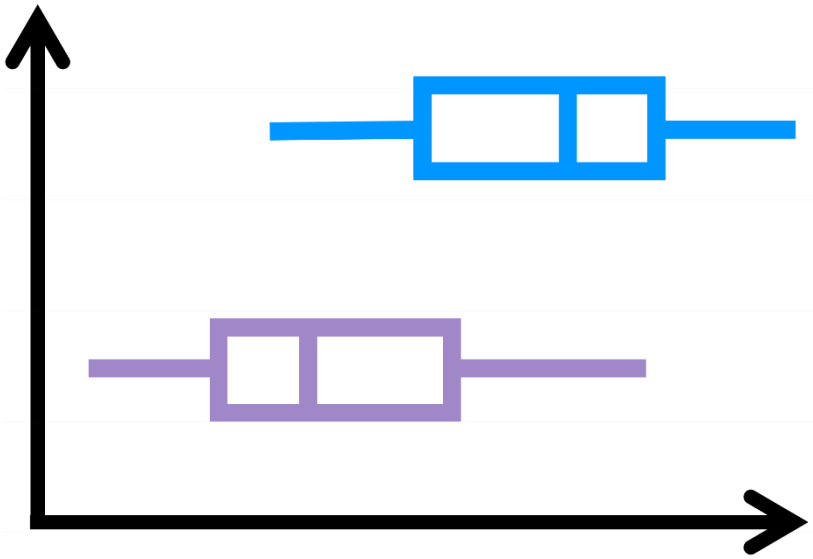
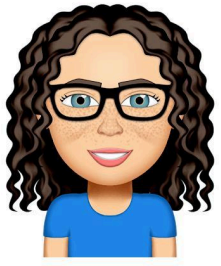
Year 9 Comparison Investigations Workbook



Name:



By Liz Sneddon

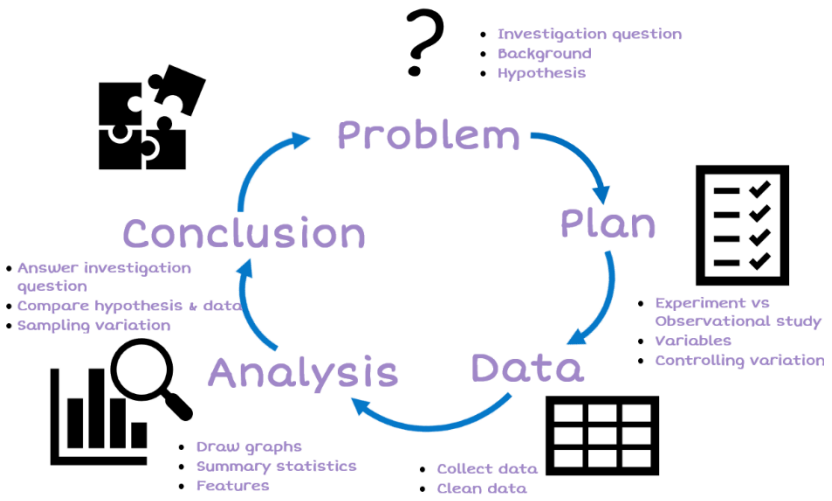
© Liz Sneddon 2020

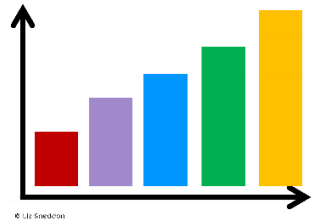
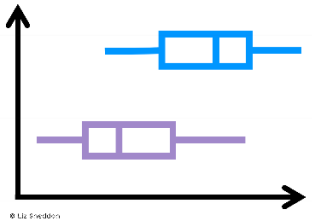
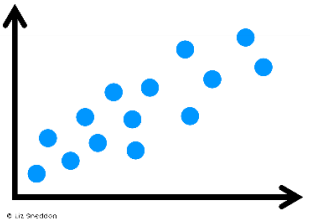
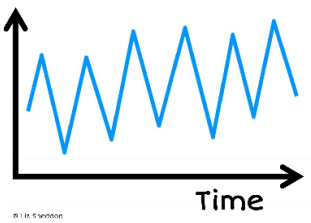
Statistical Investigations

When doing a Statistical investigation, we want to go through the PPDAC cycle (shown below).

There are several different types of Investigations.

Statistical Investigations



Summary	Comparison	Relationship	Time Series
			
Investigate **one variable** at a time	Investigate **two variables**	Investigate **two variables**	Investigate **two variables**
Types of variables: • Categorical, **OR** • Numerical.	Types of variables: • One Categorical, **AND** • One Numerical.	Types of variables: • Two Numerical.	Types of variables: • Time, **AND** • One Numerical.

© Liz Sneddon 2020

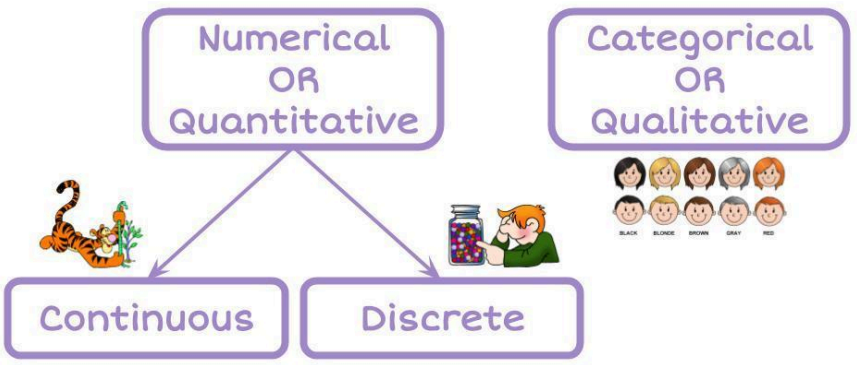
Comparison Investigations

We are now going to focus on Comparison Investigations.

This means that in a dataset we will investigate two variables at a time, one **categorical** and one **numerical**.

Reminder:

Categorical (groups) variables are characteristics, that cannot be described by numbers e.g. gender, ethnicity, apple variety.



Numerical (numbers) variables are characteristics described by numbers e.g. height, age, number of apples, weight. Numerical variables are either **discrete** or **continuous**.

Example:

Here is a sample of students from the Wall sit spreadsheet:

Students First Name	Age	Gender	Taking PE this year?	Wall sit time (seconds)	Height (cm)
Jessie	17	Female	No	114	161
Caleb	18	Male	Yes	640	185
Amisha	16	Female	No	352	155
Alena	18	Female	Yes	238	169
Luke	17	Male	Yes	421	182

Exercise:

Choose what data type the following variables are:

Students First Name	Age	Gender	Taking PE this year?	Wall sit time (seconds)	Height (cm)
• Categorical • Numerical	• Categorical • Numerical	• Categorical • Numerical	• Categorical • Numerical	• Categorical • Numerical	• Categorical • Numerical

Problem

Writing Investigation Questions

A comparison question needs:

- **One categorical variable**,
- **One numerical variable**,
- Uses the word
    - "**tend**" or "**typical**"
- **Population** (use the word **ALL** to describe it).



Example:

Here is a sample of students from the Wall sit spreadsheet:

Students First Name	Age	Gender	Taking PE this year?	Wall sit time (seconds)	Height (cm)
Jessie	17	Female	No	114	161
Caleb	18	Male	Yes	640	185
Amisha	16	Female	No	352	155
Alena	18	Female	Yes	238	169
Luke	17	Male	Yes	421	182
…	…	…	…	…	…

The sample is some Year 11 students at Saint Kentigern College, so the population is **ALL Year 11 students at Saint Kentigern College**.

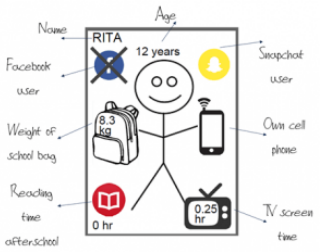
Variables:

- Categorical: gender
- Numerical: height

Comparison question:
I wonder whether the boys tend to be taller than girls for **ALL Year 11 students at Saint Kentigern College**?

-

Exercise:



1) What are some comparison investigation questions you could ask for the Stickland dataset?

   (The population is **ALL high school students in NZ**).

Variable	Description
Age	How old the student is (years)
Gender	Is the student male or female.
Facebook	Whether or not the student has Facebook (yes / no)
Snapchat	Whether or not the student has Snapchat (yes / no)
School bag weight	How heavy the students school bag is (kg)
Cellphone	Whether or not the student has a cellphone (yes / no)
Reading time	How much time the student spent reading a book per day (hours)
TV time	How much time the student spent watching TV per day (hours)

Comparison 1:

- Variables:
  - Categorical:
  - Numerical:
- Investigation question:
  - I wonder …

Comparison 2:

- Variables:
  - Categorical:
  - Numerical:
- Investigation question:
  - I wonder …

Comparison 3:

- Variables:
    - Categorical:
    - Numerical:
- Investigation question:
    - I wonder …

Comparison 4:

- Variables:
    - Categorical:
    - Numerical:
- Investigation question:
    - I wonder …

Comparison 5:

- Variables:
    - Categorical:
    - Numerical:
- Investigation question:
    - I wonder …

Comparison 6:

- Variables:
    - Categorical:
    - Numerical:
- Investigation question:
    - I wonder …

Hypothesis

We can use our general knowledge to make a hypothesis about what we might find **BEFORE** we look at the data.

Example:

Problem:

I wonder whether the boys tend to be taller than girls for **ALL Year 11 students at Saint Kentigern College**?

Hypothesis:

I think that the girls might be taller than boys at Year 11, because the girls have mostly finished their growth spurt, while boys are usually only starting their growth spurt and don't finish until they are about 18 years old.

Exercise:

Make a prediction for the following investigation questions from the Wall sit dataset.

1) **Problem:**

I wonder whether girls tend to be able to hold a wall sit for longer than the boys for **ALL Year 11 students at Saint Kentigern College**?

Hypothesis:

I predict that …

2) **Problem:**

I wonder whether students who take PE tend to be able to hold a wall sit for longer than students who **don't** take PE, for **ALL Year 11 students at Saint Kentigern College**?

Hypothesis:

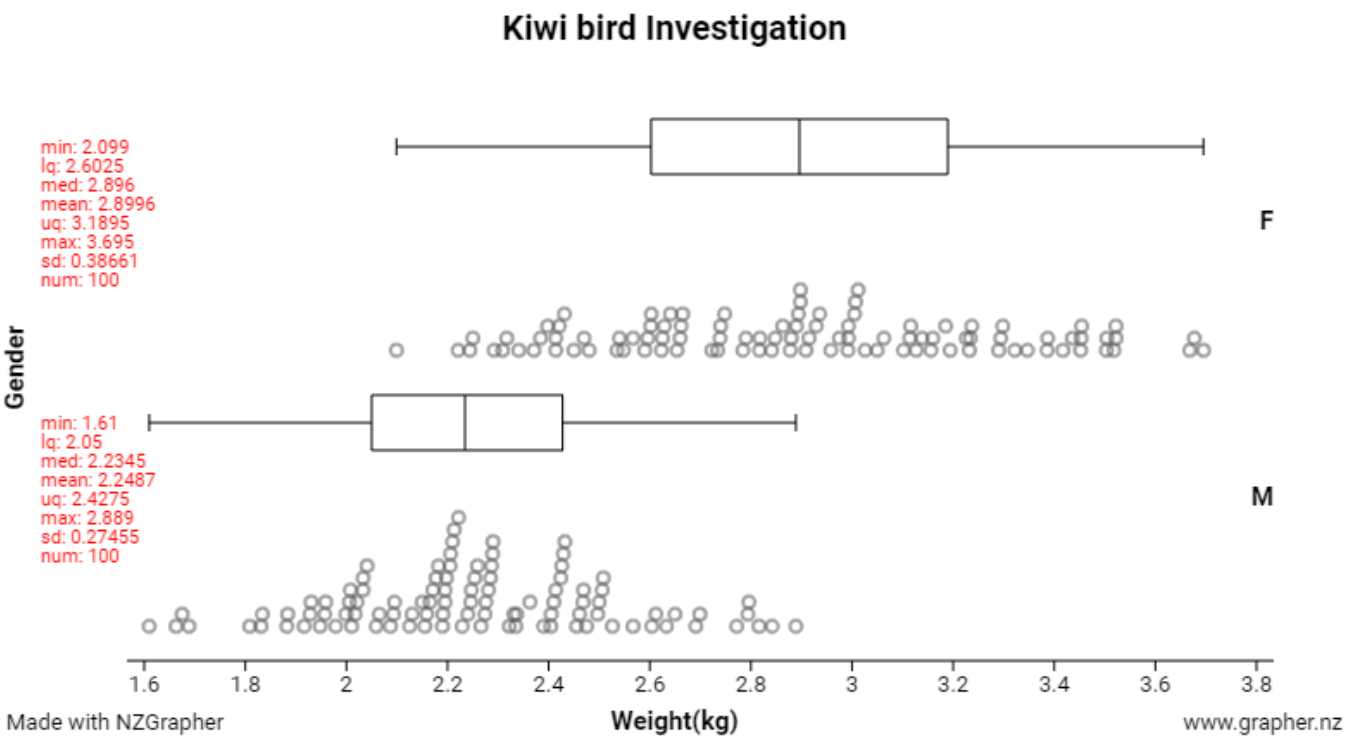
I predict that …

In the first workbook we discovered how to write a plan to collect data, how to answer a survey or questionnaire and what the difference between a random and biased sample is.

We are going to skip over the plan section and go to the Data section to make the graphs and summary statistics using NZGrapher. If you can't remember how to do this, watch the video : **http://tiny.cc/ComparisonDotPlot**

Here are a few reminders to display data and draw graphs.

The most appropriate graphs to do is a comparison dot plot and box plot.

Dot plot & Box plot


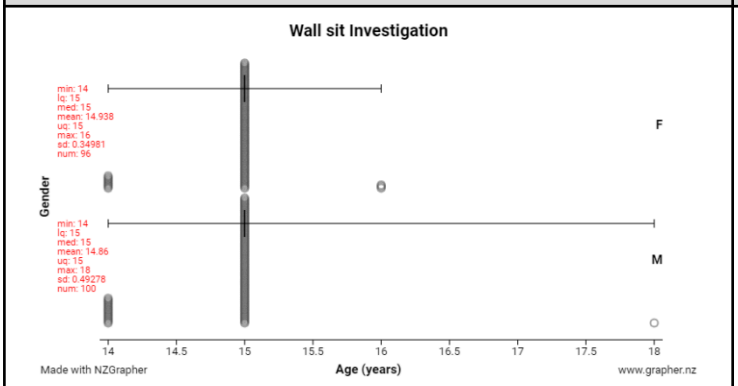
Example:

Here is a sample of students from the Wall sit spreadsheet:

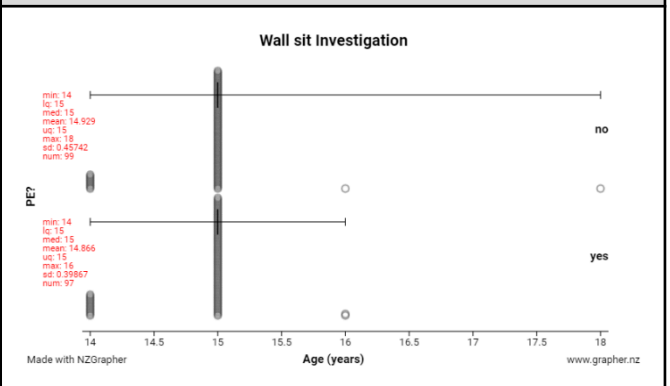
Students First Name	Age	Gender	Taking PE this year?	Wall sit time (seconds)	Height (cm)
Jessie	17	Female	No	114	161
Caleb	18	Male	Yes	640	185
...	...	...	...	...	...

