Distributions

Week 8

AEM 2850 / 5850 : R for Business Analytics
Cornell Dyson
Fall 2025

Acknowledgements: Andrew Heiss, Claus Wilke

Announcements

Welcome back from Fall Break!

Prelim 1 grades will be released on this afternoon

The average grade was 74%. Great work -- it was a tough prelim!

I plan to curve final letter grades so that the average is in the B+ to A- range

Please see the canvas announcement and gradescope for more information

We will accept regrade requests through Thursday, October 23

Please see me if you are concerned about your ability to succeed in this course

Announcements

We will provide details on the group project soon

Questions before we get started?

Plan for this week

Tuesday

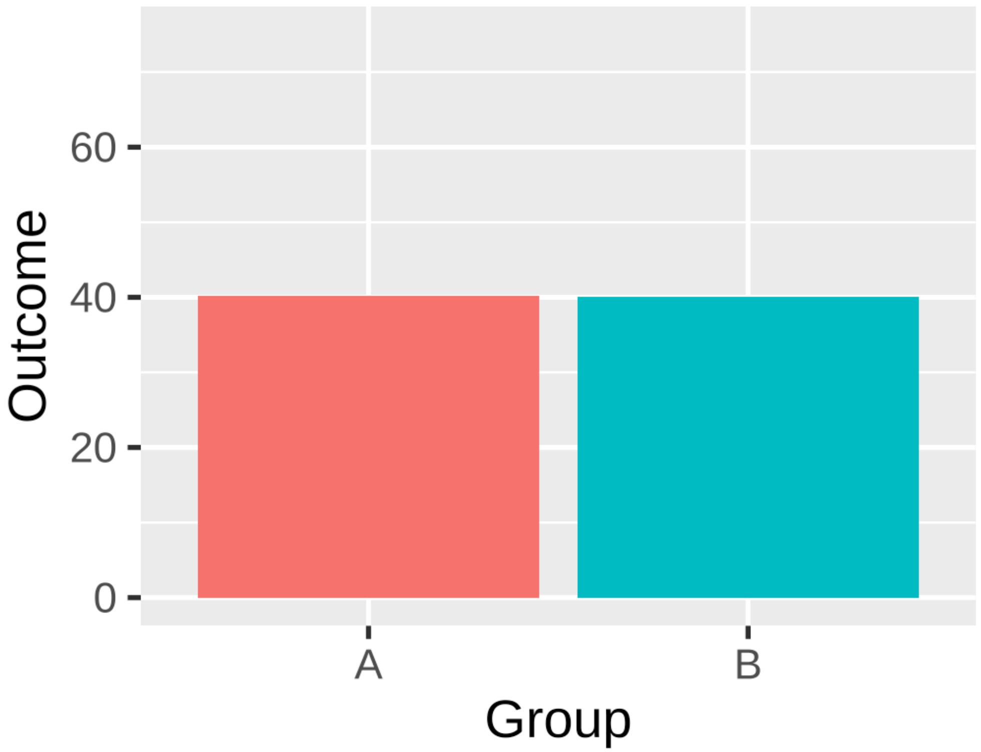
- *Fall Break: No class on Oct 14*

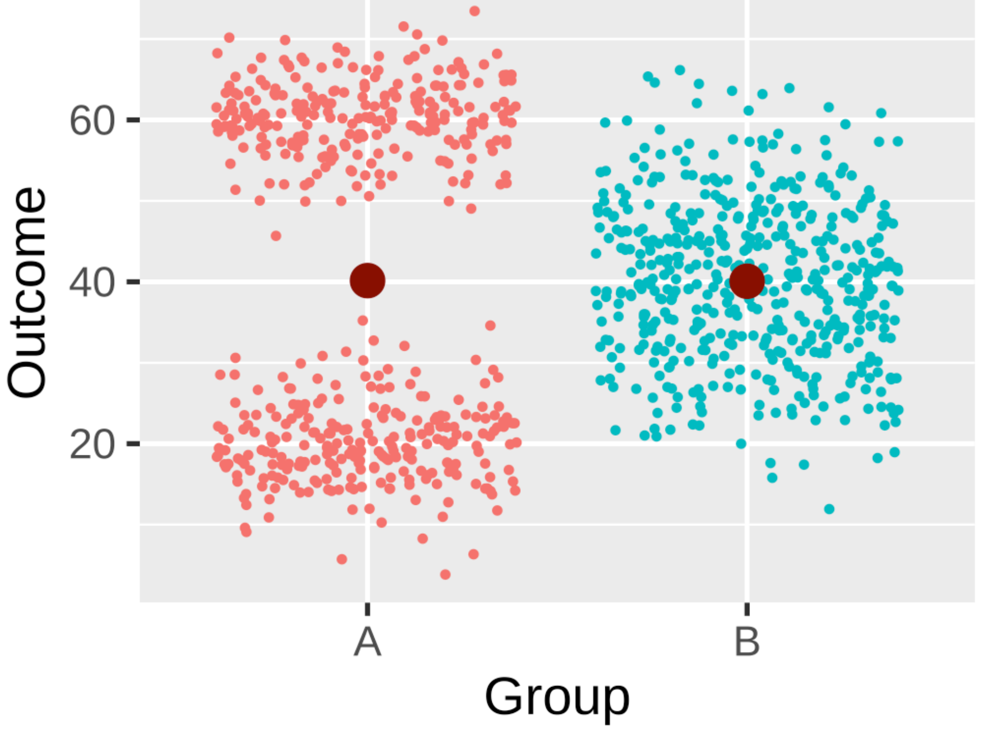
Thursday

- Distributions
- example-08-2

Distributions

Problems with single numbers





More information is (almost) always better

Avoid visualizing single numbers when you have a whole range or distribution of numbers

Uncertainty in single variables

Uncertainty across multiple variables

Uncertainty in models and simulations

What are some common methods for visualizing distributions?

