Data visualization and ggplot2

7 октября 2015

Виноградова Светлана

Do's and Don'ts for Effective Graphs

Goal: create more effective graphs

According to Naomi Robbins, effective graphs "improve understanding of data". They do not confuse or mislead.

We want to:

- Facilitate comparisons
- Reveal trends

Why we shouldn't use pie charts

The most loathed graph of all and yet surprisingly common.

Give your average person a bunch of numbers that add up to one and they want to make a pie chart.

Why?

Why are they wrong?

http://www.richardhollins.com/blog/why-pie-charts-suck/

"The only worse design than a pie chart is several of them."

This animation created by Darkhorse Analytics illustrates how communication can be greatly enhanced by eliminating clutter and de-emphasizing supporting elements. Every aspect of a figure should be there on a “need to have it” basis.

http://stat545-ubc.github.io/img/less-is-more-darkhorse-analytics.gif

from Wickham's *Tidy Data*

In **tidy data**:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

messy

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

tidy

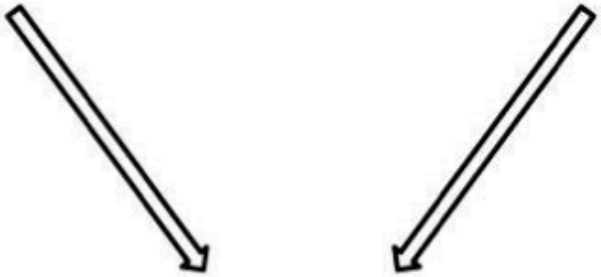
name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

https://github.com/jennybc/ggplot2-tutorial/tree/master/ggplot2-tutorial-slides

from White et al's *Nine simple ways* ...

	Habitat		
Species	X	Y	Z
A	0	3	0
B	1	0	2

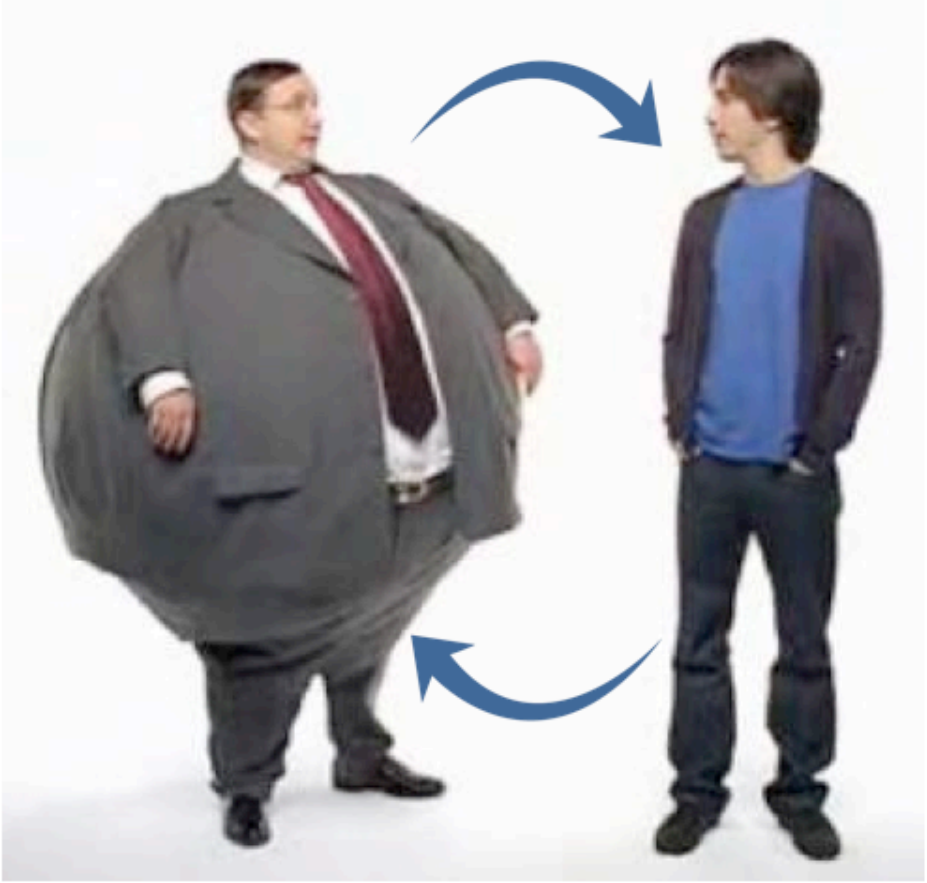
Species	HabitatX	HabitatY	HabitatZ
A	0	3	0
B	1	0	2



