

# Data visualization with ggplot2

Joshua Cheung (PID: A15441585)

## Section 1: Overview

No questions or code given for this section

## Section 2: Background

**Q.** For which phases is data visualization important in our scientific workflows?

All of the above

**Q.** True or False? The ggplot2 package comes already installed with R?

False

## Section 3: Getting Organized

No questions or code given for this section

## Section 4: Common Plot Types

**Q.** Which plot types are typically NOT used to compare distributions of numeric variables?

Network graphs

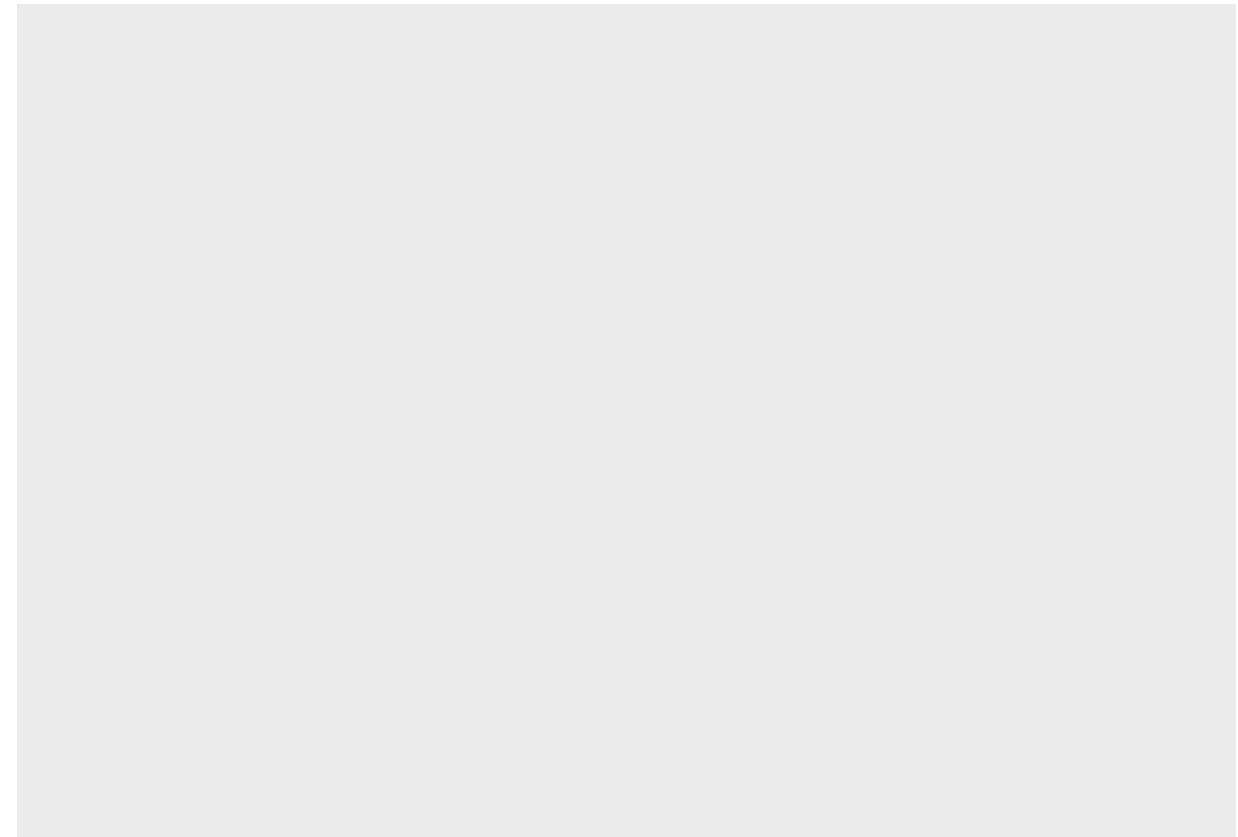
**Q.** Which statement about data visualization with ggplot2 is incorrect?

ggplot2 is the only way to create plots in R

## Section 5: Creating Scatter Plots

Specifying a dataset with `ggplot()`

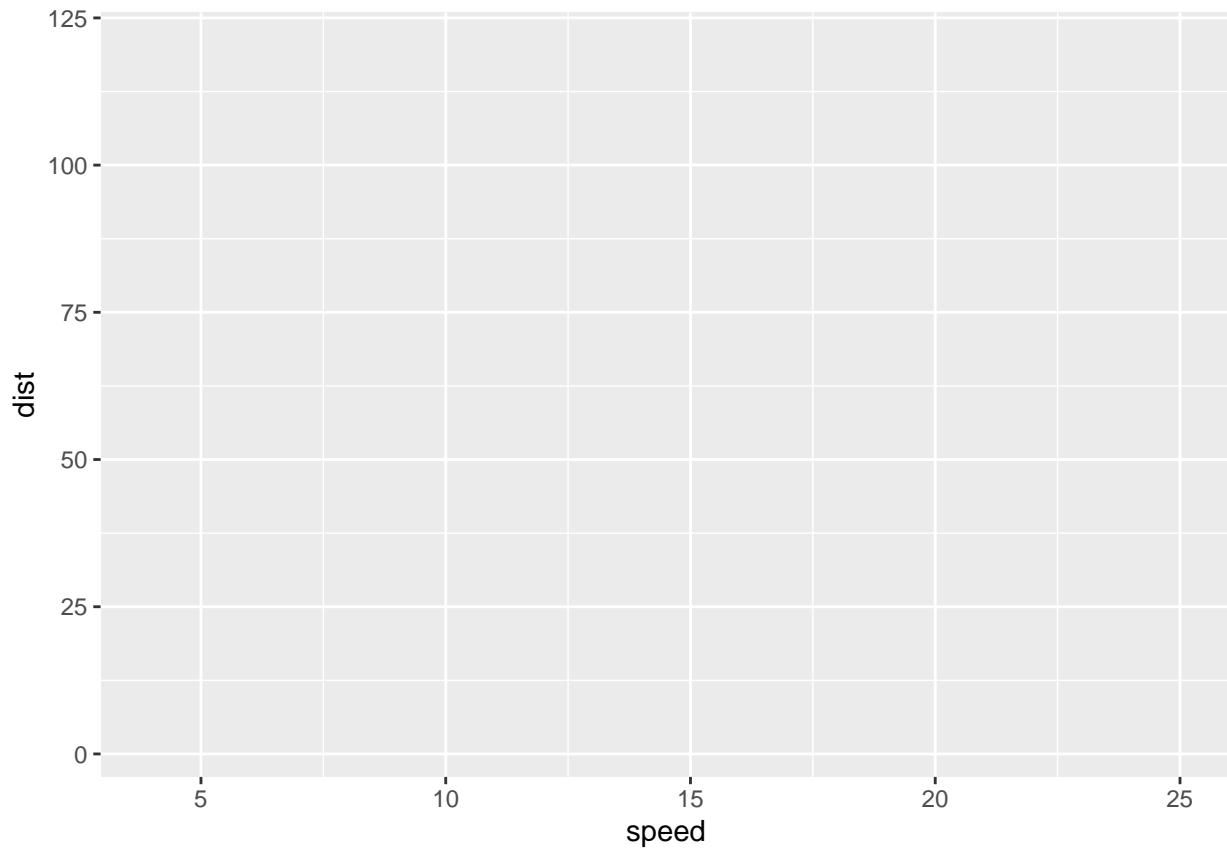
```
# install.packages("ggplot2") ## un-comment if necessary
library(ggplot2)
ggplot(cars)
```



## Specifying aesthetic mappings with aes()

Now set the x and y axis coordinates

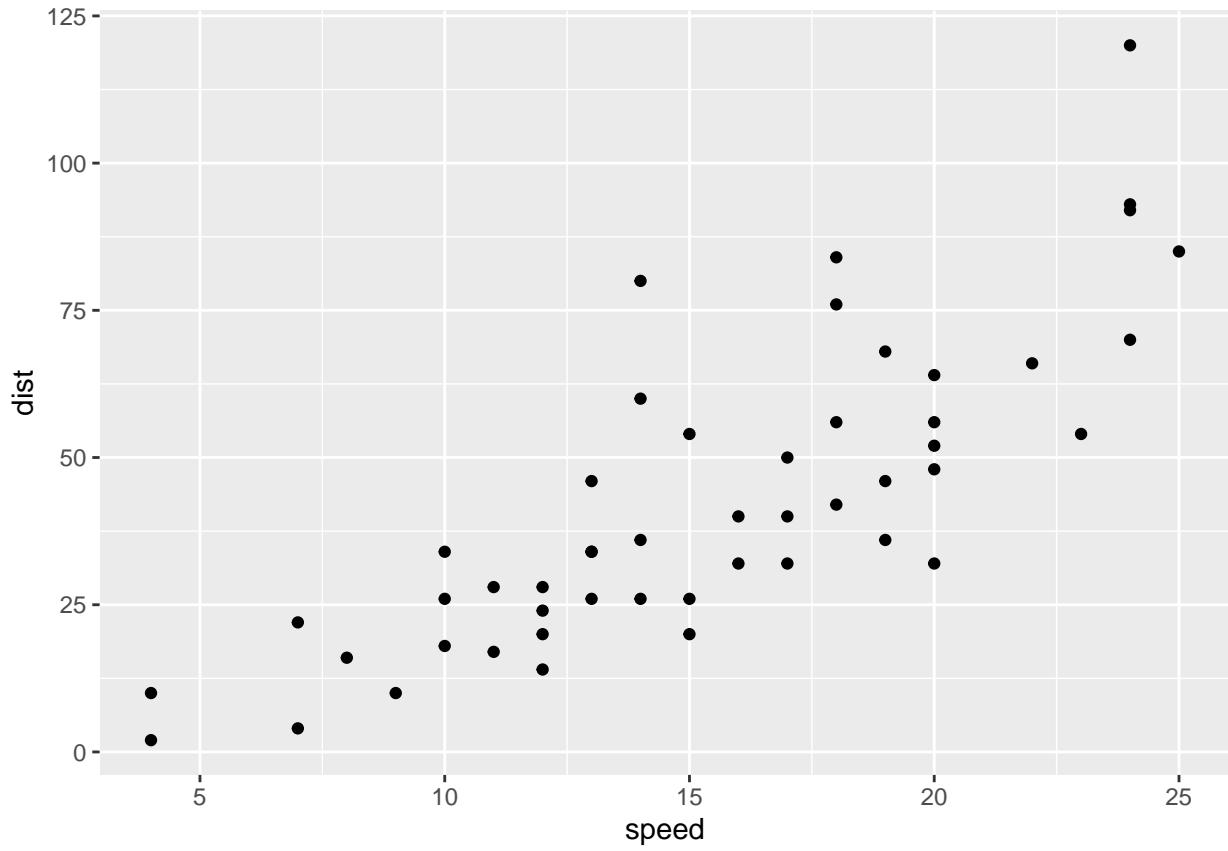
```
ggplot(cars) +  
  aes(x=speed, y=dist)
```



### Specifying a geom layer with geom\_point()

Add geom\_point() to get a point layer

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



**Q.** Which geometric layer should be used to create scatter plots in ggplot2?

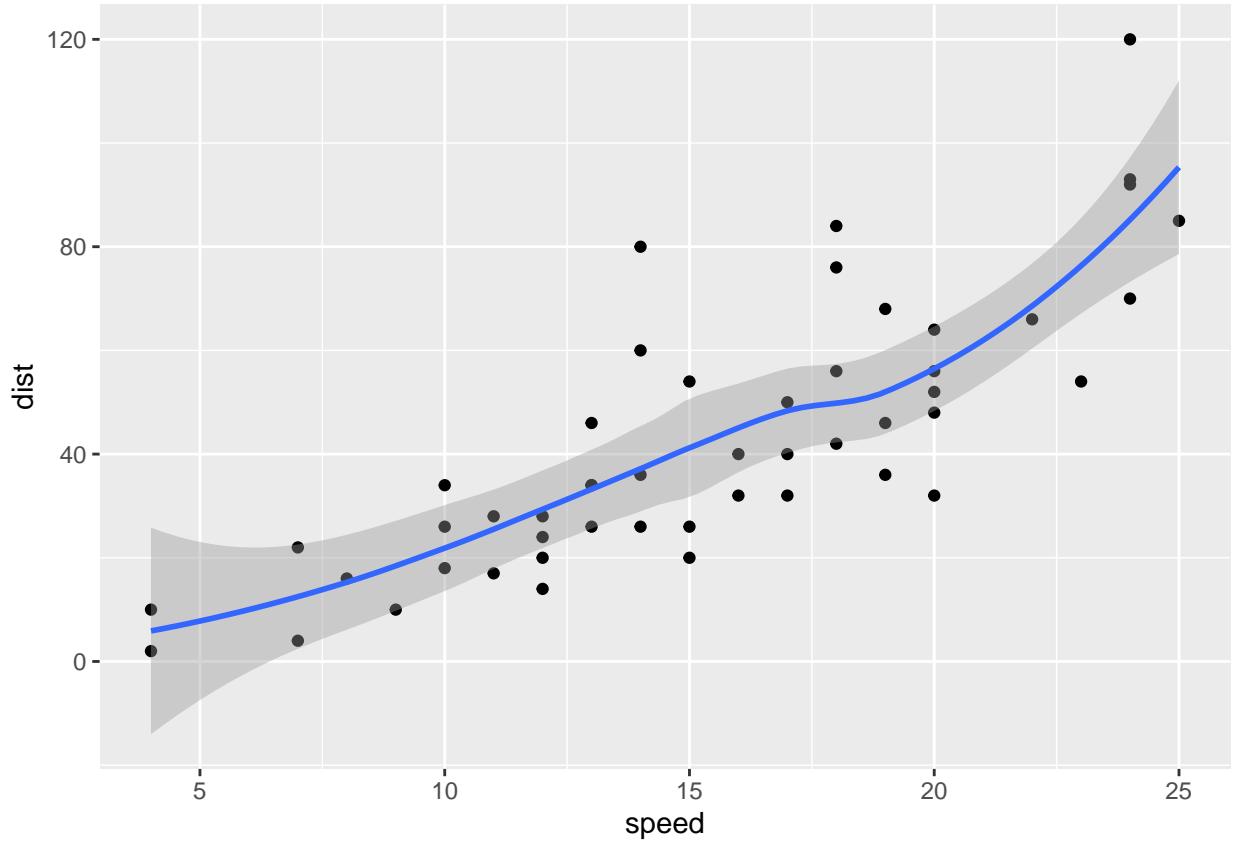
`geom_point()`

**Q.** In your own RStudio can you add a trend line layer to help show the relationship between the plot variables with the `geom_smooth()` function?

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

p + geom_smooth()

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

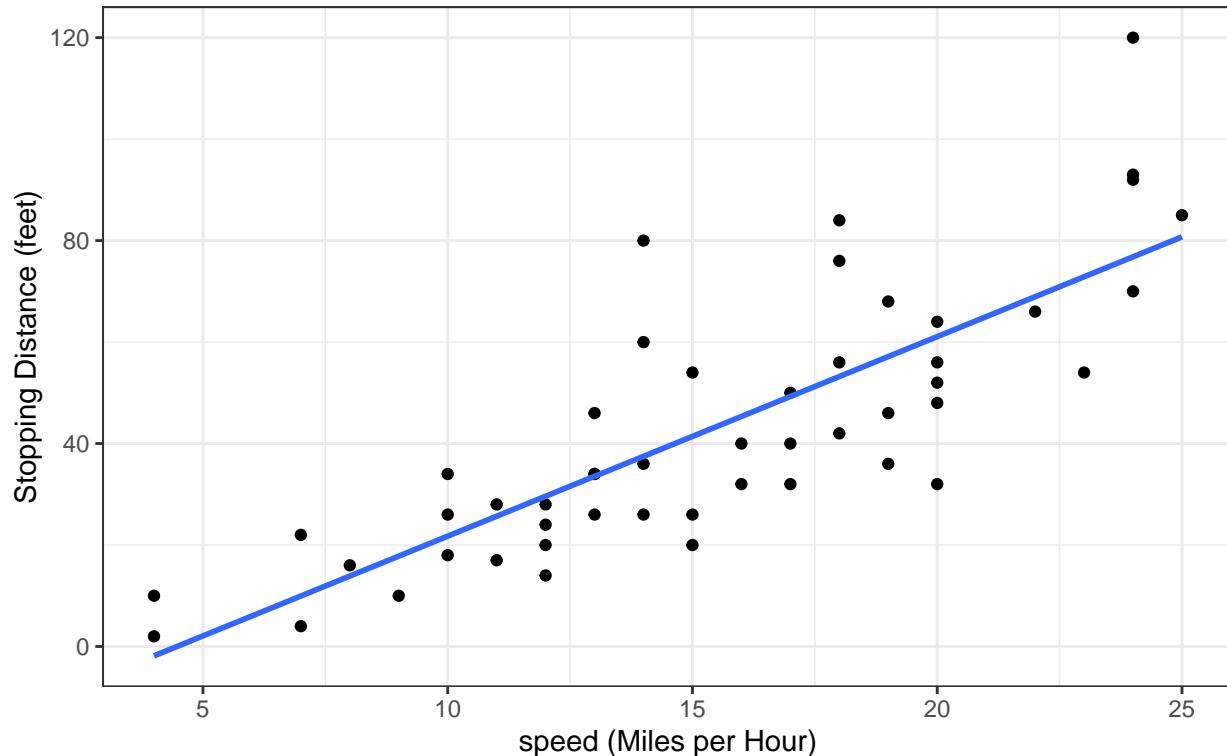


Q. Can you finish this plot by adding various label annotations with the labs() function and changing the plot look to a more conservative “black & white” theme by adding the theme\_bw() function:

```
p + labs(title = "Speed and Stopping distances of Cars",
         x = "speed (Miles per Hour)",
         y = "Stopping Distance (feet)",
         caption = "Dataset Source: 'cars'") +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```

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

## Speed and Stopping distances of Cars



Dataset Source: 'cars'

## Adding more plot aesthetics through aes()

```
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
```

**Q.** Use the nrow() function to find out how many genes are in this dataset. What is your answer?

There are 5196 genes in this dataset.

```
nrow(genes)
```

```
## [1] 5196
```

**Q.** Use the `colnames()` function and the `ncol()` function on the `genes` data frame to find out what the column names are (we will need these later) and how many columns there are. How many columns did you find?

There are 4 columns.

```
colnames(genes)

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

ncol(genes)

## [1] 4
```

**Q.** Use the `table()` function on the `State` column of this `data.frame` to find out how many ‘up’ regulated genes there are. What is your answer?

There are 127 up-regulated genes.

```
table(genes$State)

##
##      down unchanged      up
##      72        4997     127
```

**Q.** Using your values above and 2 significant figures. What fraction of total genes is up-regulated in this dataset?

The fraction of the total genes that are up-regulated in this data set rounded at two significant figures is 2.44

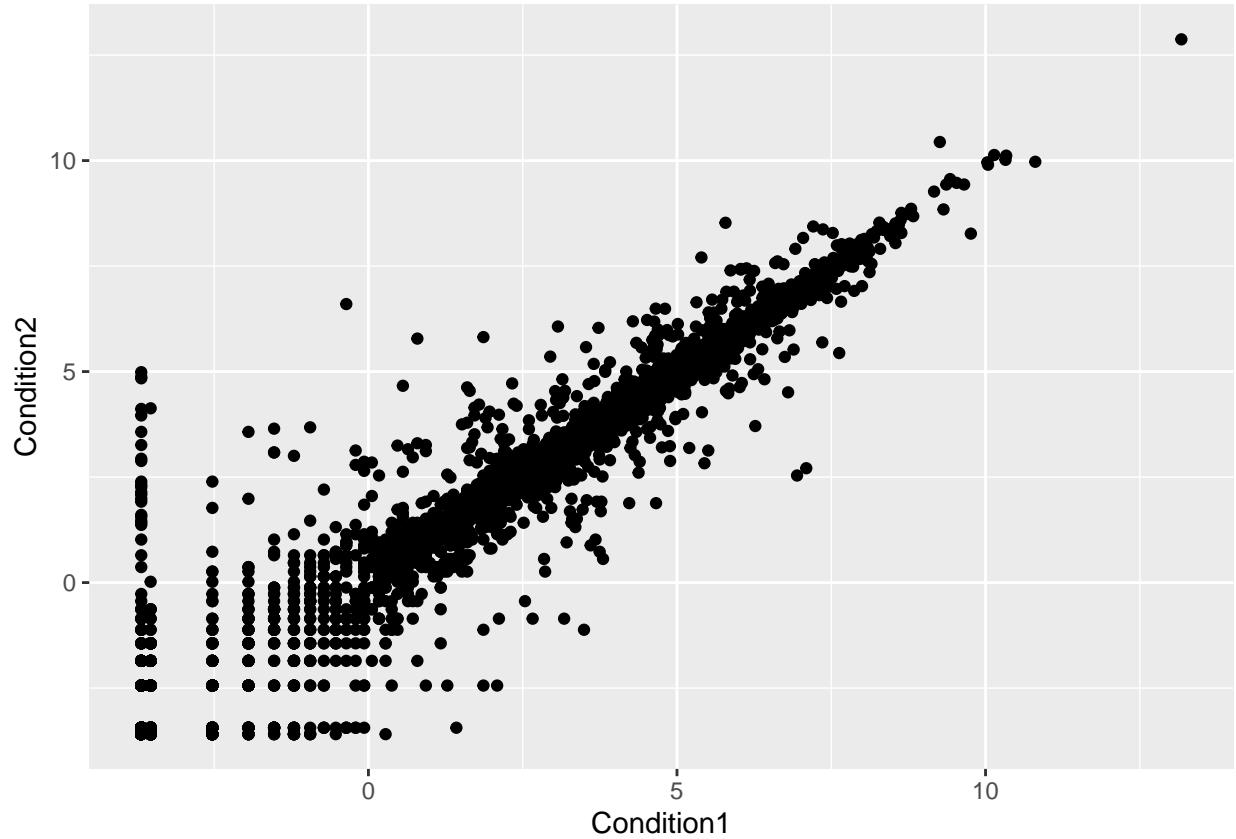
```
q <- table(genes$State)/nrow(genes)*100
round(q, 2)
```

```
##
##      down unchanged      up
##      1.39        96.17    2.44
```

**Q.** Complete the code below to produce the following plot

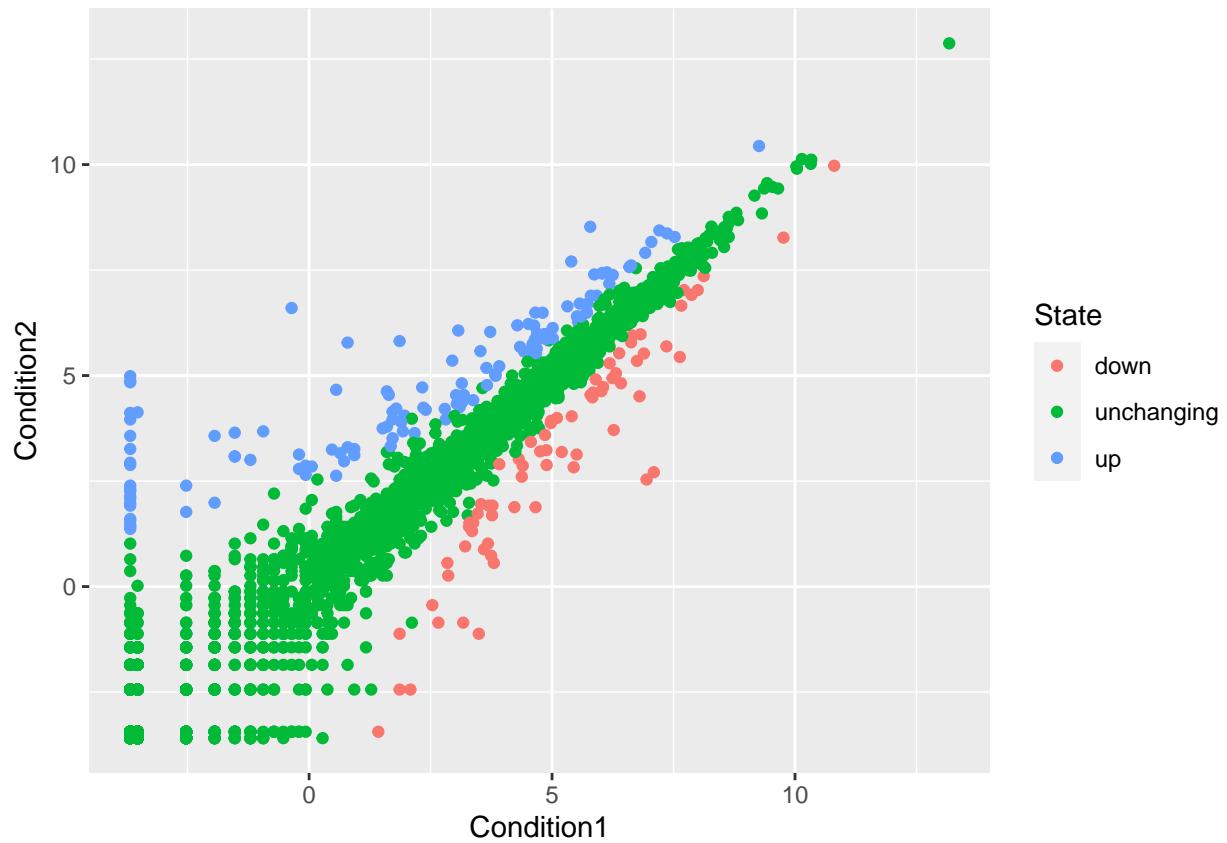
The completed code is:

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



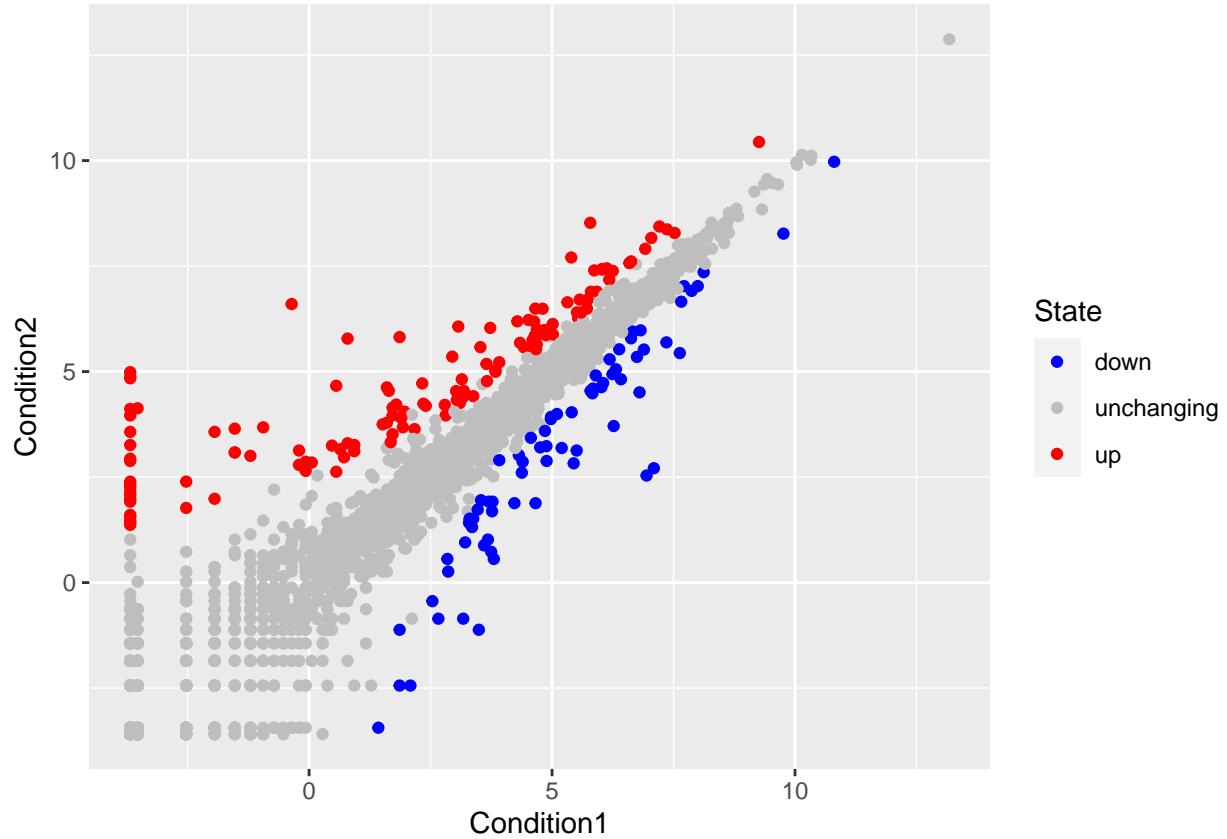
Map the state column to point color

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



Change the colors to blue, gray, red.

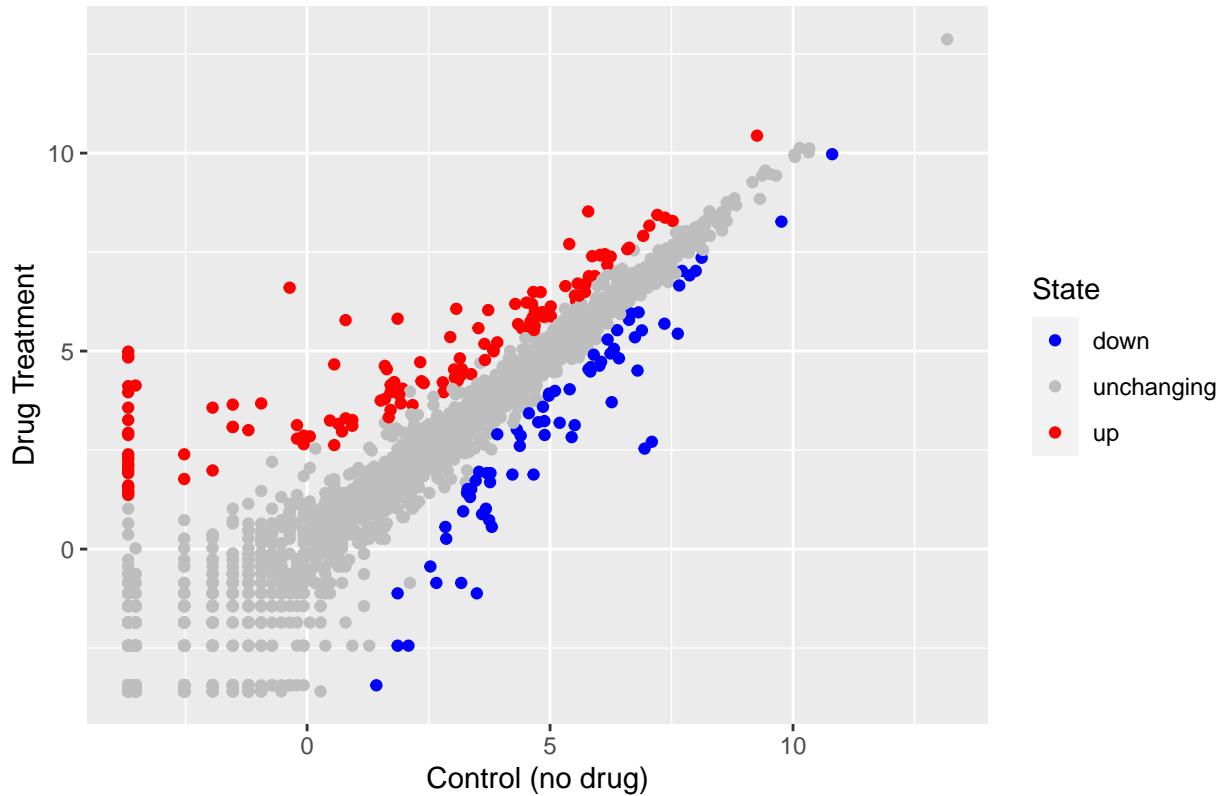
```
p + scale_colour_manual( values=c("blue","gray","red") )
```



Q. Nice, now add some plot annotations to the p object with the labs() function so your plot looks like the following:

```
r <- ggplot(genes) +
  aes(x=Condition1, y=Condition2, colour=State) +
  geom_point()
r + 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 6 OPTIONAL: Going Further

```
# install.packages("gapminder") ## un-comment to install if necessary
# install.packages("dplyr") ## un-comment to install if necessary
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

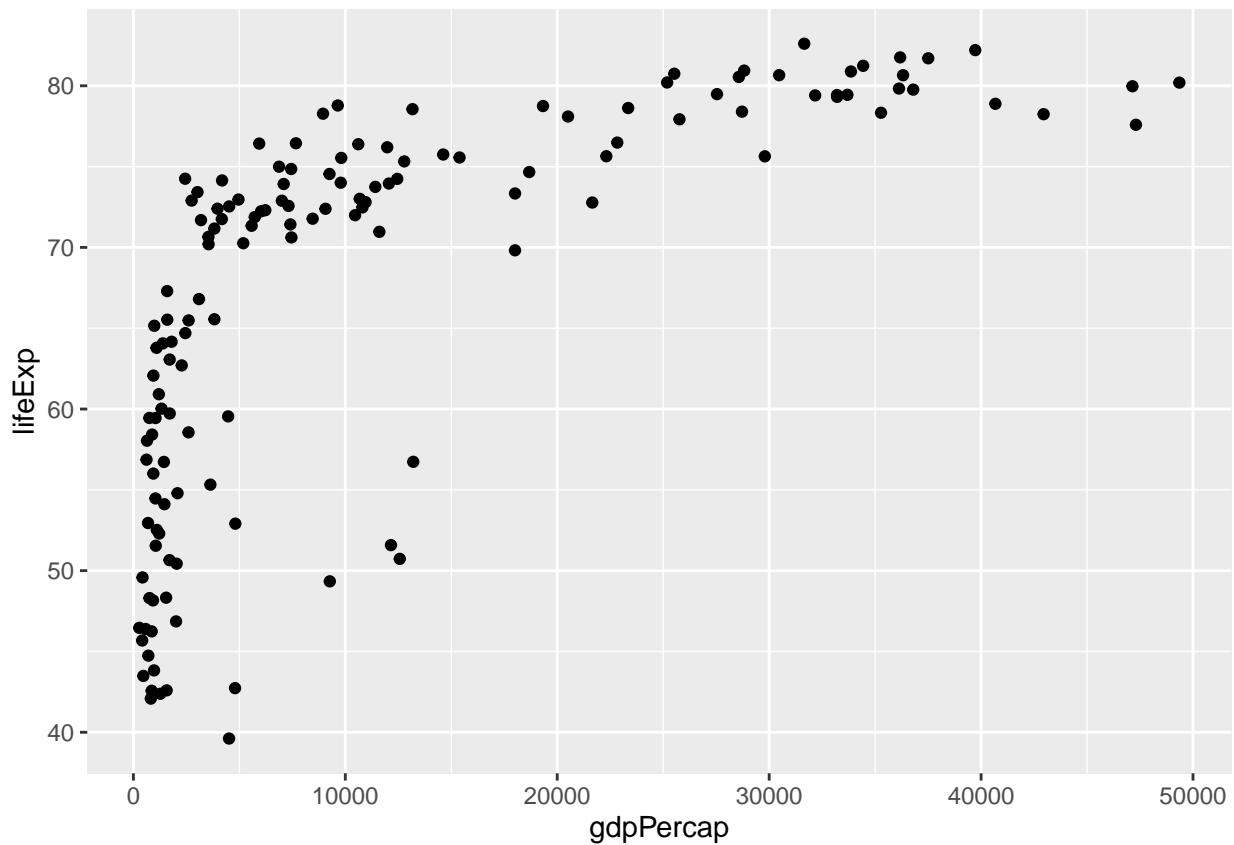
gapminder_2007 <- gapminder%>%filter(year==2007)
```

Let's consider the gapminder\_2007 dataset which contains the variables GDP per capita gdp-Percap and life expectancy lifeExp for 142 countries in the year 2007

Q. Complete the code below to produce a first basic scatter plot of this gapminder\_2007 dataset:

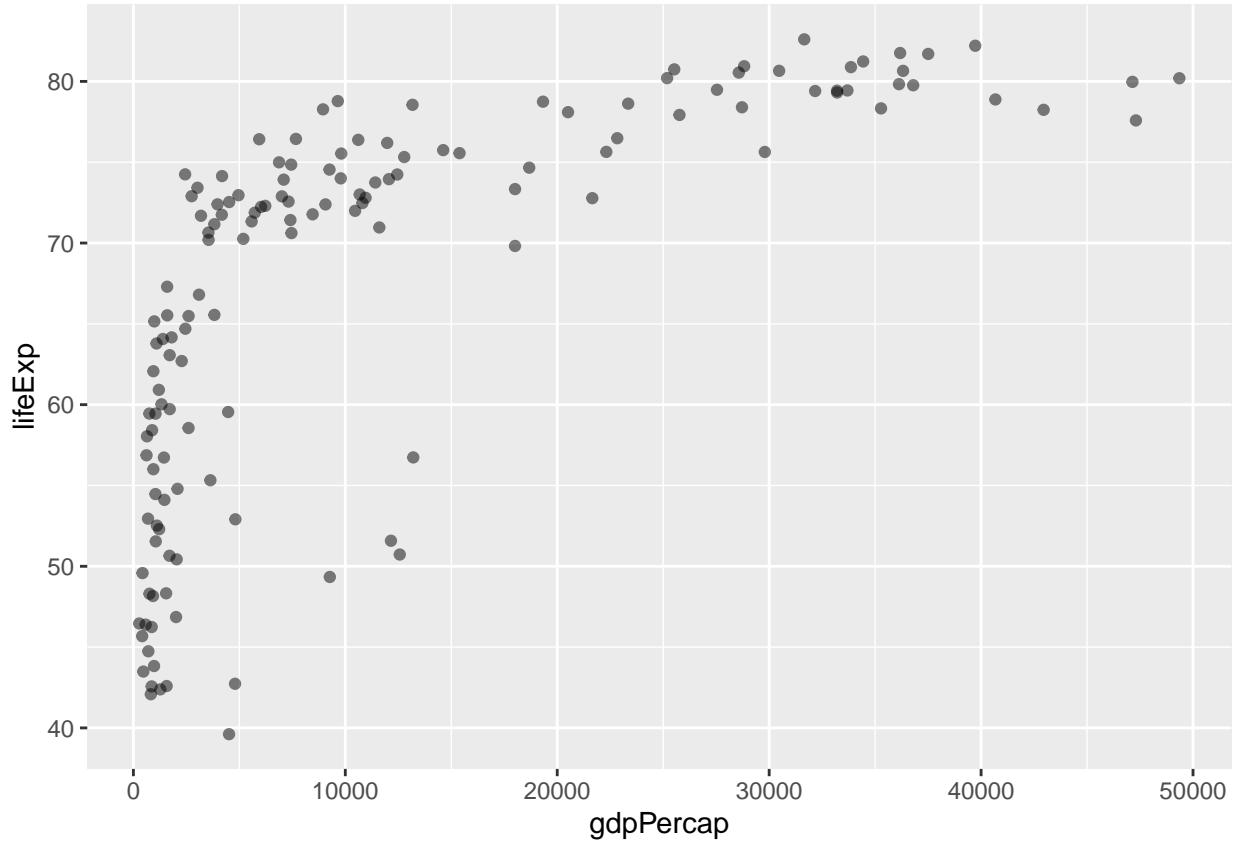
The completed code is:

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



Now make the points slightly transparent (set alpha=0.5)

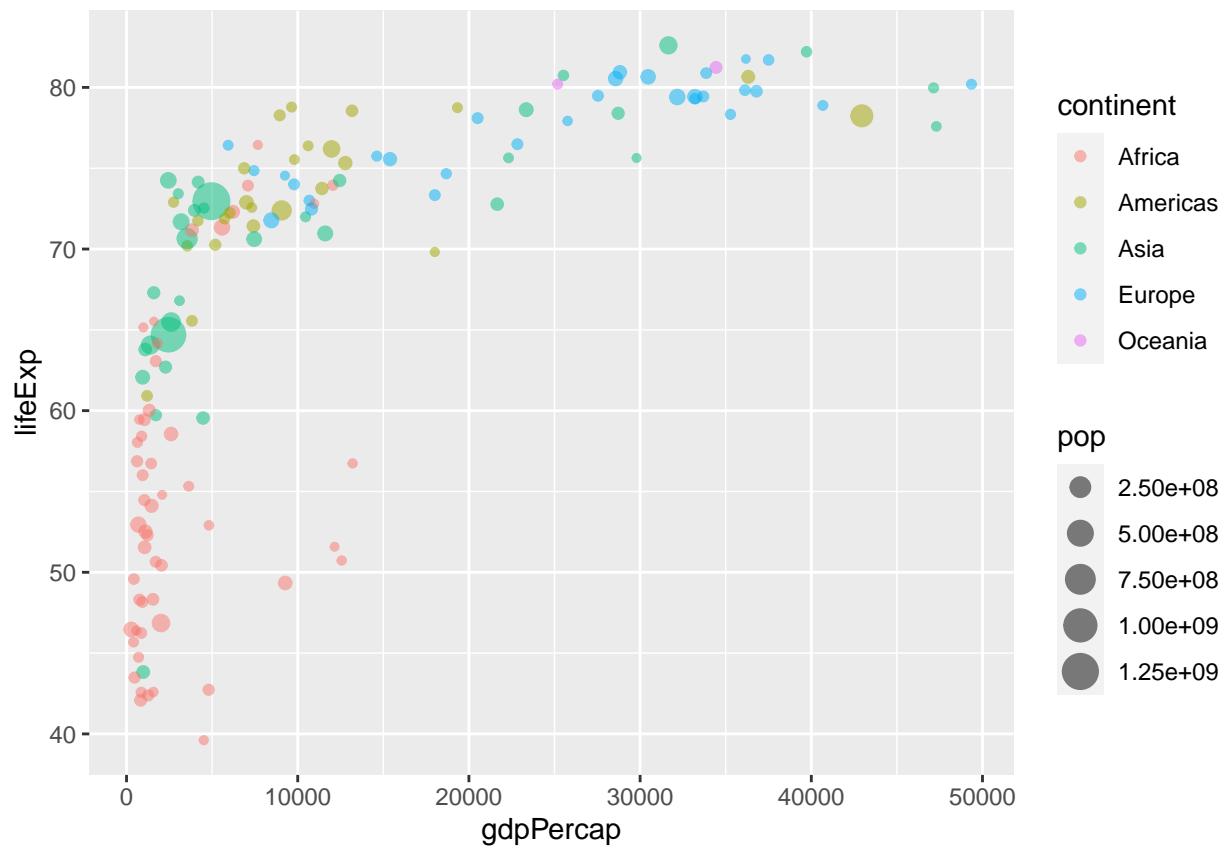
```
ggplot(gapminder_2007) +  
  aes(x=gdpPercap, y=lifeExp) +  
  geom_point(alpha=0.5)
```



### Adding more variables to aes()

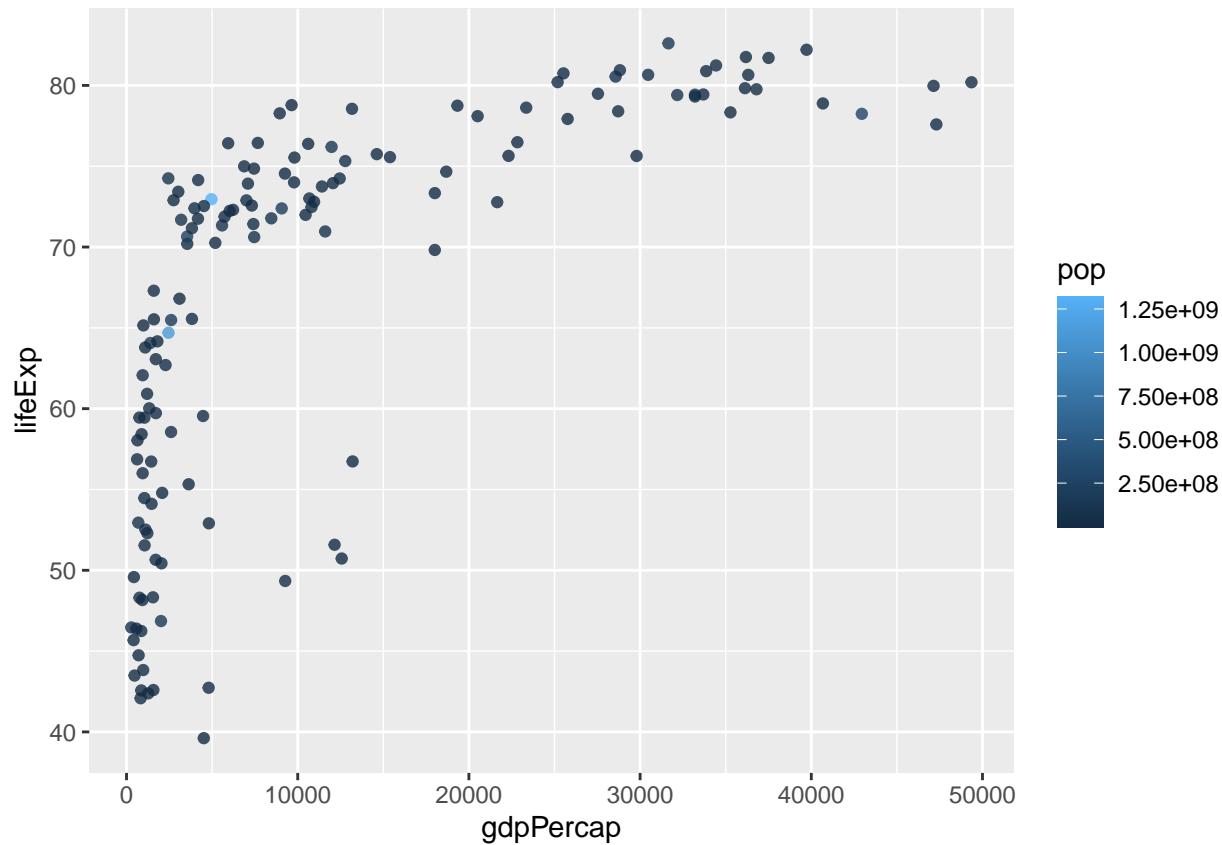
Now map the continent variable to the point color aesthetic and population through the point size argument of aes().

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



Now change the plot so that the points are colored by the numeric variable: population.

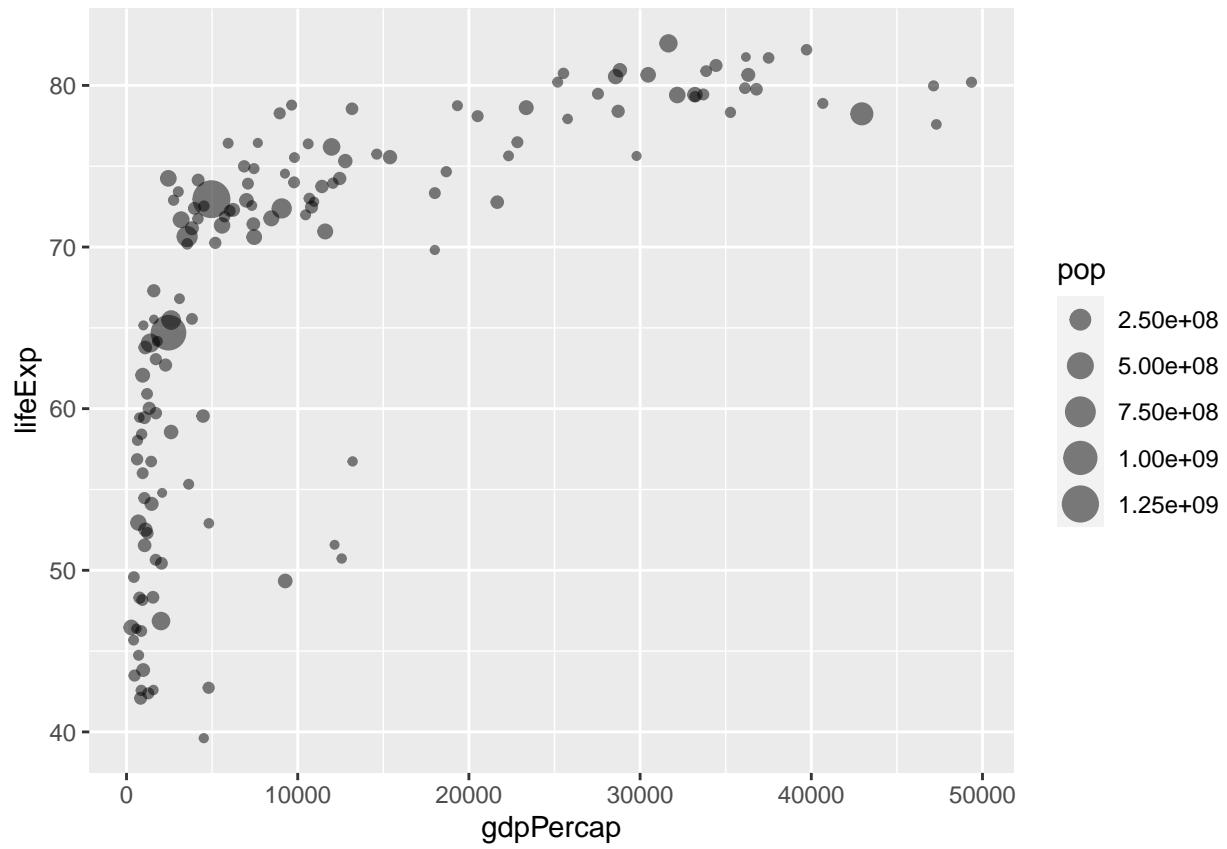
```
ggplot(gapminder_2007) +
  aes(x = gdpPercap, y = lifeExp, color = pop) +
  geom_point(alpha=0.8)
```



## Adjusting point size

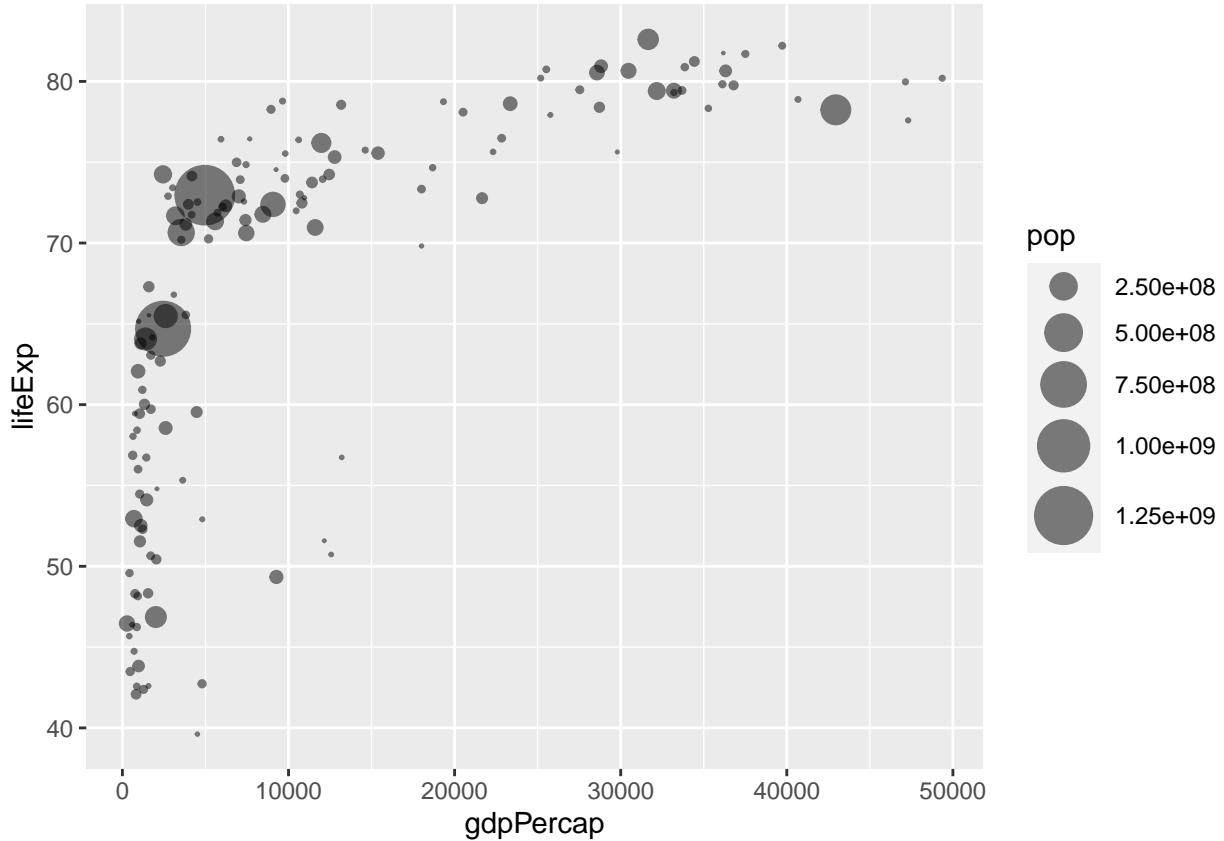
Now change the plot so that point size is based on the population of each country.

