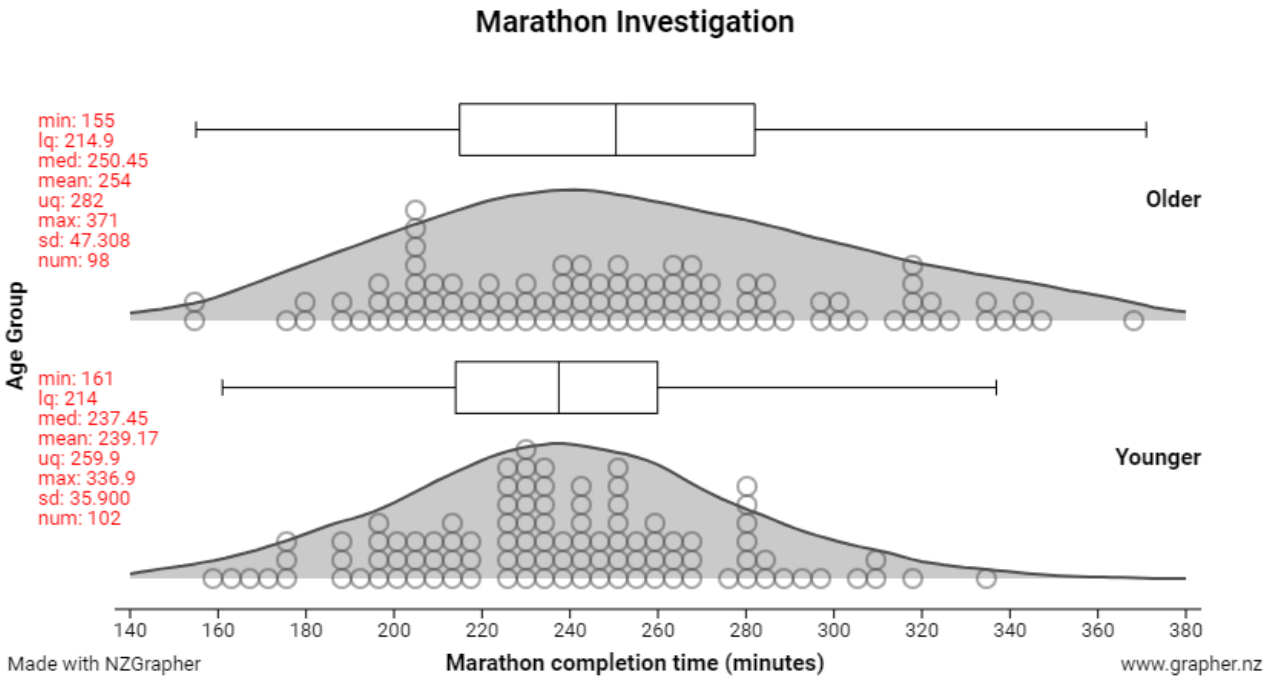
IQR (backs) = UQ – LQ = 96 – 88 = 8 kg The IQR of weights for back rugby players is 8kg.	IQR (forwards) = UQ – LQ = 117 – 1.405 = 12.5 kg The IQR of weights for forwards rugby players is 12.5kg.

In the sample, the spread of the middle 50% of weights of back rugby players is **wider** than the spread of the middle 50% of weights of forward rugby players.

Exercise 16:

Compare the spread for the graphs below.

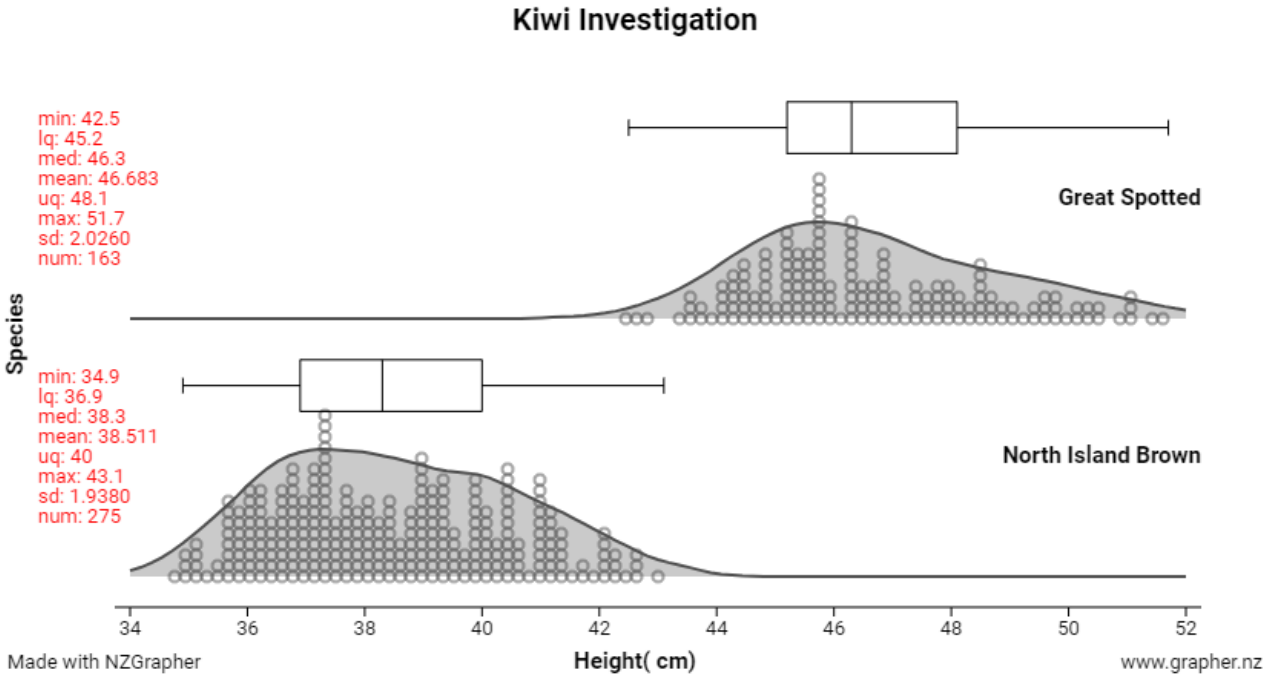
1)



IQR for older runners	
IQR for younger runners	
Compare spread	☐ About the same ☐ Older runners' times spread a little wider ☐ Older runners' times spread a lot wider ☐ Younger runners' times spread a little wider ☐ Younger runners' times spread a lot wider

In the sample, the spread of the middle 50% of marathon times for older runners is _____ than the spread of the middle 50% of marathon times for younger runners.

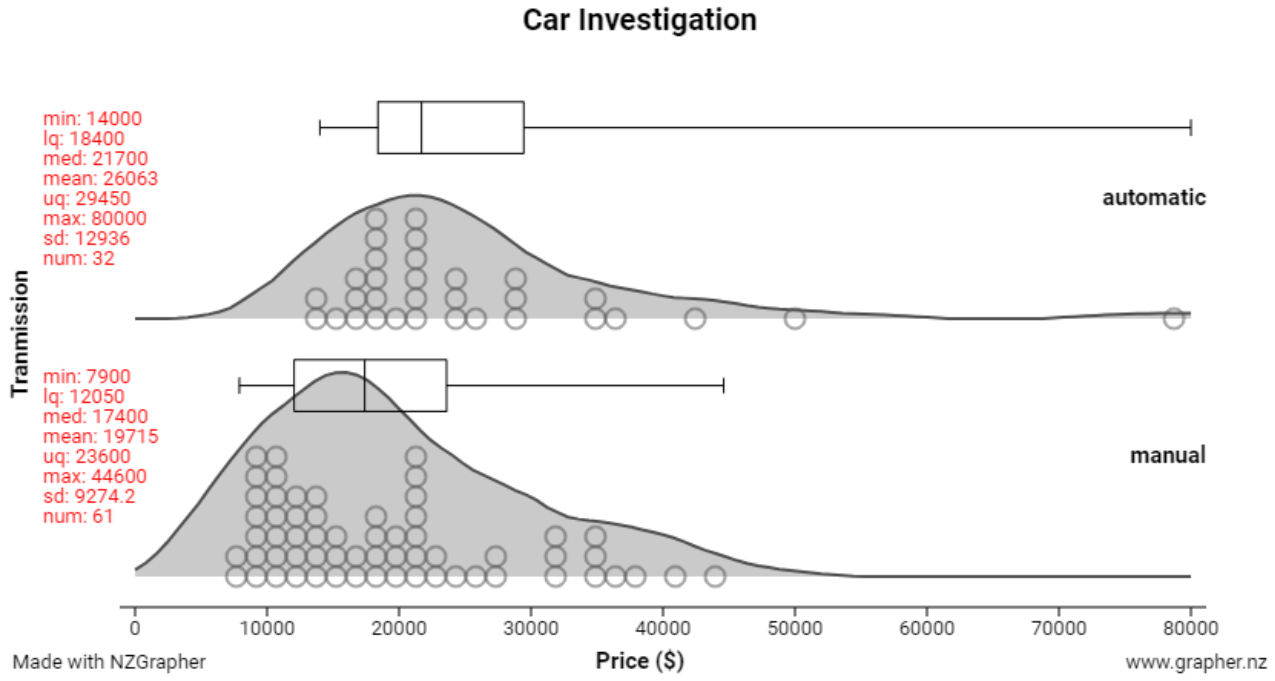
2)



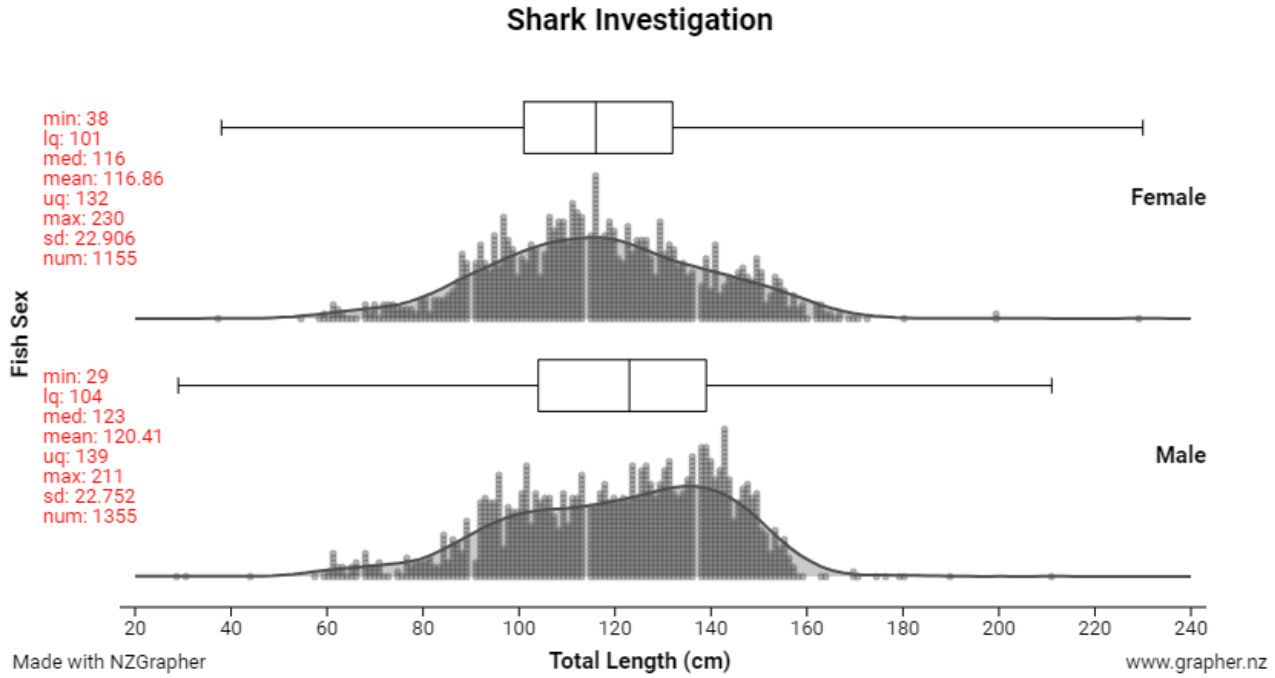
IQR for Great Spotted kiwis	
IQR for North Island Brown kiwis	
Compare spread	☐ About the same ☐ Great Spotted heights spread a little wider ☐ Great Spotted heights spread a lot wider ☐ North Island Brown heights spread a little wider ☐ North Island Brown heights spread a lot wider

In the sample, the spread of the middle 50% of heights for Great Spotted kiwis is _____ than the spread of the middle 50% of heights for North Island Brown kiwis.

3)


Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Made with NZGrapher
Price ($)
www.grapher.nz

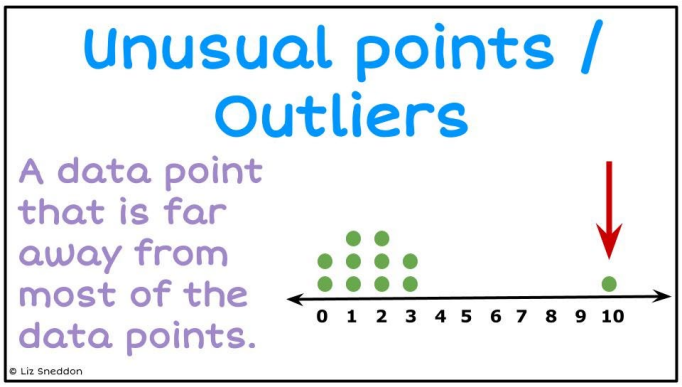
4)


Shark Investigation
min: 38
lq: 101
med: 116
mean: 116.86
uq: 132
max: 230
sd: 22.906
num: 1155
Female
Fish Sex
min: 29
lq: 104
med: 123
mean: 120.41
uq: 139
max: 211
sd: 22.752
num: 1355
Male
20
40
60
80
100
120
140
160
180
200
220
240
Made with NZGrapher
Total Length (cm)
www.grapher.nz

Outliers

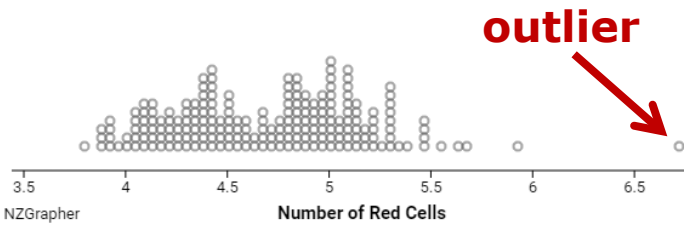
When we look for outliers, you are looking for data that is a long way away from the rest of the data. For example, if a student with a height of 2 meters entered the classroom, their height would be unusual. Or if a person who was 100 years old entered the classroom, that would be unusual.

This is not one of the main patterns we look for, but if you notice an outlier, identify it.



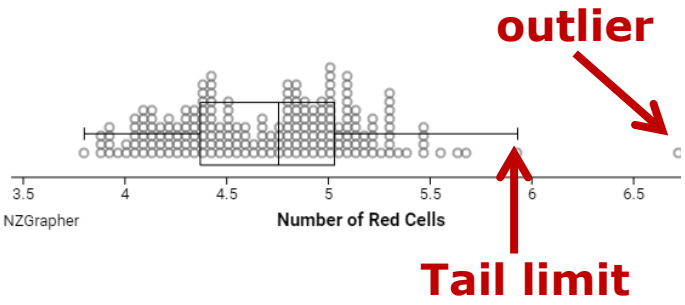
Example 1:

First look at the graph and identify any points that are a **long way away**.



Then confirm it by adding the **Box plot (no outlier)**.

Notice that the whisker stops at the value 5.8, so any data points after this are outliers.



