CSSS 508, Lecture 2

Visualizing Data

Michael Pearce
(based on slides from Chuck Lanfear)

October 6, 2022

Topics

Last time, we learned about,

1. R and RStudio
2. RMarkdown headers, syntax, and chunks
3. Basics of functions, objects, and vectors
4. Dataframes and basic plots

Today, we will cover,

1. Useful coding tips: packages, directories, and saving data
2. Advanced data manipulation tools
3. Basics of ggplot: layers and aesthetics
4. Advanced ggplot tools

1. Useful Coding Tips

Packages

Packages are collections of functions and tools to make your life easier! The large number of user-created packages in R are one of the best aspects of the language! You can see the packages installed on your machine by clicking on the `Packages` tab in the bottom-left pane of RStudio.

To install a new package in R, run the line of code:

```
install.packages("gapminder")
```

We always install packages in the **console**, because we only want to do it **once**

Installing a packages *does not* mean it's loaded in our R session. To do so, we call the package:

```
library(gapminder)
```

We need to run this code every time we open a new R session: **Where should we put this code?**

File Types

We mainly work with four file types in this class:

- `.Rmd`: These are **markdown** *syntax* files, where you write code to *make documents*.
- `.R`: These are **R** *syntax* files, where you write code to process and analyze data *without making an output document*.
- `.html` or `.pdf`: These are the output documents created when you *knit* a markdown document.

Make sure you understand the difference!

Working Directories

R saves files and looks for files to open in your current **working directory**. You can ask R what this is:

```
getwd()
```

```
## [1] "/Users/pearce790/CSSS508/Lectures/Lecture2"
```

Similarly, we can set a working directory like so:

```
setwd("C:/Users/pearce790/CSSS508/HW2")
```

Don't set a working directory in R Markdown documents! They automatically set the directory they are in as the working directory.

Managing Files

When managing R projects, it is normally best to give each project (such as a homework assignment) its own folder. I use the following system:

- Every class or project has its own folder
- Each assignment or task has a folder inside that, which is the working directory for that item.
- `.Rmd` and `.R` files are named clearly and completely

For example, this presentation is located and named this:
`GitHub/CSSS508/Lectures/Lecture2/CSSS508_Lecture2_ggplot2.Rmd`

You can use whatever system you want, but be consistent so your projects are organized! You don't want to lose work by losing or overwriting files!

Saving Data

You can save an R object on your computer as a file to open later:

```
save(new.object, file="new_object.RData")
```

(Since we didn't specify a folder, R will save the file to our current working directory!)

You can open saved files in R as well:

```
load("new_object.RData")
```

2. Advanced Data Manipulation Tools

Gapminder Data

We'll be working with data from Hans Rosling's Gapminder project. An excerpt of these data can be accessed through an R package called `gapminder`, cleaned and assembled by Jenny Bryan at UBC.

If you didn't already, run in the console: `install.packages("gapminder")`, and then load the package:

```
library(gapminder)
```

Check Out Gapminder

The data frame we will work with is called `gapminder`, available once you have loaded the package. Let's see its structure:

```
str(gapminder)
```

```
## tibble [1,704 × 6] (S3: tbl_df/tbl/data.frame)
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

What's Interesting Here?

- **Factor** variables `country` and `continent`
  - Factors are categorical data with an underlying numeric representation
  - We'll spend a lot of time on factors later!
- Many observations: $n = 1704$ rows
- For each observation, a few variables: $p = 6$ columns
- A nested/hierarchical structure: `year` in `country` in `continent`
  - These are panel data!

Into the **Tidyverse**

The **Tidyverse** is a modern collection of data science tools introduced by famed statistician Hadley Wickham. All the *cool* stats people use it!

The tidyverse is a package of packages, including:

- dplyr (data manipulation)
- ggplot2 (plotting)
- stringr (working with text data)
- and many, many more!

If you have not already installed the tidyverse, type, in the console:
`install.packages("tidyverse")`

(This may take a few seconds, because there are a *lot* of packages to install!)

Loading dplyr

We'll first use the `dplyr` package in the tidyverse, which will help us filter/slice data!

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Wait, was that an error?

When you load packages in R that have functions sharing the same name as functions you already have, the more recently loaded functions overwrite the previous ones ("masks them").

This **message** is just letting you know that. To avoid showing this in your R Markdown file, add `message=FALSE` or `include=FALSE` to your chunk options when loading packages.

Sometimes you may get a **warning message** when loading packages---usually because you aren't running the latest version of R:

```
Warning message:
package `gapminder' was built under R version 3.5.3
```

Chunk options `message=FALSE` or `include=FALSE` will hide this. *Update R* to deal with it completely!

Pipes

`dplyr` allows us to use the "pipe" data between functions using the (`%>%`) operator. So instead of nesting functions like this:

```
log(mean(gapminder$pop))
```

```
## [1] 17.20333
```

We can pipe them like this:

```
gapminder$pop %>% mean() %>% log()
```

```
## [1] 17.20333
```

Read this as, "send `gapminder$pop` to `mean()`, then send the output of that to `log()`." In essence, pipes read "left to right" while nested functions read "inside to out." This may be confusing... we'll cover it more later!

`filter` Data Frames

```
gapminder %>% filter(country == "Algeria")
```

```
## # A tibble: 12 × 6
##    country continent  year lifeExp      pop gdpPercap
##    <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Algeria Africa     1952    43.1  9279525     2449.
##  2 Algeria Africa     1957    45.7 10270856     3014.
##  3 Algeria Africa     1962    48.3 11000948     2551.
##  4 Algeria Africa     1967    51.4 12760499     3247.
##  5 Algeria Africa     1972    54.5 14760787     4183.
##  6 Algeria Africa     1977    58.0 17152804     4910.
##  7 Algeria Africa     1982    61.4 20033753     5745.
##  8 Algeria Africa     1987    65.8 23254956     5681.
##  9 Algeria Africa     1992    67.7 26298373     5023.
## 10 Algeria Africa     1997    69.2 29072015     4797.
## 11 Algeria Africa     2002    71.0 31287142     5288.
## 12 Algeria Africa     2007    72.3 33333216     6223.
```

What is this doing?

How Expressions Work

What does `country == "Algeria"` actually do?

```
head(gapminder$country == "Algeria", 50) # display first 50 elements
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [23] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [34]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [45] FALSE FALSE FALSE FALSE FALSE FALSE
```

It returns a vector of `TRUE` or `FALSE` values.

Check your understanding: How are `==`, `=` and `<-` the same or different?

Logical Operators

We used `==` for testing "equals": `country == "Algeria"`.

There are many other logical operators:

- `!=`: not equal to
- `>`, `>=`, `<`, `<=`: less than, less than or equal to, etc.
- `%in%`: used with checking equal to one of several values

Or we can combine multiple logical conditions:

- `&`: both conditions need to hold (AND)
- `|`: at least one condition needs to hold (OR)
- `!`: inverts a logical condition (`TRUE` becomes `FALSE`, `FALSE` becomes `TRUE`)

We'll practice these a lot, so don't worry right now if it's unclear!

Multiple Conditions Example

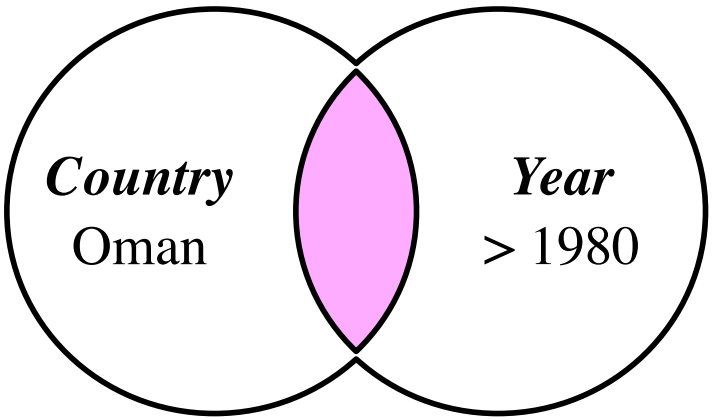
```
gapminder %>%
    filter(country == "Oman" & year > 1980)
```

```
## # A tibble: 6 × 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Oman    Asia       1982    62.7 1301048    12955.
## 2 Oman    Asia       1987    67.7 1593882    18115.
## 3 Oman    Asia       1992    71.2 1915208    18617.
## 4 Oman    Asia       1997    72.5 2283635    19702.
## 5 Oman    Asia       2002    74.2 2713462    19775.
## 6 Oman    Asia       2007    75.6 3204897    22316.
```

Multiple Conditions

And: `&`

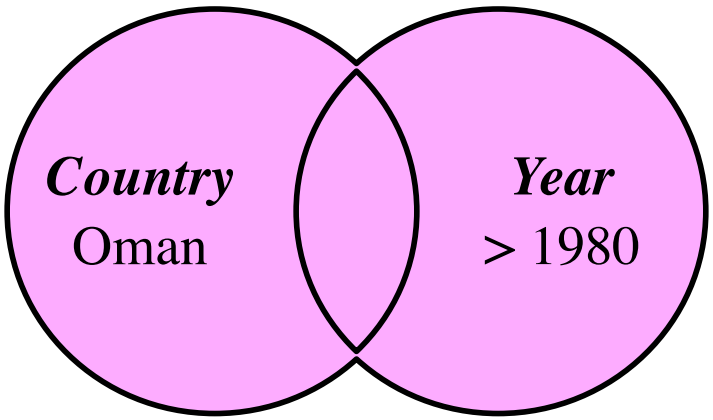
```
gapminder %>%
  filter(country == "Oman" &
         year > 1980)
```



*Give me rows where the country is Oman **and** the year is after 1980.*

Or: `|`

```
gapminder %>%
  filter(country == "Oman" |
         year > 1980)
```



*Give me rows where the country is Oman **or** the year is after 1980... or **both**.*

Saving a Subset

If we think a particular subset will be used repeatedly, we can save it and give it a name like any other object:

```
China <- gapminder %>% filter(country == "China")
head(China, 4)
```

```
## # A tibble: 4 × 6
##   country continent  year lifeExp      pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 China   Asia       1952    44   556263527      400.
## 2 China   Asia       1957    50.5 637408000      576.
## 3 China   Asia       1962    44.5 665770000      488.
## 4 China   Asia       1967    58.4 754550000      613.
```

Summary

We've discussed a lot so far:

- Saving/Loading files (`save()`, `load()`)
- Working Directories (`setwd()`, `getwd()`)
- Pipes (`gapminder %>% filter(country == "Algeria")`)
- Logical Operators (`!=`, `==`, `>`, `<=`)
- Combining Logical Operators (AND using `&`, OR using `|`)

With time and practice, these tools will become second nature!

Test Your Knowledge

Question 1: How could I filter the gapminder data to observations from Mexico after 1980?

Question 2: What's the mean life expectancy in Mexico after 1980, based on this dataset?[1]

Take 1 minute to work on this problem by yourself, then talk through your ideas with your neighbor.

Hint: Save the result from the first question in one line of code, then write a second line of code to answer the second question.

My Solution

Question 1: How could I filter the gapminder data to observations from Mexico after 1980?

```
Mexico1980 <- gapminder %>% filter(country == "Mexico" &
                                   year >= 1980)
head(Mexico1980,2)
```

```
## # A tibble: 2 × 6
##   country continent  year lifeExp      pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Mexico  Americas   1982    67.4 71640904     9611.
## 2 Mexico  Americas   1987    69.5 80122492     8688.
```

Question 2: What's the mean life expectancy in Mexico after 1980, based on this dataset?

```
Mexico1980$lifeExp %>% mean()
```

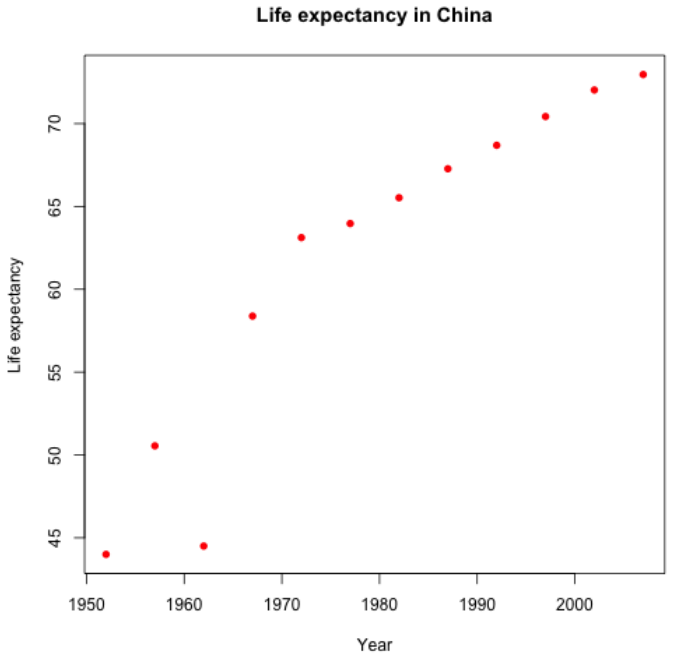
```
## [1] 72.1875
```

3. Basics of `ggplot2`

Base R Plots from Last Week

```
plot(lifeExp ~ year,
     data = China,
     xlab = "Year",
     ylab = "Life expectancy",
     main = "Life expectancy in China",
     col = "red",
     cex.lab = 1.5,
     cex.main= 1.5,
     pch = 16)
```

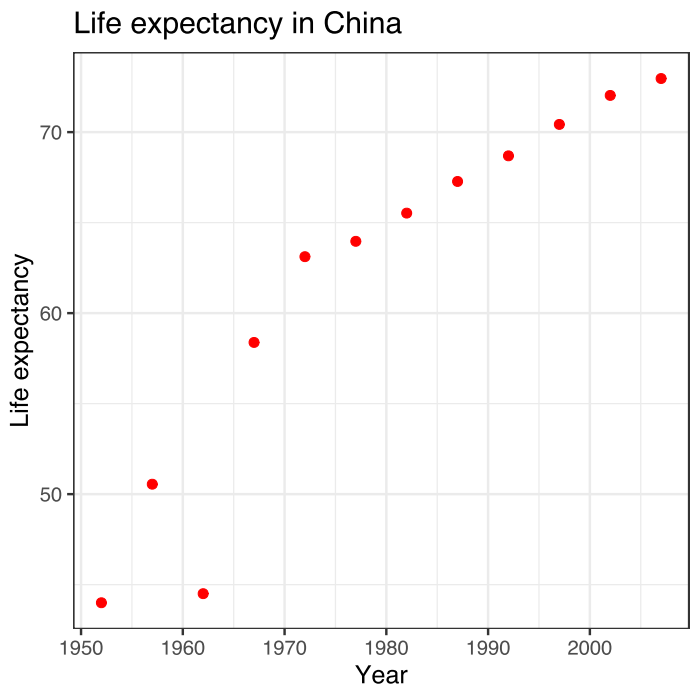
The plot is made with *one function* and *many arguments*



Starter Example in `ggplot`

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy in China") +
  theme_bw(base_size=18)
```

This `ggplot` is made with *many functions* and *fewer arguments* in each.



ggplot2

An alternative way of plotting uses the `ggplot2` package in R, which is part of the `tidyverse`.

```
library(ggplot2)
```

The core idea underlying this package is the **layered grammar of graphics**: we can break up elements of a plot into pieces and combine them.