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



Improve the plot to reflect the actual population differences by the point size. Use `scale_size_area()` instead.

```
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?

Steps to produce your 1957 plot should include

Use dplyr to filter the gapminder dataset to include only the year 1957 (check above for how we did this for 2007).

Save your result as gapminder\_1957.

Use the ggplot() function and specify the gapminder\_1957 dataset as input

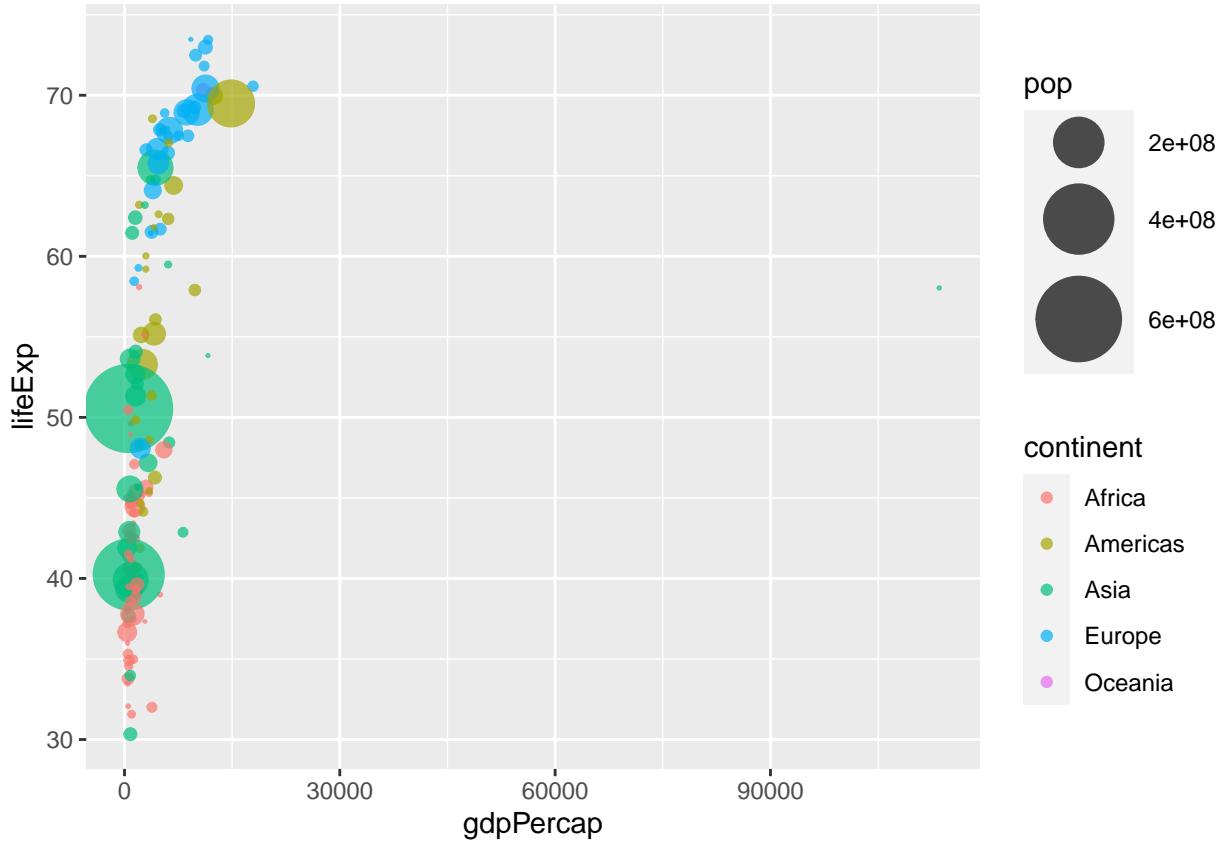
Add a geom\_point() layer to the plot and create a scatter plot showing the GDP per capita gdpPercap on the x-axis and the life expectancy lifeExp on the y-axis

Use the color aesthetic to indicate each continent by a different color

Use the size aesthetic to adjust the point size by the population pop

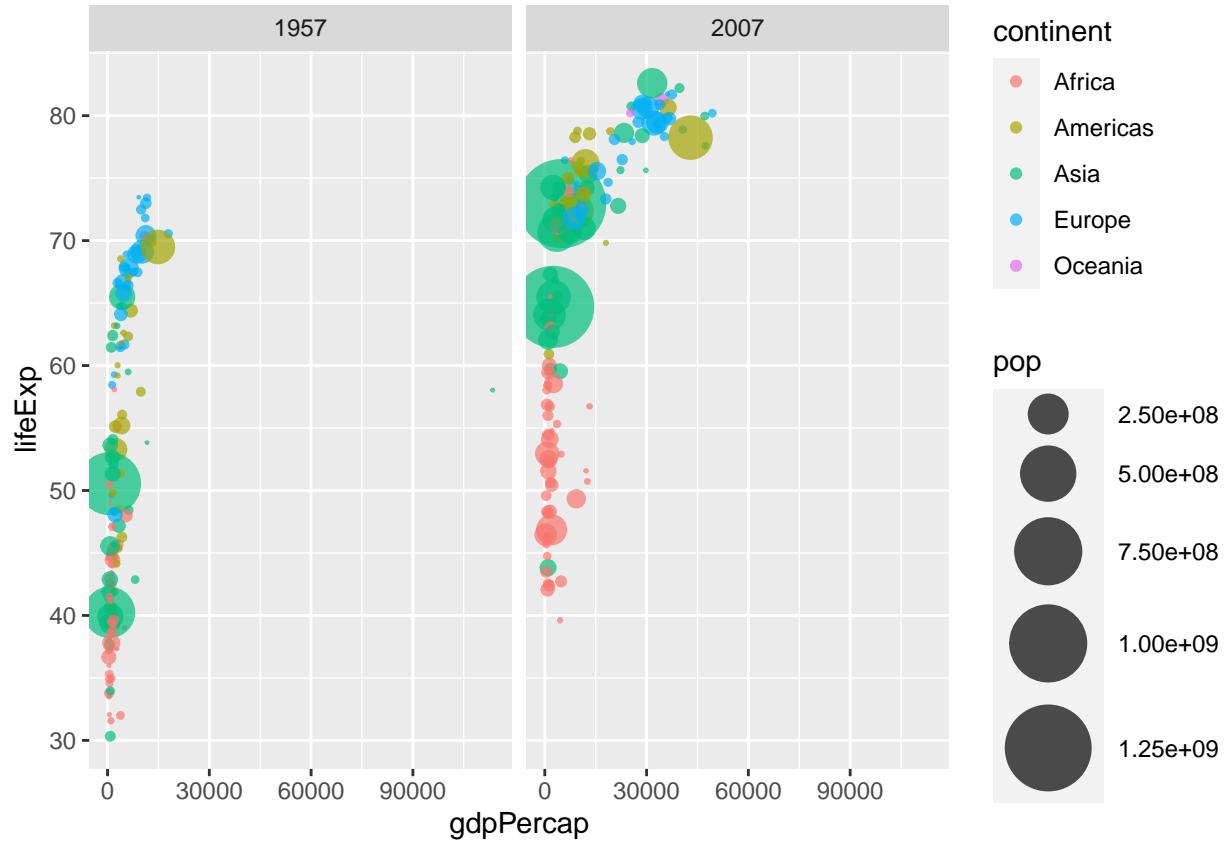
Use scale\_size\_area() so that the point sizes reflect the actual population differences and set the max\_size of each point to 15 -Set the opacity/transparency of each point to 70% using the alpha=0.7 parameter

```
gapminder_1957 <- gapminder %>% filter(year == 1957)
ggplot(gapminder_1957) +
  aes(x = gdpPercap, y = lifeExp, colour = continent, size = pop) +
  scale_size_area(max_size = 15) +
  geom_point(alpha = 0.7)
```



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:

```
gapminder_1957_2007 <- gapminder%>%filter(year==1957 | year==2007)
ggplot(gapminder_1957_2007) +
  aes(x=gdpPercap, y=lifeExp, colour=continent, size=pop) +
  scale_size_area(max_size = 15) +
  geom_point(alpha=0.7) +
  facet_wrap(~year)
```



## Section 7 OPTIONAL: Bar Charts

### Introduction to bar charts

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

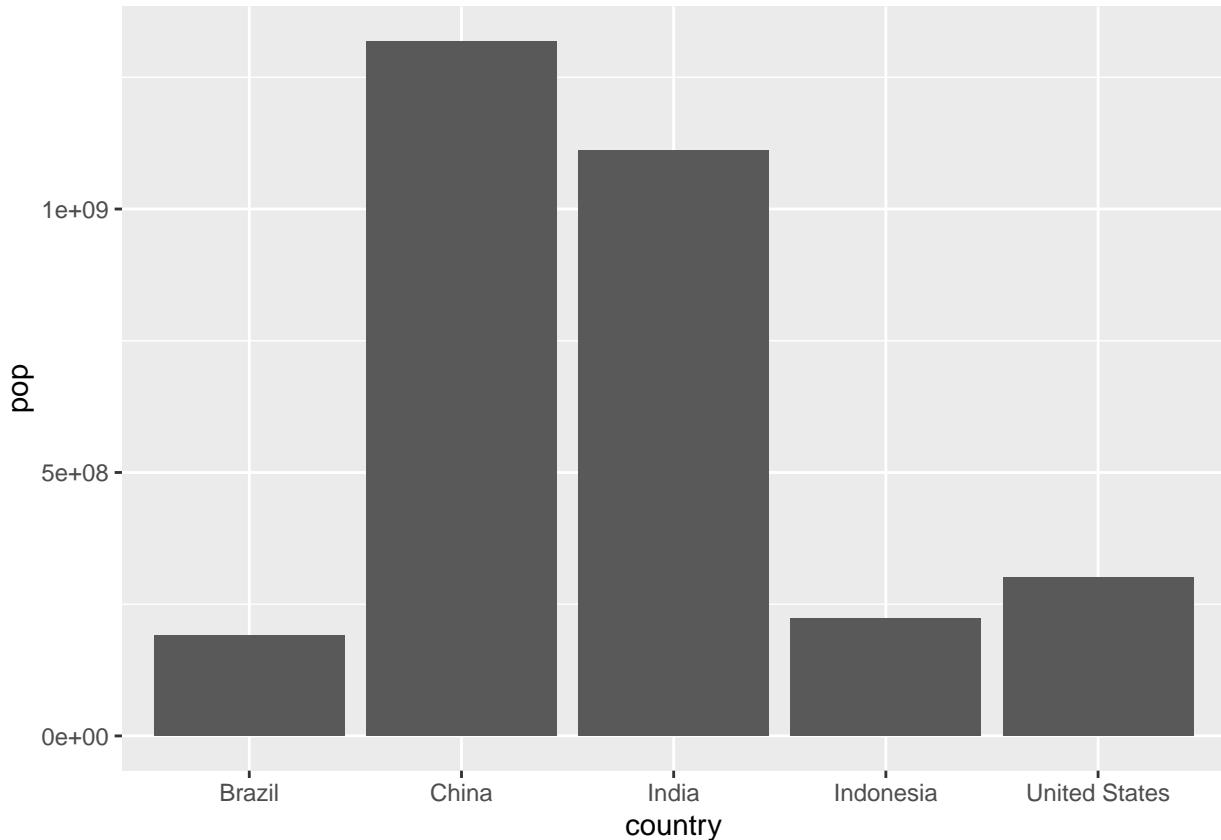
```
gapminder_top5
```

```
## # A tibble: 5 x 6
##   country      continent year lifeExp      pop gdpPerCap
##   <fct>        <fct>    <int>   <dbl>     <int>     <dbl>
## 1 China        Asia      2007    73.0 1318683096     4959.
## 2 India         Asia      2007    64.7 1110396331     2452.
## 3 United States Americas  2007    78.2 301139947     42952.
## 4 Indonesia    Asia      2007    70.6 223547000     3541.
## 5 Brazil        Americas 2007    72.4 190010647     9066.
```

## Creating a simple bar chart

Now create a bar chart with gapminder\_top5 dataset (countries are in alphabetical order by default).

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



### Q. Plot life expectancy by country

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

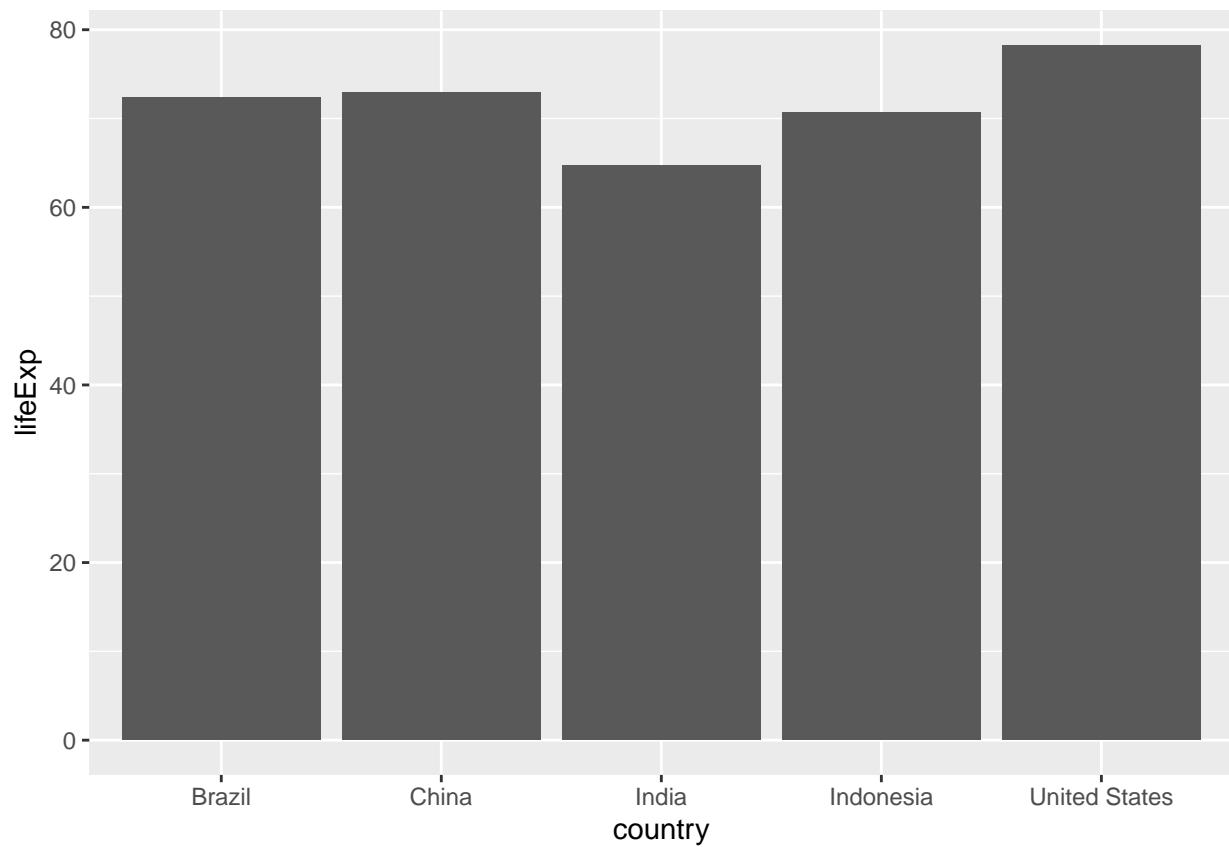
Use the `ggplot()` function and specify the `gapminder_top5` dataset as input

Add a `geom_col()` layer to the plot

Plot one bar for each country (`x` aesthetic)

Use life expectancy `lifeExp` as bar height (`y` aesthetic)

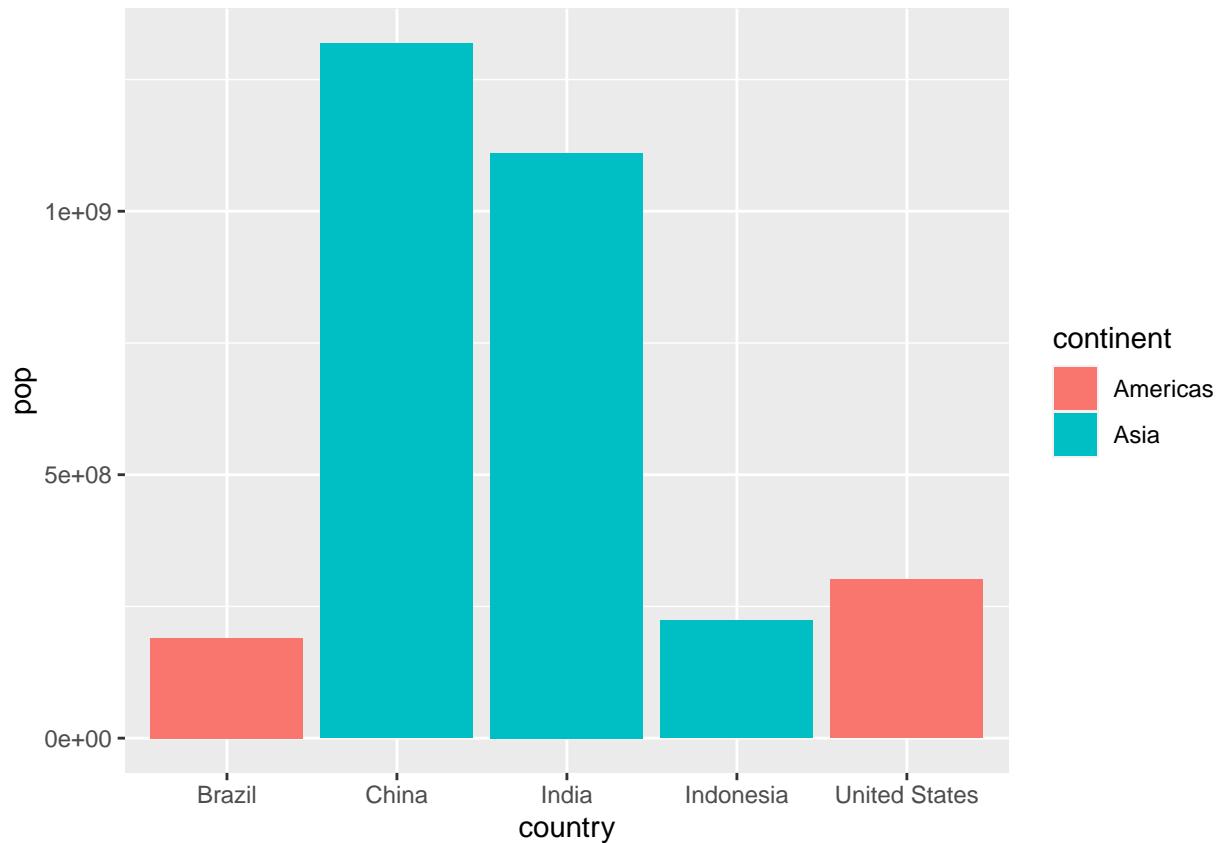
```
ggplot(gapminder_top5) +  
  geom_col(aes(x = country, y = lifeExp))
```



### Filling bars with color

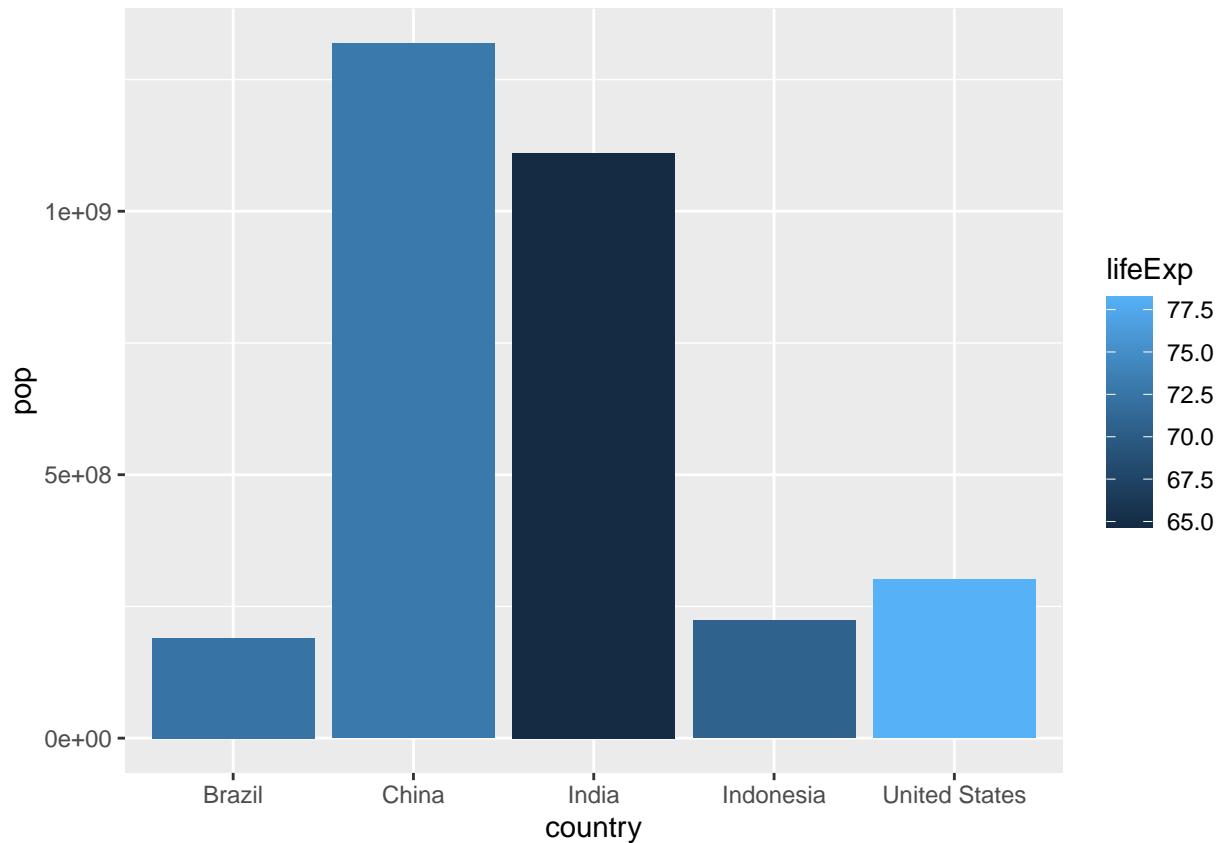
Plot the population of the biggest countires and use the continent variable to color each bar.

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



Now use the life expectancy variable to fill the colors for the bars.

```
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.

Use the `ggplot()` function and specify the `gapminder_top5` dataset as input

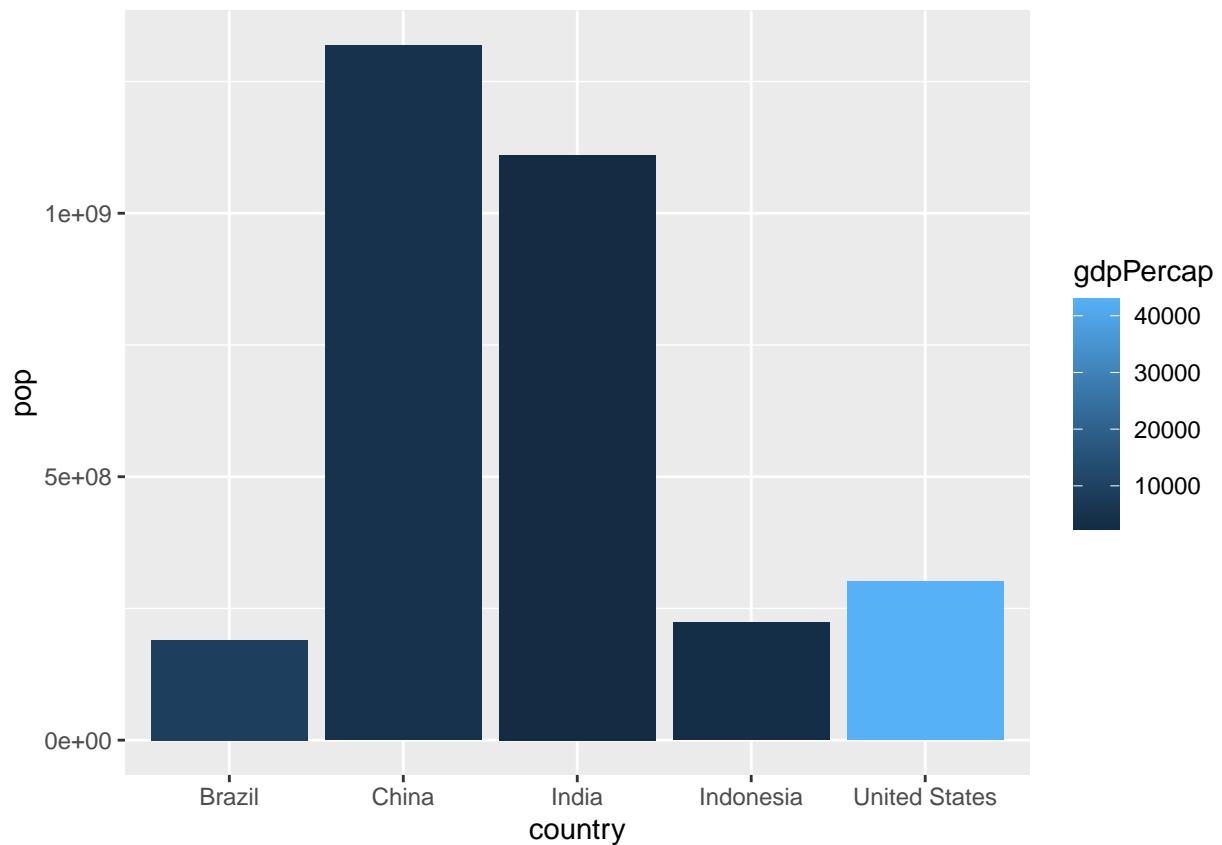
Add a `geom_col()` layer to the plot

Plot one bar for each country (`x` aesthetic)

Use population `pop` as bar height (`y` aesthetic)

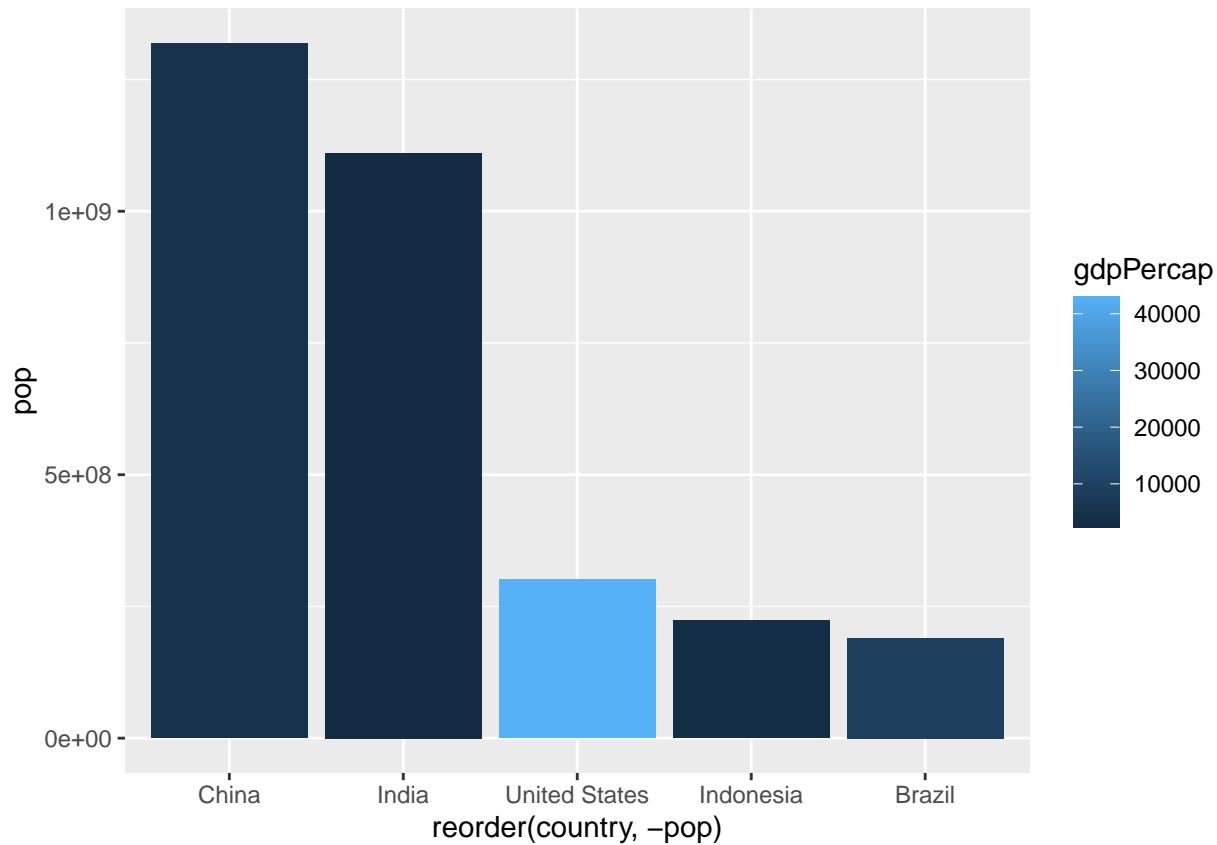
Use the GDP per capita `gdpPercap` as fill aesthetic

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



Now change the orders of the bars

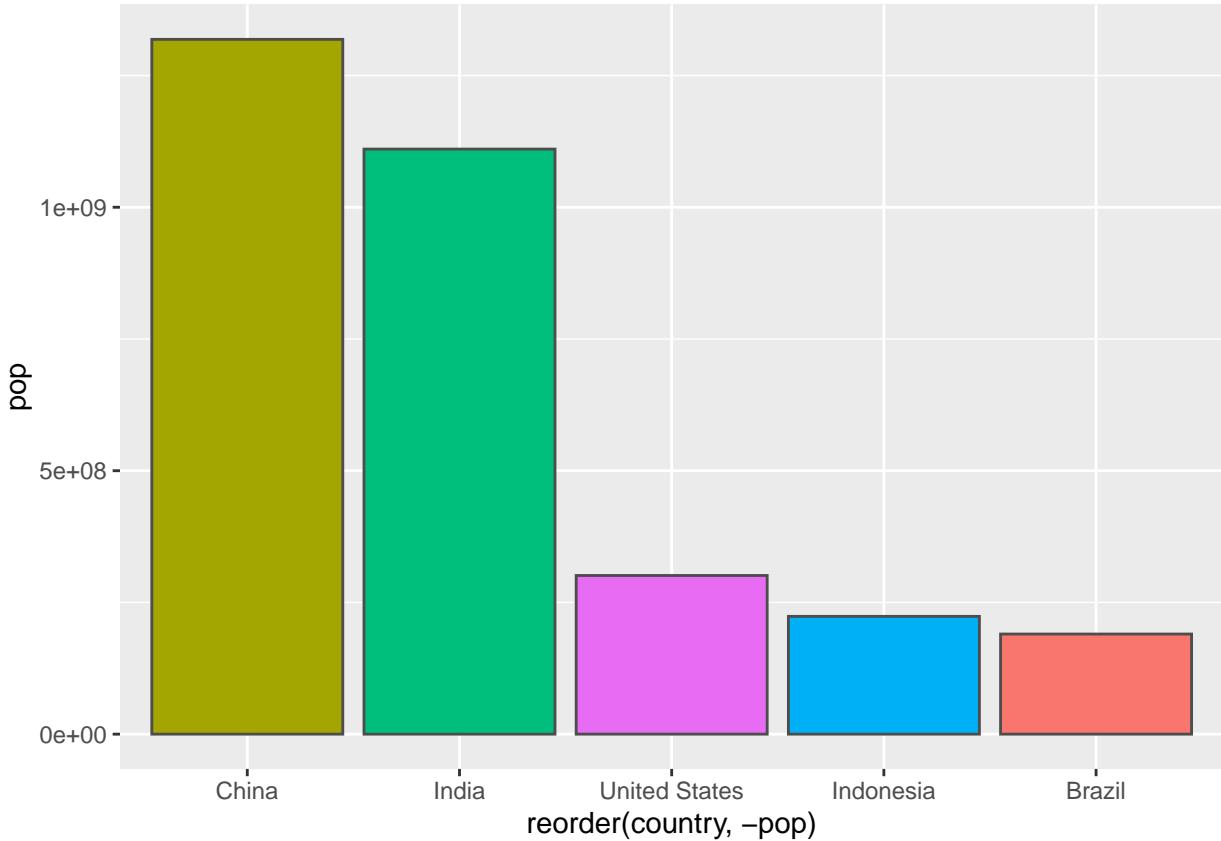
```
ggplot(gapminder_top5) +  
  geom_col(aes(x=reorder(country, -pop), y=pop, fill=gdpPercap))
```



Now just fill by country

```
ggplot(gapminder_top5) +
  aes(x=reorder(country, -pop), y=pop, fill=country) +
  geom_col(col="gray30") +
  guides(fill=FALSE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

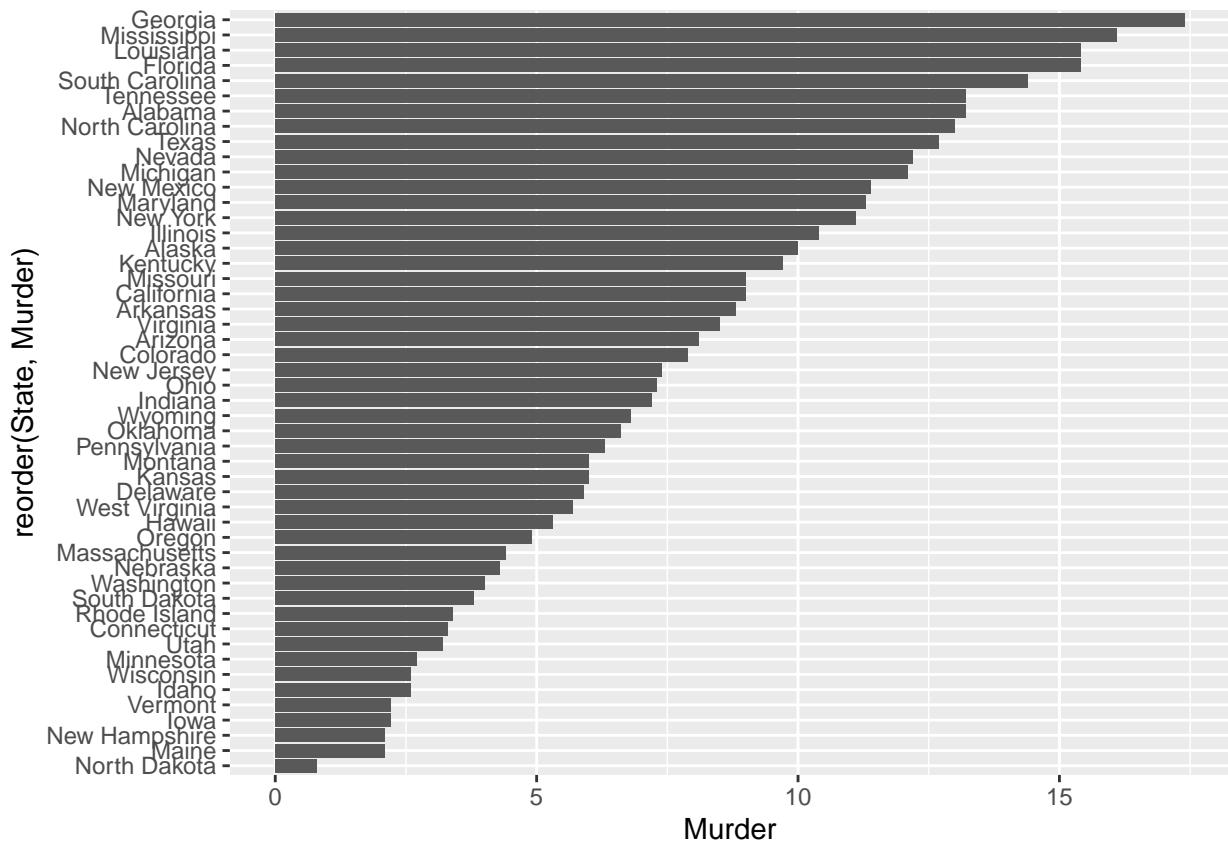


## Flipping bar charts

```
head(USArrests)

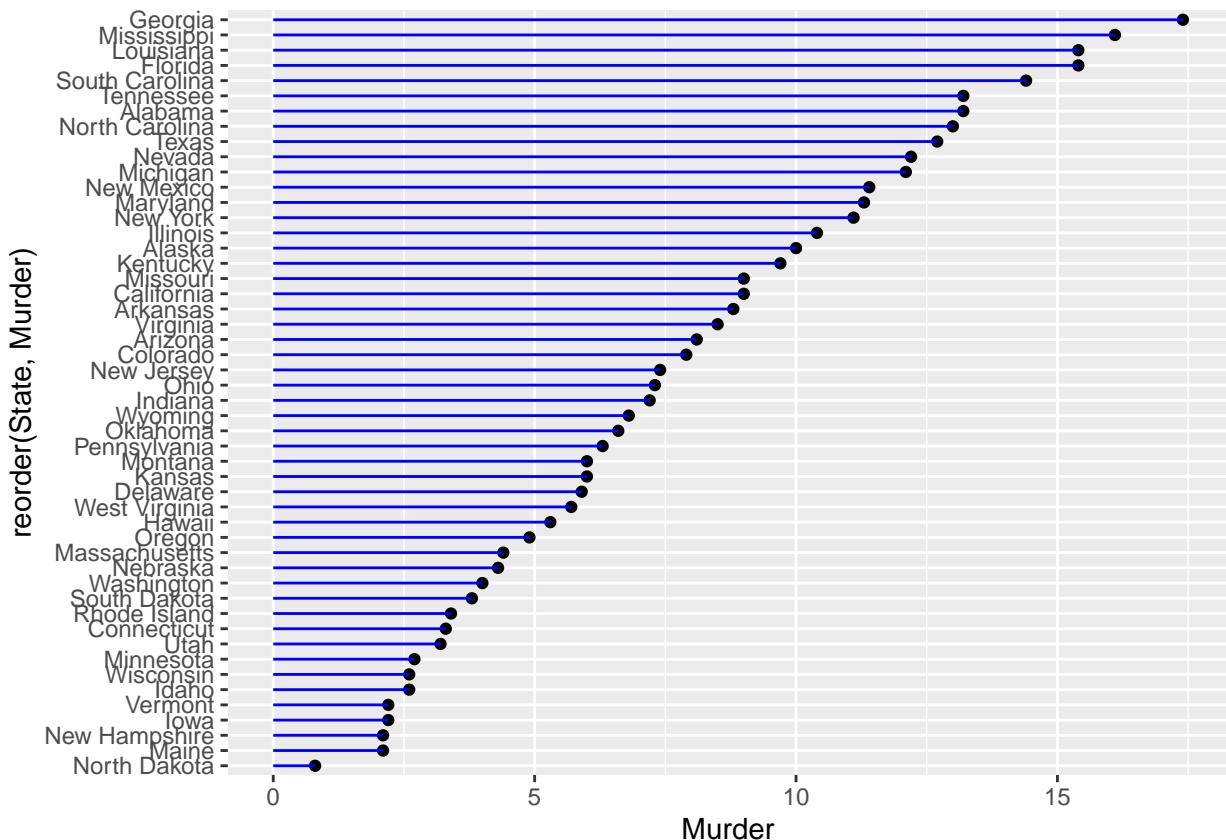
##          Murder Assault UrbanPop Rape
## Alabama     13.2    236      58 21.2
## Alaska      10.0    263      48 44.5
## Arizona      8.1    294      80 31.0
## Arkansas     8.8    190      50 19.5
## California   9.0    276      91 40.6
## Colorado     7.9    204      78 38.7
```

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



A bit crowded, use an alternative custom visualization. Combine `geom_point()` and `geom_segment()` to do this

```
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()
```



## SECTION 8 ADVANCED: Plot Animation

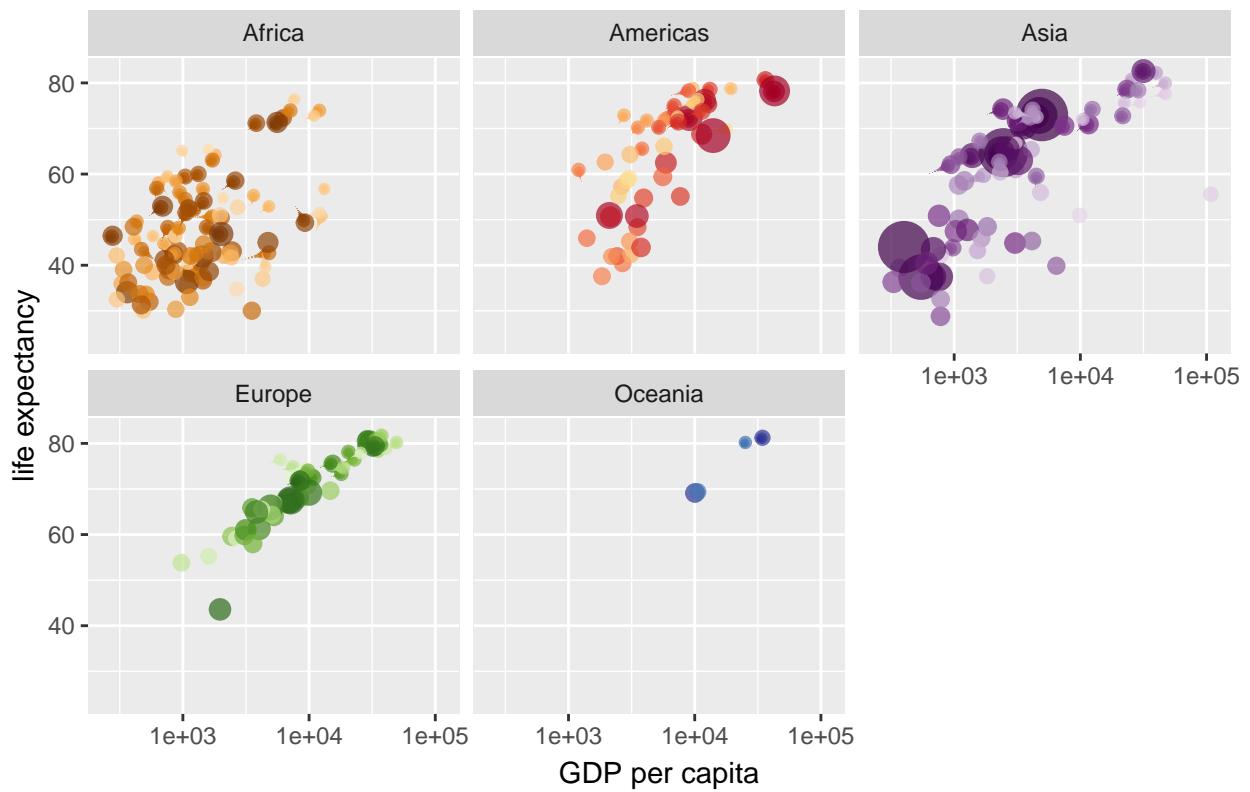
```
# install.packages("gifsSKI") ## un-comment to install if necessary
# install.packages("gganimate") ## un-comment to install if necessary
```

For this exercise, we only show the images first and last year generated for this pdf (1952 and 2007 respectively).

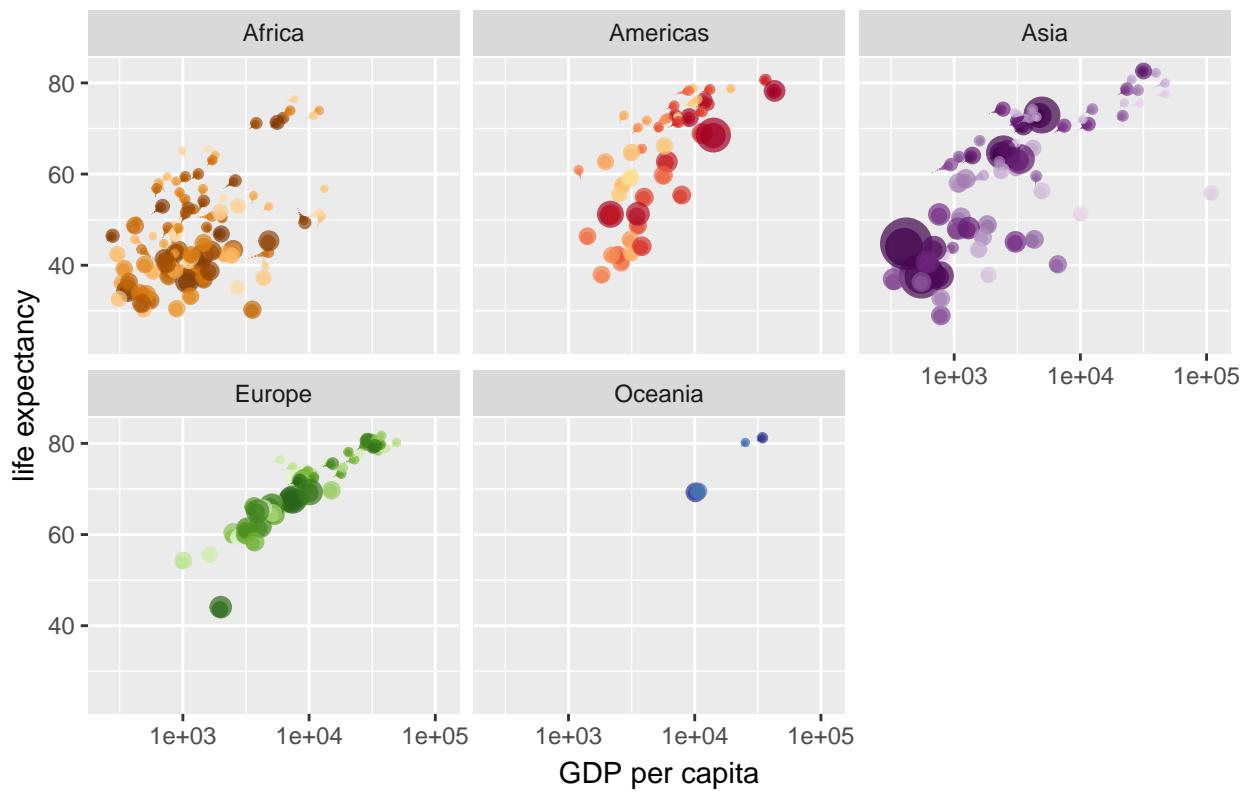
```
library(gganimate)

