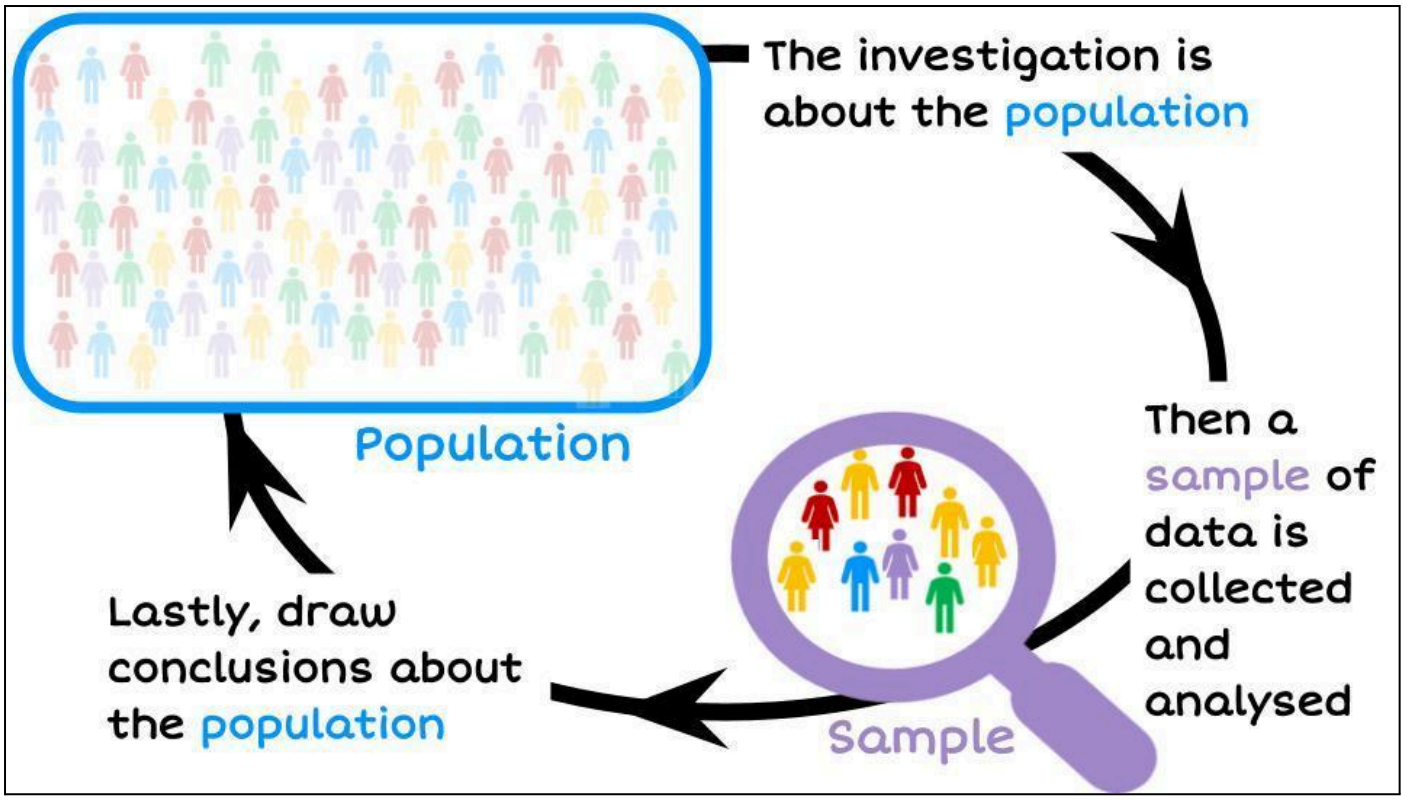
Level 3 Inference Workbook

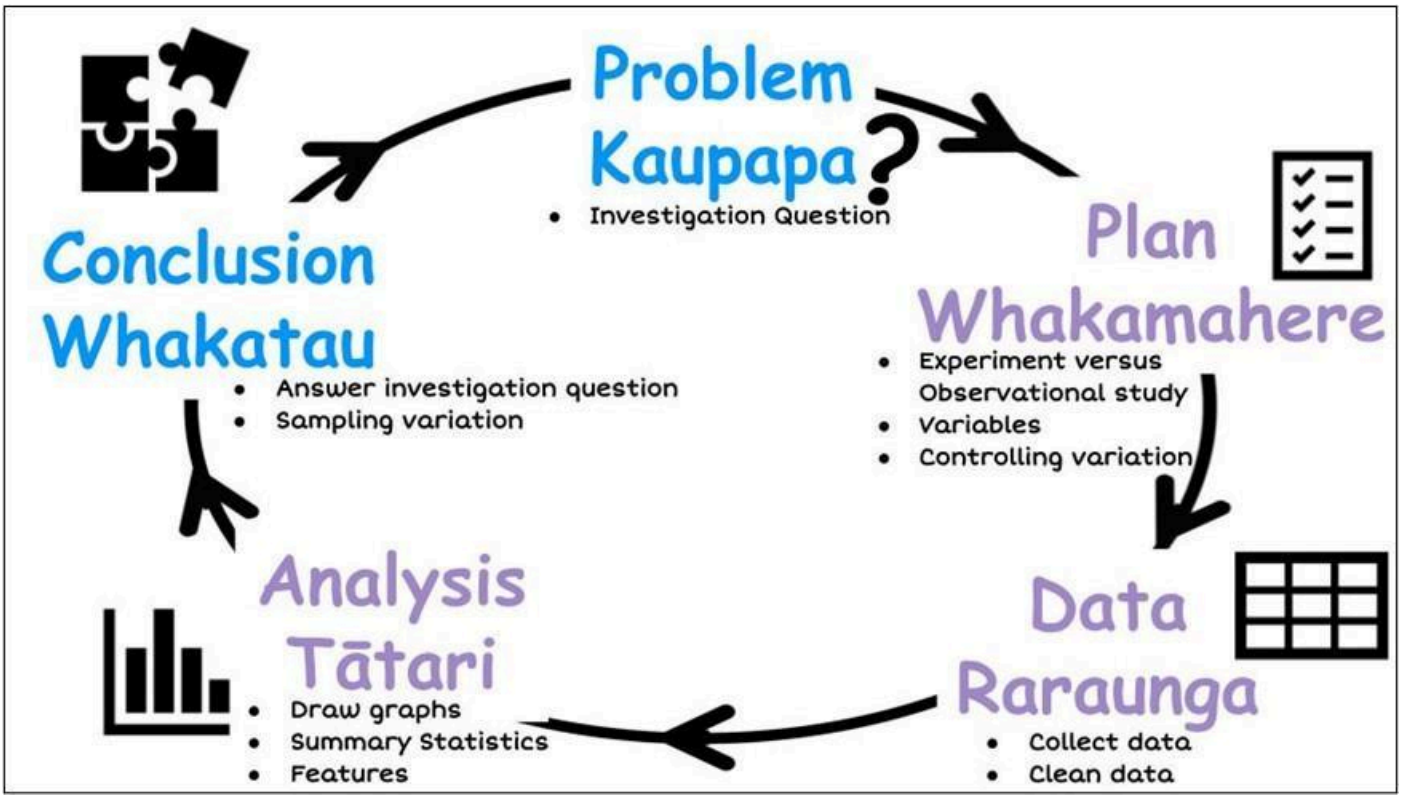
Name:

By Liz Sneddon

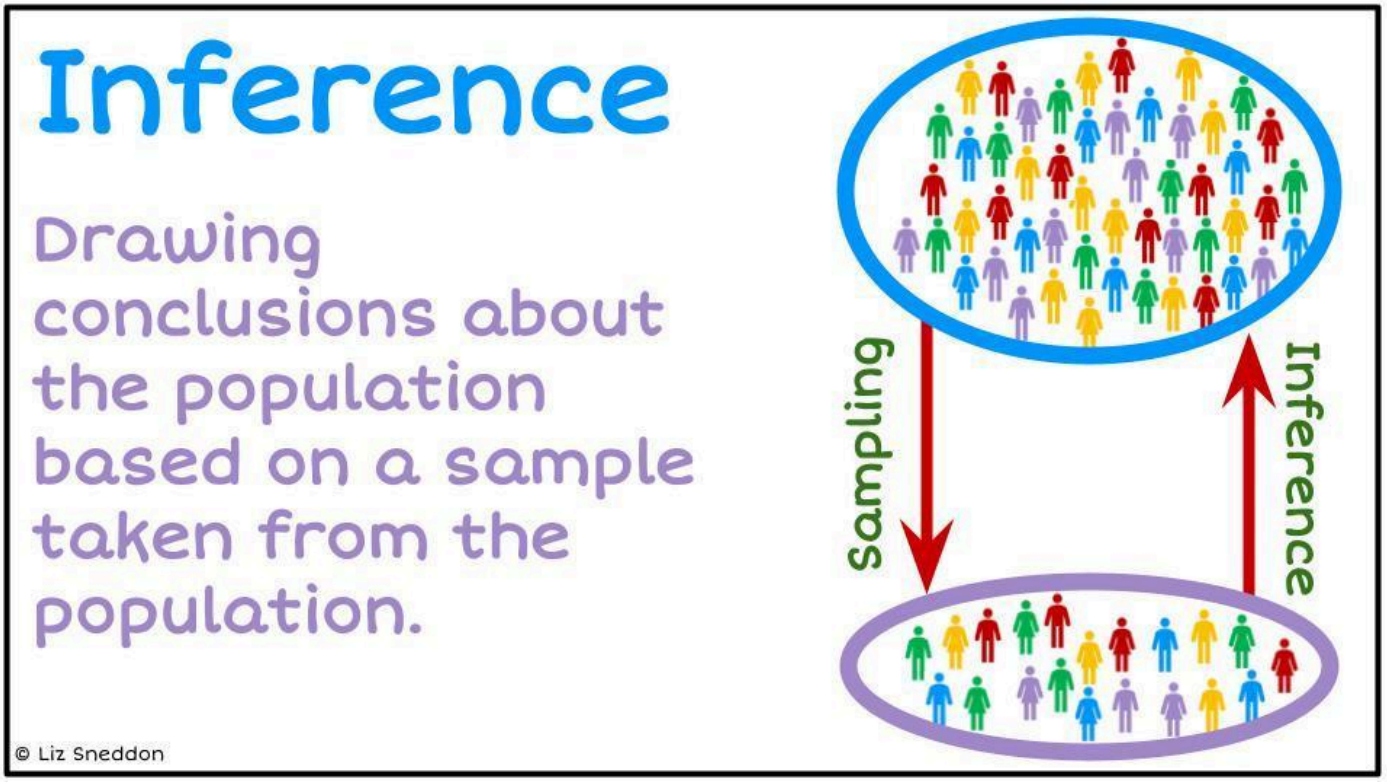
Investigative Process



The PPDAC cycle is the core of all statistical investigations.



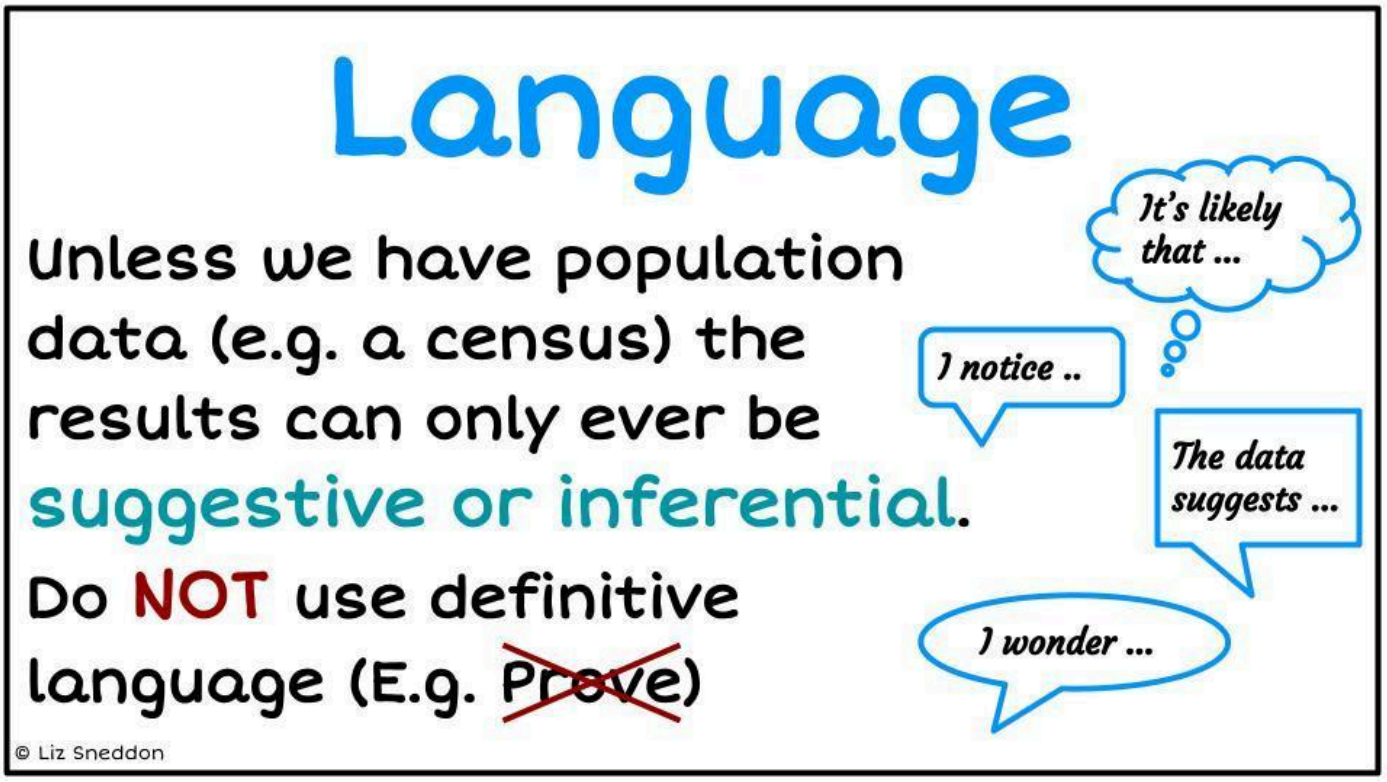
At the end of this investigation, you will need to make an Inference.



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.

Language



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*.

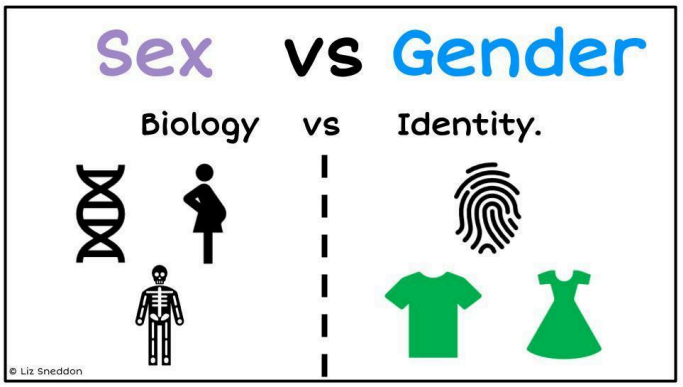
As you can see, there are two different definitions of the word average. Therefore, in order to avoid confusion, **please do not use the word average**. Instead use the specific terminology of **mean, median or mode** as appropriate.

Gender versus Sex for data collection

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

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 1:

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

A comparison question needs:

- Categorical variable,
- Numerical variable,
- The words "difference between median",
- Population (ALL).

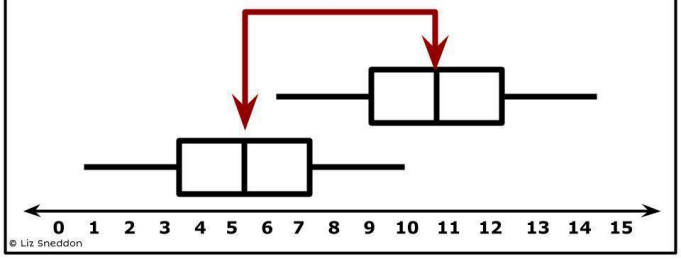
© Liz Sneddon

FORMAL investigation question

This means that we DON'T have a direction, but instead write a question asking WHAT the difference between the medians is.

© Liz Sneddon

Focus on the difference between the medians



© Liz Sneddon

Population: all the individual members or items that make up a group.

© Liz Sneddon

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.
- 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 of **ONE** population.
- 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.
- The step up from Level 2 Statistics is that instead of just focusing on which median is bigger or smaller, we now accept that there will be some difference between the medians, so at Level 3 the question now shifts to **what the difference** is and we will analyse **the size of the difference**. This means that the investigation question is always "what is the difference between the medians …".

Example:

What is the difference between the median weight (kg) of forward and back rugby players from ALL top players listed on the website http://www.rugby-sidestep-central.com.

Exercise 2:

1) In the example given above, identify the following:

Categorical variable	
Two groups being compared	
Numerical variable and units	
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	
Difference between median	Yes / No
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	
Difference between median	Yes / No
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	
Difference between median	Yes / No
Population definition	
Rewritten question:	

d) How does the number of text messages all teenage girls send daily compare with the number of text messages all teenage boys send daily, in Auckland?

Categorical variable	
Numerical variable and units	
Difference between median	Yes / No
Population definition	
Rewritten question:	

3) The dataset described below is a sample taken from marathons in NZ. It is a simple random sample of 200 athletes.

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.

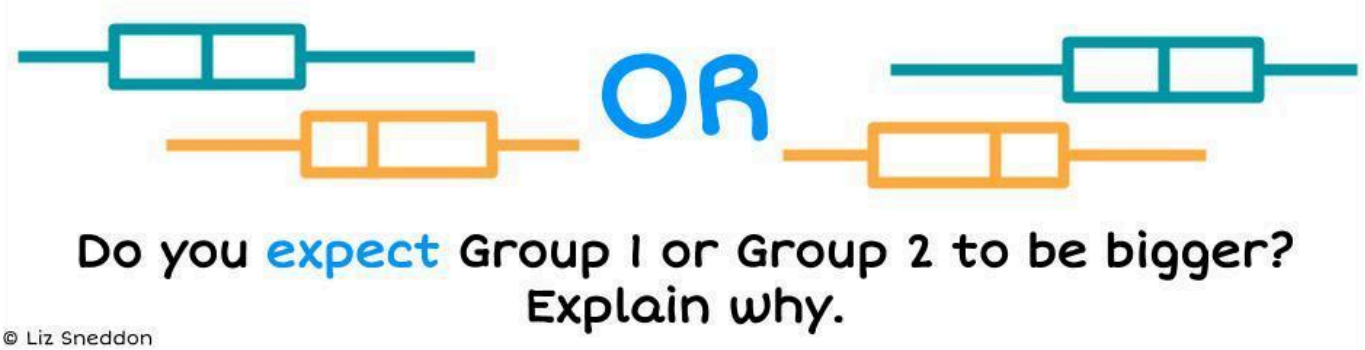
4) The dataset described below is a sample of rugby data taken from http://www.rugby-sidestep-central.com/

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)

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

Hypothesis & research

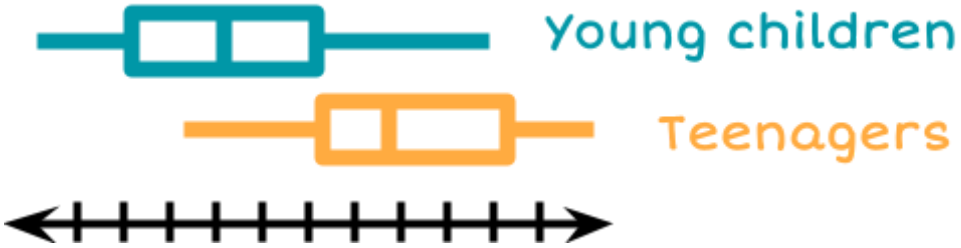
In the **Problem** section, you want to add a hypothesis about what you might see **BEFORE** you explore the data. You then want to use **research** to support this hypothesis and explain **WHY** there may be a difference between the two groups that you are comparing. Later in both your **Analysis** 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

Exercise 3:

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

1) **Problem:**
What is the difference between the median time that teenagers spend on phone calls each day compared to adults (18 years and older), 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:**

What is the difference between the median lung capacity of teenagers who vape and those that don't, for ALL teenagers in NZ.

Hypothesis:
Draw the graph you expect to see.

Use research to explain why you expect to find this.

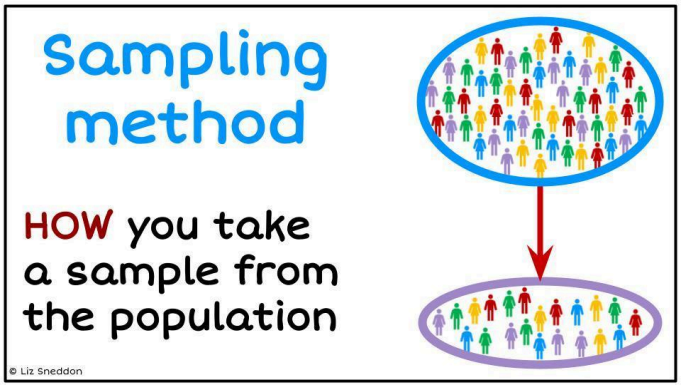
3) I wonder if the median house price for a 2-bedroom home tends to sell for higher prices than the median house price of 4-bedroom homes, for ALL homes in NZ.

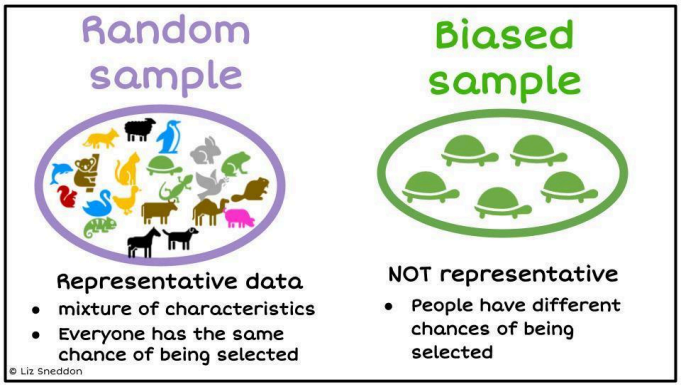
Hypothesis:
Draw the graph you expect to see.

Use research to explain why you expect to find this.

Plan

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**.



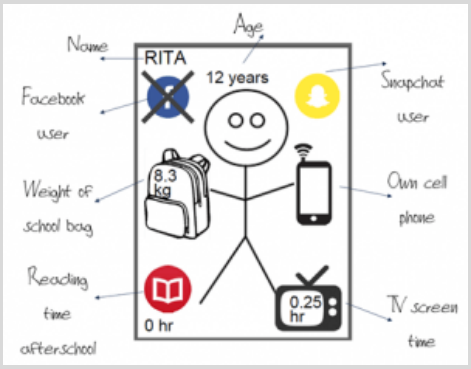


Example

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 4:

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)**

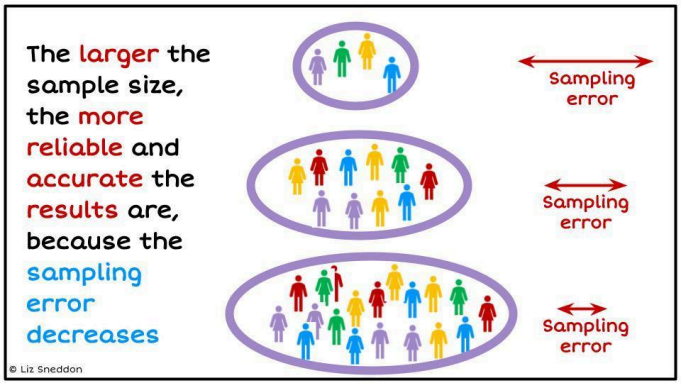
[4] Created by Anna Ferguson: http://teaching.statistics-is-awesome.org/census-at-school-stick-people-data-cards/

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 5:

1) Circle the words that complete the sentences below.

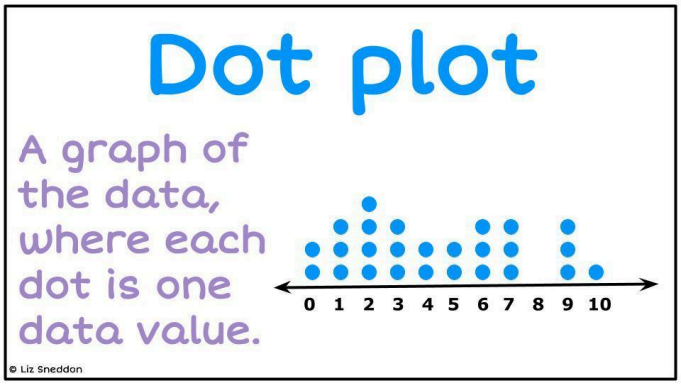
a) Smaller sample sizes take a **shorter / longer** time to collect data, but results are **more / less** precise.

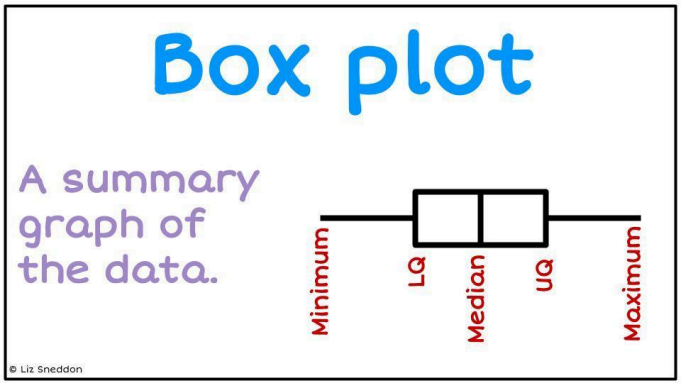
b) Larger sample sizes take a **shorter / longer** time to collect data and results are **more / less** precise.

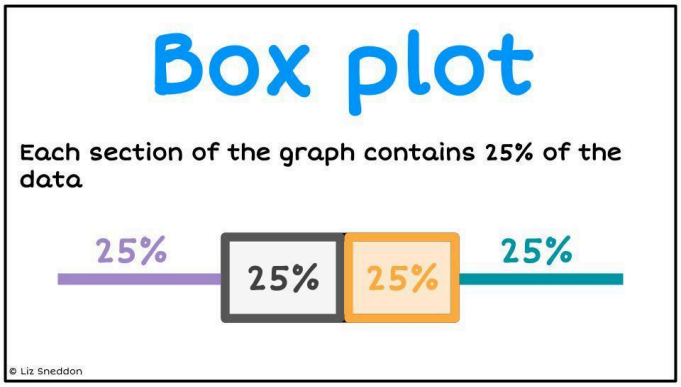
2) Mrs Sneddon is going to survey 35 girls and 40 boys to investigate their use of iPads at home. Explain why it is ok for the sample sizes to be different, and how these sample sizes affect the variation.

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.



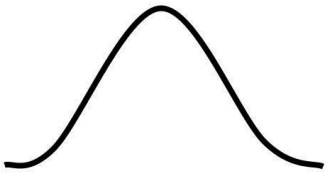




Summary Statistics

Measures of centre

Centre



- Mean
- Median
- Mode

© Liz Sneddon

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

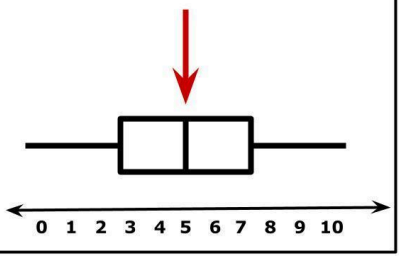
Median = middle value

Mode = most frequent value

© Liz Sneddon

Median

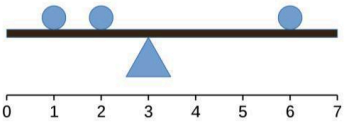
A value where **50%** of the data lies **above** & **below** it.



© Liz Sneddon

Mean

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



© Liz Sneddon

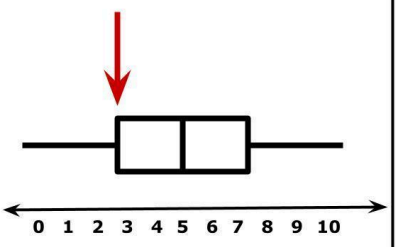
For numerical data, the mean and median are the preferred measures of centre.

For categorical data, the mode is the preferred measure of centre.

Measures of spread

Lower Quartile (LQ)

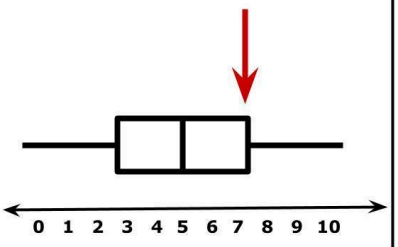
A value where **25%** of the data lies **below** it.



© Liz Sneddon

Upper Quartile (UQ)

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

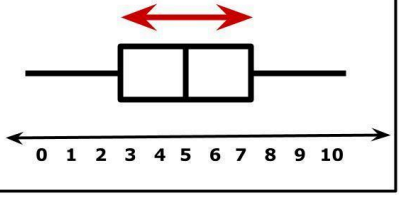


© Liz Sneddon

Interquartile range

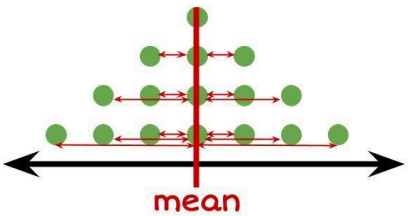
IQR = UQ - LQ

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



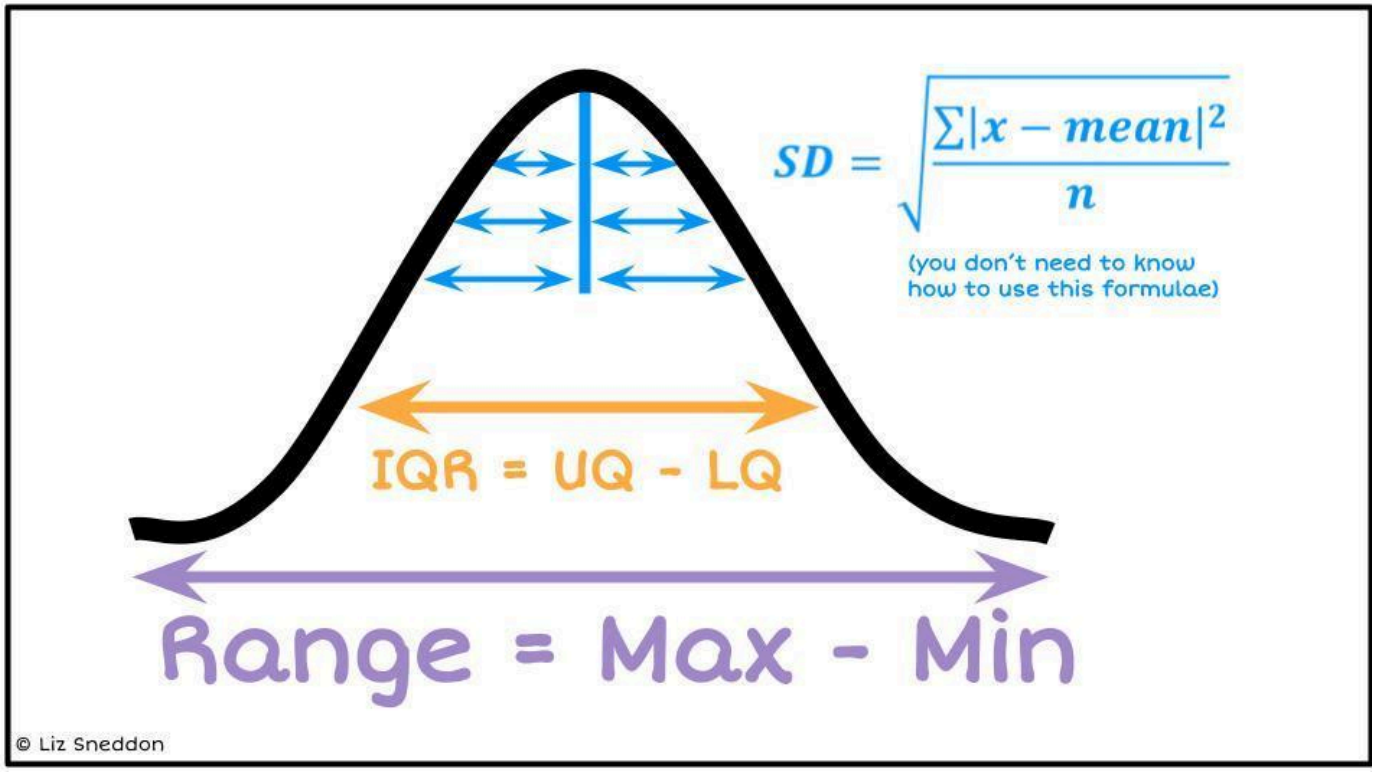
© Liz Sneddon

Standard deviation



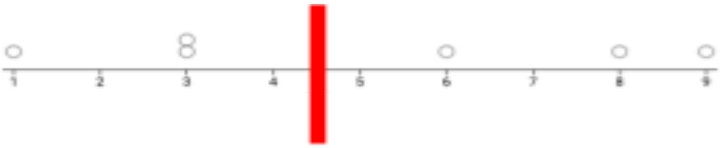
This measures how far away the data values are from the mean.

