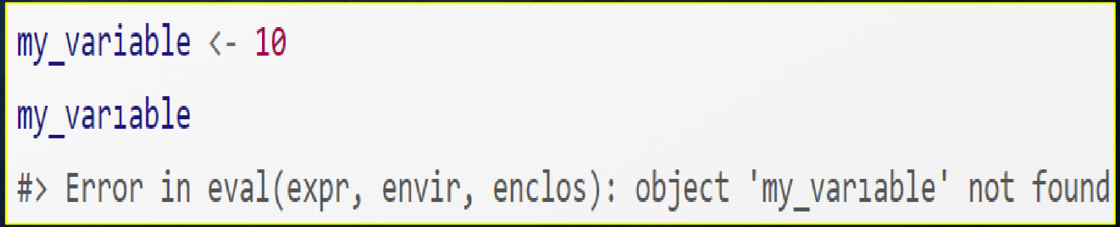
GETTING STARTED WITH R AND R STUDIO

[CEOBS, UNIVERSITY OF BRIGHTON]

Dr. Biji Shibulal

Look carefully! Training your brain to notice even the tiniest difference will pay off when programming.



```
my_variable <- 10

my_varıable

#> Error in eval(expr, envir, enclos): object 'my_varıable' not found
```

Why R

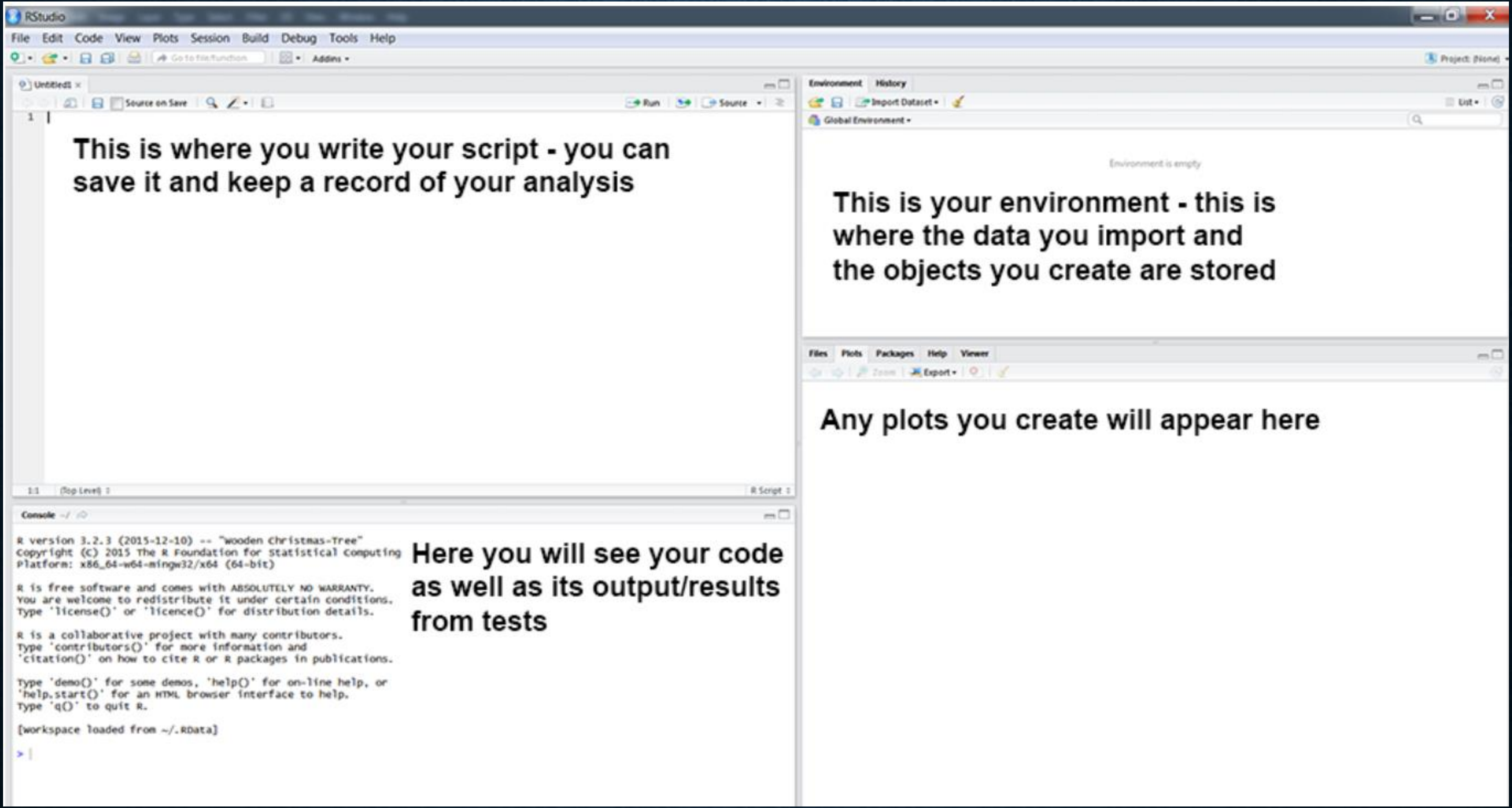
- R is free and open source

- Because R is a (statistical) programming language rather than a graphical interface, the user can easily save scripts as small text files for use in the future, or share them with collaborators.

****

- R itself does not have a graphical interface

- We will be using a program called **RStudio** as a graphical front-end to R, so that we can access our scripts and data, find help, and preview plots and outputs **all in one place.**



This is where you write your script - you can save it and keep a record of your analysis

This is your environment - this is where the data you import and the objects you create are stored

Any plots you create will appear here

Here you will see your code as well as its output/results from tests

```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

>
```

WHAT YOU HAVE TO DO

- **Type in these scripts at the script pane**, evaluate what you typed, and to look at and think critically about the output.

- **You will make mistakes and generate errors!**

  - In the process of fixing those errors, you'll learn more about how R works, and how to avoid such errors, or correct bugs in your own code in the future.

- REMEMBER to click **Ctl+S** very often

- It's useful to add comments to describe what you are doing by inserting a **hasthag (#)** in front of a line of text. R will see anything that begins with # as text instead of code, so it will not try to run it, but the text will provide valuable information about the code for whoever is reading your script (including future you!)

