Fundamentals of Data Visualization

K Arnold, based on IntroDS.org

Review

What do each of the following operations do in RStudio?

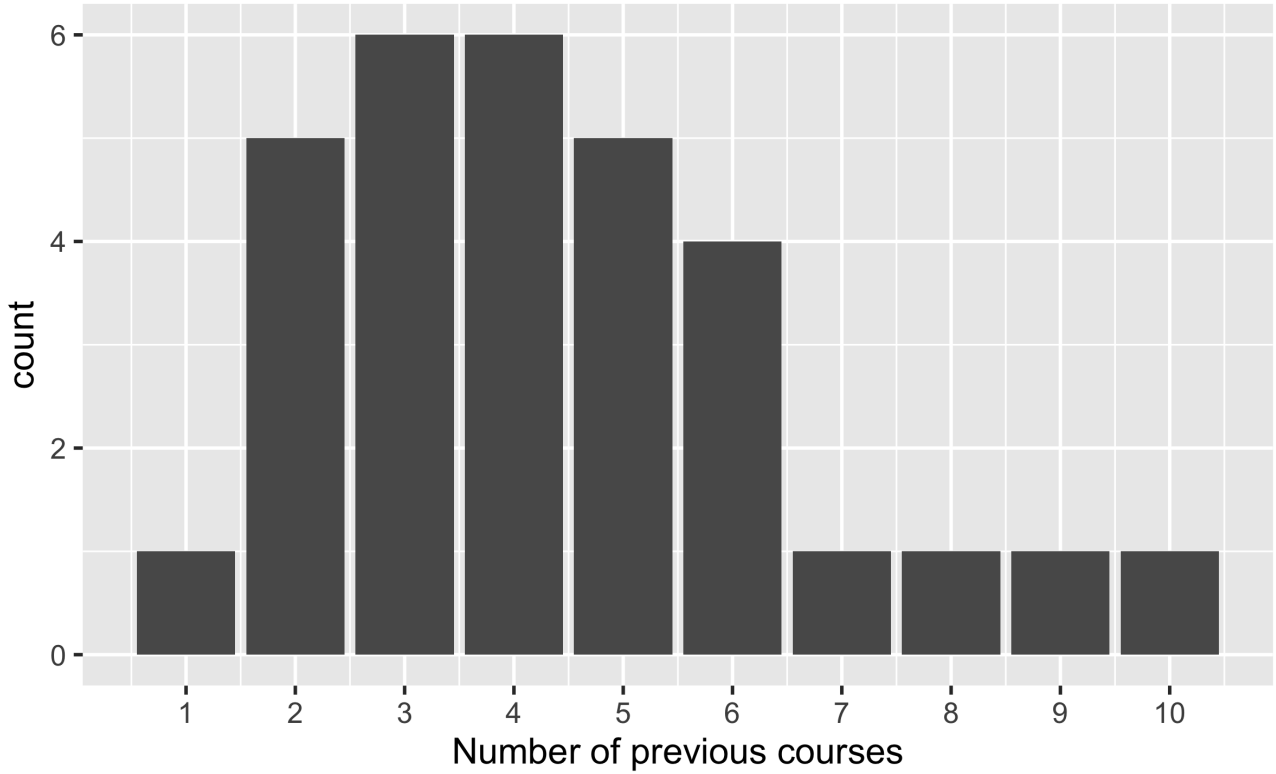
- Run Chunk
- Knit
- Commit
- Push

Think-Pair-Share.

From the survey (Quiz 1)

covid	n
I'd be in favor of using such examples.	10
I'm neutral.	17
I'd prefer something different but I'd be ok with it.	2
I feel strongly that we should not use such examples.	2

```
quiz_responses %>%
  mutate(num_courses = str_count(courses, "[0-9][0-9][0-9]")) %>%
  ggplot(aes(x = num_courses)) +
    geom_bar() +
    scale_x_continuous(breaks = 1:10) +
    labs(x = "Number of previous courses")
```



-> My office hours **Wed 2:30-3:30 in Gold Lab** or by request.

Q & A

> Can we edit directly on GitHub?

Technically yes, but GitHub won't show plots, documentation, etc., so don't.

> What programming languages?

- We're using R (`#rstats`) because it has a big community, it's popular in industry, and is good pedagogically.
- Other players: Python (good at many things), Stata / SPSS / JMP (popular in some disciplines), Tableau / Microsoft PowerBI (popular in business)

> Are post-class quizzes part of grade?

They count towards "Prep and Participation" (10%).

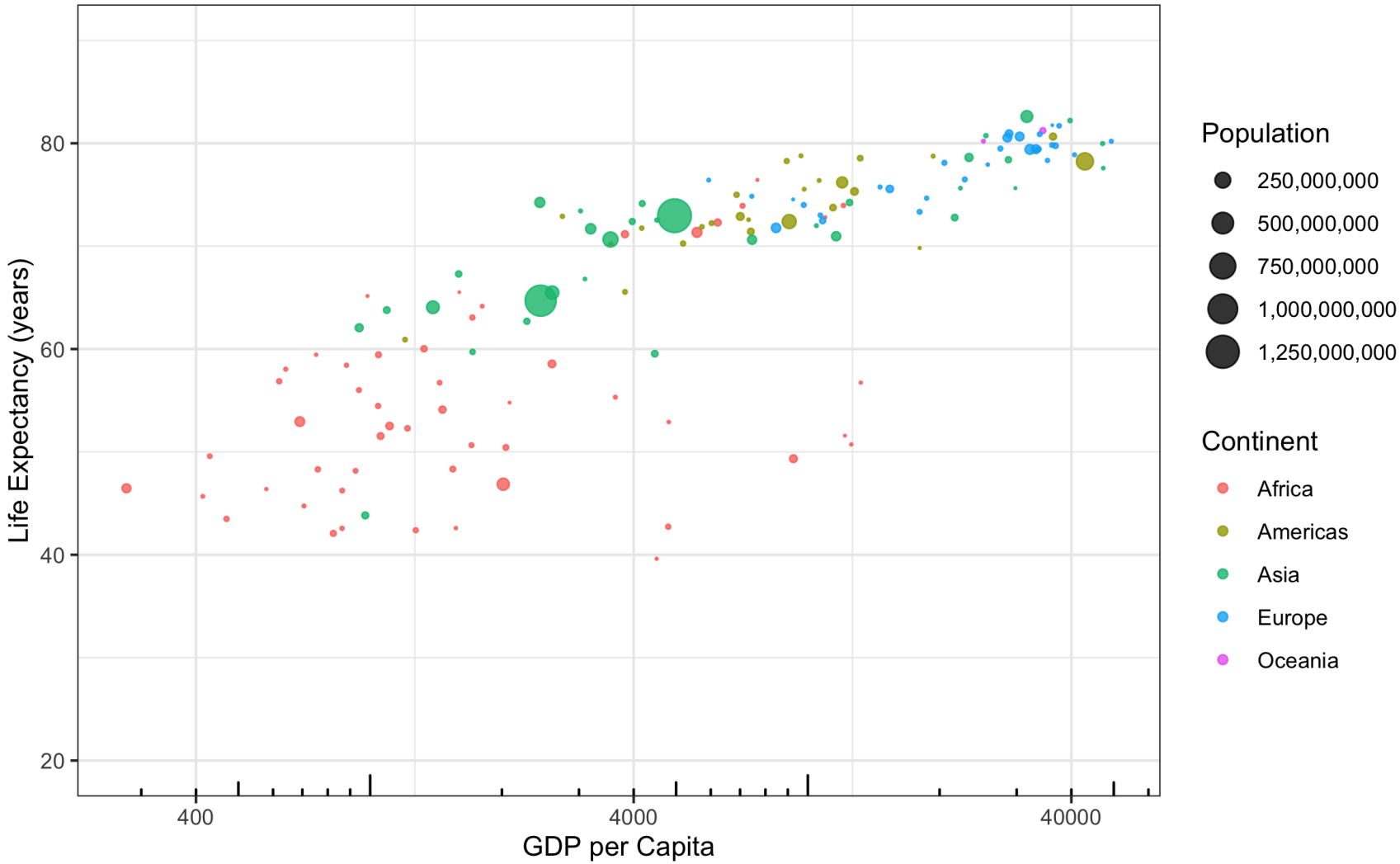
🧶 knit

✅ commit

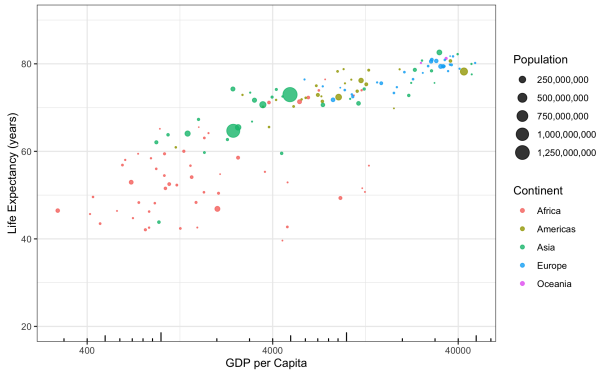
⬆️ push


BBC NEWS
Hans Rosling: 200 years in 4 minutes - ...
Watch later
Share
75 years
50 years
1948
income
$4 000
$40 000
BBC NEWS
Watch on YouTube