© Liz Sneddon


$SD = \sqrt{\frac{\sum|x - mean|^2}{n}}$
(you don't need to know how to use this formulae)
IQR = UQ - LQ
Range = Max - Min
© Liz Sneddon

Example:

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

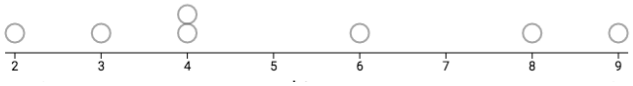


Data	9, 3, 1, 8, 3, 6
	Put the numbers in order: 1, 3, 3, 6, 8, 9
Minimum	1
Maximum	9
Median	Find the number(s) in the middle: 1, 3, 3, 6, 8, 9 Find the median = $\frac{3+6}{2}$ = 4.5
Mean	$\frac{9+3+1+8+3+6}{6}$ = 5
Mode	3
LQ	Take the numbers **below** the median: 1, 3, 3 Find the middle of these numbers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

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	

Which measures should be used?

In order of sophistication, our preference for which measure to use (if our data meets certain criteria) is as follows:

Centre:	**Spread:**
1) Mean 2) Median 3) Mode	1) Standard deviation, 2) IQR, 3) Range.

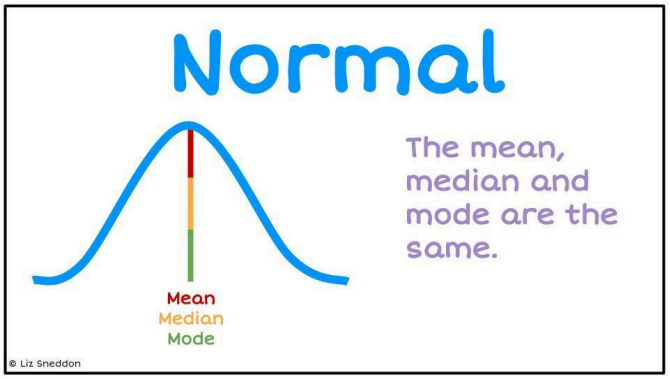
It's important to think about what the best way is to measure the centre and spread, depending on what **shape** the distribution is. Consider how the centre and spread are measured, and which measures of centre and spread will be affected by outliers and skewness, and which measures are the most stable.

We are going to focus on 3 main types of shaped distributions, normal, left skewed, and right skewed.

For **normally** distributed data:

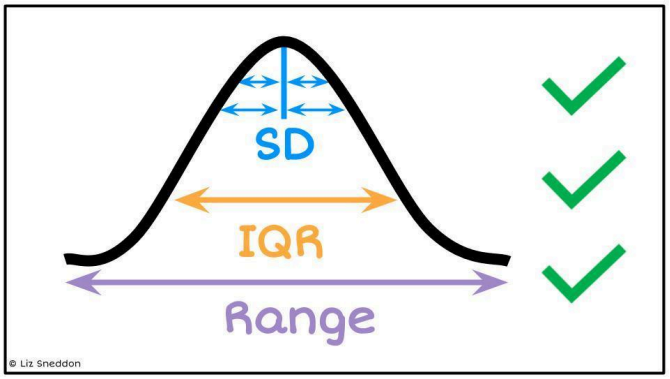
Centre:

- The mean, median and mode are all **very similar**, therefore they are all **good** measures of centre.
- Therefore, we choose to analyse the **most sophisticated measure**, which is the **mean**.



Spread:

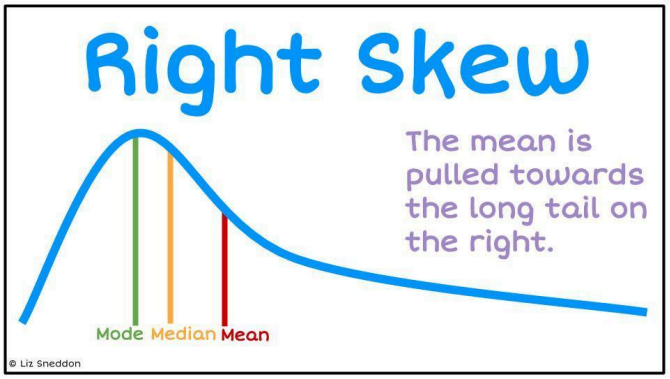
- The standard deviation, IQR and range are all **good** measures of spread.
- Therefore, we choose to analyse the **most sophisticated measure**, which is the **standard deviation**.

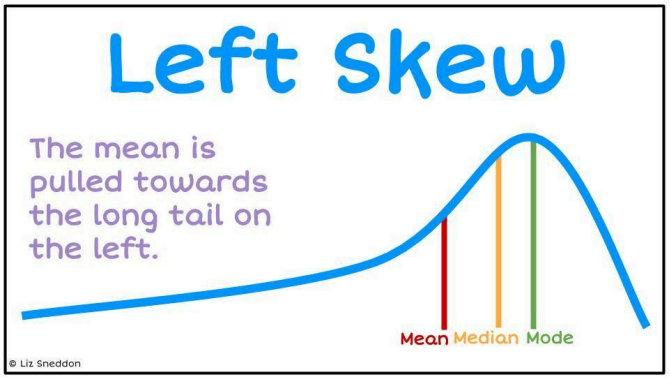


For **skewed** distributions:

Centre:

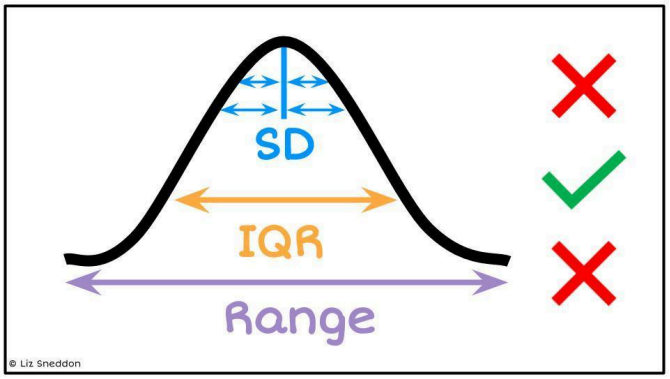
- The mean, median and mode are **different**.
- The mean is pulled in the direction of the longer tail. For right skewed data, the mean is pulled towards the right. For left skewed data, the mean is pulled towards the left.
- The mean is more affected by outliers and skewness, due to the way that it is calculated.
- Therefore, we choose to analyse the 2nd most sophisticated measure, and one that is **NOT** affected by outliers or skewness (due to the way that it is calculated), which is the **median**.





Spread:

- The range and the standard deviation are both affected by outliers and skewness, due to the way that it is calculated.
- The IQR is not affected by skewness or outliers, due to the way that it is calculated.
- Therefore, we analyse the one that is **NOT** affected by outliers or skewness, which is the **IQR**.



Exercise 7:

1) Explain how the mean and median are calculated.

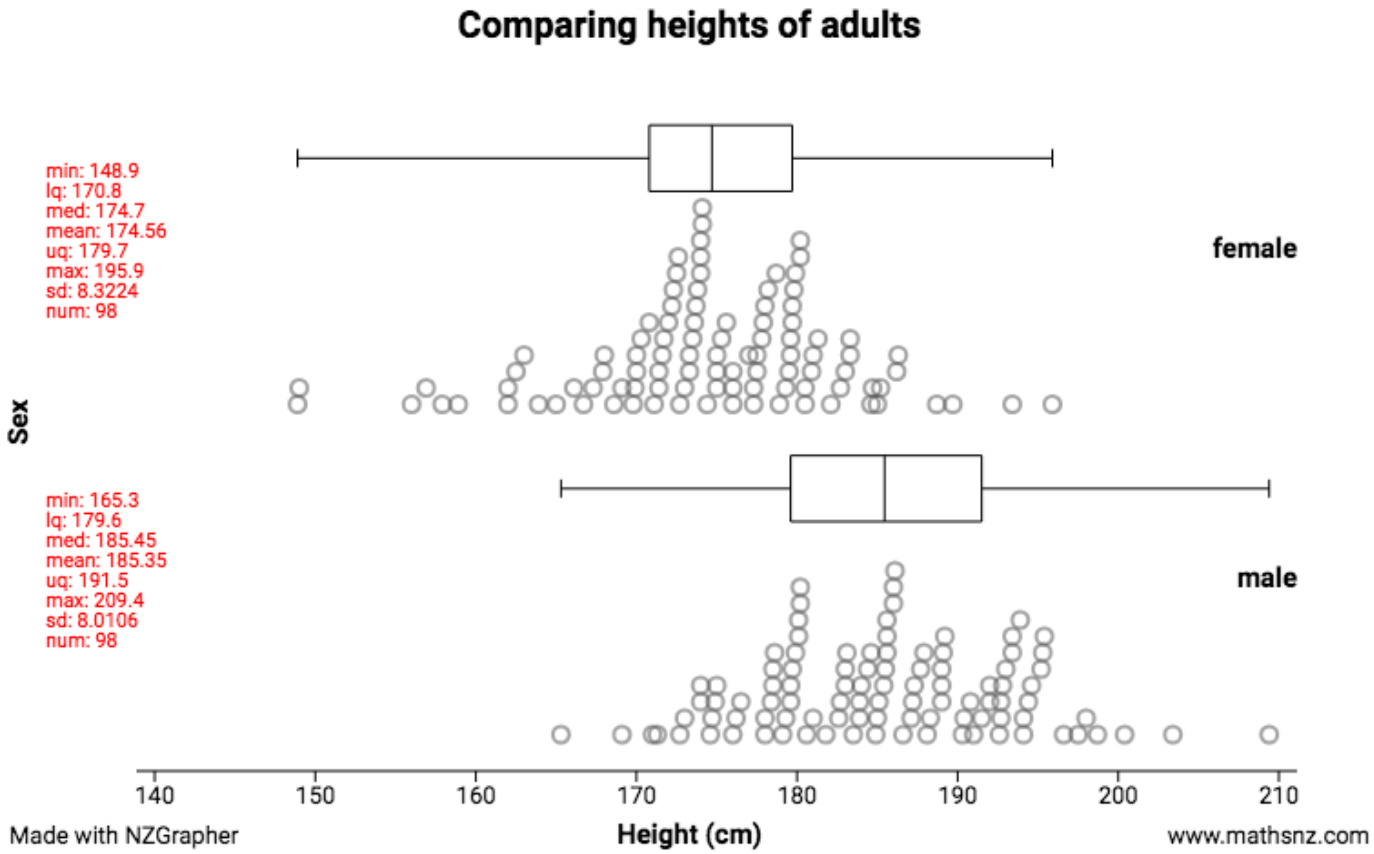
2) Explain why the mean is affected by outliers but the median is not.

3) Explain how the range and IQR are calculated.

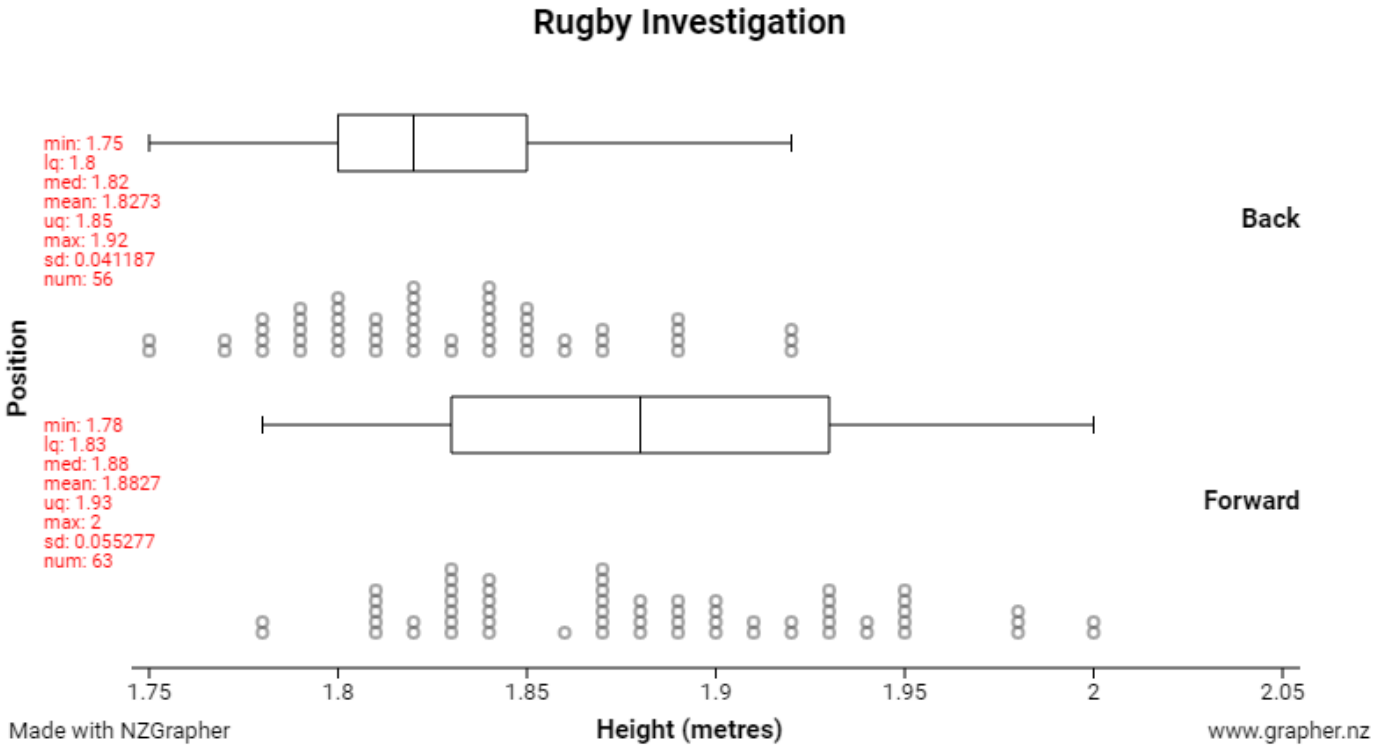
4) Explain why the Range is affected by outliers but the IQR is not.

5) Explain why the IQR is a more stable measure of the spread than the standard deviation if we have outliers.

6) For the data shown below, which measures of centre and spread would you recommend to analyse and why?



7) For the data shown below, which measures of centre and spread would you recommend to analyse and why?

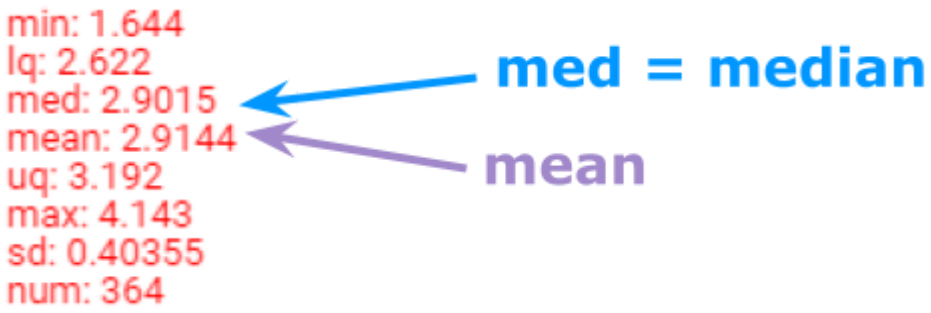


Dot Plots & Boxplots in NZGrapher

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

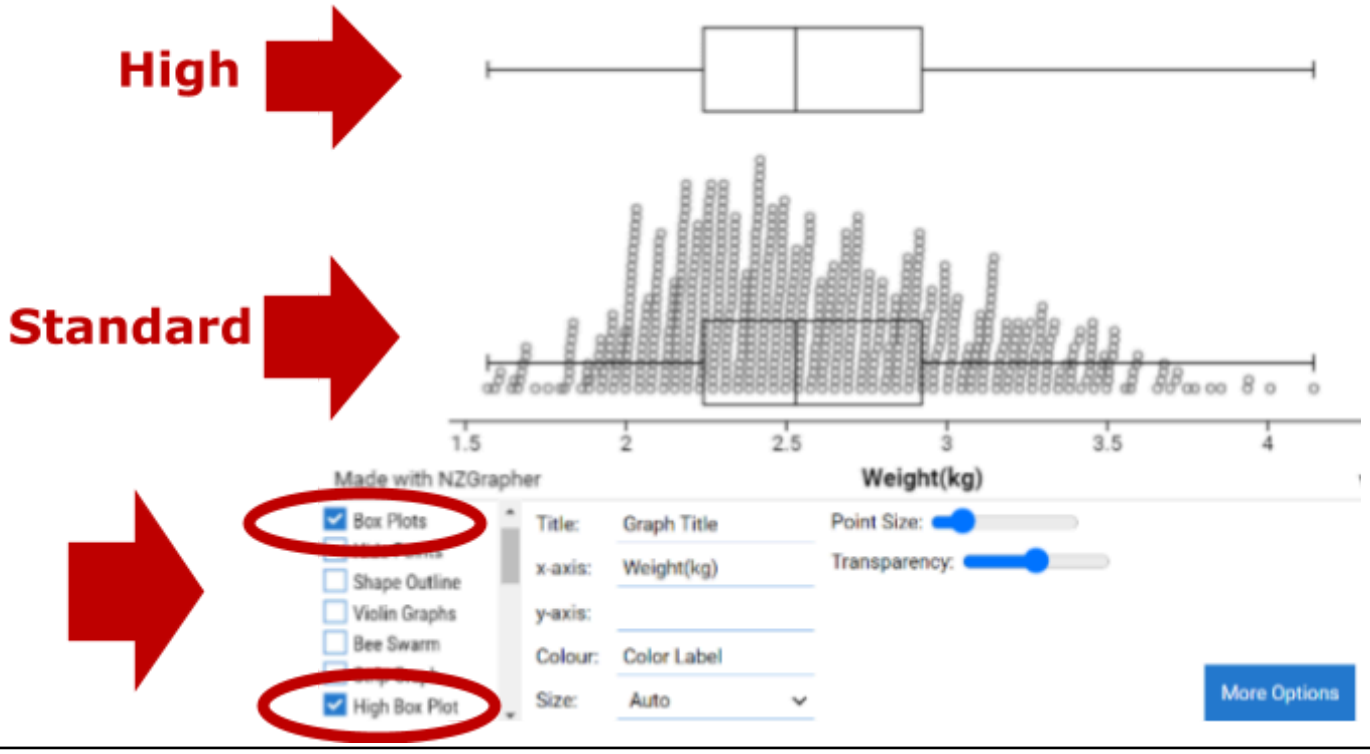
1) Add summary statistics to your graph.

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.

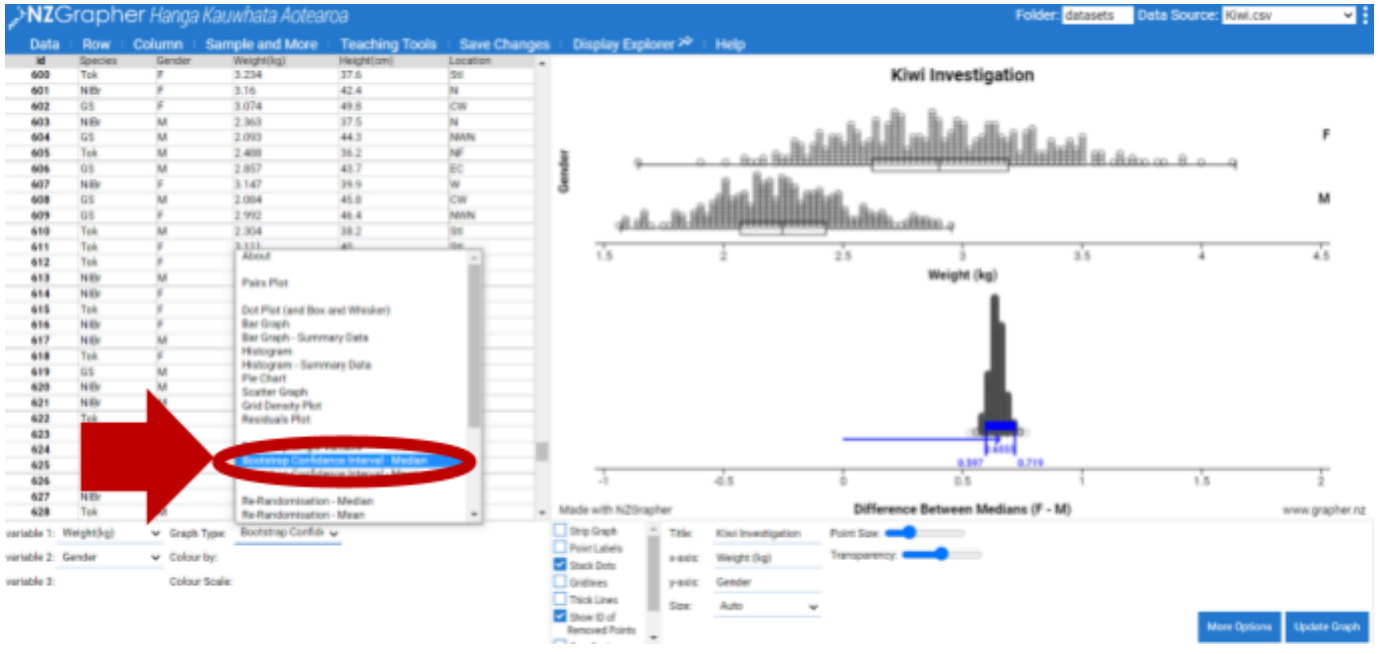
3) There are several options for box plots, one that sits **on top of** the data, and one that sits **above** the data (tick the High Box plot option).



Bootstrap Graph in NZGrapher

At Level 3 there is an additional graph that we need to add, and that is the bootstrap graph – median. We will explore what this graph is and how to interpret the results in the Conclusion section.

Change the graph type to "**Bootstrap Confidence Interval - Median**"



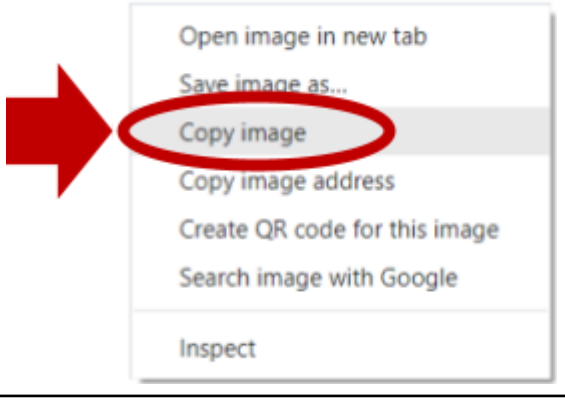
NZGrapher Tips

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.

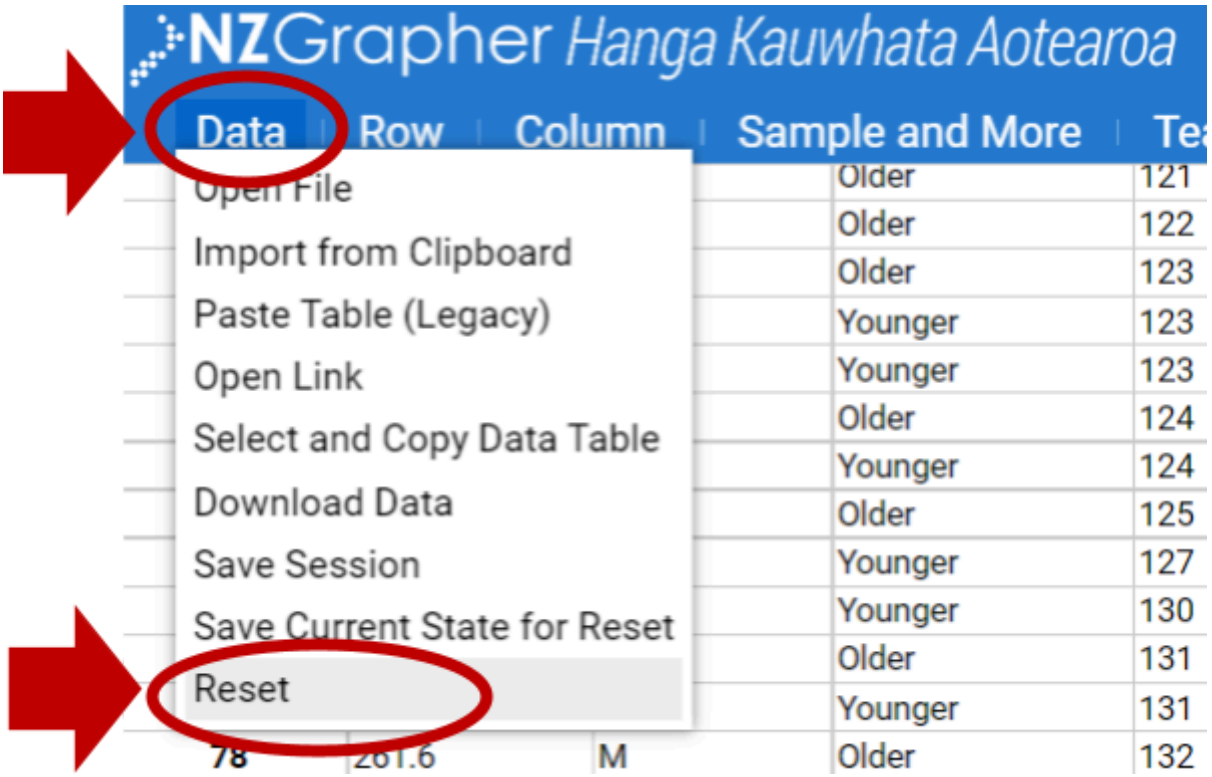
Tip 2: There are some new graphs that you can change / add. I like the Shape outline as the overall shape can be seen more easily.

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.

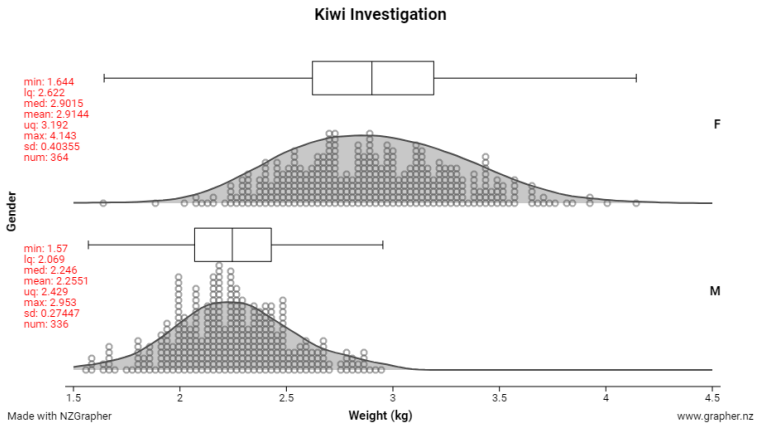
Tip 5: If **you** make a mistake when you take a random sample, then go to the Data menu and select the **Reset** button. This will reload the full dataset and you can start again.



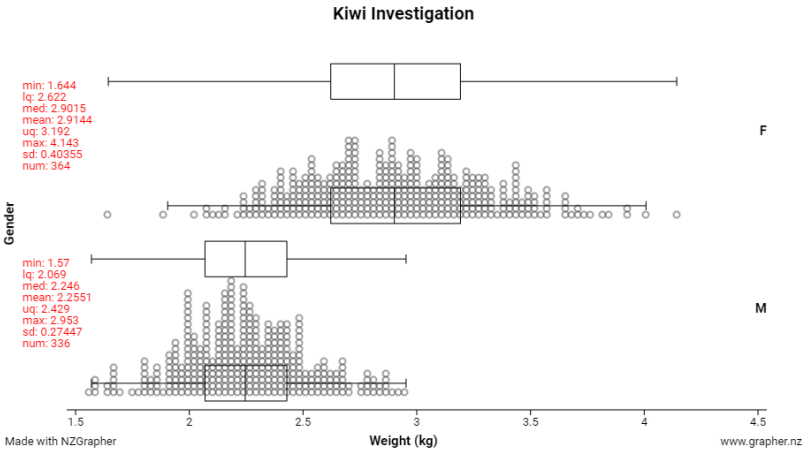
Putting NZGrapher together

These are the three graphs I usually produce to help me analyse and draw conclusions using the data.

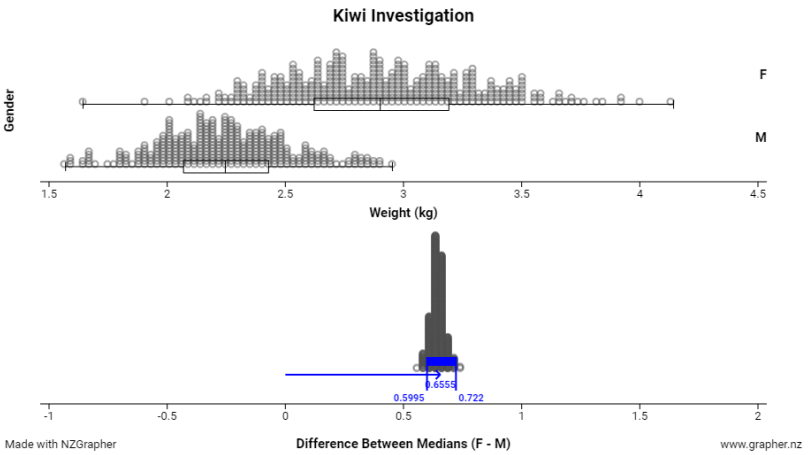
Graph 1: Dot Plot – Analysis of features

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	 Don't forget that you can change the point size if you have small sample sizes: Point Size: Transparency:

Graph 2: Dot Plot – 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: Bootstrap Confidence Interval – Medians

Select the following options:	Graph
☐ Strip Graph ☐ Point Labels ☑ Stack Dots ☐ Gridlines ☐ Thick Lines ☑ Show ID of Removed Points	

Exercise 8:

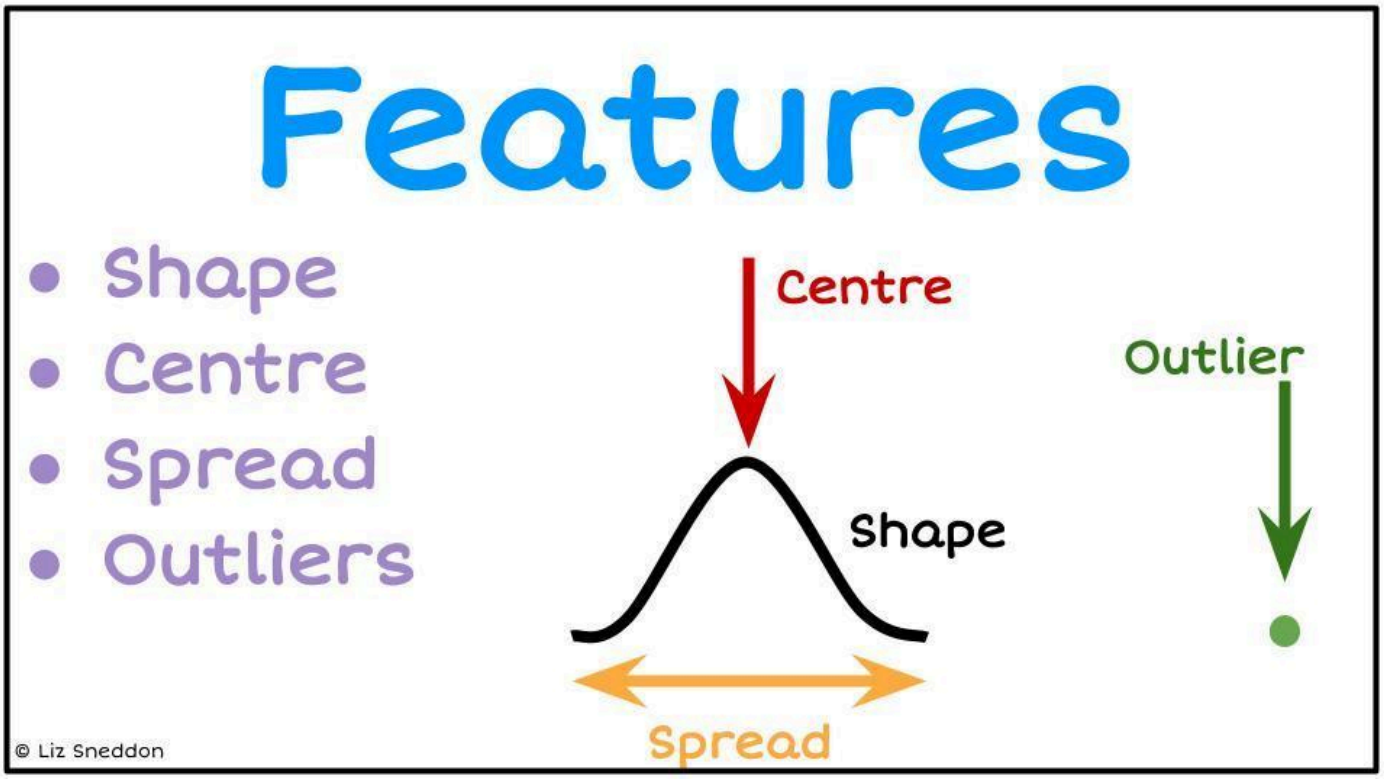
Go to NZGrapher and form appropriate graphs for each of the datasets below (don't forget titles, axis labels etc):

1) Select the Marathon dataset and use the variables "**Minutes**" and "**AgeGroup**".

2) Select the Kiwi dataset and use the variables "**Weight**" and "**Gender**".