Here are the comparison graphs I can make:

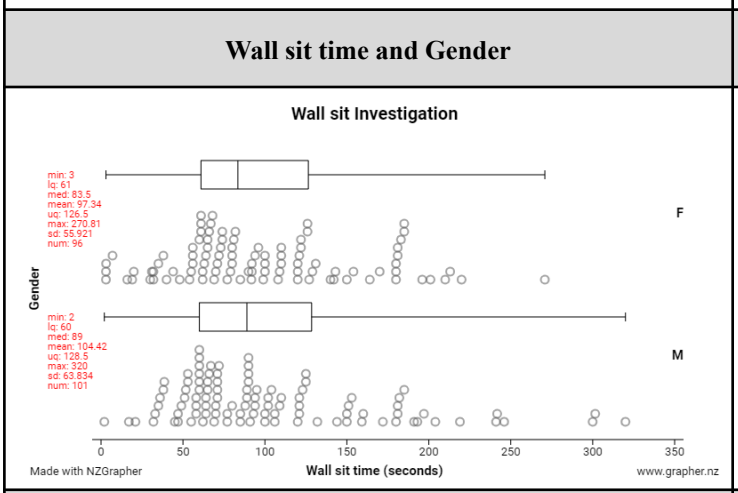
Age and Gender



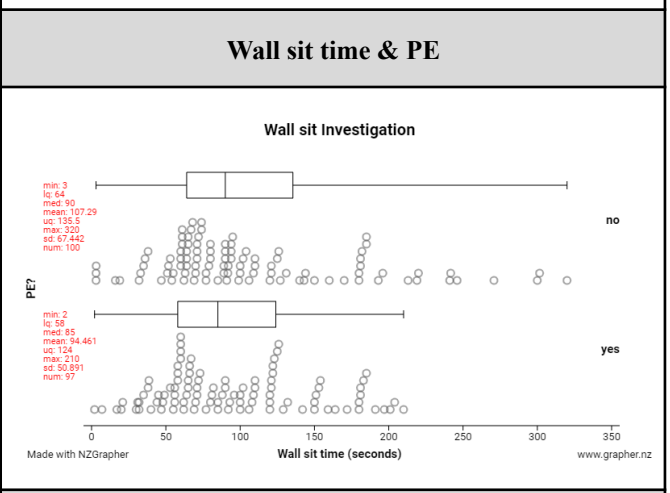
Age and PE



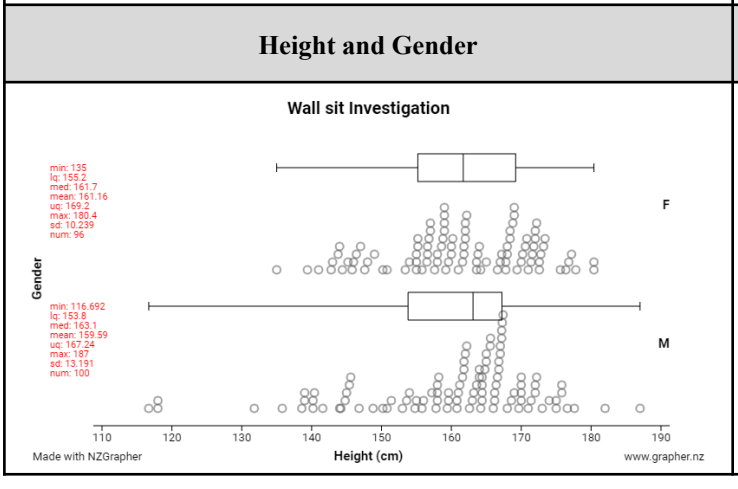
Wall sit time and Gender



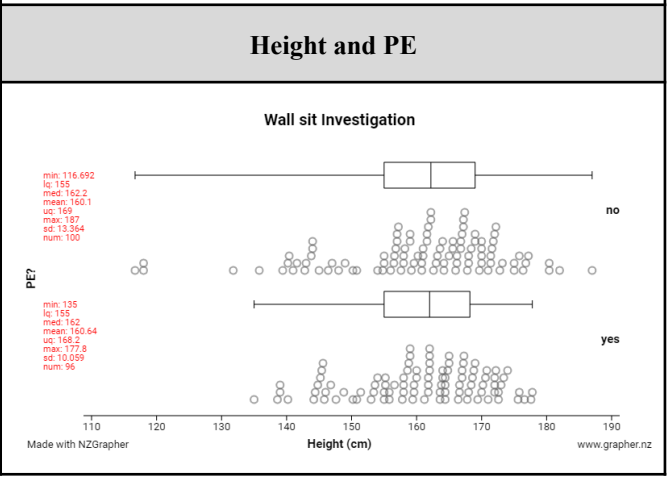
Wall sit time & PE



Height and Gender

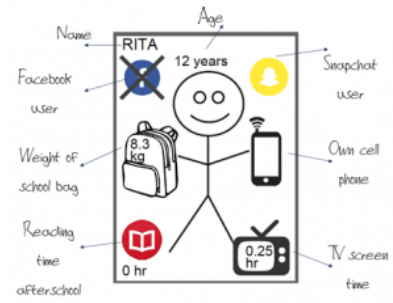


Height and PE



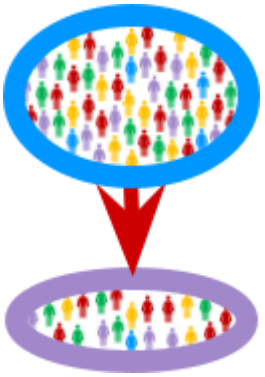
Exercise:

1) [Click here] to NZGrapher and make Comparison graphs for all variables (don't do the Name variable) in the Stickland dataset.

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

Analysis

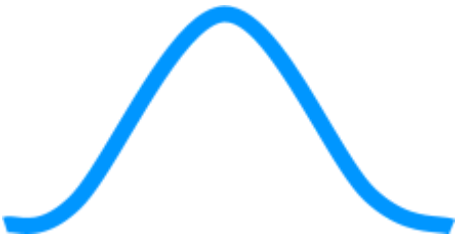
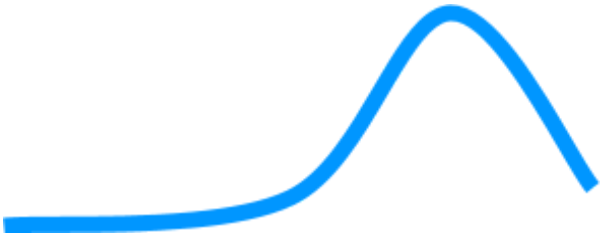
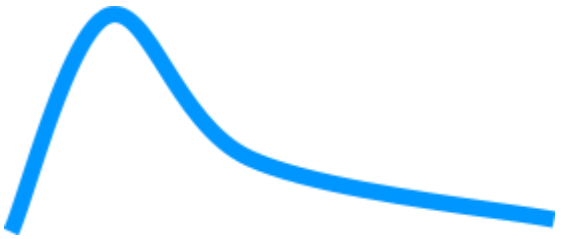
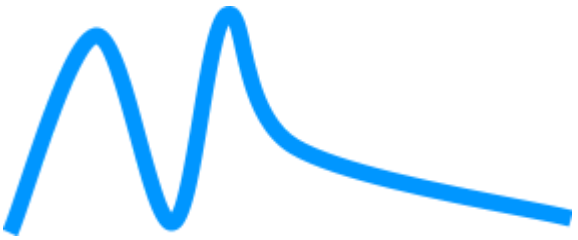
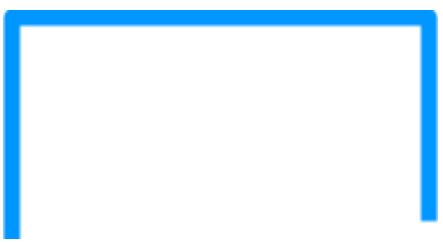
In the Analyse section we will explore the following features in the **sample**.



1. Shape
2. Center
3. Spread

We will now go through each feature, before putting it all together.

Shape

Normal distribution (hill/mound shapes, symmetric, bell shaped curve)	
Left skewed (Long tail on the left-hand side)	
Right Skewed (Long tail on the right-hand side)	
Bimodal (there are two peaks)	
Uniform (the sides are straight, and it looks like a box)	

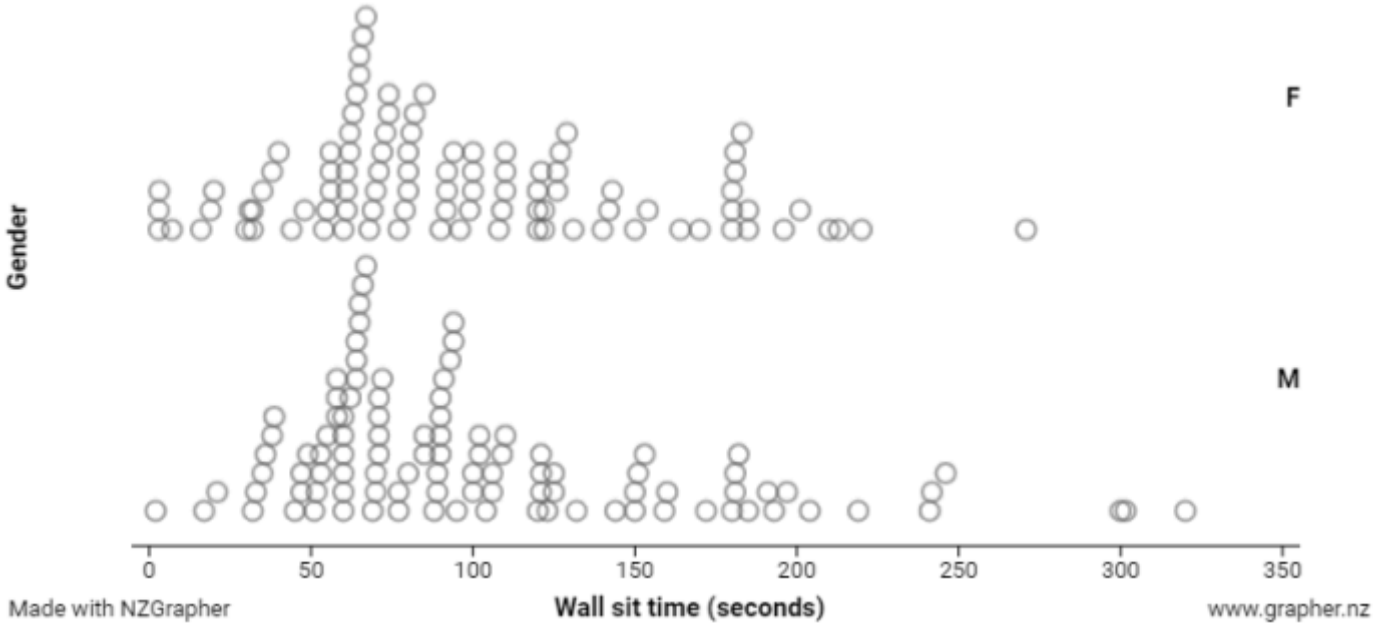
Writing about the shape

Now that we can identify the shape, we want to write this in a sentence in the context of the data.

We only discuss the shape of **Numerical** variables (not **categorical**).

Example:

Here are graphs of the Wall sit times and Heights, for our sample of students at Saint Kentigern College.

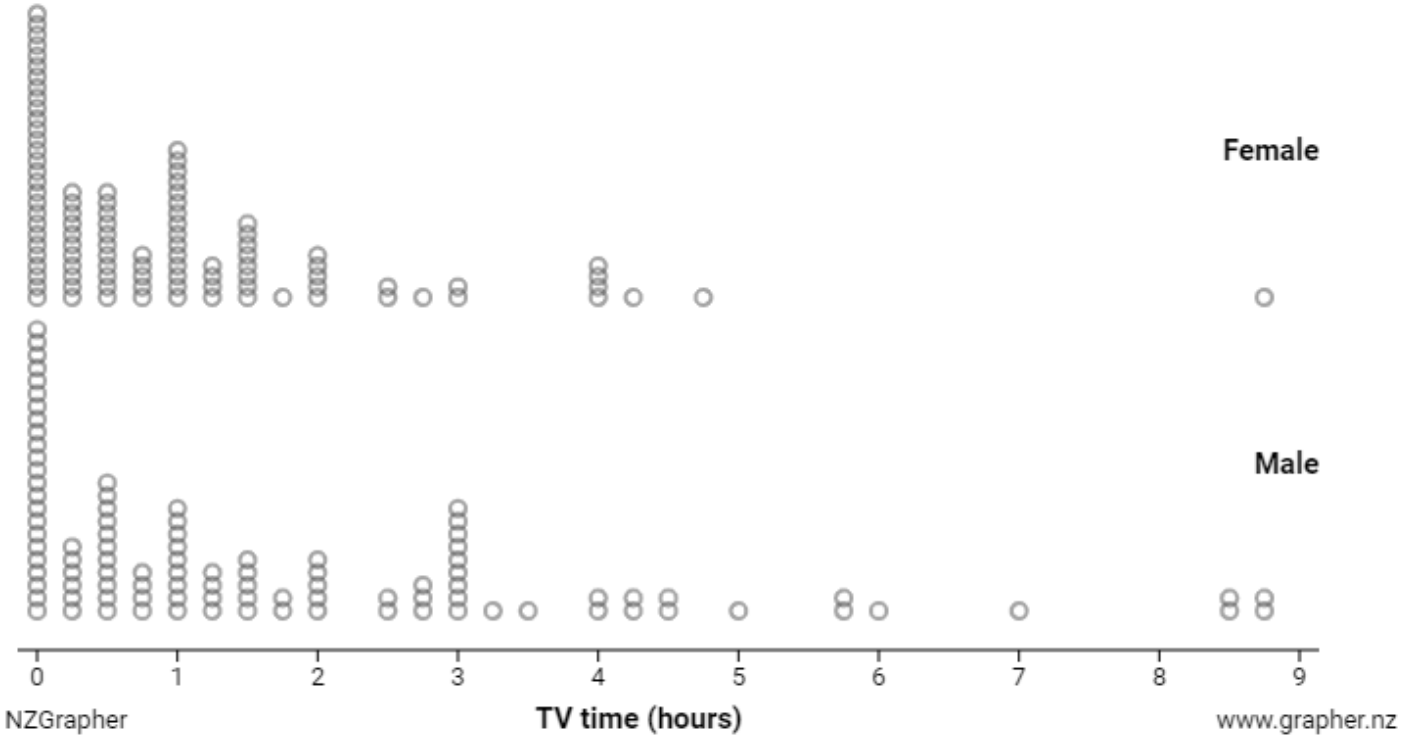


In my sample, I notice that the shape of the distribution of the time that **female** students can do a wall sit for is right skewed as there is a longer tail on the right-hand side, and the shape of the distribution of the time that **male** students can do a wall sit for is right skewed as there is a longer tail on the right-hand side

Exercise:

Write a sentence describing the shape for the sample of data from the Stickland dataset (which is a sample of high school students across New Zealand).

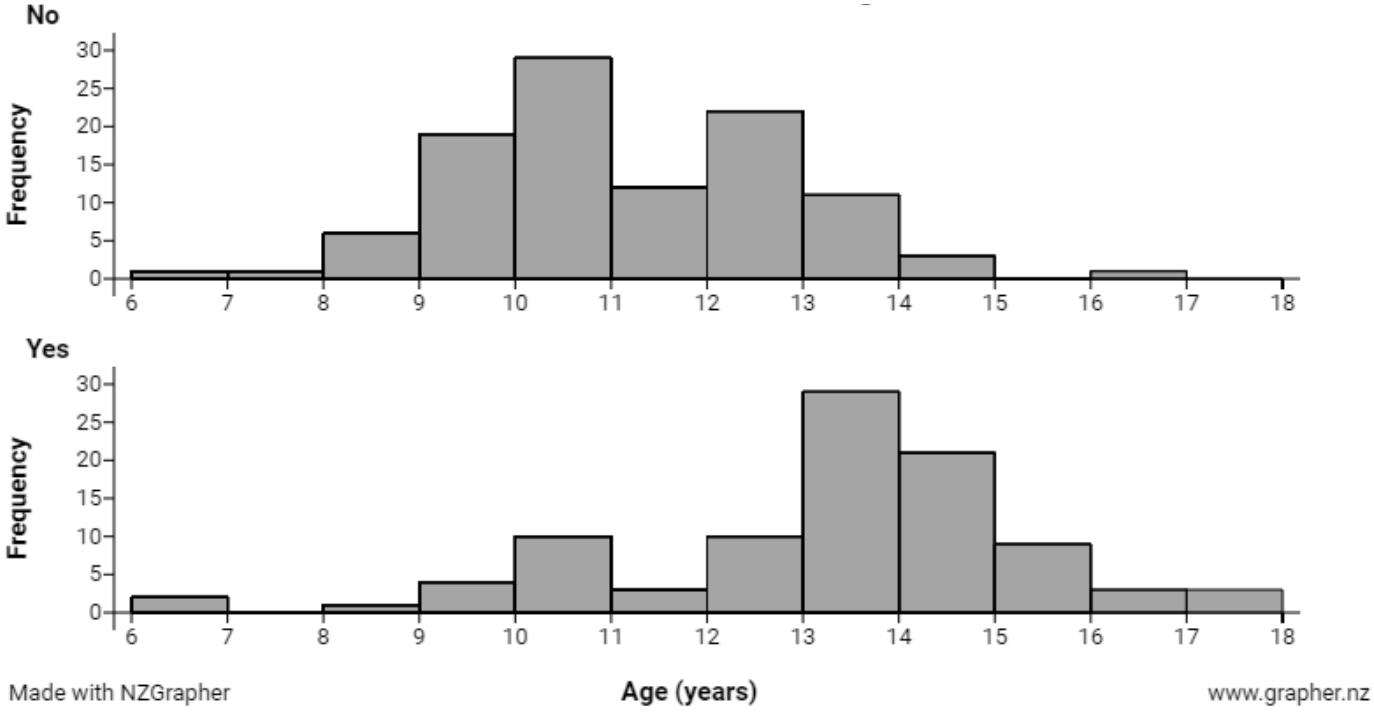
1) Comparing the amount of time boys and girls spend watching TV each day.