Species	Habitat	Abundance
A	Y	3
B	X	1
B	Z	2

https://github.com/jennybc/ggplot2-tutorial/tree/master/ggplot2-tutorial-slides

reshape your data

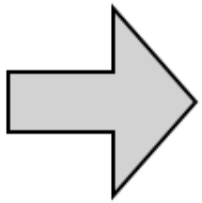


data has a tendency to get shorter and wider, but tall and thin often better for analysis + visualization

from Wickham's *Tidy Data*
see also reshape2

reshape2::melt
tidyr::gather

row	a	b	c
a	1	4	7
b	2	5	8
c	3	6	9



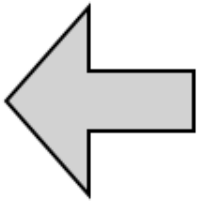
row	column	value
a	a	1
b	a	2
c	a	3
a	b	4
b	b	5
c	b	6
a	c	7
b	c	8
c	c	9

https://github.com/jennybc/ggplot2-tutorial/tree/master/ggplot2-tutorial-slides

from Wickham's *Tidy Data*
see also reshape2

reshape2::cast
tidyr::spread

row	a	b	c
a	1	4	7
b	2	5	8
c	3	6	9

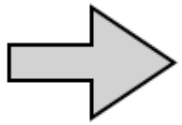


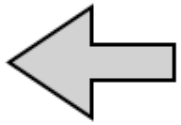
row	column	value
a	a	1
b	a	2
c	a	3
a	b	4
b	b	5
c	b	6
a	c	7
b	c	8
c	c	9

https://github.com/jennybc/ggplot2-tutorial/tree/master/ggplot2-tutorial-slides

gather

row	a	b	c
a	1	4	7
b	2	5	8
c	3	6	9





row	column	value
a	a	1
b	a	2
c	a	3
a	b	4
b	b	5
c	b	6
a	c	7
b	c	8
c	c	9

spread

typical usage pattern:

gather to facilitate analysis and visualization

spread to make compact tables that are nicer for eyeballs

https://github.com/jennybc/ggplot2-tutorial/tree/master/ggplot2-tutorial-slides

Пакет reshape2

```
install.packages("reshape2")
library(“reshape2”)

names(airquality) <- tolower(names(airquality))
head(airquality)

aql <- melt(airquality)
head(aql)
```

Контролируем, какие переменные являются id

```
aql <- melt(airquality, id.vars = c("month", "day"),
  variable.name = "climate_variable",
  value.name = "climate_value")
head(aql)
```

И обратно…

```
aql <- melt(airquality, id.vars = c("month", "day"))
aqw <- dcast(aql, month + day ~ variable)
head(aqw)
```

ggplot2

Author

ggplot2 was developed by Hadley Wickham, assistant professor of statistics at Rice University, Houston. In July 2010 the latest stable release (Version 0.8.8) was published.

Hadley Wickham
Dobelman Family Junior Chair
Statistics, Rice University
6100 Main St MS#138
Houston TX 77005-1827