3) Select the BallWear dataset and use the variables "**Gender**" and "**AmountSpent**".

Analysis / Tātari

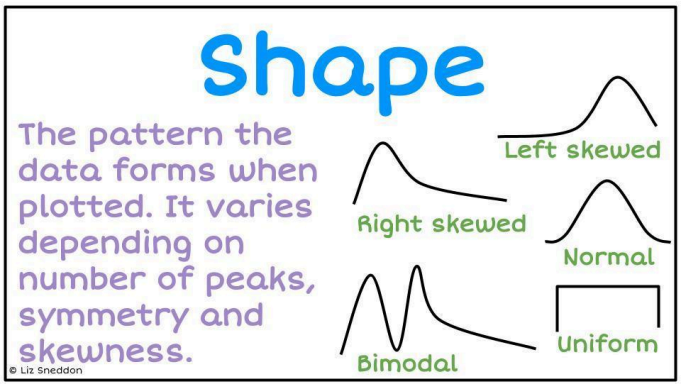


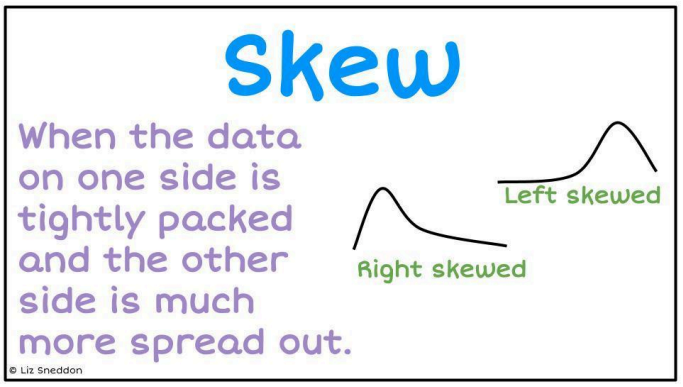
Remember to use **suggestive language** when you analyse the features and patterns in your **sample data**.

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

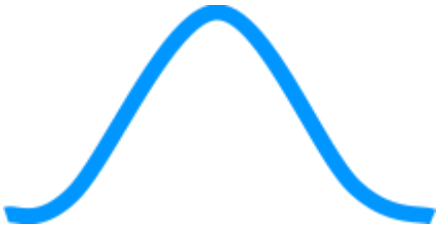
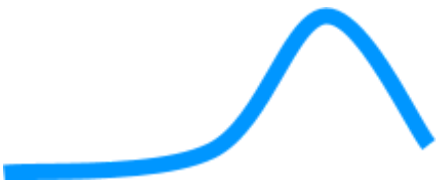
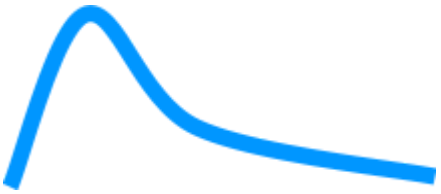
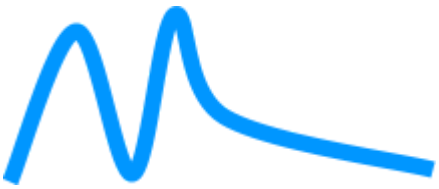
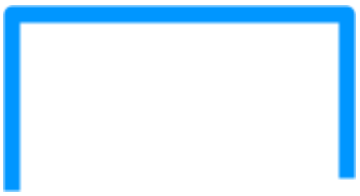
Shape

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:

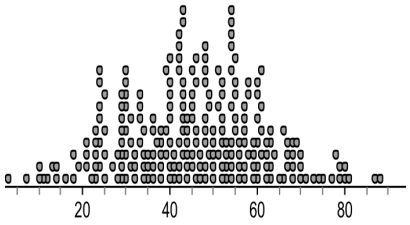
Not like this:

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

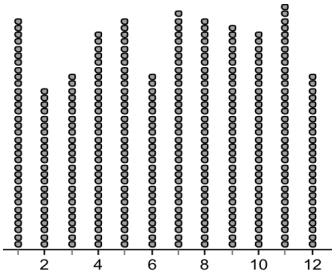
Exercise 9:[5]

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

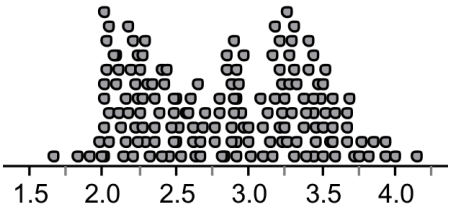
1.



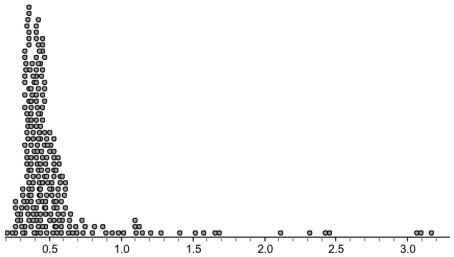
2.



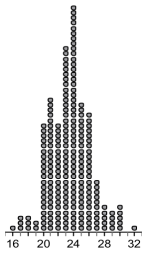
3.



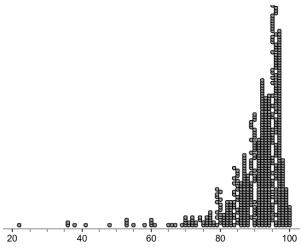
4.



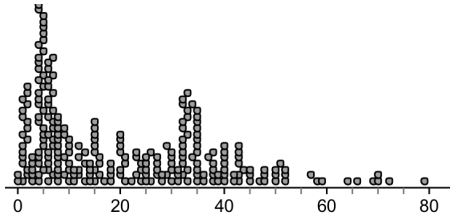
5.



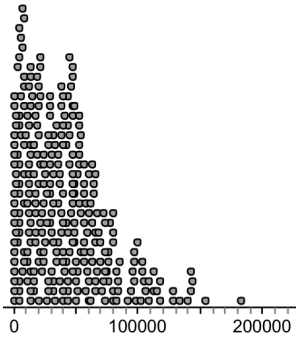
6.



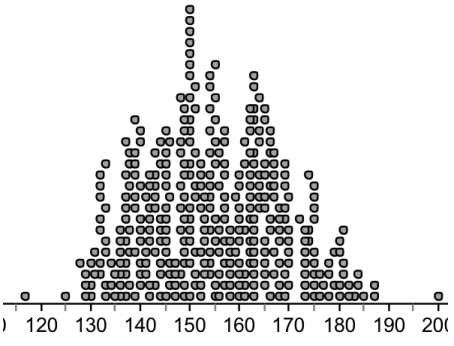
7.



8.

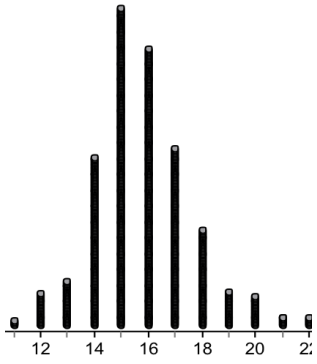


9.

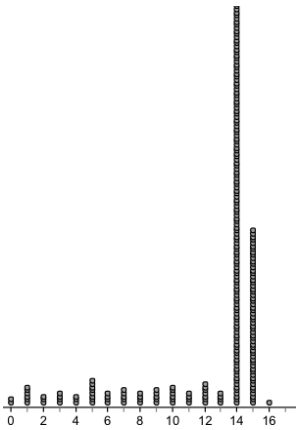


[5] Thanks to Dr Pip Arnold for the graphs.

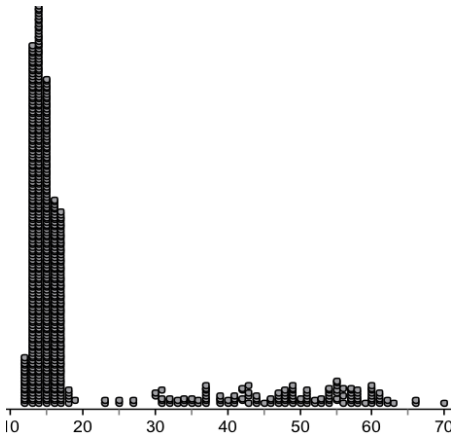
10.



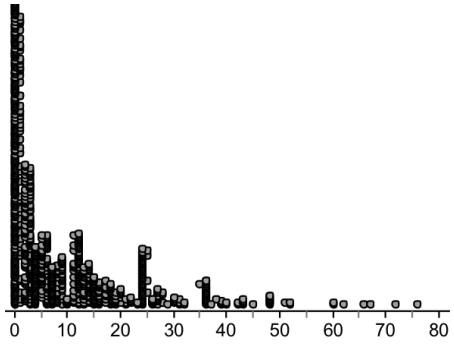
11.



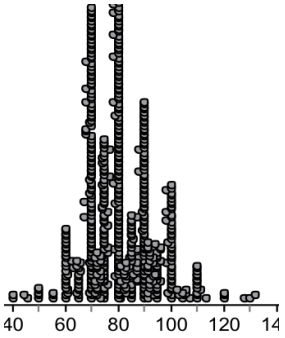
12.



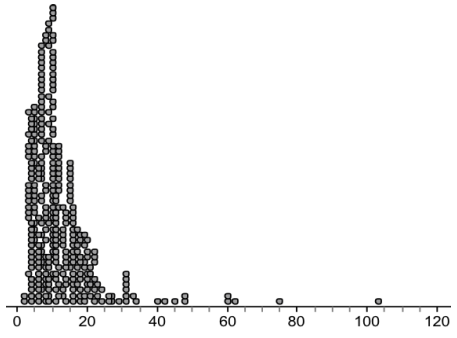
13.



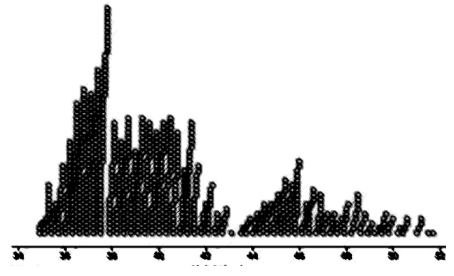
14.



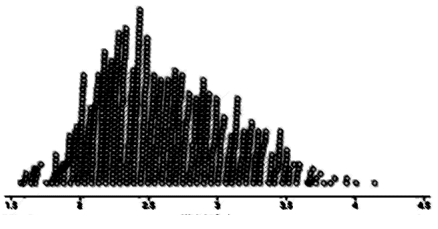
15.



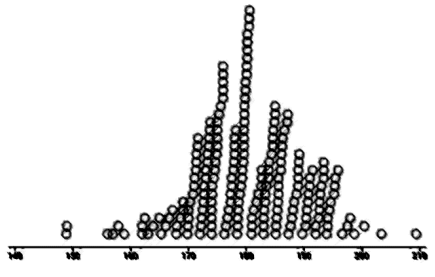
16.

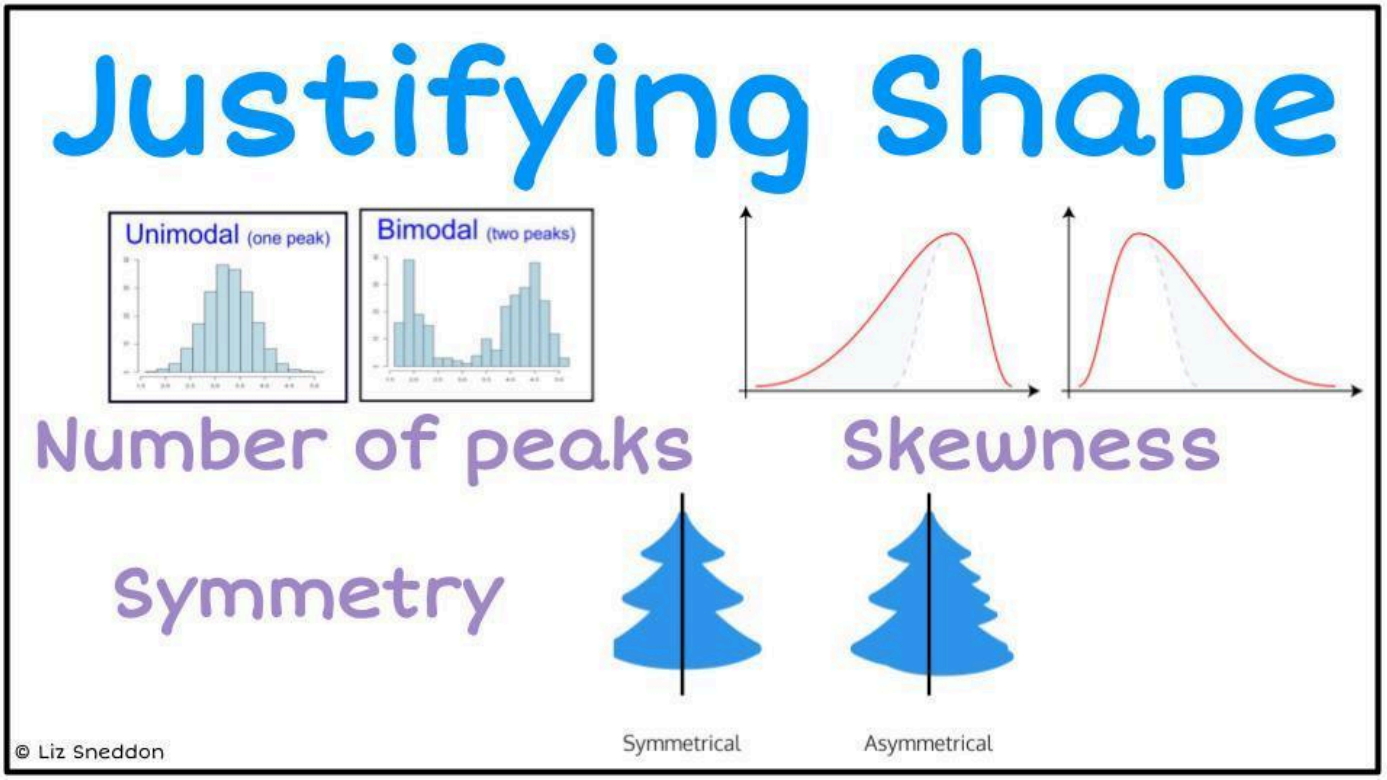


17.



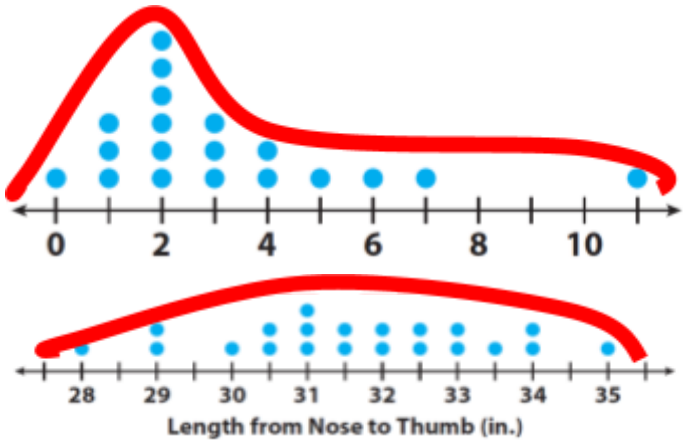
18.





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.

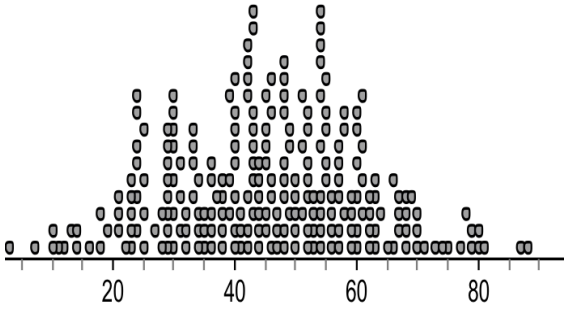


The shape of the data looks right skewed, as there is a longer tail on the right, with one hill / peak. It is not symmetric.

The shape of the data looks approximately normal, as there is one hill / peak, it is roughly symmetric and the tails on both sides are similar.

Exercise 10:

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



Symmetric (circle one): Yes / No

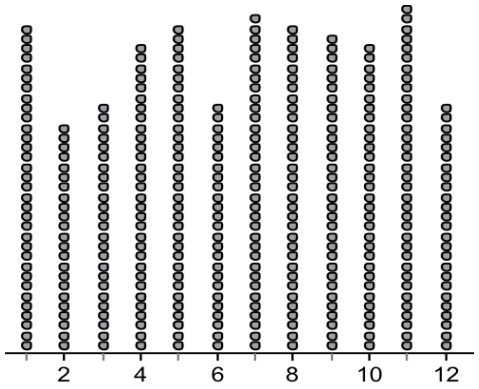
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

Left hand tail longer

Right hand tail longer



Symmetric (circle one): Yes / No

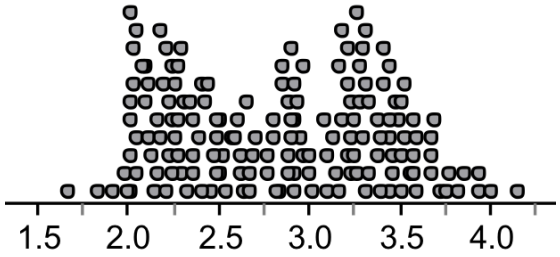
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

Left hand tail longer

Right hand tail longer



Symmetric (circle one): Yes / No

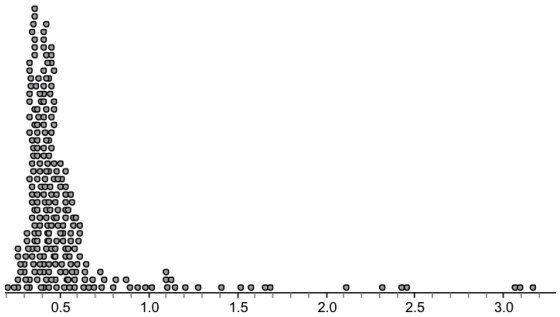
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

Left hand tail longer

Right hand tail longer



Symmetric (circle one): Yes / No

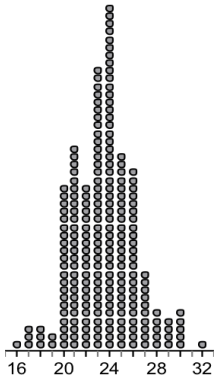
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

Left hand tail longer

Right hand tail longer



Symmetric (circle one): Yes / No

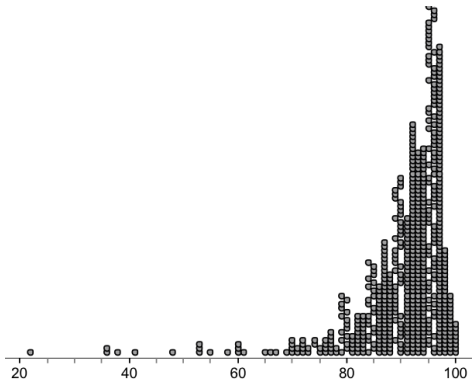
Number of hills / peaks: 0 1 2

Bell shaped curve: Yes / No

Tails (tick only if appropriate):

Left hand tail longer

Right hand tail longer



Symmetric (circle one): Yes / No

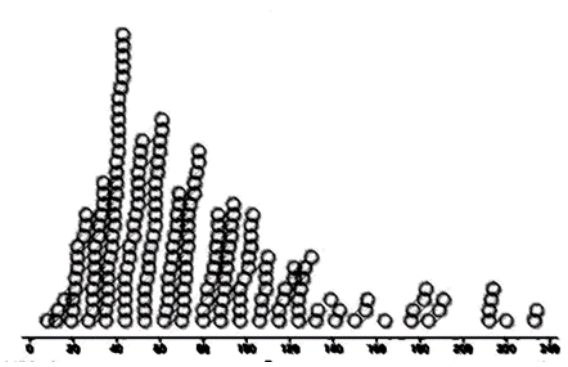
Number of hills / peaks: 0 1 2

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

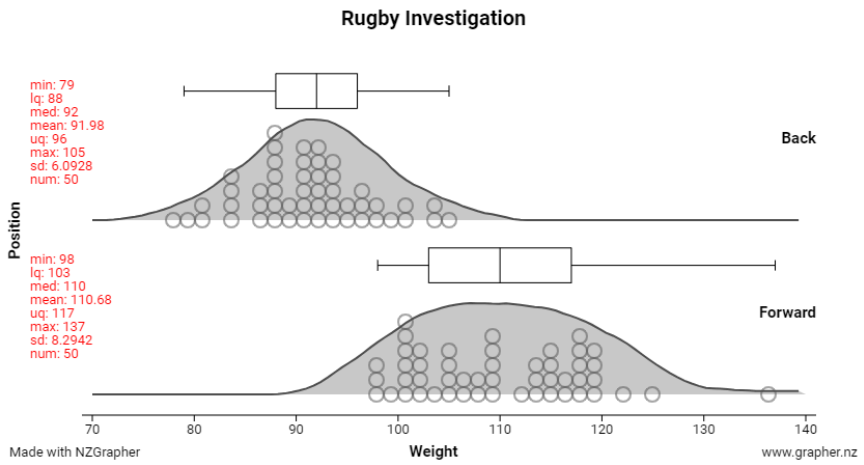
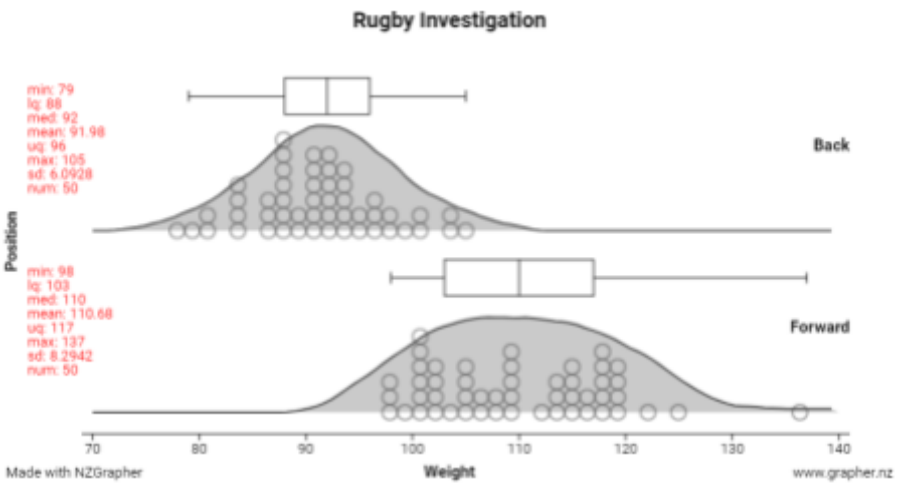
When NZGrapher draws the shape outline over each dataset, it is endeavoring to show what the **shape of the population BEHIND the sample of data** might look like. This is why the curves are smooth, and this is the shape we are trying to identify.

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.

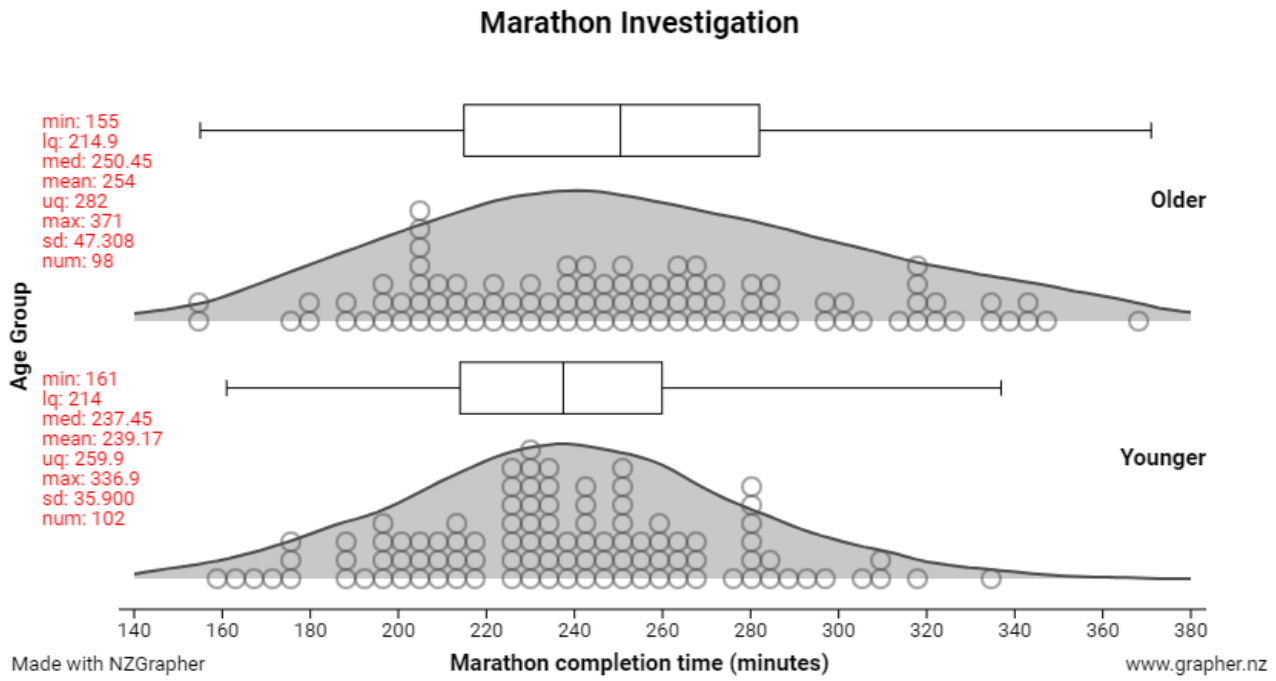
Example:

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.

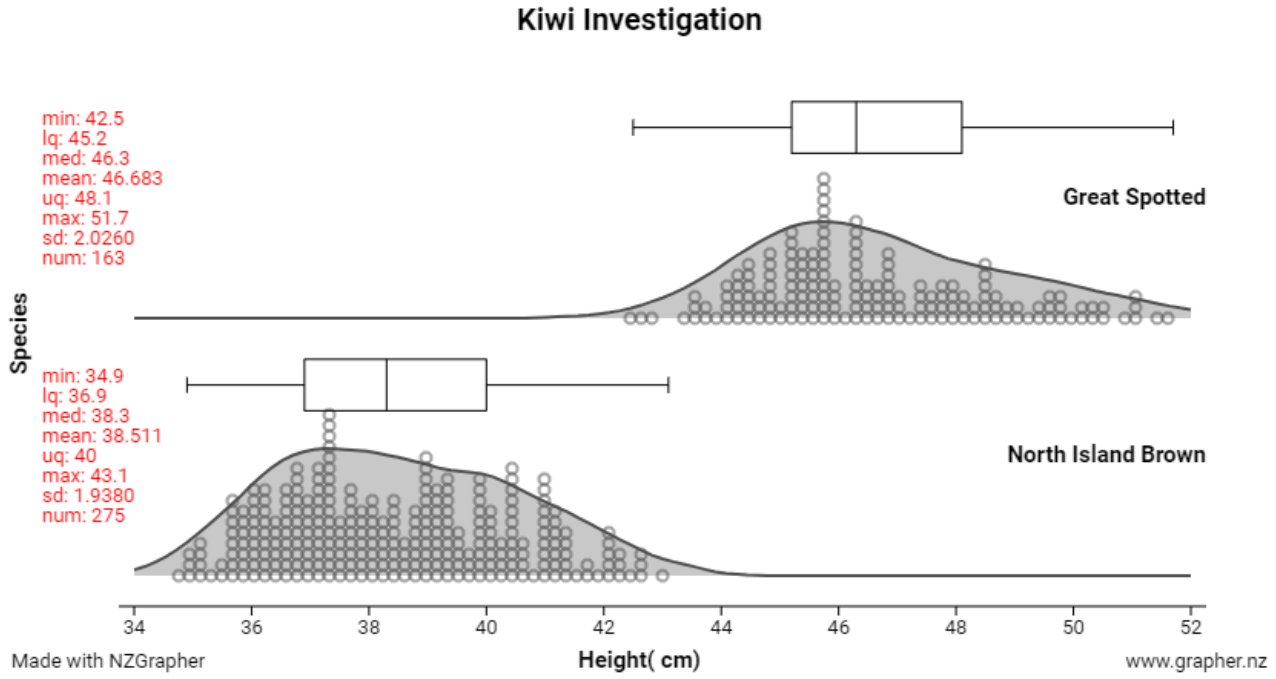
Exercise 11:

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

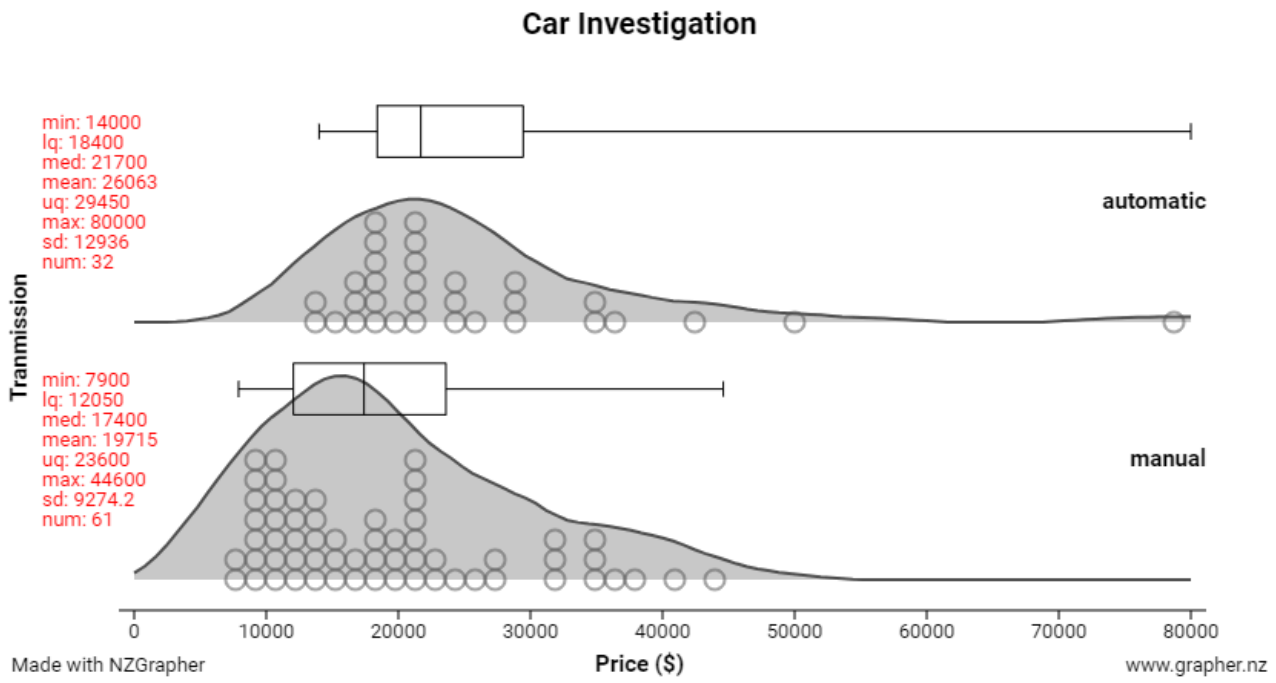
a)



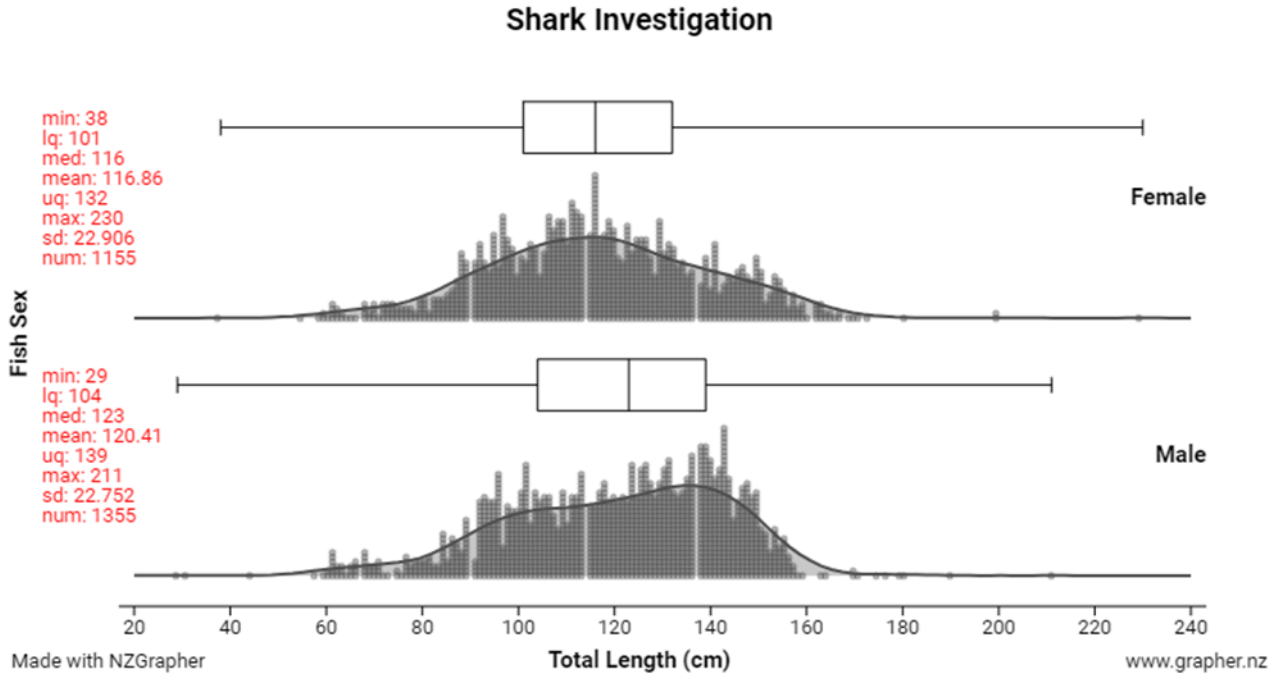
b)


