Level 1
Data Collection & Exploration

Name:

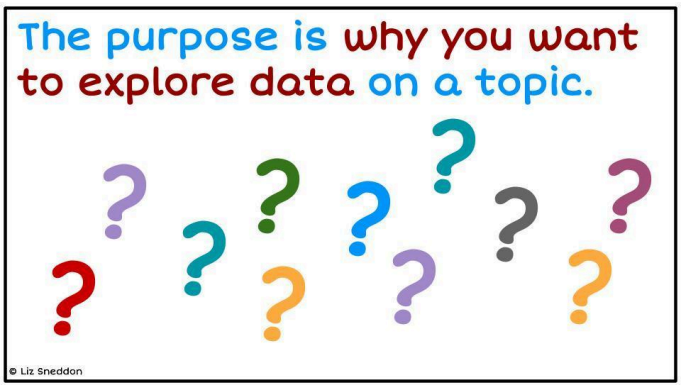
By Liz Sneddon

Introduction

Purpose

As people go through their life, their jobs, their interests, there will be times when questions arise and in order to have an informed understanding, data is often collected and explored, statistical investigations are carried out, analysis is done, and conclusions are formed which may include inferences or predictions.

At NCEA Level 1, ākonga will explore a topic, forming an individual purpose and then collecting (or sourcing) a range of data and variables which they will then explore and carry out an investigation.



Example:

Kaiako were having a discussion on teaching statistics to Level 1 ākonga. At the end of this discussion, Mrs Sneddon was interested in finding out more information about what examples, exercises and activities kaiako use for teaching Time series as she wanted to get new ideas to use in her classroom.

Exercise 1:

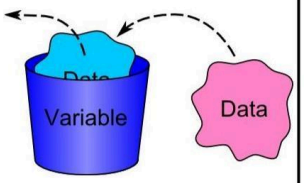
For each of the topics below, identify an area within the topic that you might be interested in and explain **why** you may be interested in it.

	Topic	Purpose identified
1)	Co-curricular activities that ākonga at your school enjoy and are involved in.	
2)	Ākonga, their cellphones and apps used on their phones.	

Variables & Data

Variable

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



Data

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

Organising data

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

Example:

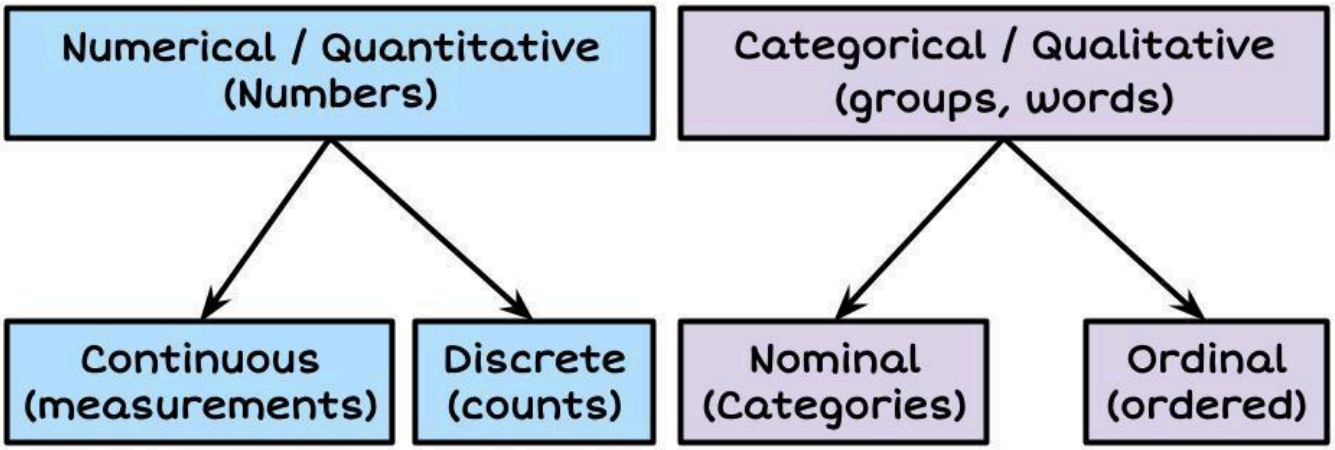
Here is a spreadsheet:

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

Each **row** is a set of **data** belonging to one ākonga.

Each column is a **variable**.

Data types



Numerical data

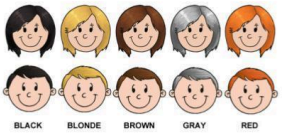
Numerical (numerical) data is data described by numbers.

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

Categorical data

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

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



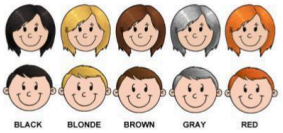
Continuous data

Continuous data is data obtained by measuring.

Nominal data

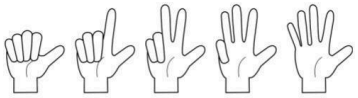
Nominal data is categorical data (groups/words) where the categories have NO order.

E.g. Black, Blonde, Brown, Grey, Red, etc



Discrete data

Discrete data is data obtained by counting.



Ordinal data

Orninal data is categorical data (groups/words) where the categories HAVE order.

E.g. 1st, 2nd, 3rd

Example:

Identify the data type for each variable in the spreadsheet below.

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

Variable	Data Type
Age	Numeric (Discrete)
Sex	Categorical
Takes PE?	Categorical
Wall sit time	Numeric (Continuous)
Leg up wall sit time	Numeric (Continuous)
Height	Numeric (Continuous)

Exercise 2:

For each of the datasets below, identify the data type for each variable.

1) The Bungee spreadsheet:

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

Variable	Elastic type	Length	Number of marbles
Data Type	Numeric Discrete Continuous Categorical Nominal Ordinal	Numeric Discrete Continuous Categorical Nominal Ordinal	Numeric Discrete Continuous Categorical Nominal Ordinal

Variable	Stretched length	Weight	Obstacles
Data Type	Numeric Discrete Continuous Categorical Nominal Ordinal	Numeric Discrete Continuous Categorical Nominal Ordinal	Numeric Discrete Continuous Categorical Nominal Ordinal

2) The Marathon dataset:

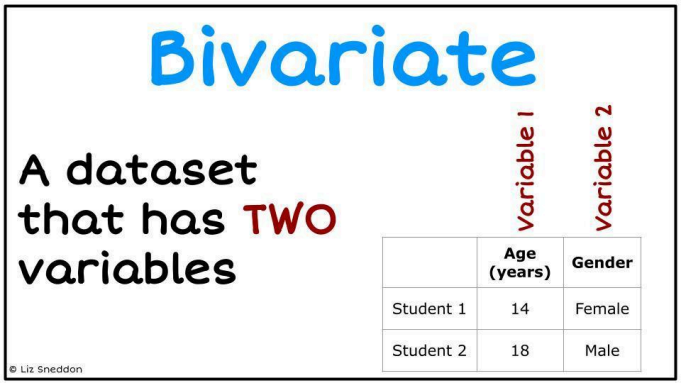
Variable	Description
Minutes	How many minutes they completed the marathon in
Gender	Male (M) or Female (F)
AgeGroup	Younger (under 40) or older (over 40)
Stridelength	The persons average stride length over the marathon in cm.

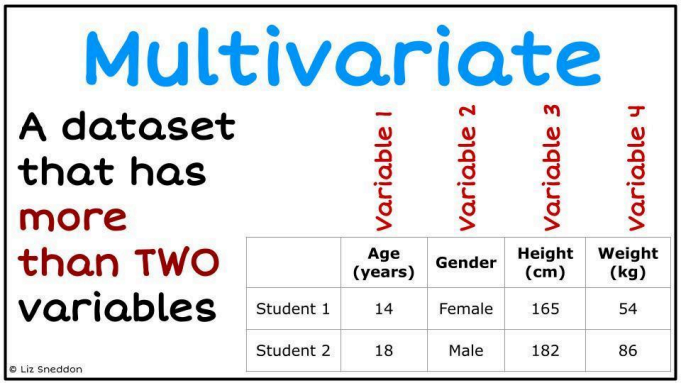
Variable	Minutes	Gender
Data Type	Numeric Discrete Continuous Categorical Nominal Ordinal	Numeric Discrete Continuous Categorical Nominal Ordinal

Variable	AgeGroup	Stridelength
Data Type	Numeric Discrete Continuous Categorical Nominal Ordinal	Numeric Discrete Continuous Categorical Nominal Ordinal

Datasets

A data set usually contains multiple variables. (Be aware that the previous Achievement Standards used the words bivariate and multivariate incorrectly).





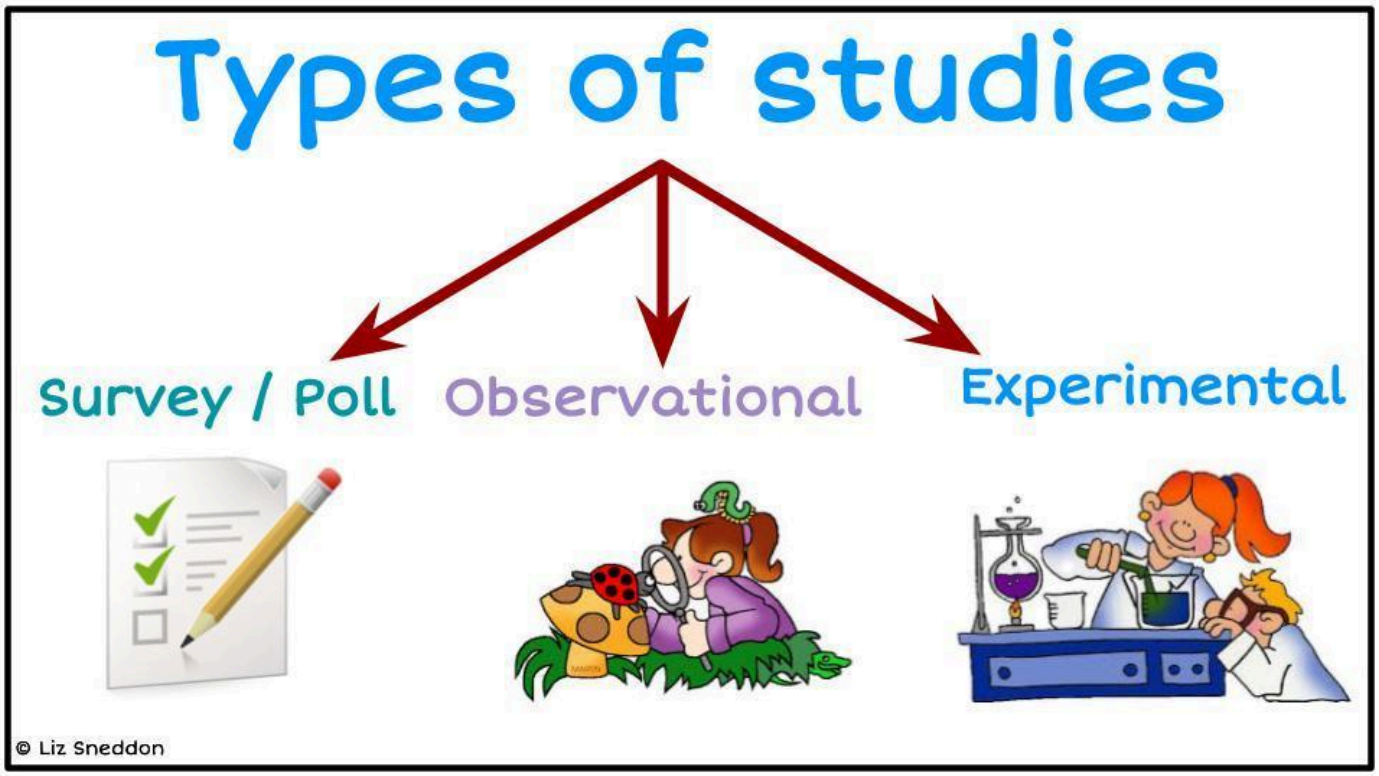
Example:

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

The dataset is **multivariate** as there are **more than 2 variables**.

Study types

When we collect data, there are three types of studies we can use – Survey / Polls, Observational or Experimental studies.



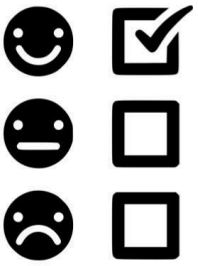
Survey

The process of collecting data from people in the population using a questionnaire.

© Liz Sneddon

Poll

Asking people in the population **ONE** question where they give their **opinion**.



© Liz Sneddon

Observational study

People or objects are observed without altering or controlling their behaviour in any way.

© Liz Sneddon

Experimental study

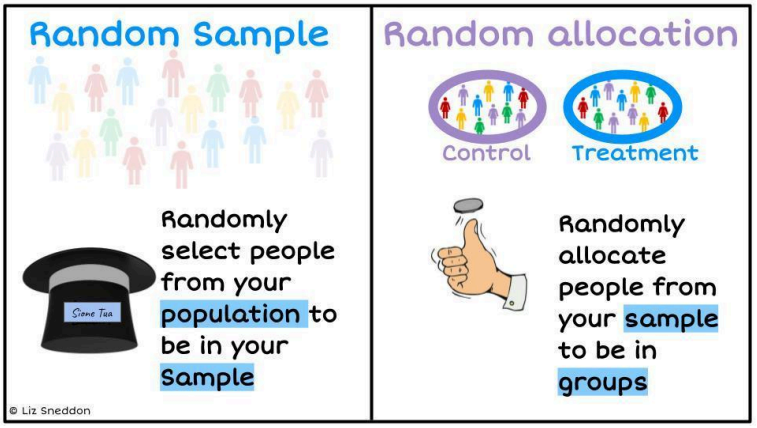
The investigator applies a treatment to participants.

© Liz Sneddon

The key difference between surveys / polls / observational and experimental studies is how people or objects are **selected** for the study. If people are **randomly allocated into groups**, then this is a key **experimental** design feature. However, if people are (preferably **randomly**) **selected from the population**, then this is an **observational** study / **survey** / **poll**. Simulations are also a type of experiment

(more on this later).

Let's have a look at the difference between **random allocation** and **random samples**.



Exercise 3:

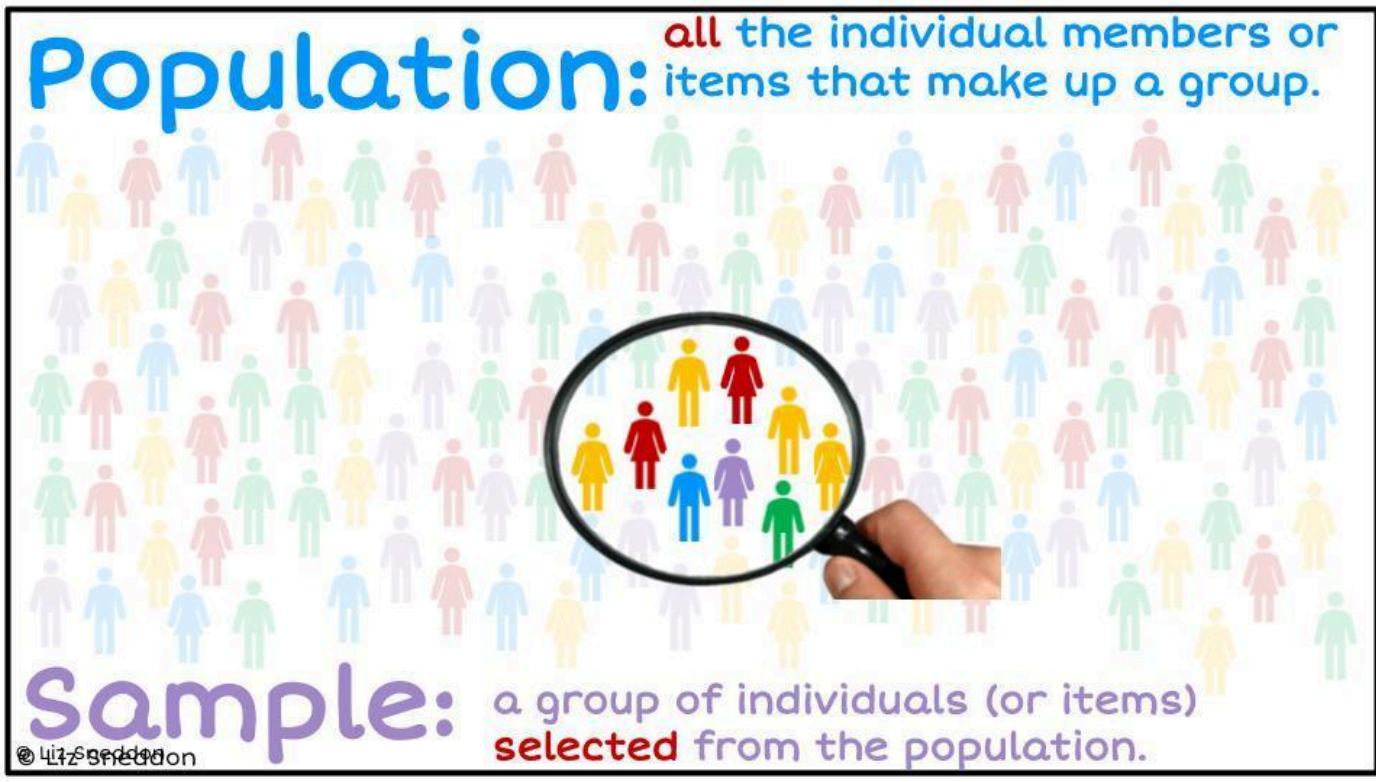
Identify which of the following are surveys, polls, observational study, or experimental study, then justify your selection.