Histograms, densities, box plots

Histograms

What are they?

Put data into equally spaced buckets (or "bins") based on values of a variable, plot how many rows of the data frame are in each bucket

Histograms

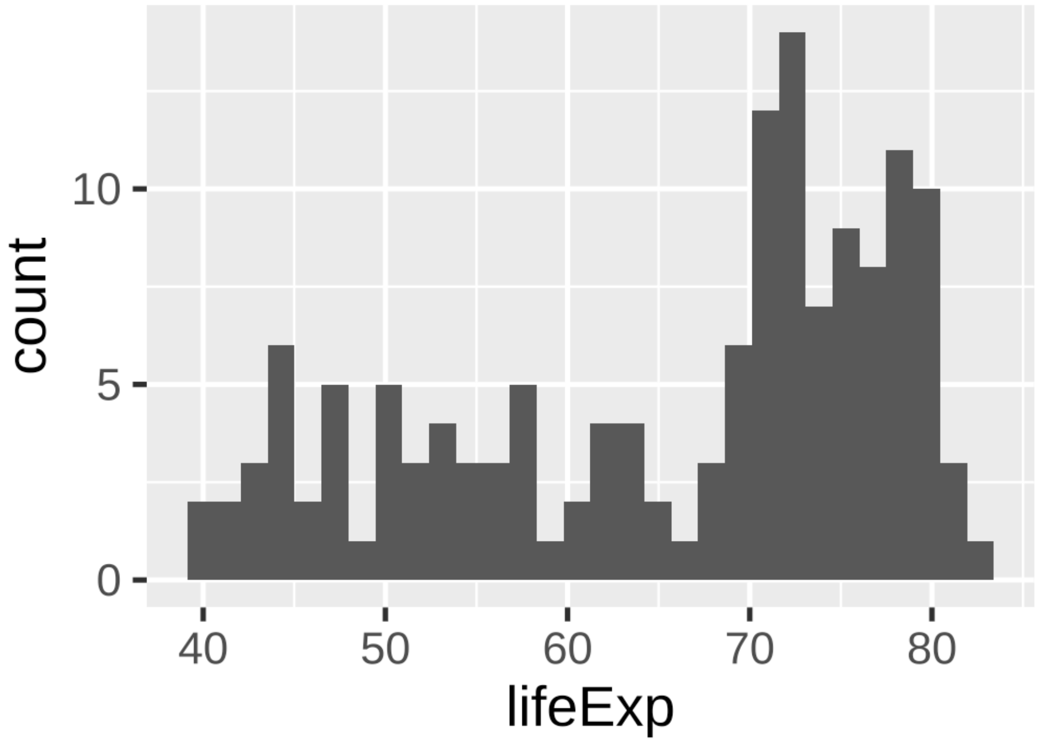
How would we use the grammar of graphics to make a histogram of `lifeExp`?

```r
library(gapminder)

gapminder_2002 <- gapminder |>
  filter(year == 2002)

head(gapminder_2002)
```

```
## # A tibble: 6 × 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       2002    42.1 25268405      727.
## 2 Albania     Europe     2002    75.7  3508512     4604.
## 3 Algeria     Africa     2002    71.0 31287142     5288.
## 4 Angola      Africa     2002    41.0 10866106     2773.
## 5 Argentina   Americas   2002    74.3 38331121     8798.
## 6 Australia   Asia       2002    80.4 19546792    30688.
```