February 3, 2010

515 450 8171
hadley@rice.edu
http://had.co.nz

2008 Ph.D. (Statistics), Iowa State University, Ames, IA. “Practical tools for exploring data and models.”

2004 M.Sc. (Statistics), First Class Honours, The University of Auckland, Auckland, New Zealand.

2002 B.Sc. (Statistics, Computer Science), First Class Honours, The University of Auckland, Auckland, New Zealand.

1999 Bachelor of Human Biology, First Class Honours, The University of Auckland, Auckland, New Zealand.

http://www.ceb-institute.org/bbs/wp-content/uploads/2011/09/handout_ggplot2.pdf

```
### устанавливаем и загружаем пакет
install.package("ggplot2")
library("ggplot2")
```

data, in data.frame form

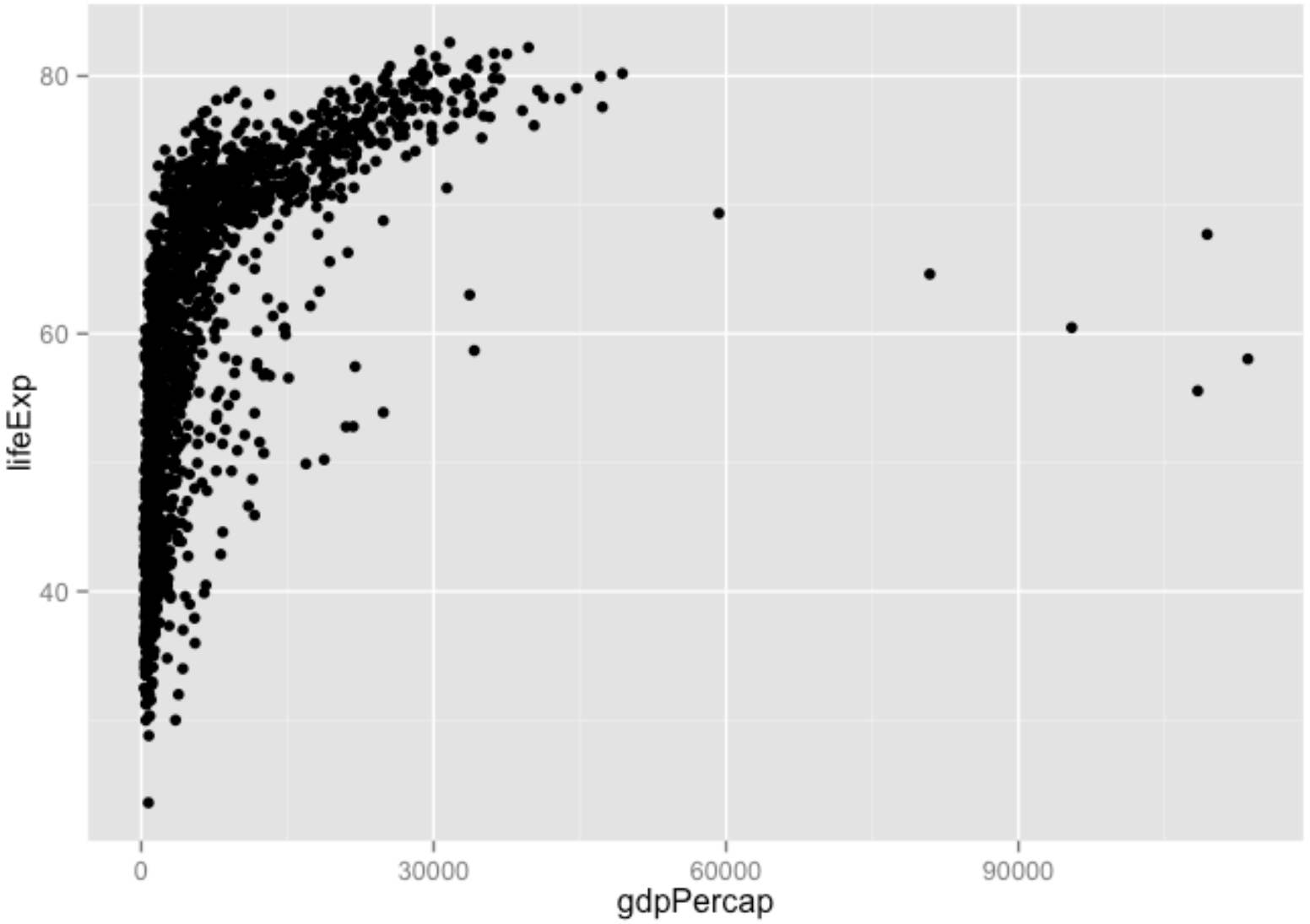
aesthetic: map variables into properties people can perceive visually ... position, color, line type?

geom: specifics of what people see ... points? lines?

