Computer Practical 3: Data Analysis in R

Dr Alice Risely [a.risely@salford.ac.uk](mailto:a.risely@salford.ac.uk)

Table of Contents

## Learning outcomes for today

What you will learn in this practical:

* How to upload a dataset into R
* How to explore a new dataset
* How to conduct the five data analysis steps to answer a scientific question
* How to write hypotheses
* How to conduct and interpret Mann-Whitney U tests and t-tests
* How to write conclusions based on evidence

## Practical 2 learning outcomes

In last week’s practical you learned:

* How to do more advanced plotting in ggplot2, including how to colour plots by a variable; add a linear trend line to a scatter plot: **+geom\_smooth()**; and add jittered points to a boxplot: **+ geom\_jitter()**.
* How to identify whether a numerical variable is normally distributed or not.
* How to conduct a Shapiro Wilk test: **shapiro.test(data$column).**
* How to interpret the output of a Shapiro Wilk test.

## Preliminary tasks

1. **If you have not completed practicals 1 and 2, you need to do these first.**

2. If you are ready to start Practical 3, **DOWNLOAD the third practical worksheet** from Blackboard, do not just view it. To download it, click the “…” next to the file and click ‘Download original file’. **Save this file in an appropriately named folder on your OneDrive (e.g., “MODULE NAME/WEEK X”) so you can access it at home.**

3. **Navigate on your web browser to** [**<https://login.rstudio.cloud/>**](https://login.rstudio.cloud/)**.** Assuming you have already registered for practical 1 - 3, then sign in. If prompted, click “Posit Cloud”.

4. **You should already have an existing R studio project from the first practical called ‘R workshop’,** click on this to enter your saved workspace.

5. **Start a new R script for this practical.** Once your project has loaded (it takes a few seconds), click ‘File’ then ‘New file’, then ‘R Script’.

6. **Press the Save button and save your script as “Practical\_3.R”.** This will save all your code you write today.

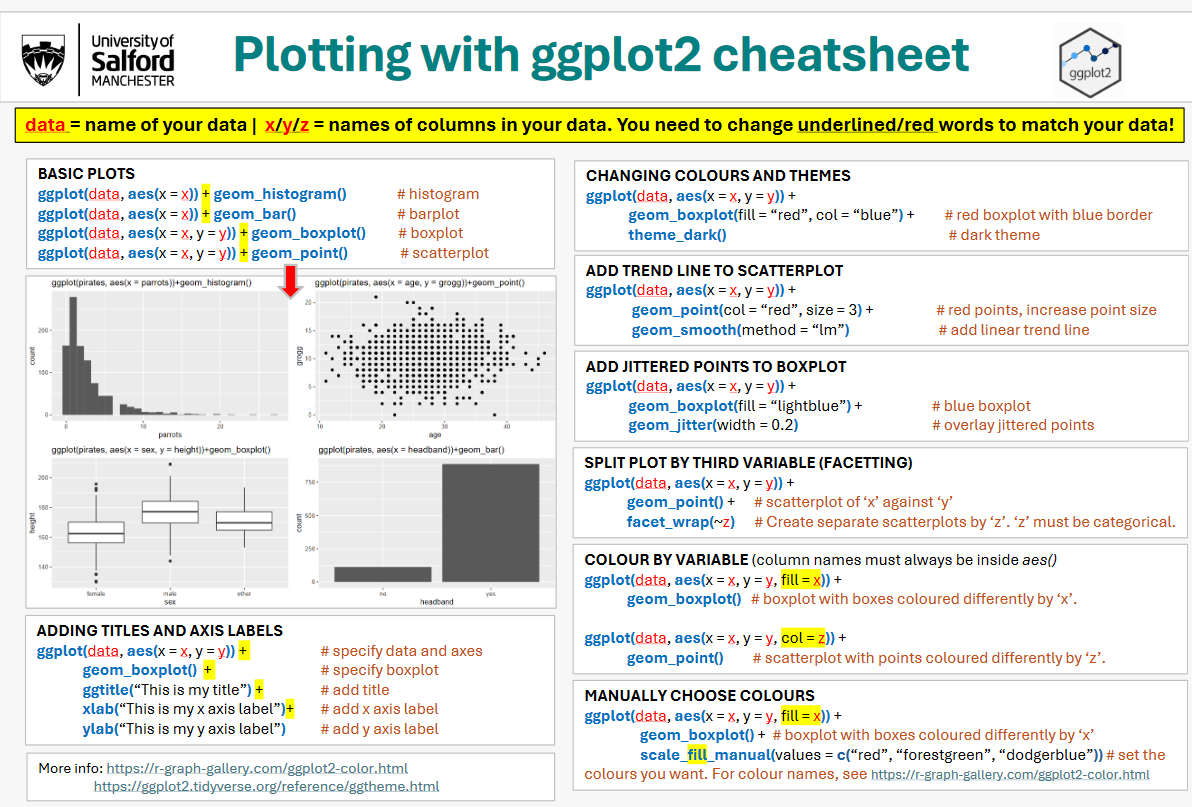
## Recap: How to run code in R

R considers each \*numbered line\* as a stand-alone line of code. To run a line of code in R, put your mouse cursor on the line you want to run and press \*\*Control and Enter\*\* (this is the most common method). Alternatively, press the ‘Run’ button above the code editor.

**Anything after a # key is NOT code**, it is comments/notes. R will ignore any writing after the # key. You do not need to copy out the comments into R.

## Recap: Plotting in ggplot2

In practicals 1 and 2, we focused mainly on learning how to visualise data using the R package *ggplot2*. You need to be familiar with the following code:



**NOTE:** There are a lot more things you can do with ggplot2 that we haven’t had time to explore. More examples of all the kinds of plots you can make with ggplot2, and the code to make them, can be found at <https://r-graph-gallery.com/>. If you click on the plot you want, it will provide the code to make it.

## Uploading a new data set to R

We will now analyse a new biological data set that is based on data collected from tigers in zoos.

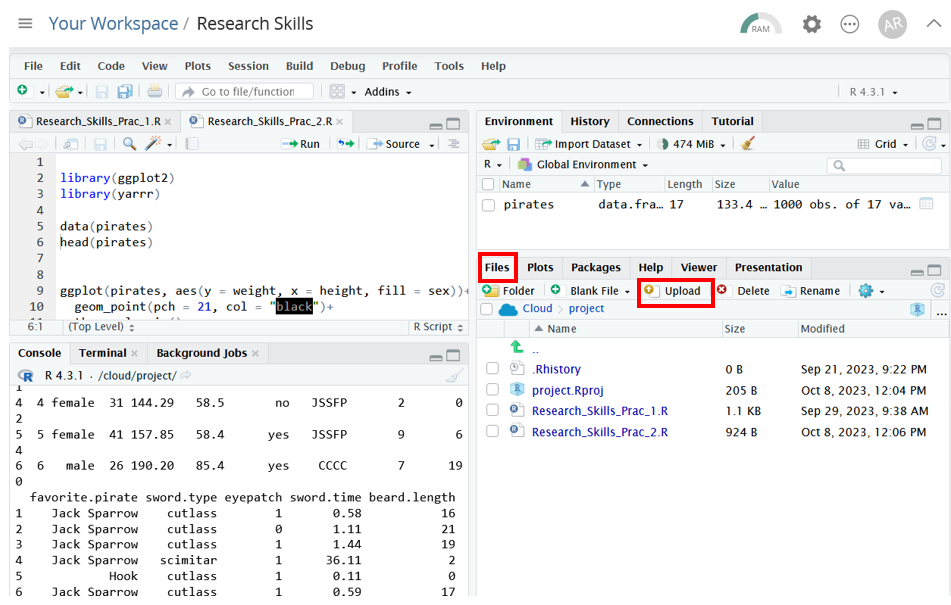


This dataset includes data from a study that looked at the effect on antibiotic treatment and diet treatment on captive tiger weight and parasite load across zoos in the United States. Parasite load was measured as the number of parasite eggs found in tiger faecal samples.

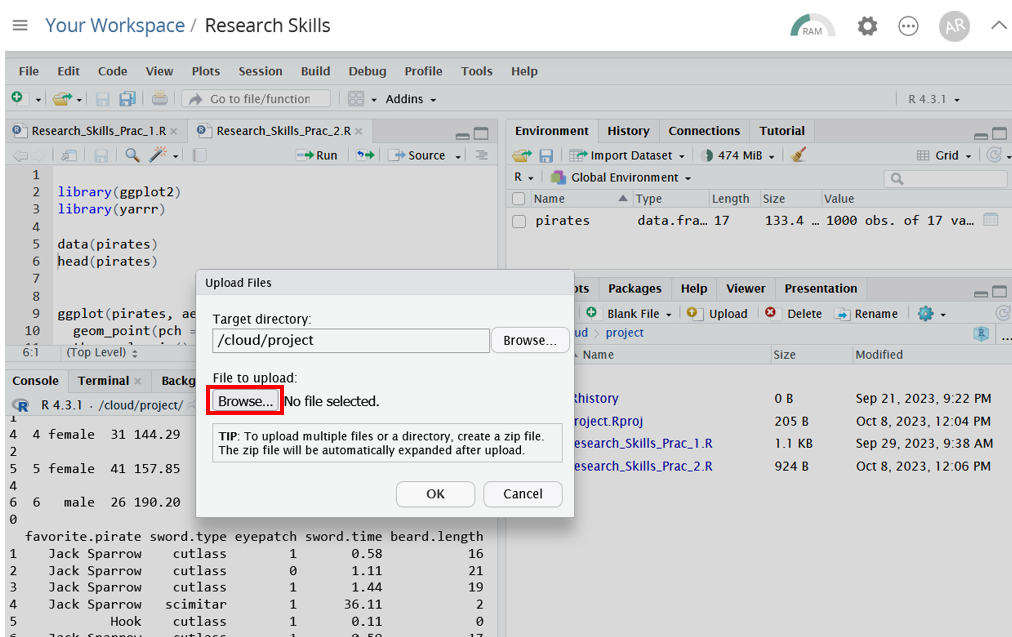
The scientists also collected data on tiger age and sex.

The new data set (*tiger\_data.csv*) is on Blackboard alongside the practical 3 tutorial. Instructions for how to load it into your R studio cloud project are below. Note, pay attention as you will need to do this for your assignment data! *If you are using R studio software, you can skip steps 2 and 3.*

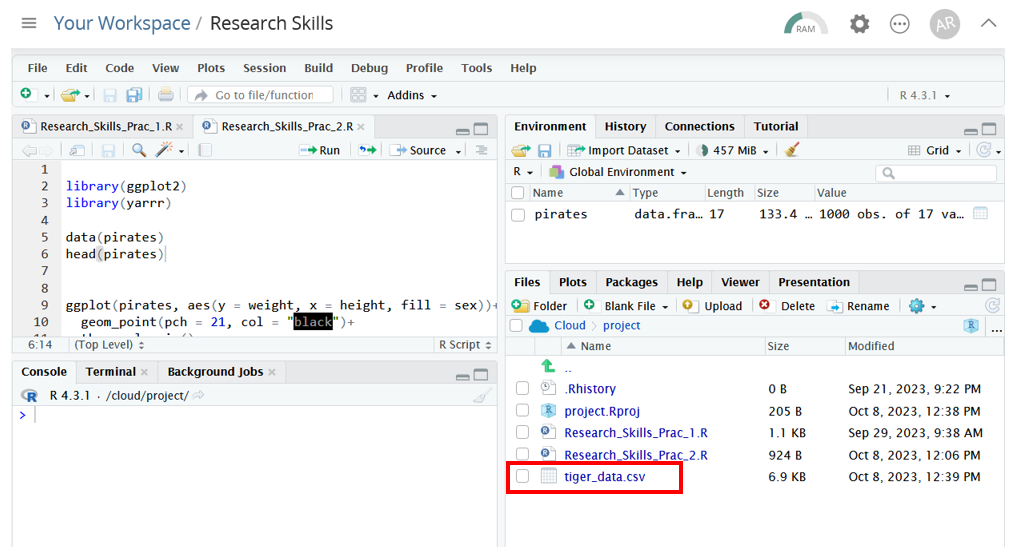
1. Download the data file from Blackboard to your OneDrive folder for this session.
2. From within R studio cloud, in your bottom right window navigate to the File tab and click ‘Upload’



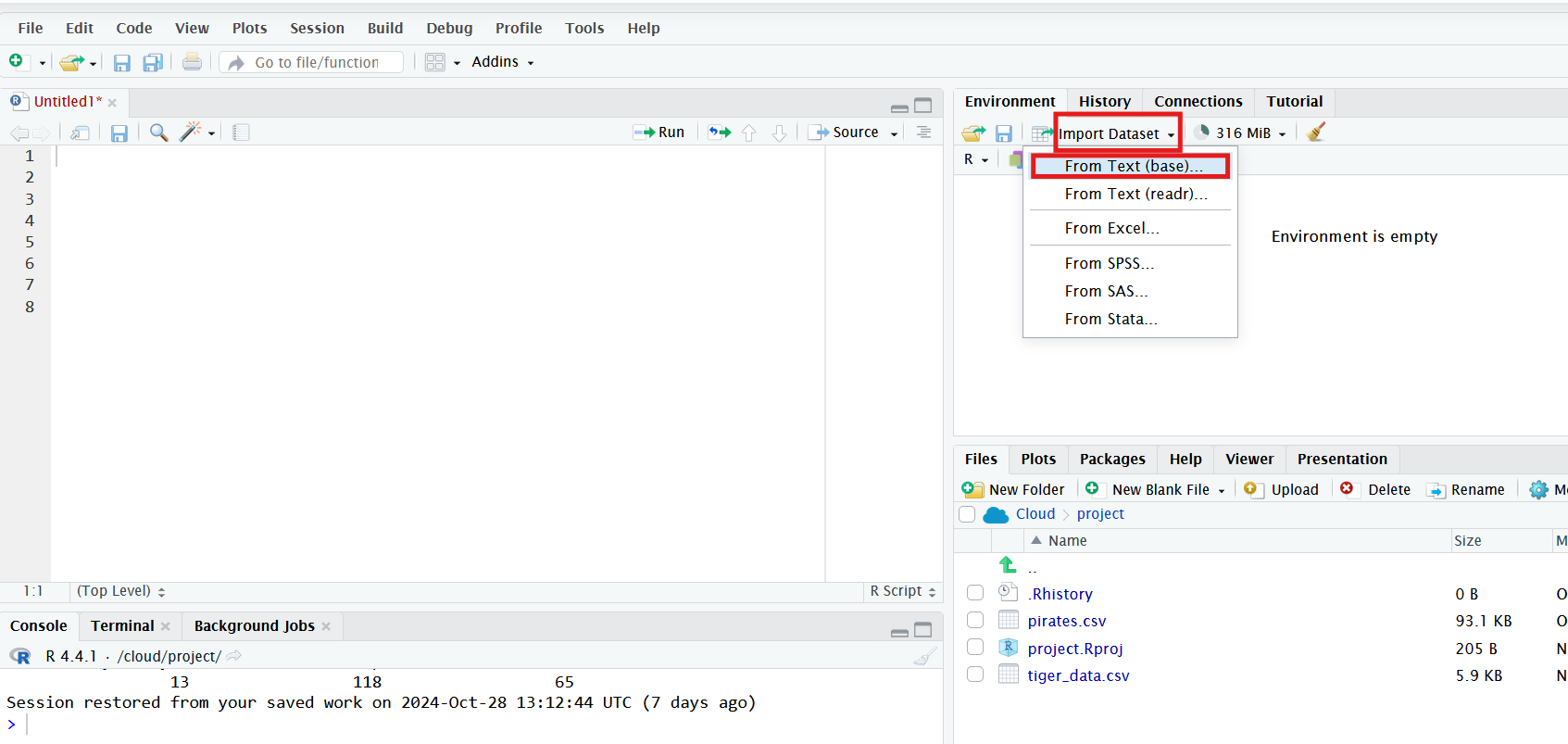
1. Under ‘file to upload’, navigate to the folder where you saved the data set



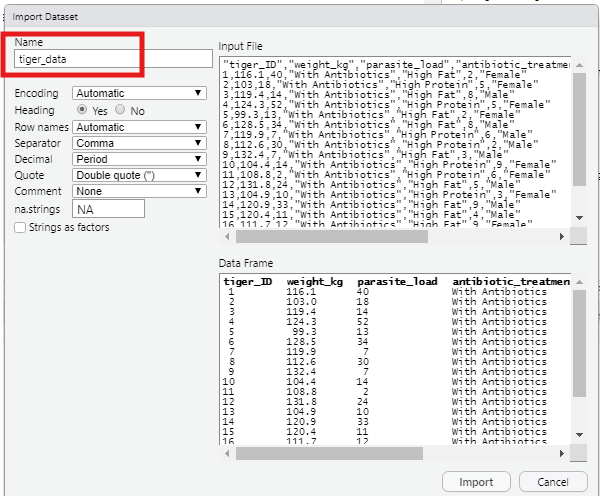
You should now see your data set tiger\_data.csv in your working directory:



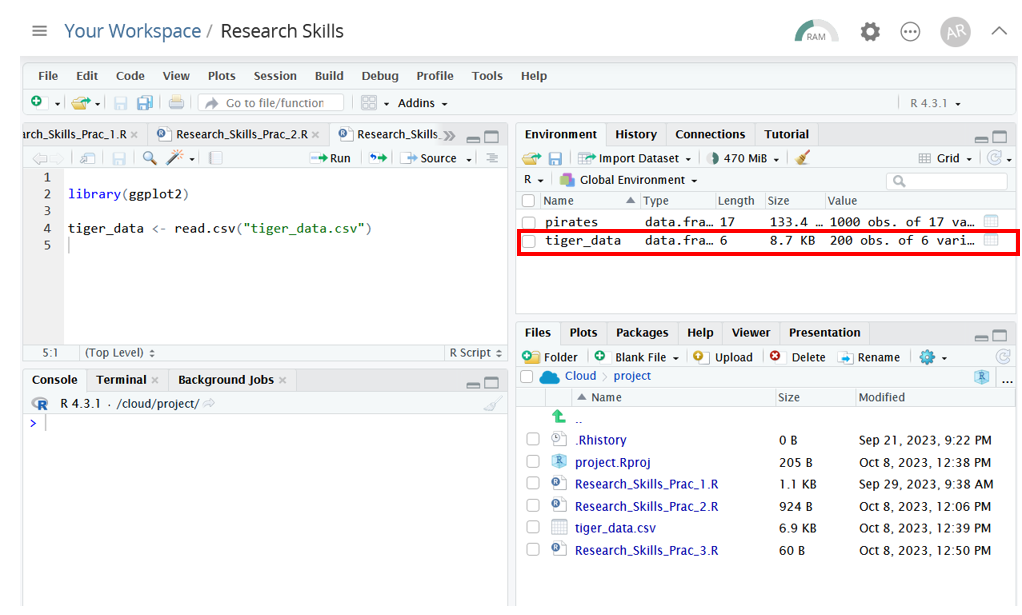
1. Lastly you need to load the new dataset into your working environment, and call it ‘*tiger\_data*’. **You can do this by pressing the ‘Import dataset’ button above Environment space, then ’From Text (base).**

* 

1. **Open your tiger\_data.csv file, make sure it is being imported as ‘tiger\_data’ (top left box in pop-up window), and press ‘Import’:**



You should now have have your data object tiger\_data in your working environment.



1. Copy the import code from your console (bottom left window) into your R script, so that in the future you can just run one line of code to import the dataset again, eg:

**tiger\_data <- read.csv(“OneDrive/Research Skills/Week 7/tiger\_data.csv”)**

The pathway above should be where you saved your data on your OneDrive.

## Exploring the new tiger dataset

First lets have a look at what this dataset looks like by using the head() and View() function.

head(tiger\_data) # Look at the first six rows in the terminal  
  
View(tiger\_data) # examine the dataset. Take note of the column names  
  
dim(tiger\_data) # Examine how many rows (obervations) and columns (variables) there are  
  
  
names(tiger\_data) # What are the column/variable names?

Now lets explore the dataset and apply some of the things you’ve learned in previous practicals.

#### Exercise 1

Check if your variables (ie column names) have been automatically assigned the correct data type by R (catagorical/continuous/count). Write out how each variable is treated by R, and the correct data type of each variable.

Remember, categorical data in R is called “factor” or “character”, count data is called “integer”, and continous data is called “numeric”.

**Hint: Use the *str()* function to examine data types.**

**Hint 2: tigers are individuals, not numbers.**

\*\* WRITE YOUR ANSWERS HERE \*\*  
  
a) What data types each variable is treated as in R:  
  
Tiger ID = ?  
Weight = ?  
Antibiotic treatment = ?  
Diet treatment = ?  
Parasite load = ?  
Age = ?  
Sex = ?  
  
  
b) What data types each variable should be:  
  
Tiger ID = ?  
Weight = ?  
Antibiotic treatment = ?  
Diet treatment = ?  
Parasite load = ?  
Age = ?  
Sex = ?  
  
---

#### Exercise 2

Use the function *as.factor()* to change tiger ID into categorical data, so that R does not treat is as numerical or count data.

\*\* COPY YOUR CODE HERE \*\*  
  
  
---

#### Exercise 3

In practicals 1 and 2 you learned how to summarise categorical and numerical data using **summary()** and **table()**. Use these functions to answer the following questions:

1. How many tigers were given antibiotics, and how many tigers were not given antibiotics?
2. Given the above, what is your total sample size?
3. How many tigers were given the high fat diet, and how many were given the high protein diet?
4. What is the mean parasite load across tigers?
5. What is the mean weight across tigers?

* \*\* WRITE ANSWERS HERE \*\*

## Introduction to statistics

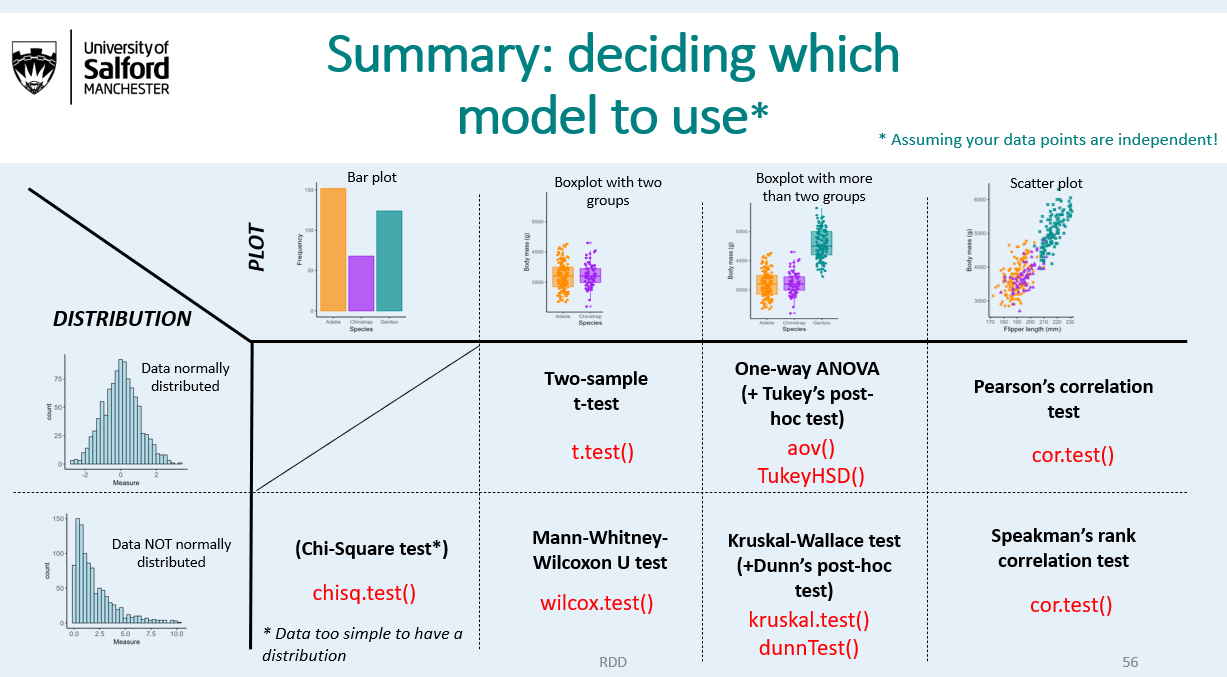
We will now begin to statistically analyse this new dataset. Statistics is the process of fitting statistical models to you data in order to generate statistical evidence for or against your hypotheses. Unfortunately, there is no one model that fits all data. The model needs to be tailored to the kind of data you have.

The type of statistical model you fit depends on three things:

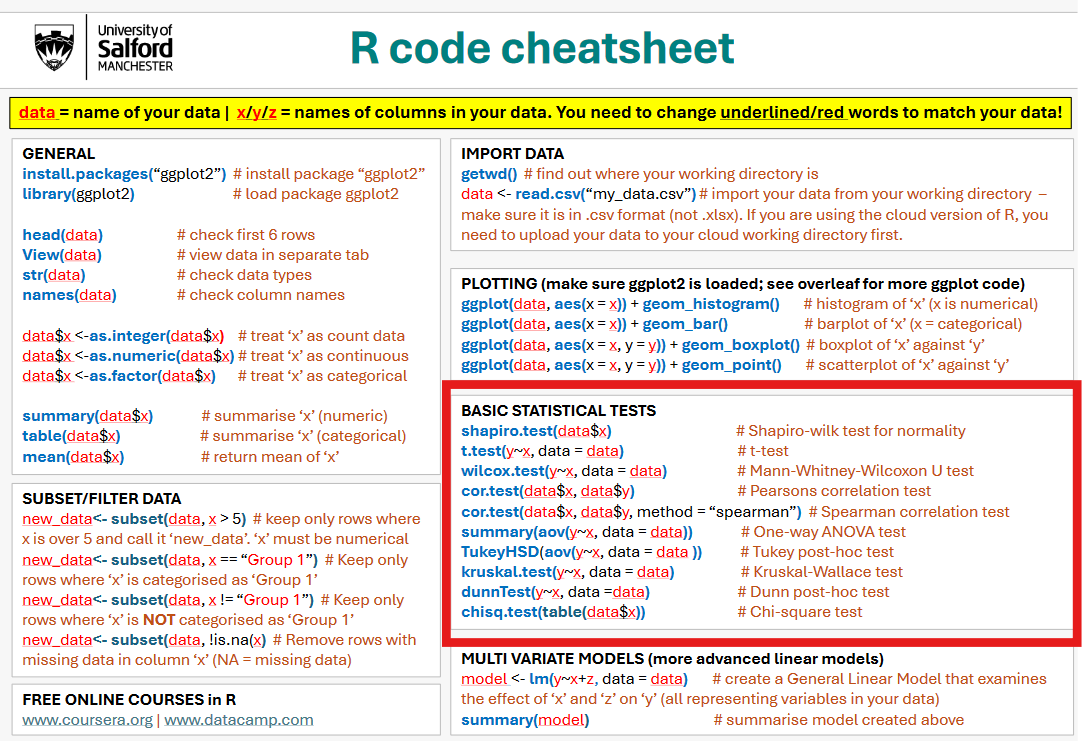
1. The type of data. This can be visualized by what **kind of plot** you can generate.
2. Whether your numerical variable of interest is **normally distributed** or not.
3. Whether your data points are **independent** from each other or not. **For now, though, we will ignore this point and assume all our data points are independent.**

Below is a summary of the names of models you can use for what kind of data, and the R function needed to perform the test.

Over the next practicals, we will focus on learning how to fit these models to our data, how to interpret the outputs, and how to write out our conclusions.



In addition, the code for these statistical tests is also on your cheatsheet handout:



## The five data analysis steps

For every analysis you do, you need to go through five steps:

1. *Write out your null and alternative hypotheses using IF/THEN statements.*
2. *Plot the relationship between your variables of interest.*
3. *Check if your numerical data has a normal (bell shaped) distribution.*
4. *Conduct the appropriate statistical test.*
5. *Write out your results using evidence generated from the statistical model.*

We will now go through these steps to choose what model we need when presented with a scientific question.

## Question 1: Does antibiotic treatment affect parasite load?

For this question, I will guide you through the 5 analysis steps. **For the next question, I will expect you to complete the steps yourself. You will also need to go through these steps for your assignment.**

Background reasoning: Antibiotics are used to kill bacteria, and also sometimes kill parasites too. Therefore, they are sometimes used as an anti-parasitic treatment. You therefore think that antibiotic treatment should be linked to parasite load in tigers.

### **Step 1: Hypotheses**

Write down our null and alternative hypotheses. Remember, hypotheses should provide a biological explanation of your predictions:

*NULL: IF [explanation or mechanism not true], THEN [what pattern you expect in your dataset]*

*ALTERNATIVE: IF [explanation or mechanisms is true], THEN [what pattern you expect in your dataset]*

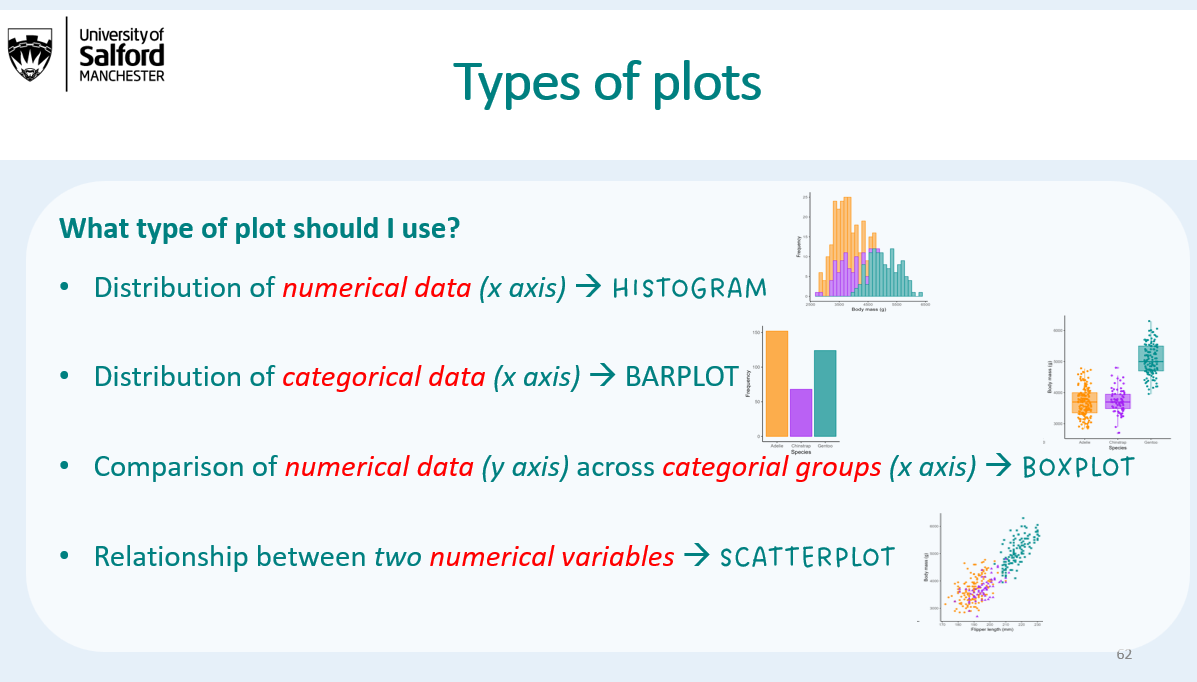
Lets have a go:

**NULL:** **If** antibiotics do not kill parasites, **then** there will be no difference in parasite load between tigers that were treated with antibiotics and those that weren’t.

**ALTERNATIVE: If** antibiotics kills parasites, **then** tigers that were treated with antibiotics will have fewer parasites, on average, than those that no antibiotic treatment.

### **Step 2: Plot relationship**

Step 2 is to plot the relationship you are interested in testing. You need to work out what kind of plot you need using the following guide:



#### Exercise 4

Plot the relationship between antibiotic treatment and parasite load.

Remember, you always need to load the ggplot2 package before plotting using *library()*, so don’t forget this line of code.

In addition, the variable that you are interested in predicting, in this case parasite load, should be on the y axis.

\*\* COPY YOUR PLOT HERE \*\*  
  
  
---

We can see from our plot that tigers that were treated with antibiotics have lower parasite load than those that were not treated with antibiotics.

Therefore, there is already some indication that the evidence supports our alternative hypothesis. BUT, we need to quantify this difference with a statistical model.

### **Step 3: Test for normality**

Step 3 involves checking if our numerical variable of interest has a normal distribution.

Our variable of interest is parasite load, which is numeric data. We need to check if it is normally distributed to know what statistical model we can use.

We can do this by:

*1) Plotting a histogram of our numerical data; and*

*2) Performing the Shapiro-Wilk test of normality on the weight data.*

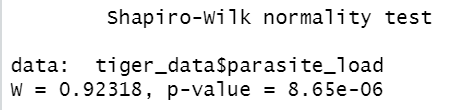
You did this in last week’s practical so you should be familiar with this step already.

#### Exercise 5

Make a histogram of parasite load and conduct a Shapiro-Wilk test. Is parasite load normally distributed?

\*\* COPY YOUR PLOT AND ANSWERS HERE \*\*  
  
  
---

You should get the following output from the shapiro-wilk test:



The p-val is 8.65e\_06, which is equal to 0.00000865 (move the decimal point 6 places to the left!), which is smaller 0.05. **This means the parasite load is NOT normally distributed.**

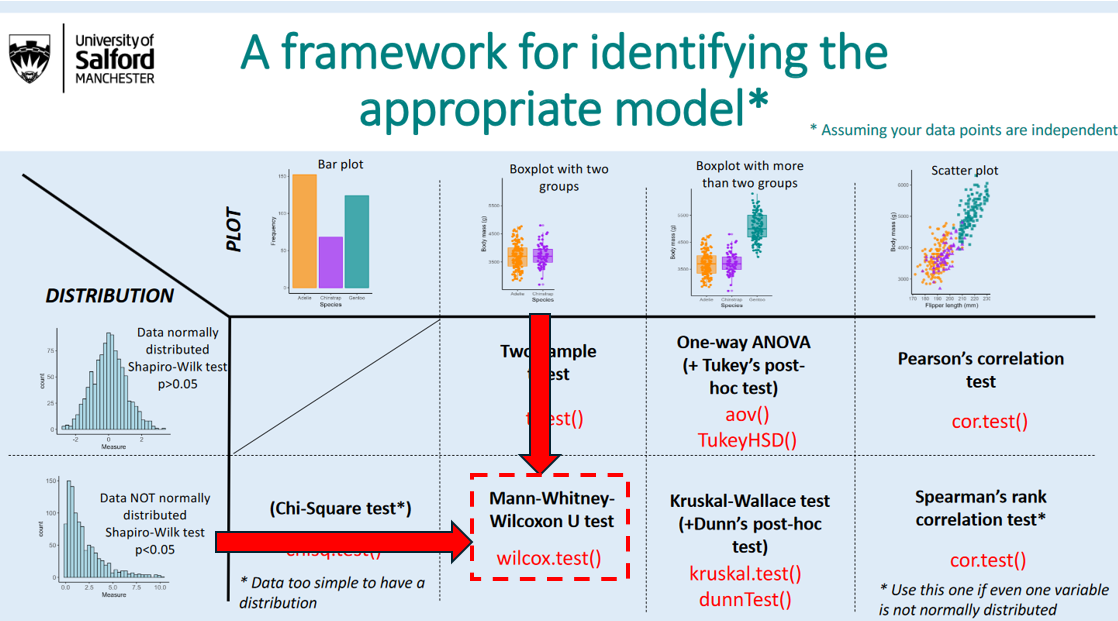
Use this table to help you interpret the results of the Shapiro-Wilk test:

|  |  |
| --- | --- |
| p-val | Conclusion |
| < 0.05 | Data is NOT normally distributed |
| > 0.05 | Data IS normally distributed |

### **Step 4: Statistical test**

Step 4 involves using the framework above to identify the correct statistical test for the data.

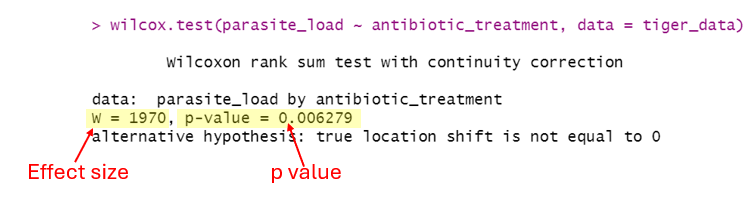
The parasite data is not normally distributed, and our boxplot compares two categories (Antibiotics and no antibiotics). We therefore need a Man–Whitney Wilcoxon U test (sometimes referred to as just “Mann-Whitney U test”, or “Wilcoxon Rank Sum test”).



Run the following code:

# conduct Mann-Whitney-Wilcoxon U test  
  
wilcox.test(parasite\_load ~ antibiotic\_treatment, data = tiger\_data)  
  
# that's it! You've fitted a statistical model.

You should get the following output with the effect size and the p value.



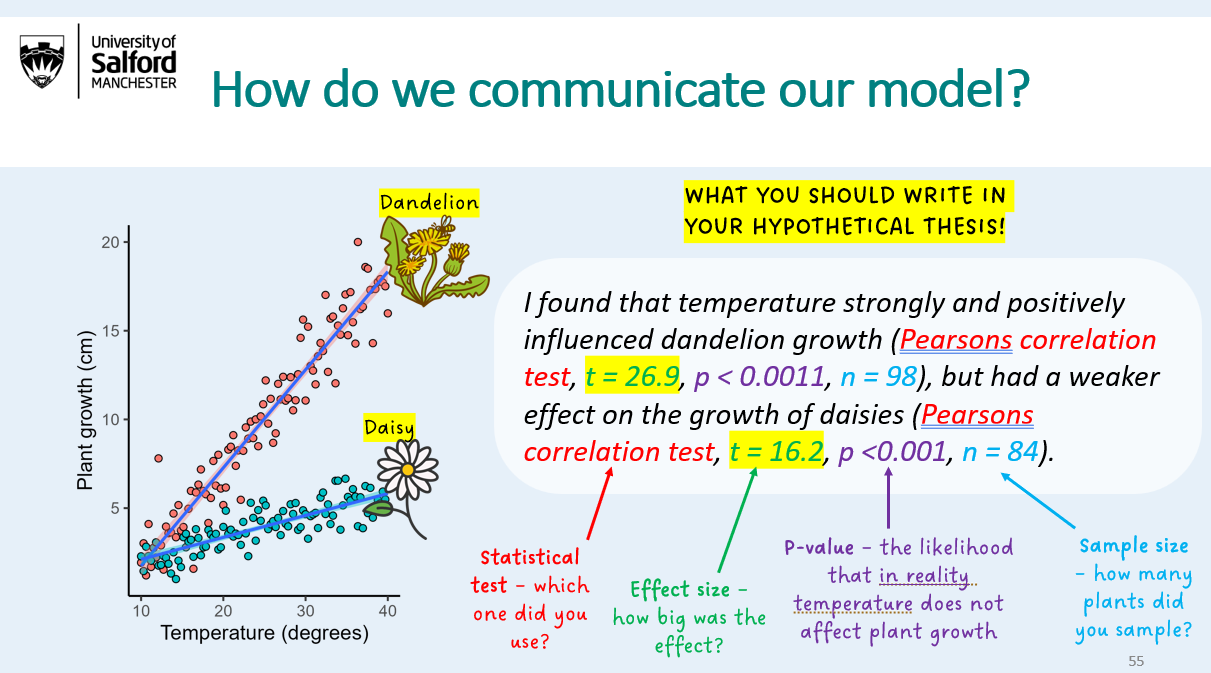
The effect size (W) is 1970. Because this is a non-parametric test, you can’t really interpret this number logically.

The P value is 0.006.

The p value is under 0.05, which means there is a SIGNIFICANT statistical association between antibiotic treatment and parasite load, so we reject our null hypothesis.