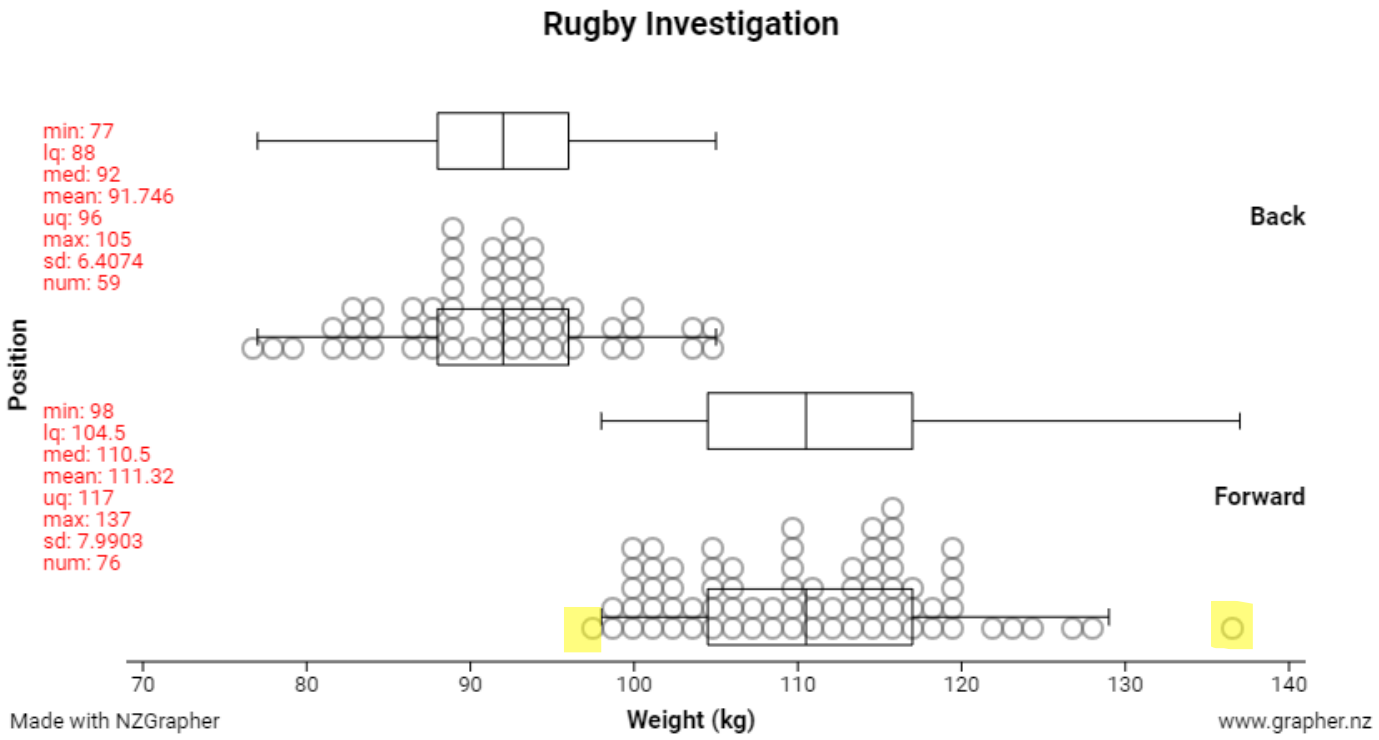
There is an outlier, a person with around 6.8 red blood cells per microliter.

Example 2:

In the graph below, you can see 2 box plots for each group. The high box plot covers the weights from the minimum to the maximum value. The box plot that is overlaid over the data points, is the Box plot (No outliers).

The key is to compare the ends of the Box plot (No outliers) and see if there are any data points (weights) that are outside of the whiskers.

I have highlighted the 2 outliers in the forwards. Notice that they are outside the ends of the Box plot (No Outliers) – but the high box plot still extends to the minimum and maximum values.

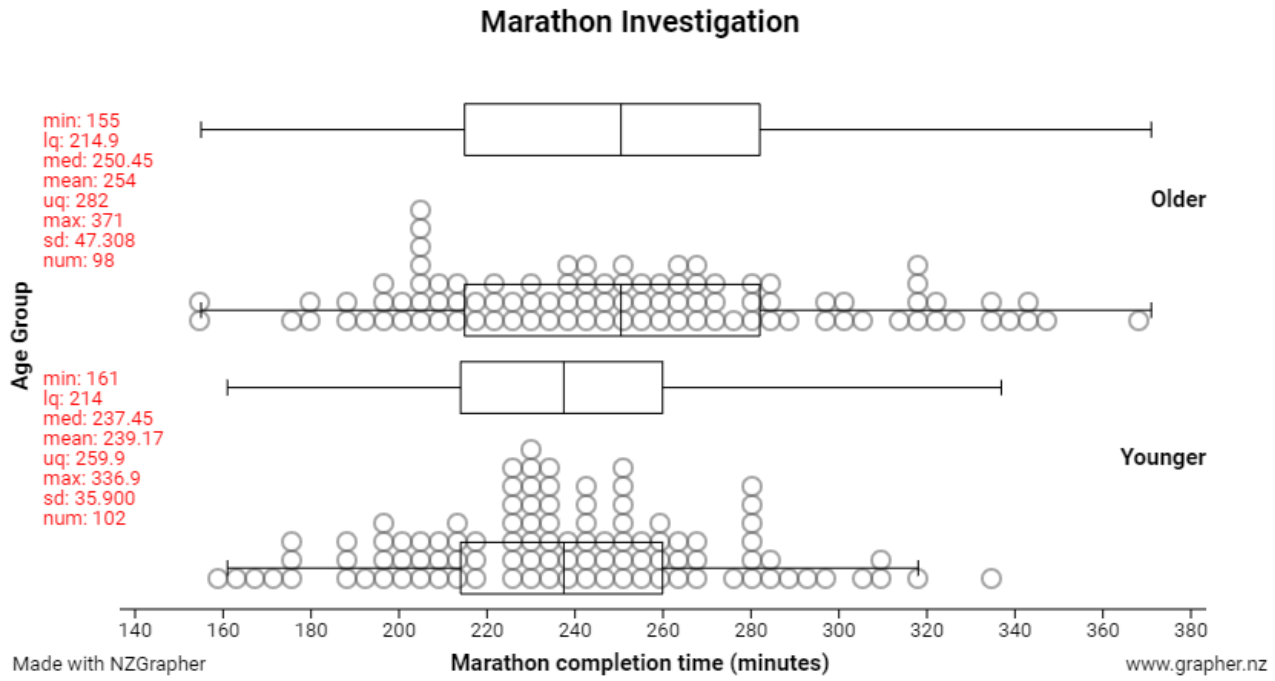


There are two outliers, who both play in a forwards position in rugby. One has a smaller than expected weight of around 96 kg, and another with a weight higher than expected at around 136 kg.

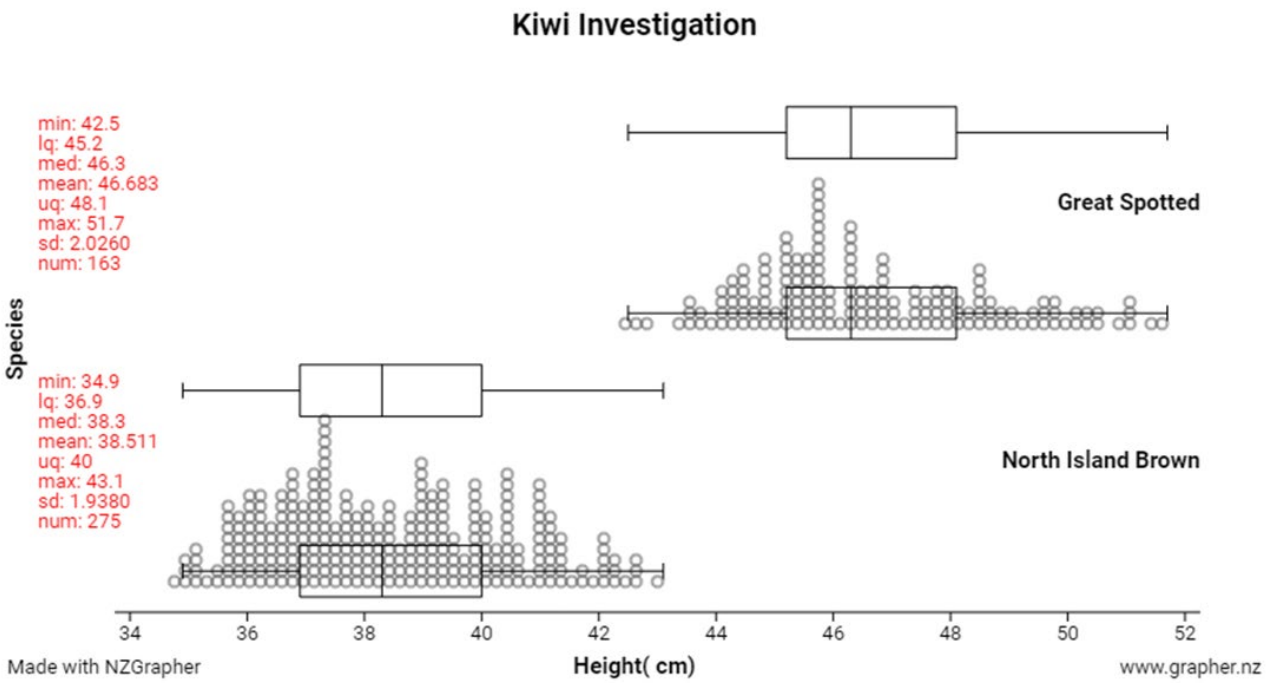
Exercise 17:

Identify and describe any outliers in the graphs below.

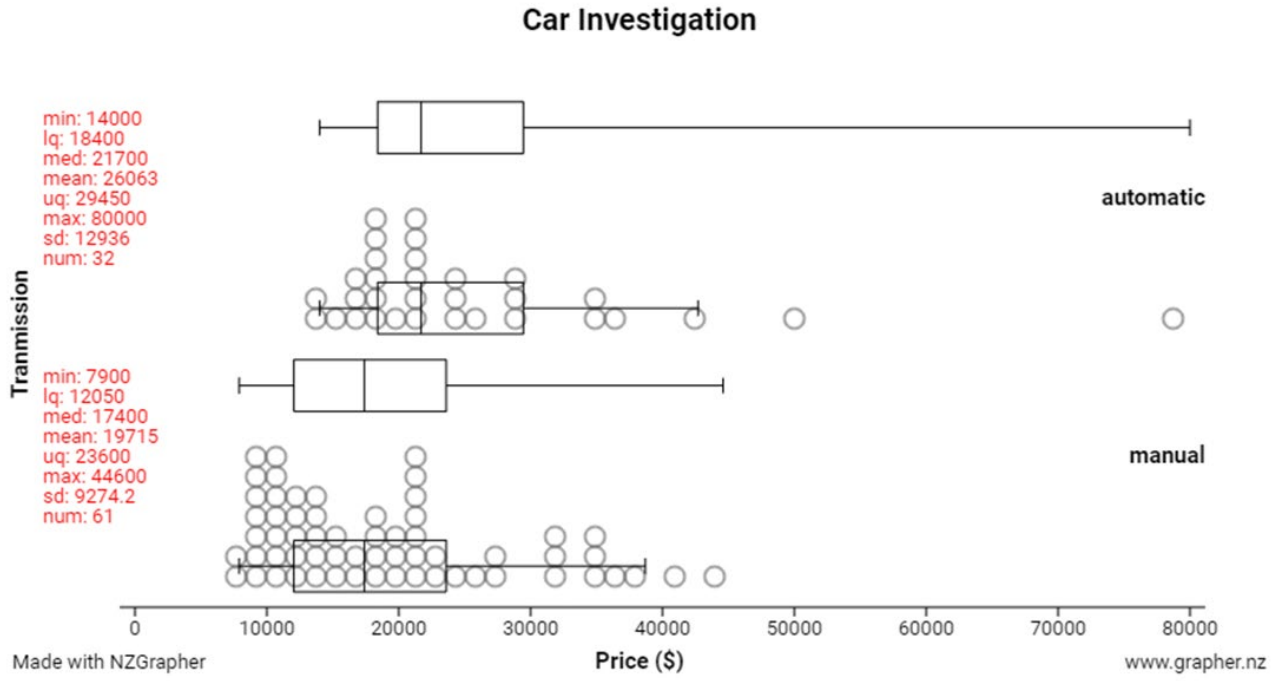
1)



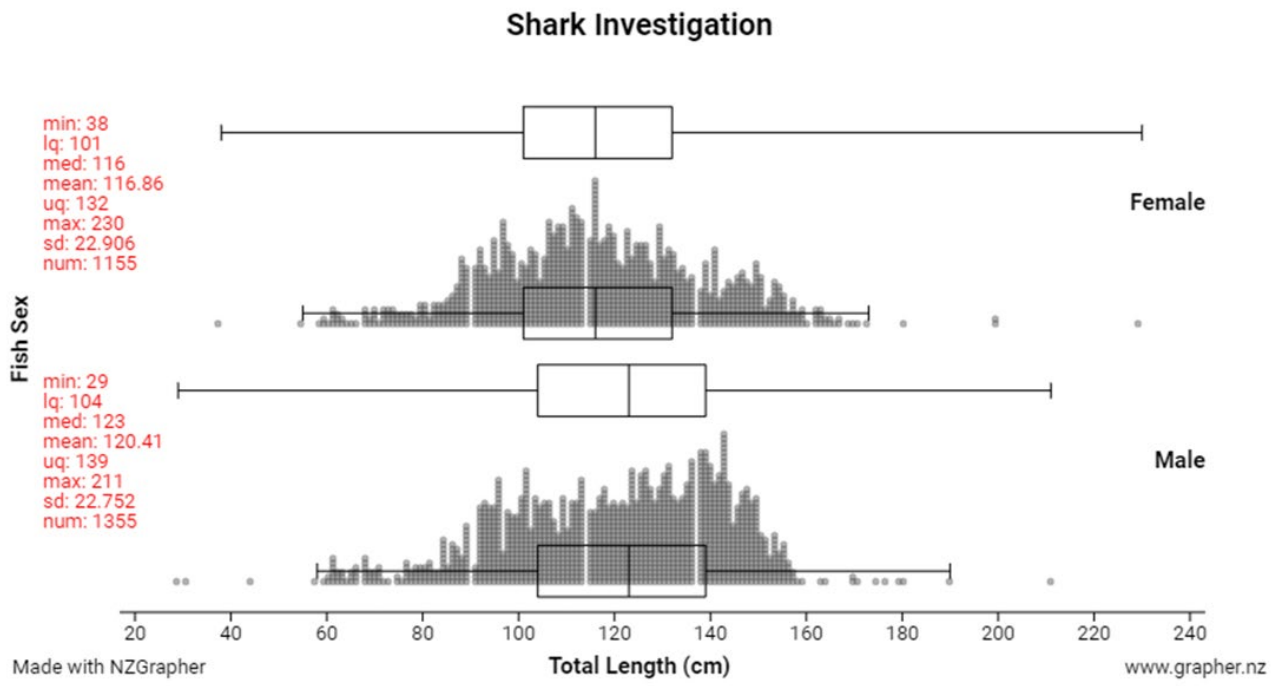
2)


Kiwi Investigation
min: 42.5
lq: 45.2
med: 46.3
mean: 46.683
uq: 48.1
max: 51.7
sd: 2.0260
num: 163
Great Spotted
Species
min: 34.9
lq: 36.9
med: 38.3
mean: 38.511
uq: 40
max: 43.1
sd: 1.9380
num: 275
North Island Brown
34
36
38
40
42
44
46
48
50
52
Made with NZGrapher
Height(cm)
www.grapher.nz

3)