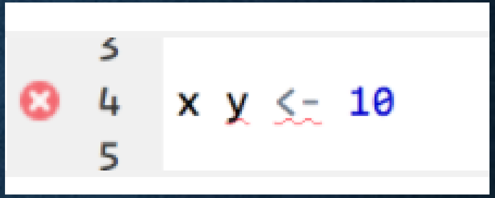
- When you exit, R will ask you if you want to **save the current workspace - NO** (The workspace will have everything you have used in a session floating around your computer memory)

- Running code

  Cmd/Ctrl + Enter

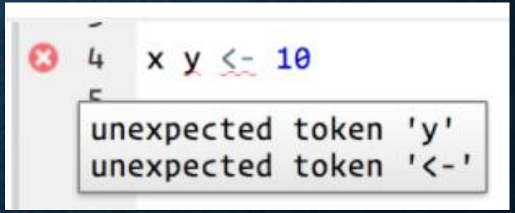
- RStudio diagnostics

✓ The script editor will also highlight syntax errors with a red squiggly line and a cross in the sidebar:



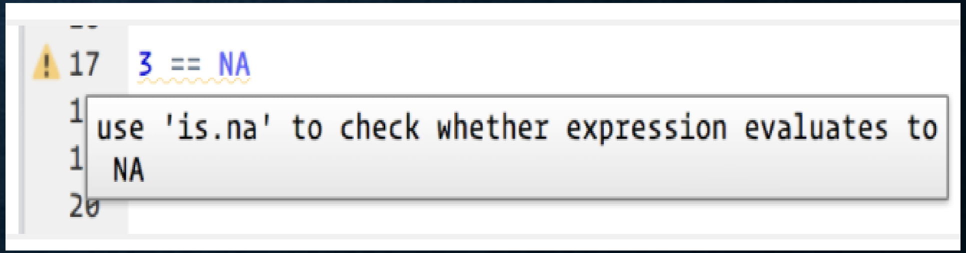
```
x y <- 10
```

✓ Hover over the cross to see what the problem is:



```
x y <- 10
unexpected token 'y'
unexpected token '<-'
```

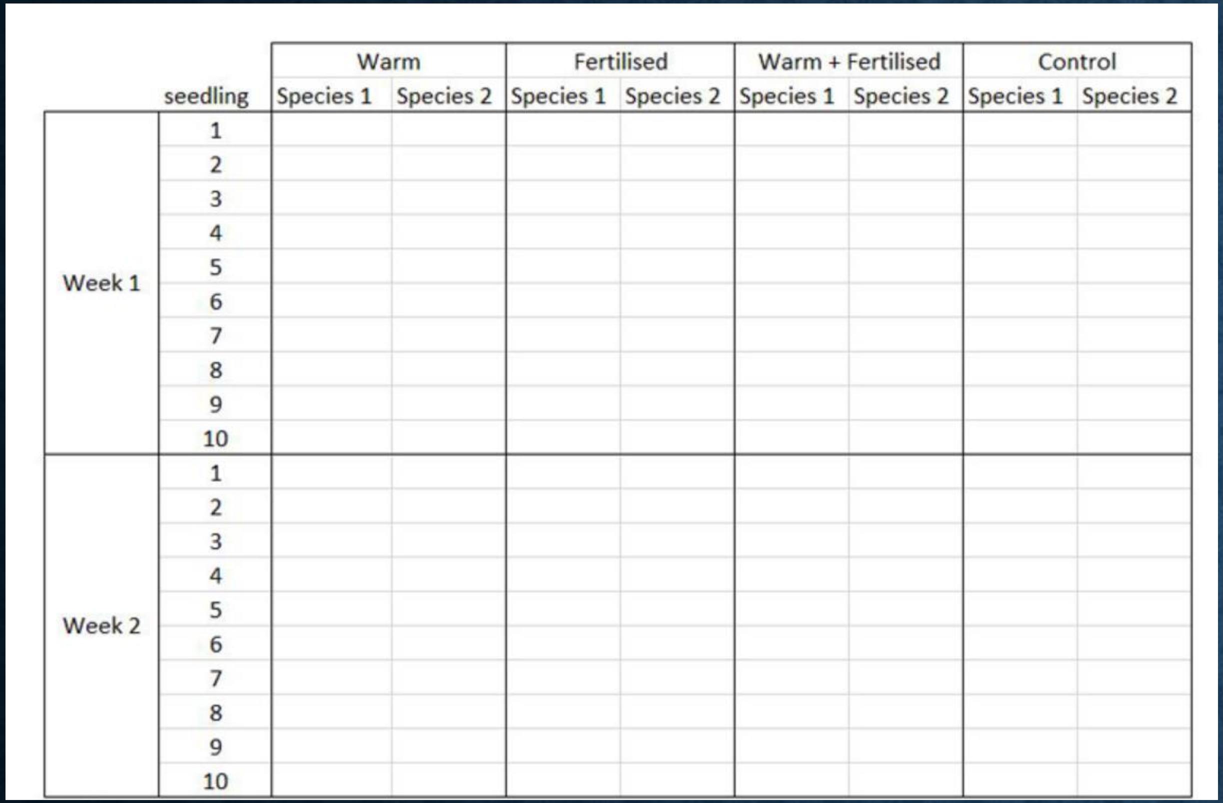
✓ RStudio will also let you know about potential problems:



```
3 == NA
use 'is.na' to check whether expression evaluates to
NA
```

Basic operators

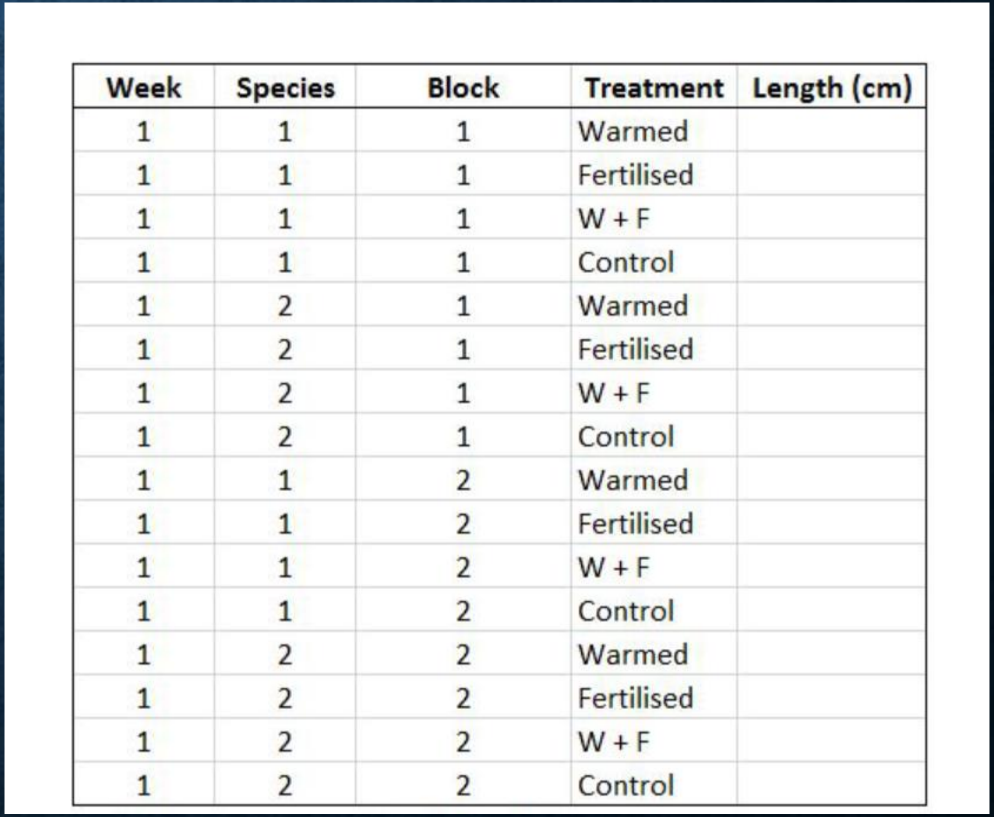
==	equals exactly
<, <=	is smaller than, is smaller than or equal to
>, >=	is bigger than, is bigger than or equal to
!=	not equal to
%in%	belongs to one of the following (usually followed by a vector of possible values)