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() +
  # Continent facet wrap
  facet_wrap(~continent) +
  # The following code is for gganimate
  labs(title = 'Year: {frame_time}', x = 'GDP per capita', y = 'life expectancy') +
  transition_time(year) +
  shadow_wake(wake_length = 0.1, alpha = FALSE)
```

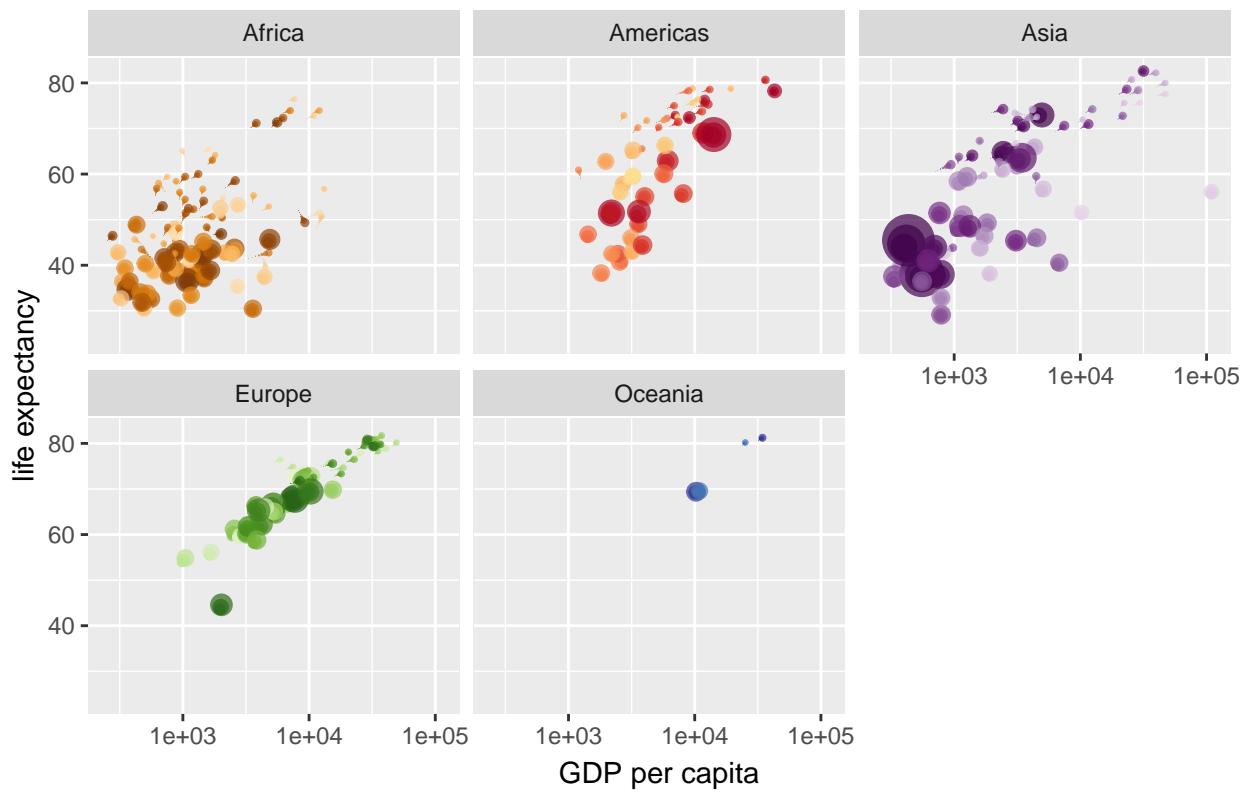
Year: 1952



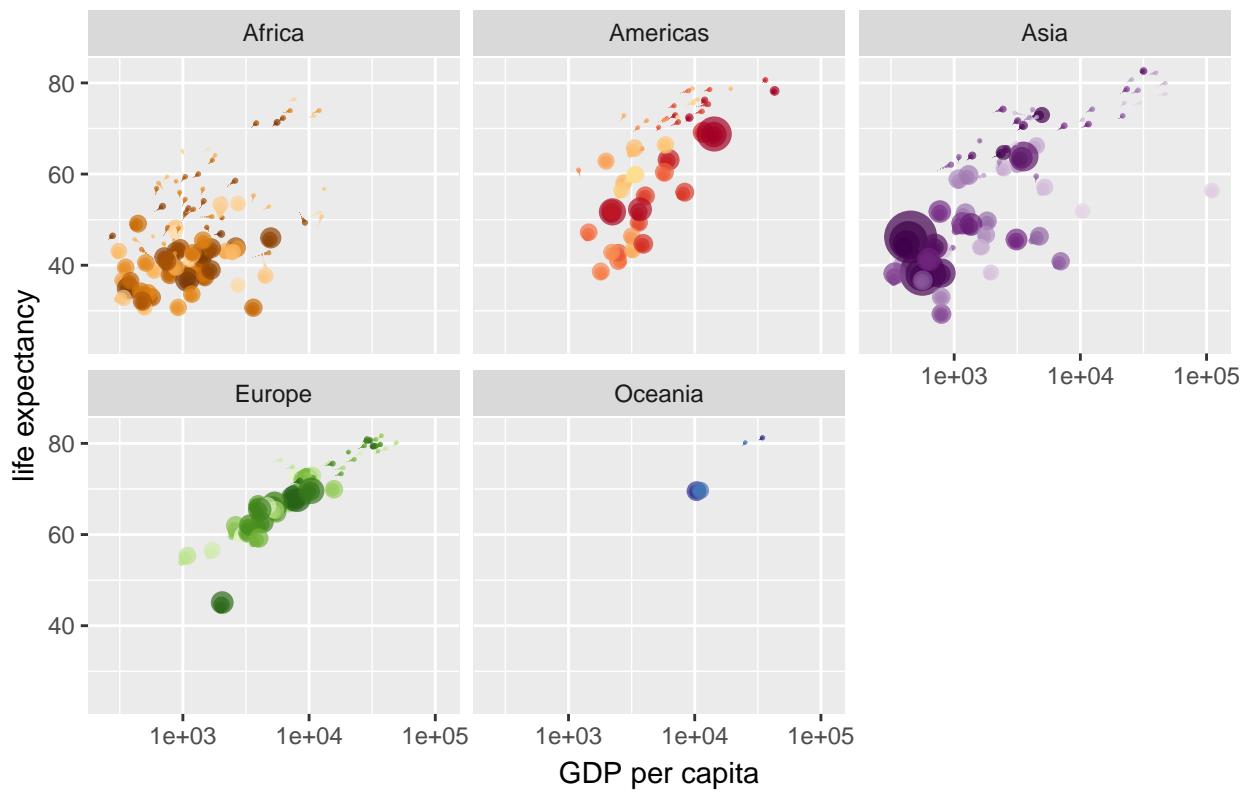
Year: 1953



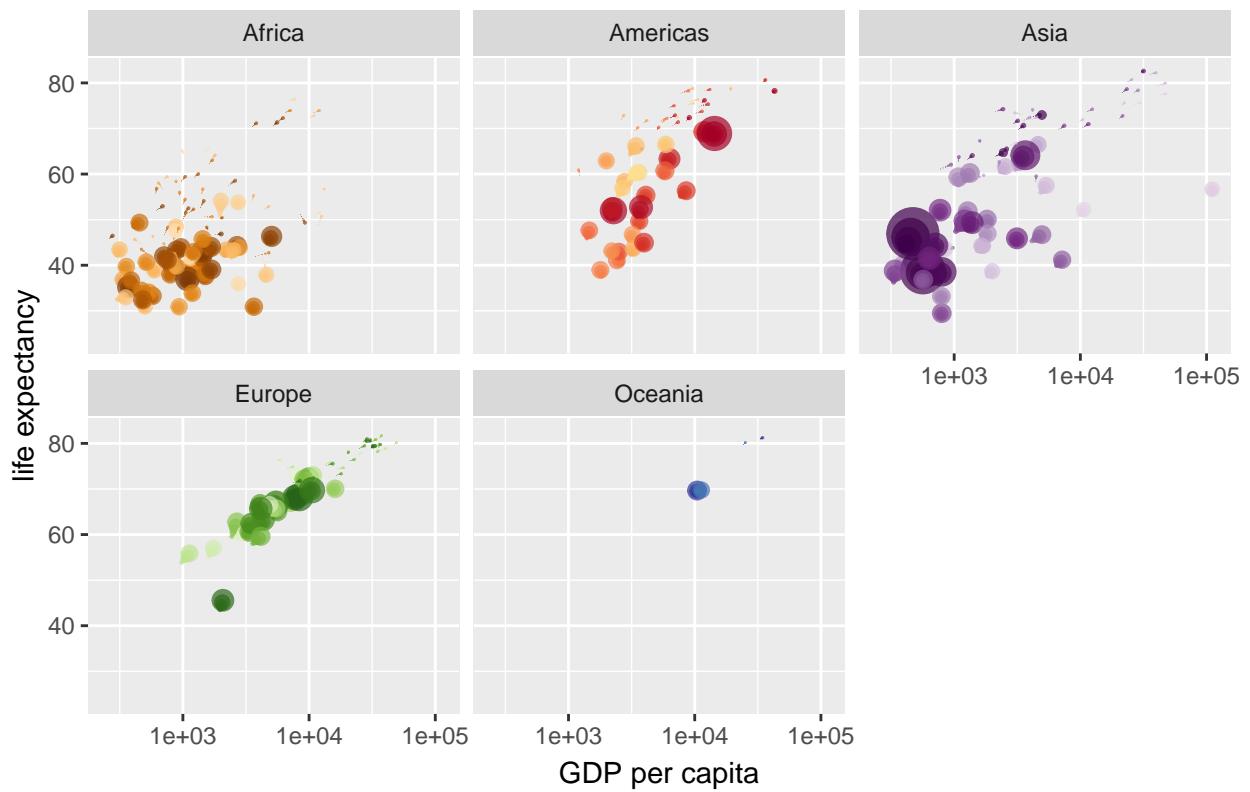
Year: 1953



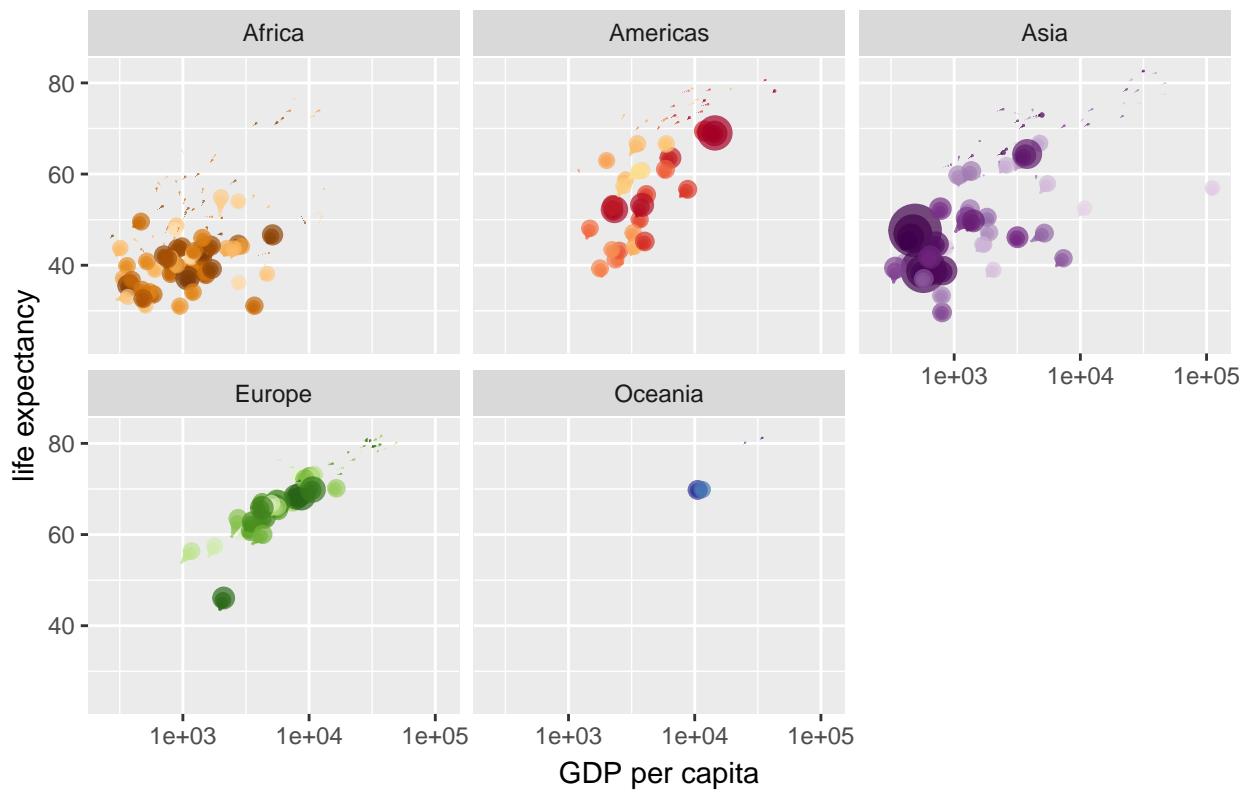
Year: 1954



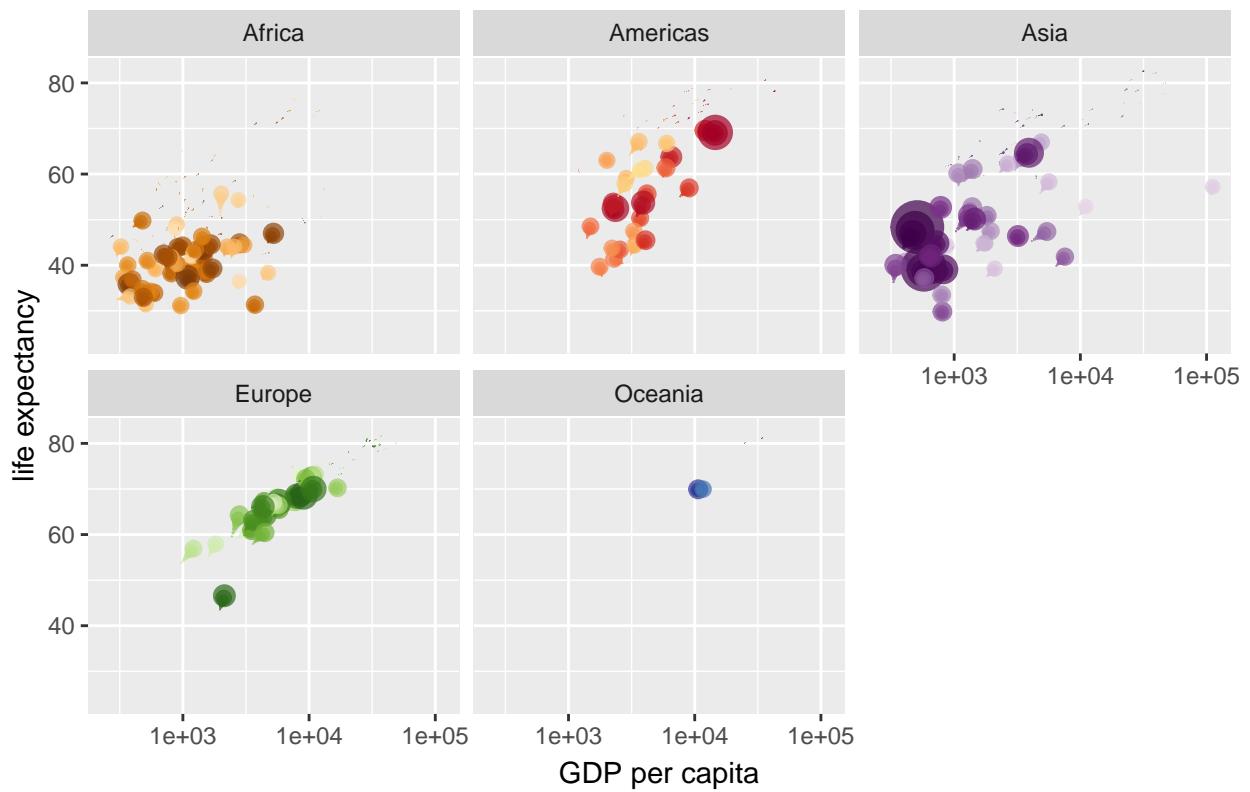
Year: 1954



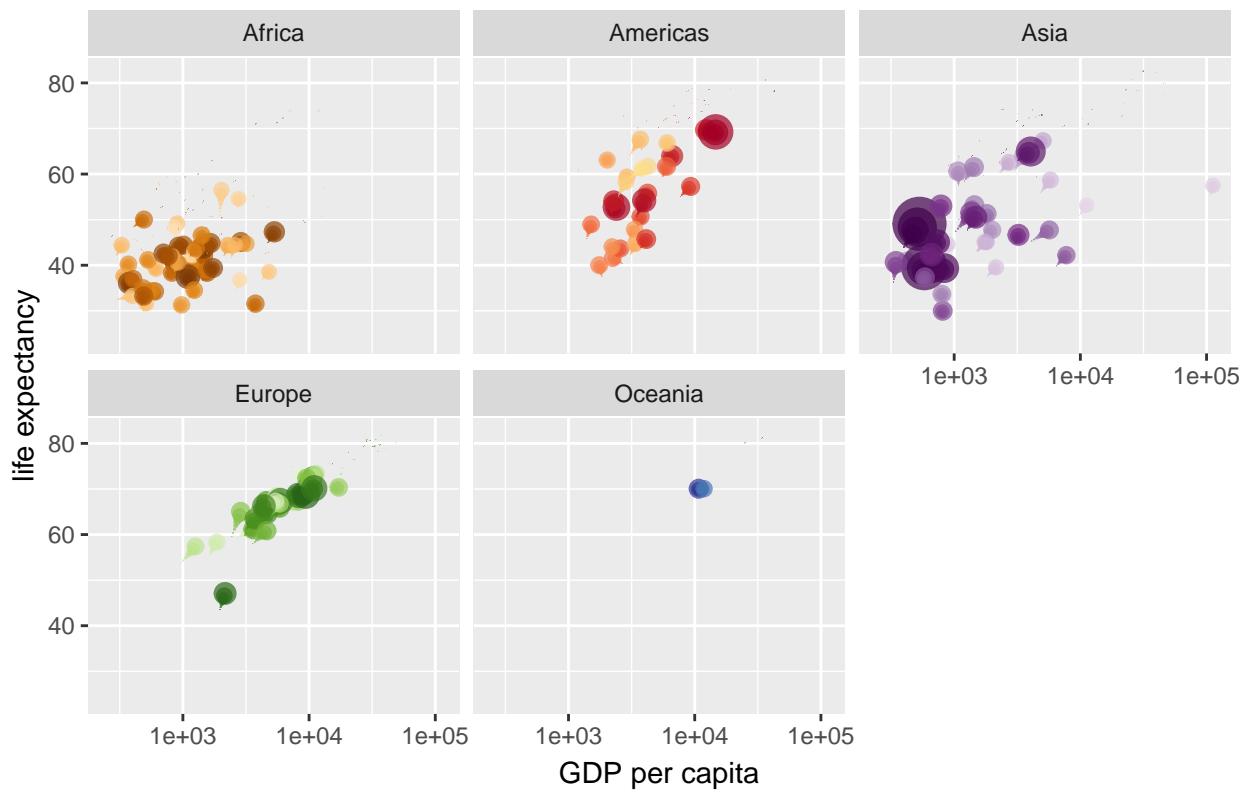
Year: 1955



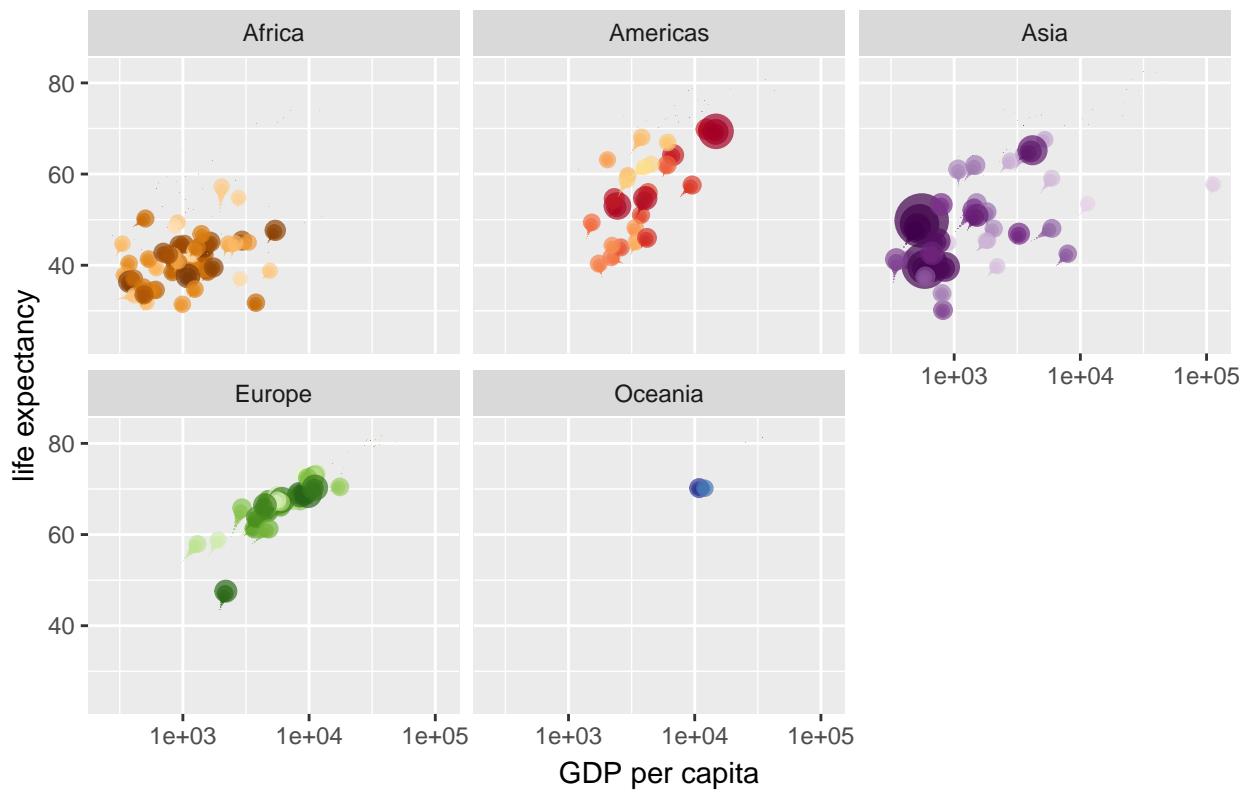
Year: 1955



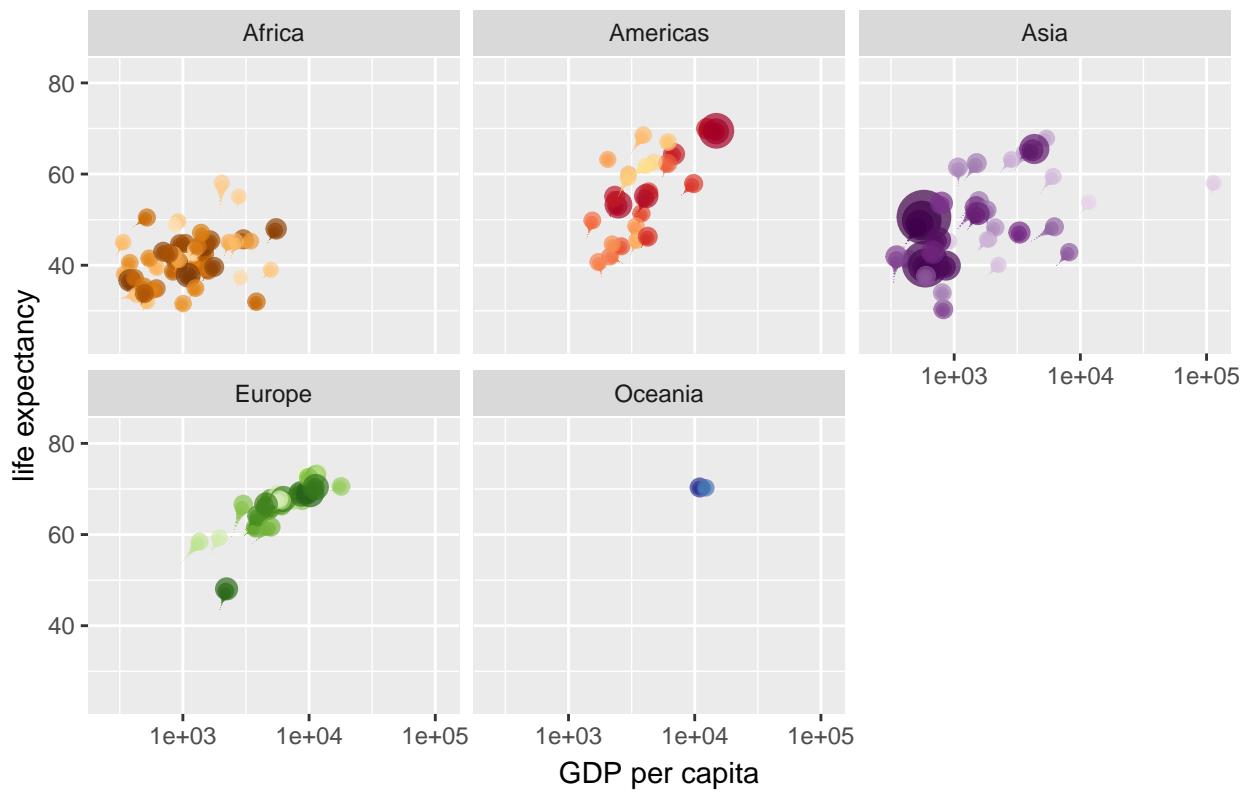
Year: 1956



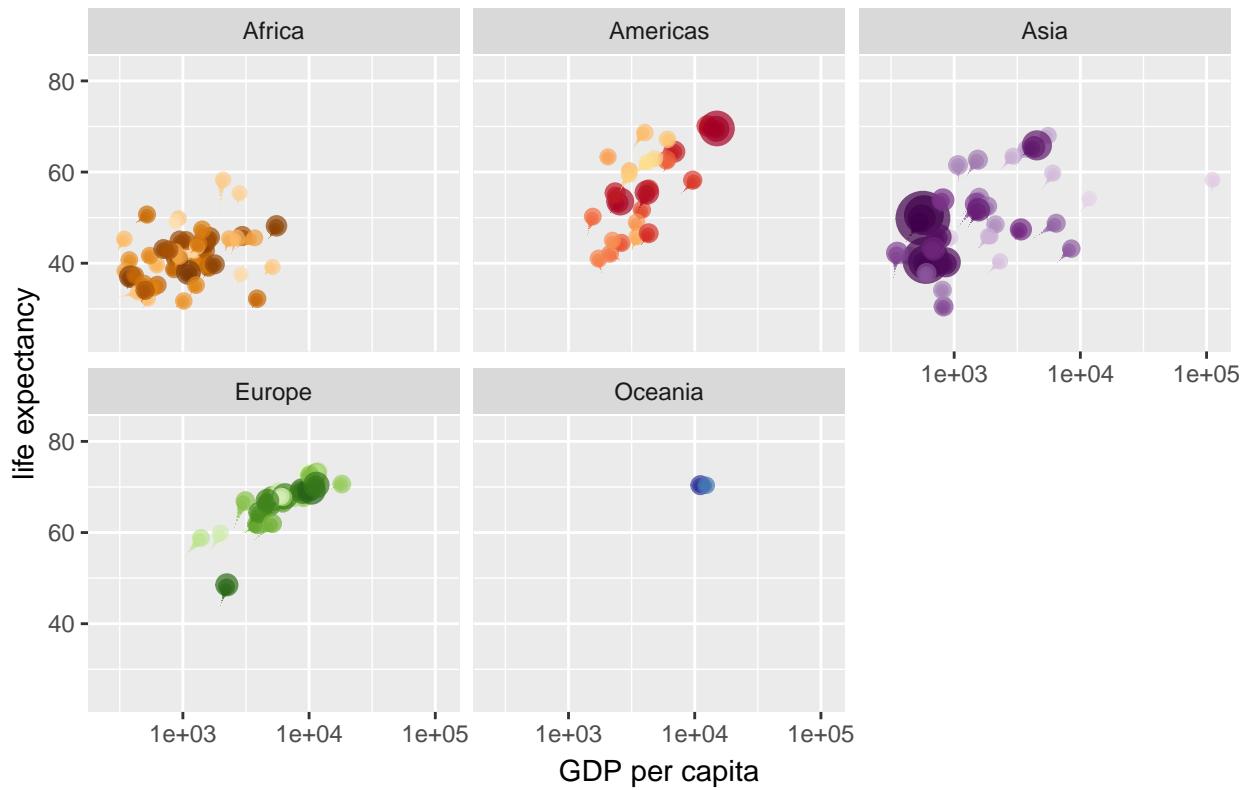
Year: 1956



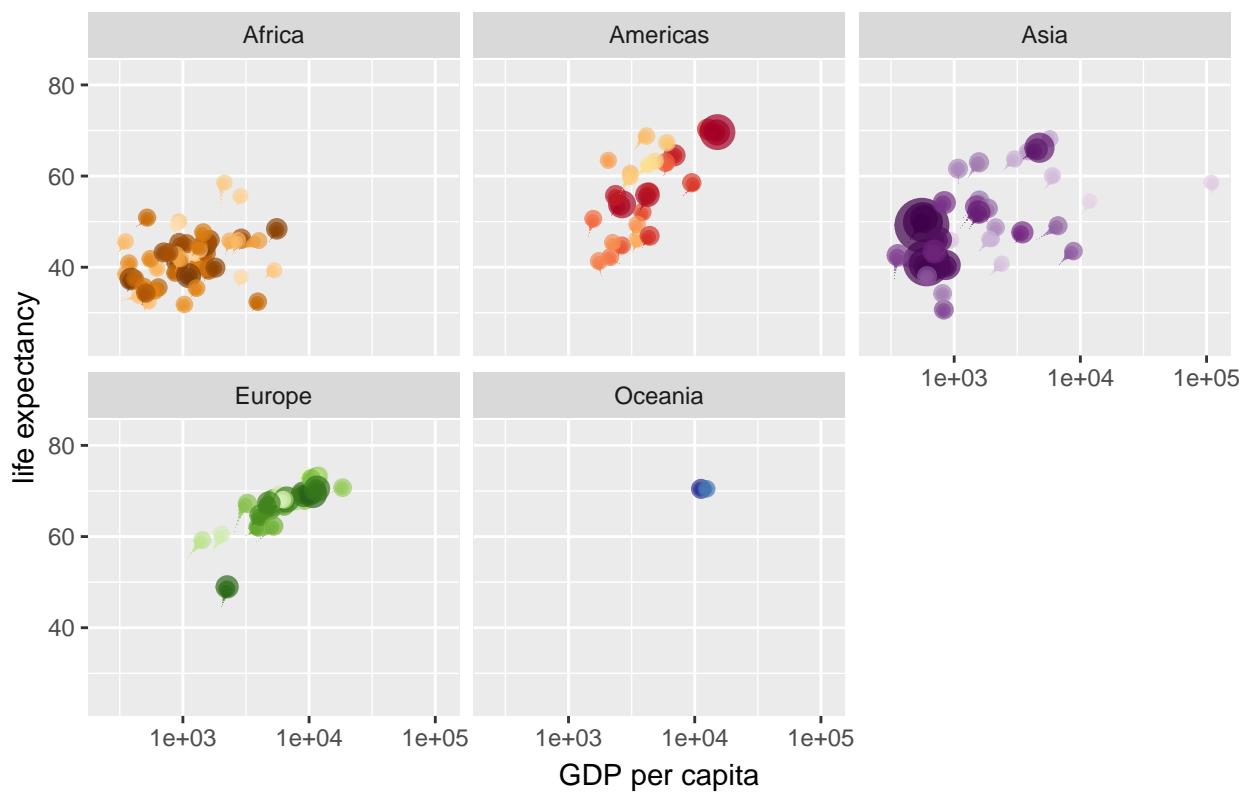
Year: 1957



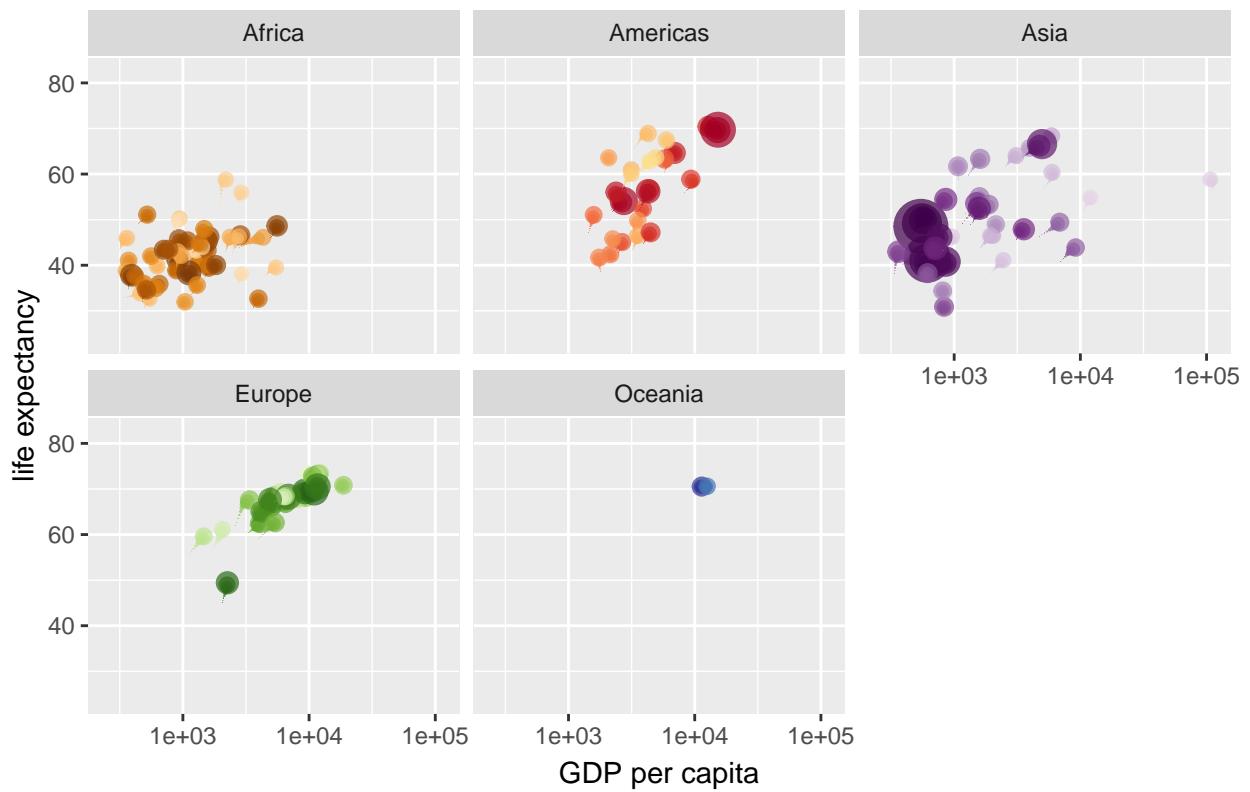
Year: 1958



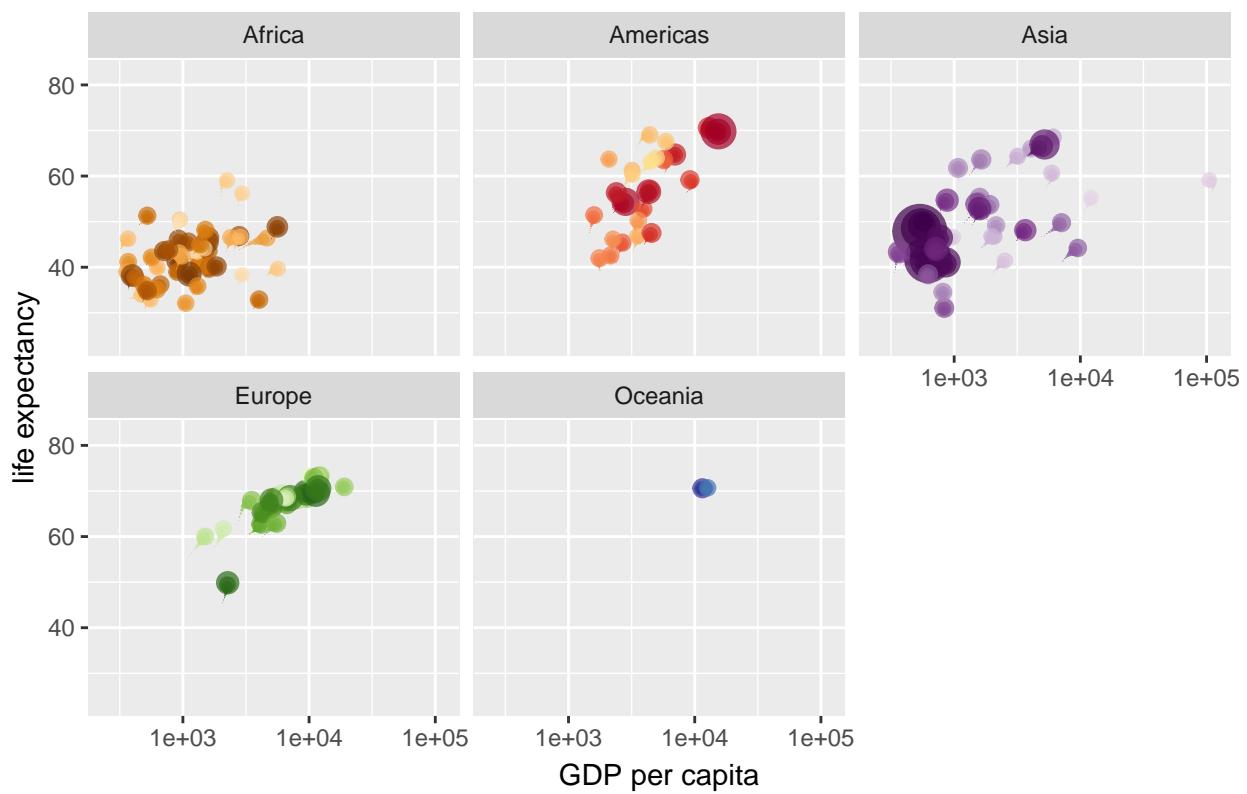
Year: 1958



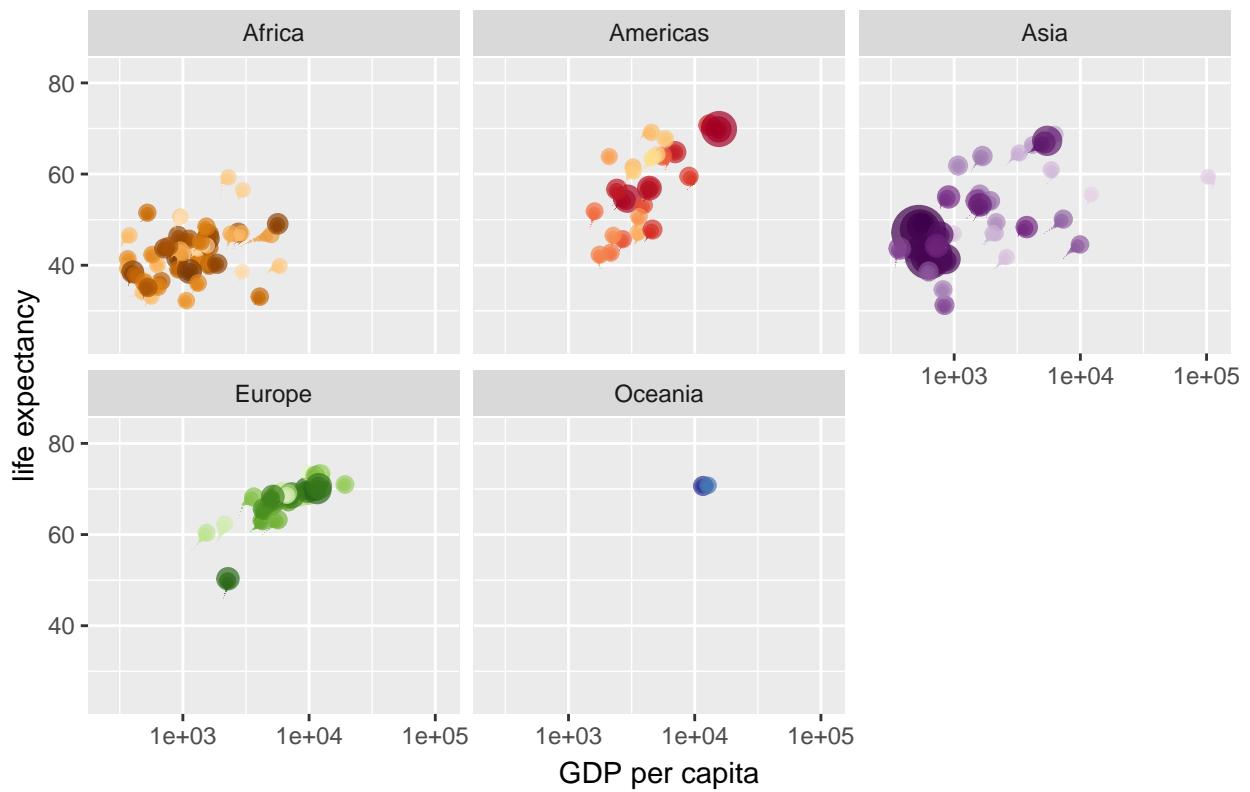
