SOC6707: Intermediate Data Analysis

Monica Alexander

Week 4: Data Visualization

Announcements

- Assignment 1 extension until end of Wednesday

Data visualization principles

- Choose the right graph
- Know your audience
- Emphasize important patterns without being misleading
- Clear, effective designs

Choose the right graph

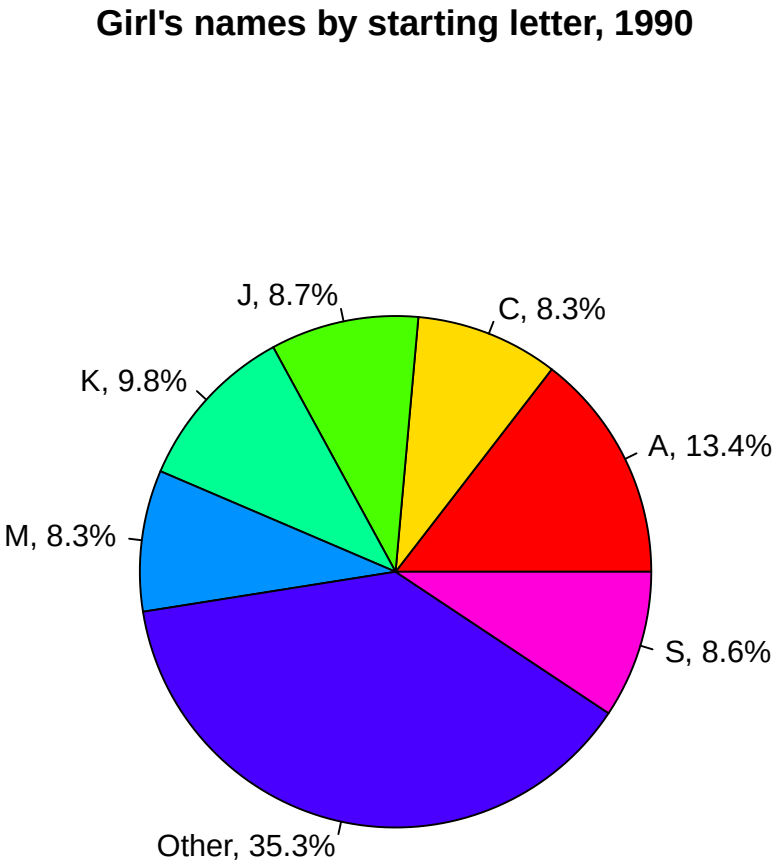
Choosing the right graph primarily depends on the type of variables that you are trying to visualize:

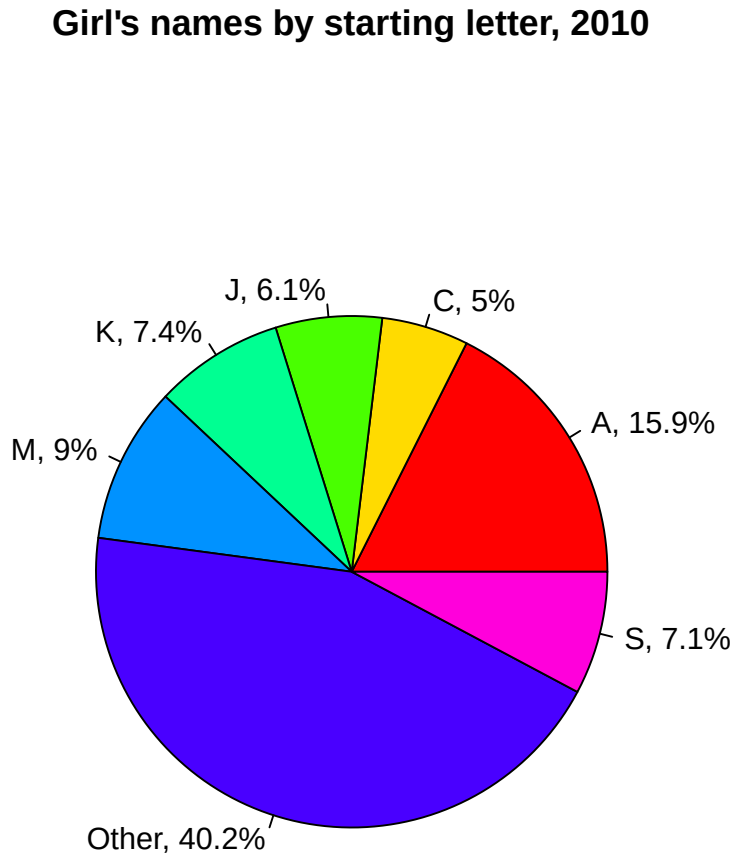
- Quantitative variables e.g. histograms, scatter plots
- Qualitative variables e.g. barcharts

Choose the graph based on the kind of data and the message to be conveyed.

- Do not use different graphs just for variety, as specific graphs convey certain types of information more effectively than others.
- If not required, do not use any chart — show only numbers.

Pie charts


Girl's names by starting letter, 1990
J, 8.7%
C, 8.3%
K, 9.8%
A, 13.4%
M, 8.3%
S, 8.6%
Other, 35.3%

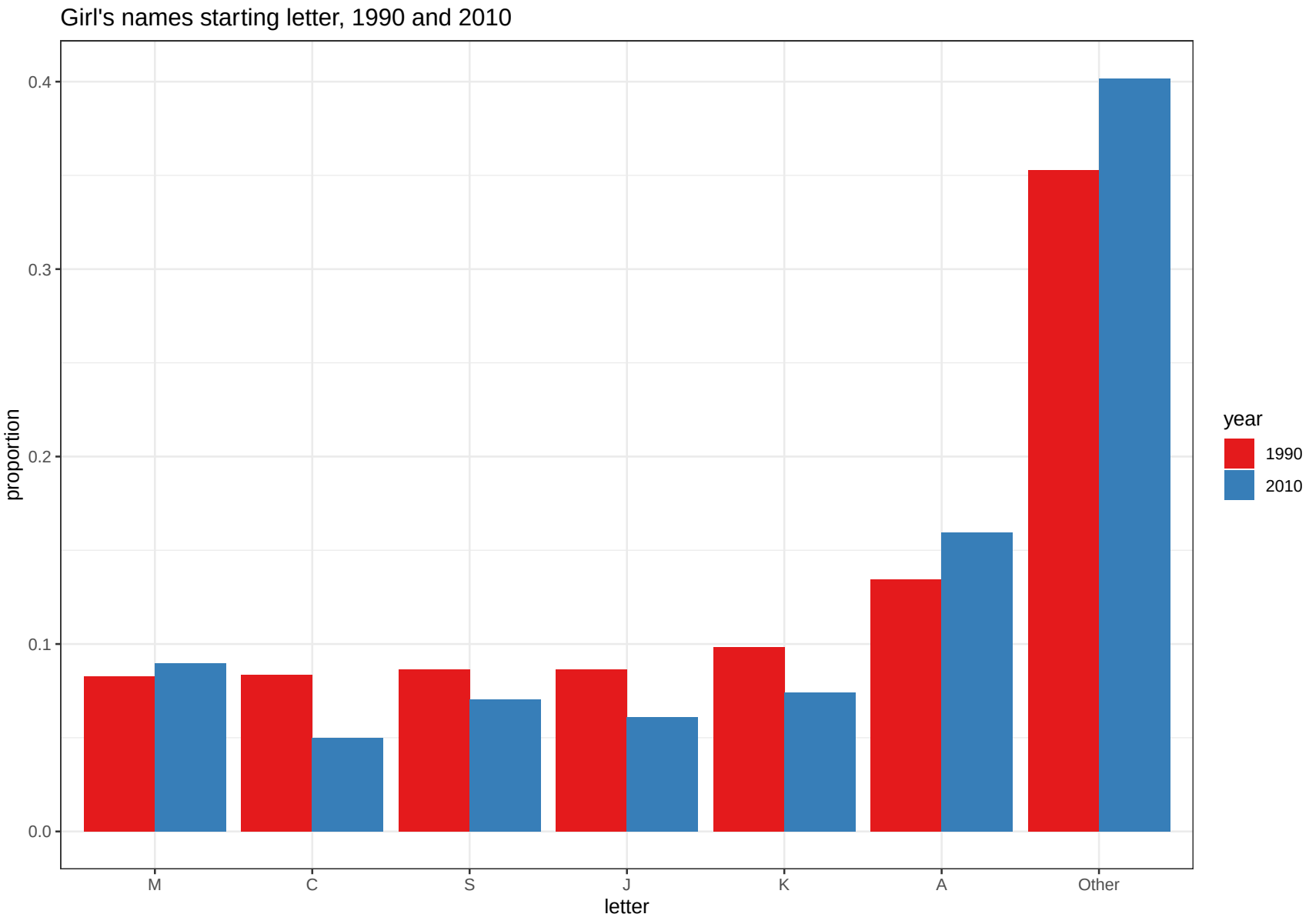

Girl's names by starting letter, 2010
J, 6.1%
C, 5%
K, 7.4%
A, 15.9%
M, 9%
S, 7.1%
Other, 40.2%

Pie charts

```
?pie
```

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

Alternative



Know your audience

Graphs can be used for

- our own exploratory data analysis
- to convey a message to experts,
- to help tell a story to a general audience.

Make sure that the intended audience understands each element of the plot.

Examples: spiral plot, log scales

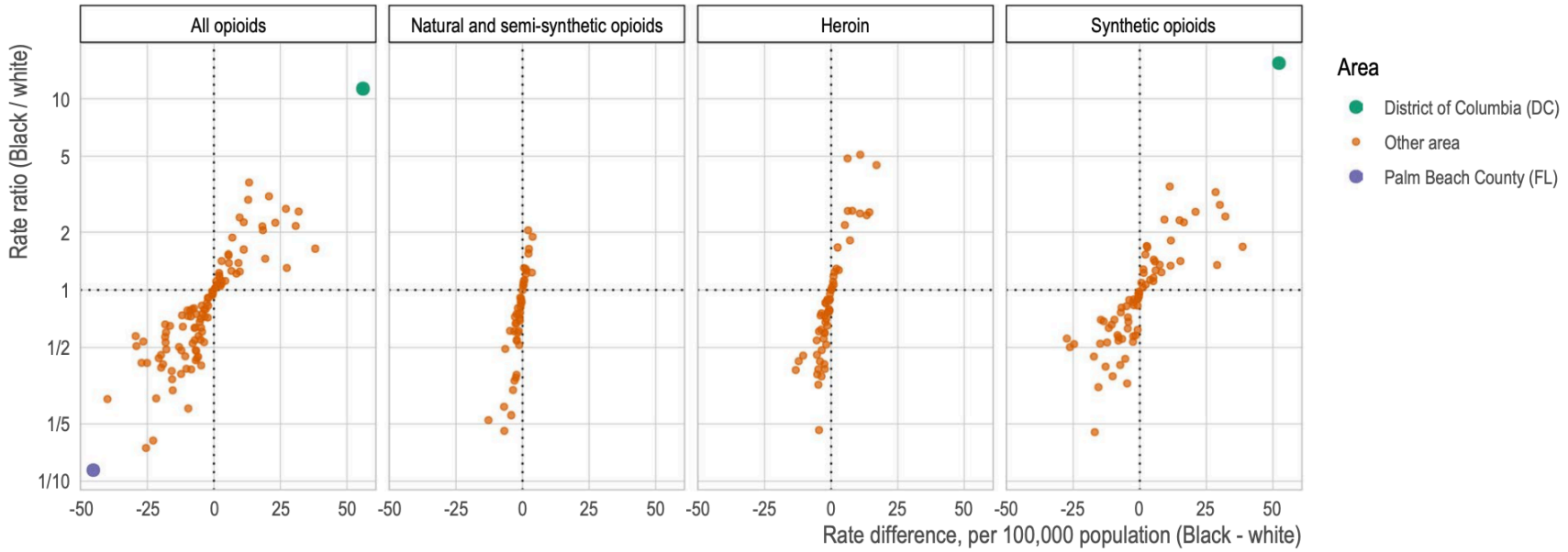
- Think of the color blind. In R, `viridis` and `brewer` palettes give colorblind-friendly options

Emphasize important patterns without being misleading

There is no such thing as information overload. There is only bad design. — Edward Tufte

- Eliminate distractions
- Highlight the essential
- Use color and text strategically
- Avoid pseudo-3D plots

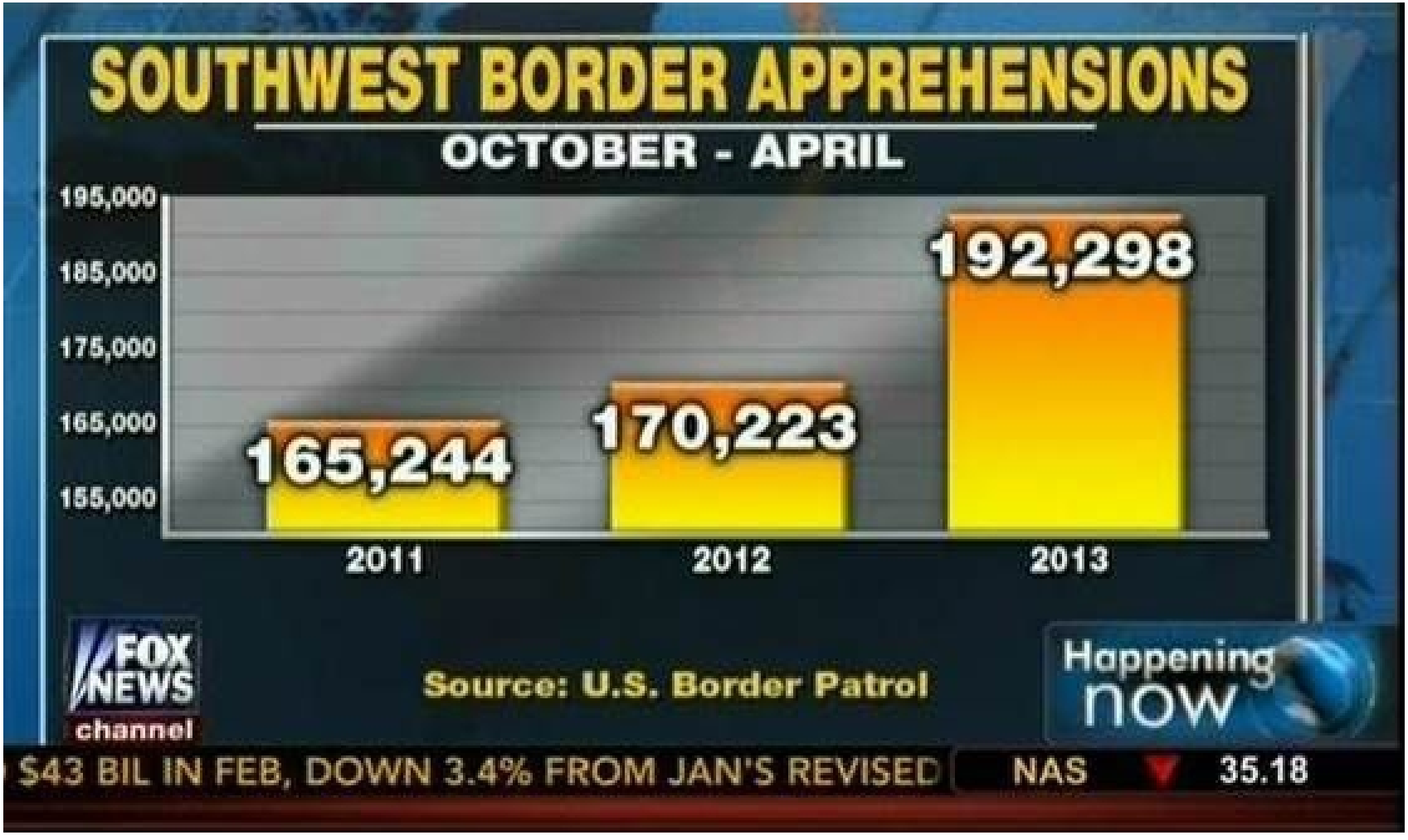
Highlight the essential



Source:
https://link.springer.com/article/10.1007/s11524-021-00573-8

When to start the axis at zero?

When to start the axis at zero?



