Contents

1 Preface

The primary purpose of this handbook is to show you how to create clear and informative plots in the R package *ggplot*. We will look at how *ggplot* is structured, how to prepare datasets for use in *ggplot* and how to customise plots. We will develop plots in R Markdown documents to aid reproducibility and re-use of plots. We will briefly look at the accessory *ggtree* package for phylogenetic trees and at some standard R stats functions.

1.1 Prerequisites

No specific knowledge prerequisites for this book but it will help if you are familiar with some common statistical tests, `t`, `ANOVA` and regression for the later parts. You will also find a knowledge of how to write computer file paths helpful.

You need to install the following stuff for this book:

1. R
2. RStudio
3. Some R packages: *ggplot2*, *ddply*, *psych*, *effsize*, *bioconductor* and *ggtree*
4. You will need to download these files and save them to somewhere on your computer: example_ros_data_flg22.xlsx, pinf_mtDNA.newick

1.2 Installing R

Follow this link and install the right version for your operating system https://www.stats.bris.ac.uk/R/

1.3 Installing RStudio

Follow this link and install the right version for your operating system https://www.rstudio.com/products/rstudio/download/

1.4 Installing R packages in RStudio.

1.4.1 Standard packages

For *ggplot2*, *ddply*, *psych* and *effsize* start RStudio and use the `Packages` tab in lower right panel. Click the install button (top left of the panel) and enter the package name, then click install as in this picture

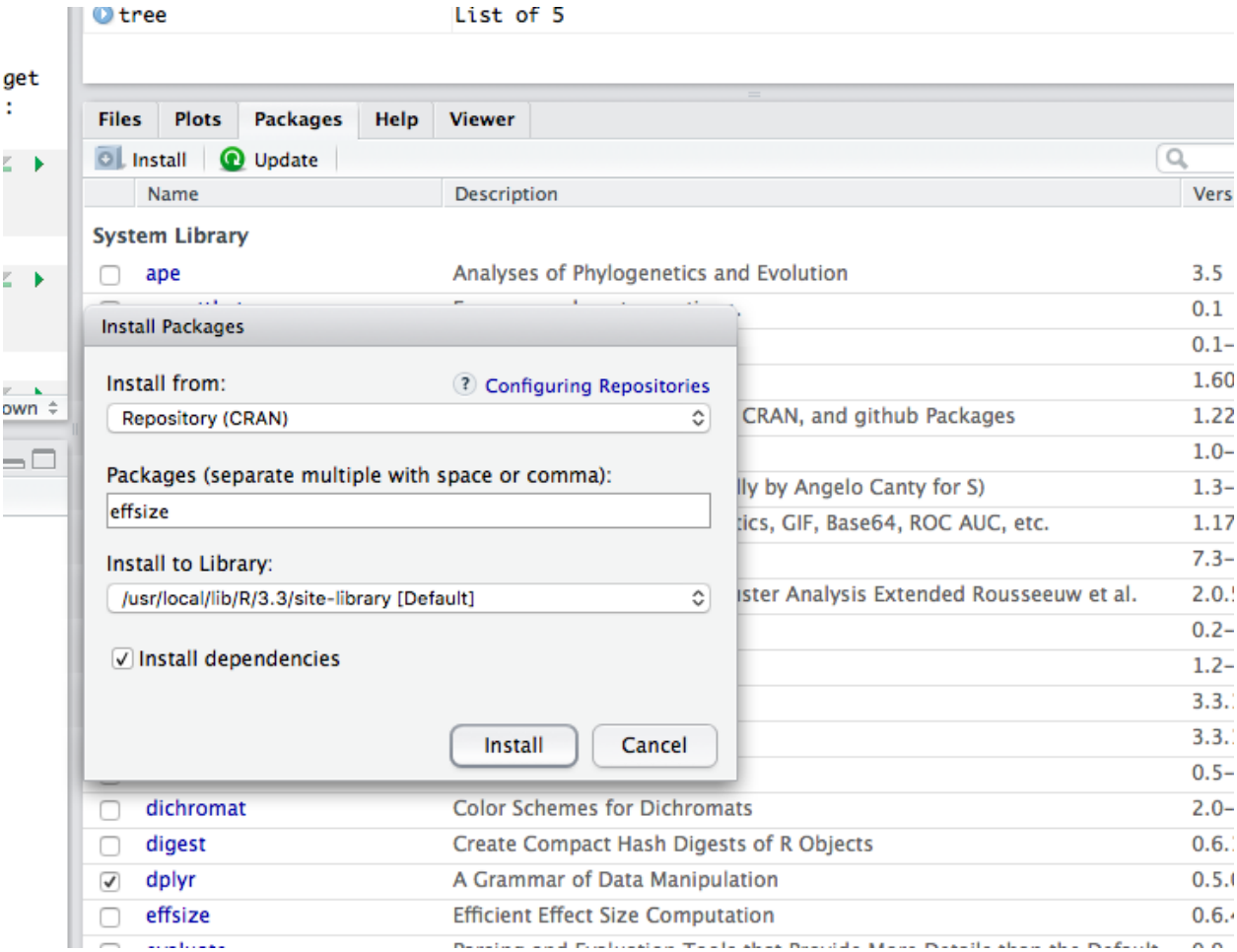


Figure 1: Installing Packages

1.4.2 Bioconductor

For the *Bioconductor* and *ggtree* package, first install *Bioconductor* using its special script.

1. Copy and paste this into the RStudio 'Console' panel on the bottom left. `source("https://bioconductor.org/biocLit`
   Press enter.
2. Copy and paste this into the 'Console' panel: `biocLite()`.
3. Wait (go for a tea-break, this bit takes ages).
4. Copy and paste this into the 'Console' panel: `biocLite("ggtree")`

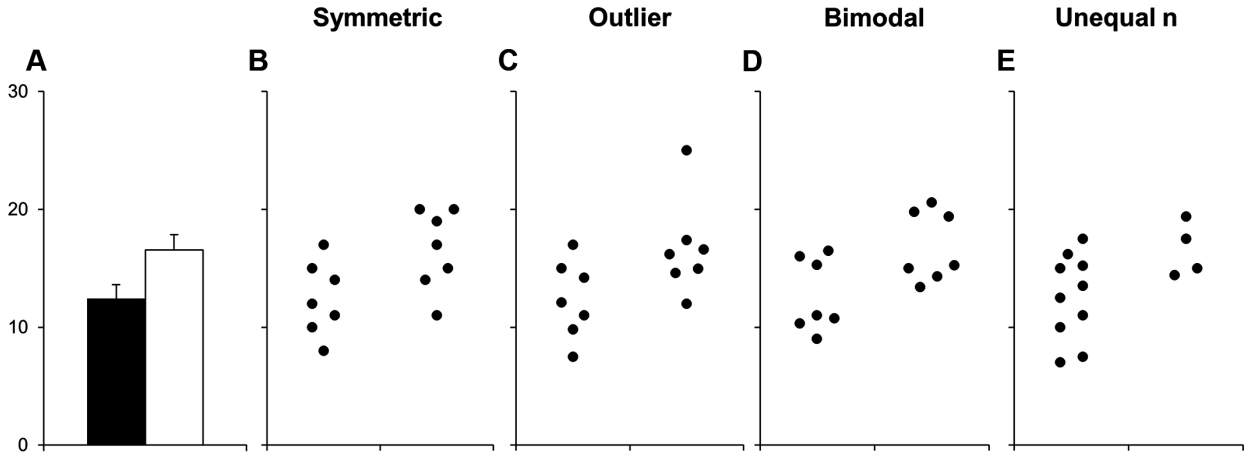
2 Motivation

2.1 Variability in measurements

Variability in measurements is a thing that happens as a natural consequence of working with complex systems that are affected by many variables in stochastic ways. Biological systems are some of the most variable we know. The variability could be a function of the behaviour of the system yet it is common practice to hide that variability when we start to analyse our data by using summary plots like box-plots. Ultimately, that's bad news for our science, because the variability could be telling us something.

2.2 Summarising your data can lead to wrong conclusions

We all know that when you create a bar chart and put some error bars on it, you're really only representing two numbers, usually a mean and standard deviation. People create bar plots instinctively, and in doing so can miss important stuff. Look at this figure:



Test			p value	
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

source: Weissgerber et al

The bar chart in panel A is one that came out of all those sets of numbers in the other panels. But it really hides some important stuff, like the fact the numbers are clearly separating into two groups in panel D, or that the two samples have different sizes in panel E.

Worse than any of these is that the significant difference in the t-test is coming from just one point in panel C. From this data set you might be tempted to conclude that there is a significant difference in the two samples and if you relied on the bar chart as a visualisation then you'd never suspect there was something funny.

Some enthusiastic young science communicators have even started a Kickstarter to lobby journals to stop using, in particular, bar charts! These people, calling themselves Bar Barplots, have a nice video on one of the main problems with bar charts.

```
<source src="https://ksr-video.imgix.net/projects/2453455/video-665338-h264_high.mp4" type="video/mp4">
    Your browser does not support the video tag
```

source: Kickstarter - Barbarplots

So ignoring your data visualisation and just making bar plots could be an error! It's important that you spend a little time getting to know, and presenting your data as clearly and thoroughly as possible.

2.3 *ggplot2* An R package for beautiful visualisations

In this tutorial we are going to use *ggplot2* to make some clear, informative, thorough visualisations. Here's an example:

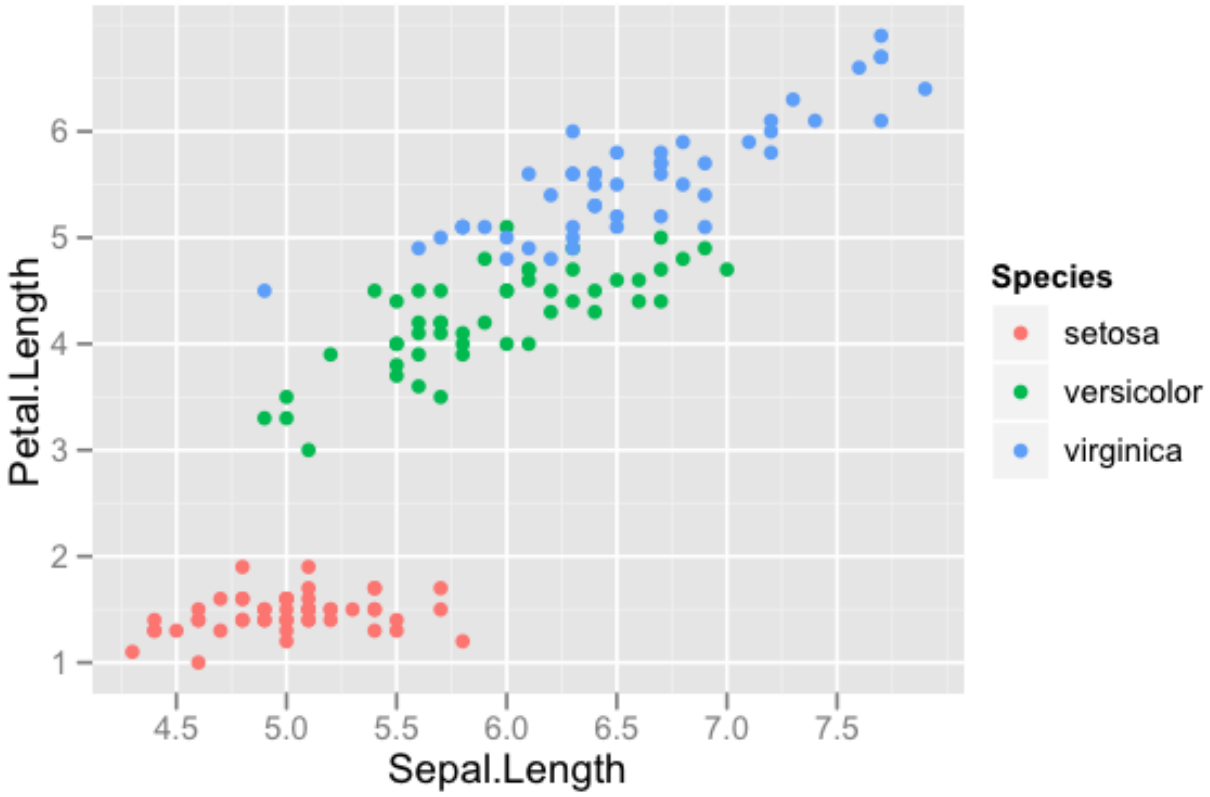


Figure 2: ggplot 2 iris data

ggplot2 is a library in the R statistical programming language - but we won't be learning to program here. The *gg* part stands for 'grammar of graphics', *ggplot2* is a small grammar that describes plots that should be built on top of data - effectively allowing a user to write their own plot description and have the computer work out what to do.

3 R Fundamentals

3.1 About this chapter

1. Questions:

- How do I use R?

2. Objectives:

- Become familiar with R syntax
- Understand the concepts of objects and assignment
- Get exposed to a few functions

3. Keypoints:

- R's capabilities are provided by functions
- R users call functions and get results

3.2 Working with R

In this workshop we'll use R in the extremely useful RStudio package. For the most part we'll work interactively, meaning we'll type stuff straight into the R console in RStudio and get our results there too. That's what you see in these R's most basic job is as a calculator:

```
3 + 5
```

```
## [1] 8
```

```
12 * 2
```

```
## [1] 24
```

```
1 / 3
```

```
## [1] 0.3333333
```

```
12 * 2
```

```
## [1] 24
```

```
 3 / 0
```

```
## [1] Inf
```

Fairly straightforward, except for the `[1]` in the output, this tells us how far through the output we are. Often R will return long lists of numbers and it can be helpful to have this extra information

3.3 Variables

We can save the output of operations for later use by giving it a name using the assignment symbol `<-`. Read this symbol as 'gets', so `x <- 5` reads as 'x gets 5'. These names are called variables, because the value they are associated with can change.

Let's give five a name, `x` then refer to the value 5 by it's name. We can then use the name in place of the value. In the jargon of computing we say we are assigning a value to a variable.

```
x <- 5
x
```

```
## [1] 5
```

```
x * 2
```

```
## [1] 10
```

```
y <- 3
x * y
```

```
## [1] 15
```

This is of course of limited value with just numbers but is of great value when we have large datasets, as the whole thing can be referred to by the variable.

3.3.1 Using objects and functions

At the top level, R is a simple language with two types of thing: functions and objects. As a user you will use functions to do stuff, and get back objects as an answer. Functions are easy to spot, they are a name followed by a pair of brackets like `mean()` is the function for calculating a mean. The options (or arguments) for the function go inside the brackets:

```
sqrt(16)
```

```
## [1] 4
```

Often the result from a function will be more complicated than a simple number object, often it will be a vector (simple list), like from the `rnorm()` function that returns lists of random numbers

```
rnorm(100)
```

```
##   [1] -0.3440279972  1.0528565977  1.0951463161  1.6146287767  0.0643405126
##   [6]  0.8150520401  1.1131992229  1.5564818953 -1.0730800700 -0.0286836864
##  [11]  0.2134190446 -0.3337661278 -0.0527924235 -0.2630346624  2.4030564722
##  [16]  0.0141123664  0.8043080602 -0.0006774788  1.3591747213 -0.3004262764
##  [21] -1.3653661614  0.8051350481 -0.9113255135 -0.4463958514  0.2009897343
##  [26] -2.0629684044 -0.2548826087  0.3811335218  0.7010657598  1.0928761912
##  [31]  1.8899795460 -1.9316535381 -0.6621049686 -0.7652779105 -0.8194588153
##  [36] -1.3832535471 -1.0422260900  0.7268774406  0.1336925037 -0.0025233944
##  [41]  0.0655083809 -0.2211289639  0.0011990119 -0.7132239788  0.1128713217
```

```
##  [46]  1.5541678420  2.4038229197  2.3546010950 -0.6923038842 -0.3636402581
##  [51]  0.7324728773  0.3815717569 -0.9882116183 -0.7688368971  0.6202001920
##  [56] -0.8252200355  0.1567816861 -1.4575353153 -2.1821313473 -0.2329023457
##  [61] -1.6828180555  0.2151405134  0.1168080434  2.0159347262 -0.9265134101
##  [66] -0.6525359119  0.4757340770  0.5217919797  0.0585857209  1.6779806891
##  [71] -0.7154721146  1.9364878384 -0.8551767883  0.4350736408  0.6199049399
##  [76] -0.6449422326  1.0069026206  0.1673614251 -0.6180526246  0.9589750941
##  [81]  2.0913448707  0.1083010549 -1.1409995556 -0.4219437408 -1.4425869694
##  [86]  0.8424045388  0.3343154062  1.2022374158  0.3152616161  0.0048723969
##  [91] -0.5274305980 -1.0437469925  0.9863818090 -1.3098056383  0.3830374045
##  [96] -1.3918549084 -1.1026361148 -0.4177647183 -0.3859381352 -0.4623020641
```

We can combine objects, variables and functions to do more complex stuff in R, here's how we get the mean of 100 random numbers.

```
numbers <- rnorm(100)
mean(numbers)
```

```
## [1] -0.0219514
```

Here we created a vector object with `rnorm(100)` and assigned it to the variable `numbers`. We than used the `mean()` function, passing it the variable `numbers`. The `mean()` function returned the mean of the hundred random numbers.

3.4 Bracket notation in this document

I'm going to use the following descriptions for the symbols `()`, `[]` and `{}`:

`()` are brackets, `[]` are square brackets `{}` are curly brackets

3.5 Quiz

1. Create two variables, `a` and `b`: Add them. What happens if we change a and then re-add a and b?
2. We can also assign `a + b` to a new variable, `c`. How would you do this?
3. Try some R functions: `round()`, `c()`, `range()`, `plot()` hint: Get help on a function by typing `?function_name` e.g `?c()`. Use the `mean()` function to calculate the average age of everyone in your house (Invent a housemate if you have to).

4 *ggplot2* Tour

4.1 About this chapter

1. Questions:

- How does ggplot2 work?

2. Objectives:

- Explain the structure of a ggplot2
- Explain the flexibilty of the structure

3. Keypoints:

- ggplot2 plots are made in user defined layers
- Using layers helps us to change plot types quickly or build progressively more complex charts

4.2 Building a plot with *ggplot2*

Loading *ggplot2* into memory so we can use it is very easy. With RStudio started, and in the console window type:

```
library(ggplot2)
```

Nothing should happen, thats a good sign!

4.3 It didn't load - I got an error

You need to go back and look at the install instructions, using the packages tab in the bottom right hand window of R studio, click `install` and type `ggplot2` into the window that appears. Select `install` and it should automatically install. If this doesn't work seek some expert help. {: .callout }

4.3.1 Loading the iris test data

R has some datasets built in that allow us to easily develop analysis. Let's look at the `iris` data

```
iris
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1            5.1         3.5          1.4         0.2     setosa
## 2            4.9         3.0          1.4         0.2     setosa
## 3            4.7         3.2          1.3         0.2     setosa
## 4            4.6         3.1          1.5         0.2     setosa
## 5            5.0         3.6          1.4         0.2     setosa
## 6            5.4         3.9          1.7         0.4     setosa
## 7            4.6         3.4          1.4         0.3     setosa
## 8            5.0         3.4          1.5         0.2     setosa
## 9            4.4         2.9          1.4         0.2     setosa
## 10           4.9         3.1          1.5         0.1     setosa
## 11           5.4         3.7          1.5         0.2     setosa
## 12           4.8         3.4          1.6         0.2     setosa
## 13           4.8         3.0          1.4         0.1     setosa
## 14           4.3         3.0          1.1         0.1     setosa
## 15           5.8         4.0          1.2         0.2     setosa
## 16           5.7         4.4          1.5         0.4     setosa
## 17           5.4         3.9          1.3         0.4     setosa
## 18           5.1         3.5          1.4         0.3     setosa
## 19           5.7         3.8          1.7         0.3     setosa
## 20           5.1         3.8          1.5         0.3     setosa
## 21           5.4         3.4          1.7         0.2     setosa
## 22           5.1         3.7          1.5         0.4     setosa
## 23           4.6         3.6          1.0         0.2     setosa
```

```
## 24          5.1         3.3          1.7         0.5     setosa
## 25          4.8         3.4          1.9         0.2     setosa
## 26          5.0         3.0          1.6         0.2     setosa
## 27          5.0         3.4          1.6         0.4     setosa
## 28          5.2         3.5          1.5         0.2     setosa
## 29          5.2         3.4          1.4         0.2     setosa
## 30          4.7         3.2          1.6         0.2     setosa
## 31          4.8         3.1          1.6         0.2     setosa
## 32          5.4         3.4          1.5         0.4     setosa
## 33          5.2         4.1          1.5         0.1     setosa
## 34          5.5         4.2          1.4         0.2     setosa
## 35          4.9         3.1          1.5         0.2     setosa
## 36          5.0         3.2          1.2         0.2     setosa
## 37          5.5         3.5          1.3         0.2     setosa
## 38          4.9         3.6          1.4         0.1     setosa
## 39          4.4         3.0          1.3         0.2     setosa
## 40          5.1         3.4          1.5         0.2     setosa
## 41          5.0         3.5          1.3         0.3     setosa
## 42          4.5         2.3          1.3         0.3     setosa
## 43          4.4         3.2          1.3         0.2     setosa
## 44          5.0         3.5          1.6         0.6     setosa
## 45          5.1         3.8          1.9         0.4     setosa
## 46          4.8         3.0          1.4         0.3     setosa
## 47          5.1         3.8          1.6         0.2     setosa
## 48          4.6         3.2          1.4         0.2     setosa
## 49          5.3         3.7          1.5         0.2     setosa
## 50          5.0         3.3          1.4         0.2     setosa
## 51          7.0         3.2          4.7         1.4 versicolor
## 52          6.4         3.2          4.5         1.5 versicolor
## 53          6.9         3.1          4.9         1.5 versicolor
## 54          5.5         2.3          4.0         1.3 versicolor
## 55          6.5         2.8          4.6         1.5 versicolor
## 56          5.7         2.8          4.5         1.3 versicolor
## 57          6.3         3.3          4.7         1.6 versicolor
## 58          4.9         2.4          3.3         1.0 versicolor
## 59          6.6         2.9          4.6         1.3 versicolor
## 60          5.2         2.7          3.9         1.4 versicolor
## 61          5.0         2.0          3.5         1.0 versicolor
## 62          5.9         3.0          4.2         1.5 versicolor
## 63          6.0         2.2          4.0         1.0 versicolor
## 64          6.1         2.9          4.7         1.4 versicolor
## 65          5.6         2.9          3.6         1.3 versicolor
## 66          6.7         3.1          4.4         1.4 versicolor
## 67          5.6         3.0          4.5         1.5 versicolor
## 68          5.8         2.7          4.1         1.0 versicolor
## 69          6.2         2.2          4.5         1.5 versicolor
## 70          5.6         2.5          3.9         1.1 versicolor
## 71          5.9         3.2          4.8         1.8 versicolor
## 72          6.1         2.8          4.0         1.3 versicolor
## 73          6.3         2.5          4.9         1.5 versicolor
## 74          6.1         2.8          4.7         1.2 versicolor
## 75          6.4         2.9          4.3         1.3 versicolor
## 76          6.6         3.0          4.4         1.4 versicolor
## 77          6.8         2.8          4.8         1.4 versicolor
```

```
## 78          6.7         3.0          5.0         1.7 versicolor
## 79          6.0         2.9          4.5         1.5 versicolor
## 80          5.7         2.6          3.5         1.0 versicolor
## 81          5.5         2.4          3.8         1.1 versicolor
## 82          5.5         2.4          3.7         1.0 versicolor
## 83          5.8         2.7          3.9         1.2 versicolor
## 84          6.0         2.7          5.1         1.6 versicolor
## 85          5.4         3.0          4.5         1.5 versicolor
## 86          6.0         3.4          4.5         1.6 versicolor
## 87          6.7         3.1          4.7         1.5 versicolor
## 88          6.3         2.3          4.4         1.3 versicolor
## 89          5.6         3.0          4.1         1.3 versicolor
## 90          5.5         2.5          4.0         1.3 versicolor
## 91          5.5         2.6          4.4         1.2 versicolor
## 92          6.1         3.0          4.6         1.4 versicolor
## 93          5.8         2.6          4.0         1.2 versicolor
## 94          5.0         2.3          3.3         1.0 versicolor
## 95          5.6         2.7          4.2         1.3 versicolor
## 96          5.7         3.0          4.2         1.2 versicolor
## 97          5.7         2.9          4.2         1.3 versicolor
## 98          6.2         2.9          4.3         1.3 versicolor
## 99          5.1         2.5          3.0         1.1 versicolor
## 100         5.7         2.8          4.1         1.3 versicolor
## 101         6.3         3.3          6.0         2.5  virginica
## 102         5.8         2.7          5.1         1.9  virginica
## 103         7.1         3.0          5.9         2.1  virginica
## 104         6.3         2.9          5.6         1.8  virginica
## 105         6.5         3.0          5.8         2.2  virginica
## 106         7.6         3.0          6.6         2.1  virginica
## 107         4.9         2.5          4.5         1.7  virginica
## 108         7.3         2.9          6.3         1.8  virginica
## 109         6.7         2.5          5.8         1.8  virginica
## 110         7.2         3.6          6.1         2.5  virginica
## 111         6.5         3.2          5.1         2.0  virginica
## 112         6.4         2.7          5.3         1.9  virginica
## 113         6.8         3.0          5.5         2.1  virginica
## 114         5.7         2.5          5.0         2.0  virginica
## 115         5.8         2.8          5.1         2.4  virginica
## 116         6.4         3.2          5.3         2.3  virginica
## 117         6.5         3.0          5.5         1.8  virginica
## 118         7.7         3.8          6.7         2.2  virginica
## 119         7.7         2.6          6.9         2.3  virginica
## 120         6.0         2.2          5.0         1.5  virginica
## 121         6.9         3.2          5.7         2.3  virginica
## 122         5.6         2.8          4.9         2.0  virginica
## 123         7.7         2.8          6.7         2.0  virginica
## 124         6.3         2.7          4.9         1.8  virginica
## 125         6.7         3.3          5.7         2.1  virginica
## 126         7.2         3.2          6.0         1.8  virginica
## 127         6.2         2.8          4.8         1.8  virginica
## 128         6.1         3.0          4.9         1.8  virginica
## 129         6.4         2.8          5.6         2.1  virginica
## 130         7.2         3.0          5.8         1.6  virginica
## 131         7.4         2.8          6.1         1.9  virginica
```

```
## 132          7.9         3.8          6.4         2.0  virginica
## 133          6.4         2.8          5.6         2.2  virginica
## 134          6.3         2.8          5.1         1.5  virginica
## 135          6.1         2.6          5.6         1.4  virginica
## 136          7.7         3.0          6.1         2.3  virginica
## 137          6.3         3.4          5.6         2.4  virginica
## 138          6.4         3.1          5.5         1.8  virginica
## 139          6.0         3.0          4.8         1.8  virginica
## 140          6.9         3.1          5.4         2.1  virginica
## 141          6.7         3.1          5.6         2.4  virginica
## 142          6.9         3.1          5.1         2.3  virginica
## 143          5.8         2.7          5.1         1.9  virginica
## 144          6.8         3.2          5.9         2.3  virginica
## 145          6.7         3.3          5.7         2.5  virginica
## 146          6.7         3.0          5.2         2.3  virginica
## 147          6.3         2.5          5.0         1.9  virginica
## 148          6.5         3.0          5.2         2.0  virginica
## 149          6.2         3.4          5.4         2.3  virginica
## 150          5.9         3.0          5.1         1.8  virginica
```

R just printed the whole thing to screen and we end up looking at just the bottom end of it. Let's look at just the top.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

We can see that we have the top six rows and we can see that the data is a list of measurments of the sepals and petals for some species of iris. Let's get a summary of the data set:

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Alright, that's quite clear, some summary values for each numeric column and note how R has calculated the number of rows of each distinct label for the text column.

4.3.2 A first plot

ggplot2 plots are built up of layers, the foundation layer is the data layer, thats the whole data set containing the bits we would want to plot. We define that with the `ggplot2` command.

```
library(ggplot2)
ggplot(data=iris)
```

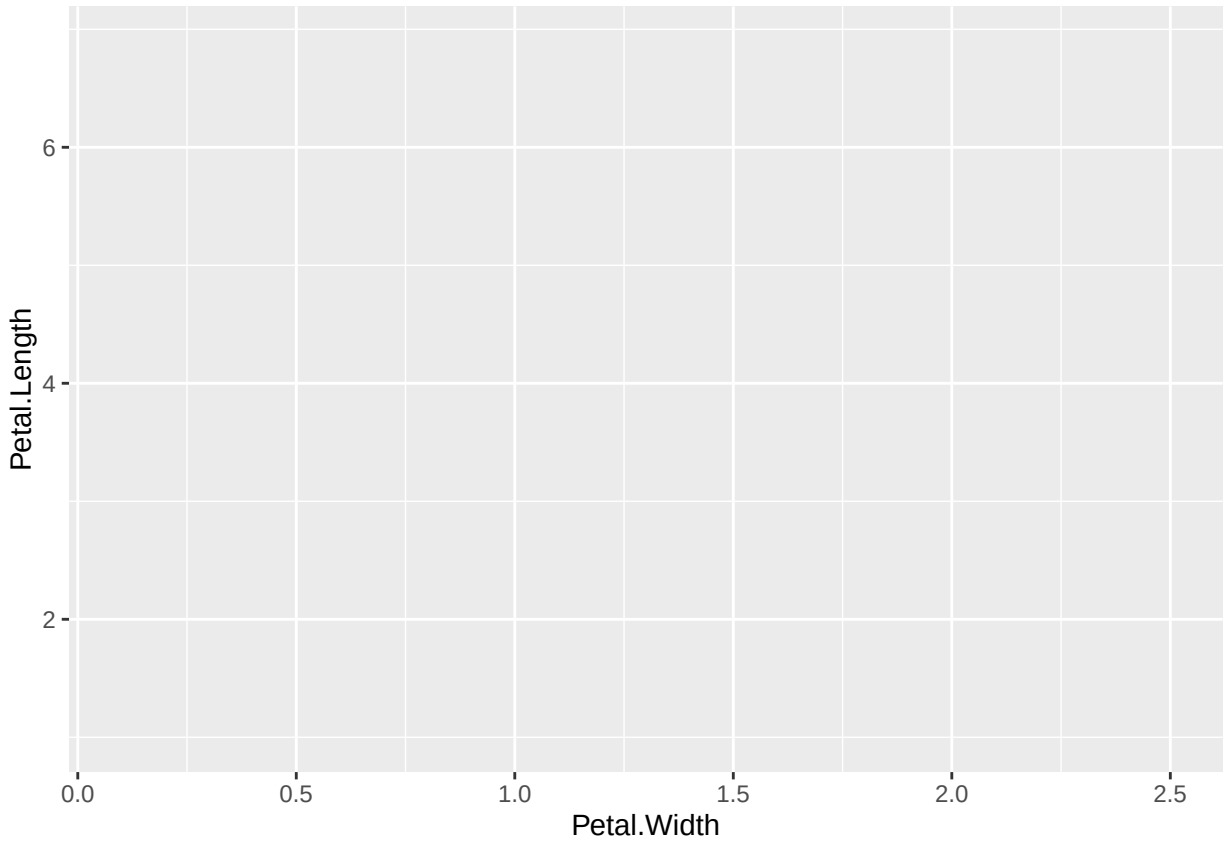

Nothing happened, you got a blank screen in the plot window to the right. That's because a data layer alone doesn't tell us what, or how to plot. It's just the source of the numbers we'll use.