1)	Ākonga in the class were asked to answer the following question: Rate your opinion of cartoon movies: 	Survey Poll Observational study Experimental study	Because
2)	A school gave a questionnaire to all ākonga in the school asking about which celebrations and religious days they celebrate each year, in order to better understand the school community.	Survey Poll Observational study Experimental study	Because
3)	A class of ākonga were randomly allocated either to sit an exam on paper or sit the exam on the computer to compare their results.	Survey Poll Observational study Experimental study	Because
4)	Ākonga in a class were all asked to throw the basketball at the hoop 3 times in order find the percentage of throws that	Survey Poll Observational study Experimental study	Because

went into the hoop.		

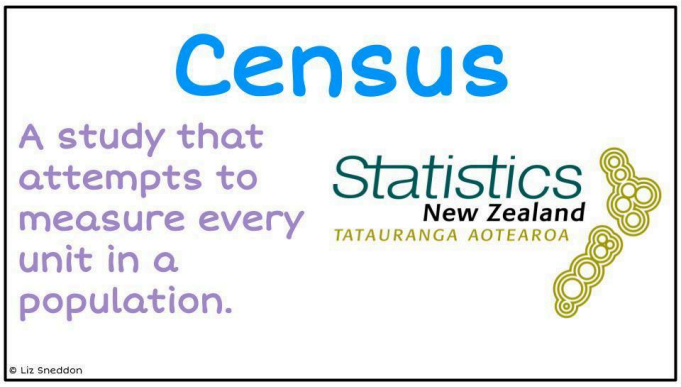
Populations and samples

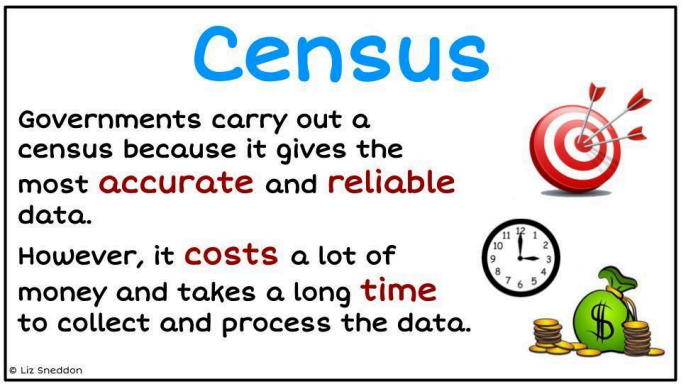
When we plan to collect data, we need to identify the population we want to collect data from, and how a representative sample (if possible) can be selected.



Example:

An example of a population is the census, where data is collected from the whole population. However this only occurs in NZ every 5 years and takes a long time to collect and process the data, along with costing millions of dollars.





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?

Exercise 4:

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

   A. A selection of logo artists.

   B. Every person in the United States.

   C. A selection of shoppers from different states.

   D. 3,800 children aged 5 - 15

2) 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.

3) 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.

4) 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.

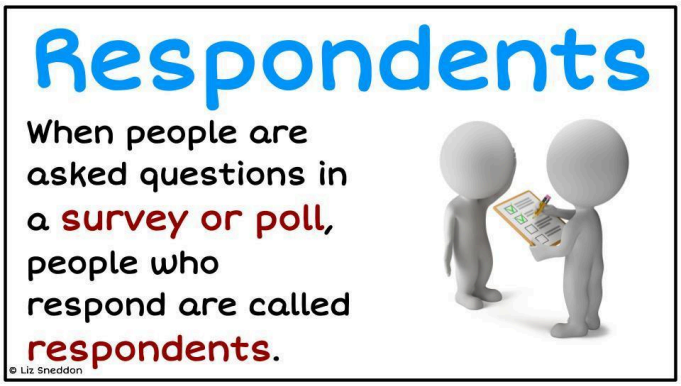
   B. A selection of male voters.

   C. A selection of voters of different ages.

   D. All voters

Respondents or Participants

When collecting data, one of the first decisions to make is who or what we are collecting data from (**our sample**). For example, are we collecting information from photographs, websites, people, dice, or other objects.



Example:

A sample of kiwi birds in Aotearoa was selected and data such as heights, weights, species, sex, and location was gathered.

This is an **observational study** (as information about each kiwi bird was simply collected), and the **participants** are kiwi birds.

Exercise 5:

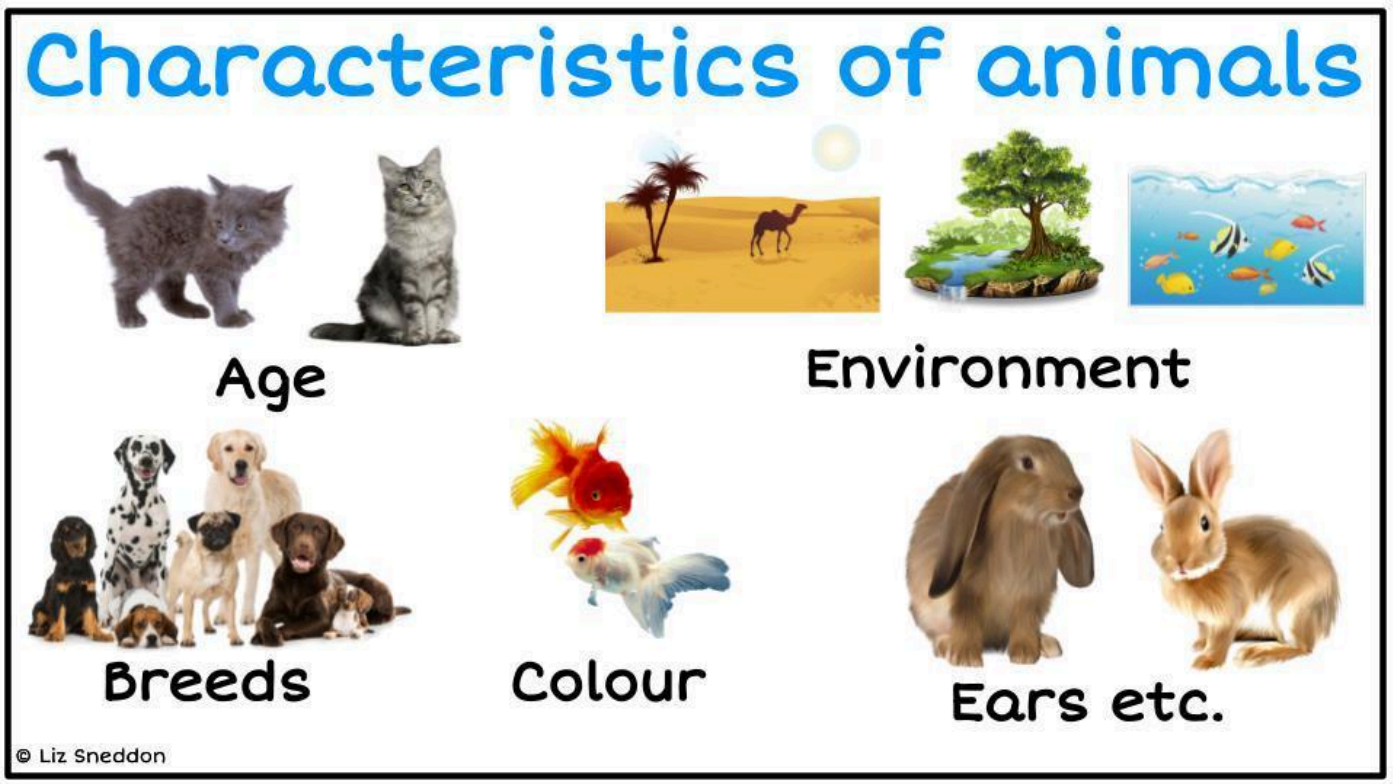
Identify who the respondents or participants are for the following studies:

	Question	Study type	The sample
1)	Ākonga in the class were asked a series of questions about their movie preferences, including genre types (drama, horror, etc), lead actors/actresses, directors, and more.	Survey Poll Observational study Experimental study	Respondent Participant **Identify:** _____ _____
2)	A science teacher (kaiako) asked ākonga to collect samples of water from their local stream, in order to test the water quality (including pH levels, acidity, oxygen levels, etc.)	Survey Poll Observational study Experimental study	Respondent Participant **Identify:** _____ _____

Representative data

There are several reasons that we collect data. It may be to explore or to carry out an investigation or experiment. One of the things we need to consider is how people or objects are selected or chosen to participate or respond. The key concept behind these sampling or selection methods is about whether the data is representative of the population or not.

Think about the population of animals in the world, and their different characteristics. Animals can be different sizes, shapes, ages, sex, height, weight, colour, environments, food preferences, and many more. A sample that has a fair mixture of characteristics from a population would be representative.

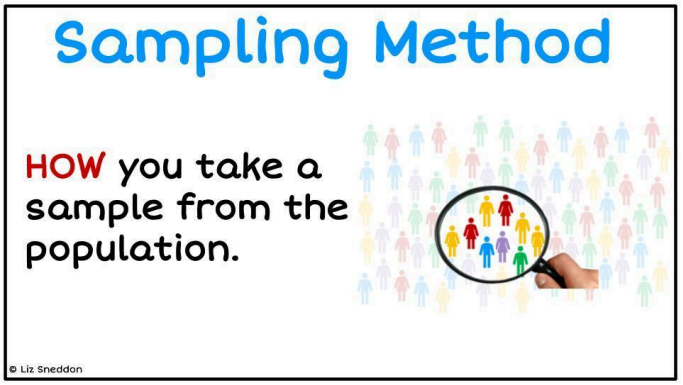


Exercise 6:

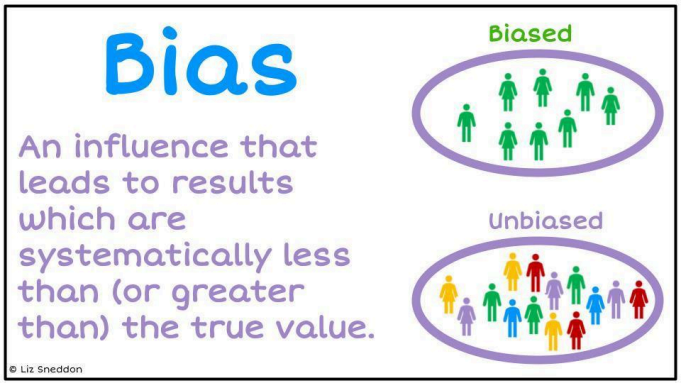
Describe the different characteristics of ākonga in your maths class.

Sampling method

Due to issues such as time, money and availability, a sample of data is usually collected. We now want to think about how we choose to select people/objects for our study. That is the sampling method.



When we carry out surveys, polls or observational studies, the sample can be selected using a random or biased method.

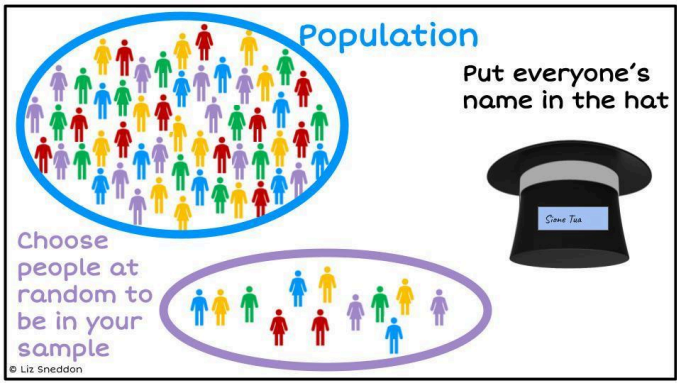


Random sampling methods

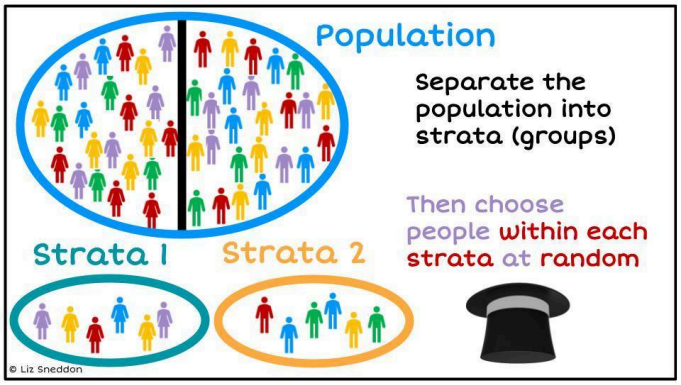
Here are two commonly used random sampling methods:

Simple random sample

Everyone in the population has an equal chance of being selected, and people or objects are selected at random.



Stratified random sample
Separate the population into groups (e.g. age, ethnicity, gender, sex, etc.) and then take a simple rand sample from **each group**.



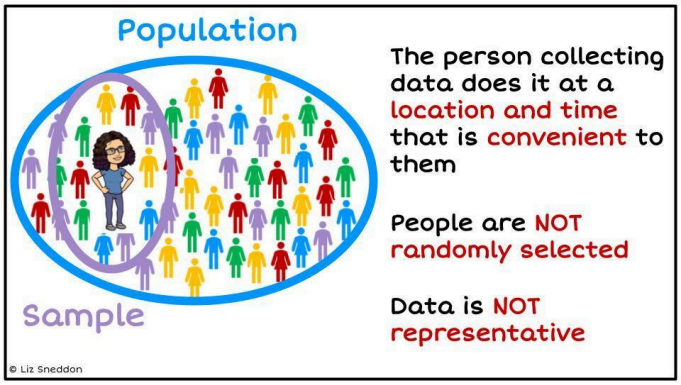
Biased sampling methods

Here are two common biased sampling methods:

Convenience sample

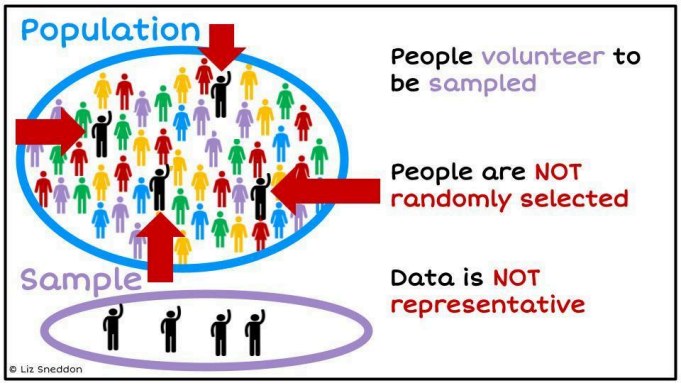
People or objects are selected by what is convenient to the person collecting the data.