Source

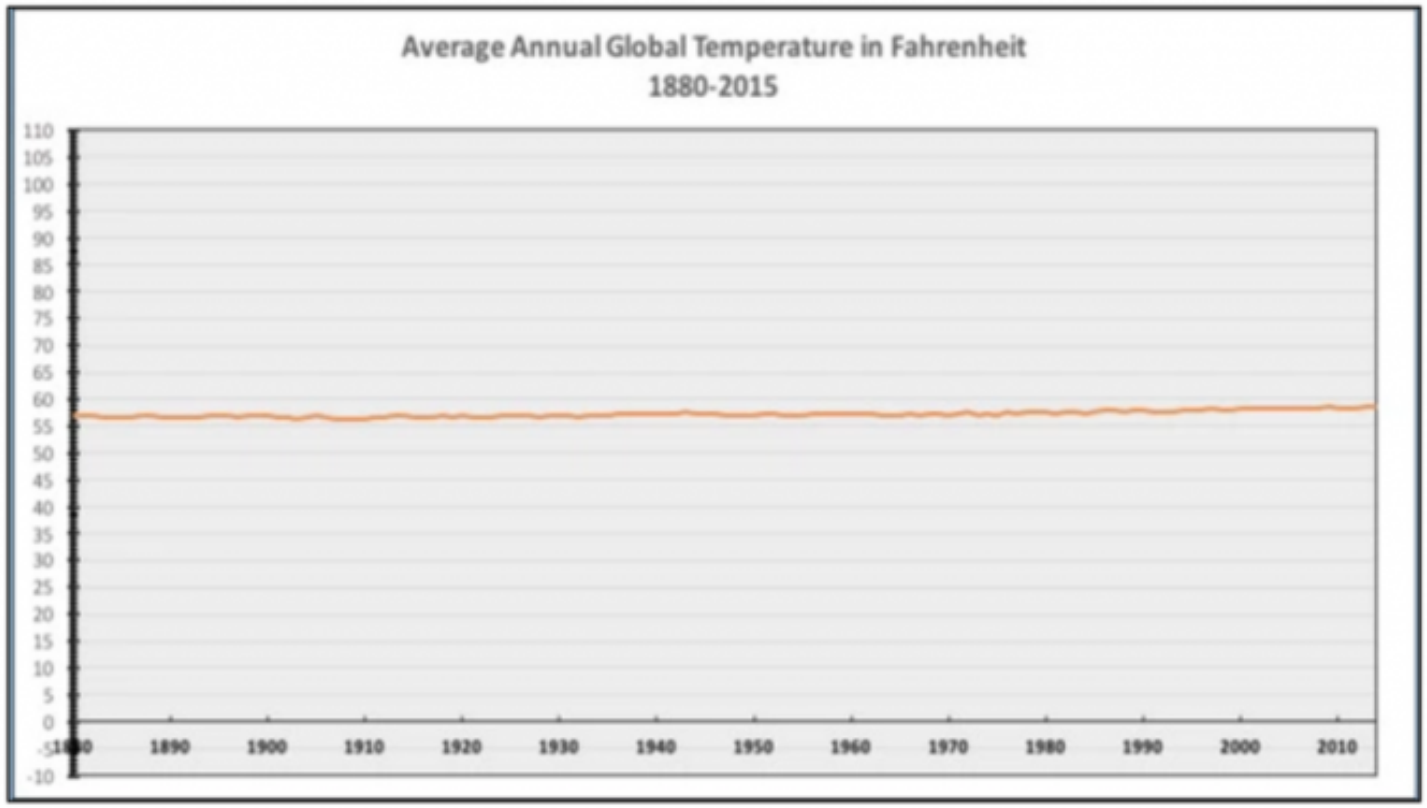
When to start the axis at zero?

National Review ✓
@NRO

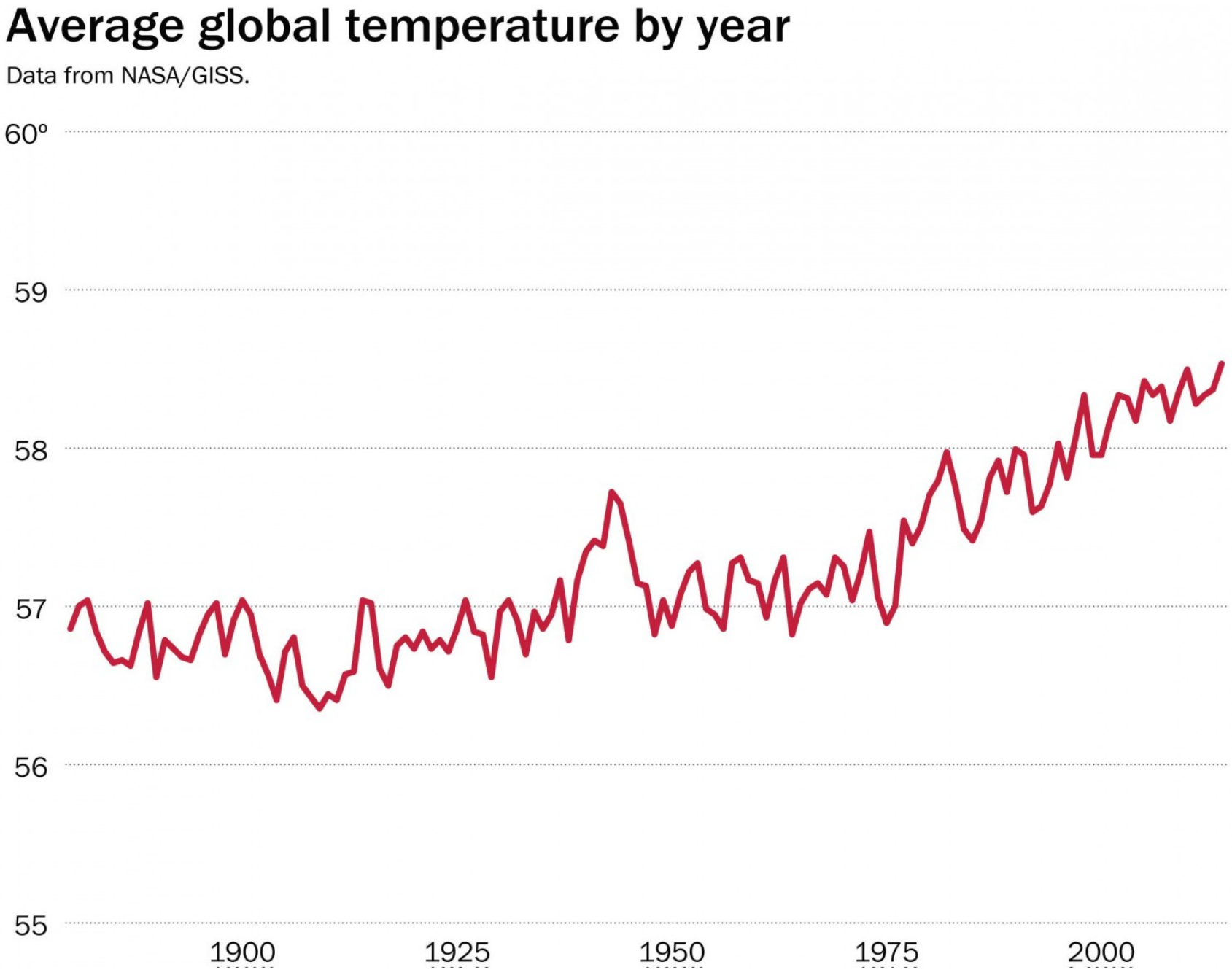
Follow

The only #climatechange chart you need to see. natl.re/wPKpro

(h/t @powerlineUS)

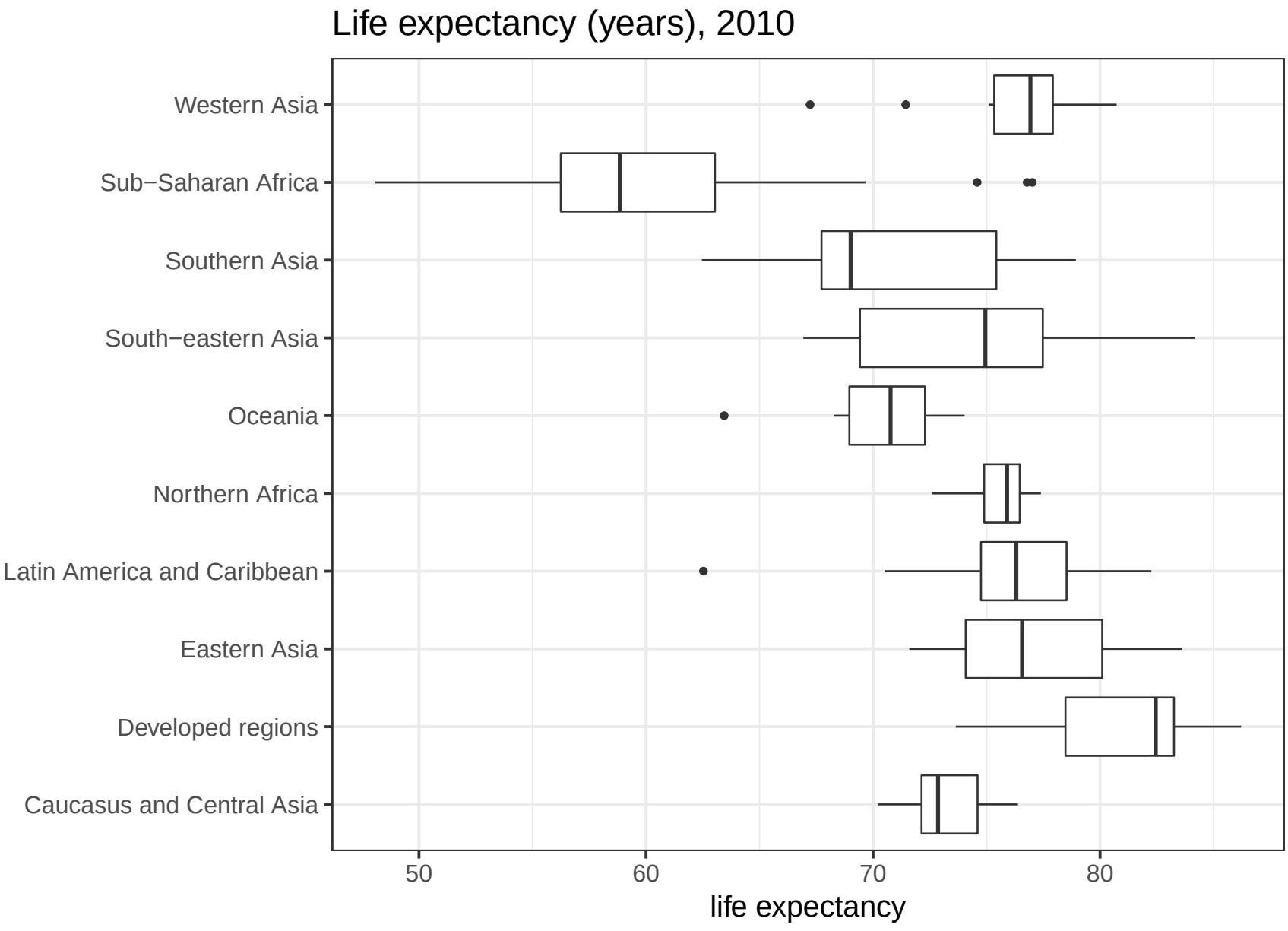


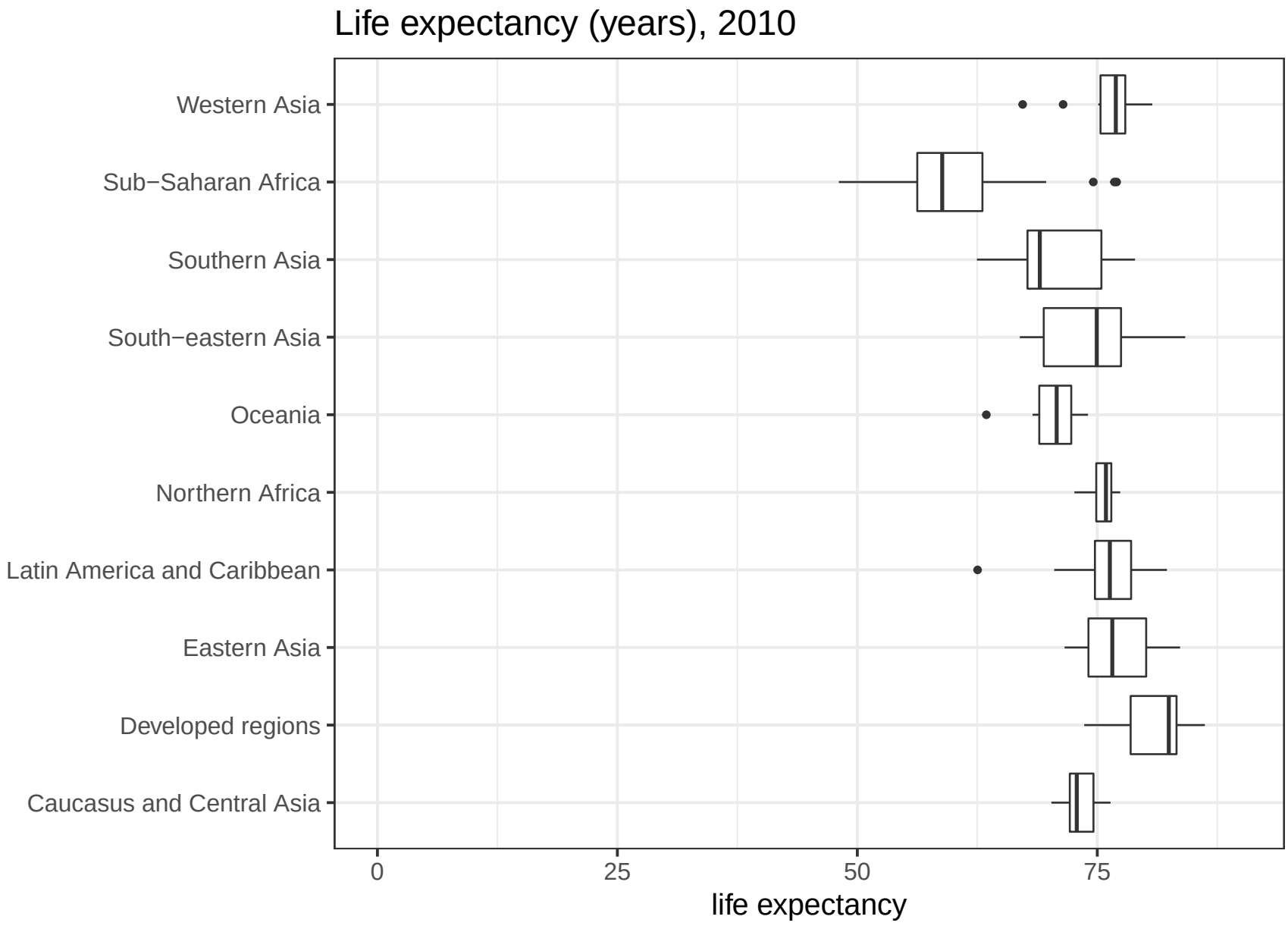
When to start the axis at zero?



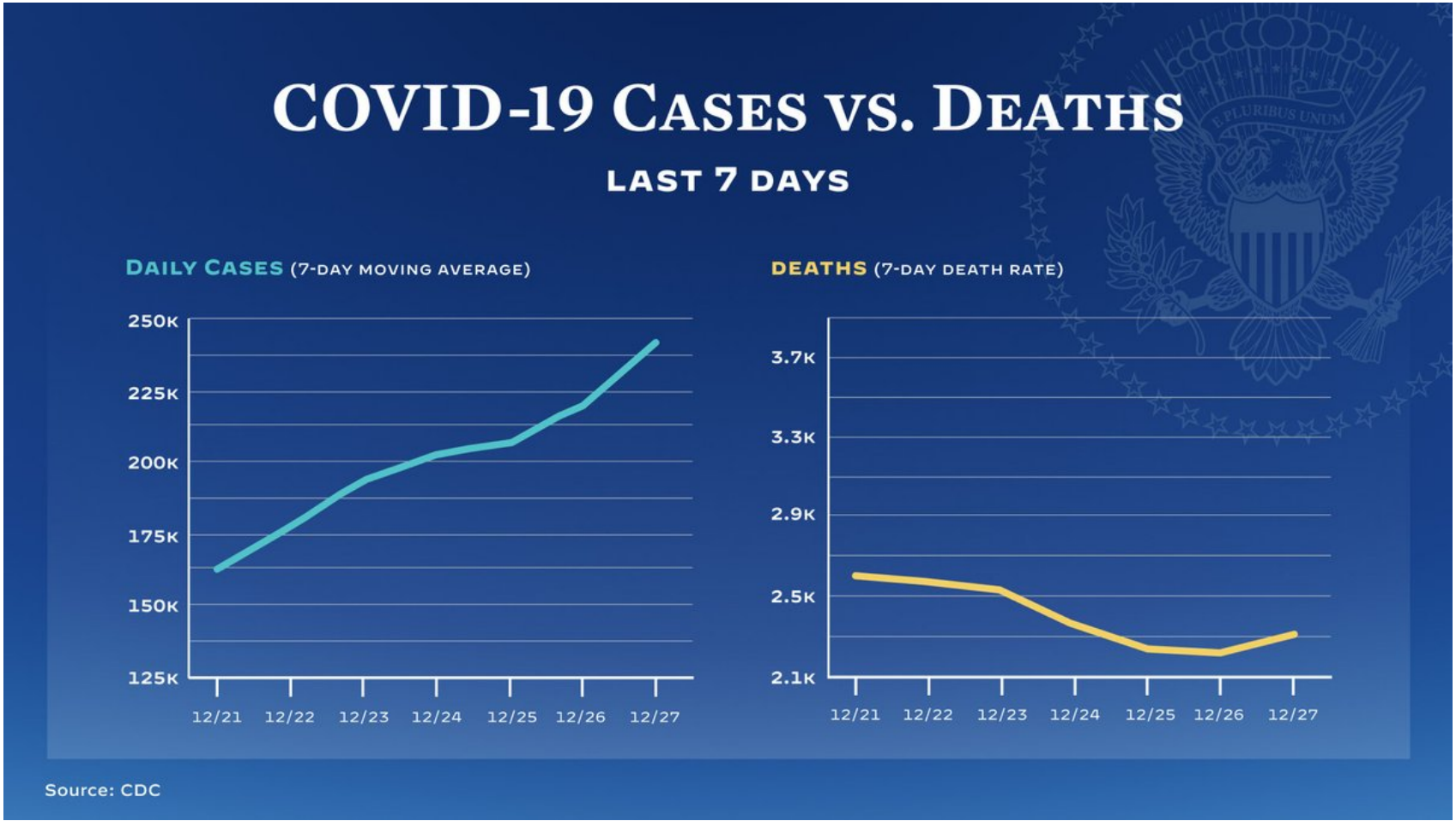
When to include zeroes

- With bar plots, we are implying the length is proportional to the quantities being displayed. By avoiding 0, relatively small differences can be made to look much bigger than they actually are.
- With line plots or plots that use position, it is not neccessary to start the axis at zero (and could be misleading)

