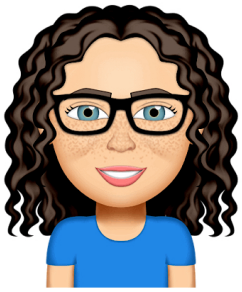
Year 9
Introduction to Data Workbook

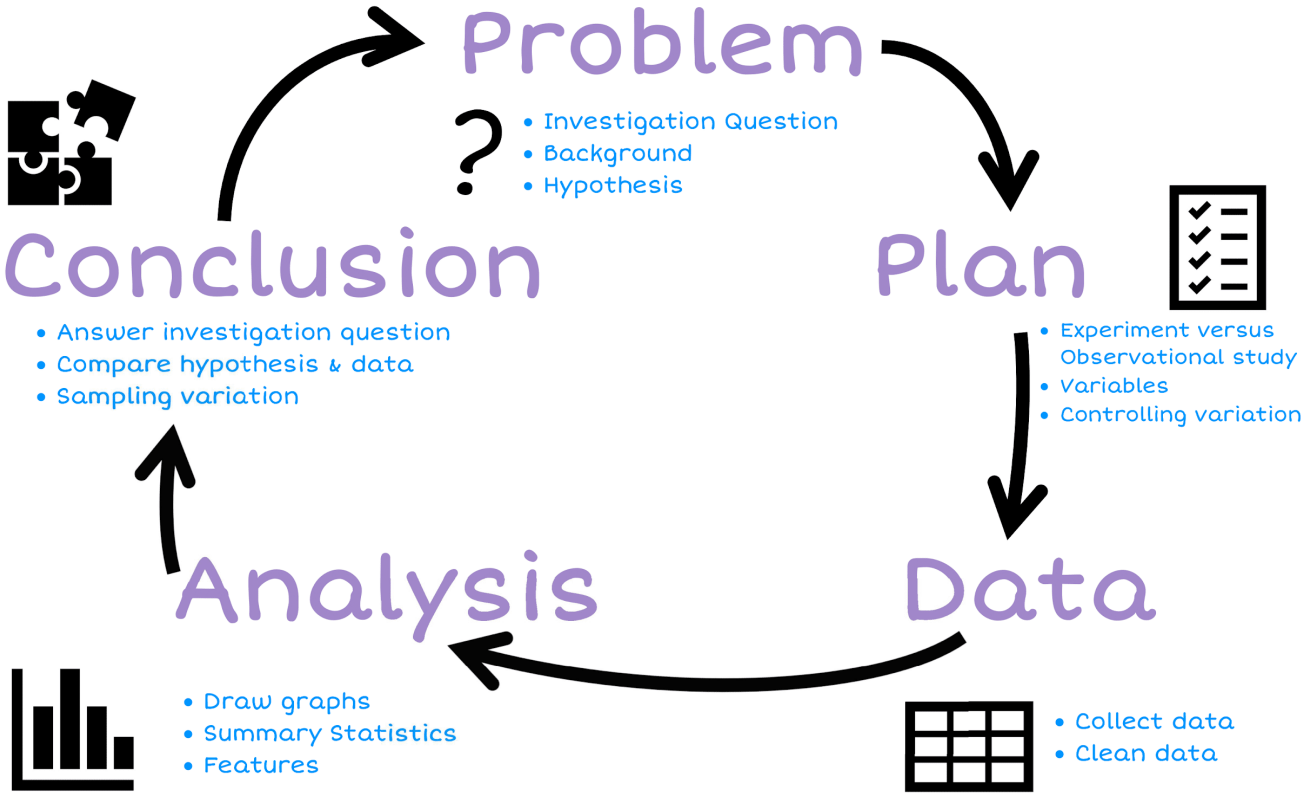
Name:



By Liz Sneddon

Investigative Process

The PPDAC cycle is the core of all statistical investigations.



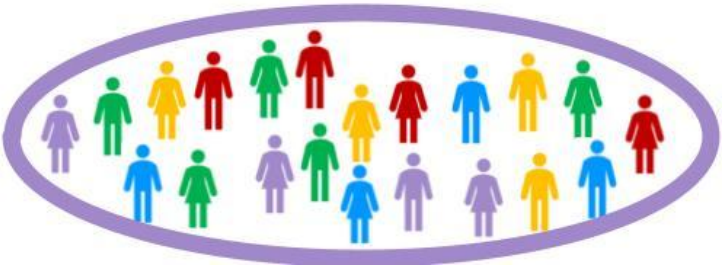
Plan

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.

A population is all the individual members or items that make up a group.

A sample is a group of individuals (or items) selected from the population.

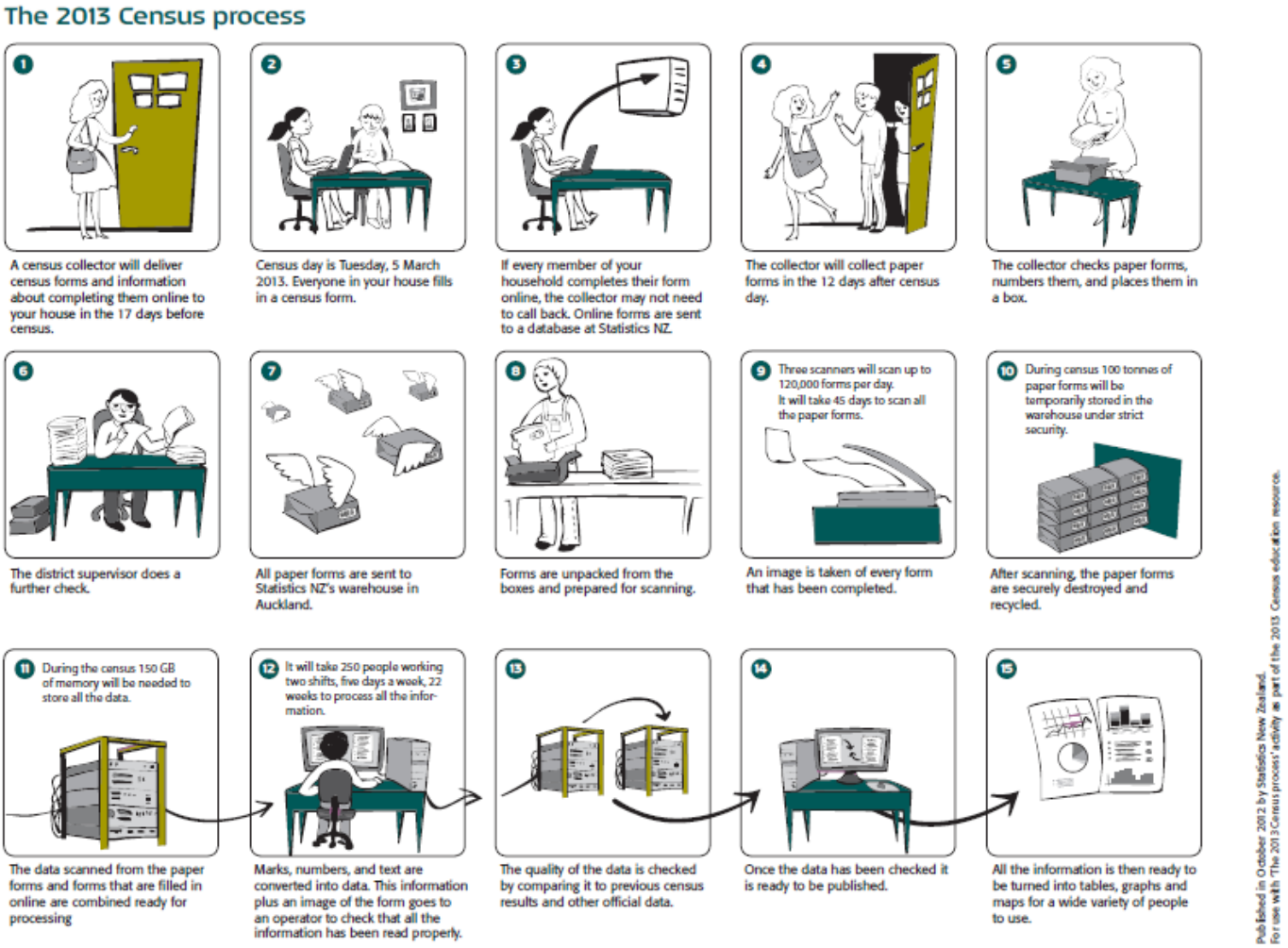


A **census** is a study that attempts to measure every unit in a population.

- The government collects data every 4 years.
- It helps the government, councils and businesses to plan for the future.
- The last census in 2013 cost over $100 million dollars.
- It took more than 6 months to collect the data from every person in NZ (over 4.2 million people).

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

- Where to build new schools (if there are a lot of young children in one area, they will need a school)
- How many hospitals do we need?
- Do families need more financial assistance?



The reasons we take samples are:

- It takes a long **time** to do a census.
- It costs a lot of **money** to collect that much data.

Exercise:

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

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

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

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

A. A selection of logo artists.

B. Every person in the United States.

C. A selection of shoppers from different states.

D. 3,800 children age 5 - 15

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

A. Every person who bought the album.

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

C. 250 girls who bought the album.

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

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

A. Visitors to the 3DS section.

B. All of the website visitors.

C. Visitors to the PS4 section.

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

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

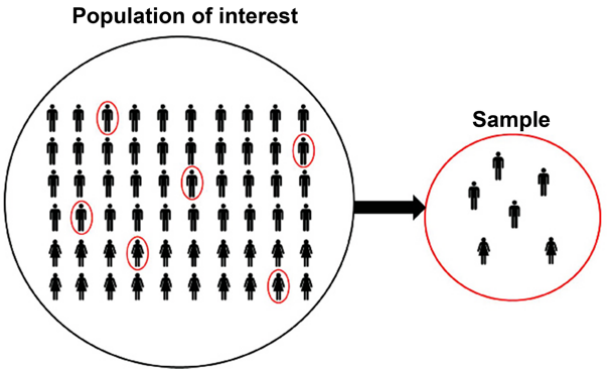
A. A selection of voters over age 50.

B. A selection of male voters.

C. A selection of voters of different ages.

D. All voters

Sampling methods



Data needs to be collected by taking a sample. The sample data will allow us to make estimates about the population without needing the time, money and effort to collect a census.

The sampling method is **HOW** we take a sample from the population.

Samples are selected **randomly** so the characteristics of the sample are typical (**representative**) of the population.

A **random sample** means that each member of the population has the same chance of being selected.

A **biased sample** is not typical of the population. It has a bias for particular members.

A **representative** sample is a group of people who have been selected **randomly**, so that there is a mix of characteristics in the sample that match the population.

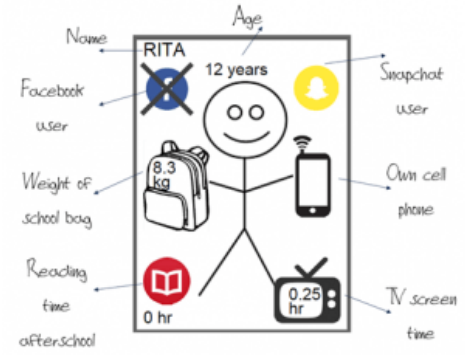
Characteristics may include: a mix of genders, ethnicities, socio-economic status, eye colour, sporting preferences, etc.

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:

1) Go to the Stickland website (https://learning.statistics-is-awesome.org/stickland/). There is an animation running where people in the population are randomly selected to go across the screen (they have a random number on their shirt). Take a random sample of 10 students (by clicking on 10 people) and 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 Cellphone? Yes / No**	**Reading time yesterday (hours)**	**TV time yesterday (hours)**

2) Why is it important for our samples to be randomly selected?

3) A school has about 690 students. The school wants to do a survey on the use of phones by students. For the methods below, state if they are biased or representative samples, and explain why.

a) Interviewing all students in a Year 9 class.