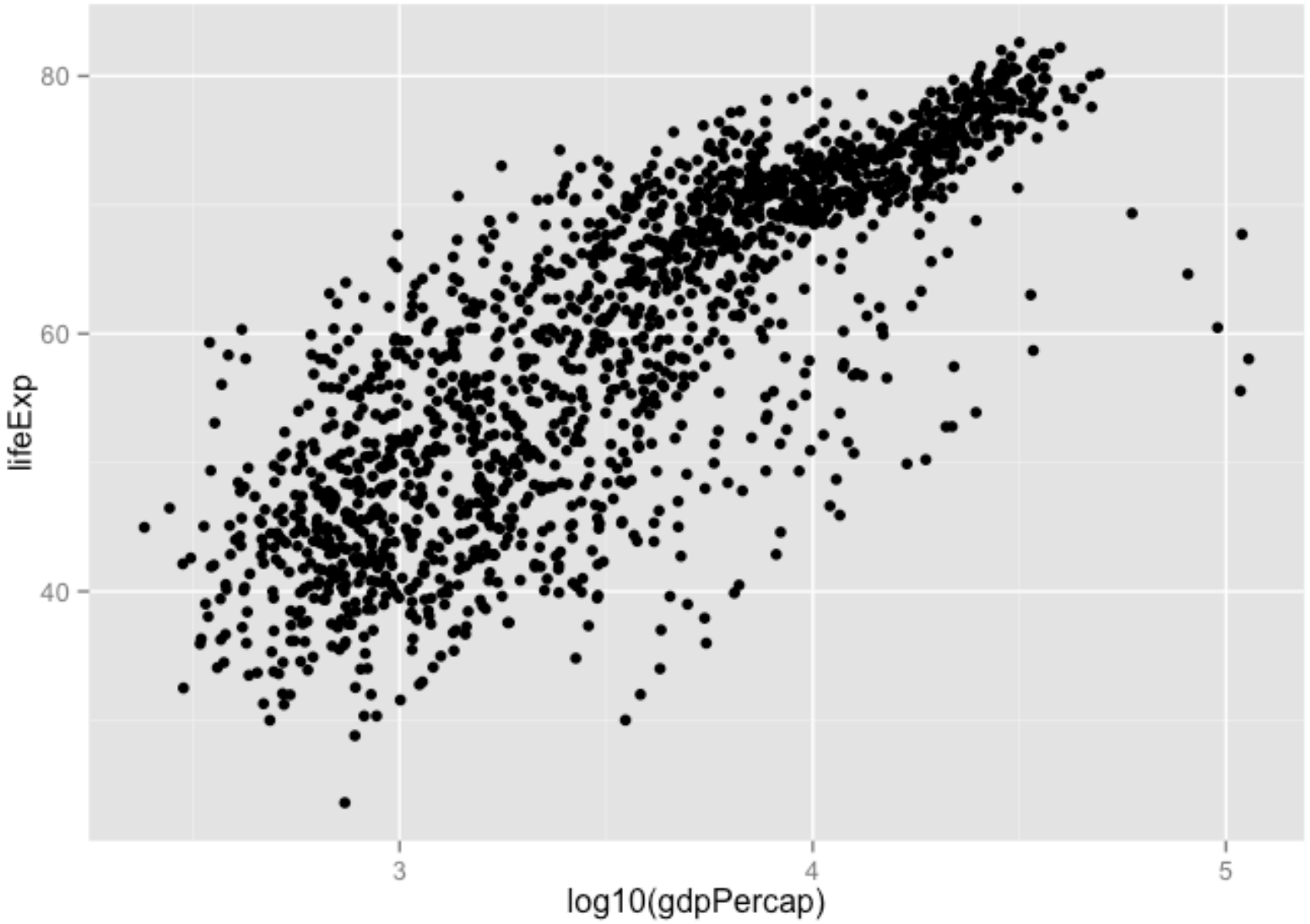
scale: map data values into "computer" values