		Warm		Fertilised		Warm + Fertilised		Control	
	seedling	Species 1	Species 2	Species 1	Species 2	Species 1	Species 2	Species 1	Species 2
Week 1	1								
	2								
	3								
	4								
	5								
	6								
	7								
	8								
	9								
	10								
Week 2	1								
	2								
	3								
	4								
	5								
	6								
	7								
	8								
	9								
	10								

You might record your data like this in field/wet lab

R needs the data to be in this format- **row** represents an **observation** and each **column** represents a **variable**



Week	Species	Block	Treatment	Length (cm)
1	1	1	Warmed	
1	1	1	Fertilised	
1	1	1	W + F	
1	1	1	Control	
1	2	1	Warmed	
1	2	1	Fertilised	
1	2	1	W + F	
1	2	1	Control	
1	1	2	Warmed	
1	1	2	Fertilised	
1	1	2	W + F	
1	1	2	Control	
1	2	2	Warmed	
1	2	2	Fertilised	
1	2	2	W + F	
1	2	2	Control	

Missing values in your dataset

- The first obstacle in any data analysis - it's important to master the methods to overcome them

- In most statistical methods the default option is listwise deletion- leads to information loss

- Some incredible R packages for missing values imputation.

- 5 R packages popularly known for missing value imputation

1. MICE
2. Amelia
3. missForest
4. Hmisc
5. mi

MICE

- MICE = Multivariate Imputation via Chained Equations
- Creates multiple imputations as compared to a single imputation (such as mean) - takes care of uncertainty in missing values.
- MICE assumes that the missing data are Missing at Random (MAR), which means that the probability that a value is missing depends only on observed value and can be predicted using them. It imputes data on a variable by variable basis by specifying an imputation model per variable.

Contd…

MICE

- By default, Predictive mean matching or linear regression is used to predict continuous missing values & logistic regression is used for categorical missing values
- Once this cycle is complete, multiple datasets are generated.
- These datasets differ only in imputed missing values. Generally, it’s considered to be a good practice to build models on these datasets separately and combining their results.

END OF DAY 1

GGPLOT

- The GG in Ggplot = Grammar of Graphics
- Scripting your code for plot is like constructing a sentence from different parts that logically follow each other, which means -
  - plotting workflow will therefore be something like creating an empty plot, adding a layer with your data points, then the axis labels, and so on.

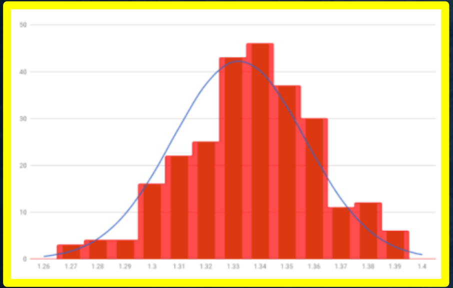
GGPLOT

- **Decide** on the right type of plot
  - A very key part of making any data visualisation is making sure that it is **appropriate to your data type** (e.g. discrete vs continuous), and **fits your purpose**, i.e. what you are trying to communicate!