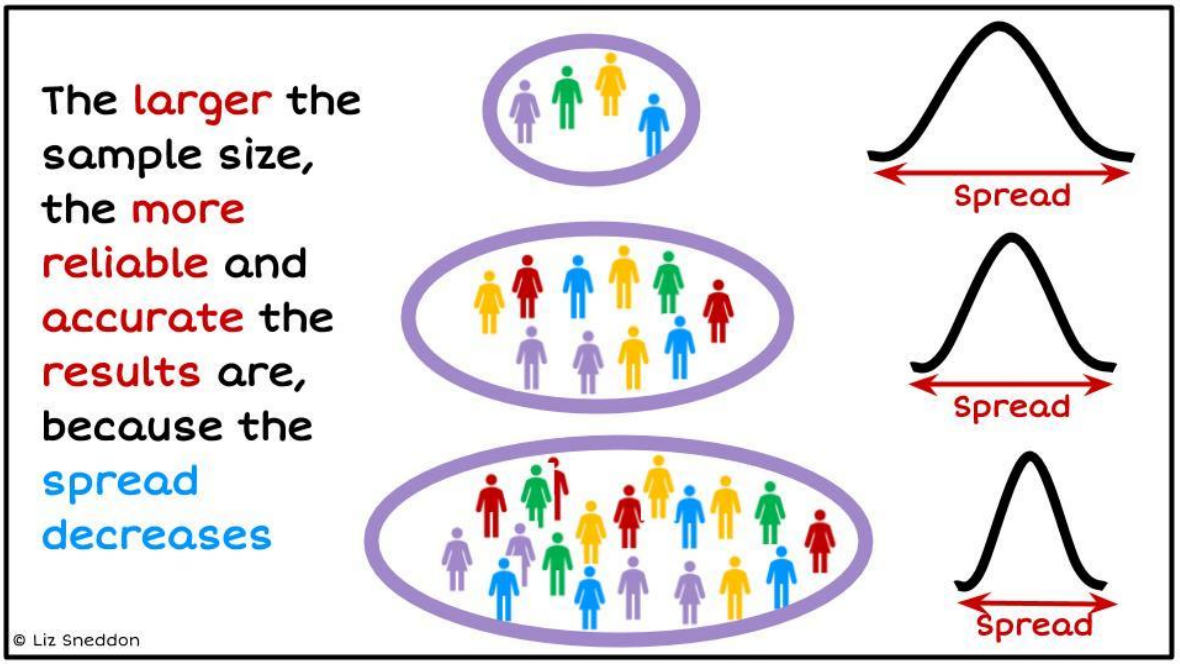
b) Interviewing every 20th student leaving school at the end of the day.

c) Asking for 40 volunteers to fill in a questionnaire.

d) Interviewing 40 students at a sports game on Wednesday afternoon.

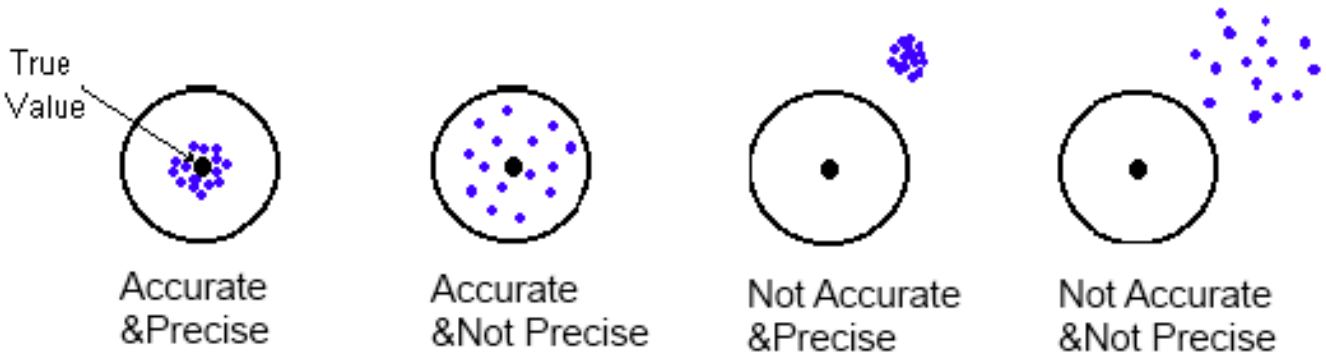
Sample size

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



The image below talks about accuracy and precision.

In statistics we want to be accurate (we want our sample to be a close estimate of the population), and we want to be precise (when we collect data we want to control the variation as much as possible).



Exercise:

1) Circle the words that complete the sentences below.

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

   b) Larger sample sizes take a _____**shorter / longer**_____ time to collect data, and are ___**more / less** reliable.

2) Mrs Sneddon is going to survey 35 girls and 40 boys at a local primary school to investigate their use of ipads at home. Does it matter that there are a different number of girls and boys? Why / why not?

Observational versus Experimental studies

We need to understand what methods we can use to collect data, what the different data types are and how to organise our data.

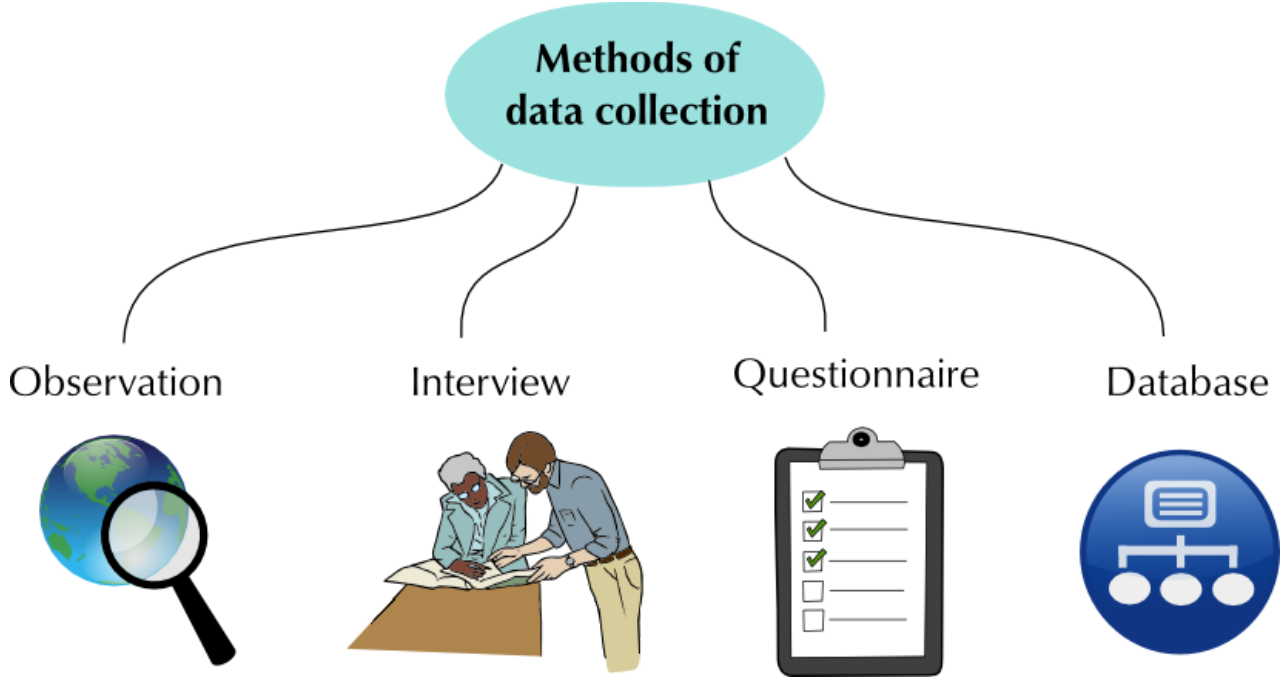
Data can either be from an **observational study** or an **experimental study**.

An **observational study** is where the population is observed without any interference by the investigation.

An **experimental study** is where the investigator randomly assigns people into one of two groups, controlling all other conditions.

Data Collection Methods

Observational data can be collected in several different ways:



Exercise:

A questionnaire or survey is one way to collect data. Complete the survey questions below.

How old are you (in years)?	What is your gender?
❑ 12 ❑ 13 ❑ 14 ❑ 15 ❑ 16	❑ Male ❑ Female

Which of the following devices do you have? (Tick all that apply)	Which of the following social media platforms do you use? (Tick all that apply.)
❑ Own cell phone ❑ Own computer or laptop ❑ Family computer or laptop ❑ None of the above	❑ Facebook ❑ Twitter ❑ Instagram ❑ Snapchat ❑ None of the above

For the last school day, estimate how many **minutes** you spent on the following:

Computer time	**TV time**	**Gaming time**	**Phone time**

Now fill in the following Google Form:

http://bit.ly/Year9MediaSurvey

The data will be recorded automatically on a Google Sheet.

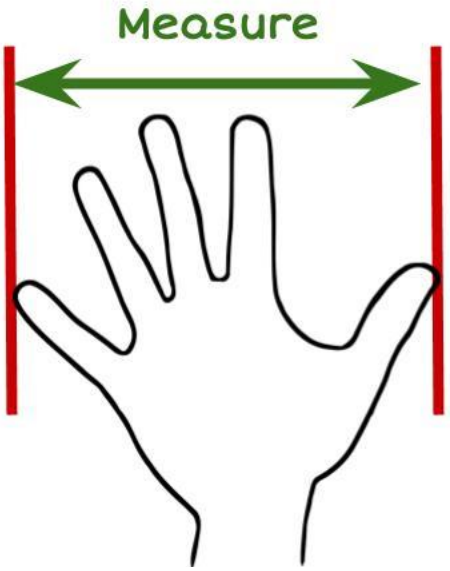
Exercise:

Follow the instructions below to measure your handspan and right foot length.

Instructions

For measuring foot length:

1) Collect a piece of paper, a pen and a ruler.
2) If you have shoes on, remove your right shoe.
3) Place the paper on the floor so it is flat.
4) Place your right foot on the paper.
5) Using the pen, draw a line at the back of your heel and at your longest toe.
6) Using your ruler, measure the distance (in cm) from the heel to the longest toe.
7) Record this measurement here: _____



For measuring hand span:

1) Collect a piece of paper, a pen, and a ruler.
2) Put the piece of paper on a flat surface (e.g. a table)
3) Place your right hand flat on the paper, palm down.
4) Spread your fingers as wide as they can.
5) Using the pen, draw a line at the edge of your smallest finger and at the edge of your thumb.
6) Using your ruler, measure the distance (in cm) from the smallest finger to the thumb.
7) Record this measurement here: _____

Now enter the data into the spreadsheet (link below), so we can collect data from all the students in the class.

http://bit.ly/HandSpan2021

Exercise:

Look at the instructions for measuring the length of your foot and answer these questions.

1) Why would I ask you to remove your shoe? Explain.

2) Is your right foot the same length as your left foot? Can you explain why/why not?

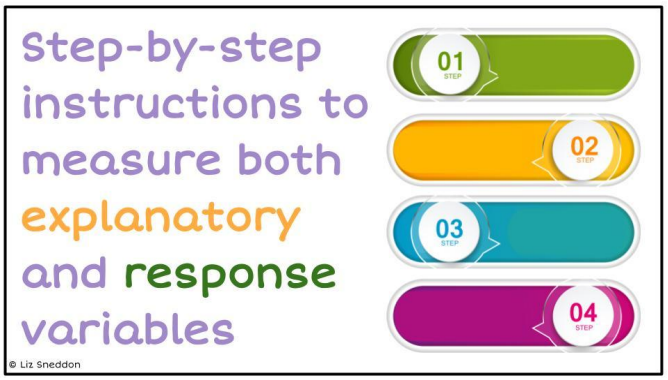
   (**Hint:** think about whether you are right or left handed and which side of your body would be stronger and used more)

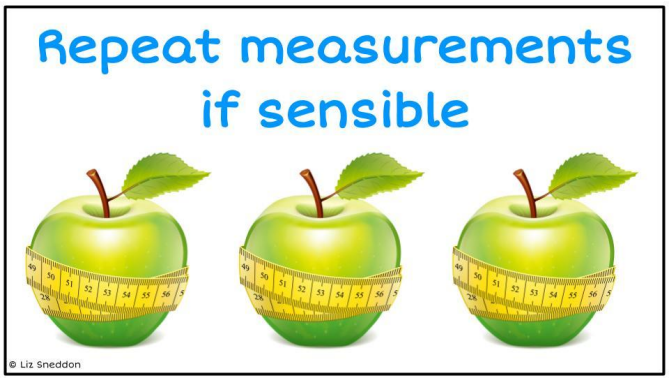
3) Why should we take measurements from people with small **and** big feet? (E.g. young and older people).

