

Class05: Data visualization with ggplot

Yi-Hung Lee (PID: A16587141)

Table of contents

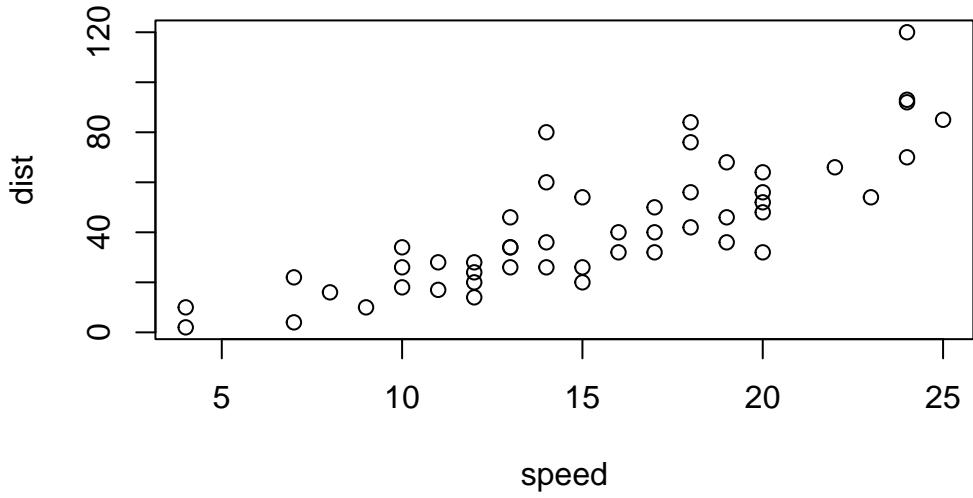
ggplot introduction

Today we will have our first play with the **ggplot2** package - one of the most popular graphics packages on the planet.

There are many plotting systems in R. These include so-called “*base*” plot/graphs.

Base plot is generally rather short code and dull plots - but it is always for you and is fast for big dataset.

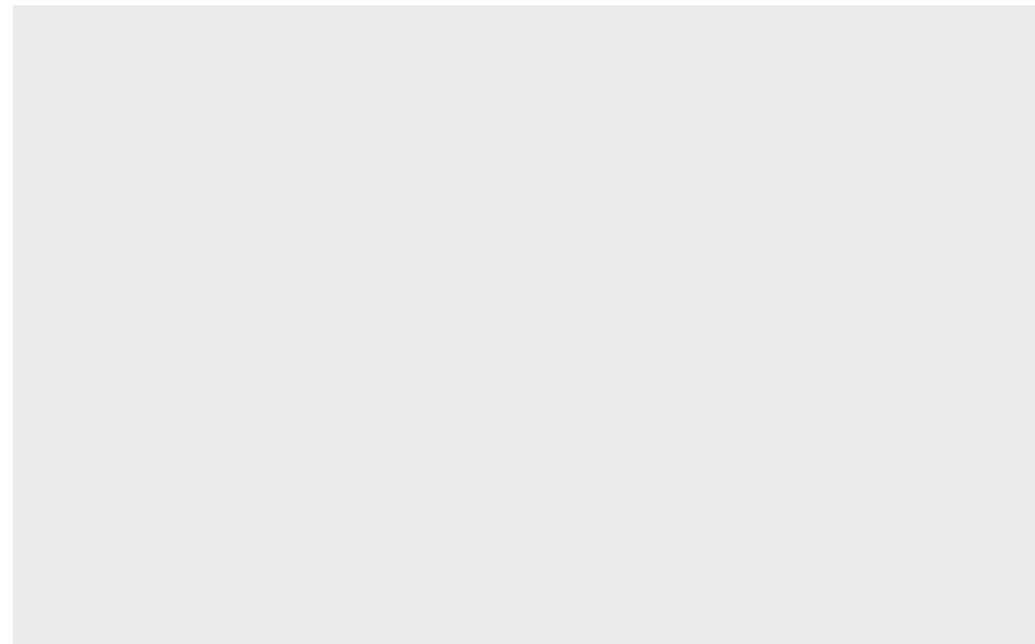
```
plot(cars)
```



Try ggplot: Need to install `ggplot` package, use function `install.packages()`. Already installed.

Every time I want to use a package, I need to call the library by function `library(ggplot2)`

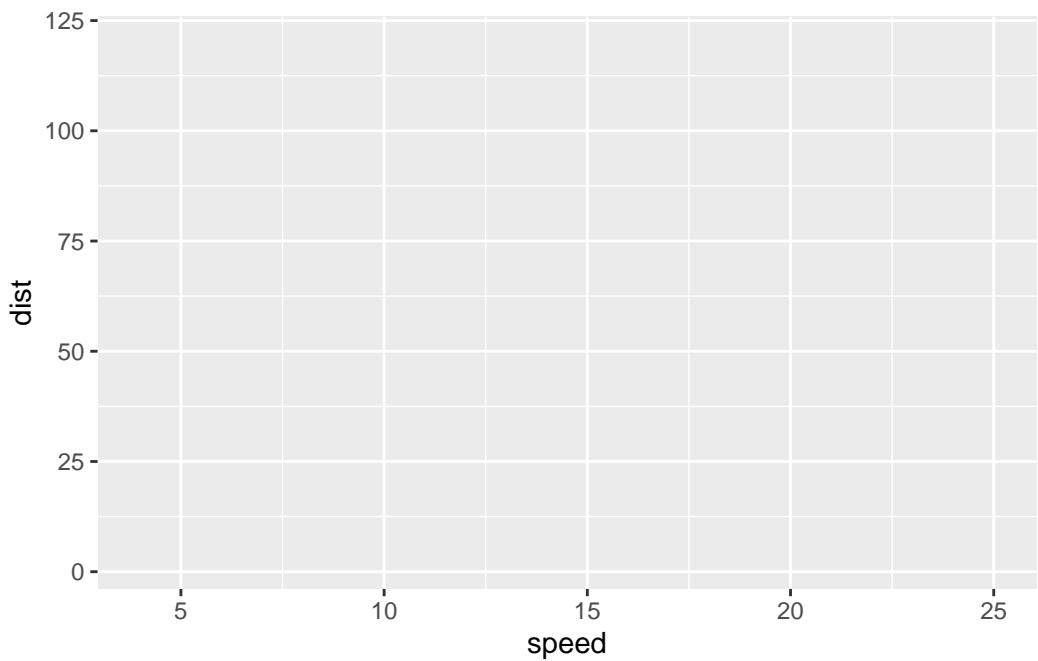
```
library(ggplot2)  
ggplot(cars)
```



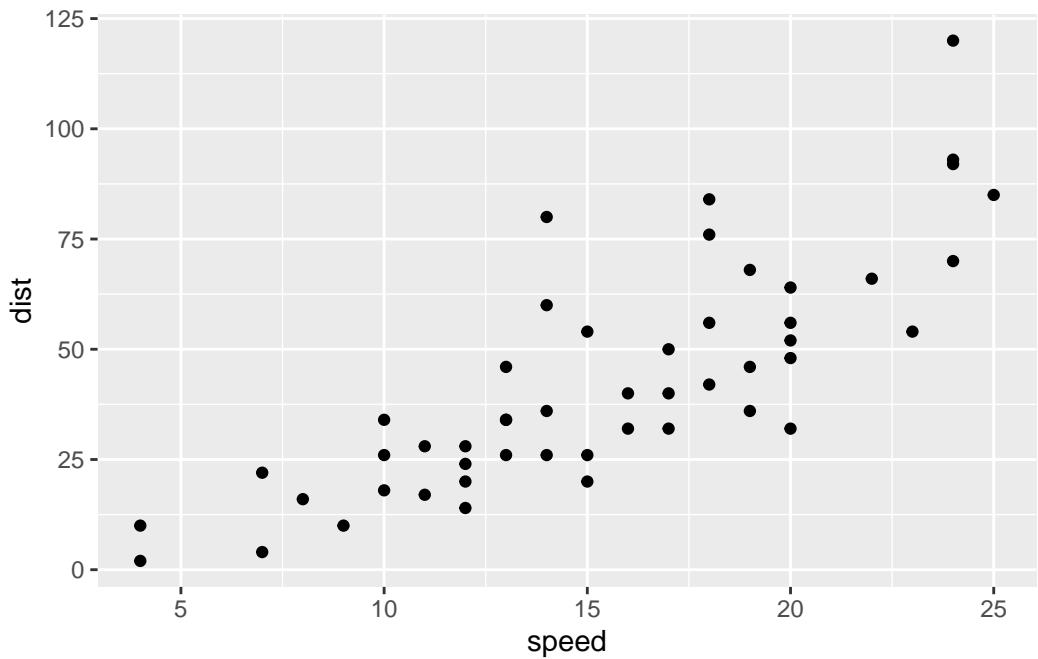
Every **ggplot** has at least 3 things:

- **data**: the data. frame with the data you want to plot
- **aes**: the aesthetic mapping of the data to the plot
- **geom**: how do you want the plot to look

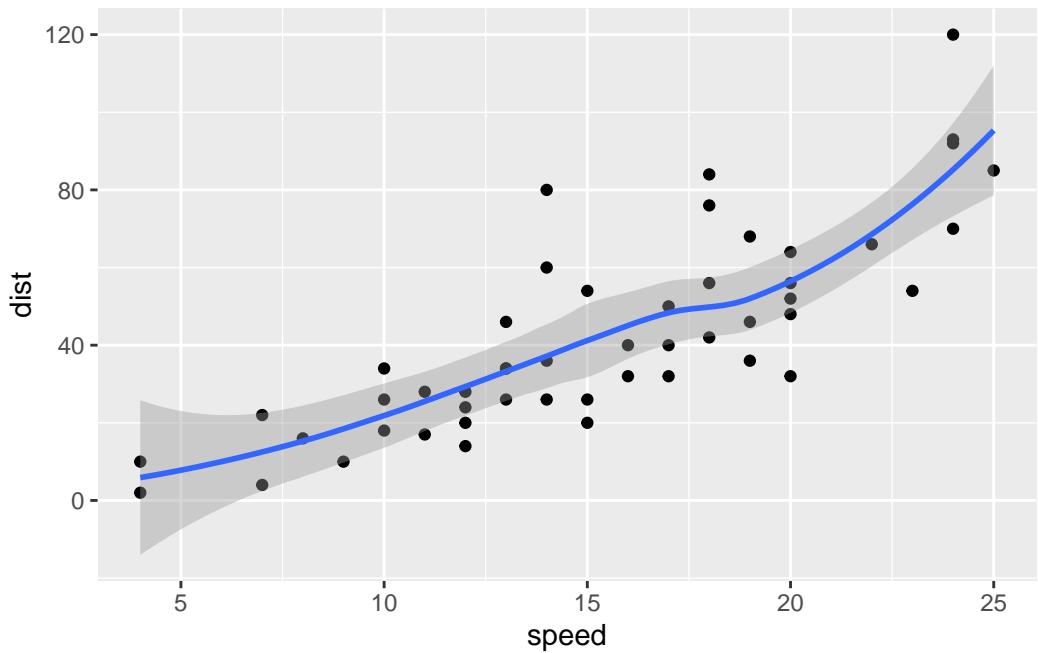
```
ggplot(cars) +  
  aes(x=speed, y=dist)
```



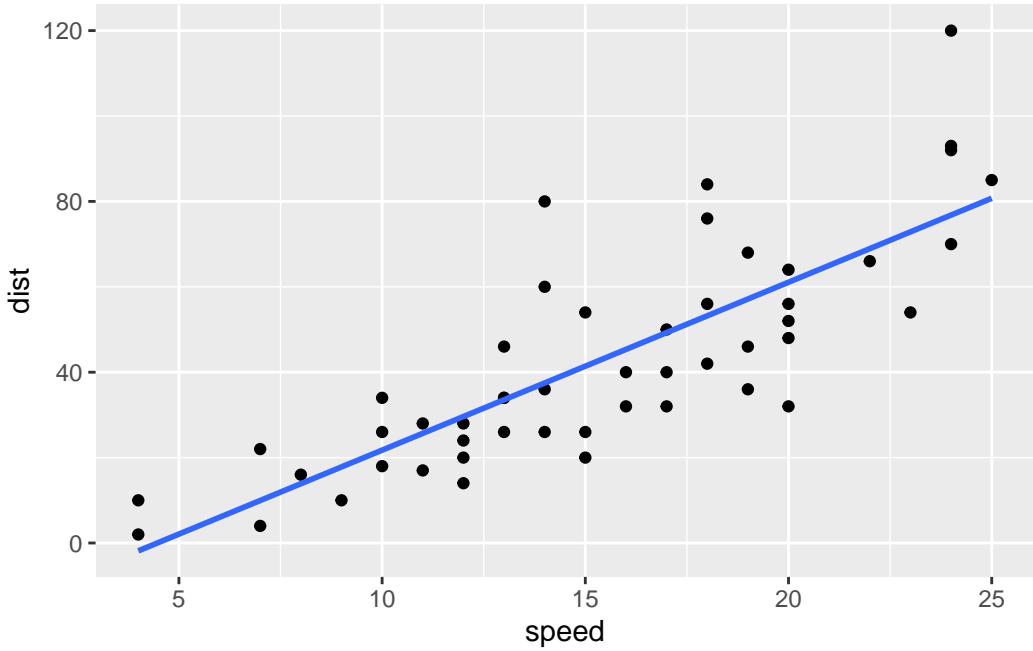
```
ggplot(cars) +  
  aes(x = speed, y = dist) +  
  geom_point()
```



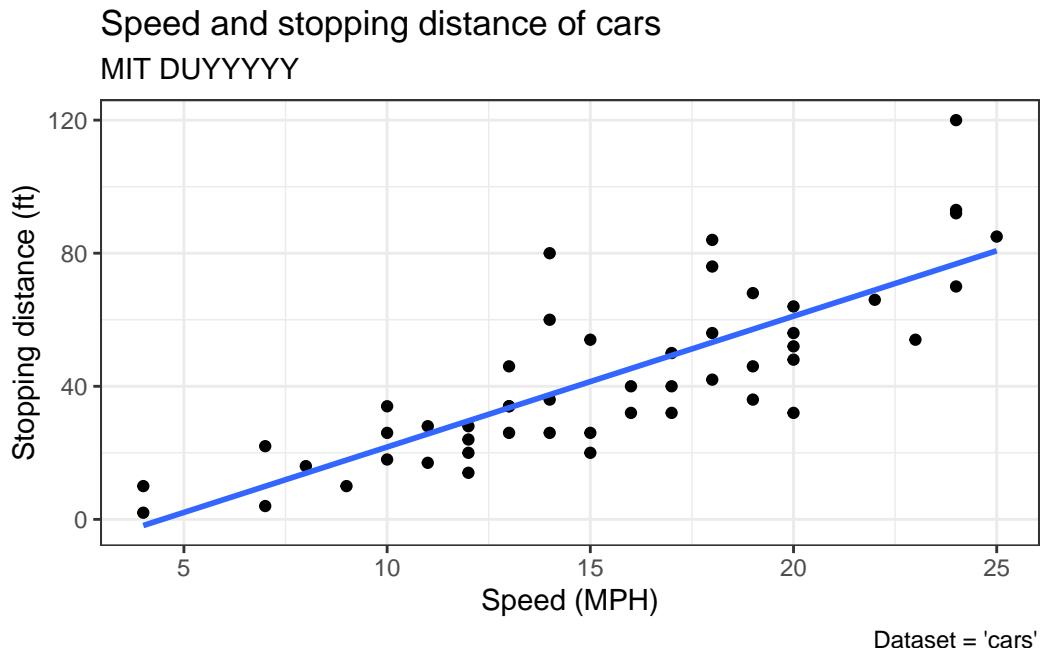
```
ggplot(cars) +  
  aes(x = speed, y = dist) +  
  geom_point() +  
  geom_smooth()  
  
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



```
ggplot(cars) +  
  aes(x = speed, y = dist) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE )  
  
`geom_smooth()` using formula = 'y ~ x'
```



```
bp <- ggplot(cars) +  
  aes(x = speed, y = dist) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE )+  
  labs(title = "Speed and stopping distance of cars",  
    x = "Speed (MPH)",  
    y = "Stopping distance (ft)",  
    subtitle = "MIT DUYYYYY",  
    caption = "Dataset = 'cars'") +  
  theme_bw()  
  
bp  
  
`geom_smooth()` using formula = 'y ~ x'
```



A more complicated scatter plot

Here we make a plot of gene expression data:

```
url <- "https://bioboot.github.io/bimm143_S20/class-material/up_down_expression.txt"
genes <- read.delim(url)
head(genes)
```

| | Gene | Condition1 | Condition2 | State |
|---|------------|------------|------------|------------|
| 1 | A4GNT | -3.6808610 | -3.4401355 | unchanging |
| 2 | AAAS | 4.5479580 | 4.3864126 | unchanging |
| 3 | AASDH | 3.7190695 | 3.4787276 | unchanging |
| 4 | AATF | 5.0784720 | 5.0151916 | unchanging |
| 5 | AATK | 0.4711421 | 0.5598642 | unchanging |
| 6 | AB015752.4 | -3.6808610 | -3.5921390 | unchanging |

```
nrow(genes)
```

```
[1] 5196
```

```

colnames(genes)

[1] "Gene"          "Condition1" "Condition2" "State"

ncol(genes)

[1] 4

table(genes$State)

down unchanging      up
72      4997       127

round(table(genes$State == 'up') / nrow(genes) * 100, 2)

FALSE  TRUE
97.56 2.44

n.gene <- nrow(genes)
n.up <- sum(genes$State == 'up')

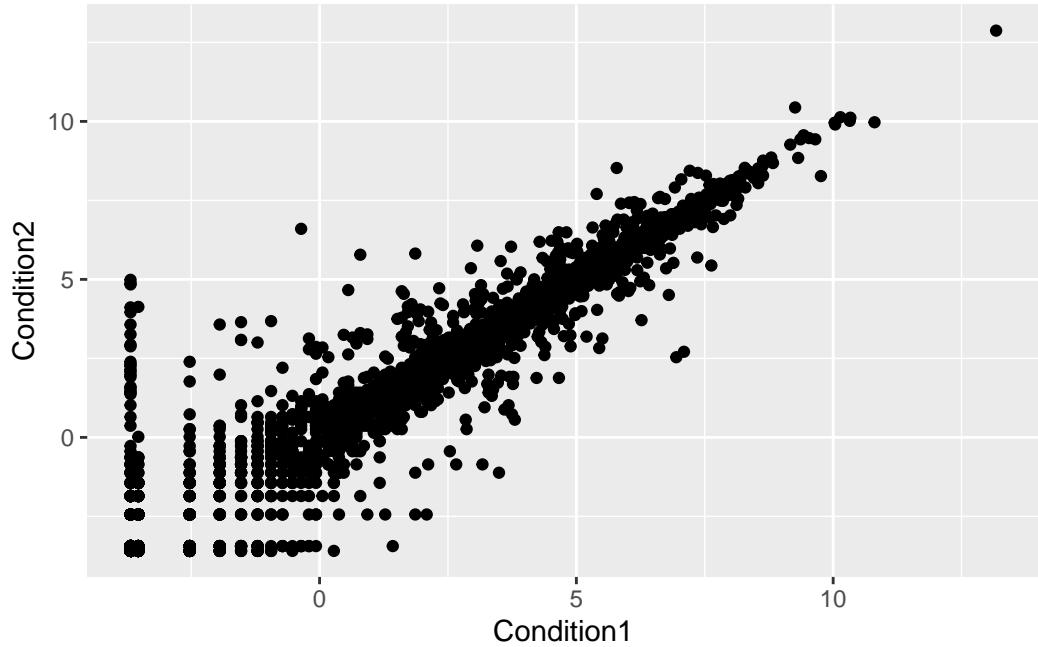
up.percent <- n.up/n.gene *100
round(up.percent, 2)

[1] 2.44

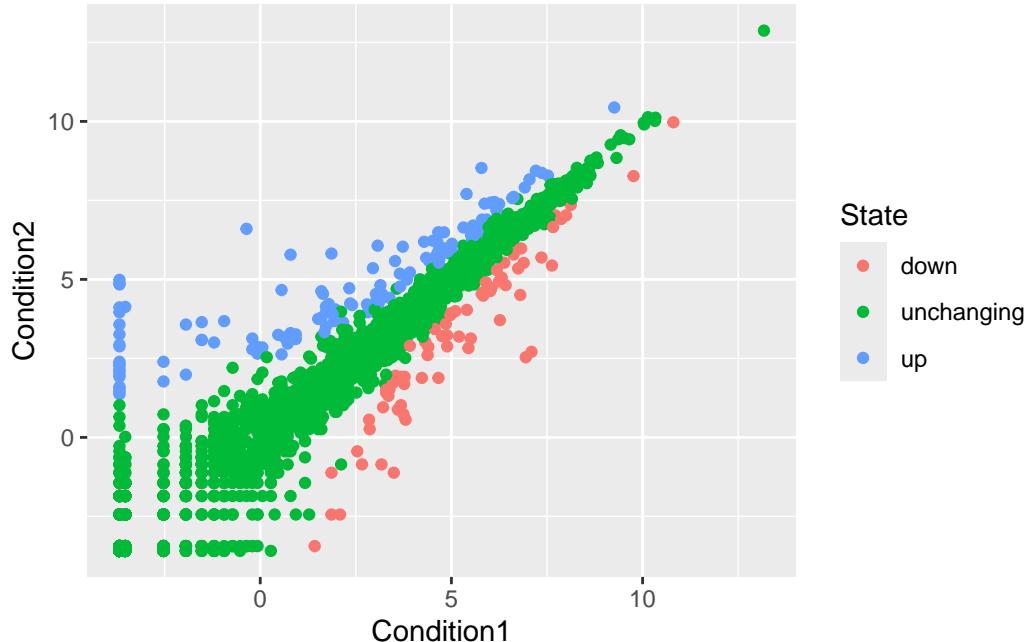
n <-
  ggplot(genes) +
  aes(x = Condition1, y = Condition2) +
  geom_point()

n

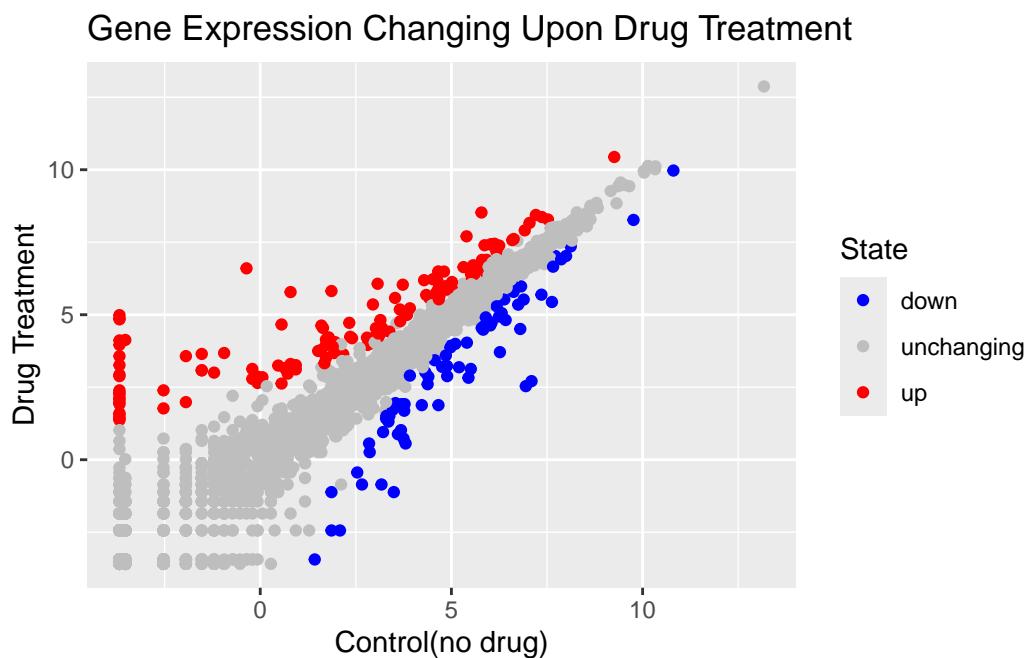
```



```
gene_plot <- n + aes(col = State)  
gene_plot
```



```
gene_plot + scale_color_manual(values = c("blue", "gray", "red")) +
  labs(x = "Control(no drug)", y = "Drug Treatment", title = "Gene Expression Changing Upon Drug Treatment")
```



Exploring the gapminder dataset

```
library(gapminder)
# File location online
url <- "https://raw.githubusercontent.com/jennybc/gapminder/master/inst/extdata/gapminder.csv"
gapminder <- read.delim(url)

head(gapminder)
```

| | country | continent | year | lifeExp | pop | gdpPercap |
|---|-------------|-----------|------|---------|----------|-----------|
| 1 | Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| 2 | Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| 3 | Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| 4 | Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| 5 | Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| 6 | Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |

Q. How many continent are there? Use function `unique()`

```
num <- unique(gapminder$continent)
length(num)
```

[1] 5

Q. How many countries are there?