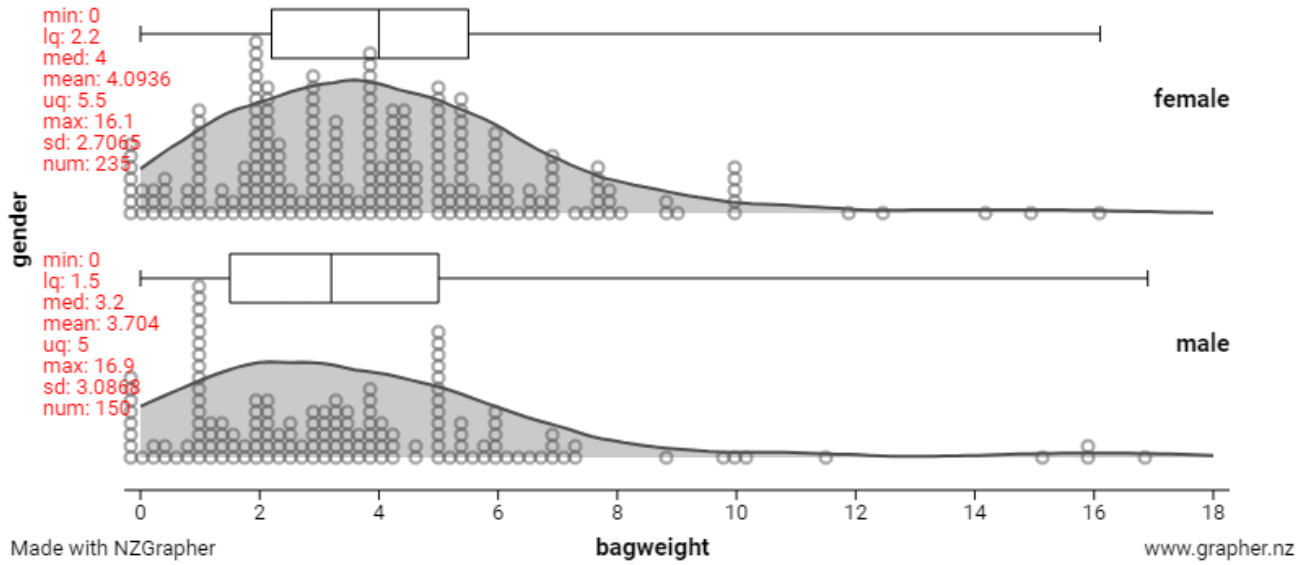

Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Made with NZGrapher
Price ($)
www.grapher.nz

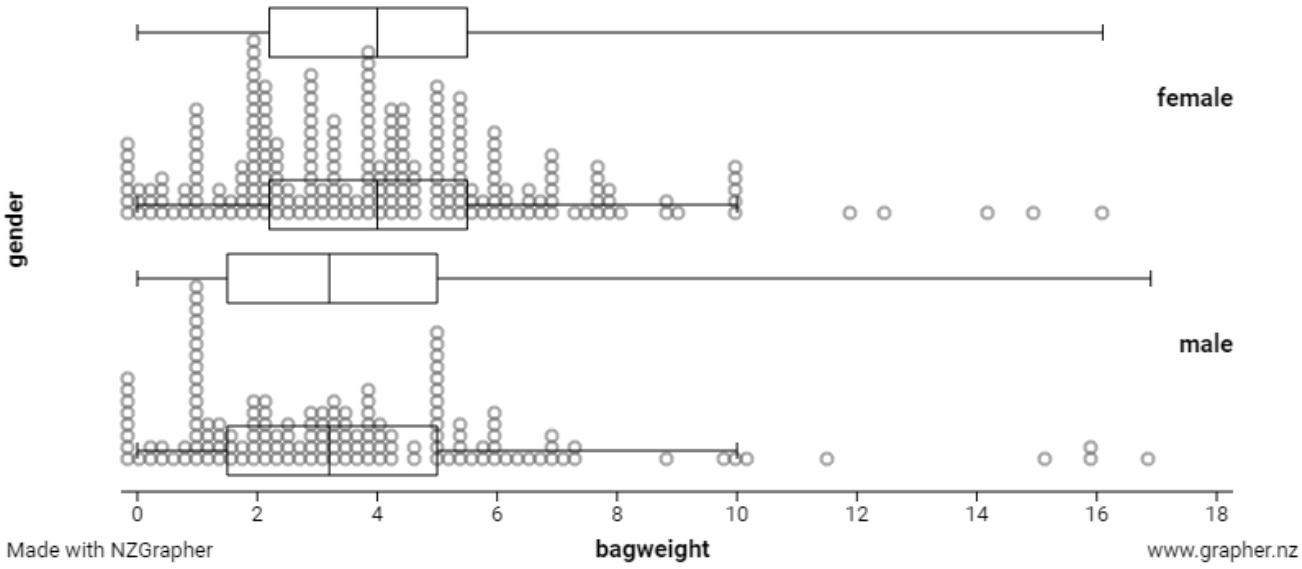
4)


Shark Investigation
min: 38
lq: 101
med: 116
mean: 116.86
uq: 132
max: 230
sd: 22.906
num: 1155
Female
Fish Sex
min: 29
lq: 104
med: 123
mean: 120.41
uq: 139
max: 211
sd: 22.752
num: 1355
Male
20
40
60
80
100
120
140
160
180
200
220
240
Made with NZGrapher
Total Length (cm)
www.grapher.nz

Full Analysis

Example:





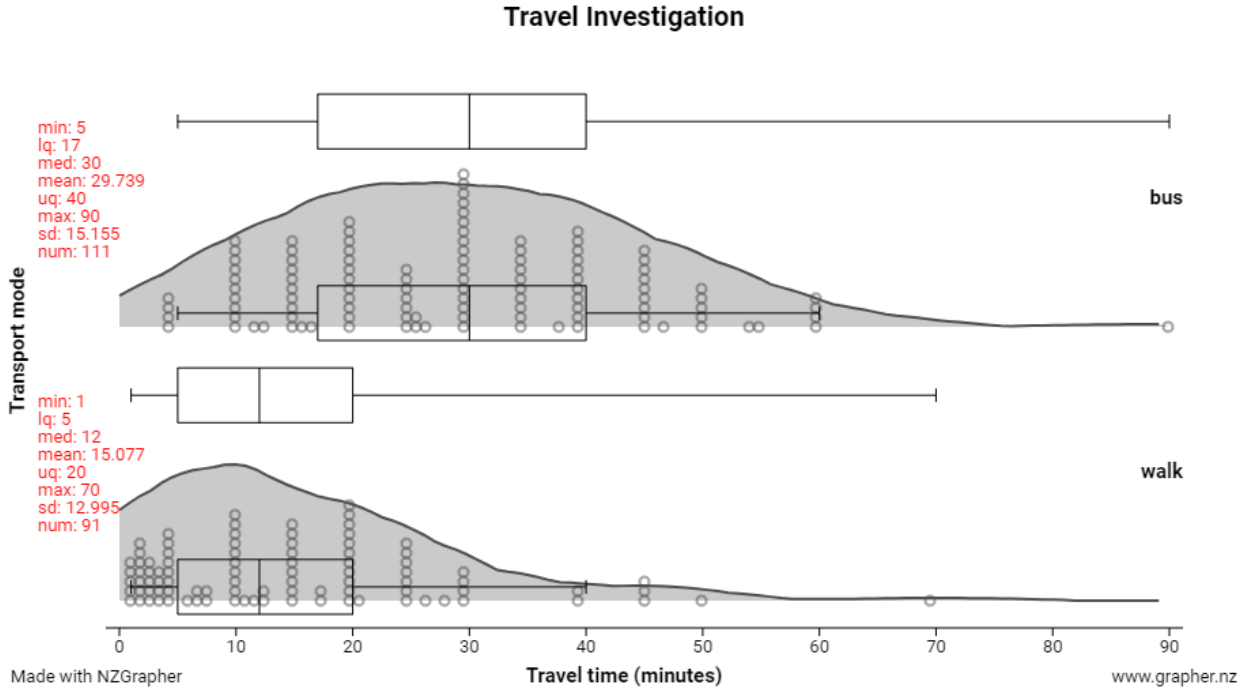
For my sample, I notice:

- The shape of the females and male bag weights have the same right skewed shape. The females and male bag weights are right skewed because they have one peak on the left-hand side, are asymmetric, and there is a longer tail on the right-hand side.
- The median of the female bag weights is a little heavier than the bag weights for males by 0.8kg. My evidence is that the median bag weight for females is around 4kg while the median bag weight for males is around 3.2kg.
- The spread of the middle 50% of males bag weights is slightly larger than the spread of females' bag weights, because the IQR of the females is approximately 3.3 kg compared to the IQR for males of 3.5 kg.
- There are several outliers in the females bag weights: at 12kg, 12.5kg, 14.2kg, 15.1kg and 16.1kg. There are several outliers in the males bag weights at 11.6kg, 15.2kg, two bags at 16kg and 16.9kg.

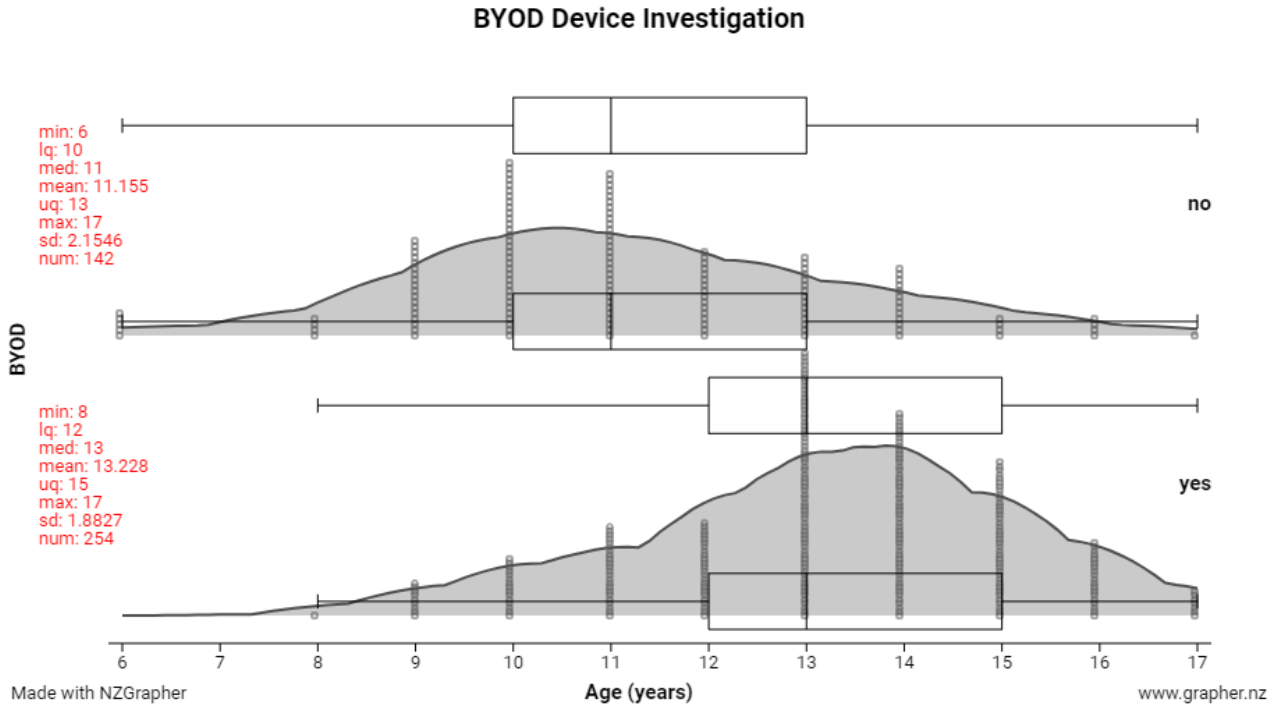
Exercise 18:

For the samples below, write a complete analysis. Discuss features such as: shape, center and spread.

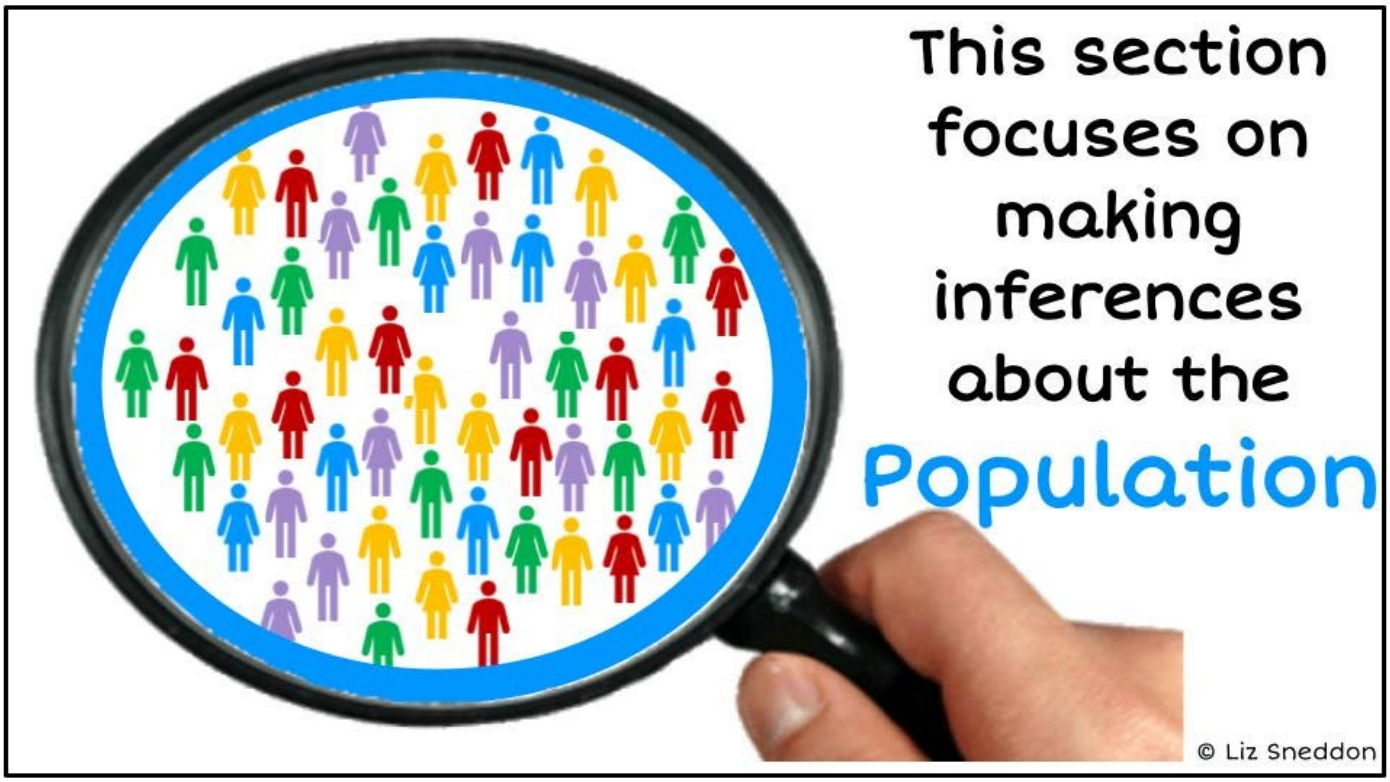
1) The dataset is about high school students and both the method of transport and the length of time it takes them to get to school.