Instructions & controlling sources of variation

When coming up with your plan, you need to think about how you can minimise the amount of variation - making sure that all the measurements are done in the same way.

Here are some things to think about:





Example:

When measuring foot length and handspan, some of the factors I will control are:

- Using the same measuring tape, so that all the measurements are consistent.
- Getting students to put their hand down on a piece of paper, so that their hand is as flat as possible. This will make the measurements consistent.
- Get students to take their shoe off when I measure the length of their foot, because the different shoes people wear could have a different end (e.g. pointed, flat, curved) which would change the measurements and not be an accurate measurement of the length of their foot.

Exercise:

Problem 1

I wonder if there is a relationship between a person's **height** and **weight** for students in your class?

Plan

Write a detailed set of instructions of how you would measure a students height and weight, describe some sources of variation that you will control and explain how you will control them.

Variable 1: Height_____

Variable 2: Weight_____

Instructions:

Controlling sources of variation:

Problem 2
I wonder if girls' hair tends to be longer than boys' hair, for students in our class?

Plan

Write a detailed set of instructions of how you would measure gender and hair length, describe some sources of variation that you will control and explain how you will control them.

Variable 1: Gender

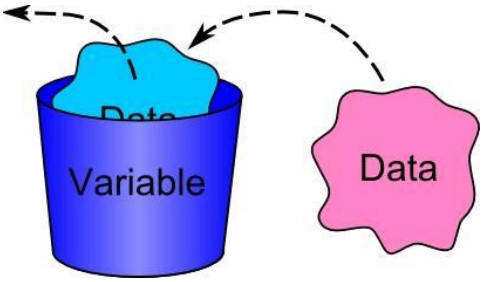
Variable 2: Hair length

Instructions:

Controlling sources of variation:

Variables and Data

A **variable** describes a characteristic of an individual from the population. The characteristic changes or varies from one individual to another.



Data is collected when the values of variables are recorded for individuals.

Example:

Here is a spreadsheet that collected fitness information from students:

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

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

Each column is a variable.

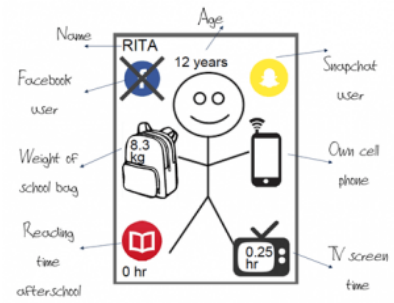
In this example there are **6 variables**:

- Students first name,
- Age,
- Gender,
- Whether they take PE this year,
- Wall sit time, and
- Height.

Exercise:

Name the variables for the data you collected a sample from previously.

1) 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)**
KATIE	10	1.5	no	no	no	1.25	1.5
EMILY	12	yes	no	3.2	yes	0.75	2

Variables:

2) The Social Media survey:

Age	Gender	Devices	Social Media	Computer Time	TV Time	Gaming Time	Phone Time
❑ 12 ❑ 13 ❑ 14 ❑ 15 ❑ 16	❑ Male ❑ Female	❑ Own cell phone ❑ Own computer or laptop ❑ Family computer or laptop ❑ None of the above	❑ Facebook ❑ Twitter ❑ Instagram ❑ Snapchat ❑ None of the above				

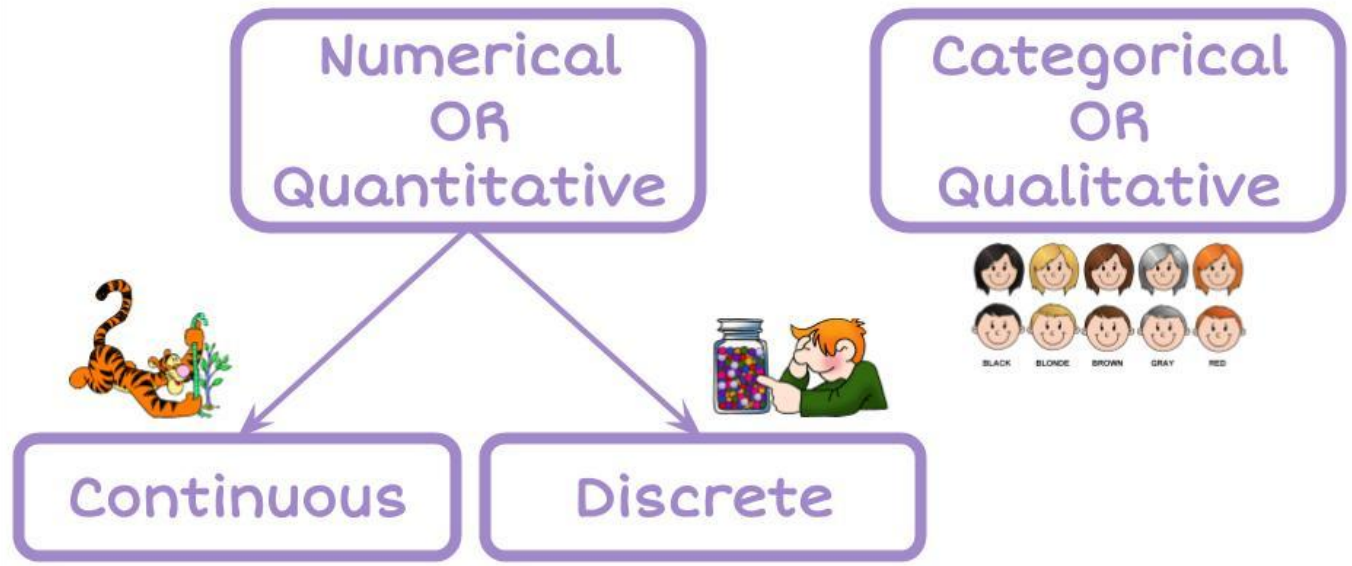
Variables:

Data Types

Categorical (groups) variables are characteristics, that cannot be described by numbers e.g. gender, ethnicity, apple variety. They can also be called **qualitative variables.**

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

- **Discrete variables** (whole numbers), values obtained by counting.
- **Continuous variables** (measurement), values obtained by measuring.



Example:

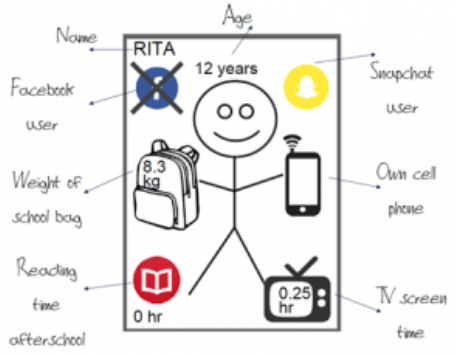
Here is the Wall sit spreadsheet:

	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
Data type	Categorical	Numerical • Discrete	Categorical	Categorical	Numerical • Continuous	Numerical • Continuous

Exercise:

1) Name the data types for the Stickland spreadsheet:

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)



2) Social Media dataset

Age	Gender	Devices	Social Media	Computer Time	TV Time	Gaming Time	Phone Time

3) Hand span and Foot length dataset

Gender	Right or left handed	Foot length (cm)	Hand span (cm)

Cleaning data

Look for the following issues:

- Data entry mistakes
- Incorrect units
- Missing data

But, you **CANNOT** change/delete data unless you **KNOW** that it is a mistake.

If you are **CERTAIN** the data is wrong, then make the cell blank (or enter a 0).

Exercise:

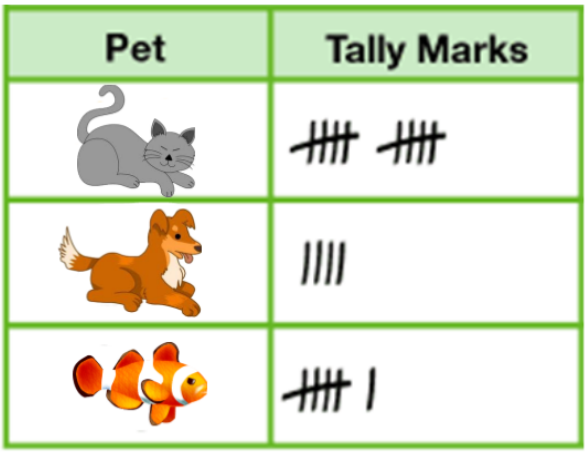
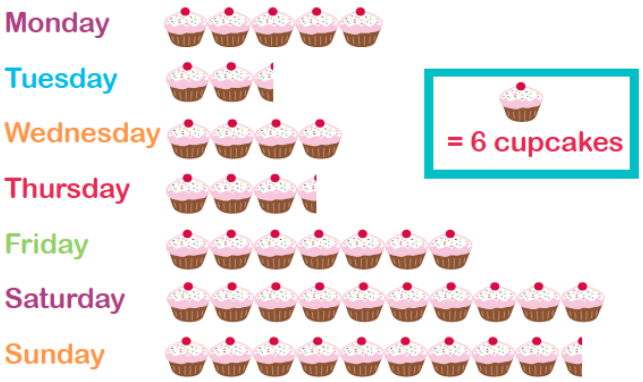
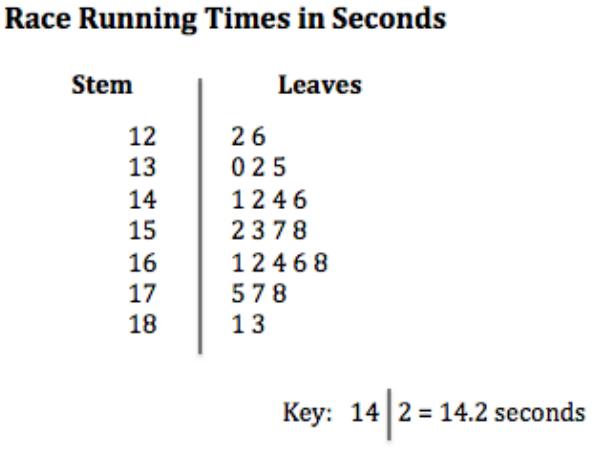
Find any data that doesn't make sense and highlight the values. State what corrections or changes you would make.

Gender	Age	Country of birth	Languages spoken	Height	Right foot length	Arm span	Index finger length
girl	14	Russia	1	149	220	115	5
boy	11	NZ	1	141	22	142	65
girl	14	NZ	2	175	255	176	81
girl	13	NZ	1	162	25	64	80
girl	1	NZ	1	158	25	163	97
girl	12	NZ	1	164	28	1	80
	13	NZ	2	166	26	180	100
girl	12	cookisl	1	154	23	156	49
girl	14	NZ	-1	170	26	1	70
girl	12	India	1	0	21	153	8

Corrections / changes

Data displays

Here are some common ways to display data.

Tally chart	Frequency table
	
Pictogram	**Stem and leaf**
	

Let's look at these in more detail, so that you know how to make them.

Tally chart & Frequency table

Tally charts and frequency tables are useful when you want to summarise data into categories (e.g. groups or word answers)

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

Tally chart

Gender	Tally marks
Female	III
Male	II

Frequency table

Gender	Tally marks	Frequency
Female	III	3
Male	II	2

Exercise:

1) Using the Wall sit data above, make a tally chart and frequency table of the PE variable.