We'll make this chart

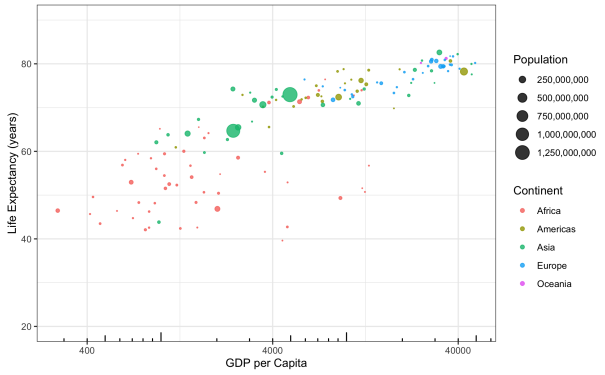


Composing a plot



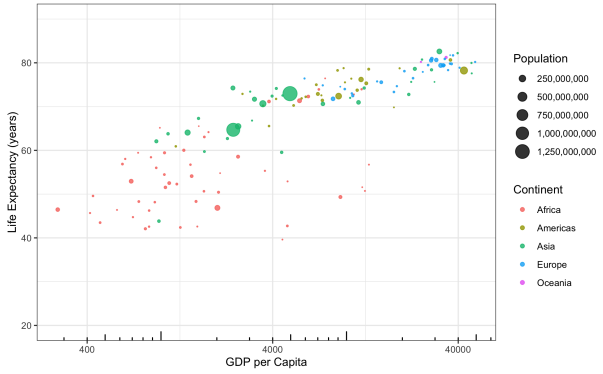
Composing a plot

- What's the data? "Each row is a _"

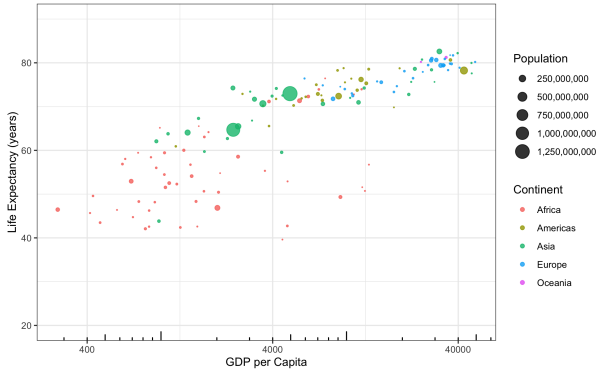


Composing a plot

- What's the data? "Each row is a _"
- What is the coordinate system? (What's x and y?)

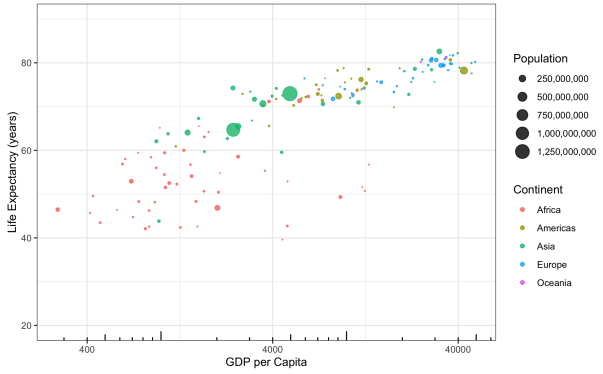


Composing a plot



- What's the data? "Each row is a _"
- What is the coordinate system? (What's x and y?)
- What graphical symbols are used? (dot? bar? line?)

Composing a plot



- What's the data? "Each row is a _"
- What is the coordinate system? (What's x and y?)
- What graphical symbols are used? (dot? bar? line?)
- What data variables are mapped to what visual cues (aesthetics)?
  - What *scales* are used? (Any transformations?)
  - What *guides* are shown? (What labels for values?)
- What labels and annotations are added?

```
gapminder
```

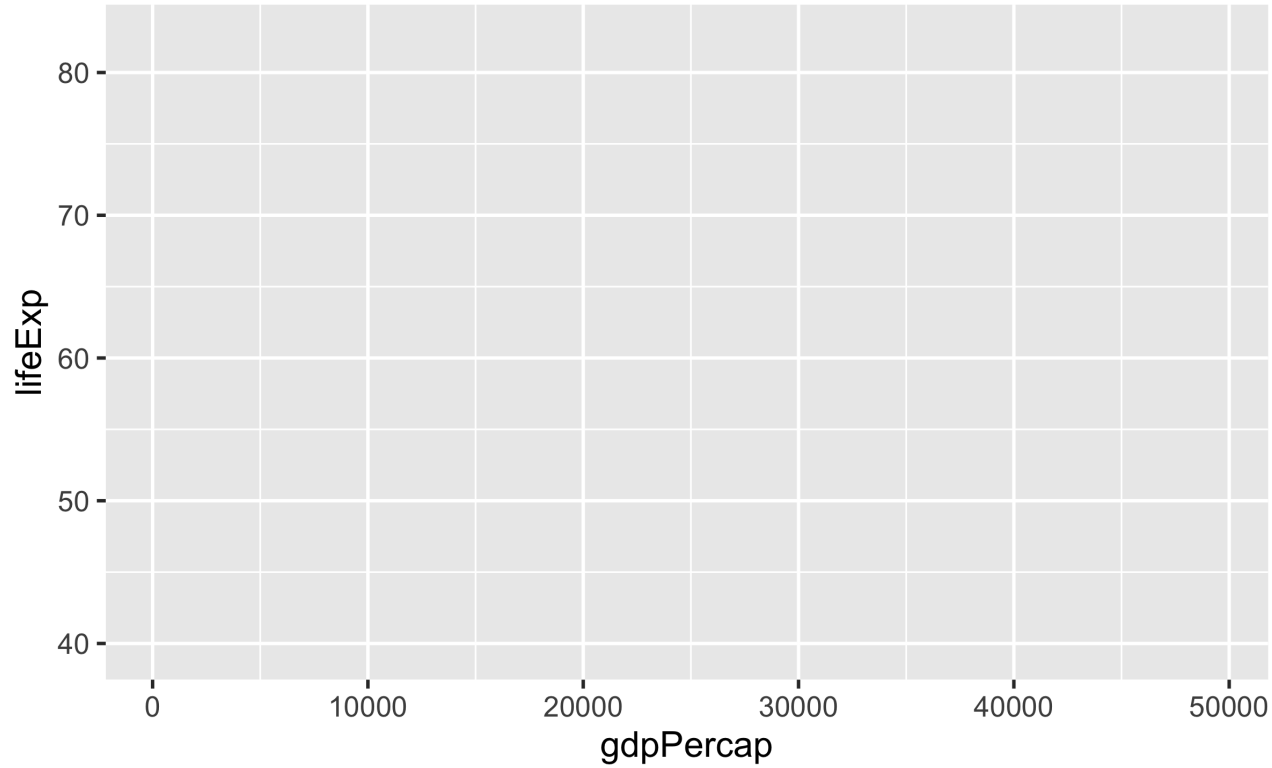
```
## # A tibble: 1,704 × 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
## # … with 1,698 more rows
```

```
gapminder %>%
  filter(year == 2007)
```