stat: summarization/transformation of data

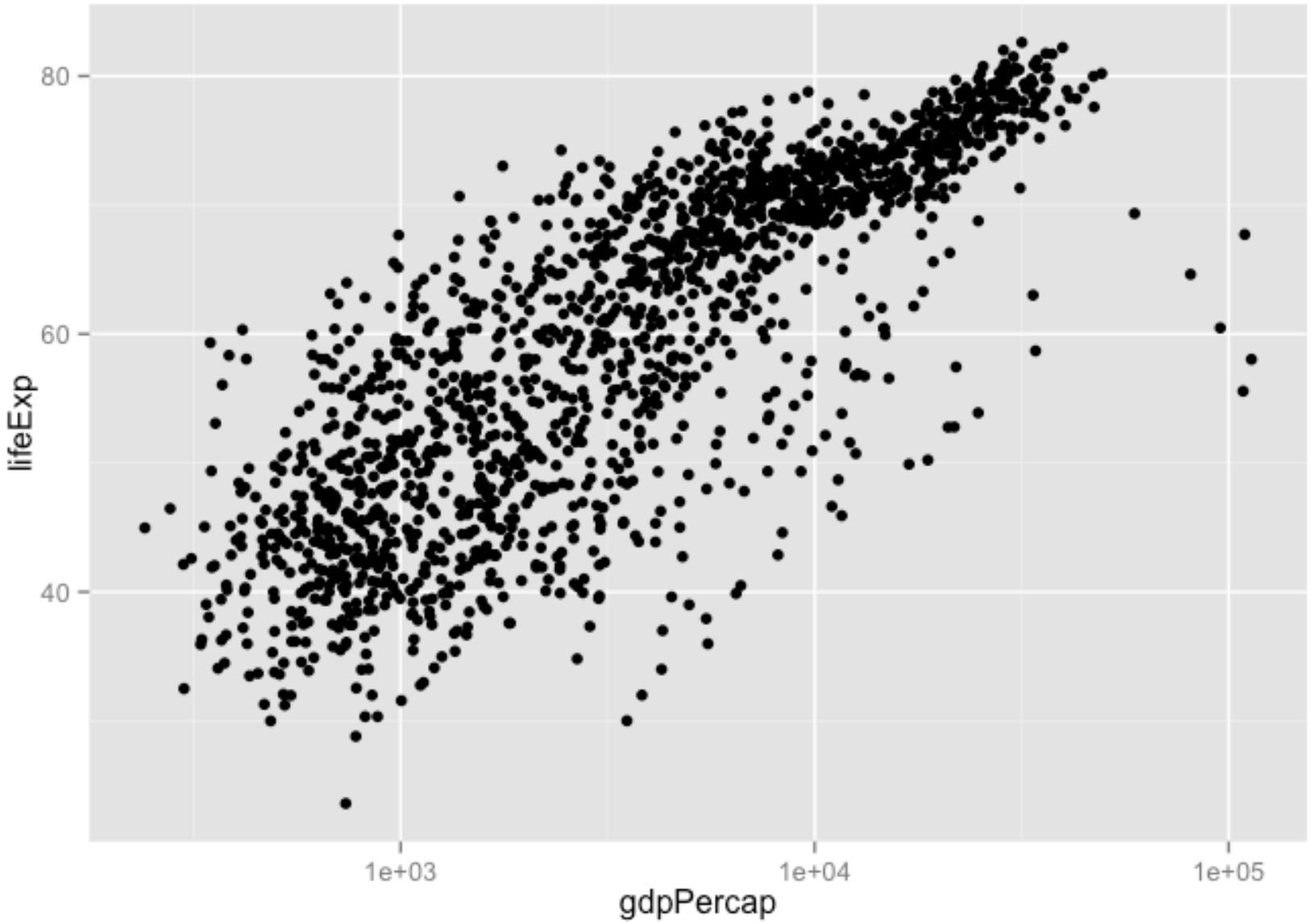
facet: juxtapose related mini-plots of data subsets




```
p <- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp))
p + geom_point()
```