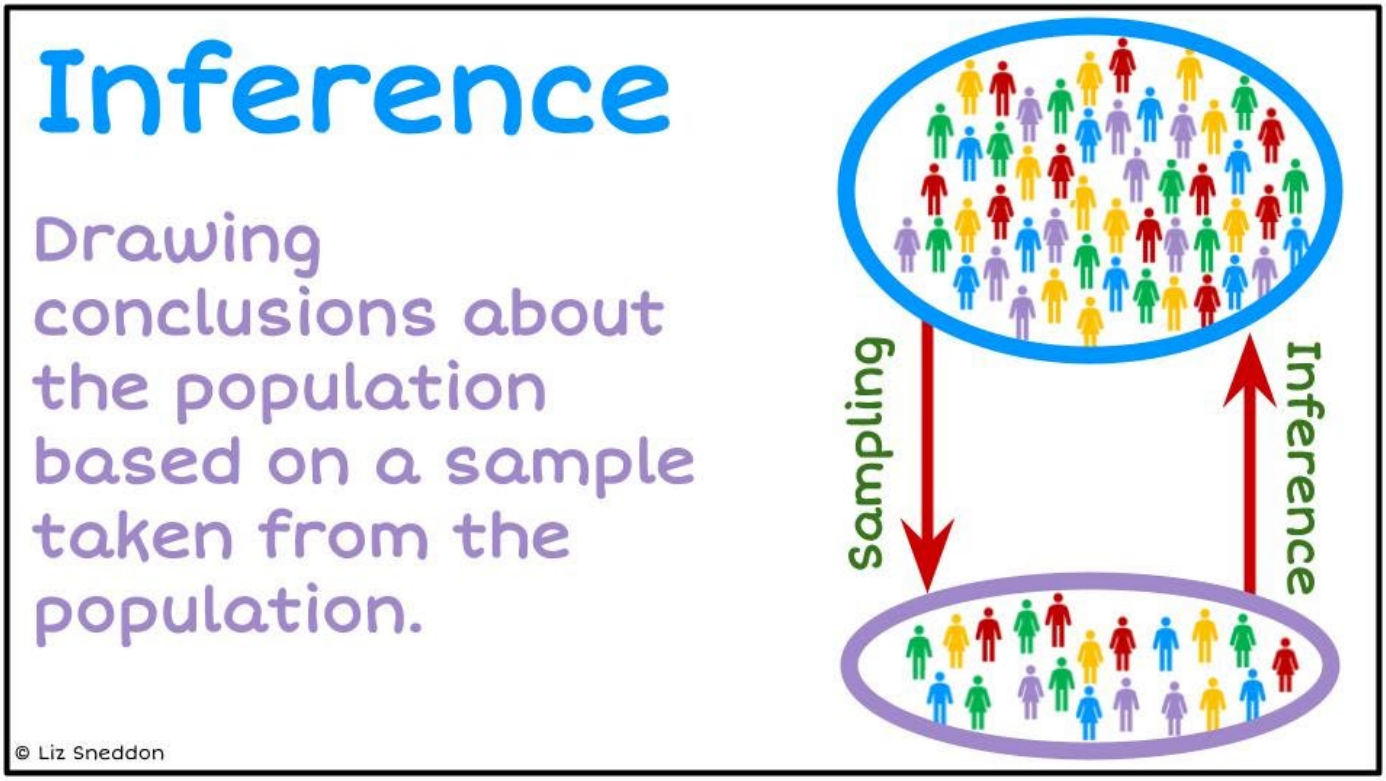


2) This investigation is looking into whether the ages of students who have a device is older than the age of students who don't have a device.



Conclusion / Whakatau



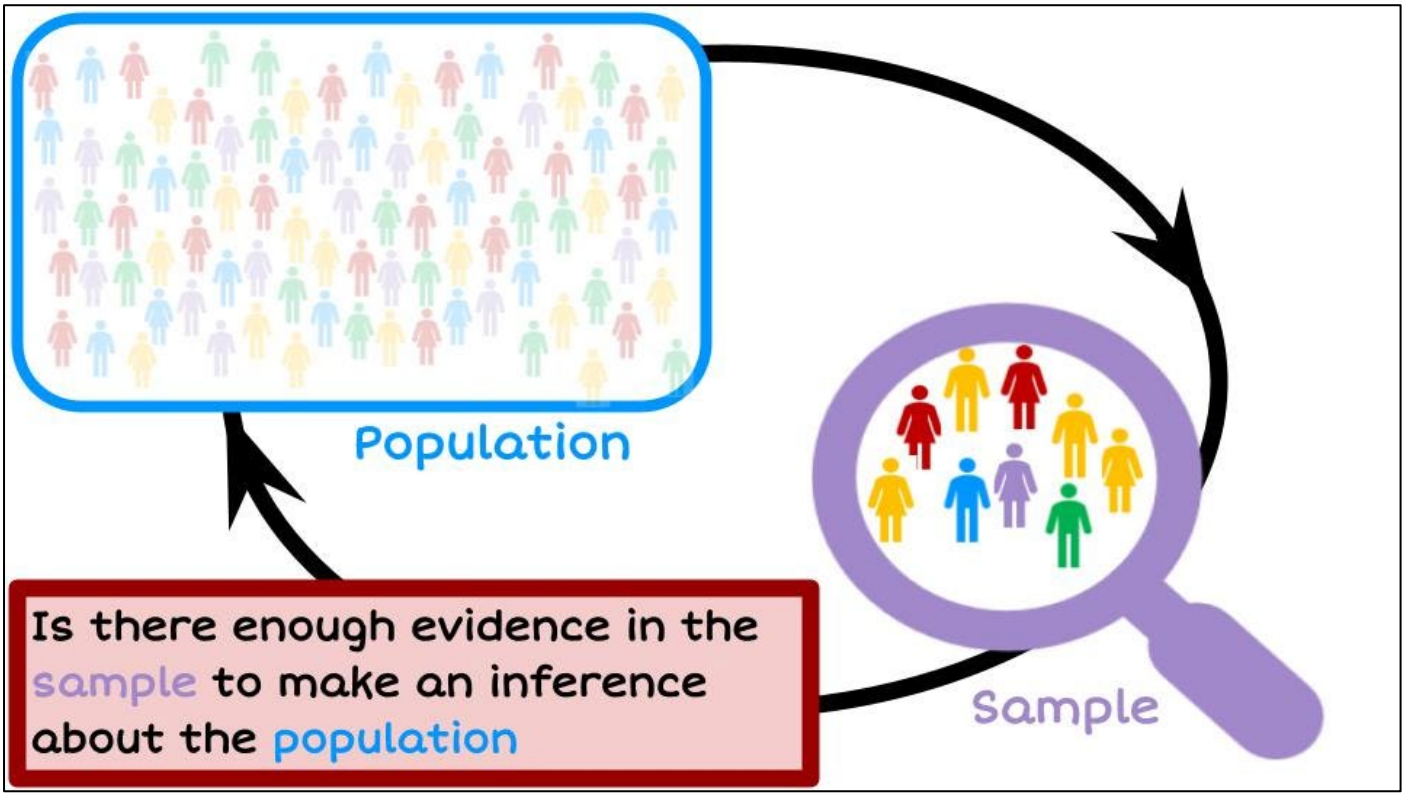


You need to include the following in your conclusion:

- Make an inference about the population using your confidence interval,
- Answer your investigation (can you make the call?),
- Discuss sampling variability.

Answering the investigation question

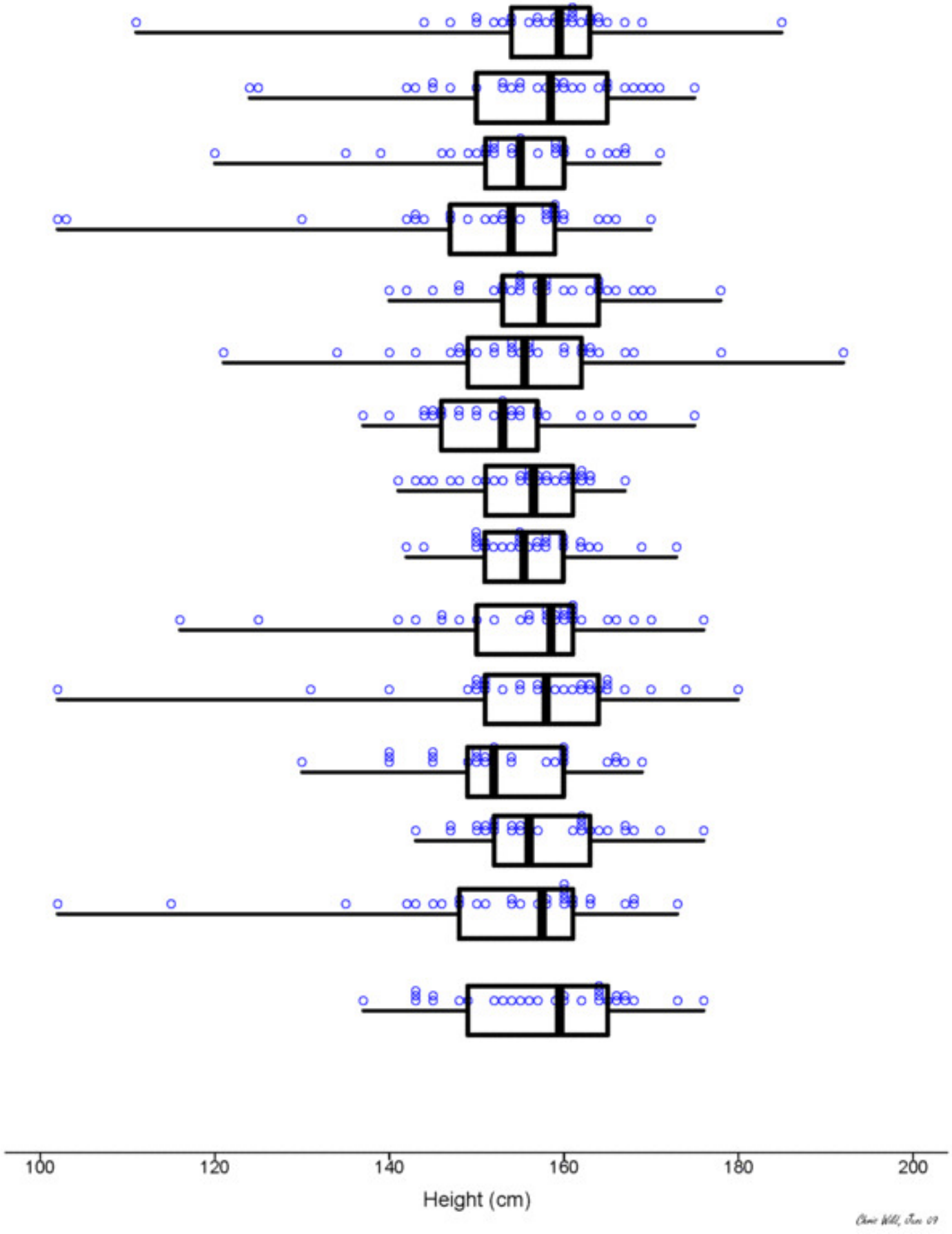
We are looking to see if we have **enough evidence** to say that the median of one group **tends to be larger** than the median of the other group, **back in the population**.



All of these things require an understanding of sampling variability, so we will start by exploring this in more depth.

Exercise 19:

Below are graphs all of which are samples of size 30, from a population of height of 12-year-old girls in NZ.[6] Here is the animation: bit.ly/VariationAnimation



[6] Thanks to Prof. Chris Wild https://www.stat.auckland.ac.nz/~wild/WPRH/

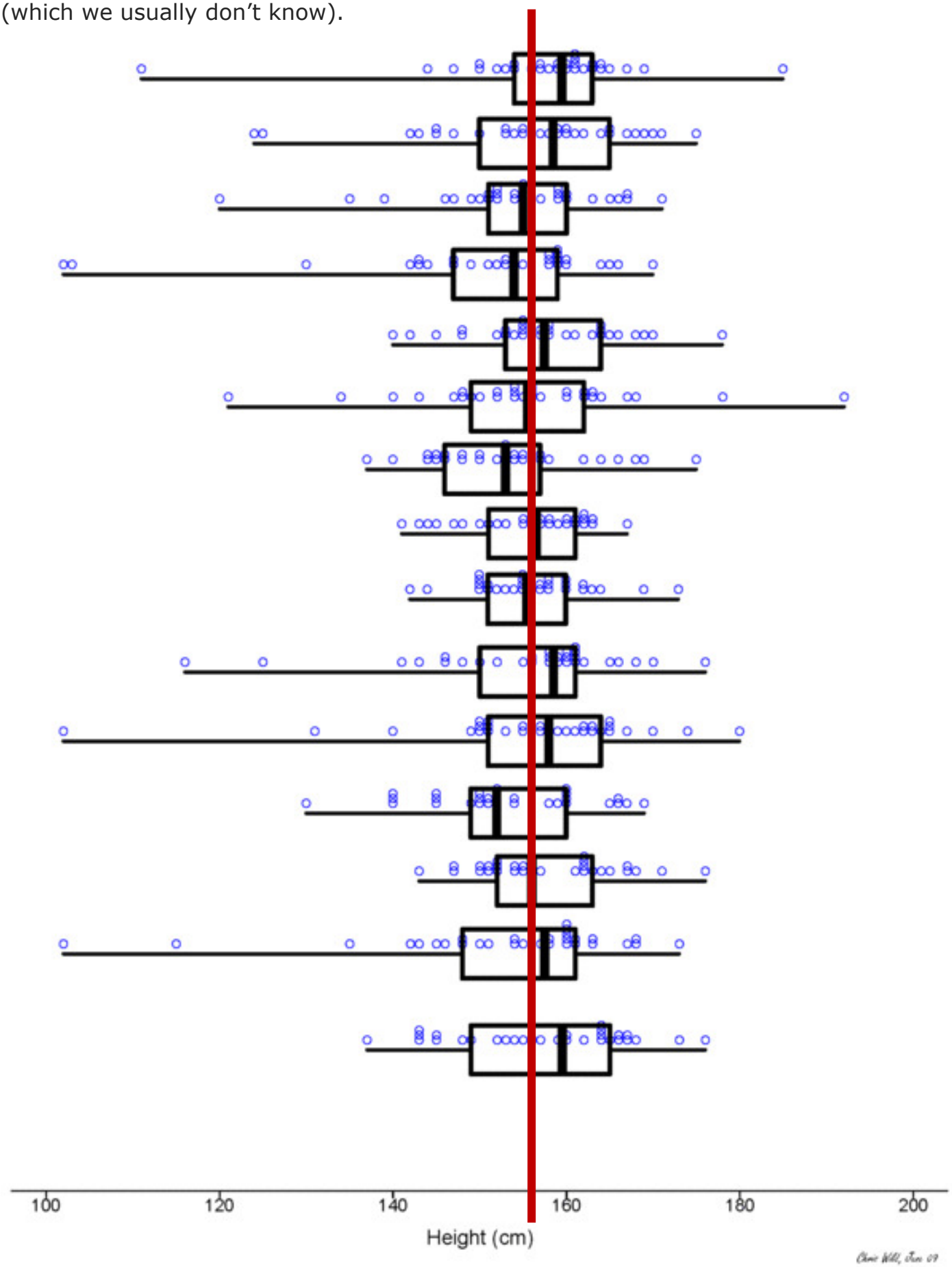
a) Highlight the medians for each of the samples above.

b) Each time we take a sample, would we randomly select the same people each time?

c) Explain why the data is different for each of the samples on the previous page.

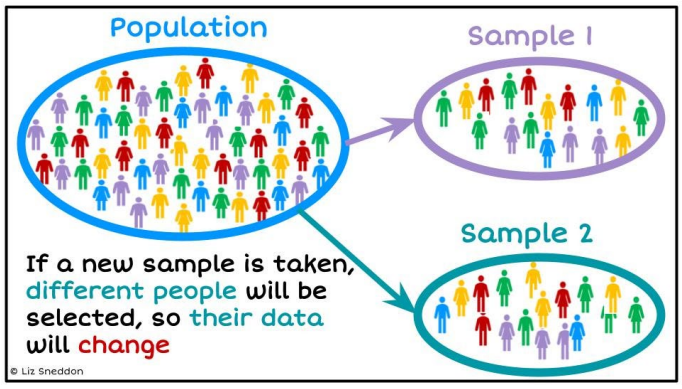
d) Find the smallest and largest estimate for the sample medians. Estimate how much do these values vary by, e.g., 1cm?

