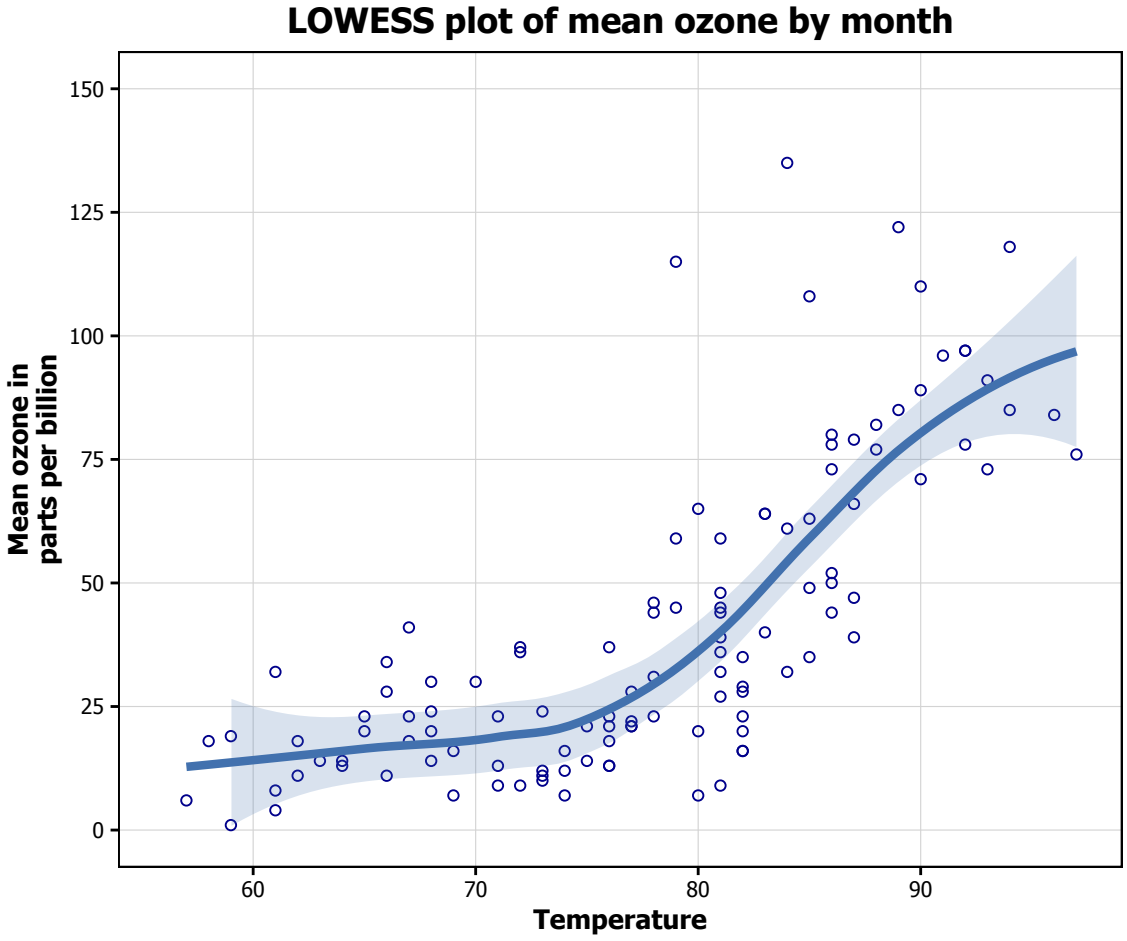
Creating plots in R using ggplot2 - part 12: LOWESS plots

Jodie Burchell
Mauricio Vargas Sepúlveda

2016-05-30

This is the twelfth and final tutorial in a series on using `ggplot2` I am creating with Mauricio Vargas Sepúlveda. In this tutorial we will demonstrate some of the many options the `ggplot2` package has for creating and customising lowess plots. LOWESS, or _lo_cally _we_ighted _s_catterplot _s_moothing, is a form of regression that creates different models for different subsets of the data. LOWESS plots represent this graphically by breaking down a continuous x-variable into a number of small bins and plotting the relationship with the y-variable individually for each subsection. Even if you don't intend to fit a LOWESS model, LOWESS plots are really useful for getting an initial feel for the relationship between your outcome and predictor, especially when you think that relationship isn't linear.

In this tutorial, we will work towards creating the LOWESS plot below using R's airquality dataset in the `datasets` package. We will take you from a basic LOWESS plot and explain all the customisations we add to the code step-by-step.



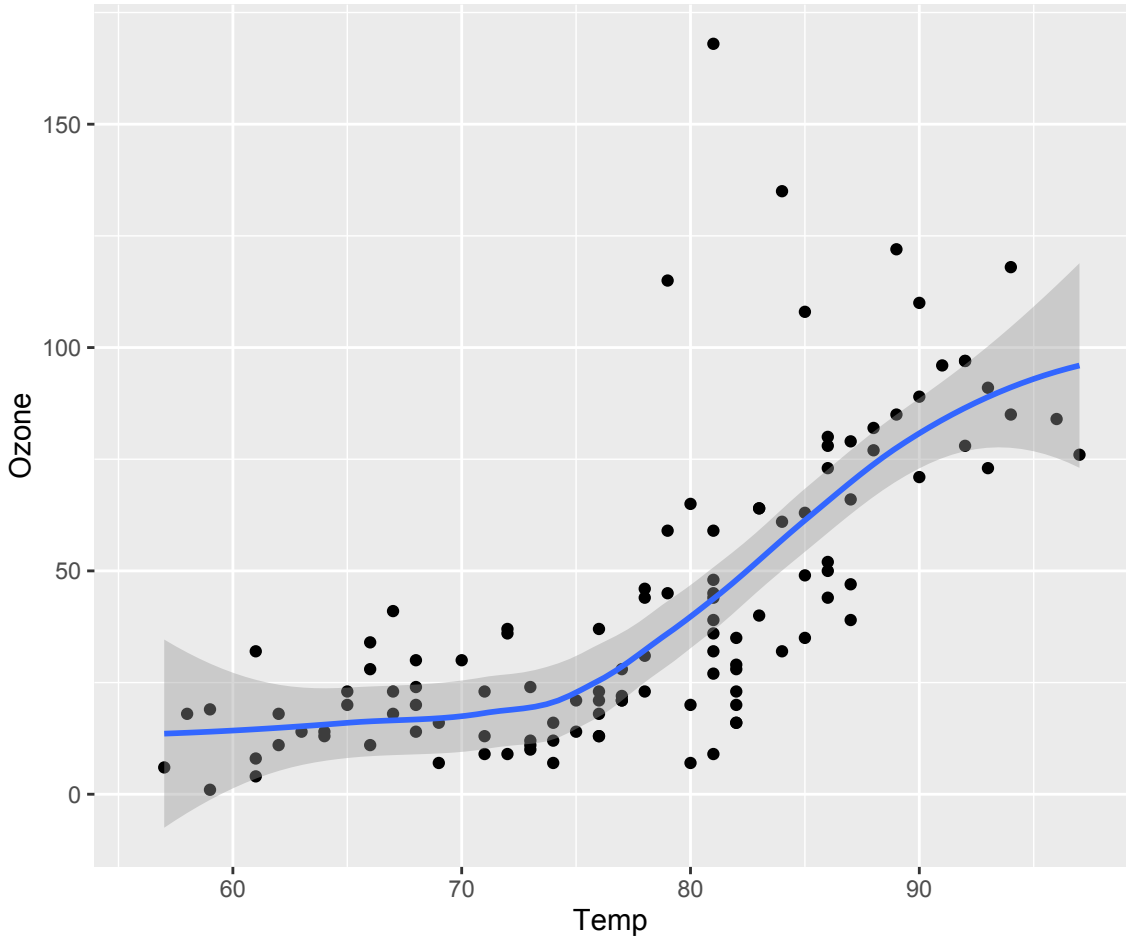
The first thing to do is load in the data and the libraries, as below. We'll convert `Month` into a labelled factor in order to use it as our grouping variable.

```
library(datasets)
library(ggplot2)
library(ggthemes)
library(grid)
library(RColorBrewer)

data(airquality)
```

Creating a basic LOWESS plot, and what it can tell us about our data

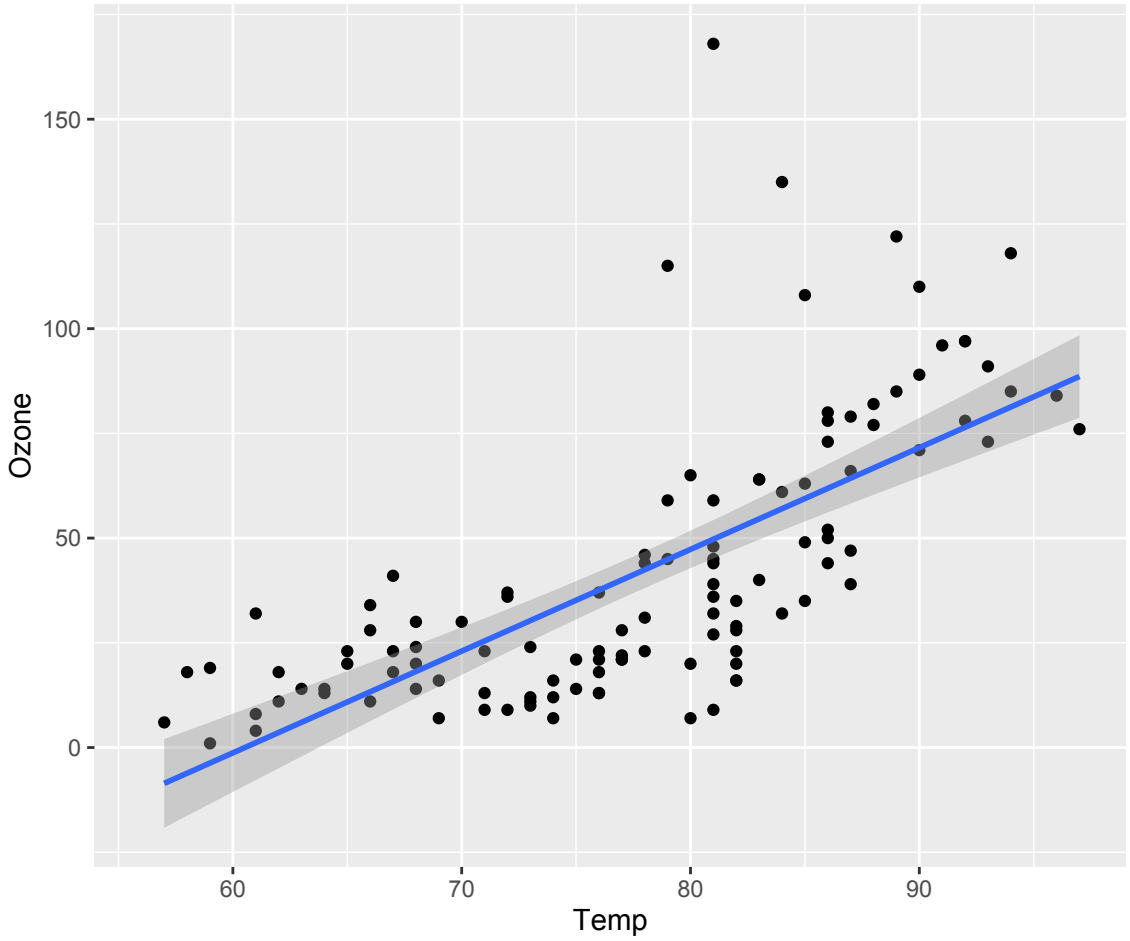
In order to initialise a plot we tell ggplot that `airquality` is our data, and specify that our x-axis plots the `Temp` variable and our y-axis plots the `Ozone` variable. We then instruct ggplot to render this as a LOWESS curve by adding the `stat_smooth(method = "loess")` option. Note that the default for `stat_smooth` is to include the confidence interval.

```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  stat_smooth(method = "loess")
p12
```