Year: 1959



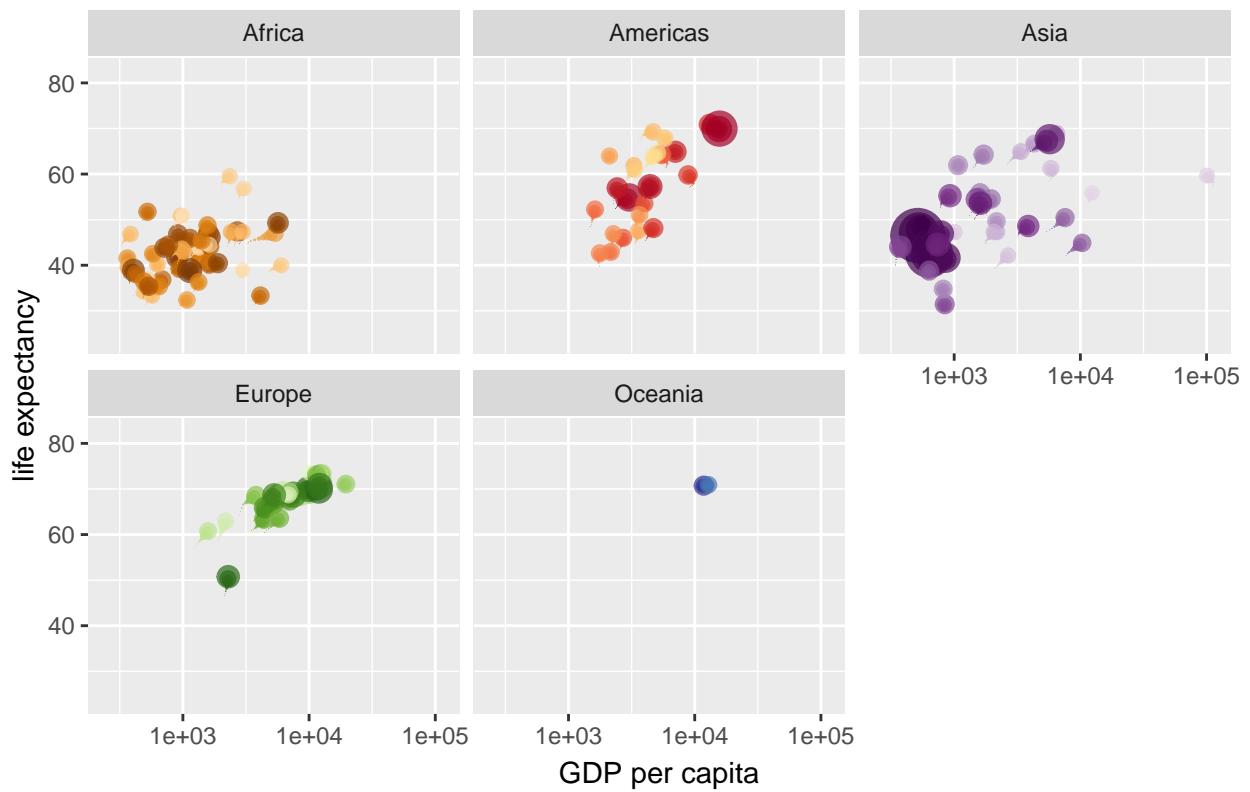
Year: 1959



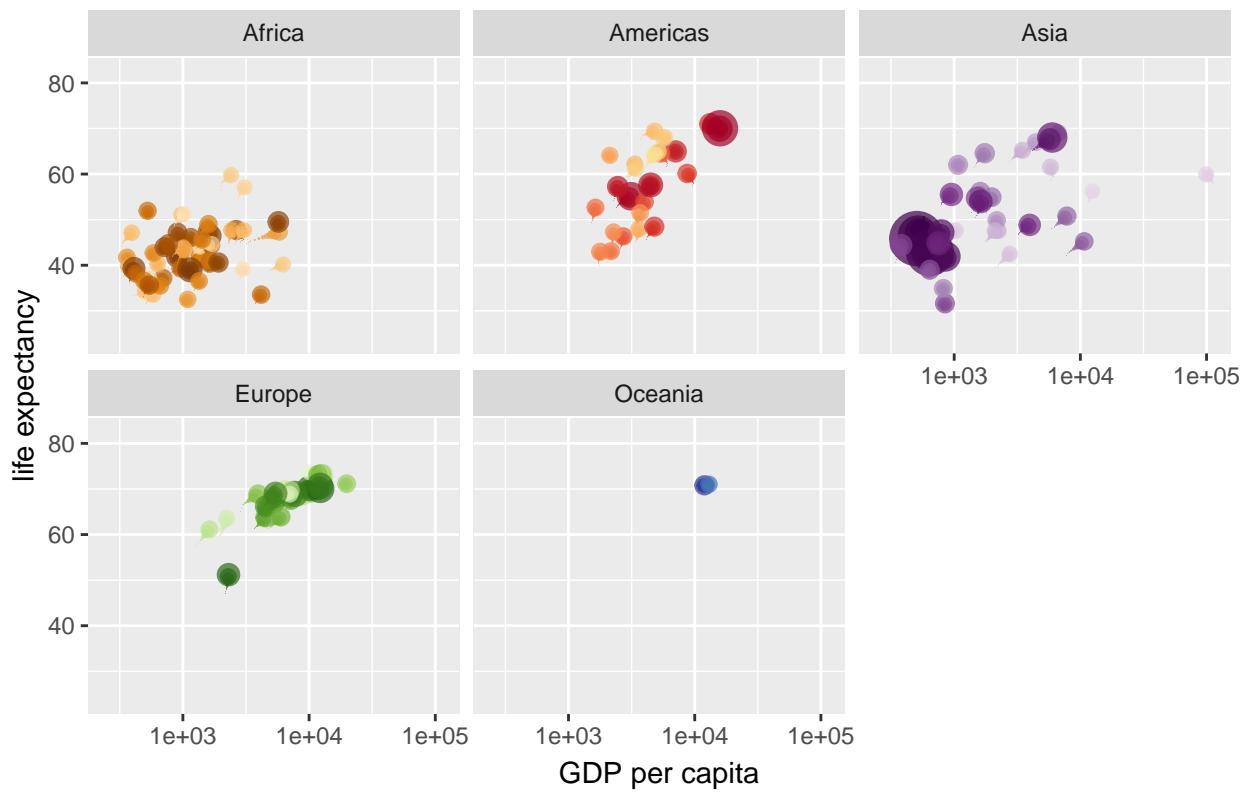
Year: 1960



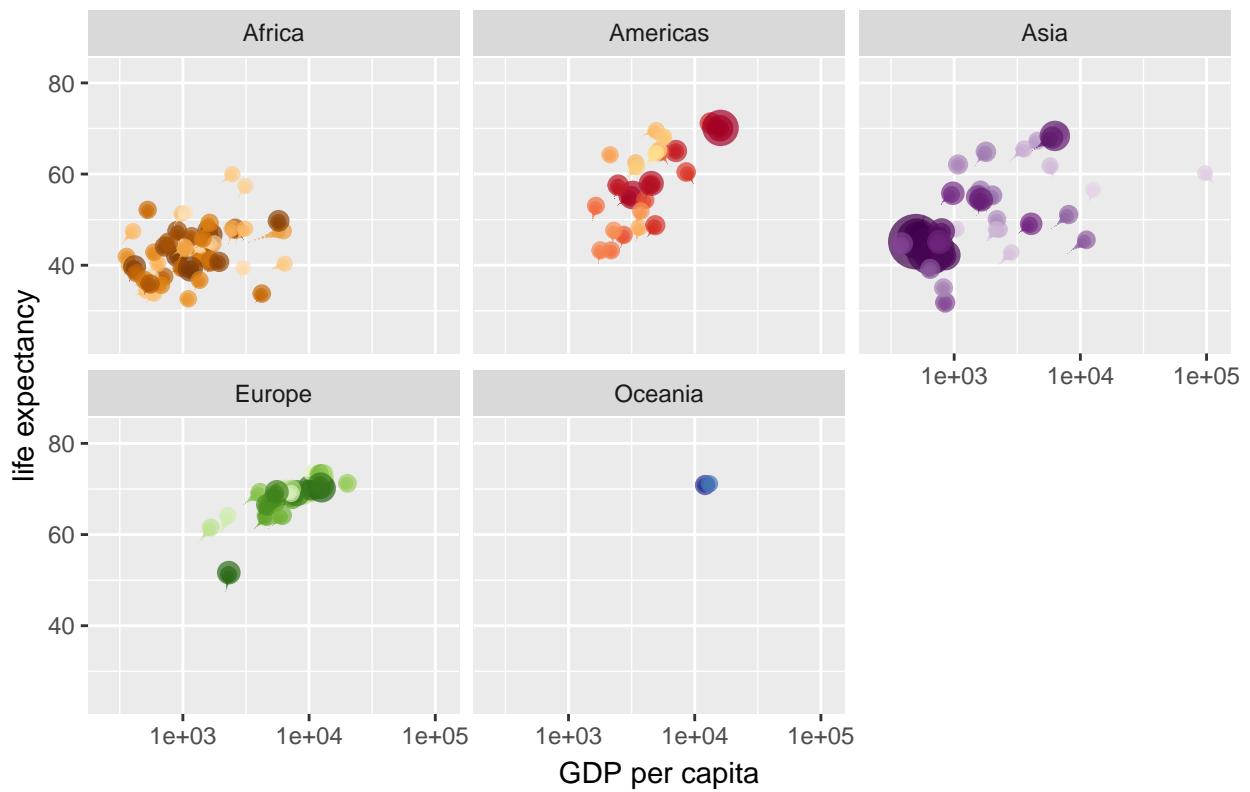
Year: 1960



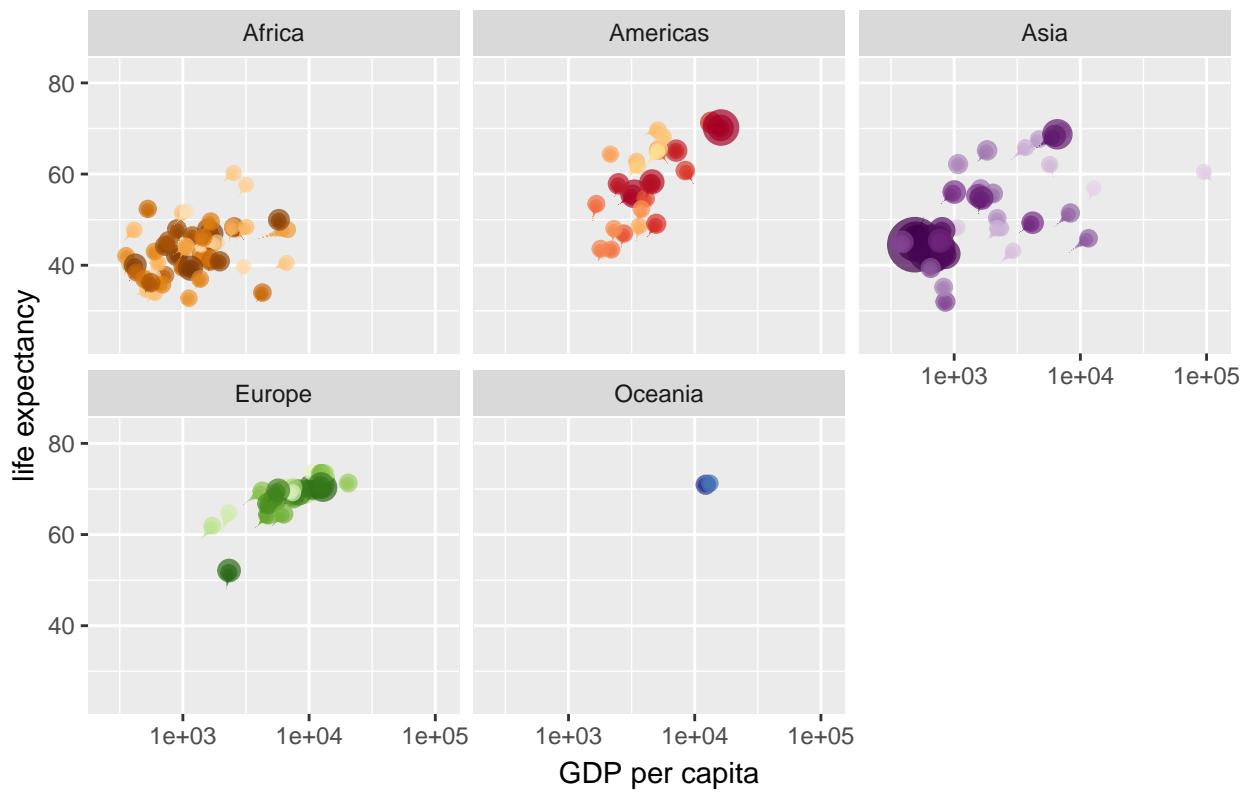
Year: 1961



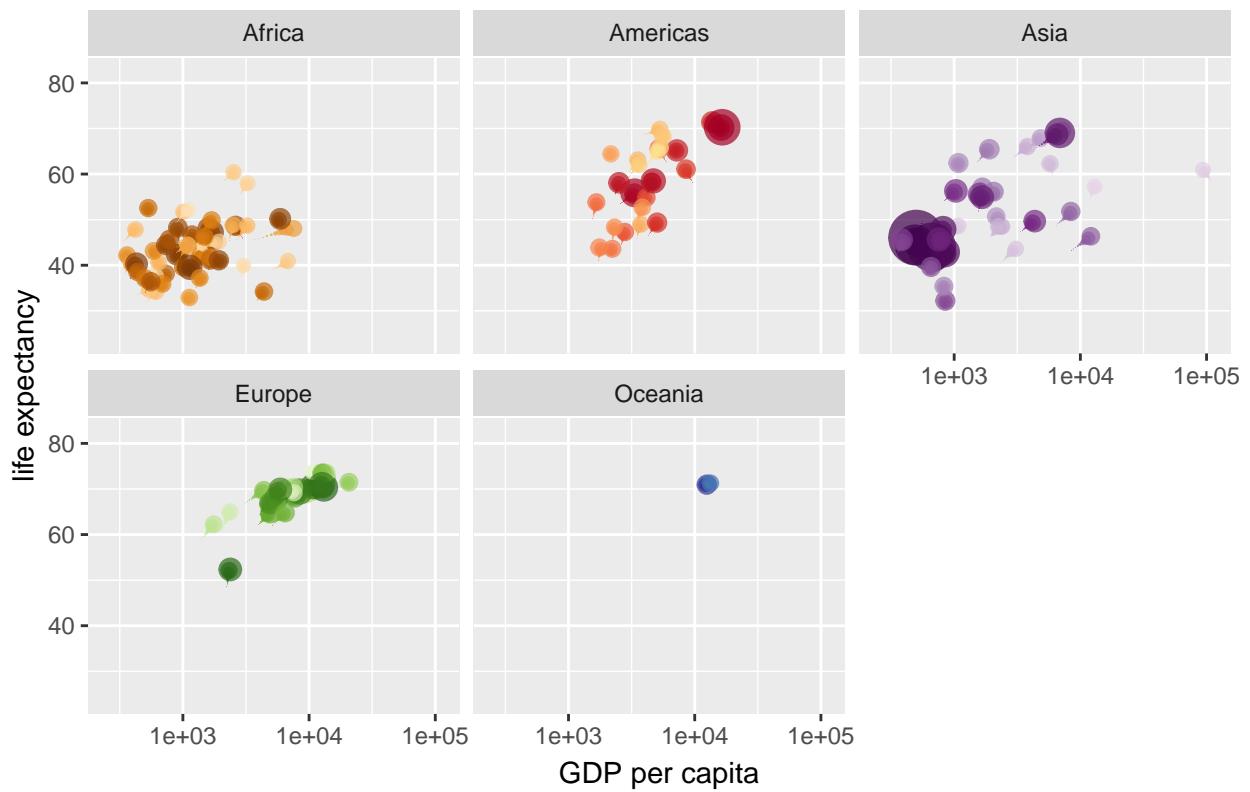
Year: 1961



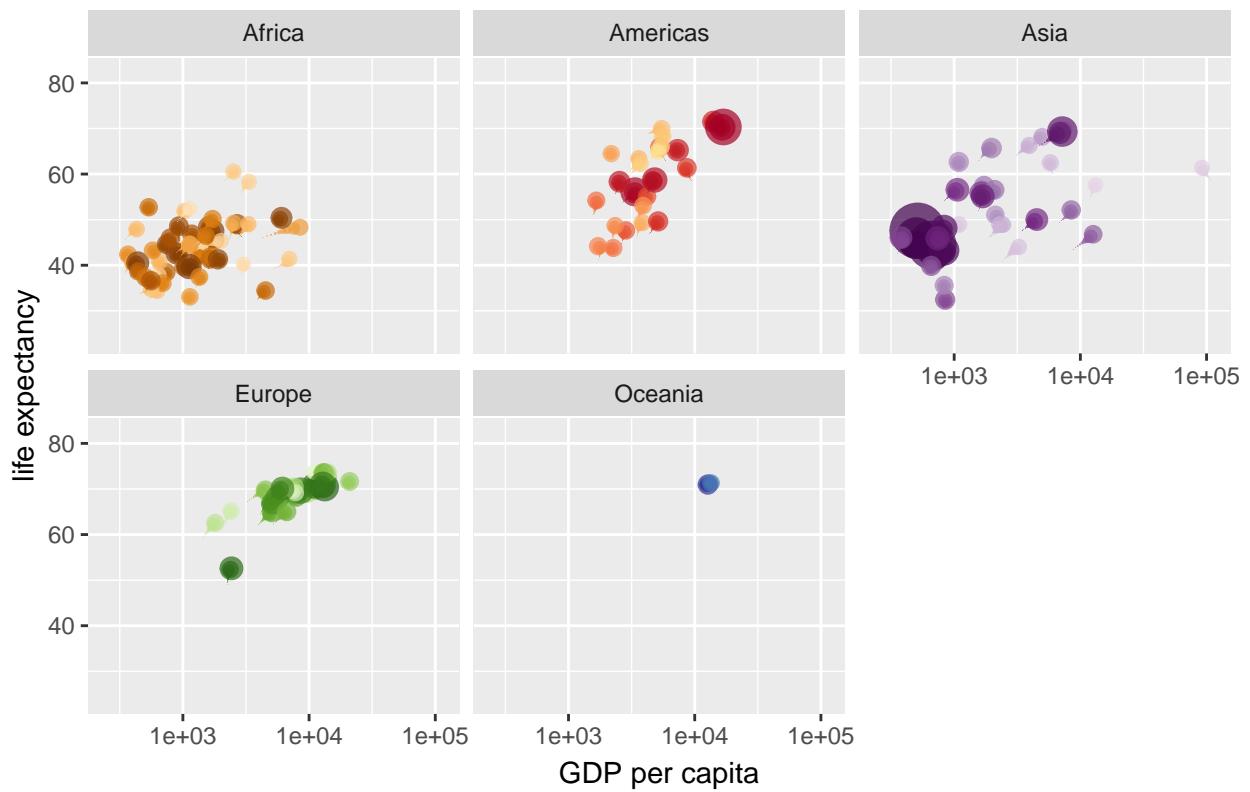
Year: 1962



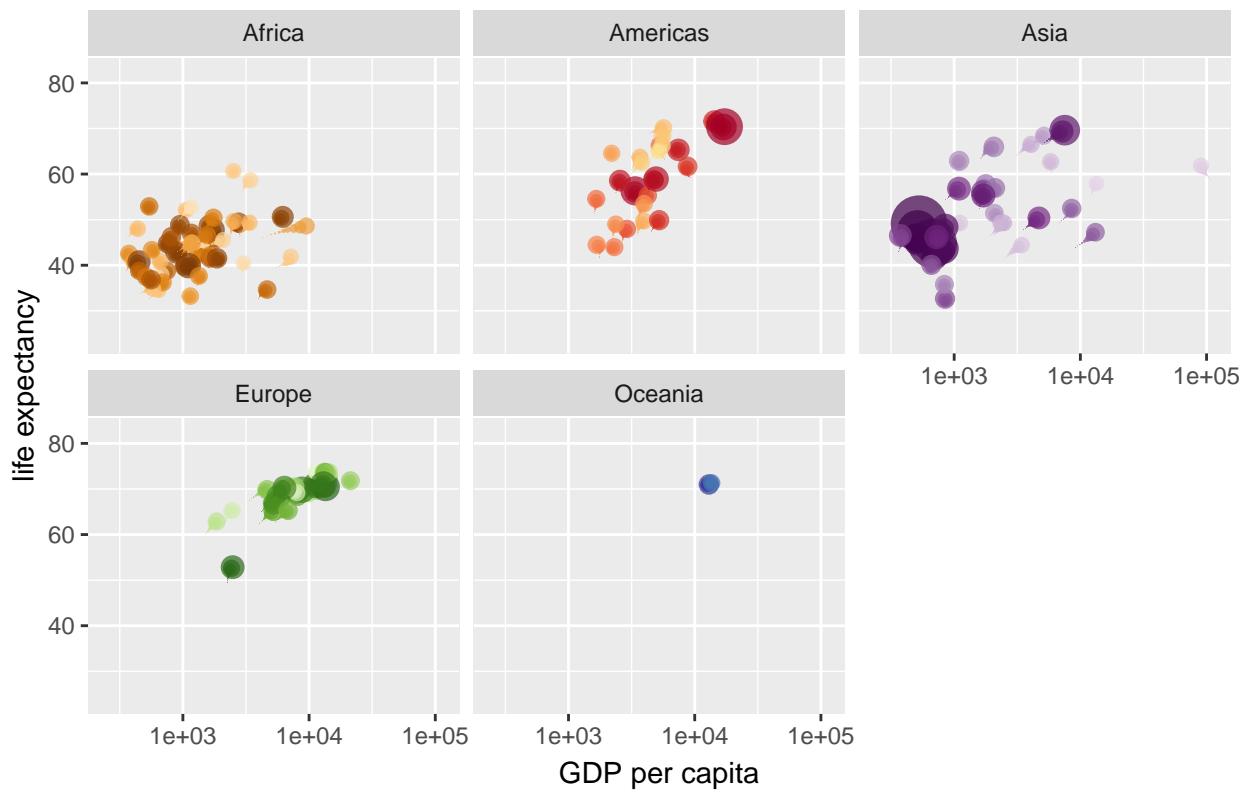
Year: 1963



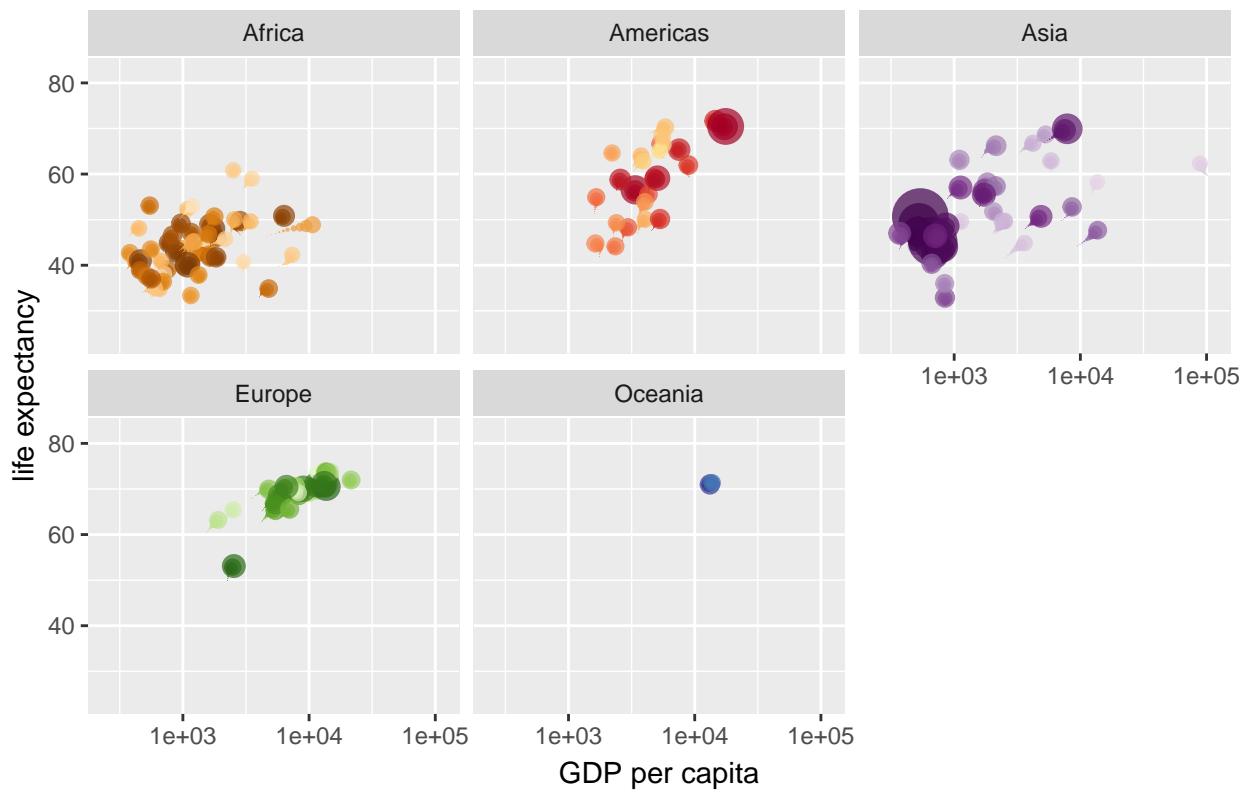
Year: 1963



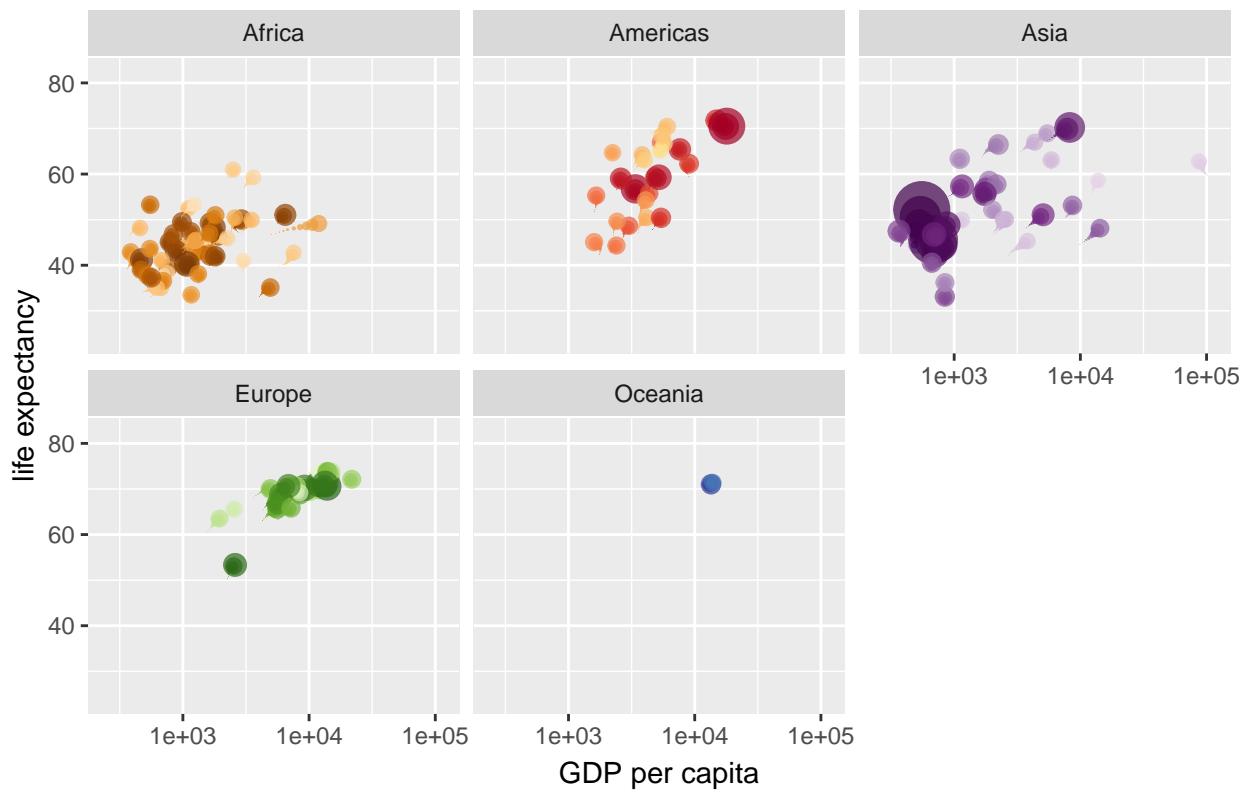
Year: 1964



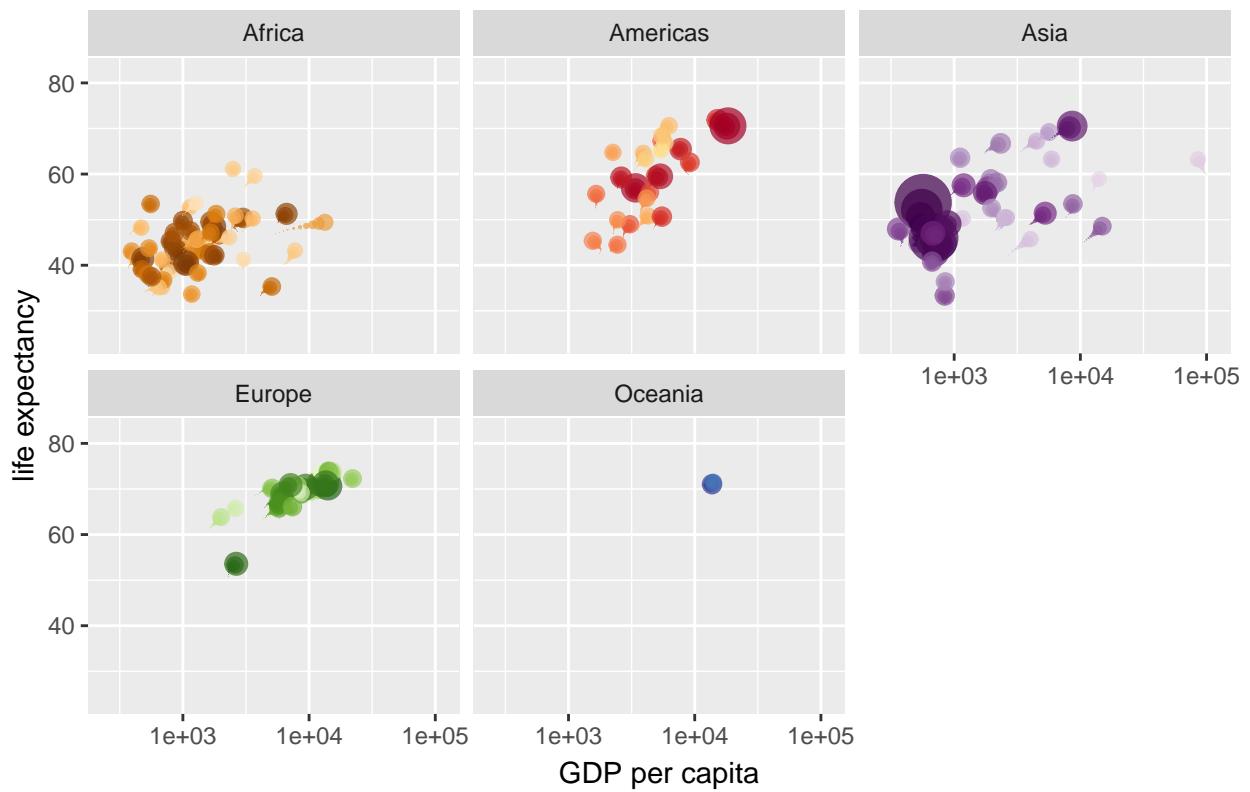
Year: 1964



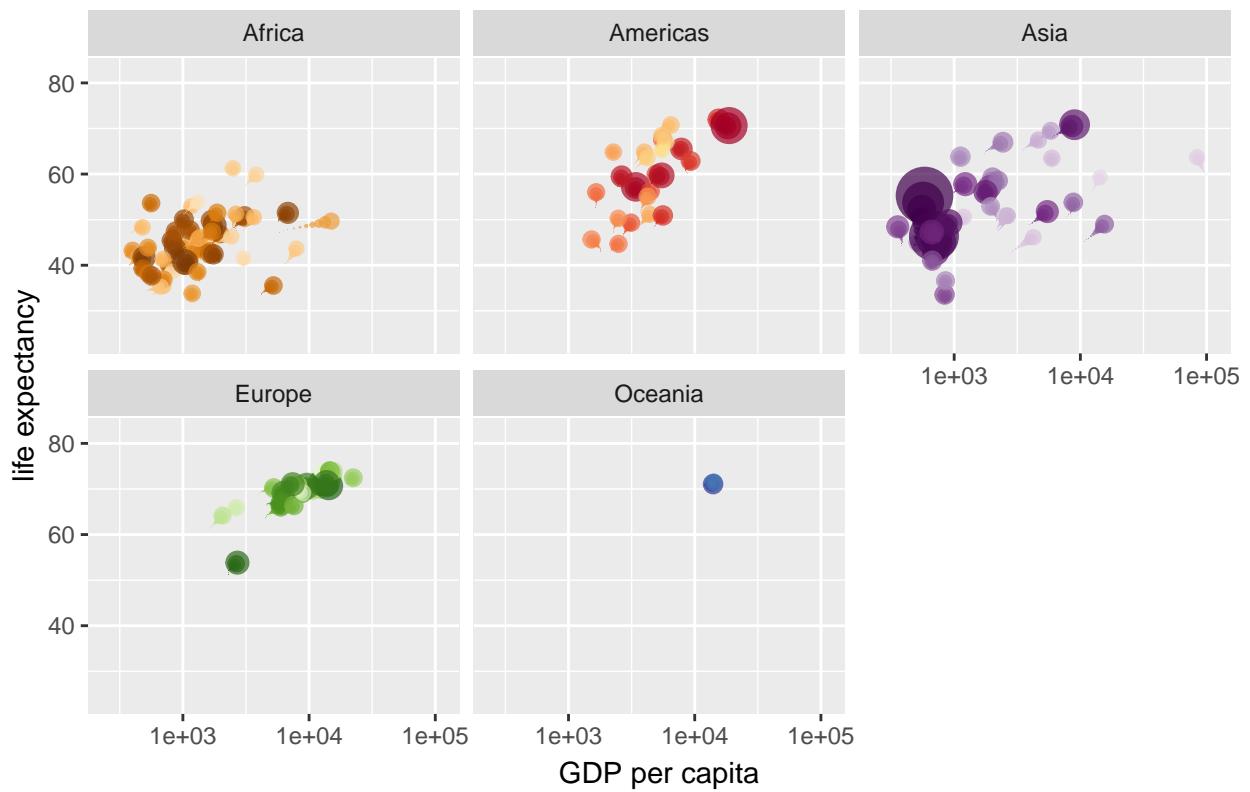
Year: 1965



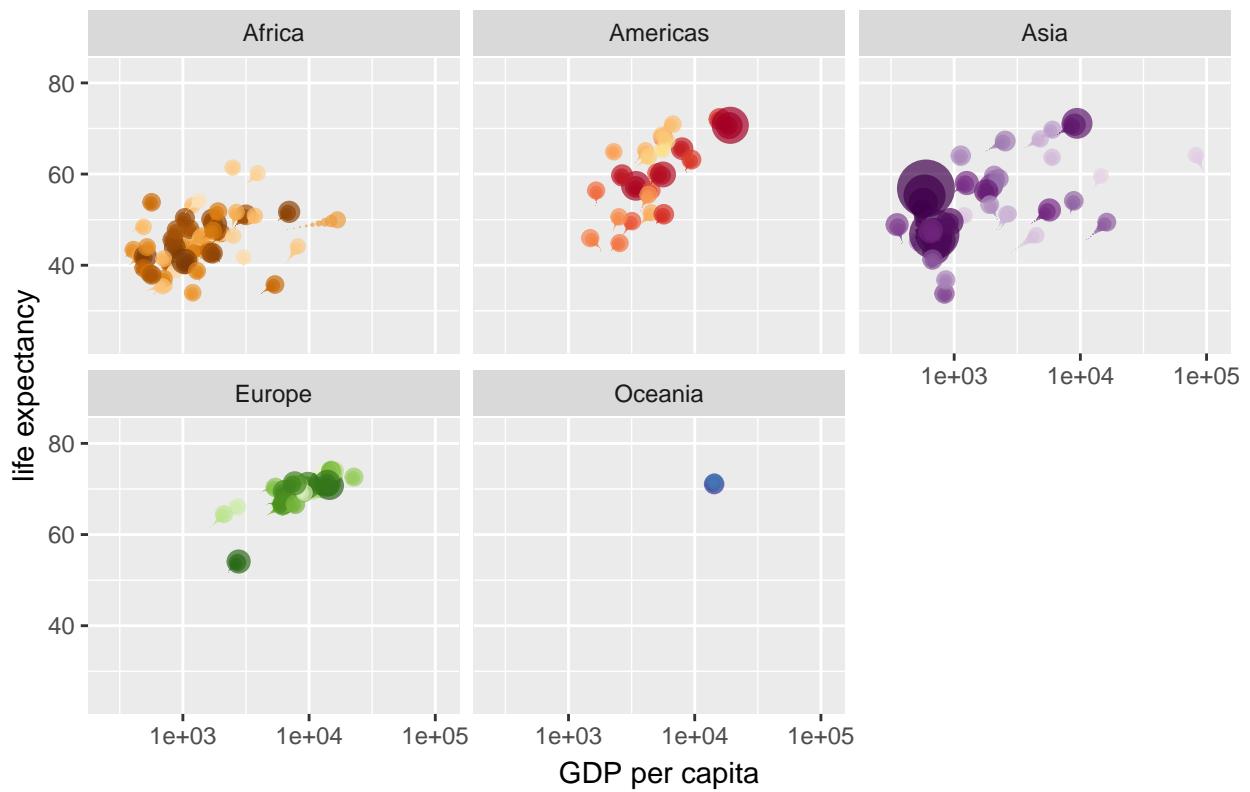
Year: 1965



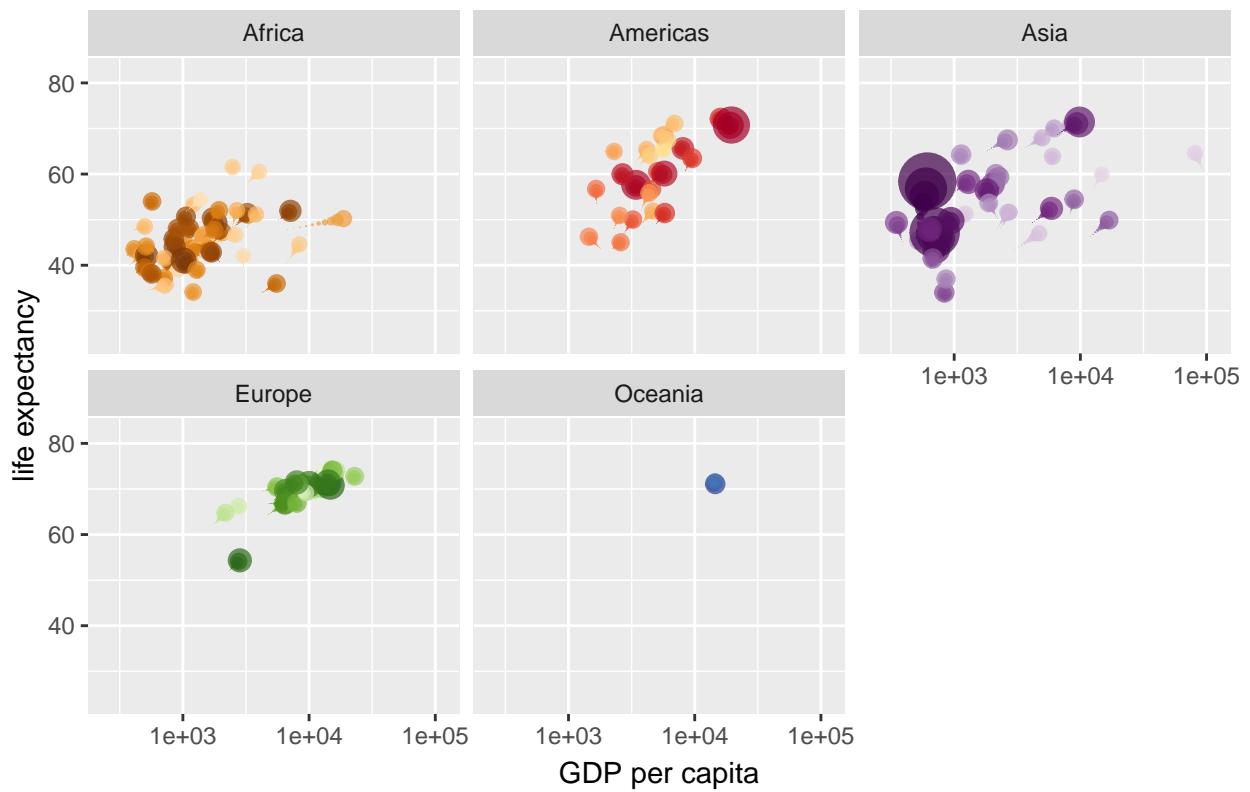
Year: 1966



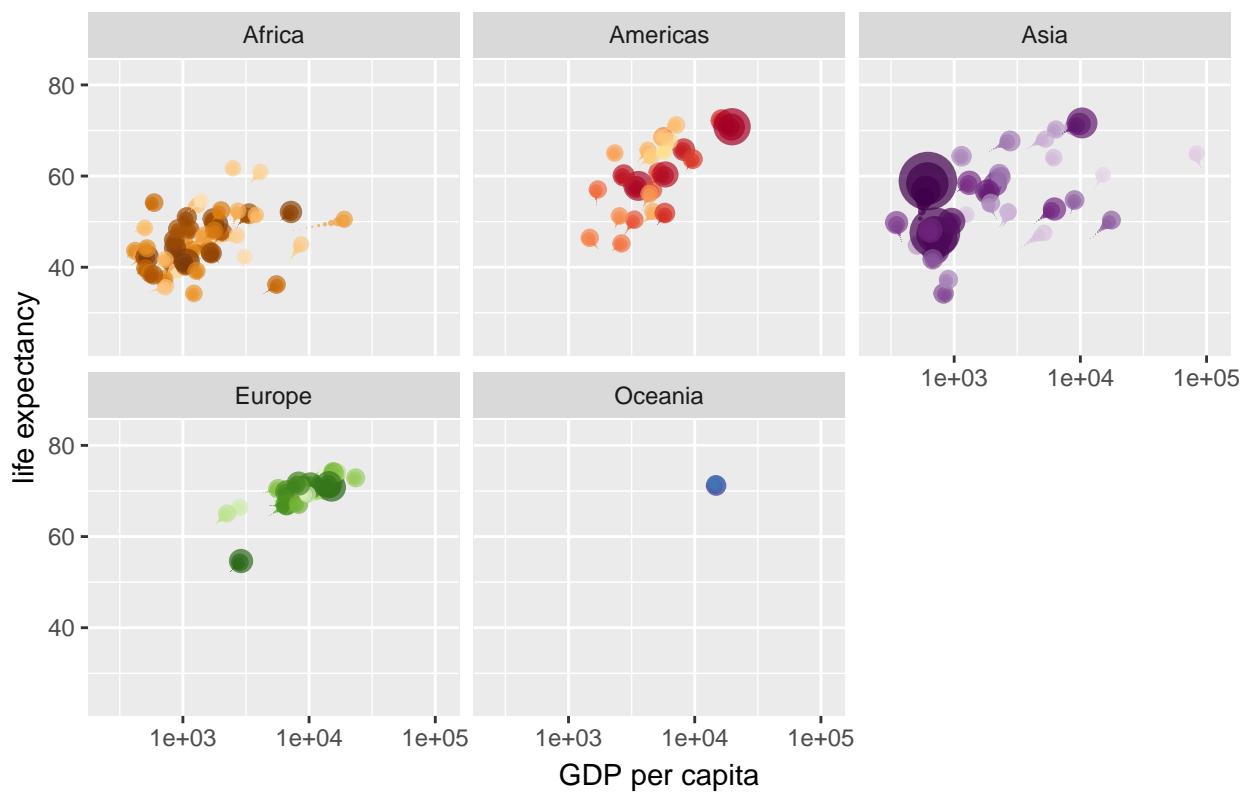
Year: 1966



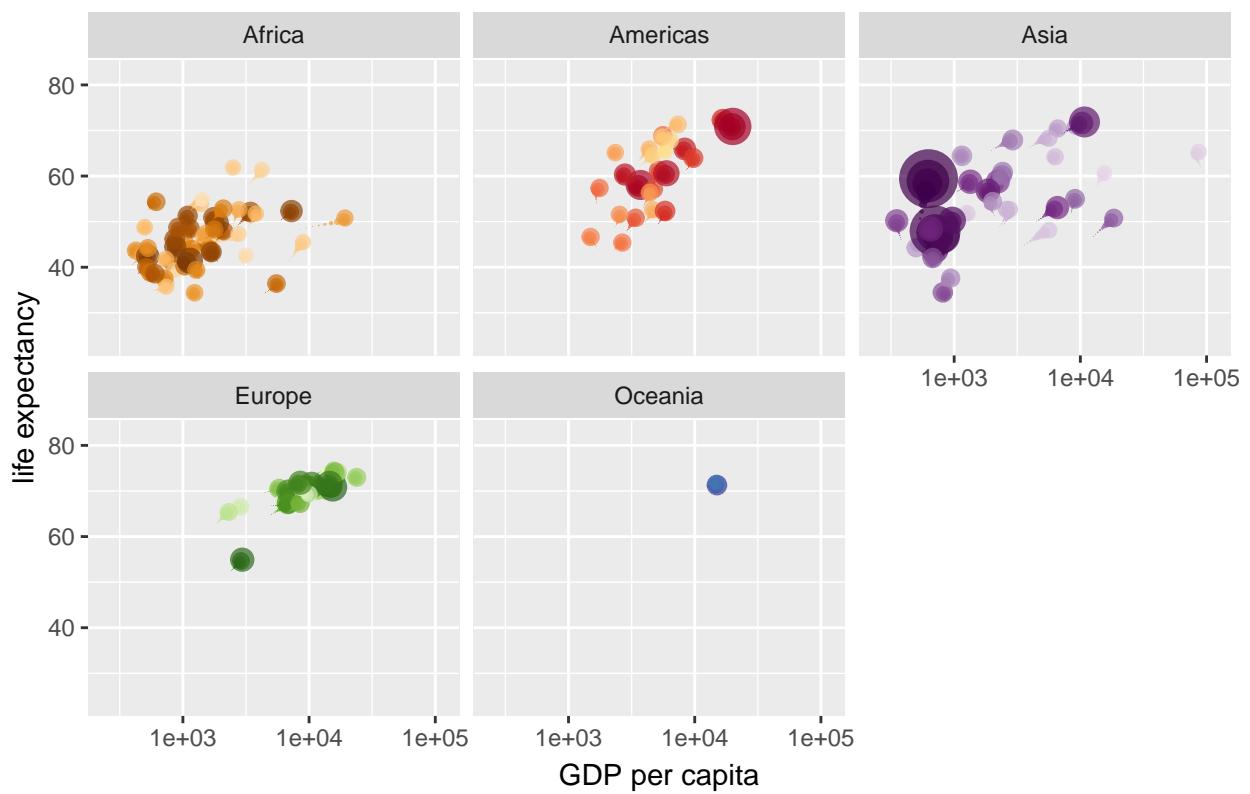
Year: 1967



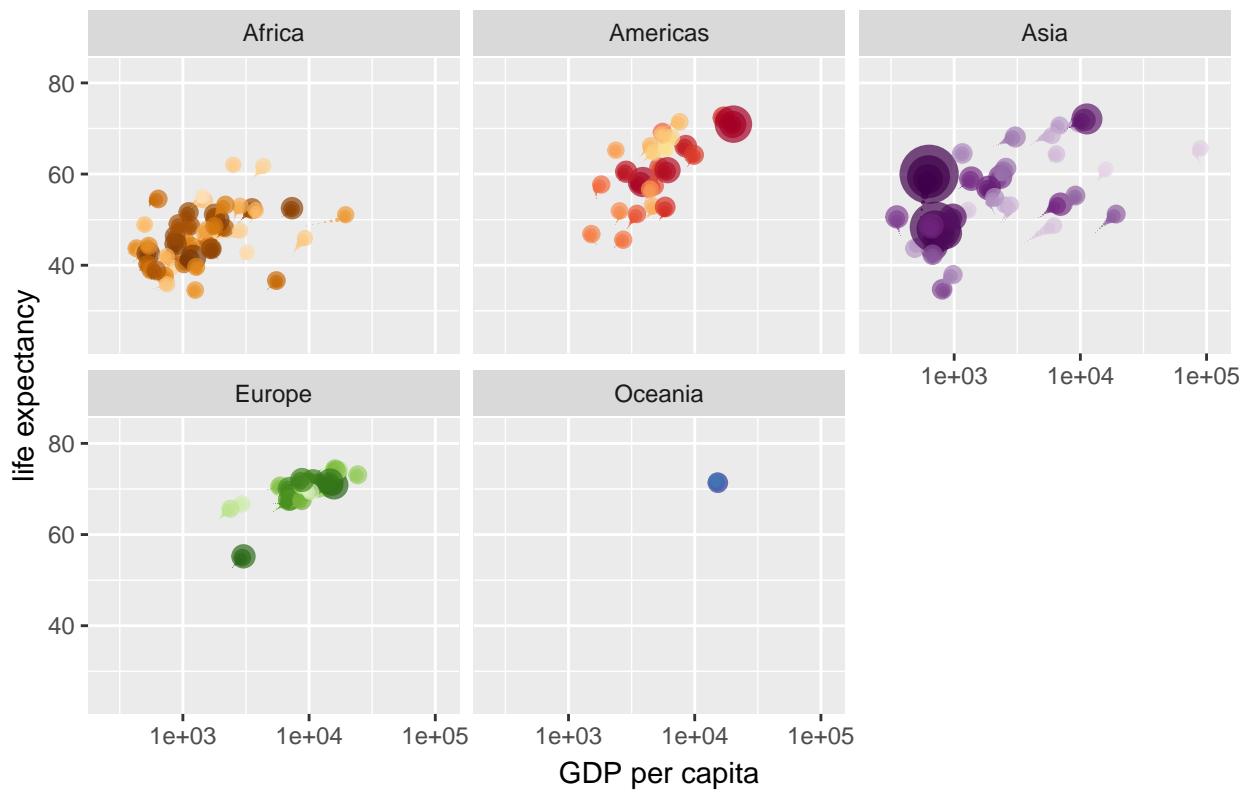
Year: 1968



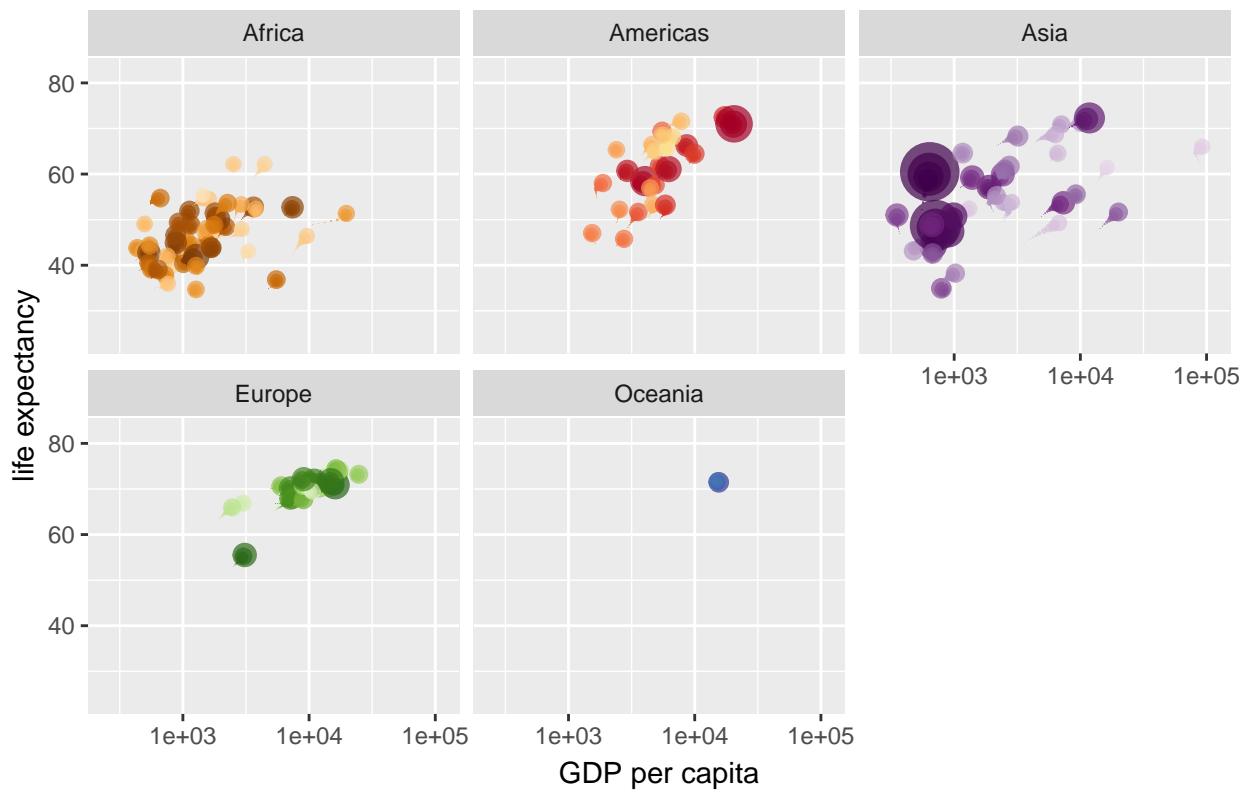
Year: 1968



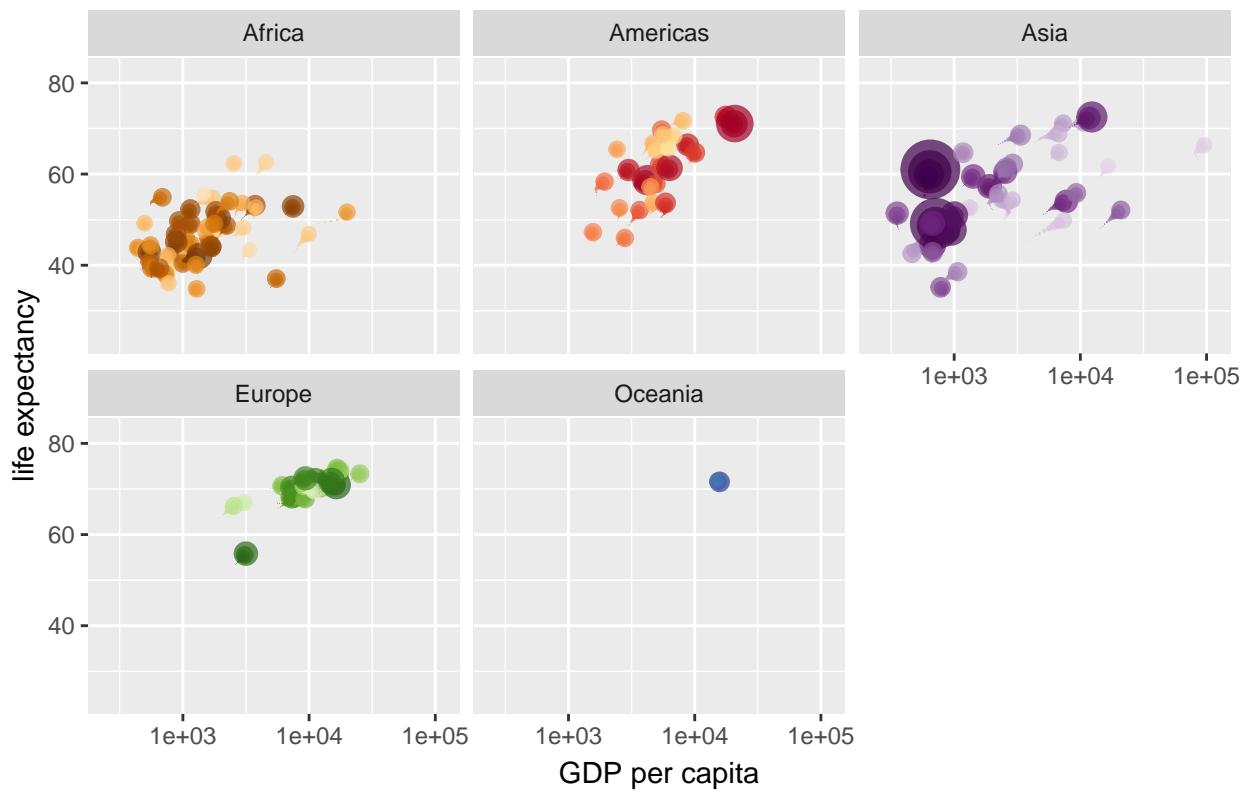
Year: 1969