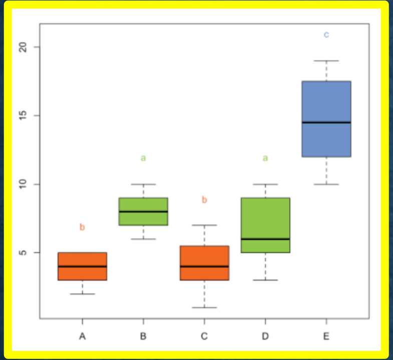
Show the distribution of your data

Compare values across groups

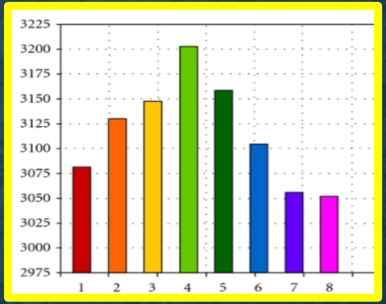
Show the relationship between variables



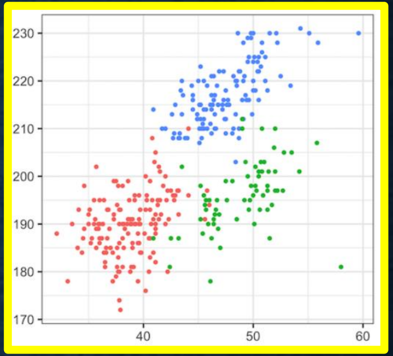
HISTOGRAM



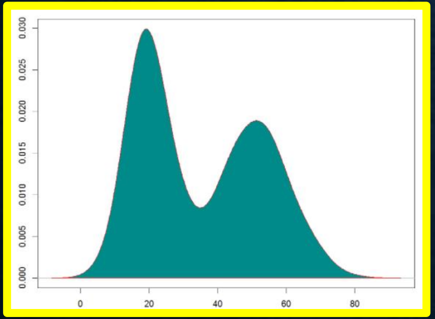
BOXPLOT



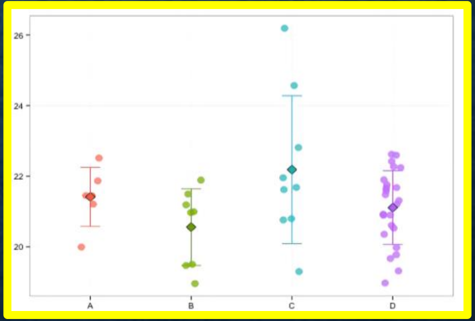
BARPLOT



SCATTERPLOT