`ggplot`s are a bit harder to create, but are usually:

- prettier,
- more professional, and
- more customizable!

Structure of a ggplot

`ggplot2` graphics objects consist of two primary components:

1. **Layers**, the components of a graph.
    - We *add* layers to a `ggplot2` object using `+`.
    - This includes adding lines, shapes, and text to a plot.
2. **Aesthetics**, which determine how the layers appear.
    - We *set* aesthetics using *arguments* (e.g. `color="red"`) inside layer functions.
    - This includes modifying locations, colors, and sizes of the layers.

Layers

Layers are the components of the graph, such as:

- `ggplot()`: initializes basic plotting object, specifies input data
- `geom_point()`: layer of scatterplot points
- `geom_line()`: layer of lines
- `geom_histogram()`: layer of a histogram
- `ggtitle()`, `xlab()`, `ylab()`: layers of labels
- `facet_wrap()`: layer creating multiple plot panels
- `theme_bw()`: layer replacing default gray background with black-and-white

Layers are separated by a `+` sign. For clarity, I usually put each layer on a new line.

Aesthetics

Aesthetics control the appearance of the layers:

- `x`, `y`: x and y coordinate values to use
- `color`: set color of elements based on some data value
- `group`: describe which points are conceptually grouped together for the plot (often used with lines)
- `size`: set size of points/lines based on some data value (greater than 0)
- `alpha`: set transparency based on some data value (between 0 and 1)

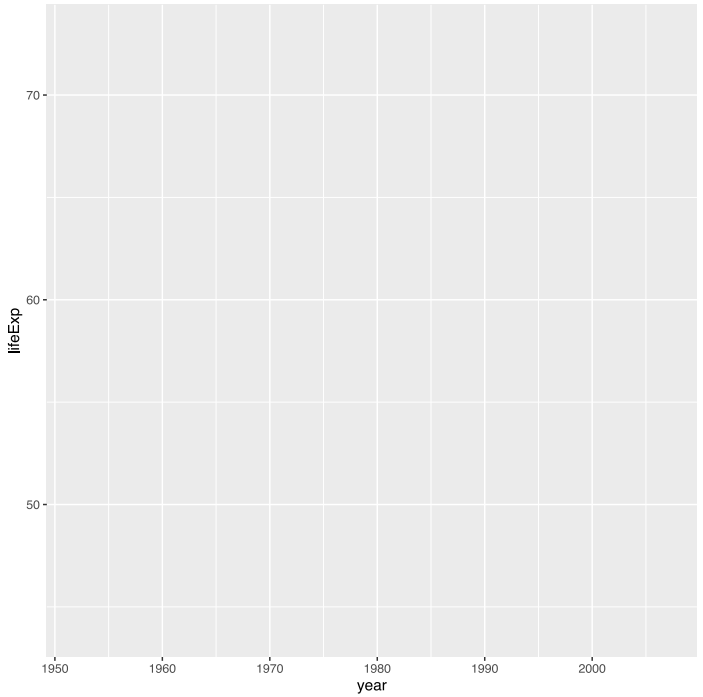
Examples: Basic Jargon in Action!

We'll now build up two `ggplot`s together that demonstrate common layers and aesthetics.

Axis Labels, Points, No Background

1: Base Plot

```
ggplot(data = China,
       aes(x = year, y = lifeExp))
```

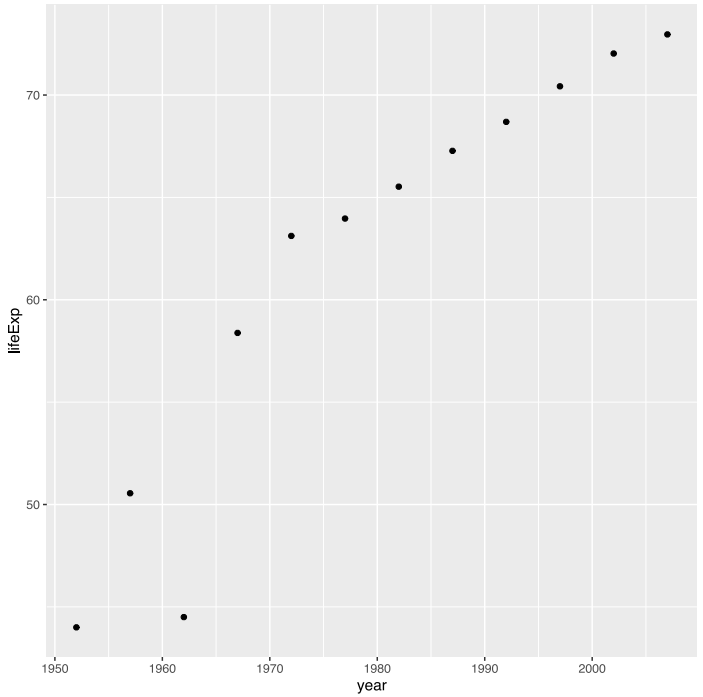


Initialize the plot with `ggplot()` and `x` and `y` aesthetics **mapped** to variables. These aesthetics will be accesible to any future layers since they're in the primary layer.

Axis Labels, Points, No Background

2: Scatterplot

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point()
```

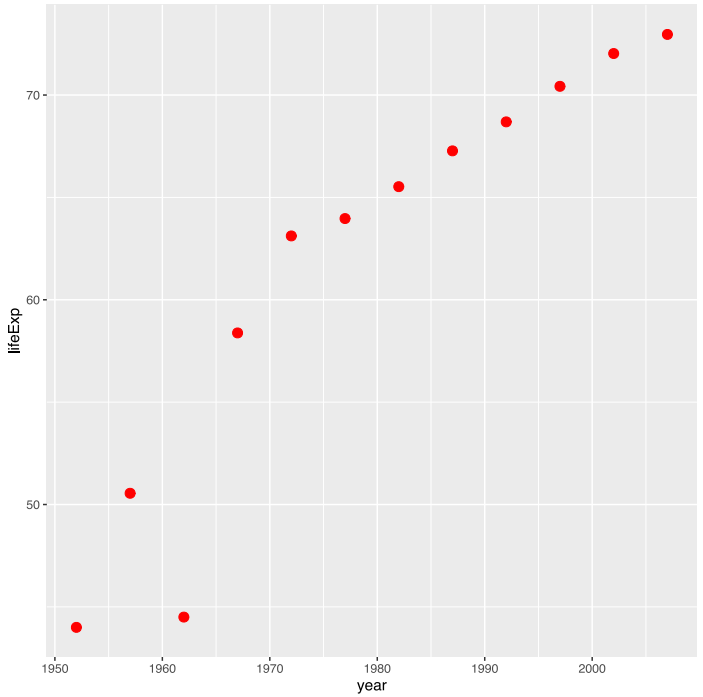


Add a scatterplot **layer**.

Axis Labels, Points, No Background

3: Point Color and Size

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3)
```

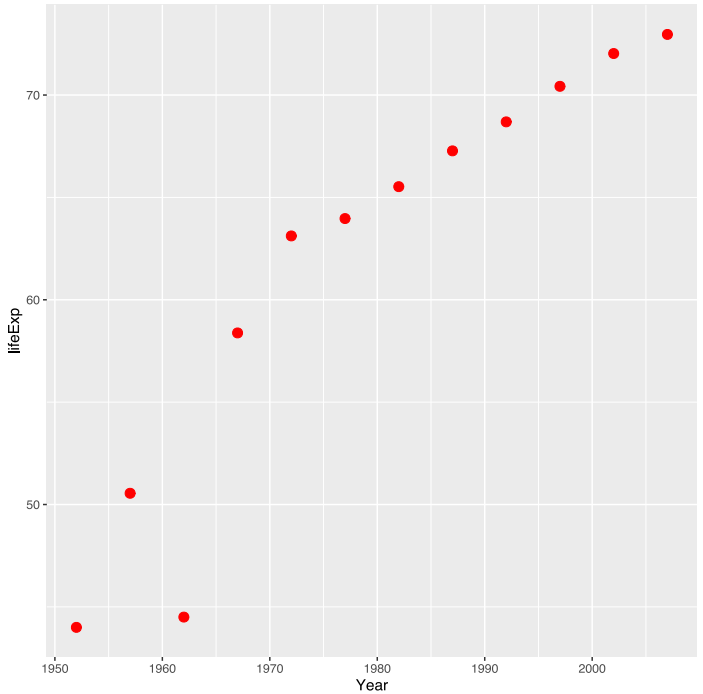


Set aesthetics to make the points large and red.

Axis Labels, Points, No Background

4: X-Axis Label

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3) +
  xlab("Year")
```

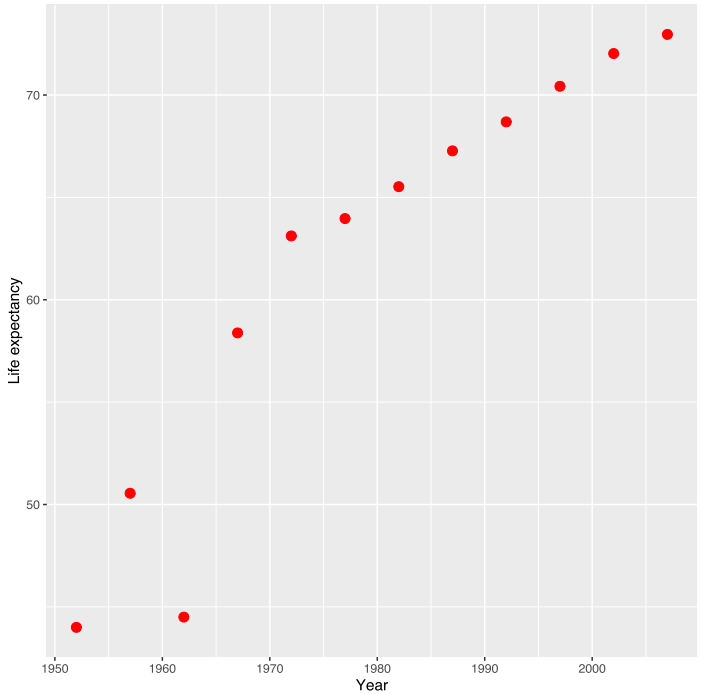


Add a layer to capitalize the x-axis label.

Axis Labels, Points, No Background

5: Y-Axis Label

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy")
```

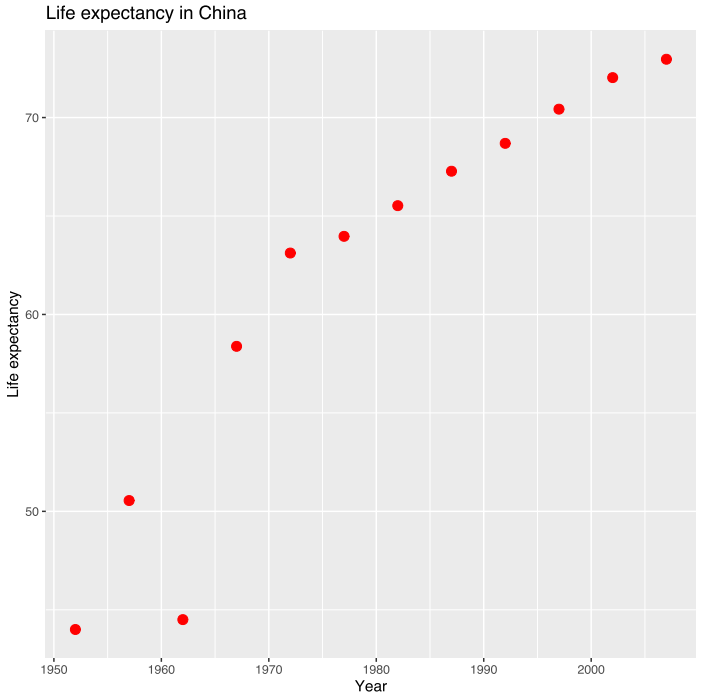


Add a layer to clean up the y-axis label.

