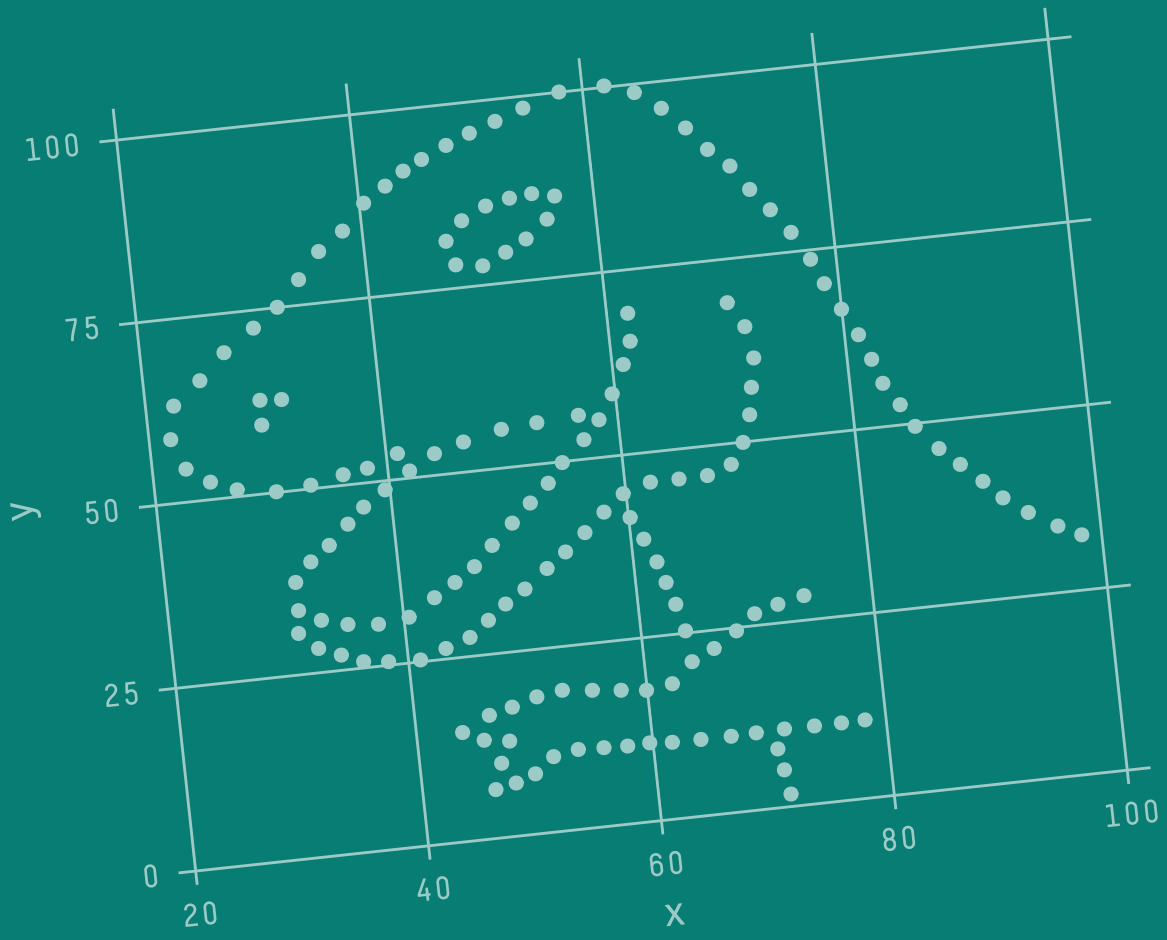
THE HITCHHIKER'S GUIDE TO GGPLOT2

Don't panic and create beautiful plots with R



Jodie Burchell
Mauricio Vargas

Leanpub

THE HITCHHIKER'S GUIDE TO GGPLOT2

JODIE BURCHELL & MAURICIO VARGAS

The Hitchhiker's Guide to Ggplot2

Jodie Burchell & Mauricio Vargas

This book is for sale at https://leanpub.com/hitchhikers_ggplot2

This version was published on 6th September, 2019.

ISBN: 978-956-362-693-3

Contents

What to expect from this book

This is a technical book. The book aims to get straight to the point, and the writing style is similar to a recipe with detailed instructions. It is assumed that you know the basics of R and that you want to learn to create beautiful plots.

Each chapter will explain how to create a different type of plot, and will take you step-by-step from a basic plot to a highly customised graph. The chapters are ordered by degree of graph complexity.

Every chapter is self contained. You can read the whole book or go to a section of your interest, and we are sure that it will be easy to understand the instructions and reproduce our examples without reading the earlier chapters.

We invite you to stay in touch and read the authors' blogs where they publish articles about R and Statistics. Jodie's blog is Standard error and Mauricio's blog is Pachá (Batteries Included).

This book is written using R Markdown and exported to LaTeX. In the present version we are using this R version and packages:

R version 3.6.1 (2019-07-05)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: *LC_CTYPE=en_US.UTF-8*, *LC_NUMERIC=C*, *LC_TIME=en_US.UTF-8*, *LC_COLLATE=en_US.UTF-8*, *LC_MONETARY=en_US.UTF-8*, *LC_MESSAGES=en_US.UTF-8*, *LC_PAPER=en_US.UTF-8*, *LC_NAME=C*, *LC_ADDRESS=C*, *LC_TELEPHONE=C*, *LC_MEASUREMENT=en_US.UTF-8* and *LC_IDENTIFICATION=C*

attached base packages: *grid*, *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *pander(v.0.6.3)*, *scales(v.1.0.0)*, *readr(v.1.3.1)*, *showtext(v.0.7)*, *showtextdb(v.2.0)*, *sysfonts(v.0.8)*, *RColorBrewer(v.1.1-2)*, *HistData(v.0.8-4)*, *dplyr(v.0.8.3)*, *forcats(v.0.4.0)*, *ggfortify(v.0.4.7)*, *ggthemes(v.4.2.0)*, *ggplot2(v.3.2.1)* and *pacman(v.0.5.1)*

loaded via a namespace (and not attached): *Rcpp(v.1.0.2)*, *pillar(v.1.4.2)*, *compiler(v.3.6.1)*, *tools(v.3.6.1)*, *zeallot(v.0.1.0)*, *digest(v.0.6.20)*, *evaluate(v.0.14)*, *tibble(v.2.1.3)*, *gtable(v.0.3.0)*, *pkgconfig(v.2.0.2)*, *rlang(v.0.4.0)*, *yaml(v.2.2.0)*, *xfun(v.0.8)*, *gridExtra(v.2.3)*, *withr(v.2.1.2)*, *stringr(v.1.4.0)*, *knitr(v.1.24)*, *vctrs(v.0.2.0)*, *hms(v.0.5.0)*, *tidyselect(v.0.2.5)*, *glue(v.1.3.1)*, *R6(v.2.4.0)*, *rmarkdown(v.1.14)*, *purrr(v.0.3.2)*, *tidyr(v.0.8.3)*, *magrittr(v.1.5)*, *backports(v.1.1.4)*, *htmltools(v.0.3.6)*, *assertthat(v.0.2.1)*, *colorspace(v.1.4-1)*, *stringi(v.1.4.3)*, *lazyeval(v.0.2.2)*, *munsell(v.0.5.0)* and *crayon(v.1.3.4)*

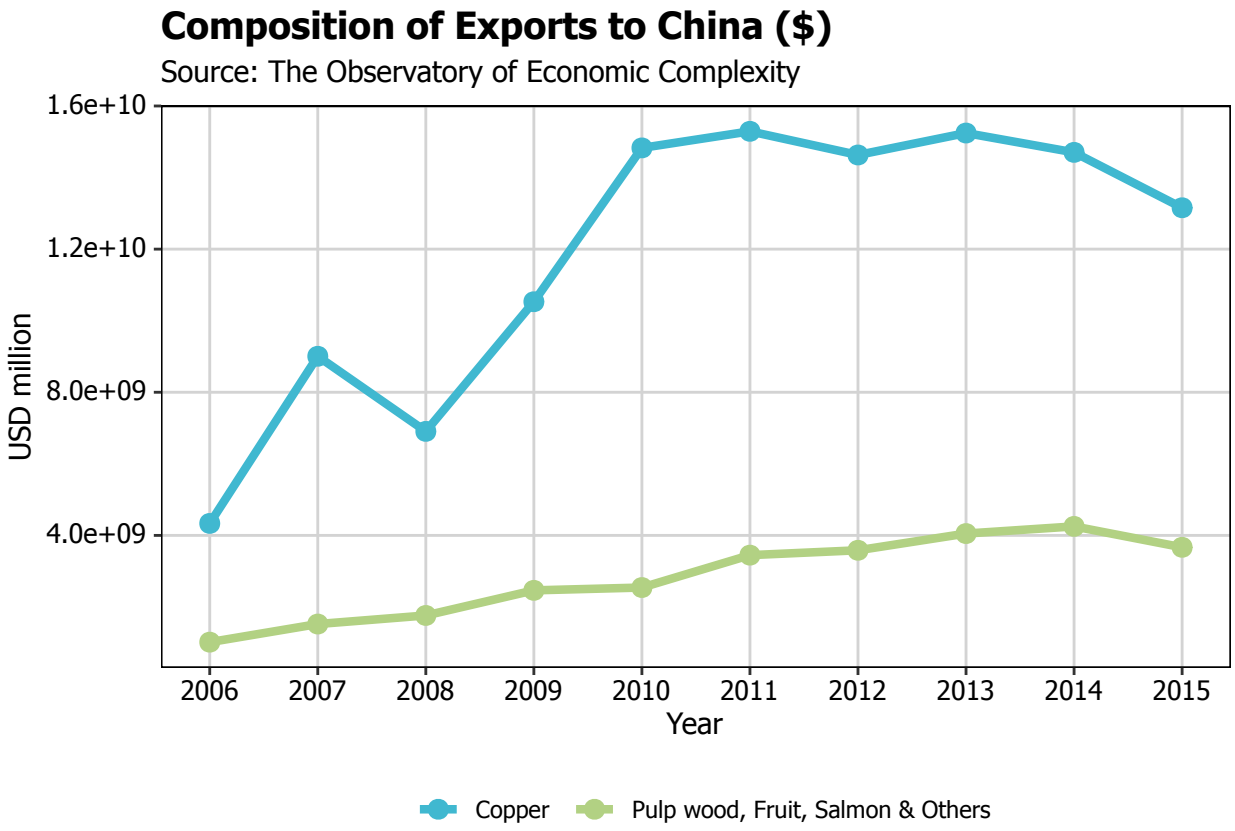
Don’t panic!

CHAPTER 1

Line plots

1.1. Introduction

In this chapter, we will work towards creating the line plot below. We will take you from a basic line plot and explain all the customisations we add to the code step-by-step.



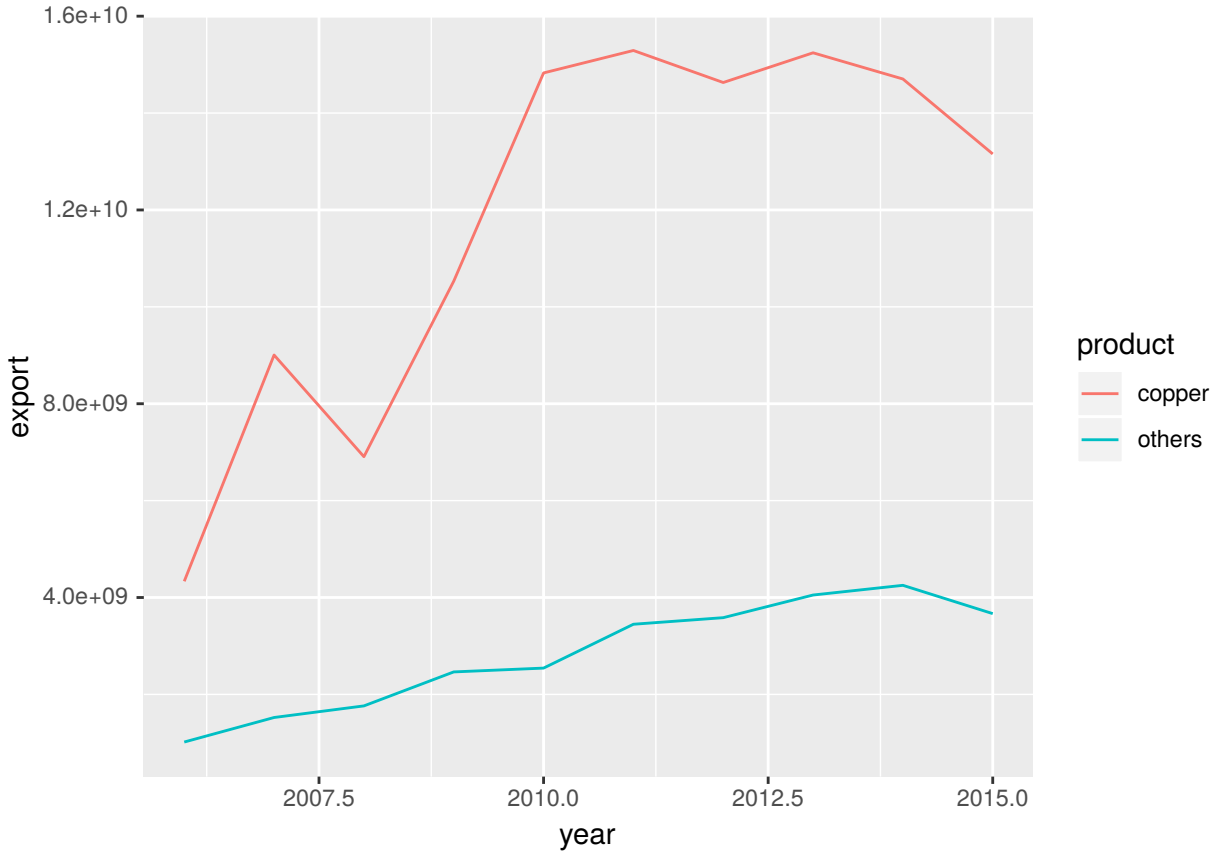
1.2. Basic graph

The first thing to do is load in the data and the libraries, as below:

```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, dplyr, readr)

chilean_exports <- "year,product,export,percentage
  2006,copper,4335009500,81
  2006,others,1016726518,19
  2007,copper,9005361914,86
  2007,others,1523085299,14
  2008,copper,6907056354,80
  2008,others,1762684216,20
  2009,copper,10529811075,81
  2009,others,2464094241,19
  2010,copper,14828284450,85
  2010,others,2543015596,15
  2011,copper,15291679086,82
  2011,others,3447972354,18
  2012,copper,14630686732,80
  2012,others,3583968218,20
  2013,copper,15244038840,79
  2013,others,4051281128,21
  2014,copper,14703374241,78
  2014,others,4251484600,22
  2015,copper,13155922363,78
  2015,others,3667286912,22
"

exports_data <- read_csv(chilean_exports)
```

In order to initialise a plot we tell ggplot that `exports_data` is our data, and specify the variables on each axis. We then instruct ggplot to render this as a line plot by adding the `geom_line` command.

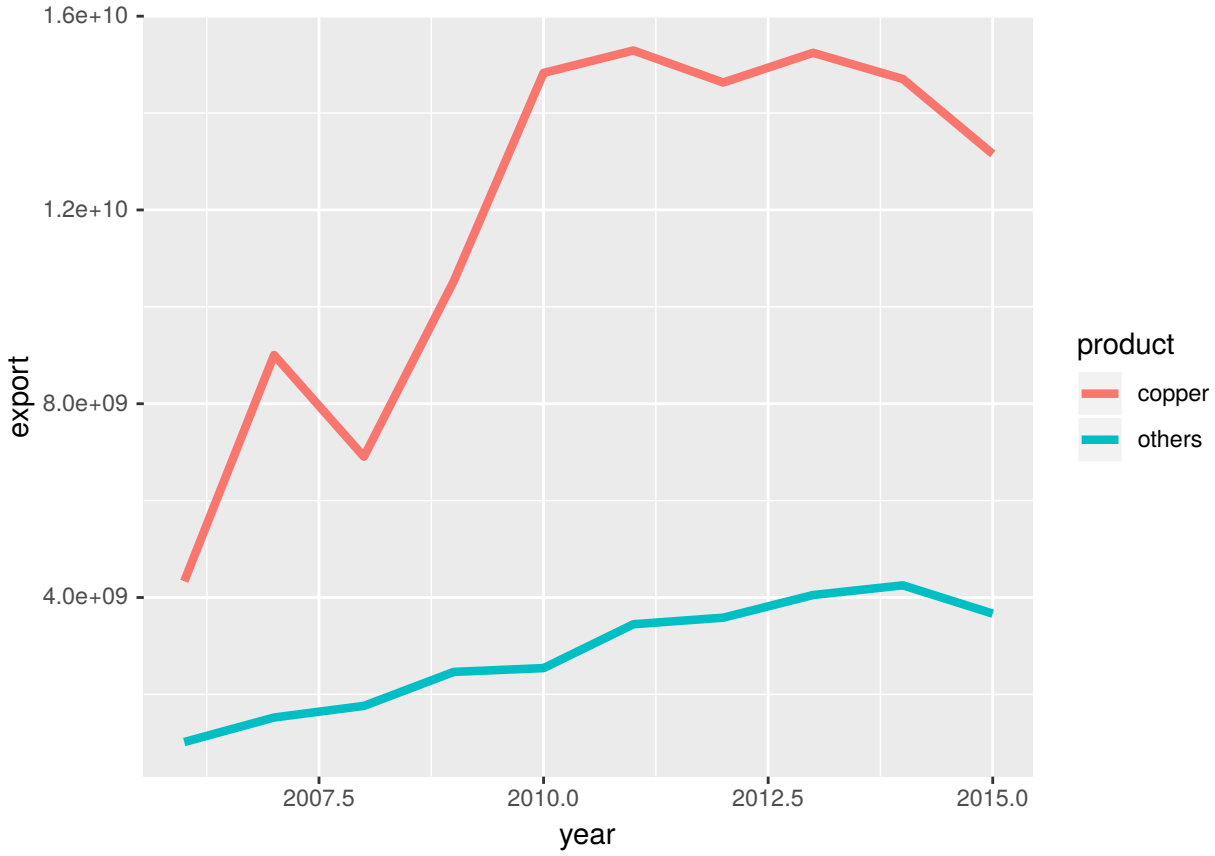
```
p1 <- ggplot(aes(y = export, x = year, colour = product), data = exports_data) +
  geom_line()
p1
```



1.3. Adjusting line width

To change the line width, we add a `size` argument to `geom_line`.

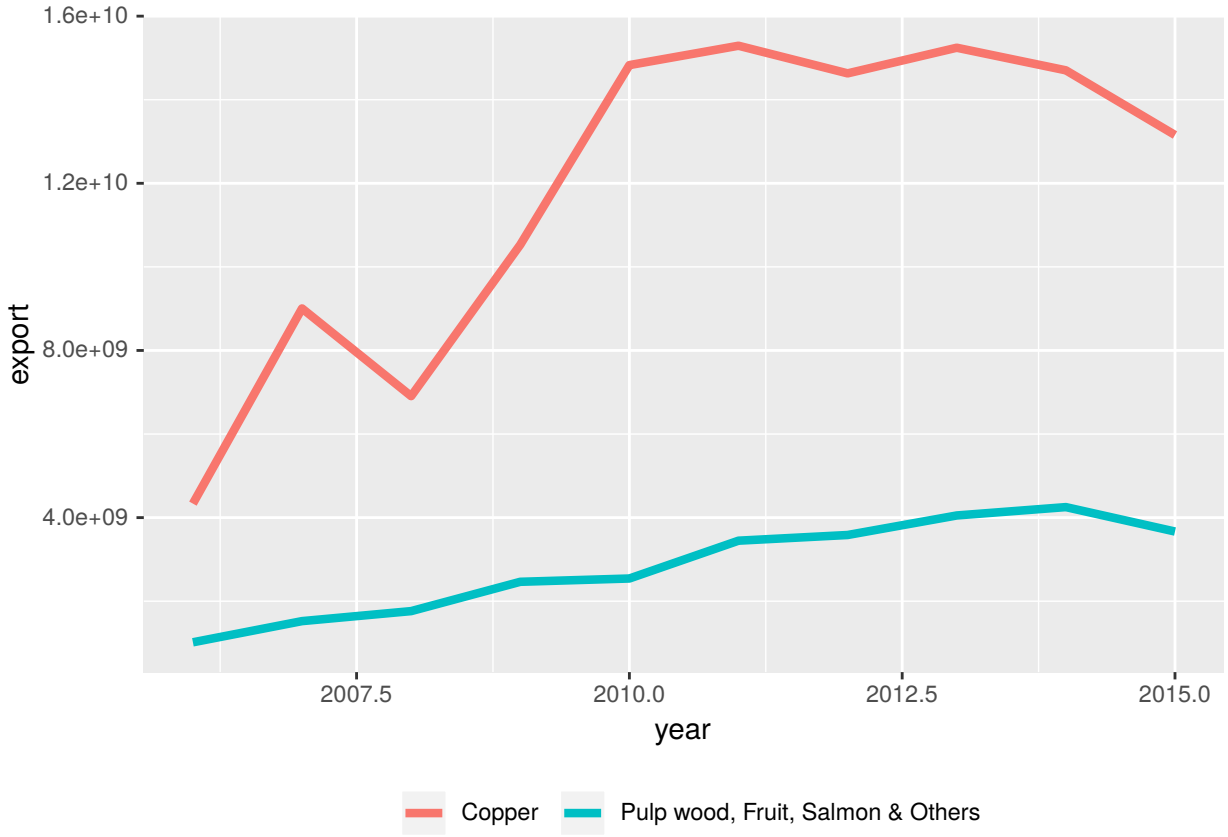
```
p1 <- ggplot(aes(y = export, x = year, colour = product), data = exports_data) +
  geom_line(size = 1.5)
p1
```



1.4. Changing variables display

To change the variables displayed name, we need to re-factor our data labels in `exports_data` tibble. Then we move the legend to the bottom using the `theme` command.

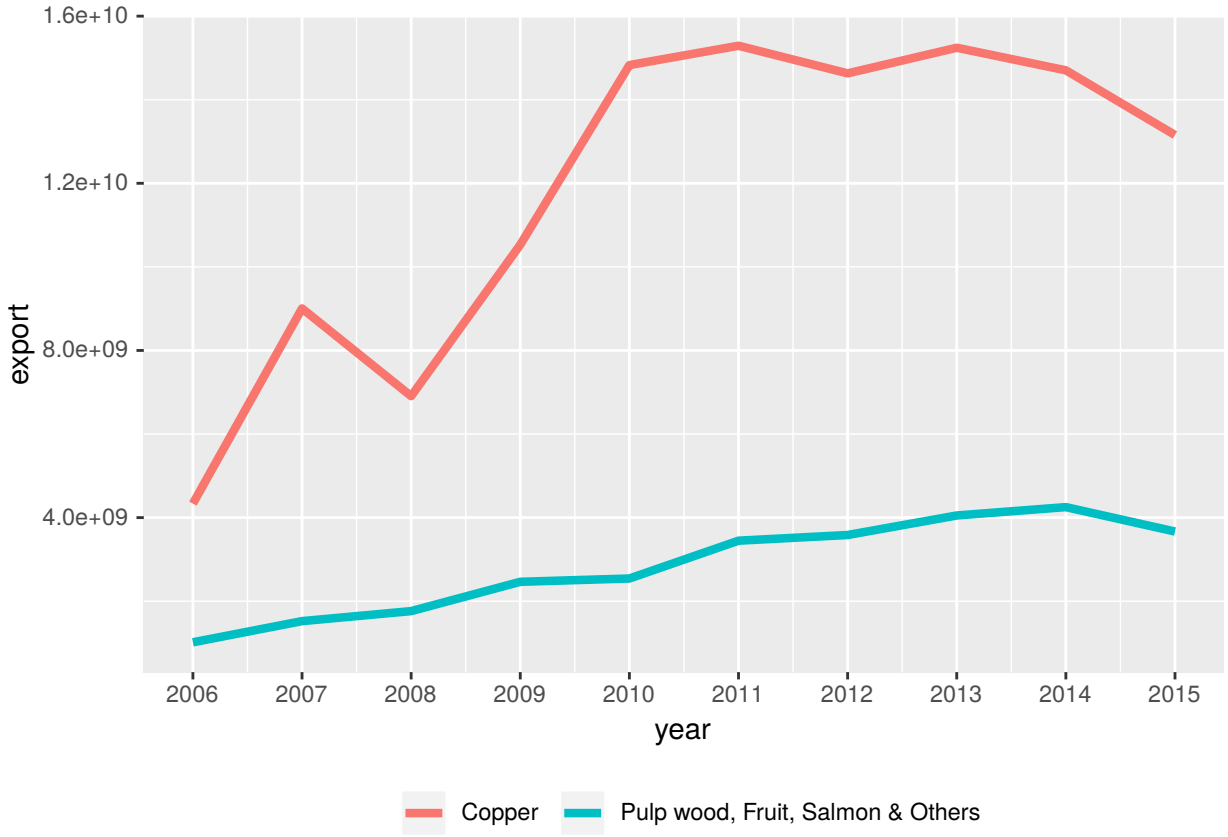
```
exports_data <- exports_data %>%
  mutate(product = factor(product, levels = c("copper","others"),
                          labels = c("Copper ","Pulp wood, Fruit, Salmon & Others")))

p1 <- ggplot(aes(y = export, x = year, colour = product),
             data = exports_data) +
  geom_line(size = 1.5) +
  theme(legend.position = "bottom", legend.direction = "horizontal",
        legend.title = element_blank())
p1
```



1.5. Adjusting x-axis scale

To change the axis tick marks, we use the `scale_x_continuous` and/or `scale_y_continuous` commands.

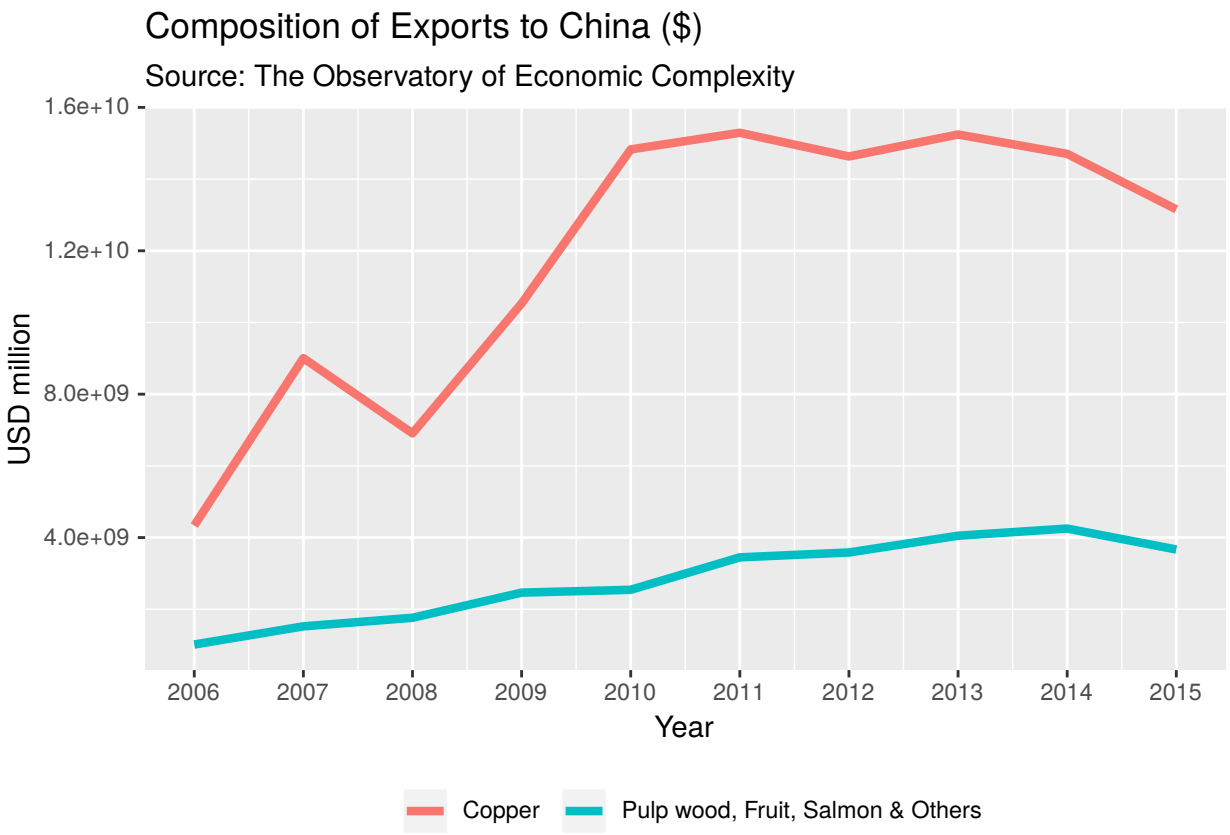
```
p1 <- p1 + scale_x_continuous(breaks = seq(2006,2015,1))
p1
```



1.6. Adjusting axis labels & adding title

To add a title, we include the option `labs` and include the name of the graph as a string argument, and to change the axis names we use the `labs` command too.

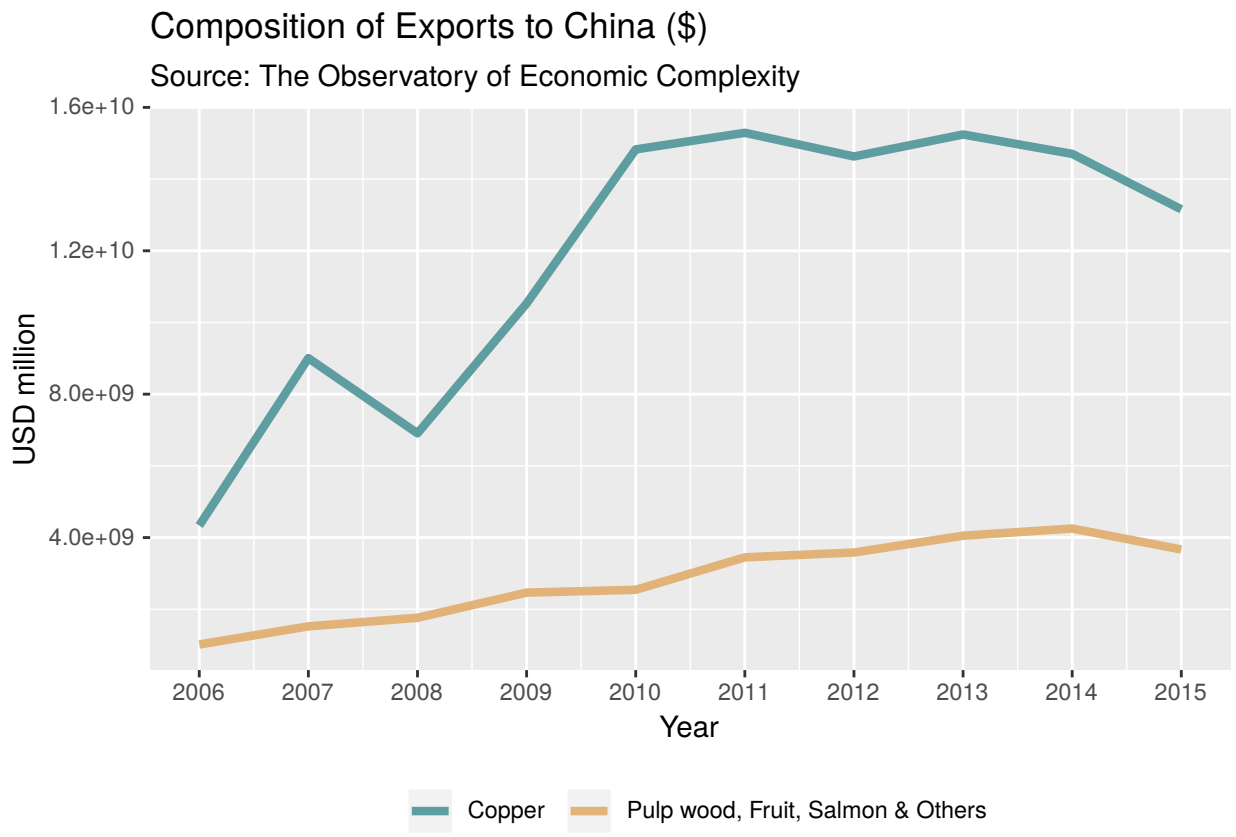
```
p1 <- p1 +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million")
p1
```



1.7. Adjusting color palette

To change the colours, we use the `scale_colour_manual` command.

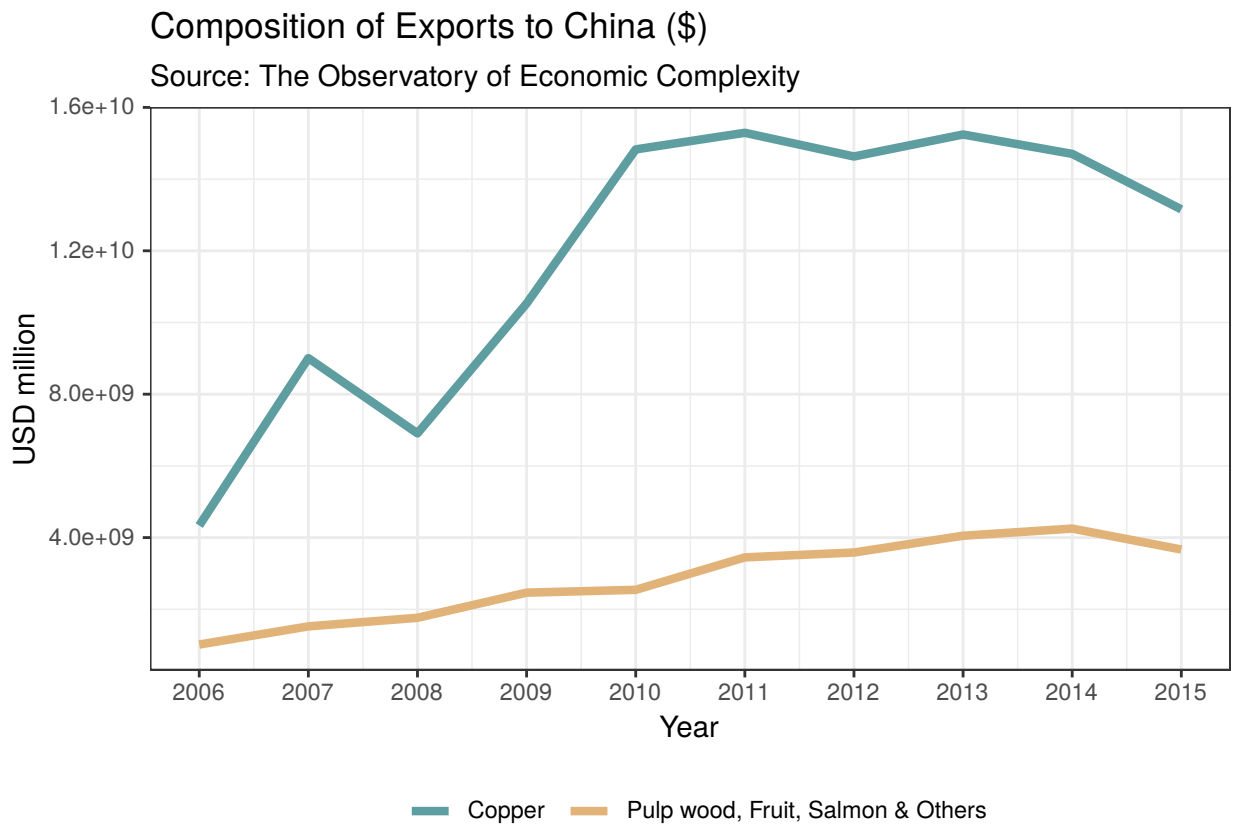
```
colour <- c("#5F9EA0", "#E1B378")
p1 <- p1 + scale_colour_manual(values = colour)
p1
```



1.8. Using the white theme

We'll start using a simple theme customisation made adding `theme_bw()` after `ggplot()`. That theme argument can be modified to use different themes.

```
p1 <- ggplot(aes(y = export, x = year, colour = product),
             data = exports_data) +
  geom_line(size = 1.5) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_colour_manual(values = colour) +
  theme_bw() +
  theme(legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank())
p1
```



1.9. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

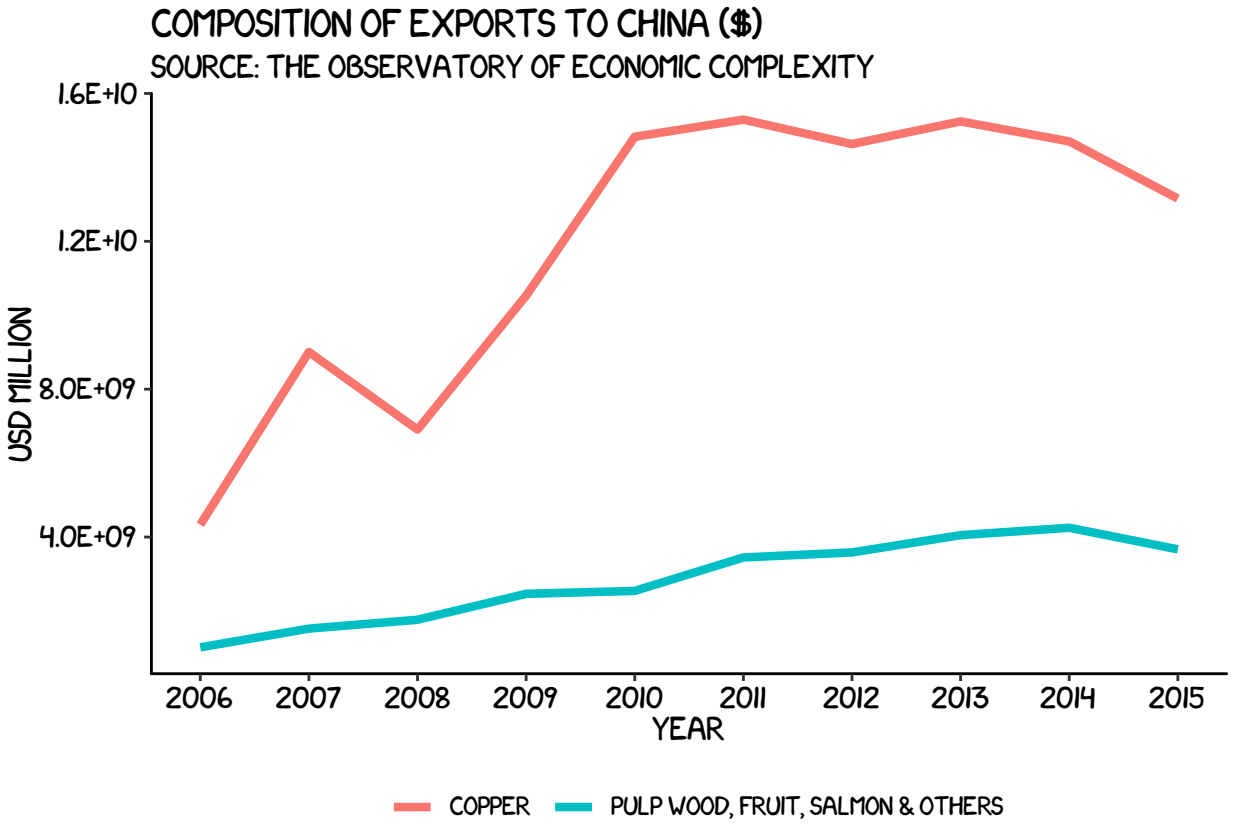
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

1.10. 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.

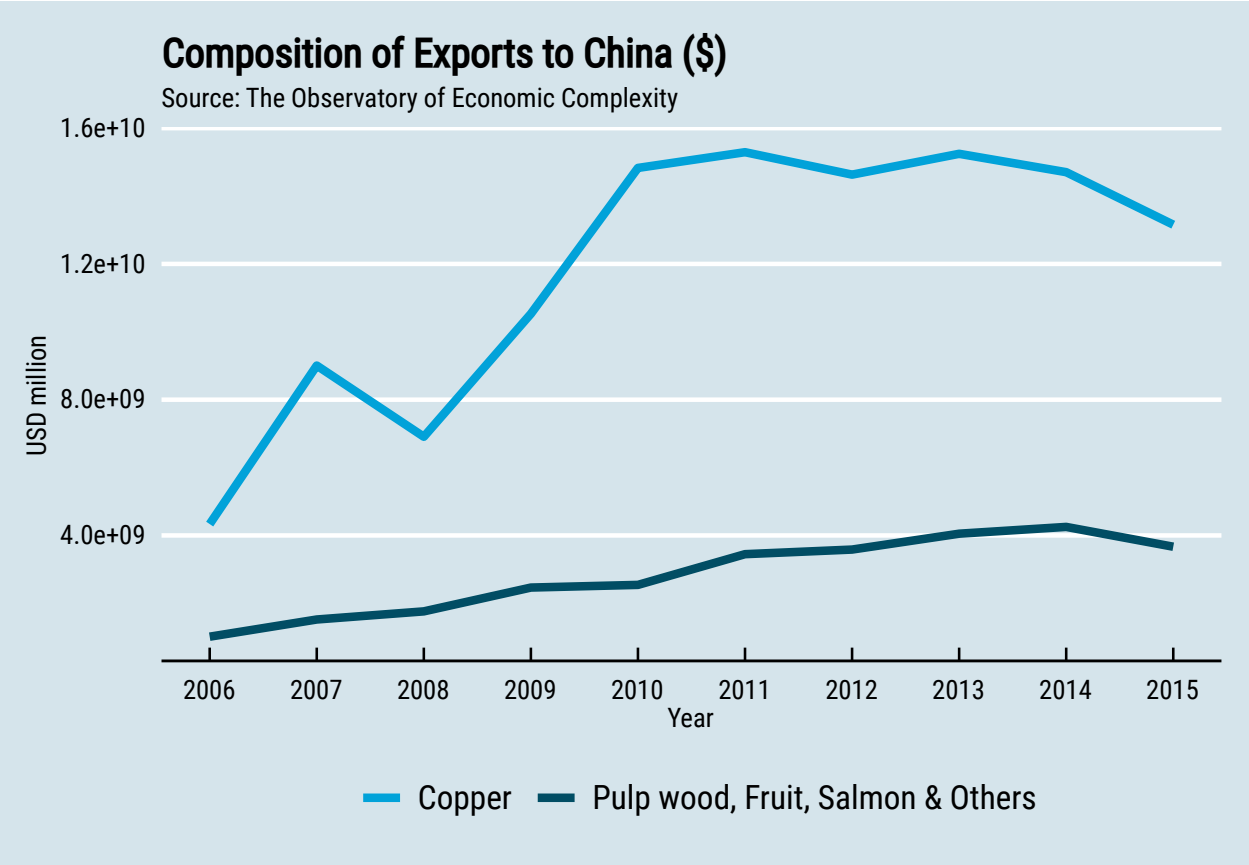
```
fill <- c("#40b8d0","#b2d183")

p1 <- ggplot(aes(y = export, x = year, colour = product),
             data = exports_data) +
  geom_line(size = 1.5) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_fill_manual(values = fill) +
  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.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = 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"),
    text = element_text(family = "XKCD"))
p1
```



1.11. 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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

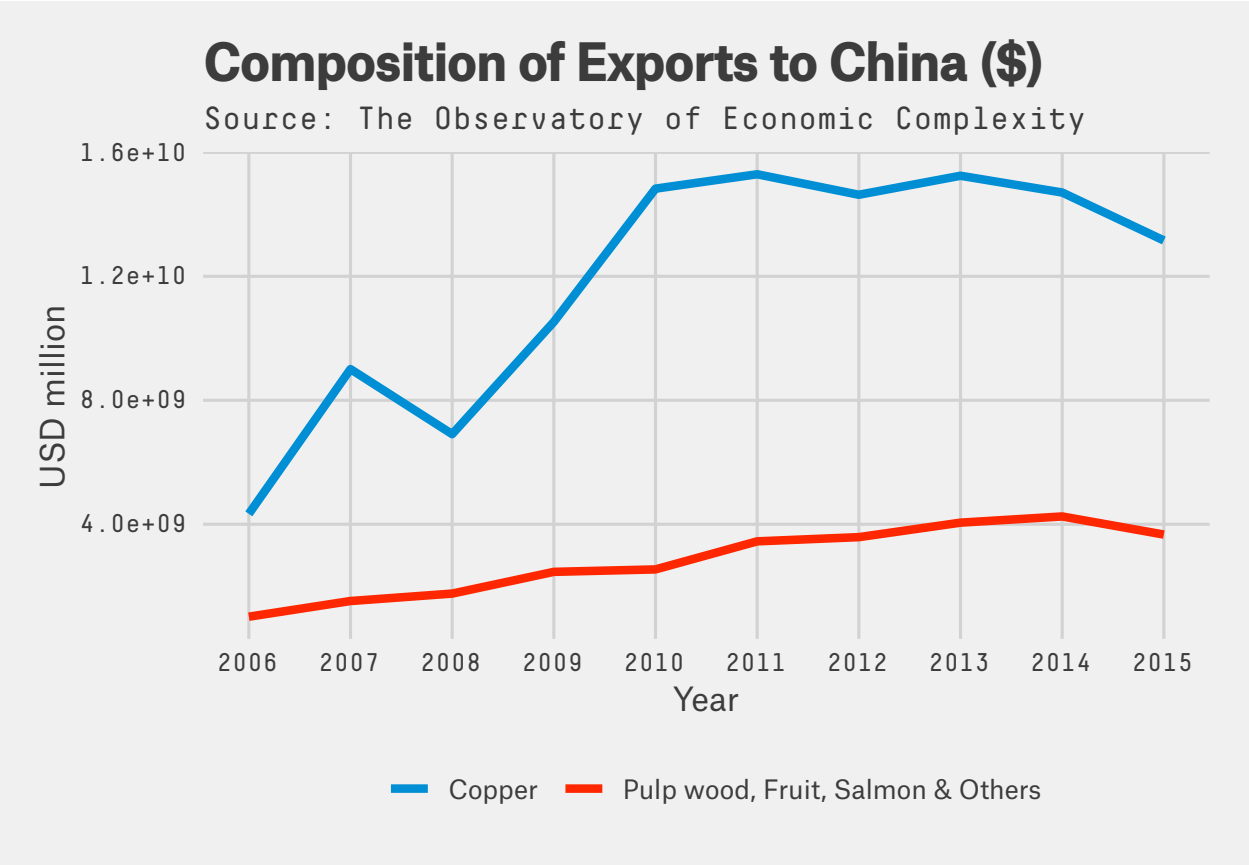
```
p1 <- ggplot(aes(y = export, x = year, colour = product),
             data = exports_data) +
  geom_line(size = 1.5) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  theme_economist() + scale_colour_economist() +
  theme(axis.line.x = element_line(size = .5, colour = "black"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank(),
    plot.title = element_text(family = "Roboto Condensed"),
    text = element_text(family = "Roboto Condensed"))
p1
```



1.12. 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'.

```
p1 <- ggplot(aes(y = export, x = year, colour = product),
             data = exports_data) +
  geom_line(size = 1.5) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  theme_fivethirtyeight() + scale_colour_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    legend.text = element_text(family = "Atlas Grotesk Regular"),
    text = element_text(family = "Decima Mono Pro"))
p1
```

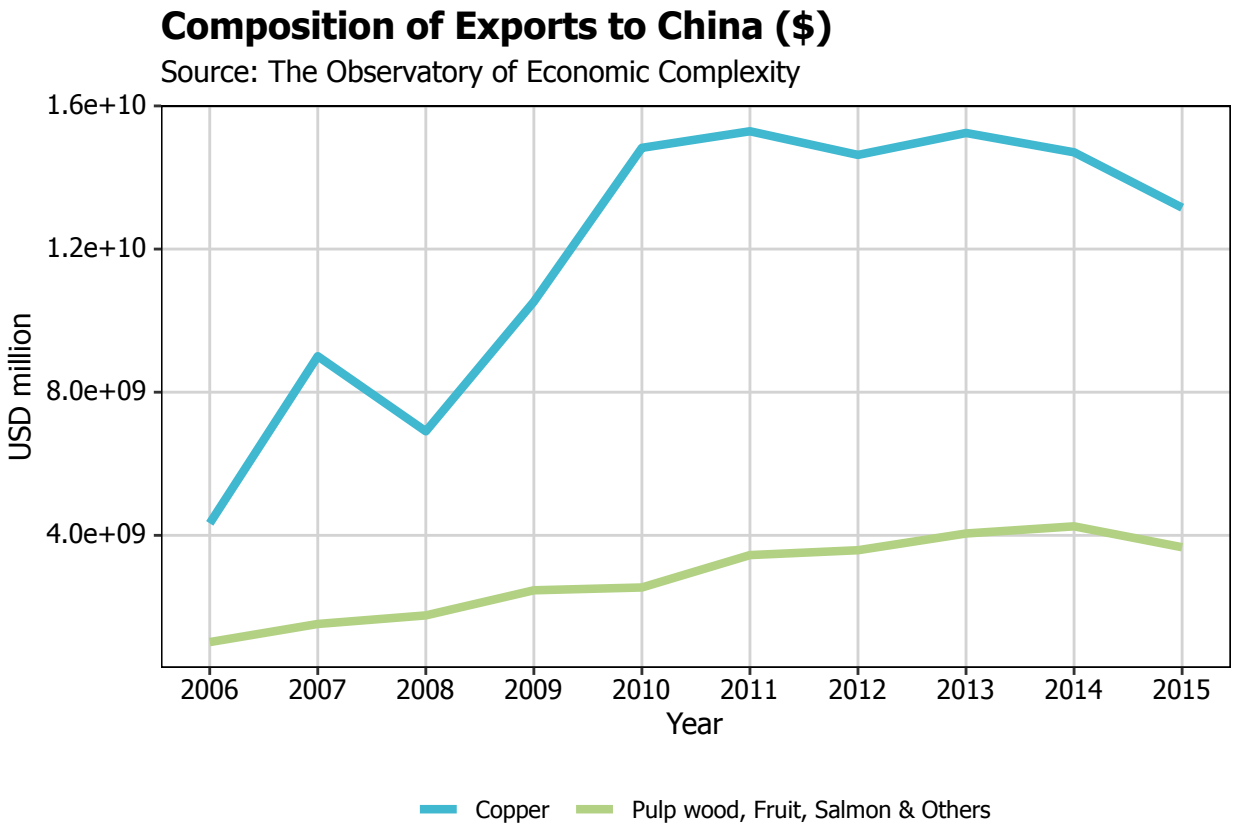


1.13. Creating your own theme

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.

```
colour <- c("#40b8d0", "#b2d183")

p1 <- ggplot(aes(y = export, x = year, colour = product),
             data = exports_data) +
  geom_line(size = 1.5) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_colour_manual(values = colour) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 10),
    axis.text.y = element_text(colour = "black", size = 10),
    legend.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
```

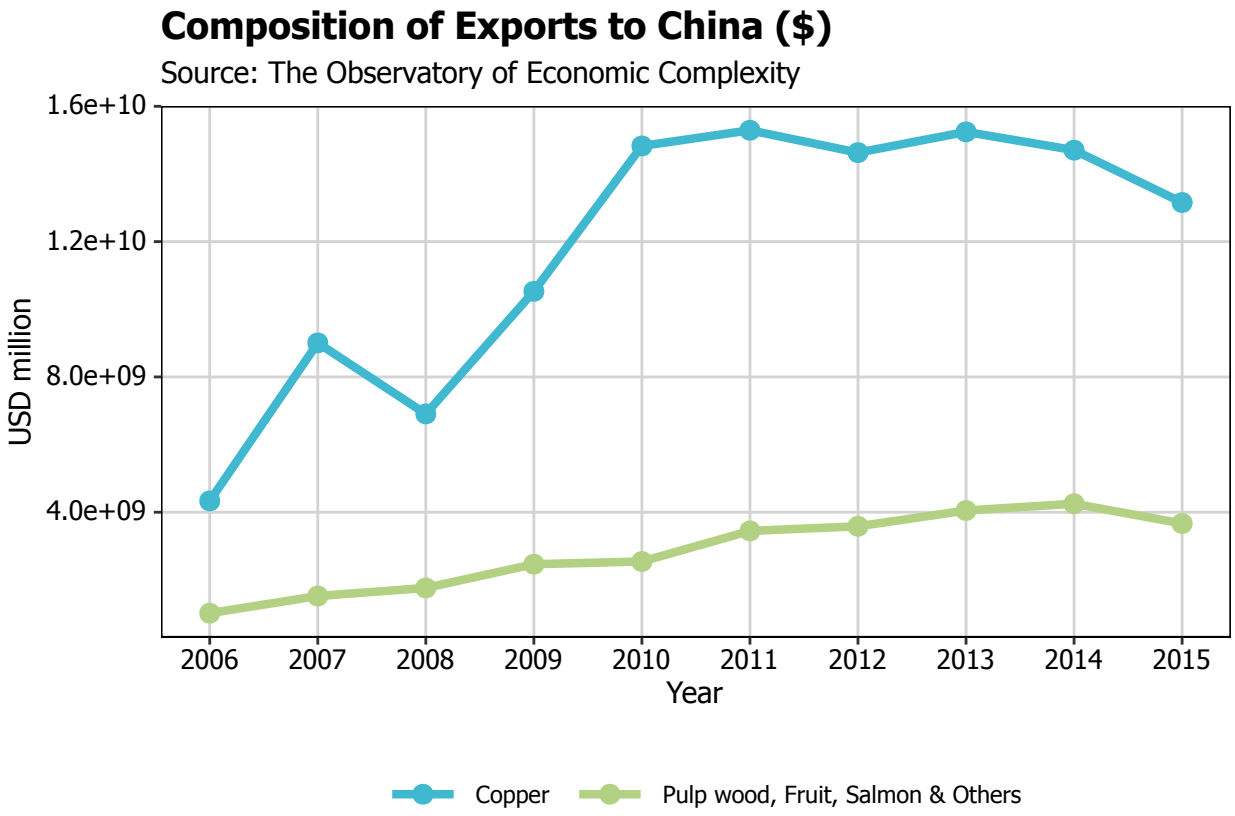
```
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p1
```



Finally, to add points to create a marked line we use `geom_point`.

```
colour <- c("#40b8d0", "#b2d183")

p1 <- ggplot(aes(y = export, x = year, colour = product),
             data = exports_data) +
  geom_line(size = 1.5) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_colour_manual(values = colour) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 10),
    axis.text.y = element_text(colour = "black", size = 10),
    legend.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
```

```
    legend.title = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"),
    legend.key.size = unit(2, 'lines')) +
  geom_point(size = 3)
p1
```

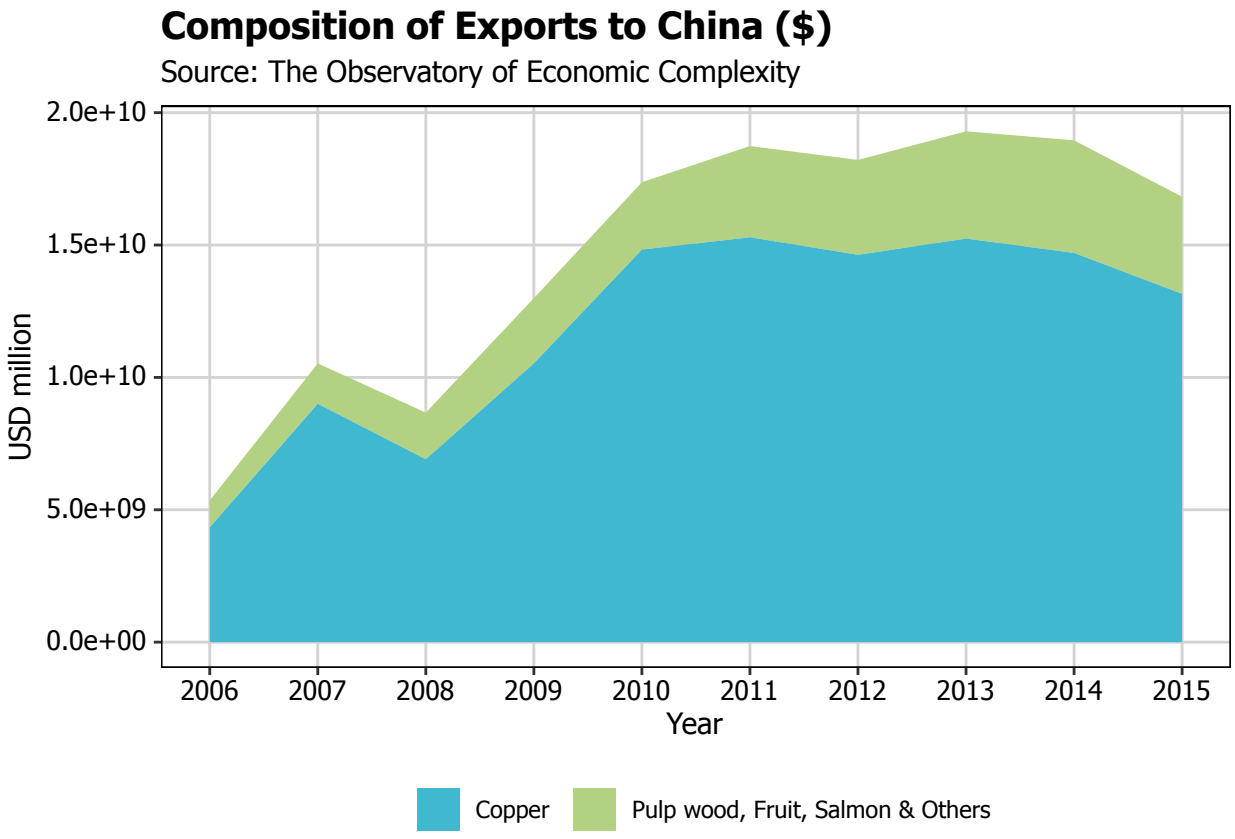


CHAPTER 2

Area plots

2.1. Introduction

In this chapter, we will work towards creating the area plot below. We will take you from a basic area plot and explain all the customisations we add to the code step-by-step.



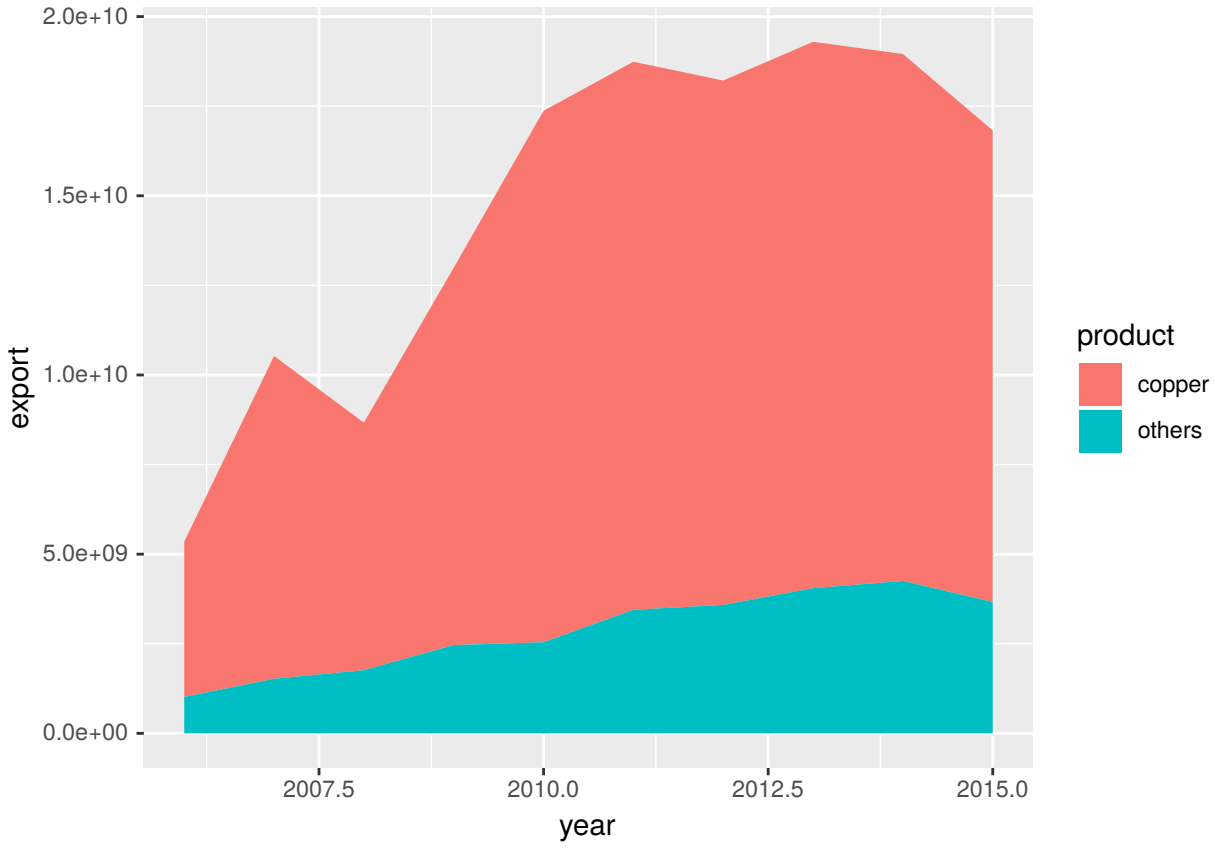
2.2. Basic graph

The first thing to do is load in the data and the libraries, as below:

```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, dplyr, readr, forcats)

chilean_exports <- "year,product,export,percentage
  2006,copper,4335009500,81
  2006,others,1016726518,19
  2007,copper,9005361914,86
  2007,others,1523085299,14
  2008,copper,6907056354,80
  2008,others,1762684216,20
  2009,copper,10529811075,81
  2009,others,2464094241,19
  2010,copper,14828284450,85
  2010,others,2543015596,15
  2011,copper,15291679086,82
  2011,others,3447972354,18
  2012,copper,14630686732,80
  2012,others,3583968218,20
  2013,copper,15244038840,79
  2013,others,4051281128,21
  2014,copper,14703374241,78
  2014,others,4251484600,22
  2015,copper,13155922363,78
  2015,others,3667286912,22
"

exports_data <- read_csv(chilean_exports)
```

In order to initialise a plot we tell ggplot that `exports_data` is our data, and specify the variables on each axis. We then instruct ggplot to render this as an area plot by adding the `geom_area` command.

```
p2 <- ggplot(aes(y = export, x = year, fill = product), data = exports_data) +
  geom_area()
p2
```



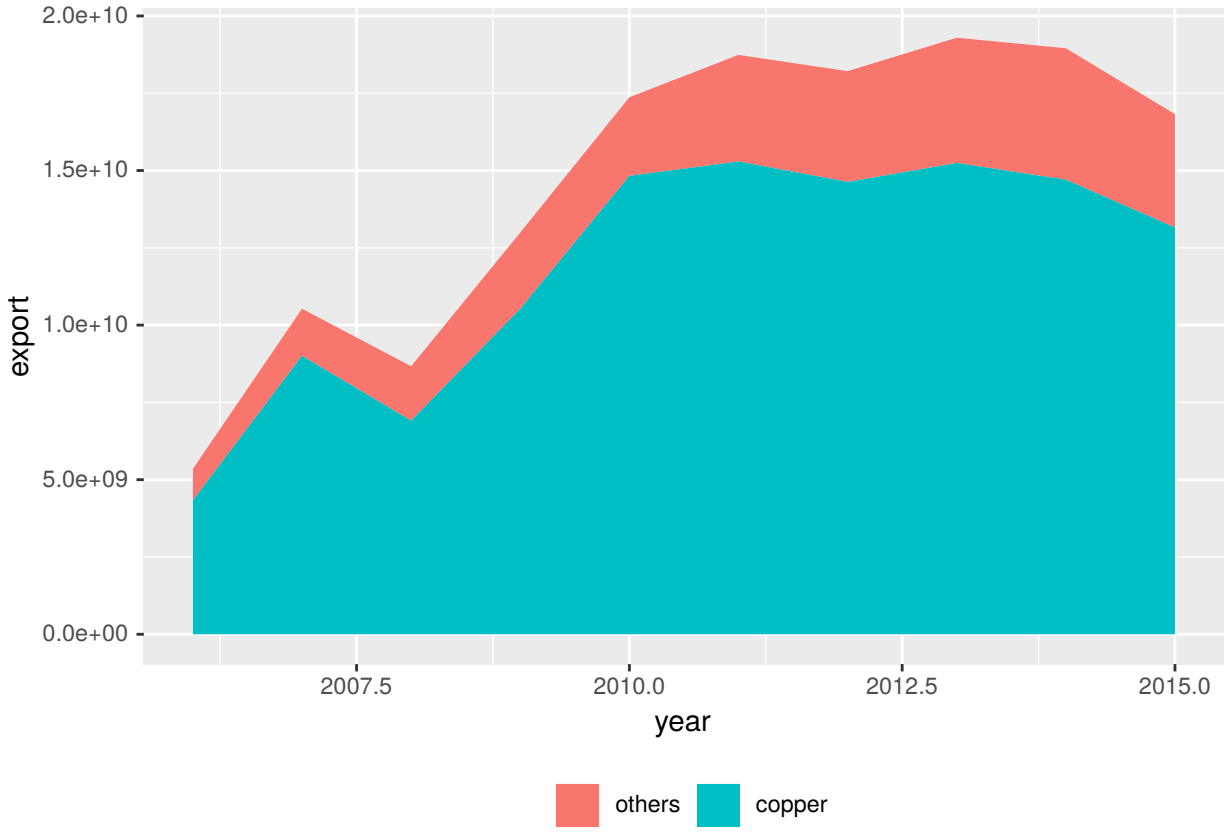
From now and ongoing we will stack in the opposite order. Changing the fill option to `fill = fct_rev(product)` allows us to do that.

```
p2 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_area()
p2
```

2.3. Adjusting legend position

To adjust the position of the legend from the default spot of right of the graph, we add the `theme` option and specify the `legend.position = "bottom"` argument. We can also change the title to blank using the `legend.title = element_blank()` argument and change the legend shape using the `legend.direction = "horizontal"` argument.

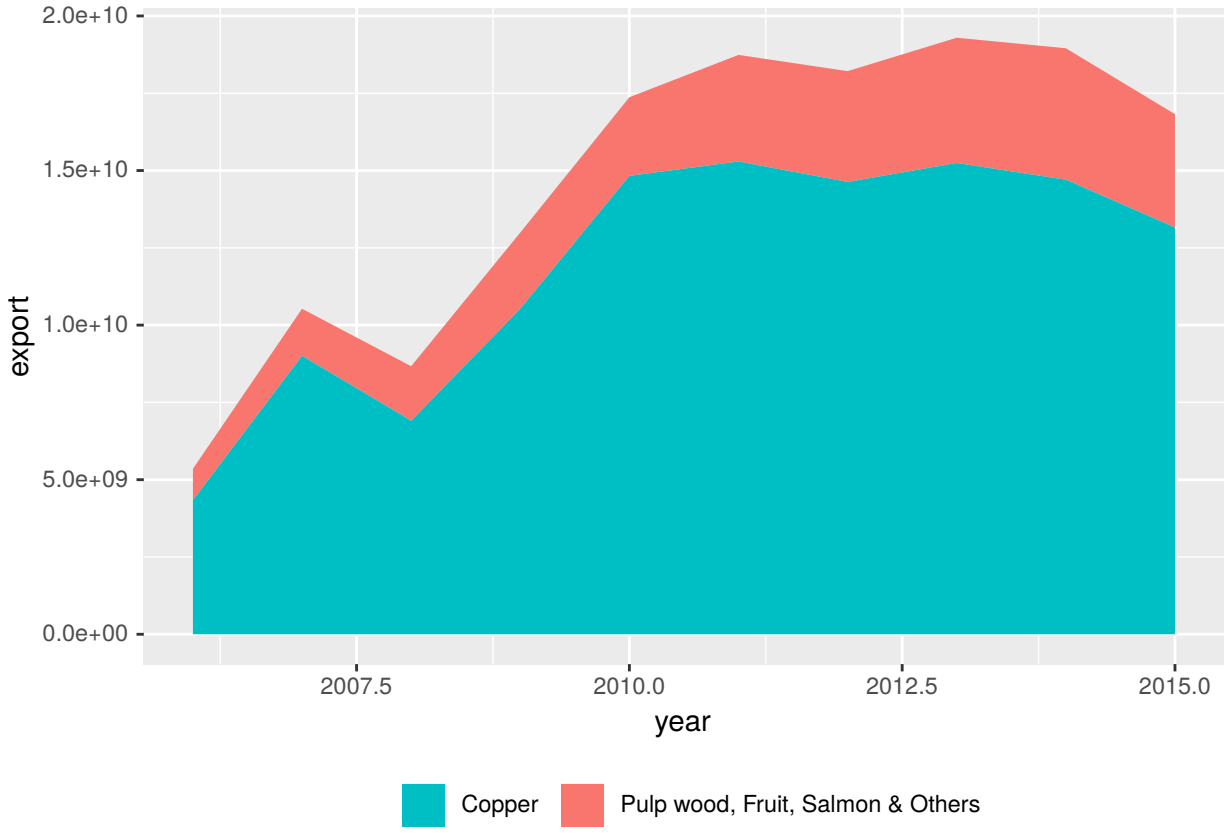
```
p2 <- p2 + theme(legend.position = "bottom", legend.direction = "horizontal",
                 legend.title = element_blank())
p2
```



2.4. Changing variables display

To change the variables displayed name, we need to re-factor our data labels in `exports_data` tibble. We'll also move the legend.

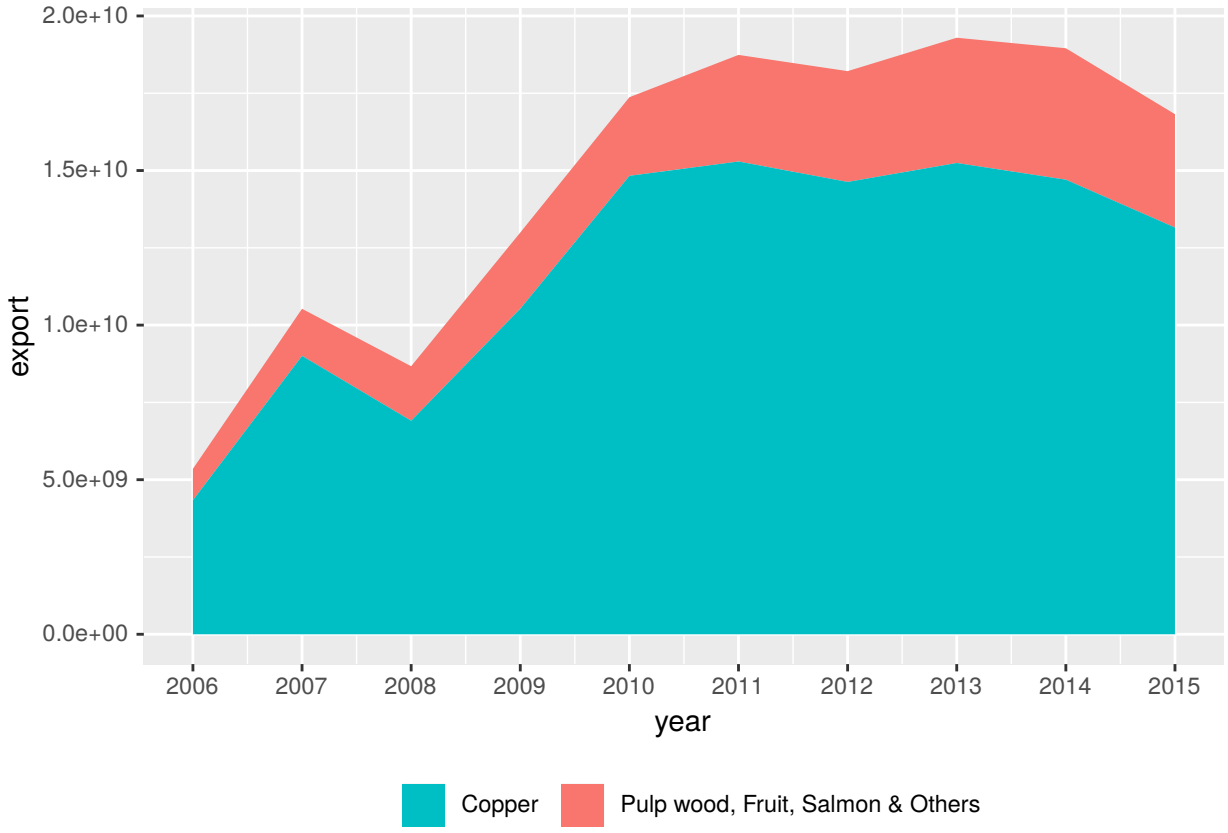
```
exports_data <- exports_data %>%
  mutate(product = factor(product, levels = c("copper","others"),
                          labels = c("Copper ","Pulp wood, Fruit, Salmon & Others")))

p2 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_area() +
  theme(legend.position = "bottom", legend.direction = "horizontal",
        legend.title = element_blank()) +
  guides(fill = guide_legend(reverse = T))
p2
```



2.5. Adjusting x-axis scale

To change the axis tick marks, we use the `scale_x_continuous` and/or `scale_y_continuous` commands.

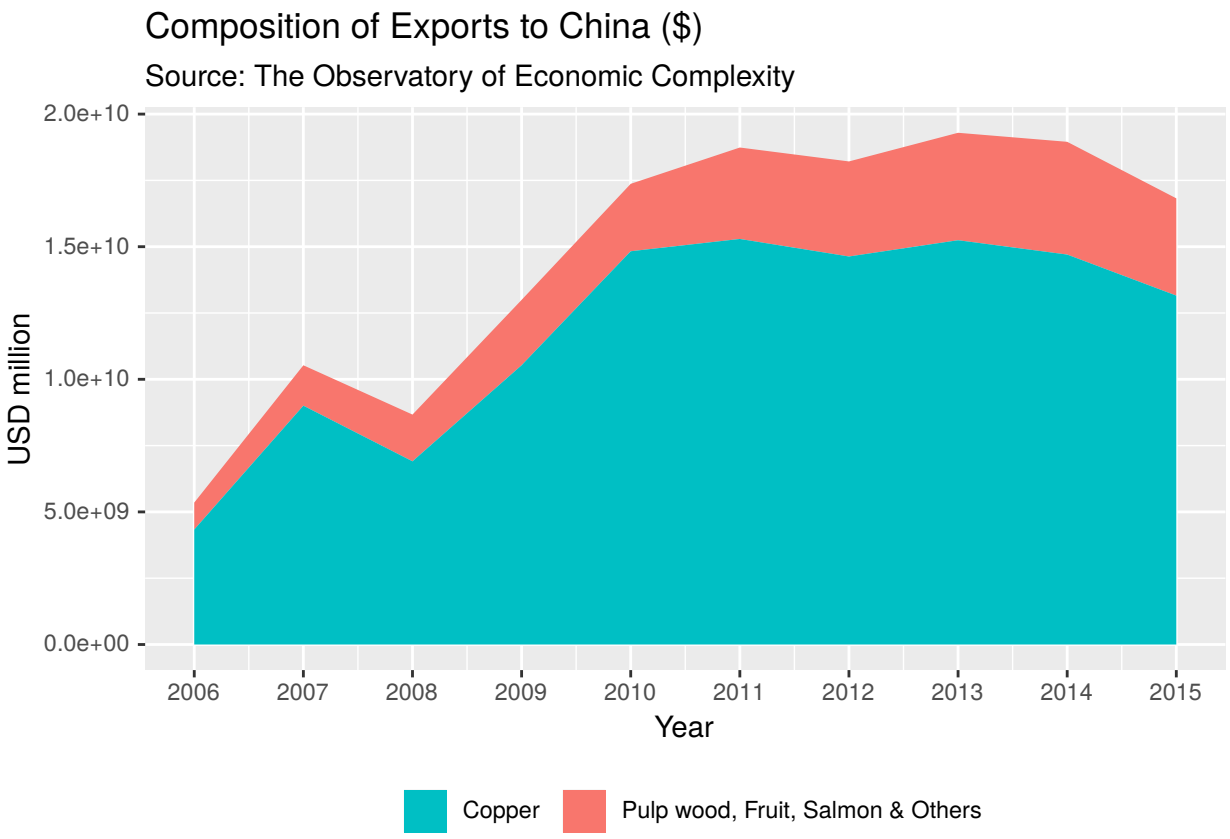
```
p2 <- p2 + scale_x_continuous(breaks = seq(2006,2015,1))
p2
```



2.6. Adjusting axis labels & adding title

To add a title, we include the option `labs` and include the name of the graph as a string argument, and to change the axis names we use the `labs` command too.

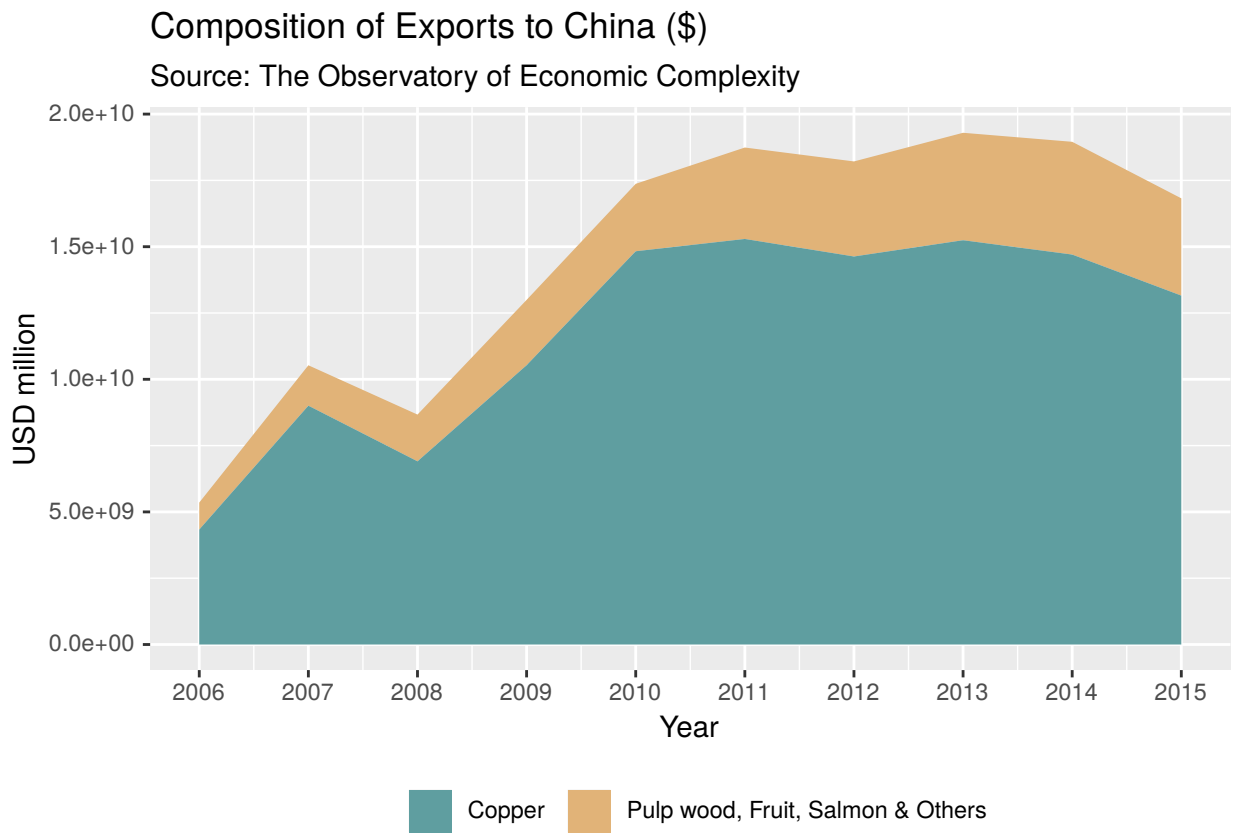
```
p2 <- p2 +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million")
p2
```



2.7. Adjusting color palette

To change the colours, we use the `scale_colour_manual` command. Note that you can reference the specific colours you'd like to use with specific HEX codes. You can also reference colours by name, with the full list of colours recognised by R here.

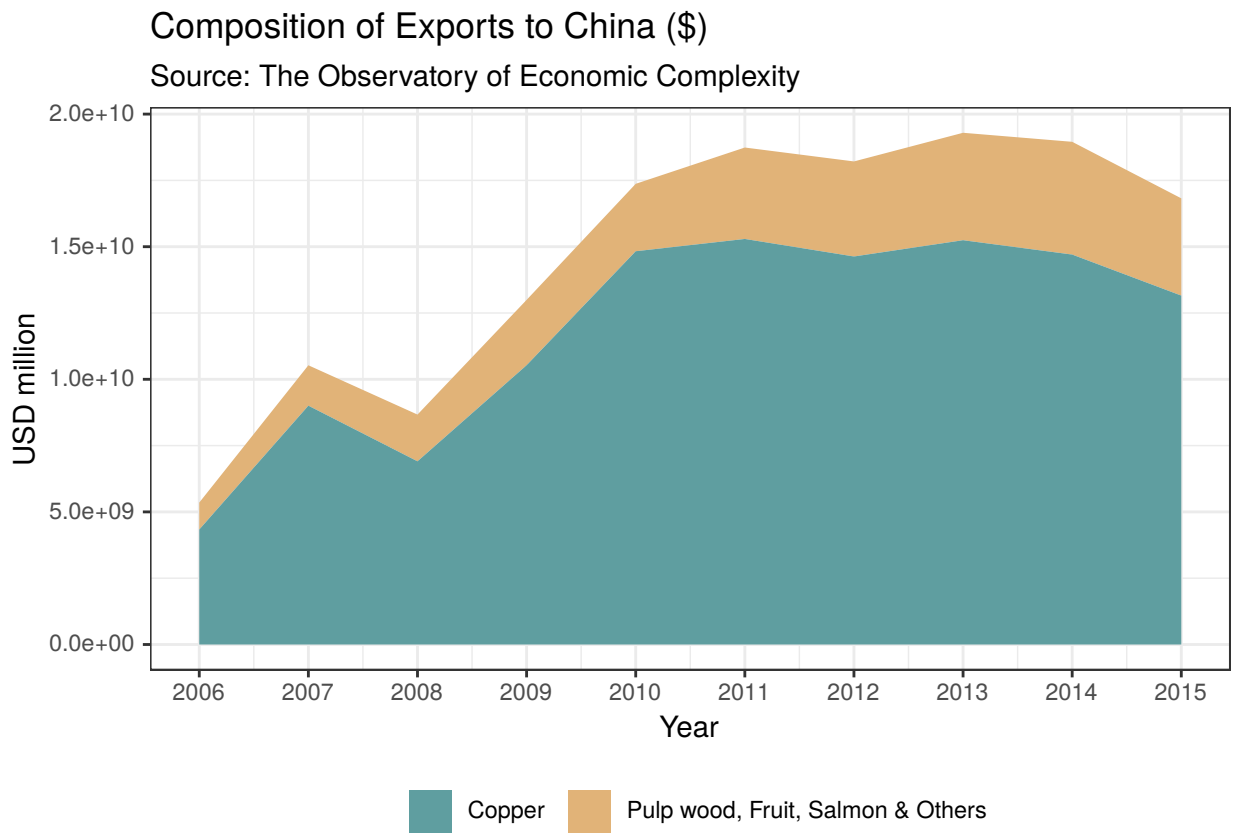
```
fill <- c("#E1B378","#5F9EA0")
p2 <- p2 + scale_fill_manual(values = fill)
p2
```



2.8. Using the white theme

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

```
p2 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_area() +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_fill_manual(values = fill) +
  theme_bw() +
  theme(legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank()) +
  guides(fill = guide_legend(reverse = T))
p2
```



2.9. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

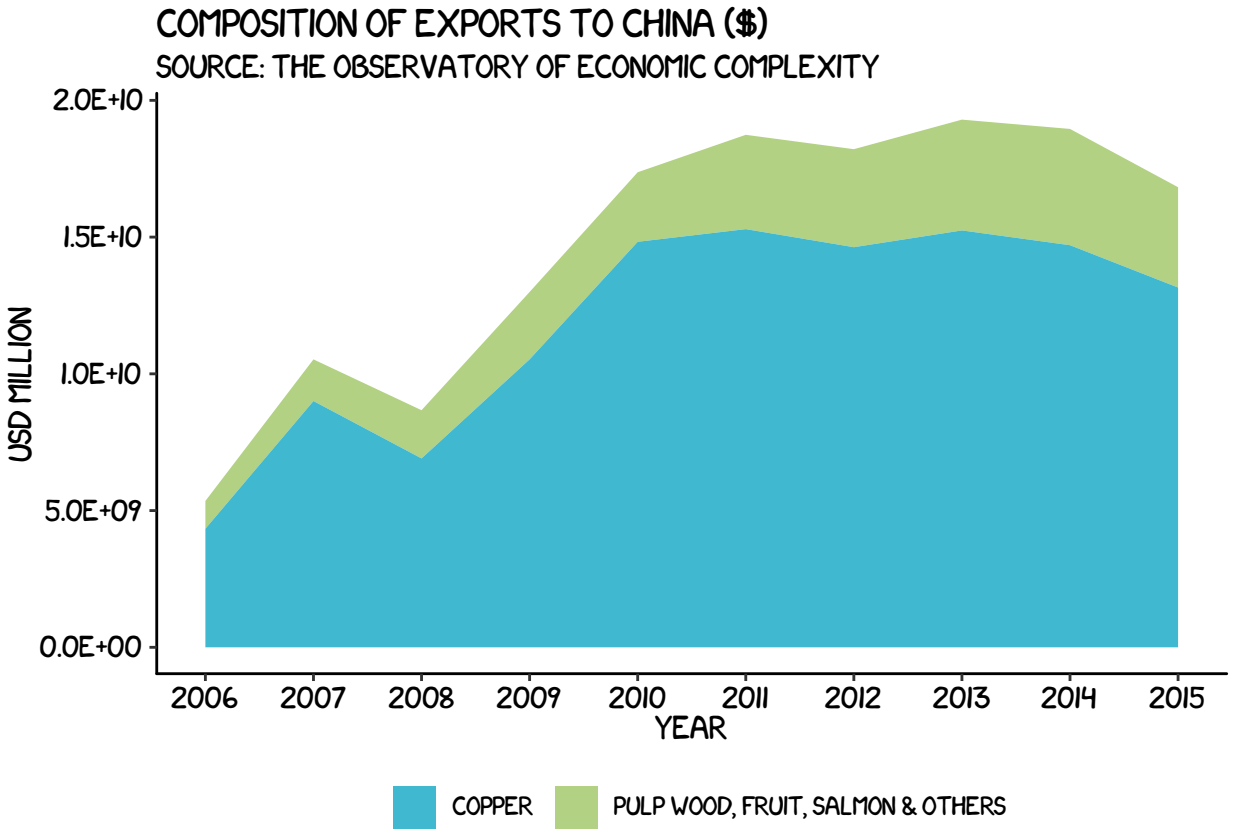
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

2.10. 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.

```
fill <- c("#b2d183","#40b8d0")

p2 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_area() +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_fill_manual(values = fill) +
  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.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = 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"),
    text = element_text(family = "XKCD")) +
  guides(fill = guide_legend(reverse = T))
p2
```

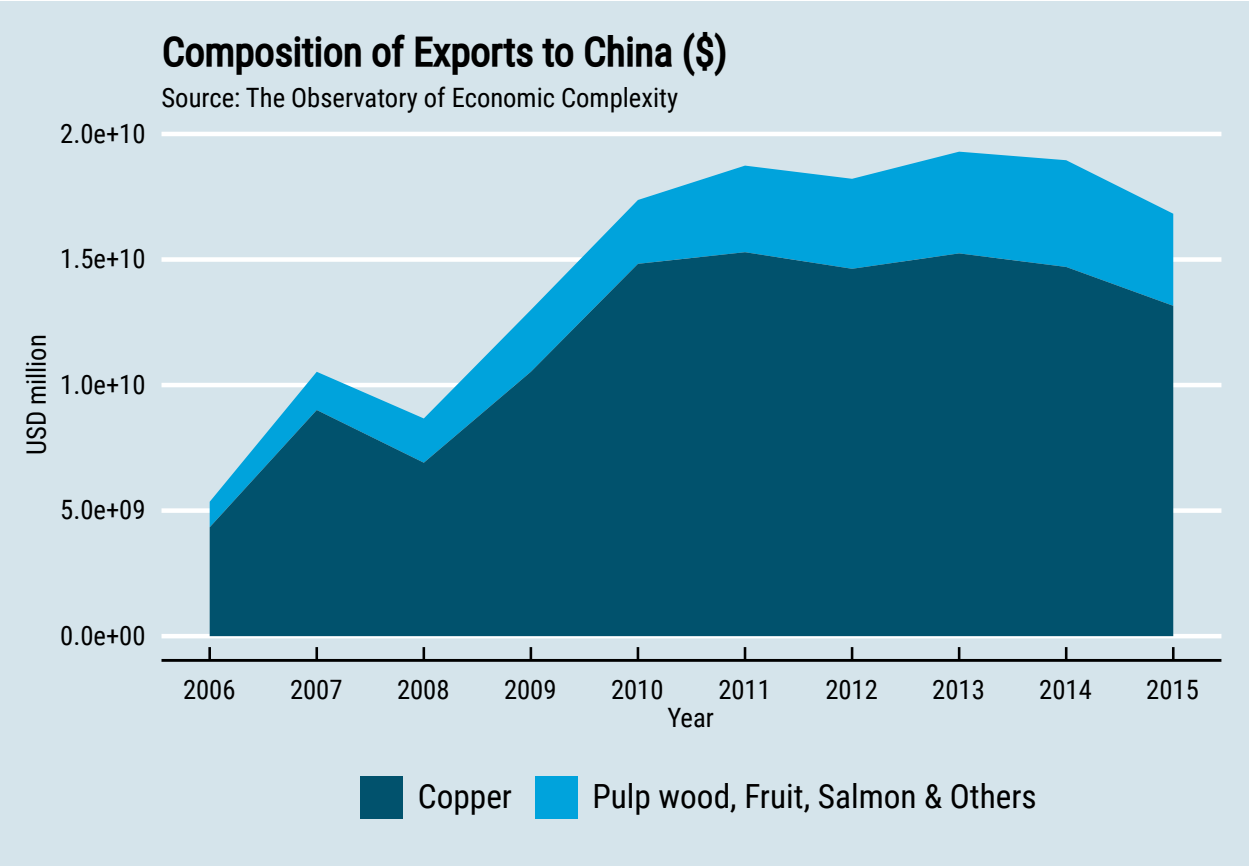


2.11. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
fill <- c("#00a3dc","#01526d")

p2 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_area() +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  theme_economist() + scale_fill_manual(values = fill) +
  theme(axis.line.x = element_line(size = .5, colour = "black"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank(),
    plot.title = element_text(family = "Roboto Condensed"),
    text = element_text(family = "Roboto Condensed")) +
```

```
  guides(fill = guide_legend(reverse = T))
p2
```

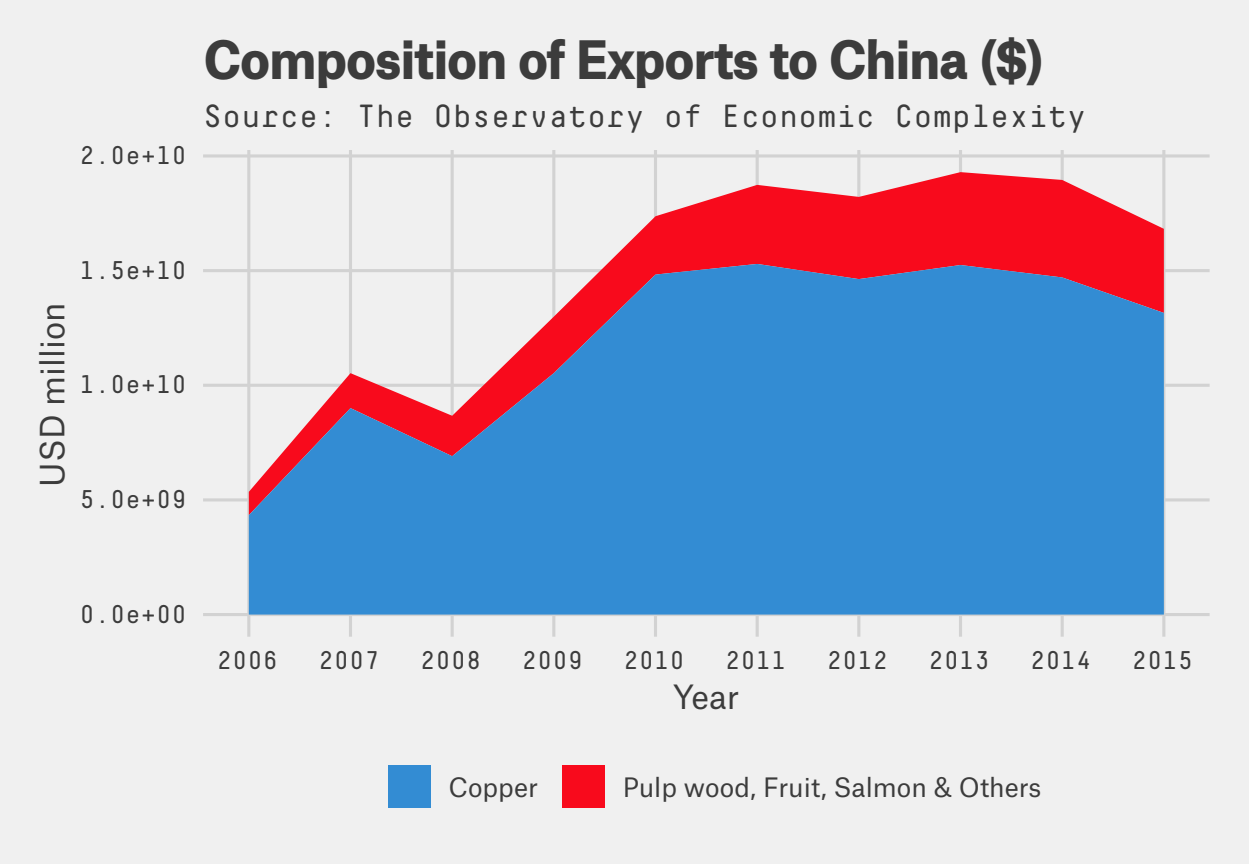


2.12. Using 'Five Thirty Eight' theme

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'.

```
fill <- c("#f80a1c","#338cd3")

p2 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_area() +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  theme_fivethirtyeight() + scale_fill_manual(values = fill) +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
```

```
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    legend.text = element_text(family = "Atlas Grotesk Regular"),
    text = element_text(family = "Decima Mono Pro")) +
  guides(fill = guide_legend(reverse = T))
p2
```

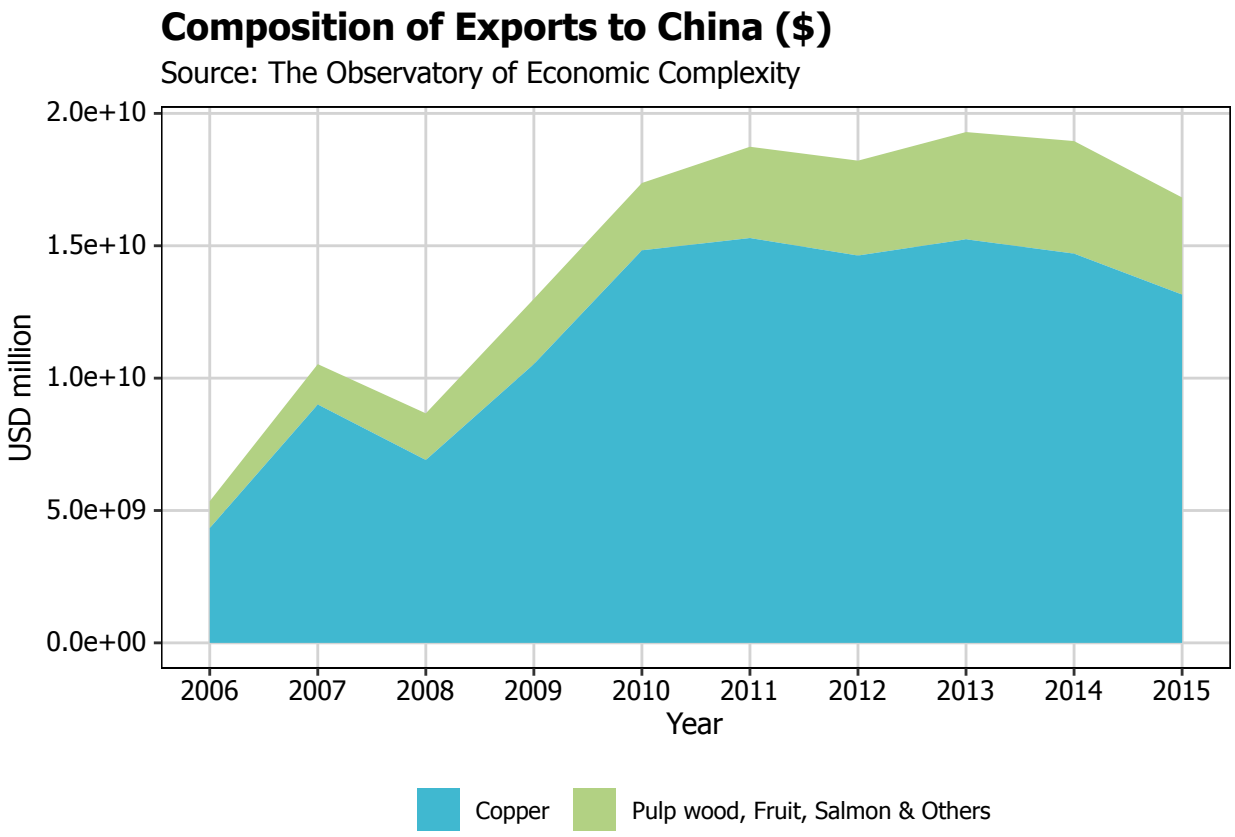


2.13. Creating your own theme

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 <- c("#b2d183","#40b8d0")

p2 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_area() +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_fill_manual(values = fill) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
```

```
    axis.text.x = element_text(colour = "black", size = 10),
    axis.text.y = element_text(colour = "black", size = 10),
    legend.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma")) +
  guides(fill = guide_legend(reverse = T))
p2
```

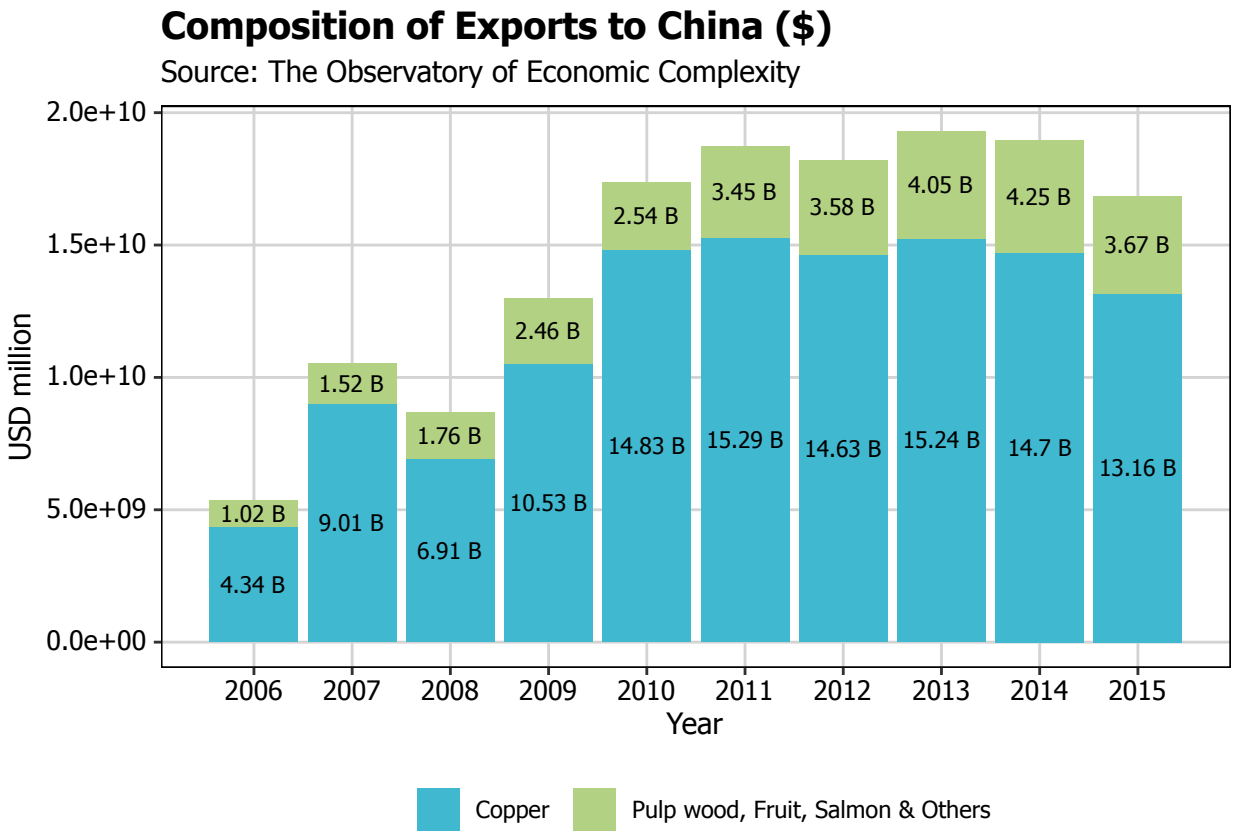


CHAPTER 3

Bar plots

3.1. Introduction

In this chapter, we will work towards creating the area plot below. We will take you from a basic bar plot and explain all the customisations we add to the code step-by-step.