```
num <- unique(gapminder$country)
length(num)
```

[1] 142

```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

`filter`, `lag`

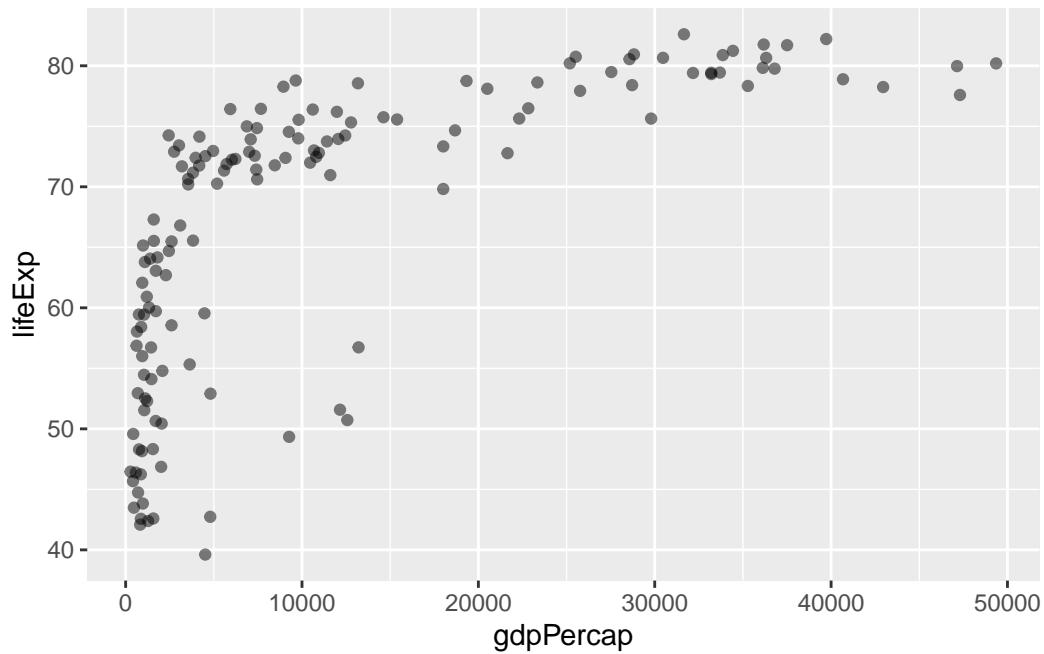
```
The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

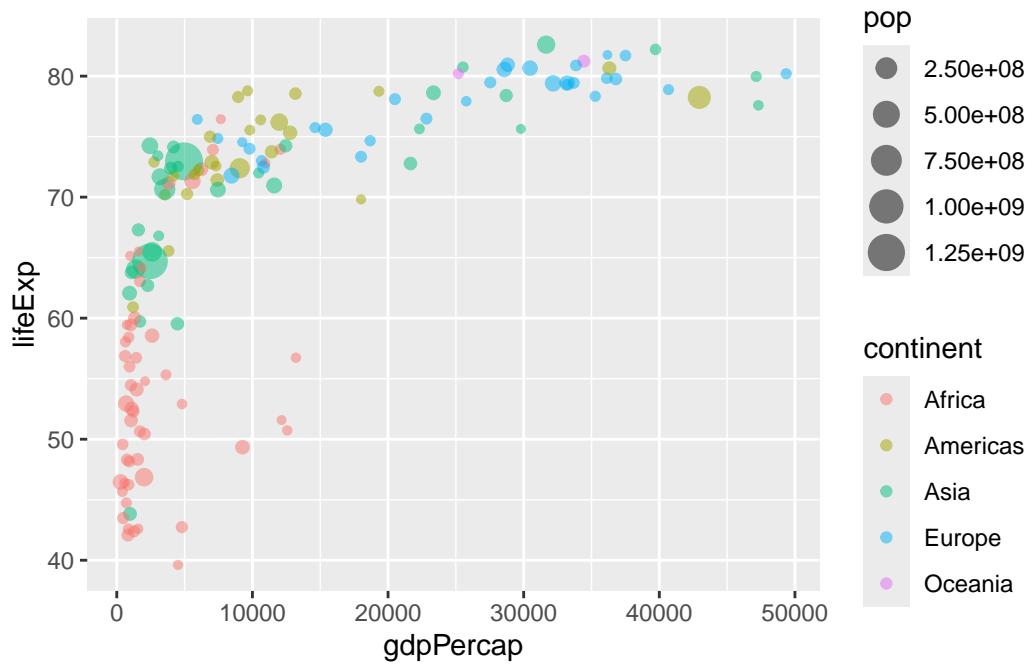
```
gapminder_2007 <- gapminder %>% filter(year==2007)  
head(gapminder_2007)
```

| | country | continent | year | lifeExp | pop | gdpPercap |
|---|-------------|-----------|------|---------|----------|------------|
| 1 | Afghanistan | Asia | 2007 | 43.828 | 31889923 | 974.5803 |
| 2 | Albania | Europe | 2007 | 76.423 | 3600523 | 5937.0295 |
| 3 | Algeria | Africa | 2007 | 72.301 | 33333216 | 6223.3675 |
| 4 | Angola | Africa | 2007 | 42.731 | 12420476 | 4797.2313 |
| 5 | Argentina | Americas | 2007 | 75.320 | 40301927 | 12779.3796 |
| 6 | Australia | Oceania | 2007 | 81.235 | 20434176 | 34435.3674 |

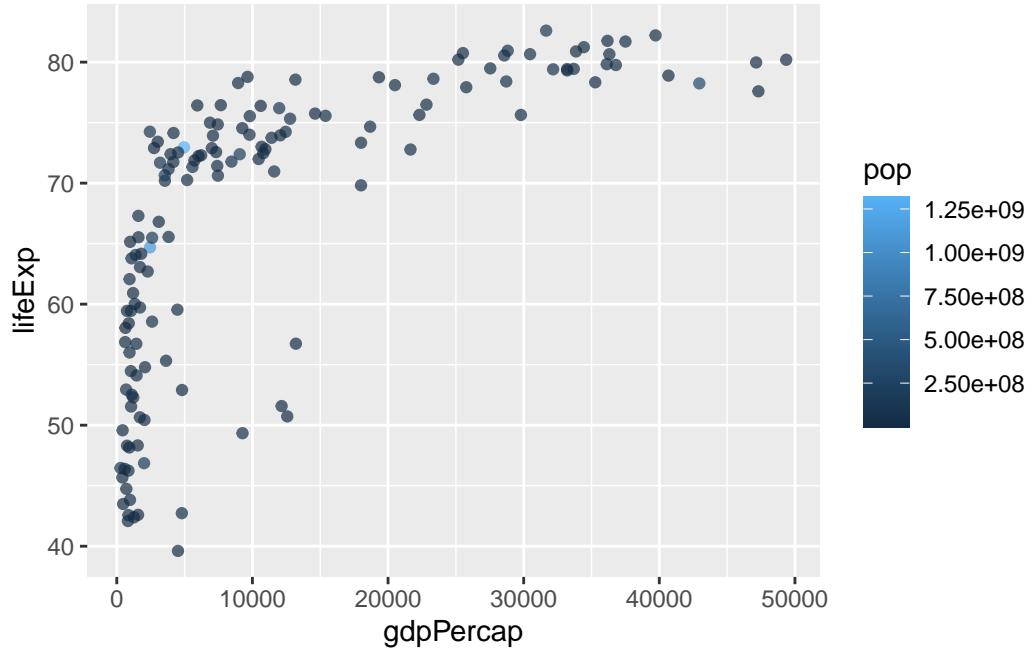
```
library(ggplot2)  
ggplot(gapminder_2007) +  
  aes(x = gdpPercap, y = lifeExp) +  
  geom_point(alpha=0.5)
```



```
ggplot(gapminder_2007) +
  aes(x = gdpPercap, y = lifeExp, color = continent, size = pop) +
  geom_point(alpha=0.5)
```

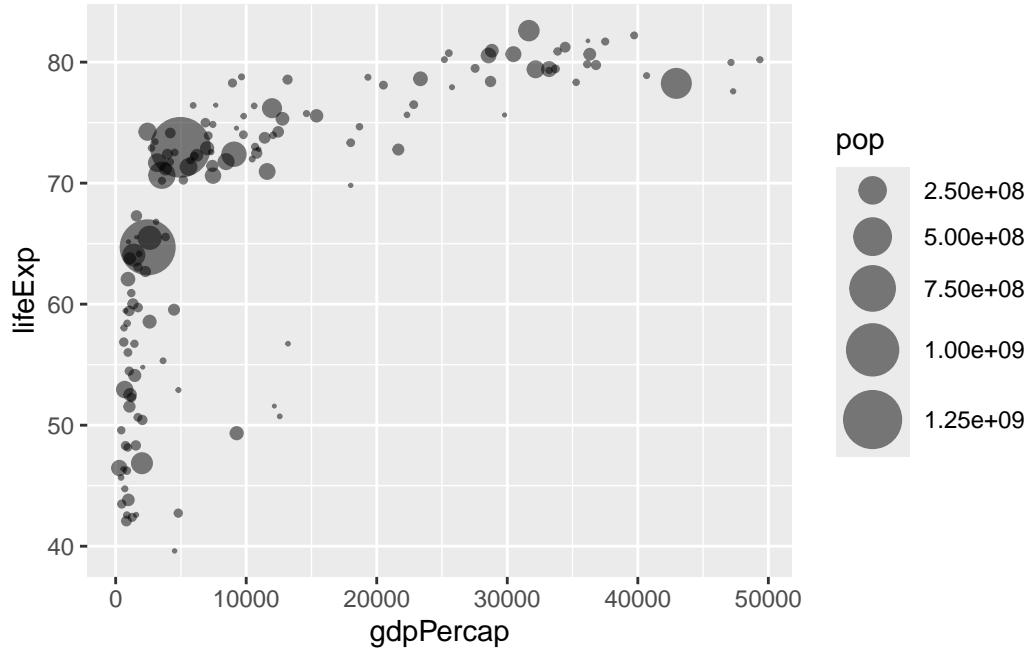


```
ggplot(gapminder_2007) +
  aes(x = gdpPercap, y = lifeExp, color = continent, size = pop) +
  geom_point(alpha=0.7)
```



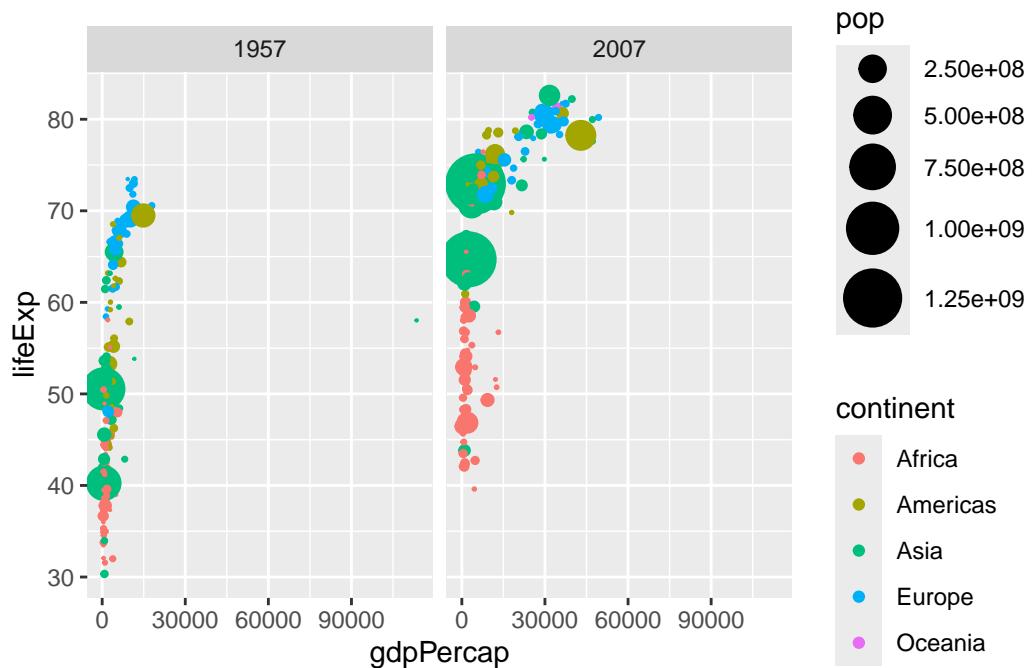
Q. How to reflect the real population on the size of the points

```
ggplot(gapminder_2007) +  
  geom_point(aes(x = gdpPercap, y = lifeExp,  
                 size = pop), alpha=0.5) +  
  scale_size_area(max_size = 10)
```



Q. Can you adapt the code you have learned thus far to reproduce our gapminder scatter plot for the year 1957? What do you notice about this plot is it easy to compare with the one for 2007?

```
gapminder_1957 <- gapminder %>% filter(year == 2007 | year==1957)
ggplot(gapminder_1957) +
  aes(x = gdpPercap, y = lifeExp, color = continent, size = pop) +
  geom_point() +
  scale_size_area(max_size = 10) +
  facet_wrap(~year)
```



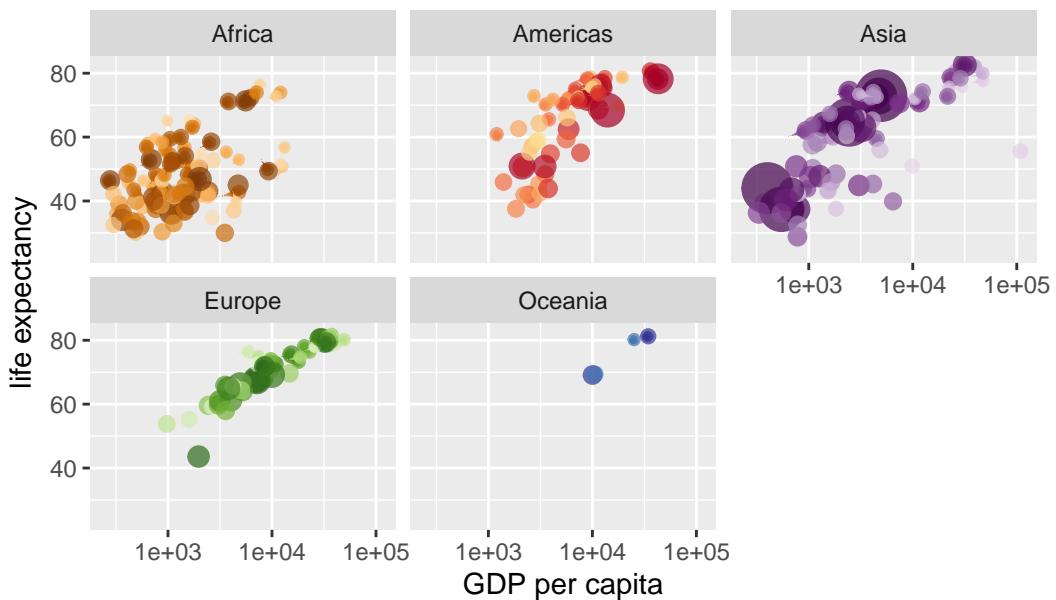
Animation

```
library(gapminder)
library(gganimate)