The next thing we need is an `aesthetic` layer. This is basically the things to look at, and includes:

1. x and y axes (sometimes called position)
2. colour (the line colour of a thing)
3. fill (the block colour of a thing)
4. shape (e.g of points)
5. line type
6. size (e.g of points)

Let's decide to look at petal width and length. We use the `aes()` function for the aesthetic and we can add layers together with the `+` operator.

```
ggplot(data=iris) + aes(x=Petal.Width, y=Petal.Length)
```

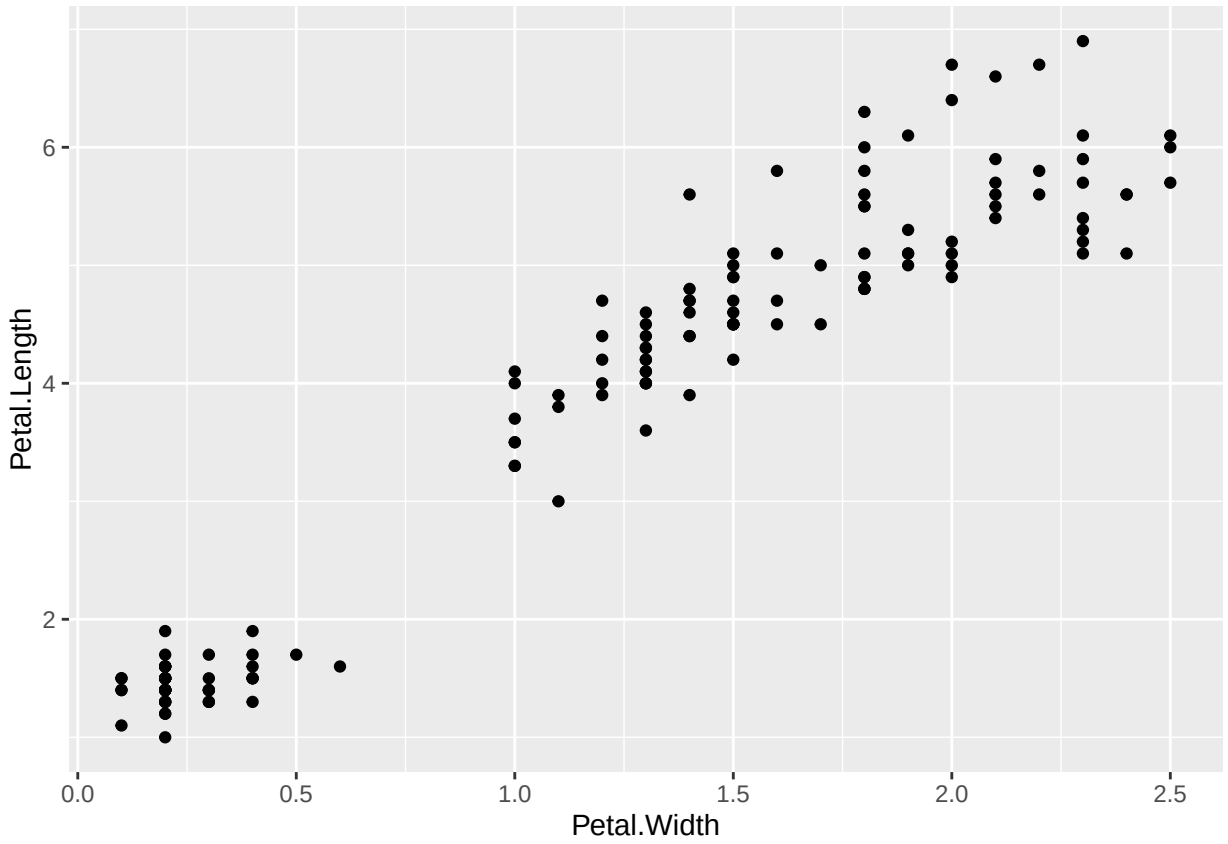


This time we get axes in the plot window. *ggplot* now knows the data source and the things that should be displayed on the axis, but it doesn't fully know *how* to display them. That is done in the `geom` (or geometric objects) layer. There are loads of geoms e.g

```
1. `geom_point()` for scatter plots
2. `geom_line()` for trend lines
3. `geom_boxplot()` for boxplots!
```

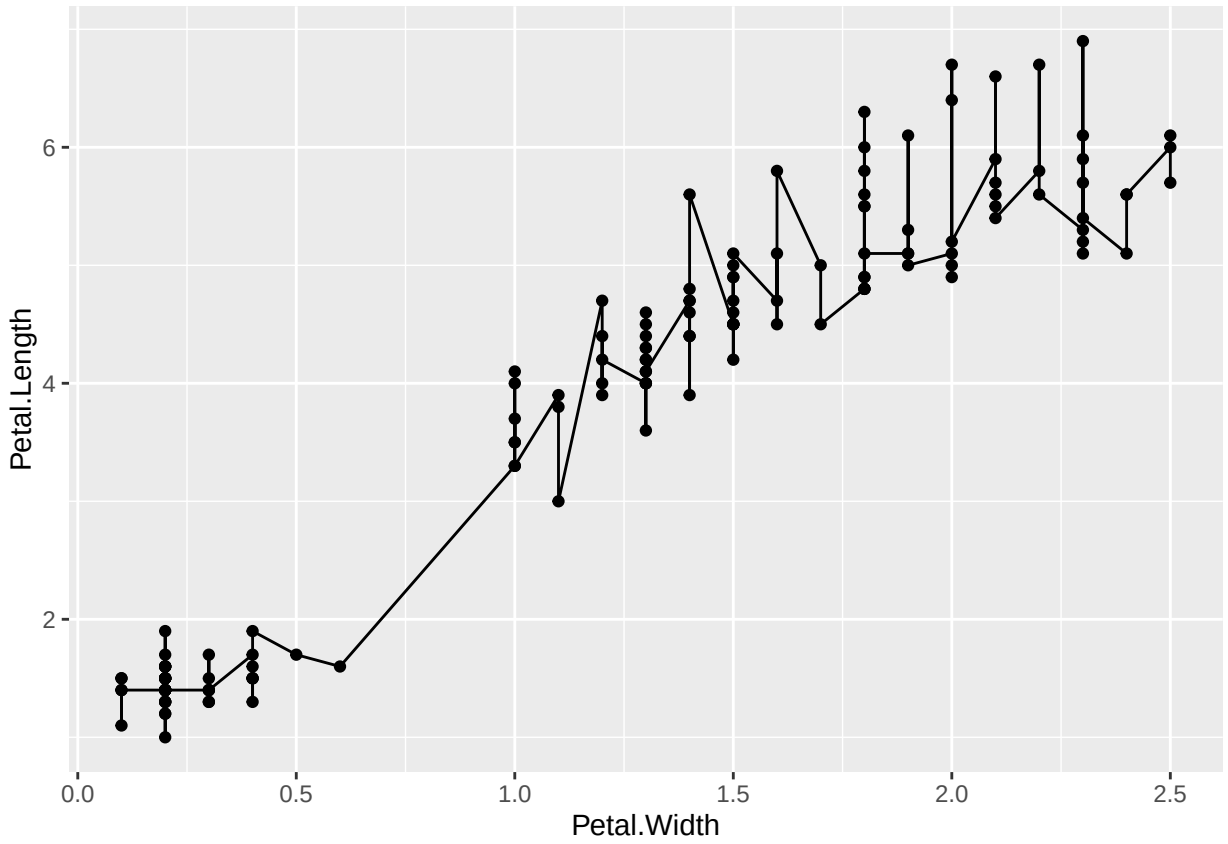
Let's add a geom layer.

```
ggplot(data=iris) + aes(x=Petal.Width, y=Petal.Length) + geom_point()
```



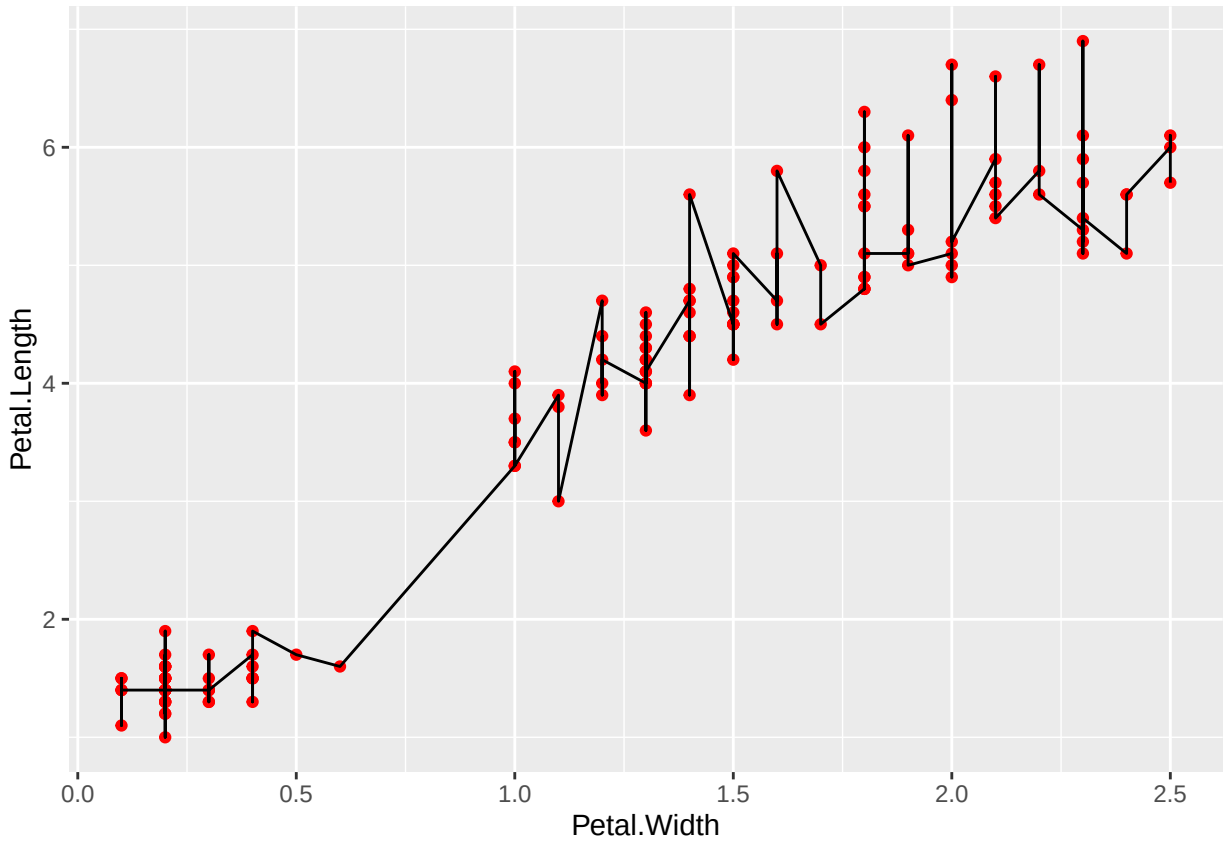
Now we see the whole plot. The data has been mapped onto the right axes and the geometric objects on top of that. Let's go crazy and add more layers.

```
ggplot(data=iris) + aes(x=Petal.Width, y=Petal.Length) + geom_point() + geom_line()
```



You can see the new geom just adds straight on top of the old one. By default, `geom_line()` is a simple join the dots sort of line, so it looks really squiggly. Different layers can have their own options set, e.g the points can be coloured.

```
ggplot(data=iris) + aes(x=Petal.Width, y=Petal.Length) + geom_point(colour="Red") + geom_line()
```



4.4 Making and saving a base plot

There is actually no need to go round typing in the whole command above repetively all the time. *ggplot2* layers can be saved to R variables like this:

```
p <- ggplot(data=iris) + aes(x=Petal.Width, y=Petal.Length)
```

and the bits we want to add or change stuck on top:

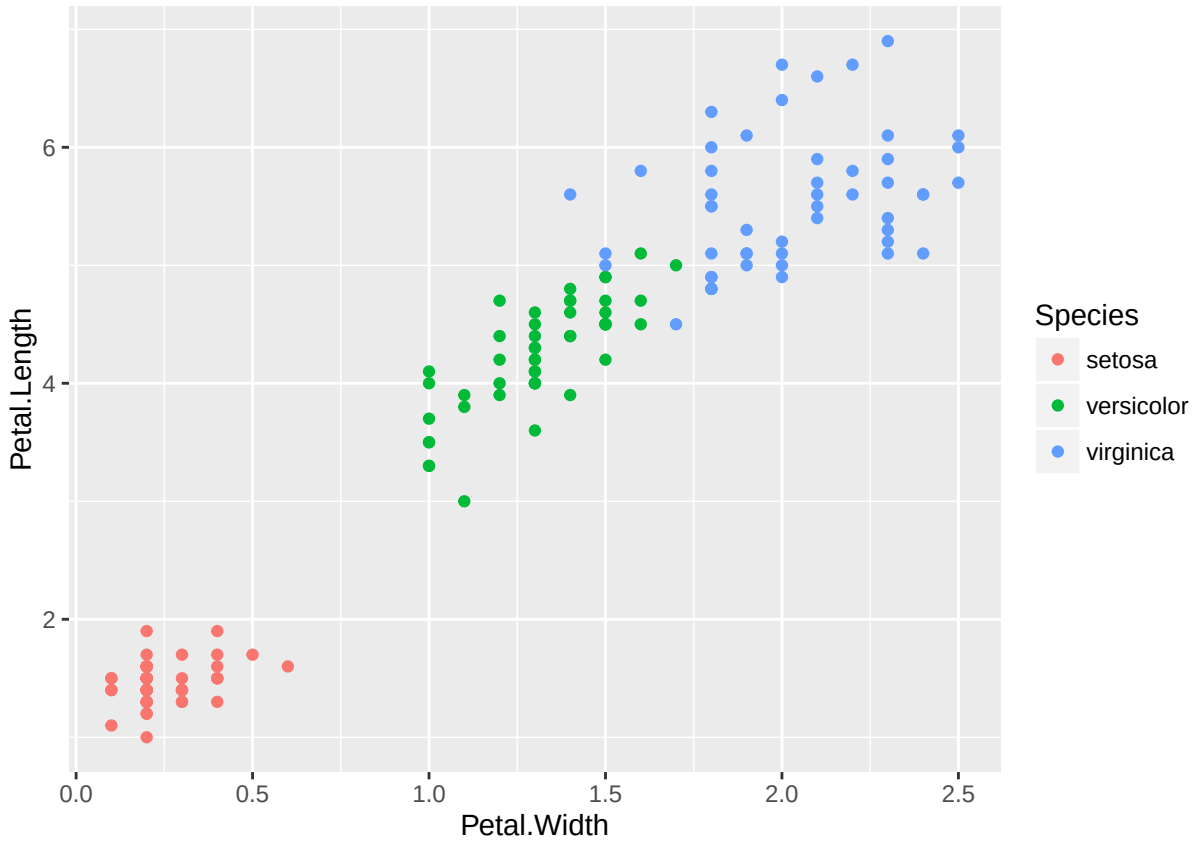
```
p + geom_point()
```

4.5 Mappings versus assignment

The power of *ggplot* to 'just do the right thing' comes from its use of mappings, these can be thought of as rules for what to do when it meets a bit of data in a particular place.

Above we set the colour, `geom_point()` to `"Red"`. This set all the points to red, it was an assignment, since ggplot didn't have anything to work out, every point is just red. By setting the colour to a column in the data we can make ggplot work colours out for us dependent on the information in that column. Try:

```
p <- ggplot(data=iris) + aes(x=Petal.Width, y=Petal.Length)
p + geom_point(aes(colour=Species))
```

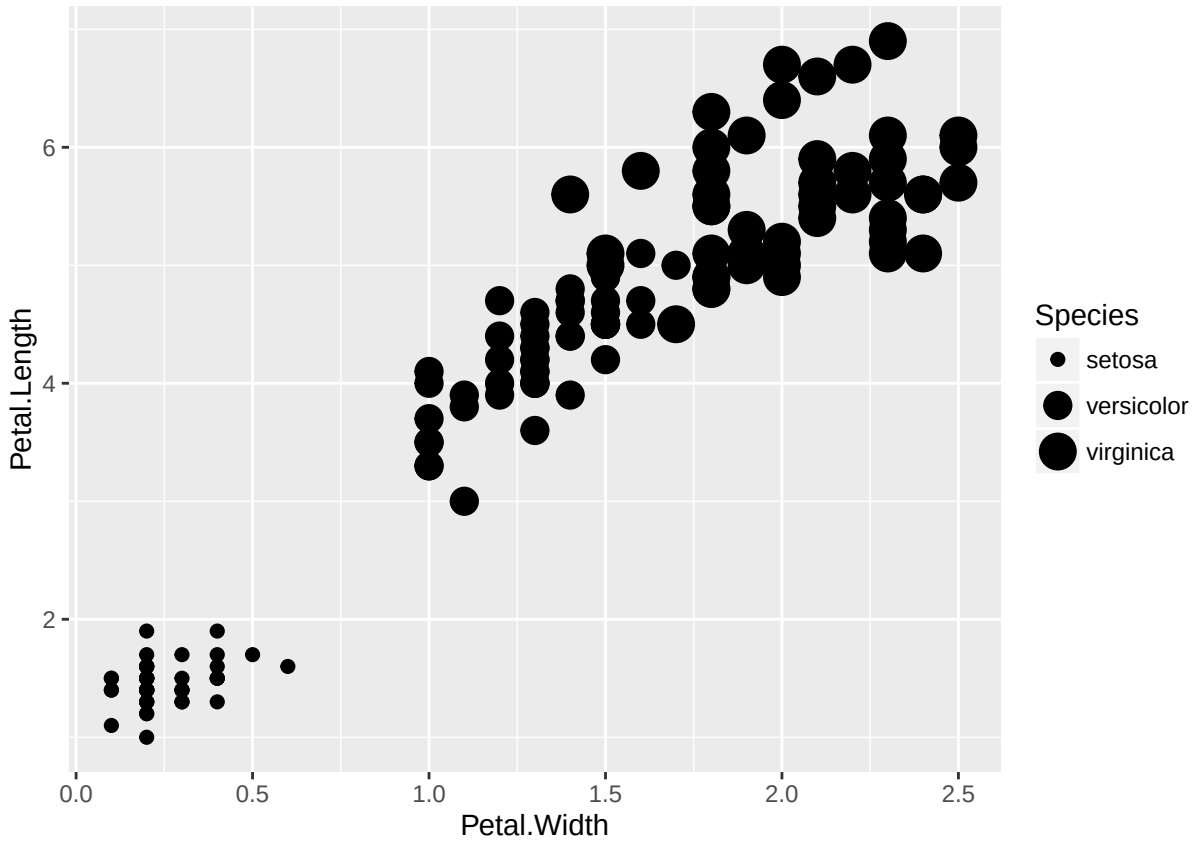


This time we told ggplot to use the value of the Species column to colour each data point, ggplot decided on a mapping for a list of colours to each different value in the Species column and drew that on the plot for us. Only aesthetics can be mappings, so we had to use an `aes()` function inside the geom.

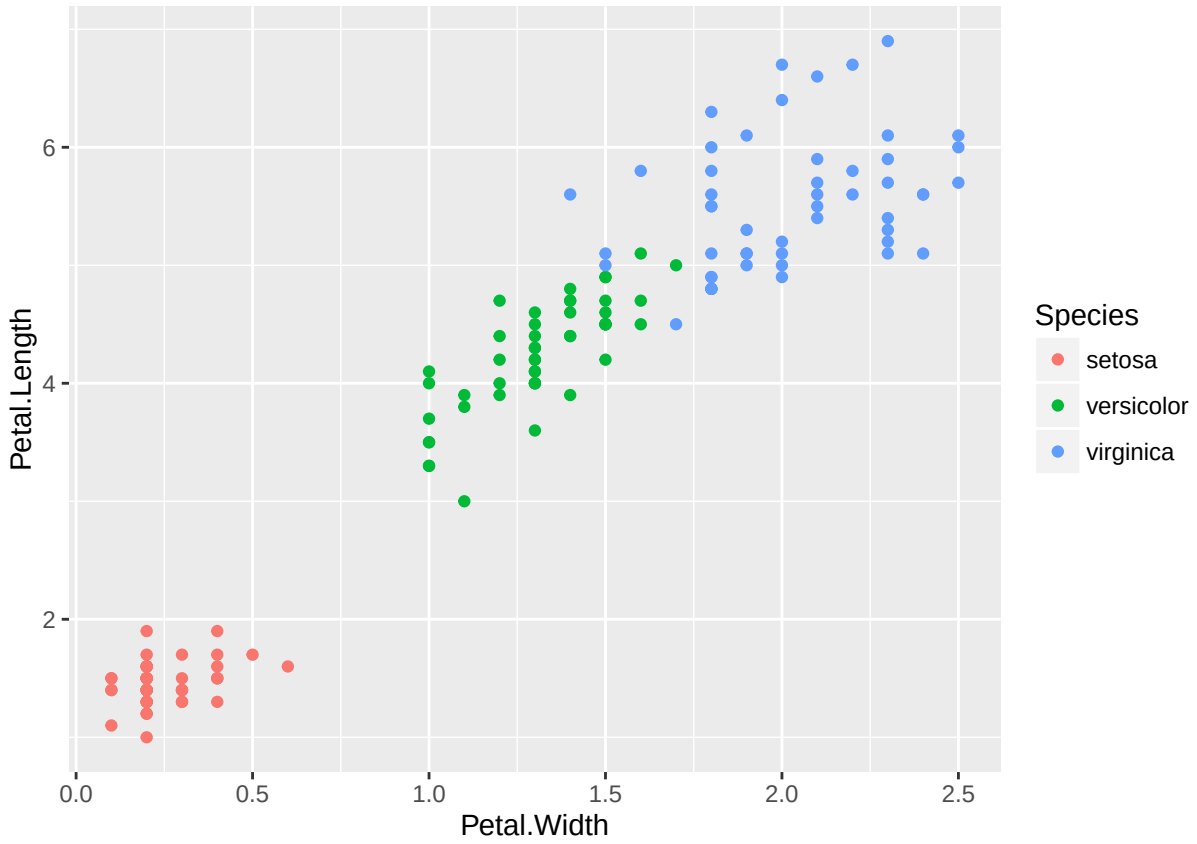
Lots of aesthetic features can be mapped to data, try `size` and `shape`, and try mixing them.

```
p + geom_point(aes(size=Species))
```

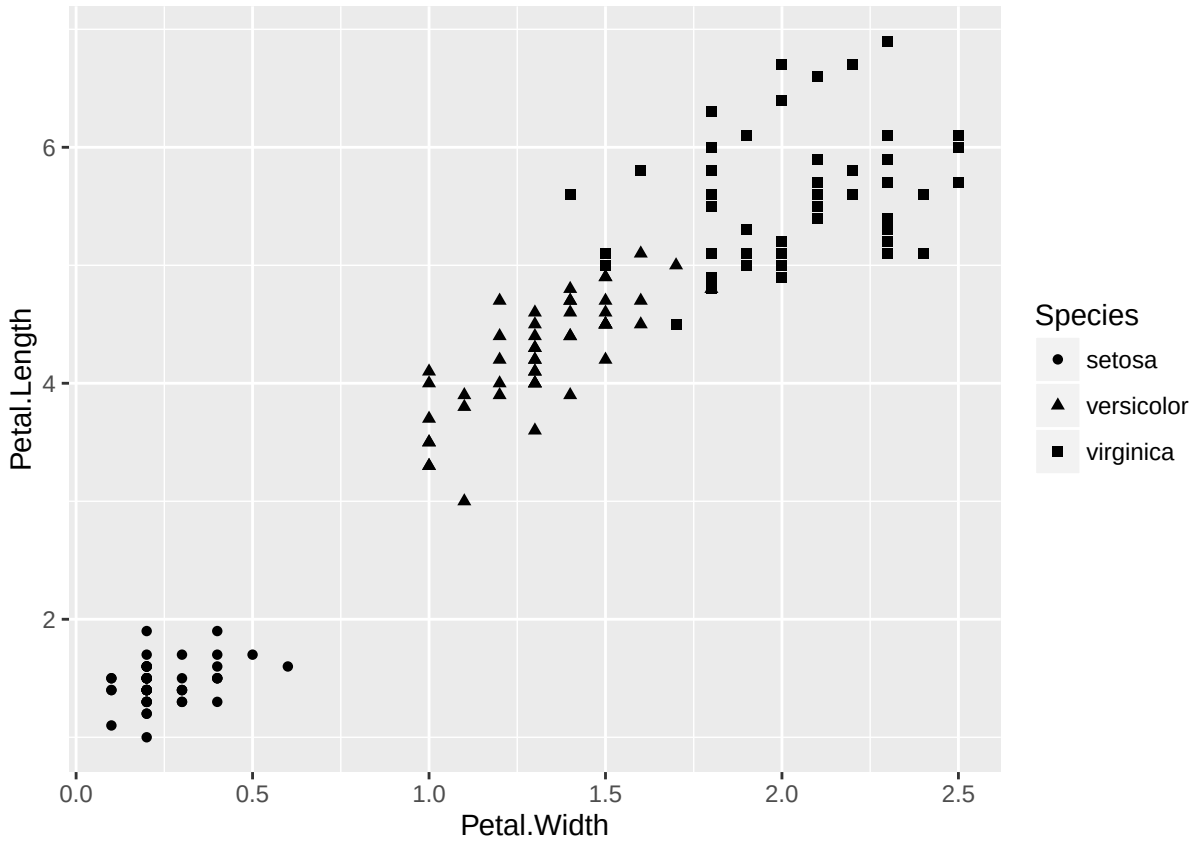
```
## Warning: Using size for a discrete variable is not advised.
```




```
p + geom_point(aes(colour=Species))
```

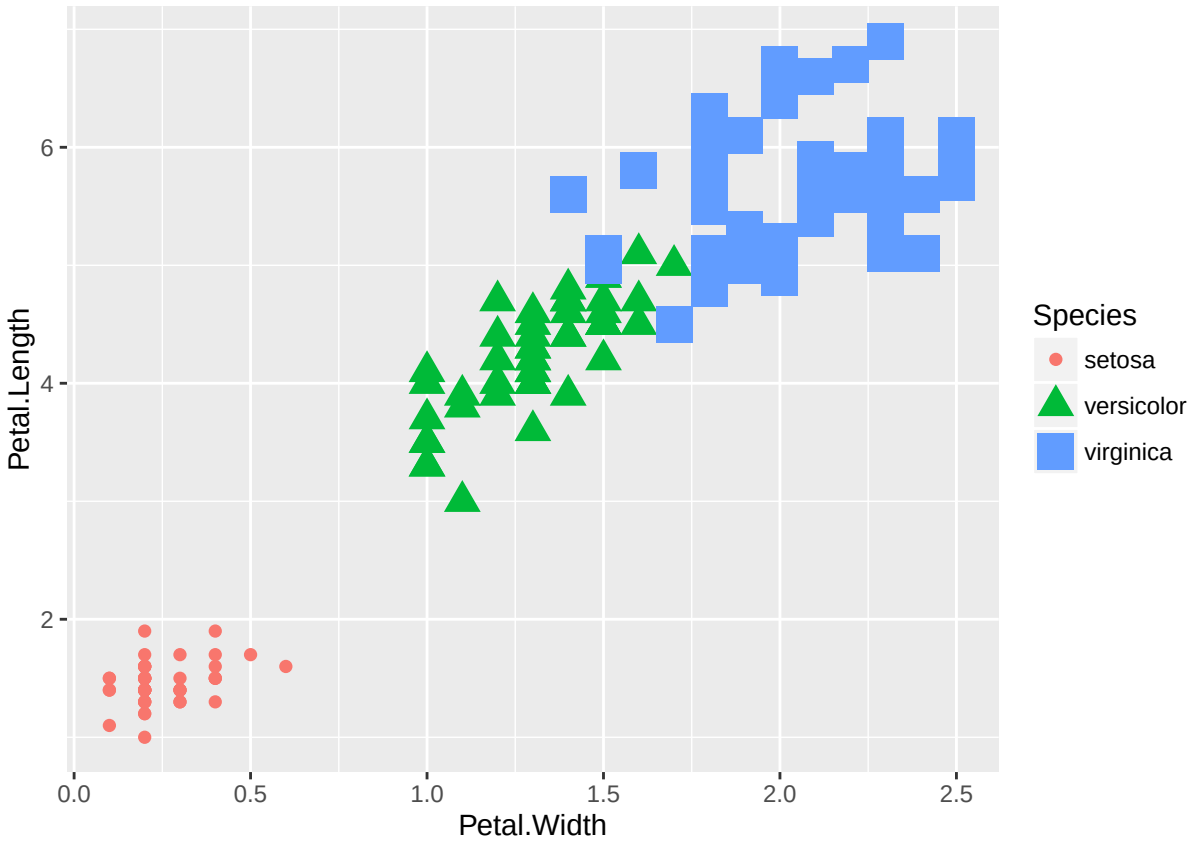



```
p + geom_point(aes(shape=Species))
```




```
p + geom_point(aes(size=Species,colour=Species,shape=Species))
```

```
## Warning: Using size for a discrete variable is not advised.
```



4.6 Quiz

1. Use the docs at http://docs.ggplot2.org/current/ to examine the geoms that are available. Try `geom_jitter()`, why choose this over `geom_point()`?
2. Use this base plot `p <- ggplot(data=iris) + aes(x=Petal.Width, y=Petal.Length)`
3. What happens if you map a continuous variable to an aesthetic like colour? EG `aes(color=Petal.Width)`
4. Try combining `geom_smooth()` with `geom_jitter()`
5. Why doesn't `geom_boxplot()` work? (Hint: you need to think about the difference between categorical or discrete and continuous data).
6. How could you make `geom_boxplot()` show you boxplots for the three species Petal.Width. (Hint: you need to think about the aesthetic and where you set it.

5 Common Geoms

5.1 About this chapter

1. Questions:

- What sorts of plot can I do?

2. Objectives:

- Demonstrate the main types of plot

3. Keypoints:

- There are geoms for continuous and discrete data
- Selecting and mixing these properly can give a nice representation of your data

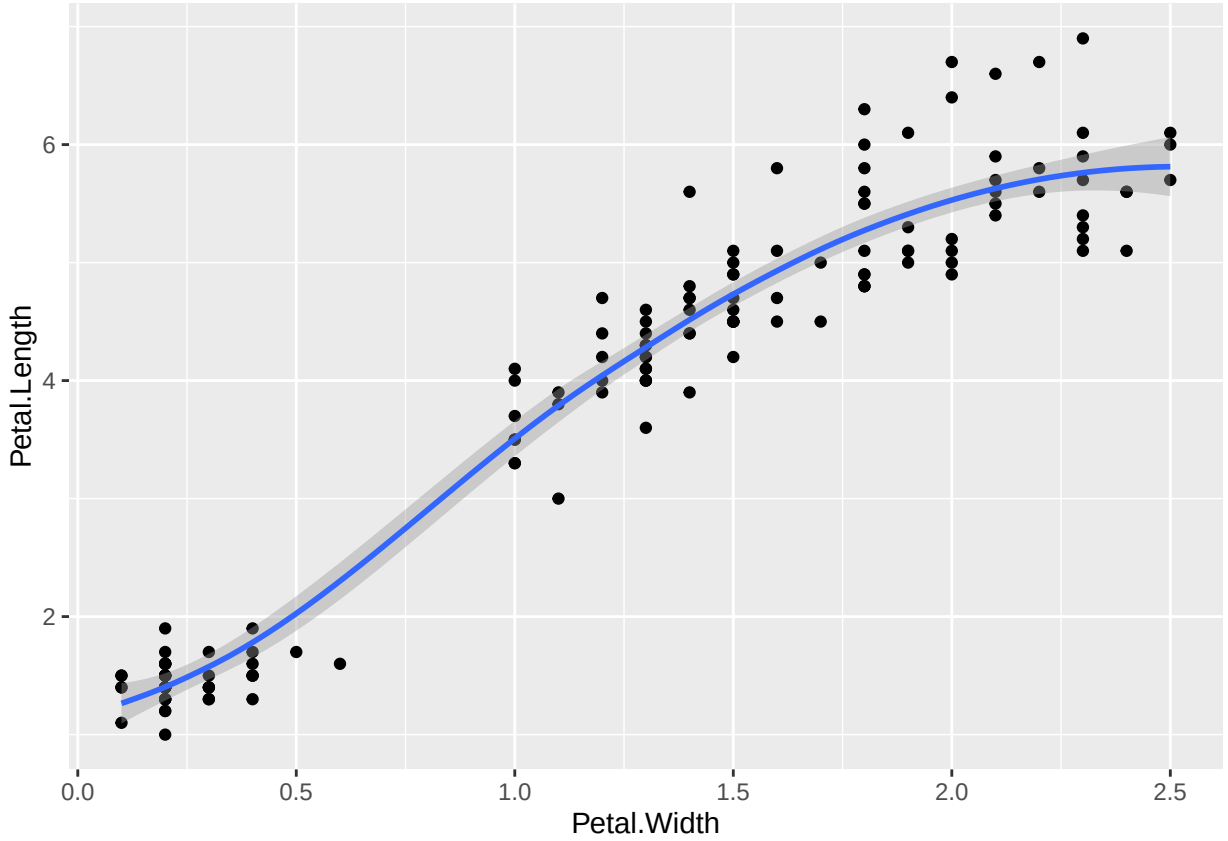
5.2 Continuous geoms

Let's look at some geoms that use continuous data on the x and y axis.