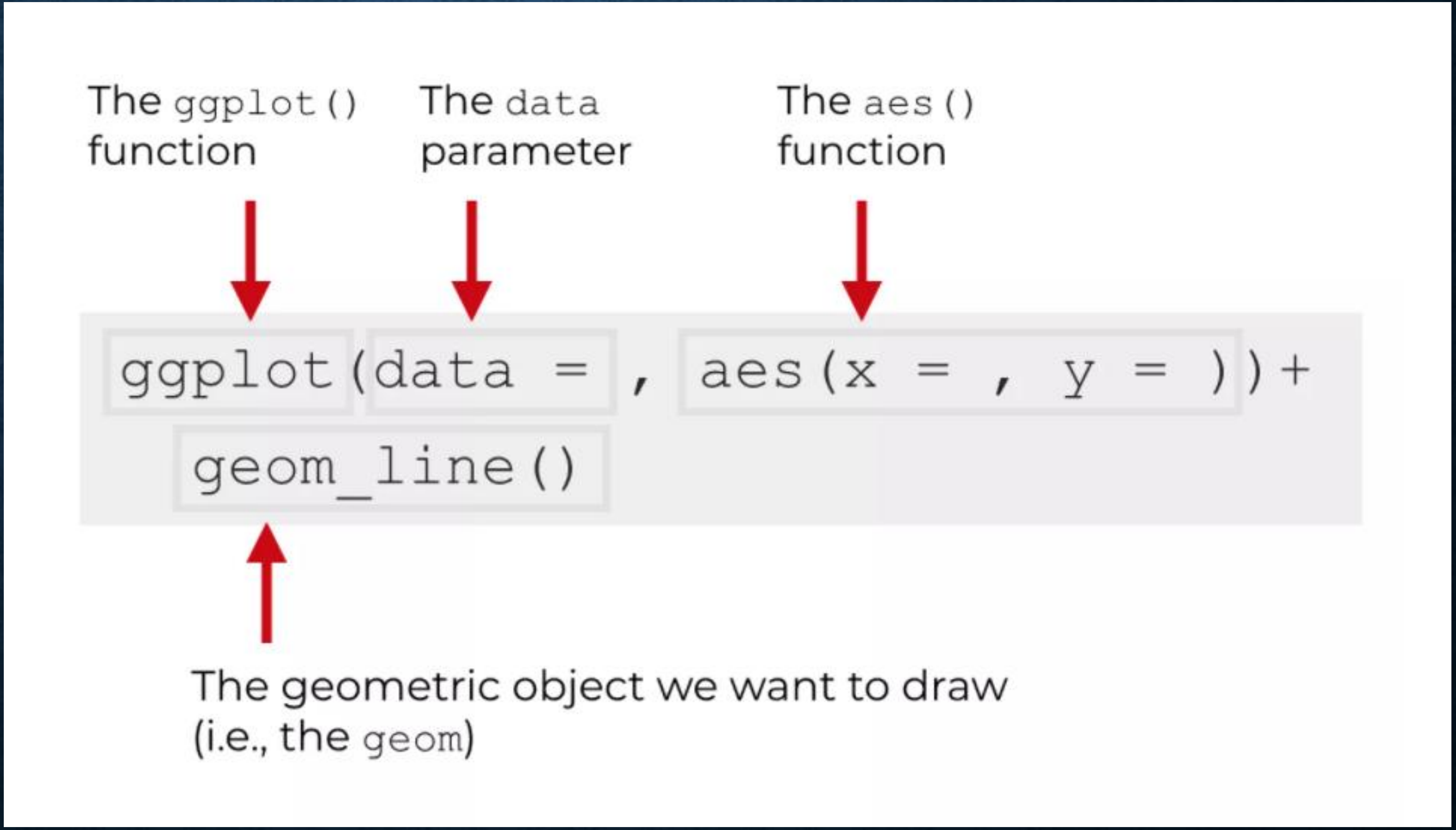
DENSITY PLOT



DOTPLOT

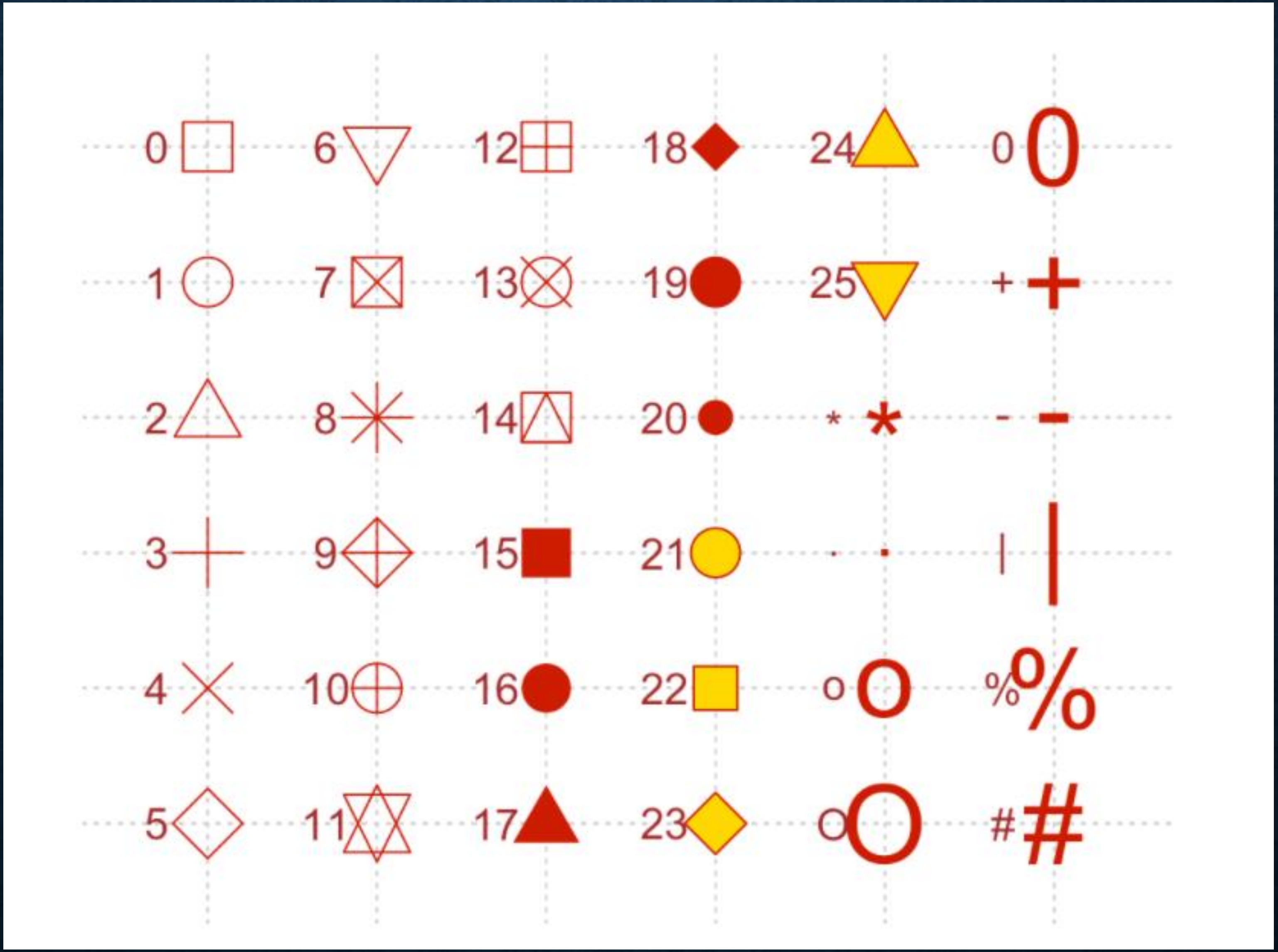
GGPLOT_BASIC SYNTAX

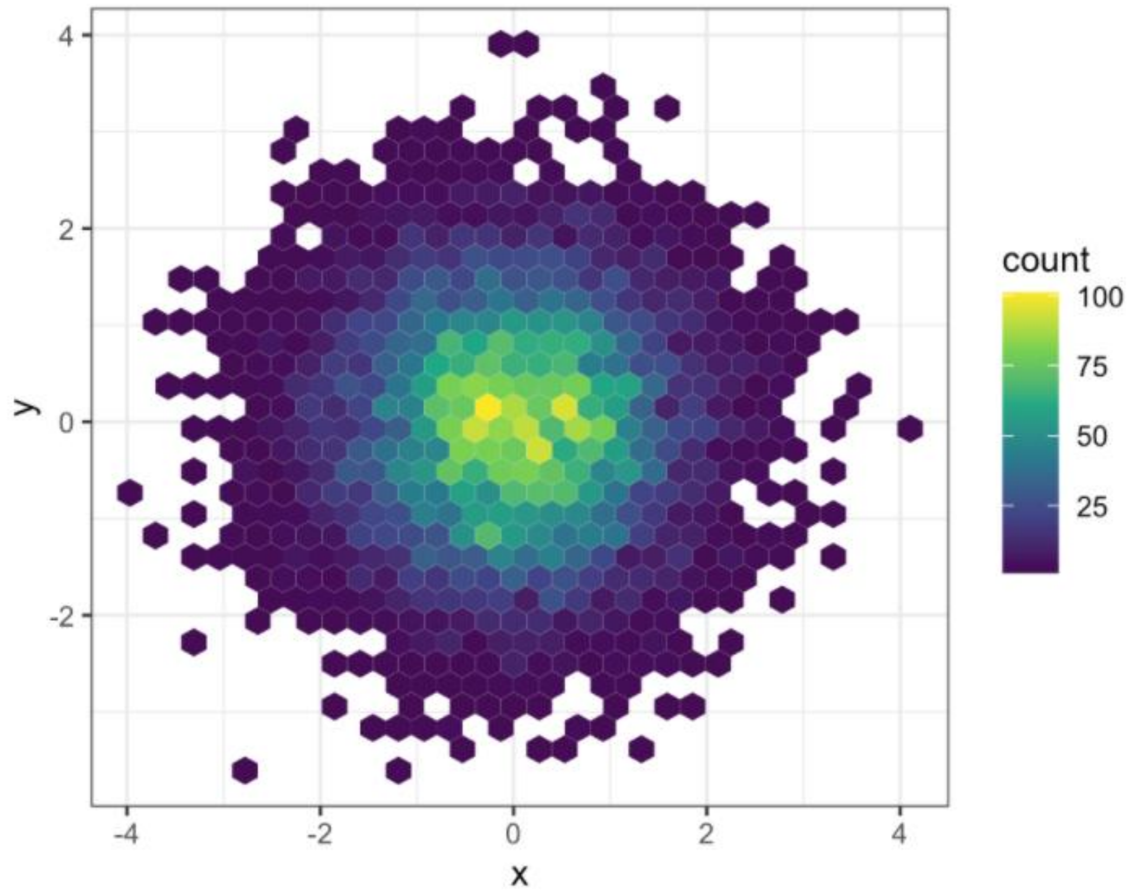
Plot = data + Aesthetics + Geometry



```
ggplot(data = , aes(x = , y = ))+
   geom_line()
```

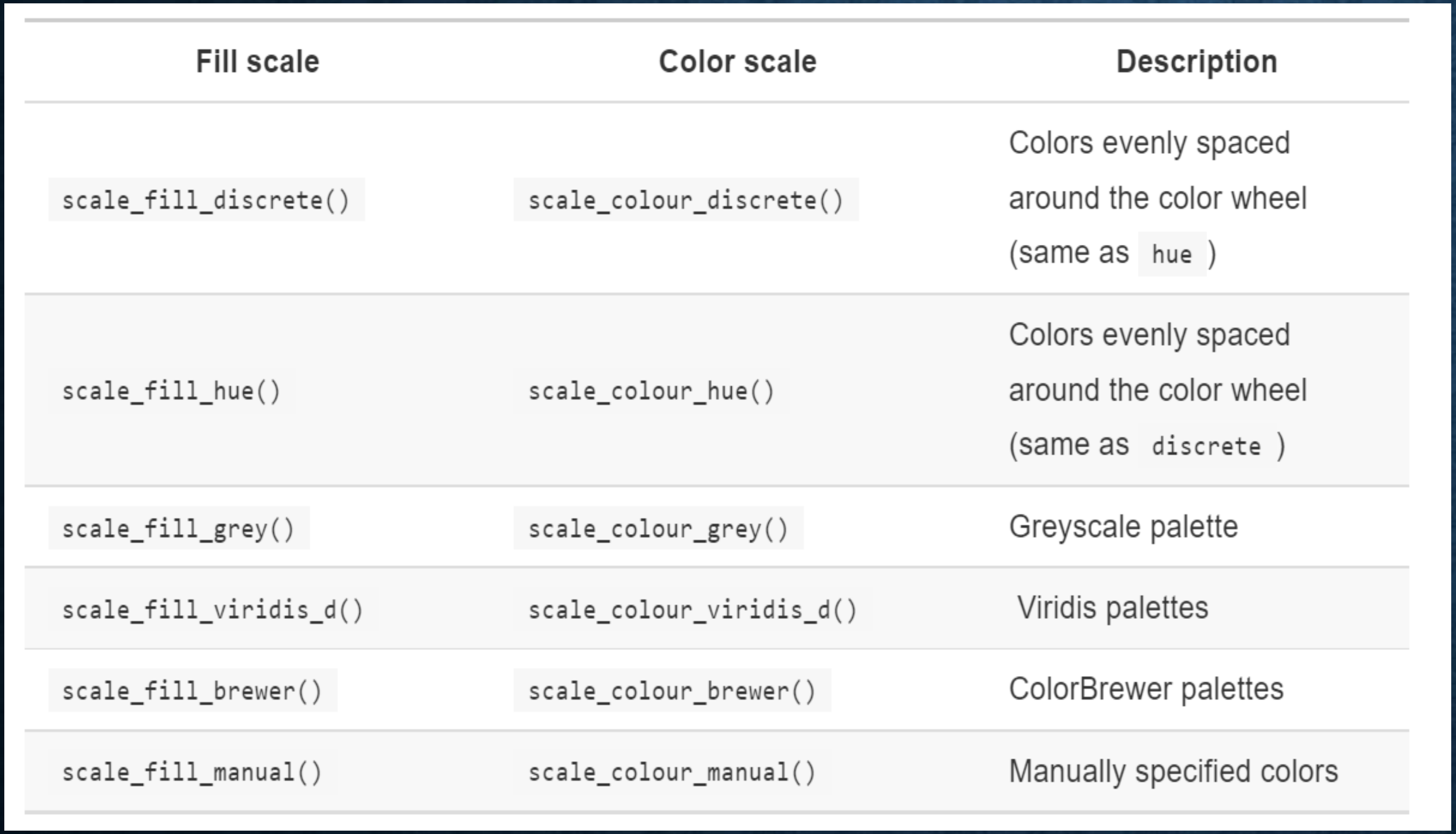
Different shapes available in R graphics