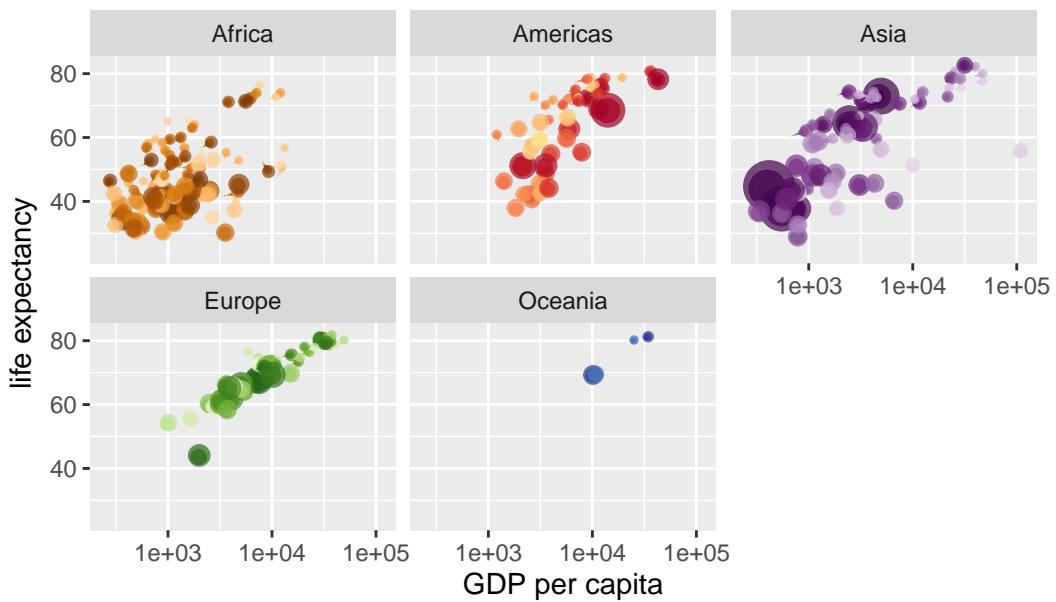
# Setup nice regular ggplot of the gapminder data
ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, colour = country)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors) +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  # Facet by continent
  facet_wrap(~continent) +
  # Here comes the gganimate specific bits
  labs(title = 'Year: {frame_time}', x = 'GDP per capita', y = 'life expectancy') +
  transition_time(year) +
  shadow_wake(wake_length = 0.1, alpha = FALSE)
```

Warning in formals(fun): argument is not a function

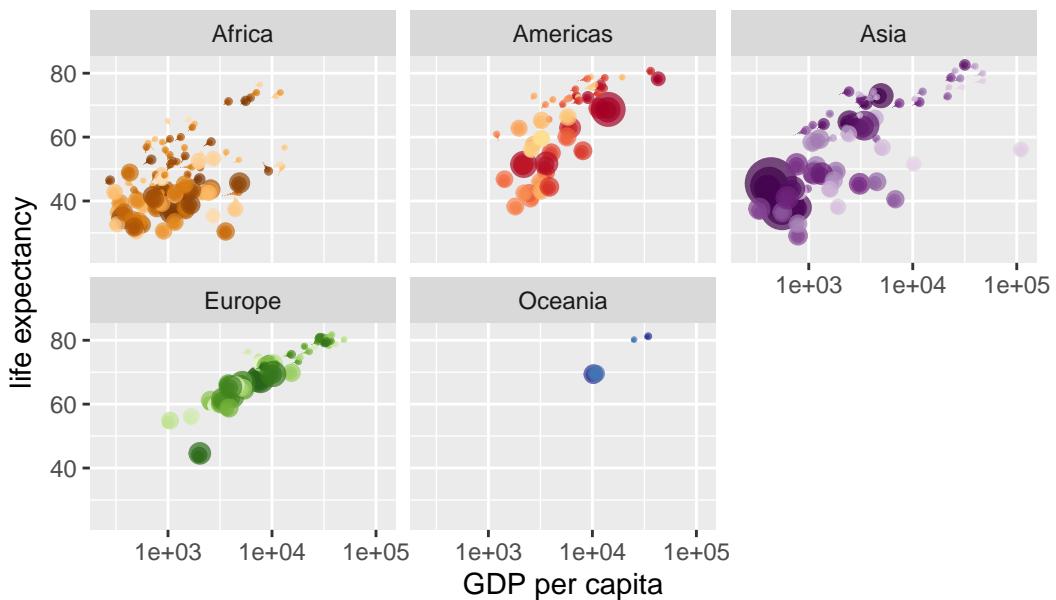
Year: 1952



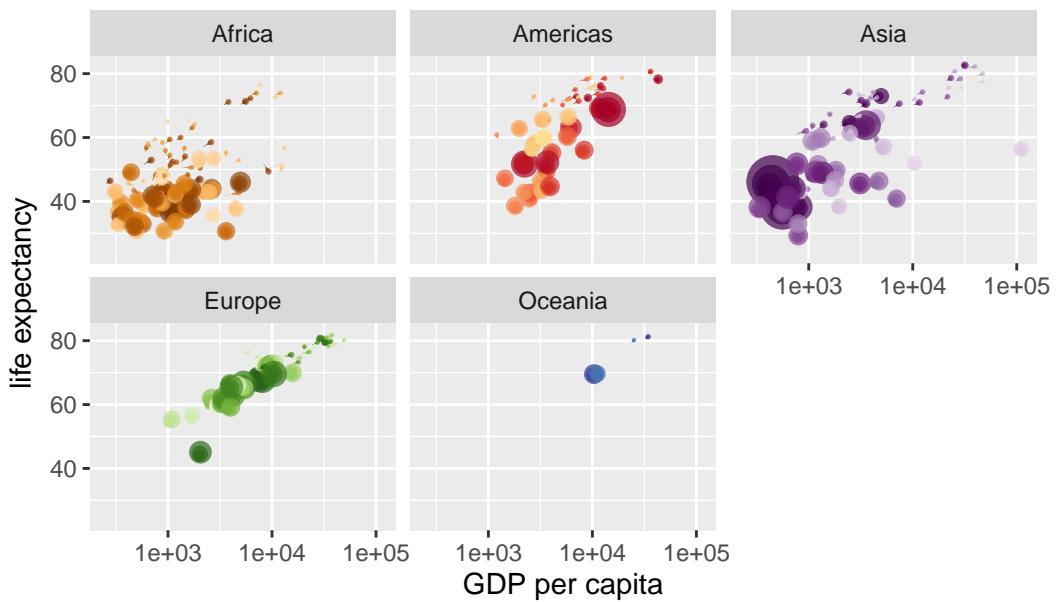
Year: 1953



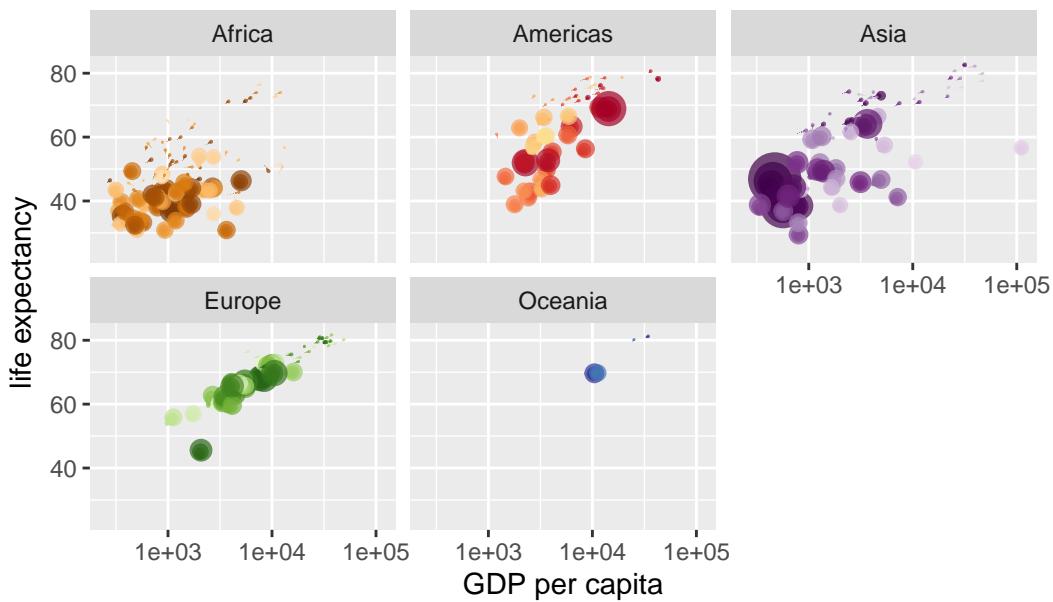
Year: 1953



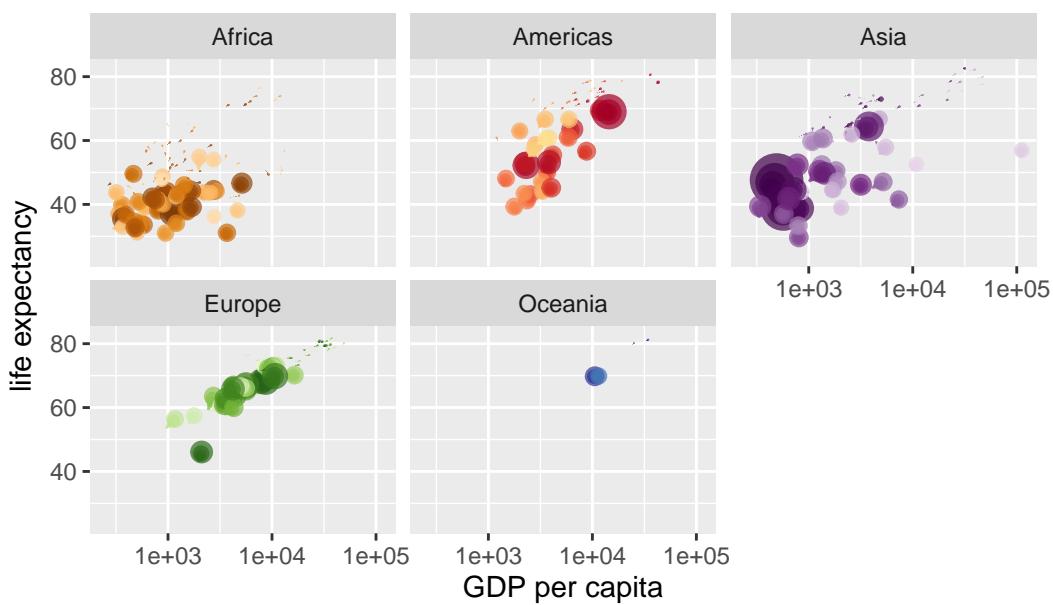
Year: 1954



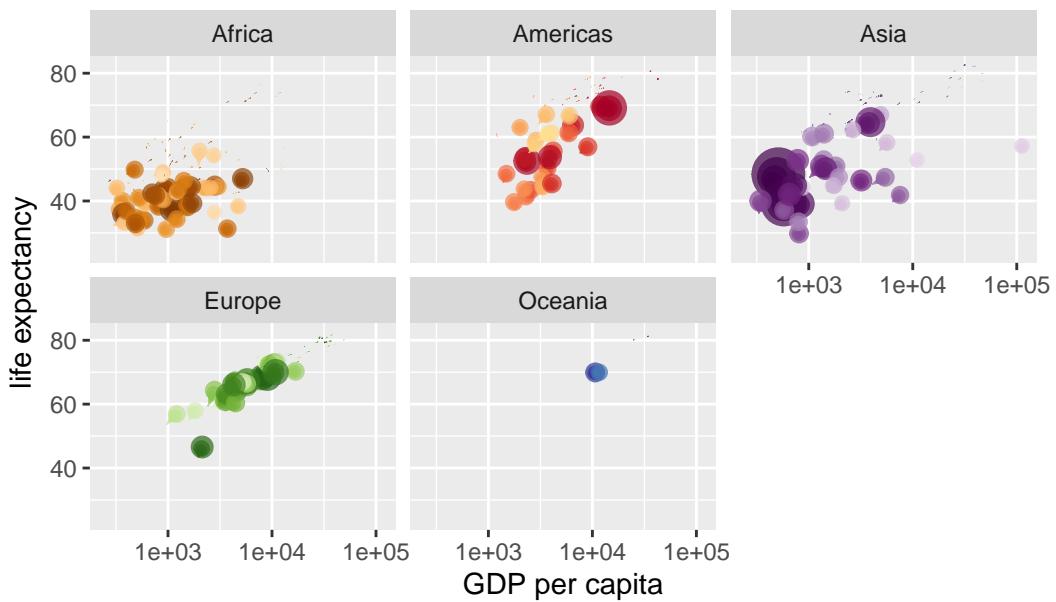
Year: 1954



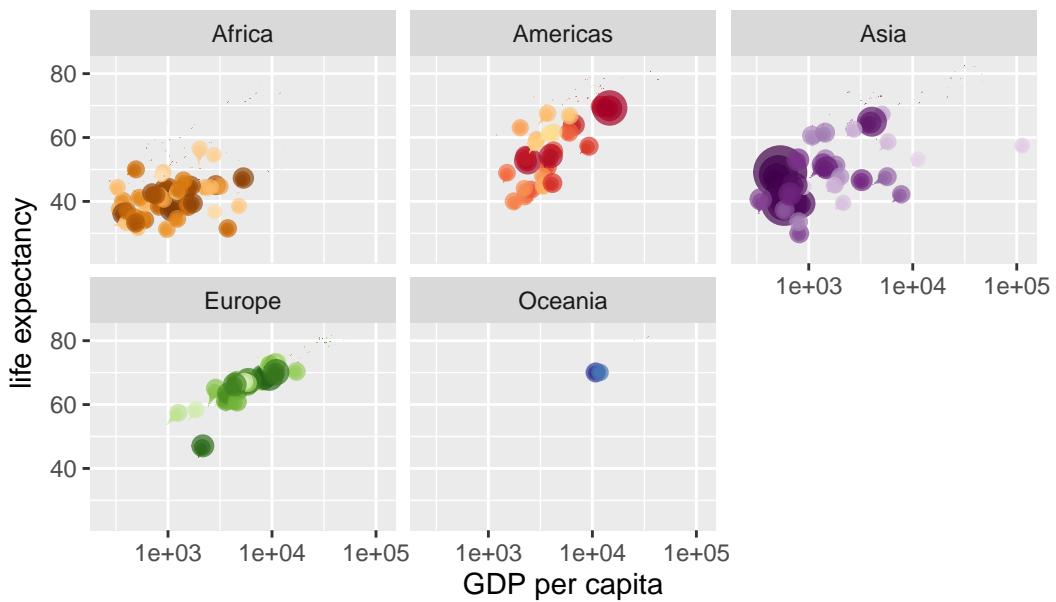
Year: 1955



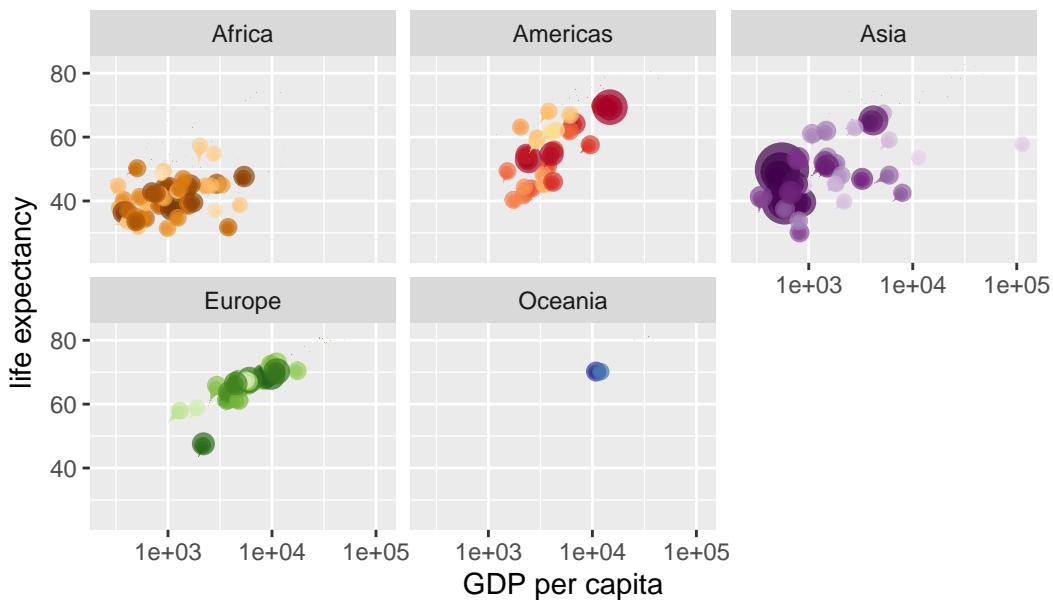
Year: 1955



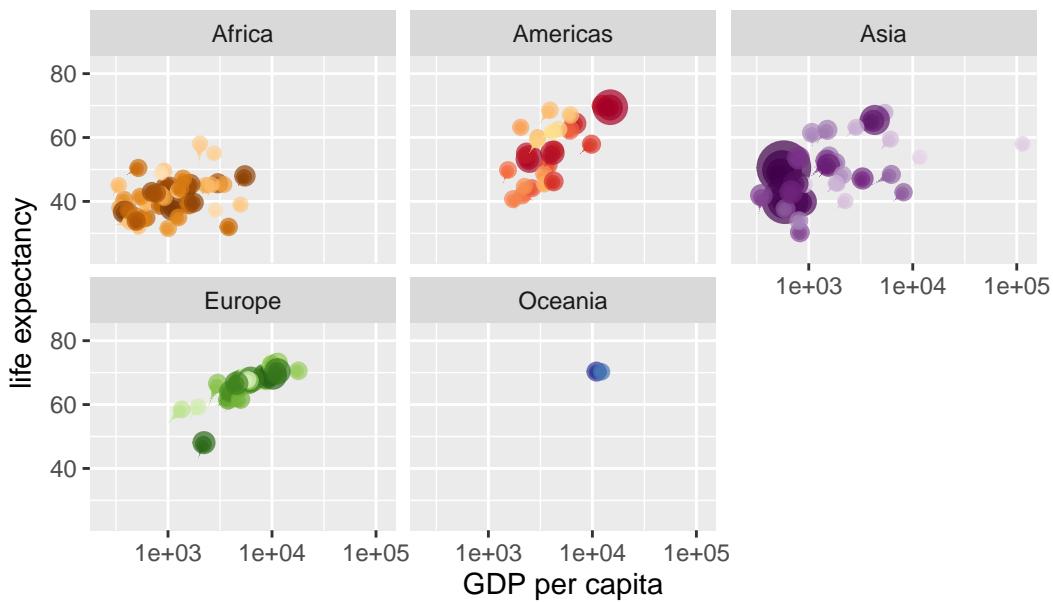
Year: 1956



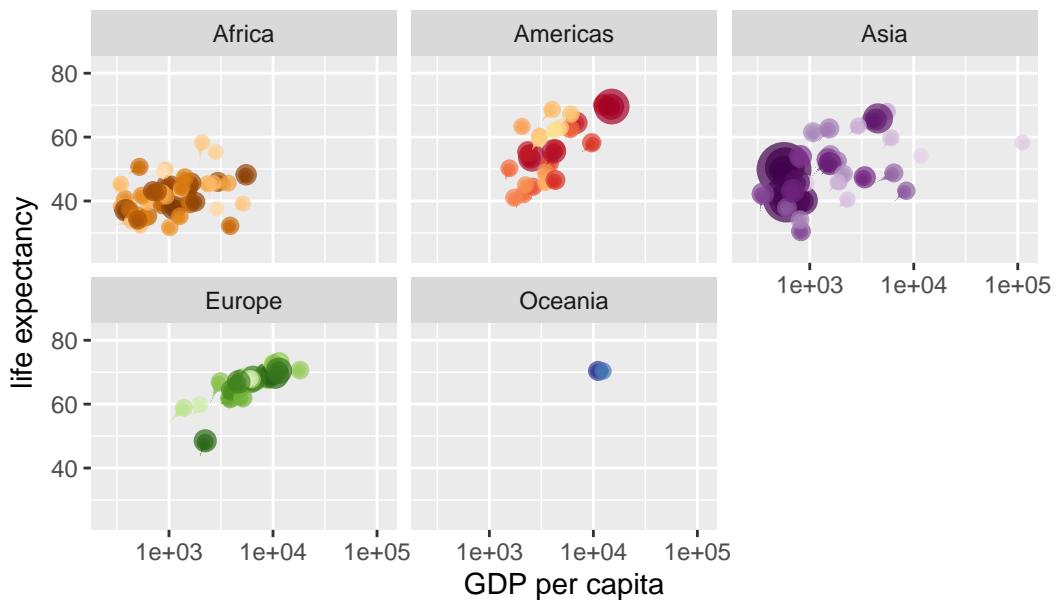
Year: 1956



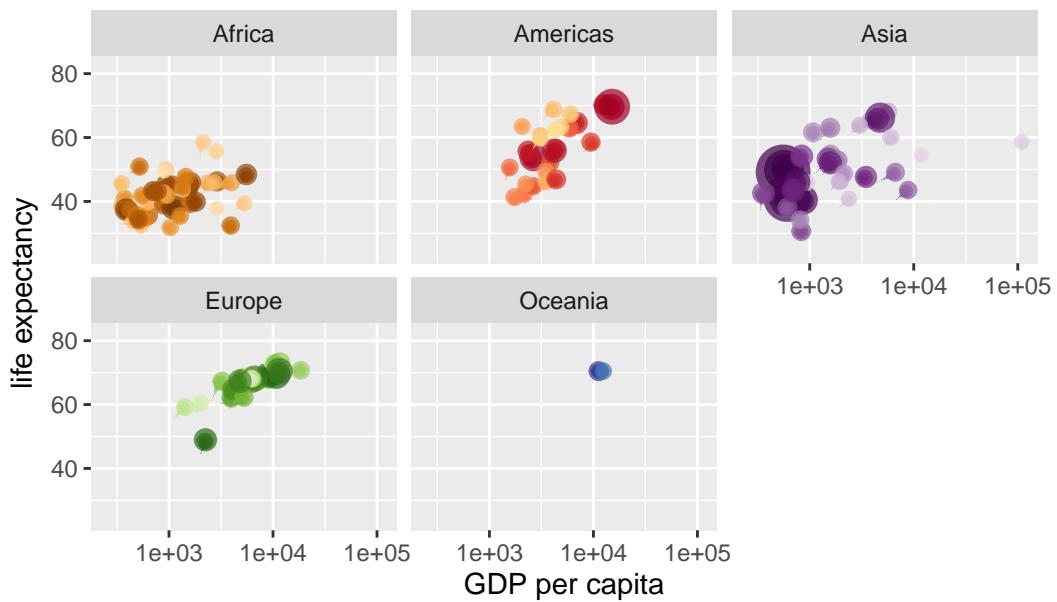
Year: 1957



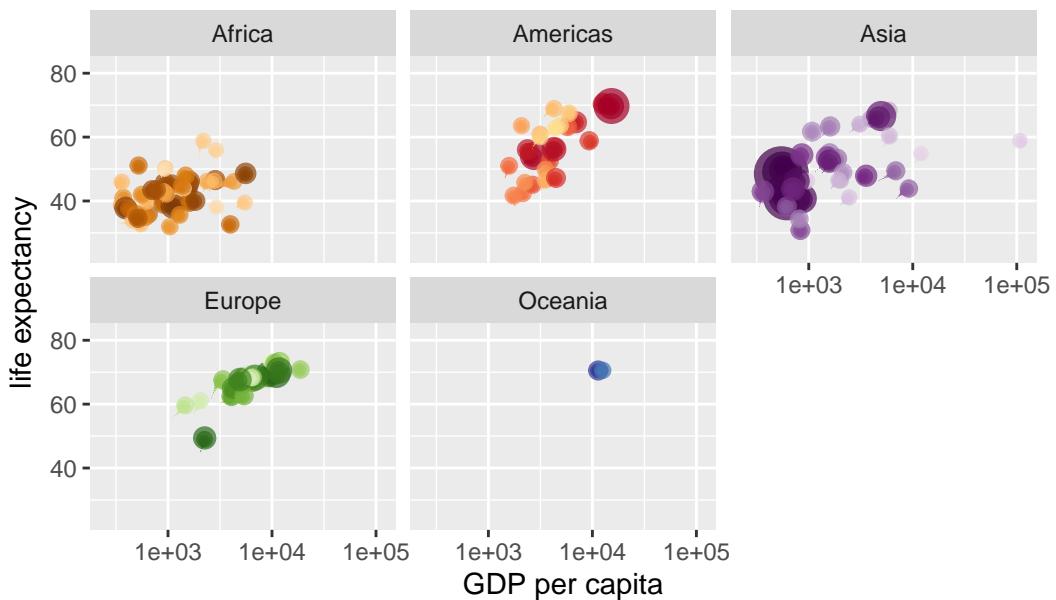
Year: 1958



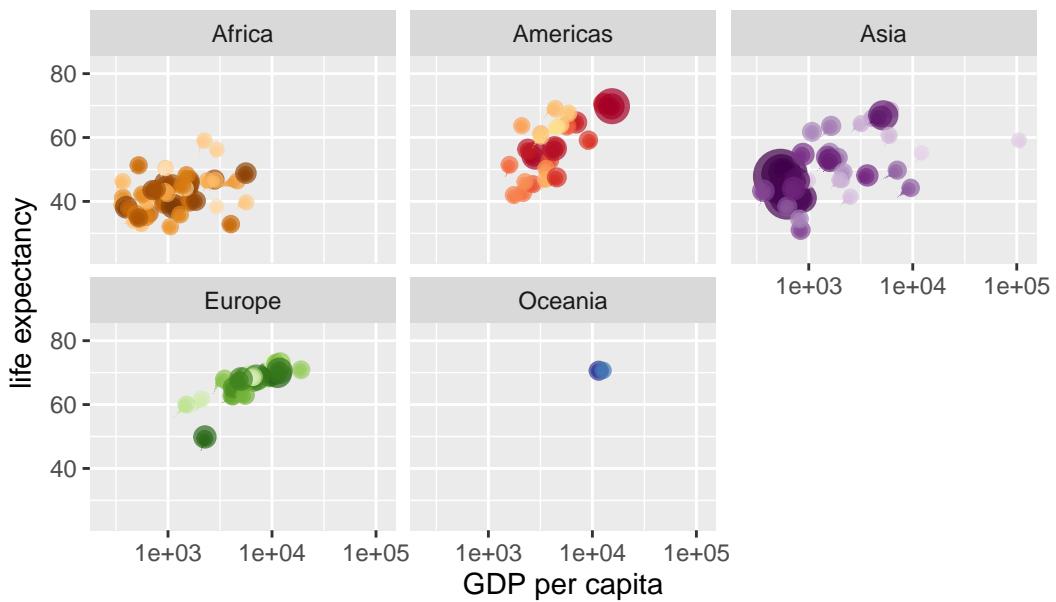
Year: 1958



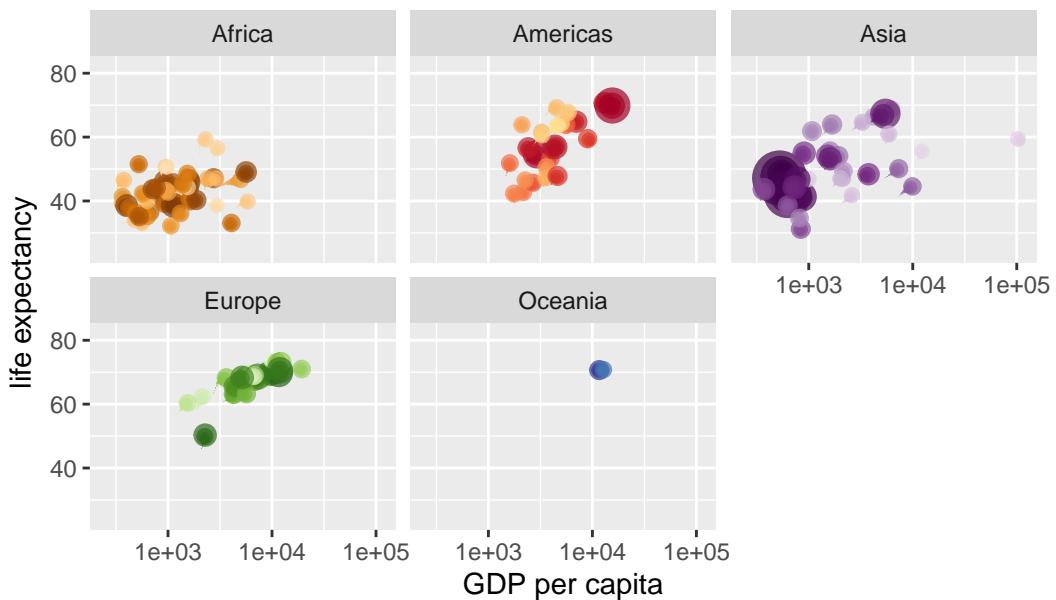
Year: 1959



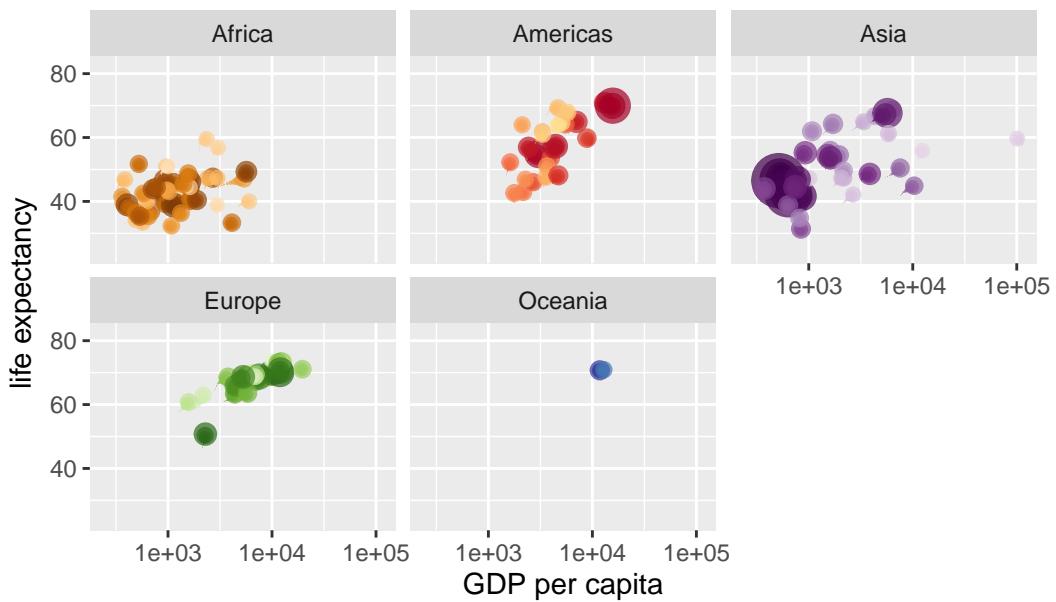
Year: 1959



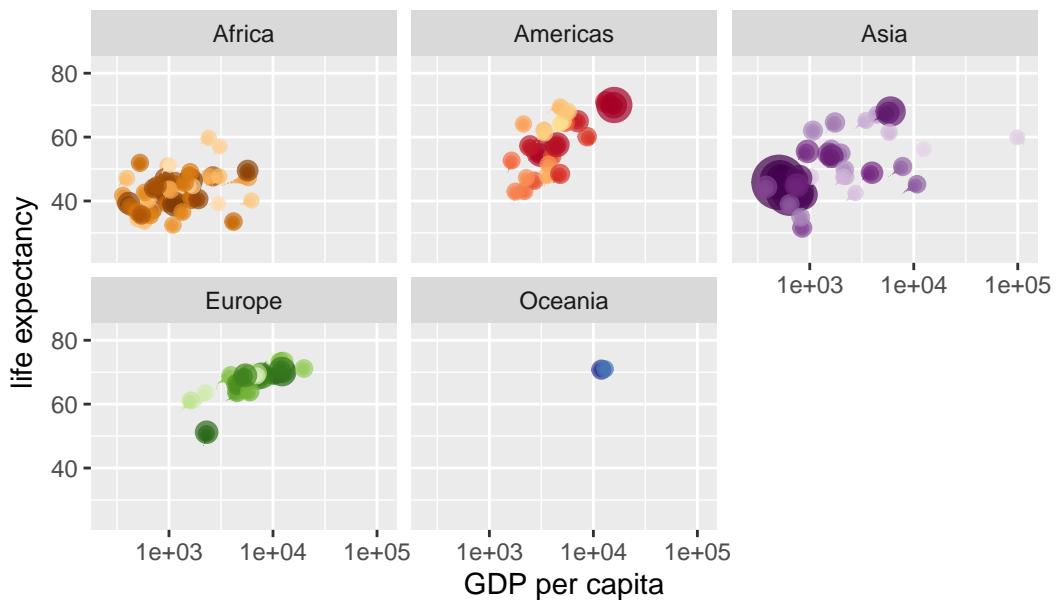
Year: 1960



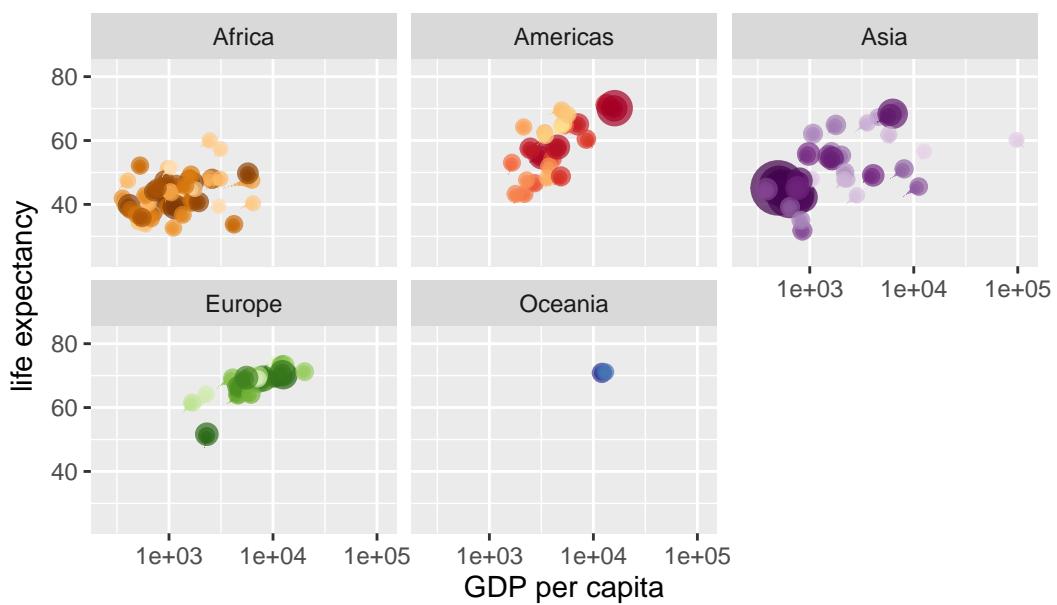
Year: 1960



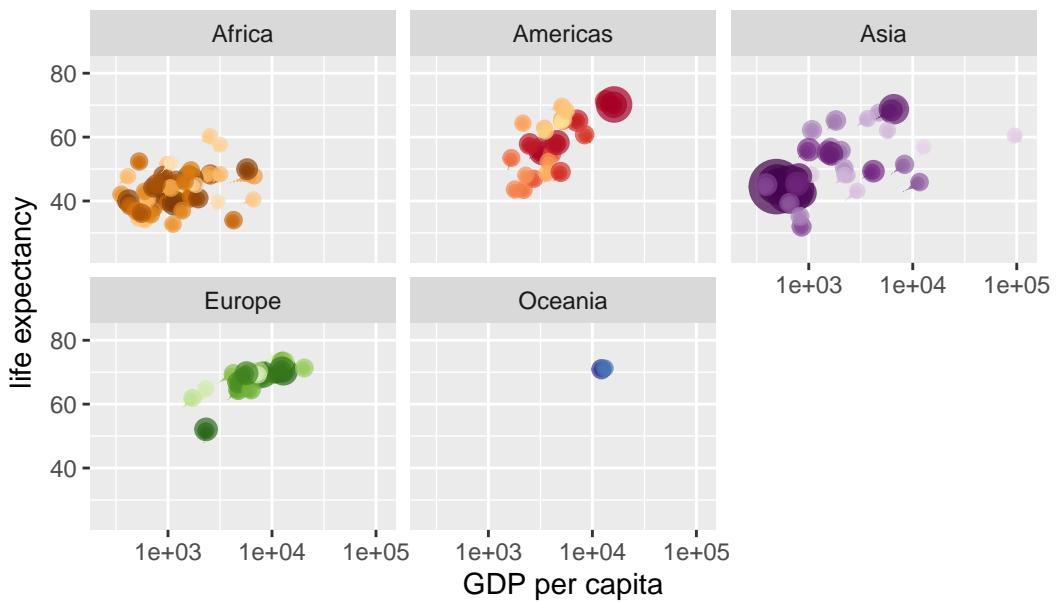
Year: 1961



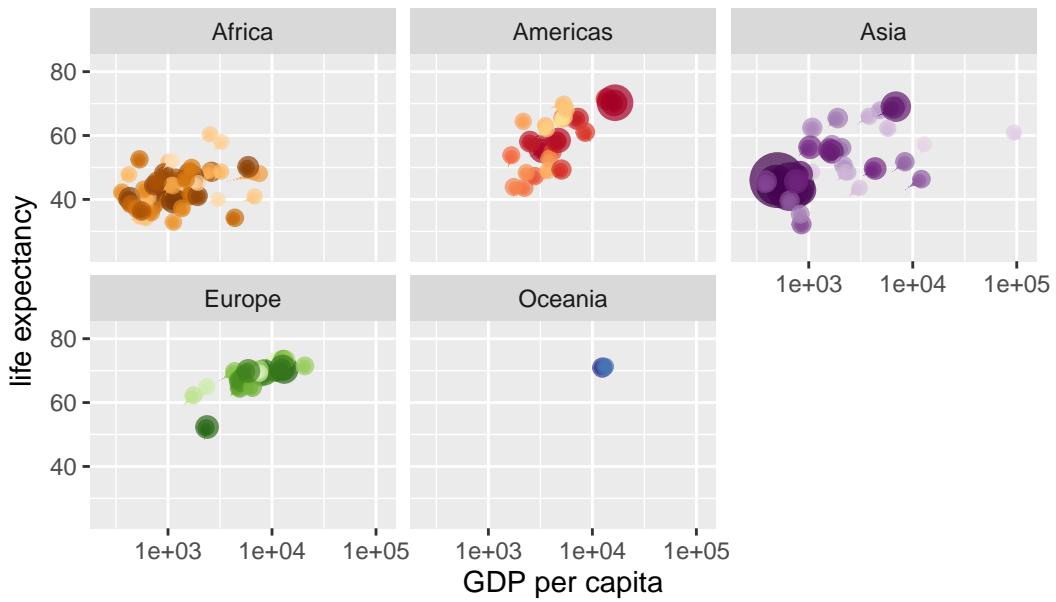
Year: 1961



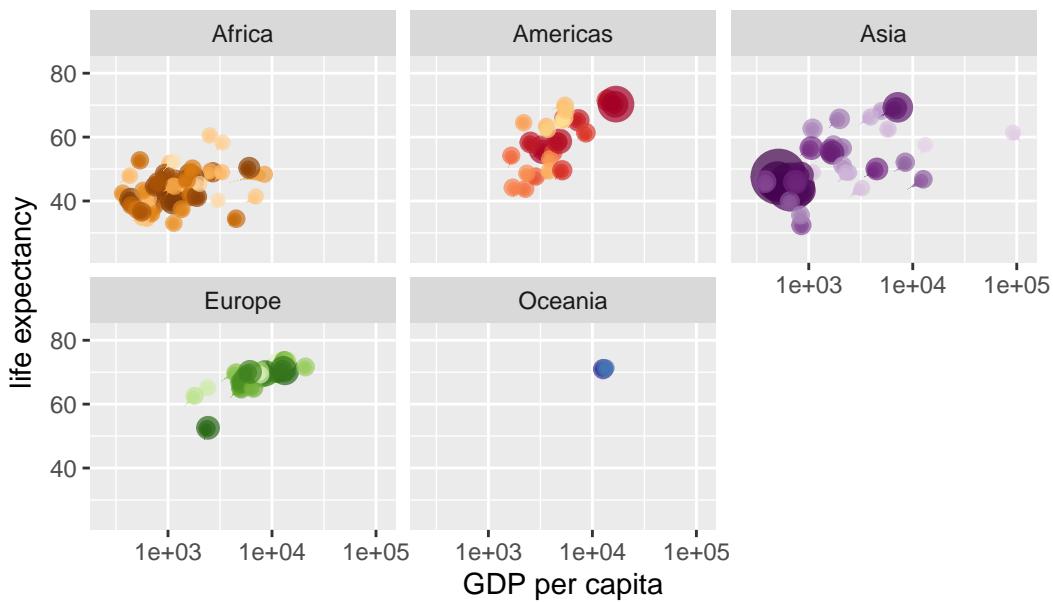
Year: 1962



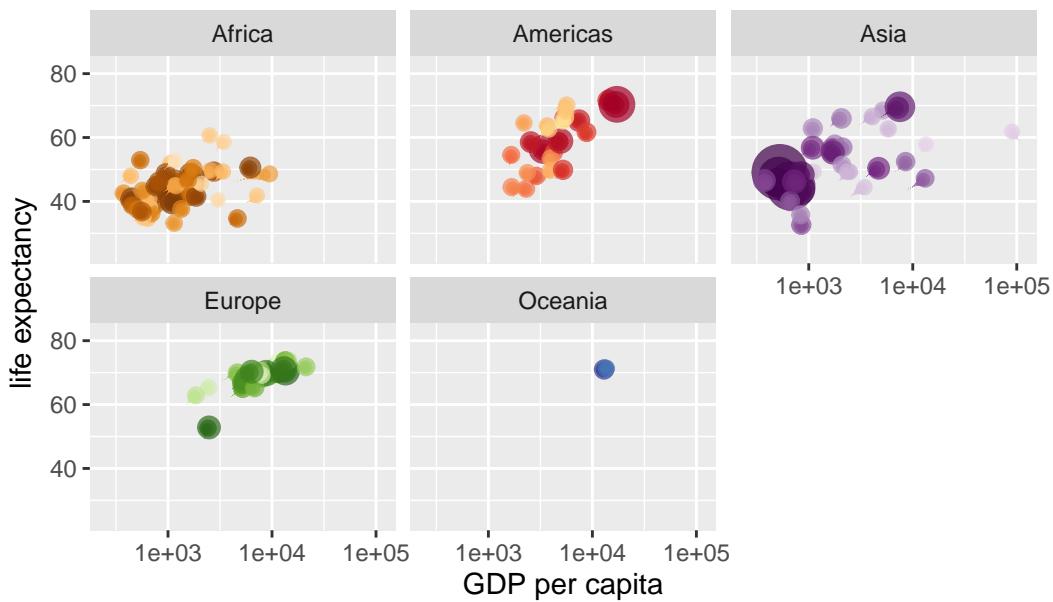
Year: 1963



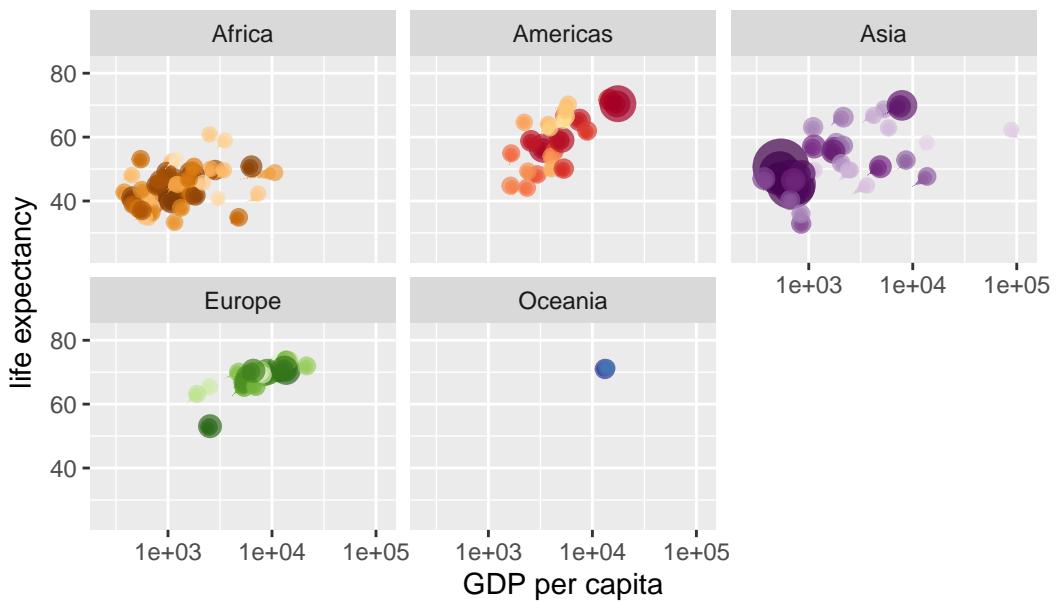
Year: 1963



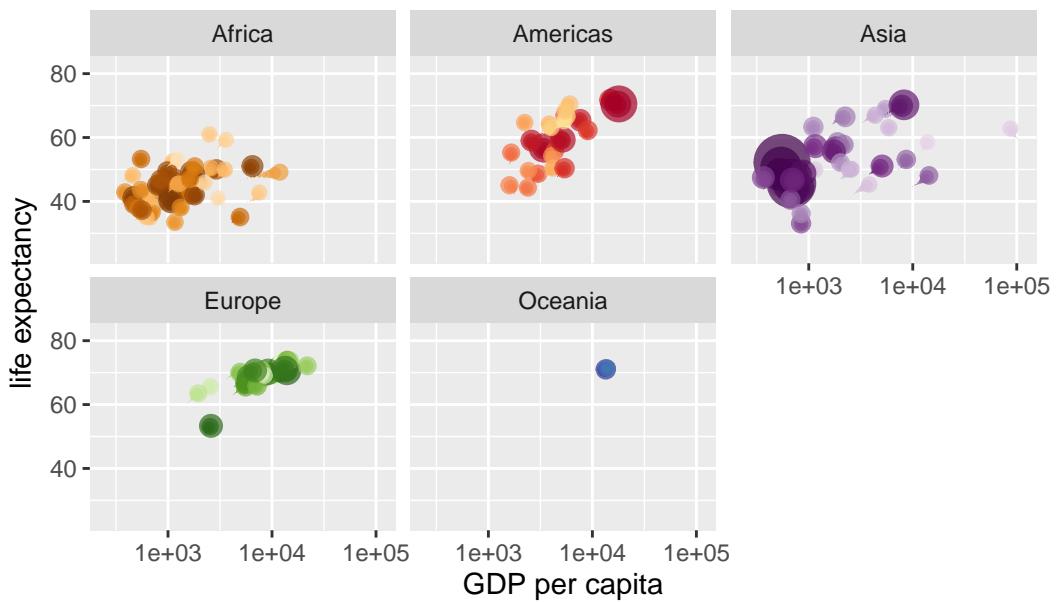
Year: 1964



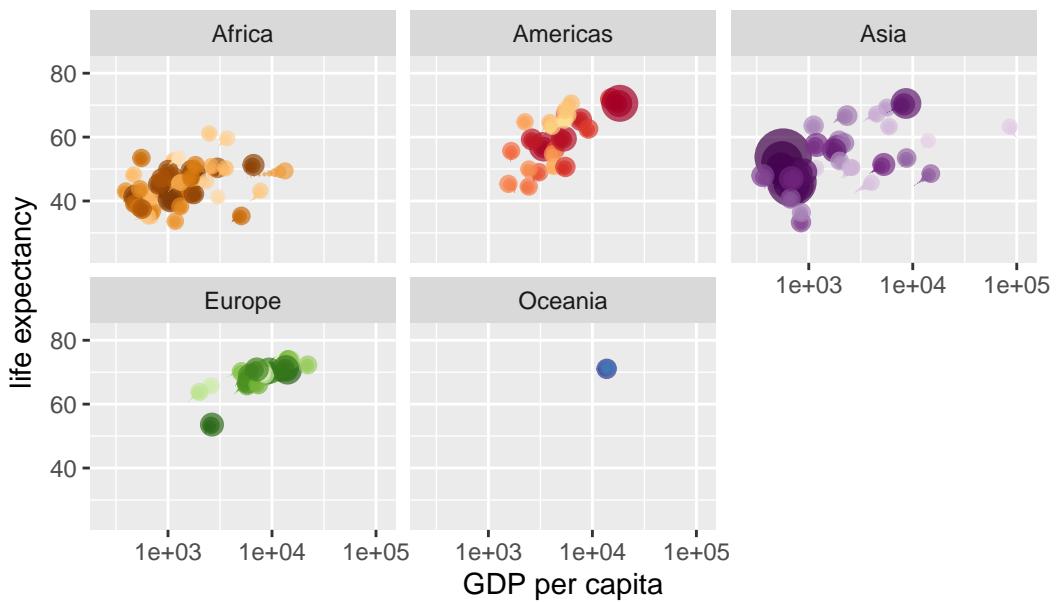
Year: 1964



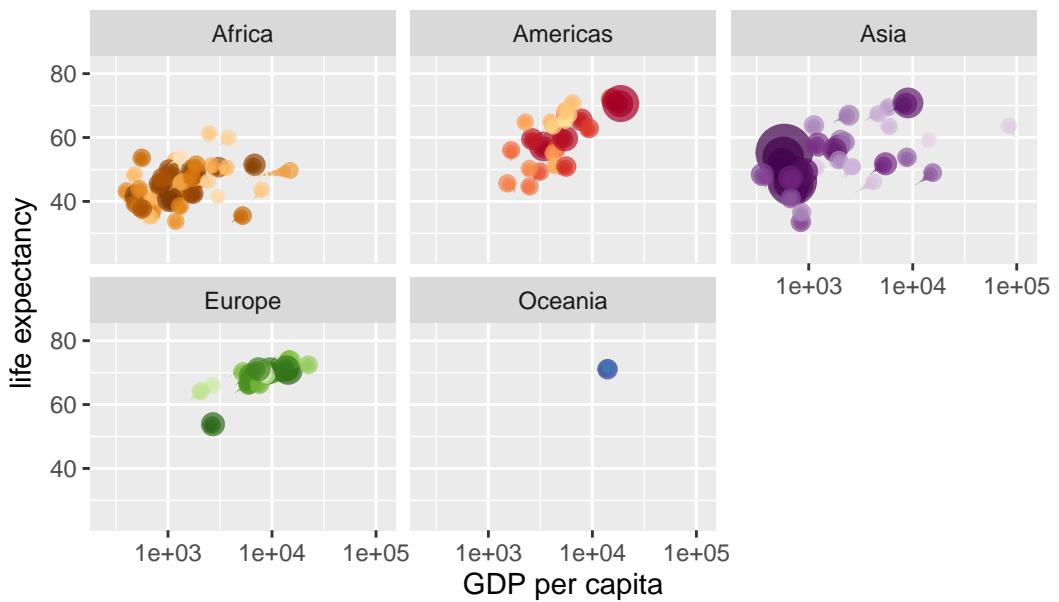
Year: 1965



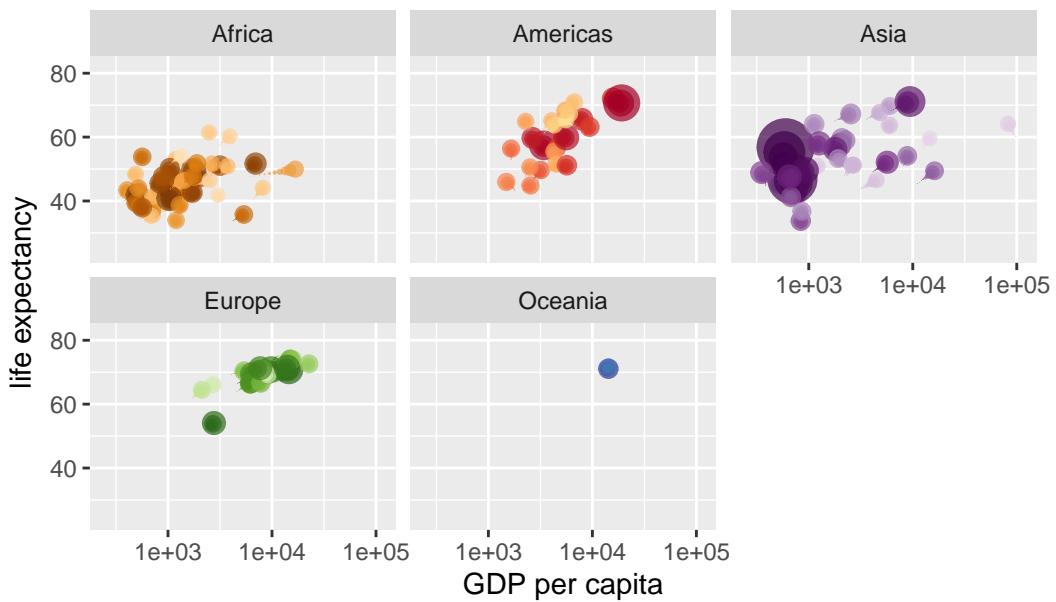
Year: 1965



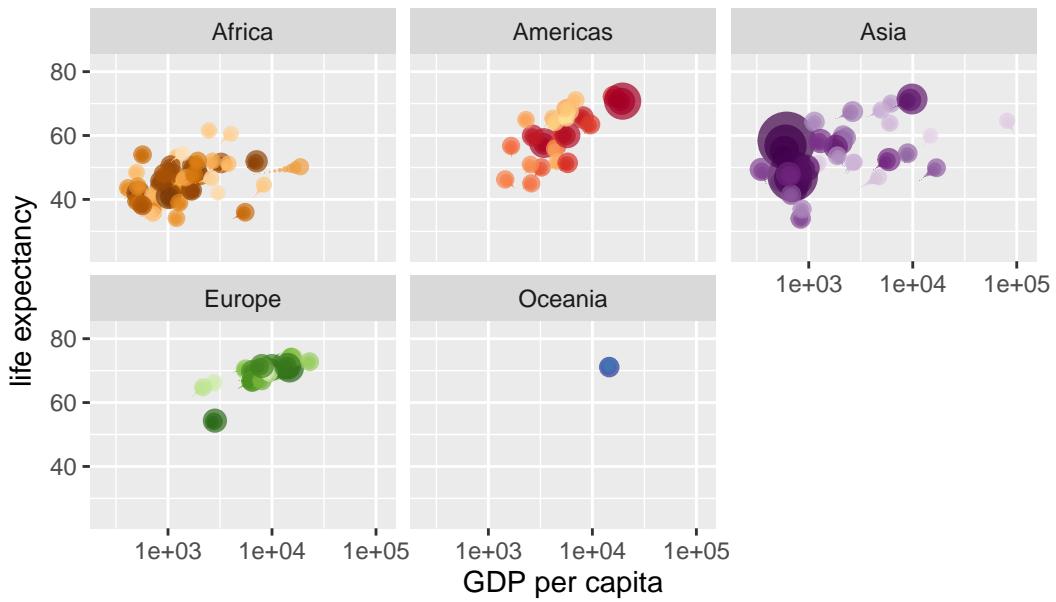
Year: 1966



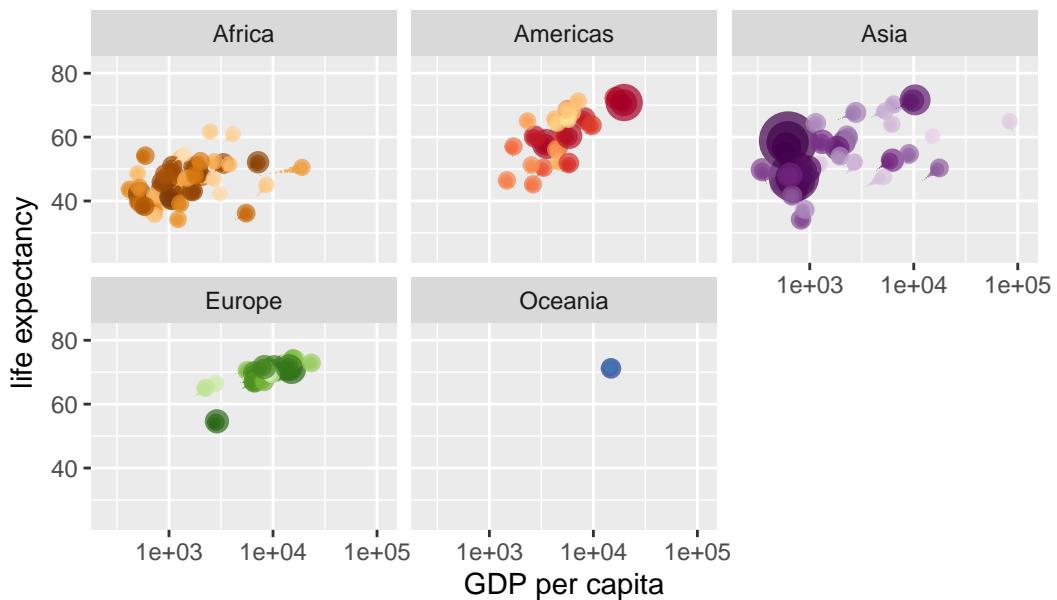
Year: 1966



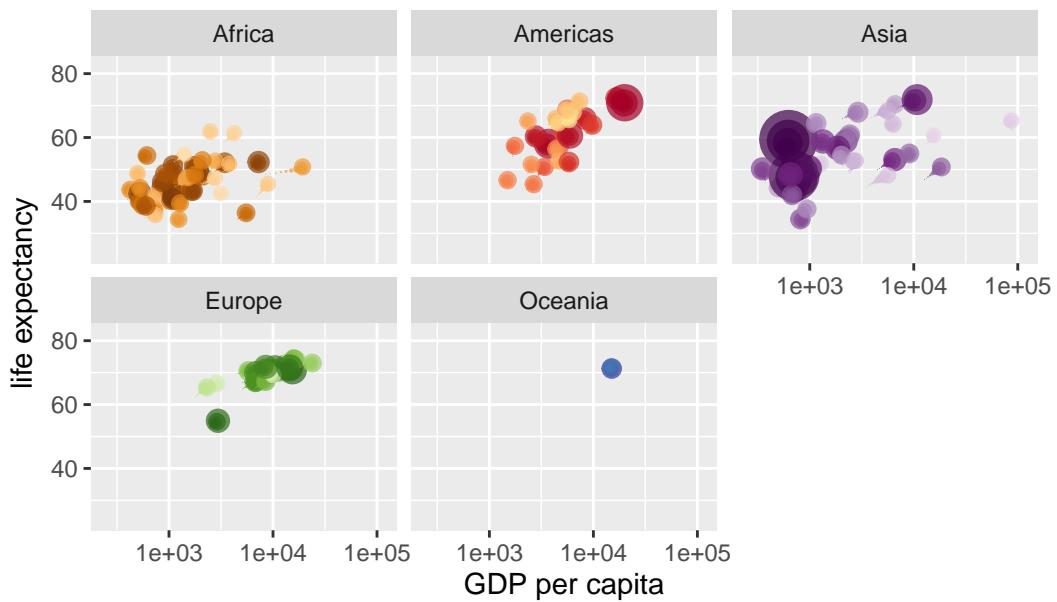
Year: 1967



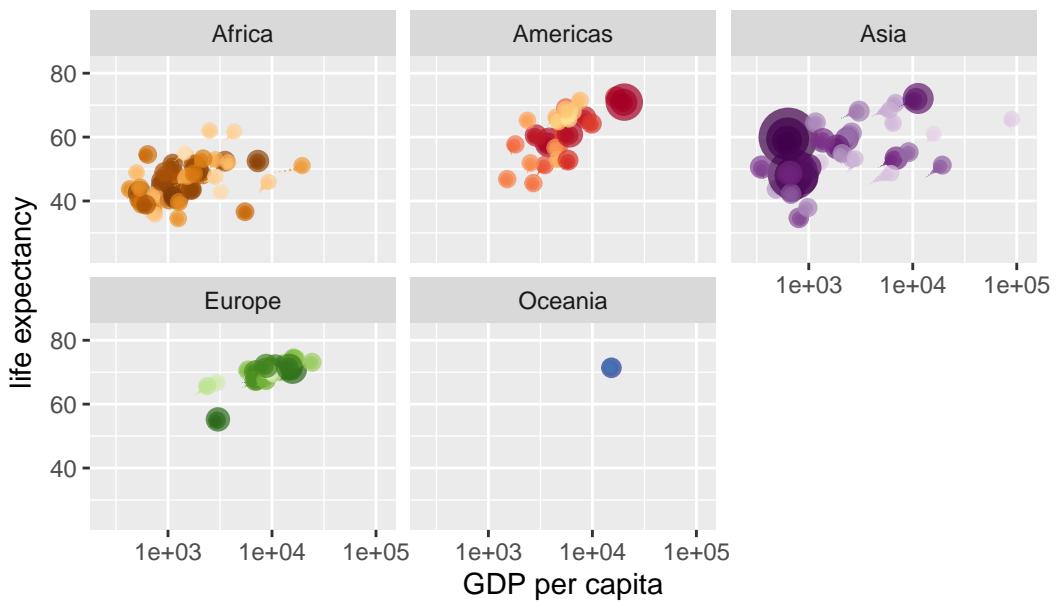
Year: 1968



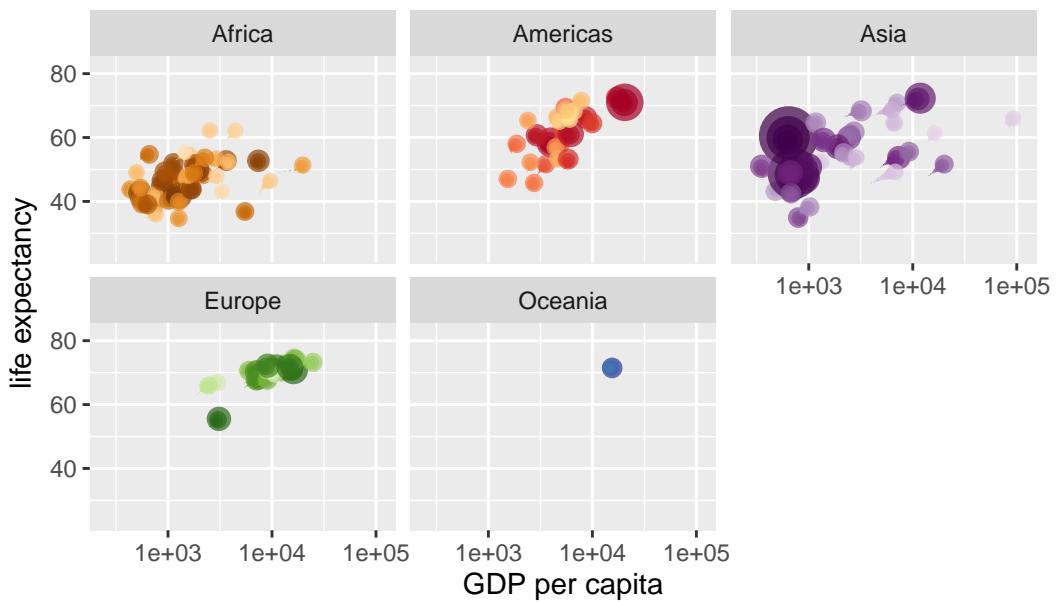
Year: 1968



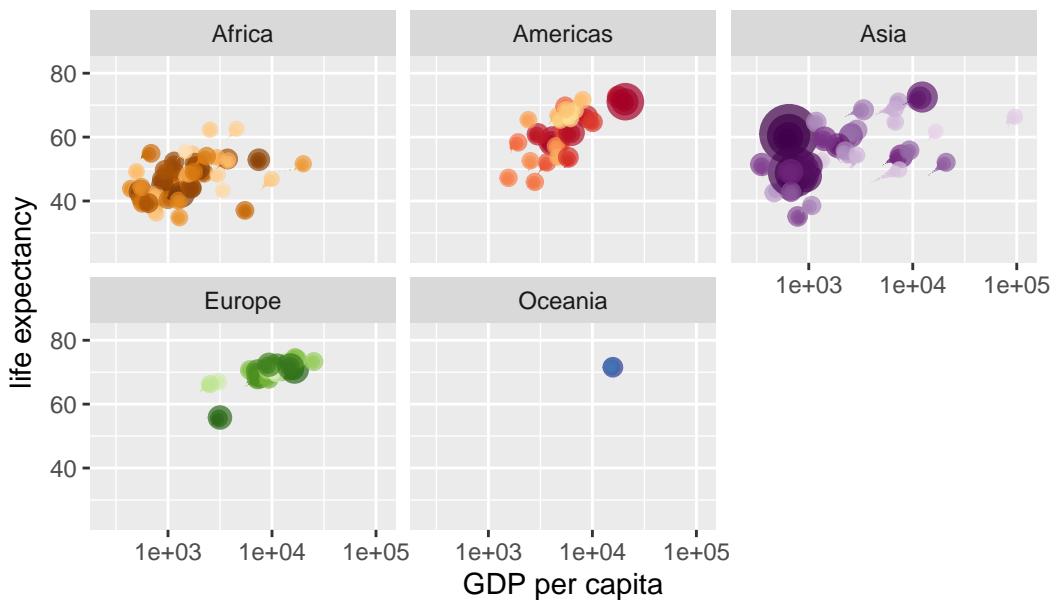
Year: 1969



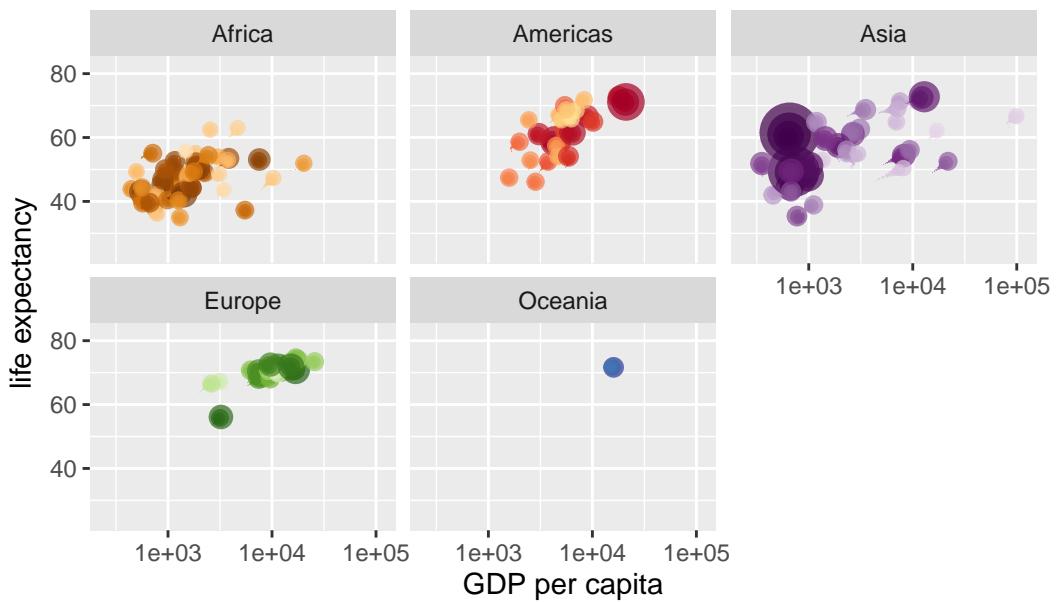
Year: 1969



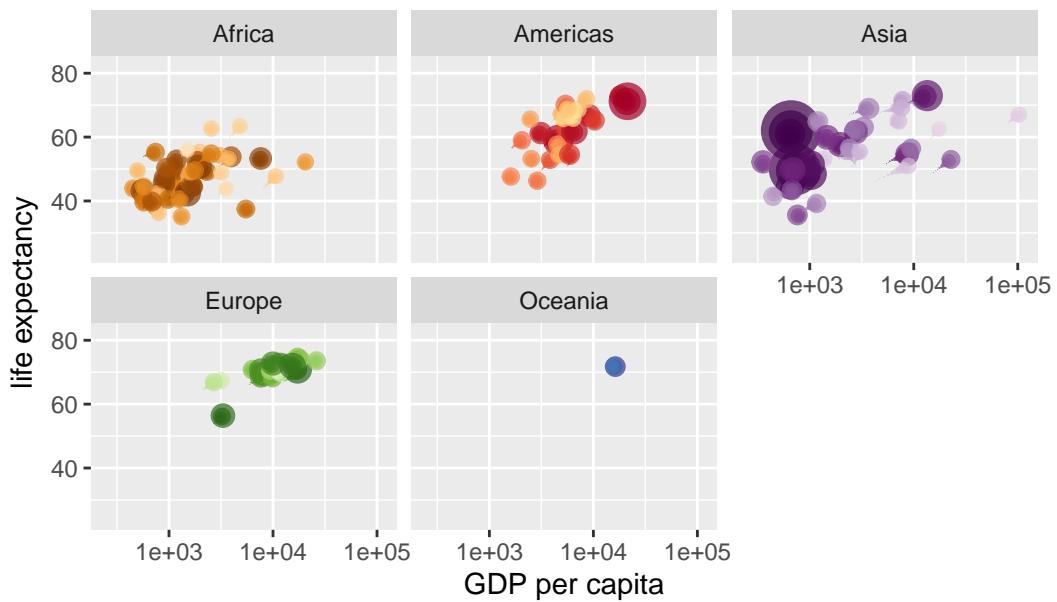
Year: 1970



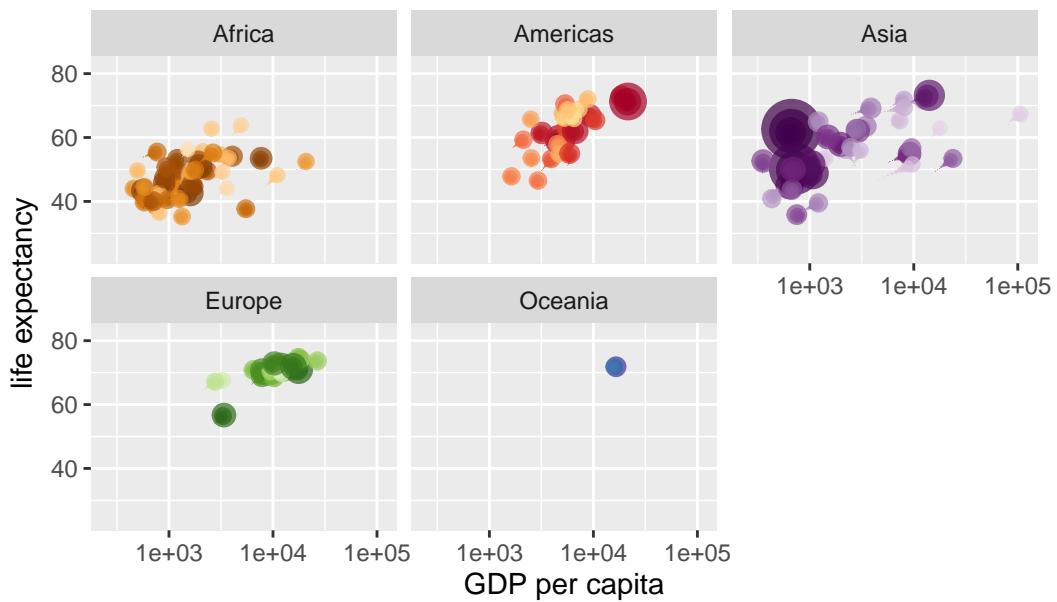
Year: 1970



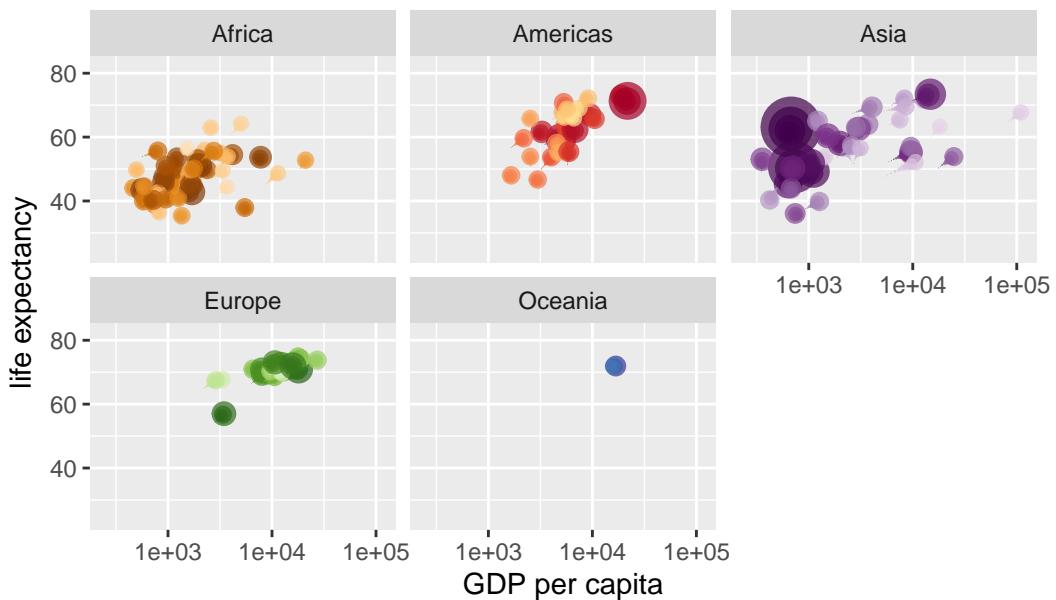
Year: 1971



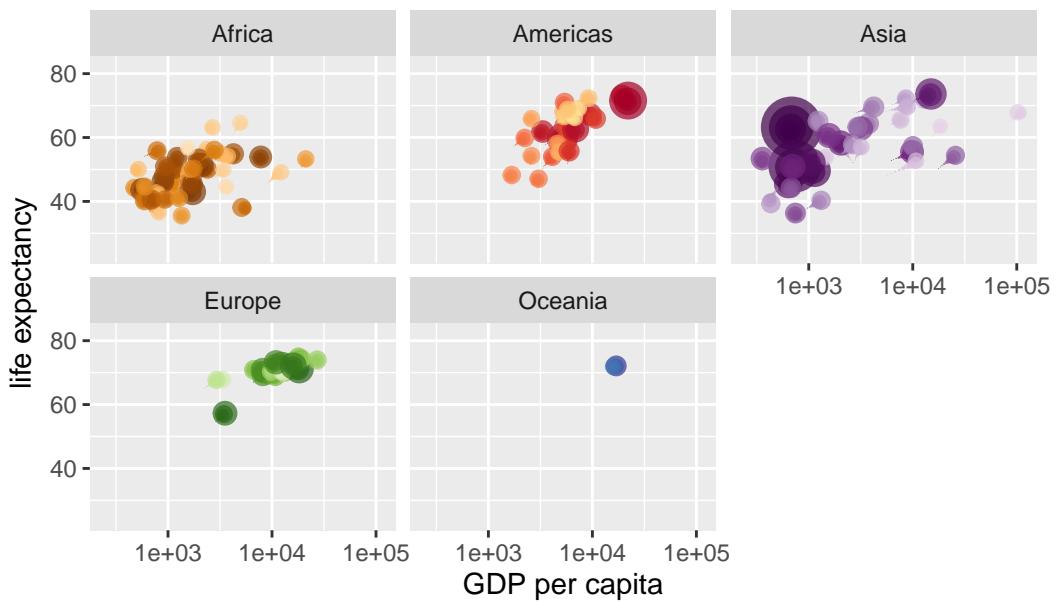
Year: 1971



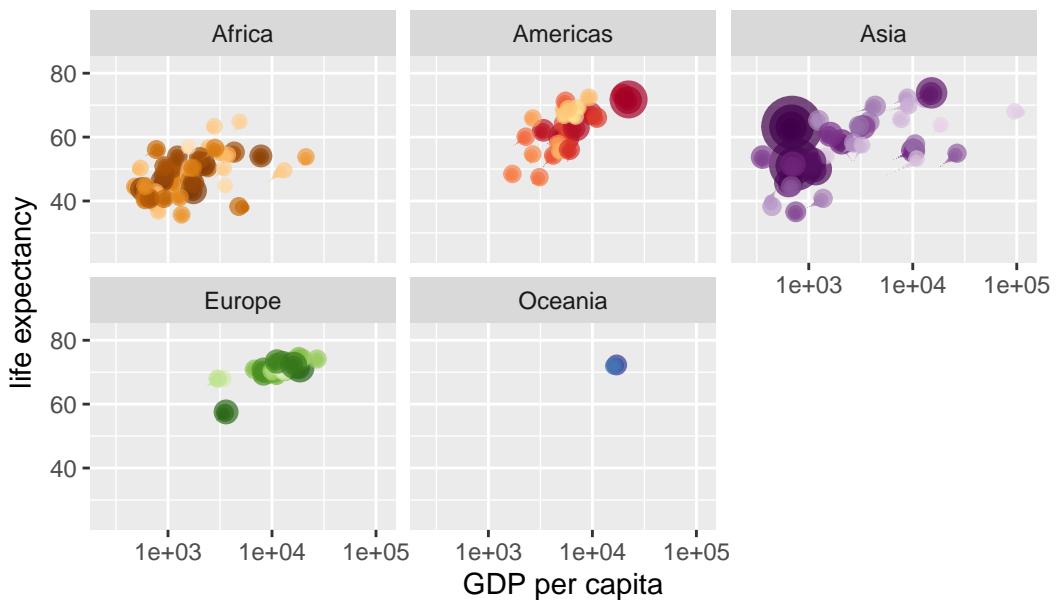
Year: 1972



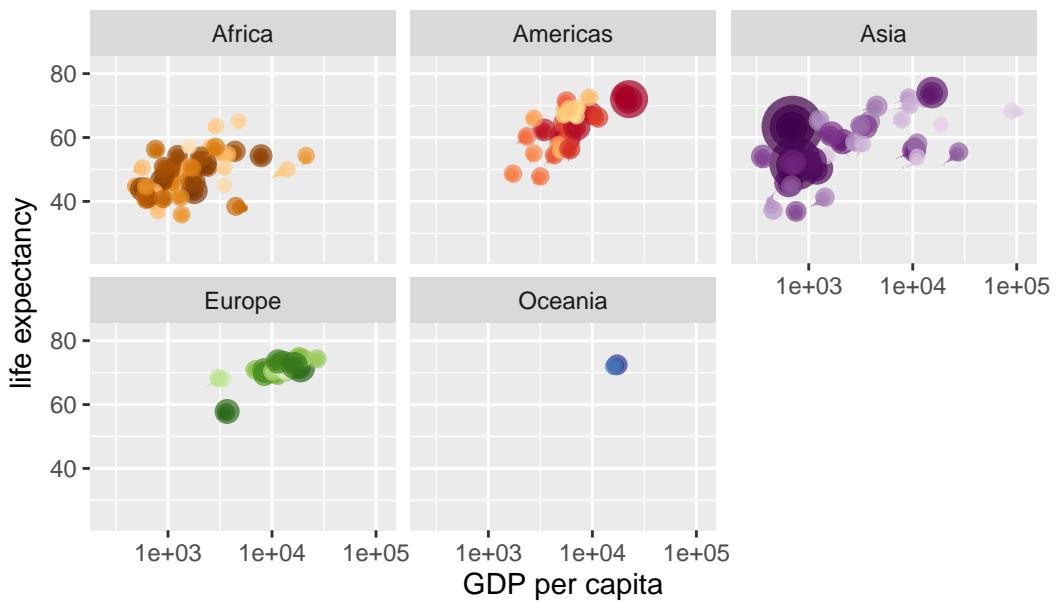
Year: 1973



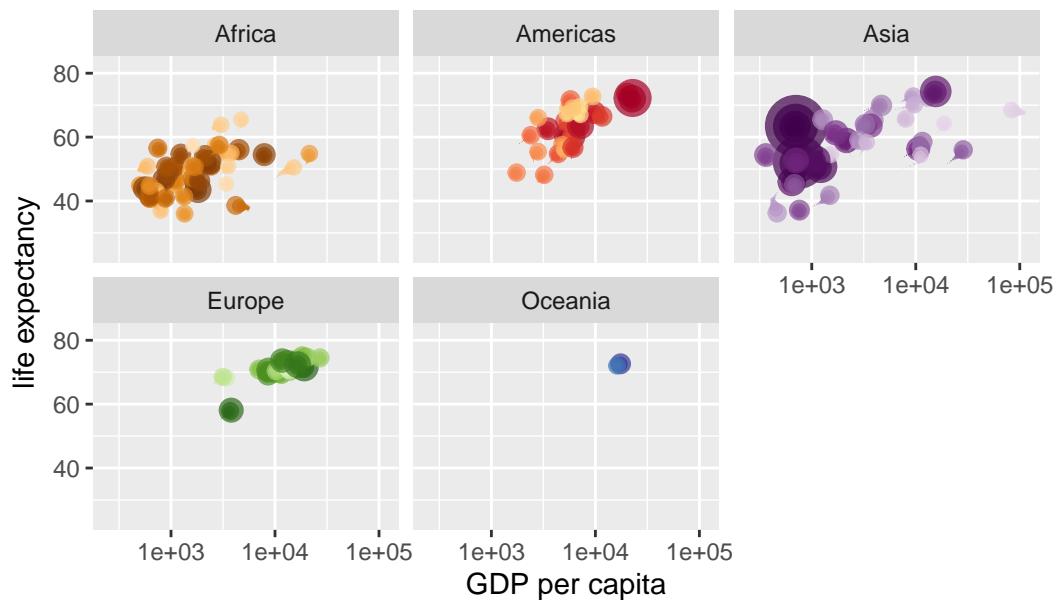
Year: 1973



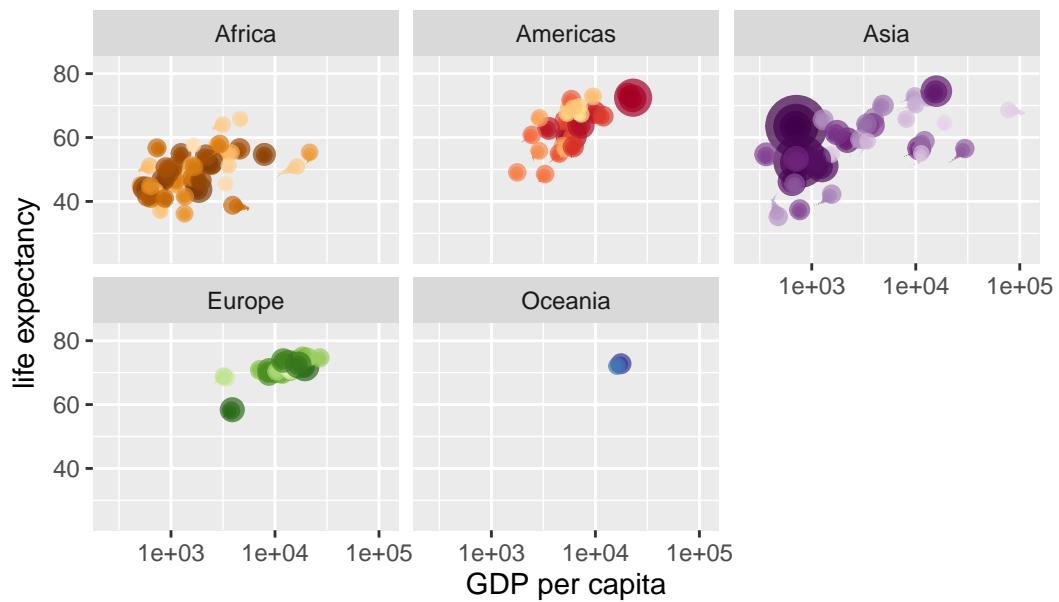
Year: 1974



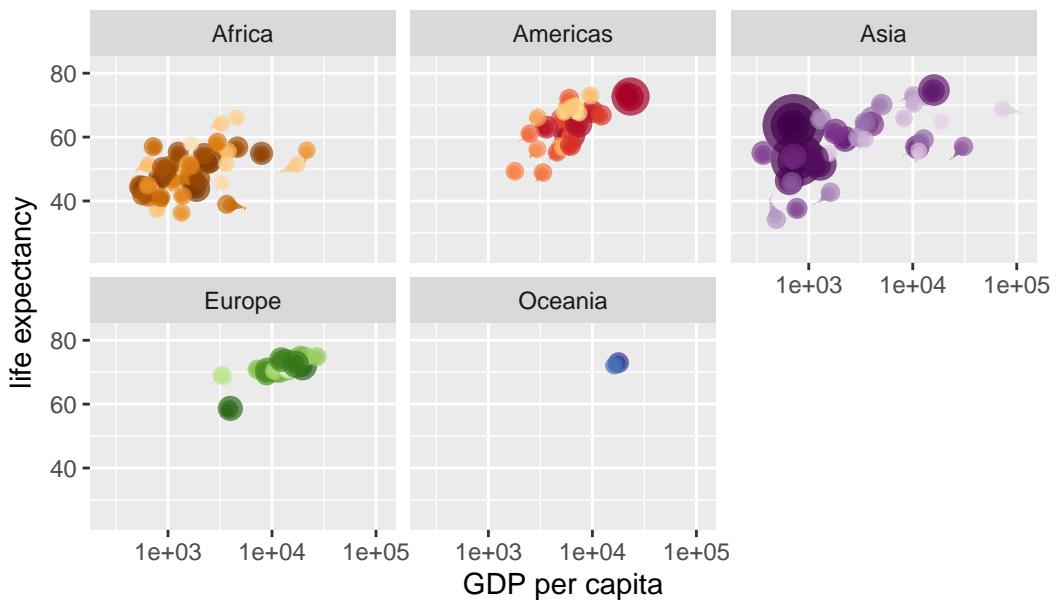
Year: 1974



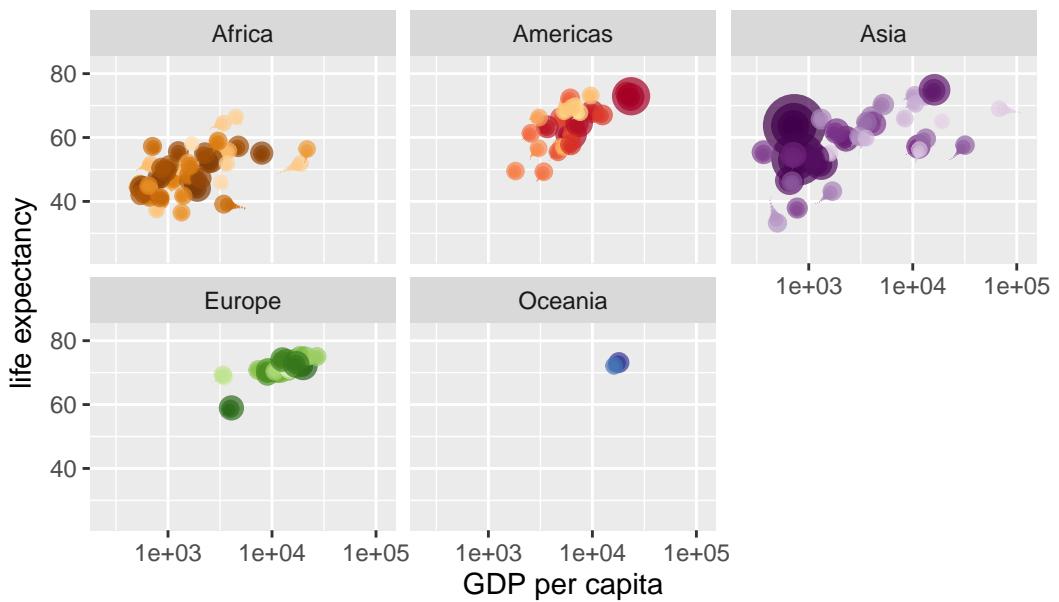
Year: 1975



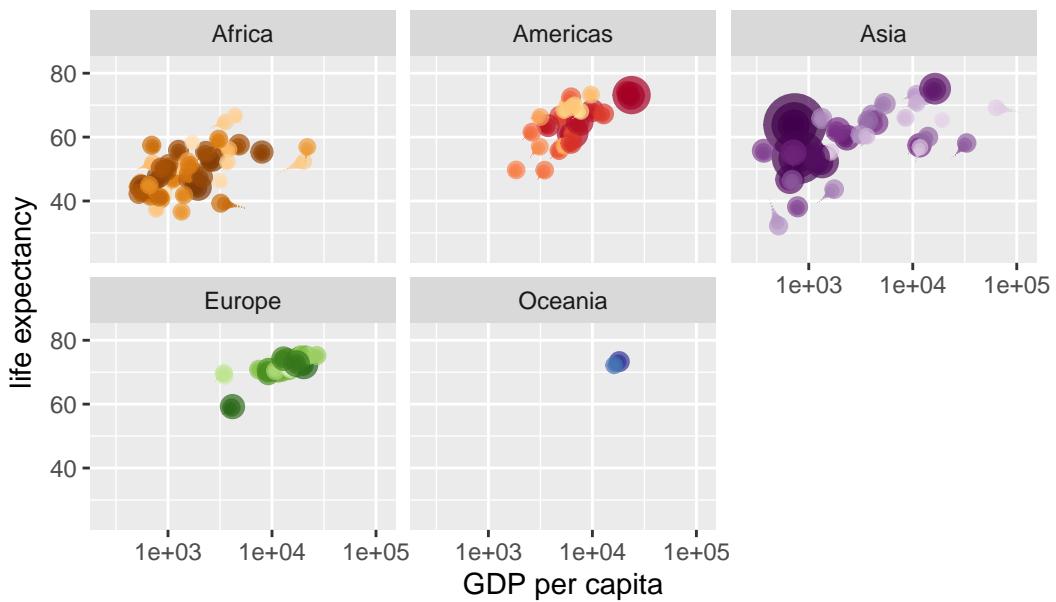
Year: 1975



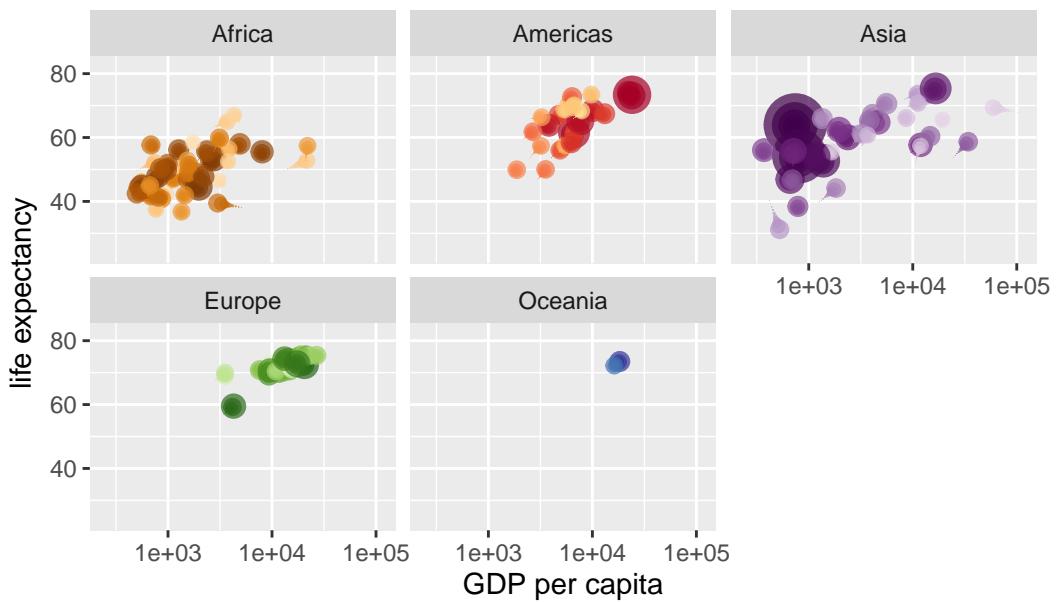
Year: 1976



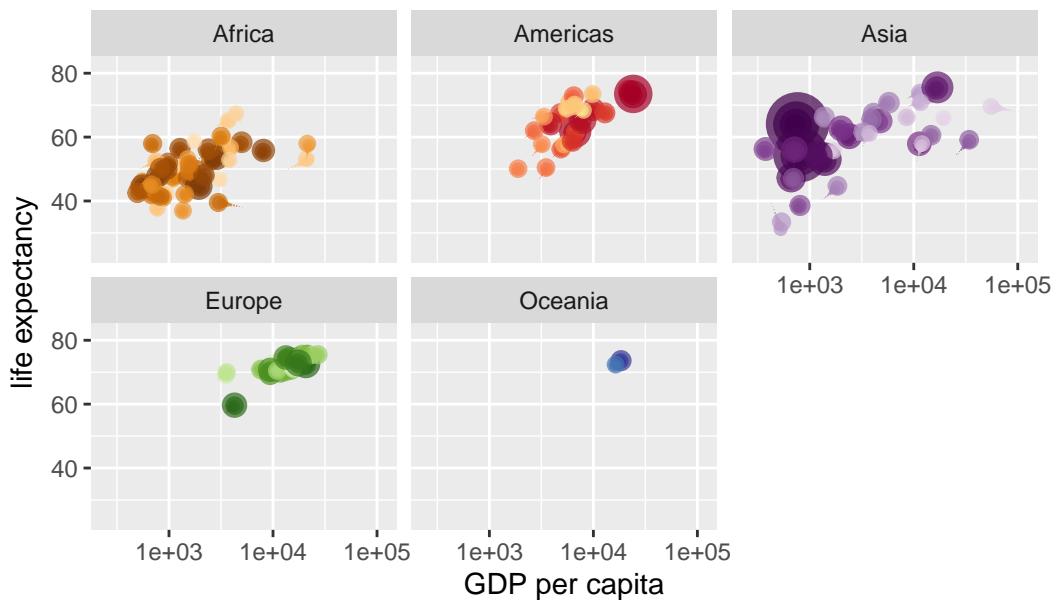
Year: 1976



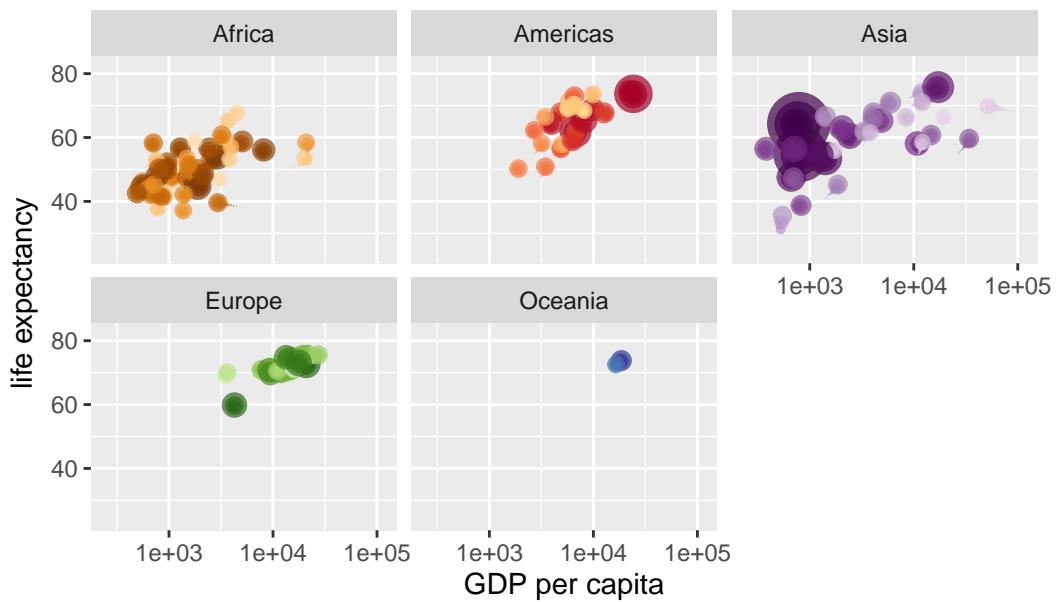
Year: 1977



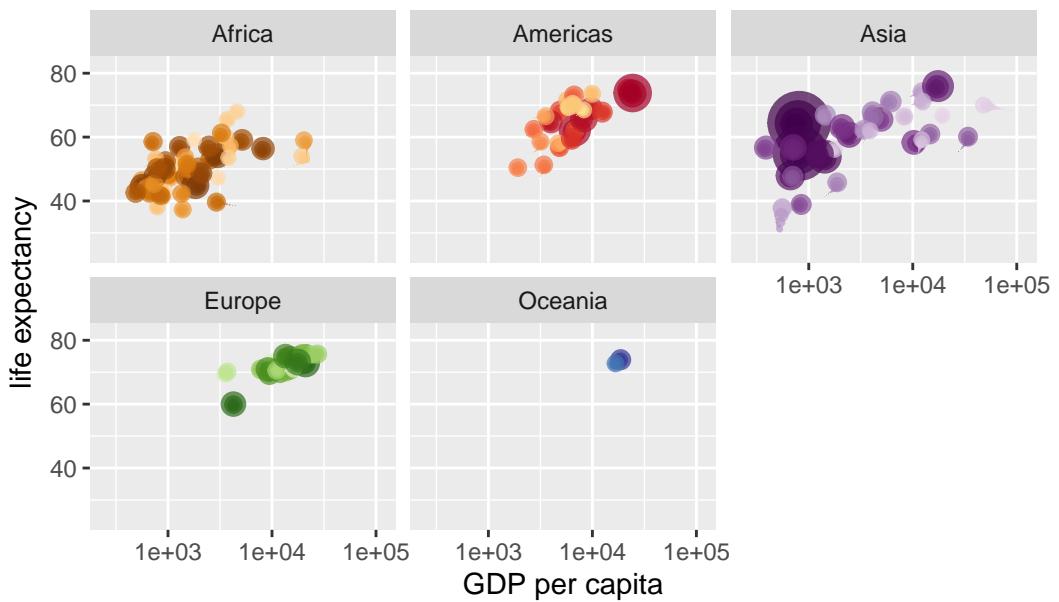
Year: 1978



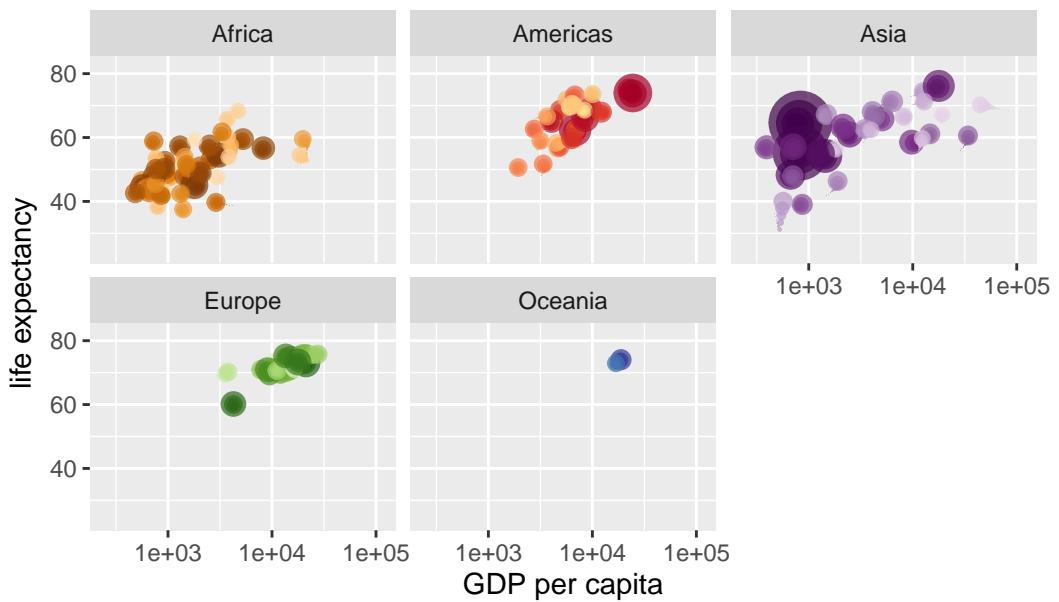
Year: 1978



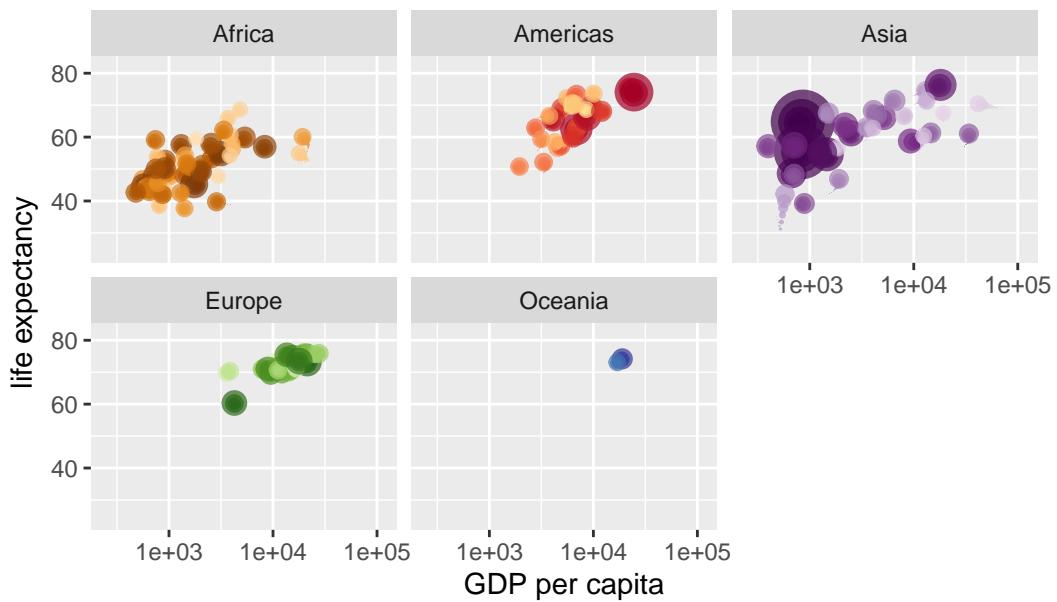
Year: 1979



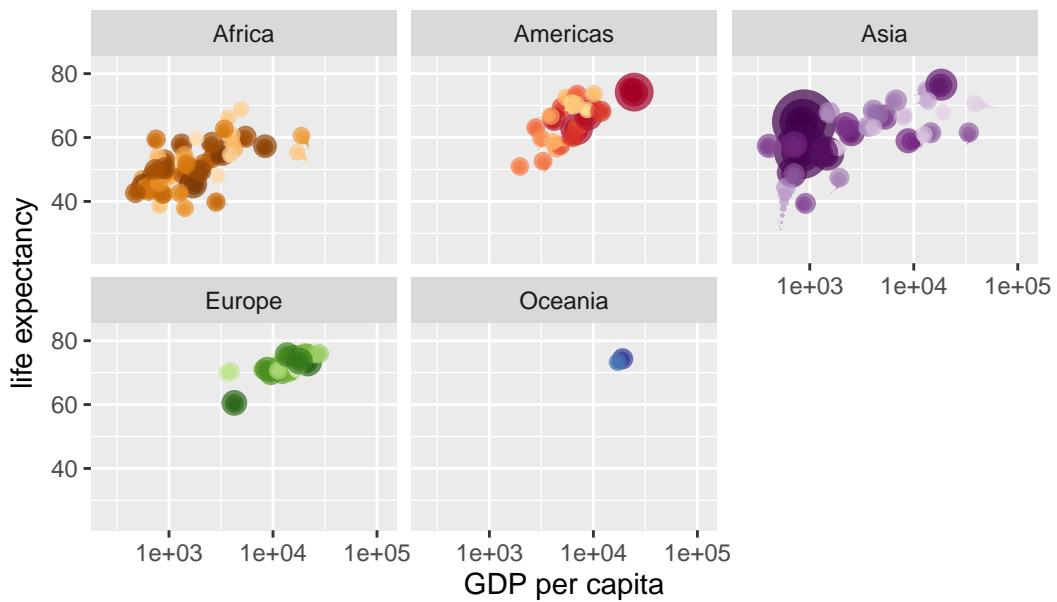
Year: 1979



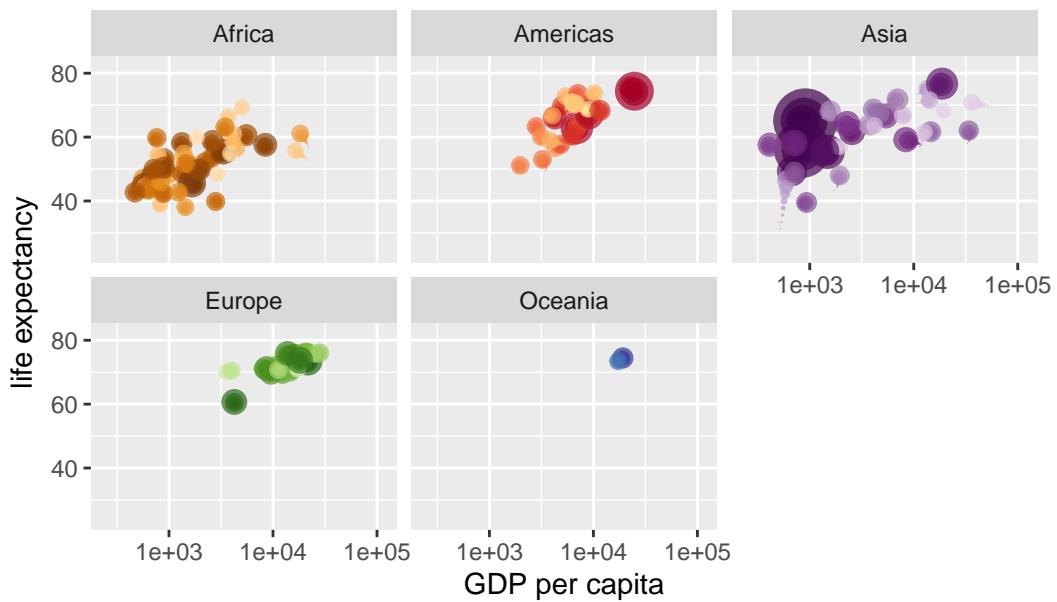
Year: 1980



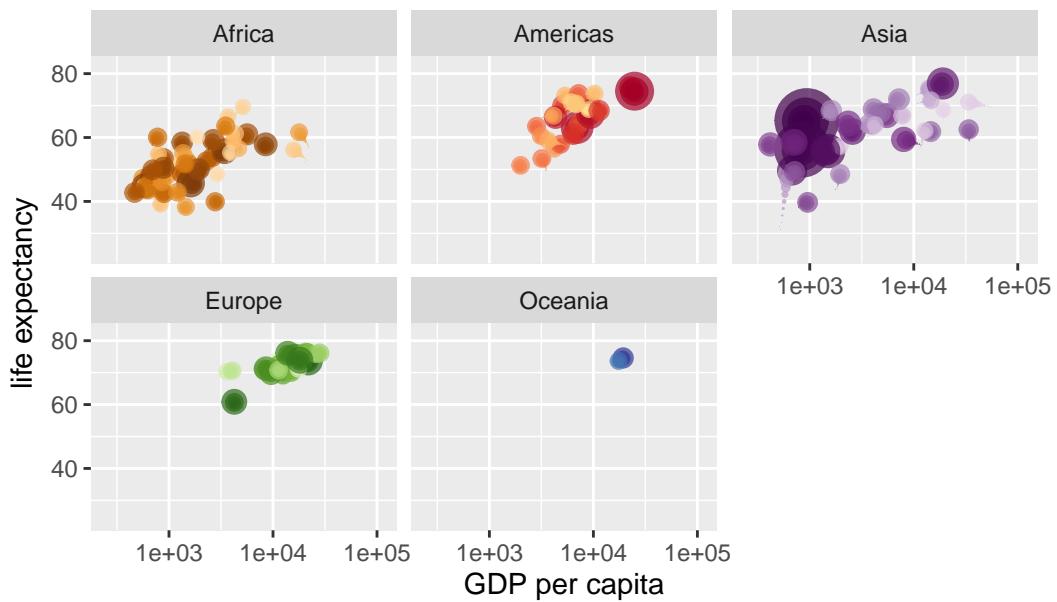
Year: 1980



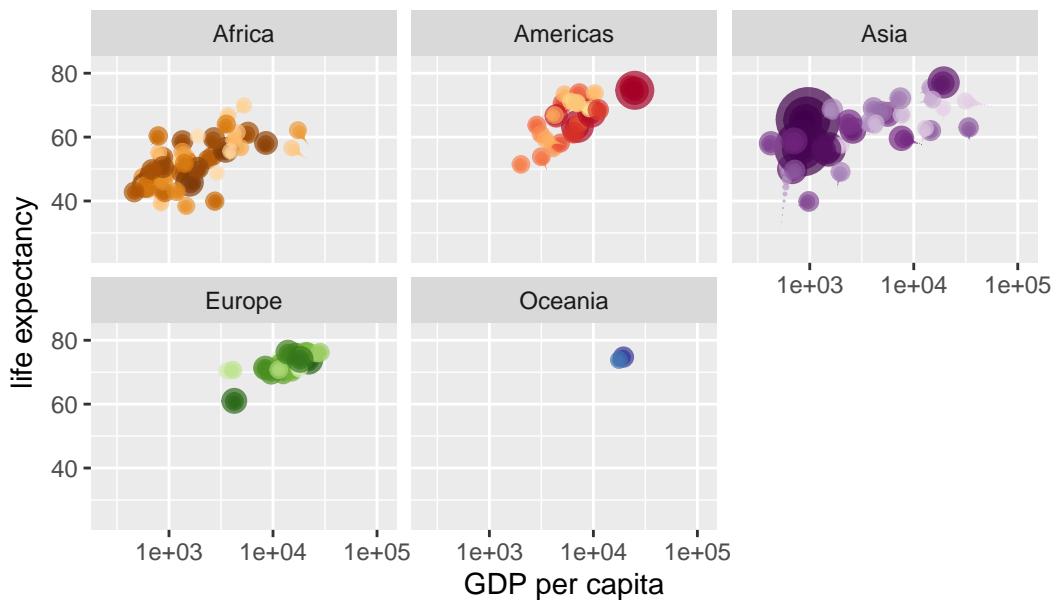
Year: 1981



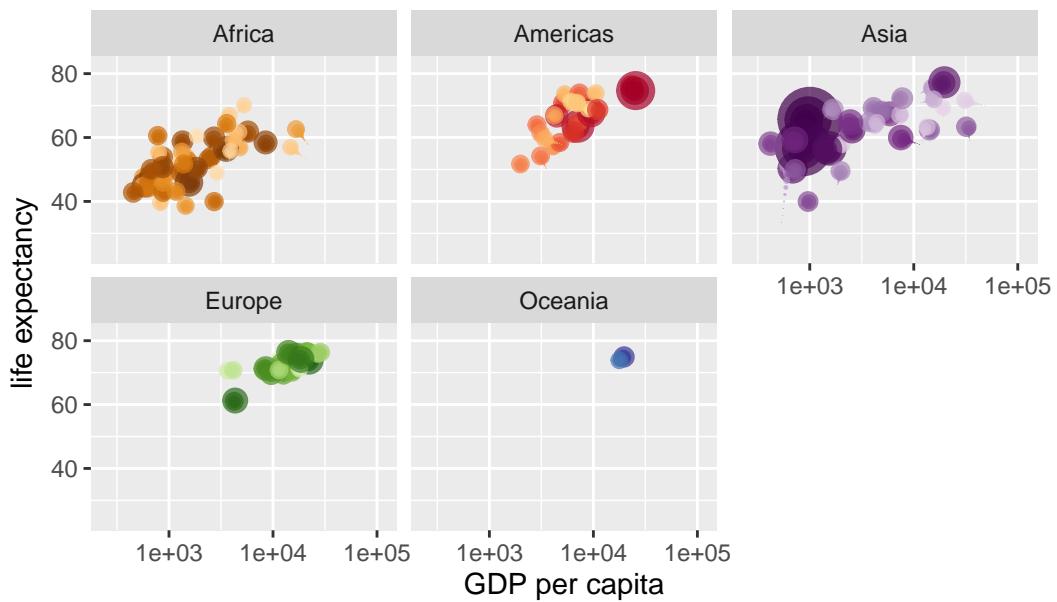
Year: 1981



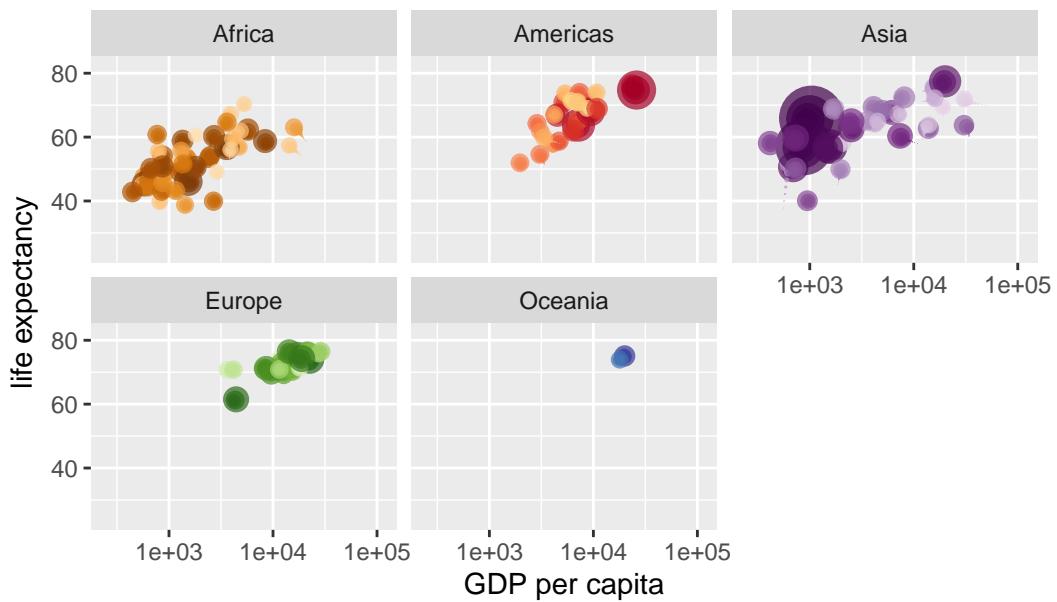
Year: 1982



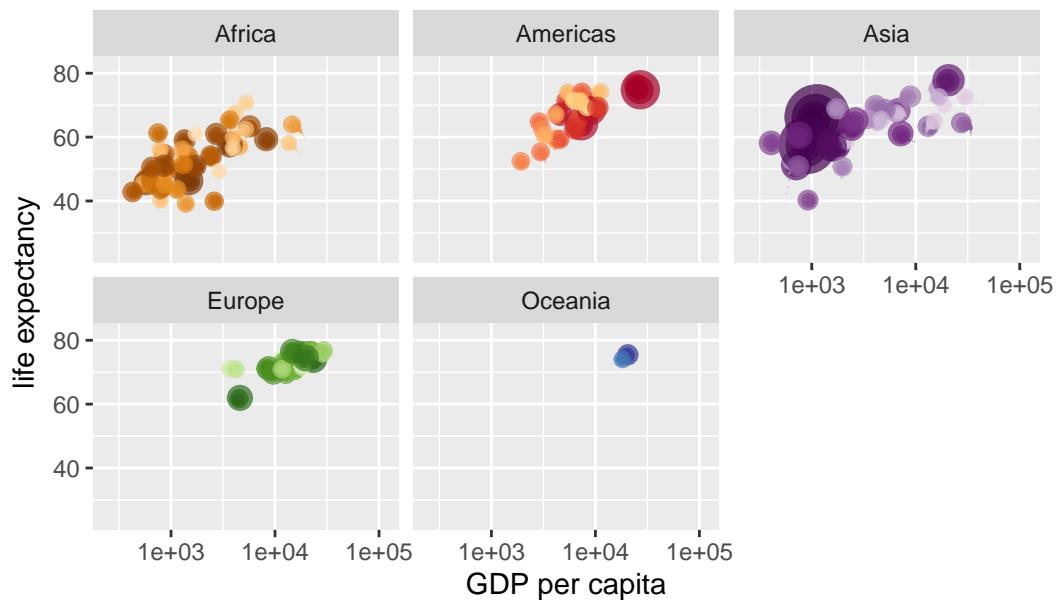
Year: 1983



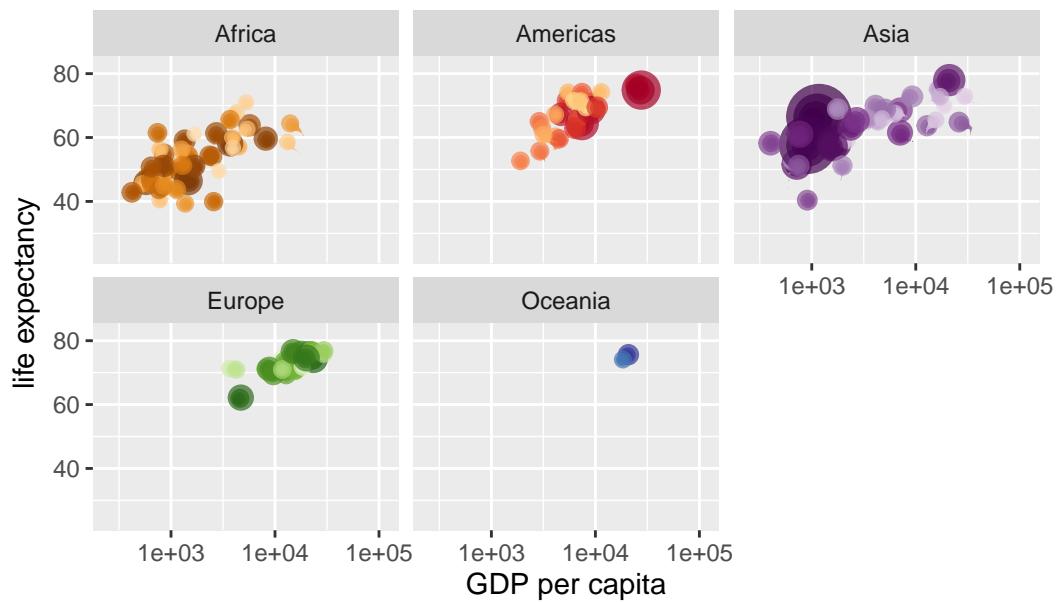
Year: 1983



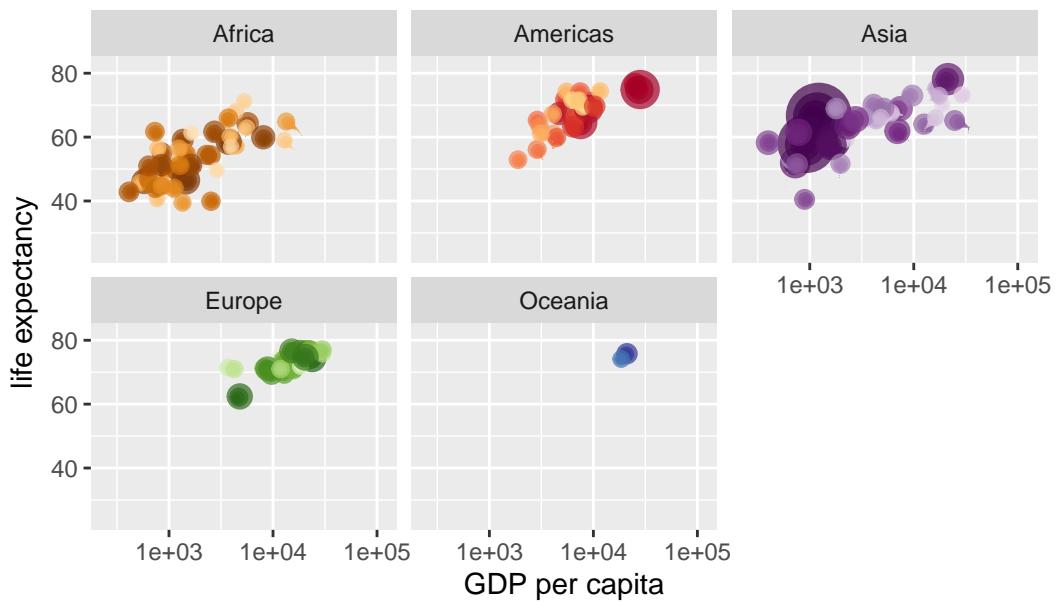
Year: 1984



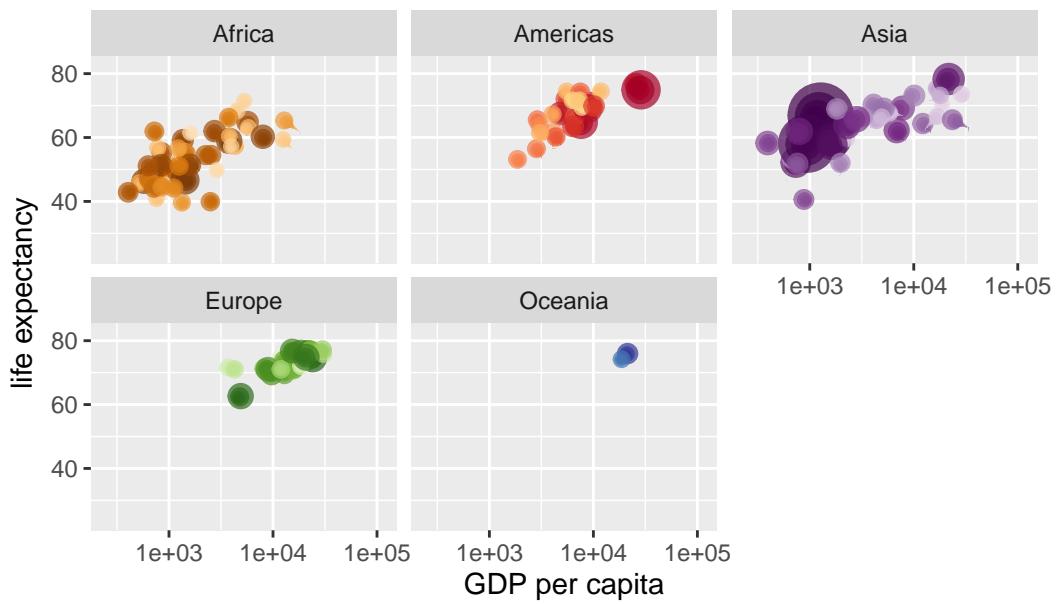
Year: 1985



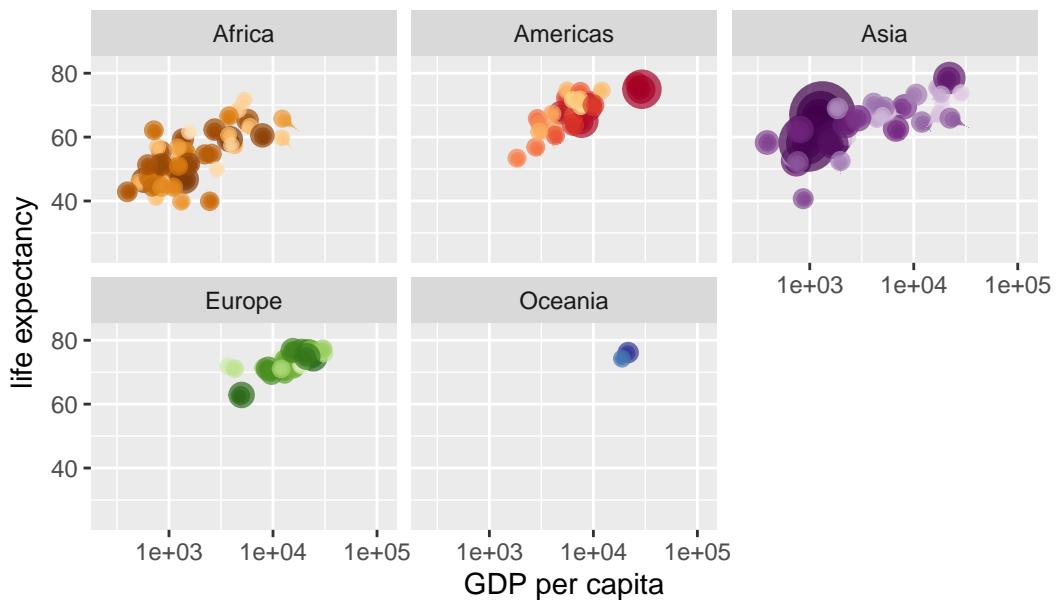
Year: 1985



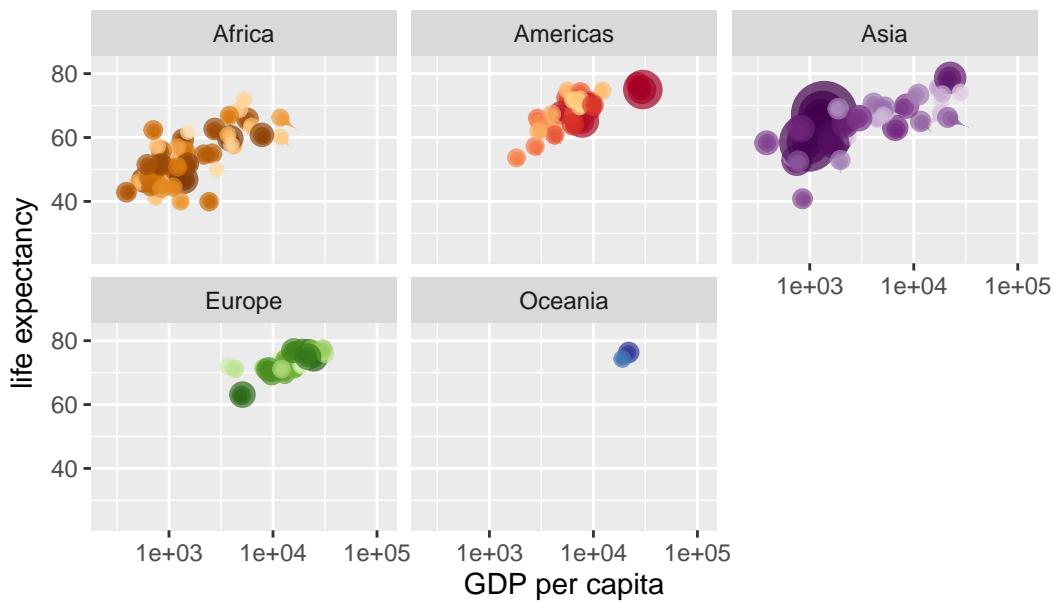
Year: 1986



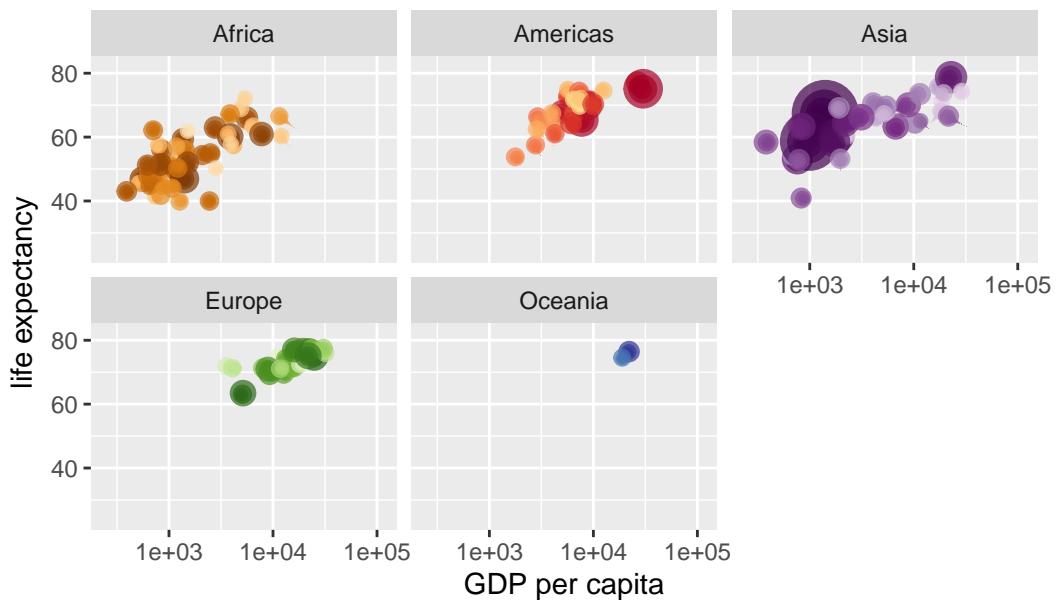
Year: 1986



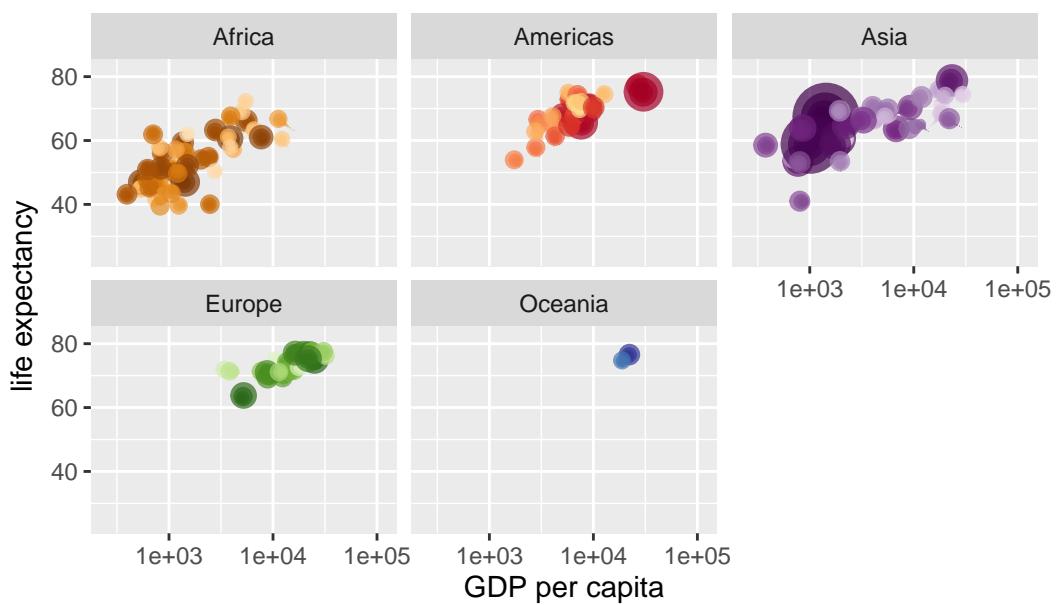
Year: 1987



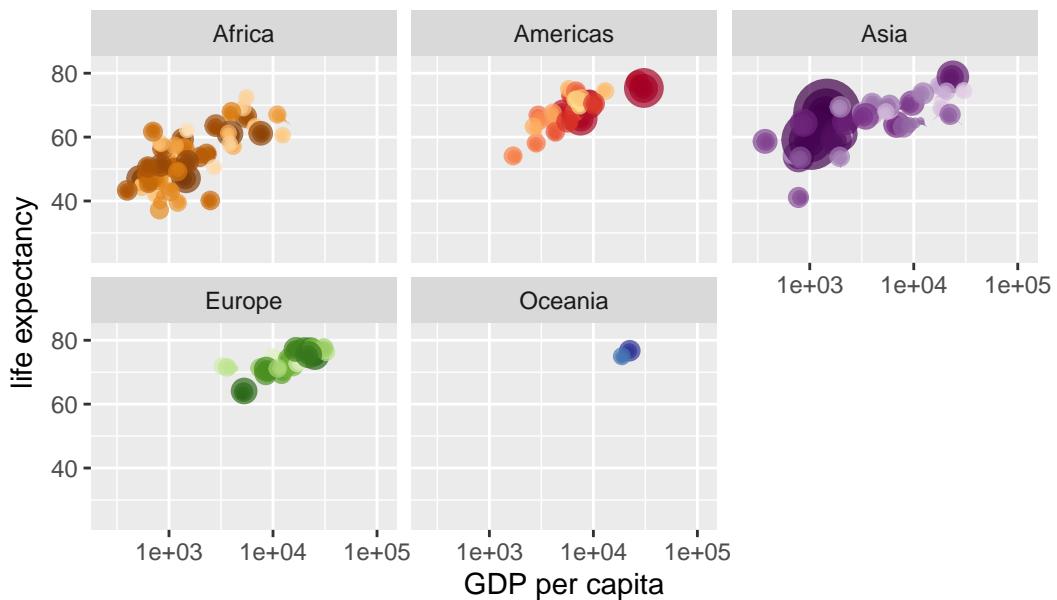
Year: 1988



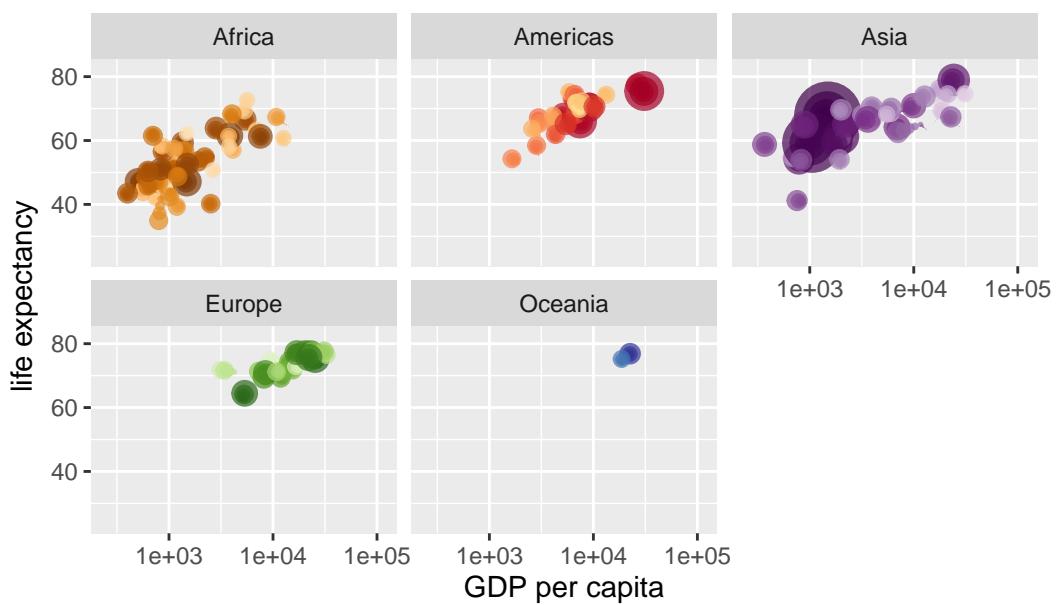
Year: 1988



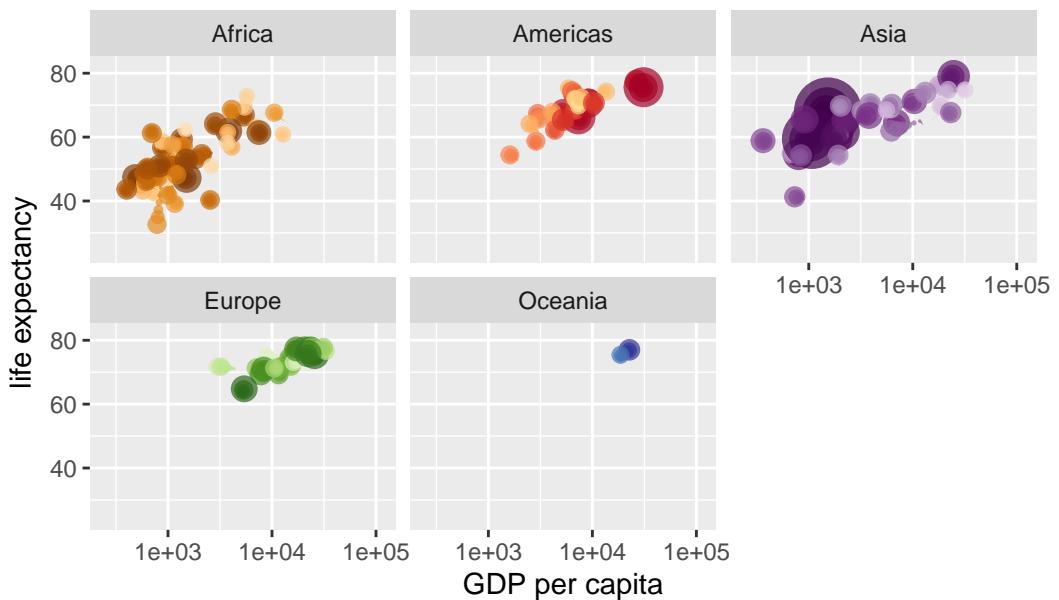
Year: 1989



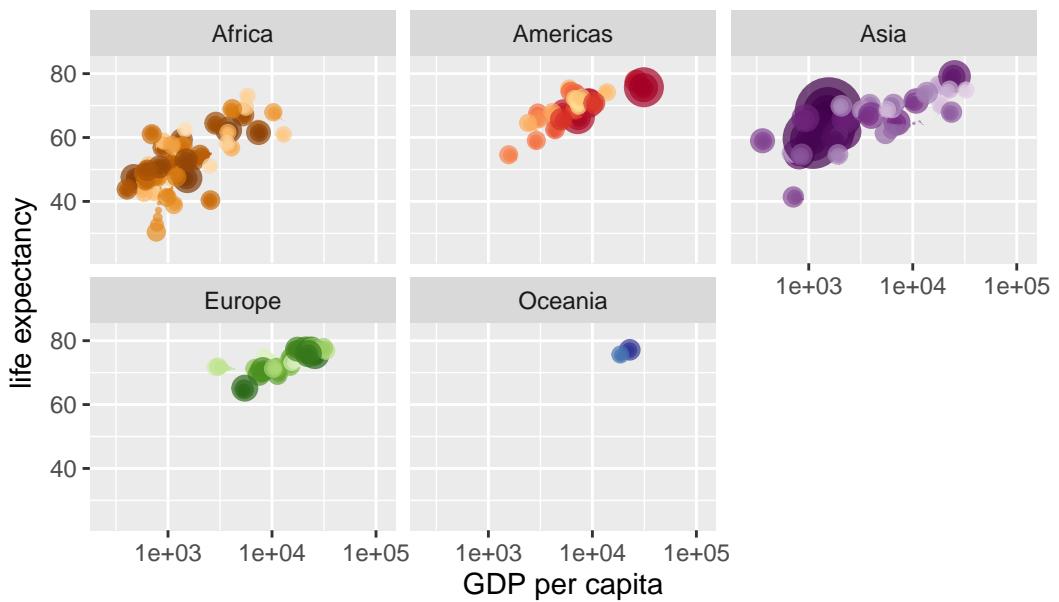
Year: 1989



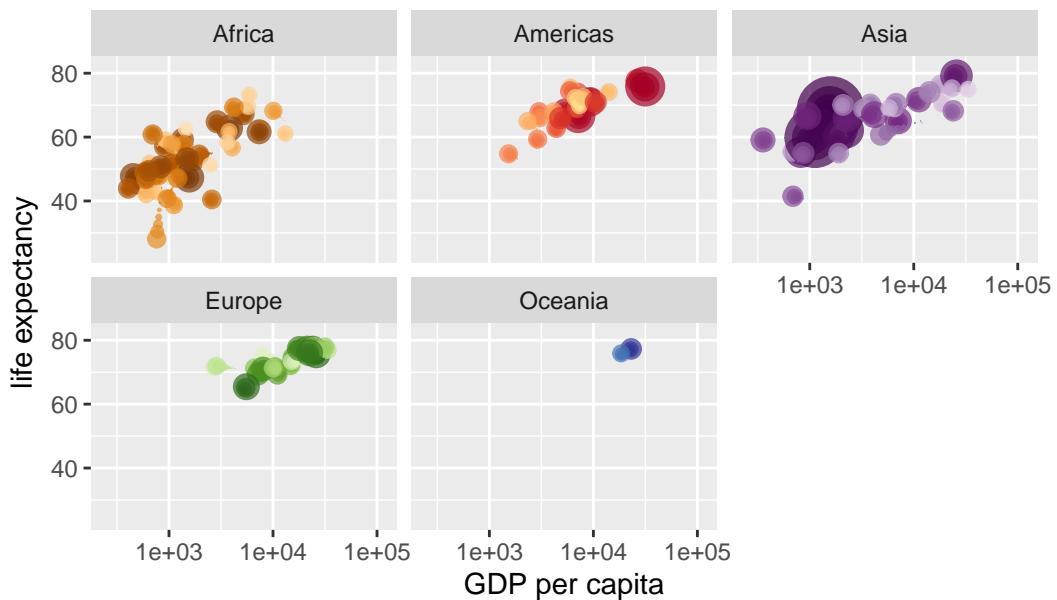
Year: 1990



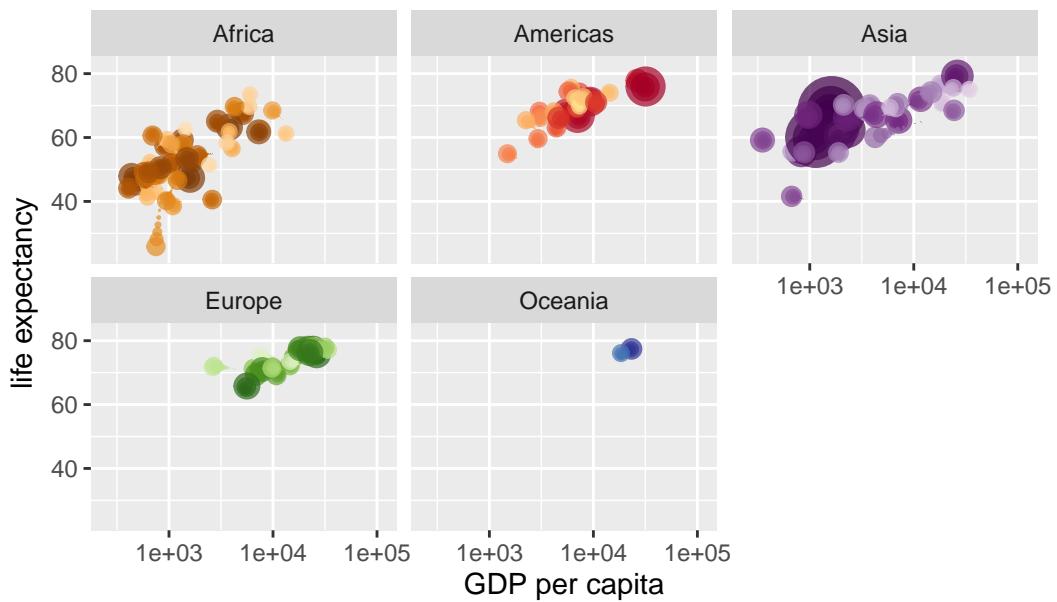
Year: 1990



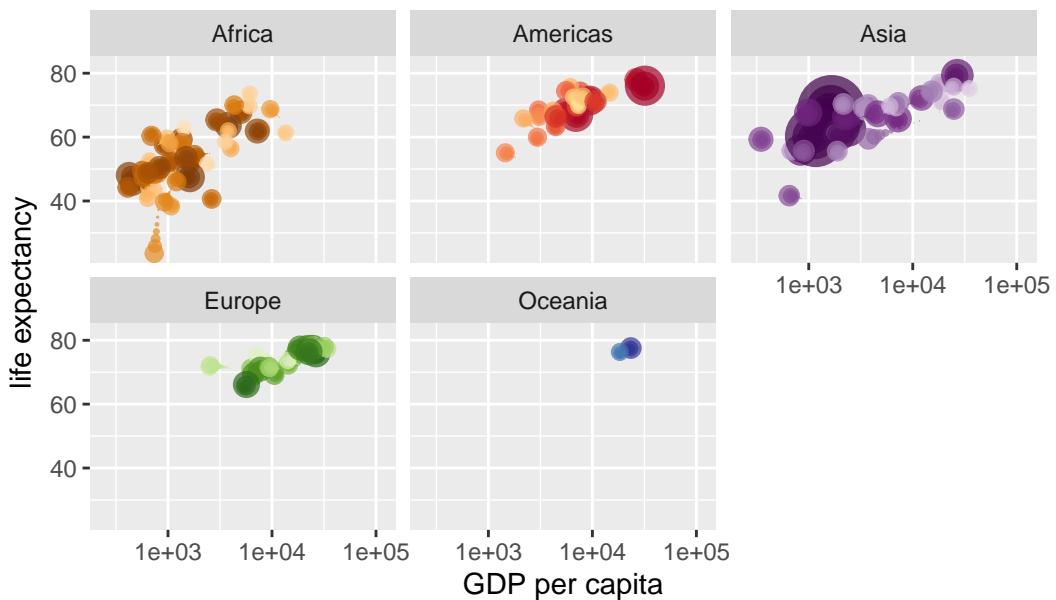
Year: 1991



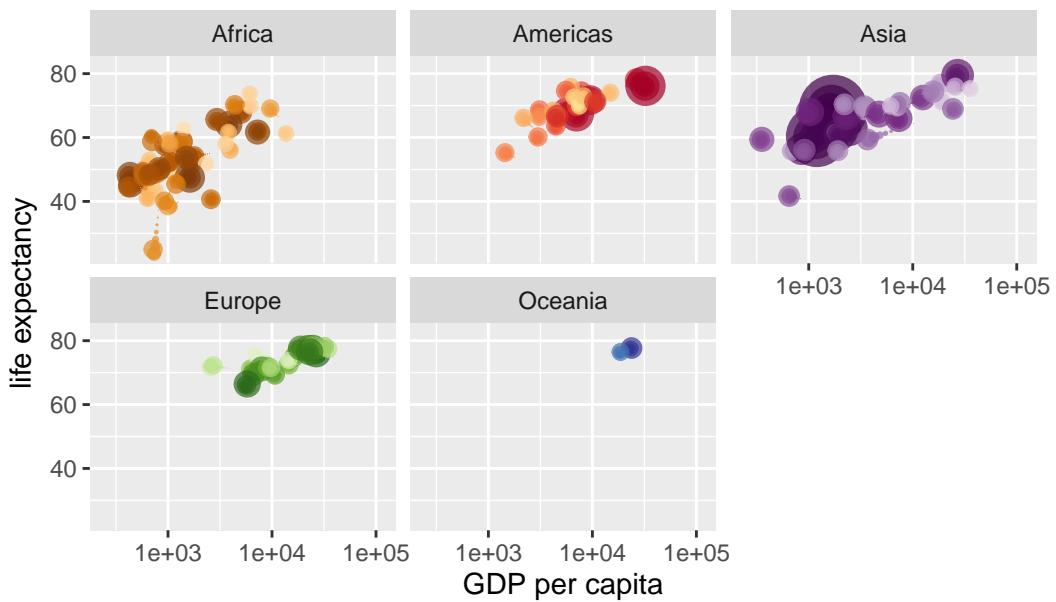
Year: 1991



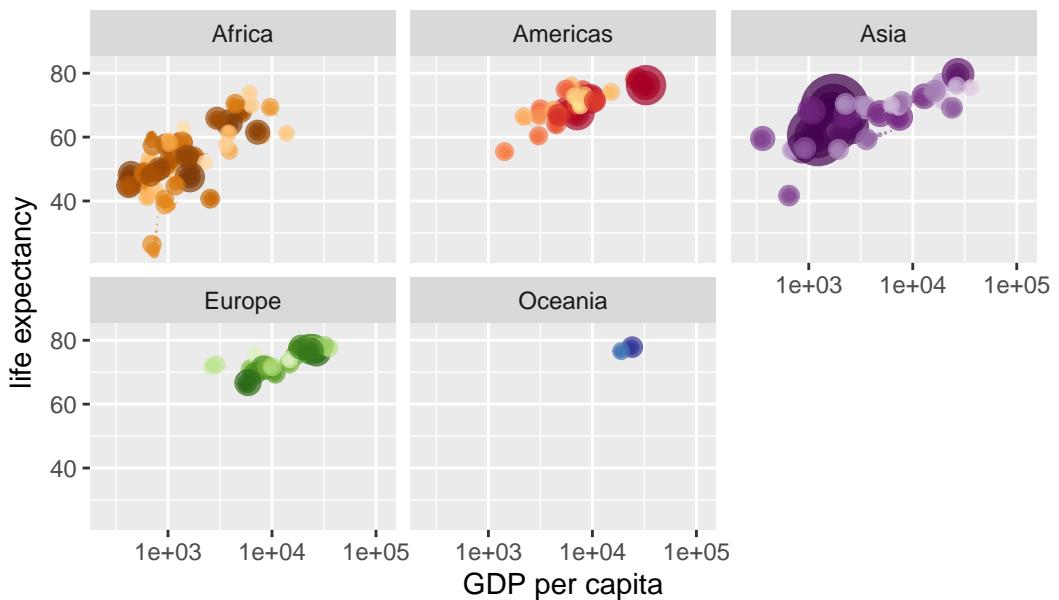
Year: 1992



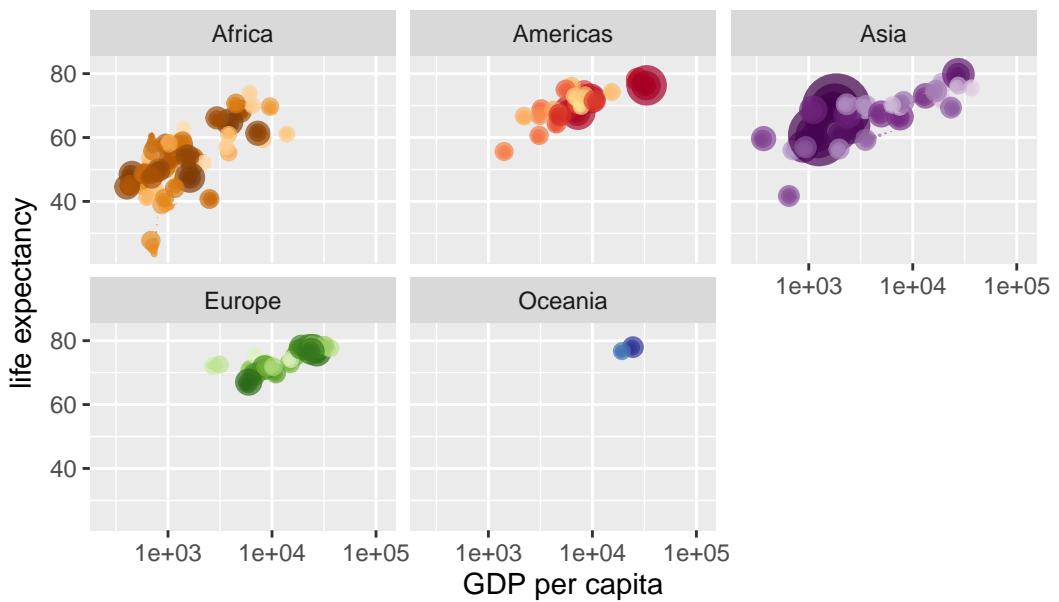
Year: 1993



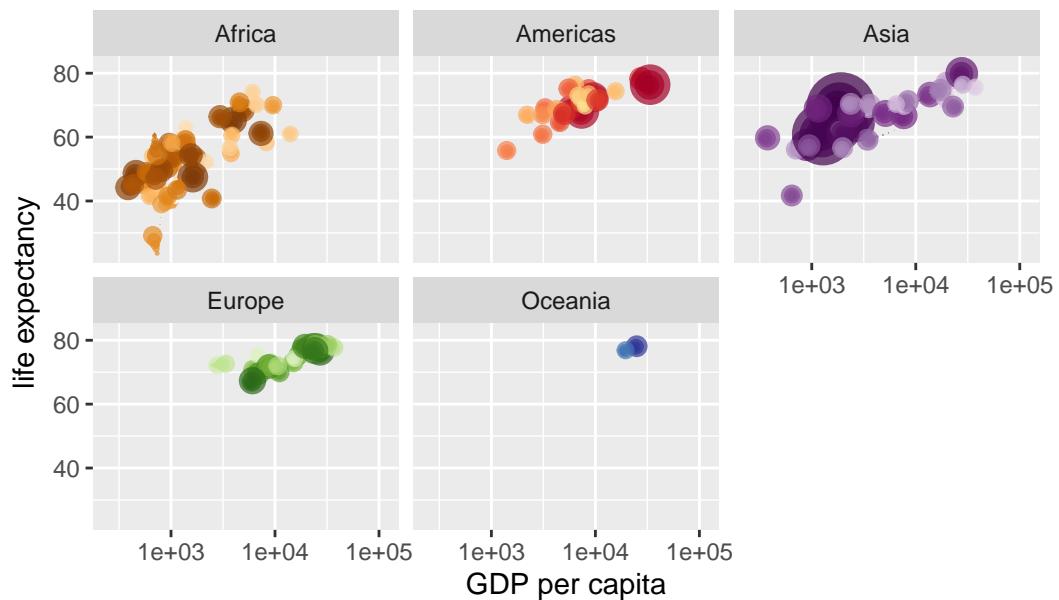
Year: 1993



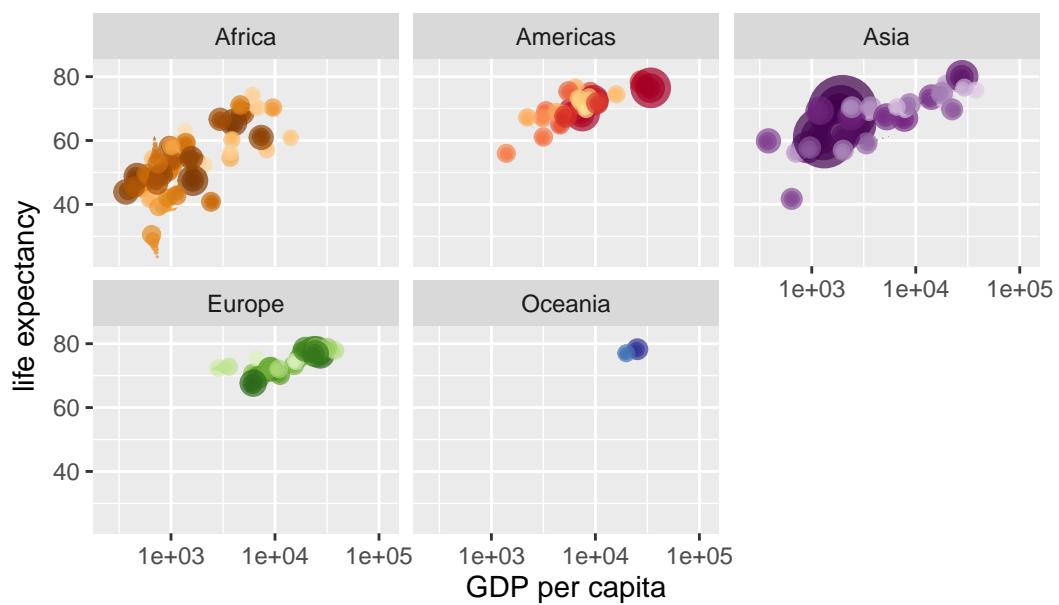
Year: 1994



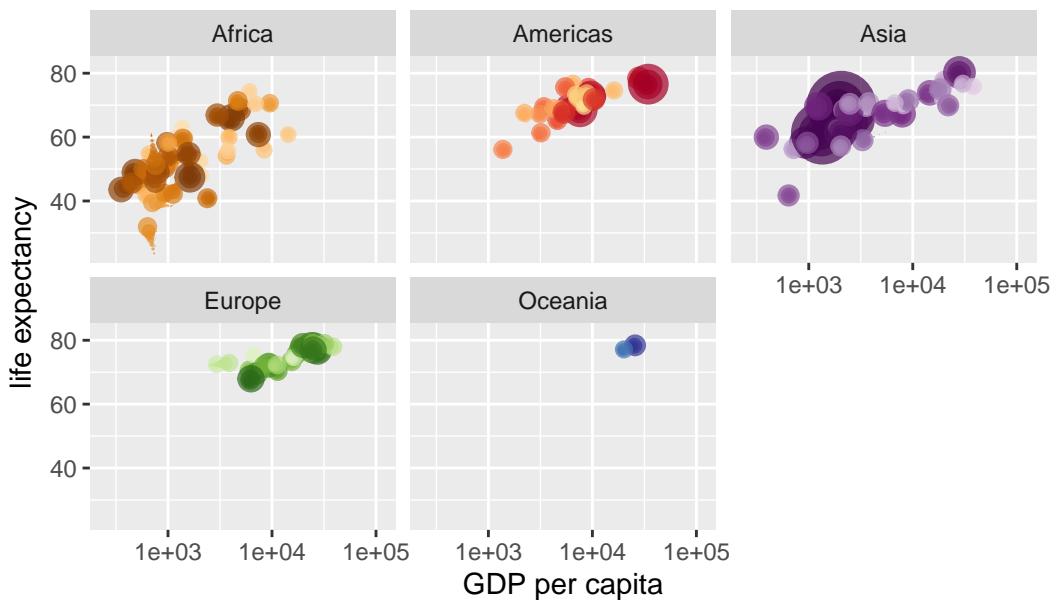
Year: 1994



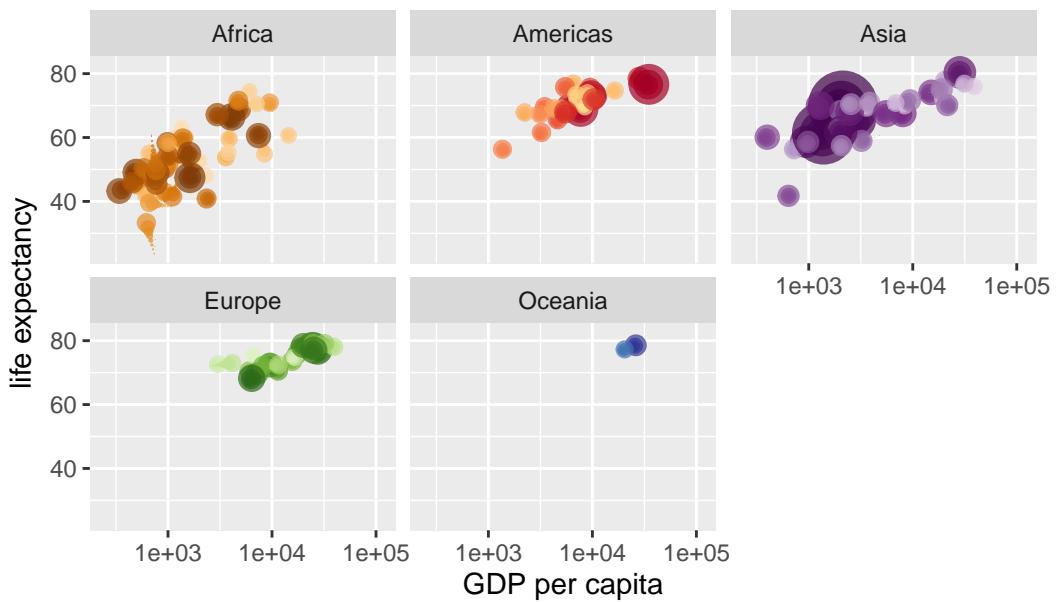
Year: 1995



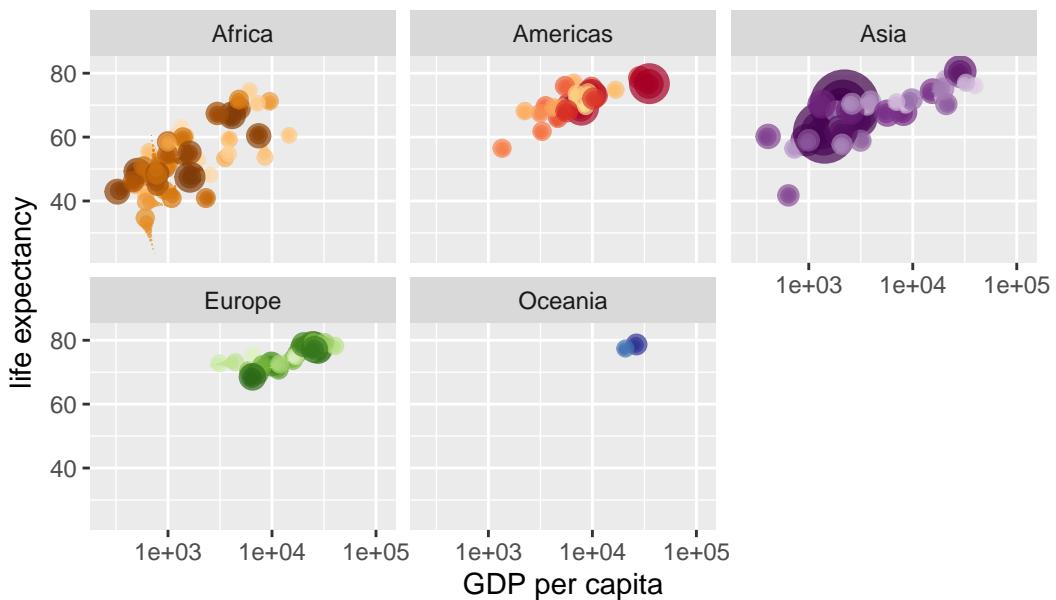
Year: 1995



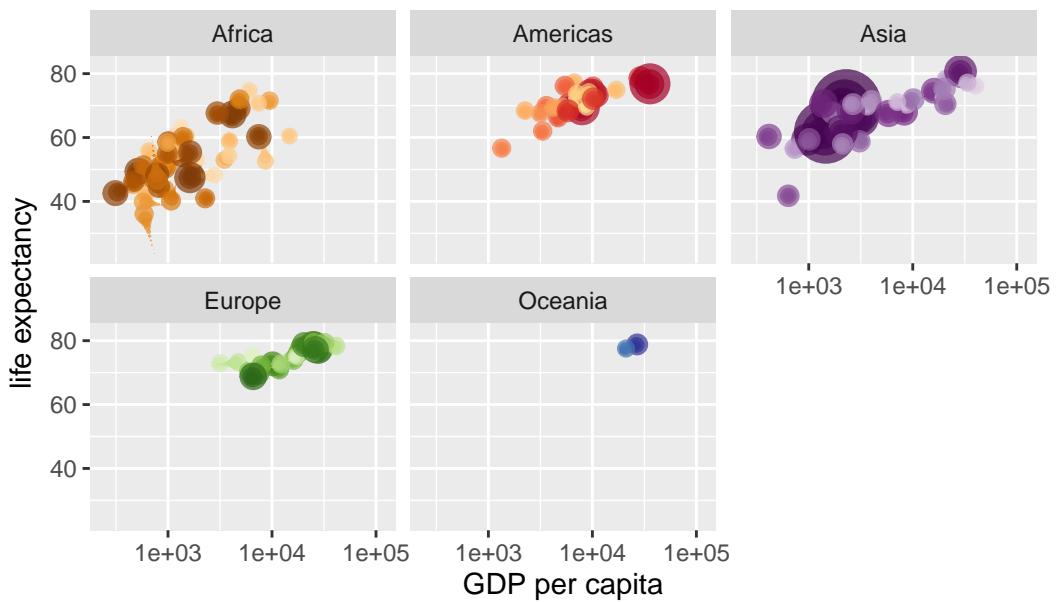
Year: 1996



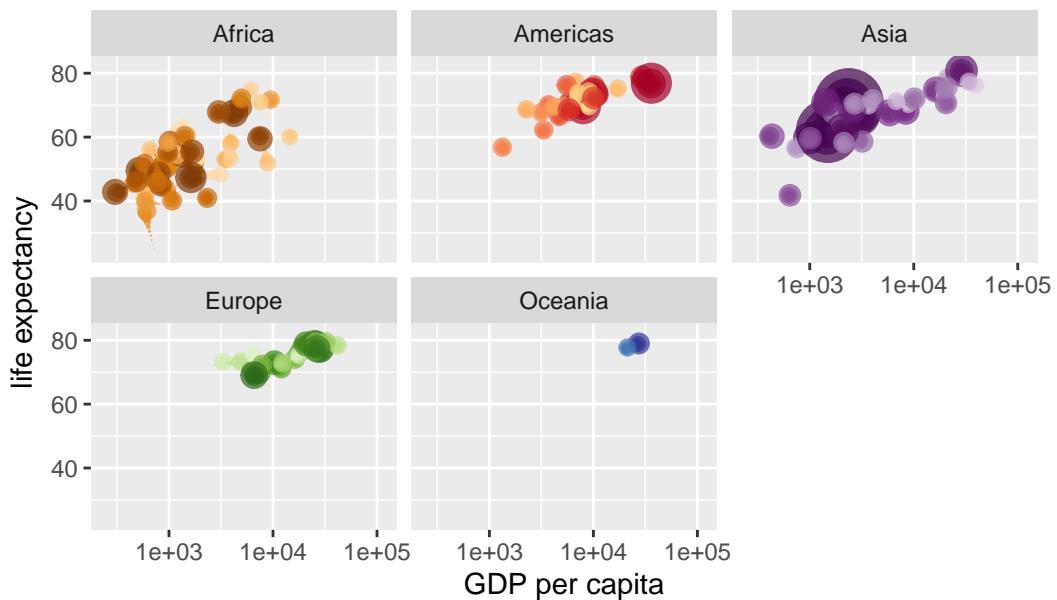
Year: 1996



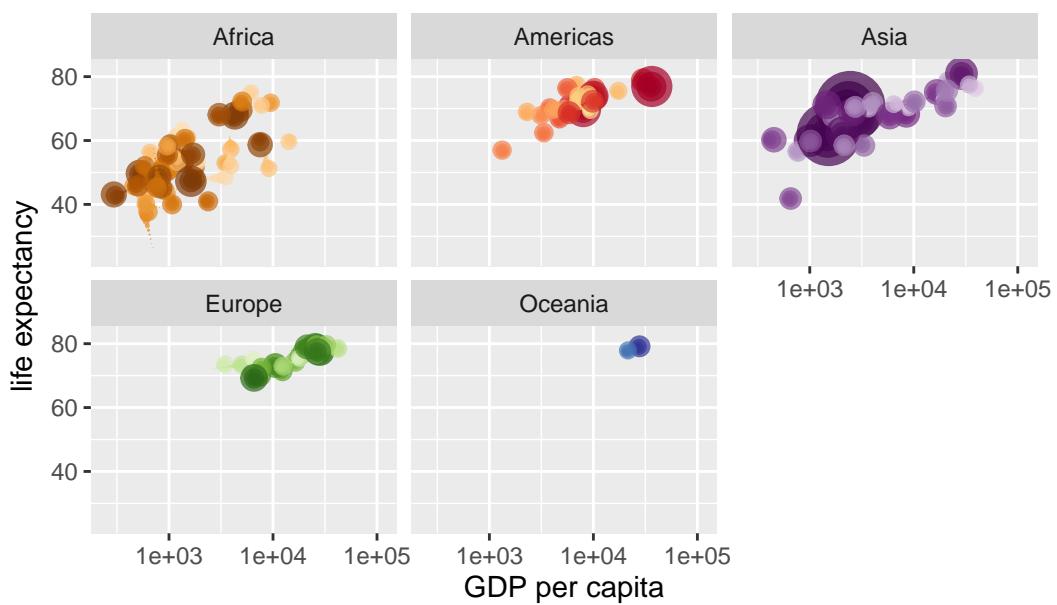
Year: 1997



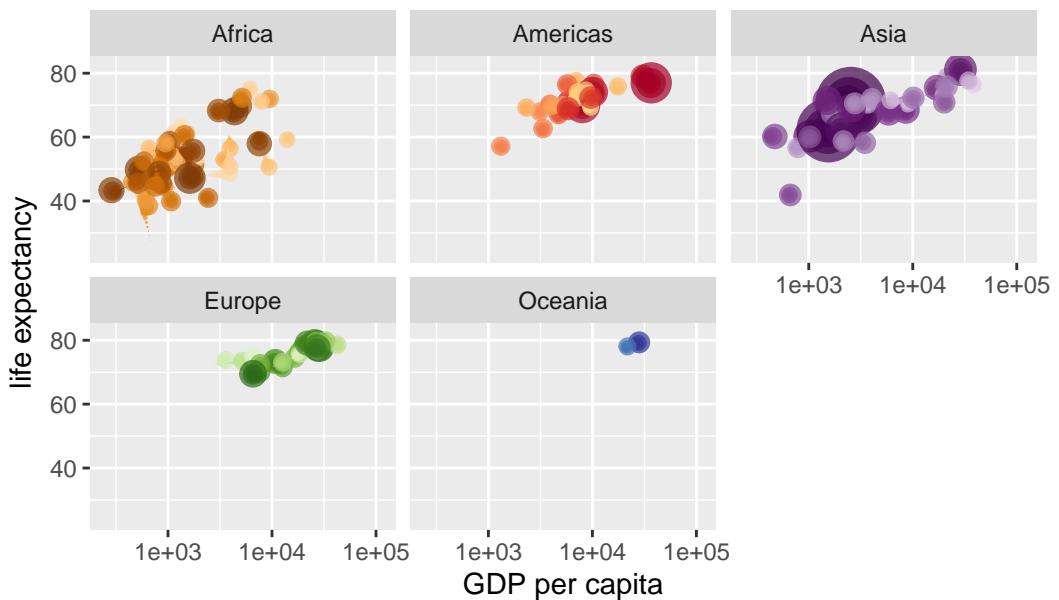
Year: 1998



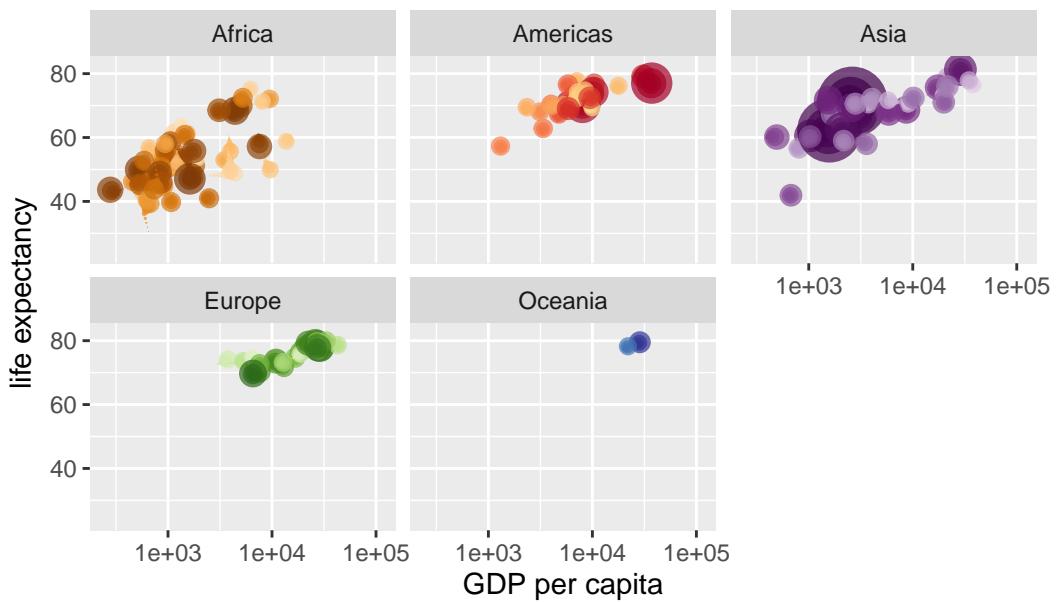
Year: 1998



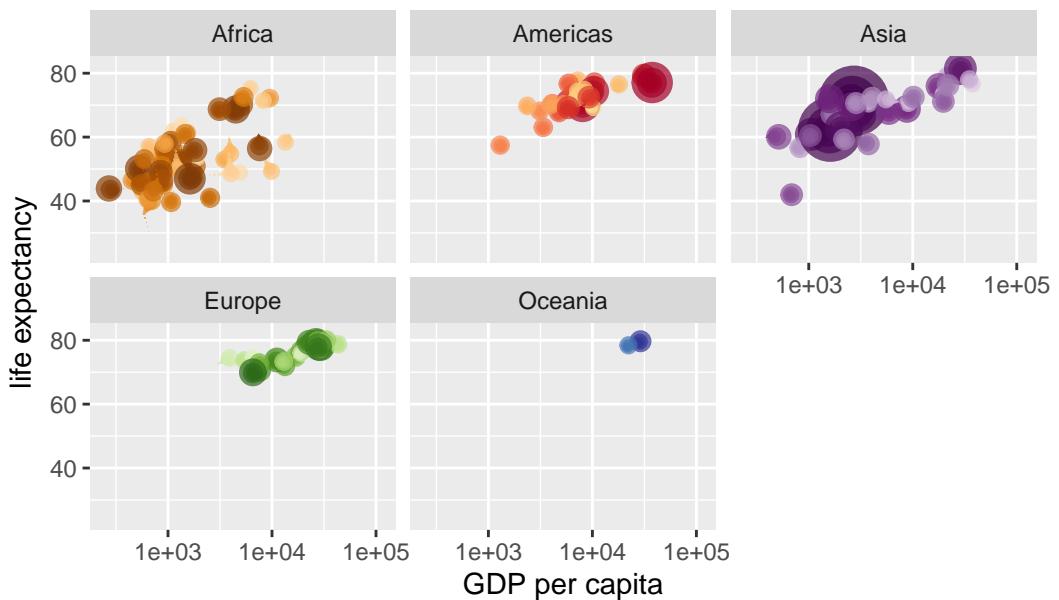
Year: 1999



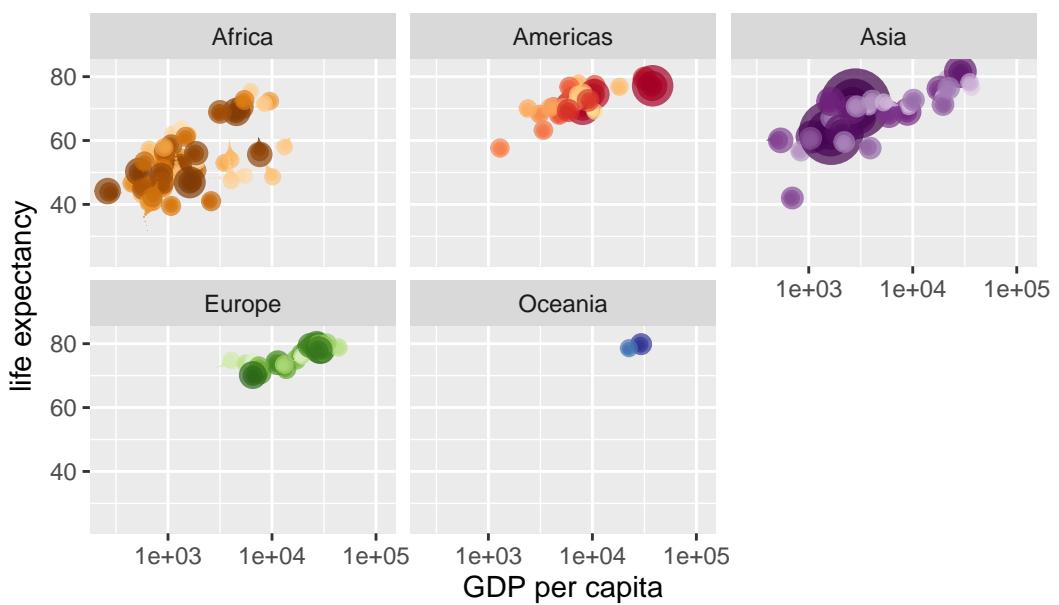
Year: 1999



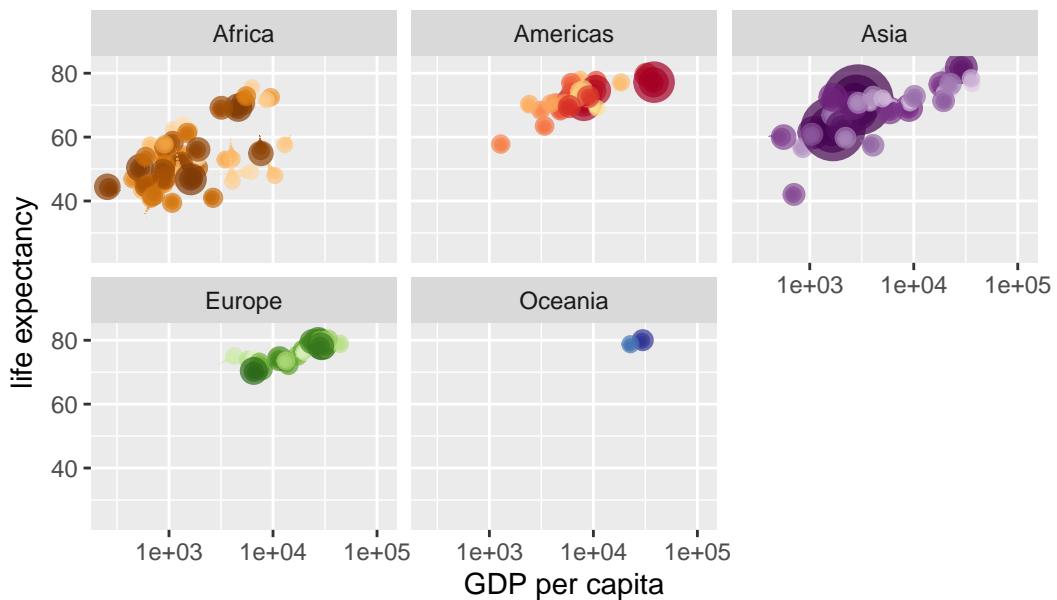
Year: 2000



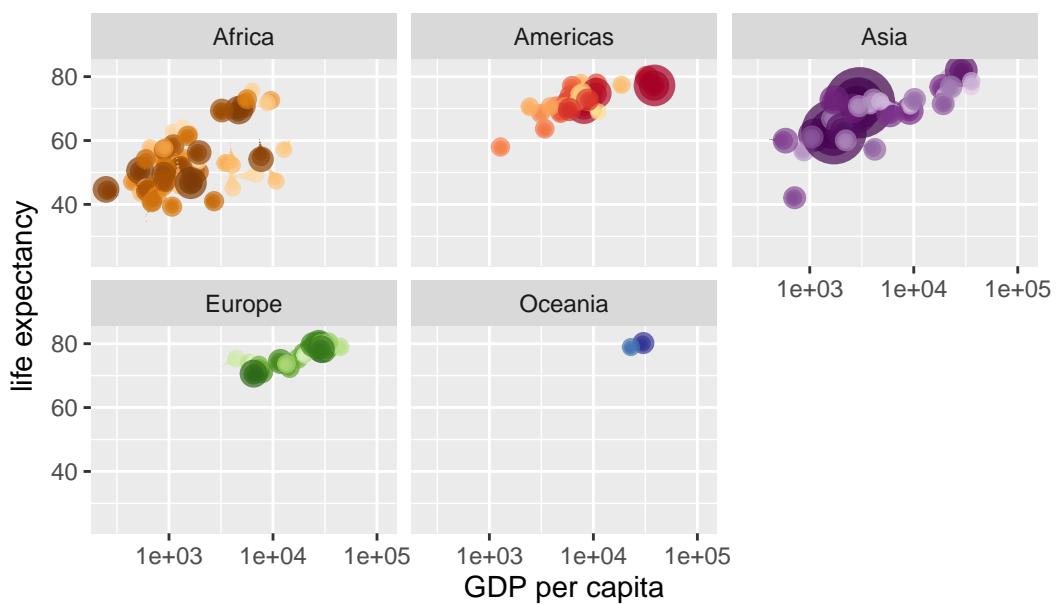
Year: 2000



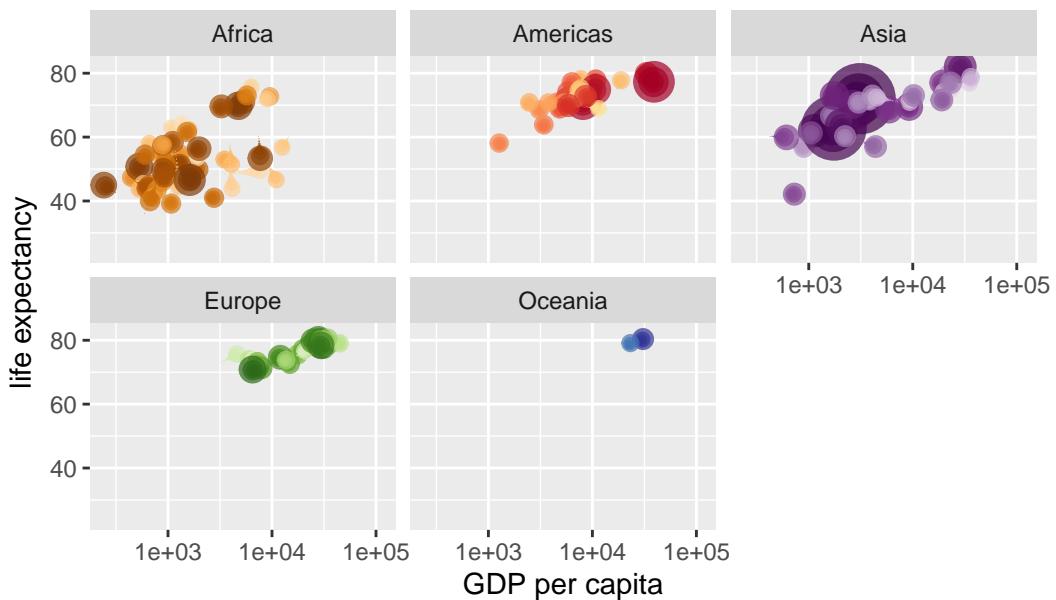
Year: 2001



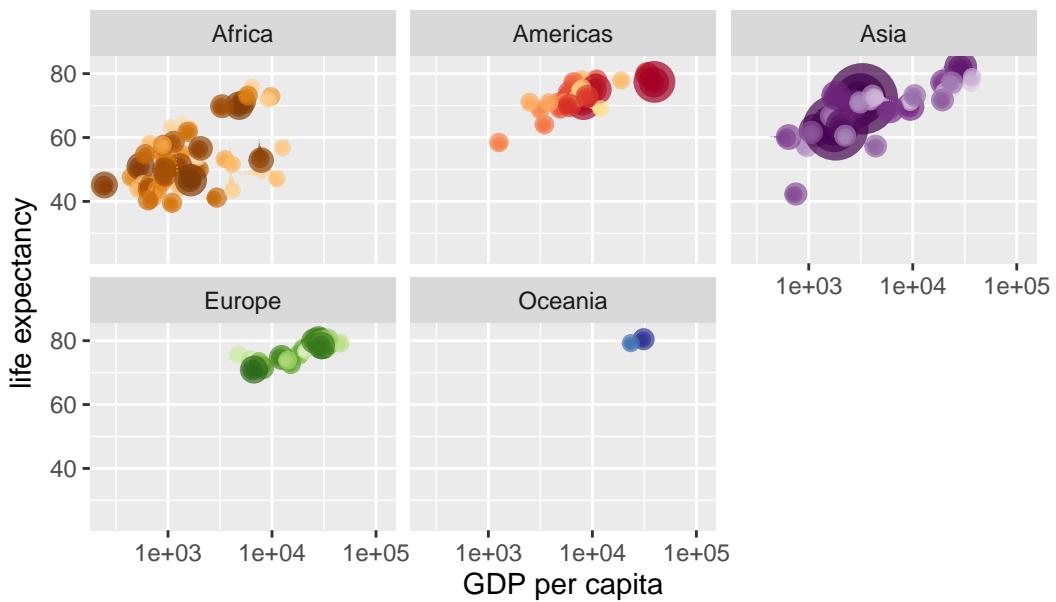
Year: 2001



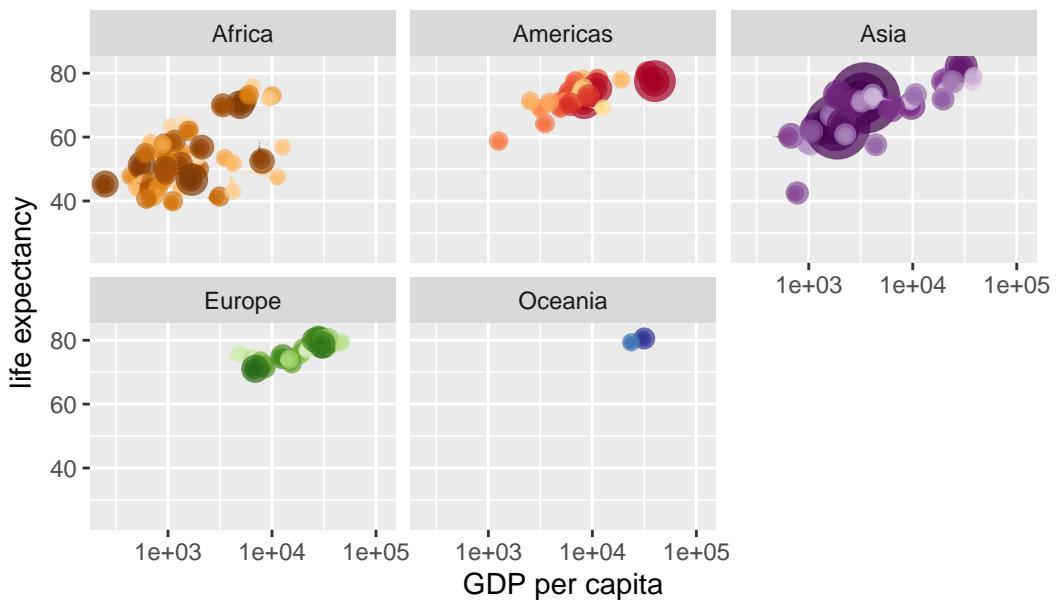
Year: 2002



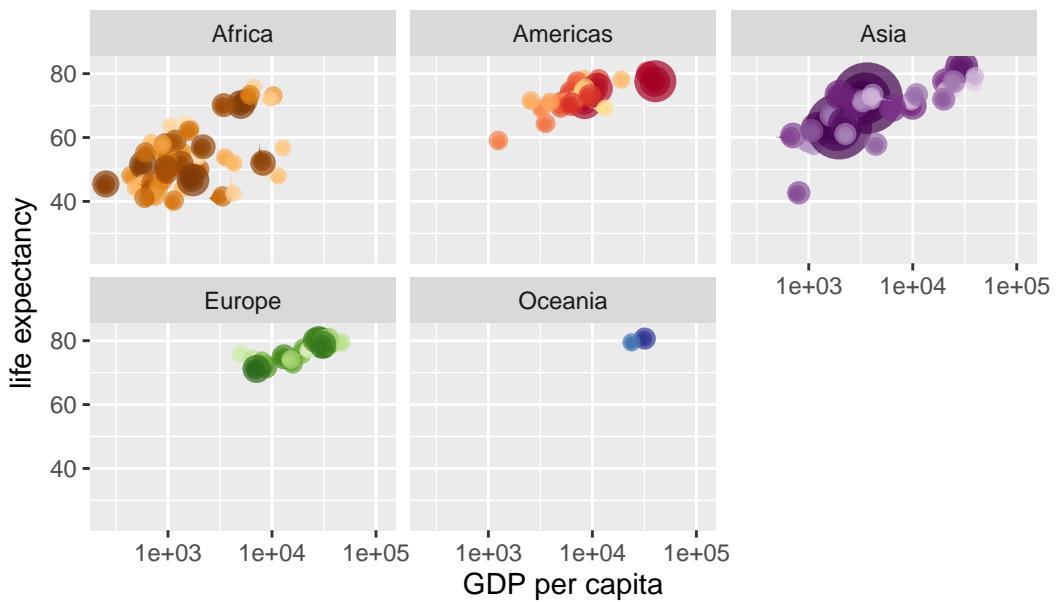
Year: 2003



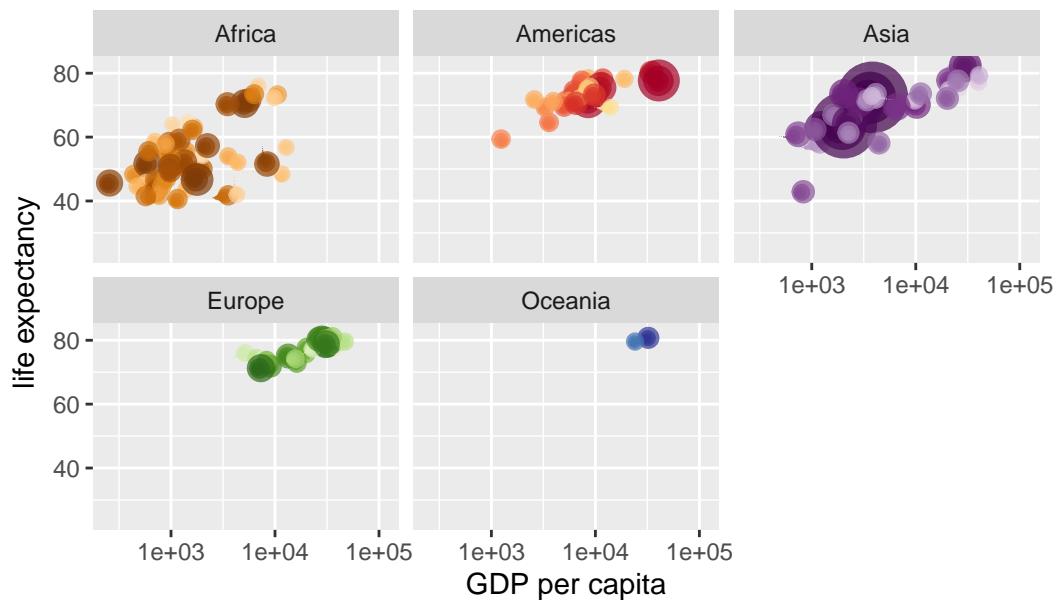
Year: 2003



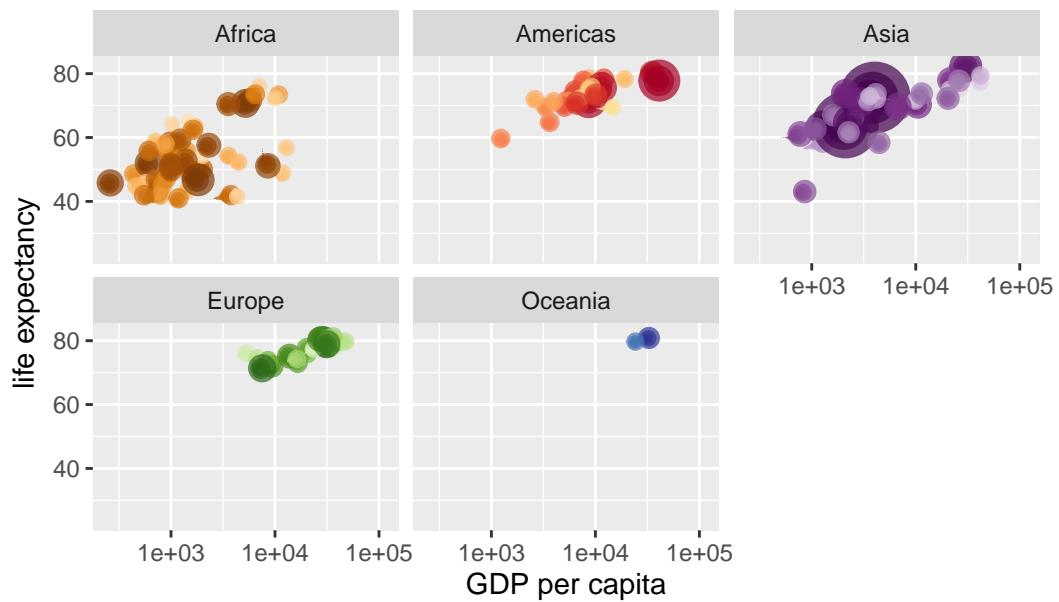
Year: 2004



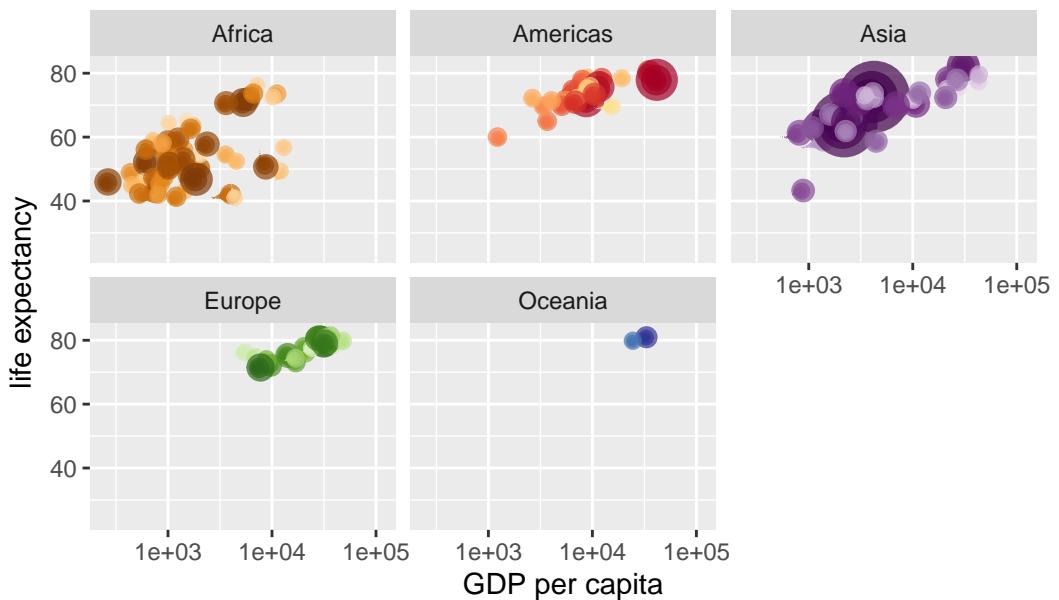
Year: 2004



Year: 2005



Year: 2005



Year: 2006