For example, kaiako selects all the ākonga who are in class in a specific lesson to collect data.



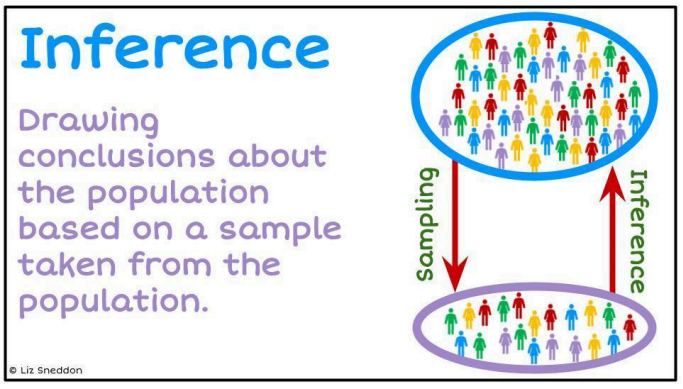
Volunteer or Self-selected sample

This is a common method where people volunteer to participate. This is commonly used in experiments, and may be used in surveys, polls, or observational studies.



Conclusions

If we collect data **randomly**, the data is likely to be **representative** of the population, which means that we can make **inferences** about the population.



If the sample is **not randomly selected**, then we **CANNOT** make an **inference** about the population.

Exercise 7:

Below are descriptions of situations where sampling methods were used.

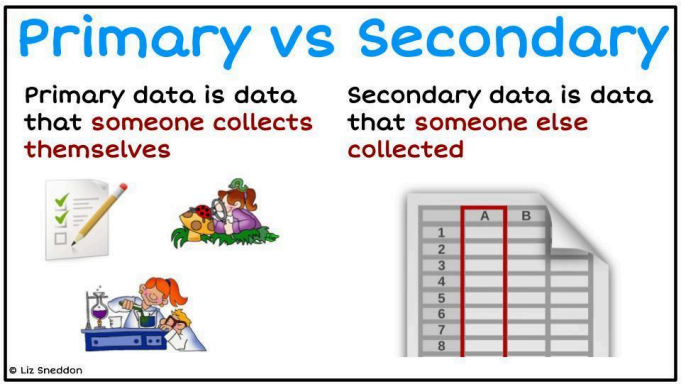
Identify if this is a random or biased method, then identify which sampling method has been described.

	Situation	Method	Sampling method
1)	At assembly, the principal asks ākonga who want to take part in the blood drive later this year to stand up.	Random Biased	Simple random Stratified Convenience Self-selected
2)	Mrs Sneddon separates her class into boys and girls, and randomly choose 3 girls and 3 boys to participate in an experiment.	Random Biased	Simple random Stratified Convenience Self-selected
3)	Mrs Sneddon randomly chooses one ākonga from her class to collect messages from the office.	Random Biased	Simple random Stratified Convenience Self-selected
4)	Mrs Sneddon needs help setting up classrooms for an assessment and asks five ākonga seated nearby to help.	Random Biased	Simple random Stratified Convenience Self-selected

Data collection methods

When a person **collects the data themselves** (could be through a survey, poll, observational study, or experimental study), then this is called **primary** data.

When data that has been **collected by someone else** (for example, a spreadsheet containing data, etc.) this is called secondary data.



Example:

On the World Health Organization's website data is the number of Covid cases in different countries are displayed, as can be seen from a sample of the table.

This is **secondary data**, as countries and the WHO have collected data themselves, and I would simply be using data already collected.

Name	Cases - cumulative total	Cases - newly reported in last 7 days	Deaths - cumulative total	Deaths - newly reported in last 7 days	Vaccines - Total doses administered per 100 population	Vaccines - Persons vaccinated with a complete primary series per 100 population	Vaccines - Persons vaccinated with at least one booster or additional dose 100 population
Global	772,052,752	1,163	6,985,278	23	174.42	66.25	31.9
New Zealand	2,416,777		3,522		254.08	84.69	56.39
Hungary	2,212,994		48,848		169.6	63.49	39.93
Bangladesh	2,046,060	12	29,477		219.95	86.34	41.61
Slovakia	1,868,801		21,167		132.3	51.37	31.03

Exercise 8:

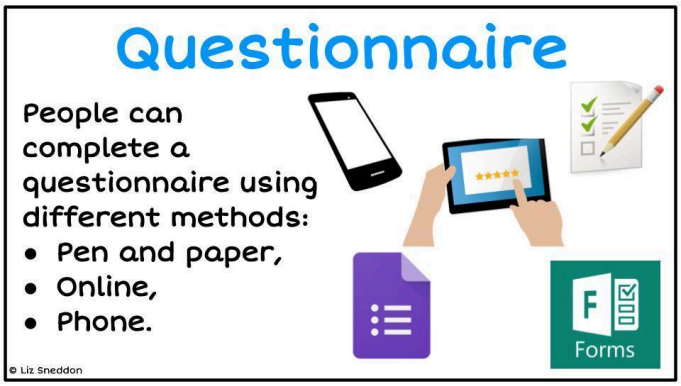
Identify and explain whether the data described below is primary or secondary data.

1)	Measuring the heights and ages of ākonga in your class.	Primary Secondary	
2)	Data on qualifications of New Zealanders was download from Statistics NZ.	Primary Secondary	
3)	Collecting data from a website which has information about movies	Primary Secondary	

Collecting primary data

When we want to collect primary data, there are several design factors to consider. We want to think about the type of study, who we collect data from (the respondents or participants), how we select them (sampling or volunteers), and the process we use to collect data (instructions, ethics, etc.).

Primary data may be collected by a variety of methods, such as:

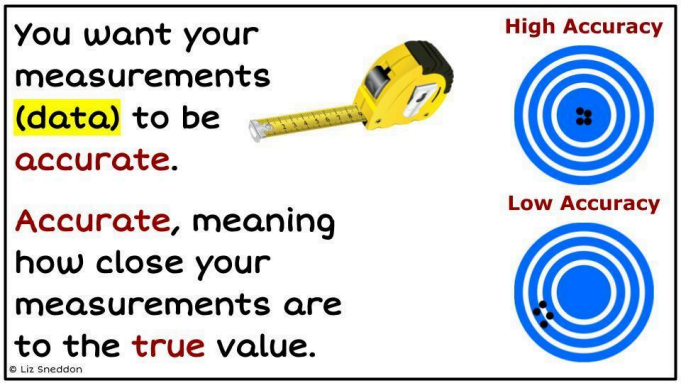


If you are collecting primary data for your assessment, is it more likely that you will collect data using a questionnaire (such as Google Forms or Microsoft Forms), or by observation.

Accuracy & Consistency

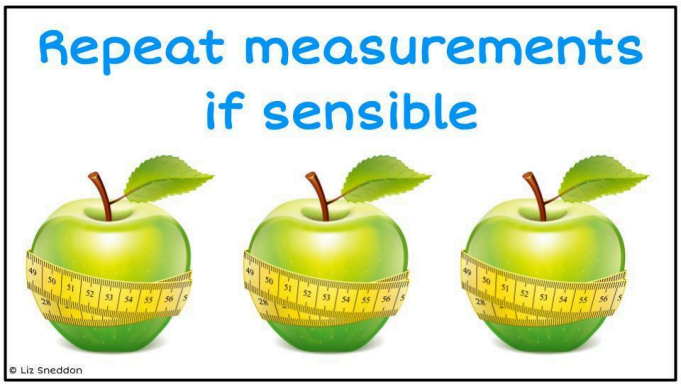
In order to collect data that is accurate we need to think about how detailed the instructions that we write are, what equipment to use, the sources of variation that affect the data and if/how these may be able to be controlled.

For data to be consistent, we need to make sure that the same method is used when collecting each piece of data.



To collect consistent data, here are some important things to remember:

- Anyone collecting data should use the same steps and instructions.
- Clearly identify and apply categories (e.g. happiness scale: depressed, sad, neutral, happy, exuberant).



Exercise 9:

1) What equipment would you suggest using to measure the following?

Time to run 100 meters	
Weight of an apple	
Capacity of liquid in a bottle	
Volume of music on the radio	
Angle of a wheelchair ramp	

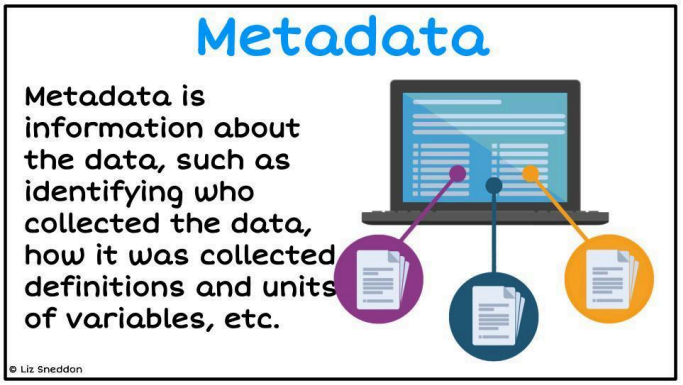
2) Explain how using different scales to measure the weight of pens would affect the data that is collected.

3) When measuring the length of a pen, explain why the instructions should include making sure that all pens have their lids on (if they have a lid).

Metadata for secondary data

If we are using secondary data (sourcing it from a website etc), then we need to look at the **metadata** (or **data dictionary**) to identify and define the variables, population, participants, data collection method, etc.

Data is raw information, whereas **metadata is the context** of that information[1].



Example 1:

One of the datasets on NZGrapher is the Kiwi dataset. Here is the metadata about this dataset.

A sample of kiwi birds around New Zealand was collected in order to help with conservation efforts. The original data is from: http://www.kiwisforkiwi.org/ and was sourced from the secondary school guides (http://seniorsecondary.tki.org.nz/Mathematics-and-statistics/Achievement objectives/AOs-by-level/AO-S7-1)

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-Northwest 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

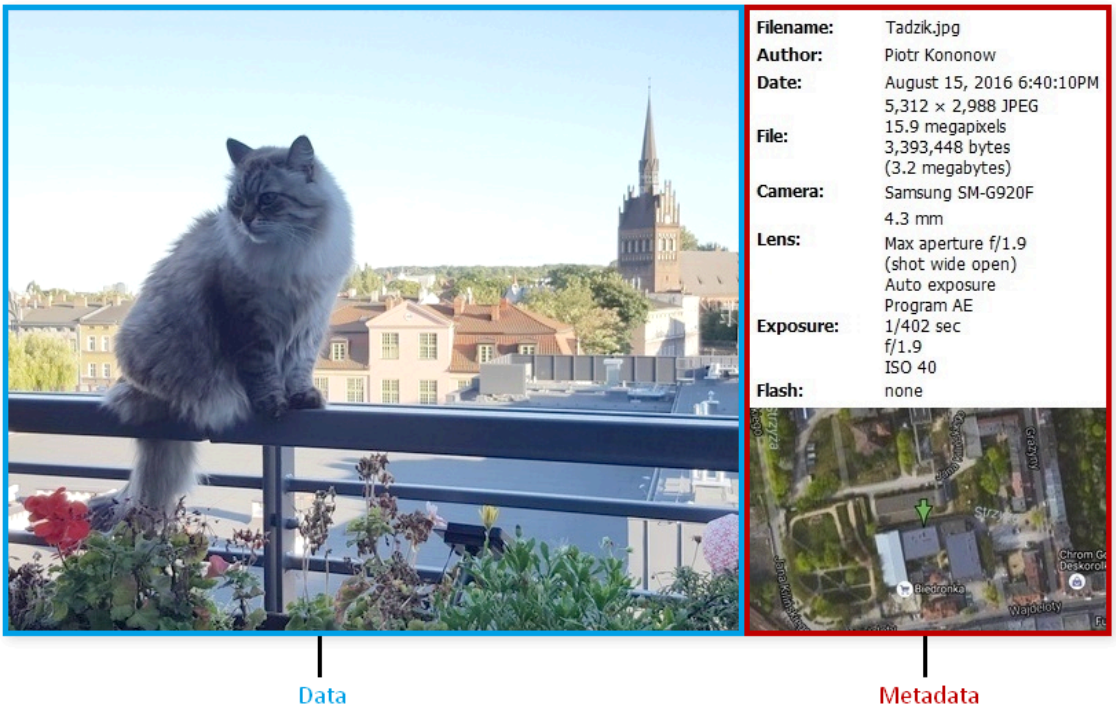
The bird itself is a taonga (treasure) to Māori, who have strong cultural, spiritual, and historic associations with kiwi. Its feathers are valued in weaving kahukiwi (kiwi feather cloak) for people of high rank. Due to the cultural significance to Māori and the traditional knowledge about the bird, tangata whenua are a key stakeholder in

[1] https://atlan.com/what-is-metadata/

kiwi management.

Example 2:

Data can be collected from photographs. Here is metadata about this photograph[2].



Exercise 10:

Go to https://info.grapher.nz/dataset-info/ which provides metadata about the datasets available on NZGrapher. Scroll down to find the following information about the **Babies** dataset.

Sample size	
Location of data collection	
Year data was collected	
Purpose of the study	

Variables	**Description**

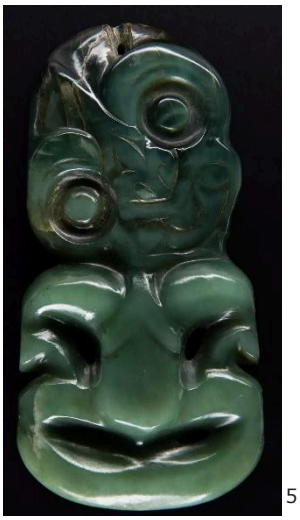
[2] https://dataedo.com/kb/data-glossary/what-is-metadata#toc_1

Ethics

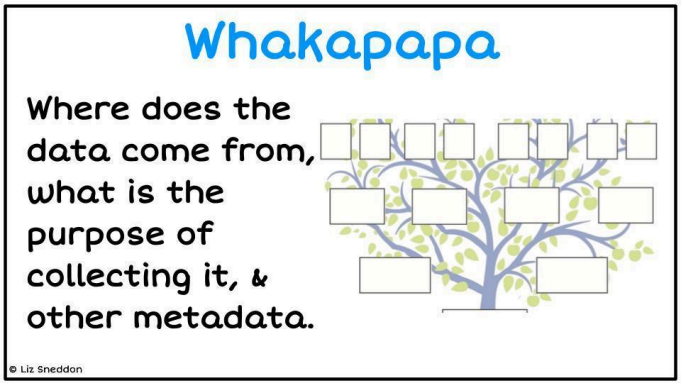
This section has been written using resources from Statistics NZ Tatauranga Aotearoa along with Te Mana Raraunga, Māori Data Sovereignty Network, which is an organisation run by Statistics researchers.

It is important that consideration is given to why the data is being collected, how it is collected, how it is stored, who can access the data, how the data is published, and more.

Dr Warren Williams[3] stated "Data is a living taonga (treasured possession) for me. It is something to be cherished, protected, and cared for. And with that comes responsibility". This quote summarises both the responsibility and value that anyone dealing with data has.


5

Some of the important ethical factors that ākonga should consider when planning data collection (both for primary and secondary data), are[4]:



3 https://www.biometricupdate.com/202202/maori-data-experts-want-indigenous-data-to-be-classified-as-a-treasured-possession

4 https://www.temanararaunga.maori.nz/s/TMR-Maori-Data-Sovereignty-Principles-Oct-2018.pdf

5 https://www.aucklandmuseum.com/discover/research/crafting-aotearoa/recovering-a-precious-heirloom

Example:

Mrs Sneddon's class is planning to write a questionnaire for ākonga to fill in, asking questions about friends and friendships. Here is a brainstorm of some of the ethical considerations they considered when writing the questionnaire, and the instructions for the collection of data:

- Make sure that there is a statement at the start of the questionnaire, letting respondents know that their data will be anonymous, stored securely, and confidential. Also need to make sure that we have a process to ensure this, which might include kaiako storing the filled in questionnaire forms in a locked drawer, names are not written on the questionnaire forms so that we don't know who filled in each form, etc.
- The language we use in the questionnaire needs to be neutral and use respectful language, E.g. don't ask "are you a loser with no friends?". Once the questionnaire is written, we will ask an English kaiako to check the wording before we print it out.
- When approaching ākonga to ask if they would like to fill in the questionnaire, we need to be respectful and use manners, including "please" and "thank you". We can't pressure anyone to fill in the questionnaire and if they don't want to fill it in, we need to be polite and accept this without asking any more questions.