VIRIDIS COLOR PALETTE

Discrete fill and color scales



Fill scale	Color scale	Description
`scale_fill_discrete()`	`scale_colour_discrete()`	Colors evenly spaced around the color wheel (same as `hue`)
`scale_fill_hue()`	`scale_colour_hue()`	Colors evenly spaced around the color wheel (same as `discrete`)
`scale_fill_grey()`	`scale_colour_grey()`	Greyscale palette
`scale_fill_viridis_d()`	`scale_colour_viridis_d()`	Viridis palettes
`scale_fill_brewer()`	`scale_colour_brewer()`	ColorBrewer palettes
`scale_fill_manual()`	`scale_colour_manual()`	Manually specified colors

Same result

plot + scale_fill_discrete()
plot + scale_fill_hue()
plot + scale_color_viridis()

Error in plot.new() : figure margins too large

Script to execute:

```
par(mar=c(1, 1, 1, 1))
```

This will adjust margin parameters automatically by R

t- test

to compare the means of two groups

One-sample t-test

Two sample/ Independent sample /unpaired t- test

The basic application is:

- ***A/B Testing**: Compare two variants*
- ***Case Control Studies:** Before/after treatment*

Paired/dependant/ matched pairs t-test

For a t-test to produce valid results, the following assumptions should be met:

•**Random**: A random sample or random experiment should be used to collect data for both samples.

•**Normal**: The sampling distribution is normal or approximately normal.

Two sample t- test

The standard *Student's t-test*

- assumes that the variance of the two groups are equal

The *Welch's t-test*

- less restrictive compared to the original Student's test
- do not assume that the variance is the same in the two groups, which results in the fractional degrees of freedom

t.test()

```
One Sample t-test

data:  mice$weight

t = -8.1045, df = 9, p-value = 1.995e-05

alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 18.78346 21.49654

sample estimates:
mean of x
    20.14
```

Degree of freedom= n-1

p-value<0.05; reject H_0

The true mean is bet. This interval with a probability of 95%

Mean of the variable, Weight

stat.test()

```
stat.test
# A tibble: 1 x 7
  .y.       group1   group2        n     statistic     df          p
* <chr>     <chr>    <chr>       <int>   <dbl>        <dbl>     <dbl>
1 weight      1      null model   10     -8.10          9     0.00002
```

The outcome variable

The compared groups in the pairwise tests. Here, we have null model (one-sample test).

The t-test value

dbl stands for double class. A double-precision floating point number. It is a data type defined to hold numeric values with decimal points

Cohen’s d test

Recall that, t-test conventional effect sizes, proposed by Cohen J. (1998), are:
0.2 (small effect),
0.5 (moderate effect) and
0.8 (large effect) (Cohen 1998).

As the effect size, d, is 2.56 you can conclude that there is a large effect.

ANOVA (analysis of variance)

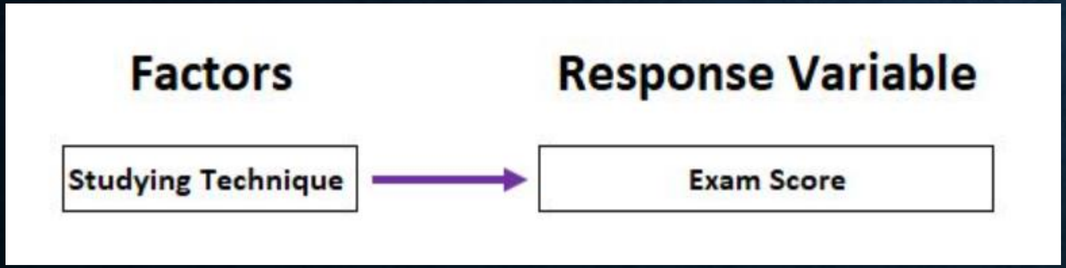
centred on the different sources of variation in a typical variable

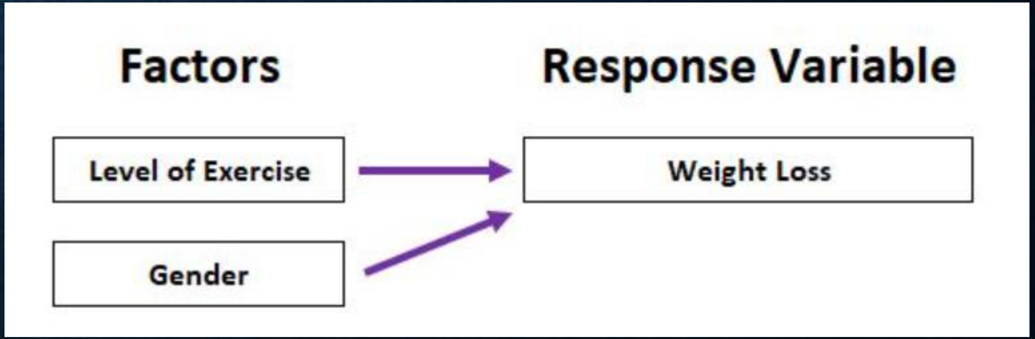
to determine whether or not there is a statistically significant difference between the means of <u>three or more groups</u>.

- Normality
- Equal variance
- Independence- randomized design

One-way ANOVA

Two-way ANOVA





t-test is type I error prone

3 groups; A B C   =   t-tests to be done are A-B, B-C, A-C

- The probability that we commit a type I error with one t-test is 1 – 0.95 = **0.05**.

- The probability that we commit a type I error with three t-tests is 1 – (0.95^3) = **0.1427**.

- An ANOVA controls for these errors so that the Type I error remains at just 5%. This allows us to be more confident that a statistically significant test result is actually meaningful and not just a result that we got from performing a lot of tests.