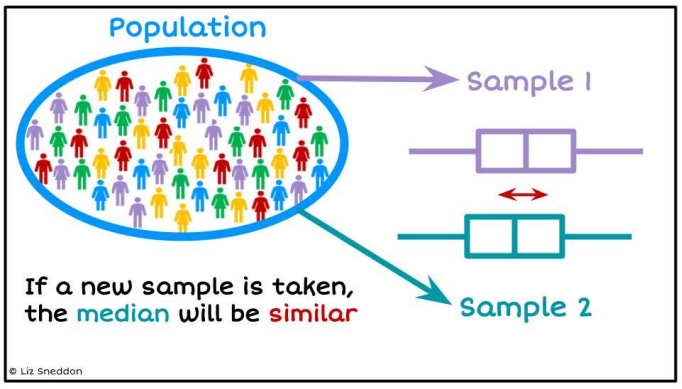
I've added the line drawn from the axis to the top, which is the population median (which we usually don't know).



e) What do you notice about the sample medians compared to the population medians?

This leads us to a big idea, that the **sample median** is **likely** to be **close** to the **population median,** but our estimate is not exact.





Sample size and sampling variation

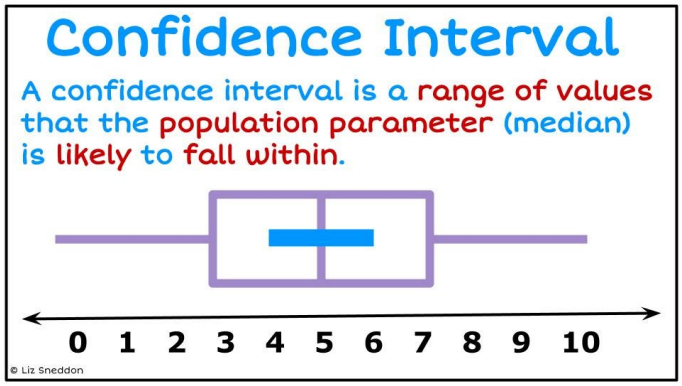
Exercise 20:

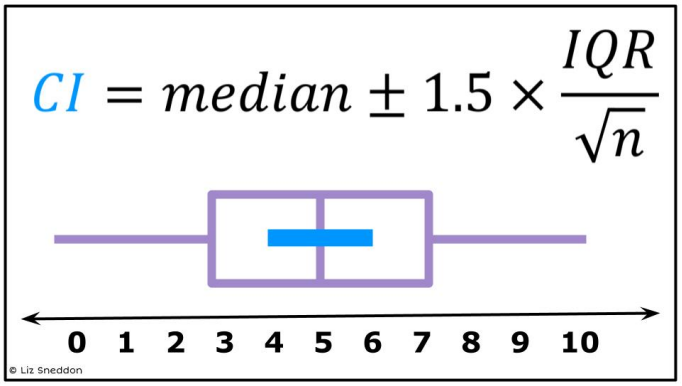
Go to the animation: bit.ly/SamplingVariation

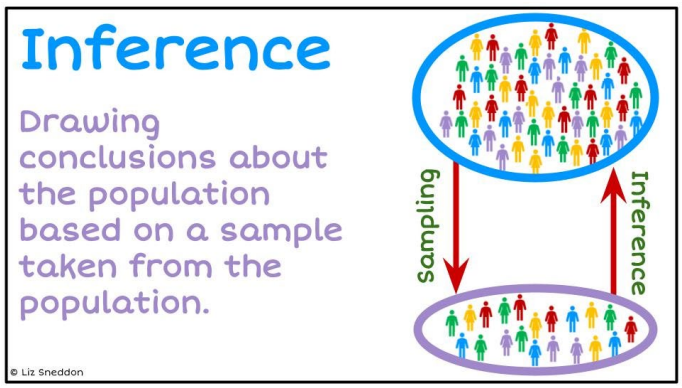
This shows you different samples being taken, with different sample sizes. Look at the blue line which is the median. What do you notice about the amount of variation in the medians with the different sample sizes?

Confidence Intervals

Our next step is to form an interval around our sample median, to try and estimate how much variation there is likely to be in the sample median. To form this interval, we need to consider the sample size, and the sampling variation that is always present when we take a sample.



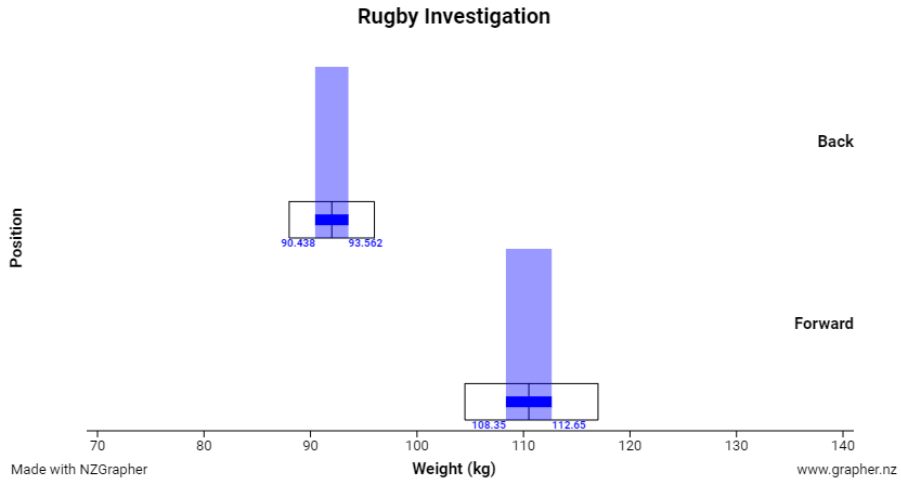




To interpret the confidence interval, we need to remember that we are making an Inference. We are using sample data to estimate the population median.
The confidence interval gives us a range of values that the population median weight of kiwi birds is likely to lie within.

As we are estimating the population median value, remember that we need to round the values.

Example:

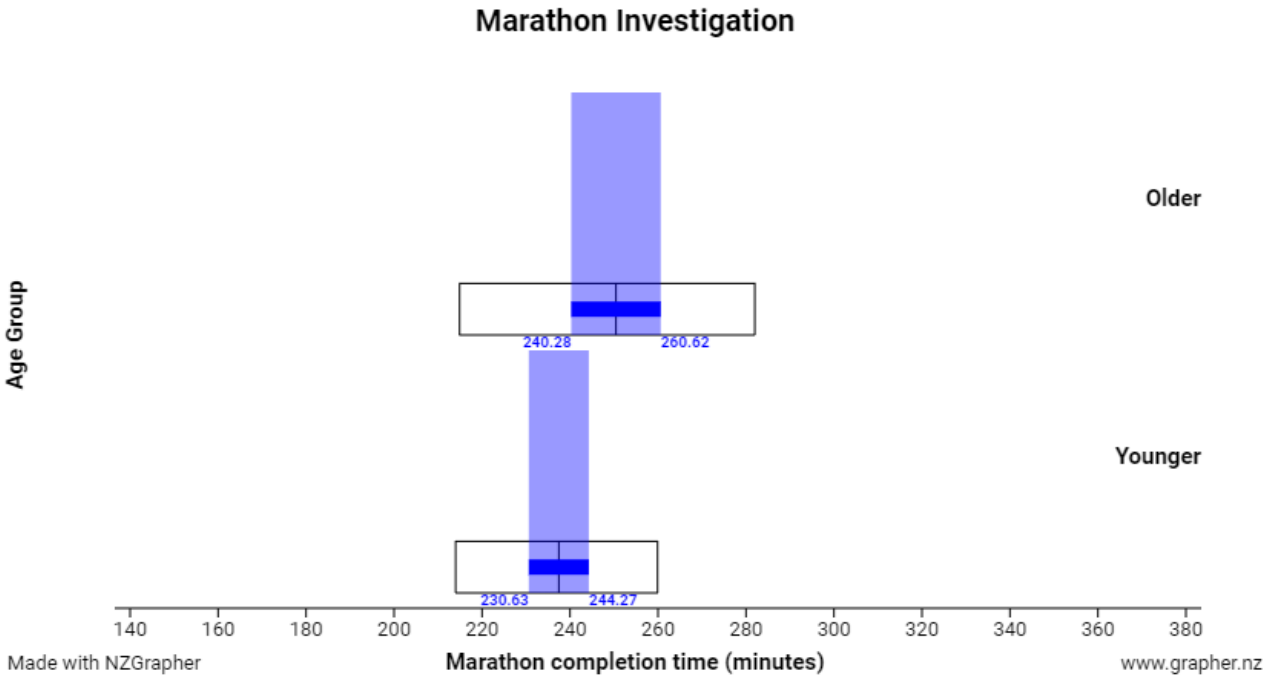


The median weight of rugby players who play in the backs, is likely to be between 90.5kg and 93.6kg, for **ALL** rugby players from the website http://www.rugby-sidestep-central.com/.

The median weight of rugby players who play in the forwards, is likely to be between 108.4kg and 112.7kg, for **ALL** rugby players from the website http://www.rugby-sidestep-central.com/.

Exercise 21:

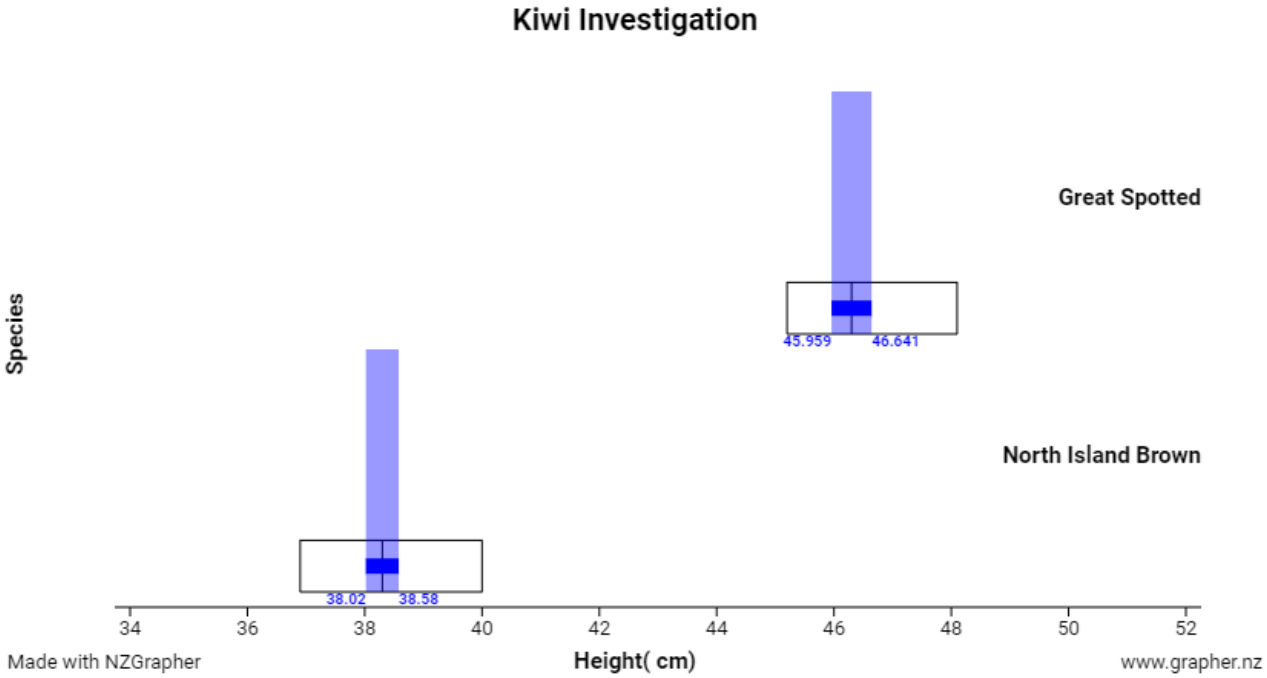
1) This is a sample of marathon runners in NZ.



Confidence interval for marathon completion time for **older** athletes:

Confidence interval for marathon completion time for **younger** athletes:

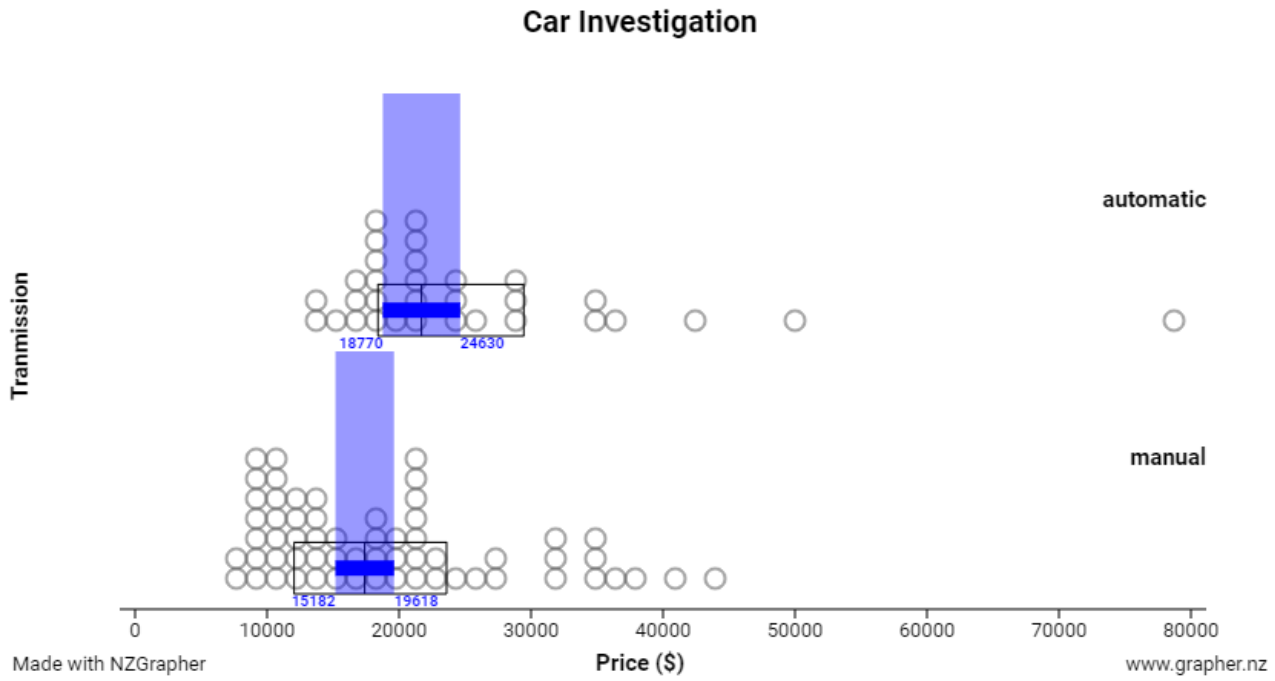
2) Here is a sample of kiwi birds from around NZ.