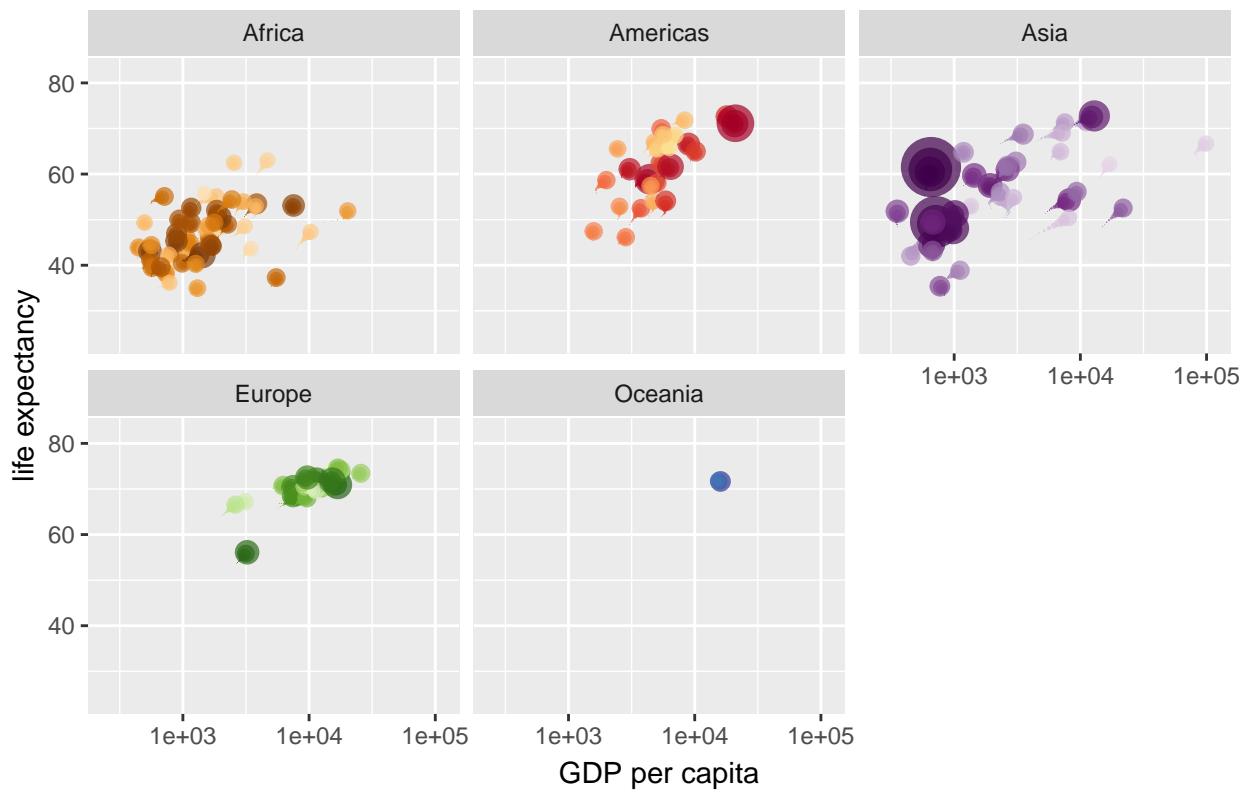
Year: 1969



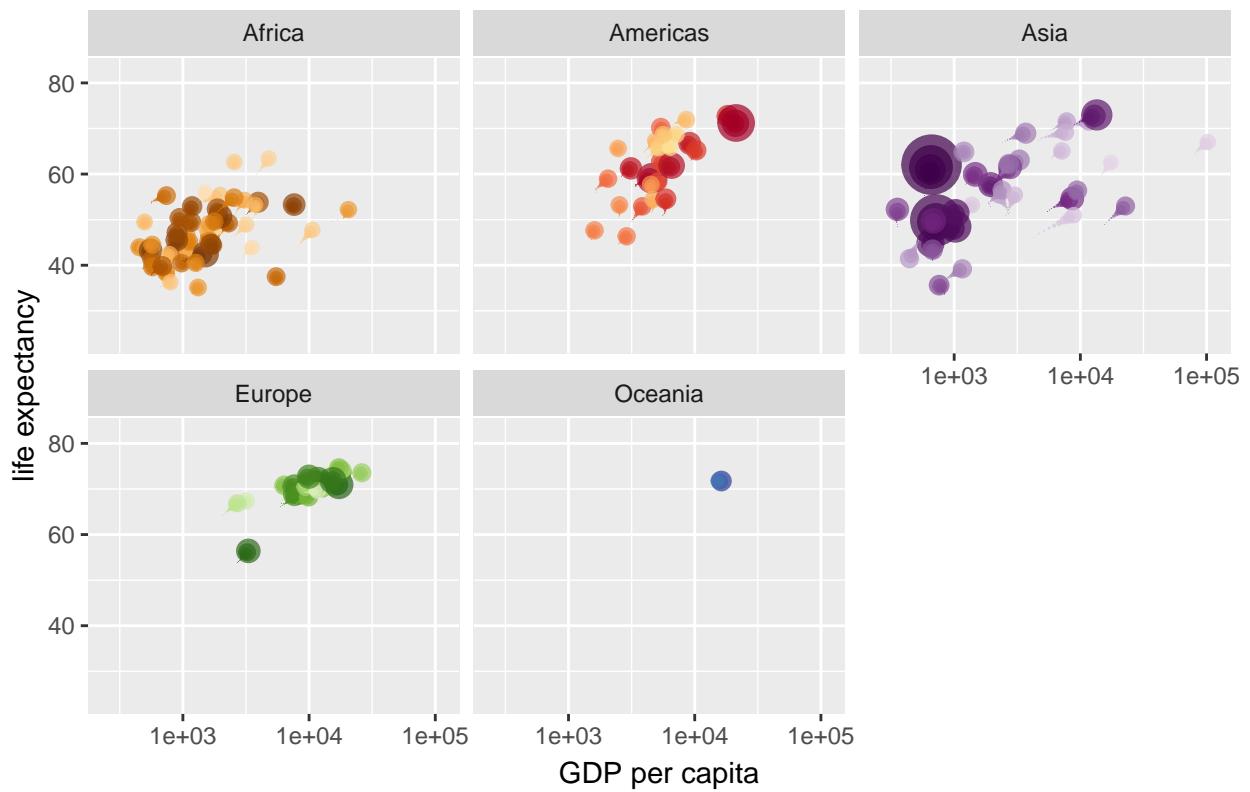
Year: 1970



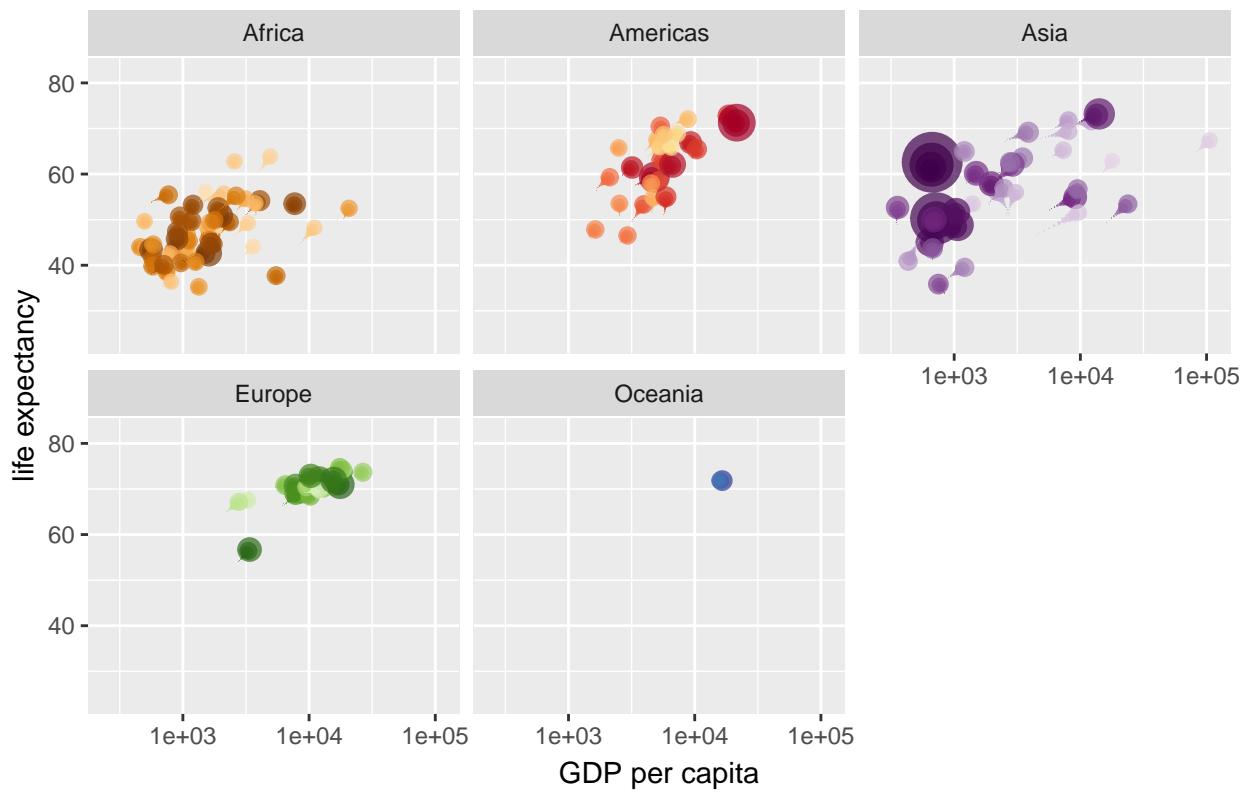
Year: 1970



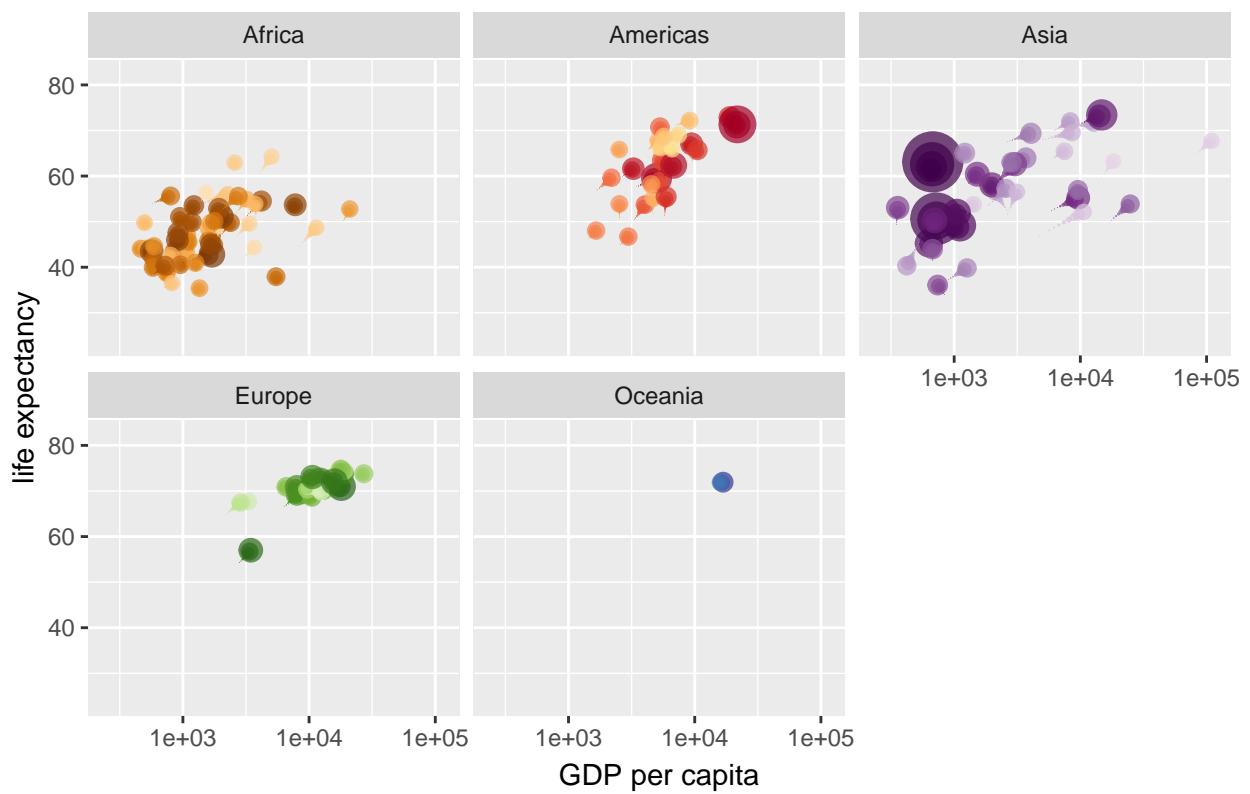
Year: 1971



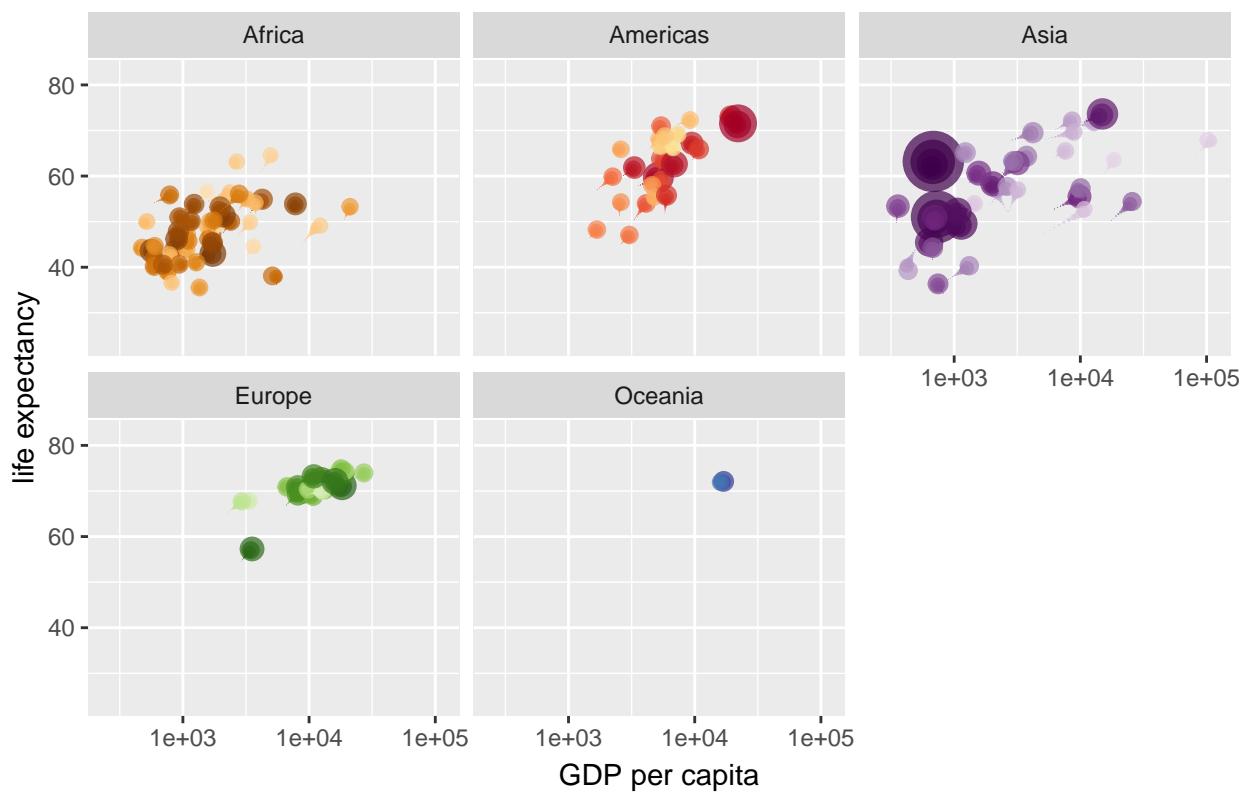
Year: 1971



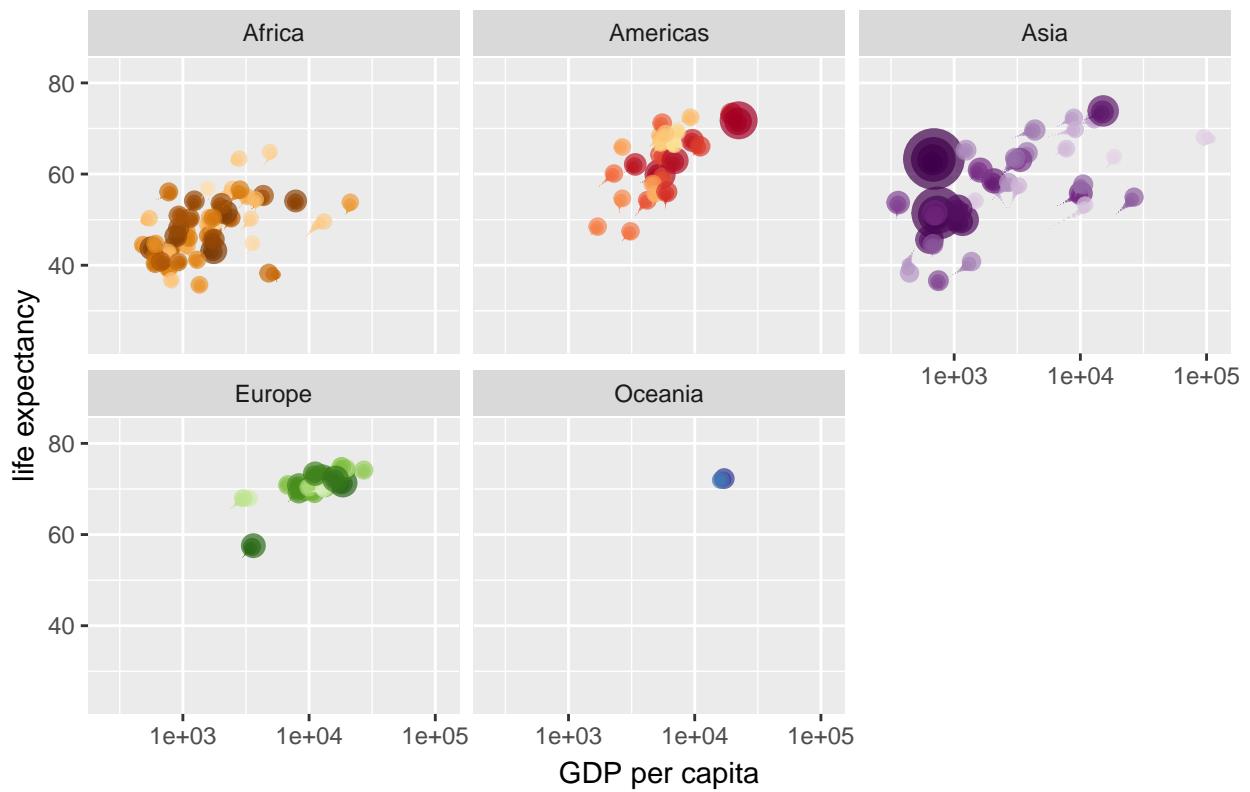
Year: 1972



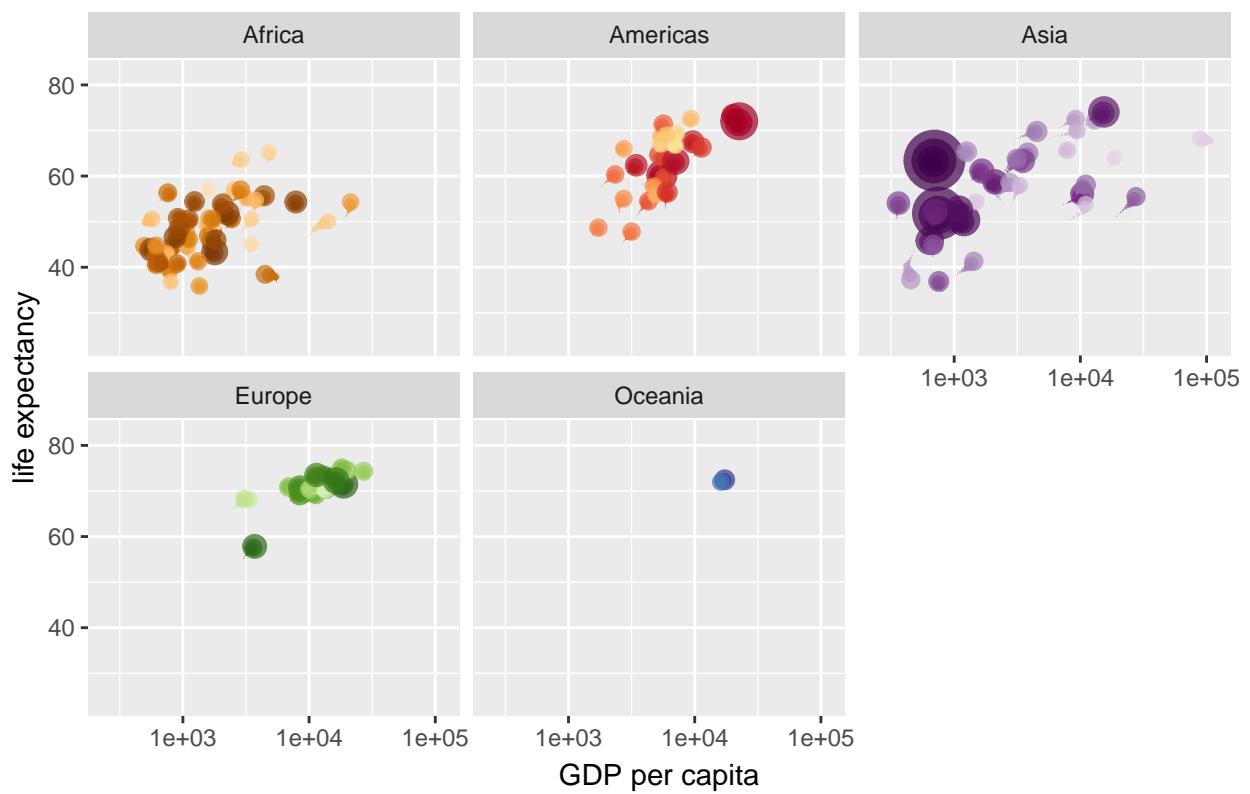
Year: 1973



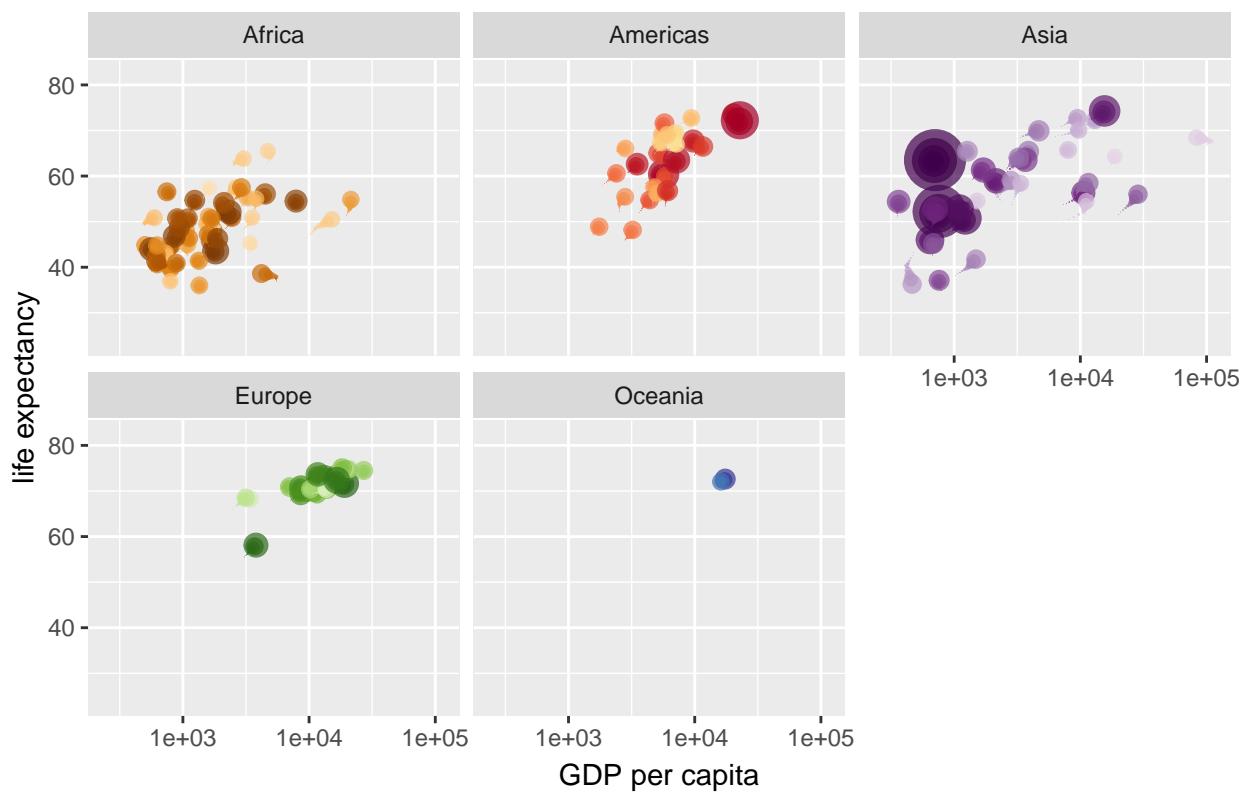
Year: 1973



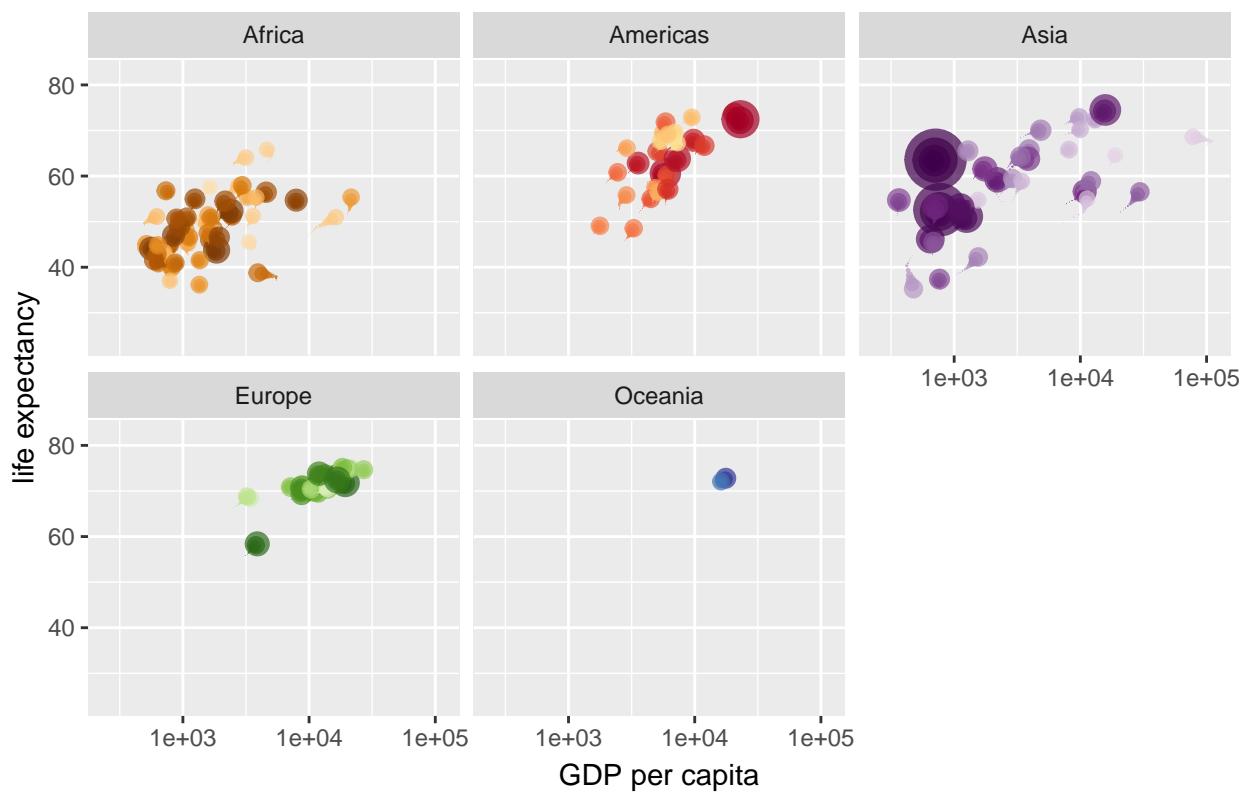
Year: 1974



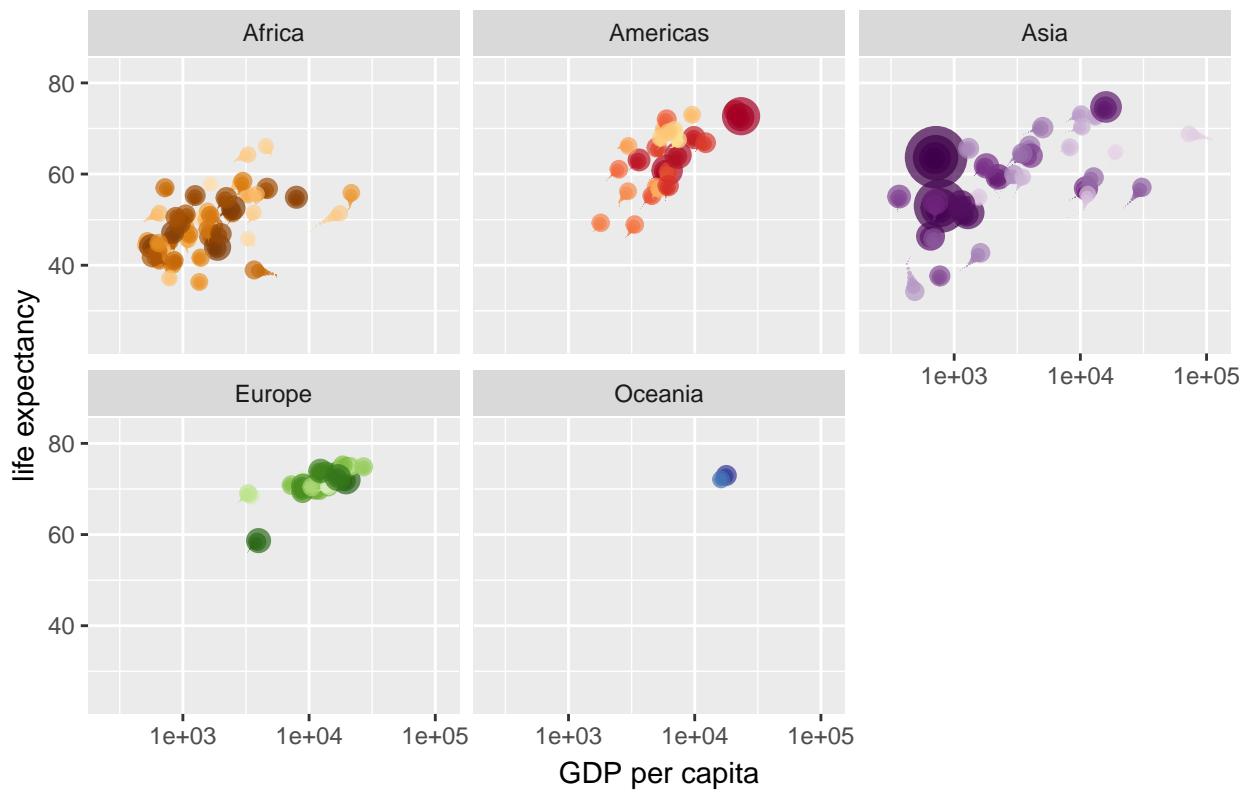
Year: 1974



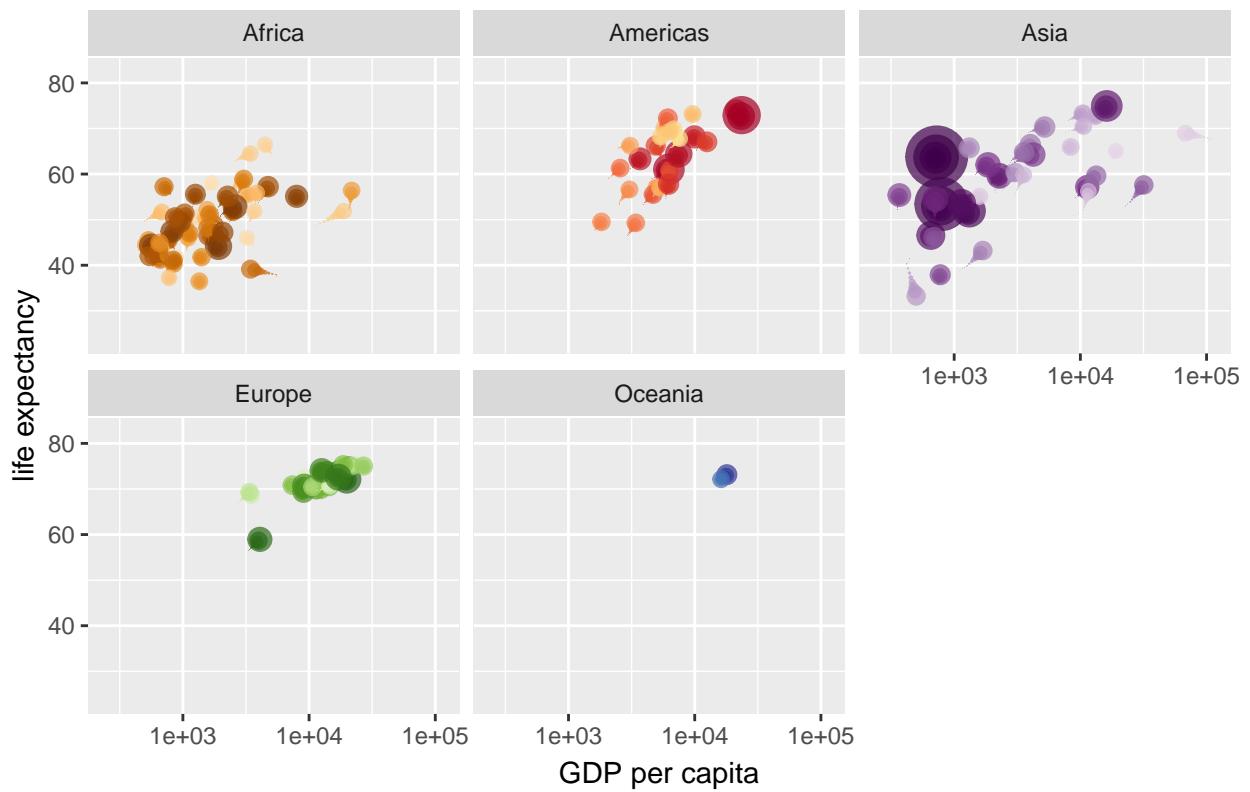
Year: 1975



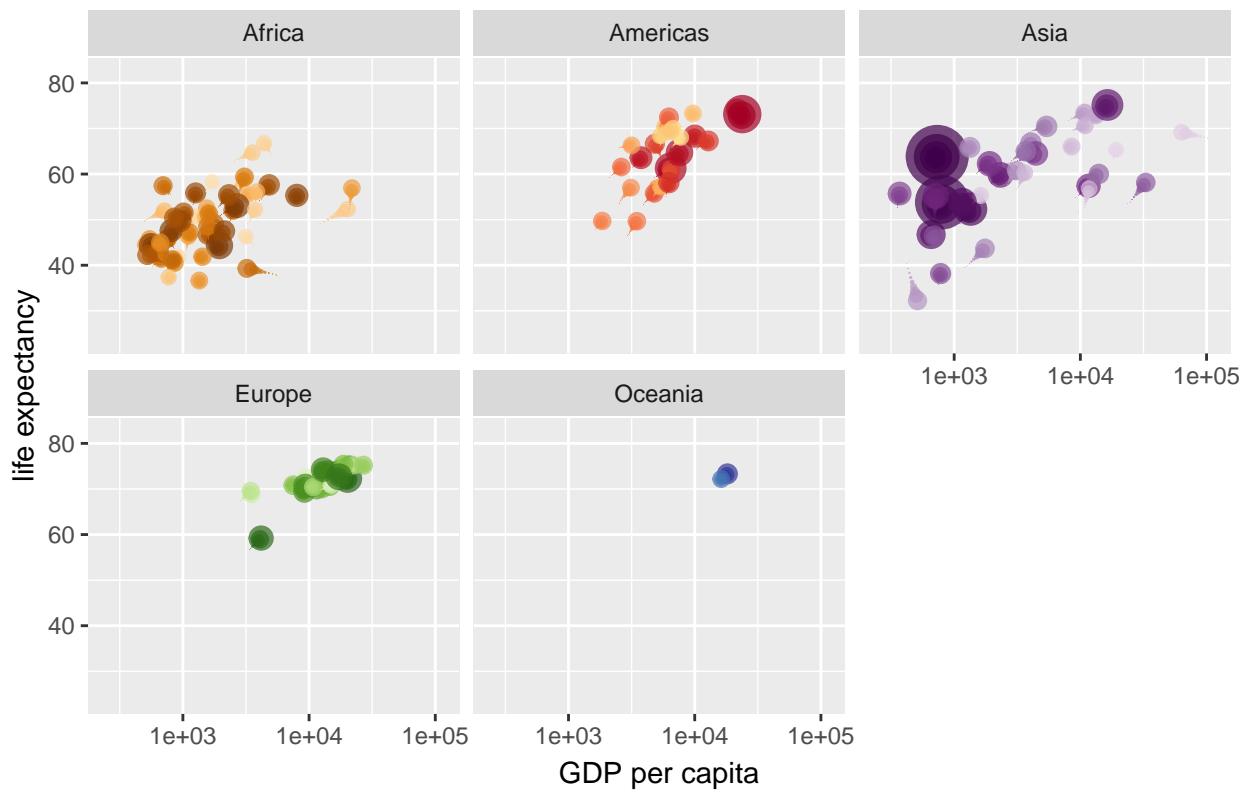
Year: 1975



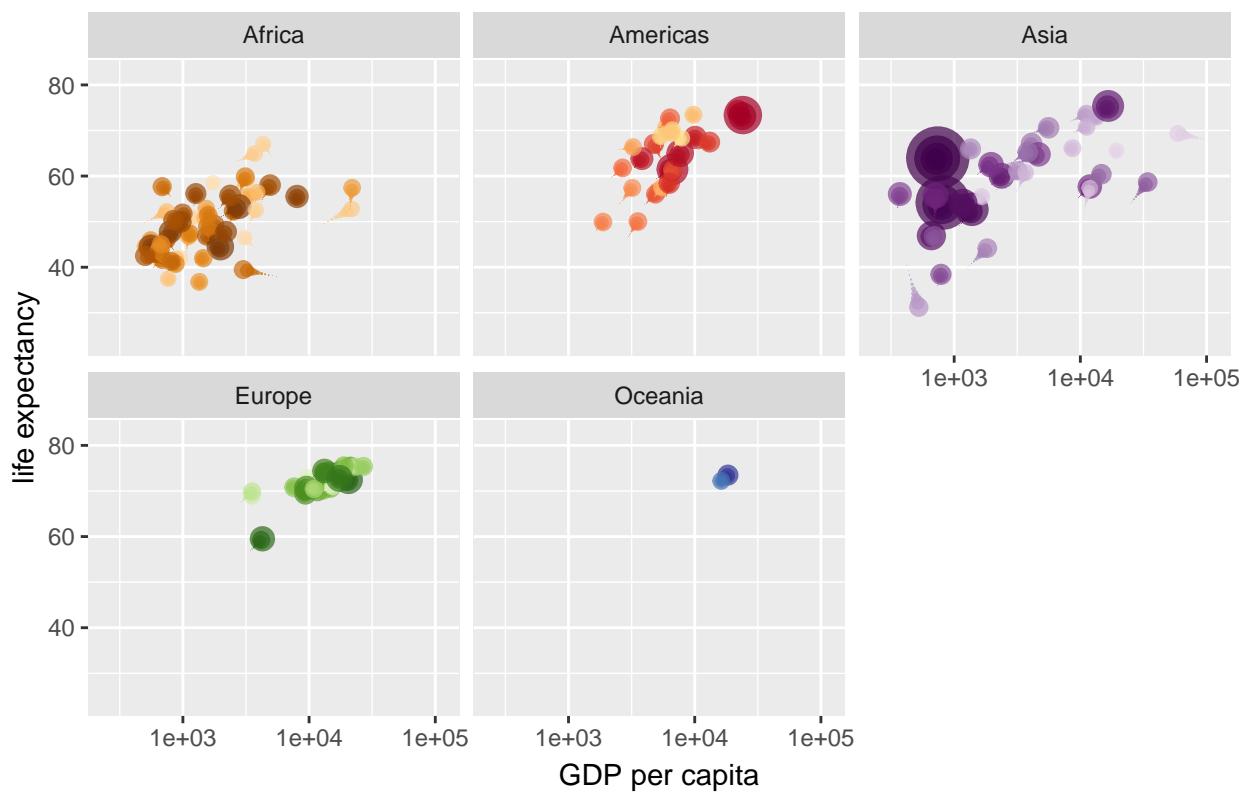
Year: 1976



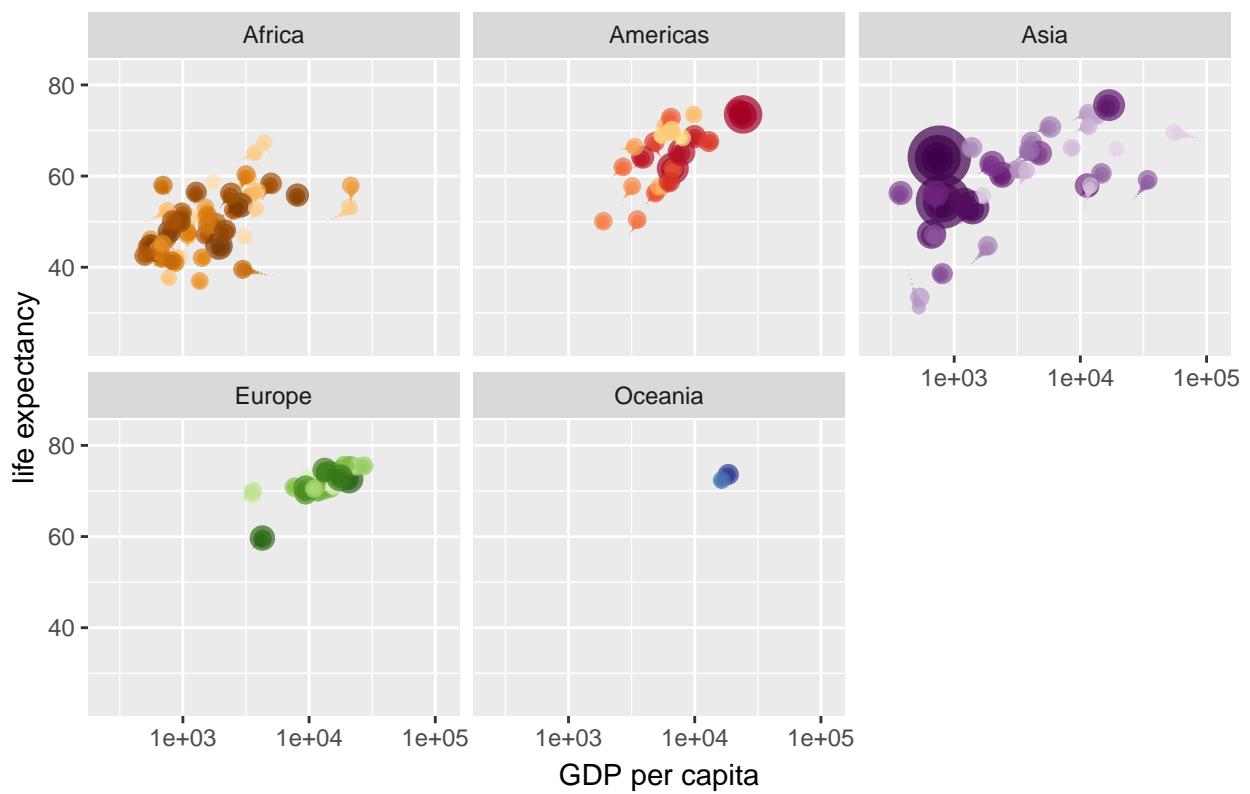
Year: 1976



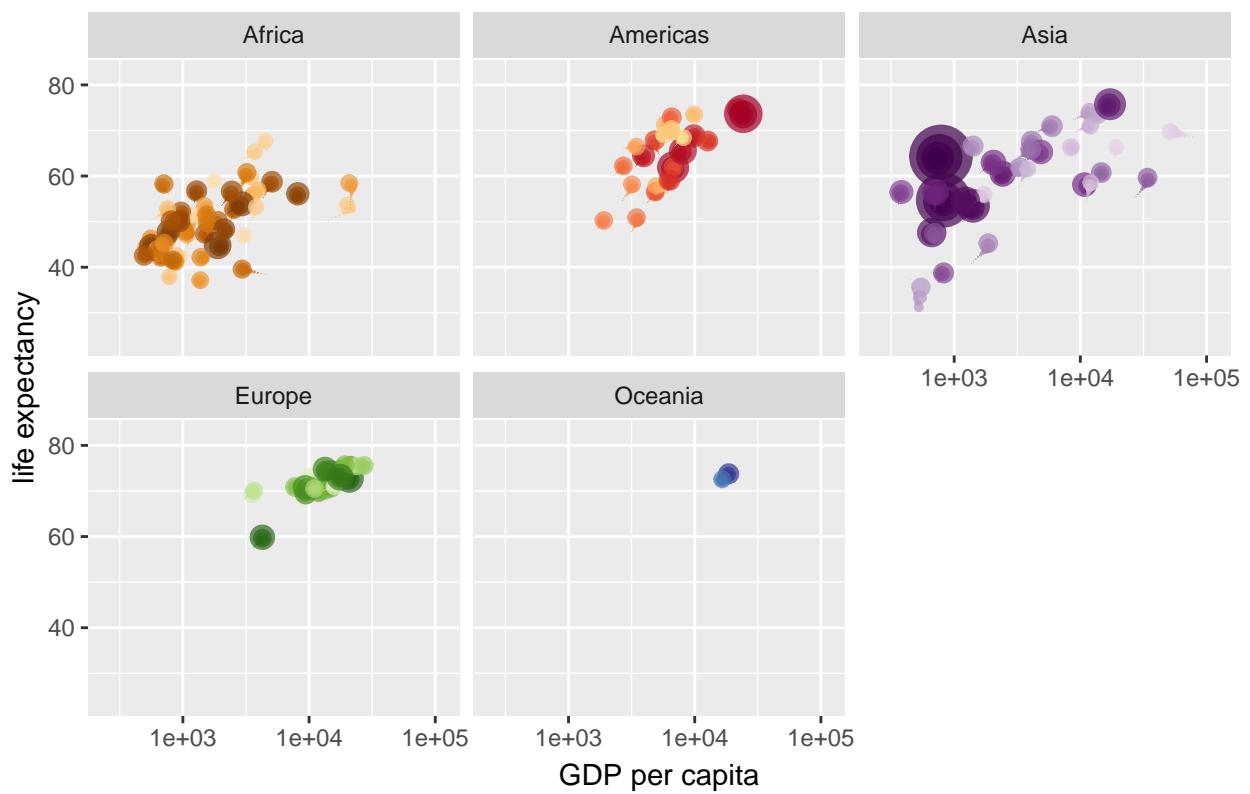
Year: 1977



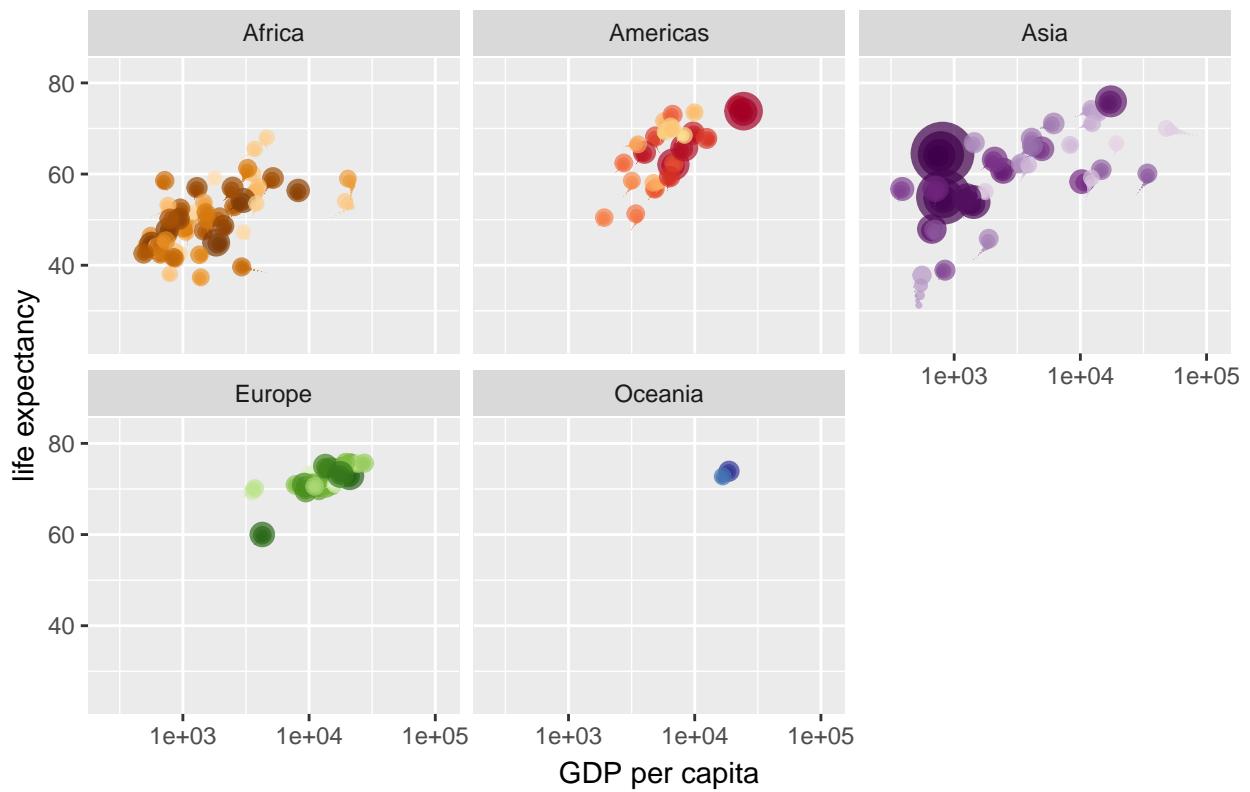
Year: 1978



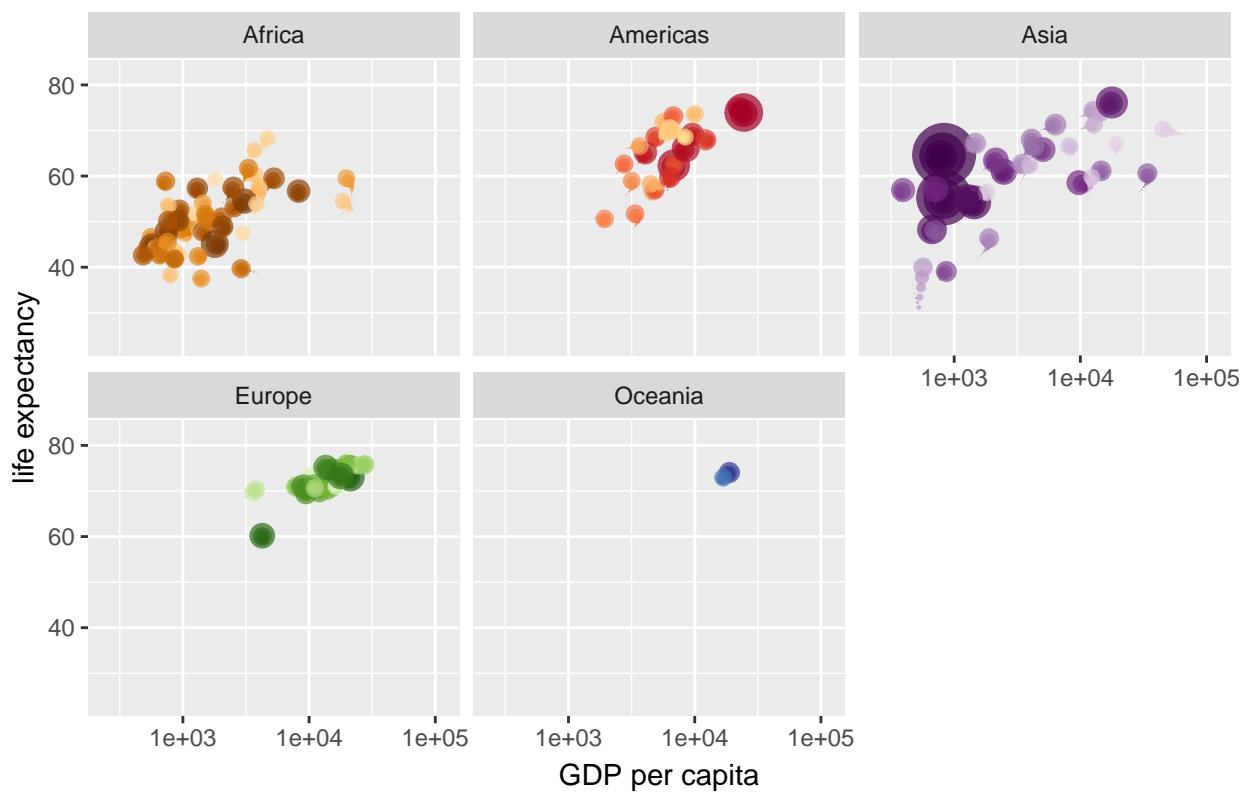
Year: 1978



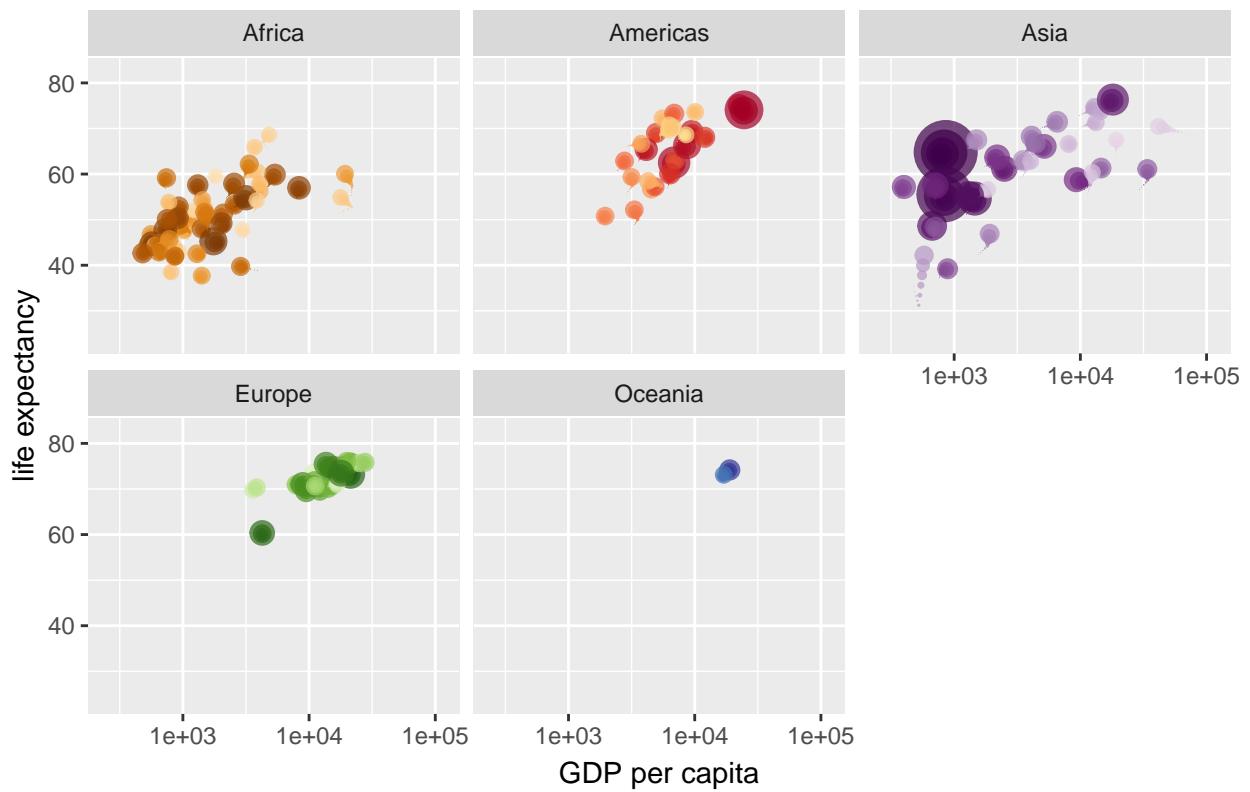
Year: 1979