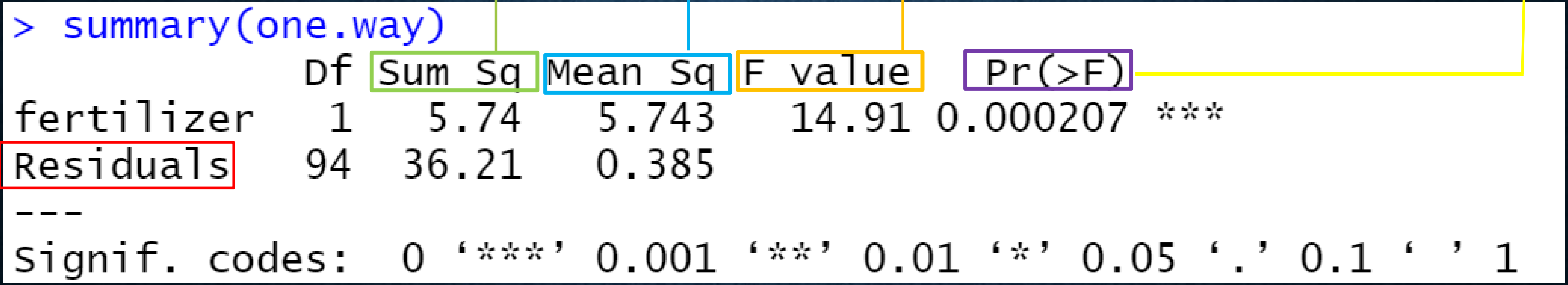
- When we want to understand whether there is a difference between the means of three or more groups, we **must** use an ANOVA so that our results are statistically valid and reliable.

The total variation = SUM SQ. between the group means and the overall mean explained by that variable. The sum of squares for the fertilizer variable is 5.74, while the sum of squares of the residuals is 36.21.

The mean of the sum of squares, calculated by dividing the sum of squares by the degrees of freedom.

The test statistics = the mean square of each independent variable divided by the mean square of the residuals. The larger the F value, the more likely it is that the variation associated with the independent variable is real and not due to chance.

P- value



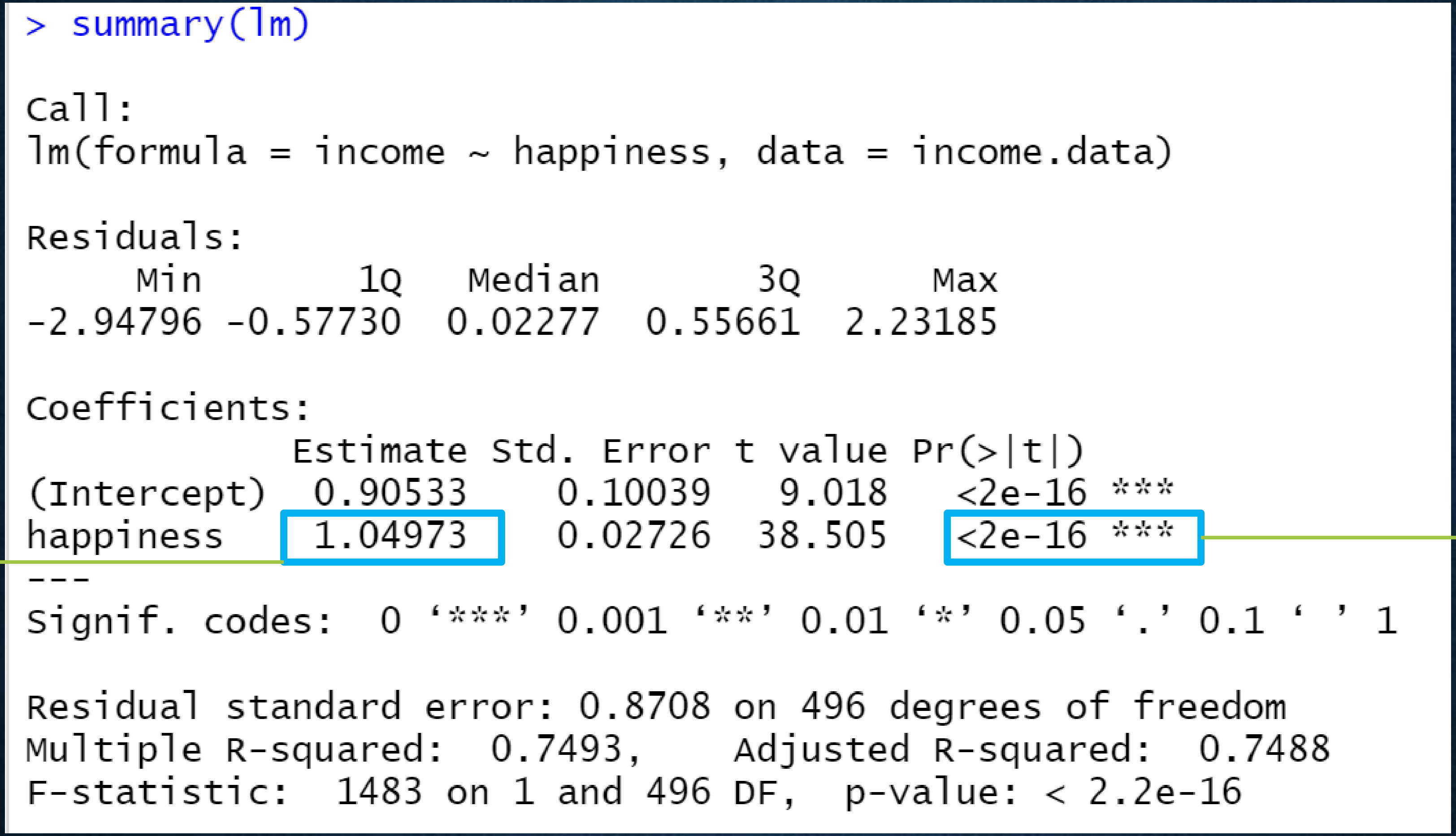
```
> summary(one.way)
            Df Sum Sq Mean Sq F value   Pr(>F)    
fertilizer   1   5.74   5.743   14.91 0.000207 ***
Residuals   94  36.21   0.385                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model error

Linear regression

- A statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).

- Check that our data meet the four main assumptions of linear regression
    - Independence of observations (or no autocorrelation)
    - Normality
    - Linearity
    - Homoscedasticity



```
> summary(lm)