For my sample, I notice that the shape of the distribution of time that girls watch TV for each day is _____

For my sample, I notice that the shape of the distribution of time that boys watch TV for each day is _____

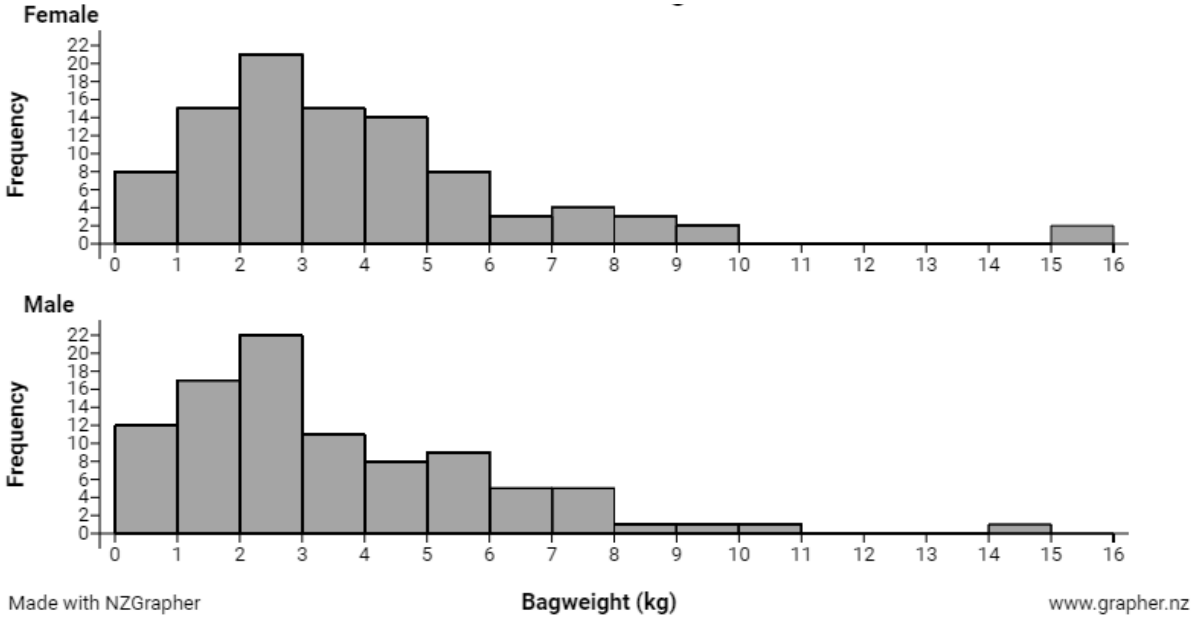
2) Comparing students age and whether or not they have Facebook.



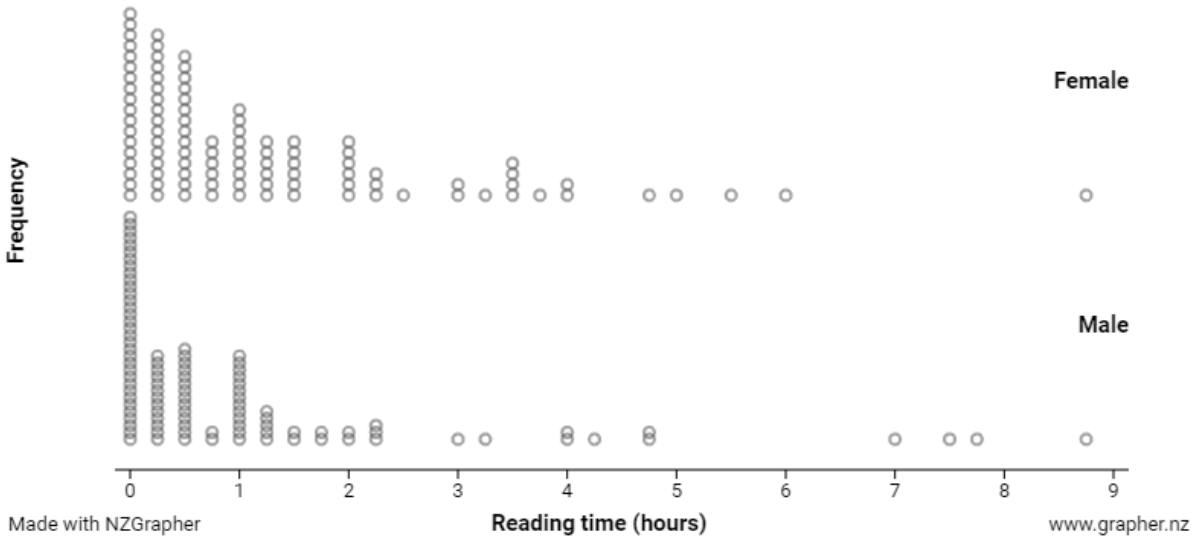
For my sample, I notice that the shape of the distribution of students who have Facebook is _____

For my sample, I notice that the shape of the distribution of students who **don't** have Facebook is _____

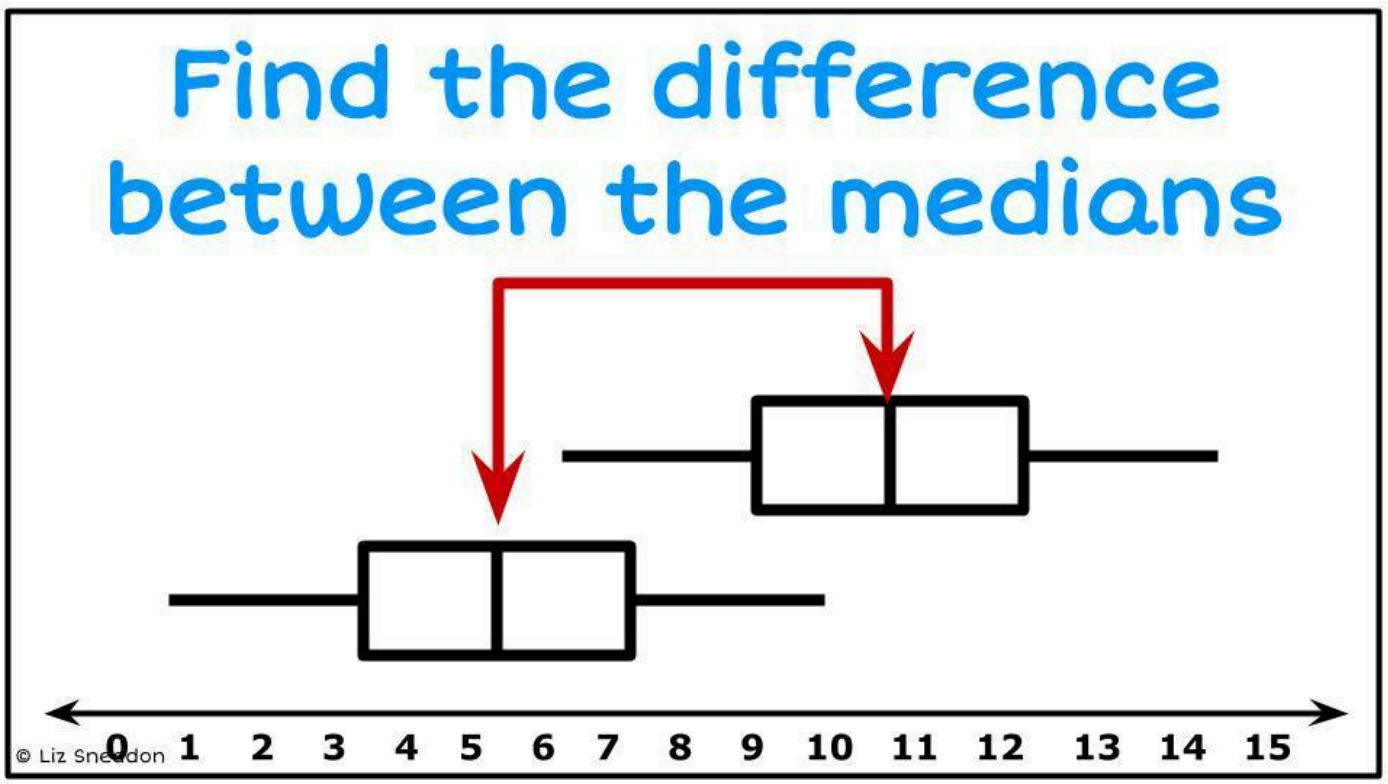
3) Comparing how heavy school bags are for girls and boys.



4) Comparing how much time girls and boys spend each day reading.



Centre

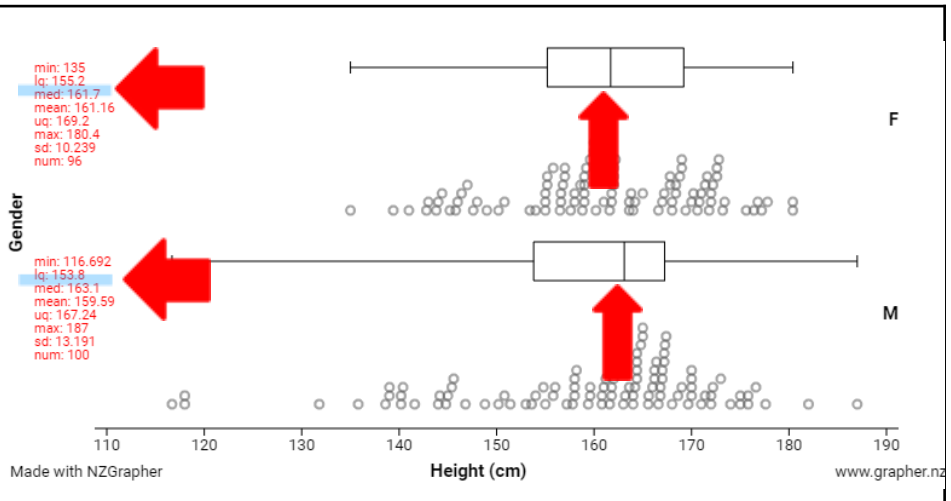


Writing about the centre

Remember that we can add summary statistics in NZGrapher, and this will calculate the Mean and Median for us. Because we have **numerical** data, we want to write one sentence comparing the **medians**.

Example:

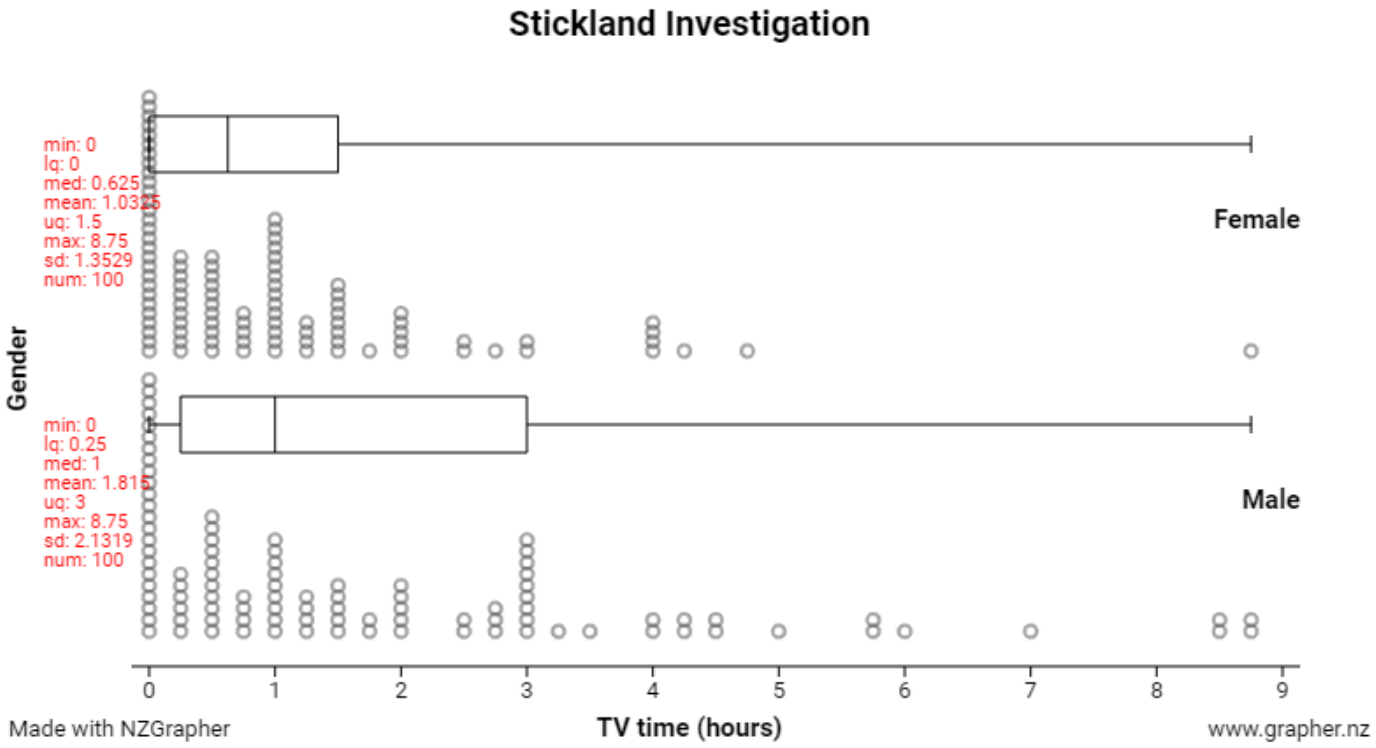
Here are graphs from the Wall sit dataset, for our sample of students at Saint Kentigern College.

In my sample, I notice that the median height of year 11 girls is 161.7 cm, and the median height of the year 11 boys is 163.1 cm. We can see that the median height of the boys is a little bit bigger than the girls, by 1.4 cm.	

Exercise:

Write a sentence comparing the centre for the sample of data from the Stickland dataset (which is a sample of high school students across New Zealand).

1) Comparing the amount of time boys and girls spend watching TV each day.

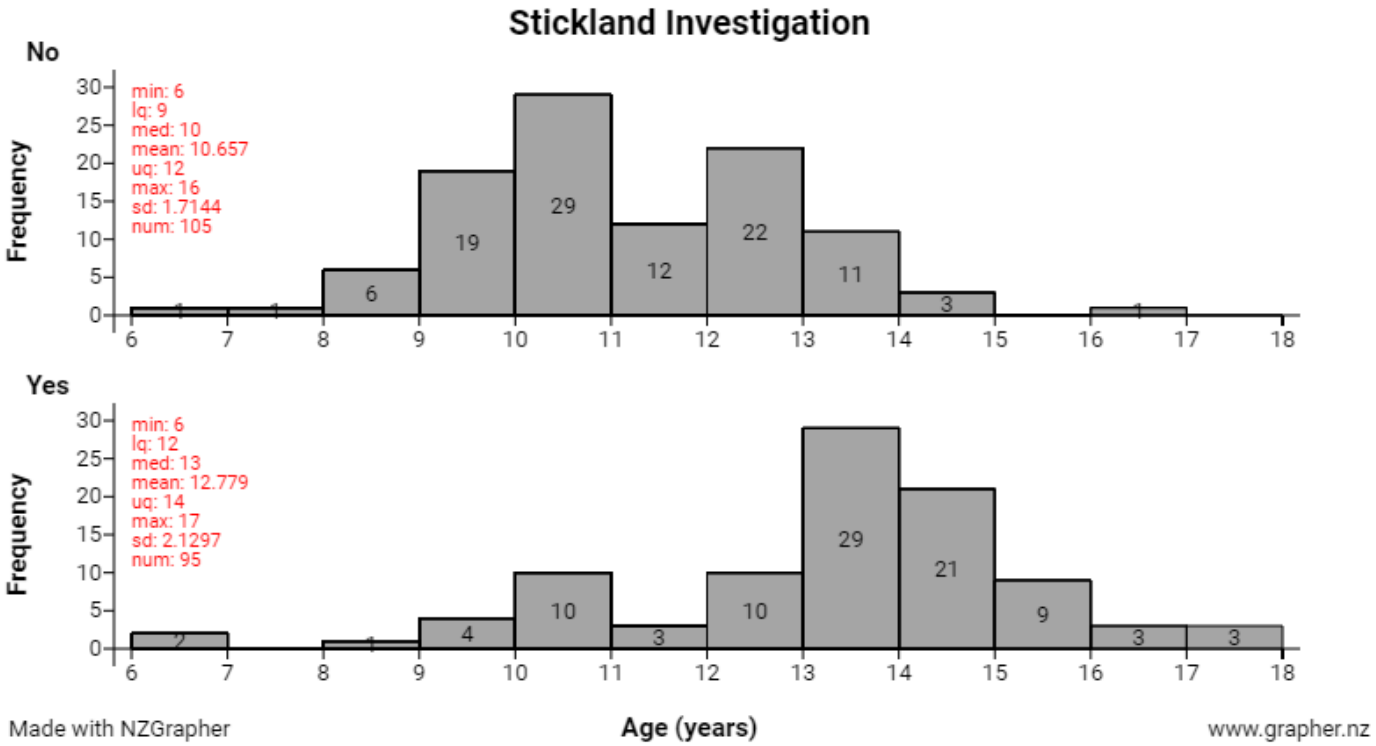


For my sample, I notice that the median amount of time that girls watch TV for each day is _____

For my sample, I notice that the median amount of time that boys watch TV for each day is _____

The median TV time for _____ is more than the median TV time for _____ by _____

2) Comparing students age and whether or not they have Facebook.