Histograms

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_histogram()
```

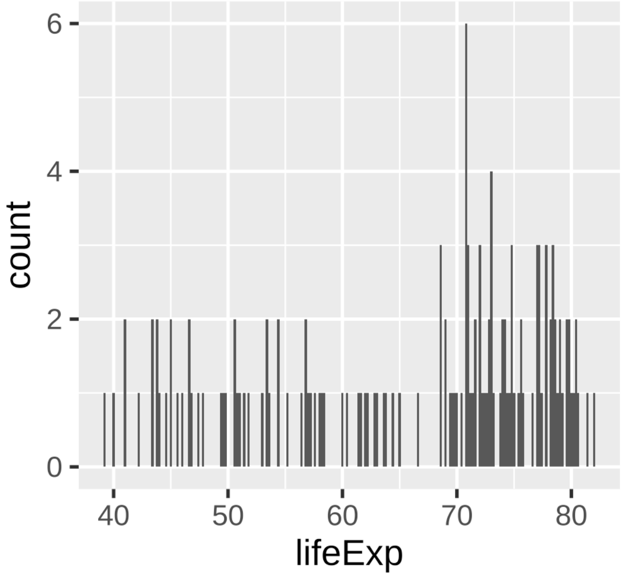


Histograms: binwidth argument

No official rule for what makes a good bin width

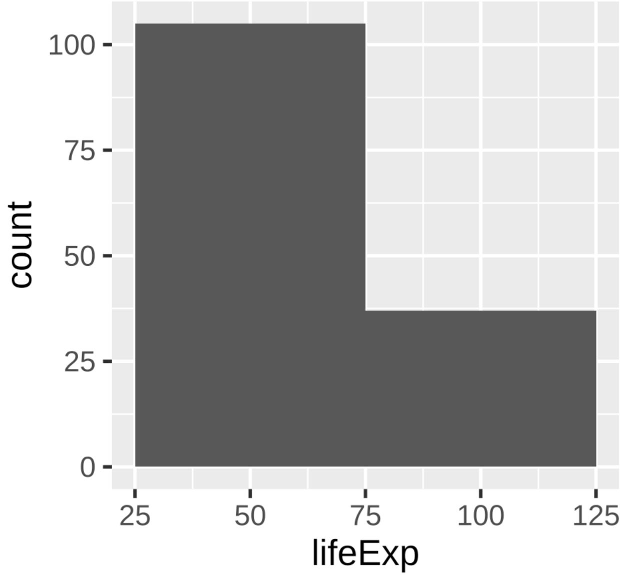
Too narrow:

`geom_histogram(binwidth = .2)`



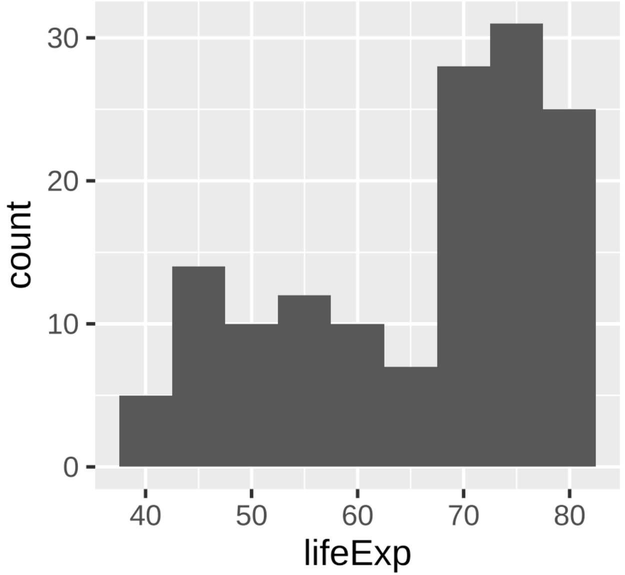
Too wide:

`geom_histogram(binwidth = 50)`



(One type of) just right:

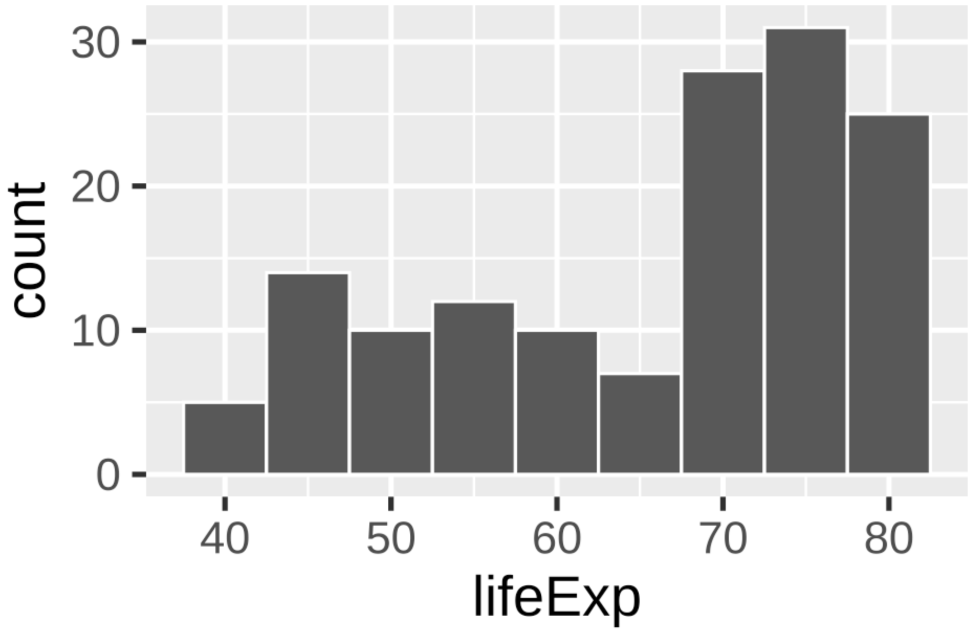
`geom_histogram(binwidth = 5)`



Histograms: tips using other arguments

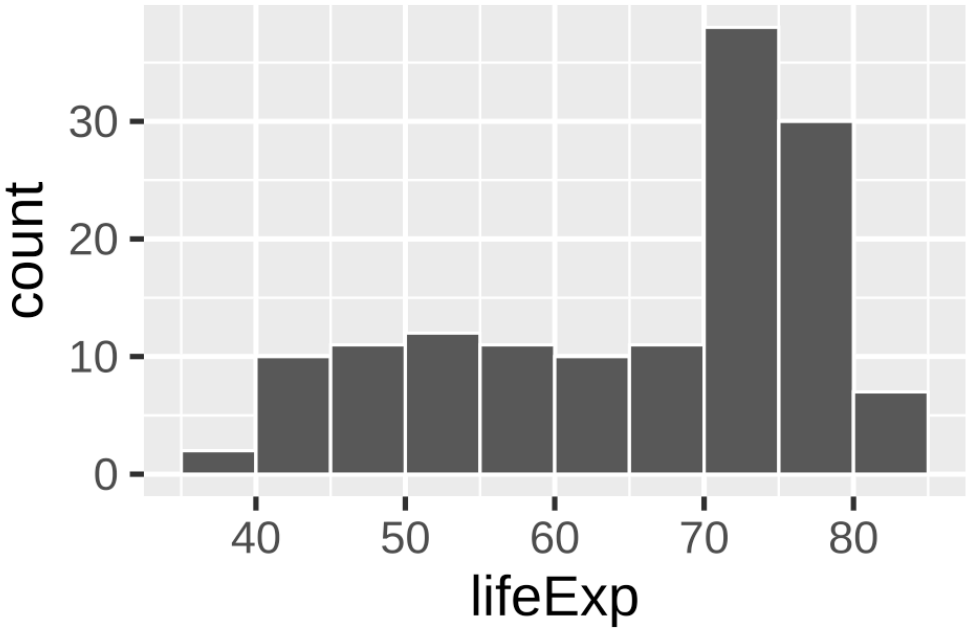
Add a border to the bars for readability

```
geom_histogram(..., color = "white")
```



Set the boundary;
bucket now 50–55, not 47.5–52.5