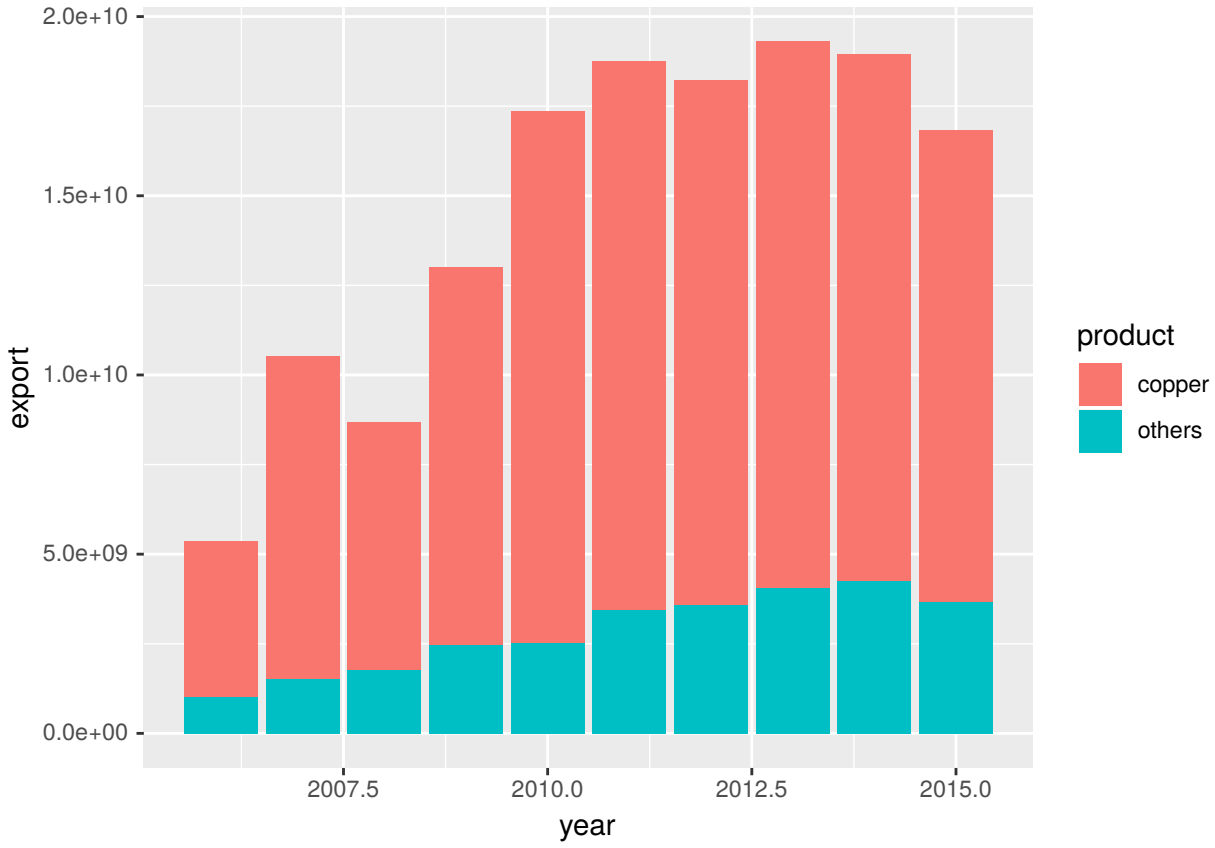
3.2. Basic graph

The first thing to do is load in the data and the libraries, as below:

```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, dplyr, readr, scales, forcats)

chilean_exports <- "year,product,export,percentage
  2006,copper,4335009500,81
  2006,others,1016726518,19
  2007,copper,9005361914,86
  2007,others,1523085299,14
  2008,copper,6907056354,80
  2008,others,1762684216,20
  2009,copper,10529811075,81
  2009,others,2464094241,19
  2010,copper,14828284450,85
  2010,others,2543015596,15
  2011,copper,15291679086,82
  2011,others,3447972354,18
  2012,copper,14630686732,80
  2012,others,3583968218,20
  2013,copper,15244038840,79
  2013,others,4051281128,21
  2014,copper,14703374241,78
  2014,others,4251484600,22
  2015,copper,13155922363,78
  2015,others,3667286912,22
"

exports_data <- read_csv(chilean_exports)
```

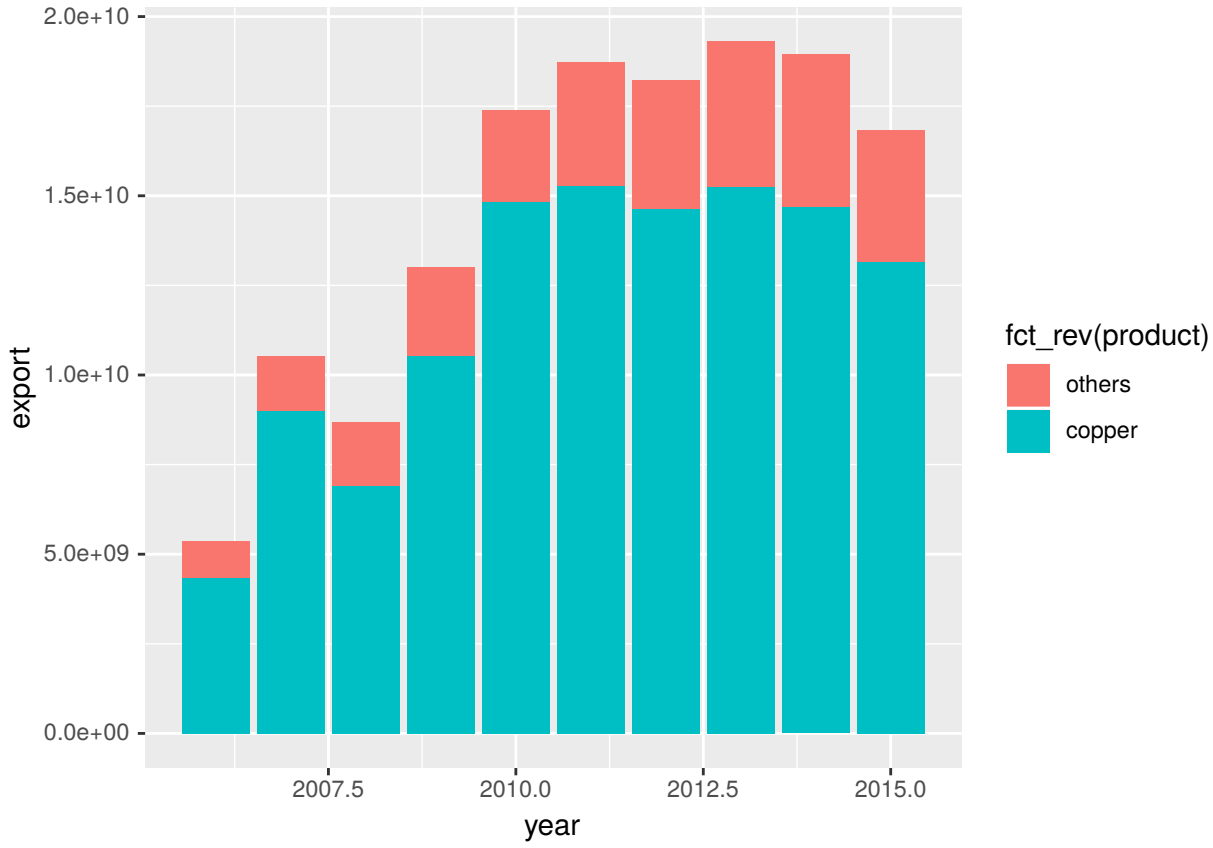
In order to initialise a plot we tell ggplot that `exports_data` is our data, and specify the variables on each axis. We then instruct ggplot to render this as an bar plot by adding the `geom_col` command.

```
p3 <- ggplot(aes(y = export, x = year, fill = product), data = exports_data) +
  geom_col()
p3
```



From now and ongoing we will stack in the opposite order. Changing the fill option to `fill = fct_rev(product)` allows us to do that.

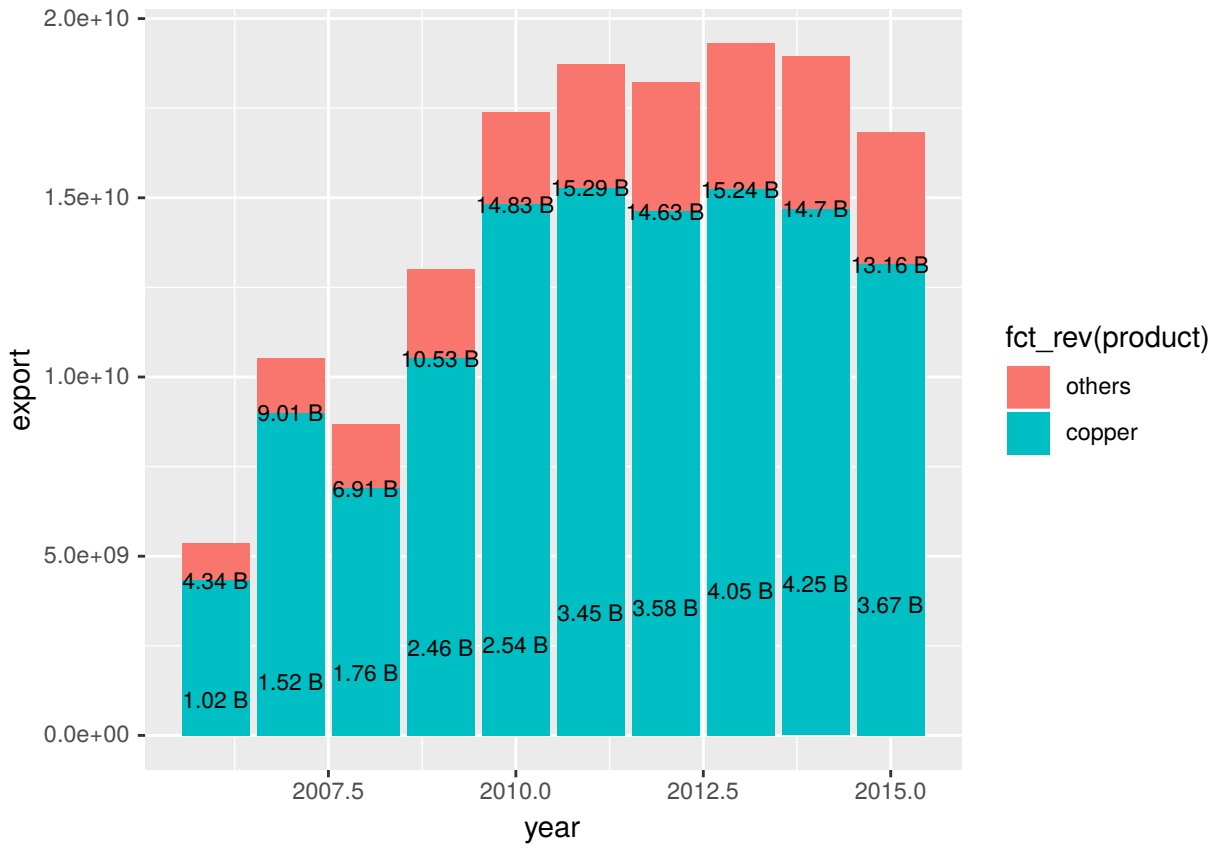
```
p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col()
p3
```



3.3. Adding data labels

To label the bars according to some variable in the data, we add the `label` argument to the `ggplot(aes())` option. In this case, we have labelled the bars with a transformation of the numbers from the `export` variable.

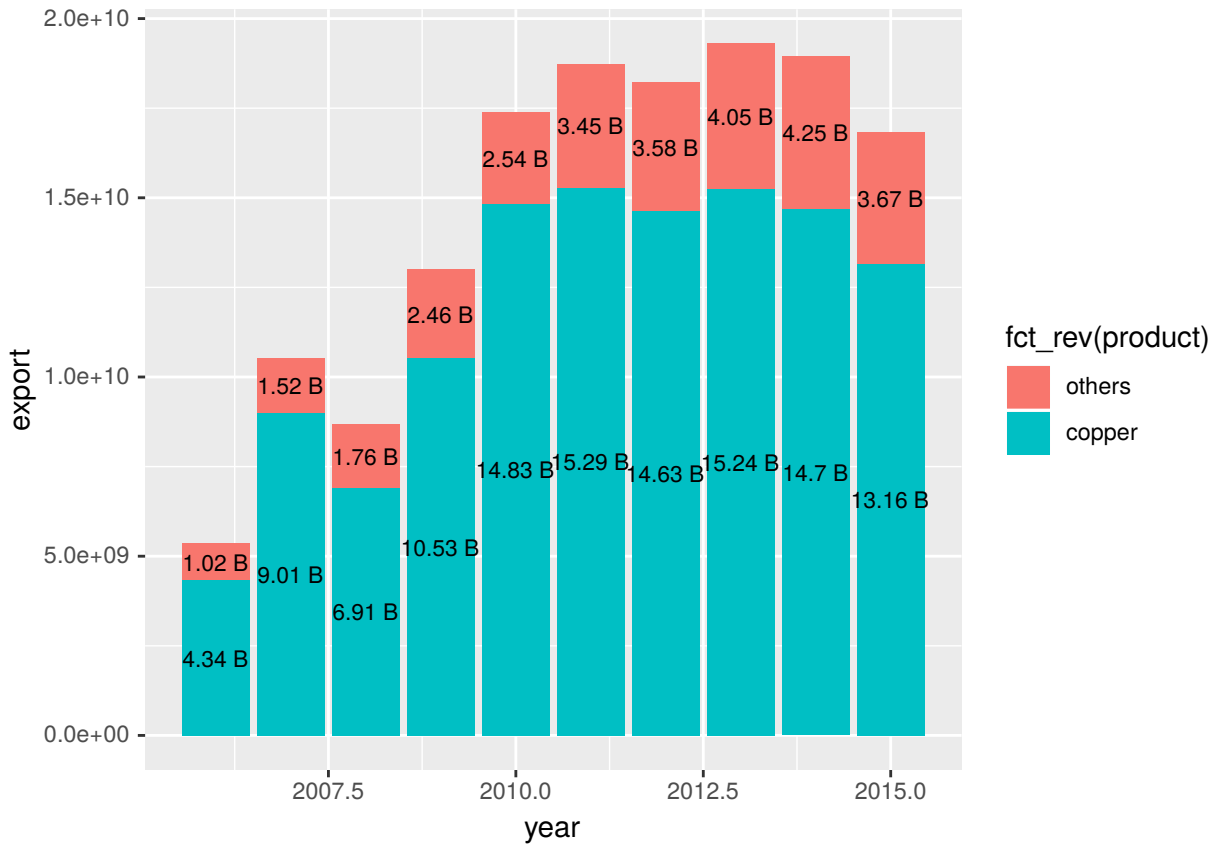
```
exports_data <-  exports_data %>%
  mutate(export_label = paste(round(export/1000000000,2), "B"))
p3 = p3 + geom_text(data = exports_data, aes(x = year, y = export,
                    label = export_label), size = 3)
p3
```



3.4. Adjusting data labels position

One of the easiest possibilities is to use the `position` parameter in `geom_text(aes())`.

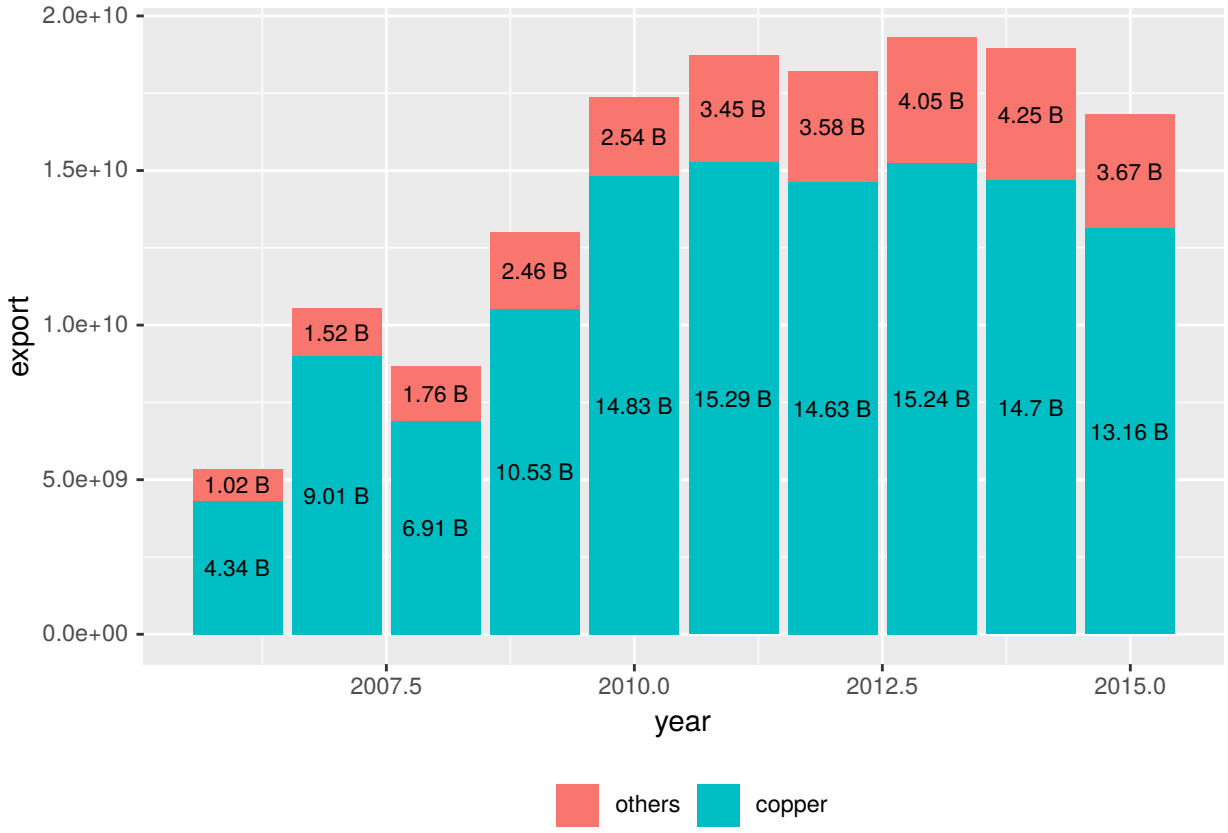
```
p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col()
p3 <- p3 +
  geom_text(aes(label = export_label), position = position_stack(vjust = 0.5),
            size = 3)
p3
```



3.5. Adjusting legend position

To adjust the position of the legend from the default spot of right of the graph, we add the `theme` option and specify the `legend.position = "bottom"` argument. We can also change the title to blank using the `legend.title = element_blank()` argument and change the legend shape using the `legend.direction = "horizontal"` argument.

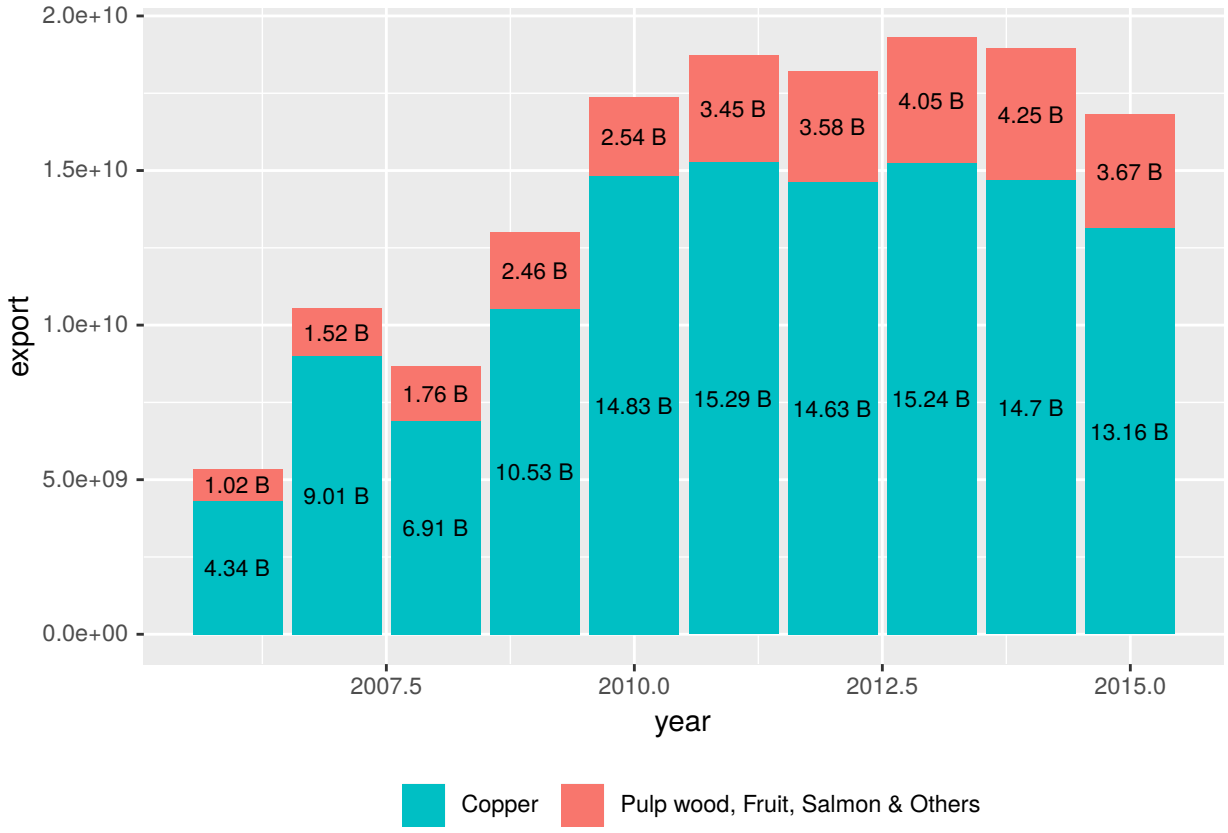
```
p3 <- p3 + theme(legend.position = "bottom", legend.direction = "horizontal",
  legend.title = element_blank())
p3
```



3.6. Changing variables display

To change the variables' displayed name, we need to re-factor our data labels in `exports_data` tibble.

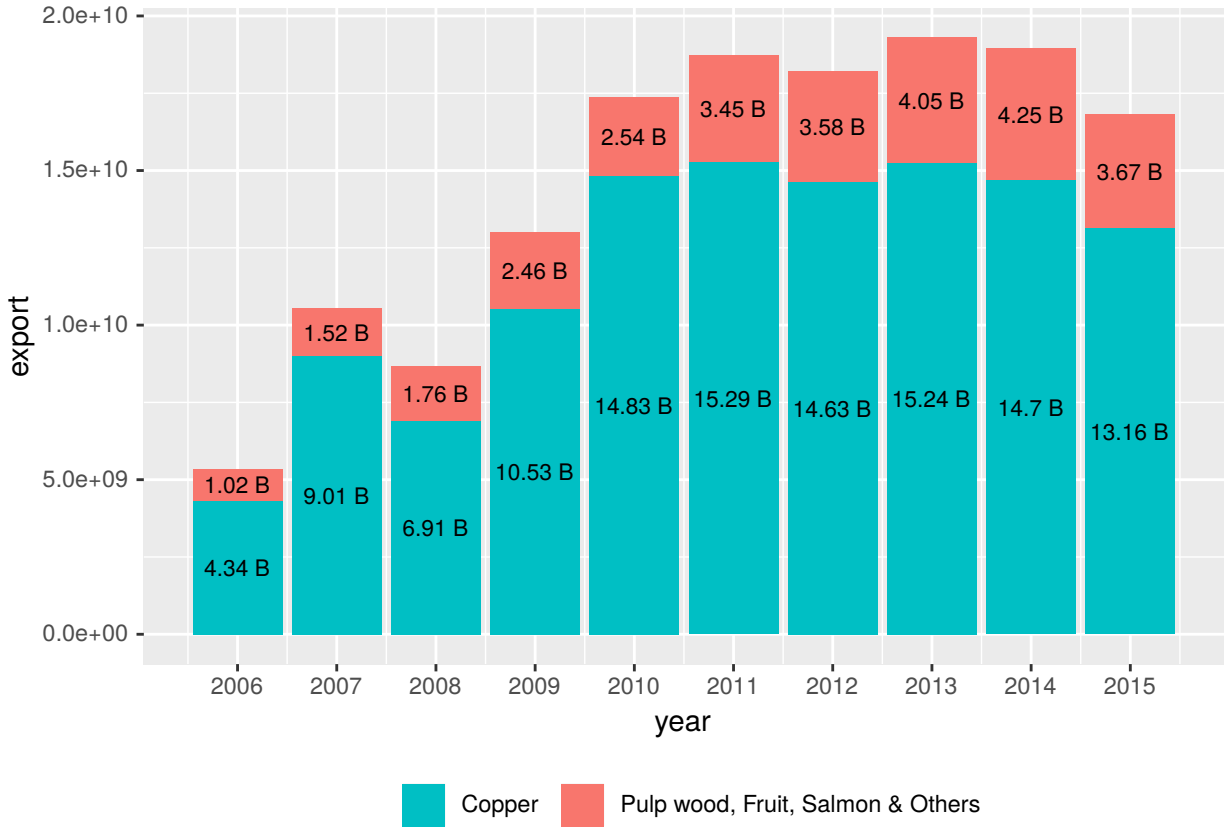
```
exports_data <- exports_data %>%
  mutate(product = factor(product, levels = c("copper","others"),
                          labels = c("Copper ","Pulp wood, Fruit, Salmon & Others")))

p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col() +
  geom_text(aes(label = export_label), position = position_stack(vjust = 0.5),
    size = 3, show.legend = F) +
  theme(legend.position = "bottom", legend.direction = "horizontal",
        legend.title = element_blank()) +
  guides(fill = guide_legend(reverse = T))
p3
```



3.7. Adjusting x-axis scale

To change the axis tick marks, we use the `scale_x_continuous` and/or `scale_y_continuous` commands.

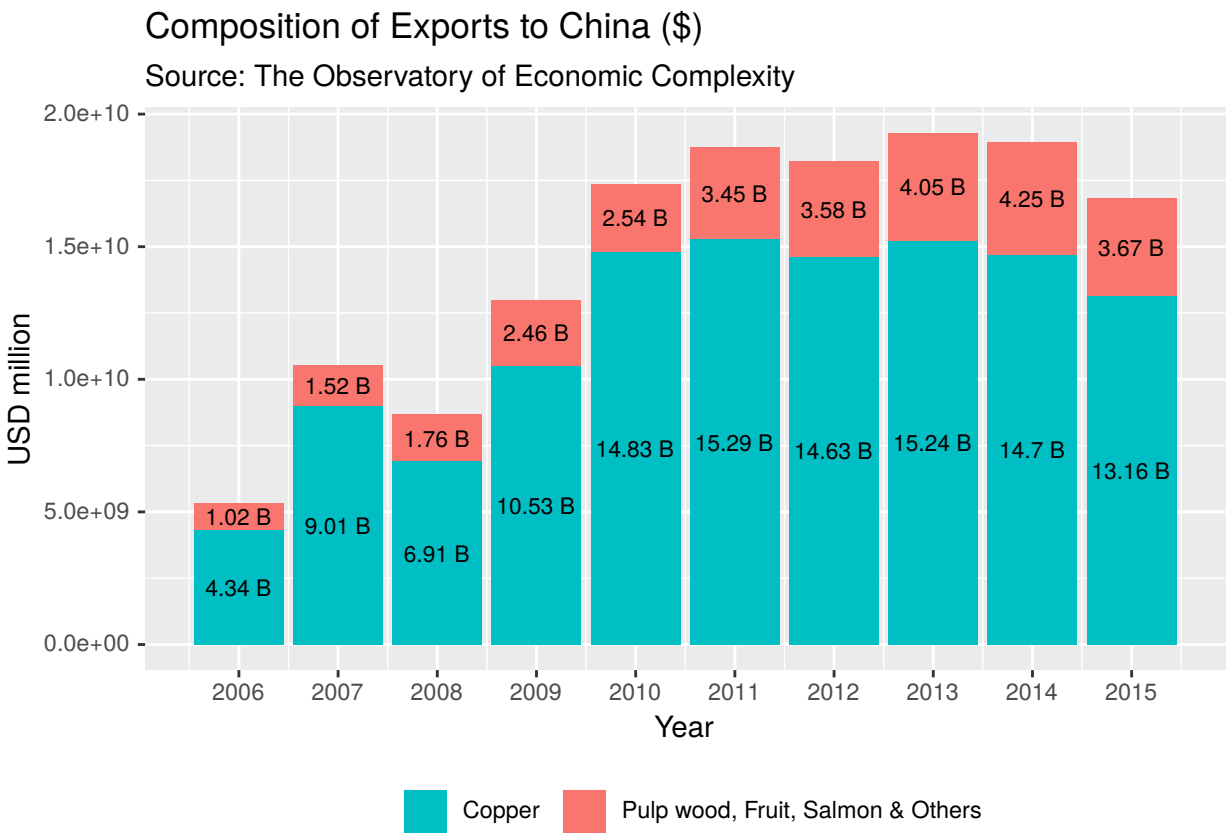
```
p3 <- p3 + scale_x_continuous(breaks = seq(2006,2015,1))
p3
```



3.8. Adjusting axis labels & adding title

To add a title, we include the option `labs` and include the name of the graph as a string argument, and to change the axis names we use the `labs` command too.

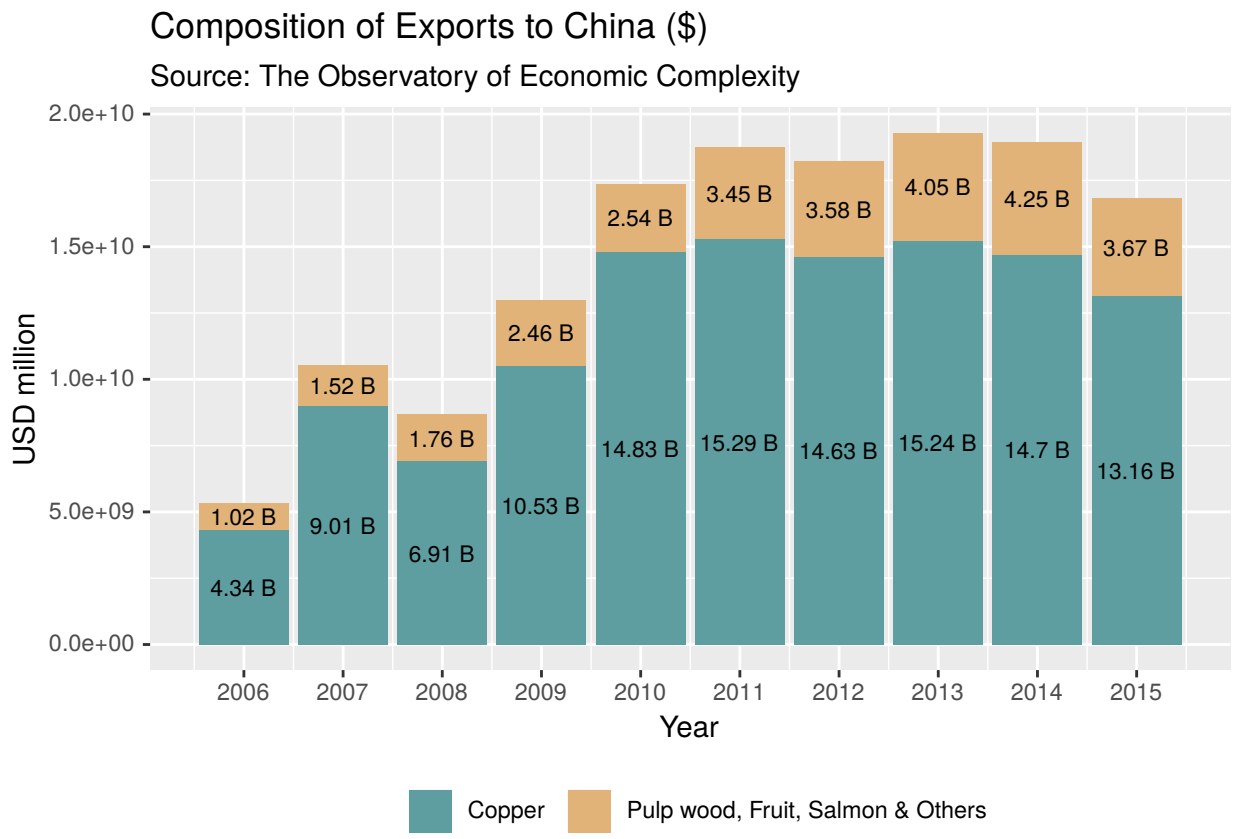
```
p3 <- p3 +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million")
p3
```



3.9. Adjusting color palette

To change the colours, we use the `scale_colour_manual` command. Note that you can reference the specific colours you'd like to use with specific HEX codes. You can also reference colours by name, with the full list of colours recognised by R here.

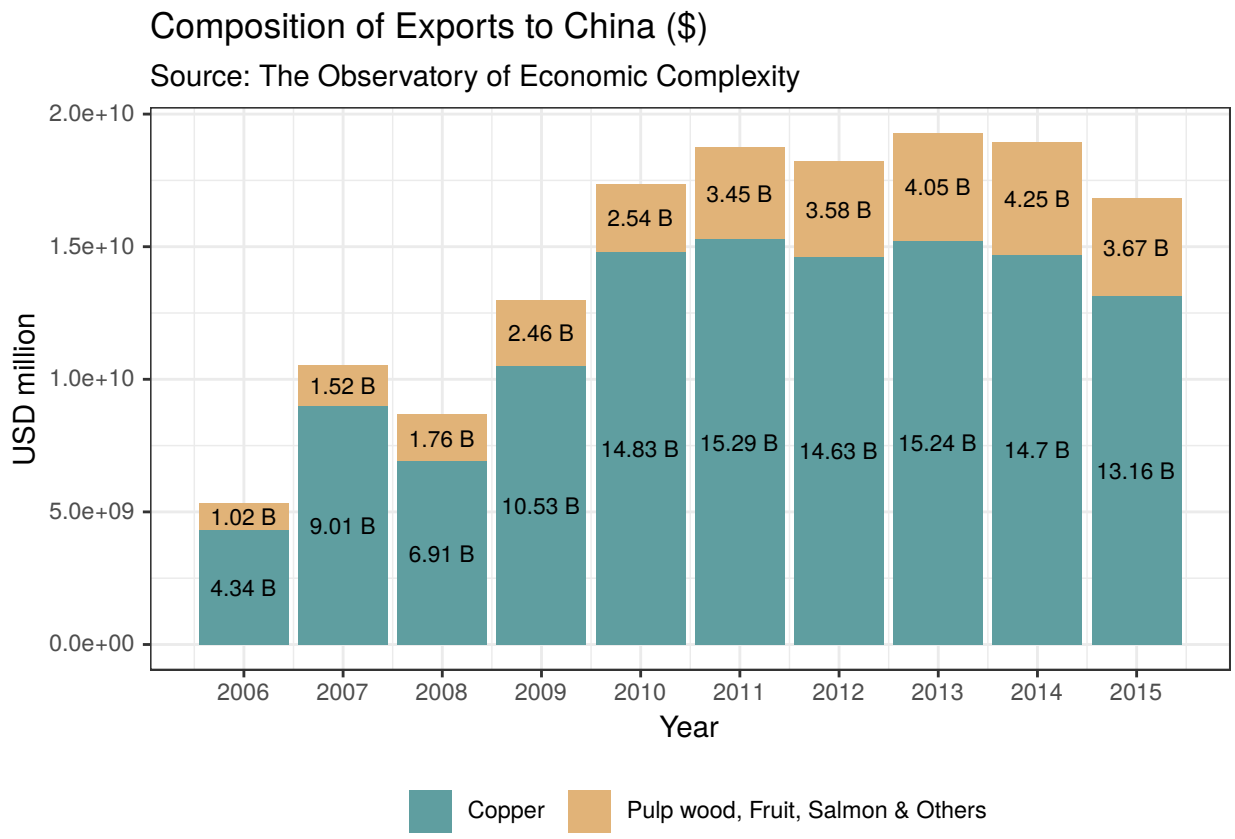
```
fill <- c("#E1B378","#5F9EA0")
p3 <- p3 + scale_fill_manual(values = fill)
p3
```



3.10. Using the white theme

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

```
p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col() +
  geom_text(aes(label = export_label), position = position_stack(vjust = 0.5),
            size = 3, show.legend = F) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_fill_manual(values = fill) +
  theme_bw() +
  theme(legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank()) +
  guides(fill = guide_legend(reverse = T))
p3
```



3.11. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

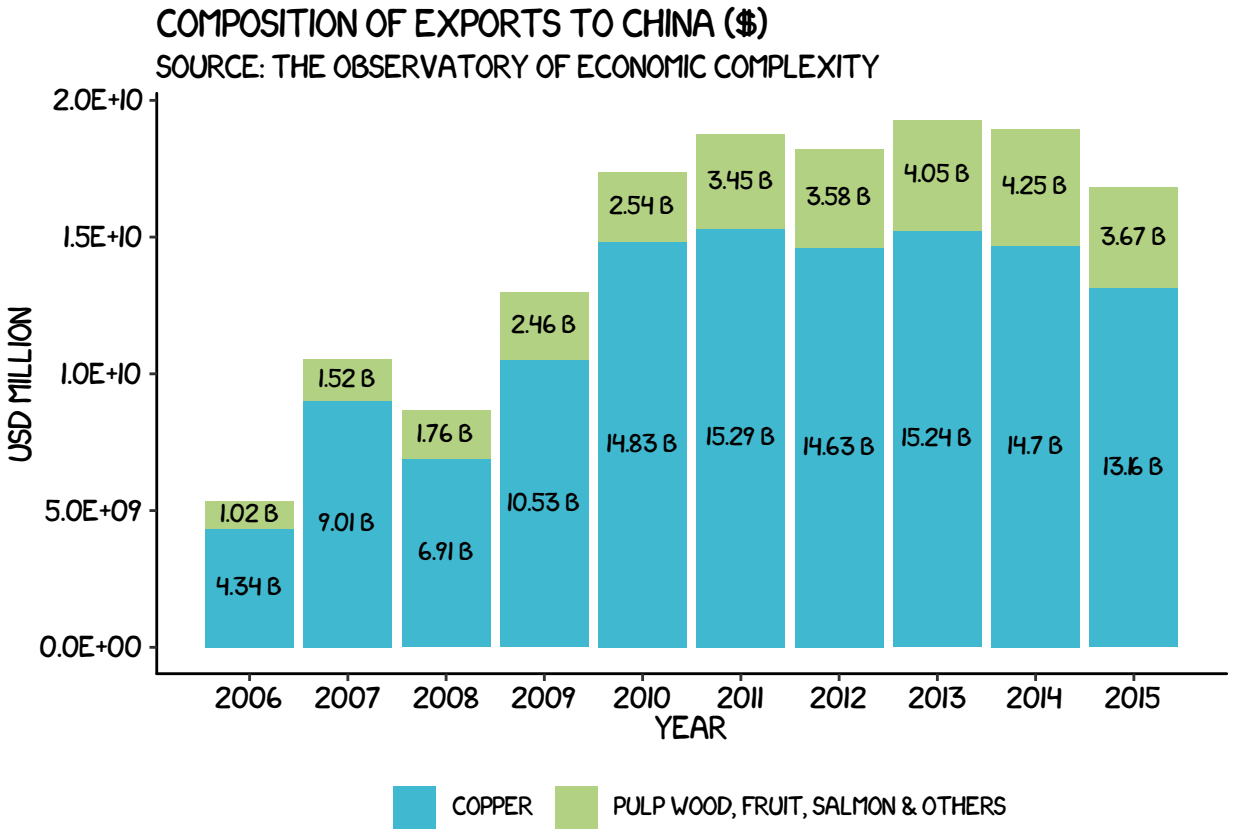
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

3.12. 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.

```
fill <- c("#b2d183","#40b8d0")

p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col() +
  geom_text(aes(label = export_label), position = position_stack(vjust = 0.5),
            size = 3, family = "XKCD", show.legend = F) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_fill_manual(values = fill) +
  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.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = 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"),
    text = element_text(family = "XKCD")) +
  guides(fill = guide_legend(reverse = T))
p3
```

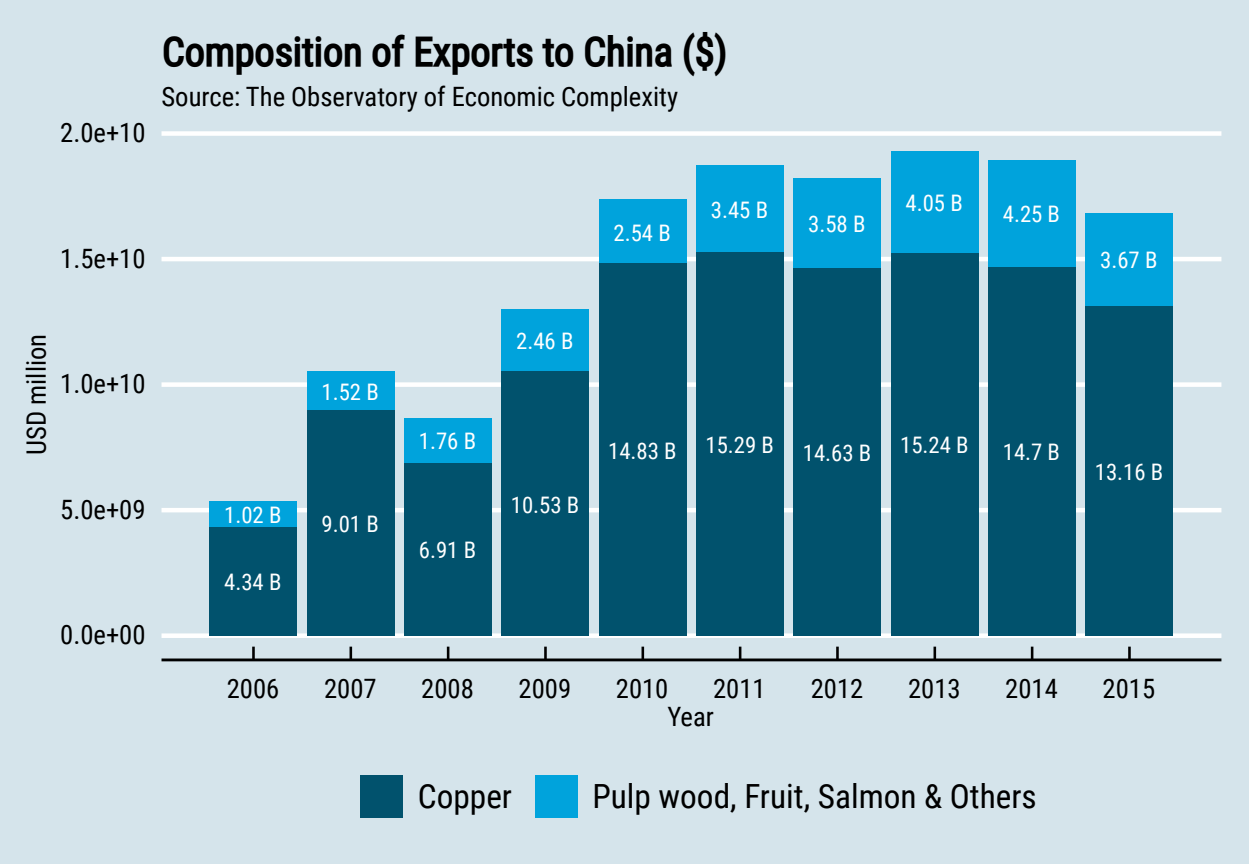


3.13. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
fill <- c("#00a3dc","#01526d")

p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col() +
  geom_text(aes(label = export_label), position = position_stack(vjust = 0.5),
            colour = "white", size = 3, family = "Roboto Condensed",
            show.legend = F) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  theme_economist() + scale_fill_manual(values = fill) +
  theme(axis.line.x = element_line(size = .5, colour = "black"),
    legend.position = "bottom",
    legend.direction = "horizontal",
```

```
    legend.title = element_blank(),
    plot.title = element_text(family = "Roboto Condensed"),
    text = element_text(family = "Roboto Condensed")) +
  guides(fill = guide_legend(reverse = T))
p3
```

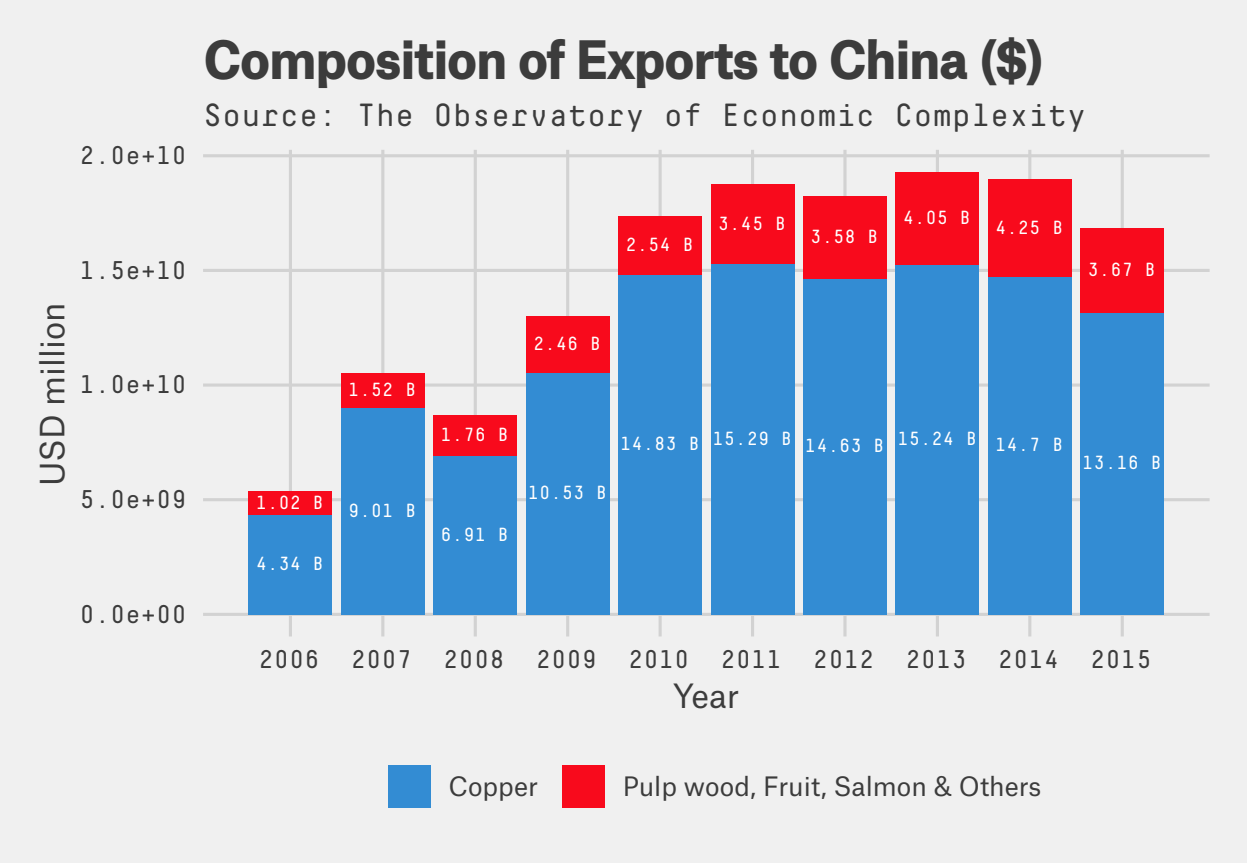


3.14. Using 'Five Thirty Eight' theme

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'.

```
fill <- c("#f80a1c","#338cd3")

p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col() +
  geom_text(aes(label = export_label), position = position_stack(vjust = 0.5),
            colour = "white", size = 2.5, family = "Decima Mono Pro",
            show.legend = F) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
```

```
      subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  theme_fivethirtyeight() + scale_fill_manual(values = fill) +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    legend.text = element_text(family = "Atlas Grotesk Regular"),
    text = element_text(family = "Decima Mono Pro")) +
  guides(fill = guide_legend(reverse = T))
p3
```

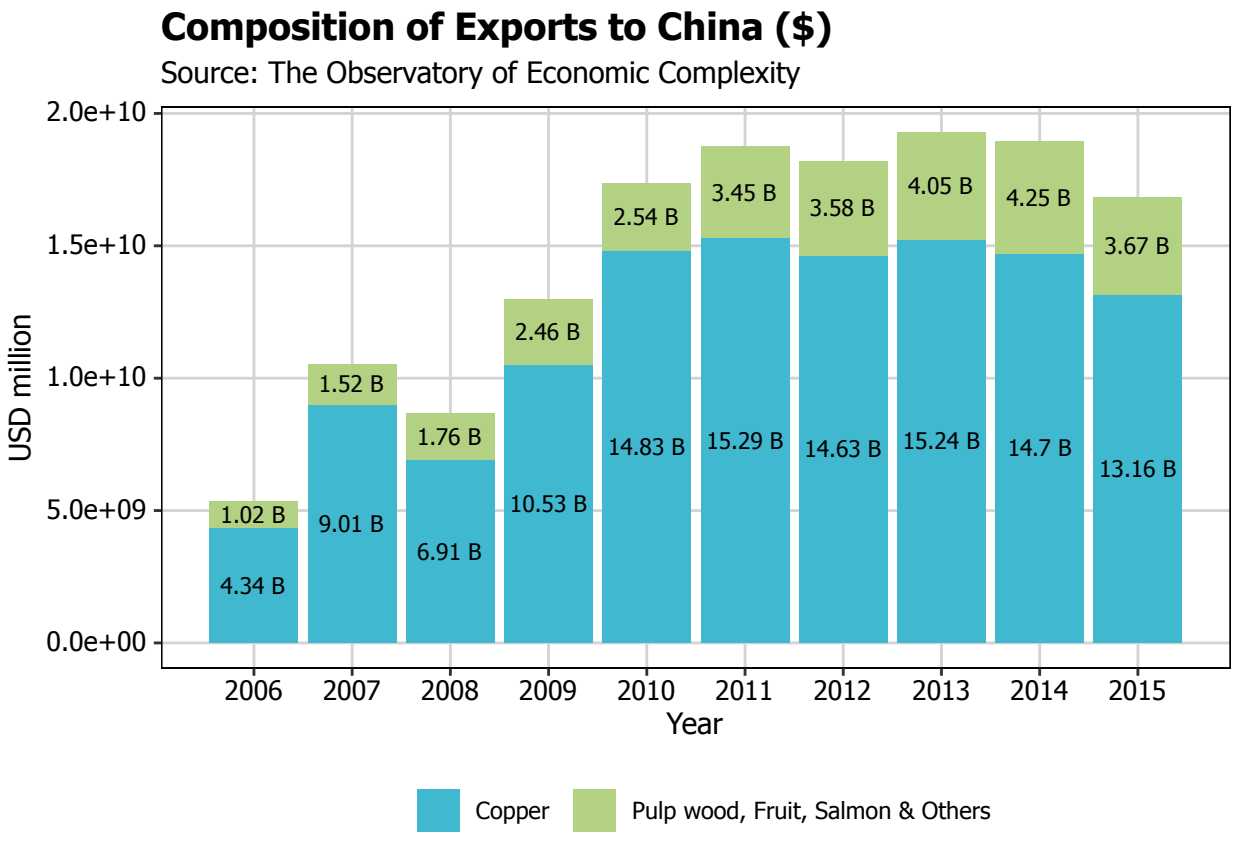


3.15. Creating your own theme

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 <- c("#b2d183","#40b8d0")

p3 <- ggplot(aes(y = export, x = year, fill = fct_rev(product)), data = exports_data) +
  geom_col() +
```

```
  geom_text(aes(label = export_label), position = position_stack(vjust = 0.5),
            colour = "black", family = "Tahoma", size = 3, show.legend = F) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "USD million") +
  scale_fill_manual(values = fill) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 10),
    axis.text.y = element_text(colour = "black", size = 10),
    legend.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma")) +
  guides(fill = guide_legend(reverse = T))
p3
```

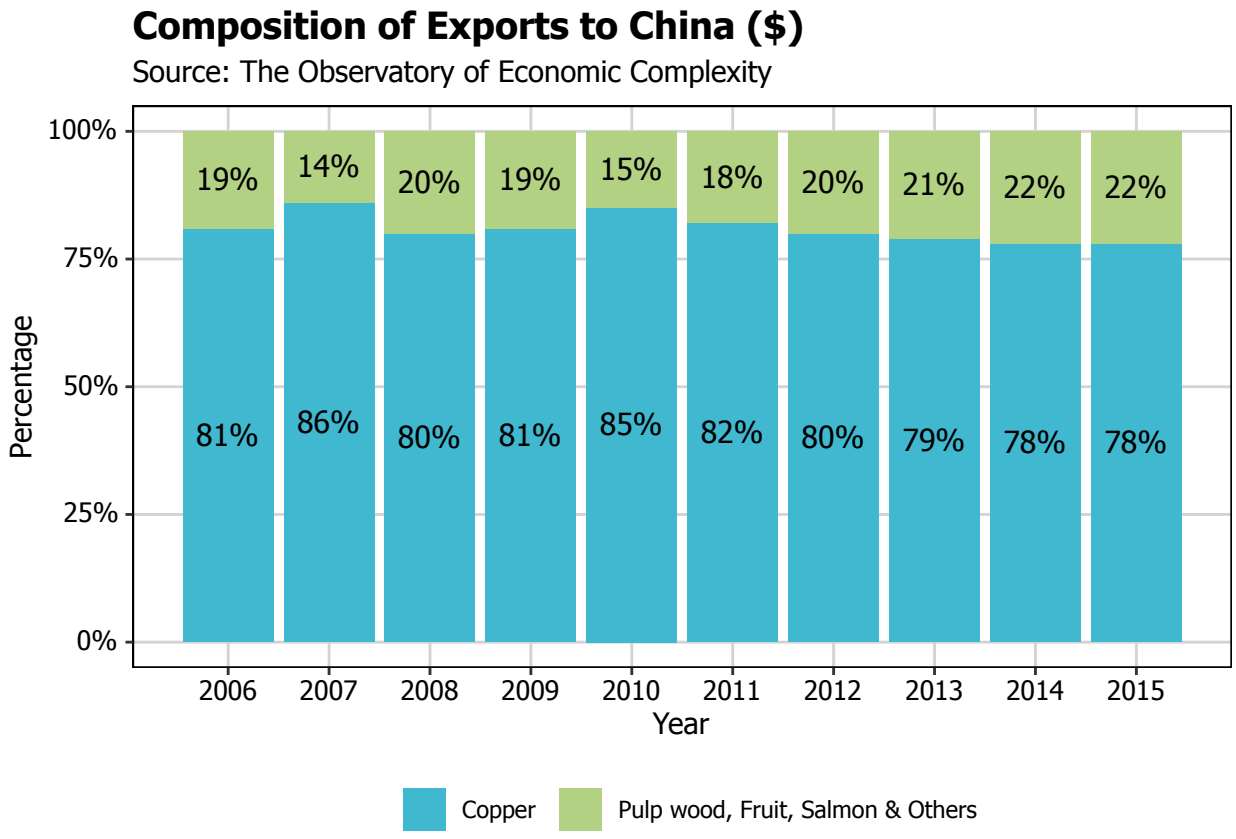


CHAPTER 4

Stacked bar plots

4.1. Introduction

In this chapter, we will work towards creating the bar plot below. We will take you from a basic stacked bar plot and explain all the customisations we add to the code step-by-step.



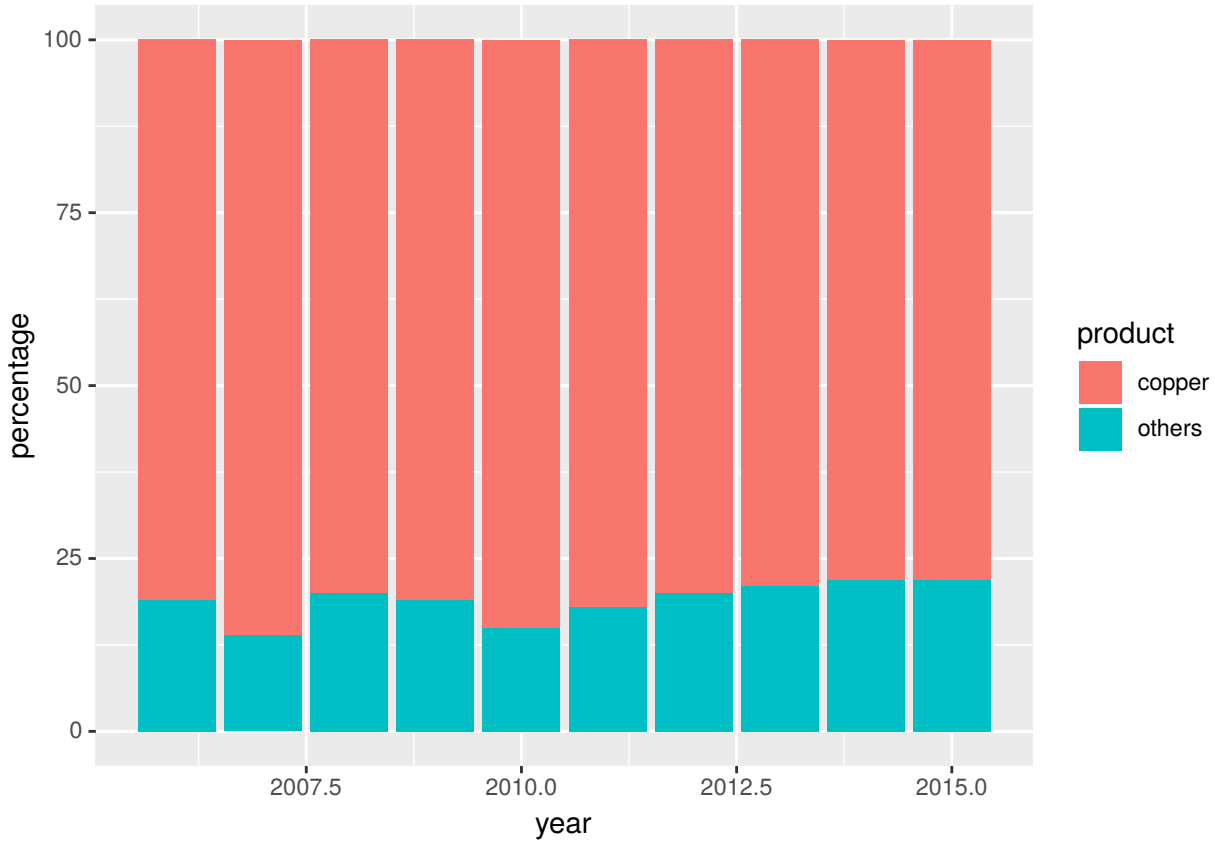
4.2. Basic graph

In order to initialise a plot we tell ggplot that `exports_data` is our data, and specify the variables on each axis. We then instruct ggplot to render this as a stacked bar plot by adding the `geom_col` command.

The first thing to do is load in the data and the libraries, as below:

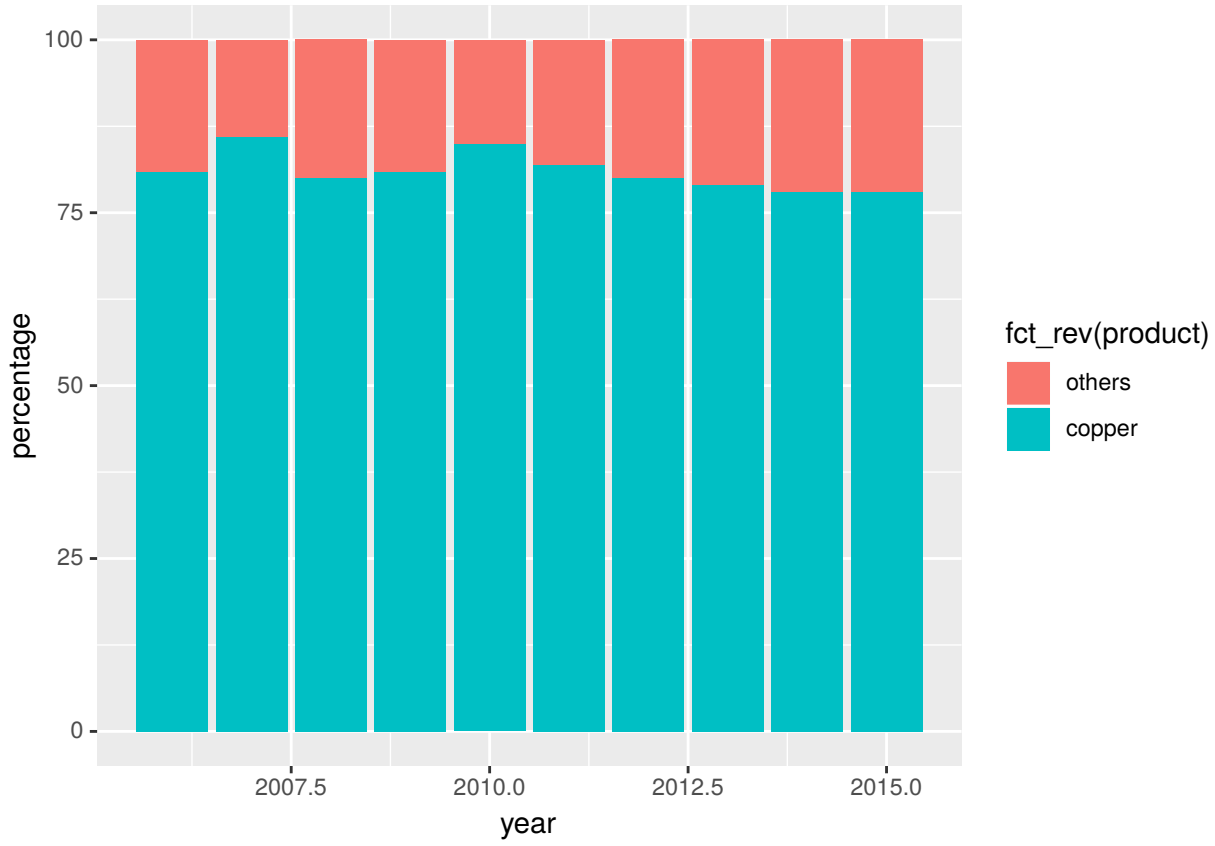
```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, dplyr, readr, scales, forcats)

chilean_exports <- "year,product,export,percentage
  2006,copper,4335009500,81
  2006,others,1016726518,19
  2007,copper,9005361914,86
  2007,others,1523085299,14
  2008,copper,6907056354,80
  2008,others,1762684216,20
  2009,copper,10529811075,81
  2009,others,2464094241,19
  2010,copper,14828284450,85
  2010,others,2543015596,15
  2011,copper,15291679086,82
  2011,others,3447972354,18
  2012,copper,14630686732,80
  2012,others,3583968218,20
  2013,copper,15244038840,79
  2013,others,4051281128,21
  2014,copper,14703374241,78
  2014,others,4251484600,22
  2015,copper,13155922363,78
  2015,others,3667286912,22
"

exports_data <- read_csv(chilean_exports)
```

```
p4 <- ggplot(aes(y = percentage, x = year, fill = product), data = exports_data) +
  geom_col()
p4
```



From now and ongoing we will stack in the opposite order. Changing the fill option to `fill = fct_rev(product)` allows us to do that.

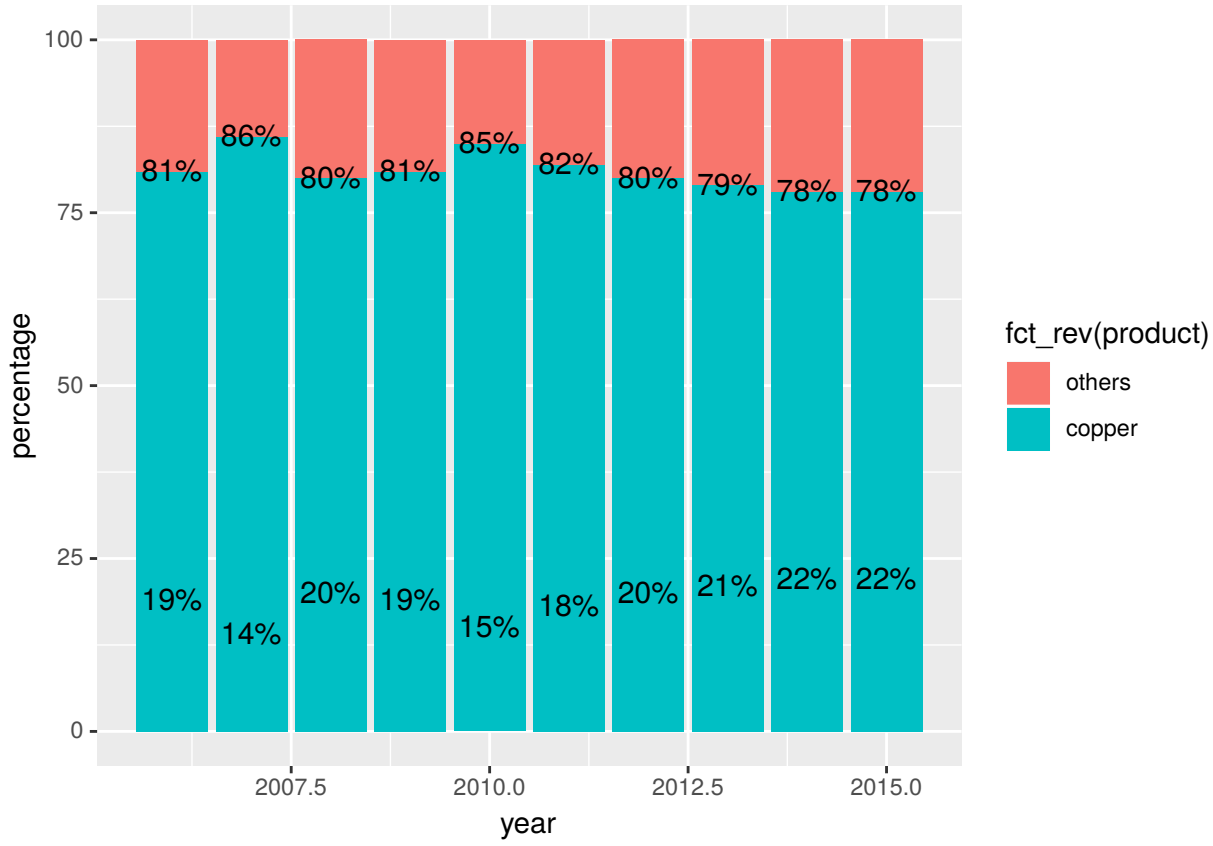
```
p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col()
p4
```



4.3. Adding data labels

To label the bars according to some variable in the data, we add the `label` argument to the `ggplot(aes())` option. In this case, we have labelled the bars with numbers from the `export` variable.

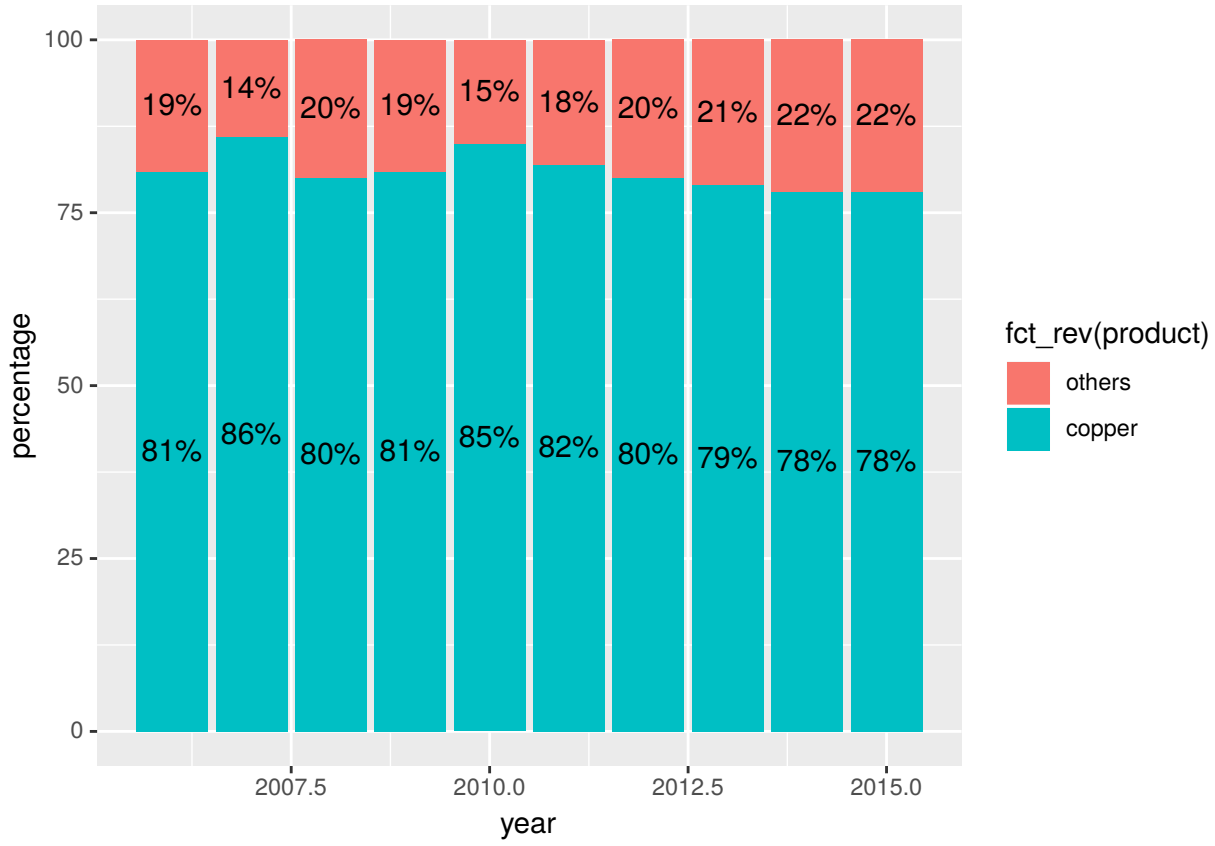
```
p4 <- p4 + geom_text(data = exports_data, aes(x = year, y = percentage,
  label = paste0(percentage,"%")), size = 4)
p4
```



4.4. Adjusting data labels position

To adjust the position of the data labels from the default placement, we use the `ddply` function on the data, and create a new variable called `pos`. This variable is at the centre of each bar and can be used to specify the position of the labels by assigning it to the `y` argument in `geom_text(aes())`.

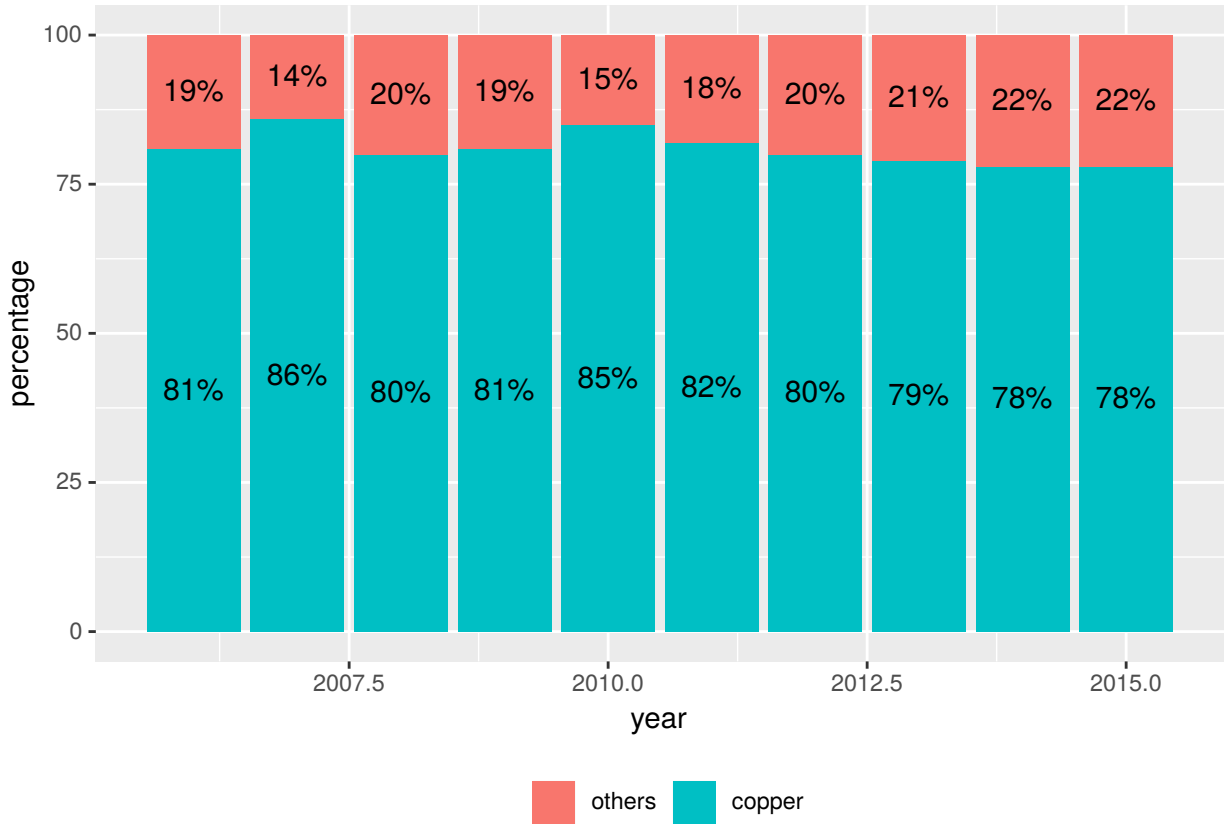
```
p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col()
p4 <- p4 + geom_text(aes(label = paste0(percentage,"%")),
                     position = position_stack(vjust = 0.5), size = 4)
p4
```



4.5. Adjusting legend position

To adjust the position of the legend from the default spot of right of the graph, we add the `theme` option and specify the `legend.position = "bottom"` argument. We can also change the title to blank using the `legend.title = element_blank()` argument and change the legend shape using the `legend.direction = "horizontal"` argument.

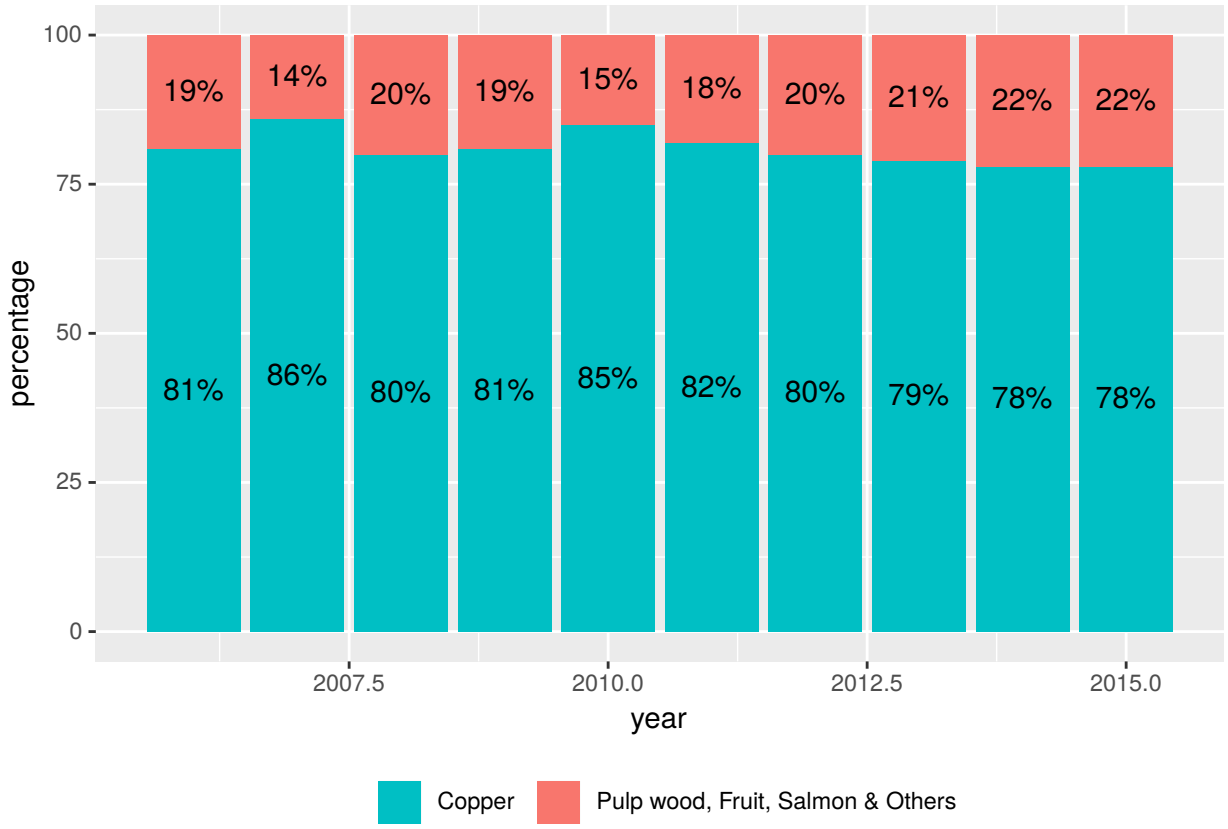
```
p4 <- p4 + theme(legend.position = "bottom", legend.direction = "horizontal",
  legend.title = element_blank())
p4
```



4.6. Changing variables display

To change the variables' displayed name, we need to re-factor our data labels in `exports_data` tibble.

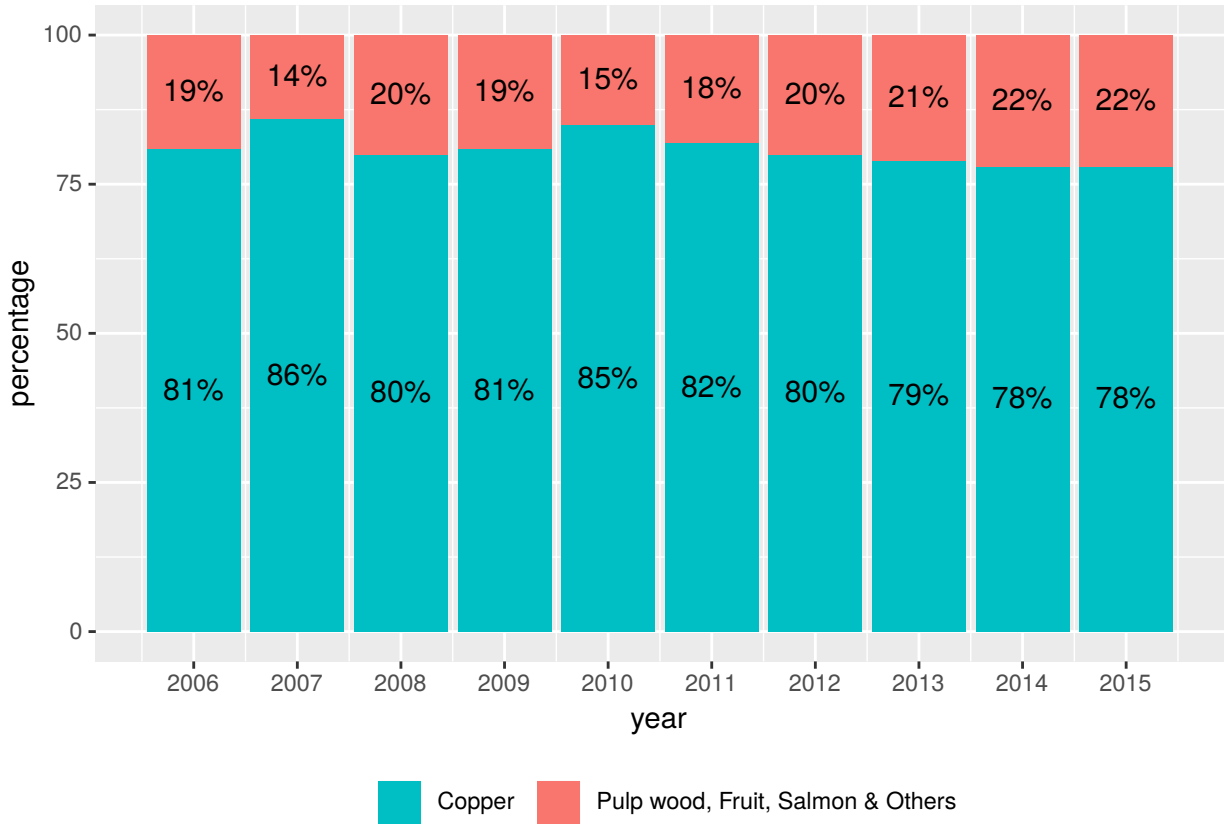
```
exports_data <- exports_data %>%
  mutate(product = factor(product, levels = c("copper","others"),
                          labels = c("Copper ","Pulp wood, Fruit, Salmon & Others")))

p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col() +
  geom_text(aes(label = paste0(percentage,"%")),
            position = position_stack(vjust = 0.5), size = 4) +
  theme(legend.position = "bottom", legend.direction = "horizontal", legend.title =
    element_blank()) +
  guides(fill = guide_legend(reverse = T))
p4
```



4.7. Adjusting x-axis scale

To change the axis tick marks, we use the `scale_x_continuous` and/or `scale_y_continuous` commands.

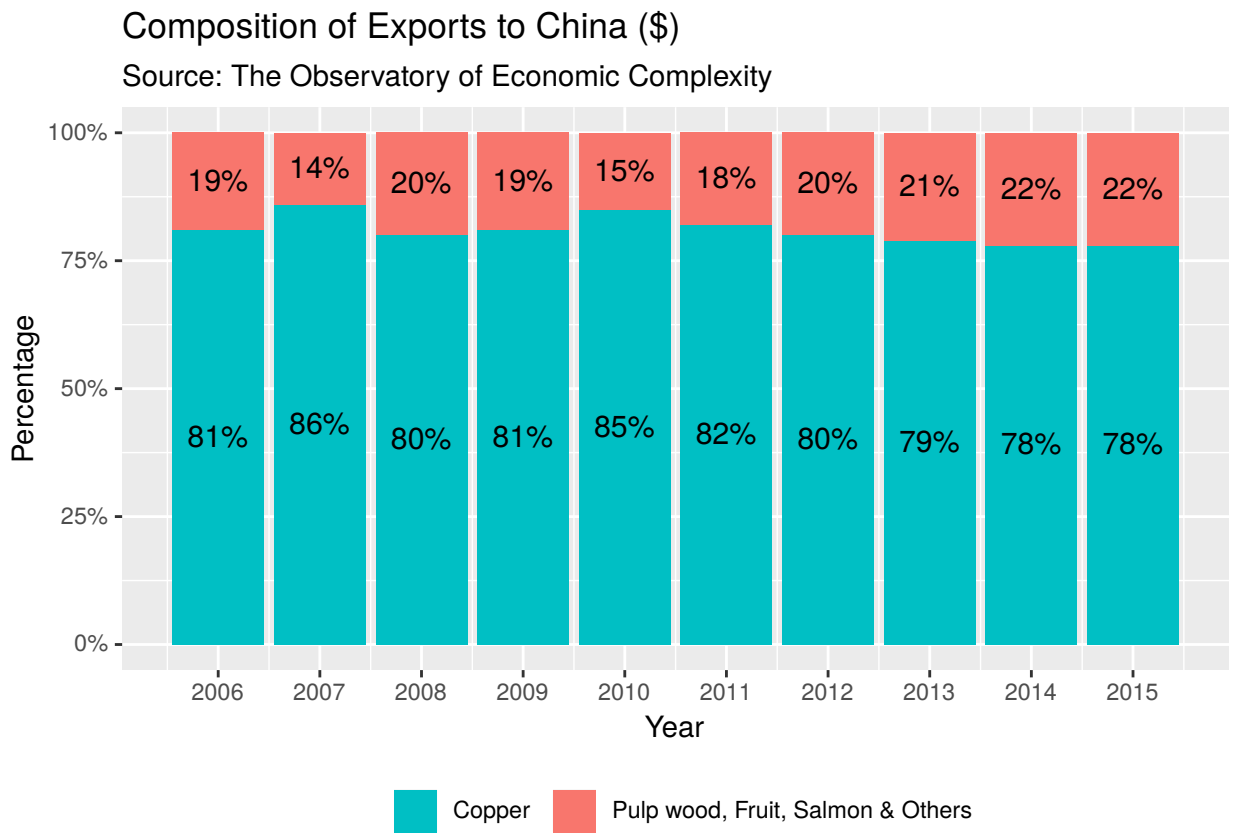
```
p4 <- p4 + scale_x_continuous(breaks = seq(2006,2015,1))
p4
```



4.8. Adjusting axis, title & units

To add a title, we include the option `labs` and include the name of the graph as a string argument, and to change the axis names we use the `labs` command too.

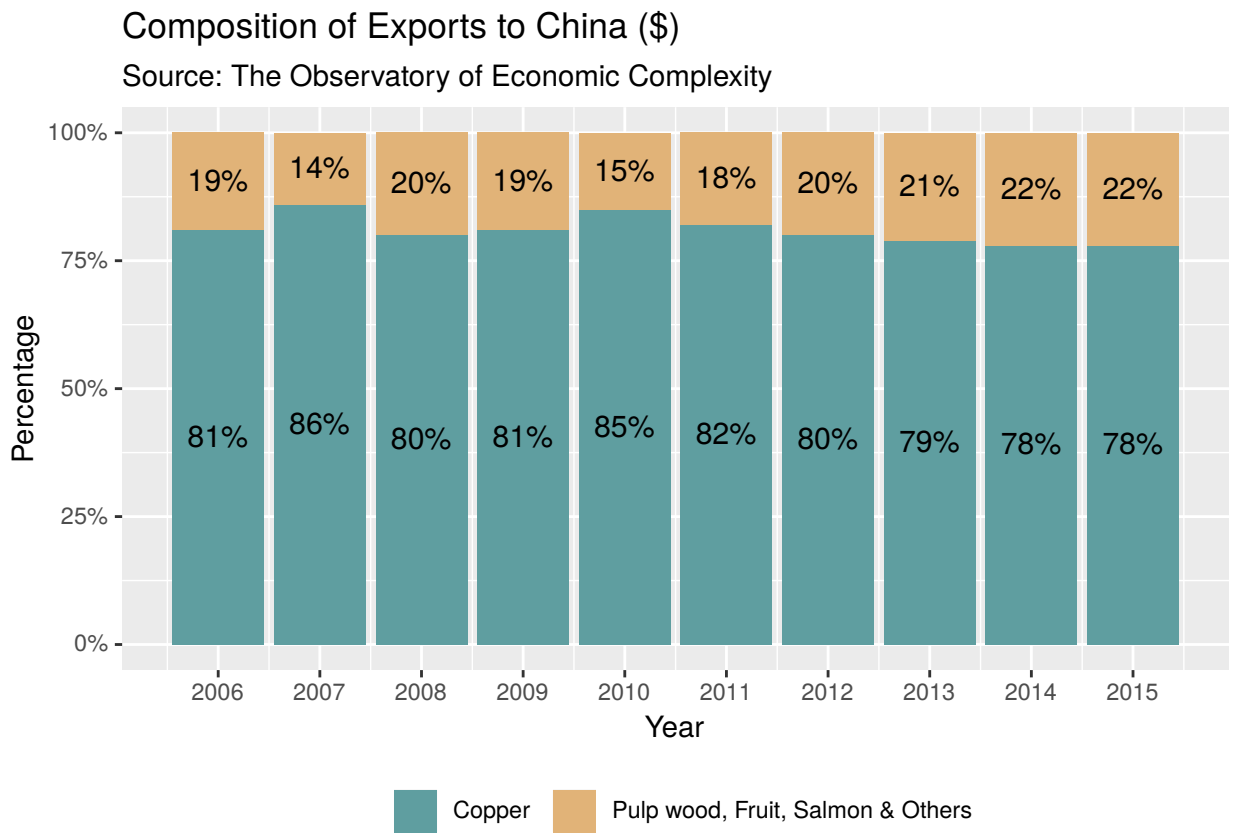
```
p4 <- p4 +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "Percentage") +
  scale_y_continuous(labels = dollar_format(suffix = "%", prefix = ""))
p4
```



4.9. Adjusting color palette

To change the colours, we use the `scale_colour_manual` command. Note that you can reference the specific colours you'd like to use with specific HEX codes. You can also reference colours by name, with the full list of colours recognised by R here.

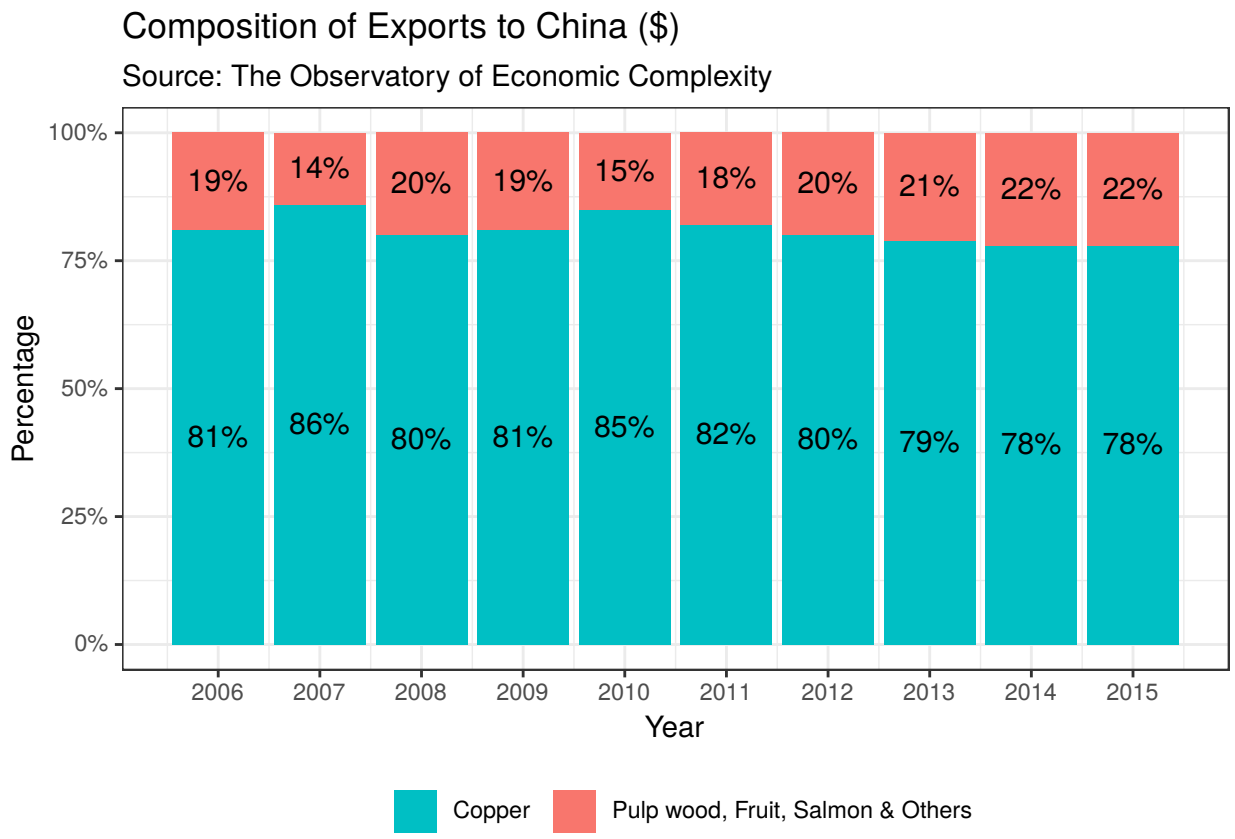
```
fill <- c("#E1B378","#5F9EA0")
p4 <- p4 + scale_fill_manual(values = fill)
p4
```



4.10. Using the white theme

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

```
p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col() +
  geom_text(aes(label = paste0(percentage,"%")),
            position = position_stack(vjust = 0.5), size = 4) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  scale_y_continuous(labels = dollar_format(suffix = "%", prefix = "")) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "Percentage") +
  theme_bw() +
  theme(legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank()) +
  guides(fill = guide_legend(reverse = T))
p4
```



4.11. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

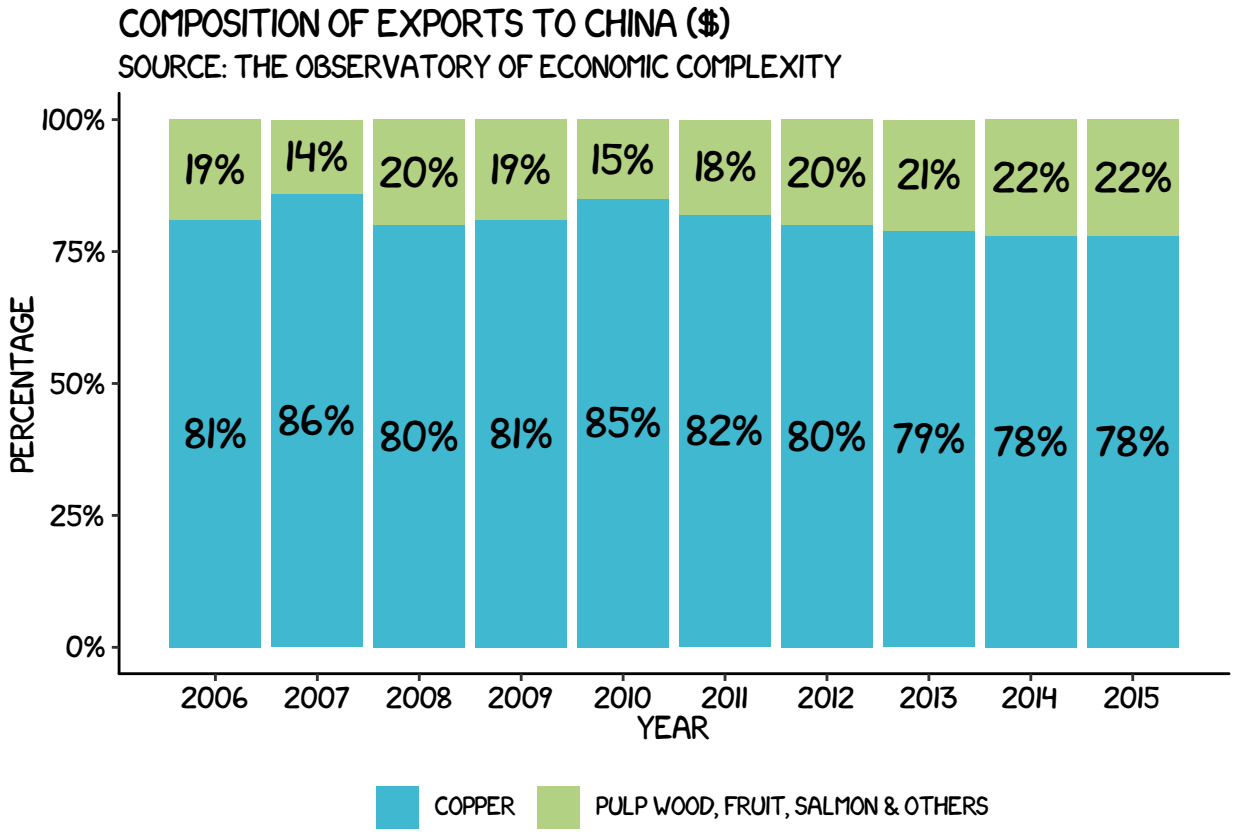
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

4.12. 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.

```
fill <- c("#b2d183","#40b8d0")

p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col() +
  geom_text(aes(label = paste0(percentage,"%")),
            position = position_stack(vjust = 0.5), colour = "black",
            family = "XKCD", size = 5, show.legend = F) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  scale_y_continuous(labels = dollar_format(suffix = "%", prefix = "")) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "Percentage") +
  scale_fill_manual(values = fill) +
  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.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD")) +
  guides(fill = guide_legend(reverse = T))
p4
```

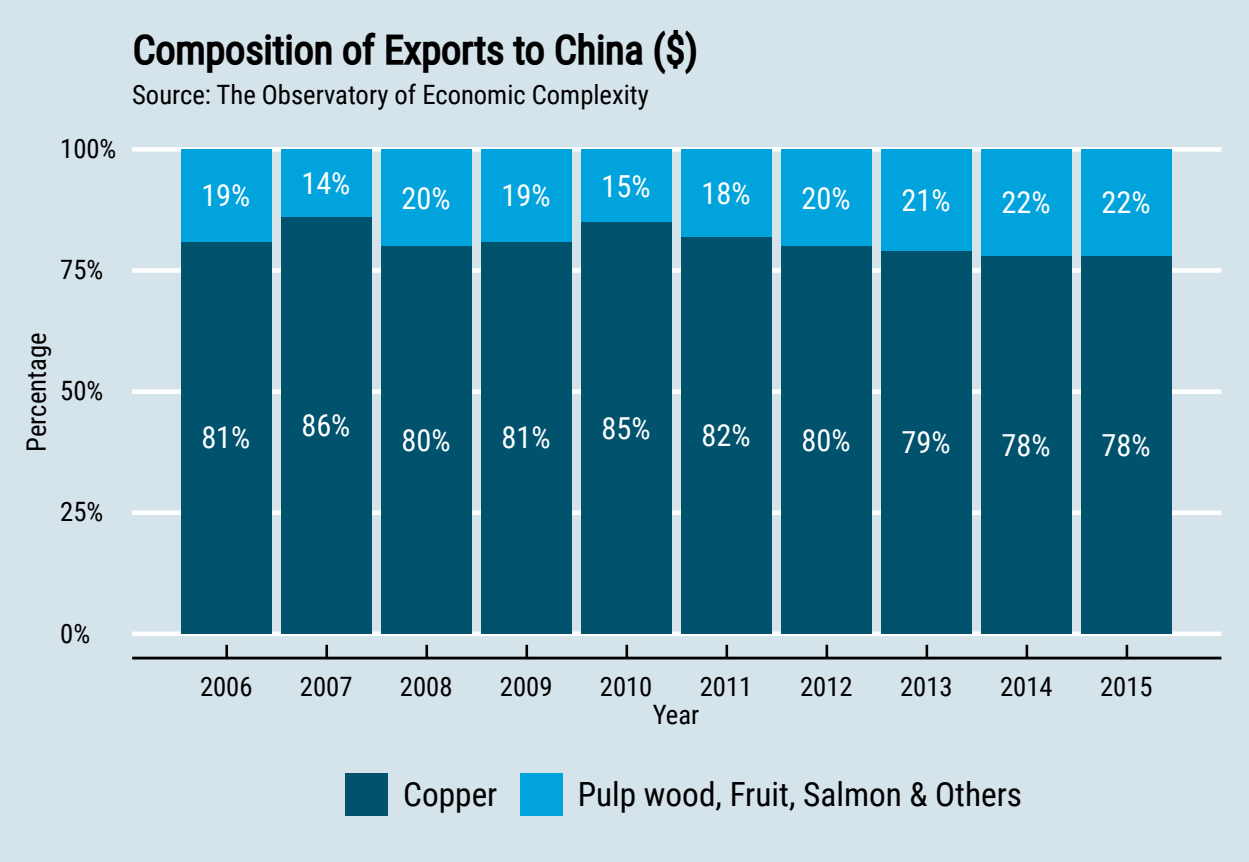


4.13. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
fill <- c("#00a3dc","#01526d")

p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col() +
  geom_text(aes(label = paste0(percentage,"%")),
            position = position_stack(vjust = 0.5), colour = "white",
            family = "Roboto Condensed", size = 4) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  scale_y_continuous(labels = dollar_format(suffix = "%", prefix = "")) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "Percentage") +
  theme_economist() + scale_fill_manual(values = fill) +
  theme(axis.line.x = element_line(size = .5, colour = "black"),
```

```
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_blank(),
    plot.title = element_text(family = "Roboto Condensed"),
    text = element_text(family = "Roboto Condensed")) +
  guides(fill = guide_legend(reverse = T))
p4
```

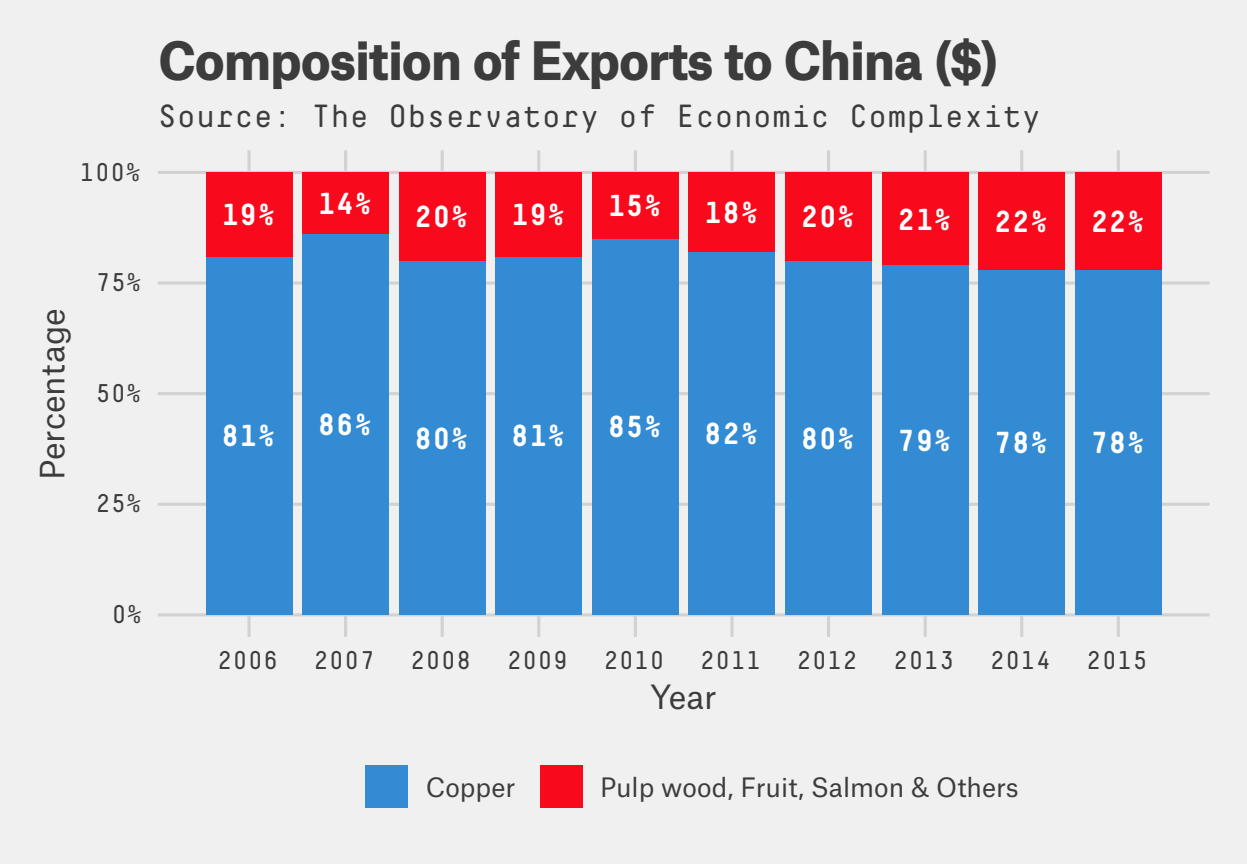


4.14. Using 'Five Thirty Eight' theme

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'.