Confidence interval for marathon completion time for **Great Spotted** kiwis:

Confidence interval for marathon completion time for **North Island Brown** kiwis:

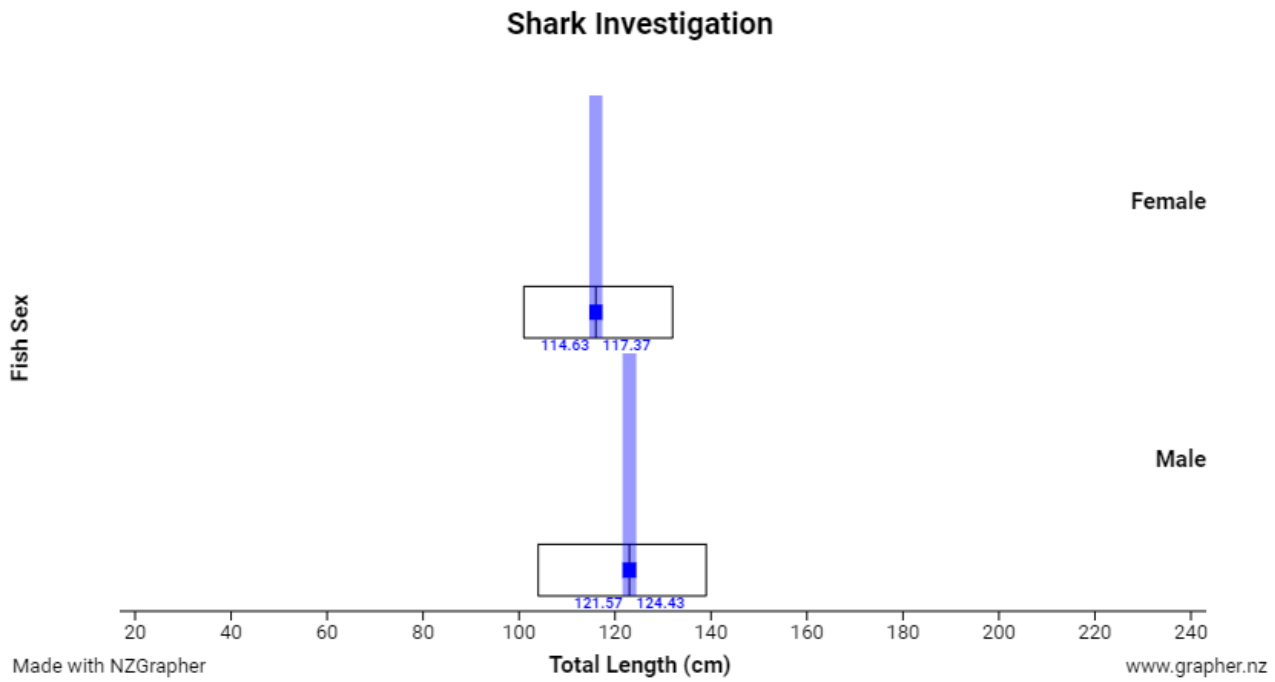
3) Here is a sample of cars sold in America in 1993.



Confidence interval for cost of **Automatic** cars:

Confidence interval for cost of **Manual** cars:

4) Here is a sample of sharks from the MPI (Ministry of Primary Industries – a government organisation) database in NZ.



Confidence interval for **female** sharks on the MPI database:

Confidence interval for **male** sharks on the MPI database:

More on CI's

Confidence intervals use sample data to estimate values that the population parameter (usually the median) is likely to fall between. The formula to calculate the limits (shown below) takes into account both the amount of variation in the sample (the ***IQR***), as well as the size of the sample (***n***).

$$CI = median \pm margin\ of\ error$$

$$= median \pm 1.5 \times \frac{IQR}{\sqrt{n}}$$

Example:

Calculate the confidence interval if the IQR = 1, n = 10 and median = 10.

$$CI = median \pm margin\ of\ error$$

$$= median \pm 1.5 \times \frac{IQR}{\sqrt{n}}$$

$$= 10 \pm 1.5 \times \frac{1}{\sqrt{10}}$$

$$= 10 \pm 0.4743$$

$$= (10 - 0.4743,\ \ 10 + 0.4743)$$

$$= (9.5257,\ \ 10.4743)$$

Notice that I separated out the $\pm$ into 2 calculations, and when I wrote the answer in brackets, I put the smaller number on the left, and the larger number on the right.

This is set notation: (small, large)

Exercise 22:

Consider what happens when the sample size (***n***) increases.

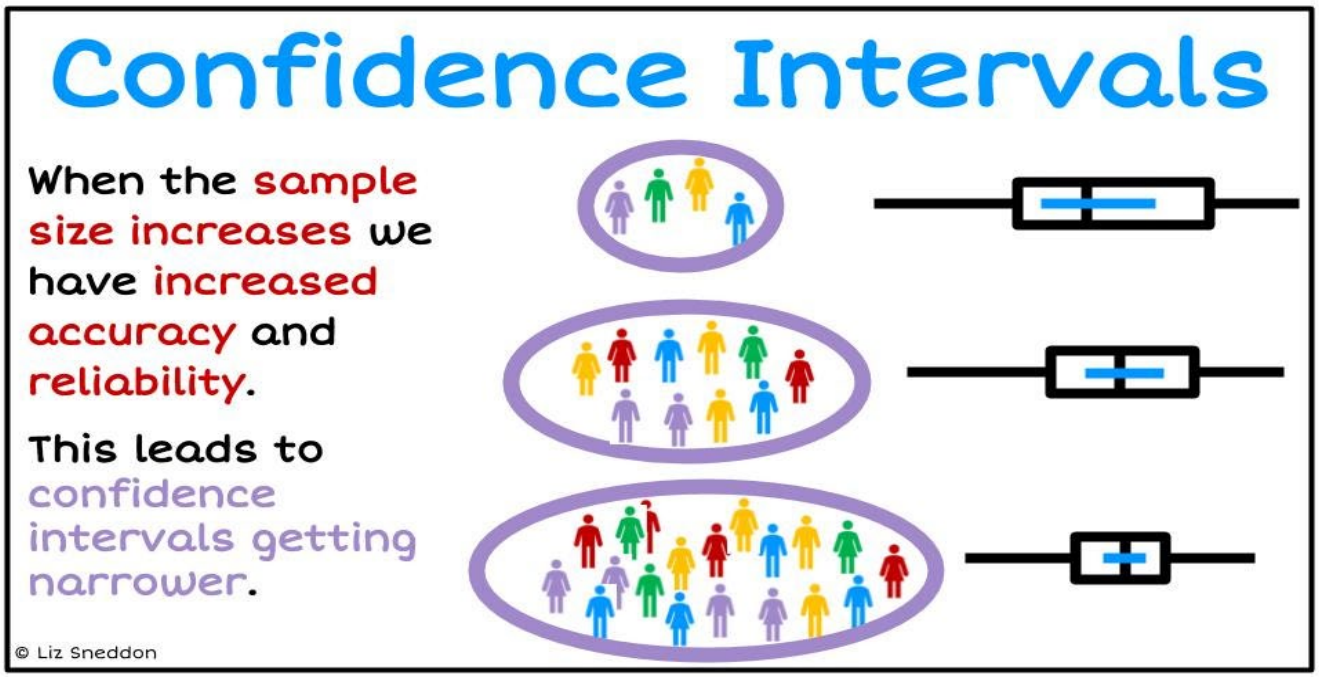
Using the formulae above calculate the margin of error, the confidence interval, and the width of the confidence interval to 4.d.p.

Assume IQR = 1 and median = 10.

Sample size, *n*	**Margin of error** $\mathbf{1.5 \times \frac{1}{\sqrt{n}}}$	**Confidence Interval, (Limit 1, Limit 2)**	**Width of CI, Limit 2 - Limit 1**
10	$1.5 \times \frac{1}{\sqrt{10}} =$		
100	$1.5 \times \frac{1}{\sqrt{100}} =$		
1000	$1.5 \times \frac{1}{\sqrt{1000}} =$		
10,000	$1.5 \times \frac{1}{\sqrt{10{,}000}} =$		
100,000	$1.5 \times \frac{1}{\sqrt{100{,}000}} =$		

a) Explain what happens to the margin of error as the sample size increases.

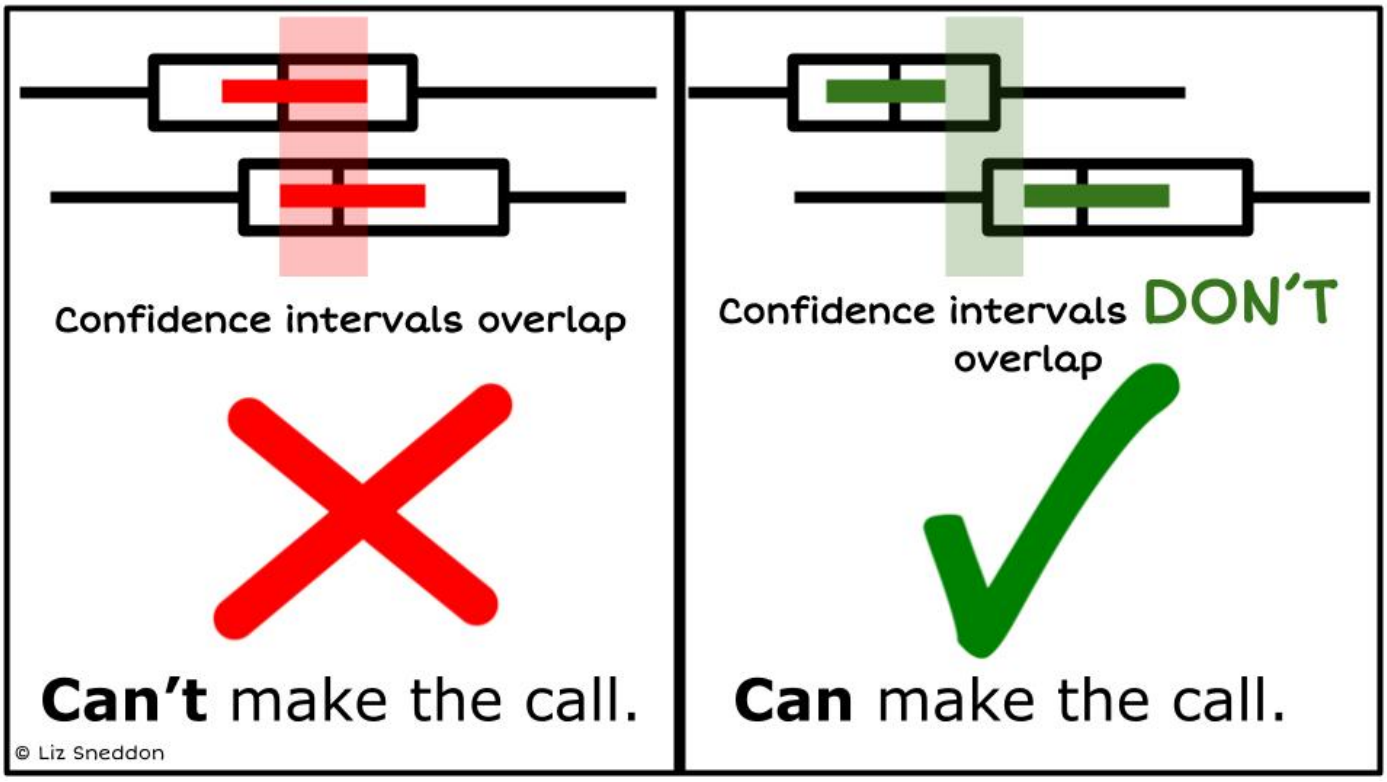
b) Explain what happens to the width of the confidence interval as the sample size increases.



Answering the investigation question

Because we have a sample, and don't know the whole population, we need to decide if we can make the call.

Look at the confidence intervals. If they overlap, you **CAN'T** make the call. If they **DON'T** overlap, you **CAN** make the call.



To answer the comparison question you need:

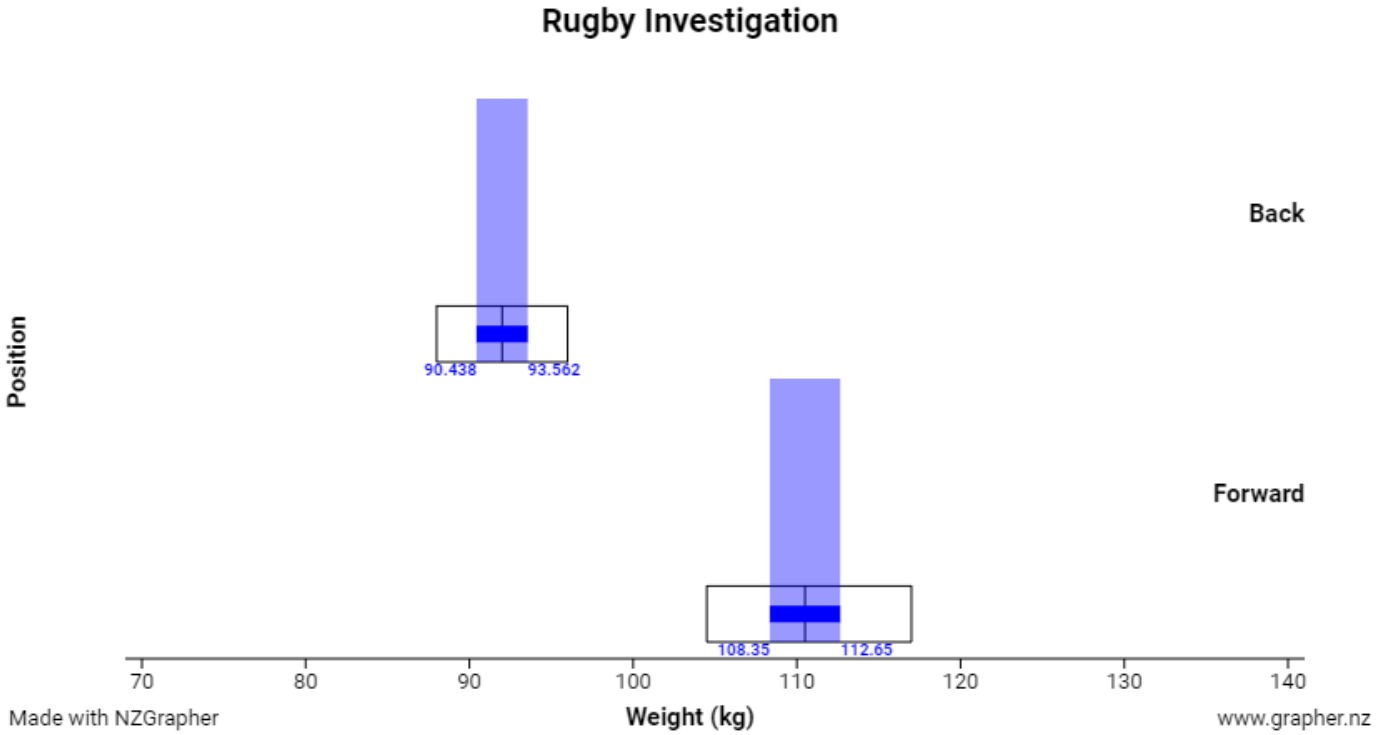
- Categorical variable,
- Numerical variable,
- The word "median",
- Direction,
- Population (ALL).

© Liz Sneddon

Example:

Problem:

I wonder if median weight of rugby players in the back positions tends to be lighter than the median weight of players in the forwards positions, for ALL rugby players on the website http://www.rugby-sidestep-central.com/.