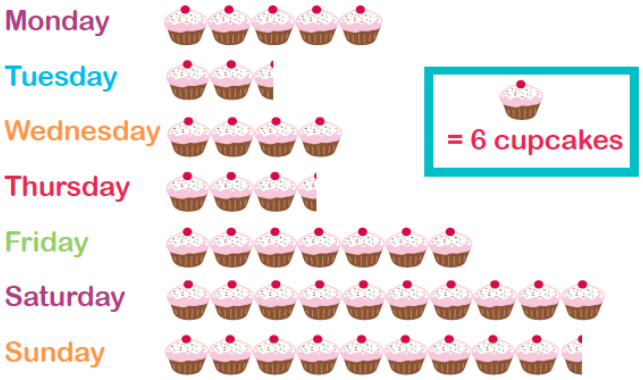
PE	Tally marks	Frequency

2) Using the Stickland data given below, make a tally chart and frequency table of each categorical variable (Facebook, Snapchat and Cellphone).

Age	Facebook	Snapchat	Bag weight	Cellphone	Reading time	TV time
6	no	yes	5	yes	0.25	4.25
13	no	no	5.8	yes	1	0.25
8	yes	yes	2.4	yes	0	0
14	yes	yes	1.1	yes	0	0
16	yes	yes	4.1	yes	0	2
6	yes	no	1	yes	3.25	3
11	yes	yes	3	yes	0	1
10	yes	no	3	yes	0.25	1
14	no	no	5	yes	0	1.5
10	no	no	3.7	no	0.25	3

Pictogram

Pictograms use pictures to represent data. You need to have a key/legend to state how many units each picture represents.

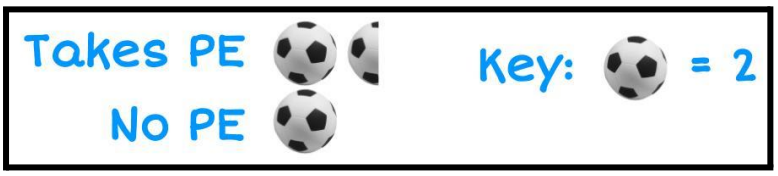


Example:

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

Pictogram of the PE variable



Exercise:

1) Using the Wall sit data above, make a pictogram of the Gender variable.

Male	
Female	

2) Using the Stickland data given below, make a pictogram for each categorical variable (Facebook, Snapchat and Cellphone).

Age	Facebook	Snapchat	Bag weight	Cellphone	Reading time	TV time
6	no	yes	5	yes	0.25	4.25
13	no	no	5.8	yes	1	0.25
8	yes	yes	2.4	yes	0	0
14	yes	yes	1.1	yes	0	0
16	yes	yes	4.1	yes	0	2
6	yes	no	1	yes	3.25	3
11	yes	yes	3	yes	0	1
10	yes	no	3	yes	0.25	1
14	no	no	5	yes	0	1.5
10	no	no	3.7	no	0.25	3

Stem and leaf

A stem and leaf is a way to summarise a lot of **numeric** data in a graphical type format. It works well when you don't have too much data.

A key is necessary.

Each piece of data is split into a stem part and a leaf part. The leaf part will only have 1 digit in it, and the rest of the number goes into the stem.

Then you put the data in order with the smallest number on the left of the leaf.

Race Running Times in Seconds

Stem	Leaves
12	2 6
13	0 2 5
14	1 2 4 6
15	2 3 7 8
16	1 2 4 6 8
17	5 7 8
18	1 3

Key: 14 \| 2 = 14.2 seconds

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

Stem and leaf of height variable

E.g. split 161 into a stem of 16 | 1

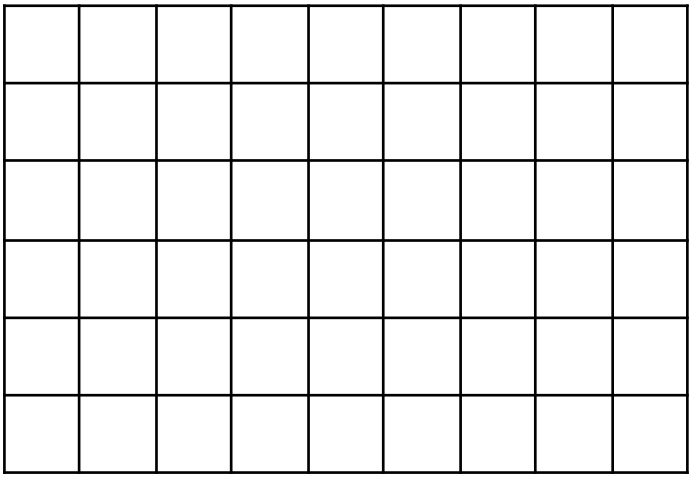
15	5
16	1 9
17	
18	2 5

Key: 15 \| 5 = 155cm

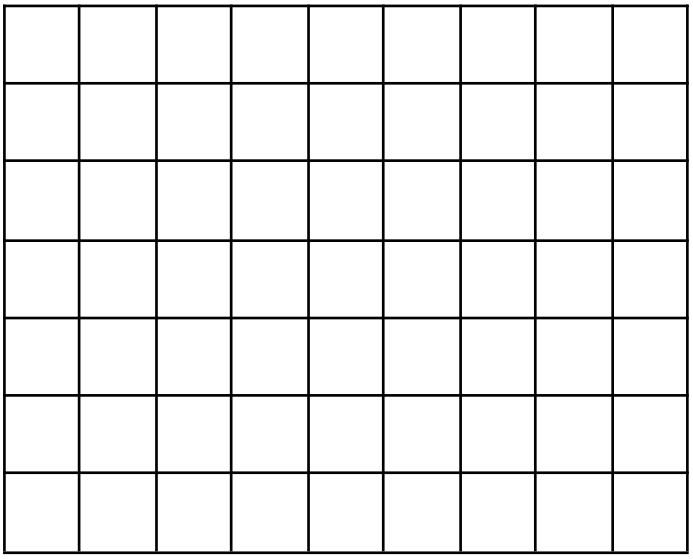
Exercise:

Make stem and leaf plots of the data in each question below.

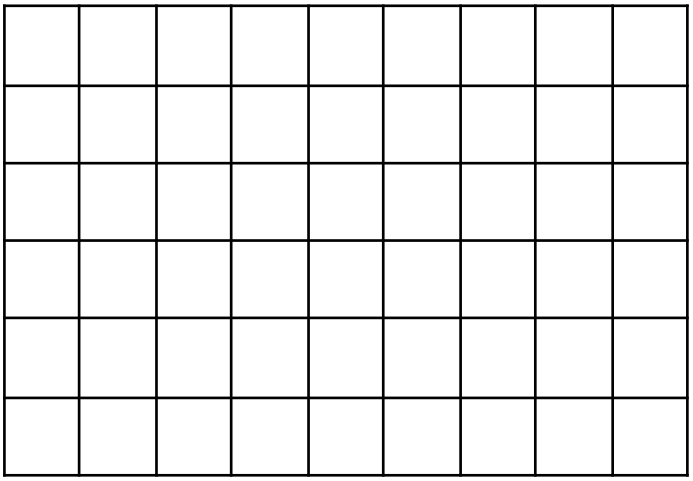
1) 25, 28, 30, 31, 33, 35, 37, 38, 37, 40, 41, 42, 42 ,43, 45, 45, 47



2) 12, 18, 22, 24, 29, 31, 37, 39, 42, 45, 49, 52, 57, 60, 62, 64, 66, 71, 73, 75



3) 255, 258, 262, 262, 267, 268, 269, 271, 276, 281, 293, 295, 301, 307



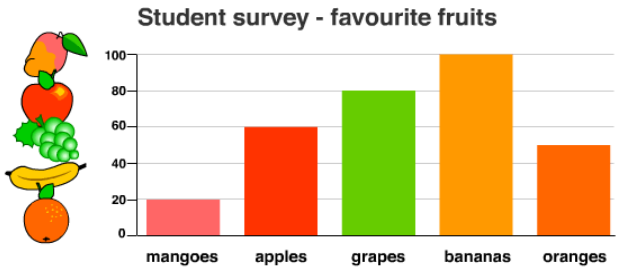
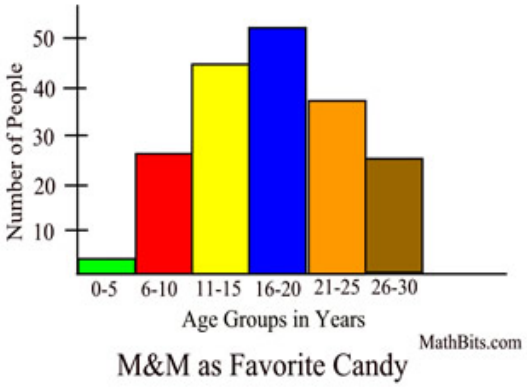
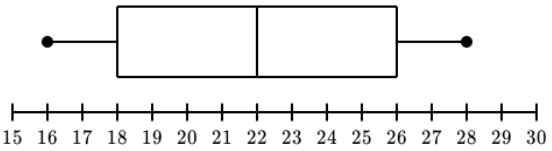
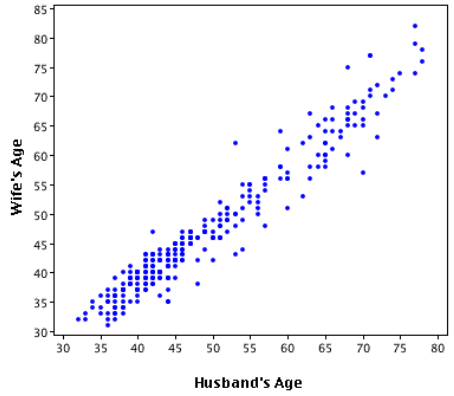
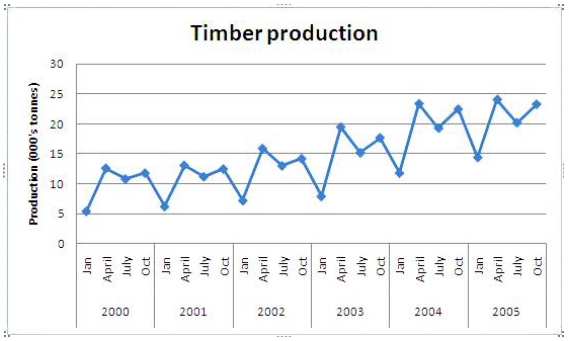
4) 402, 458, 461, 465, 466, 468, 468, 472, 473, 474, 475, 478, 479, 482, 491

5) 1.4, 1.6, 1.8, 1.9, 2.1, 2.2, 2.4, 2.6, 2.9, 3.1, 3.4, 3.5, 3.7, 4.1, 4.3

6) 1.35, 1.37, 1.39, 1.42, 1.45, 1.46, 1.46, 1.48, 1.51, 1.52, 1.57, 1.60

Graphs

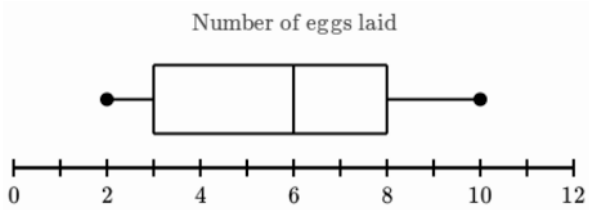
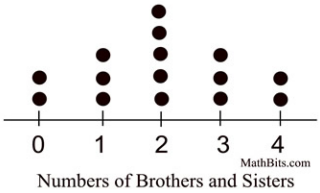
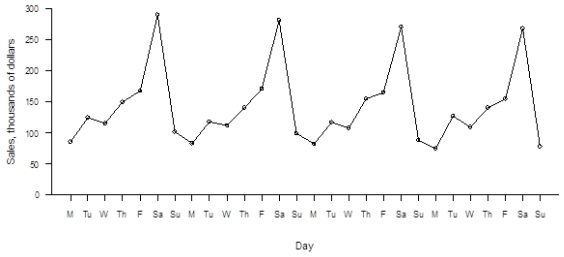
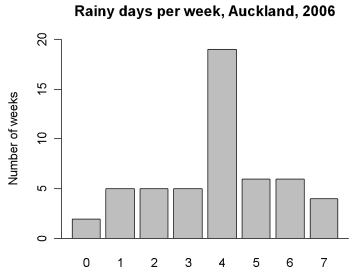
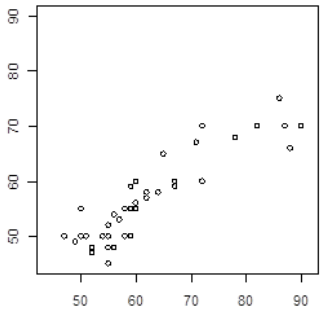
Here are some common ways to graph data.