```
fill <- c("#f80a1c","#338cd3")

p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col() +
  geom_text(aes(label = paste0(percentage,"%")),
            position = position_stack(vjust = 0.5), colour = "white",
```

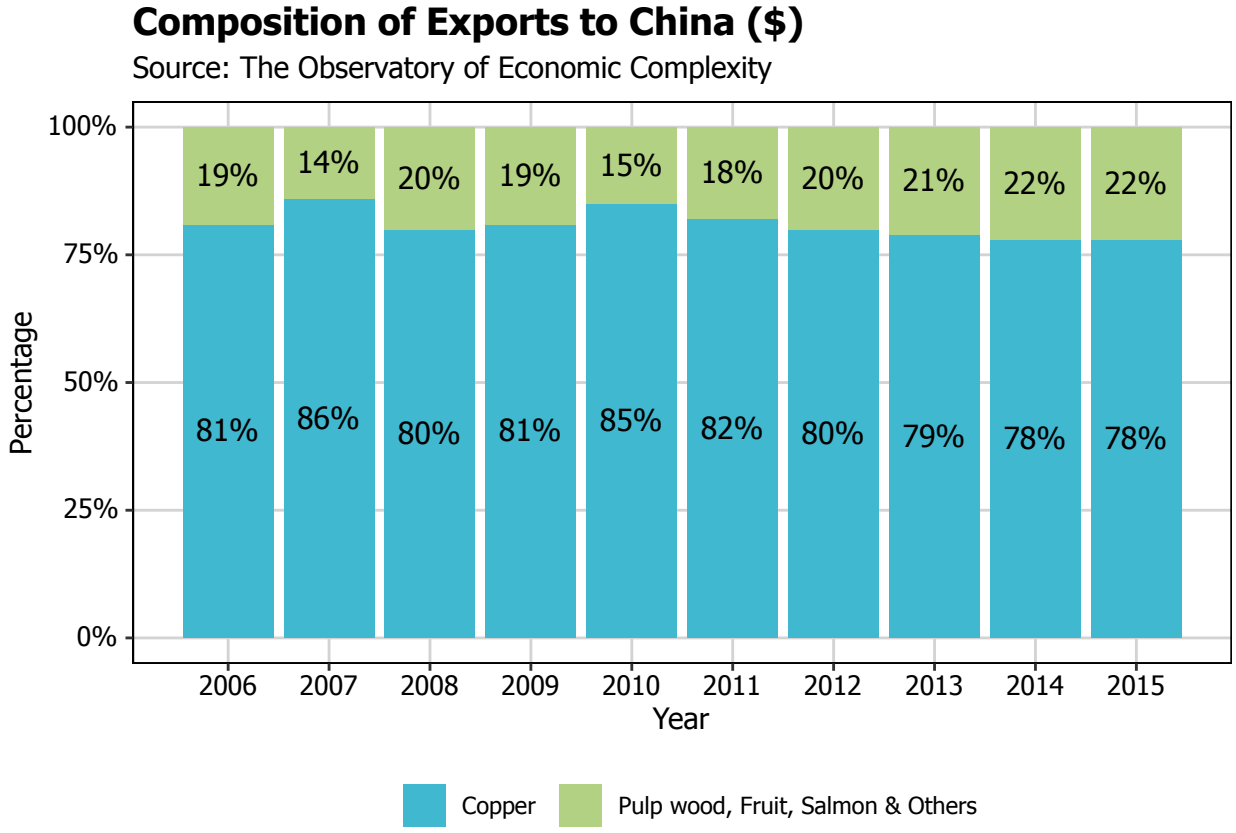
```
            family = "DecimaMonoPro-Bold", size = 4) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  scale_y_continuous(labels = dollar_format(suffix = "%", prefix = "")) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "Percentage") +
  theme_fivethirtyeight() + scale_fill_manual(values = fill) +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    legend.text = element_text(family = "Atlas Grotesk Regular"),
    text = element_text(family = "Decima Mono Pro")) +
  guides(fill = guide_legend(reverse = T))
p4
```



4.15. 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.

```r
fill <- c("#b2d183","#40b8d0")

p4 <- ggplot(aes(y = percentage, x = year, fill = fct_rev(product)),
             data = exports_data) +
  geom_col() +
  geom_text(aes(label = paste0(percentage,"%")),
            position = position_stack(vjust = 0.5), colour = "black",
            family = "Tahoma", size = 4) +
  scale_x_continuous(breaks = seq(2006,2015,1)) +
  scale_y_continuous(labels = dollar_format(suffix = "%", prefix = "")) +
  labs(title = "Composition of Exports to China ($)",
       subtitle = "Source: The Observatory of Economic Complexity") +
  labs(x = "Year", y = "Percentage") +
  scale_fill_manual(values = fill) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 10),
    axis.text.y = element_text(colour = "black", size = 10),
    legend.key = element_rect(fill = "white", colour = "white"),
    legend.position = "bottom", legend.direction = "horizontal",
    legend.title = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma")) +
  guides(fill = guide_legend(reverse = T))
p4
```

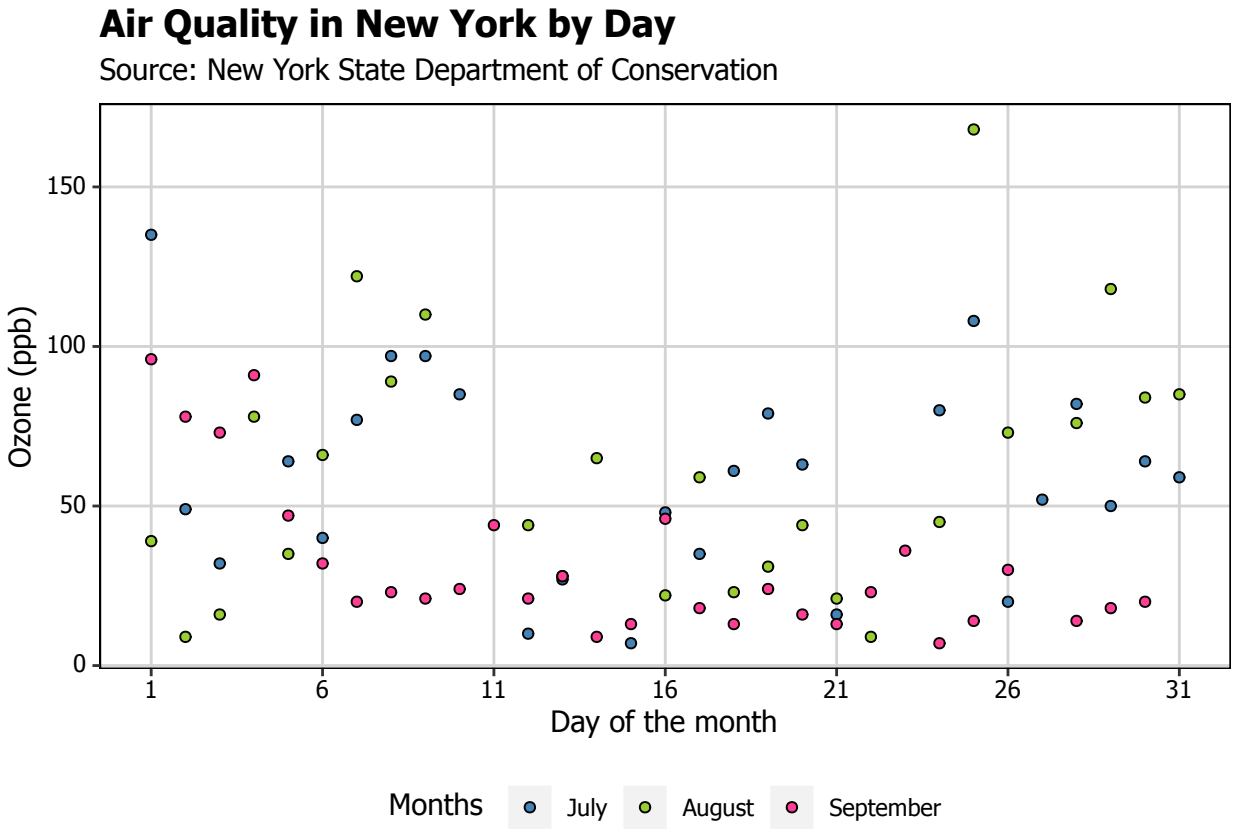

Composition of Exports to China ($)
Source: The Observatory of Economic Complexity
100%
75%
50%
25%
0%
Percentage
19%
14%
20%
19%
15%
18%
20%
21%
22%
22%
81%
86%
80%
81%
85%
82%
80%
79%
78%
78%
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Year
Copper
Pulp wood, Fruit, Salmon & Others

CHAPTER 5

Scatterplots

5.1. Introduction

In this chapter, we will work towards creating the scatterplot below. We will take you from a basic scatterplot 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:

```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, dplyr, scales, grid)

data(airquality)
```

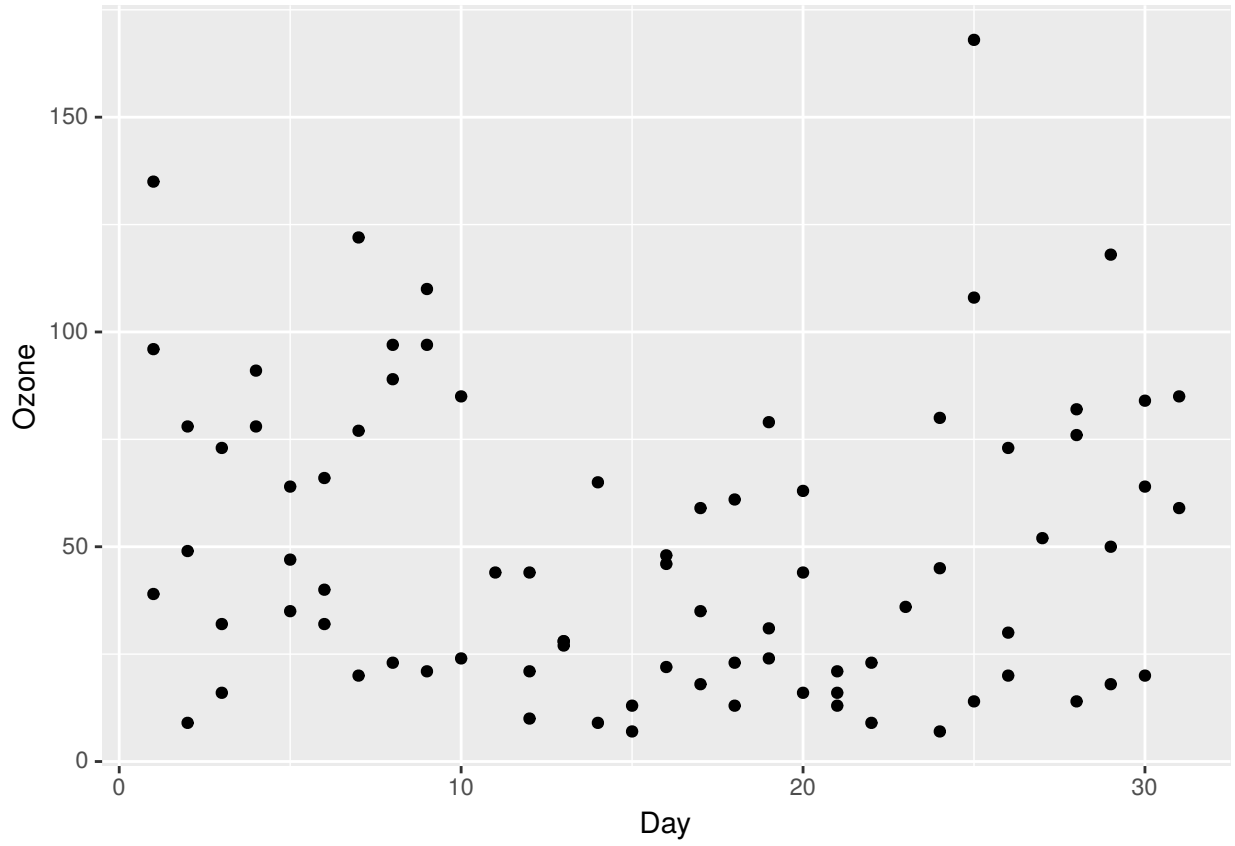
We will then trim the data down to the final three months and turn the `Month` variable into a labelled factor variable. We end up with a new dataset called `aq_trim`.

```
aq_trim <- airquality %>%
  filter(Month %in% c(7,8,9)) %>%
  mutate(Month = factor(Month,
                        labels = c("July", "August", "September")))
```

5.2. Basic scatterplot

In order to initialise a scatterplot we tell ggplot that `aq_trim` is our data, and specify that our x-axis plots the `Day` variable and our y-axis plots the `Ozone` variable. We then instruct ggplot to render this as a scatterplot by adding the `geom_point()` option.

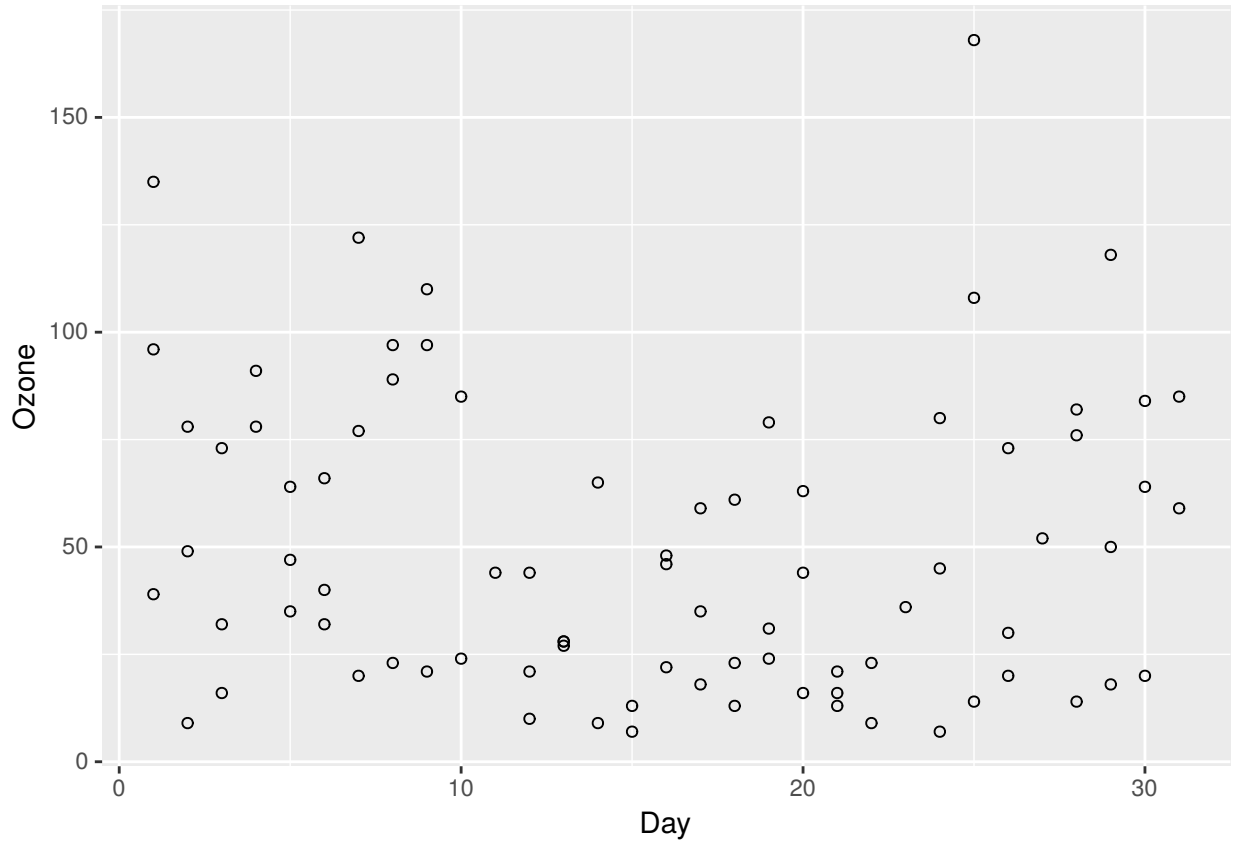
```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone)) +
  geom_point()
p5
```



5.3. Changing the shape of the data points

Perhaps we want the data points to be a different shape than a solid circle. We can change these by adding the `shape` argument to `geom_point`. An explanation of the allowed arguments for shape are described in this article. In this case, we will use shape 21, which is a circle that allows different colours for the outline and fill.

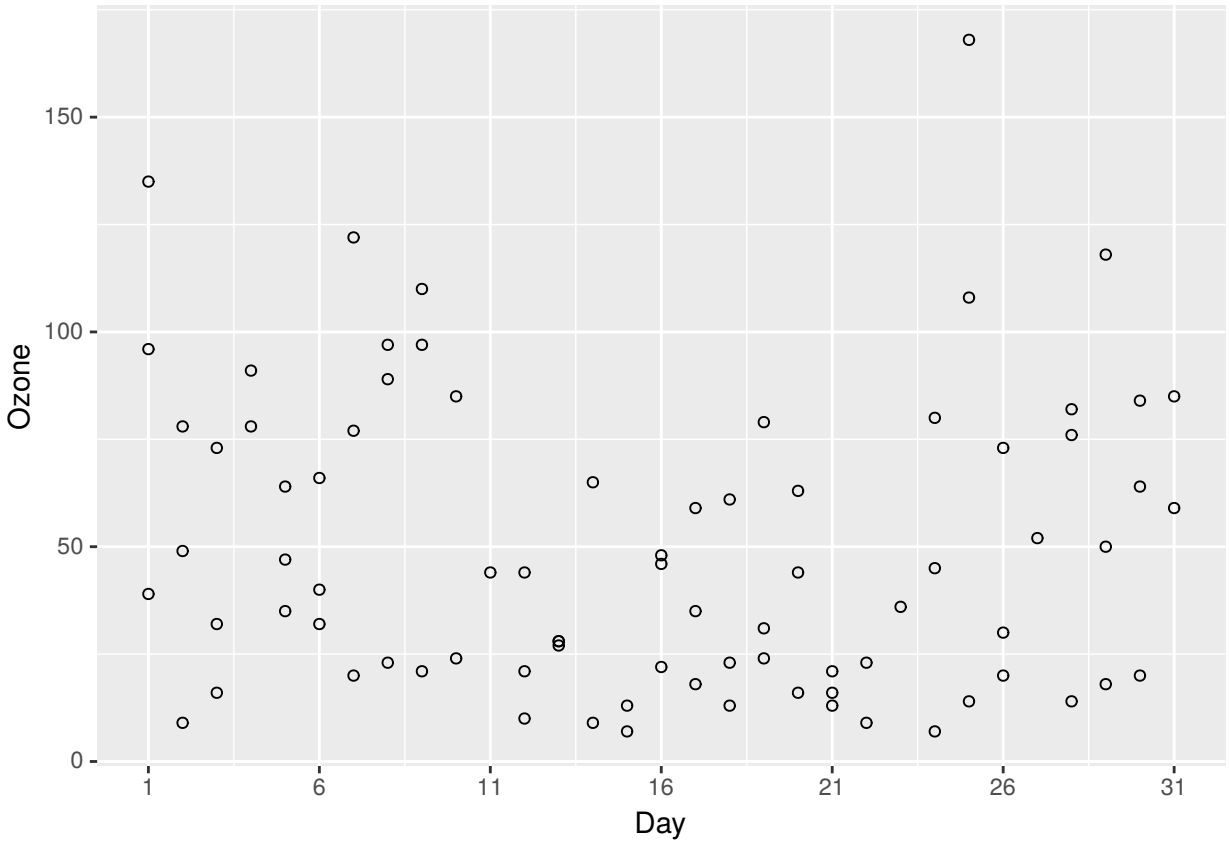
```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone)) + geom_point(shape = 21)
p5
```



5.4. Adjusting the axis scales

To change the x-axis tick marks, we use the `scale_x_continuous` option. Similarly, to change the y-axis we use the `scale_y_continuous` option. Here we will change the x-axis to every 5 days, rather than 10, and change the range from 1 to 31 (as 0 is not a valid value for this variable).

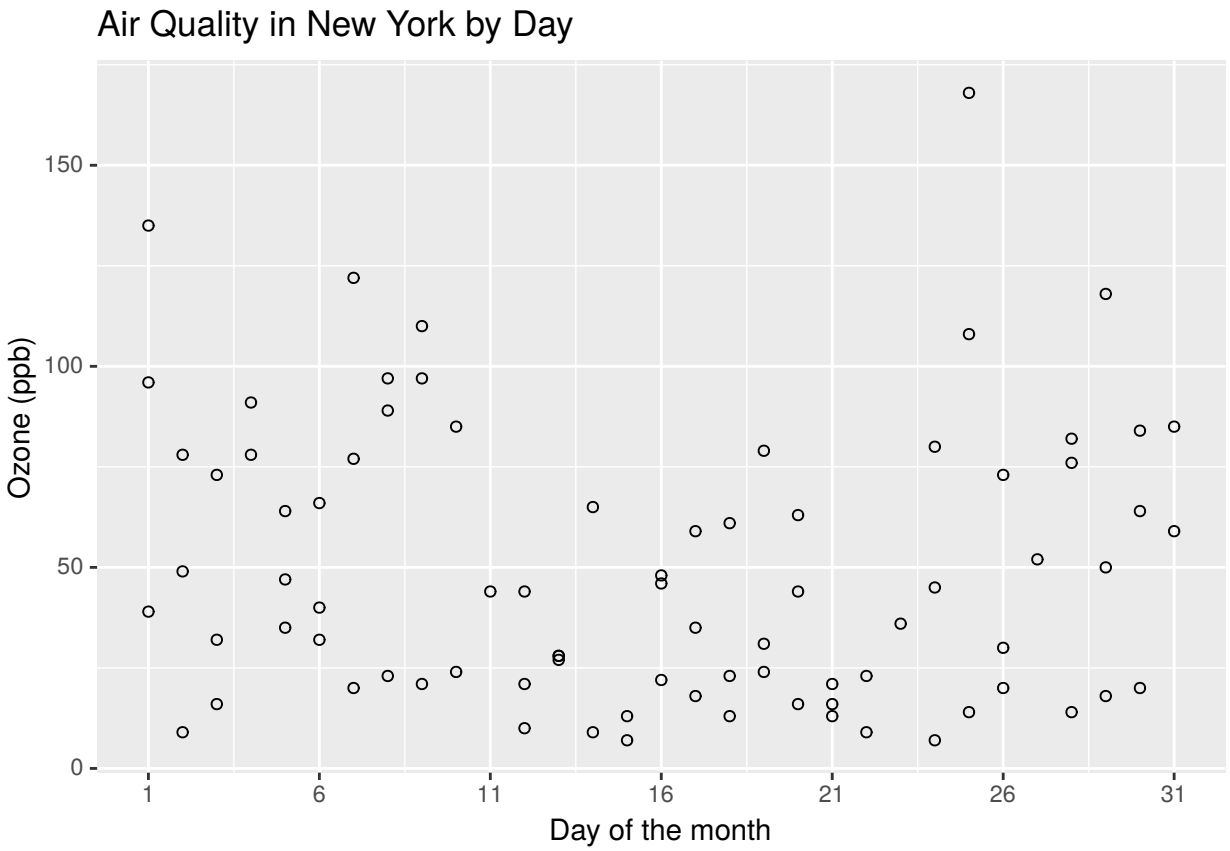
```
p5 <- p5 + scale_x_continuous(breaks = seq(1, 31, 5))
p5
```



5.5. Adjusting axis labels & adding title

To add a title, we include the option `labs` and include the name of the graph as a string argument. To change the axis names we add `x` and `y` arguments to the `labs` command.

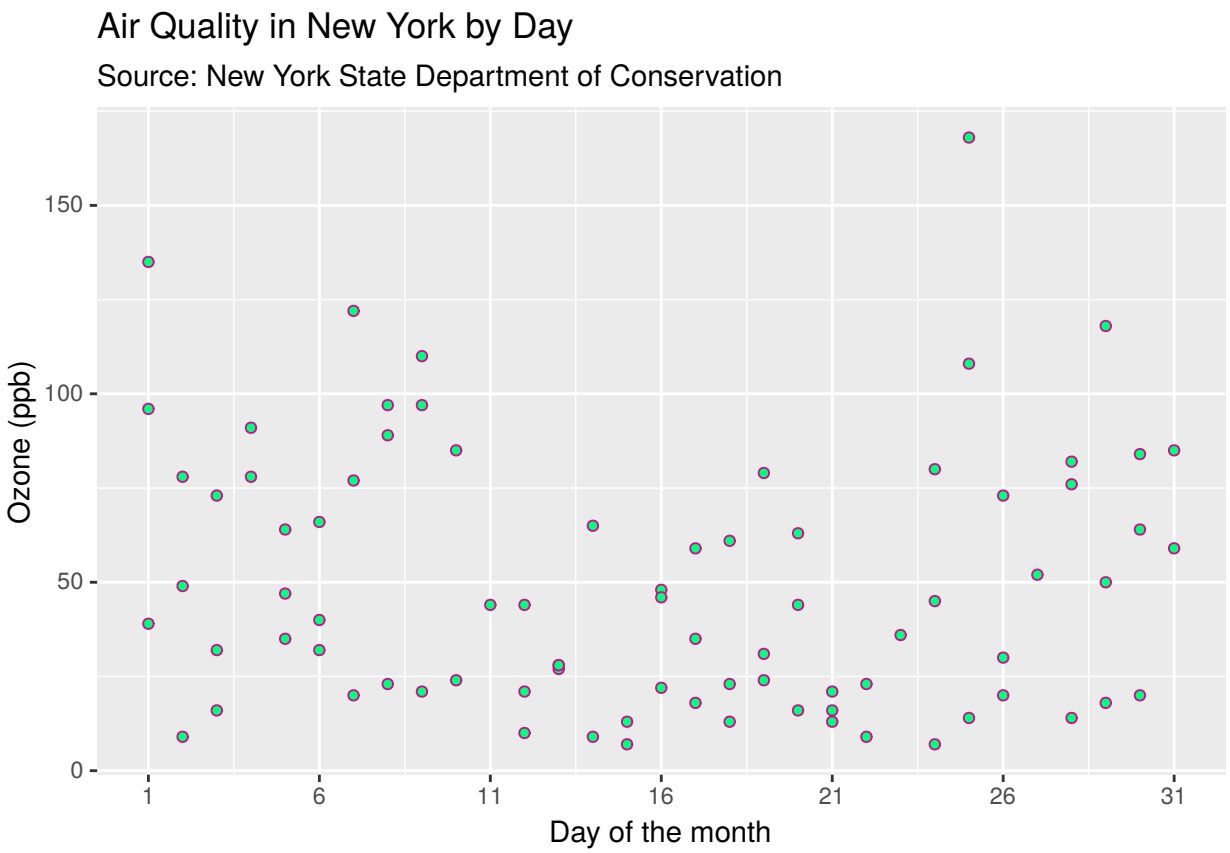
```
p5 <- p5 + labs(title = "Air Quality in New York by Day") +
  labs(x = "Day of the month", y = "Ozone (ppb)")
p5
```



5.6. Adjusting the colour palette

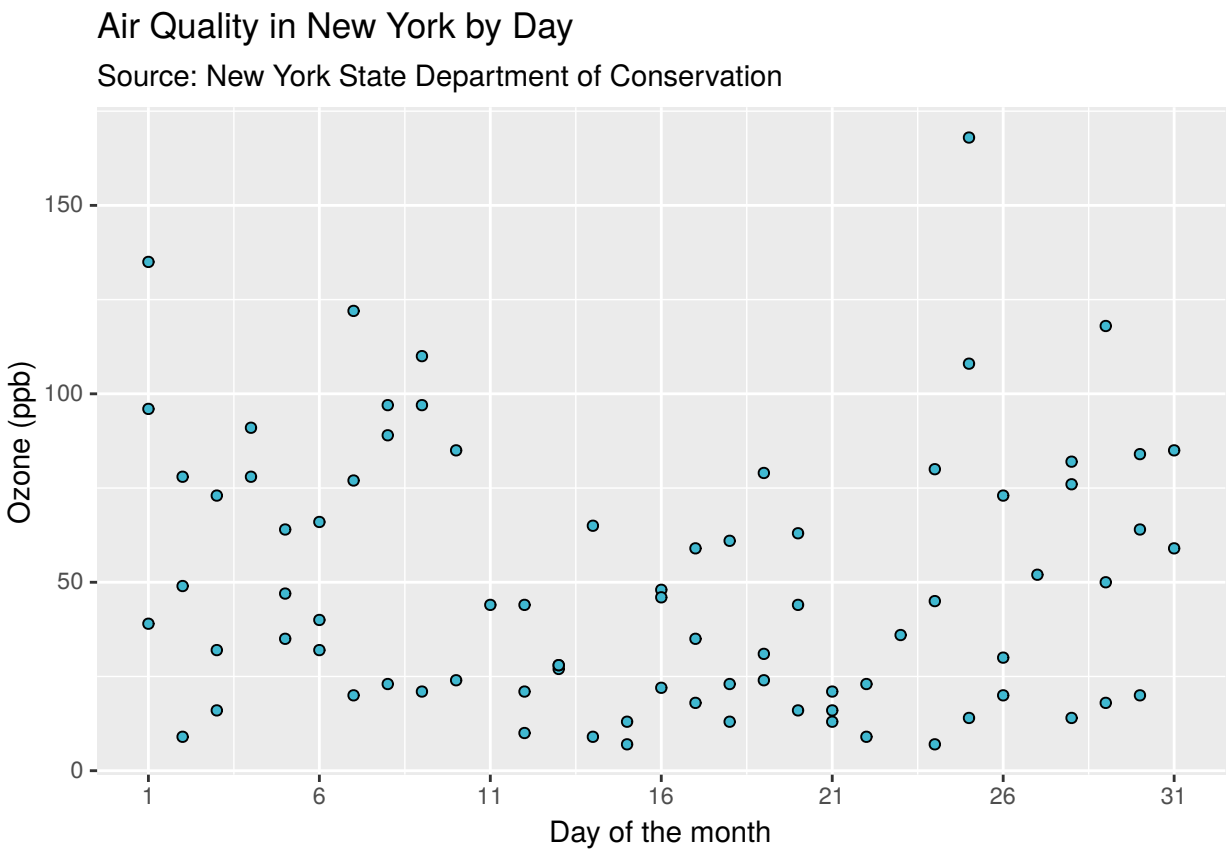
There are a few options for adjusting the colour. The most simple is to make every point one fixed colour. You can reference colours by name, with the full list of colours recognised by R here. Let's try making the outline `mediumvioletred` and the fill `springgreen`.

```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone)) +
  geom_point(shape = 21, colour = "mediumvioletred", fill = "springgreen") +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p5
```



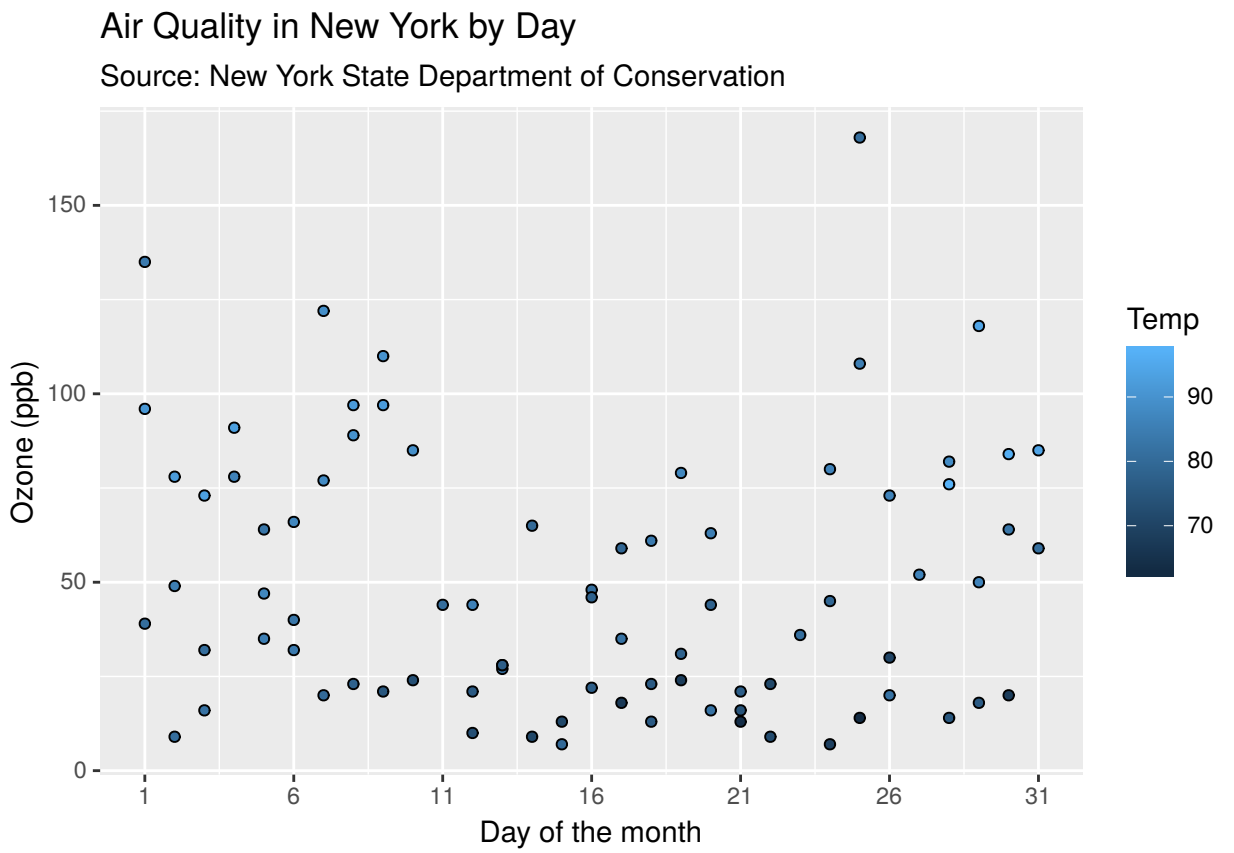
You can change the colours using specific HEX codes instead. Here we have made the outline #000000 (black) and the fill "#40b8d0 (vivid cyan).

```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone)) +
  geom_point(shape = 21, colour = "#000000", fill = "#40b8d0") +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p5
```



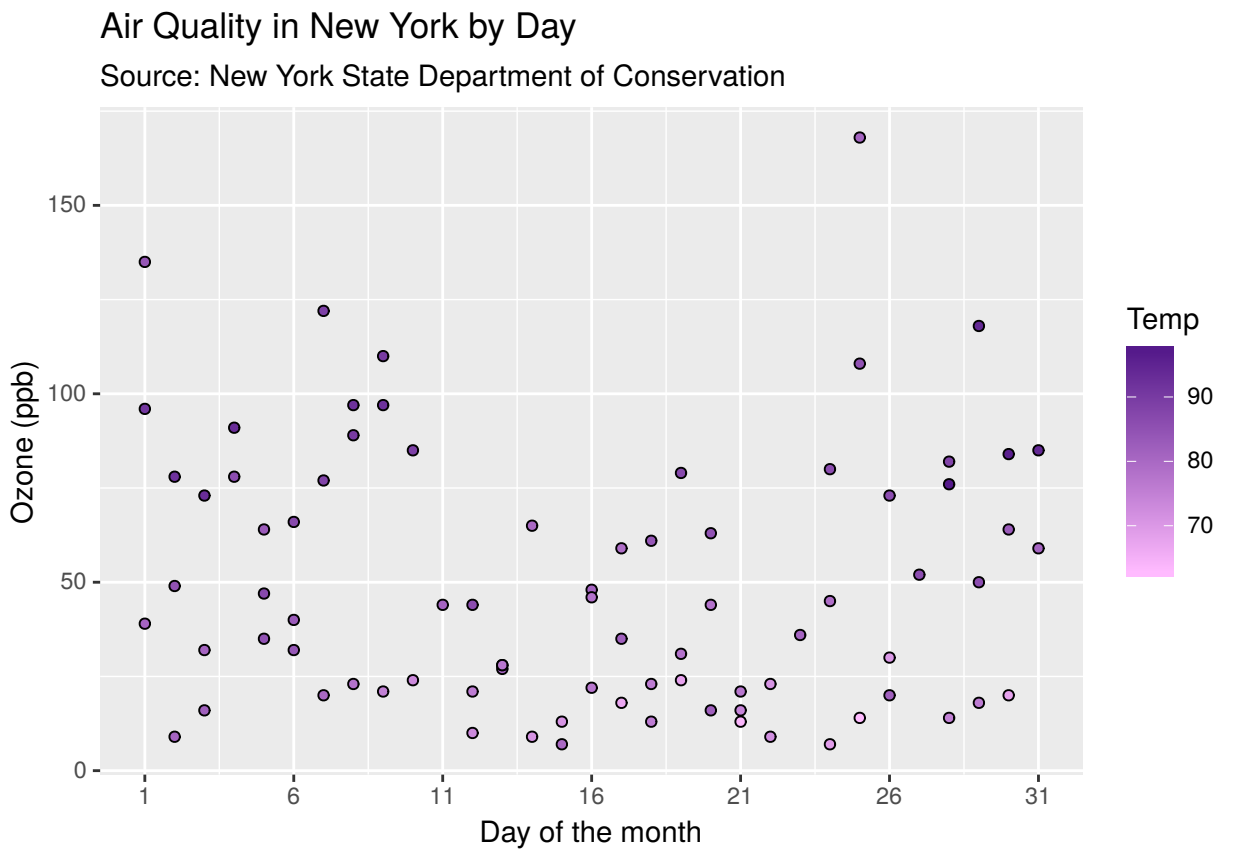
You can also change the colour of the data points according to the levels of another variable. This can be done either as a continuous gradient, or as a levels of a factor variable. Let's change the colour by the values of temperature:

```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone, fill = Temp)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p5
```



We can change the gradient's colours by adding the `scale_fill_continuous` option. The `low` and `high` arguments specify the range of colours the gradient should transition between.

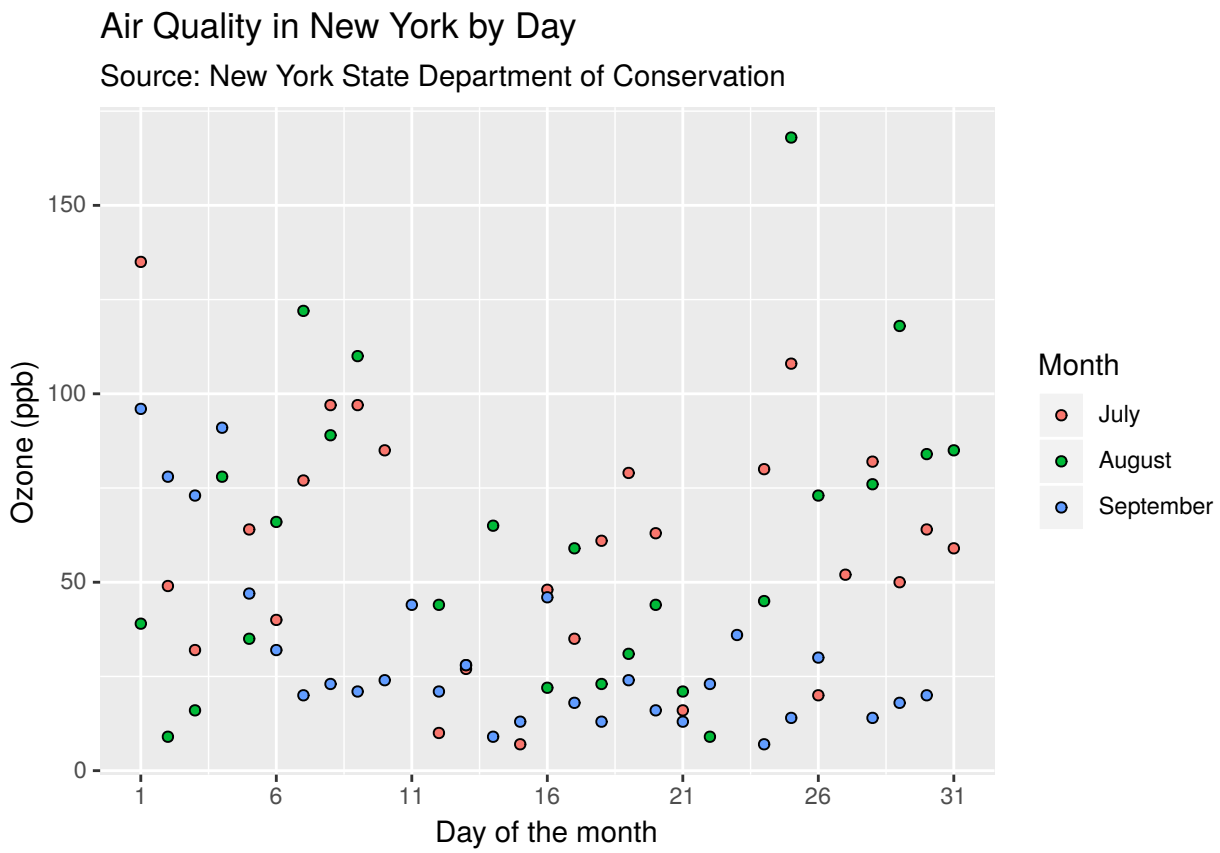
```
p5 <-  p5 + scale_fill_continuous(low = "plum1", high = "purple4")
p5
```



We can see that higher temperatures seem to have higher ozone levels.

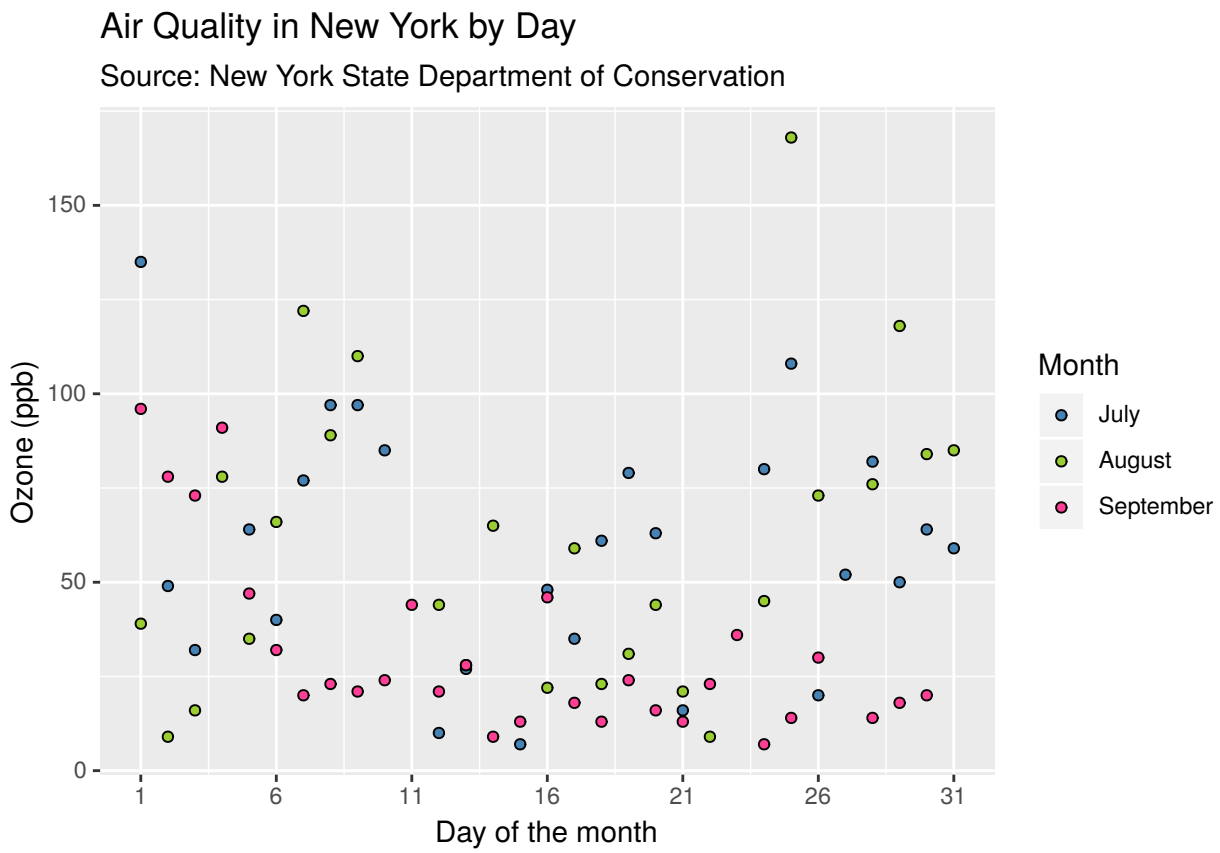
Let's now change the colours of the data points by a factor variable, `Month`.

```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p5
```



Again, we can change the colours of these data points, this time using `scale_fill_manual`.

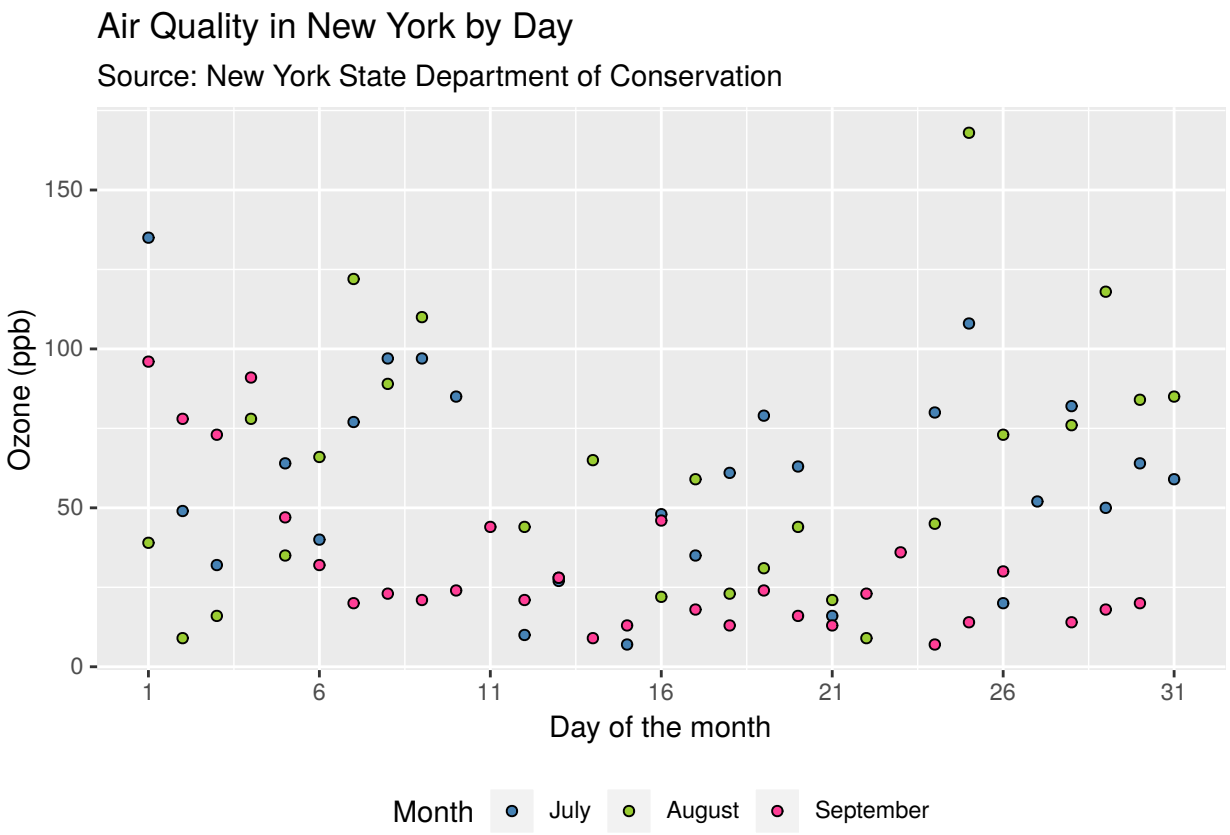
```
fill <- c("steelblue", "yellowgreen", "violetred1")

p5 <- p5 + scale_fill_manual(values = fill)
p5
```



5.7. Adjusting legend position

To adjust the position of the legend from the default spot of right of the graph, we add the `theme` option and specify the `legend.position = "bottom"` argument. We can also change the legend shape using the `legend.direction = "horizontal"` argument.

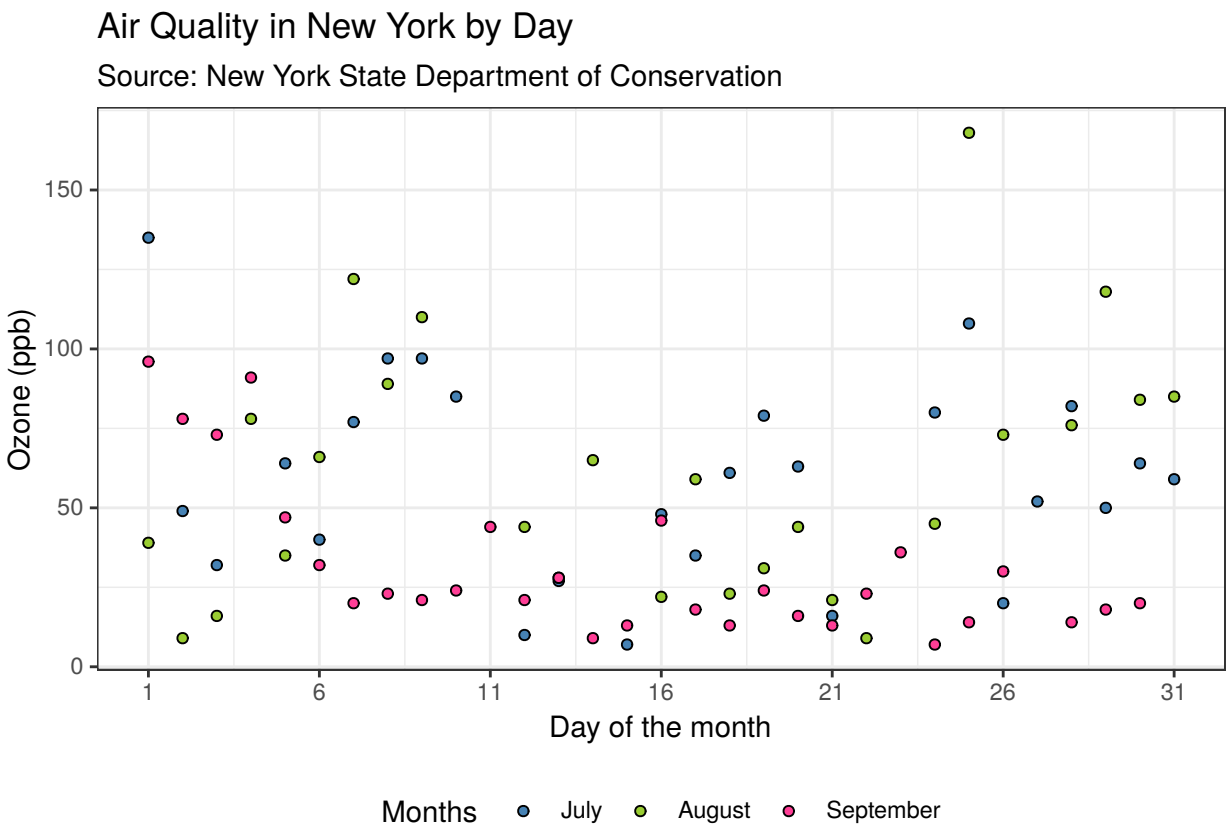
```
p5 <- p5 + theme(legend.position = "bottom", legend.direction = "horizontal")
p5
```



5.8. 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()` after `ggplot()`. As you can see, we can further tweak the graph using the `theme` option, which we've used so far to change the legend.

```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)", fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_fill_manual(values = fill) +
  scale_size(range = c(1, 10)) +
  theme_bw() +
  theme(legend.position = "bottom", legend.direction = "horizontal")
p5
```



5.9. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

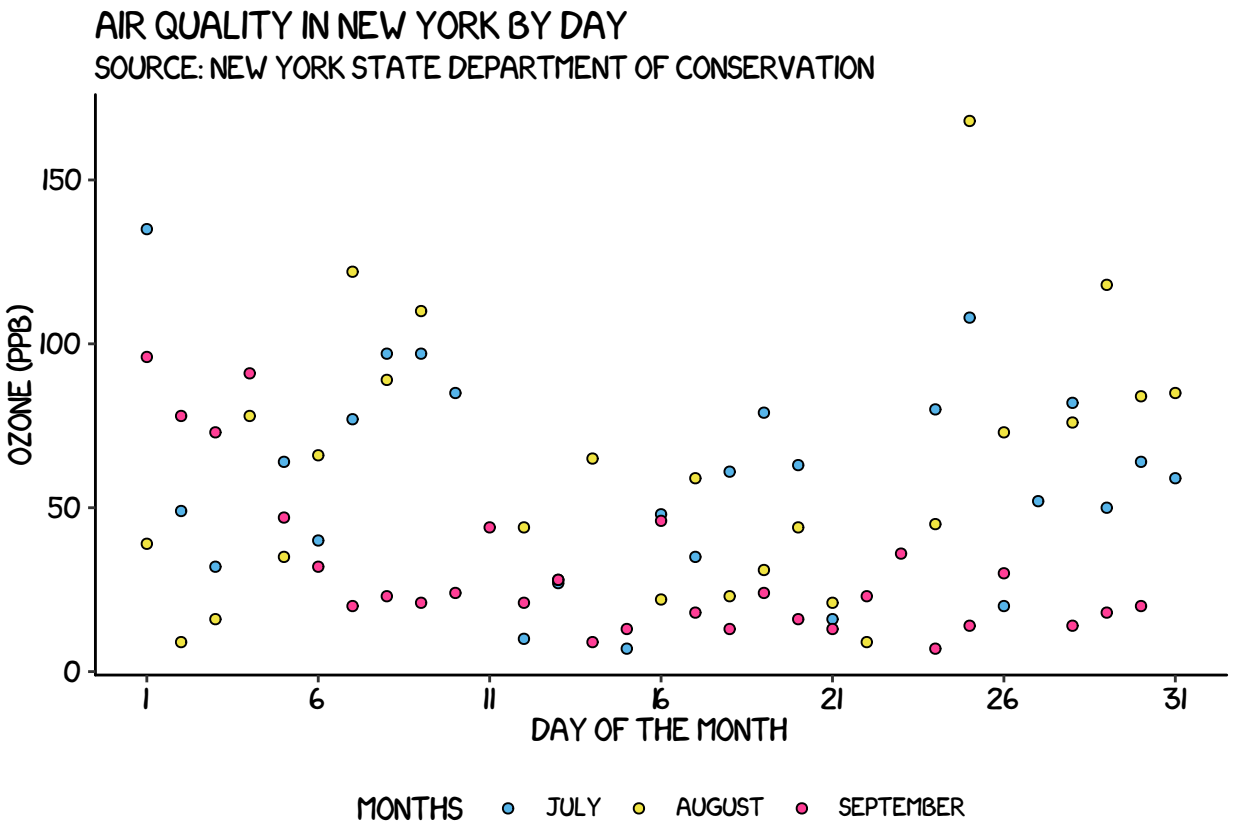
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

5.10. 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.

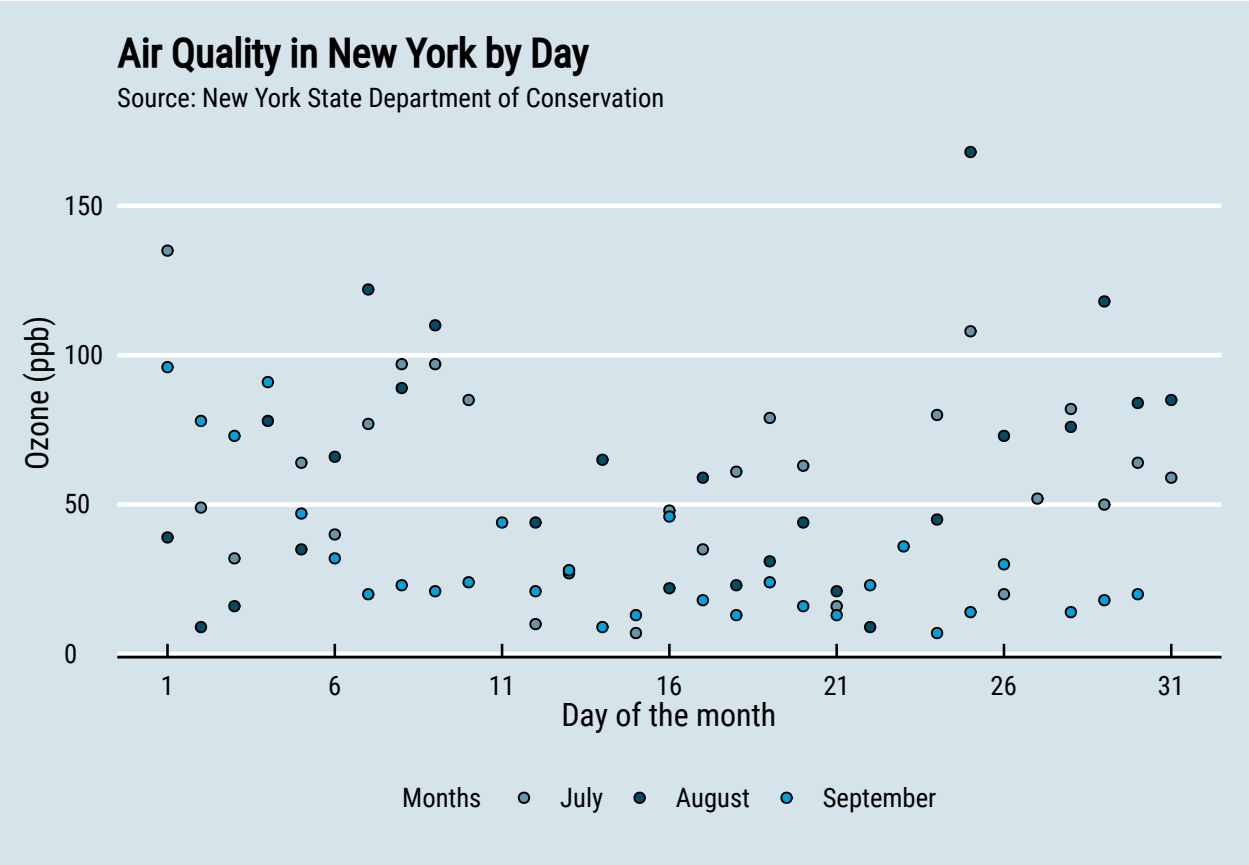
```
fill <- c("#56B4E9","#F0E442","violetred1")

p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)", fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_fill_manual(values = fill) +
  scale_size(range = c(1, 10)) +
  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.key = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p5
```



5.11. 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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)", fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_size(range = c(1, 10)) +
  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.text = element_text(size = 10),
    plot.title = element_text(family = "Roboto Condensed"),
    text = element_text(family = "Roboto Condensed"))
p5
```

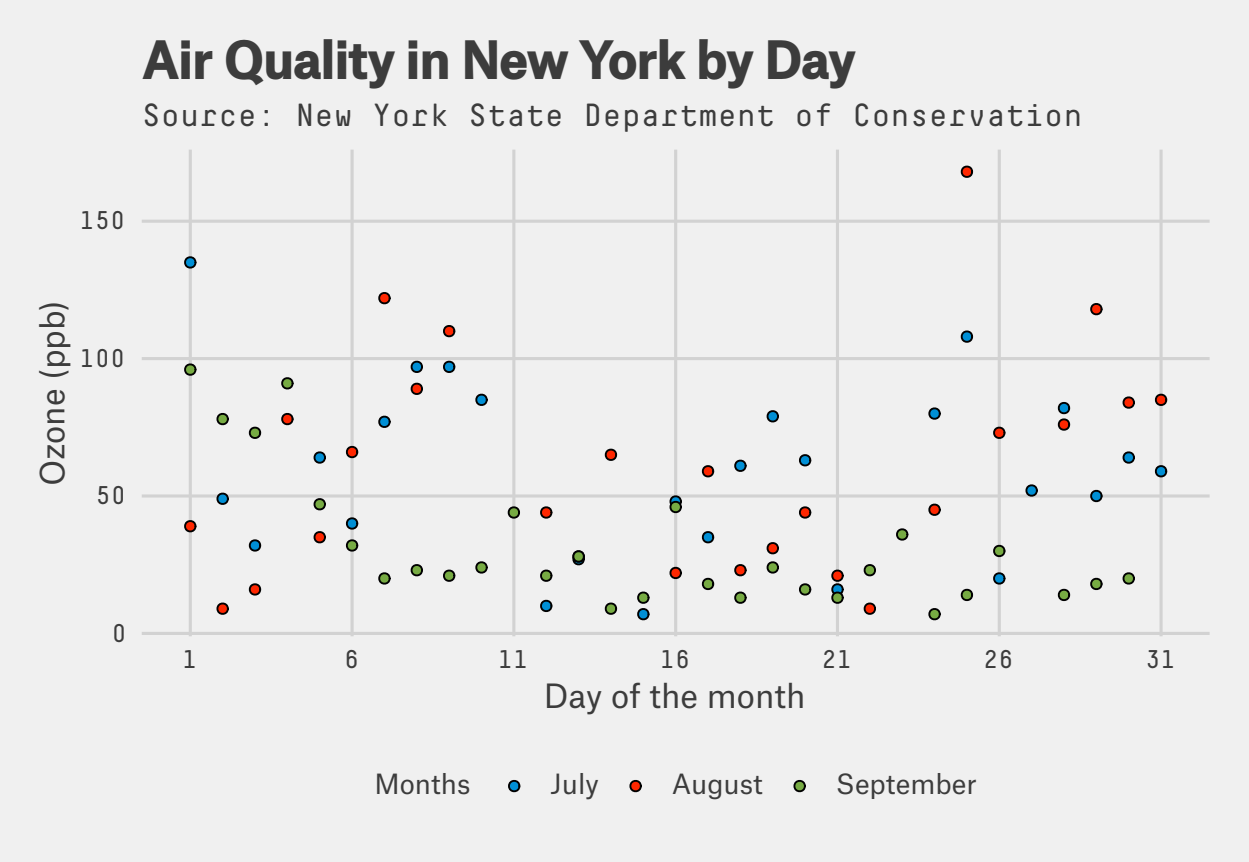


5.12. Using 'Five Thirty Eight' theme

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'.

```
p4 <- ggplot(aq_trim, aes(x = Day, y = Ozone, fill = Month)) +
  scale_fill_economist() +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)", fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_size(range = c(1, 10)) +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 10),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 10),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
```

```
    text = element_text(family = "Decima Mono Pro"))
p4
```

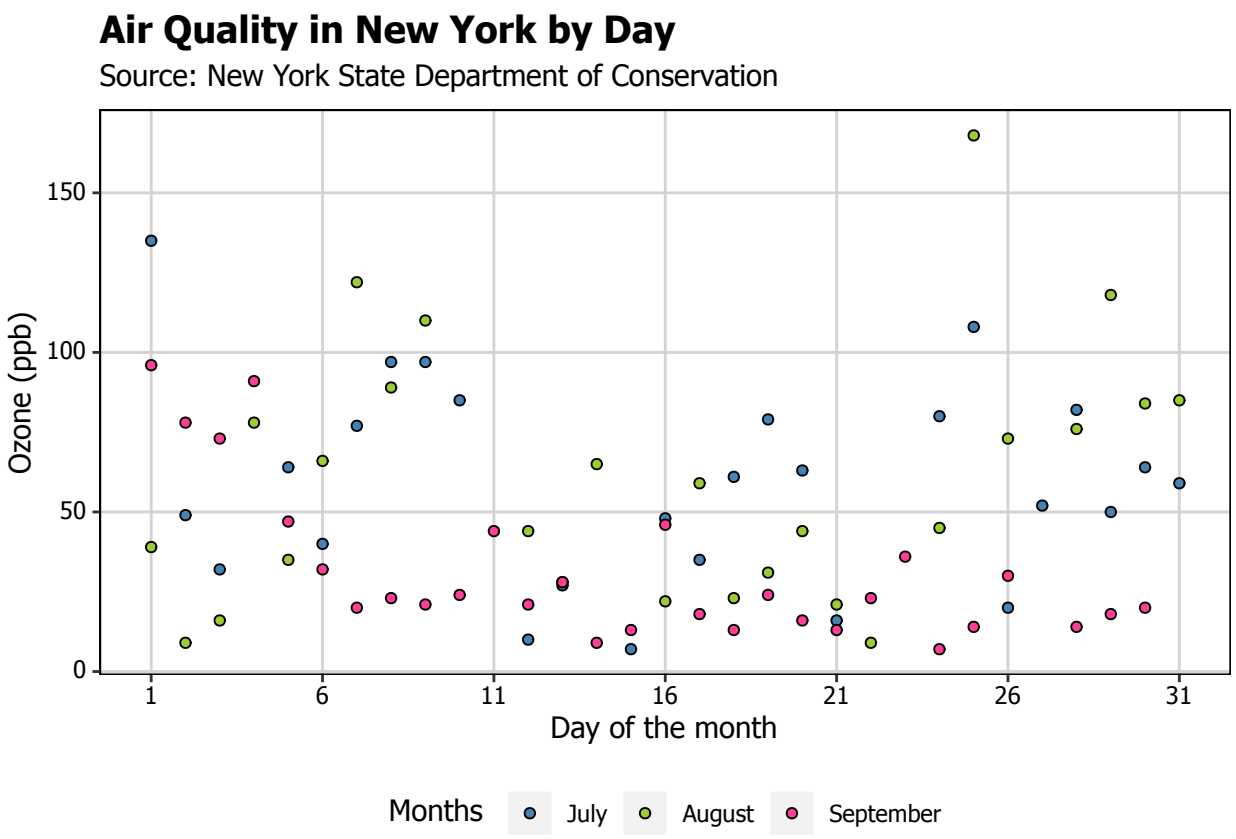


5.13. Creating your own theme

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 <- c("steelblue", "yellowgreen", "violetred1")

p5 <- ggplot(aq_trim, aes(x = Day, y = Ozone, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)", fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_size(range = c(1, 10)) +
  scale_fill_manual(values = fill) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
```

```
    legend.position = "bottom", legend.direction = "horizontal",
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p5
```

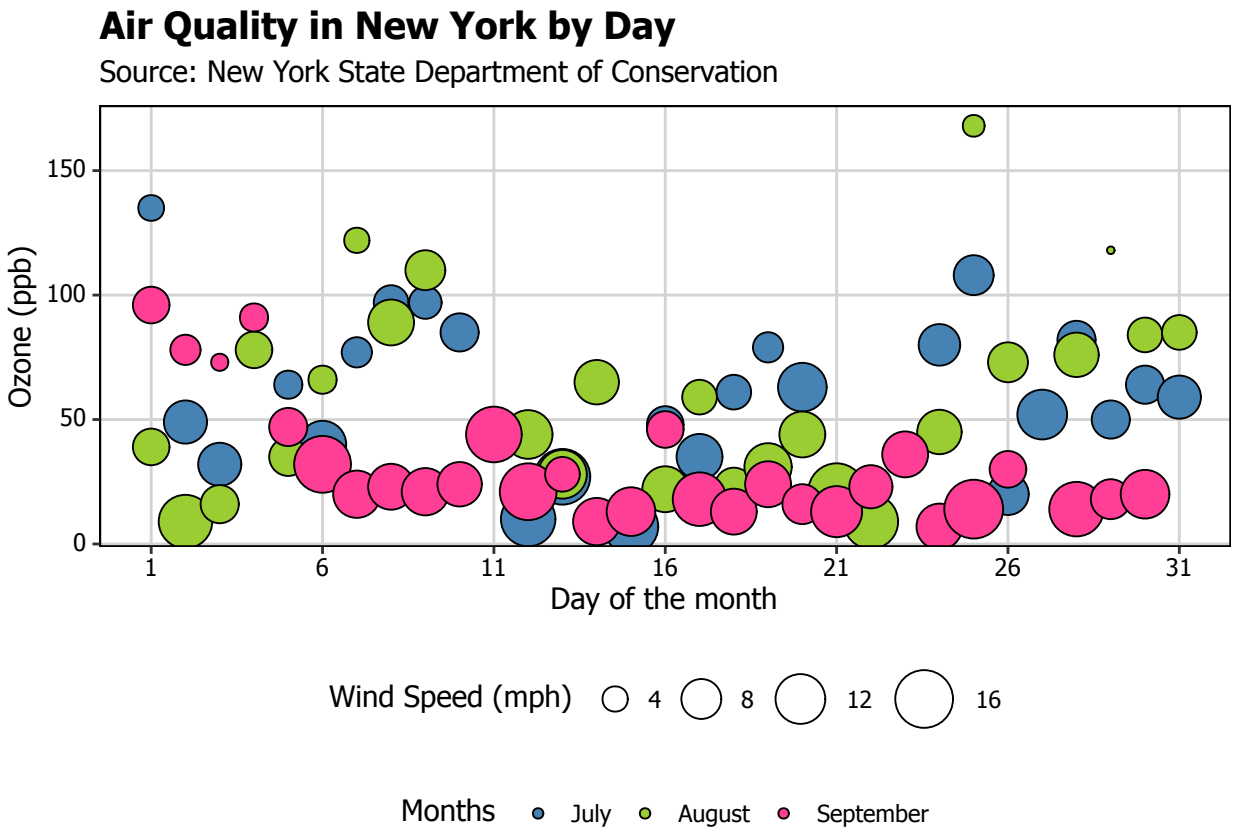


CHAPTER 6

Weighted scatterplots

6.1. Introduction

In this chapter, we will work towards creating the weighted scatterplot below. We will take you from a basic scatterplot 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:

```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, dplyr, scales, grid)

data(airquality)
```

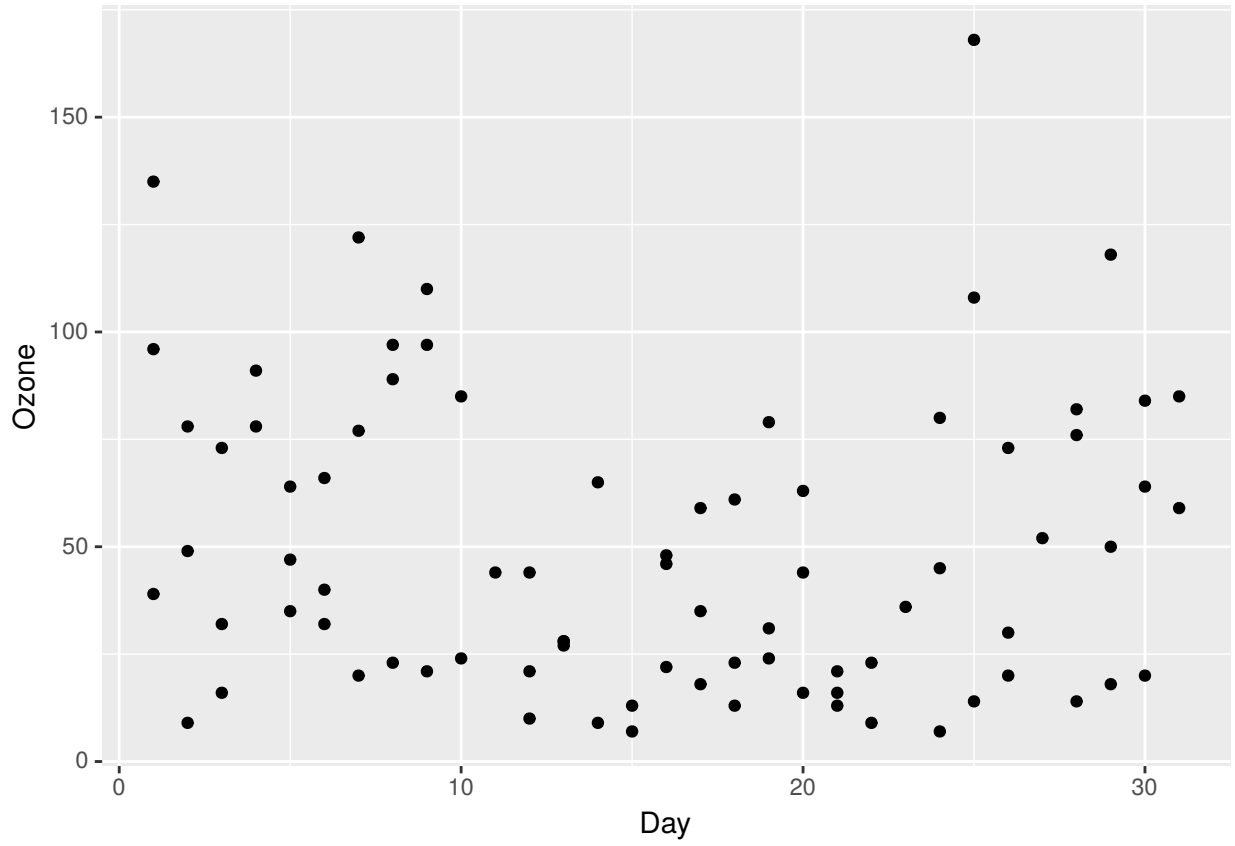
We will then trim the data down to the final three months and turn the `Month` variable into a labelled factor variable. We end up with a new dataset called `aq_trim`.

```
aq_trim <- airquality %>%
  filter(Month %in% c(7,8,9)) %>%
  mutate(Month = factor(Month,
                        labels = c("July", "August", "September")))
```

6.2. Basic weighted scatterplot

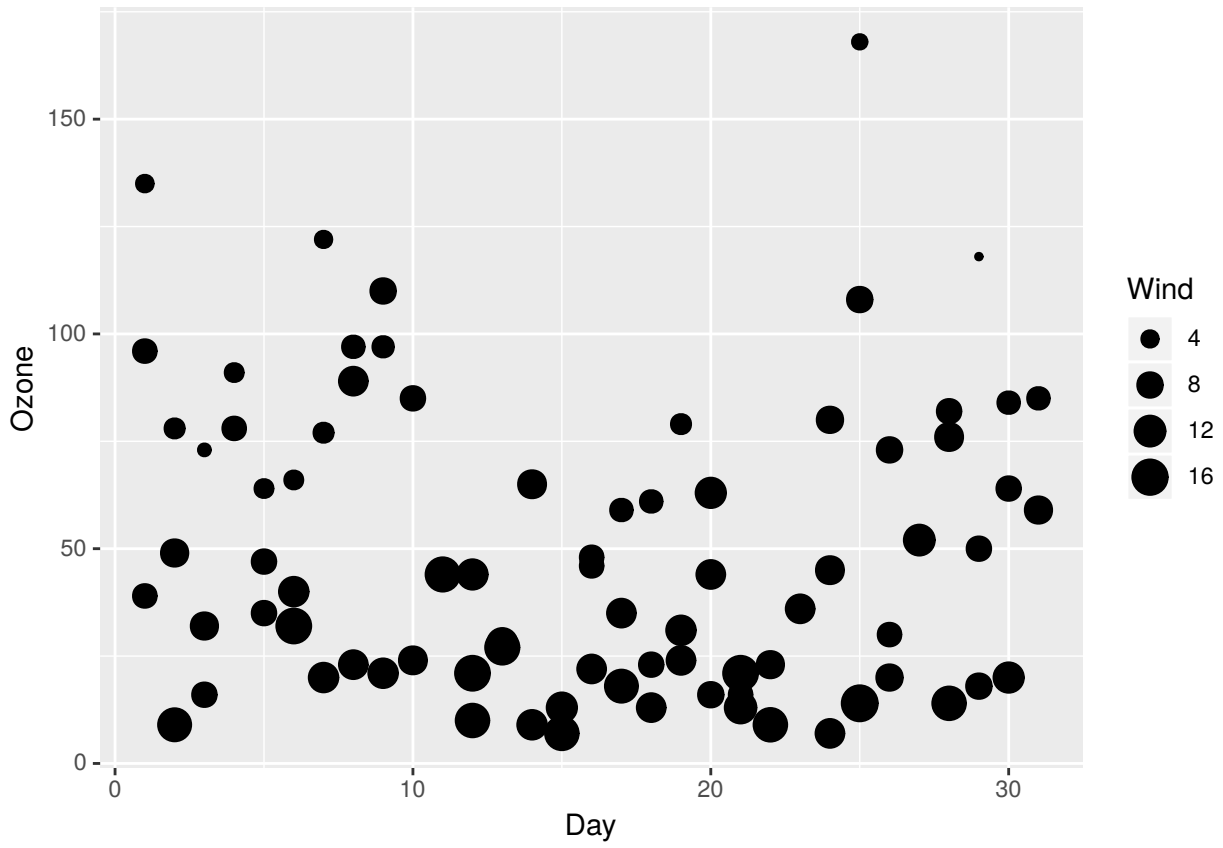
Let's start really slowly by revisiting how to create a basic scatterplot. In order to initialise this plot we tell ggplot that `aq_trim` is our data, and specify that our x-axis plots the `Day` variable and our y-axis plots the `Ozone` variable. We then instruct ggplot to render this as a scatterplot by adding the `geom_point()` option.

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone)) +
      geom_point()
p6
```



In order to turn this into a weighted scatterplot, we simply add the `size` argument to `ggplot(aes())`. In this case, we want to weight the points by the `Wind` variable.

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind)) +
  geom_point()
p6
```

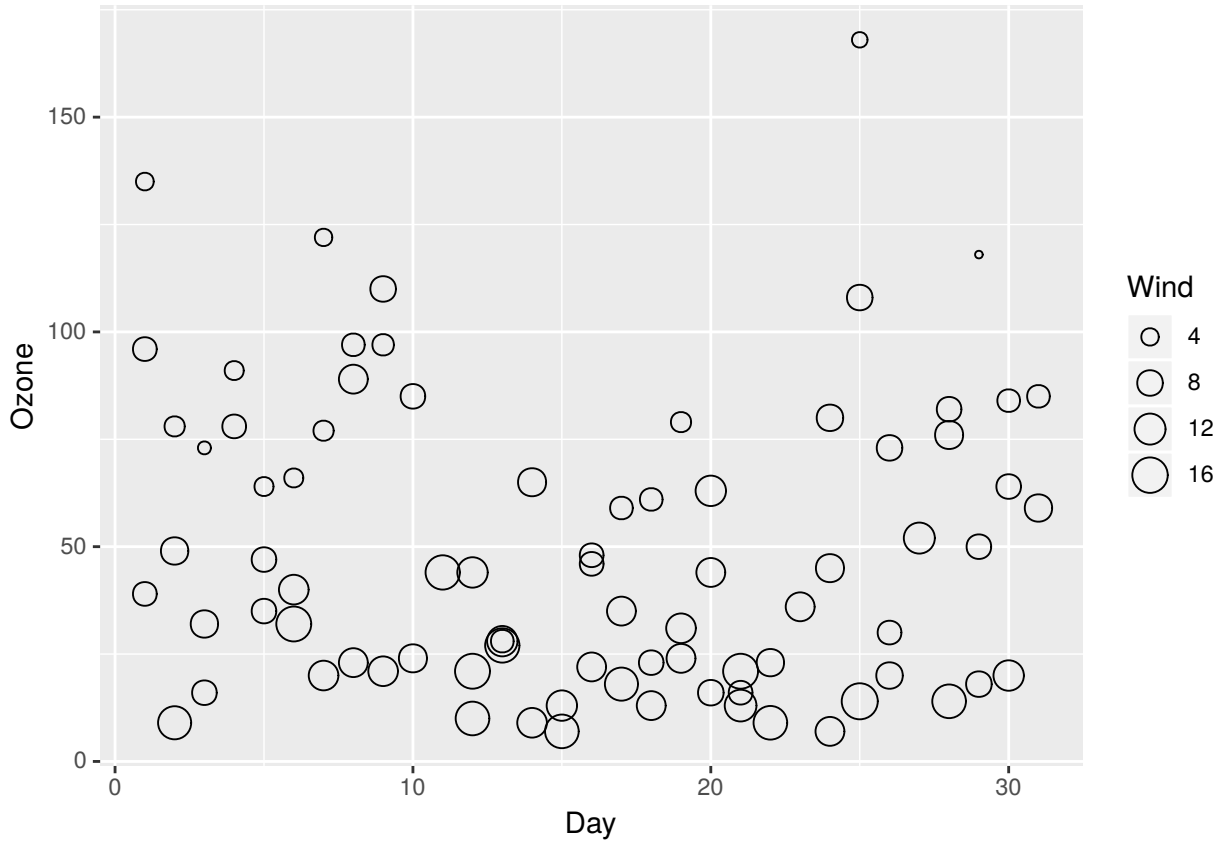


You can see we already have an interesting looking pattern, where days with higher wind speed tend to have lower ozone (or in other words, better air quality). Now let's make it beautiful!

6.3. Changing the shape of the data points

Perhaps we want the data points to be a different shape than a solid circle. We can change these by adding the `shape` argument to `geom_point`. An explanation of the allowed arguments for shape are described in this article. In this case, we will use shape 21, which is a circle that allows different colours for the outline and fill.

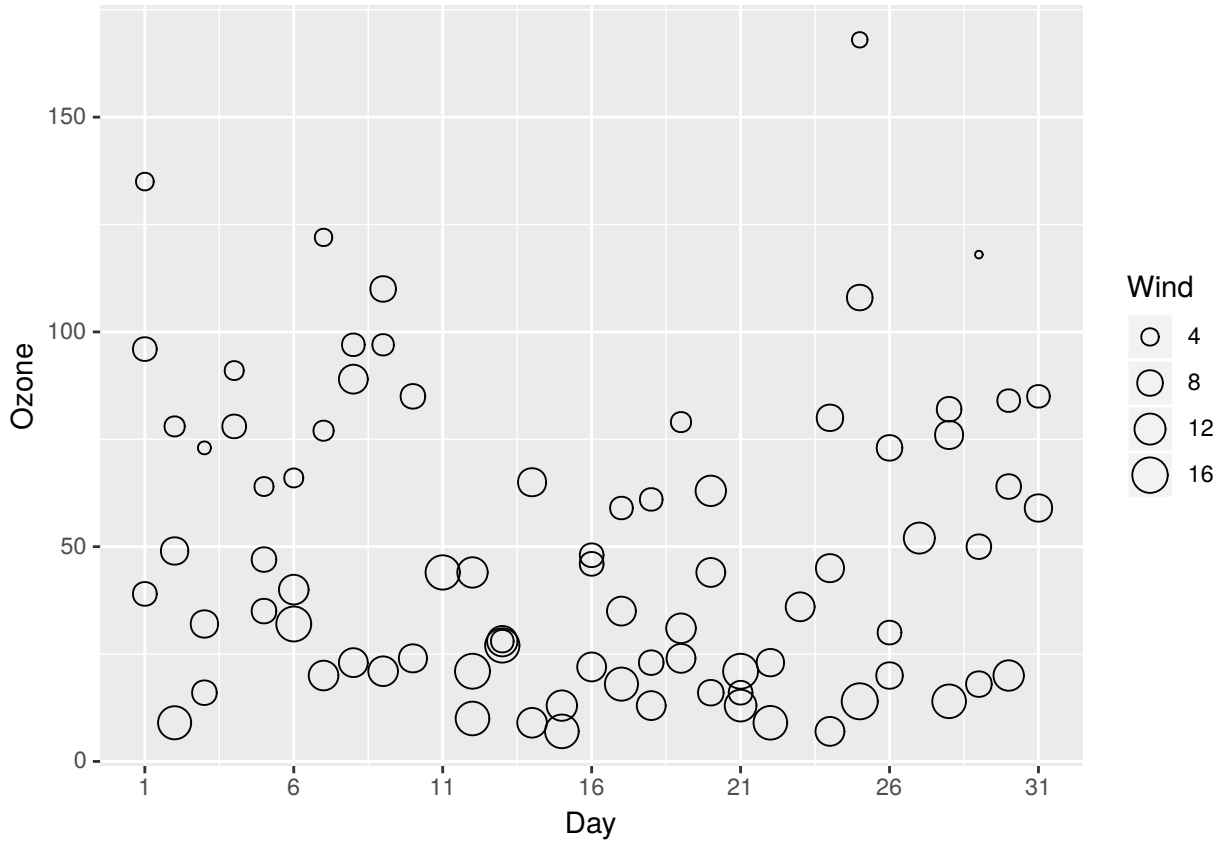
```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind)) +
  geom_point(shape = 21)
p6
```



6.4. Adjusting the axis scales

To change the x-axis tick marks, we use the `scale_x_continuous` option. Similarly, to change the y-axis we use the `scale_y_continuous` option. Here we will change the x-axis to every 5 days, rather than 10, and change the range from 1 to 31 (as 0 is not a valid value for this variable).

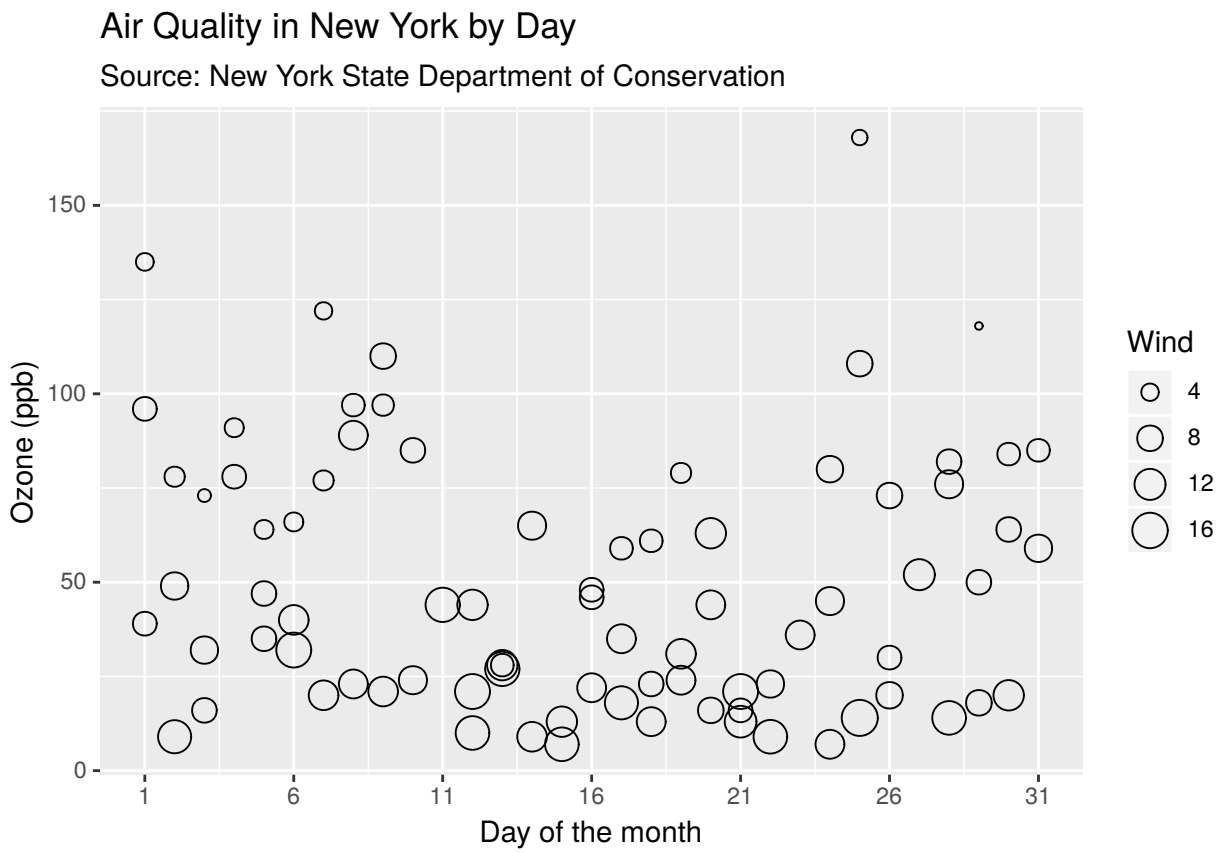
```
p6 <- p6 + scale_x_continuous(breaks = seq(1, 31, 5))
p6
```



6.5. Adjusting axis labels & adding title

To add a title, we include the option `labs` and include the name of the graph as a string argument. To change the axis names we add `x` and `y` arguments to the `labs` command.

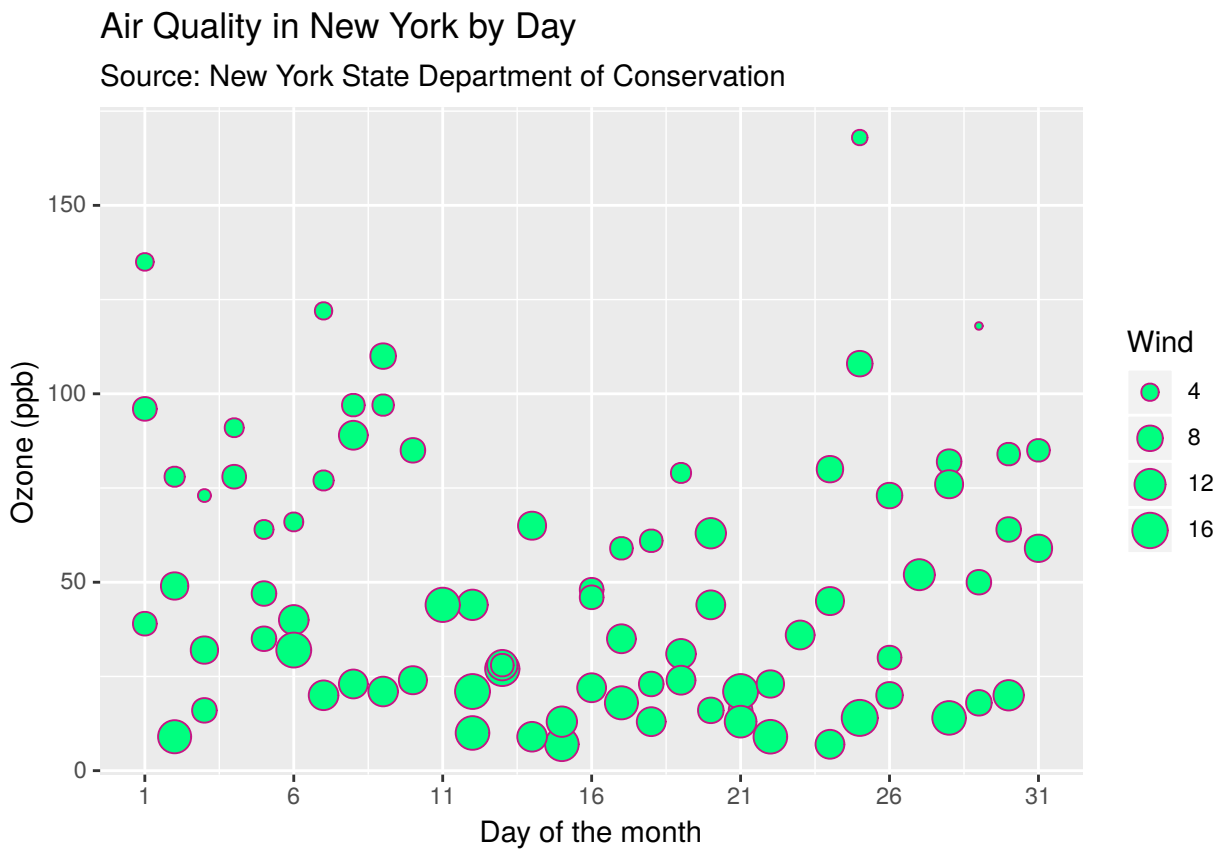
```
p6 <- p6 +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)")
p6
```



6.6. Adjusting the colour palette

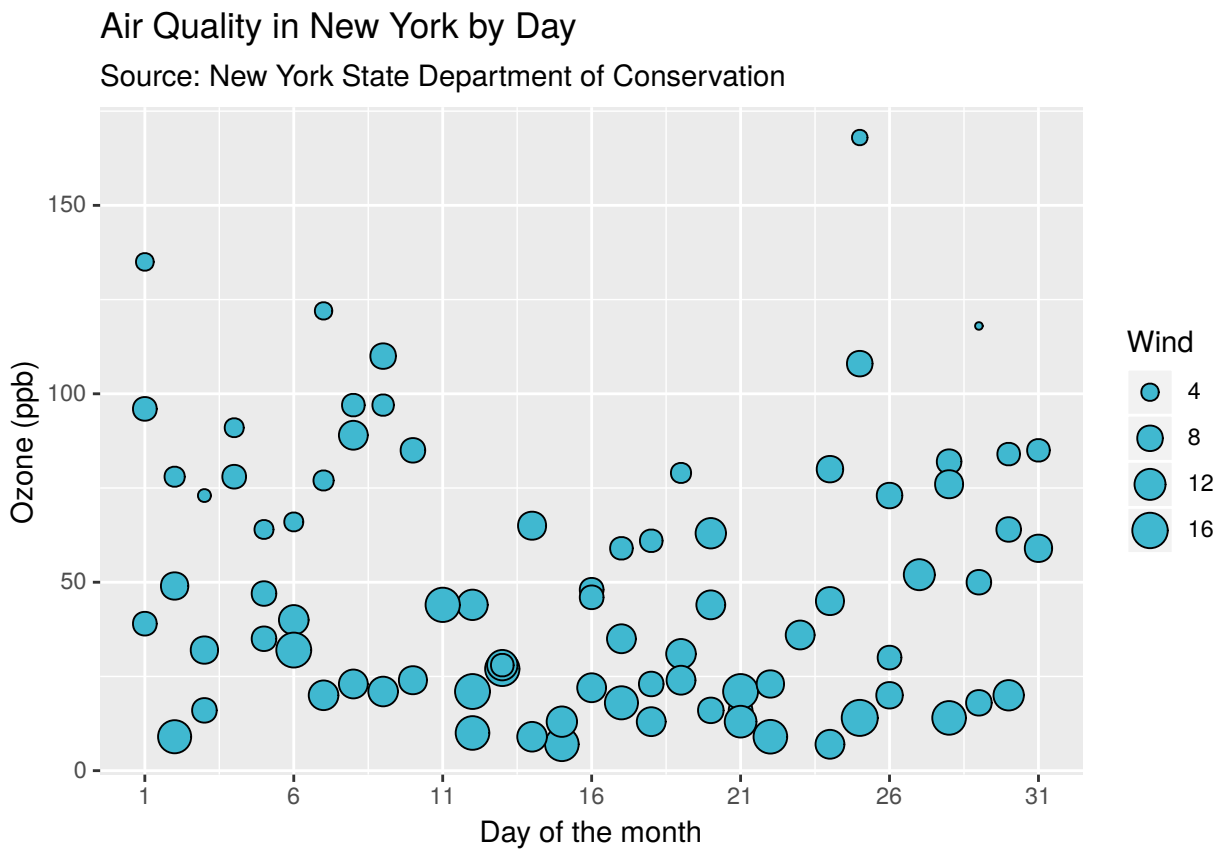
There are a few options for adjusting the colour. The most simple is to make every point one fixed colour. You can reference colours by name, with the full list of colours recognised by R here. Let's try making the outline `mediumvioletred` and the fill `springgreen`.

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind)) +
  geom_point(shape = 21, colour = "mediumvioletred", fill = "springgreen") +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p6
```



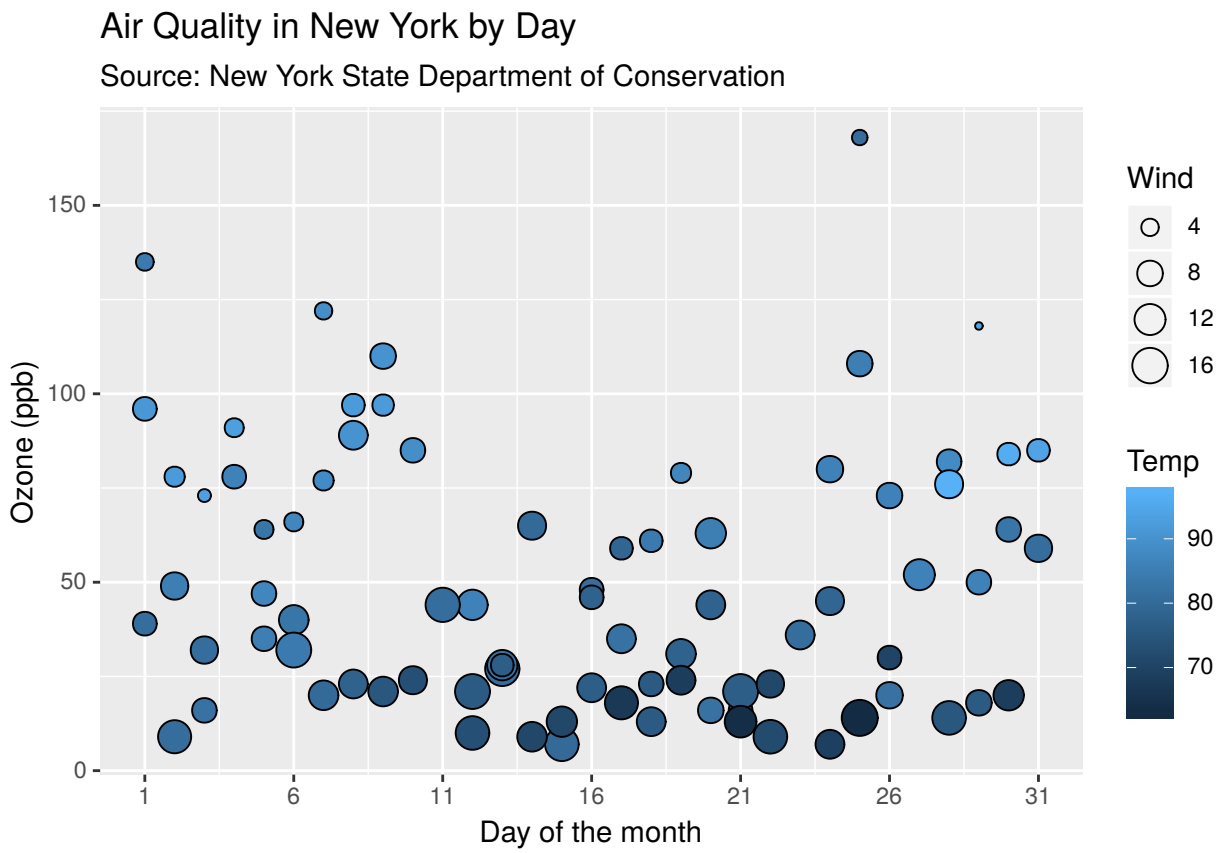
You can change the colours using specific HEX codes instead. Here we have made the outline #000000 (black) and the fill "#40b8d0 (vivid cyan).