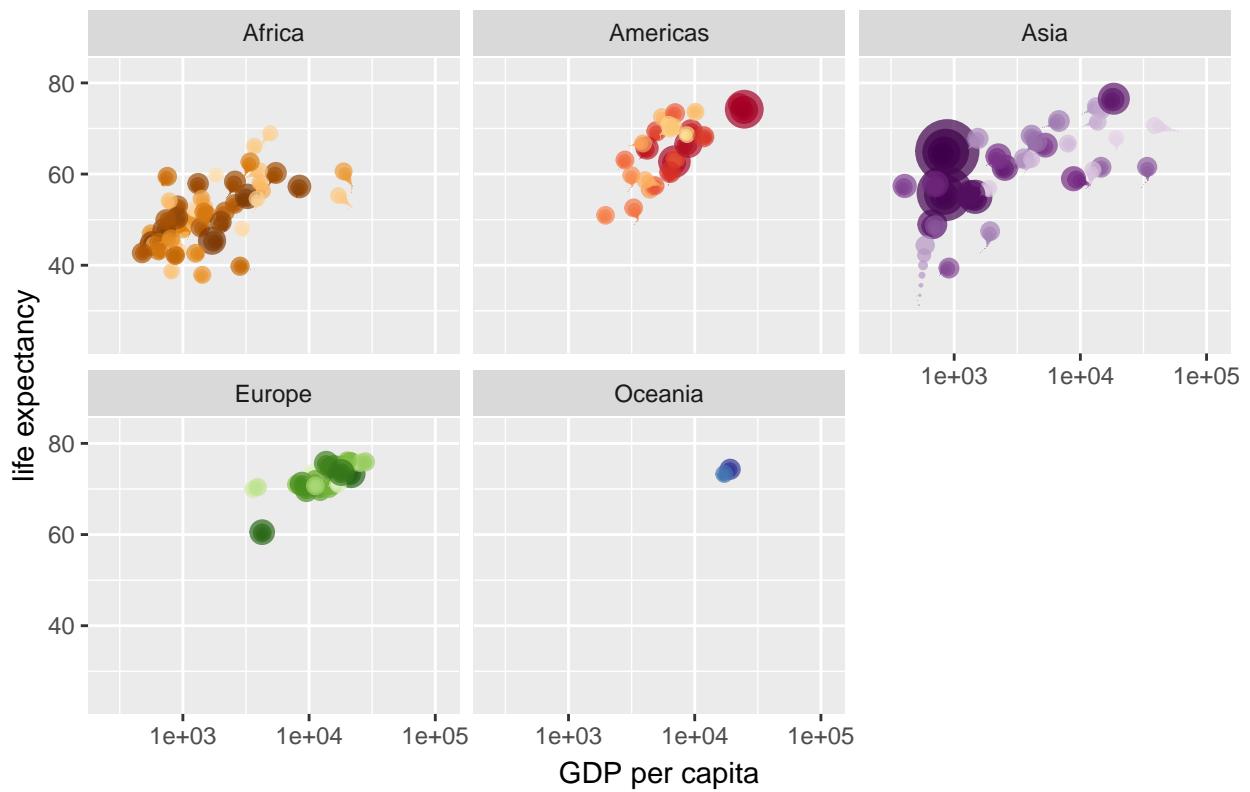
Year: 1979



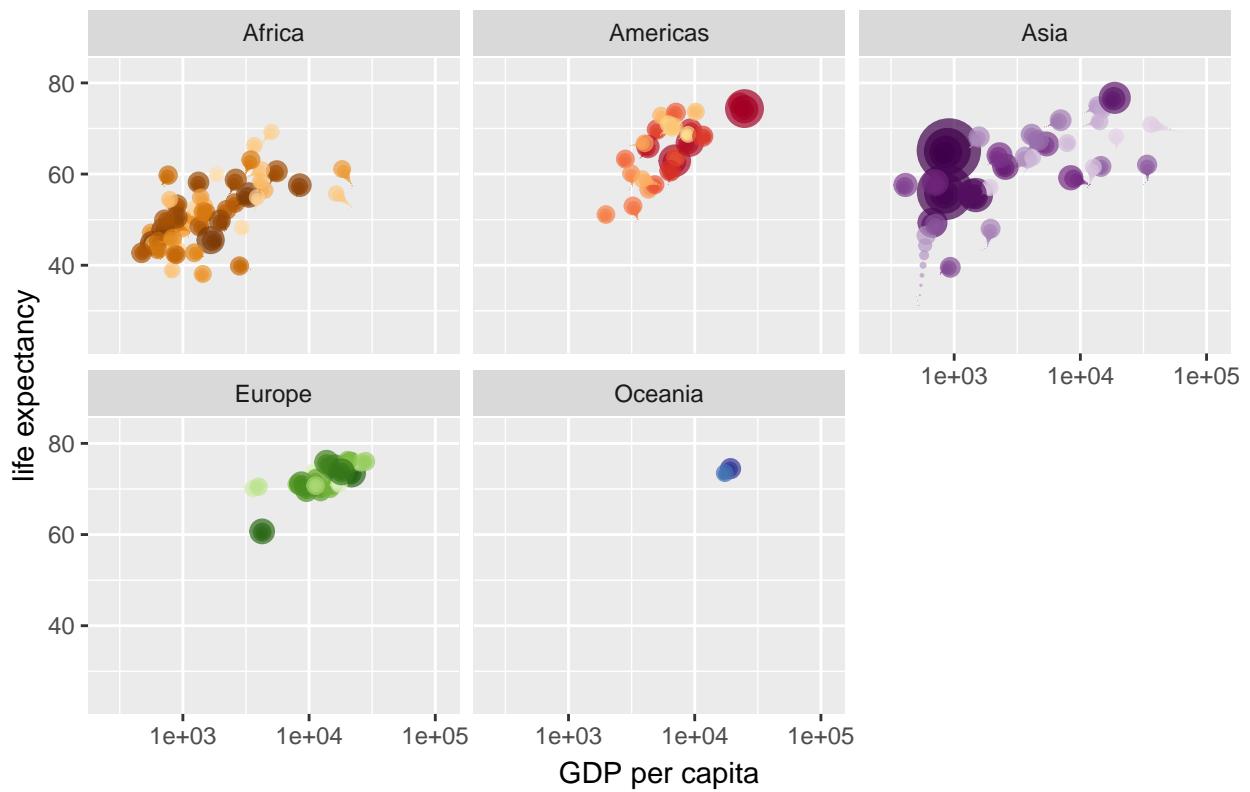
Year: 1980



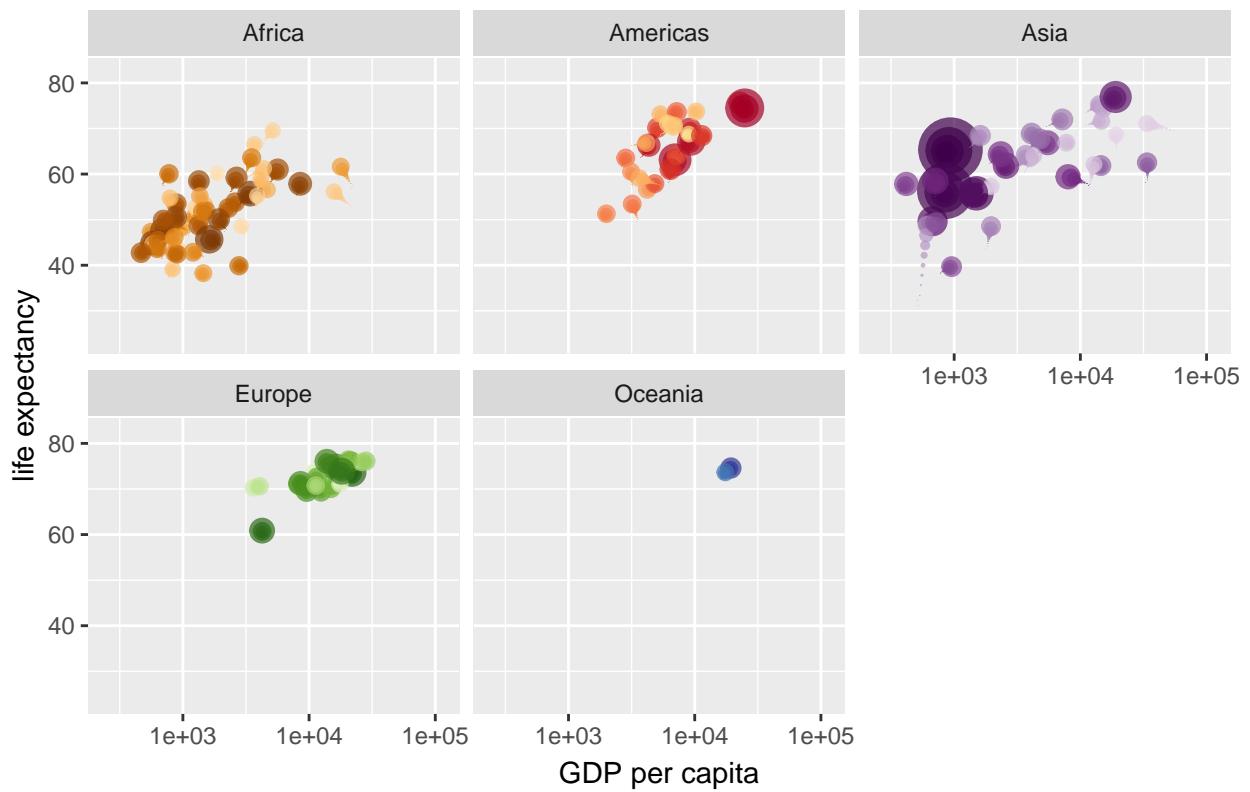
Year: 1980



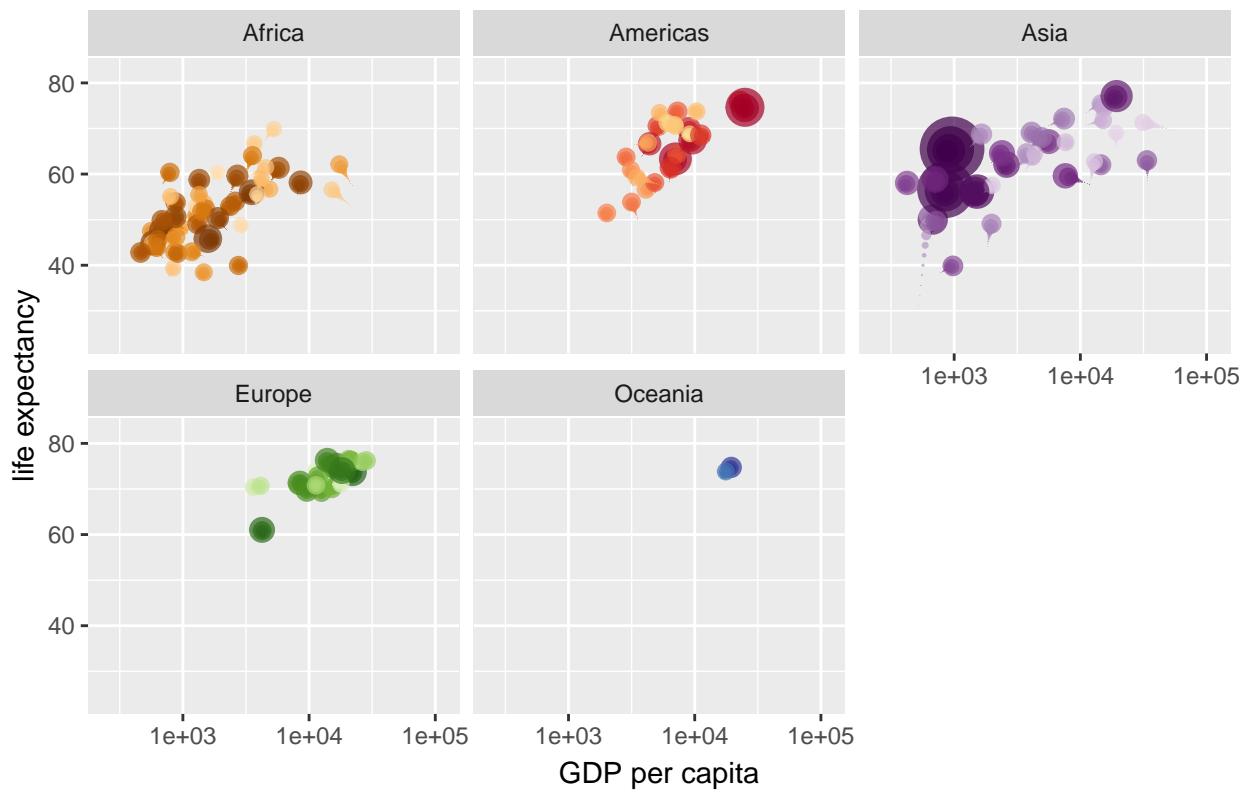
Year: 1981



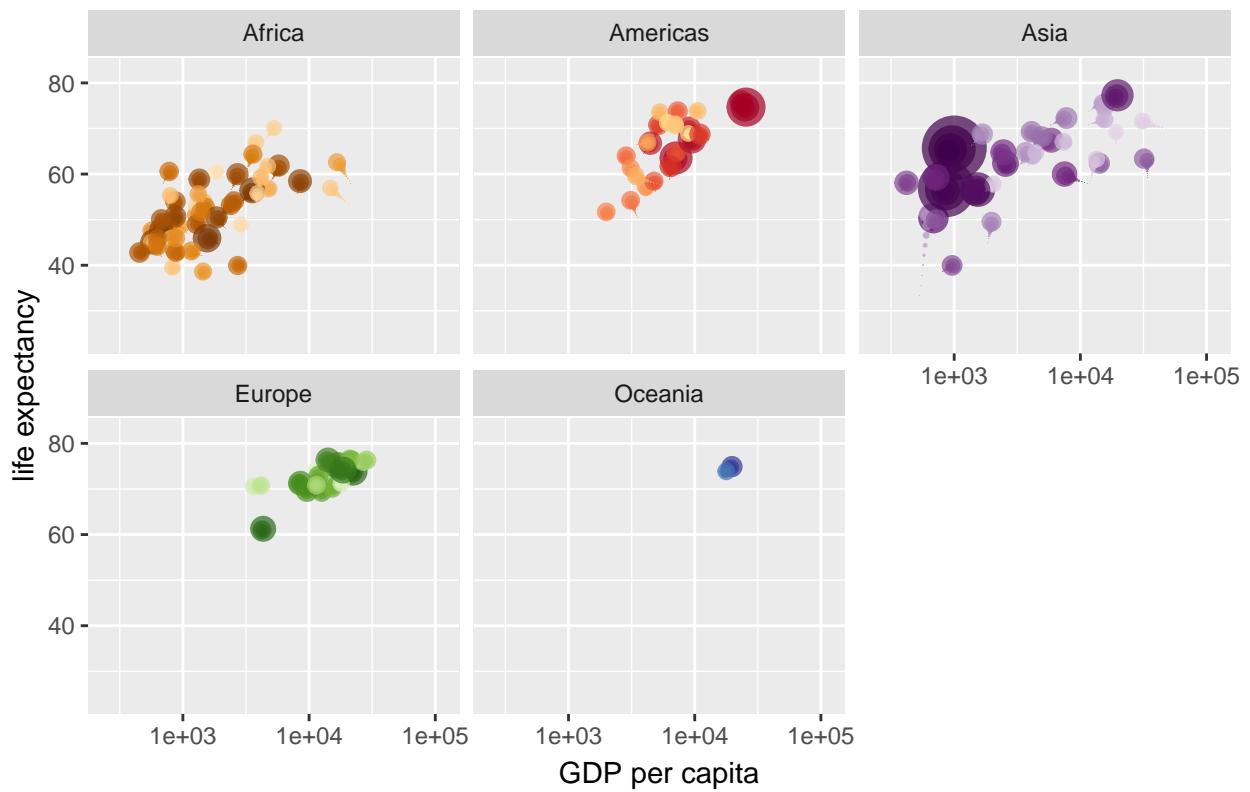
Year: 1981



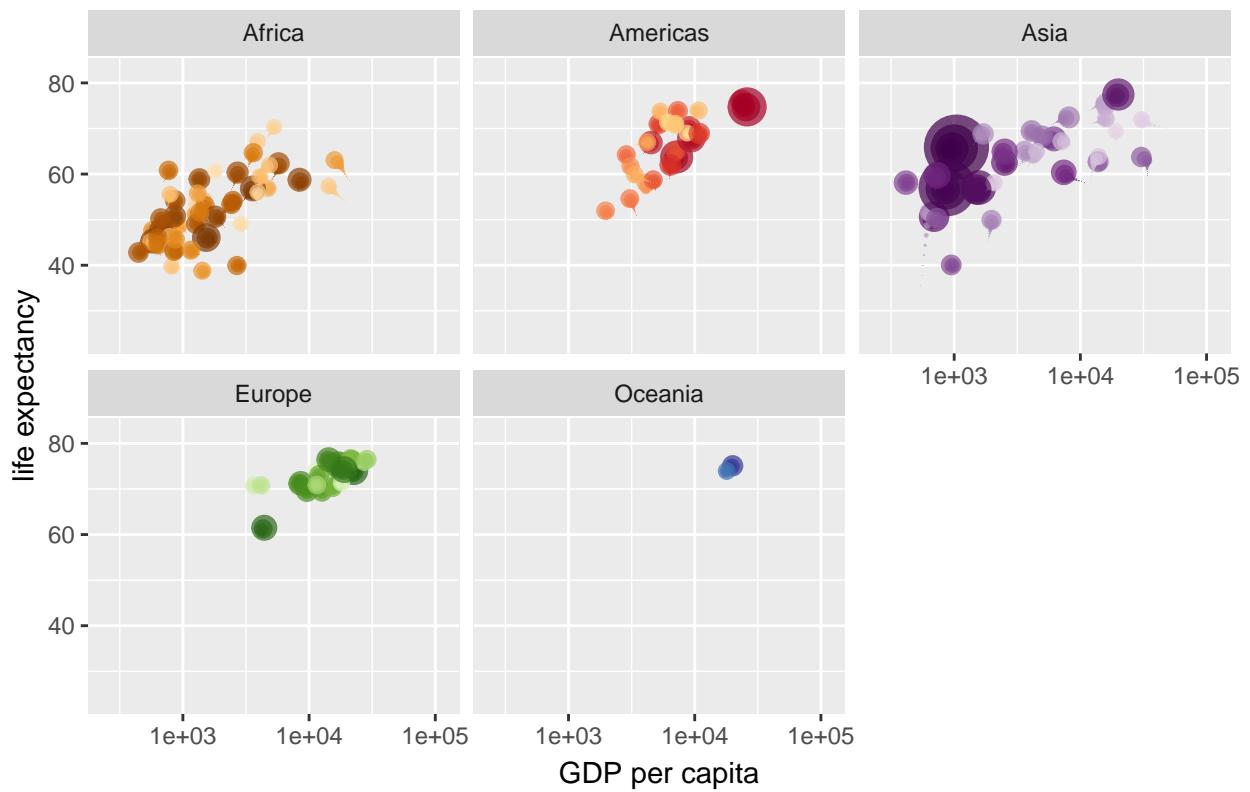
Year: 1982



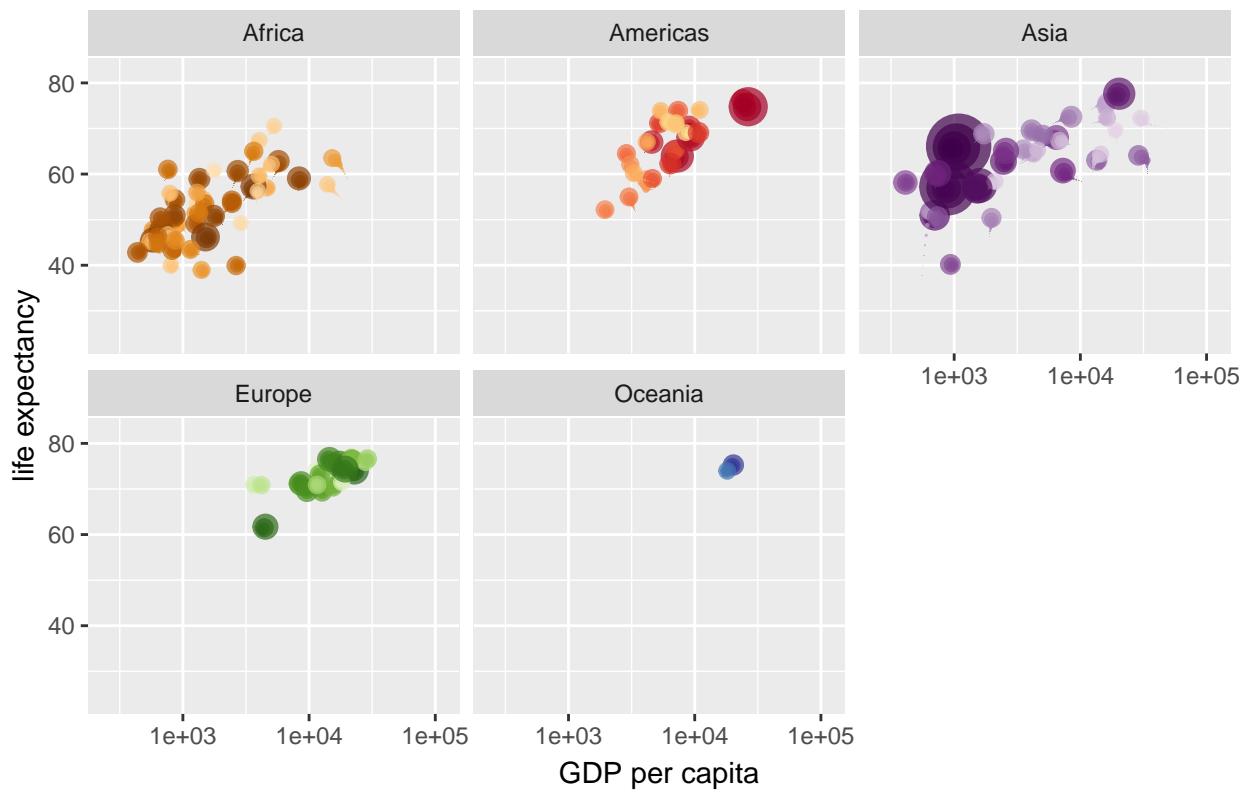
Year: 1983



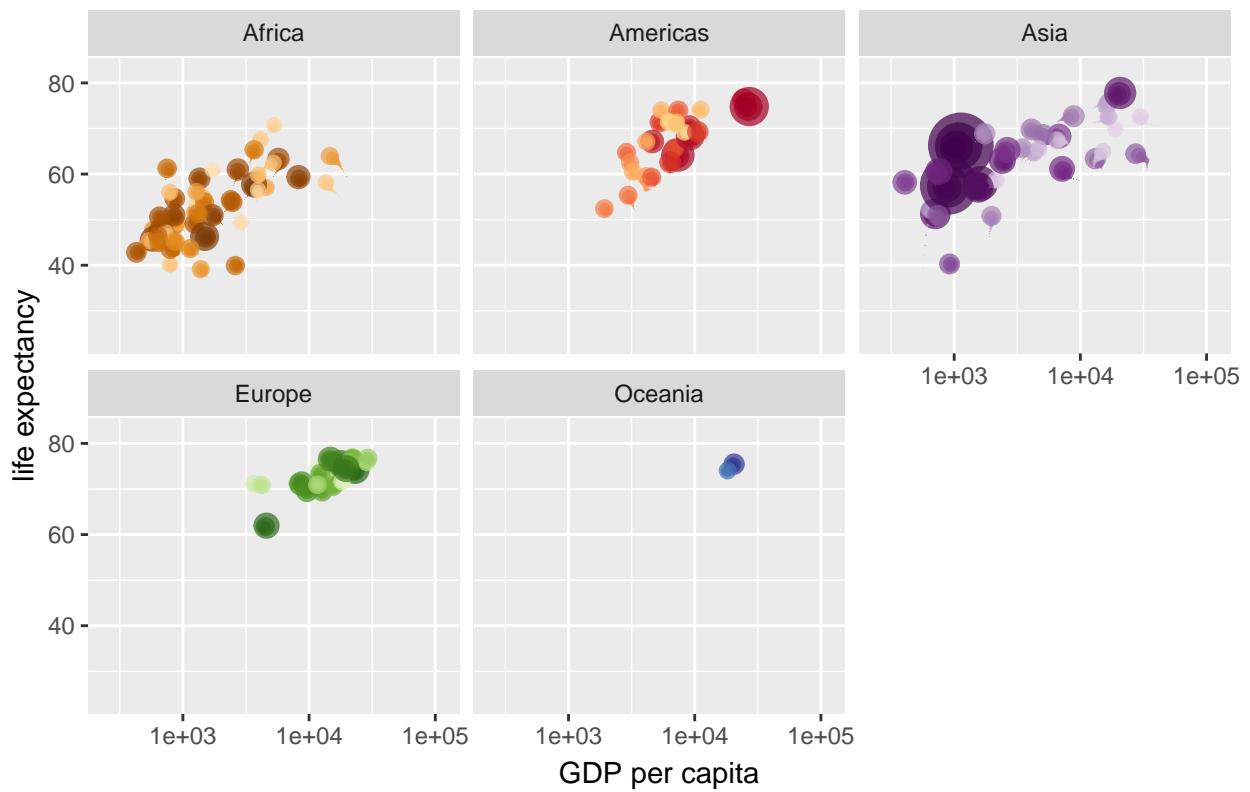
Year: 1983



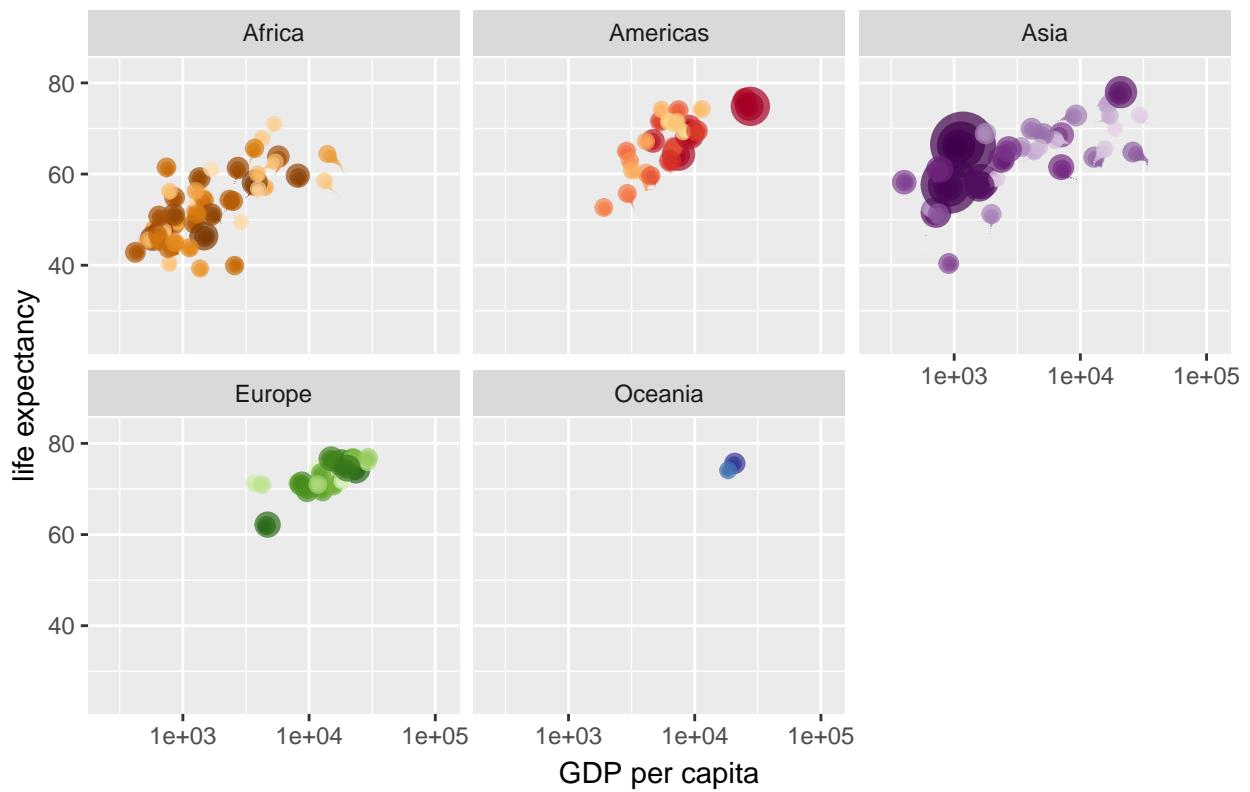
Year: 1984



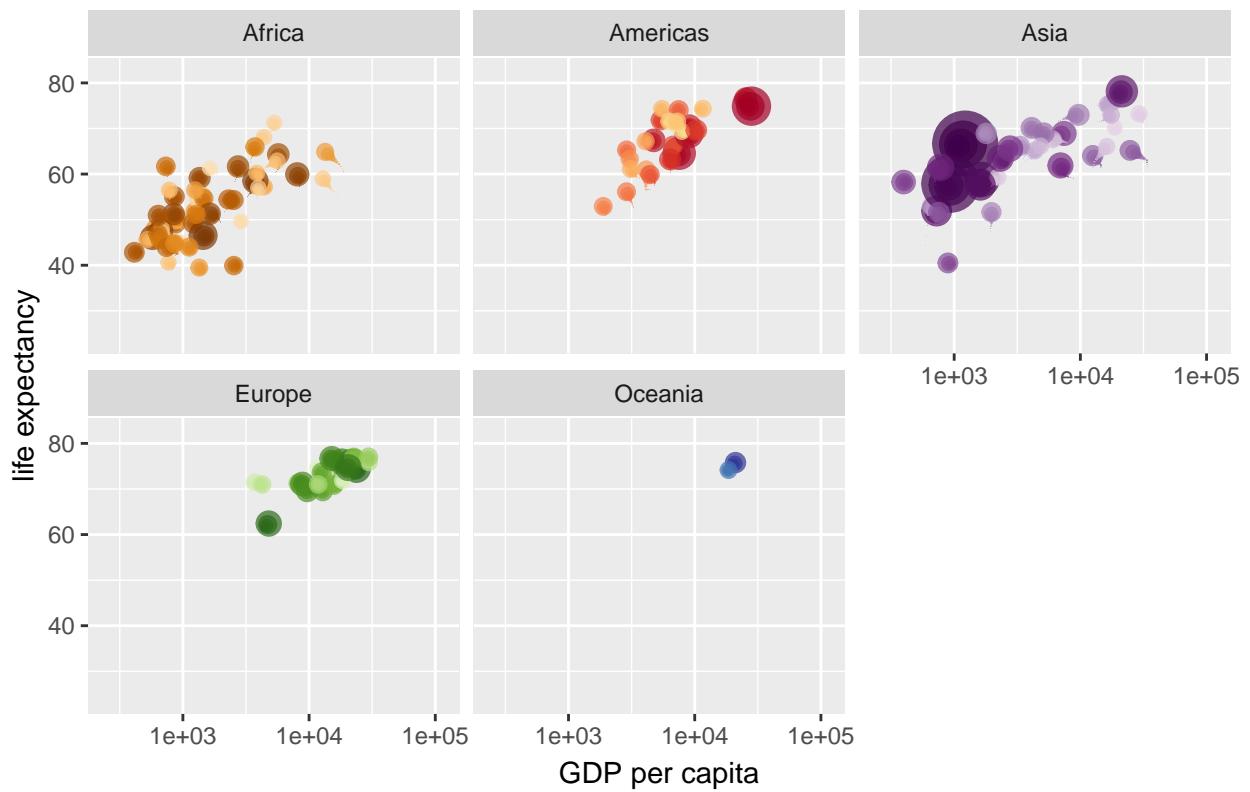
Year: 1984



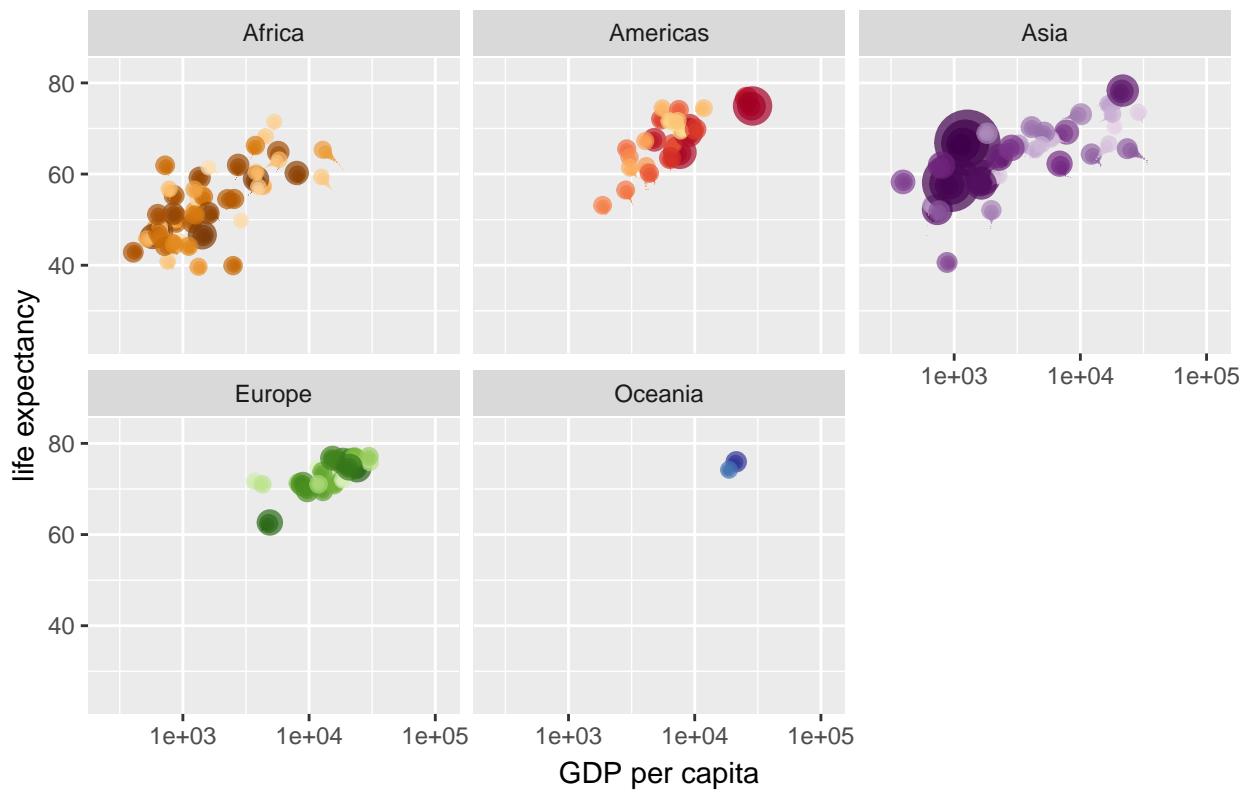
Year: 1985



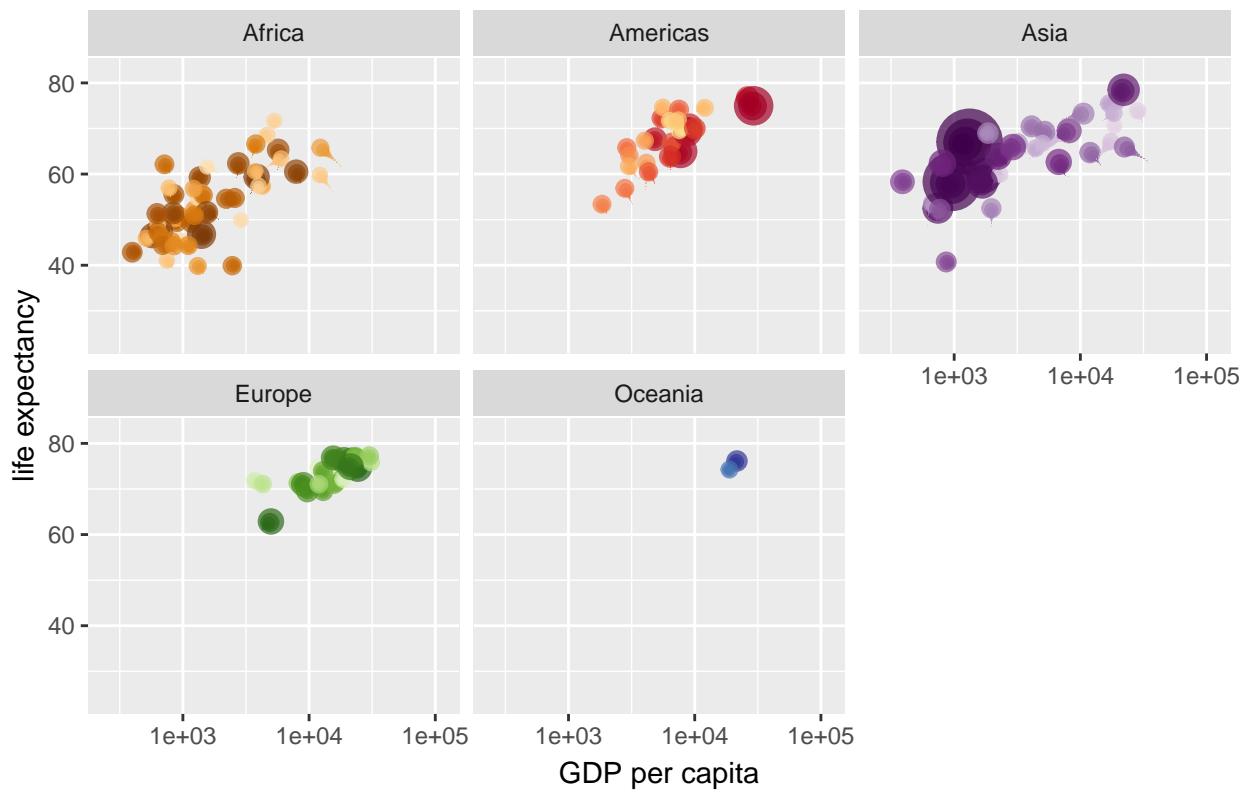
Year: 1985



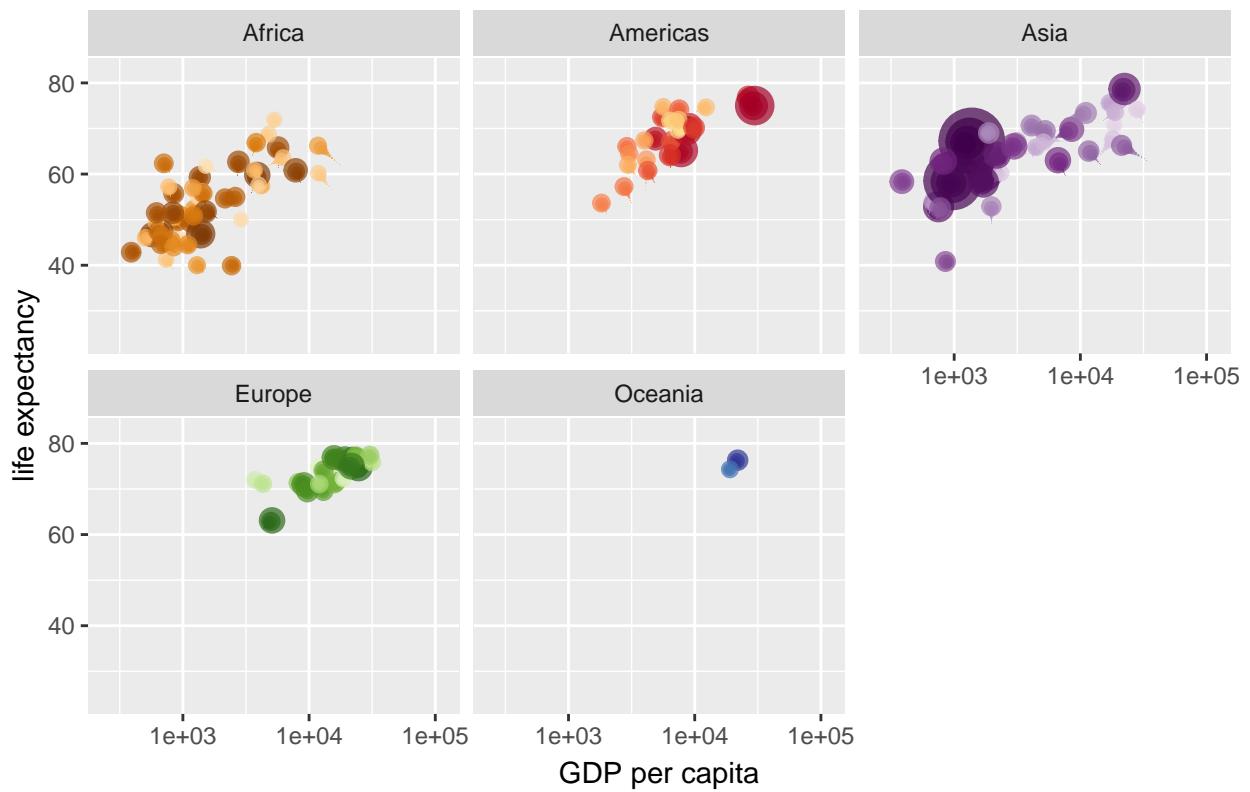
Year: 1986



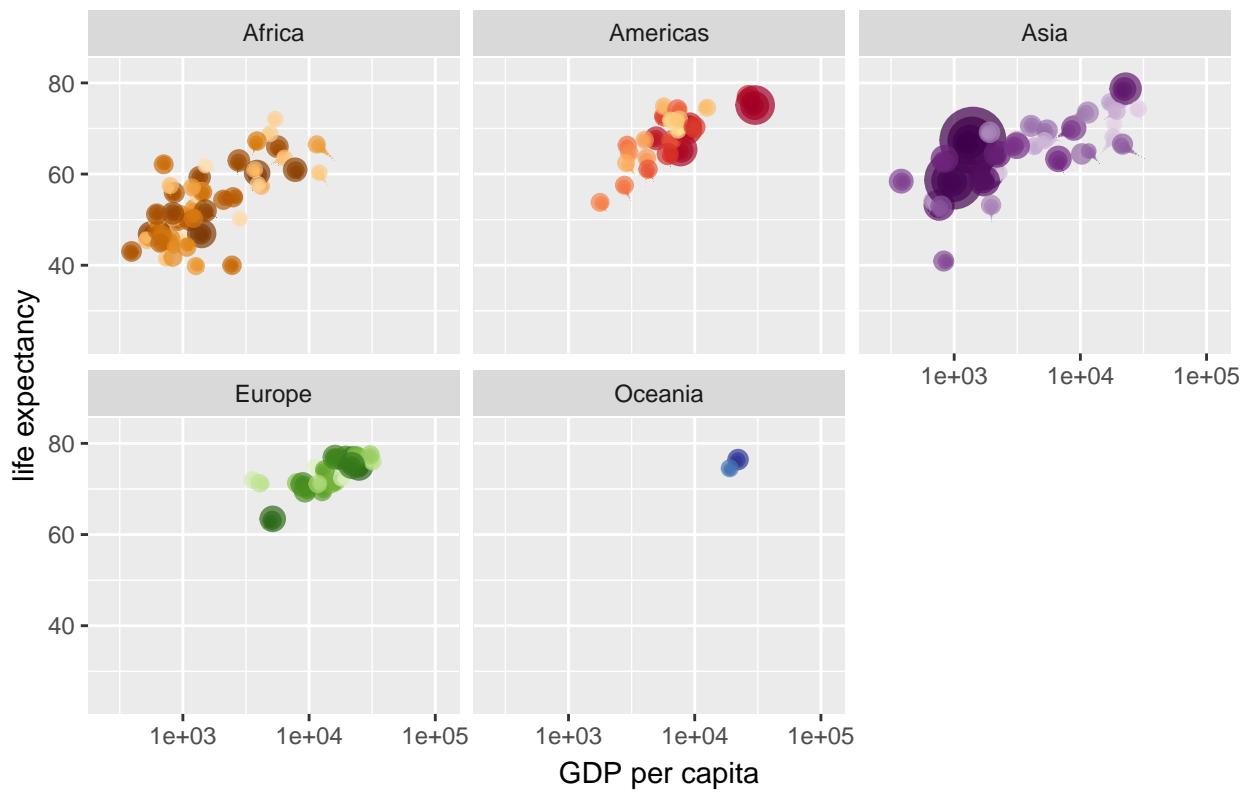
Year: 1986



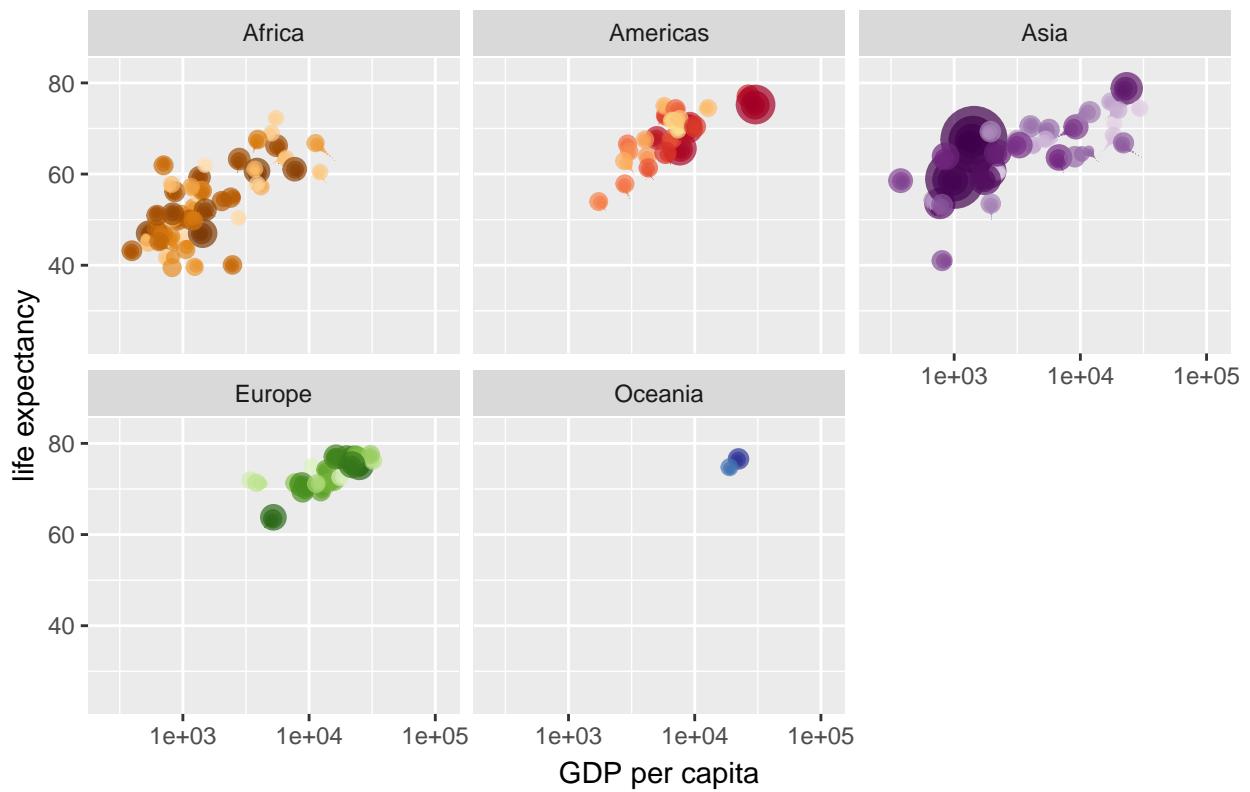
Year: 1987



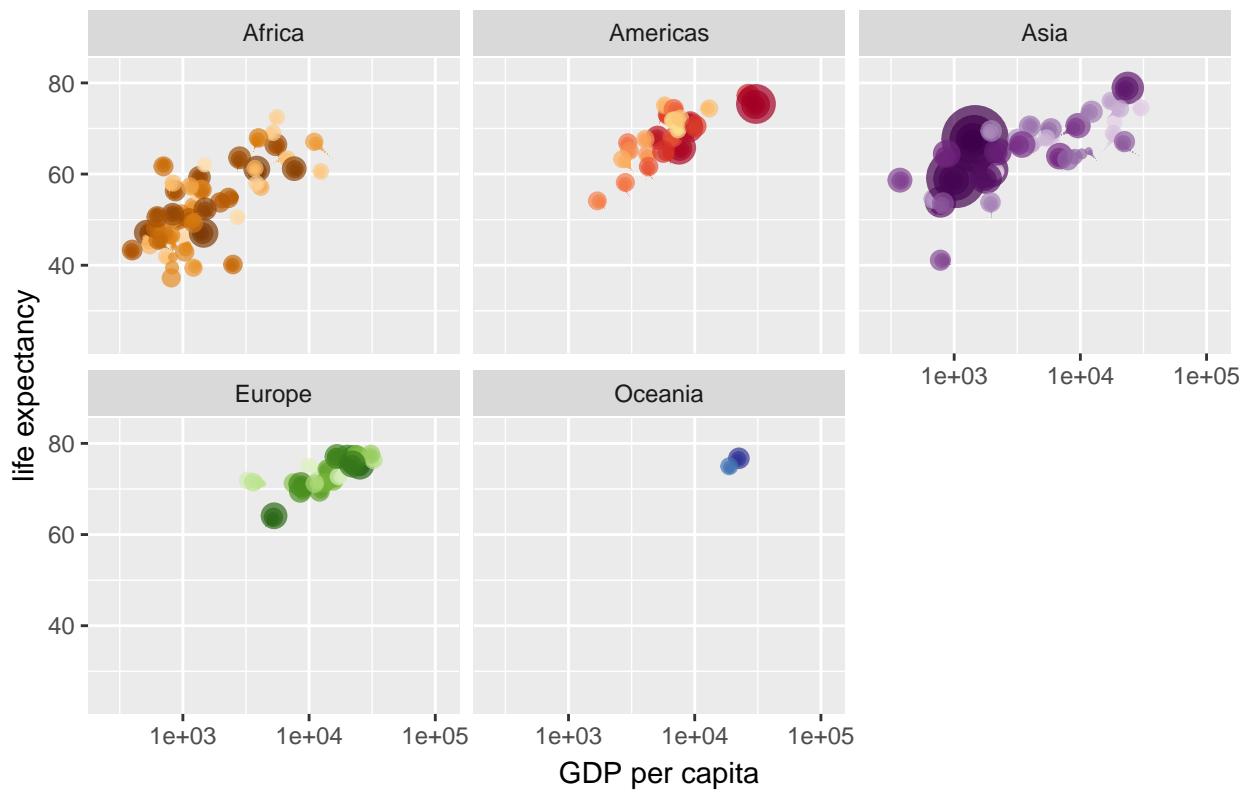
Year: 1988



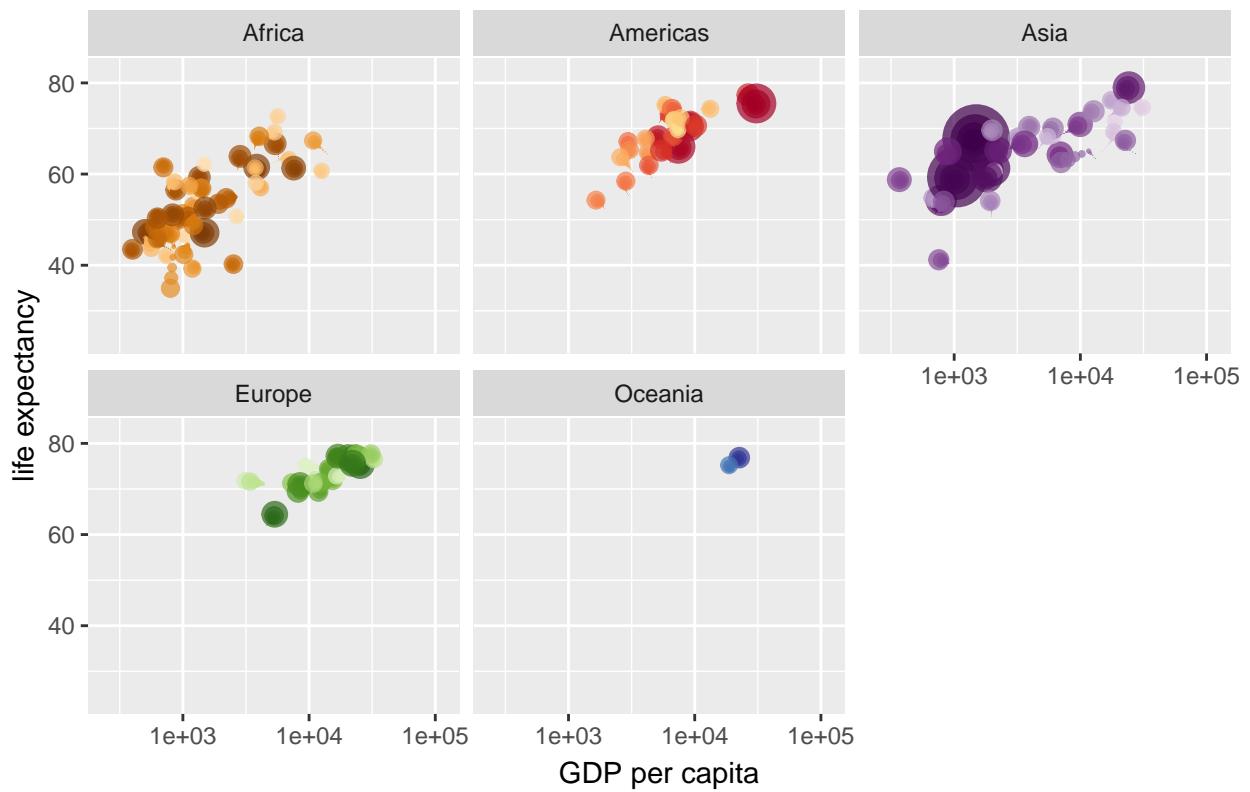
Year: 1988



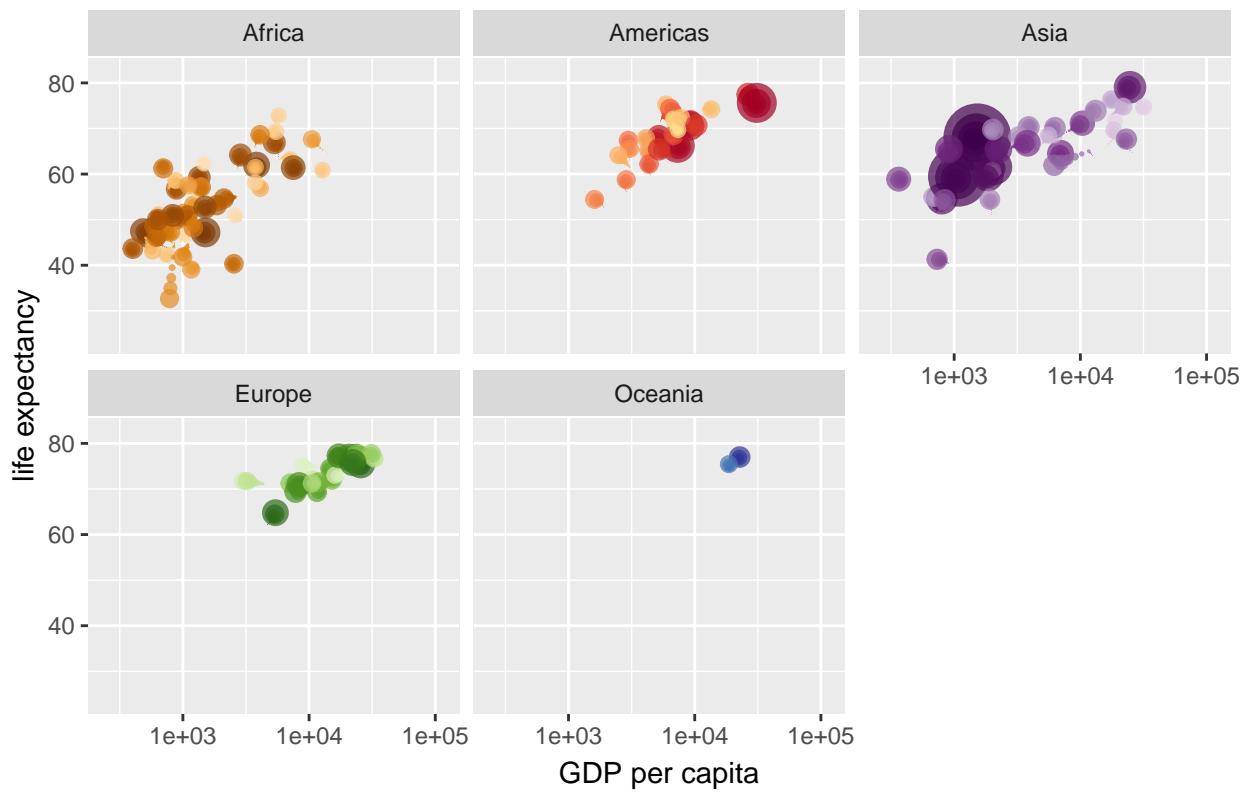
Year: 1989



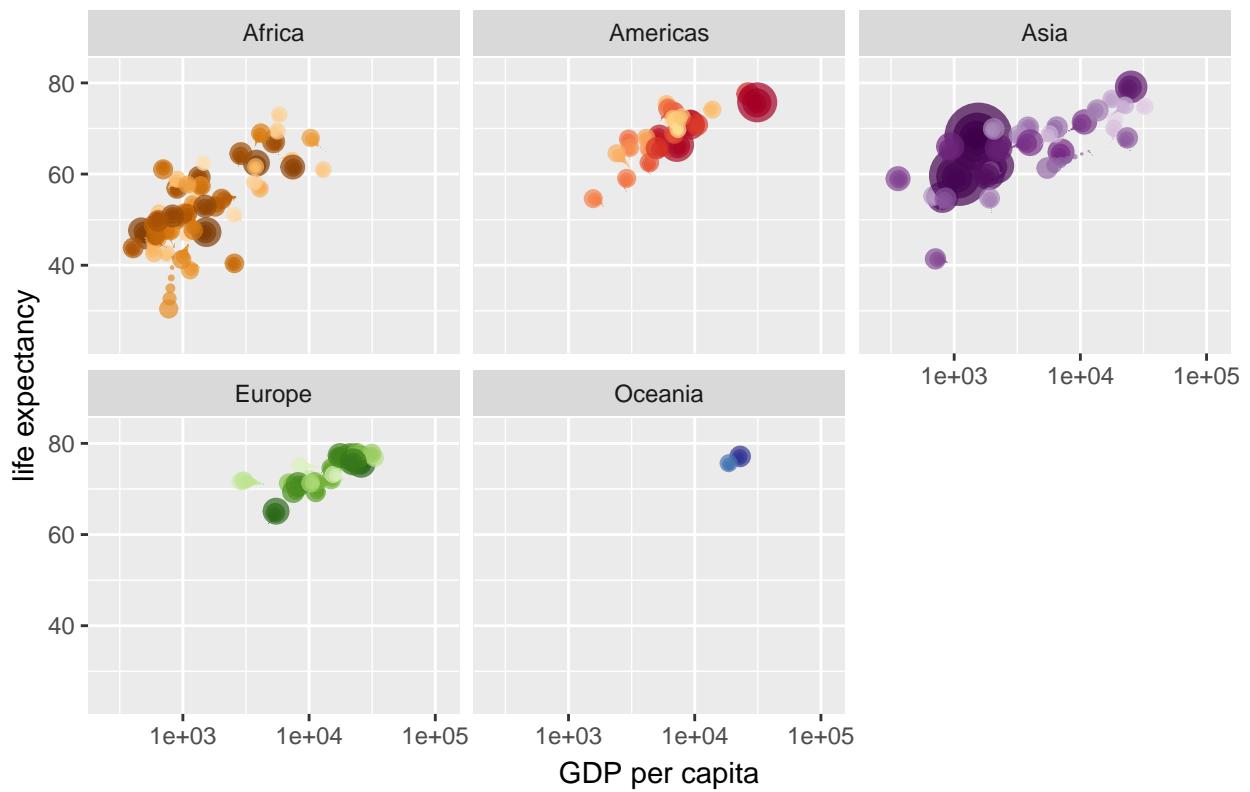