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)

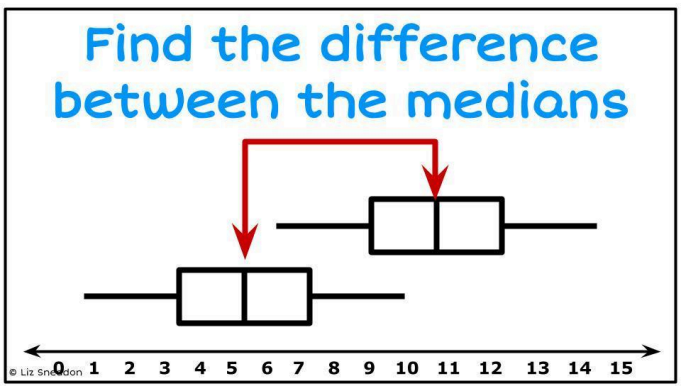

Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Made with NZGrapher
Price ($)
www.grapher.nz

d)



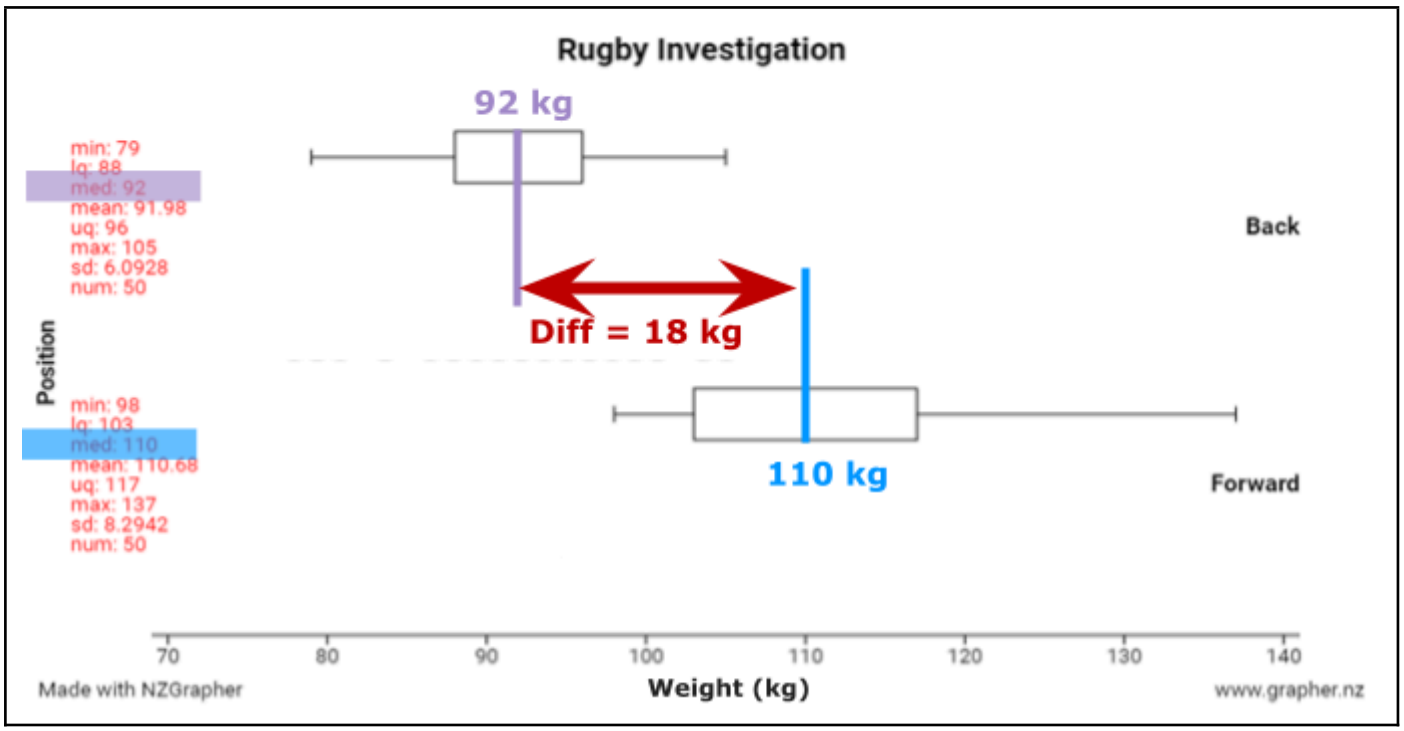
Comparing Centre

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

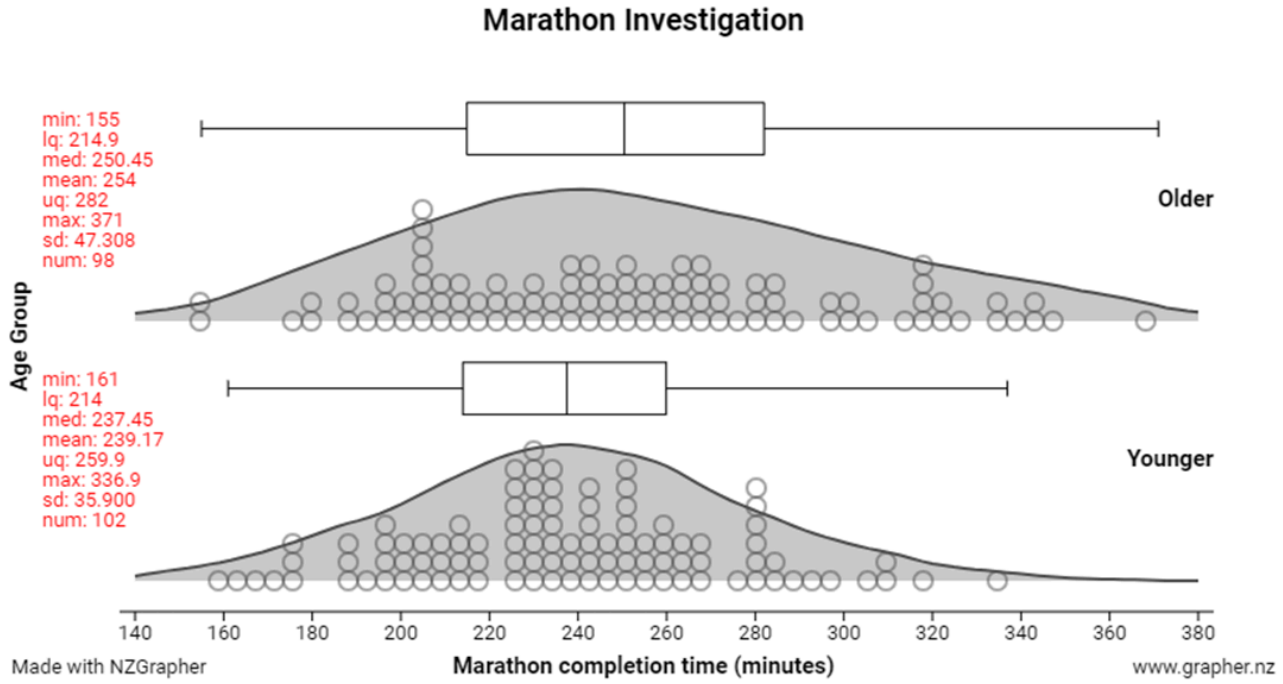
= 18 kg

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.

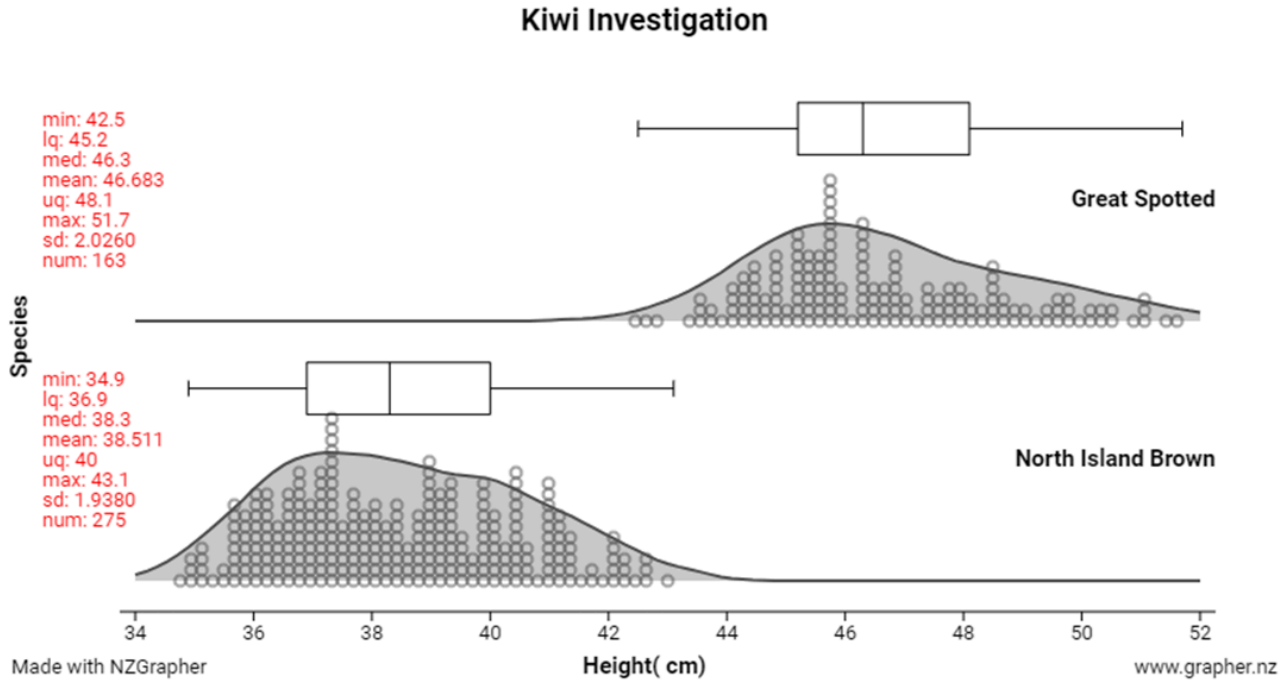
Exercise 12:

Compare the medians for the graphs below.

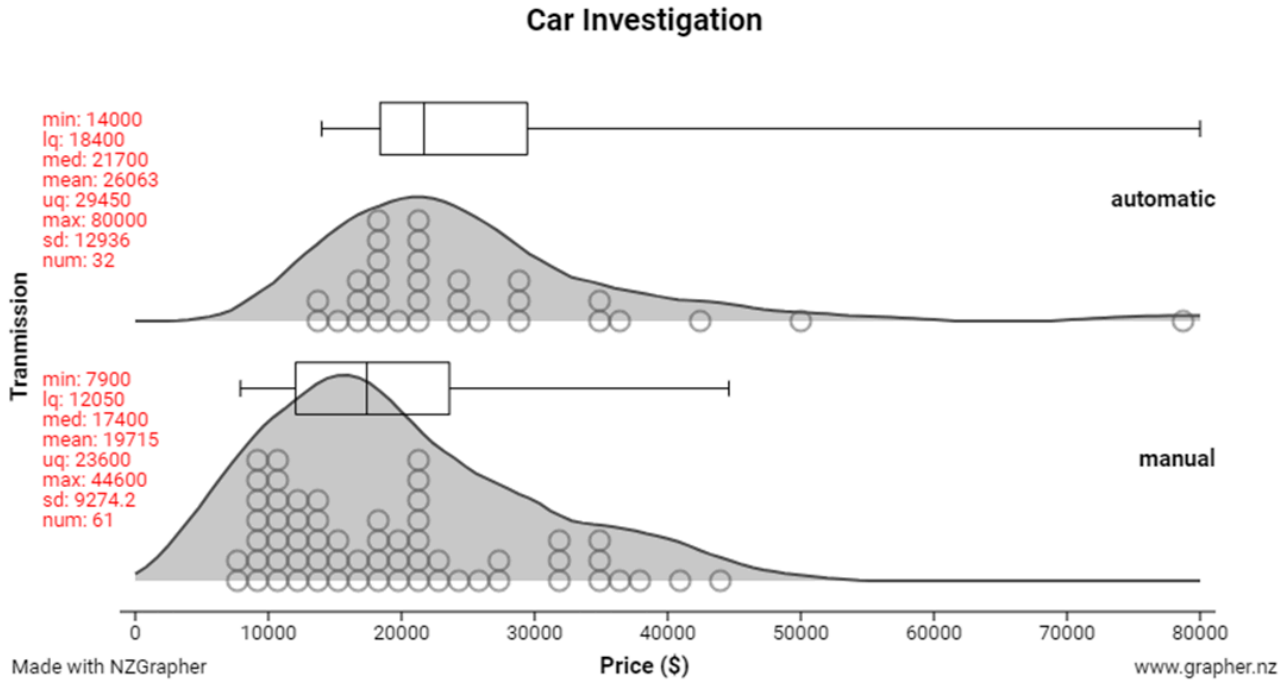
1)


Marathon Investigation
min: 155
lq: 214.9
med: 250.45
mean: 254
uq: 282
max: 371
sd: 47.308
num: 98
Older
Age Group
min: 161
lq: 214
med: 237.45
mean: 239.17
uq: 259.9
max: 336.9
sd: 35.900
num: 102
Younger
140
160
180
200
220
240
260
280
300
320
340
360
380
Made with NZGrapher
Marathon completion time (minutes)
www.grapher.nz

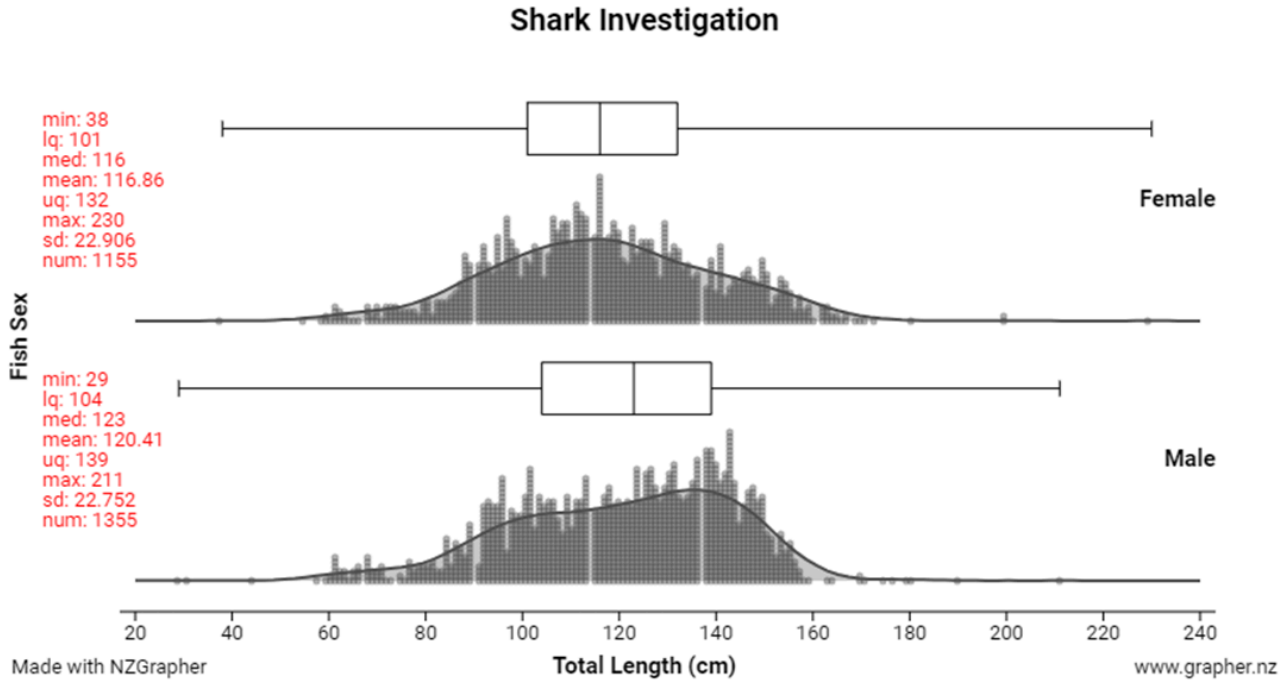
2)


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

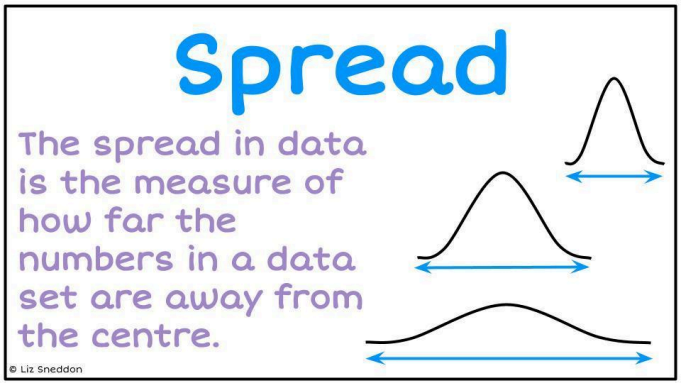
3)

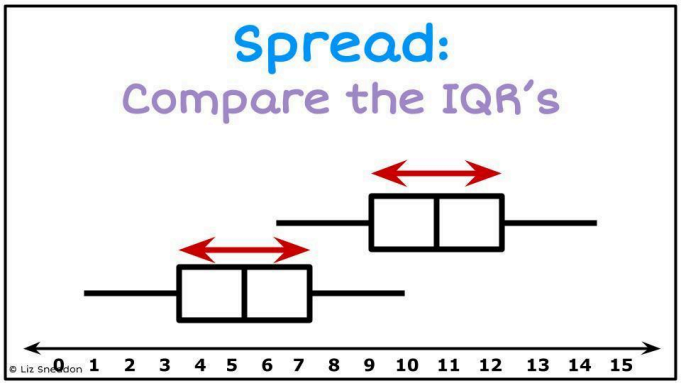

Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Made with NZGrapher
Price ($)
www.grapher.nz

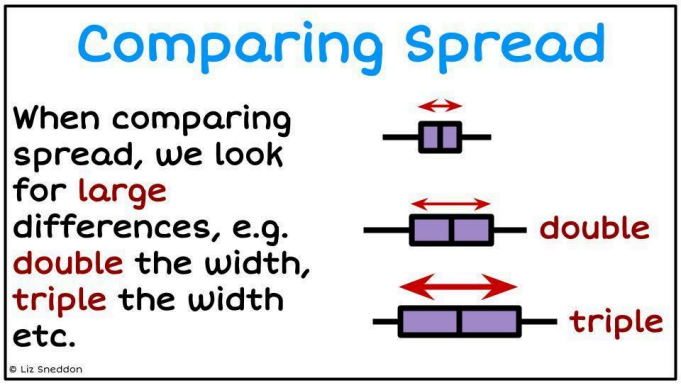
4)

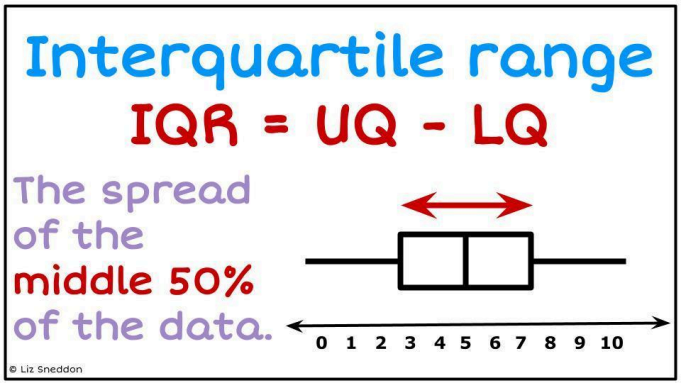

Shark Investigation
min: 38
lq: 101
med: 116
mean: 116.86
uq: 132
max: 230
sd: 22.906
num: 1155
Female
Fish Sex
min: 29
lq: 104
med: 123
mean: 120.41
uq: 139
max: 211
sd: 22.752
num: 1355
Male
20
40
60
80
100
120
140
160
180
200
220
240
Total Length (cm)
Made with NZGrapher
www.grapher.nz

Comparing Spread









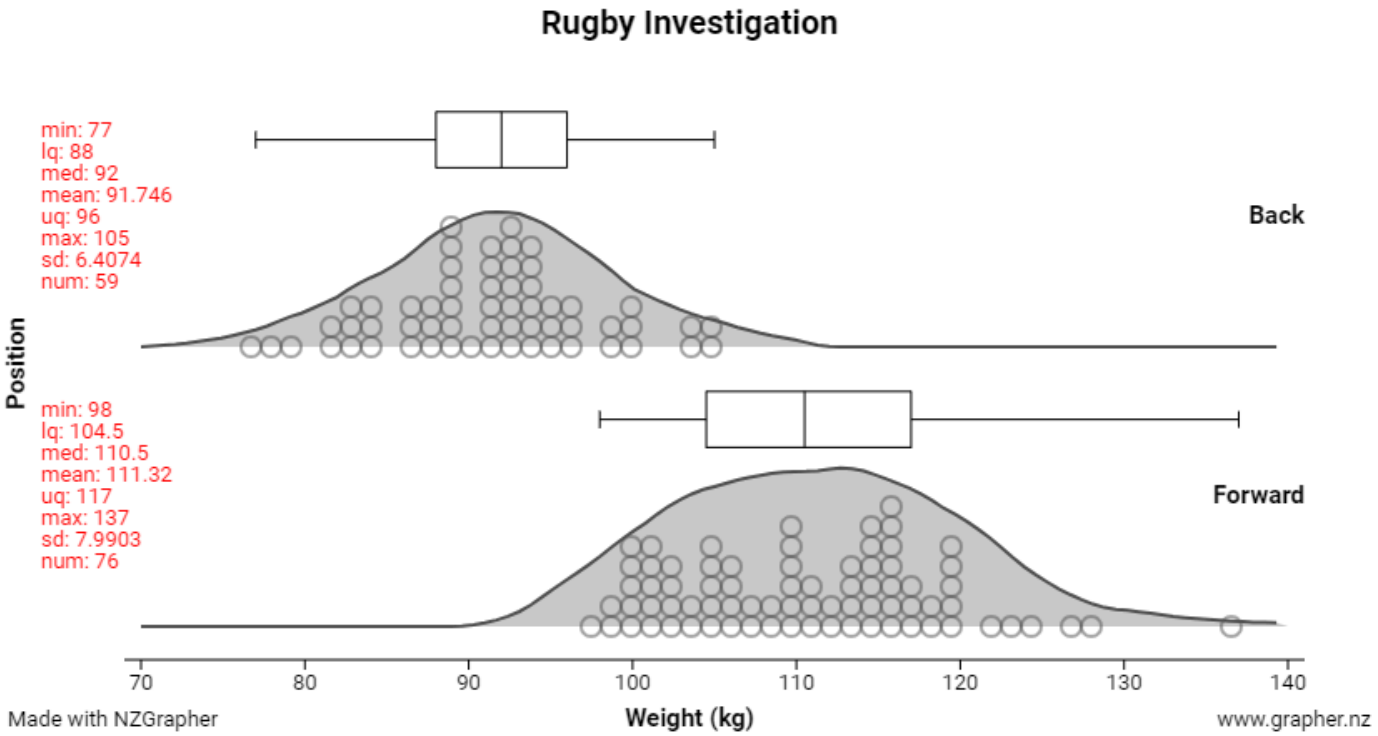
The IQR is the middle 50% of your sample data, and in our boxplot, it is the box.

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:



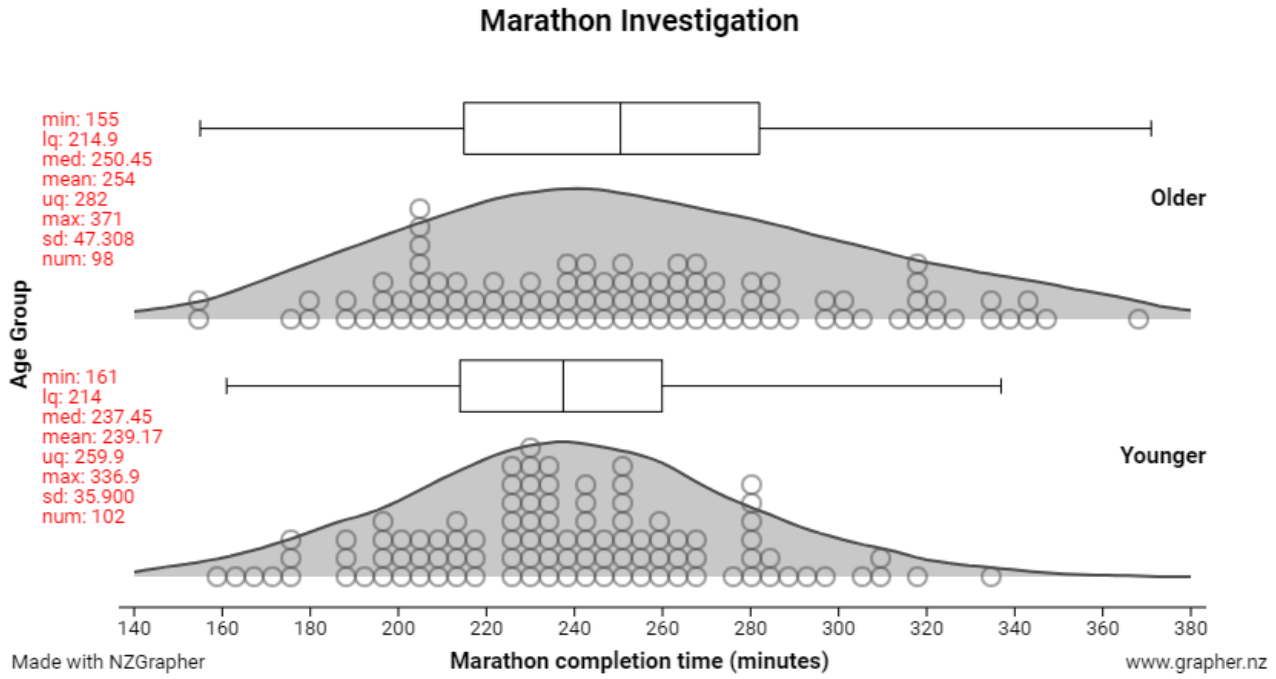
IQR (backs) = UQ – LQ = 96 – 88 = 8 kg The IQR of weights for back rugby players is 8kg.	IQR (forwards) = UQ – LQ = 117 – 110.5 = 6.5 kg The IQR of weights for forwards rugby players is 6.5kg.

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

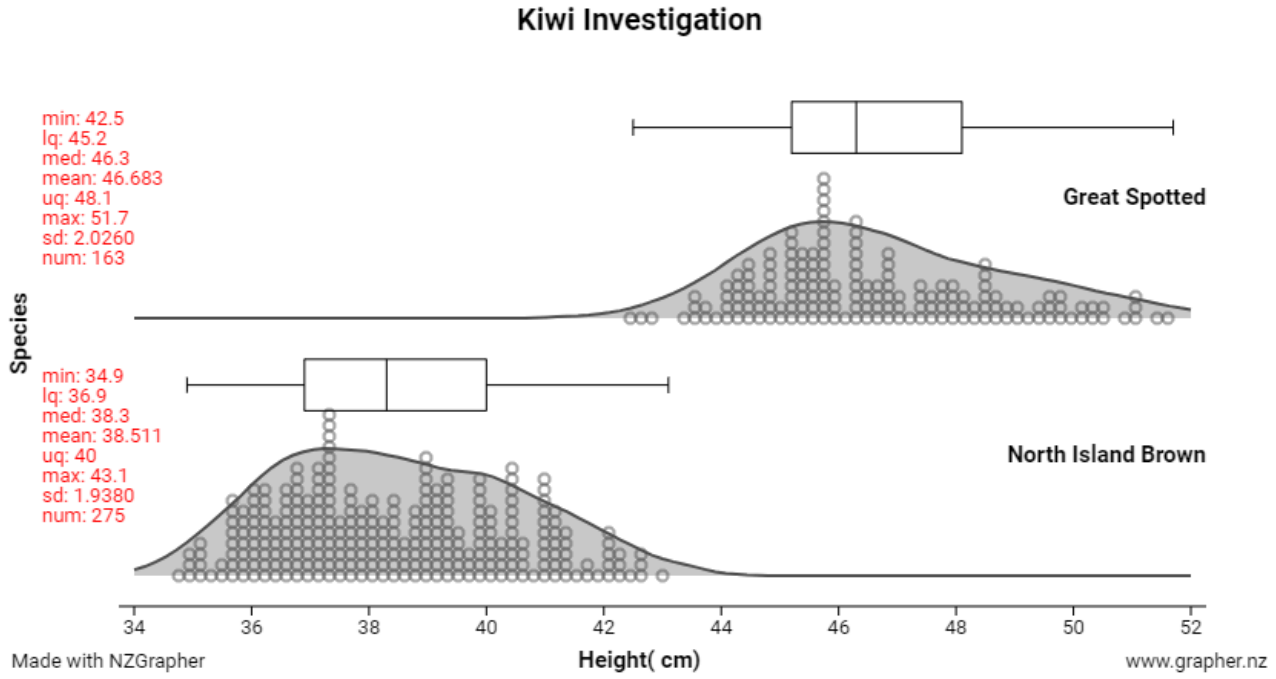
Exercise 13:

Compare the spread for the graphs below.

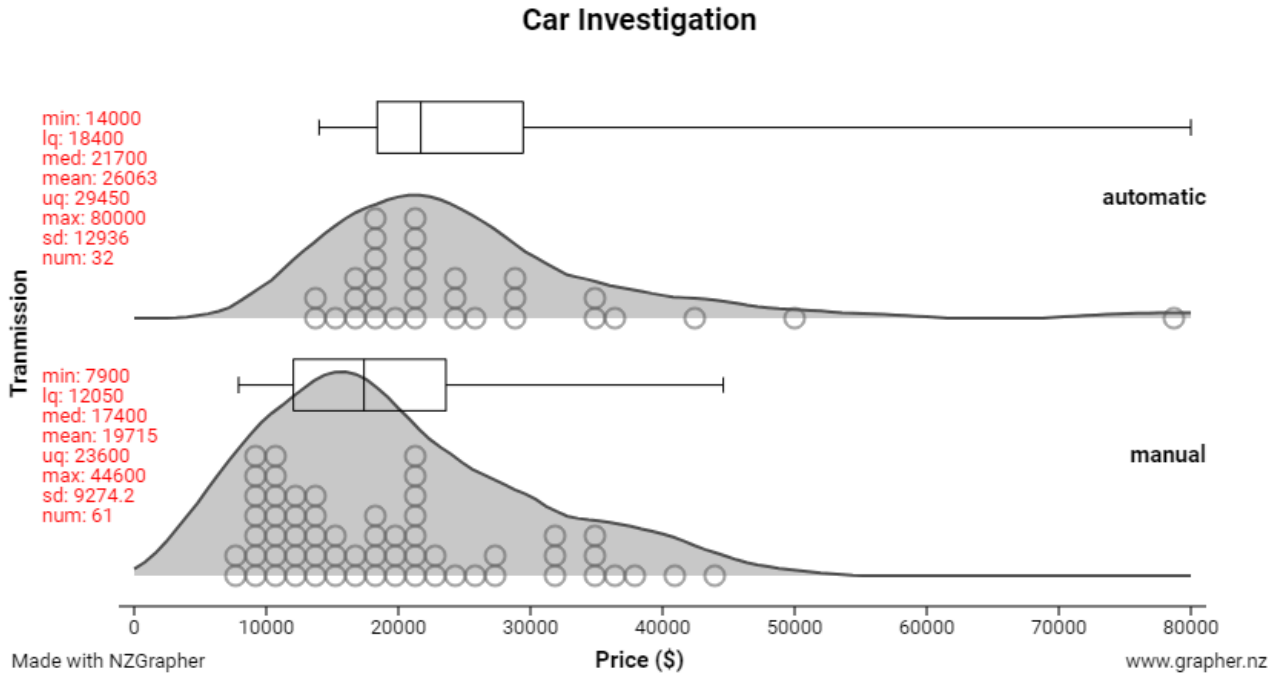
1)



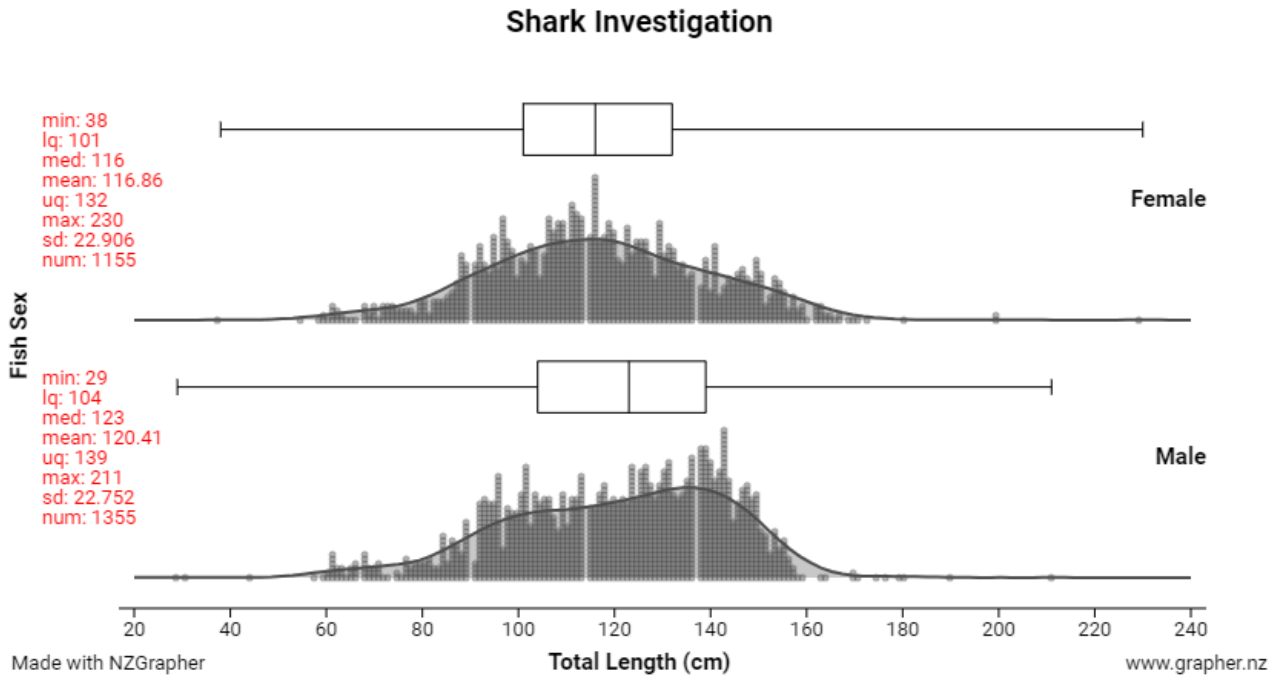
2)


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)

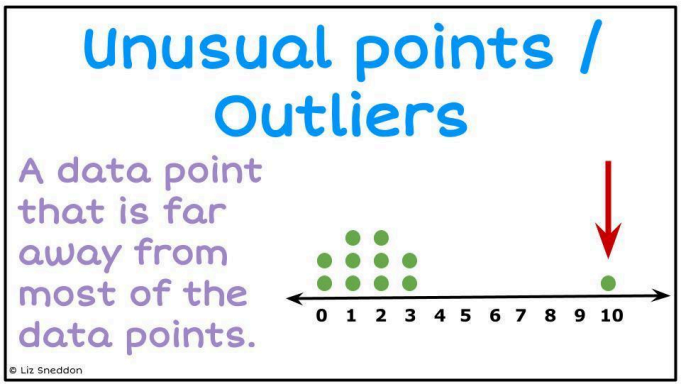

Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Made with NZGrapher
Price ($)
www.grapher.nz

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

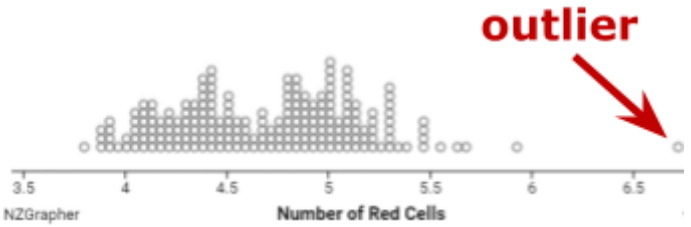
Outliers

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.