We can see that while the relationship between `Temp` and `Ozone` is fairly linear, the LOWESS plot is demonstrating there may be a threshold effect where ozone only starts increasing as temperatures pass around 75 degrees Fahrenheit. To assess whether this is the case, let's see how a standard linear fit between these variables looks.

```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method=lm)
p12
```



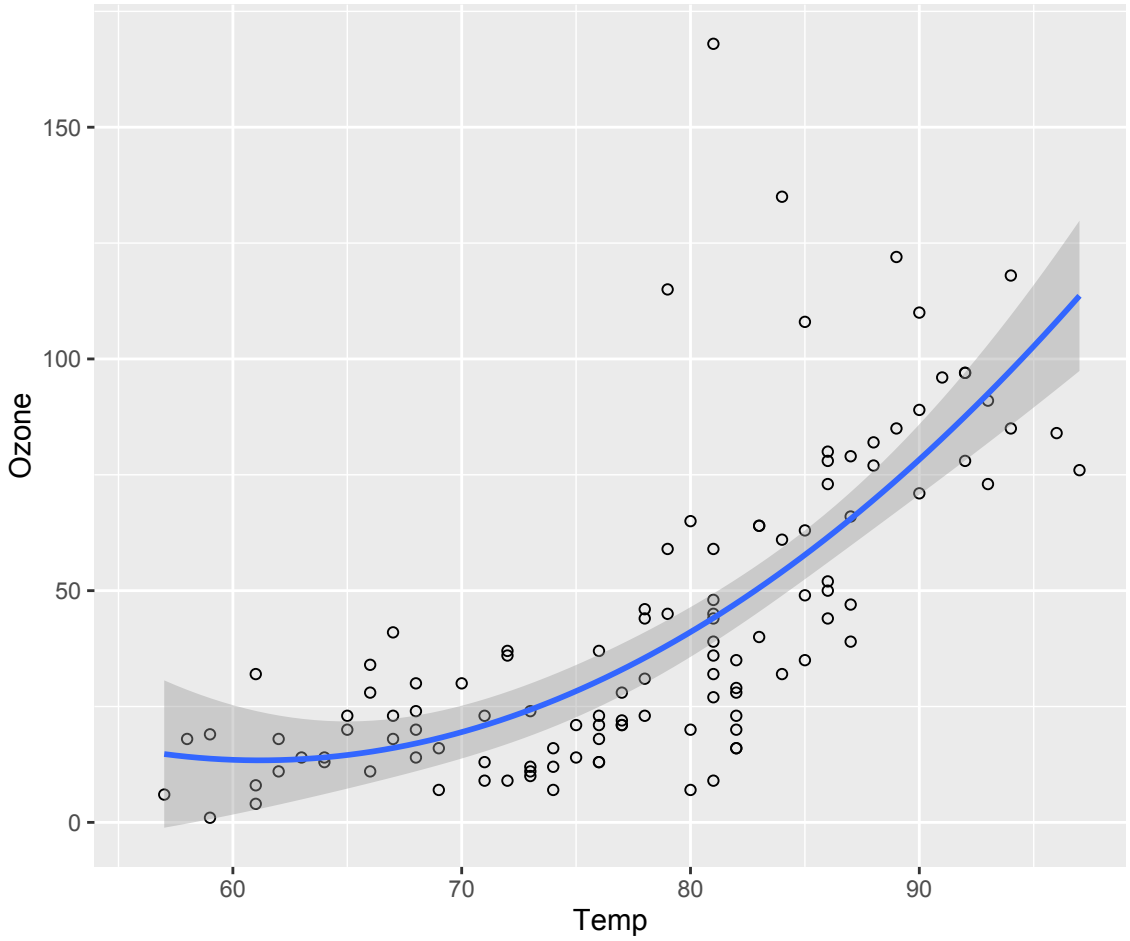
Let's now have a look at the amount of variance it explains in ozone levels by extracting the adjusted R^2 from the linear regression model between these two variables.

```
m1 <- summary(lm(Ozone ~ Temp, data = airquality))
m1$adj.r.squared
```

```
[1] 0.4832134
```

You can see that the line comes away from the data at several points, which will have increased the error in the regression model and brought down the overall R^2. Let's see whether we can get a better result by fitting a quadratic model.

```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point(shape=1) +
  stat_smooth(method = "lm", formula = y ~ x + I(x^2))
p12
```



You can see this fits the data *much* better. Let's see if the regression model confirms this:

```
m2 <- summary(lm(Ozone ~ Temp + I(Temp^2), data = airquality))
m2$adj.r.squared
```

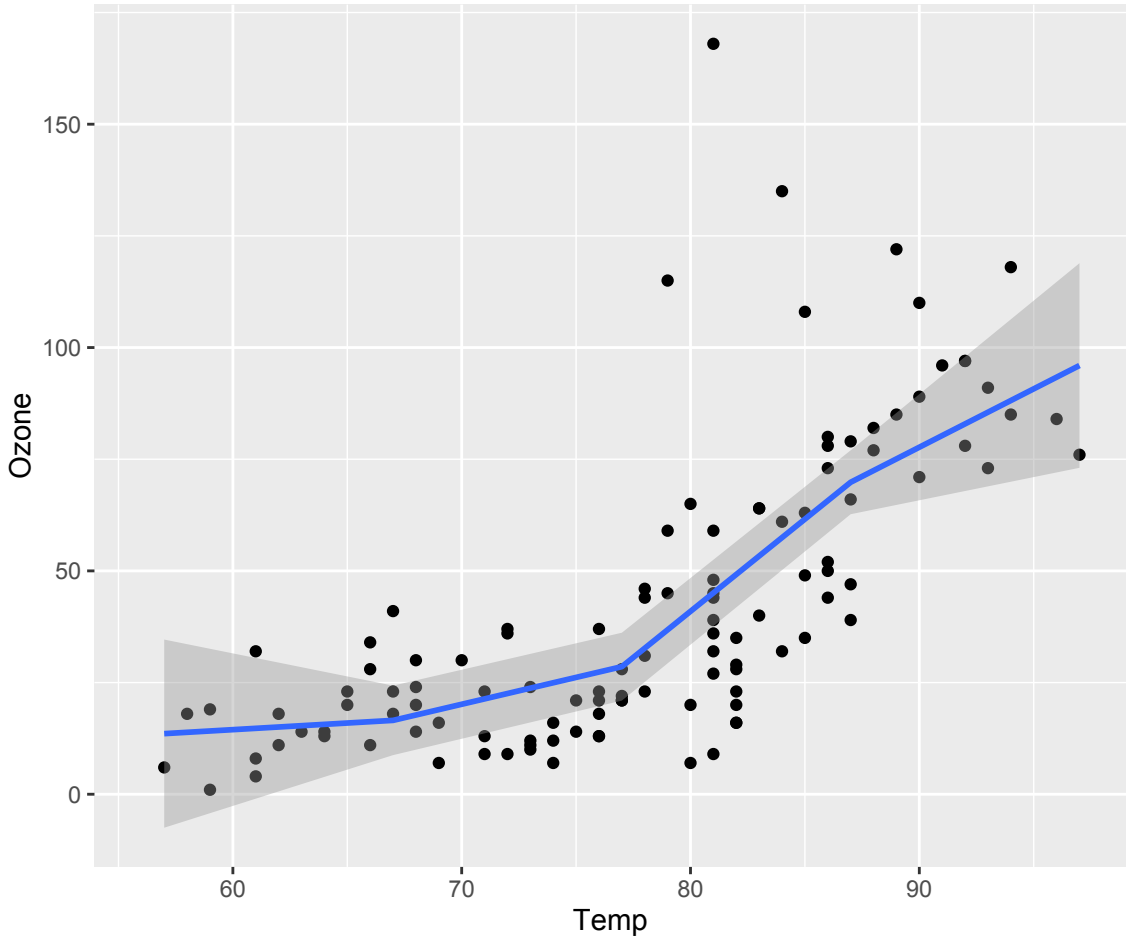
```
[1] 0.5361501
```

You can see that we've managed to explain an additional 5% of variance in ozone levels by fitting a quadratic model rather than defaulting to a linear model. Using LOWESS plots to explore the relationships between your variables can therefore guide you in choosing the the right regression model in a fairly pain-free way.

Changing the width of the bins

An important part of fitting LOWESS curves is that you can change the number of bins that the x-axis is divided into by using the argument `n`. More bins smooth out the line more, while less make it closer to linear. The default number is 80, and here we will change it to 5 so you can see the difference.

```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method = "loess", n = 5)
p12
```