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind)) +
  geom_point(shape = 21, colour = "#000000", fill = "#40b8d0") +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p6
```



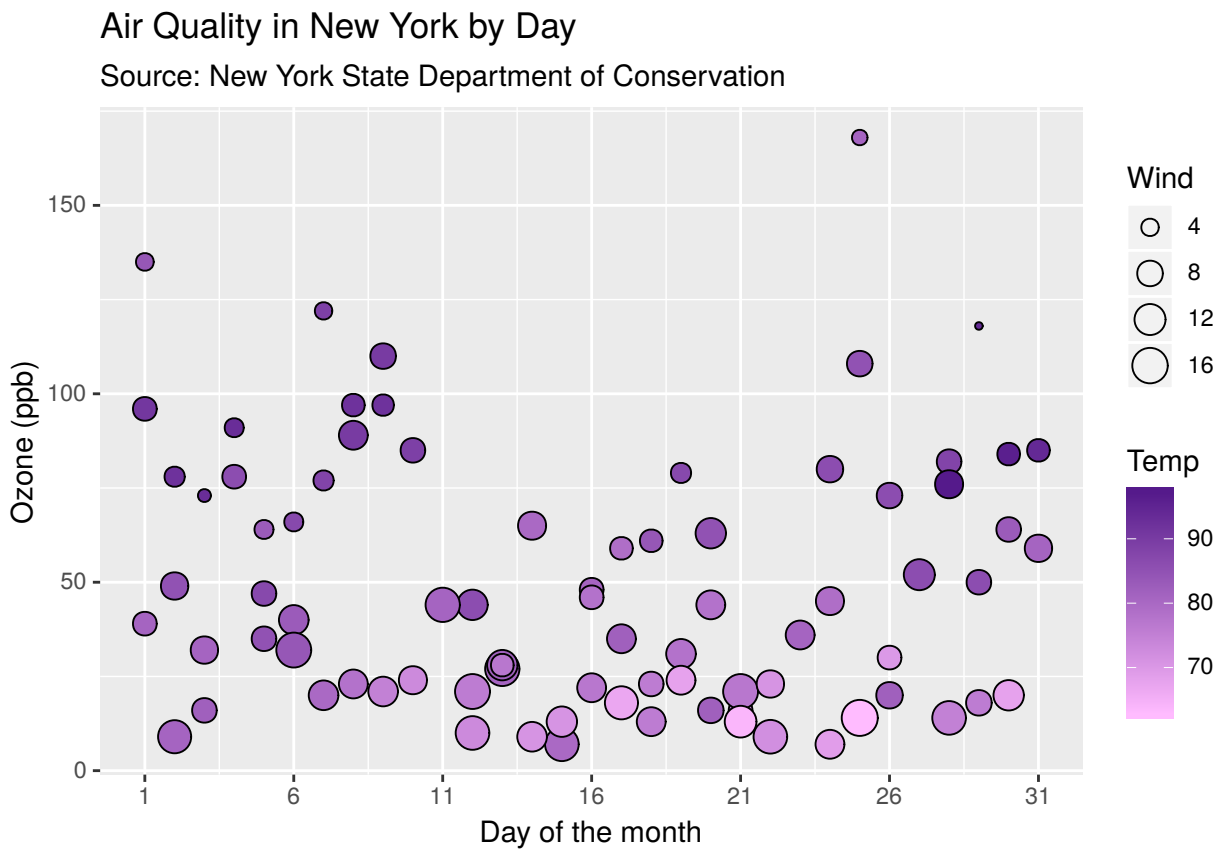
You can also change the colour of the data points according to the levels of another variable. This can be done either as a continuous gradient, or as a levels of a factor variable. Let's change the colour by the values of temperature:

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Temp)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p6
```



We can change the gradient's colours by adding the `scale_fill_continuous` option. The `low` and `high` arguments specify the range of colours the gradient should transition between.

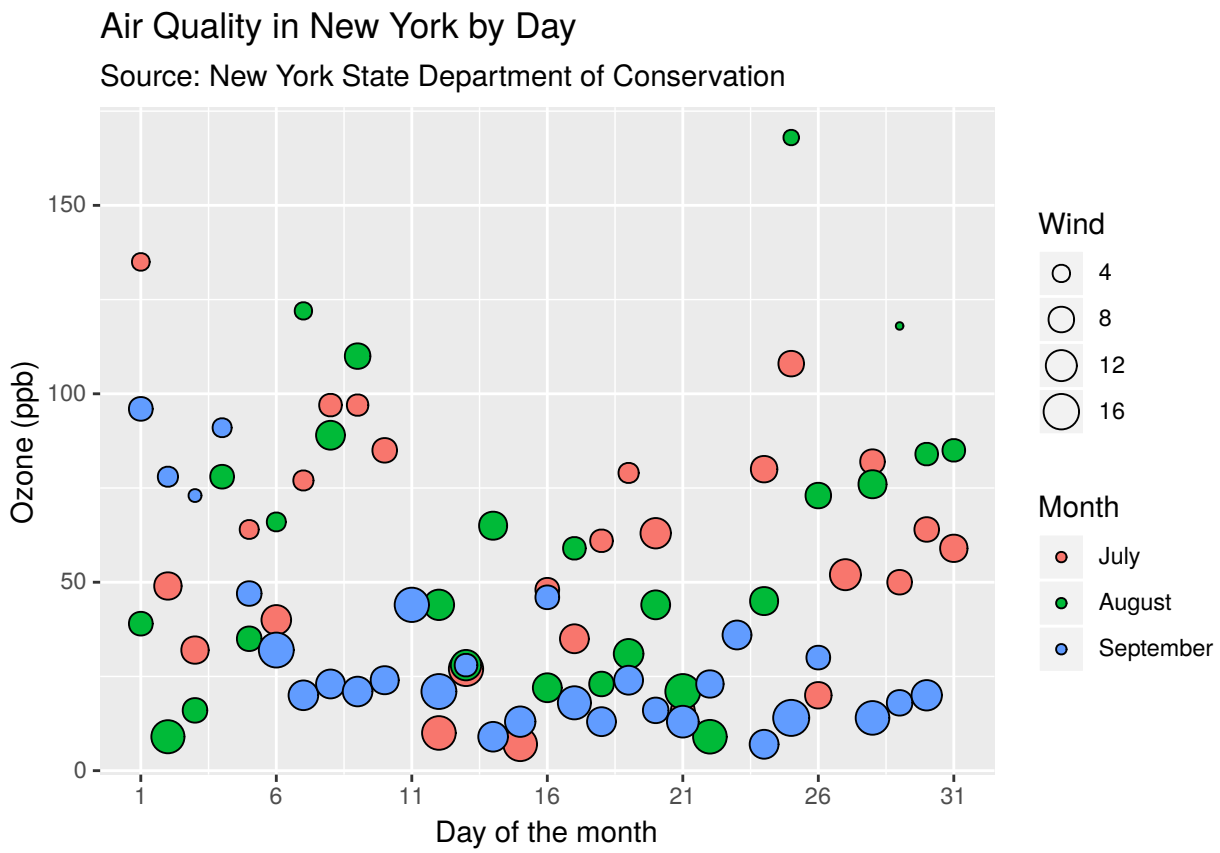
```
p6 <- p6 + scale_fill_continuous(low = "plum1", high = "purple4")
p6
```



We can see that higher temperatures seem to have higher ozone levels.

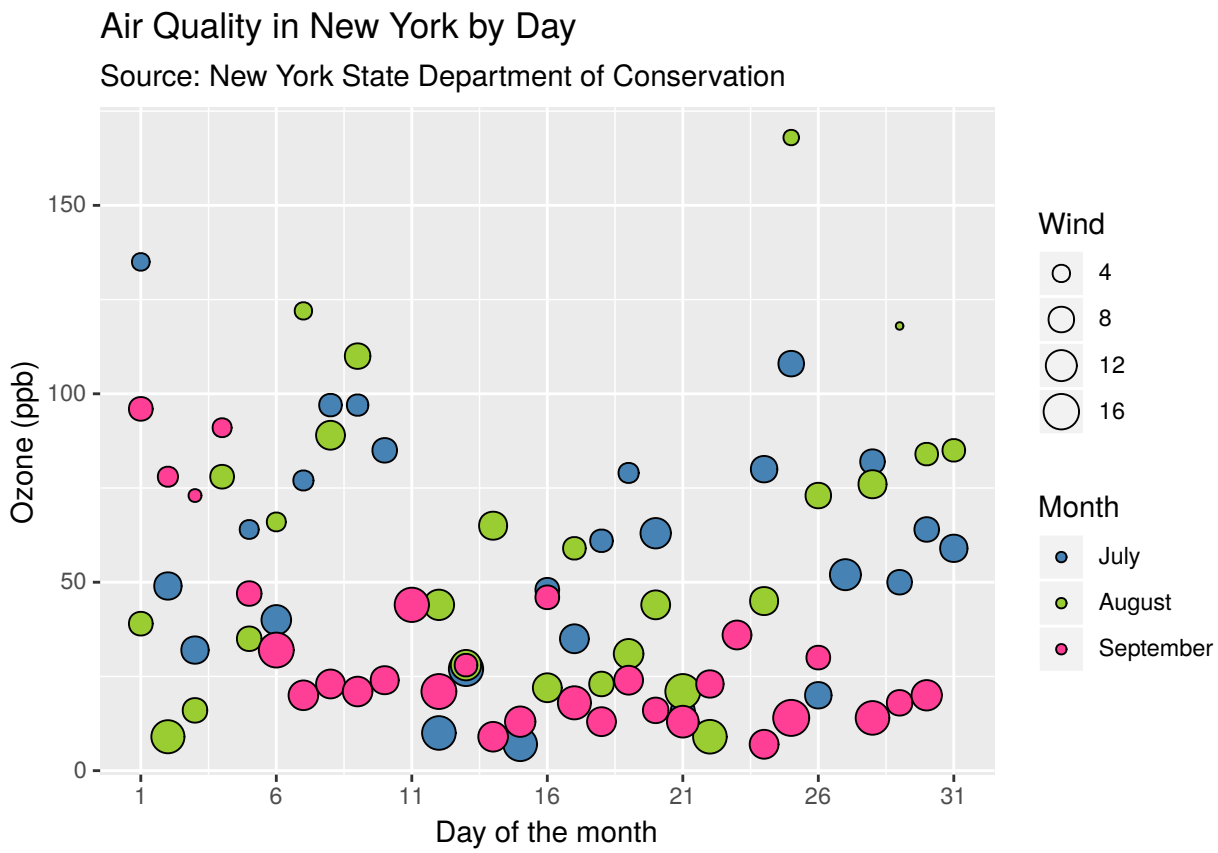
Let's now change the colours of the data points by a factor variable, `Month`.

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5))
p6
```



Again, we can change the colours of these data points, this time using `scale_fill_manual`.

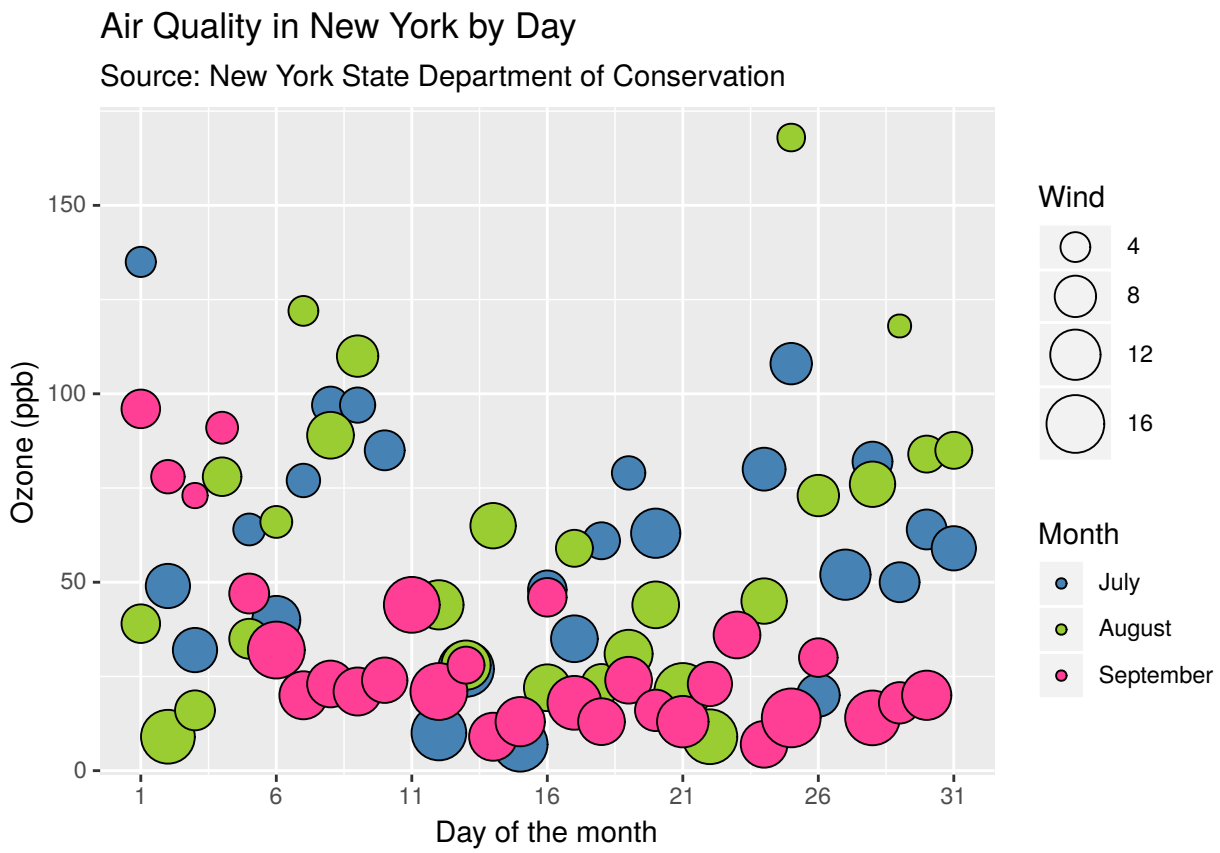
```
fill <- c("steelblue", "yellowgreen", "violetred1")

p6 <- p6 + scale_fill_manual(values = fill)
p6
```



6.7. Adjusting the size of the data points

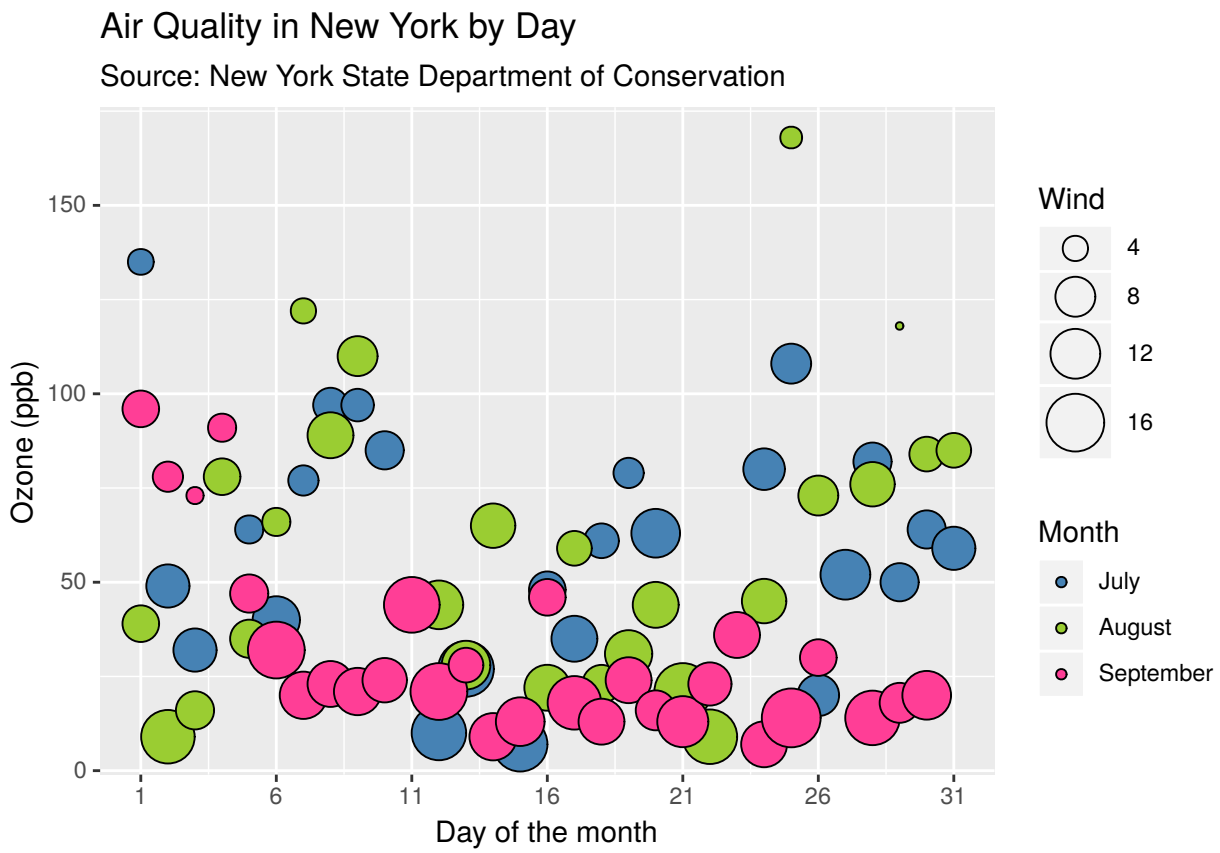
The default size of the the data points in a weighted scatterplot is mapped to the radius of the plots. If we want the data points to be proportional to the value of the weighting variable (e.g., a wind speed of 0 mph would have a value of 0), we need to use the `scale_size_area`.

```
p6 <- p6 + scale_size_area(max_size = 10)
p6
```



For our graph, this makes the pattern for `Wind` a little hard to see. Another way to adjust the size of the data points is to use `scale_size` and specify a desired range.

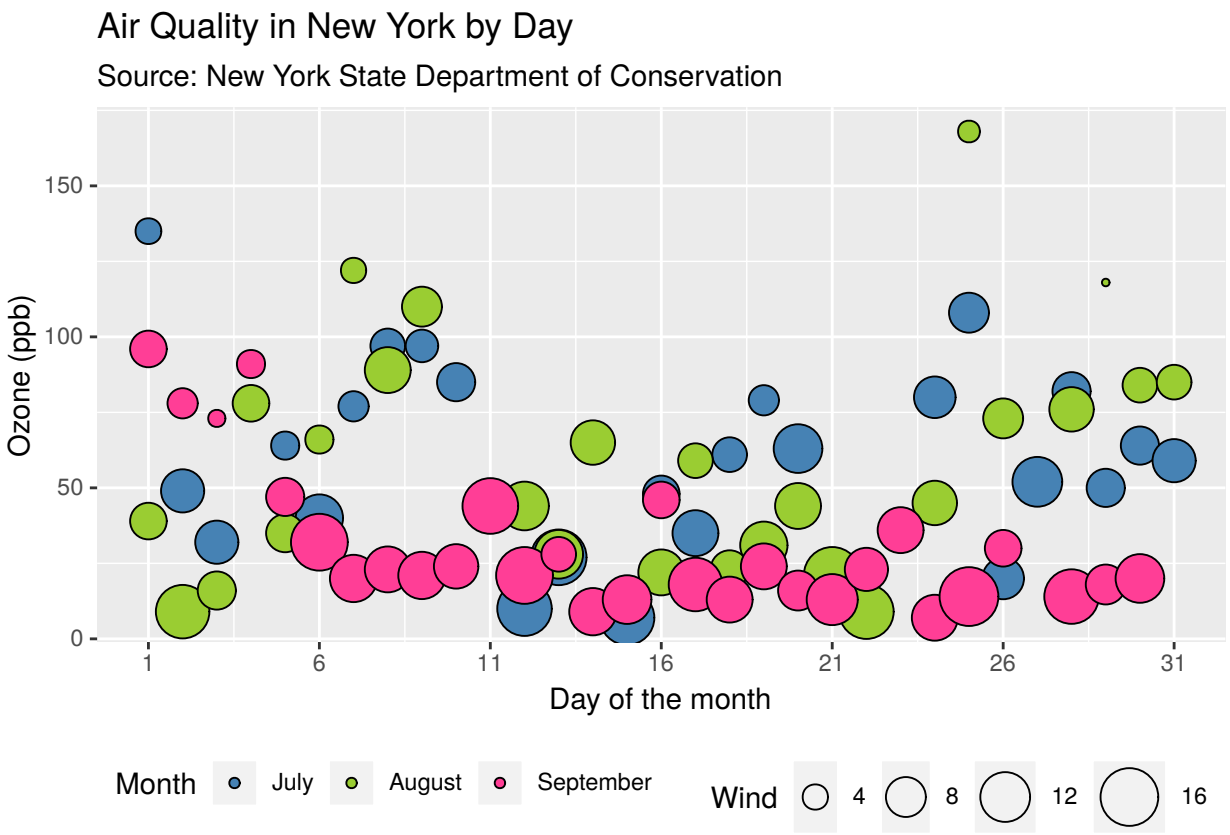
```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_fill_manual(values = fill) +
  scale_size(range = c(1, 10))
p6
```



6.8. Adjusting legend position

To adjust the position of the legend from the default spot of right of the graph, we add the `theme` option and specify the `legend.position = "bottom"` argument. We can also change the legend shape using the `legend.direction = "horizontal"` argument.

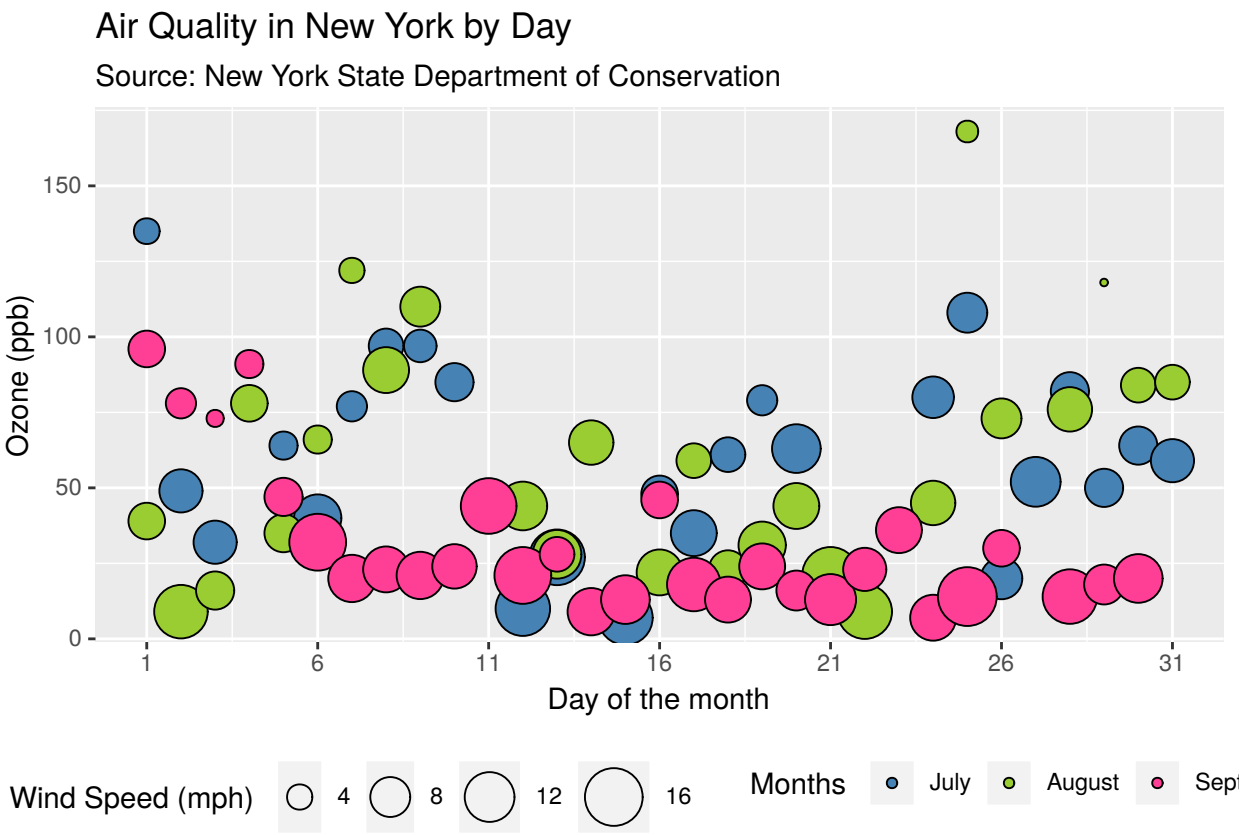
```
p6 <- p6 + theme(legend.position = "bottom", legend.direction = "horizontal")
p6
```



6.9. Changing the legend titles

To change the titles of the two legends, we use the `labs` option. In order to tell ggplot2 exactly what legend you're referring to, just have a look in the `ggplot` option and see what argument you used to create the legend in the first place. In this case we used the `size` argument for "Wind" and `fill` for "Month", so we pass these to `labs` with our new titles.

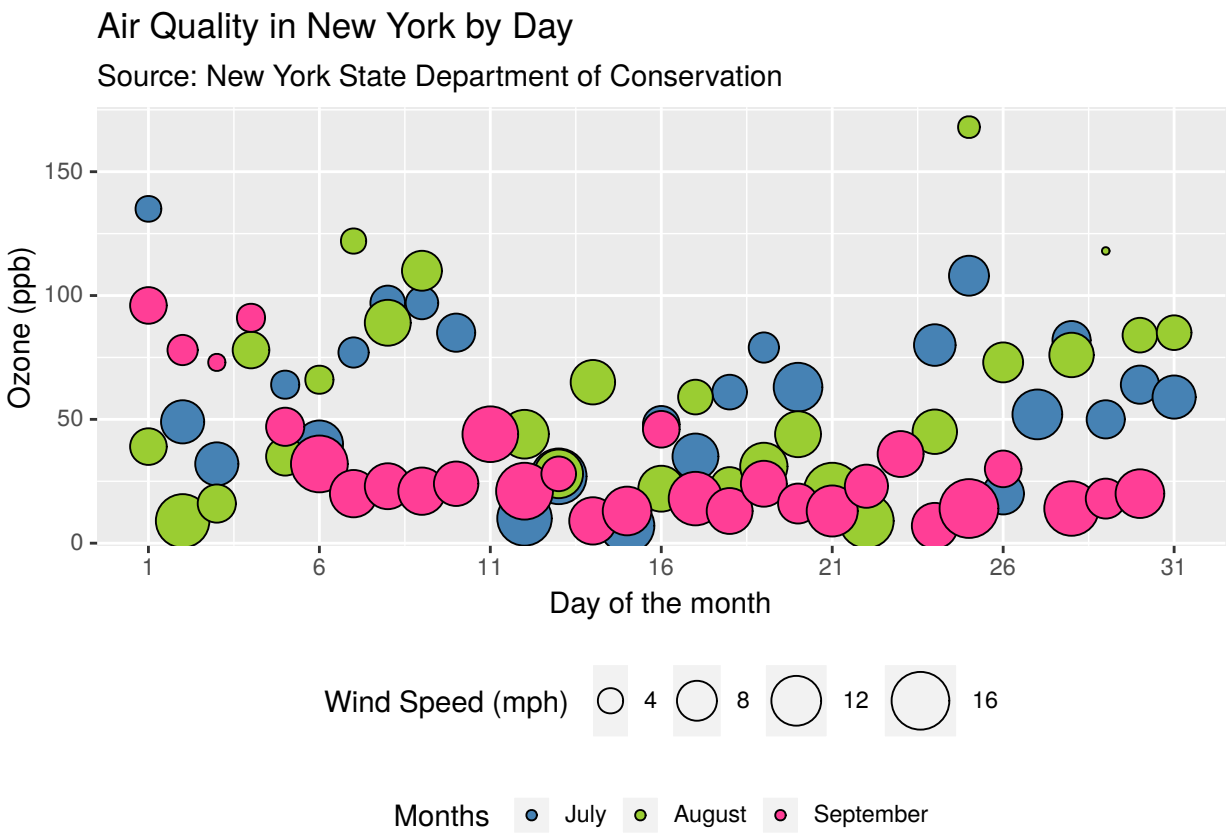
```
p6 <- p6 + labs(size = "Wind Speed (mph) ", fill = "Months ")
p6
```



6.10. Creating horizontal legends

It looks a little awkward having the two titles sitting on top of each other, as well as taking up unnecessary space. To place the legends next to each other, we use the `legend.box = "vertical"` argument in `theme`. Because the boxes around the legend keys aren't even in each of the legends, this means the legends don't align properly. To fix this, we change the box size around the legend keys using `legend.key.size`. We need to load in the `grid` package to get this argument to work.

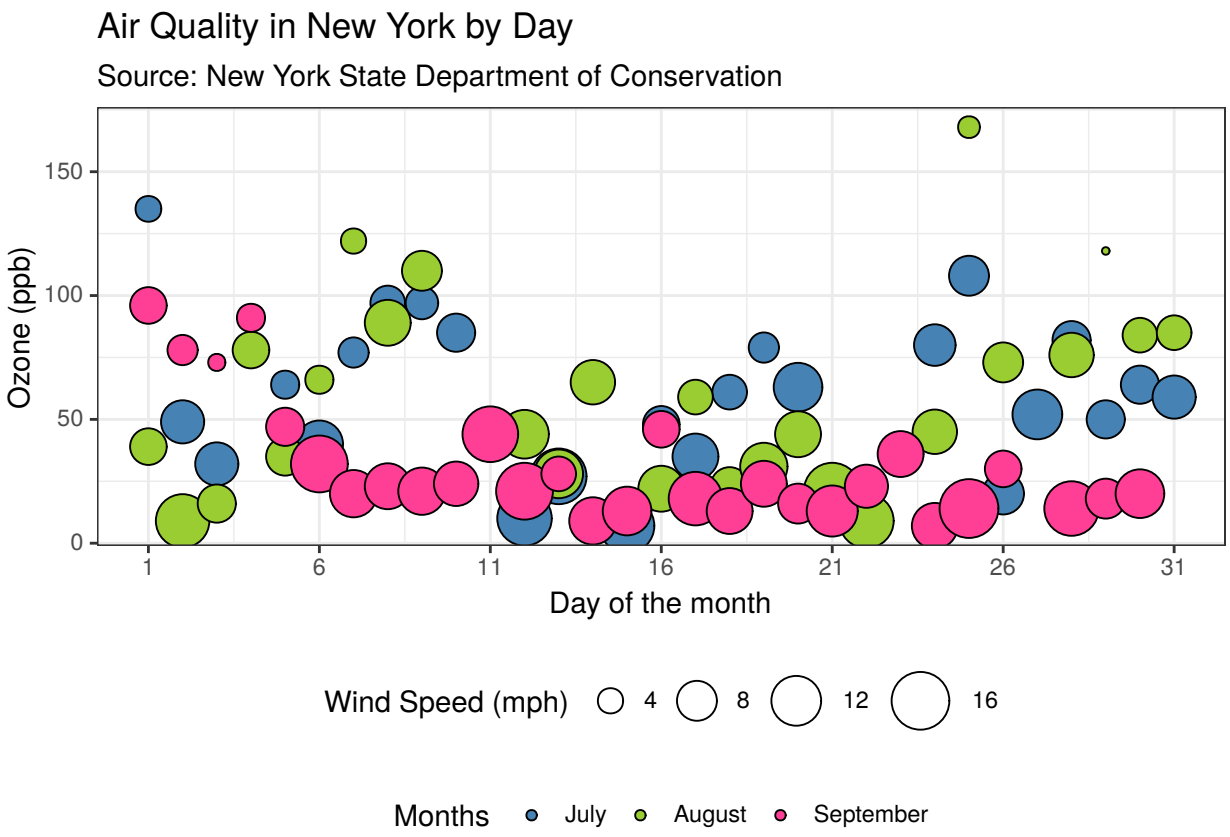
```
p6 <- p6 + theme(legend.box = "vertical", legend.key.size = unit(0.5, "cm"))
p6
```



6.11. 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()` after `ggplot()`. As you can see, we can further tweak the graph using the `theme` option, which we've used so far to change the legend.

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)",
    size = "Wind Speed (mph) ", fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_fill_manual(values = fill) +
  scale_size(range = c(1, 10)) +
  theme_bw() +
  theme(legend.position = "bottom", legend.direction = "horizontal",
    legend.box = "vertical",
    legend.key.size = unit(0.5, "cm"))
p6
```



6.12. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

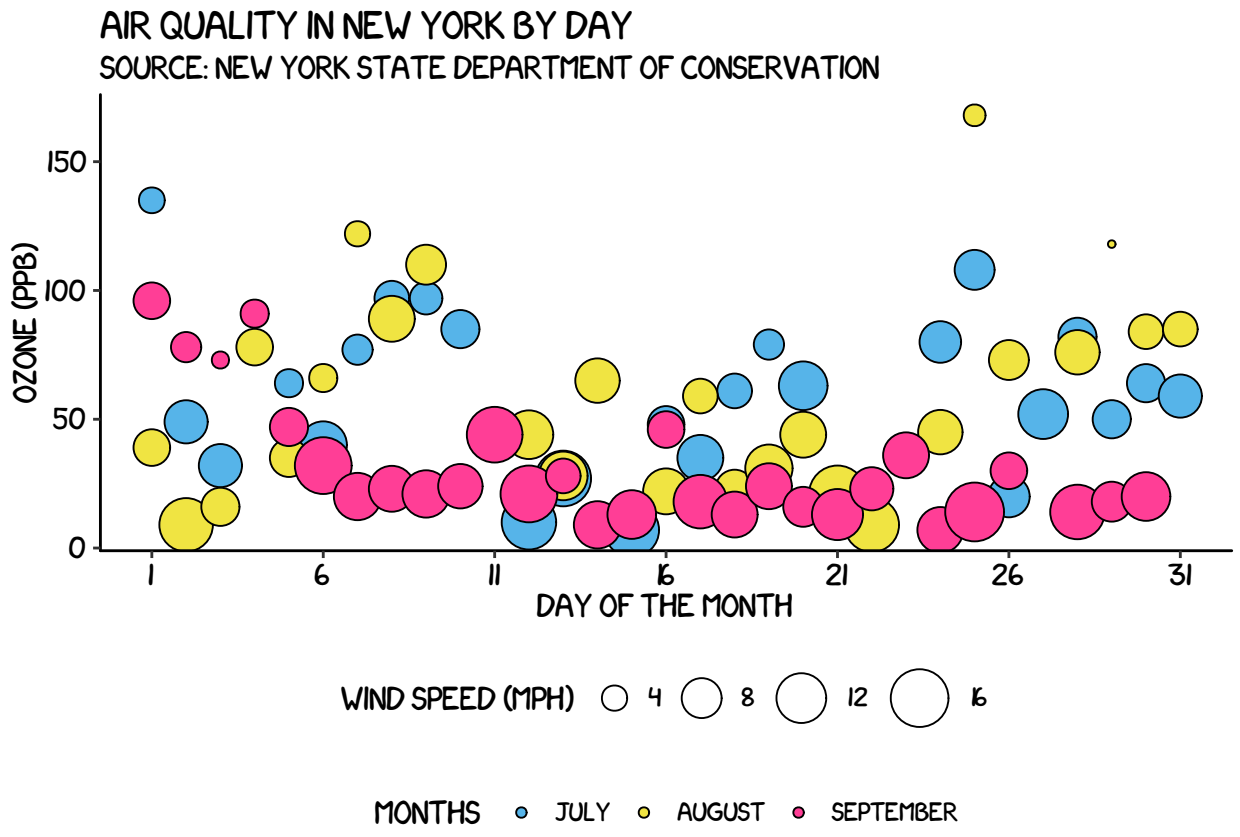
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

6.13. 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.

```
fill <- c("#56B4E9", "#F0E442", "violetred1")

p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)",
    size = "Wind Speed (mph) ", fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_fill_manual(values = fill) +
  scale_size(range = c(1, 10)) +
  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 = "vertical",
    legend.key.size = unit(0.5, "cm"),
    legend.key = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p6
```

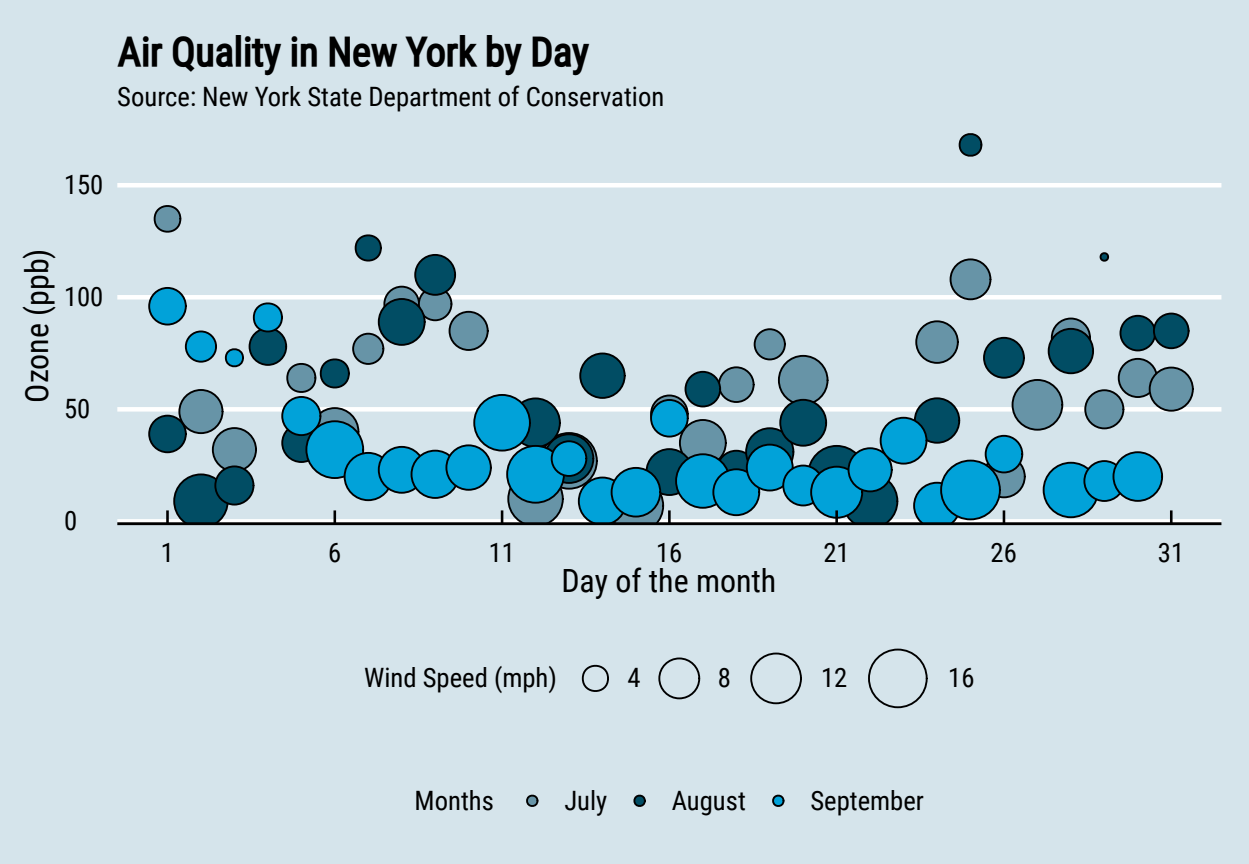


6.14. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)", size = "Wind Speed (mph) ",
       fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_size(range = c(1, 10)) +
  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 = "vertical",
    legend.key.size = unit(0.5, "cm"),
```

```
    legend.text = element_text(size = 10),
    text = element_text(family = "Roboto Condensed"),
    plot.title = element_text(family = "Roboto Condensed"))
p6
```

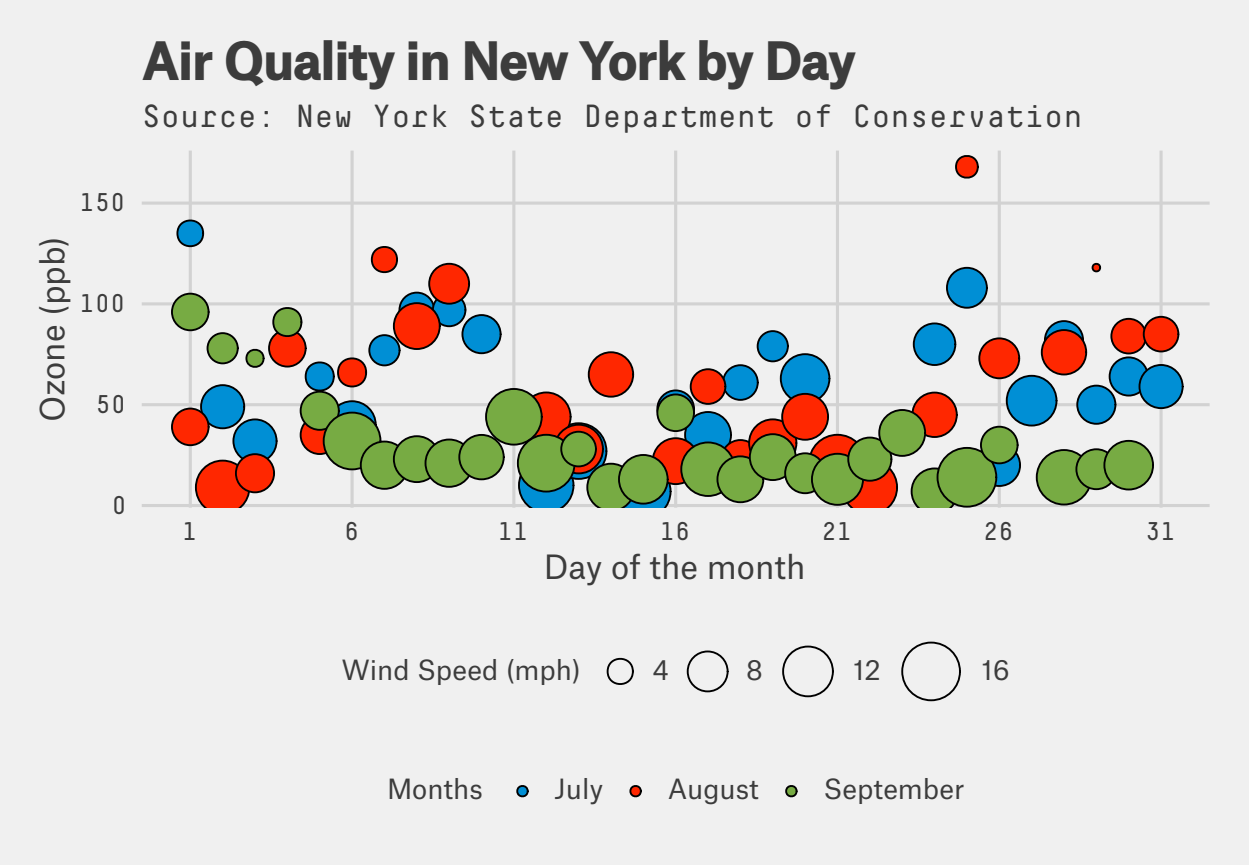


6.15. Using 'Five Thirty Eight' theme

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'.

```
p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
  labs(x = "Day of the month", y = "Ozone (ppb)", size = "Wind Speed (mph) ",
       fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_size(range = c(1, 10)) +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
```

```
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "vertical",
    legend.key.size = unit(0.5, "cm"),
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 10),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 10),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    text = element_text(family = "Decima Mono Pro"))
p6
```

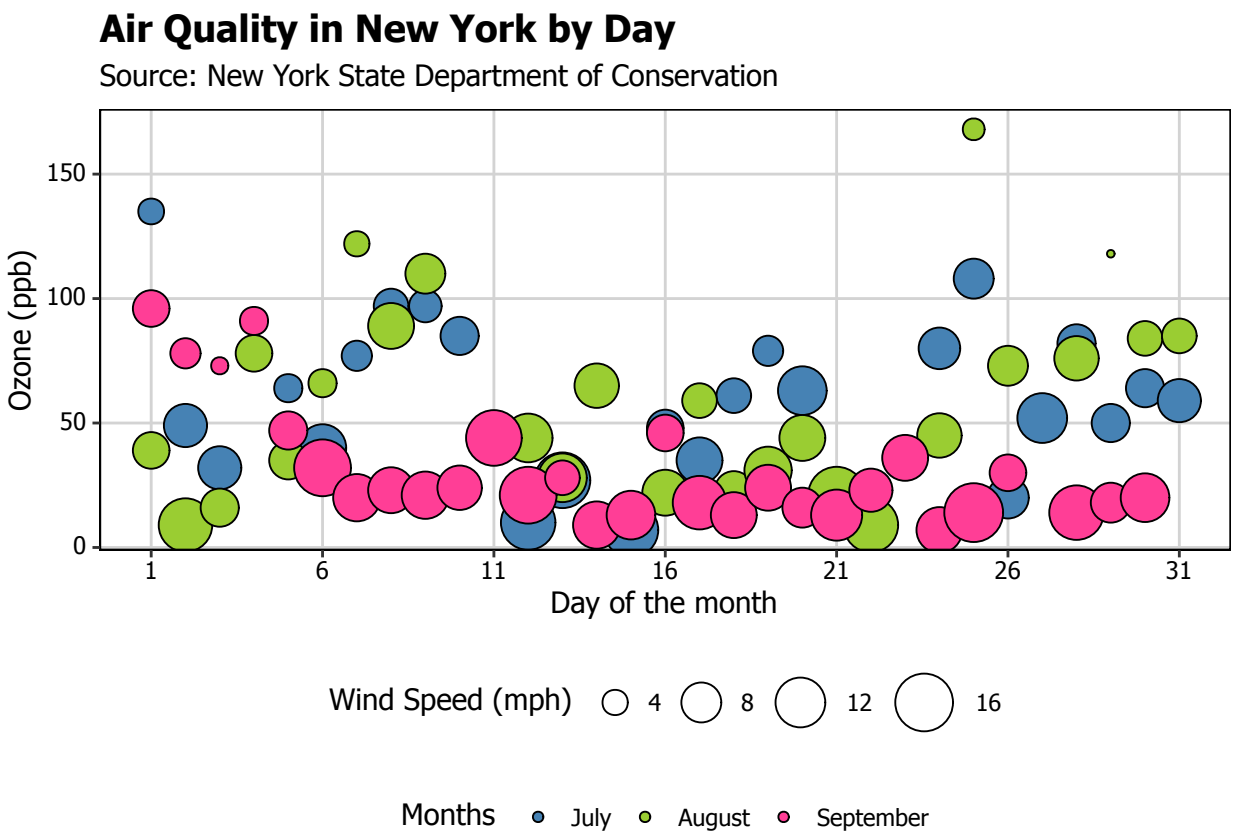


6.16. Creating your own theme

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 <- c("steelblue", "yellowgreen", "violetred1")

p6 <- ggplot(aq_trim, aes(x = Day, y = Ozone, size = Wind, fill = Month)) +
  geom_point(shape = 21) +
  labs(title = "Air Quality in New York by Day",
       subtitle = "Source: New York State Department of Conservation") +
```

```
  labs(x = "Day of the month", y = "Ozone (ppb)", size = "Wind Speed (mph) ",
       fill = "Months ") +
  scale_x_continuous(breaks = seq(1, 31, 5)) +
  scale_size(range = c(1, 10)) +
  scale_fill_manual(values = fill) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "vertical",
    legend.key.size = unit(0.5, "cm"),
    legend.key = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p6
```

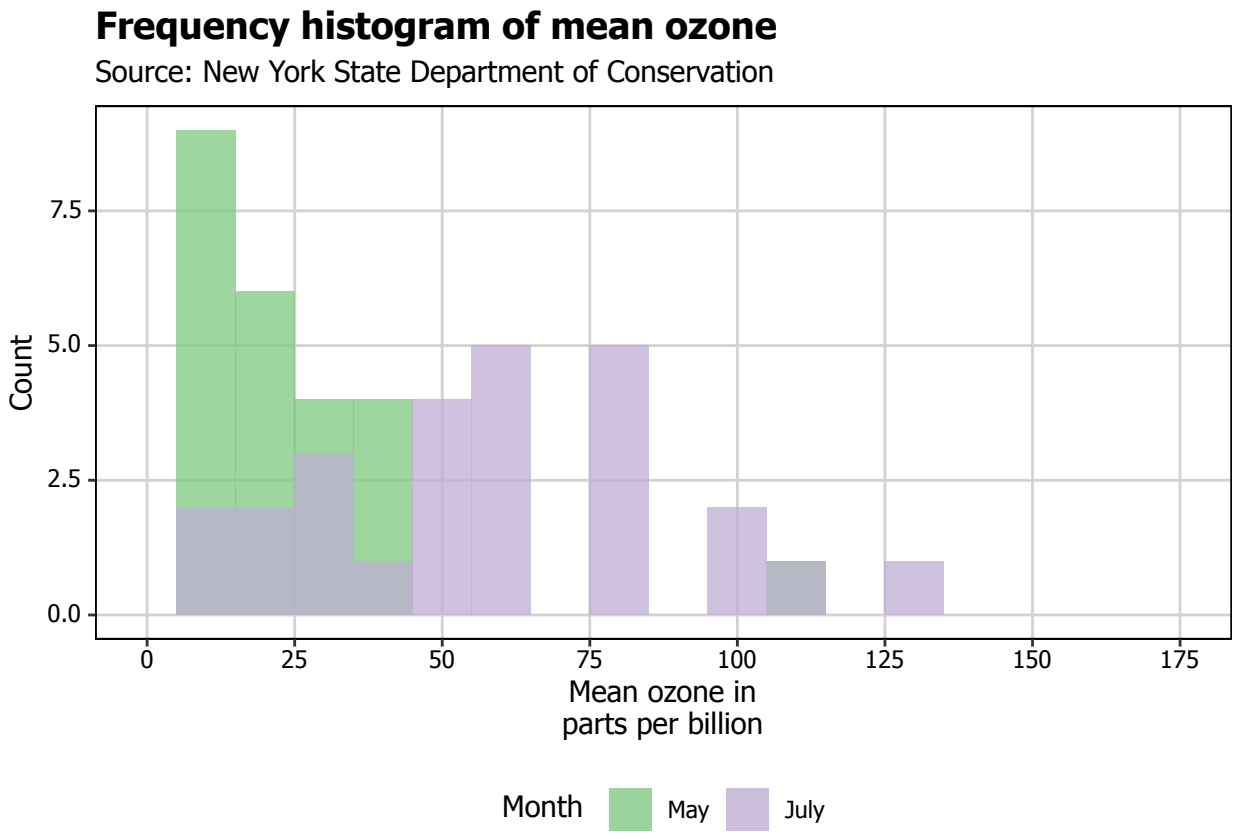


CHAPTER 7

Histograms

7.1. Introduction

In this chapter, we will work towards creating the histogram below. We will take you from a basic histogram and explain all the customisations we add to the code step-by-step.



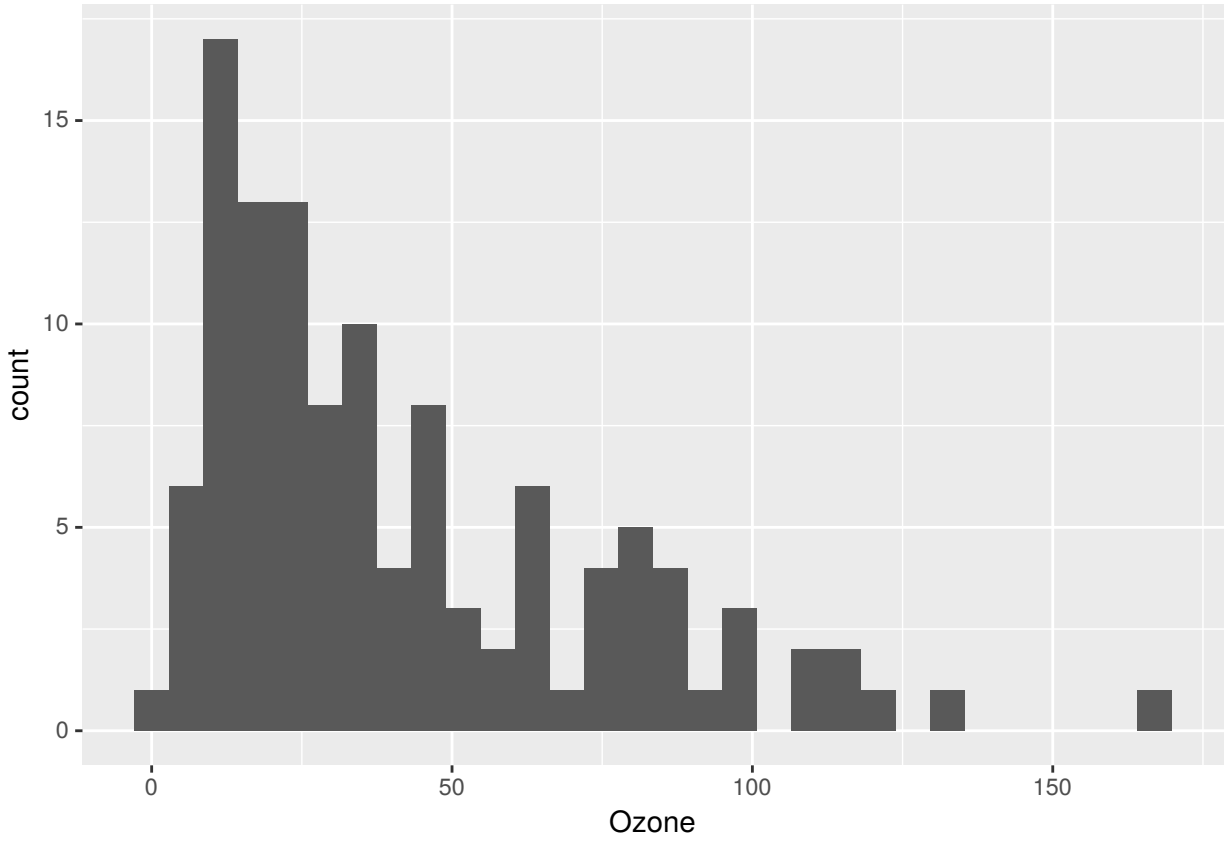
The first thing to do is load in the data and the libraries, as below:

```
if (!require("pacman")) install.packages("pacman")
p_load(datasets, ggplot2, ggthemes, dplyr, grid, RColorBrewer)

data(airquality)
```

7.2. Basic histogram

In order to initialise a plot we tell ggplot that `airquality` is our data, and specify that our x axis plots the `Ozone` variable. We then instruct ggplot to render this as a histogram by adding the `geom_histogram()` option.

```
p7 <- ggplot(airquality, aes(x = Ozone)) + geom_histogram()
p7
```



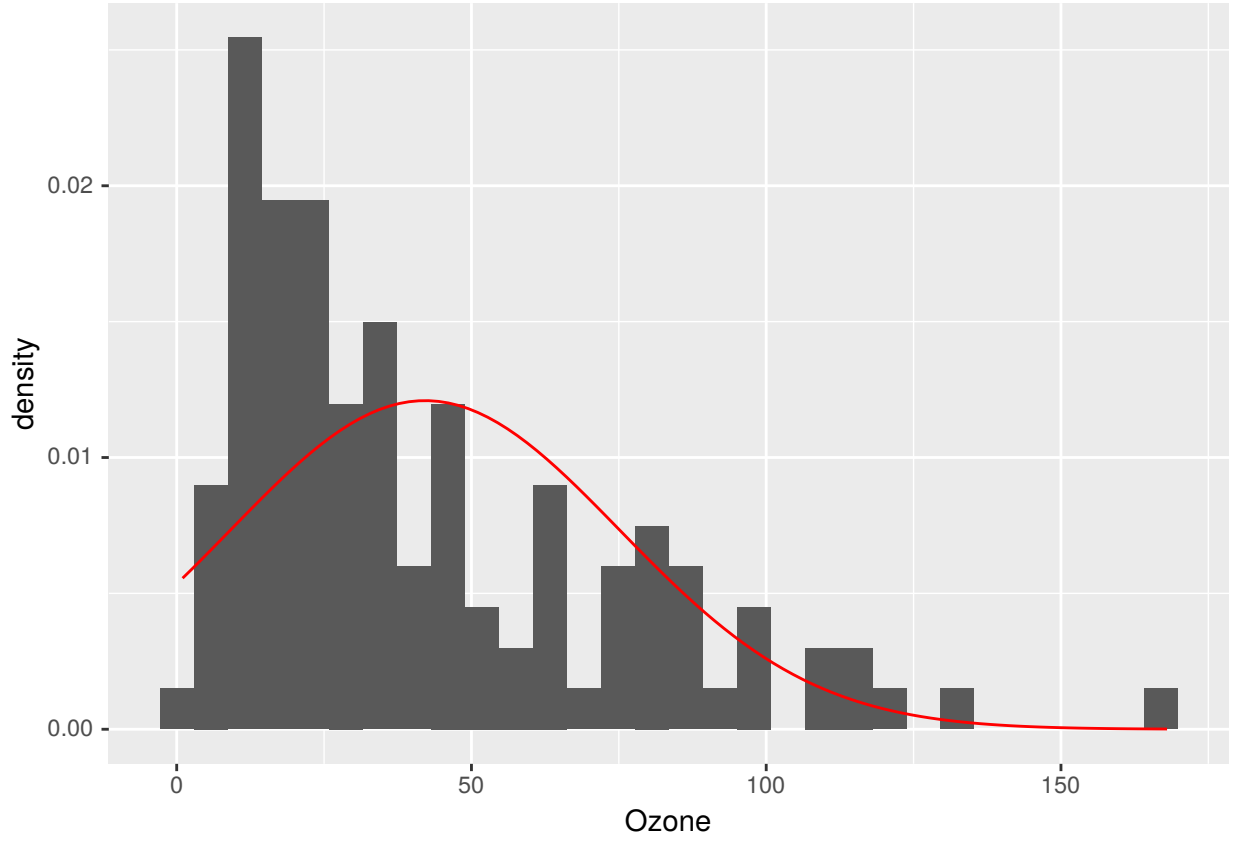
7.3. Adding a normal density curve

We can overlay a normal density function curve on top of our histogram to see how closely (or not) it fits a normal distribution. In this case, we can see it deviates from a normal distribution, showing marked positive skew. In order to overlay the function curve, we add the option `stat_function(fun = dnorm)`, and specify the shape using the `mean = mean(airquality$Ozone)` and `sd = sd(airquality$Ozone)`

arguments. If you have missing data like we did, make sure you pass the `na.rm = TRUE` argument to the mean and sd parameters. Finally, you can change the colour using the `colour = "red"` argument. We will discuss how to customise colours further below.

One further change we must make to display the normal curve correctly is adding `aes(y = ..density..)` to the `geom_histogram` option. Note that the normal density curve will not work if you are using the frequency rather than the density, which we are changing in our next step.

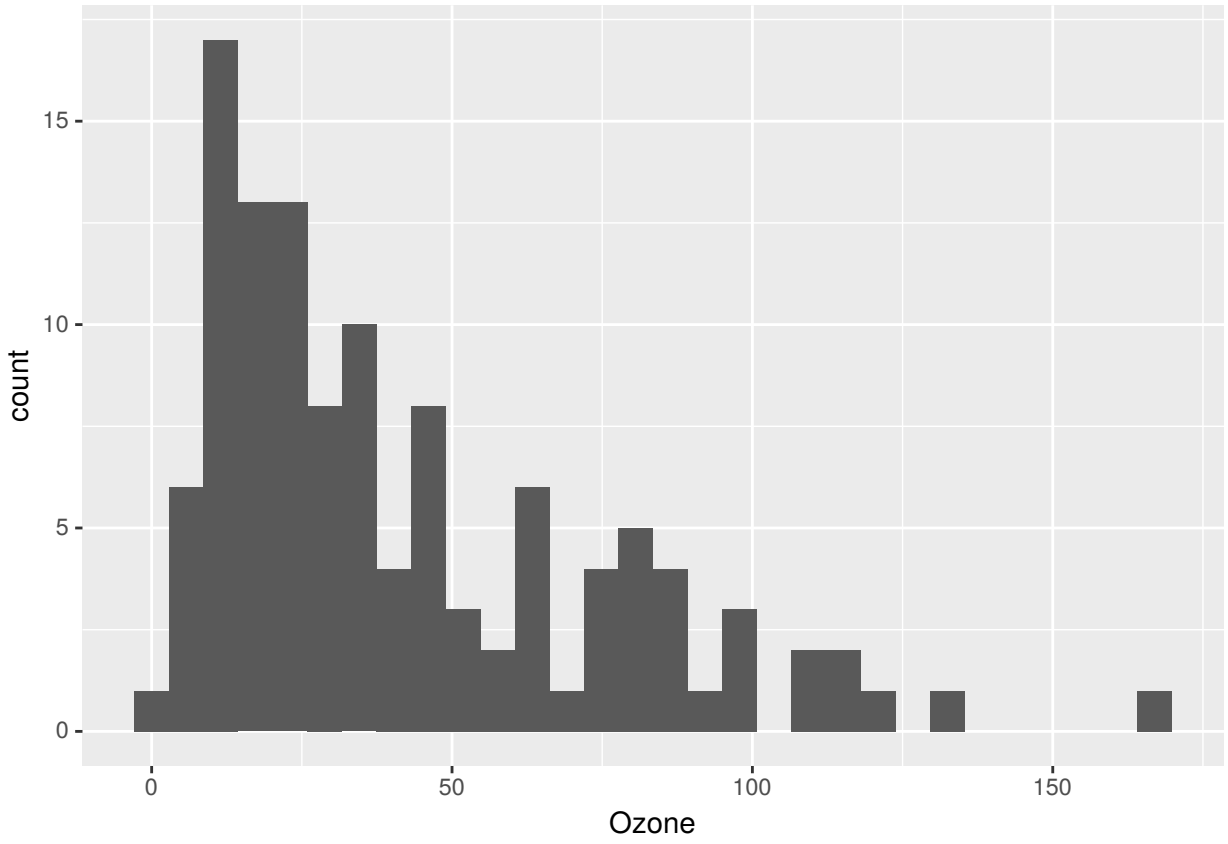
```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..density..)) +
  stat_function(fun = dnorm, colour = "red",
    args = list(mean = mean(airquality$Ozone, na.rm = TRUE),
    sd = sd(airquality$Ozone, na.rm = TRUE)))
p7
```



7.4. Changing from density to frequency

Let's go back to the basic plot and lose the function curve. To change the y-axis from density to frequency, we add the `aes(y = ..count..)` option to `geom_histogram`.

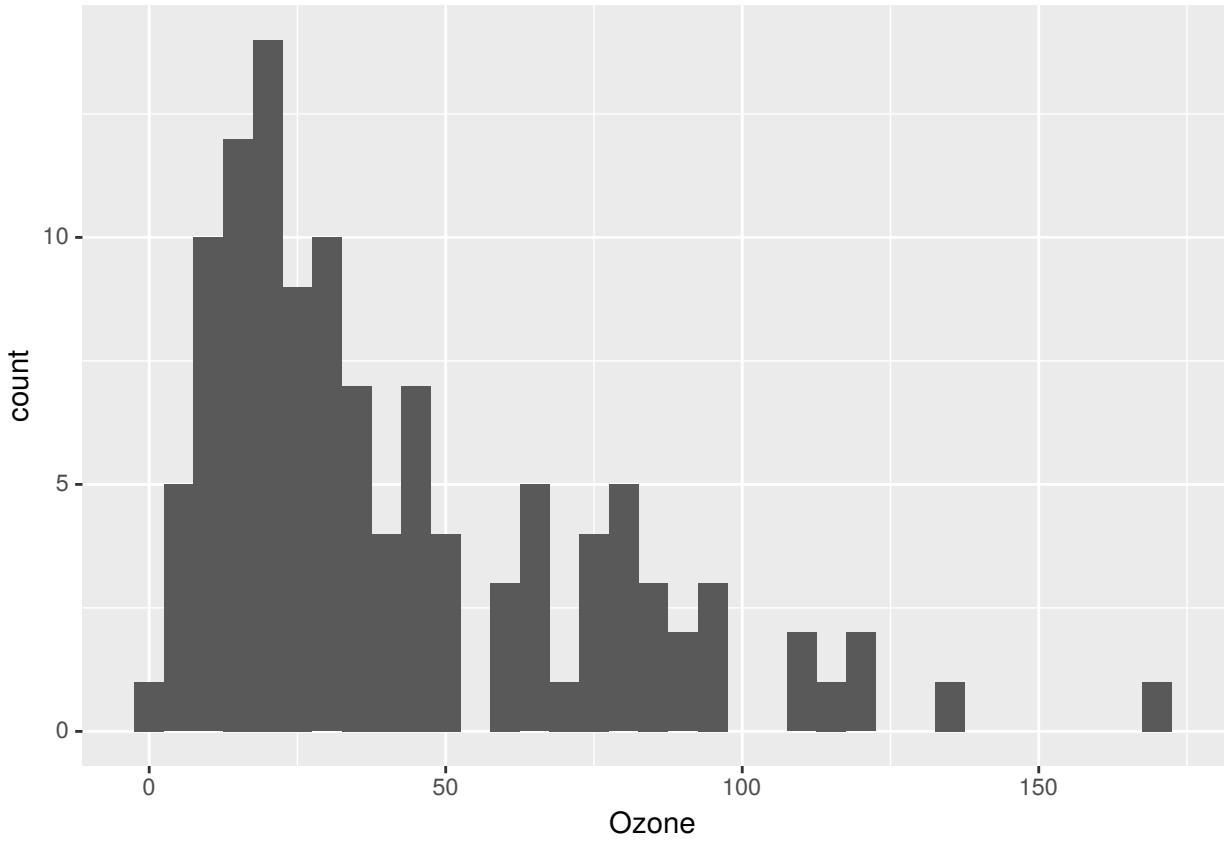
```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..))
p7
```



7.5. Adjusting binwidth

To change the binwidth, we add a `binwidth` argument to `geom_histogram`. In this case, we will make binwidth 5 units of the `Ozone` variable.

```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5)
p7
```

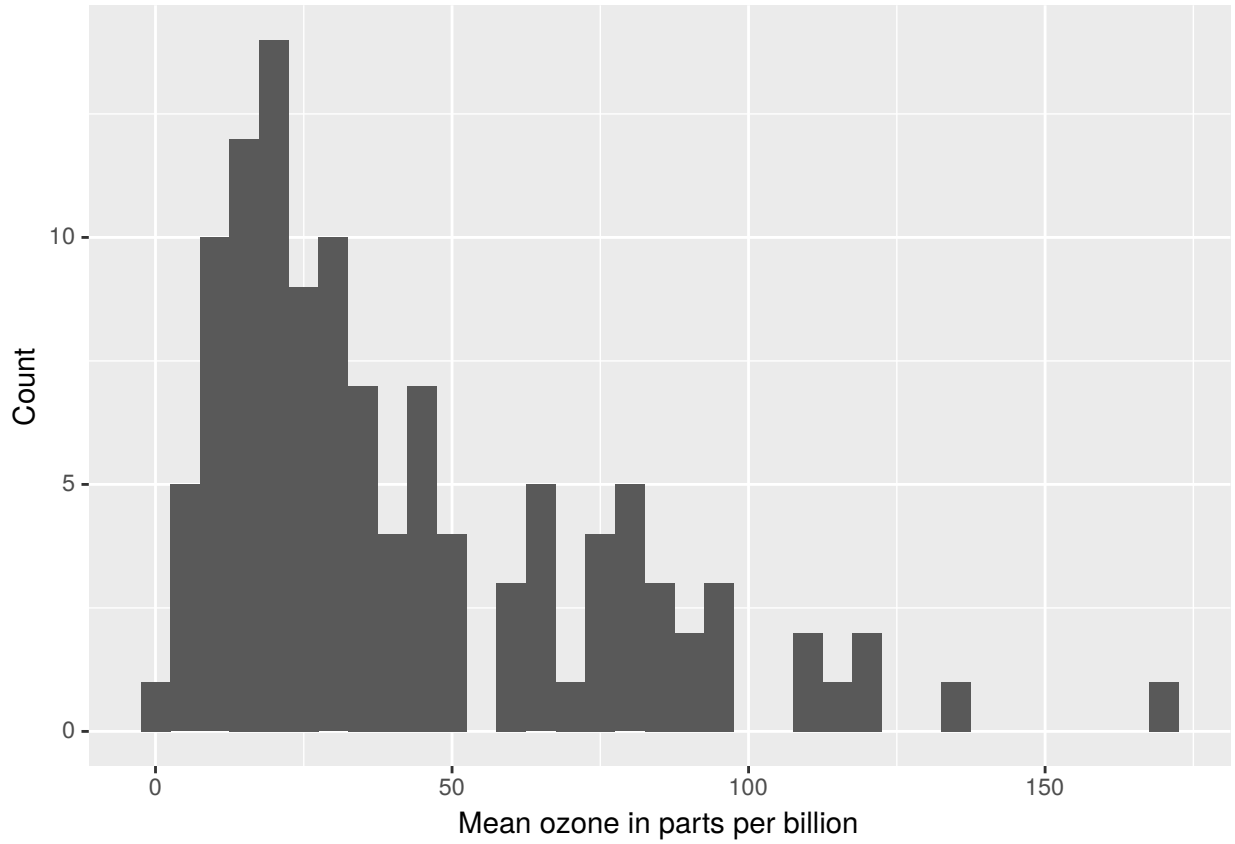


7.6. Customising axis labels

7.6.1. Single line labels

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.

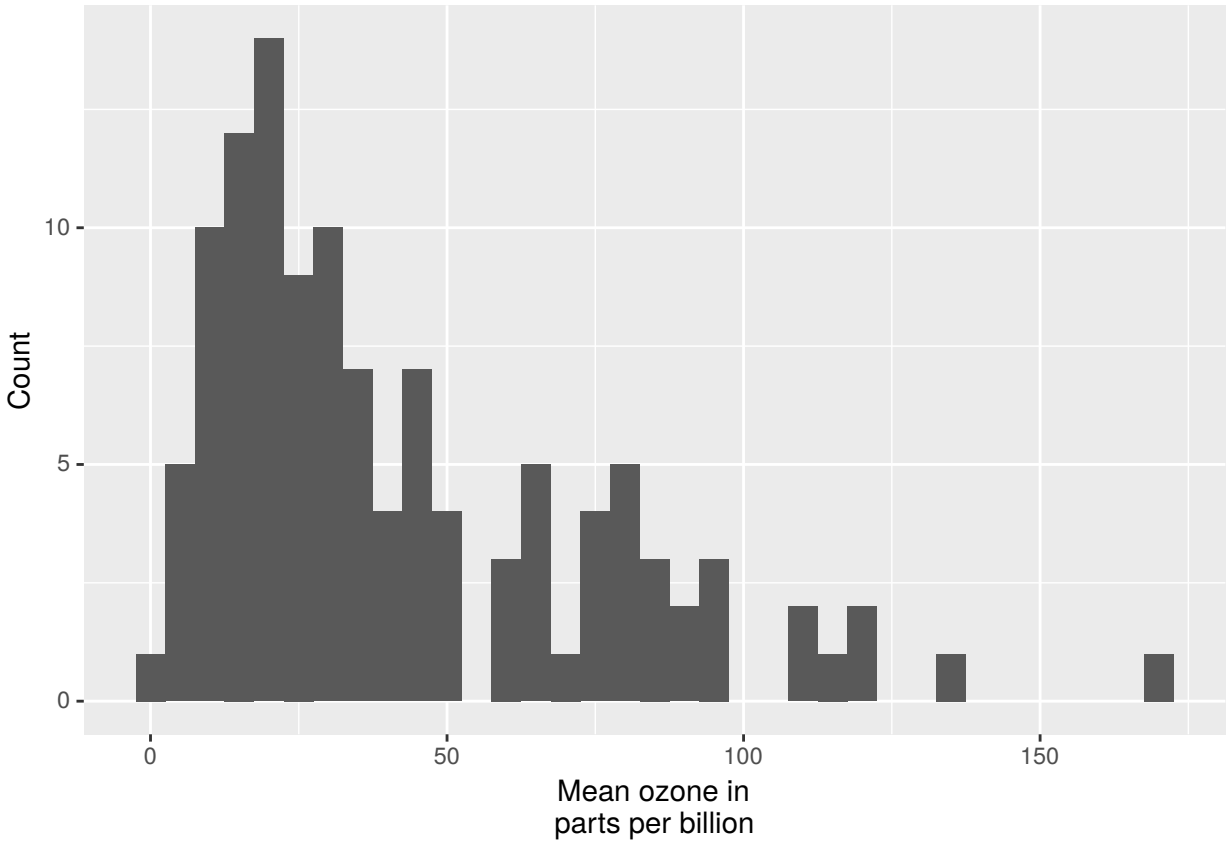
```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5) +
  scale_x_continuous(name = "Mean ozone in parts per billion") +
  scale_y_continuous(name = "Count")
p7
```



7.6.2. Multiline labels

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

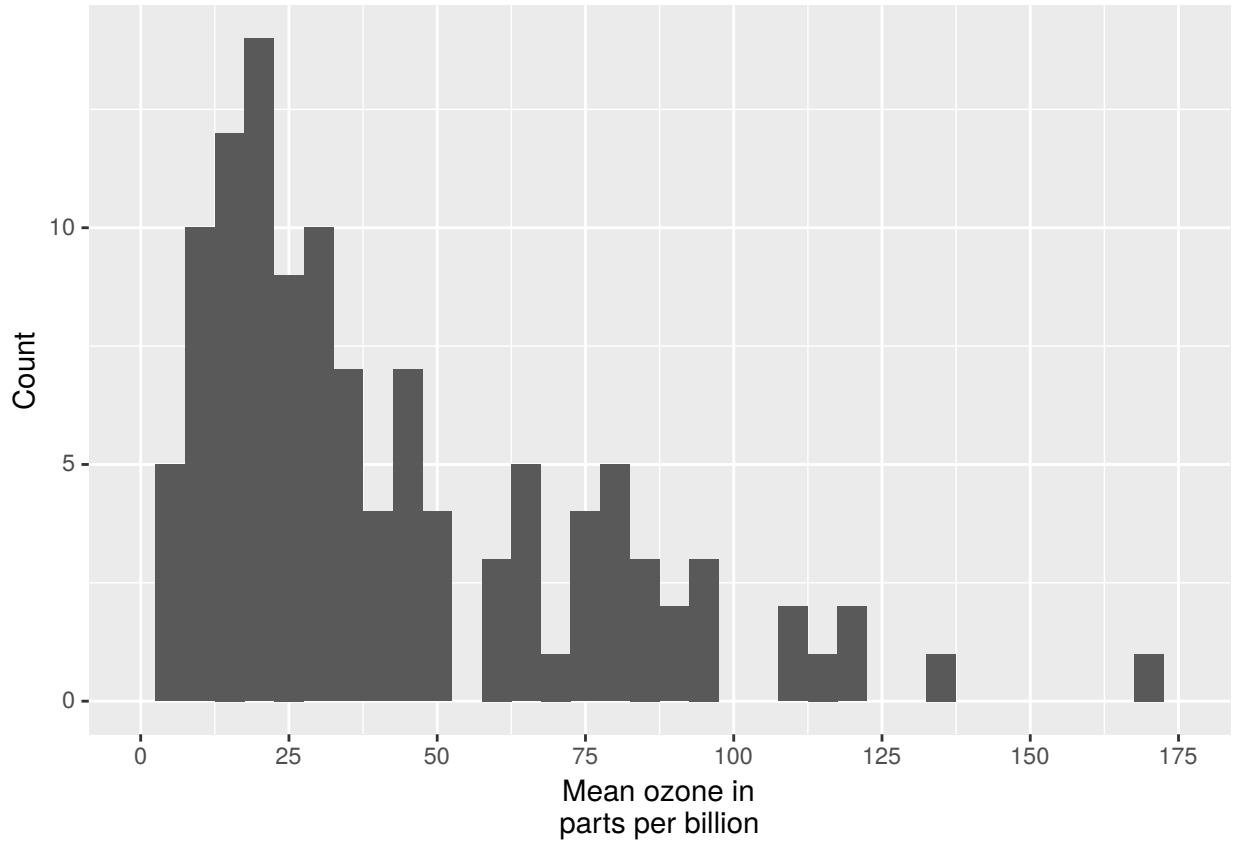
```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion") +
  scale_y_continuous(name = "Count")
p7
```



7.7. Changing axis ticks

The next thing we will change is the axis ticks. Let's make the x-axis ticks appear at every 25 units rather than 50 using the `breaks = seq(0, 175, 25)` argument in `scale_x_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 x-axis begins and ends where we want by also adding the argument `limits = c(0, 175)` to `scale_x_continuous`.

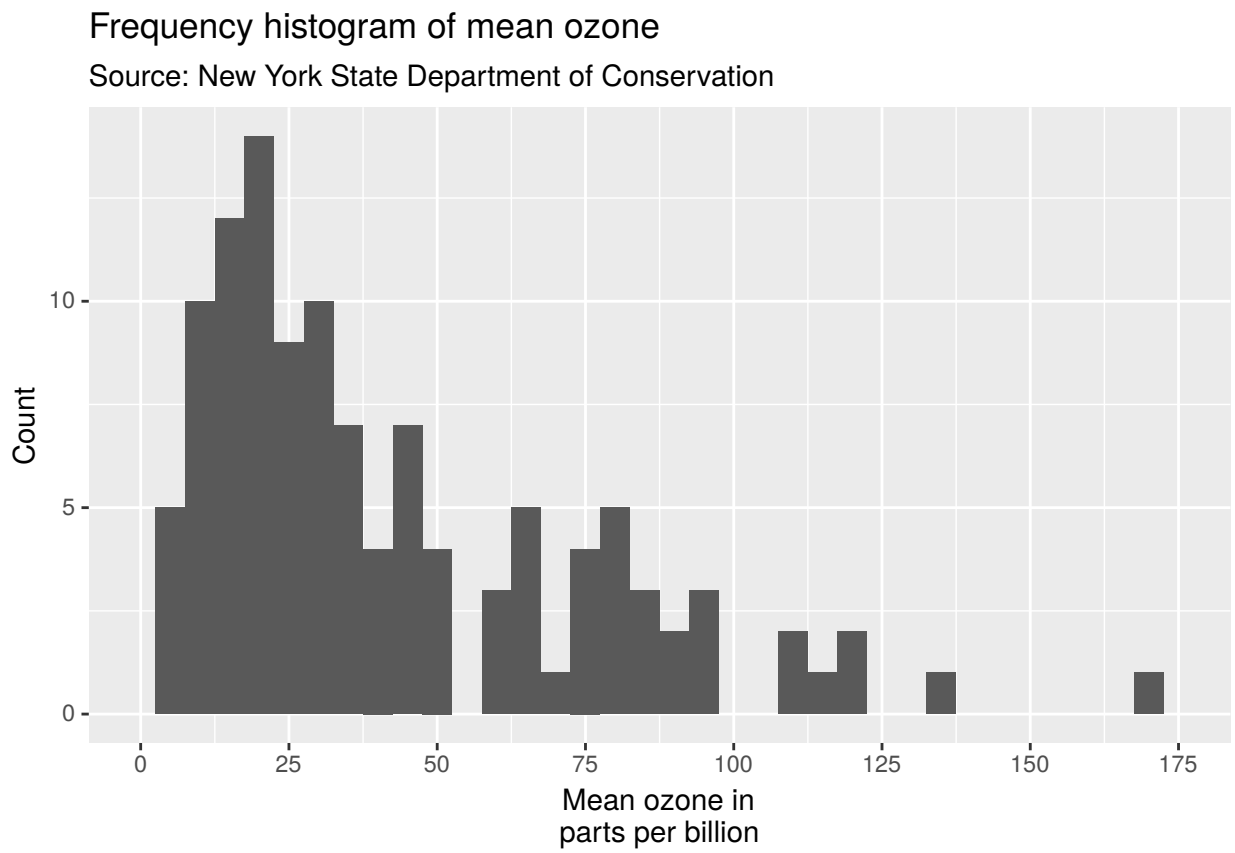
```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count")
p7
```



7.8. Adding a title

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

```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p7
```

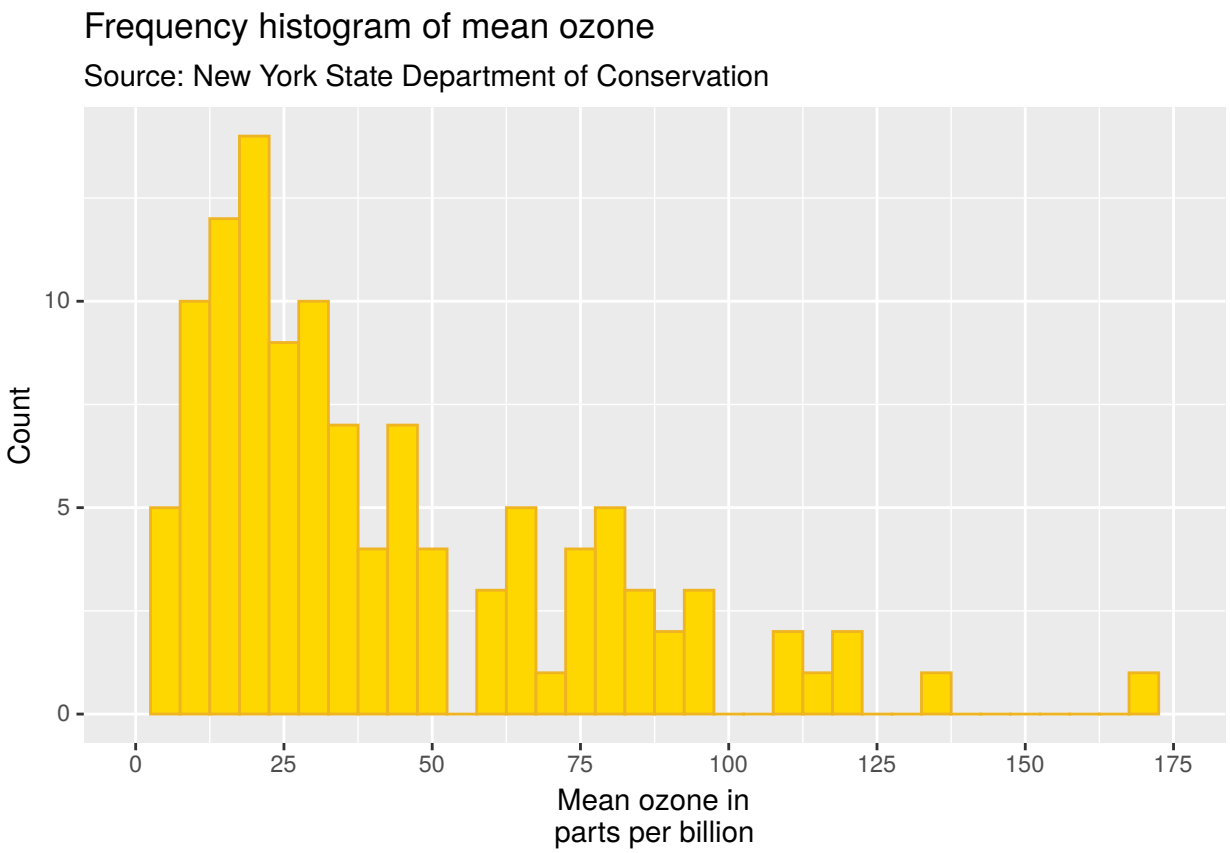


7.9. Changing the colour of the bars

7.9.1. By colour name

To change the line and fill colours of the bars, we add a valid colour to the `colour` and `fill` arguments in `geom_histogram` (note that I assigned these colours to variables outside of the plot to make it easier to change them). A list of valid colours is here.

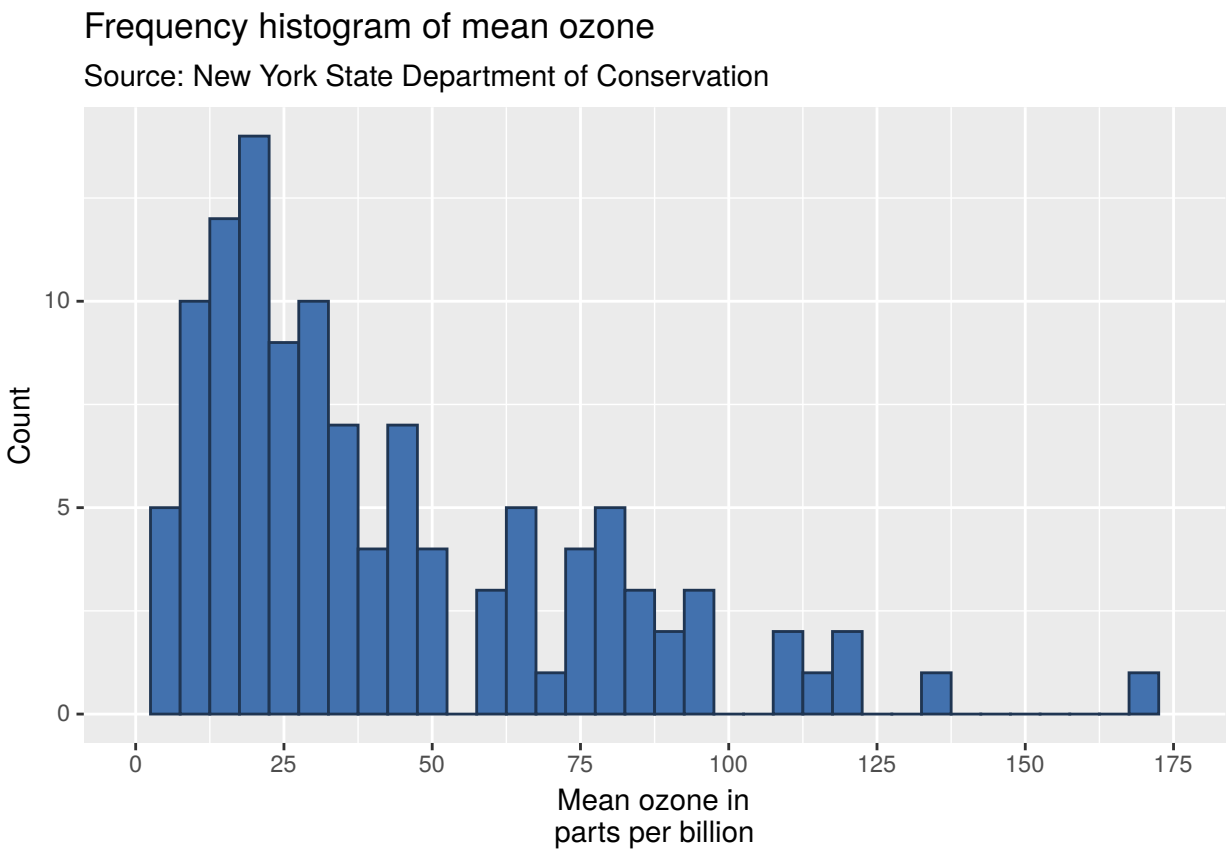
```
barfill = "gold1"; barlines = "goldenrod2"

p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5,
    colour = barlines, fill = barfill) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p7
```



7.9.2. By HEX code

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 two shades of blue for the fill and lines using their HEX codes.

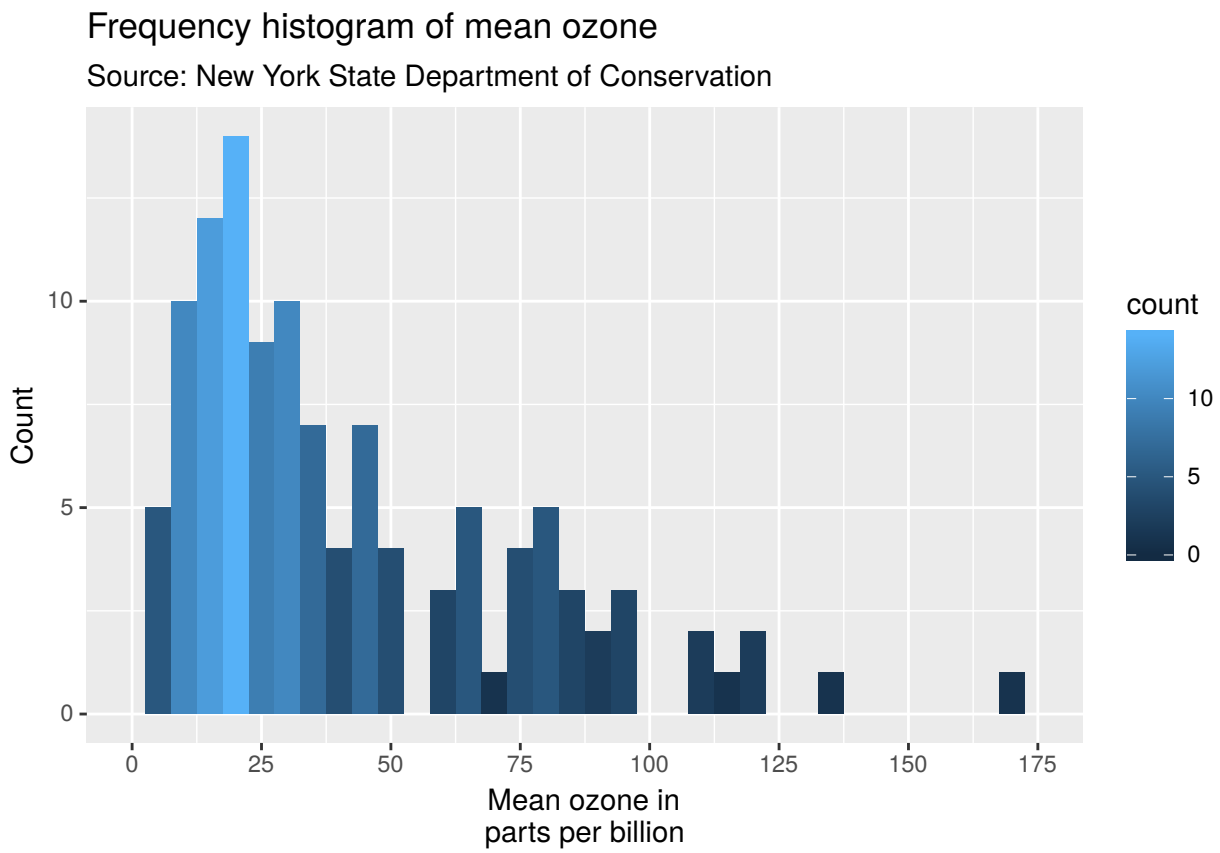
```
barfill <- "#4271AE"; barlines <- "#1F3552"

p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5,
    colour = barlines, fill = barfill) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p7
```



7.10. Colour gradients

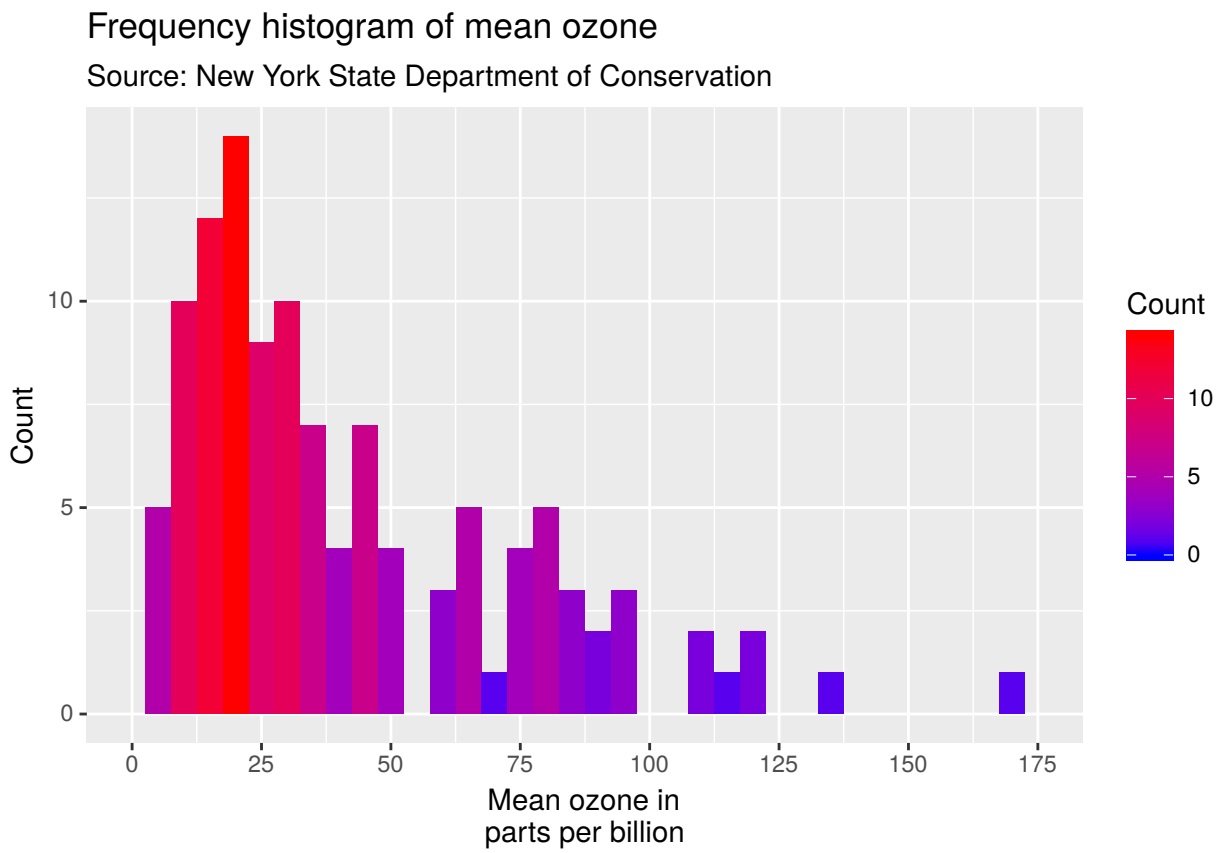
You can also add a gradient to your colour scheme that varies according to the frequency of the values. Below is the default gradient colour scheme. In order to do this, you can see we have changed the `aes(y = ..count..)` argument in `geom_histogram` to `aes(fill = ..count..)`.

```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(fill = ..count..), binwidth = 5) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p7
```



You can customise the gradient by changing the anchoring colours for high and low. To do so, we have added the option `scale_fill_gradient` to the plot with the arguments `Count` (the name of the legend), `low` (the colour for the least frequent values) and `high` (the colour for the most frequent values).