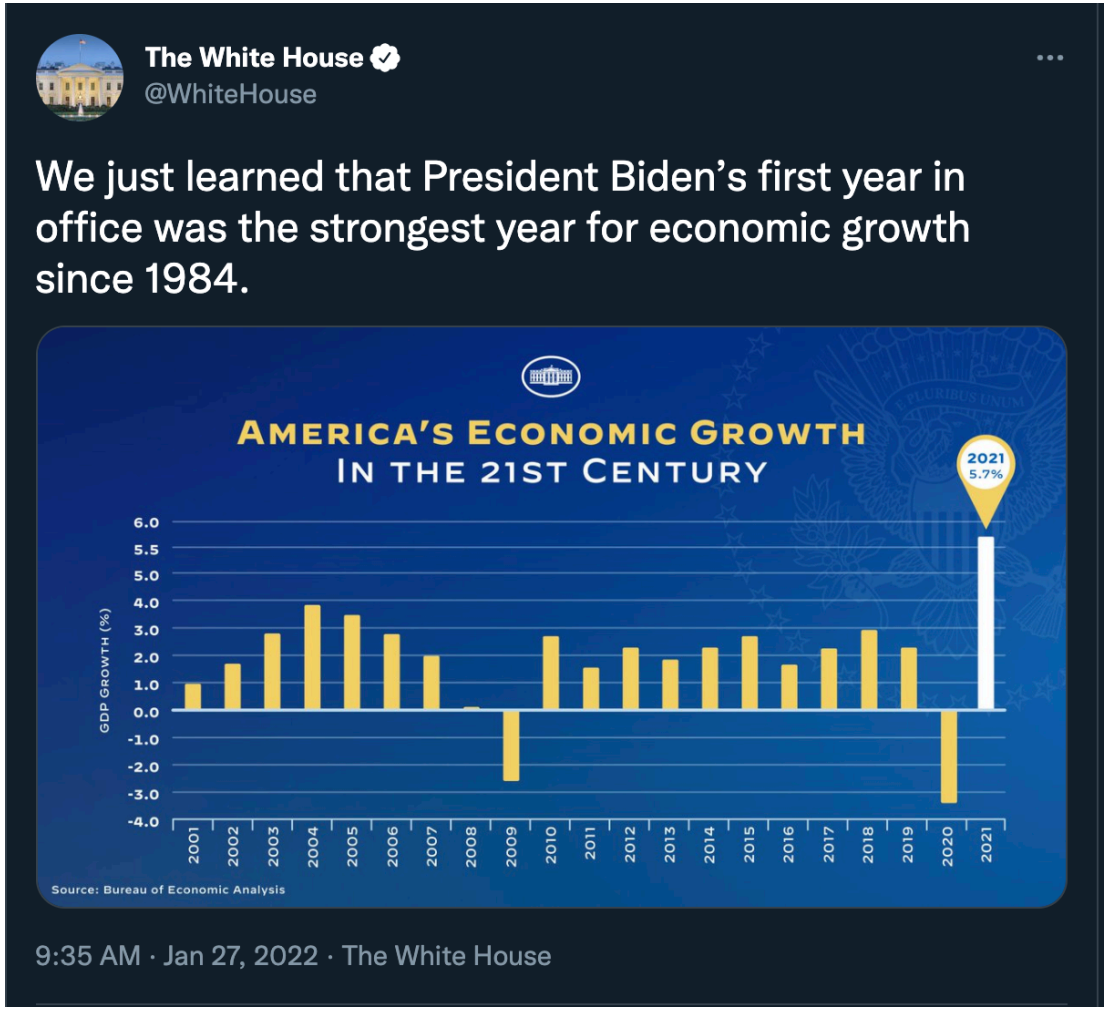

Life expectancy (years), 2010
Western Asia
Sub−Saharan Africa
Southern Asia
South−eastern Asia
Oceania
Northern Africa
Latin America and Caribbean
Eastern Asia
Developed regions
Caucasus and Central Asia
50
60
70
80
life expectancy


Life expectancy (years), 2010
Western Asia
Sub−Saharan Africa
Southern Asia
South−eastern Asia
Oceania
Northern Africa
Latin America and Caribbean
Eastern Asia
Developed regions
Caucasus and Central Asia
0
25
50
75
life expectancy

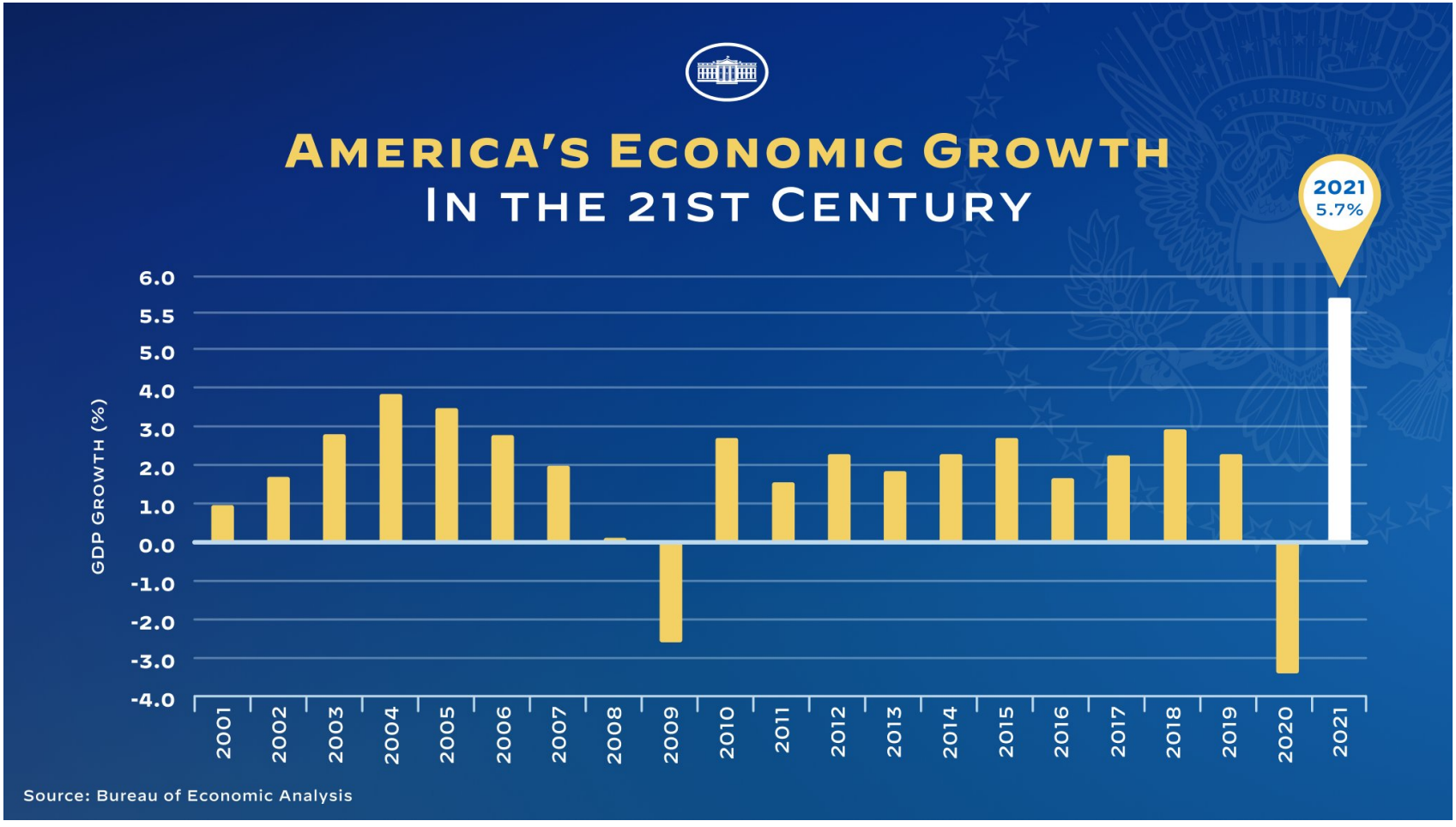
Emphasize important patterns without being misleading



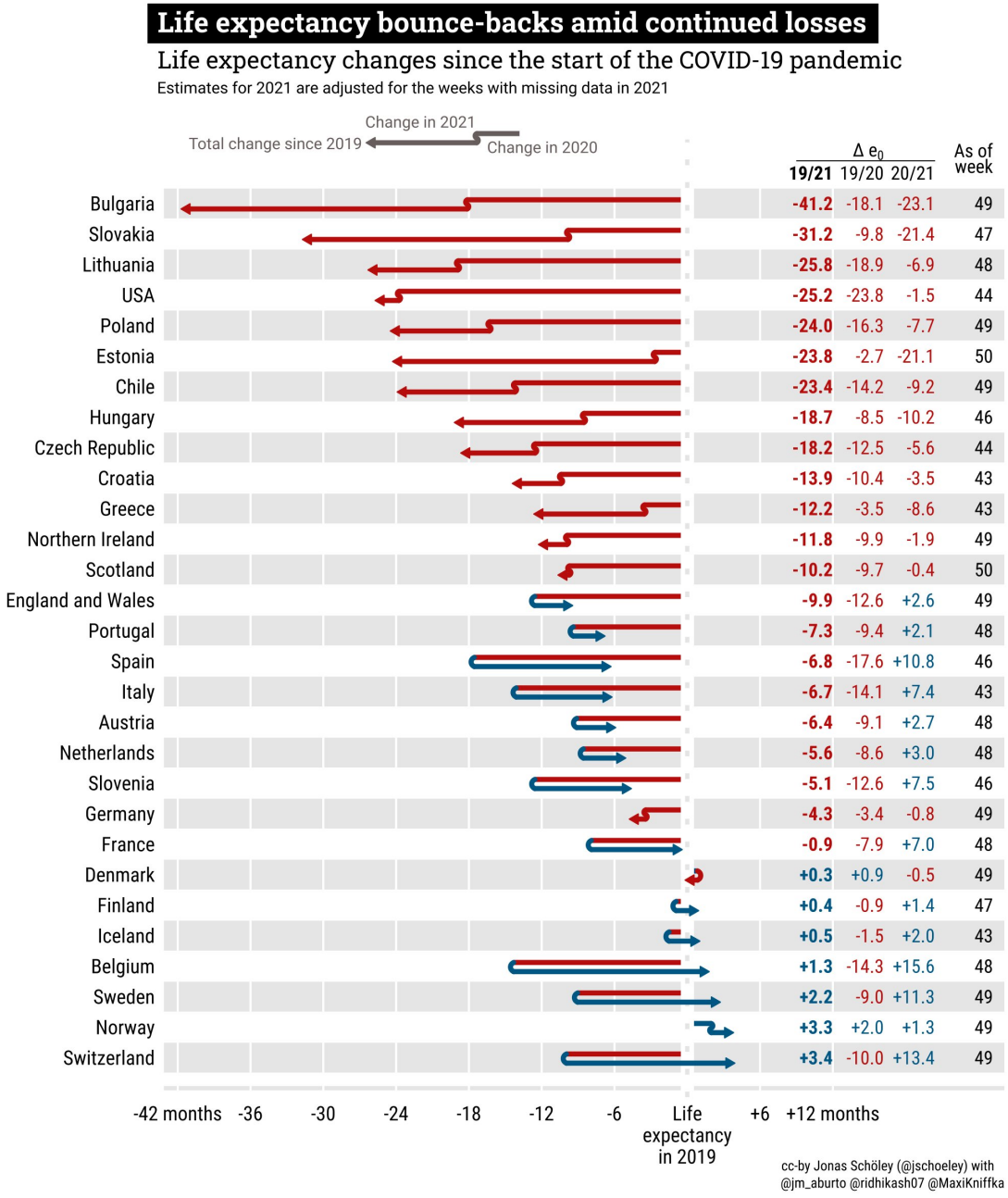
Emphasize important patterns without being misleading



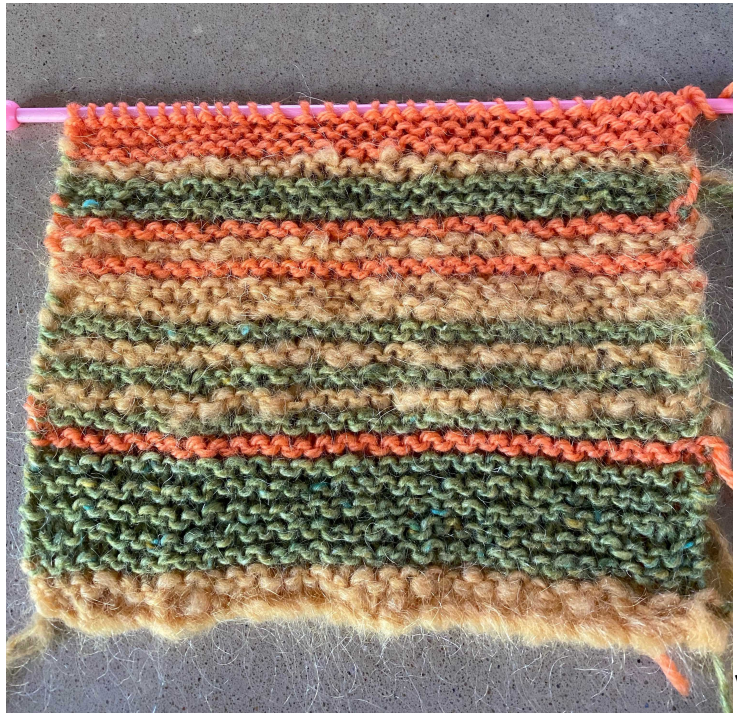
Emphasize important patterns without being misleading