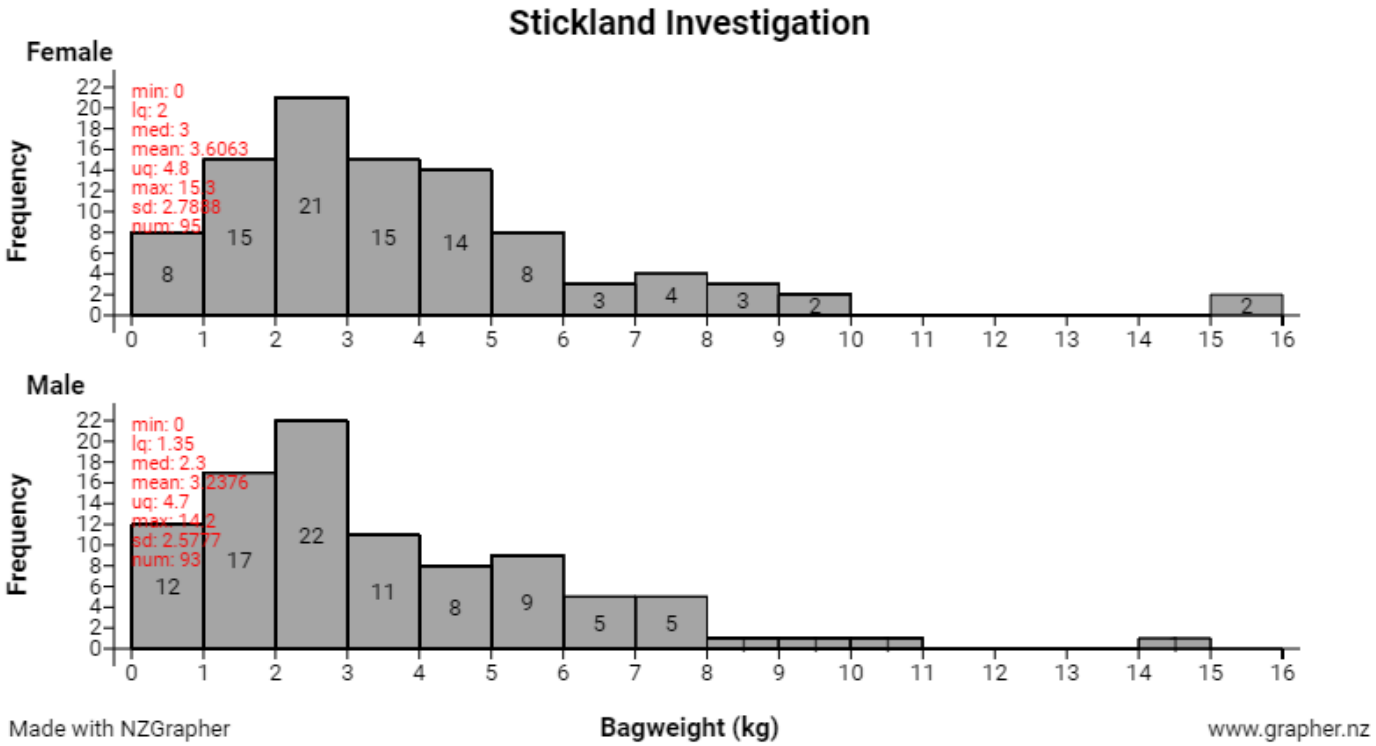


For my sample, I notice that the median age of students who **do NOT** have Facebook is _____

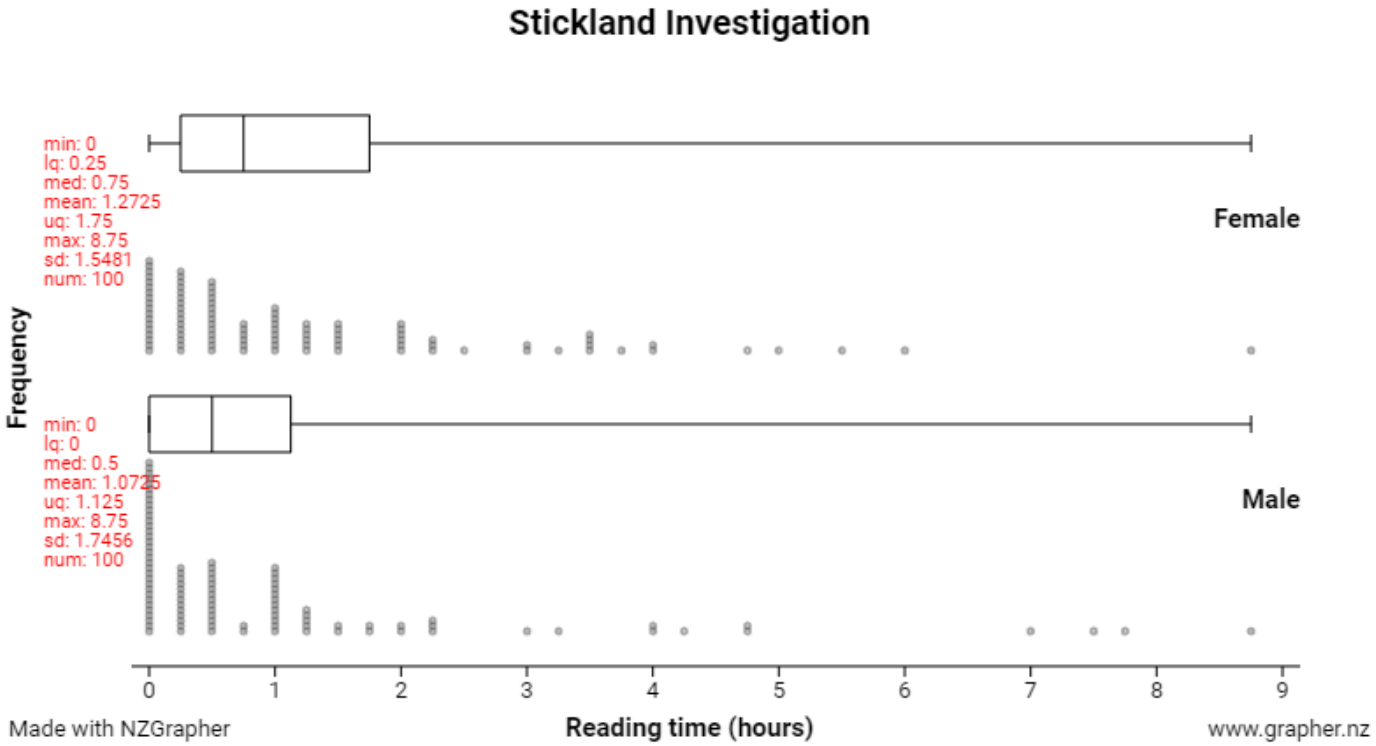
For my sample, I notice that the median age of students who have Facebook is _____

The median age for _____ is more than the median age for _____ by _____

3) Comparing how heavy school bags are for girls and boys.



4) Comparing how much time girls and boys spend each day reading.

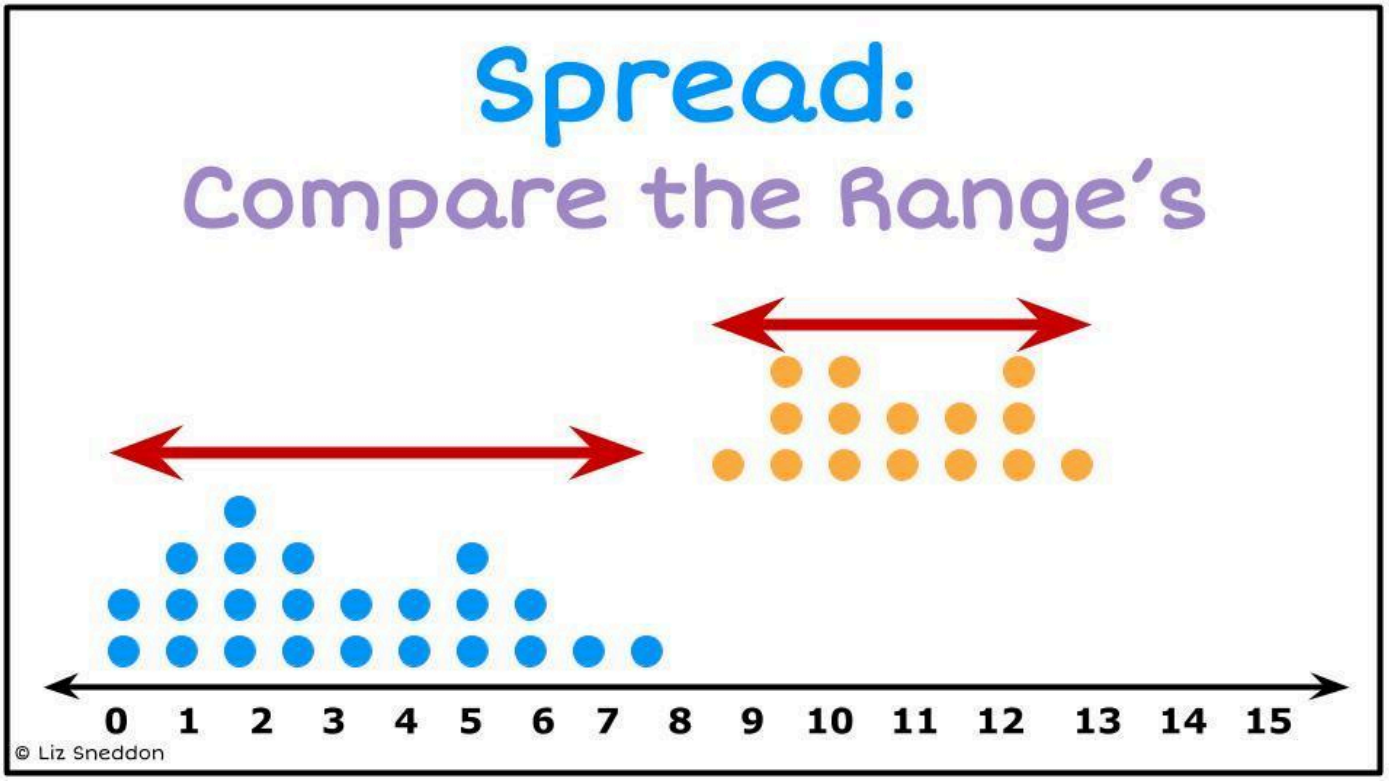


Spread

Remember that we can add summary statistics in NZGrapher, and this will calculate the Minimum, Maximum, UQ and LQ for us.

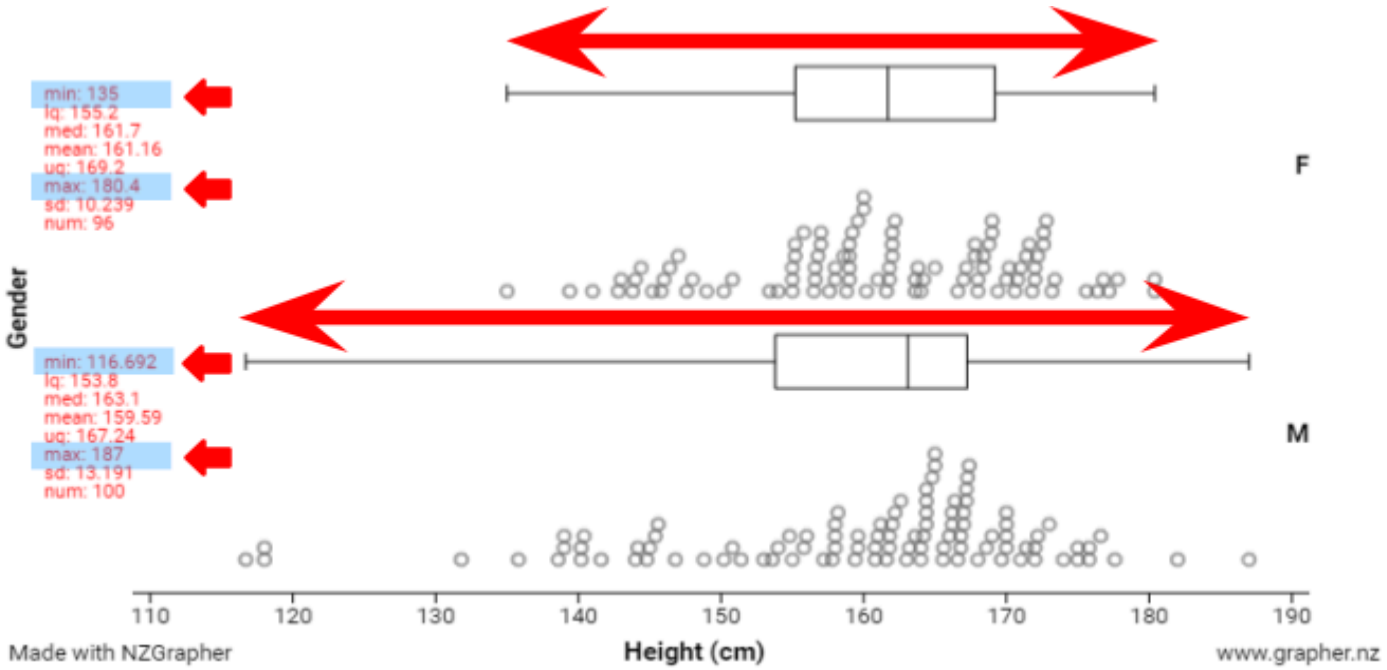
We just need to calculate the Range for each group using the summary statistics and write a sentence in context **comparing** the spread.





Example:

Here are graphs from the Wall sit dataset, for our sample of students at Saint Kentigern College.



From a visual look at the graph, we can see that the spread of the boy's heights is much more than the spread of the girl's heights, for our sample of year 11 students.

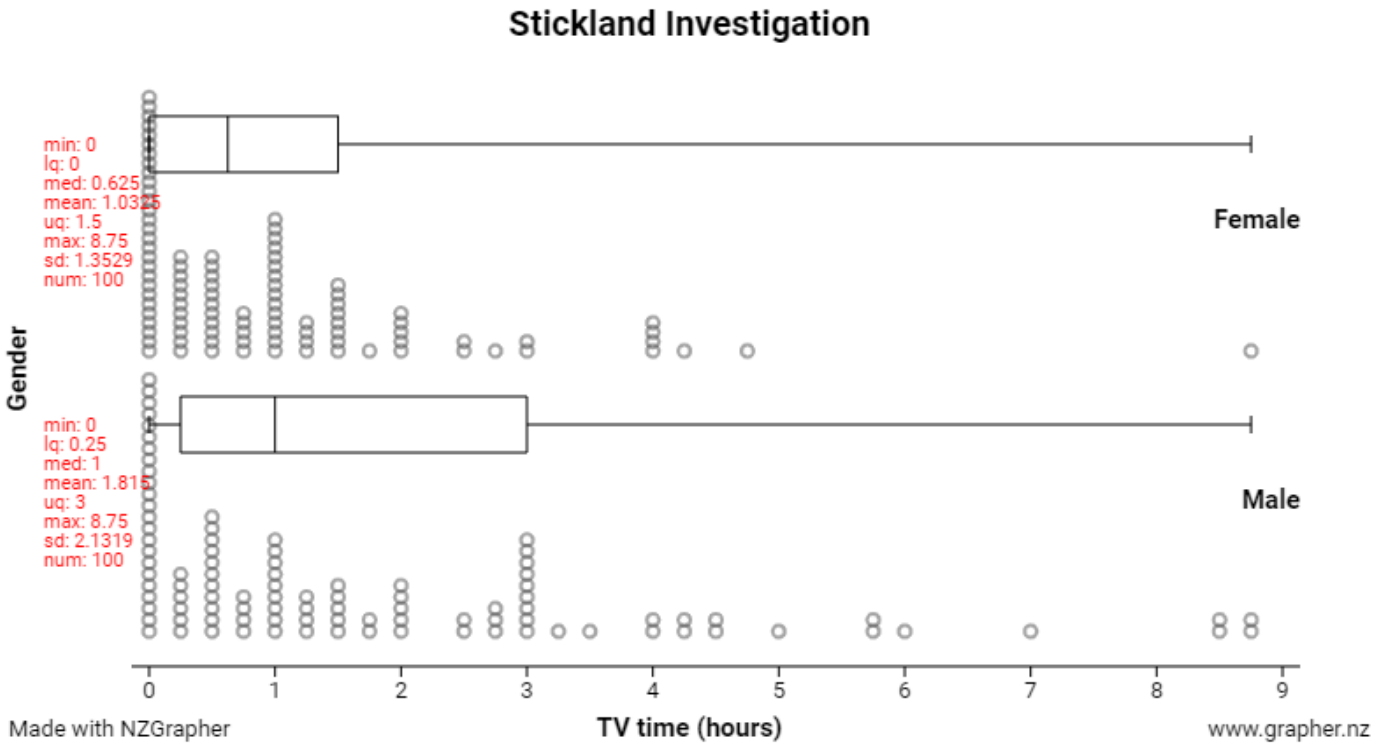
Girls	**Boys**
Range = Maximum – Minimum = 180.4 - 135 = 45.4 cm	Range = Maximum – Minimum = 187 – 116.692 = 70.308 cm

In my sample, I notice that the range of heights for Year 11 girls is 45.4 cm and the boys' range of heights is 70.308 cm. I can see that the heights of boys are much more spread out than the girls.