Bar graph **(categorical or discrete data)**	**Histogram** **(continuous data)**
	
Dot plot **(numerical data)**	**Box plot (or Box and whisker)** **(numerical data)**
	
Scatter Graph **(two numerical variables)**	**Time series graph** **(one numerical variable and one variable about time)**
	

Let's look at these in more detail, so that you know how to draw them by hand and on the computer.

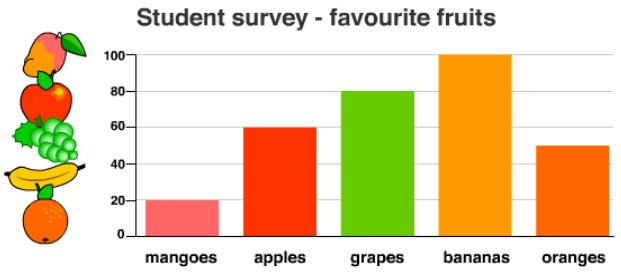
Exercise:

For each of the graphs below, state what type of graph it is.

	**Graph type:**
	**Graph type:**
	**Graph type:**
	**Graph type:**
	**Graph type:**
Actual weights of male university students (kg) 5 | 1577 6 | 0000002223557889 7 | 00012233455 8 | 00344589999 9 | 008 10 | 0009 11 | 12 | 0 The stem unit is 10kg	**Graph type:**

Bar graphs

You need your data in a frequency table, and you can draw bar graphs either from categorical data (groups) or discrete data (counting numbers).



When drawing bar graphs, you need to make sure that the bars **DO NOT** touch each other. This is because the data is not connected to each other.

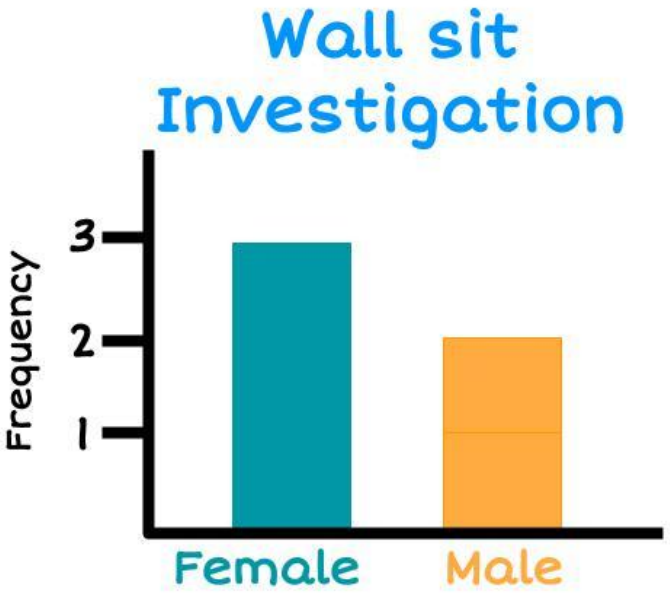
Example:

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

First make a Frequency table of the Gender variable, then make a Bar graph

Gender	Frequency
Female	3
Male	2



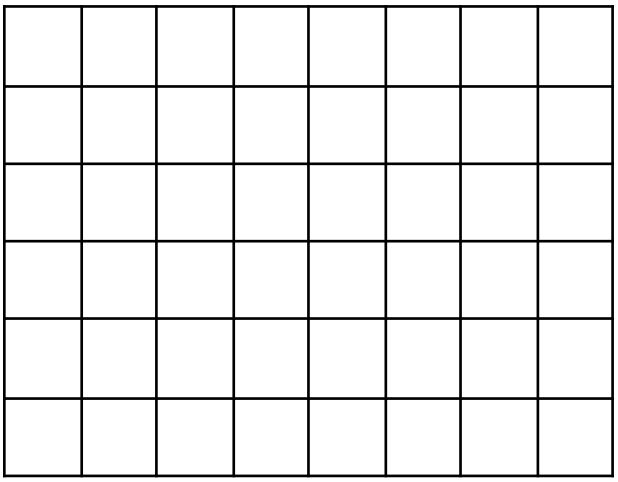
Exercise:

1) Using the Wall sit data above, make a bar graph of the PE variable.

Frequency table

PE	Frequency
Yes	
No	

Bar graph



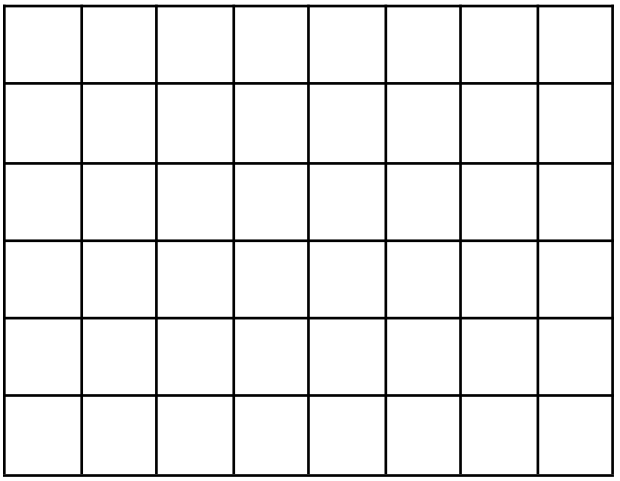
2) Using the Stickland data given below, make a bar graph for each categorical variable (Facebook, Snapchat, and Cellphone).

Age	Facebook	Snapchat	Bag weight	Cellphone	Reading time	TV time
6	no	yes	5	yes	0.25	4.25
13	no	no	5.8	yes	1	0.25
8	yes	yes	2.4	yes	0	0
14	yes	yes	1.1	yes	0	0
16	yes	yes	4.1	yes	0	2
6	yes	no	1	yes	3.25	3
11	yes	yes	3	yes	0	1
10	yes	no	3	yes	0.25	1
14	no	no	5	yes	0	1.5
10	no	no	3.7	no	0.25	3

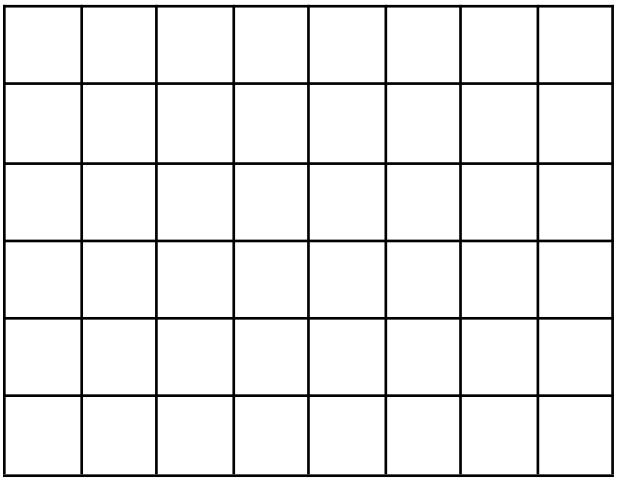
Frequency table

Facebook	Frequency

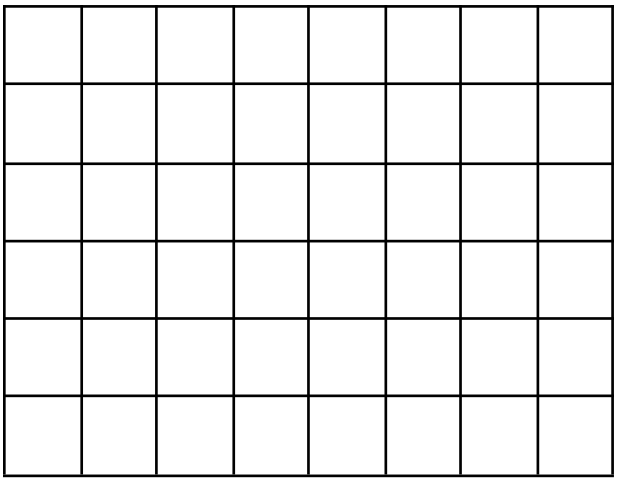
Bar graph



Snapchat	Frequency



Cellphone	Frequency

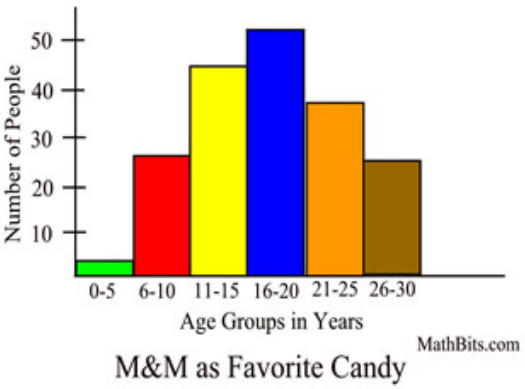


Histogram

You can draw histograms with continuous data (measurements).

You usually need to make a frequency table where you group the data into ranges first.

When drawing histograms, you need to make sure that the bars **DO** touch each other. This is because the data is continuous.



Example:

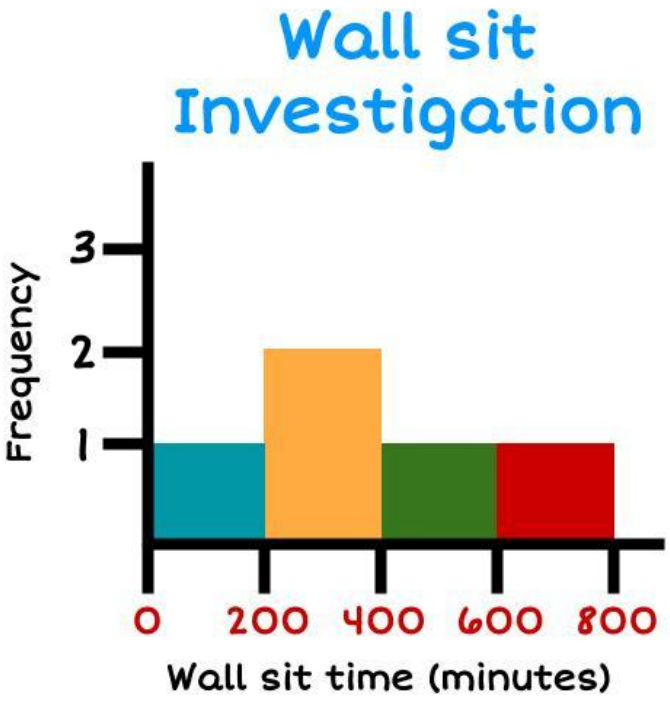
Here is a sample of students from the Wall sit spreadsheet:

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

Frequency table

Wall sit time (seconds)	Frequency
0 - 199	1
200 - 399	2
400 - 599	1
600 - 799	1

Histogram



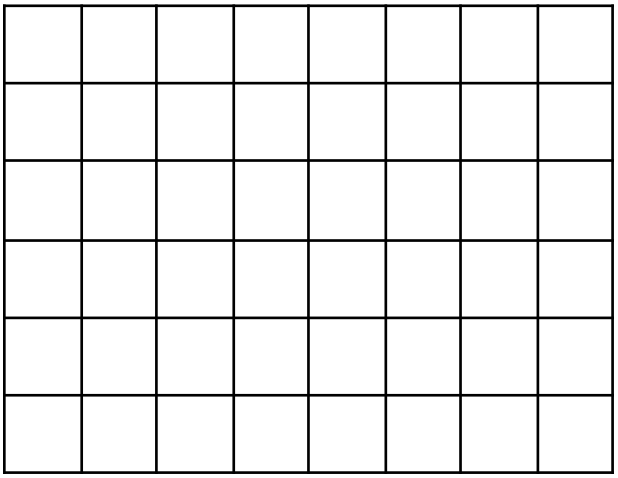
Exercise:

1) Using the Wall sit data above, make a histogram of the Height variable.