Call:
lm(formula = income ~ happiness, data = income.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.94796 -0.57730  0.02277  0.55661  2.23185 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.90533    0.10039   9.018   <2e-16 ***
happiness    1.04973    0.02726  38.505   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8708 on 496 degrees of freedom
Multiple R-squared:  0.7493,	Adjusted R-squared:  0.7488 
F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

estimated effect of income on happiness

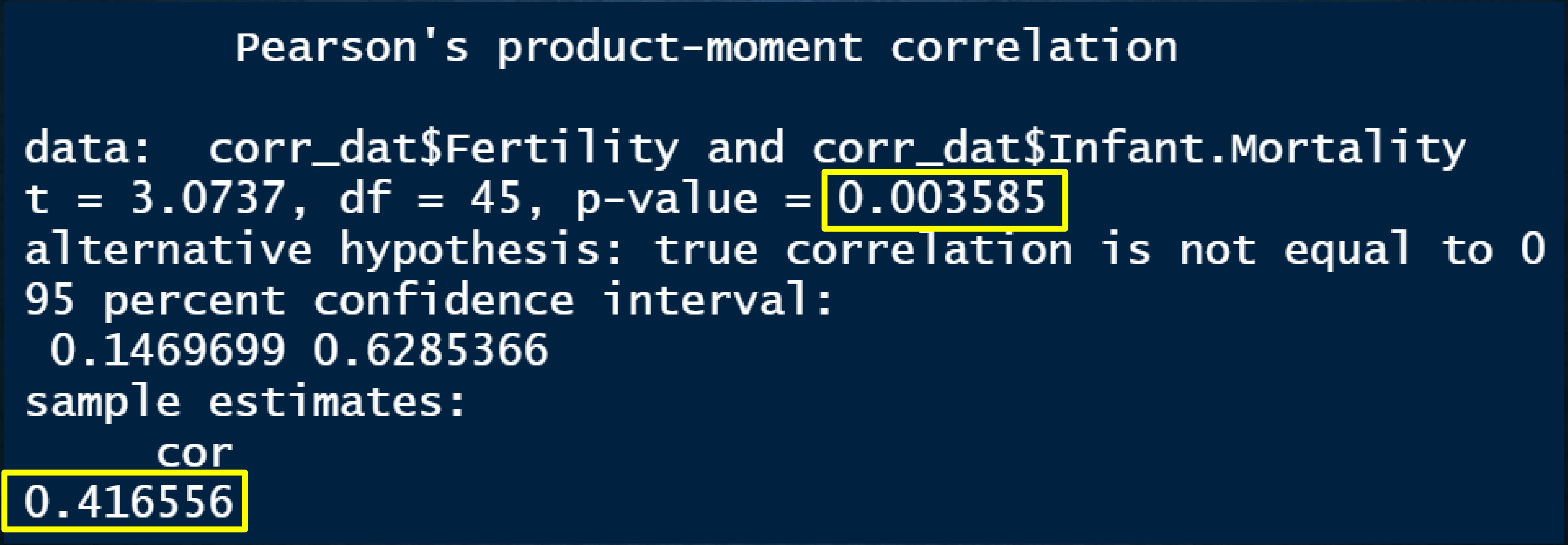
P-value

CORRELATION ANALYSIS

Test the assumptions:

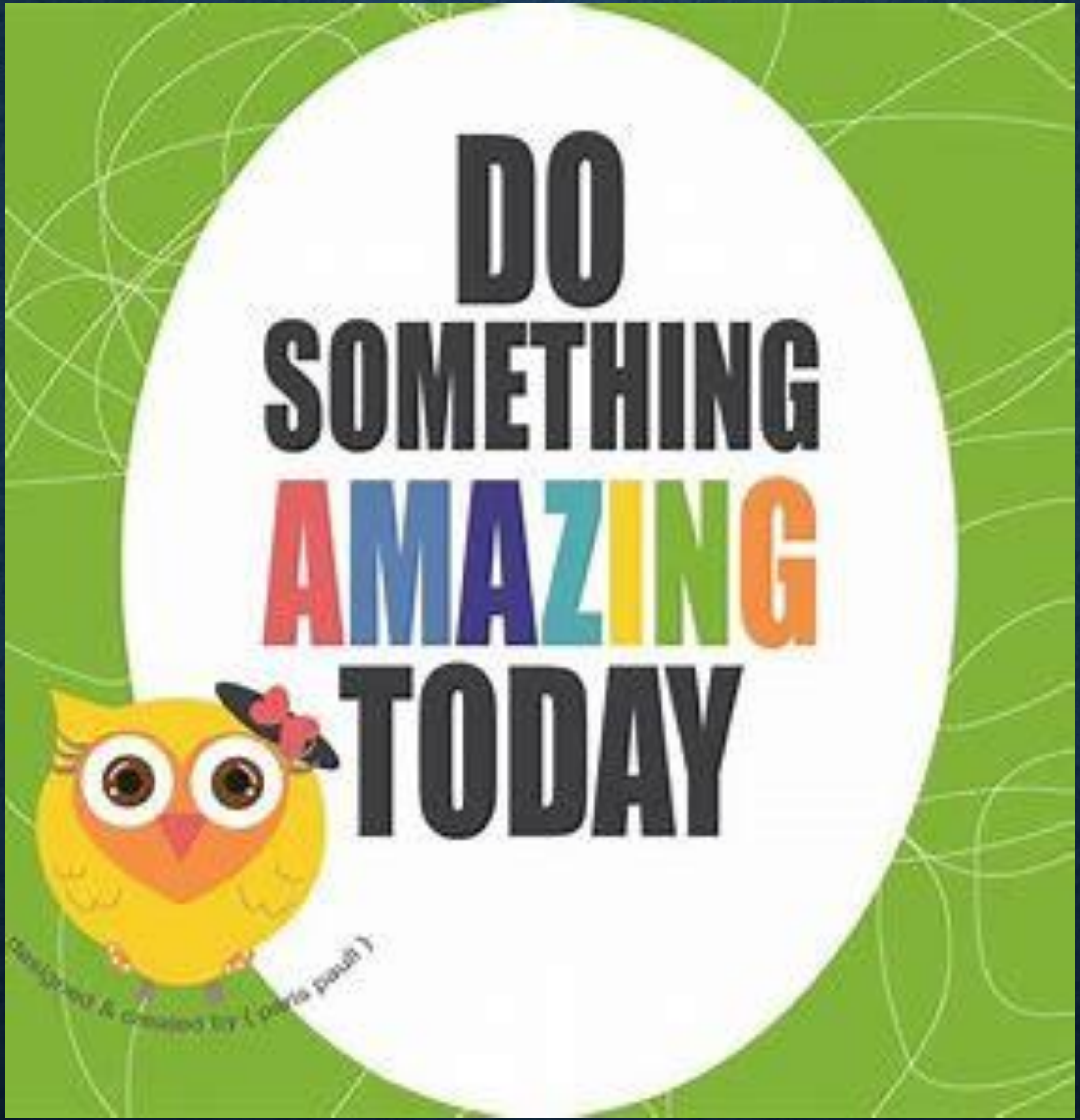
- **Linearity**: Visible from the plot itself
- **Normality**: Using Shapiro test

If the data are not normally distributed, it's recommended to use the non-parametric correlation, including Spearman and Kendall rank-based correlation tests.



```
        Pearson's product-moment correlation

data:  corr_dat$Fertility and corr_dat$Infant.Mortality
t = 3.0737, df = 45, p-value = 0.003585
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1469699 0.6285366
sample estimates:
     cor
0.416556
```


DO
SOMETHING
AMAZING
TODAY
designed & created by (gloria paull)