Exercise:

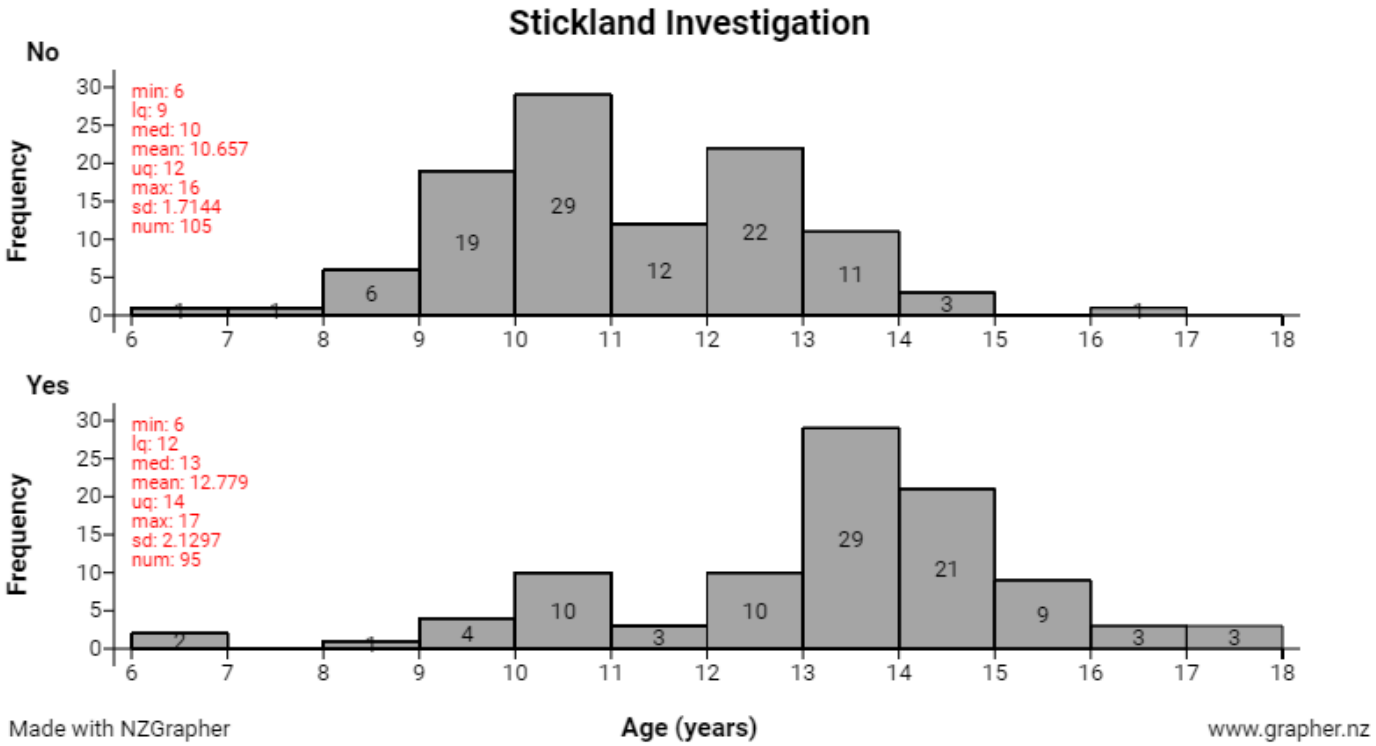
Write a sentence comparing the spread for the sample of data from the Stickland dataset (which is a sample of high school students across New Zealand).

1) Comparing the amount of time boys and girls spend watching TV each day.



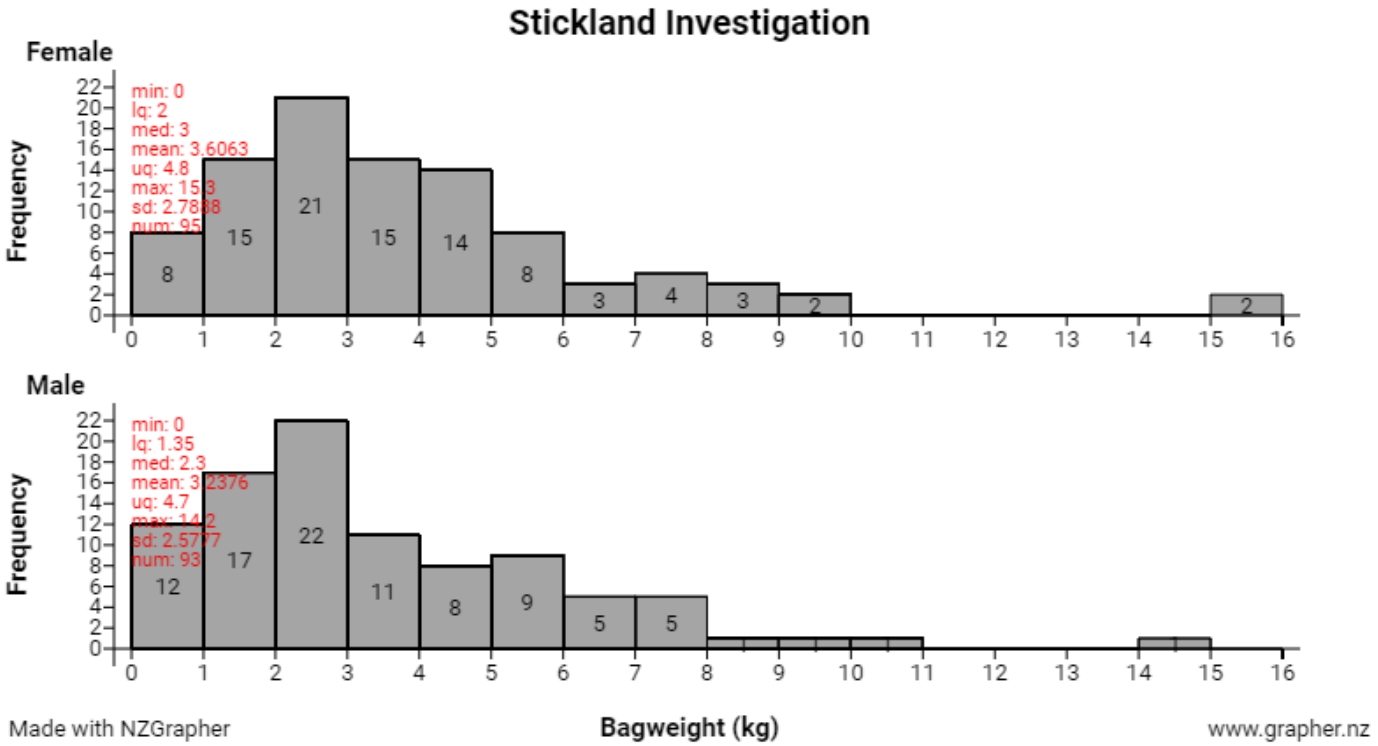
Girls	Boys
Range = Max – Min = =	Range = Max – Min = =
For my sample, I notice …	

2) Comparing students age and whether or not they have Facebook.



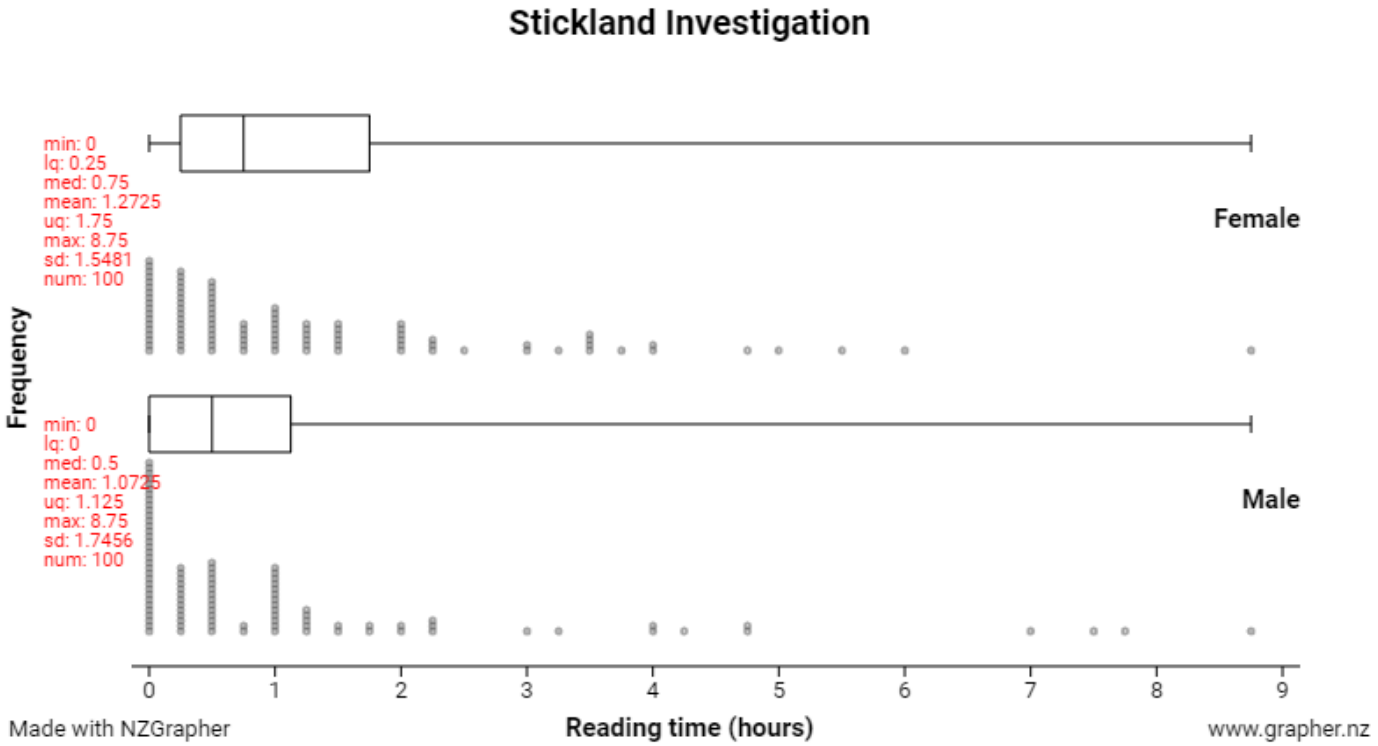
No Facebook	Yes Facebook
Range = Max – Min = =	Range = Max – Min = =
For my sample, I notice …	

3) Comparing how heavy school bags are for girls and boys.



Girls	Boys
Range = Max – Min = =	Range = Max – Min = =
For my sample, I notice …	

4) Comparing how much time girls and boys spend each day reading.

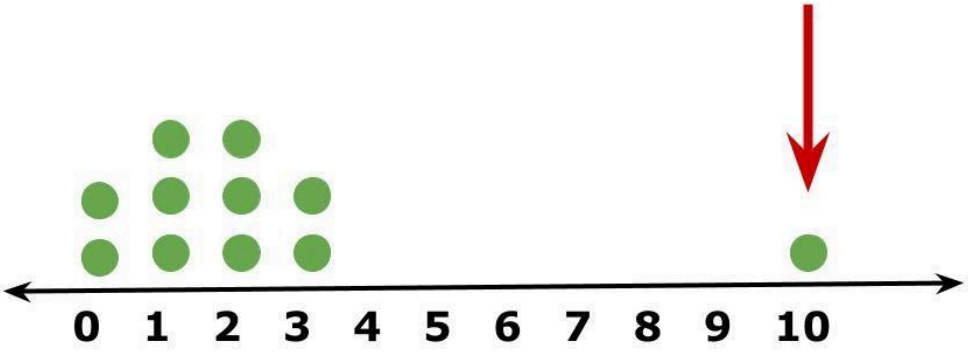


Girls	Boys
Range = Max – Min = =	Range = Max – Min = =
For my sample, I notice …	

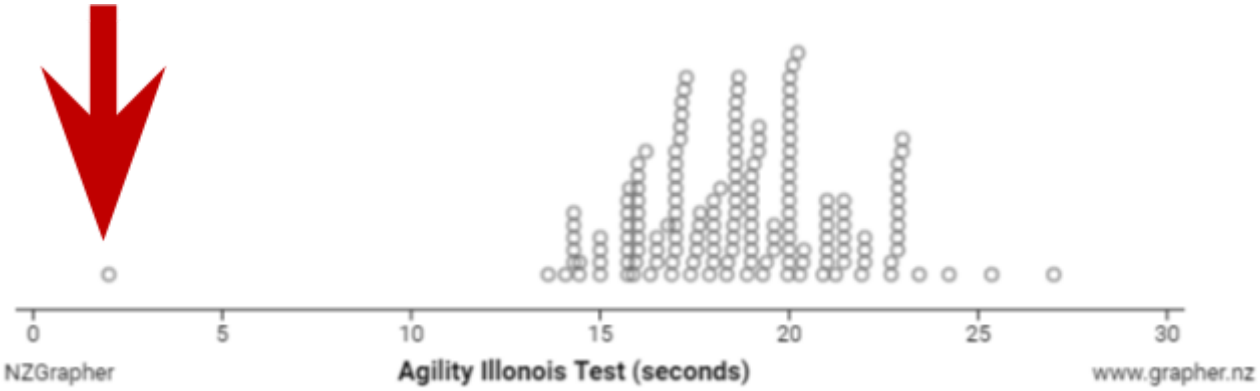
Outliers

An outlier is a data point that is a **LONG** way away from the rest of the data.

Identify if there are any outliers, then state their numerical value.



Example:

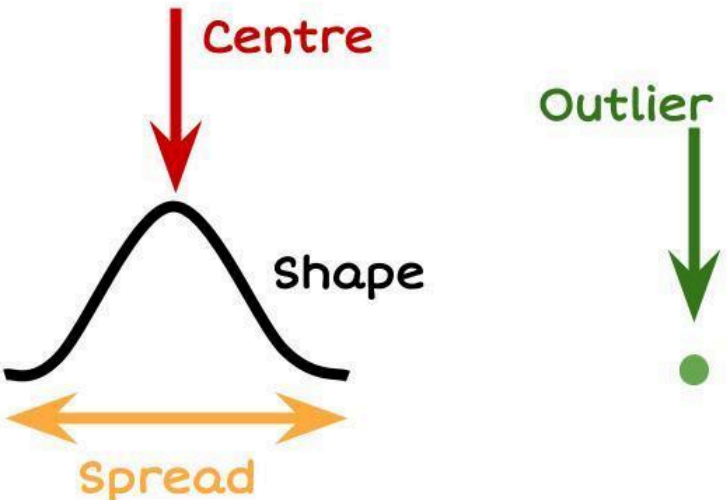


In this sample, there is one outlier, a student who has an agility test time of about 2 seconds.

Putting it altogether

Now we need to put all the features together:

- Shape,
- Centre,
- Spread,
- Outlier(s).

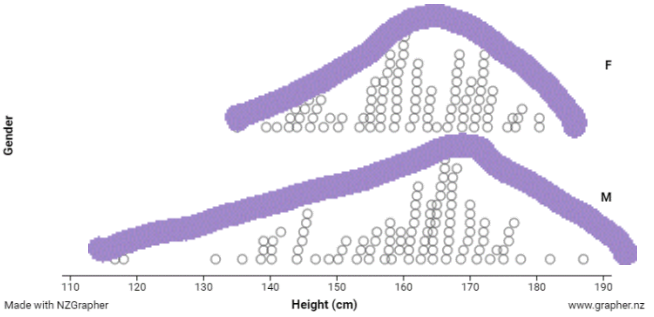


Example:

Write an analysis of the features (shape, centre, spread & outliers) for the **Gender** and **Heights** of students in the Wall sit dataset (Year 11 students at Saint Kentigern College).