Axis Labels, Points, No Background

6: Title

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy in China")
```

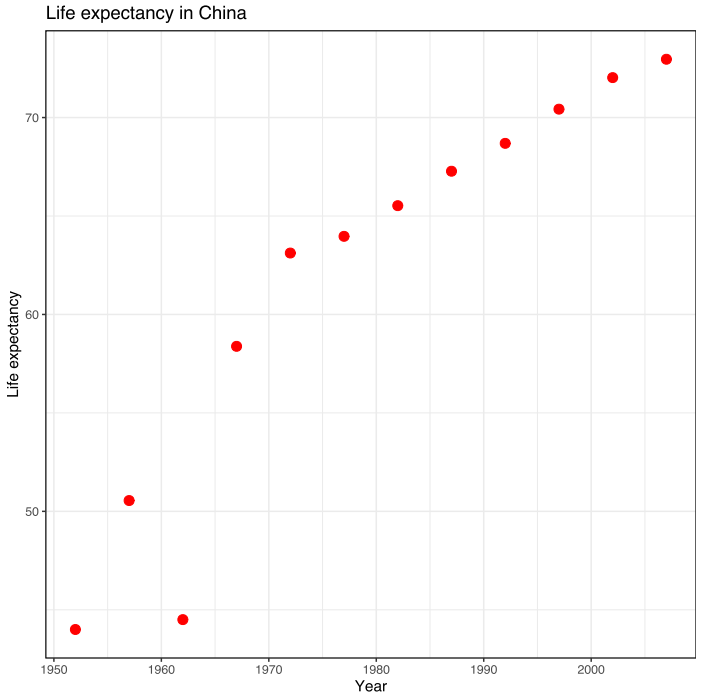


Add a title layer.

Axis Labels, Points, No Background

7: Theme

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy in China") +
  theme_bw()
```

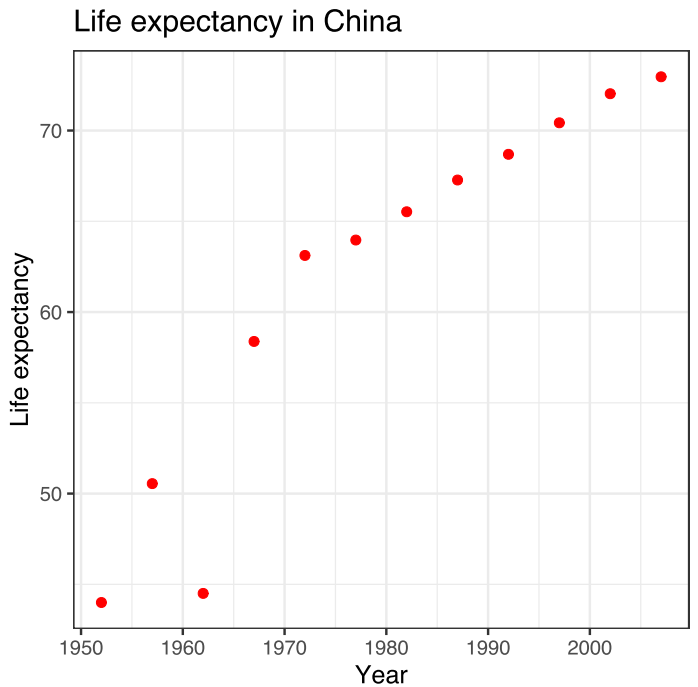


Pick a nicer theme with a new layer.

Axis Labels, Points, No Background

8: Text Size

```
ggplot(data = China,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy in China") +
  theme_bw(base_size=18)
```



Increase the base text size.

Plotting All Countries

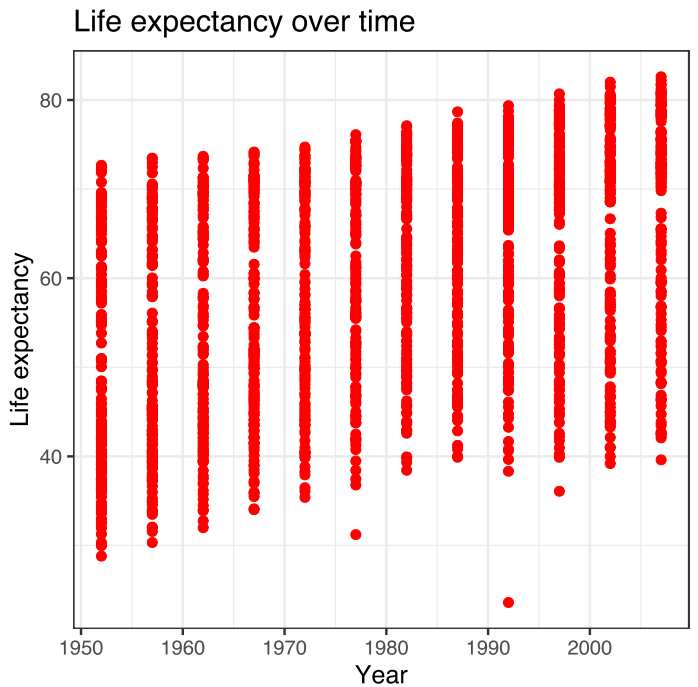
We have a plot we like for China…

… but what if we want *all the countries*?

Plotting All Countries

1: A Mess!

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp)) +
  geom_point(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw(base_size=18)
```

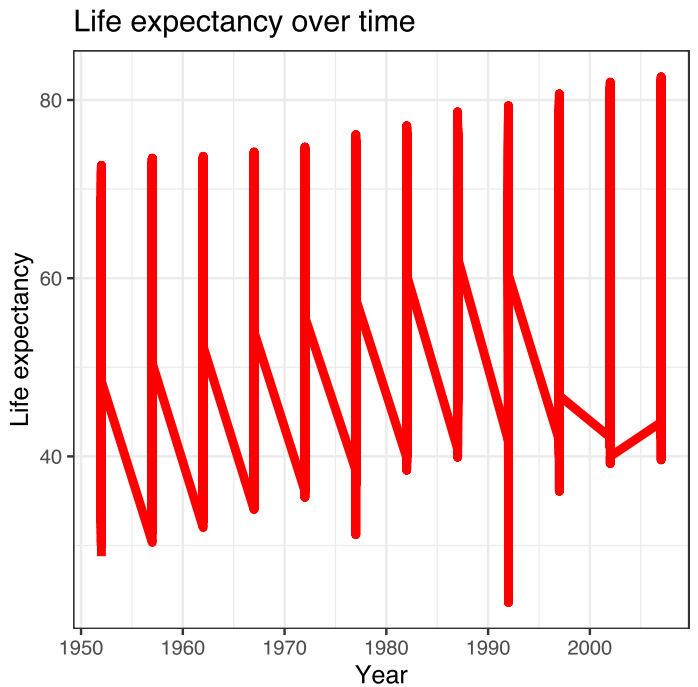


We can't tell countries apart! Maybe we could follow *lines*?

Plotting All Countries

2: Lines

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp)) +
  geom_line(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw(base_size=18)
```

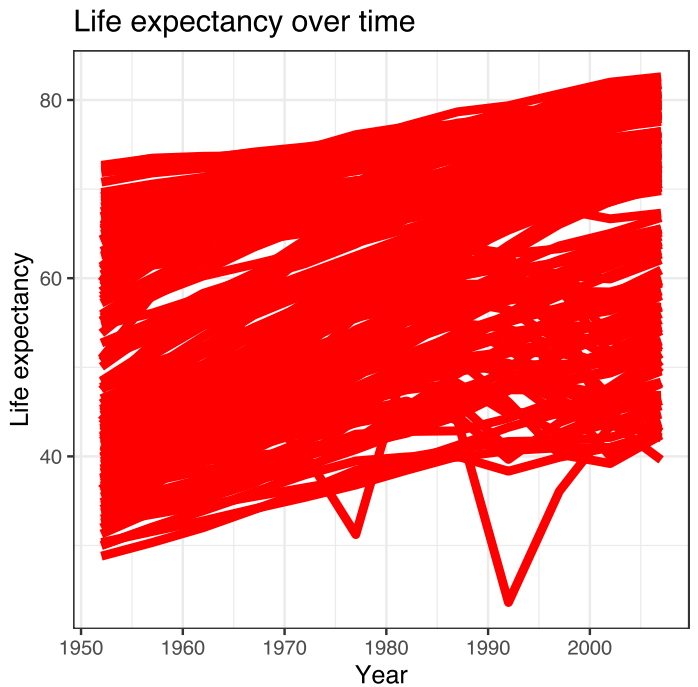


`ggplot2` doesn't know how to connect the lines!

Plotting All Countries

3: Grouping

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp,
           group = country)) +
  geom_line(color = "red", size = 3) +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw(base_size=18)
```

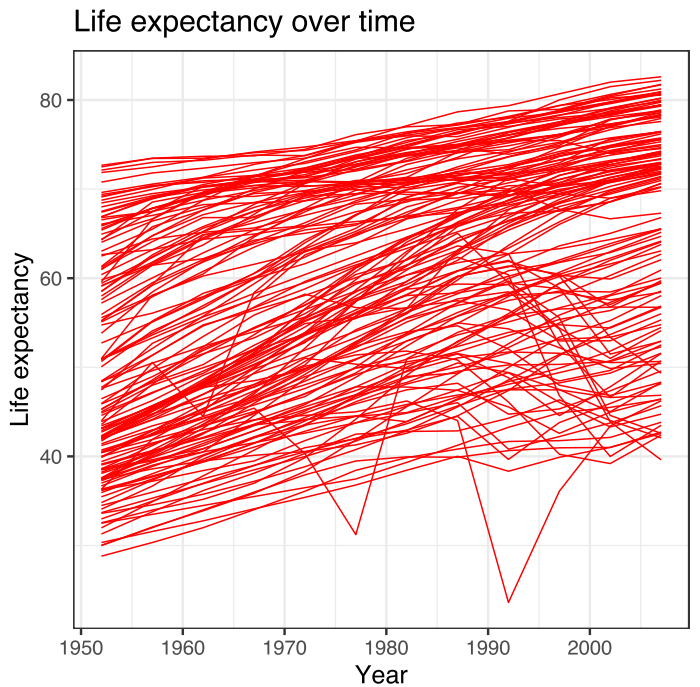


That looks more reasonable... but the lines are too thick!

Plotting All Countries

4: Size

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp,
           group = country)) +
  geom_line(color = "red") +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw(base_size=18)
```

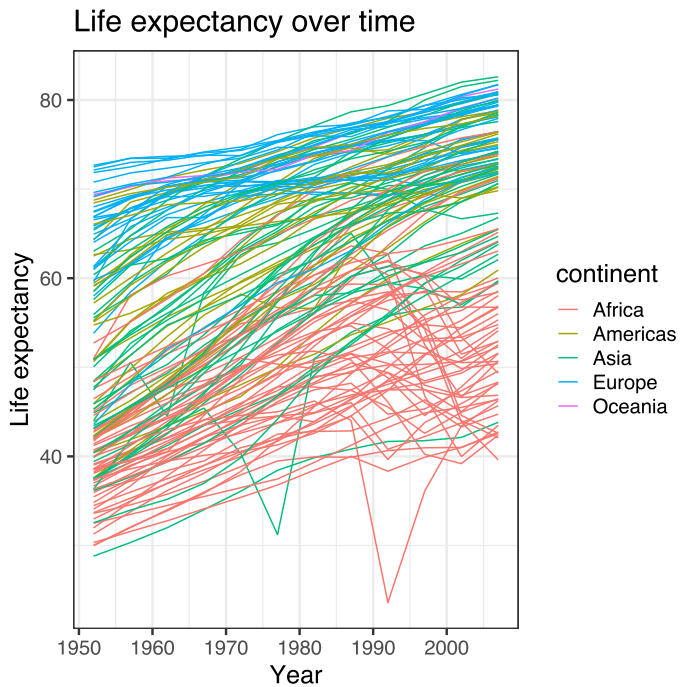


Much better... but maybe we can do highlight regional differences?

Plotting All Countries

5: Color

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp,
           group = country,
           color = continent)) +
  geom_line() +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw(base_size=18)
```

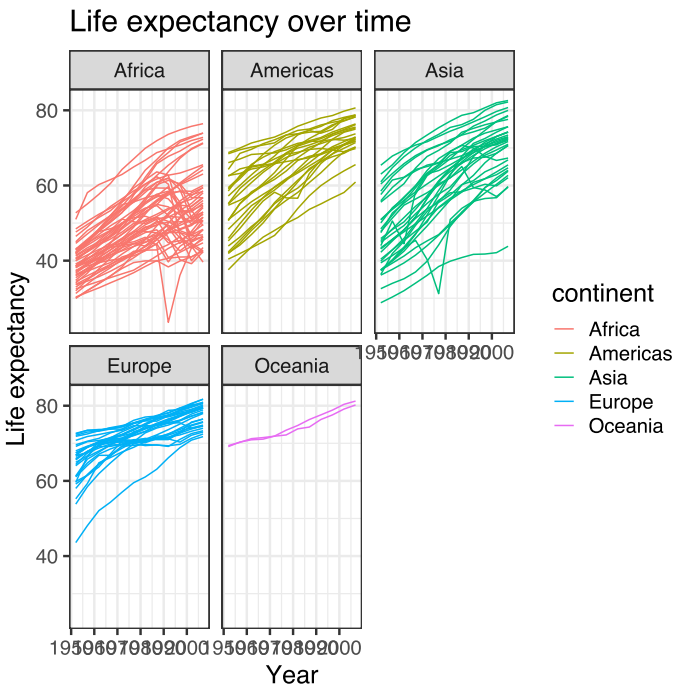


Patterns are obvious... but why not separate continents completely?

Plotting All Countries

6: Facets

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp,
           group = country,
           color = continent)) +
  geom_line() +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw(base_size=18) +
  facet_wrap(~ continent)
```

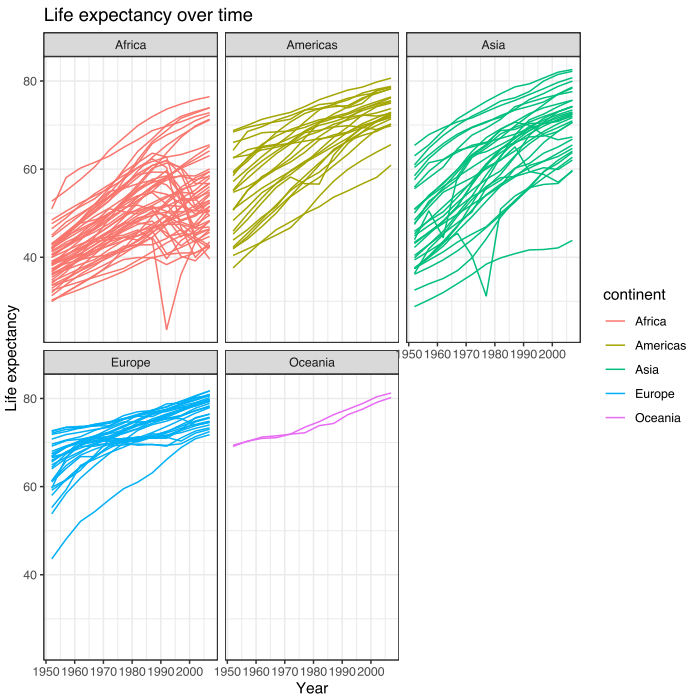


Now the text is too big!