```
geom_histogram(..., boundary = 50)
```



Density plots

What are they?

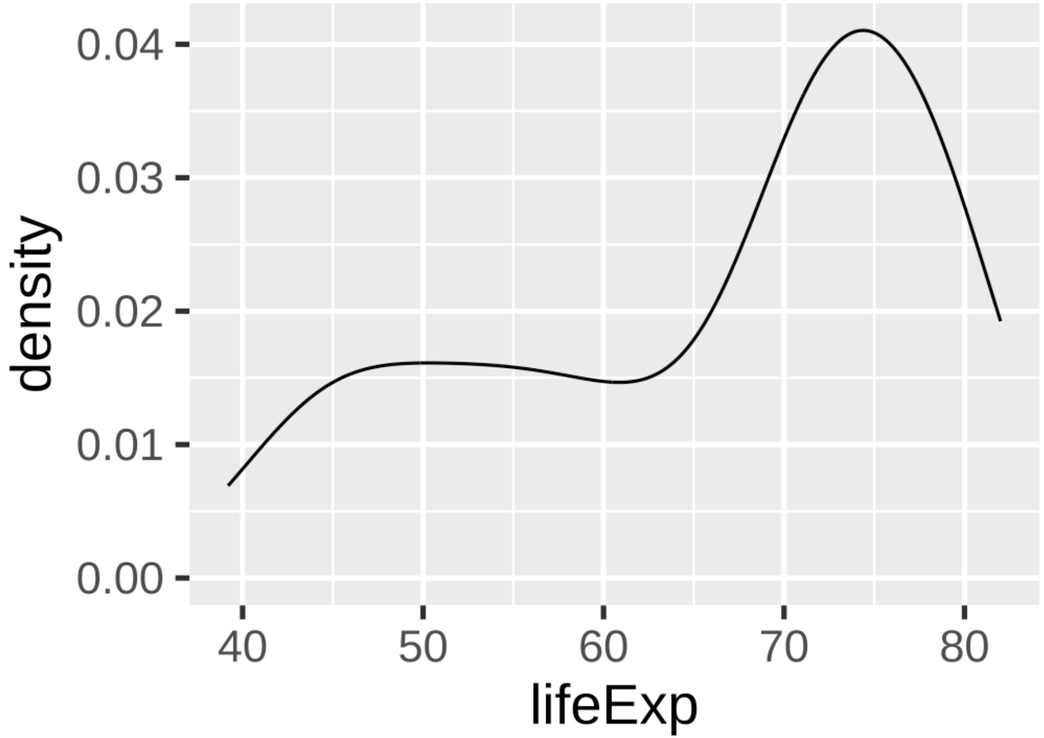
Estimates of the **probability *density* function** of a random variable

Histograms show raw counts; density plots show proportions (integrate to 1)

How would we use the grammar of graphics to make a density plot of `lifeExp`?

Density plots

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_density()
```