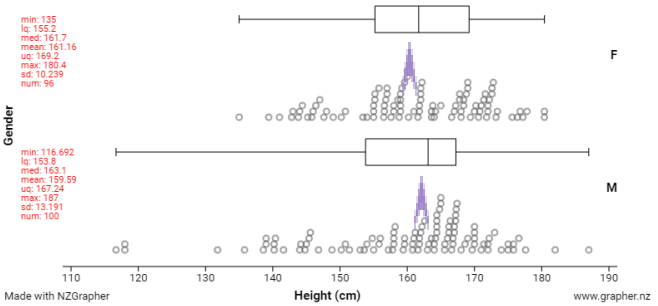
Shape:

In my sample, I notice that the shape of the distribution of the height of **female** students is normal as it is roughly symmetric with one peak, and the shape of the distribution of the height of **male** students is left skewed as there is a longer tail on the left-hand side



Centre:

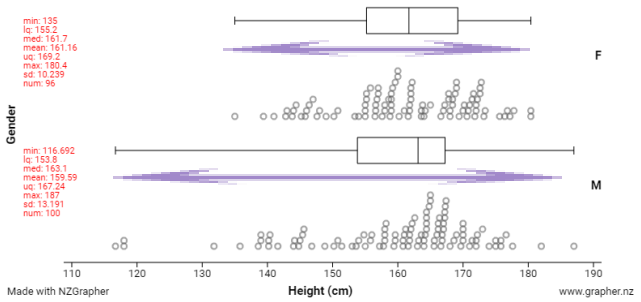
In my sample, I notice that the median height of year 11 girls is 161.7 cm, and the median height of the year 11 boys is 163.1 cm. We can see that the median height of the boys is a little bit bigger than the girls, by 1.4 cm.



Spread:

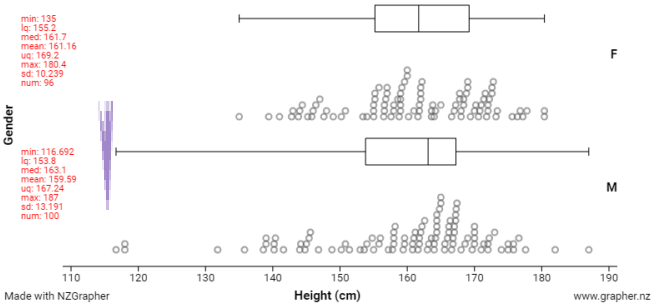
Girls	Boys
Range = Max – Min = 180.4 - 135 = 45.4 cm	Range = Max – Min = 187 – 116.692 = 70.308 cm

In my sample, I notice that the range of heights for Year 11 girls is 45.4 cm and the boys' range of heights is 70.308 cm. I can see that the heights of boys are much more spread out than the girls.



Outliers:

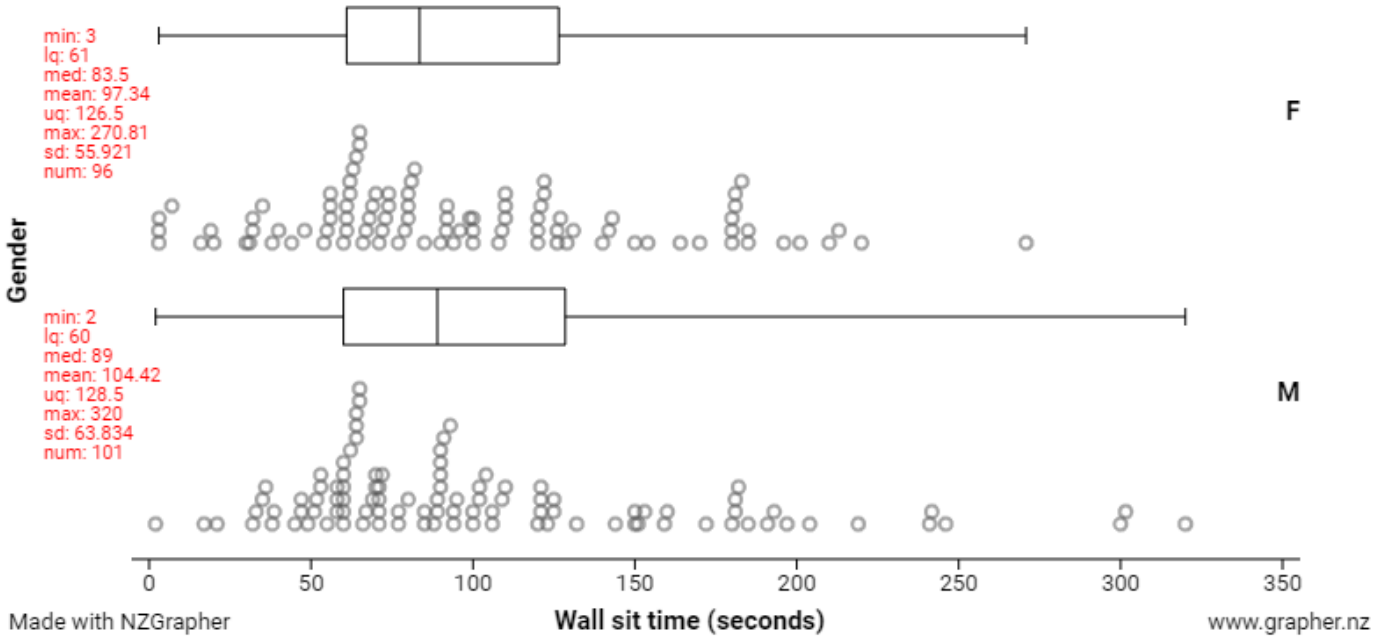
In my sample, I notice that there are 3 possible outliers, male students who have heights of around 116 and 118cm.



Exercise:

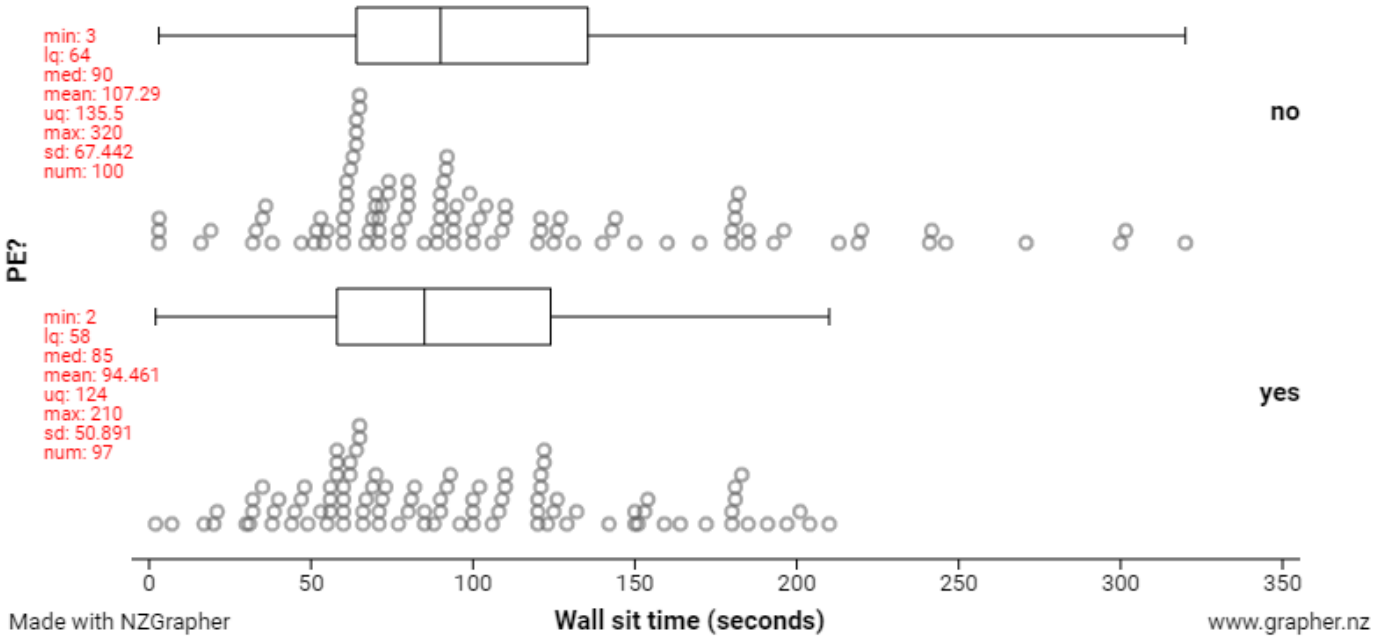
Write an analysis of the features (shape, centre, spread & outliers) for the sample of data from the Wall sit dataset (Year 11 students at Saint Kentigern College).

1) Amount of time girls and boys are able to do a wall sit for.



Girls	Boys
Range = Max – Min = =	Range = Max – Min = =
Shape: **Centre:** **Spread:** **Outliers:**	

2) Amount of time PE students and non-PE students are able to do a wall sit for.



No PE	Yes PE
Range = Max – Min = =	Range = Max – Min = =
Shape: **Centre:** **Spread:** **Outliers:**	

Conclusion

There are two things you need to do in your conclusion:

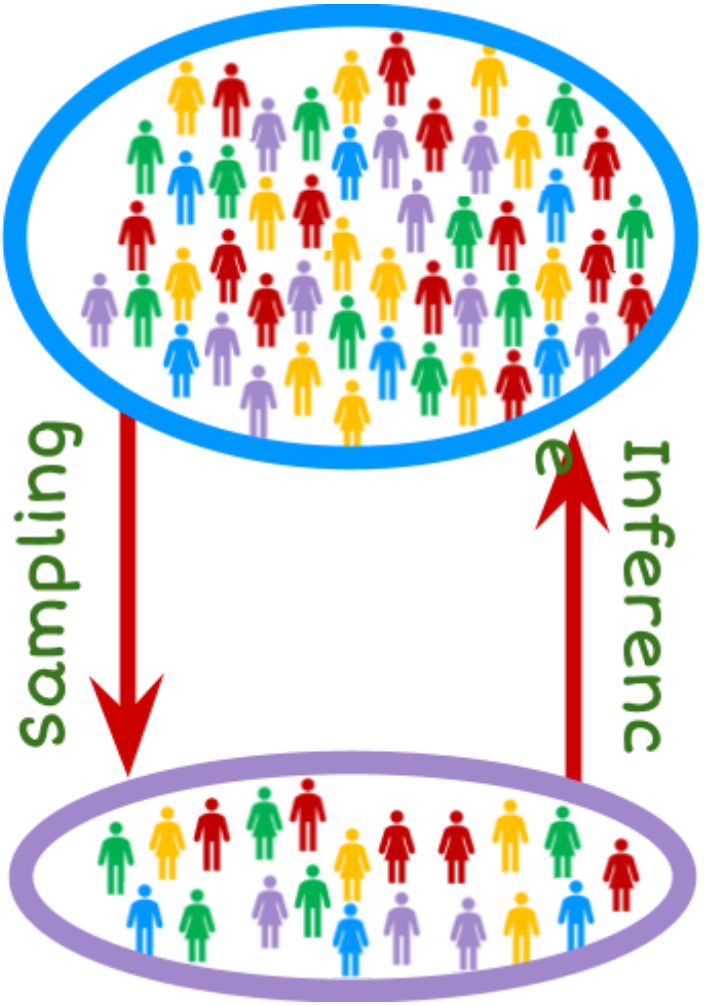
1) Answer the investigation problem and make an inference,
2) Reflect on the data collection process,

Answer the investigation problem

Remember back in the beginning in the Problem section, we wrote an investigation question about the **POPULAITON**.

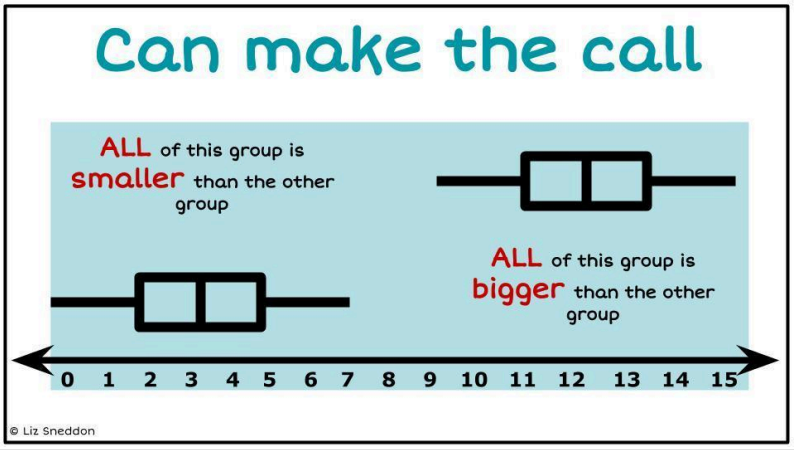
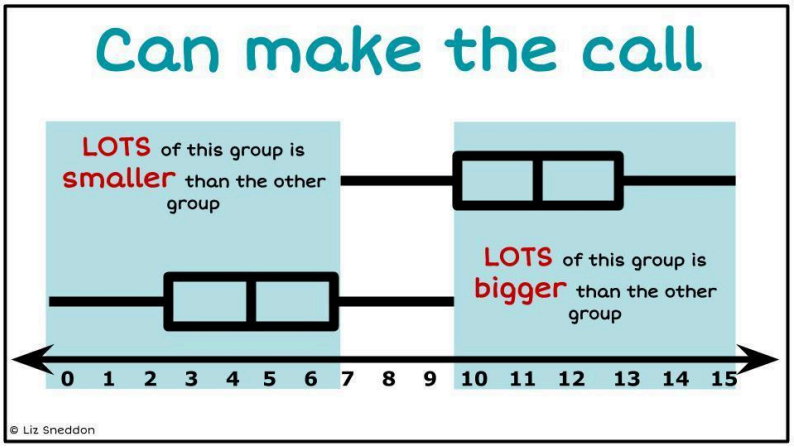
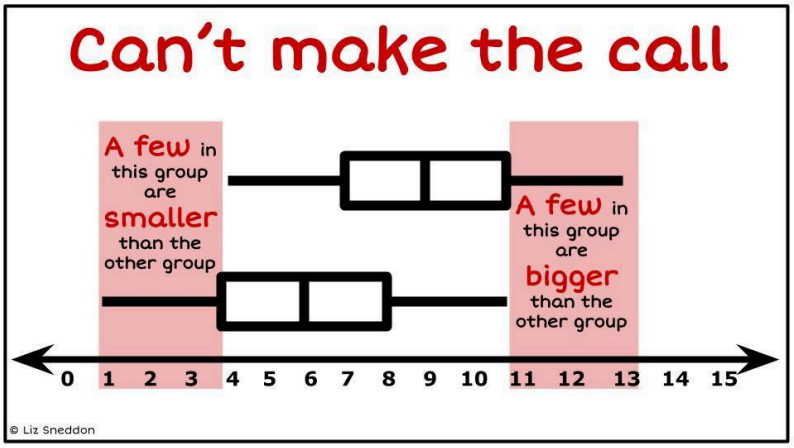
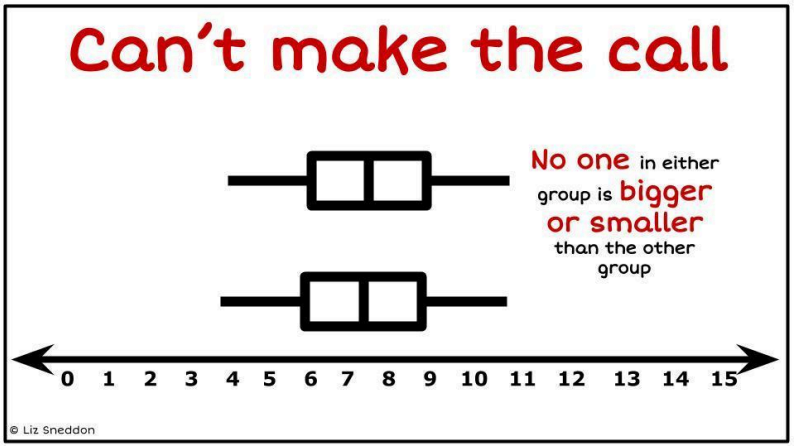
Then in the Plan, Data and Analysis sections we investigated our **SAMPLE**.

Now in the Conclusion section we want to make an Inference (suggestion) about the **population** based on what we found in the **sample**.