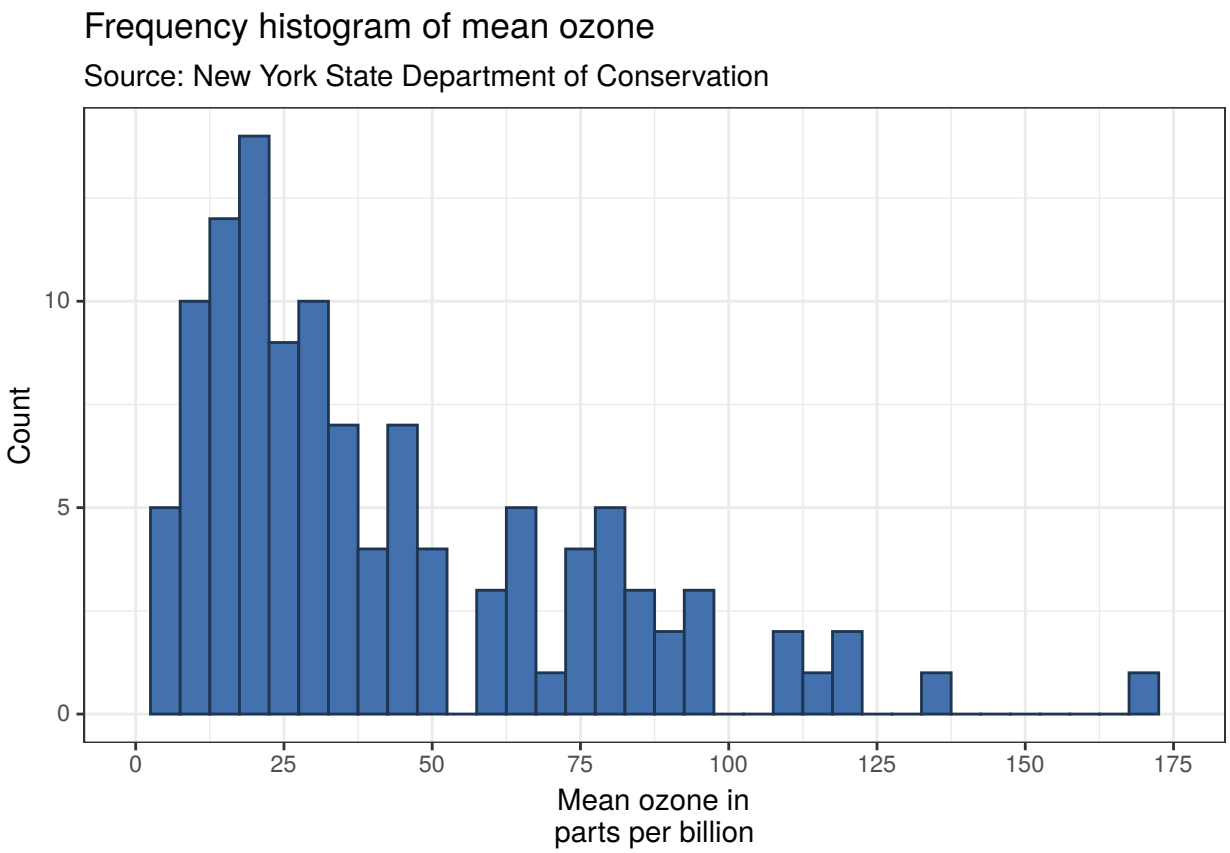
```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(fill = ..count..), binwidth = 5) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  scale_fill_gradient("Count", low = "blue", high = "red")
p7
```



7.11. 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()` after `ggplot()`. As you can see, we can further tweak the graph using the `theme` option, which we've used so far to change the legend.

```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5,
  colour = barlines, fill = barfill) +
    scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25),
    limits = c(0, 175)) +
    scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  theme_bw()
p7
```



7.12. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

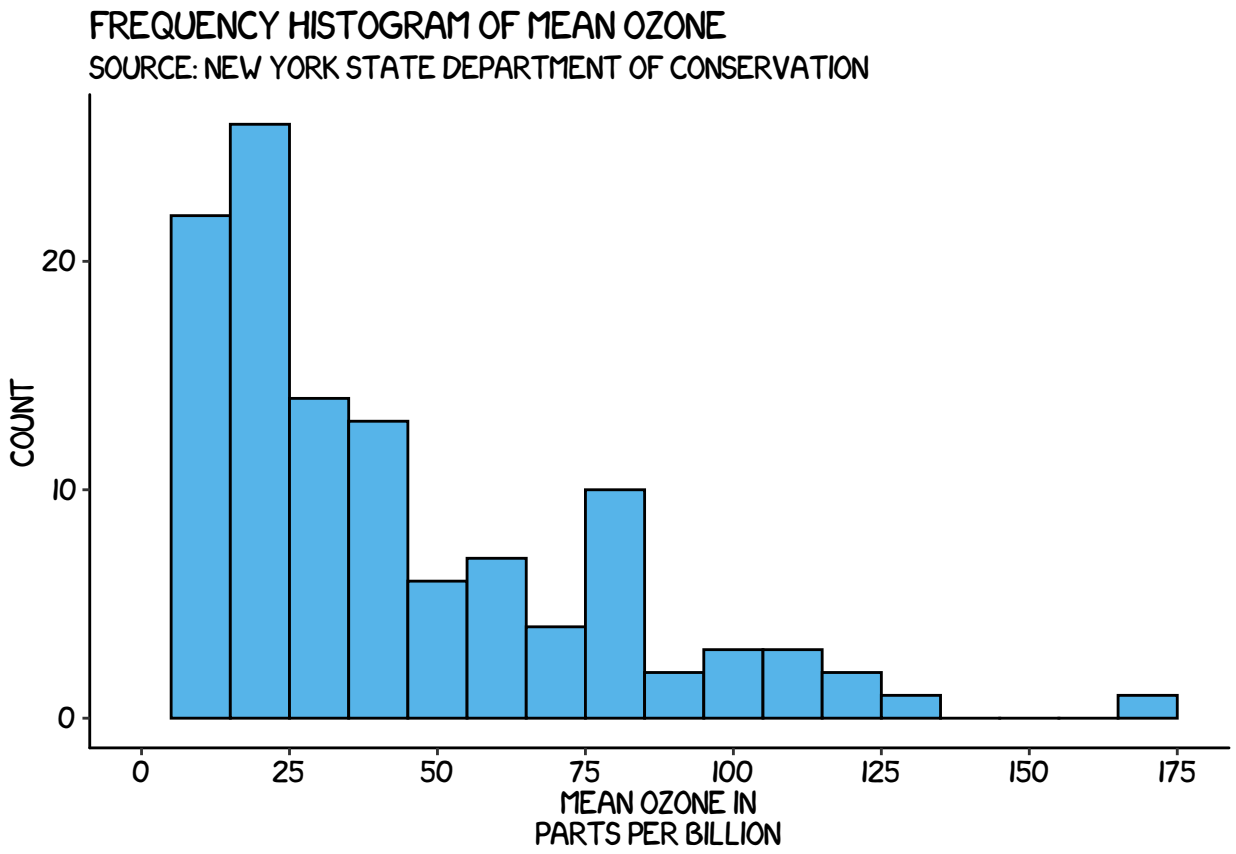
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

7.13. 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.

```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 10,
    colour = "black", fill = "#56B4E9") +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  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.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p7
```

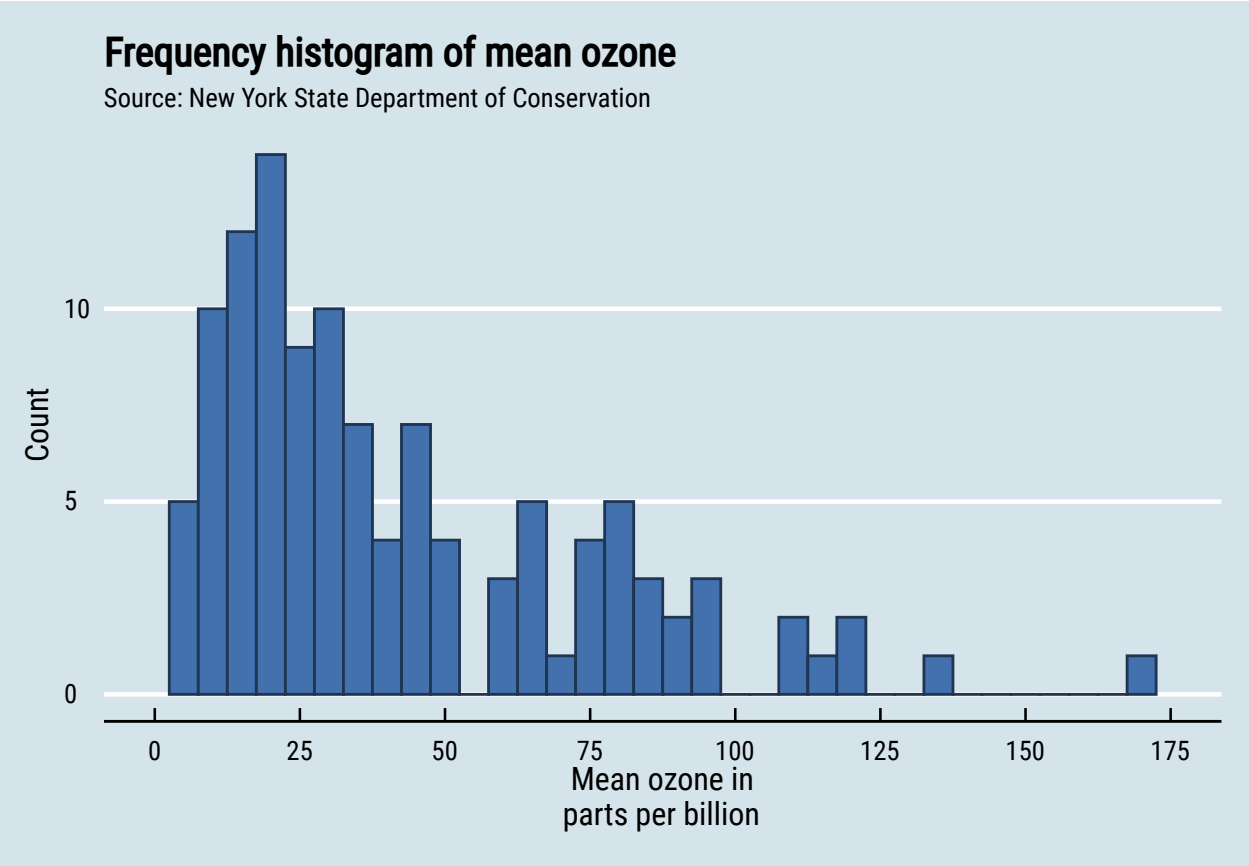


7.14. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5,
    colour = barlines, fill = barfill) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  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 = "Roboto Condensed"),
    plot.title = element_text(family = "Roboto Condensed"))
p7
```

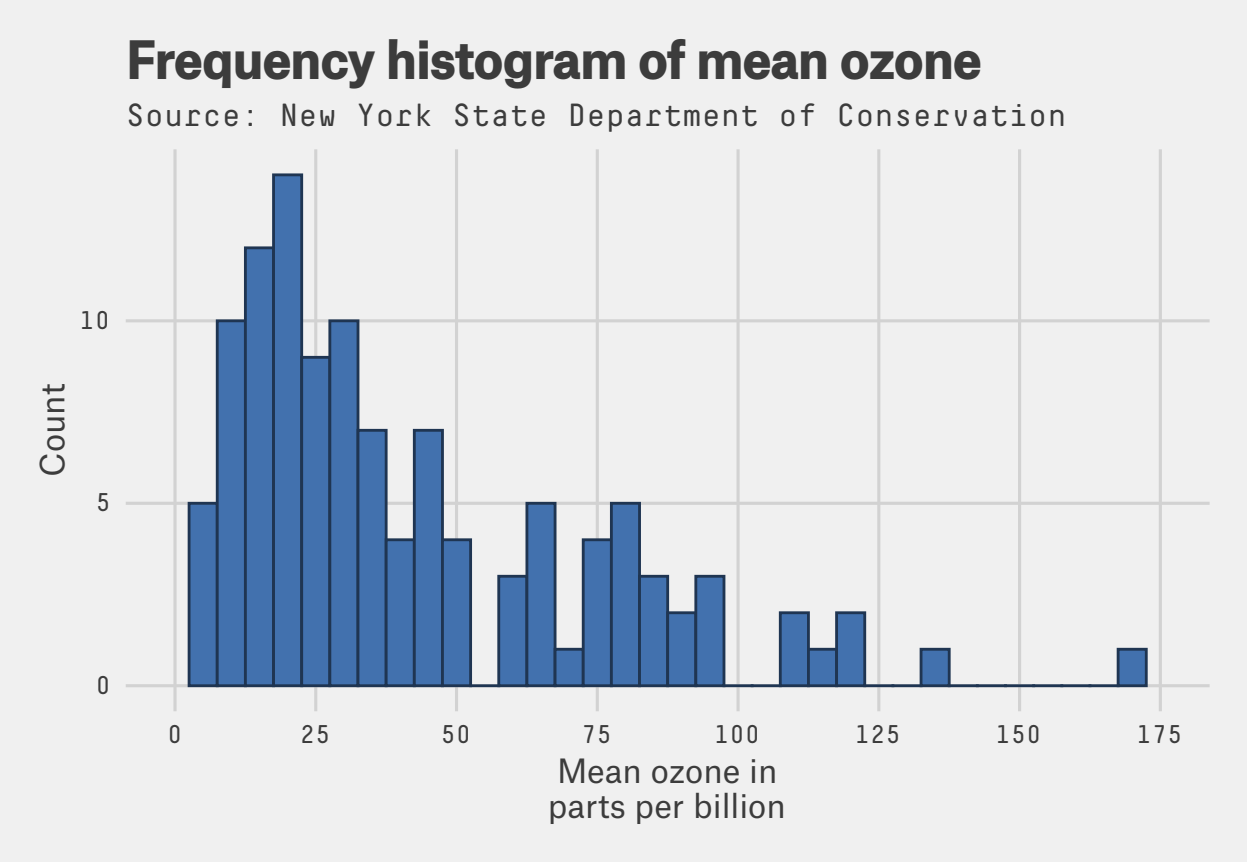


7.15. Using 'Five Thirty Eight' theme

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'.

```
p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5,
    colour = barlines, fill = barfill) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom",
```

```
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 10),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 10),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    text = element_text(family = "Decima Mono Pro"))
p7
```

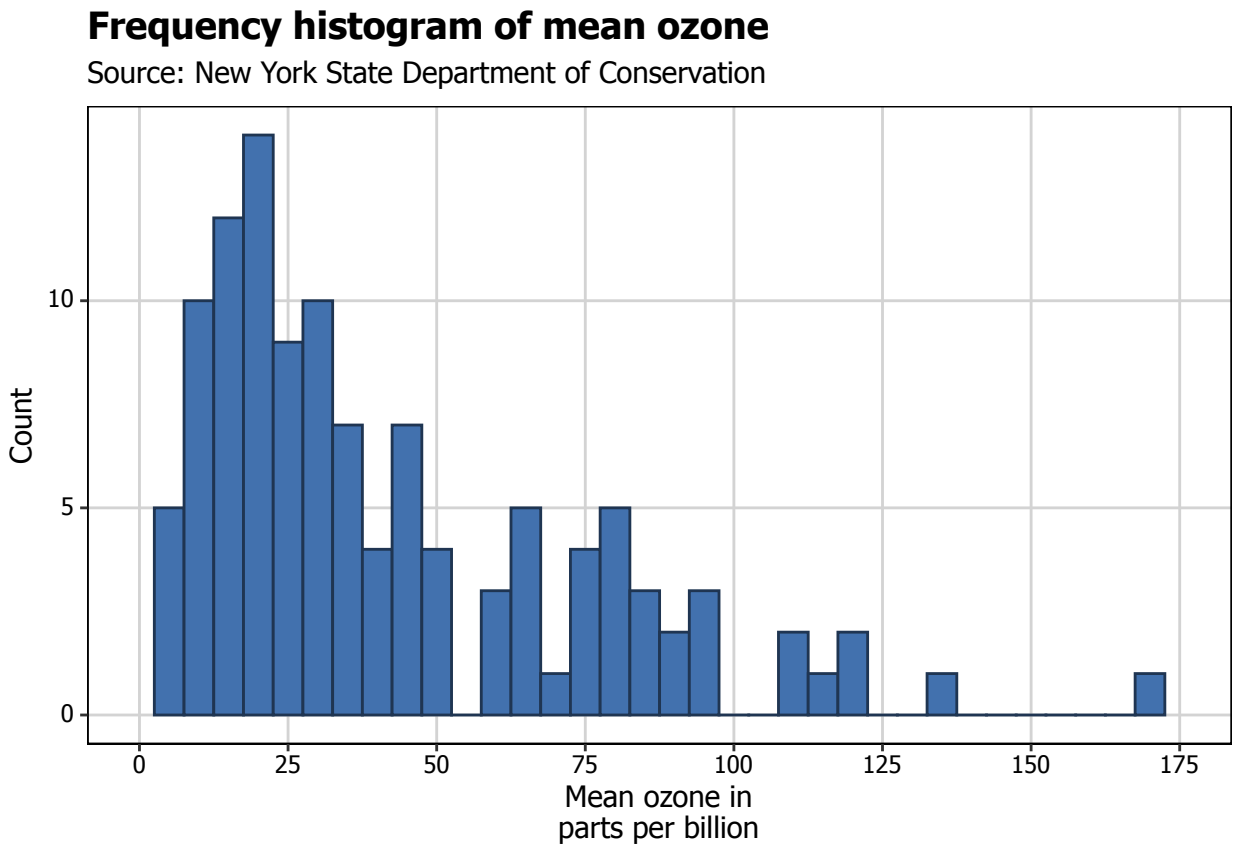


7.16. Creating your own theme

As before, you can modify your plots a lot as `ggplot2` allows many customisations. Here is a custom plot where we have modified the axes, background and font.

```
barfill <- "#4271AE"; barlines <- "#1F3552"

p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5,
    colour = barlines, fill = barfill) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25),
    limits = c(0, 175)) +
```

```
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.key = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p7
```

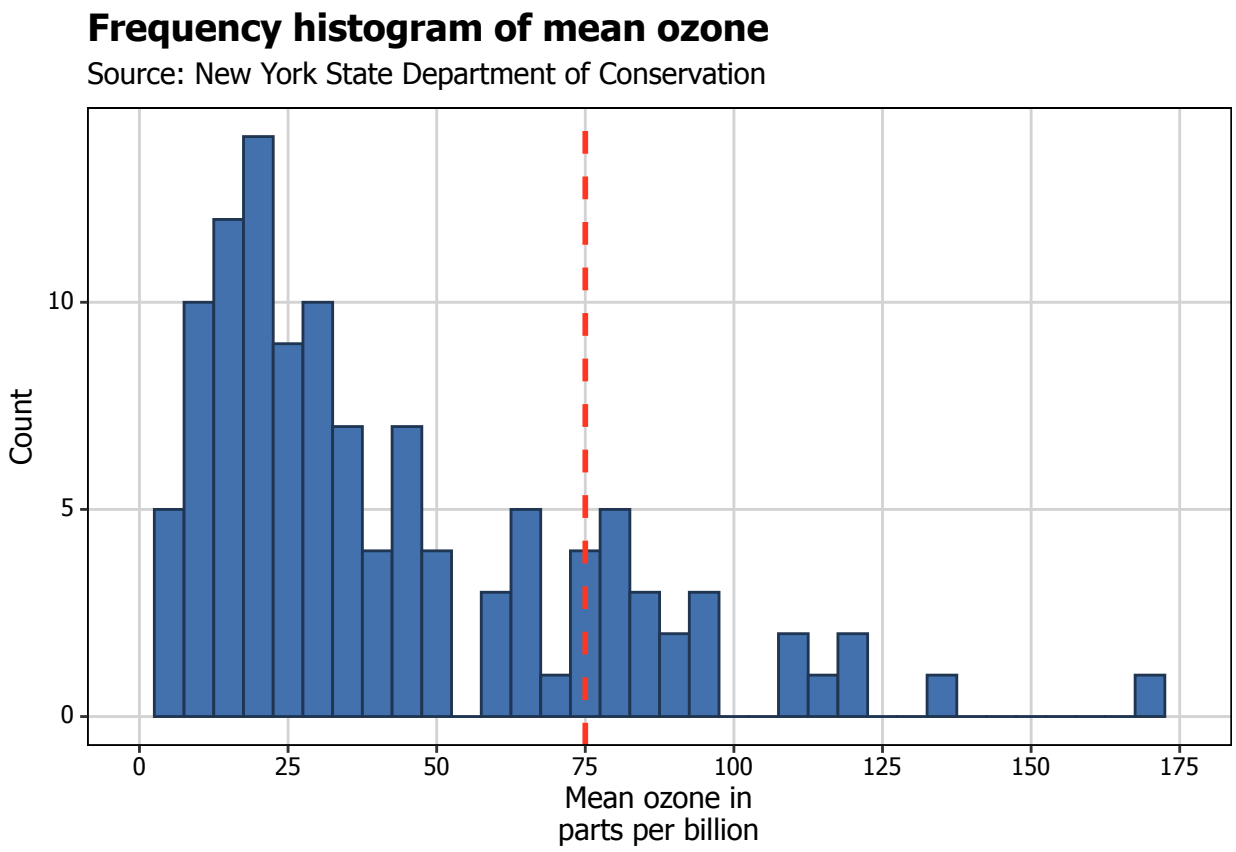


7.17. Adding lines

Let's say that we want to add a cutoff value to the chart (75 parts of ozone per billion). We add the `geom_vline` option to the chart, and specify where it goes on the x-axis using the `xintercept` argument.

We can customise how it looks using the `colour` and `linetype` arguments in `geom_vline`. (In the the same way, horizontal lines can be added using the `geom_hline`.)

```
barfill <- "#4271AE"; barlines <- "#1F3552"

p7 <- ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(aes(y = ..count..), binwidth = 5,
    colour = barlines, fill = barfill) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  geom_vline(xintercept = 75, size = 1, colour = "#FF3721",
    linetype = "dashed") +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.key = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p7
```



7.18. Multiple histograms

You can also easily create multiple histograms by the levels of another variable. There are two options, in separate (panel) plots, or in the same plot.

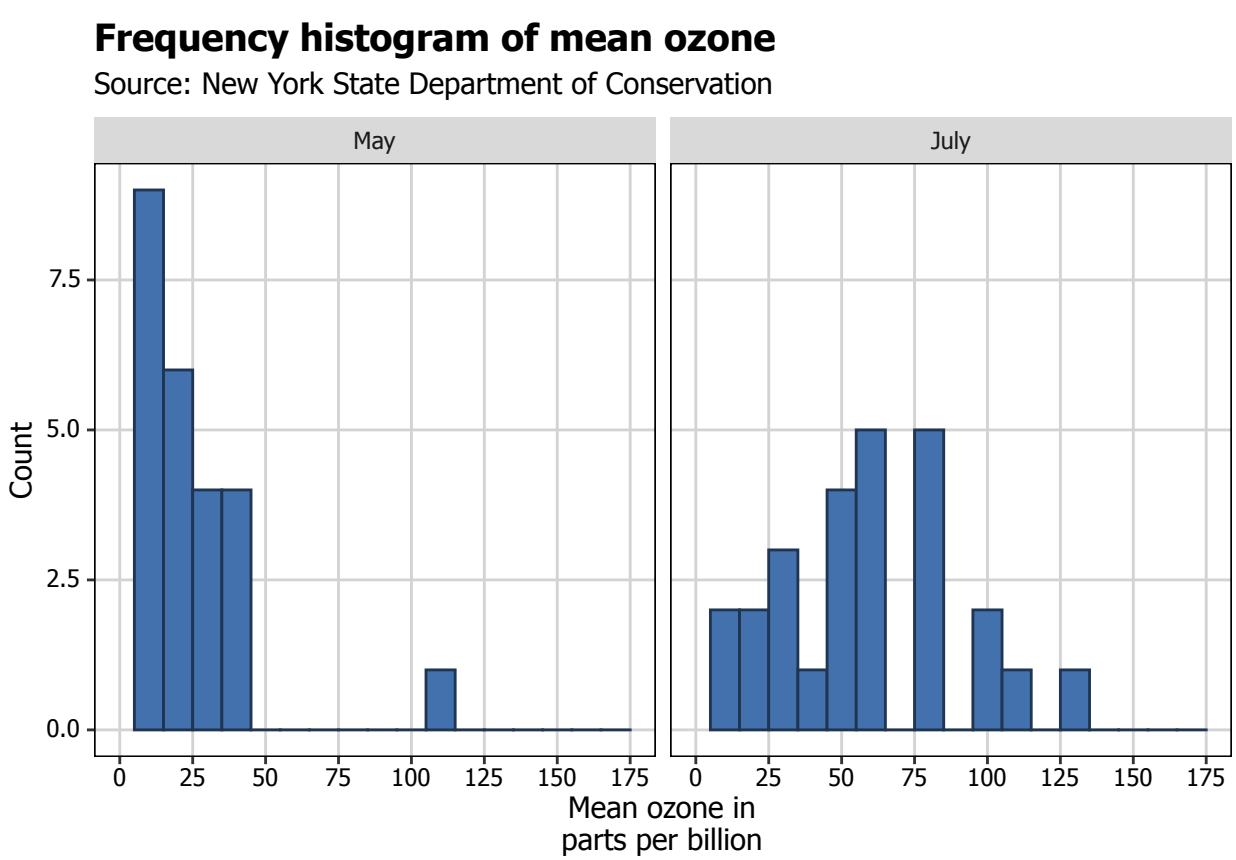
7.18.1. In panel plots

We first need to do a little data wrangling. In order to make the graphs a bit clearer, we've kept only months "5" (May) and "7" (July) in a new dataset `airquality_trimmed`. We also need to convert this variable into either a character or factor variable. We have created a new factor variable `Month.f`.

In order to produce a panel plot by month, we add the `facet_grid(. ~ Month.f)` option to the plot. The additional `scale = free` argument in `facet_grid` means that the y-axes of each plot do not need to be the same.

```
airquality_trimmed <- airquality %>%
  filter(Month %in% c(5,7)) %>%
  mutate(Month.f = factor(Month,
                          labels = c("May", "July")))

p7 <- ggplot(airquality_trimmed, aes(x = Ozone)) +
```

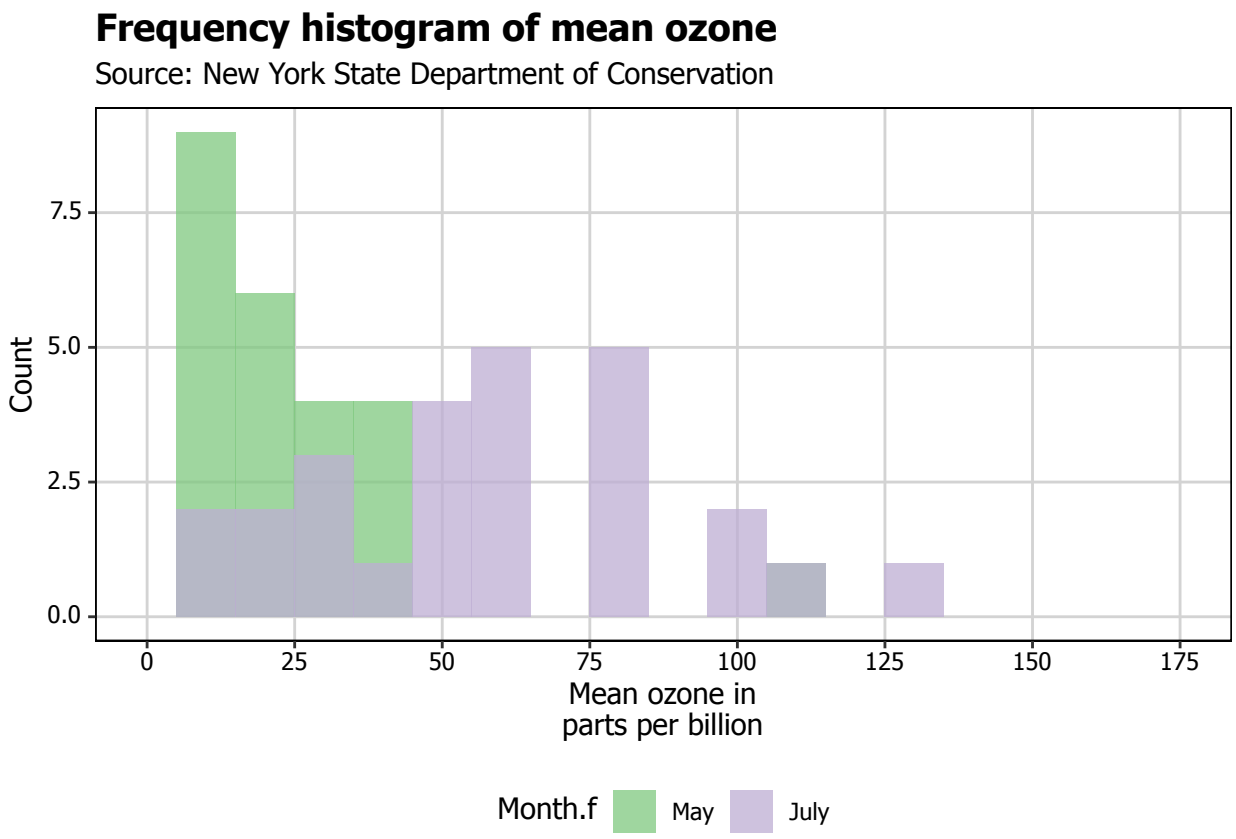
```
  geom_histogram(aes(y = ..count..), binwidth = 10,
    colour = barlines, fill = barfill) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  facet_grid(. ~ Month.f, scales = "free") +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.key = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p7
```



7.18.2. In the same plot

In order to plot the two months in the same plot, we add several things. Firstly, in the `ggplot` function, we add a `fill = Month.f` argument to `aes`. Secondly, in order to more clearly see the graph, we add two arguments to the `geom_histogram` option, `position = "identity"` and `alpha = 0.75`. This controls the position and transparency of the curves respectively. Finally, you can customise the colours of the histograms by adding the `scale_fill_brewer` to the plot from the `RColorBrewer` package. This blog post describes the available packages.

```
p7 <- ggplot(airquality_trimmed, aes(x = Ozone, fill = Month.f)) +
  geom_histogram(aes(y = ..count..), binwidth = 10,
    position = "identity", alpha = 0.75) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  scale_fill_brewer(palette = "Accent") +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.key = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p7
```

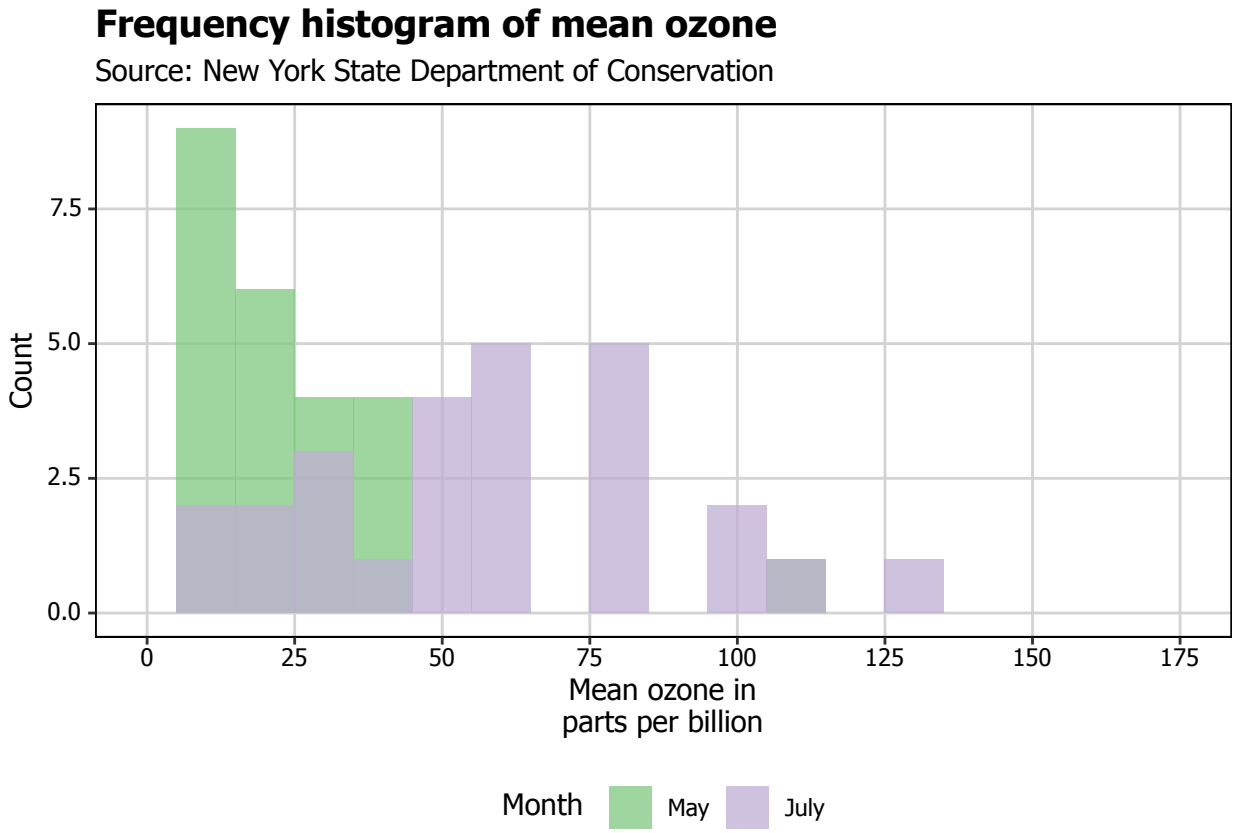


7.19. Formatting the legend

Finally, we can format the legend. Firstly, we can change the position by adding the `legend.position = "bottom"` argument to the `theme` option, which moves the legend under the plot. Secondly, we can fix the title by adding the `labs(fill = "Month ")` option to the plot.

```
p7 <- ggplot(airquality_trimmed, aes(x = Ozone, fill = Month.f)) +
  geom_histogram(aes(y = ..count..), binwidth = 10,
    position = "identity", alpha = 0.75) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_y_continuous(name = "Count") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  scale_fill_brewer(palette = "Accent") +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.key = element_blank(),
```

```
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma")) +
  labs(fill = "Month ")
p7
```

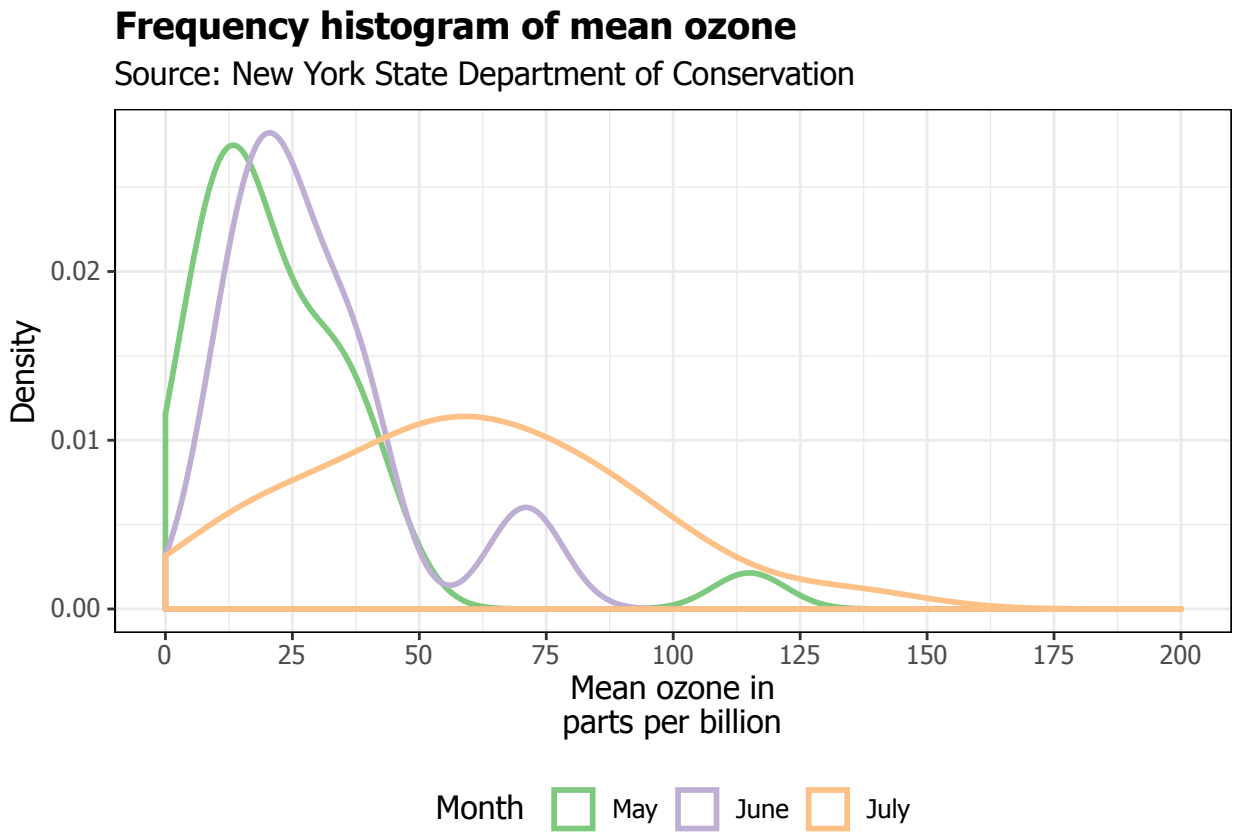


CHAPTER 8

Density plots

8.1. Introduction

In this chapter, we will work towards creating the density plot below. We will take you from a basic density plot and explain all the customisations we add to the code step-by-step.



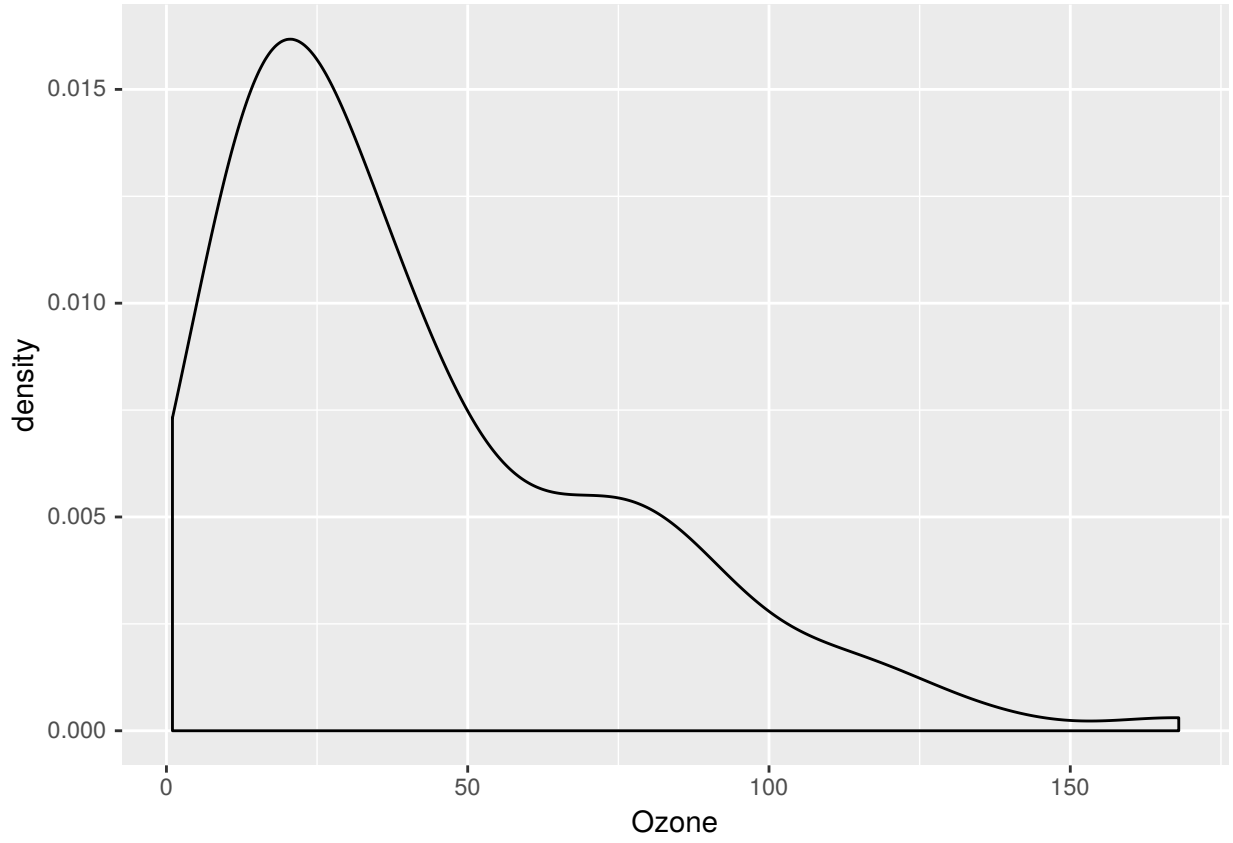
The first thing to do is load in the data and the libraries, as below:

```
if (!require("pacman")) install.packages("pacman")
p_load(datasets, ggplot2, ggthemes, dplyr, RColorBrewer, grid)

data(airquality)
```

8.2. Basic density plot

In order to initialise a plot we tell ggplot that `airquality` is our data, and specify that our x axis plots the `Ozone` variable. We then instruct ggplot to render this as a density plot by adding the `geom_density()` option.

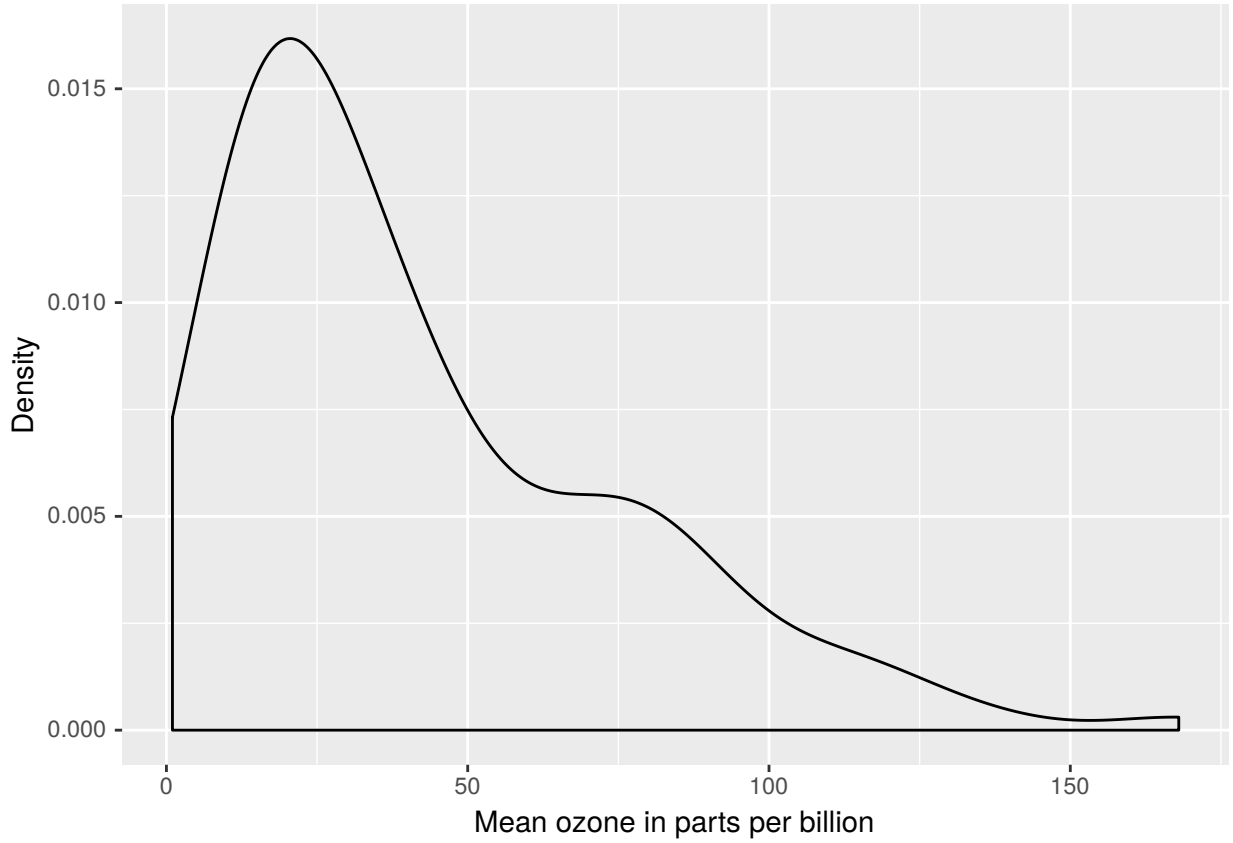
```
p8 <- ggplot(airquality, aes(x = Ozone)) + geom_density()
p8
```



8.3. Customising axis labels

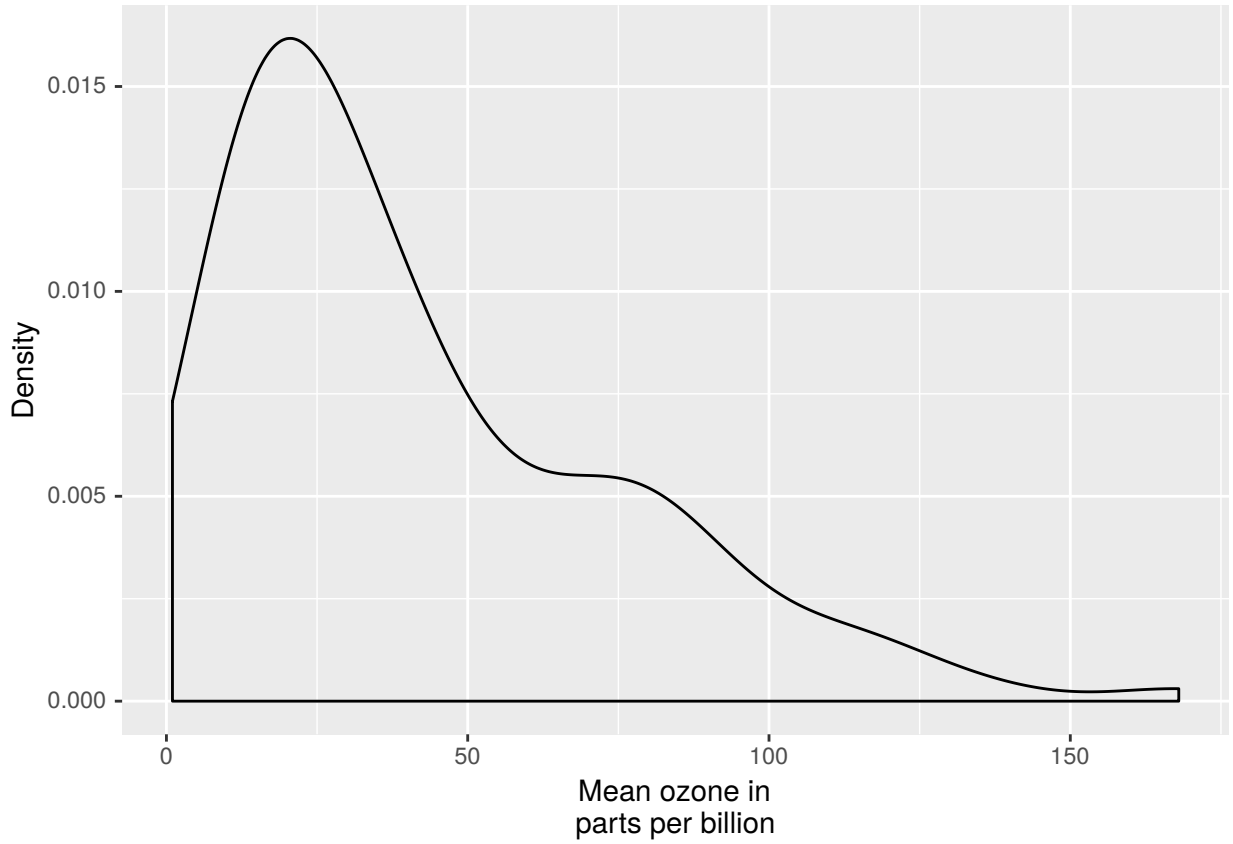
In order to change the axis labels, we have a couple of options. In this case, we have used the `scale_x_continuous` and `scale_y_continuous` options, as these have further customisation options for the axes we will use below. In each, we add the desired name to the `name` argument as a string.

```
p8 <- p8 + scale_x_continuous(name = "Mean ozone in parts per billion") +
    scale_y_continuous(name = "Density")
p8
```



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

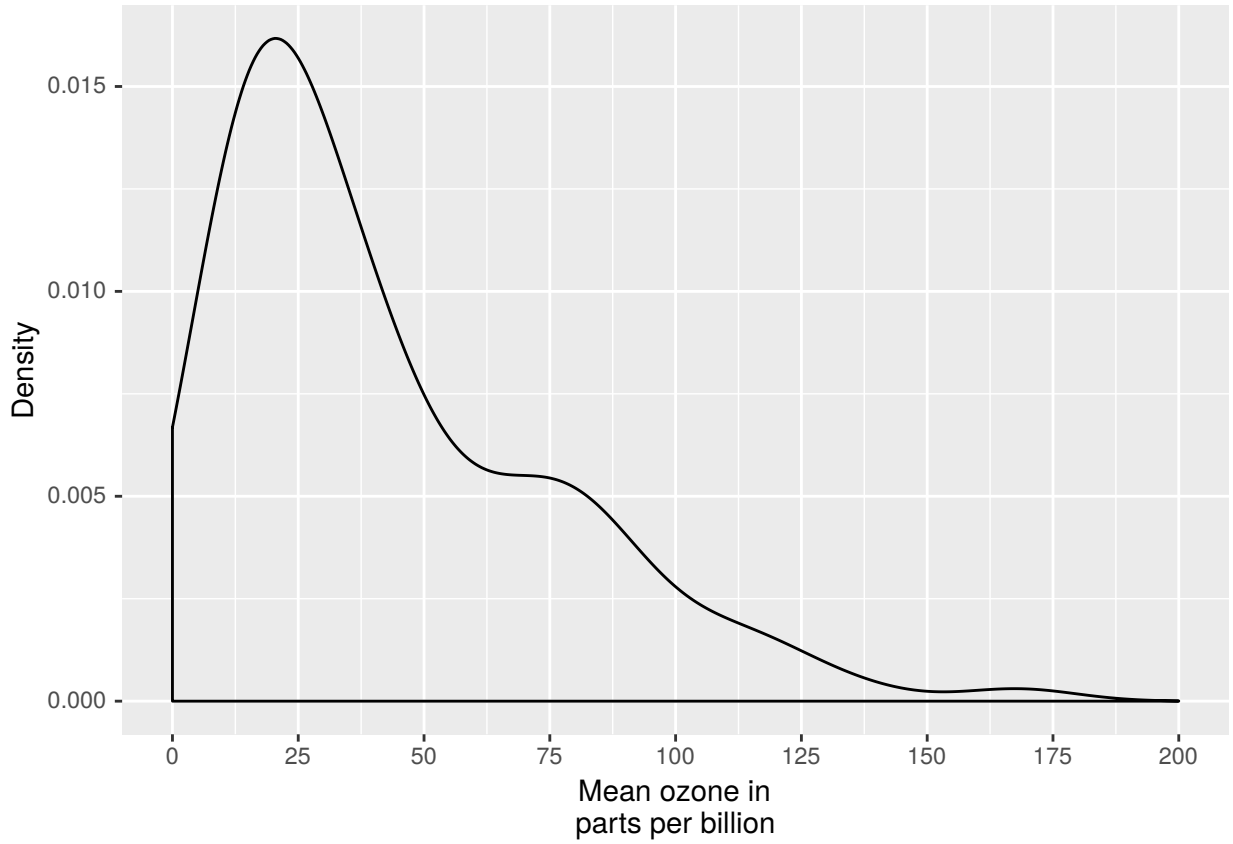
```
p8 <- p8 + scale_x_continuous(name = "Mean ozone in\nparts per billion")
p8
```



8.4. Changing axis ticks

The next thing we will change is the axis ticks. Let's make the x-axis ticks appear at every 25 units rather than 50 using the `breaks = seq(0, 200, 25)` argument in `scale_x_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 x-axis begins and ends where we want by also adding the argument `limits = c(0, 200)` to `scale_x_continuous`.

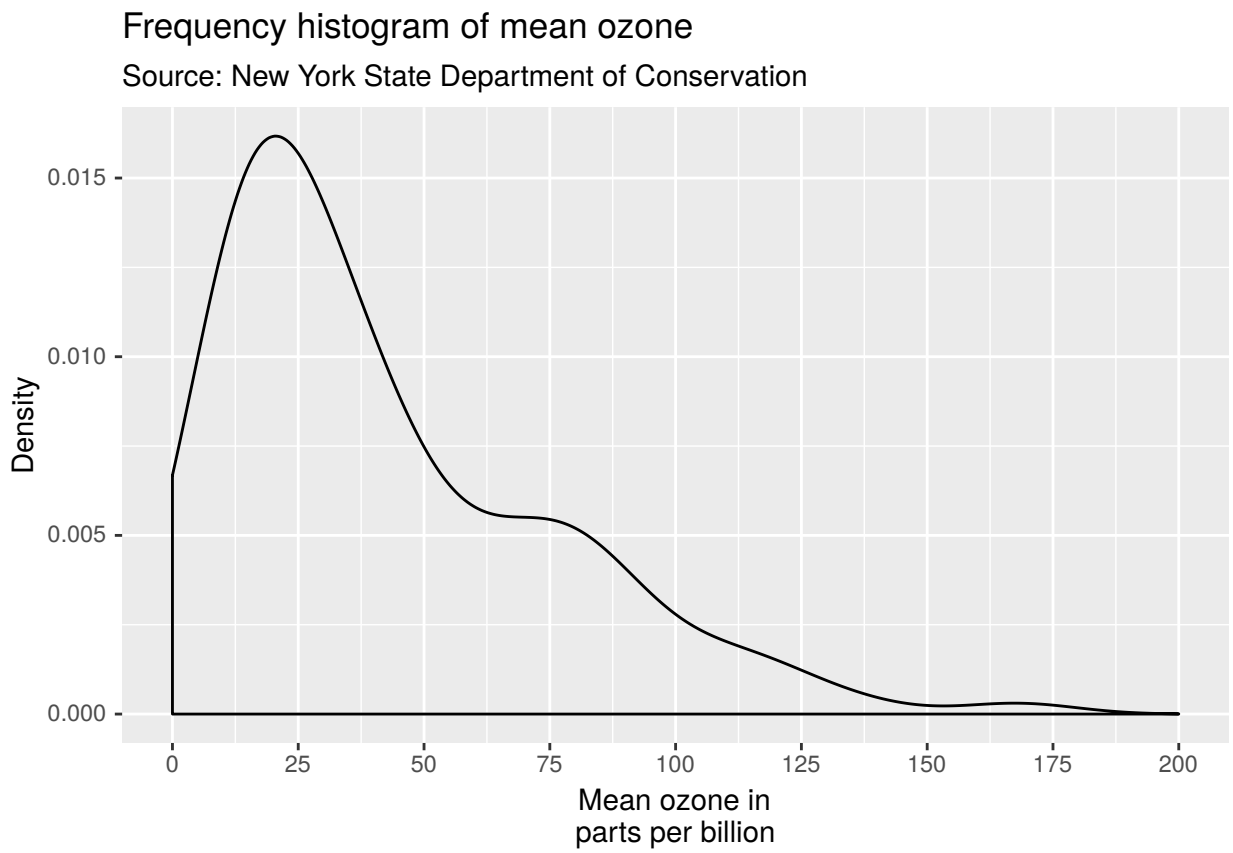
```
p8 <- p8 + scale_x_continuous(name = "Mean ozone in\nparts per billion",
     breaks = seq(0, 200, 25), limits = c(0, 200))
p8
```



8.5. Adding a title

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

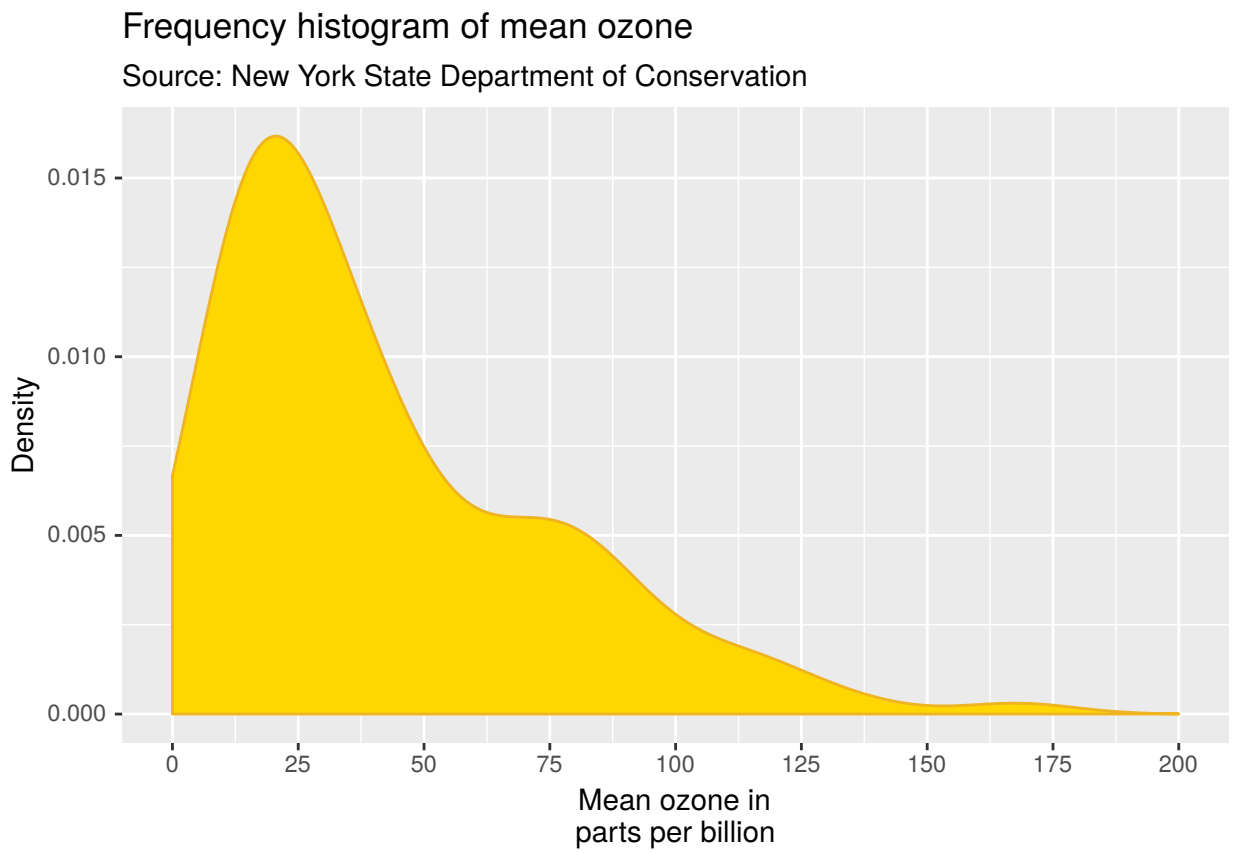
```
p8 <- p8 +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p8
```



8.6. Changing the colour of the curves

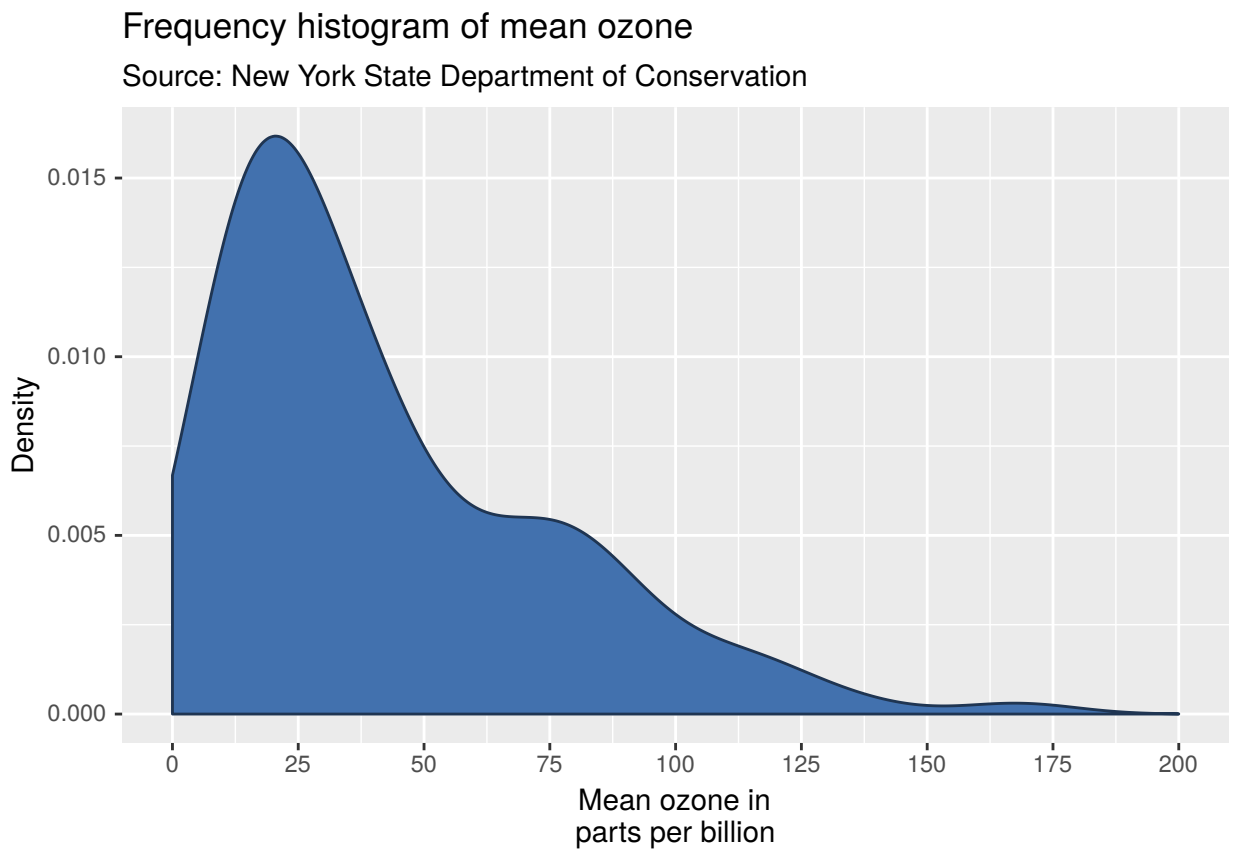
To change the line and fill colours of the density plot, we add a valid colour to the `colour` and `fill` arguments in `geom_density()` (note that I assigned these colours to variables outside of the plot to make it easier to change them). A list of valid colours is here.

```
fill <- "gold1"; line <- "goldenrod2"

p8 <- ggplot(airquality, aes(x = Ozone)) +
  geom_density(fill = fill, colour = line) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p8
```



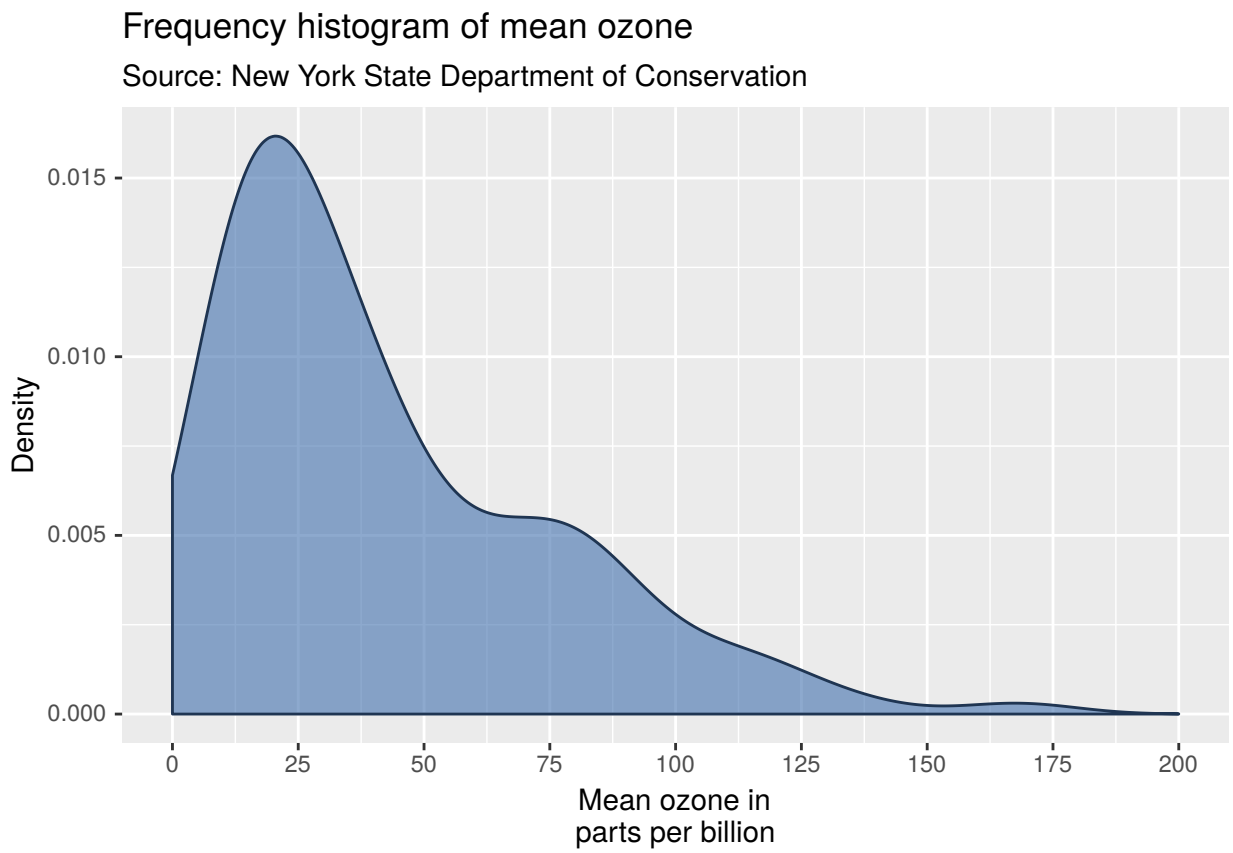
If you want to go beyond the options in the list above, you can also specify exact HEX colours by including them as a string preceded by a hash, e.g., "#FFFFFF". Below, we have called two shades of blue for the fill and lines using their HEX codes.

```
fill <- "#4271AE"; line <- "#1F3552"

p8 <- ggplot(airquality, aes(x = Ozone)) +
  geom_density(fill = fill, colour = line) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p8
```



You can also specify the degree of transparency in the density fill area using the argument `alpha` in `geom_density`. This ranges from 0 to 1.

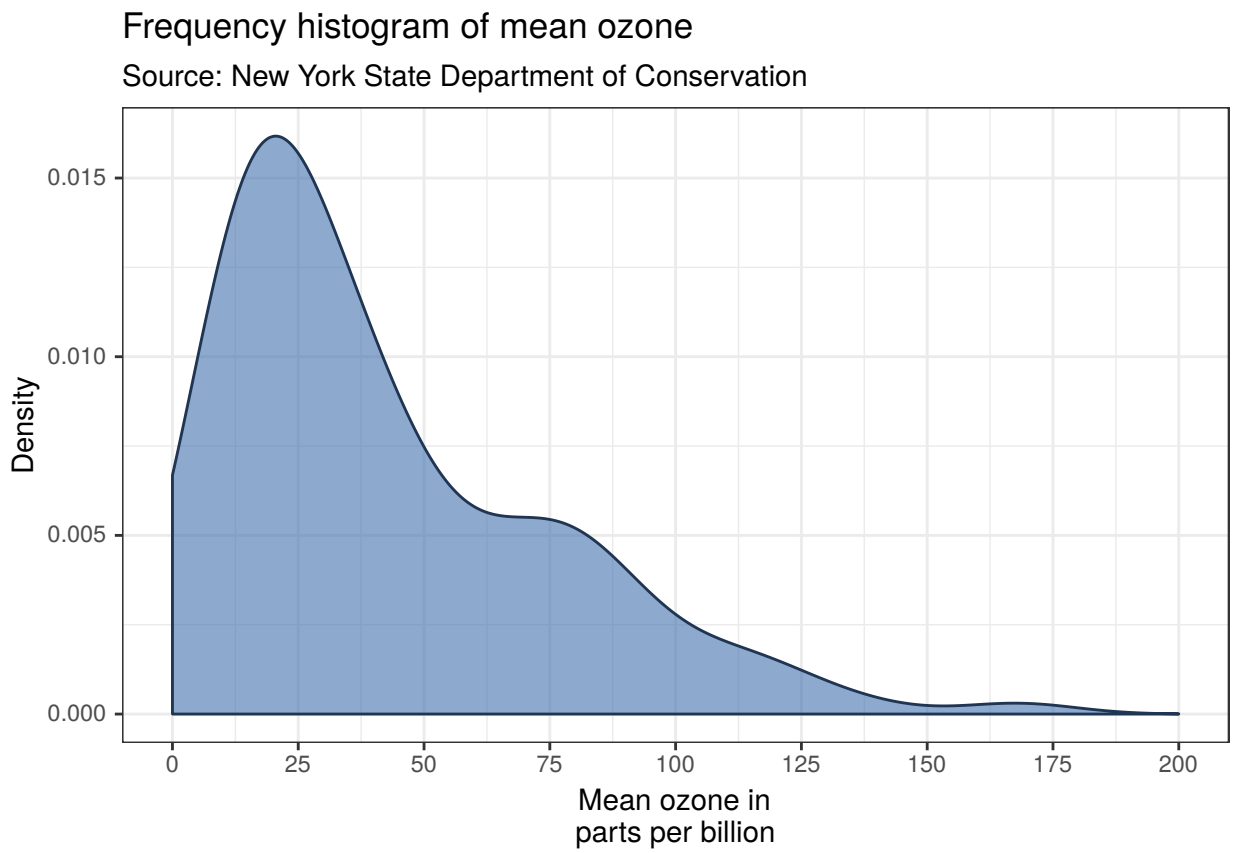
```
p8 = ggplot(airquality, aes(x = Ozone)) +
  geom_density(fill = fill, colour = line, alpha = 0.6) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p8
```



8.7. 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.

```
p8 <- p8 + theme_bw()
p8
```



8.8. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

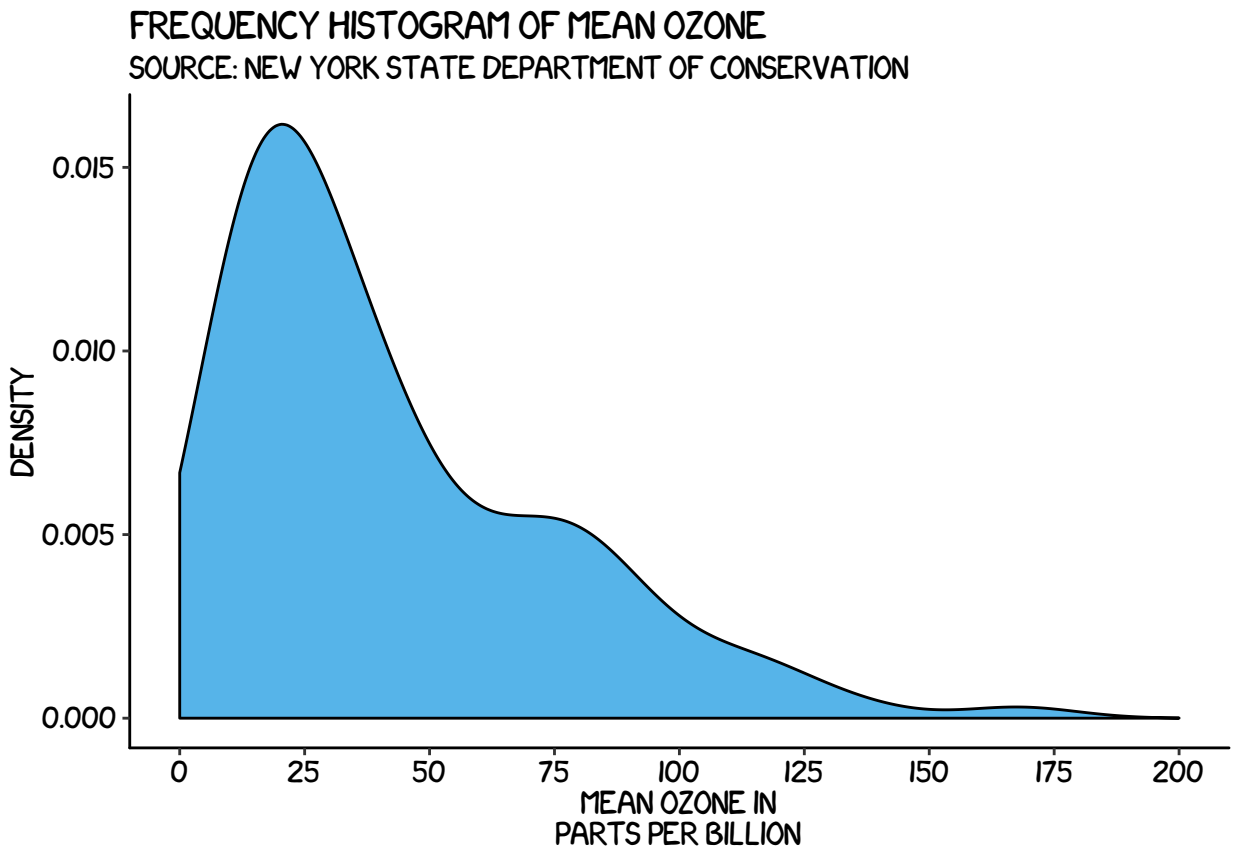
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

8.9. 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.

```
p8 <- ggplot(airquality, aes(x = Ozone)) +
  geom_density(colour = "black", fill = "#56B4E9") +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  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.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p8
```

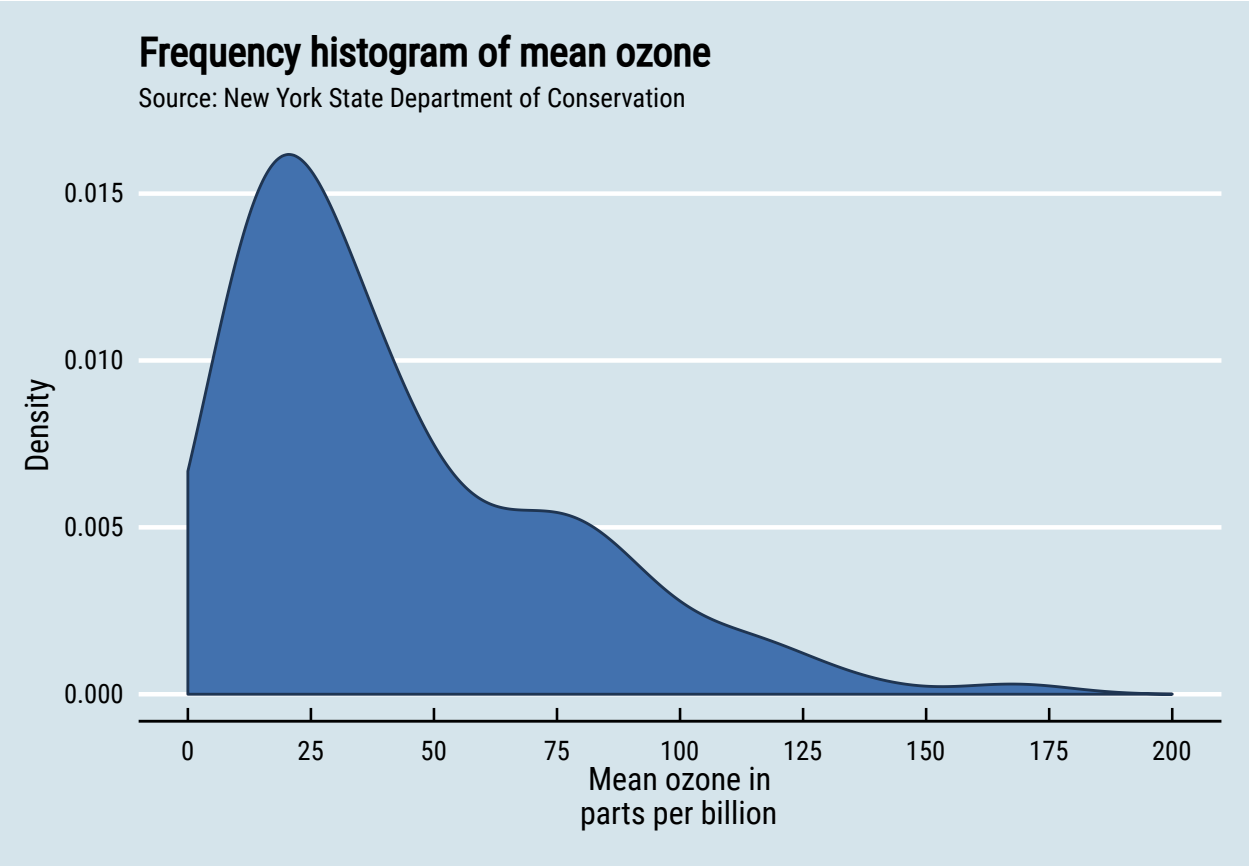


8.10. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
p8 <- ggplot(airquality, aes(x = Ozone)) +
  geom_density(fill = fill, colour = line) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  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 = "Roboto Condensed"),
```

```
    plot.title = element_text(family = "Roboto Condensed"))
p8
```

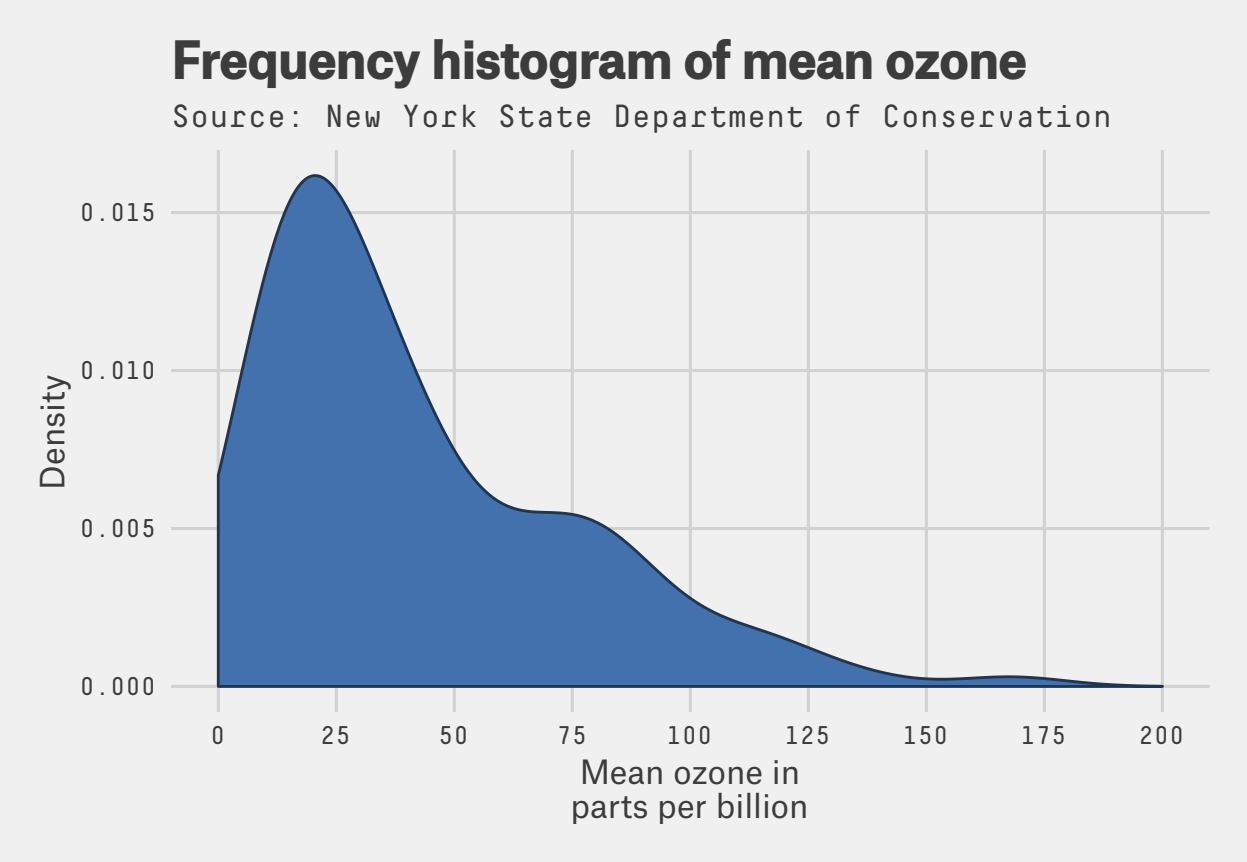


8.11. Using 'Five Thirty Eight' theme

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'.

```
p8 <- ggplot(airquality, aes(x = Ozone)) +
  geom_density(fill = fill, colour = line) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
```

```
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 10),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 10),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    text = element_text(family = "Decima Mono Pro"))
p8
```

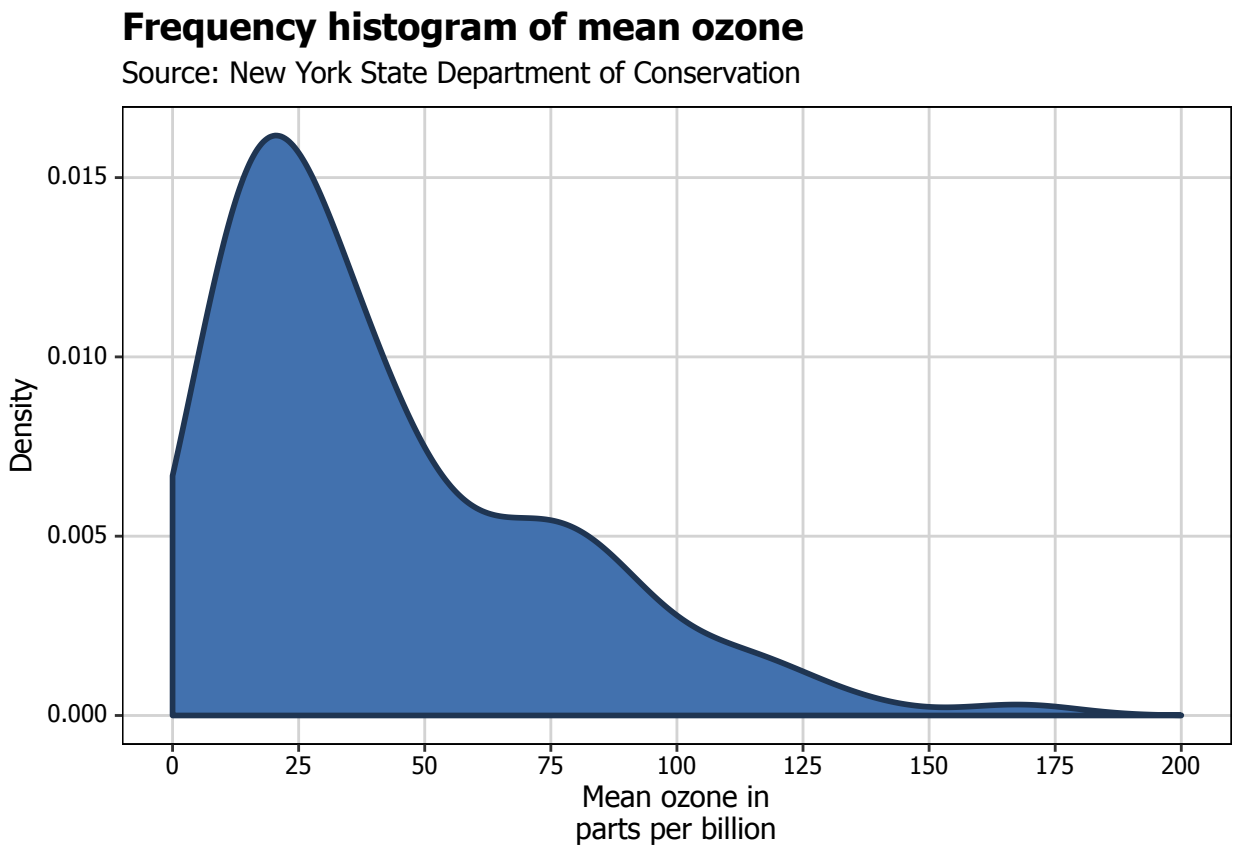


8.12. Creating your own theme

As before, you can modify your plots a lot as `ggplot2` allows many customisations. Here is a custom plot where we have modified the axes, background and font.

```
fill <- "#4271AE"; lines <- "#1F3552"

p8 <- ggplot(airquality, aes(x = Ozone)) +
  geom_density(colour = lines, fill = fill, size = 1) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
```

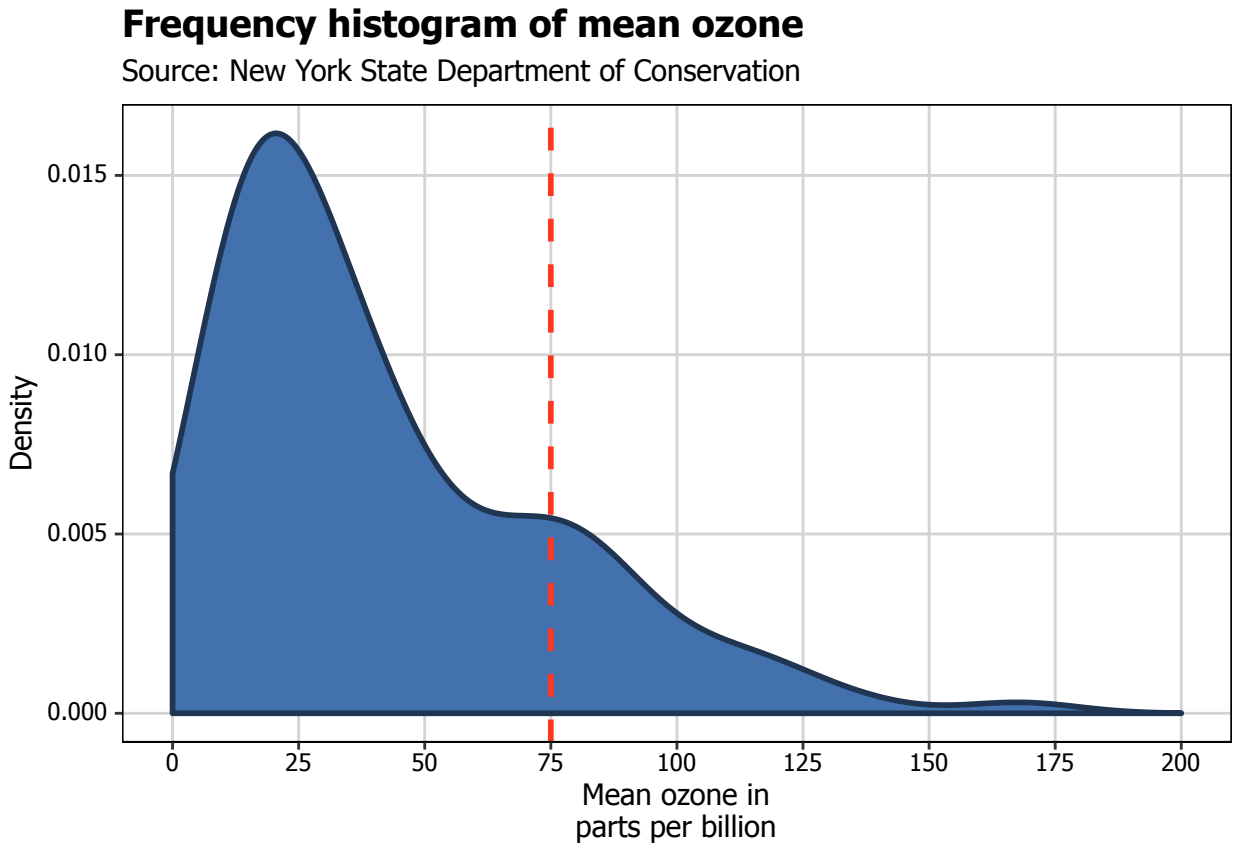
```
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.key = element_blank(),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p8
```



8.13. Adding lines

Let's say that we want to add a cutoff value to the chart (75 parts of ozone per billion). We add the `geom_vline` option to the chart, and specify where it goes on the x-axis using the `xintercept` argument. We can customise how it looks using the `colour` and `linetype` arguments in `geom_vline`. (In the the same way, horizontal lines can be added using the `geom_hline`.)

```
p8 <- p8 + geom_vline(xintercept = 75, size = 1, colour = "#FF3721",
     linetype = "dashed")
p8
```



8.14. Multiple densities

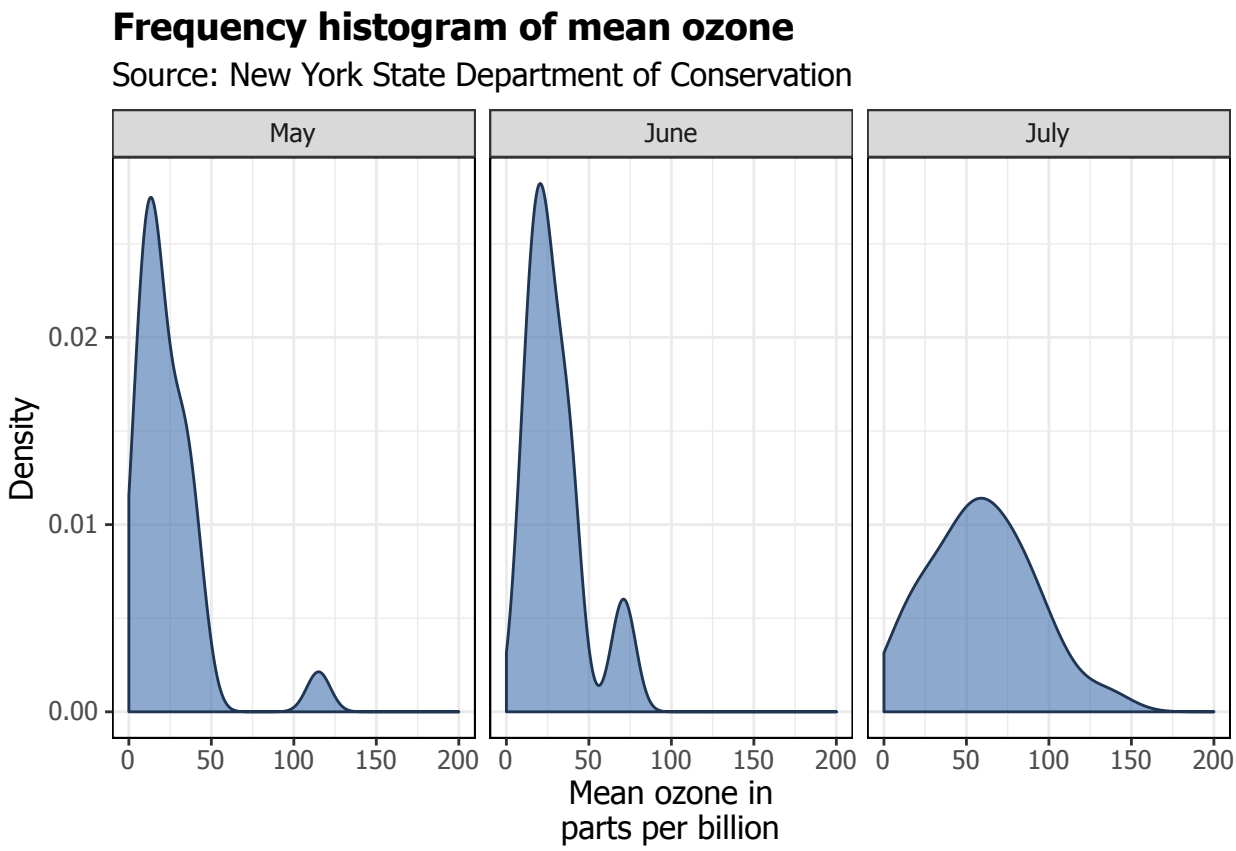
You can also easily create multiple density plots by the levels of another variable. There are two options, in separate (panel) plots, or in the same plot. There are also a couple of variations on these we'll discuss below.

We first need to do a little data wrangling. In order to make the graphs a bit clearer, we've kept only months "5" (May), "6" (June) and "7" (July) in a new dataset `airquality_trimmed`. We also need to convert this variable into either a character or factor variable. We have created a new factor variable `Month.f`.

In order to produce a panel plot by month, we add the `facet_grid(. ~ Month.f)` option to the plot. Note that we've also changed the scale of the x-axis to make it fit a little more neatly in the panel format.

```
airquality_trimmed <- airquality %>%
  filter(Month %in% c(5,6,7)) %>%
  mutate(Month.f = factor(Month,
                          labels = c("May", "June", "July")))
```

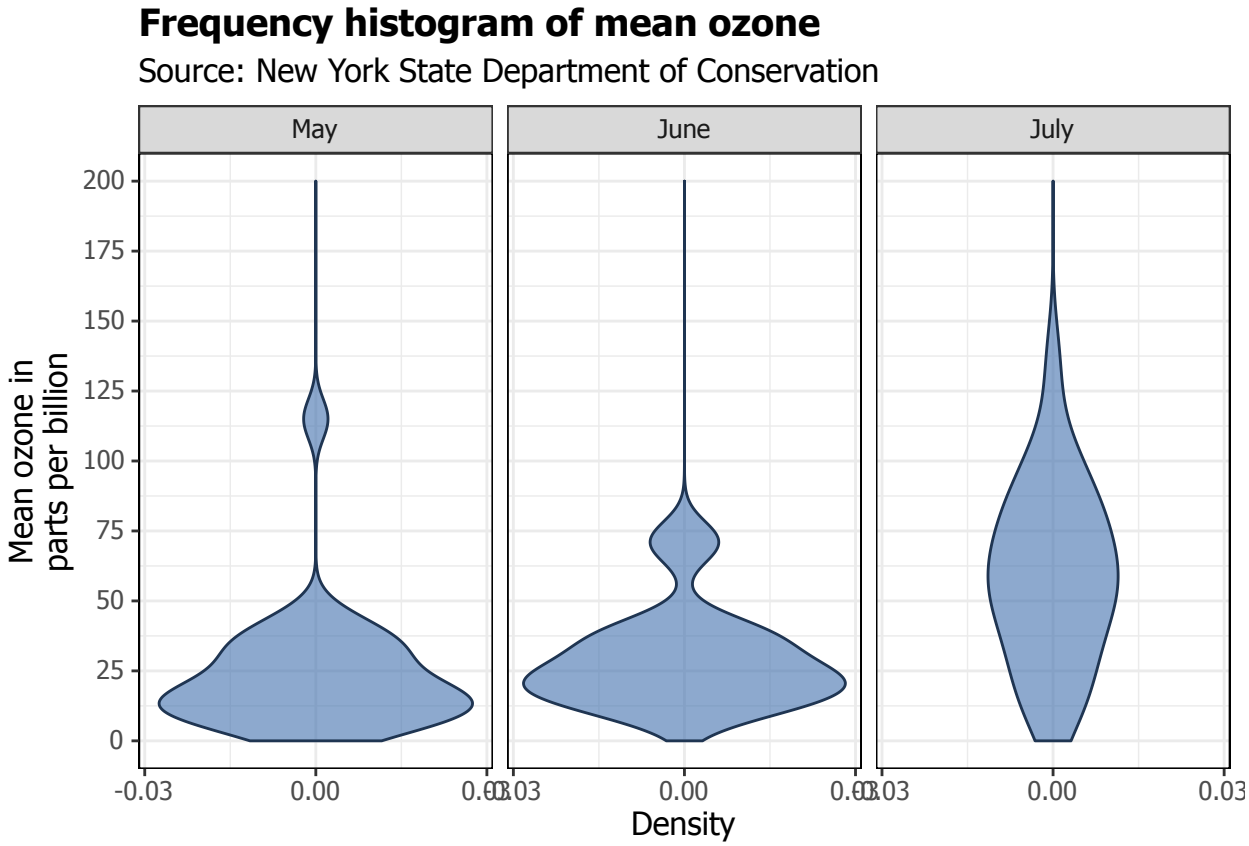
```
p8 <- ggplot(airquality_trimmed, aes(x = Ozone)) +
  geom_density(fill = fill, colour = line, alpha = 0.6) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 50), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  facet_grid(. ~ Month.f) +
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(size = 12, family = "Tahoma"))
p8
```



An alternative to a panel plot is the *volcano plot*. This plot swaps the axes (so the variable of interest is on the y-axis and the density is on the x-axis), and reflects the density. In order to create this plot, we replace `geom_density` with `stat_density`, and include the arguments `aes(ymax = ..density.., ymin = -..density..)` and `geom = "ribbon"` to create a density plot, the usual `fill`, `colour` and `alpha` arguments, and `position = "identity"`. We also need to add a `coord_flip()` option to the plot.

```
p8 <- ggplot(airquality_trimmed, aes(x = Ozone)) +
  stat_density(aes(ymax = ..density..,  ymin = -..density..),
    geom = "ribbon",
    fill = fill, colour = line, alpha = 0.6,
```

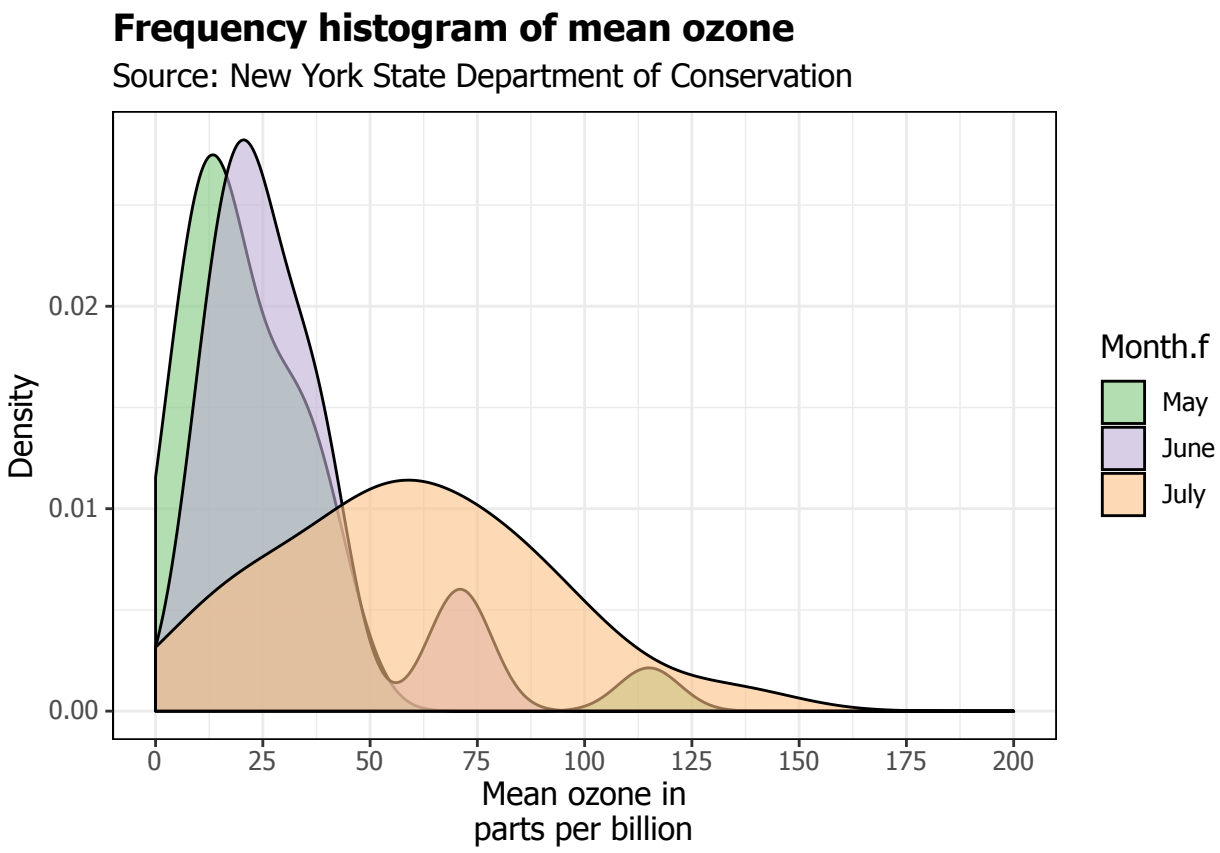
```
    position = "identity") +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density",
    breaks = seq(-0.03, 0.03, 0.03)) +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  facet_grid(. ~ Month.f) +
  coord_flip() +
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(size = 12, family = "Tahoma"))
p8
```



In order to plot the three months in the same plot, we add several things. Firstly, in the `ggplot` function, we add a `fill = Month.f` argument to `aes`. Secondly, in order to more clearly see the graph, we add the argument `position = "identity"` to the `geom_density` option. This controls the position of the curves respectively. Finally, you can customise the colours of the histograms by adding the `scale_fill_brewer` to the plot from the `RColorBrewer` package. This blog post describes the available packages.

```
p8 <- ggplot(airquality_trimmed, aes(x = Ozone, fill = Month.f)) +
  geom_density(position = "identity", alpha = 0.6) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
```

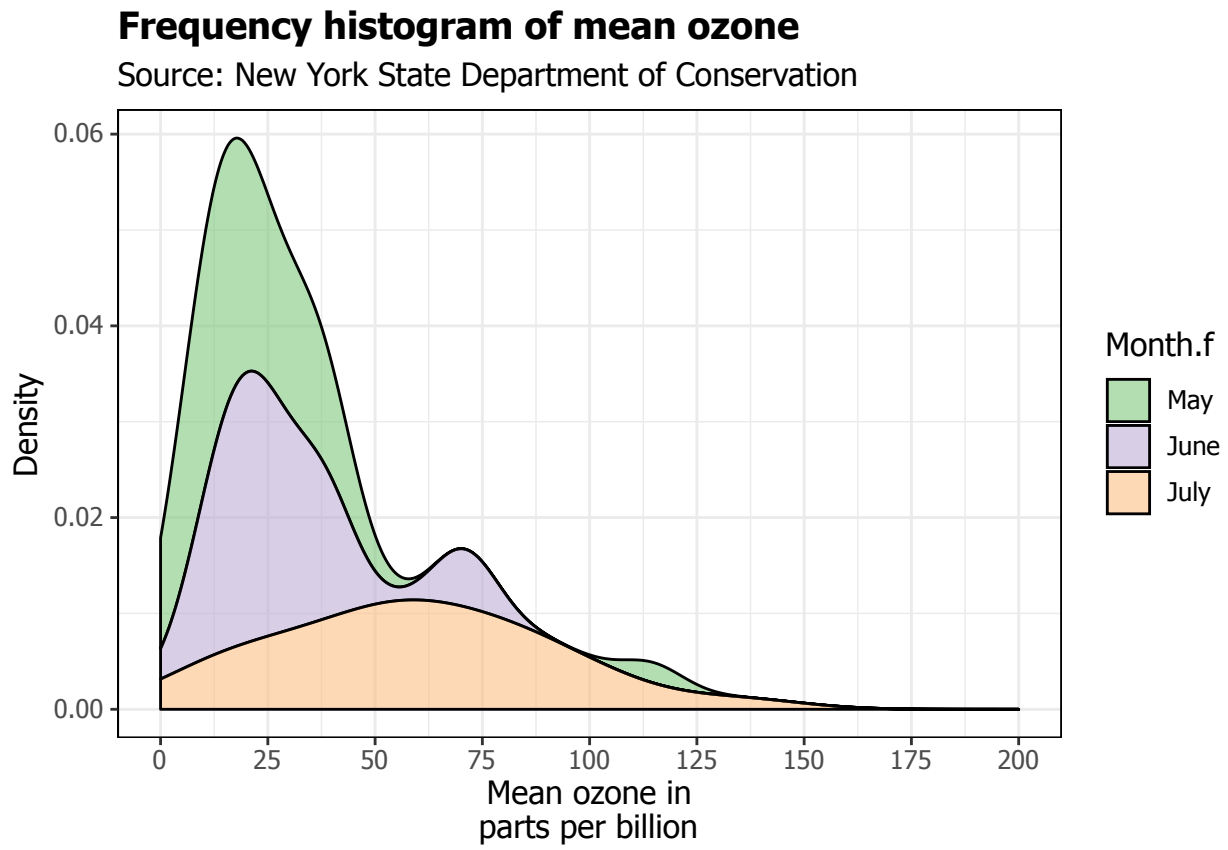
```
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  scale_fill_brewer(palette = "Accent") +
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(size = 12, family = "Tahoma"))
p8
```



These densities are a little hard to see. One way we can make it easier to see them is to stack the densities on top of each other. To do so, we swap `position = "stack"` for `position = "identity"` in `geom_density`.