5.2.1 geom_smooth()

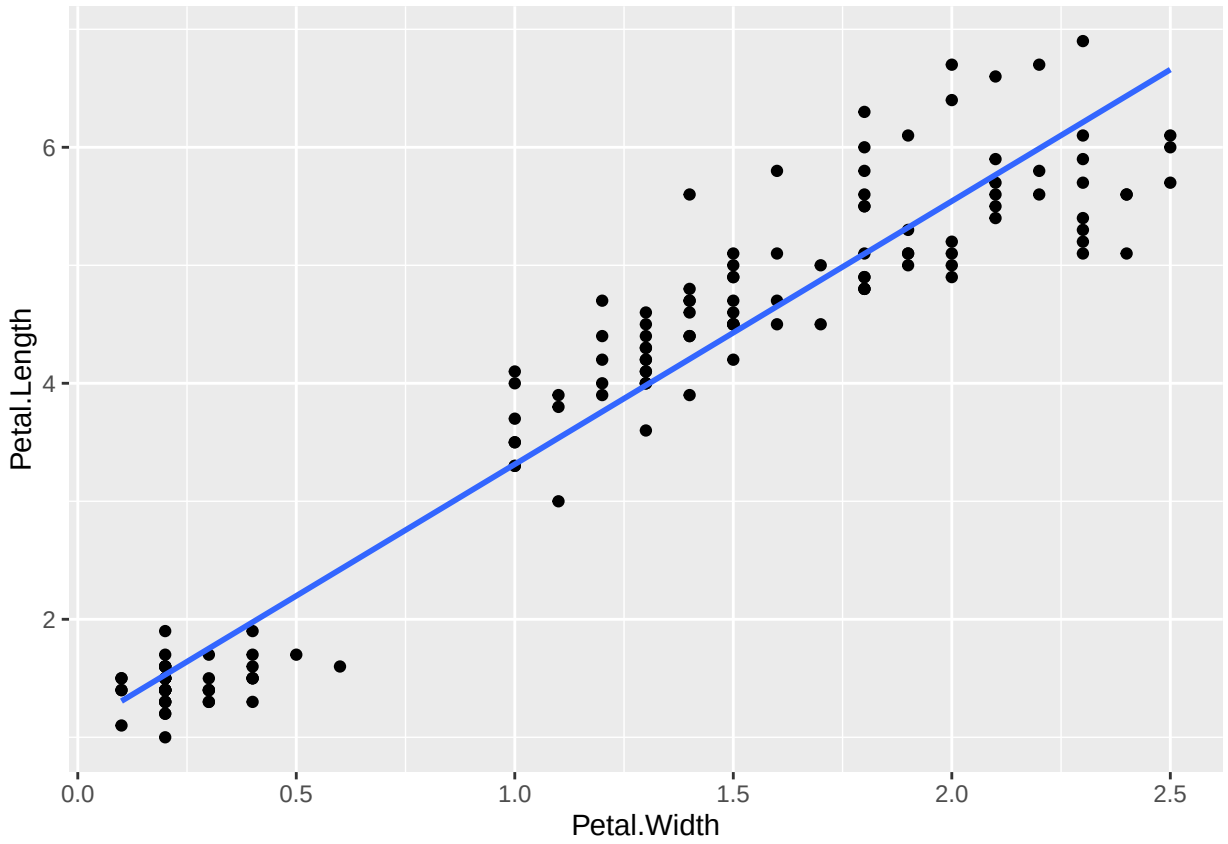
The built in geom `geom_smooth()` is a great one for getting a nice summary line through the data

```
p <- ggplot(iris) + aes(Petal.Width,Petal.Length) + geom_point()
p + geom_smooth()
```



By default, this isn't a simple line of best fit, as you can see the smoothed line has curves! And it has a grey region that shows the standard error of the line. To get the standard line of the form `y=mx+c`, use

```
p + geom_smooth(method = "lm", se = FALSE)
```



5.2.2 What's the r2?

Having shown you how to put the line of best fit on the graph, you probably want to know how to get the equation and r2 value. That takes a little bit of pure R. Here's how, using the `lm` linear model function. The syntax for this is `lm(y ~ x, dataset)` so for the iris data and the graph we just made (note the order Y and X is used in not the order X and Y)

```
model <- lm(Petal.Length ~ Petal.Width, iris)
```

The result is now saved in the `model` variable we just created. This is a complex R object, which we can see a summary of using

```
summary(model)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <2e-16 ***
```

```
## Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

This is complex, but we want model coefficients, that is the `m` value - the slope (here 2.22994) and the `c` value - the intercept (here 1.08356), and the adjusted R-squared (0.9266)

5.3 Shorthand notation

A shorthand in ggplot allows you to leave out the `data=` part of the function call, if you put the data in the first position so

`ggplot(iris)` is the same as `ggplot(data=iris)`

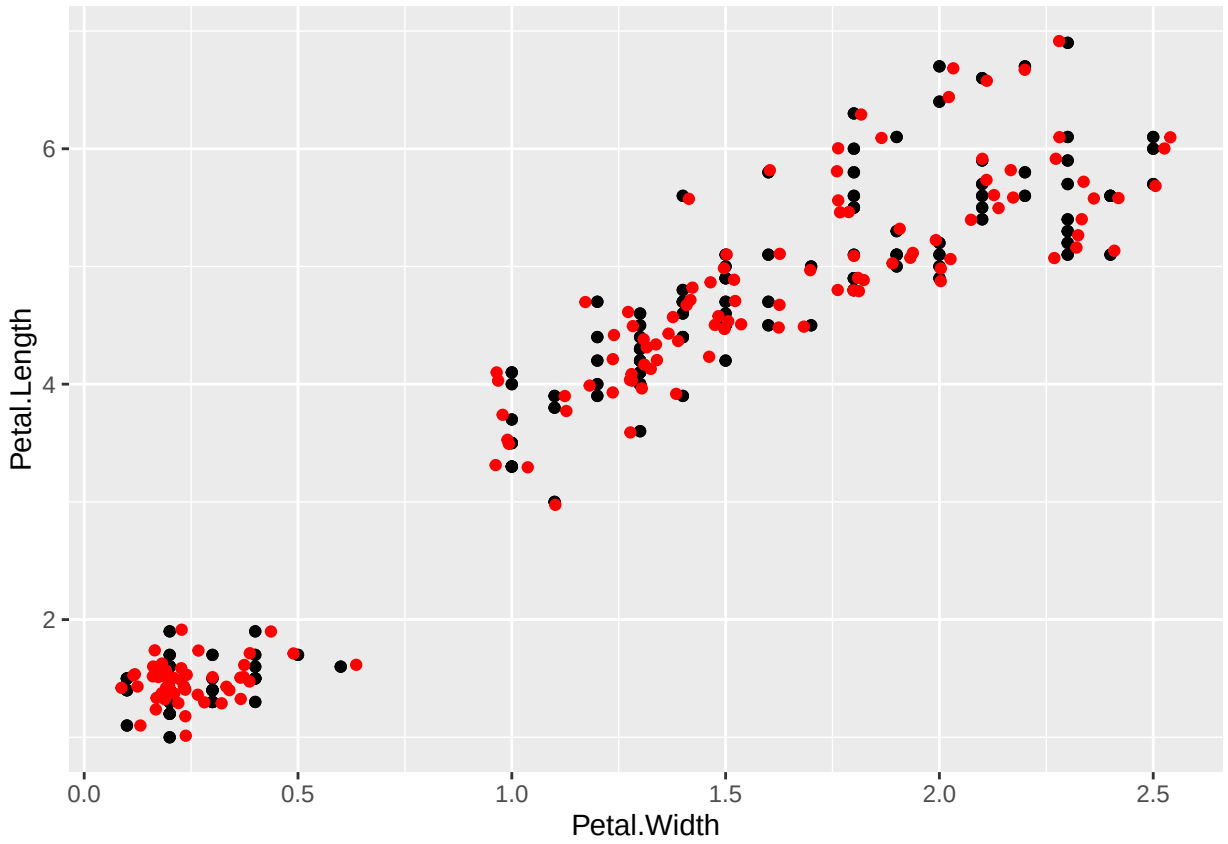
In the `aes()` function we can do the same. We can leave out the `x=` and `y=` parts and instead use the first two things in the function call for the x and y axis.

So `aes(Petal.Length, Petal.Width)` is the same as `aes(x=Petal.Length, y=Petal.Width)`

5.3.1 geom_jitter()

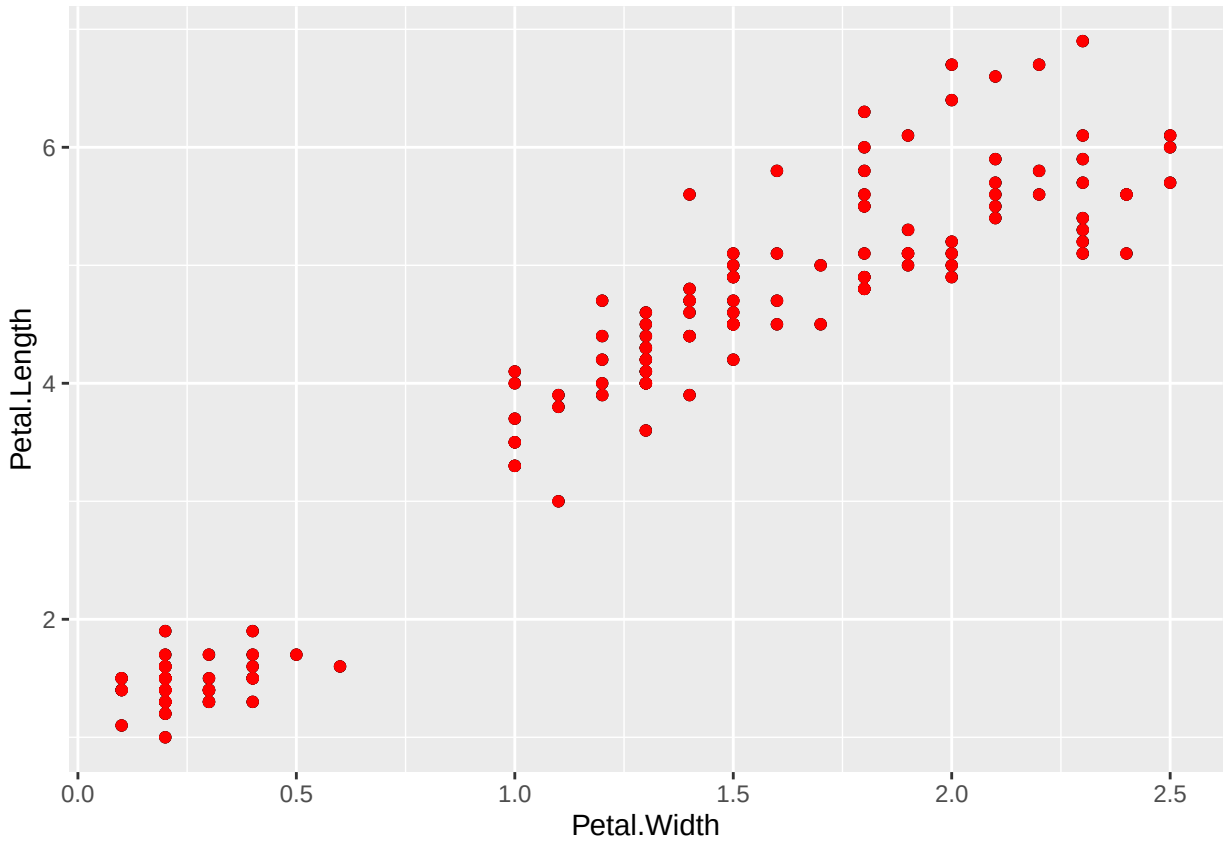
Sometimes, point plots get crowded, the points can get too close together, a visual problem called overplotting. A jitter plot lets us get over this by adding a random bit of noise to the position of the points. Here the points from the jitter geom are set to red.

```
p + geom_point() + geom_jitter(colour="Red")
```



We can fiddle with the range of the jitter with `width` and `height` options

```
p + geom_point() + geom_jitter(colour="Red", width=0.001, height=0.001)
```



conversely,

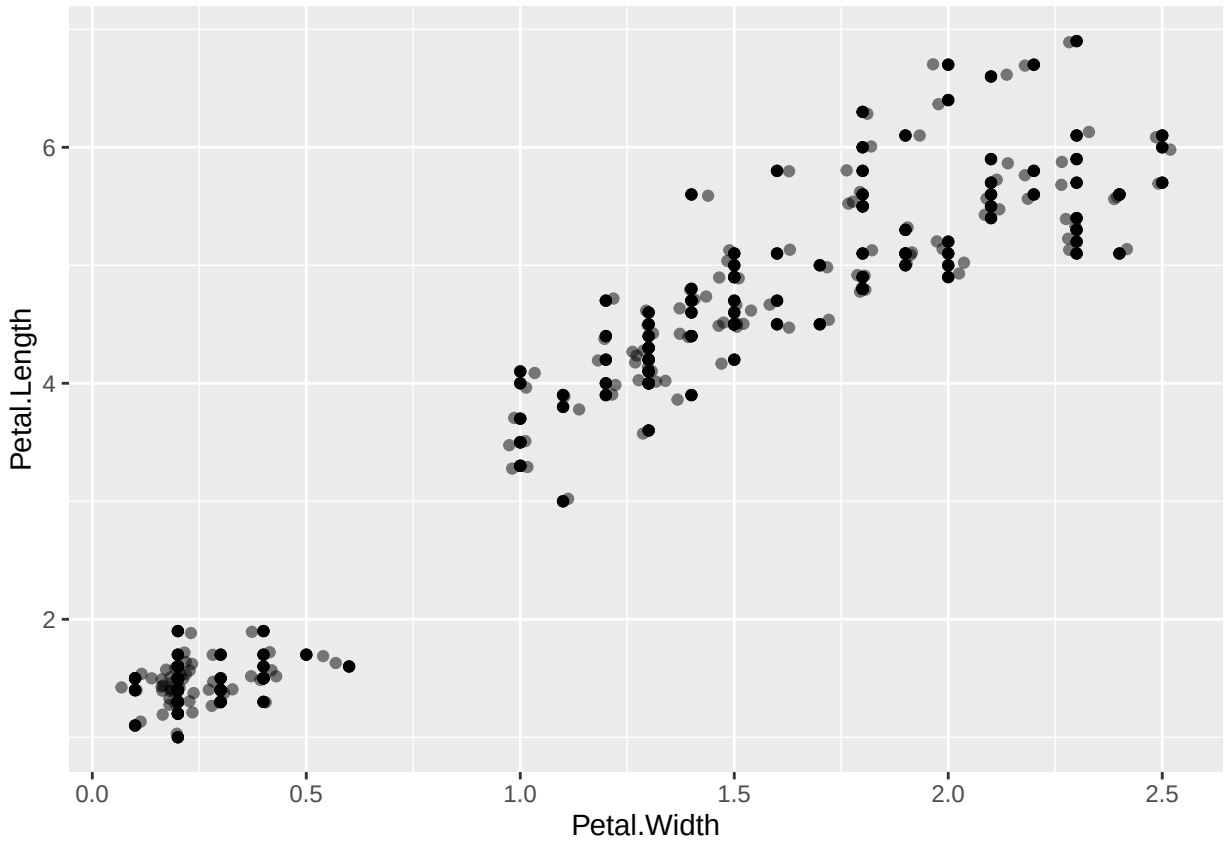
```
p + geom_point() + geom_jitter(colour="Red", width=10, height=10)
```

The defaults are usually a good choice though.

5.3.2 Changing opacity

Overplotting can be dealt with in other ways, changing the opacity of the geom is another. This is the `alpha` option. Choose the value in the range `0` to `1`, where `0` is invisible and `1` is solid

```
p + geom_point() + geom_jitter(alpha=0.5)
```

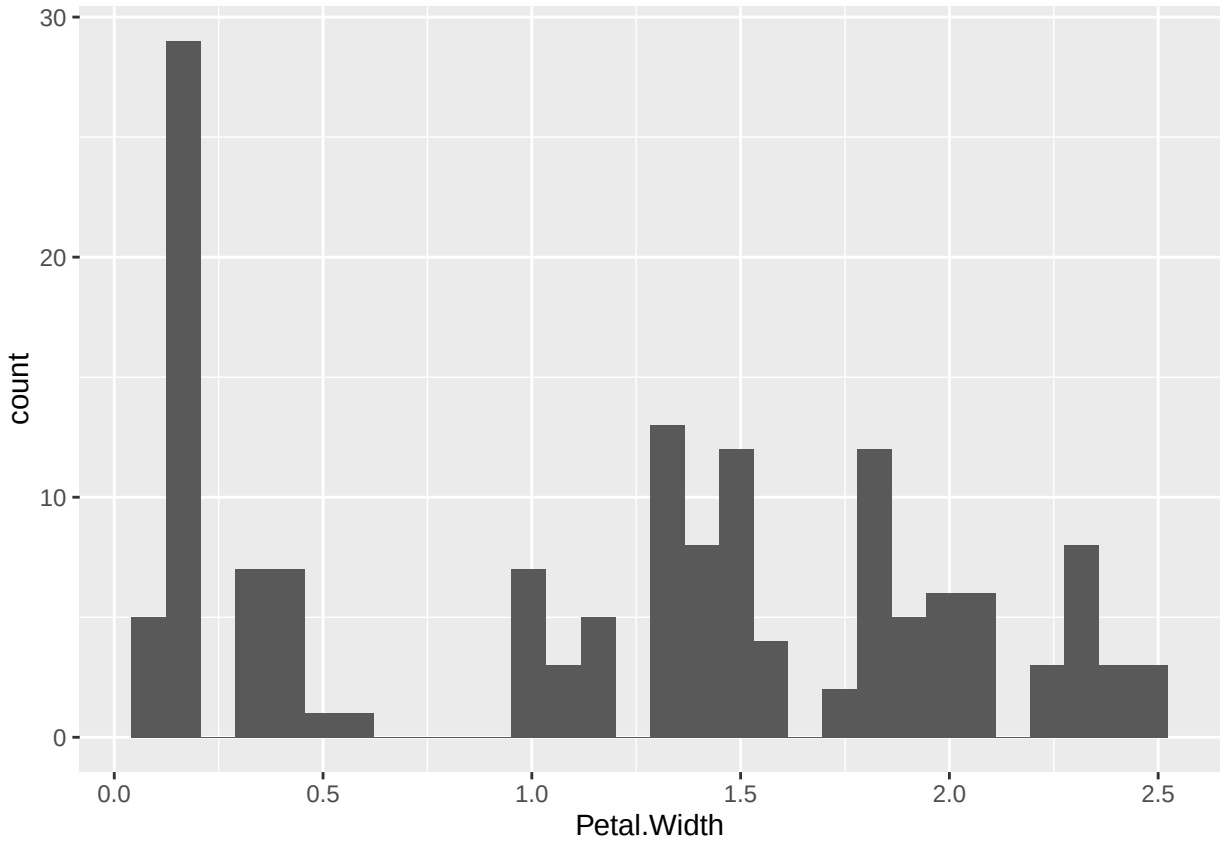


5.3.3 geom_histogram()

Plotting a histogram is done with the `geom_histogram()`. The y value for this is calculated automatically, so you provide the x value.

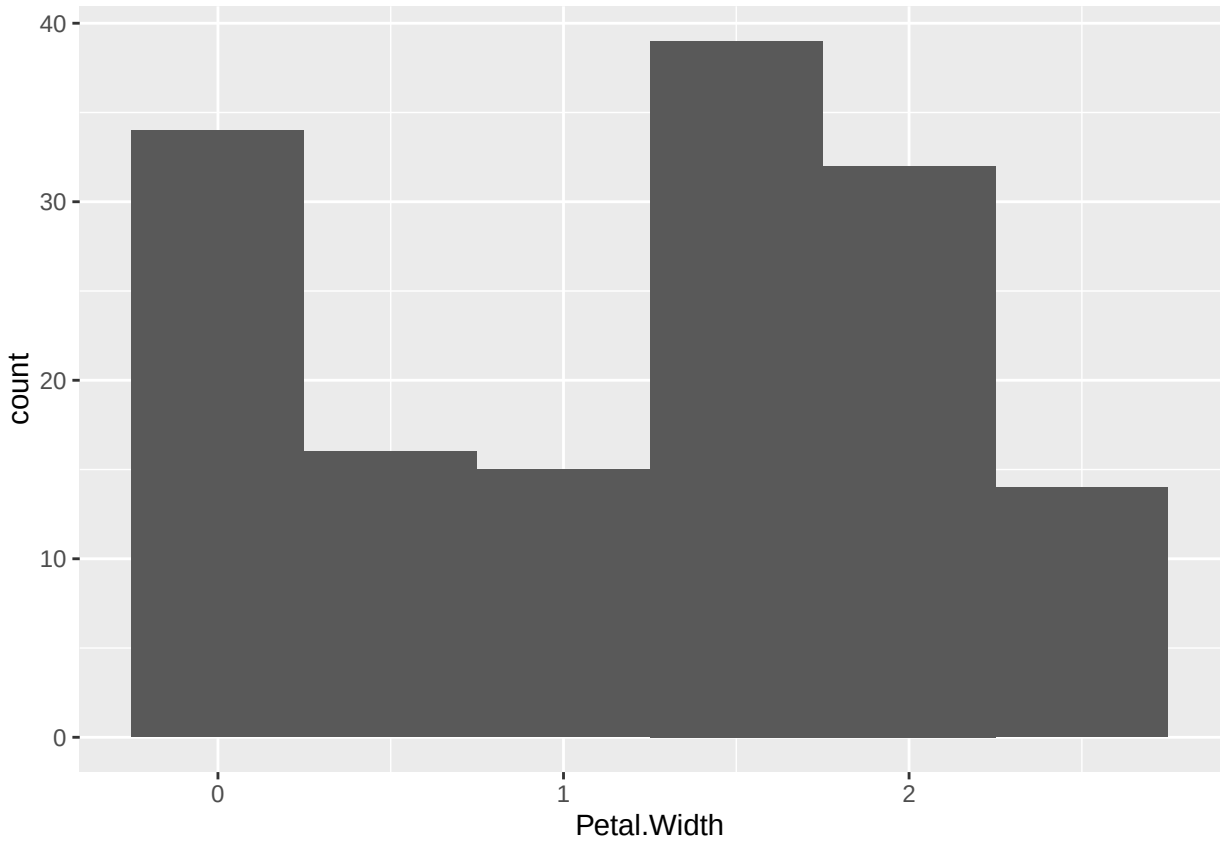
```
p <- ggplot(iris) + aes(Petal.Width)
p + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

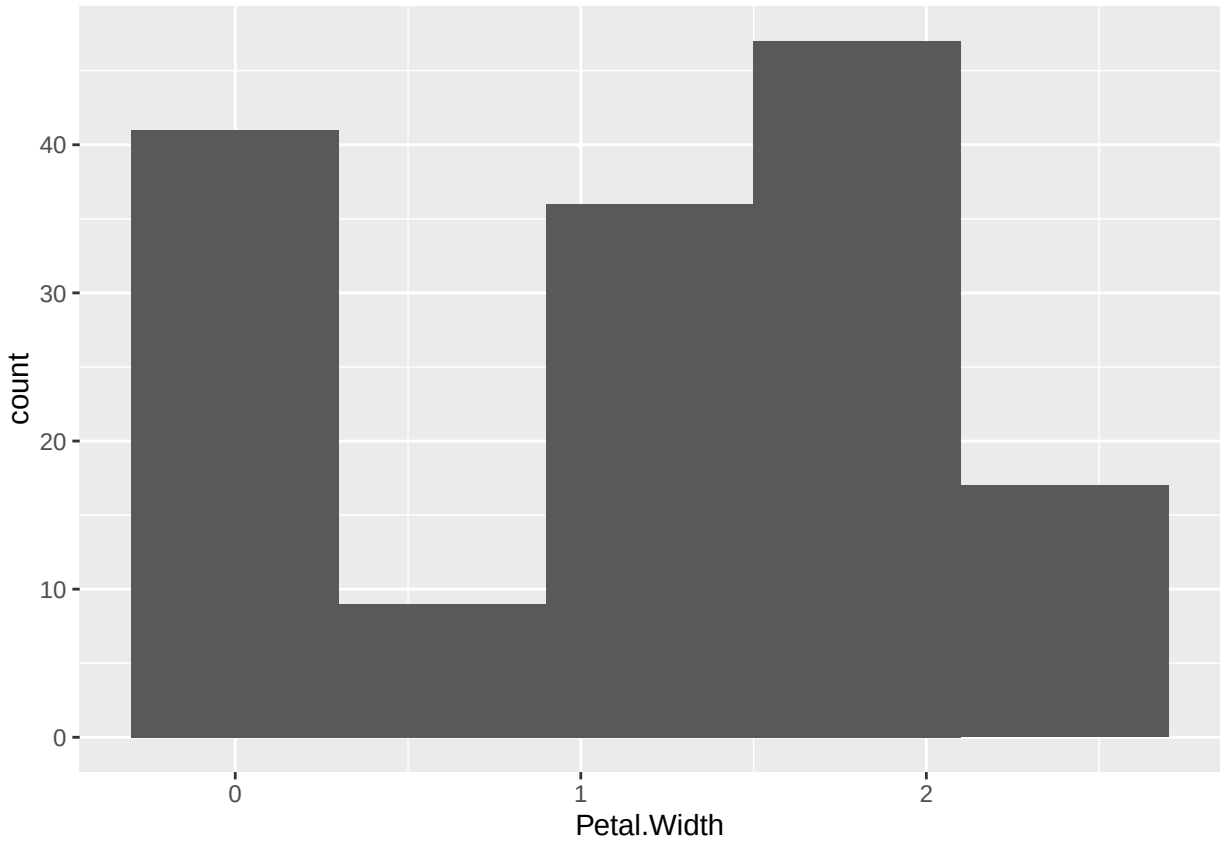


You can change the width of the bins with `binwidth`, or set the number of bins with `bins`

```
p + geom_histogram(binwidth=0.5)
```

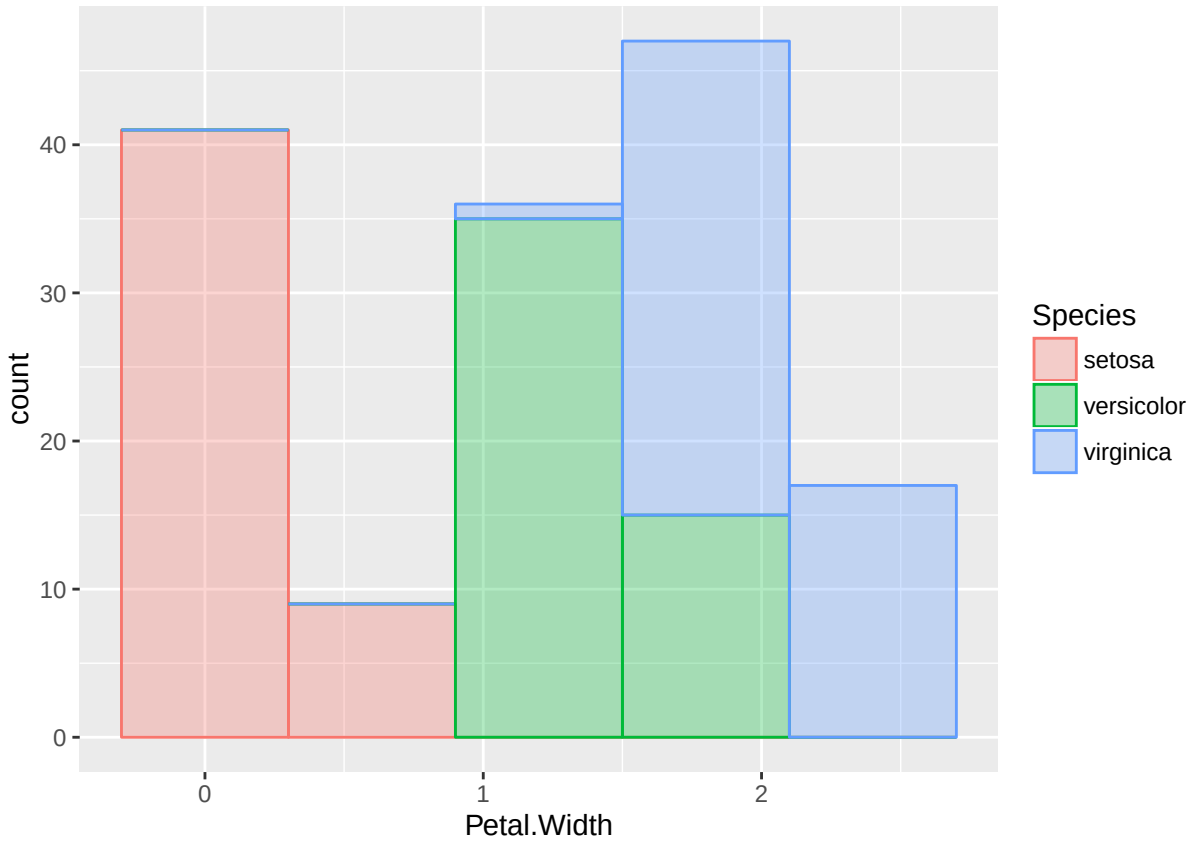



```
p + geom_histogram(bins=5)
```



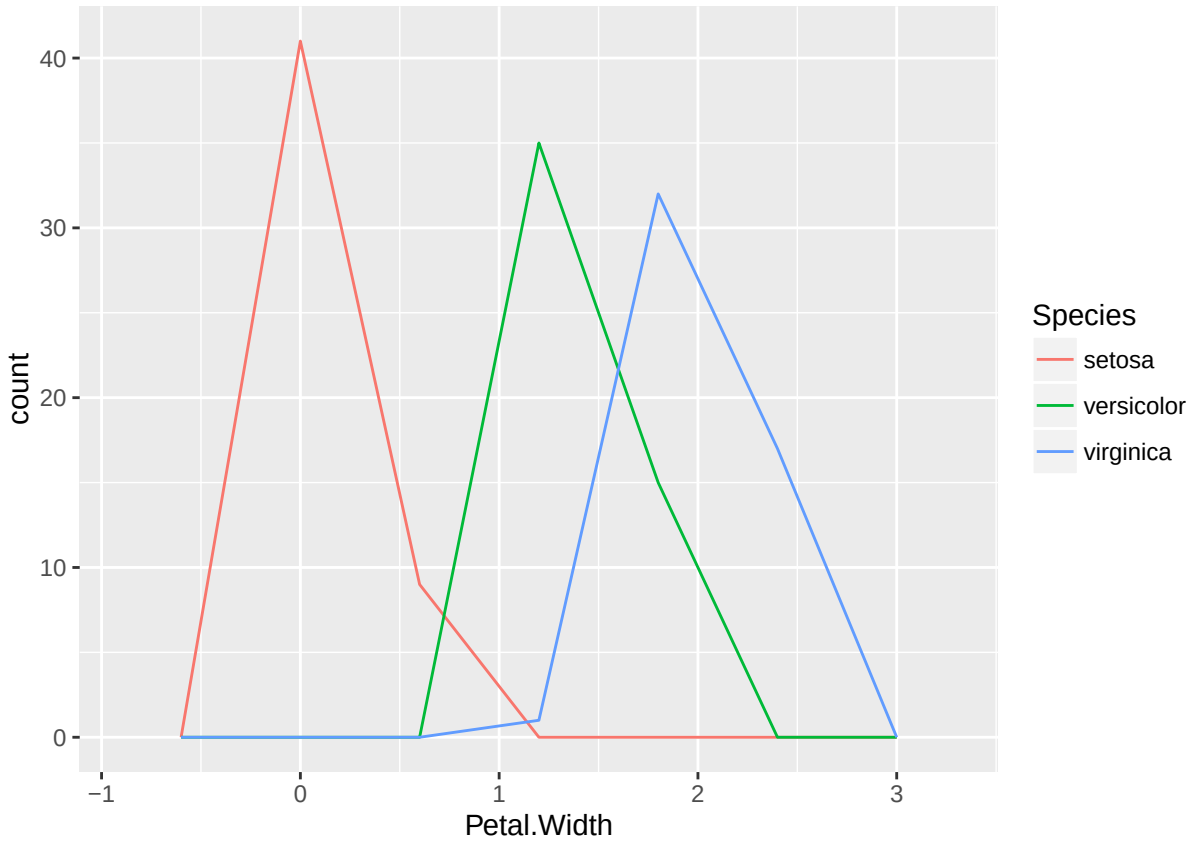
Trying to map the species to colour in this one gives us a weird sort of stacked histogram.

```
p + geom_histogram(bins=5, aes(colour=Species, fill=Species),alpha=0.3 )
```



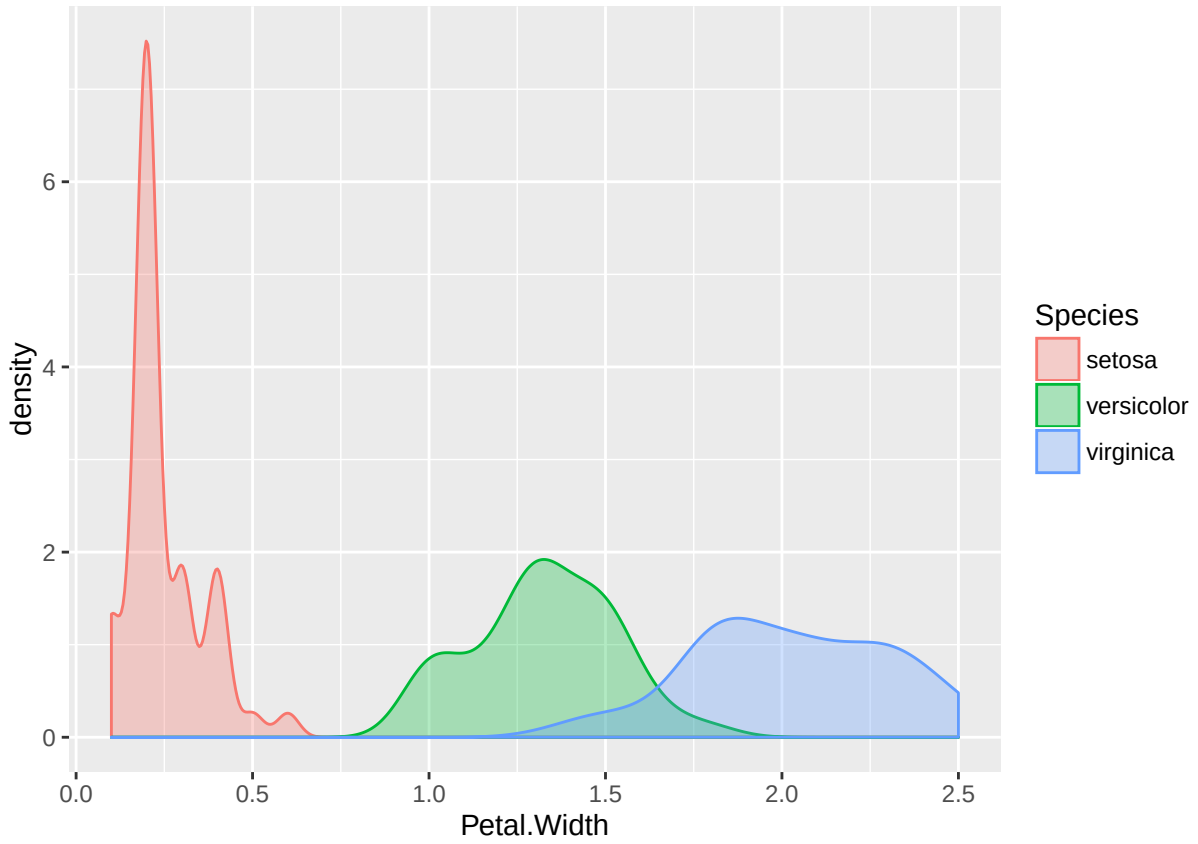
We can avoid this in a few ways, one is by using `geom_freqpoly()`, which is a line graph joining the tops of the bars of the histogram.

```
p + geom_freqpoly( aes(colour=Species), bins=5 )
```



or with `geom_density()` which gives us smoothed lines from a kernel density estimate of the data (which is a way of generating a smooth curve over histograms).

```
p + geom_density( aes(colour=Species, fill=Species), alpha=0.3)
```



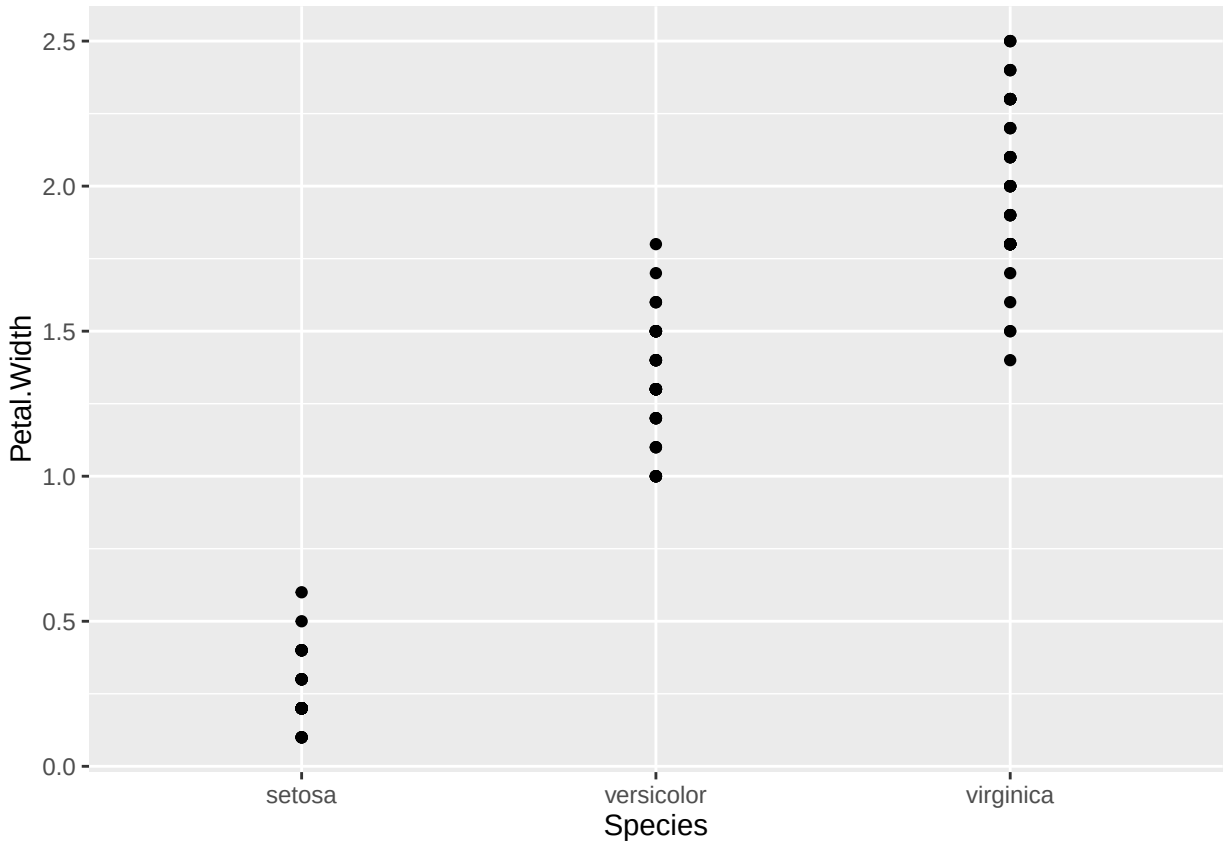
5.4 Discrete geoms

Let's look at some geoms with categories on the x and numbers on the y axis.

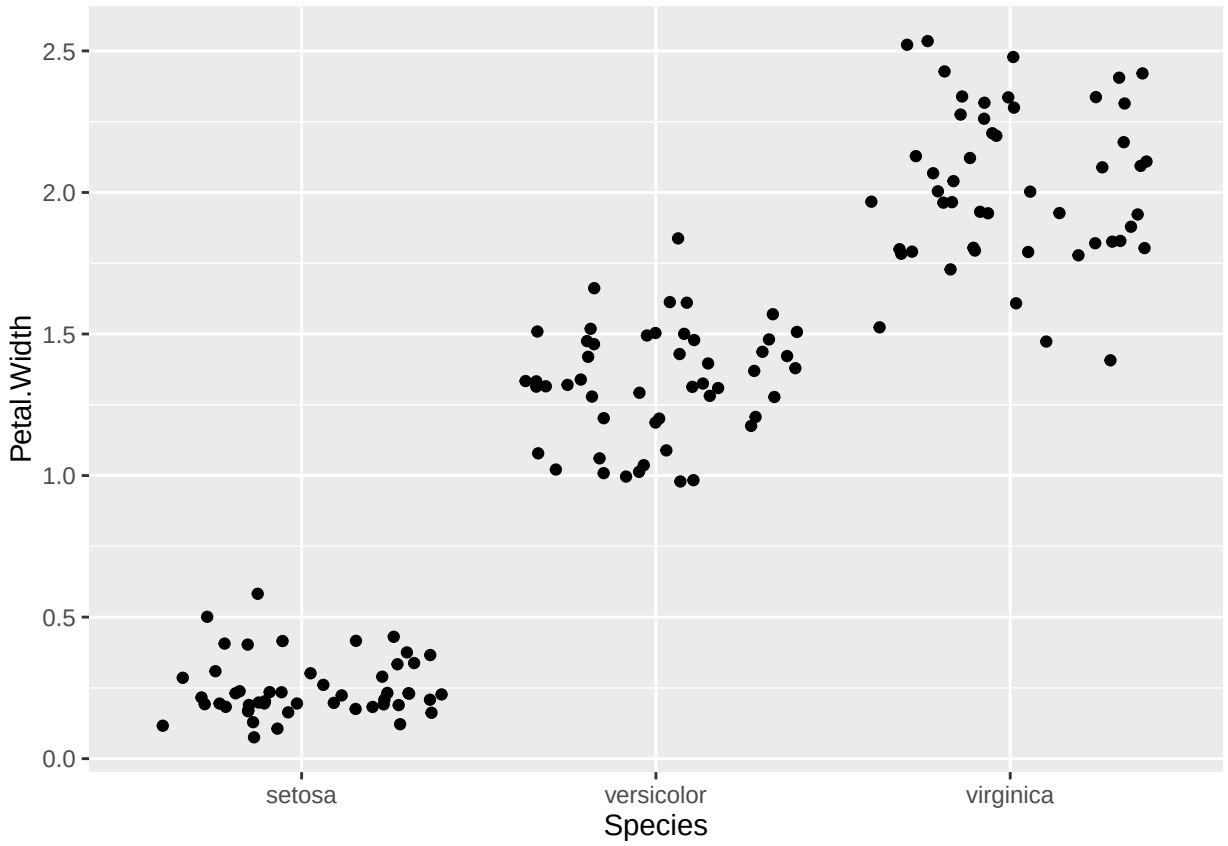
5.4.1 geom_point() and geom_jitter()

Both these geoms can be used with categoric data in one dimension. This is a useful and very honest way of showing your data points.

```
p <- ggplot(iris) + aes(x=Species, y=Petal.Width)
p + geom_point()
```



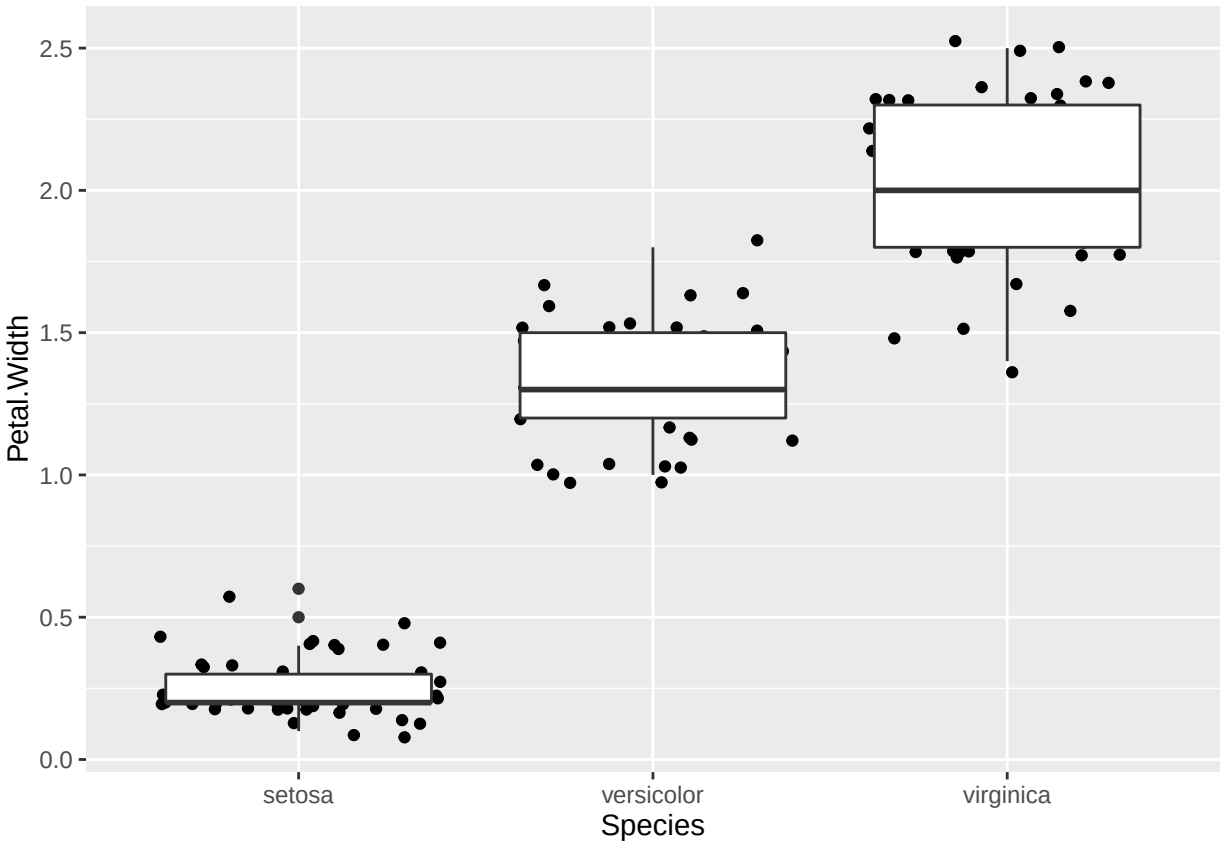

```
p + geom_jitter()
```



5.4.2 geom_boxplot() and geom_violin()

A great way to summarise the distributions of points is to use a boxplot in conjunction with the dots.

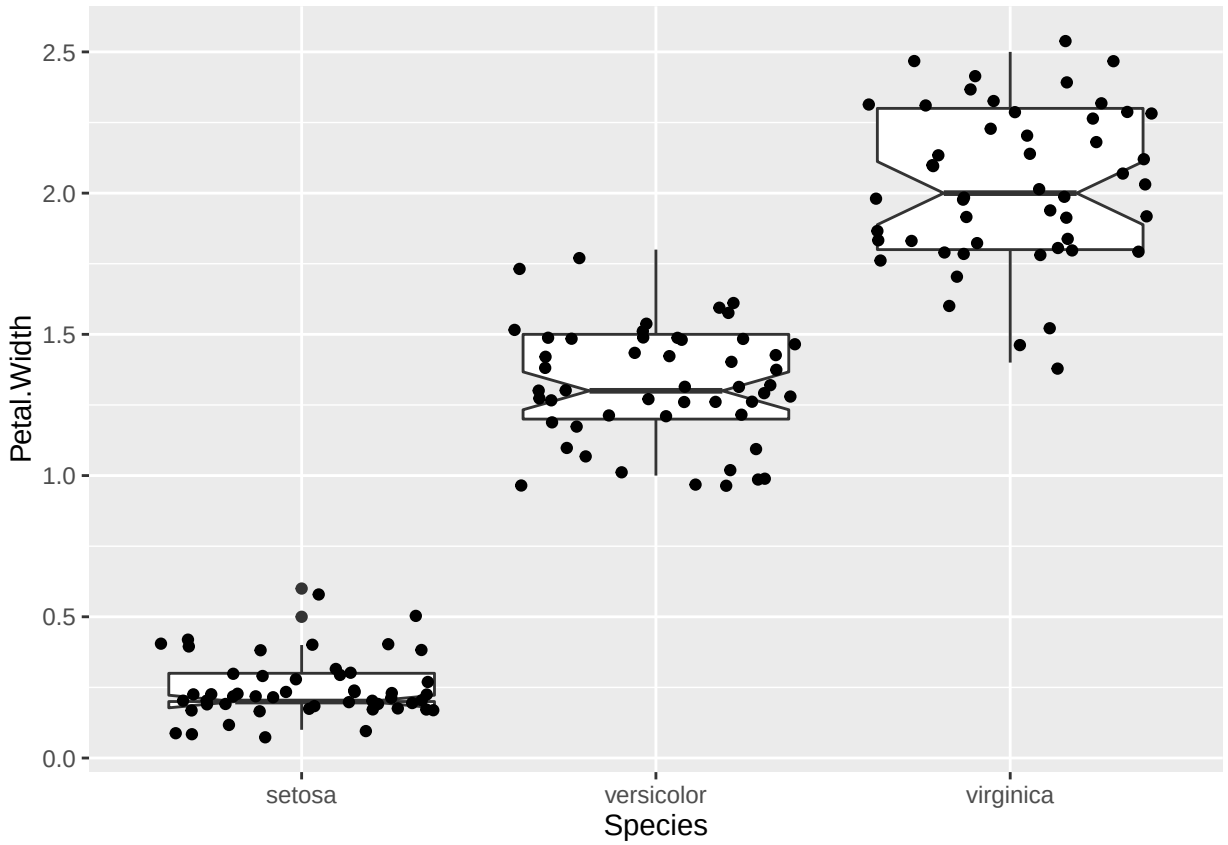
```
p <- ggplot(iris) + aes(x=Species, y=Petal.Width)
p + geom_jitter() + geom_boxplot()
```



Which unhelpfully puts the newest layer on top. Reverse the order to see the points

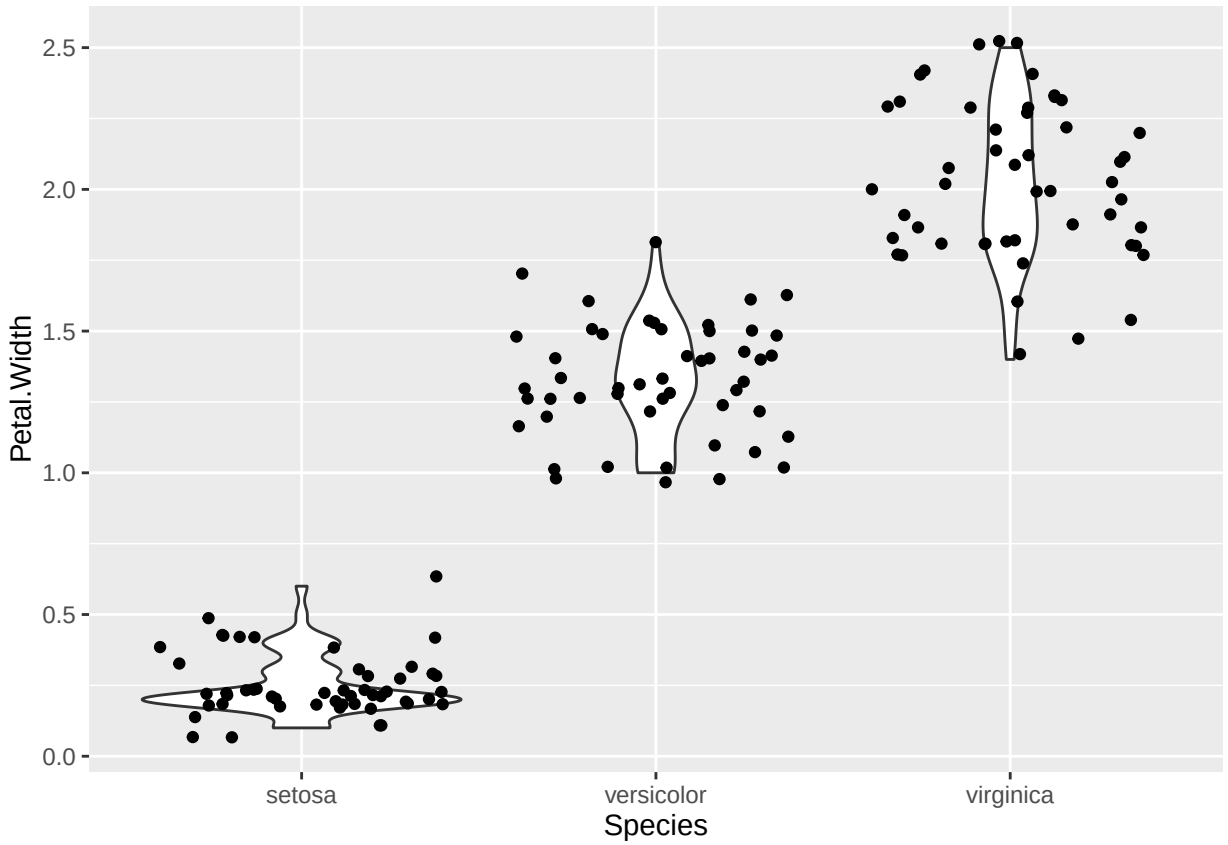
```
p + geom_boxplot(notch=TRUE) + geom_jitter()
```

```
## notch went outside hinges. Try setting notch=FALSE.
```



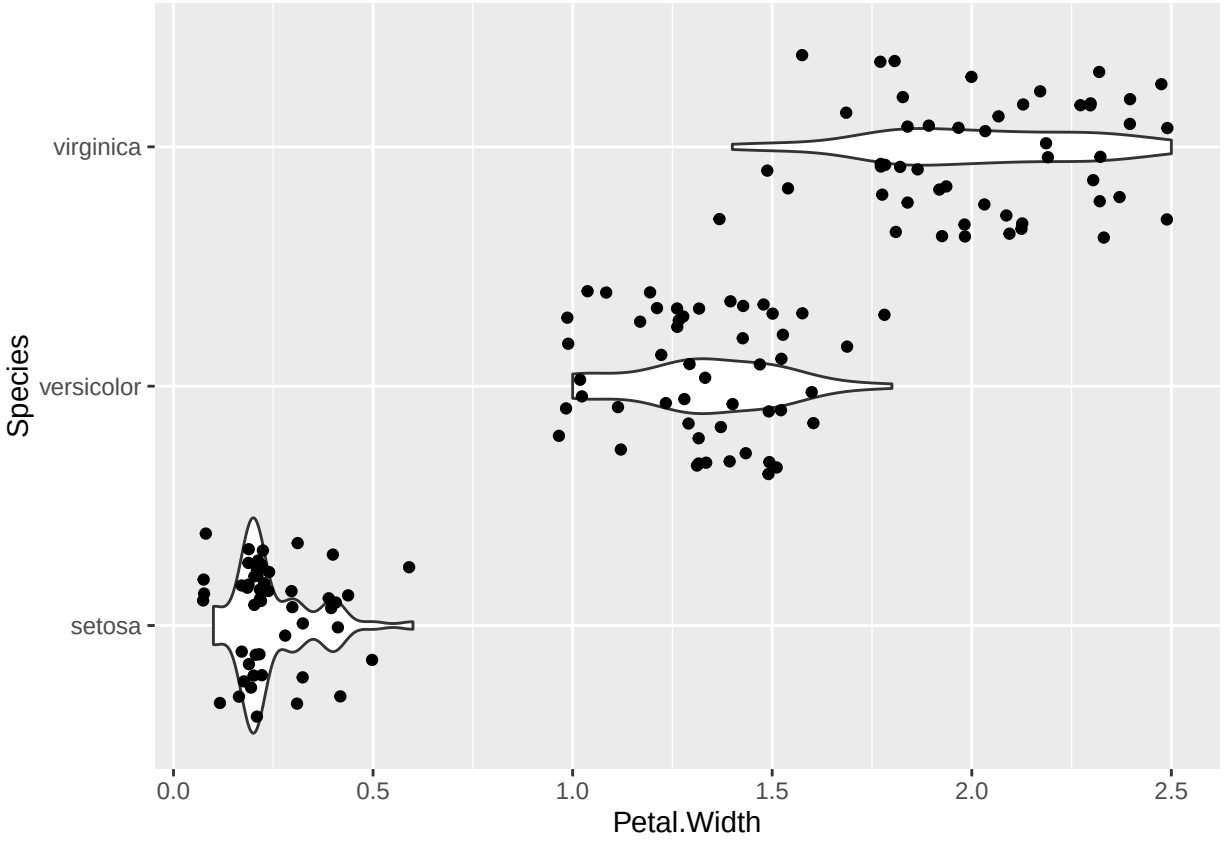
A common reason for using the boxplot is to use those notches to show the significant differences in the data. But really, these only help you assess a difference meaningfully if the data are normally distributed. In other circumstances you should be aware that the notches are misleading. Instead you can see the spread of your data much better with a violin plot.

```
p <- ggplot(iris) + aes(x=Species, y=Petal.Width)
p + geom_violin() + geom_jitter()
```



By turning your head to the side you can see the histogram curve / density distribution a bit more clearly. In fact ggplot has a way to flip a plot, one of a set of things called a transformation.

```
p + geom_violin() + geom_jitter() + coord_flip()
```



Now you can see clearly that the setosa numbers are really badly bunched down at the lower end and a bit skewed by that.

5.5 Boxplots are best for normally distributed data.

Really, these boxplots, especially the ones with the notches only help you assess a difference if the data is nicely normally distributed ggplot

5.6 Quiz

1. Incorporate a jitter and notched boxplot into the Petal.Width and Species plots we already used: `ggplot(iris) + aes(Species,Petal.Width)...`

6 Using Factors to Subset Data and Plots

6.1 About this chapter

1. Questions:

- How can I make plots that compare multiple categories?"

2. Objectives:

- Understand factors

- Understand colouring and faceting on factors
- Use factors for summaries and plot design

3. Keypoints:

- A factor is a value of a categorical variable, or the different values a label can take
- Factors are needed to subset and add attributes to data dynamically

6.2 Factors

In previous plots we've been using categories, specifically the `Species` category to split our data, colour our plots etc. These categorical columns are called Factors in R. Looking at the `diamonds` data set we can see how this is set up in R.

```
head(diamonds)
```

```
## # A tibble: 6 × 10
##   carat       cut color clarity depth table price     x     y     z
##   <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23     Ideal     E     SI2  61.5    55   326  3.95  3.98  2.43
## 2  0.21   Premium     E     SI1  59.8    61   326  3.89  3.84  2.31
## 3  0.23      Good     E     VS1  56.9    65   327  4.05  4.07  2.31
## 4  0.29   Premium     I     VS2  62.4    58   334  4.20  4.23  2.63
## 5  0.31      Good     J     SI2  63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good     J    VVS2  62.8    57   336  3.94  3.96  2.48
```

Here we can see the `cut`, `color` and `clarity` columns are all non-numeric, textual data. These are the factor variables of this dataset. We can confirm that by asking for the `class` of the column, that is, the type of data in it. We use the dataset `$` column name syntax for this.

```
class(diamonds$color)
```

```
## [1] "ordered" "factor"
```

```
class(diamonds$depth)
```

```
## [1] "numeric"
```

We can also ask for all the different values of the factor, in R called the levels

```
levels(diamonds$color)
```

```
## [1] "D" "E" "F" "G" "H" "I" "J"
```

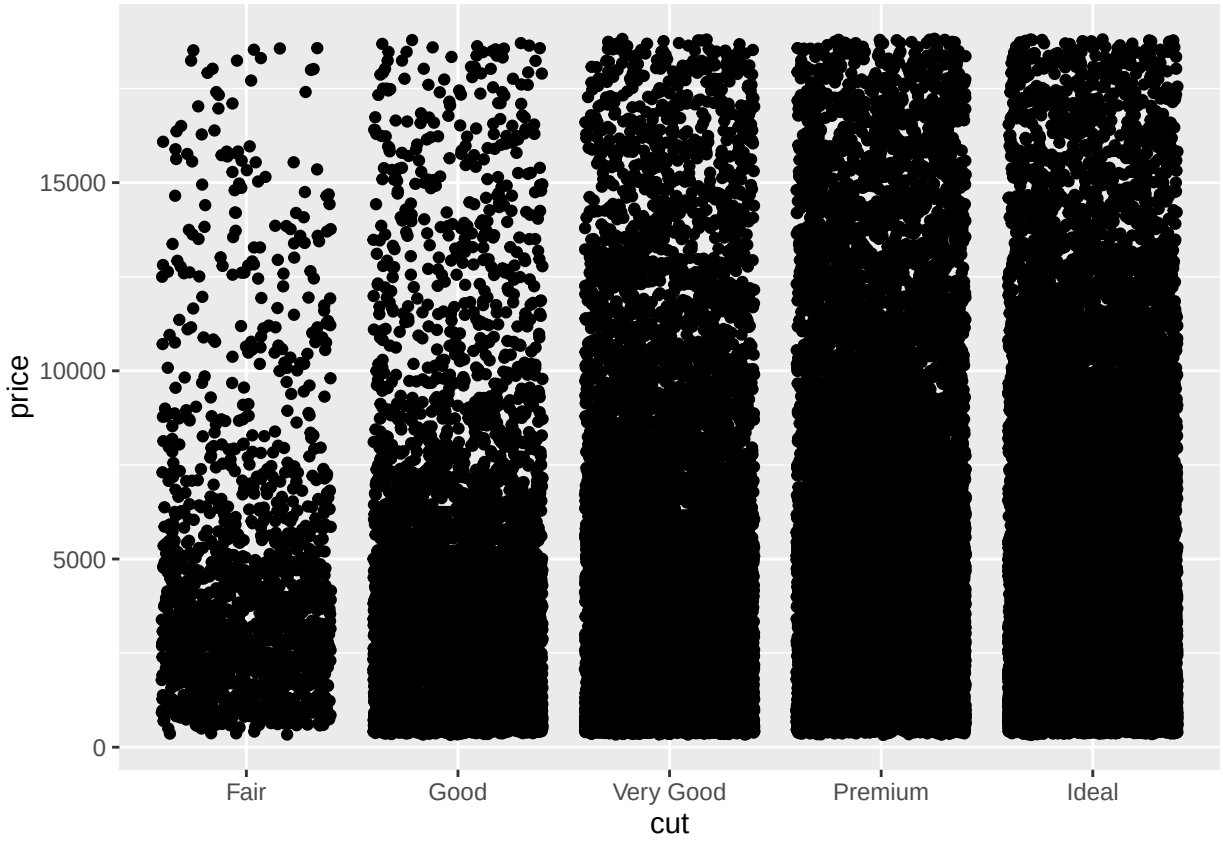
```
levels(diamonds$cut)
```

```
## [1] "Fair"      "Good"      "Very Good" "Premium"   "Ideal"
```

6.3 Colouring by factors

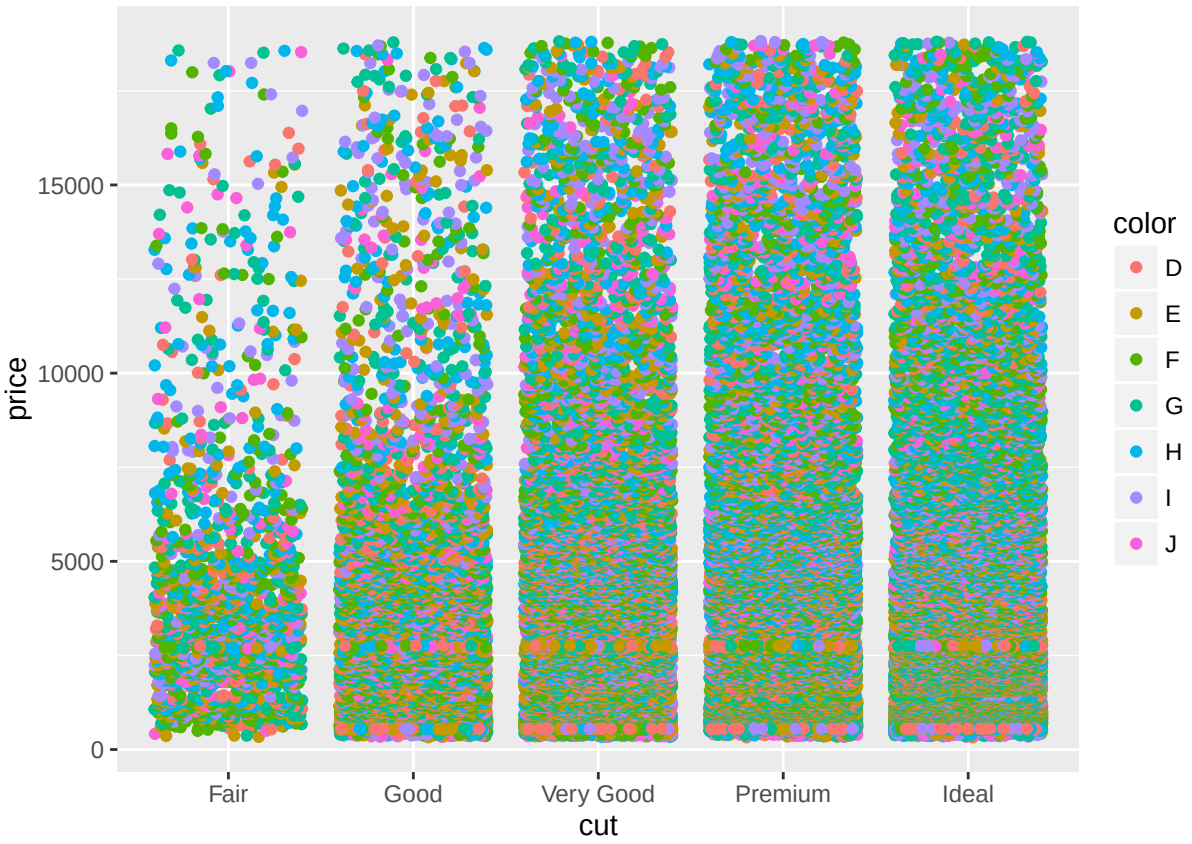
Let's look at applying mappings by a factor. Let's look at how price varies by cut.

```
p <- ggplot(diamonds) +  aes(cut,price)
p + geom_jitter()
```



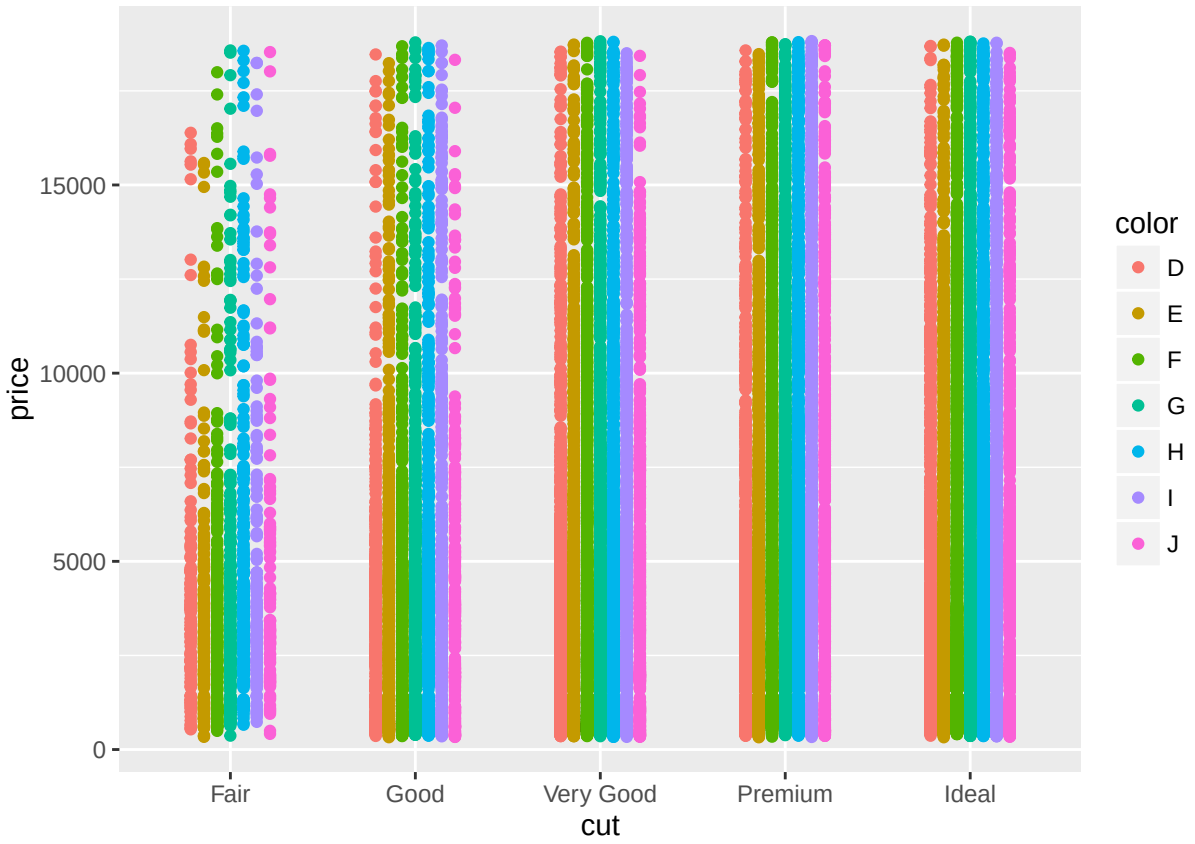
Now let's throw a second variable in there, lets see how color varies within each cut. We do this by creating a new aesthetic mapping within the geom_jitter()

```
p + geom_jitter(aes(colour=color))
```



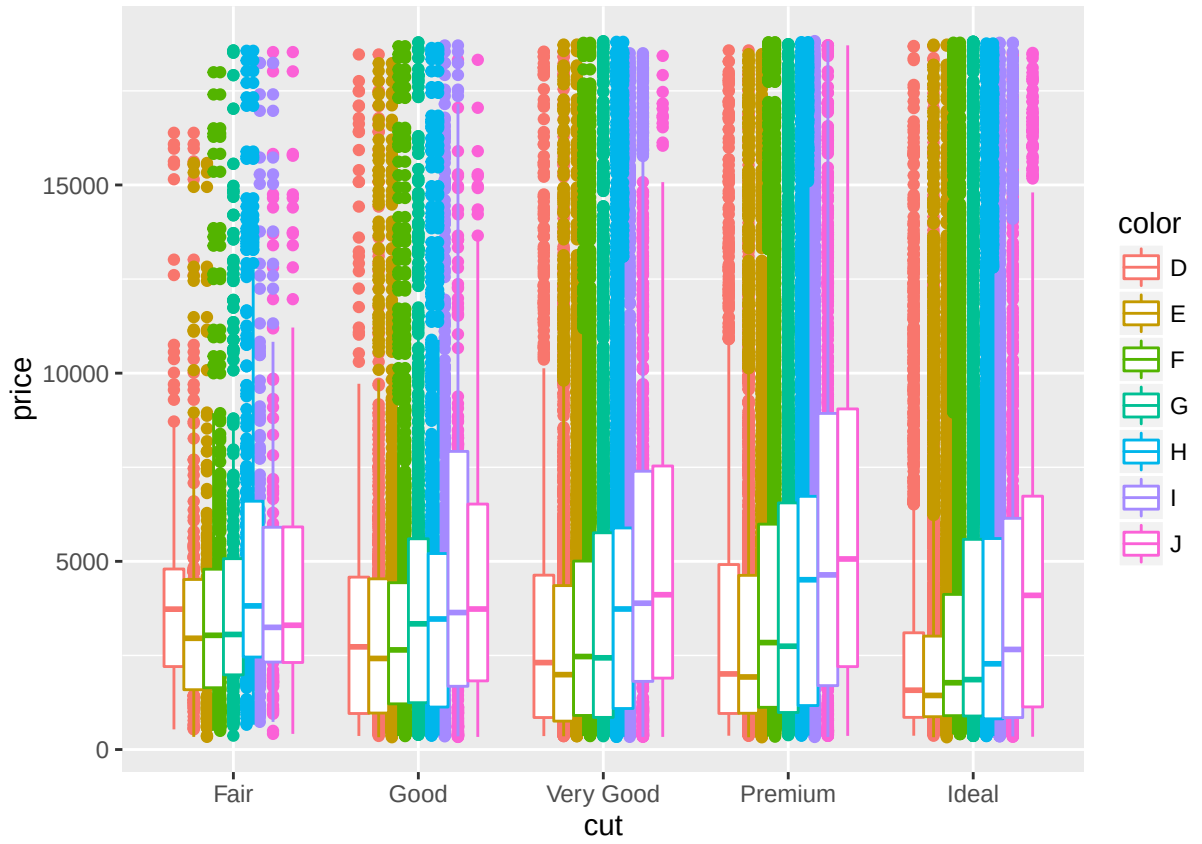
The spots are all overlapping, we can force the different colours to stay separate with the `position` option. We use `position_dodge()` to make them dodge each other. The width option tells the spots how far to stay apart.

```
p + geom_jitter(aes(colour=color), position=position_dodge(width=0.5) )
```



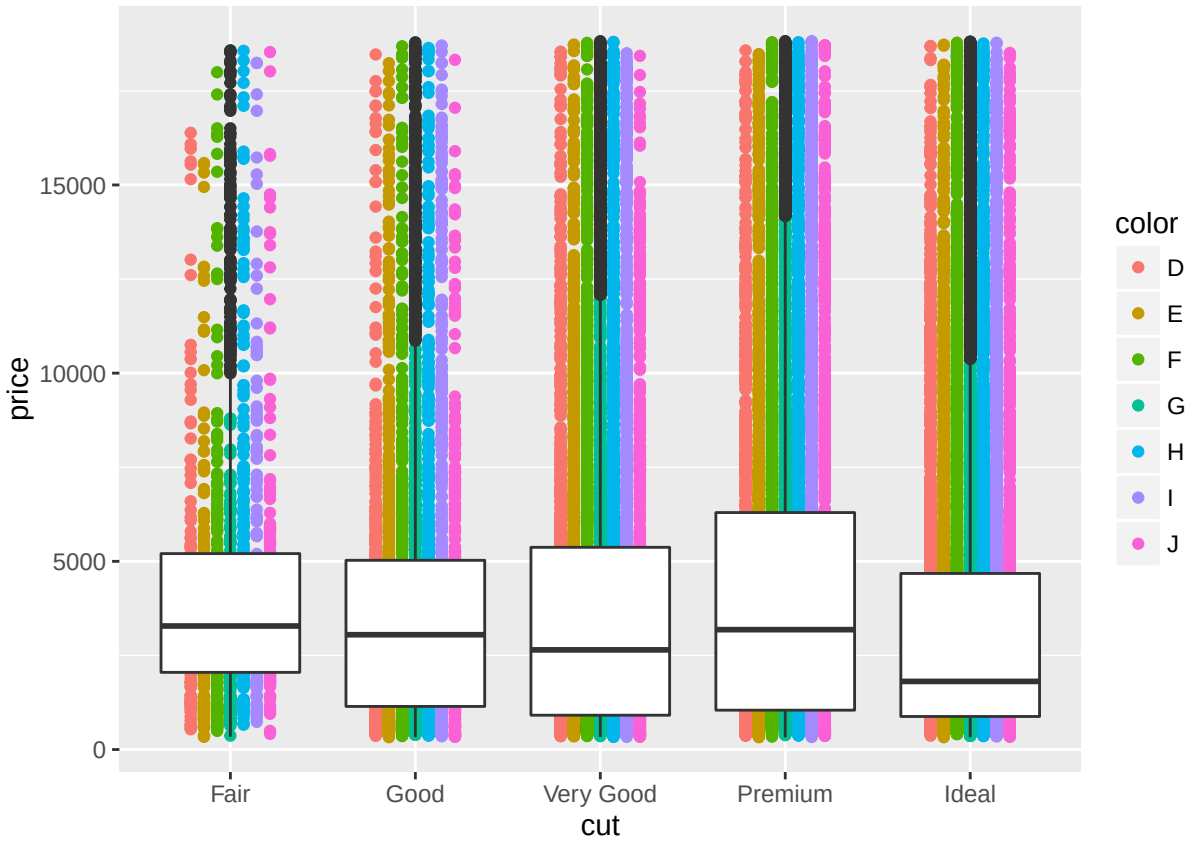
We can also throw other geoms on top in the same way. EG Boxplots for each cut and colour

```
p + geom_jitter(aes(colour=color), position=position_dodge(width=0.5) ) + geom_boxplot( aes(colour=colo
```



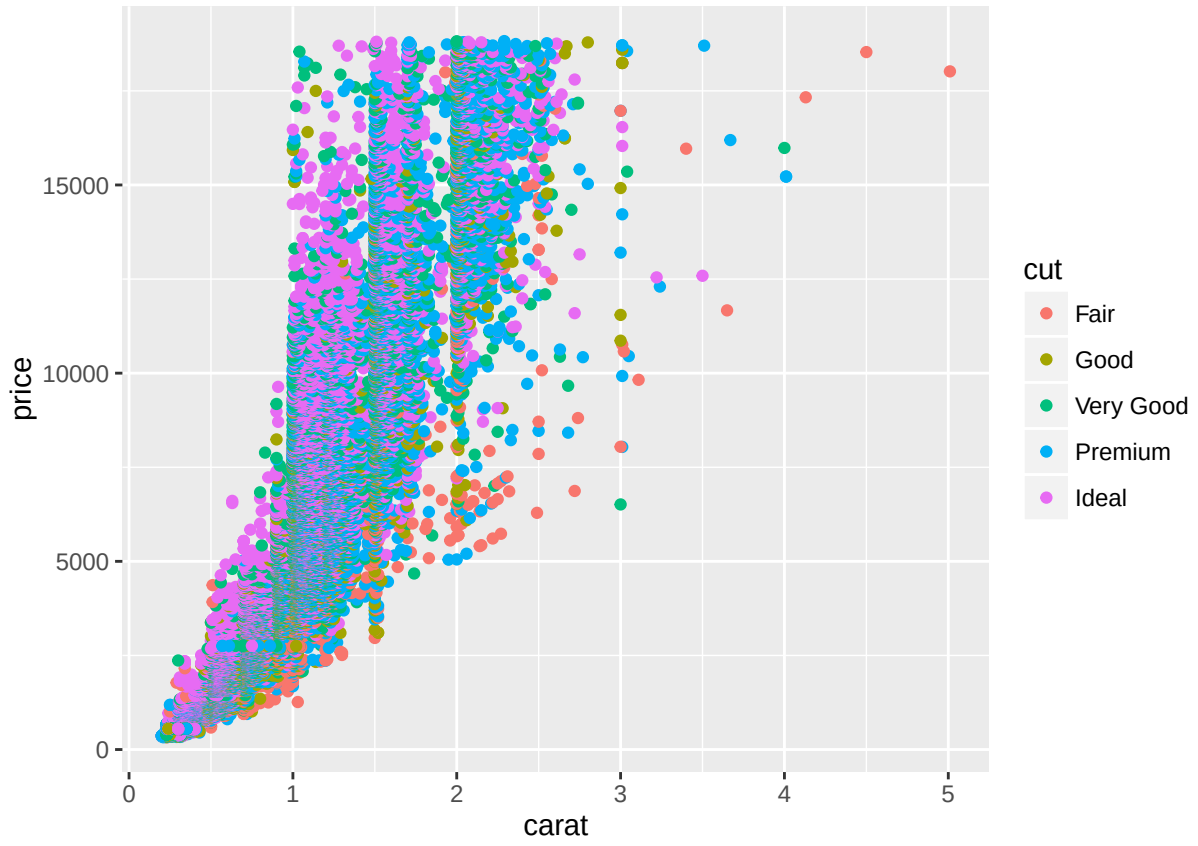
Remember layers/geoms are independent, so can be set up to show individual aspects of the data. Let's have a boxplot for the whole of the cut, irrespective of the colour.

```
p + geom_jitter(aes(colour=color),position=position_dodge(width=0.5)) + geom_boxplot()
```



And of course, the whole thing still works even if we are comparing two numerical columns. We can still use the aesthetic mapping in the geom to colour our points by a factor

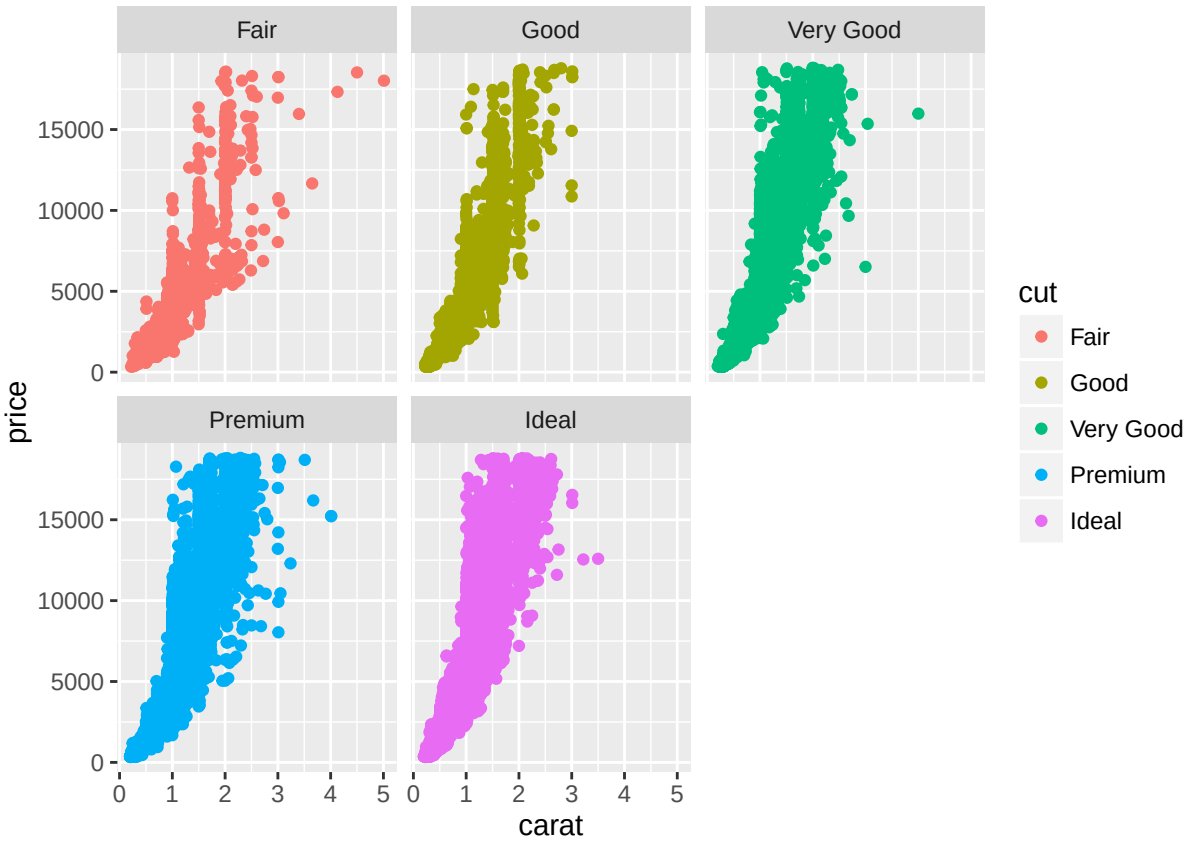
```
ggplot(diamonds) + aes(carat, price) + geom_point(aes(colour=cut))
```



6.4 Small multiple plots

Sometimes, trying to squeeze a lot of data into one plot isn't the clearest way to show it. Instead small multiple plots (different data, same settings) can be used. In ggplot, this is called faceting and is done with the `facet_wrap()` or `facet_grid()` function. We use the factors to define the facet. Let's add faceting to the previous plot

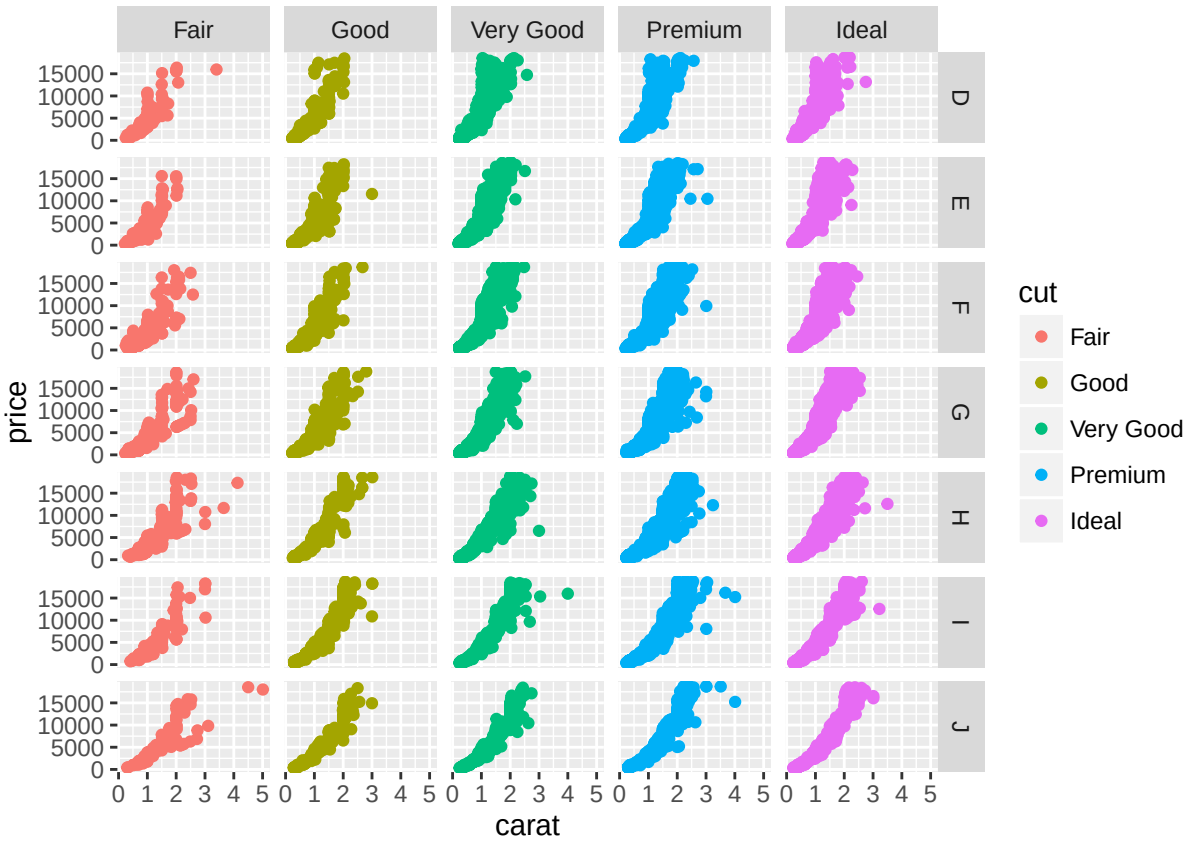
```
p <- ggplot(diamonds) + aes(carat, price)
p + geom_point(aes(colour=cut)) + facet_wrap( ~ cut)
```



Here we see the plot is divided into panels, one for each 'cut'. The `facet_wrap()` function puts all the panels into a single row, but wll wrap that row as space demands. The syntax is a bit odd, we used the `~` operator to mean 'varies by' , even though we only used one variable. It's just a quirk of ggplot.

The `facet_grid()` function forces a grid structure and can take more than one factor. Now the `~` 'varies by' syntax makes more sense:

```
p + geom_point(aes(colour=cut)) + facet_grid(color ~ cut)
```



6.5 Summary Statistics

Factors are powerful things for helping us to quickly get summary statistics, and not just plots out of the data. We already saw how to generate summary statistics on a whole dataset using the `summary()` function.

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

But a better way to summarise by factor is with the `describeBy()` function in the `psych` package. Note you need to use `$` notation to describe the column with the factor you want to subset with.

```
library(psych)
describeBy(iris, iris$Species)
```

```
## $setosa
##              vars  n mean   sd median trimmed  mad min max range skew
## Sepal.Length    1 50 5.01 0.35    5.0    5.00 0.30 4.3 5.8   1.5 0.11
## Sepal.Width     2 50 3.43 0.38    3.4    3.42 0.37 2.3 4.4   2.1 0.04
## Petal.Length    3 50 1.46 0.17    1.5    1.46 0.15 1.0 1.9   0.9 0.10
## Petal.Width     4 50 0.25 0.11    0.2    0.24 0.00 0.1 0.6   0.5 1.18
## Species*        5 50 1.00 0.00    1.0    1.00 0.00 1.0 1.0   0.0  NaN
##              kurtosis   se
## Sepal.Length    -0.45 0.05
## Sepal.Width      0.60 0.05
## Petal.Length     0.65 0.02
## Petal.Width      1.26 0.01
## Species*          NaN 0.00
##
## $versicolor
##              vars  n mean   sd median trimmed  mad min max range  skew
## Sepal.Length    1 50 5.94 0.52   5.90    5.94 0.52 4.9 7.0   2.1  0.10
## Sepal.Width     2 50 2.77 0.31   2.80    2.78 0.30 2.0 3.4   1.4 -0.34
## Petal.Length    3 50 4.26 0.47   4.35    4.29 0.52 3.0 5.1   2.1 -0.57
## Petal.Width     4 50 1.33 0.20   1.30    1.32 0.22 1.0 1.8   0.8 -0.03
## Species*        5 50 2.00 0.00   2.00    2.00 0.00 2.0 2.0   0.0   NaN
##              kurtosis   se
## Sepal.Length    -0.69 0.07
## Sepal.Width     -0.55 0.04
## Petal.Length    -0.19 0.07
## Petal.Width     -0.59 0.03
## Species*          NaN 0.00
##
## $virginica
##              vars  n mean   sd median trimmed  mad min max range  skew
## Sepal.Length    1 50 6.59 0.64   6.50    6.57 0.59 4.9 7.9   3.0  0.11
## Sepal.Width     2 50 2.97 0.32   3.00    2.96 0.30 2.2 3.8   1.6  0.34
## Petal.Length    3 50 5.55 0.55   5.55    5.51 0.67 4.5 6.9   2.4  0.52
## Petal.Width     4 50 2.03 0.27   2.00    2.03 0.30 1.4 2.5   1.1 -0.12
## Species*        5 50 3.00 0.00   3.00    3.00 0.00 3.0 3.0   0.0   NaN
##              kurtosis   se
## Sepal.Length    -0.20 0.09
## Sepal.Width      0.38 0.05
## Petal.Length    -0.37 0.08
## Petal.Width     -0.75 0.04
## Species*          NaN 0.00
##
## attr(,"call")
## by.data.frame(data = x, INDICES = group, FUN = describe, type = type)
```

With this you can get a nice, comprehensive table of summary statistics across all the numerical columns, divided by the chosen factor.

For combinations of factors, you can use the `ddply()` function in the `plyr` package. Here you can choose a list of factors to summarise, but you must name the output columns and the R function to use. Helpfully the R function for a mean is `mean()` and the function for standard deviation is `sd()`.

Here, we divide up on `cut` and `colour` using the make-a-list function `c()`, we tell `ddply` we want to `summarise` and that it should add a `mean` column using the `mean()` function and an `sd` column using the `sd(function)`

```
library(plyr)
ddply(diamonds, c('cut', 'color'), summarise, mean=mean(price), sd=sd(price) )
```

```
##          cut color     mean       sd
## 1       Fair     D 4291.061 3286.114
## 2       Fair     E 3682.312 2976.652
## 3       Fair     F 3827.003 3223.303
## 4       Fair     G 4239.255 3609.644
## 5       Fair     H 5135.683 3886.482
## 6       Fair     I 4685.446 3730.271
## 7       Fair     J 4975.655 4050.459
## 8       Good     D 3405.382 3175.149
## 9       Good     E 3423.644 3330.702
## 10      Good     F 3495.750 3202.411
## 11      Good     G 4123.482 3702.505
## 12      Good     H 4276.255 4020.660
## 13      Good     I 5078.533 4631.702
## 14      Good     J 4574.173 3707.791
## 15 Very Good     D 3470.467 3523.753
## 16 Very Good     E 3214.652 3408.024
## 17 Very Good     F 3778.820 3786.124
## 18 Very Good     G 3872.754 3861.375
## 19 Very Good     H 4535.390 4185.798
## 20 Very Good     I 5255.880 4687.105
## 21 Very Good     J 5103.513 4135.653
## 22   Premium     D 3631.293 3711.634
## 23   Premium     E 3538.914 3794.987
## 24   Premium     F 4324.890 4012.023
## 25   Premium     G 4500.742 4356.571
## 26   Premium     H 5216.707 4466.190
## 27   Premium     I 5946.181 5053.746
## 28   Premium     J 6294.592 4788.937
## 29     Ideal     D 2629.095 3001.070
## 30     Ideal     E 2597.550 2956.007
## 31     Ideal     F 3374.939 3766.635
## 32     Ideal     G 3720.706 4006.262
## 33     Ideal     H 3889.335 4013.375
## 34     Ideal     I 4451.970 4505.150
## 35     Ideal     J 4918.186 4476.207
```

6.6 Quiz

The built in dataset `CO2` describes measurement of CO2 uptake versus concentration for Quebec and Mississippi grasses in chilled and nonchilled tests. The dataset is as follows:

- `Type` is a factor column with two levels `Quebec` and `Mississippi`
- `Treatment` is a factor colum with two levels `nonchilled` and `chilled`
- `Uptake` is a numerical colum with CO2 uptake rate in micromoles per metre squared per second
- `Plant` is a factor with twelve levels, one for each individual plant assayed.

1. Create a plot with `geom_point()` that shows the Plant on the x-axis and the Uptake on the y-axis. Colour the points by 'Type' and `facet_wrap()` by `Treatment` to get a subplot for chilled and nonchilled.

7 Using RMarkdown for Reproducible Publishable Plots

7.1 About this chapter

1. Questions:

- How can I design a plot once and use it for many experiments?

2. Objectives:

- Use RMarkdown documents to build a plot.

3. Keypoints:

- Reproducible work is good work.
- R Markdown can helps us be reproducible and transparent

7.2 Being lazy is a virtue. Work hard to be lazy.

Writing reproducible code will save you time and effort. Computers are especially good at carrying out commands and if you are smart enough to put those commands in an executable document, rather than run the whole thing by hand every time, you'll save time, you can ensure that you'll do the same thing everytime and those who look at your work later will be absolutely clear about what you did.

Of course, this takes a little bit more effort up front, but it will pay off. And R Studio has plenty of ways to help you do just this. R Markdown documents are one such way. For the rest of this course we'll be putting our code into R Markdown.

7.3 R Markdown

Markdown is a way of adding little tags to text, to define parts of the structure of it, so that when a file written in Markdown is sent to a program that knows how to interpret it, the program can render the text as you intended.

R Studio has the ability to take a document written in Markdown, squeeze R code into it and produce the output in a pretty format. The flavour is called R Markdown. By combining this with our plotting knowledge, we can make a dynamic document that can be re-run every time we get a new dataset.

You can find more information on R Markdown in this handy cheat sheet https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf

7.3.1 A new R Markdown document

Creating an R Markdown document is easy. In R Studio, use the menu `File -> New File -> R Markdown` and you'll get a dialogue box like

Leave everything as default (making a document with output format html) and click OK. A new panel should appear in R Studio. The header looks like:

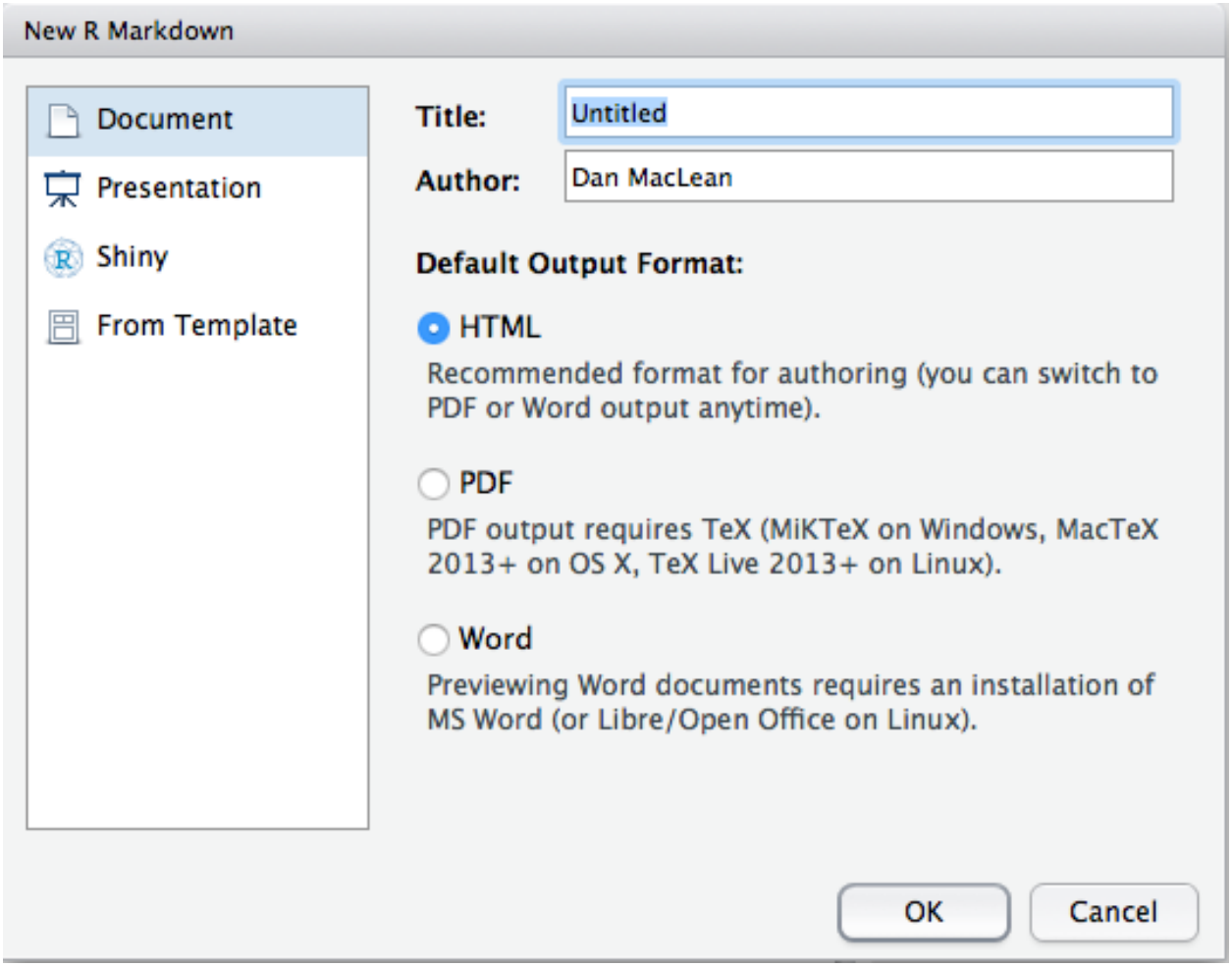


Figure 3: dialogue box

```
---
title: "Untitled"
author: "Dan MacLean"
date: "21 September 2016"
output: html_document
---
```

The top four lines are metadata about the document, R will use this to make an automatic header. You can change the values of `title`, `author` and `date` if you like. The last bit about output defines the type of document you can get.

The rest of the document is straightforward text right up to the parts with the three backticks ““’ (weird quote things). These are the blocks of R code that will get evaluate in our R markdown. Anything between two sets of three backticks is sent to R and treated as R code, so that

```
```r
summary(cars)
```
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```
```

Gets the output of the `summary()` function in that position. To see this work, click the `Knit HTML` button and choose a filename for the R markdown document.

The eventual document produced is nicely formatted markdown with R code and results added in the proper place.

### 7.4 Markdown tags

Markdown provides a rich set of tags to mark up the document to make it look as pretty as you like. Here's a cheat-sheet you can use to make your own Markdown documents https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet

### 7.5 Quiz

1. Make a new R Markdown document that creates and renders a plot of your choice - any of the ones you've already done will be fine. Hint: Every time you run a markdown document the computer's memory is cleared for that document, it doesn't know about what goes on outside of the document. You need to load libraries and files in the document, even if they are loaded in R Studio already

## 8 Visual Customisation

1. Questions:
```

- How do I make my plot look the way I want it to?

2. Objectives:

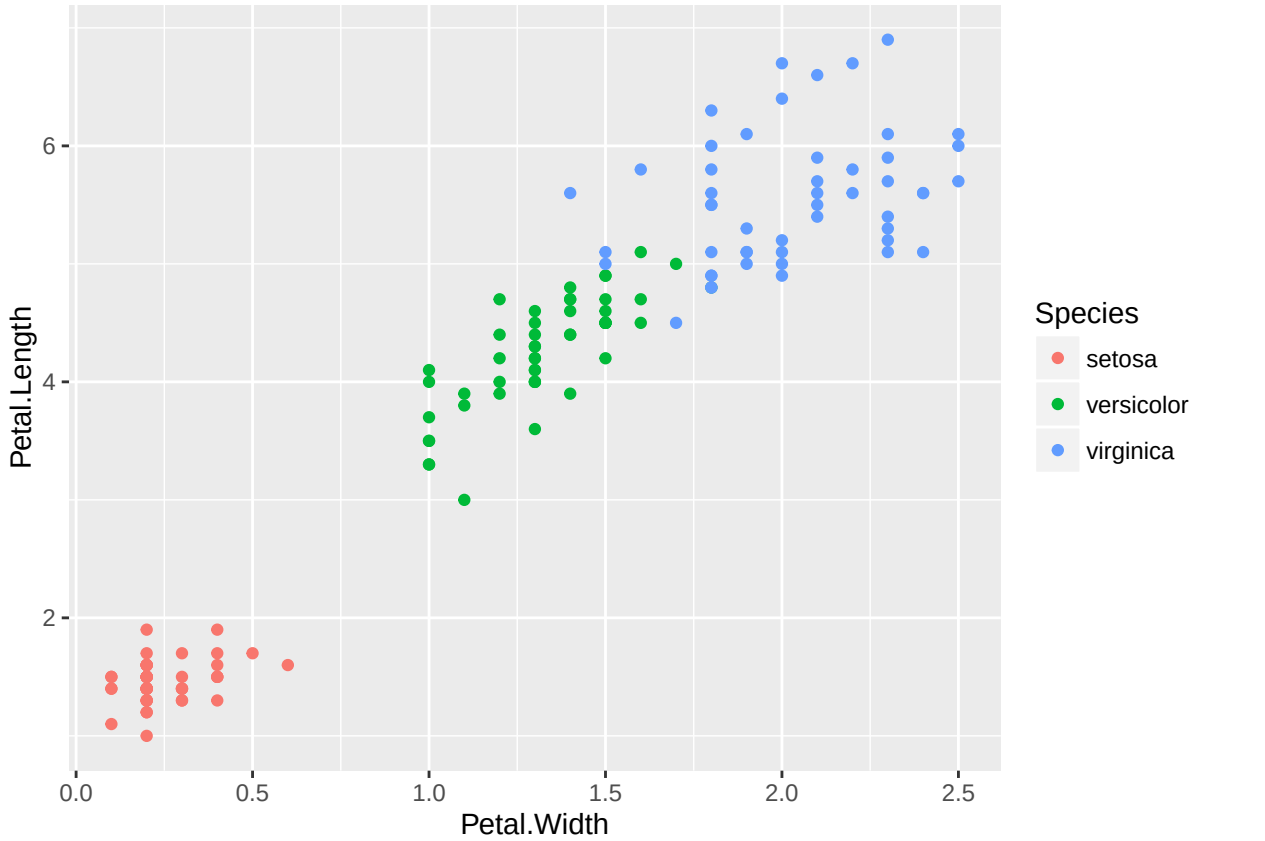
- Explain how themes are applied
- Explain how individual plot elements can be changed
- Set the order and limits on scales keypoints:
- There are a wide range of themes that can be modified
- The `theme()` function allows us to set individual theme elements
- The `scale` family of functions allows us to specify the scales

8.1 Themes

At some point you're going to want to custom/personalise or generally improve the look of your plots. So far we've concentrated on getting the data shown in the right place, now we'll look at finessing the plot to make a final version. `ggplot2` and a companion package `ggthemes` have a wide variety of ready to go themes that can be applied and modified.

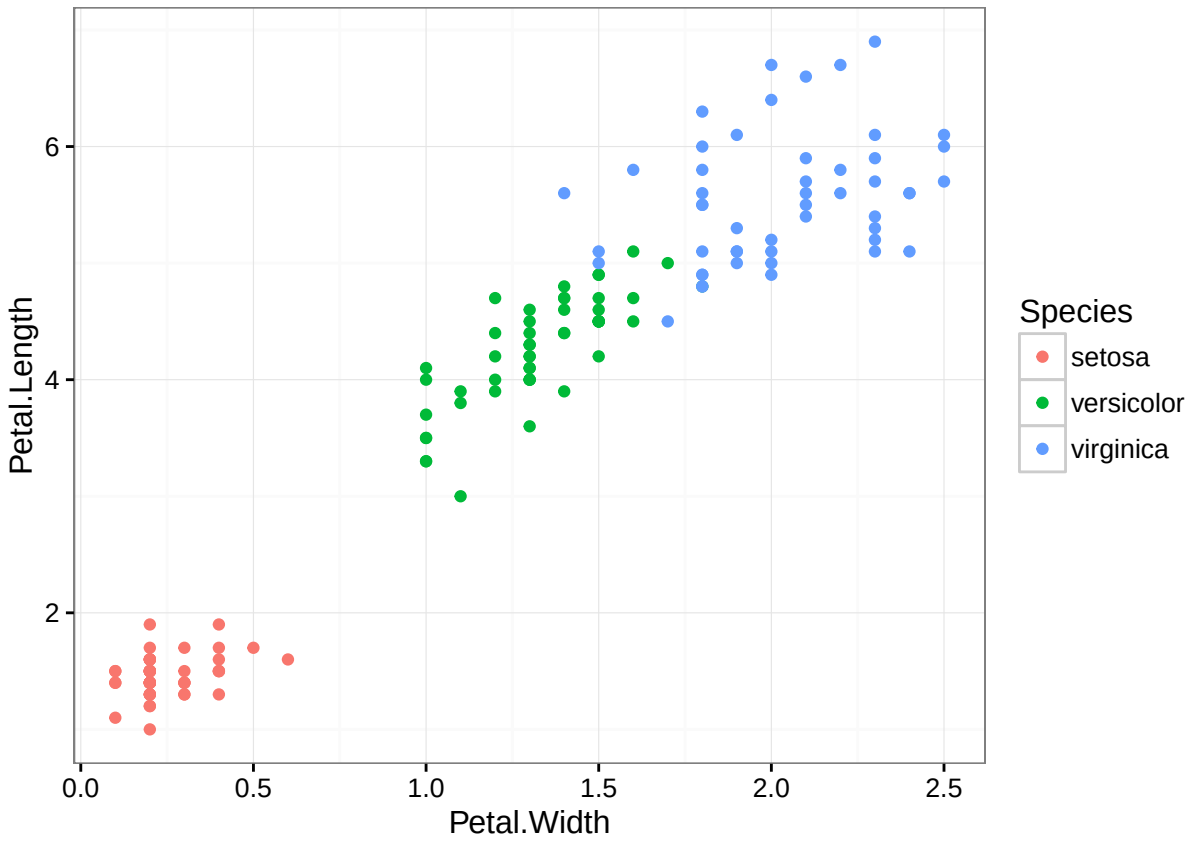
Applying a built-in theme is very easy, we can think of the theme as a new layer to add. This code will give us a standard plot

```
p <- ggplot(iris) + aes(Petal.Width,Petal.Length) + geom_point(aes(colour=Species))
p
```



Let's add a `theme_bw()` layer. Which is really a simple theme that takes away all colour you didn't explicitly ask for - so the points stay coloured.

```
p <- ggplot(iris) + aes(Petal.Width,Petal.Length) + geom_point(aes(colour=Species))
p + theme_bw()
```



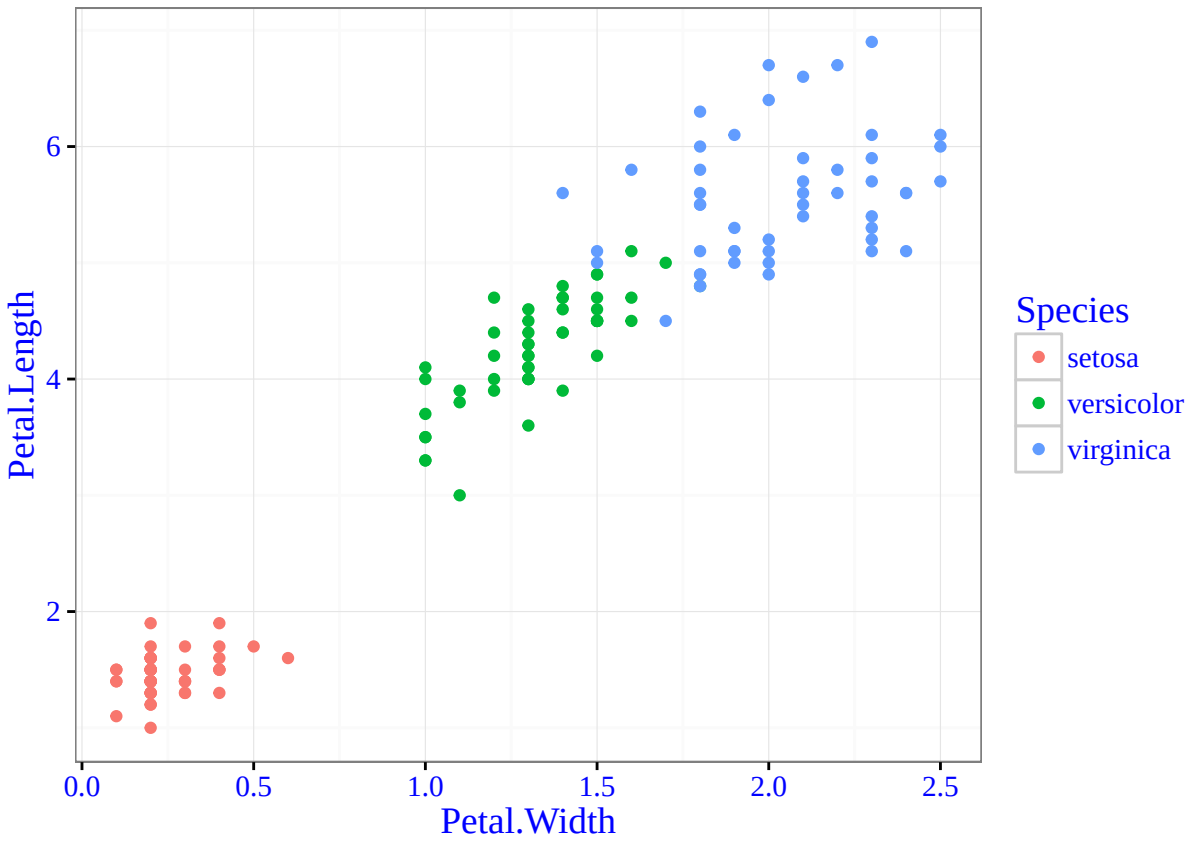
8.2 Quiz

1. `ggplot` itself only has a few themes built in. Try `theme_minimal()`, `theme_grey()` and `theme_dark()`.
2. Use the docs at https://github.com/jrnold/ggthemes to examine the themes that are available in this external package. Try loading the package and using some of the themes. Don't miss `theme_excel()`.

8.3 The `theme()` function

Changing the theme wholesale by applying a theme layer is great, but you'll usually want to change individual theme elements. This is possible too, and is done using the `theme()` function.

```
p <- ggplot(iris) + aes(Petal.Width,Petal.Length) + geom_point(aes(colour=Species))
p + theme_bw() + theme(text = element_text(family = "Times", colour = "blue", size = 14))
```



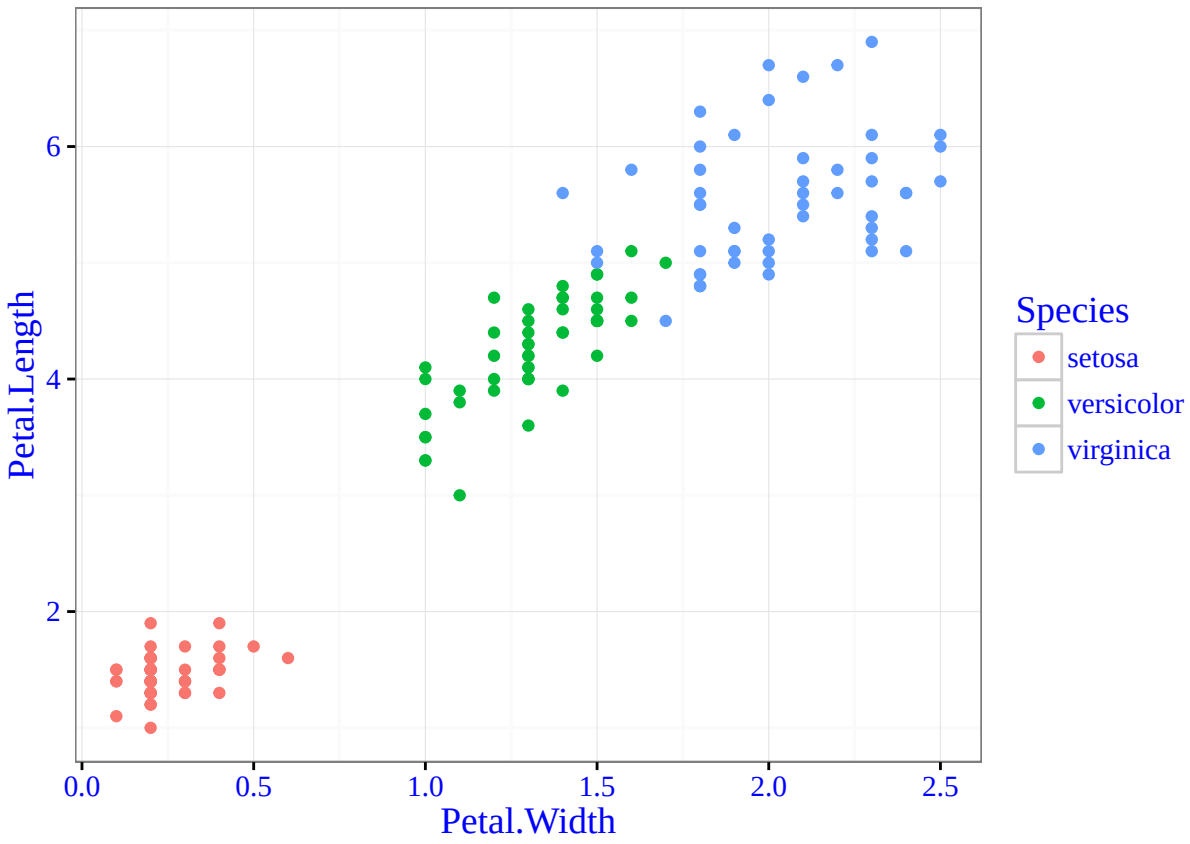
So here we built a plot, applied a theme layer and then modified an element of that theme. The theme layer is really just a list of plot elements and their current settings. Conceptually it looks like this:

```
- line = element_line(colour="black", size=0.5),
- text = element_text(family="Arial", colour="black", size=12)
```

With the thing on the left of the equals being the attribute of the plot e.g the `line` or the `text` and the thing on the right of the equals being the function that does the changing. Each plot element can be reset by using the proper function and setting the options for that function appropriately. Just like we did above!

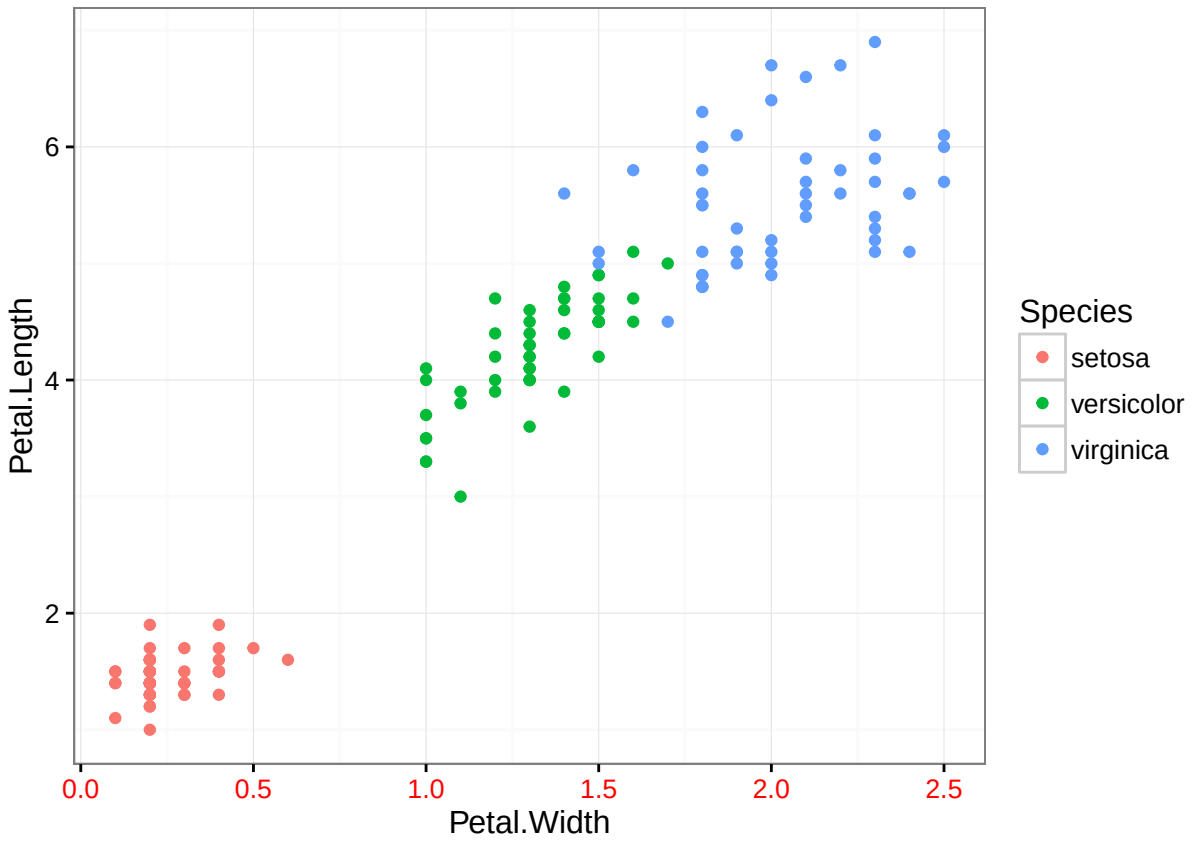
Some of the attributes apply across the whole plot, this one `text` applies to all text in the plot.

```
p + theme_bw() + theme(text = element_text(family = "Times", colour = "blue", size = 14))
```



But this one `axis.text.x` changes only the x axis text.

```
p + theme_bw() + theme(axis.text.x = element_text(colour="red"))
```



A full list of plot elements and the functions to set them are in http://docs.ggplot2.org/dev/vignettes/themes.html and here are the most important ones. The options for each element function are in the ggplot2 docs http://docs.ggplot2.org/dev/element.html

```
line =               element_line(),
rect =               element_rect(),
text =               element_text(),
axis.text =          element_text(),
strip.text =         element_text(),

axis.line =          element_blank(),
axis.text.x =        element_text(),
axis.text.y =        element_text(),
axis.ticks =         element_line(),
axis.title.x =       element_text(),
axis.title.y =       element_text(),
axis.ticks.length =  unit(),
axis.ticks.margin =  unit(),

legend.background =  element_rect(),
legend.margin =      unit(),
legend.key =         element_rect(),
legend.key.size =    unit(),
legend.text =        element_text(),
legend.title =       element_text(),
```

```
legend.position =    "right",
legend.justification = "center",

panel.background =   element_rect(),
panel.border =       element_blank(),
panel.grid.major =   element_line(),
panel.grid.minor =   element_line(),
panel.margin =       unit(),

strip.background =   element_rect(),
strip.text.x =       element_text(),
strip.text.y =       element_text(),

plot.background =    element_rect(),
plot.title =         element_text(),
plot.margin =        unit(),
```

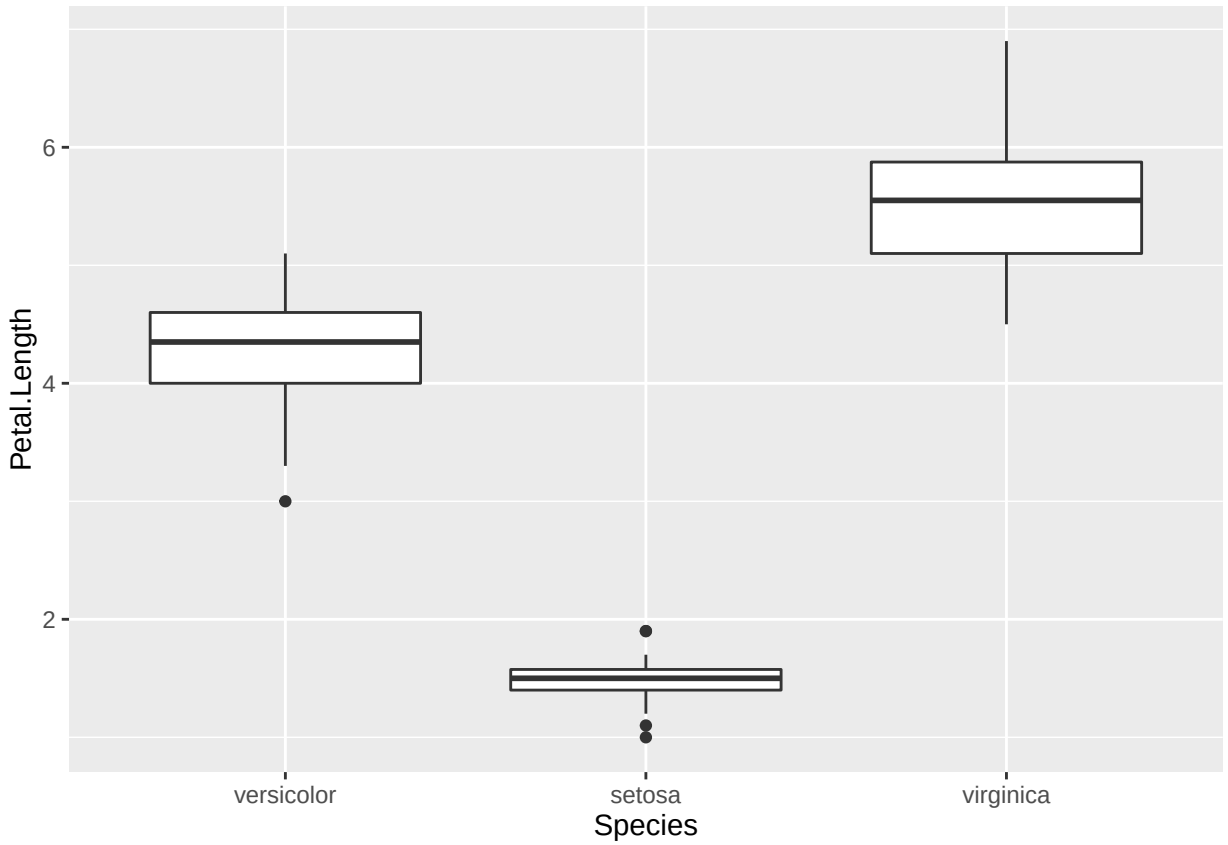
Putting these together if we want to make our legend text a bit bigger, and use Helvetica font, in green, we'd follow this scheme:

1. Use the list above to find which element is the right one for legend text. Here it will be `legend.text`
2. Read off the element function for the `legend.text`, here it is `element_text()`
3. Use the ggplot2 docs to see the options for that element function: http://docs.ggplot2.org/dev/element.html
4. Form the theme function: `theme( legend.text = element_text(size = 20, family="Helvetica", colour="green") )`
5. Add it to the plot `plot + theme( legend.text = element_text(size = 20, family="Helvetica", colour="green") )`

8.4 Changing the order of categories in the plot

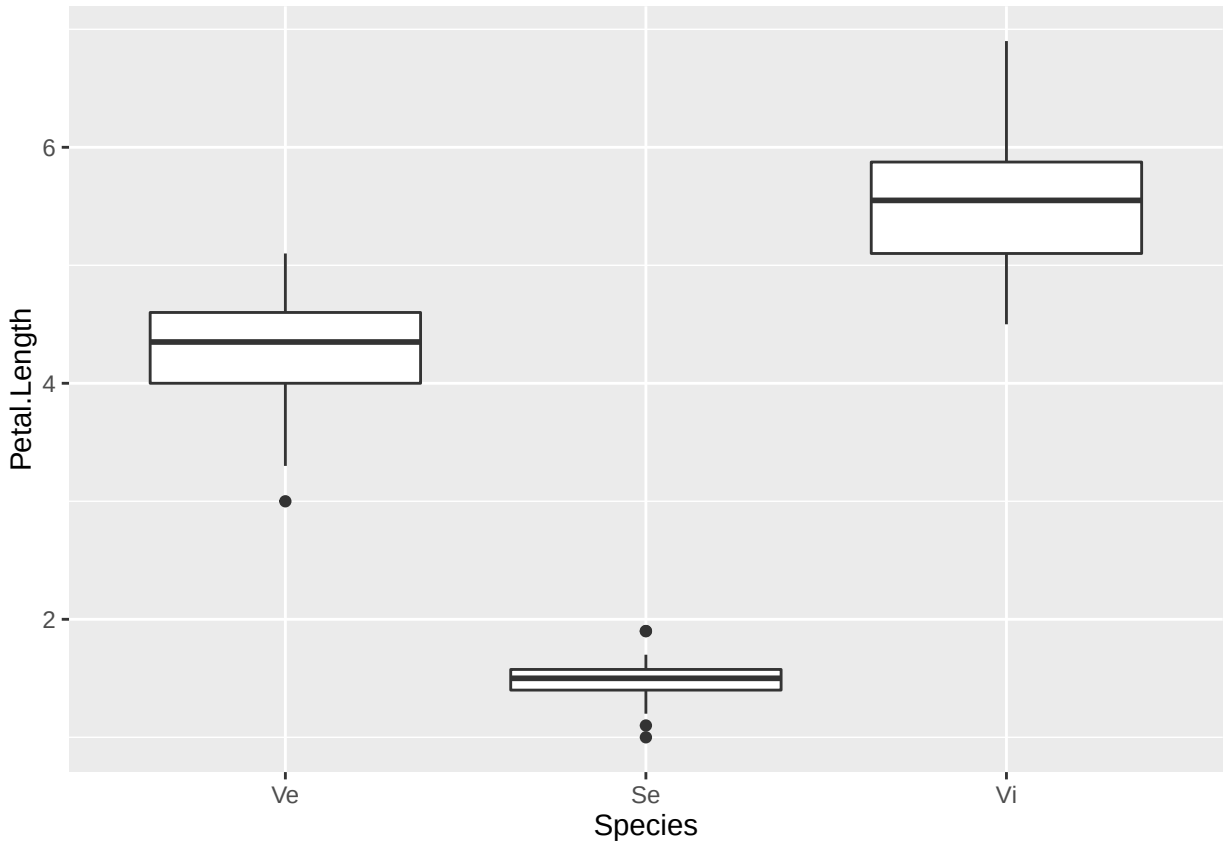
The list of options above doesn't provide anything we can use to specify the order in which the different categories are displayed. Instead this is done with a new type of function, the `scale` family of functions. By using the `scale_x_discrete()` function and options (especially the `limits`) we can set the way the scale on the axis is set. For a discrete (or categorical) variable this includes the order.

```
p <- ggplot(iris) + aes(Species,Petal.Length) + geom_boxplot()
p + scale_x_discrete(limits=c("versicolor", "setosa", "virginica"))
```



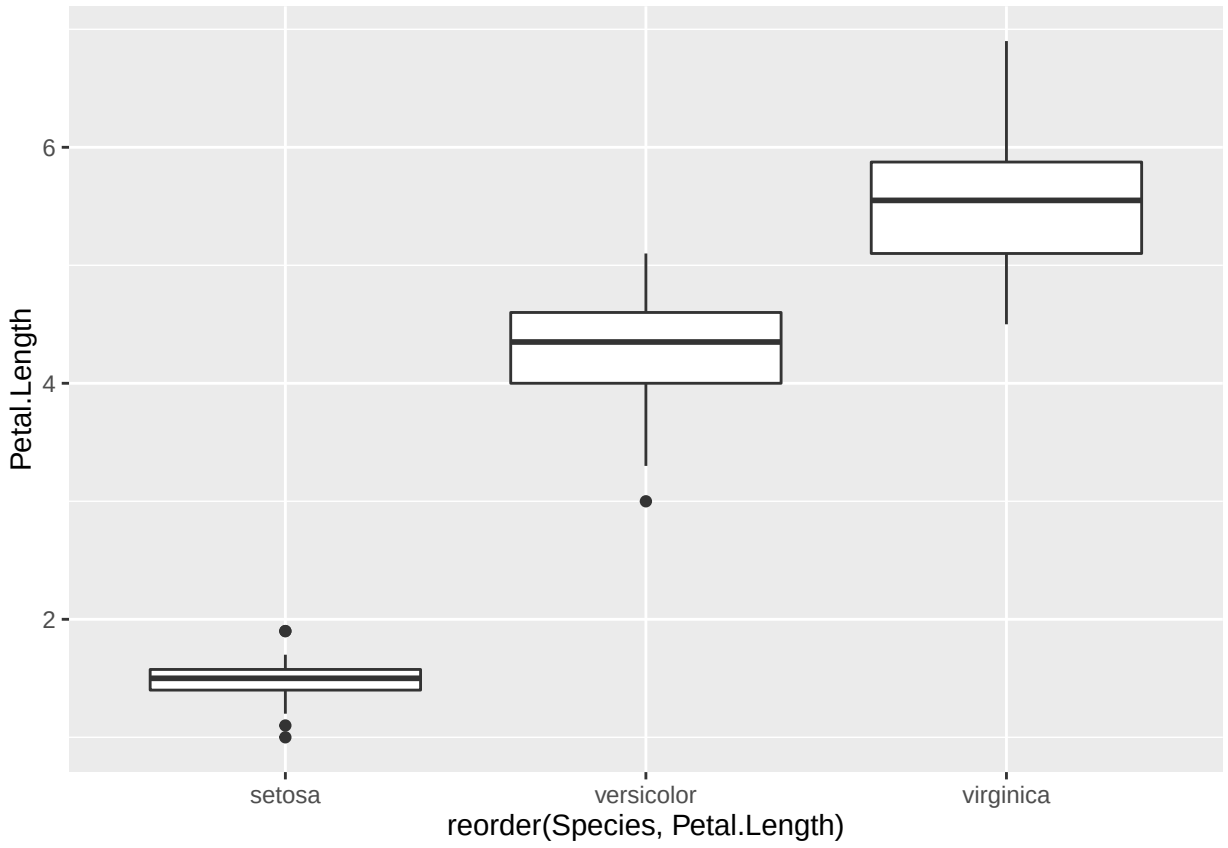
You can change labels in the same way with the `labels` option,

```
p + scale_x_discrete(limits=c("versicolor", "setosa", "virginica"), labels=c("Ve", "Se", "Vi"))
```



You can reorder based on the value of some other value, e.g get the boxes ordered by the `Petal.Length` variable by squeezing in the `reorder()` function. Unusually, this is done in the `aes()` function in the aesthetic layer. We want to reorder the x-axis so we use the `reorder()` function on that. The syntax is `reorder(<variable to reorder>, <variable to reorder by)`, so here we're changing the order of the Species on the x-axis according to what is in `Petal.Length`.

```
p <- ggplot(iris) + aes(x=reorder(Species, Petal.Length), y=Petal.Length ) + geom_boxplot()
p
```



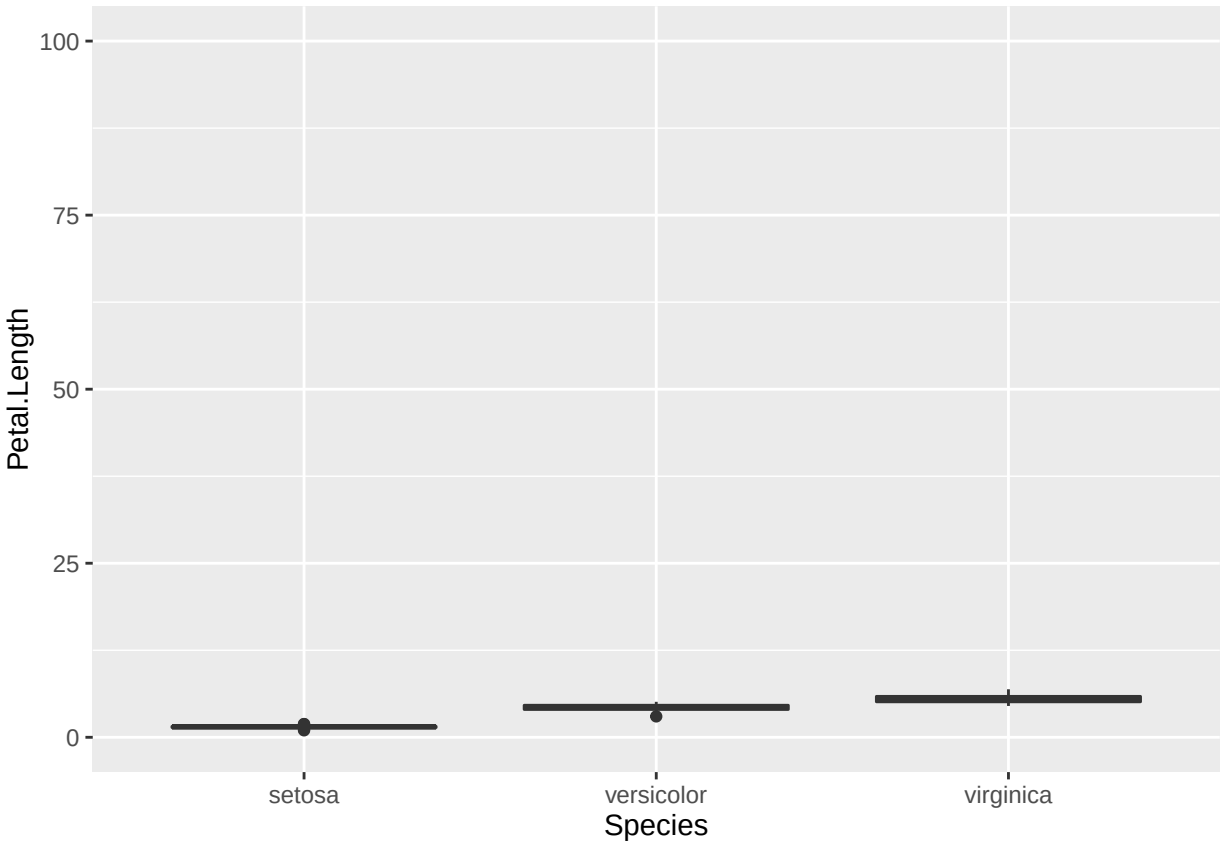
8.5 Text formatting in plots

Biological notation is frustrating because it uses text formatting to express differences between things. So the wild-type allele is referred to in italics or underlined capitals whereas a mutant is referred to in italic or underlined lower case. Programming languages have a hard time with text formatting, so tend to deal with plain text. `ggplot` is no exception and there isn't a way to make your labels italic. The best way to achieve this, therefore is to save the plot as a `.svg` file, then edit the labels manually in a graphics program like Inkscape.

8.6 Changing the limits of a continuous scale

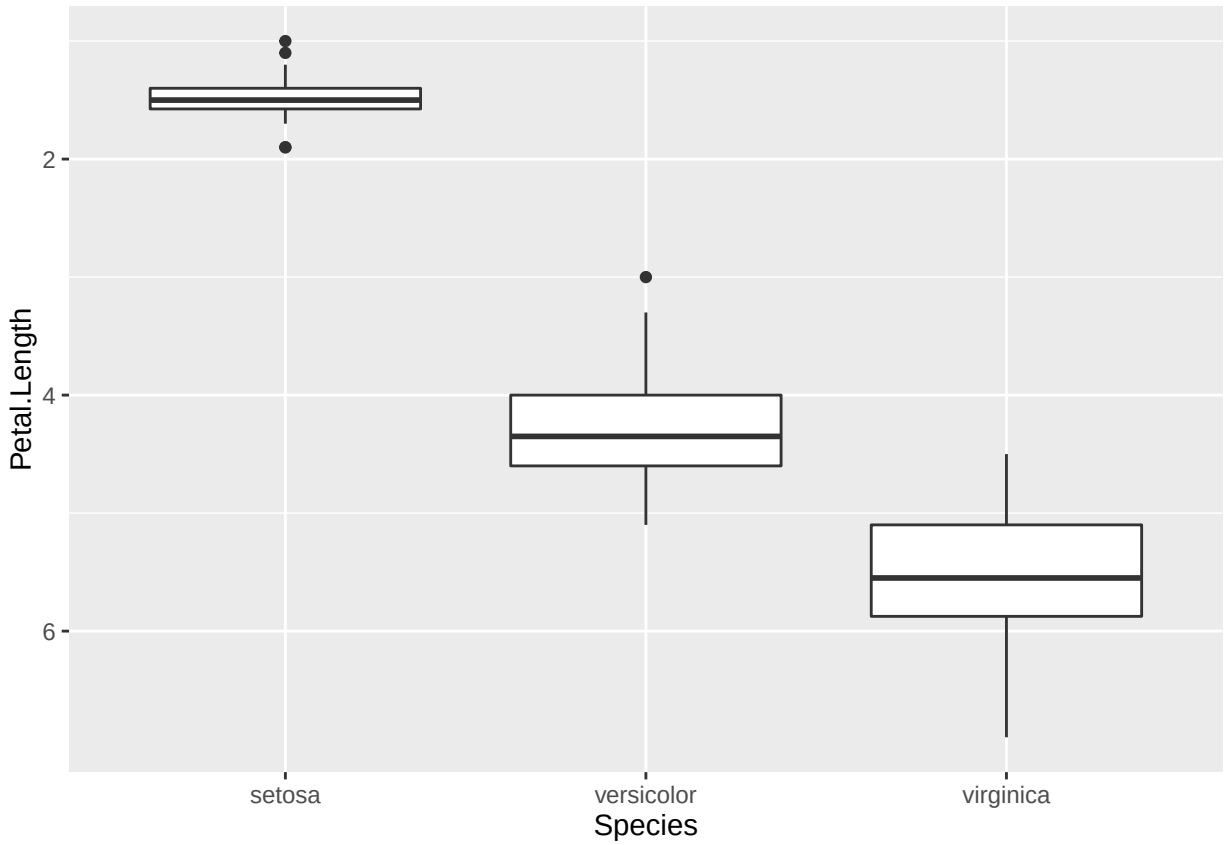
The `scale_x_discrete()` function has analogous functions for the y-axis and for continuous axes - I.E. `scale_y_discrete()` and `scale_x_continuous()` and `scale_y_continuous()`. The most common thing to want to do with a continuous scale is set the limits, the start and end points.

```
p <- ggplot(iris) + aes(Species, Petal.Length ) + geom_boxplot()
p + scale_y_continuous(limits =c(0,100))
```

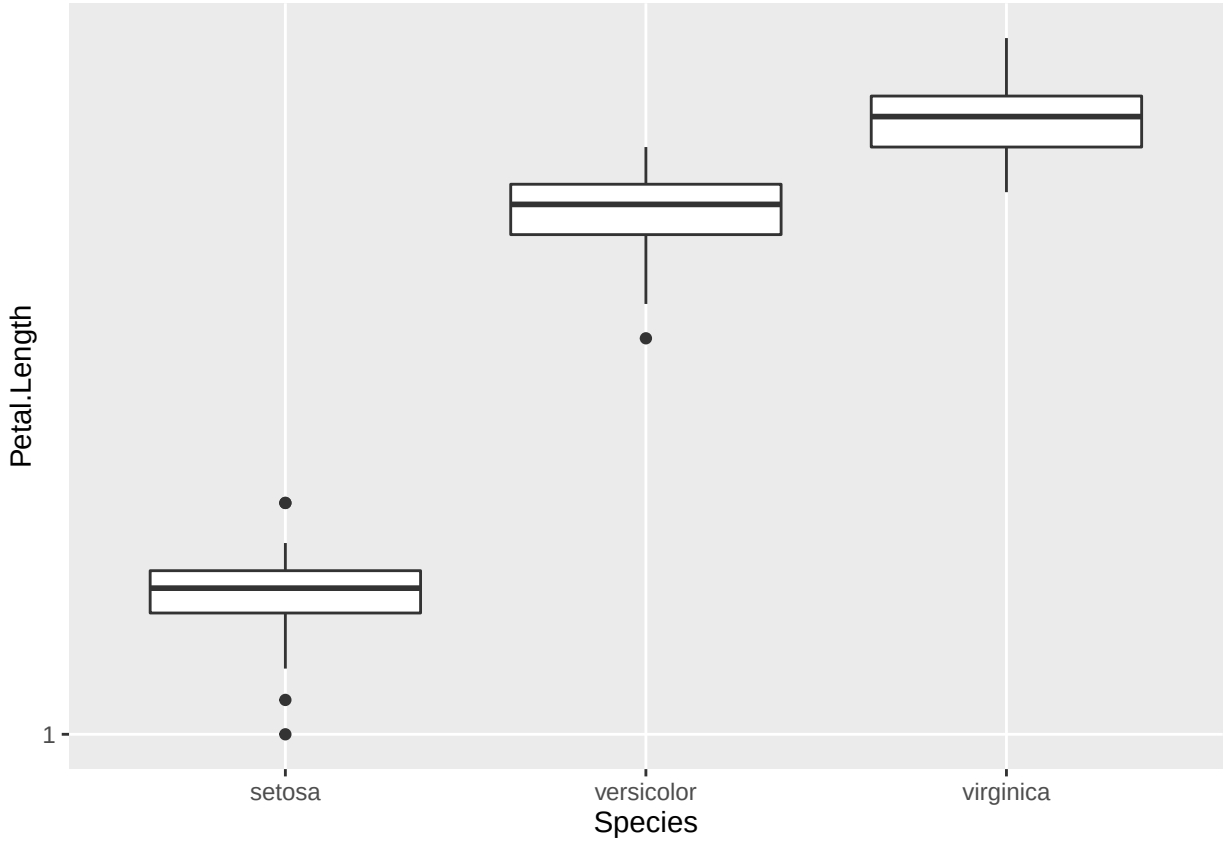


It is also possible to change the scale to a logarithmic one with the `scale_y_log10()`, function, reverse it with `scale_y_reverse()` functions.

```
p + scale_y_reverse()
```




```
p + scale_y_log10()
```



8.7 Quiz

1. Using the iris dataset, create a boxplot of Petal Width for each species
2. Overlay the actual data by adding a jitter plot
3. Remove the grey background of the plot (Hint: try `element_blank()` and `panel.background`)
4. Change the Y axis title to 'Petal Width'
5. Remove the X axis title
6. Make the species names bigger
7. Make the thick panel grid lines black, remove the thin panel grid lines.
8. Set the order of species to 'virginica', 'setosa', 'versicolor' Extra Credit: Set the values on the species axis to 'Iris virginica', 'Iris setosa', 'Iris versicolor'

9 Loading your own data

1. Questions:

- How do I use *my* data?

2. Objectives:

- Preparing a csv file in 'tidy' format
- Understanding file system paths
- Loading a file to a dataframe

- Explicitly describing the file contents

3. Keypoints:

- Data needs to be in a particular format for `ggplot` to work
- Specifying the data type is sometimes necessary when creating a data frame.

9.1 Tidy data

There are many ways to structure data. Here are two quite common ones.

treatmenta

treatmentb

John Smith

11

2

Jane Doe

16

11

Mary Johnson

3

1

John Smith

Jane Doe

Mary Johnson

treatementa

11

16

3

treatementb

2

11

1

source: Hadley Wickham

Tables contain two things, variables and values for those variables. In these two tables there are only three variables. `treatment` is one, with the values `a` and `b` . The second is 'name', with three values hidden in plain sight, and the third is `result` which is the value of the thing actually measured for each person and treatment.

For human reading purposes, we don't need to state the variables explicitly, we can see them by interpolating between the columns and rows and adding a bit of common sense. This mixing up of variables and values across tables like this has led some to call these tables 'messy'. A computer finds it hard to make sense of a messy table.

Working with R is made much less difficult if we get the data into a 'tidy' format. This format is distinct because each variable has its own column explicitly, like this

name

treatment

result

John Smith

a

11

Jane Doe

a

16

Mary Johnson

a

3

John Smith

b

2

Jane Doe

b

11

Mary Johnson

b

1

Now each variable has a column, and each seperate observation of the data has its own row. It is *much* more verbose for a human, but R can use this easily because we are now explicit about what is called what and how it relates to everything else.

9.2 Getting your data into tidy format

The bad news here is that there is no magic function to make your data tidy. If you have an existing table then you can do this manually in Excel or some other spreadsheet package. If you have lots of data, it is possible to do it programmatically in R, see the `dplyr` and `tidyr` packages, which are complex but designed for this purpose. Also have a look at the cheat-sheet here https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf.

9.3 Loading in a CSV file

R can deal with a lot of file formats, but the most common and easily used one is 'csv', a comma-separated value file. These can be exported from virtually any spreadsheet program so its a good interchange format to get data into R from wherever you already have it. Loading a file is done easily with the `read.csv()` function, just give it the file to be read.

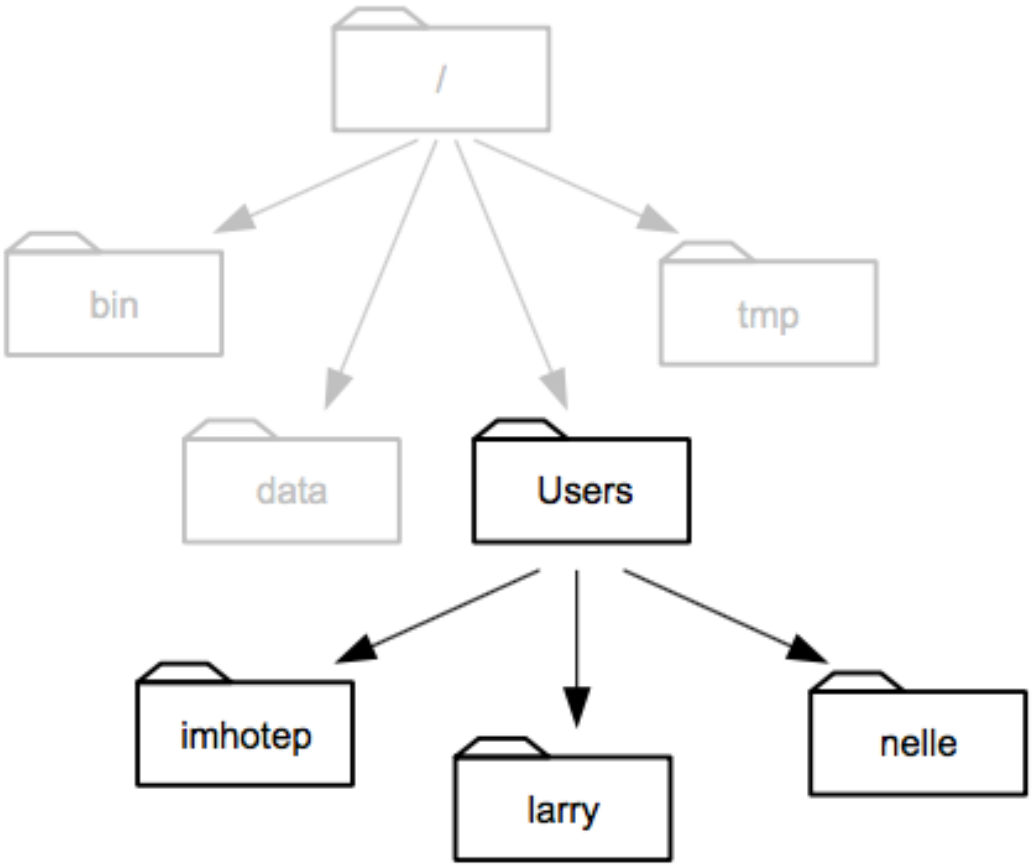
```
my_data <- read.csv('my_data.csv')
```

This will create an object called a dataframe that can be used just like the iris data.

9.4 Finding the file

R needs to be given the correct and full path to a file. This means the full address of the file on the hard disk of your computer. R doesn't have a file chooser so you need to know how to write this down.

Computer file systems are laid out in folders and sub-folders with files inside them. Conceptually, this results in a tree of folders and a path down the branches from the root of the tree to everything else. The root gets called '/' on Mac/Linux computers and 'C:' on Windows computers



source: Software Carpentry

This picture of an example file system shows how that is formed. When we write this down, everytime we go inside a new folder we use a slash to show we've changed folder. Most computer systems have a 'Users' or similar folder in which each users stuff is stored. Supposing we're in Larry's folder then the path would be `/Users/larry`. And a file called `my_file.txt` in that folder would be `/Users/larry/my_file.txt`.

So to write the full file path for R we can use this pattern, the first bit would be `/Users/username/` (or `C:\Documents and Settings\username\` or `C:\Users\username\`) and then the set of folders

within that user area follows on. If your file `my_file.txt` is on the Desktop the full path would be `/Users/username/Desktop/my_file.txt` (or `C:/Documents and Settings/username/Desktop/my_file.txt`)

9.4.1 Make it easy on yourself

The easiest way not to have to think too hard about this stuff is to set up a consistent folder and file structure for every analysis and use RMarkdown documents to run your analysis. Here's an example scheme:

1. Create a new folder and call it something relevant to your experiment, e.g `disease_incidence_2016-11-01`
2. Within the folder create a sub-folder called `raw` and a sub-folder called `output_images`.
3. Put your tidy csv file in the `raw` folder.
4. Create a new R Markdown document and save it in the `disease_incidence_2016-11-01` folder.

Now whenever you open and run that R Markdown document, the path of your input file is `"raw/my_input_filename.csv"`. You can save your plots with the `ggsave()` function to `"output_images/filename.png"` (don't forget the quotes).

If you never mess around with the relative positions of the files and folders described, then the paths will always be the same. You can move the whole folder without worrying, just don't jumble it's contents.

9.5 Making sure the data types are correct

When we load new data we need to make sure that any header has been properly parsed as column names, and that the columns have been identified as the right sort of data

Once the file is loaded, you need to ensure that the created dataframe is correct. We can examine a dataframe with the `str()` function.

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The output tells us that this dataframe has 150 observations of 5 variables (or columns) the column names come after the `$` sign and the type of each column comes after that. So here the Length and Width columns are of type 'num' - which means numeric and the Species column is of type factor (and that factor has 3 different values)

If the header isn't parsed properly we can force the first line in the file to be taken as the header with

```
data <- read.csv('my_data.csv', header=TRUE)
```

If a column should be numeric but hasn't been loaded that way, you can change it with

```
data$column_name <- as.numeric(data$column_name)
```

and if a column should be a factor, but isn't you can change it with

```
data$column_name <- as.factor(data$column_name)
```

Once the output of `str()` shows what you expect, then you are good to start analysing.

9.6 Quiz

1. Set up an analysis folder:
    1. Make a new folder called `analysis` on the Desktop
    2. Inside `analysis` make a new folder called `raw` and put `example_ros_data_flg22.xlsx` into it.
    3. Start a new R Markdown document and save it in `analysis`
2. Convert `raw/example_ros_data_flg22.xlsx` into a 'tidy' format .csv file and save to `raw`
3. Load in the data from the tidy file using `read.csv()` (Hint: You may need to save a csv version from Excel - R won't read .xlsx files.)
4. Check the datatypes and headers using `str()`, change them if necessary.
5. Create a plot that shows each data point in each treatment (Col, pp2c38, pp2c48 pp2c38/pp2c48) in each day the experiment was done.
6. Make sure the plot you generate gets saved to a folder inside `analysis` called `output_images`

10 *ggtree* a package for plotting phylogenetic trees

1. Questions:

- How do I render and annotate a tree from an existing tree file?

2. Objectives:

- Understanding the basics of *ggtree*
- Annotating Nodes
- Grouping Clades

3. Keypoints:

- ggtree is a relative of ggplot for trees
- ggtree contains many geoms for annotating trees

10.1 `ggtree` - a Bioconductor package for displaying phylogenetic trees

Bioconductor is a (very) large set of libraries for operating on biological data types https://www.bioconductor.org/ . `ggtree` is a library inspired by `ggplot` for drawing trees. Much of what we've already seen in `ggplot` is transferrable to `ggtree` so the syntax should be familiar.

10.2 Installing `ggtree`

Installing bioconductor takes a long time and isn't done in the same way as with other R packages. To install first run this special script in the R console (making sure you have an internet connection) : `source("https://bioconductor.org/biocLite.R")` in the R console (making sure you have an internet connection). Bioconductor libraries can then be installed with the `biocLite()` function 'biocLite("ggtree")

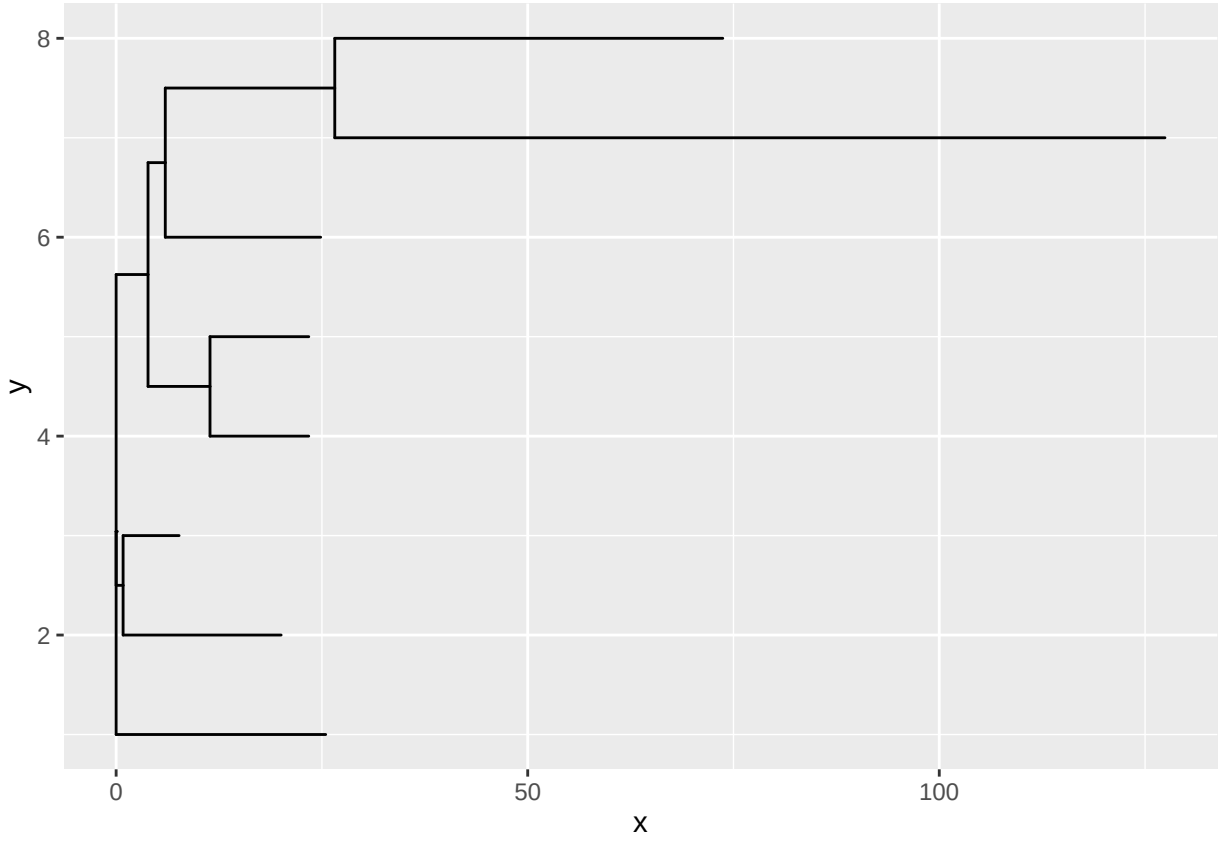
`ggtree` takes a range of common tree formats as input. We'll use a sample file in `newick` format. This creates a special sort object called a 'phylo' object that knows all sorts of information about the tree.

```
library(ggtree)
tree <- read.tree("data/mammals.nwk")
str(tree)
```

```
## List of 4
##  $ edge       : int [1:13, 1:2] 9 10 10 9 11 12 12 11 13 14 ...
##  $ Nnode      : int 6
##  $ tip.label  : chr [1:8] "raccoon" "bear" "sea_lion" "seal" ...
##  $ edge.length: num [1:13] 0.846 19.2 6.8 3.874 7.53 ...
##  - attr(*, "class")= chr "phylo"
##  - attr(*, "order")= chr "cladewise"
```

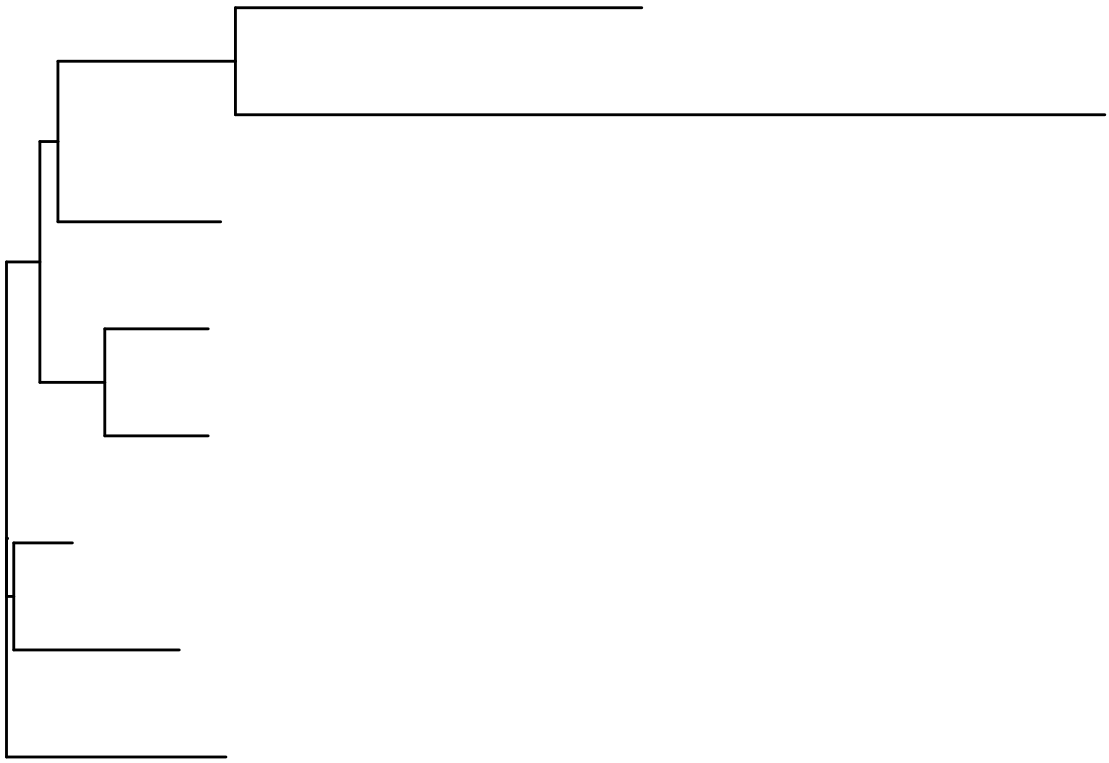
The tree can be drawn using the `geom_tree()` function.

```
ggplot(tree) + geom_tree()
```



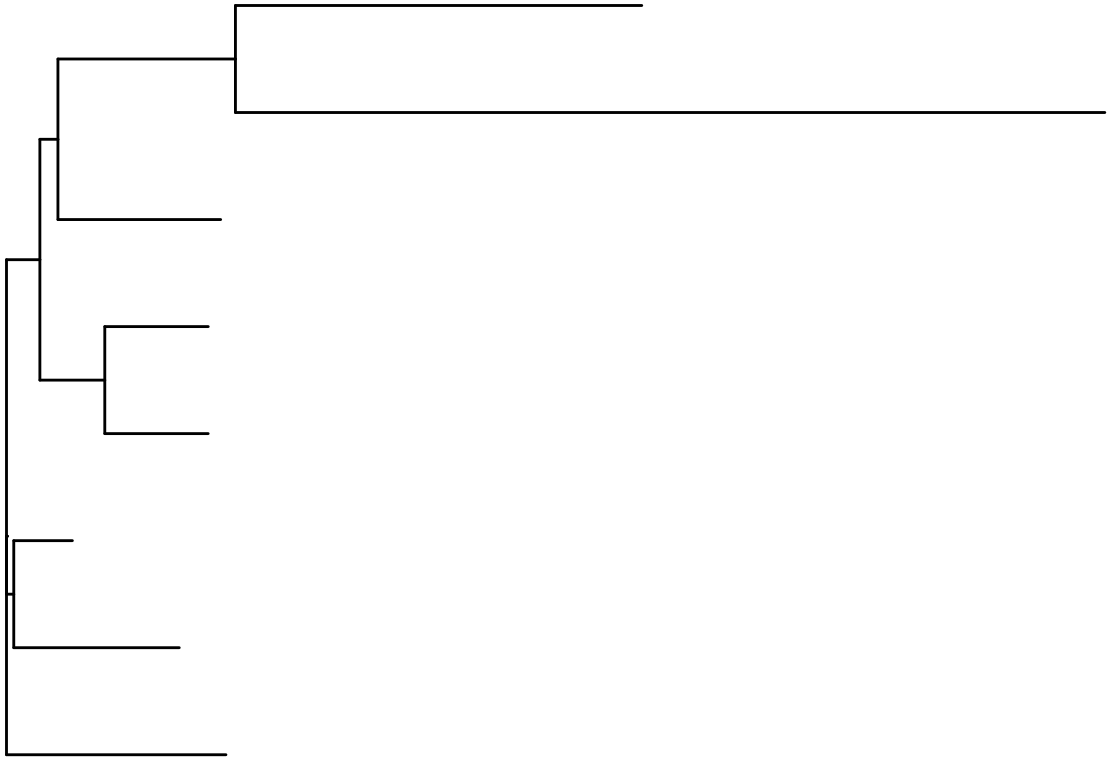
There is a special theme that sorts out the background for trees:

```
ggplot(tree) + geom_tree() + theme_tree()
```



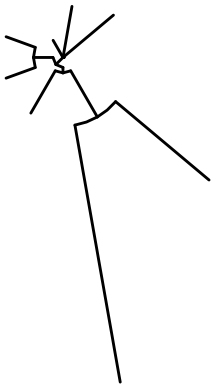
And because you nearly always want these three `ggtree` provides a utility function to do all of that - `ggtree()`

```
ggtree(tree)
```



With this function we can add layout options

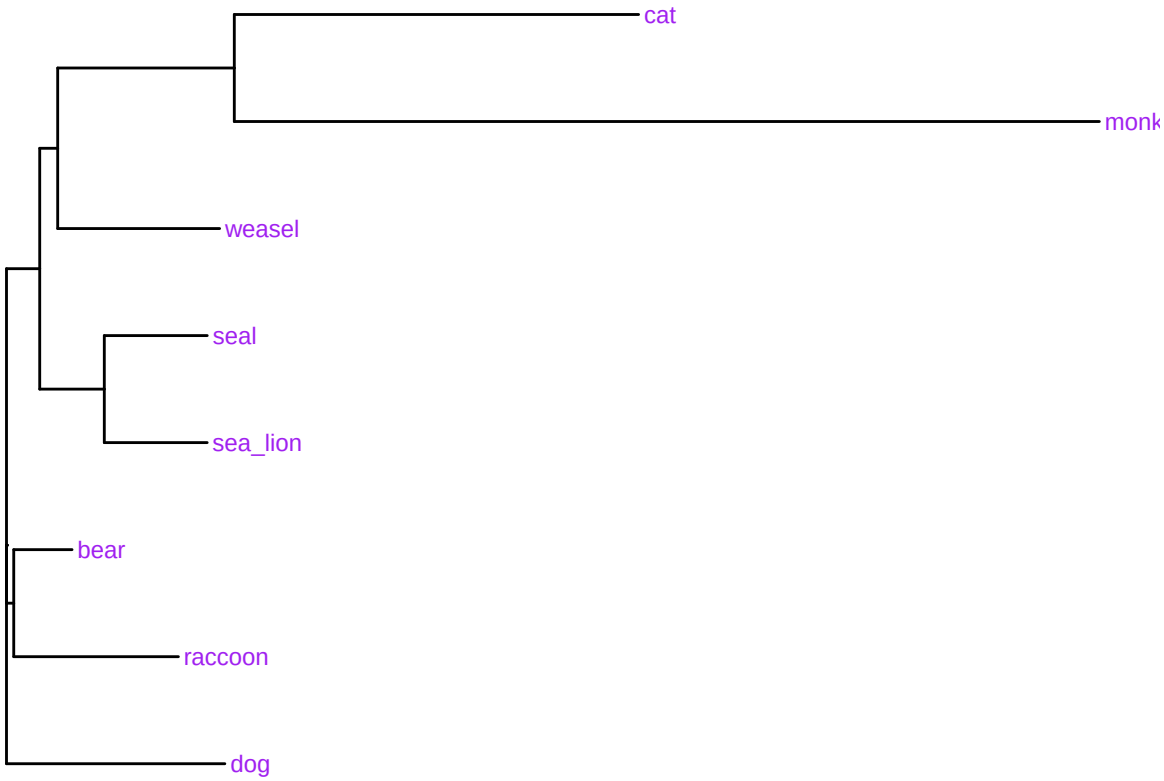
```
ggtree(tree, layout = "circular")
```



10.2.1 Labels

Adding labels to treetips is done with the `geom_tiplab()` geom.

```
ggtree(tree) + geom_tiplab(size=3, color="purple")
```



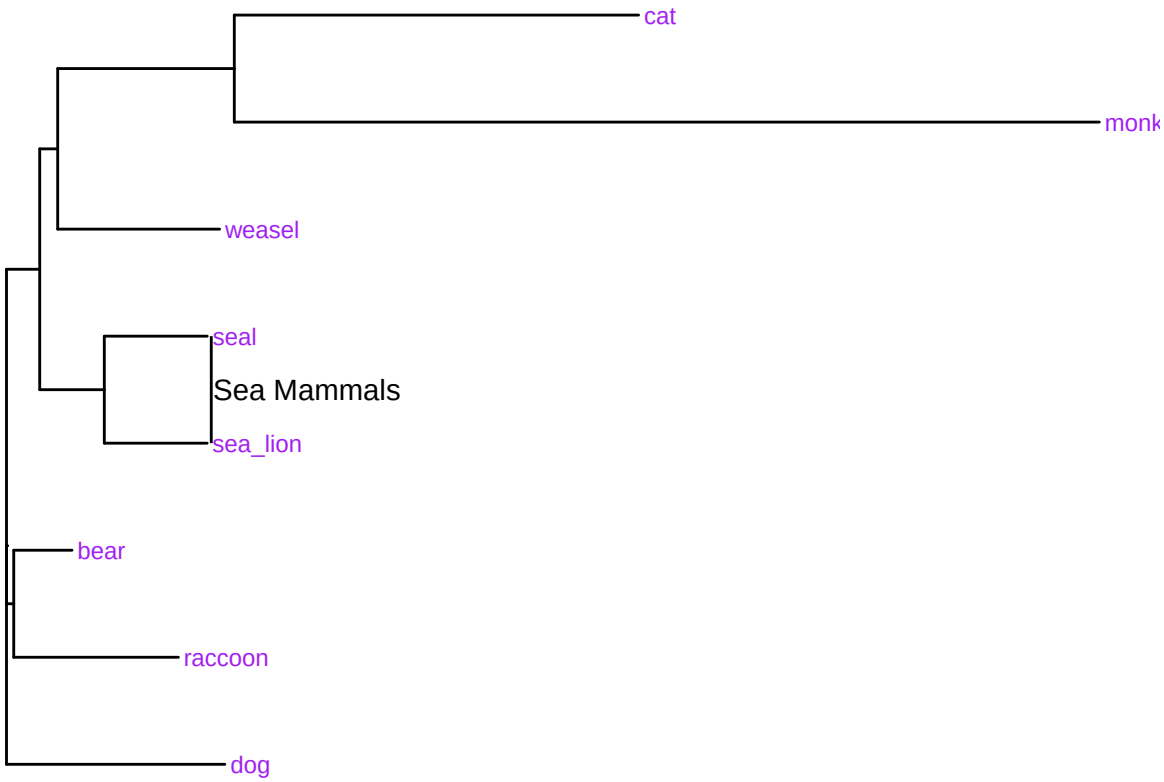
Adding the nodes is done with special options to the `geom_point()` geom. The shape and colour aesthetics are set to the variable `isTip` which is an internal variable in the tree object.

```
p <-  ggtree(tree) + geom_tiplab(size=3, color="purple")
p + geom_point(aes(shape=isTip, color=isTip), size=3)
```

10.2.2 Annotating and colouring clades

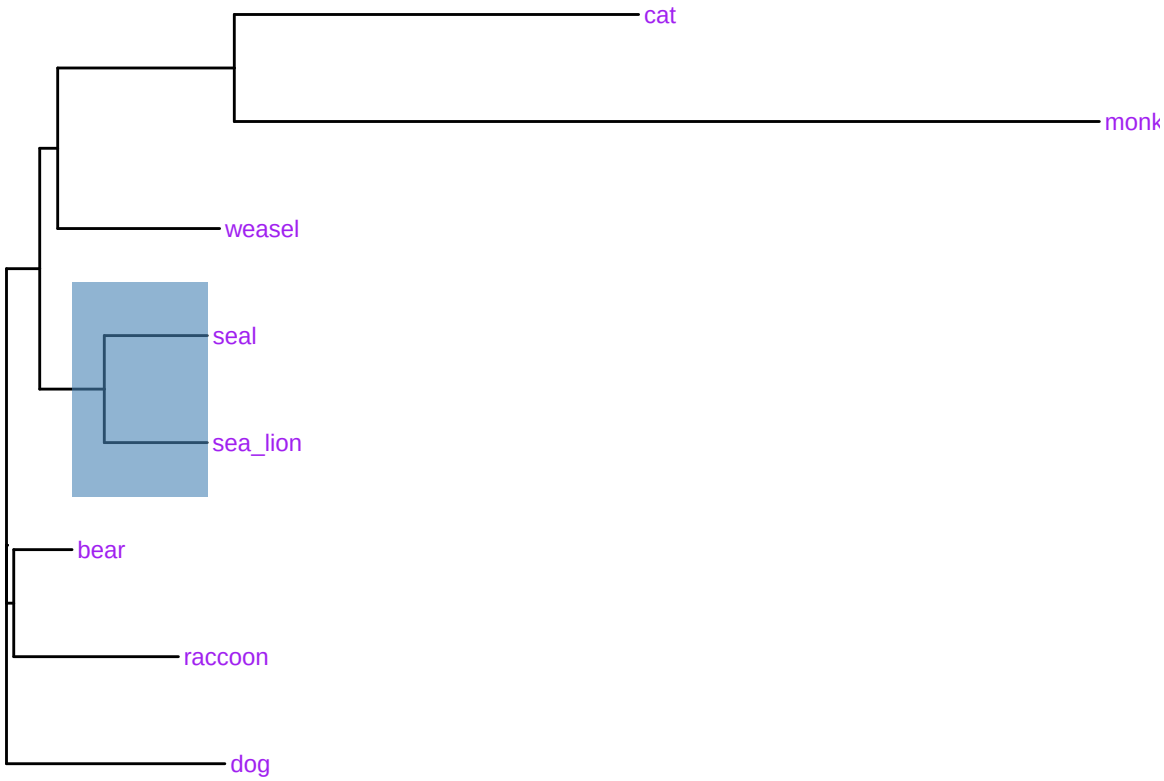
More useful is the ability to colour particular bits of the tree. First let's add a highlight bar to the side of the tree highlighting the sea mammal family. To do this we need to find the first node in the tree common to the clade we want to highlight. For this it's node number 12 and we use the `geom_cladelabel()` geom to add. We can use multiple `geom_cladelabel()` layers for more labels.

```
p + geom_cladelabel(node=12, label="Sea Mammals")
```



You can also use blocks of colour for this:

```
p + geom_hilight(node=12, fill="steelblue", alpha=.6)
```



To actually find the node number we need we can use the `MRCA()` function and pass it a list of labels we want

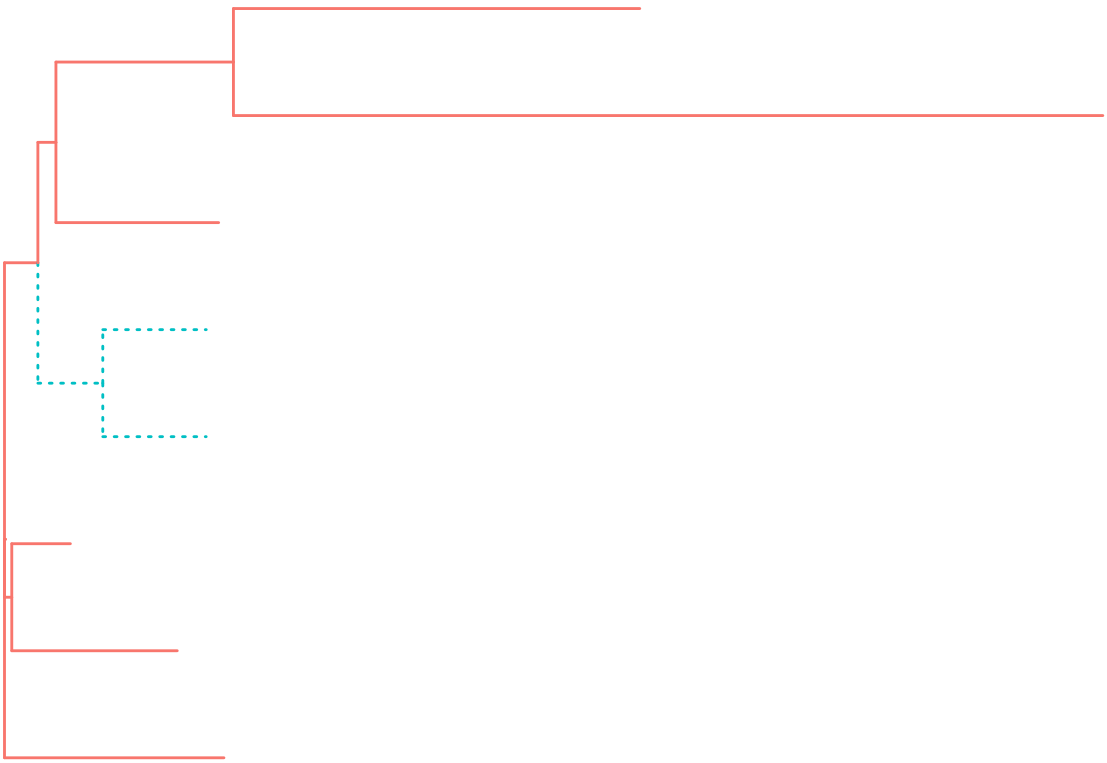
the most recent common ancestor for

```
MRCA(tree, tip=c('seal','sea_lion'))
```

```
## [1] 12
```

You can use this information to colour different parts of the tree, too. First you need to mark the tree objects as having a new `group` factor with the `groupClade()`, function and then dynamically colour by the new group factor

```
tree <- groupClade(tree, node=12)
ggtree(tree, aes(color=group, linetype=group))
```

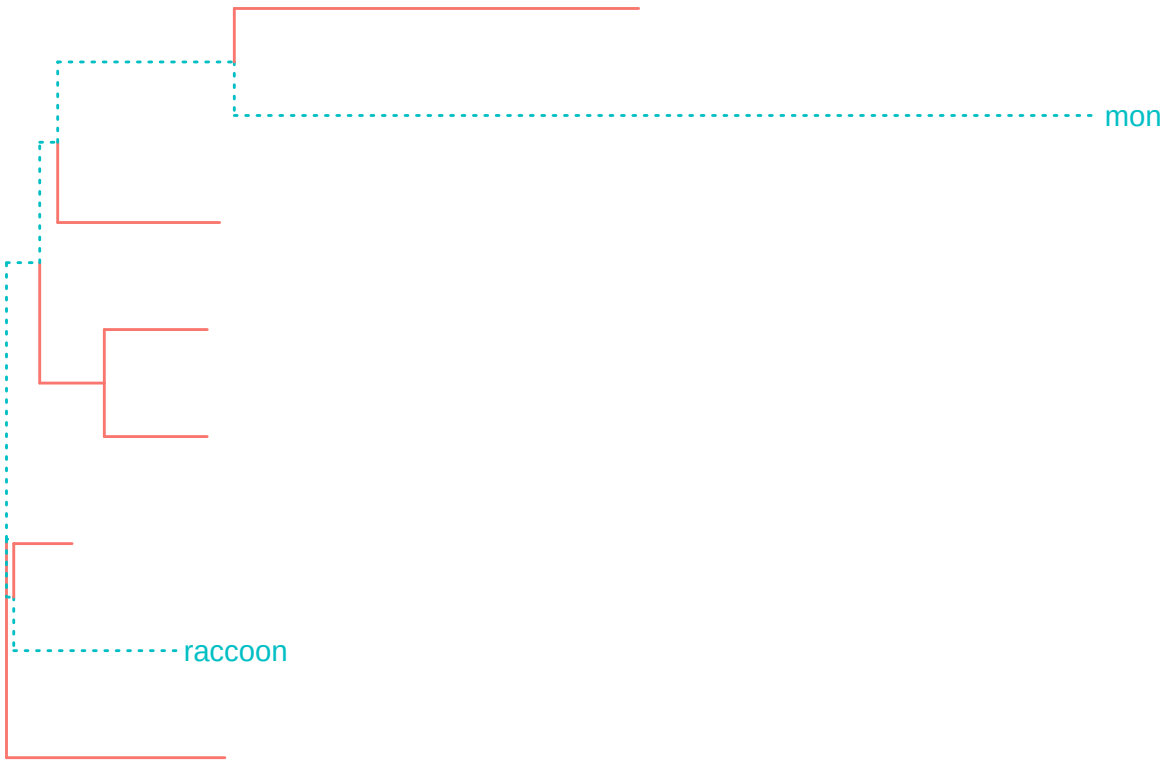


And you can preferentially operate on a clade by using the geoms on particular group numbers, here adding a label only to group 1

```
tree <- groupClade(tree, node=c(12, 13) )
ggtree(tree, aes(color=group, linetype=group)) + geom_tiplab(aes(subset=(group==1)))
```

You can work on arbitrarily defined groups with the `groupOTU()` function

```
tree <- groupOTU(tree, focus=c("monkey", "raccoon"))
ggtree(tree, aes(color=group, linetype=group)) + geom_tiplab(aes(subset=(group==1)))
```



10.3 Quiz

1. Set up an analysis folder:
2. Make a new folder called `analysis` on the Desktop
3. Inside `analysis` make a new folder called `raw` and put `pinf_mtDNA.newick` into it.
4. Start a new R Markdown document and save it in `analysis`
5. Create a circular cladogram of the tree and annotate the tips with tip labels. Rotate them to layout better. Hint: See the documentation here
6. Can you add the bootstrap values for each branchpoint?
7. Find the most recent common ancestor of US5 and PE_6096. Highlight the clade with a coloured box.

11 Using the stats functions in R

11.1 About this chapter

1. Questions:

- How do I do some common test?

2. Objectives:

- Using `t.test()` and calculating effect sizes
- Using `lm()` for regression
- Doing ANOVA with a linear model

3. Keypoints:

- There are a range of functions for doing every statistical test in R.

R is built for statistics so in this section we'll look at some common statistics functions.

First just getting the summary or descriptive statistics.

11.2 Summary Statistics

Earlier we saw that R has very many ways to get summary statistics like the mean and standard deviation from entire datasets. These ranged from the `summary()` function.

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
```

```
##  virginica :50
##
##
##
```

But a better way to summarise by factor is with the `describeBy()` function in the `psych` package. Note you need to use `$` notation to describe the column with the factor you want to subset with.

```
library(psych)
describeBy(iris, iris$Species)
```

```
## $setosa
##              vars  n mean   sd median trimmed  mad min max range skew
## Sepal.Length    1 50 5.01 0.35    5.0    5.00 0.30 4.3 5.8   1.5 0.11
## Sepal.Width     2 50 3.43 0.38    3.4    3.42 0.37 2.3 4.4   2.1 0.04
## Petal.Length    3 50 1.46 0.17    1.5    1.46 0.15 1.0 1.9   0.9 0.10
## Petal.Width     4 50 0.25 0.11    0.2    0.24 0.00 0.1 0.6   0.5 1.18
## Species*        5 50 1.00 0.00    1.0    1.00 0.00 1.0 1.0   0.0  NaN
##              kurtosis   se
## Sepal.Length    -0.45 0.05
## Sepal.Width      0.60 0.05
## Petal.Length     0.65 0.02
## Petal.Width      1.26 0.01
## Species*          NaN 0.00
##
## $versicolor
##              vars  n mean   sd median trimmed  mad min max range  skew
## Sepal.Length    1 50 5.94 0.52   5.90    5.94 0.52 4.9 7.0   2.1  0.10
## Sepal.Width     2 50 2.77 0.31   2.80    2.78 0.30 2.0 3.4   1.4 -0.34
## Petal.Length    3 50 4.26 0.47   4.35    4.29 0.52 3.0 5.1   2.1 -0.57
## Petal.Width     4 50 1.33 0.20   1.30    1.32 0.22 1.0 1.8   0.8 -0.03
## Species*        5 50 2.00 0.00   2.00    2.00 0.00 2.0 2.0   0.0   NaN
##              kurtosis   se
## Sepal.Length    -0.69 0.07
## Sepal.Width     -0.55 0.04
## Petal.Length    -0.19 0.07
## Petal.Width     -0.59 0.03
## Species*          NaN 0.00
##
## $virginica
##              vars  n mean   sd median trimmed  mad min max range  skew
## Sepal.Length    1 50 6.59 0.64   6.50    6.57 0.59 4.9 7.9   3.0  0.11
## Sepal.Width     2 50 2.97 0.32   3.00    2.96 0.30 2.2 3.8   1.6  0.34
## Petal.Length    3 50 5.55 0.55   5.55    5.51 0.67 4.5 6.9   2.4  0.52
## Petal.Width     4 50 2.03 0.27   2.00    2.03 0.30 1.4 2.5   1.1 -0.12
## Species*        5 50 3.00 0.00   3.00    3.00 0.00 3.0 3.0   0.0   NaN
##              kurtosis   se
## Sepal.Length    -0.20 0.09
## Sepal.Width      0.38 0.05
## Petal.Length    -0.37 0.08
## Petal.Width     -0.75 0.04
## Species*          NaN 0.00
##
```

```
## attr(,"call")
## by.data.frame(data = x, INDICES = group, FUN = describe, type = type)
```

With this you can get a nice, comprehensive table of summary statistics across all the numerical columns, divided by the chosen factor.

For combinations of factors, you can use the `ddply()` function in the `plyr` package. Here you can choose a list of factors to summarise, but you must name the output columns and the R function to use. Helpfully the R function for a mean is `mean()` and the function for standard deviation is `sd()`.

Here, we divide up on `cut` and `colour` using the make-a-list function `c()`, we tell `ddply` we want to `summarise` and that it should add a `mean` column using the `mean()` function and an `sd` column using the `sd(function)`

```
ddply(diamonds, c('cut', 'color'), summarise, mean=mean(price), sd=sd(price) )
```

```
##           cut color     mean       sd
## 1       Fair     D 4291.061 3286.114
## 2       Fair     E 3682.312 2976.652
## 3       Fair     F 3827.003 3223.303
## 4       Fair     G 4239.255 3609.644
## 5       Fair     H 5135.683 3886.482
## 6       Fair     I 4685.446 3730.271
## 7       Fair     J 4975.655 4050.459
## 8       Good     D 3405.382 3175.149
## 9       Good     E 3423.644 3330.702
## 10      Good     F 3495.750 3202.411
## 11      Good     G 4123.482 3702.505
## 12      Good     H 4276.255 4020.660
## 13      Good     I 5078.533 4631.702
## 14      Good     J 4574.173 3707.791
## 15 Very Good     D 3470.467 3523.753
## 16 Very Good     E 3214.652 3408.024
## 17 Very Good     F 3778.820 3786.124
## 18 Very Good     G 3872.754 3861.375
## 19 Very Good     H 4535.390 4185.798
## 20 Very Good     I 5255.880 4687.105
## 21 Very Good     J 5103.513 4135.653
## 22   Premium     D 3631.293 3711.634
## 23   Premium     E 3538.914 3794.987
## 24   Premium     F 4324.890 4012.023
## 25   Premium     G 4500.742 4356.571
## 26   Premium     H 5216.707 4466.190
## 27   Premium     I 5946.181 5053.746
## 28   Premium     J 6294.592 4788.937
## 29     Ideal     D 2629.095 3001.070
## 30     Ideal     E 2597.550 2956.007
## 31     Ideal     F 3374.939 3766.635
## 32     Ideal     G 3720.706 4006.262
## 33     Ideal     H 3889.335 4013.375
## 34     Ideal     I 4451.970 4505.150
## 35     Ideal     J 4918.186 4476.207
```

11.3 `mean()` and `sd()`

R has even simpler functions for getting these values from lists of numbers, the `mean()` and `sd()` functions. These take lists of numbers, rather than whole datasets: `mean(c(1,2,3))`

```
[1] 2
```

or

```
mean(iris$Sepal.Length)
```

```
[1] 5.843333
```

These are useful for quick calculations, but less so for real datasets since you have to apply manually to every subset that you are interested in.

11.4 The independent *t* test

Let's look at the standard independent t test. This is done with the `t.test()` function. Let's look at the `iris` built in data set.

The general form for a tidy dataset with factors like `iris` is `t.test(measured_thing ~ grouping_factor, data=dataset)`, so to compare Petal Widths between Species in the `iris` dataset

```
t.test(Petal.Width ~ Species, data = iris)
```

```
## Error in t.test.formula(Petal.Width ~ Species, data = iris): grouping factor must have exactly 2 leve
```

Recall that this data set has 3 levels in the Species factor. The t test, is built for one to one comparisons only, so we must have just two things to compare (and if you end up doing more than one t-test to compare lots of groups, you probably should be using ANOVA!).

So lets use the `filter()` function from the `dplyr` package to select our two

```
library(dplyr)
small_iris <- filter(iris, Species == c("versicolor", "setosa"))
t.test(Petal.Width ~ Species, data=small_iris)
```

```
##
##  Welch Two Sample t-test
##
## data:  Petal.Width by Species
## t = -23.556, df = 41.092, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1812722 -0.9947278
## sample estimates:
##     mean in group setosa mean in group versicolor
##                    0.264                    1.352
```

The output is fairly clear, the `p-value` is significant at some number less than 0.000000000000000022 (which is a built in lower limit that is caused by R running out of decimal places).

The mean in each group is presented for each group, is the limits of the confidence interval of the range we would expect the true difference to fall between 95% of the time (this is not the CI of the mean of each data set!).

11.4.1 Effect size for a t test.

Even though the p value is significant at the 95% level, we do not know whether it is a substantial difference. This is an important thing to assess too since small differences can be statistically significant simply by recording a difference many times (IE by having a large number of replicates), but small differences may not be meaningful in the real world. Effect size is one way of getting a measure of the size of the difference that is more meaningful and comparable between different experiments and answering whether the difference you've observed is substantive as well as significant. A responsible scientist will use both p and effect size for deciding whether the difference between means has any real world meaning.

There are loads of effect sizes, we will look at two common ones for a t test, Cohen's d and Rosenthal's `r`.

11.4.1.1 Cohen's d

Cohen's d is a standardized measure of effect and can be thought of the number of standard deviations between the means. Different d mean different things:

~0.2

small effect

~0.5

medium effect

~0.8

large effect

The d can be calculated using the `cohen.d()` function in the `effsize` package and it works just like the `t.test()` function

```
library(effsize)
cohen.d(Petal.Width ~ Species, data=small_iris)
```

```
## Error in cohen.d.default(d, f, ...): Factor should have only two levels
```

Doing this throws an error, because of an inconsistency between the programmers of the `t.test()` function, the `filter()` function and the `cohen.d()` function. Even though we filtered our `iris` dataset to just two levels, the `levels` attribute of the `small_iris` dataset still says there are 3 levels!

```
levels(small_iris$Species)
```

```
## [1] "setosa"     "versicolor" "virginica"
```

Which is annoying. The `t.test()` function just worked out the right number of levels, but the `cohen.d()` checks the `levels` attribute instead. This inconsistency is just a thing that happens when different people write different code. So to fix the levels after filtering we must manually set them

```
small_iris$Species <- factor(small_iris$Species, levels = c("setosa", "versicolor"))
cohen.d(Petal.Width ~ Species, data=small_iris)
```

```
##
## Cohen's d
##
```

```
## d estimate: -6.662612 (large)
## 95 percent confidence interval:
##       inf       sup
## -8.173936 -5.151289
```

And `cohen.d()` tells us this is a large difference, with d = -6.663 (you can ignore the sign of the d).

11.4.1.2 Rosenthal's *r*

Another well used effect size is r. It is calculated with this formula, using t and df (degrees of freedom from the `t.test()`):

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Figure 4: R

Different values of r indicate different effect sizes, here's some rough rules.

~0.1

small effect

~0.3

medium effect

~0.5

large effect

The values for t and df can be extracted from the result of `t.test()` and the value of `r` calculated. Let's specify a custom function to do this for us.

Making a function Functions are the building blocks of R and it allows you to make your own. Its actually quite easy, its just a call to the function called `function()` and some regular R code surrounded by a couple of curly brackets - `{ }`. Here's how we'd define our own squaring function:

```
my_square <- function(x){ return(x * x) }
```

Now `my_square()` is a new function in R. We can use it like:

```
my_square(2) [1] 4
```

- The `x` is a temporary name for any data you want to use within the function.
- The `return()` keyword simply tells the function what to send you back here
- To send data into the function you simply put it in the brackets after the function name.
- The R code in the function can be as long as you want and split over many lines

Here's a custom function for r,

```
rosenthal_r <- function(t_test_result){
    t <- t_test_result$statistic[[1]]
    df <- t_test_result$parameter[[1]]
    r <- sqrt(t^2/(t^2+df))
    return(r)
}
```

Here's how we use it:

```
my_t <- t.test(Petal.Width ~ Species, data=small_iris)
rosenthal_r(my_t)
```

```
## [1] 0.9649093
```

Functions are useful because they are reuseable and save lots of typing. The actual value of r we get here is 0.96, which is considered a large effect.

11.5 Linear regression

Linear regression (or as some think of it - correlation analysis) is at the base of a whole lot of statistics. In R it's done with the `lm()` function. `lm` stands for 'linear model' as the resulting formula gives us a line in the form `y=mx+c` that describes the relationship between our input data and can be used as a model to make predictions. Doing an `lm()` is similar to the `t.test()`.

```
lm(Petal.Width ~ Petal.Length, data=iris)
```

```
##
## Call:
## lm(formula = Petal.Width ~ Petal.Length, data = iris)
##
## Coefficients:
##  (Intercept)  Petal.Length
##      -0.3631        0.4158
```

Which is the minimal information we need the first number (Intercept) -0.3631 is the `c` of `y=mx+c`, the place where the axis is crossed and the other (Petal.Length, 0.4158) is the gradient of the line. But we also want the p and R values, these are obtained by using the summary function on the result of `lm()`.

```
l <- lm(Petal.Width ~ Petal.Length, data=iris)
summary(l)
```

```
##
## Call:
## lm(formula = Petal.Width ~ Petal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56515 -0.12358 -0.01898  0.13288  0.64272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
## Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2065 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

Where we can see that the R is 0.9266 and p is again $<$ 2.2e-16.

The `lm()` function works fine for a dataset with only one factor, in this example we've completely ignored the fact that the `Petal.Length` and `Petal.Width` are from 3 species. If we want to do the linear model species wise, we need to filter the data first

```
ve <- filter(iris, Species == "versicolor")
lm(Petal.Width ~ Petal.Length, data=ve)
```

```
##
## Call:
## lm(formula = Petal.Width ~ Petal.Length, data = ve)
##
## Coefficients:
##  (Intercept)  Petal.Length
##     -0.08429       0.33105
```

11.6 One Way ANOVA

The ANOVA is actually based on a linear model so we need to construct that first, again with the `lm()` function. This time we use the factor we want to use to split the data up as the second variable. We do the actual ANOVA with `aov()` function and get the answers with the `summary.lm()` function on that.

```
l <- lm(Petal.Width ~ Species, data = iris)
a <- aov(l)
summary.lm(a)
```

```
##
## Call:
## aov(formula = l)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.626 -0.126 -0.026  0.154  0.474
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.24600    0.02894    8.50 1.96e-14 ***
## Speciesversicolor  1.08000    0.04093   26.39  < 2e-16 ***
## Speciesvirginica   1.78000    0.04093   43.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2047 on 147 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.9279
## F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

Then we can apply Tukey's HSD to get the actual p-values between groups.

```
TukeyHSD(a)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = l)
##
## $Species
##                       diff       lwr       upr p adj
## versicolor-setosa    1.08 0.9830903 1.1769097     0
## virginica-setosa     1.78 1.6830903 1.8769097     0
## virginica-versicolor 0.70 0.6030903 0.7969097     0
```

where the column `diff` gives us the difference between means, the `lwr` and `upr` bounds are the confidence interval and `p adj` is the p value rounded off. Here Petal Widths are significantly different between the species.

11.7 Quiz

1. Using the `mtcars` dataset, work out whether the number of cylinders (cyl) in a car has an effect on the mpg (miles per gallon), specifically answer whether any difference is statistically significant and substantial. Hints: Remember to use the `str()` function to check factors are factors. A one-way ANOVA should be appropriate for this sort of data. The `F` statistic or ratio is a good estimate of effect size. Extra Credit: Do any of the warnings about 'unbalanced' design make you worry about the results of the test?