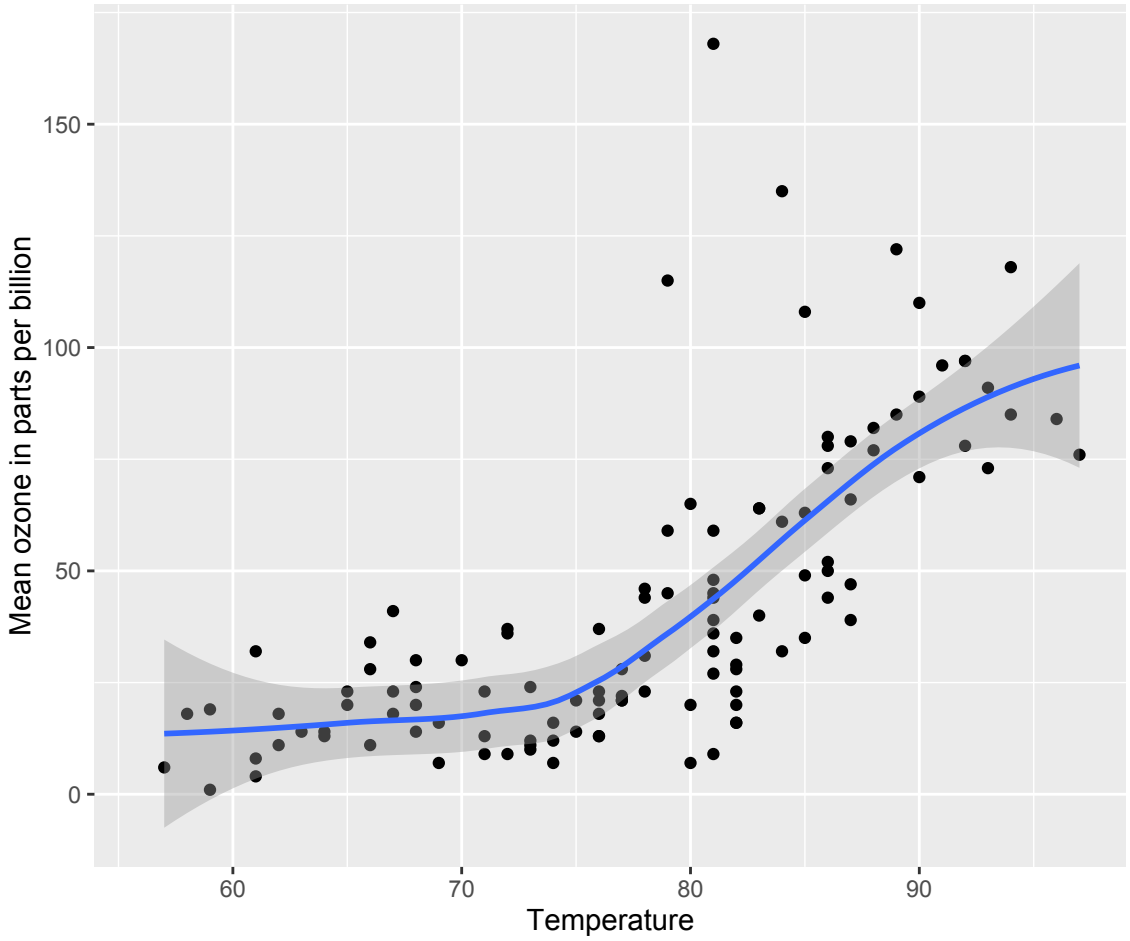


Customising axis labels

Now that we've established the rationale for using them, let's get down to customising our basic LOWESS plot.

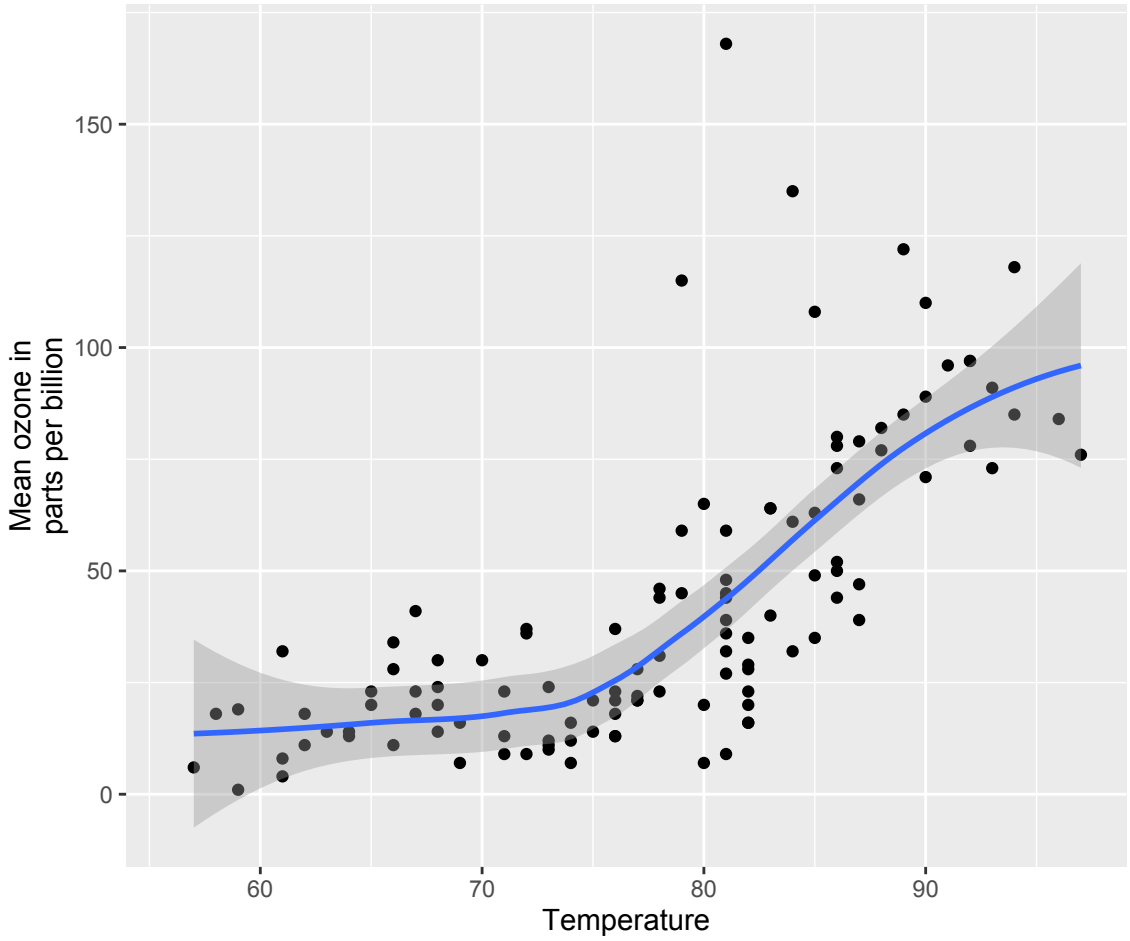
In order to change the axis labels, we have a couple of options. In this case, we have used the `scale_x_continuous` and `scale_y_continuous` options, as these have further customisation options for the axes we will use below. In each, we add the desired name to the `name` argument as a string.

```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  stat_smooth(method = "loess") +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in parts per billion")
p12
```



ggplot also allows for the use of multiline names (in both axes and titles). Here, we've changed the y-axis label so that it goes over two lines using the `\n` character to break the line.

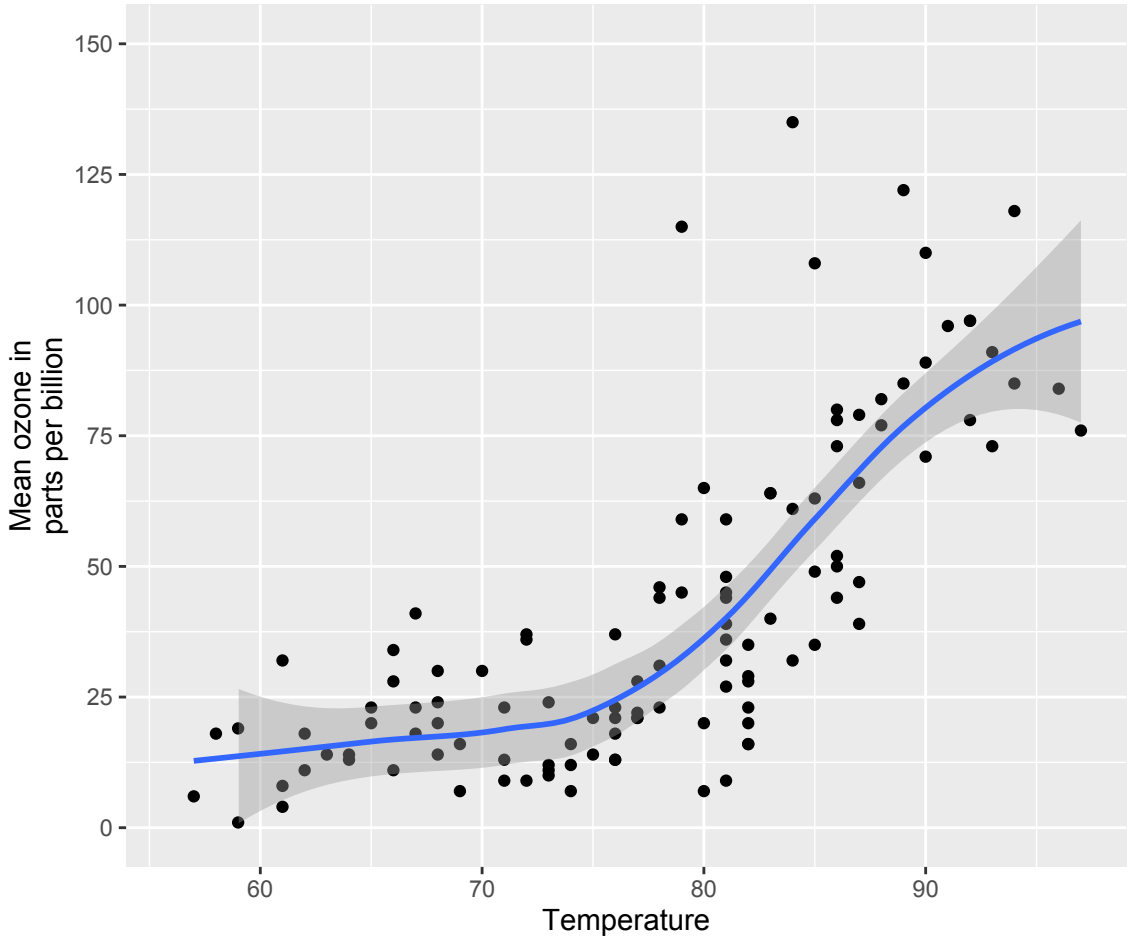
```
p12 <- p12 + scale_y_continuous(name = "Mean ozone in\nparts per billion")
p12
```



Changing axis ticks

The next thing we will change is the axis ticks. Let's make the y-axis ticks appear at every 25 units rather than 50 using the `breaks = seq(0, 150, 25)` argument in `scale_y_continuous`. (The `seq` function is a base R function that indicates the start and endpoints and the units to increment by respectively. See `help(seq)` for more information.) We ensure that the y-axis begins and ends where we want by also adding the argument `limits = c(0, 150)` to `scale_y_continuous`.