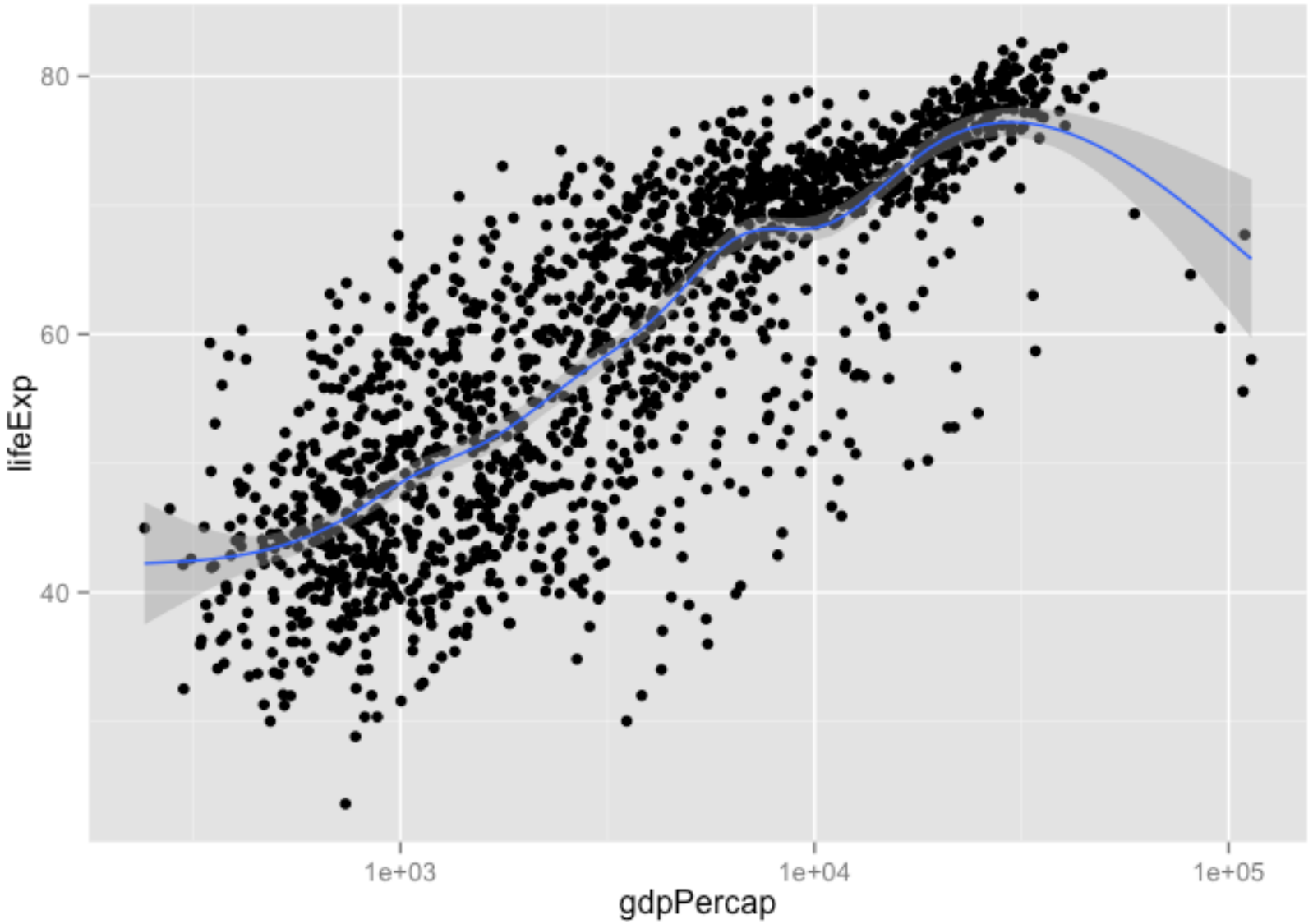



```
ggplot(gapminder, aes(x = log10(gdpPercap), y = lifeExp)) +
geom_point()
```

