

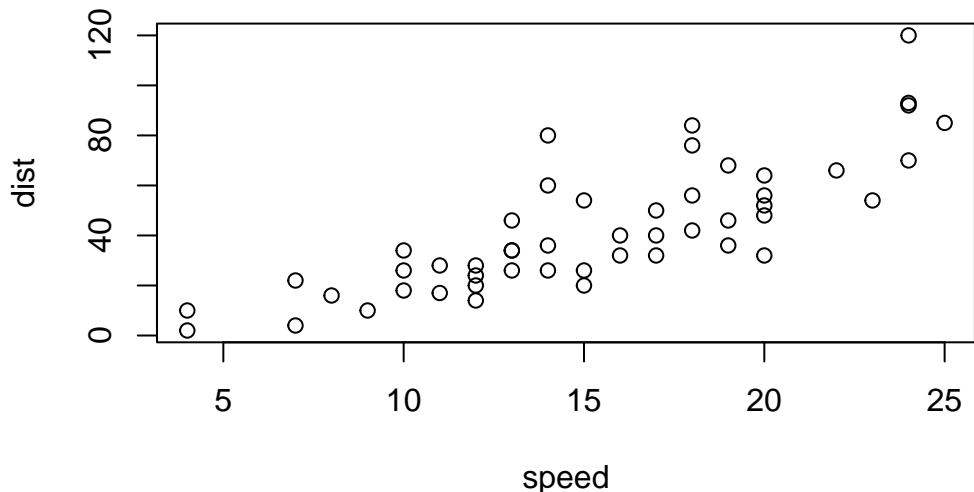
Class 5: Data Viz with ggplot

Sabrina Wu (A16731683)

Plotting in R

R has lots of ways to make plots and figures. This includes so-called **base** graphics and packages like **ggplot2**

```
plot(cars)
```



This is a **base** R plot of the in-built **cars** dataset that has only two columns:

```
head(cars)
```

```
speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
```

Q. How would we plot this wee dataset with **ggplot2**

All ggplot figures have at least 3 layers:

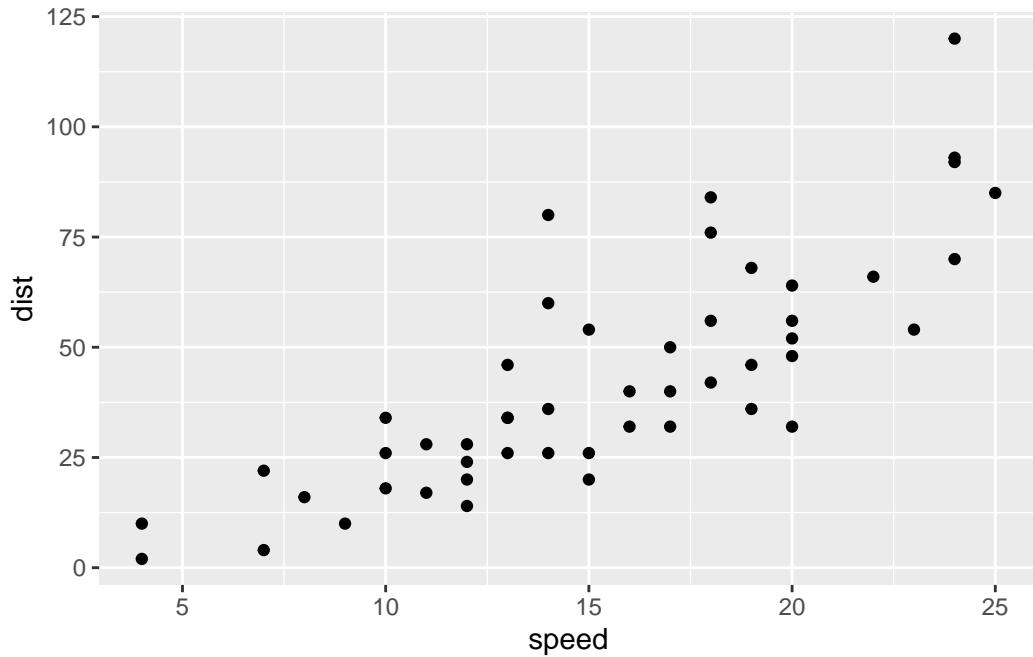
- **data**
- **aesthetics** (how the data map to the plot)
- **geomtry** (how we draw the plot, line, points, etc)

Before I use any new package I need to download and install it with the `install.packages()` command.

I never use `install.package()` within my quarto document otherwise I will install the package over and over and over again - which is silly!

Once a package is installed I can load it up with the `library()` function.

```
# install.packages("ggplot2")
library(ggplot2)
ggplot(cars) +
  aes(x=speed, y=dist) +
  geom_point()
```



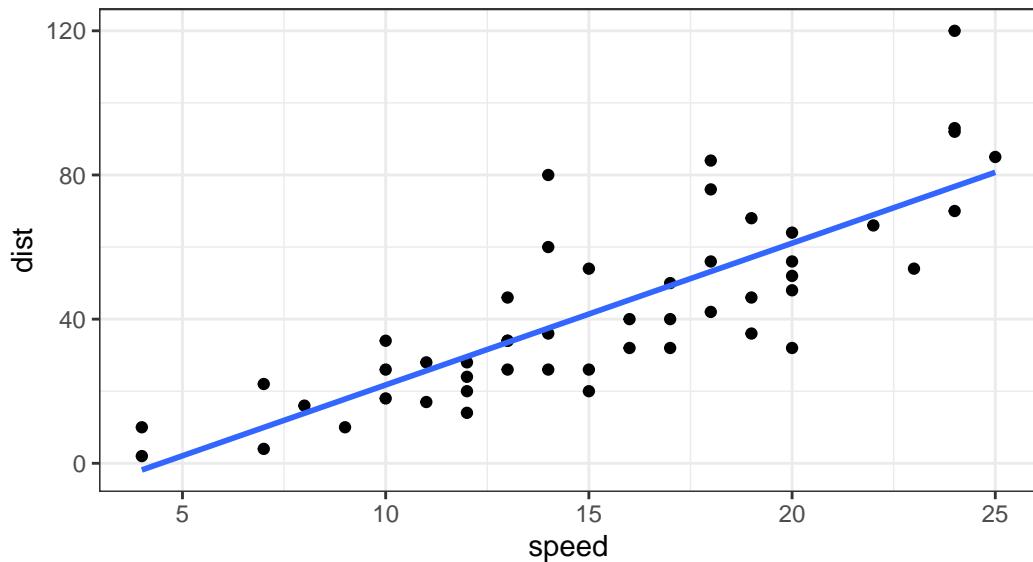
Key-point: FOr simple plots (like the one above) ggplot is more verbose (we need to do more typing) but as plots get more complicated ggplot starts to be more clear and simplae than base R plot()

```
p <- ggplot(cars) +
  aes(speed, dist) +
  geom_point() +
  geom_smooth(method = lm, se=FALSE) +
  labs(title="Stopping distance of old cars",
       subtitle = "From the inbuilt cars dataset") +
  theme_bw()
p

`geom_smooth()` using formula = 'y ~ x'
```

Stopping distance of old cars

From the inbuilt cars dataset



Graph for expression analysis for anti-viral drugs

Calling up the data to plot

```
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 |

Finding number of rows, column, column names, number and fraction of up-regulated genes in dataset

```
nrow(genes)
```

```
[1] 5196
```

```
colnames(genes)
```

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

```
ncol(genes)
```

```
[1] 4
```

```
table(genes$State)
```

| | down | unchanging | up |
|--|------|------------|-----|
| | 72 | 4997 | 127 |

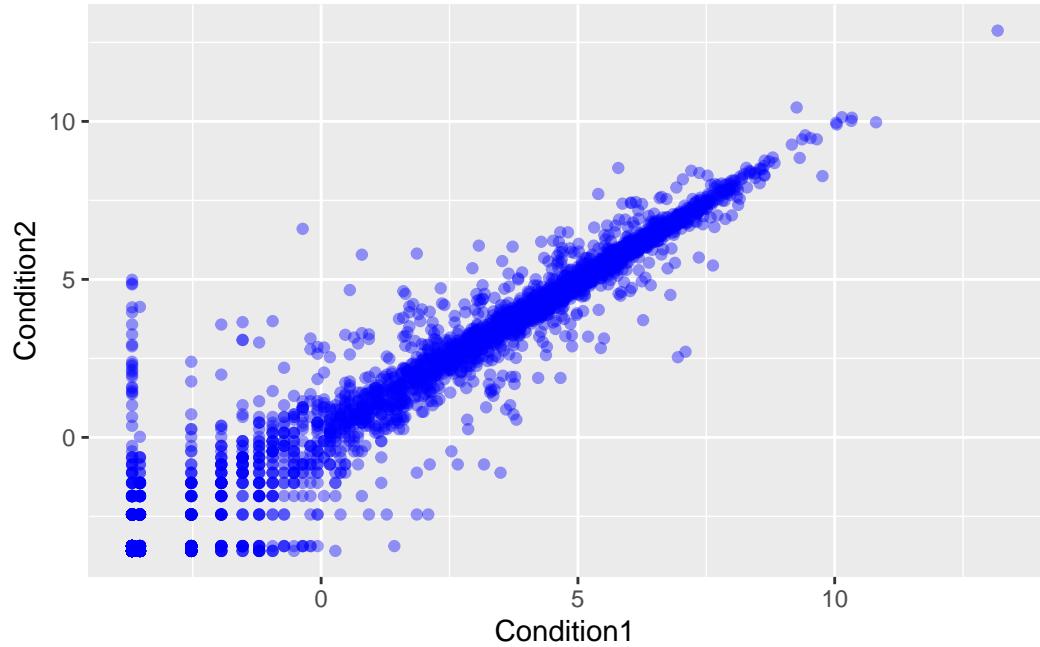
```
round(table(genes$State)/nrow(genes)*100,2)
```

| | down | unchanging | up |
|--|------|------------|------|
| | 1.39 | 96.17 | 2.44 |

The key functions here were: `nrow()` and `ncol` `table()` is very useful for getting counts finally `round()`

Plotting the two conditions against each other

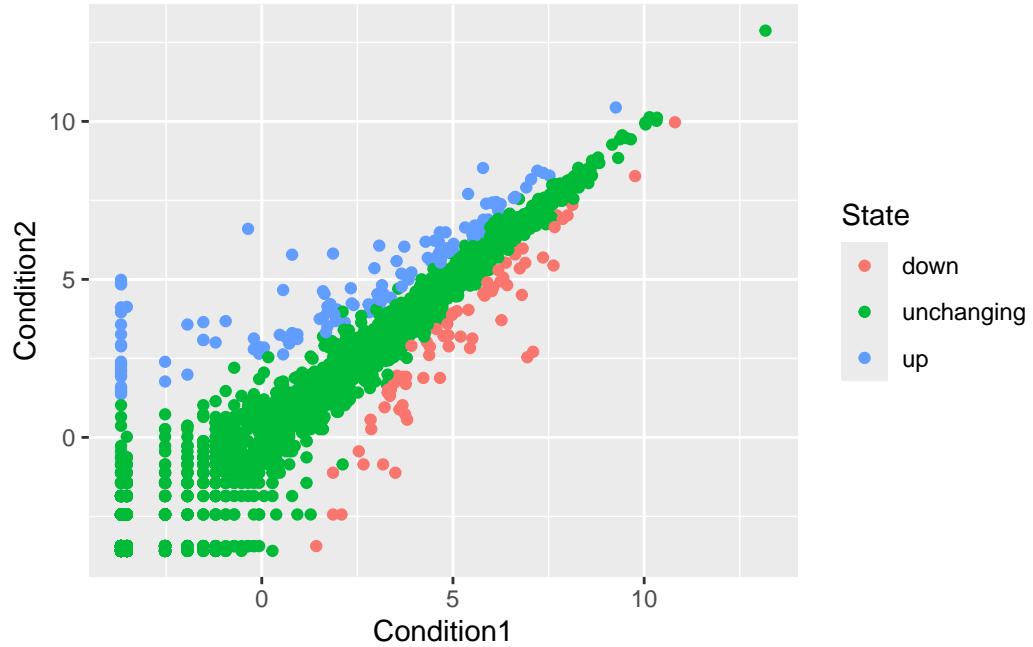
```
ggplot(genes) +  
  aes(x=Condition1,y=Condition2) +  
  geom_point(col="blue", alpha=0.4)
```



Note: The color function goes in the geom not aes section. Alpha in the geom lets it be transparent.

Plotting with colors

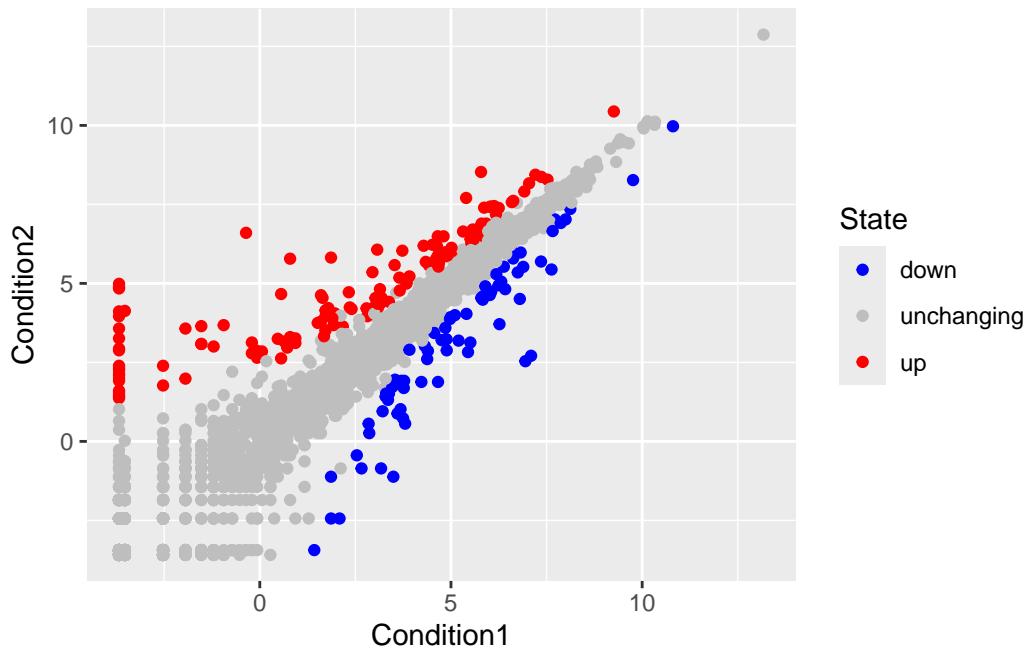
```
p <- ggplot(genes) +  
  aes(x=Condition1, y=Condition2, col=State) +  
  geom_point()  
p
```



Note: col in aes function is showing and specifying the legend, because it is coming from the data.

Specifying colors

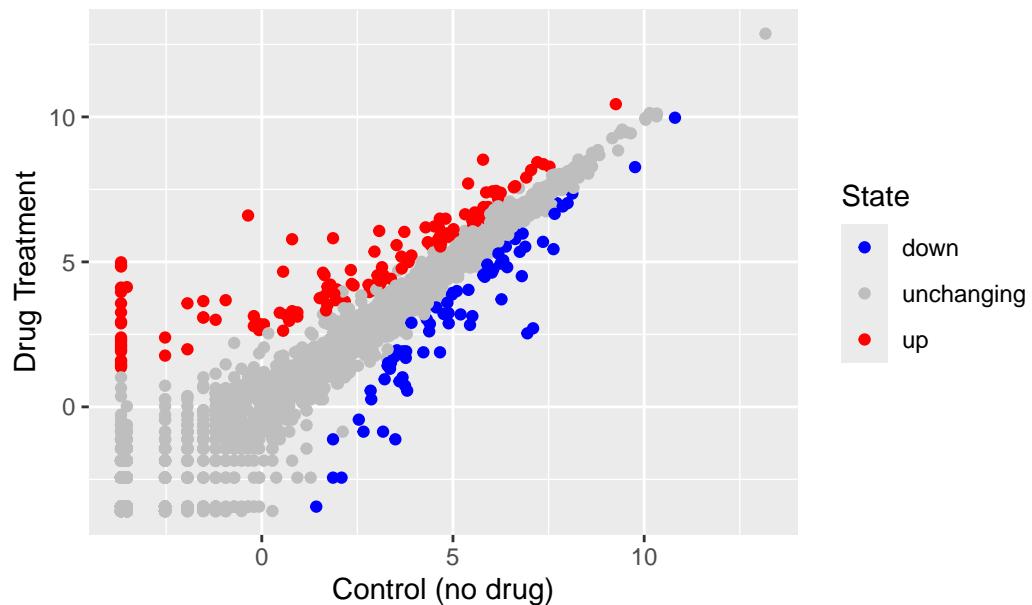
```
p + scale_colour_manual( values=c("blue","gray","red") )
```



Labeling axis

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

Gene Expression Changes Upon Drug Treatment



Section 7: Going Further

Can either install the package (`install.packages("gapminder") library(gapminder)`) or read from github

```
url <- "https://raw.githubusercontent.com/jennybc/gapminder/master/inst/extdata/gapminder.ts"
gapminder <- read.delim(url)
```

To focus on a specific year, need to download the **dplyr** code

```
# install.packages("dplyr")
library(dplyr)
```

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

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

```
filter, lag
```

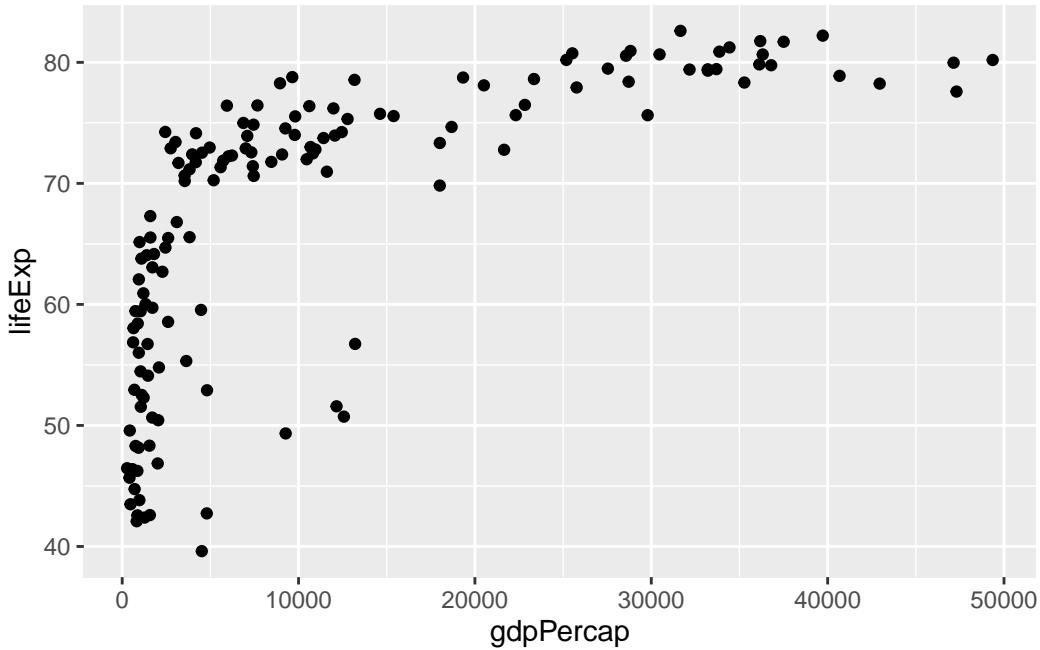
```
The following objects are masked from 'package:base':
```

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

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

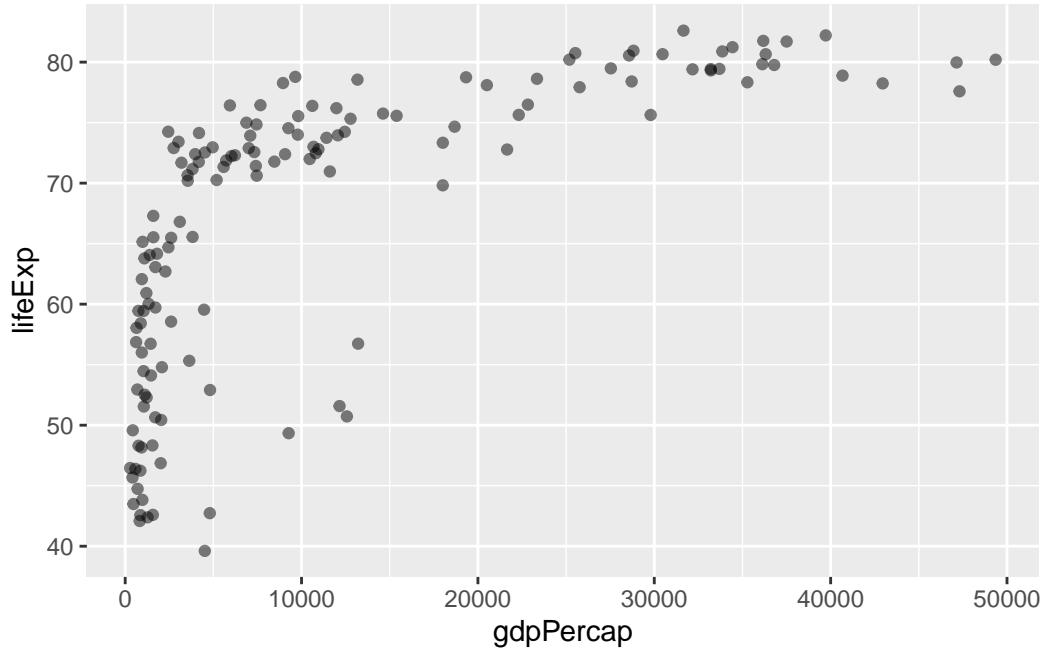
Creating scatterplot for gapminder_2007 data

```
ggplot(gapminder_2007) +  
  aes(x=gdpPercap, y=lifeExp) +  
  geom_point()
```



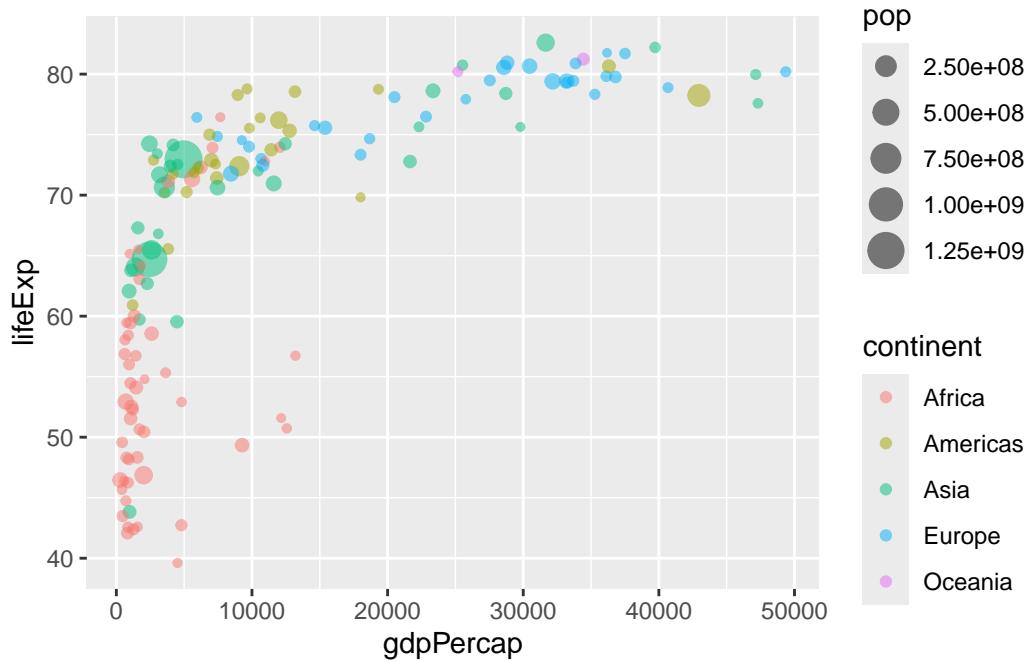
Making points transparent, use `alpha=` in the geom section

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



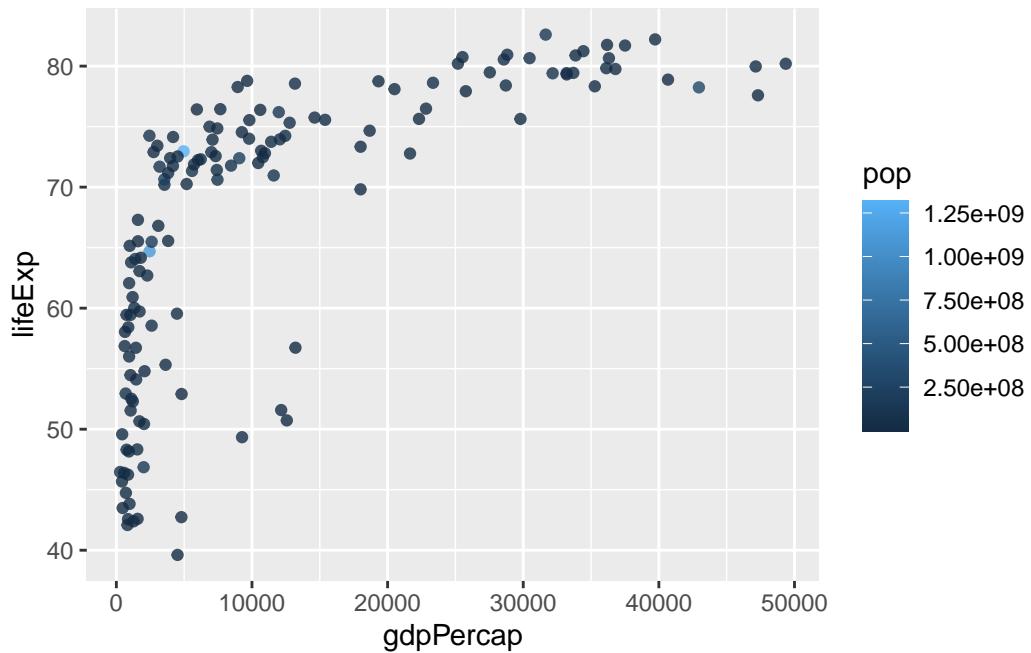
Adding more variables to `sec()` Making color based on continent and size of point on population

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



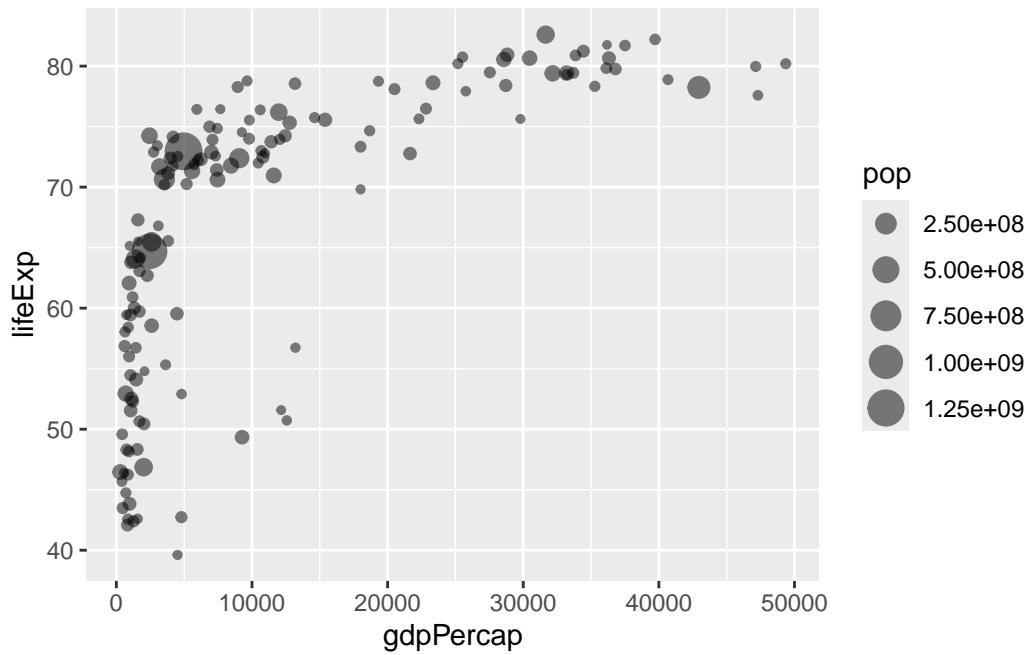
Coloring points based on population

```
ggplot(gapminder_2007) +
  aes(x = gdpPercap, y = lifeExp, color = pop) +
  geom_point(alpha=0.8)
```



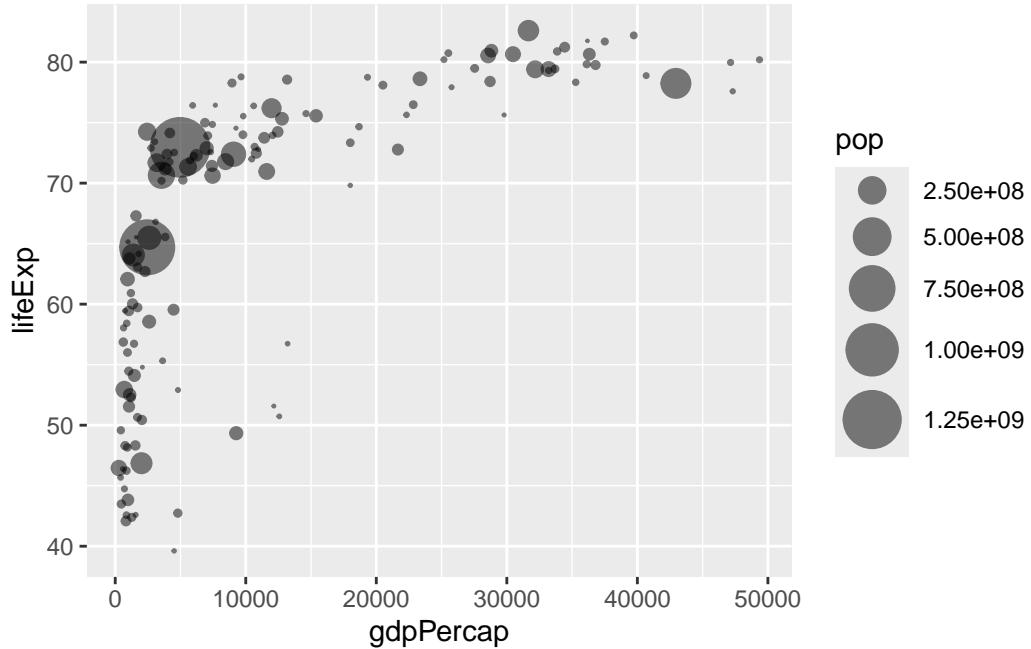
Point size based on population

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



Rescaling the point, use `scale_size_area()`

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



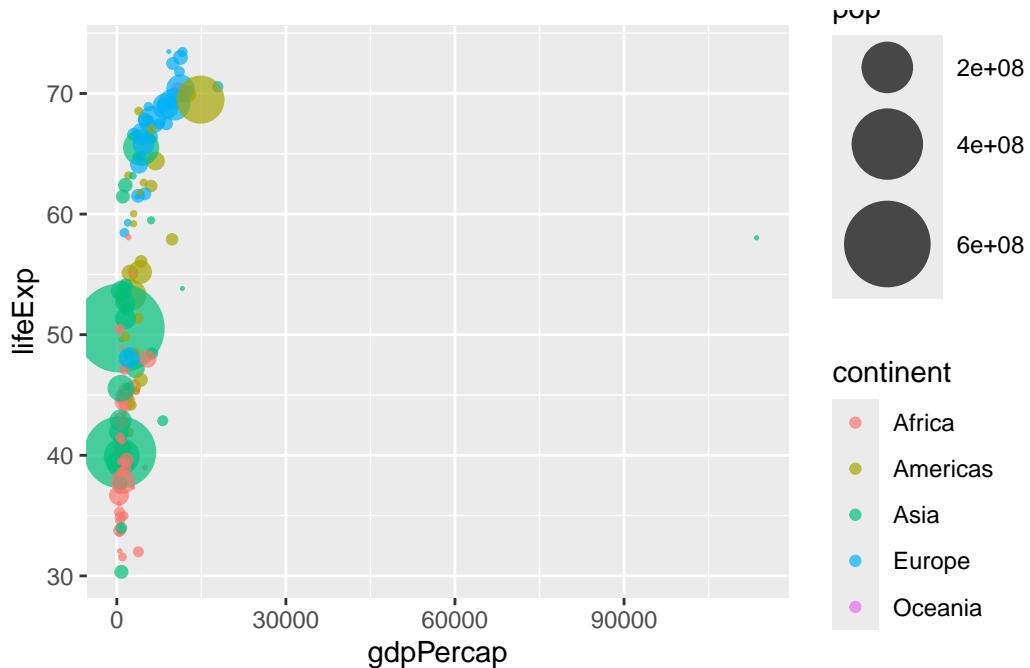
Q Can you adapt the code you have learned thus far to reproduce our gapminder scatter plot for the year 1957? What do you notice about this plot is it easy to compare with the one for 2007?

Plotting for year 1957

```
library(dplyr)

gapminder_1957 <- gapminder %>% filter(year==1957)

ggplot(gapminder_1957) +
  aes(gdpPercap,lifeExp, color=continent, size=pop) +
  geom_point(alpha=0.7) +
  scale_size_area(max_size=15)
```

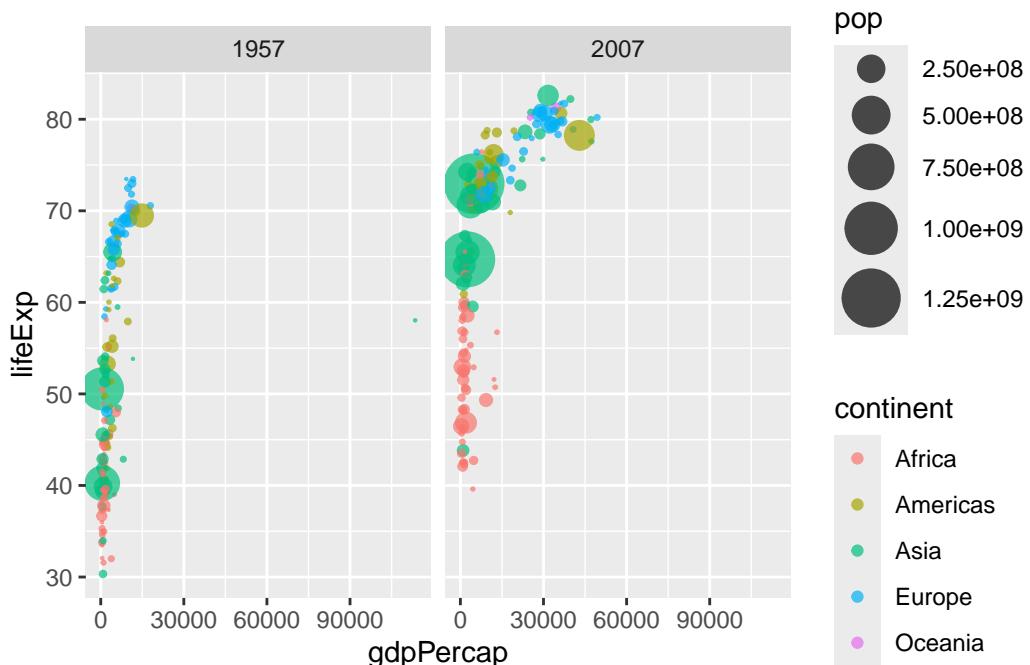


Q. Do the same steps above but include 1957 and 2007 in your input dataset for ggplot(). You should now include the layer facet_wrap(~year) to produce the following plot:

Plotting 1957 and 2007 together

```
gapminder_together <- gapminder %>% filter(year==1957 | year == 2007)

ggplot(gapminder_together) +
  aes(gdpPercap, lifeExp, color=continent, size=pop) +
  geom_point(alpha=0.7) +
  scale_size_area(max_size=10) +
  facet_wrap(~year)
```



```
table(gapminder$year)
```

```
1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 2002 2007
142   142   142   142   142   142   142   142   142   142   142   142
```

```
length(unique(gapminder$year))
```

```
[1] 12
```

```
library(dplyr)
filter(gapminder, country=="United States")
```

| | country | continent | year | lifeExp | pop | gdpPerCap |
|---|---------------|-----------|------|---------|-----------|-----------|
| 1 | United States | Americas | 1952 | 68.440 | 157553000 | 13990.48 |
| 2 | United States | Americas | 1957 | 69.490 | 171984000 | 14847.13 |
| 3 | United States | Americas | 1962 | 70.210 | 186538000 | 16173.15 |
| 4 | United States | Americas | 1967 | 70.760 | 198712000 | 19530.37 |
| 5 | United States | Americas | 1972 | 71.340 | 209896000 | 21806.04 |
| 6 | United States | Americas | 1977 | 73.380 | 220239000 | 24072.63 |

```

7 United States Americas 1982 74.650 232187835 25009.56
8 United States Americas 1987 75.020 242803533 29884.35
9 United States Americas 1992 76.090 256894189 32003.93
10 United States Americas 1997 76.810 272911760 35767.43
11 United States Americas 2002 77.310 287675526 39097.10
12 United States Americas 2007 78.242 301139947 42951.65

```

What is the population of Ireland in the last year we have dataa for?

```
filter(gapminder, country=="Ireland", year ==2007)
```

```

country continent year lifeExp      pop gdpPercap
1 Ireland     Europe 2007 78.885 4109086      40676

```

What countries in data set had pop smaller than Ireland in 2007

```

#limit the dataset to year 2007
gap07 <- filter(gapminder, year ==2007)

#find the `pop` value of Ireland
ire_pop <- filter(gap07, country == "Ireland")["pop"]
ire_pop

```

```

pop
1 4109086

```

```
#Extract all rows with `pop` less than Ireland's
filter(gap07, pop<4109086)
```

```

country continent year lifeExp      pop gdpPercap
1       Albania    Europe 2007 76.423 3600523 5937.0295
2       Bahrain     Asia 2007 75.635 708573 29796.0483
3     Botswana    Africa 2007 50.728 1639131 12569.8518
4     Comoros     Africa 2007 65.152 710960  986.1479
5   Congo, Rep.    Africa 2007 55.322 3800610 3632.5578
6     Djibouti    Africa 2007 54.791 496374 2082.4816
7 Equatorial Guinea    Africa 2007 51.579 551201 12154.0897
8       Gabon     Africa 2007 56.735 1454867 13206.4845
9      Gambia     Africa 2007 59.448 1688359  752.7497
10  Guinea-Bissau    Africa 2007 46.388 1472041  579.2317

```

| | | | | | | |
|----|-----------------------|----------|------|--------|---------|------------|
| 11 | Iceland | Europe | 2007 | 81.757 | 301931 | 36180.7892 |
| 12 | Jamaica | Americas | 2007 | 72.567 | 2780132 | 7320.8803 |
| 13 | Kuwait | Asia | 2007 | 77.588 | 2505559 | 47306.9898 |
| 14 | Lebanon | Asia | 2007 | 71.993 | 3921278 | 10461.0587 |
| 15 | Lesotho | Africa | 2007 | 42.592 | 2012649 | 1569.3314 |
| 16 | Liberia | Africa | 2007 | 45.678 | 3193942 | 414.5073 |
| 17 | Mauritania | Africa | 2007 | 64.164 | 3270065 | 1803.1515 |
| 18 | Mauritius | Africa | 2007 | 72.801 | 1250882 | 10956.9911 |
| 19 | Mongolia | Asia | 2007 | 66.803 | 2874127 | 3095.7723 |
| 20 | Montenegro | Europe | 2007 | 74.543 | 684736 | 9253.8961 |
| 21 | Namibia | Africa | 2007 | 52.906 | 2055080 | 4811.0604 |
| 22 | Oman | Asia | 2007 | 75.640 | 3204897 | 22316.1929 |
| 23 | Panama | Americas | 2007 | 75.537 | 3242173 | 9809.1856 |
| 24 | Puerto Rico | Americas | 2007 | 78.746 | 3942491 | 19328.7090 |
| 25 | Reunion | Africa | 2007 | 76.442 | 798094 | 7670.1226 |
| 26 | Sao Tome and Principe | Africa | 2007 | 65.528 | 199579 | 1598.4351 |
| 27 | Slovenia | Europe | 2007 | 77.926 | 2009245 | 25768.2576 |
| 28 | Swaziland | Africa | 2007 | 39.613 | 1133066 | 4513.4806 |
| 29 | Trinidad and Tobago | Americas | 2007 | 69.819 | 1056608 | 18008.5092 |
| 30 | Uruguay | Americas | 2007 | 76.384 | 3447496 | 10611.4630 |
| 31 | West Bank and Gaza | Asia | 2007 | 73.422 | 4018332 | 3025.3498 |

Section 8: Bar Charts

Viewing five largest countries by population in 2007

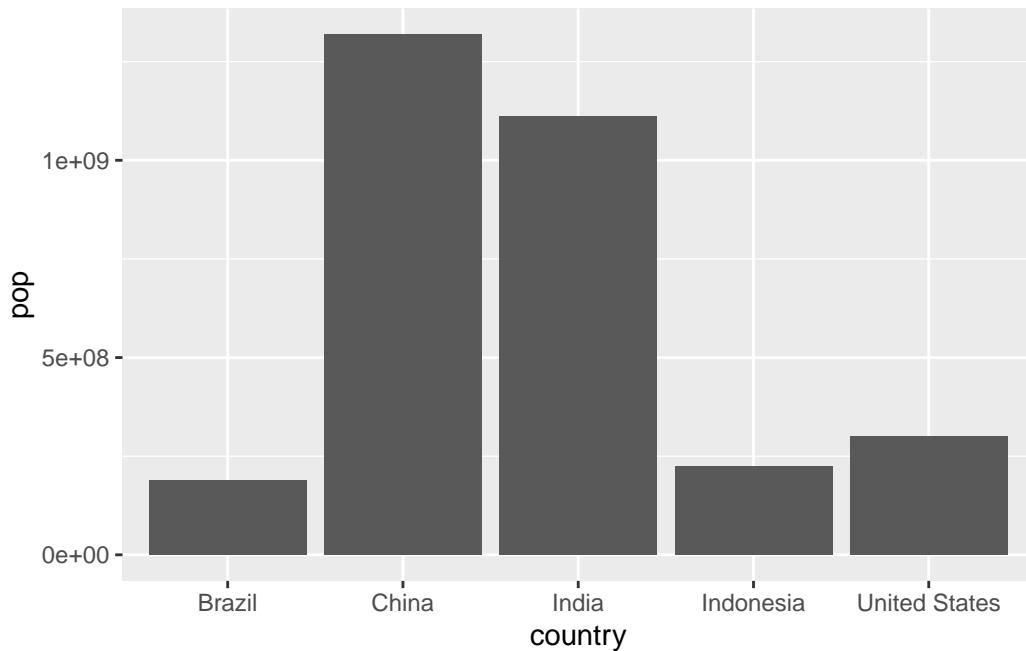
```
gapminder_top5 <- gapminder %>%
  filter(year==2007) %>%
  arrange(desc(pop)) %>%
  top_n(5, pop)

gapminder_top5
```

| | country | continent | year | lifeExp | pop | gdpPercap |
|---|---------------|-----------|------|---------|------------|-----------|
| 1 | China | Asia | 2007 | 72.961 | 1318683096 | 4959.115 |
| 2 | India | Asia | 2007 | 64.698 | 1110396331 | 2452.210 |
| 3 | United States | Americas | 2007 | 78.242 | 301139947 | 42951.653 |
| 4 | Indonesia | Asia | 2007 | 70.650 | 223547000 | 3540.652 |
| 5 | Brazil | Americas | 2007 | 72.390 | 190010647 | 9065.801 |

Simple bar chart: use geom_col

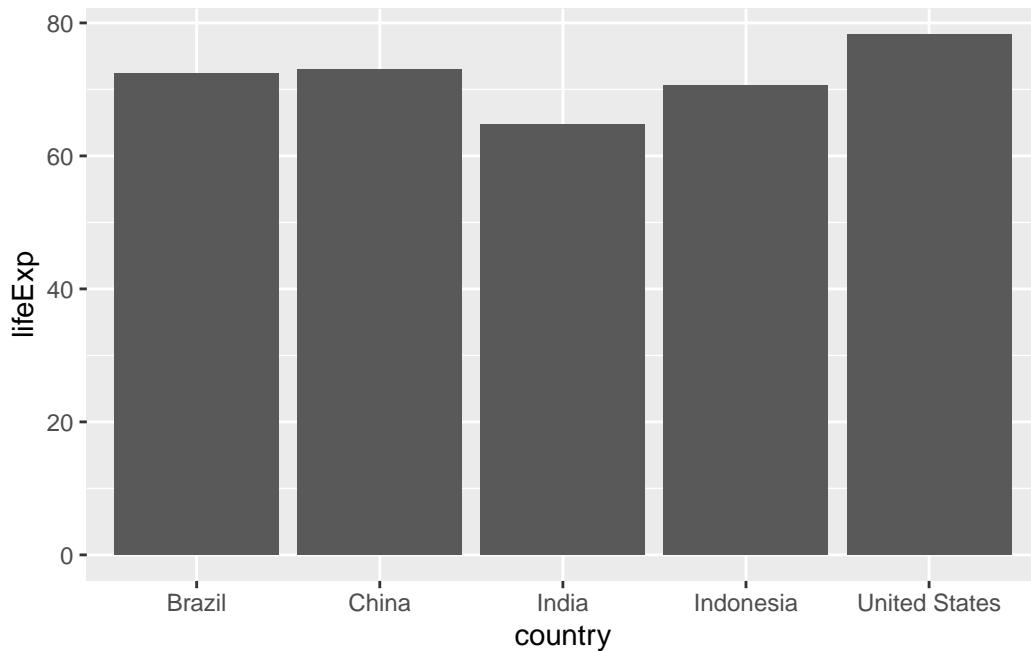
```
ggplot(gapminder_top5) +  
  geom_col() +  
  aes(x = country, y = pop)
```



Q Create a bar chart showing the life expectancy of the five biggest countries by population in 2007.

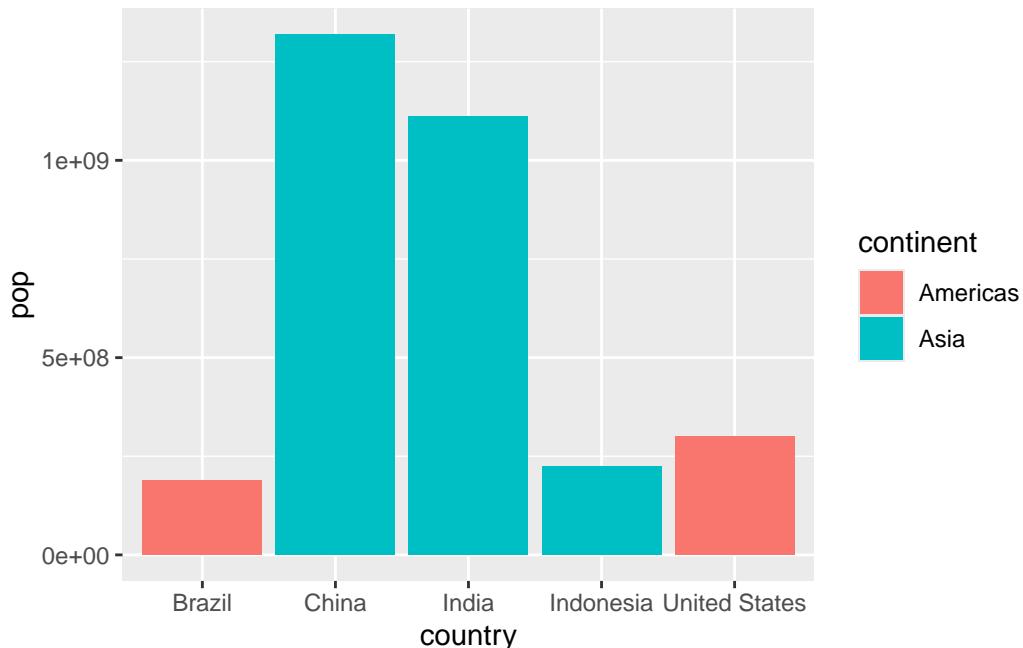
Simple bar chart of life expectancy

```
ggplot(gapminder_top5) +  
  geom_col() +  
  aes(x = country, y = lifeExp)
```



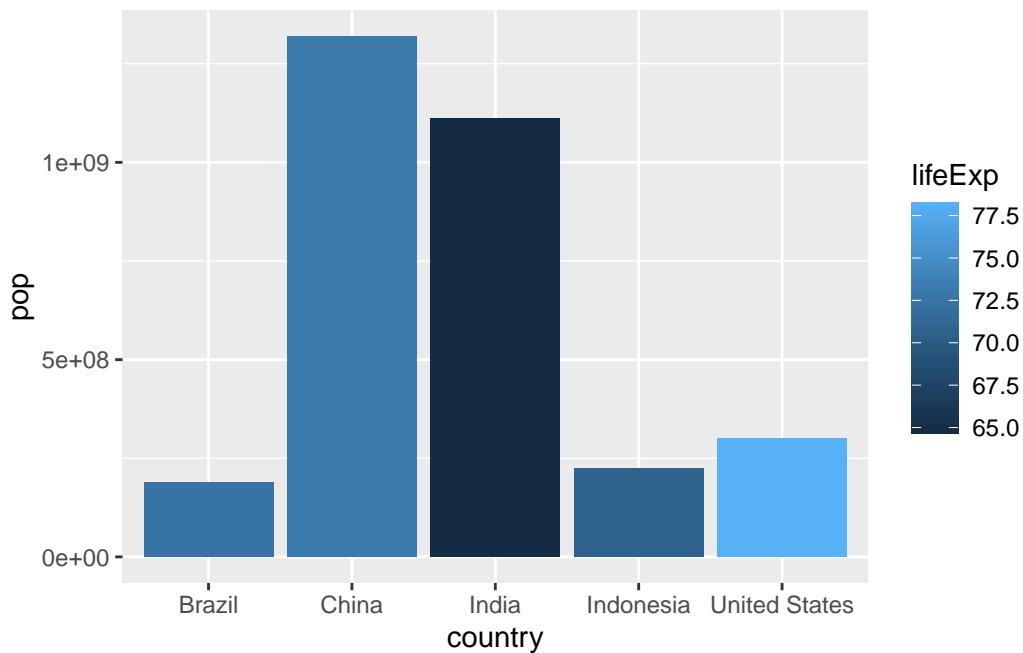
Filling bars with color corresponding to continent (categorical variable) using `fill` aesthetic

```
ggplot(gapminder_top5) +  
  geom_col() +  
  aes(x = country, y = pop, fill = continent)
```



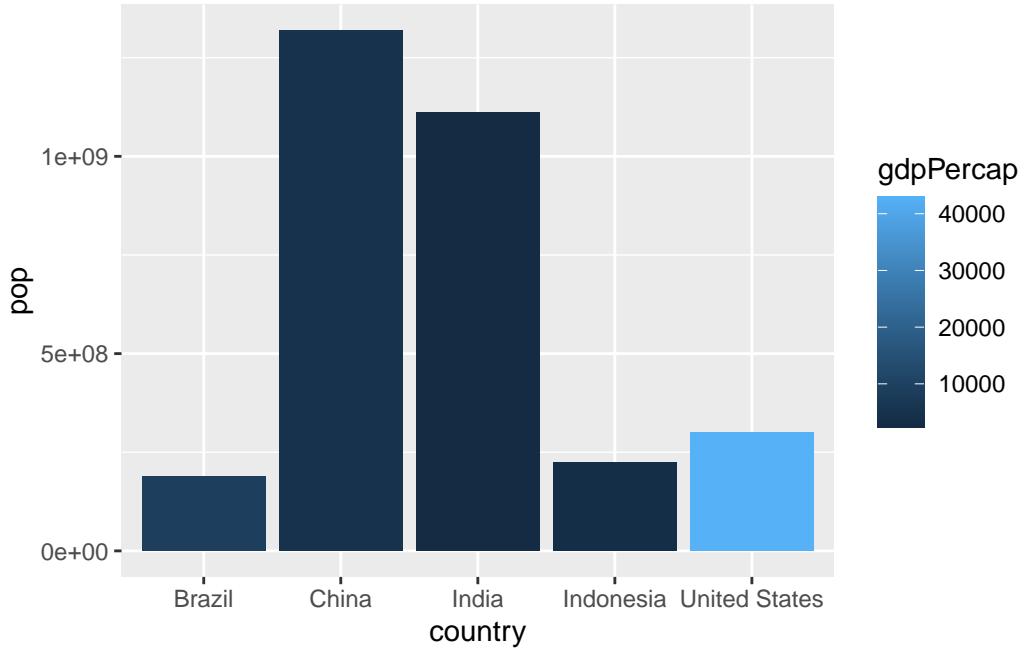
Filling bars with color corresponding to life exp (numerical variable) using `fill` aesthetic

```
ggplot(gapminder_top5) +  
  geom_col() +  
  aes(x = country, y = pop, fill = lifeExp)
```



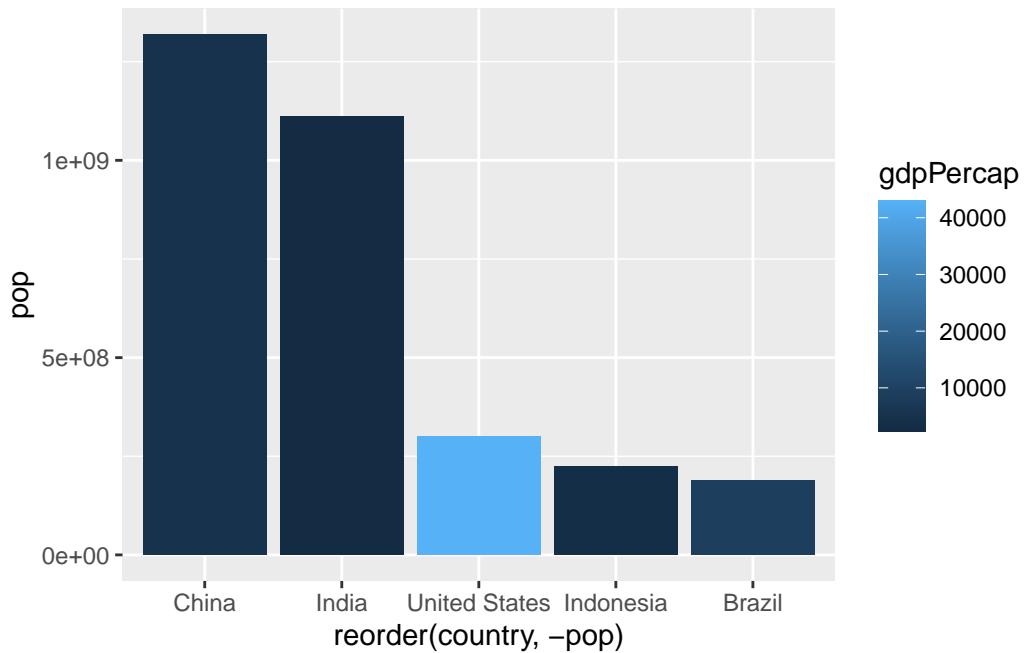
Q. Plot population size by country. Create a bar chart showing the population (in millions) of the five biggest countries by population in 2007.

```
ggplot(gapminder_top5)+  
  geom_col()  
  aes(country, pop, fill=gdpPercap)
```



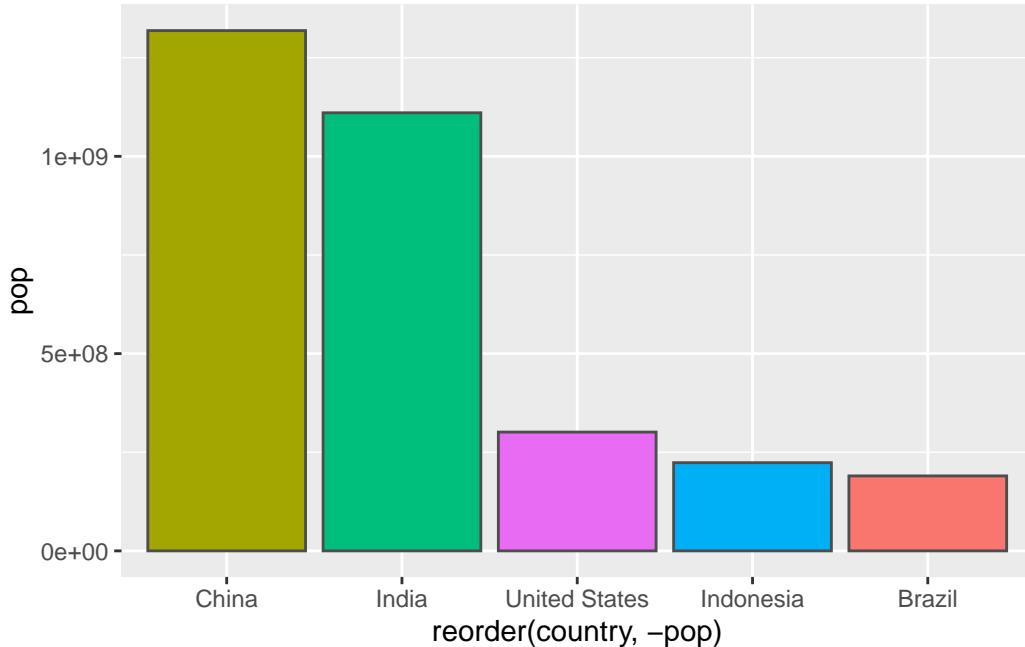
Changing the order of the bars use `reorder()` aesthetic

```
ggplot(gapminder_top5) +  
  aes(x=reorder(country, -pop), y=pop, fill=gdpPerCap) +  
  geom_col()
```



Filling by country

```
ggplot(gapminder_top5) +  
  aes(x=reorder(country, -pop), y=pop, fill=country) +  
  geom_col(col="gray30") +  
  guides(fill="none")
```

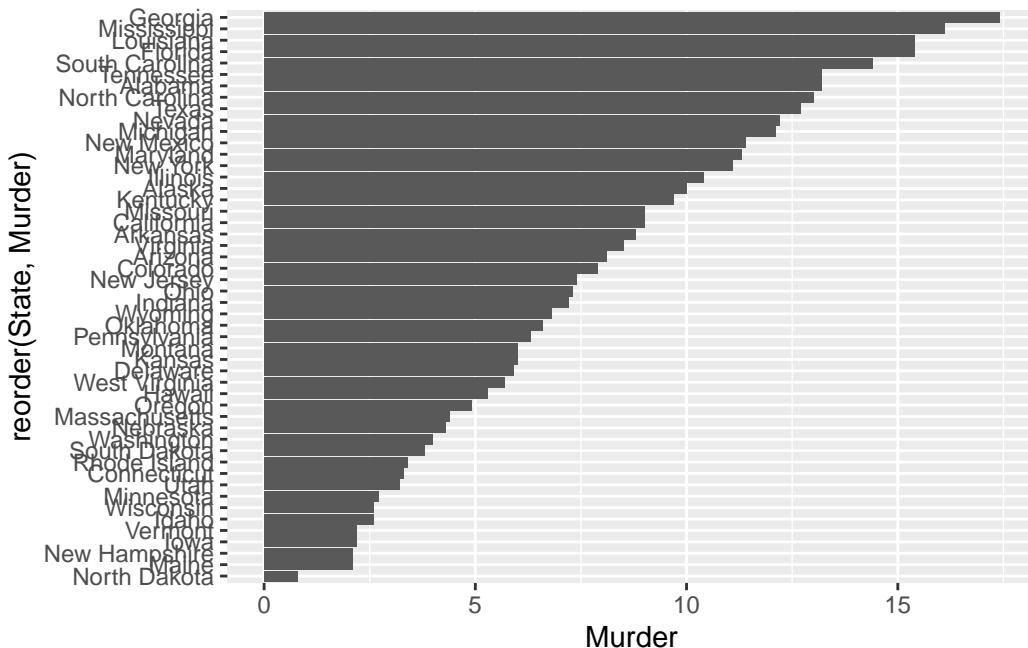


Flipping Bar Charts Use `coord_flip()` function to flip

```
head(USArrests)
```

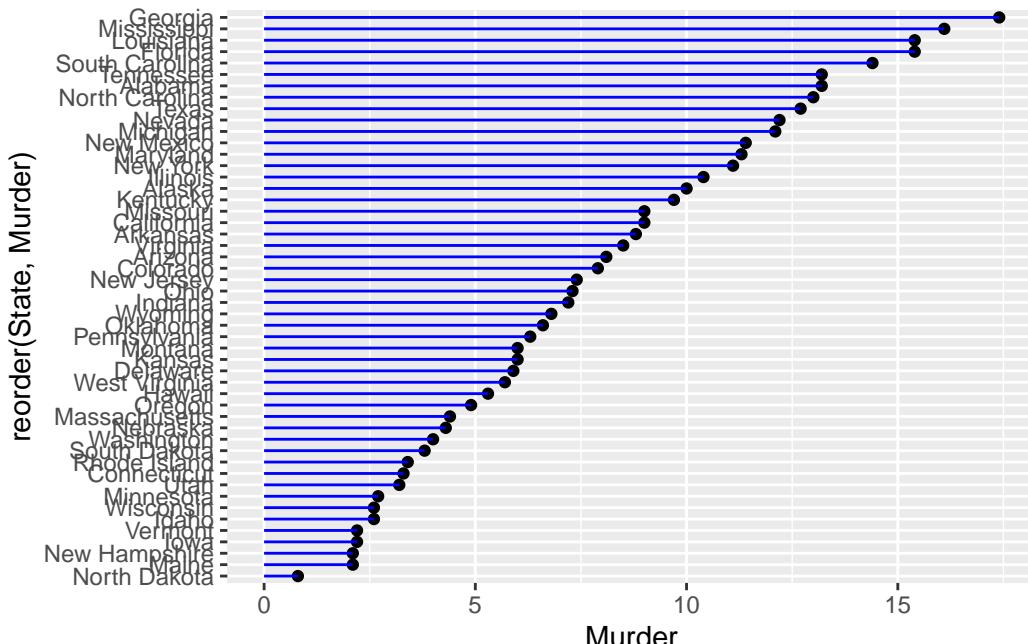
| | Murder | Assault | UrbanPop | Rape |
|------------|--------|---------|----------|------|
| Alabama | 13.2 | 236 | 58 | 21.2 |
| Alaska | 10.0 | 263 | 48 | 44.5 |
| Arizona | 8.1 | 294 | 80 | 31.0 |
| Arkansas | 8.8 | 190 | 50 | 19.5 |
| California | 9.0 | 276 | 91 | 40.6 |
| Colorado | 7.9 | 204 | 78 | 38.7 |

```
USArrests$State <- rownames(USArrests)
ggplot(USArrests) +
  aes(x=reorder(State,Murder) , y=Murder) +
  geom_col() +
  coord_flip()
```



Customizing more with both `geom_point()` and `geom_segment()`

```
ggplot(USArrests) +
  aes(x=reorder(State,Murder), y=Murder) +
  geom_point() +
  geom_segment(aes(x=State,
                   xend=State,
                   y=0,
                   yend=Murder), color="blue") +
  coord_flip()
```



```
##9: extensions: Animation Install ganimate and gifski packages
```

```
#install.packages("gifski")
#install.packages("ganimate")
```

```
library(gapminder)
```

```
Attaching package: 'gapminder'
```

```
The following object is masked _by_ '.GlobalEnv':
```

```
gapminder
```

```
library(gganimate)

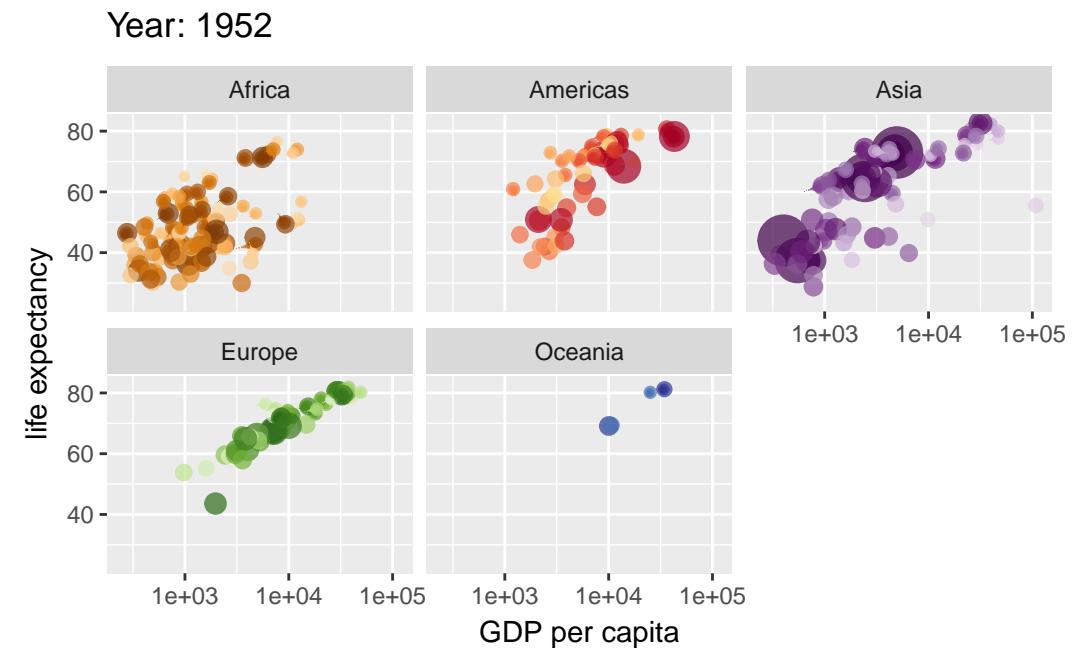
# Setup nice regular ggplot of the gapminder data
ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, colour = country)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors) +
  scale_size(range = c(2, 12)) +
  scale_x_log10()
```

```

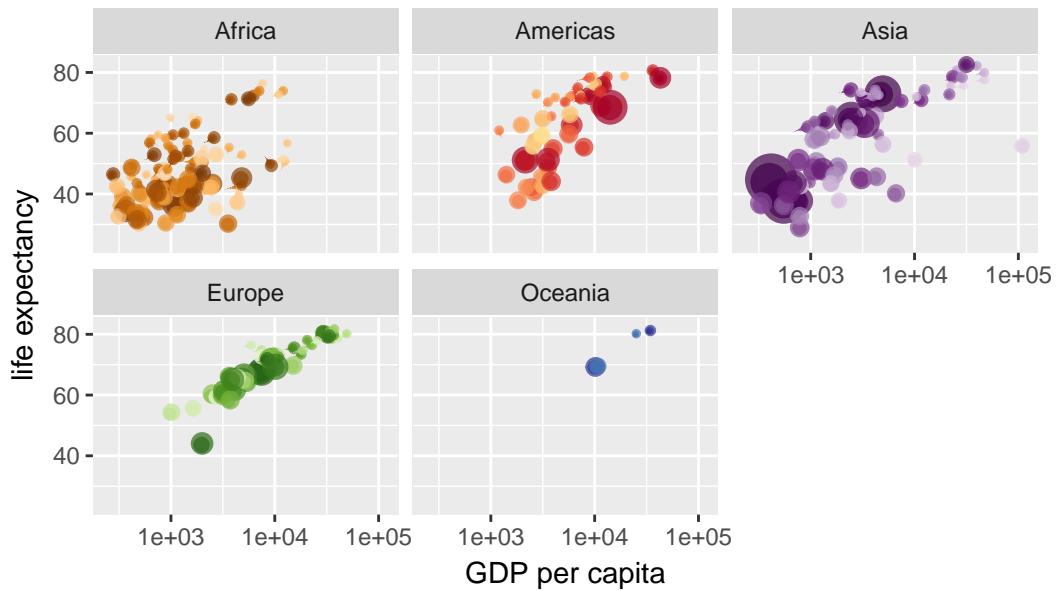
# Facet by continent
facet_wrap(~continent) +
# Here comes the gganimate specific bits
labs(title = 'Year: {frame_time}', x = 'GDP per capita', y = 'life expectancy') +
transition_time(year) +
shadow_wake(wake_length = 0.1, alpha = FALSE)

```

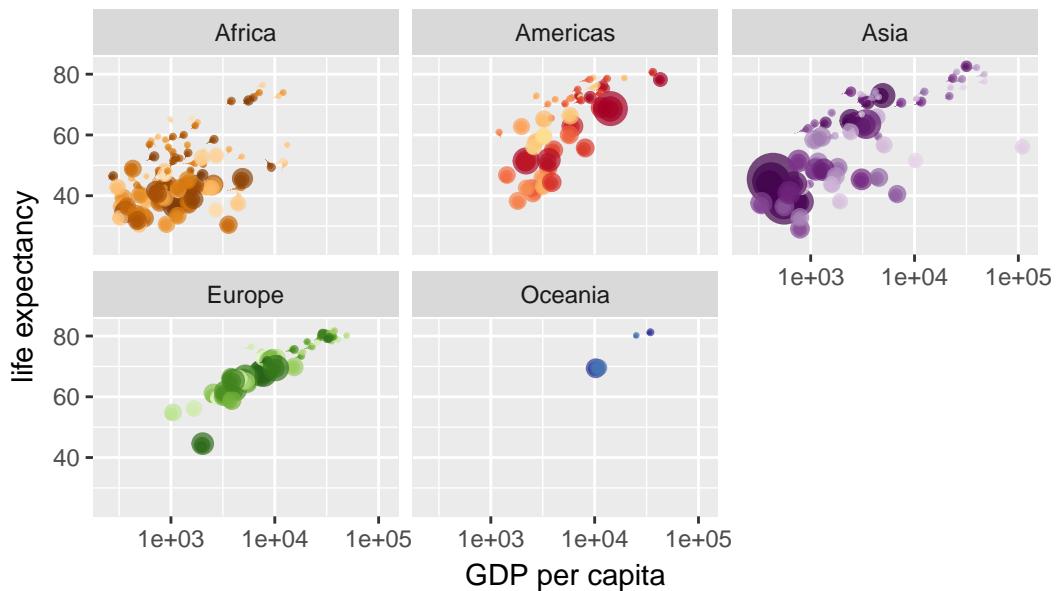
Warning in formals(fun): argument is not a function



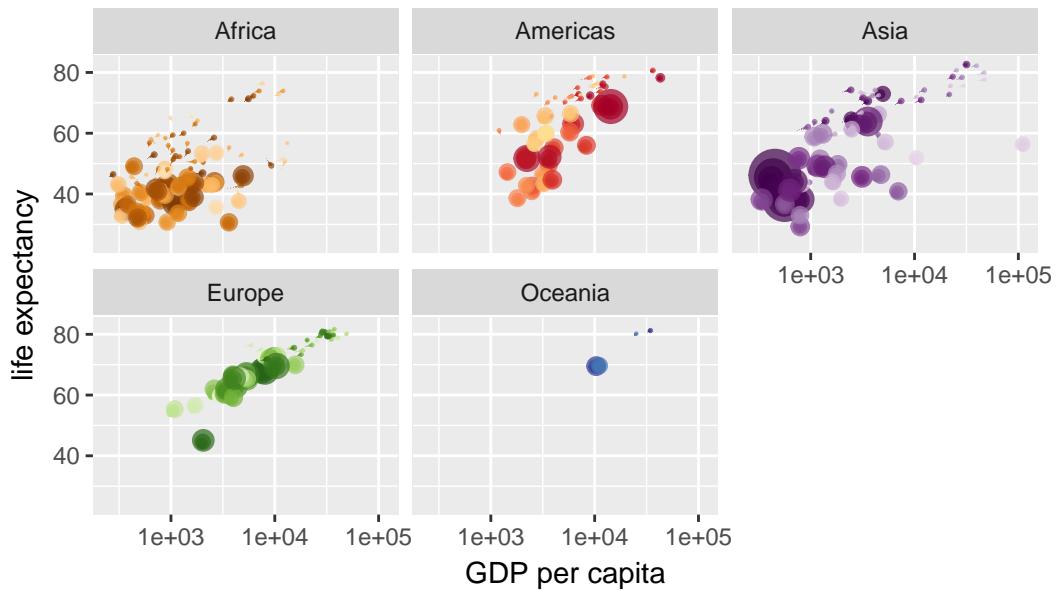
Year: 1953



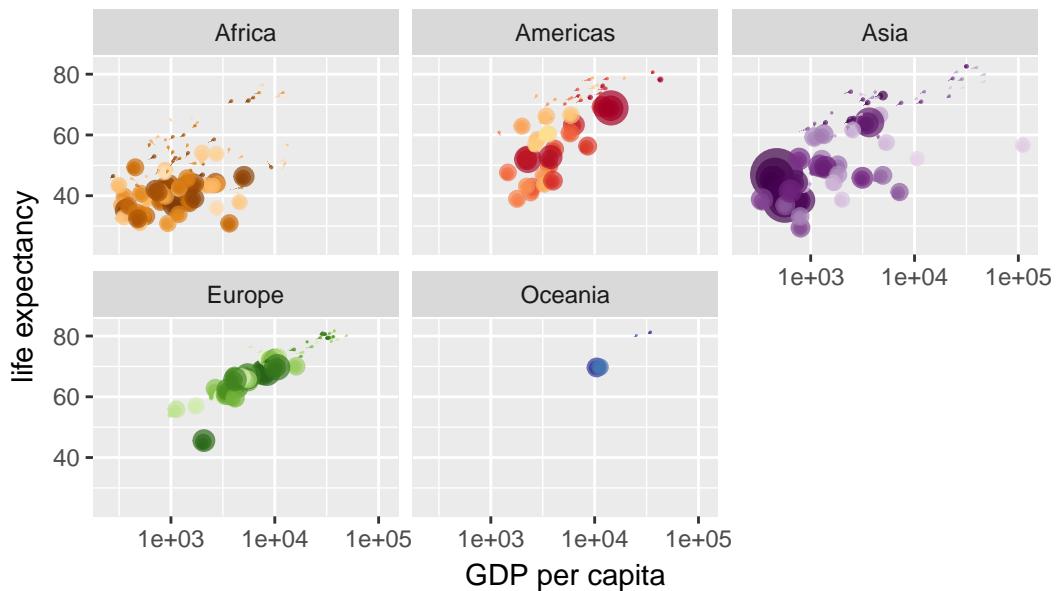
Year: 1953



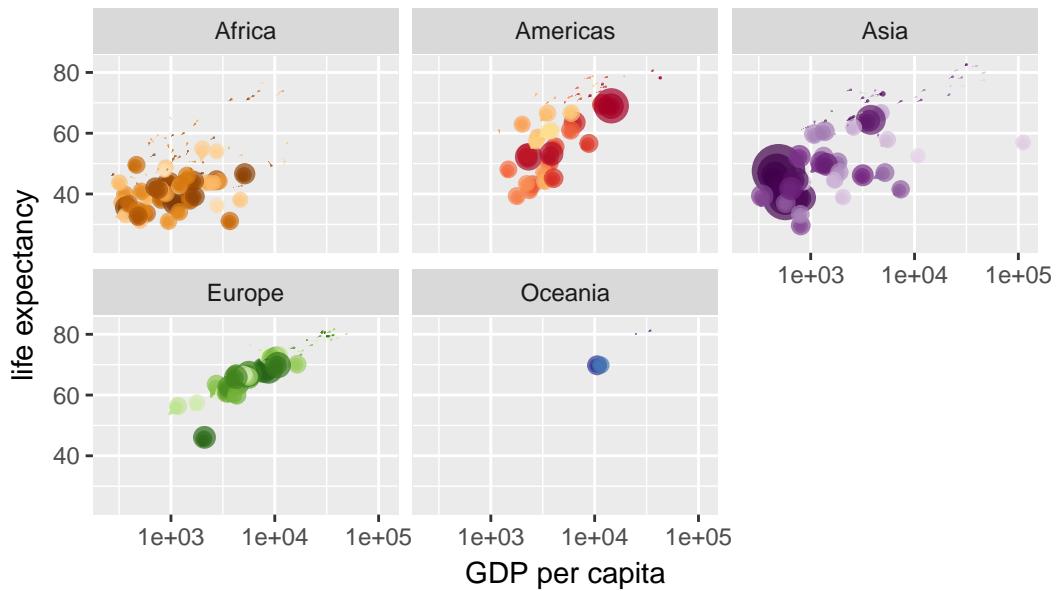
Year: 1954



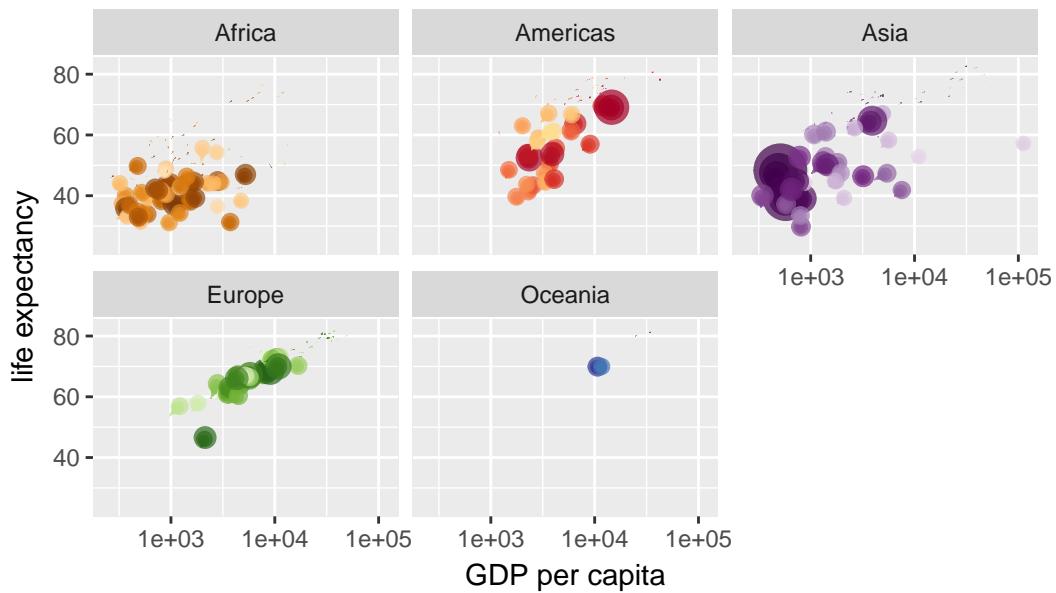
Year: 1954



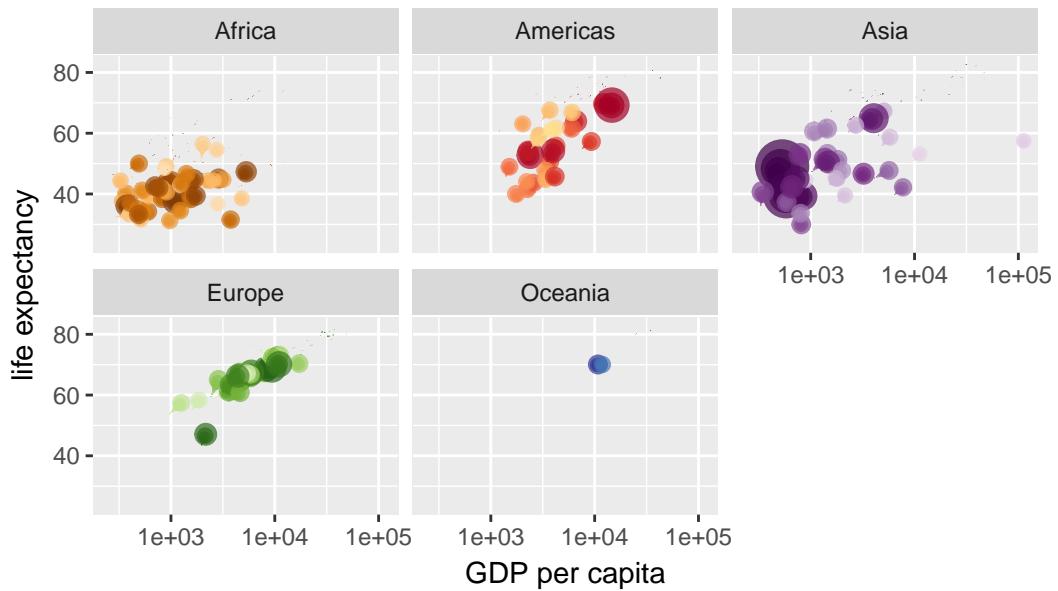
Year: 1955



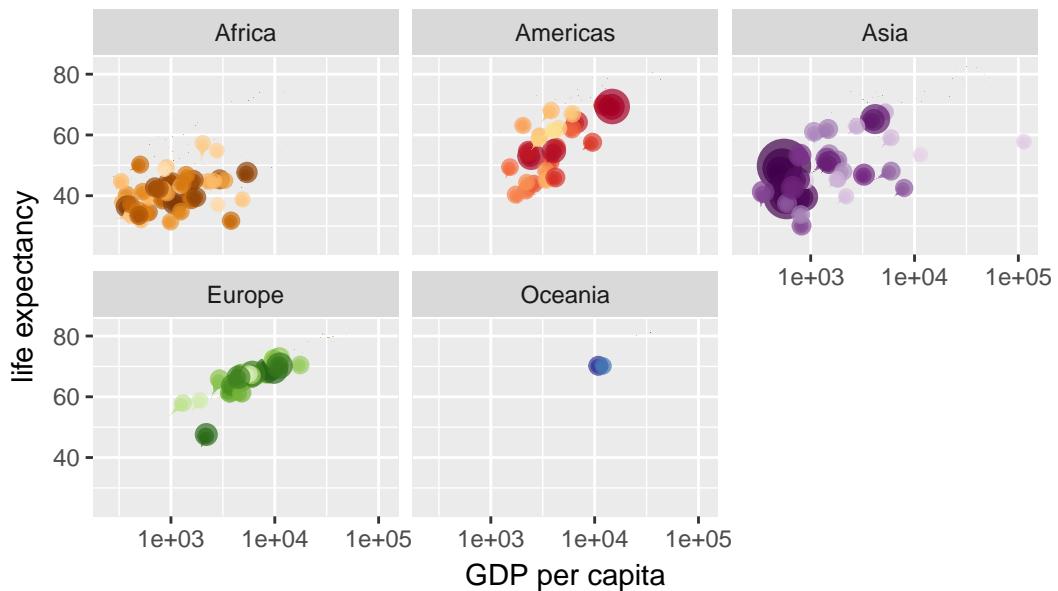
Year: 1955



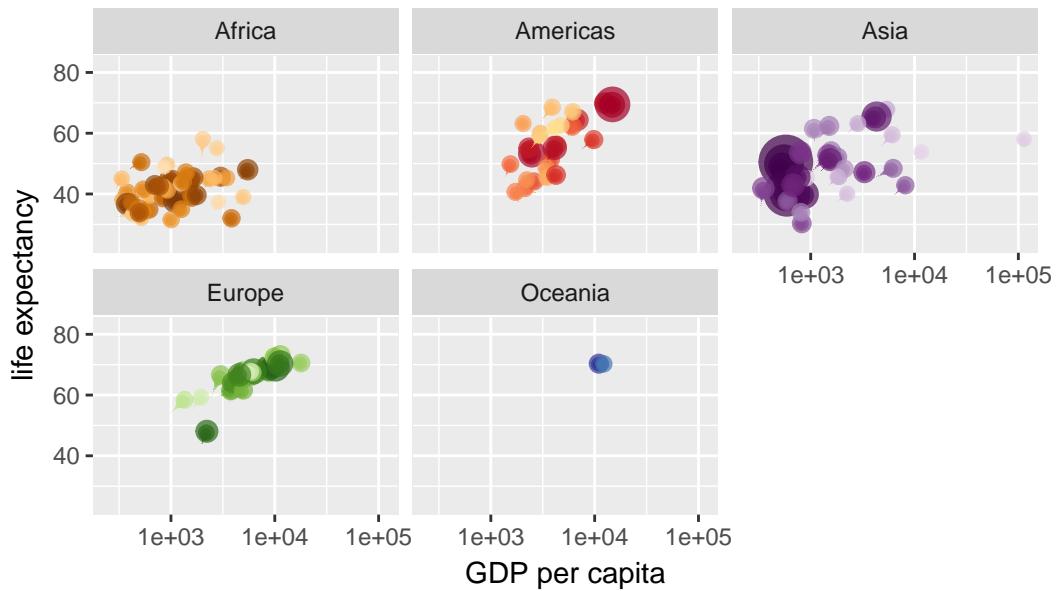
Year: 1956



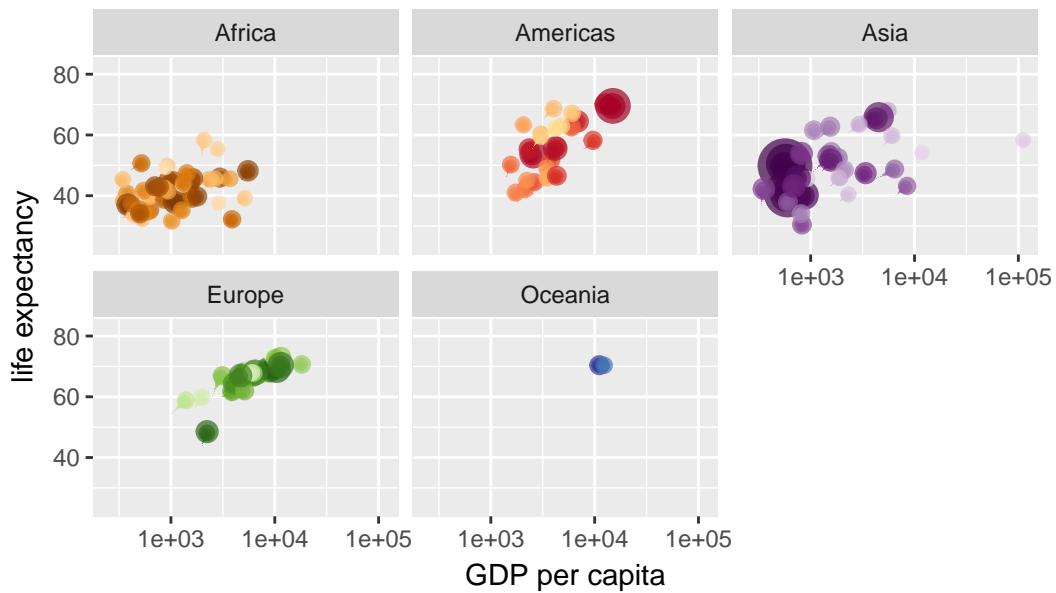
Year: 1956



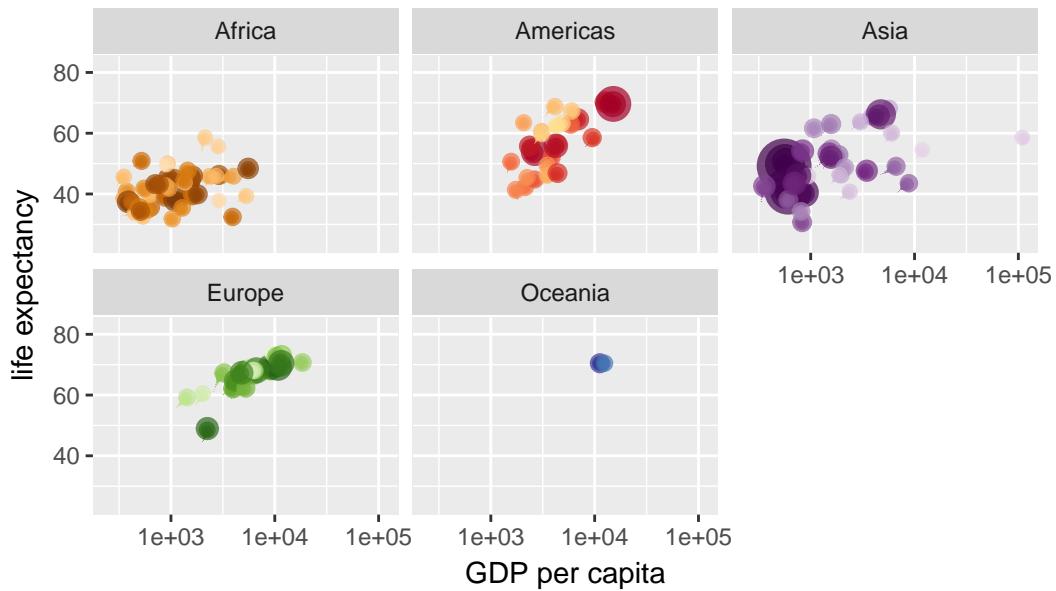
Year: 1957



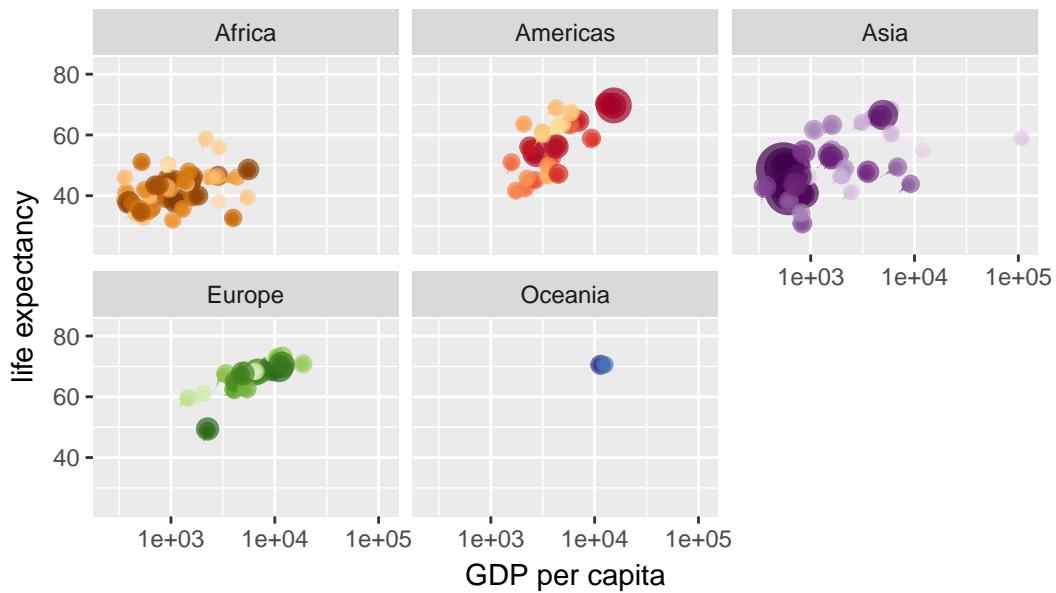
Year: 1958



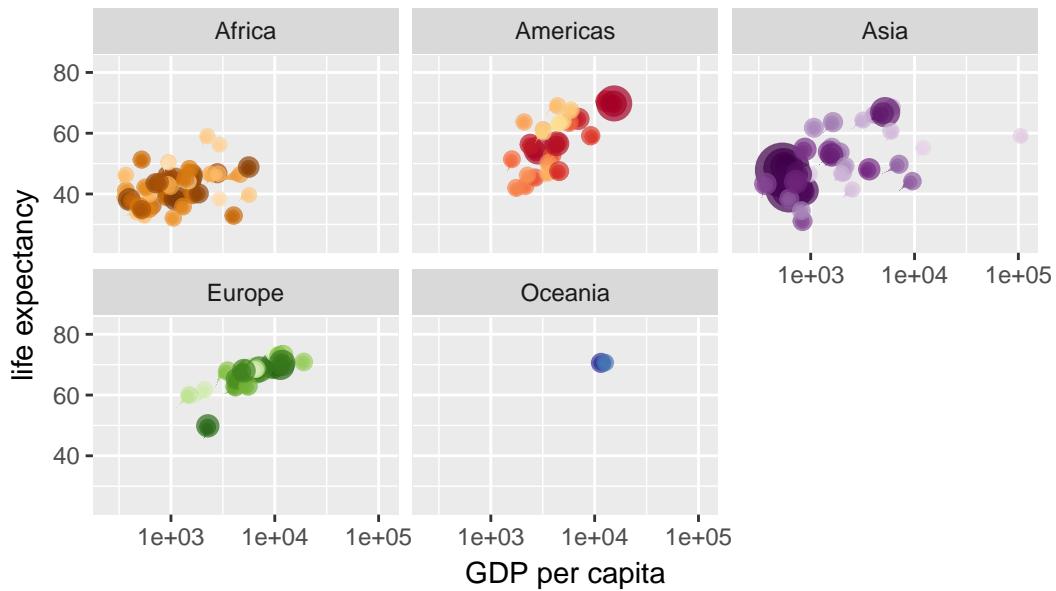
Year: 1958



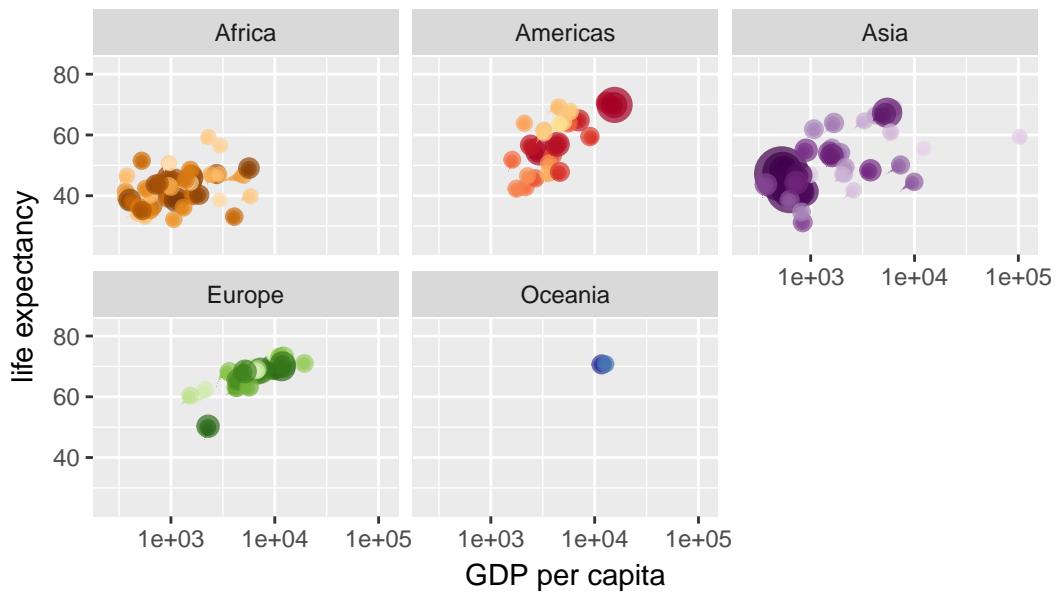
Year: 1959



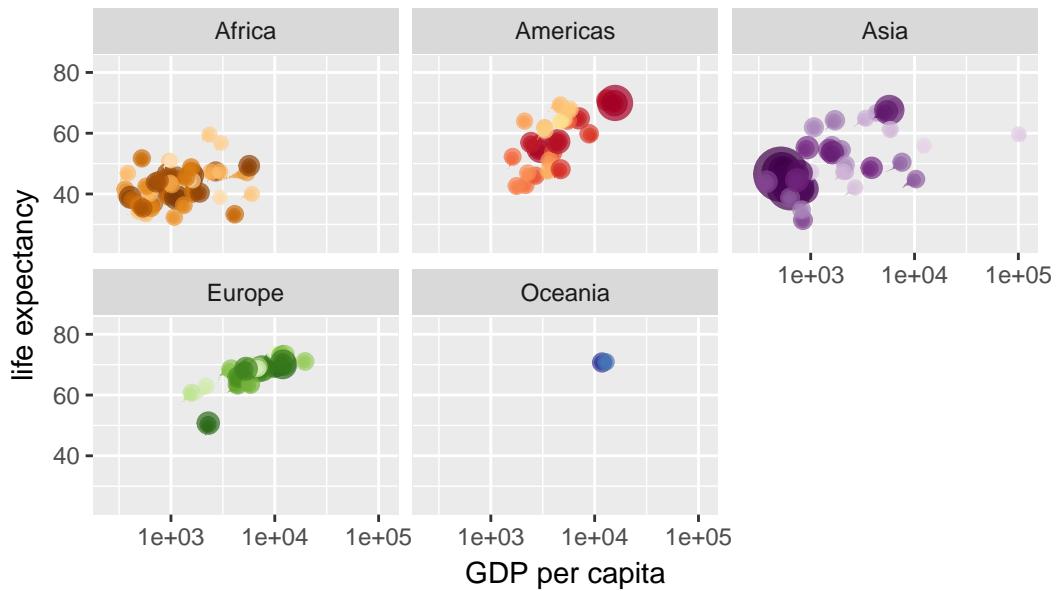
Year: 1959



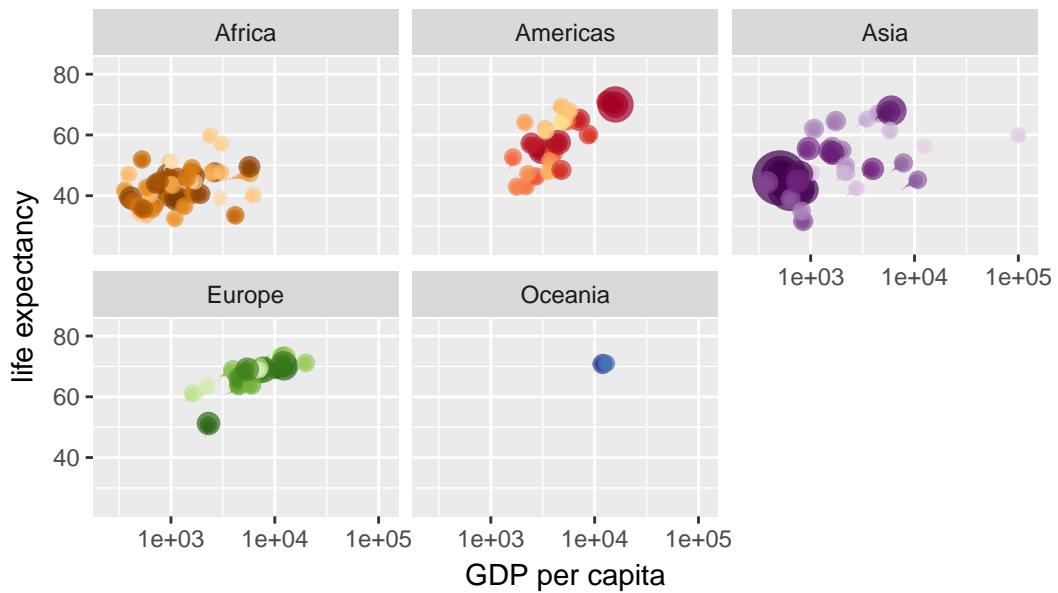
Year: 1960



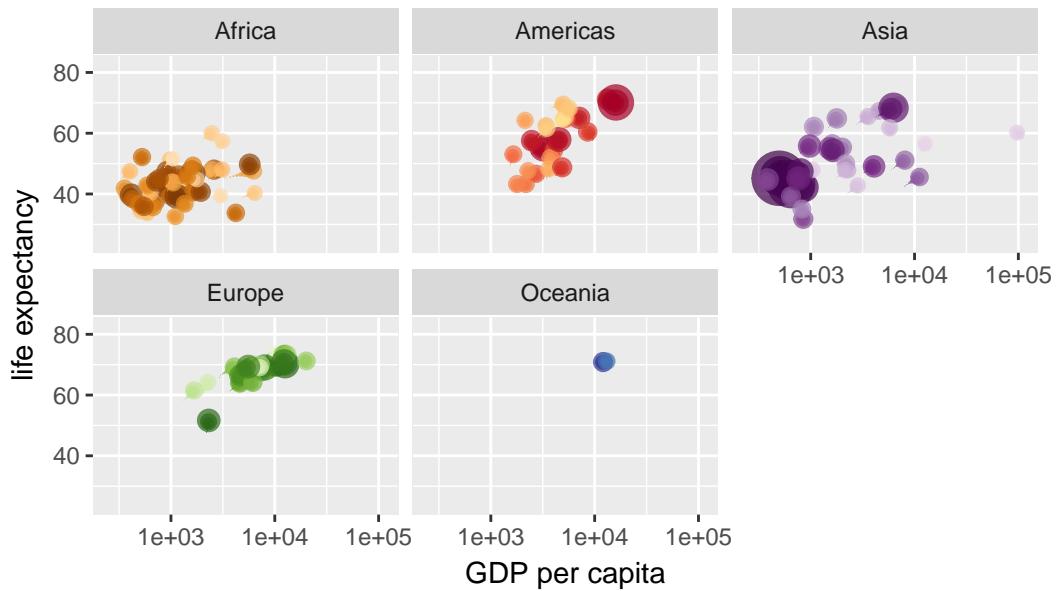
Year: 1960



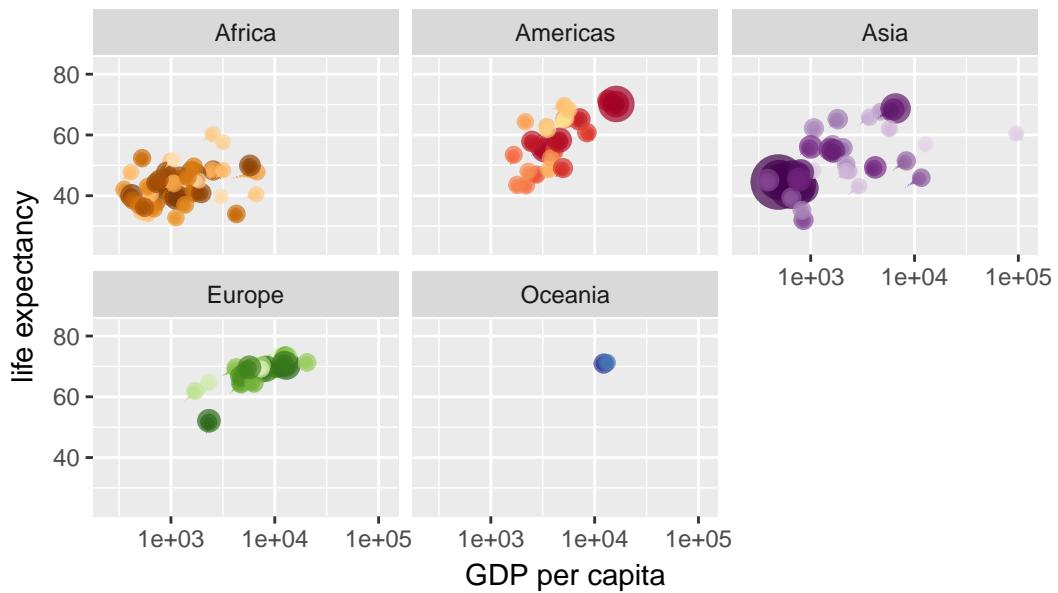
Year: 1961



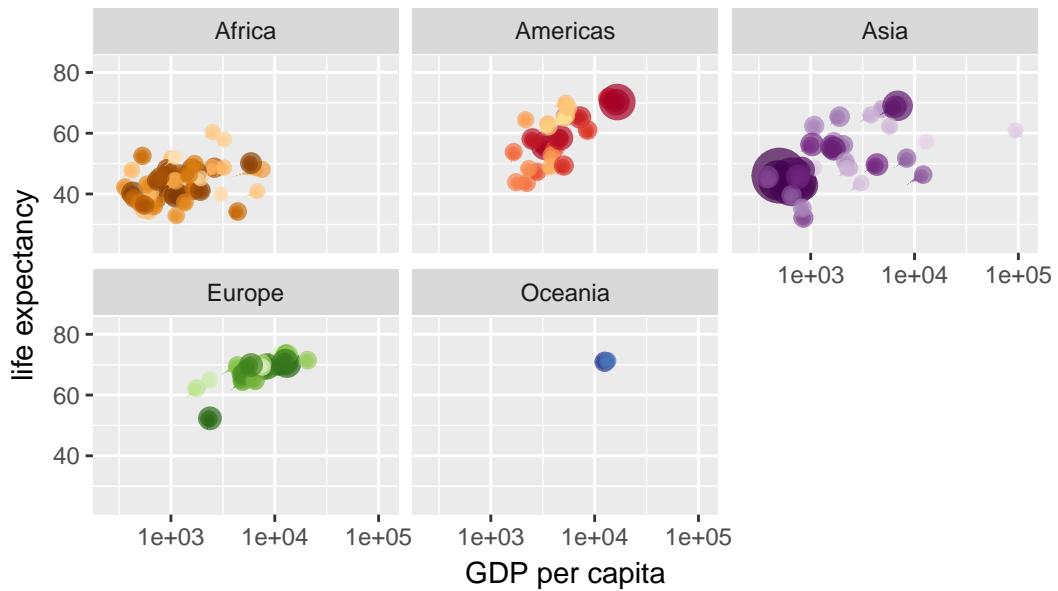
Year: 1961



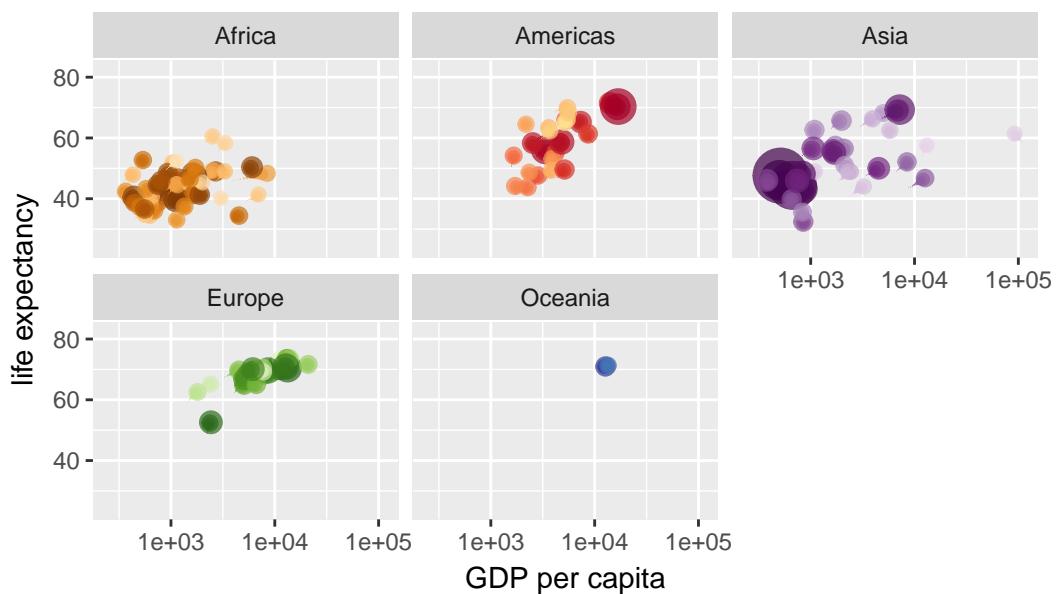
Year: 1962



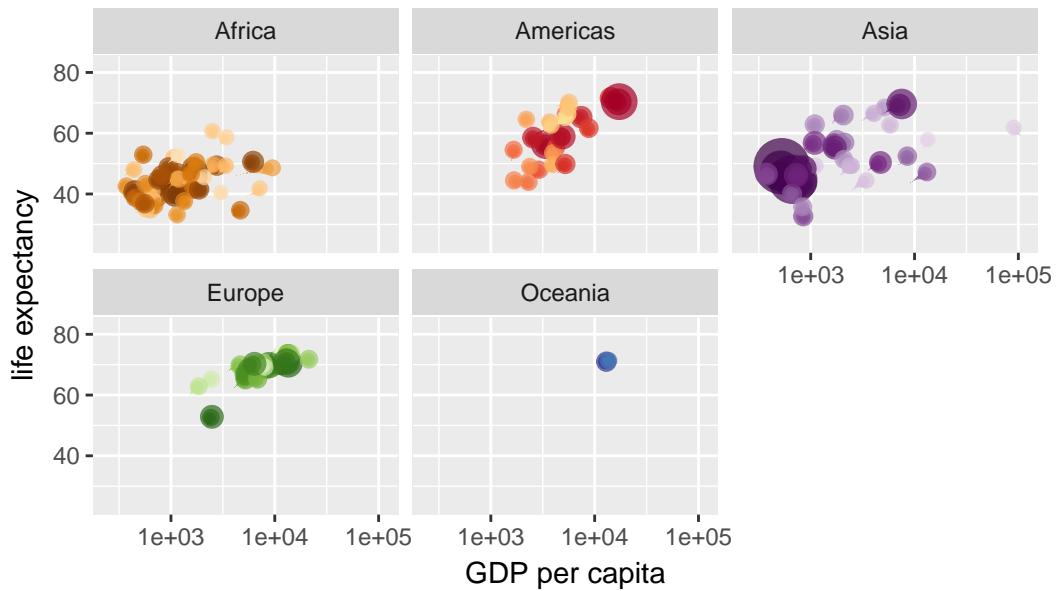
Year: 1963



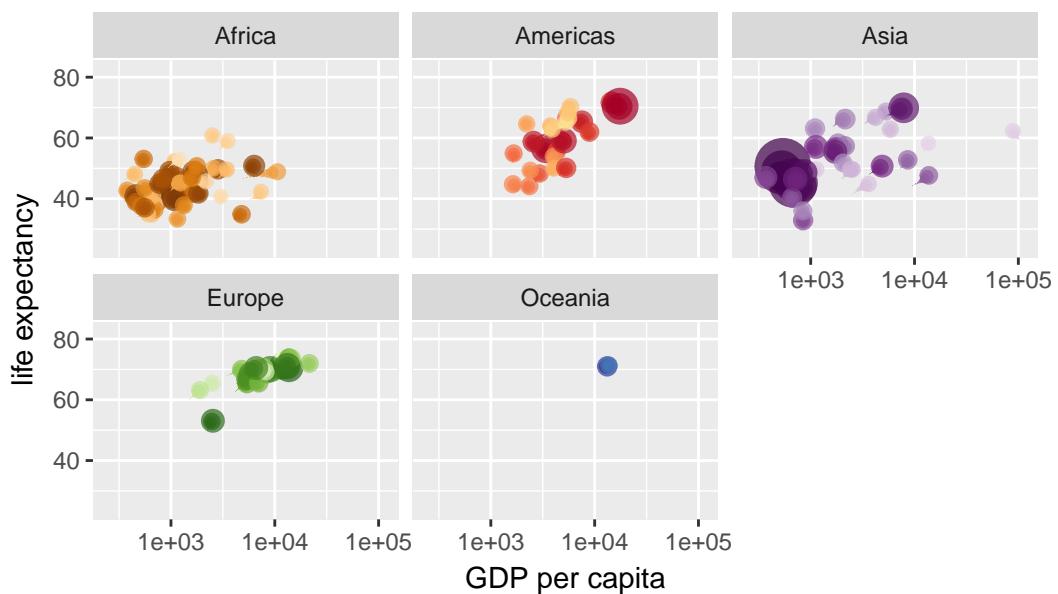
Year: 1963



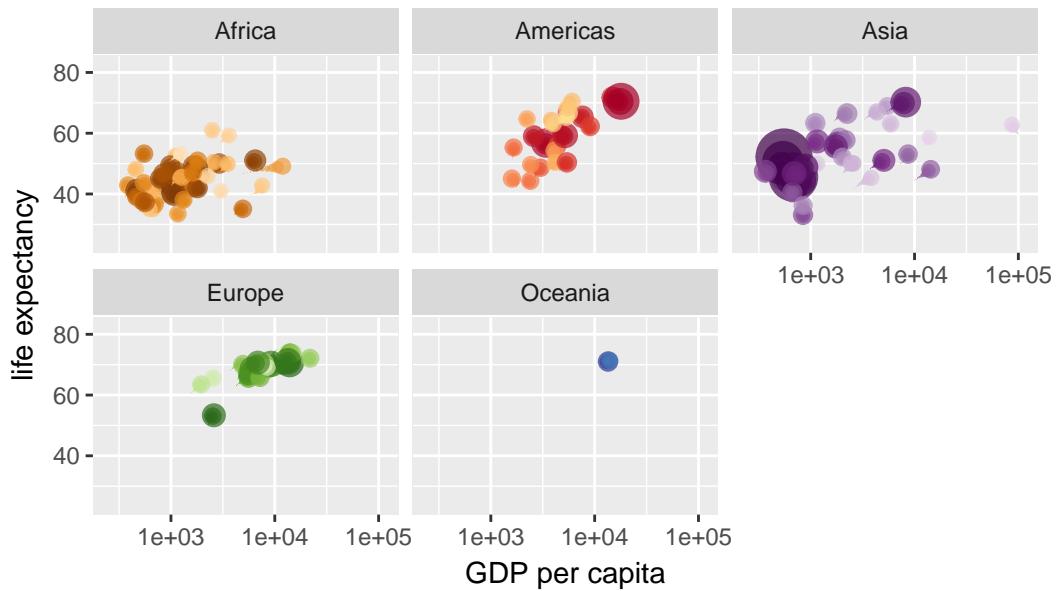
Year: 1964



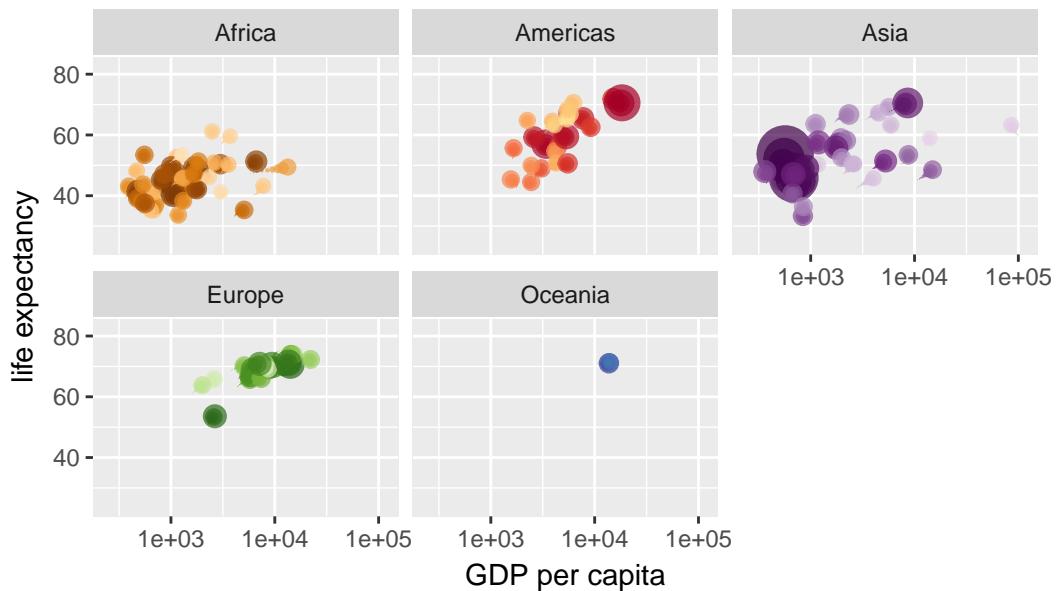
Year: 1964



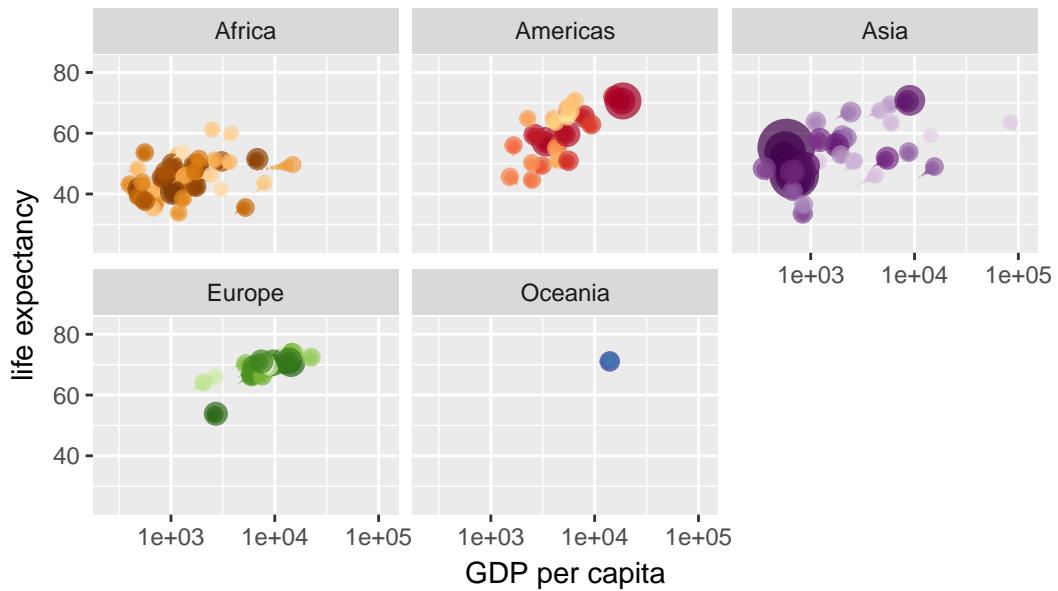
Year: 1965



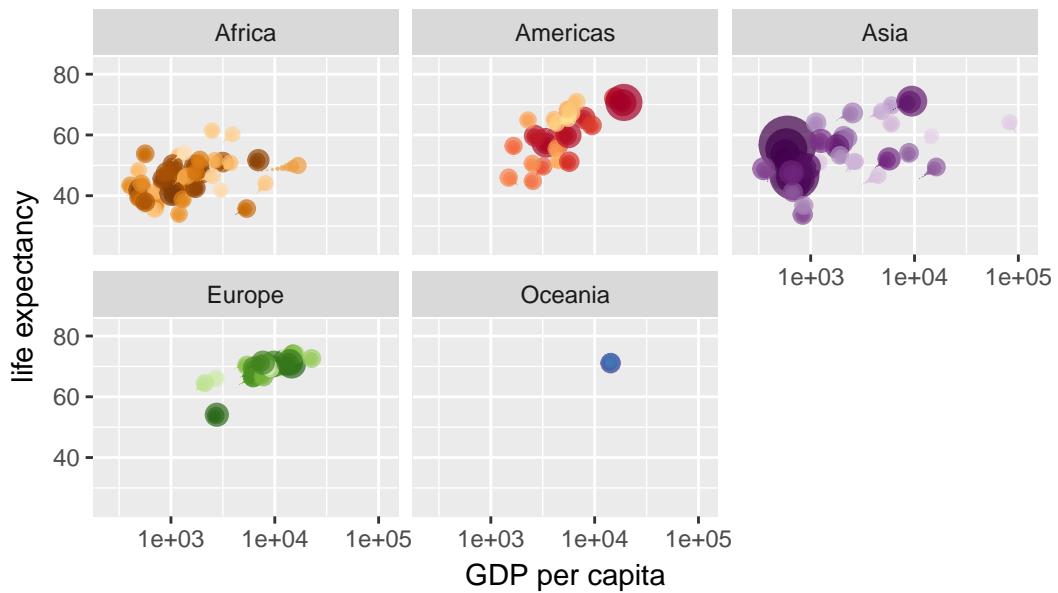
Year: 1965



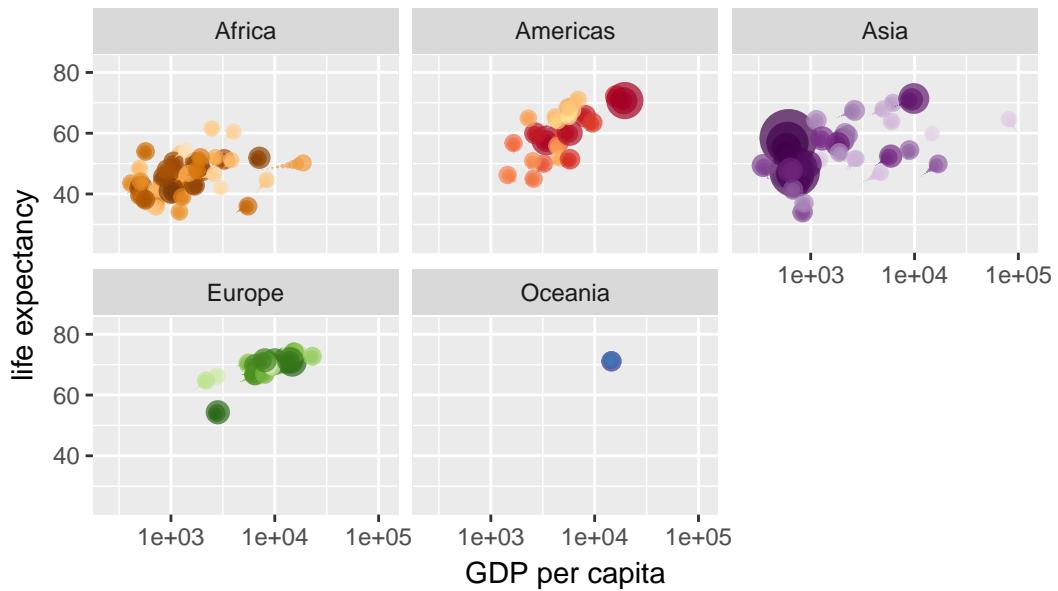
Year: 1966



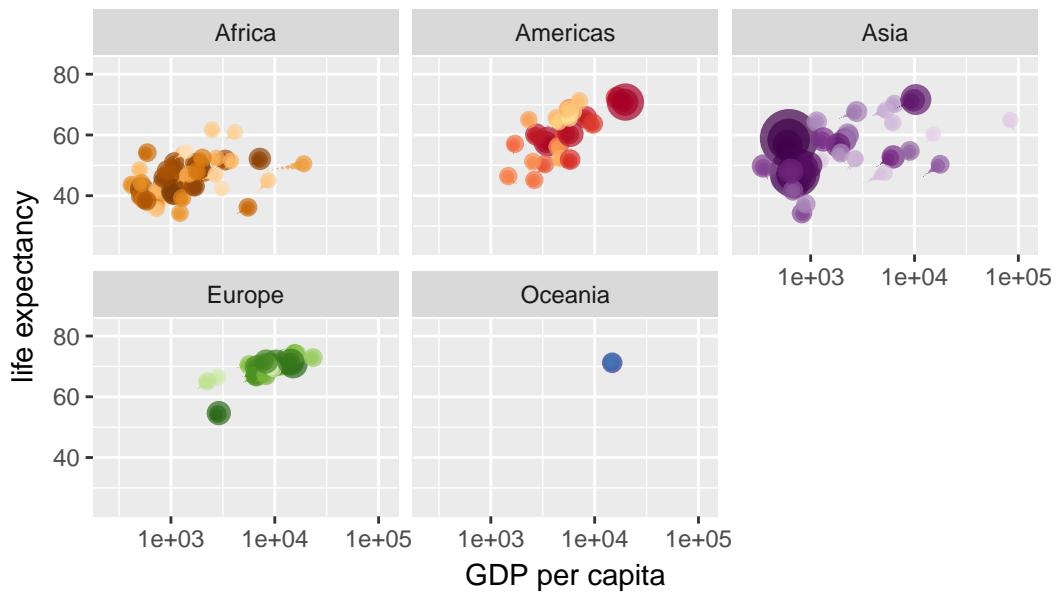
Year: 1966



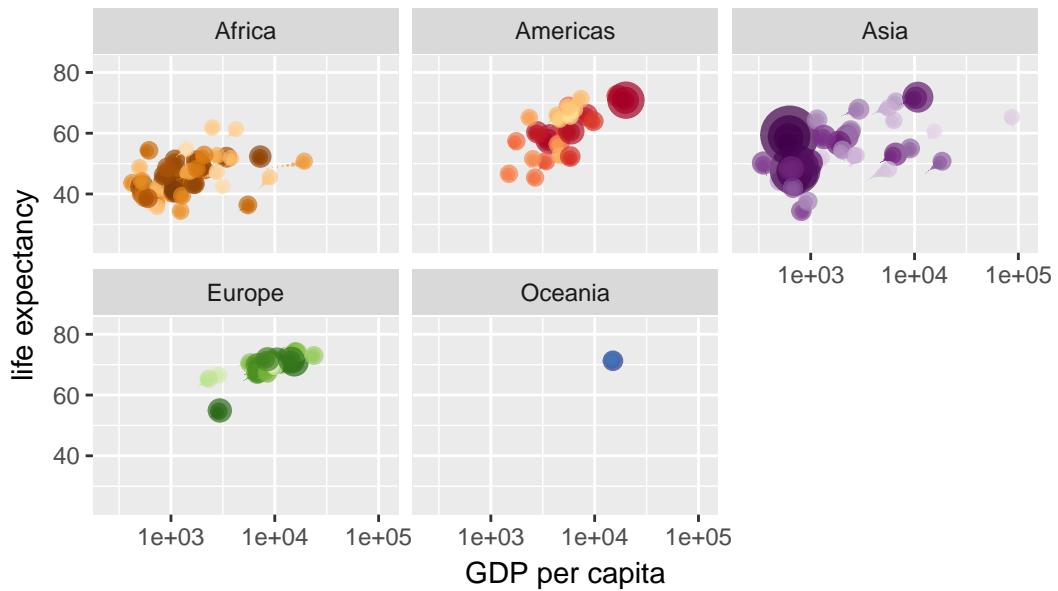
Year: 1967



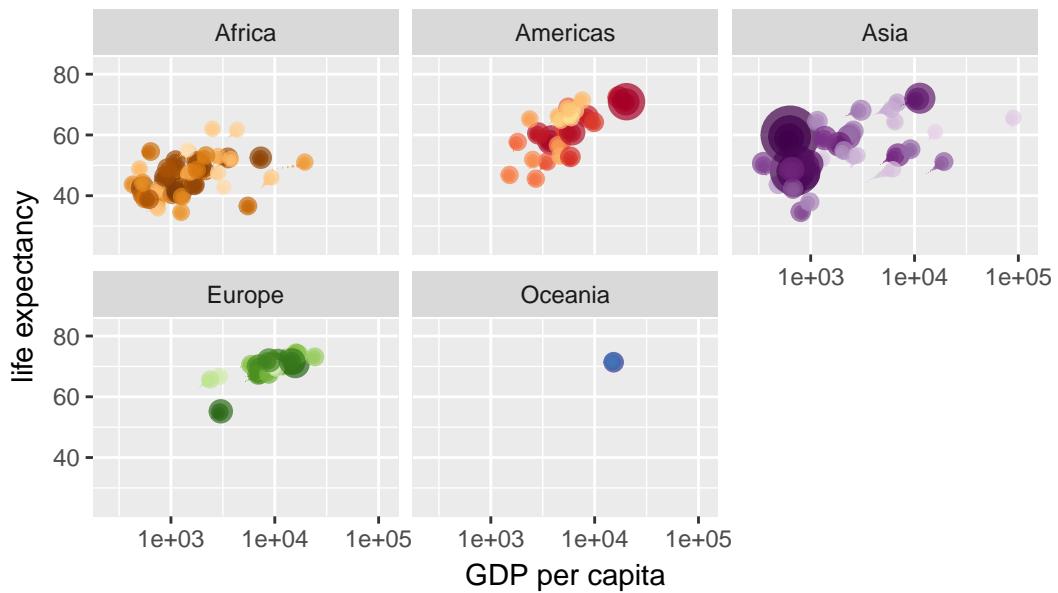
Year: 1968



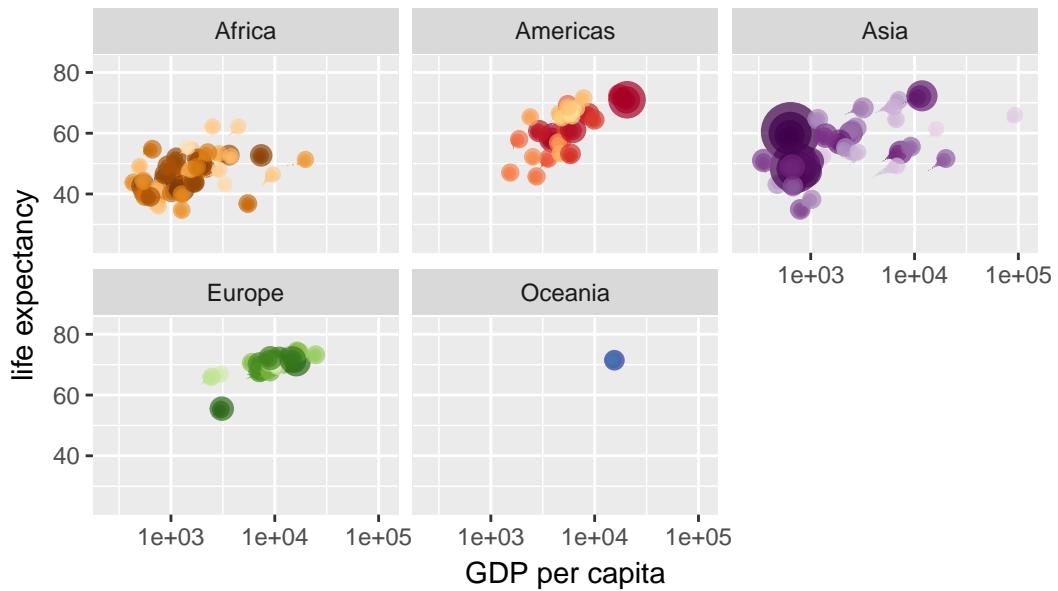
Year: 1968



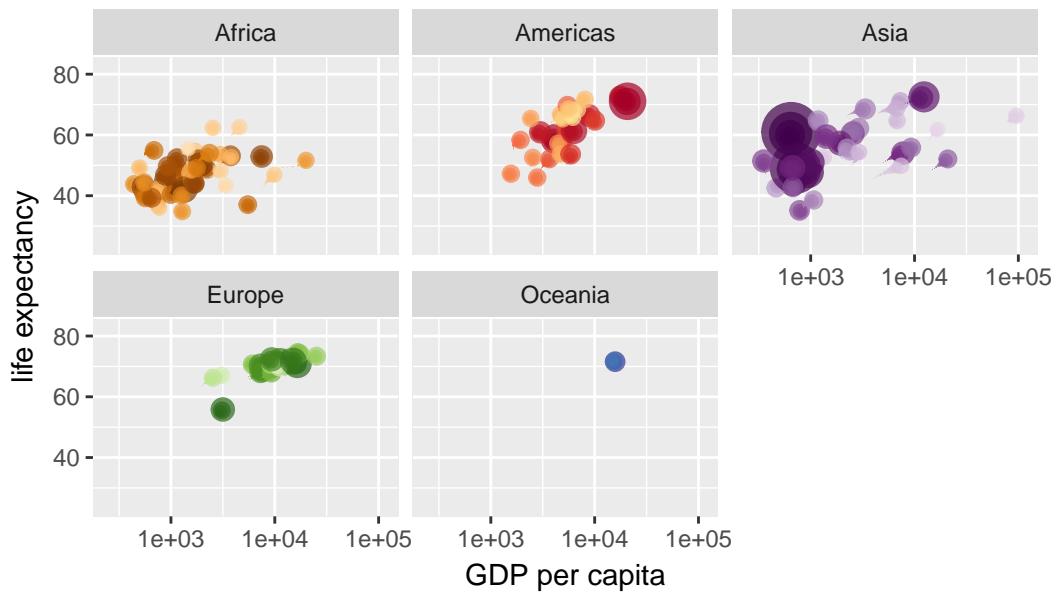
Year: 1969



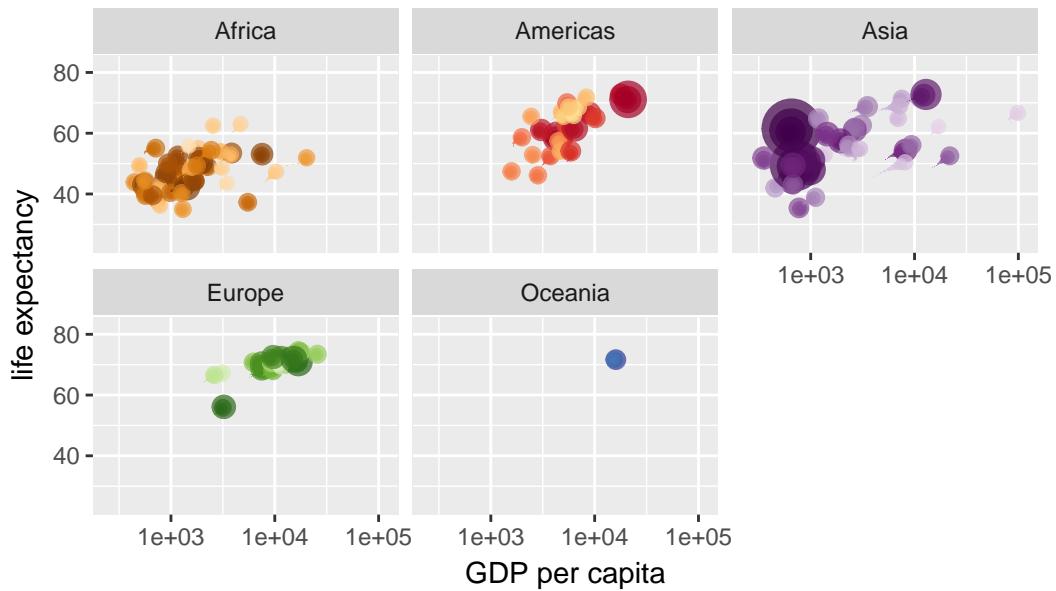
Year: 1969



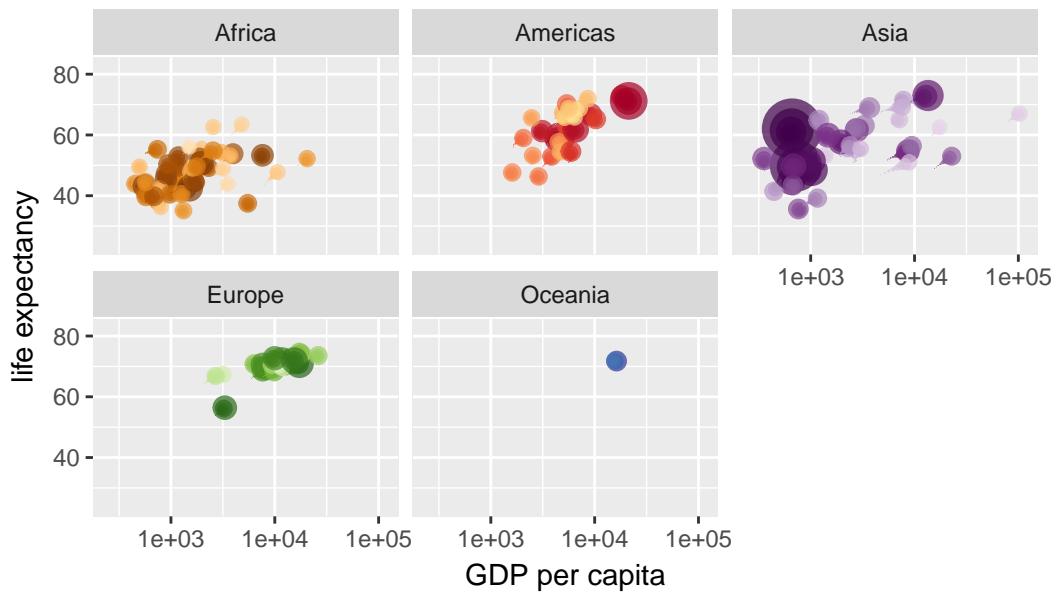
Year: 1970



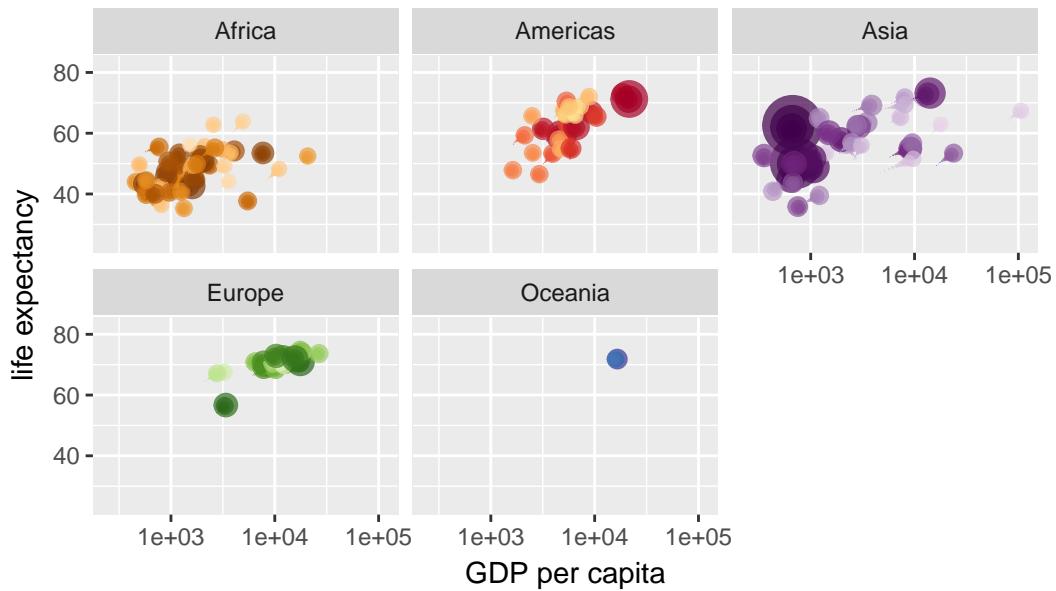
Year: 1970



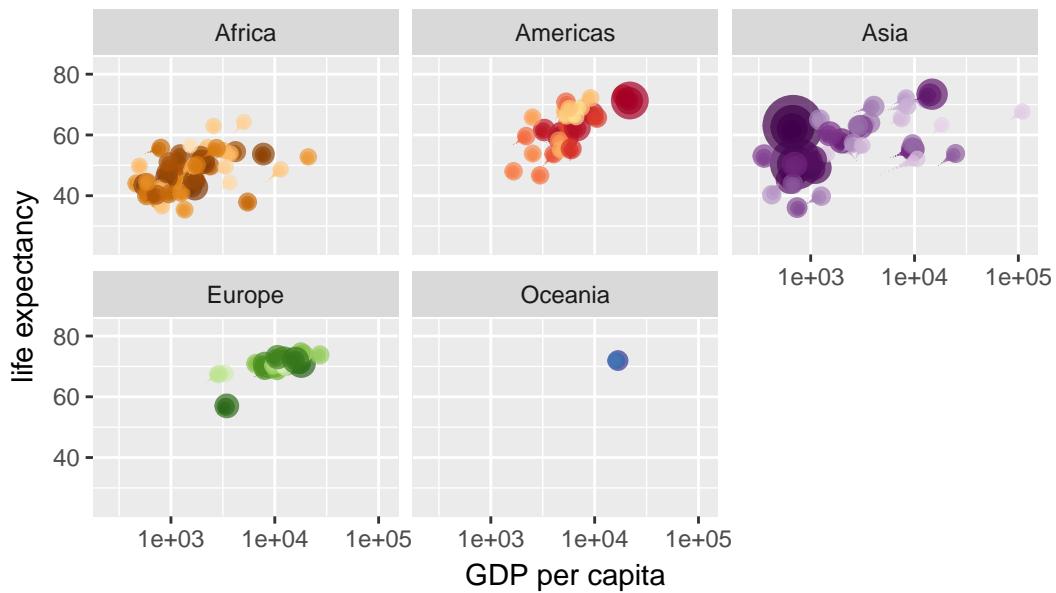
Year: 1971



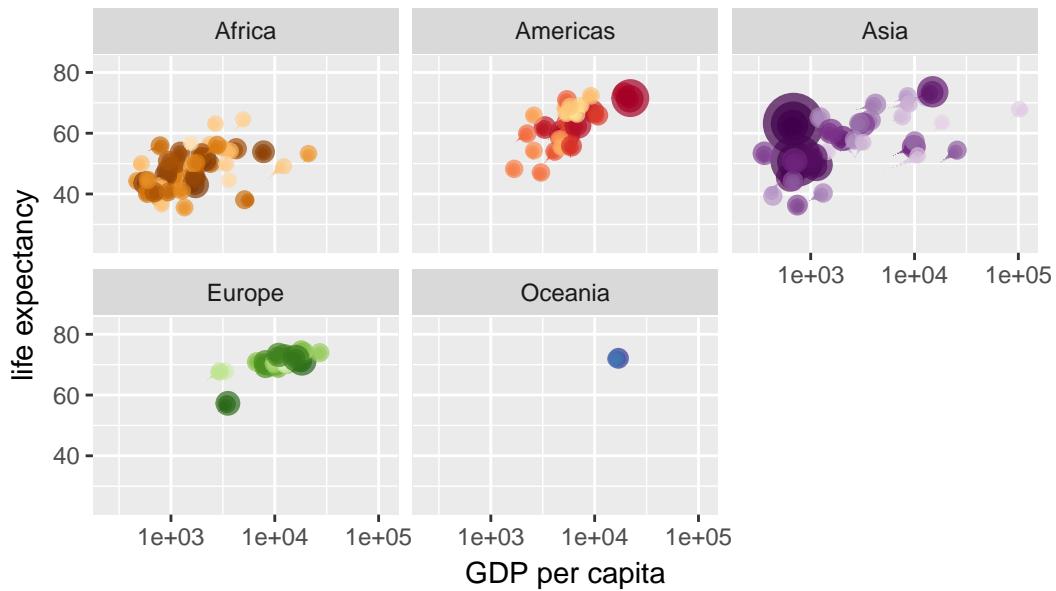
Year: 1971



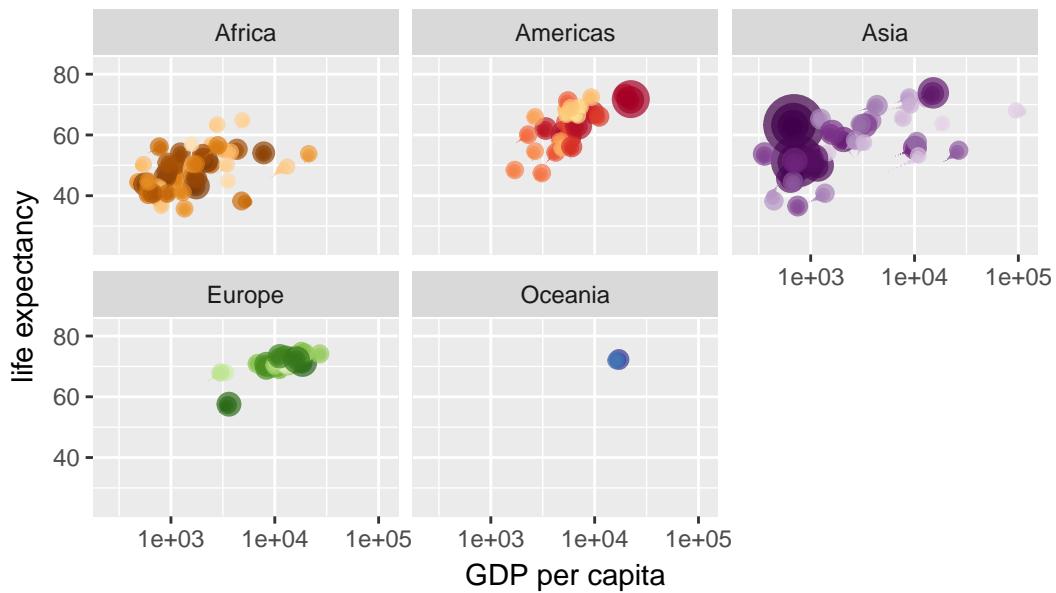
Year: 1972



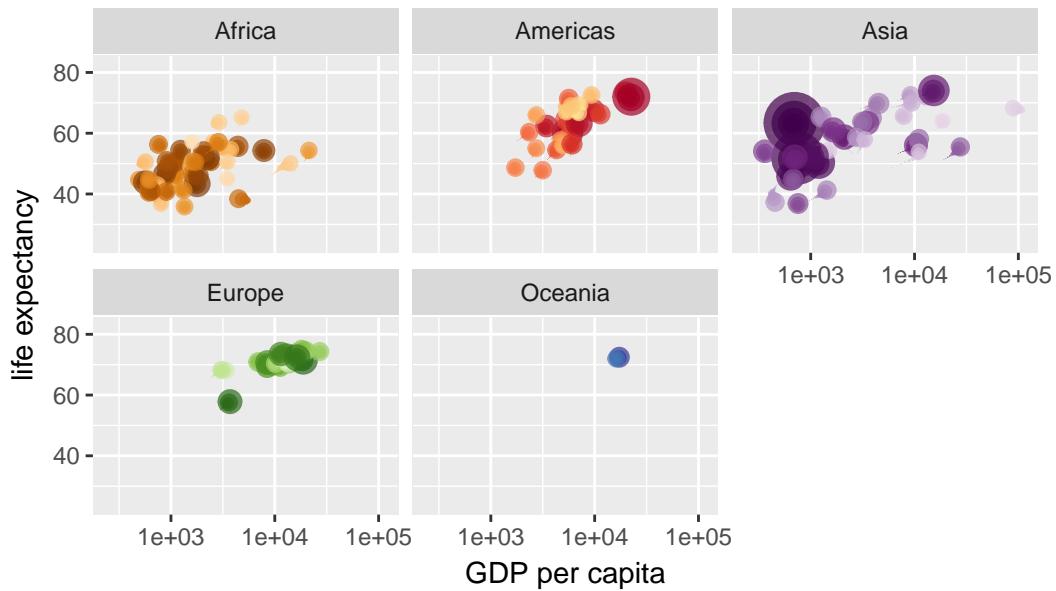
Year: 1973



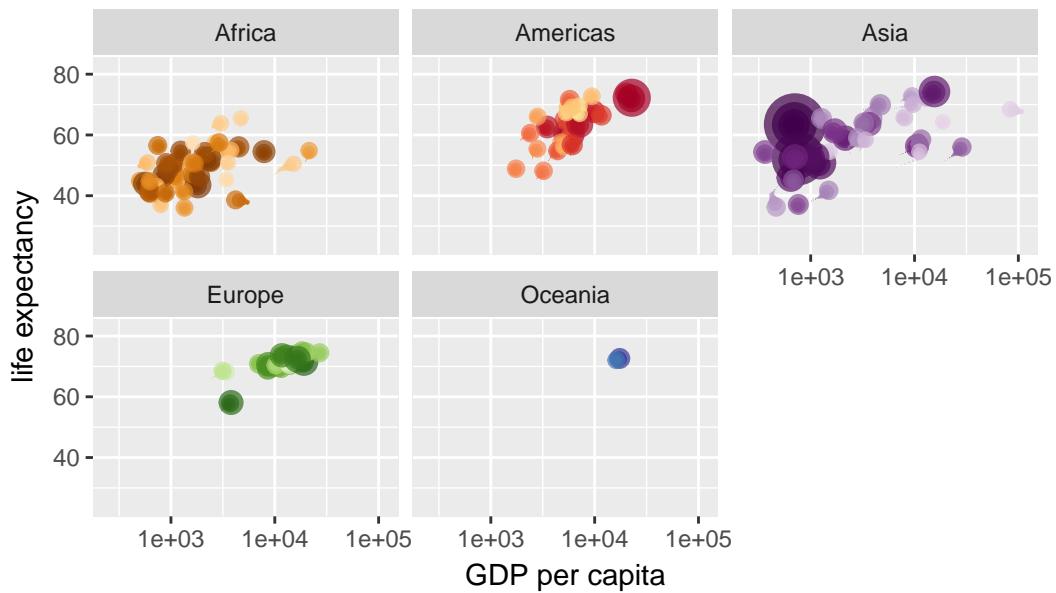
Year: 1973



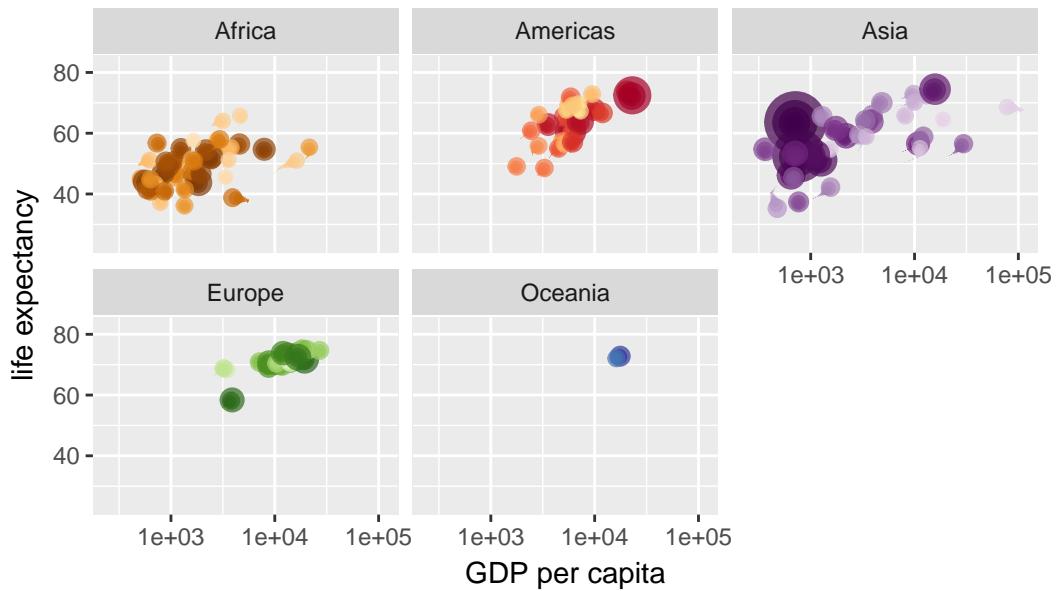
Year: 1974



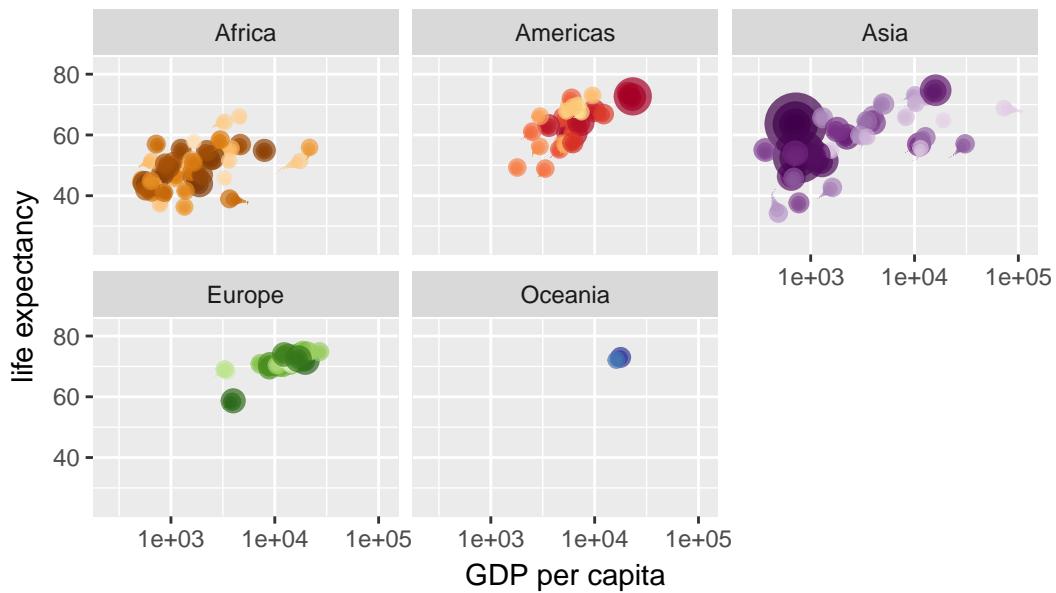
Year: 1974



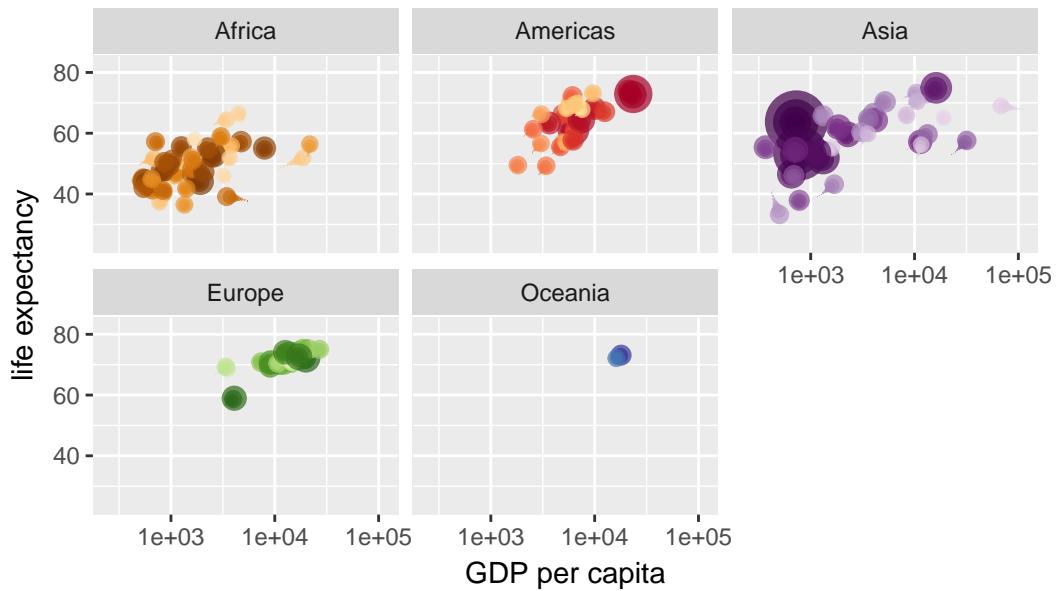
Year: 1975



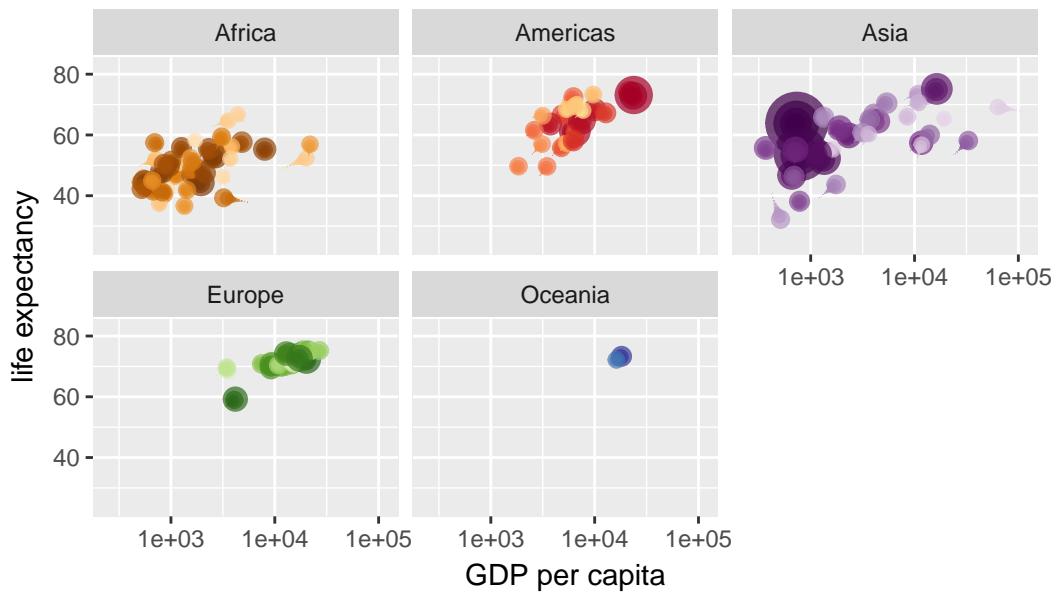
Year: 1975



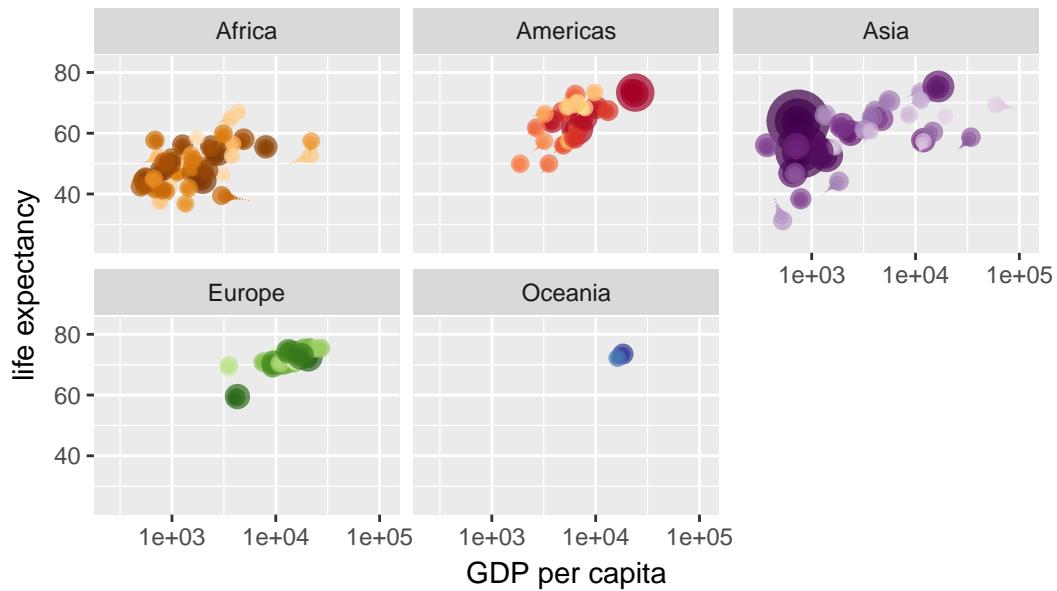
Year: 1976



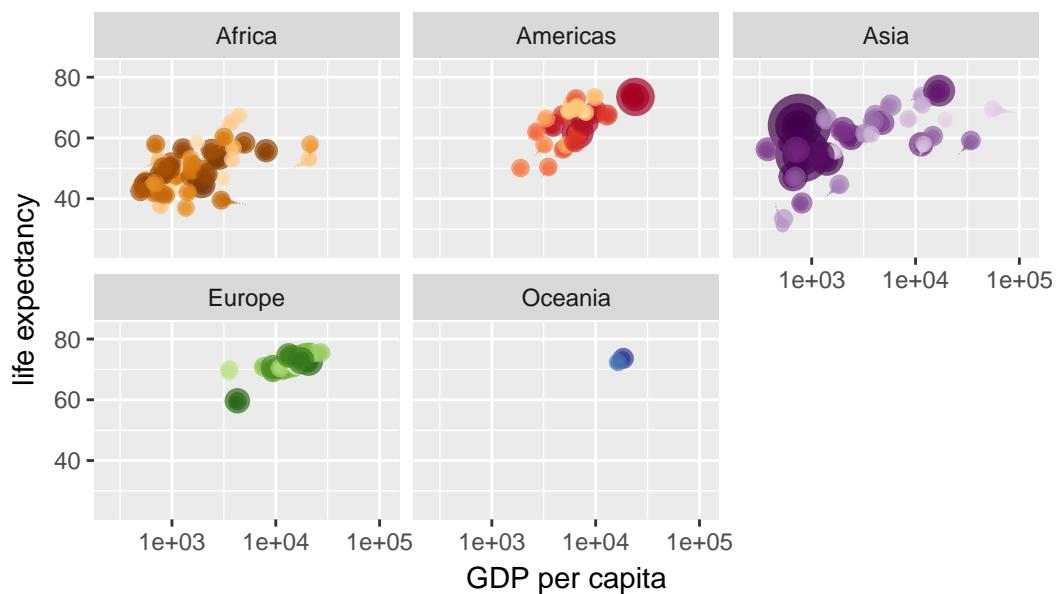
Year: 1976



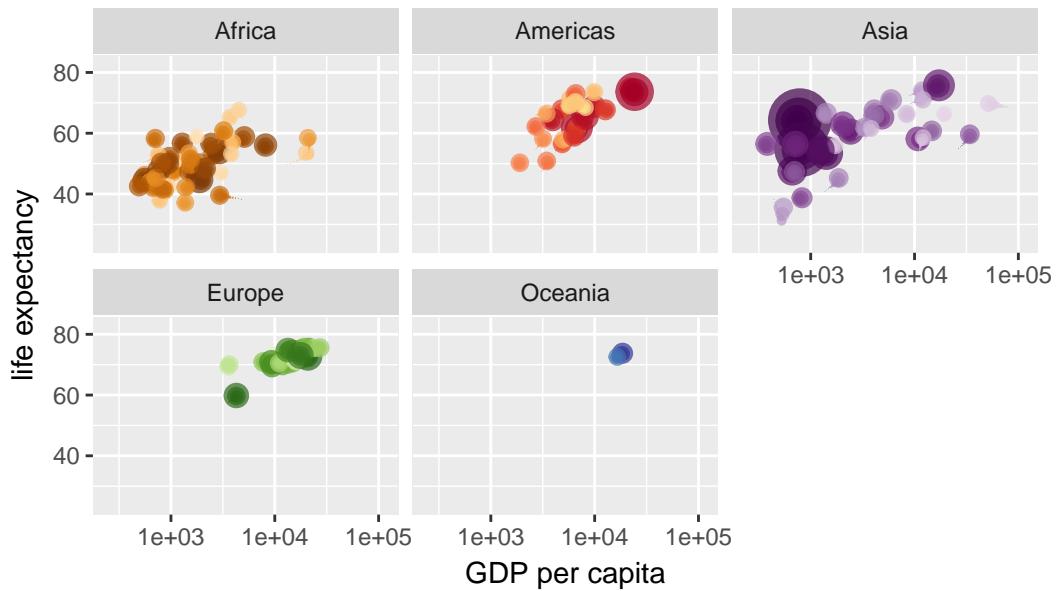
Year: 1977



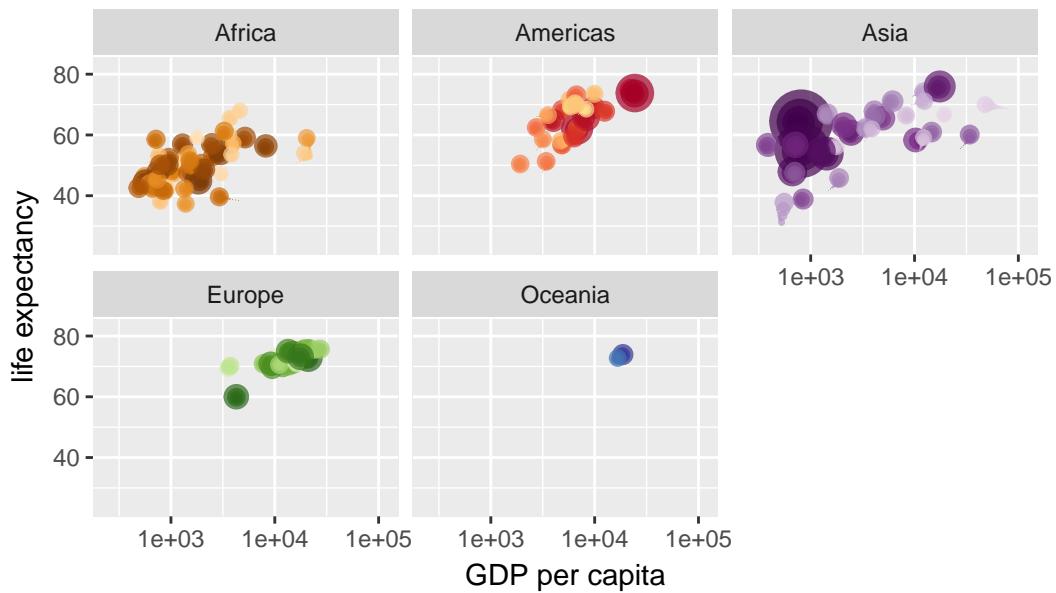
Year: 1978



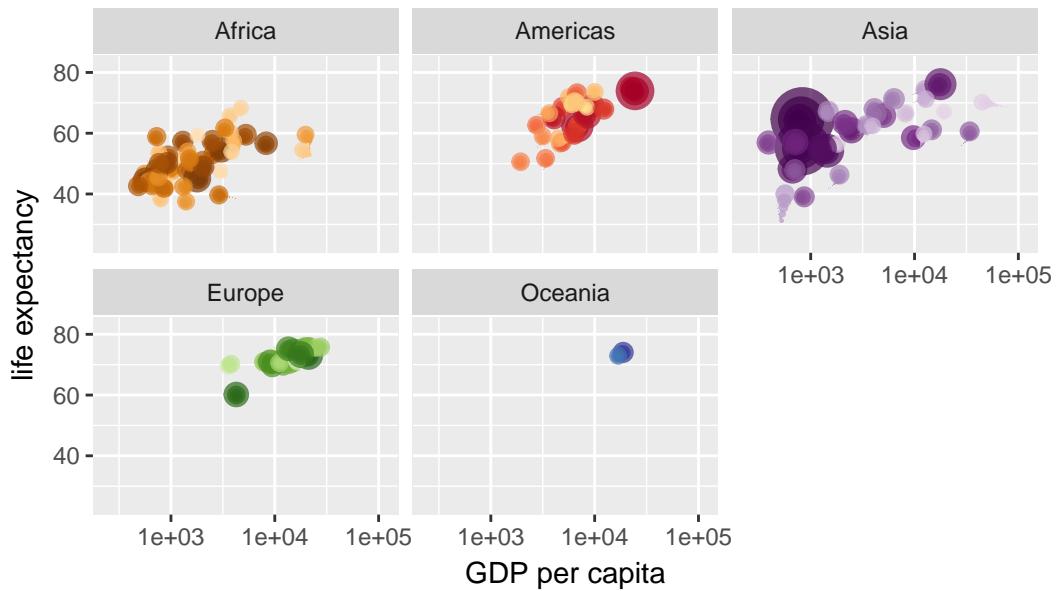
Year: 1978



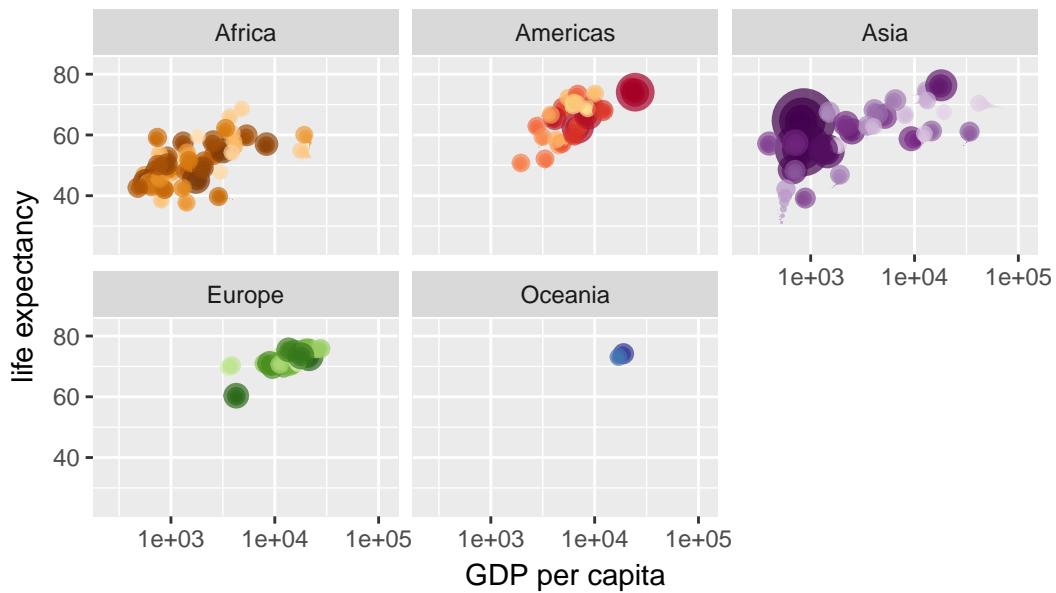
Year: 1979



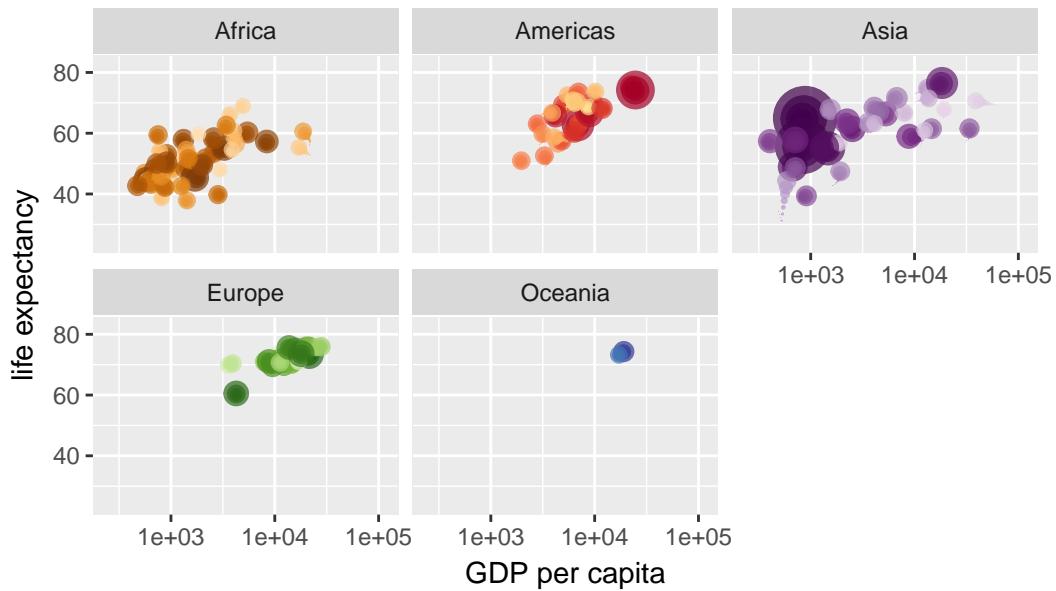
Year: 1979



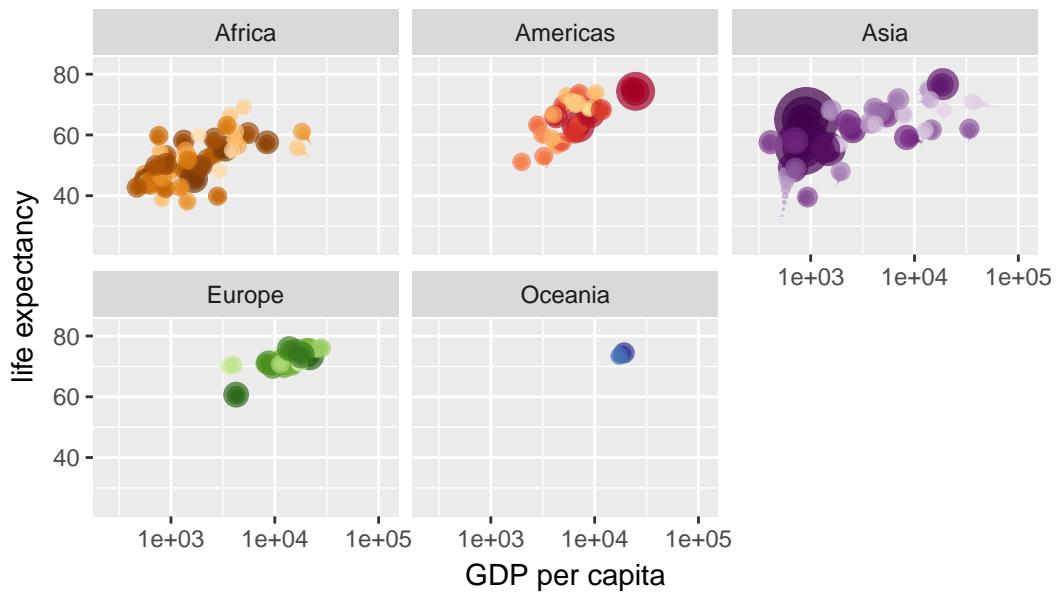
Year: 1980



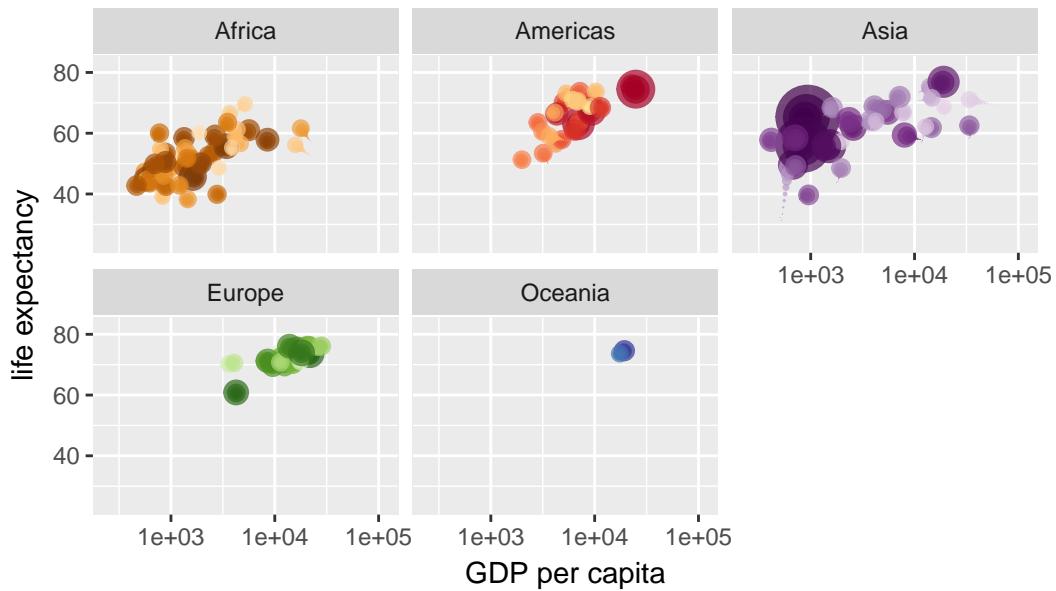
Year: 1980



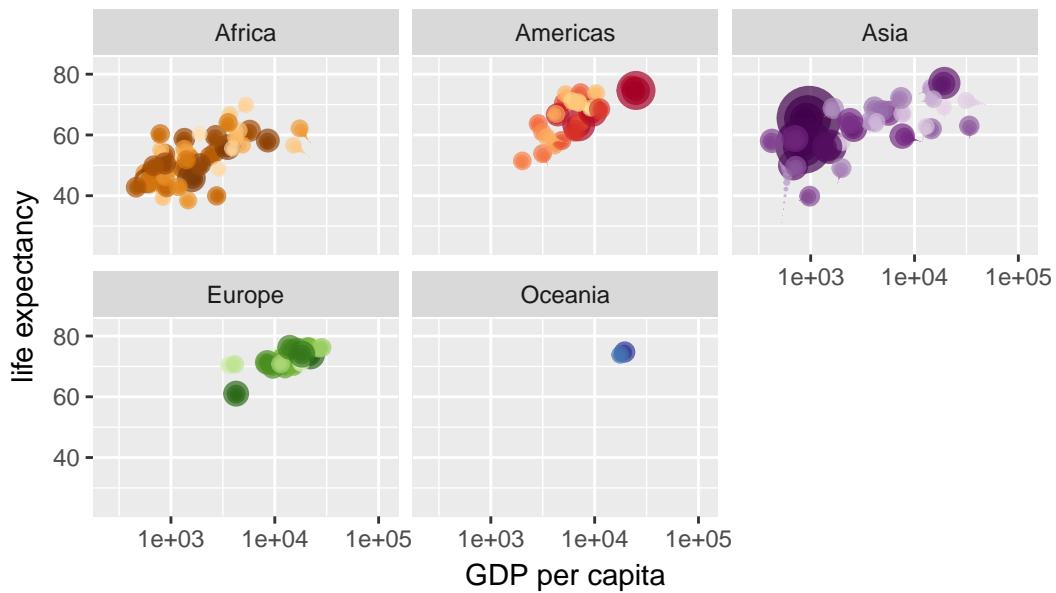
Year: 1981



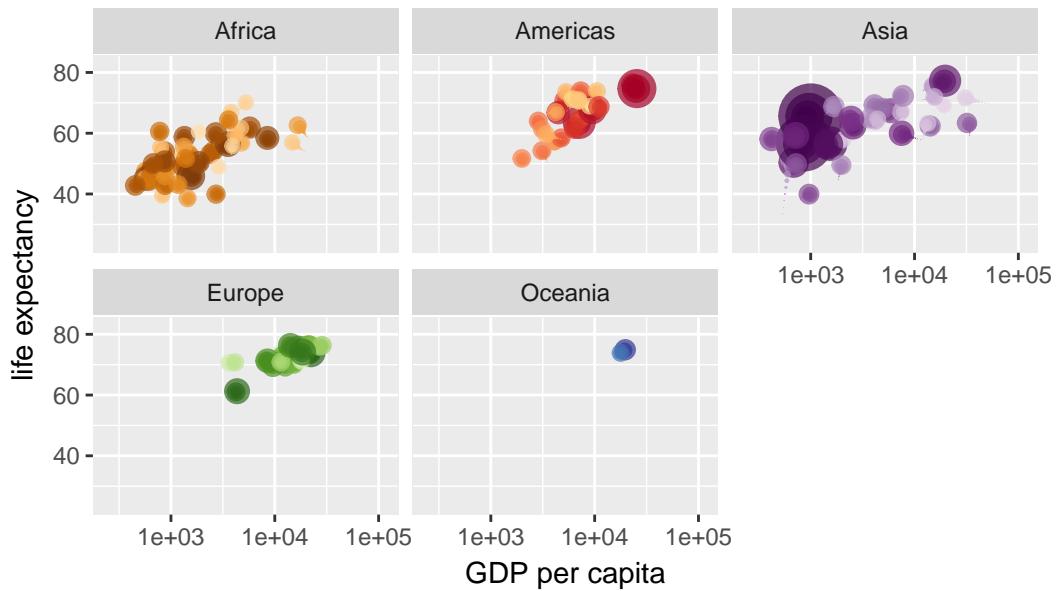
Year: 1981



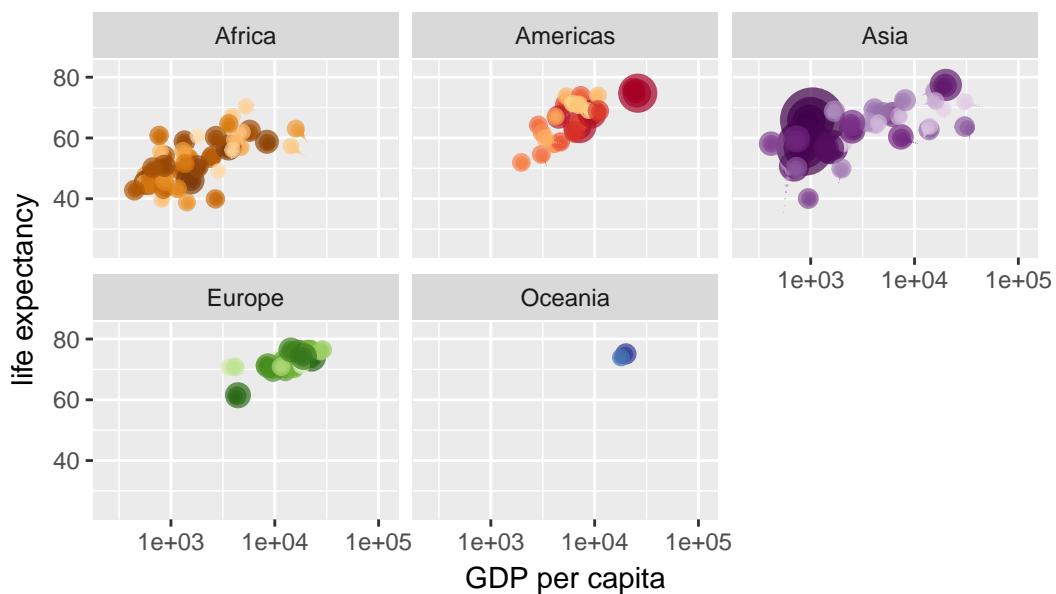
Year: 1982



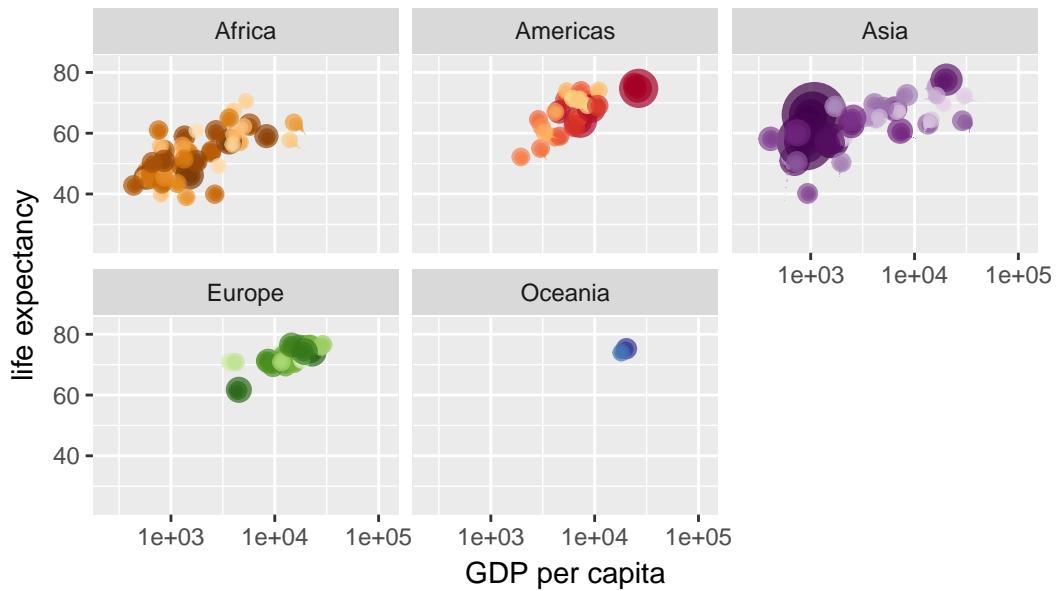
Year: 1983



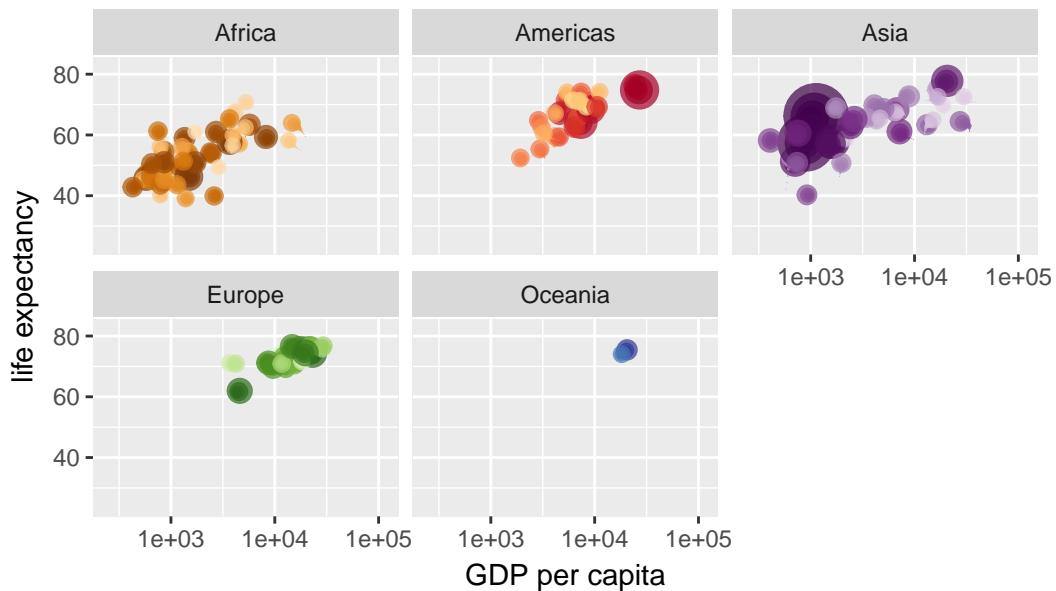
Year: 1983



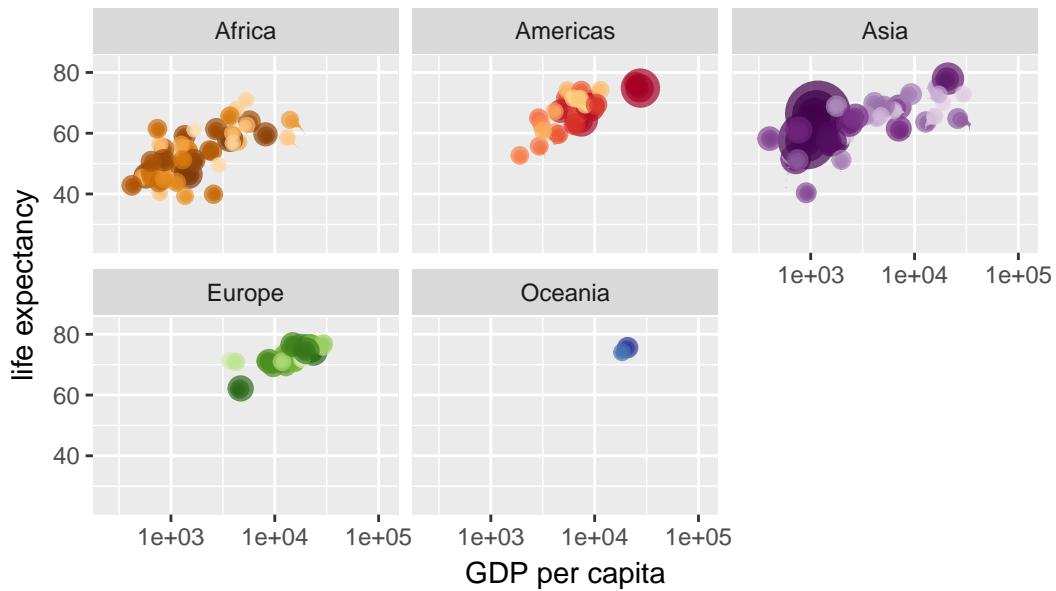
Year: 1984



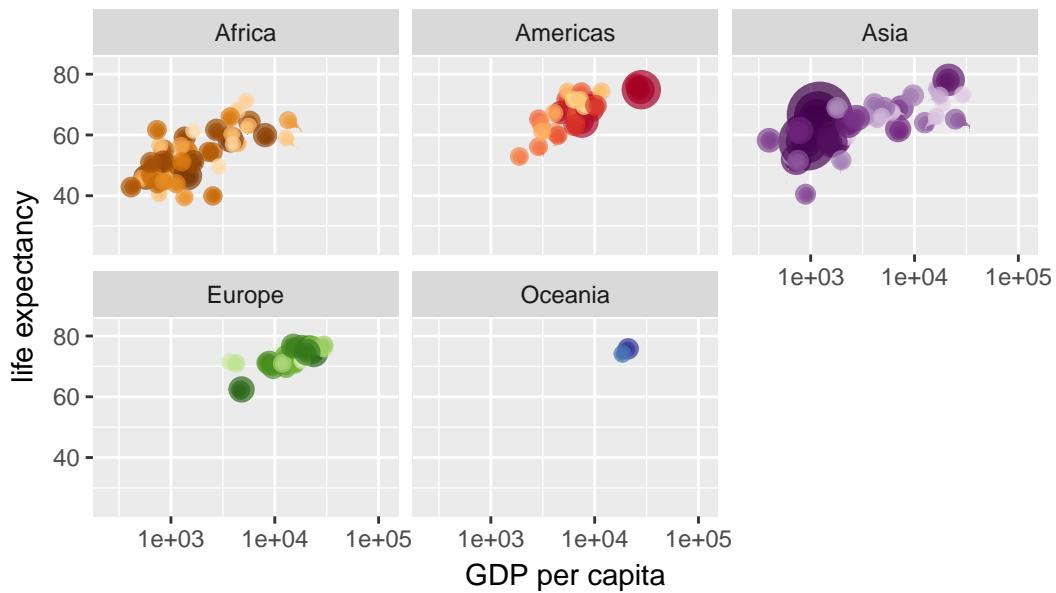
Year: 1984



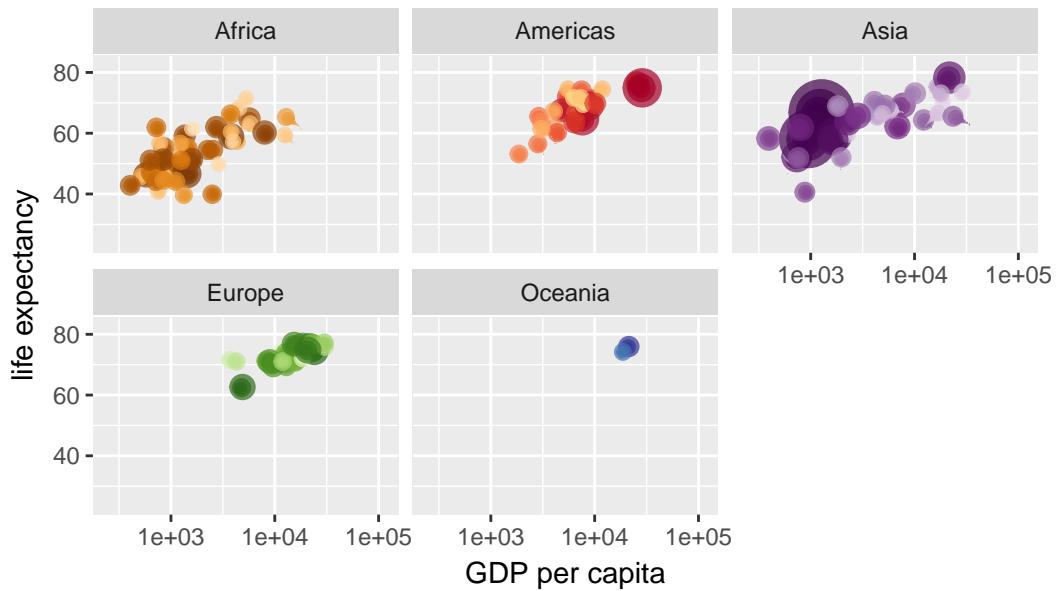
Year: 1985



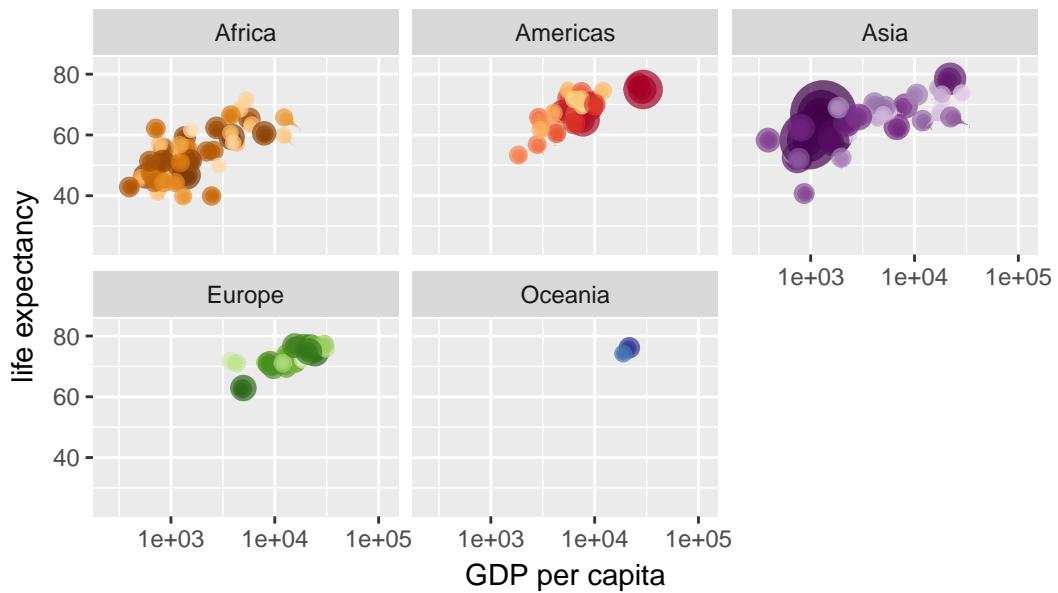
Year: 1985



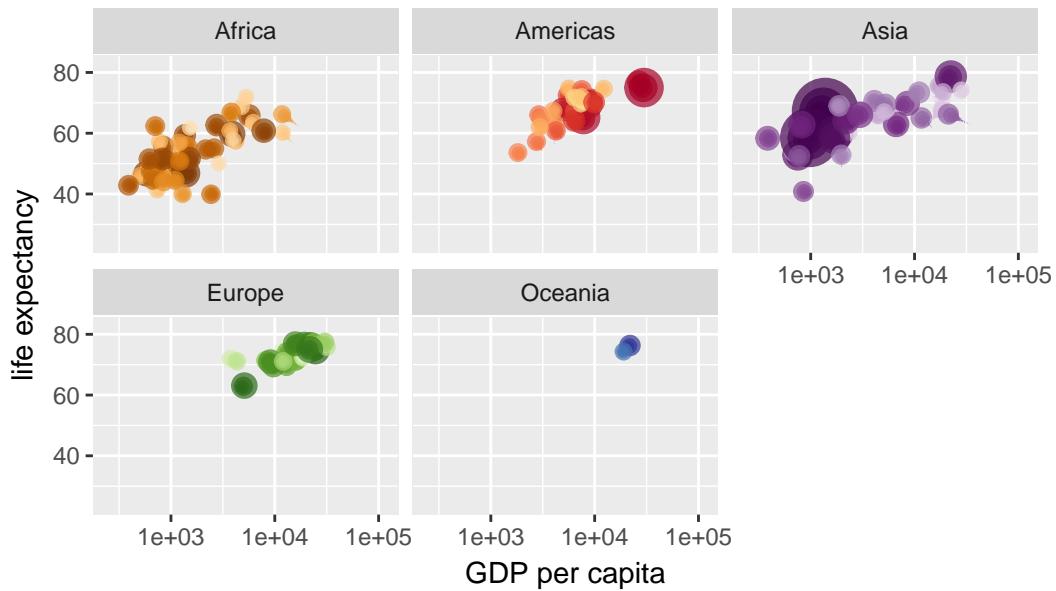
Year: 1986



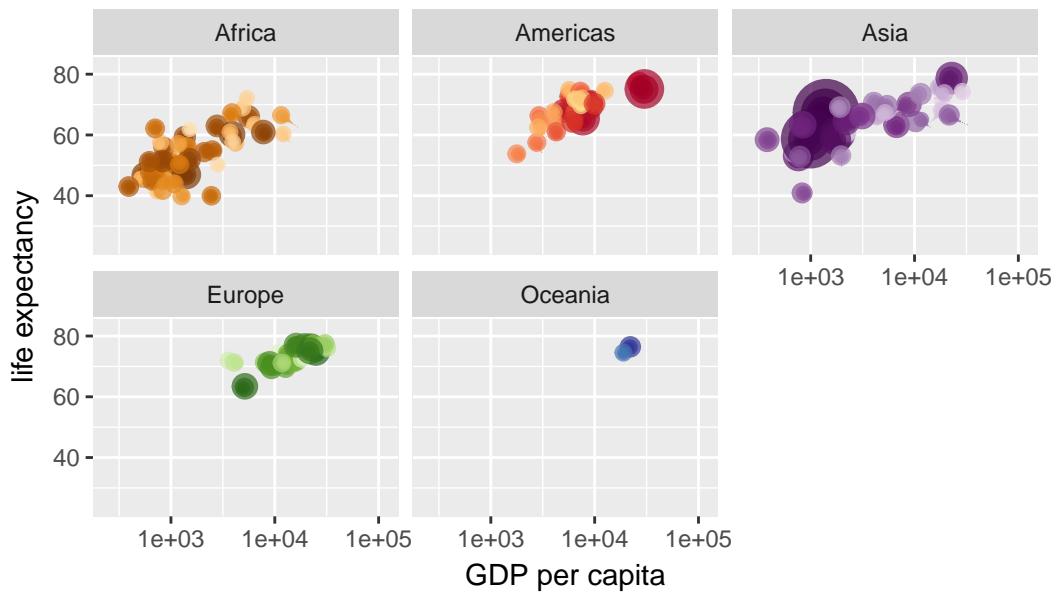
Year: 1986



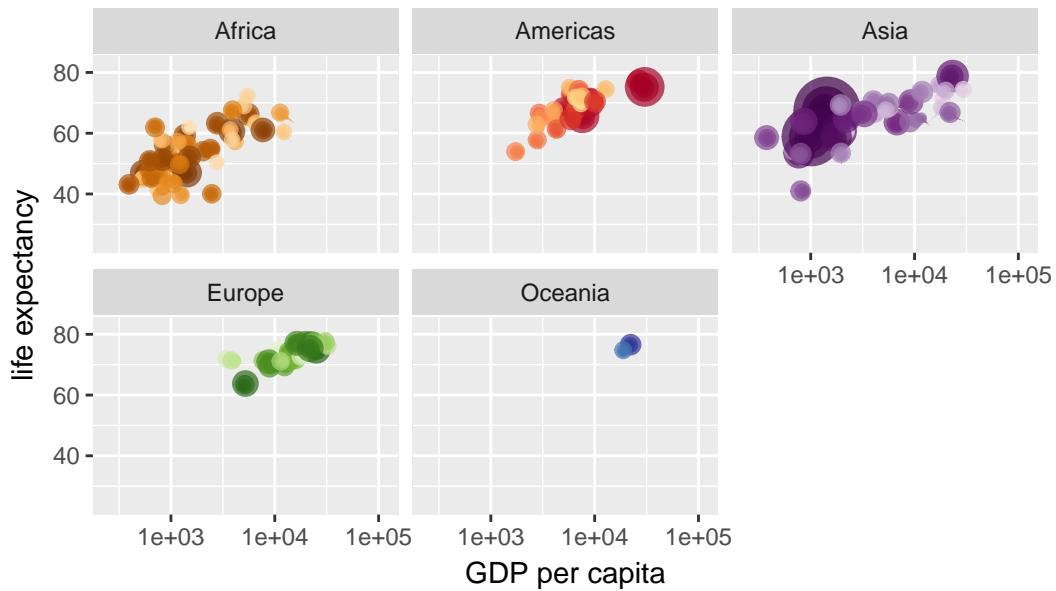
Year: 1987



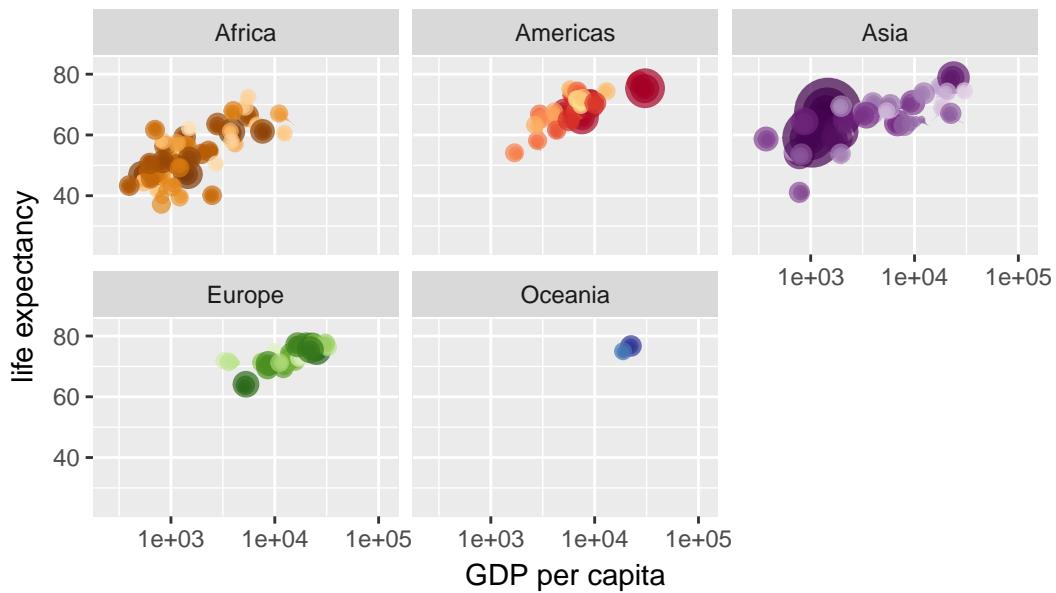
Year: 1988



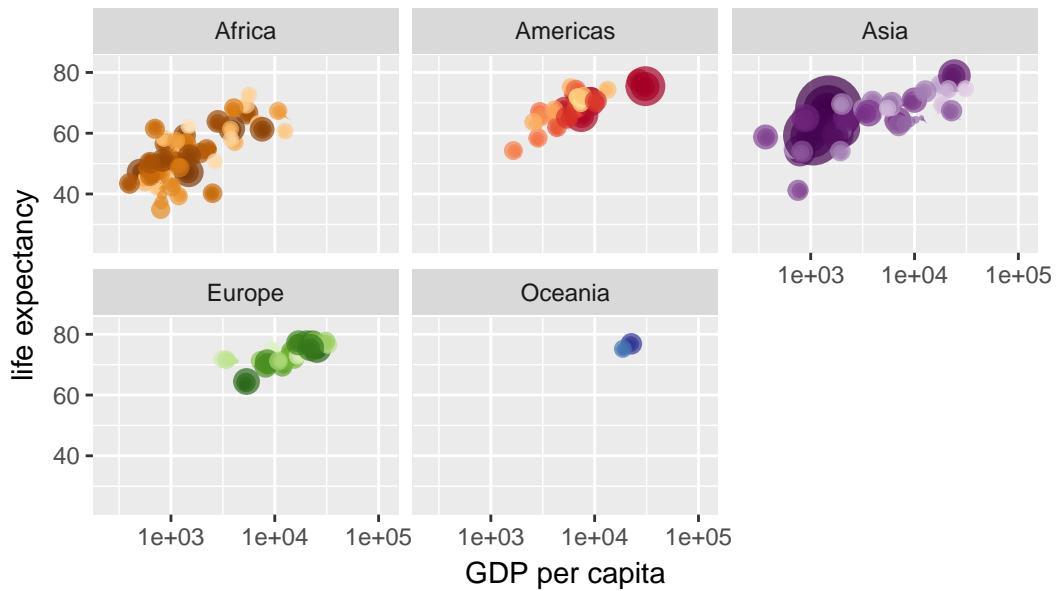
Year: 1988



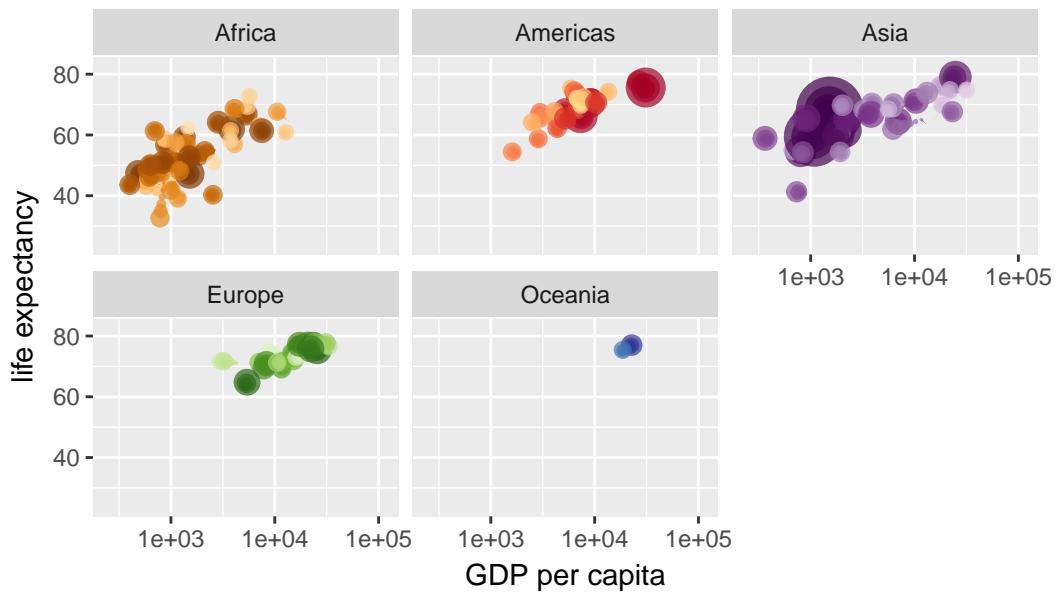
Year: 1989



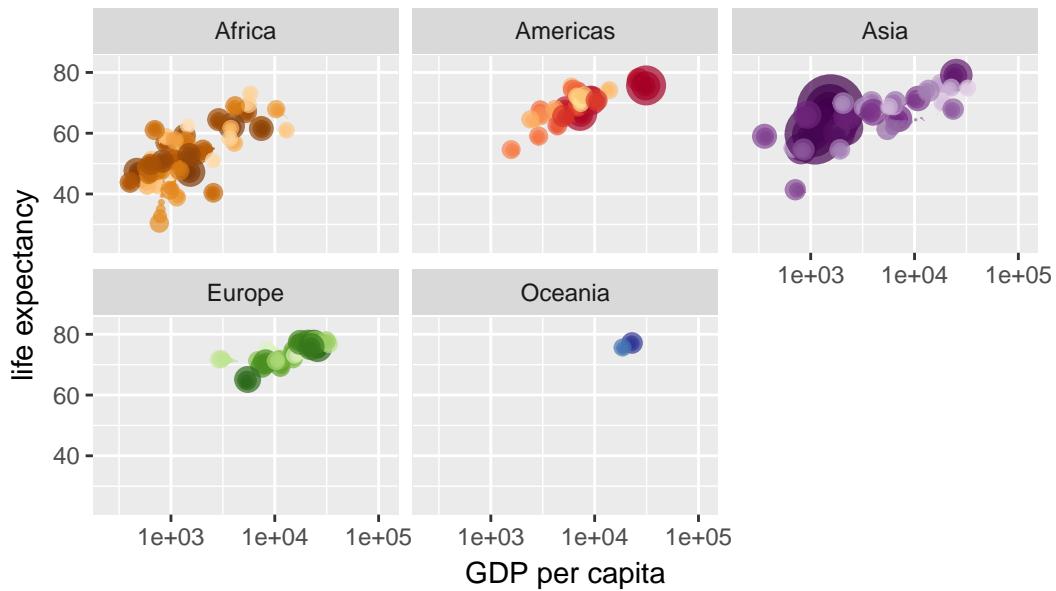
Year: 1989



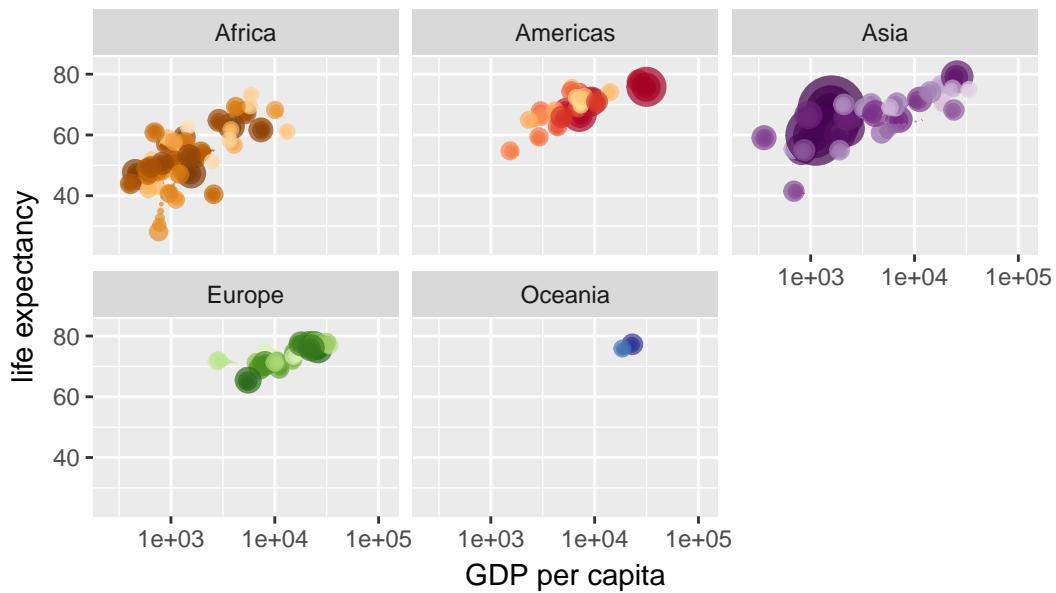
Year: 1990



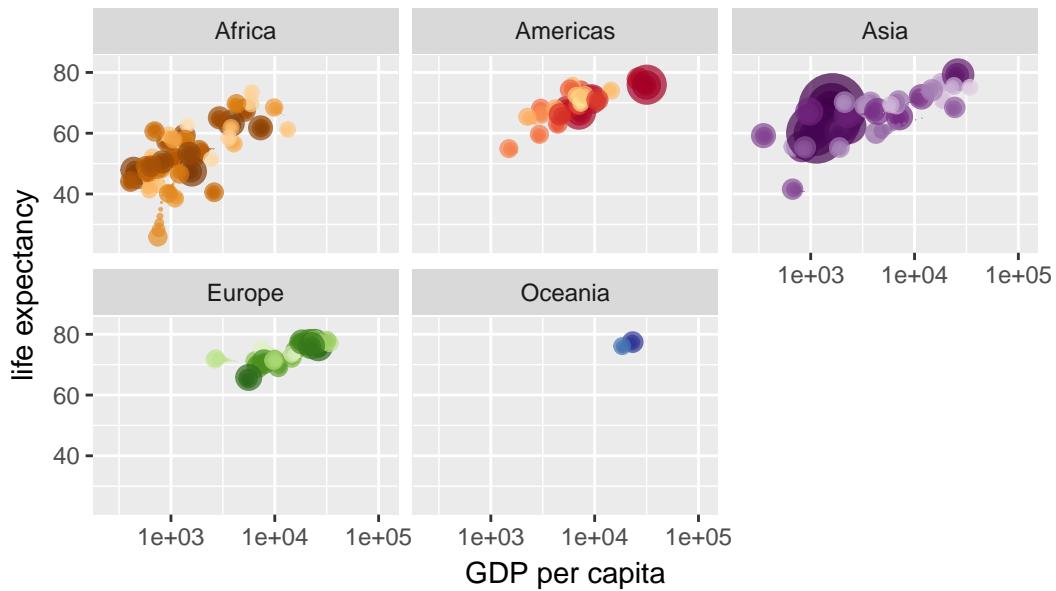
Year: 1990



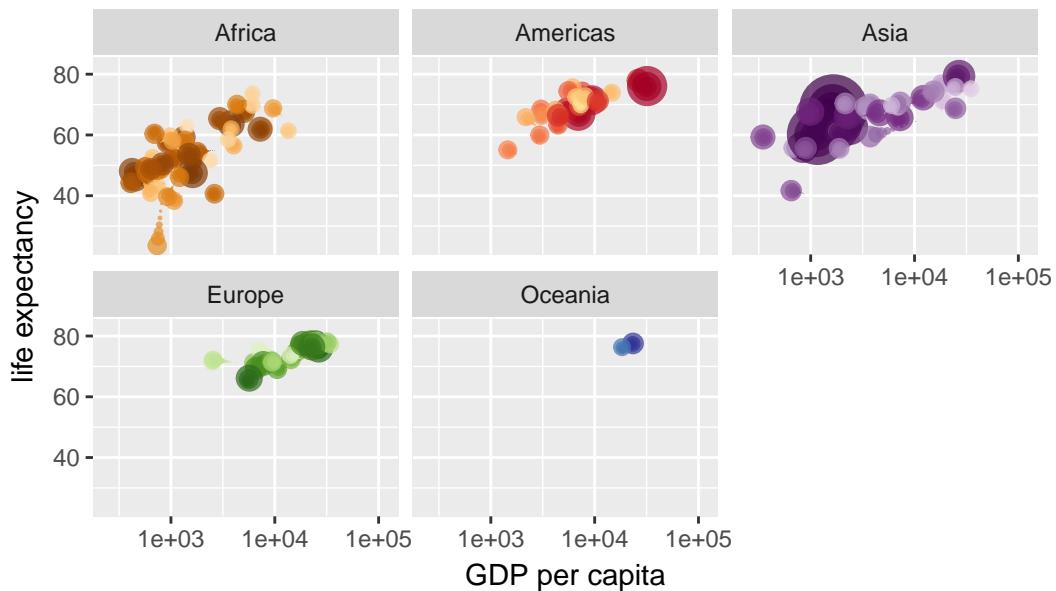
Year: 1991



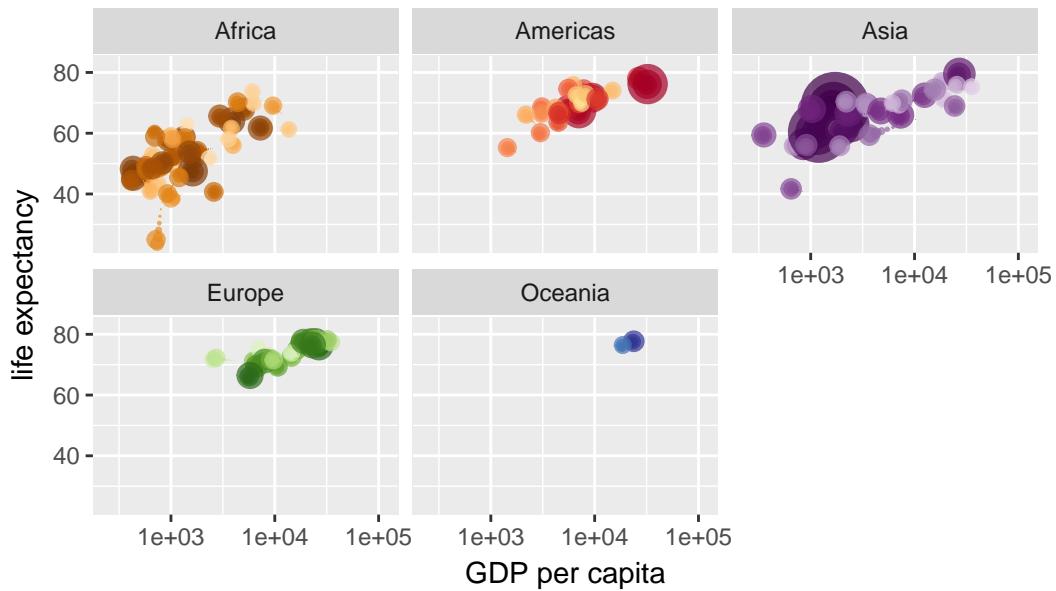
Year: 1991



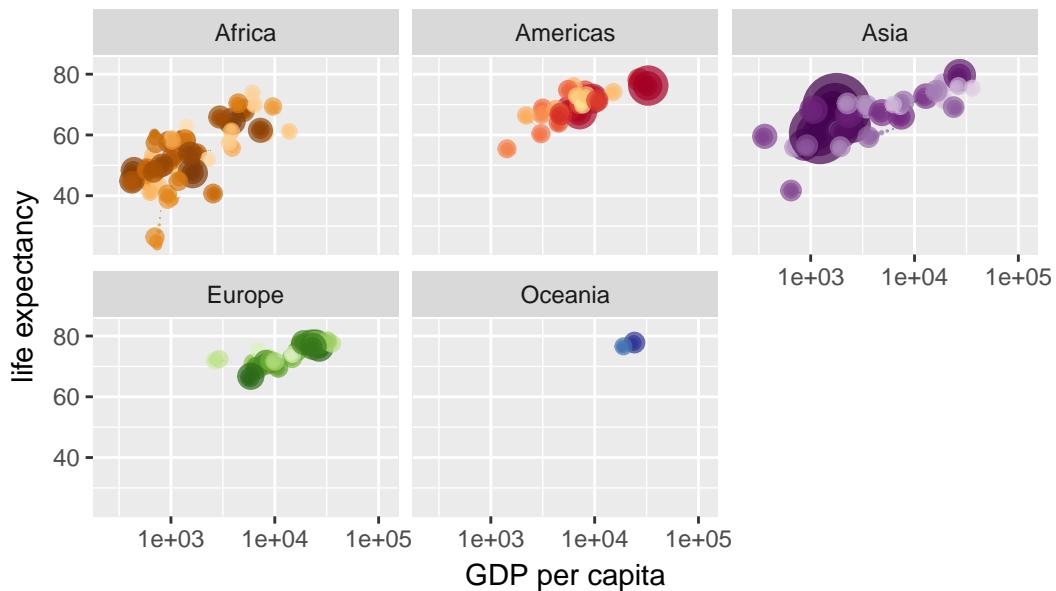
Year: 1992



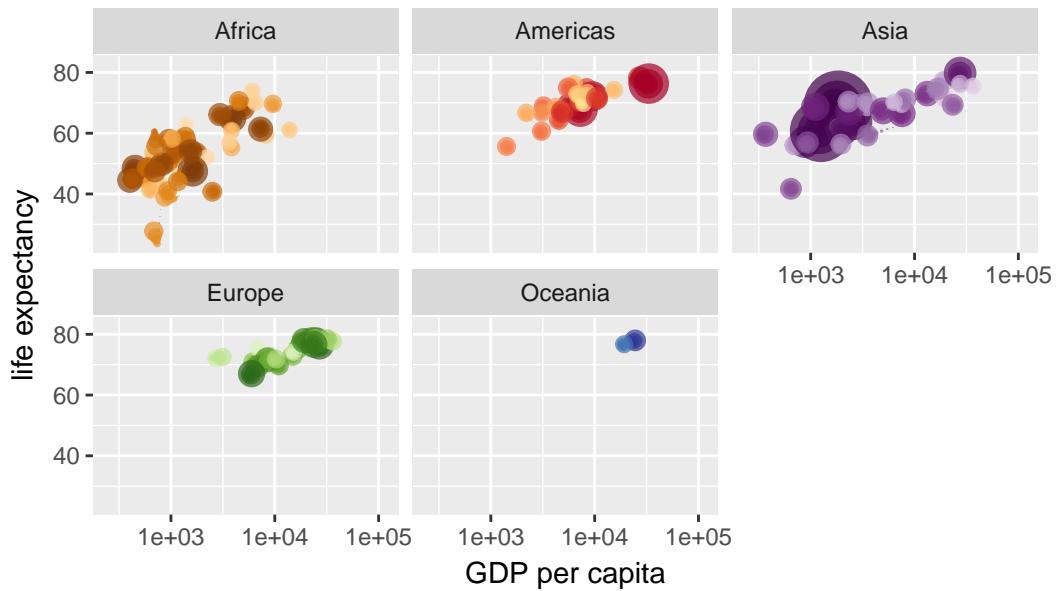
Year: 1993



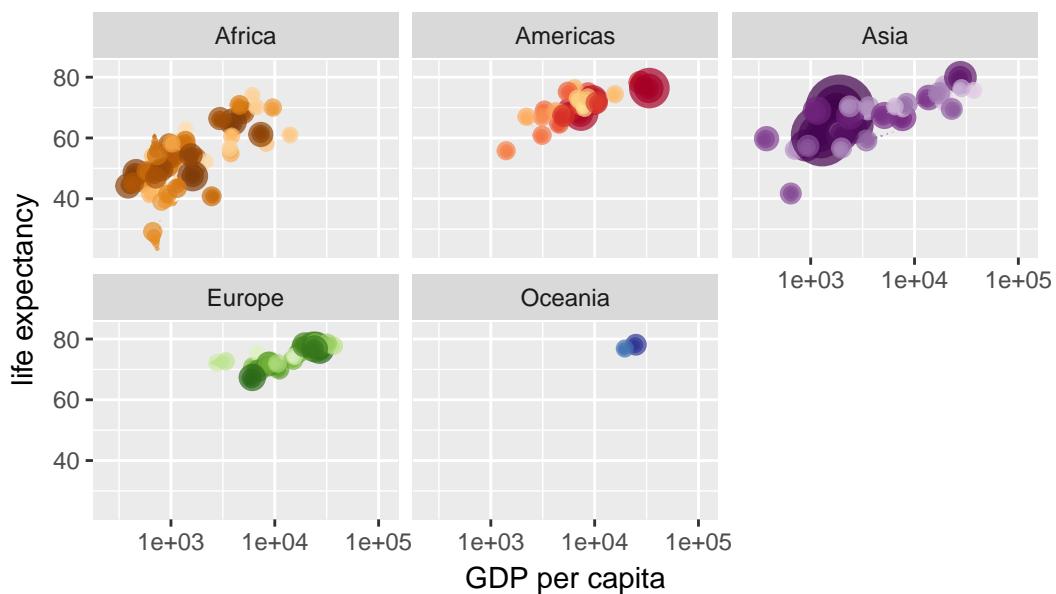
Year: 1993



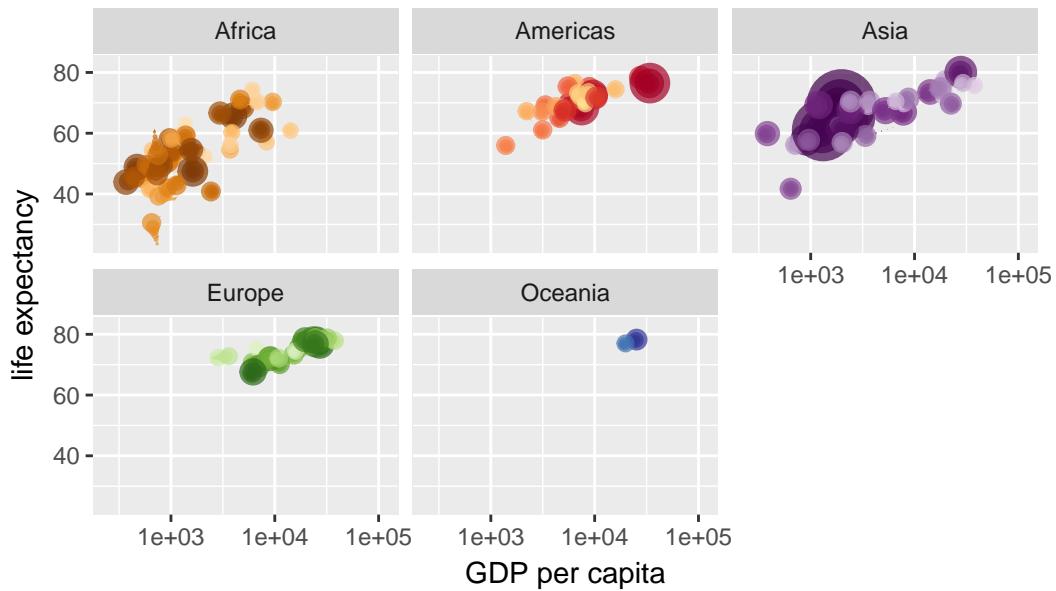
Year: 1994



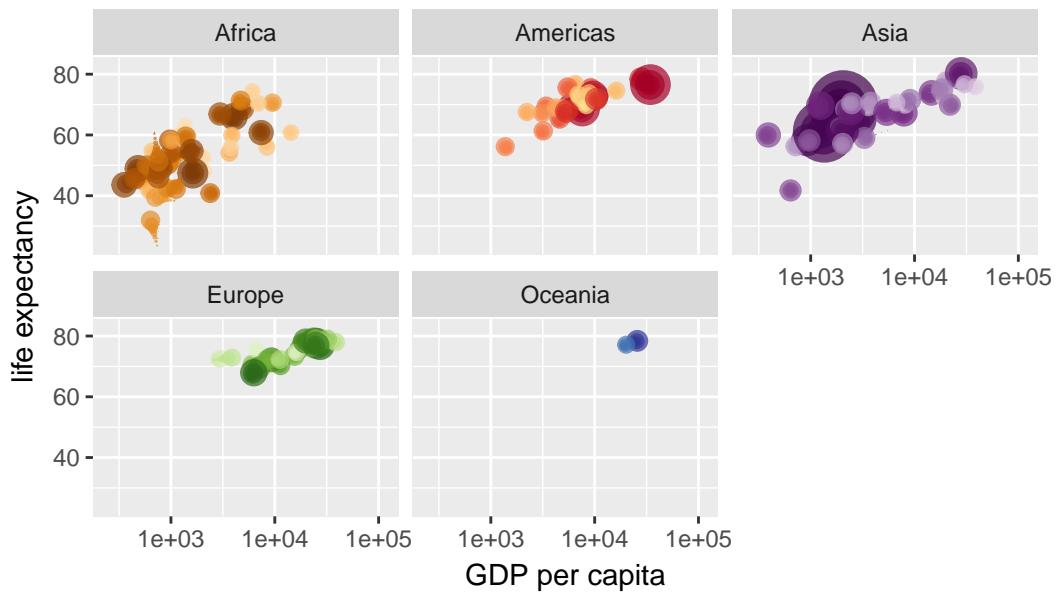
Year: 1994



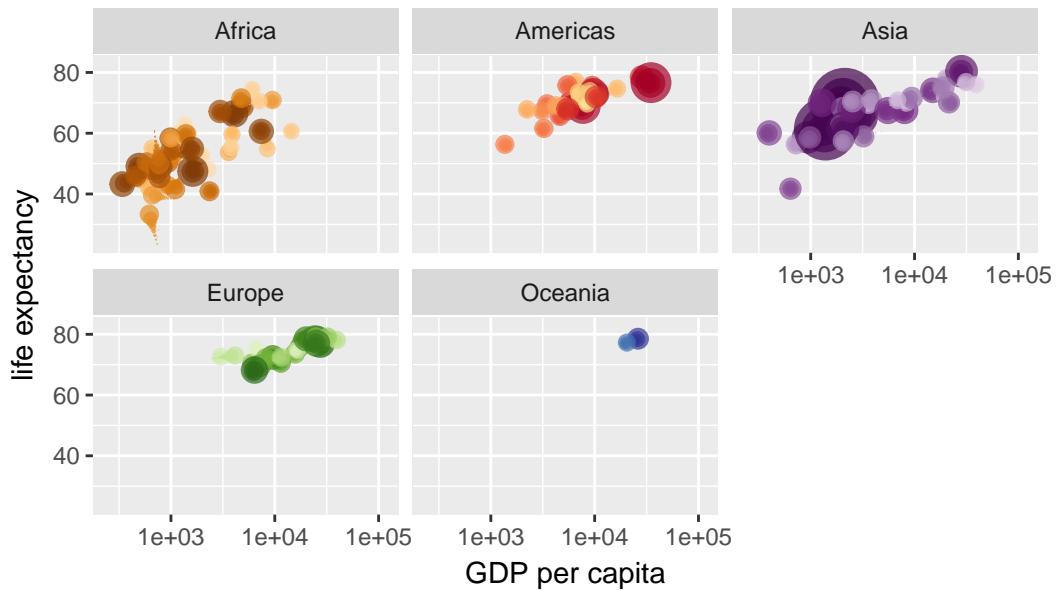
Year: 1995



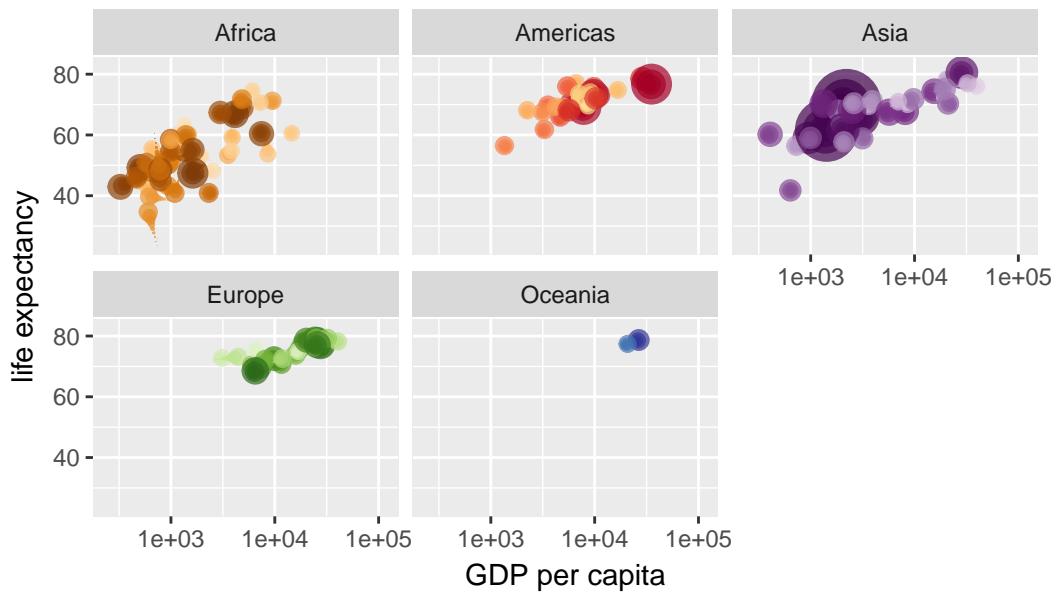
Year: 1995



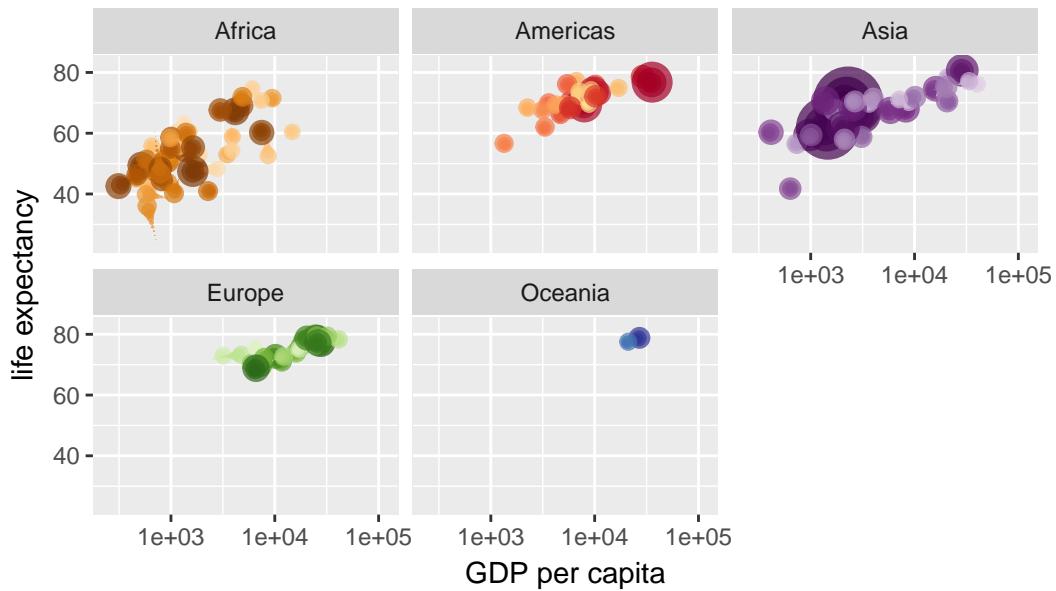
Year: 1996



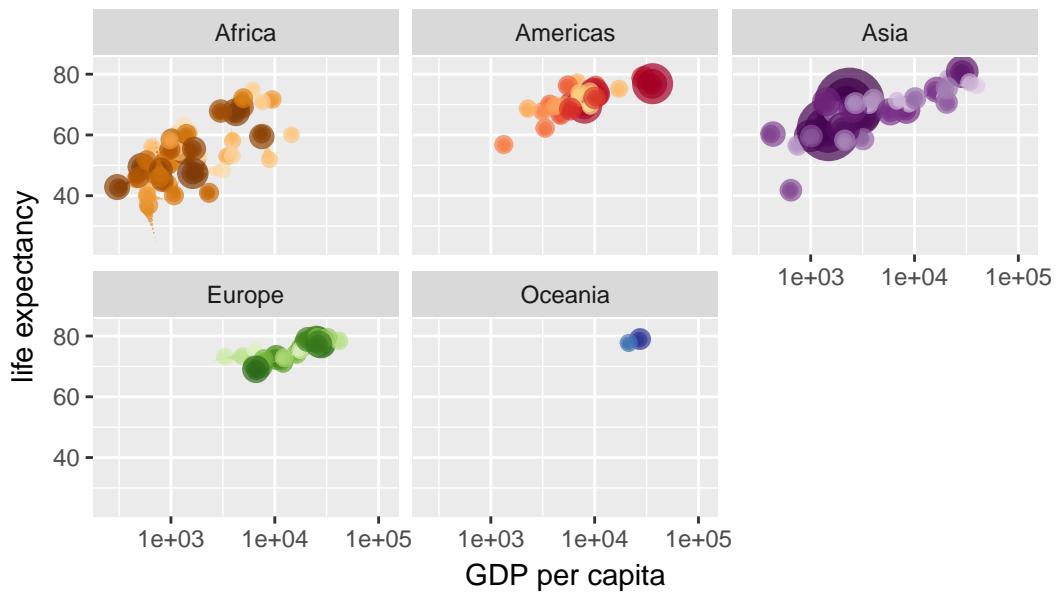
Year: 1996



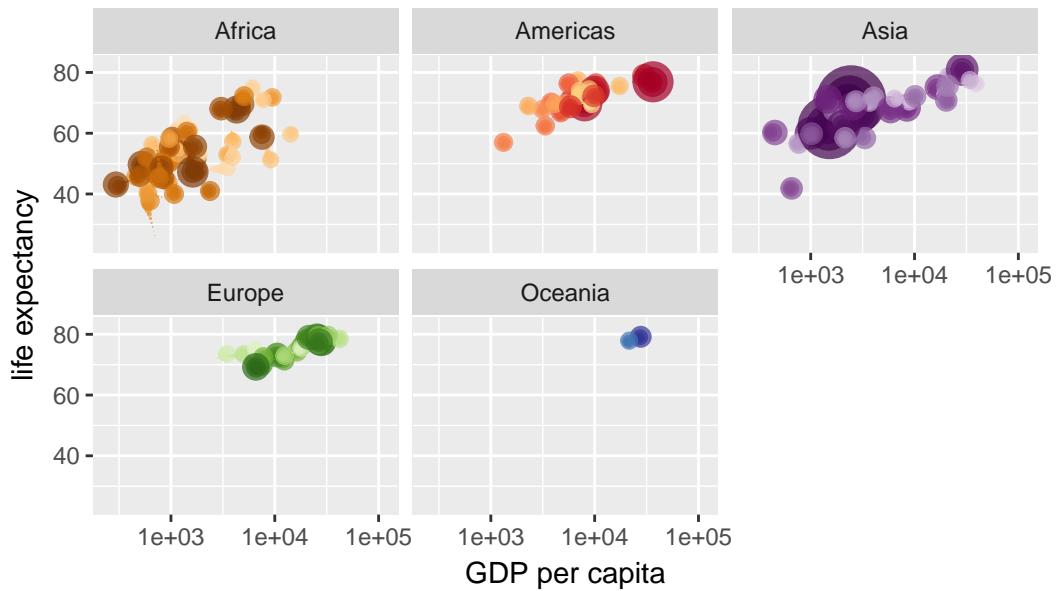
Year: 1997



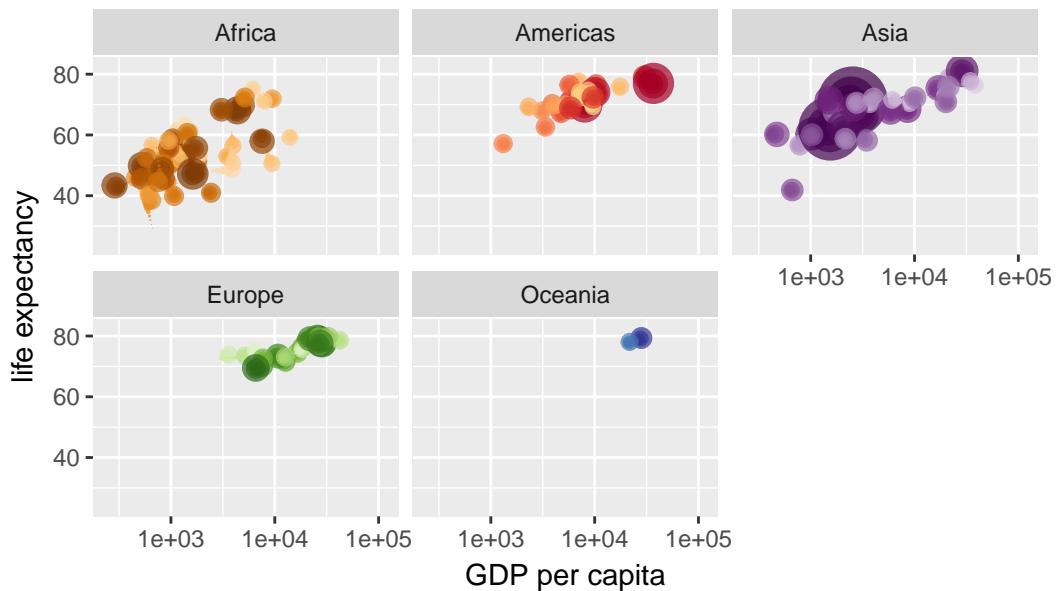
Year: 1998



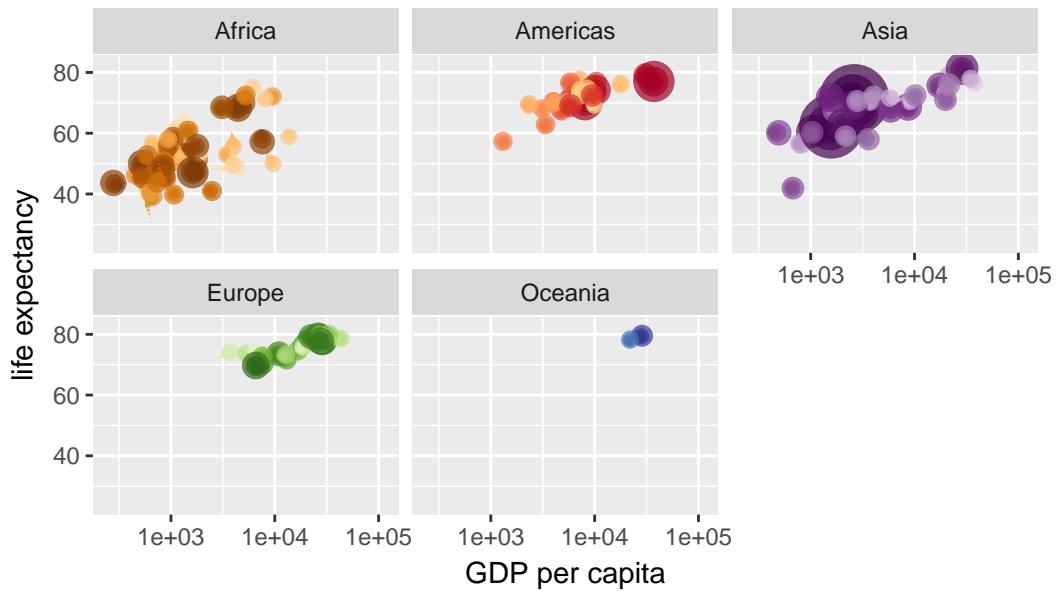
Year: 1998



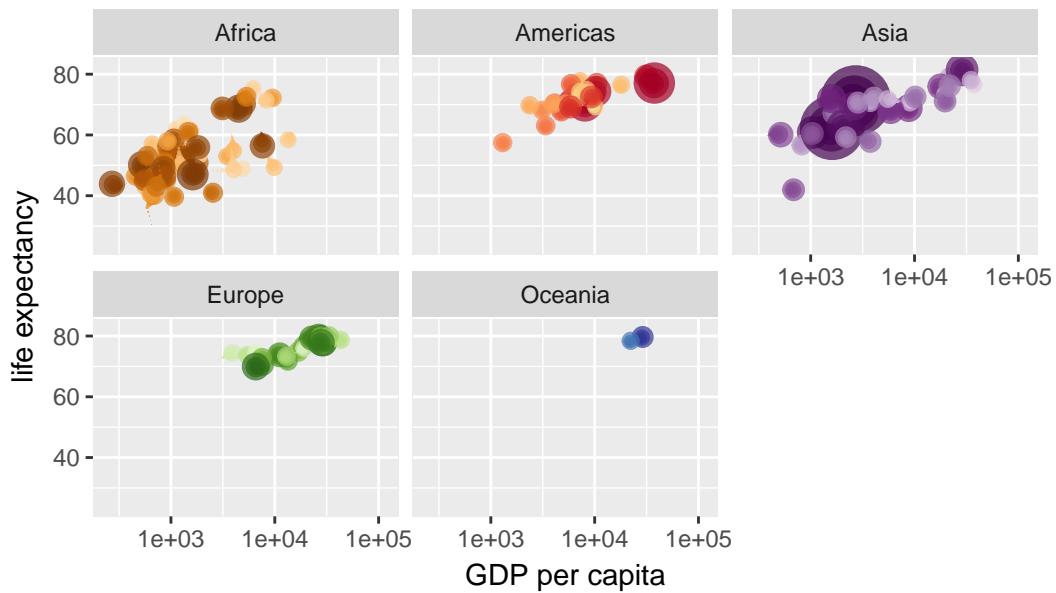
Year: 1999



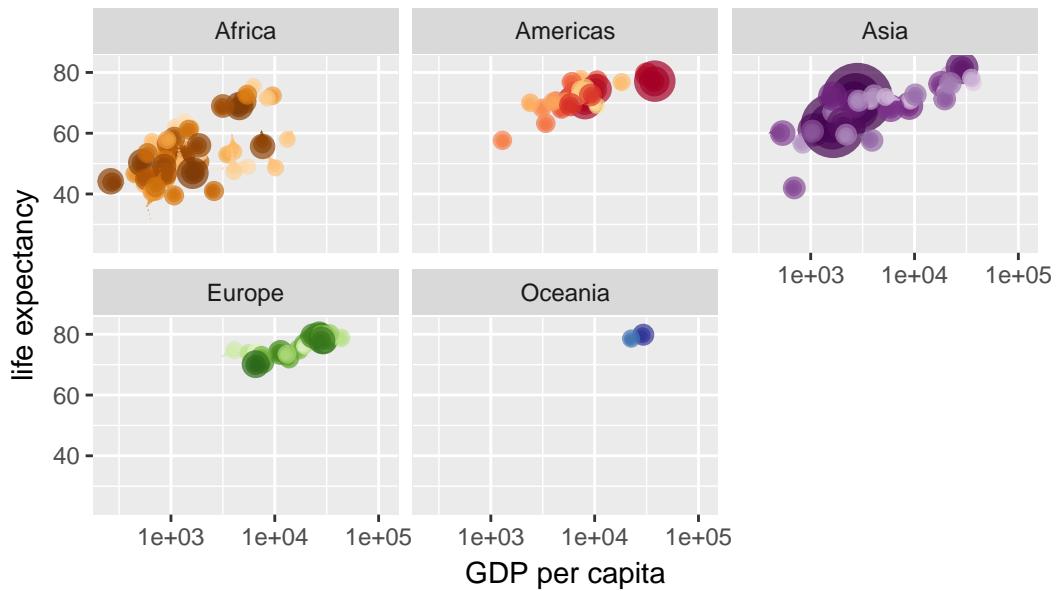
Year: 1999



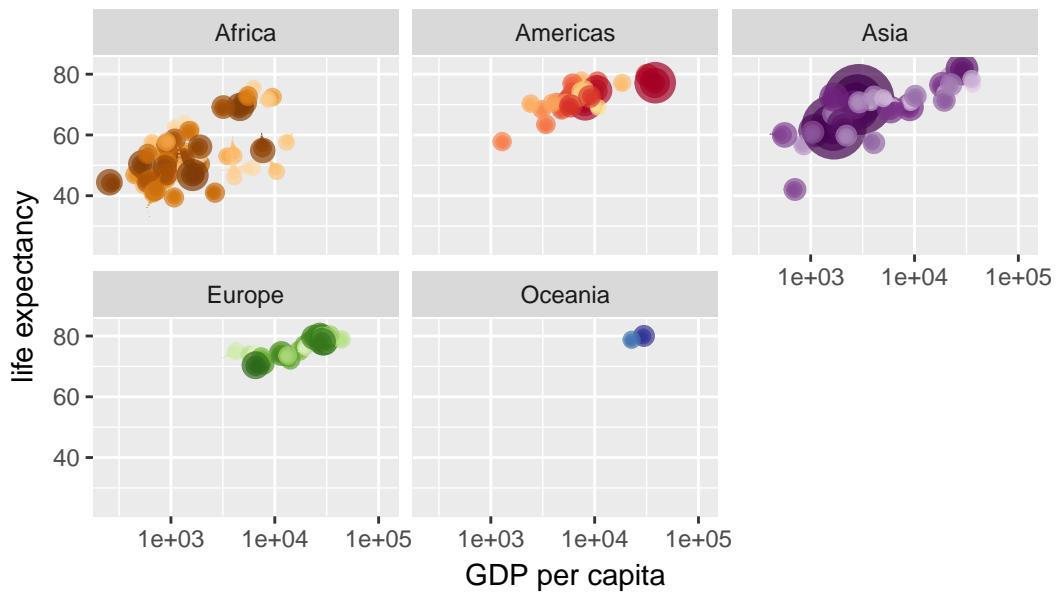
Year: 2000



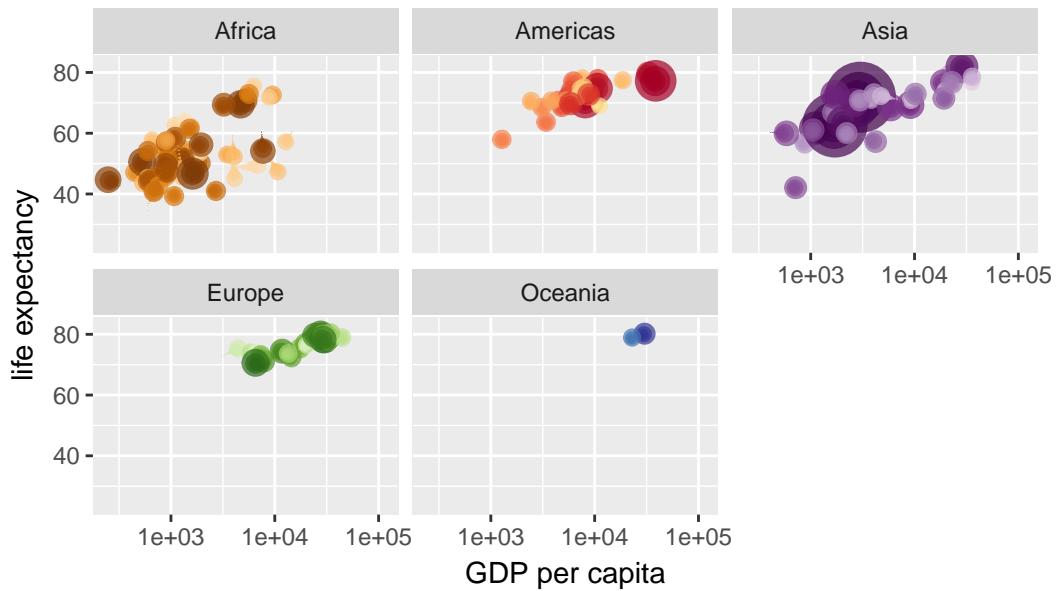
Year: 2000



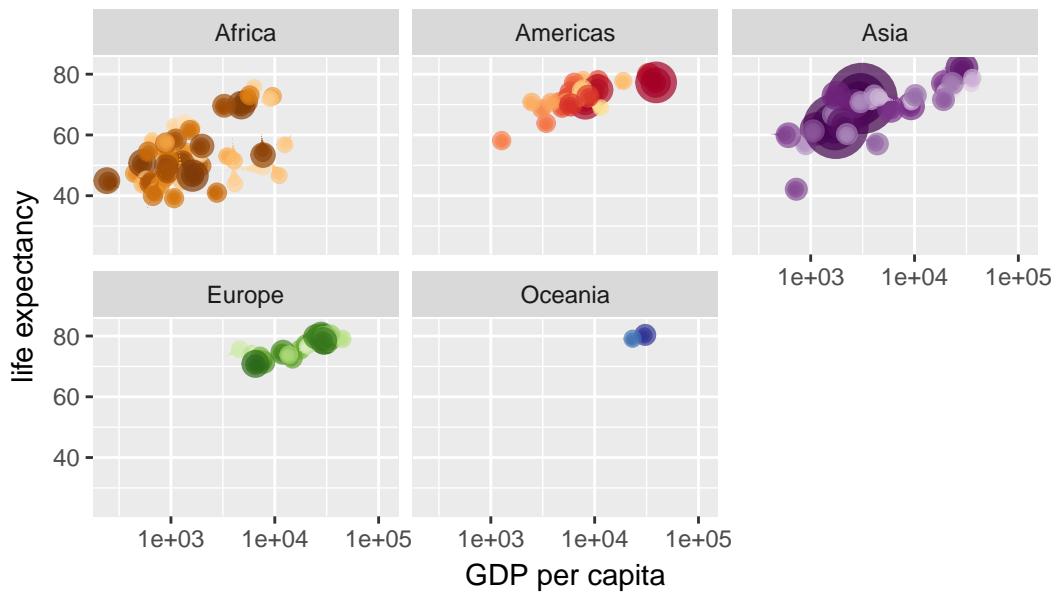
Year: 2001



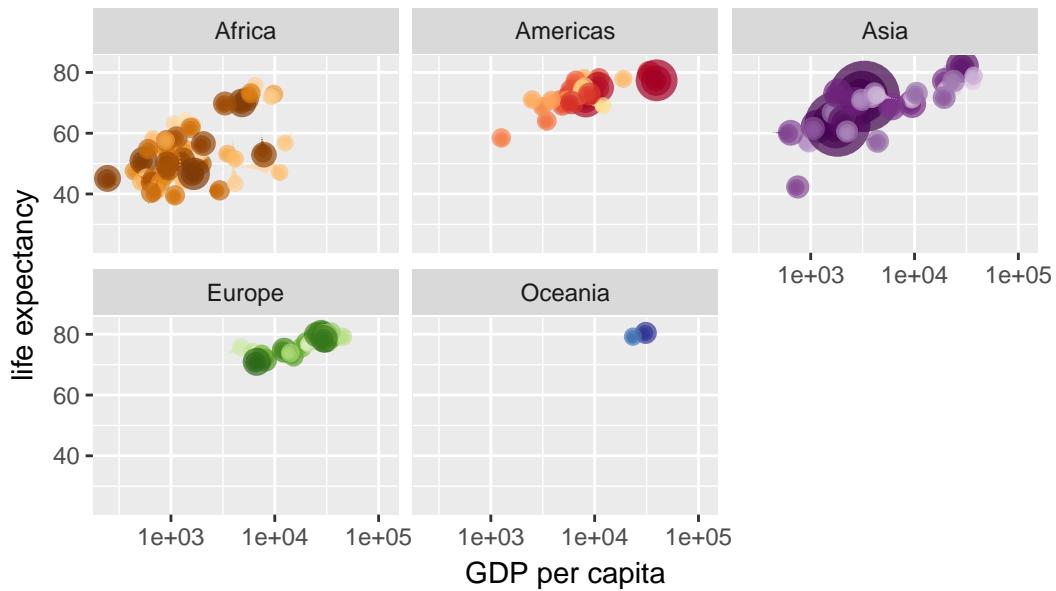
Year: 2001



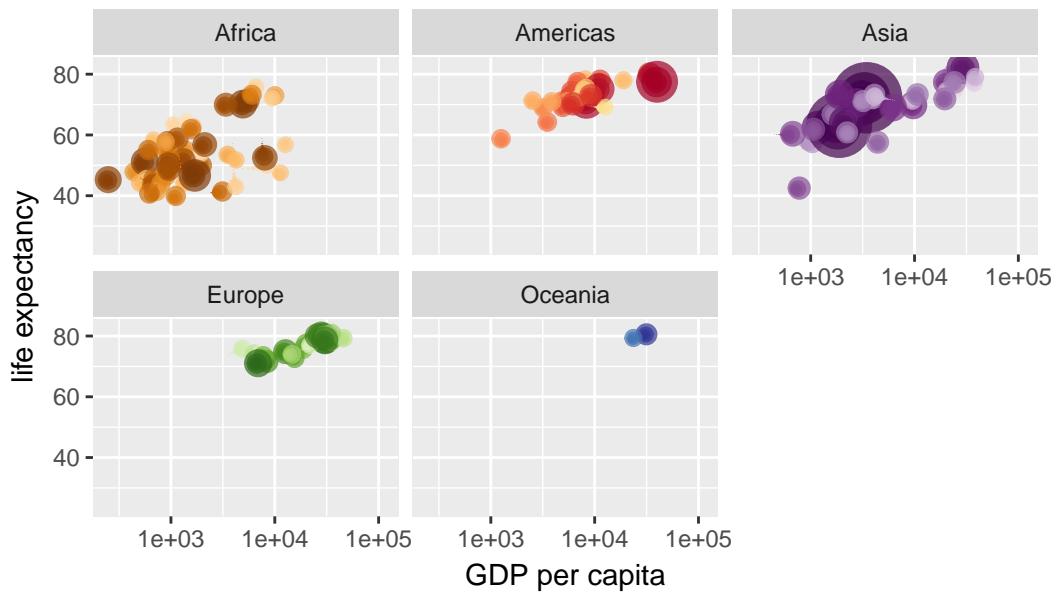
Year: 2002



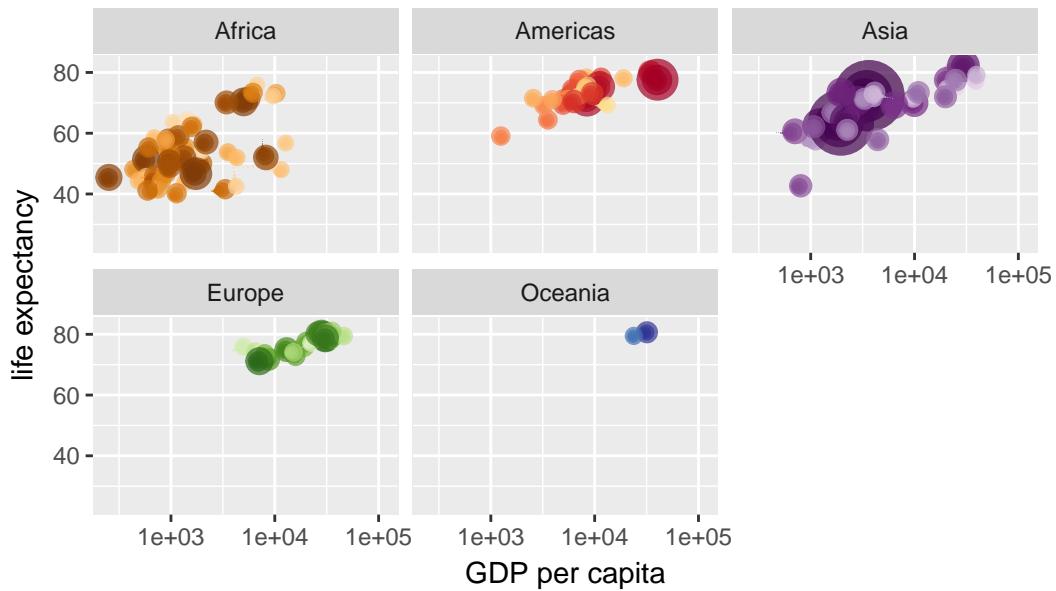
Year: 2003



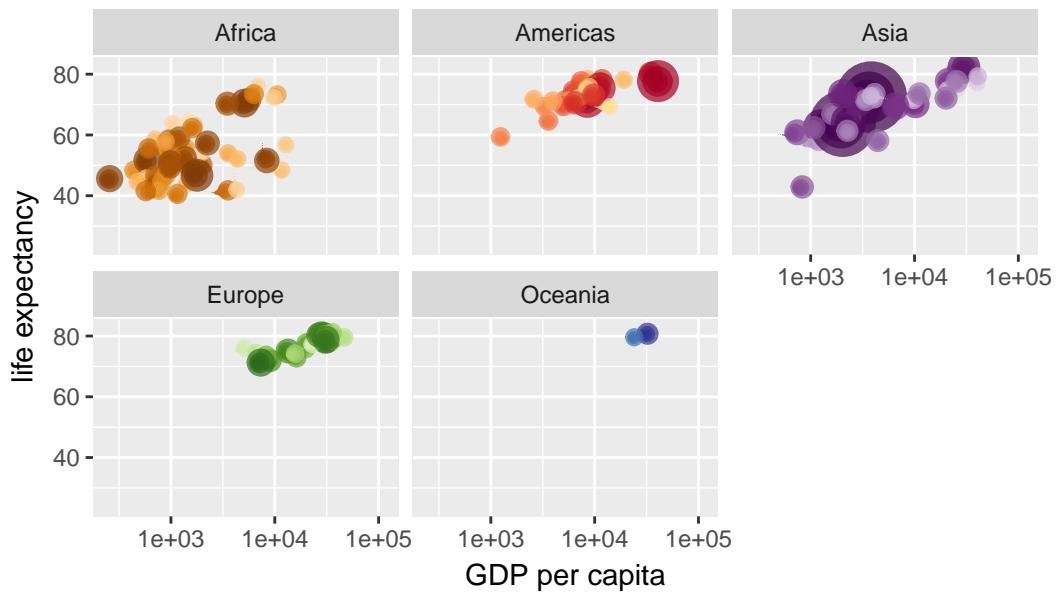
Year: 2003



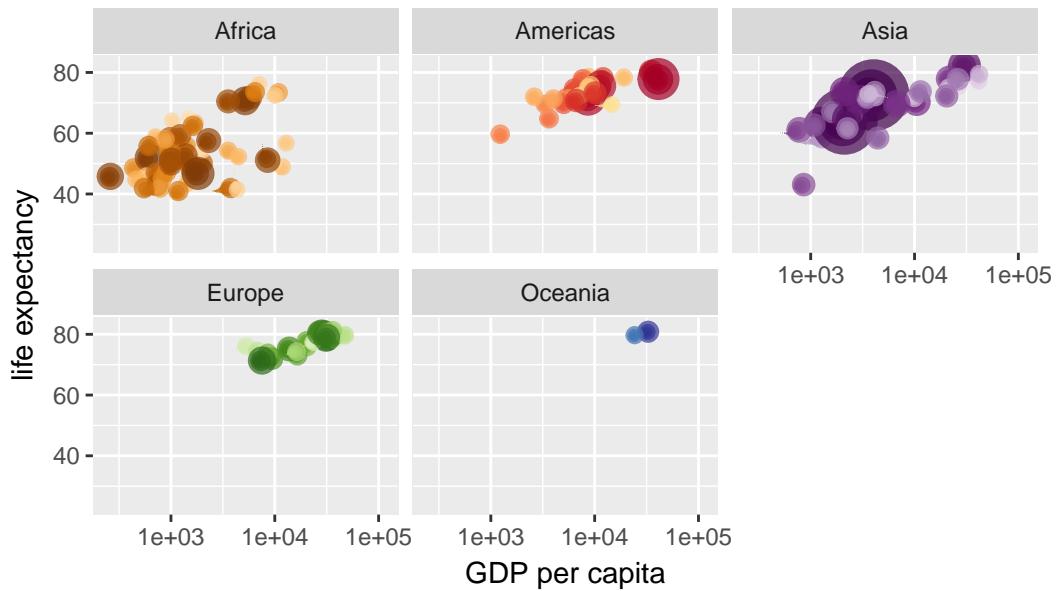
Year: 2004



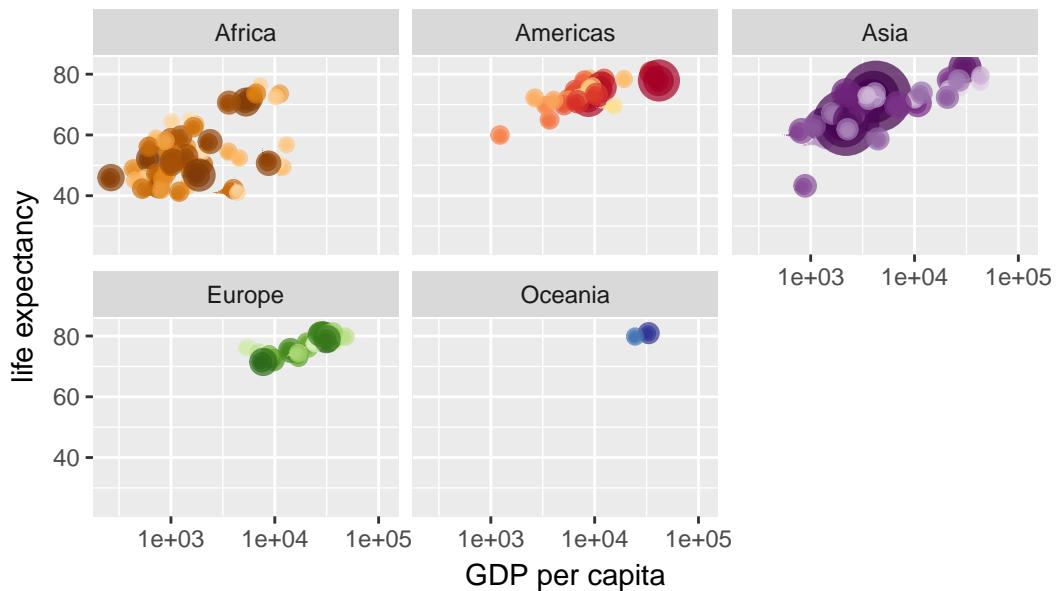
Year: 2004



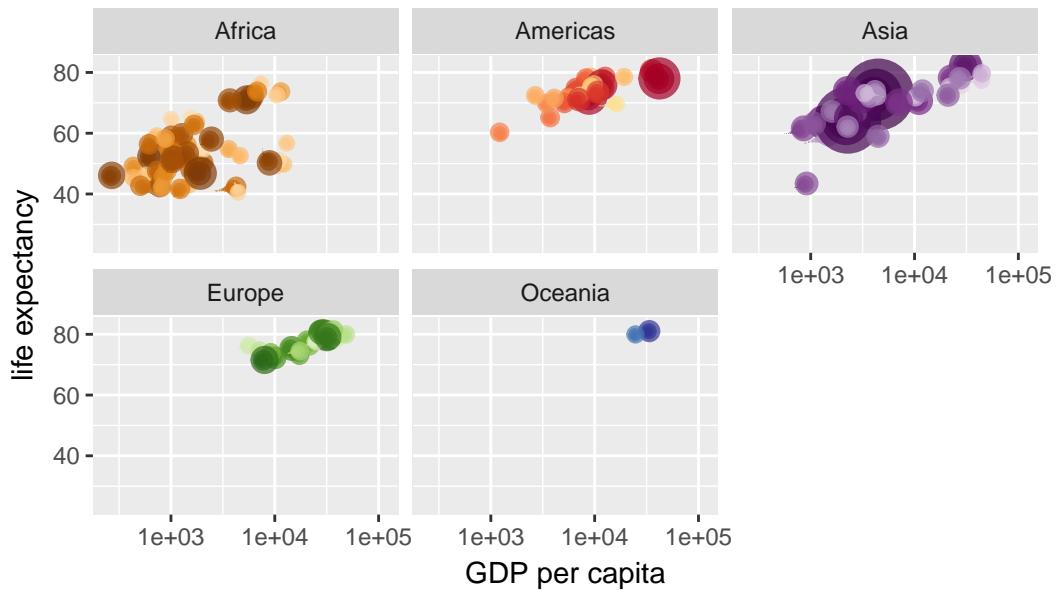
Year: 2005



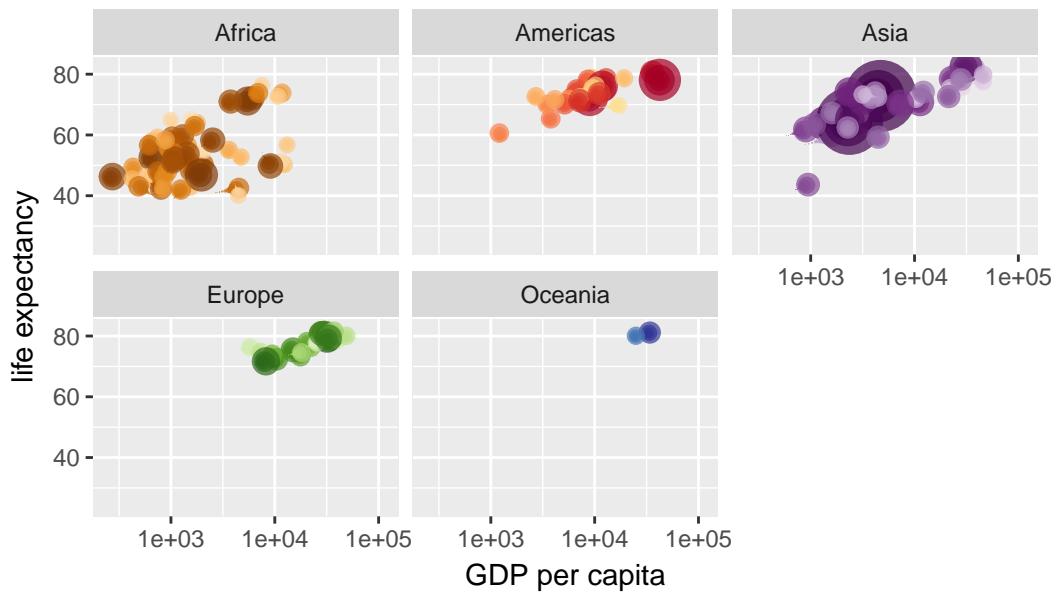
Year: 2005



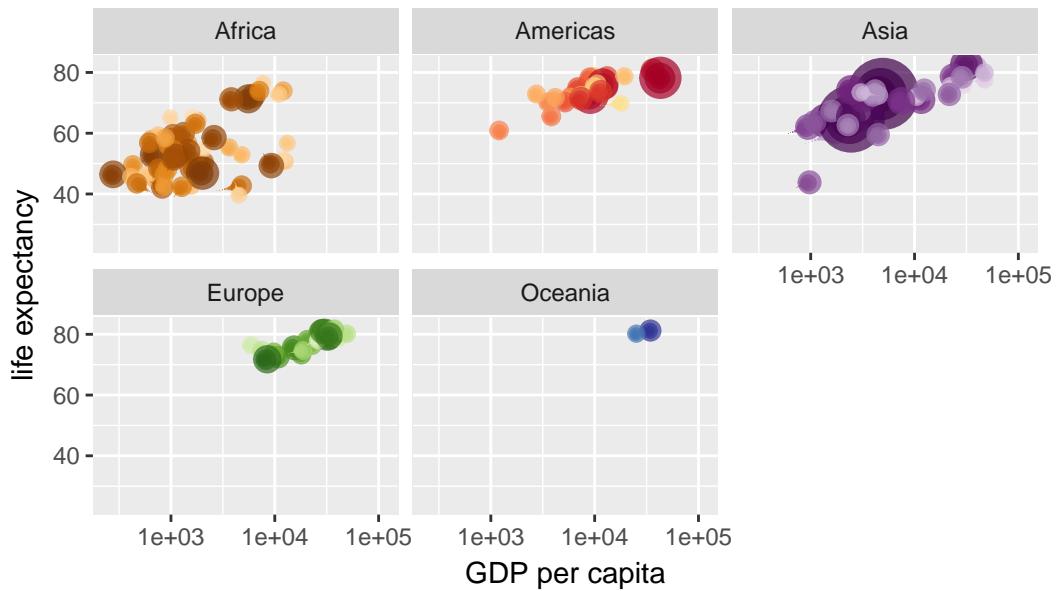
Year: 2006



Year: 2006



Year: 2007



##Combining plots Install patchwork package

```
#install.packages("patchwork")
library(patchwork)

# Setup some example plots
p1 <- ggplot(mtcars) + geom_point(aes(mpg, disp))
p2 <- ggplot(mtcars) + geom_boxplot(aes(gear, disp, group = gear))
p3 <- ggplot(mtcars) + geom_smooth(aes(disp, qsec))
p4 <- ggplot(mtcars) + geom_bar(aes(carb))

# Use patchwork to combine them here:
(p1 | p2 | p3) /
  p4

`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