Exercise 11:

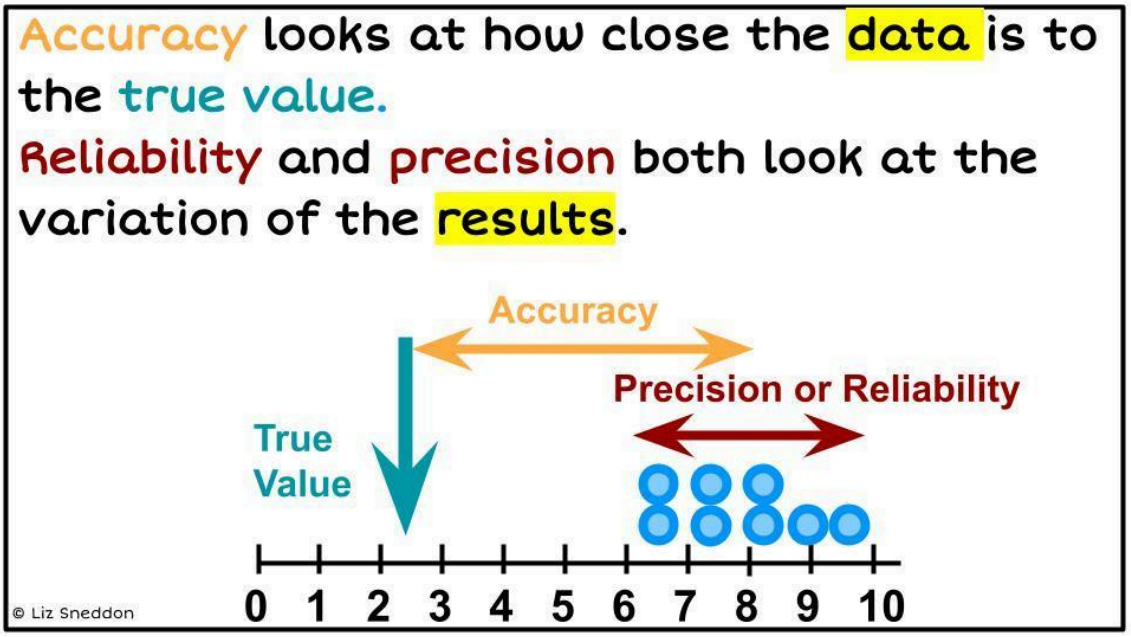
For each of the situations below, identify at least two ethical issues that need to be taken into consideration when planning for the data collection.

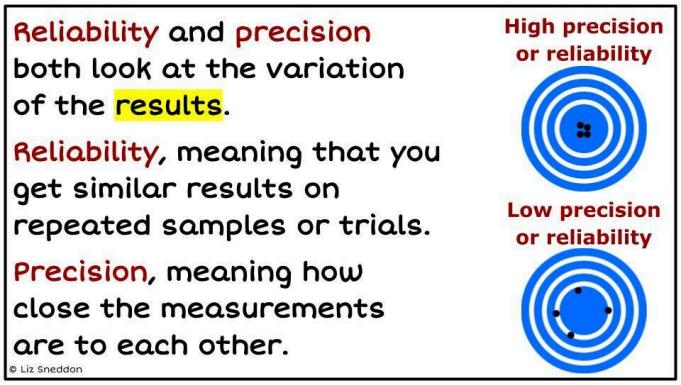
1) A class is planning on collecting data on the fitness levels of ākonga, including how far they can jump, how fast they can run 100m, agility testing, etc.

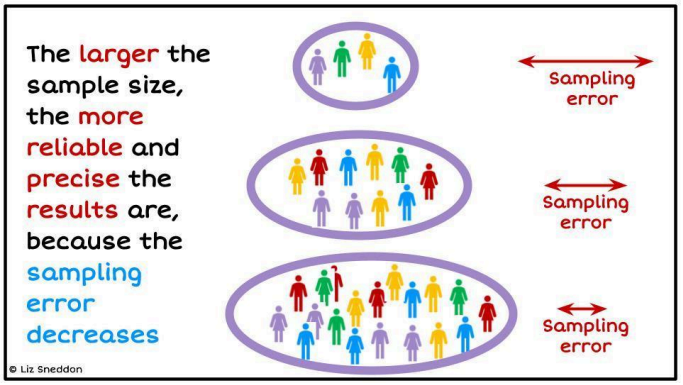
2) Ākonga in Year 13 are planning for a final lunch together before the end of the

year, and need to collect data on dietary information, tikanga or protocols for the meal, attendance, and more.

Reliability & Precision







Exercise 12:

Circle the words that complete the sentences below.

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

2) Larger sample sizes take a **shorter / longer** time to collect, but the data will be _____**more / less**____ reliable.

Sample size for assessment

The more data we collect, the more precise and reliable the results will be, however there is also a balance on primary data collection in terms of how much time it will take to collect the data. The minimum requirements for this standard depends on the type of investigation being done (more on this later), but as a general rule, we aim for 30 or more pieces of data (relationships, experimental probabilities, simulations), 100 for comparisons, and 5 cycles for time series.

Sources of variation

Our world is full of variation. When we collect data, we want to identify key sources of variation that might affect the variables that we are interested in exploring and investigating.

Primary data

There are several different types of sources of variation that we need to consider when writing instructions:

Natural or real variation

Every person or object is different, which leads to differences in measurements for any particular variable.

© Liz Sneddon

Occasion-to-occasion variation

Repeated measurements on the same person or object can change at different occasions.

© Liz Sneddon

Measurement variation

There can be (small) differences in how the measuring equipment is used each time.

© Liz Sneddon

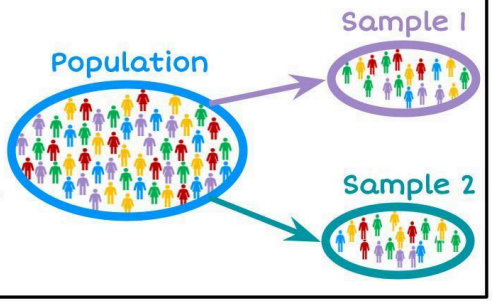
Induced variation

Factors other than natural variation can lead to differences in measurements of the same quantity for different individuals. E.g. plants growth changes due to sunshine, water, soil, etc.

© Liz Sneddon

Sampling variation

If a new sample is taken, the summary statistics will differ



© Liz Sneddon

For your assessment, you will need to **identify, manage, and explain two different** sources of variation in the data that you are collecting and exploring. Ākonga need to focus on the **occasion-to-occasion, measurement, and induced variation**, as both the natural and sampling variation can't be managed and controlled, therefore don't meet the requirement for the standard. You also need to explain the effect if the variation is not controlled.

Example:

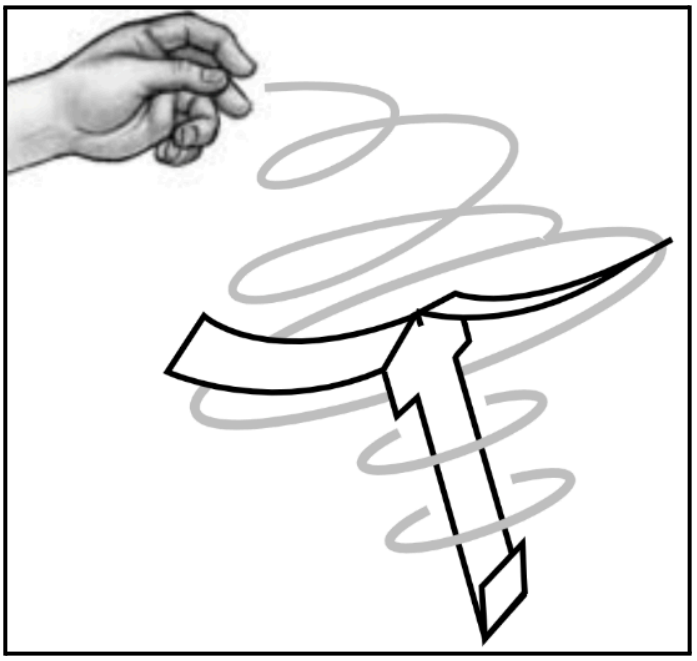
Mrs Sneddon's class will be measuring the foot length of ākonga in her class. The class needed to **identify, explain and manage** at least two different sources of variation.

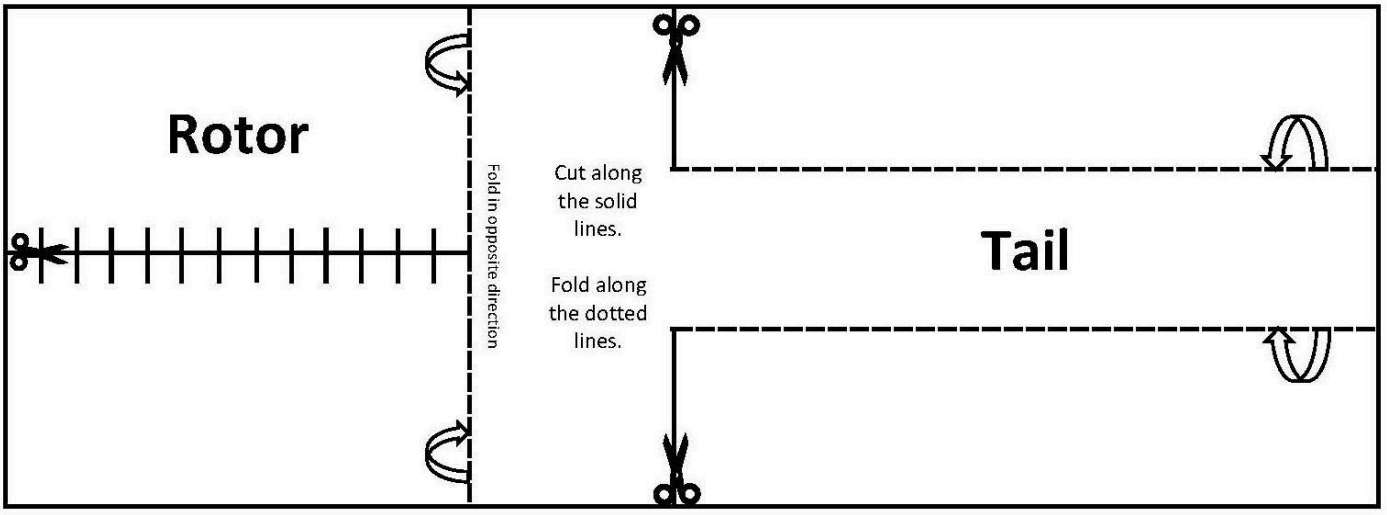
Variation	Example
Natural or real	The foot length of each ākonga varies due to natural variation between different individuals. It is not possible to control or limit this variation, it will always be present.
Occasion-to-occasion	When measuring foot length, one factor to control is the time of day that the measurements are taken. During the day, as we walk on our feet they swell a little, so by the end of the day they are likely to be slightly longer than at the start of the day. To manage this, we should measure the foot lengths all in the same period in the morning, so that the measurements are more accurate and consistent.
Measurement	Using the same measuring tape, so that all the measurements are consistent. Get ākonga to take their shoes 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.
Induced	One factor that can affect the length of people's feet is the age of ākonga. I expect that younger ākonga will have smaller feet than older ākonga, as we go through a series of growth spurts up until around the age of 18 years old. To manage this, we are measuring the foot length of ākonga who are approximately the same age (e.g. they are all in a Year 11 maths class). (Unless we want to explore the relationship between foot length and another numeric variable. In that case, we would want to have a range of feet sizes, so we would want to collect data from younger and older ākonga to get a wide enough range to be able to see if there is a relationship or not.)
Sampling	Depending on whether ākonga in the dataset are a representative sample or not, will determine whether the data is likely to represent the foot lengths of ākonga in the population. If another sample was taken using the same method, then the summary statistics (e.g. minimum, LQ, median, UQ, maximum) are likely to be similar. The only thing that can be done to control and limit this variation is to have a larger sample size.

Class Exercise:

Information:

- See the helicopter template below. Notice that the helicopters have marks on the rotors every 0.5 cm, and you need to put a paper clip at the bottom.
- Make sure you drop the helicopters with a variety of sized rotor blades at least 30 different times. You can do several measurements with each rotor blade length.
- If made correctly, the helicopters will spin as they fall to the ground.
- Kaiako will find an appropriate place for you to drop your helicopters.





Plan

What are some sources of variation that need to be managed / controlled to measure the rotor length and the time it takes to drop to the ground accurately? Remember that you need to identify at least two different types.

Exercise 13:

Here is a recipe for making chocolate brownies from Chelsea Sugar.

Ingredients

250g **Tararua Butter**
½ cup cocoa powder
1 ½ cups **Chelsea White Sugar**
4 eggs
1 tsp vanilla essence
1 cup **Edmonds Standard Grade Flour**
1 tsp **Edmonds Baking Powder**

Method

Preheat oven to 180°C bake. Line an 18 x 28cm sponge roll tin with baking paper.

Melt **Tararua Butter** in a saucepan large enough to mix all ingredients in. Mix in cocoa, remove from heat and stir in **Chelsea White Sugar**.

Add eggs and mix well, then add vanilla essence. Sift in **Edmonds Standard Grade Flour** and **Edmonds Baking Powder** and mix to combine. Pour into prepared tin.

Bake for 25-30 minutes or until brownie springs back when touched lightly.

These instructions are quite detailed which helps to manage any sources of variation. Here are a few questions for you to answer about the recipe.

1) Why is it important to measure the ingredients accurately? Explain how the brownie might be affected if the measurements weren't accurate.

2) Why do the ingredients need to be added in the order given in the recipe? Explain how the brownie might be affected if the ingredients were added in a different order.

3) One of the instructions says to use "a saucepan large enough to mix all ingredients in". Explain why this instruction was included in the recipe.

4) Identify one source of variation that has been managed in the instructions (other than the examples used in the questions above).

For each of the questions below, identify and describe at least two different sources of variation (from occasion-to-occasion, measurement, and induced variation).

5) Measuring the time to run 100 meters for ākonga in your class.

6) Measuring the weight of a bag of apples.

7) Measuring the volume of music on different radio stations.

Writing a Plan

When planning to collect primary data, after we have identified sources of variation, respondents/participants, ethics, data collection method, sample size, the next step is to write a set of instructions that everyone who collects the data will follow the same processing, helping us to collect consistent and accurate data.

Example:

Here are instructions for measuring handspan.

1) Ask the person to place their right hand flat on a piece of paper on a desk, palm down.
2) Ask the person to spread their fingers as wide as they can.
3) Using a pen, mark the edge of the persons' smallest finger and thumb.
4) The person can now remove their hand.
5) Using a tape measure, measure the distance (in mm) between the two marks.
6) Record this measurement on a data table.

Exercise 14:

1) Working in groups of 2 or 3, write a set of instructions to walk from the doorway of the classroom to the opposite corner (without walking into the furniture, tripping over, or walking into other ākonga).

2) Give your instructions to your teacher to test them out.

As your teacher tests them, think about what improvements you could make and write these down.

Secondary data

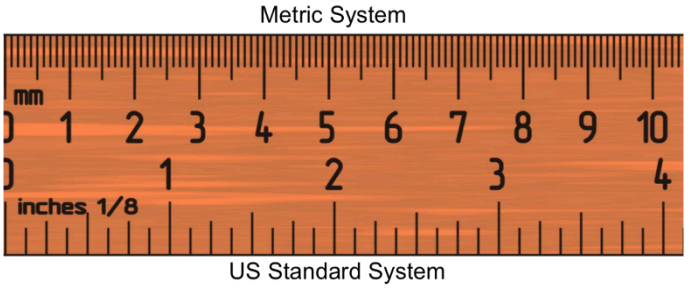
For data that has been collected by other people, we need to identify sources of possible variation from the information provided. Think about the ideas of **representative sample, accurate and consistent data**. Any sources need to be reasonable and likely, not trivial.