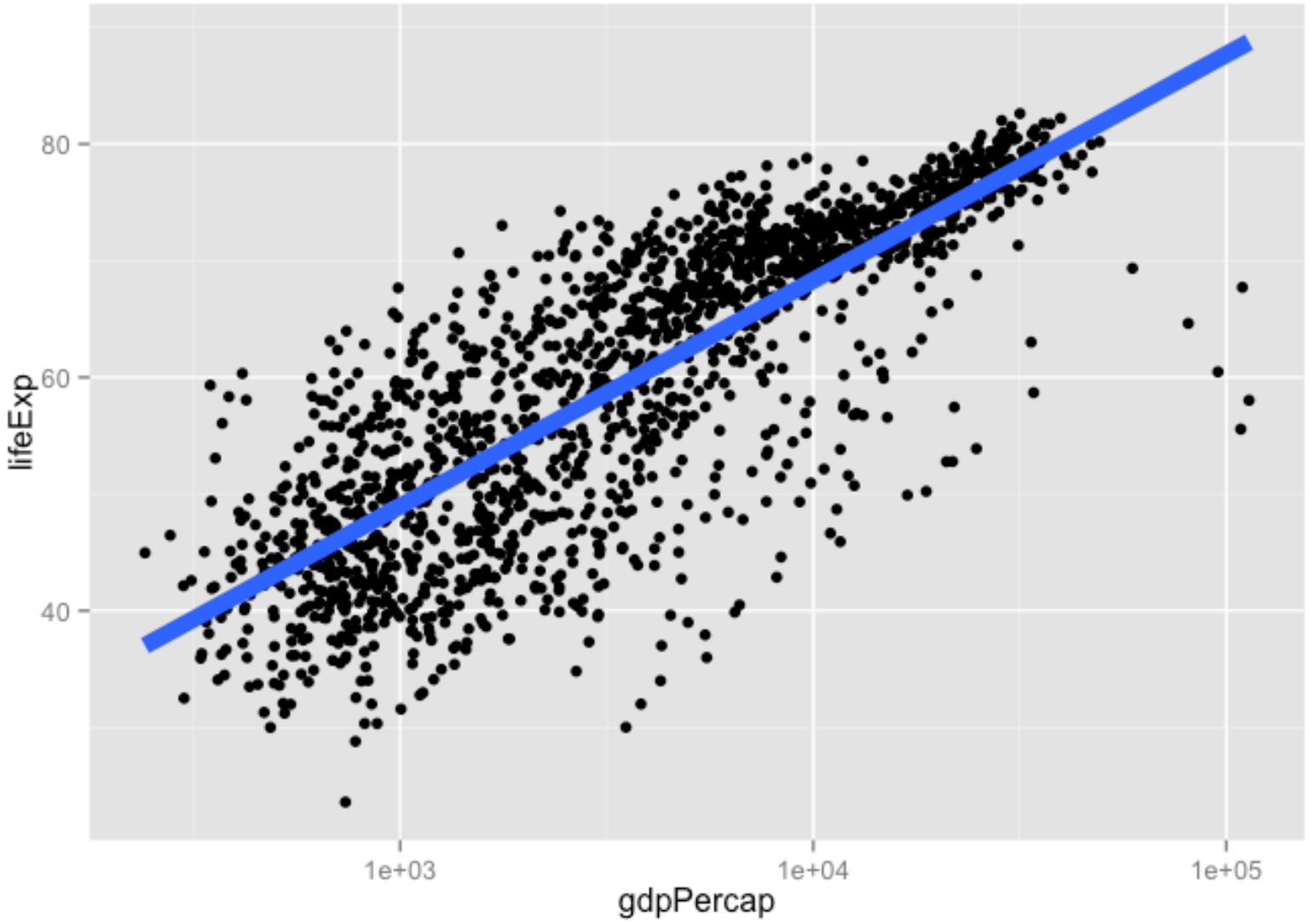



```
p + geom_point() + scale_x_log10()
```


```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, color =
continent)) +
  geom_point() + scale_x_log10()
```