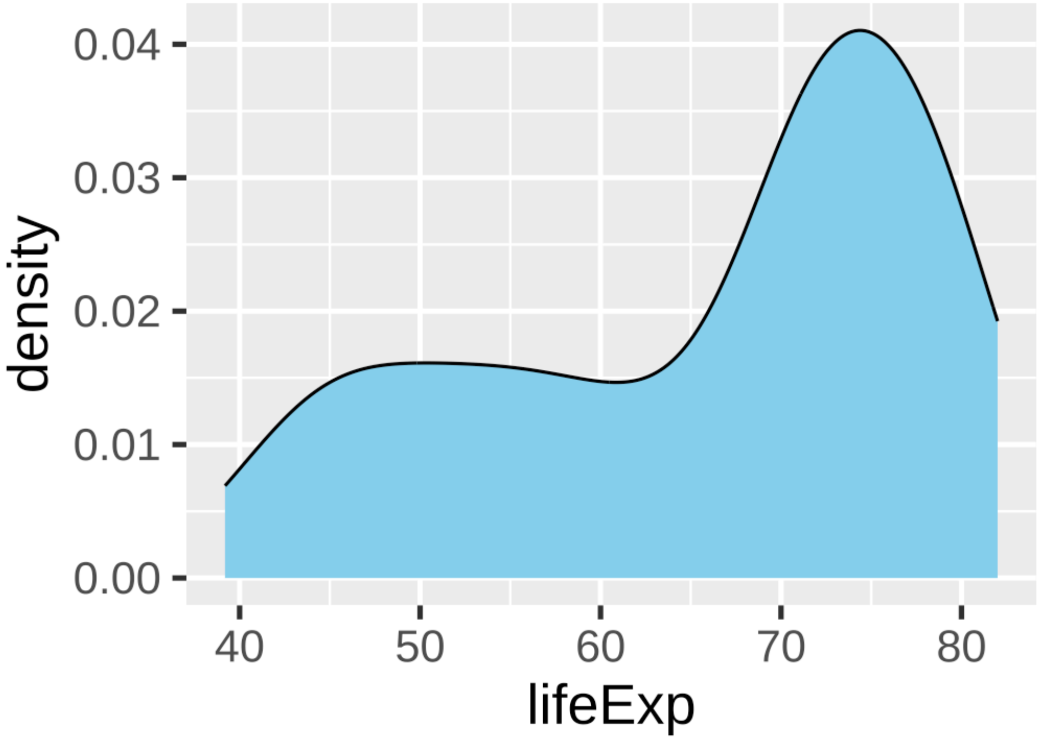


Density plots: add some color

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_density(fill = "skyblue")
```

We can use aesthetics as *parameters* inside a geom rather than inside an **aes()** statement

Here we used **fill = "skyblue"**

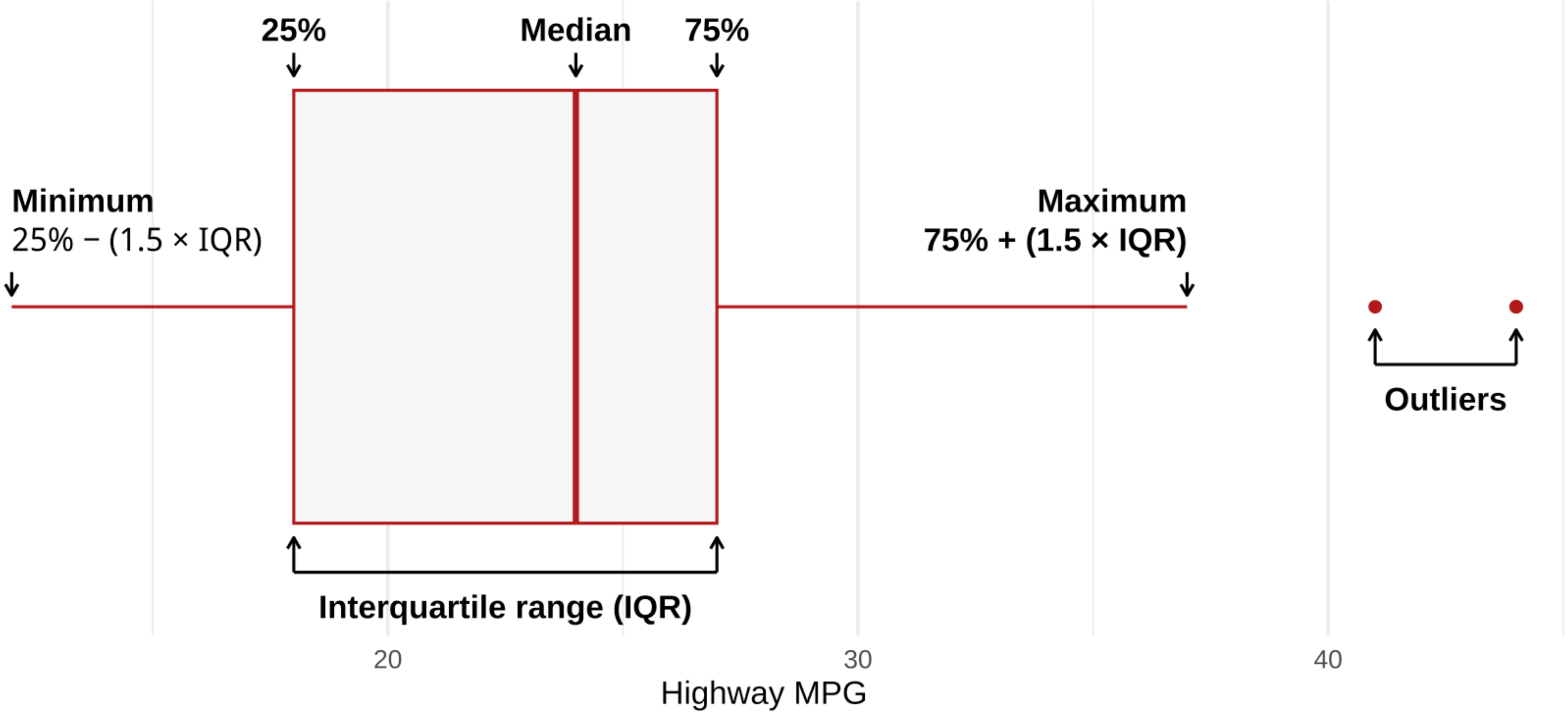


Box and whisker plots

What are they?

Graphical representations of specific points in a distribution

Box and whisker plots



Box and whisker plots

What are they?

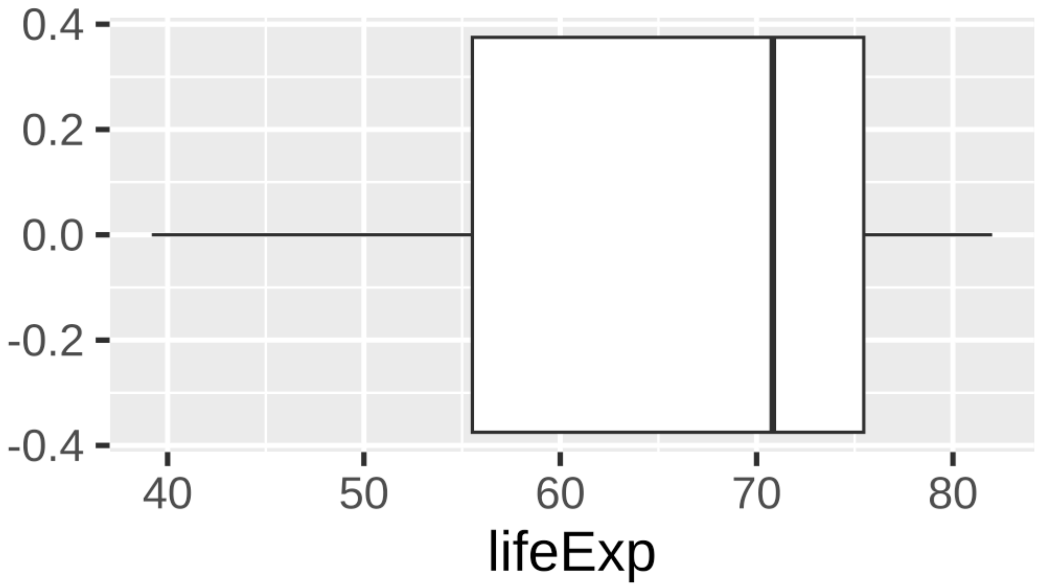
Graphical representations of specific points in a distribution

How could we use ggplot to make a boxplot of `lifeExp`?

Box and whisker plots

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_boxplot()
```