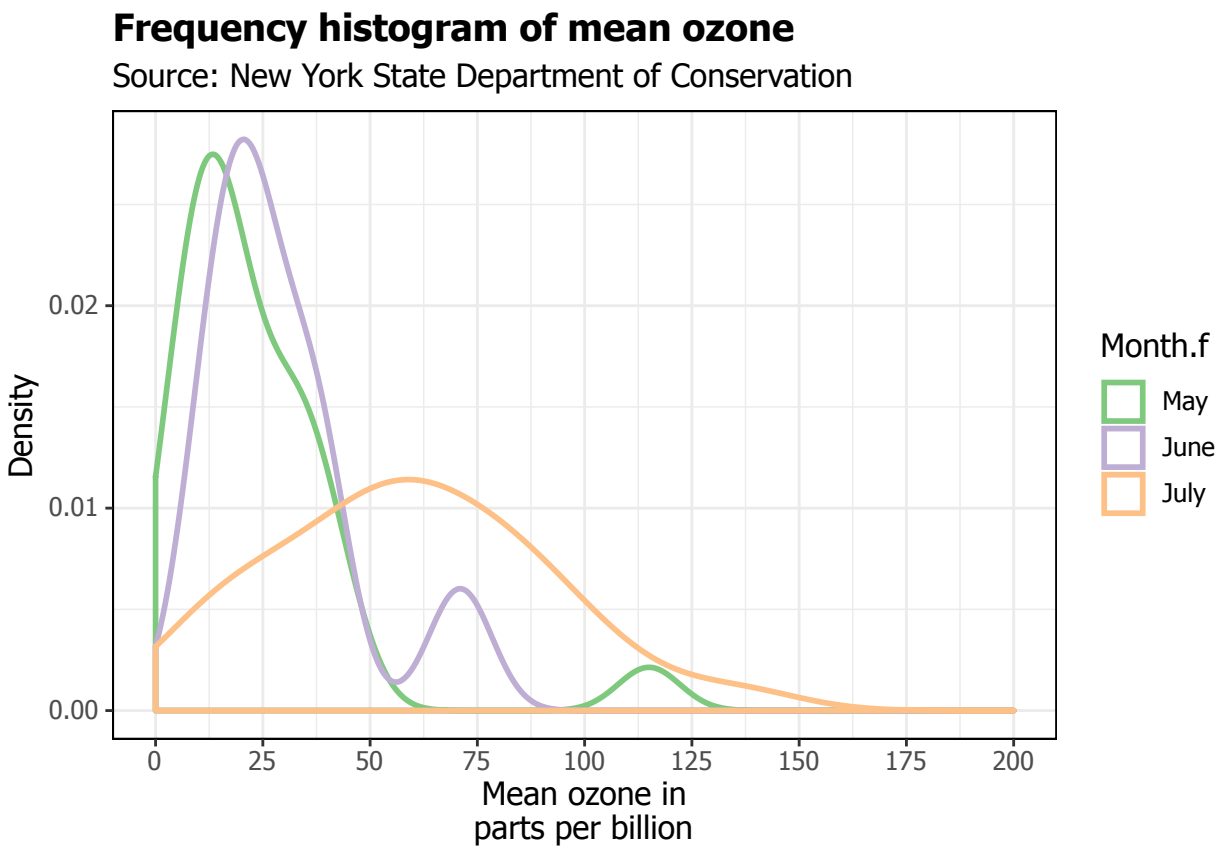
```
p8 <- ggplot(airquality_trimmed, aes(x = Ozone, fill = Month.f)) +
  geom_density(position = "stack", alpha = 0.6) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  scale_fill_brewer(palette = "Accent") +
  theme_bw() +
```

```
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(size = 12, family = "Tahoma"))
p8
```



Another way to make it a little easier to see the densities by dropping out the fill. To do this need a few changes. We need to swap the option `fill = Month.f` in `ggplot` for `colour = Month.f`. We add the `fill = NA` to `geom_density`, and we've also added `size = 1` to make it easier to see the lines. Finally, we change the `scale_fill_brewer()` option for `scale_colour_brewer()`.

```
p8 <- ggplot(airquality_trimmed, aes(x = Ozone, colour = Month.f)) +
  geom_density(position = "identity", fill = NA, size = 1) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  scale_colour_brewer(palette = "Accent") +
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(size = 12, family = "Tahoma"))
p8
```



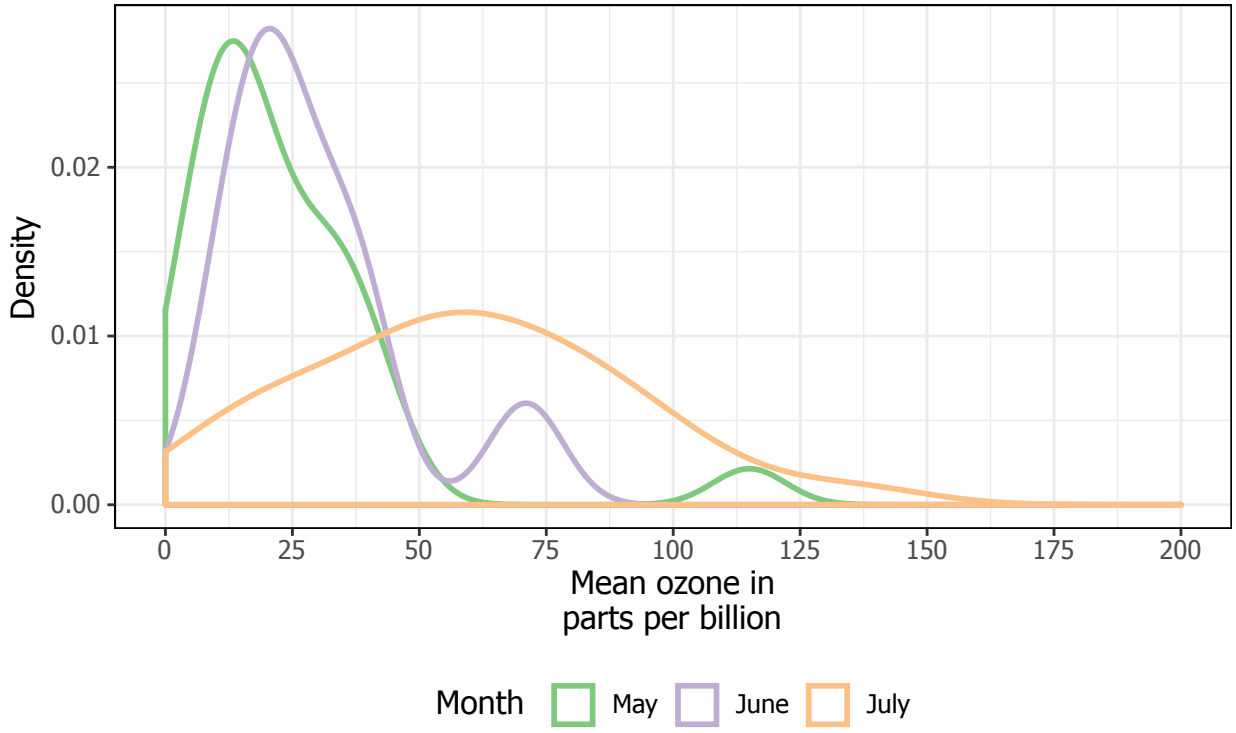
8.15. Formatting the legend

Finally, we can format the legend. Firstly, we can change the position by adding the `legend.position = "bottom"` argument to the `theme` option, which moves the legend under the plot. Secondly, we can fix the title by adding the `labs(fill = "Month ")` option to the plot.

```
p8 <- ggplot(airquality_trimmed, aes(x = Ozone, colour = Month.f)) +
  geom_density(position = "identity", fill = NA, size = 1) +
  scale_x_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 200, 25), limits = c(0, 200)) +
  scale_y_continuous(name = "Density") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation") +
  scale_colour_brewer(palette = "Accent") +
  labs(colour = "Month ") +
  theme_bw() +
  theme(legend.position = "bottom",
    panel.border = element_rect(colour = "black", fill = NA, size = .5),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(size = 12, family = "Tahoma"))
p8
```

Frequency histogram of mean ozone

Source: New York State Department of Conservation

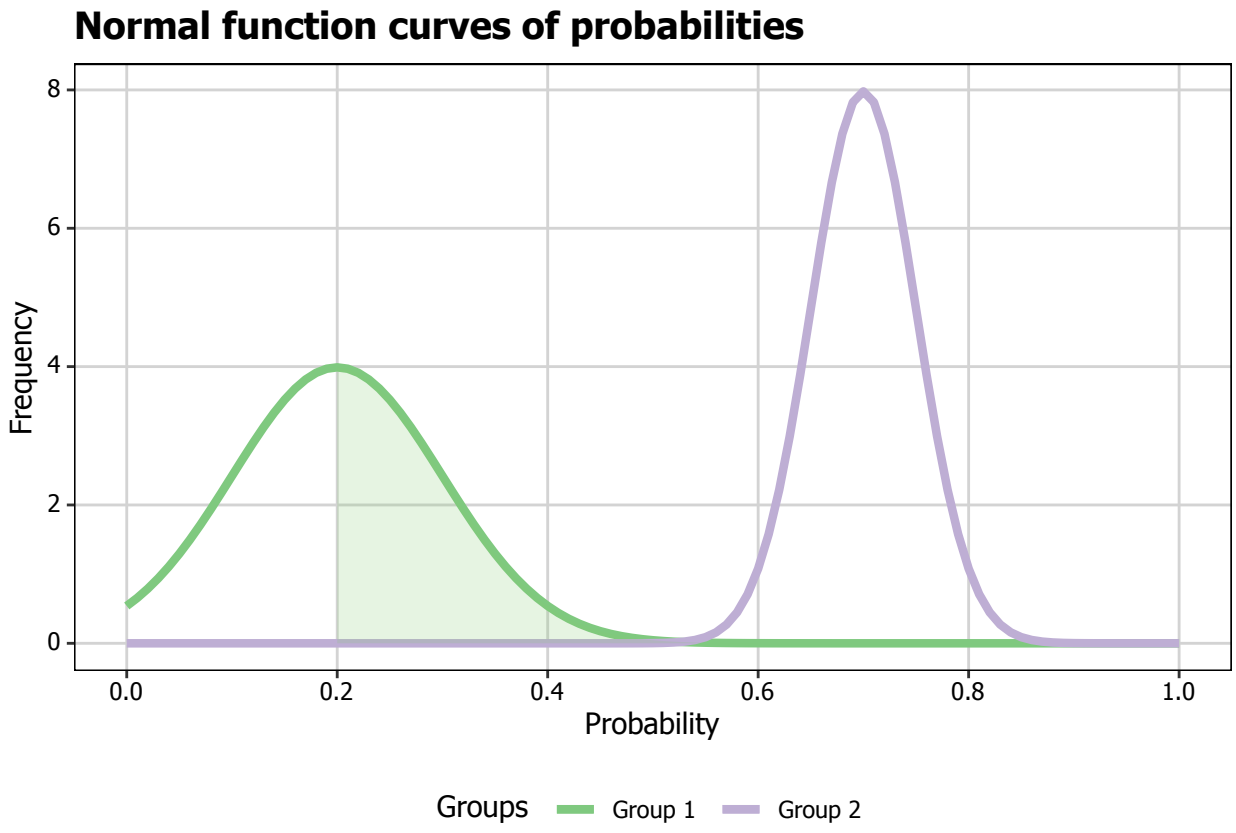


CHAPTER 9

Function plots

9.1. Introduction

In this chapter, we will work towards creating the function plot below. We will take you from a basic function plot and explain all the customisations we add to the code step-by-step.



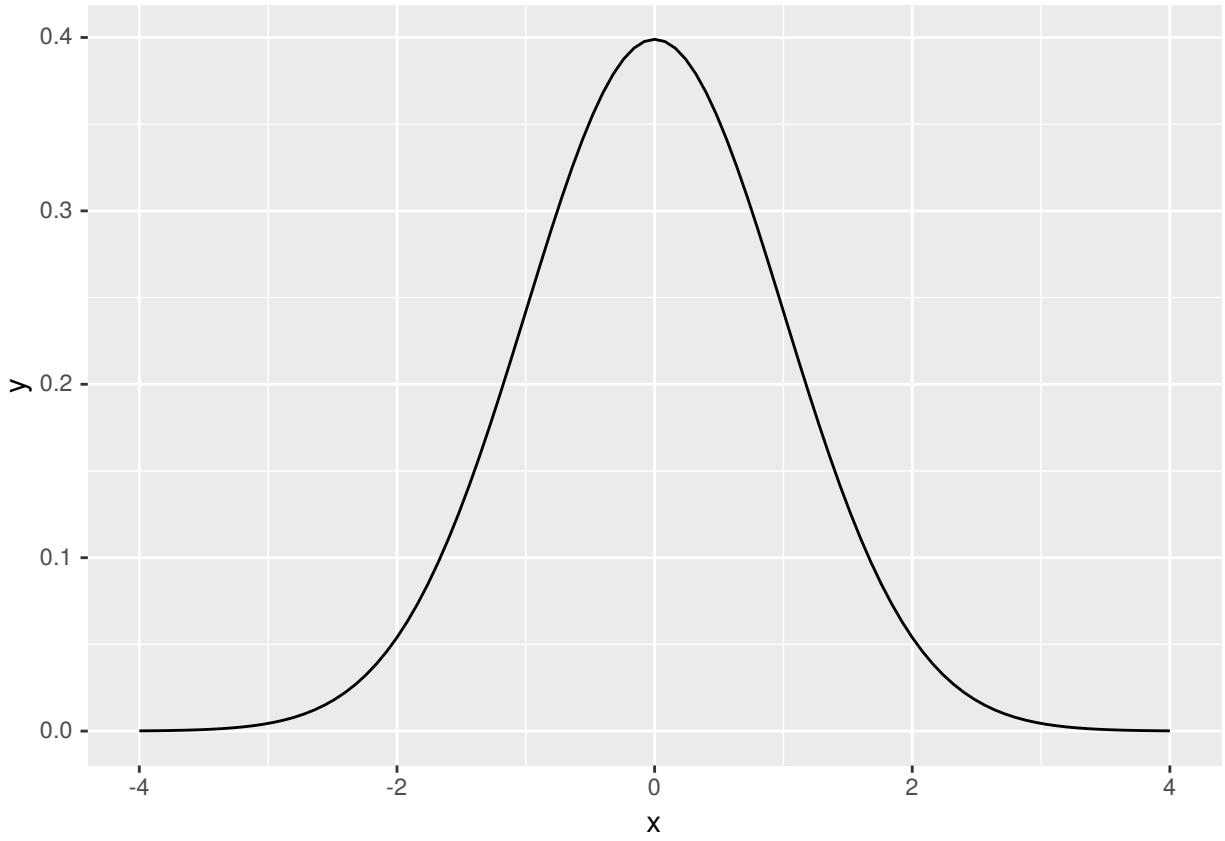
The first thing to do is load in the libraries, as below:

```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, grid)
```

9.2. Basic normal curve

In order to create a normal curve, we create a ggplot base layer that has an x-axis range from -4 to 4 (or whatever range you want!), and assign the x-value aesthetic to this range (`aes(x = x)`). We then add the `stat_function` option and add `dnorm` to the function argument to make it a normal curve.

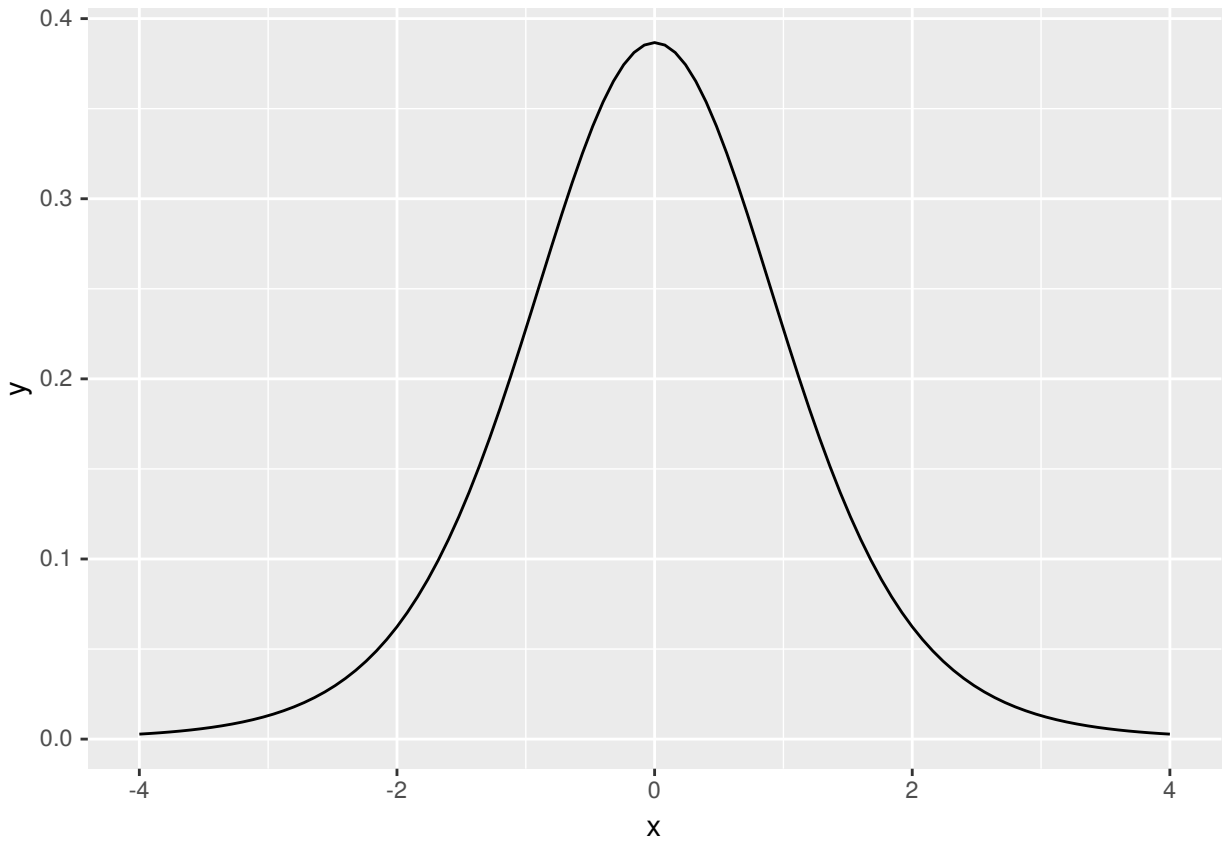
```
p9 <- ggplot(tibble(x = c(-4, 4)), aes(x = x)) +
  stat_function(fun = dnorm)
p9
```



9.3. Basic t-curve

`stat_function` can draw a range of continuous probability density functions, including t (`dt`), F (`df`) and Chi-square (`dchisq`) PDFs. Here we will plot a t-distribution. As the shape of the t-distribution changes depending on the sample size (indicated by the degrees of freedom, or df), we need to specify our df value as part of defining our curve.

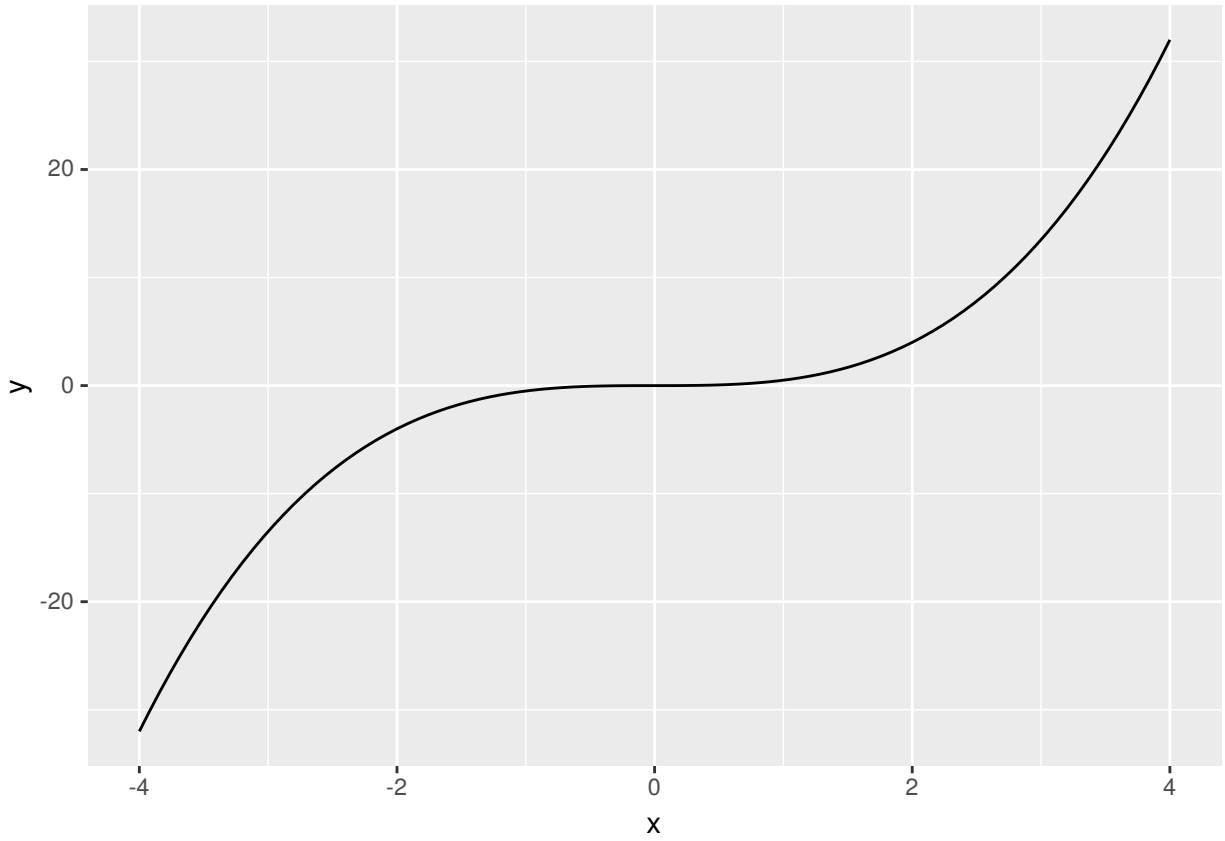
```
p9 <- ggplot(tibble(x = c(-4, 4)), aes(x = x)) +
  stat_function(fun = dt, args = list(df = 8))
p9
```



9.4. Plotting your own function

You can also draw your own function, as long as it takes the form of a formula that converts an x-value into a y-value. Here we have plotted a curve that returns y-values that are the cube of x times a half:

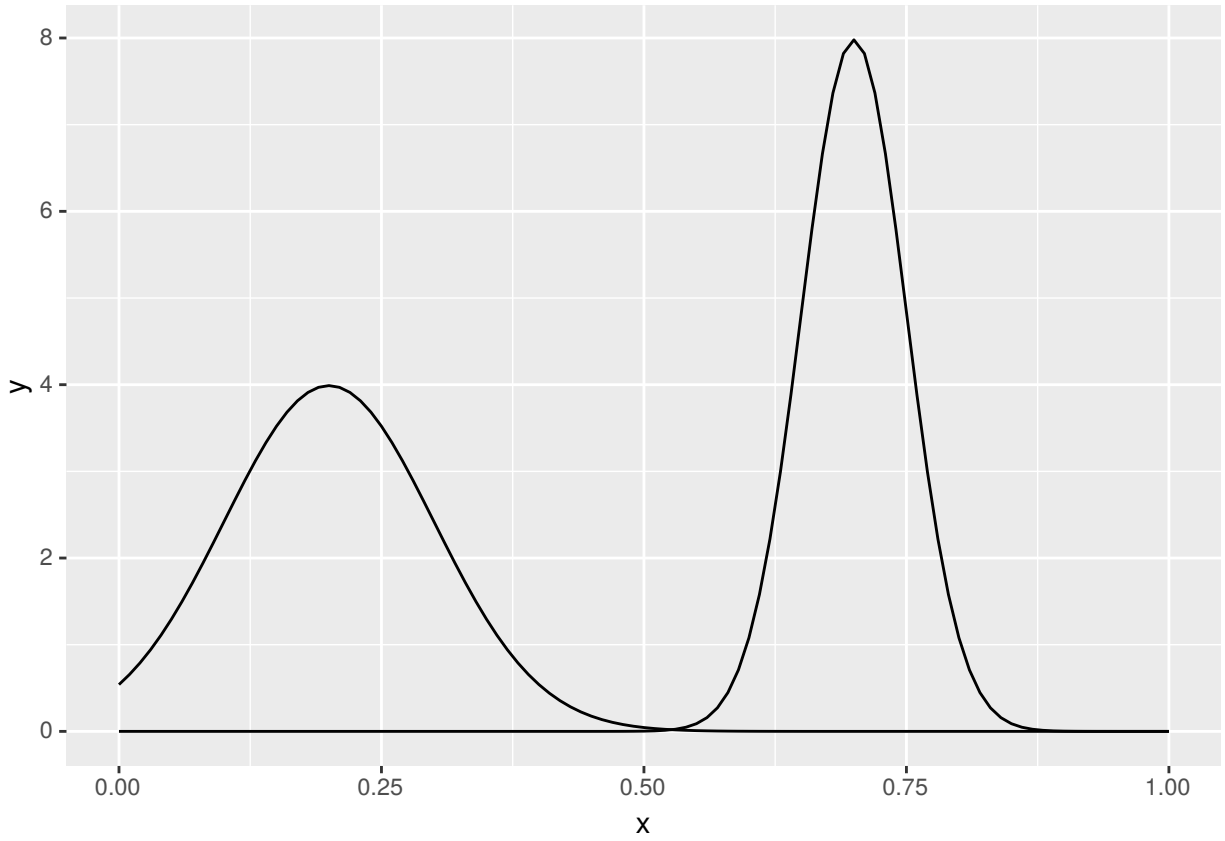
```
cubeFun <- function(x) {
    x^3 * 0.5
}

p9 <- ggplot(tibble(x = c(-4, 4)), aes(x = x)) +
  stat_function(fun = cubeFun)
p9
```



9.5. Plotting multiple functions on the same graph

You can plot multiple functions on the same graph by simply adding another `stat_function()` for each curve. Here we have plotted two normal curves on the same graph, one with a mean of 0.2 and a standard deviation of 0.1, and one with a mean of 0.7 and a standard deviation of 0.05. (Note that the `dnorm` function has a default mean of 0 and a default standard deviation of 1, which is why we didn't need to explicitly define them in the first normal curve we plotted above.) You can also see we've changed the range of the x-axis to between 0 and 1.

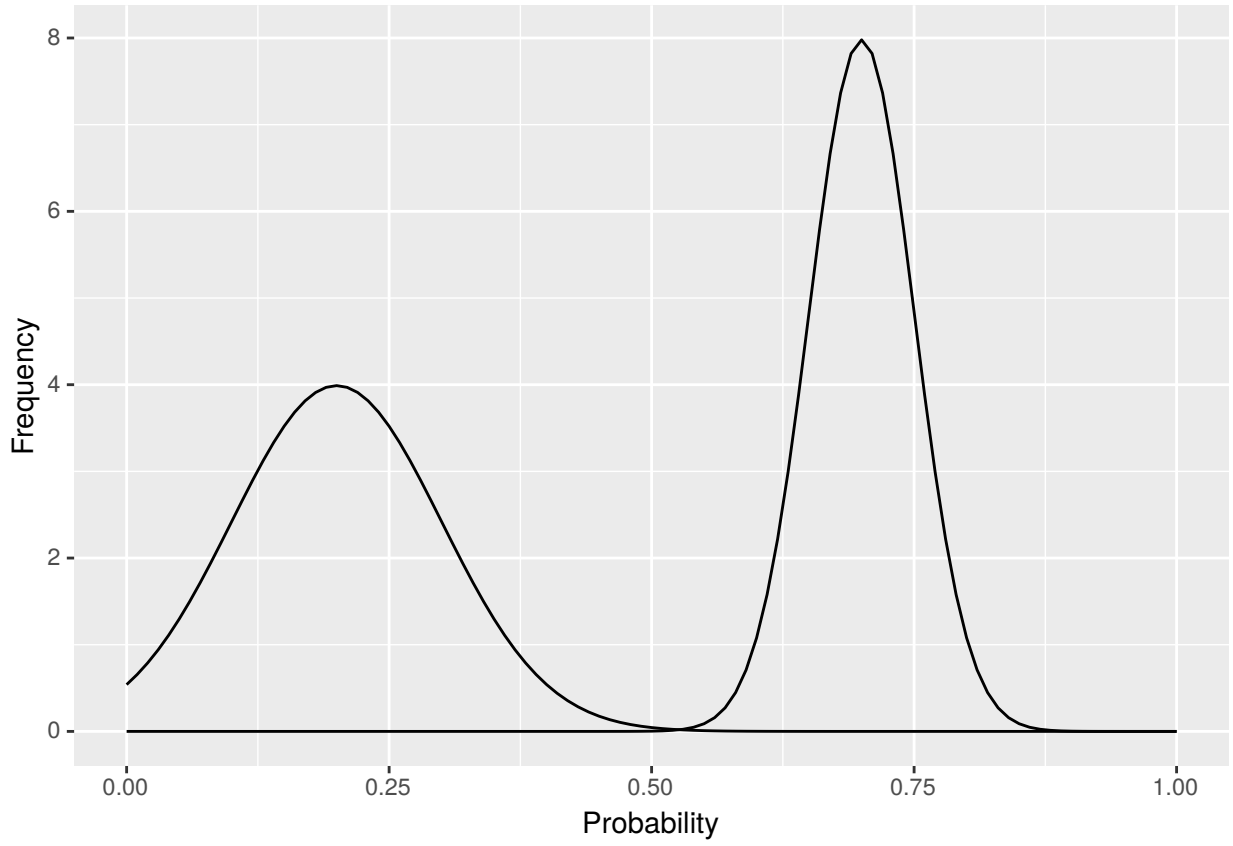
```
p9 = ggplot(tibble(x = c(0, 1)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(0.2, 0.1)) +
  stat_function(fun = dnorm, args = list(0.7, 0.05))
p9
```



9.6. Customising axis labels

Let's move forward with this two function graph, and start tweaking the appearance. 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.

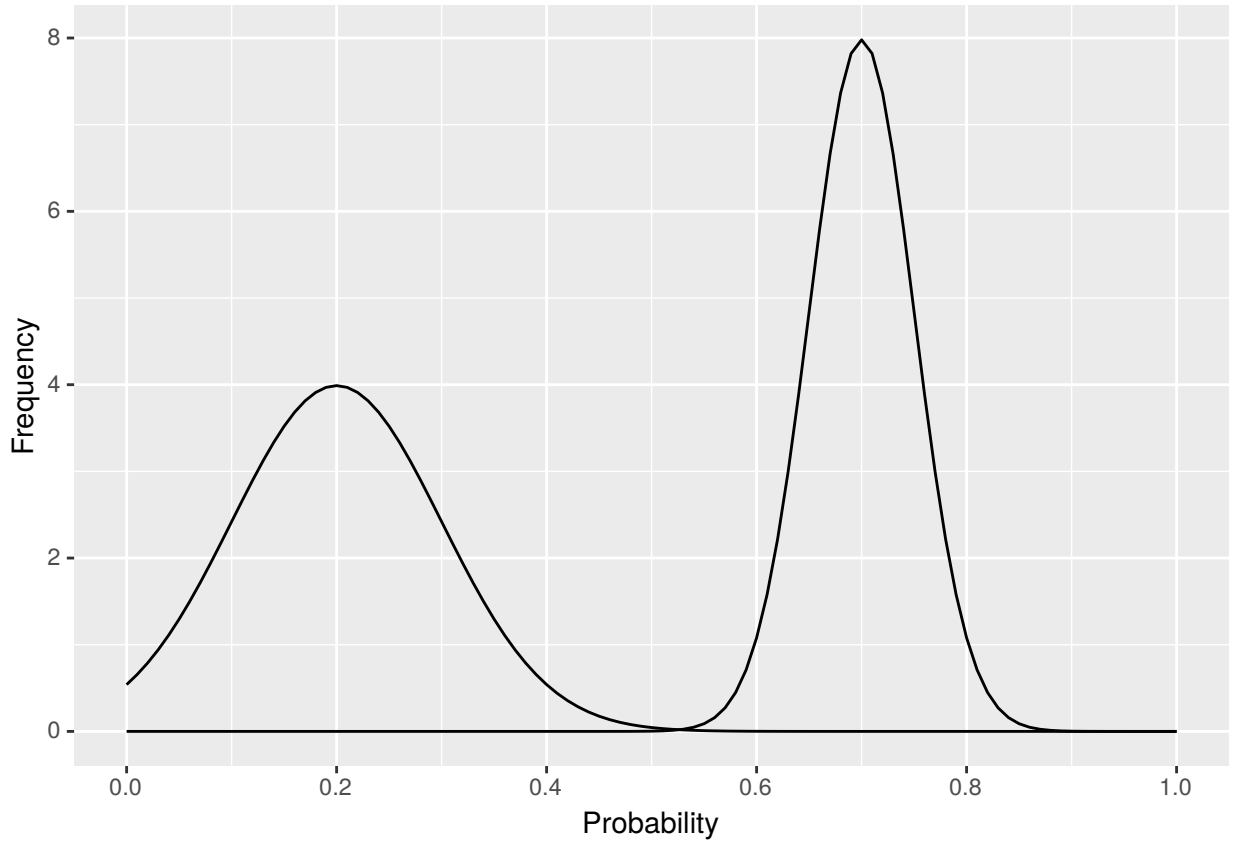
```
p9 <- p9 + scale_x_continuous(name = "Probability") +
  scale_y_continuous(name = "Frequency")
p9
```



9.7. Changing axis ticks

The next thing we will change is the axis ticks. Let's make the x-axis ticks appear at every 0.2 units rather than 0.25 using the `breaks = seq(0, 1, 0.2)` argument in `scale_x_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 x-axis begins and ends where we want by also adding the argument `limits = c(0, 1)` to `scale_x_continuous`.

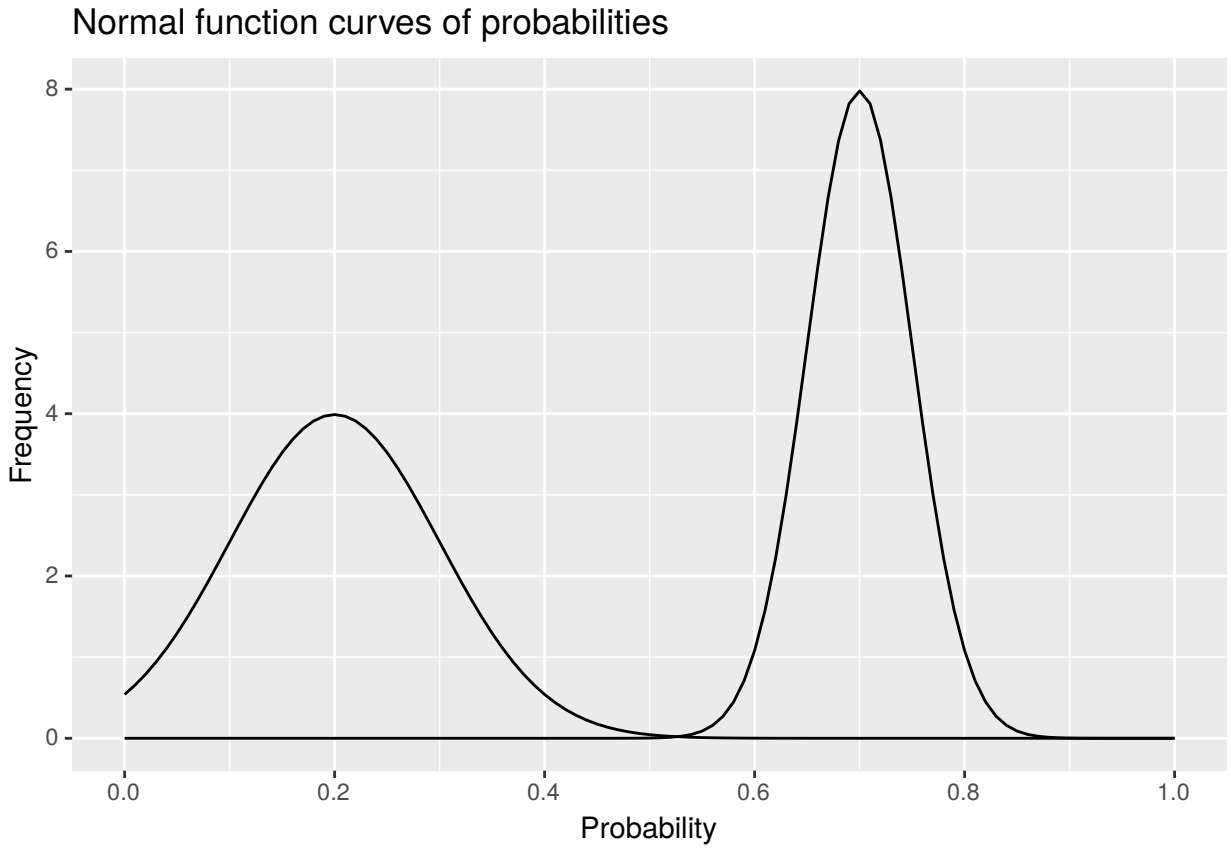
```
p9 <- p9 + scale_x_continuous(name = "Probability",
    breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
  scale_y_continuous(name = "Frequency")
p9
```



9.8. Adding a title

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

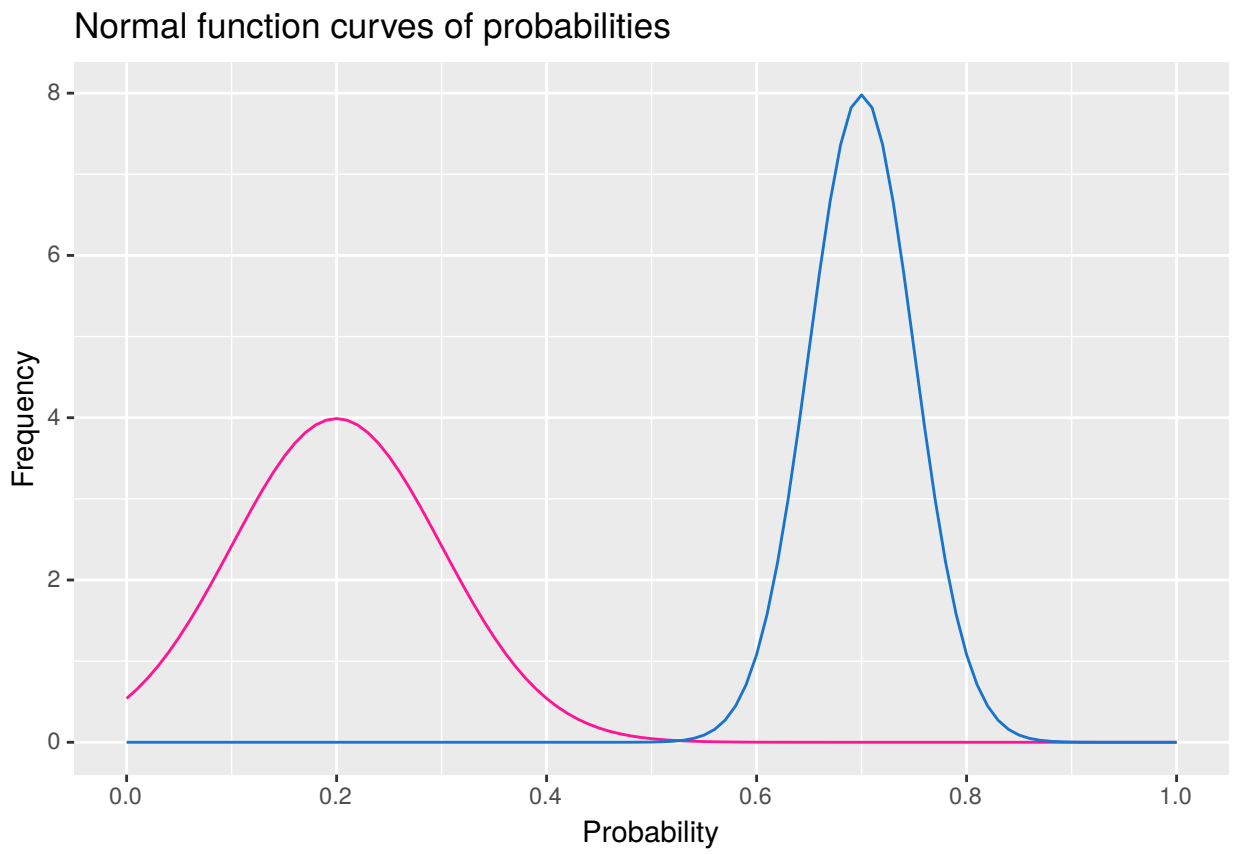
```
p9 <- p9 + labs(title = "Normal function curves of probabilities")
p9
```



9.9. Changing the colour of the curves

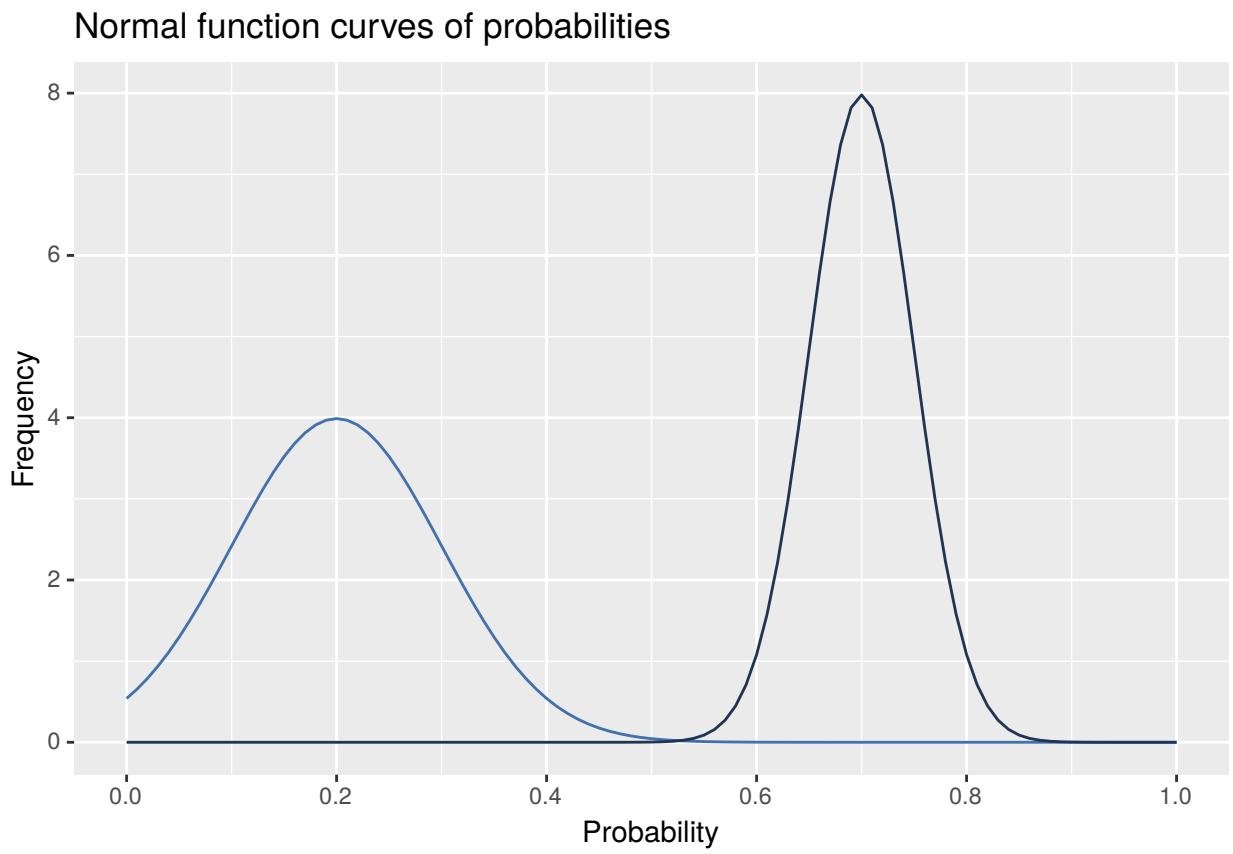
To change the line colours of the curves, we add a valid colour to the `colour` arguments in `stat_function`. A list of valid colours is here.

```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
    stat_function(fun = dnorm, args = list(0.2, 0.1),
      colour = "deeppink") +
    stat_function(fun = dnorm, args = list(0.7, 0.05),
      colour = "dodgerblue3") +
    scale_x_continuous(name = "Probability",
      breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
    scale_y_continuous(name = "Frequency") +
    labs(title = "Normal function curves of probabilities")
p9
```



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 two shades of blue for the lines using their HEX codes.

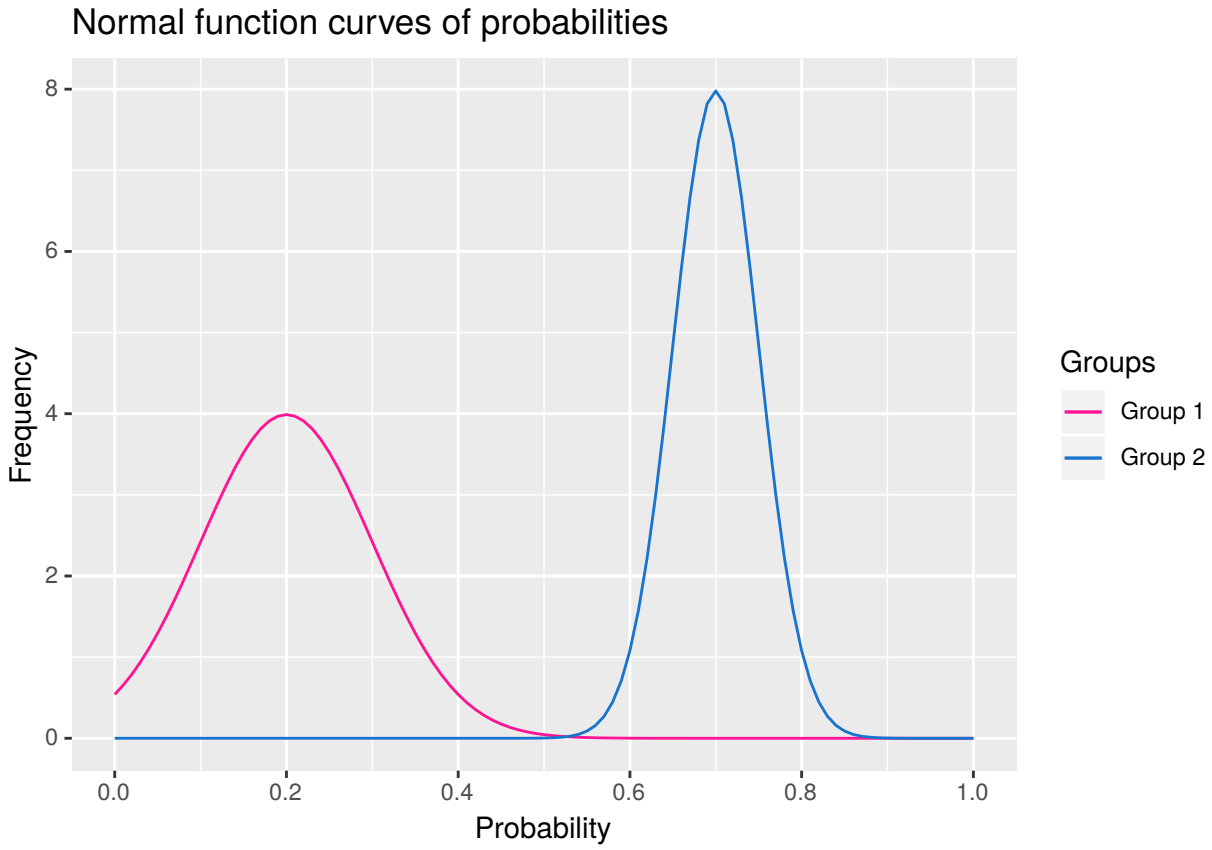
```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
    stat_function(fun = dnorm, args = list(0.2, 0.1),
      colour = "#4271AE") +
    stat_function(fun = dnorm, args = list(0.7, 0.05),
      colour = "#1F3552") +
    scale_x_continuous(name = "Probability",
      breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
    scale_y_continuous(name = "Frequency") +
    labs(title = "Normal function curves of probabilities")
p9
```



9.10. Adding a legend

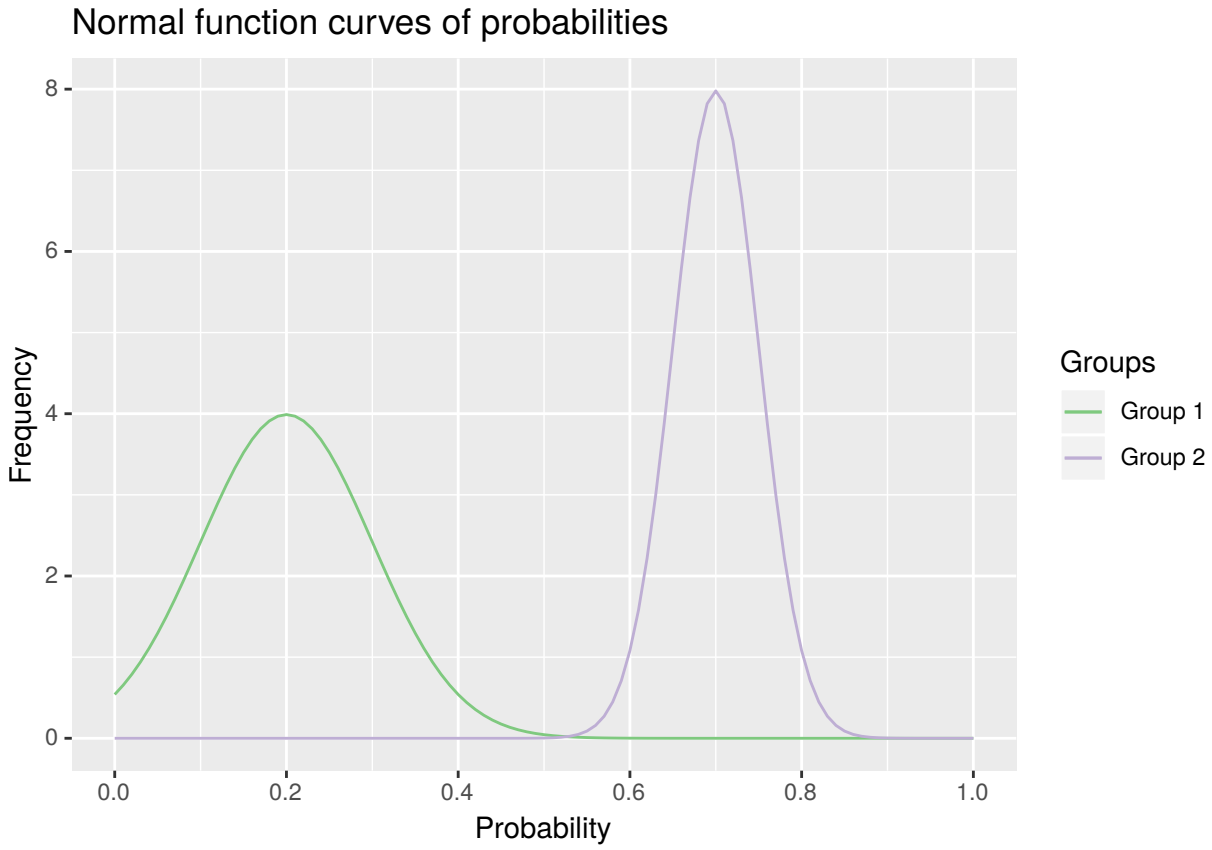
As we have added two separate commands to plot the two function curves, ggplot does not automatically recognise that it needs to create a legend. We can make a legend by swapping out the `colour` argument in each of the `stat_function` commands for `aes(colour = )`, and assigning it the name of the group. We also need to add the `scale_colour_manual` command to make the legend appear, and also assign colours and a title.

```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
    stat_function(fun = dnorm, args = list(0.2, 0.1),
      aes(colour = "Group 1 ")) +
    stat_function(fun = dnorm, args = list(0.7, 0.05),
      aes(colour = "Group 2 ")) +
    scale_x_continuous(name = "Probability",
      breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
    scale_y_continuous(name = "Frequency") +
    labs(title = "Normal function curves of probabilities") +
    scale_colour_manual("Groups ", values = c("deeppink", "dodgerblue3"))
p9
```



If you want to use one of the automatic brewer palettes, you can swap `scale_colour_manual` for `scale_colour_brewer`, and call your favouite brewer colour scheme. You can see all of the brewer palettes using `display.brewer.all(5)` As this command doesn't allow you to assign a title to the legend, you can assign a title using `labs(colour = "Groups ")`.

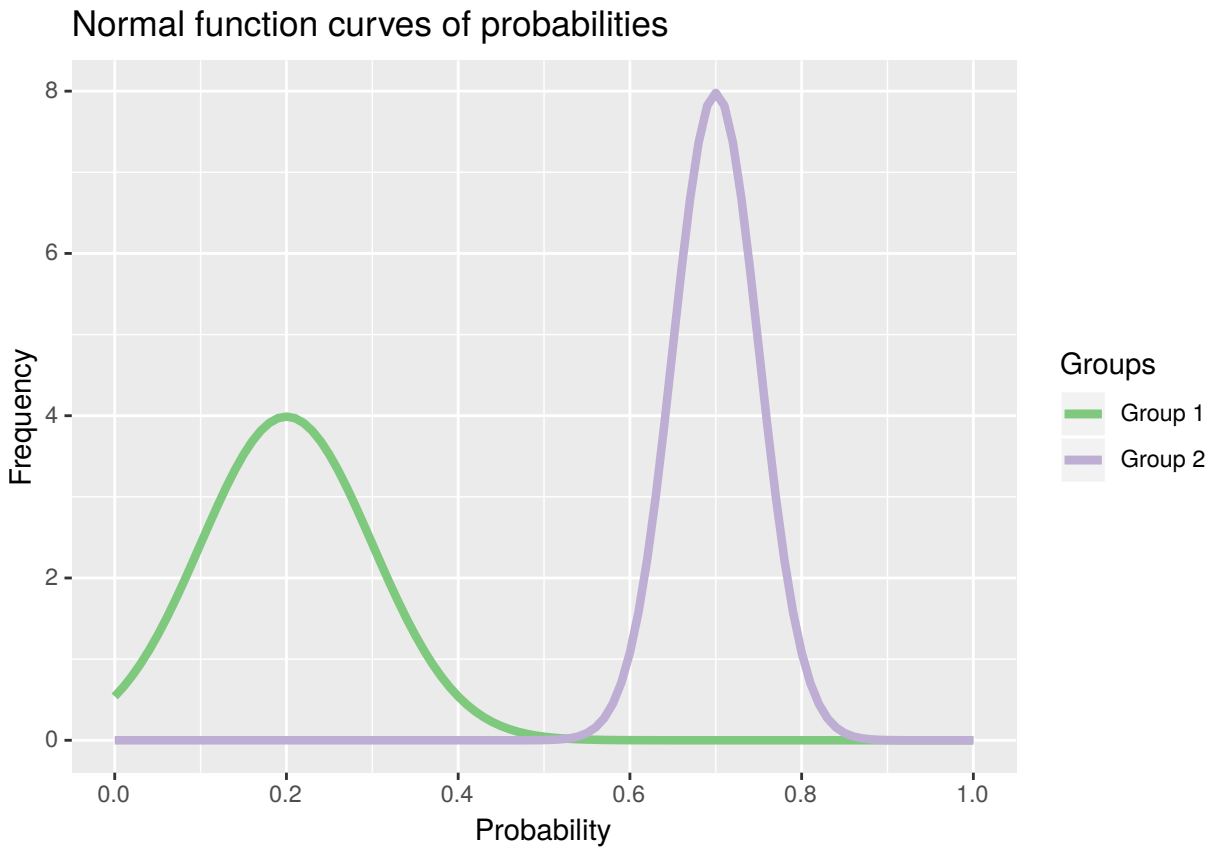
```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
    stat_function(fun = dnorm, args = list(0.2, 0.1),
      aes(colour = "Group 1 ")) +
    stat_function(fun = dnorm, args = list(0.7, 0.05),
      aes(colour = "Group 2 ")) +
    scale_x_continuous(name = "Probability",
      breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
    scale_y_continuous(name = "Frequency") +
    labs(title = "Normal function curves of probabilities") +
    labs(colour = "Groups ") +
    scale_colour_brewer(palette = "Accent")
p9
```



9.11. Changing the size of the lines

As you can see, the lines are a little difficult to see. You can make them thicker (or thinner) using the argument `size` argument within `stat_function`. Here we have changed the thickness of each line to size 2.

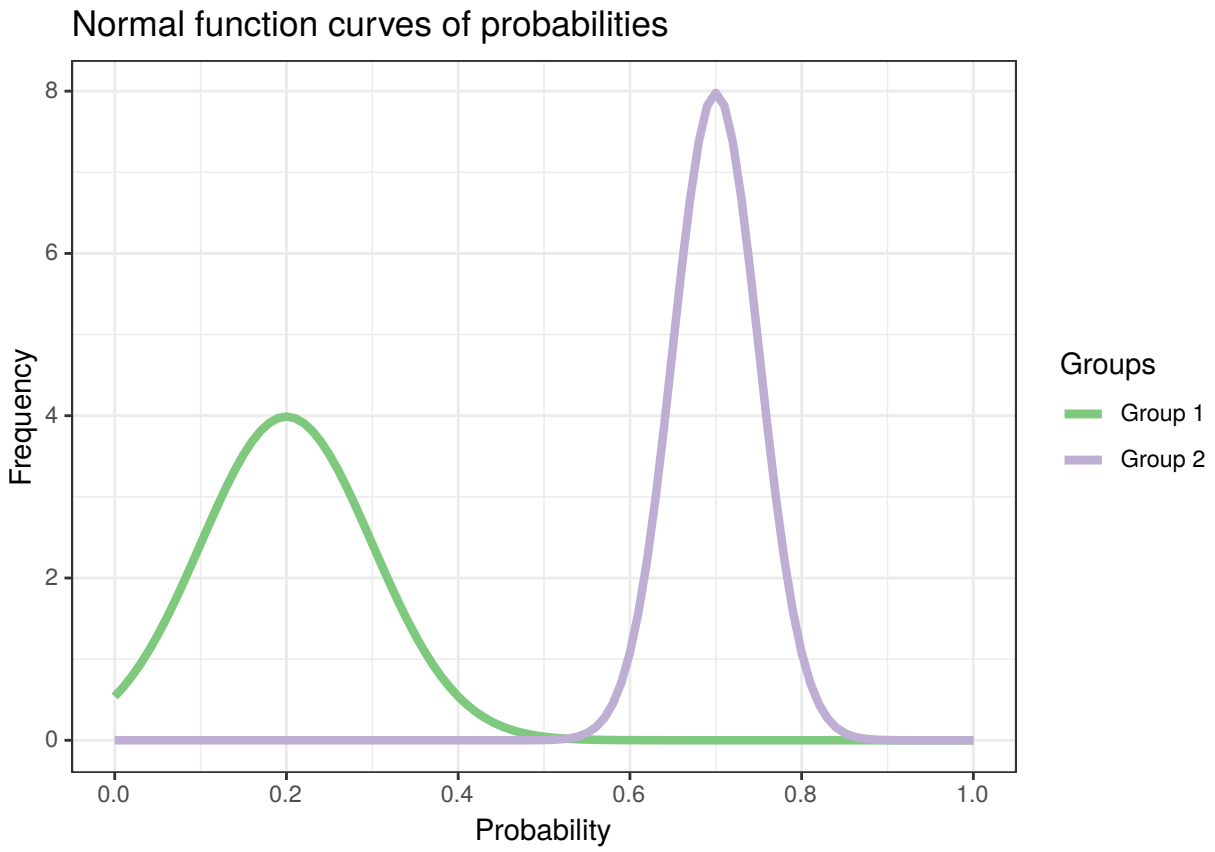
```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
    stat_function(fun = dnorm, args = list(0.2, 0.1),
      aes(colour = "Group 1 "), size = 1.5) +
    stat_function(fun = dnorm, args = list(0.7, 0.05),
      aes(colour = "Group 2 "), size = 1.5) +
    scale_x_continuous(name = "Probability",
      breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
    scale_y_continuous(name = "Frequency") +
    labs(title = "Normal function curves of probabilities") +
    labs(colour = "Groups ") +
    scale_colour_brewer(palette = "Accent")
p9
```



9.12. 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()` after `ggplot()`. As you can see, we can further tweak the graph using the `theme` option, which we've used so far to change the legend.

```
p9 <- p9 + theme_bw()
p9
```



9.13. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

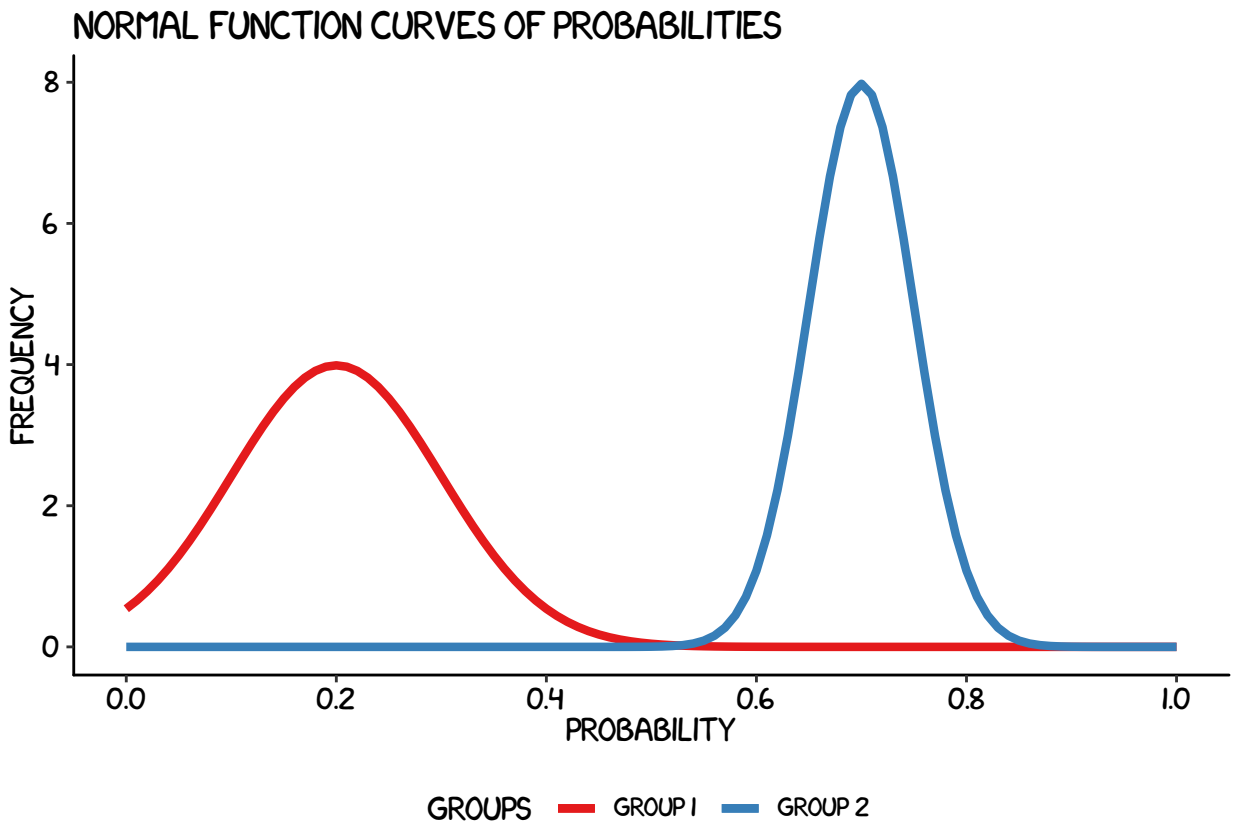
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

9.14. 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.

```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(0.2, 0.1),
    aes(colour = "Group 1 "), size = 1.5) +
  stat_function(fun = dnorm, args = list(0.7, 0.05),
    aes(colour = "Group 2 "), size = 1.5) +
  scale_x_continuous(name = "Probability",
    breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
  scale_y_continuous(name = "Frequency") +
  labs(title = "Normal function curves of probabilities") +
  scale_colour_brewer(palette = "Set1") +
  labs(colour = "Groups ") +
  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.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p9
```

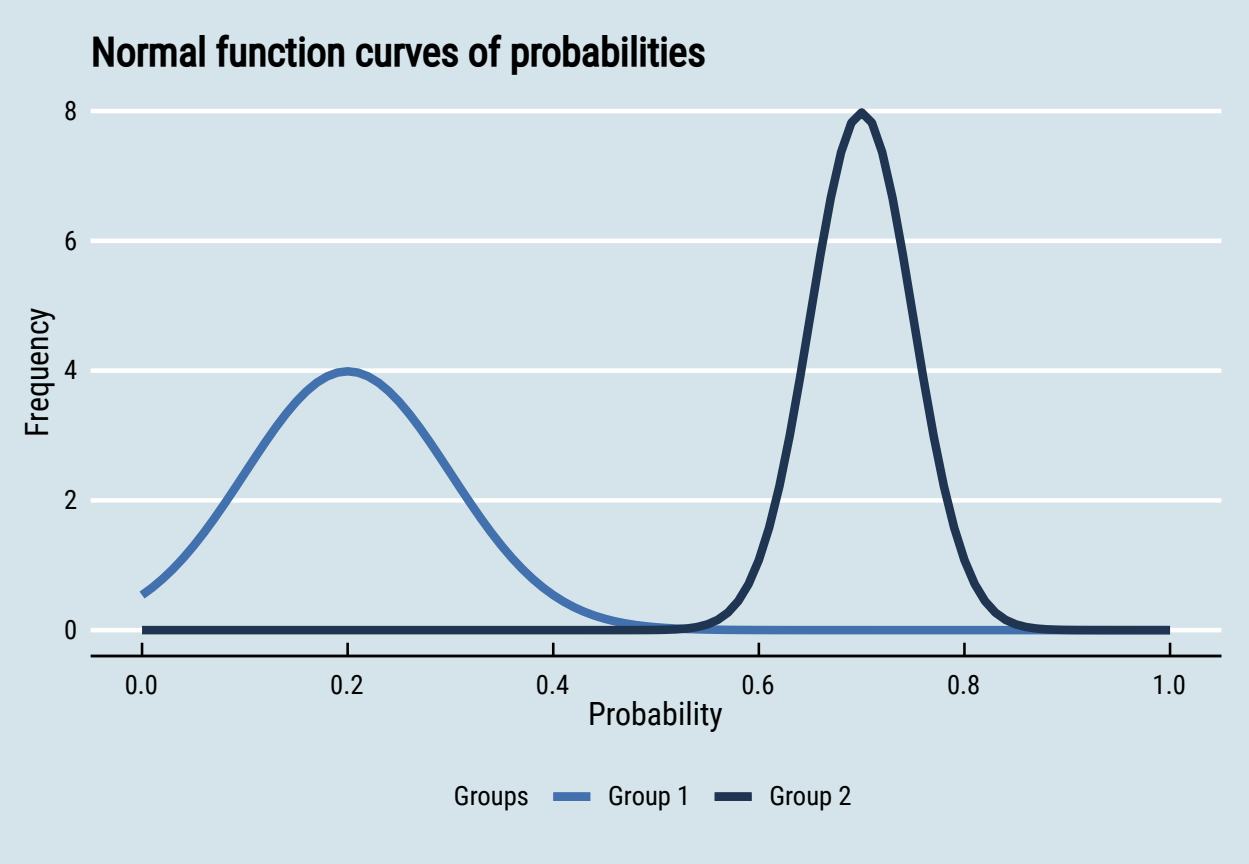


9.15. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(0.2, 0.1),
    aes(colour = "Group 1 "), size = 1.5) +
  stat_function(fun = dnorm, args = list(0.7, 0.05),
    aes(colour = "Group 2 "), size = 1.5) +
  scale_x_continuous(name = "Probability",
    breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
  scale_y_continuous(name = "Frequency") +
  labs(title = "Normal function curves of probabilities") +
  scale_colour_manual("Groups ", values = c("#4271AE", "#1F3552")) +
  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 = "Roboto Condensed"),
    plot.title = element_text(family = "Roboto Condensed"))
p9
```

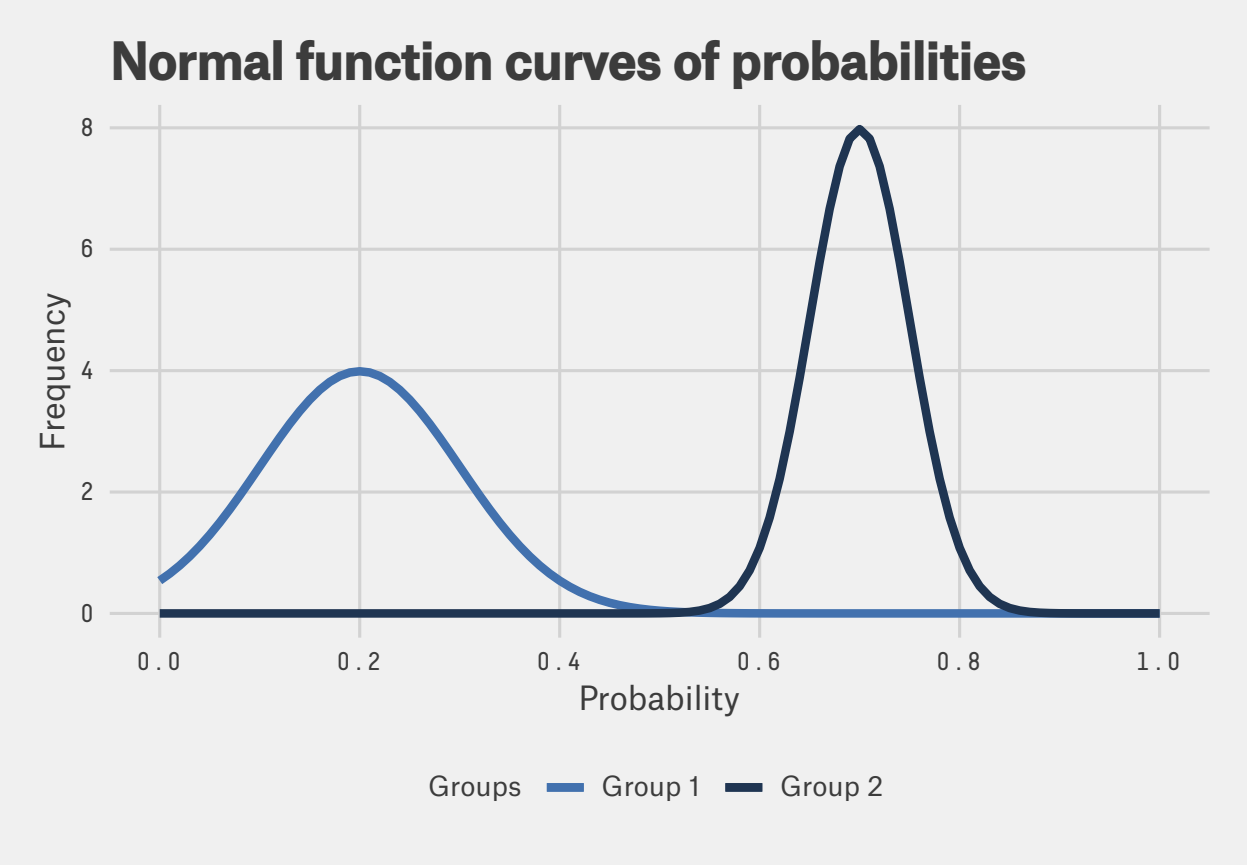


9.16. Using 'Five Thirty Eight' theme

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'.

```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(0.2, 0.1),
    aes(colour = "Group 1 "), size = 1.5) +
  stat_function(fun = dnorm, args = list(0.7, 0.05),
    aes(colour = "Group 2 "), size = 1.5) +
  scale_x_continuous(name = "Probability",
    breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
  scale_y_continuous(name = "Frequency") +
  labs(title = "Normal function curves of probabilities") +
```

```
  scale_colour_manual("Groups ", values = c("#4271AE", "#1F3552")) +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 10),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 10),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    text = element_text(family = "Decima Mono Pro"))
p9
```

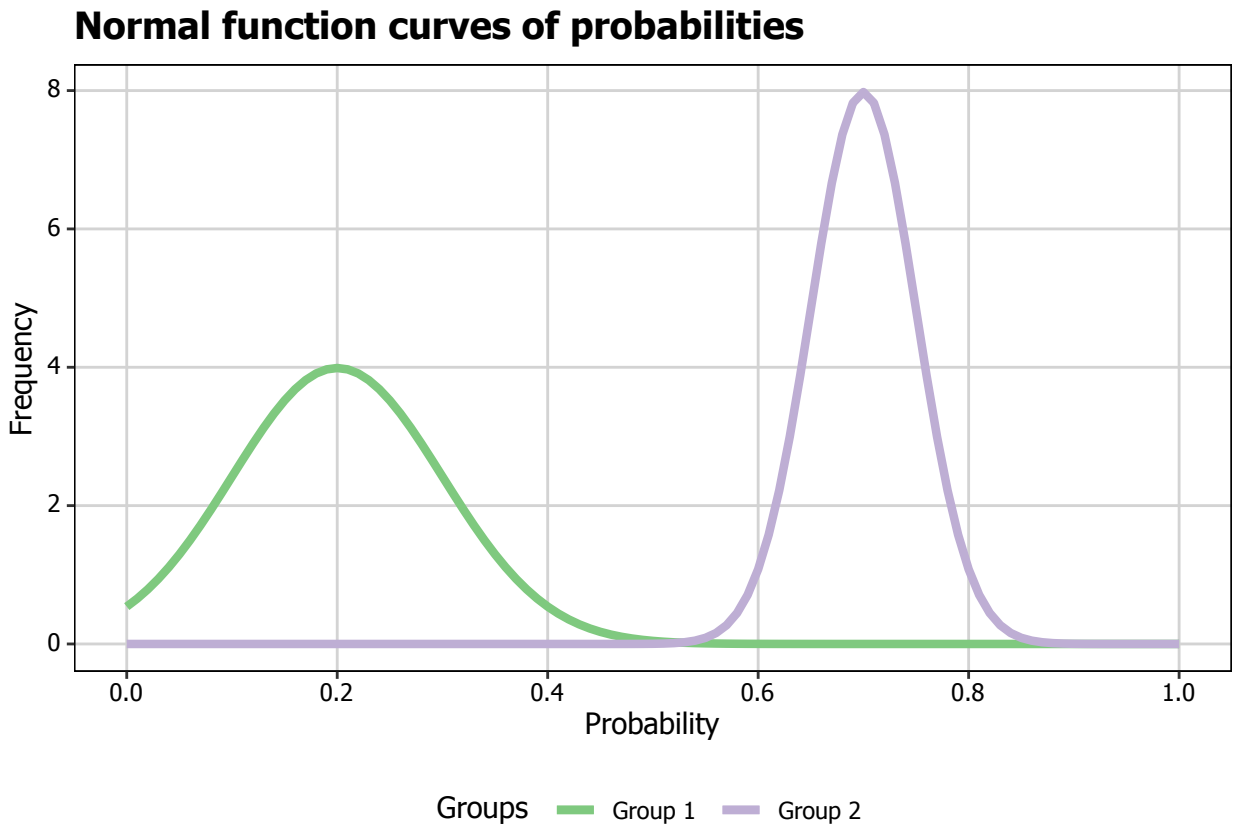


9.17. Creating your own theme

As before, you can modify your plots a lot as `ggplot2` allows many customisations. Here is a custom plot where we have modified the axes, background and font.

```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(0.2, 0.1),
    aes(colour = "Group 1 "), size = 1.5) +
  stat_function(fun = dnorm, args = list(0.7, 0.05),
```

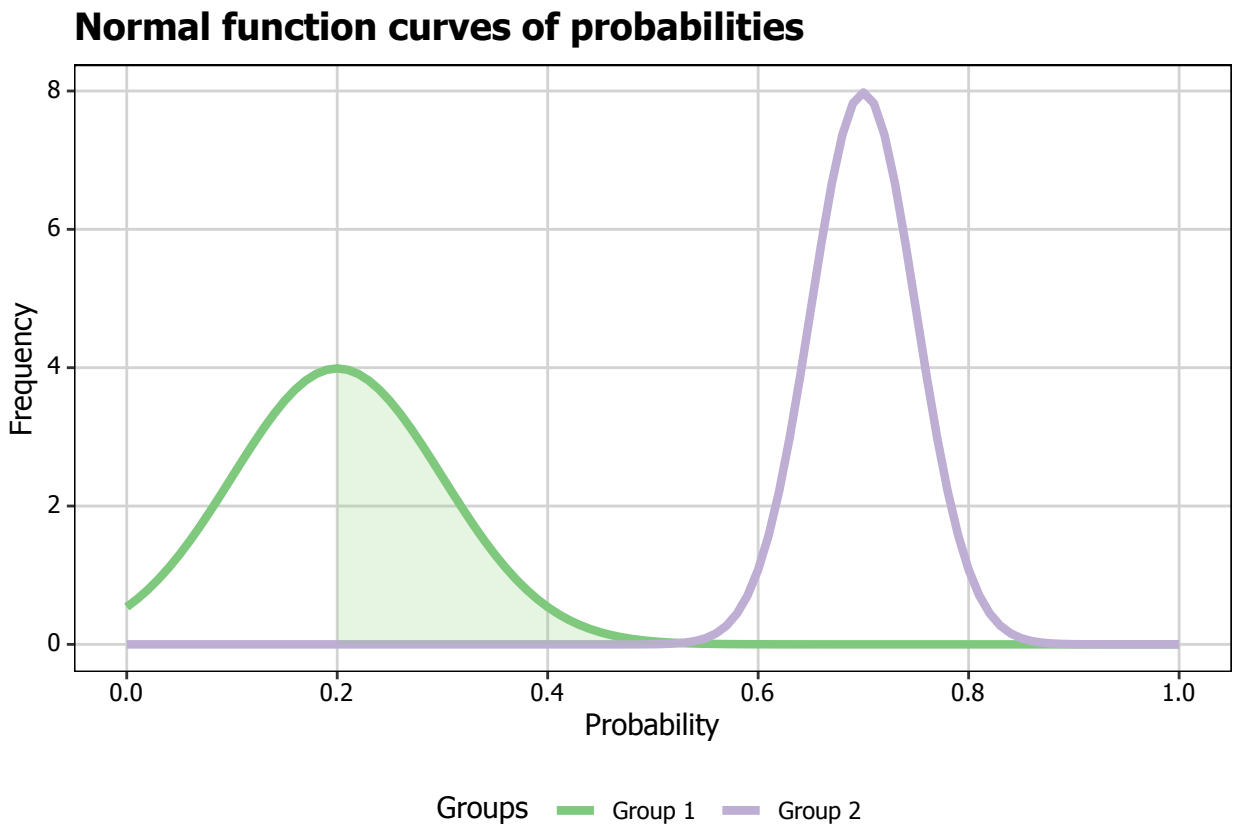
```
    aes(colour = "Group 2 "), size = 1.5) +
  scale_x_continuous(name = "Probability",
    breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
  scale_y_continuous(name = "Frequency") +
  labs(title = "Normal function curves of probabilities") +
  labs(colour = "Groups ") +
  scale_colour_brewer(palette = "Accent") +
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p9
```



9.18. Adding areas under the curve

If we want to shade an area under the curve, we can do so by creating a function that generates a range of normal values with a given mean and standard deviation, and then only retains those values that lie within the desired range (by assigning NAs to everything outside of the range). In this case, we have created a shaded area under the group 1 curve which covers between the mean and 4 standard deviations above the mean (as given by `0.2 + 4 * 0.1`). We then add another `stat_function` command to the graph which plots the area specified by this function, indicates it should be an `area` plot, and makes it semi-transparent using the `alpha` argument.

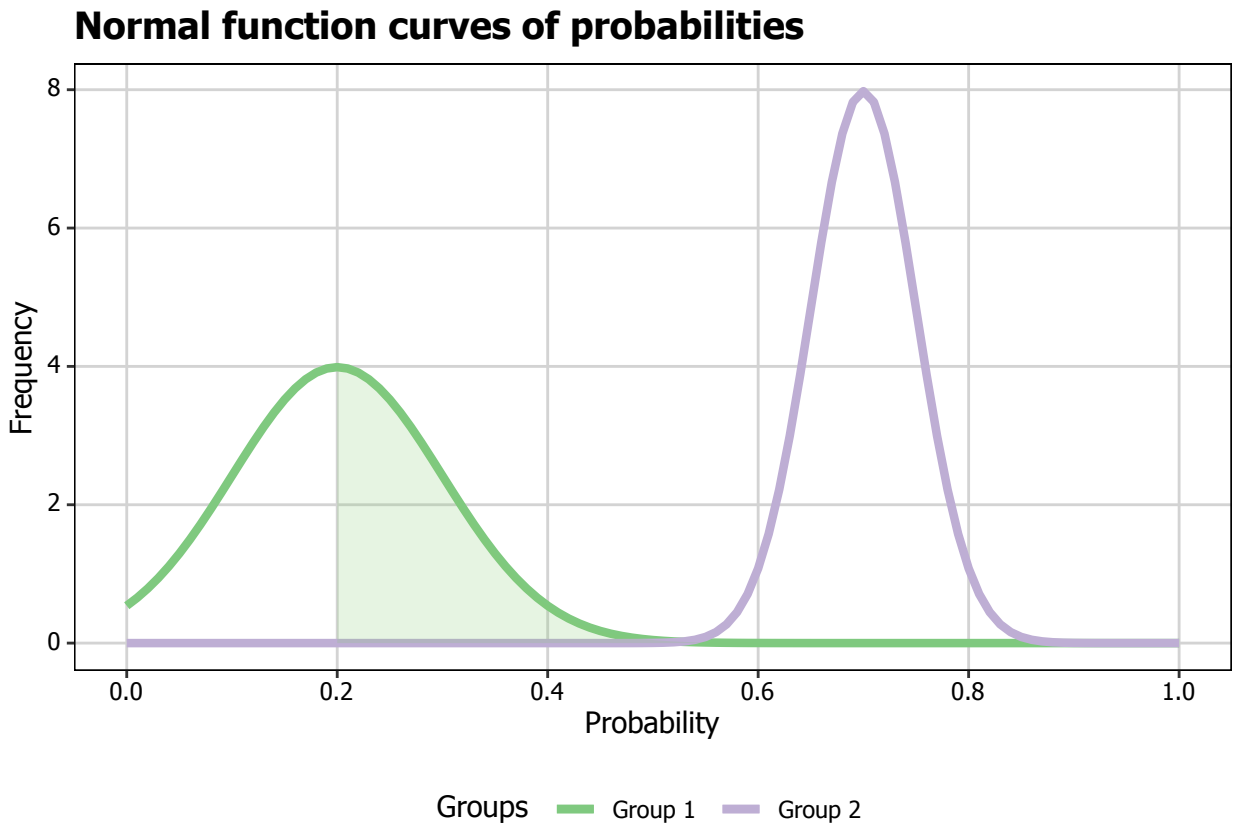
```
funcShaded <- function(x) {
    y <- dnorm(x, mean = 0.2, sd = 0.1)
    y[x < 0.2 | x > (0.2 + 4 * 0.1)] <- NA
    return(y)
}

p9 <- p9 + stat_function(fun = funcShaded, geom = "area", fill = "#84CA72", alpha = 0.2)
p9
```



9.19. Formatting the legend

Finally, we can format the legend by changing the position. We simply add the `legend.position = "bottom"` argument to the `theme` option, which moves the legend under the plot.

```
p9 <- ggplot(tibble(x = c(0, 1)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(0.2, 0.1),
    aes(colour = "Group 1 "), size = 1.5) +
  stat_function(fun = dnorm, args = list(0.7, 0.05),
    aes(colour = "Group 2 "), size = 1.5) +
  stat_function(fun = funcShaded, geom = "area", fill = "#84CA72", alpha = 0.2) +
  scale_x_continuous(name = "Probability",
    breaks = seq(0, 1, 0.2), limits = c(0, 1)) +
  scale_y_continuous(name = "Frequency") +
  labs(title = "Normal function curves of probabilities") +
  labs(colour = "Groups ") +
  scale_colour_brewer(palette = "Accent") +
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p9
```

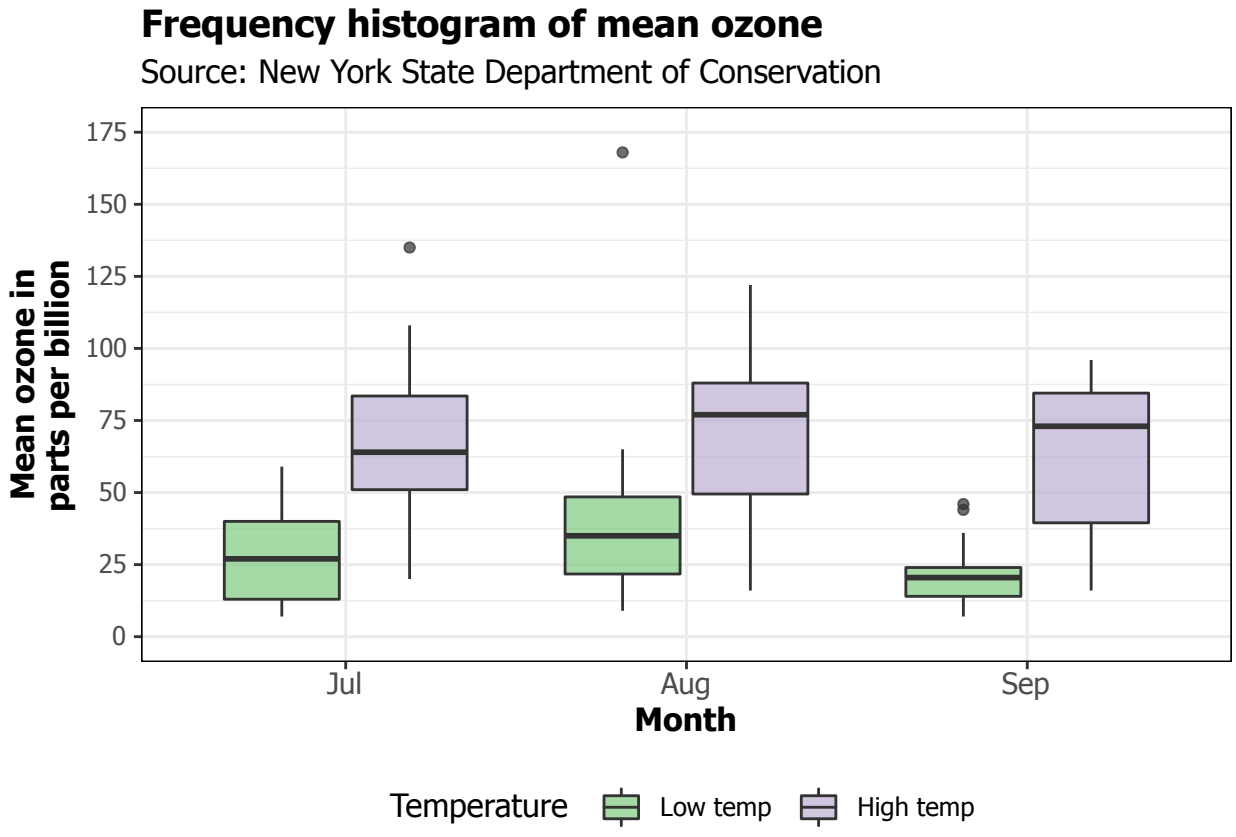

Normal function curves of probabilities
8
6
4
2
0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Probability
Groups
Group 1
Group 2

CHAPTER 10

Boxplots

10.1. Introduction

In this chapter, we will work towards creating the boxplot below. We will take you from a basic boxplot and explain all the customisations we add to the code step-by-step.



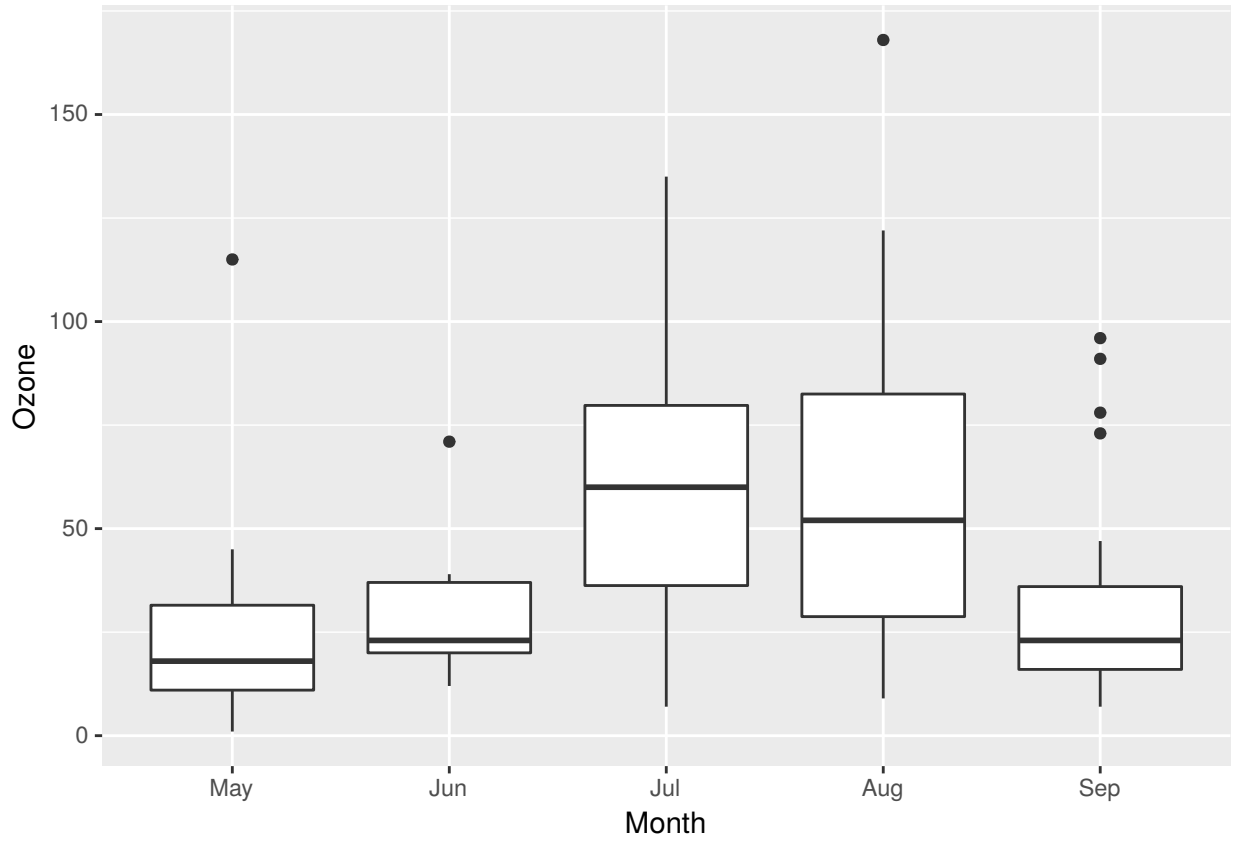
The first thing to do is load in the data and the libraries, as below. We'll convert `Month` into a labelled factor in order to use it as our grouping variable.

```
if (!require("pacman")) install.packages("pacman")
p_load(datasets, ggplot2, ggthemes, grid, dplyr, RColorBrewer)

data(airquality)
airquality <- airquality %>%
  mutate(Month = factor(Month,
                        labels = c("May", "Jun", "Jul", "Aug", "Sep")))
```

10.2. Basic boxplot

In order to initialise a plot we tell ggplot that `airquality` is our data, and specify that our x-axis plots the `Month` variable and our y-axis plots the `Ozone` variable. We then instruct ggplot to render this as a boxplot by adding the `geom_boxplot()` option.

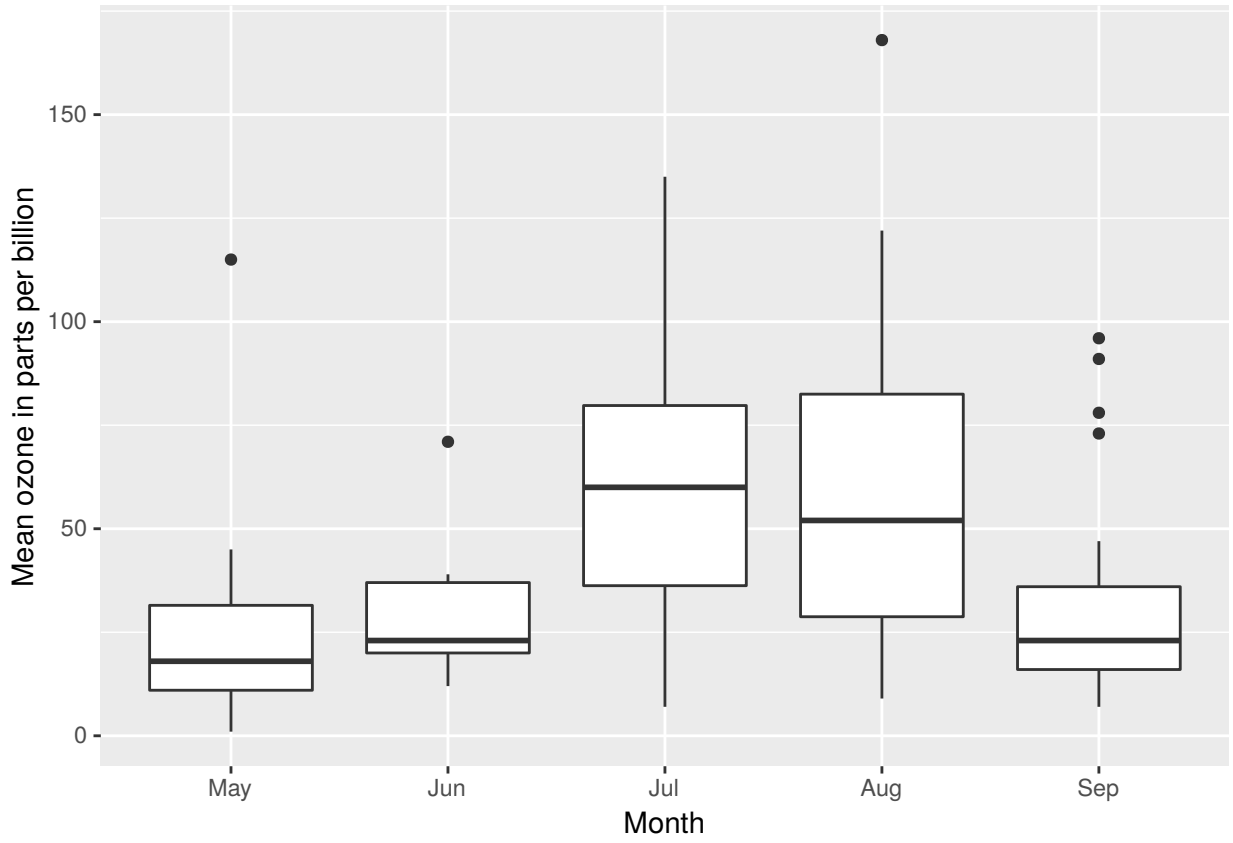
```
p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot()
p10
```



10.3. Customising axis labels

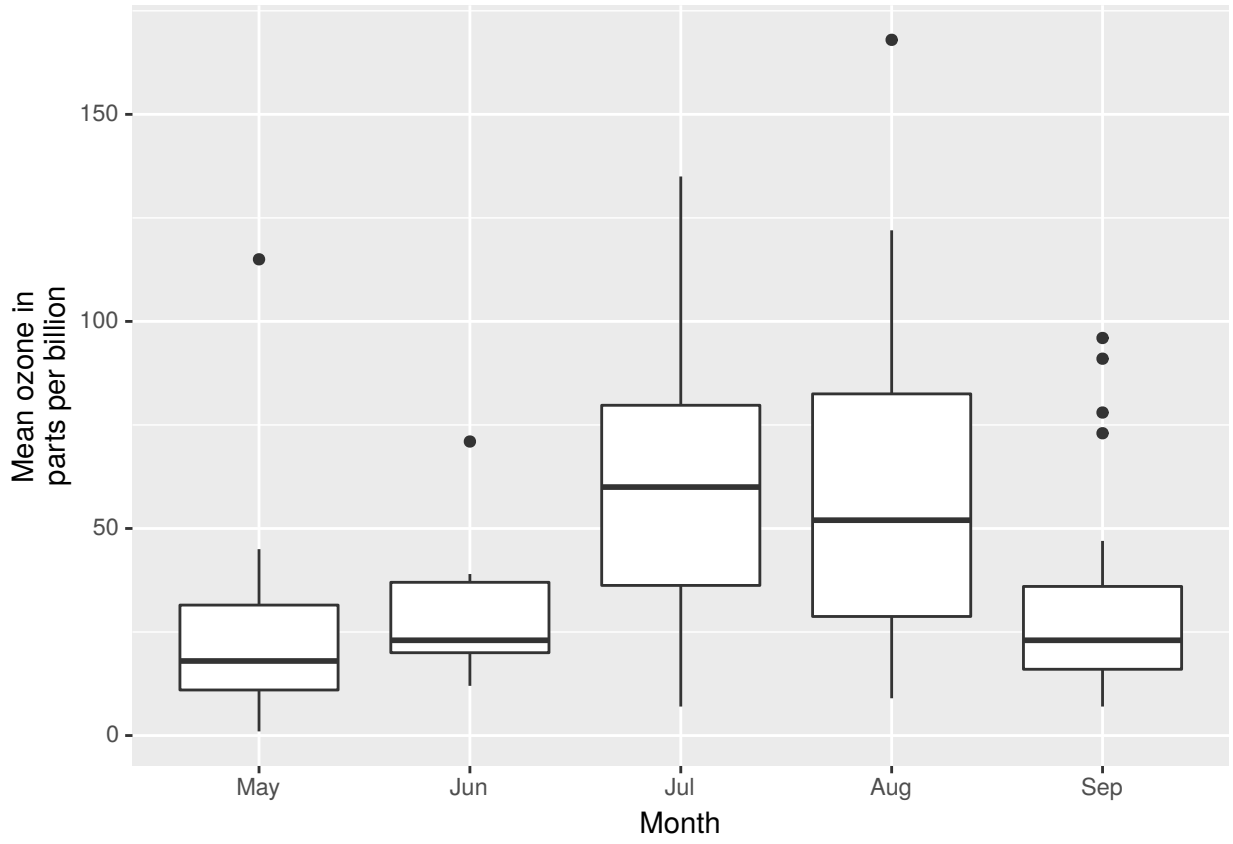
In order to change the axis labels, we have a couple of options. In this case, we have used the `scale_x_discrete` and `scale_y_continuous` options, as these have further customisation options for the axes we will use below. In each, we add the desired name to the `name` argument as a string.