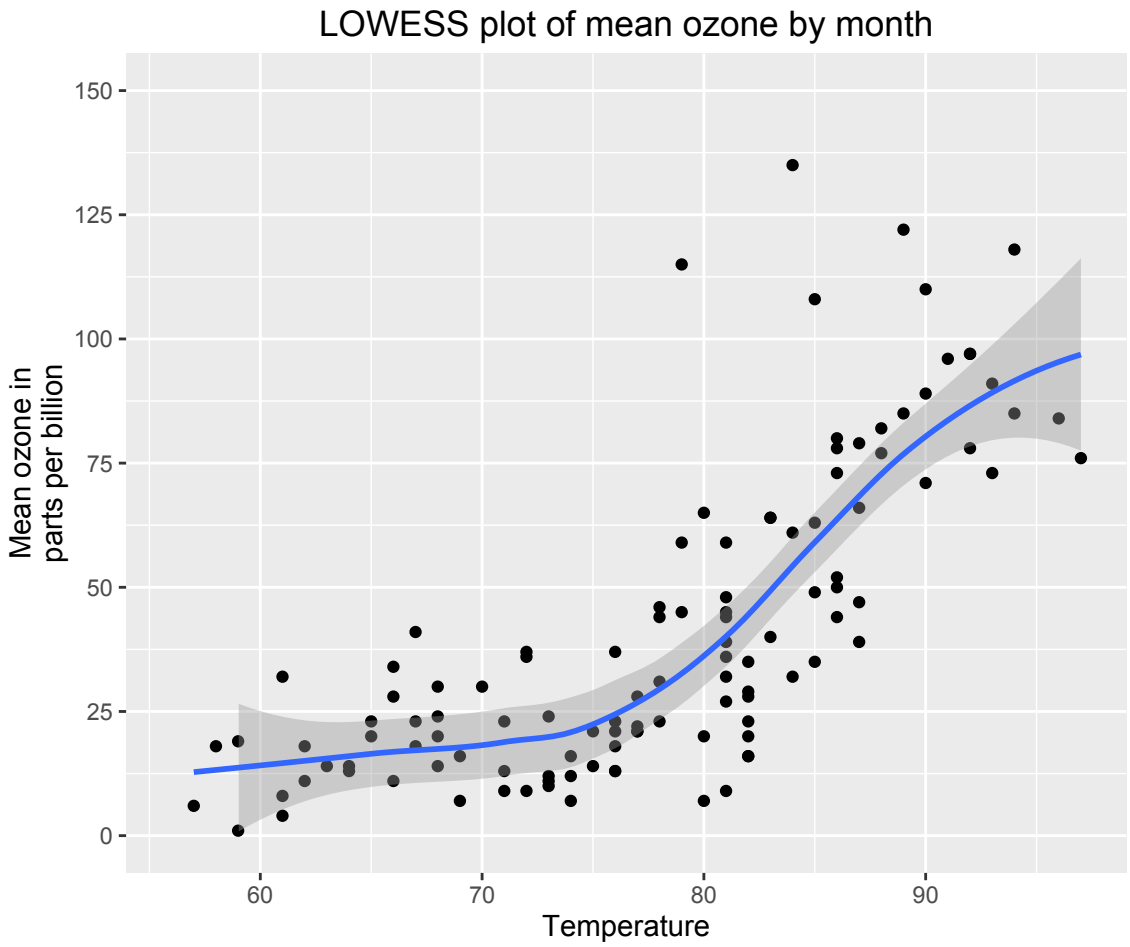
```
p12 <- p12 + scale_y_continuous(name = "Mean ozone in\nparts per billion",
                                breaks = seq(0, 150, 25), limits=c(0, 150))
p12
```



Adding a title

To add a title, we include the option `ggtitle` and include the name of the graph as a string argument.

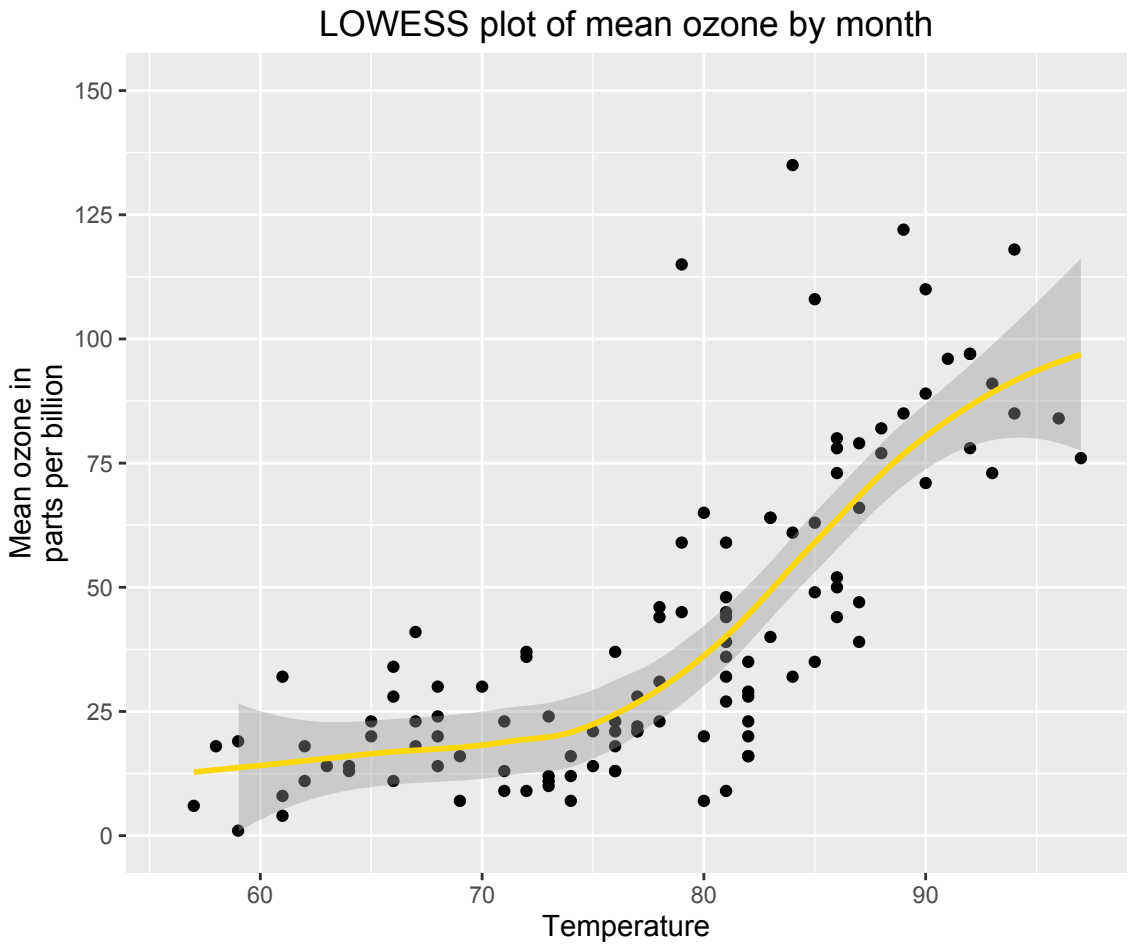
```
p12 <- p12 + ggtitle("LOWESS plot of mean ozone by month")
p12
```



Changing the colour and size of the LOWESS curve

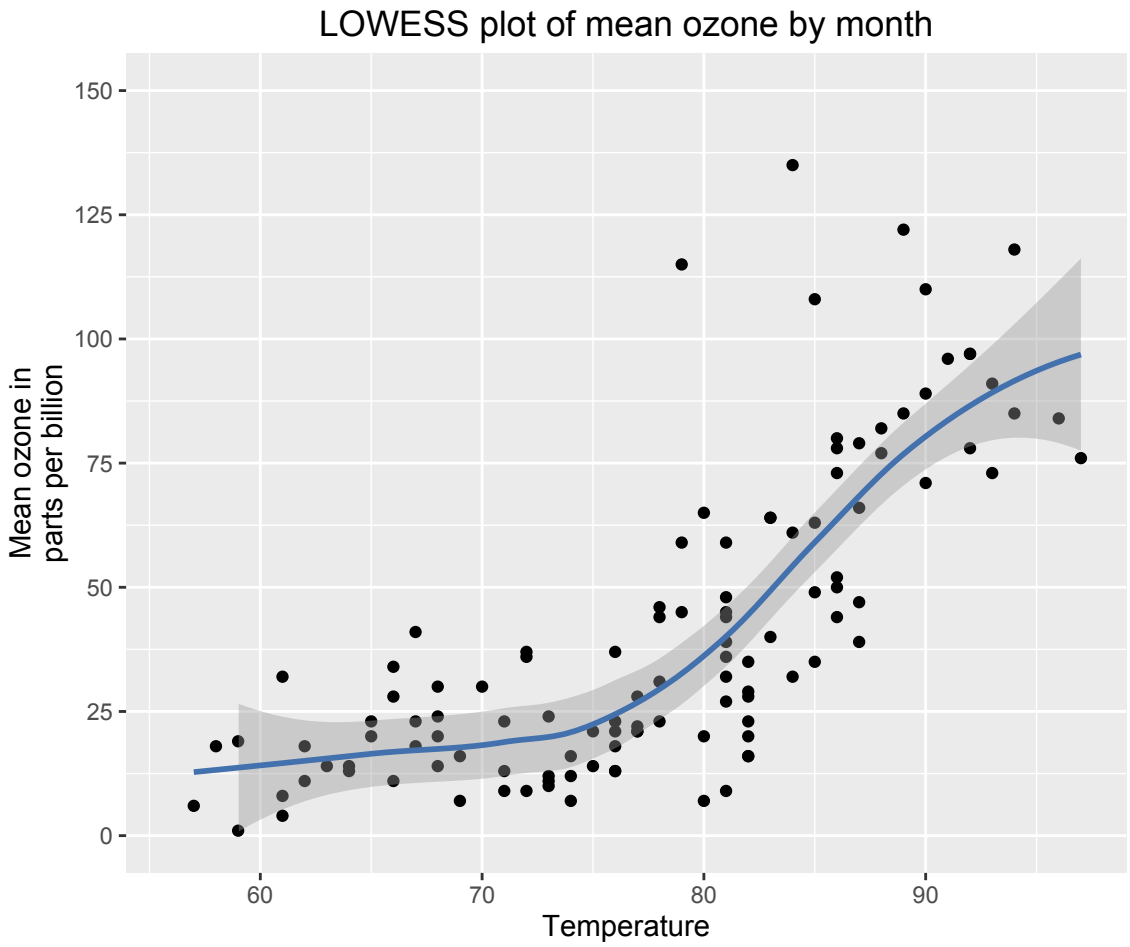
To change the colour of the LOWESS curve, we add a valid colour to the `colour` argument in `geom_smooth()` (note that we assigned this colour to a variable outside of the plot to make it easier to change it). A list of valid colours is here.

```
fill <- "gold1"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method = "loess", colour = fill) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



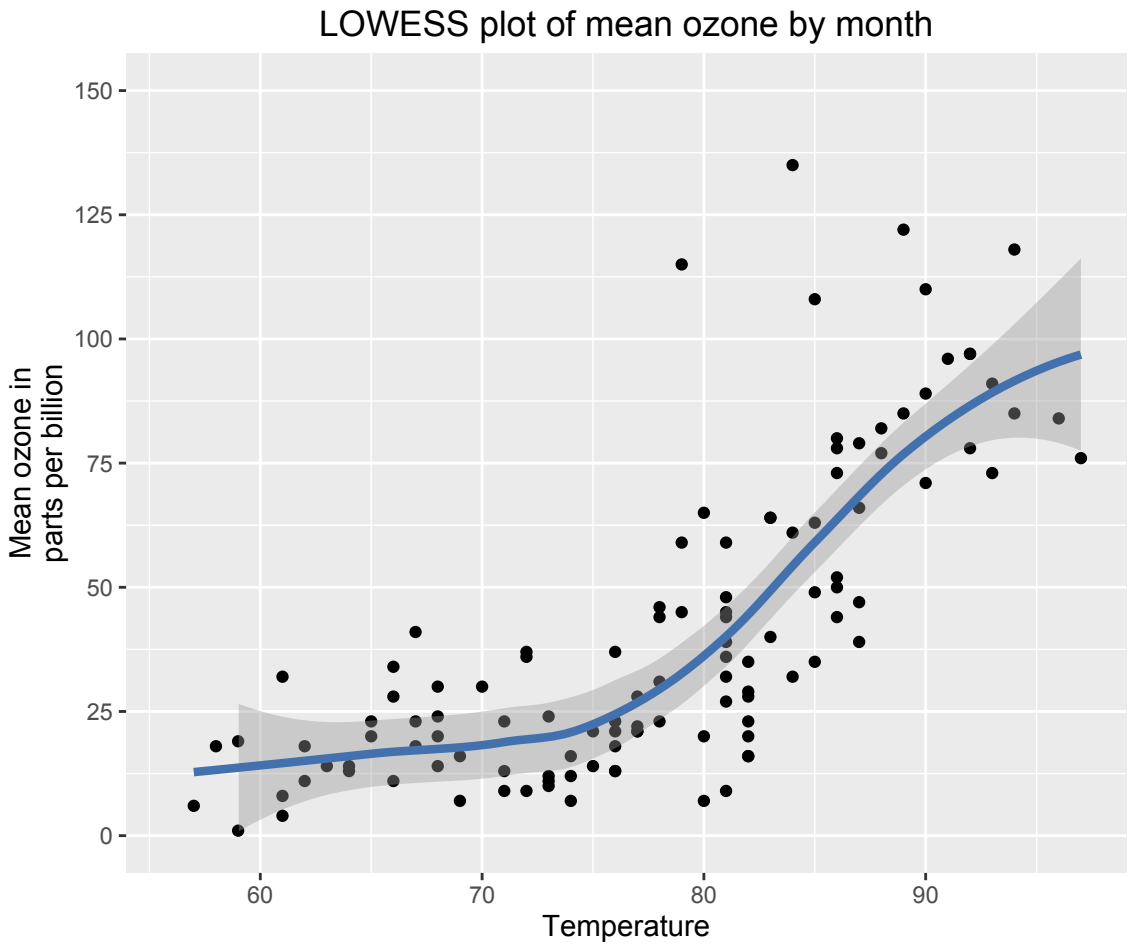
If you want to go beyond the options in the list above, you can also specify exact HEX colours by including them as a string preceded by a hash, e.g., "#FFFFFF". Below, we have called a shade of blue for the line using its HEX code.