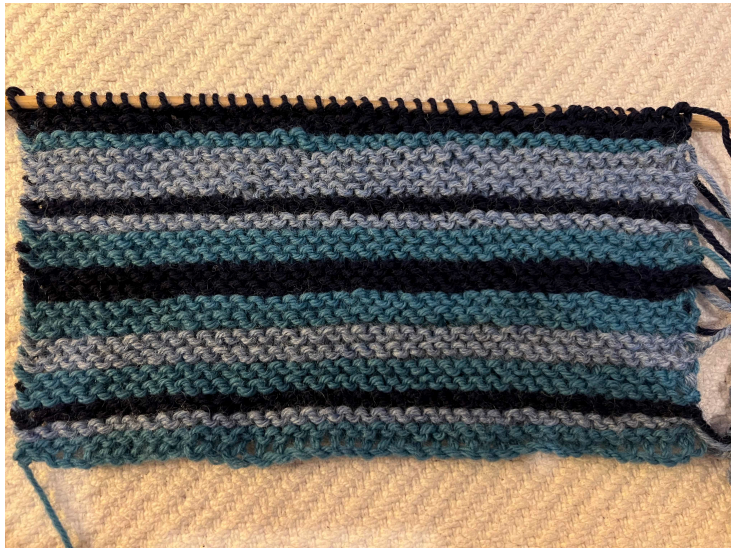
Clear, effective designs



...but data visualization need not be a graph



Hobart, Tasmania, Australia



Toronto, Ontario, Canada

Important types of graphs

Important types of graphs

- Histograms
- Bar charts
- Boxplots
- Line plots
- Scatter plots

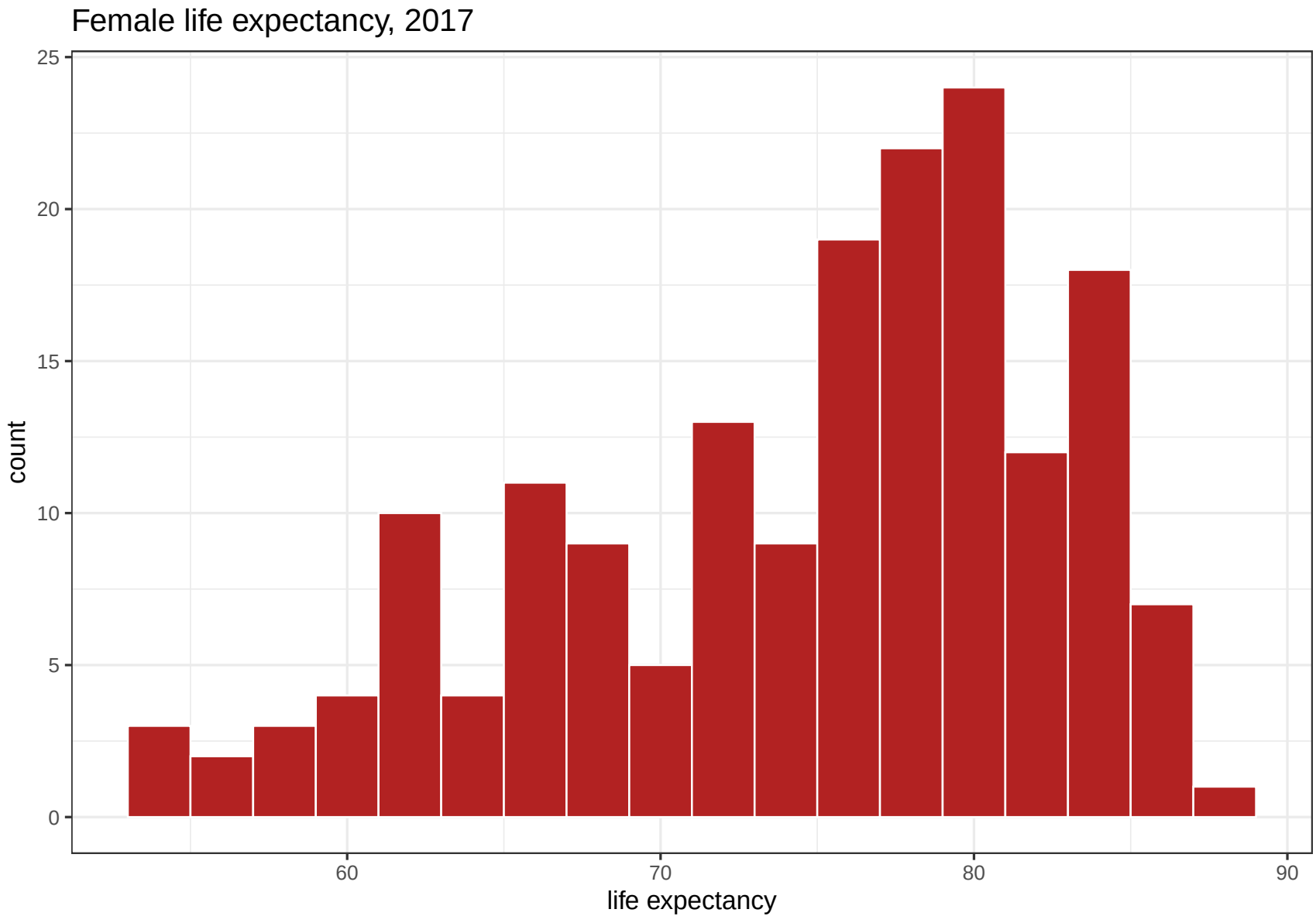
Example datasets used here

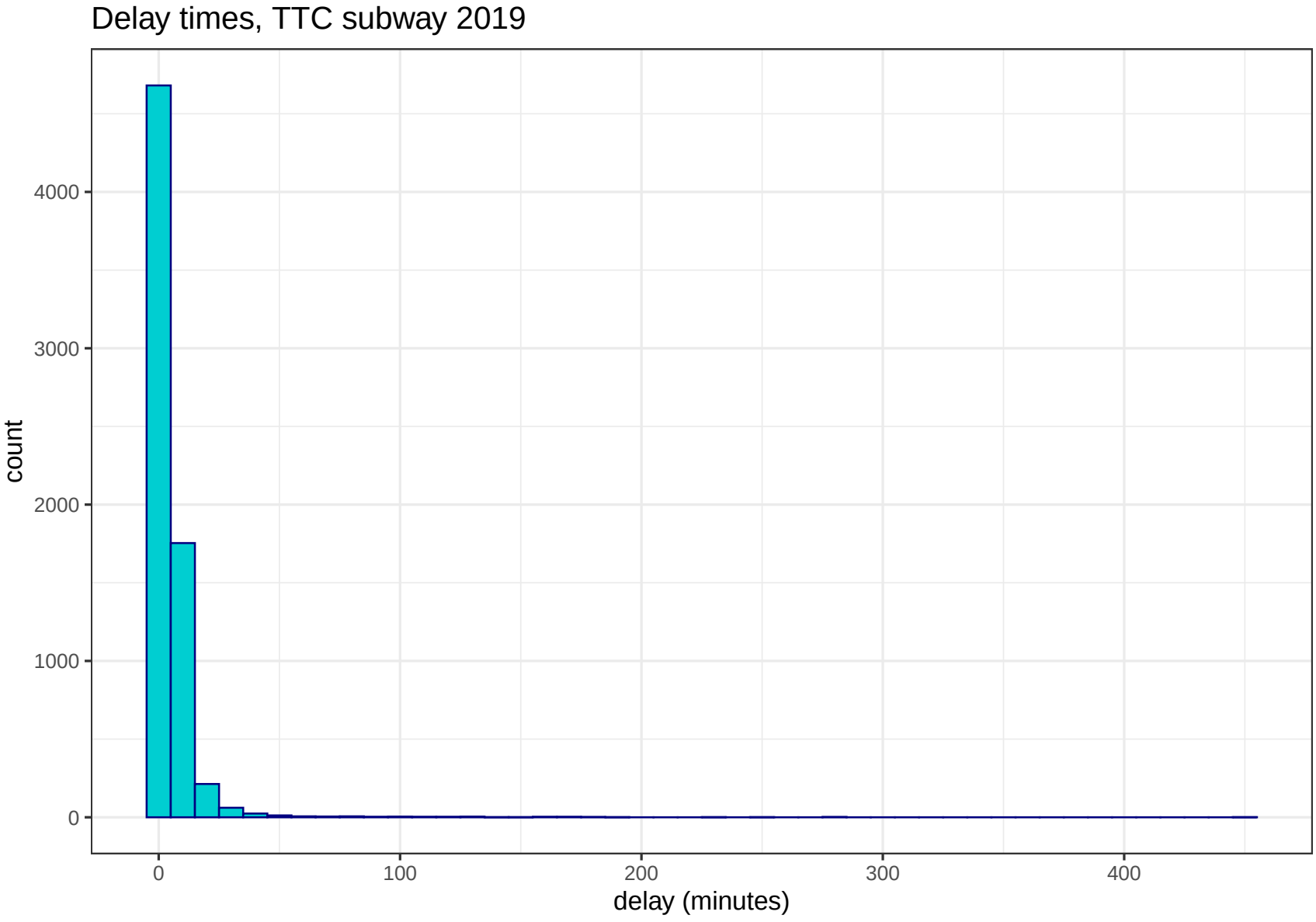
1. TTC subway delays (from above)
2. Country-level indicators, 2009-2017
   - Uploaded onto Quercus
   - TFR = total fertility rate
   - GDP = gross domestics product
   - dataset also has life expectancy (females), child mortality, maternal mortality

Histograms

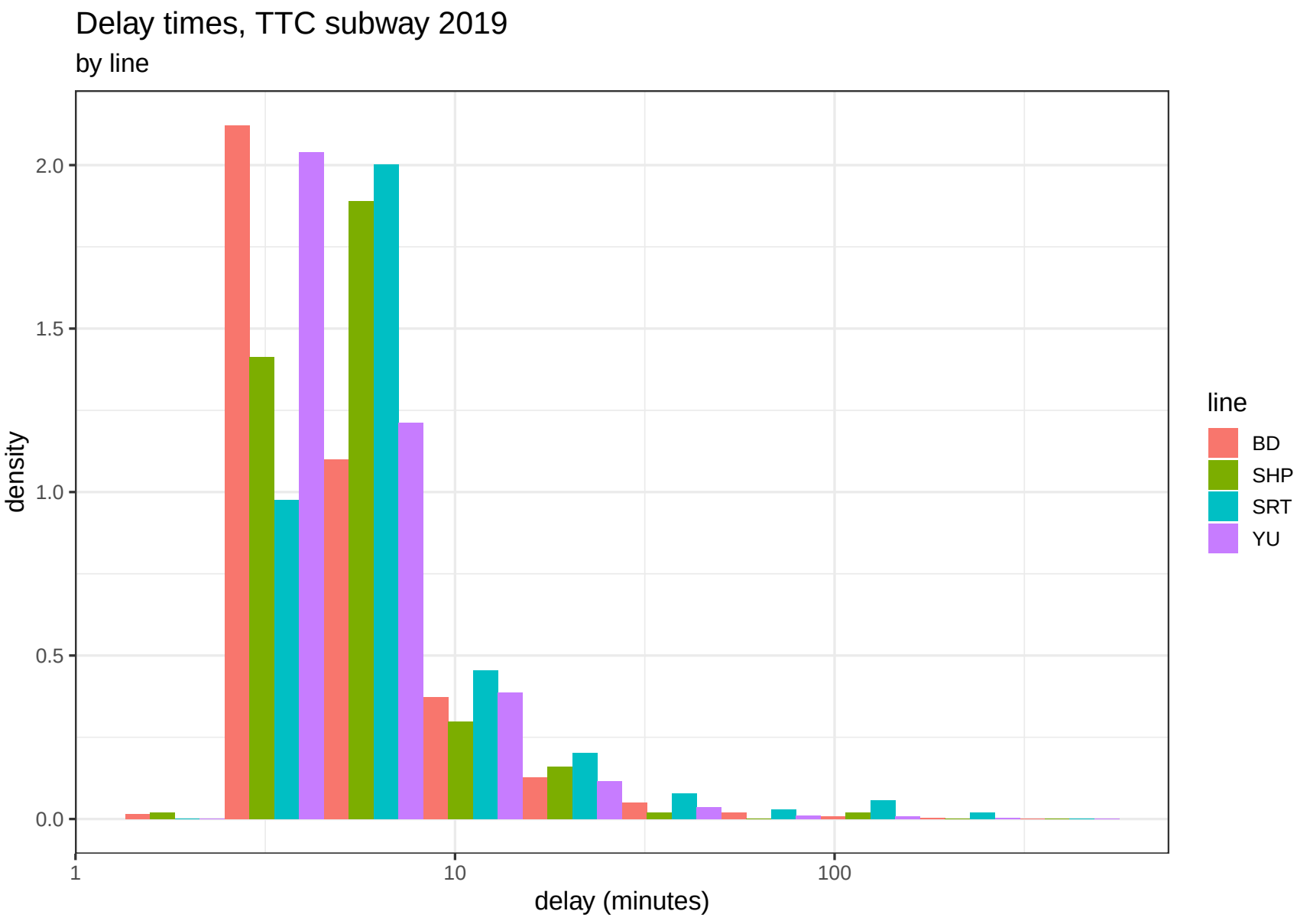
Shows the distribution of a **quantitative** variable

- Histograms show the frequency (count) of observations by value
- The range of values of a variables is divided into intervals ('bins') and then the number of observations in each bin is tabulated
- A histogram shows the count of observations in each bin with a rectangle of height equal to the count
- The x axis is the value bins, the y axis is the count/frequency (or proportion)


Female life expectancy, 2017
25
20
15
10
5
0
count
60
70
80
90
life expectancy


Delay times, TTC subway 2019
count
4000
3000
2000
1000
0
0
100
200
300
400
delay (minutes)