```
p10 <- p10 + scale_x_discrete(name = "Month") +
  scale_y_continuous(name = "Mean ozone in parts per billion")
p10
```



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.

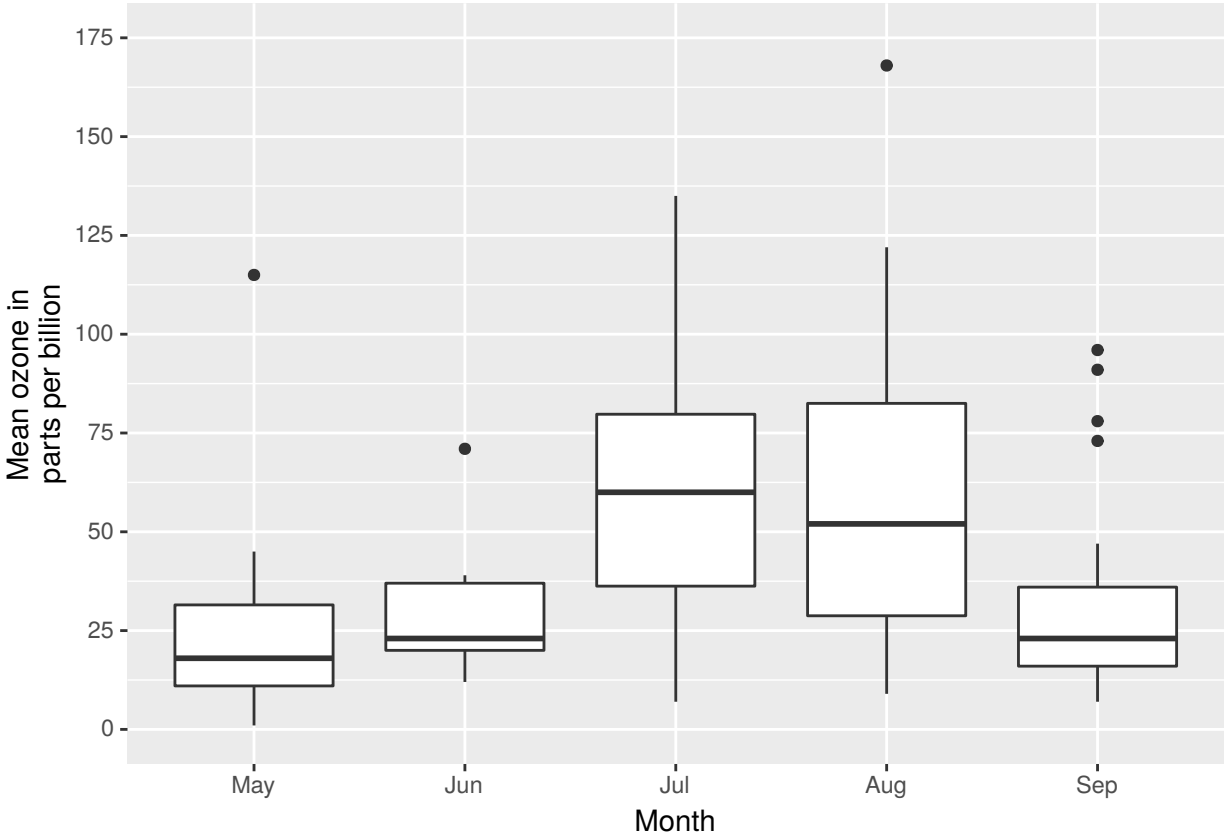
```
p10 <- p10 + scale_y_continuous(name = "Mean ozone in\nparts per billion")
p10
```



10.4. 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, 175, 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, 175)` to `scale_y_continuous`.

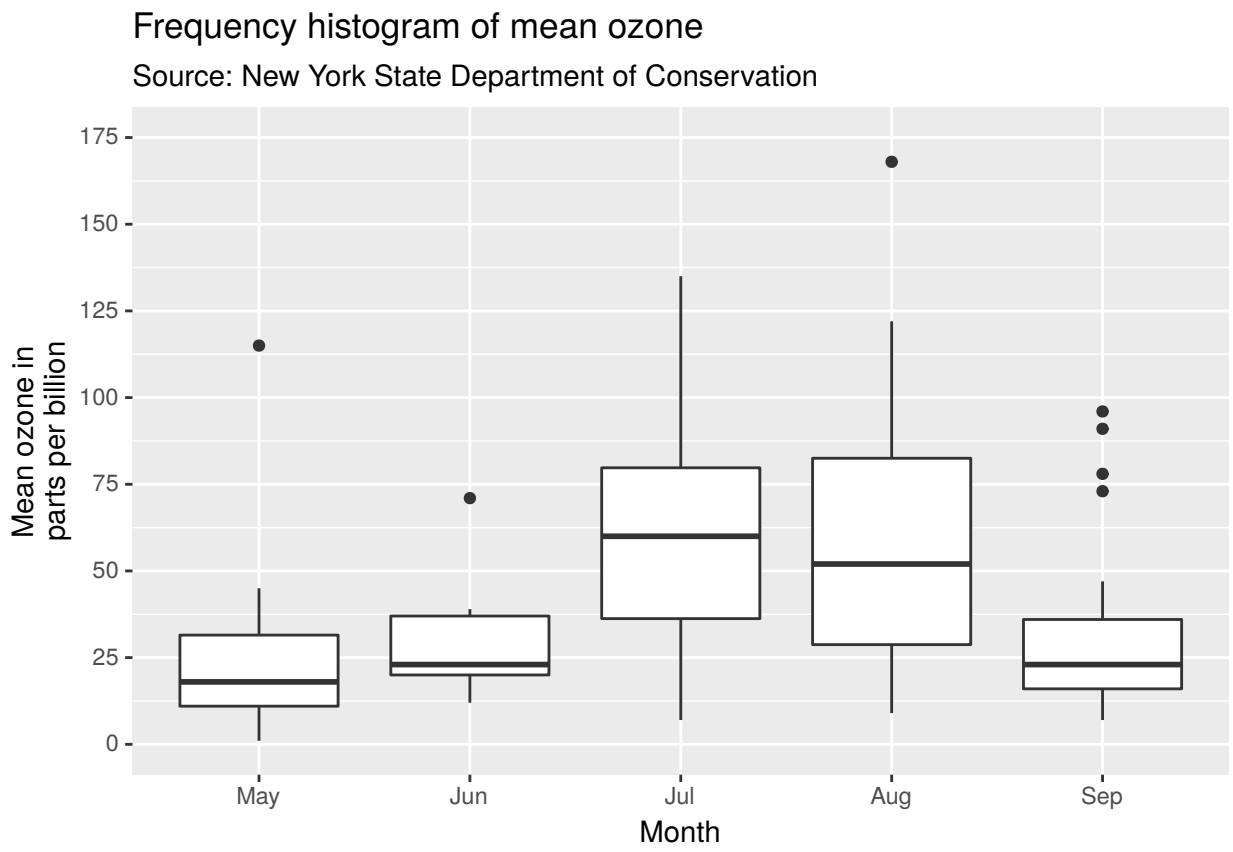
```
p10 <- p10 + scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175))
p10
```



10.5. Adding a title

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

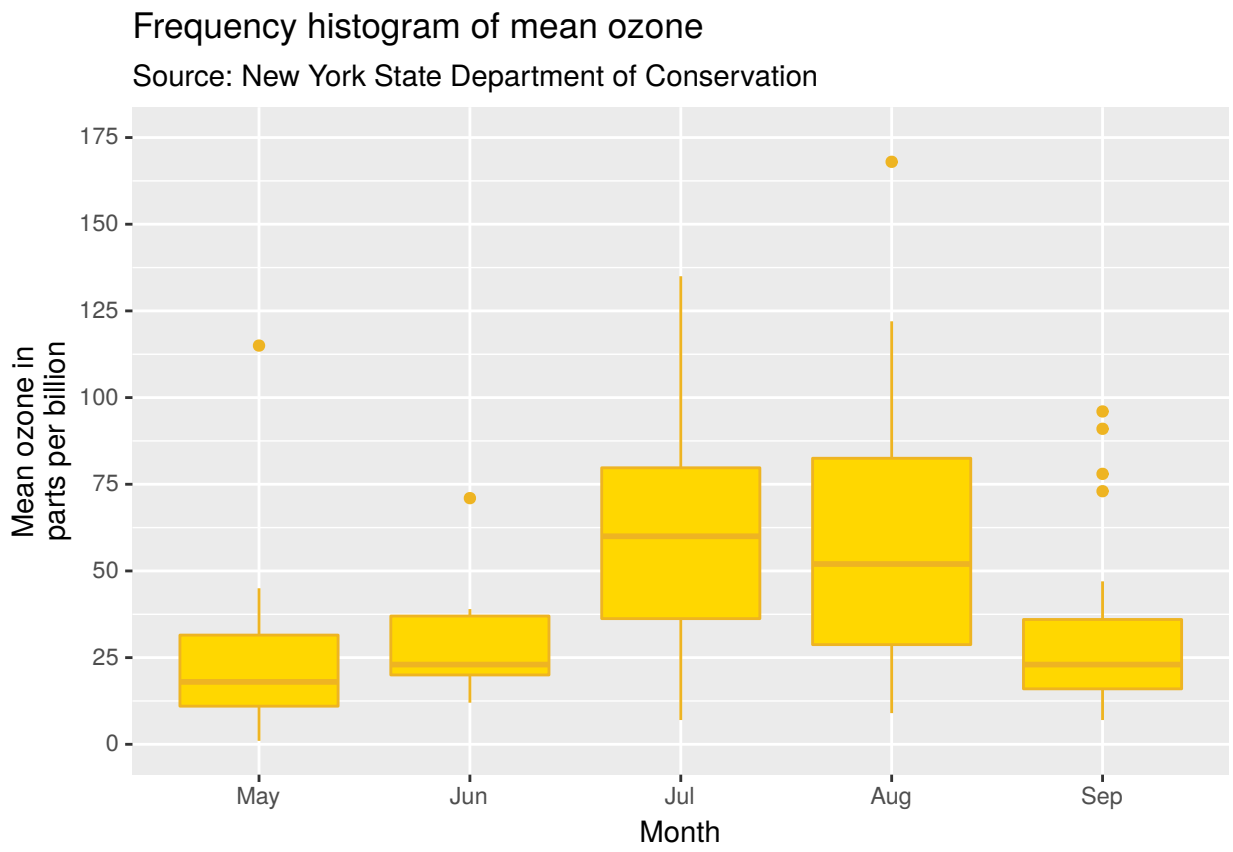
```
p10 <- p10 +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p10
```



10.6. Changing the colour of the boxes

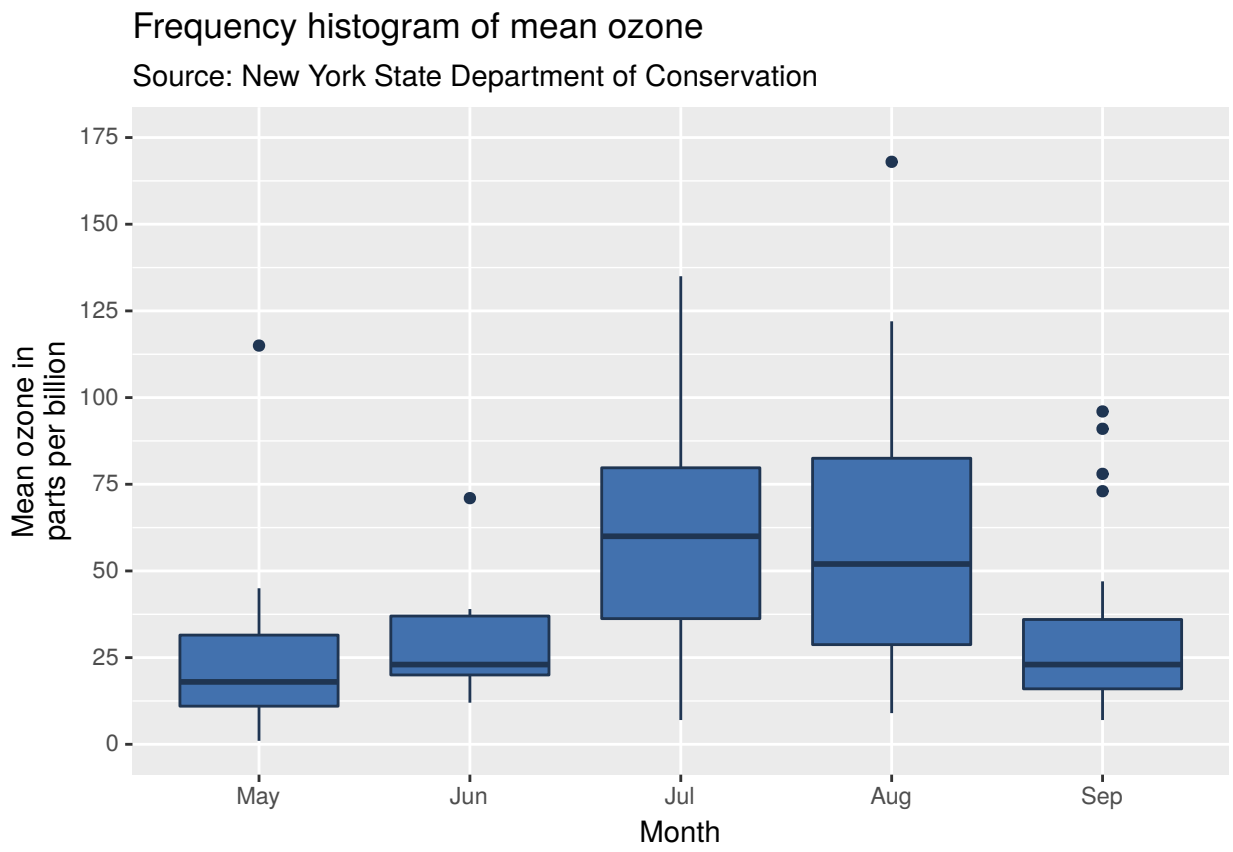
To change the line and fill colours of the box plot, we add a valid colour to the `colour` and `fill` arguments in `geom_boxplot()` (note that we assigned these colours to variables outside of the plot to make it easier to change them). A list of valid colours is here.

```
fill <- "gold1"; line <- "goldenrod2"

p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(fill = fill, colour = line) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p10
```



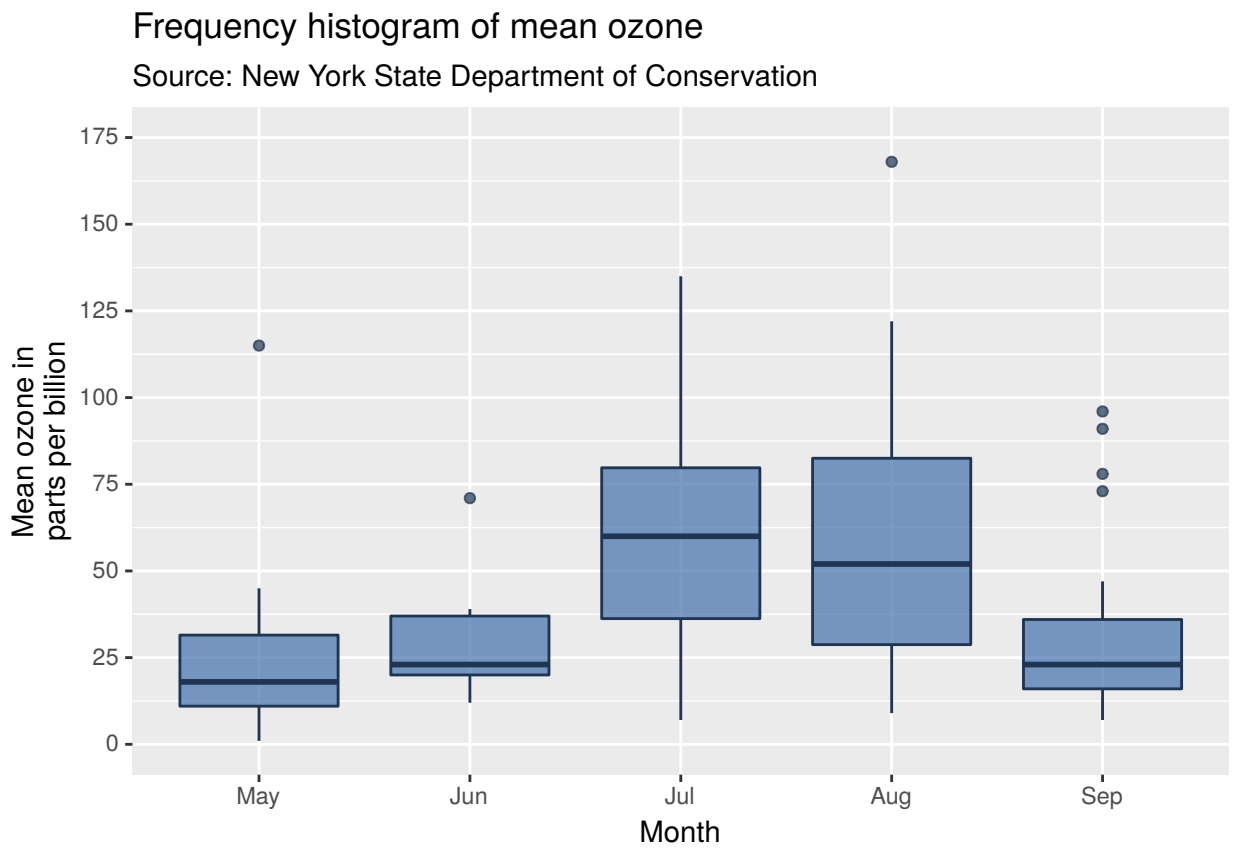
If you want to go beyond the options in the list above, you can also specify exact HEX colours by including them as a string preceded by a hash, e.g., "#FFFFFF". Below, we have called two shades of blue for the fill and lines using their HEX codes.

```
fill <- "#4271AE"; line <- "#1F3552"

p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(fill = fill, colour = line) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p10
```



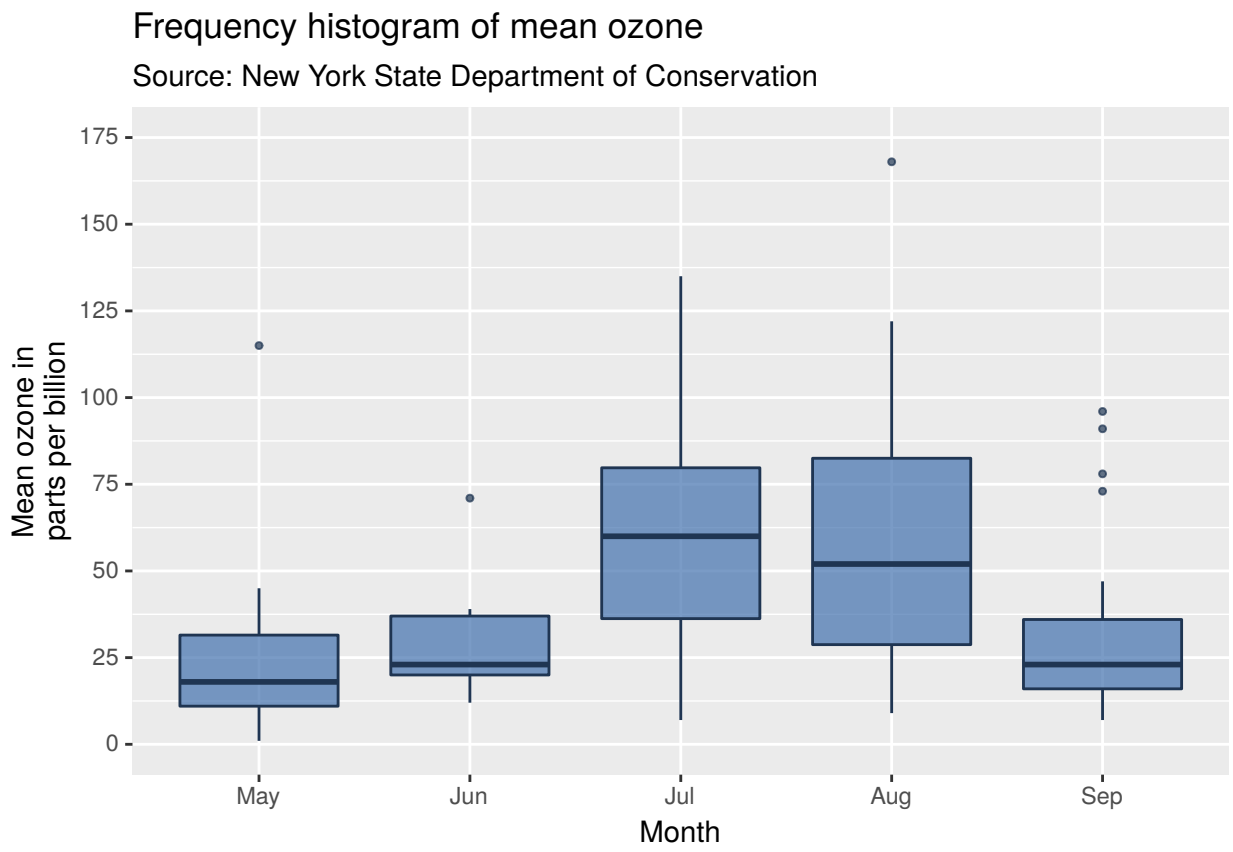
You can also specify the degree of transparency in the box fill area using the argument `alpha` in `geom_boxplot`. This ranges from 0 to 1.

```
p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(fill = fill, colour = line,
    alpha = 0.7) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p10
```



Finally, you can change the appearance of the outliers as well, using the arguments `outlier.colour` and `outlier.shape` in `geom_boxplot` to change the colour and shape respectively. An explanation of the allowed arguments for shape are described in this article, although be aware that because there is no "fill" argument for outlier, you cannot create circles with separate outline and fill colours. Here we will make the outliers small solid circles (using `outlier.shape = 20`) and make them the same colour as the box lines (using `outlier.colour = "#1F3552"`).

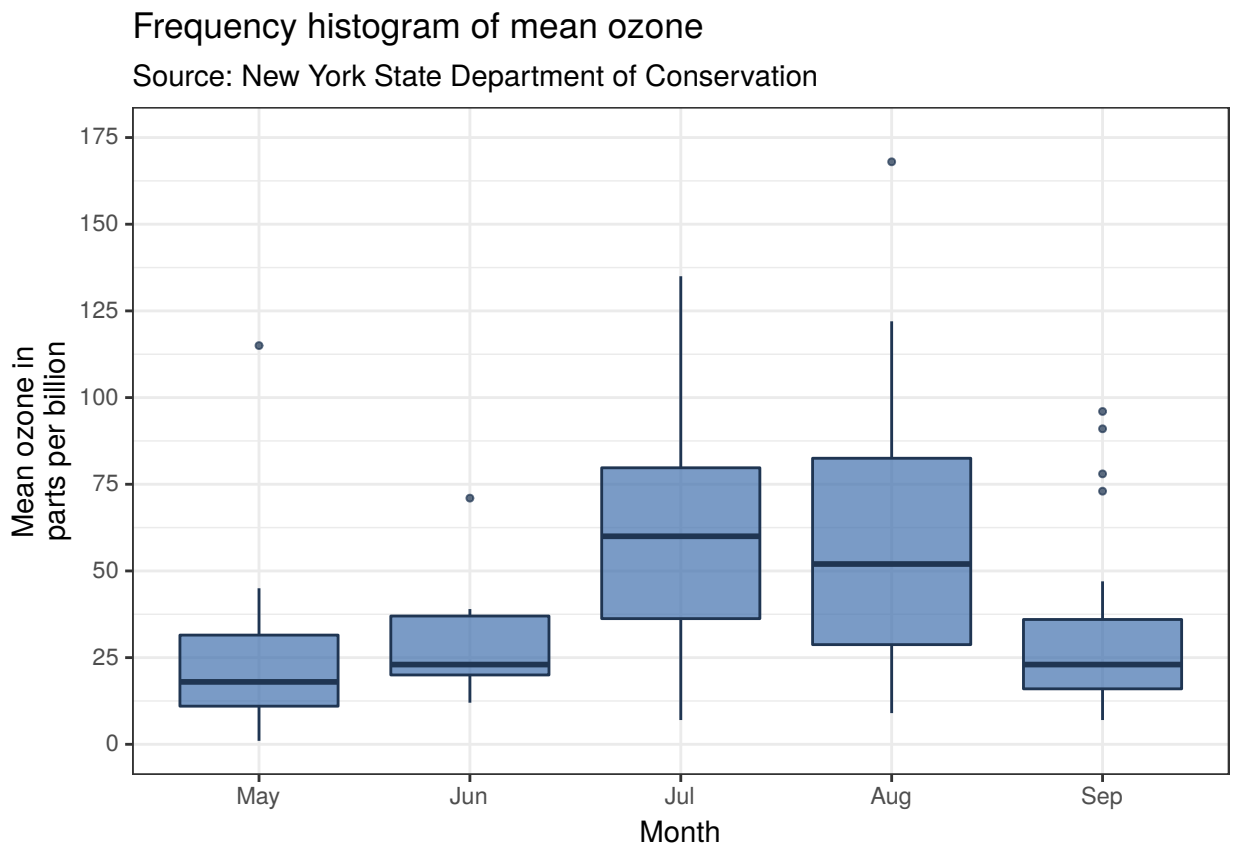
```
p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(fill = fill, colour = line, alpha = 0.7,
    outlier.colour = "#1F3552", outlier.shape = 20) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")
p10
```



10.7. 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.

```
p10 <- p10 + theme_bw()
p10
```



10.8. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

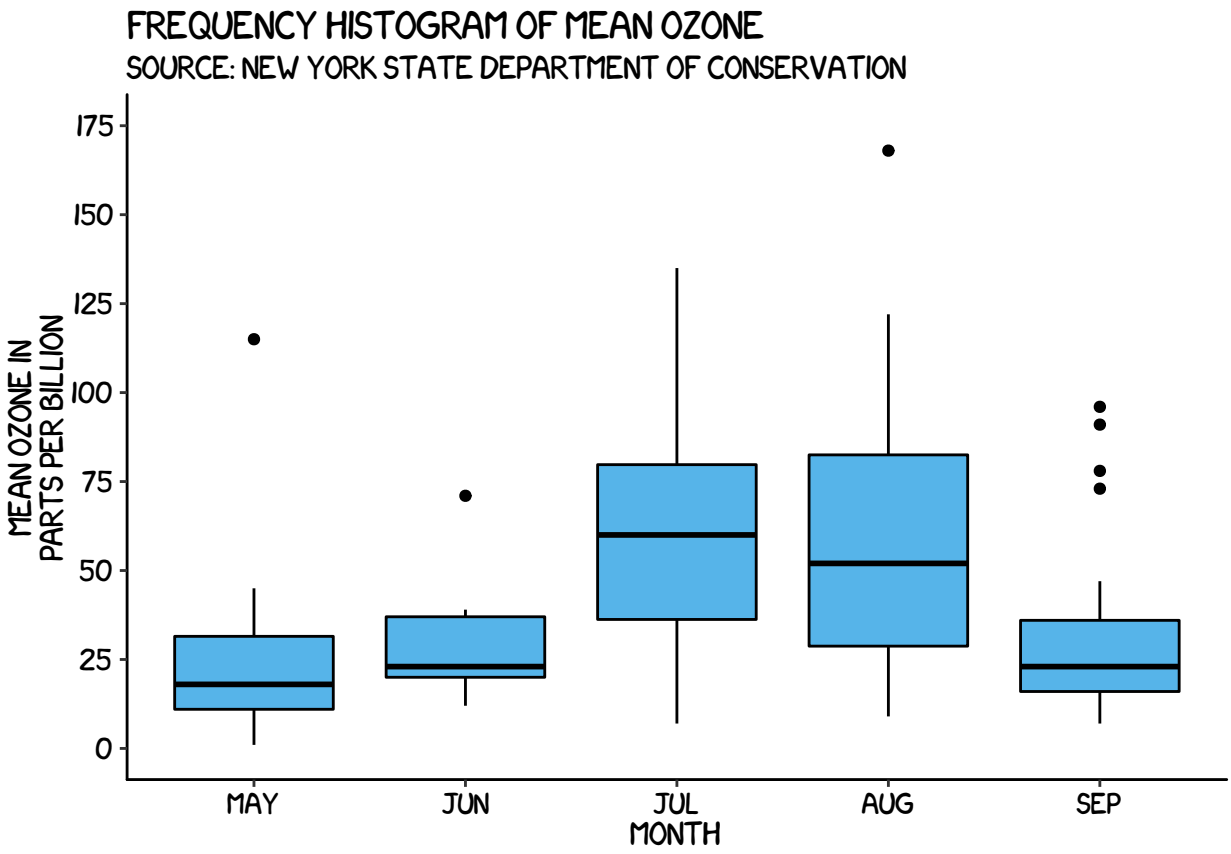
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

10.9. 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.

```
p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(colour = "black", fill = "#56B4E9") +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
  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.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p10
```

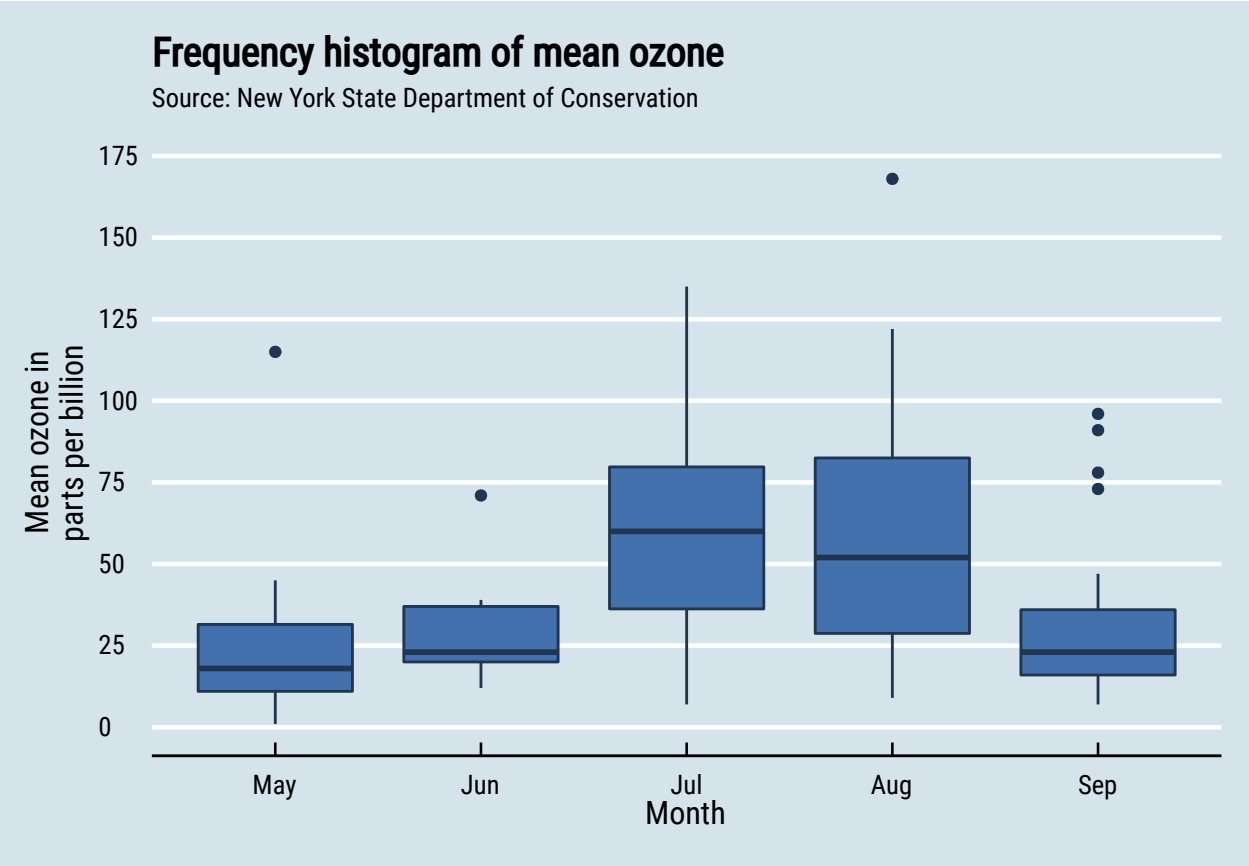


10.10. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(fill = fill, colour = line) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
  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 = "Roboto Condensed"),
```

```
    plot.title = element_text(family = "Roboto Condensed"))
p10
```

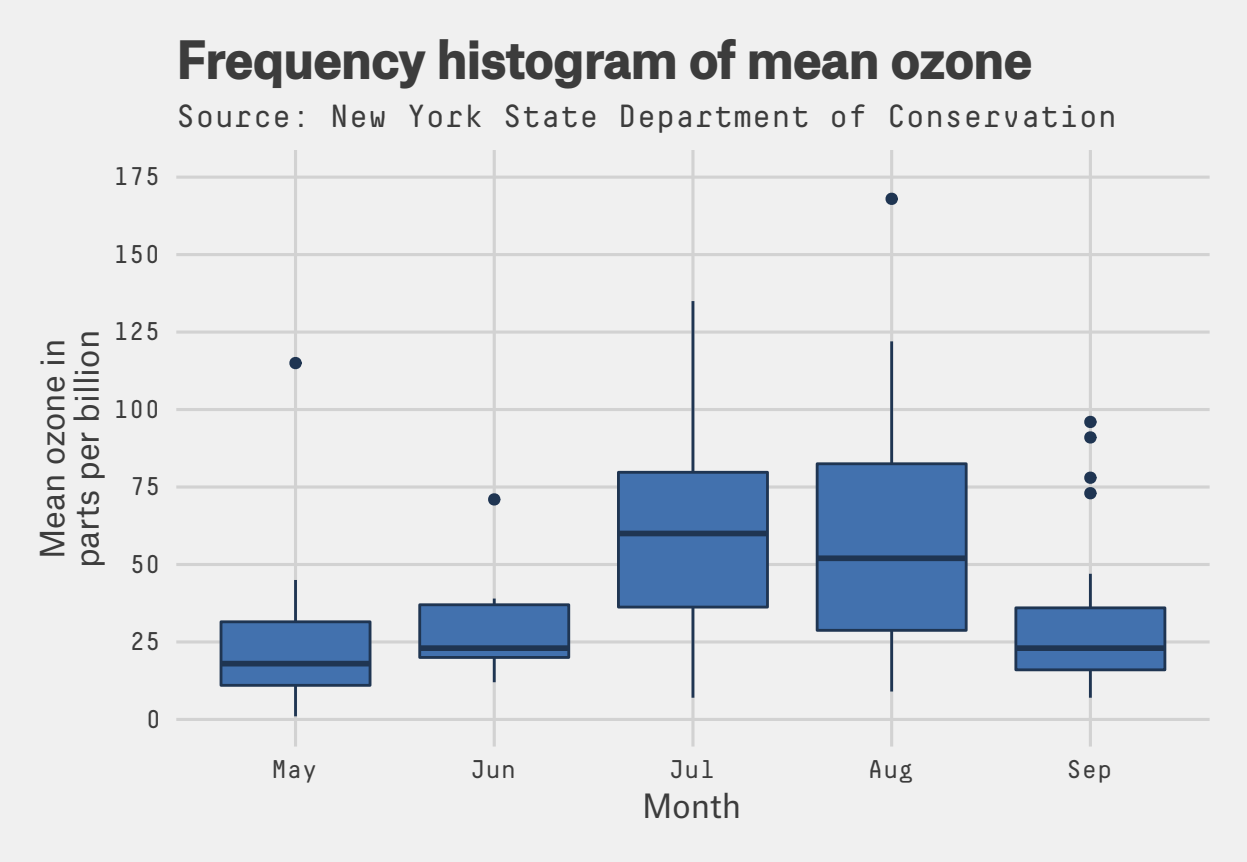


10.11. Using 'Five Thirty Eight' theme

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'.

```
p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(fill = fill, colour = line) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
```

```
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 10),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 10),
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    text = element_text(family = "Decima Mono Pro"))
p10
```

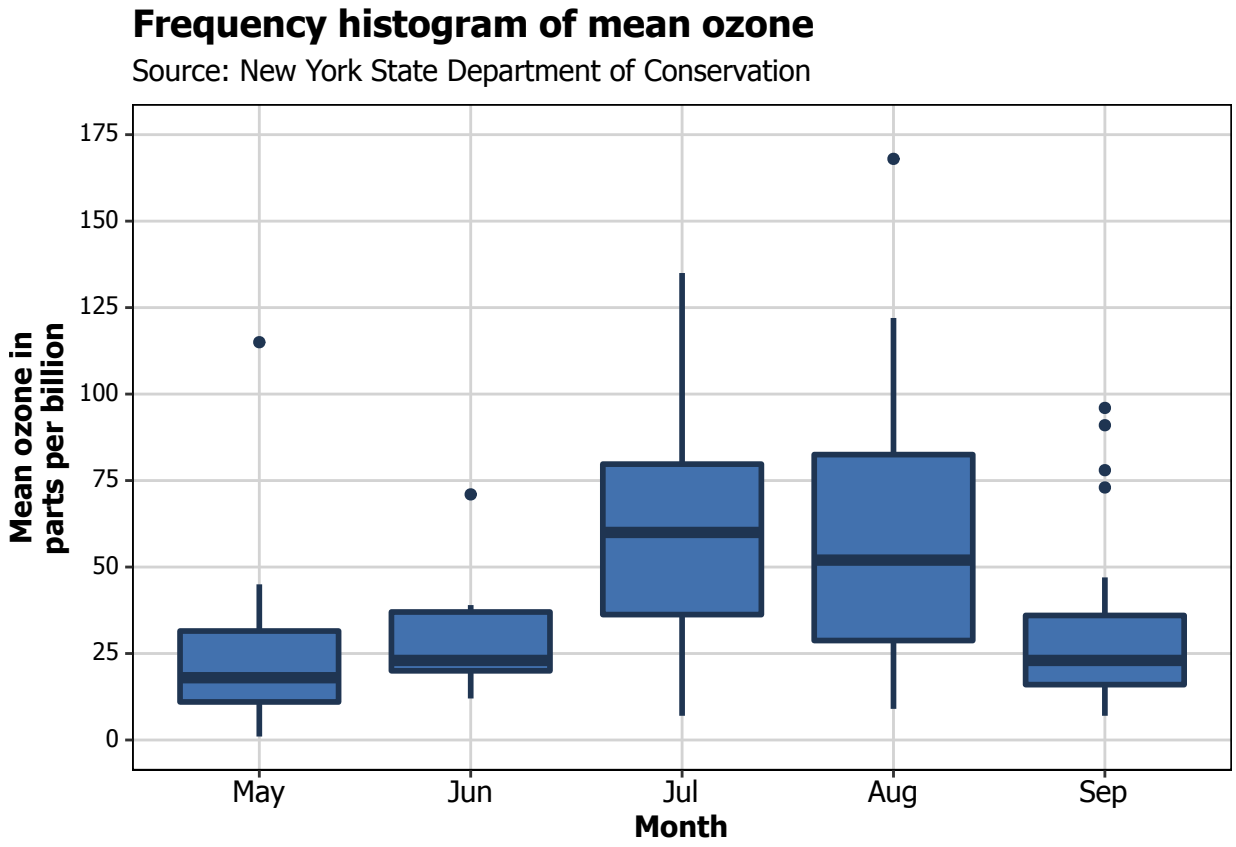


10.12. Creating your own theme

As before, you can modify your plots a lot as `ggplot2` allows many customisations. Here is a custom plot where we have modified the axes, background and font.

```
fill <- "#4271AE"; lines <- "#1F3552"

p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(colour = lines, fill = fill,
    size = 1) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
```

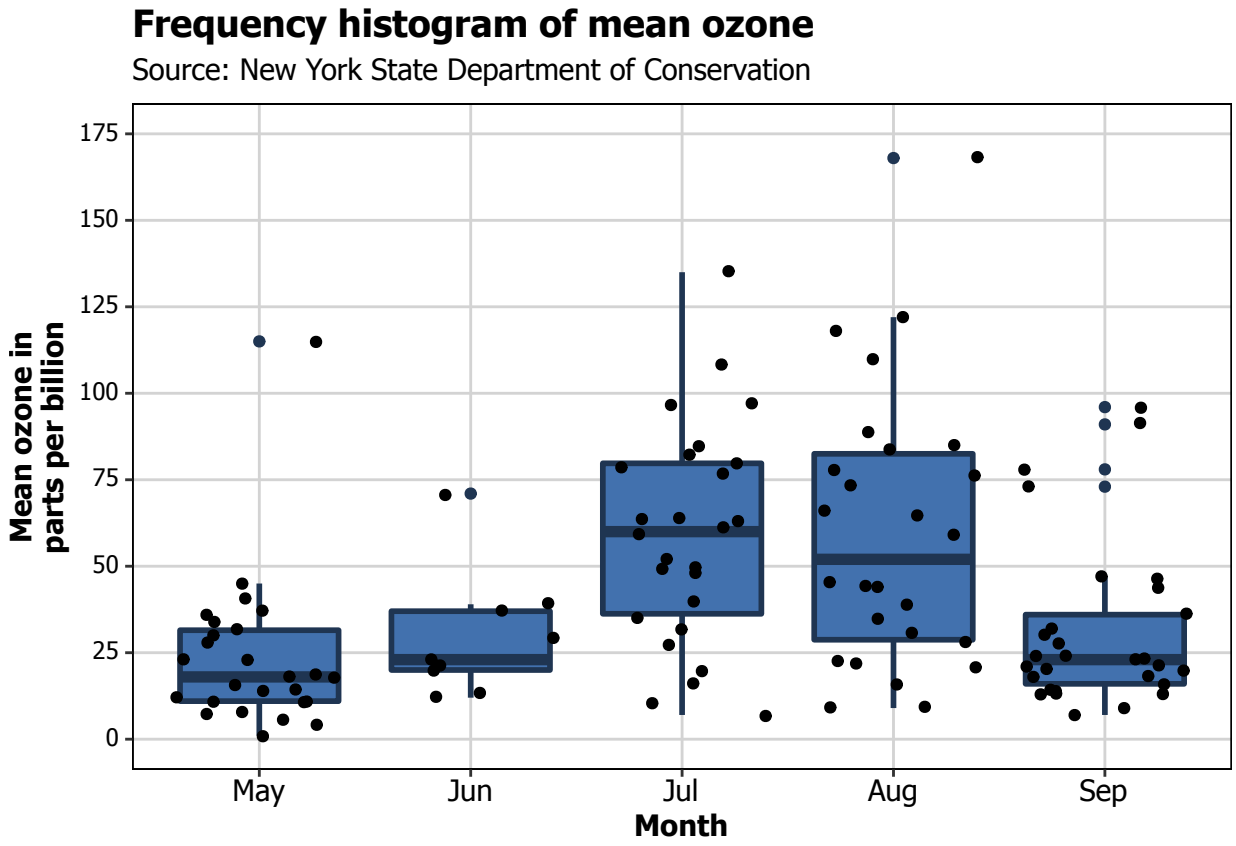
```
  theme_bw() +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"),
    axis.title = element_text(face = "bold"),
    axis.text.x = element_text(colour = "black", size = 11),
    axis.text.y = element_text(colour = "black", size = 9))
p10
```



10.13. Boxplot extras

An extra feature you can add to boxplots is to overlay all of the points for that group on each boxplot in order to get an idea of the sample size of the group. This can be achieved using by adding the `geom_jitter` option.

```
p10 <- p10 + geom_jitter()
p10
```

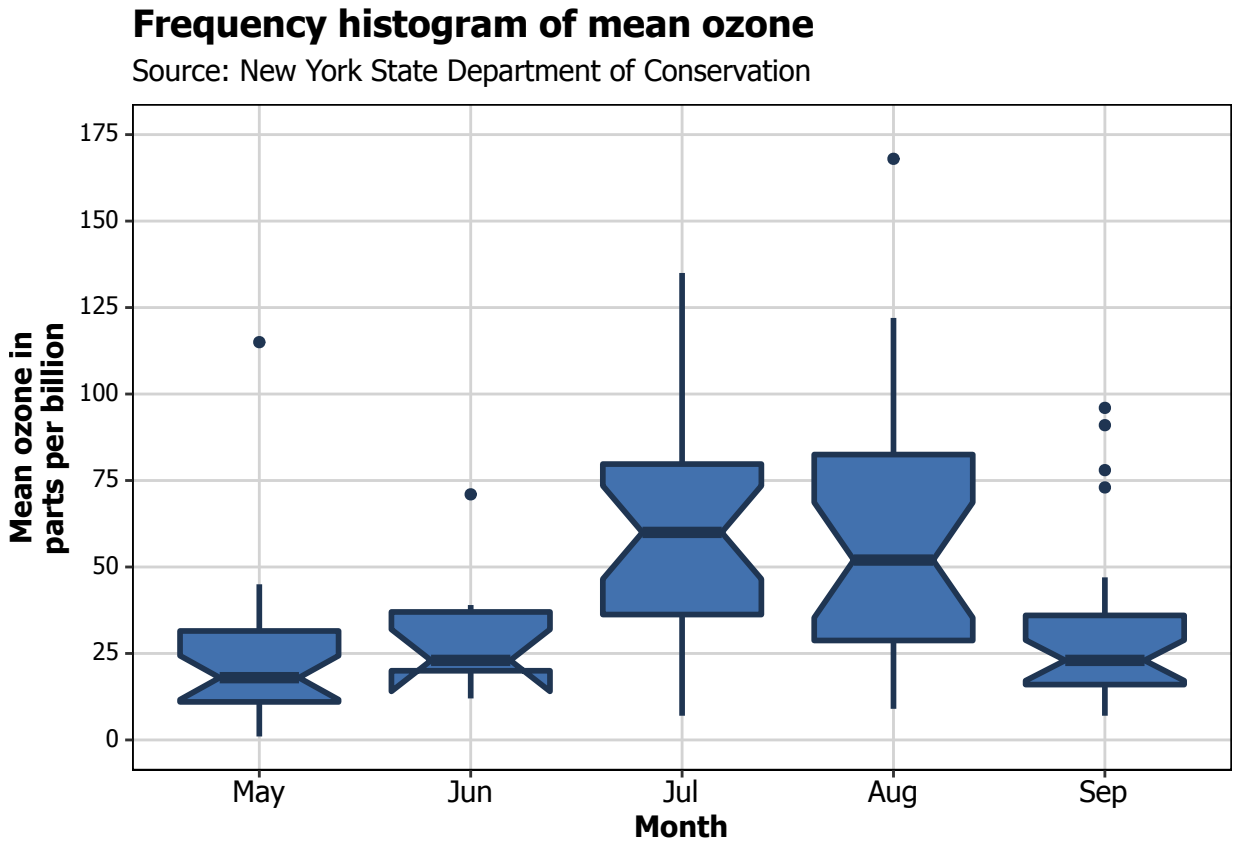


We can see that June has a pretty small sample, indicating that information based on this group may not be very reliable.

Another thing you can do with your boxplot is add a notch to the box where the median sits to give a clearer visual indication of how the data are distributed within the IQR. You achieve this by adding the argument `notch = TRUE` to the `geom_boxplot` option. You can see on our graph that the box for June looks a bit weird due to the very small gap between the 25th percentile and the median.

```
p10 <- ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot(colour = lines, fill = fill,
    size = 1, notch = TRUE) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
  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.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"),
    axis.title = element_text(face = "bold"),
```

```
    axis.text.x = element_text(colour = "black", size = 11),
    axis.text.y = element_text(colour = "black", size = 9))
p10
```



10.14. Grouping by another variable

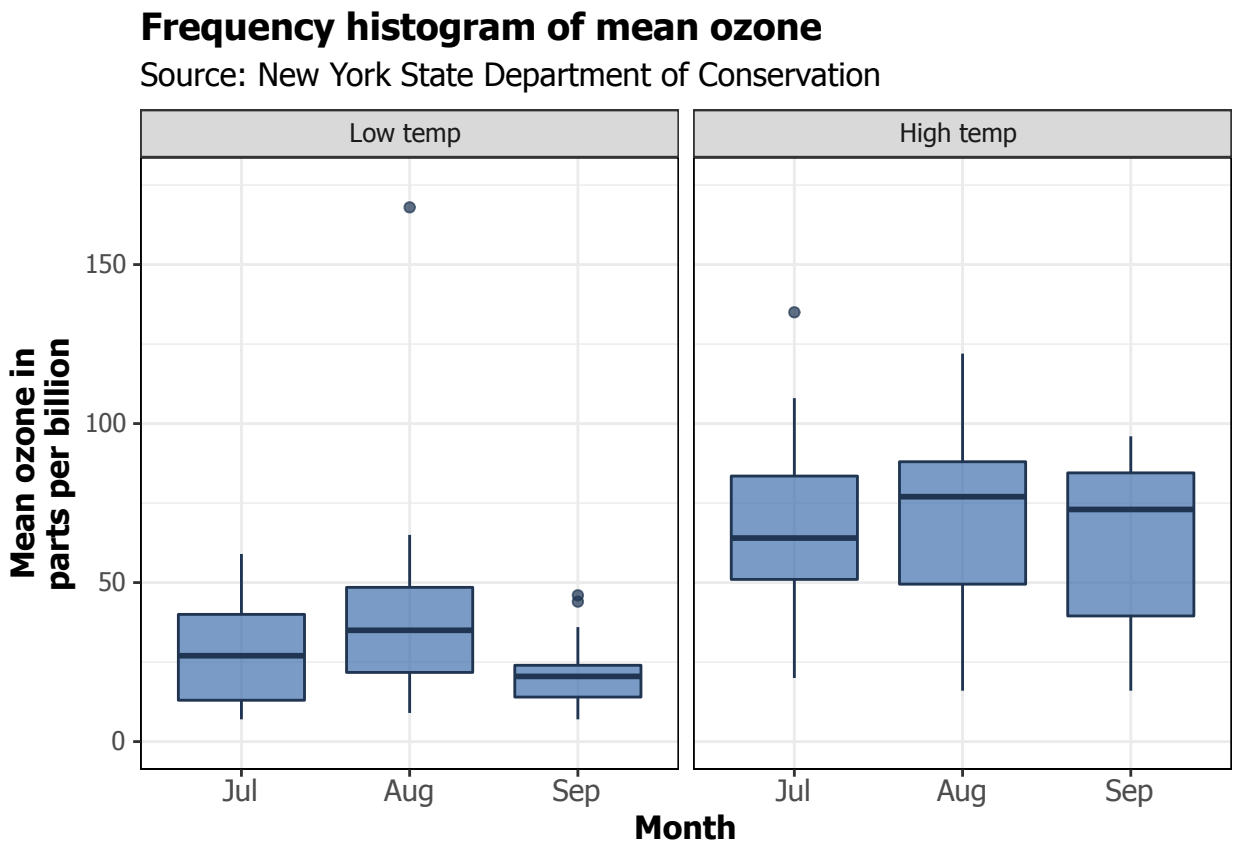
You can also easily group box plots by the levels of another variable. There are two options, in separate (panel) plots, or in the same plot.

We first need to do a little data wrangling. In order to make the graphs a bit clearer, we've kept only months "July", "Aug" and "Sep" in a new dataset `airquality_trimmed`. We've also mean-split `Temp` so that this is also categorical, and made it into a new labelled factor variable called `Temp.f`.

In order to produce a panel plot by Temperature , we add the `facet_grid(. ~ Temp.f)` option to the plot.

```
airquality_trimmed <- airquality %>%
  filter(Month %in% c("Jul", "Aug", "Sep")) %>%
  mutate(Temp.f = factor(ifelse(Temp > mean(Temp), 1, 0),
                         labels = c("Low temp ", "High temp ")))

p10 <- ggplot(airquality_trimmed, aes(x = Month, y = Ozone)) +
```

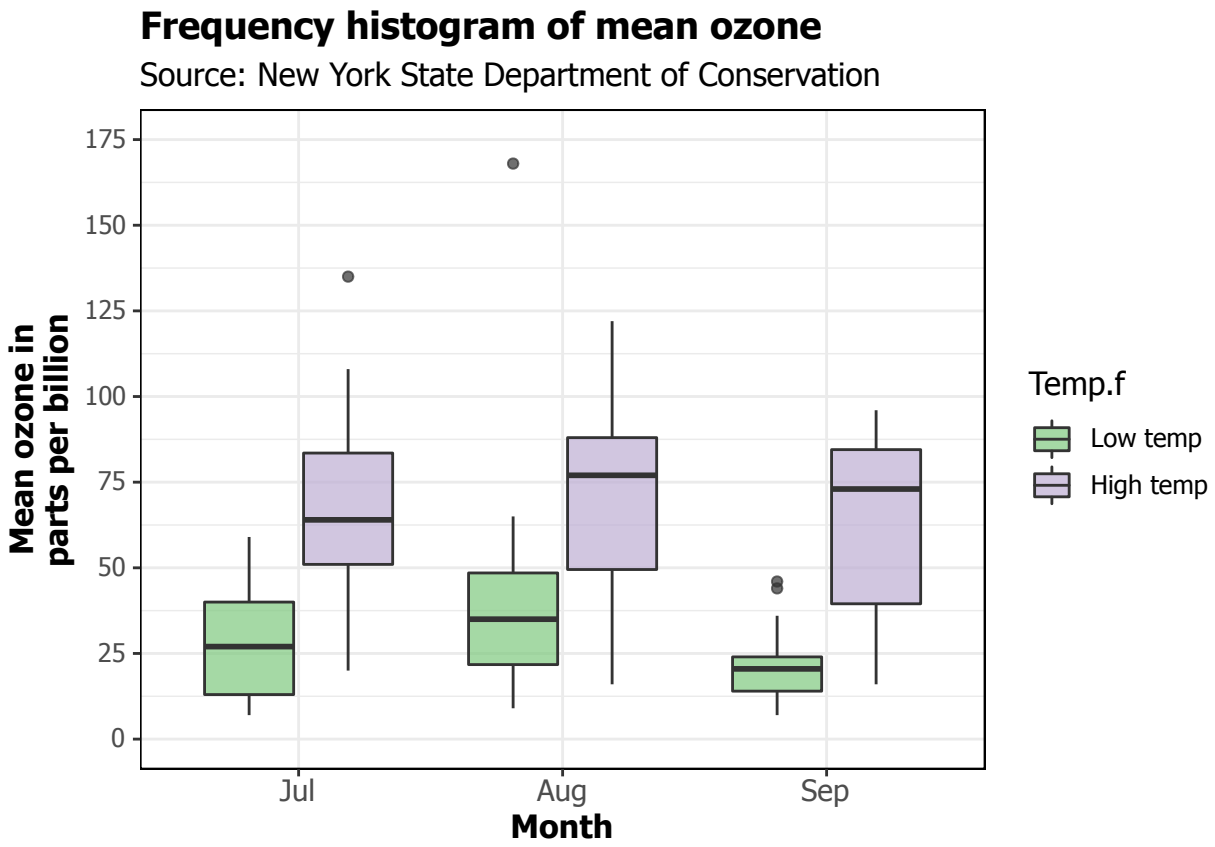
```
  geom_boxplot(fill = fill, colour = line,
    alpha = 0.7) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
    breaks = seq(0, 175, 50), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
  theme_bw() +
  theme(plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    panel.border = element_rect(colour = "black", fill = NA, size = .5),
    text = element_text(size = 12, family = "Tahoma"),
    axis.title = element_text(face = "bold"),
    axis.text.x = element_text(size = 11)) +
  facet_grid(. ~ Temp.f)
p10
```



In order to plot the two Temperature levels in the same plot, we need to add a couple of things. Firstly, in the `ggplot` function, we add a `fill = Temp.f` argument to `aes`. Secondly, we customise the colours of the boxes by adding the `scale_fill_brewer` to the plot from the `RColorBrewer` package. This blog post describes the available packages.

```
p10 <- ggplot(airquality_trimmed, aes(x = Month, y = Ozone, fill = Temp.f)) +
  geom_boxplot(alpha = 0.7) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
```

```
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
  theme_bw() +
  theme(plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    panel.border = element_rect(colour = "black", fill = NA, size = .5),
    text = element_text(size = 12, family = "Tahoma"),
    axis.title = element_text(face = "bold"),
    axis.text.x = element_text(size = 11)) +
  scale_fill_brewer(palette = "Accent")
p10
```

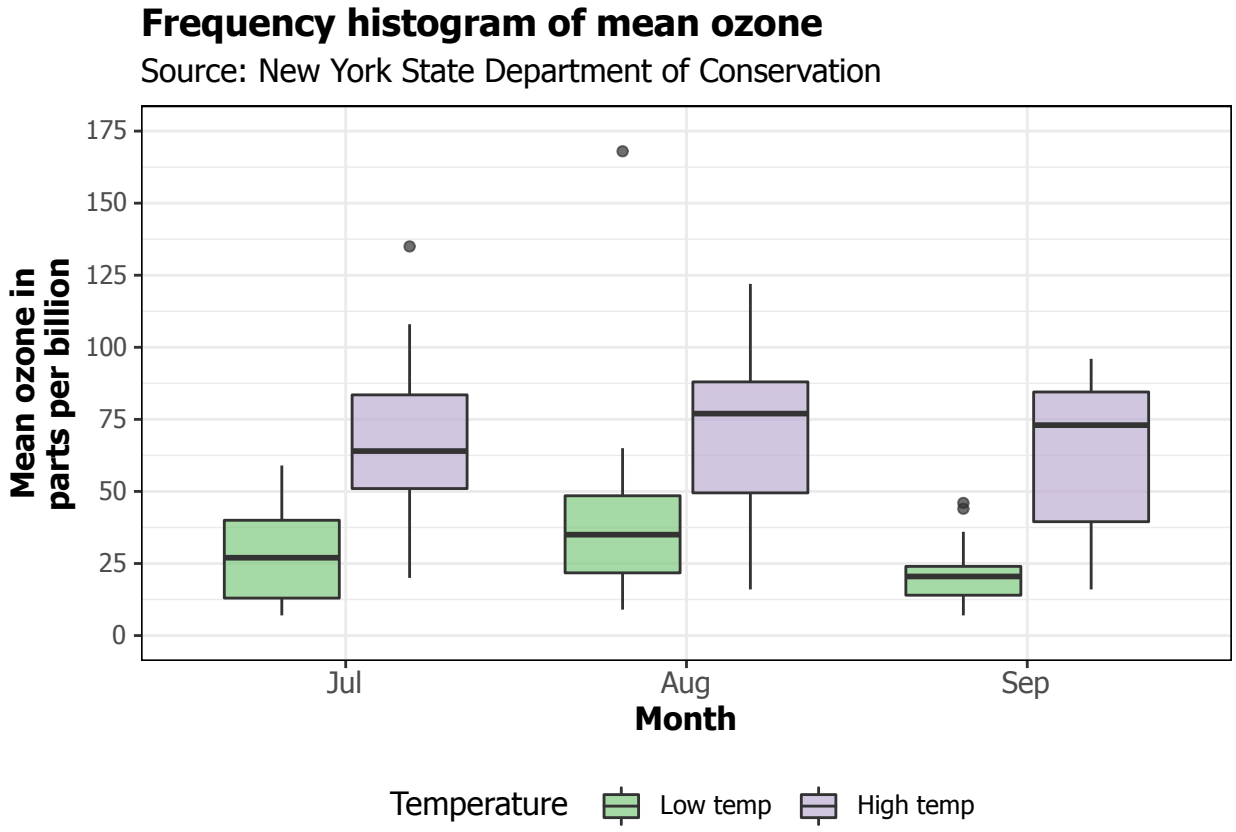


10.15. Formatting the legend

Finally, we can format the legend. Firstly, we can change the position by adding the `legend.position = "bottom"` argument to the `theme` option, which moves the legend under the plot. Secondly, we can fix the title by adding the `labs(fill = "Temperature ")` option to the plot.

```
p10 <- ggplot(airquality_trimmed, aes(x = Month, y = Ozone, fill = Temp.f)) +
  geom_boxplot(alpha = 0.7) +
  scale_y_continuous(name = "Mean ozone in\nparts per billion",
```

```
    breaks = seq(0, 175, 25), limits = c(0, 175)) +
  scale_x_discrete(name = "Month") +
  labs(title = "Frequency histogram of mean ozone",
       subtitle = "Source: New York State Department of Conservation")  +
  theme_bw() +
  theme(plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    panel.border = element_rect(colour = "black", fill = NA, size = .5),
    text = element_text(size = 12, family = "Tahoma"),
    axis.title = element_text(face = "bold"),
    axis.text.x = element_text(size = 11),
    legend.position = "bottom") +
  scale_fill_brewer(palette = "Accent") +
  labs(fill = "Temperature ")
p10
```



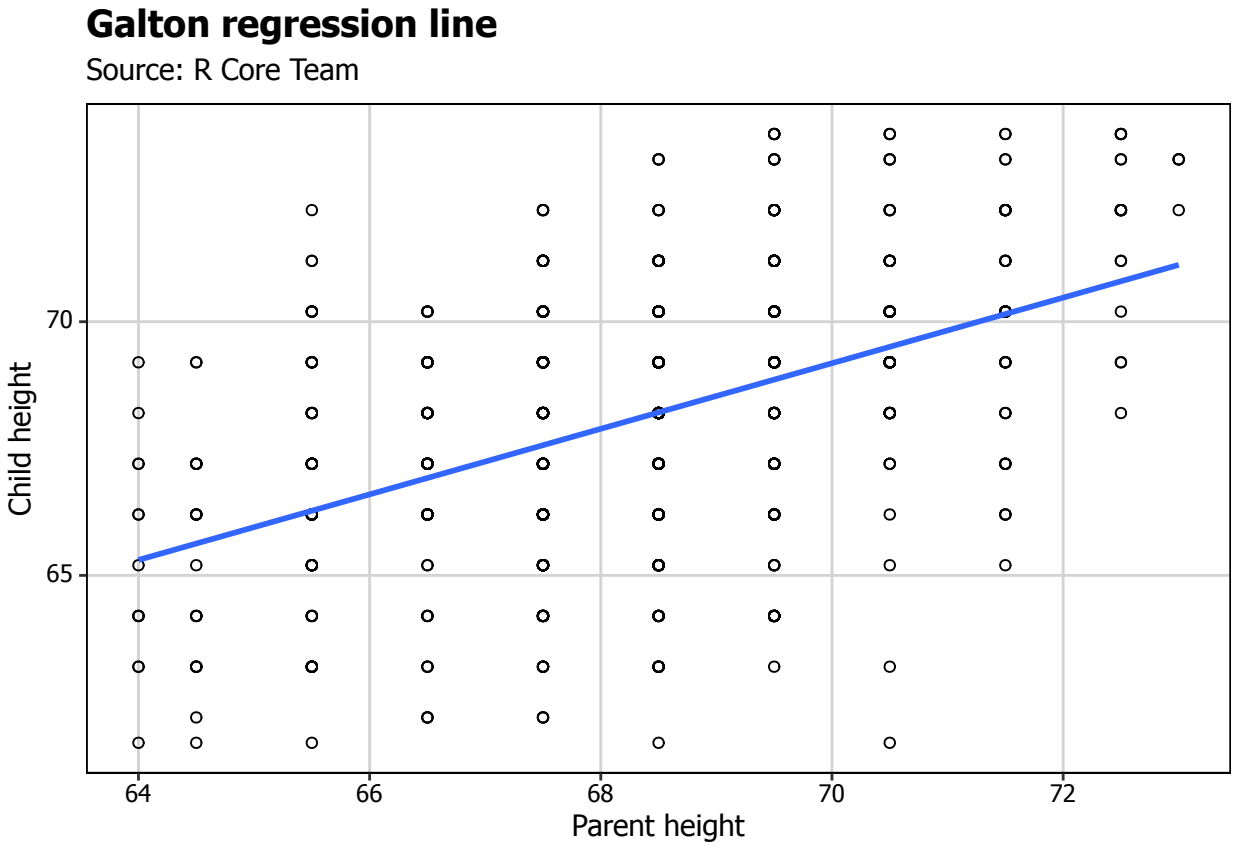
CHAPTER 11

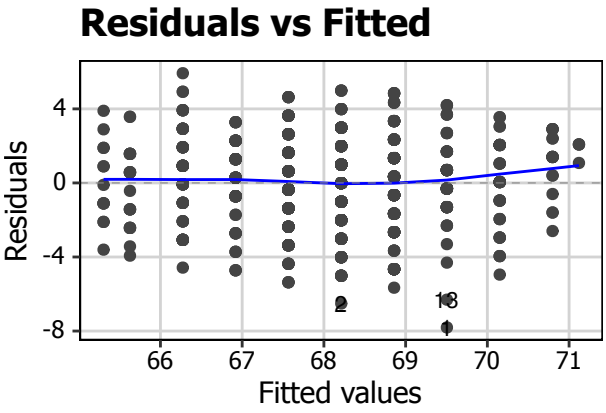
Linear regression plots

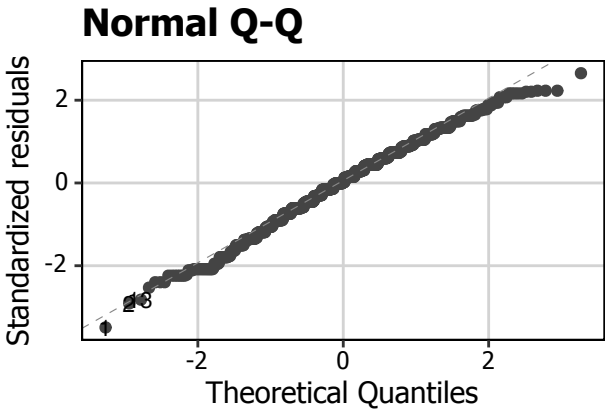
11.1. Introduction

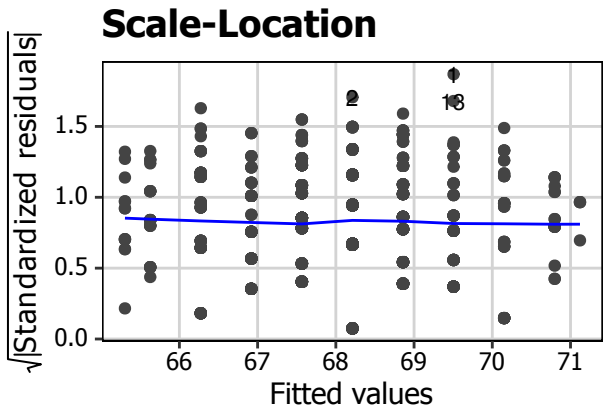
This chapter will be much more than showing you how to create regression plots. Here we are explaining regression results from Galton dataset created by Galton himself who also developed Econometrics to explore his data on the relationship between parents and children height by the end of the 19th century.

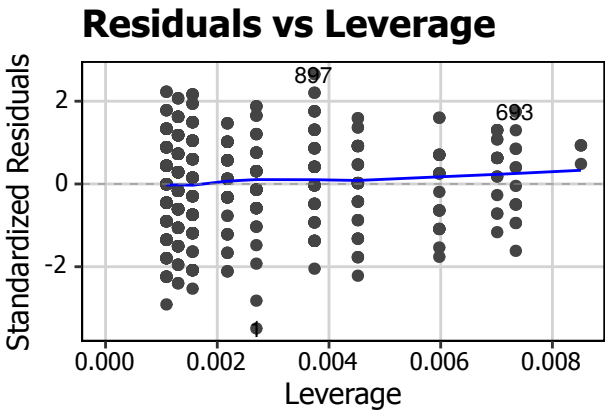
In this tutorial, we will work towards creating the trend line and diagnostics plots below. We will take you from a basic regression plot and explain all the customisations we add to the code step-by-step.


Galton regression line
Source: R Core Team
Child height
Parent height
70
65
64
66
68
70
72


Residuals vs Fitted
Residuals
Fitted values
4
0
-4
-8
66
67
68
69
70
71


Normal Q-Q
Standardized residuals
Theoretical Quantiles
2
0
-2
-2
0
2


Scale-Location
√|Standardized residuals|
Fitted values
1.5
1.0
0.5
0.0
66
67
68
69
70
71


Residuals vs Leverage
Standardized Residuals
Leverage
897
693
2
0
-2
0.000
0.002
0.004
0.006
0.008

The first thing to do is download and load in the libraries and the data.

```
if (!require("pacman")) install.packages("pacman")
p_load(ggplot2, ggthemes, grid, dplyr, HistData)
p_load_gh("sinhrks/ggfortify")

Galton <- as_tibble(Galton)
```

11.2. Trend line plot

11.2.1. Basic trend line plot

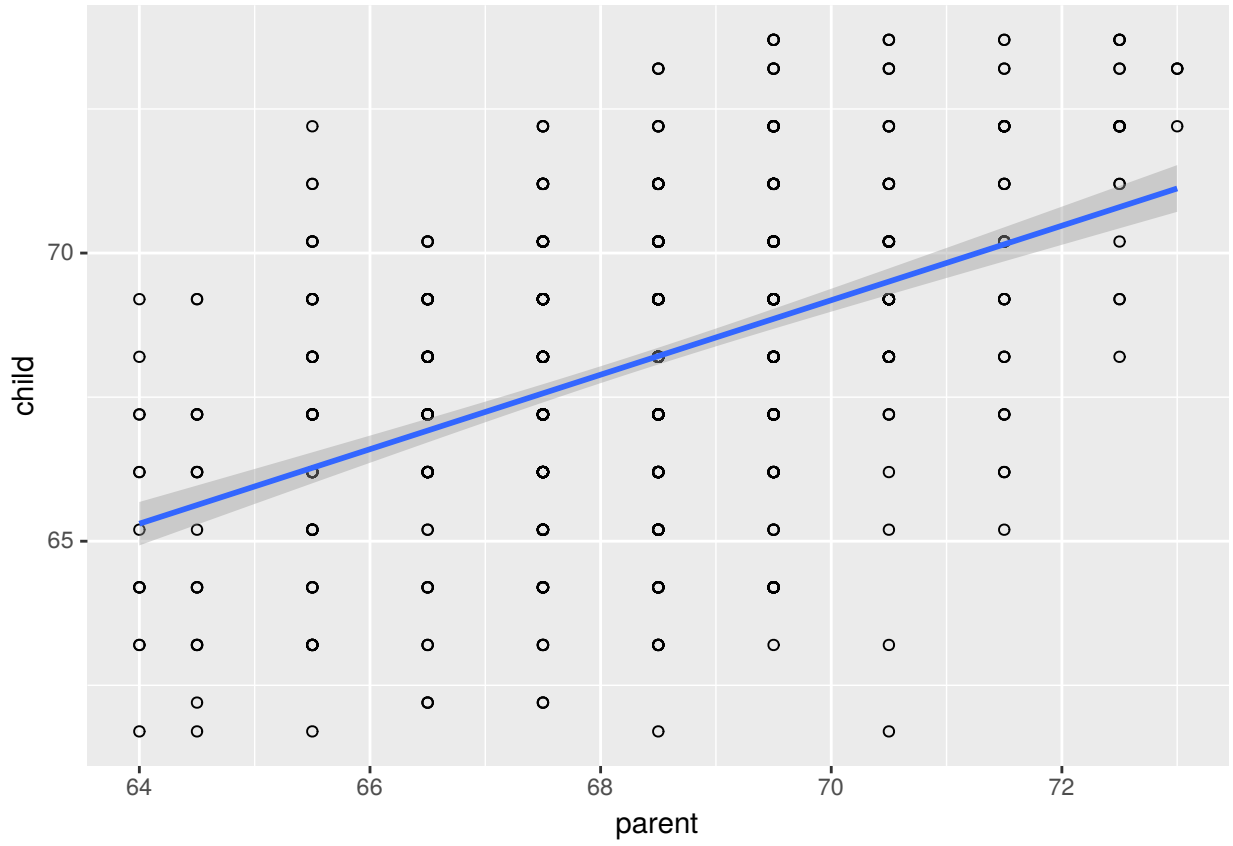
We'll be estimating a model of the form $y_i = \beta_0 + \beta_1 x_i + e_i$ where (x_i, y_i) is an observation of the height of a parent and his child.

```
fit <- lm(child ~ parent, data = Galton)
summary(fit)
```

```
Call:
lm(formula = child ~ parent, data = Galton)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.94153    2.81088   8.517   <2e-16 ***
parent       0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

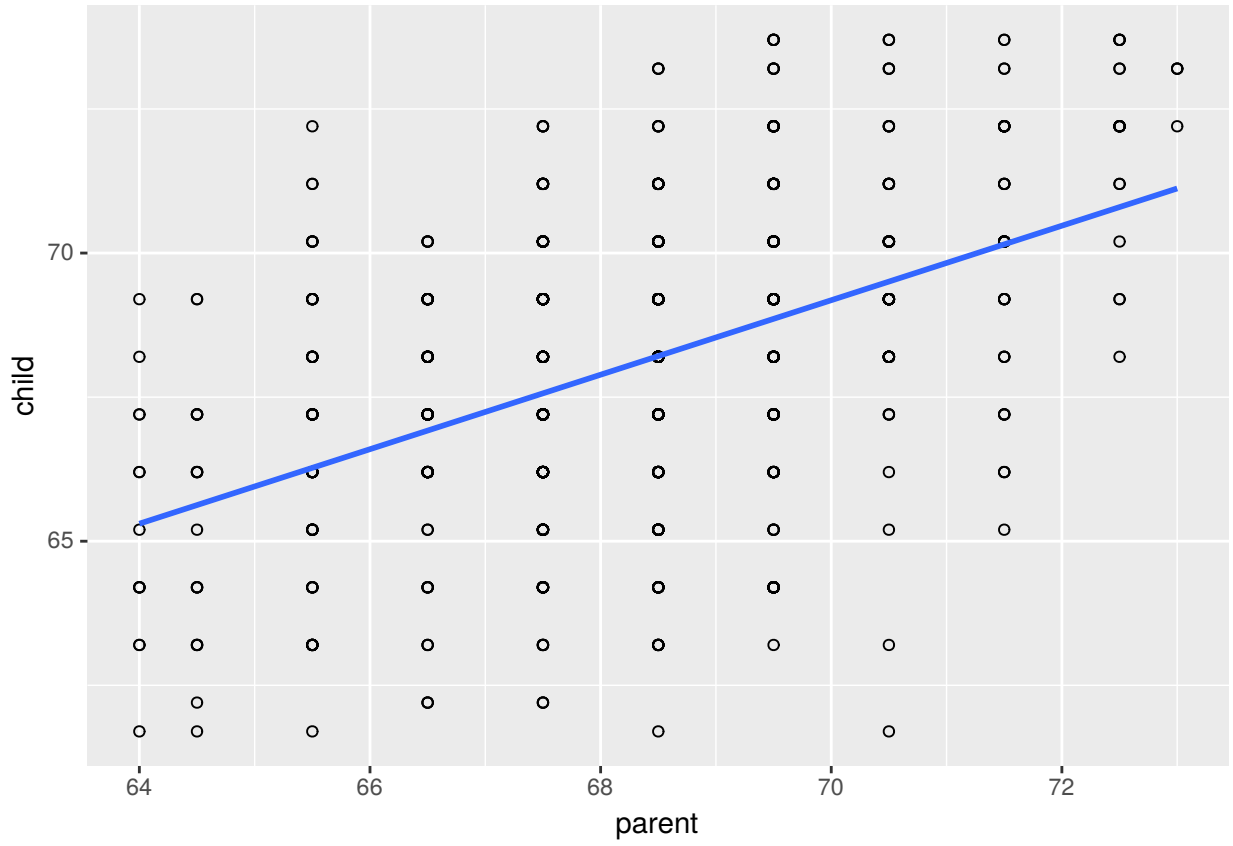
Up to this point we have all we need to plot regressions. We will start with a basic regression plot.

```
p11 <- ggplot(Galton, aes(x = parent, y = child)) +
  geom_point(shape = 1) + geom_smooth(method = lm)
p11
```



`geom_smooth` can be customized, for example, not to include the confidence region

```
p11 <- ggplot(Galton, aes(x = parent, y = child)) +
  geom_point(shape = 1) + geom_smooth(method = lm, se = FALSE)
p11
```

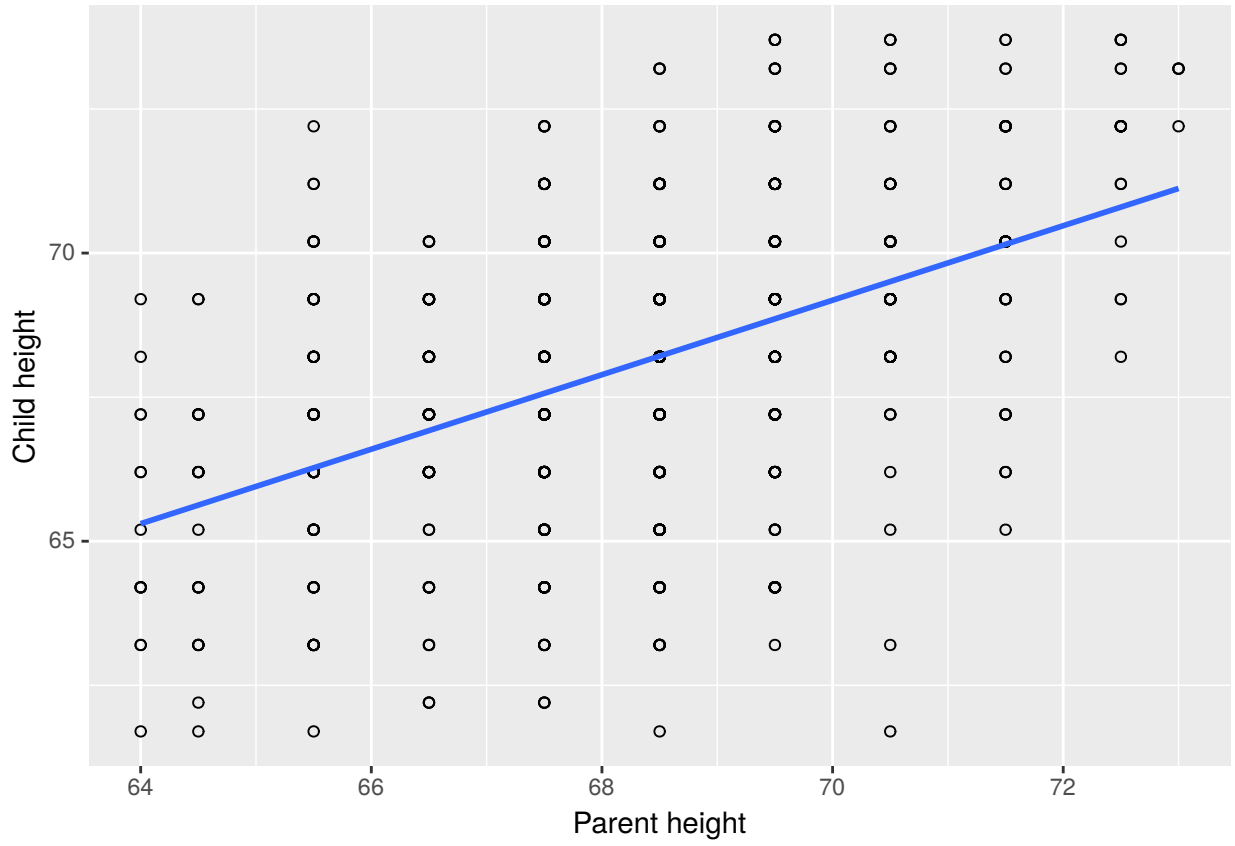


Before continuing it is a good idea to fix the axis labels and add a title.

11.2.2. Customising axis labels

We can change the text of the axis labels using the `scale_x_continuous` and `scale_y_continuous` options, with the names passed as a string to the `name` arguments in each.

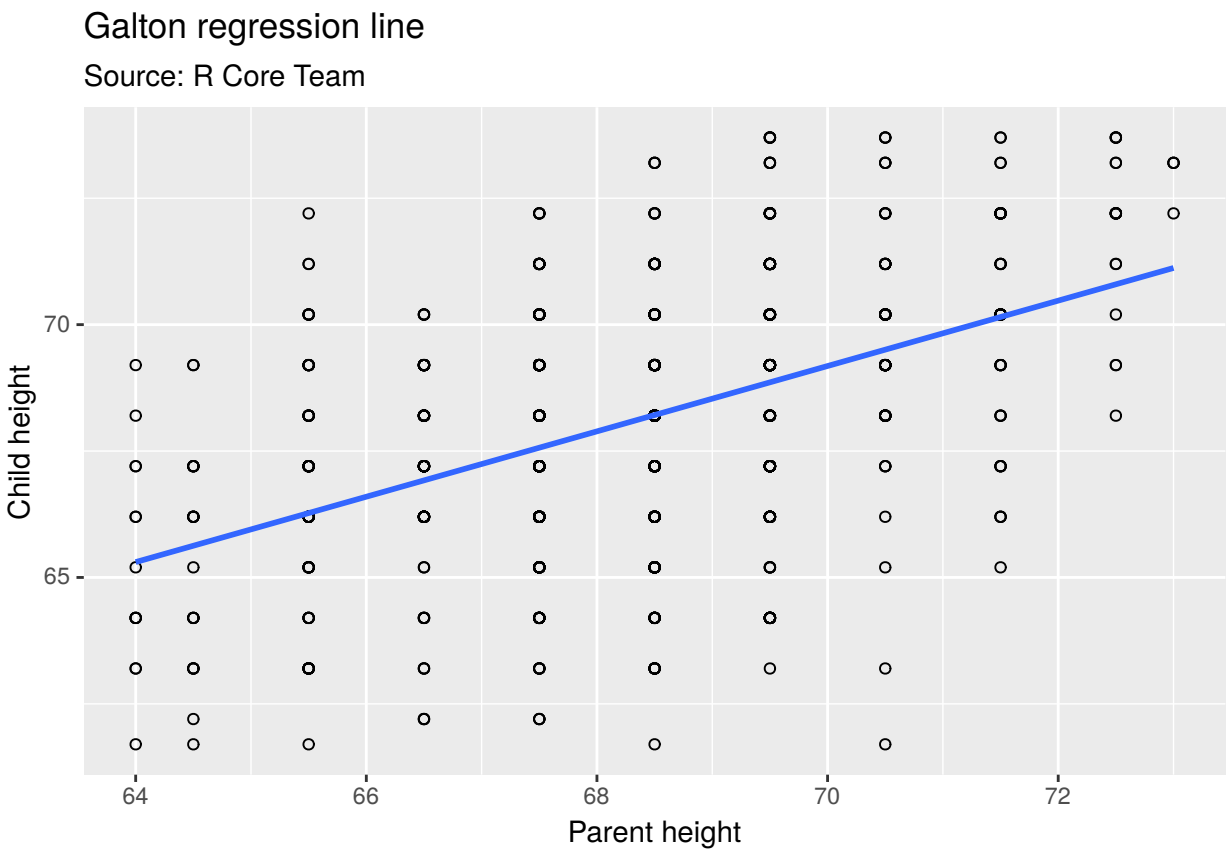
```
p11 <- p11 + scale_x_continuous(name = "Parent height") +
  scale_y_continuous(name = "Child height")
p11
```



11.2.3. Adding a title

Similarly, we can add a title using the `labs` option.

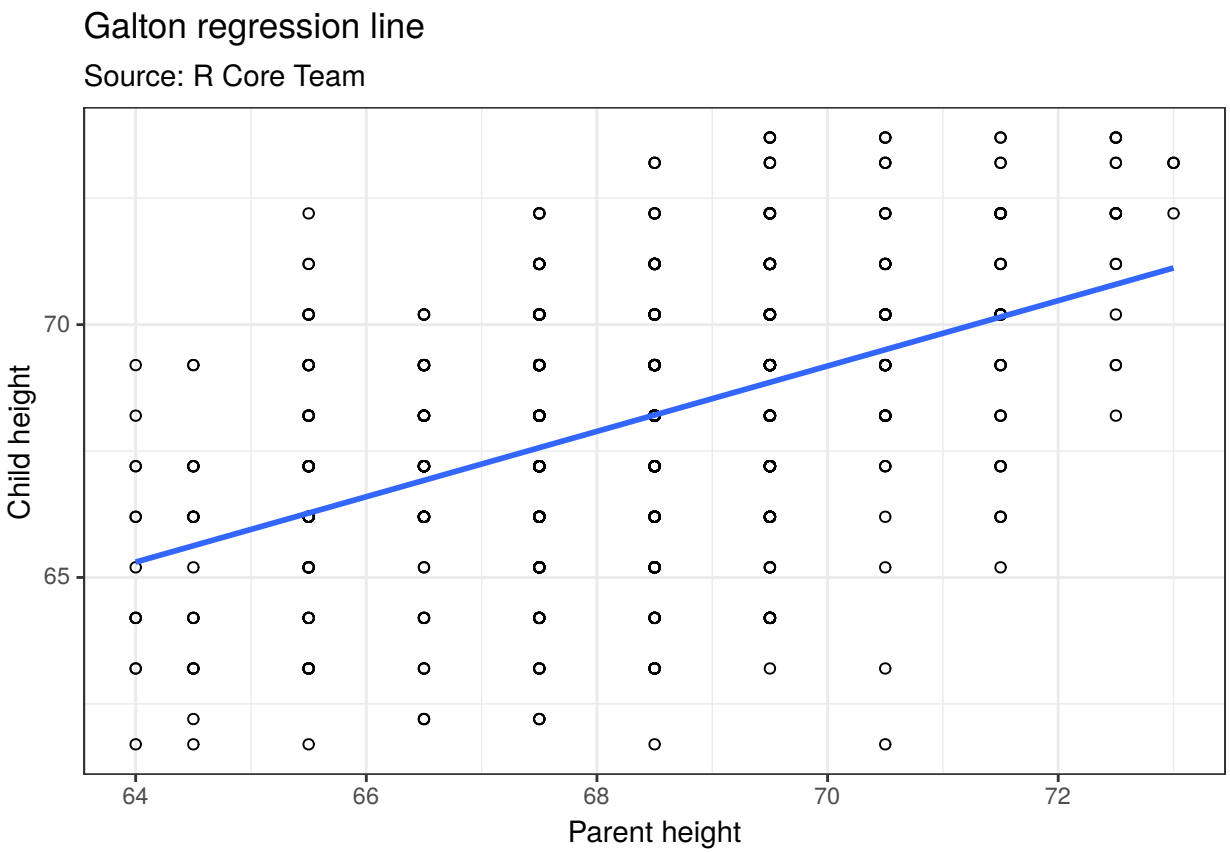
```
p11 <- p11 +
  labs(title = "Galton regression line",
       subtitle = "Source: R Core Team")
p11
```



11.2.4. Using the white theme

As explained in the previous chapters, 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.

```
p11 <- p11 + theme_bw()
p11
```



11.2.5. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

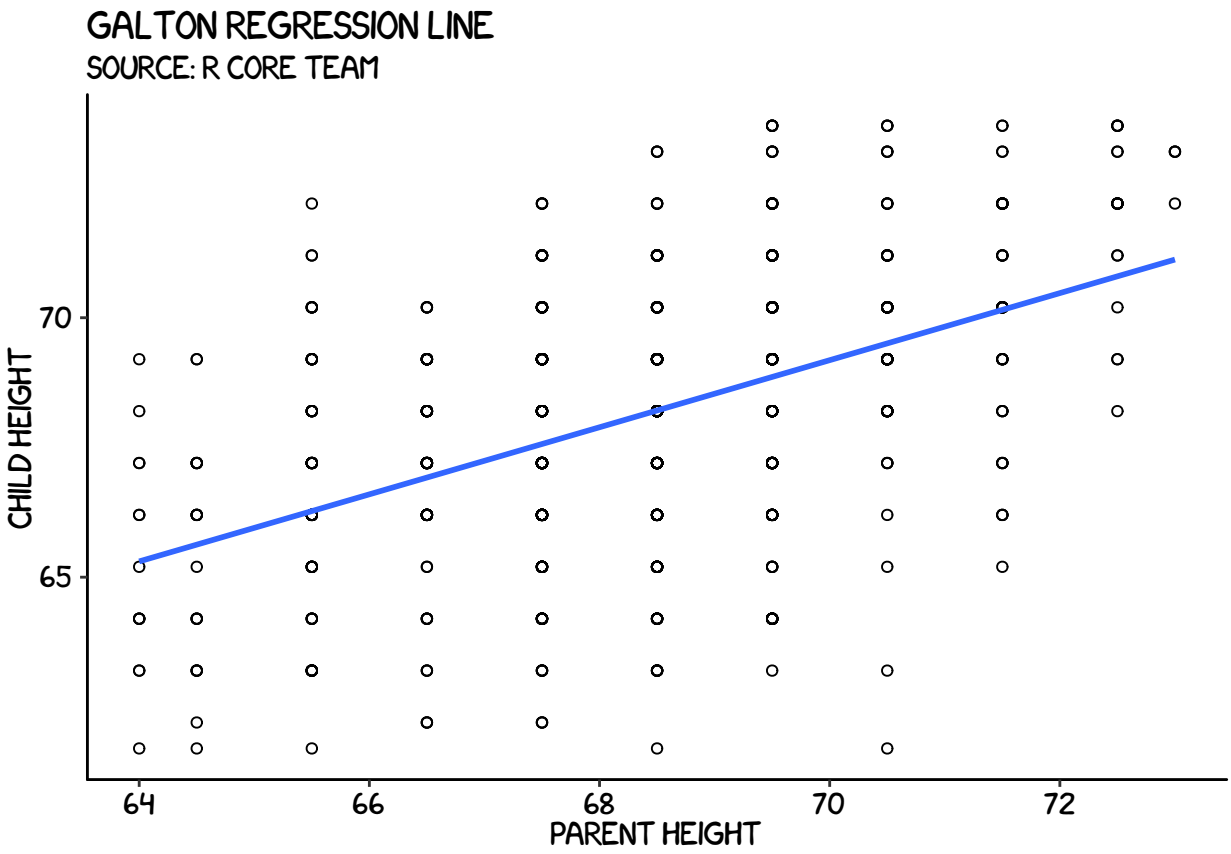
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

11.2.6. 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.

```
p11 <- ggplot(Galton, aes(x = parent, y = child)) +
  geom_point(shape = 1) + geom_smooth(method = lm, se = FALSE) +
  labs(title = "Galton regression line",
       subtitle = "Source: R Core Team") +
  scale_x_continuous(name = "Parent height") +
  scale_y_continuous(name = "Child height") +
  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.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p11
```

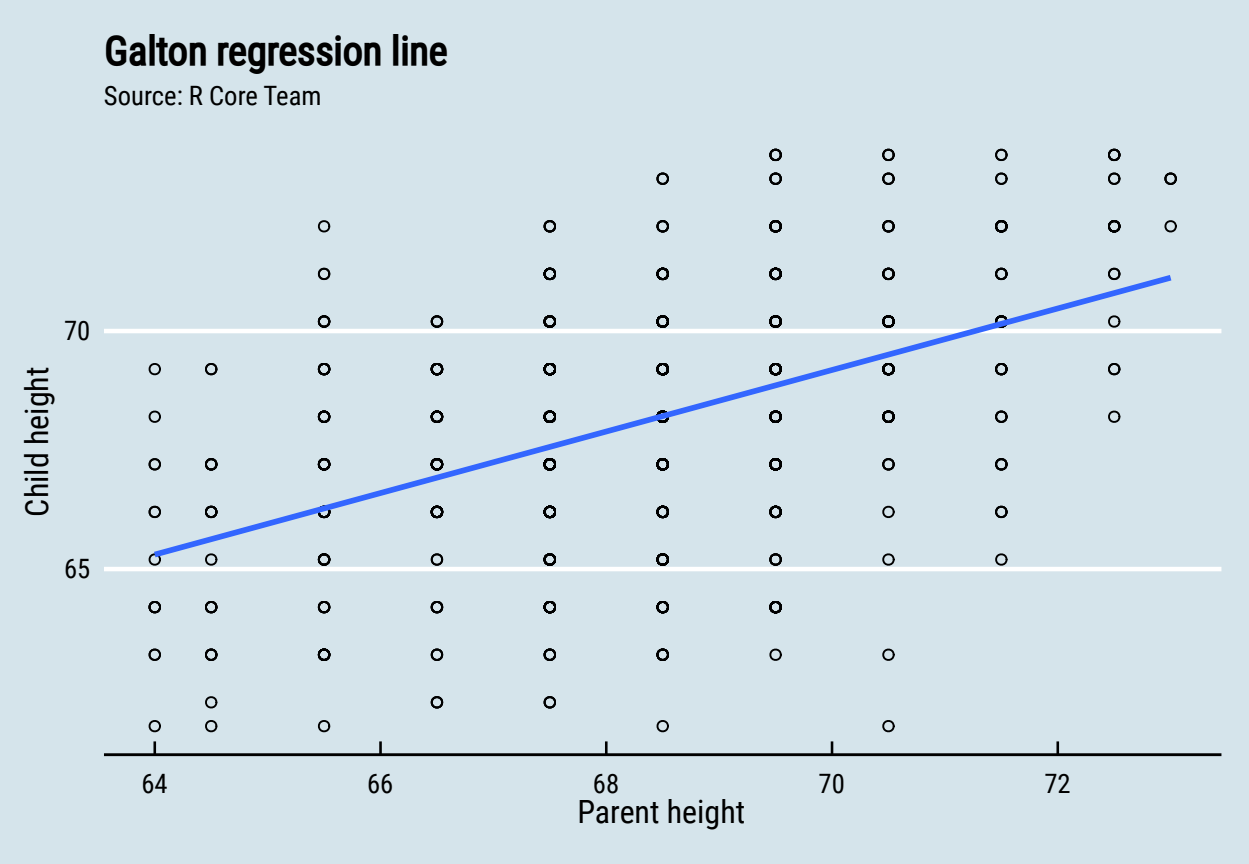


11.2.7. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans'.

```
p11 <- ggplot(Galton, aes(x = parent, y = child)) +
  geom_point(shape = 1) + geom_smooth(method = lm, se = FALSE) +
  labs(title = "Galton regression line",
       subtitle = "Source: R Core Team") +
  scale_x_continuous(name = "Parent height") +
  scale_y_continuous(name = "Child height") +
  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 = "Roboto Condensed"),
    plot.title = element_text(family = "Roboto Condensed"))
```

```
p11
```

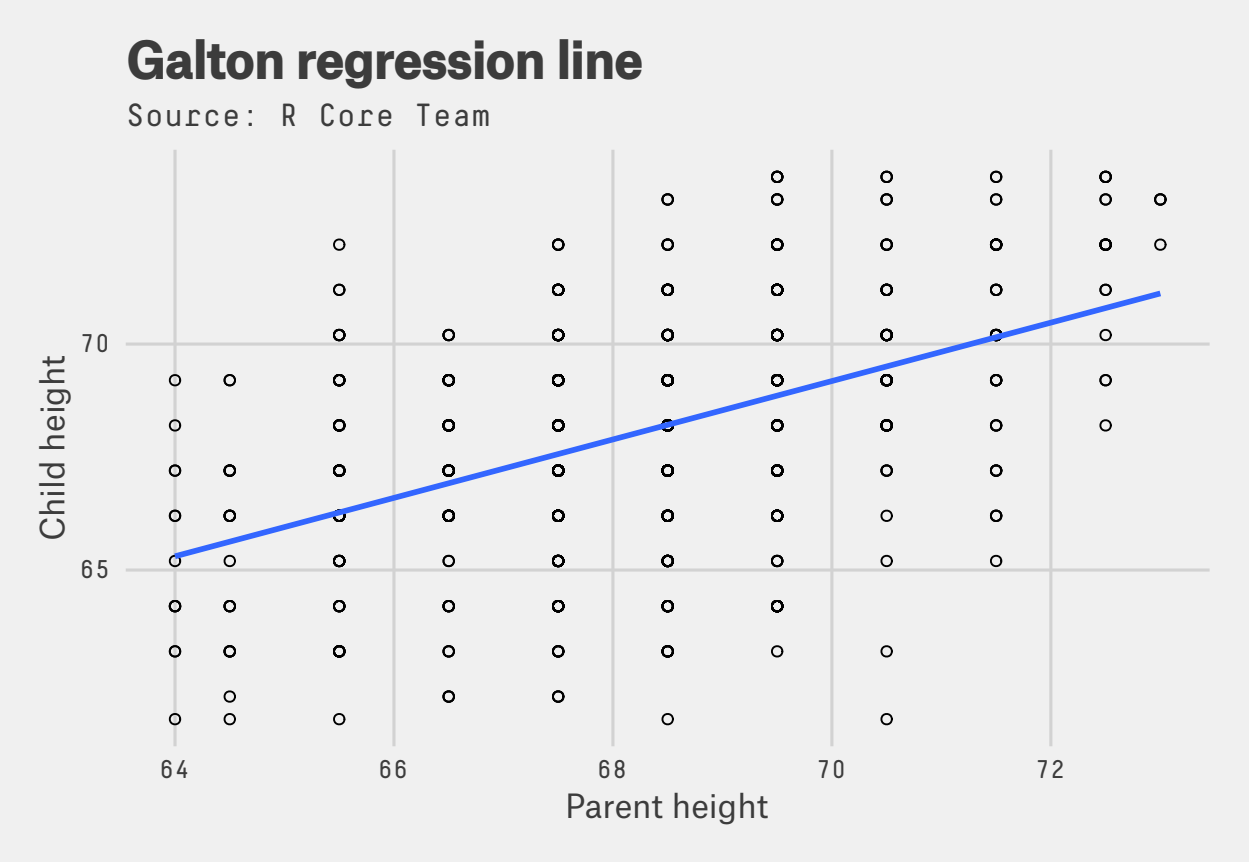


11.2.8. Using 'Five Thirty Eight' theme

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'.

```
p11 <- ggplot(Galton, aes(x = parent, y = child)) +
  geom_point(shape = 1) + geom_smooth(method = lm, se = FALSE) +
  labs(title = "Galton regression line",
       subtitle = "Source: R Core Team") +
  scale_x_continuous(name = "Parent height") +
  scale_y_continuous(name = "Child height") +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 10),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 10),
```

```
    plot.title = element_text(family = "Atlas Grotesk Medium"),
    text = element_text(family = "Decima Mono Pro"))
p11
```

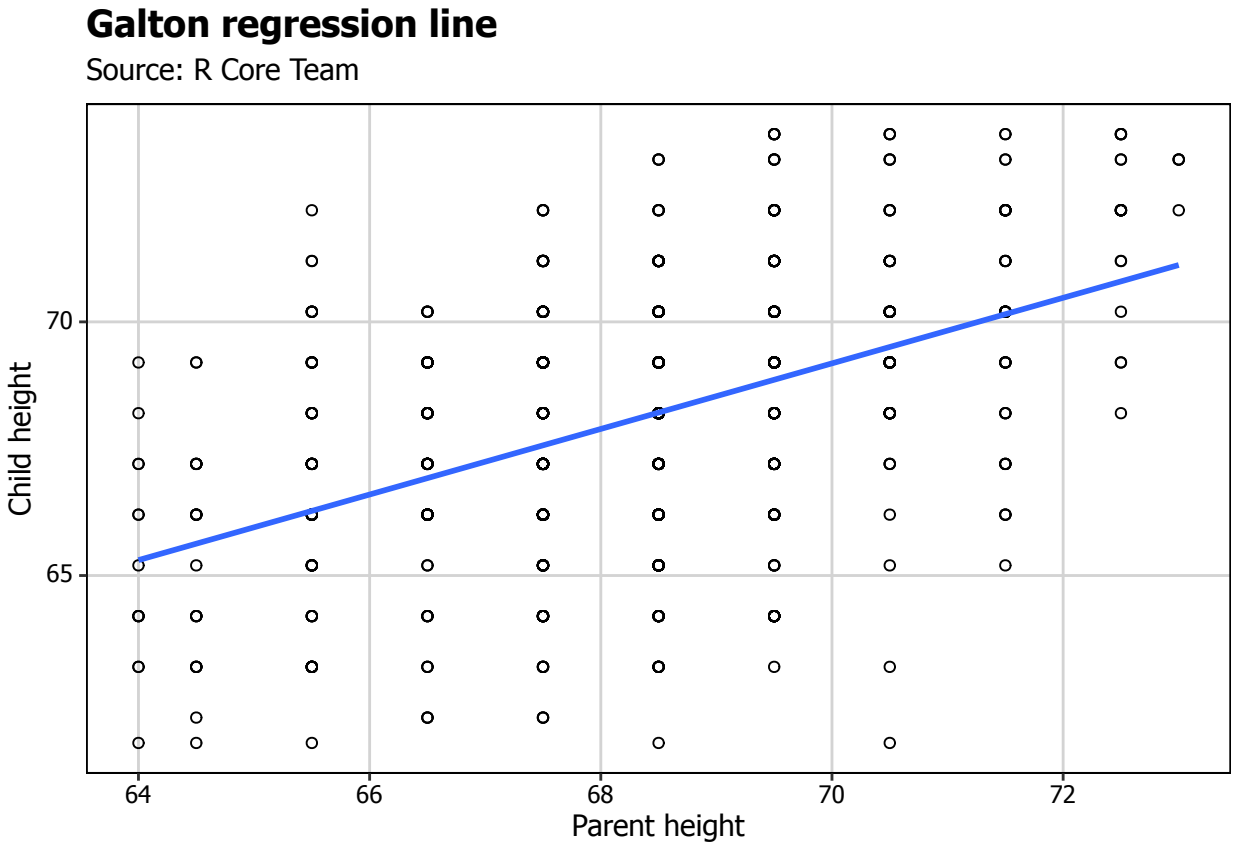


11.2.9. Creating your own theme

As before, you can modify your plots a lot as `ggplot2` allows many customisations. Here is a custom plot where we have modified the axes, background and font.