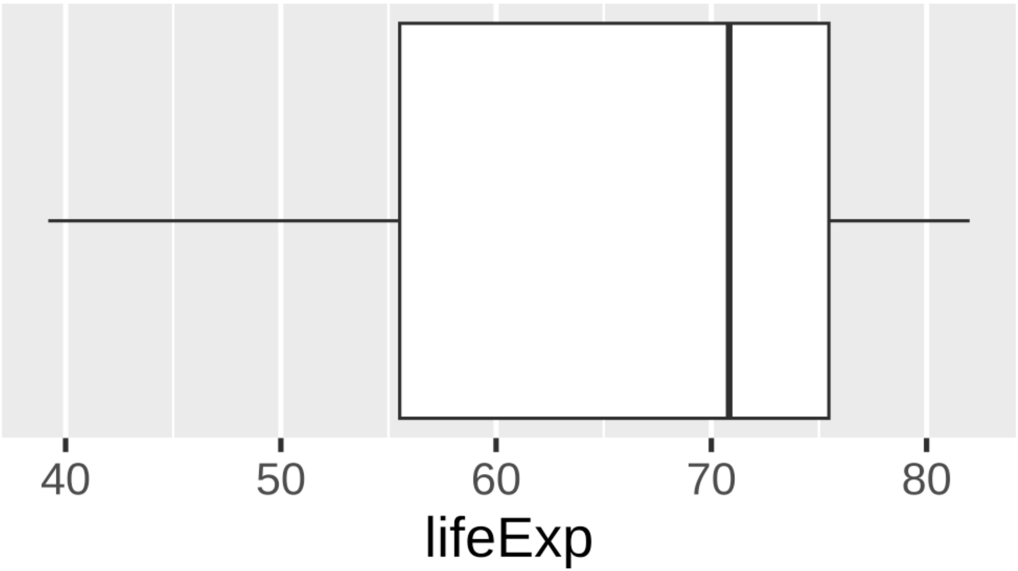
What do the y axis numbers mean?



Box and whisker plots

Use `theme()` to customize the plot for this geom

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_boxplot() +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    panel.grid.major.y = element_blank(),
    panel.grid.minor.y = element_blank()
  )
```



Uncertainty across multiple variables

How could we visualize the distribution of a single variable across groups?

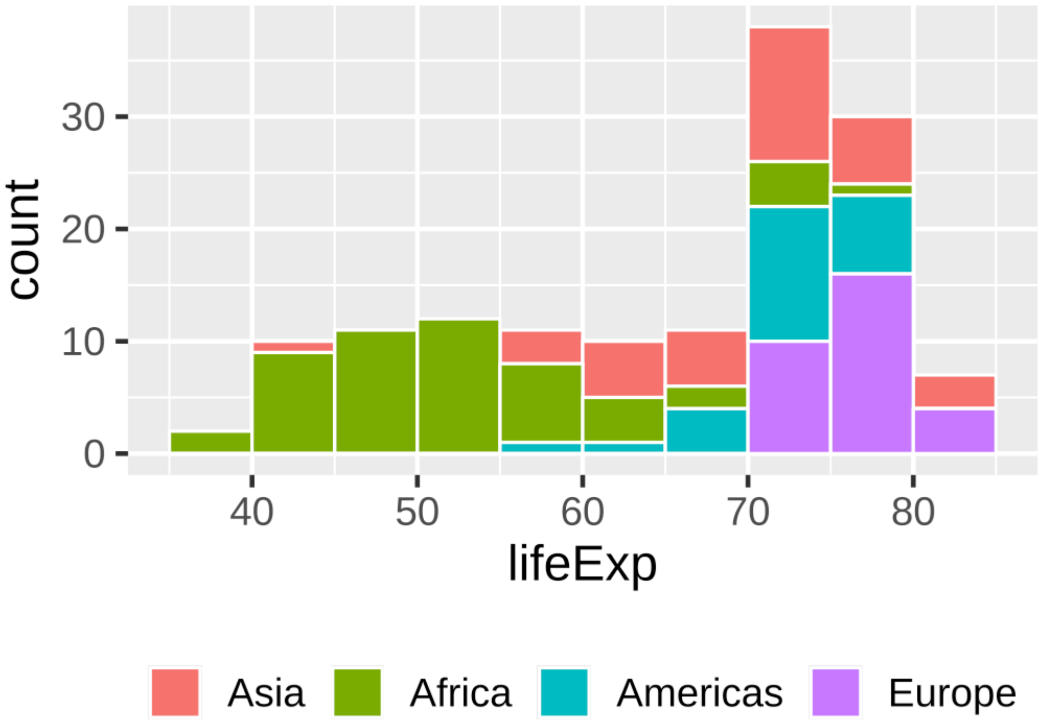
Add a `fill` aesthetic or use facets!

Multiple histograms

Fill with a different variable

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp,
             fill = continent)) +
  geom_histogram(binwidth = 5,
                 color = "white",
                 boundary = 50) +
  theme(legend.position = "bottom") +
  labs(fill = NULL)
```