Plotting All Countries

7: Text Size

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp,
           group = country,
           color = continent)) +
  geom_line() +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw() +
  facet_wrap(~ continent)
```

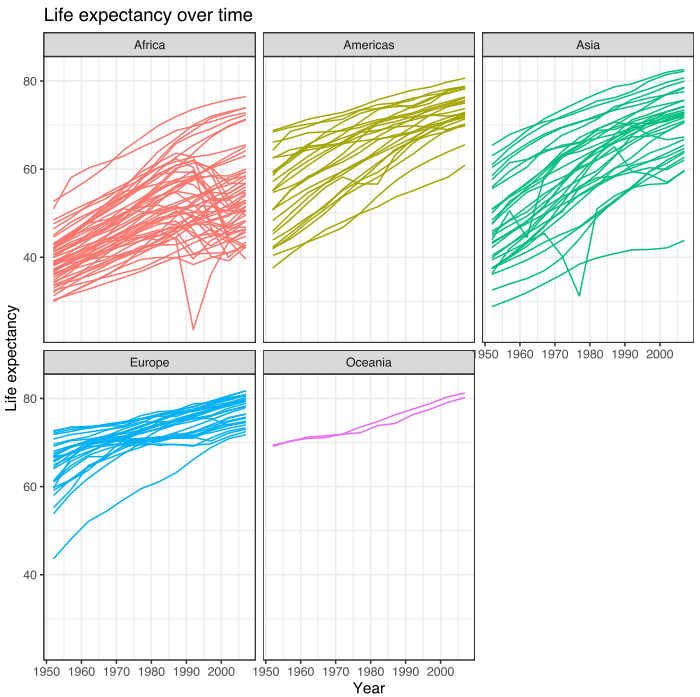


Better. Do we even need the legend anymore?

Plotting All Countries

8: No Legend

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp,
           group = country,
           color = continent)) +
  geom_line() +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw() +
  facet_wrap(~ continent) +
  theme(legend.position = "none")
```



Looking good!

4. Advanced ggplot tools

Next, we'll discuss:

- Storing, modifying, and saving ggplots
- Advanced axis changes (scales, text, ticks)
- *Legendary* legend changes (scales, colors, locations)

Storing Plots

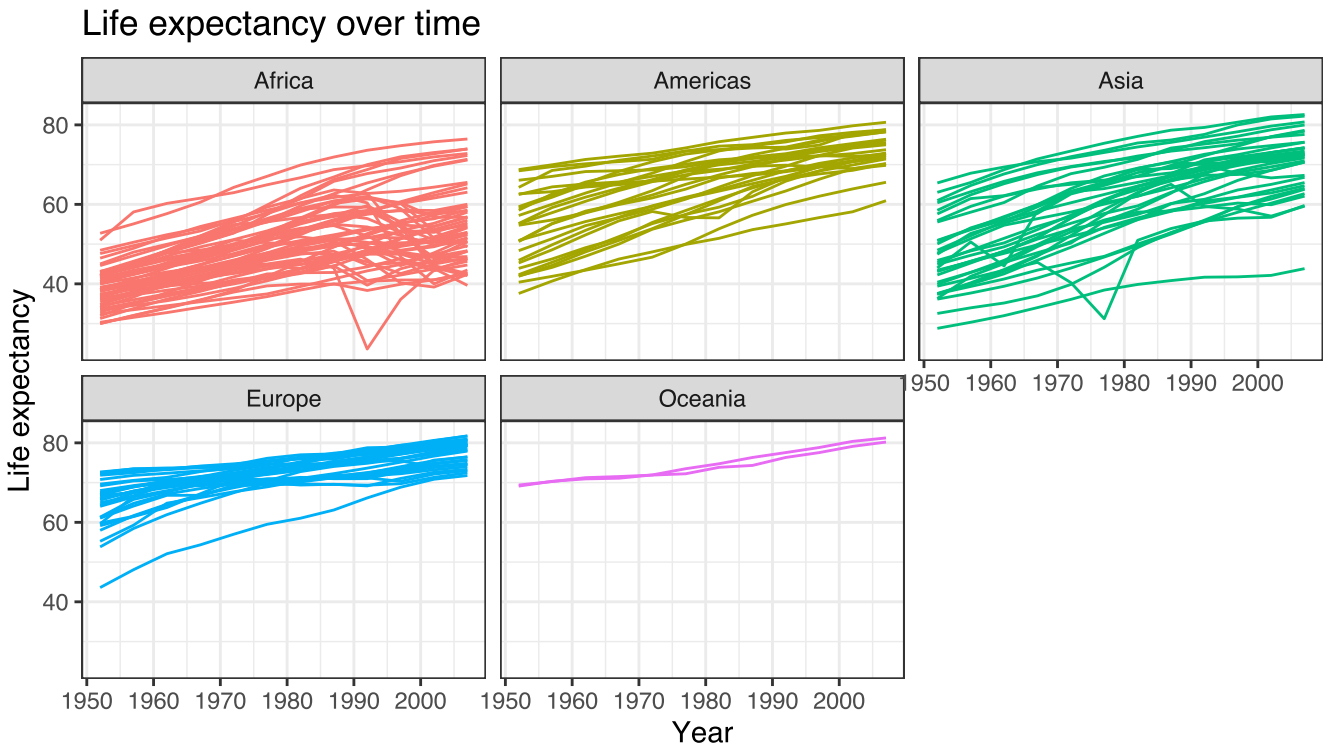
We can assign a `ggplot` object to a name:

```
lifeExp_by_year <-
  ggplot(data = gapminder,
       aes(x = year, y = lifeExp,
           group = country,
           color = continent)) +
  geom_line() +
  xlab("Year") +
  ylab("Life expectancy") +
  ggtitle("Life expectancy over time") +
  theme_bw() +
  facet_wrap(~ continent) +
  theme(legend.position = "none")
```

The graph won't be displayed when you do this. You can show the graph using a single line of code with just the object name, *or take the object and add more layers*.

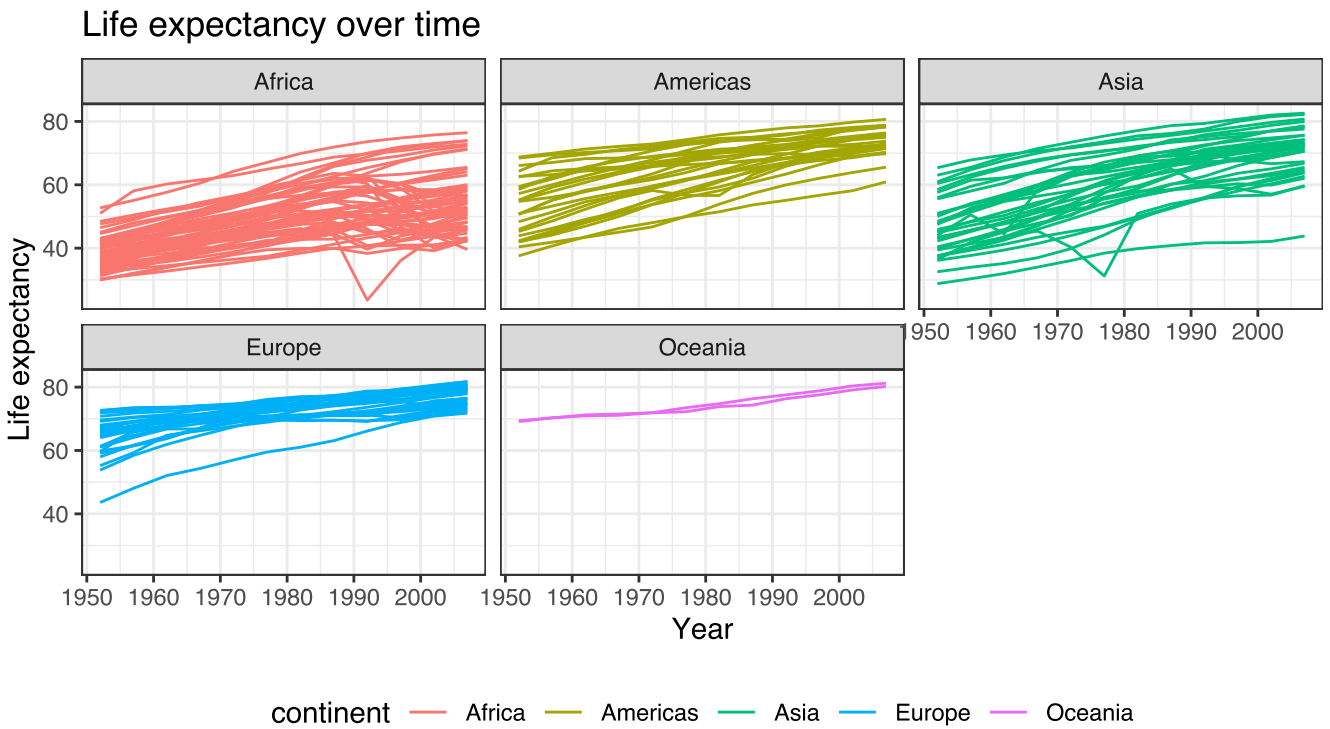
Showing a Stored Graph

```
lifeExp_by_year
```



Adding a Layer

```
lifeExp_by_year +
    theme(legend.position = "bottom")
```



Saving `ggplot` Plots

When you knit an R Markdown file, any plots you make are automatically saved in the "figure" folder in `.png` format. If you want to save another copy (perhaps of a different file type for use in a manuscript), use `ggsave()`:

```
ggsave("I_saved_a_file.pdf", plot = lifeExp_by_year,
       height = 3, width = 5, units = "in")
```

If you didn't manually set font sizes, these will usually come out at a reasonable size given the dimensions of your output file.

Bad/non-reproducible way[1]: choose *Export* on the plot preview or take a screenshot / snip.

[1] I still do this for quick emails of simple plots. Bad me!

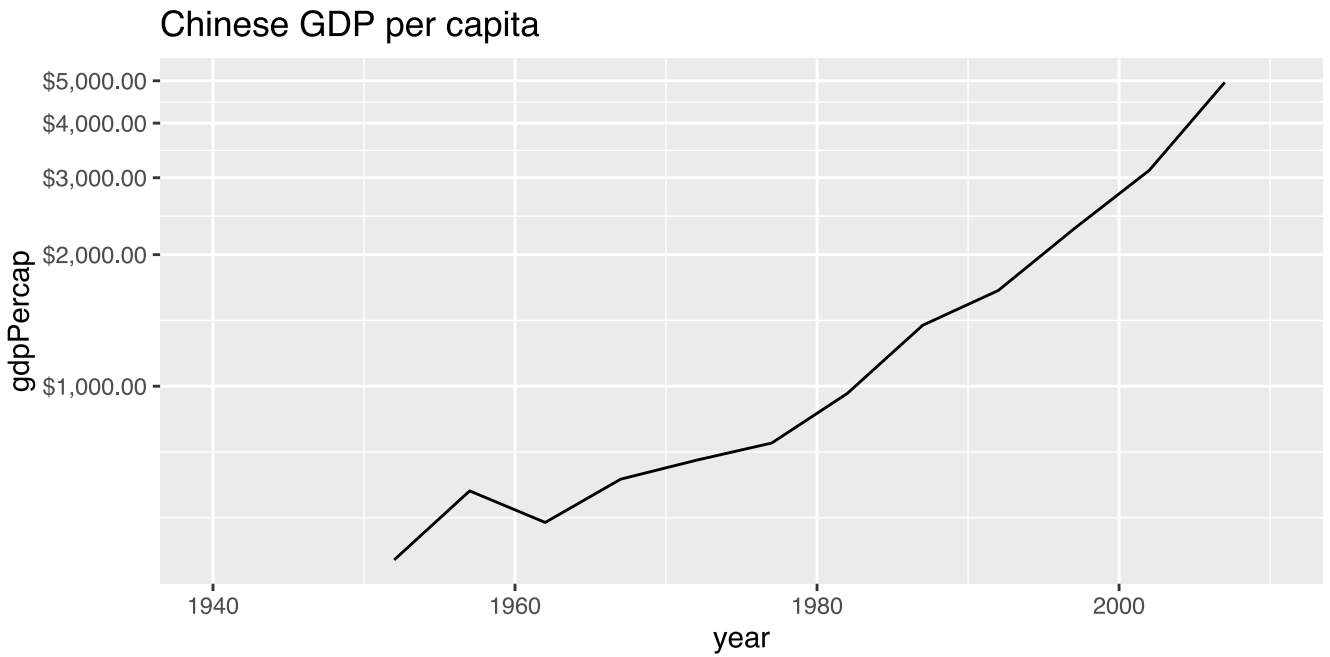
Changing the Axes

We can modify the axes in a variety of ways, such as:

- Change the x or y range using `xlim()` or `ylim()` layers
- Change to a logarithmic or square-root scale on either axis: `scale_x_log10()`, `scale_y_sqrt()`
- Change where the major/minor breaks are: `scale_x_continuous(breaks =, minor_breaks = )`