Frequency table

Height (cm)	Frequency
150 - 159	
160 - 169	
170 - 179	
180 - 189	

Histogram



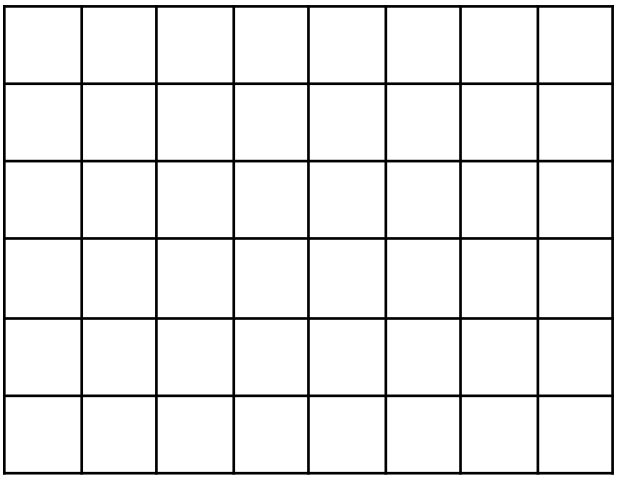
2) Using the Stickland data given below, make a histogram for each continuous variable (Age, Bag weight, Reading time and TV time).

Age	Facebook	Snapchat	Bag weight	Cellphone	Reading time	TV time
6	no	yes	5	yes	0.25	4.25
13	no	no	5.8	yes	1	0.25
8	yes	yes	2.4	yes	0	0
14	yes	yes	1.1	yes	0	0
16	yes	yes	4.1	yes	0	2
6	yes	no	1	yes	3.25	3
11	yes	yes	3	yes	0	1
10	yes	no	3	yes	0.25	1
14	no	no	5	yes	0	1.5
10	no	no	3.7	no	0.25	3

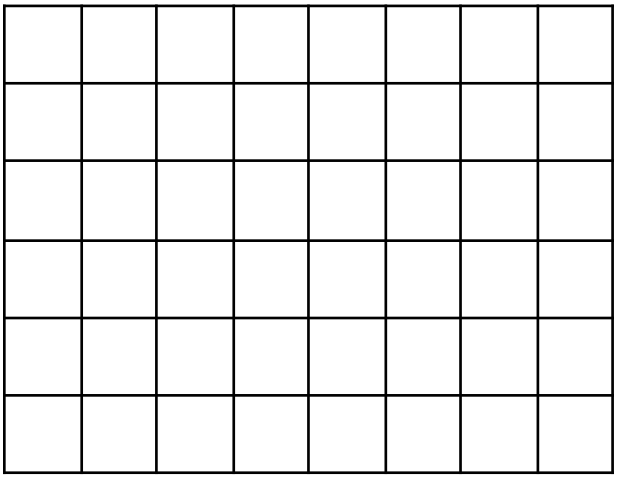
Frequency table

Age	Frequency

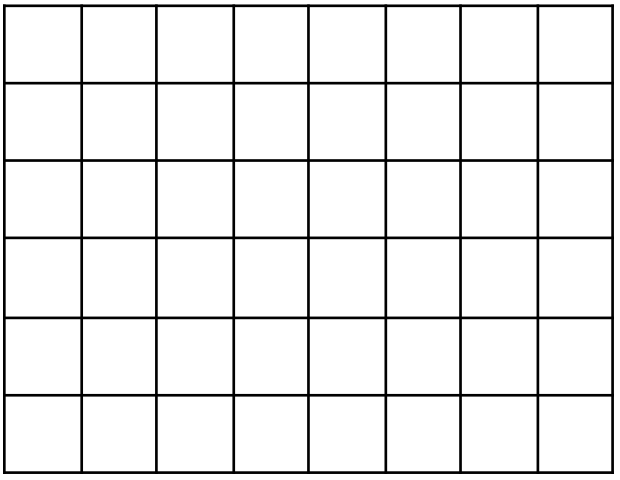
Bar graph



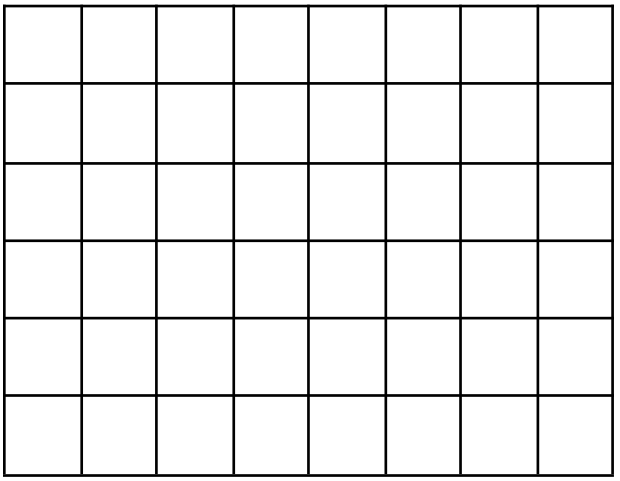
Bagweight	Frequency



Reading time	Frequency



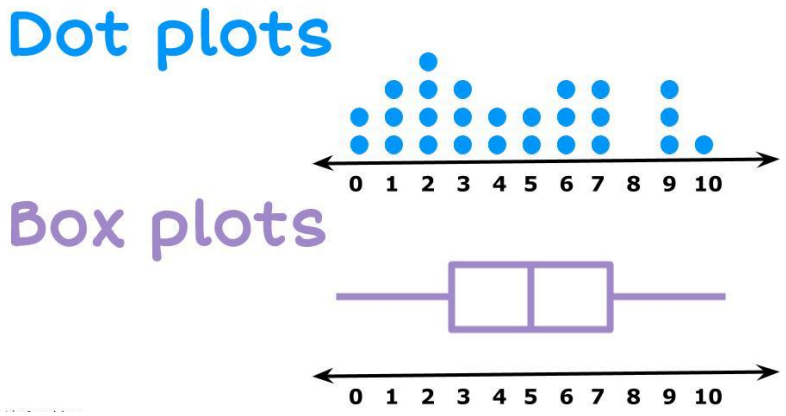
TV time	Frequency



Dot plot & Box plot (Box and whisker)

Both of these graphs are for Numeric data.

A dot plot plots every data point, and the box plot (sometimes called a box and whisker plot) is a summary of the data.



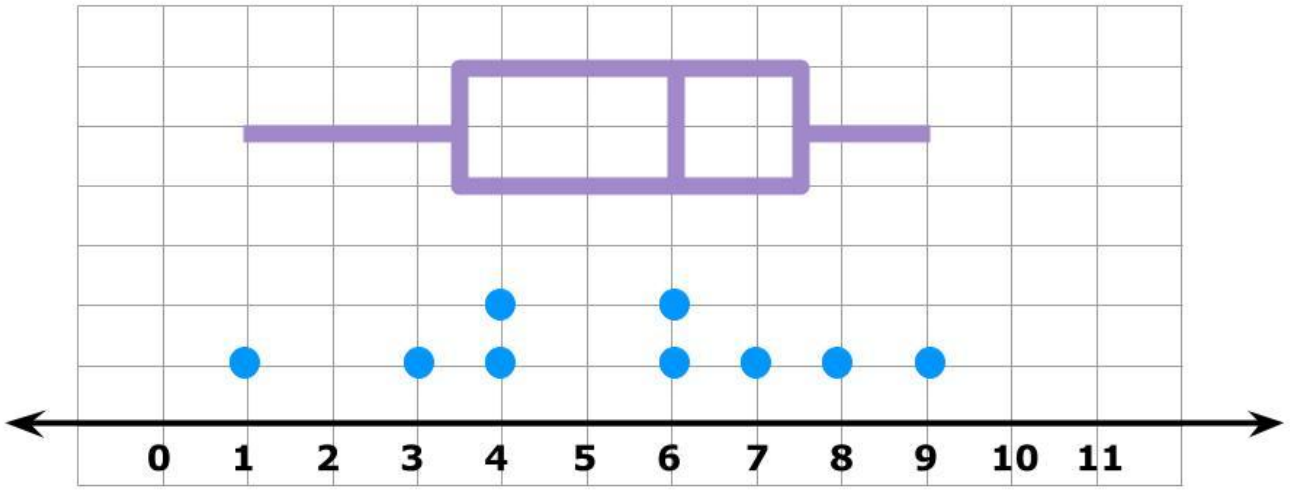
Later in this workbook we will show you how to find the summary statistics needed for the box plot. For now, we will give you these values and we want you to focus on how to draw the graph.

Example:

Draw a dot plot with the following data: 1, 3, 4, 6, 8, 9, 4, 6, 7

Then draw a box plot with the following data:

- Minimum = 1
- LQ = 3.5
- Median = 6
- UQ = 7.5
- Maximum = 9



It's very helpful when we want to analyse the data later to have the dot plot and box box plot stacked on top of each other like this example.

Exercise:

1) Using the Stickland data given below, make a dot plot and box plot for each numerical variable (Age, Bag weight, Reading time and TV time).

Age	Facebook	Snapchat	Bag weight	Cellphone	Reading time	TV time
6	no	yes	5	yes	0.25	4.25
13	no	no	5.8	yes	1	0.25
8	yes	yes	2.4	yes	0	0
14	yes	yes	1.1	yes	0	0
16	yes	yes	4.1	yes	0	2
6	yes	no	1	yes	3.25	3
11	yes	yes	3	yes	0	1
10	yes	no	3	yes	0.25	1
14	no	no	5	yes	0	1.5
10	no	no	3.7	no	0.25	3

2) Draw a dot plot of the **Age** variable. Then add a box plot using the summary data below.

- Minimum = 6
- LQ = 8
- Median = 10.5
- UQ = 14
- Maximum = 16

3) Draw a dot plot of the **Bag weight** variable. Then add a box plot using the summary data below.

- Minimum = 1
- LQ = 2.4
- Median = 3.95
- UQ = 5
- Maximum = 5.8

4) Draw a dot plot of the **Reading time** variable. Then add a box plot using the summary data below.

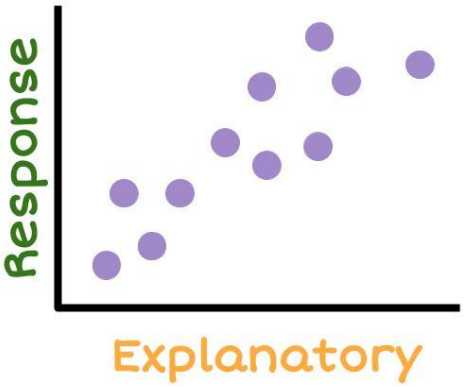
- Minimum = 0
- LQ = 0
- Median = 0.125
- UQ = 0.25
- Maximum = 3.25

5) Draw a dot plot of the **TV time** variable. Then add a box plot using the summary data below.

- Minimum = 0
- LQ = 0.25
- Median = 1.25
- UQ = 3
- Maximum = 4.25

Scatter graph

The scatter graph looks to see if there is a relationship between two numeric variables.