### **Step 5: Write out results**

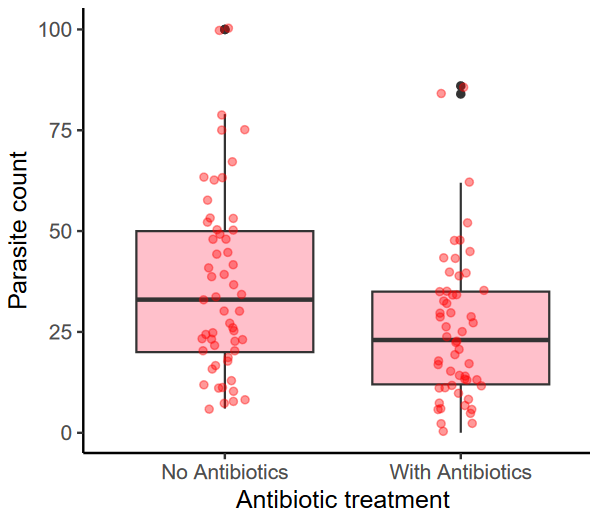
To summarise your results, you need to include the direction of the effect, and information on your sample size, effect size, and p value. Should also include a nice figure to visualise the relationship you are testing.



Note: in your results section you do not need to refer to your hypotheses. This is how you would write your results in an assignment or scientific manuscript:

*From our sample of 110 captive tigers, we found evidence that antibiotic treatment* ***significantly reduced parasite load*** *(****Mann-Whitney U test, W = 1970, p = 0.006; Fig. 1****).*

*Figure 1) Effect of antibiotic treatment on parasite load in 110 captive tigers.*



# Note, this is the code I used to genera the plot above  
  
ggplot(tiger\_data, aes(x = antibiotic\_treatment, y = parasite\_load))+   
 geom\_boxplot(fill = "pink")+   
 geom\_jitter(width = 0.1, alpha = 0.4, col = "red")+   
 theme\_classic(base\_size = 12)+   
 ylab("Parasite count")+   
 xlab("Antibiotic treatment")

## Q2: Do antibiotics affect tiger weight?

We have just tested whether antibiotics are effective at treating parasites, but do they affect tiger weight? You think that antibiotics might increase energy availability by removing parasites - which are energetically expensive for hosts - thereby allowing tigers to harvest more energy for weight gain.

Now it is your turn to go through the 5 analysis steps to answer this second question investigating the possible effect of antibiotics on tiger weight.

#### Exercise 6

Step 1: Write out your null and alternative hypotheses, making sure you use if/then statements.

\*\* Write your hypotheses below \*\*  
  
  
  
---

#### Exercise 7

Step 2: Plot the relationship between weight and antibiotic use.

\*\* Copy your plot below \*\*  
  
  
  
  
---

#### Exercise 8

Step 3: Test whether your numeric variable of interest (in this case, weight) is normally distributed. Remember, you need to generate a histogram AND do a Shapiro Wilk test.

\*\* Is weight normally distributed? What is your evidence? \*\*  
  
  
  
  
---

#### Exercise 9

Step 4: Identify and conduct the appropriate statistical test using the framework for identifying the appropriate model.

\*\* What is the appropriate test? Copy your test output below \*\*  
  
  
  
---

#### Exercise 10

Step 5: Write our your conclusions based on your plot and model below.

\*\* Write your results section below. Copy in your presentable plot that goes with the results section\*\*  
  
  
  
  
---

Note: Answers will be provided to all exercises across practicals after the last R practical, to help you prepare for your assignment.

## Optional: Advanced models

For those who want to go a step further, below I provide an example of a more sophisticated statistical analysis of the data. **However, you do NOT need this for your assignment.** You MAY need this for your final year project next year. Therefore I am including it here just in case.

**General Linear Models**

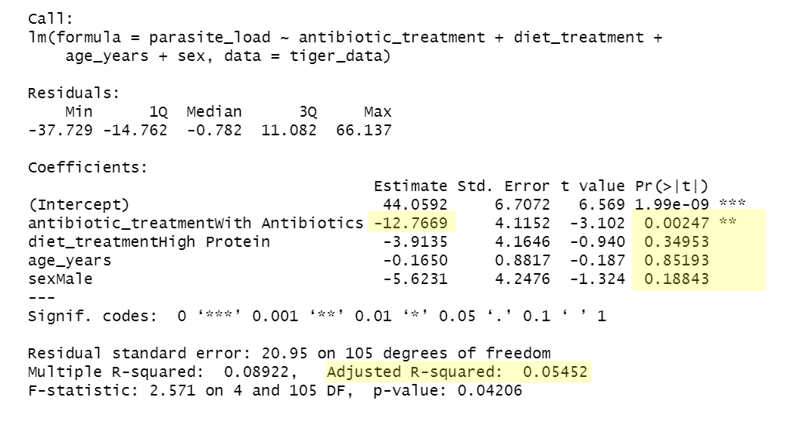
In this practical and the next practicals we have just focused on UNIVARITE analyses. This means we look at the effect of one variable on one other variable. However, in reality, lots of things are likely to affect your variable of interest.

Ideally you want to include all those variables in your model. You can do that, and the most common model to do this is called the General Linear Model.

Luckily, they are extremely easy to run. Lets look at the independent effects on antibiotic treatment, diet treatment, age and sex on parasite load.

# run model  
glm\_model<-lm(parasite\_load ~ antibiotic\_treatment + diet\_treatment + age\_years + sex, data = tiger\_data)  
  
# look at output  
summary(glm\_model)

You should get the following output:



This output tells you:

* That antibiotic treatment significantly affects parasite load (p < 0.05), but diet treatment, age, and sex and no significant effect (p > 0.05).
* On average, tigers that were given antibiotics had 12.8 fewer parasite eggs than those that didn’t not get antibiotics.
* Overall, the model explains 5.5% of variation in parasite load (this is called the Adjected R2).

GLMs have a bunch of assumptions that you need to know about if you are going to use them on your own data, but this serves as an introduction on how to use them.

#### Optional Exercise 11

Run a GLM to look at the effect of antibiotics, diet, age and sex on tiger weight.

\*\* COPY YOUR RESULTS HERE \*\*  
  
  
  
---