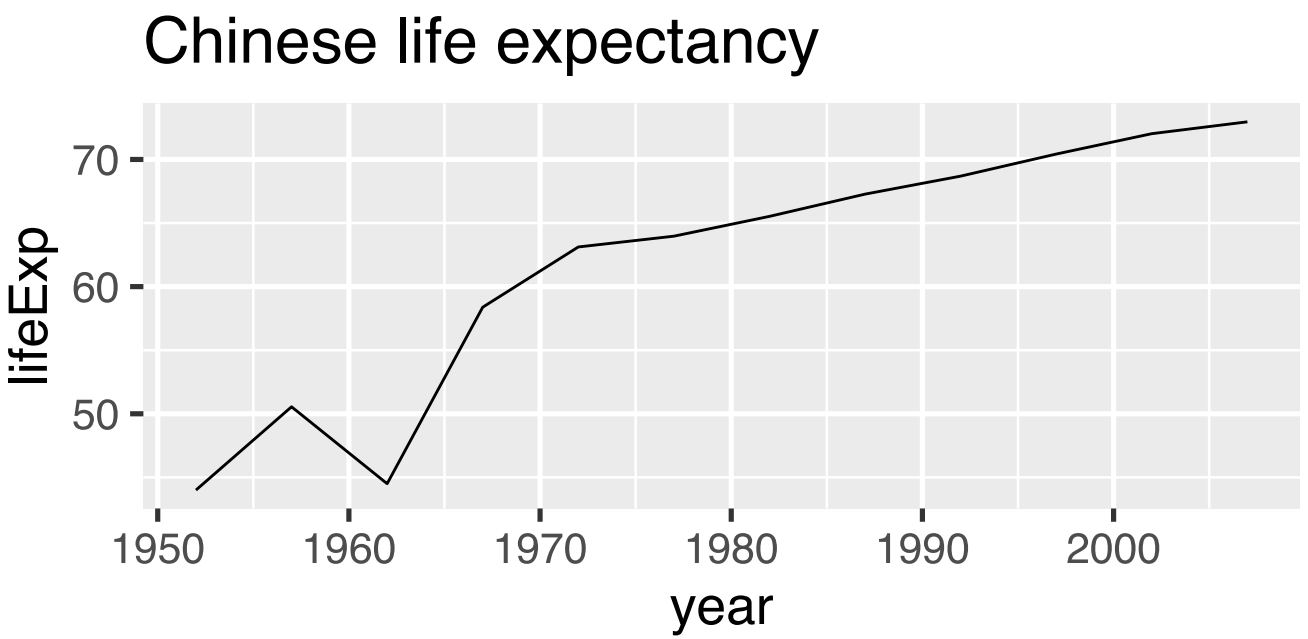
Axis Changes

```
ggplot(data = China, aes(x = year, y = gdpPercap)) +
    geom_line() +
    scale_y_log10(breaks = c(1000, 2000, 3000, 4000, 5000),
                  labels = scales::dollar) +
    xlim(1940, 2010) + ggtitle("Chinese GDP per capita")
```



Fonts Too Small?

```
ggplot(data = China, aes(x = year, y = lifeExp)) +
    geom_line() +
    ggtitle("Chinese life expectancy") +
    theme_gray(base_size = 20)
```



Text and Tick Adjustments

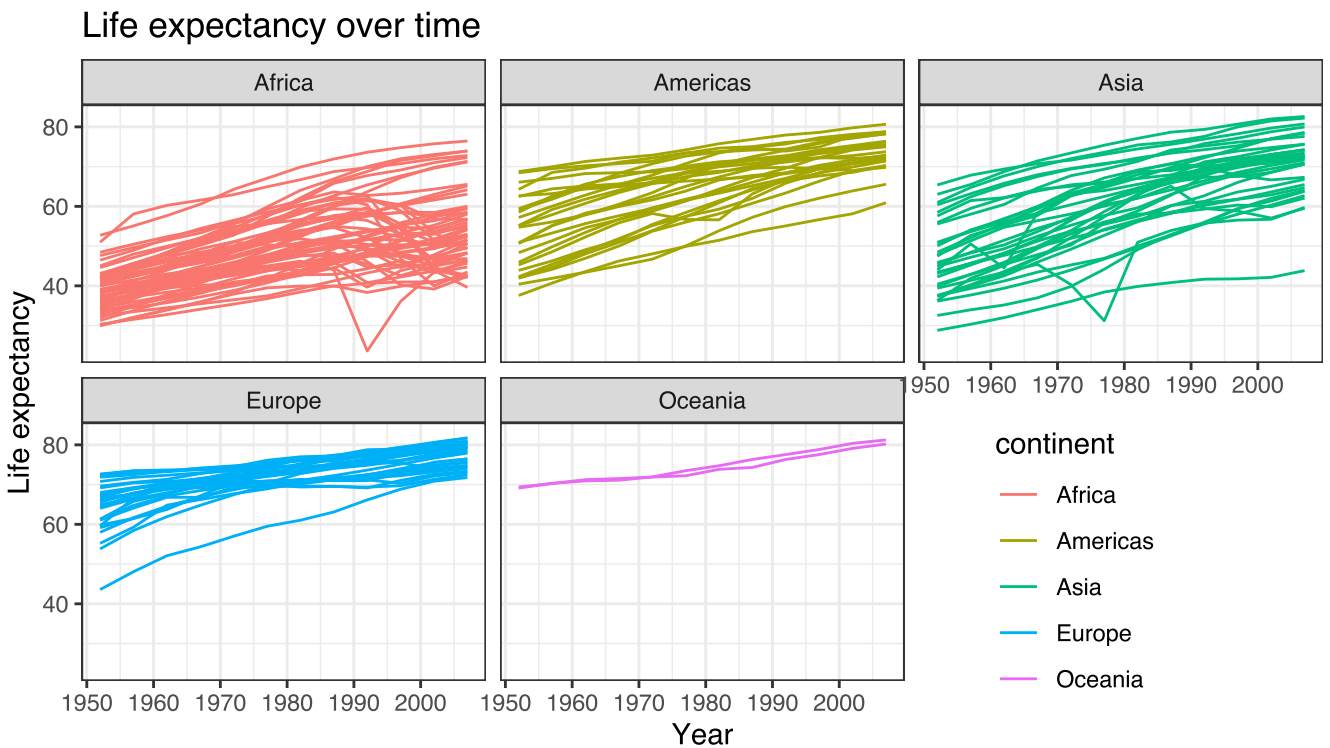
Text size, labels, tick marks, etc. can be messed with more precisely using arguments to the `theme()` layer.

- `plot.title = element_text(size = rel(2), hjust = 0)` makes the title twice as big as usual and left-aligns it
- `axis.text.x = element_text(angle = 45)` rotates x axis labels
- `axis.text = element_text(colour = "blue")` makes the x and y axis labels blue
- `axis.ticks.length = unit(.5, "cm")` makes the axis ticks longer

I recommend using `theme()` *after* `theme_bw()` or other *global themes*.

Precise Legend Position

```
lifeExp_by_year +
  theme(legend.position = c(0.8, 0.2))
```



Instead of coordinates, you could also use "top", "bottom", "left", or "right".

Scales for Color, Shape, etc.

Scales are layers that control how the mapped aesthetics appear.

You can modify these with a `scale_[aesthetic]_[option]()` layer:

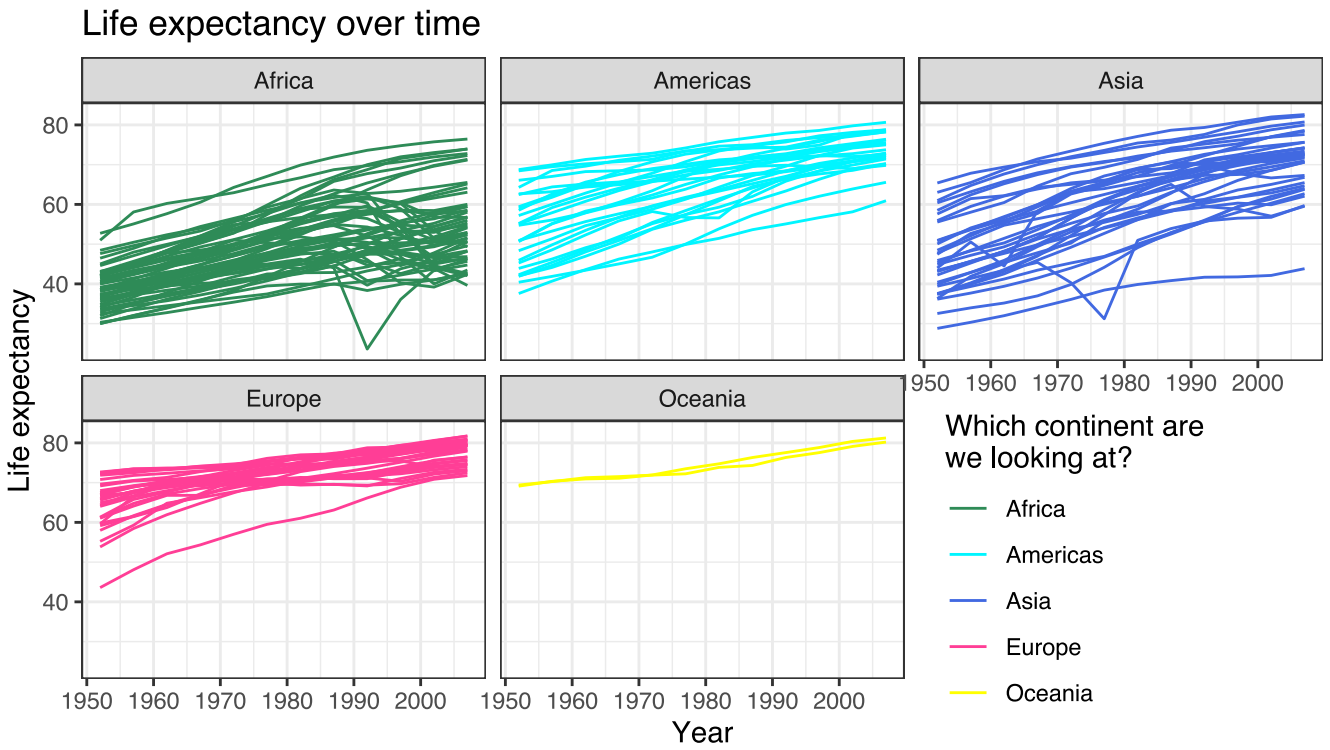
- `[aesthetic]` is `color`, `shape`, `linetype`, `alpha`, `size`, `fill`, etc.
- `[option]` is something like `manual`, `continuous` or `discrete` (depending on nature of the variable).

Examples:

- `scale_linetype_manual()`: manually specify the linetype for each different value
- `scale_alpha_continuous()`: varies transparency over a continuous range
- `scale_color_manual()`: manually specify colors

Legend Name and Manual Colors

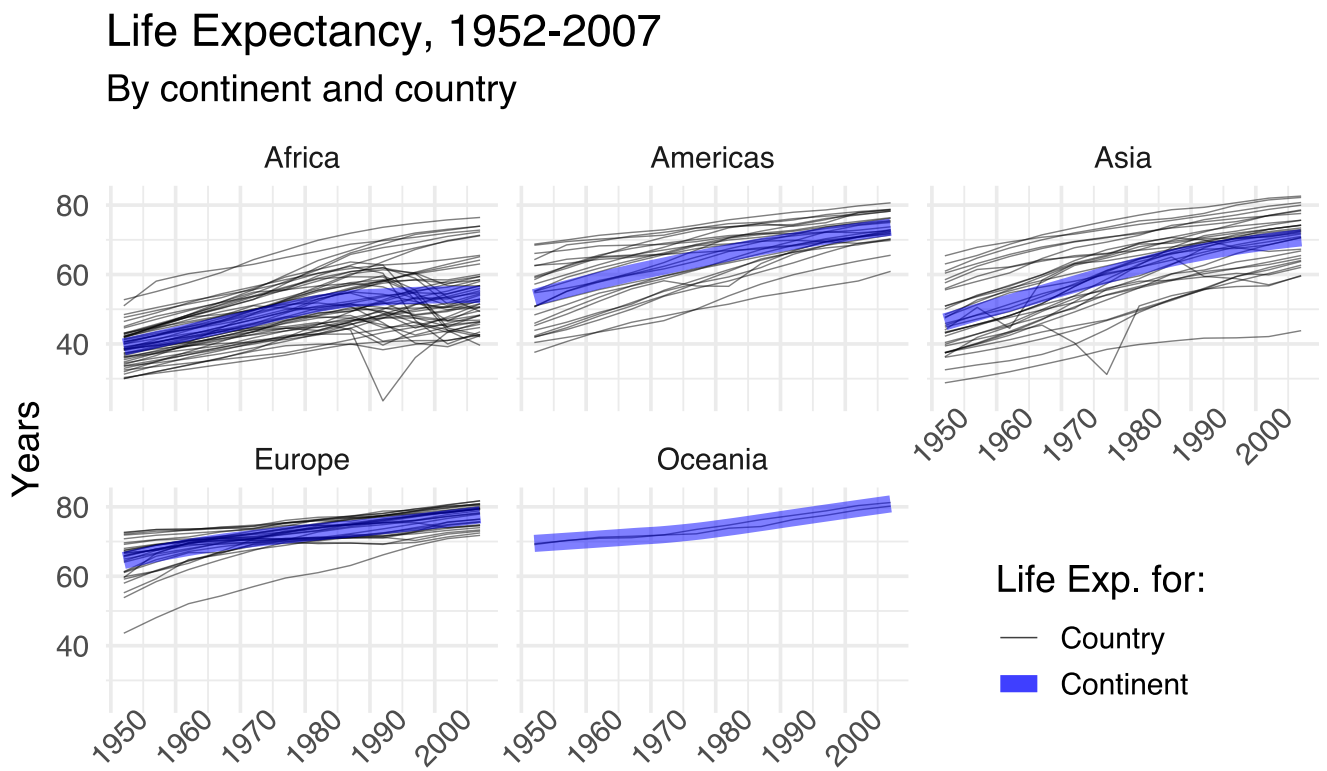
```
lifeExp_by_year +
  theme(legend.position = c(0.8, 0.2)) +
  scale_color_manual(
    name = "Which continent are\nwe looking at?", # \n adds a line break
    values = c("Africa" = "seagreen", "Americas" = "turquoise1",
               "Asia" = "royalblue", "Europe" = "violetred1", "Oceania" = "yellow"))
```



Bonus: Advanced Example!