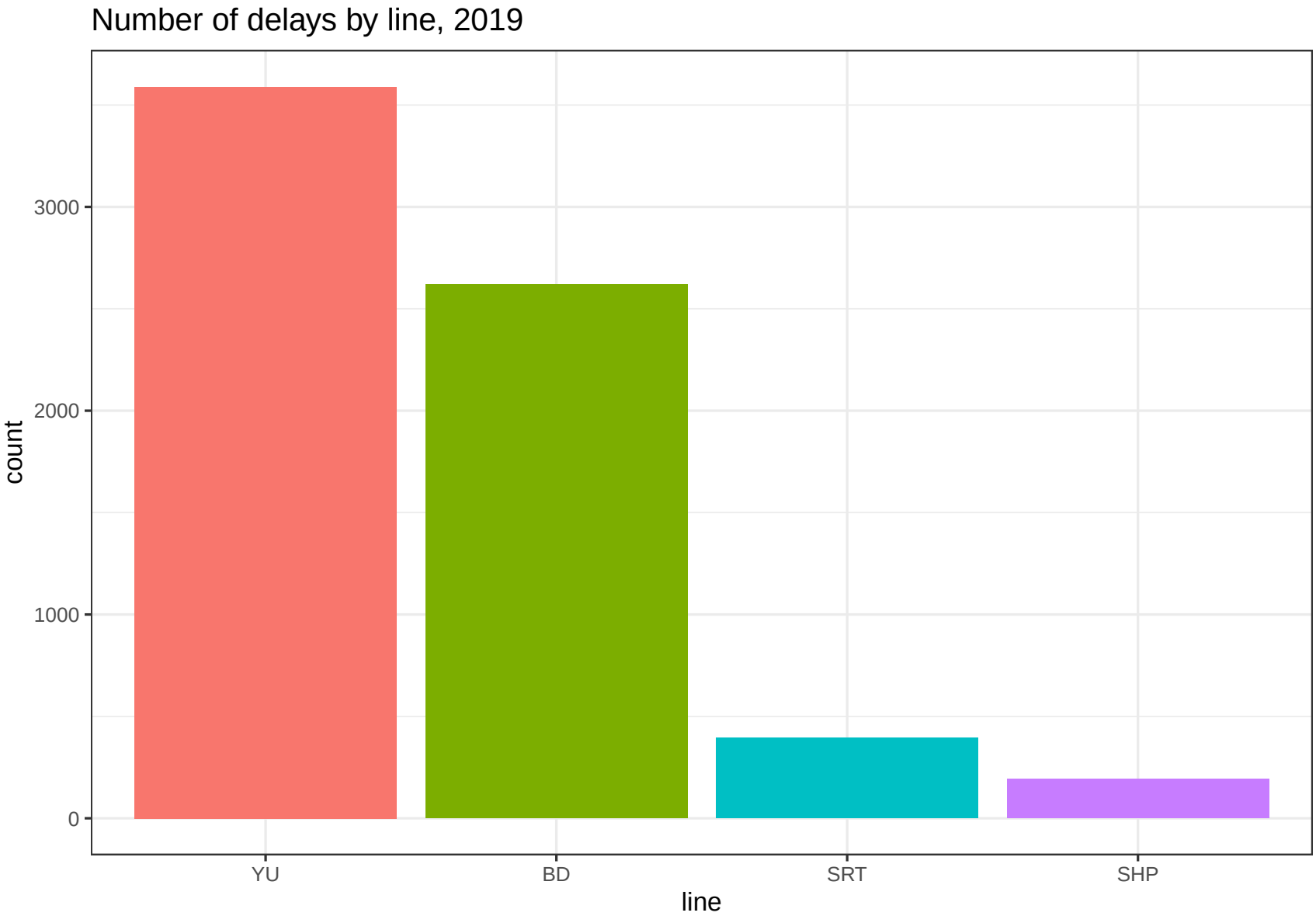
Making the histogram more informative



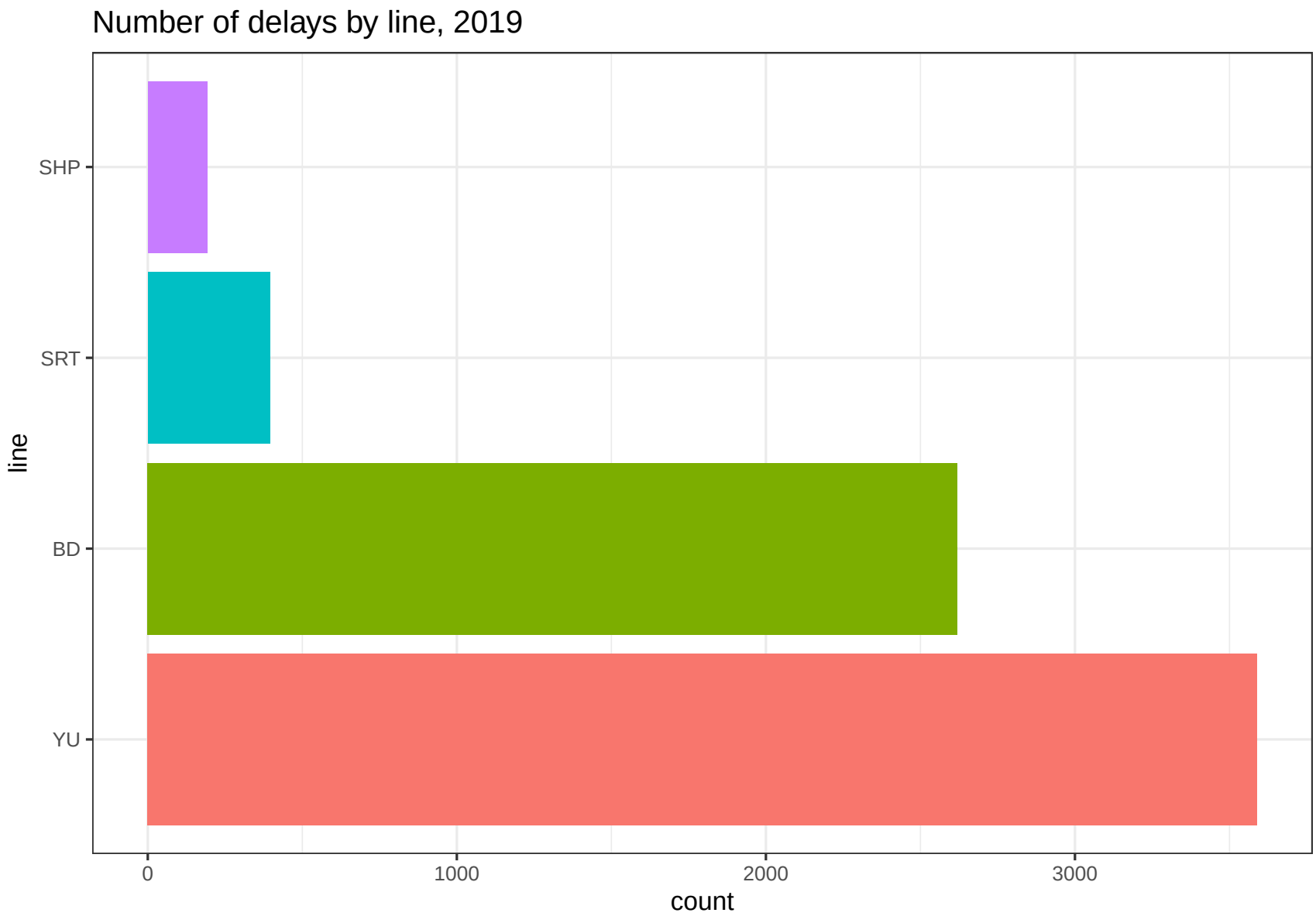
Bar charts

Shows summary measures across values of a **categorical** (qualitative) variable

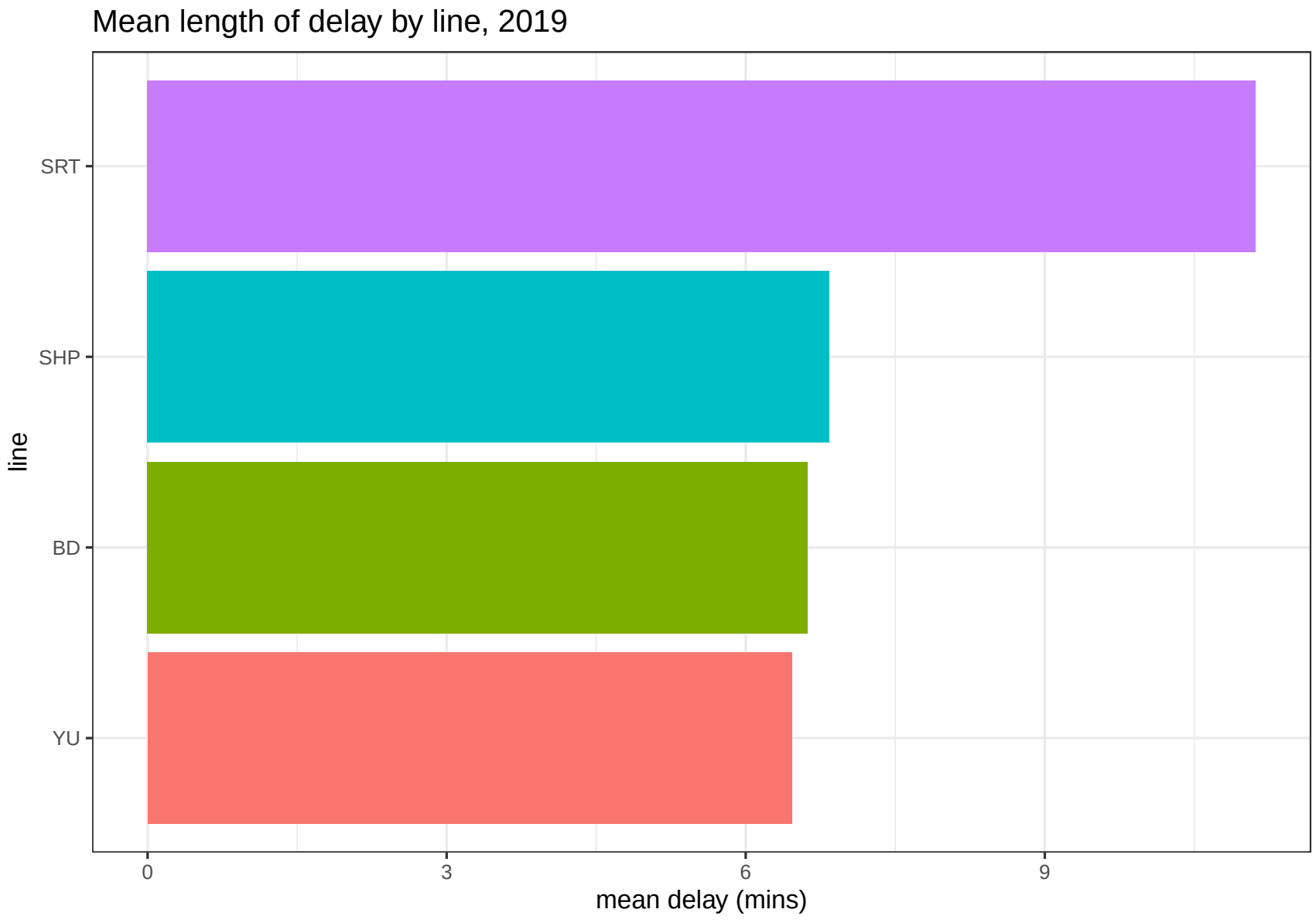
- Illustrate the value of a particular outcome in a particular category
- The 'value' can be counts, but could also be a summary measure (e.g. mean)
- The value is again shown by a rectangle of height equal to the value
- Bar carts can be plotted vertically or horizontally
- In the vertical setting, the x axis is the categories and the y axis is the value of the quantitative variable


Number of delays by line, 2019
count
3000
2000
1000
0
YU
BD
SRT
SHP
line

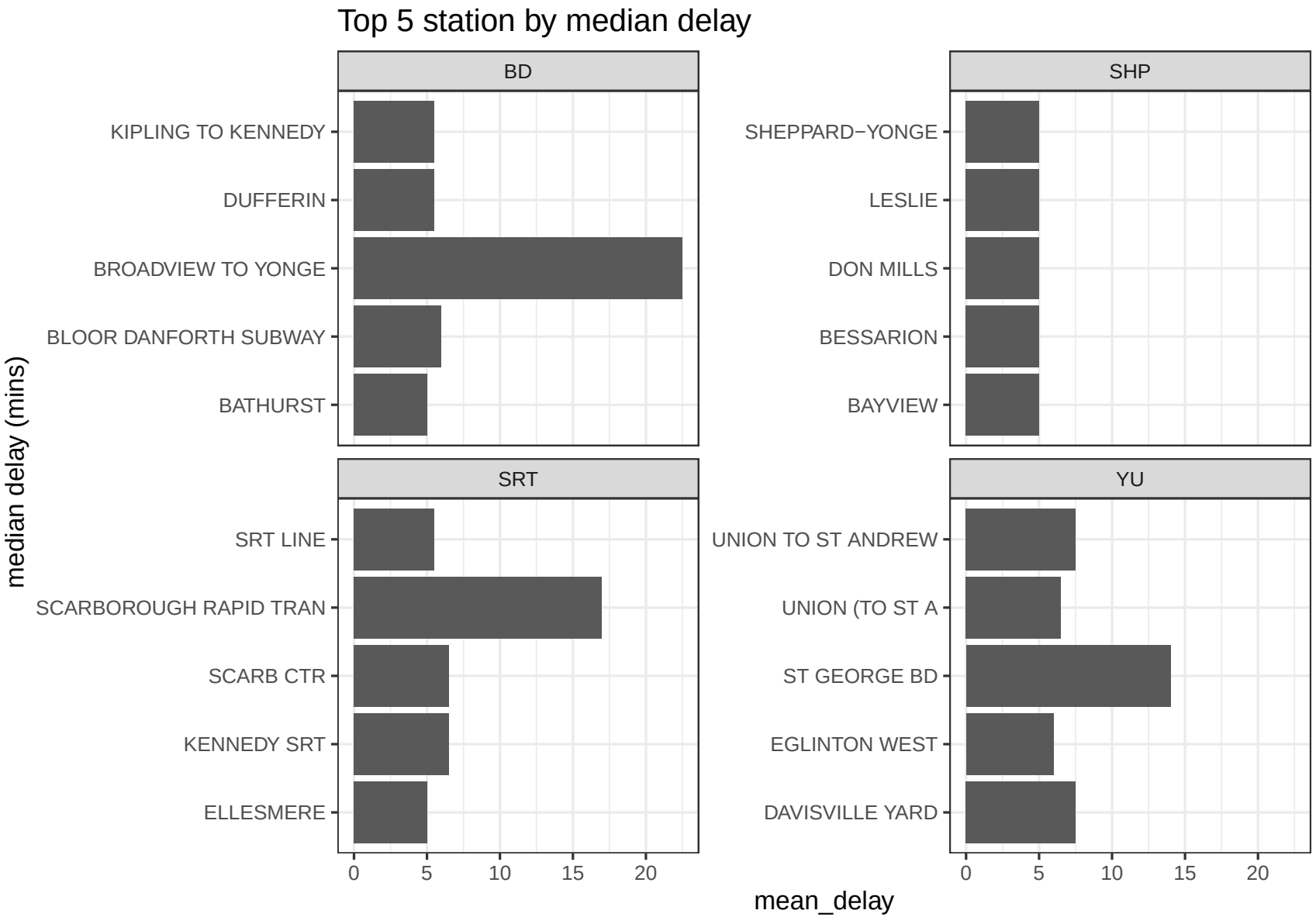
Same but horizontal



Showing mean delay time



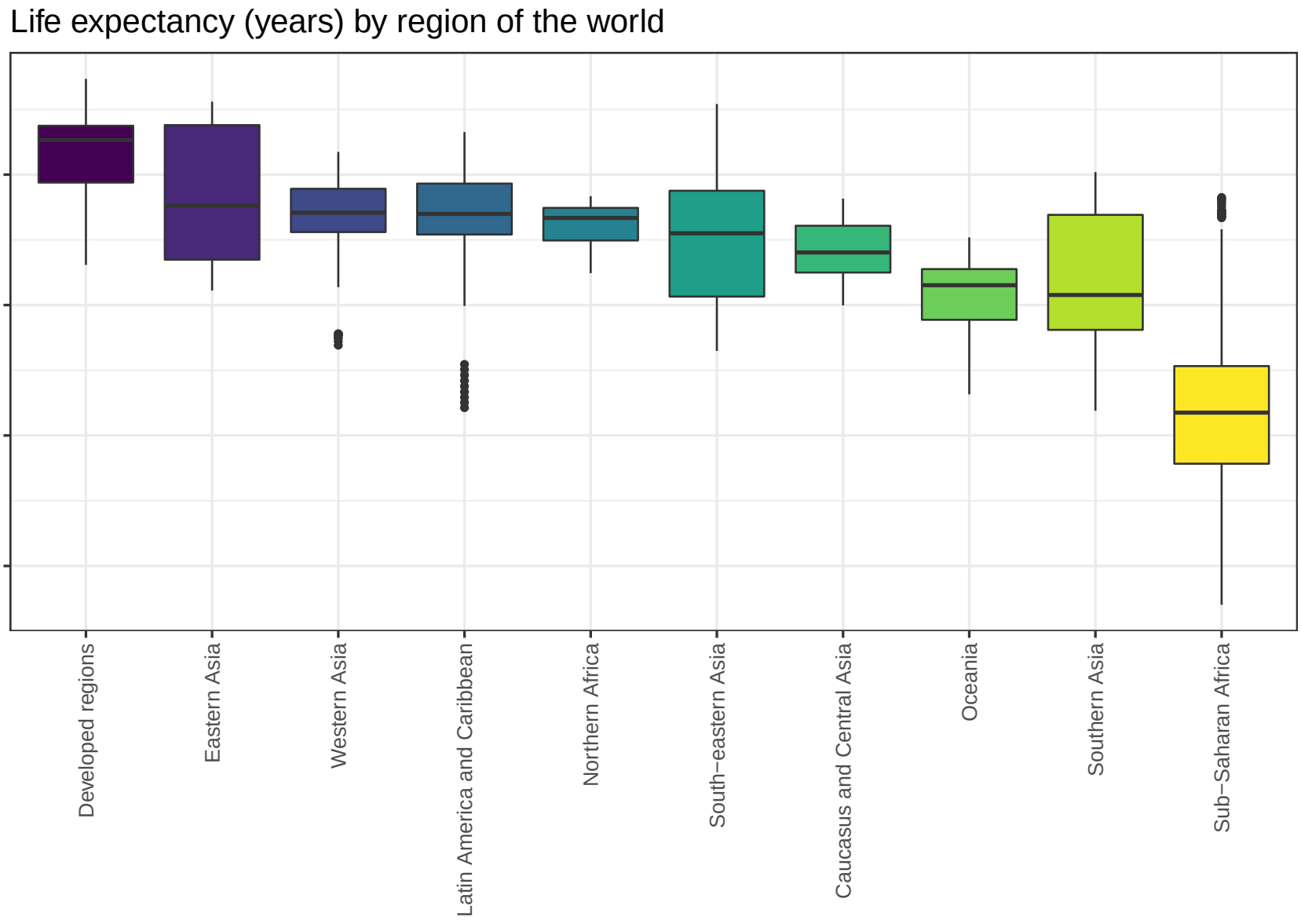
More complicated example



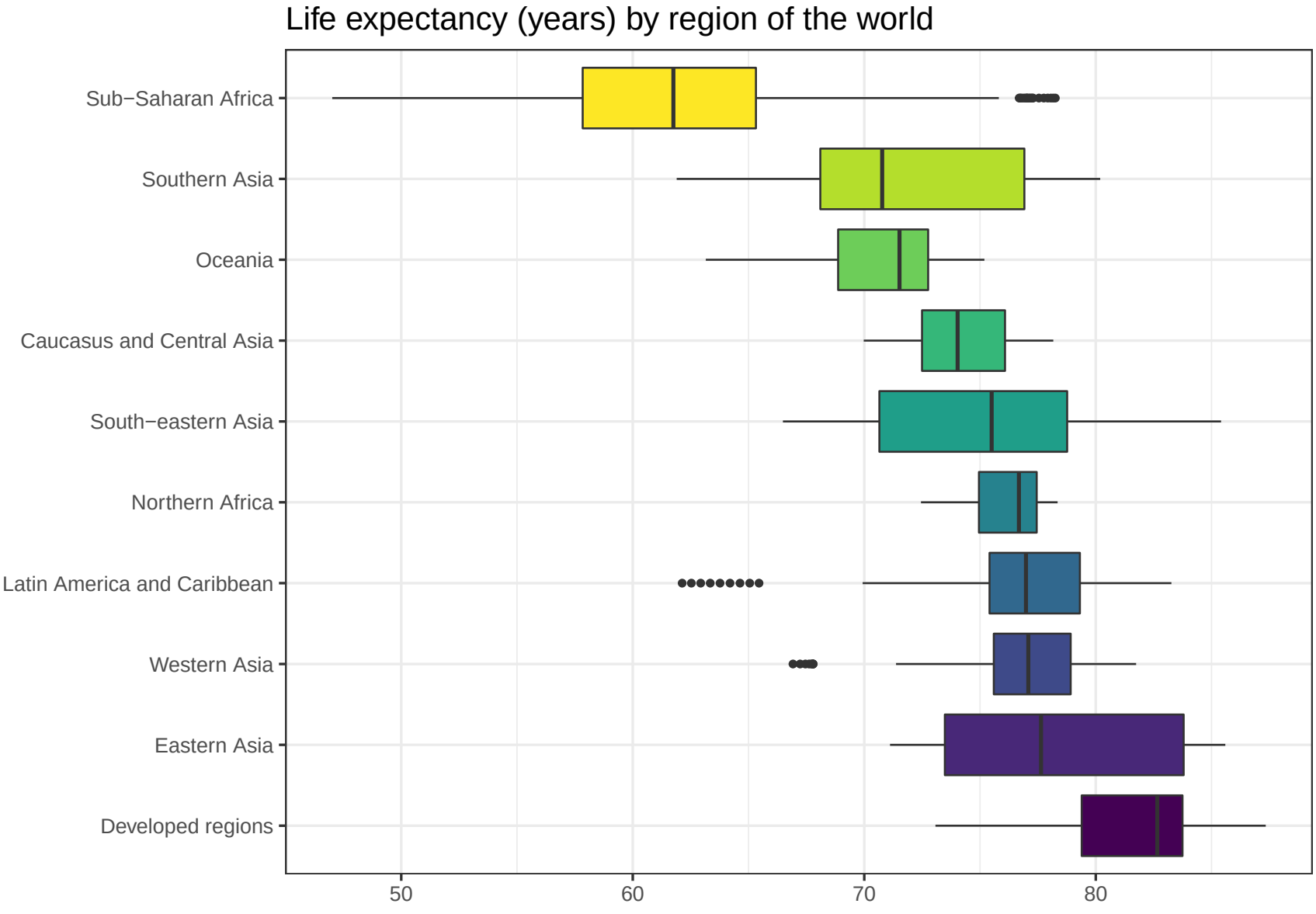
Box plots

Good for showing summaries of **quantitative** variables across different **categorical** groups.