Year: 1989



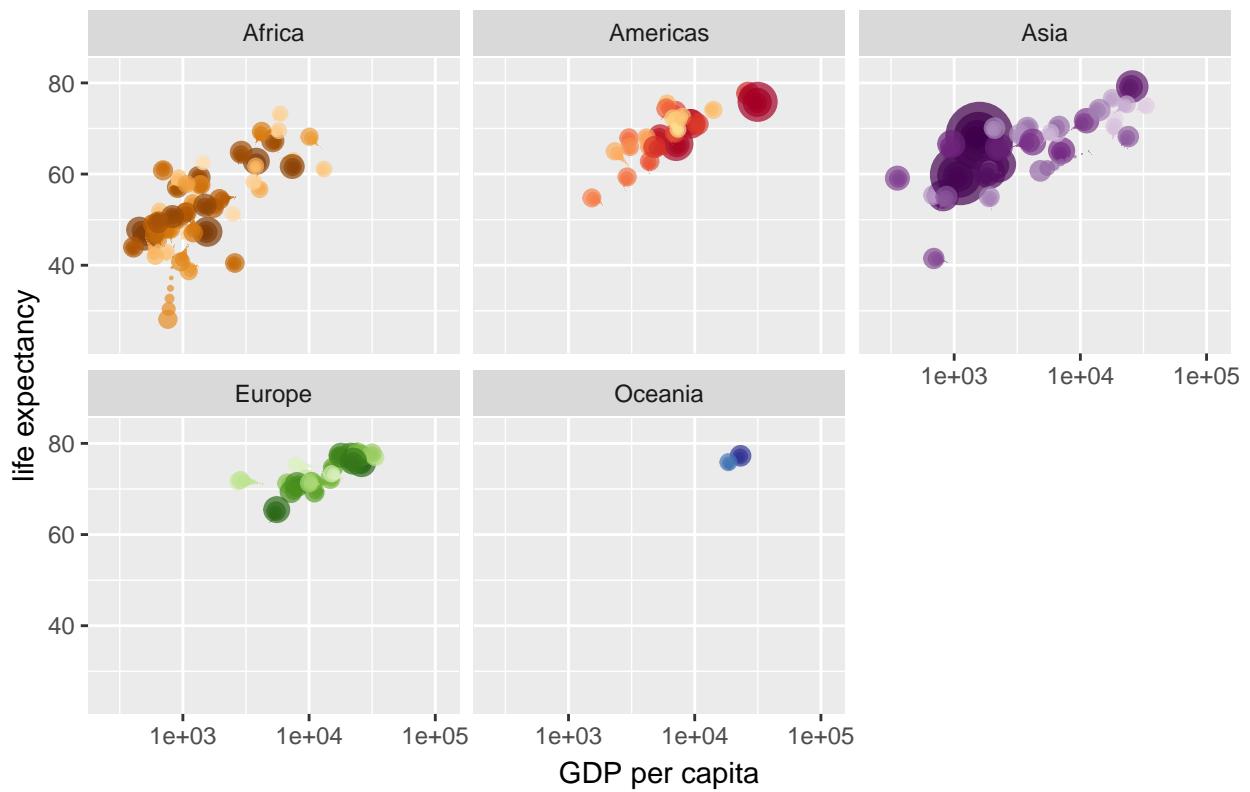
Year: 1990



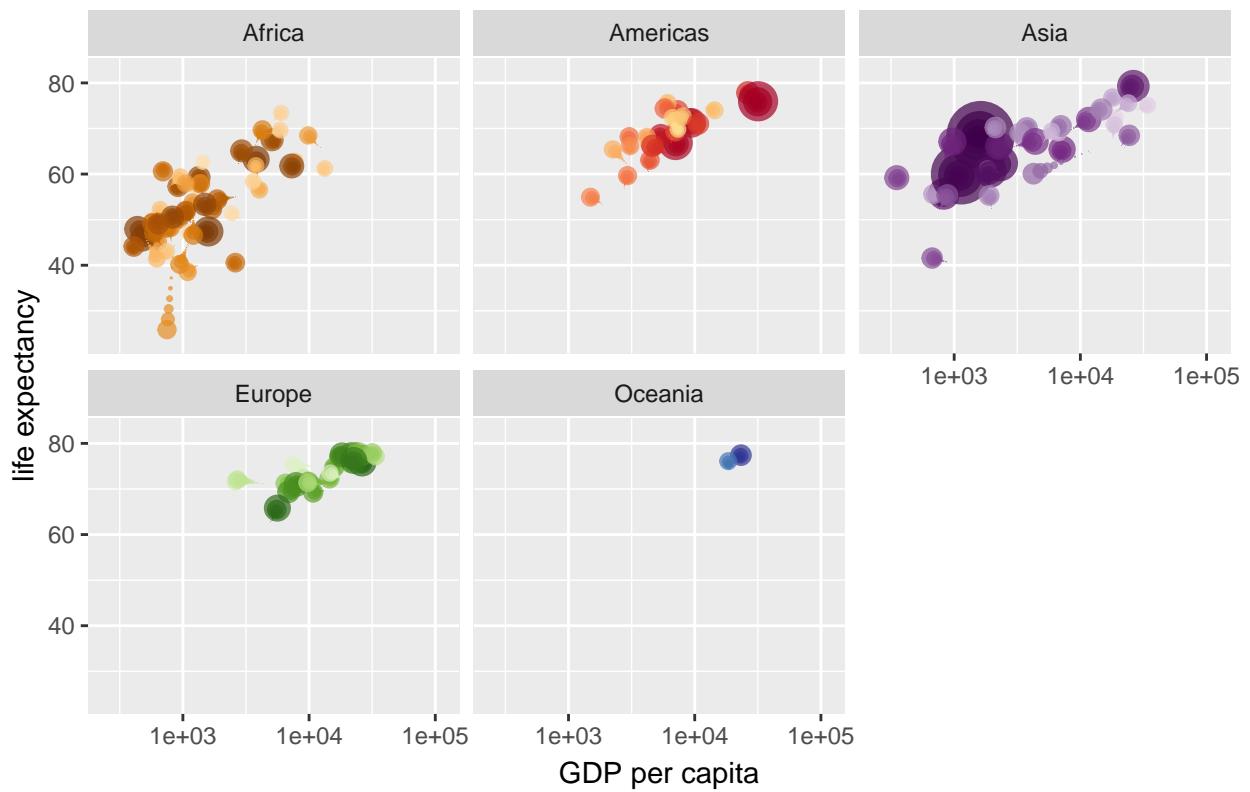
Year: 1990



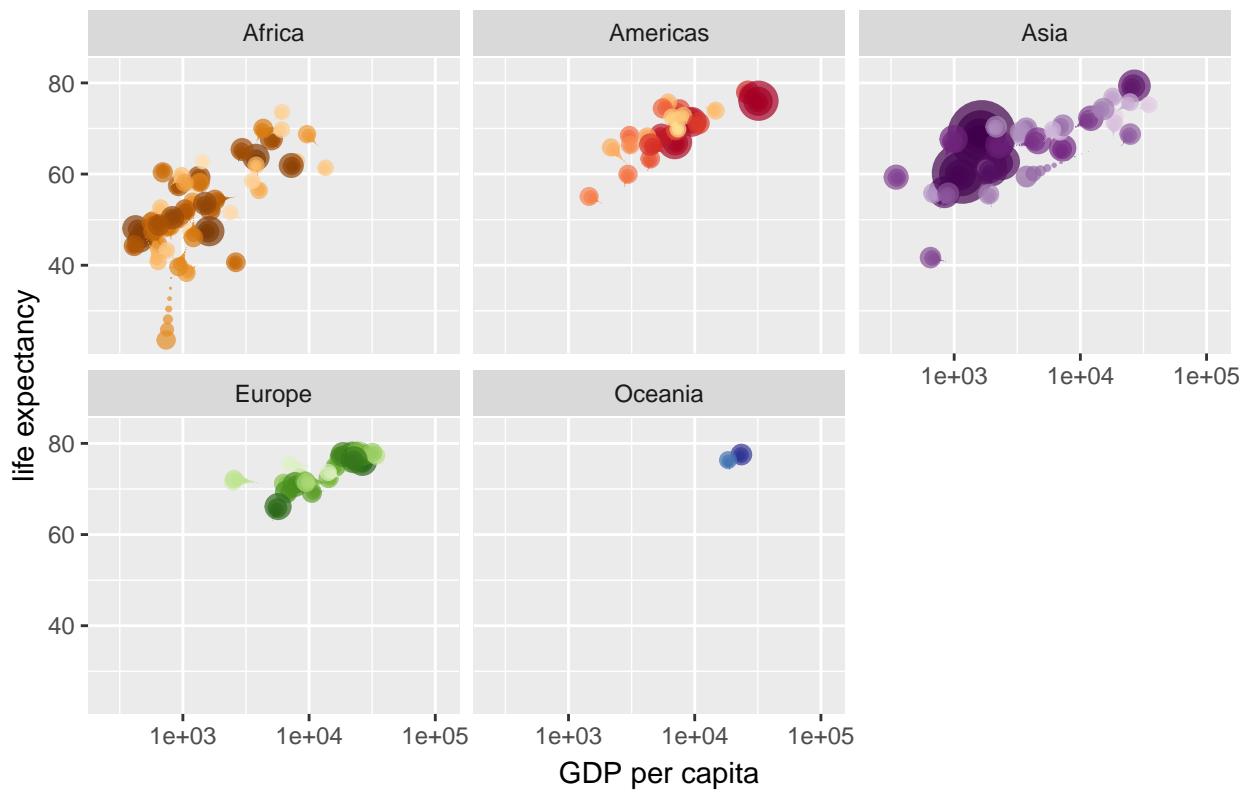
Year: 1991



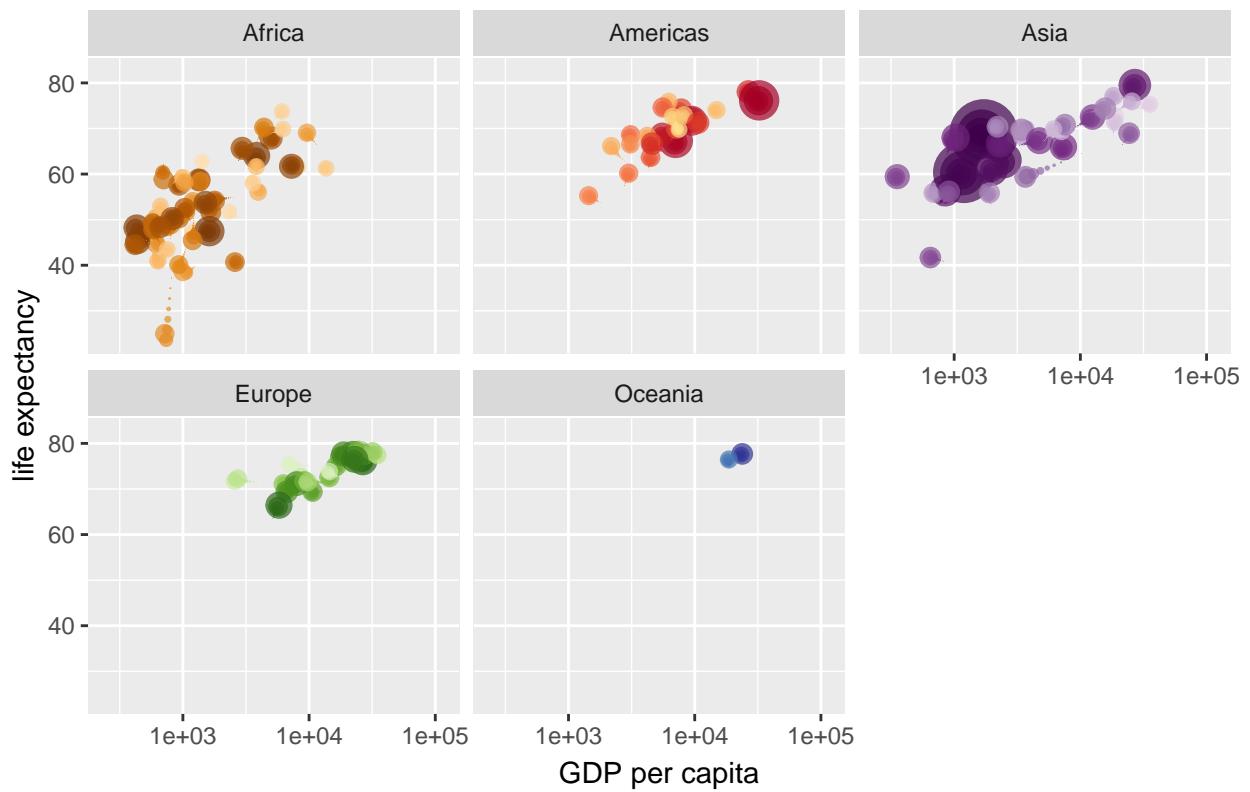
Year: 1991



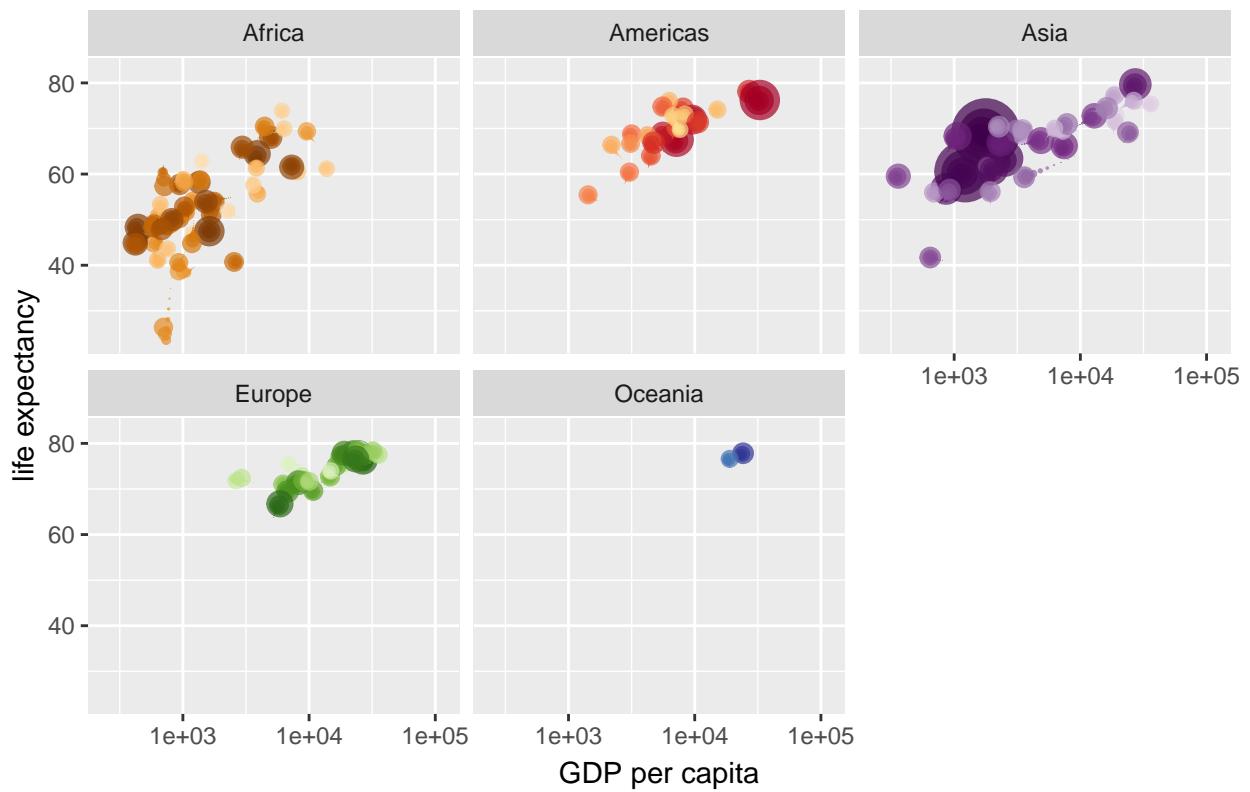
Year: 1992



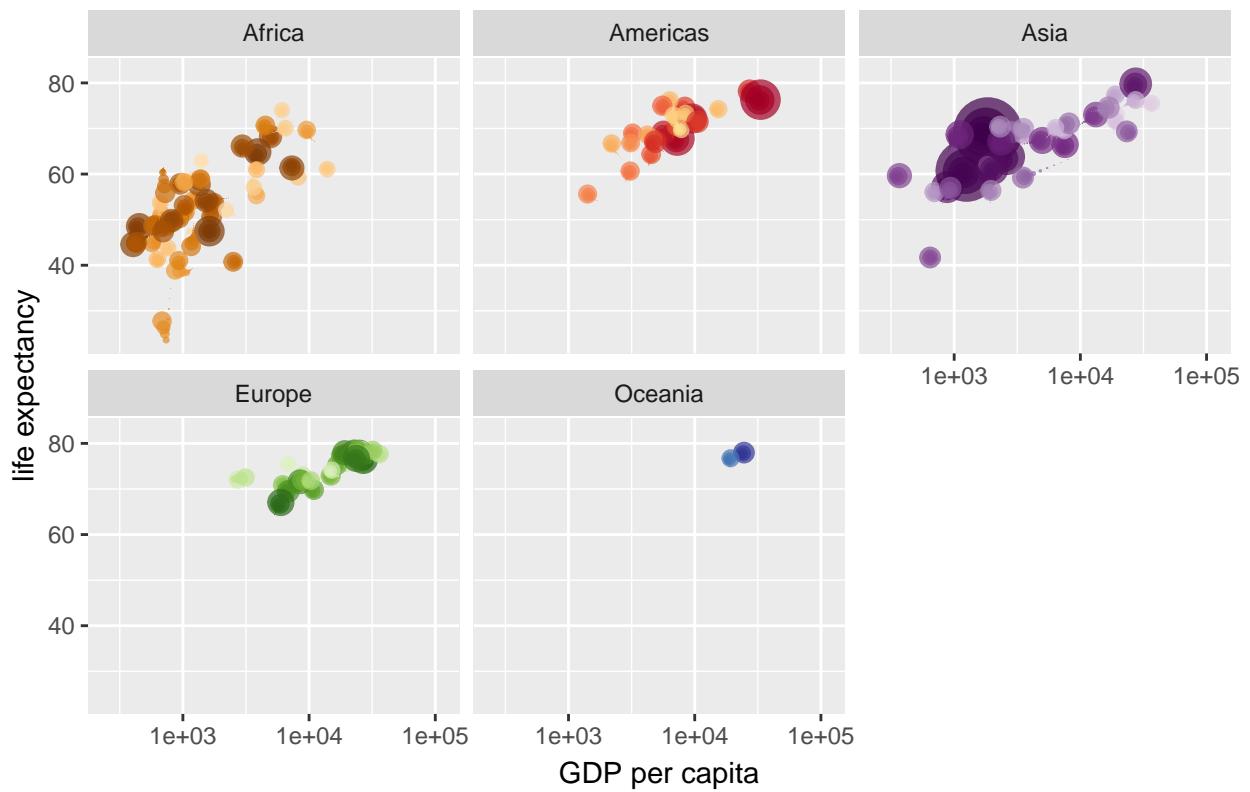
Year: 1993



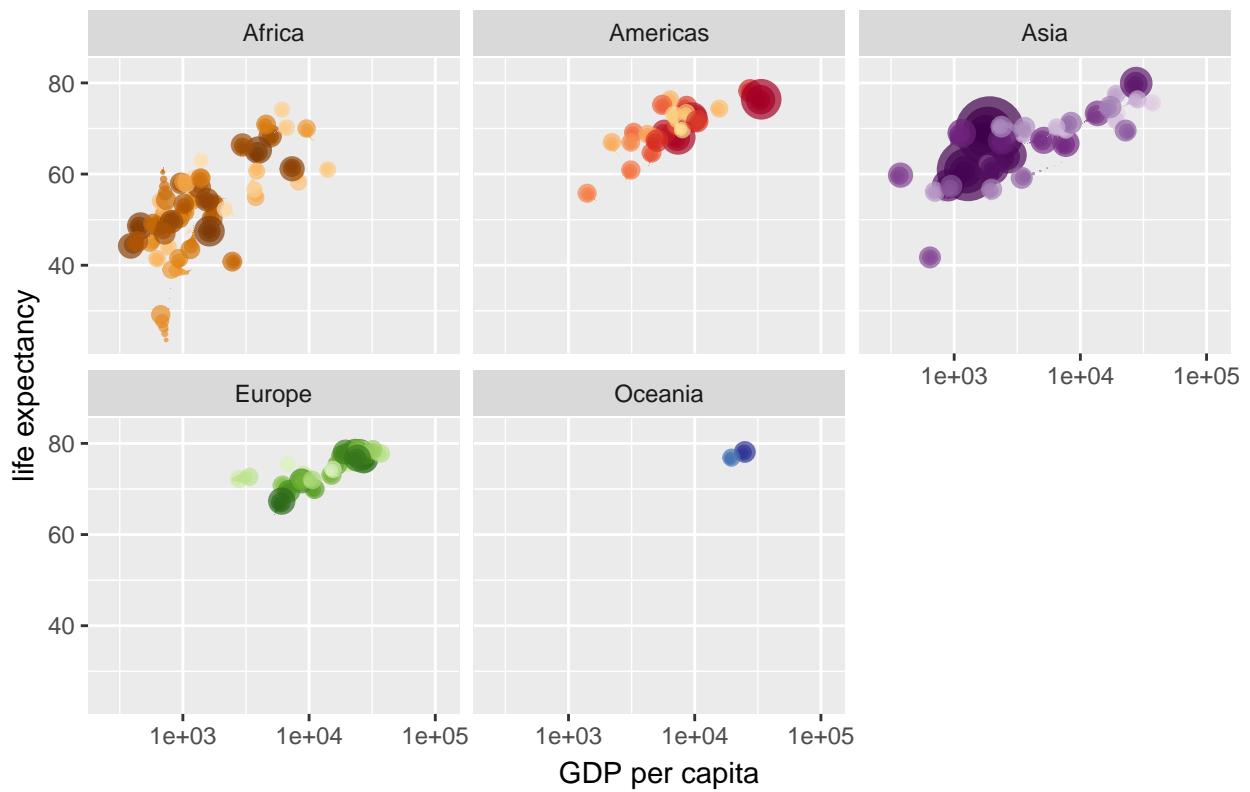
Year: 1993



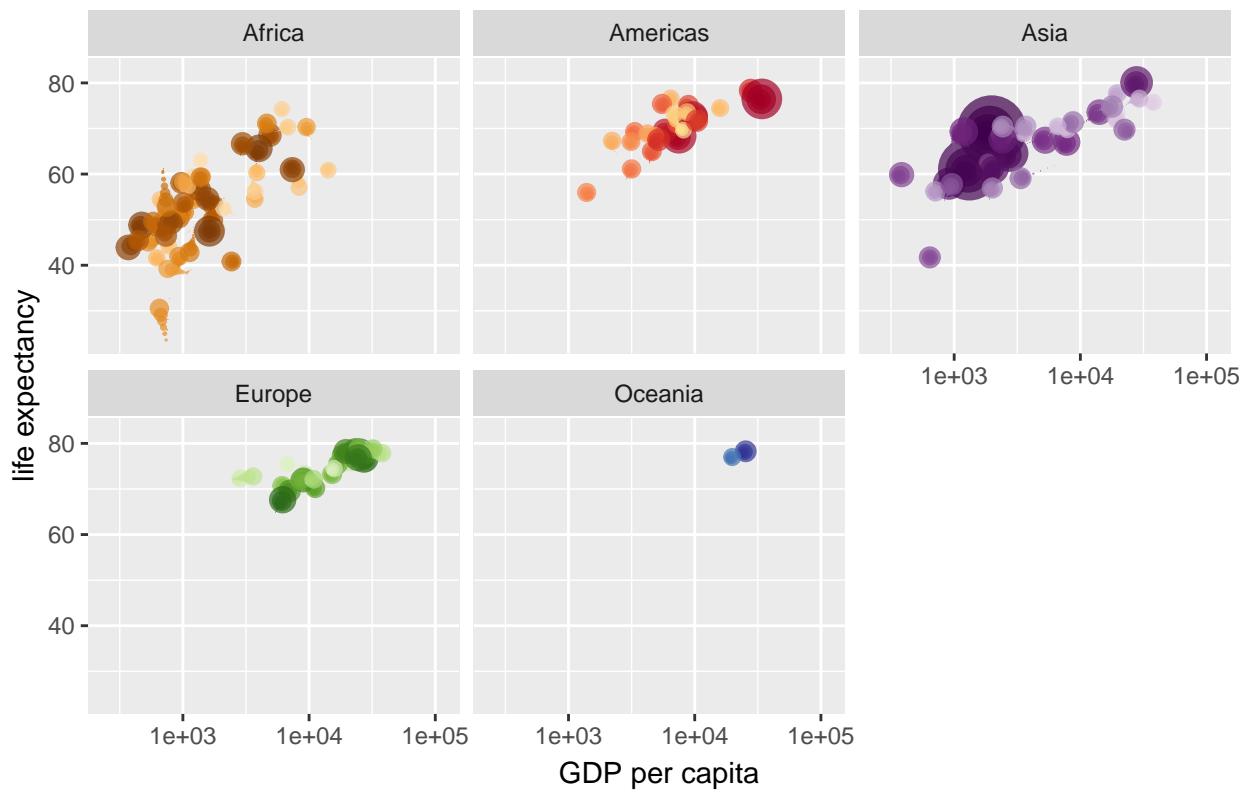
Year: 1994



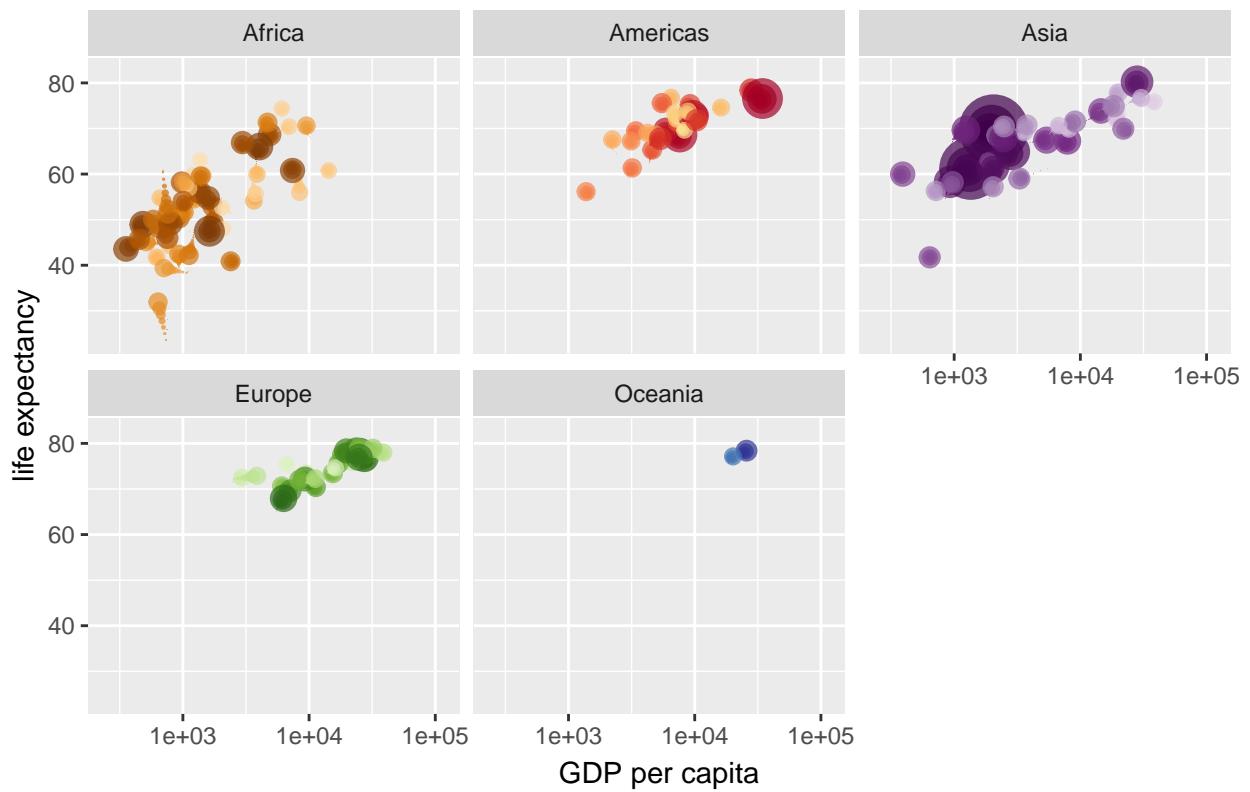
Year: 1994



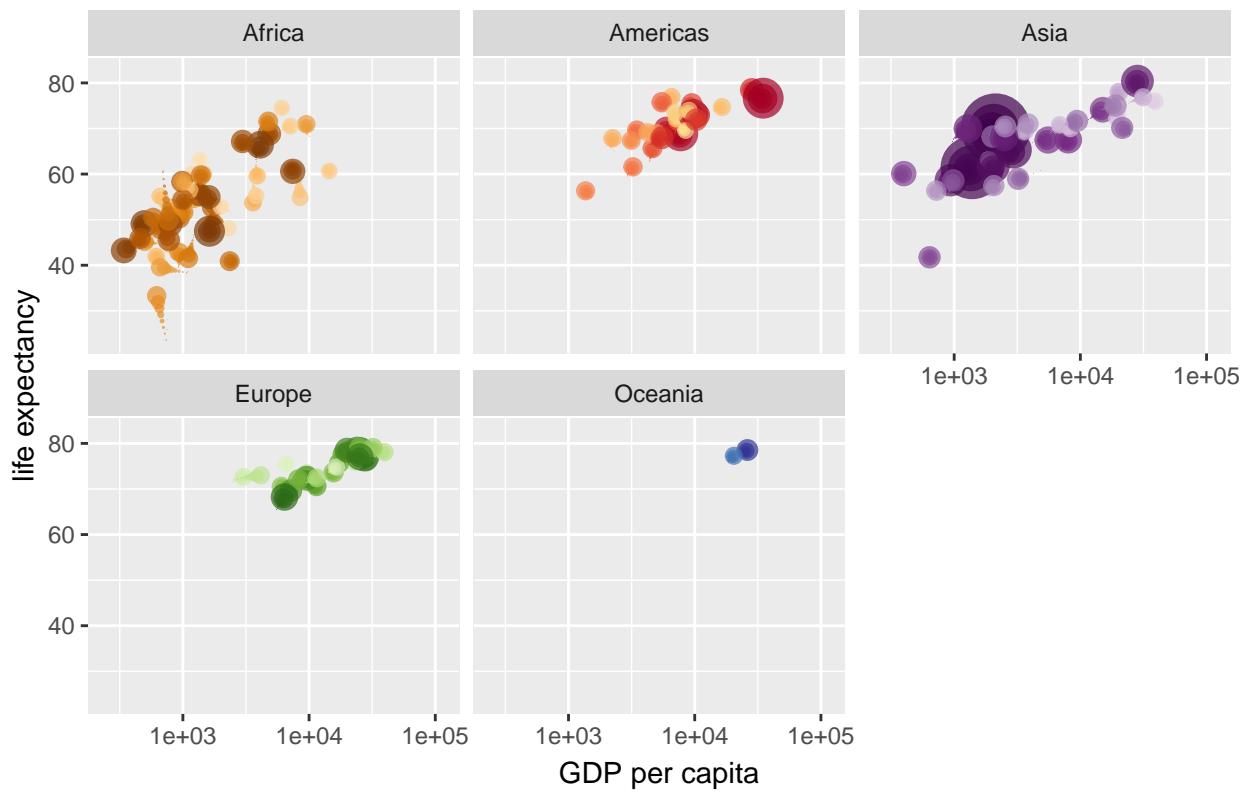
Year: 1995



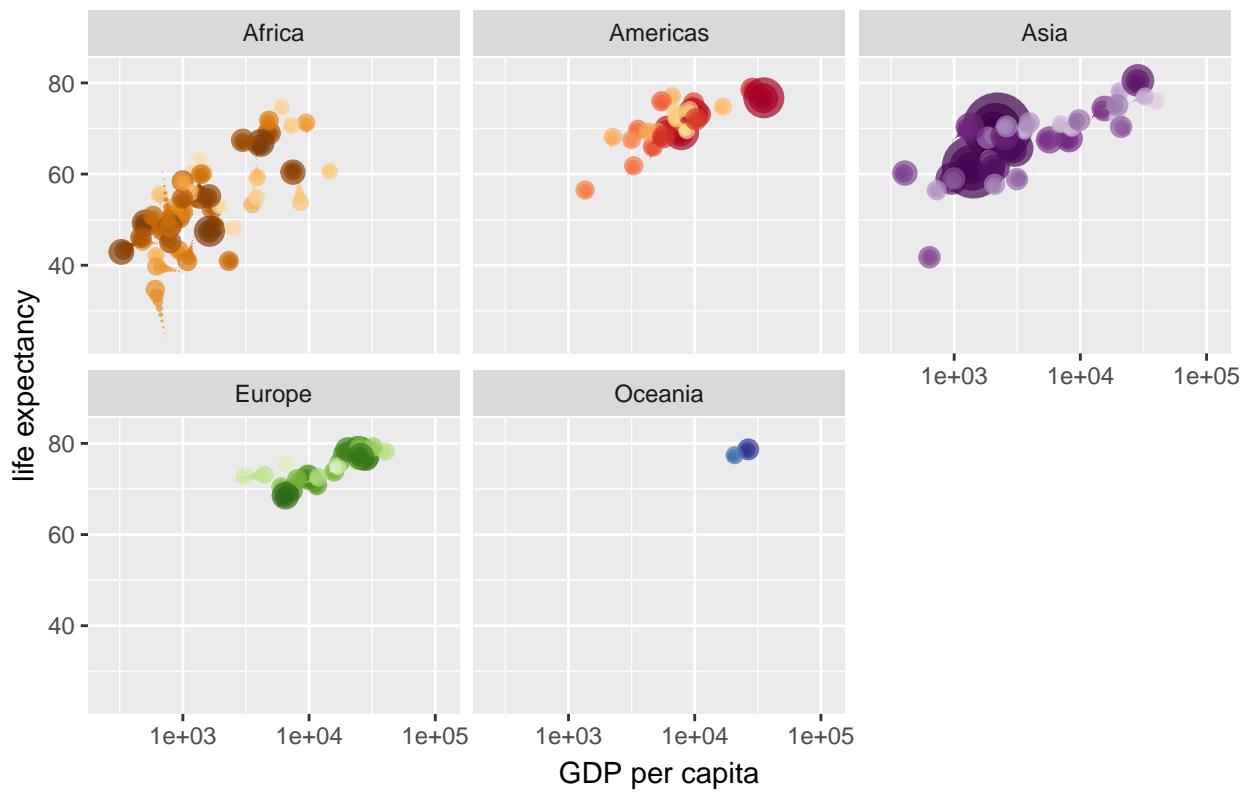
Year: 1995



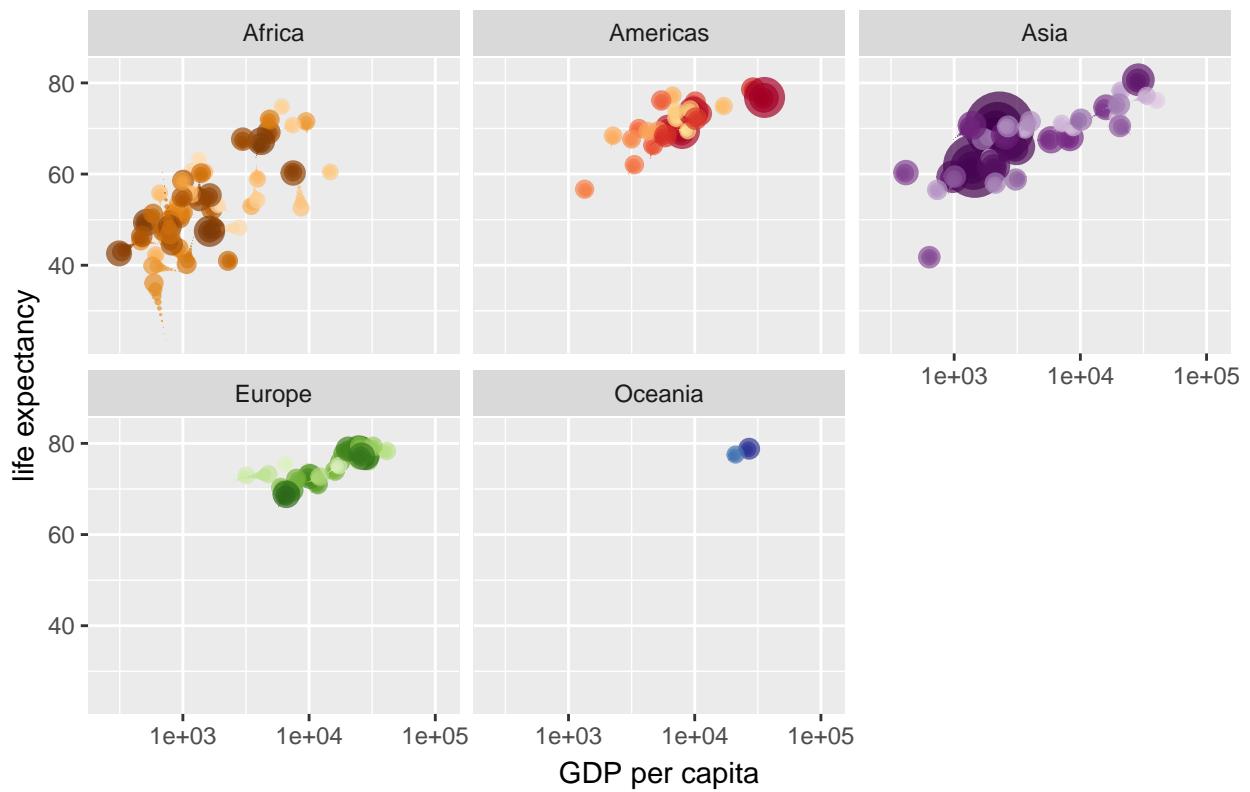
Year: 1996



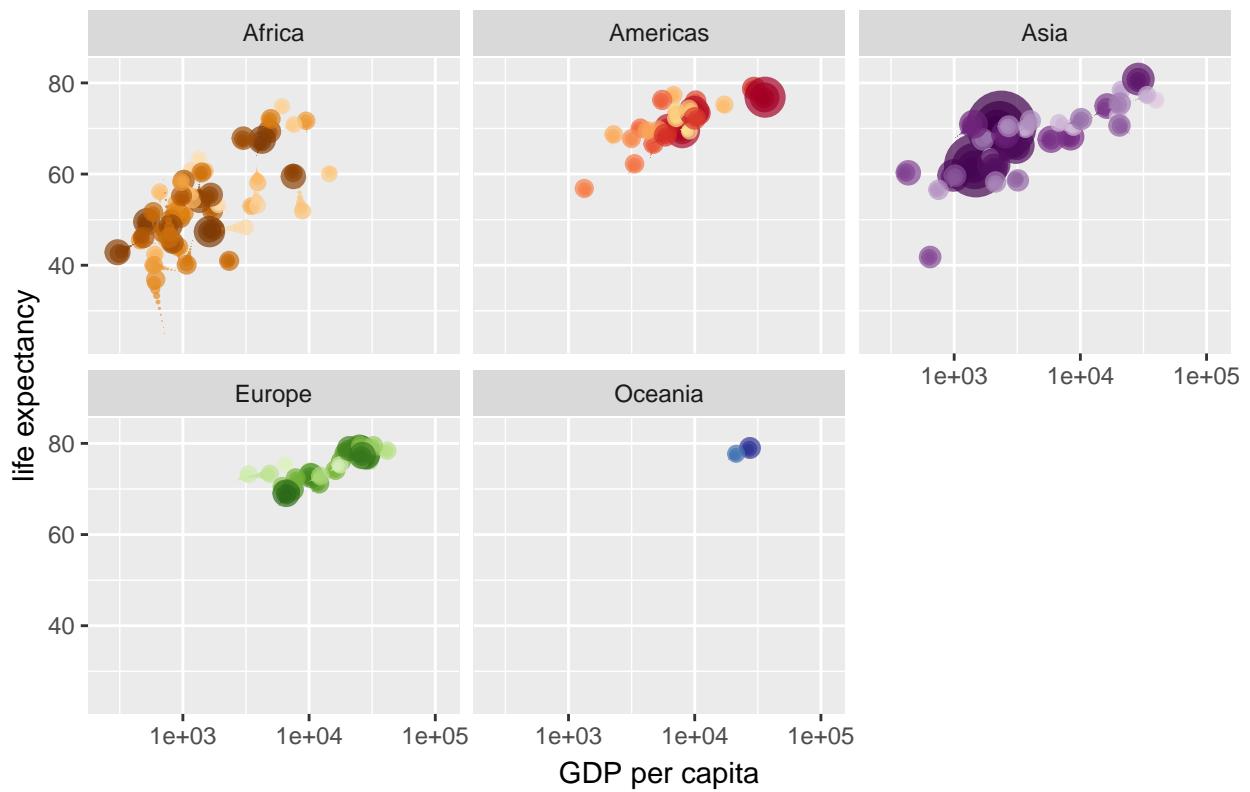
Year: 1996



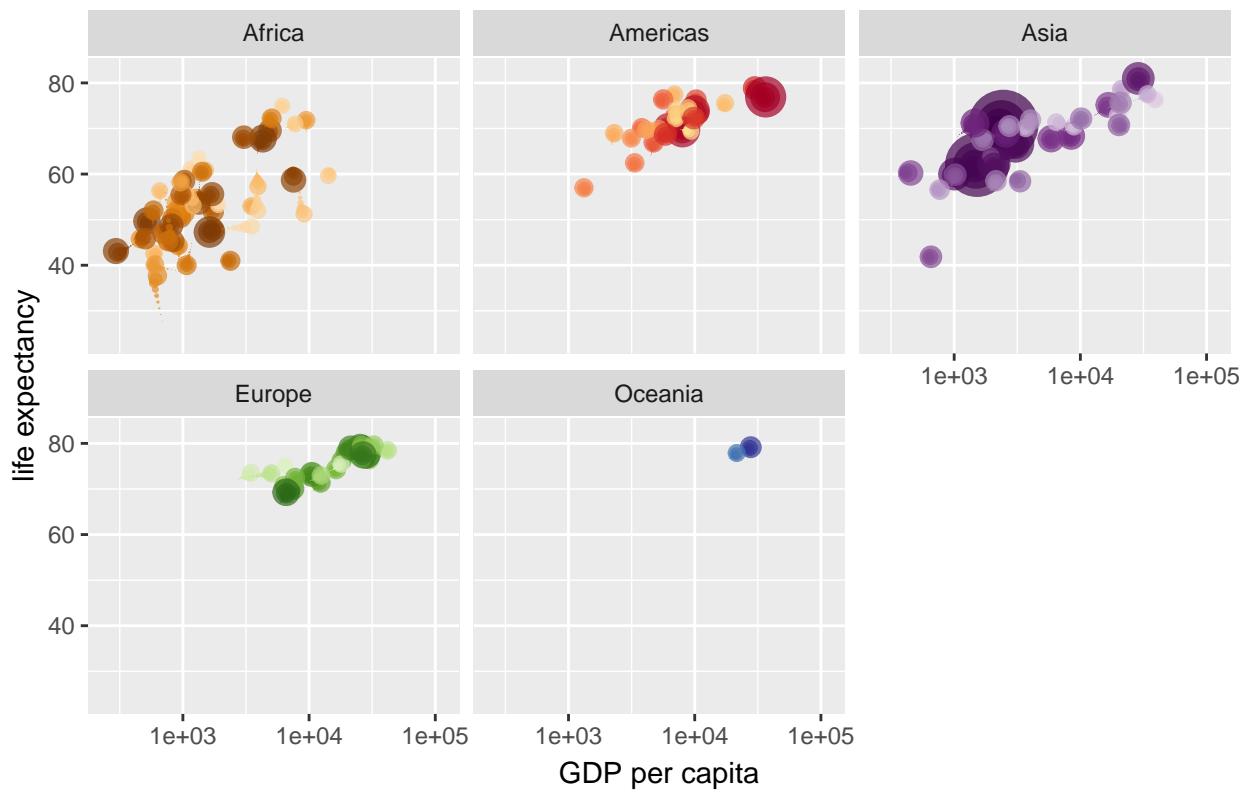
Year: 1997



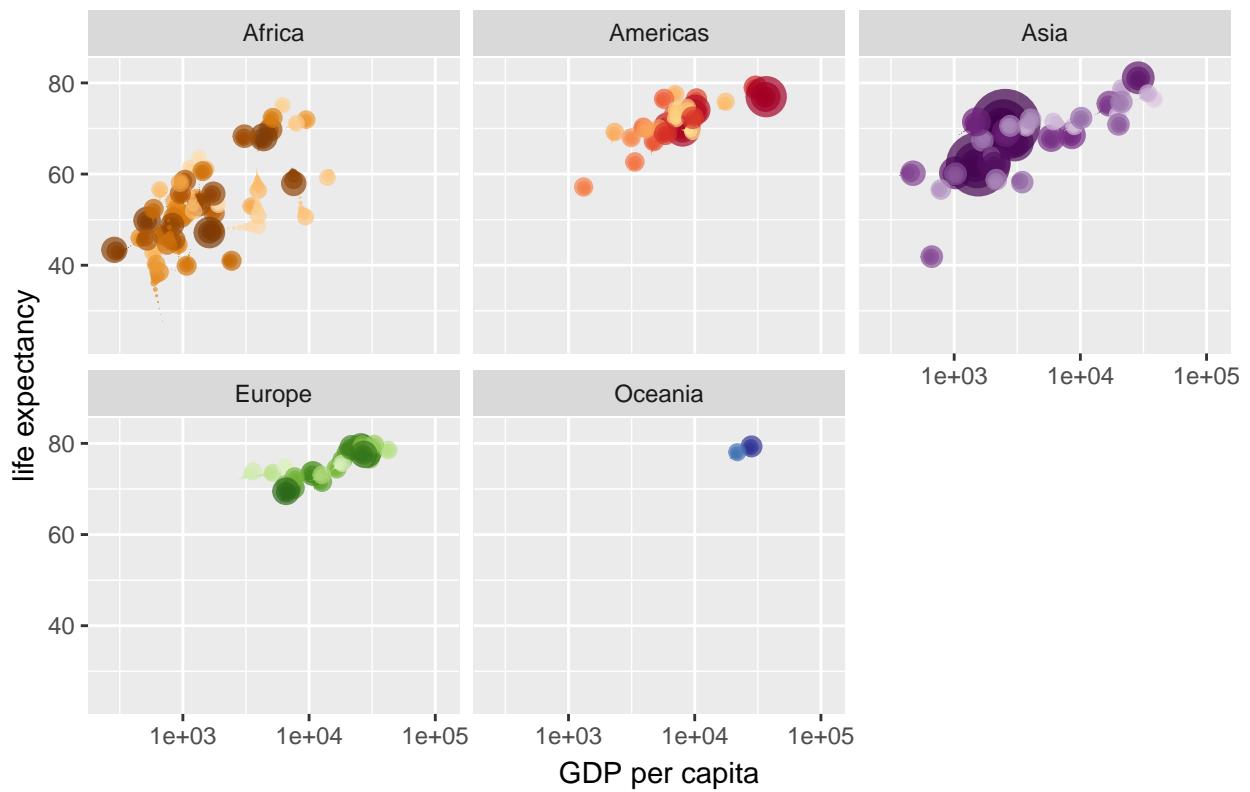
Year: 1998



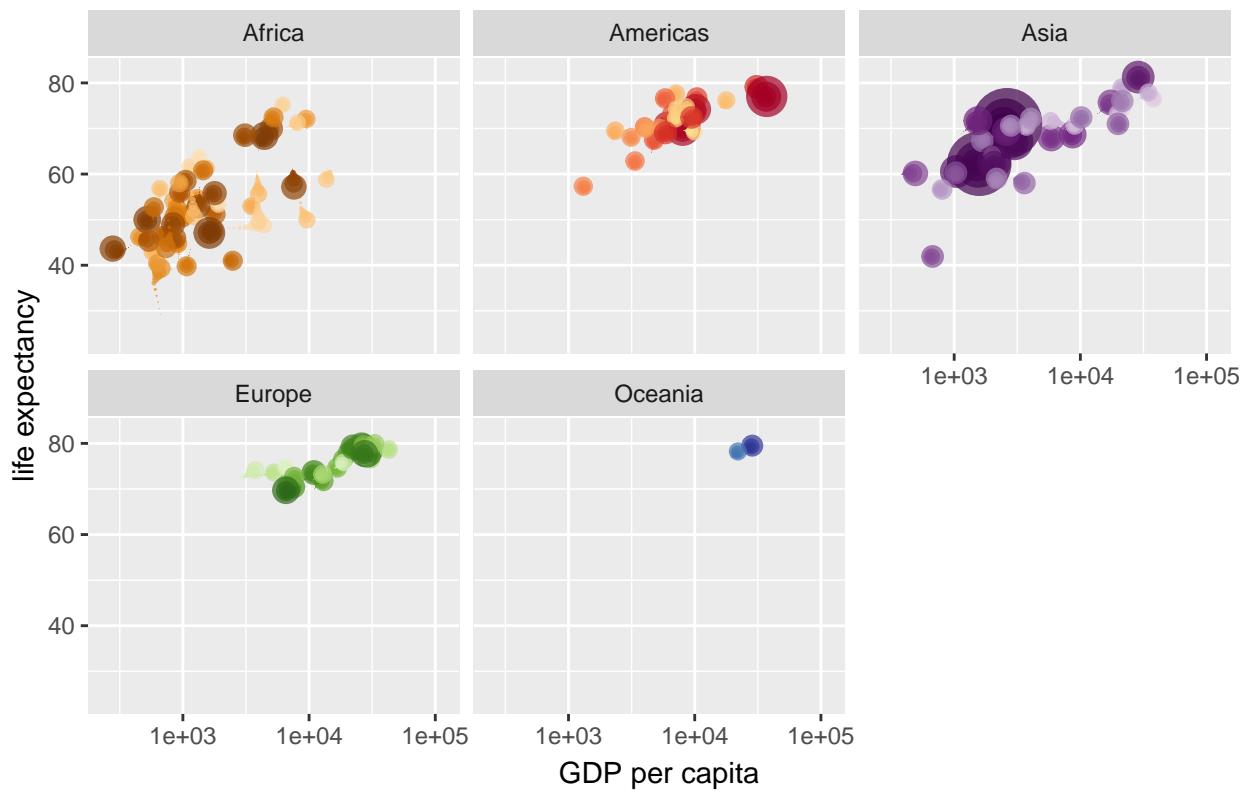
Year: 1998



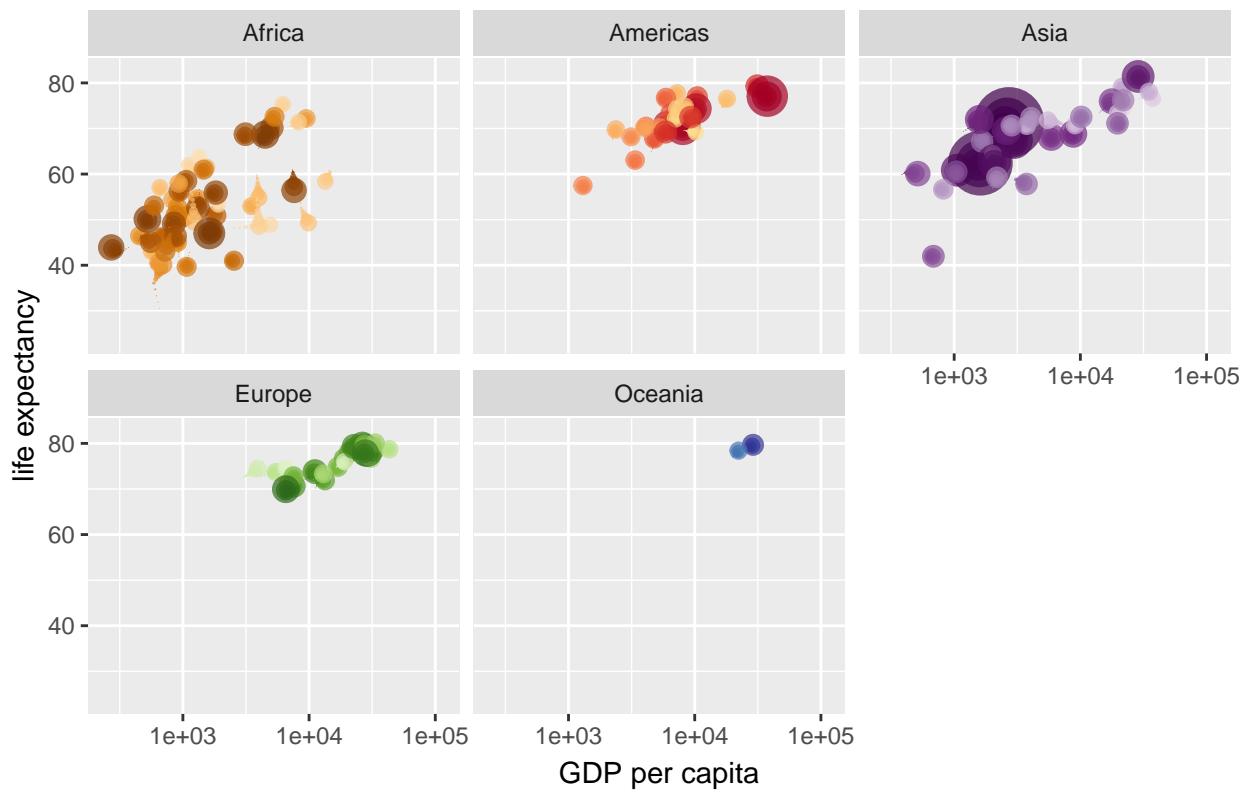
Year: 1999



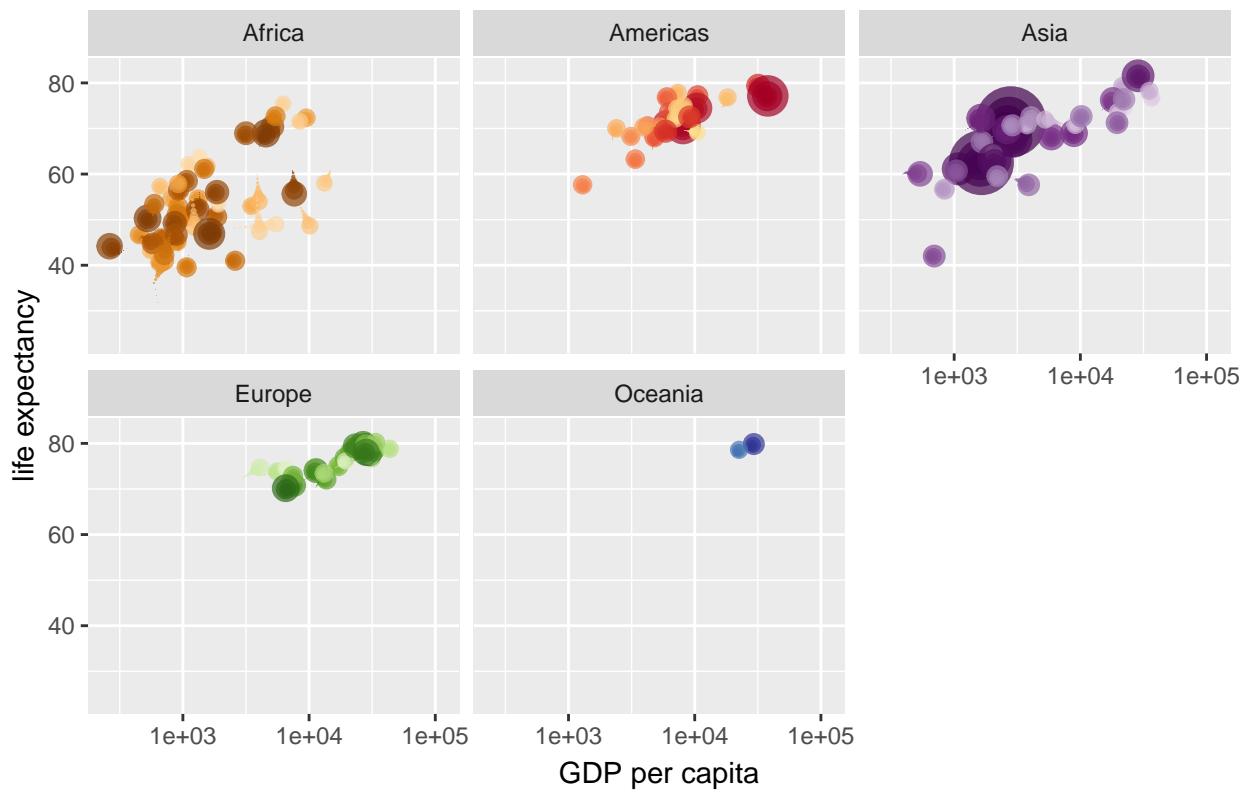