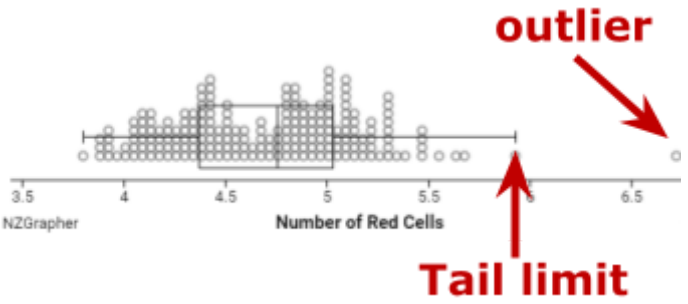


Example 1:

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.



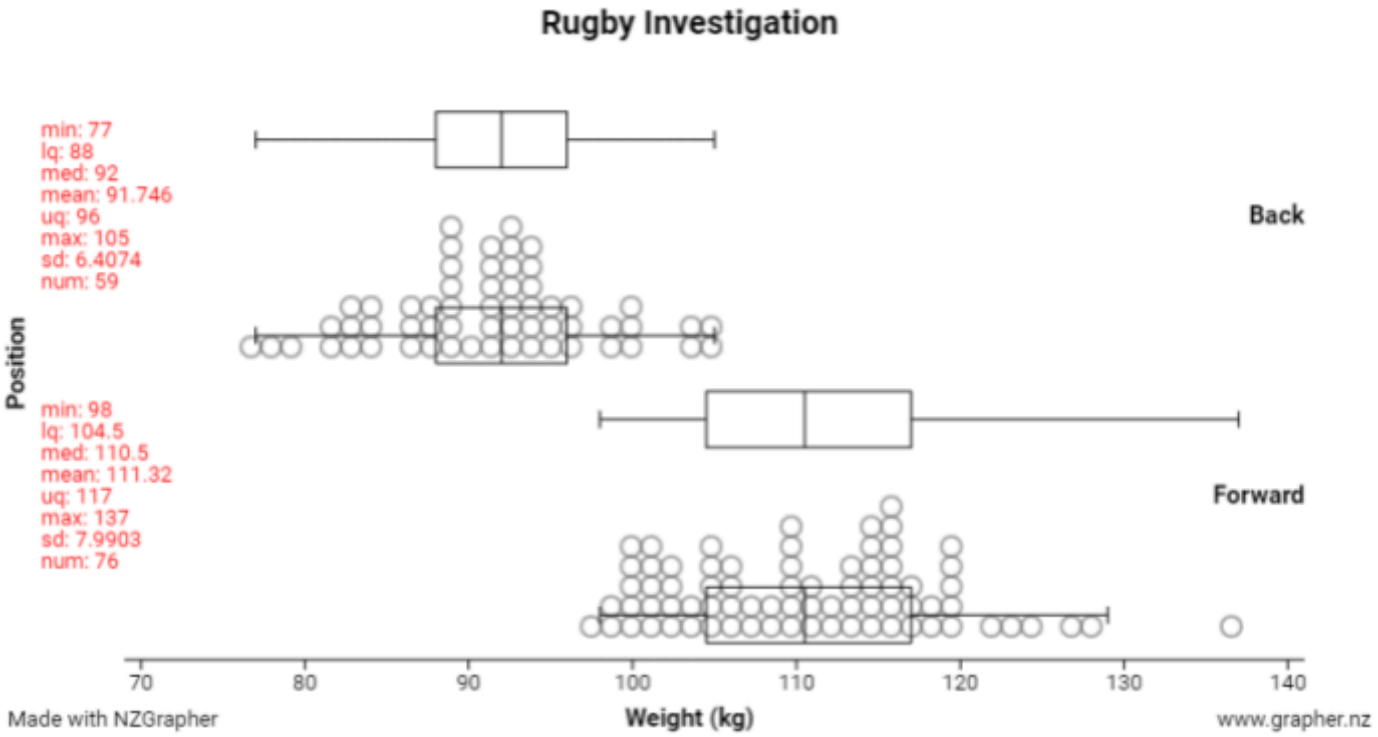
There is an outlier, a person with around 6.8 red blood cells per microliter.

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.

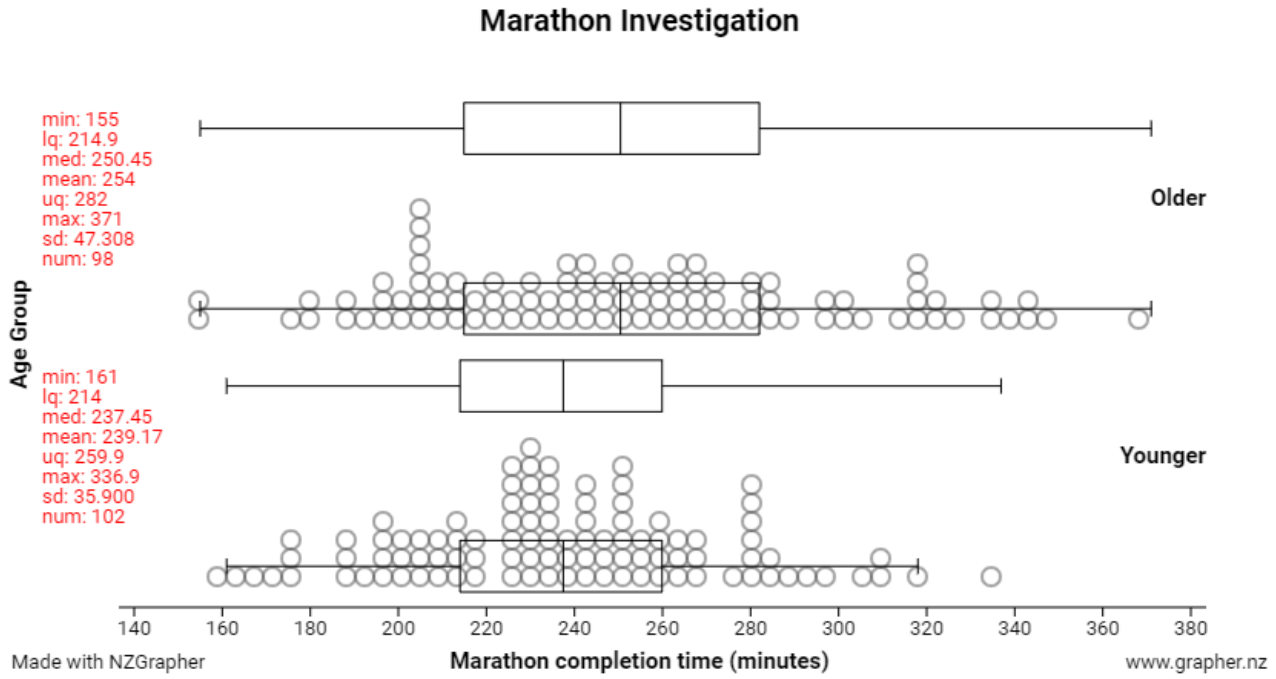


There are two outliers, who both play in a forwards position in rugby. One has a smaller than expected weight of around 96 kg, and another with a weight higher than expected at around 136 kg.

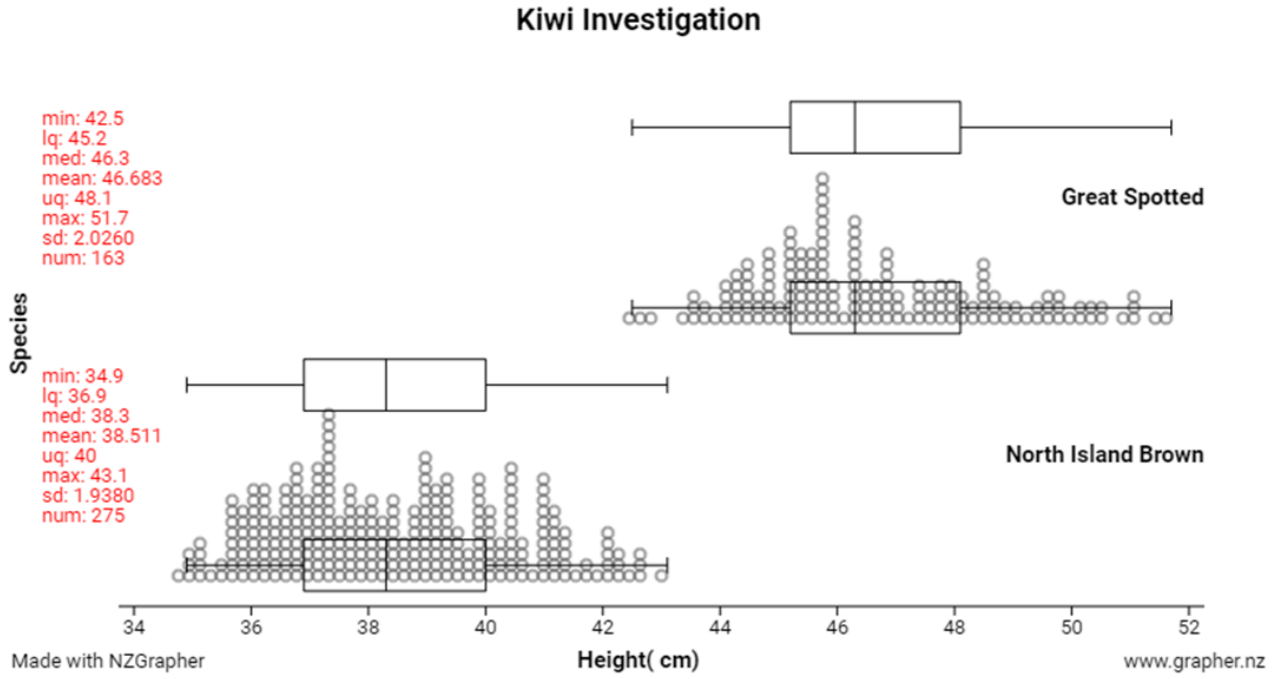
Exercise 14:

Identify and describe any outliers in the graphs below.

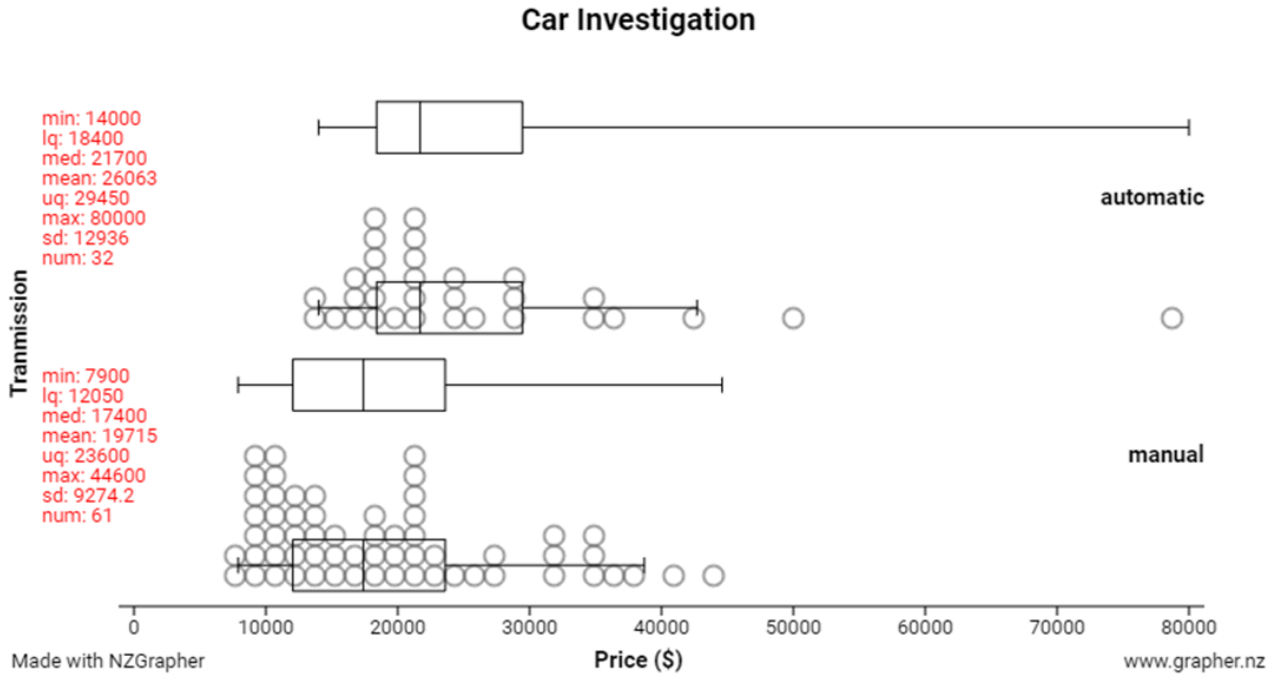
1)



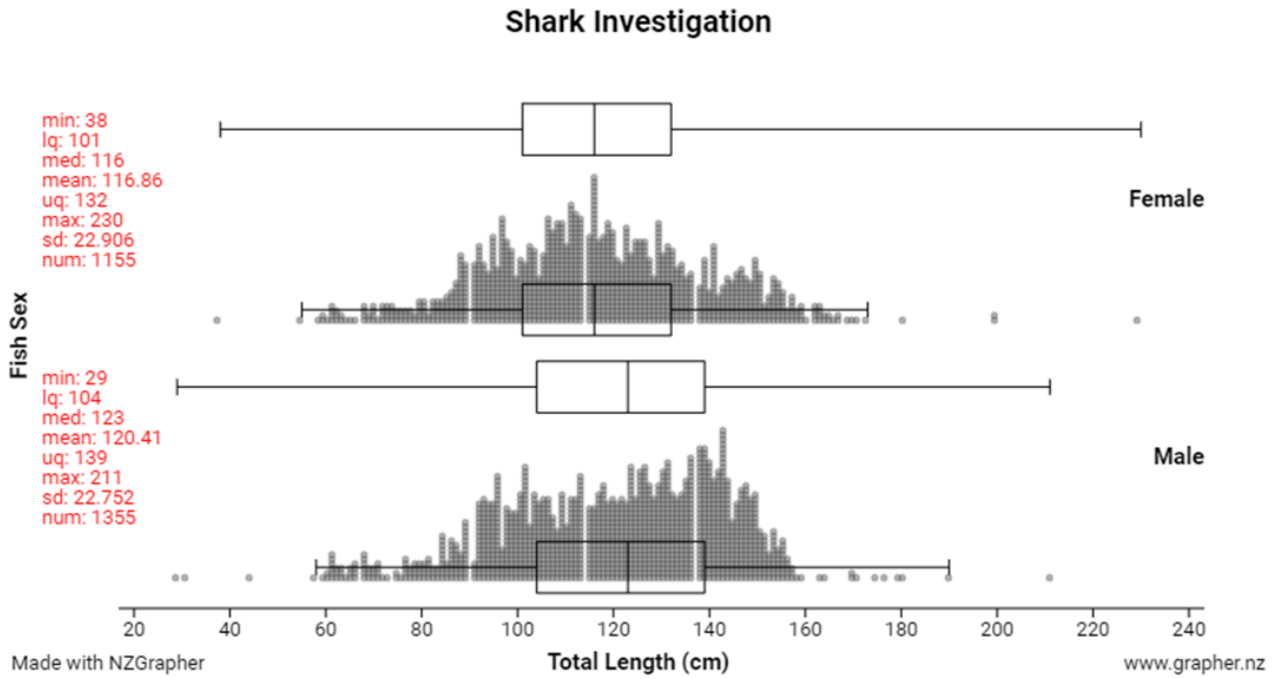
2)


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)

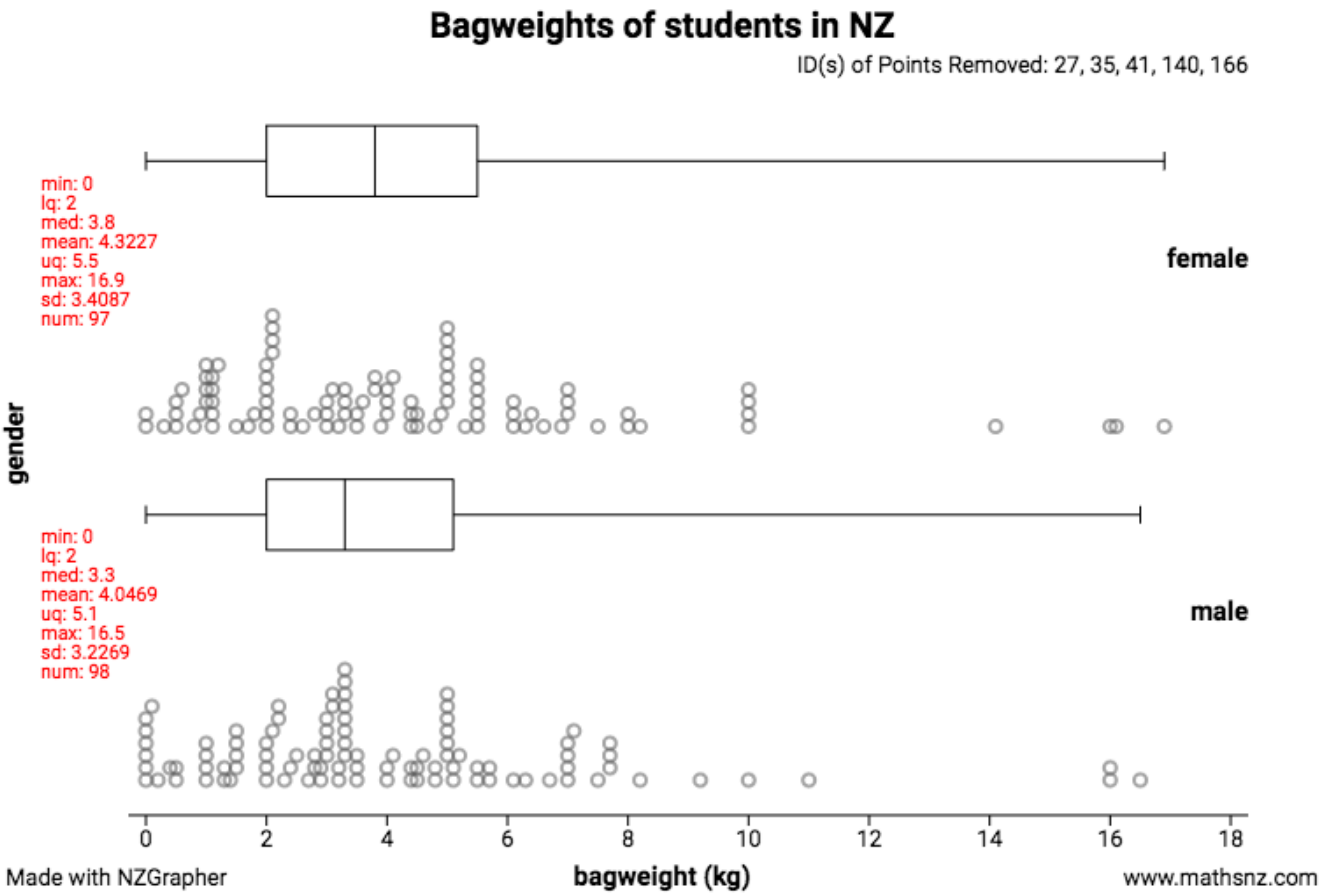

Car Investigation
min: 14000
lq: 18400
med: 21700
mean: 26063
uq: 29450
max: 80000
sd: 12936
num: 32
automatic
Tranmission
min: 7900
lq: 12050
med: 17400
mean: 19715
uq: 23600
max: 44600
sd: 9274.2
num: 61
manual
0
10000
20000
30000
40000
50000
60000
70000
80000
Price ($)
Made with NZGrapher
www.grapher.nz

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

Full Analysis

Example:



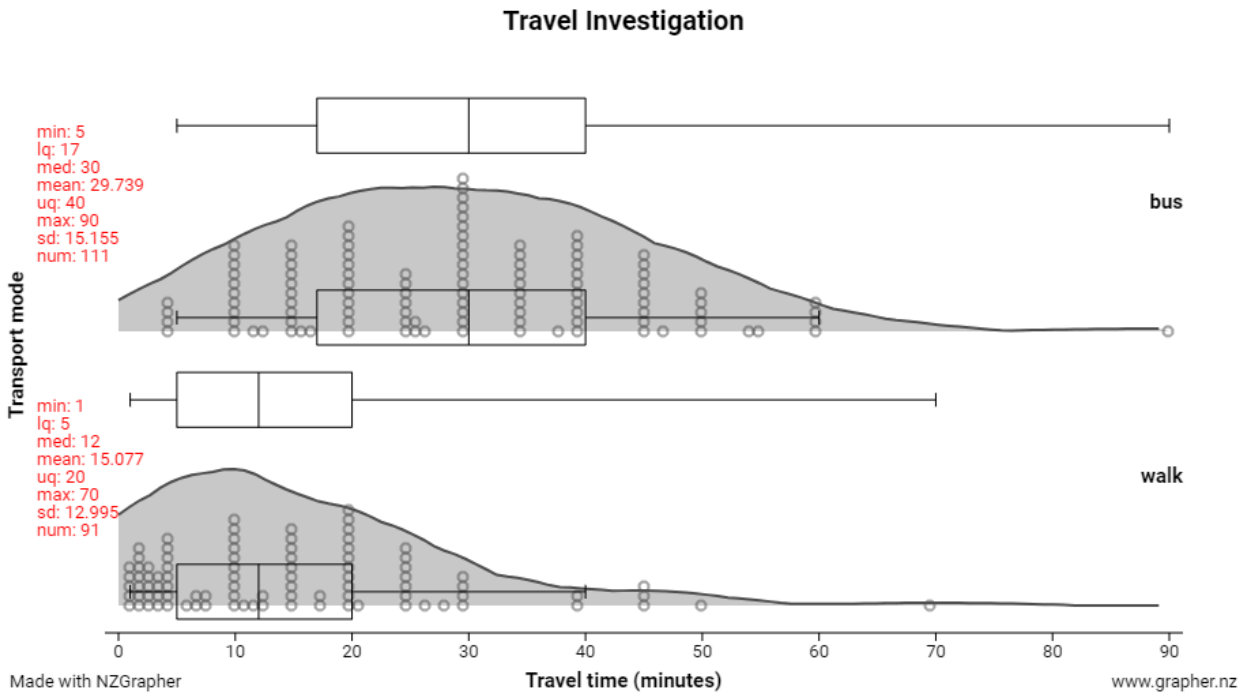
For my sample, I notice:

- 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 0.5kg. My evidence is that the median bag weight for females is around 3.8kg while the median bag weight for males is around 3.3kg.
- The spread of the middle 50% of females bag weights is slightly larger than the spread of males' bag weights, because the IQR of the females is approximately 3.5 kg compared to the IQR for males of 3.1 kg.

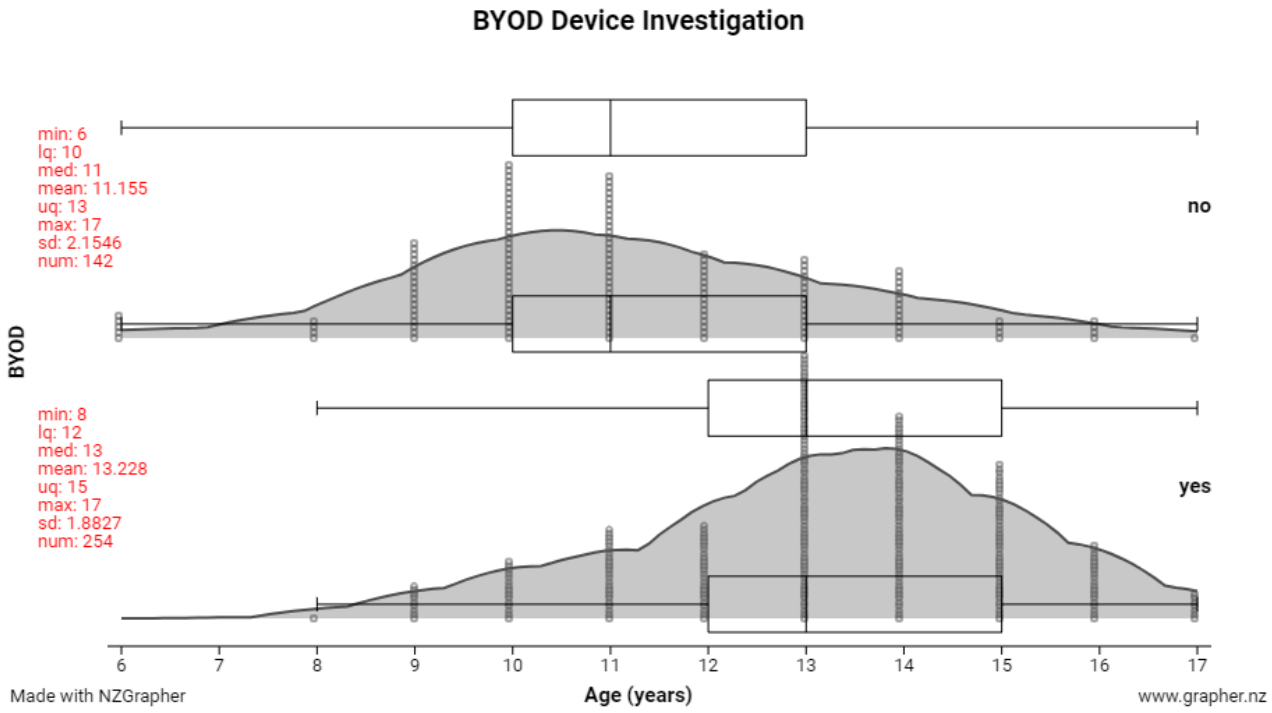
Exercise 15:

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

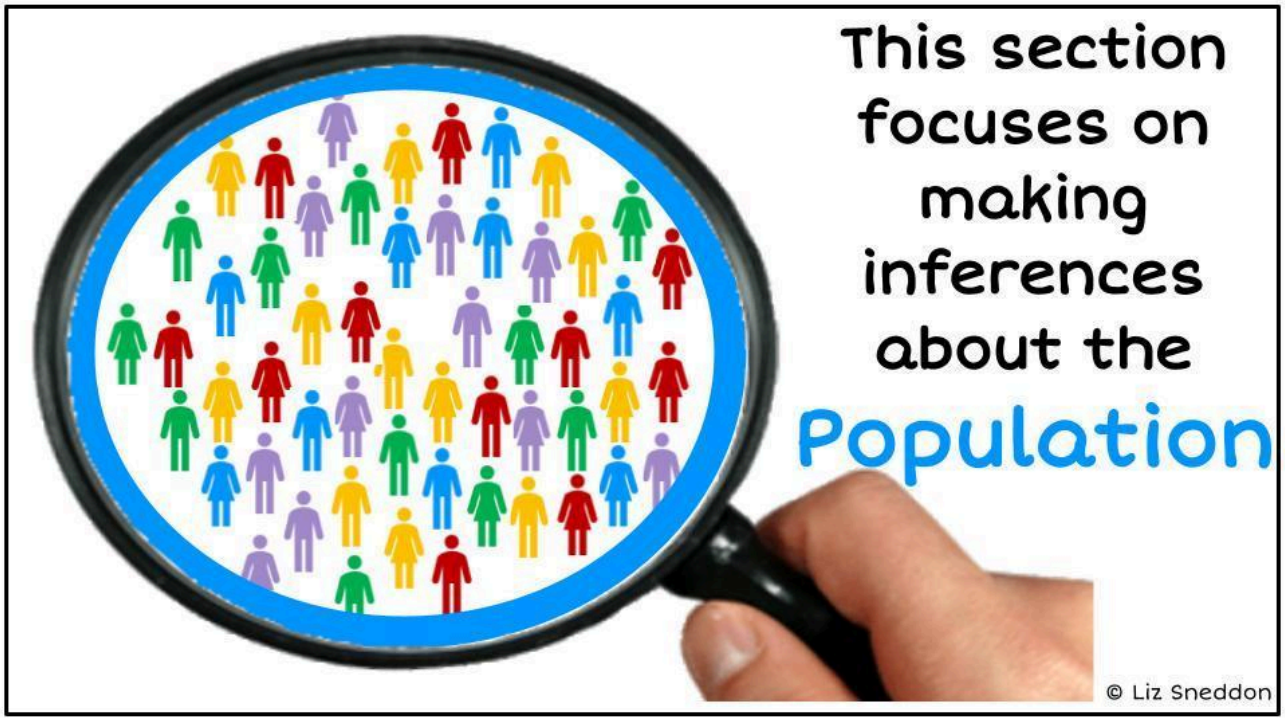
1) The dataset is about high school students and both the method of transport and the length of time it takes them to get to school.