```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method = "loess", colour = fill) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



We can also increase the thickness of the line using the `size` option in `geom_smooth()`.

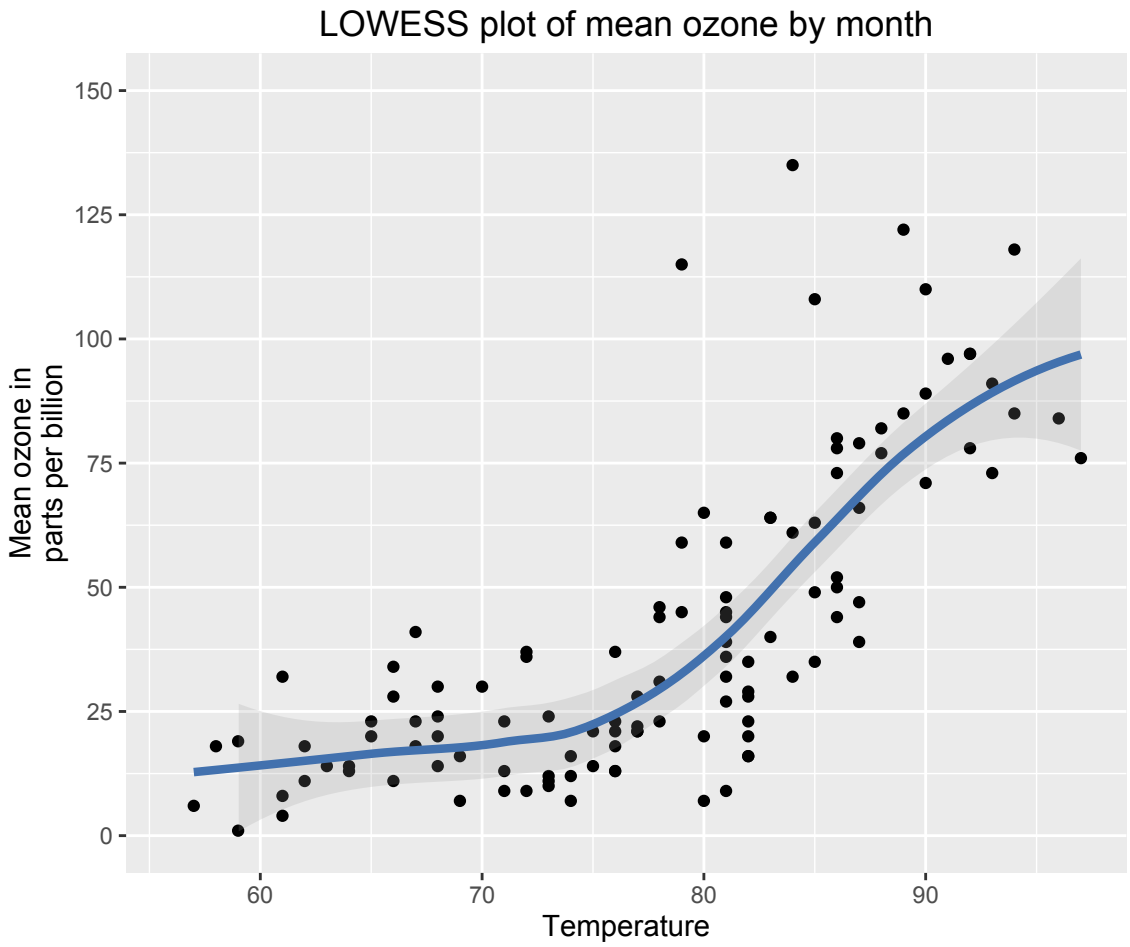
```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method = "loess", colour = fill, size = 1.5) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



Changing the appearance of the confidence interval

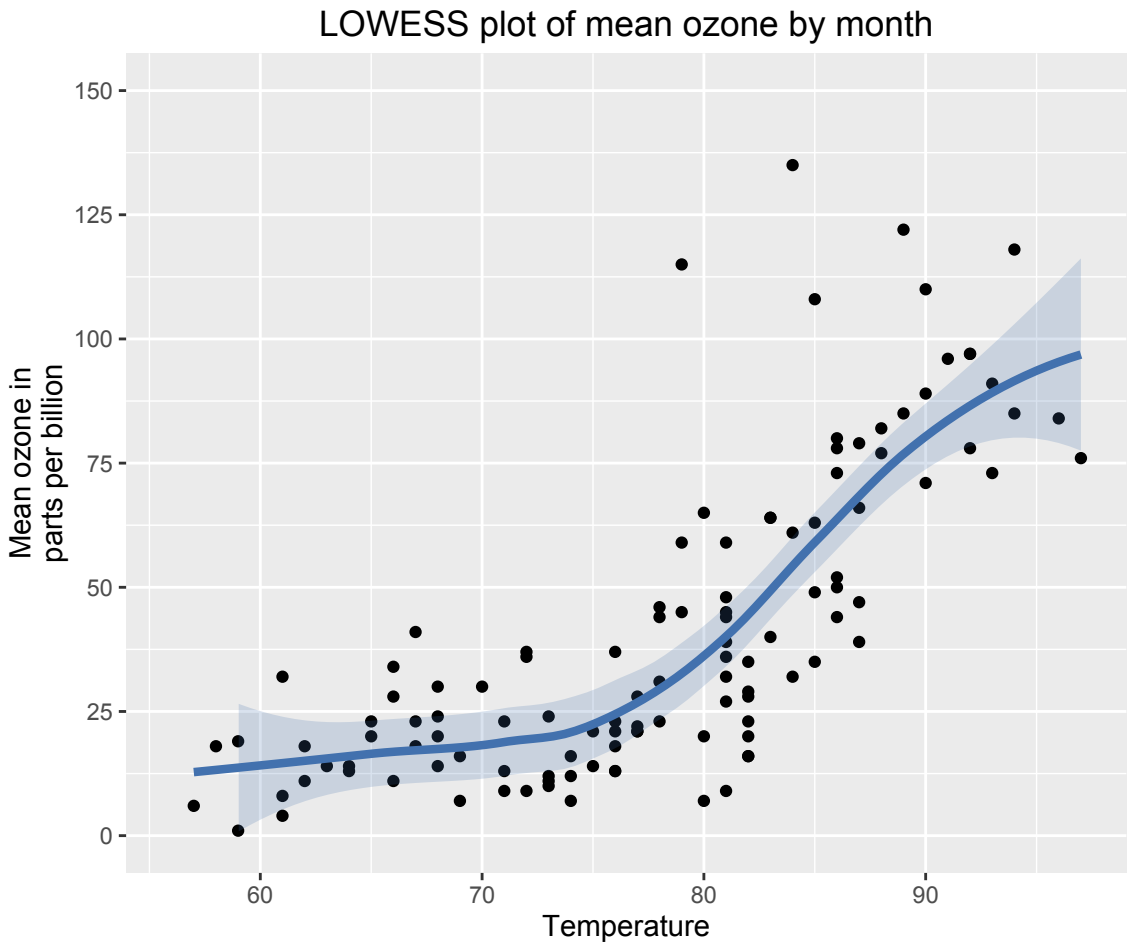
We can also alter how the confidence interval around the LOWESS curve looks. We can change the transparency using the argument `alpha` in `geom_smooth()`. This ranges from 0 to 1. Here we will increase the transparency of the confidence interval.