It can be helpful to do a little research on google to find out more, and some of the questions below[6] may be helpful:

1) Who collected the data?
2) When was the data collected?
3) What was the purpose for collecting the data?
4) What was the data collection or survey question asked to collect the data?
5) Were the survey questions appropriate?
6) How was the variable measured?
7) What are the possible outcomes for the variable?
8) What are the data types?

Incorrect Example:

An example of a **trivial** answer (this is **NOT appropriate** for NCEA Level 1) is:

To measure the length of a pen, they should always use centimeters so that the measurements are consistent, not having some in inches and some in centimeters.	

[6] Arnold, P (2022), *Statistical Investigations*. NZCER Research.

Example 1:

One of the datasets on NZGrapher is the Kiwi dataset. Here is the metadata about this dataset (I have selected two of the variables).

A sample of kiwi birds around New Zealand was collected in order to help with conservation efforts. The original data is from: http://www.kiwisforkiwi.org/ and was sourced from the secondary school guides (http://seniorsecondary.tki.org.nz/Mathematics-and-statistics/Achievement objectives/AOs-by-level/AO-S7-1)

Variable	Description
Gender	M-Male F-Female
Weight(kg)	The weight of the kiwi bird in kg

Variation	Example
Natural or real	The gender and weight of each kiwi bird varies due to natural variation between different individual kiwi birds. It is not possible to control or limit this variation, it will always be present.
Occasion-to-occasion	Measuring the heights of kiwis at different times of the day may give slightly different results. For example, in the early evening straight after the kiwi has been resting all day, its weight is likely to be slightly lower as it will have been sleeping and digesting food during the day. Whereas if it is weighed during the morning, shortly after it has spent all night foraging for food, its weight may be slightly higher. To manage this variation, kiwis should all be measured in the morning before lunch in order to have more consistent measurements.
Measurement	Here is a video I found (bit.ly/WeighingKiwi) that shows the process of weighing a kiwi bird. This may or may not be the method used for this dataset, but the equipment used is likely to be the same (a spring scale). As you can see in the video, the kiwi bird is moving around in the bag as it is being weighed, so the weight will be a good estimate, but won't be completely accurate due to the movement of the kiwi. To manage this variation, the same person and equipment should be used to measure the weight of each kiwi. 

	In other research I found[7], it is difficult to identify the sex of a kiwi bird, and there are currently two measures that are used. One is where scientists measure the length of the beak (or bill) of an adult kiwi, and depending on which species it is, if it is longer than a reference amount it is classified as female, or shorter lengths are classified as males. The second method is to identify the sex by extracting DNA from a feather, which correctly identifies the sex in 95% of samples (this takes around one week to confirm). Therefore, when people identified the sex for this dataset, there may be some data that is not as accurate due to the difficulty of determining the sex of the bird while out in the field (unless it is registered with a leg band or other identification). To manage this variation, it would best if DNA testing was used to identify the sex of all kiwi birds as this would give more accurate data.
Induced	The weight of each kiwi bird may differ during different seasons. For example, in the breeding season between June to March, the food is most plentiful[8]. When the food is more plentiful, each kiwi is likely to weigh more, whereas between March and June when food is less plentiful, each kiwi may weigh more – so each kiwi birds weight changes throughout the year. To manage this variation, it is best to collect the weights of kiwis during the same month of the year so that the weights are more consistent. A second example is if a female kiwi who is carrying an egg (which takes around 30 days to develop before she lays the egg) is weighed, then her weight is likely to be higher than when she is not pregnant. So being pregnant affects the weights of only the female kiwis. To manage this variation, people measuring the weights of kiwis should avoid collecting measurements during the breeding season. Other factors that may affect the weights of kiwi birds include the location it is found (some areas may be warmer or cooler requiring more or less energy needed to keep warm), the number of predators in the area (more predators would likely reduce the weight as kiwis will have to run away and hide, as well as have less time to feed), and many other factors.
Sampling	Depending on whether the kiwi birds in the dataset are a representative sample or not, will determine whether the data is likely to represent the sex and weights of kiwis in the population. If another sample was taken using the same method, then the summary statistics (e.g. minimum, LQ, median, UQ, maximum) are likely to be similar. The only thing that can be done to control and limit this variation is to have a larger sample size.

[7] http://avianrearingresource.co.uk/species/documents/169.pdf
[8] https://savethekiwi.nz/about-kiwi/kiwi-facts/kiwi-life-cycle/

Exercise 15:

The datasets below are secondary datasets. Identify at least two different sources of variation for each question (from occasion-to-occasion, measurement, or induced variation).

1) Here is metadata information about the **Marathon** dataset found on NZGrapher.

The data 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 persons average stride length over the marathon in cm.

Sources of variation:

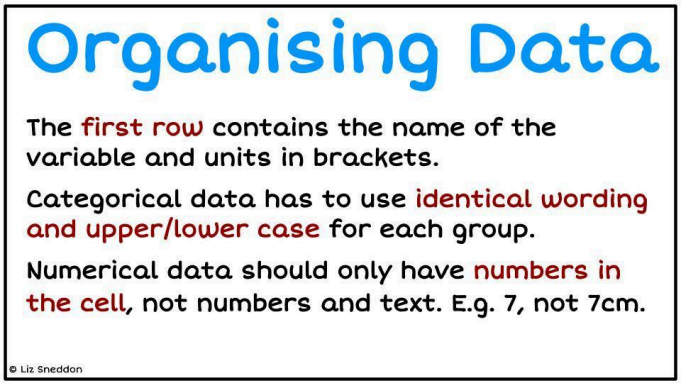
2) Here is metadata information about the **TS – Temperatures Auckland** dataset found on NZGrapher.

Temperature data from the weather station at Auckland Airport sourced from NIWA.

Variable	Description
Month	The Month of the Data
Tmax	Average Maximum Temperature for the Month
Tmin	Average Minimum Temperature for the Month

Sources of variation:

We've looked at how data needs to be organised with variables in columns in a spreadsheet. Now we need to look at data in a little more detail and make sure that we know how to setup a spreadsheet (either Google Sheets or Microsoft Excel). We also need to clean up any data that is not formatted correctly or is incorrect.



Example:

Here is a spreadsheet:

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

Notice that the **first row** contains the **name of the variable**, and I also suggest including the **units in brackets** as this helps when we draw graphs.

Another thing to notice is that the data in the column with the categorical variable **Takes PE?**, is identical in their wording and upper/lower case (e.g. "**no**", not a combination of "**NO**" or "**No**" etc.). It is important that all the **categories are identical**. Even having a space at the start or end of a word can create problems with graphing. An easy way to do this if you are entering the data manually, is to type in the first row, and then copy and paste the category for other rows.

Also notice that the numerical data only contains a number in each cell. For example, if the cell contained the number and the units, then the graphing programmes automatically classify these cells as text rather than number. So, all numerical data should not contain any text in the cells.

Formatting Time Series data

When we have time series data, there are a range of different formatting options depending on what units the data is. For example, data could be hourly, daily, weekly, monthly, quarterly (4 times per year which is common for businesses), or annual.

The information below is specifically for how the data needs to be formatting for **NZGrapher**. Other programmes may have different formatting requirements.

Firstly, if you have annual data (one data point per year), then you need to only put the Year. E.g. 2000, 2001, 2002, 2003, etc.

For any other time periods the general formatting requirements is to have **a number**, then **a letter**, then **a number**. E.g. 2000D1

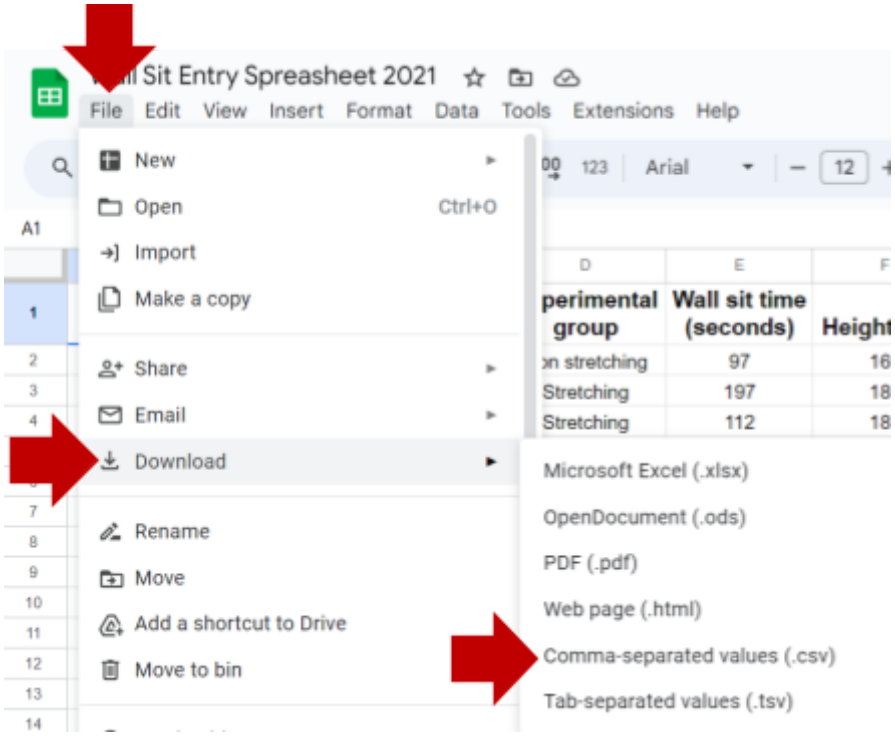
Time	Quarterly	Monthly	Week	Day	Hour	Custom
Note	Use the letter "Q"	Use the letter "M" and 2 digit numbers for the months	Use the letter "W" and 2 digit numbers for the weeks	Use the letter "D" and go up to Day 7, then restart at Day 1	Use the letter "H" and 2 digit numbers for the hours	Use the letter "C"
Example	2000Q1 2000Q2 2000Q3 2000Q4 2001Q1 2001Q2 2001Q3 2001Q4 …	2000M01 2000M02 2000M03 2000M04 2000M05 2000M06 2000M07 2000M08 2000M09 2000M10 2000M11 2000M12 2001M01 …	2000W01 2000W02 2000W03 2000W04 2000W05 2000W06 …	1D1 1D2 1D3 1D4 1D5 1D6 1D7 2D1 2D2 …	1H01 1H02 1H03 1H04 1H05 1H06 1H07 … 1H24 2H01 2H02 …	1C1 1C2 1C3 1C4 1C5 2C1 2C2 …

Saving spreadsheets

In order to be able to import datasets into graphing programmes, it needs to be saved as a "**.CSV**" file. Here are instructions to do this.

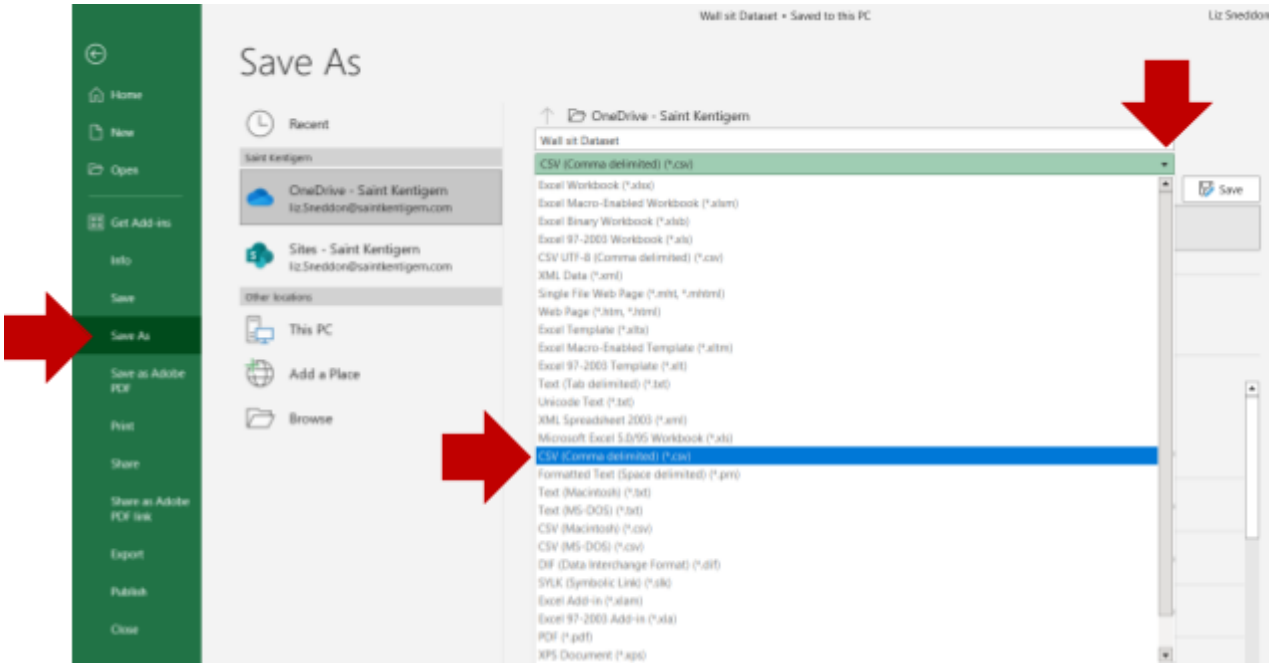
If using Google Sheets:

Click on the **File** menu, then select **Download**, and select **Comma-separated values (.csv)**



If using Microsoft Excel:

Click on the **File** menu, select **Save As**, choose **CSV (Comma delimited) (*.csv)**, and click on the **Save** button.



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 either correct the value or make the cell blank (or enter a 0).

Exercise 16:

1) Find any data that doesn't make sense and highlight the values.

Elastic type	Length (cm)	Number of marbles	Stretched length (cm)	Weight (grams)	Obstacles
WIDE	69 cm	6	890 mm	34.8gm	Yes
widE	690	8	92	46.4	No
	0.69 metres	10	95	58	YES
Cord		12	98 cm	69.6 grams	NO
CORd		14	99	81.2	No
CORD		16	1020 mm	92.8	Yes
Cord	690	18	106	104.4 gm	yes
String	690	20	1080 mm	116	
Narrow	68.03 cm	-10	1.16metres	-58	yes

Write the corrected values in the table below.

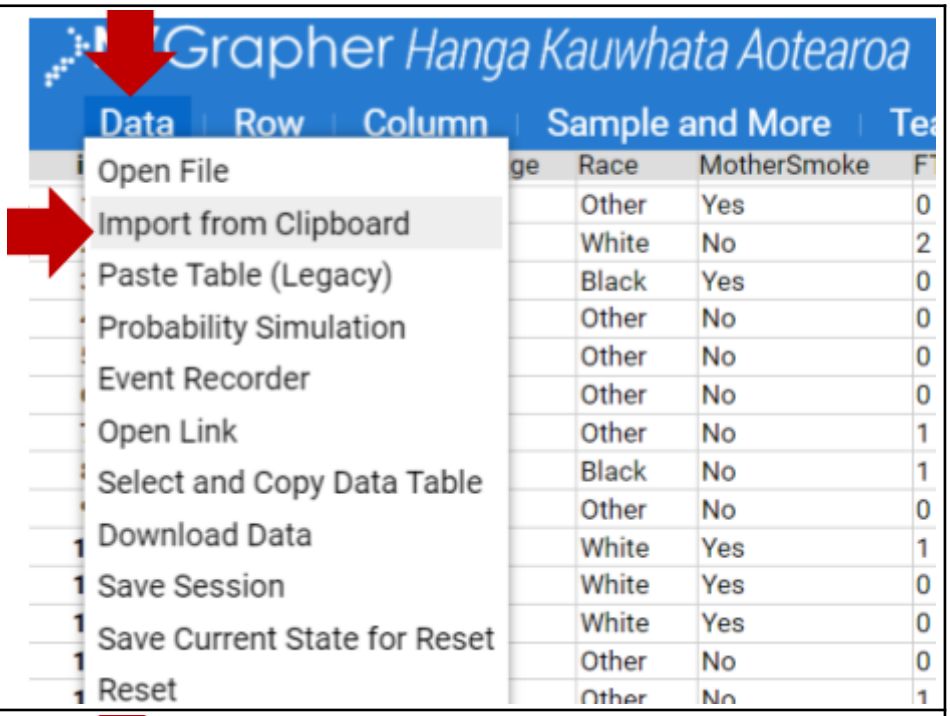
Elastic type	Length (cm)	Number of marbles	Stretched length (cm)	Weight (grams)	Obstacles

Getting data into NZGrapher

There are two different methods for getting data from your spreadsheet into NZGrapher.