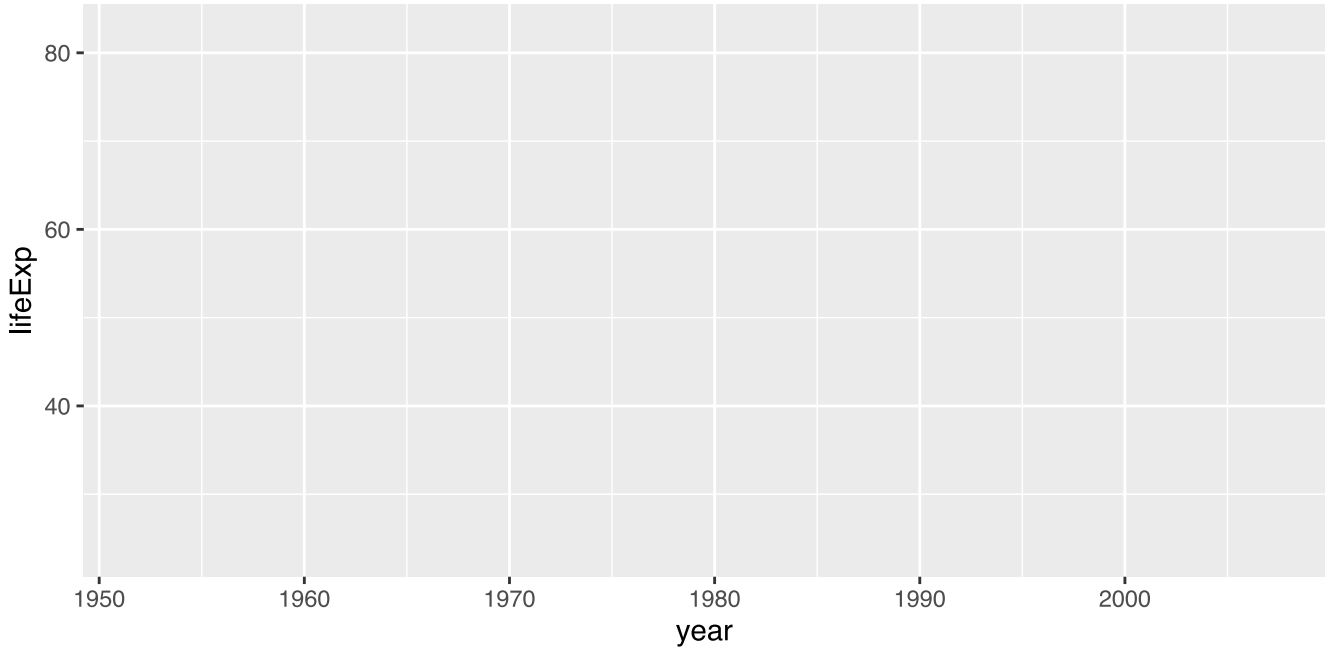
End Result

We're going to *slowly* build up a *really detailed plot* now!



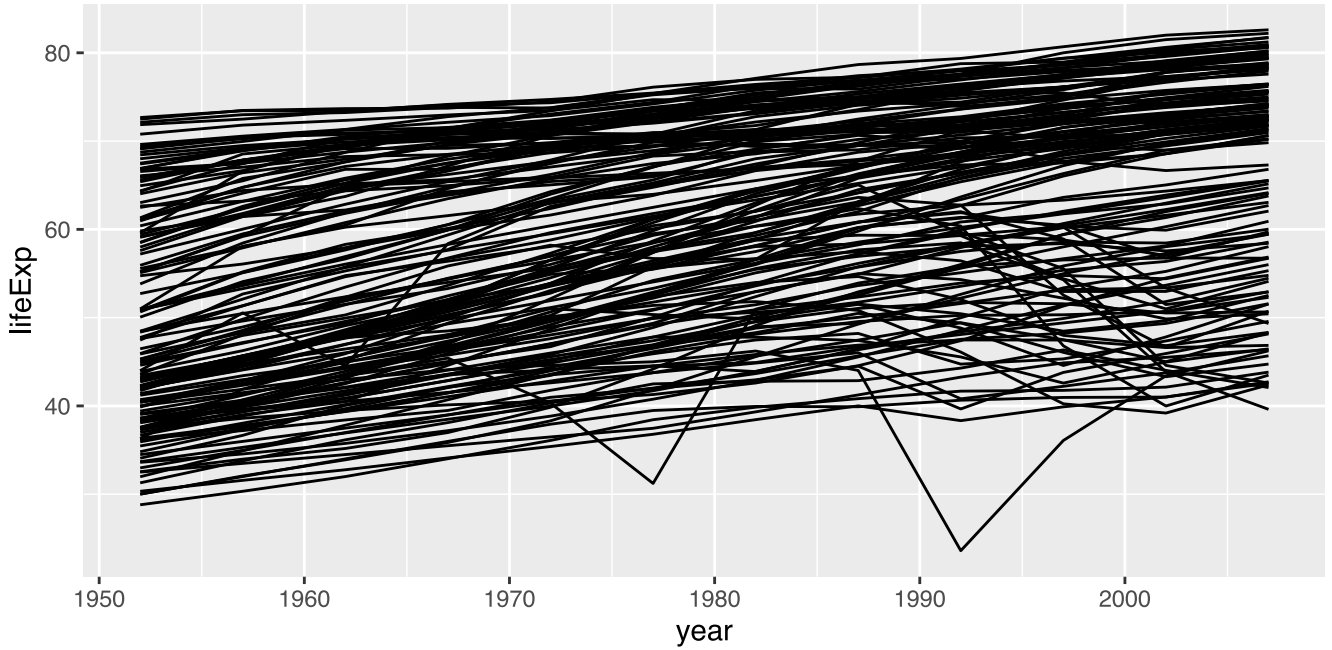
1. Base Plot

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country))
  #
  #
  #
  #
  #
  #
  #
  #
  #
```



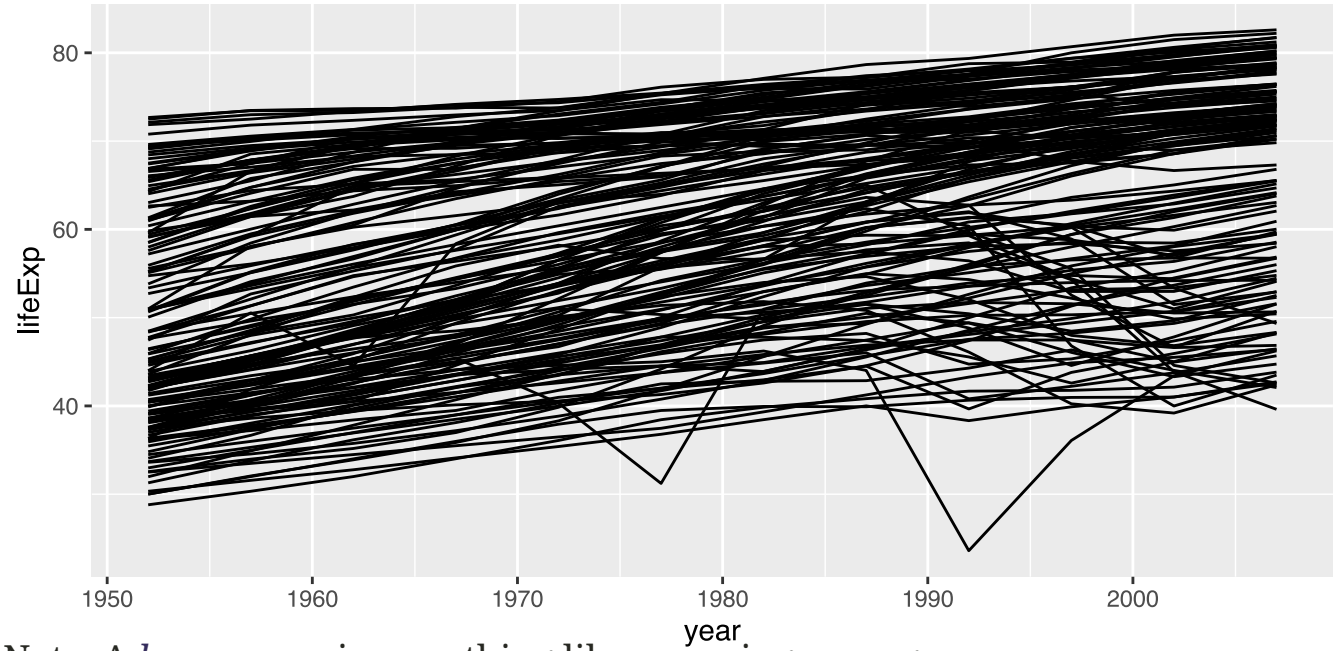
2. Lines

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line()
  #
  #
  #
  #
  #
  #
  #
  #
```



3. Continent Average

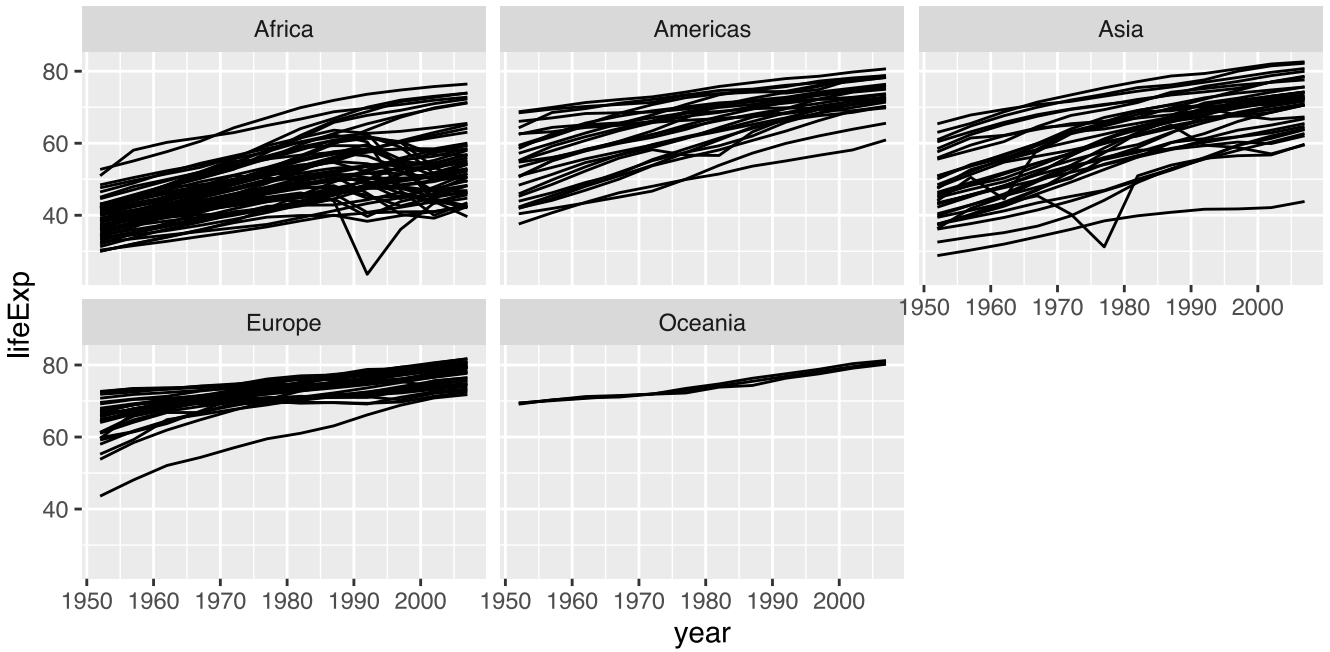
```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line() +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent))
  #
  #
  #
  #
  #
  #
```



Note: A *loess* curve is something like a moving average.