```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method = "loess", colour = fill, size = 1.5, alpha = 0.2) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



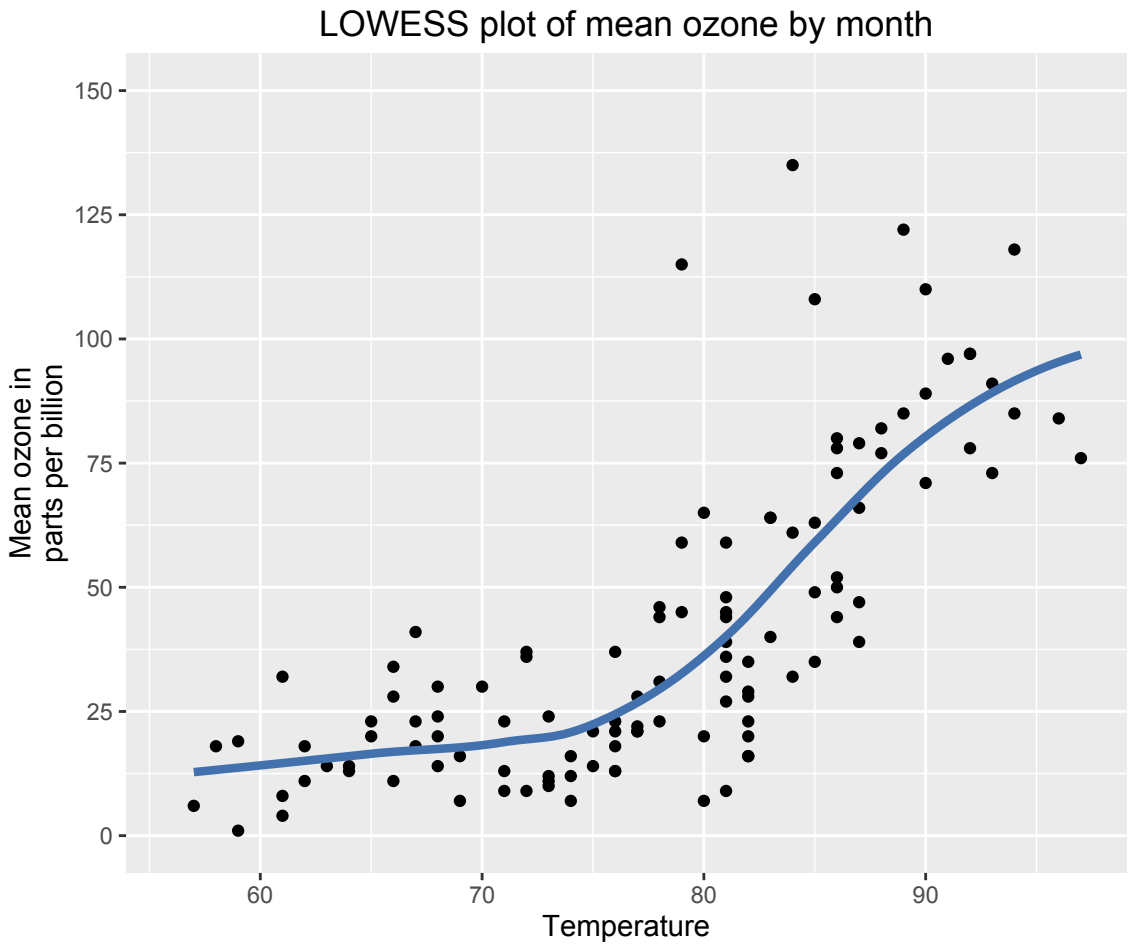
We can also change the colour of the confidence interval from the default grey using the argument `fill`, also within `geom_smooth()`. Let's change it to the same blue as our LOWESS curve.

```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method = "loess", colour = fill, size = 1.5,
              alpha = 0.2, fill = fill) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



Finally, you can also turn off the confidence altogether by adding the argument `se = FALSE` to `geom_smooth()`.

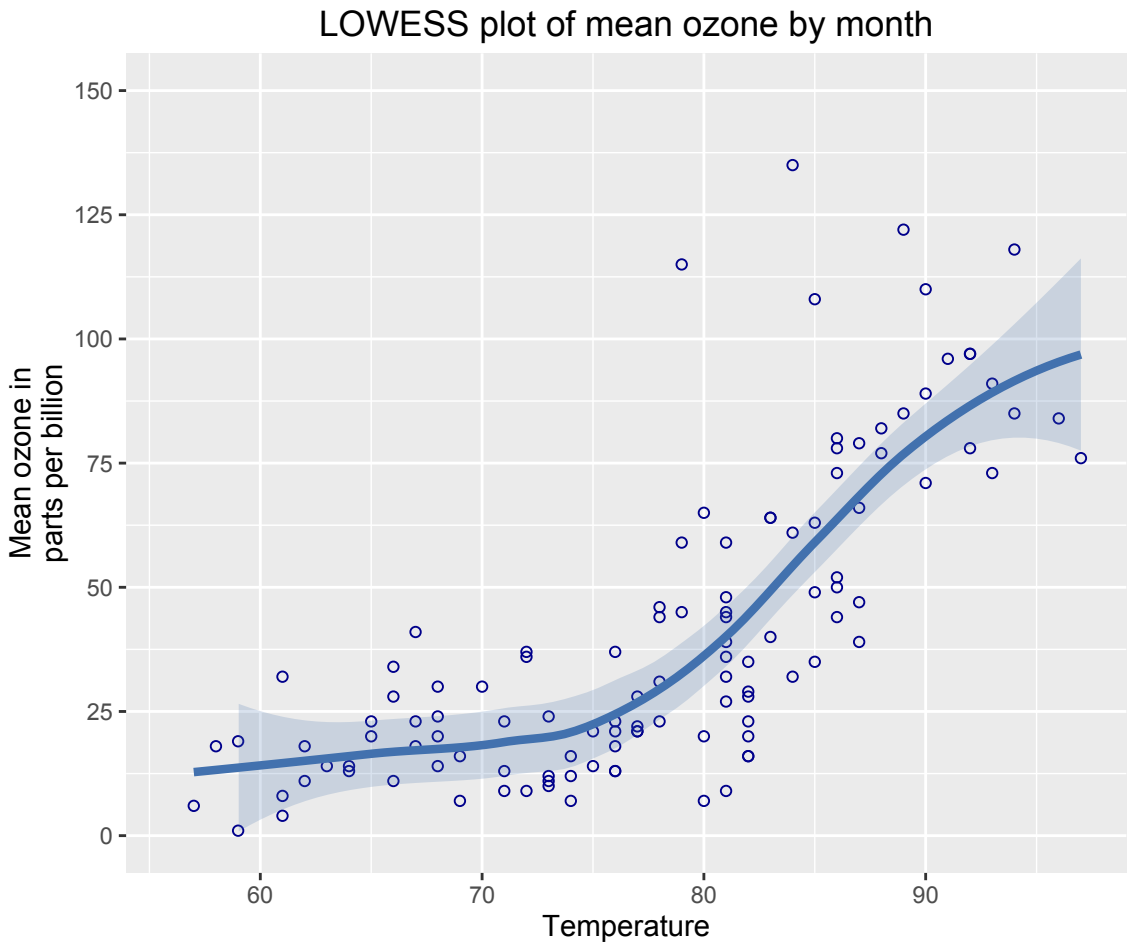
```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point() +
  geom_smooth(method = "loess", colour = fill, size = 1.5, se = FALSE) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



Changing the appearance of the scatterplot

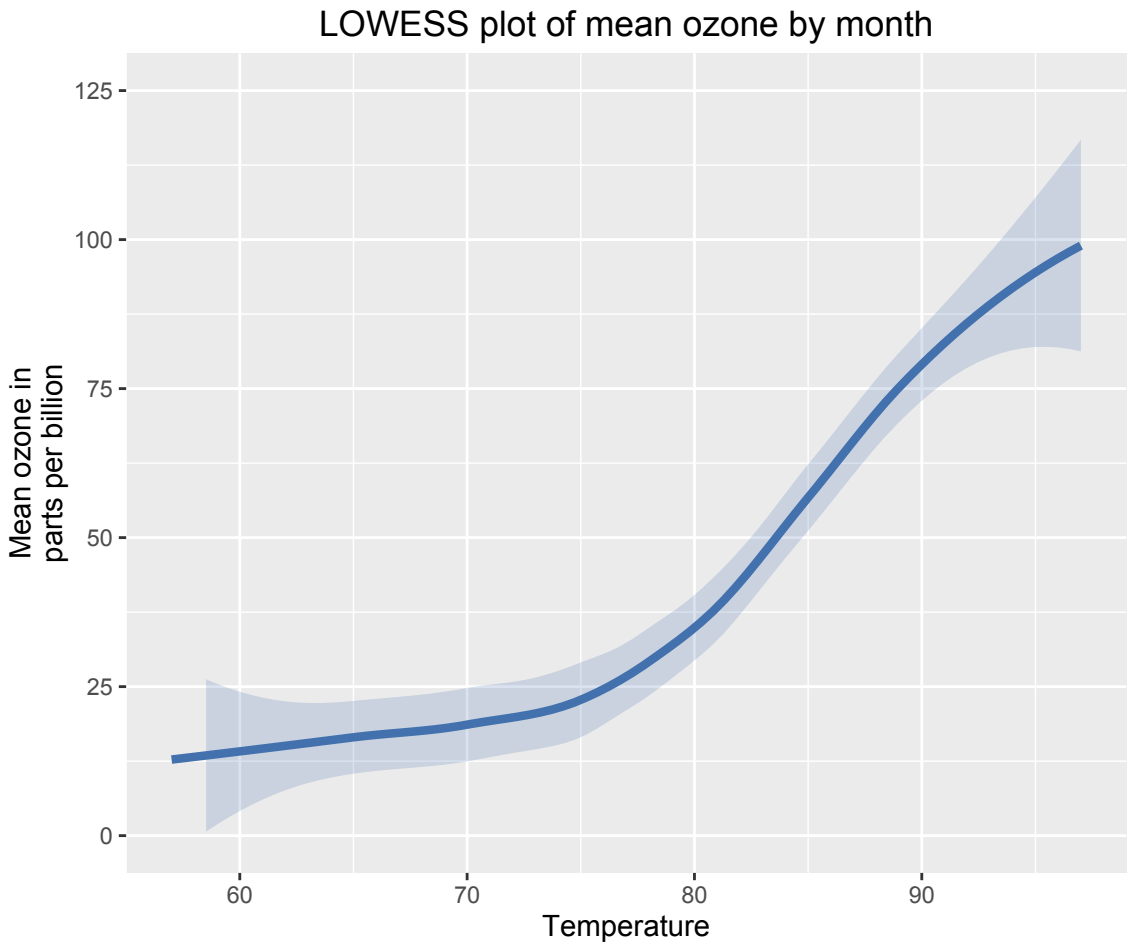
Of course, the LOWESS curve is not the only part of this plot. We can also customise the appearance of the scatterplot underlying the curve. Let's change the circles to shape 21, which is a circle that allows different colours for the outline and fill, and change the colour of the outline to dark blue. We can do this by adding the `shape` and `colour` arguments to `geom_point()` respectively.

```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point(shape = 21, colour = "darkblue") +
  geom_smooth(method = "loess", colour = fill, size = 1.5,
              alpha = 0.2, fill = fill) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



You can also get rid of the scatterplot points altogether by removing the `geom_point()` option. You can see we have also changed the range of the y-axis in `scale_y_continuous()` so the graph sits closer to the top of the LOWESS curve.

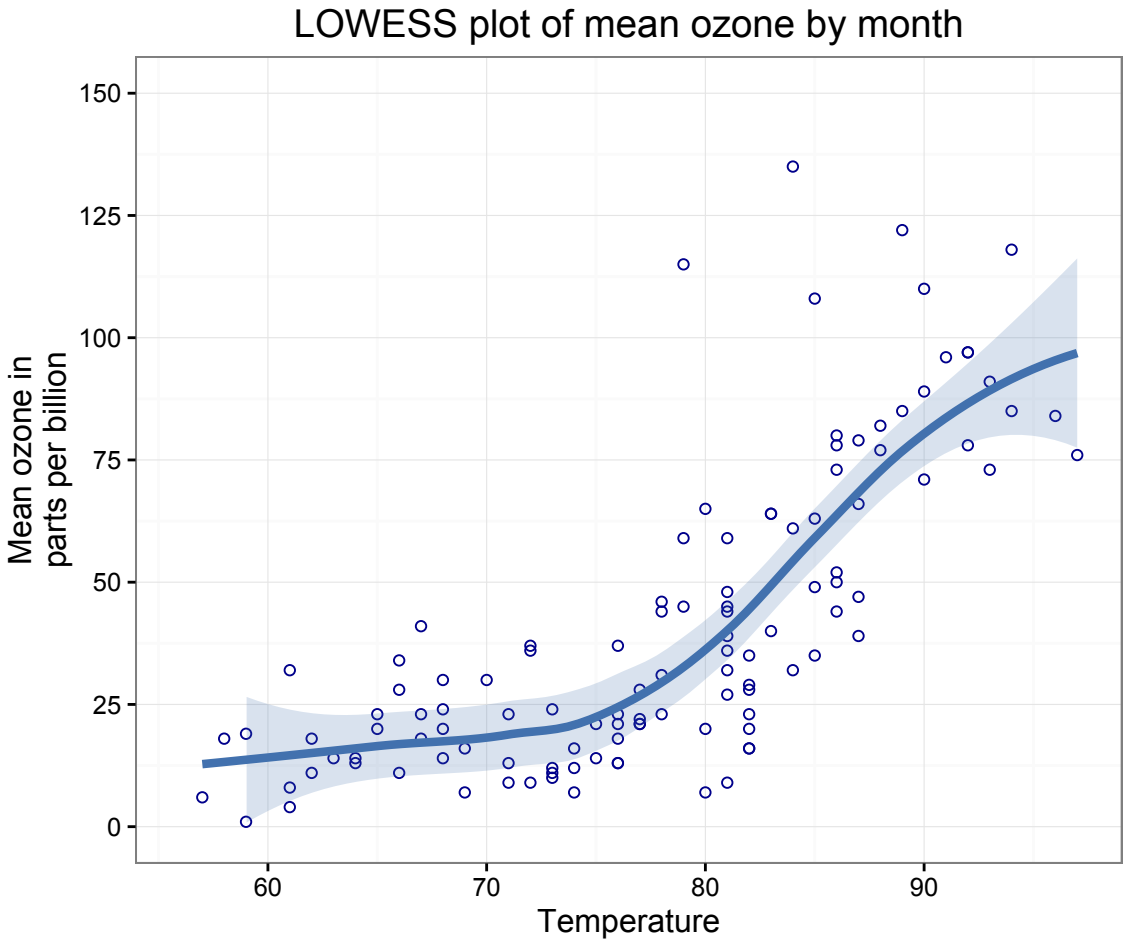
```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_smooth(method = "loess", colour = fill, size = 1.5,
              alpha = 0.2, fill = fill) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 125, 25), limits=c(0, 125)) +
  ggtitle("LOWESS plot of mean ozone by month")
p12
```



Using the white theme

As explained in the previous posts, we can also change the overall look of the plot using themes. We'll start using a simple theme customisation by adding `theme_bw()`. As you can see, we can further tweak the graph using the `theme` option, which we've used so far to change the legend.