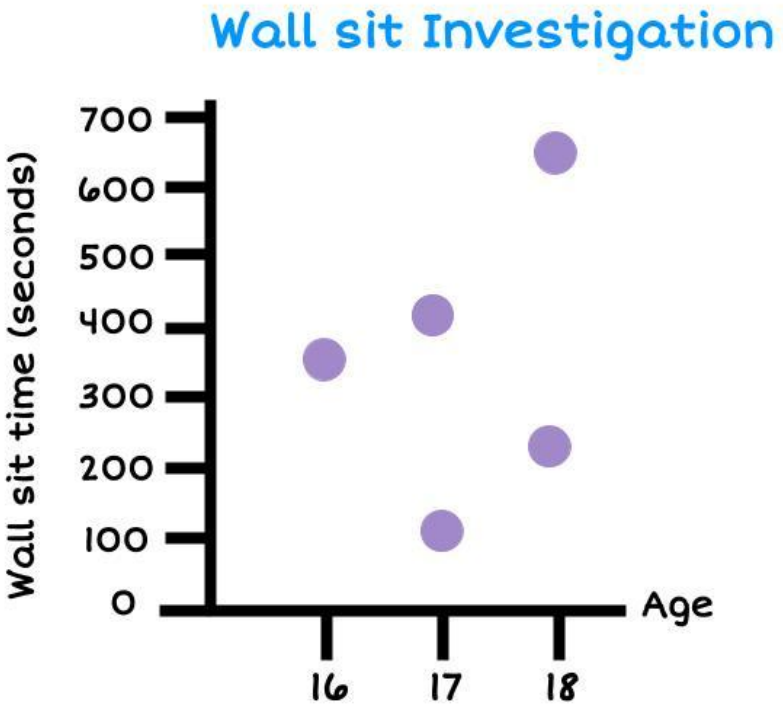


Example:

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

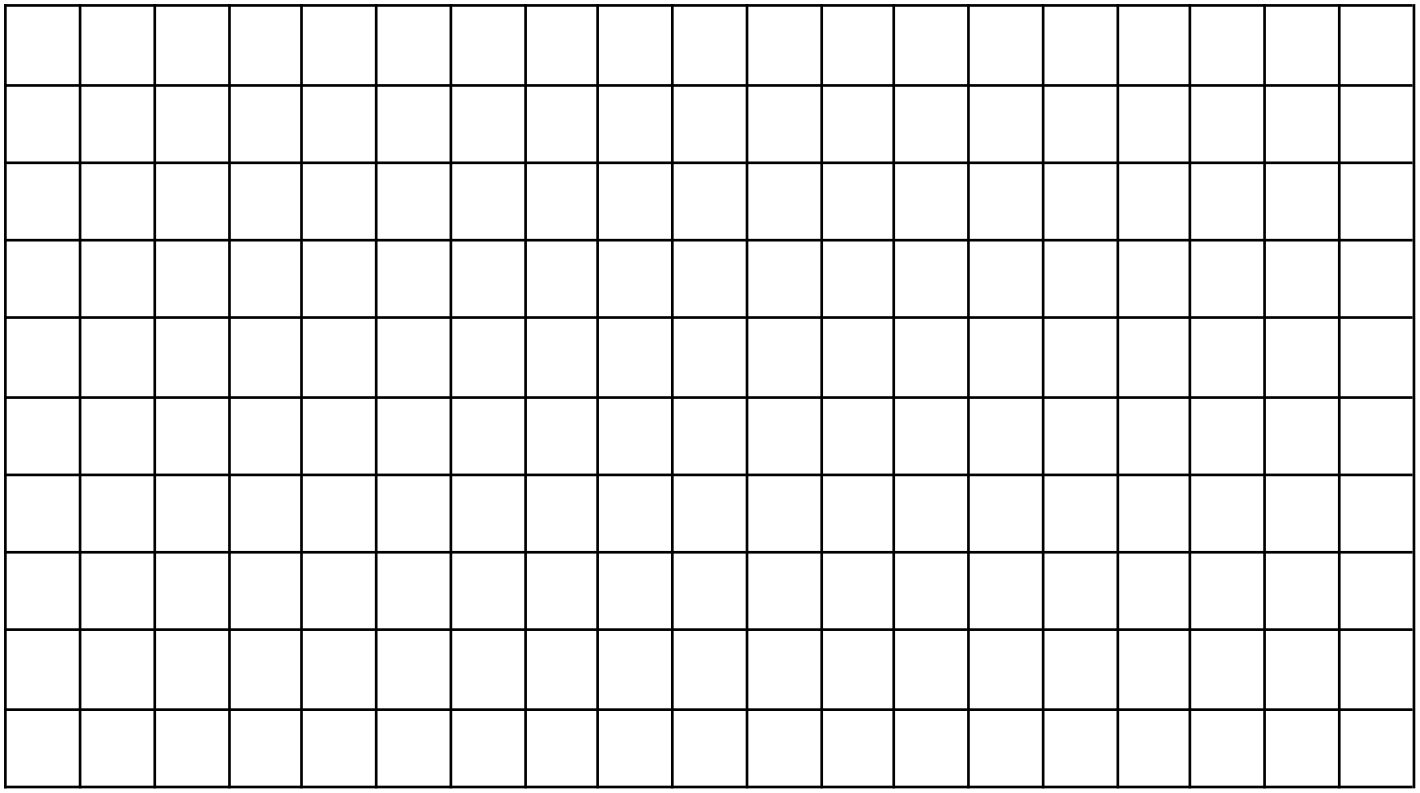
Students First Name	Age	Gender	Taking PE this year?	Wall sit time (seconds)	Height (cm)
Jessie	17	Female	No	114	161
Caleb	18	Male	Yes	640	185
Amisha	16	Female	No	352	155
Alena	18	Female	Yes	238	169
Luke	17	Male	Yes	421	182

Draw a scatter graph of the **Age** and **Wall sit** variables.

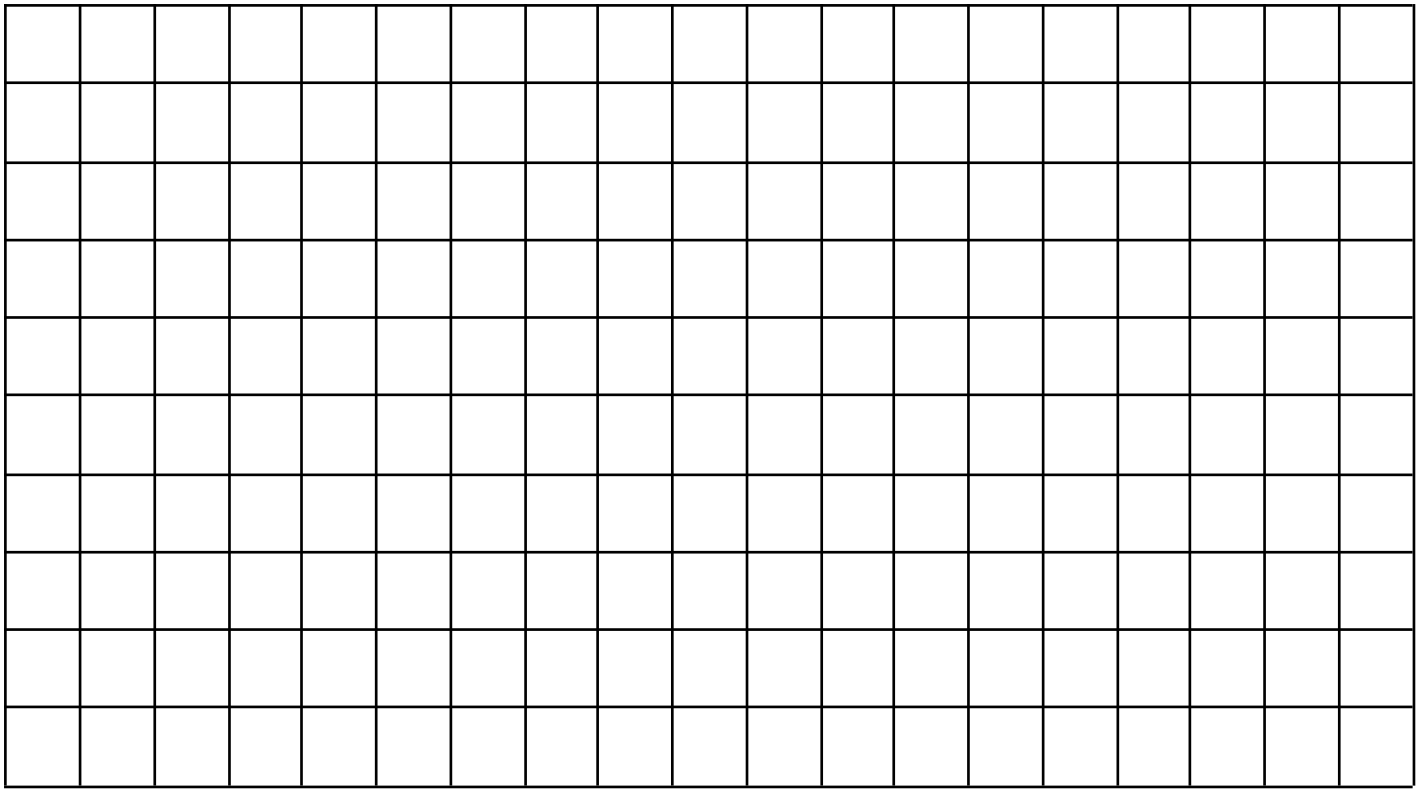


Exercise:

1) Using the Wall sit data above, make a scatter graph of the **age** and **height** variables.



2) Using the Wall sit data above, make a scatter graph of the **height** and **wall sit** variables.



2) Using the Stickland data given below, choose two numeric variables and make a scatter graph.

Age	Facebook	Snapchat	Bag weight	Cellphone	Reading time	TV time
6	no	yes	5	yes	0.25	4.25
13	no	no	5.8	yes	1	0.25
8	yes	yes	2.4	yes	0	0
14	yes	yes	1.1	yes	0	0
16	yes	yes	4.1	yes	0	2
6	yes	no	1	yes	3.25	3
11	yes	yes	3	yes	0	1
10	yes	no	3	yes	0.25	1
14	no	no	5	yes	0	1.5
10	no	no	3.7	no	0.25	3

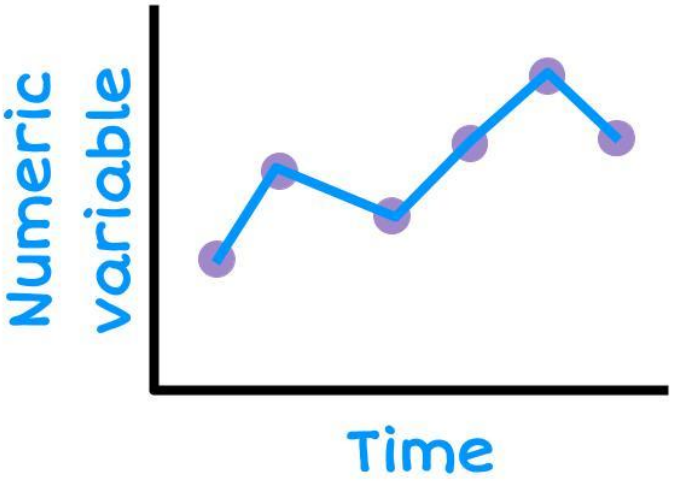
Time series graph

Time series graphs are about data that has been collected over time.

We put the time on the horizontal (x) axis.

Time can be measured in minutes, hours, days, weeks, months or years.

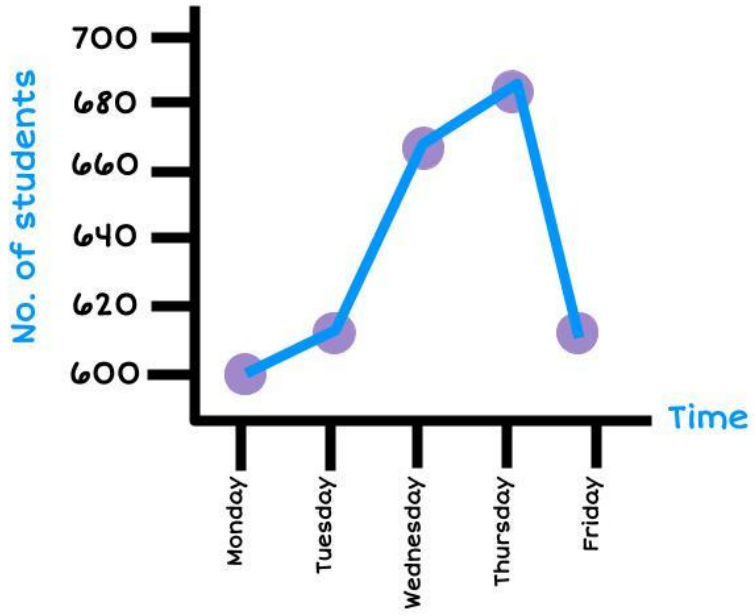
The vertical axis must be a numerical variable.