Conclusion:

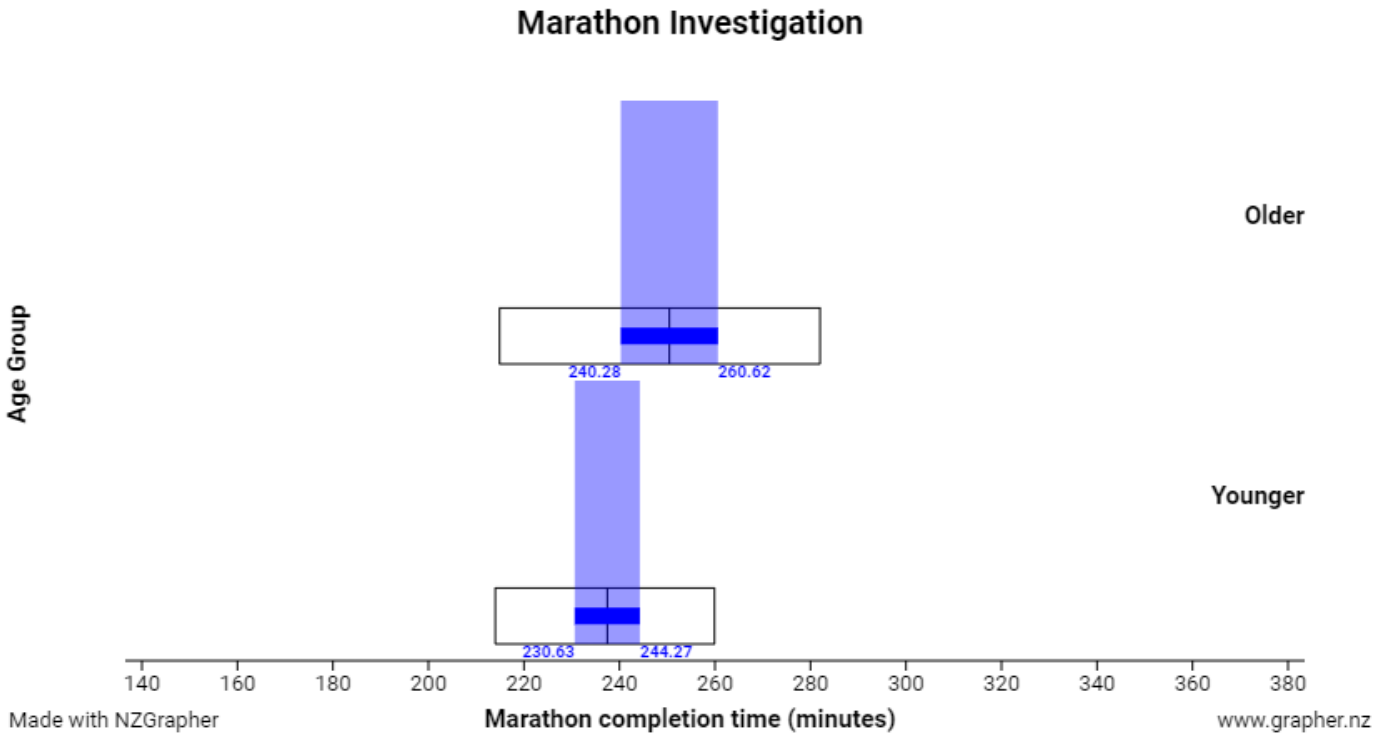
The confidence interval for weights of rugby players who are playing in the back positions **DOESN'T** overlap with the confidence interval for weights of rugby players who play in forwards positions.

(This means that my **sample suggests** that back in the population, the medians aren't likely to be the same.)

I can make the call, so I **DO have enough evidence** that the median weight of rugby players in the back position is **lighter** than the median weight of rugby players who play in forwards positions, for **ALL** rugby players on the website http://www.rugby-sidestep-central.com/.

Exercise 23:

1)



Problem:

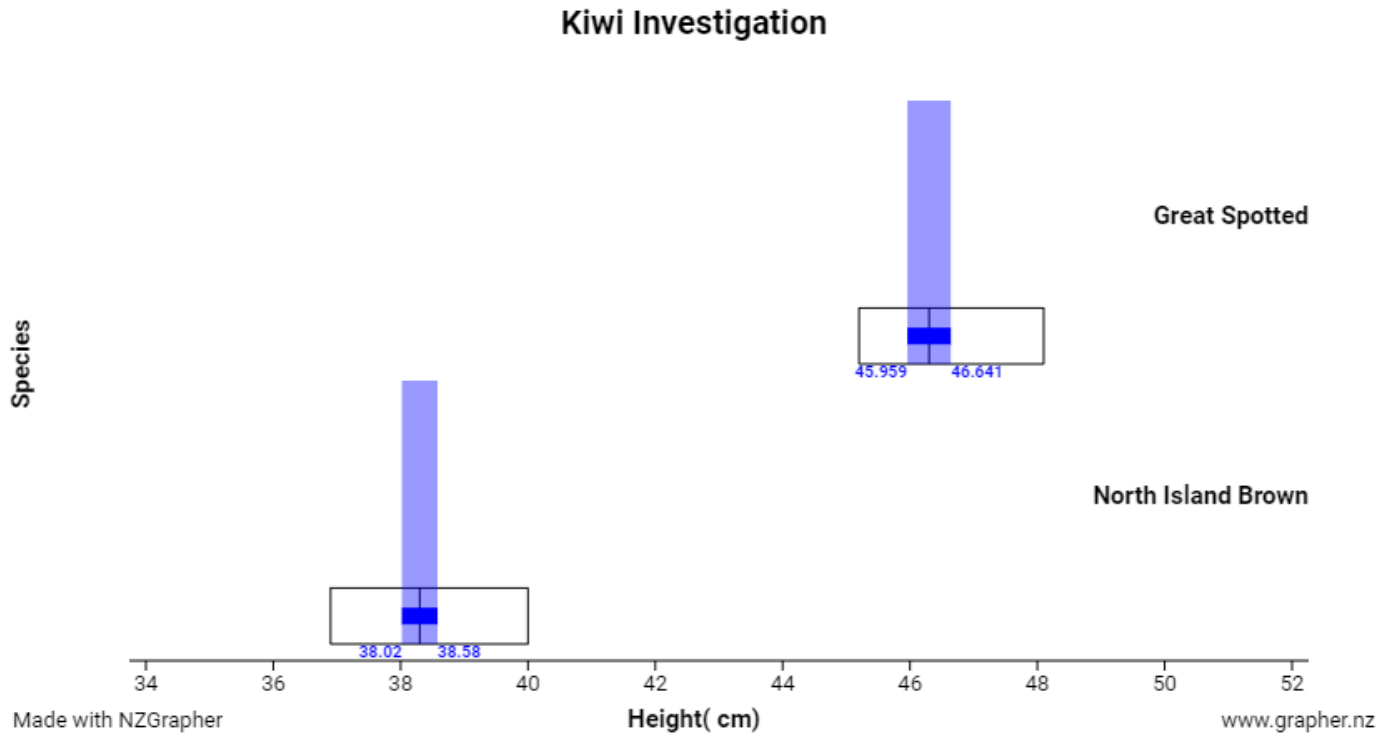
I wonder if the median time it takes older people to complete a marathon is longer than the median time it takes younger people to complete a marathon, for all marathon runners in NZ?

Conclusion:

Do the confidence intervals overlap?	Yes / No
Can you make the call?	Yes / No

I _____ have enough evidence that the median time it takes older people to complete a marathon is longer than the median time it takes younger people to complete a marathon, for **ALL** marathon runners in NZ.

2)



Problem:

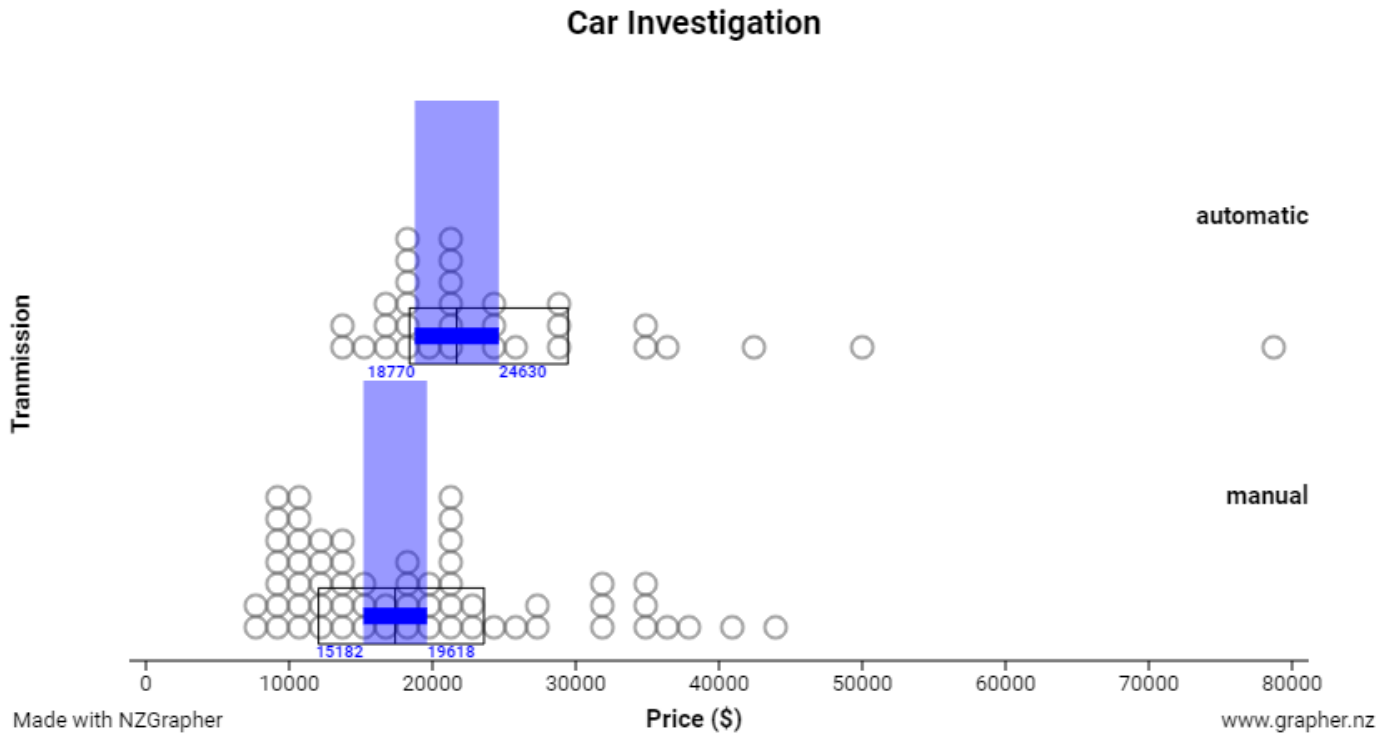
I wonder if the median height for Great Spotted kiwis is taller than the median height for North Island Brown kiwis, for ALL kiwis in NZ?

Conclusion:

Do the confidence intervals overlap?	Yes / No
Can you make the call?	Yes / No

I _____ have enough evidence that the median height for Great Spotted kiwis is taller than the median height for North Island Brown kiwis, for ALL kiwis in NZ.

3)



Problem:

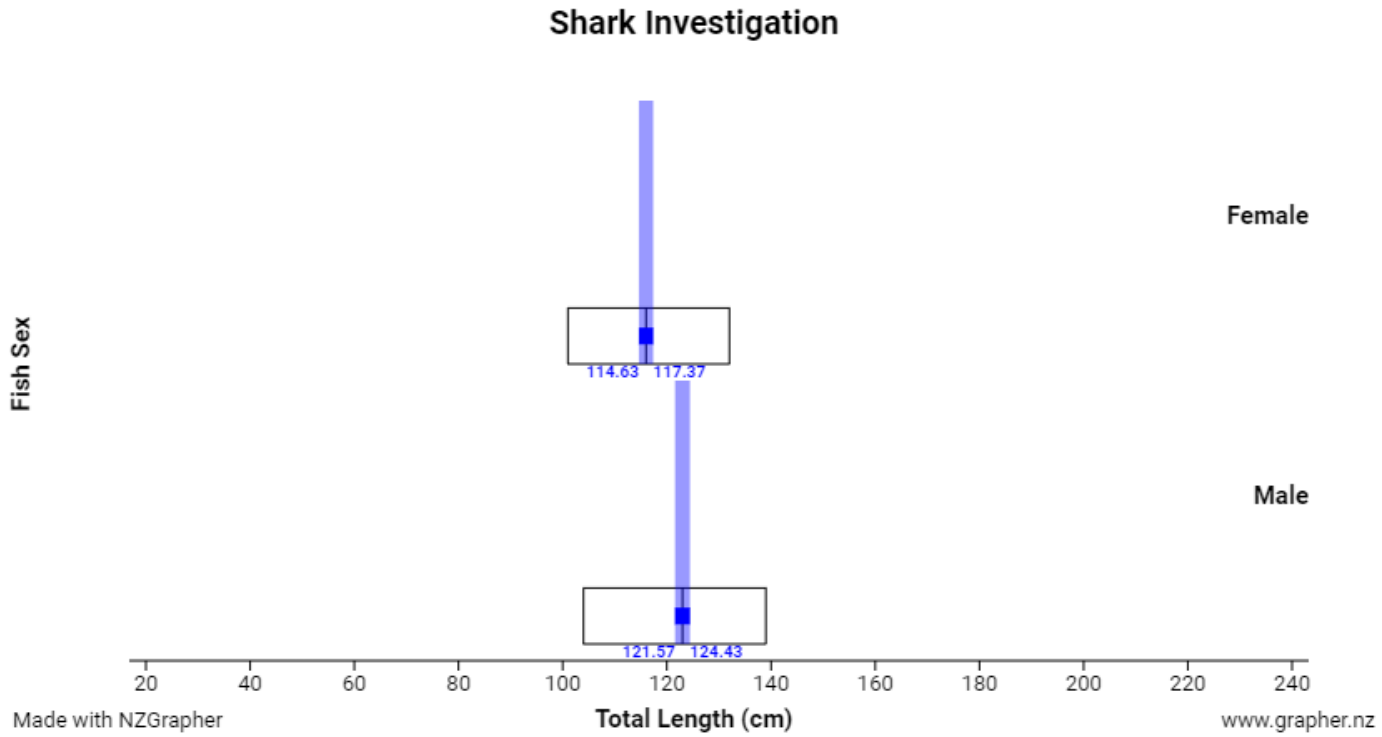
I wonder if the median cost of automatic cars tends to be more than the median cost of manual cars, for ALL cars sold in America in 1993?

Conclusion:

Do the confidence intervals overlap?	Yes / No
Can you make the call?	Yes / No

Answer the investigation question:

4)



Problem:

I wonder if the median length of female sharks tends to be shorter than the median length of male sharks, for ALL sharks on the MPI database in NZ?

Conclusion:

Do the confidence intervals overlap?	Yes / No
Can you make the call?	Yes / No

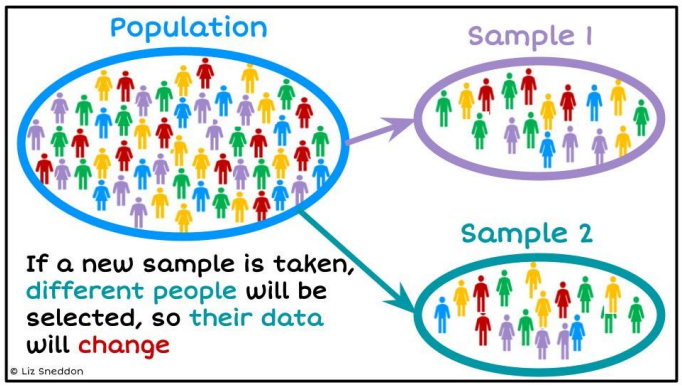
Answer the investigation question:

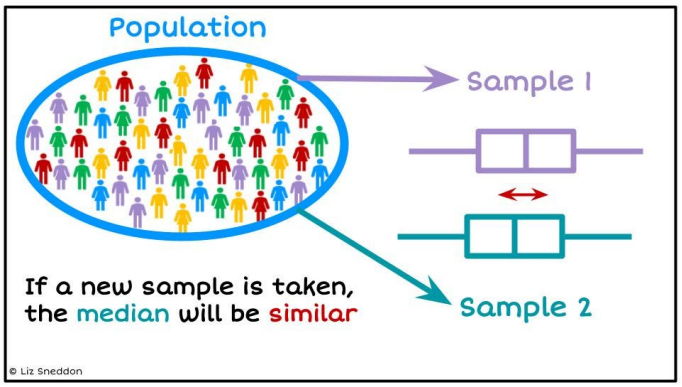
Sampling variability

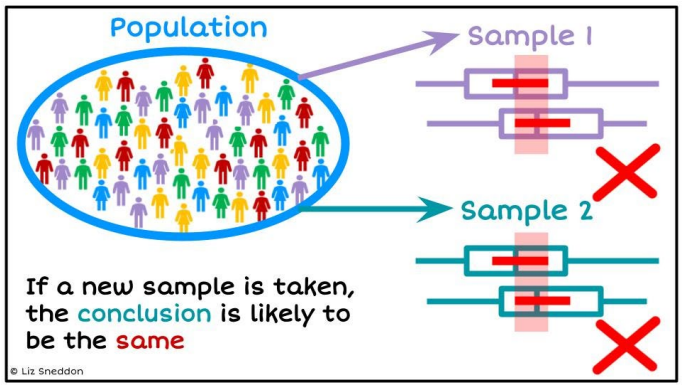
This describes the variation (differences) that happens when you take another sample.