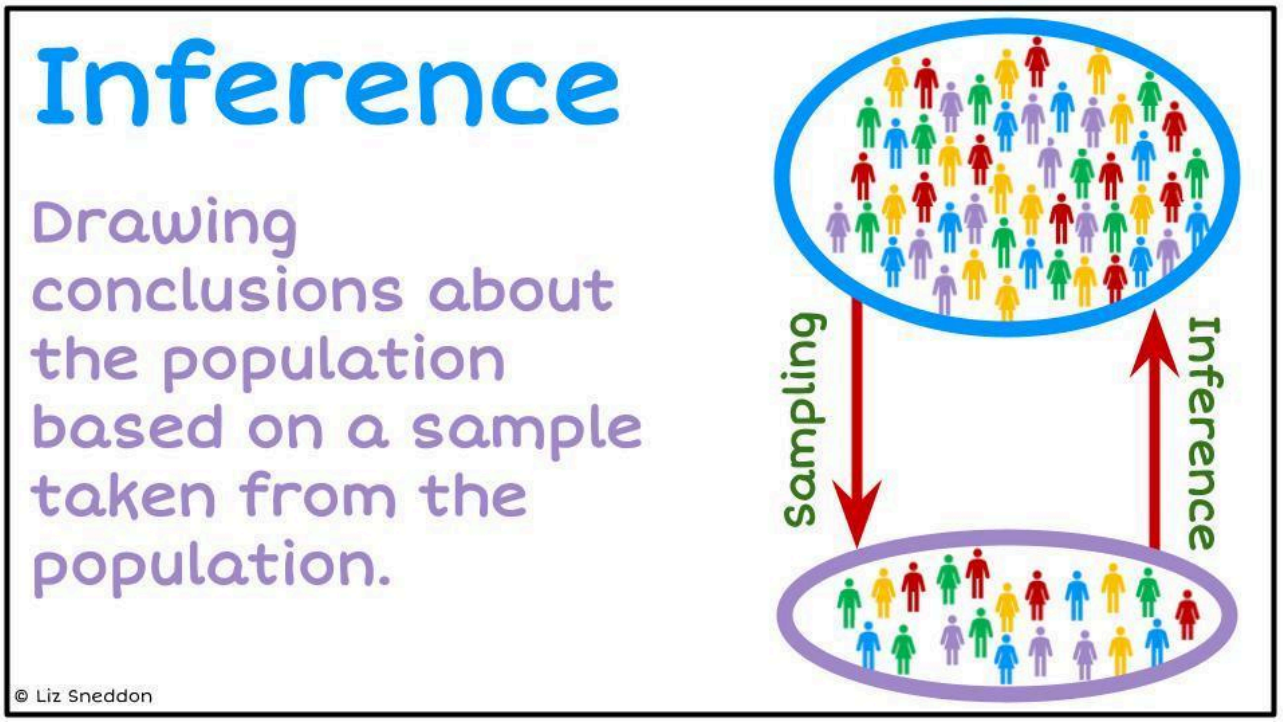


2) This investigation is looking into whether the ages of students who have a device is older than the age of students who don't have a device.



Conclusion





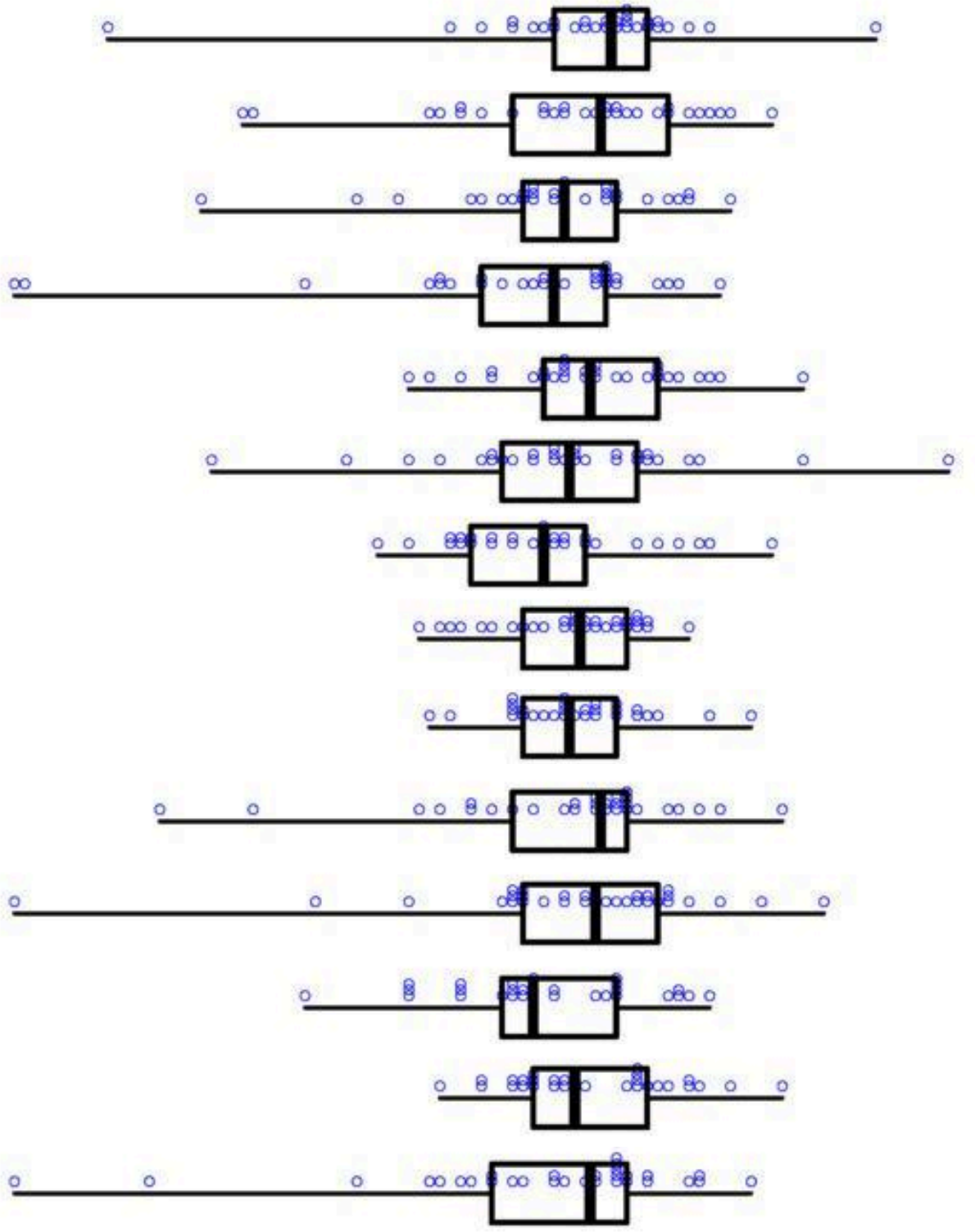
You need to include the following in your conclusion:

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

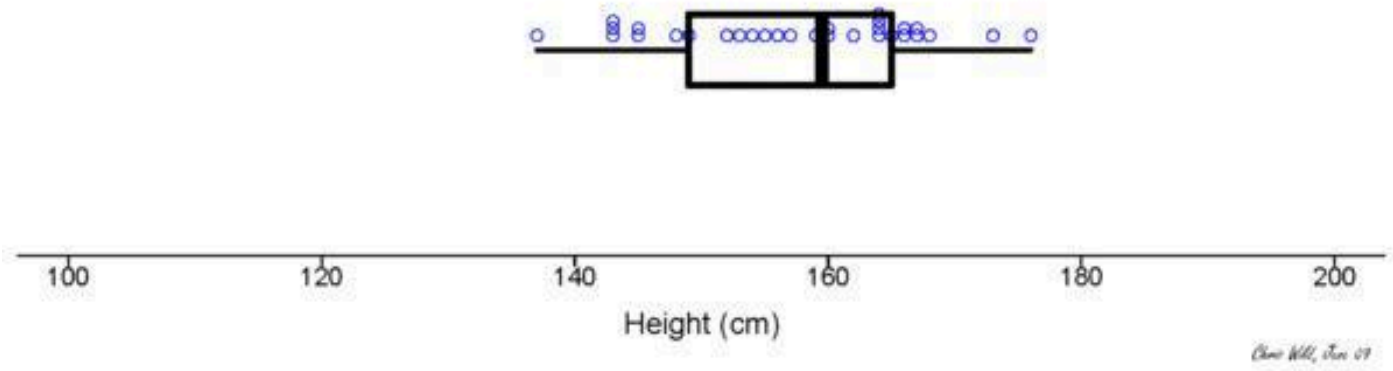
All of these things require an understanding of sampling variability, so we will start by exploring this in more depth.

Exercise 16:

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/



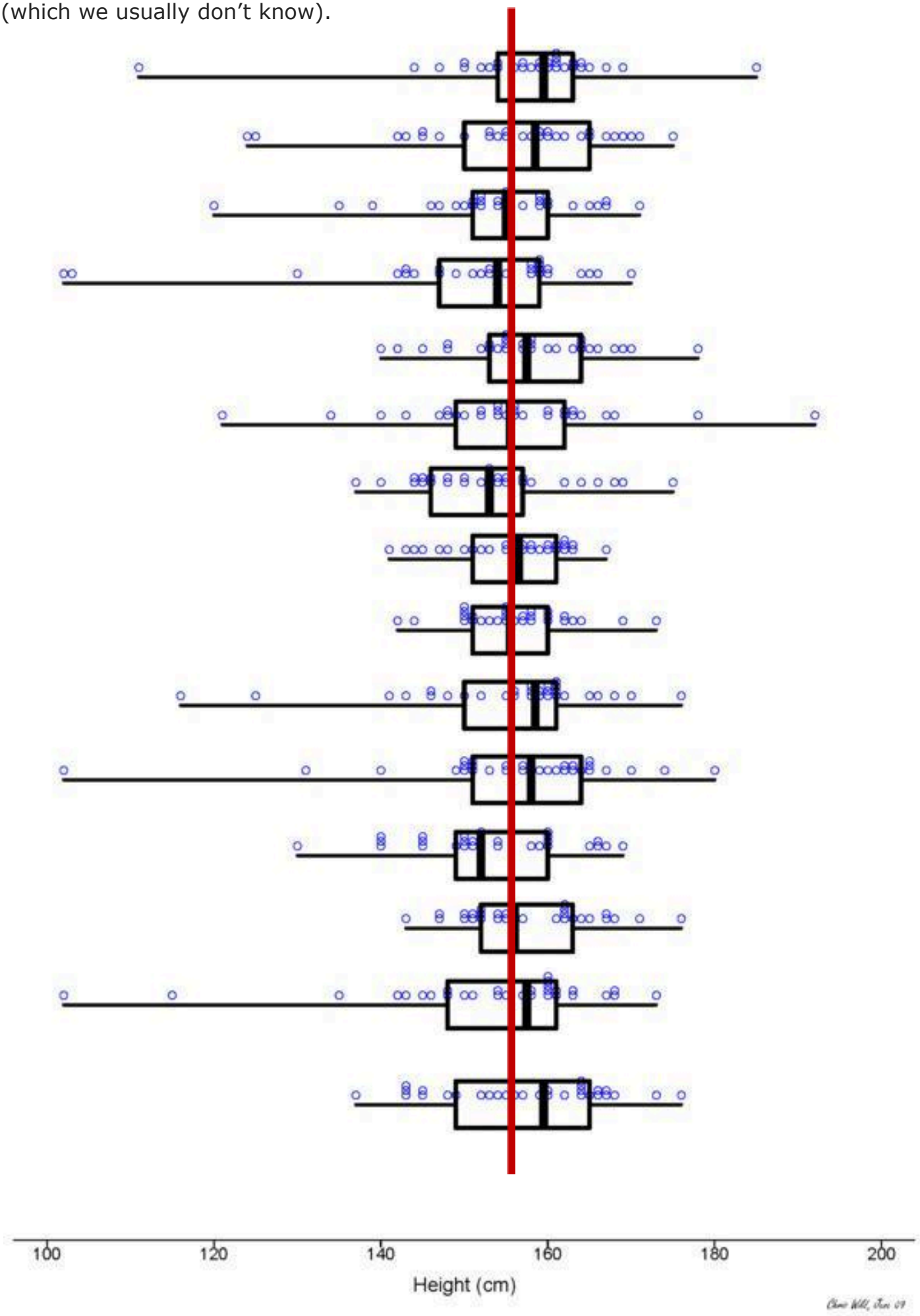
a) Highlight the medians for each of the samples above.

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

c) Explain why the medians are different for each of the samples on the previous page.

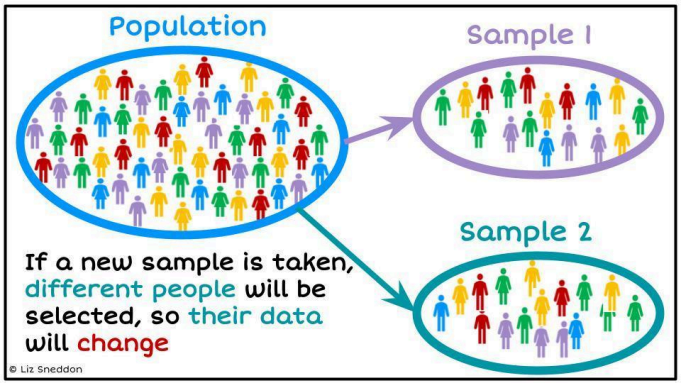
d) Find the smallest and largest estimate for the sample medians. Estimate how much do these values vary by, e.g., 1cm?

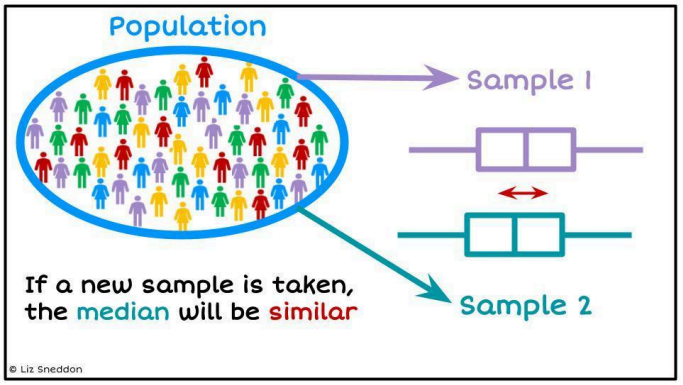
I've added the line drawn from the axis to the top, which is the population median (which we usually don't know).



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.





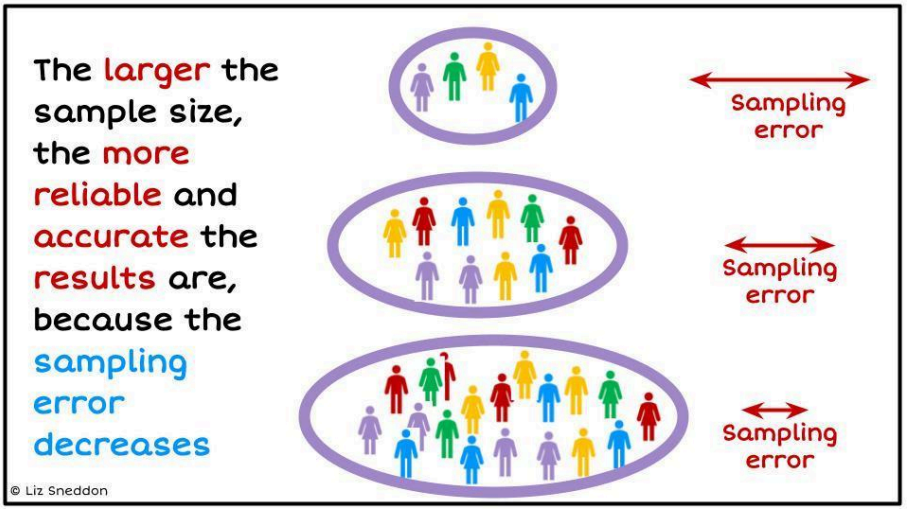
Sample size and sampling variation

Exercise 17:

Go to the animation: bit.ly/SamplingVariation

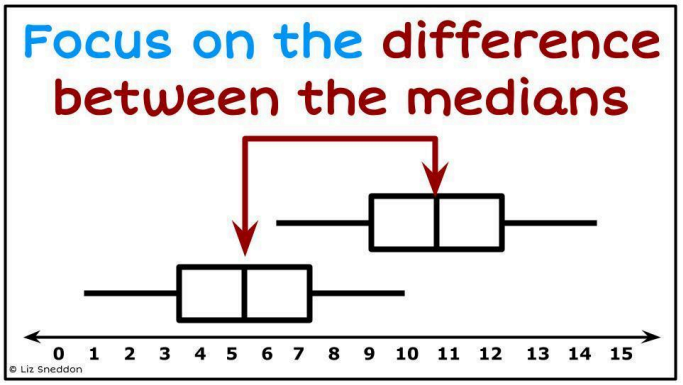
This shows you different samples being taken, with different sample sizes. Look at the blue line which is the median. What do you notice about the amount of variation in the medians with the different sample sizes?

This leads us to a big idea, that the larger the sample size, smaller the **sampling error** is.



Difference between the medians

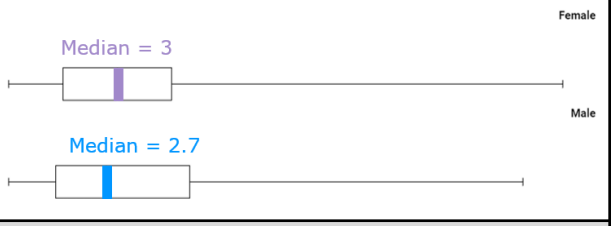
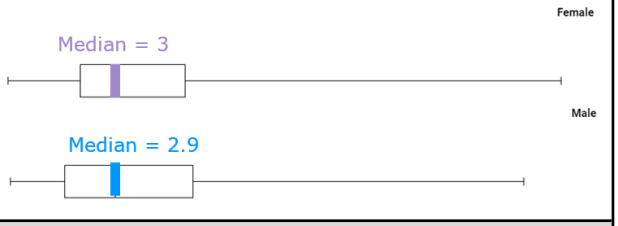
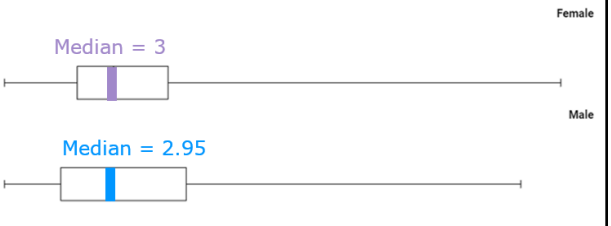
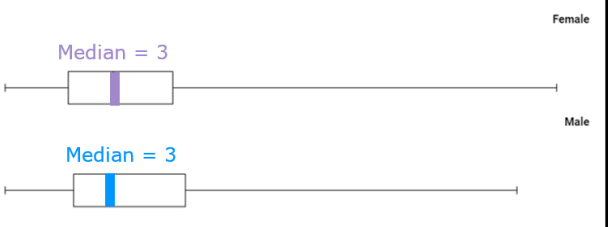
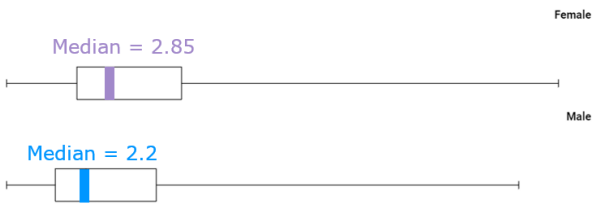
The next step is to focus on the differences between the medians.

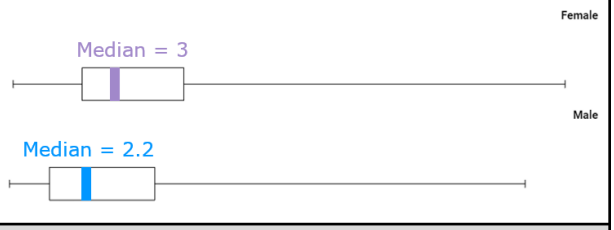
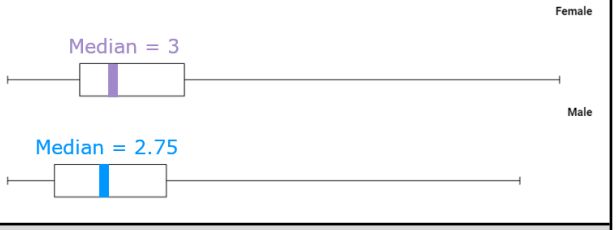
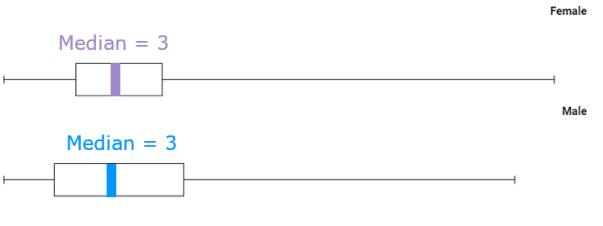
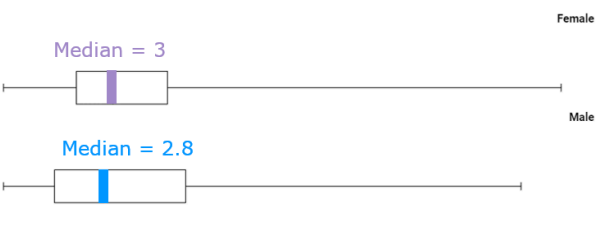


Exercise 18:

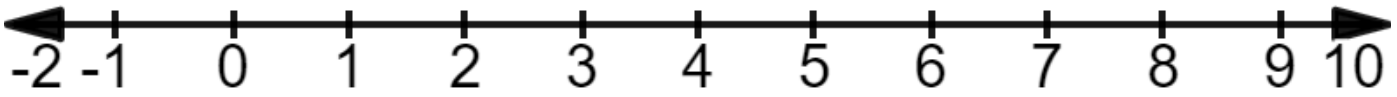
Each of the graphs below is a different sample of bag weights for teenagers at school in NZ. Calculate the difference between the medians for each graph.

Difference = Female median – Male median

1)		2)	
	Difference =		Difference =
3)		4)	
	Difference =		Difference =
5)		6)	
	Difference =		Difference =

7)		8)	
	Difference =		Difference =
9)		10)	
	Difference =		Difference =

Next, draw add the differences that you calculated onto the axis below.



Difference between medians (Female – Male)

This leads us to understanding the basis of the **Bootstrap distribution**.

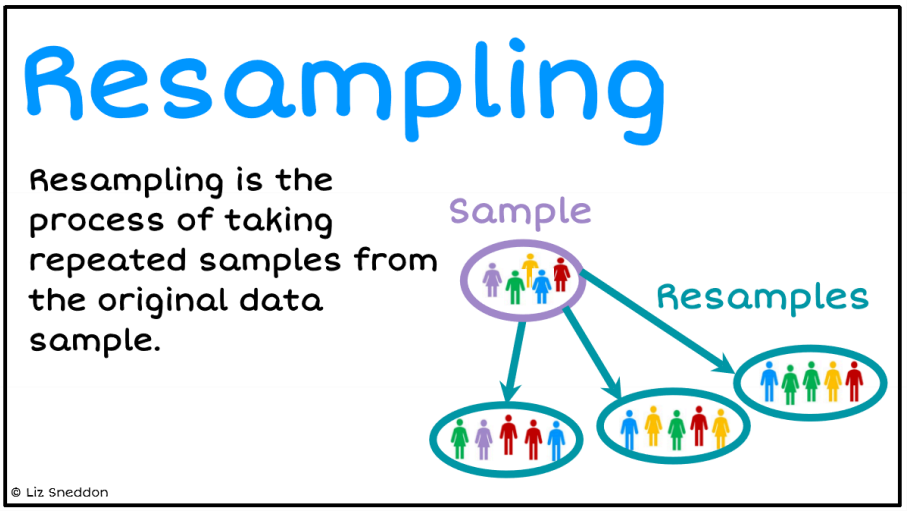
Bootstrapping Distribution

The diagram below shows the process for a bootstrap distribution.

We start by having a population that we are interested in investigating, and we then take as big a sample as time and money allows.

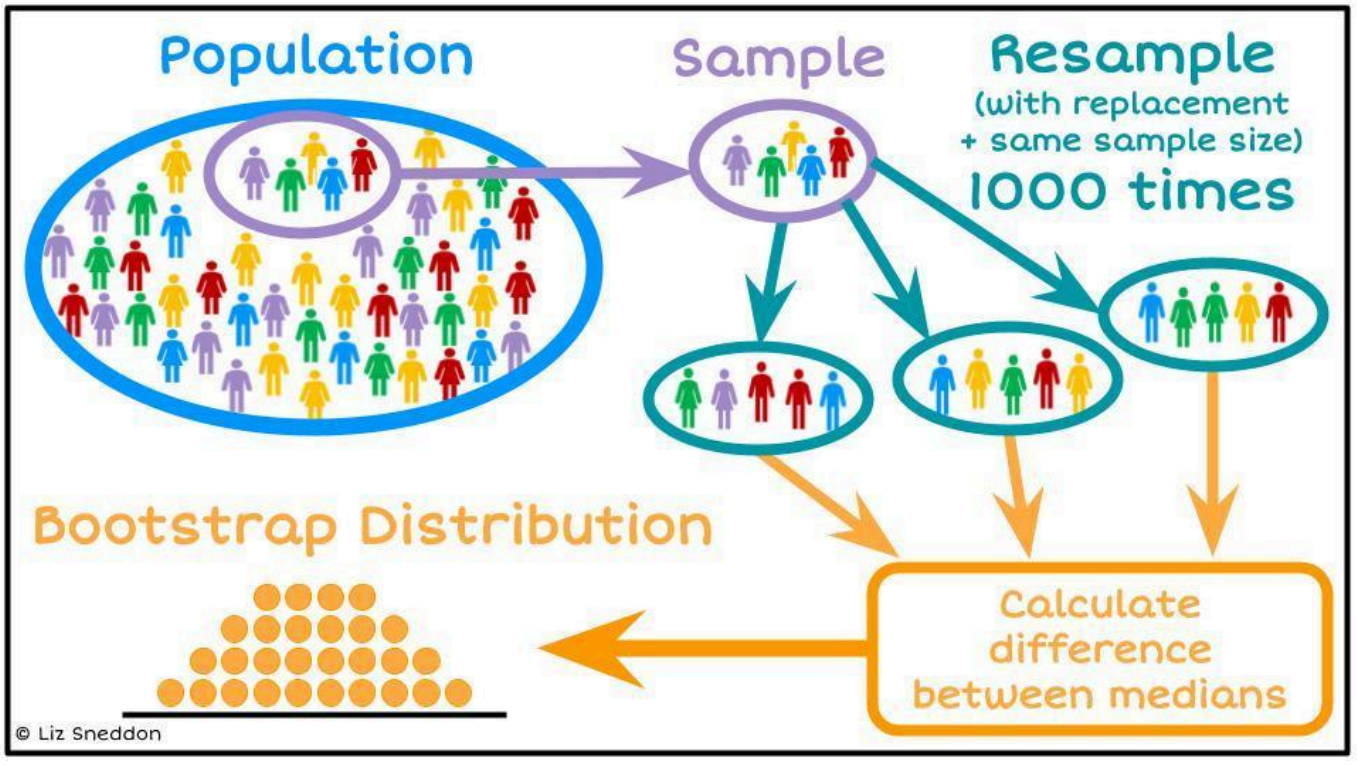
Then in order to improve the reliability and accuracy of the estimated results, we carry out a bootstrap process (this is a way of reusing the sample data to improve our estimates without needing to spend more money collecting samples - it does rely on the original sample being collected randomly and being representative of the population).

A re-sample is collected from the original sample, and people are selected randomly with replacement, where the sample size (n) is the same as the original sample.

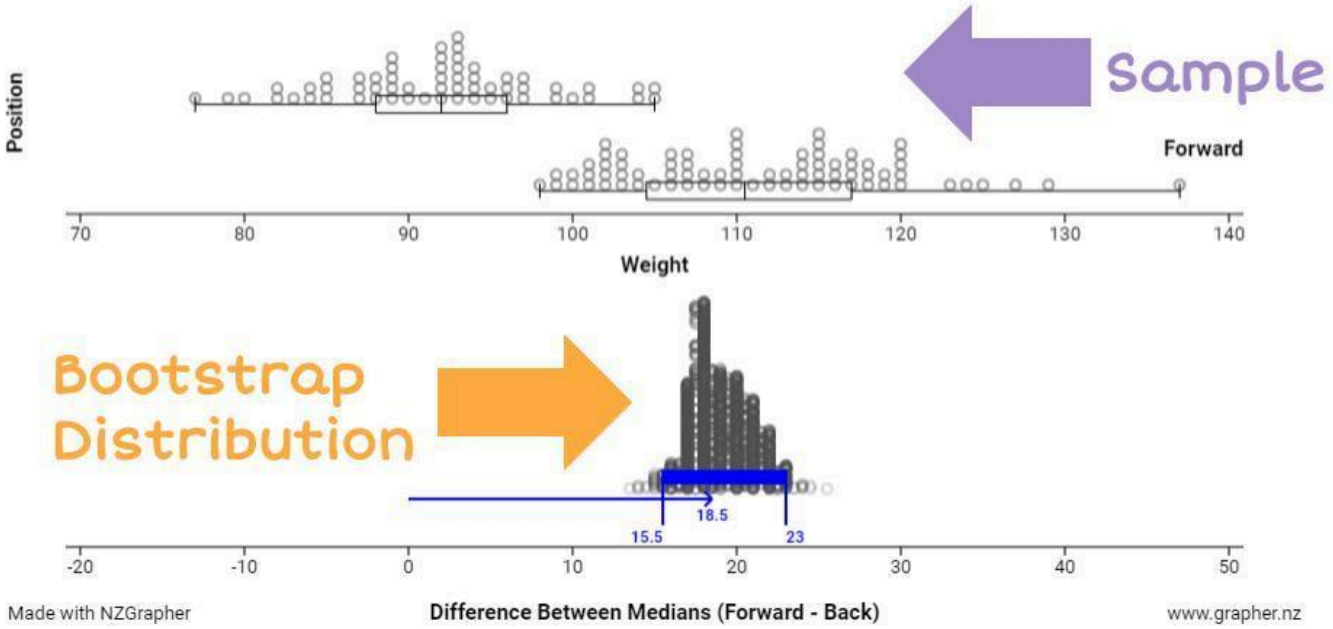


The difference between the medians of this re-sample is calculated and added to a dot plot.

This re-sampling process is repeated 1000 times to form the **Bootstrap Distribution**.

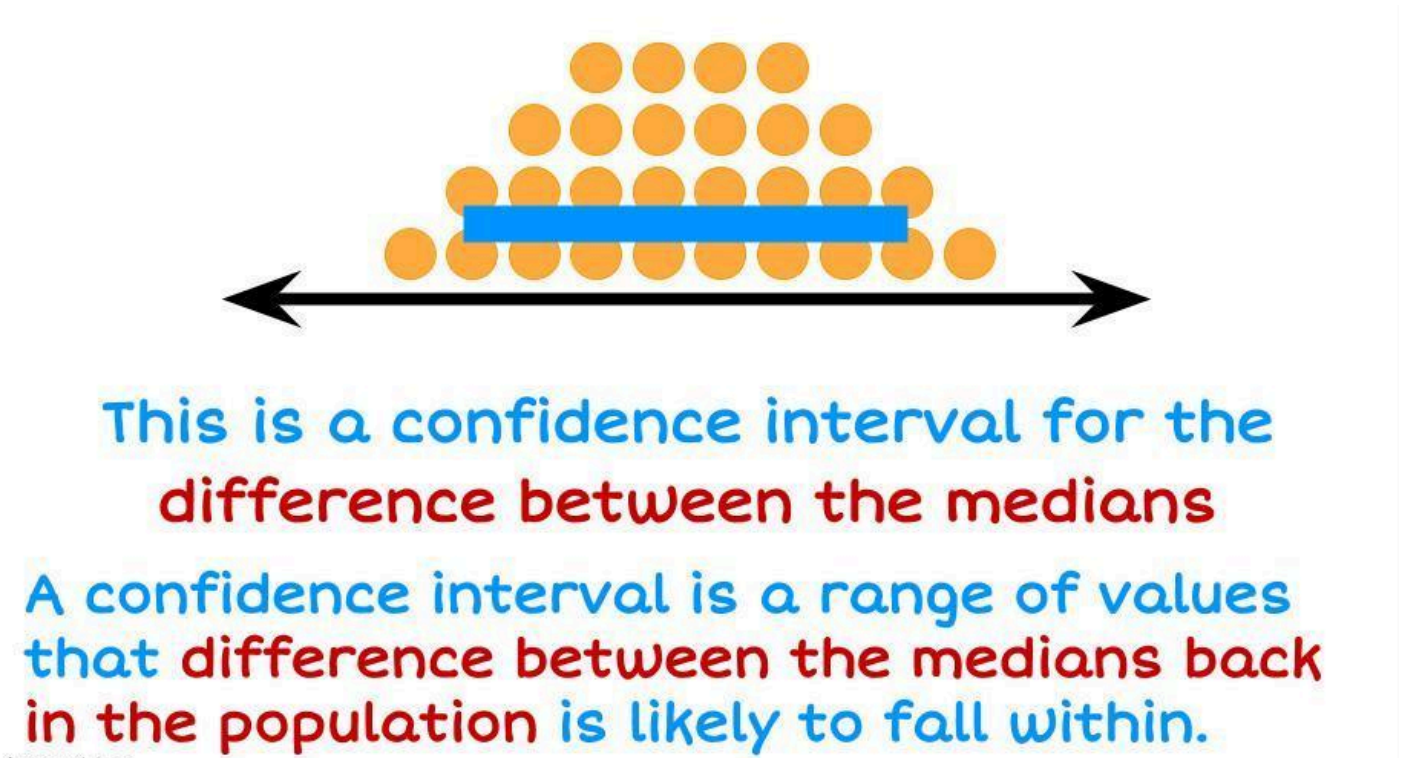


Example:

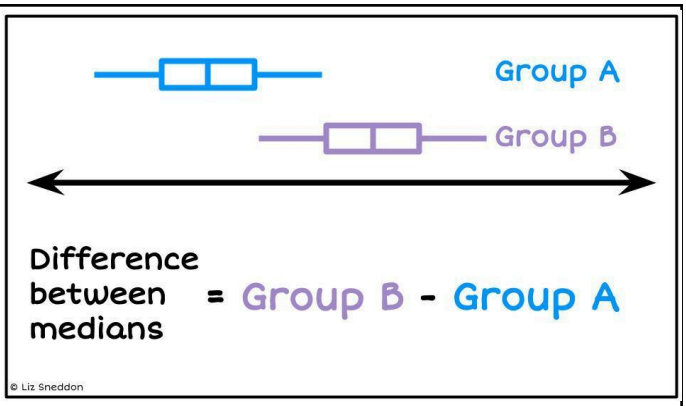


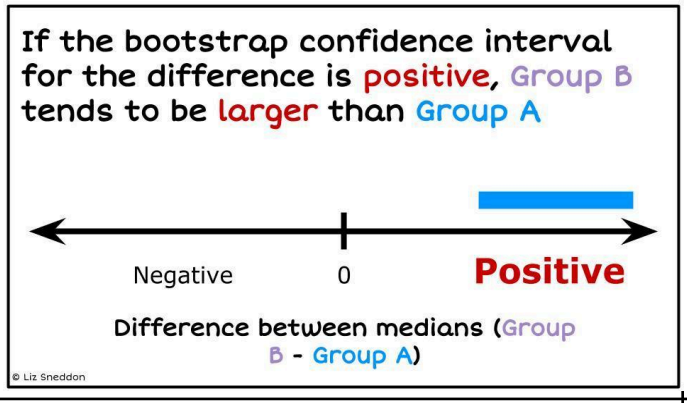
Bootstrapping Confidence Intervals

Once the Bootstrap distribution is formed, a confidence interval is then added.