- Visualizing quartiles (25/50/75 percentiles) of quantitative data
- Boxes show the IQR and median
- Whiskers show values outside IQR (in R/`ggplot`, default is 1.5*IQR)
- Outliers may be shown with individual dots
- In the vertical case, the x axis is the categories and the y axis is the quantitative variable


Life expectancy (years) by region of the world
80
70
60
50
Developed regions
Eastern Asia
Western Asia
Latin America and Caribbean
Northern Africa
South−eastern Asia
Caucasus and Central Asia
Oceania
Southern Asia
Sub−Saharan Africa

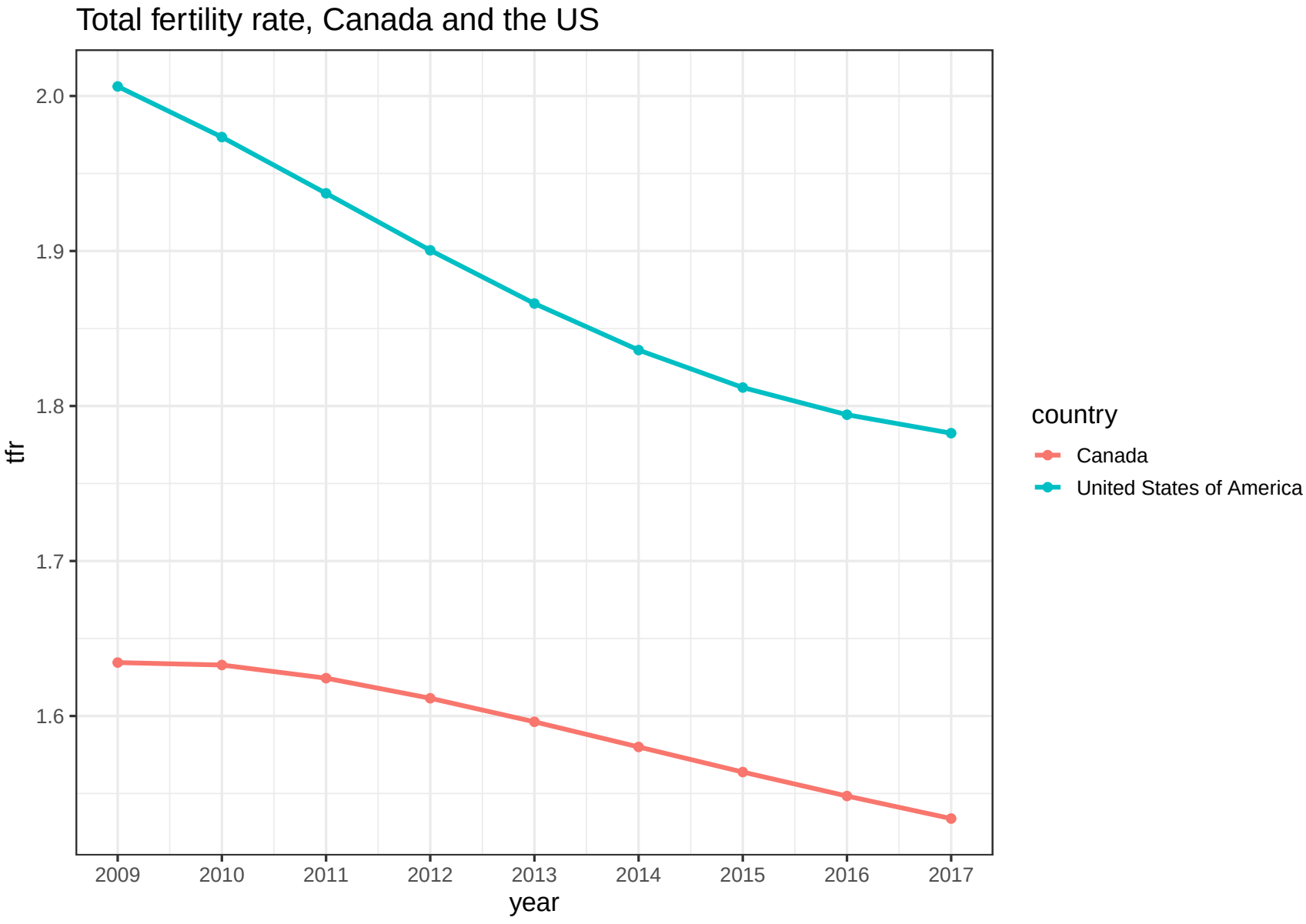
Could also do horizontal



Line plots

Best used to describe values of a **quantitative** variable (on y axis) across sequential values of another **quantitative** variable on the x axis

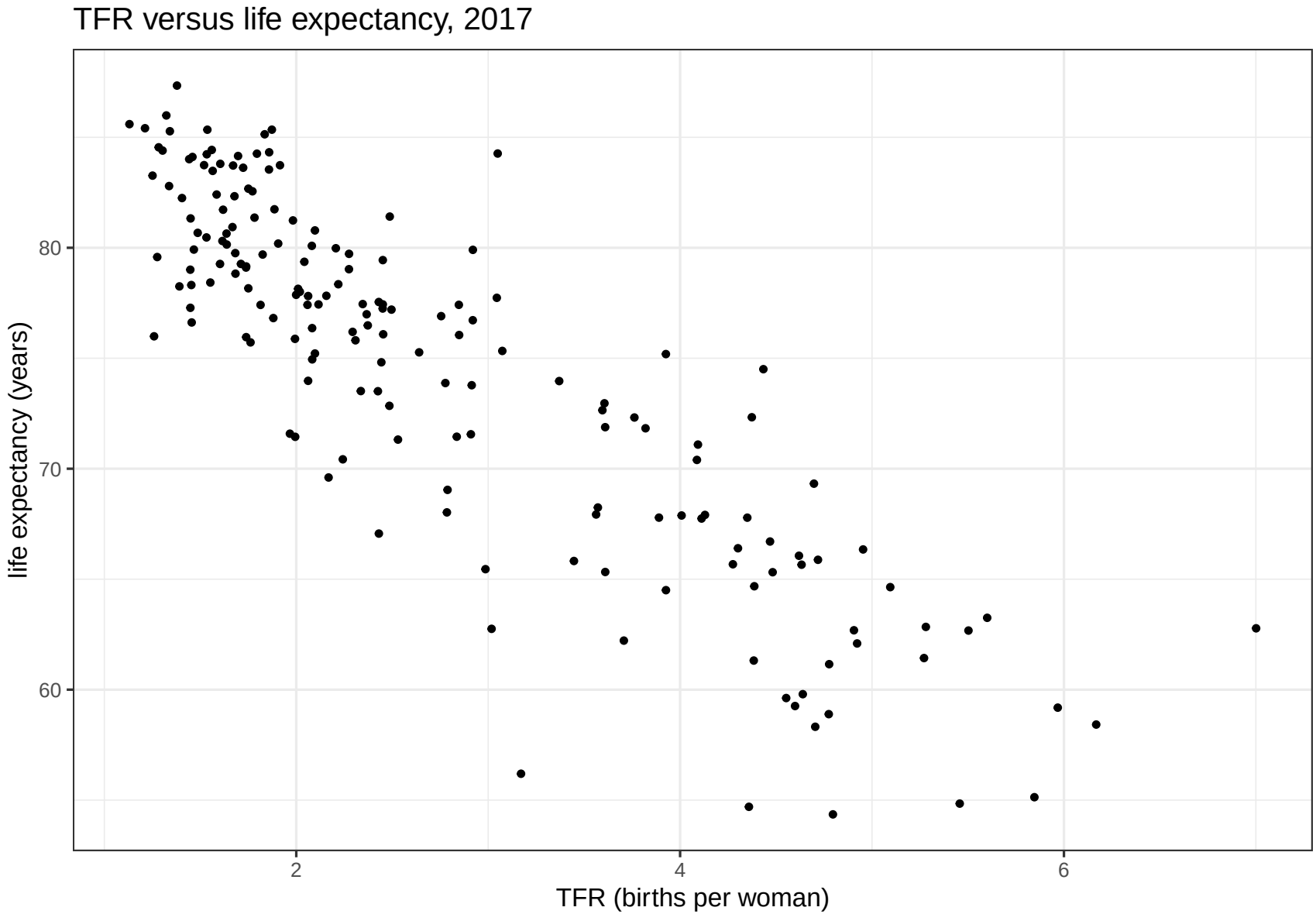
- Plots a series of values of a quantitative variable connected together by a line
- Useful to visualize trends over time

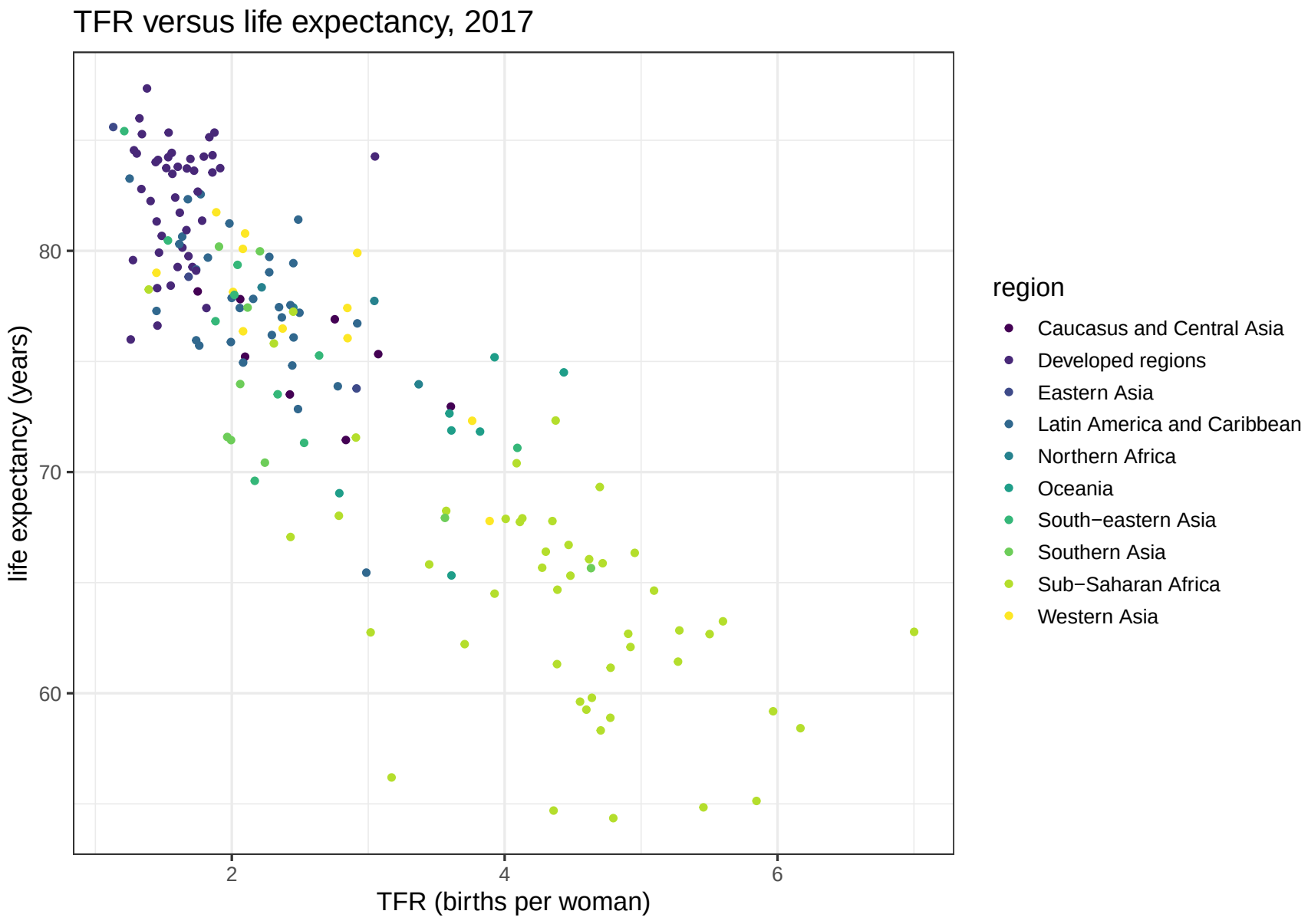

Total fertility rate, Canada and the US
2.0
1.9
1.8
1.7
1.6
tfr
2009
2010
2011
2012
2013
2014
2015
2016
2017
year
country
Canada
United States of America

Scatter plots

Shows relationship between two different **quantitative** variables

- Uses dots to represent values for two different **quantitative** values
- The position of each dot on the x and y axis indicates values for an individual data point
- Extremely useful in visualizing the relationship between two quantitative variables


TFR versus life expectancy, 2017
life expectancy (years)
80
70
60
2
4
6
TFR (births per woman)


TFR versus life expectancy, 2017
life expectancy (years)
80
70
60
2
4
6
TFR (births per woman)
region
Caucasus and Central Asia
Developed regions
Eastern Asia
Latin America and Caribbean
Northern Africa
Oceania
South−eastern Asia
Southern Asia
Sub−Saharan Africa
Western Asia

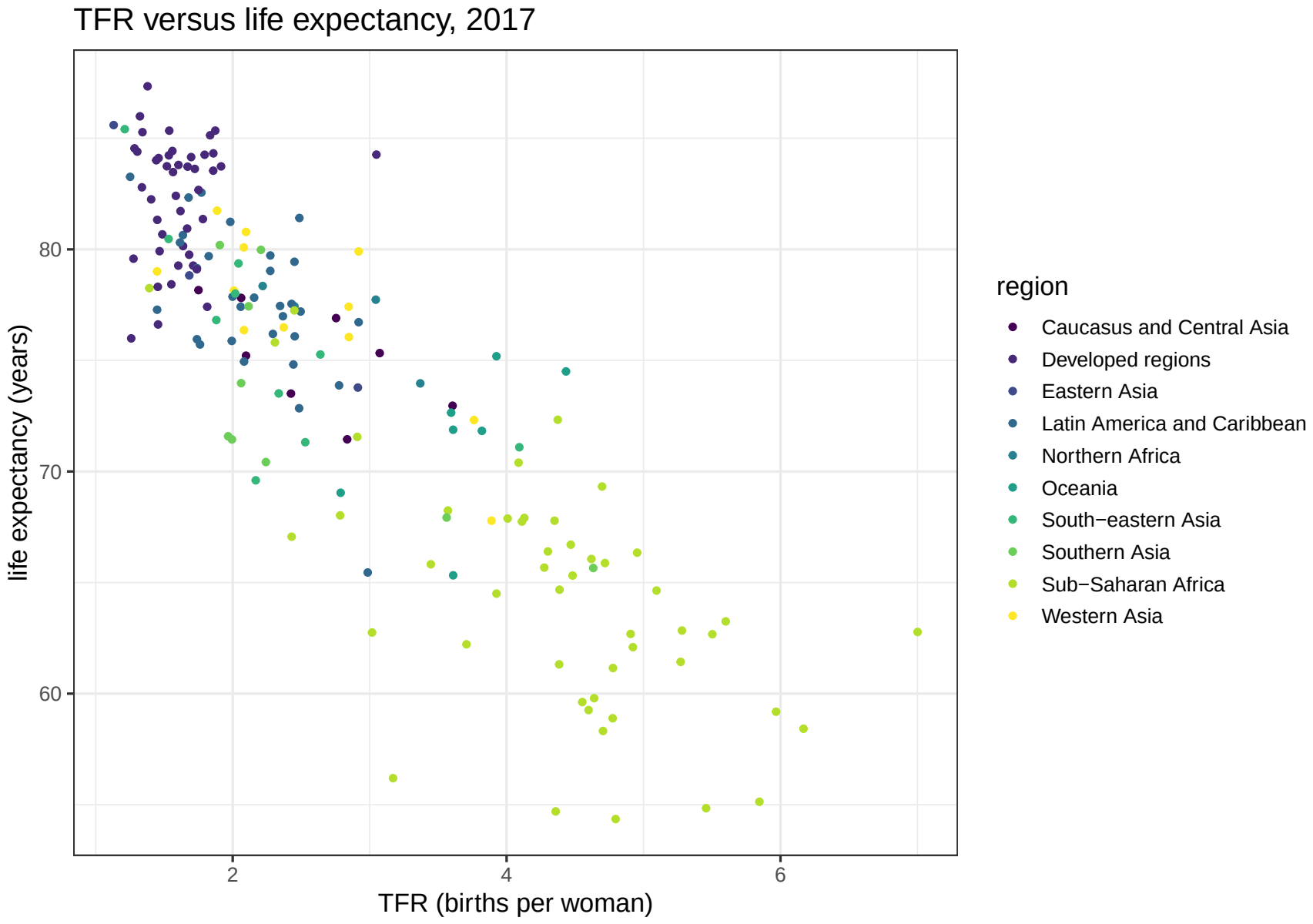
Further introduction to ggplot

ggplot

- `ggplot` is the graphing package that goes with the `tidyverse` in R
- Very powerful to make a wide range of graphics
- Every graph so far this lecture was done in `ggplot`
- `ggplot` code works in layers, with each layer adding complexity
  - start with defining dataset and different variables
  - add on type of plot
  - scales
  - layout (facets)
  - themes, fonts, sizes...