```
p11 <- ggplot(Galton, aes(x = parent, y = child)) +
  geom_point(shape = 1) + geom_smooth(method = lm, se = FALSE) +
  labs(title = "Galton regression line",
       subtitle = "Source: R Core Team") +
  scale_x_continuous(name = "Parent height") +
  scale_y_continuous(name = "Child height") +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
```

```
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
p11
```



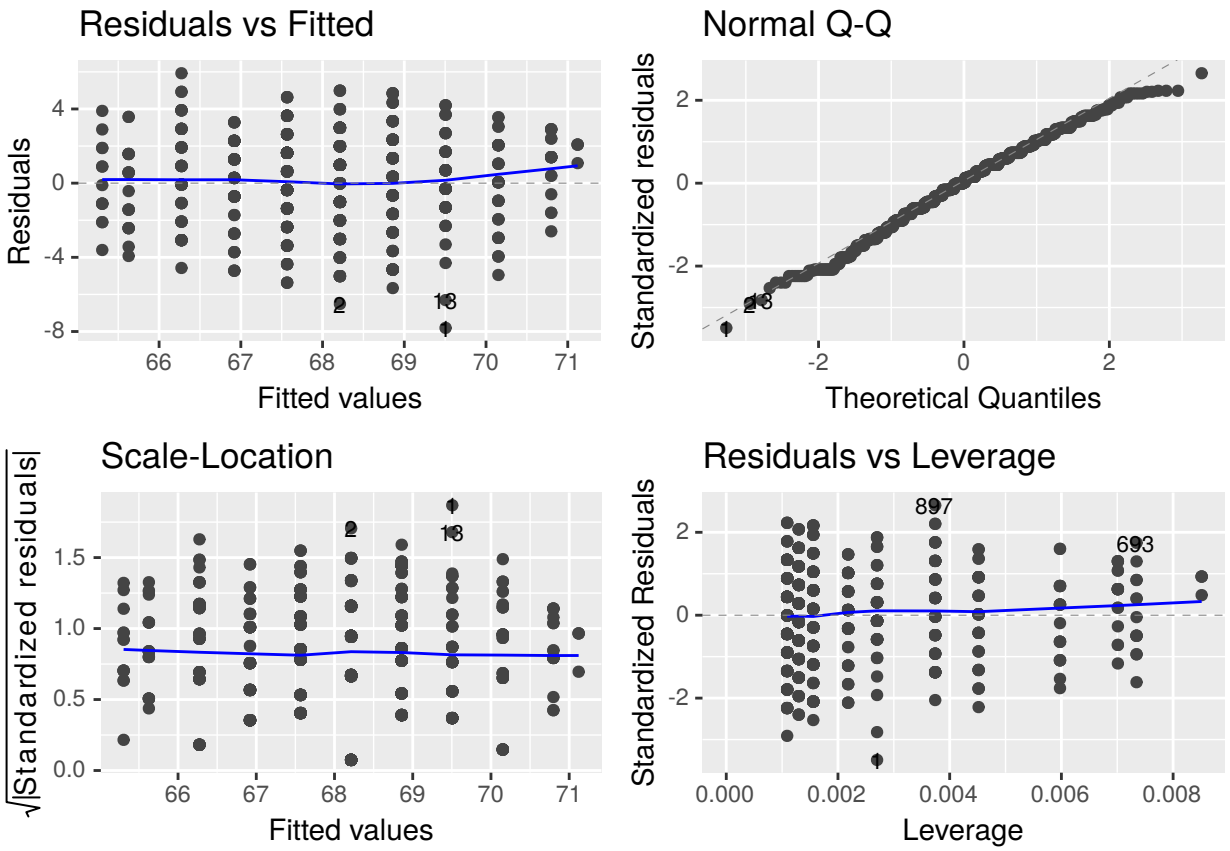
11.3. Regression diagnostics plots

11.3.1. Basic diagnostics plots

An important part of creating regression models is evaluating how well they fit the data. We can use the package `ggfortify` to let `ggplot2` interpret `lm` objects and create diagnostic plots.

To avoid many number on the y-axis, we will round the residuals.

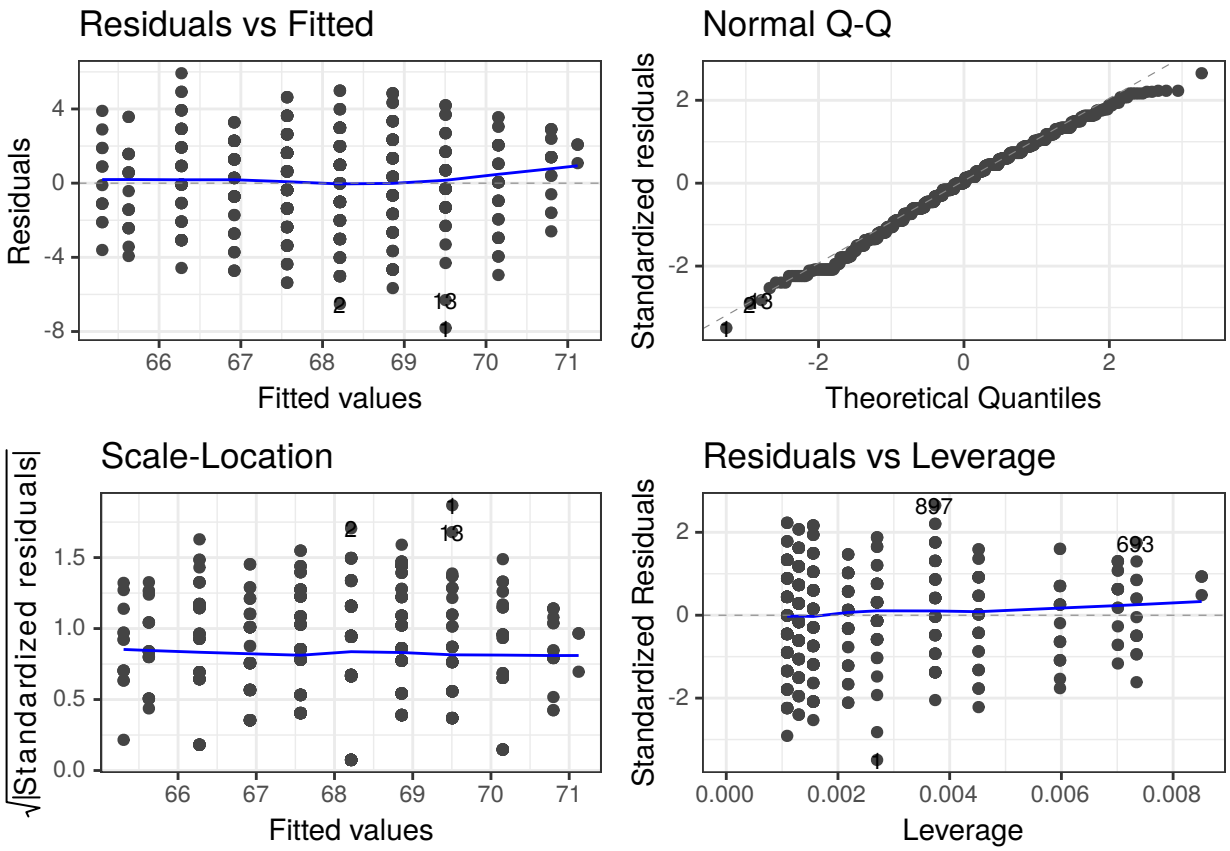
```
autoplot(fit, label.size = 3)
```



11.3.2. Using the white theme

We can also customise the appearance of our diagnostic plots. Let's first use the white theme by again adding `theme_bw()`.

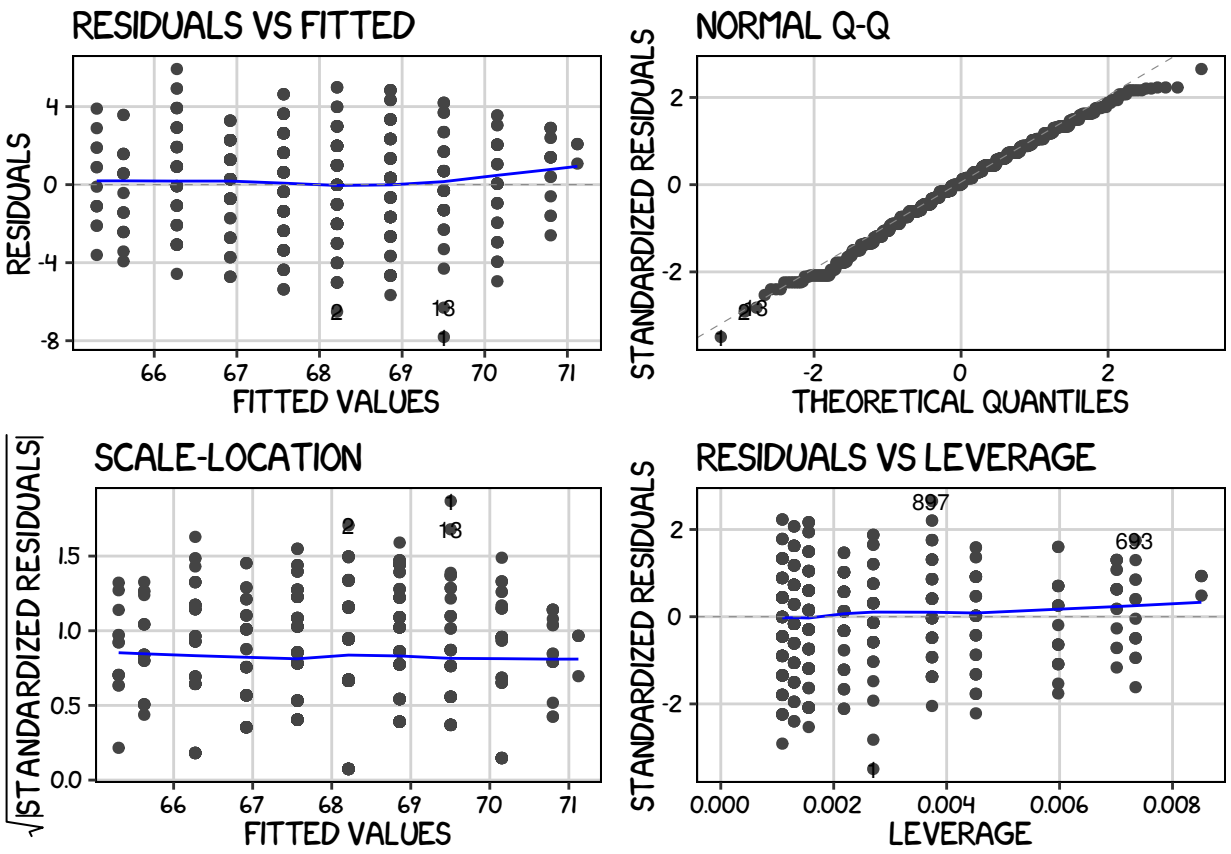
```
autoplot(fit, label.size = 3) + theme_bw()
```



11.3.3. Creating an XKCD style chart

We can of course apply our other themes as well. Let's try the XKCD theme.

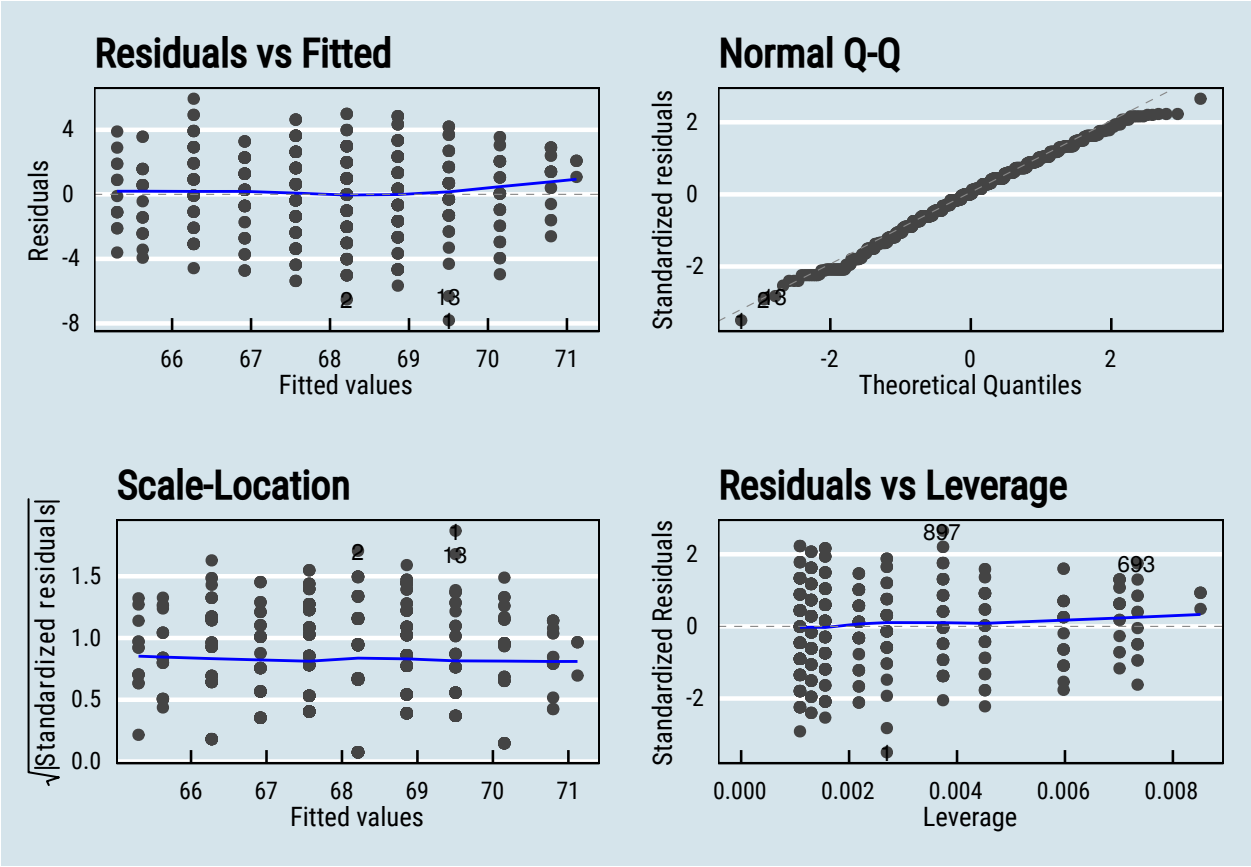
```
autoplot(fit, label.size = 3) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 8),
    axis.text.y = element_text(colour = "black", size = 8),
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
```



11.3.4. Using 'The Economist' theme

And now the Economist theme.

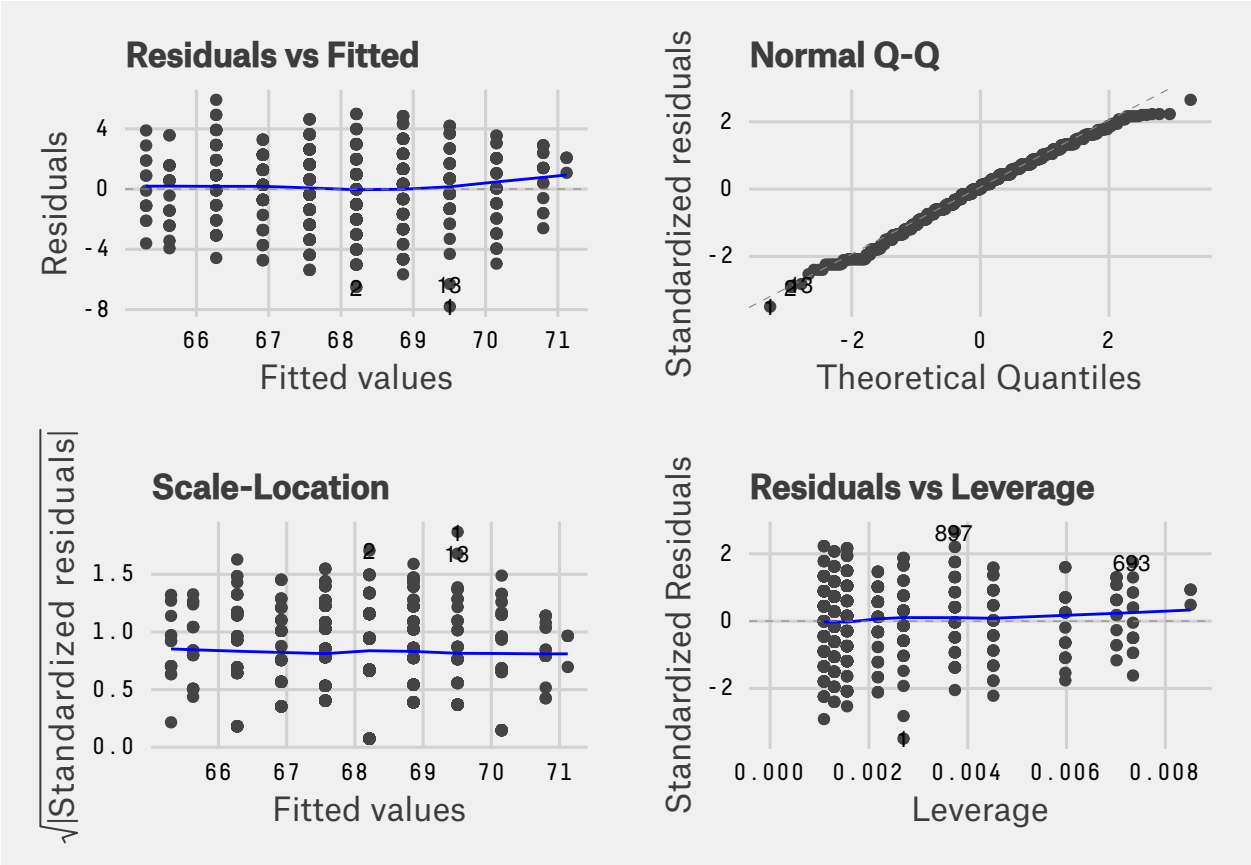
```
autoplot(fit, label.size = 3) + theme_economist() +
  theme(panel.border = element_rect(colour = "black",
                                    fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    panel.background = element_blank(),
    plot.title = element_text(family = "Roboto Condensed"),
    text = element_text(family = "Roboto Condensed"))
```



11.3.5. Using 'Five Thirty Eight' theme

And now Five Thirty Eight theme.

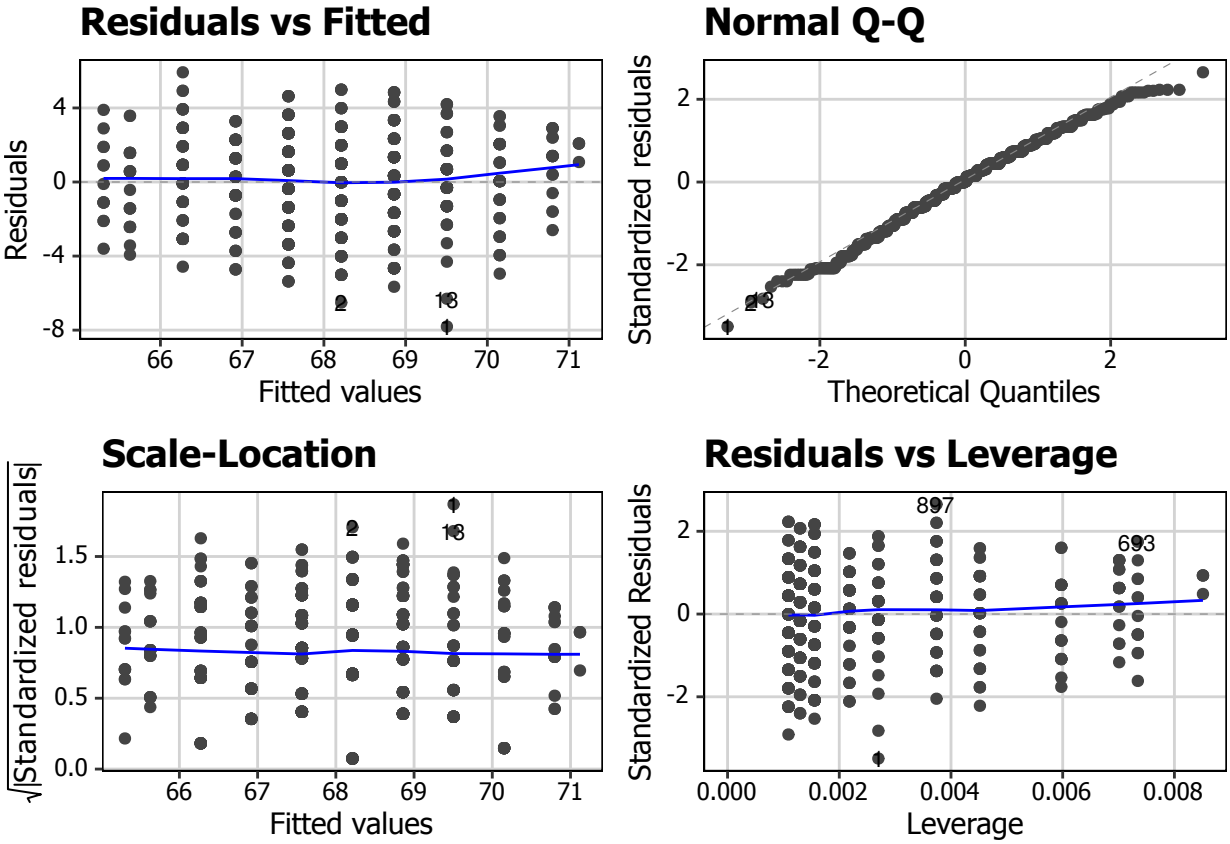
```
autoplot(fit, label.size = 3) + theme_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular"),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    plot.title = element_text(family = "Atlas Grotesk Medium", size = 12),
    text = element_text(family = "Decima Mono Pro"))
```



11.3.6. Creating your own theme

Finally, we can also fully customise the diagnostic plots to match our regression plot simply by applying all of the same theme options.

```
autoplot(fit, label.size = 3) +
  theme(panel.border = element_rect(colour = "black", fill = NA, size = .5),
    axis.text.x = element_text(colour = "black", size = 9),
    axis.text.y = element_text(colour = "black", size = 9),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    panel.grid.major = element_line(colour = "#d3d3d3"),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
    text = element_text(family = "Tahoma"))
```

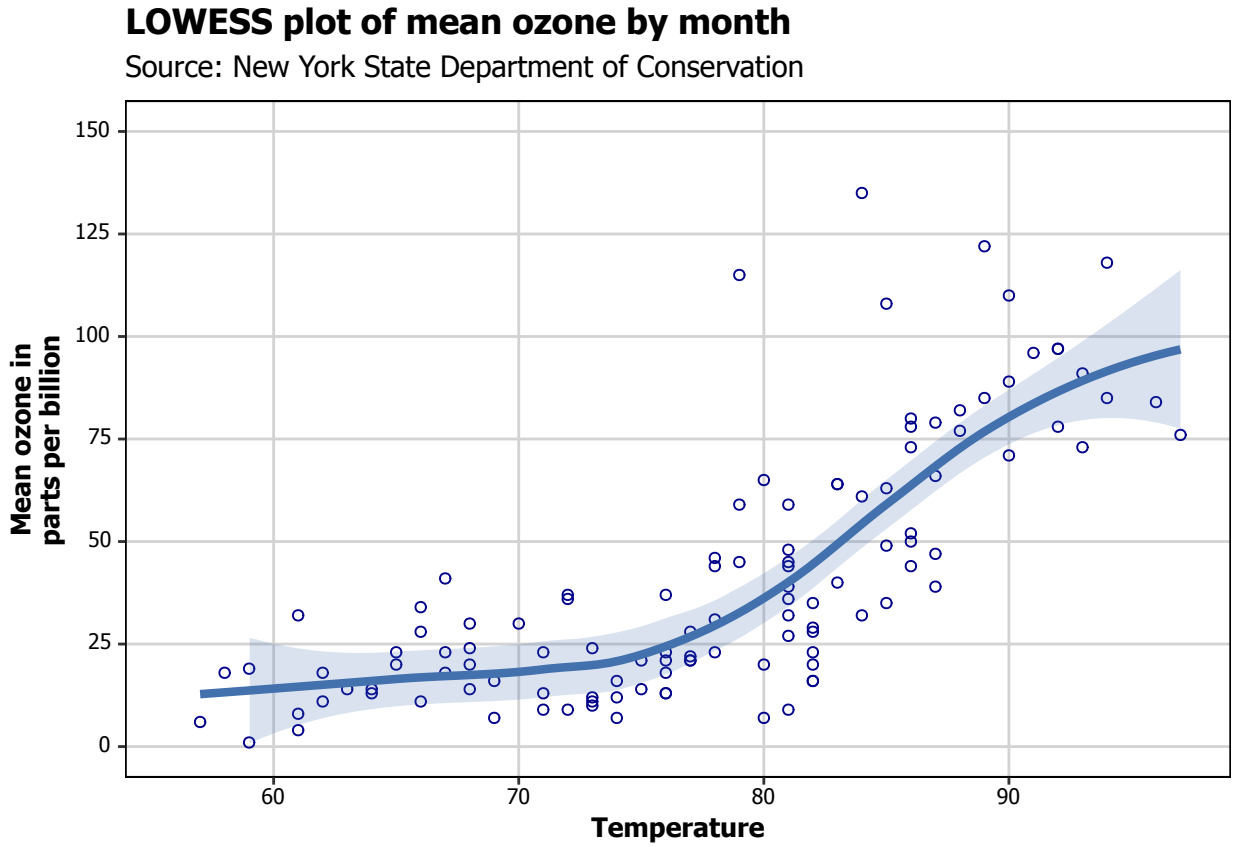

Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized Residuals
Leverage

CHAPTER 12

LOWESS plots

12.1. Introduction

This is the twelfth and final chapter. In this chapter, 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

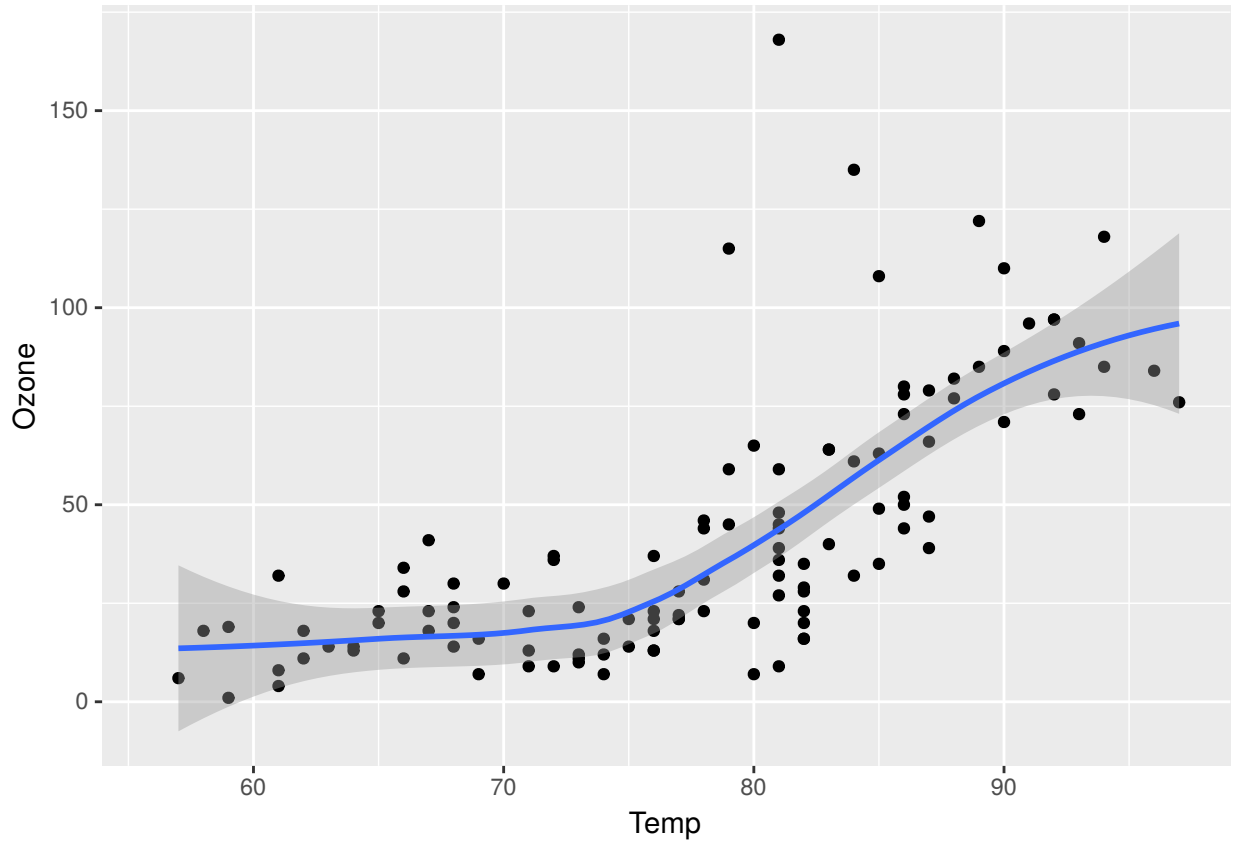
in order to use it as our grouping variable.

```
if (!require("pacman")) install.packages("pacman")
p_load(datasets, ggplot2, ggthemes, grid, RColorBrewer)

data(airquality)
```

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

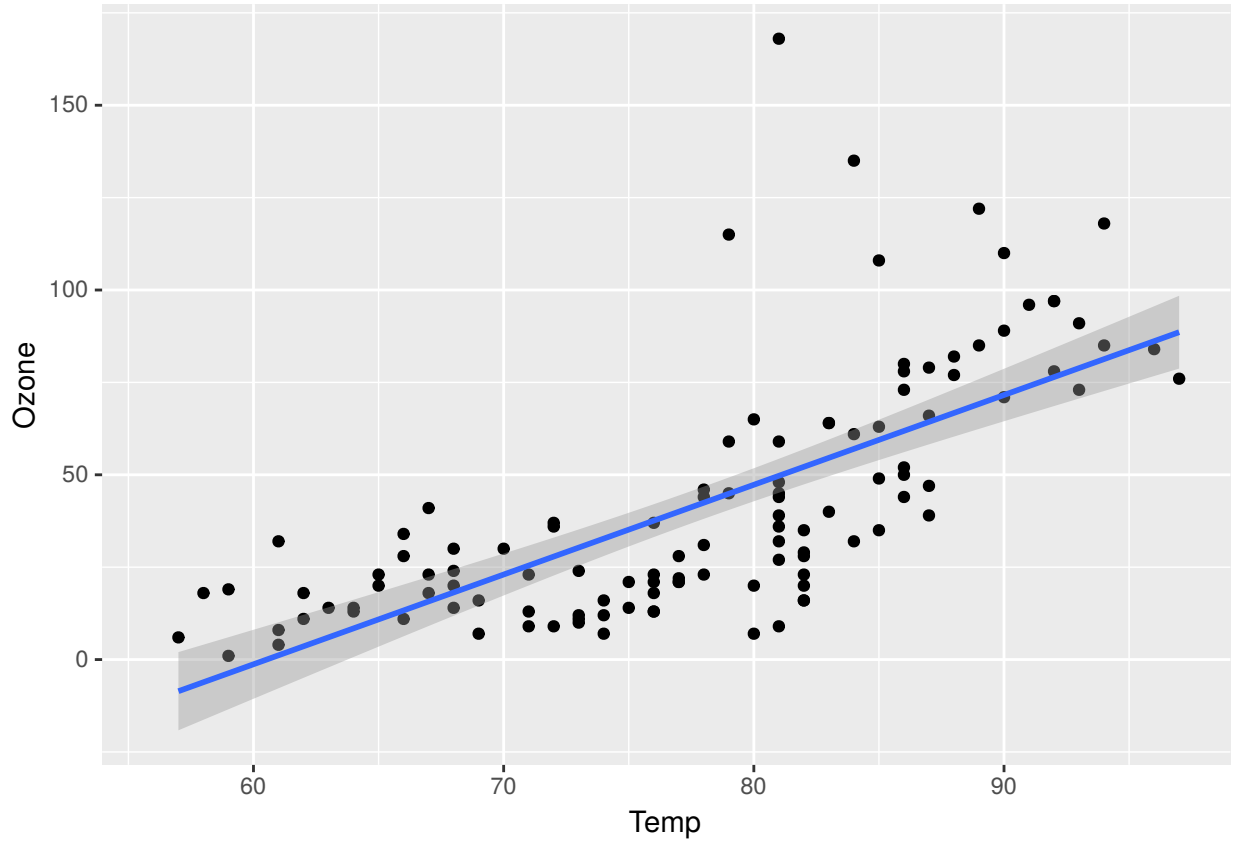
In order to initialise a plot we tell ggplot that `airquality` is our data, and specify that our x-axis plots the `Temp` variable and our y-axis plots the `Ozone` variable. We then instruct ggplot to render this as a LOWESS curve by adding the `stat_smooth(method = "loess")` option. Note that the default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



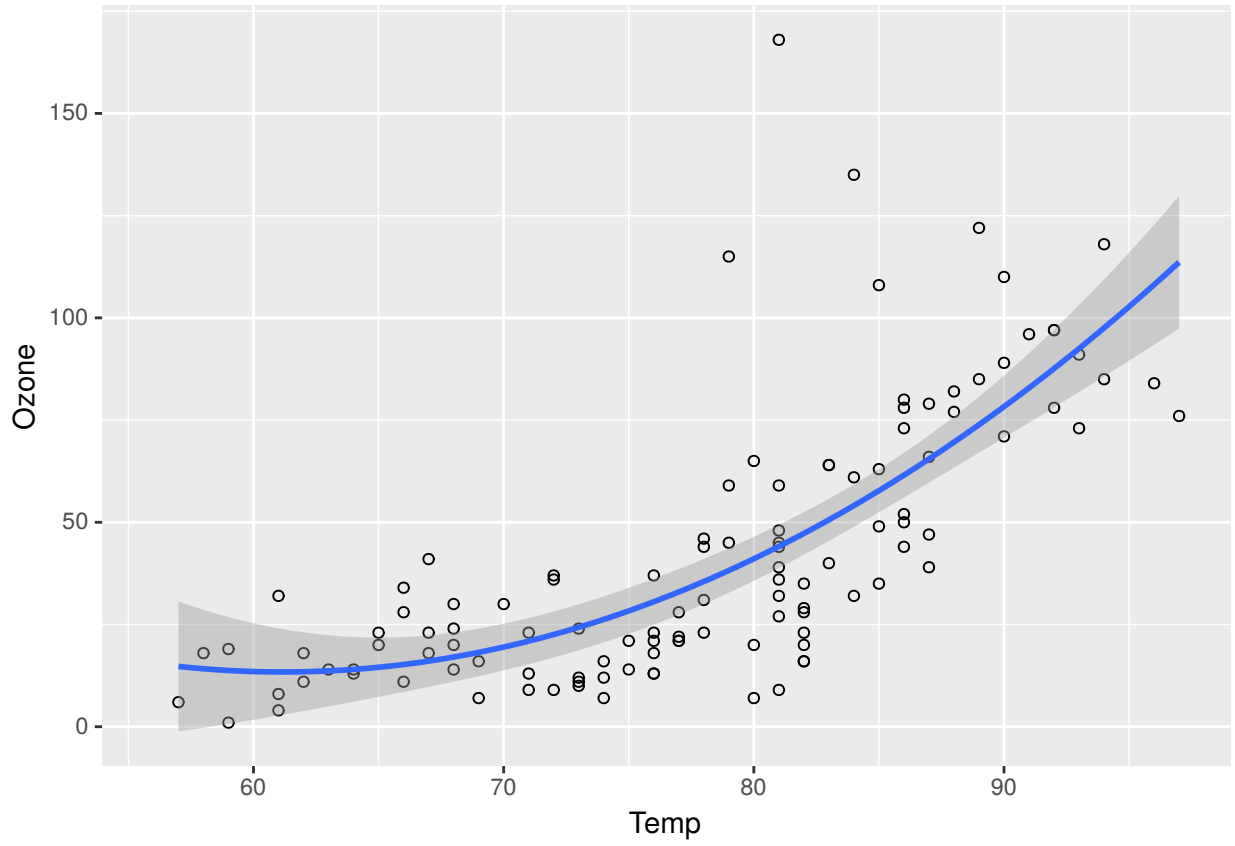
Let's now have a look at the amount of variance it explains in ozone levels by extracting the adjusted R^2 from the linear regression model between these two variables.

```
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
```

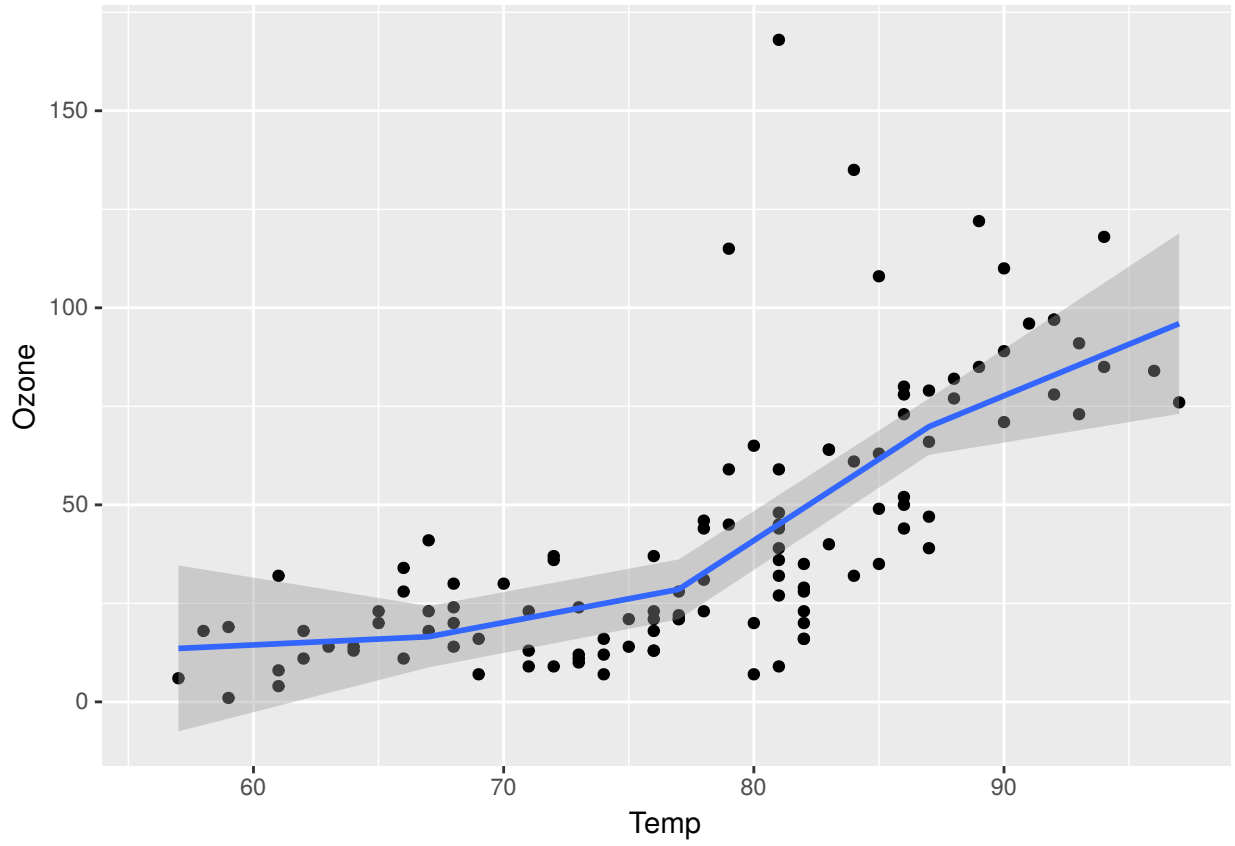
```
[1] 0.5361501
```

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.

12.3. 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
```

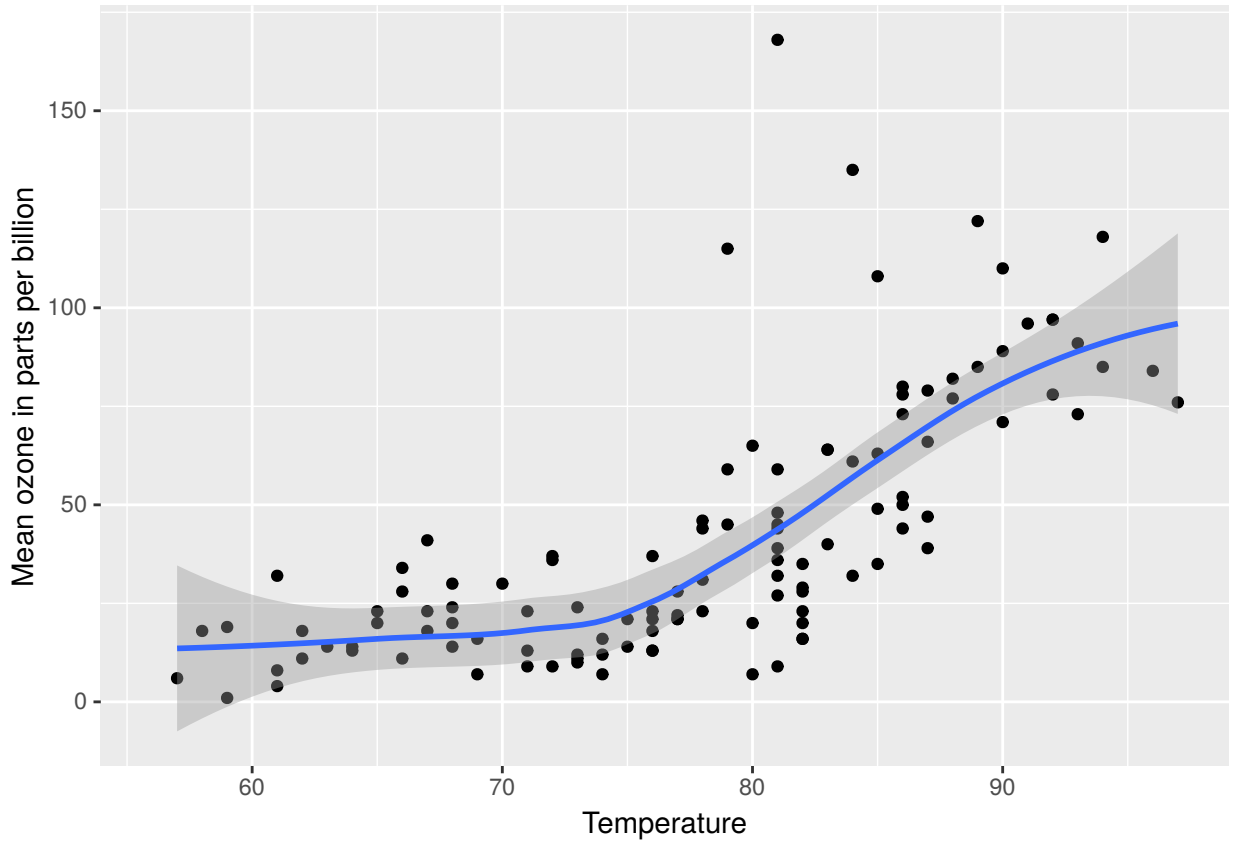


12.4. Customising axis labels

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

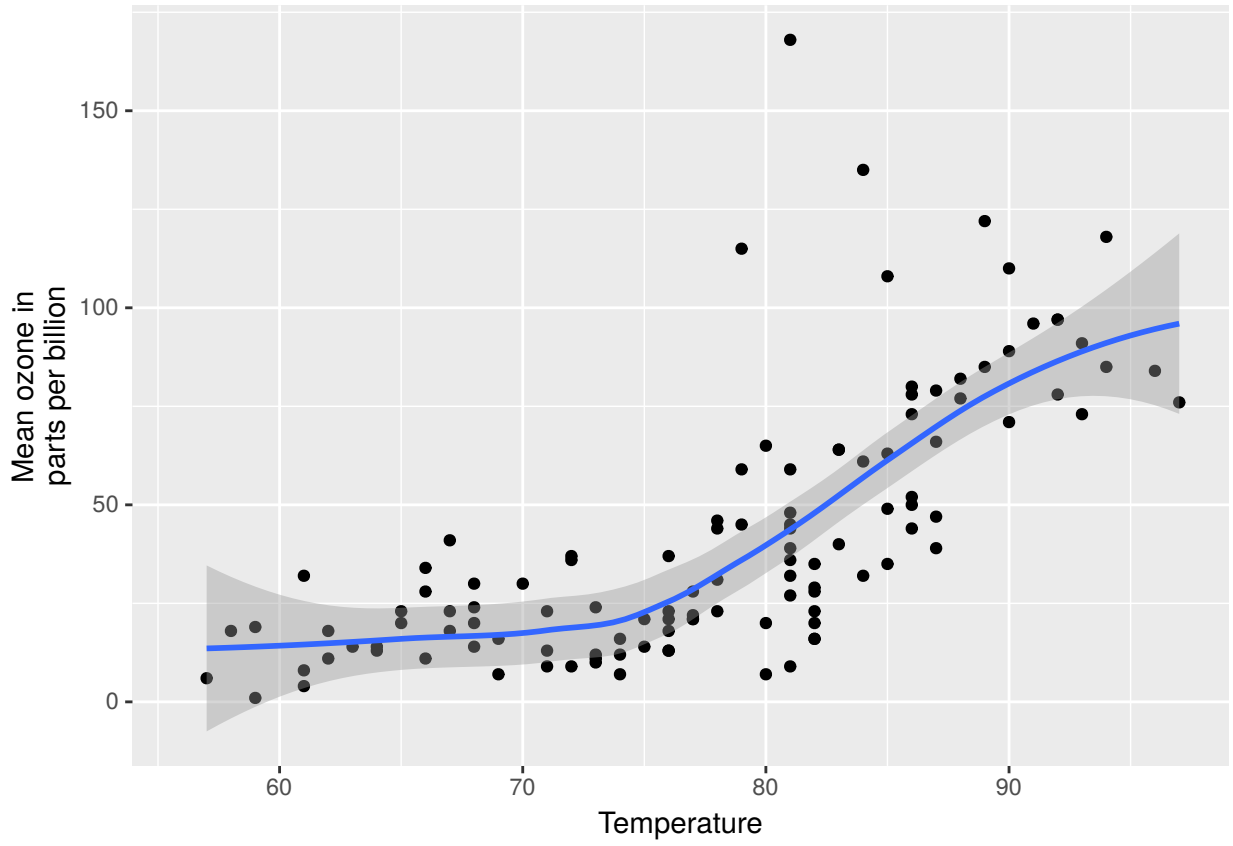
In order to change the axis labels, we have a couple of options. In this case, we have used the `scale_x_continuous` and `scale_y_continuous` options, as these have further customisation options for the axes we will use below. In each, we add the desired name to the `name` argument as a string.

```
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.

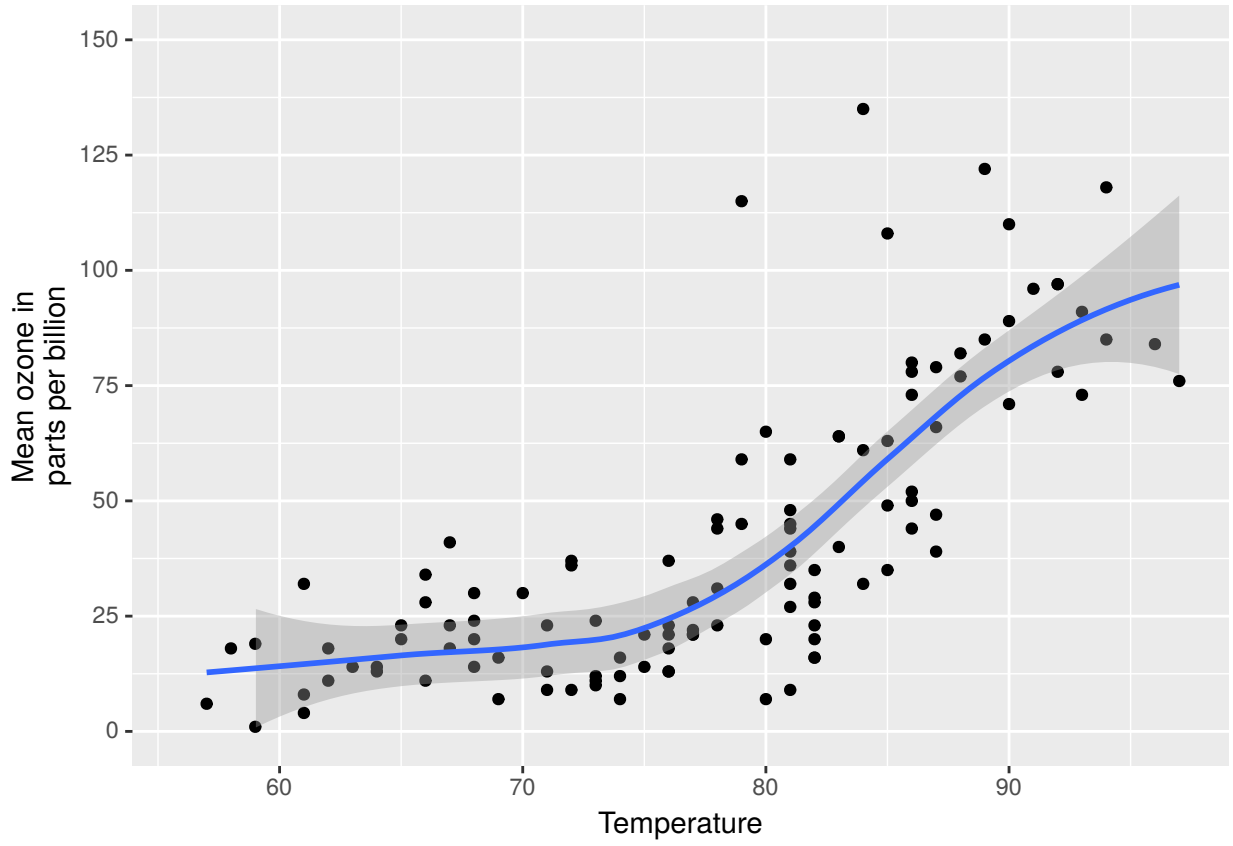
```
p12 <- p12 + scale_y_continuous(name = "Mean ozone in\nparts per billion")
p12
```



12.5. 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`.

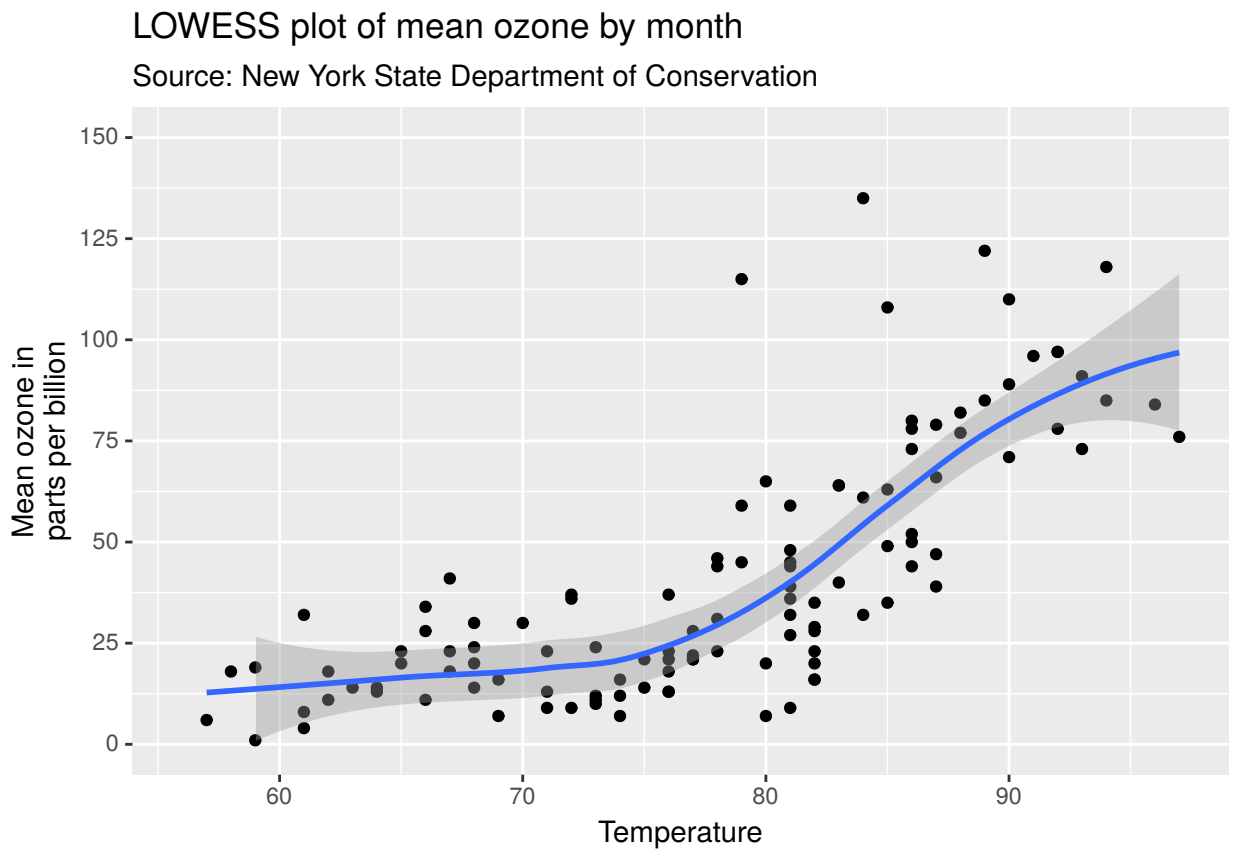
```
p12 <- p12 + scale_y_continuous(name = "Mean ozone in\nparts per billion",
                                breaks = seq(0, 150, 25), limits = c(0, 150))
p12
```



12.6. Adding a title

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

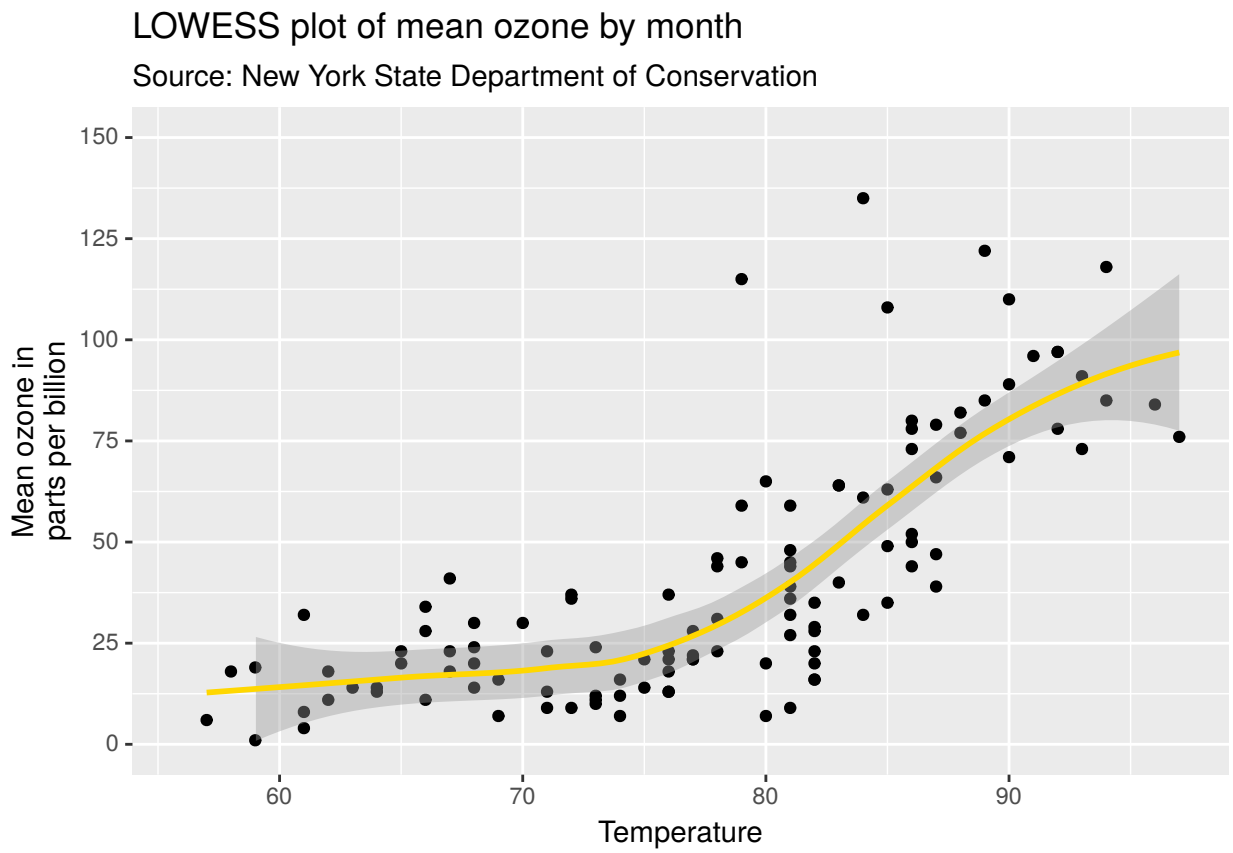
```
p12 <- p12 +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



12.7. Changing the colour and size of the LOWESS curve

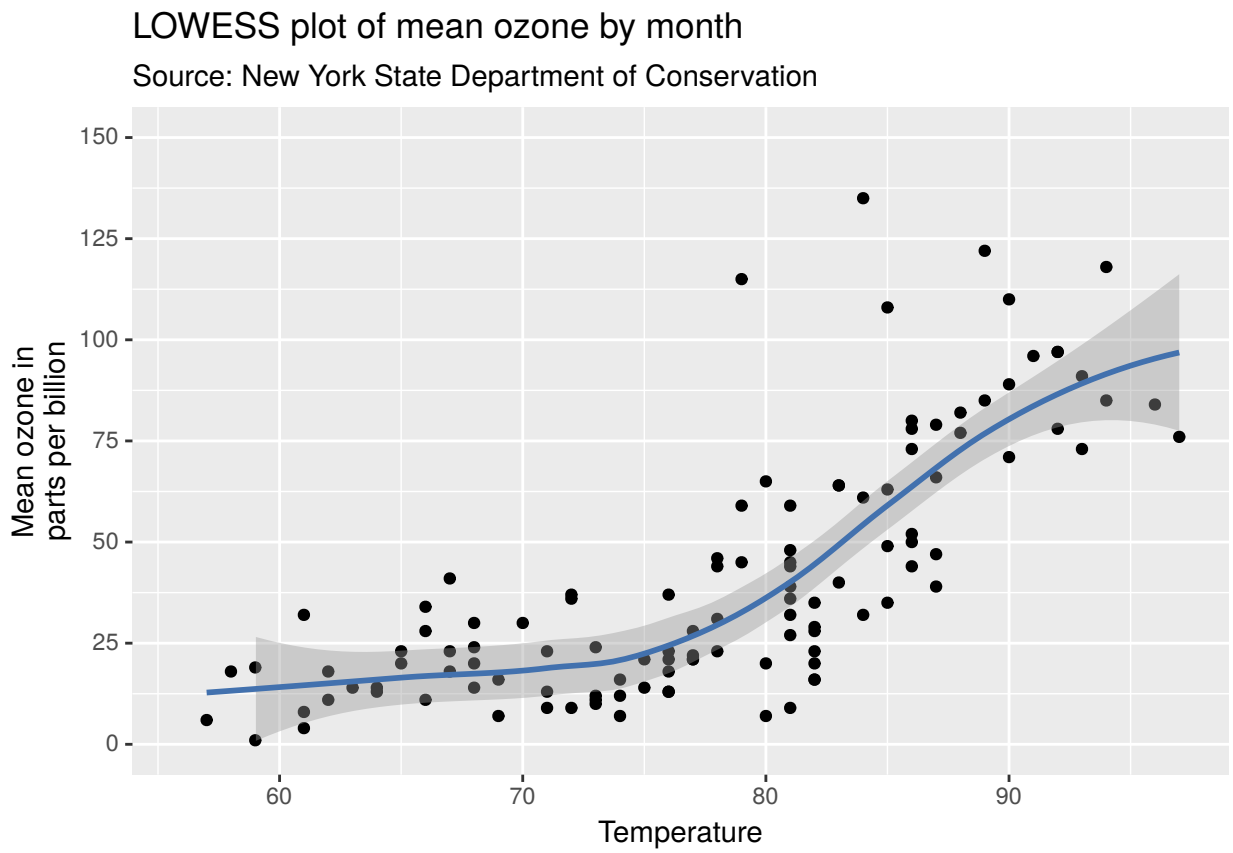
To change the colour of the LOWESS curve, we add a valid colour to the `colour` argument in `geom_smooth()` (note that we assigned this colour to a variable outside of the plot to make it easier to change it). A list of valid colours is here.

```
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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



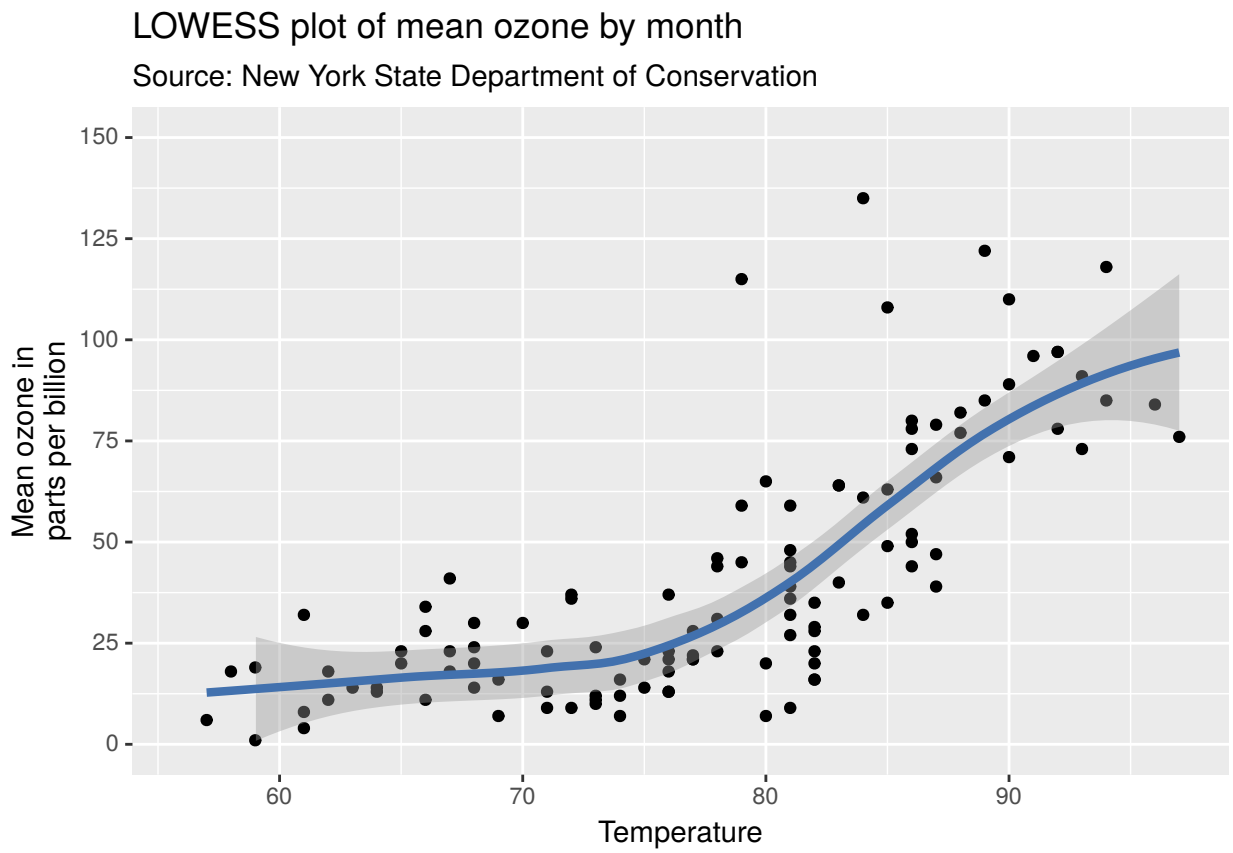
If you want to go beyond the options in the list above, you can also specify exact HEX colours by including them as a string preceded by a hash, e.g., "#FFFFFF". Below, we have called a shade of blue for the line using its HEX code.

```
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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



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

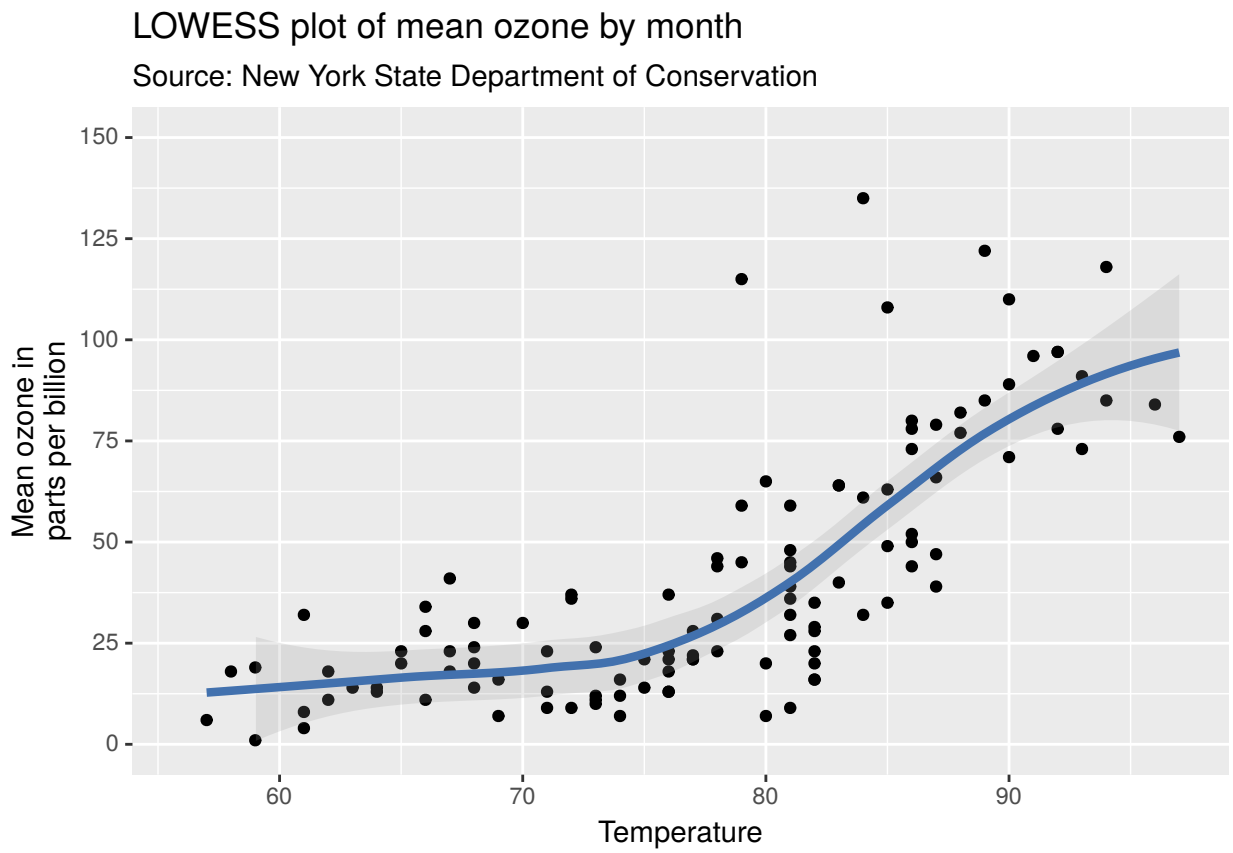
```
fill <- "#4271AE"

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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



12.8. Changing the appearance of the confidence interval

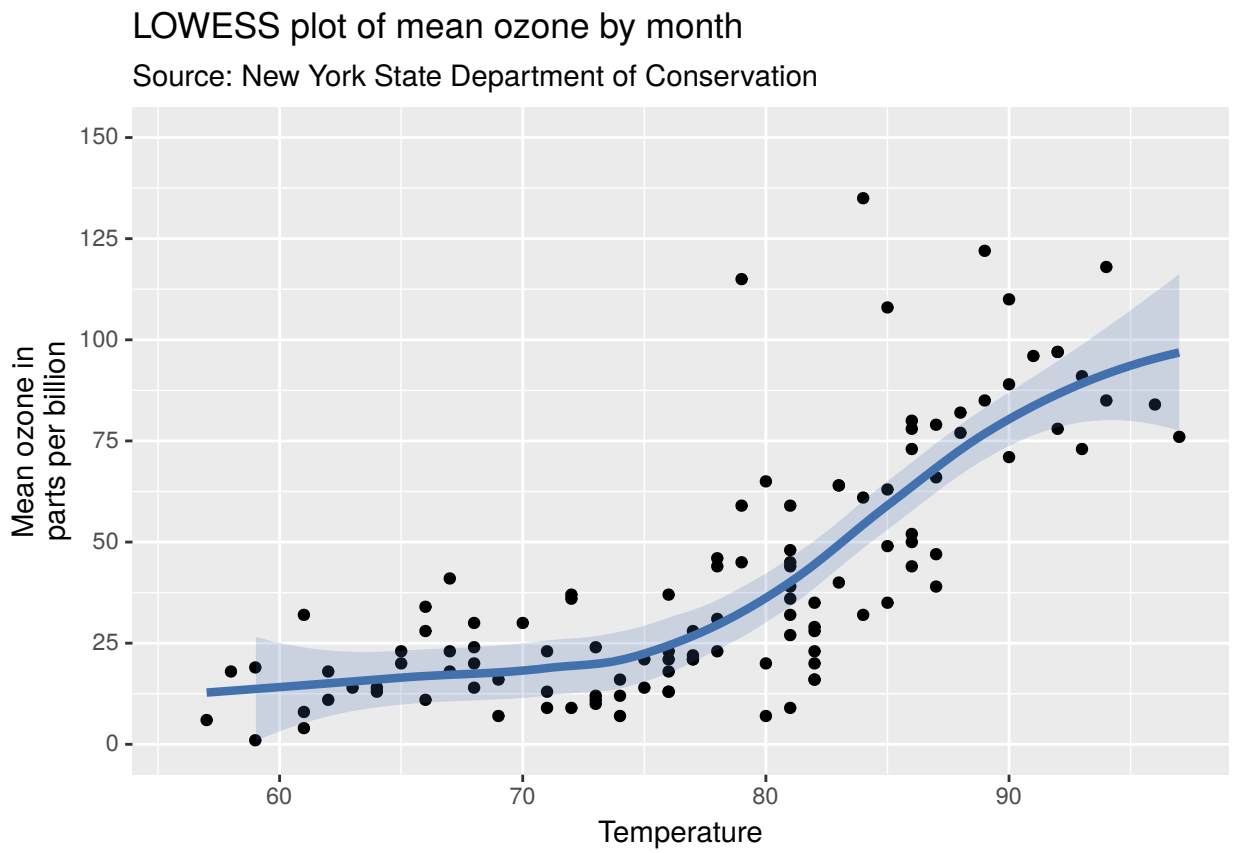
We can also alter how the confidence interval around the LOWESS curve looks. We can change the transparency using the argument `alpha` in `geom_smooth()`. This ranges from 0 to 1. Here we will increase the transparency of the confidence interval.

```
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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



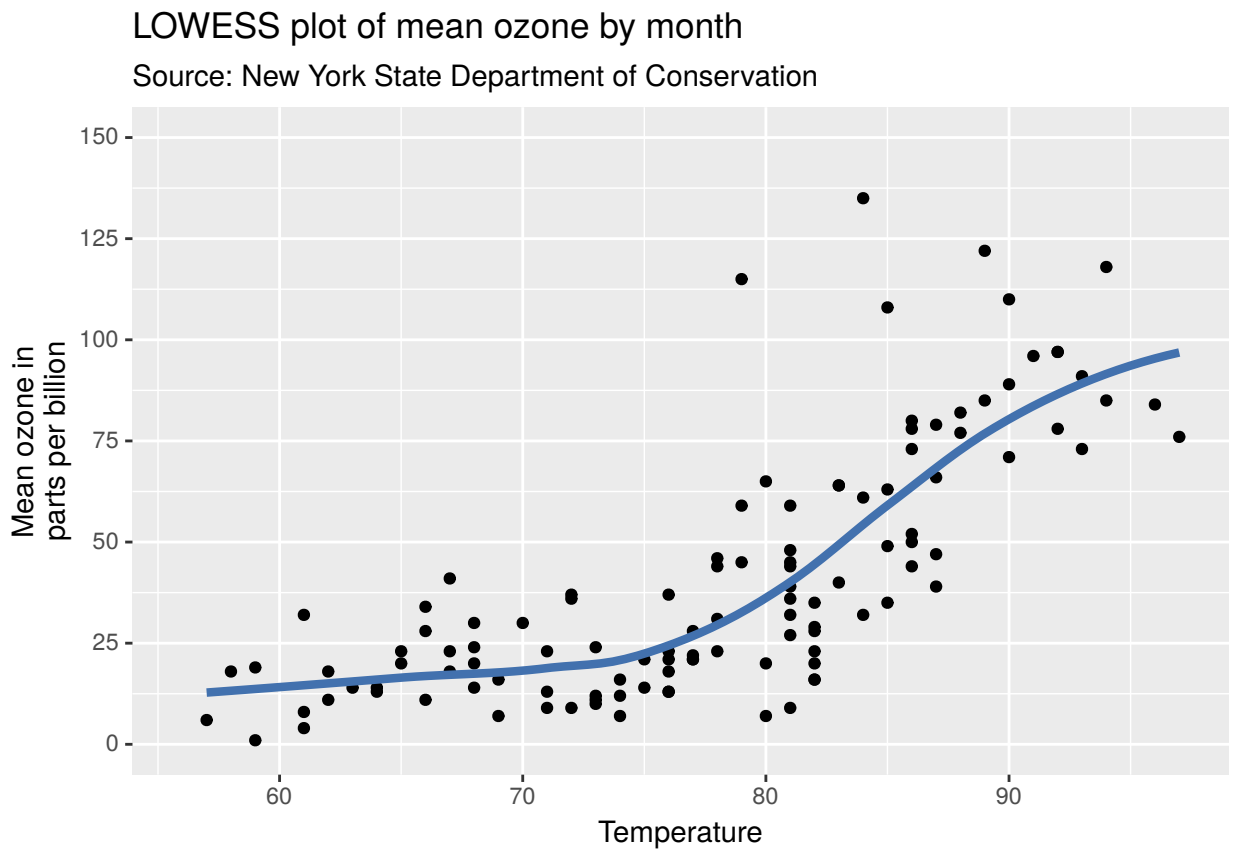
We can also change the colour of the confidence interval from the default grey using the argument `fill`, also within `geom_smooth()`. Let's change it to the same blue as our LOWESS curve.

```
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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



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

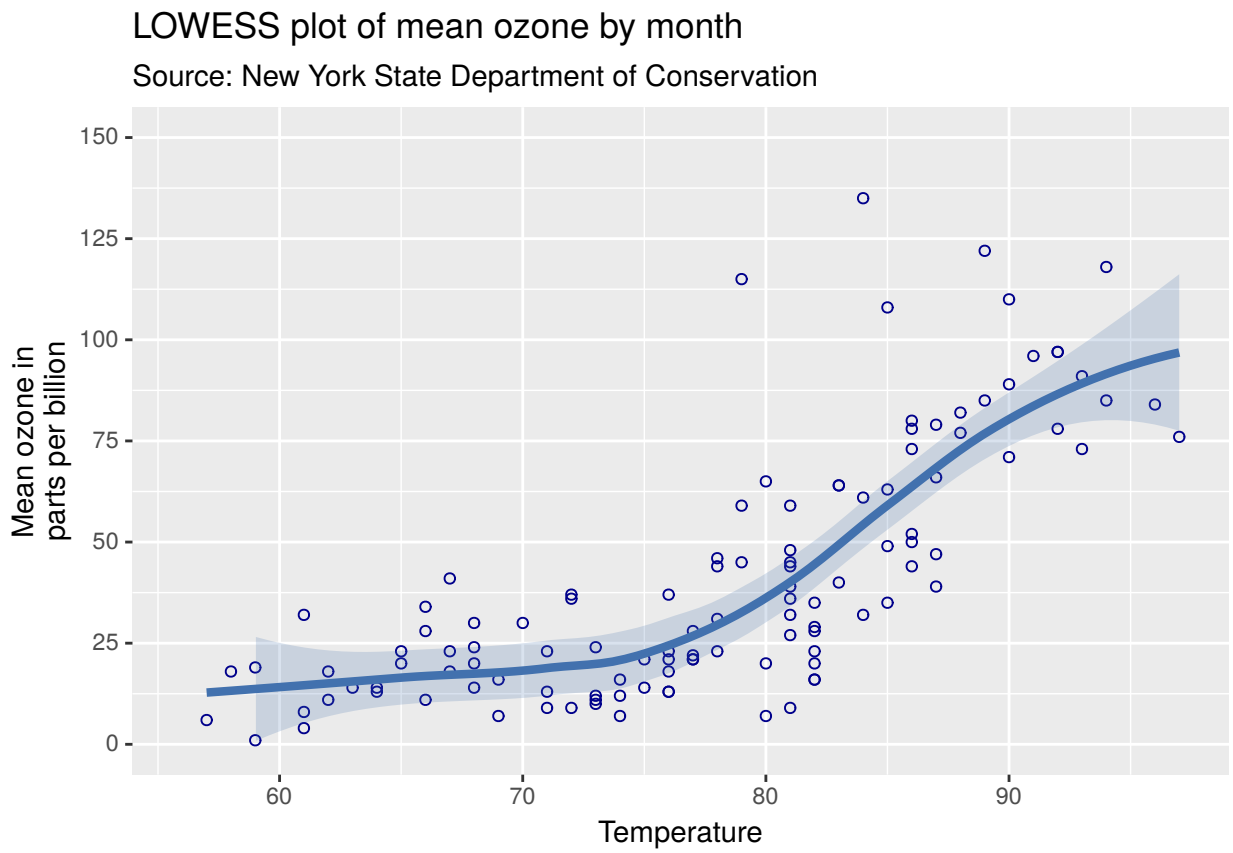
```
fill <- "#4271AE"

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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



12.9. Changing the appearance of the scatterplot

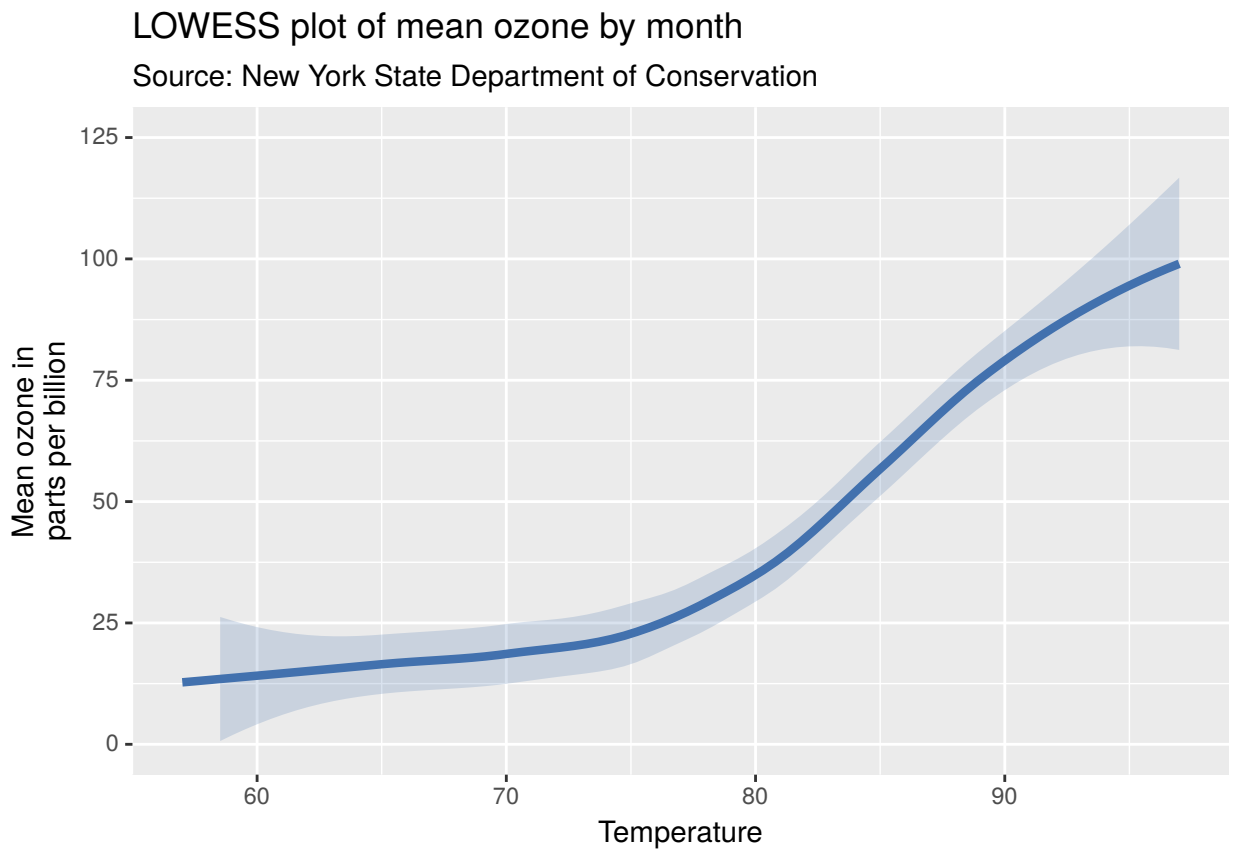
Of course, the LOWESS curve is not the only part of this plot. We can also customise the appearance of the scatterplot underlying the curve. Let's change the circles to shape 21, which is a circle that allows different colours for the outline and fill, and change the colour of the outline to dark blue. 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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
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.

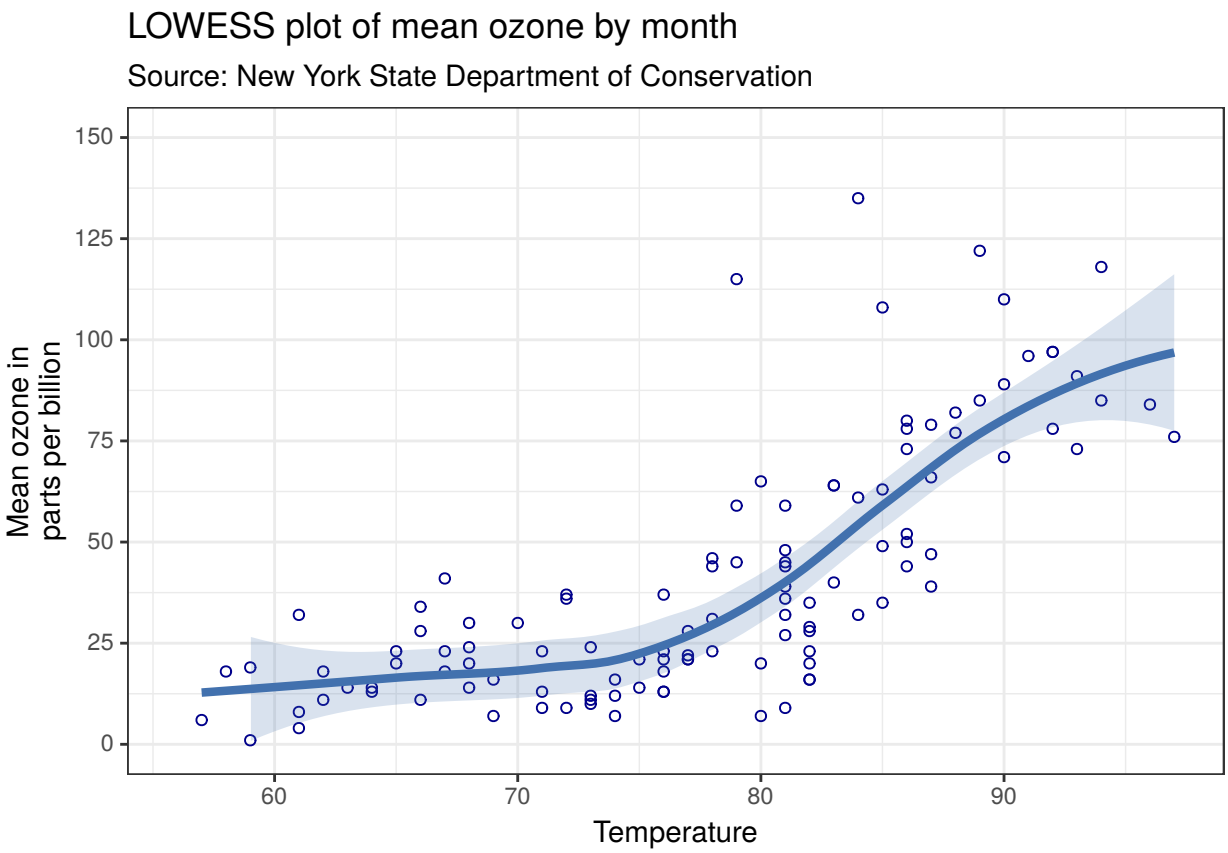
```
fill <- "#4271AE"

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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation")
p12
```



12.10. 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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation") +
  theme_bw()
p12
```



12.11. Using system fonts

You can use fonts such as Tahoma within `ggplot2`. One option is the classic `extrafont` package but here we will use `showtext` package that allows a wider variety of options including Google Fonts. The reason behind this is that its easier to make it work under Windows, Ubuntu and OS X.

We'll use XKCD, 'Roboto Condensed', 'Atlas Grotesk' and 'Decima Mono Pro' fonts. 'Officina Sans' is used in the plots that appear in The Economist while 'Atlas Grotesk' and 'Decima Mono Pro' are used by FiveThirtyEight.

The XKCD font that is available under an open license here or here.

'Officina Sans' is a commercial font and is available here. However, as a replacement we use 'Roboto Condensed' provided freely by Google Fonts.

'Atlas Grotesk' and 'Decima Mono Pro' are commercial fonts and are available here and here.

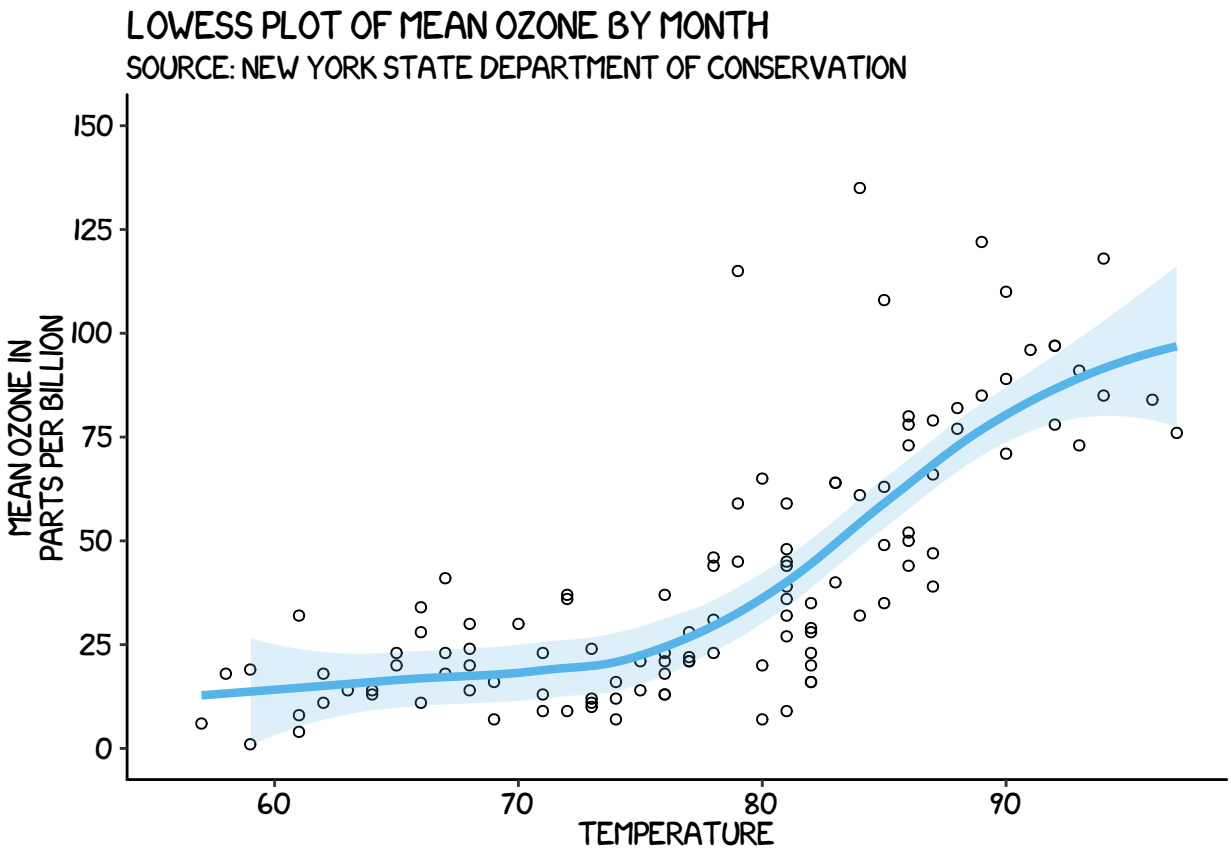
```
p_load(showtext)

font_add("Tahoma","Tahoma.ttf")
font_add("XKCD","xkcd-Regular.otf")
font_add("Roboto Condensed", "RobotoCondensed-Regular.ttf")
font_add("Decima Mono Pro","DecimaMonoPro.otf")
```

```
font_add("Atlas Grotesk Regular","AtlasGrotesk-Regular.otf")
font_add("Atlas Grotesk Medium","AtlasGrotesk-Medium.otf")
showtext_auto()
```

12.12. 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 that is available under an open license here, 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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation") +
  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.background = element_blank(),
    plot.title = element_text(family = "XKCD"),
    text = element_text(family = "XKCD"))
p12
```

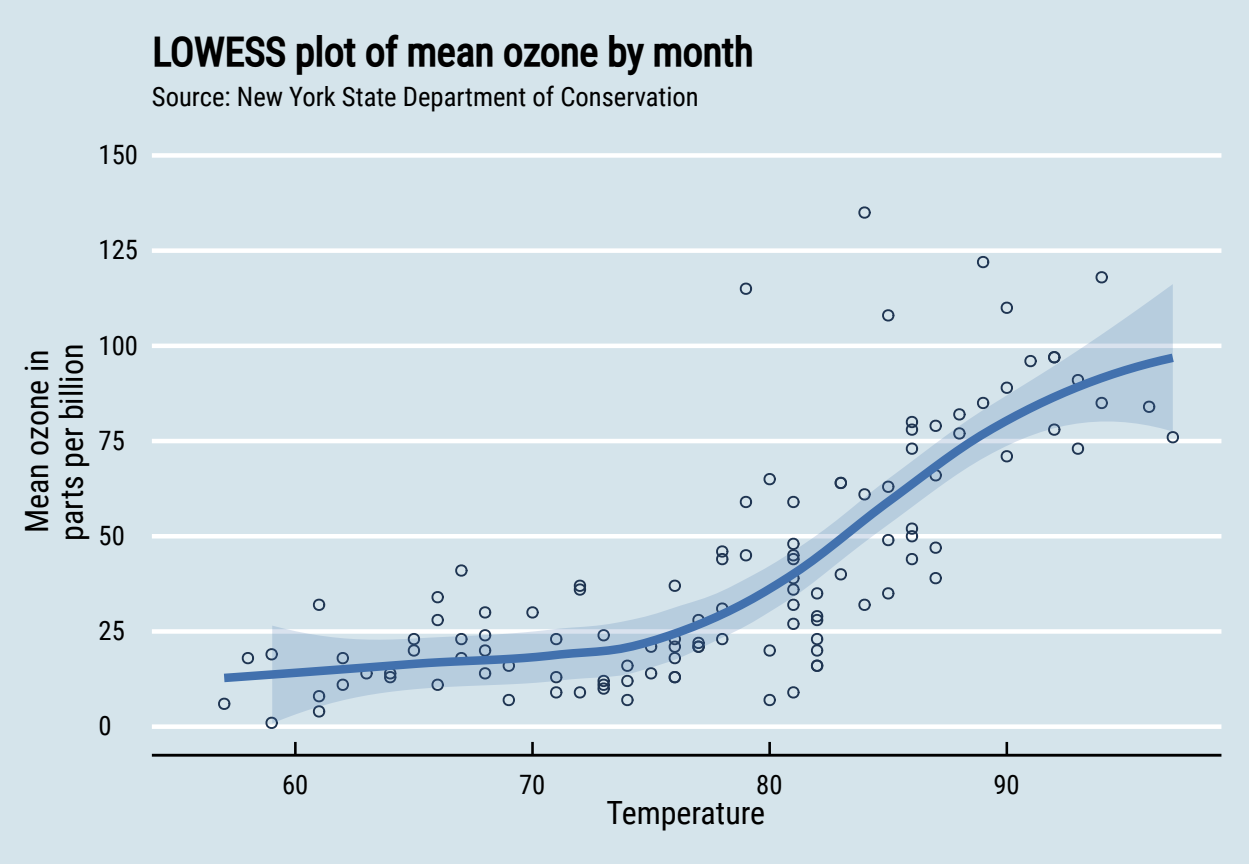


12.13. Using 'The Economist' theme

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. It is also important that the font change argument inside `theme` is optional and it's only to obtain a more similar result compared to the original. For an exact result you need 'Officina Sans' which is a commercial font and is available here.

```
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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation") +
  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 = "Roboto Condensed"),
    plot.title = element_text(family = "Roboto Condensed"))
p12
```

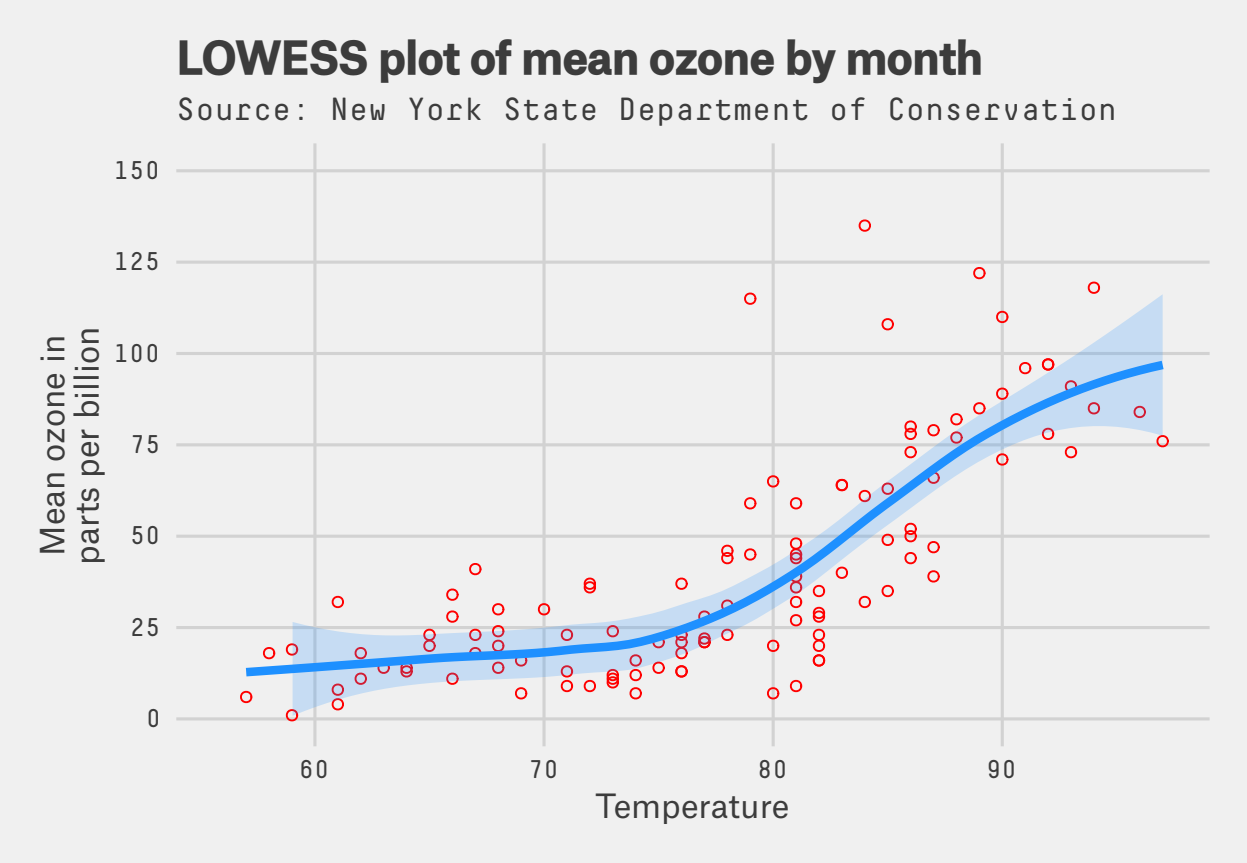


12.14. Using 'Five Thirty Eight' theme

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 fonts 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)) +
  labs(title = "LOWESS plot of mean ozone by month",
```

```
       subtitle = "Source: New York State Department of Conservation") +
  theme_fivethirtyeight() + scale_fill_fivethirtyeight() +
  theme(axis.title = element_text(family = "Atlas Grotesk Regular", size = 12),
    legend.position = "bottom",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.title = element_text(family = "Atlas Grotesk Regular", size = 8),
    legend.text = element_text(family = "Atlas Grotesk Regular", size = 8),
    plot.title = element_text(family = "Atlas Grotesk Medium", size = 16),
    text = element_text(family = "Decima Mono Pro"))
p12
```

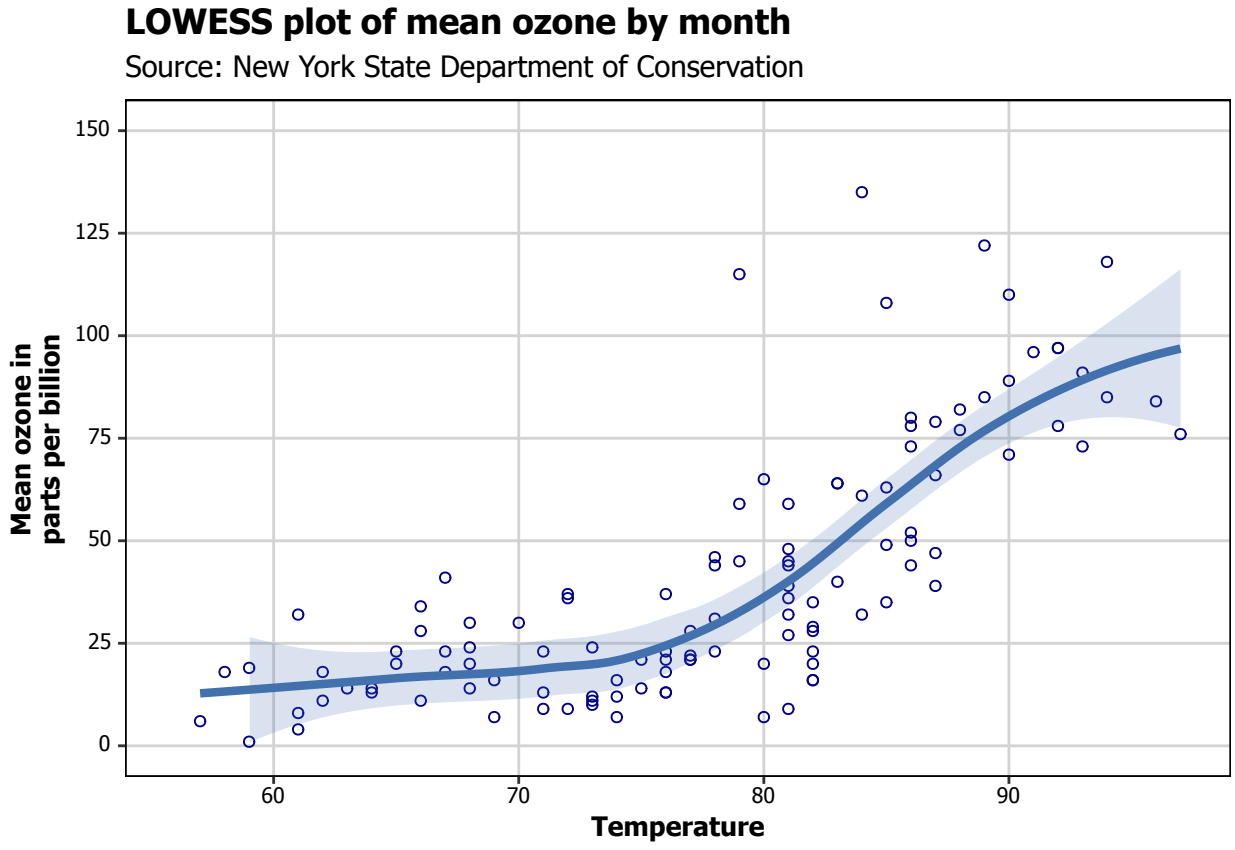


12.15. Creating your own theme

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)) +
  labs(title = "LOWESS plot of mean ozone by month",
       subtitle = "Source: New York State Department of Conservation") +
  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.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
```



Suggested material

(1) Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2009.

(2) Jenny Bryan. All the graph things.

(3) Jenny Bryan. Teaching materials for the R package ggplot2.

(4) R. Peng, J. Leek, B. Caffo. Exploratory Data Analysis.

(5) Winston Chang. *R Graphics Cookbook*. O'Reilly Media, 2012.