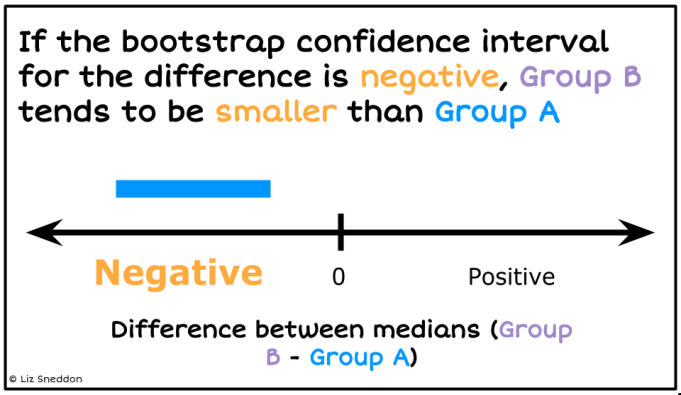
Always check the label on the axis underneath the bootstrap distribution. We will use this to work out what a positive or negative confidence interval limit means.





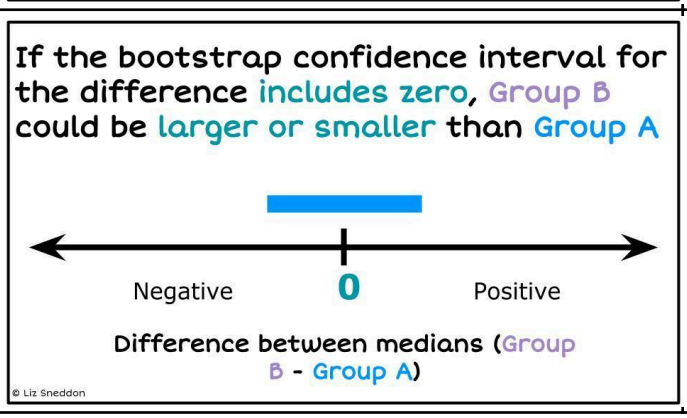
E.g.

Group B median is likely to be between _____ and _____ **more** than Group A median, for the population.



E.g.

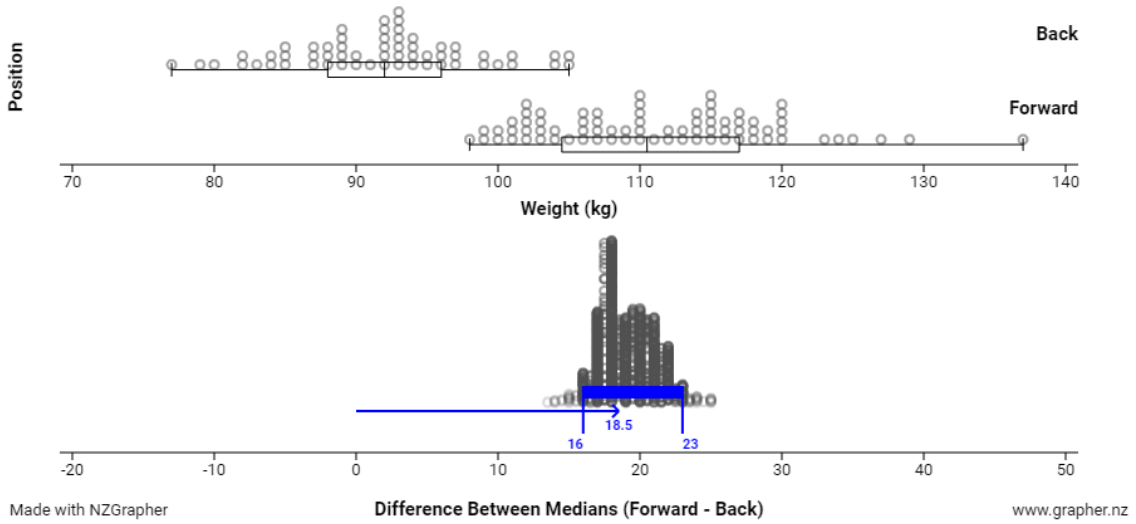
Group B median is likely to be between _____ and _____ **less** than Group A median, for the population.



E.g.

Group B median is likely to be between _____ **less** and _____ **more** than Group A median, for the population.

Example 1:

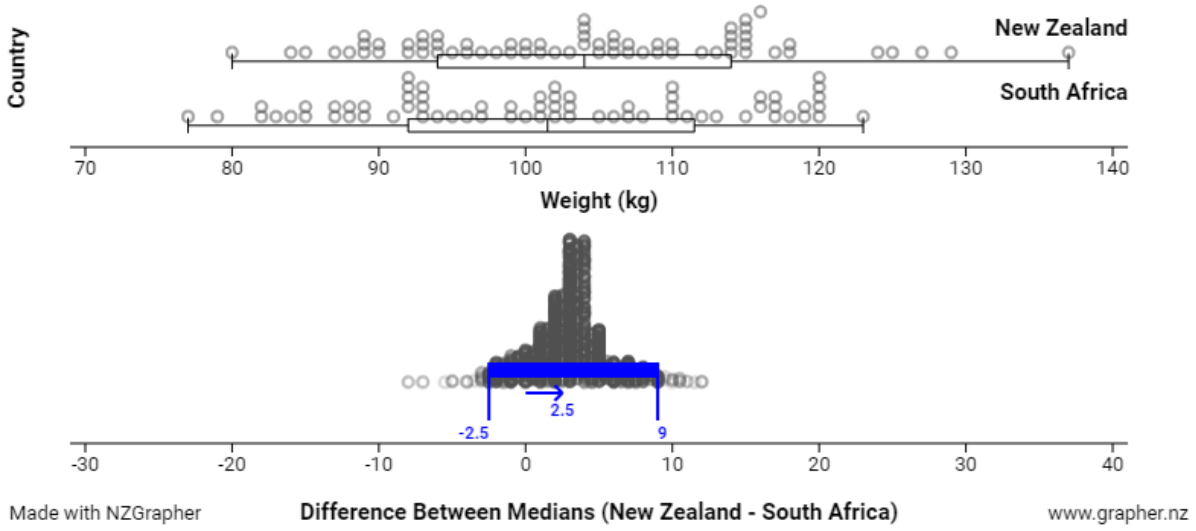


From the bootstrapping confidence interval, I am fairly sure that the median weight of forwards will be between 16 kg and 23 kg more than median weight of backs for the **POPULATION** of ALL top players listed on the website http://www.rugby-sidestep-central.com/.

Example 2:

Problem:

What is the difference between the median weight (kg) of rugby players in New Zealand and South Africa from ALL top players listed on the website http://www.rugby-sidestep-central.com/



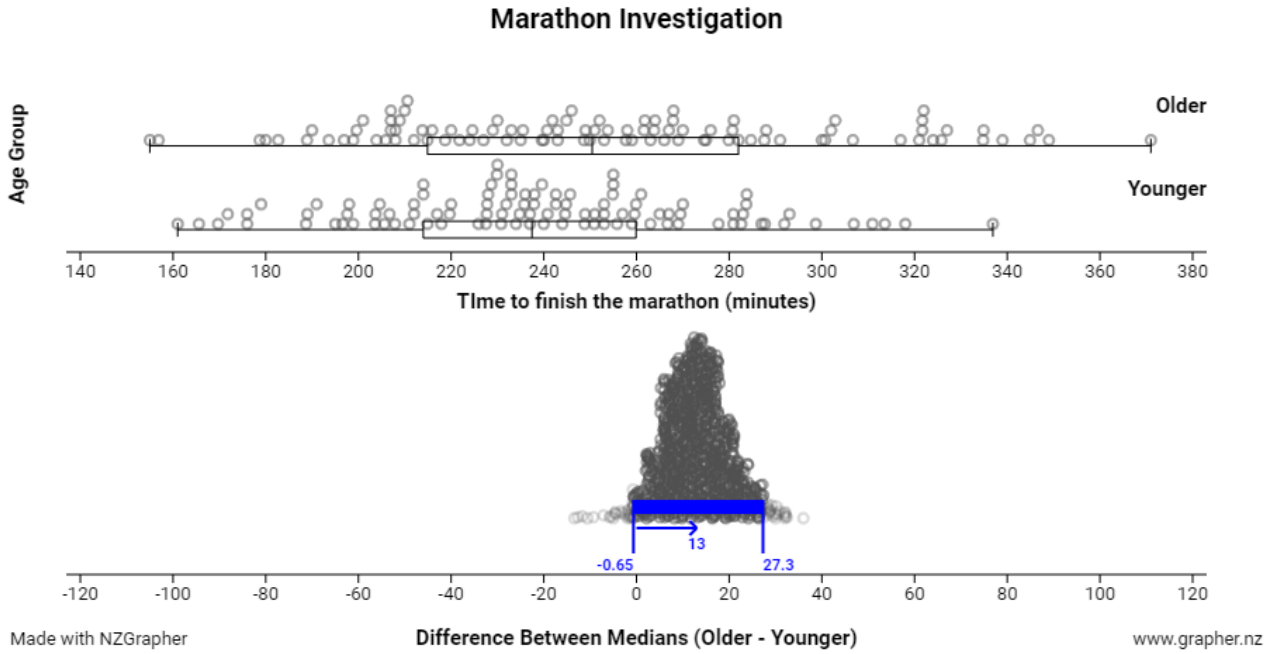
From the bootstrapping confidence interval I am fairly sure that the median weight of New Zealand rugby players could be up to 9 kg more or up to 2.5kg less than median weight of South African rugby players, for the **POPULATION** of ALL top players listed on the website http://www.rugby-sidestep-central.com/

Exercise 19:

For the samples below, interpret the bootstrap confidence intervals.

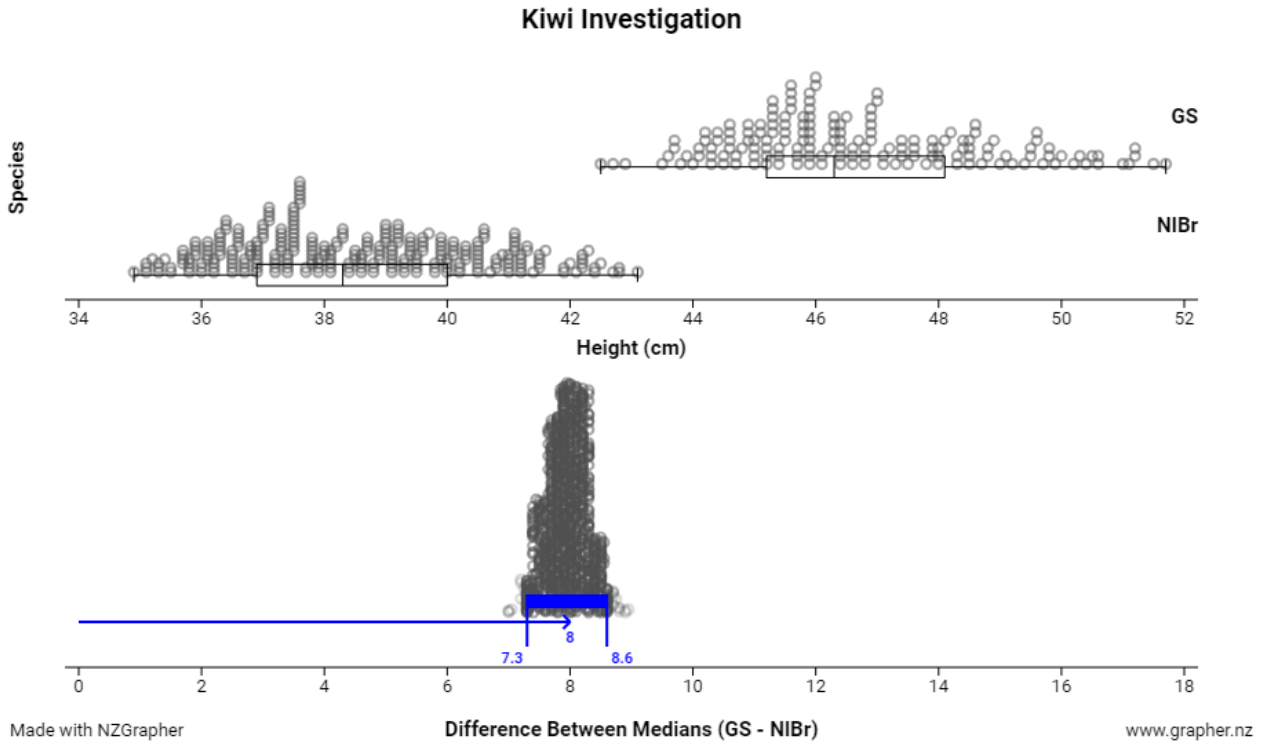
1) **Problem:**

What is the difference between the median length of time it takes older (over 40 years old) and younger (under 40 years old) people to complete a marathon, for ALL marathon runners in NZ?



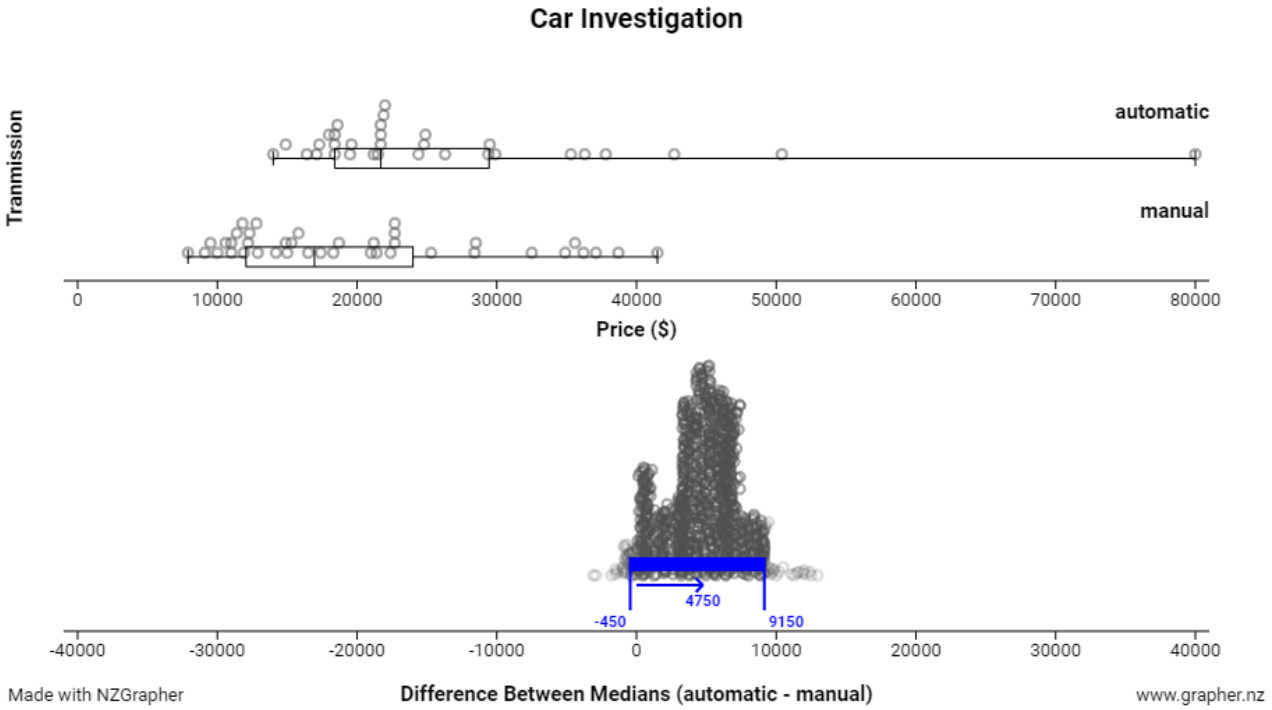
2) **Problem:**

What is the difference between the median height of the great spotted kiwi and the North Island Brown kiwis, for all kiwi birds in NZ?



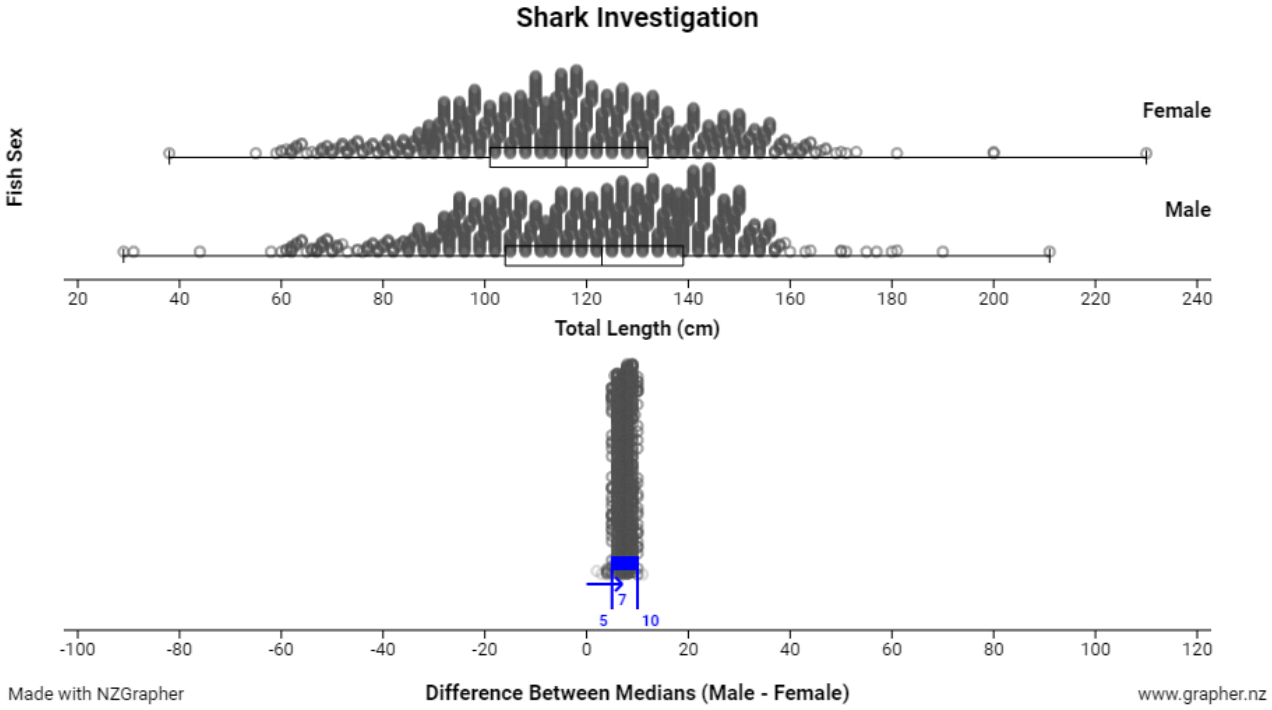
3) **Problem:**

What is the difference between the median price of automatic and manual cars, for all cars sold in America in 2012?



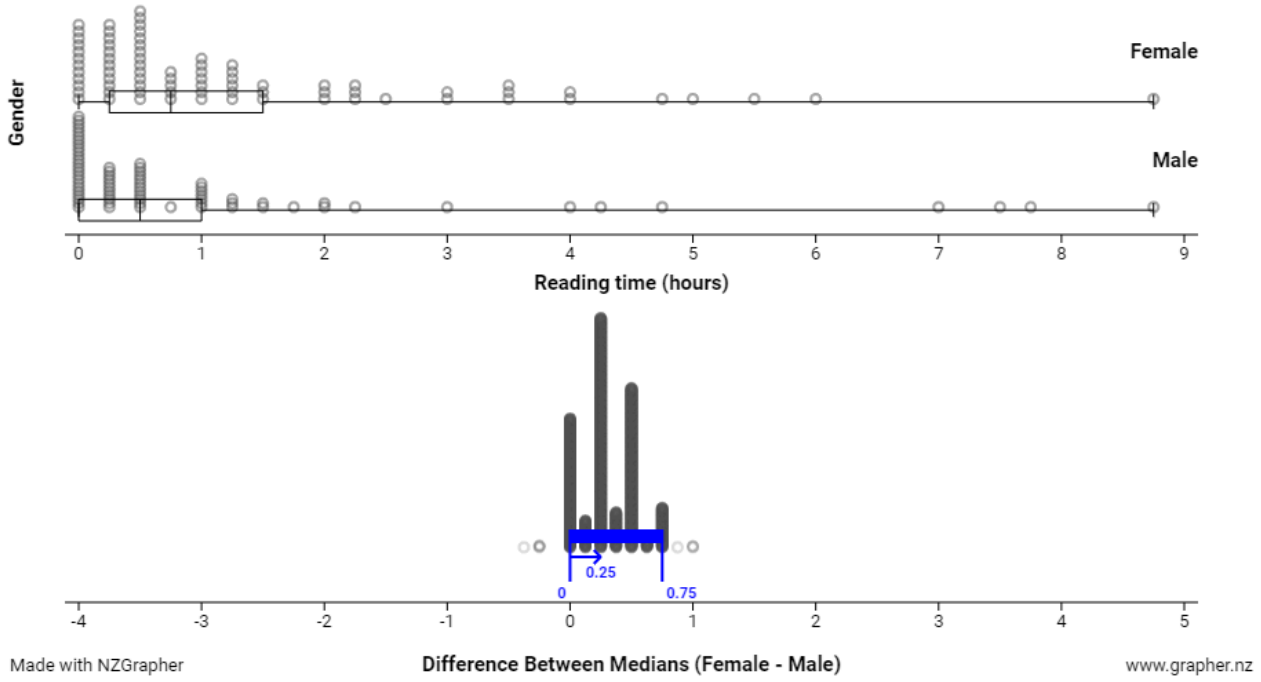
4) **Problem:**

What is the difference between the median length of male sharks compared with female sharks, for all sharks from the MPI (Ministry of Primary Industries – a government organisation) database in NZ.



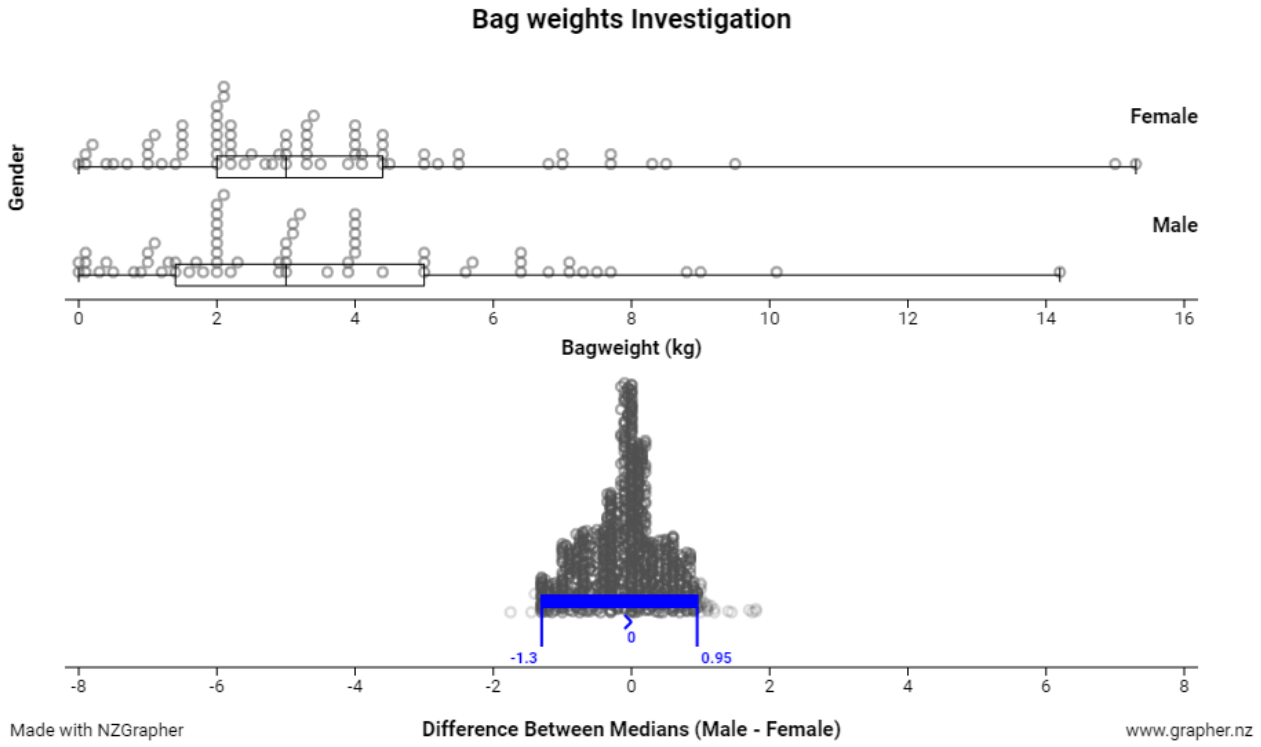
5) **Problem:**

What is the difference between the amount of time that girls and boys spend reading, for all teenagers at school in NZ?



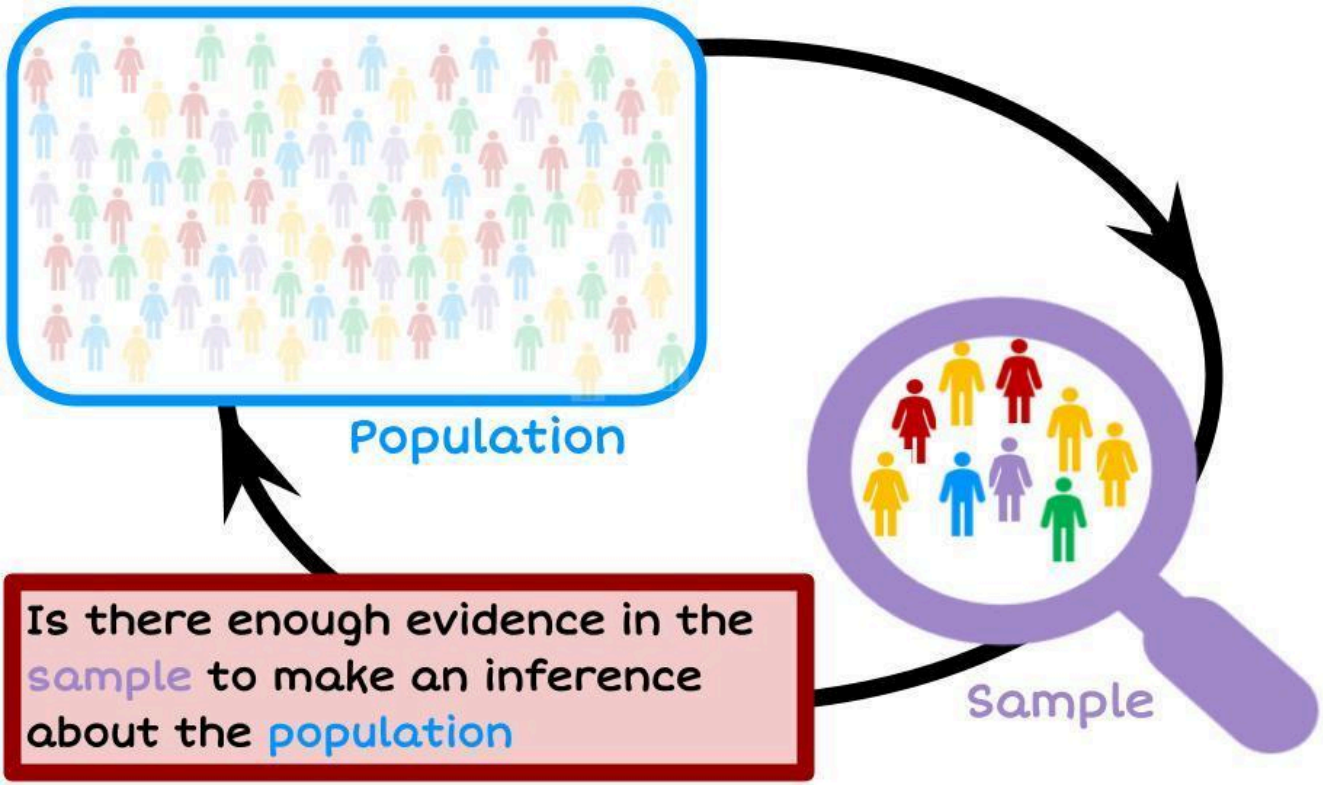
6) **Problem:**

What is the difference between the median weight of school bags for female and male teenagers in NZ?

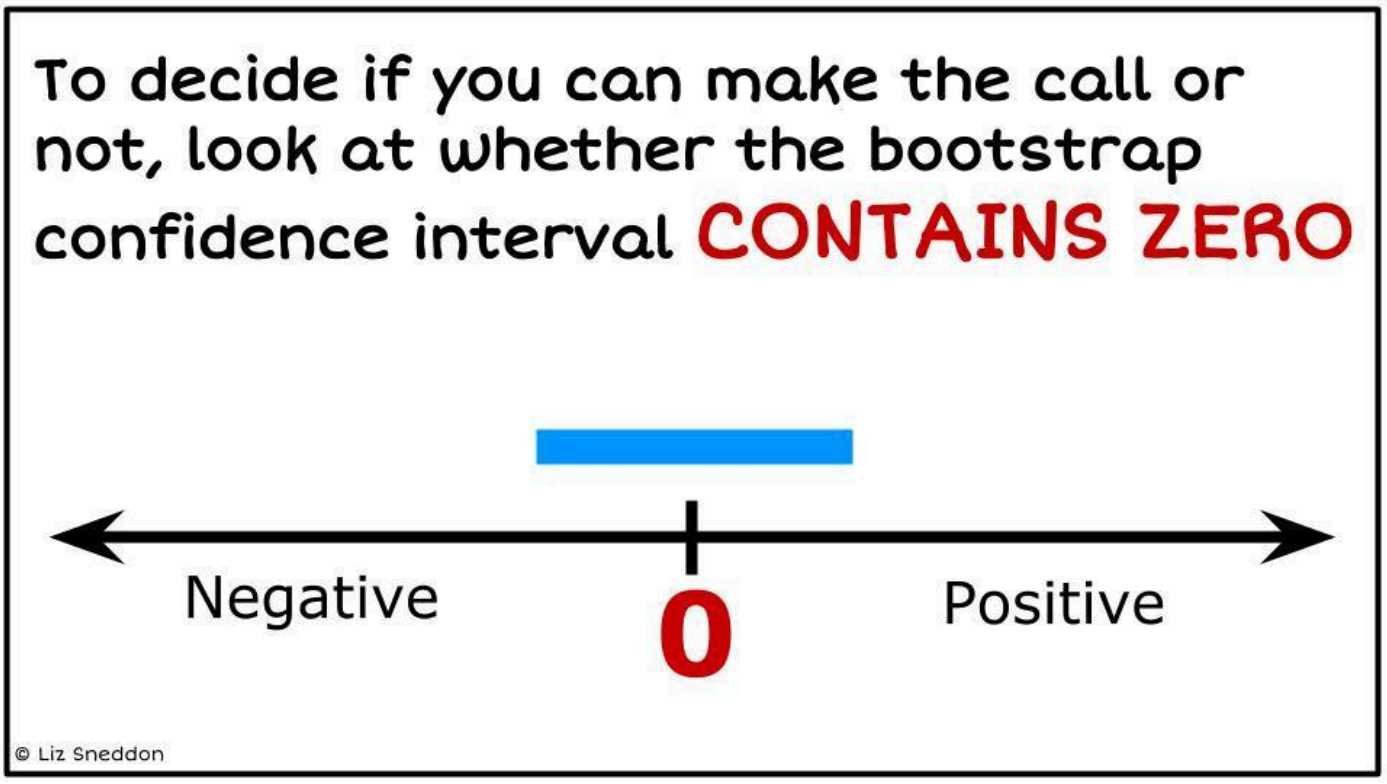


Is there enough evidence?