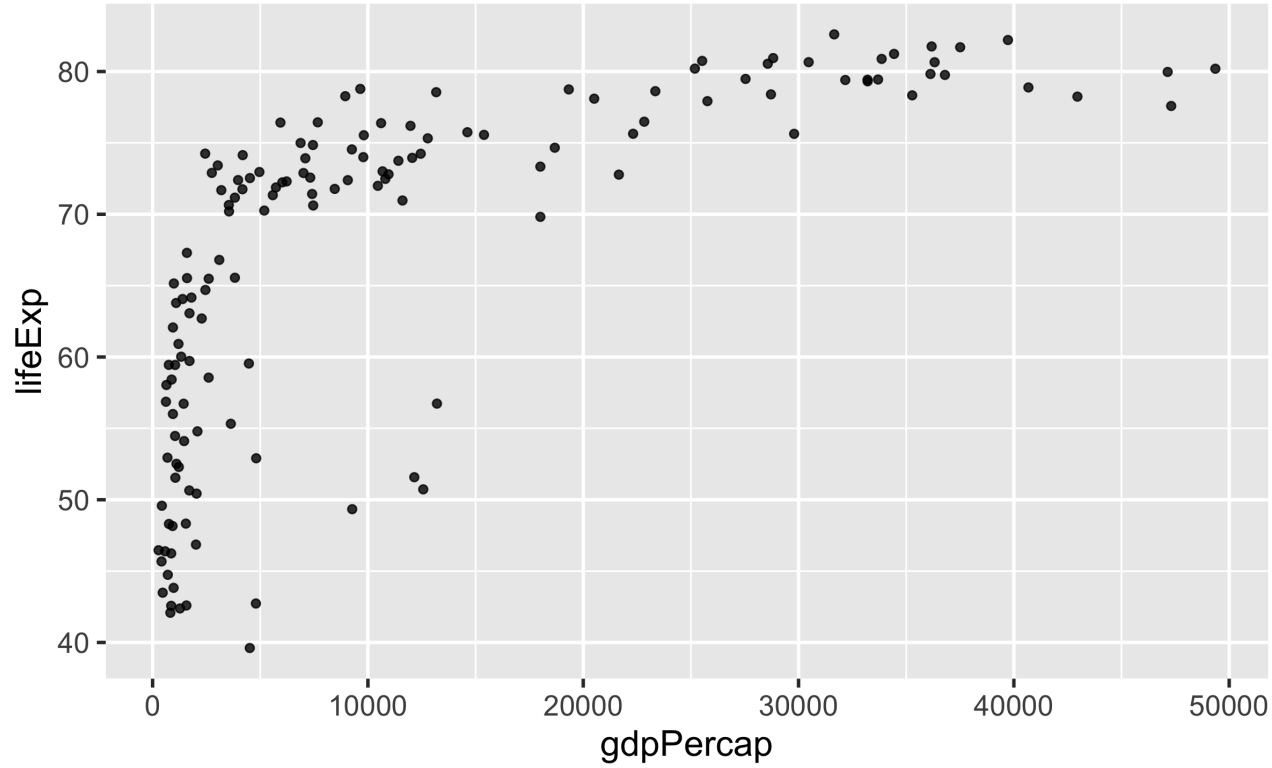
```
## # A tibble: 142 × 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       2007    43.8 31889923      975.
## 2 Albania     Europe     2007    76.4  3600523     5937.
## 3 Algeria     Africa     2007    72.3 33333216     6223.
## 4 Angola      Africa     2007    42.7 12420476     4797.
## 5 Argentina   Americas   2007    75.3 40301927    12779.
## 6 Australia   Oceania    2007    81.2 20434176    34435.
## # … with 136 more rows
```

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot()
```

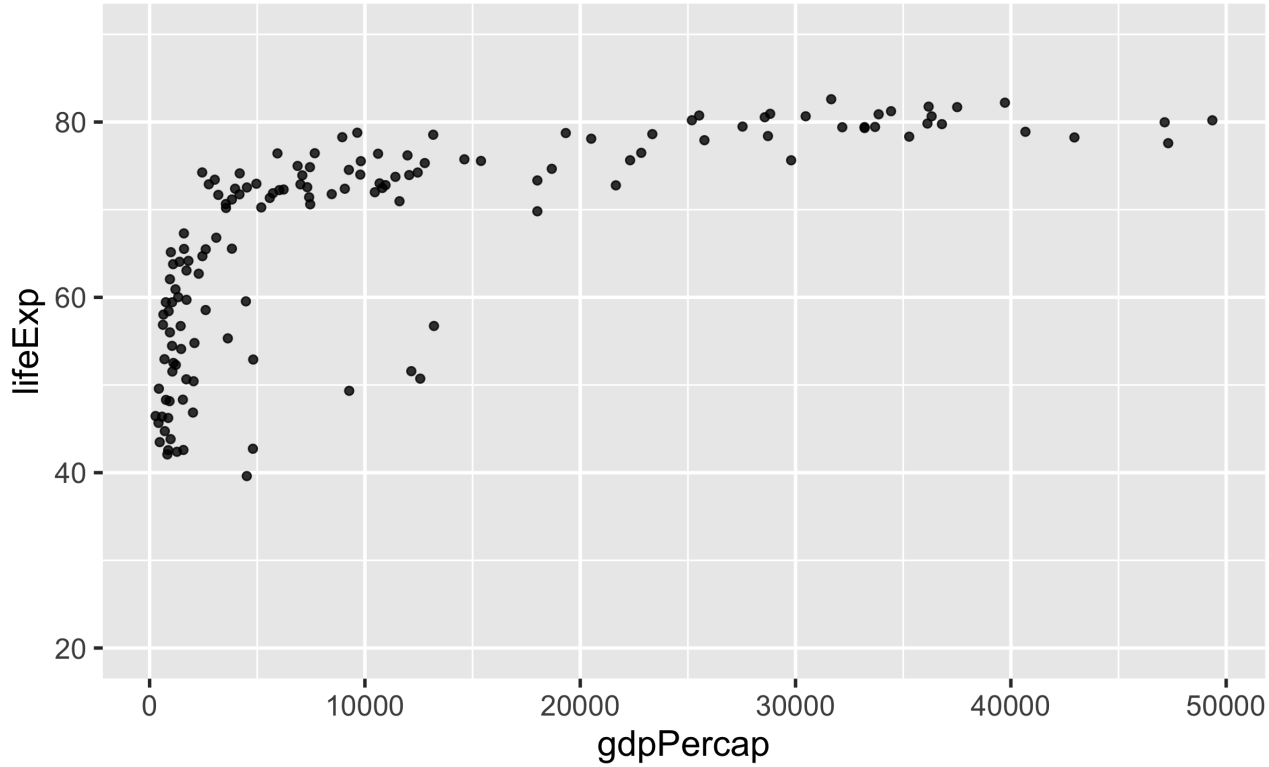
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp)
```