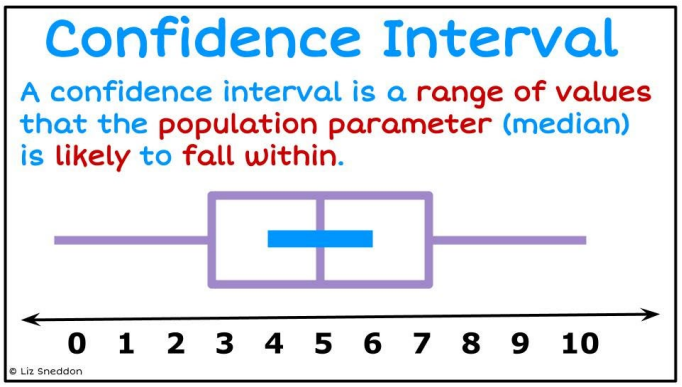
You need to consider 4 things:

- how would the **data** change?
- how would the **medians** change?
- how would your **conclusion** change?
- how likely is it that the **population median** is inside your confidence interval?









Exercise 24:

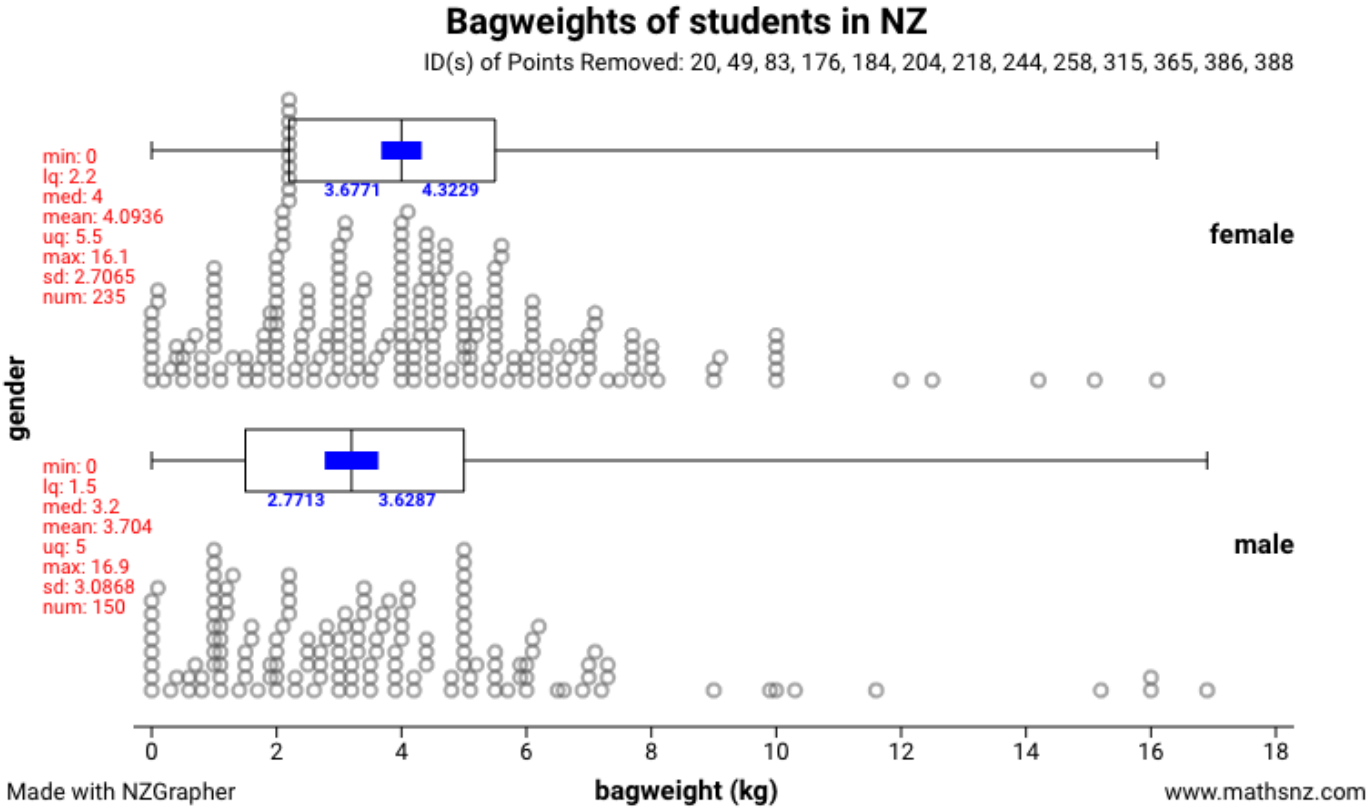
1) Complete the following sentence:

If I took another sample …

Full Example:

Problem:

Is the median weight of girls' school bags greater than the median weight of boys' school bags, for ALL students at Intermediate schools in NZ?



Conclusion:

Interpreting the confidence intervals.

I estimate that the median weight of a school bag for **ALL** female Intermediate students in NZ, is likely to be between 3.68kg and 4.32kg.

I estimate that the median weight of a school bag for **ALL** male Intermediate students in NZ, is likely to be between 2.77kg and 3.63kg.

Answering the investigation question.

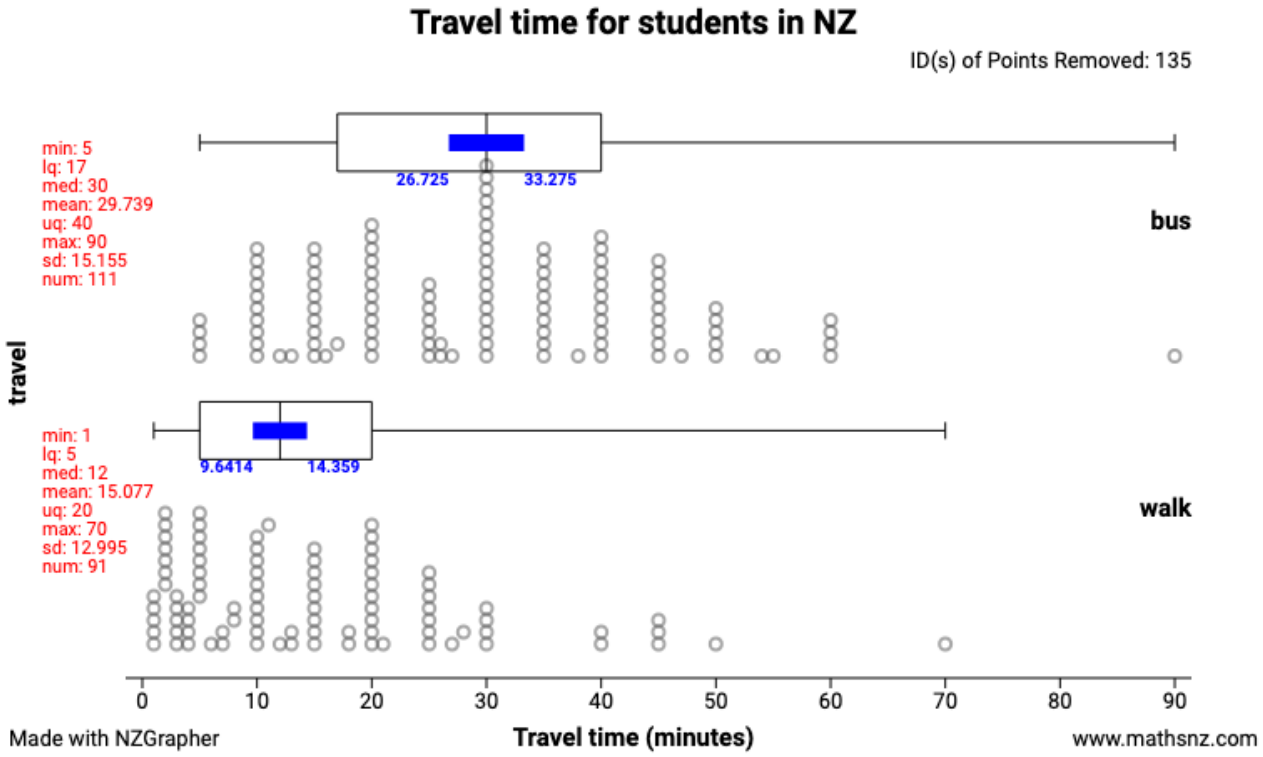
The confidence intervals do NOT overlap, which means that I can make the call. I have enough evidence that the median weight of girls' school bags is greater than the median weight of boys' school bags, for ALL students at Intermediate schools in NZ?

Sampling variability.

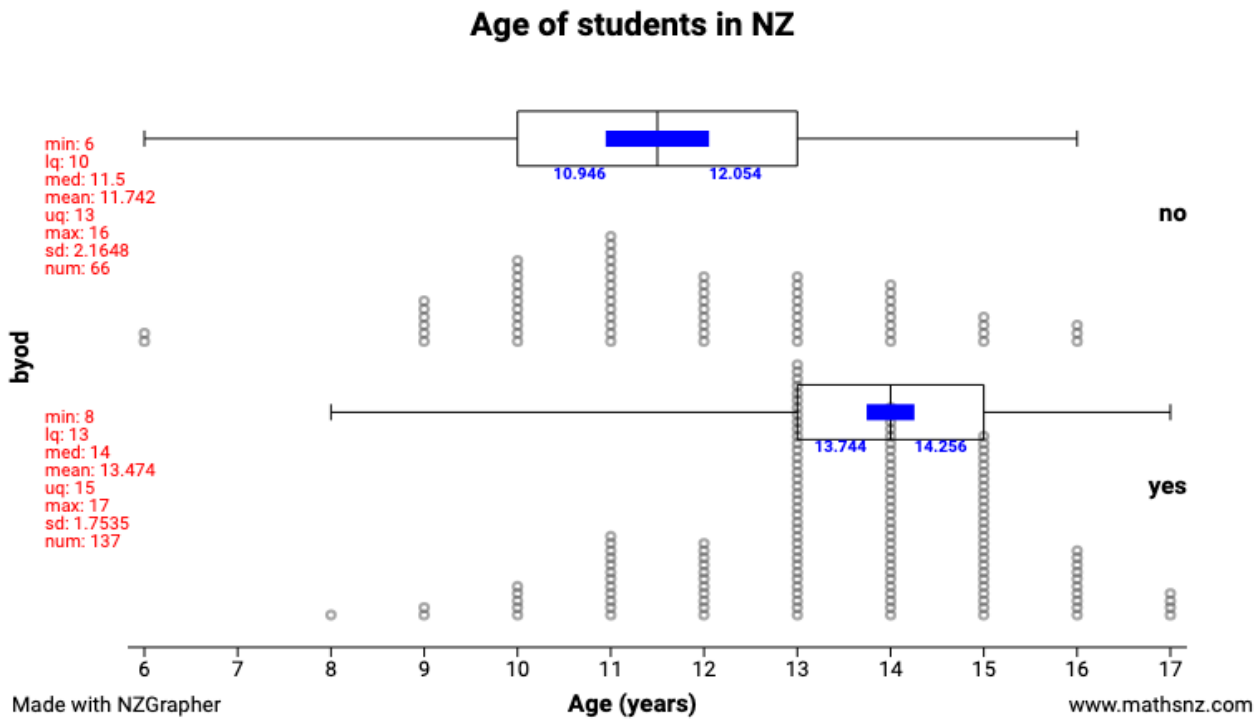
If I took another sample, I would get different weights of school bags, as I would be collecting data from different students. I would expect though that the median weights of school bags for boys and girls at Intermediate schools in NZ is likely to be similar to the values I found in my sample. I also expect that my conclusion, that the median weights of school bags for girls is heavier than the median weight of school bags for boys, for **ALL** students at Intermediate schools in NZ, to remain the same.

Exercise 25:

1) For high school students in New Zealand, is the median travel time (minutes) for **ALL** students who take the bus to school longer than the median travel time than for **ALL** high school students who walk to school?



2) For **ALL** high school students in New Zealand, is the median age for students who have a device older than the median age for those who do not have a device?



Other Factors

Think about making a puzzle. If you have one piece of the puzzle, you can learn a lot - the colours, the shapes, and guess where it might go in the big picture. But you know that this is only one piece of the puzzle, and that there are lots of other pieces that when put together make a whole different picture.

Data is like a puzzle. You might investigate one particular piece, one particular comparison, but we always need to be aware that there are many other pieces in this puzzle. Before we can paint the whole picture, we need to think about what the other pieces might be.

This is the idea of other factors. You need to consider what other factors might also affect the numerical variable that you are interested in and use research to explain **WHY** this factor might be affecting it.

Example:

Problem:
I wonder if the median weight of teenagers is heavier than the median weight of young children, for all children in NZ.

Other variables or factors:
There are a number of different factors that might affect the weight of children in NZ. For example, if a child has parents who are both slim and short in stature, then because of the genetic link it is likely that the child is also likely to be slim and short in stature. Equally, a child whose parents have bigger and heavier bones, and a wide/tall build are likely to be taller and heavier. A study published in the UK[7] supports this and discusses how there is a link between people's weight and their genetics, where being slim is a heritable trait.

Another factor that could affect the weight of children is the amount of exercise they do each week. I expect that a child who is more active and spends more time each week exercising would have less body fat than a child who is less active and spends less time exercising each week. The more body fat a child has the higher their weight will be. Kids Health[8] suggest that "kids can reach a healthy weight by eating right and being active".

[7] https://www.healthline.com/health-news/heres-how-much-your-genes-impact-your-ability-to-lose-weight
[8] https://kidshealth.org/en/parents/childs-weight.html

Exercise 26:

Discuss other factors that may affect the numerical variable and support with research for the situations below.

1) **Problem:**
I wonder if the median time that teenagers spend on phone calls each day is more than the median time that adults spend on their own phone calls each day, for all people in NZ.

Other factors:
What other factors do you think would affect the time that teenagers in NZ spend talking on phone calls each day? Use research to explain why and how this factor would affect the time on phone calls.

2) **Problem:**

I wonder if the median lung capacity of teenagers who vape is smaller than the median lung capacity of teenagers who don't vape, for ALL teenagers in NZ.

Other factors:

What other factors do you think would affect the lung capacity of teenagers? Use research to explain why and how this factor would affect the lung capacity.

3) **Problem:**

I wonder if the median house price for 2-bedroom homes tends to sell for higher prices than the median house price of 4-bedroom homes, for ALL homes in NZ.

Other factors:

What other factors do you think would affect house prices in NZ? Use research to explain why and how this factor would affect house prices.