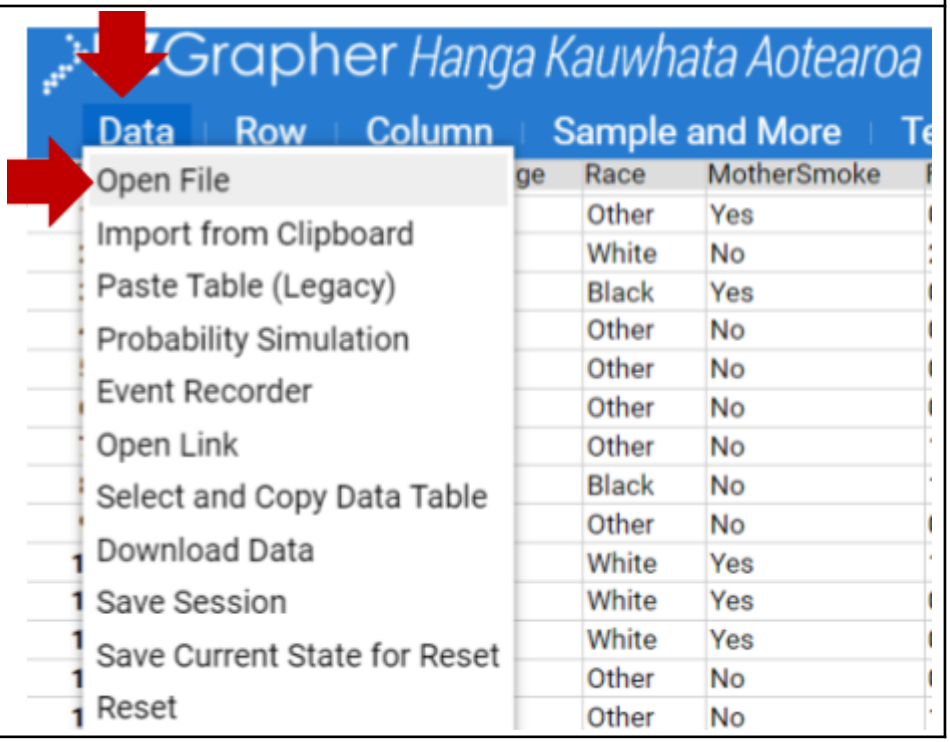
Method 1:

Select the data you want to copy, **copy** this, then go into **NZGrapher**, go to the **Data** menu, and select "**Import from Clipboard**"



Method 2:

Go to the **Data** menu, select **Open File**, then select the **.csv** file that you have saved.



It is also possible to enter data for experimental probabilities and simulations directly into NZGrapher without using a spreadsheet. The instructions for this are covered in the Probability workbook.

Exercise 17:

1) Open the Google sheet using data in this link: bit.ly/DailyMovieData

Copy the data and import it into NZGrapher.

2) Type the data shown in the table below (which has information about recipes) into an Excel spreadsheet. Save as a **.csv** file and **import** this file into NZGrapher.

Flavour	Prep time (minutes)	Cooking time (minutes)	Ingredients	Difficulty level
savoury	30	65	15	difficult
sweet	20	20	9	difficult
savoury	15	10	12	easy
sweet	20	60	8	easy
savoury	20	5	12	easy
sweet	10	15	14	easy

Data displays

Once we have our variables and have identified their data type, we can then decide how to display and summarise the data.

Some displays can be used with either categorical or numerical data, but some displays can only be used for numerical data or only categorical data. The key to knowing which graph can be used is to identify the data type.

Tally chart

Good for summarising categorical data.

𝍸 = 5

Frequency table

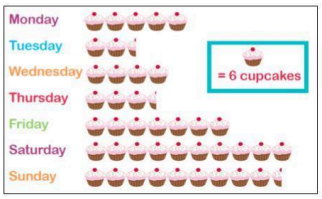
Summarises categorical data.

Gender	Tally marks	Frequency
Female	𝍸 𝍸	10
Male	𝍸 III	8

Pictogram

Pictograms use pictures to represent data.

You need to have a key/legend to state how many units each picture represents.

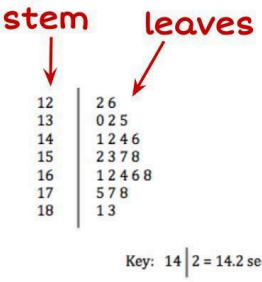


Stem-and-leaf

Used when a small number of NUMERIC variables.

Must have a key.

Split data into a stem and a leaf.

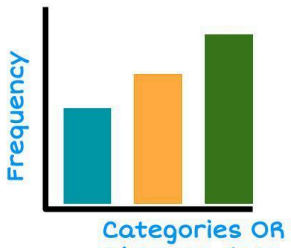


Bar Graph

A summary graph for either categorical or discrete data.

You need to first have the frequencies for each category.

Gaps between the bars.



Histogram

A summary graph for continuous data.

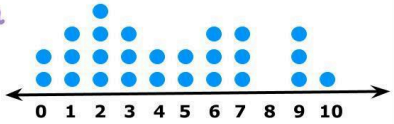
No gaps between the bars.

Summarise the data into a frequency table before making the graph (putting data into equal size ranges)

Dot plot

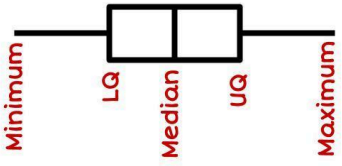
Numeric data

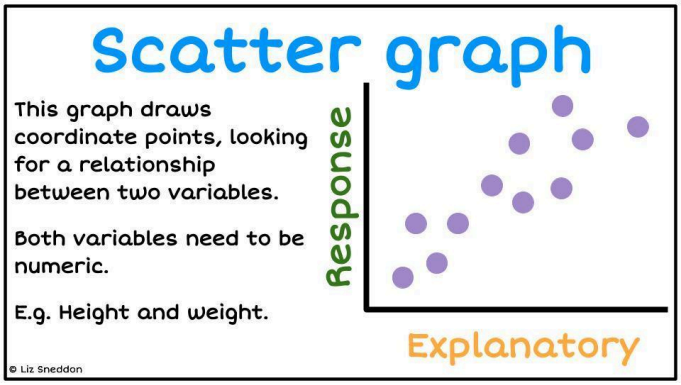
Plot every data point.

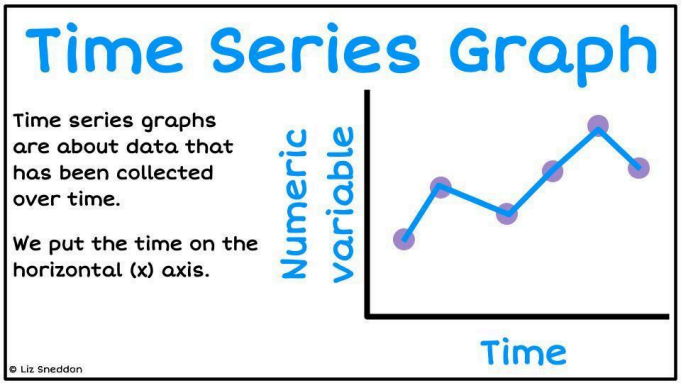


Box plot

A summary graph of the data.







Example:

Identify the name of the display, the variable(s), and the data type(s).

Shoes We Wear

Shoes	Tally	Total
	𝍸	5
	𝍸	3
	𝍸	4

This is a **frequency table**.

The variable is **the type of shoes**.

When people answer this question, they choose sneaker, loafer, or sandal. These options are word answers, so the data is **Categorical**.

There is no order to the shoe types, so the data is **Nominal**.

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

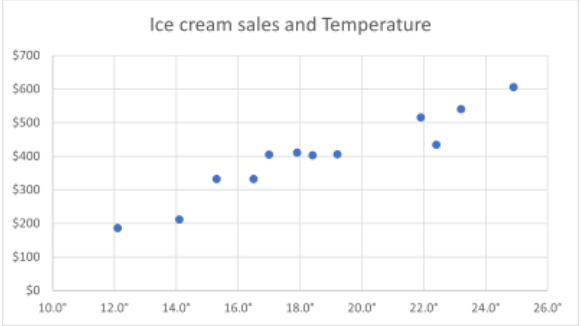
Key: 14 | 2 = 14.2 seconds

The data display is a **stem and leaf graph**.

The variable is the **time to run a race**.

The data is recorded is the time measured in seconds. Therefore, the data is **Numerical**.

Time is a measurement (and can include decimal values), so the data is **continuous**.



The data display is a **Scatter graph**.

There are two variables. One is the **Temperature** and the second is the **amount of money in sales of ice cream**.

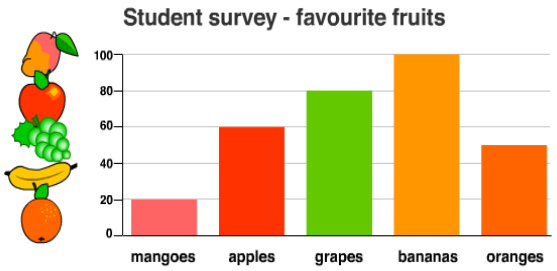
Temperature is **Numerical** data, and **continuous**.

Ice cream sales is **Numerical**, and **continuous**.

Exercise 18:

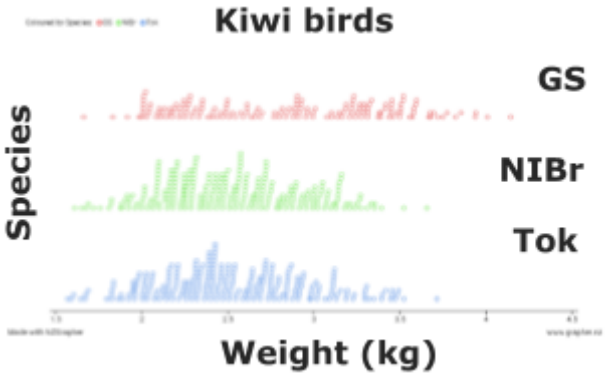
Identify the name of the data display, if there are one or two variables, what the variables are, and the data type(s) for each variable.

1)



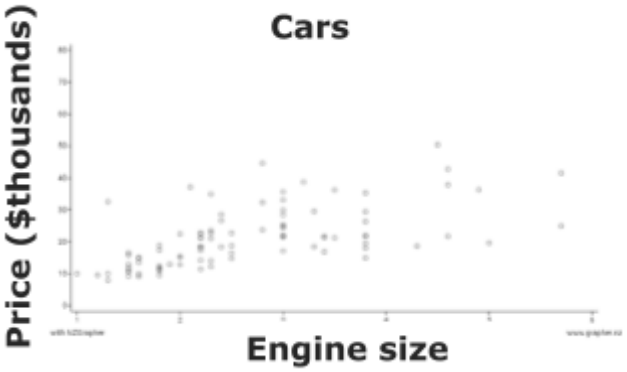
Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

2)



Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

3)



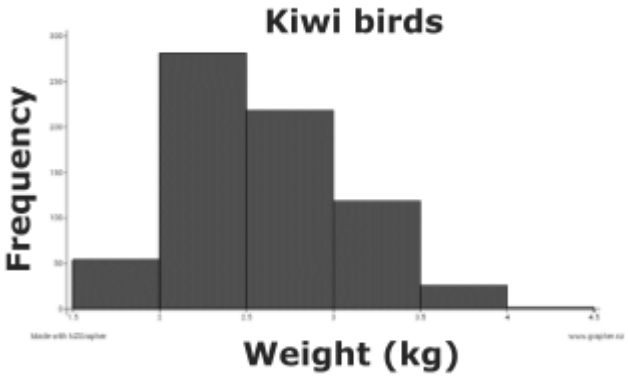
Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

4)

What flavor of ice cream would you pick?			
	Chocolate	Vanilla	Neither
Children	40	22	15
Teens	12	16	45
Adults	55	54	10
Total	107	92	70

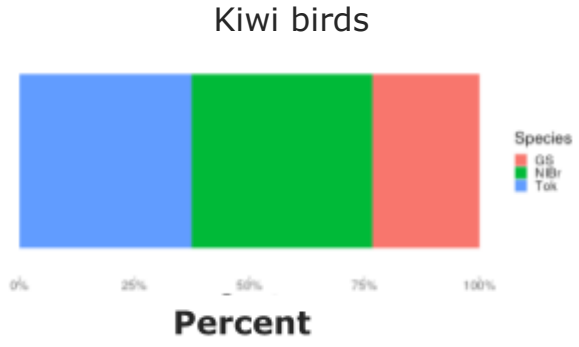
Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

5)

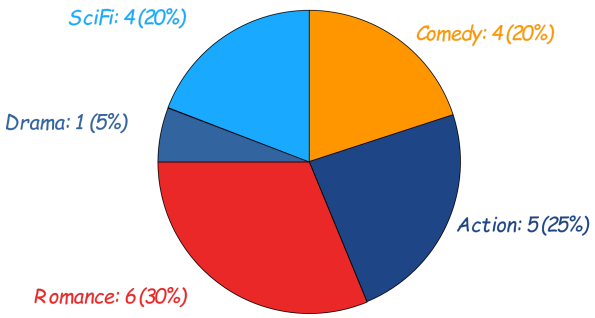
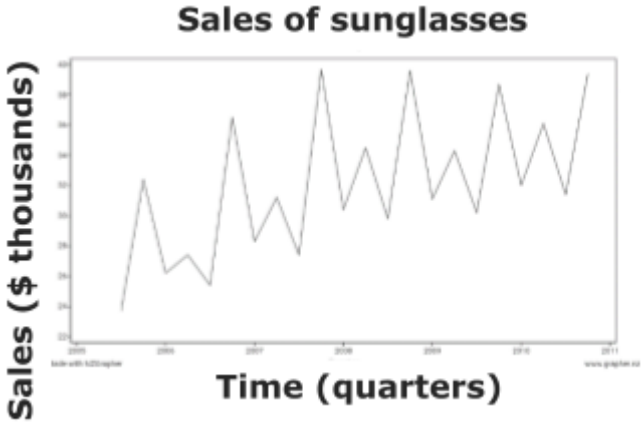
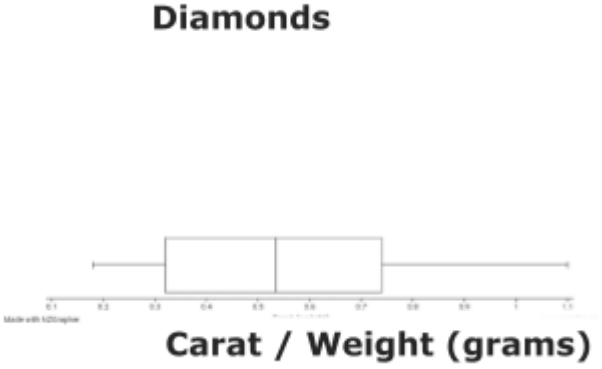


Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

6)



Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

7)		**Data display**	
		Number of variables	One / Two
		Variable definition(s)	
		Data types	
8)		**Data display**	
		Number of variables	One / Two
		Variable definition(s)	
		Data types	
9)		**Data display**	
		Number of variables	One / Two
		Variable definition(s)	
		Data types	

10)

Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

11)

Number of days with rain	Number of weeks
0	II
1	𝍸
2	𝍸
3	𝍸
4	𝍸 𝍸 𝍸 IIII
5	𝍸 I
6	𝍸 I
7	IIII

Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

12)

Speeds of cars driving in Northtown, before and after putting up a speed limit sign.