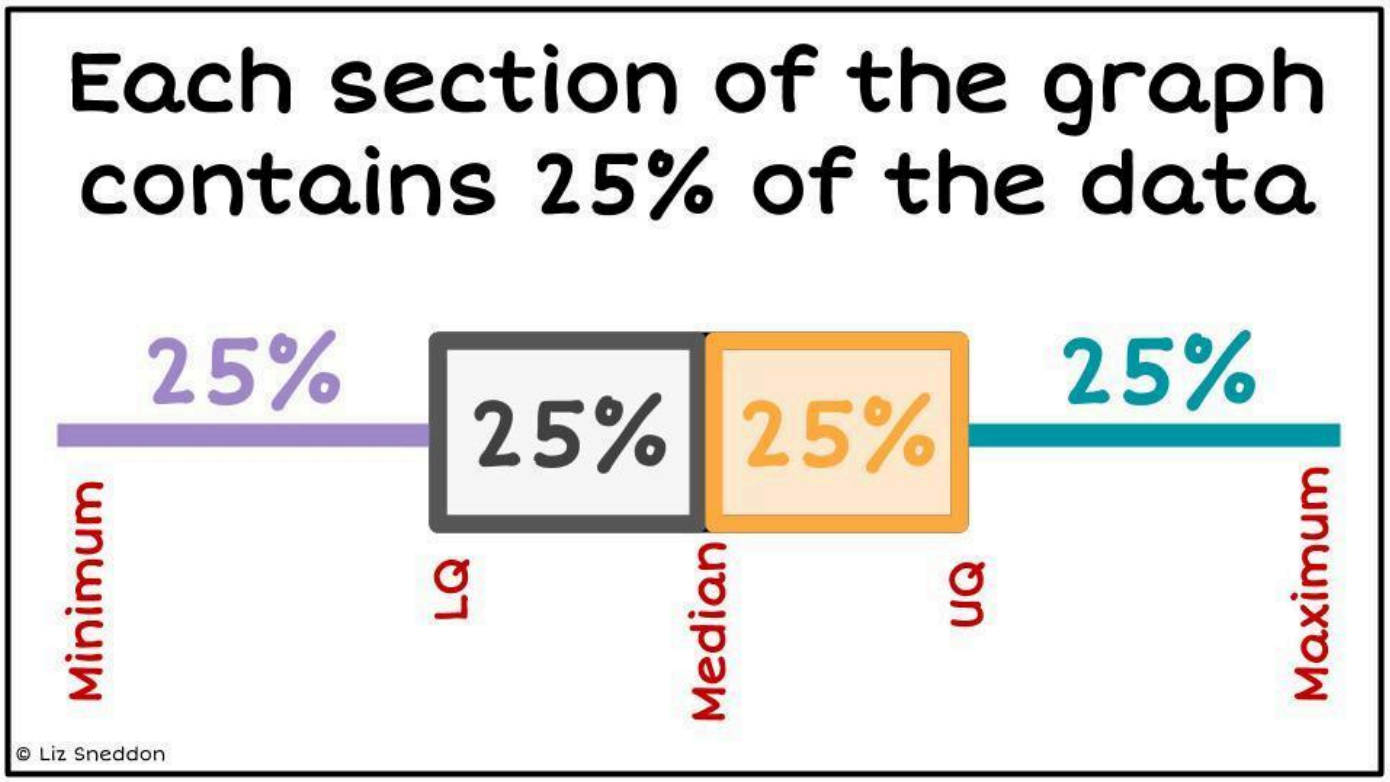
Making the call

A key idea to drawing a conclusion, is to see if LOTS of one group and bigger or smaller than LOTS of another group.

	The data in the two groups don't overlap **AT ALL**.
	The data in the two groups overlap **A LITTLE**.
	The data in the two groups overlap **A LOT**.
	The data in the two groups overlap **COMPLETLY**.

Exercise:

Let's explore this to develop a rule:

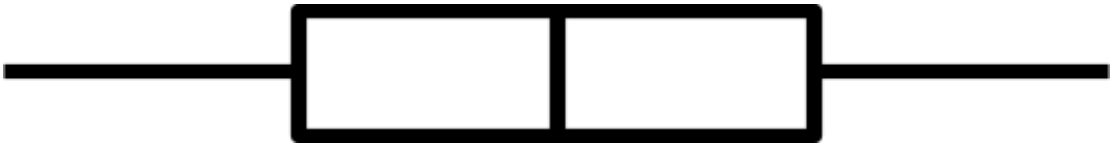


1) What percentage of the data lies **above** the Lower Quartile?

2) What percentage of the data lies **above** the median?

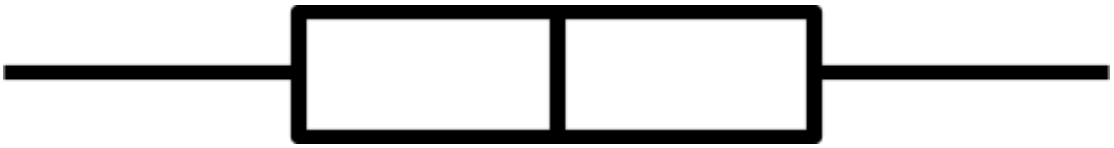
3) What percentage of the data lies **below** the Upper Quartile?

4) For the box plot below, **shade** in the data **above the Lower Quartile**. What percentage have you shaded in?



Percentage above the LQ =

5) For the box plot below, **shade** in the data **below the Median**. What percentage have you shaded in?

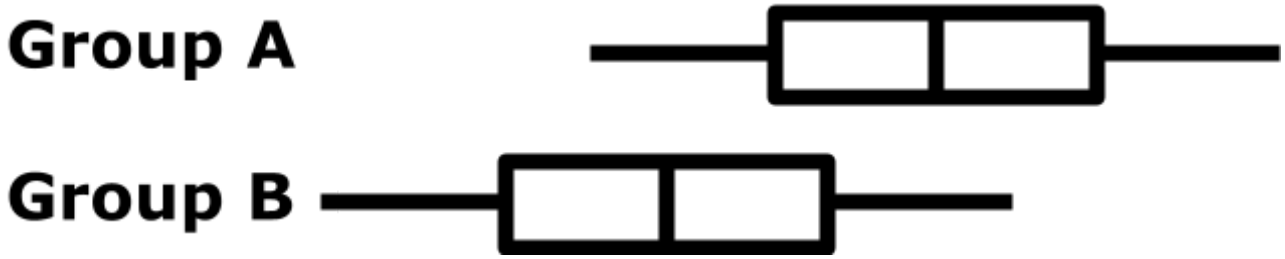


Percentage below the median =

6) Now let's combine both of these.

For the **group A**, shade in the data **above** the **Lower quartile**.

For the **group B**, shade in the data **below** the **median**.



Do you notice that **75%** of the data in the Group A is **bigger** than 50**%** of the data in the Group B? This is the evidence we are looking for. We want to know that **a LOT** of the data in one group is higher, not just a few values.

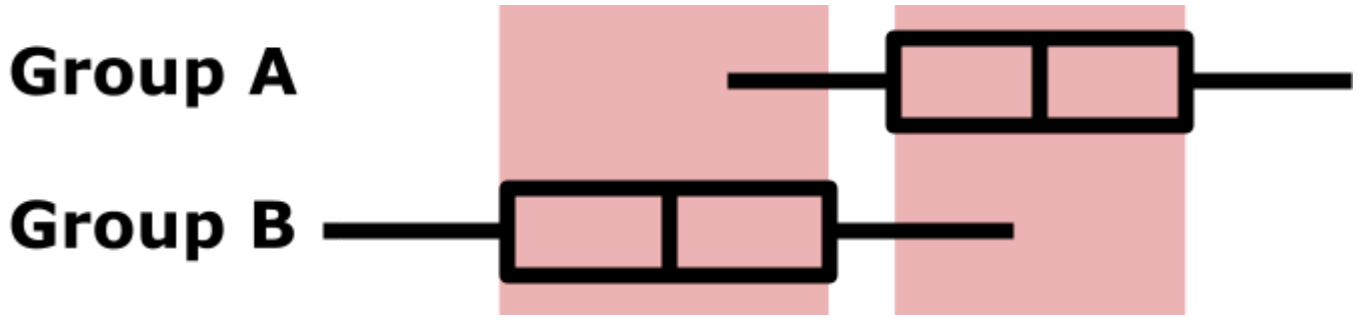
This is called the **¾ - ½** rule or the **75% - 50%** rule.

Rules

The shading idea that we explored above can be simplified to the following rules that embodies these ideas in a method that is easy to use.

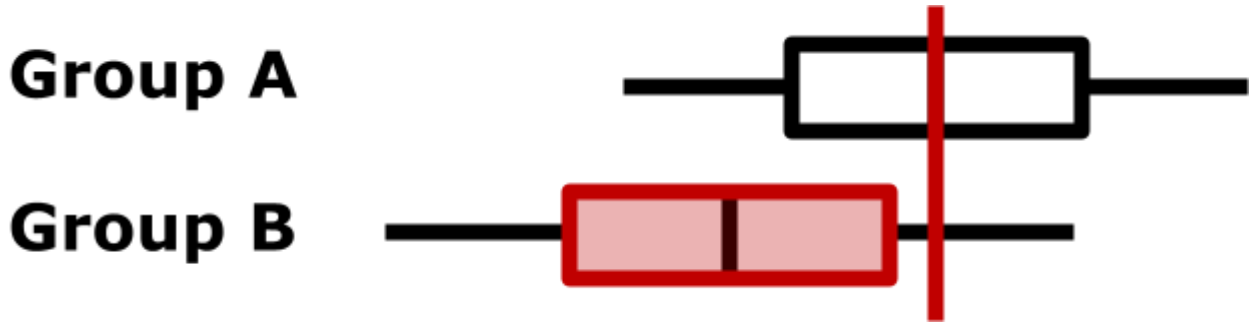
Rule 1:

If there is **no overlap of the boxes**, then the **Group A tends** to be bigger than **Group B**, **for the population**.



Rule 2:

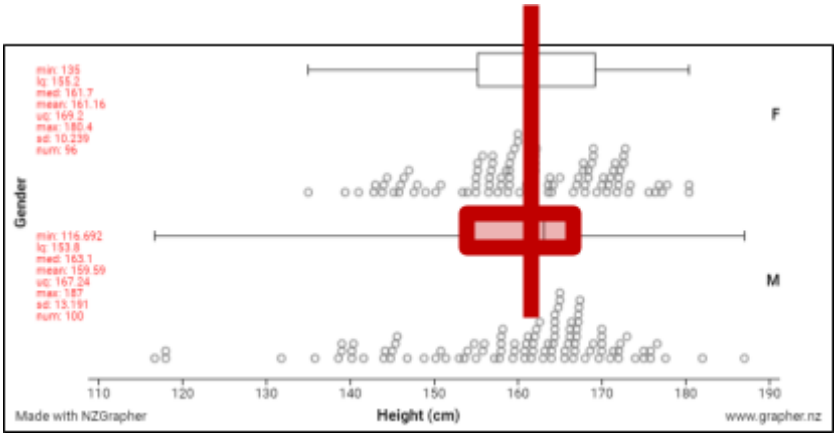
If the **median** for one of the samples **lies outside the box** for the other sample, then **Group A tends** to be bigger than **Group B**, **for the population**.



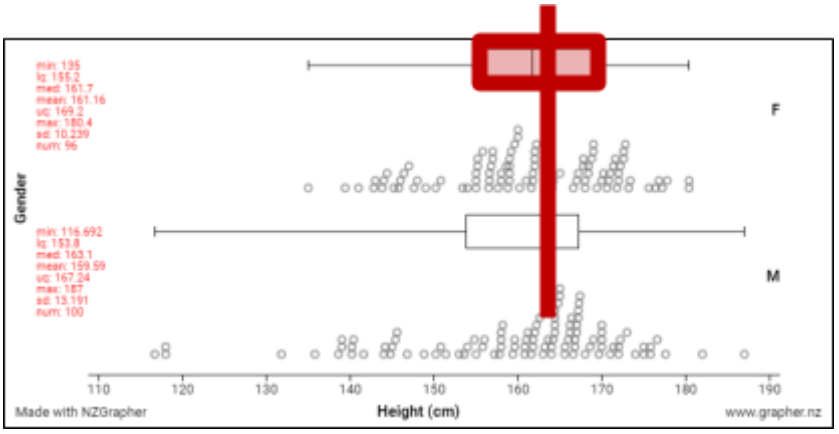
Example:

Can we make the call to say that the heights of boys tend to be taller than girls for **ALL year 11 students at Saint Kentigern College**?

Check the median of the girls. Does it lie inside or outside the box of boys' heights?



Then check for the median of the boys. Does it lie inside or outside the box of girls' heights?



Making the call:

I can't make the call, because the median of boys' height lies inside the box of the girls, and the median of the girls' height lies inside the box of the boys.

Answering the investigation question:

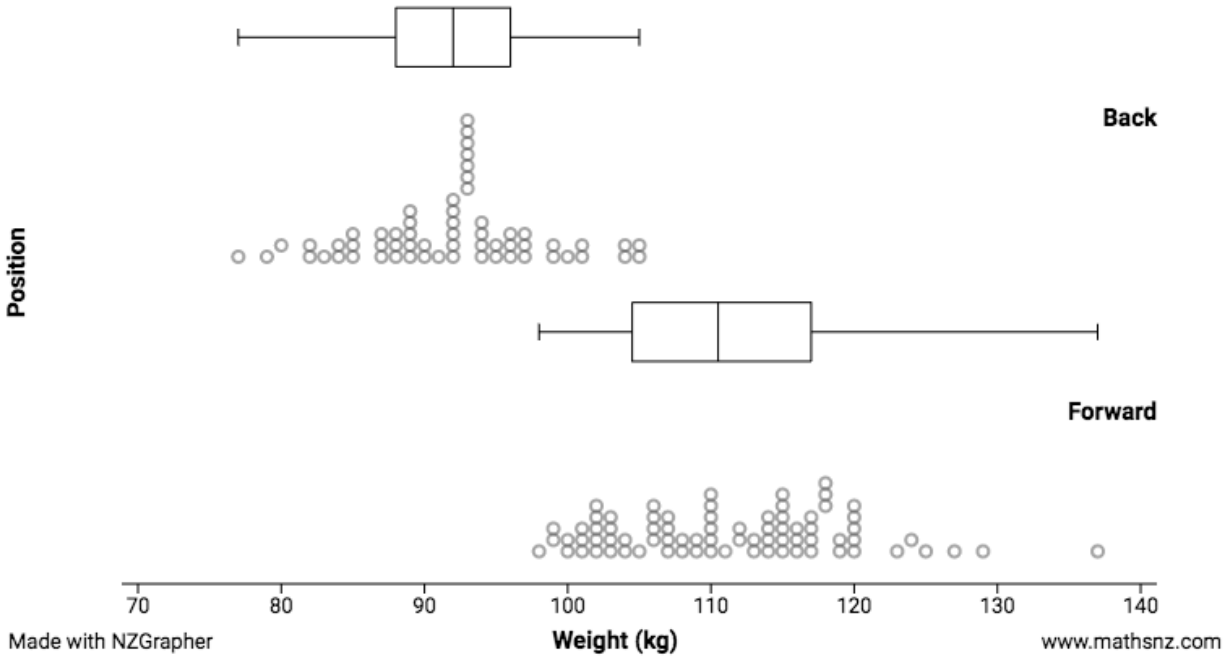
Therefore, I can't make the call that boys tend to be taller than girls, for **ALL year 11 students at Saint Kentigern College**.

Exercise:

Can you make the call for the sample data below, that back in the population, one group tends to be larger or smaller than the other? Decide if you can make the call or not and answer the investigation question.