```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point(shape = 21, colour = "darkblue") +
  geom_smooth(method = "loess", colour = fill, size = 1.5,
              alpha = 0.2, fill = fill) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month") +
  theme_bw()
p12
```

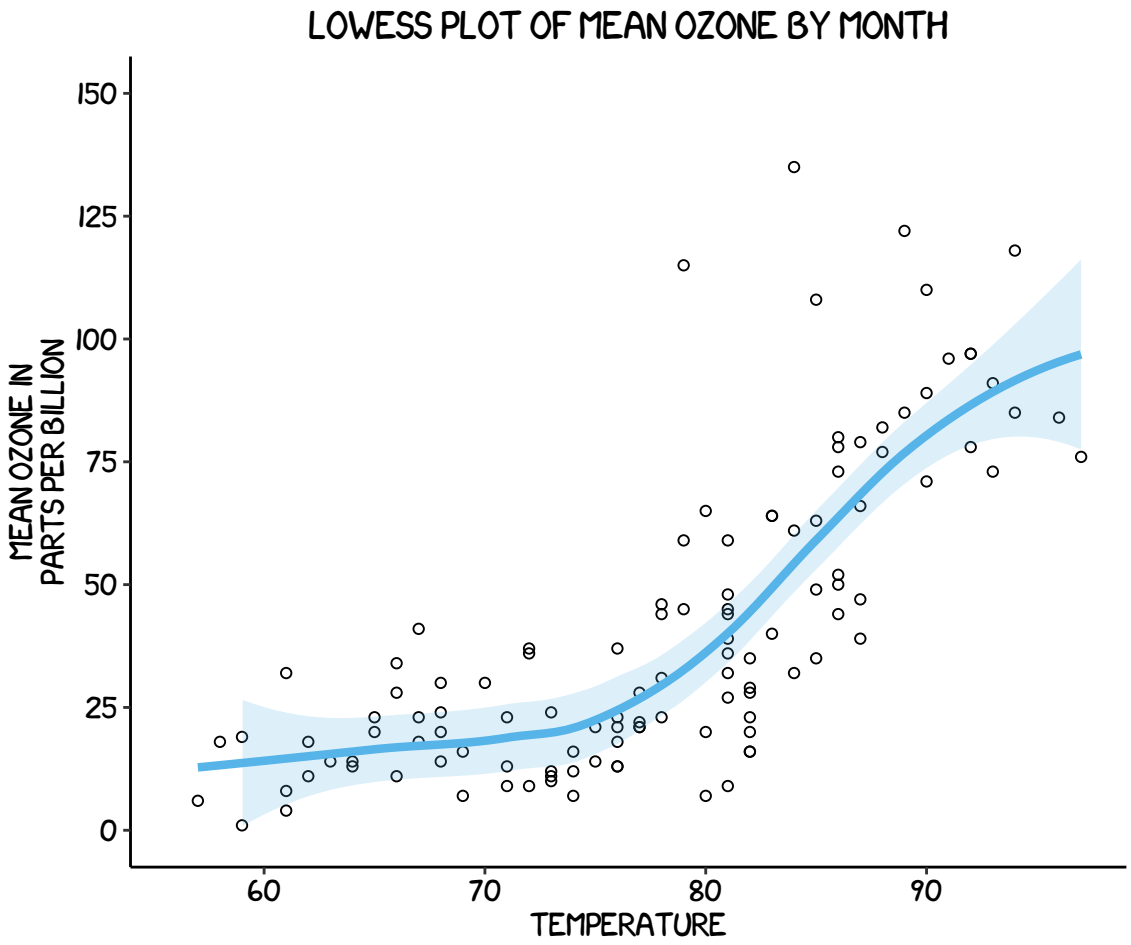


Creating an XKCD style chart

Of course, you may want to create your own themes as well. `ggplot2` allows for a very high degree of customisation, including allowing you to use imported fonts. Below is an example of a theme Mauricio was able to create which mimics the visual style of XKCD. In order to create this chart, you first need to import the XKCD font, and load it into R using the `extrafont` package.

```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point(shape = 21, colour = "black") +
  geom_smooth(method = "loess", colour = "#56B4E9", size = 1.5,
              alpha = 0.2, fill = "#56B4E9") +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month") +
  theme(axis.line.x = element_line(size=.5, colour = "black"),
    axis.line.y = element_line(size=.5, colour = "black"),
    axis.text.x=element_text(colour="black", size = 10),
    axis.text.y=element_text(colour="black", size = 10),
    legend.position="bottom",
    legend.direction="horizontal",
```

```
    legend.box = "horizontal",
    legend.key = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    plot.title=element_text(family="xkcd-Regular"),
    text=element_text(family="xkcd-Regular"))
p12
```

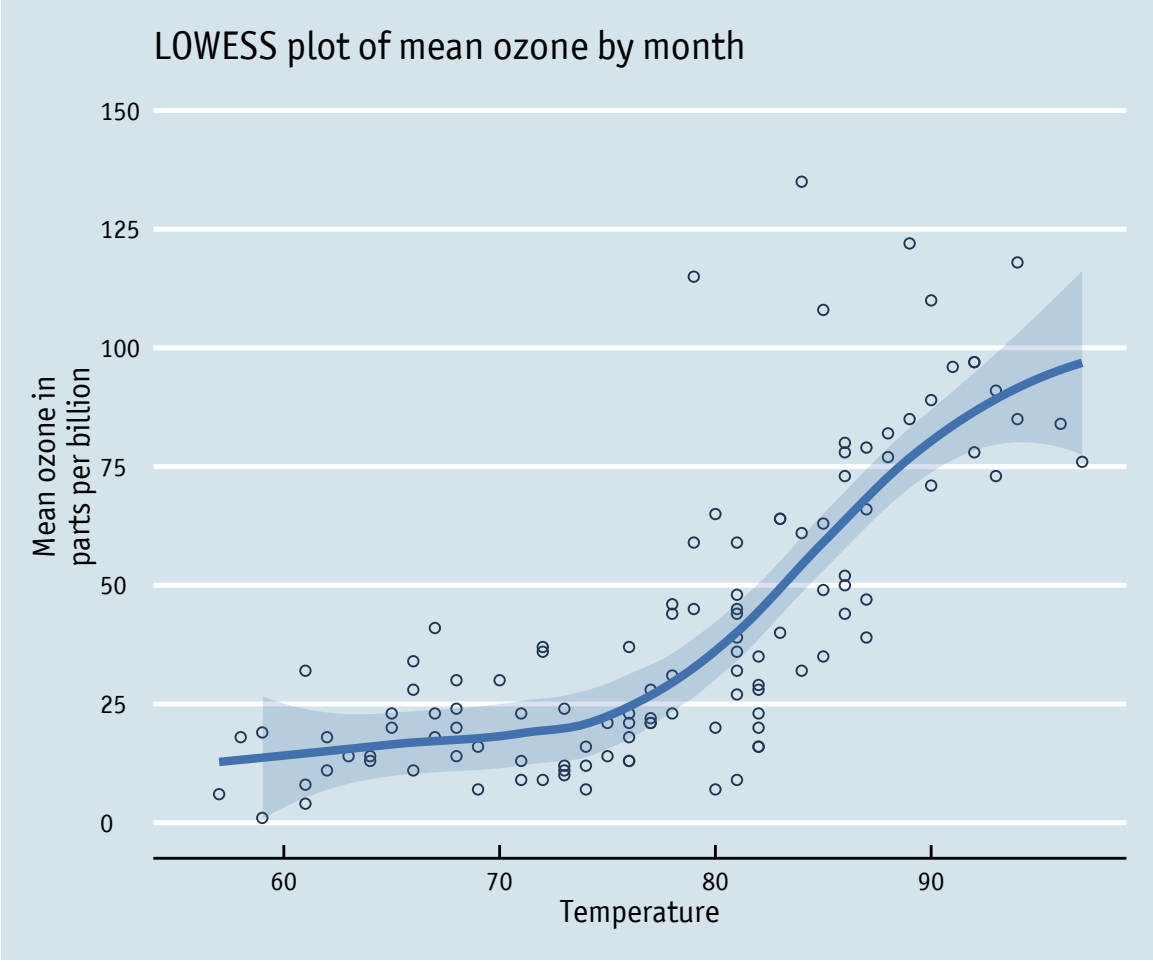


Using 'The Economist' theme

There are a wider range of pre-built themes available as part of the `ggthemes` package (more information on these here). Below we've applied `theme_economist()`, which approximates graphs in the Economist magazine.