Year: 1999



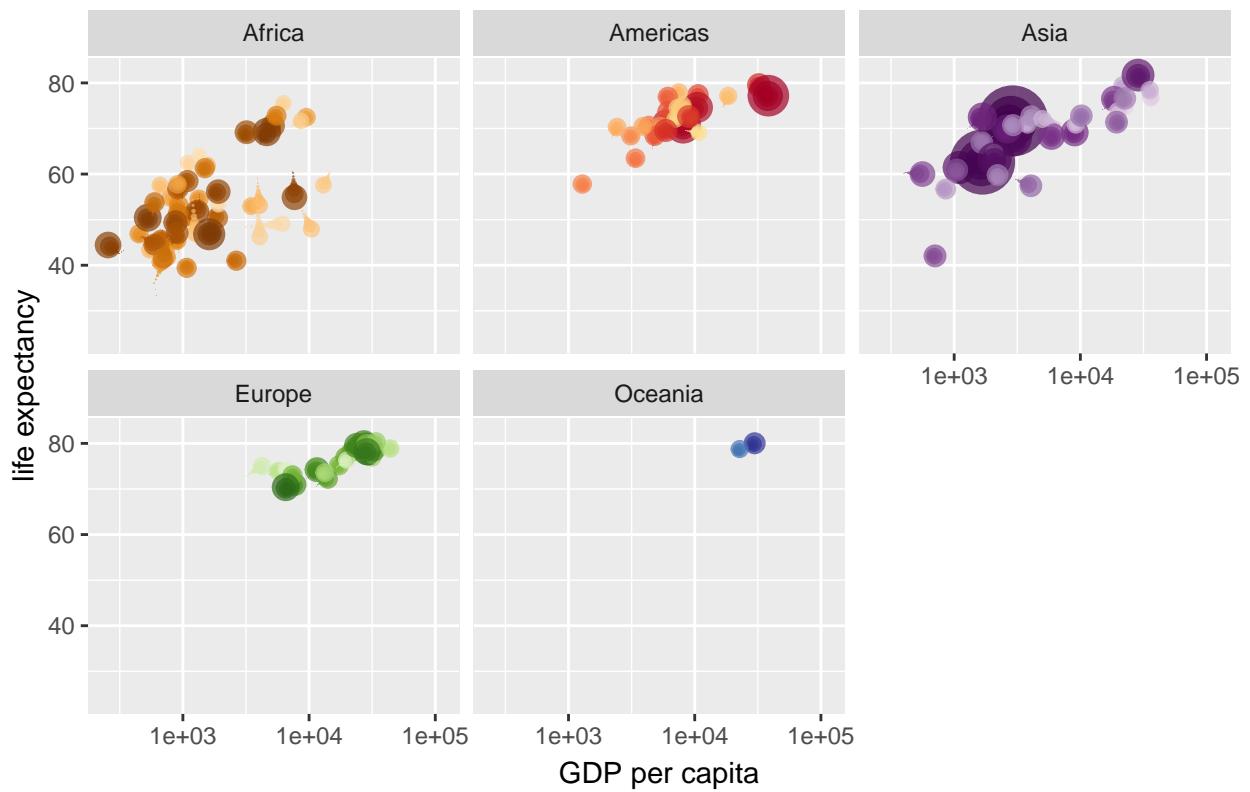
Year: 2000



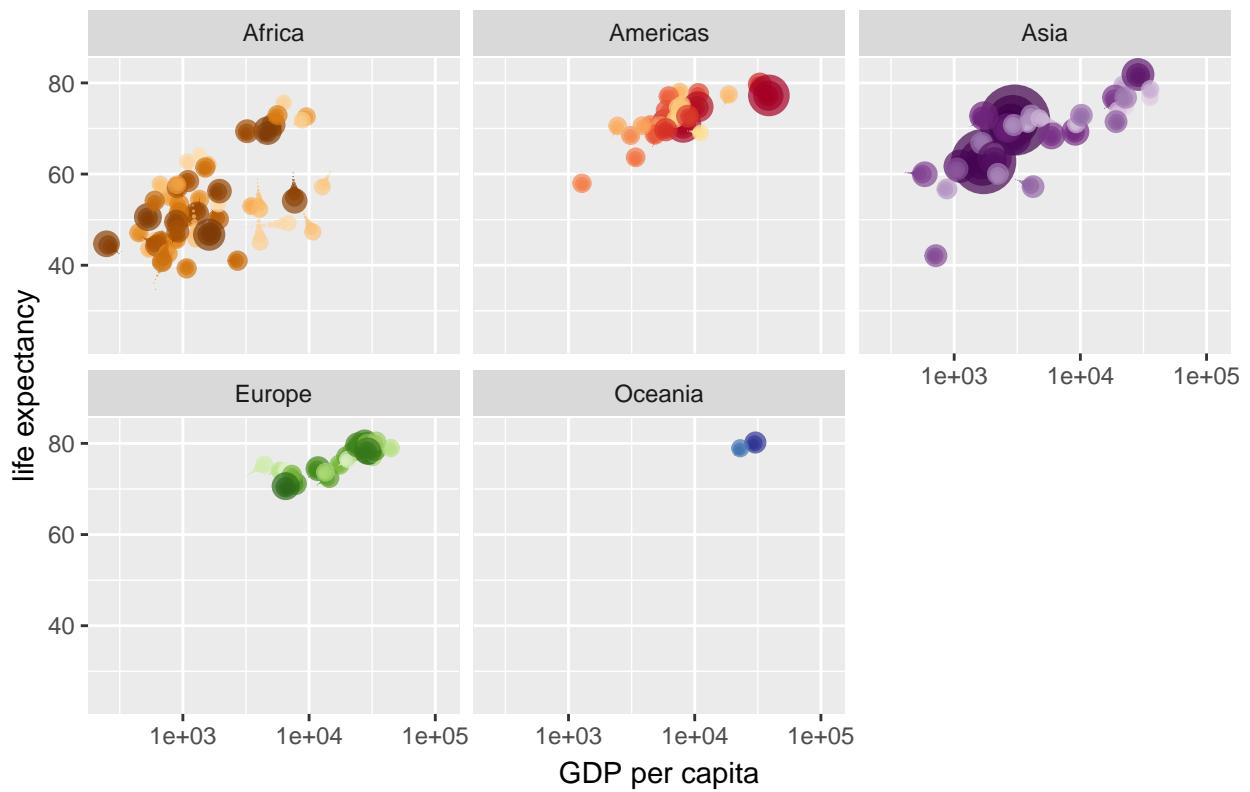
Year: 2000



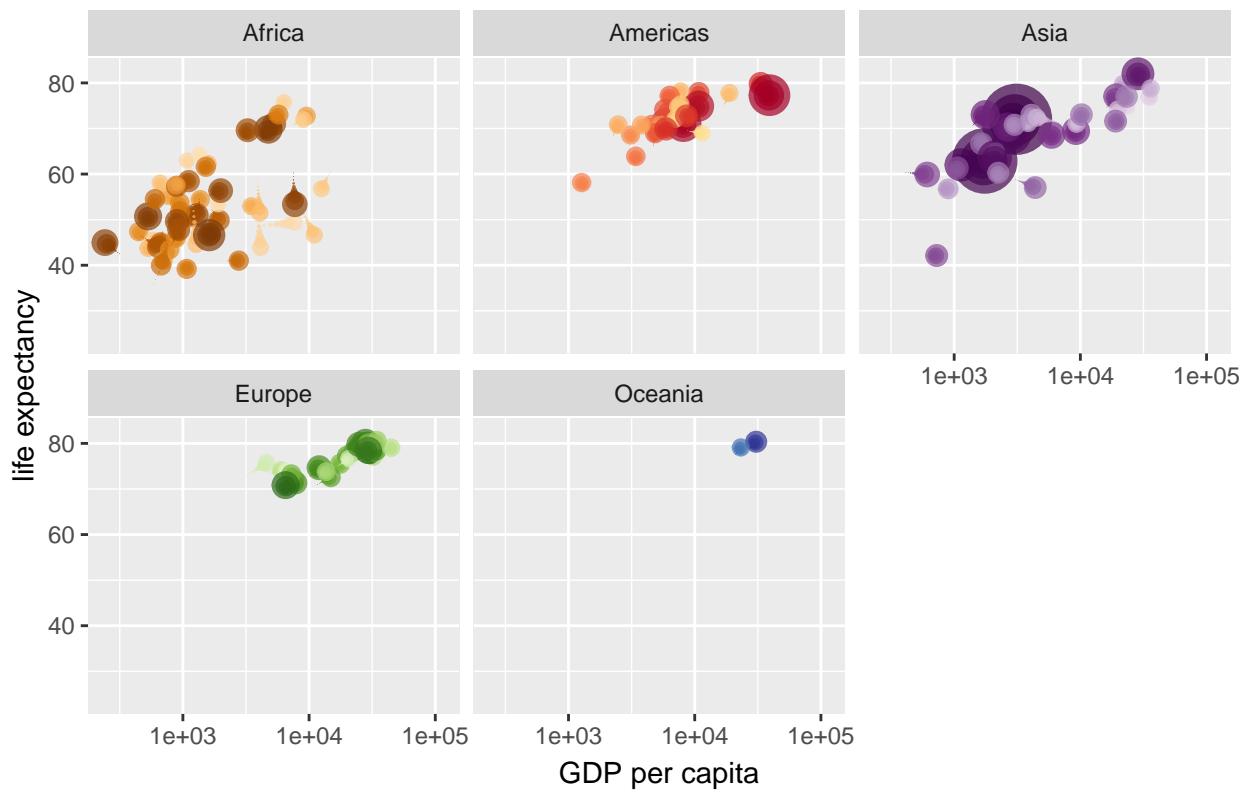
Year: 2001



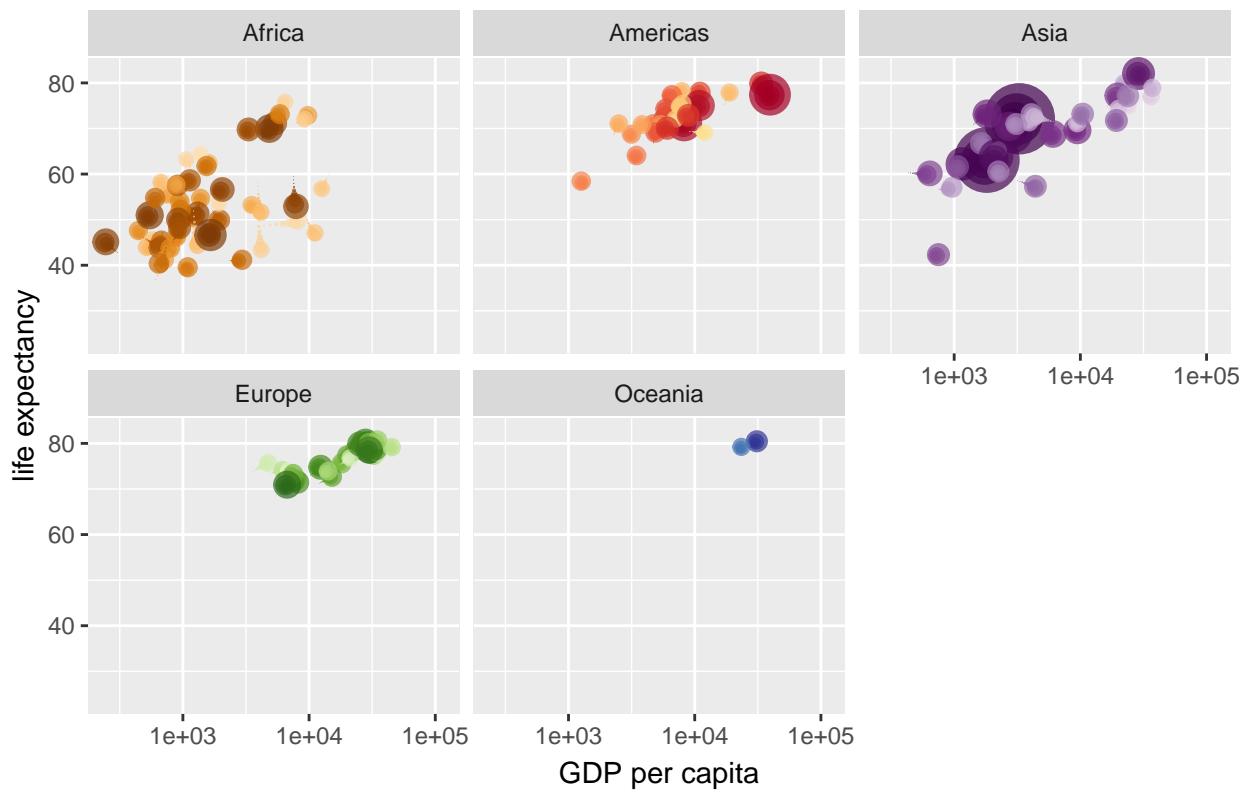
Year: 2001



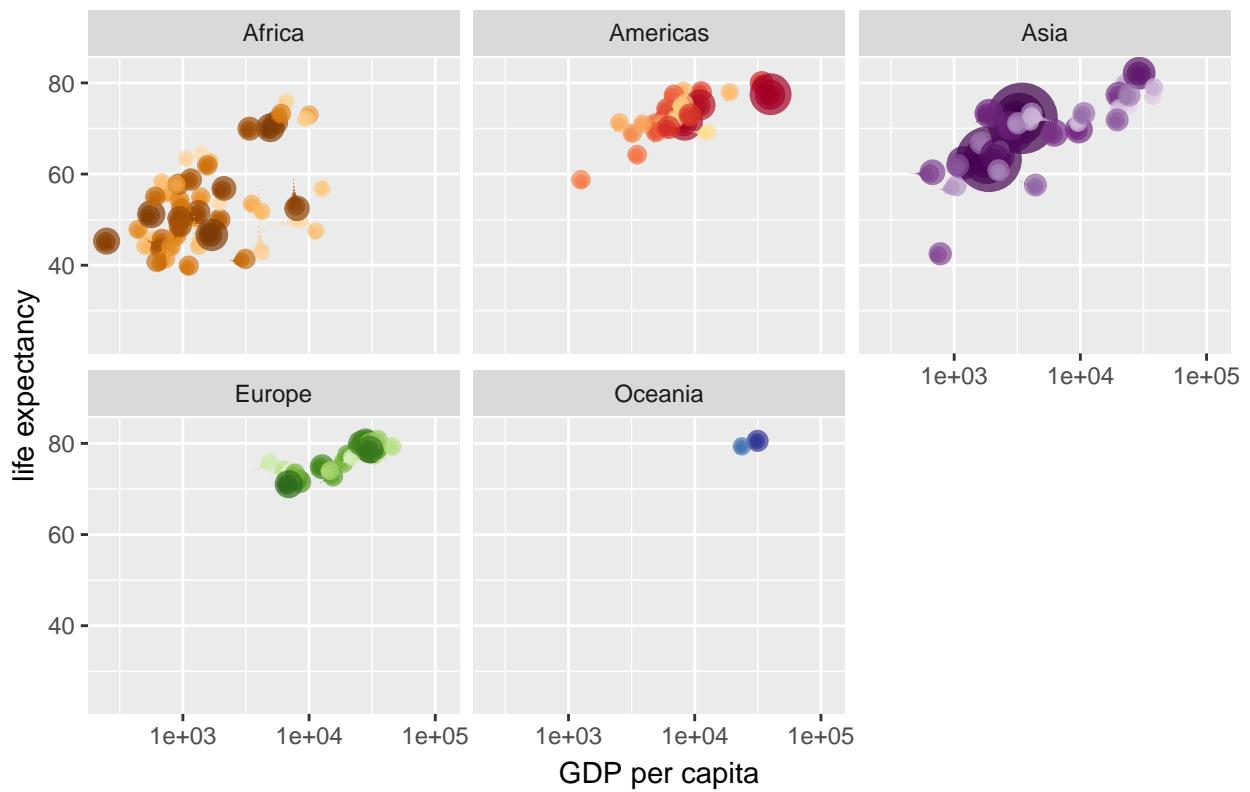
Year: 2002



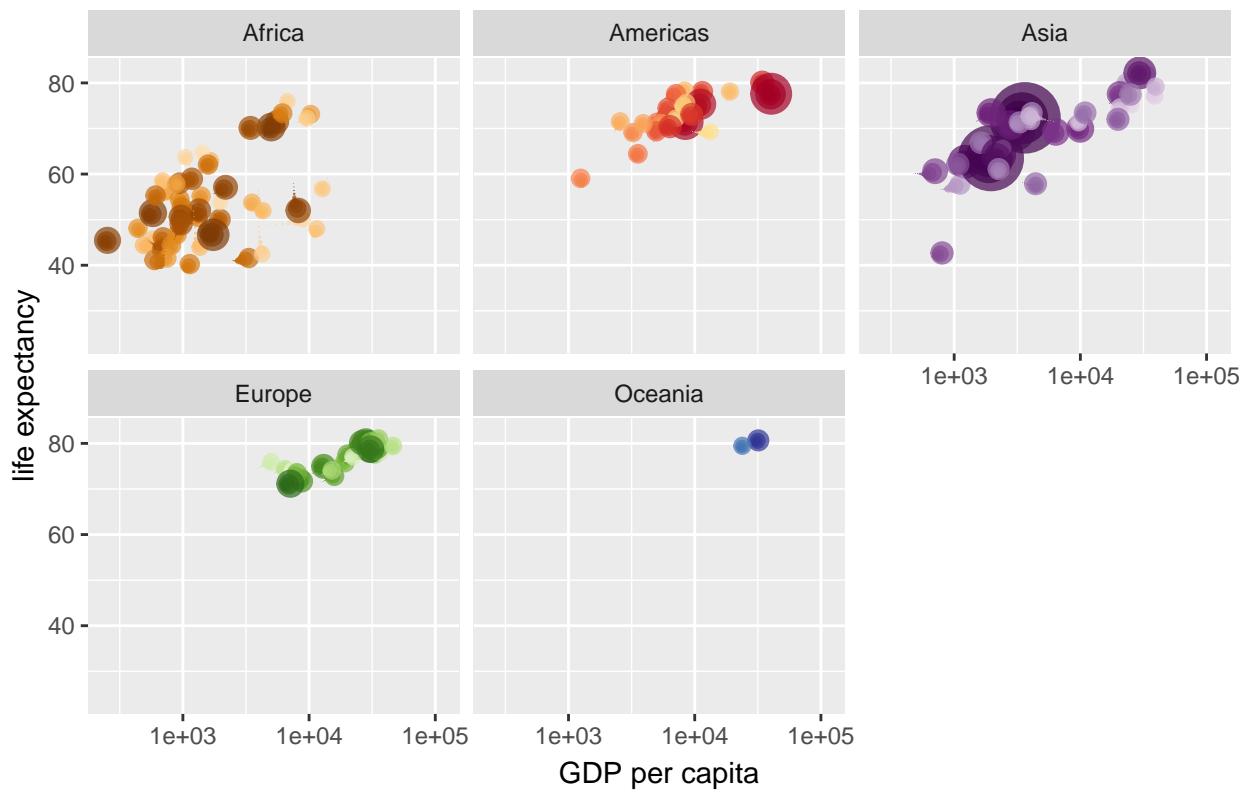
Year: 2003



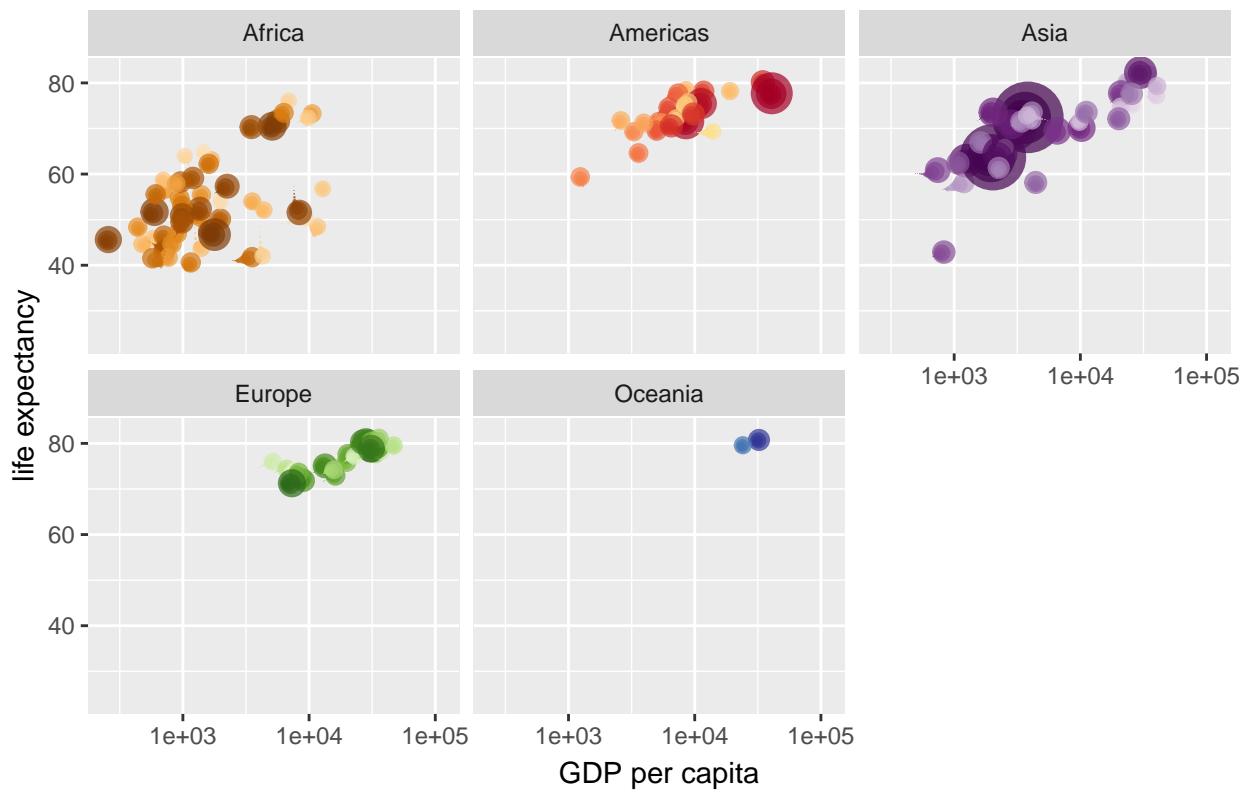
Year: 2003



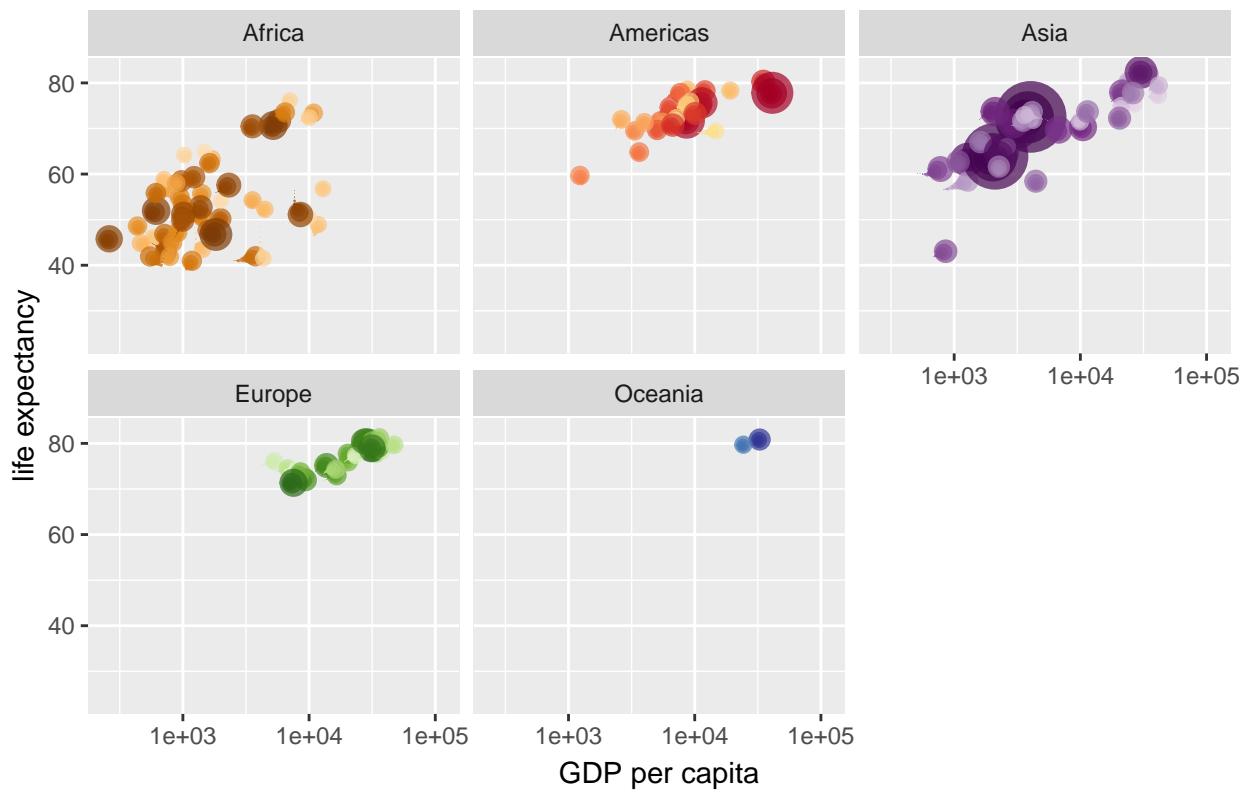
Year: 2004



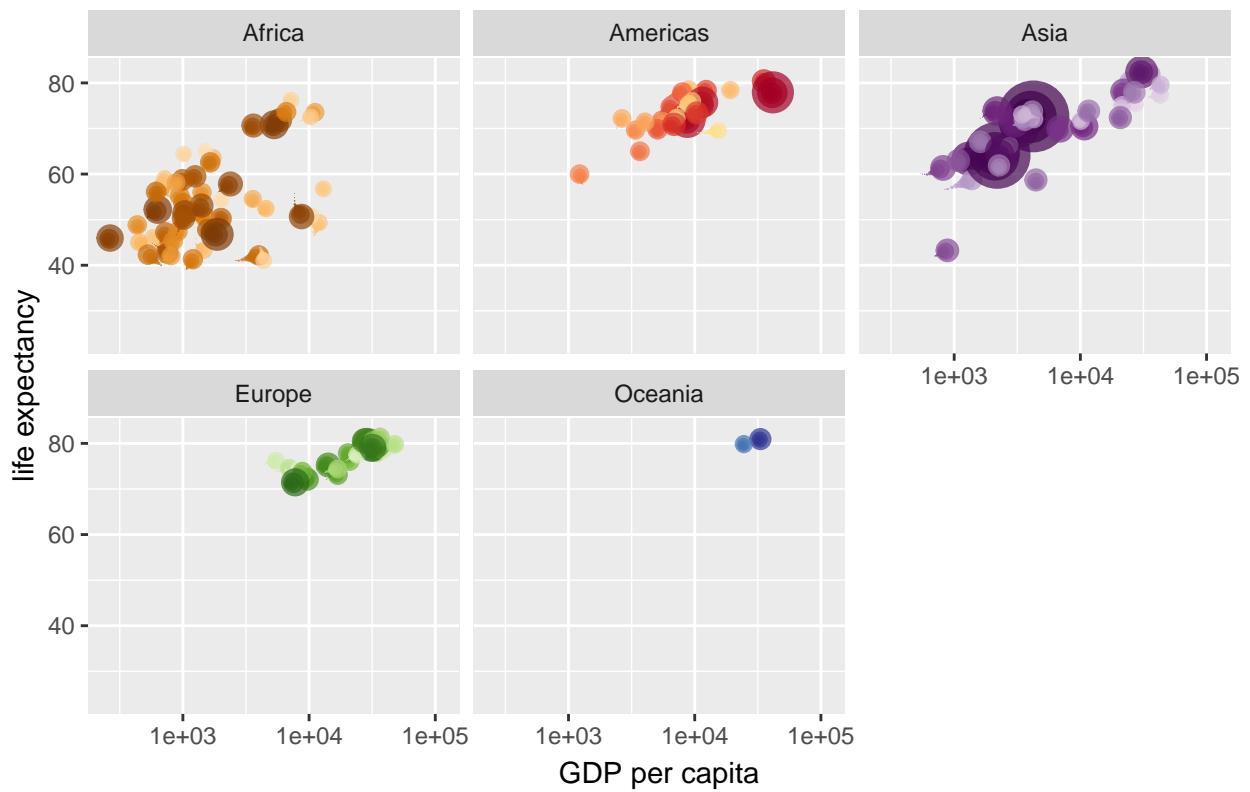
Year: 2004



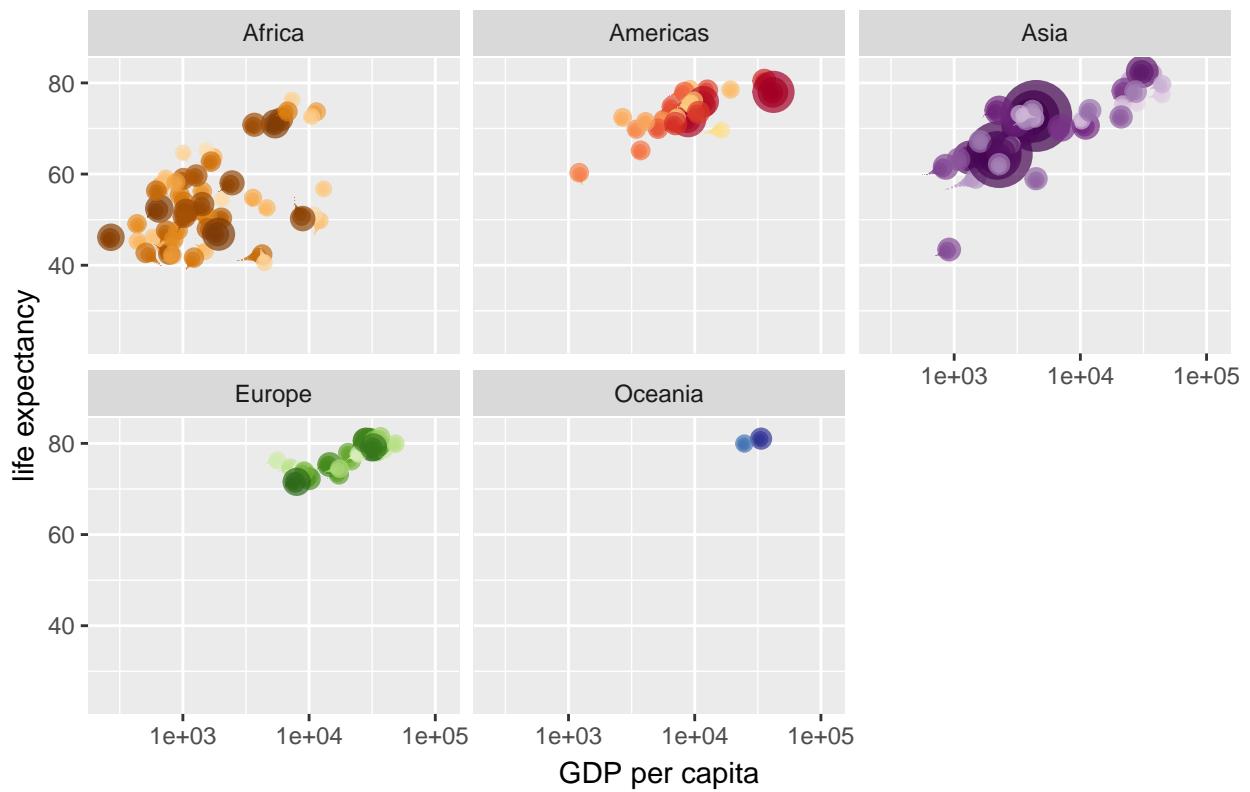
Year: 2005



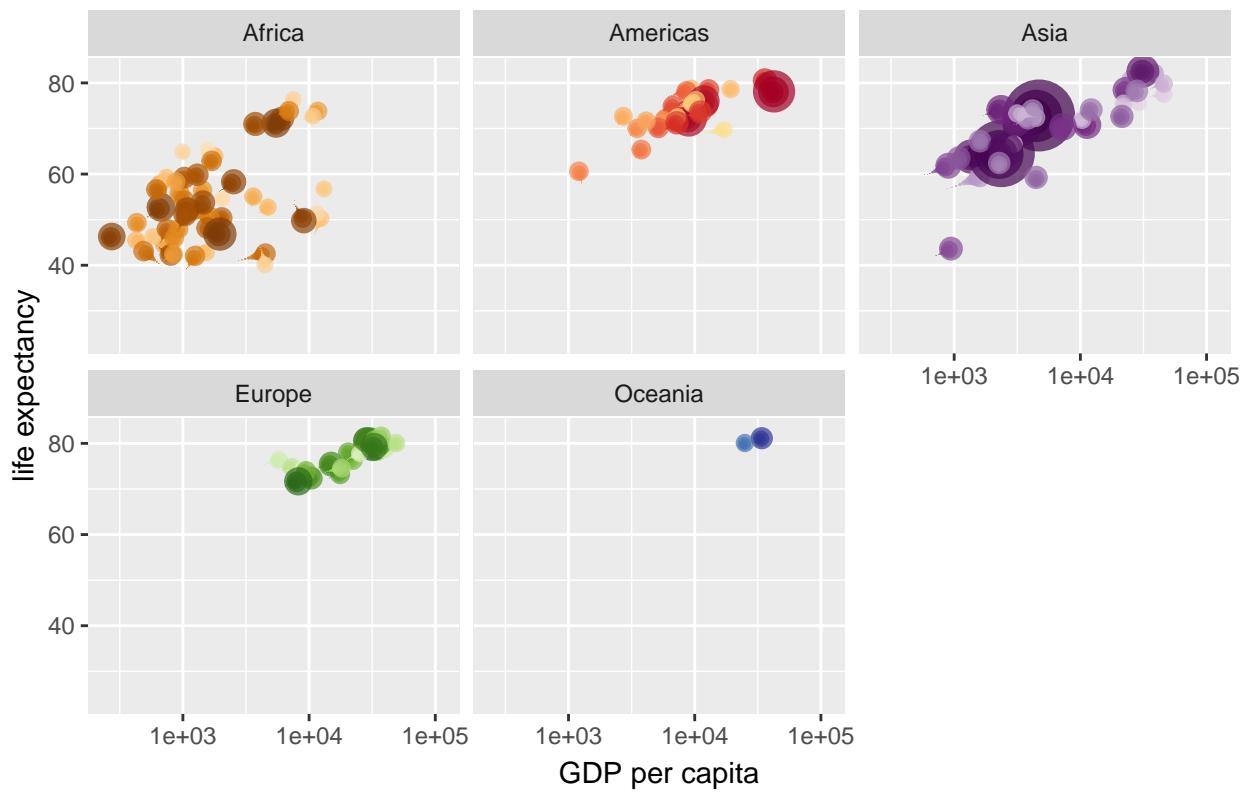
Year: 2005



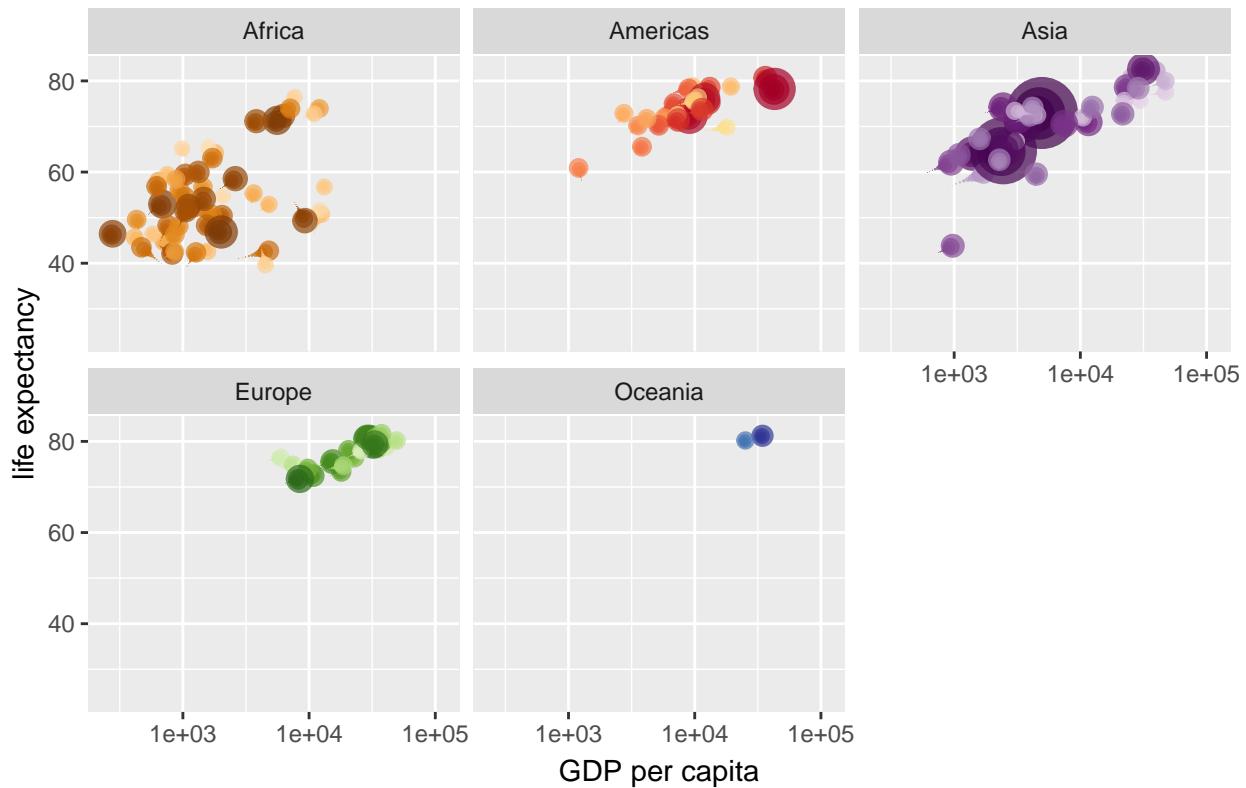
Year: 2006



Year: 2006



Year: 2007



## Section 9: Combining Plots

```
#install.packages("patchwork") ## un-comment to install if needed
```

```
library(patchwork)

# These are 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))

# This part is for patchwork
(p1 | p2 | p3) /p4
```

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