This stacked histogram is bad and hard to read though

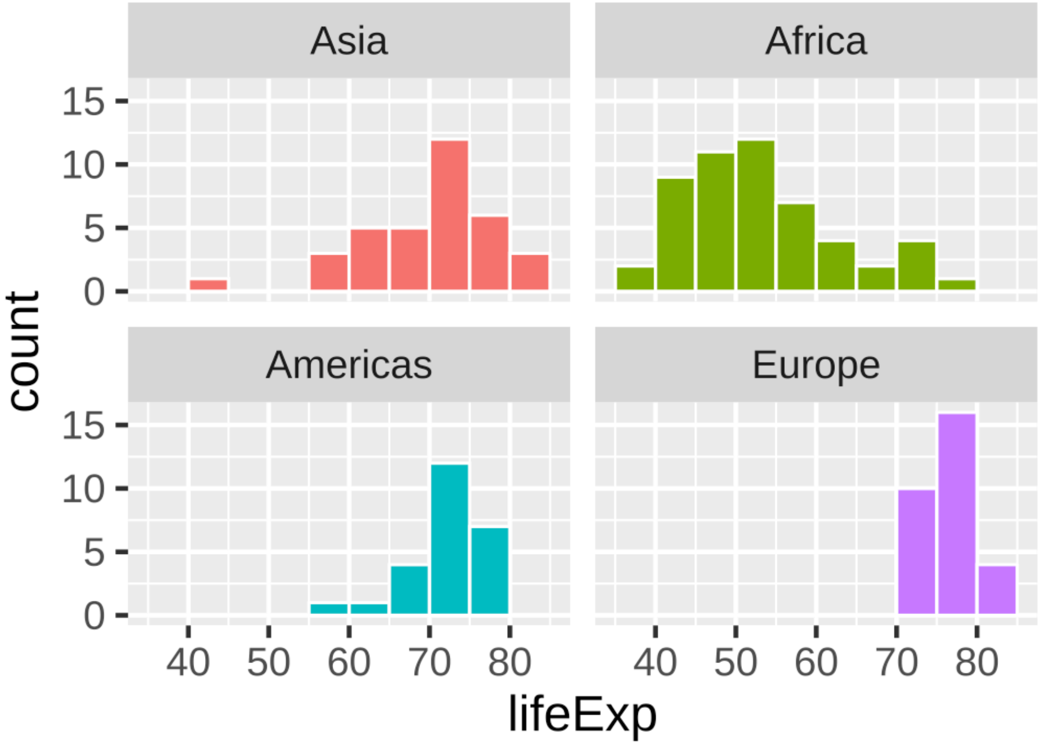


Multiple histograms

Facet with a different variable

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp,
             fill = continent)) +
  geom_histogram(binwidth = 5,
                 color = "white",
                 boundary = 50) +
  facet_wrap(vars(continent)) +
  guides(fill = "none")
```

Note: we could also omit `fill = continent`

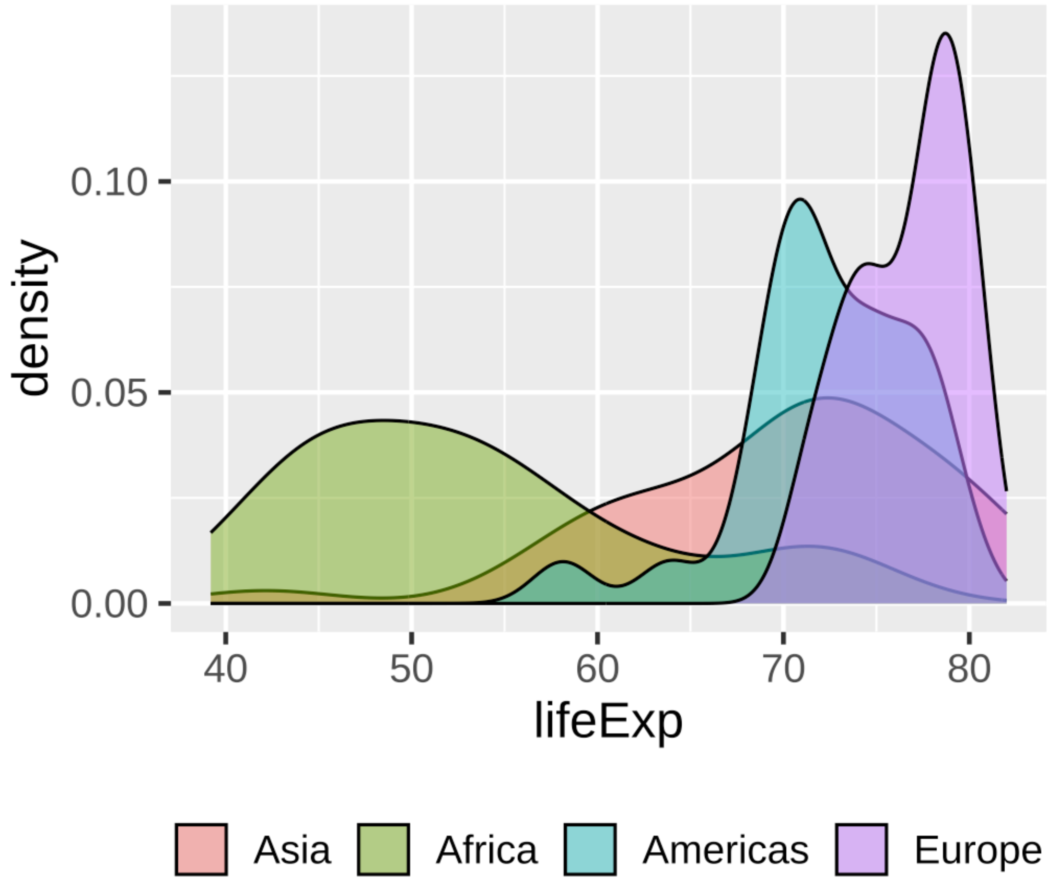


Multiple densities: Transparency

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp,
             fill = continent)) +
  geom_density(alpha = 0.5) +
  theme(legend.position = "bottom") +
  labs(fill = NULL)
```