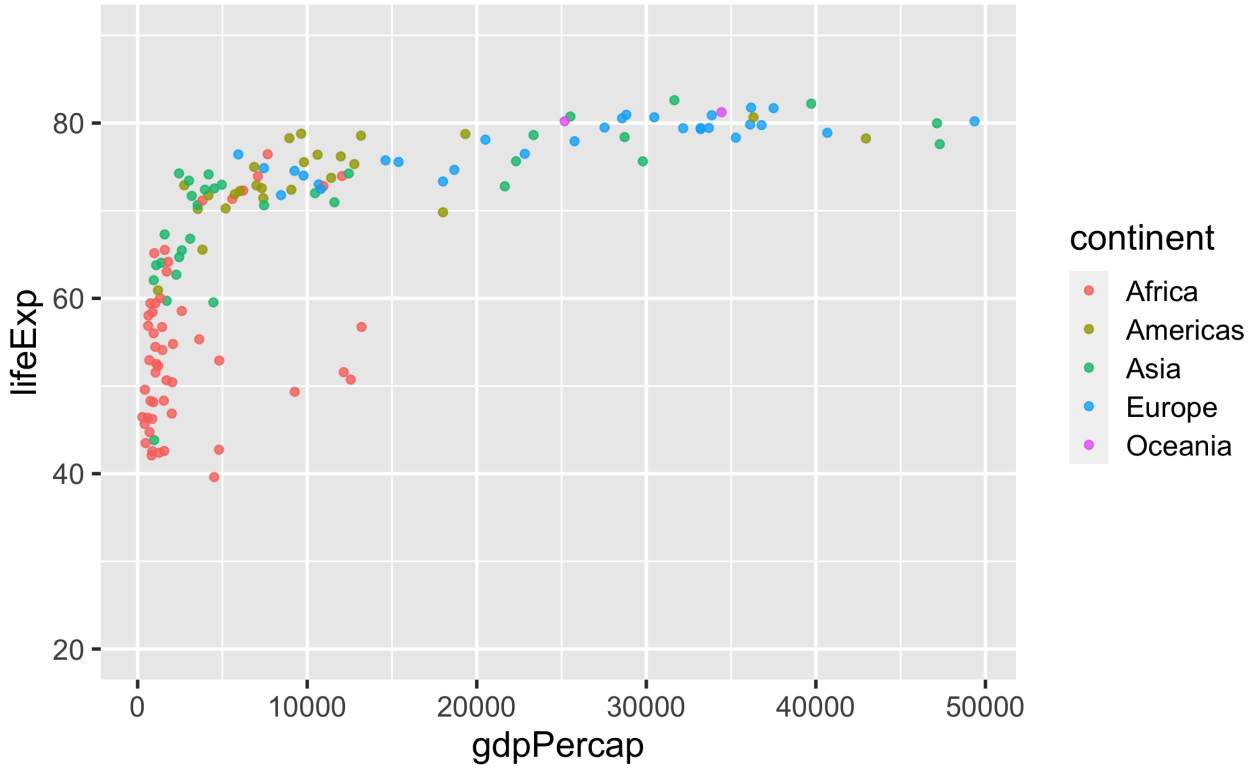
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8)
```



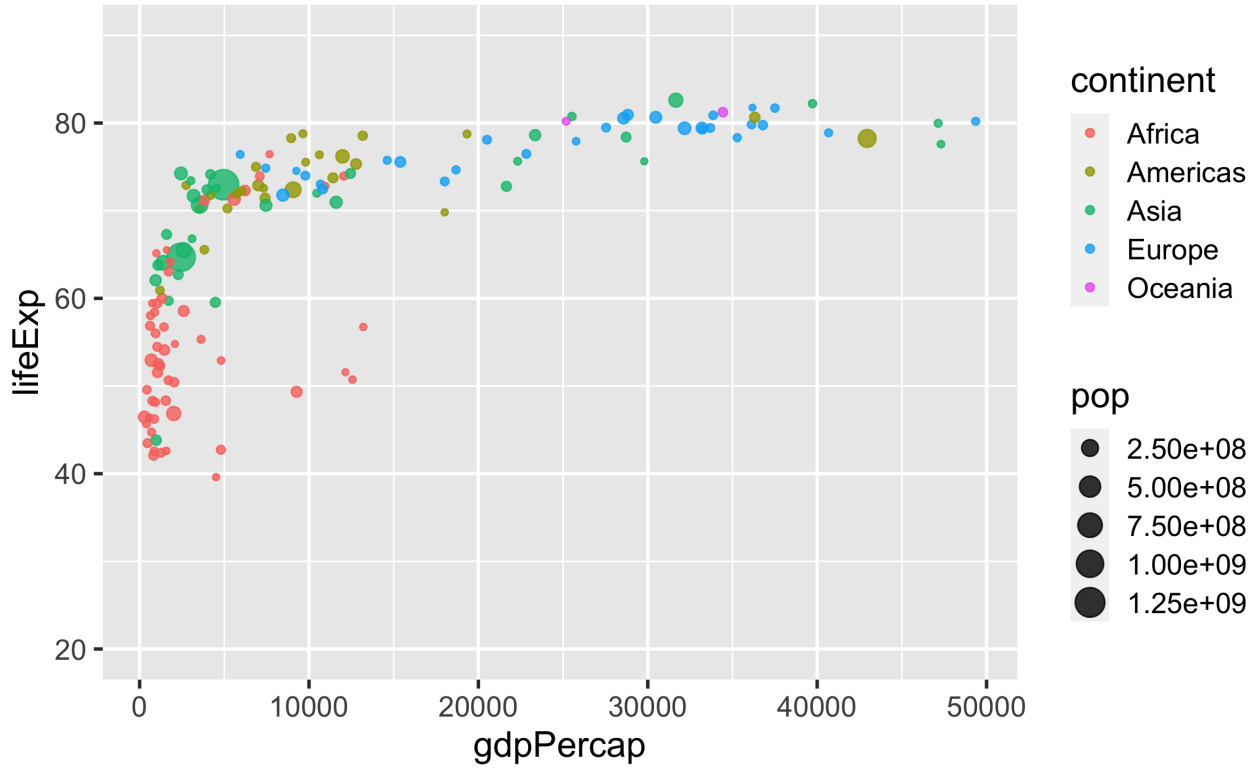
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
```