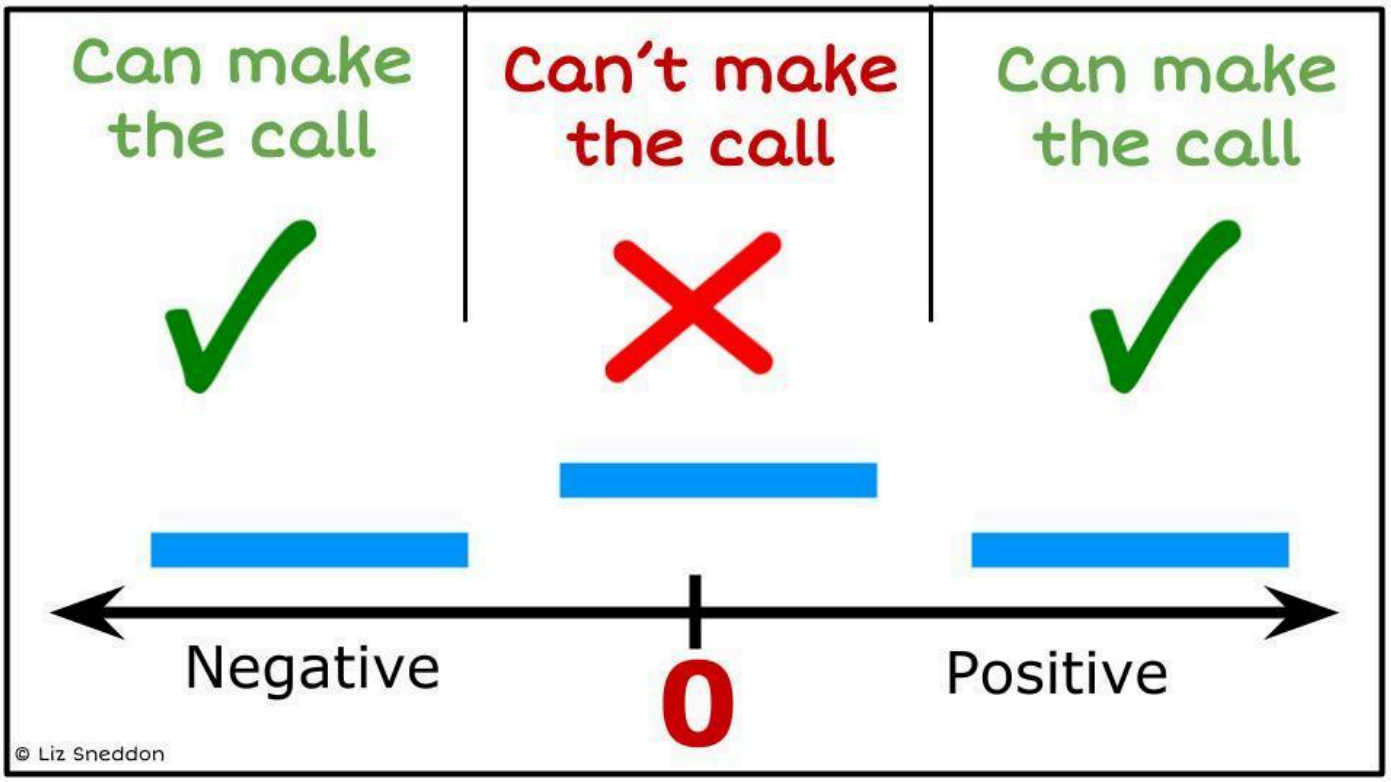
We are looking to see if we have **enough evidence** of a difference between the medians, **back in the population**.



Making the call



If the bootstrap confidence interval **contains zero**, this suggests that the **medians could be the same**, therefore we **can't make the call**.



Answering the investigation question

To answer the comparison question you need:

- Categorical variable,
- Numerical variable,
- Compare the medians,
- Direction,
- Population **(ALL)**.

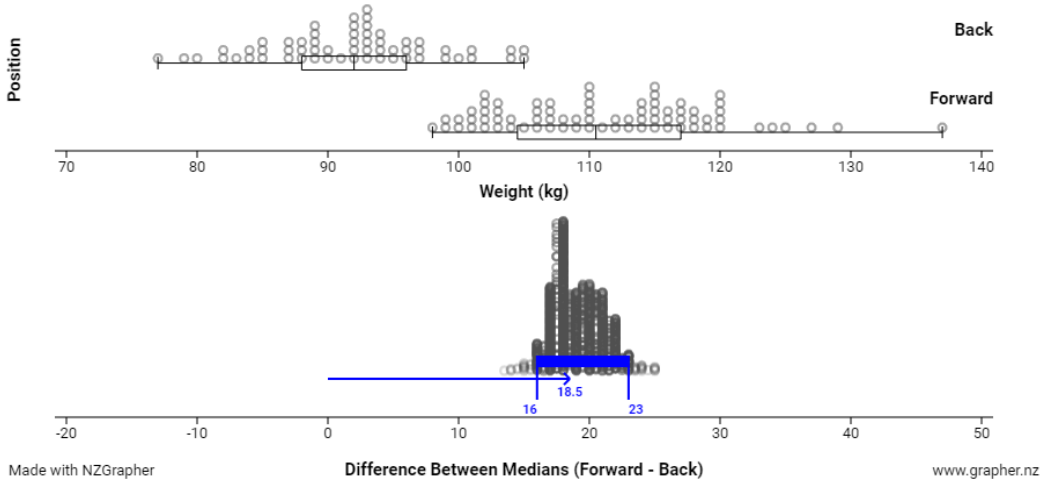
© Liz Sneddon

Don't forget to include the **direction** in your answer.

Example:

Problem:

What is the difference between the median weight (kg) of forward and back rugby players from ALL top players listed on the website http://www.rugby-sidestep-central.com.



Conclusion:

The bootstrap confidence interval **DOES NOT** contain zero, so we **CAN** make the call.

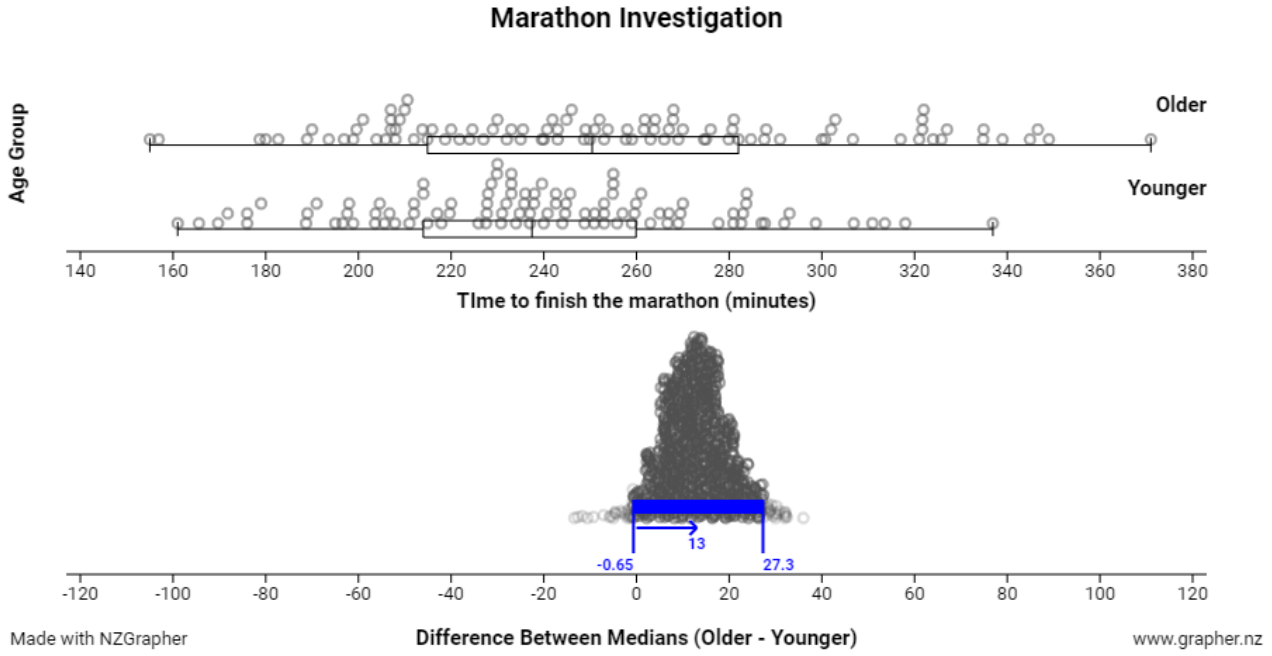
I can make the call, so I **DO** have enough evidence that the median weight of forwards is heavier than the median weight of backs, for **ALL** top players listed on the website http://www.rugby-sidestep-central.com.

Exercise 20:

For the investigations below, answer the investigation question, justifying your conclusion.

1) **Problem:**

What is the difference between the median length of time it takes older (over 40 years old) and younger (under 40 years old) people to complete a marathon, for ALL marathon runners in NZ?

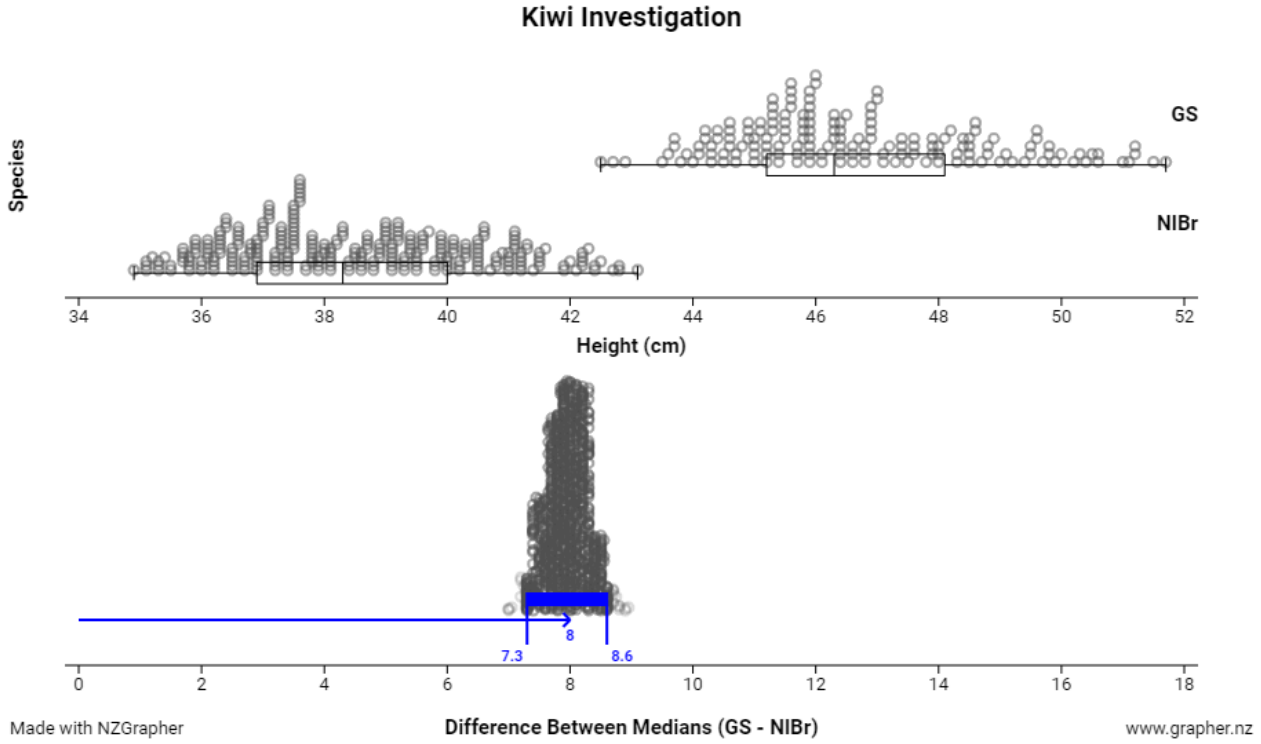


Does the CI include zero?	Yes / No
Can you make the call?	Yes / No

Answer the investigation question:

2) **Problem:**

What is the difference between the median height of the Great Spotted kiwi and the North Island Brown kiwis, for all kiwi birds in NZ?

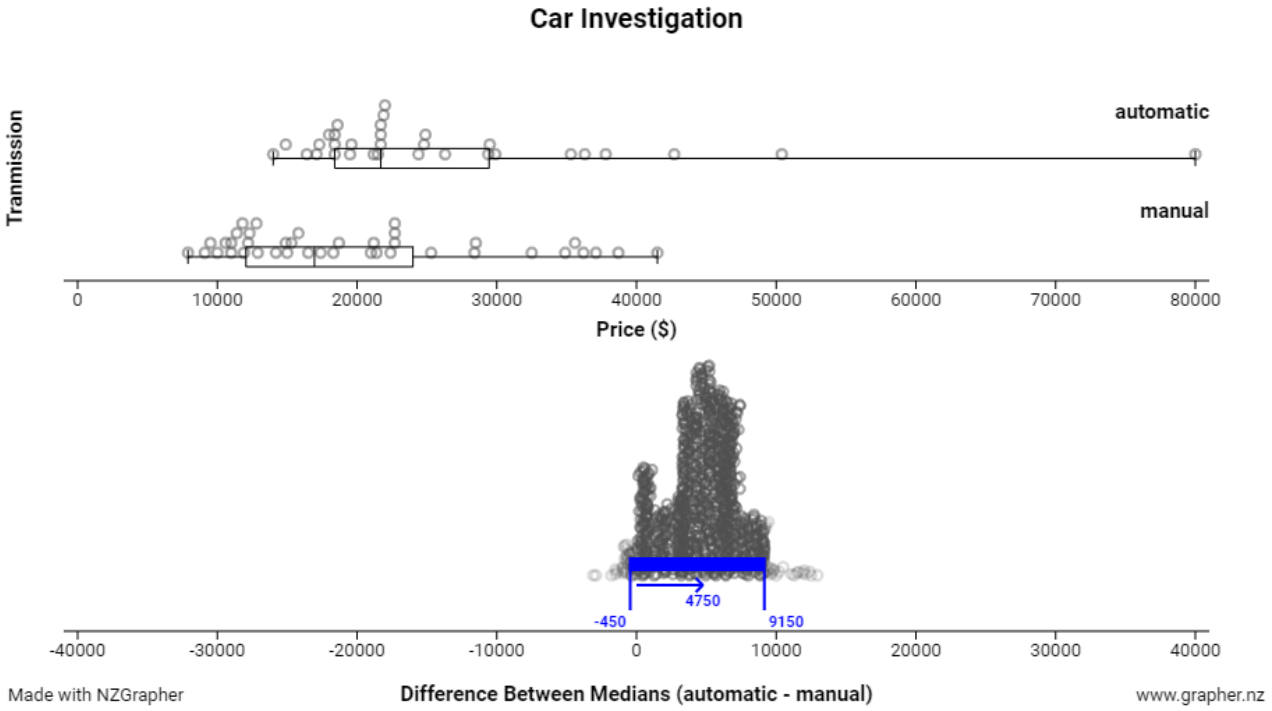


Does the CI include zero?	Yes / No
Can you make the call?	Yes / No

Answer the investigation question:

3) **Problem:**

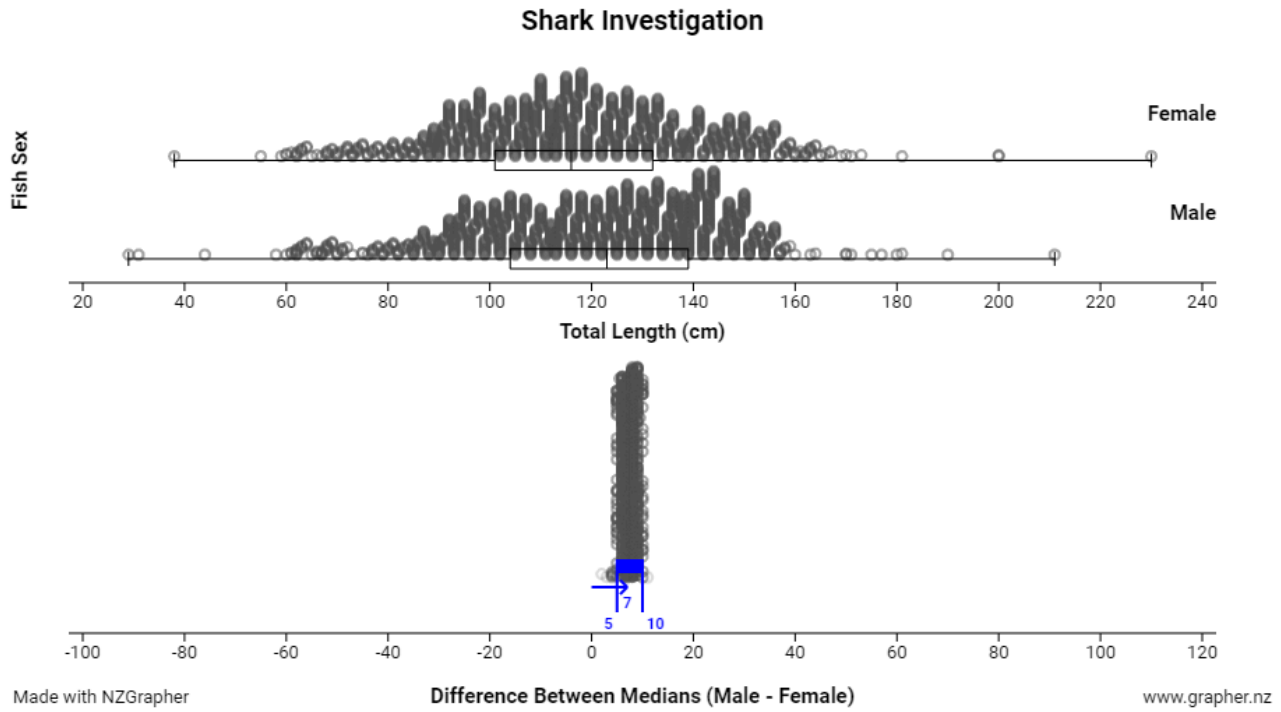
What is the difference between the median price of automatic and manual cars, for all cars sold in America in 2012?



Answer the investigation question:

4) **Problem:**

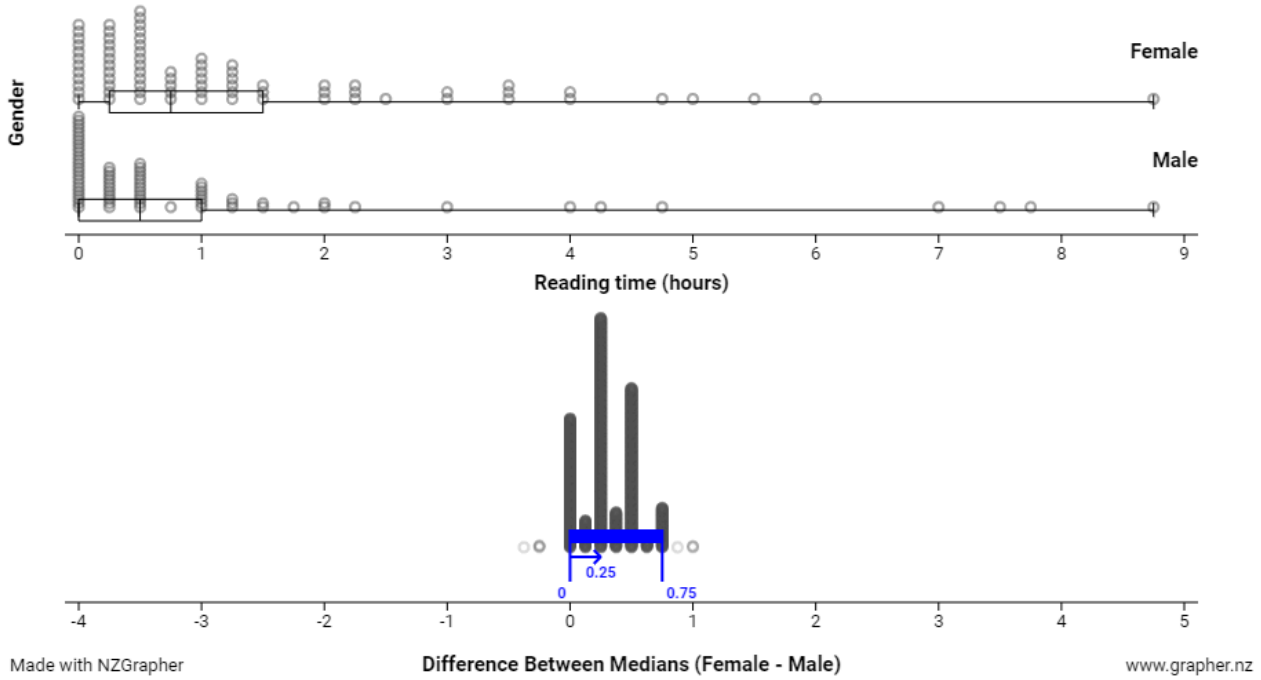
What is the difference between the median length of male sharks compared with female sharks, for all sharks from the MPI (Ministry of Primary Industries – a government organisation) database in NZ.



Answer the investigation question:

5) **Problem:**

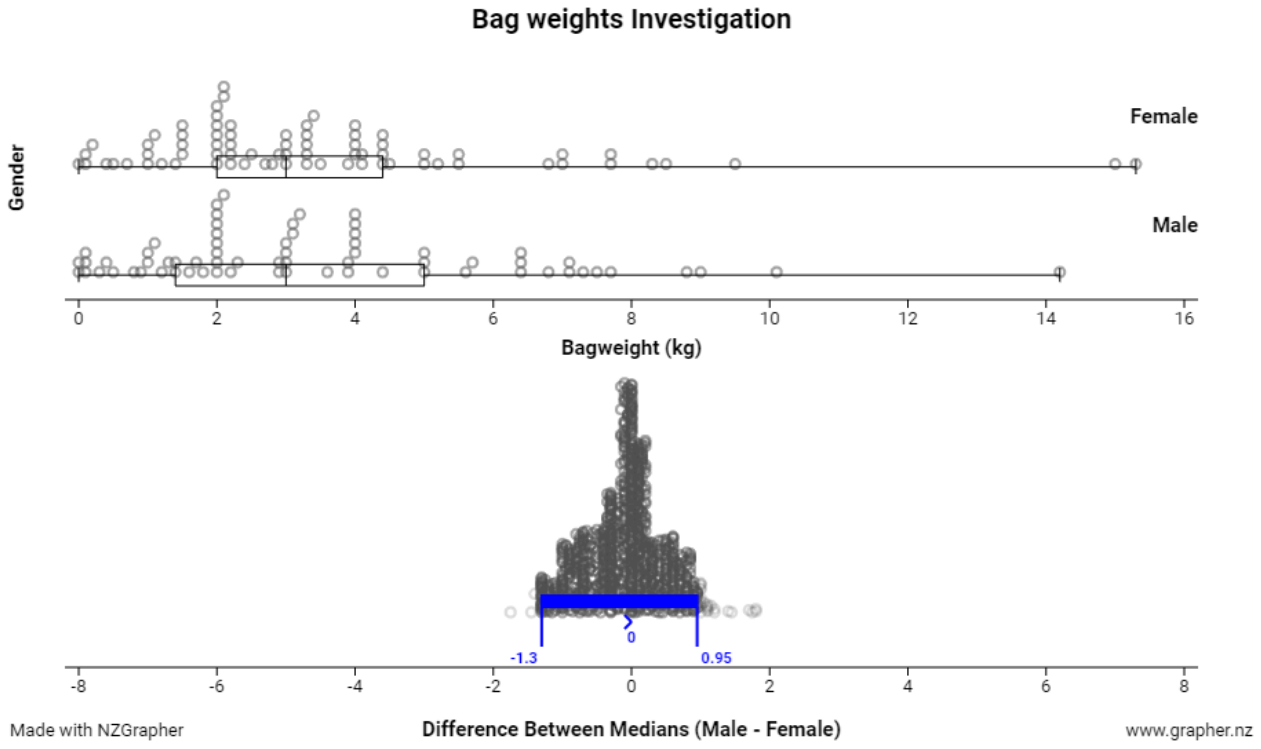
What is the difference between the amount of time that girls and boys spend reading, for all teenagers at school in NZ?



Answer the investigation question:

6) **Problem:**

What is the difference between the median weight of school bags for female and male teenagers in NZ?



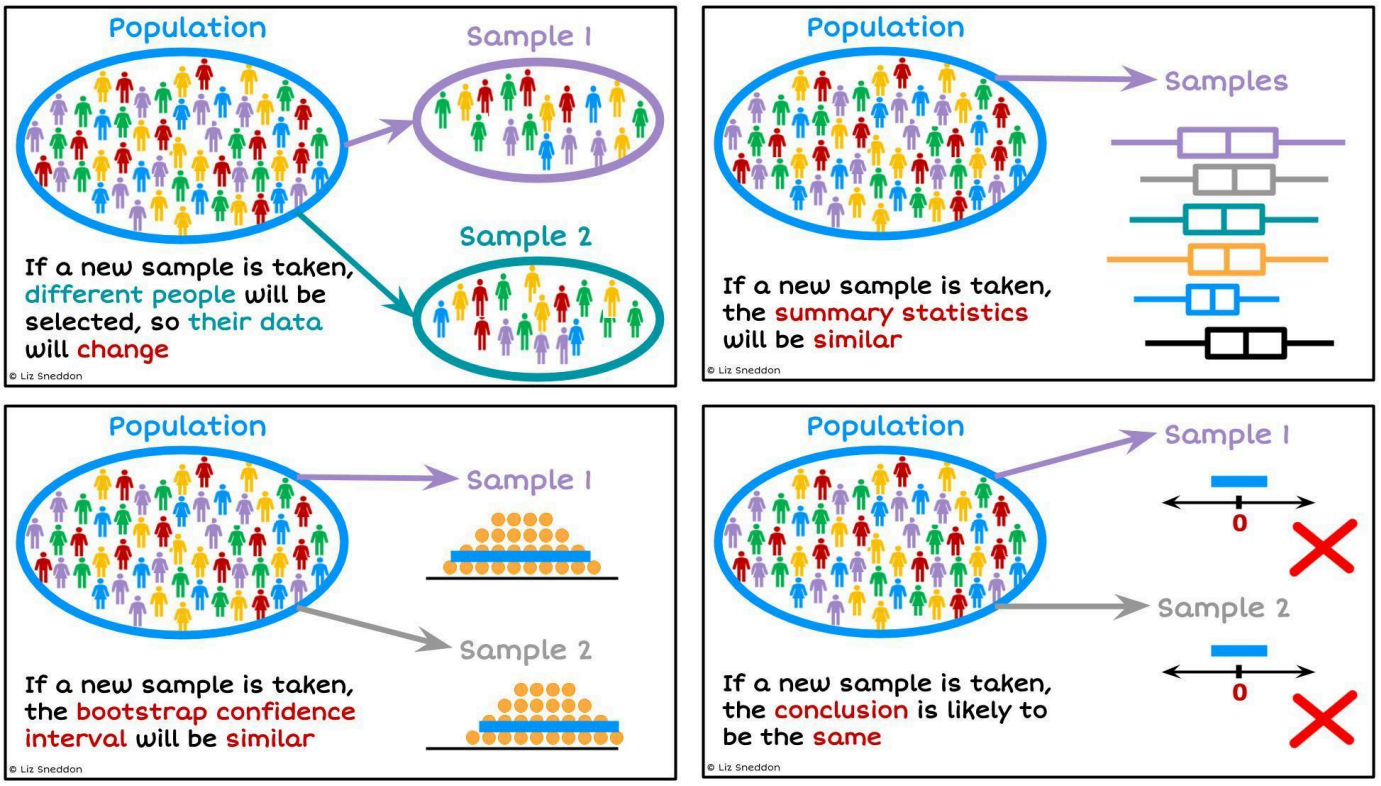
Answer the investigation question:

Sampling variability

This describes the variation (differences) that happens when you take another sample. You always have to write an explanation in context.

You need to consider:

- how would the **data** change?
- how would the **analysis** change (particularly the summary statistics)?
- how would your **bootstrap confidence interval** change?
- how would your **conclusion** change?



Example:

If I took another sample, I would get different weights for rugby players as I would be collecting data from different rugby players. I would expect though that the summary statistics (minimum, LQ, median, UQ and maximum) weights of forwards and backs to be similar to the values in my sample. Because the median weights for forwards

and backs would be similar, this would lead to a similar bootstrap confidence interval, and therefore the conclusion that the weights of forwards are larger than the weights of backs, is likely to stay the same.

Exercise 21:

1) Discuss sampling variability for the Kiwi dataset.

2) Discuss sampling variability for the Marathon dataset.

Confidence Intervals

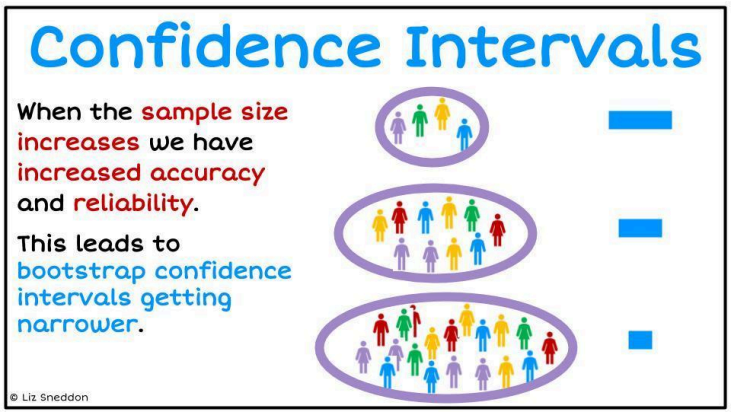
There are two things we want to consider:

- Sampling variability with different sized samples,
- Bootstrap confidence intervals and the differences between the population medians.

Confidence Intervals & Sample size

Go to: https://sites.google.com/view/lizsneddon/year-13/inference/conclusion

This has several animations for you to view.



Exercise 22:

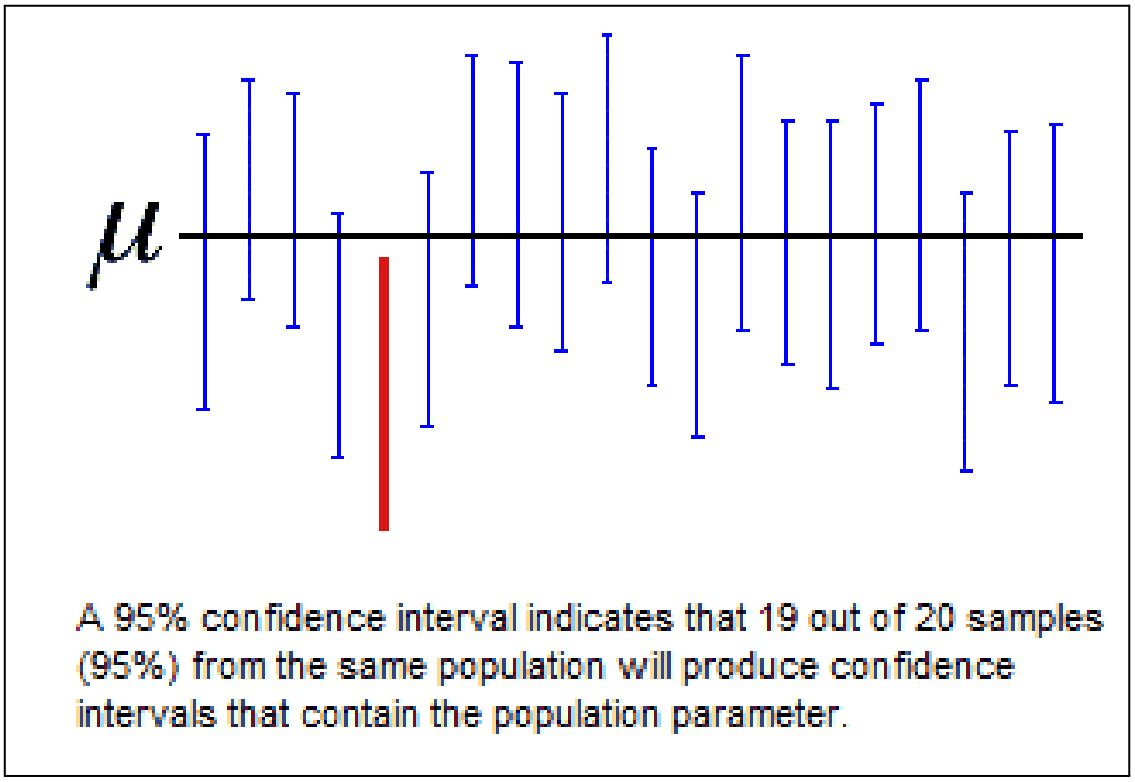
Complete the sentence below, considering how the data, the median, the confidence interval and the conclusion will be affected by a larger sample size.

If I took a bigger sample ...

Confidence intervals and population medians

The second idea we need to cover is whether or not in **our one sample**, how likely we are to have captured the true difference between the population medians in our confidence interval?

The image below shows twenty different confidence intervals, and out of these twenty intervals, there is only one that does not cover the population median.



When we take a sample, we don't know which of these confidence intervals we have, and whether or not we have the one that does not contain the difference between the population medians. However, because most of the confidence intervals **DO** contain the difference between the population medians, it is quite likely that our confidence interval **does** contain the difference between the population medians.

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:
What is the difference between the median weight of teenagers than 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 23:

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

1) **Problem:**
What is the difference between the median time that teenagers spend on phone calls each day compared to adults, for all people in NZ.

Other factors:
What other factors do you think would affect the time that teenagers and adults 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:**

What is the difference between the median lung capacity of teenagers who vape compared to those 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:**

What is the difference between the median house price for 2-bedroom homes compared to 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.