Before		After
9	1	1 8
6 5 5 4 0	2	2 3 5 8 9
9 8 6 2 1 0	3	0 0 2 2 2 5 9
8 8 7 5 5 2 2 0	4	0 3 4 5 7 7
8 6 6 5 5 4 2	5	0 1 1 3 6
8 3 2 2 1	6	2 4 5 7
1	7	3

Key to reading the speeds

Before: 9|1= 19 After: 1|1=11

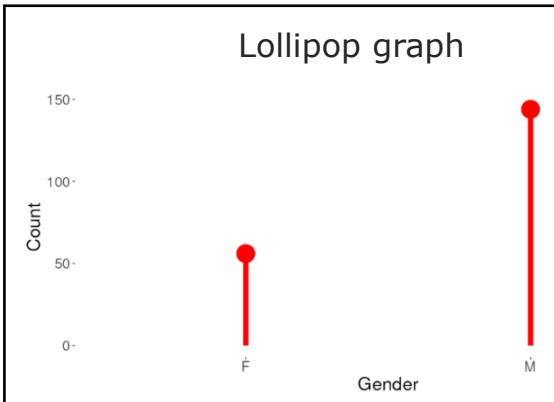
Data display	
Number of variables	One / Two
Variable definition(s)	
Data types	

Extension questions

Here are some graphs that you may not have seen before, they are all exploring data from athletes in NZ running a marathon.

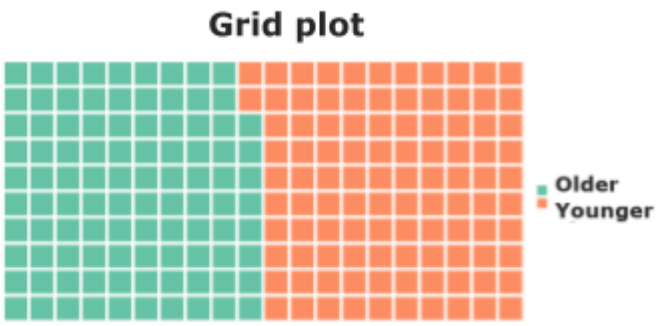
Can you identify what the variable(s) are and the data types for each variable?

13)



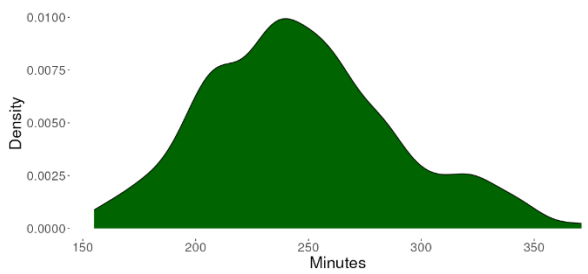
Variable(s)	
Data types	

14)



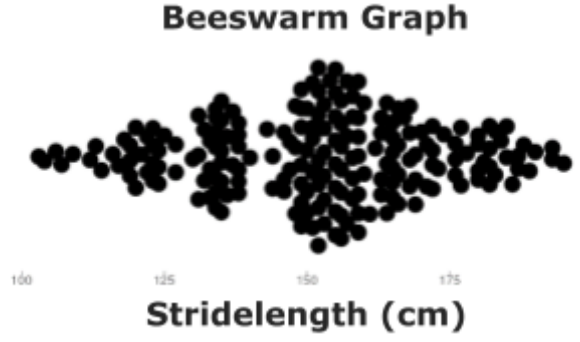
Variable(s)	
Data types	

15)



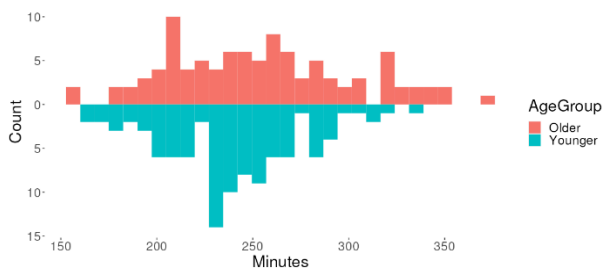
Variable(s)	
Data types	

16)



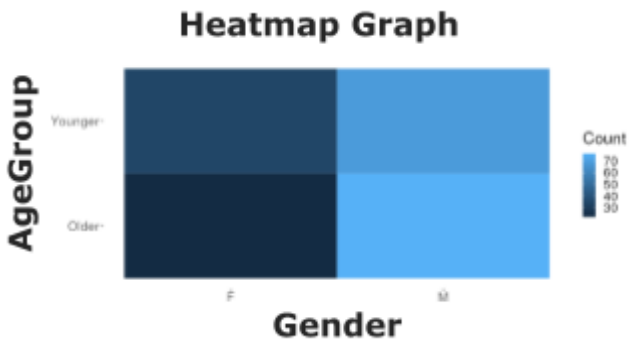
Variable(s)	
Data types	

17)



Variable(s)	
Data types	

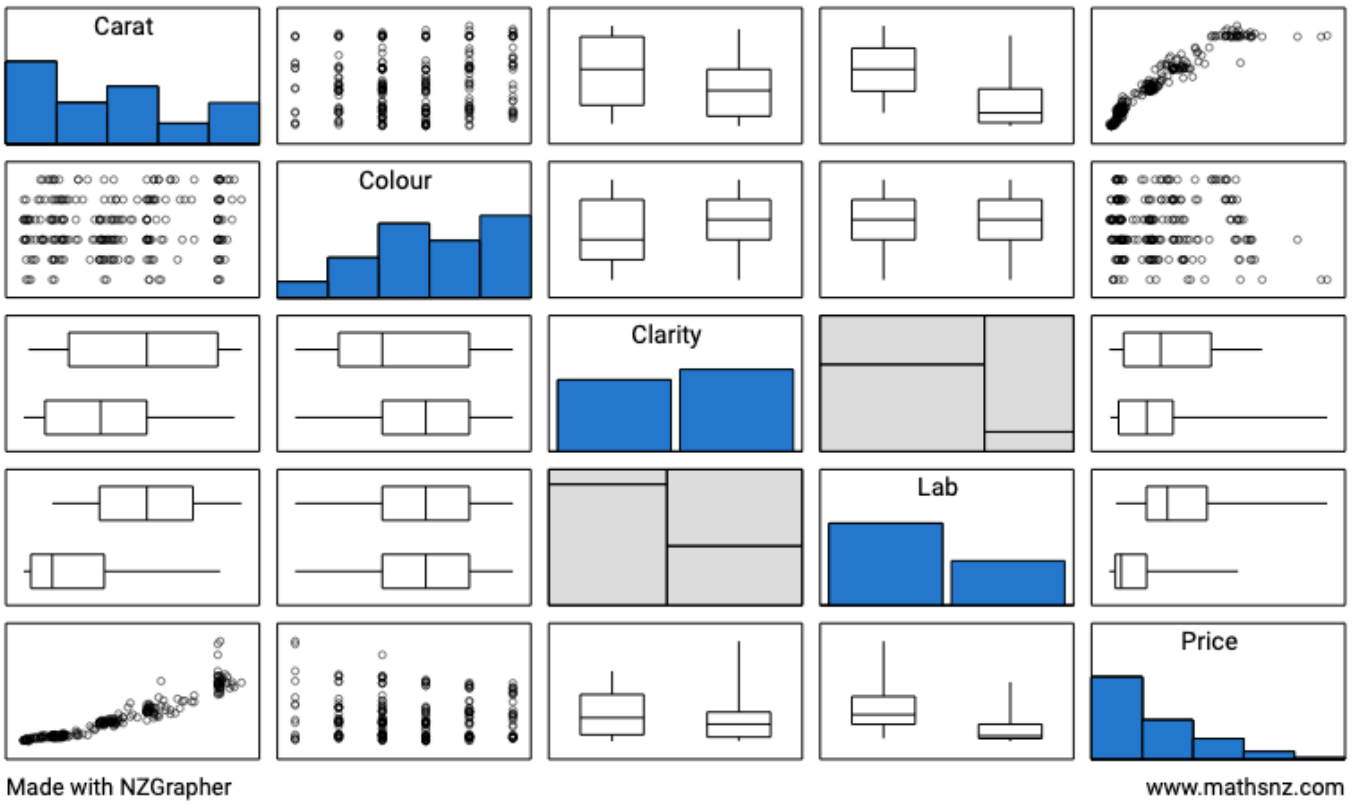
18)



Variable(s)	
Data types	

NZGrapher https://grapher.nz/

Pairs plots are useful as they give 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 19:

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.

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.

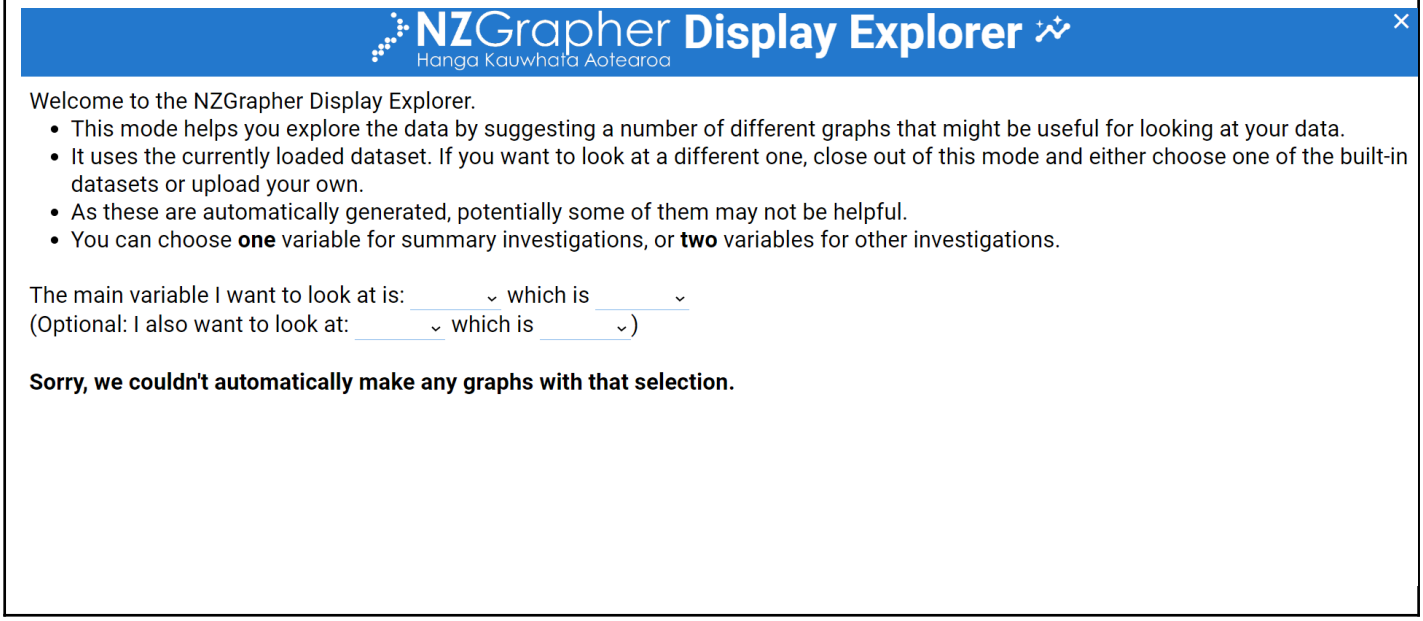
Exploring data in NZGrapher

There is another tool for exploring data in NZGrapher. One of the menus is called **Display Explorer**. Before you click on this option, make sure that the dataset you want to explore is loaded up and showing.

Class Exercise:

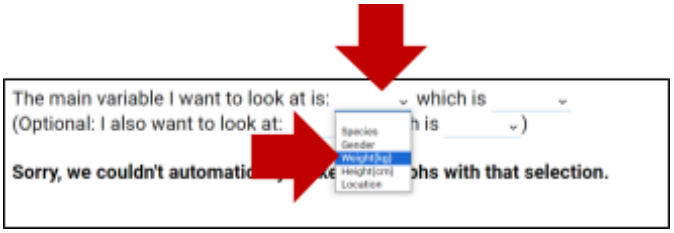
Step 1:

Open **NZGrapher**, select the **Kiwi** dataset, and click on the **Display Explorer** menu. The following screen will appear



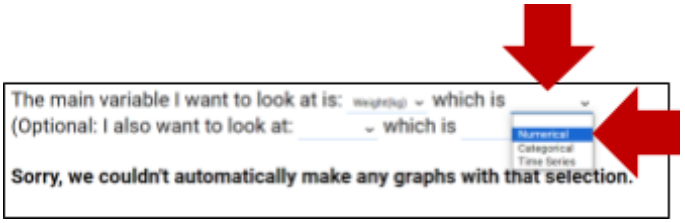
Step 2:

Select the **Weight** variable.



Step 3:

Choose **Numerical** data type.

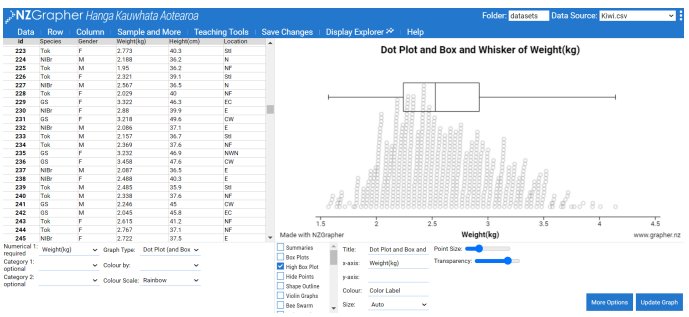


The following graphs will appear. These are all different ways to represent the weight data.

Click on any of the graphs to open in full NZGrapher mode and edit the settings.

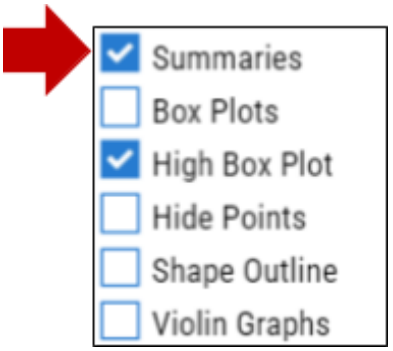
Step 4:

Click on the **Dot plot and Box and Whisker** graph. This will take you to that specific graph.



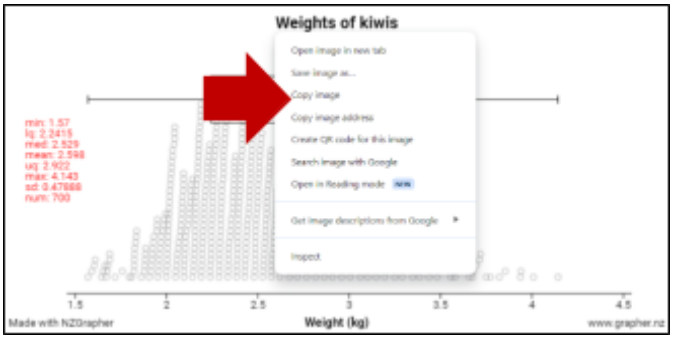
Step 5:

Click on the Summaries tick box and change the Title of the graph.



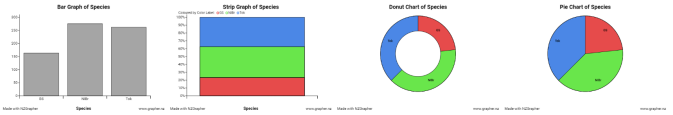
Step 6:

Hover your mouse anywhere over the graph and click on the **right-hand** button of your mouse. Select **Copy image**. Then go to a blank Word document and paste this graph.