More practice in lab, but here's a starting example

Reproducing the TFR verus life expectancy chart, colored by region



Data

```
# read in the data
country_ind <- read_csv("../../data/country_indicators.csv")
country_ind
```

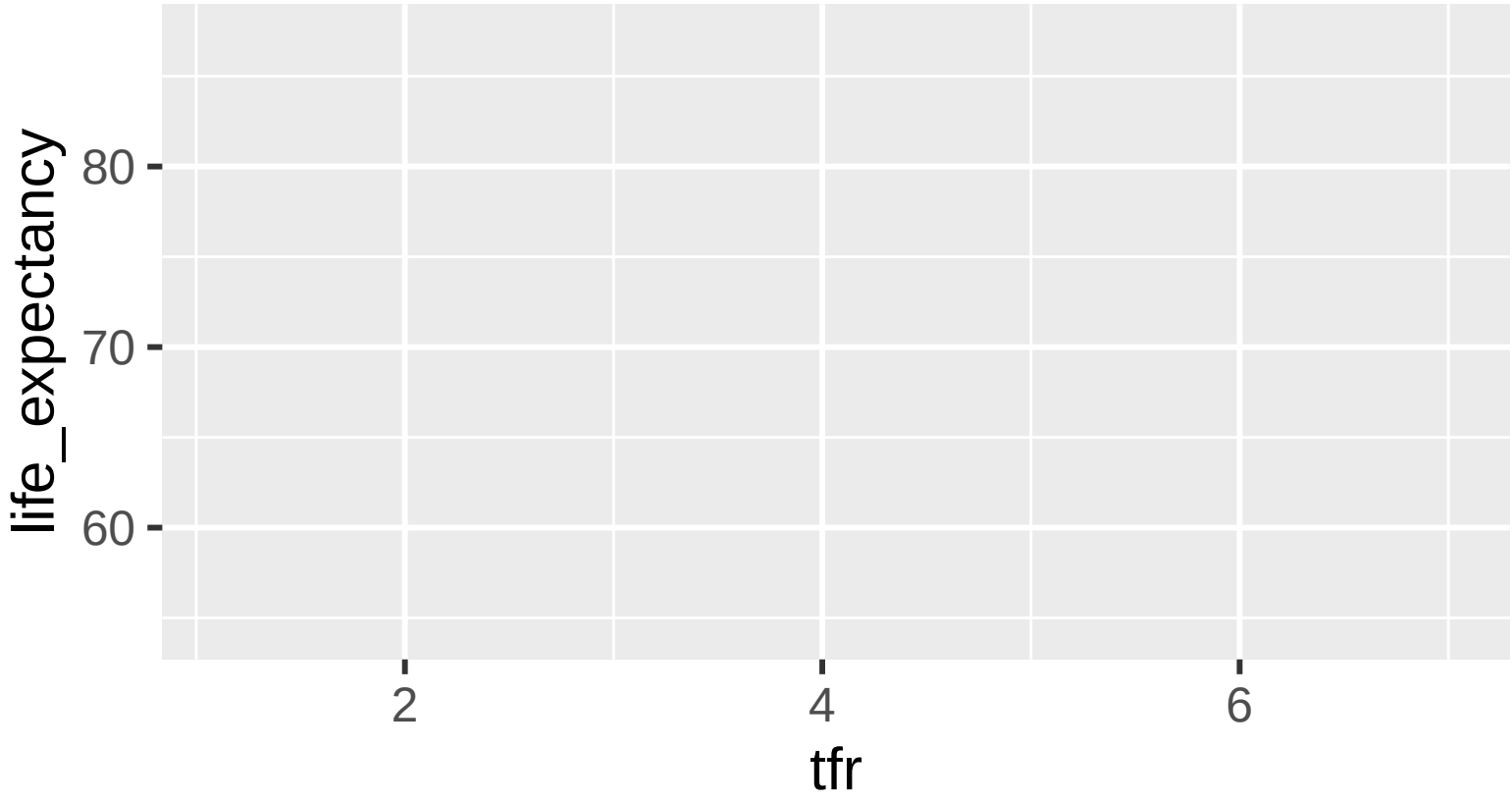
```
## # A tibble: 1,584 x 9
##    country_code country     region          year   tfr life_expectancy child_mort
##    <chr>        <chr>       <chr>          <dbl> <dbl>           <dbl>      <dbl>
##  1 AFG          Afghanistan Southern Asia   2009  6.18            61.9       93.9
##  2 AFG          Afghanistan Southern Asia   2010  5.98            62.5       90.0
##  3 AFG          Afghanistan Southern Asia   2011  5.77            63         86.3
##  4 AFG          Afghanistan Southern Asia   2012  5.56            63.5       82.9
##  5 AFG          Afghanistan Southern Asia   2013  5.36            64.0       79.6
##  6 AFG          Afghanistan Southern Asia   2014  5.16            64.5       76.6
##  7 AFG          Afghanistan Southern Asia   2015  4.98            64.9       73.8
##  8 AFG          Afghanistan Southern Asia   2016  4.80            65.3       71.2
##  9 AFG          Afghanistan Southern Asia   2017  4.63            65.7       68.8
## 10 ALB          Albania     Developed re~   2009  1.65            79.0       16.7
## # ... with 1,574 more rows, and 2 more variables: maternal_mort <dbl>,
## #   gdp <dbl>
```

```
# filter to just be 2017
country_ind_2017 <- country_ind %>% filter(year==2017)
```

A blank canvas

aes stands for aesthetic and tells ggplot the main characteristics of your plot (x, y, and if the color or fill vary by group)

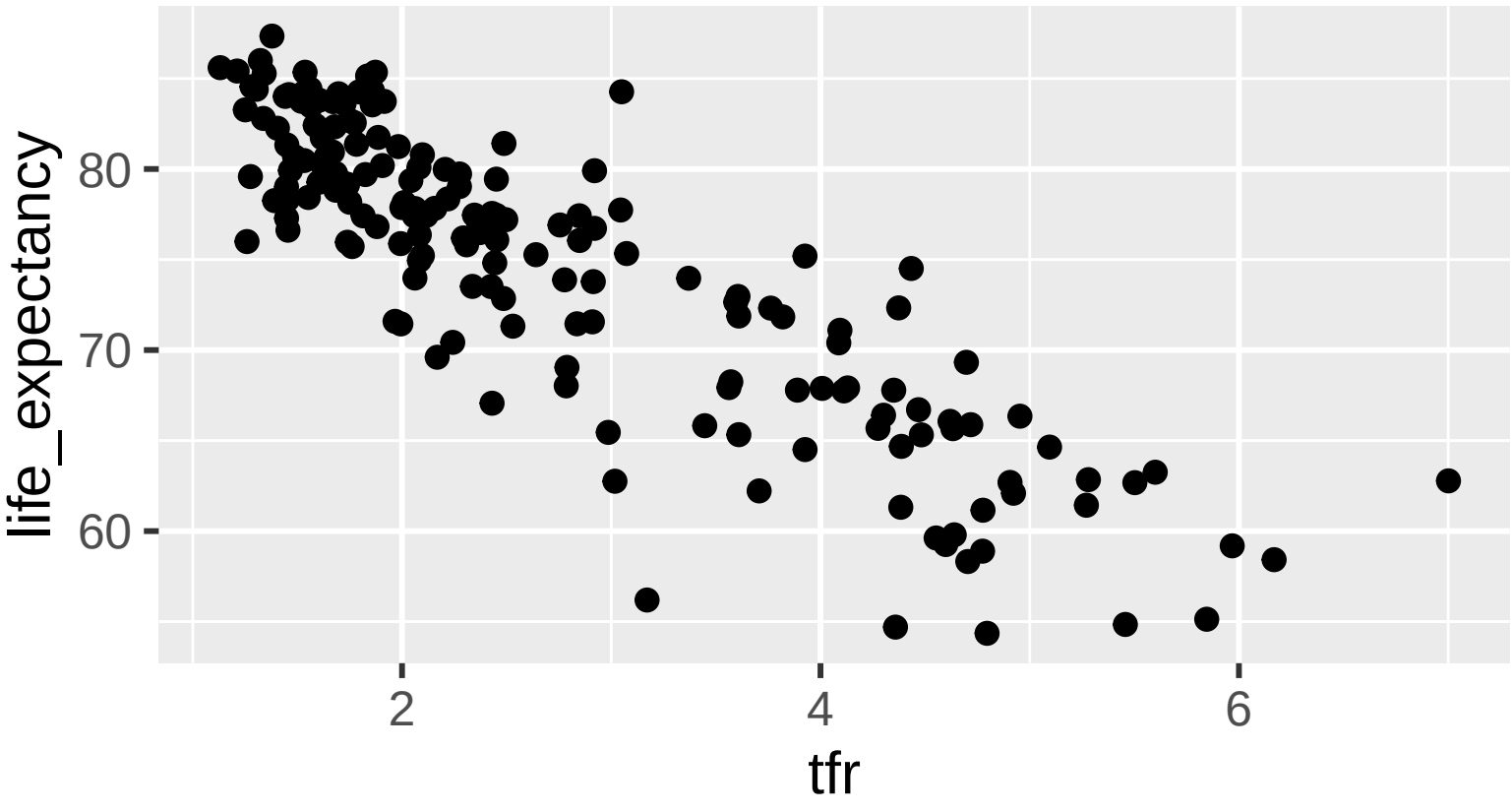
```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy))

#print
plot1
```



Add the points

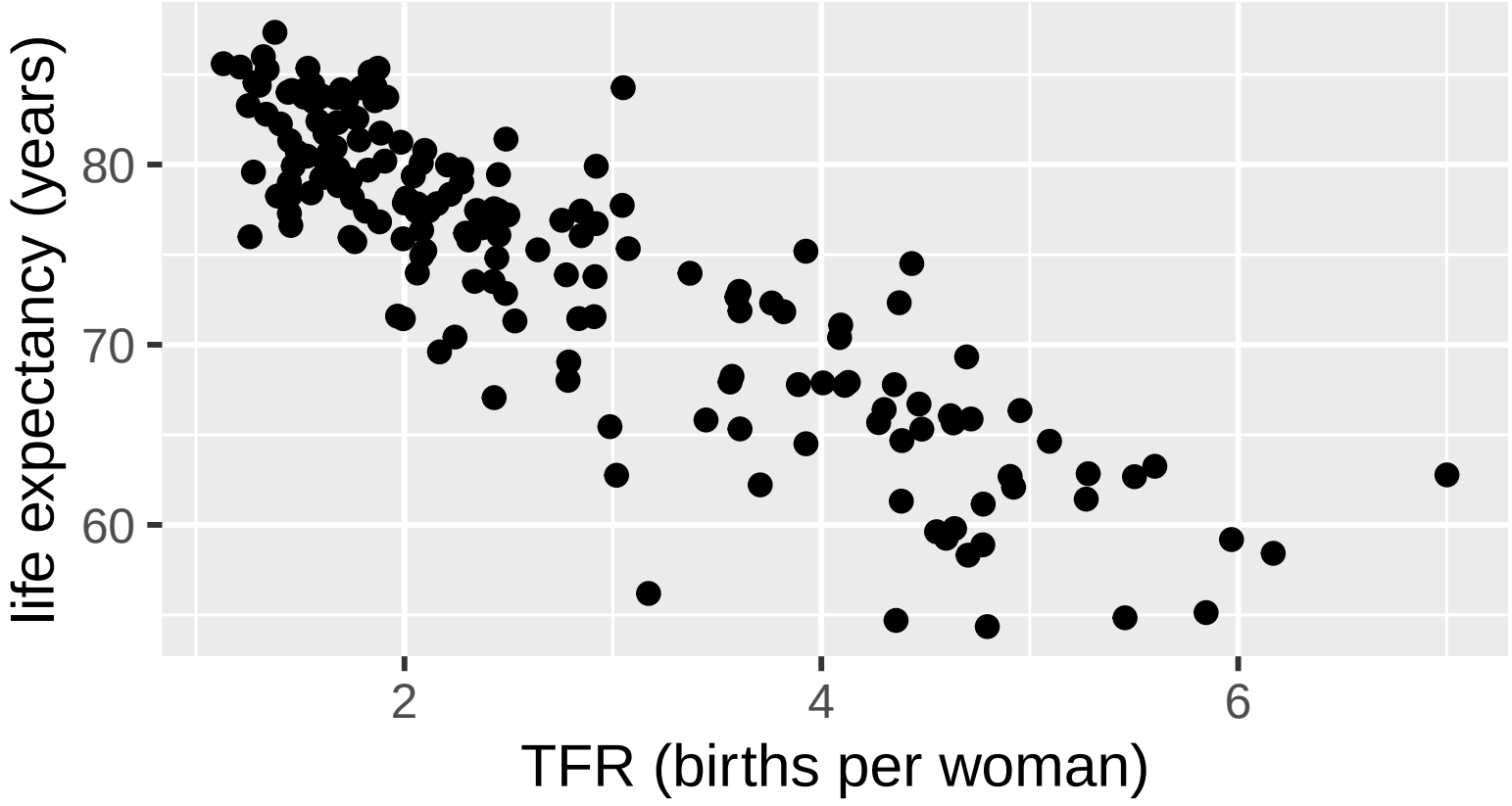
Add layers with ggplot using the +

```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy)) +
  geom_point()

plot1
```



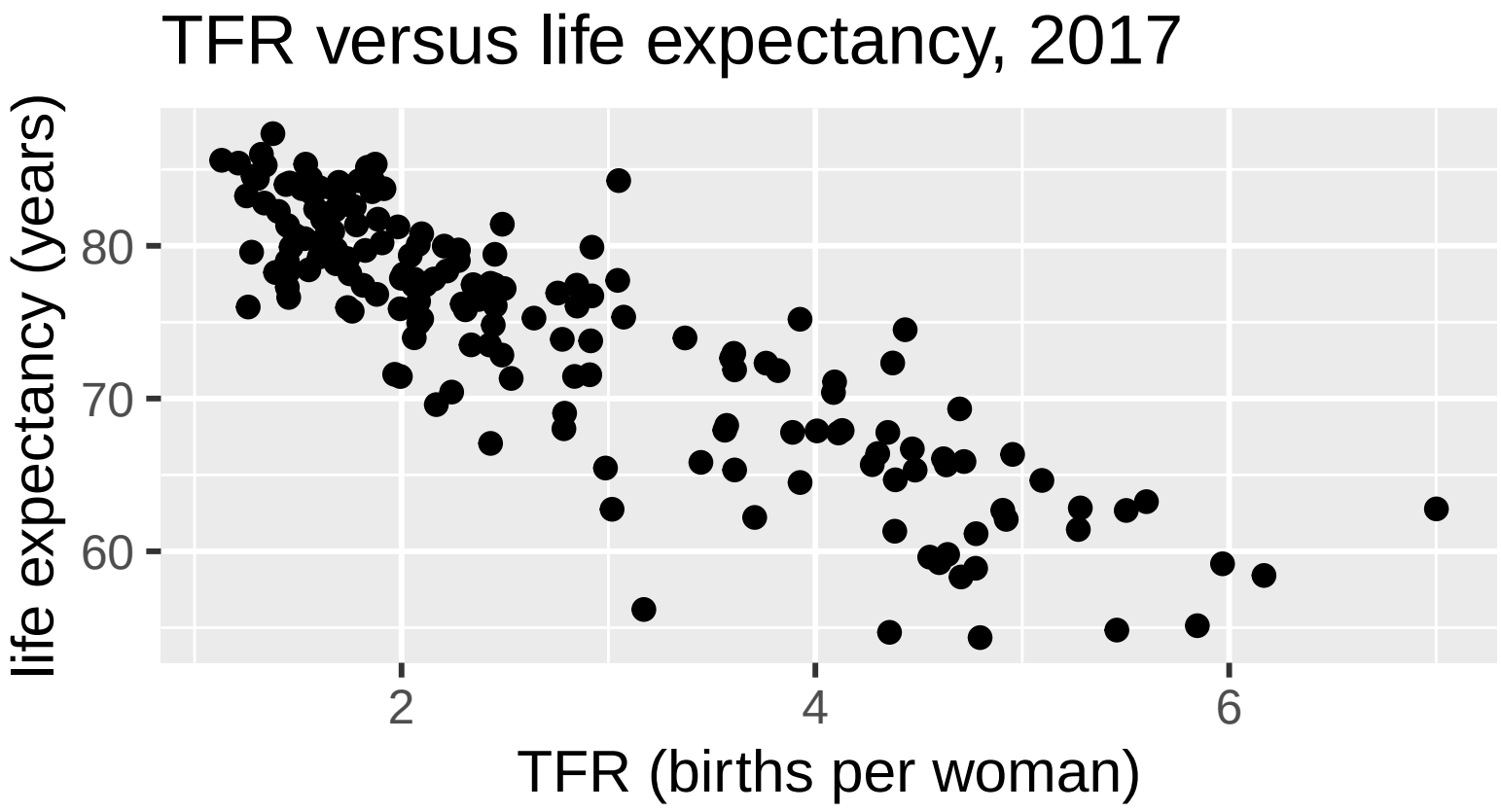
Tidy up labels

```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy)) +
  geom_point()+
  xlab("TFR (births per woman)")+
  ylab("life expectancy (years)")

plot1
```