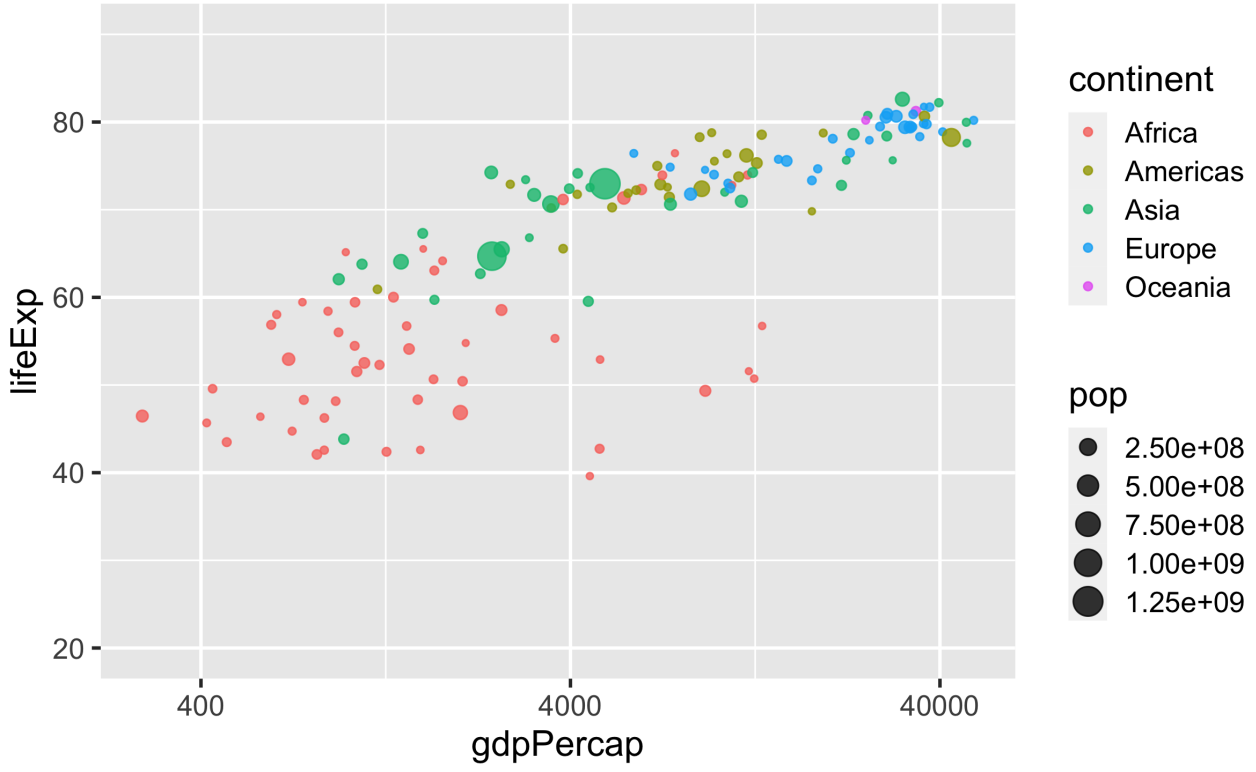
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent)
```



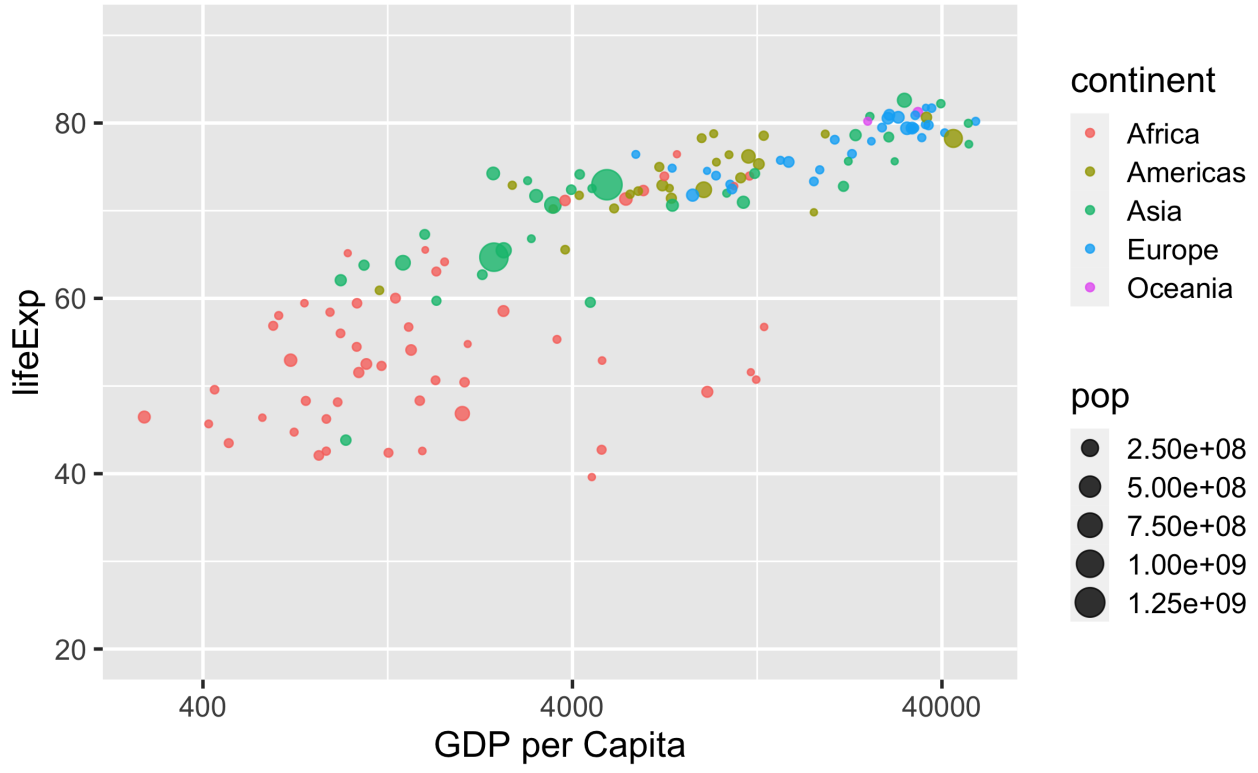
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop)
```



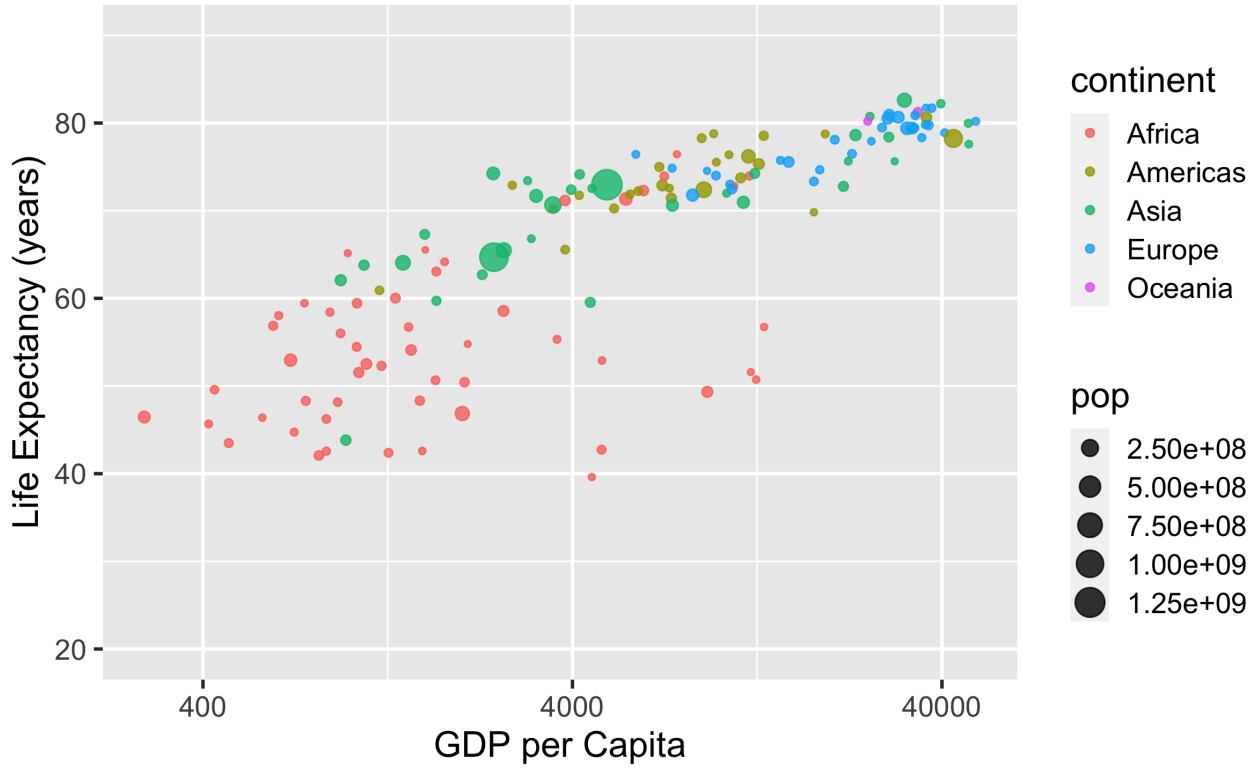
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10")
```



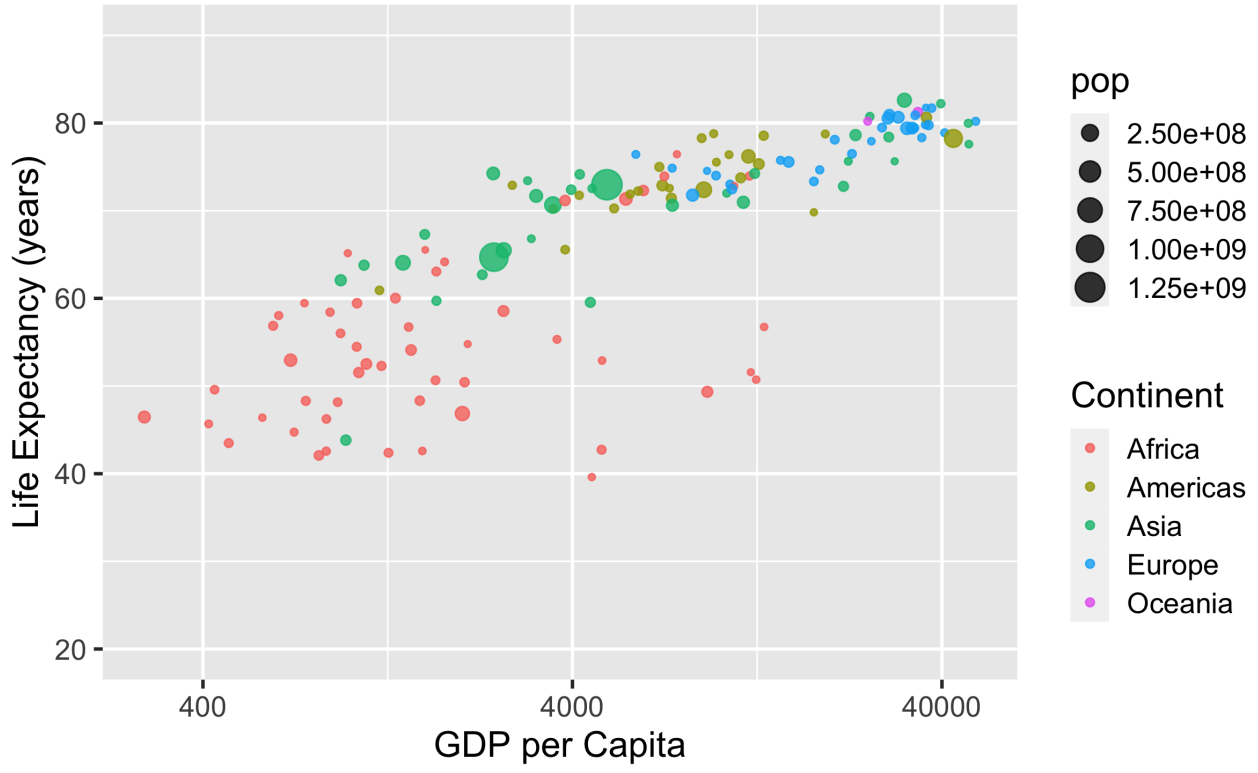
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10") +
    labs(x = "GDP per Capita")
```