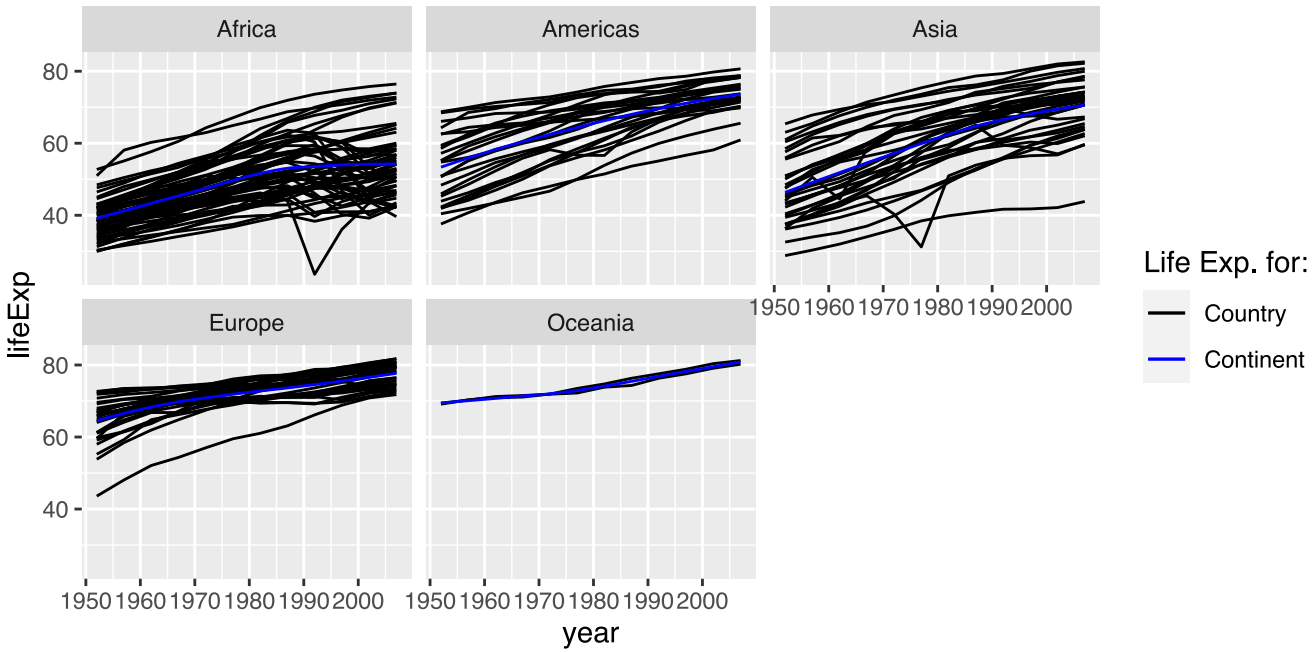
4. Facets

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line() +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent)) +
  facet_wrap(~ continent, nrow = 2)
  #
  #
  #
  #
  #
```



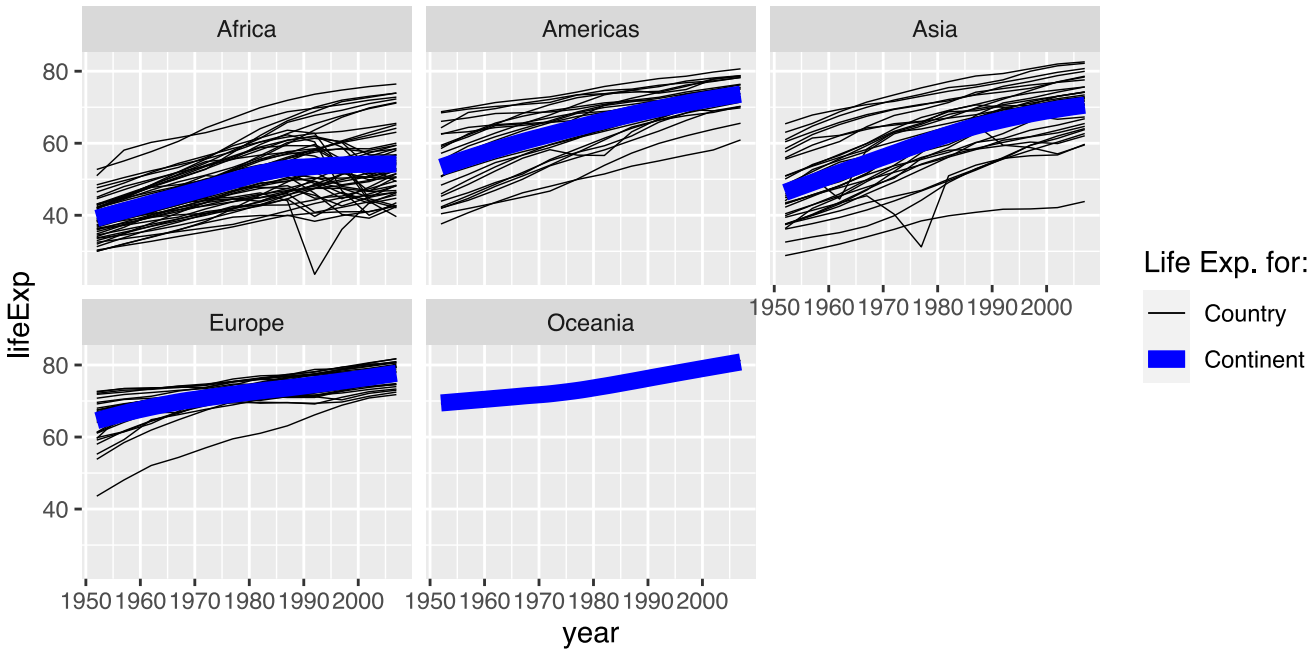
5. Color Scale

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line(aes(color = "Country")) +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent, color = "Continent")) +
  facet_wrap(~ continent, nrow = 2) +
  scale_color_manual(name = "Life Exp. for:", values = c("Country" = "black", "Continent" = "blue"))
  #
  #
  #
  #
```



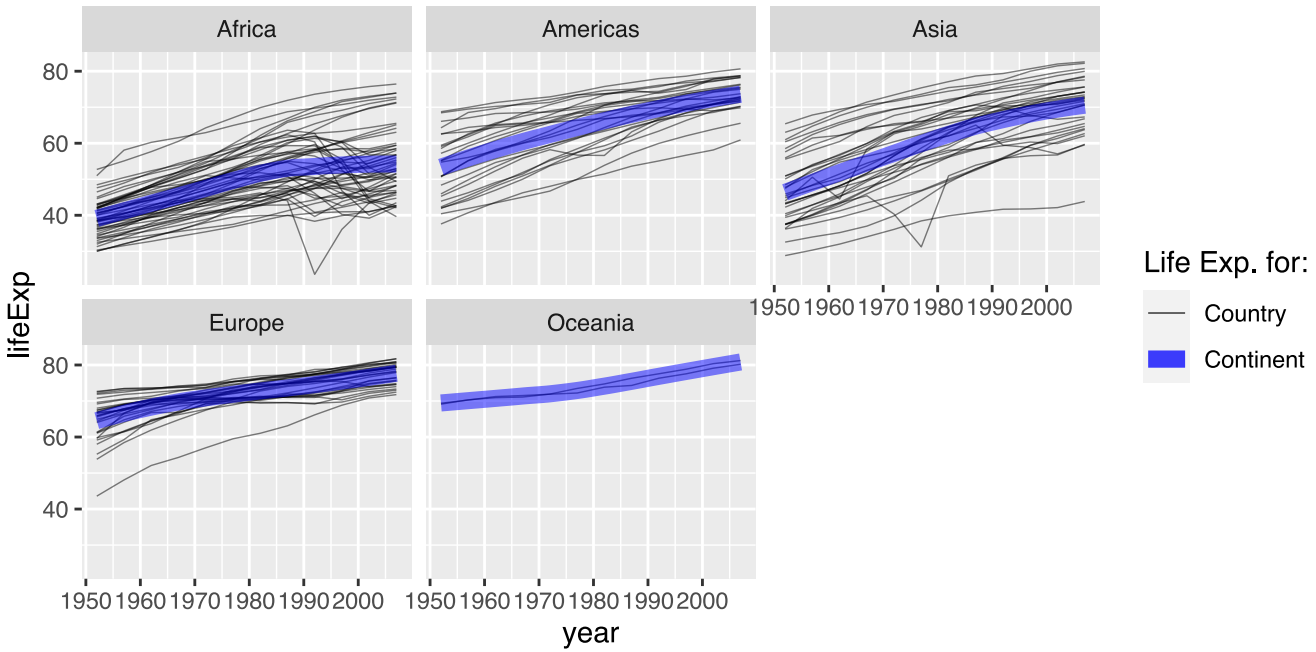
6. Size Scale

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line(aes(color = "Country", size = "Country")) +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent, color = "Continent", size = "Continent")) +
  facet_wrap(~ continent, nrow = 2) +
  scale_color_manual(name = "Life Exp. for:", values = c("Country" = "black", "Continent" = "blue")) +
  scale_size_manual(name = "Life Exp. for:", values = c("Country" = 0.25, "Continent" = 3))
  #
  #
  #
```



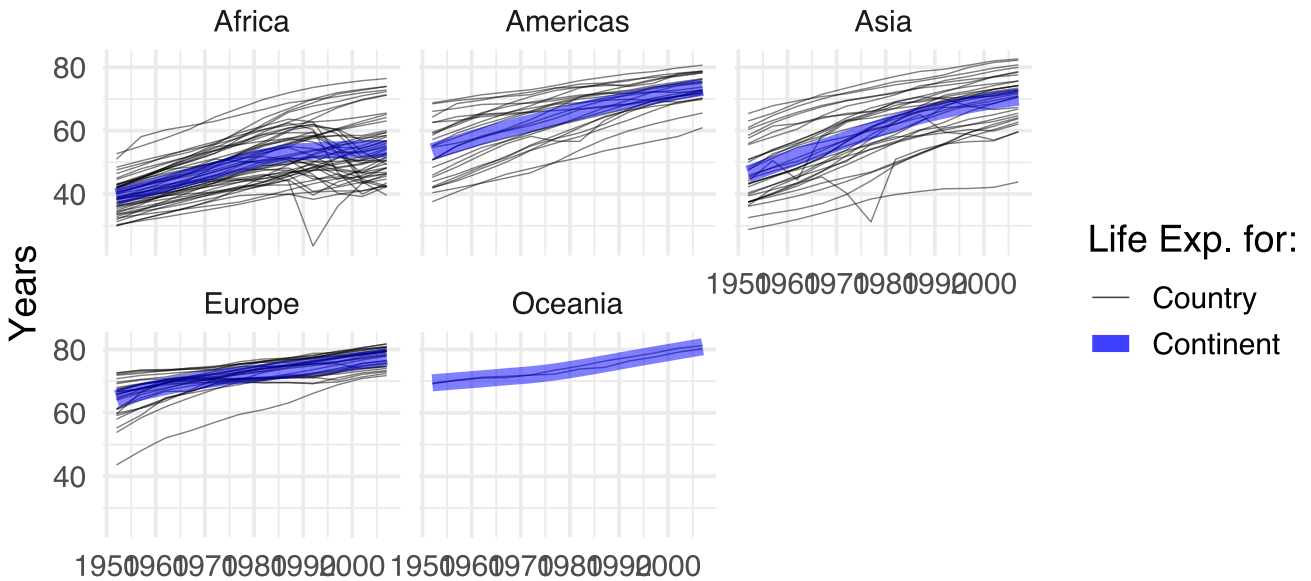
7. Alpha (Transparency)

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line(alpha = 0.5, aes(color = "Country", size = "Country")) +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent, color = "Continent", size = "Continent"), alpha = 0.5) +
  facet_wrap(~ continent, nrow = 2) +
  scale_color_manual(name = "Life Exp. for:", values = c("Country" = "black", "Continent" = "blue")) +
  scale_size_manual(name = "Life Exp. for:", values = c("Country" = 0.25, "Continent" = 3))
  #
  #
  #
```



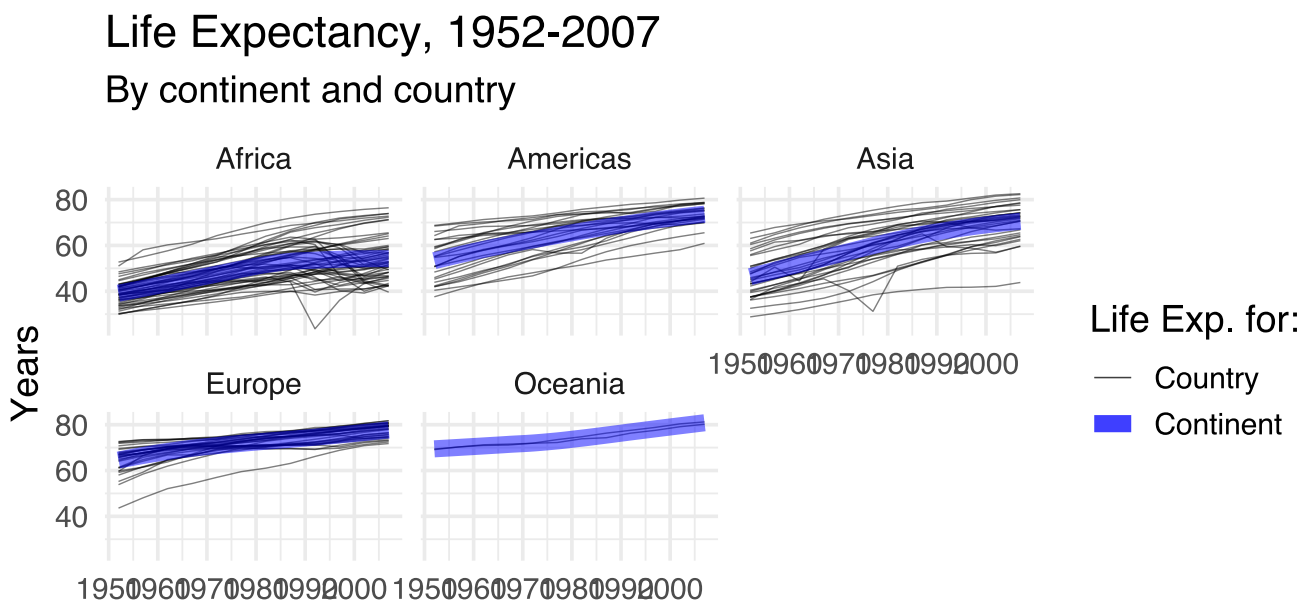
8. Theme and Labels

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line(alpha = 0.5, aes(color = "Country", size = "Country")) +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent, color = "Continent", size = "Continent"), alpha = 0.5) +
  facet_wrap(~ continent, nrow = 2) +
  scale_color_manual(name = "Life Exp. for:", values = c("Country" = "black", "Continent" = "blue")) +
  scale_size_manual(name = "Life Exp. for:", values = c("Country" = 0.25, "Continent" = 3)) +
  theme_minimal(base_size = 14) + ylab("Years") + xlab("")
  #
  #
```



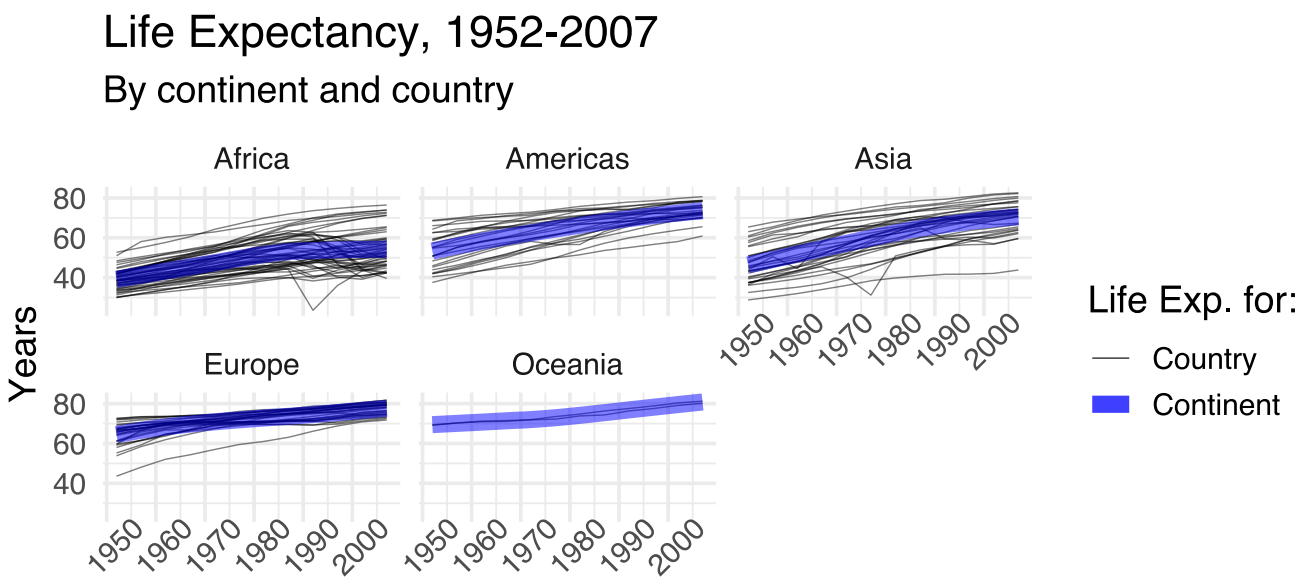
9. Title and Subtitle

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line(alpha = 0.5, aes(color = "Country", size = "Country")) +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent, color = "Continent", size = "Continent"), alpha = 0.5) +
  facet_wrap(~ continent, nrow = 2) +
  scale_color_manual(name = "Life Exp. for:", values = c("Country" = "black", "Continent" = "blue")) +
  scale_size_manual(name = "Life Exp. for:", values = c("Country" = 0.25, "Continent" = 3)) +
  theme_minimal(base_size = 14) + ylab("Years") + xlab("") +
  ggtitle("Life Expectancy, 1952-2007", subtitle = "By continent and country")
  #
```