Step 7:

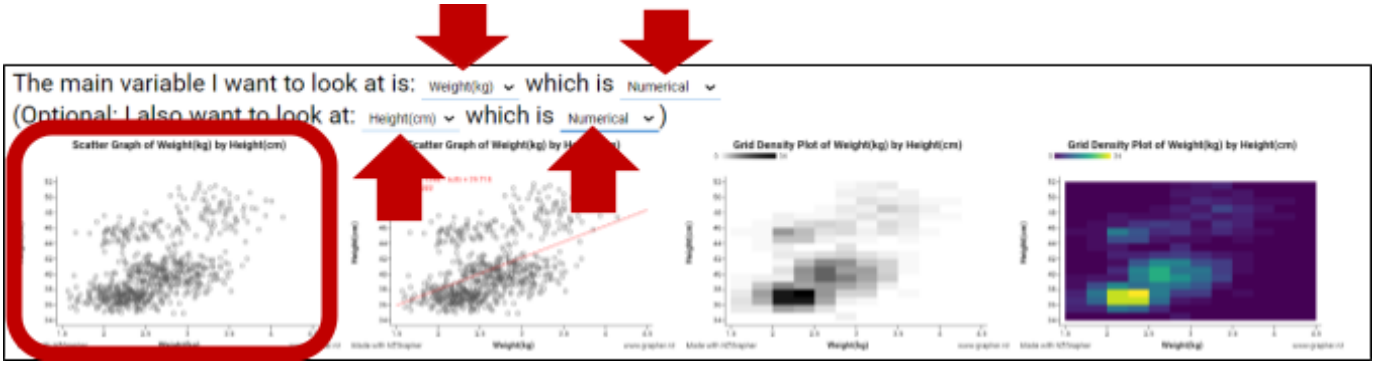
Go back to the **Display Explorer,** select the variable **Species**, select **Categorical** data type. The following graphs will appear:



Click on one of the images to explore further. Copy and paste this into your Word document.

Step 8:

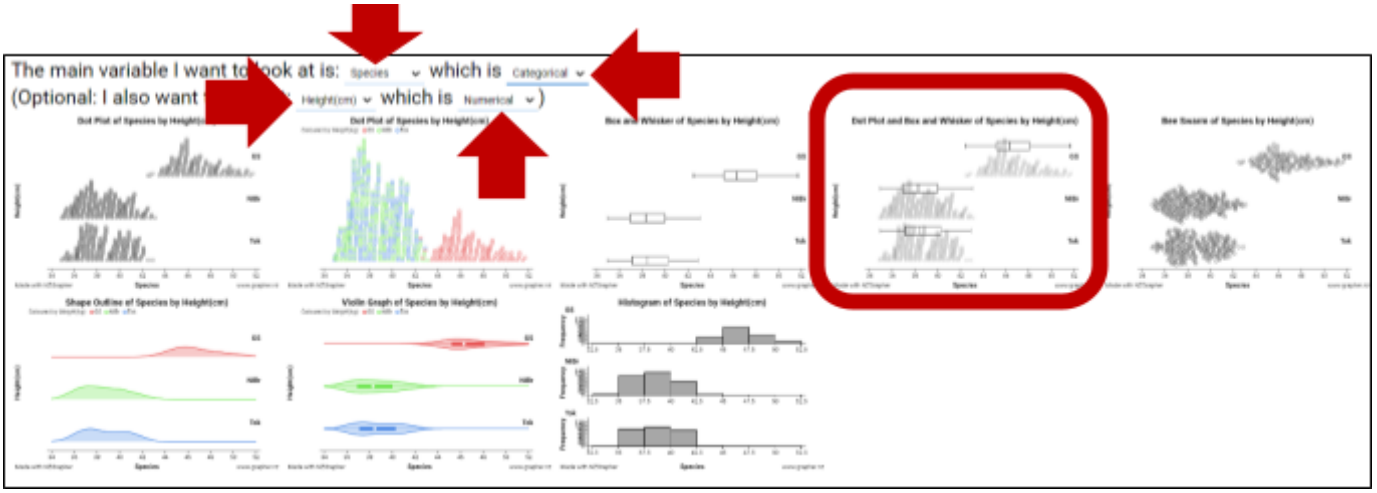
Go back to the **Display Explorer,** select the variable **Weight**, select **Numerical** data type. Then choose a second variable **Height,** select **Numerical** data type. The following graphs will appear:



Click on the first **Scatter graph** to explore further. Copy and paste this into your Word document.

Step 9:

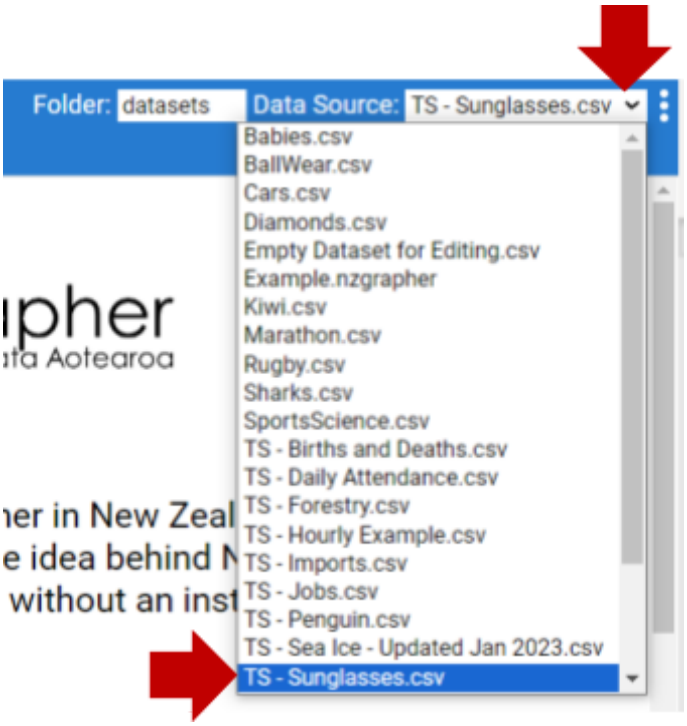
Go back to the **Display Explorer,** select the variable **Species**, select **Categorical** data type. Then choose a second variable **Weight,** select **Numerical** data type. The following graphs will appear:



Click on the fourth **Dot plot and Box and Whisker graph** to explore further. Copy and paste this into your Word document.

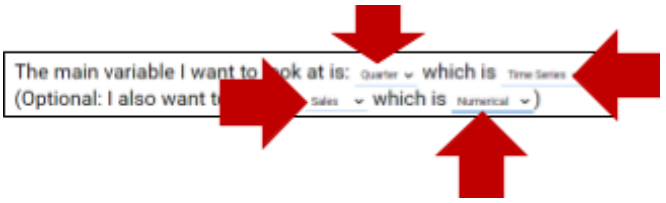
Step 10:

Change the data set to **TS – Sunglasses.csv**. (Any of the datasets that start with the letters "**TS**" are Time Series datasets.)



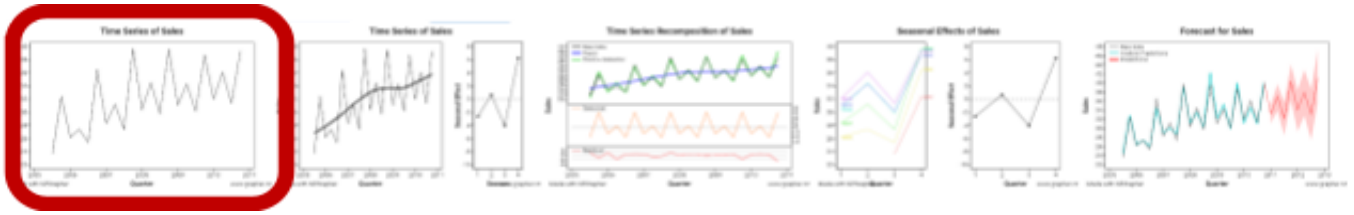
Step 11:

Click on **Display Explorer**, select the variable **Quarter**, choose **Time Series** data type. Then for the second variable, select the variable **Sales** and choose **Numerical** data type.



Step 12:

The following graphs will appear. Click on the first graph and copy and paste this into your Word document.



There are a few more tips and ideas of useful additions to graphs in NZGrapher, and we will look at these in further workbooks when looking at specific investigations.

Exercise 20:

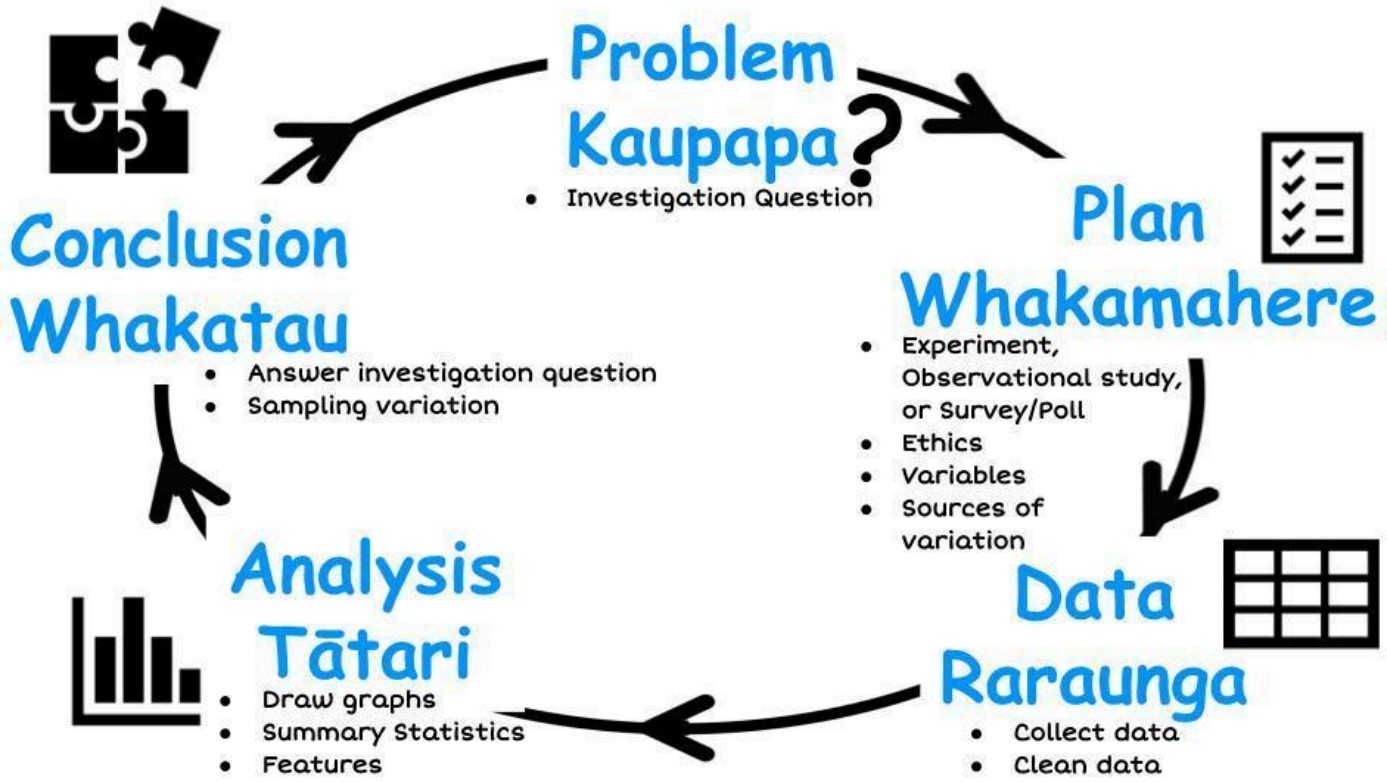
1) Explore the **Marathon** dataset in NZGrapher.

2) Explore the **TS – Temperatures Auckland** dataset in NZGrapher.

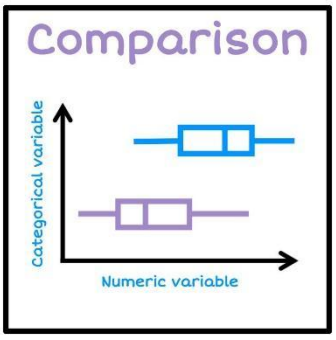
Statistical Investigations

PPDAC

This is the statistical enquiry process that we will be using to carry out statistical investigations.

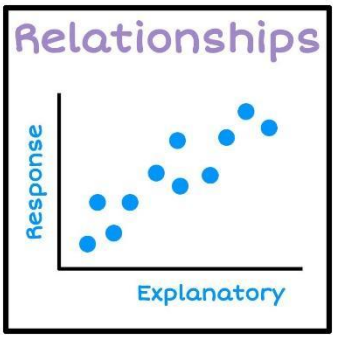


There are separate workbooks to discuss specific investigations (see below). These will focus on writing specific investigation questions (Problem), and the Analysis and Conclusion sections. We need to know what investigation problem is, because the data, graphs, and calculations we do for analysis, and the conclusions will be different for each data type.



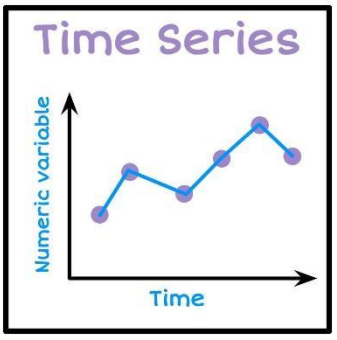
Types of variables:

- One Categorical,
- One Numerical



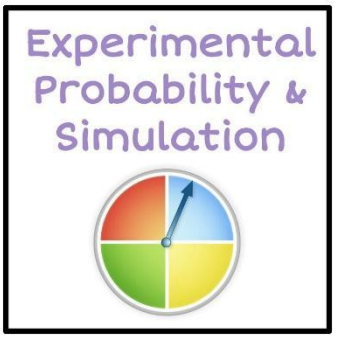
Types of variables:

- Two Numerical.



Types of variables:

- Time,
- One Numerical.



Types of variables:

- Categorical,
- Numerical.

Exercise 21:

For the following questions, use the dataset shown below to identify the data types of the variables, the different displays that could be used to display the data and the type of investigation.

Here are the variable definitions for a multivariate dataset from Rugby players:

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)

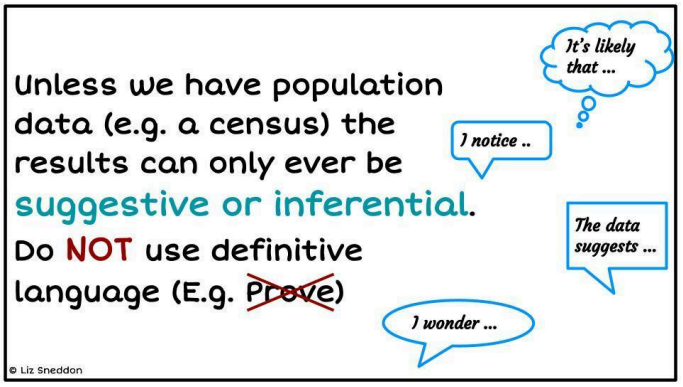
1)	Investigate the heights of NZ versus South African rugby players.	**Data types**	
		Data display(s)	
		Investigation type	
2)	Investigate the height and weight of rugby players	**Data types**	
		Data display(s)	
		Investigation type	

For each of the questions below, identify the data types for any variables and identify the type of investigation.

3)	Investigate the probability of winning a game of Monopoly.	**Data types**	
		Data display(s)	
		Investigation type	
4)	Investigate the change in the number of people living in NZ between 2010 and 2024.	**Data types**	
		Data display(s)	
		Investigation type	
5)	Investigate the relationship between the number of hours ākonga study and the number of credits they get.	**Data types**	
		Data display(s)	
		Investigation type	
6)	Investigate if ākonga who play sports tend to run faster than ākonga who don't play sports.	**Data types**	
		Data display(s)	
		Investigation type	

Language

When data is collected, we need to be aware that we data is never an exact match to the true situation. Data is collected to learn more about a situation, conduct an investigation, or to explore and extend our understanding. This means that when we write about the data, we must always use suggestive language.



Exercise 22:

Brainstorm ideas for more words that are suggestive versus definitive.

Suggestive words	Definitive words
Maybe	Prove