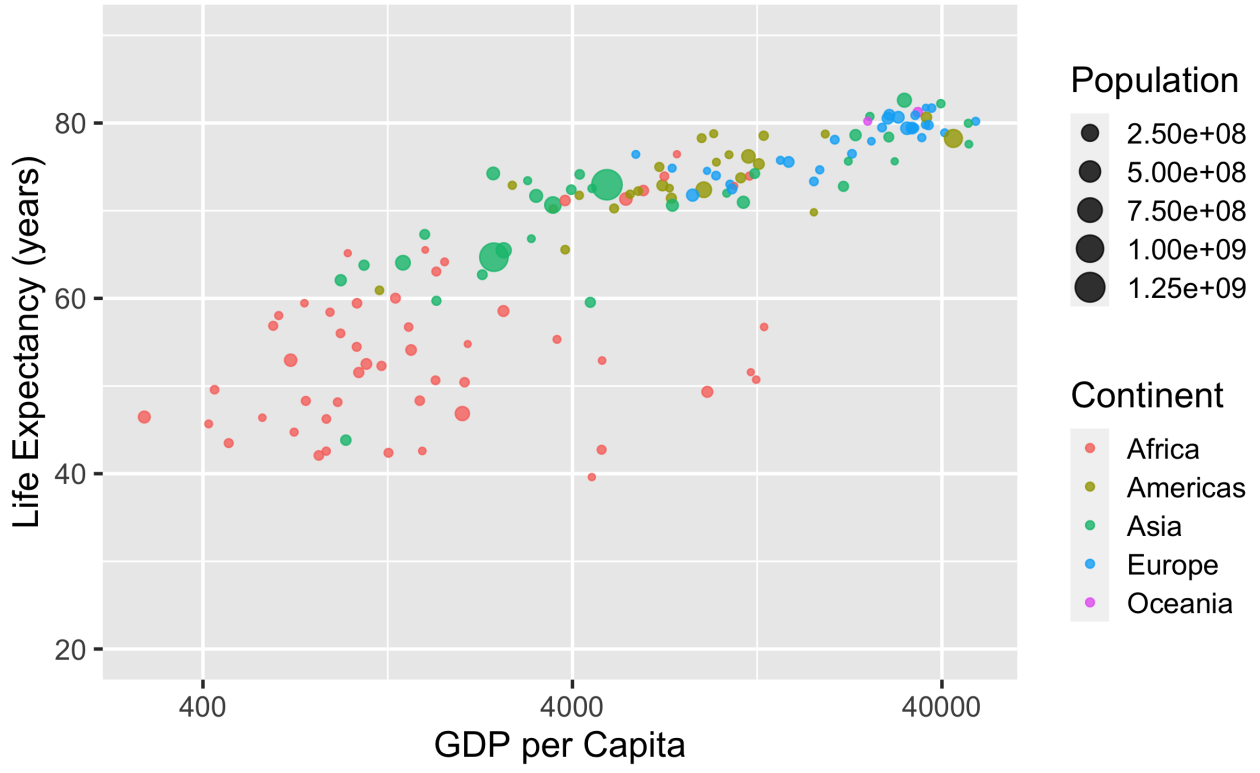
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10") +
    labs(x = "GDP per Capita") +
    labs(y = "Life Expectancy (years)"
```



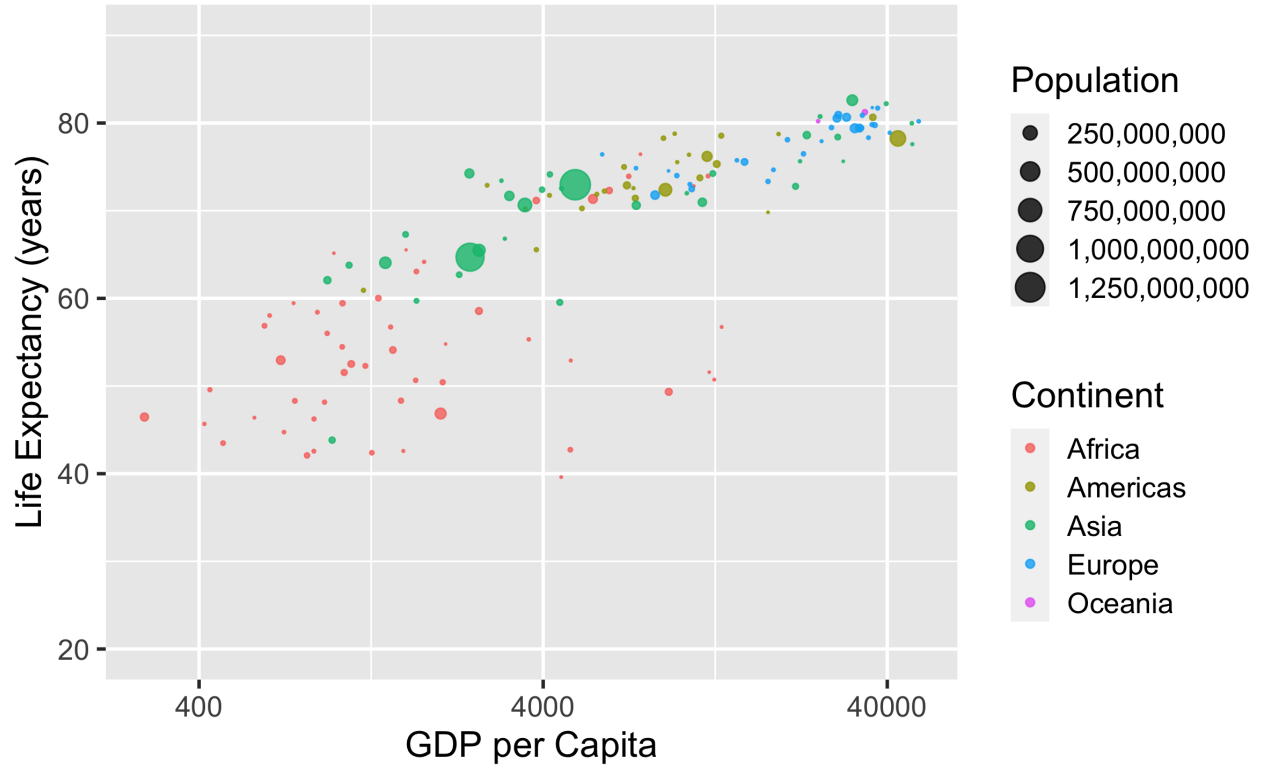
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10") +
    labs(x = "GDP per Capita") +
    labs(y = "Life Expectancy (years)"
    labs(color = "Continent")
```