1) **Problem:**

I wonder if the weight of Forwards tends to be heavier than Backs, **for ALL top rugby players in NZ.**



Does one or both medians lie outside the other box? Yes / No

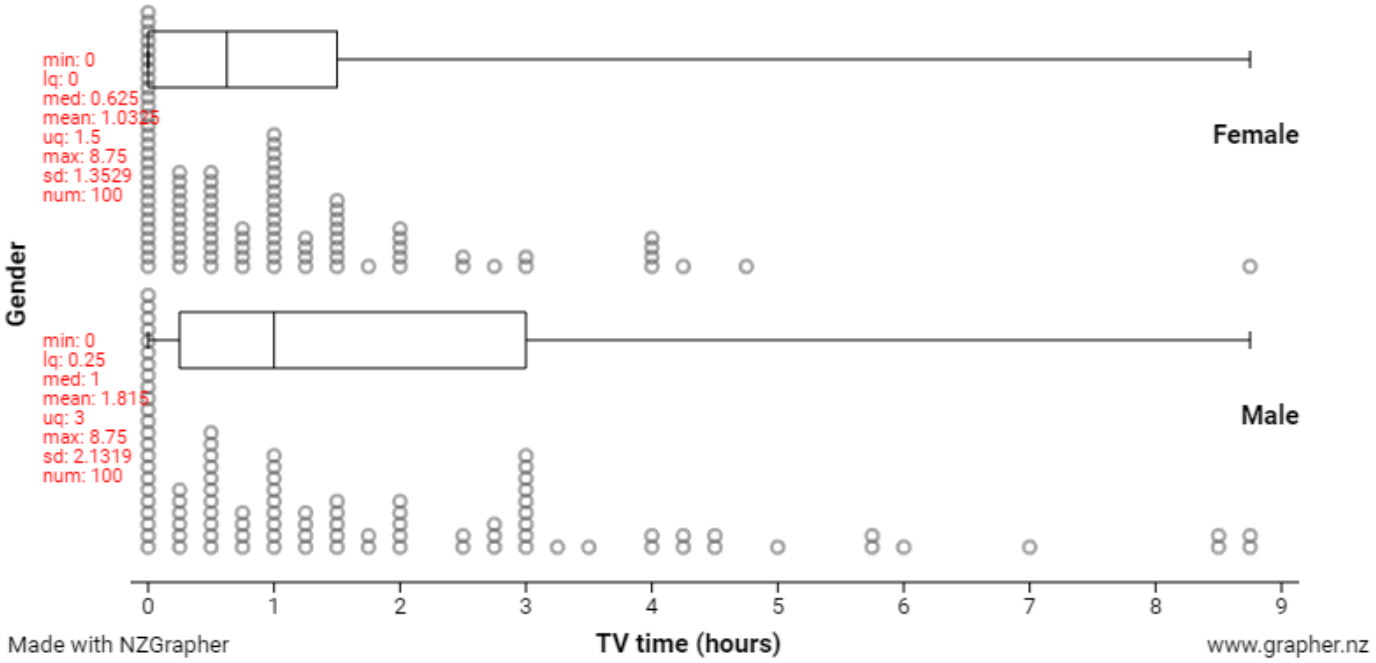
Can you make the call? Yes / No

Answering the investigation question:

I can / can't make the call that the weight of Forwards tends to be heavier than Backs, **for ALL top rugby players in NZ.**

2) **Problem:**

I wonder if the amount of time that girls tend to spend watching TV each day is more than boys, **for ALL high school students in NZ.**

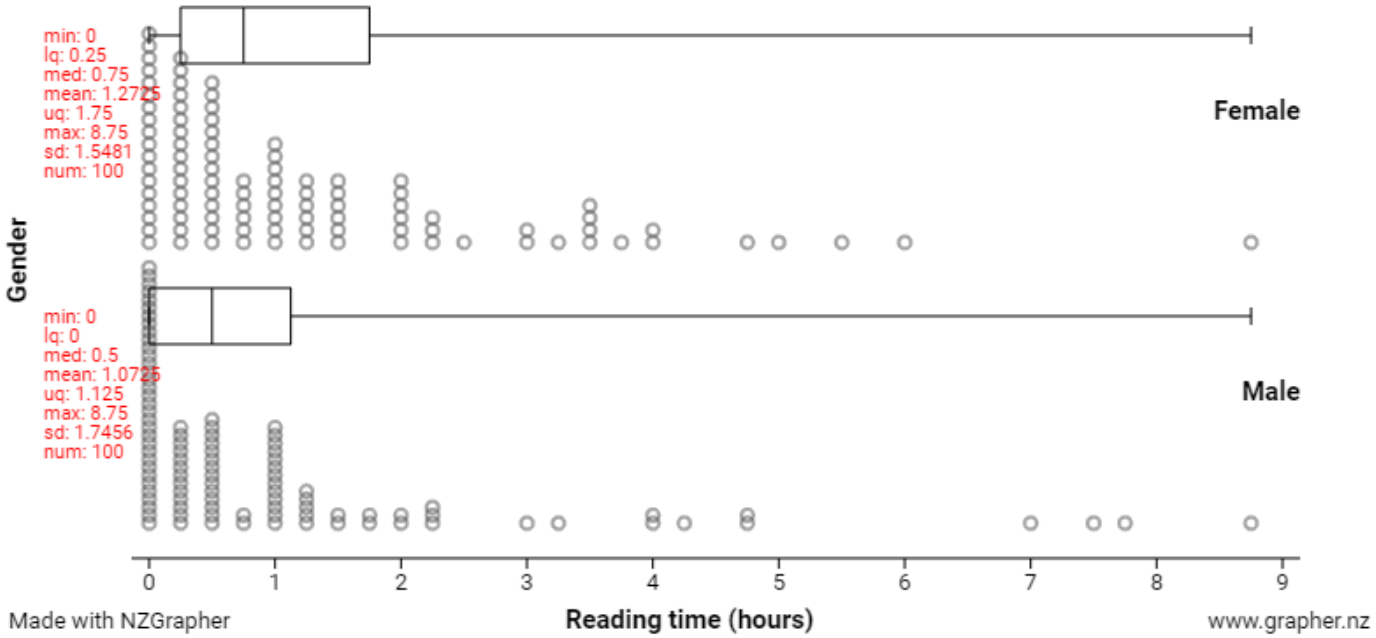


Does one or both medians lie outside the other box?	Yes / No
Can you make the call?	Yes / No

Answering the investigation question:

3) **Problem:**

I wonder if the amount of time that girls tend to spend Reading each day is more than boys, **for ALL high school students in NZ.**



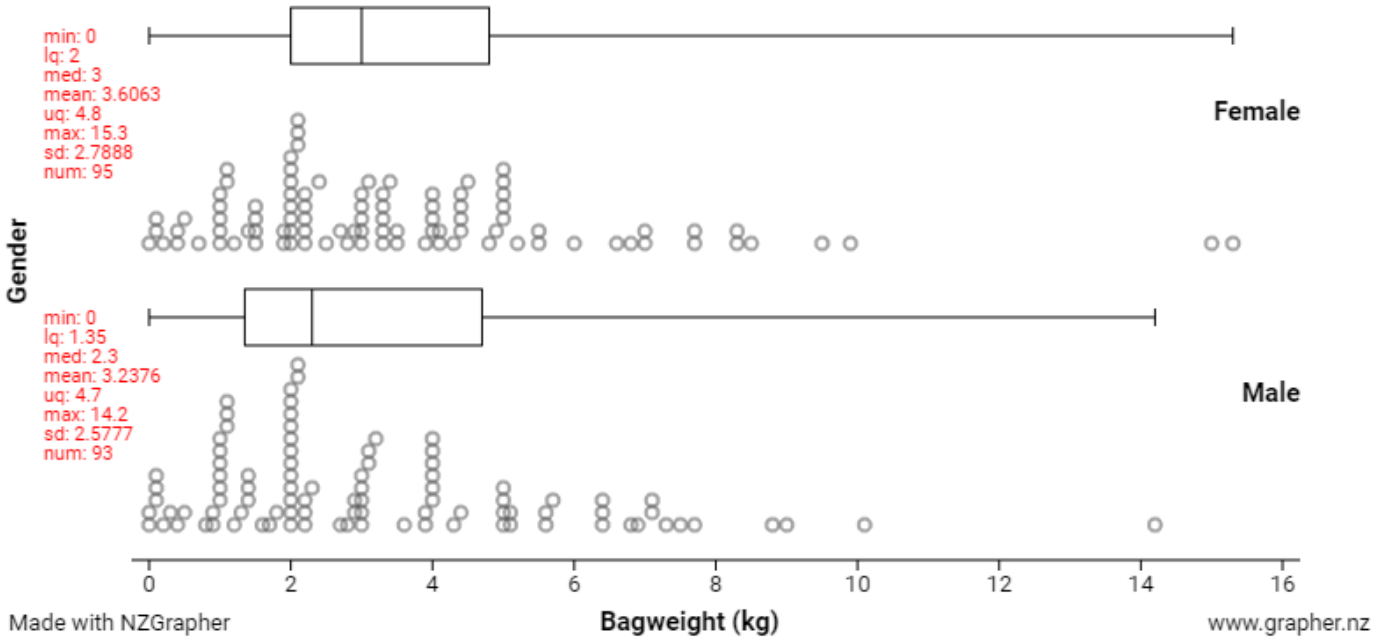
Does one or both medians lie outside the other box? Yes / No

Can you make the call? Yes / No

Answering the investigation question:

4) **Problem:**

I wonder if the weight of school bags for boys tends to be heavier than girls, **for ALL high school students in NZ.**



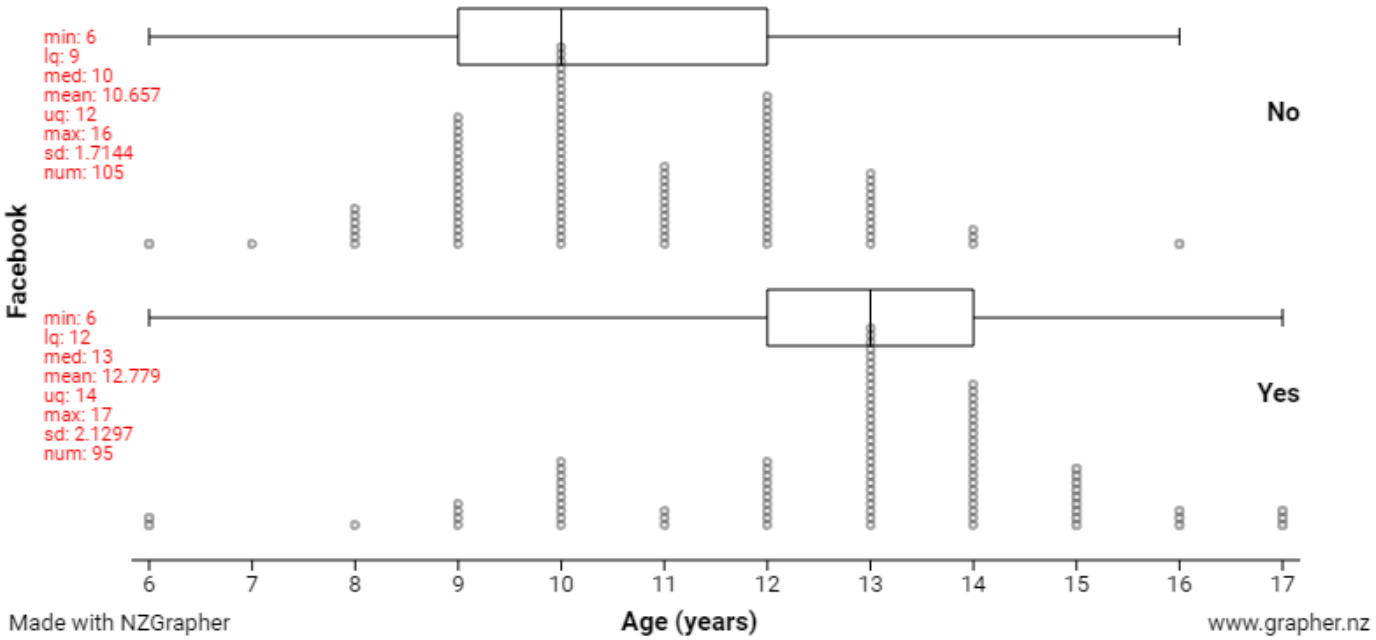
Does one or both medians lie outside the other box? Yes / No

Can you make the call? Yes / No

Answering the investigation question:

5) **Problem:**

I wonder if the age of students who have Facebook tends to be older than the age of students who don't have Facebook, **for ALL high school students in NZ.**



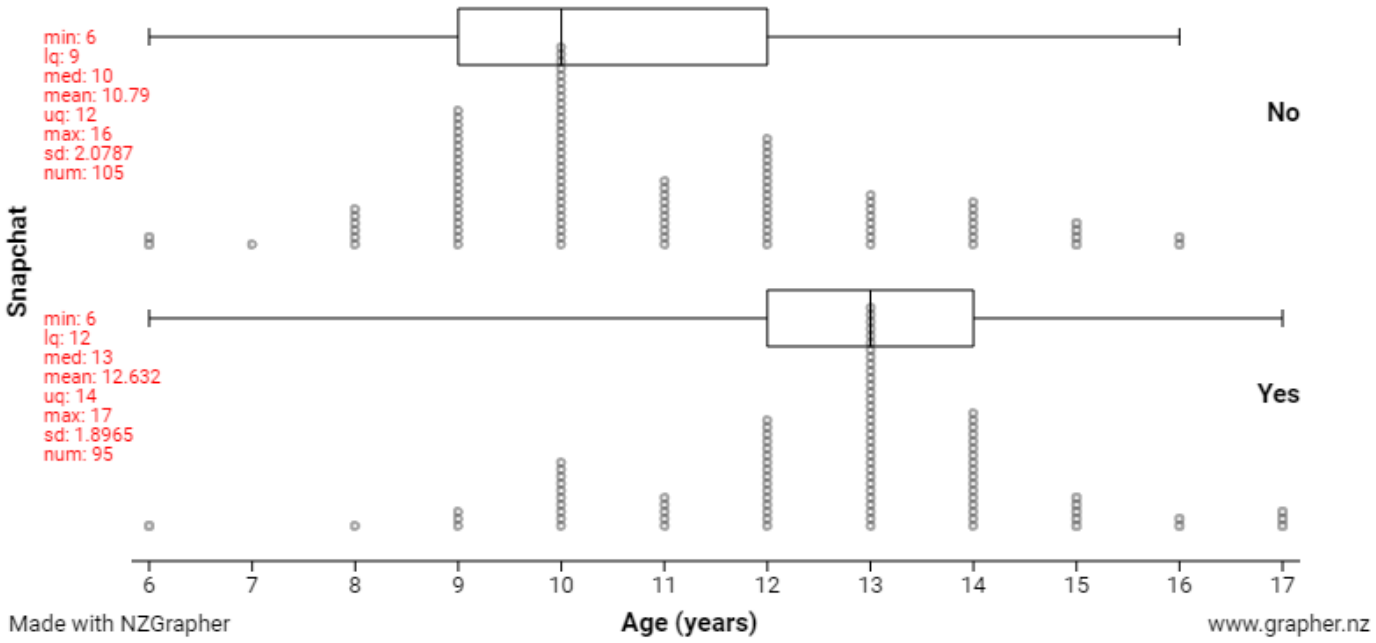
Does one or both medians lie outside the other box? Yes / No

Can you make the call? Yes / No

Answering the investigation question:

6) **Problem:**

I wonder if the age of students who have Snapchat tends to be older than the age of students who don't have Snapchat, **for ALL high school students in NZ.**



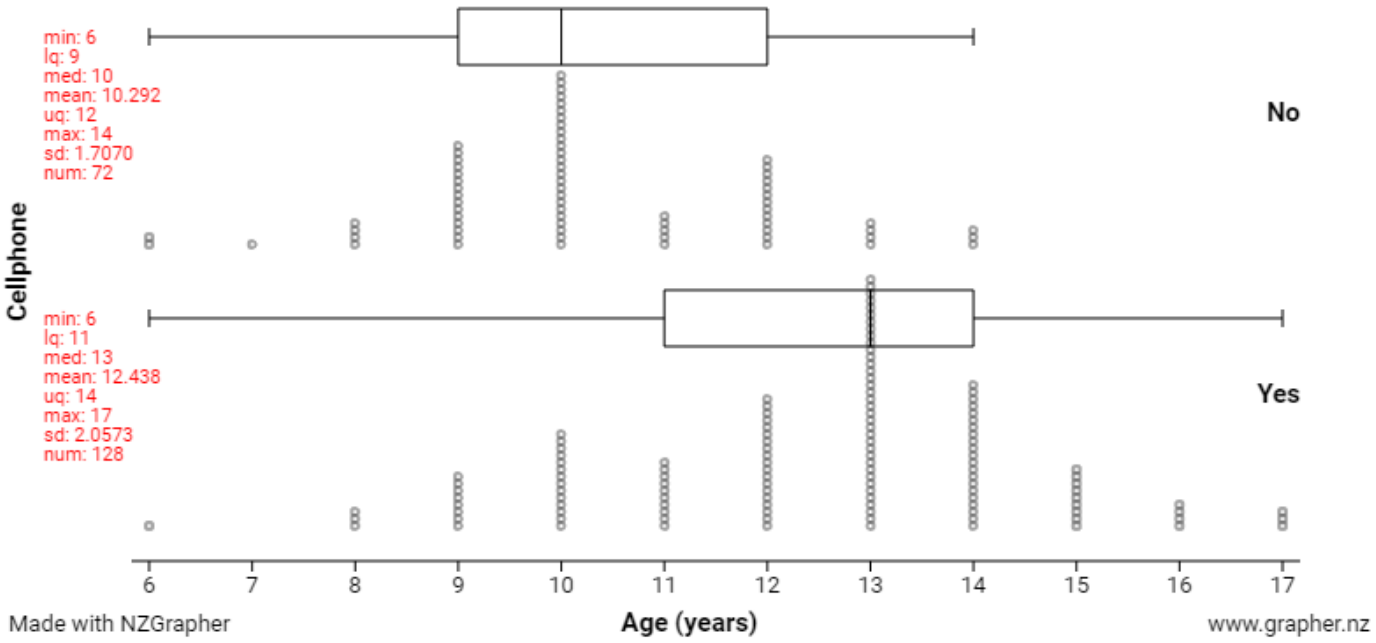
Does one or both medians lie outside the other box? Yes / No

Can you make the call? Yes / No

Answering the investigation question:

7) **Problem:**

I wonder if the age of students who have a Cellphone tends to be older than the age of students who don't have Cellphone, **for ALL high school students in NZ.**



Does one or both medians lie outside the other box? Yes / No

Can you make the call? Yes / No

Answering the investigation question: