ChickR - ggplot2

S Bashir (ilustat.com)

27 Março 2017

ChickR

Welcome to ChickR

Setup

It is assumed that the latest version of R is installed on your computer. It is recommended that you install and use Rstudio IDE (integrated development environment) for the ChickR series. Other IDEs should work fine too. The first step is to install some packages (*addons*/*apps*) by typing the following in the console:

```
install.packages("dplyr")
install.packages("ggplot2")
```

You only need to install packages **once**. Without going into details, we next need to *activate* some of these packages to use by typing the following in the console:

```
library(dplyr)
library(ggplot2)
```

Chick Weight Data Objective

Our objective is to investigate the effect of four different diets on the chick weights over a 21 day period.

Chick Weight Data

The `ChickWeight` data is one of the many datasets included as part of R and are from a weight gain experiment for chicks. Let's look at the data:

```
data("ChickWeight")
glimpse(ChickWeight)
```

```
## Observations: 578
## Variables: 4
## $ weight <dbl> 42, 51, 59, 64, 76, 93, 106, 125, 149, 171, 199, 205, 4...
## $ Time   <dbl> 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 21, 0, 2, 4, 6, ...
## $ Chick  <ord> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2...
## $ Diet   <fctr> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
```

We can see that there are 4 variables and 578 observations. Each chick in the experiment is uniquely identified by the `Chick` variable (**R is case sensitive**) and they are randomly assigned to be fed one of four diets (`Diet`). Their `weight` (in grams) is measured over `Time` (from day zero to day 21).

```
head(ChickWeight)
```

```
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

```
tail(ChickWeight)
```

```
##     weight Time Chick Diet
## 573    155   12    50    4
## 574    175   14    50    4
## 575    205   16    50    4
## 576    234   18    50    4
## 577    264   20    50    4
## 578    264   21    50    4
```

```
head(select(ChickWeight, Chick, Diet, Time, weight))
```

```
##   Chick Diet Time weight
## 1     1    1    0     42
## 2     1    1    2     51
## 3     1    1    4     59
## 4     1    1    6     64
## 5     1    1    8     76
## 6     1    1   10     93
```

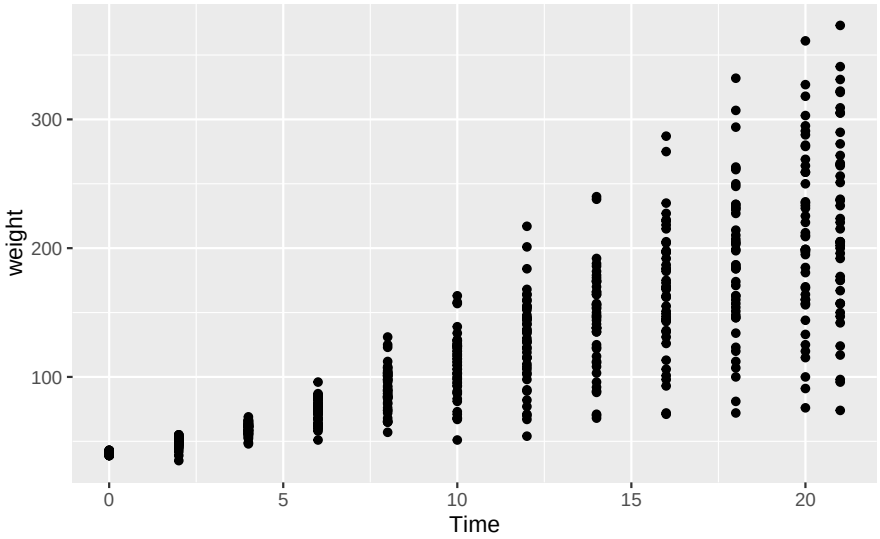
```
tail(select(ChickWeight, Chick, Diet, Time, weight))
```

```
##     Chick Diet Time weight
## 573    50    4   12    155
## 574    50    4   14    175
## 575    50    4   16    205
## 576    50    4   18    234
## 577    50    4   20    264
## 578    50    4   21    264
```

Graphical Exploration

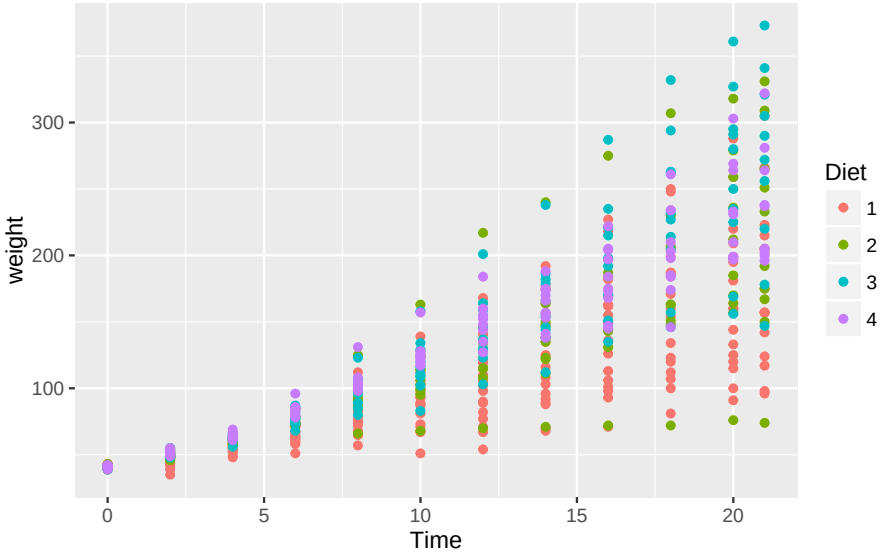
We will start by looking the raw data graphically using the `ggplot2` package using some relatively simple plots. At this stage don't worry too much about the details of the commands just try to build your own understanding.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_point()
```



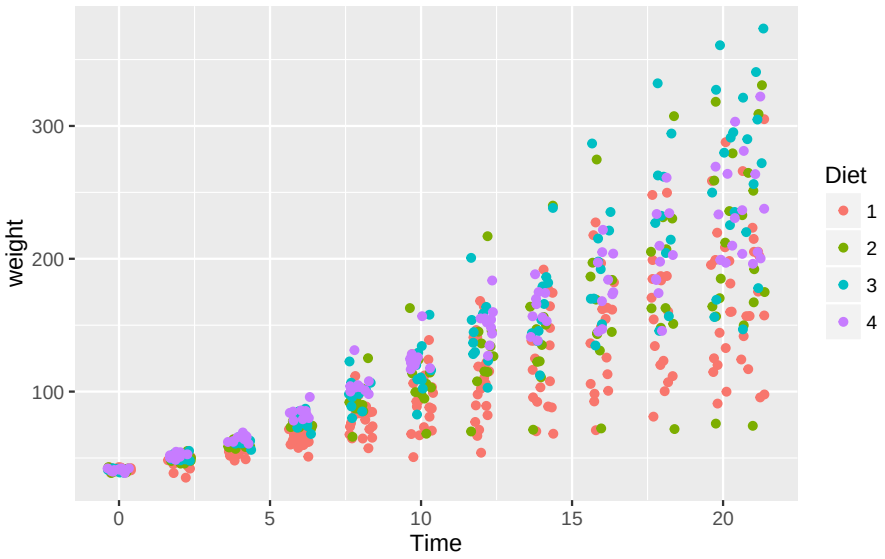
From the above ***scatter plot*** we can see that in general chick weights (vertical axis) increase over time (horizontal axis) however it does not tell us much about the effect of diet. Let's identify the different diets using some colour coding.

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet)) + geom_point()
```



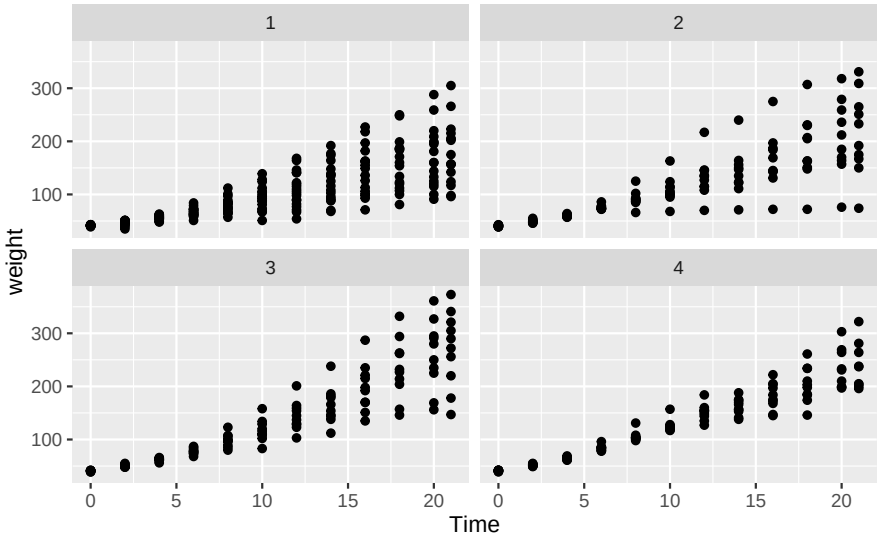
This is better but it is hard to make out the real effect of the diet as there are many overlapping points. We can introduce some jitter (i.e. *shake* the points to identify the overlapping points).

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet)) + geom_jitter()
```



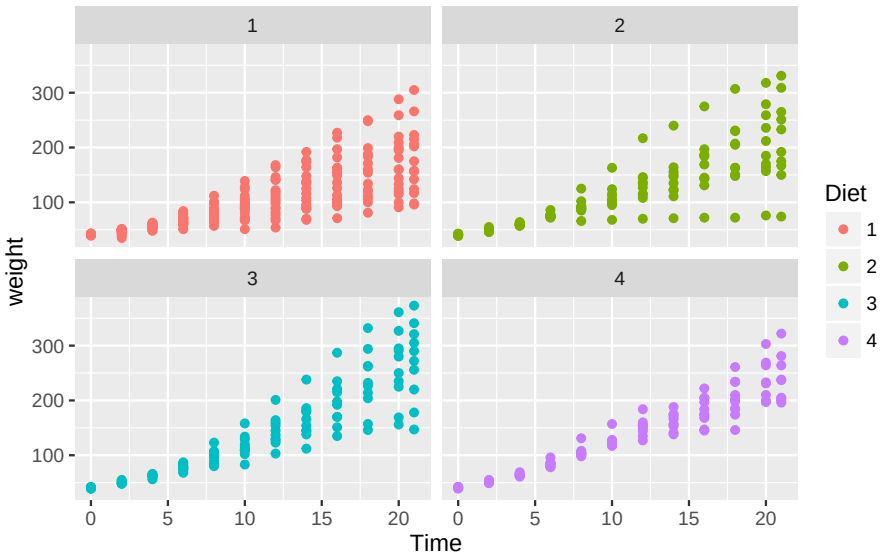
Overlapping is not a major issue here but this looks like four hives of bees spreading out so still not easy to see what the effect of diet. Perhaps we can plot the each diet in a separate scatter plot.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_point() + facet_wrap(~Diet)
```



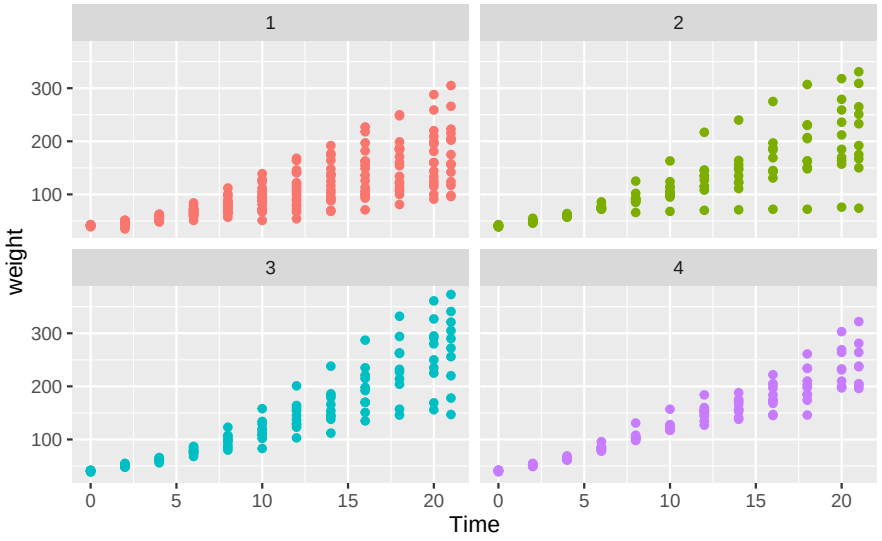
Perhaps it will look better with a bit of colour

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet)) + geom_point() + facet_wrap(~Diet)
```



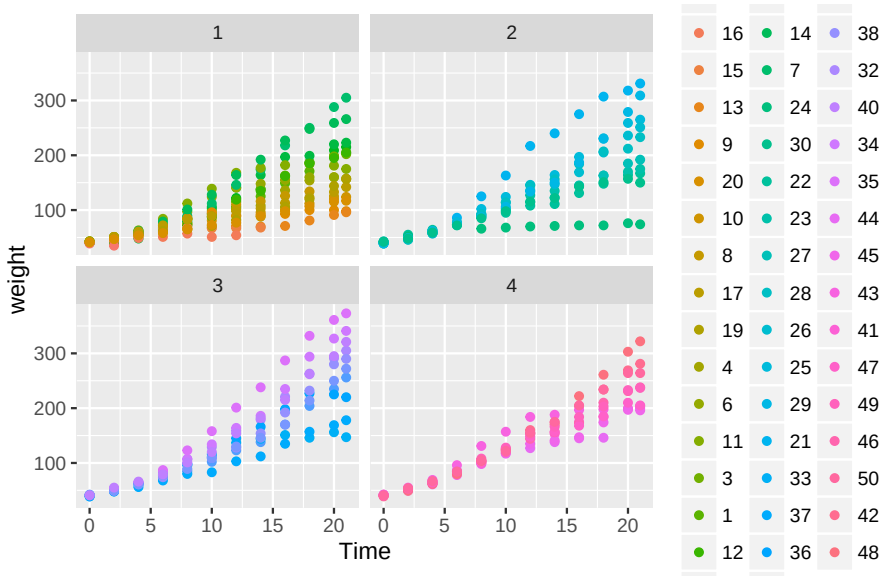
We don't need to have a legend as the `Diet` number is already given at the top of each plot so we can remove it.

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet)) + geom_point(show.legend = FALSE) +
  facet_wrap(~Diet)
```



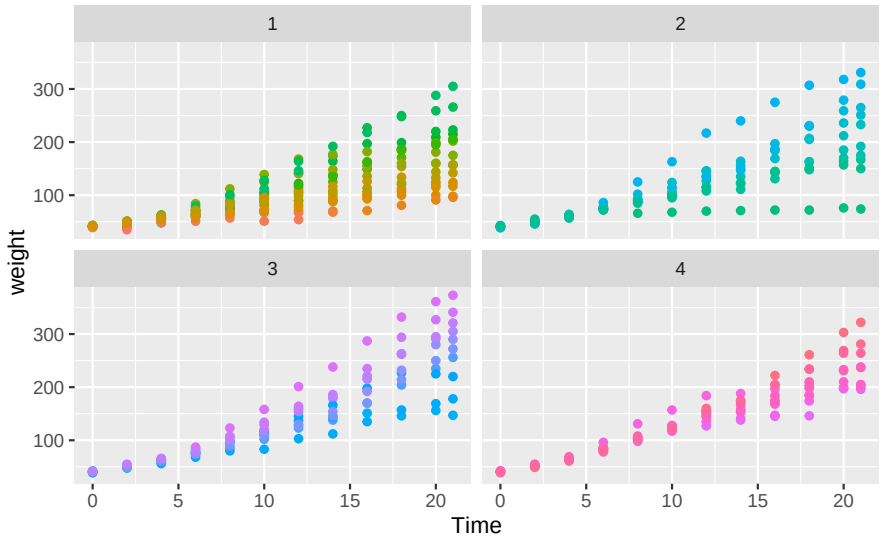
Perhaps it could help to identify each chick using a differenct colour.

```
ggplot(ChickWeight, aes(Time, weight, colour = Chick)) + geom_point() +
  facet_wrap(~Diet)
```



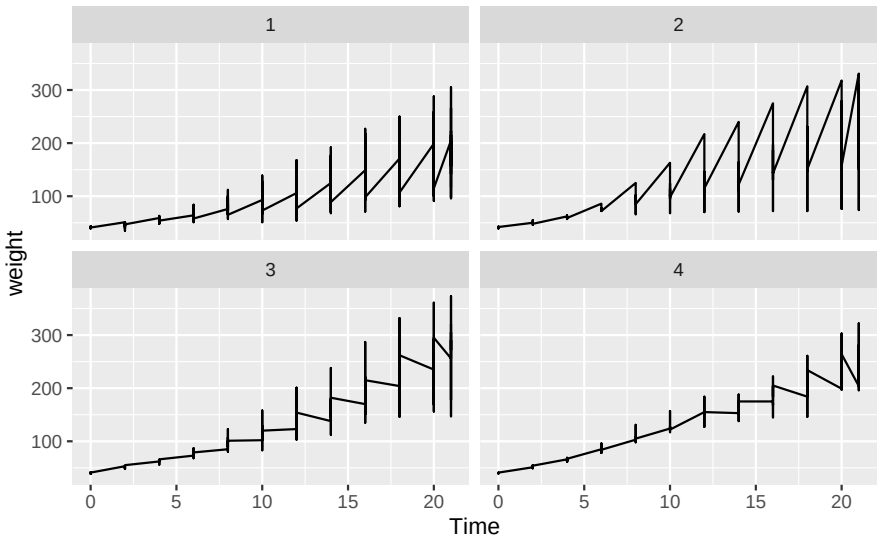
It sort of helps but there is still a lot of information to process and knowing the chick number for each colour does not really enhance understand so we can remove the legend.

```
ggplot(ChickWeight, aes(Time, weight, colour = Chick)) + geom_point(show.legend = FALSE) + face
```



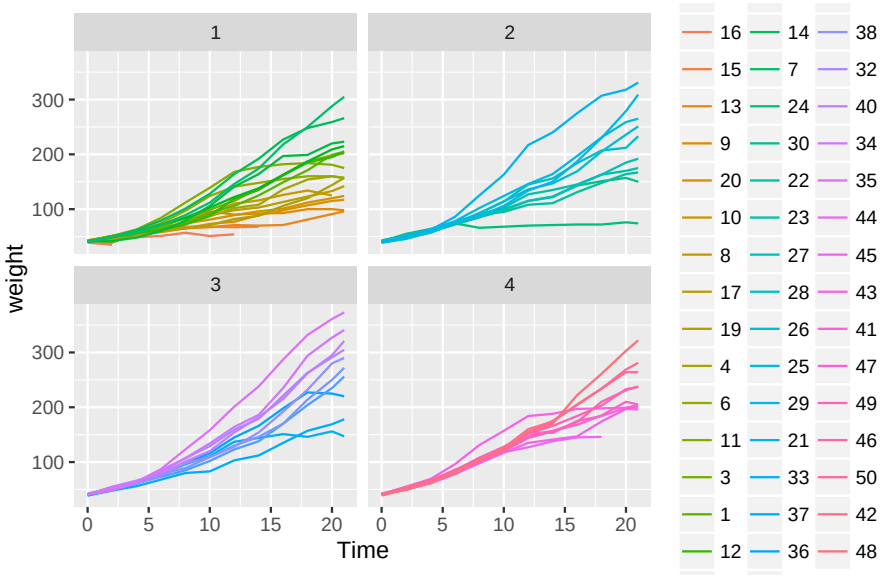
It can be hard to identify individual chicks based on the colour so let's use `lines` instead of `points`.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_line() + facet_wrap(~Diet)
```



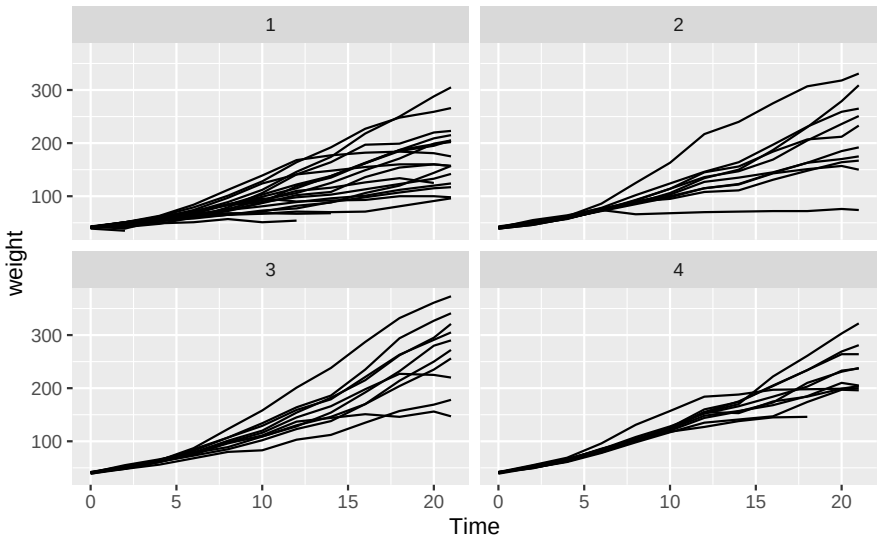
Whoops it is not what we expected. We want one line for each chick. Let's try again.

```
ggplot(ChickWeight, aes(Time, weight, colour = Chick)) + geom_line() +
  facet_wrap(~Diet)
```



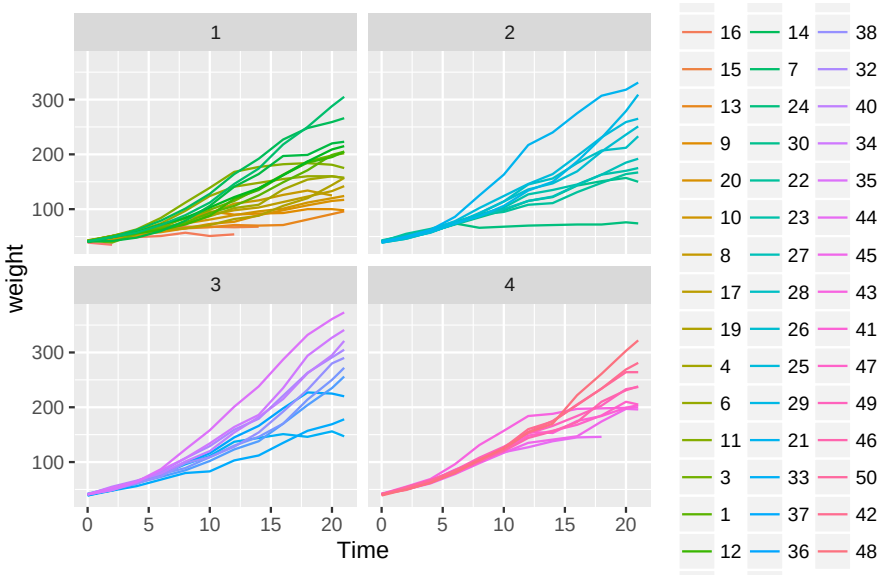
This is better but we will remove the legend.

```
ggplot(ChickWeight, aes(Time, weight, group = Chick)) + geom_line() +
  facet_wrap(~Diet)
```



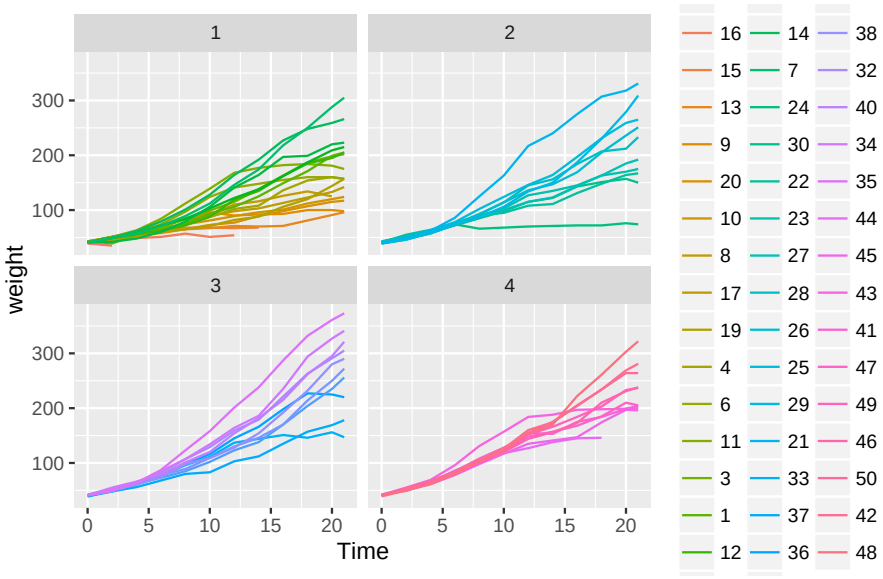
Oh... but now we've lost the colours.

```
ggplot(ChickWeight, aes(Time, weight, group = Chick, colour=Chick)) + geom_line() +
  facet_wrap(~Diet)
```



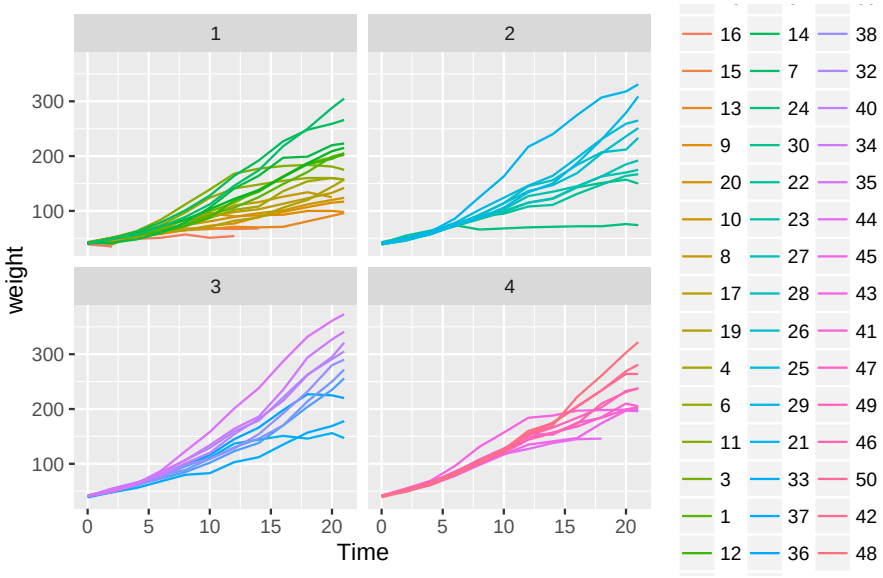
Perhaps the aesthetics (`aes`) need to be in the `geom_line` part.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_line(aes(group = Chick, colour=Chick)) +
  facet_wrap(~Diet)
```



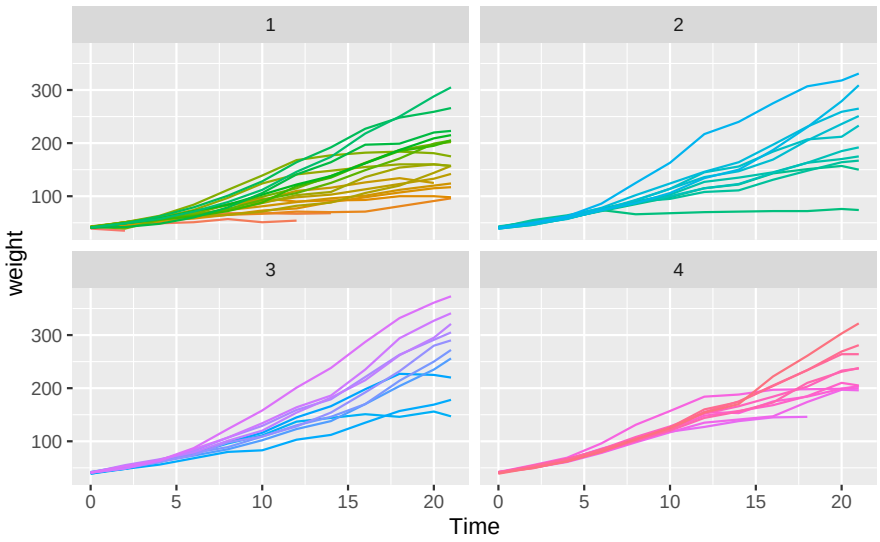
Black and white again but no legend. Yay. Let's upgrade to colour.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_line(aes(colour = Chick)) + facet_wrap(~Diet)
```



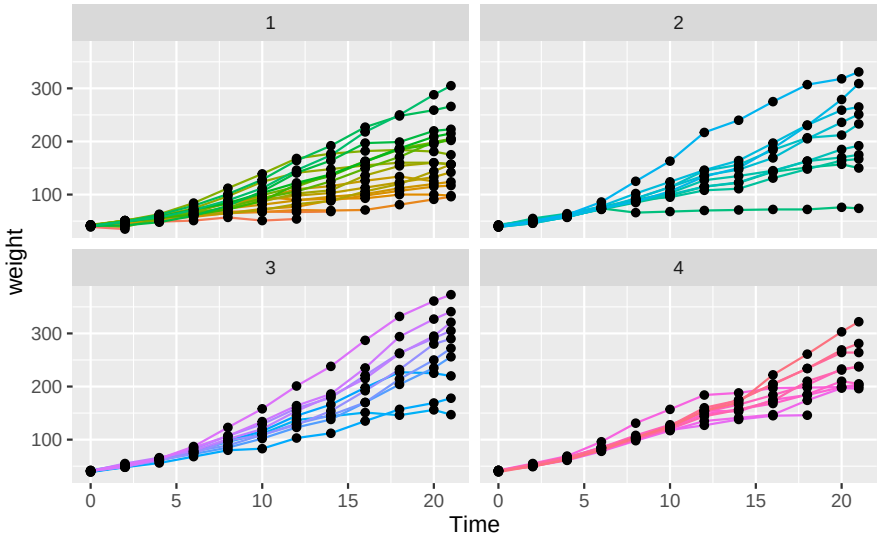
What is going on??? Colour but the legend is back. Mixing `colour` and `group` in `geom_line`.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_line(aes(colour = Chick), show.legend = FALSE) +
  facet_wrap(~Diet)
```



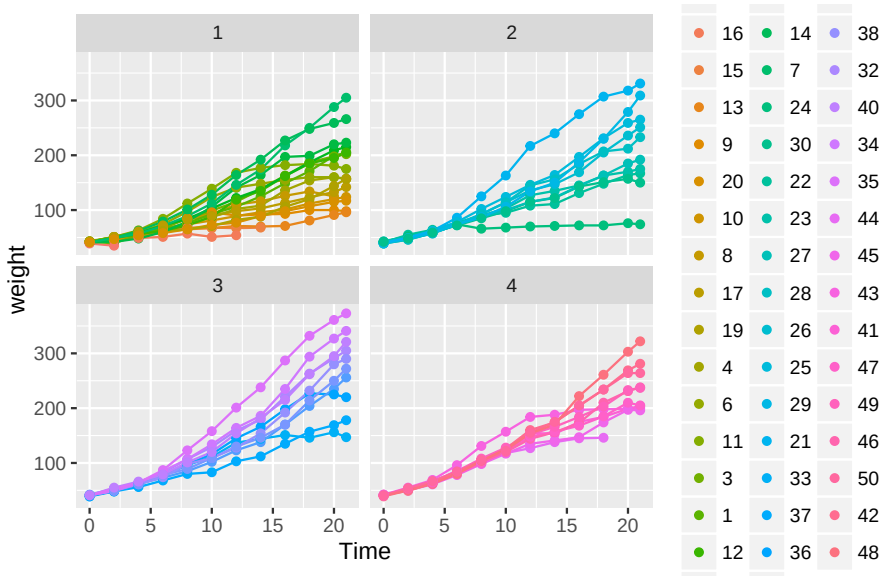
This gives us a better idea of how each chick performs. We also notice that some chicks don't have measurement up to day 21 (e.g. for diet 1 there is a chick who does not have measurements after day 12). Perhaps if we add points it will be clearer.

```
ggplot(ChickWeight, aes(Time, weight)) + geom_line(aes(colour = Chick), show.legend = FALSE) +
  geom_point() + facet_wrap(~Diet)
```



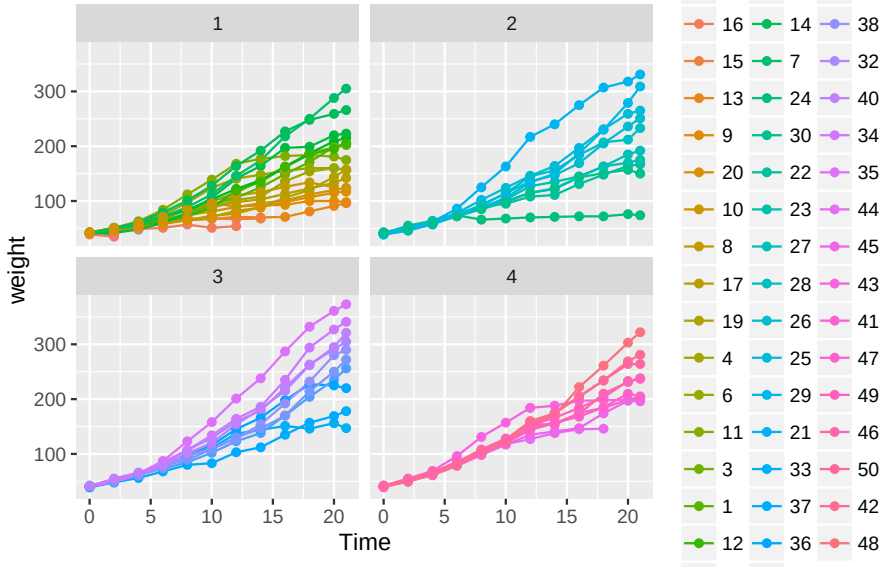
The points should be the same colour as the lines so

```
ggplot(ChickWeight, aes(Time, weight)) + geom_line(aes(colour = Chick), show.legend = FALSE) +
  geom_point(aes(colour = Chick)) + facet_wrap(~Diet)
```



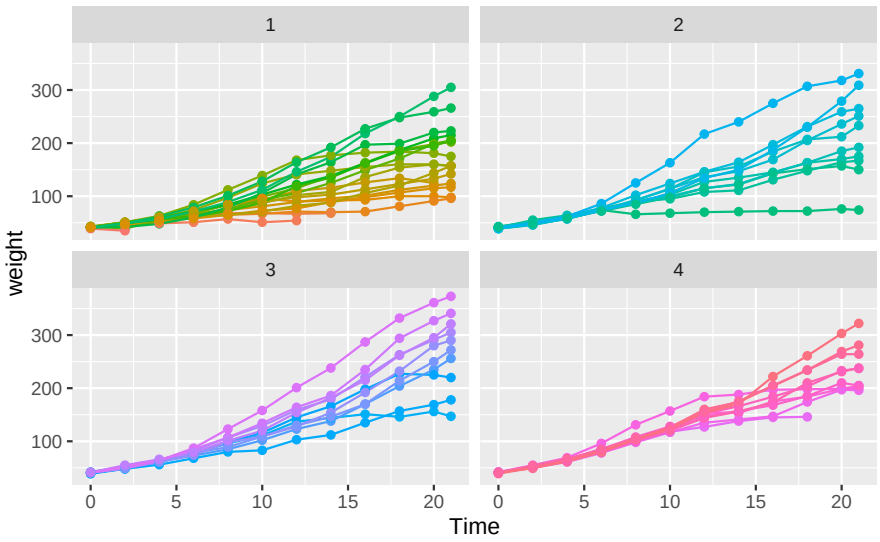
There are two problems here, the legend is back and the duplication of `aes(colour = Chick)`. Let's move it back to the `aes` in the `ggplot` part.

```
ggplot(ChickWeight, aes(Time, weight, colour = Chick)) + geom_line() + geom_point() +
  facet_wrap(~Diet)
```



Removing the legend yet again!

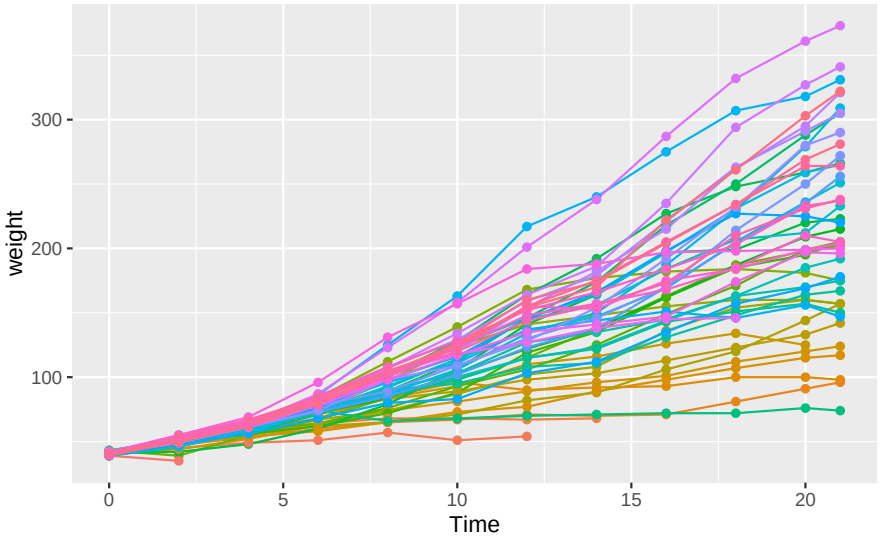
```
ggplot(ChickWeight, aes(Time, weight, colour = Chick)) + geom_line(show.legend = FALSE) +
  geom_point(show.legend = FALSE) + facet_wrap(~Diet)
```



This plot looks good enough for our initial assessments. Chicks gain weight over time for each of the diets. It seems although diet 4 has less variabilty than the others. Diet 3 could be the one that has the most weight gain over time but we would need to investigate the data further using summary statistics and some statistical modelling to confirm that.

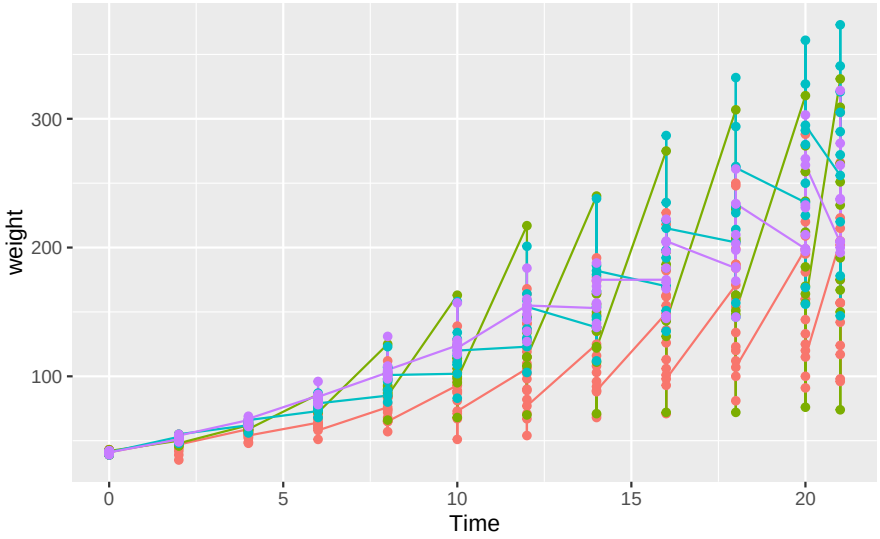
What do they look like in one graph (by removing `facet_wrap`).

```
ggplot(ChickWeight, aes(Time, weight, colour = Chick)) + geom_line(show.legend = FALSE) +
  geom_point(show.legend = FALSE)
```



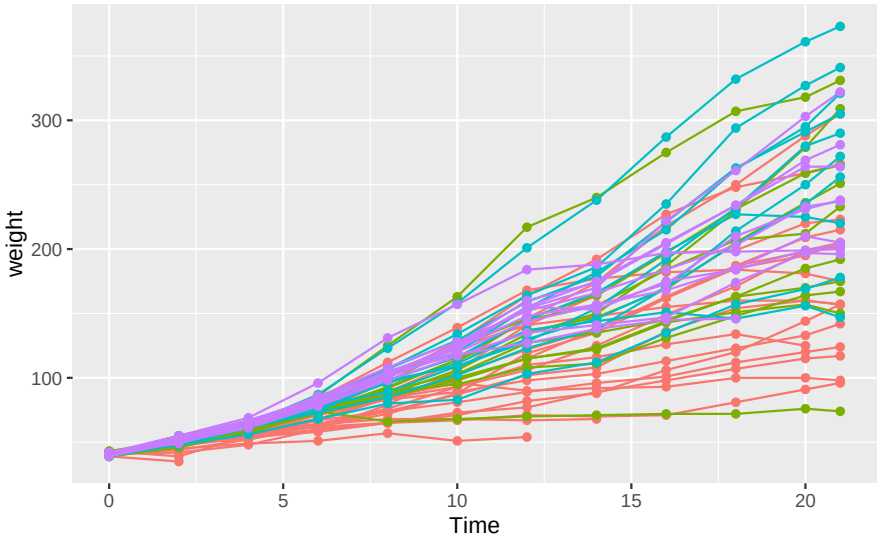
A spaghetti mess that does not tell us anything about the `diet`. Using `Diet` for `colour` instead of `Chick`.

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet)) + geom_line(show.legend = FALSE) +
  geom_point(show.legend = FALSE)
```



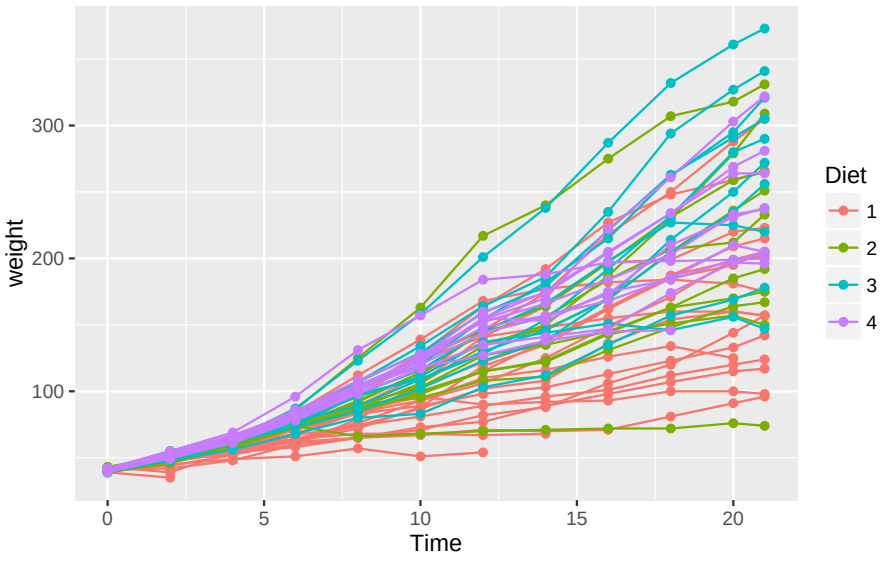
Not what was hoped for. Bring back the `Chick` variable to identify the unique lines and use `Diet` as the colour.

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet, group=Chick)) +
  geom_line(show.legend = FALSE) +
  geom_point(show.legend = FALSE)
```



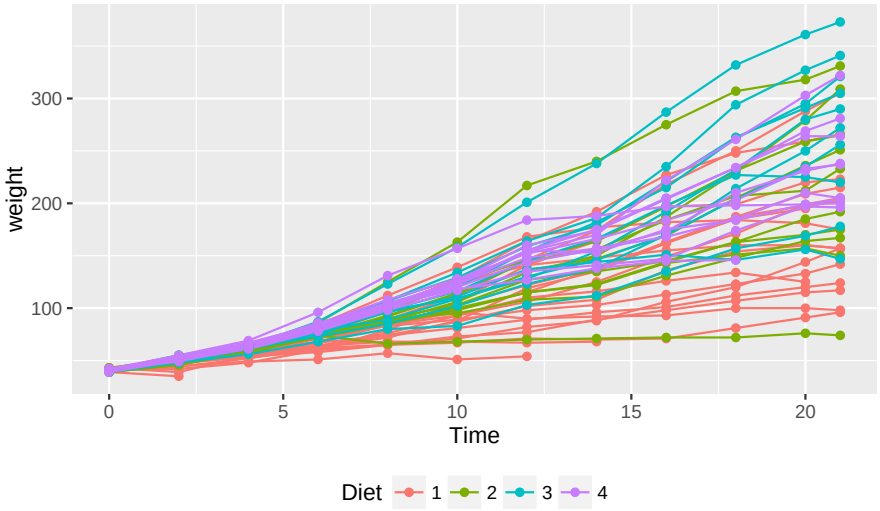
As the colour coding is diet it makes sense to bring back the legend by removing both `show.legend = FALSE`

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet, group=Chick)) +
  geom_line() + geom_point()
```



Move the legend to the bottom of the graph.

```
ggplot(ChickWeight, aes(Time, weight, colour = Diet, group=Chick)) + geom_line() +
  geom_point() + theme(legend.position = "bottom")
```



Summary

The objective is to "*investigate the effect of four different diets on the chick weights over a 21 day period*" using the chick weight (`ChickWeight`) dataset. We looked at different plots that represented the weight on the vertical axis and the time on the horizontal axis. `ggplot2` allowed us to plot different colours for the diet and/or split the data into four plots, one for each diet.