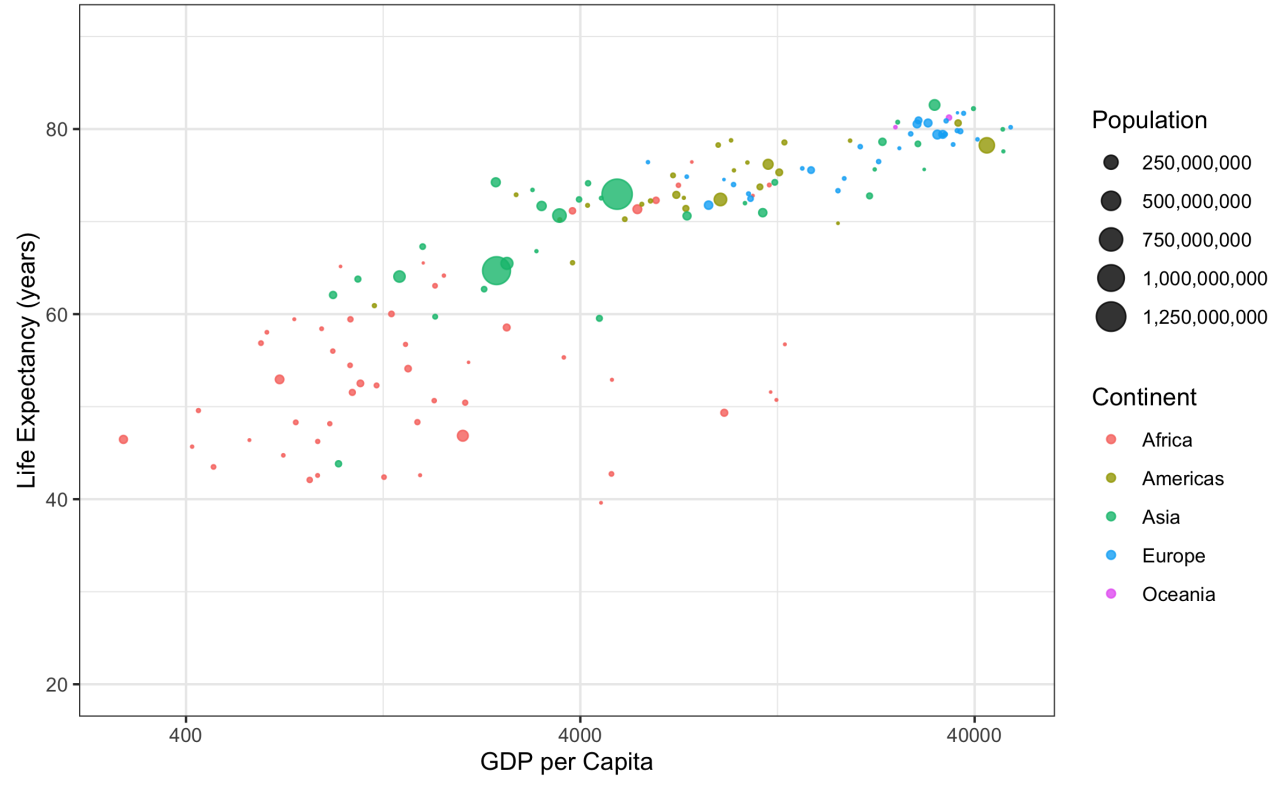
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10") +
    labs(x = "GDP per Capita") +
    labs(y = "Life Expectancy (years)"
    labs(color = "Continent") +
    labs(size = "Population")
```



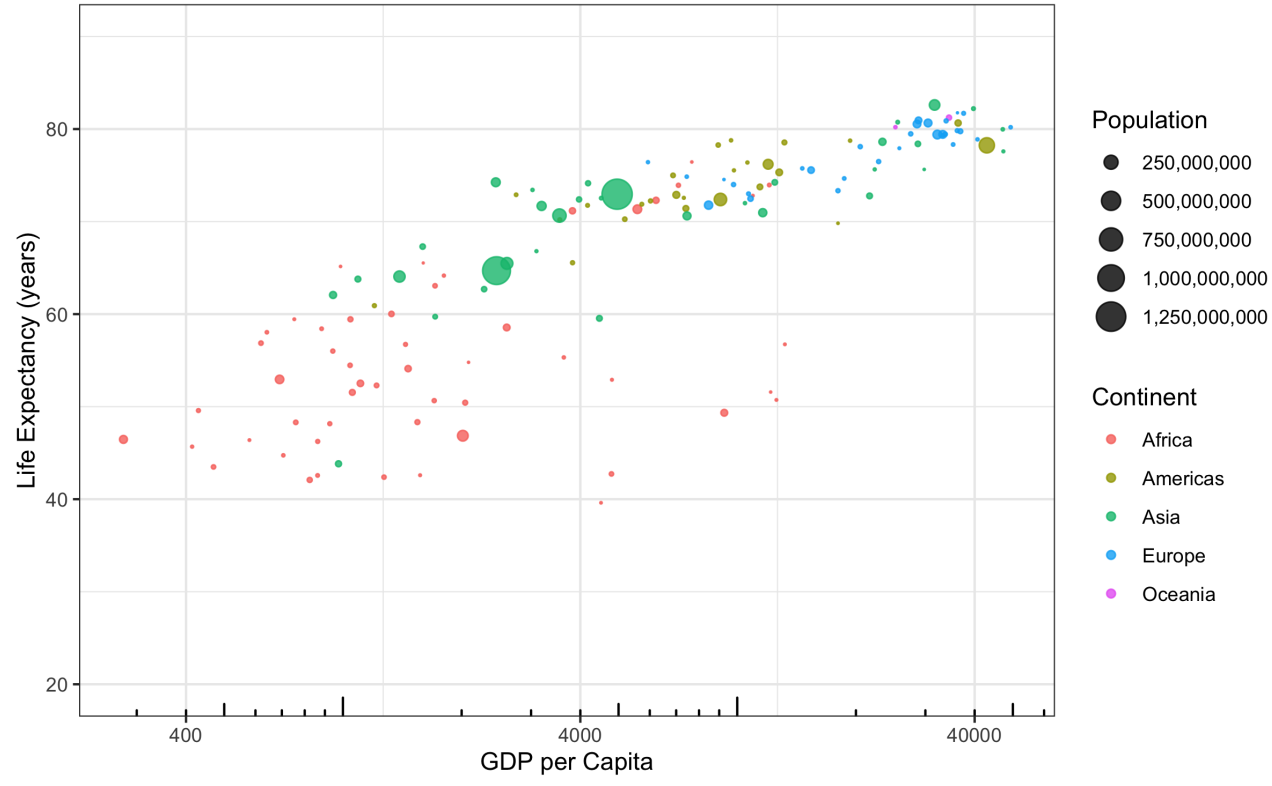
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10") +
    labs(x = "GDP per Capita") +
    labs(y = "Life Expectancy (years)"
    labs(color = "Continent") +
    labs(size = "Population") +
    scale_size_area(
      labels = label_comma())
```



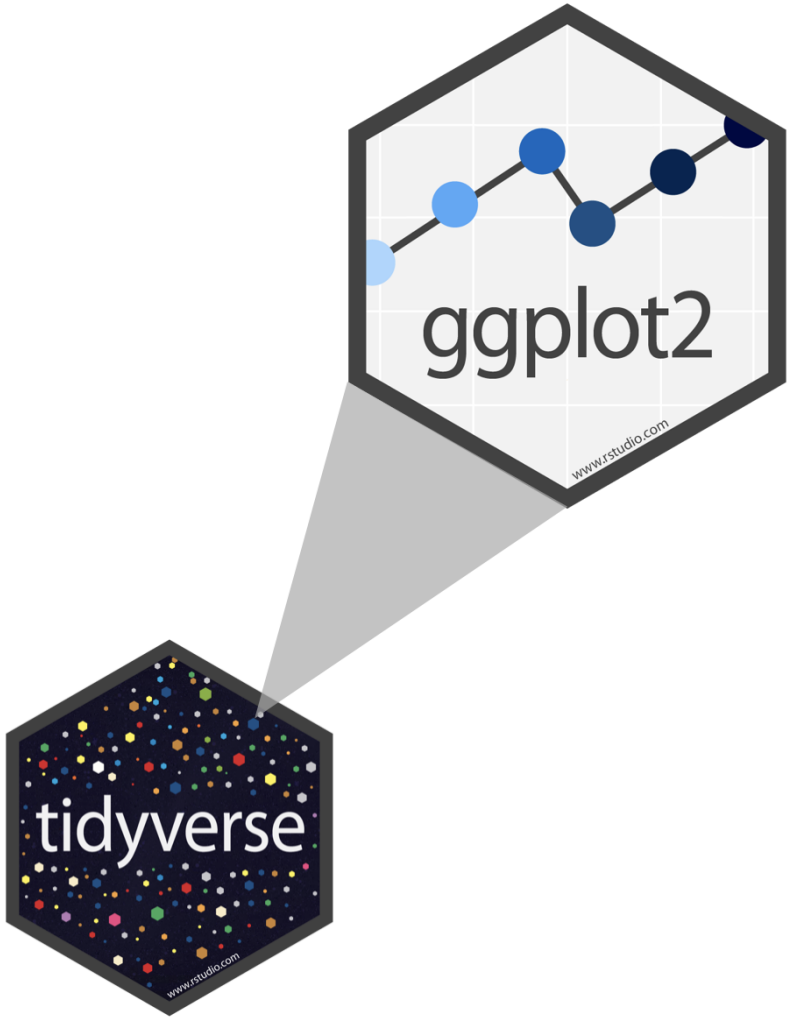
```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10") +
    labs(x = "GDP per Capita") +
    labs(y = "Life Expectancy (years)"
    labs(color = "Continent") +
    labs(size = "Population") +
    scale_size_area(
      labels = label_comma()) +
    theme_bw()
```



```
gapminder %>%
  filter(year == 2007) %>%
  ggplot() +
    aes(x = gdpPercap, y = lifeExp) +
    geom_point(alpha = .8) +
    coord_cartesian(ylim = c(20, 90))
    aes(color = continent) +
    aes(size = pop) +
    scale_x_continuous(
      breaks = c(400, 4000, 40000),
      trans = "log10") +
    labs(x = "GDP per Capita") +
    labs(y = "Life Expectancy (years)"
    labs(color = "Continent") +
    labs(size = "Population") +
    scale_size_area(
      labels = label_comma()) +
    theme_bw() +
    annotation_logticks(sides = "b")
```



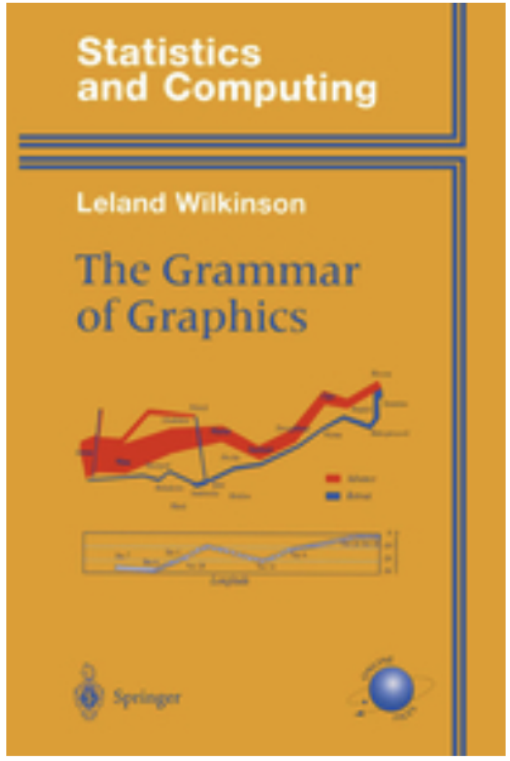
ggplot2 ∈ tidyverse

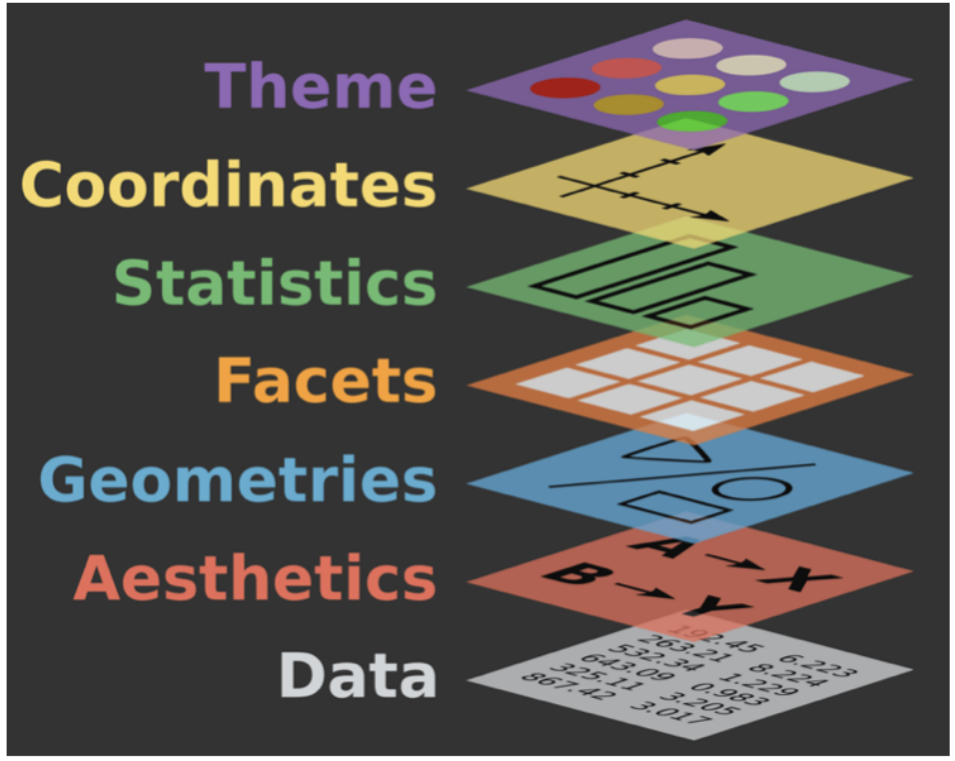


- **ggplot2** is tidyverse's data visualization package
- The gg in "ggplot2" stands for Grammar of Graphics
- It is inspired by the book **Grammar of Graphics** by Leland Wilkinson

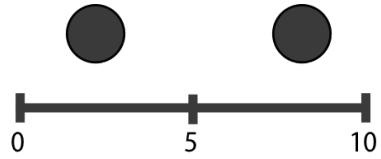
Grammar of Graphics

Concisely describe the components of a graphic

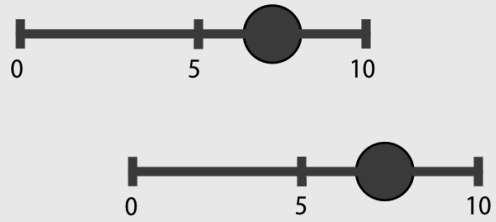




Source: BloggoType



Position on
a common scale

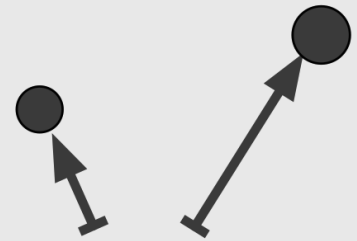


Position on
unaligned
scales

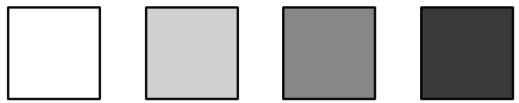
Length

Tilt or Angle

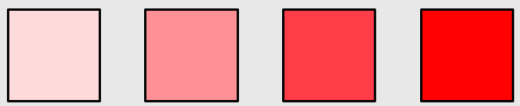
Area (2D as size)



Depth
(3D as position)

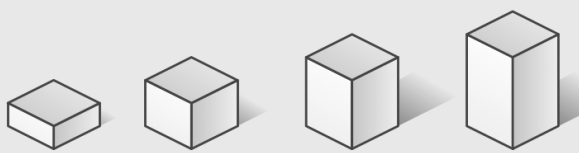


Color luminance
or brightness



Color saturation
or intensity

Curvature



Volume
(3D as size)