

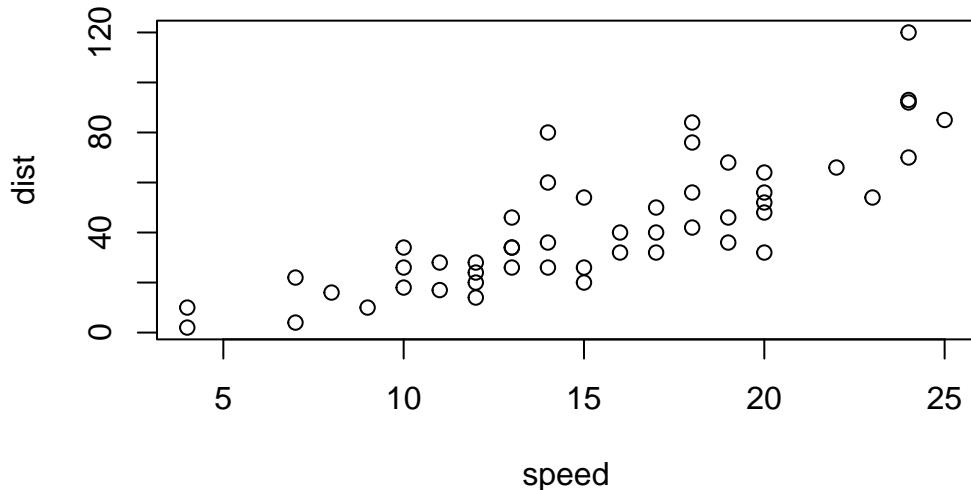
Class 5: Data Viz with ggplot

Juan Gonzalez (PID: A69036681)

R can make figures and graphs in particular.

One that comes with R out of the box is called “**base**” R - the `plot()` function.

```
plot(cars)
```



A very popular package in this area is called **ggplot2**

Before I can use any add-on package like this, I must install the package with the `install.packages("ggplot2")` command/function

Install the package in the console to avoid downloading the package every time you render your page.

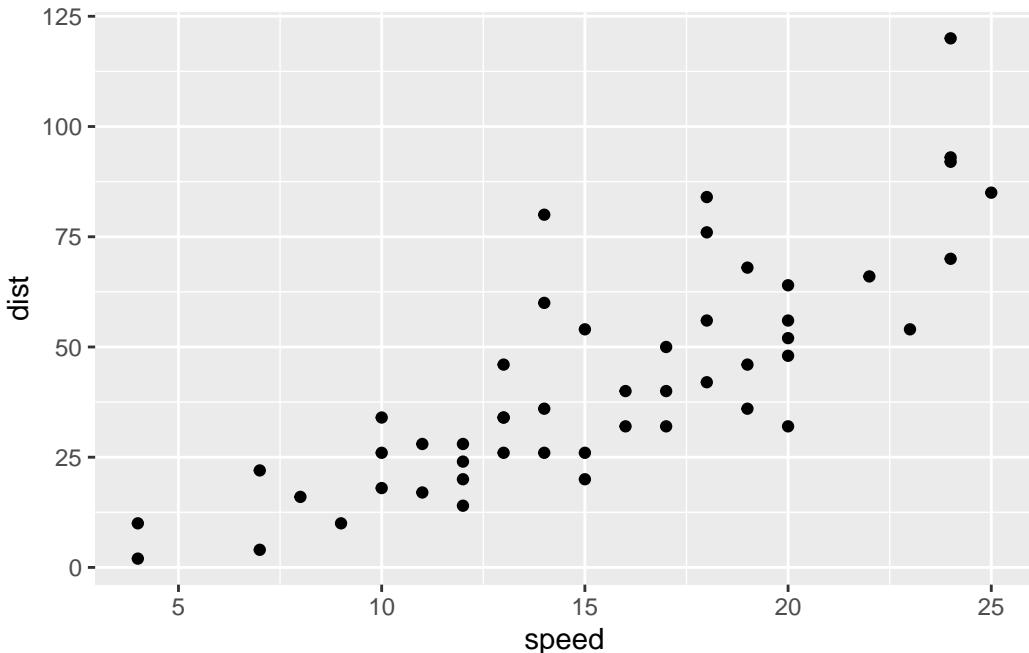
Whenever you need a package, you have to call it from a `library(ggplot2)`

Think of it like you don't want R to download everything on your computer when you start a new project. It'll be really slow. To make it efficient, you have to call the package you want to work with.

Layers of ggplot: 1) data (data.frame you want to plot); 2) aesthetic (graph mechanics); 3) geometry (types of graphs)

To make the plot, you have to call your data with 1) `ggplot()` 2) aesthetic: by adding `+` next to the and add a subsection code of `aes(x=, y=)`. 3) Geometry: by add another `+` and new subsection code `geom_point()`.

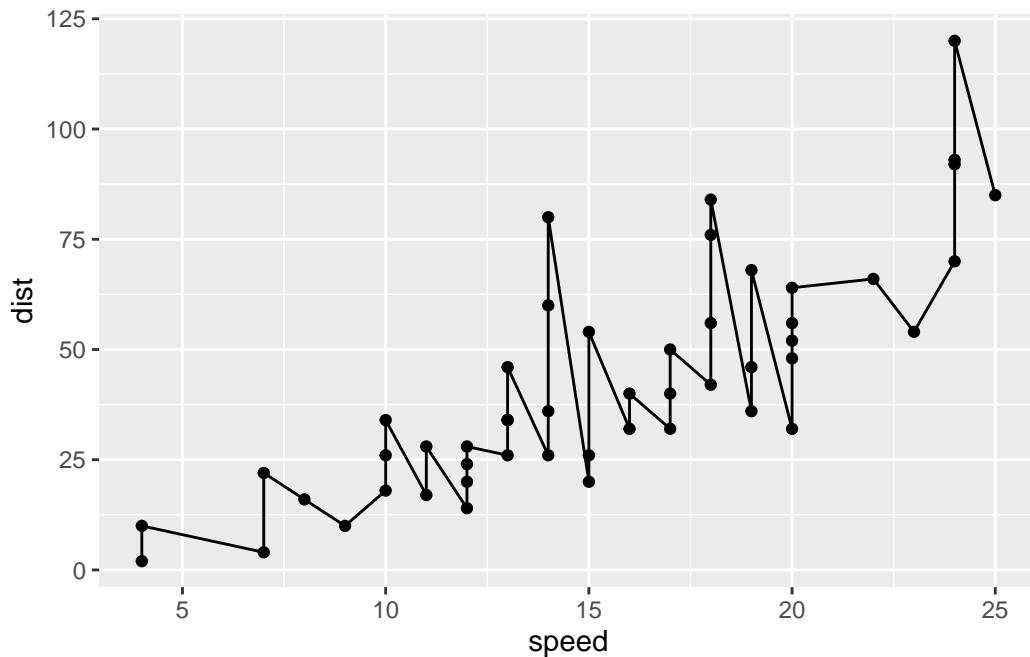
```
library(ggplot2)
ggplot(cars) +
  aes(x=speed, y=dist) +
  geom_point()
```



For “simple” plots like this one, base R code will be much shorter than ggplot code.

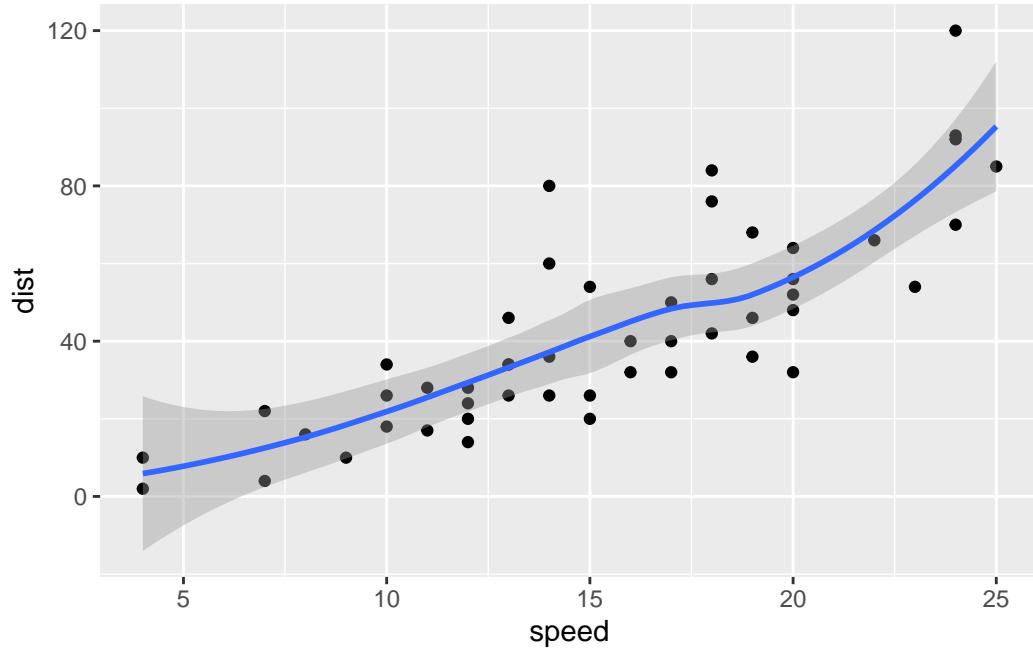
Let's fit a model and show it on my plot. 3) line: add `+`, then `geom_line`. You should see by now that ggplot works with layers.

```
library(ggplot2)
ggplot(cars) +
  aes(x=speed, y=dist) +
  geom_point() +
  geom_line()
```



But this is good for a time-series data set, but this isn't that. We want a smooth correlation line. Instead of `geom_line`, use `geom_smooth`

```
library(ggplot2)
ggplot(cars) +
  aes(x=speed, y=dist) +
  geom_point() +
  geom_smooth()  
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

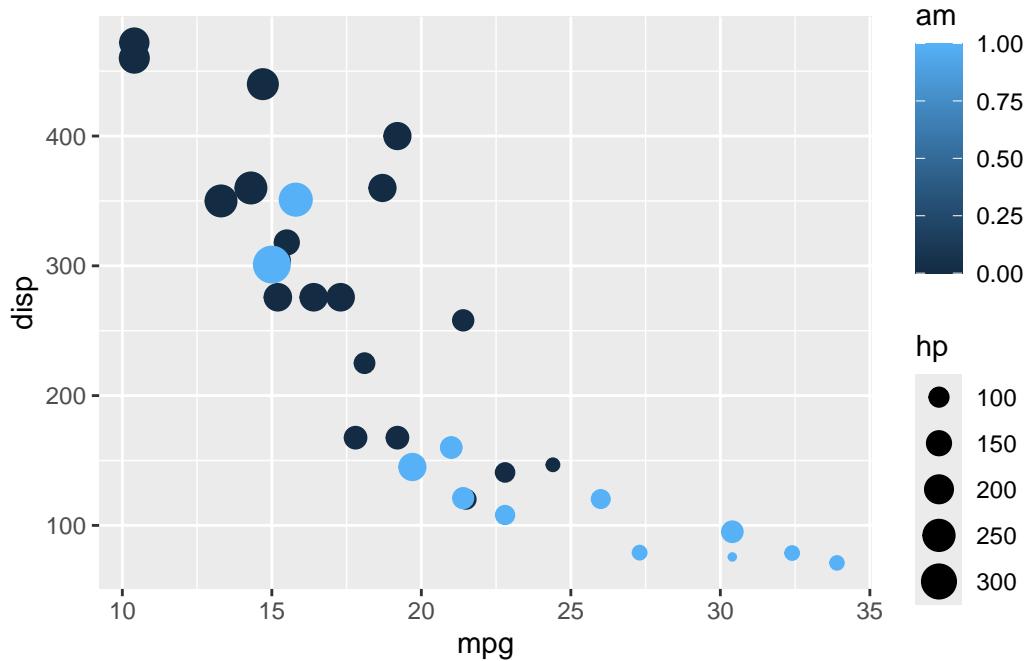


You can also use `geom_col()` for a bar graph

LETS MOVE ON!

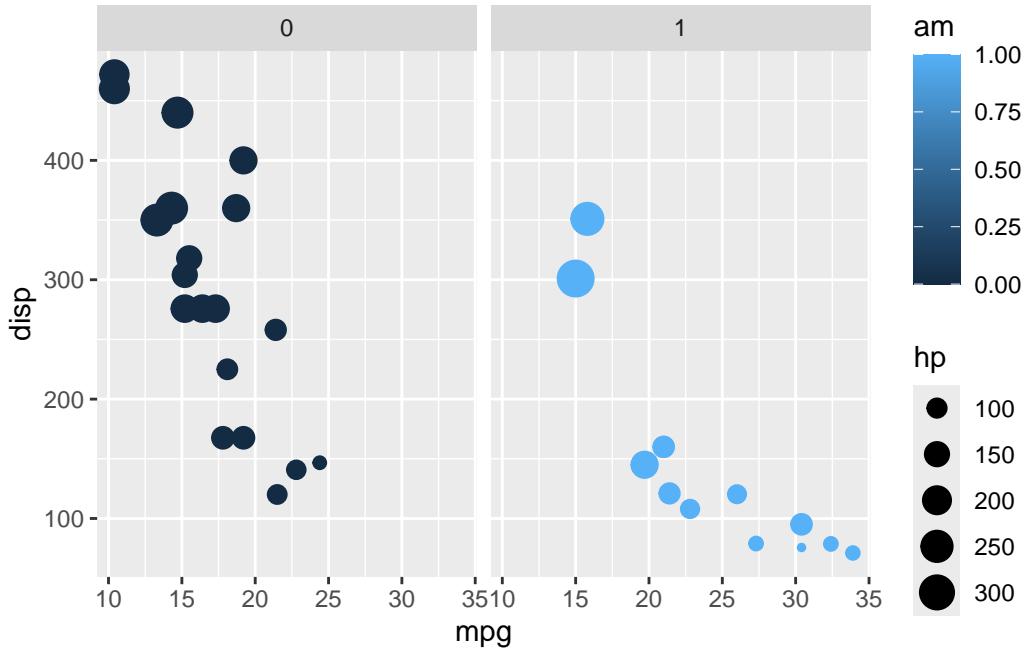
Make a ggplot of `mtcars` data set using `mpgvs.dist` AND set the size of the points to the `hp` AND set the color to `am`

```
ggplot(mtcars) +  
  aes(x=mpg, y=dist, size=hp, color=am) +  
  geom_point()
```



To separate the “ams” into two graphs, add `facet_wrap(~am)`

```
ggplot(mtcars) +  
  aes(x=mpg, y=disp, size=hp, color=am) +  
  geom_point() +  
  facet_wrap(~am)
```

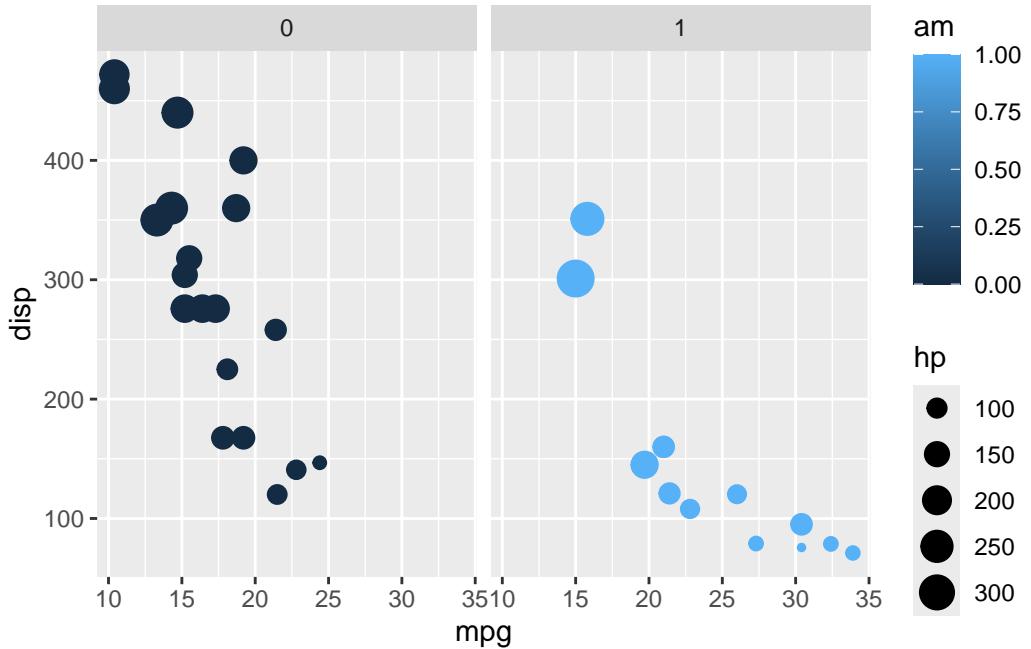


If you want labels, instead of the color distribution, add `label=rownames(mtcars)` in the `aes()` and a new subsection titled `geom_label`, but all the points will be on top of each other.

`library(ggrepel)`, will alter the overlap, which is great for graphing gene expression. The new layer code is `geom_text_repel()`

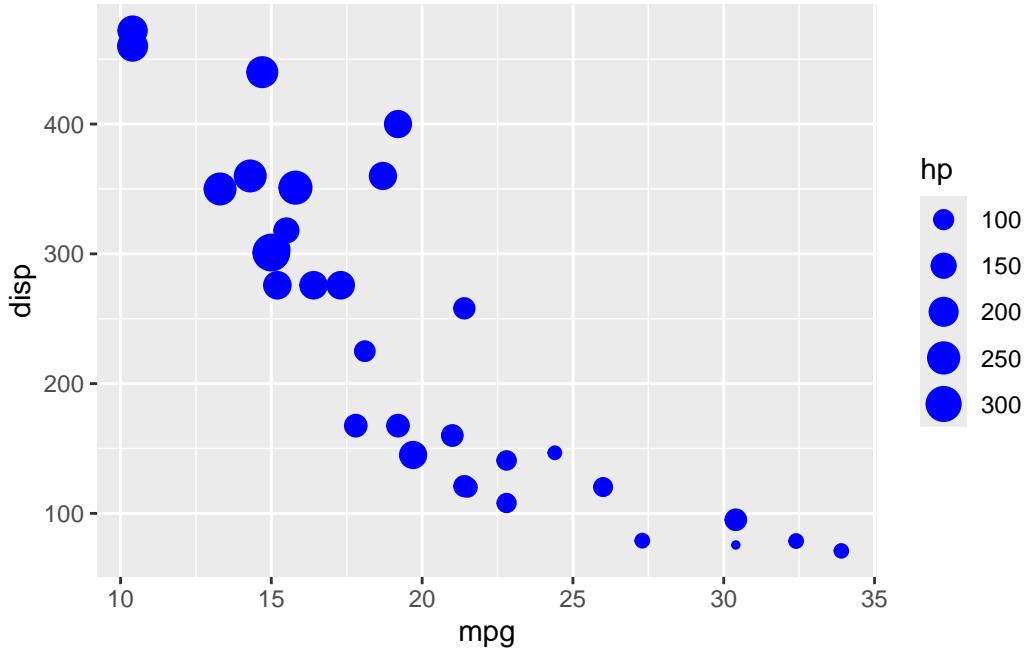
```
library (ggrepel)

ggplot(mtcars) +
  aes(x=mpg, y=disp, size=hp, col=am, label=rownames(mtcars))+
  geom_point()+
  facet_wrap(~am)
```



Now color all points “blue”. You have to put it in `geom_point` because the color blue is not a data-set, it’s just a customization. This is a graph customization feature, so you put it in `geom_point`

```
ggplot(mtcars) +  
  aes(x=mpg, y=disp, size=hp) +  
  geom_point(color="blue")
```



Now lets change data-sets. Lets look at gene expression. Here is the data for this:

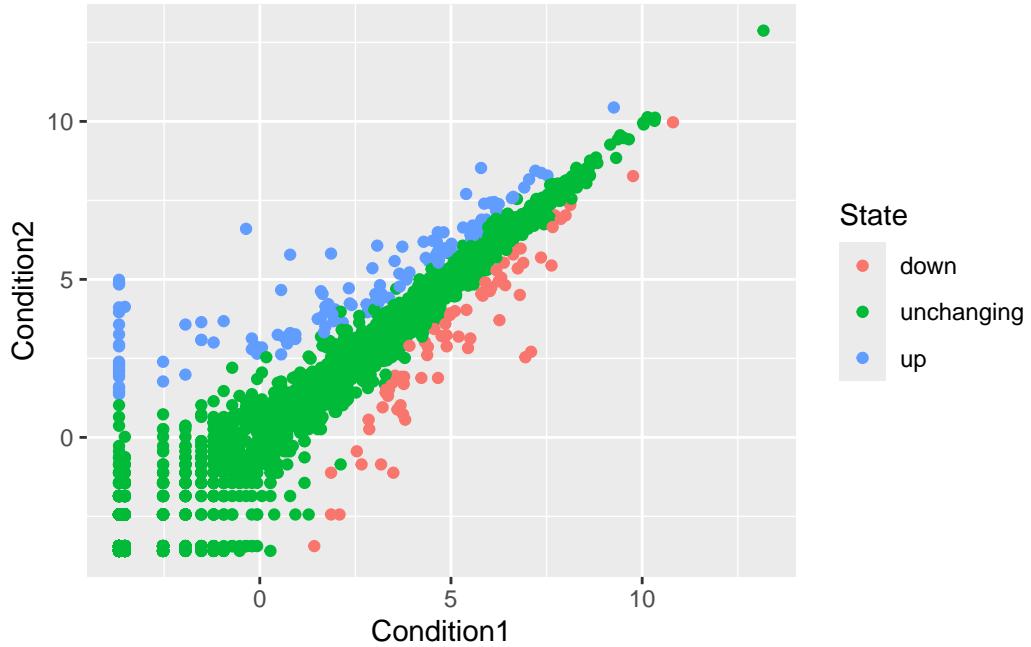
```
url <- "https://bioboot.github.io/bimm143_S20/class-material/up_down_expression.txt"
genes <- read.delim(url) head(genes)
```

Lets make a graph following the same steps as before, also adding the State column:

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

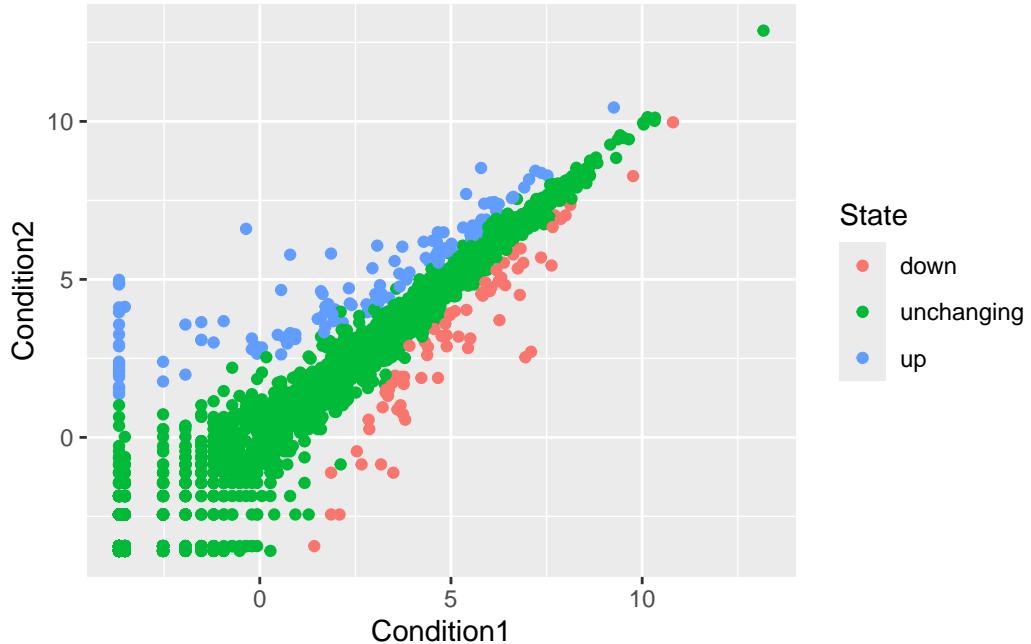
	Gene	Condition1	Condition2	State
1	A4GNT	-3.6808610	-3.4401355	unchanging
2	AAAS	4.5479580	4.3864126	unchanging
3	AASDH	3.7190695	3.4787276	unchanging
4	AATF	5.0784720	5.0151916	unchanging
5	AATK	0.4711421	0.5598642	unchanging
6	AB015752.4	-3.6808610	-3.5921390	unchanging

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



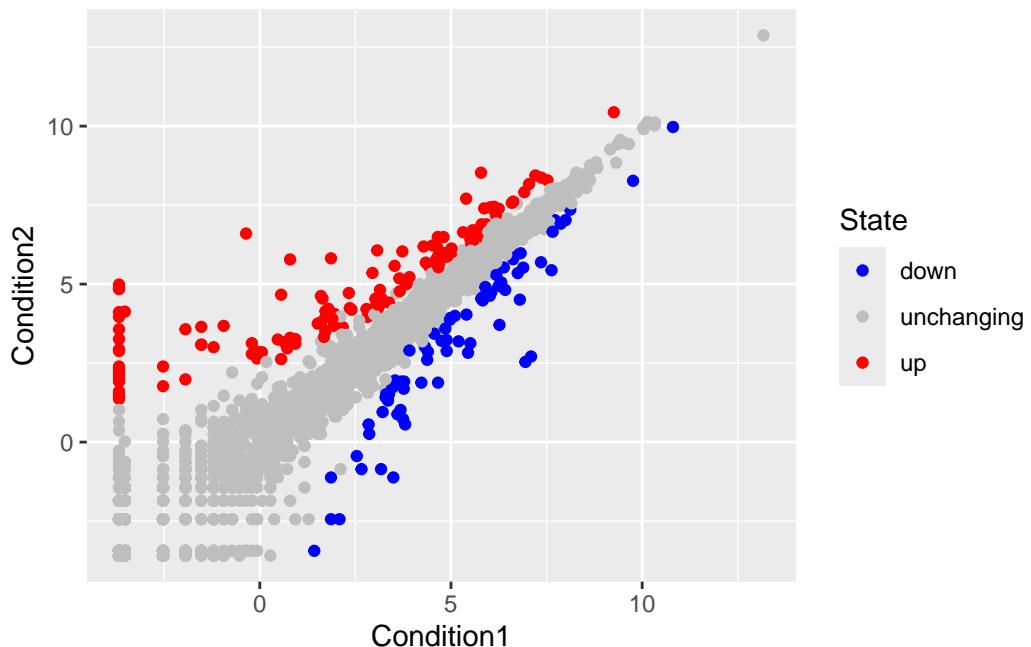
Now let's make this graph a vector to make things easier to customize from now on.

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



To change colors of the up, down, and unchanging states, lets use `scale_colour_manual(values=c("blue","gray","red"))`

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



Going Further

Now let's work with another data set. This data contains the economic and demographic data about various countries since 1952.

The data set covers many years and many countries. Let's use **dplyr** code to focus on a single year. # `install.packages("dplyr")`

```
library(gapminder)
```

```
library(dplyr)
```

Attaching package: 'dplyr'

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

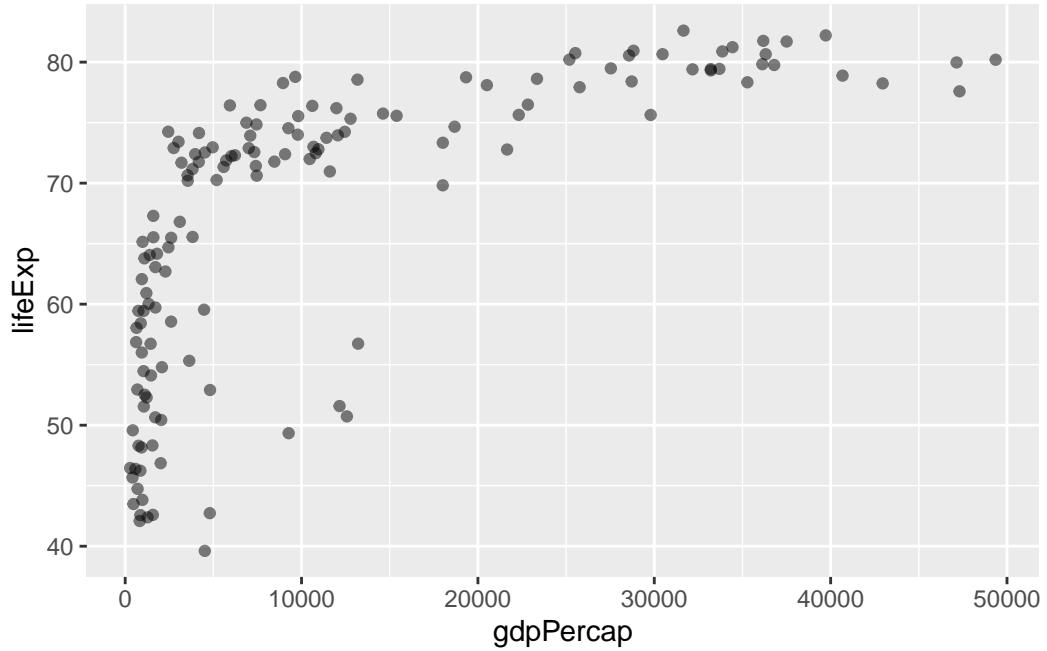
filter, lag

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

intersect, setdiff, setequal, union

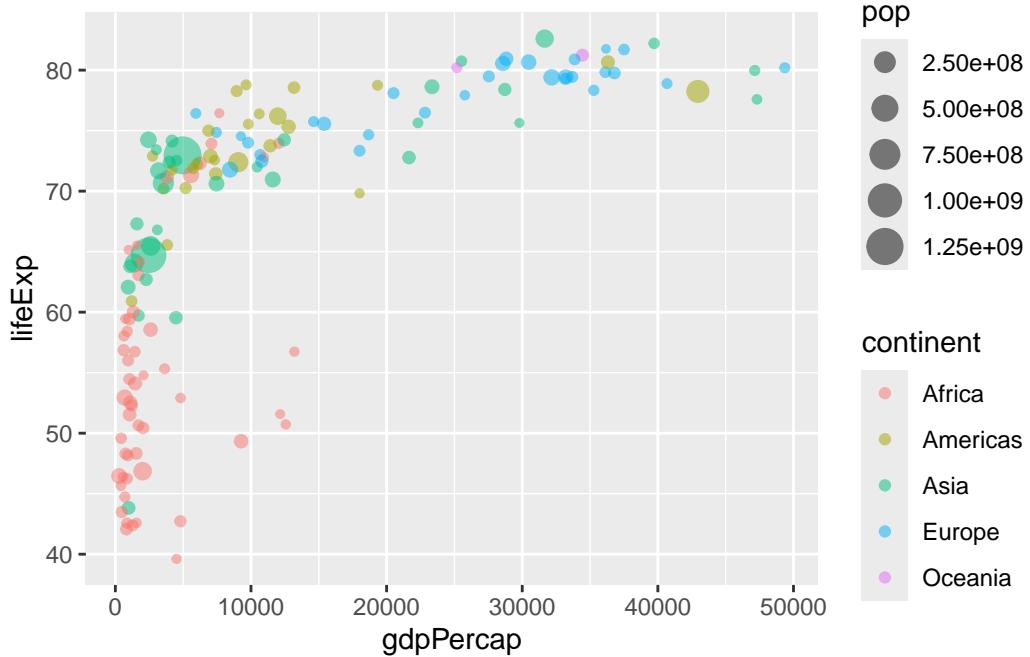
```
gapminder_2007 <- gapminder %>% filter(year==2007)

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



Now lets add the color=continent, size=population. This will obtain a much richer plot.

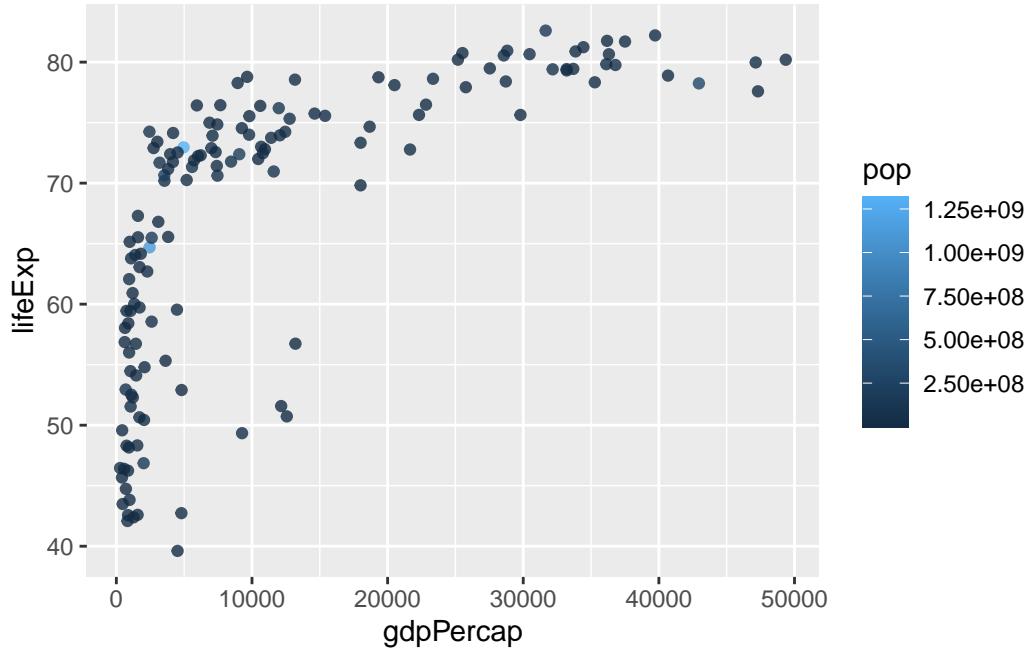
```
ggplot(gapminder_2007) +  
  aes(x=gdpPercap, y=lifeExp, color=continent, size=pop) +  
  geom_point(alpha=0.5)
```



Here each new aesthetic we add results in a new dimension to our scatter plot. We see that in the resulting plot each point is colored differently based on the continent of each country. ggplot uses the coloring scheme based on the categorical data type of the variable continent.

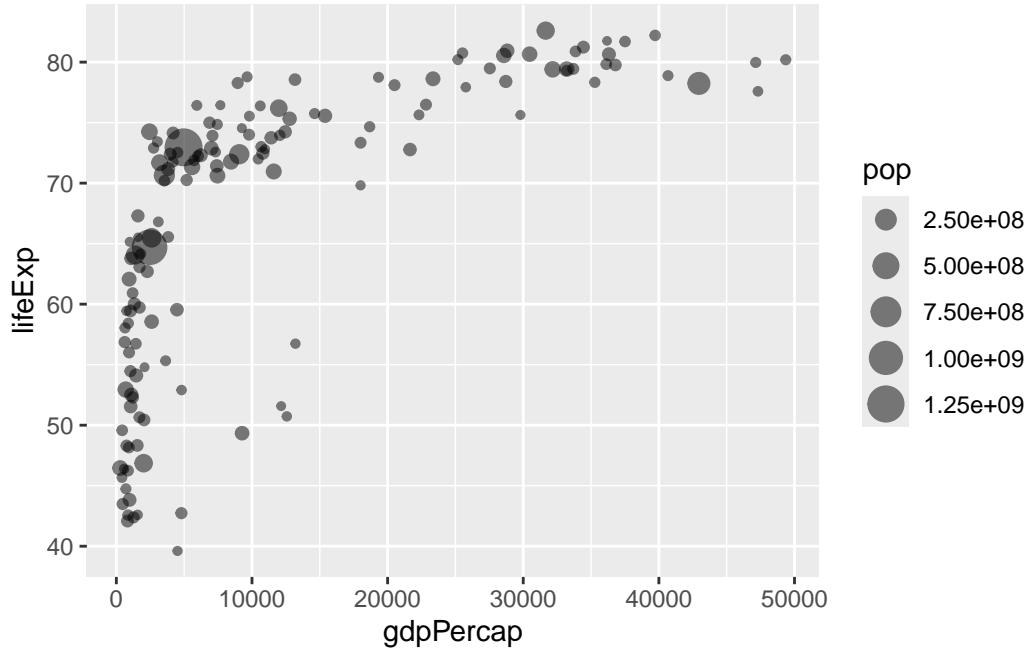
By contrast, let's see how the plot looks like if we color the points by the numeric variable population `pop`:

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



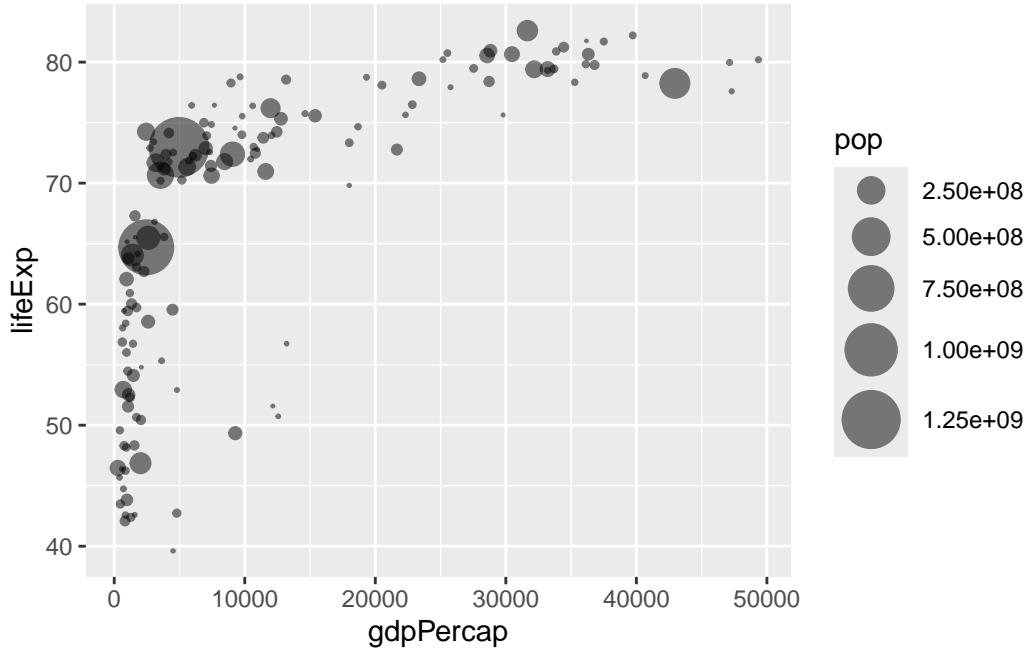
For the gapminder_2007 dataset we can plot the GDP per capita ($x=gdpPercap$) vs. the life expectancy ($y=lifeExp$) and set the point size based on the population ($size=pop$) of each country we can use:

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



However, if you look closely we see that the point sizes in the plot above do not clearly reflect the population differences in each country. If we compare the point size representing a population of 250 million people with the one displaying 750 million, we can see, that their sizes are not proportional. Instead, the point sizes are binned by default. To reflect the actual population differences by the point size we can use the `scale_size_area()` function instead. The scaling information can be added like any other ggplot object with the `+` operator:

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



In a bar graph, you can reorder by a certain visualization by `aes(x=reorder(country, -pop))` A new graph function is `coord_flip()`, which will rotate your plots to enable a clear visualization. If it's too cluttered, combine `geom_point` and `geom_segment`

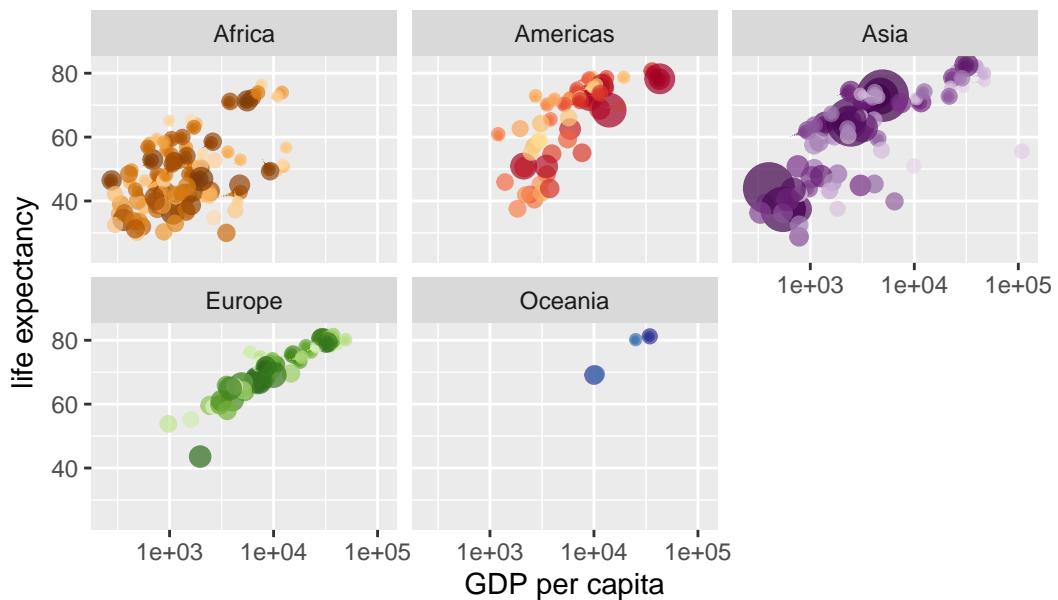
NOW LETS ANIMATE!!!!

```
library(gapminder)
library(gganimate)

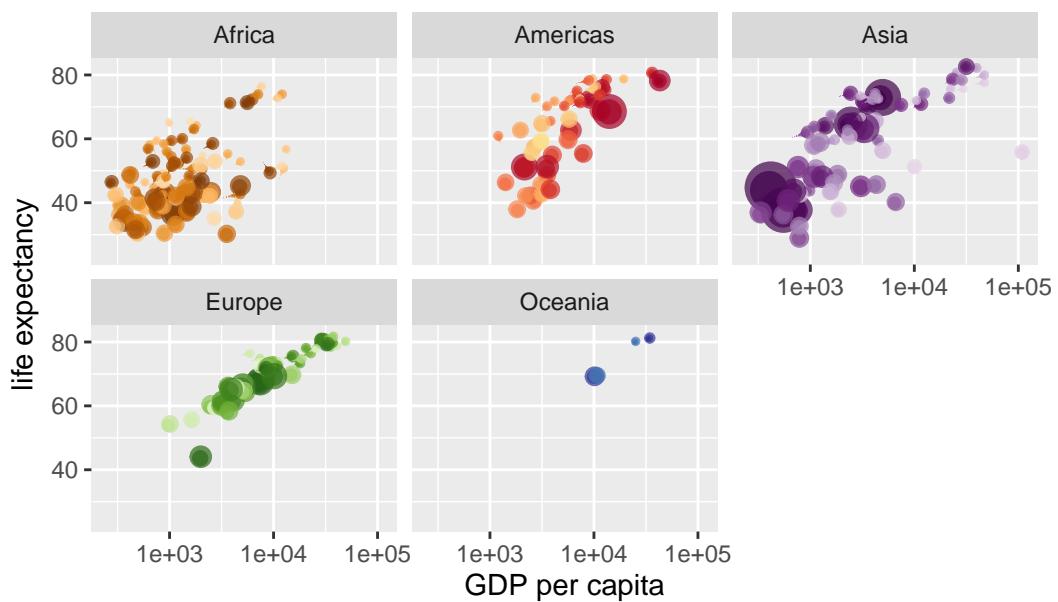
# Setup nice regular ggplot of the gapminder data
ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, colour = country)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors) +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  # Facet by continent
 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

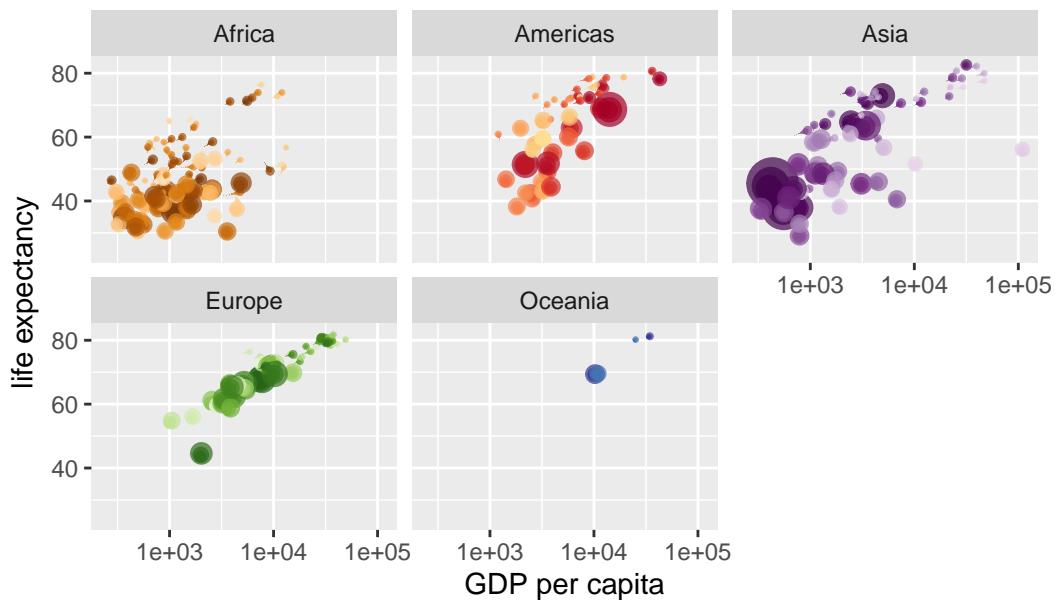
Year: 1952



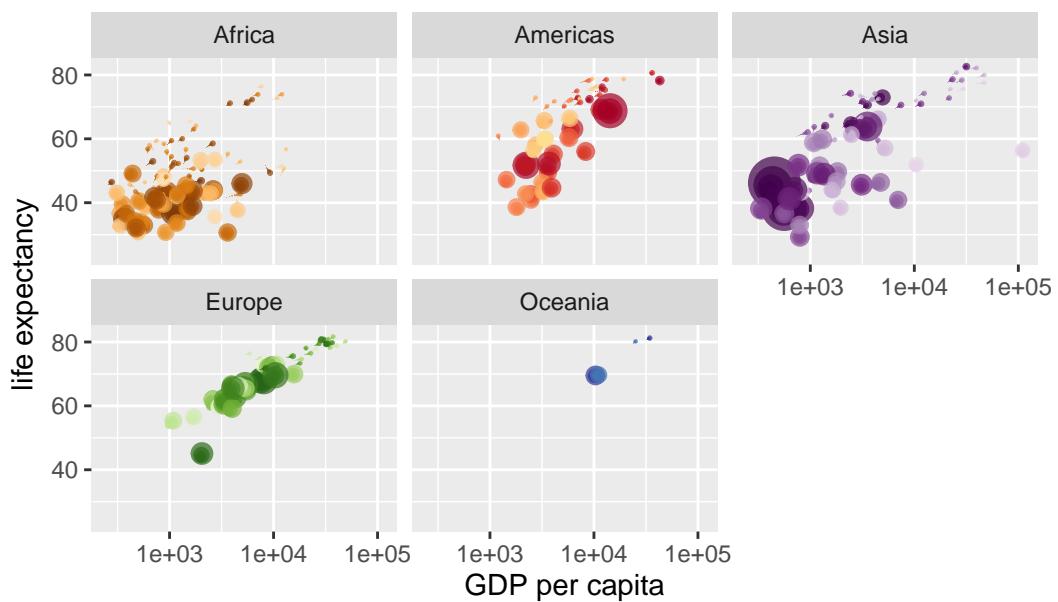
Year: 1953



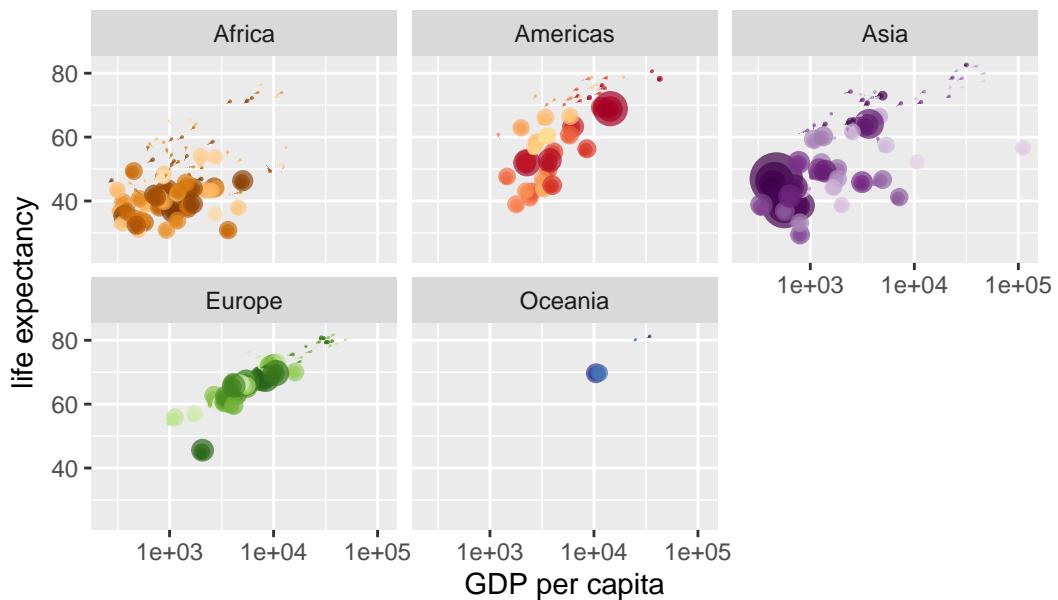
Year: 1953



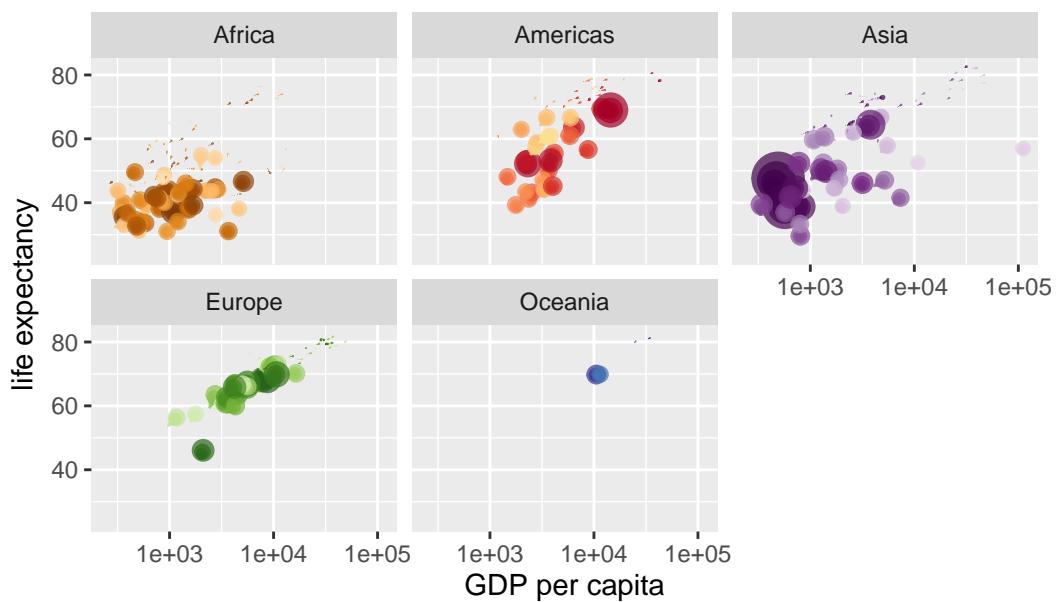
Year: 1954



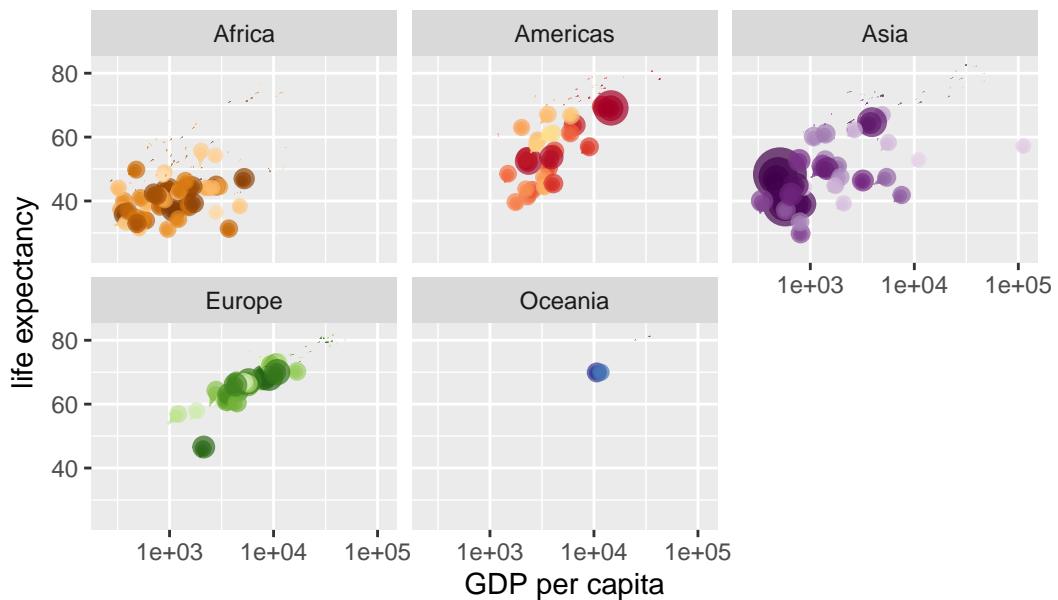
Year: 1954



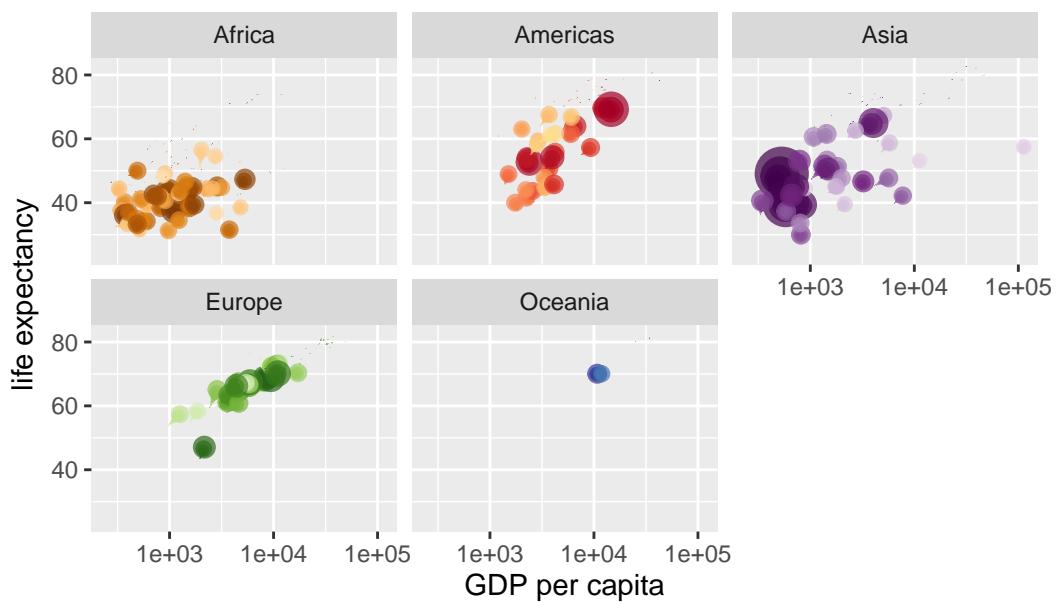
Year: 1955



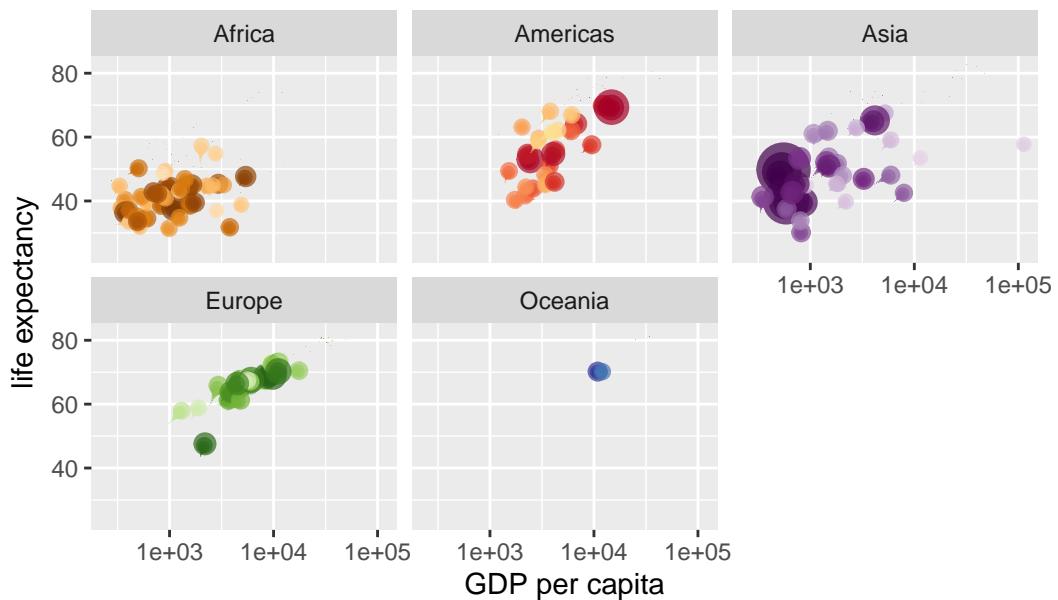
Year: 1955



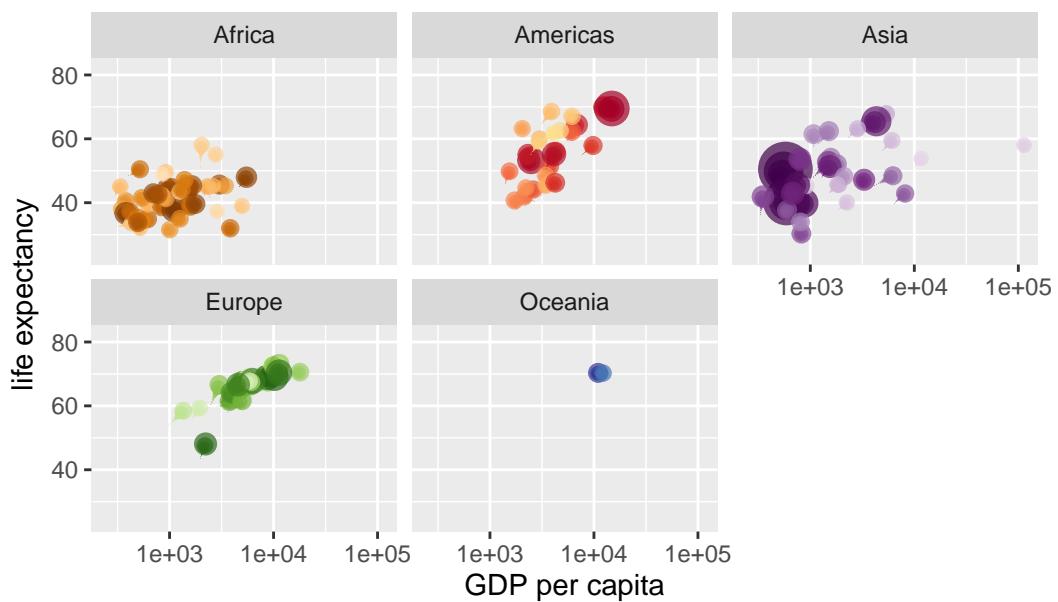
Year: 1956



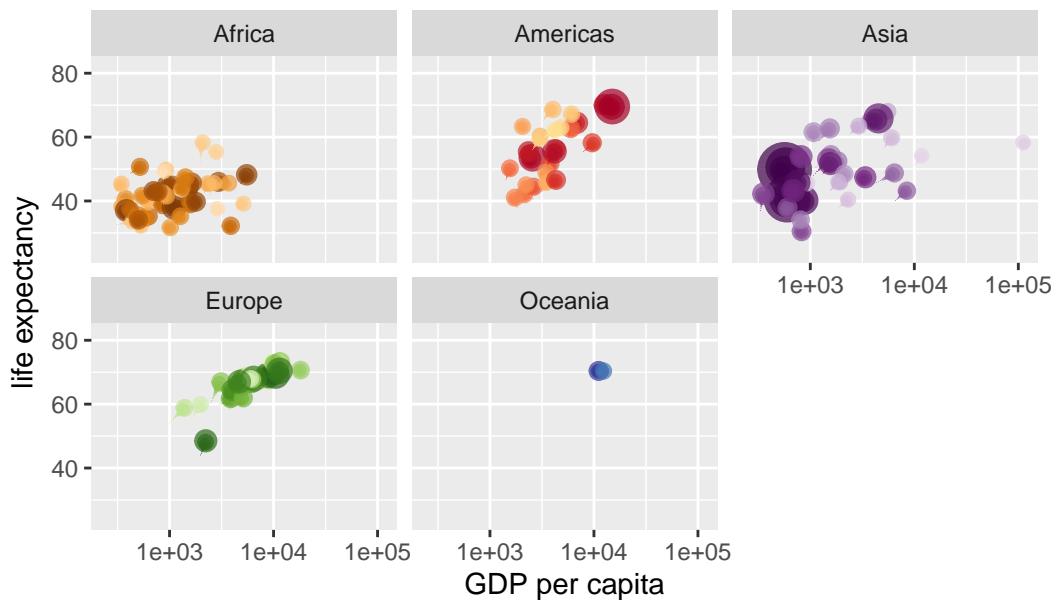
Year: 1956



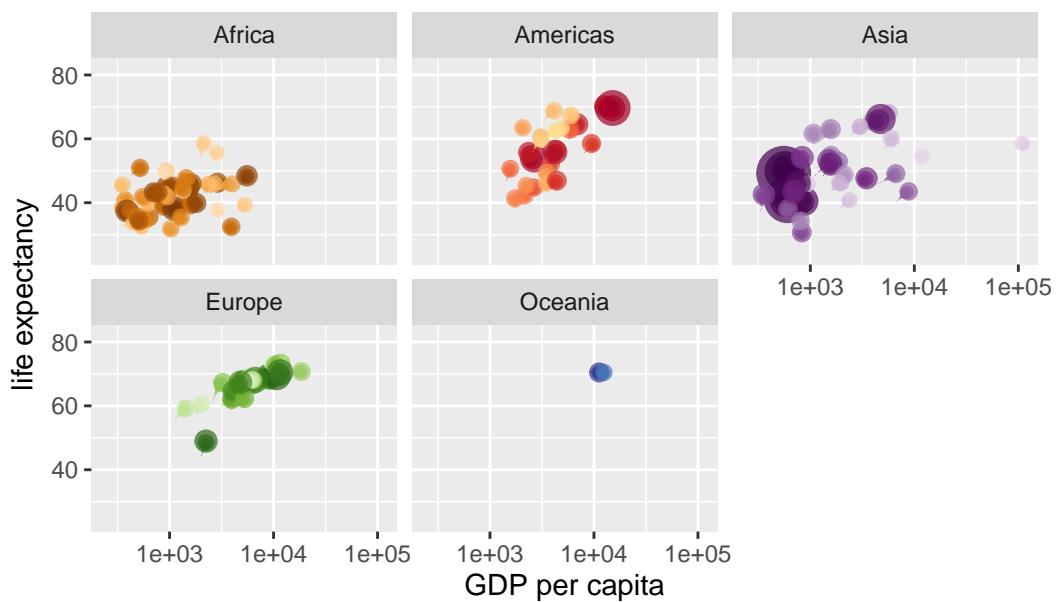
Year: 1957



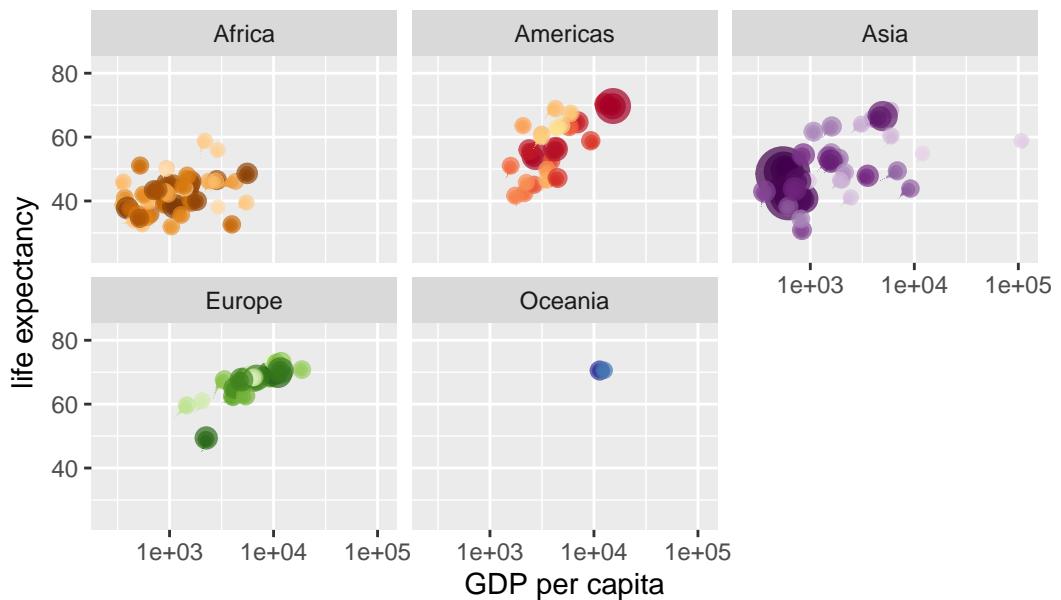
Year: 1958



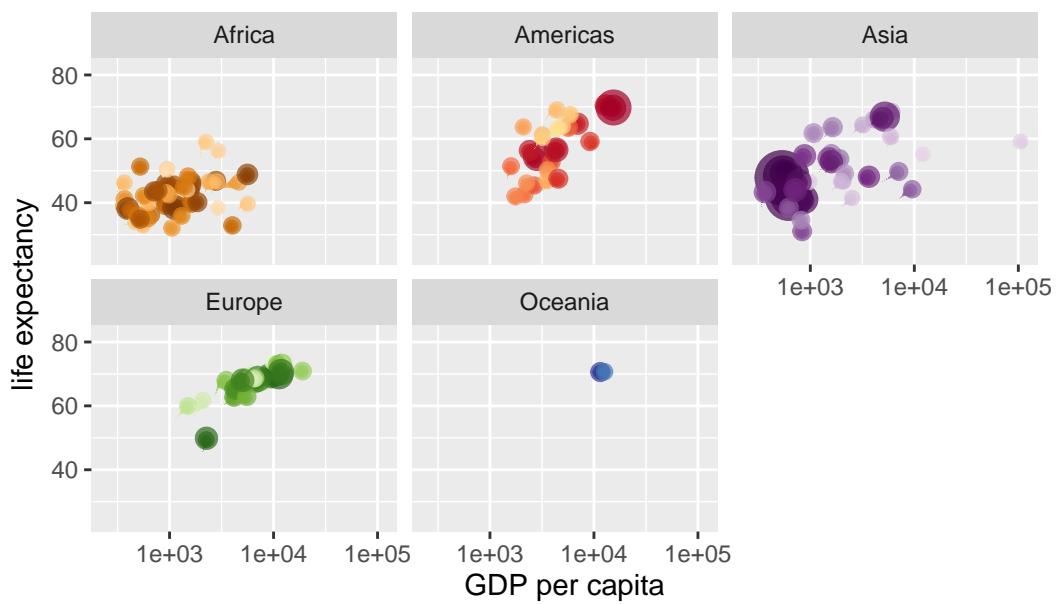
Year: 1958



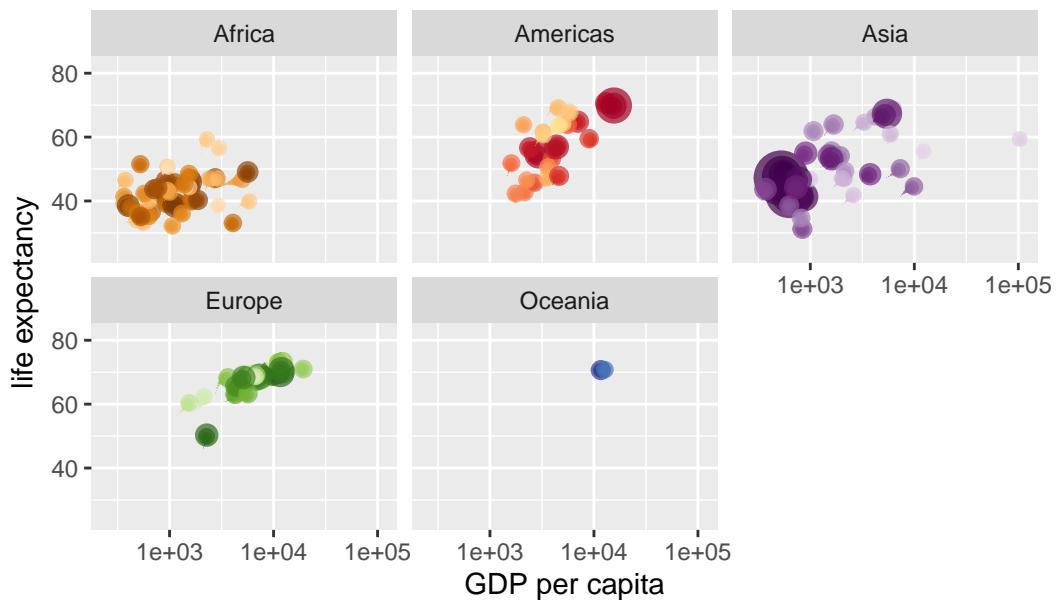
Year: 1959



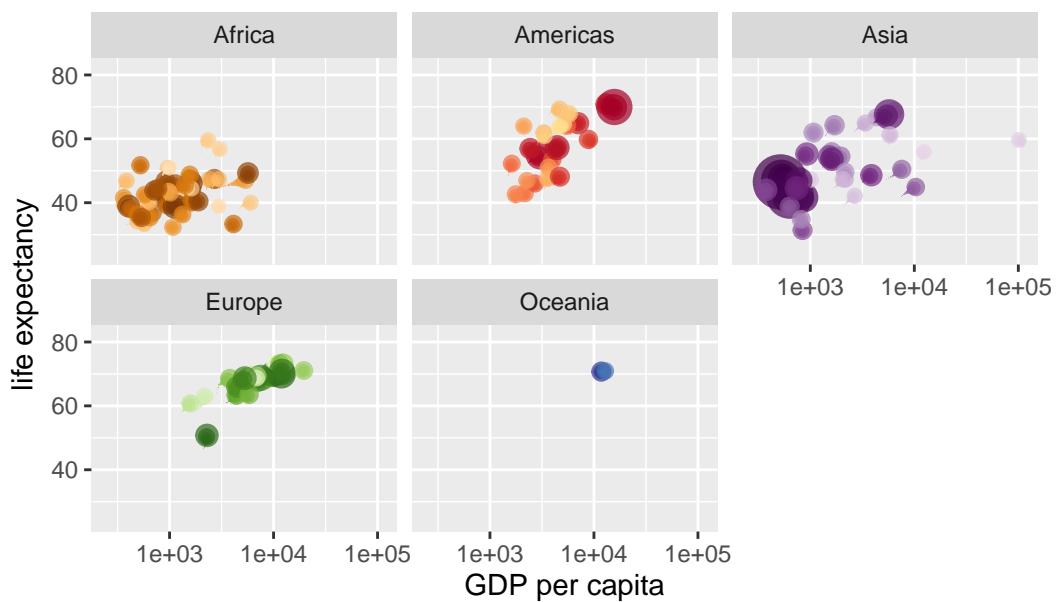
Year: 1959



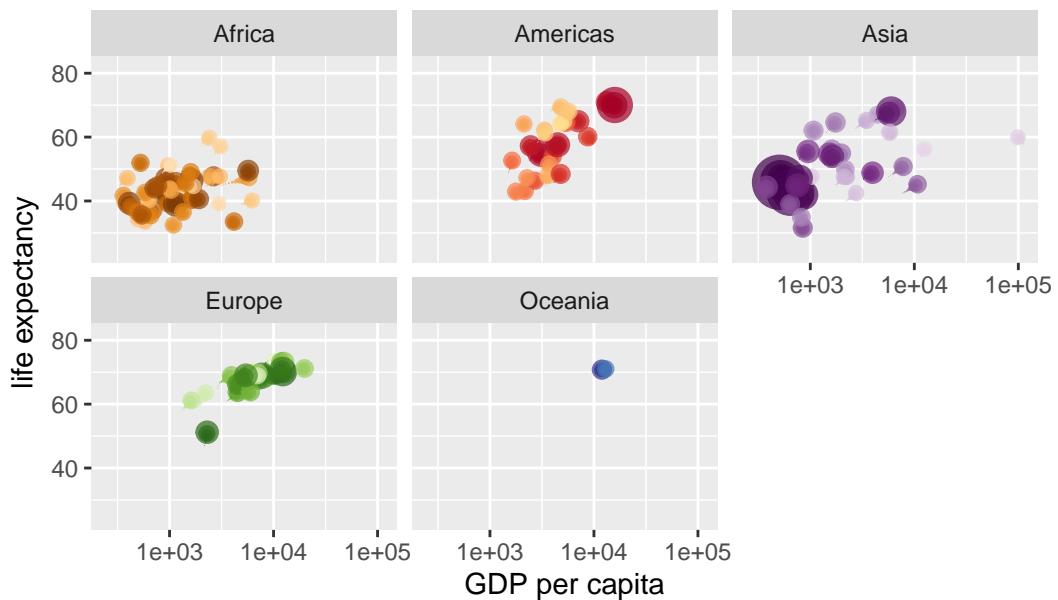
Year: 1960



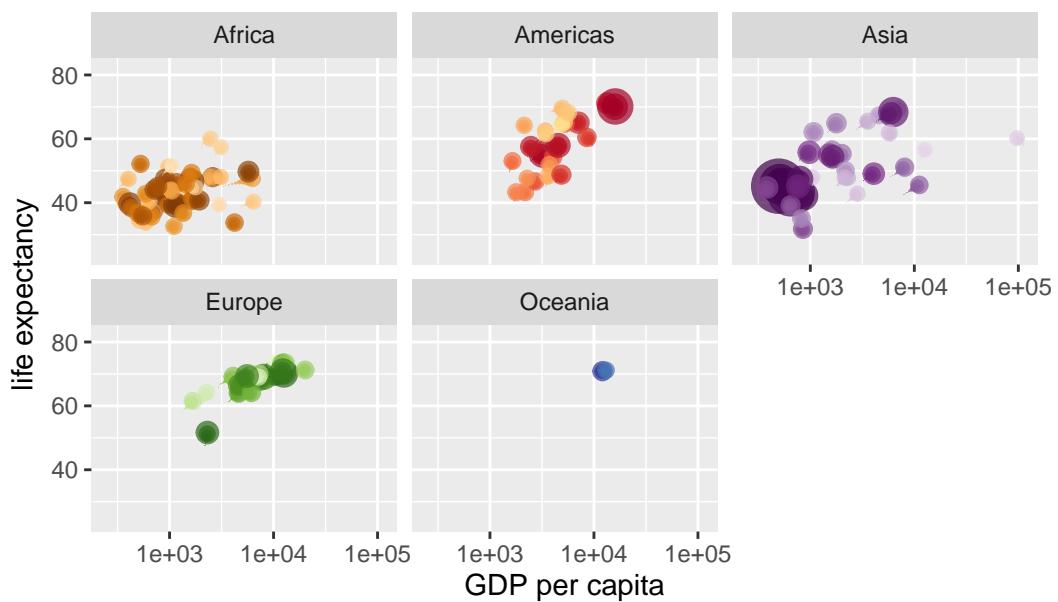
Year: 1960



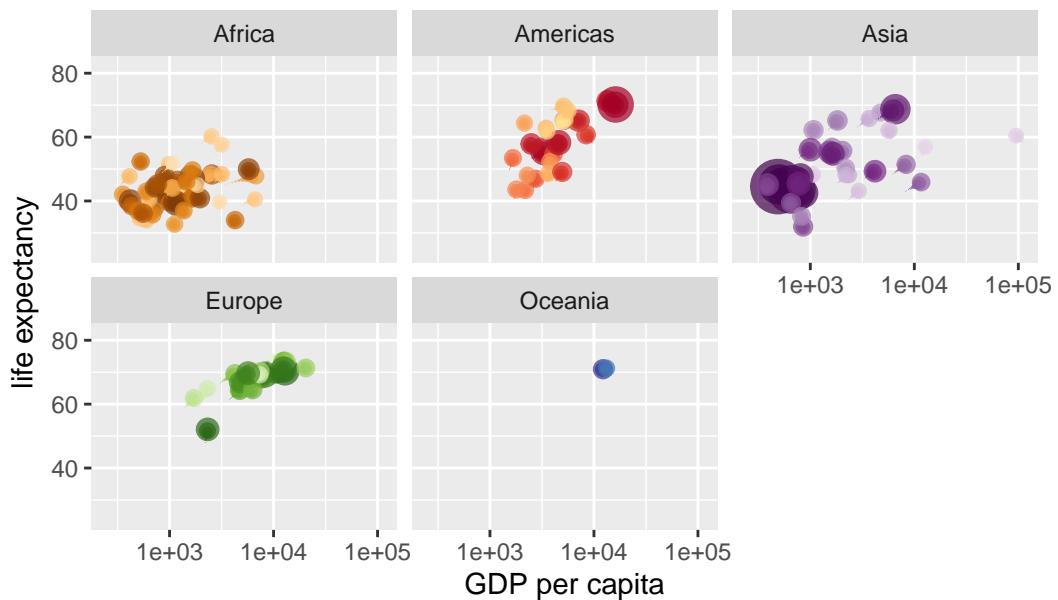
Year: 1961



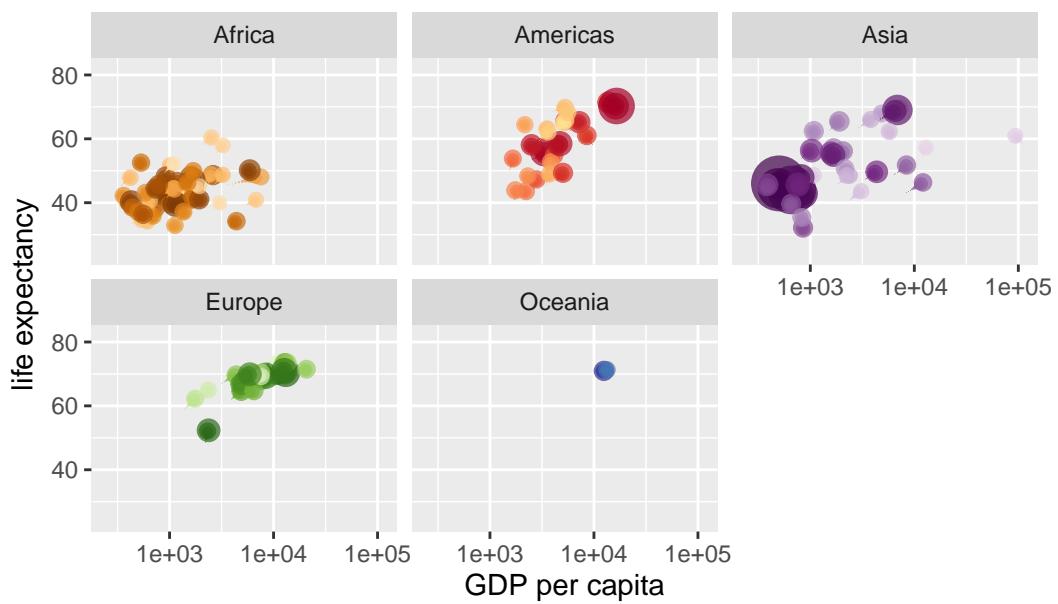
Year: 1961



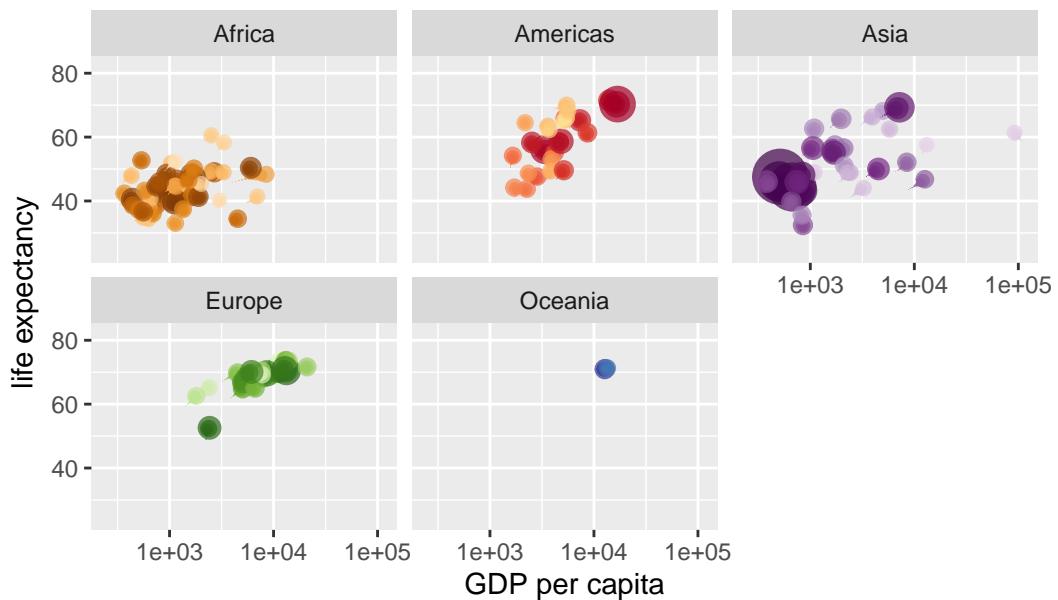
Year: 1962



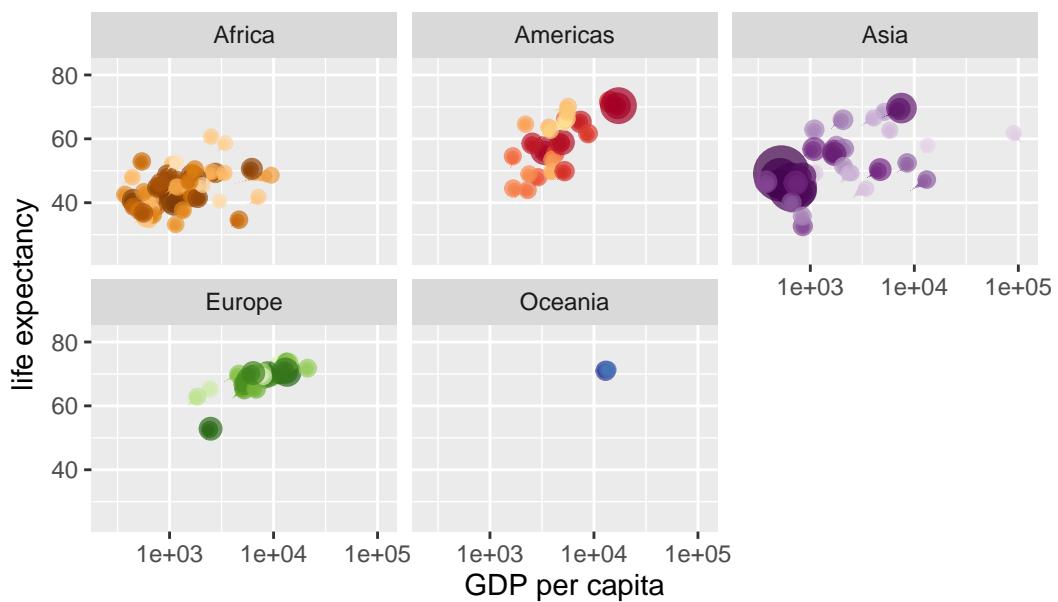
Year: 1963



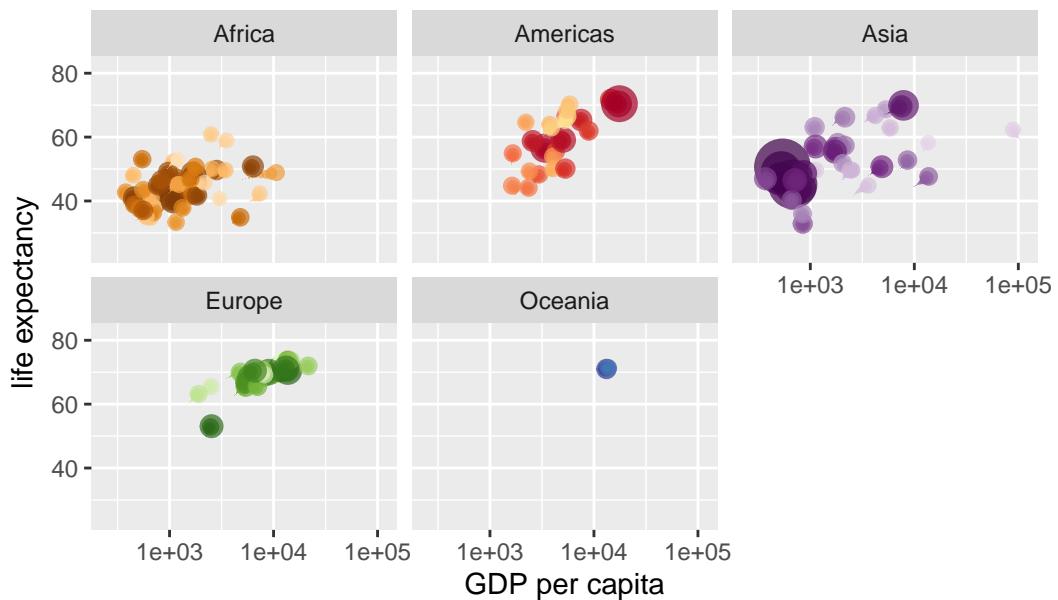
Year: 1963



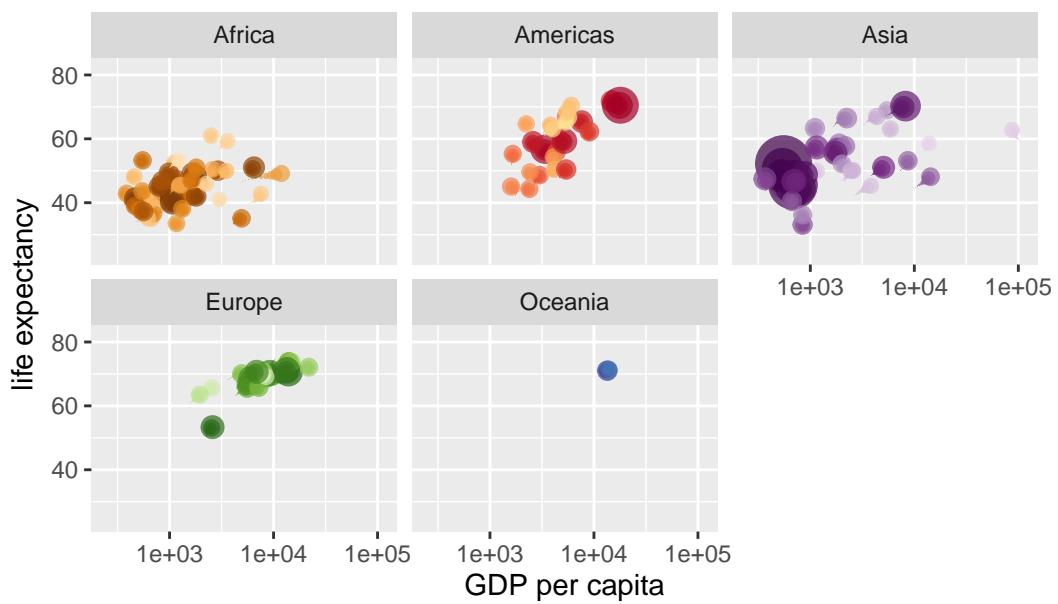
Year: 1964



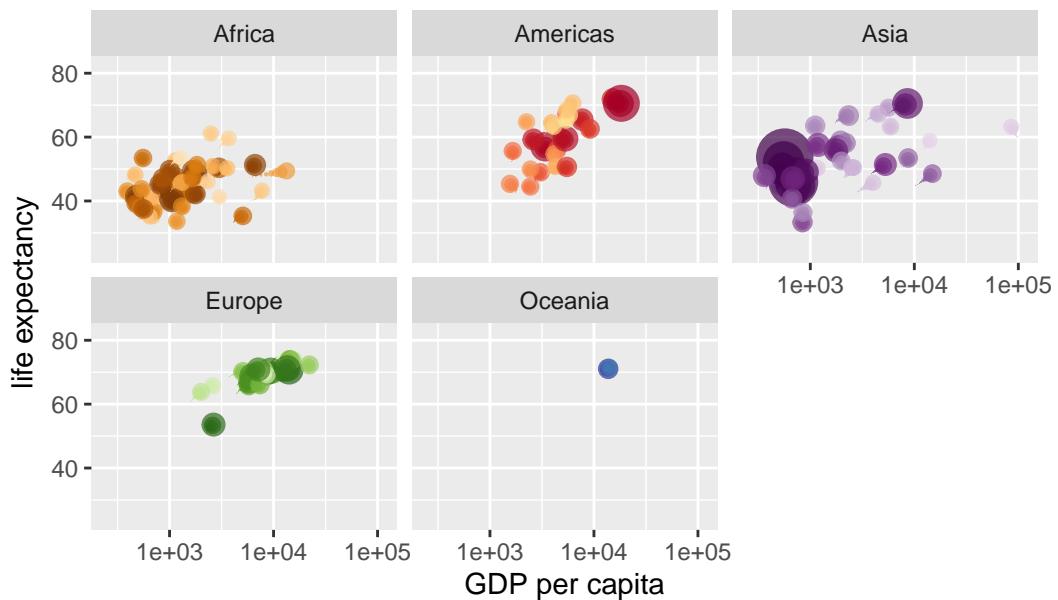
Year: 1964



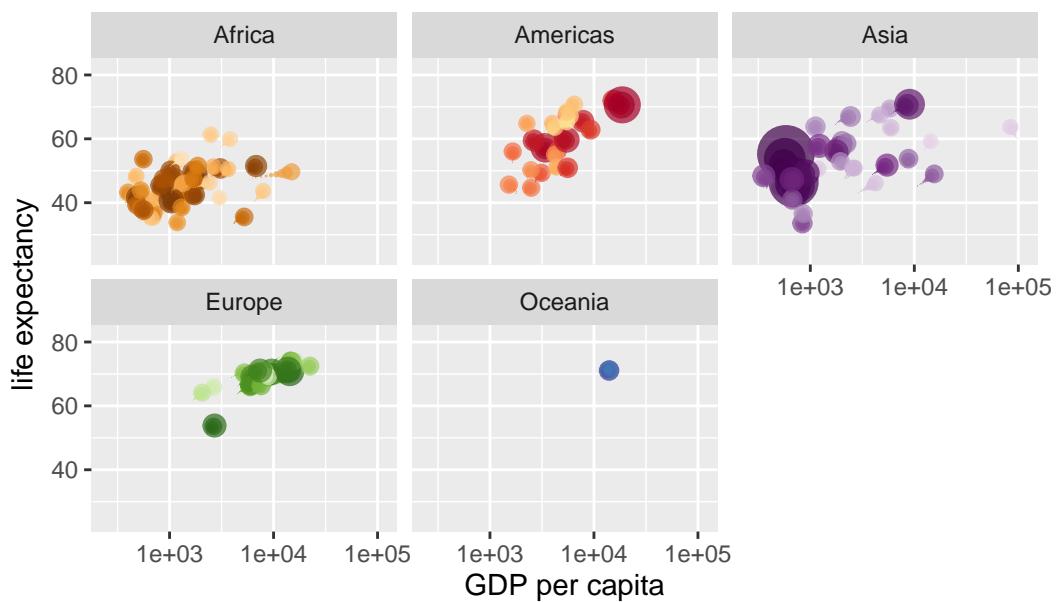
Year: 1965



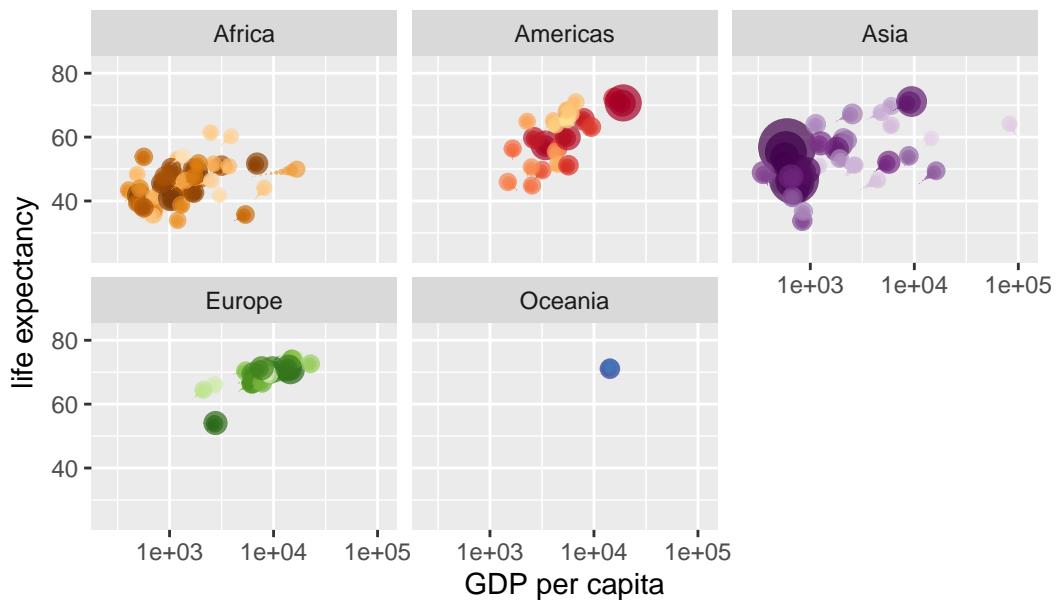
Year: 1965



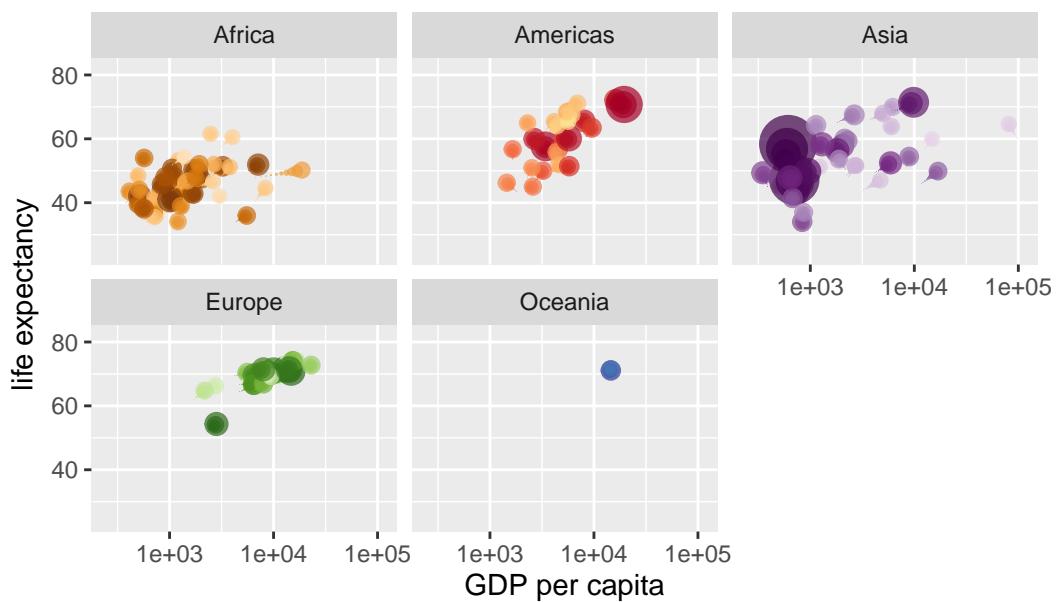
Year: 1966



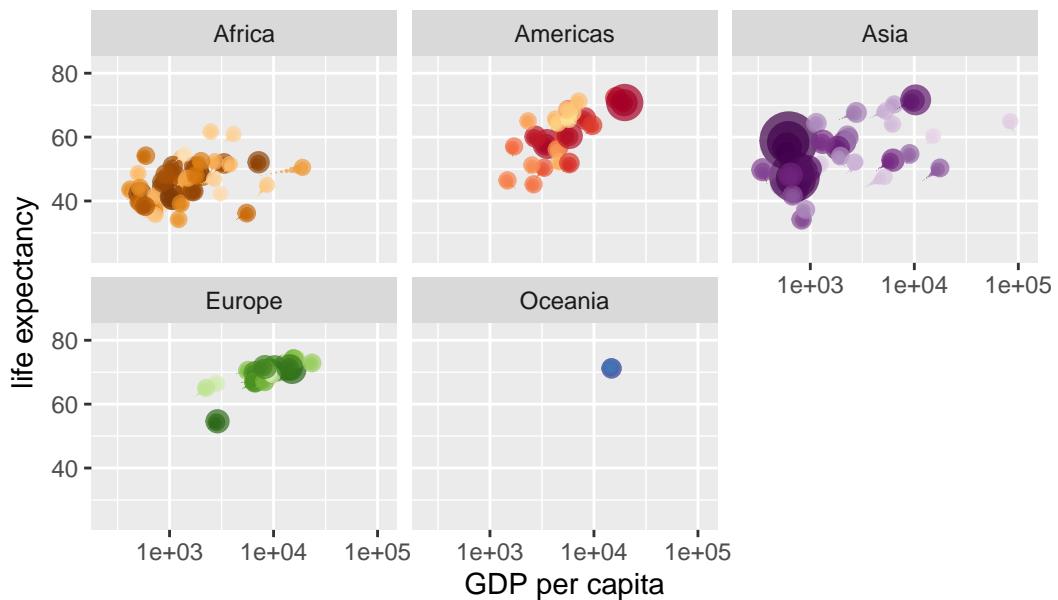
Year: 1966



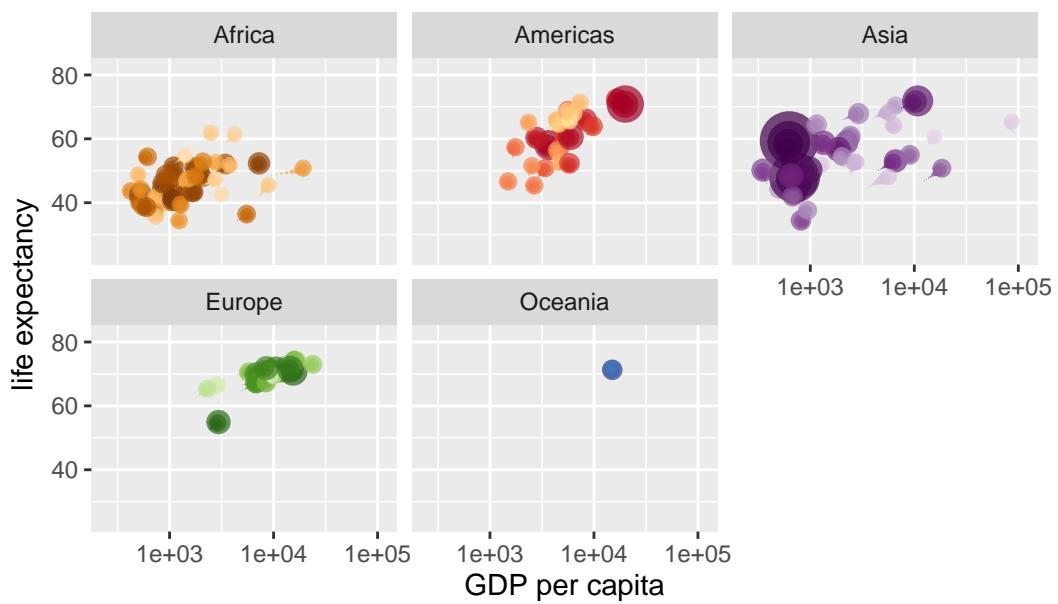
Year: 1967

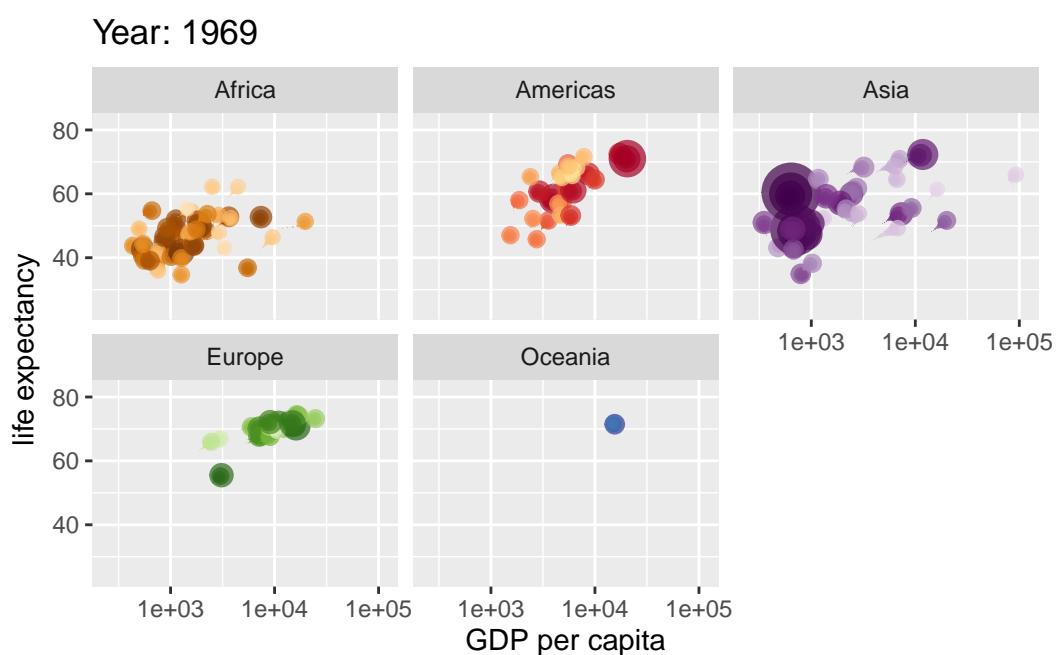
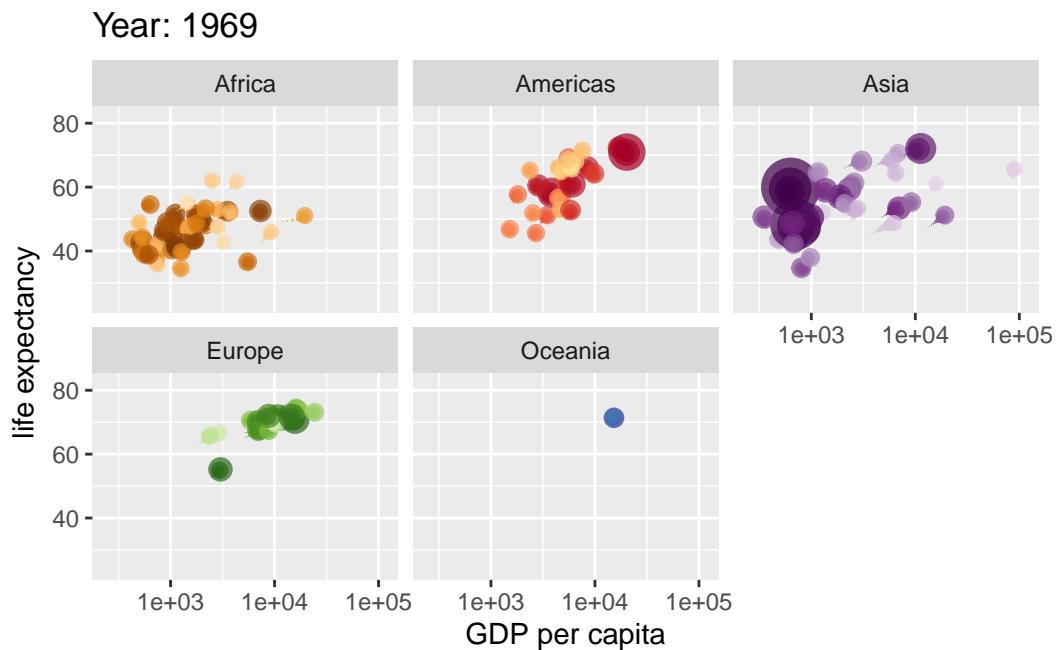


Year: 1968

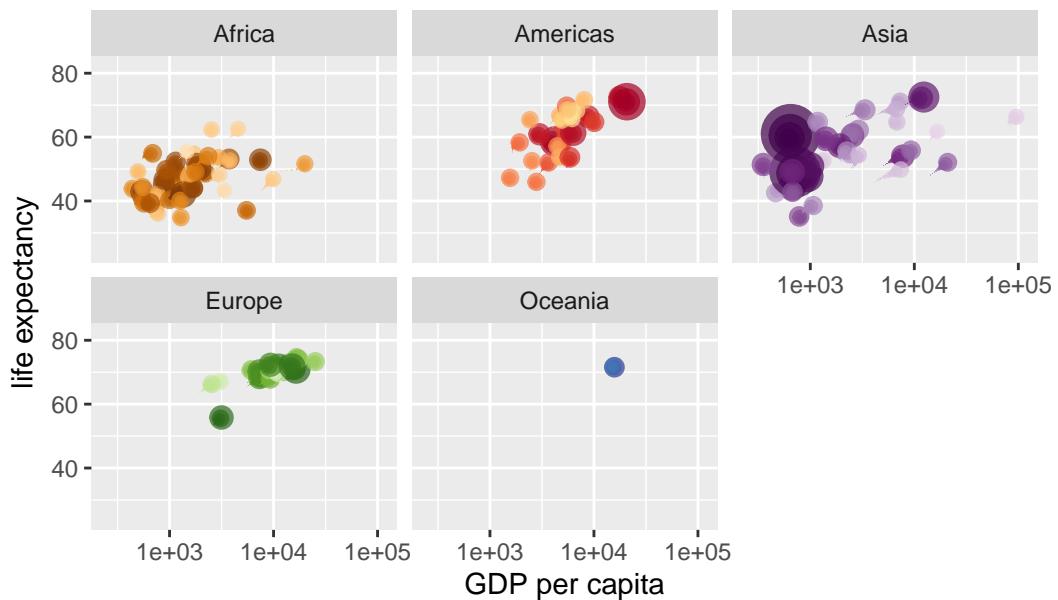


Year: 1968

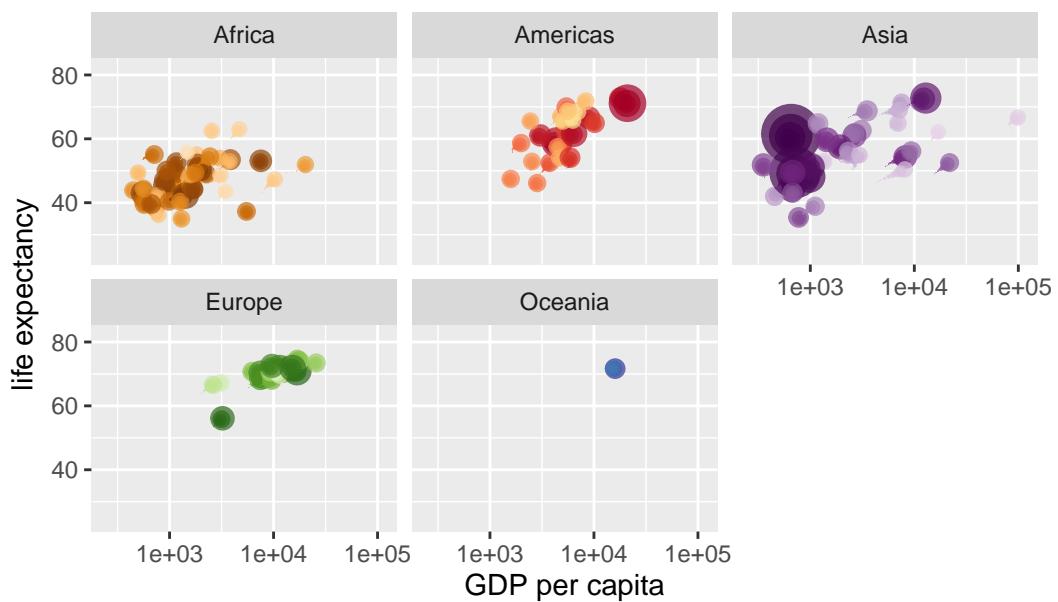




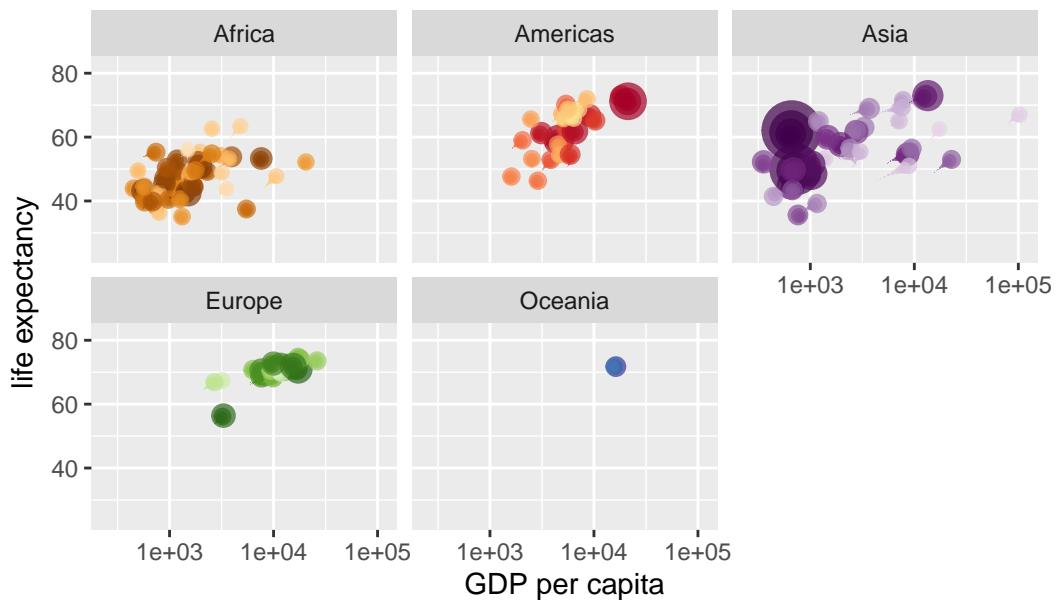
Year: 1970



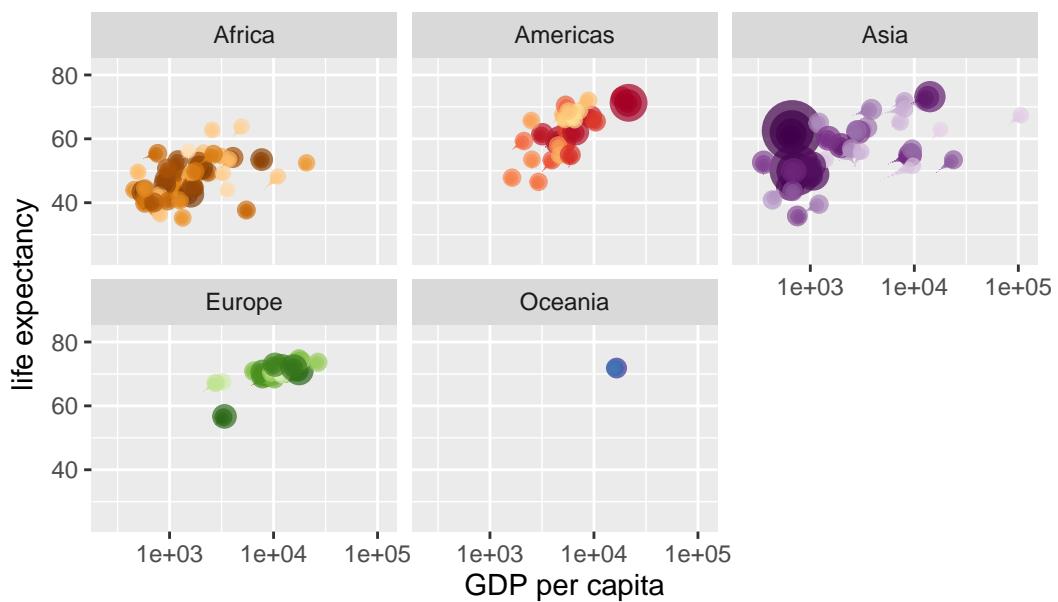
Year: 1970



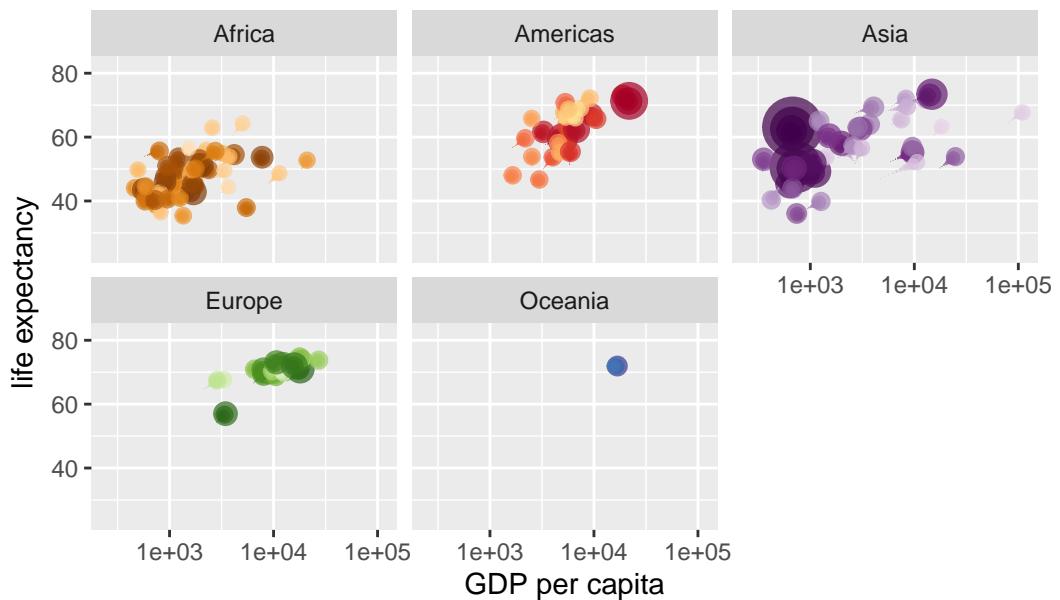
Year: 1971



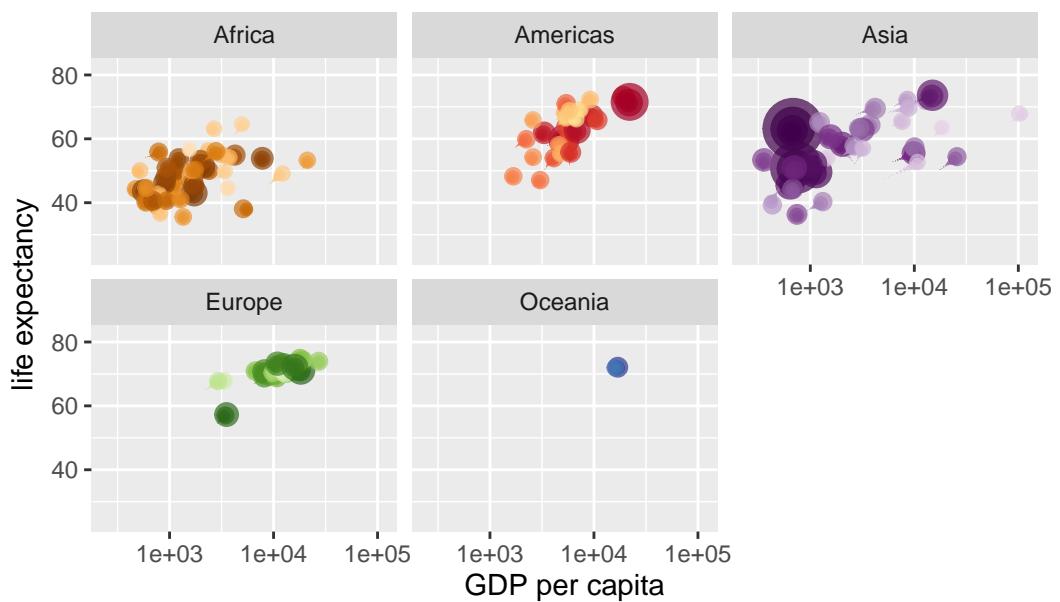
Year: 1971



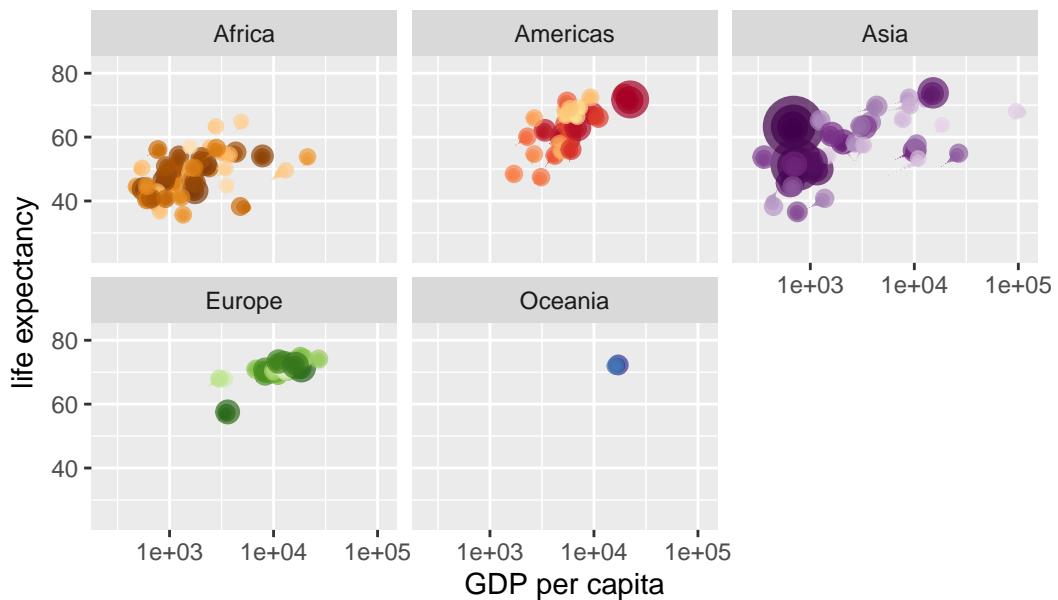
Year: 1972



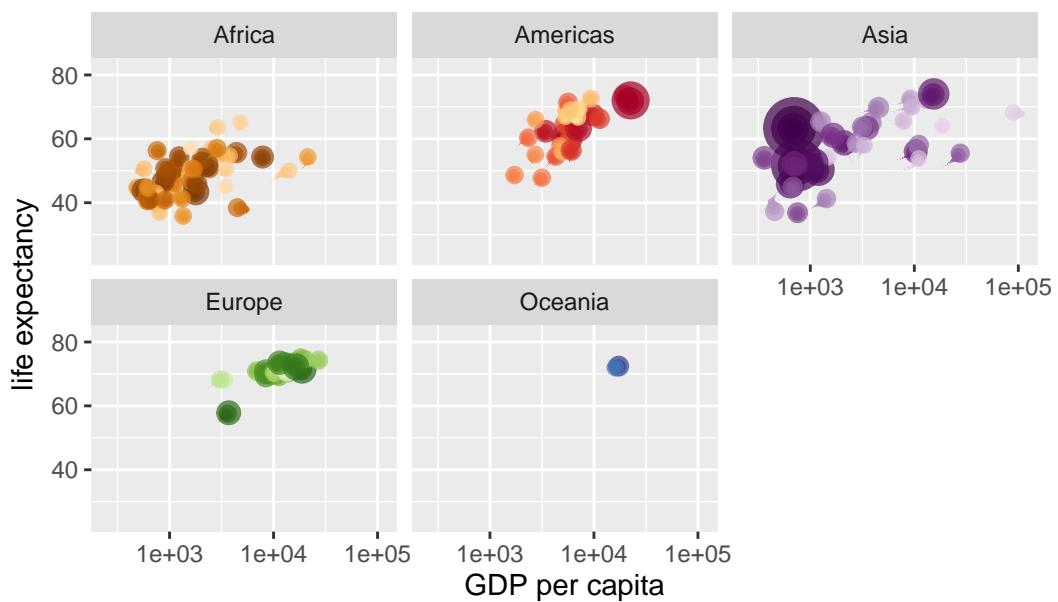
Year: 1973



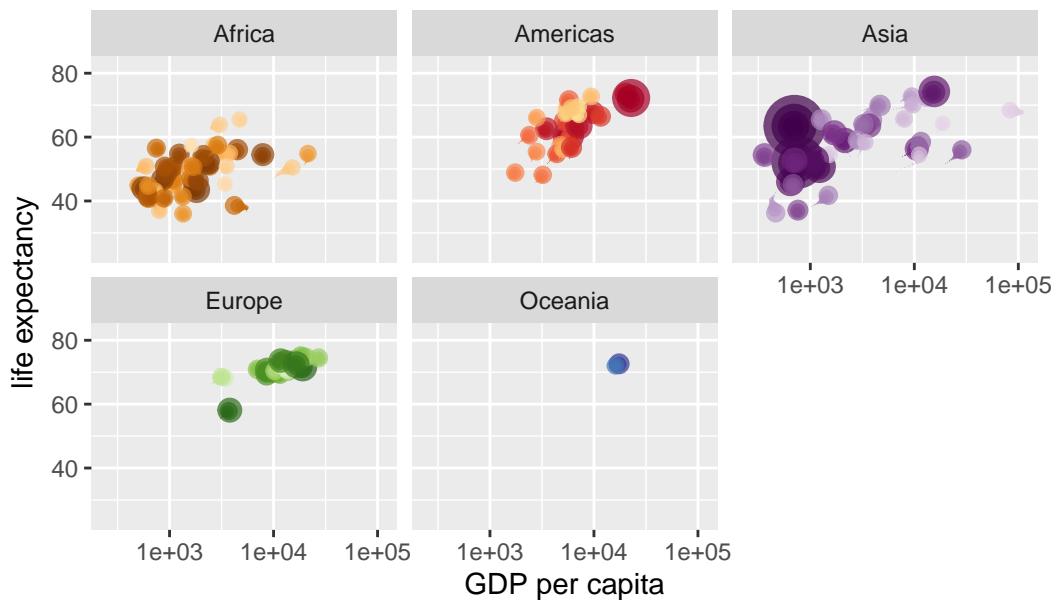
Year: 1973



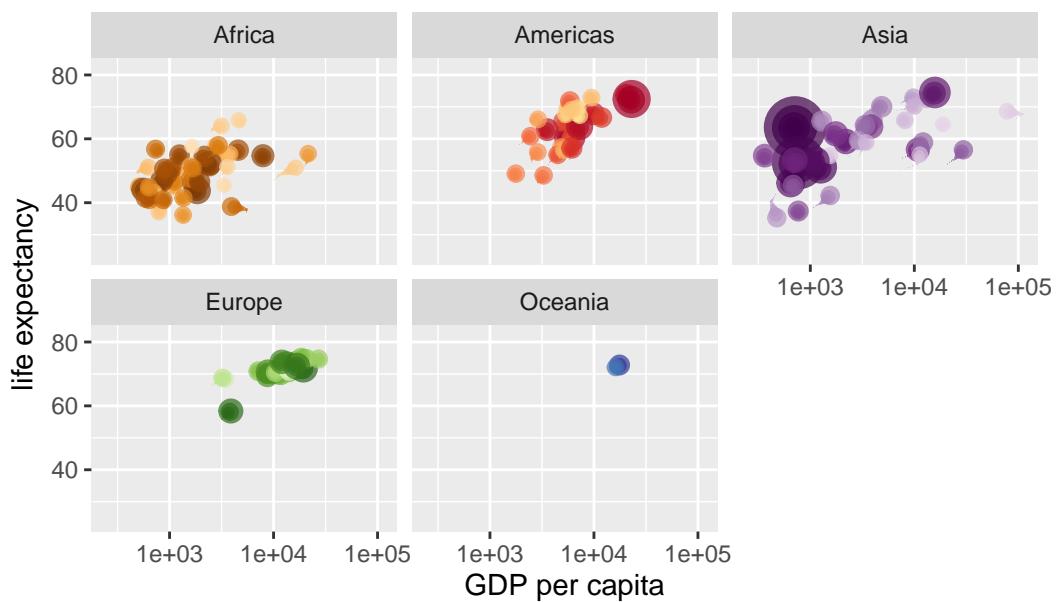
Year: 1974



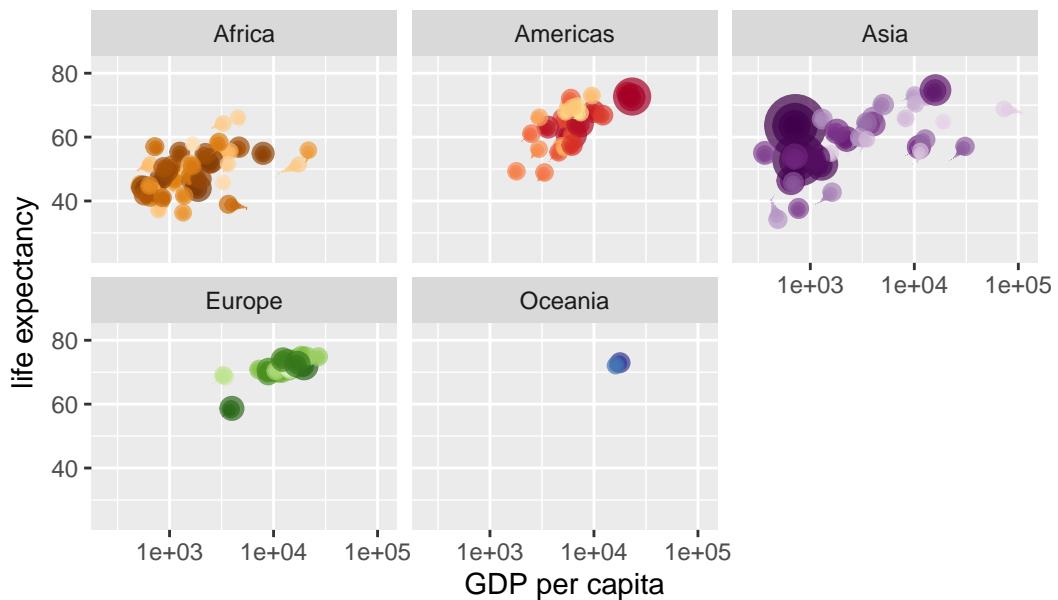
Year: 1974



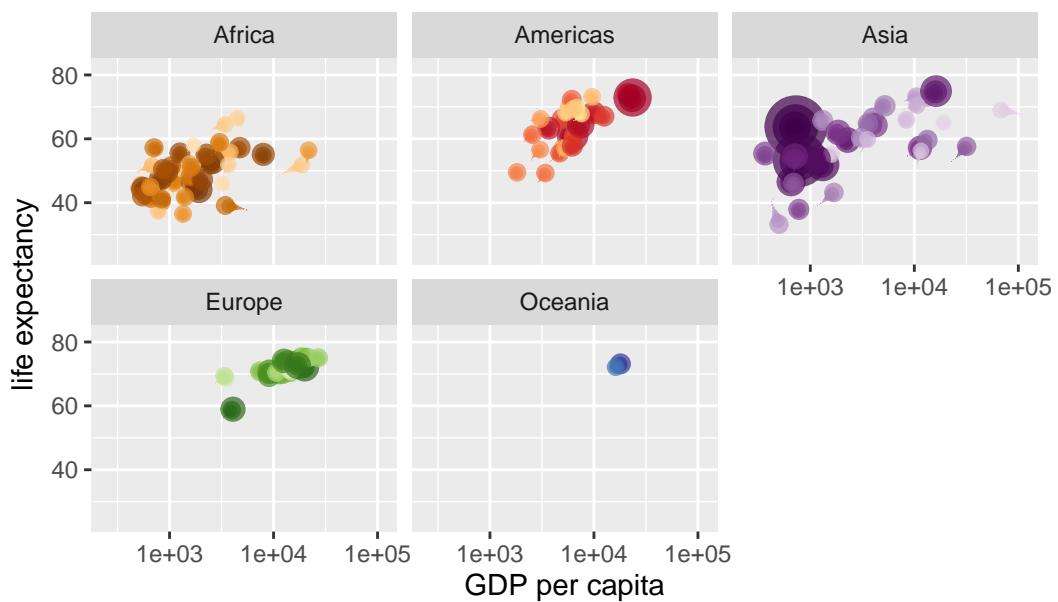
Year: 1975



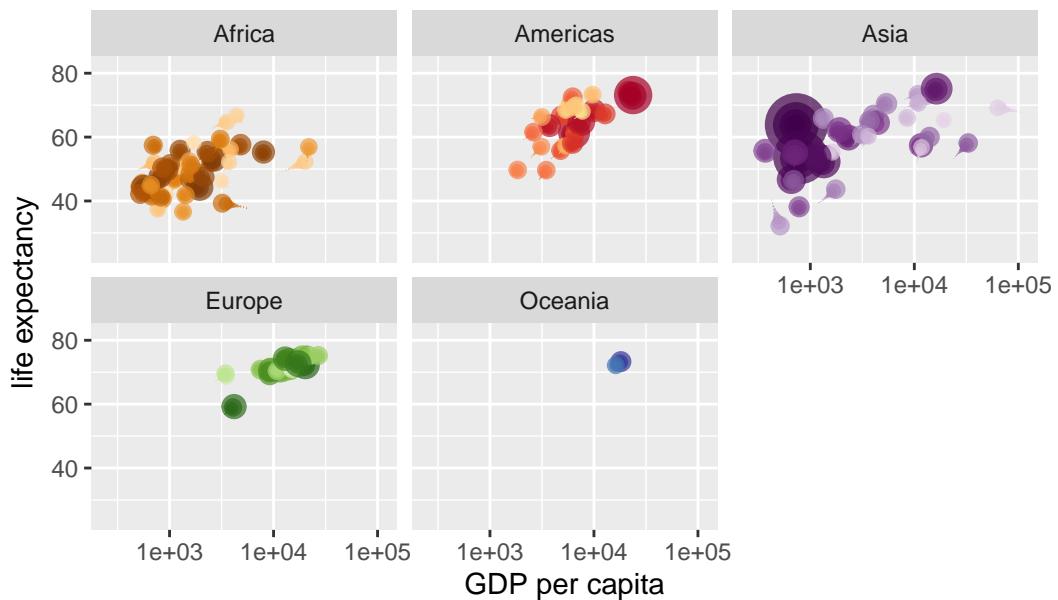
Year: 1975



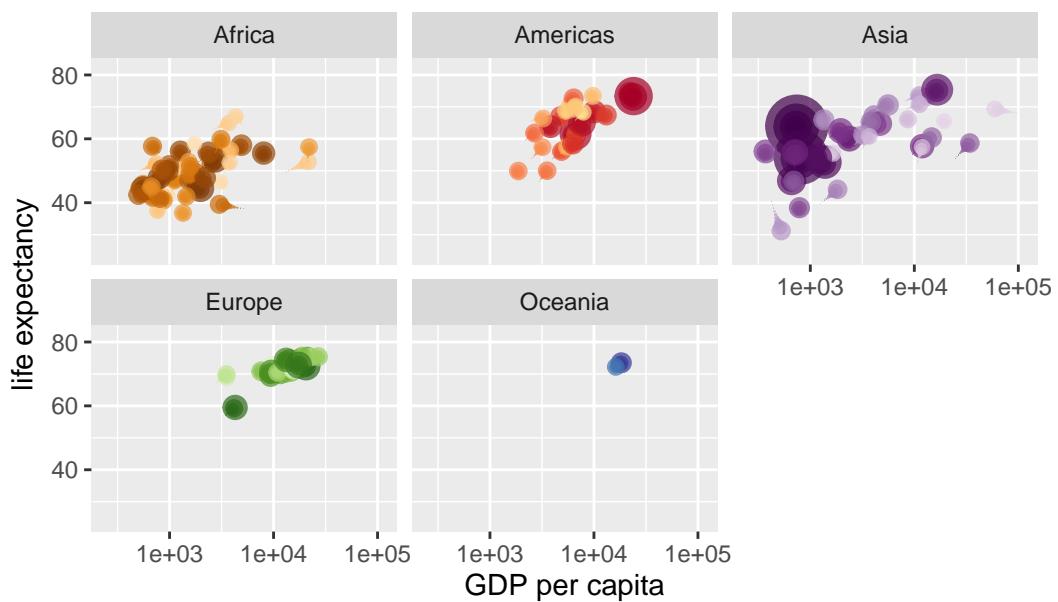
Year: 1976



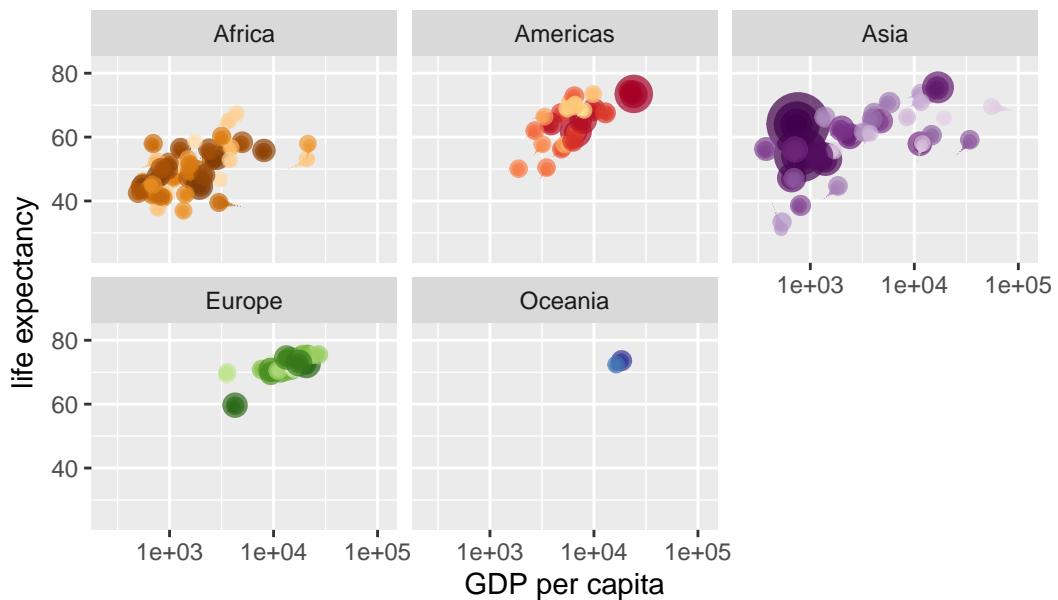
Year: 1976



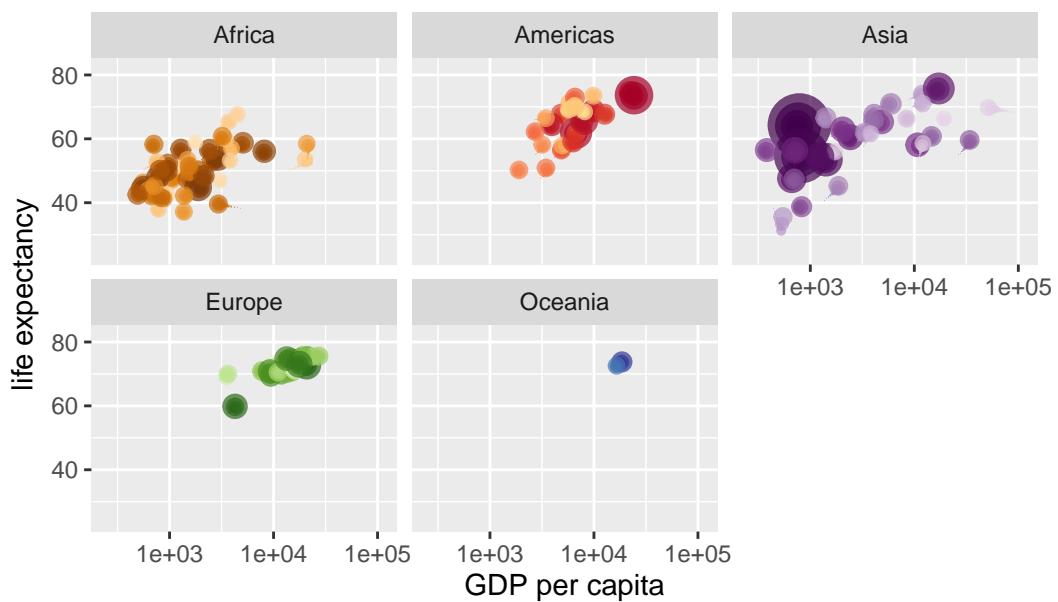
Year: 1977



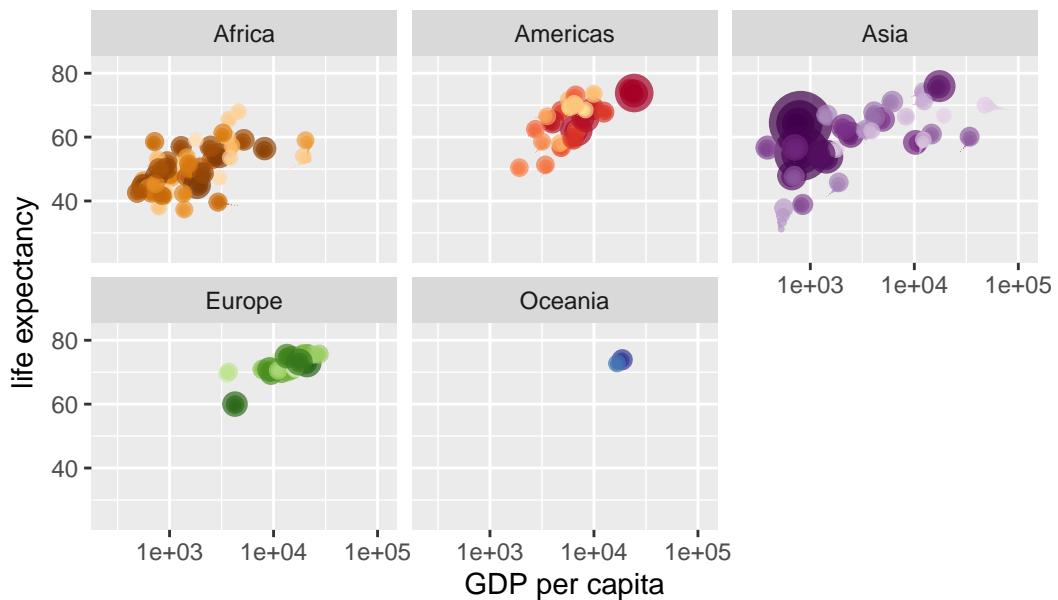
Year: 1978



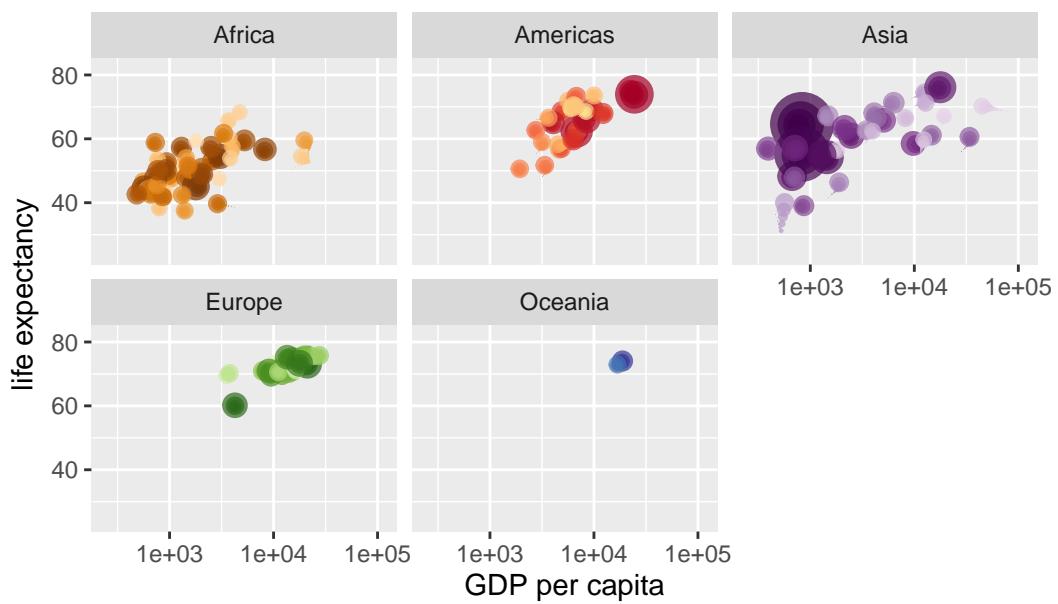
Year: 1978



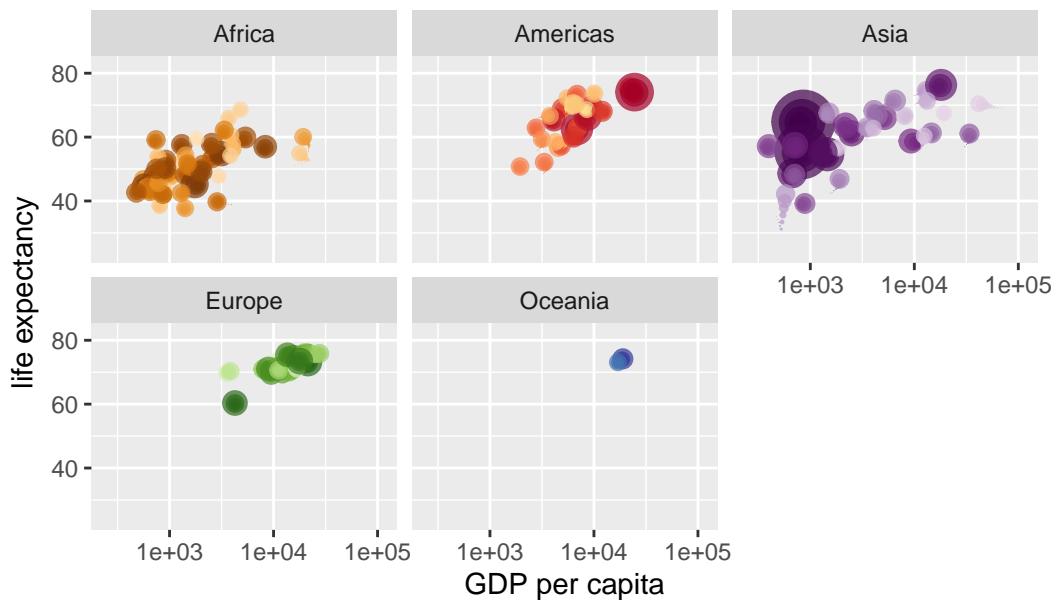
Year: 1979



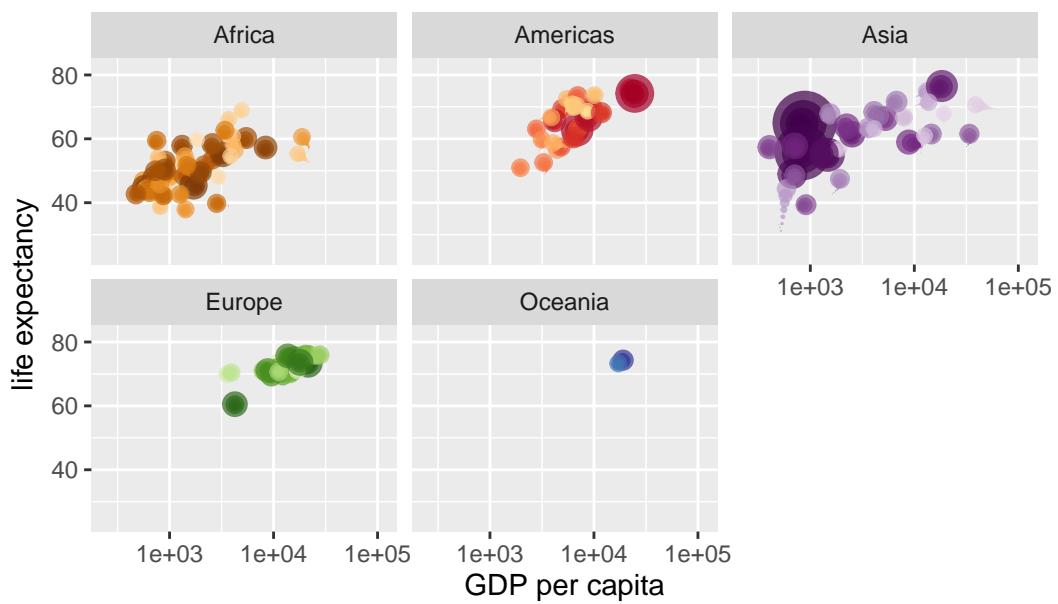
Year: 1979



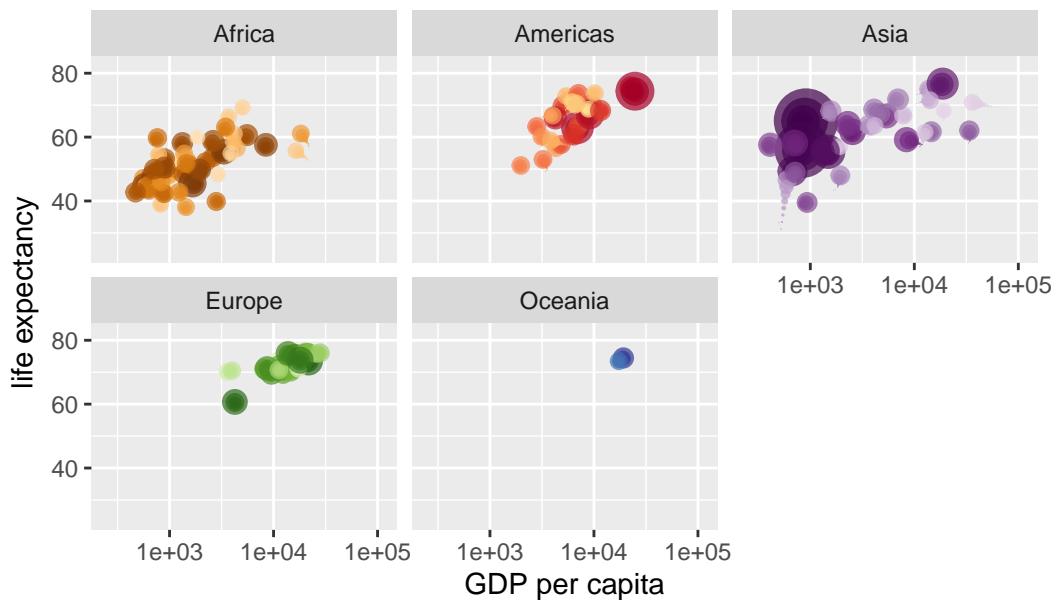
Year: 1980



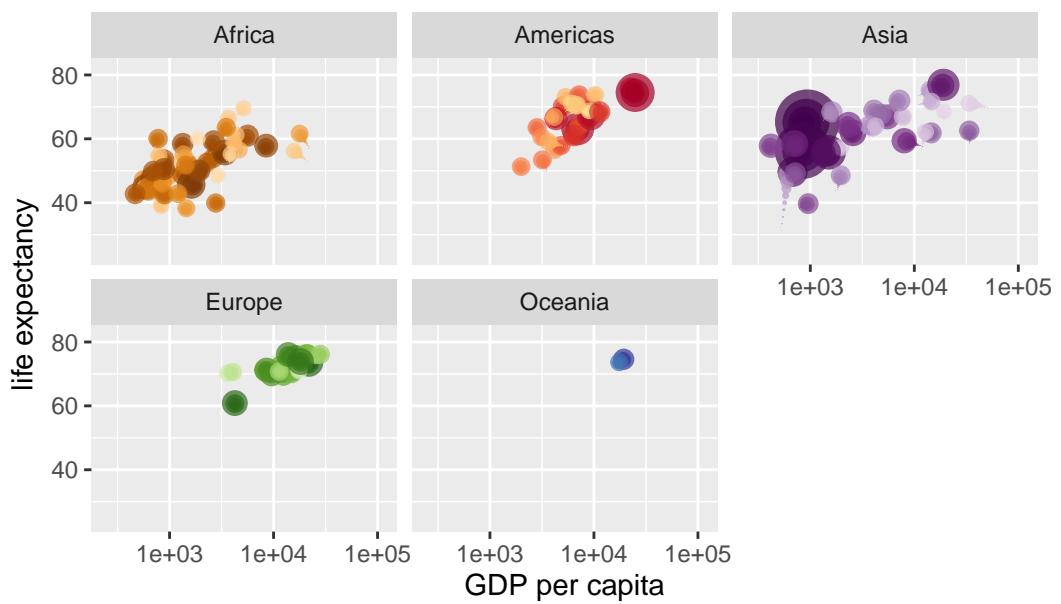
Year: 1980



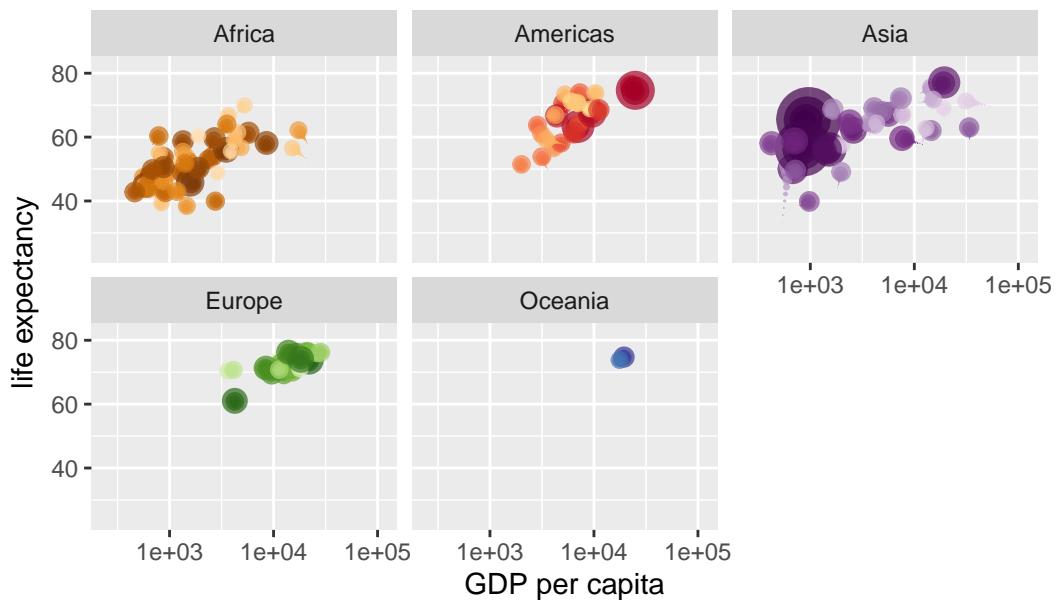
Year: 1981



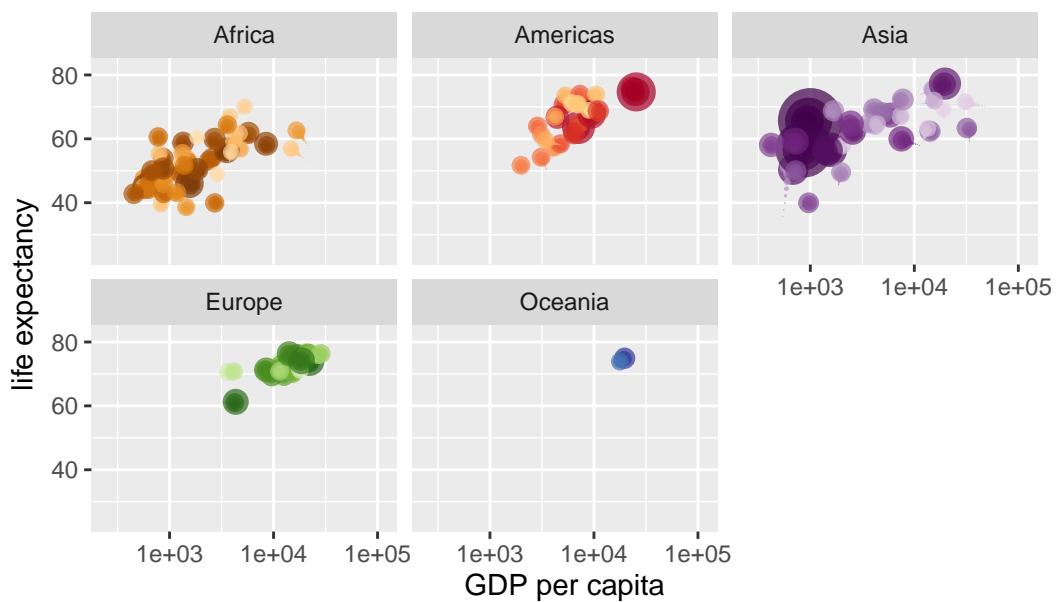
Year: 1981



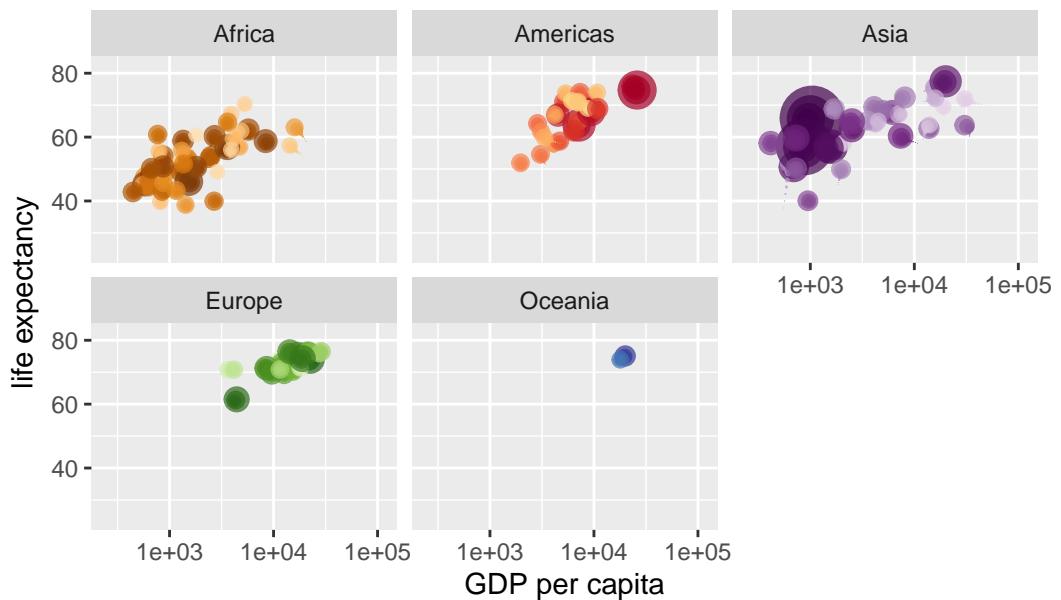
Year: 1982



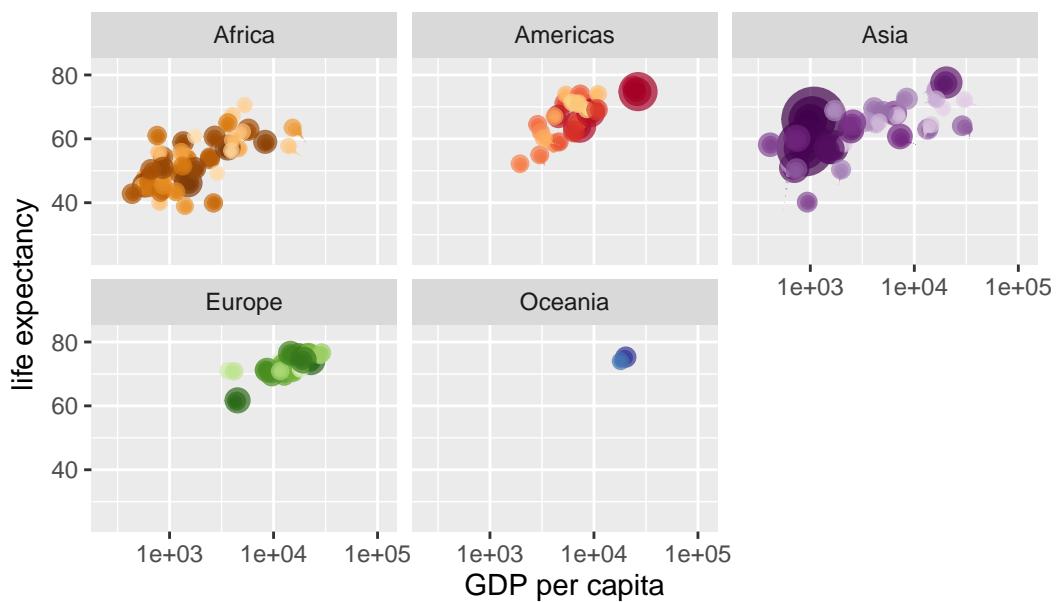
Year: 1983



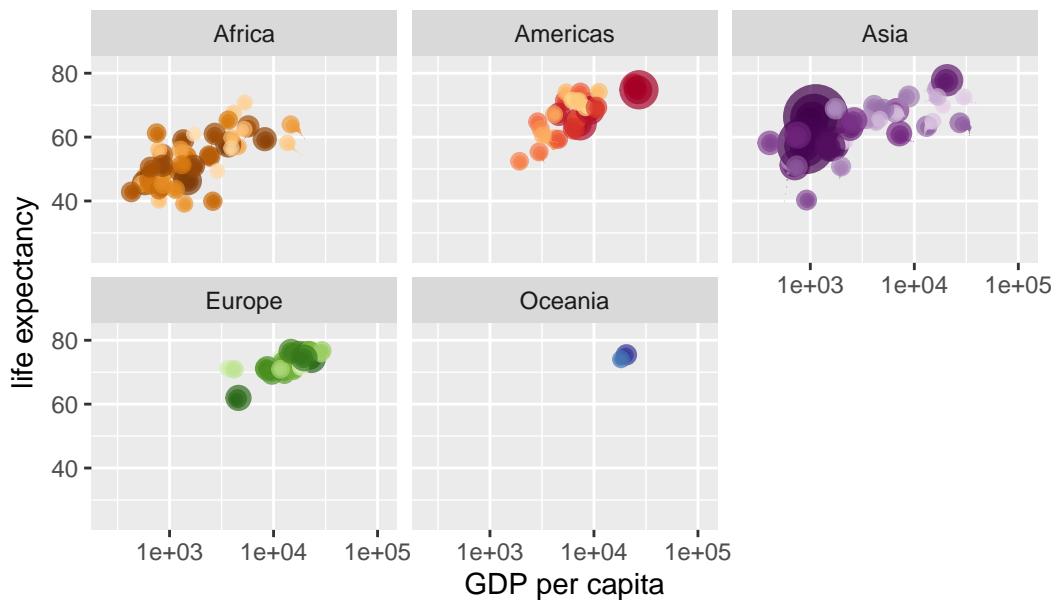
Year: 1983



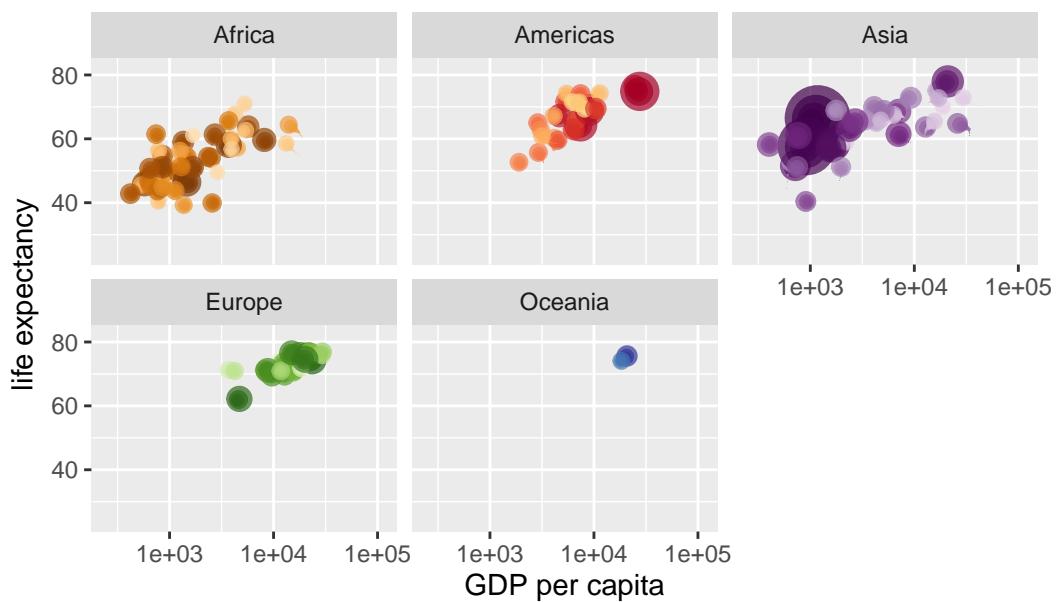
Year: 1984



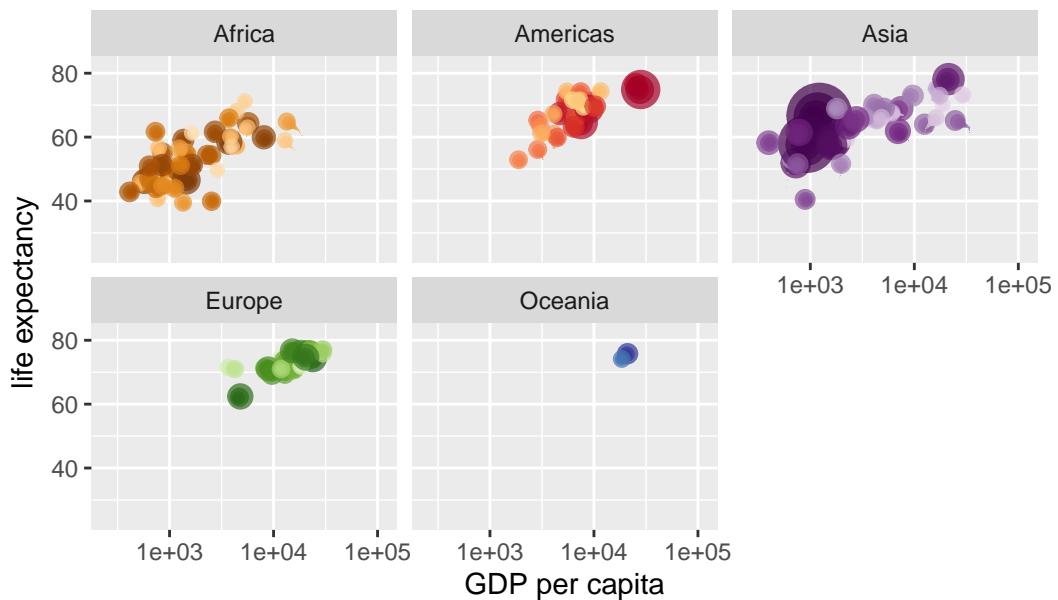
Year: 1984



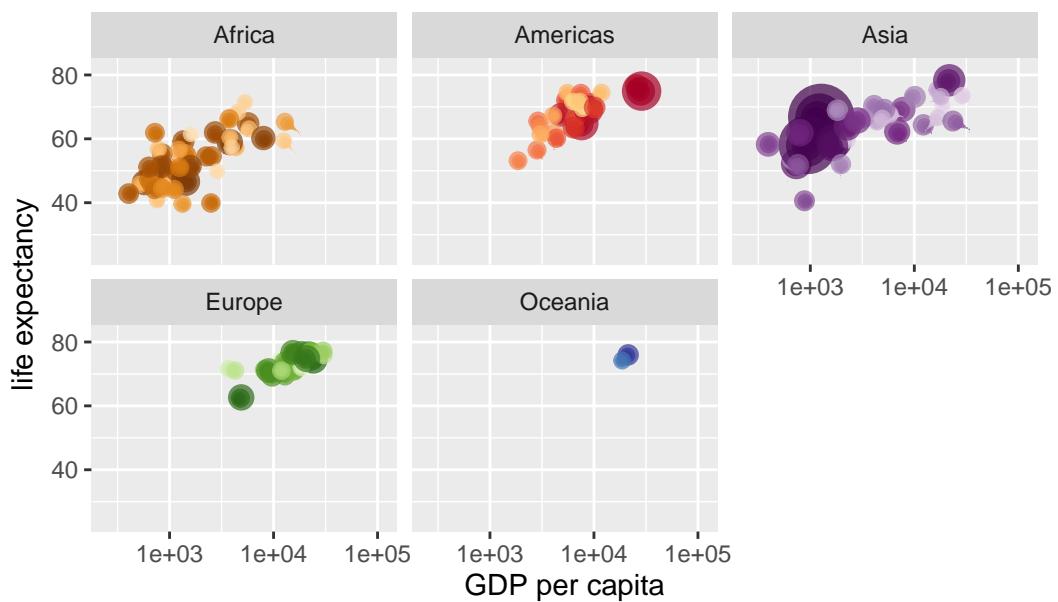
Year: 1985



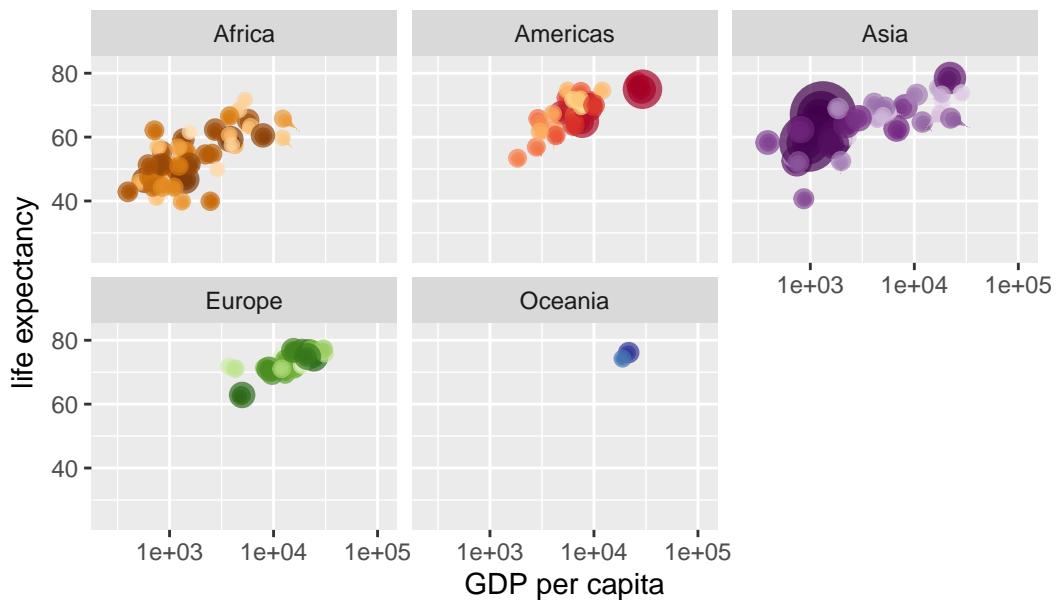
Year: 1985



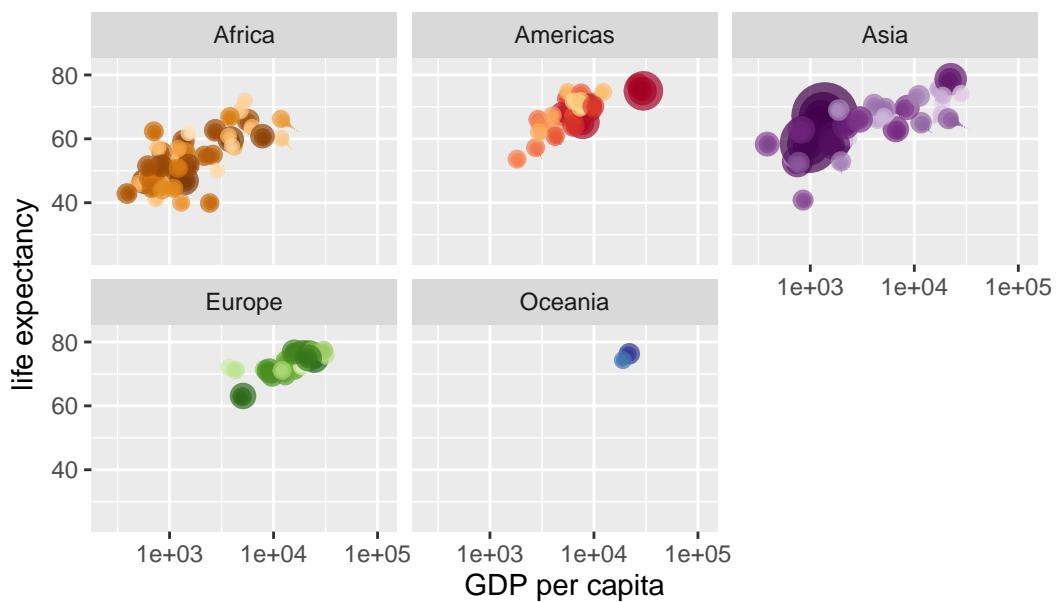
Year: 1986



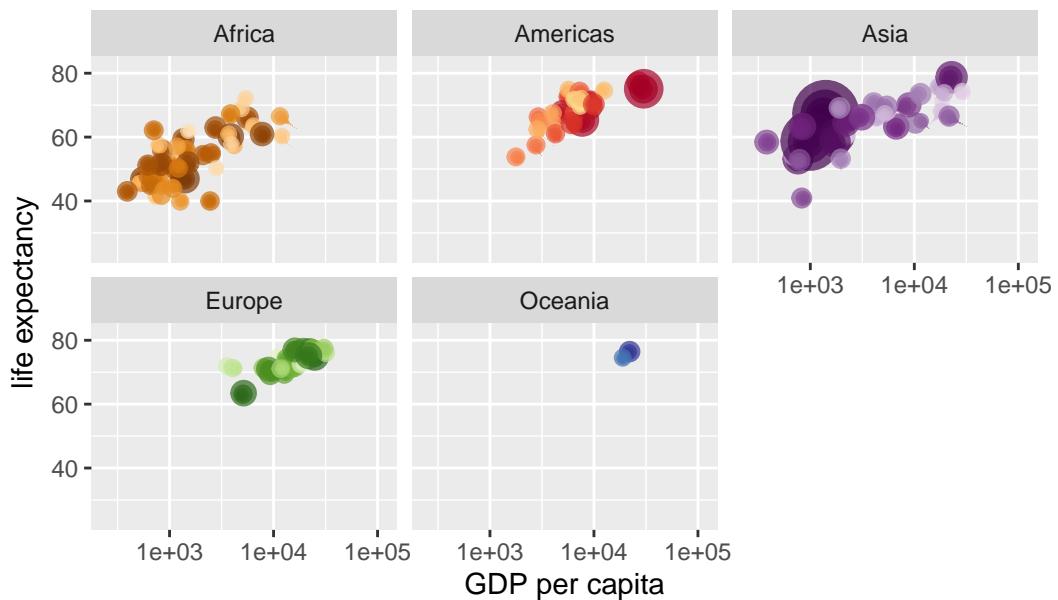
Year: 1986



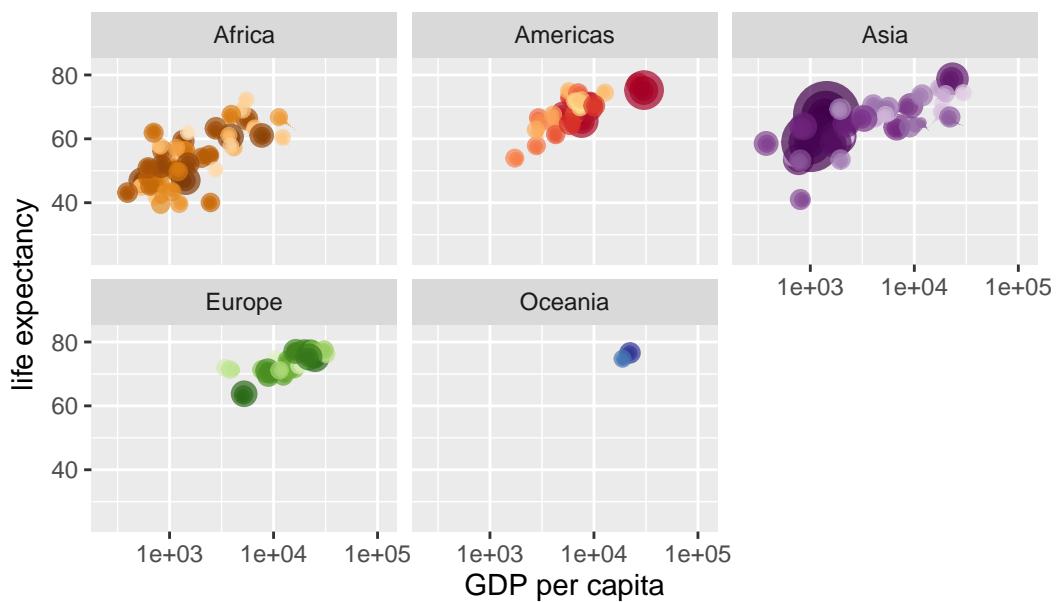
Year: 1987



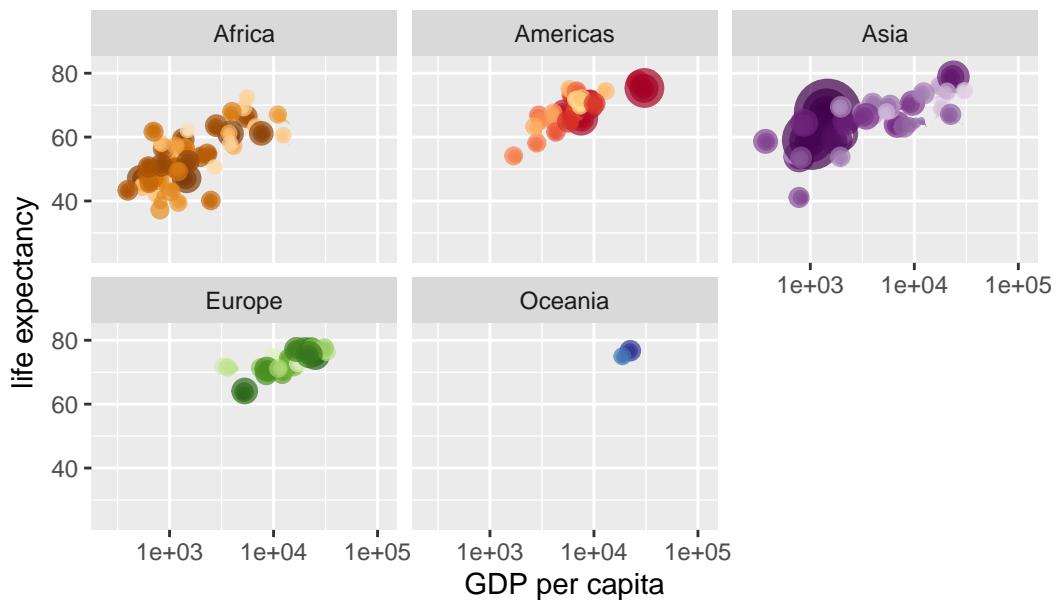
Year: 1988



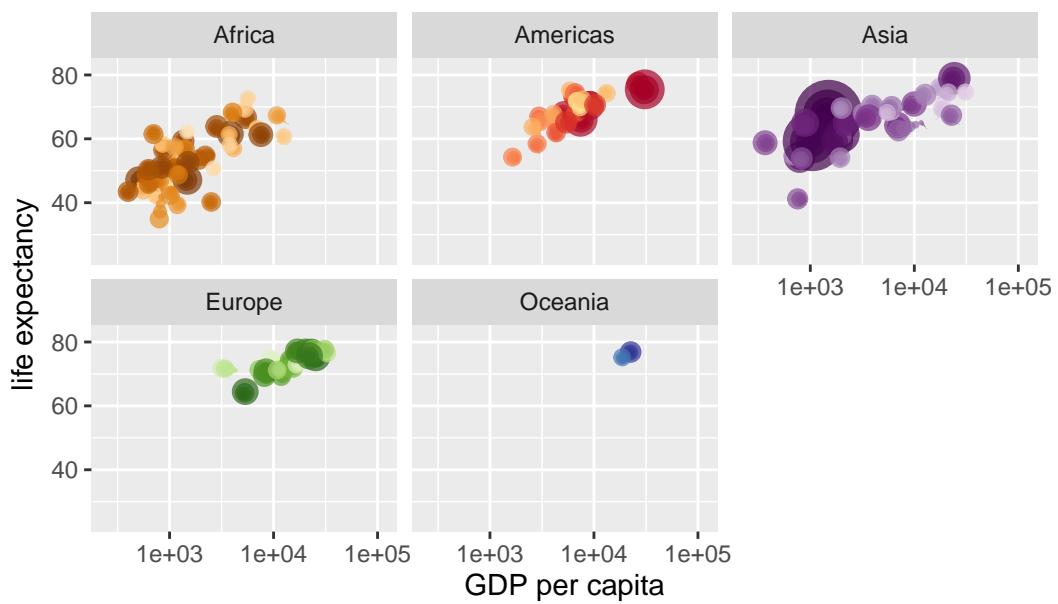
Year: 1988



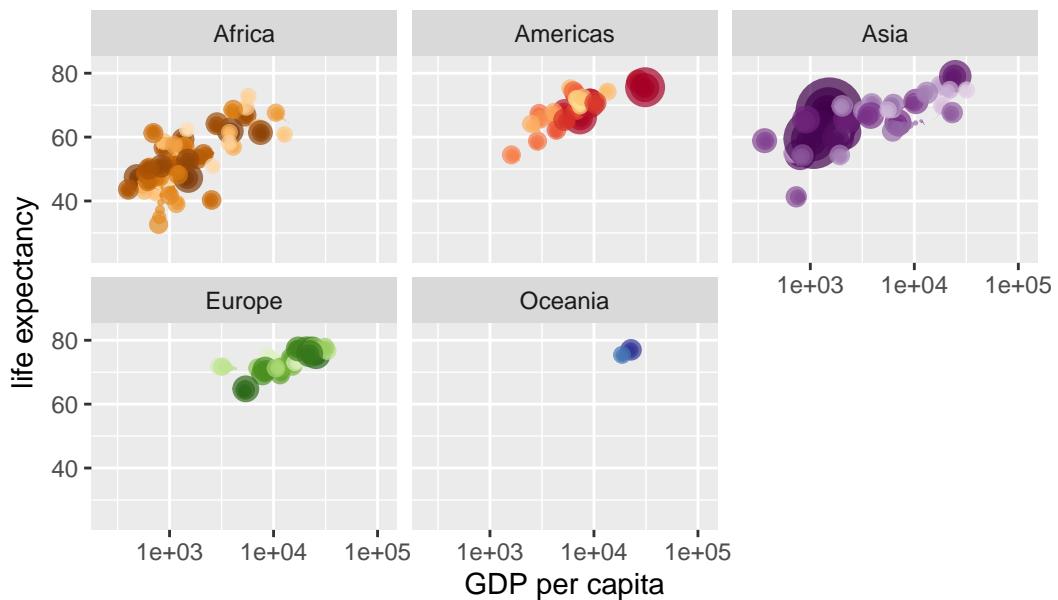
Year: 1989



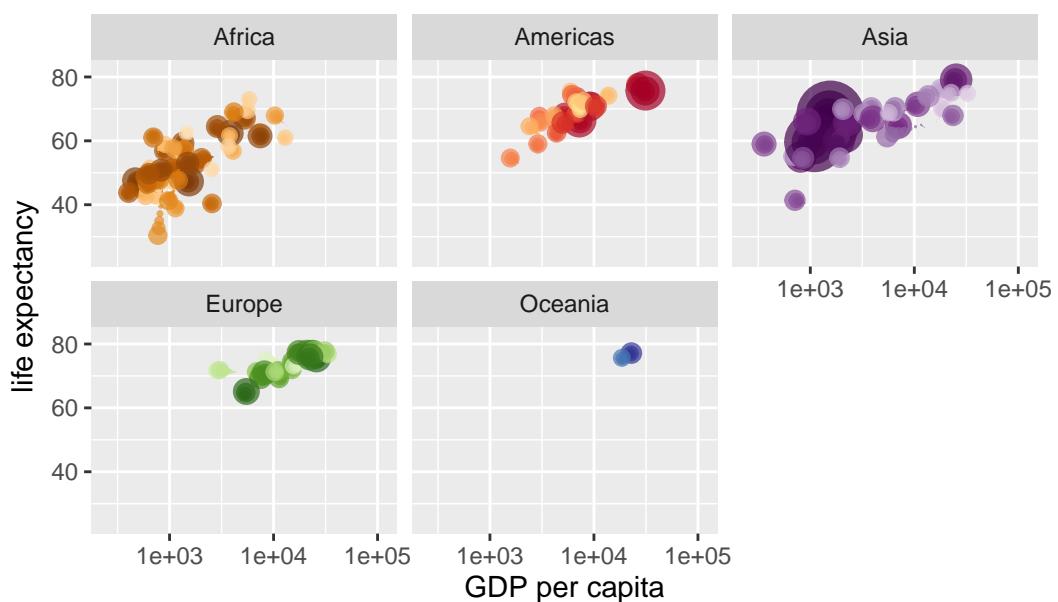
Year: 1989



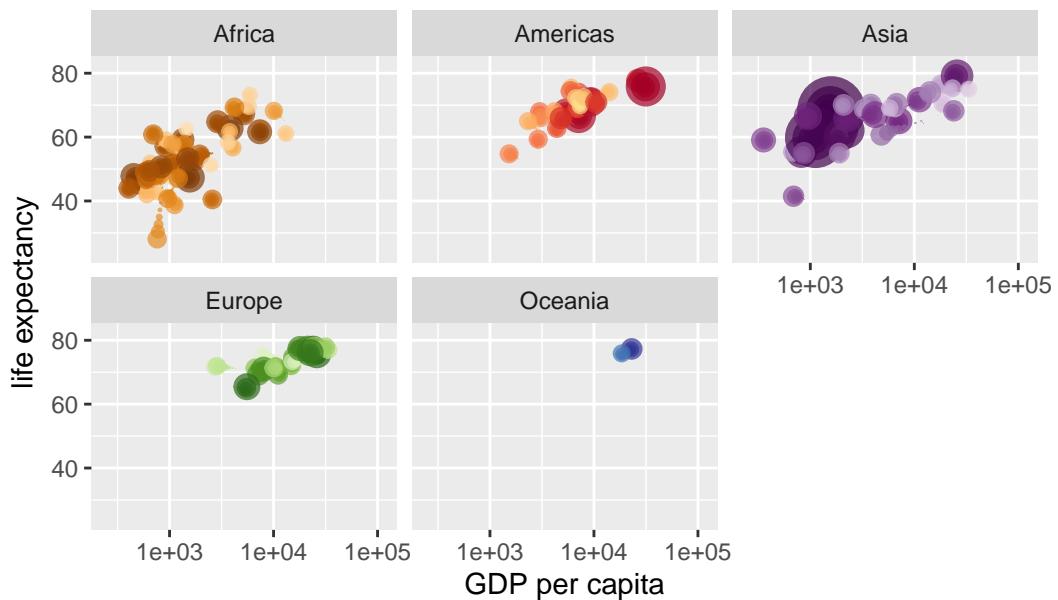
Year: 1990



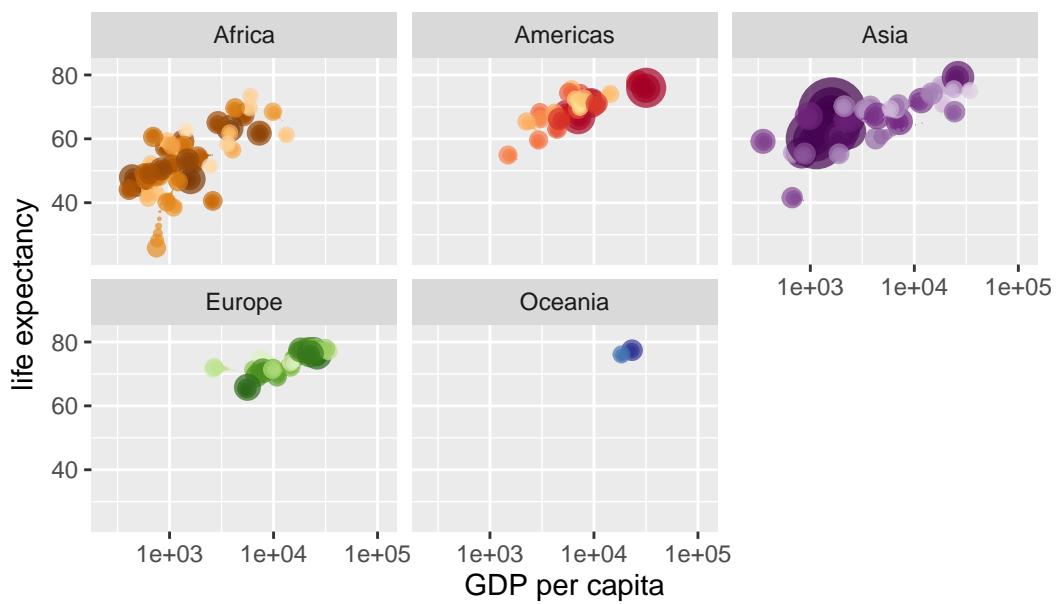
Year: 1990



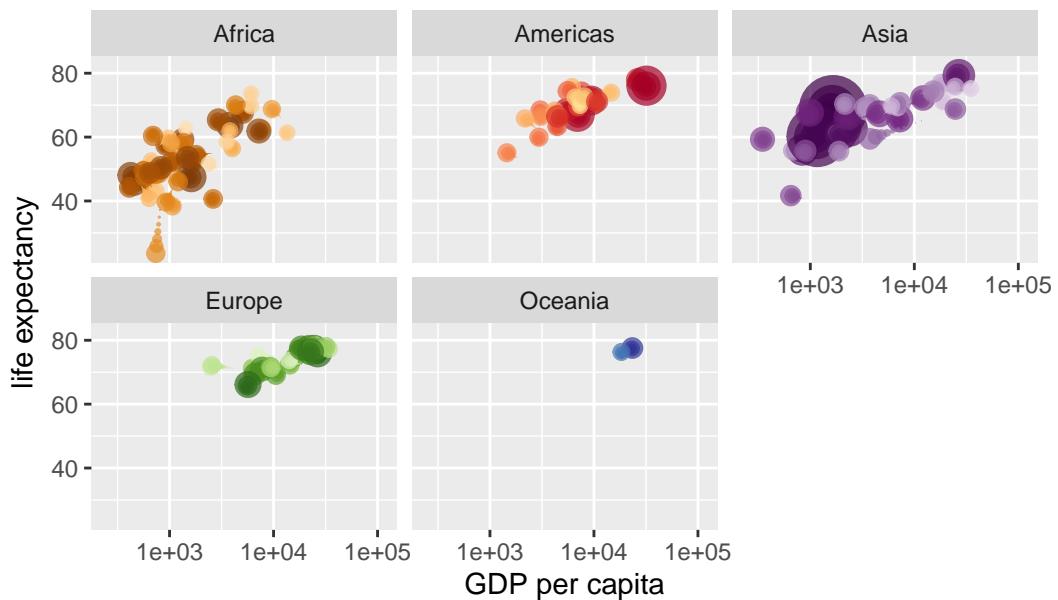
Year: 1991



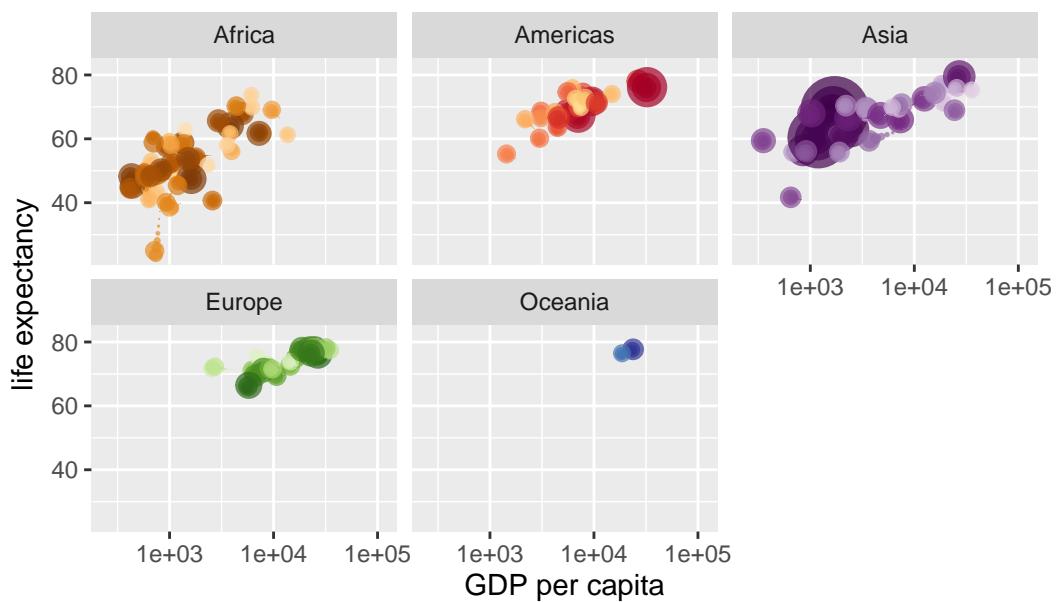
Year: 1991



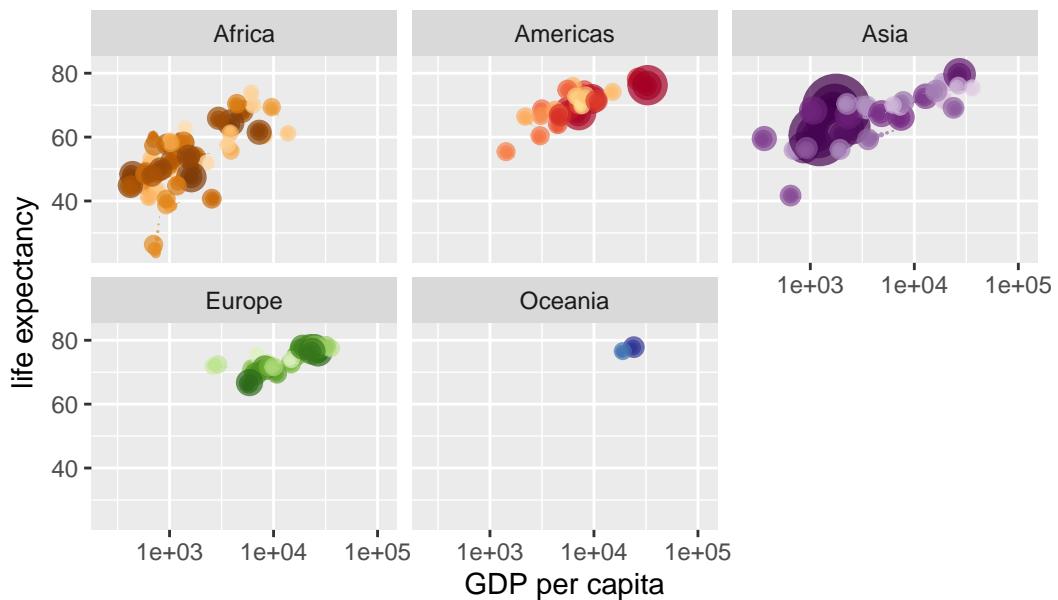
Year: 1992



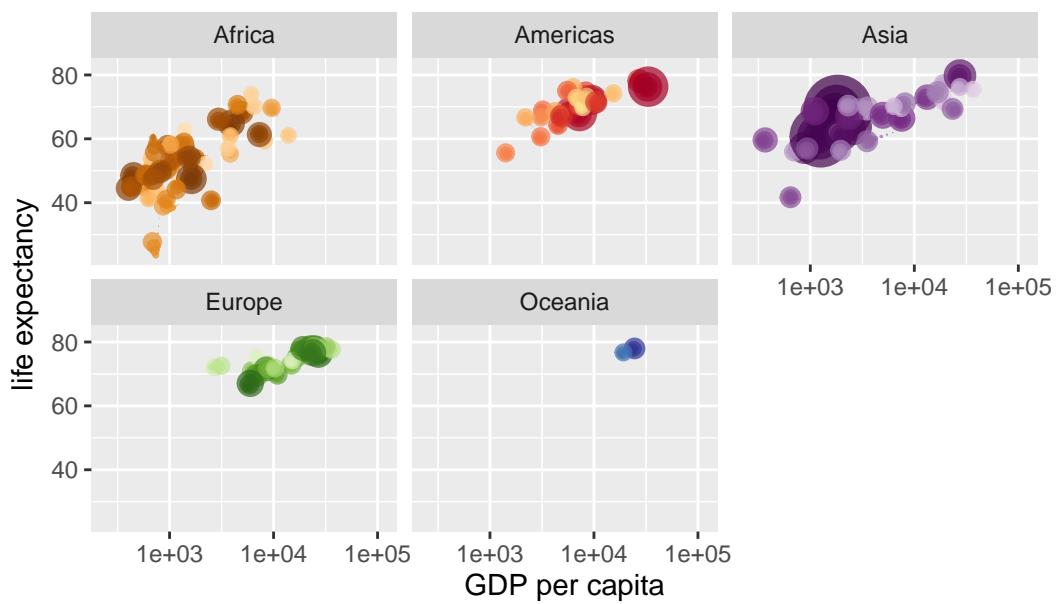
Year: 1993



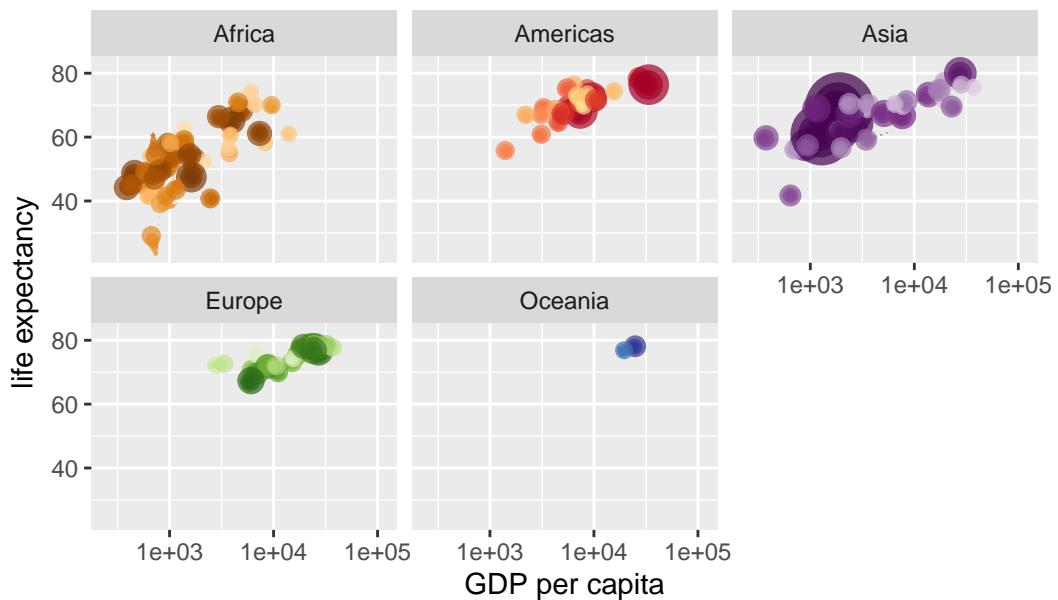
Year: 1993



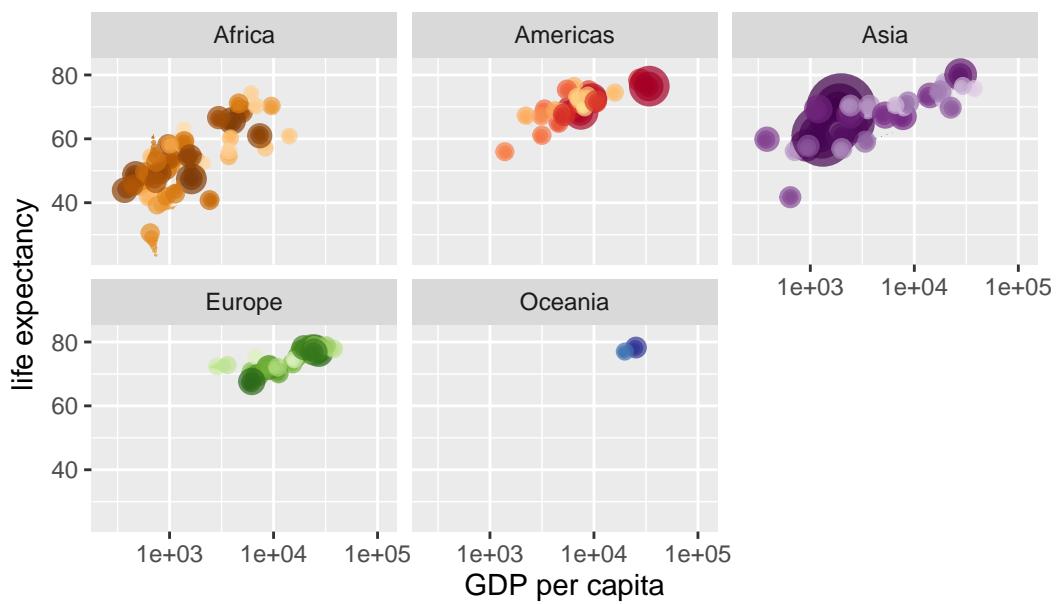
Year: 1994



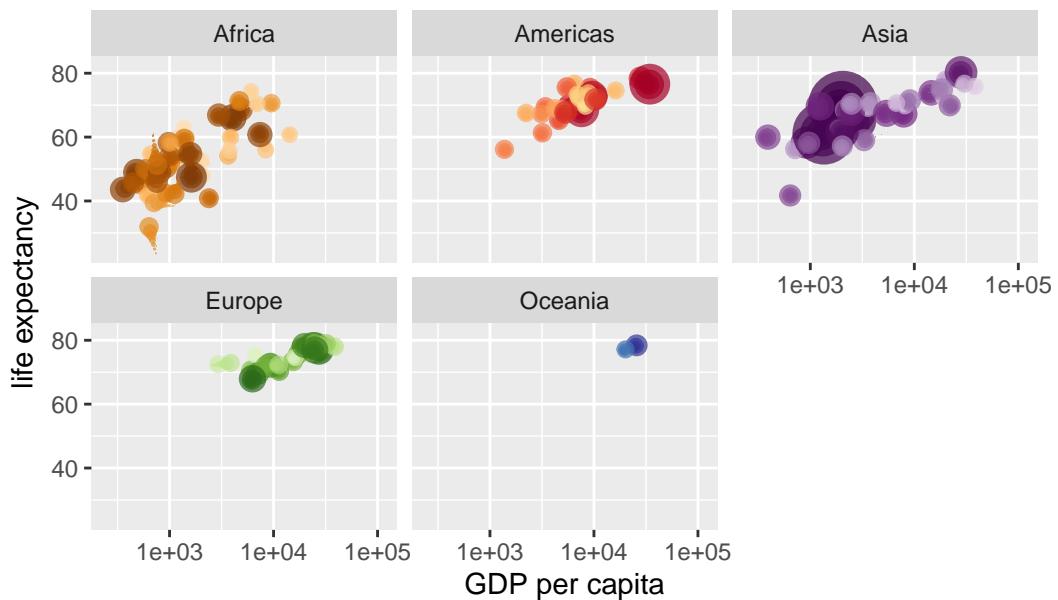
Year: 1994



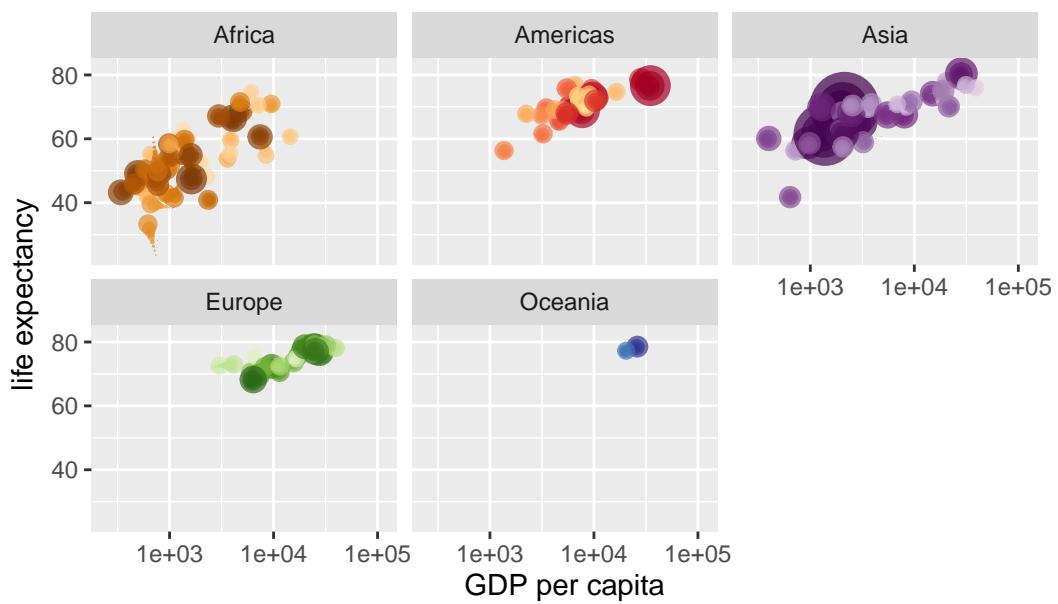
Year: 1995



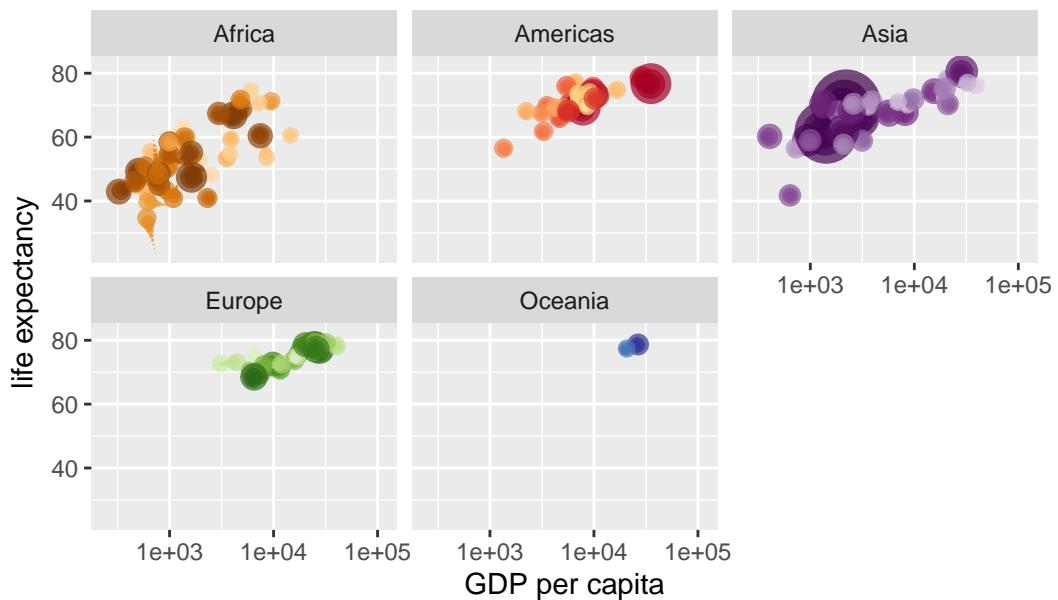
Year: 1995



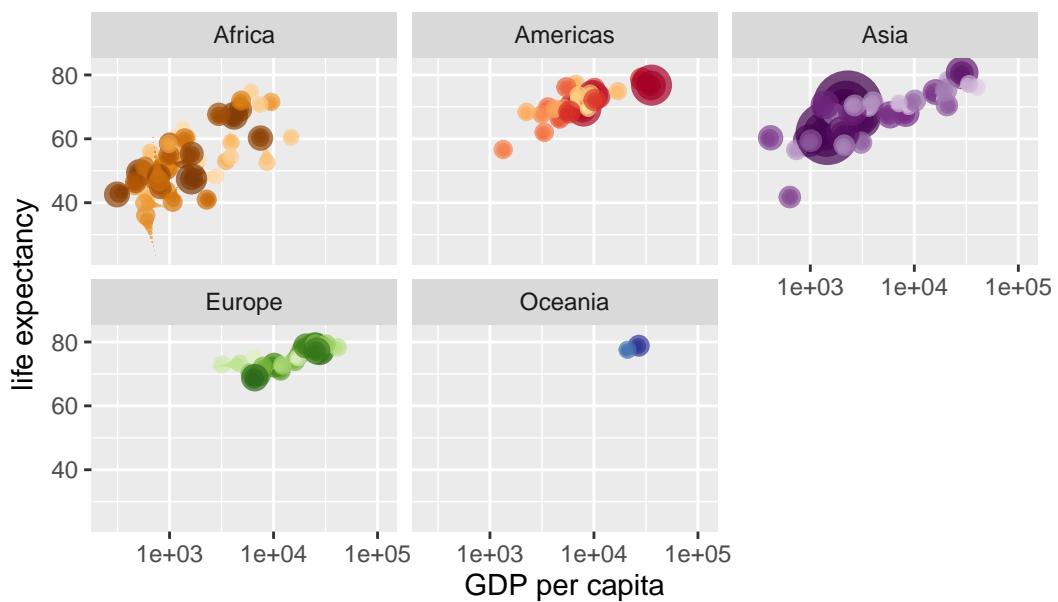
Year: 1996



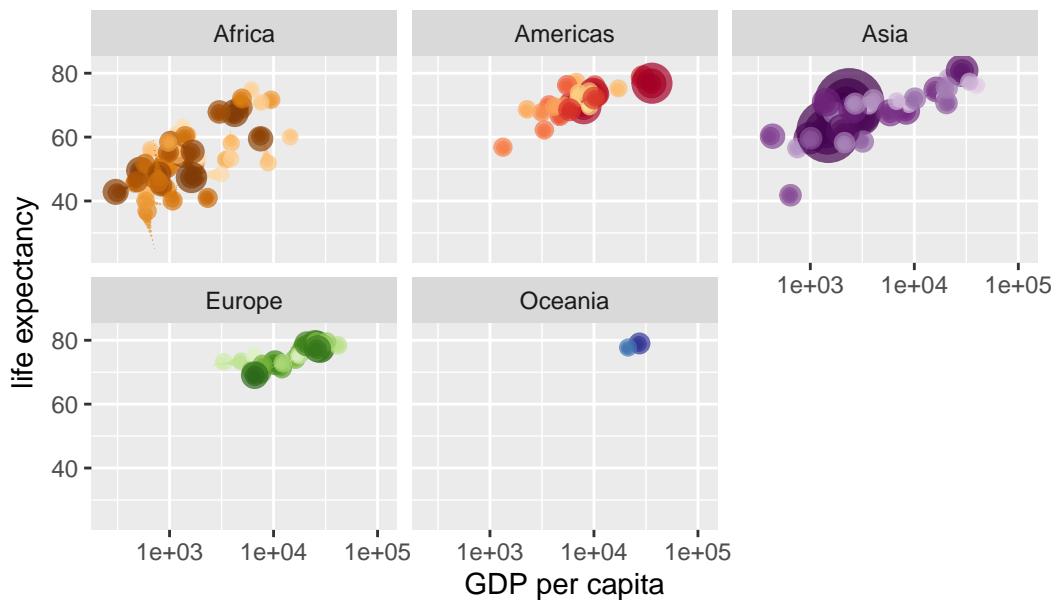
Year: 1996



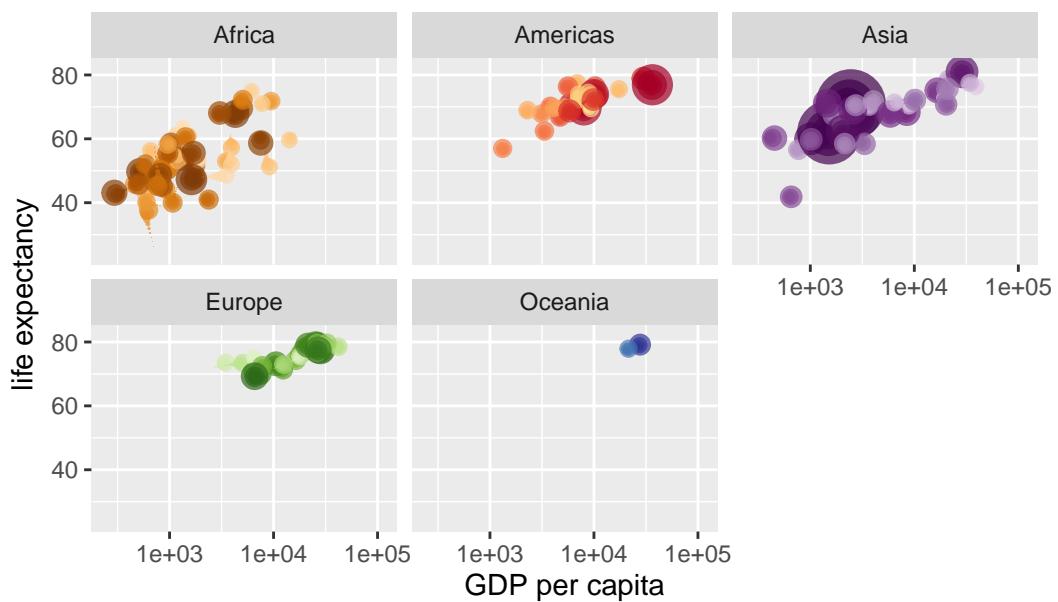
Year: 1997



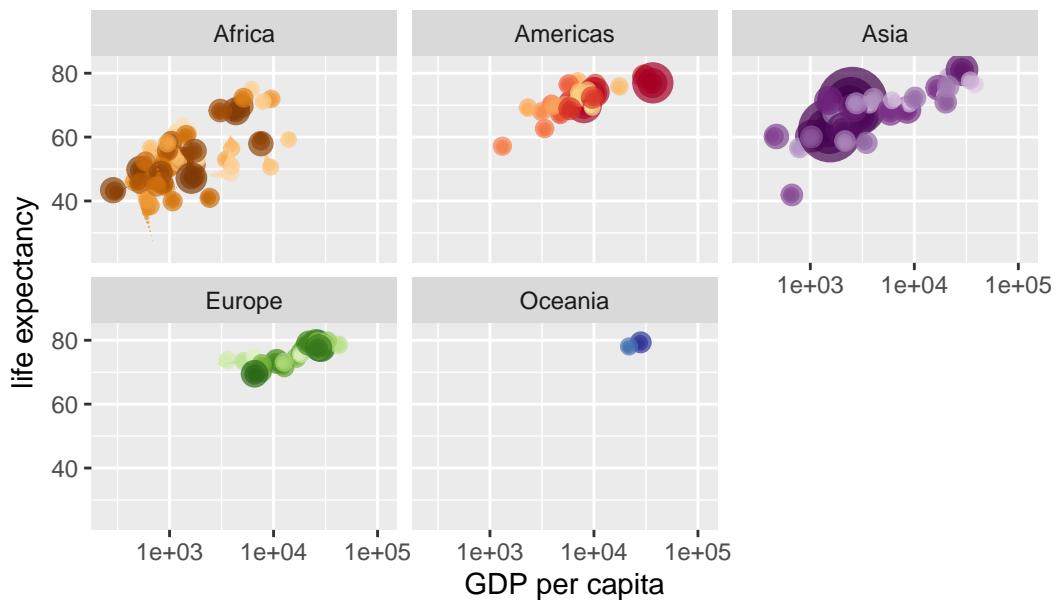
Year: 1998



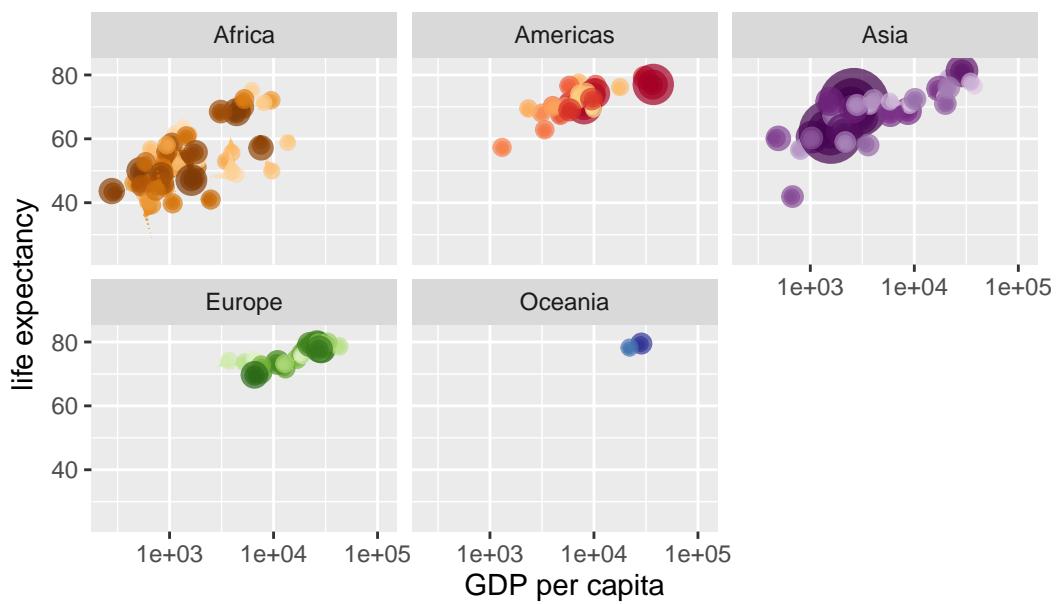
Year: 1998



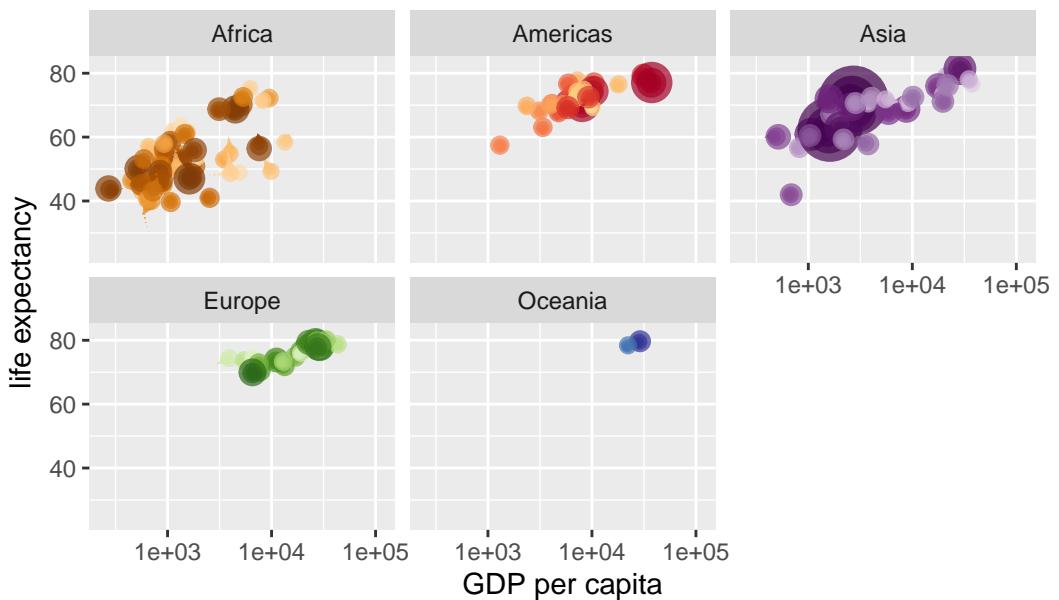
Year: 1999



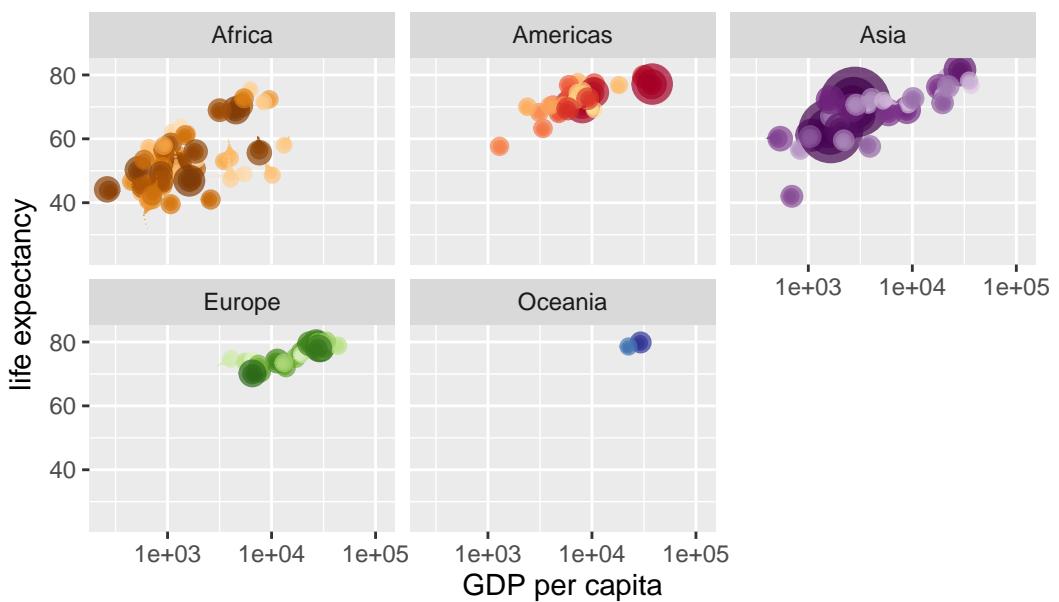
Year: 1999



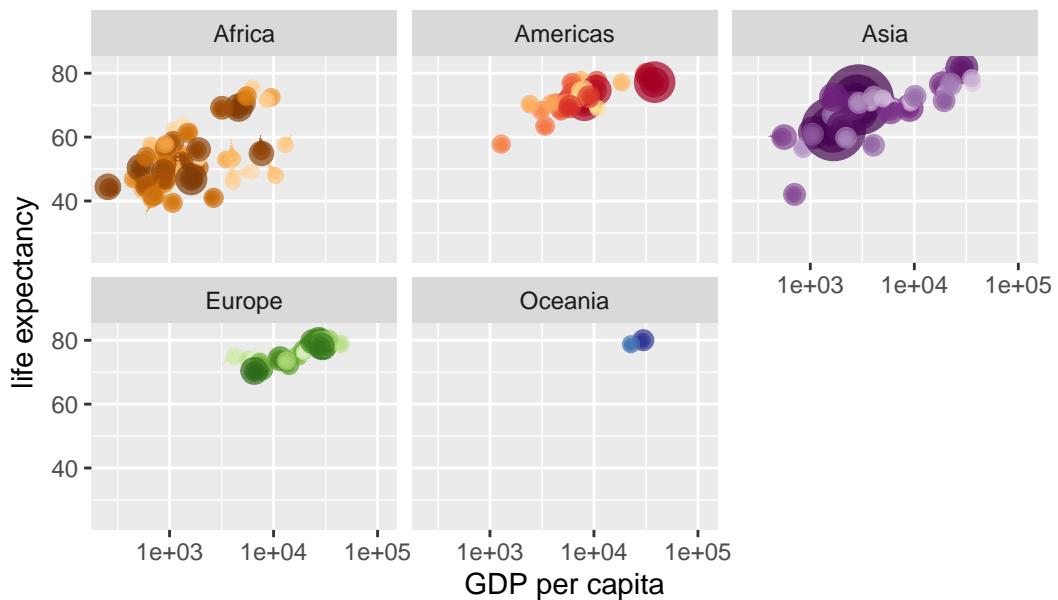
Year: 2000



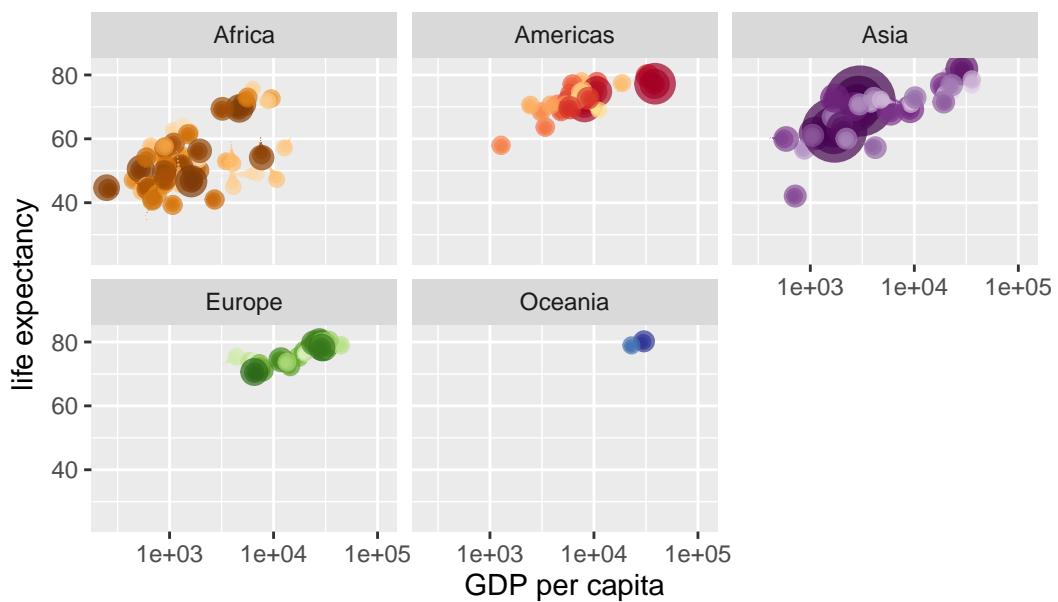
Year: 2000



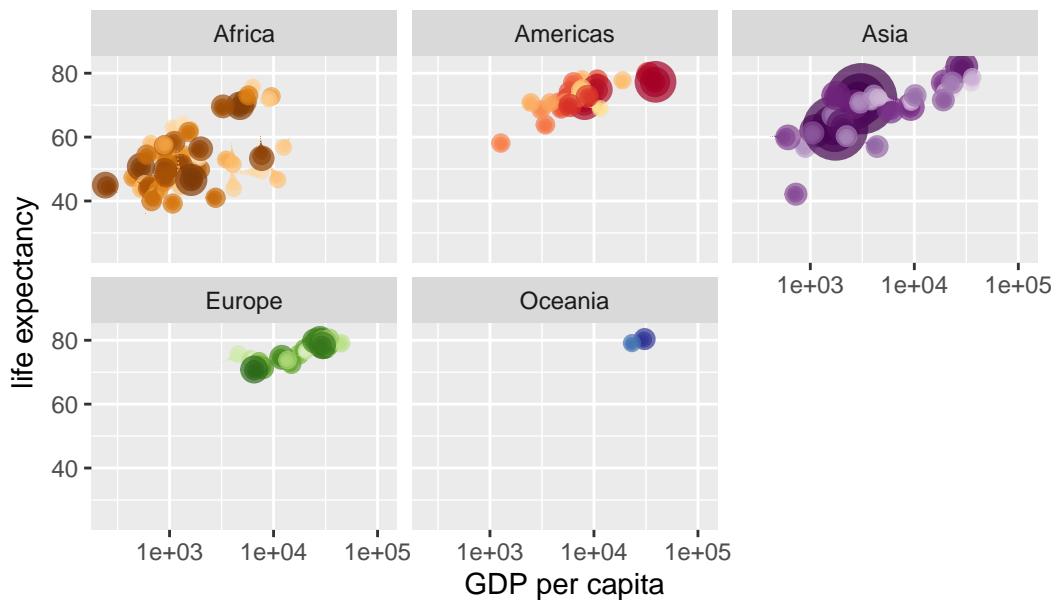
Year: 2001



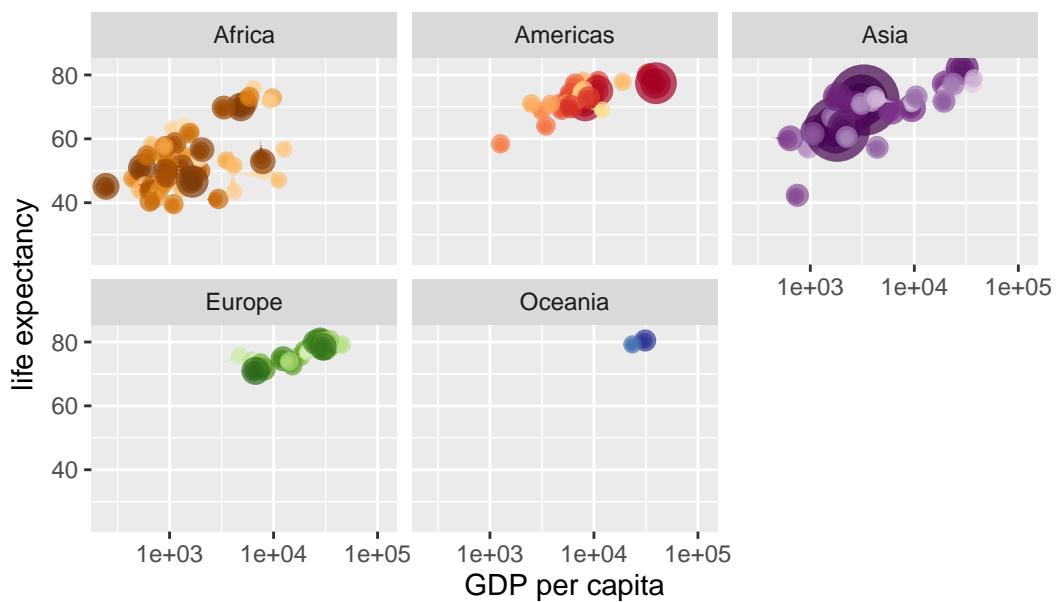
Year: 2001



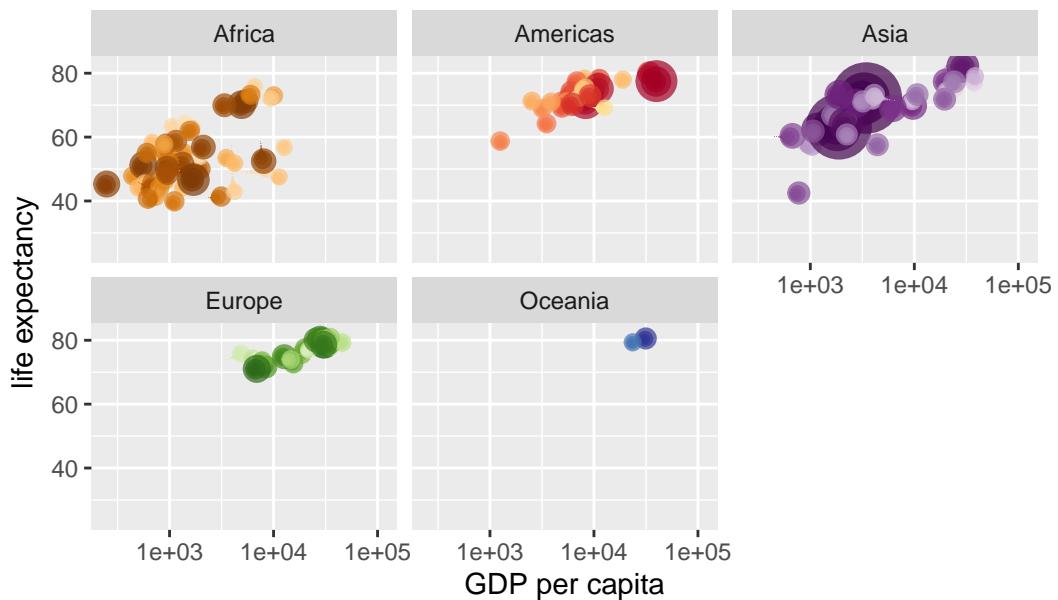
Year: 2002



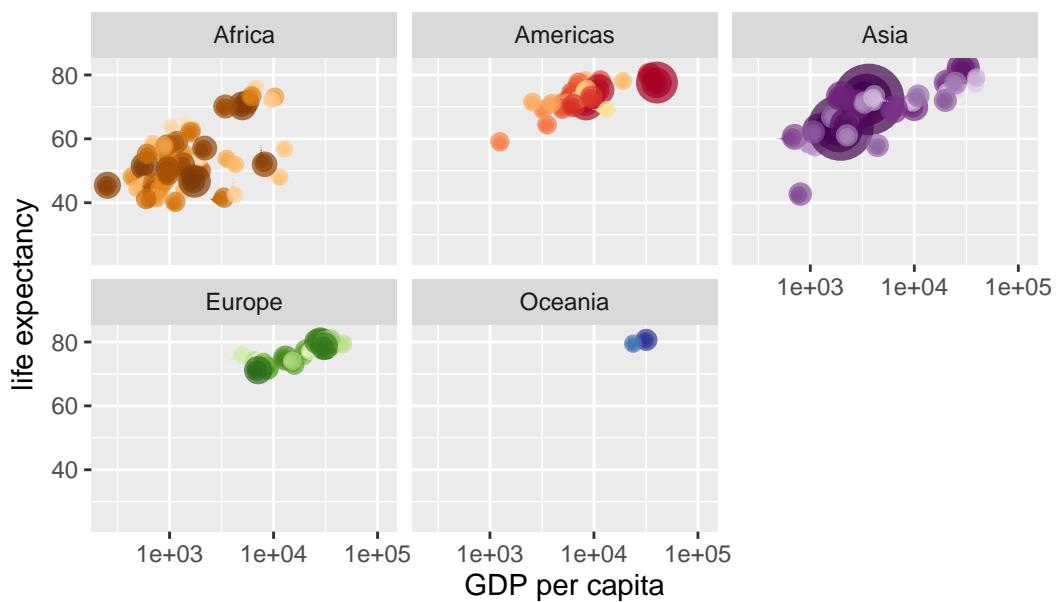
Year: 2003



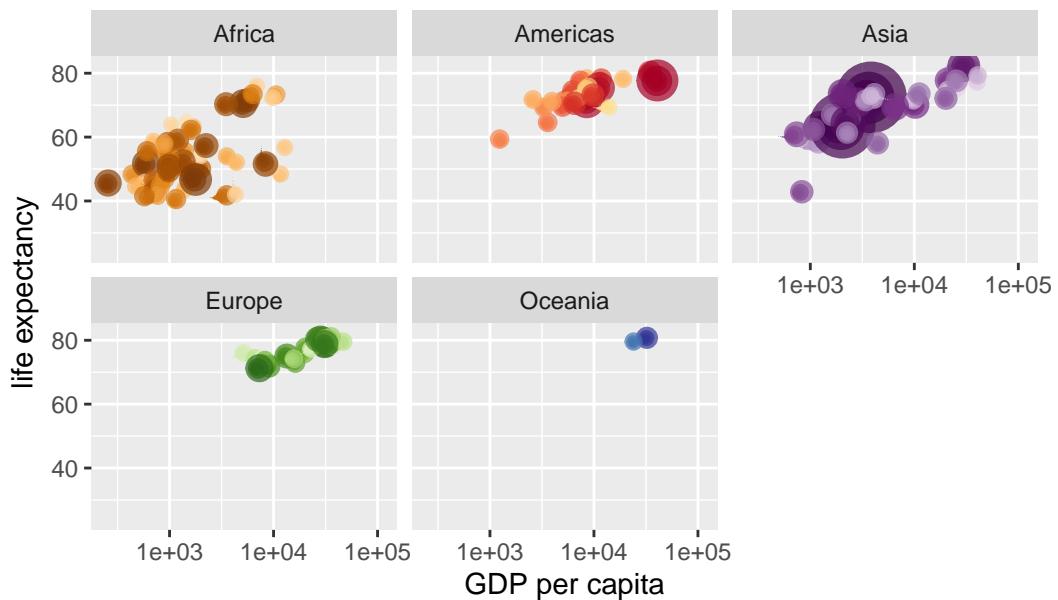
Year: 2003



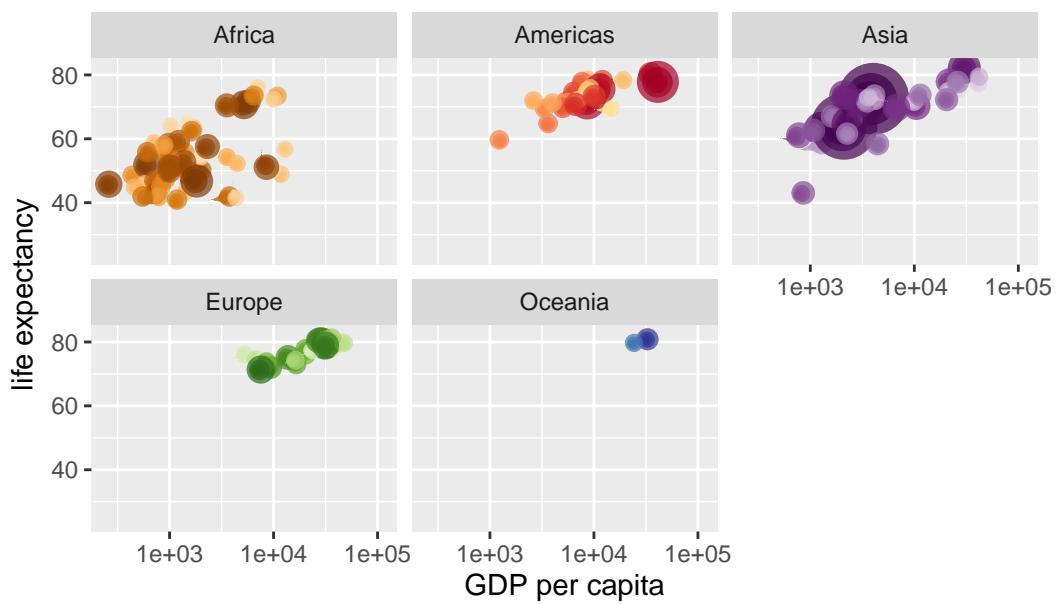
Year: 2004



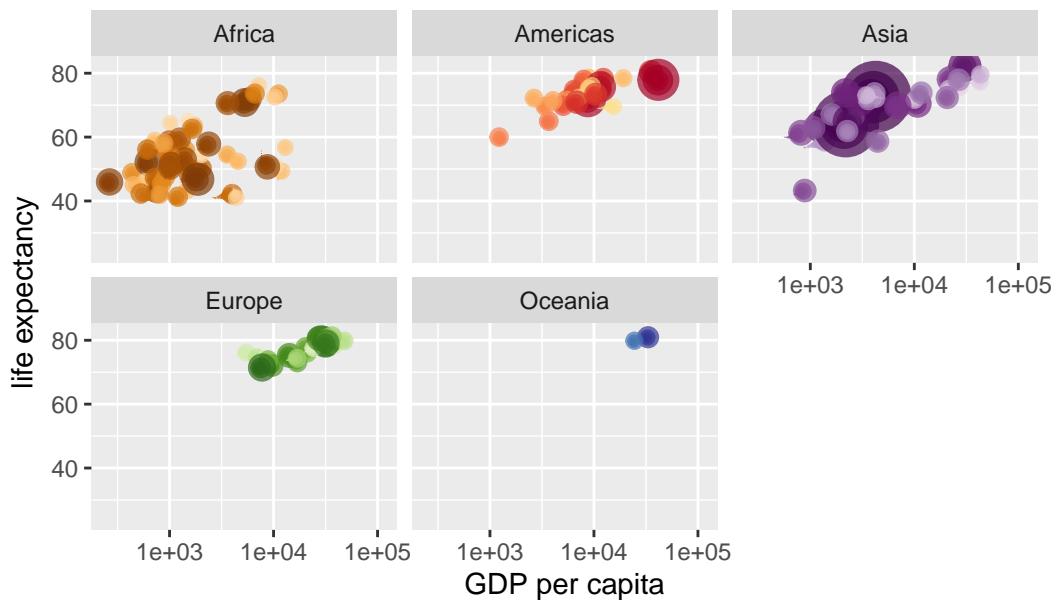
Year: 2004



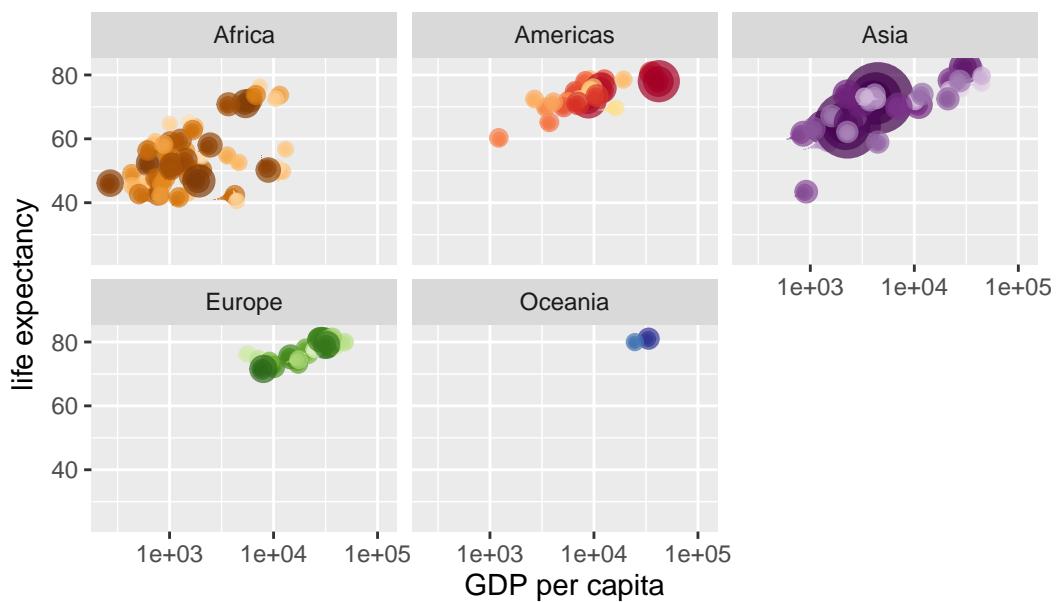
Year: 2005



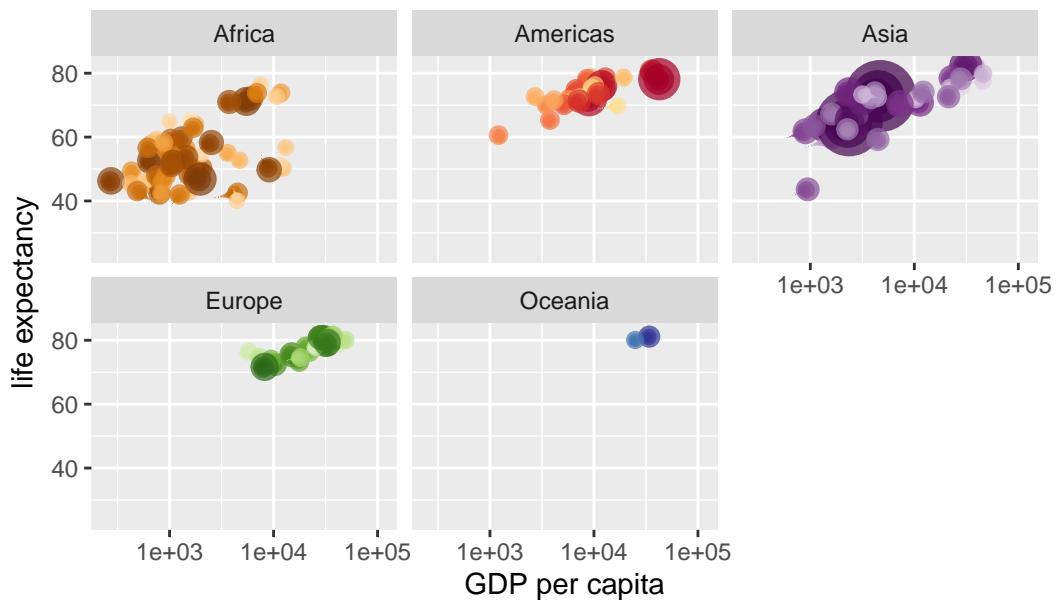
Year: 2005



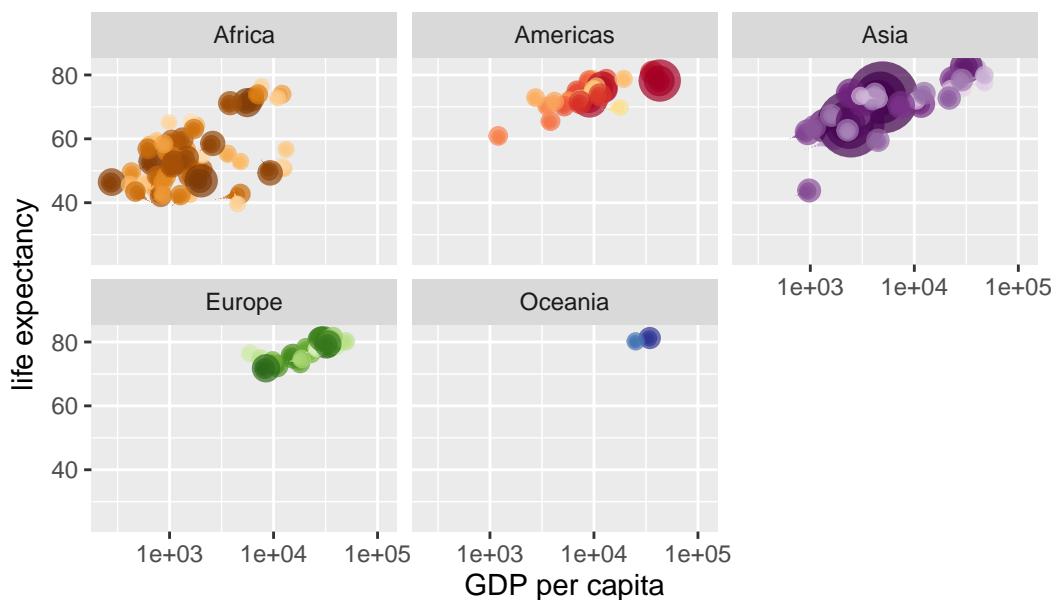
Year: 2006



Year: 2006



Year: 2007



NOW LET'S COMBINE PLOTS Another excellent package extending ggplot is patchwork that is useful for combining plots to make an all-in-one multi-panel figure. For example:

```

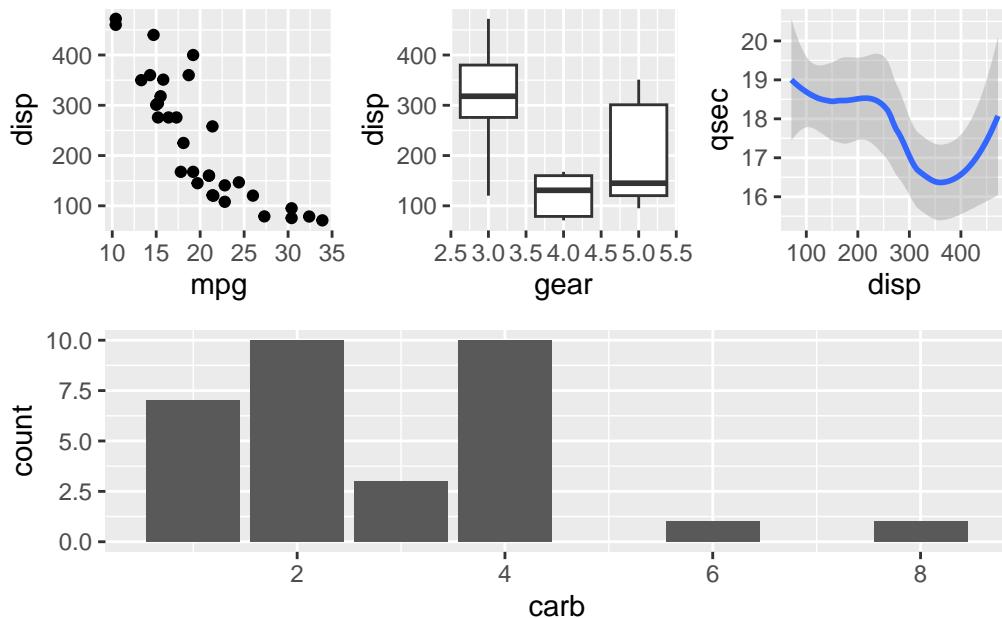
library(patchwork)

# 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'



to save use ggsave("title.pdf")