Title

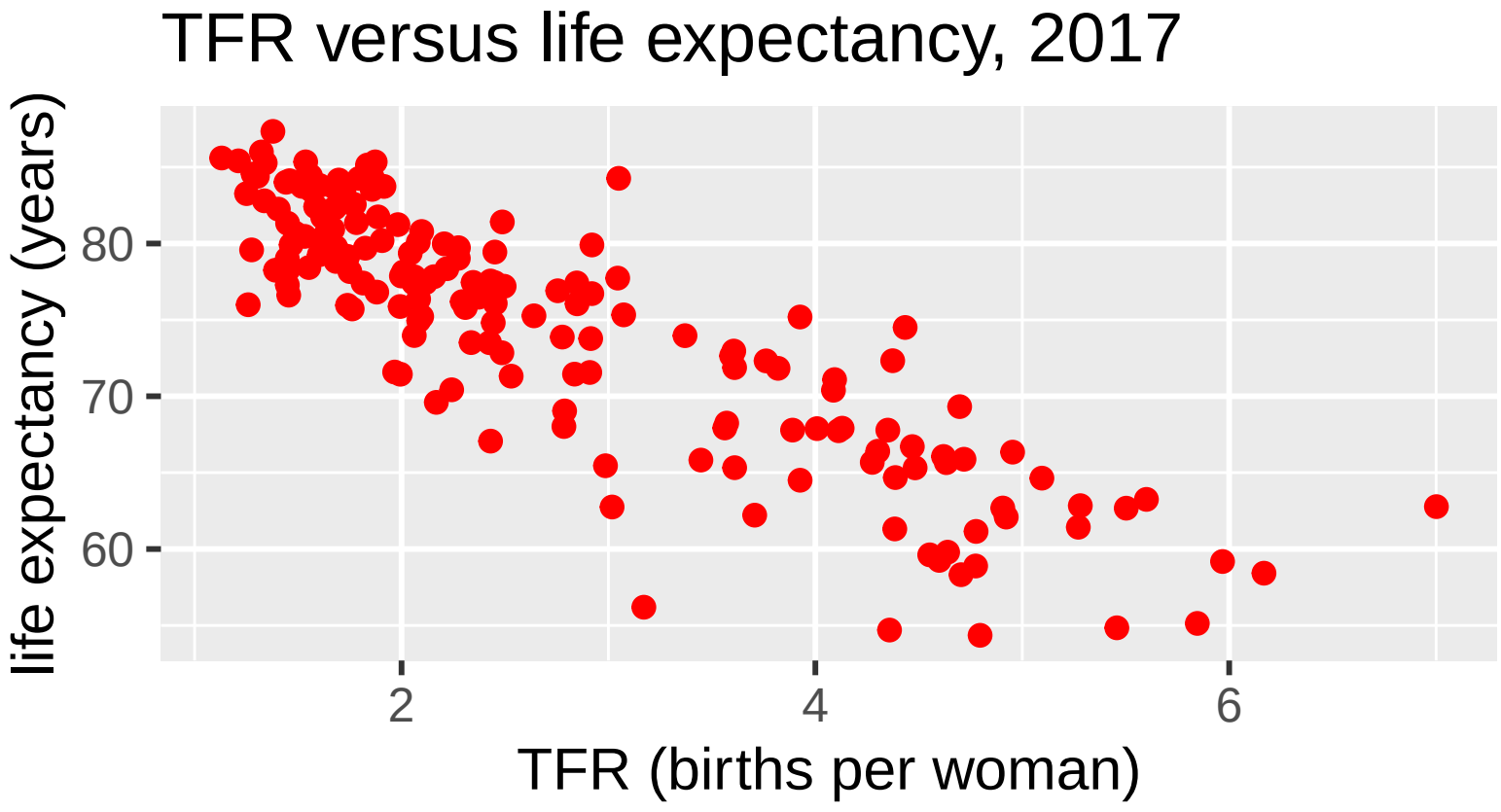
```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy)) +
  geom_point()+
  xlab("TFR (births per woman)")+
  ylab("life expectancy (years)")+
  ggtitle("TFR versus life expectancy, 2017")

plot1
```



Change color of points

to see all colors, type `colors()`

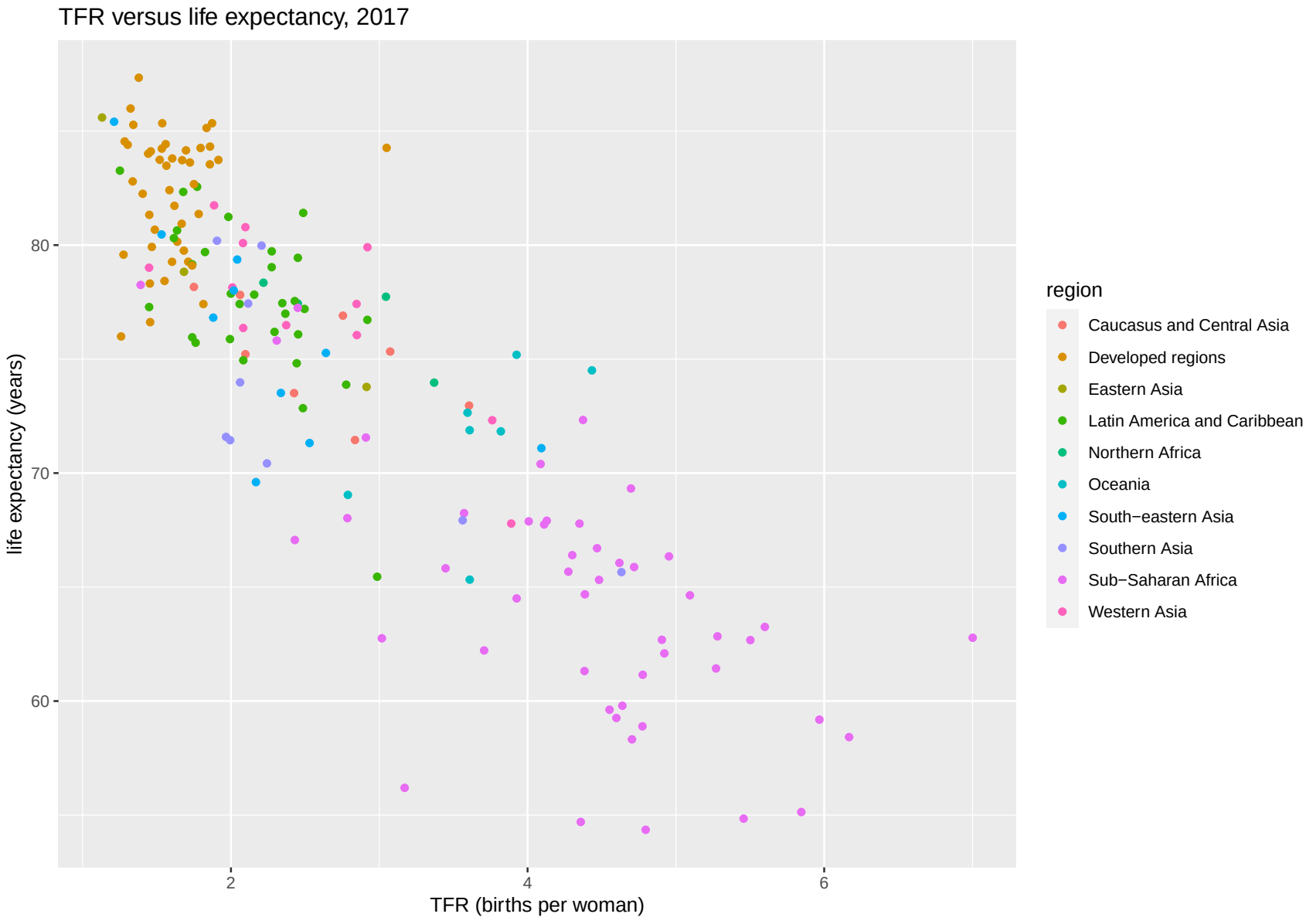
```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy)) +
  geom_point(color = "red")+
  xlab("TFR (births per woman)")+
  ylab("life expectancy (years)")+
  ggtitle("TFR versus life expectancy, 2017")

plot1
```



Coloring by group

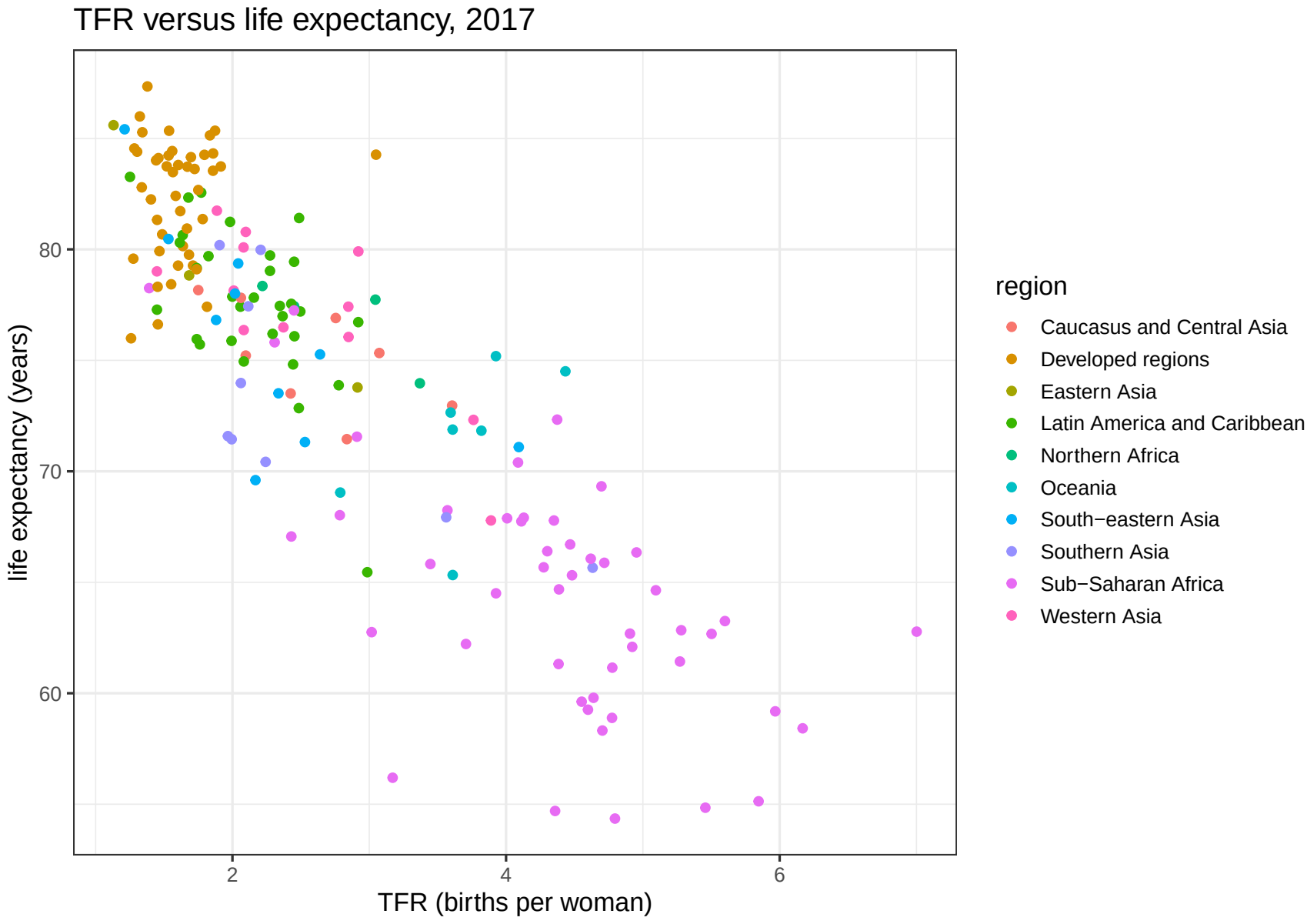
This goes in the `aes()` because it **depends on the data**

```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy, color = region)) +
  geom_point()+
  xlab("TFR (births per woman)")+
  ylab("life expectancy (years)")+
  ggtitle("TFR versus life expectancy, 2017")

plot1
```


TFR versus life expectancy, 2017
life expectancy (years)
80
70
60
2
4
6
TFR (births per woman)
region
Caucasus and Central Asia
Developed regions
Eastern Asia
Latin America and Caribbean
Northern Africa
Oceania
South−eastern Asia
Southern Asia
Sub−Saharan Africa
Western Asia

Change theme (optional) and size of points

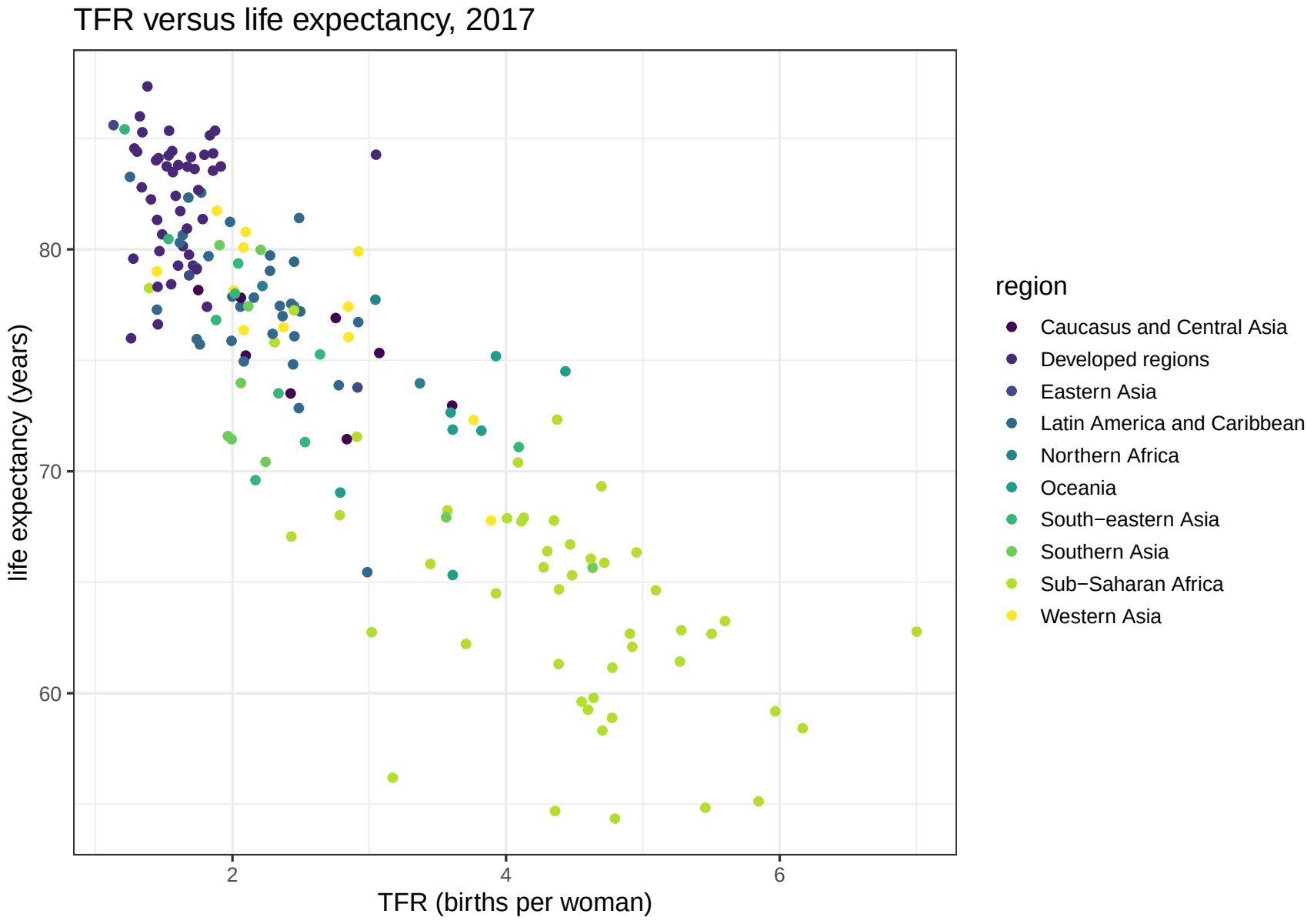
```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy, color = region)) +
  geom_point(size =2)+
  xlab("TFR (births per woman)")+
  ylab("life expectancy (years)")+
  ggtitle("TFR versus life expectancy, 2017")+
  theme_bw(base_size = 14)

plot1
```


TFR versus life expectancy, 2017
life expectancy (years)
80
70
60
2
4
6
TFR (births per woman)
region
Caucasus and Central Asia
Developed regions
Eastern Asia
Latin America and Caribbean
Northern Africa
Oceania
South−eastern Asia
Southern Asia
Sub−Saharan Africa
Western Asia

Change color scheme

`viridis` and `brewer` both good options

```
plot1 <- ggplot(data = country_ind_2017, aes(x = tfr, y = life_expectancy, color = region)) +
  geom_point(size =2)+
  xlab("TFR (births per woman)")+
  ylab("life expectancy (years)")+
  ggtitle("TFR versus life expectancy, 2017")+
  theme_bw(base_size = 14)+
  scale_color_viridis_d()

plot1
```


TFR versus life expectancy, 2017
life expectancy (years)
80
70
60
2
4
6
TFR (births per woman)
region
Caucasus and Central Asia
Developed regions
Eastern Asia
Latin America and Caribbean
Northern Africa
Oceania
South−eastern Asia
Southern Asia
Sub−Saharan Africa
Western Asia

Summary

- EDA and data visualization is often just as informative and important as statistical analysis
- It is essential to understand the structure of your data, missing-ness, any outliers/issues, and the raw patterns in your data before deciding on your statistical analysis
- Plot, plot, plot
- Practice, practice, practice

Summary

Plots:

- Bar charts for categorical/qualitative variables
- Histograms, boxplots for one quantitative variable (potentially across multiple categories)
- Line plots and scatter plots for two quantitative variables (line plot when one is sequential)

Data ideas

- IPUMS: https://ipums.org/
- ICPSR: https://www.icpsr.umich.edu/web/pages/ICPSR/thematic-collections.html
- CHASS SDA: https://datacentre.chass.utoronto.ca/
- Toronto Open Data Portal: https://open.toronto.ca/ or use `opendatatoronto` R package (ask for code)
- UN WPP: https://population.un.org/wpp/
- NBER: https://www.nber.org/research/data?page=1&perPage=50