Example:

Here is data on attendance at a school over one week. Draw a time series graph.

Weekday	No. of students attending
Monday	600
Tuesday	610
Wednesday	672
Thursday	688
Friday	608



Exercise:

1) Using the data below about the number of cars sold at a car yard, make a time series graph.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Sales	68	60	64	64	58	54	68	58	60	68

2) Using the data below on the number of phones sold per week at a phone store, make a time series graph.

Week	1	2	3	4	5	6	7	8	9	10	11	12
Phone Sales	174	183	147	174	134	156	151	138	147	129	138	116

3) Using the data below on the number of text messages sent per day, make a time series graph.

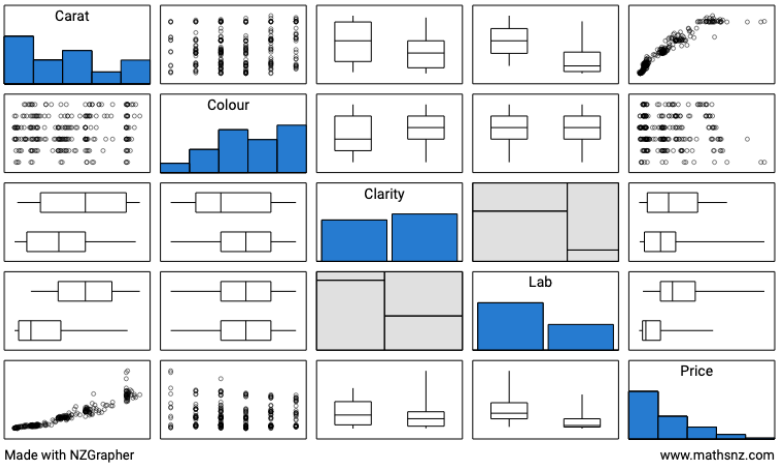
Day	1	2	3	4	5	6	7	8	9	10	11	12
Text messages	52	63	52	59	74	82	93	77	84	104	92	113

4) Using the data below on the profit (in thousands of dollars) for a company over the last 11 years, make a time series graph.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Profit ($000's)	50	36	43	44.5	39	37.5	33.5	38	42	41.5	32.5

NZGrapher https://grapher.nz/

Pairs plots are useful as it gives an overview of the dataset, the variables, and the comparative graphs. If you click on any of the graphs, it will take you to that graph.



Exercise:

For this exercise, you will use the **Kiwi** dataset in NZGrapher. Here are the variables.

Variable	Description	
Species	GS-Great Spotted NIBr-North Island Brown Tok-Southern Tokoeka	
Gender	M-Male F-Female	
Weight(kg)	The weight of the kiwi bird in kg	
Height(cm)	The height of the kiwi bird in cm	
Location	NWN-North West Nelson CW-Central Westland EC-Eastern Canterbury StI-Stewart Island NF-North Fiordland	SF-South Fiordland N-Northland E-East North Island W-West North Island

1) Go to NZGrapher and select the **Kiwi** dataset.

2) Look at the data on the left hand side. Find point number 20 and 40, and write their data values in the table below.

Data point	Species	Gender	Weight	Height	Location
20					
40					

3) Make 2 bar graphs, one with the variable **Species**, and one with the variable **Location.** Add to your graph a title, and summary statistics.

Copy the graphs (move the mouse over the image and right click, select copy) and paste them both into a Word document.

4) Make 2 histograms, one with the variable **Weight**, and one with the variable **Height.** Add to your graph a title, units onto the axis label and summary statistics.

Copy and paste the graphs into your Word document.

5) Make a pie chart and a donut graph with the variable **Gender.** Add to your graph a title, and summary statistics.

Copy and paste the graphs into your Word document.

6) Make a dot and box plot with the variable **Weight.** Add to your graph a title, units on the horizontal axis, a High box plot and summary statistics.

Copy and paste the graphs into your Word document.

Repeat this with the variable **Height**.

7) Make a scatter graph with the variables **Height** and **Weight.** Add to your graph a title, and a label (with units) on both the horizontal and vertical axis).

Copy and paste the graphs into your Word document.

Then add a regression line and copy this into the box below also.

8) Select the dataset **TS - Sunglasses.csv**.

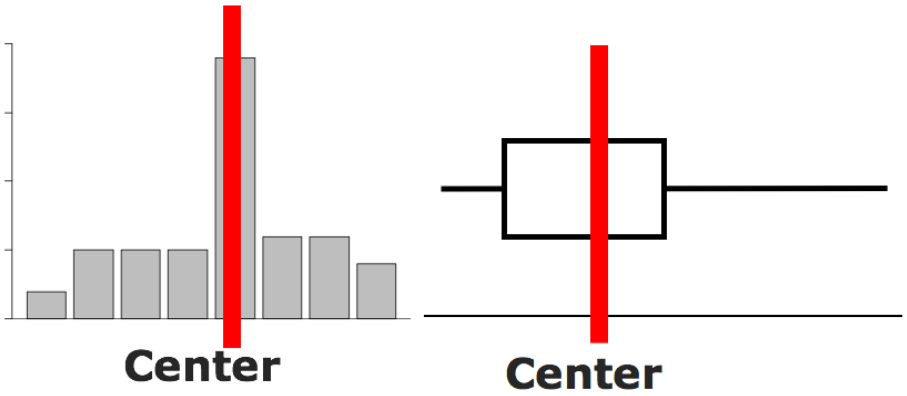
Create a Time Series graph of the variables **Quarter** and **Sales** and add a title

Copy and paste the graphs into your Word document.

Summary Statistics

Numbers calculated from a *sample* of numerical values that are used to summarise the sample. The statistics will usually include at least one ***measure of center*** and at least one ***measure of spread***.

Measures of Center



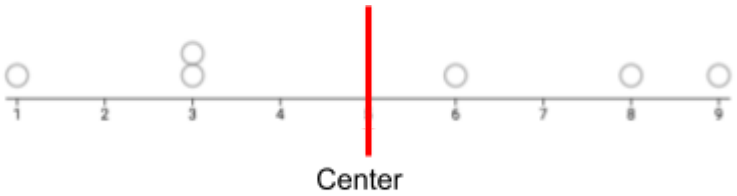
There are 3 measures of center:

- Mean = $\frac{add\ up\ all\ the\ values}{the\ number\ of\ data\ values}$
- Median = the number in the middle (when the data is in order)
- Mode = the most common number

Example:

Estimate the center, and find the mean, median and mode.

Data: 9, 3, 1, 8, 3, 6



Mean = $\frac{9+3+1+8+3+6}{6}$ = 5

Median

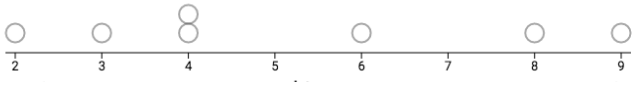
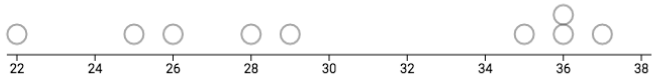
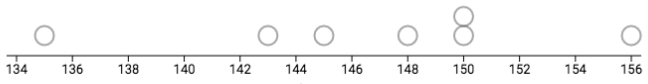
Put the numbers in order: 1, 3, 3, 6, 8, 9

Find the number(s) in the middle: 1, 3, 3, 6, 8, 9

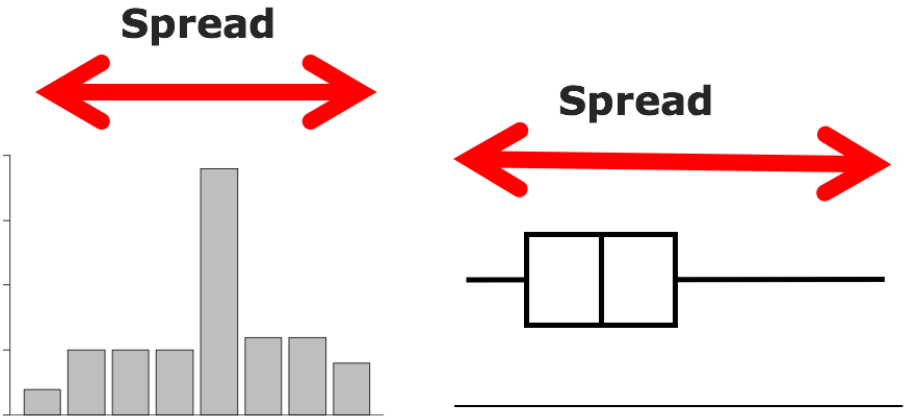
Find the median = $\frac{3+6}{2}$ = 4.5

Mode = 3

Exercises:

Estimate center on the graph.	Calculate the Mean, Median, and Mode
Data: 4, 6, 3, 8, 2, 4, 9 	
Data: 4.4 4.7 3.5 2.2 4.2 6.7 2.9 4.4 1.5 2.0 3.3 	
Data: 25, 35, 37, 36, 28, 29, 36, 26, 22 	
Data: $150, $145, $135, $150, $148, $156, $143 	

Measures of Spread



A measure of spread looks at how precise or accurate the data is. There are two measures you will use:

- Range = Maximum - Minimum
- IQR (InterQuartile Range) = UQ - LQ

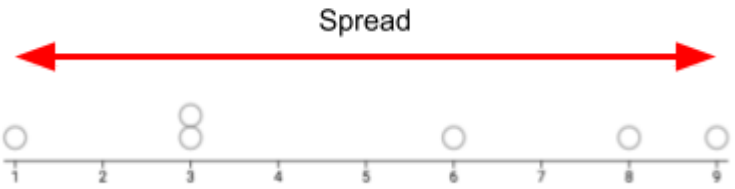
where UQ = Upper Quartile = the number where one quarter of the data lies **above** it (find the median, then find the middle of the numbers **above** the median, this is the UQ),

and LQ = Lower Quartile = the number where one quarter of the data lies **below** it (find the median, then find the middle of the numbers **below** the median, this is the LQ).

Example:

Show the spread on the graph, and find the range and IQR.

Data: 9, 3, 1, 8, 3, 6



Range = 9 - 1 = 8

IQR

Put the data in order: 1, 3, 3, 6, 8, 9

Find where the median is: 1, 3, 3 | 6, 8, 9

Find the LQ (the median of numbers below the median), the median of 1, 3, 3

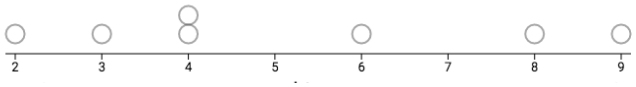
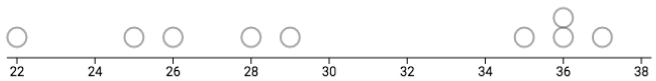
LQ = 3

Find the UQ (the median of numbers above the median), the median of 6, 8, 9

UQ = 8

IQR = UQ - LQ = 8 - 3 = 5

Exercises:

Show the spread on the graph.	Calculate the Range and Interquartile Range
Data: 4, 6, 3, 8, 2, 4, 9 	
Data: 4.4 4.7 3.5 2.2 4.2 6.7 2.9 4.4 1.5 2.0 3.3 	
Data: 25, 35, 37, 36, 28, 29, 36, 26, 22 	
Data: $150, $145, $135, $150, $148, $156, $143