But be careful, these can get confusing quickly

With many groups, better to space them out using ridgeline plots

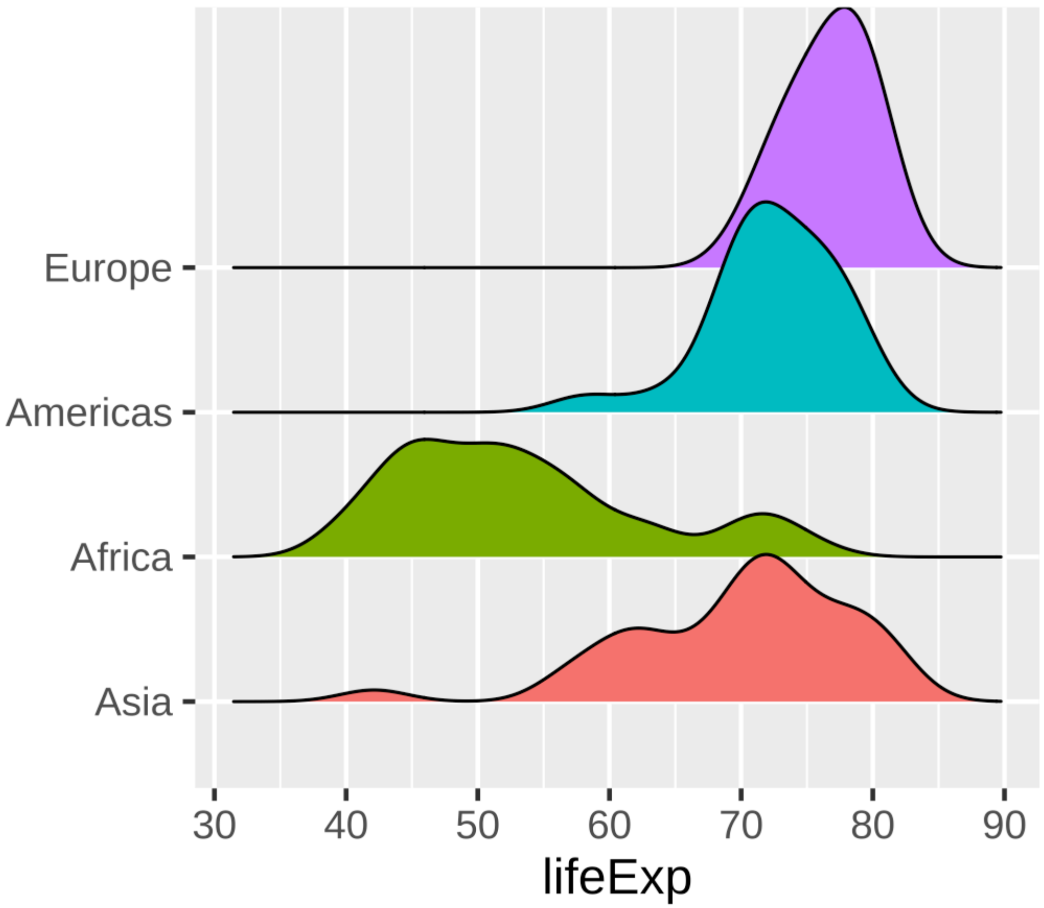


Multiple densities: Ridgeline plots

```r
library(ggridges)

gapminder_2002 |>
  ggplot(aes(x = lifeExp,
             fill = continent,
             y = continent)) +
  guides(fill = "none") +
  labs(y = NULL) +
  geom_density_ridges()
```

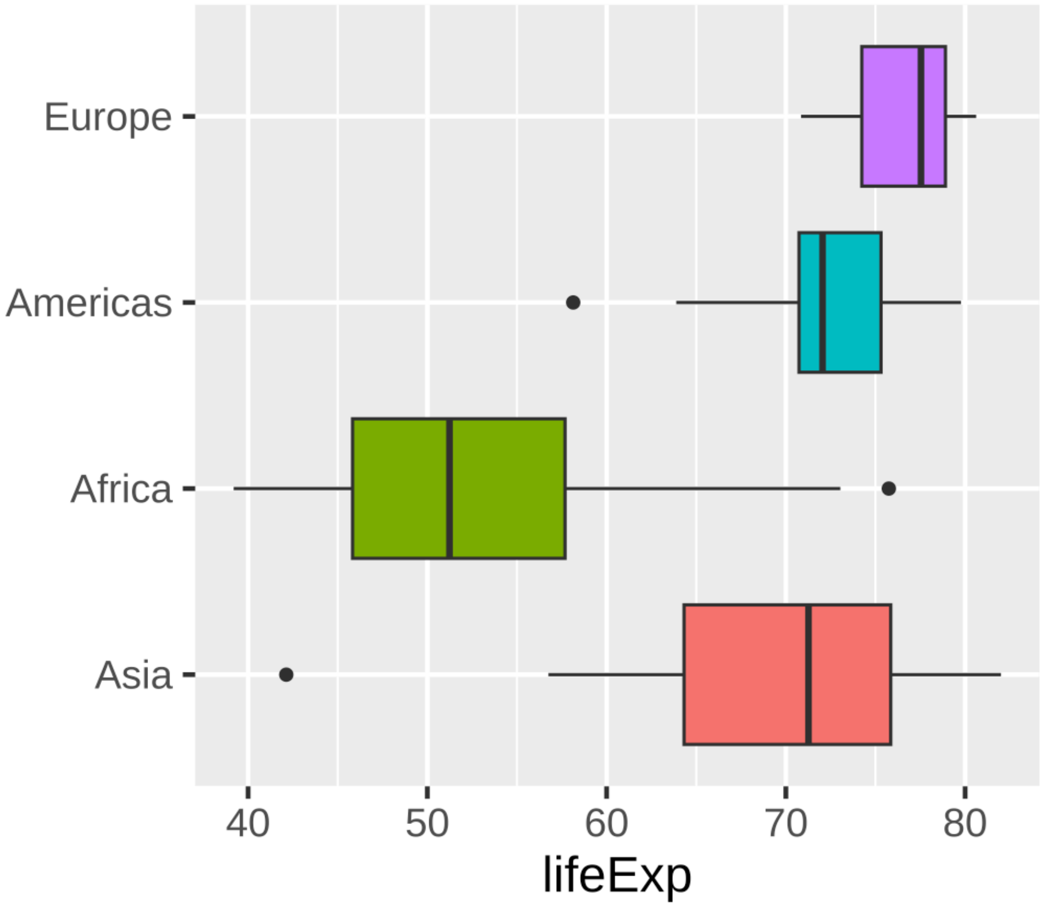
There is no explicit scale for the densities anymore (it is shared with y)

With many densities, use a single fill color to prevent distraction



Multiple box and whisker plots

```
gapminder_2002 |>
  ggplot(aes(
    x = lifeExp,
    fill = continent,
    y = continent
  )) +
  guides(fill = "none") +
  labs(y = NULL) +
  geom_boxplot()
```



example-08:
distributions-practice.R