```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point(shape = 21, colour = "#1F3552") +
  geom_smooth(method = "loess", colour = "#4271AE", size = 1.5,
              alpha = 0.2, fill = "#4271AE") +
```

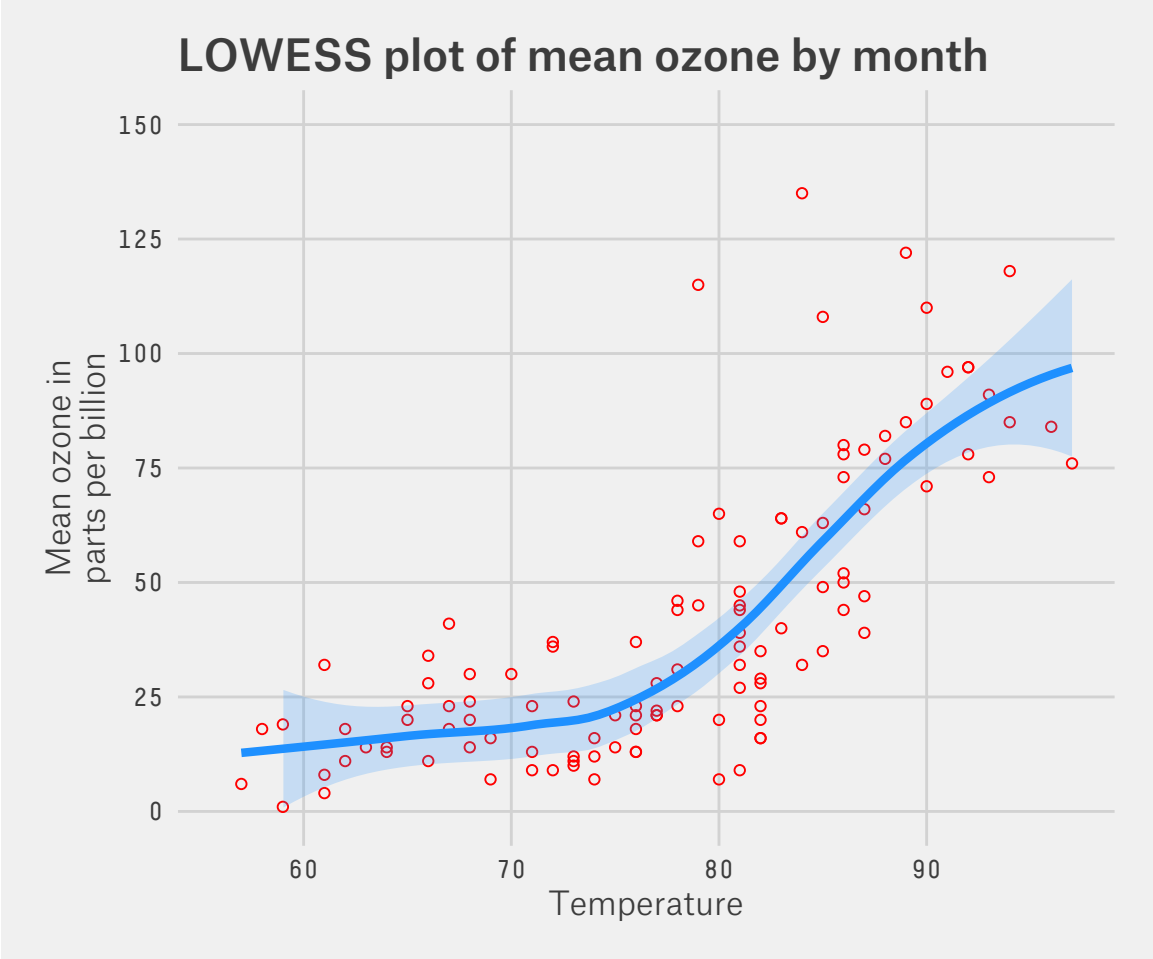
```
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month") +
  theme_economist() + scale_fill_economist() +
  theme(axis.line.x = element_line(size=.5, colour = "black"),
    axis.title = element_text(size = 12),
    legend.position="bottom",
    legend.direction="horizontal",
    legend.box = "horizontal",
    legend.text = element_text(size = 10),
    text = element_text(family = "OfficinaSanITC-Book"),
    plot.title = element_text(family="OfficinaSanITC-Book"))
p12
```



Using 'Five Thirty Eight' theme

Below we've applied `theme_fivethirtyeight()`, which approximates graphs in the nice FiveThirtyEight website. Again, it is also important that the font change is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Atlas Grotesk' and 'Decima Mono Pro' which are commercial font and are available here and here.

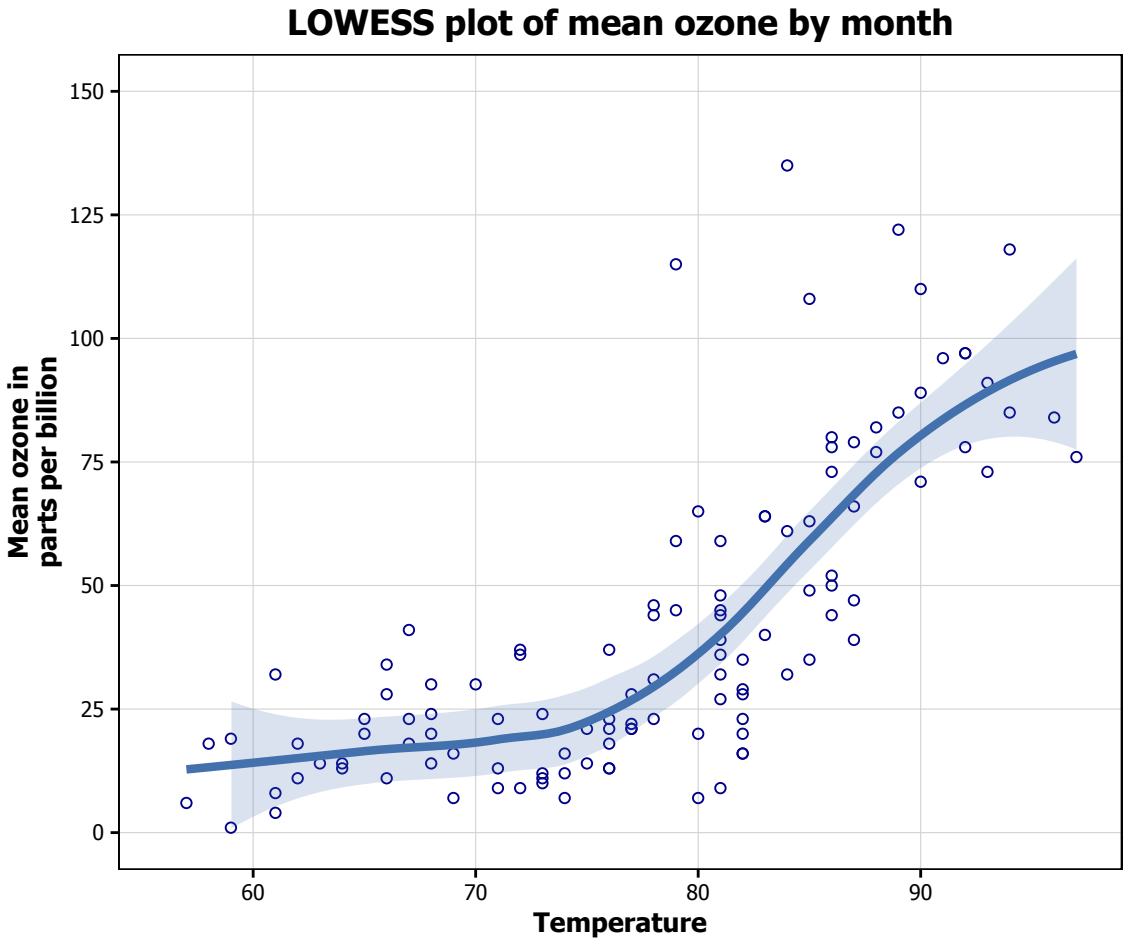
```
p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point(shape = 21, colour = "red") +
  geom_smooth(method = "loess", colour = "dodgerblue", size = 1.5,
              alpha = 0.2, fill = "dodgerblue") +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month") +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family="AtlasGrotesk-Light", size = 12),
    legend.position="bottom",
    legend.direction="horizontal",
    legend.box = "horizontal",
    legend.title=element_text(family="AtlasGrotesk-Light", size = 8),
    legend.text=element_text(family="AtlasGrotesk-Light", size = 8),
    plot.title=element_text(family="AtlasGrotesk-Medium", size = 16),
    text=element_text(family="DecimaMonoPro"))
p12
```



Creating your own theme

As before, you can modify your plots a lot as `ggplot2` allows many customisations. Here we present our original result shown at the top of page.

```
fill <- "#4271AE"

p12 <- ggplot(airquality, aes(x = Temp, y = Ozone)) +
  geom_point(shape = 21, colour = "darkblue") +
  geom_smooth(method = "loess", colour = fill, size = 1.5,
              alpha = 0.2, fill = fill) +
  scale_x_continuous(name = "Temperature") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
                     breaks = seq(0, 150, 25), limits=c(0, 150)) +
  ggtitle("LOWESS plot of mean ozone by month") +
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill=NA, size=.5),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(), panel.background = element_blank(),
    plot.title = element_text(size = 13, family = "Tahoma", face = "bold"),
    text=element_text(family = "Tahoma"),
    axis.title = element_text(face="bold", size = 10),
    axis.text.x = element_text(colour="black", size = 8),
    axis.text.y = element_text(colour="black", size = 8))
p12
```


LOWESS plot of mean ozone by month
Mean ozone in parts per billion
150
125
100
75
50
25
0
60
70
80
90
Temperature