10. Angled Tick Values

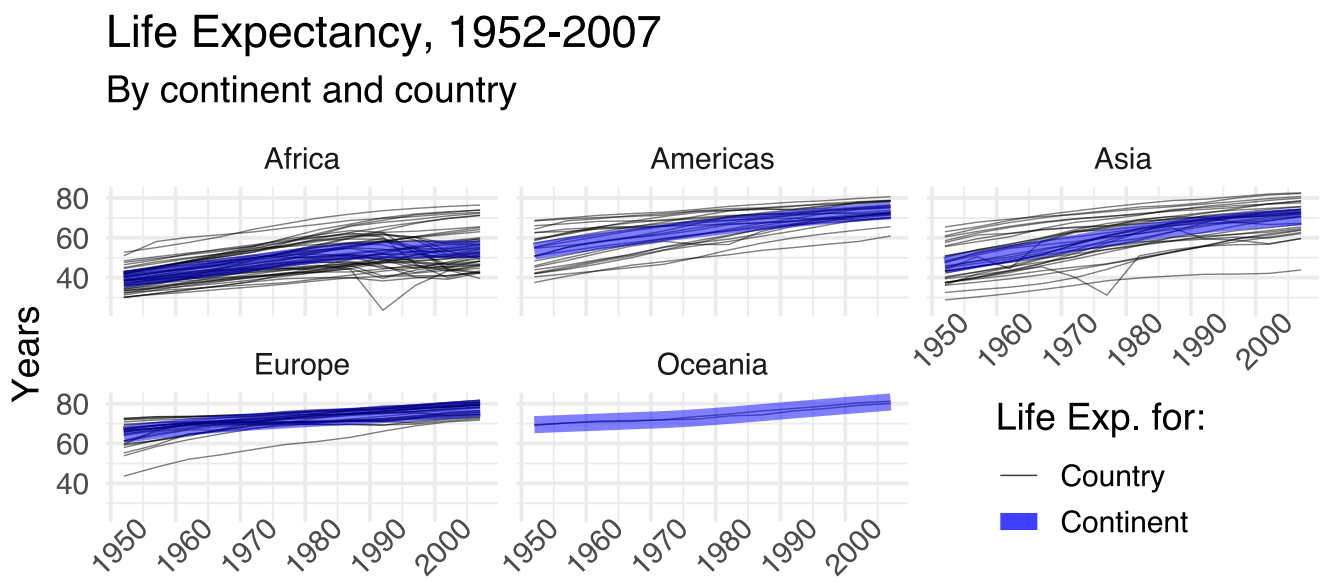
```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line(alpha = 0.5, aes(color = "Country", size = "Country")) +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent, color = "Continent", size = "Continent"), alpha = 0.5) +
  facet_wrap(~ continent, nrow = 2) +
  scale_color_manual(name = "Life Exp. for:", values = c("Country" = "black", "Continent" = "blue")) +
  scale_size_manual(name = "Life Exp. for:", values = c("Country" = 0.25, "Continent" = 3)) +
  theme_minimal(base_size = 14) + ylab("Years") + xlab("") +
  ggtitle("Life Expectancy, 1952-2007", subtitle = "By continent and country") +
  theme(axis.text.x = element_text(angle = 45))
```



Note: Fewer values might be better than angled labels!

11. Legend Position

```
ggplot(data = gapminder, aes(x = year, y = lifeExp, group = country)) +
  geom_line(alpha = 0.5, aes(color = "Country", size = "Country")) +
  geom_line(stat = "smooth", method = "loess",
            aes(group = continent, color = "Continent", size = "Continent"), alpha = 0.5) +
  facet_wrap(~ continent, nrow = 2) +
  scale_color_manual(name = "Life Exp. for:", values = c("Country" = "black", "Continent" = "blue")) +
  scale_size_manual(name = "Life Exp. for:", values = c("Country" = 0.25, "Continent" = 3)) +
  theme_minimal(base_size = 14) + ylab("Years") + xlab("") +
  ggtitle("Life Expectancy, 1952-2007", subtitle = "By continent and country") +
  theme(legend.position=c(0.82, 0.15), axis.text.x = element_text(angle = 45))
```

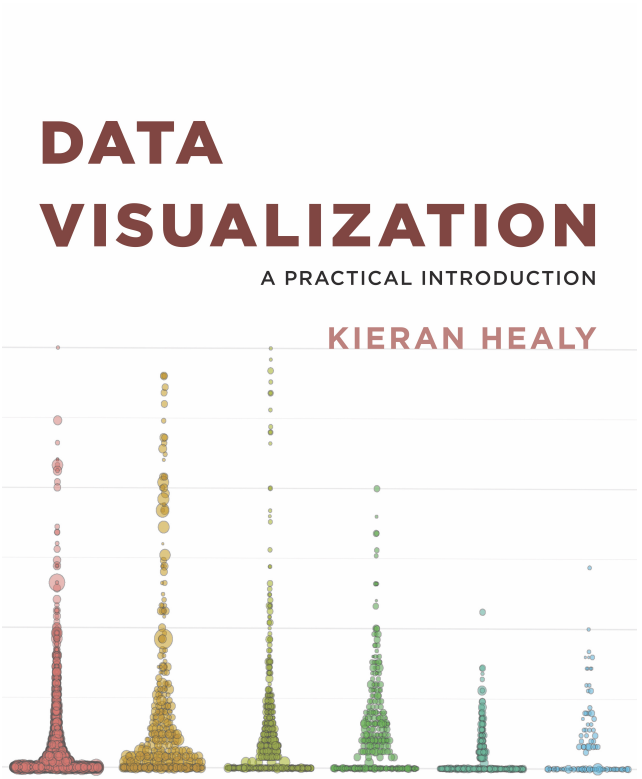


Summary

Summary

`ggplot2` can do a LOT! I won't expect you to remember all these tools.

- With time and practice, you'll start to remember the key tools
- When in doubt, Google it! ("*R ggplot rename title*")
- There are lots of great resources out there:
  - The Cookbook for R website
  - The RStudio ggplot Cheatsheets.
  - Kieran Healy's book Data Visualization: A Practical Introduction (right) which is targeted at social scientists without technical backgrounds.



Exercise: Histograms

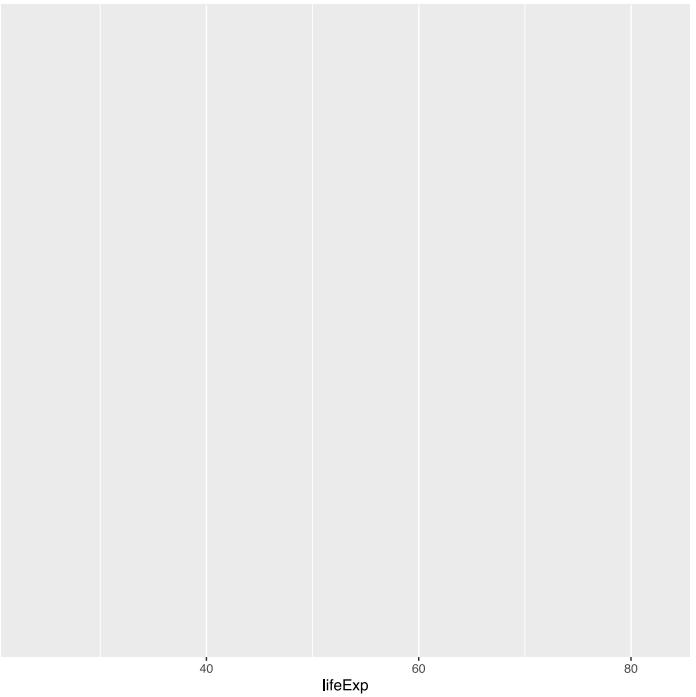
In pairs, you will create a histogram of life expectancy observations in the complete Gapminder dataset.

1. Set the base layer by specifying the data as `gapminder` and the x variable as `lifeExp`
2. Add a second layer to create a histogram using the function `geom_histogram()`
3. Customize your plot with nice axis labels and a title.

My Solution

1: Set Base Layer

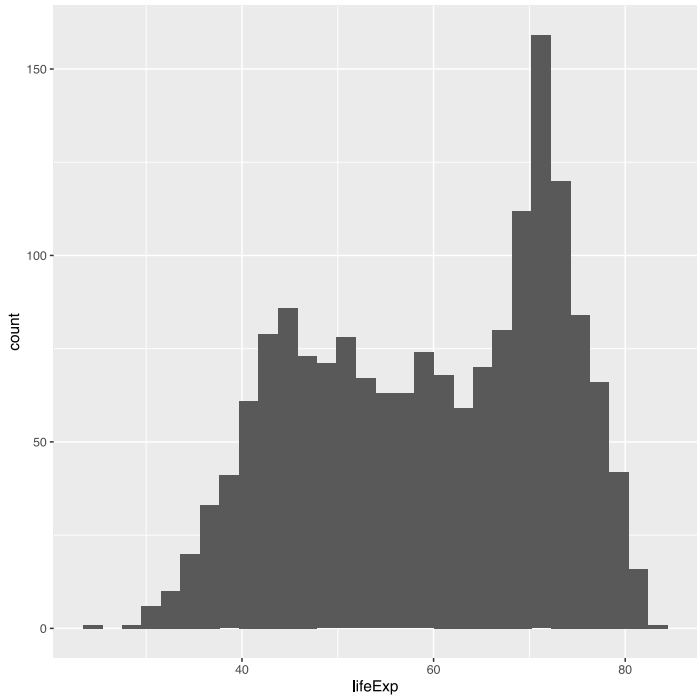
```
ggplot(gapminder,aes(x=lifeExp))
```



My Solution

2: Add Histogram Layer

```
ggplot(gapminder,aes(x=lifeExp))+
  geom_histogram(bins=30)
```

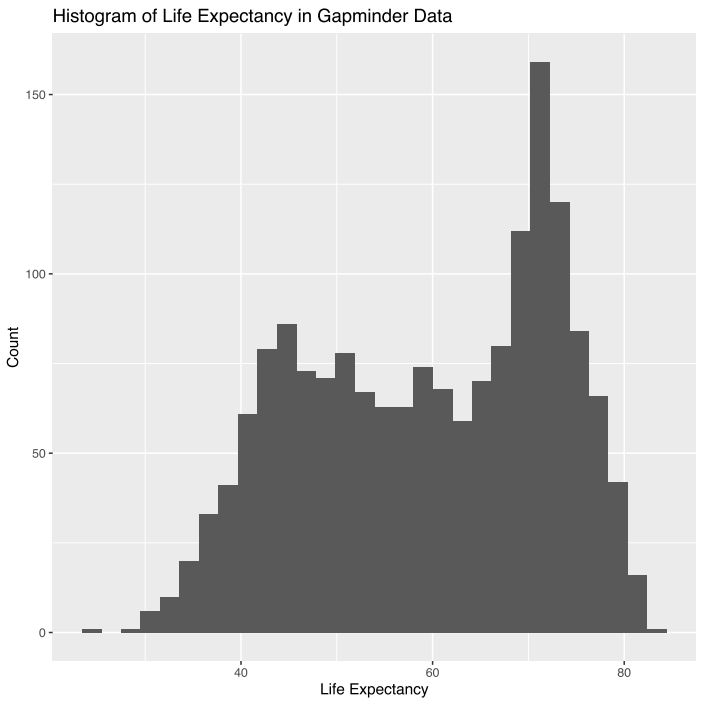


Setting the `bins` aesthetic removes a pesky message!

My Solution

3: Add Label Layers

```
ggplot(gapminder,aes(x=lifeExp))+
  geom_histogram(bins=30)+
  xlab("Life Expectancy")+
  ylab("Count")+
  ggtitle("Histogram of Life Expectancy
          in Gapminder Data")
```



Homework 2

Pick some relationship to look at in the Gapminder data and write up a .Rmd file investigating that question graphically. You might work with a subset of the data (e.g. just Africa). Upload both the `.Rmd` file and the `.html` file to Canvas.

- Include 4 to 8 plots.
- All titles, axes, and legends should be labelled clearly (no raw variable names).
- You must have at least one graph with `facet_wrap()` or `facet_grid()`.
- You must include at least one manually specified legend.
- You can use other `geoms` like histograms, bar charts, add vertical or horizontal lines, etc.

Your document should be pleasant for a peer to look at, with some organization. You must write up your observations in words as well as showing the graphs. Use chunk options like `echo=FALSE` to limit the code/output you show in the `.html`.