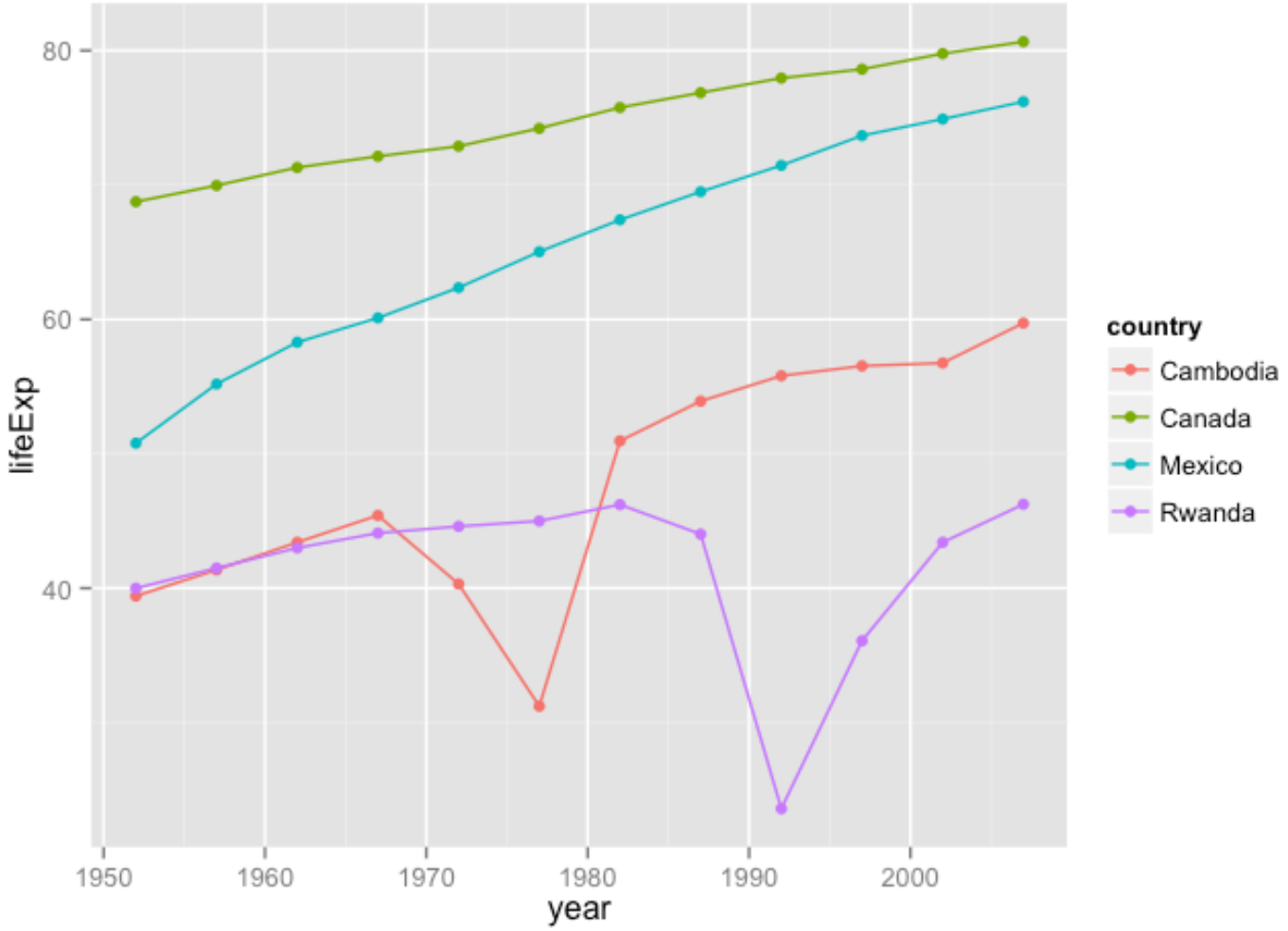



```
p + geom_point() + geom_smooth()
```

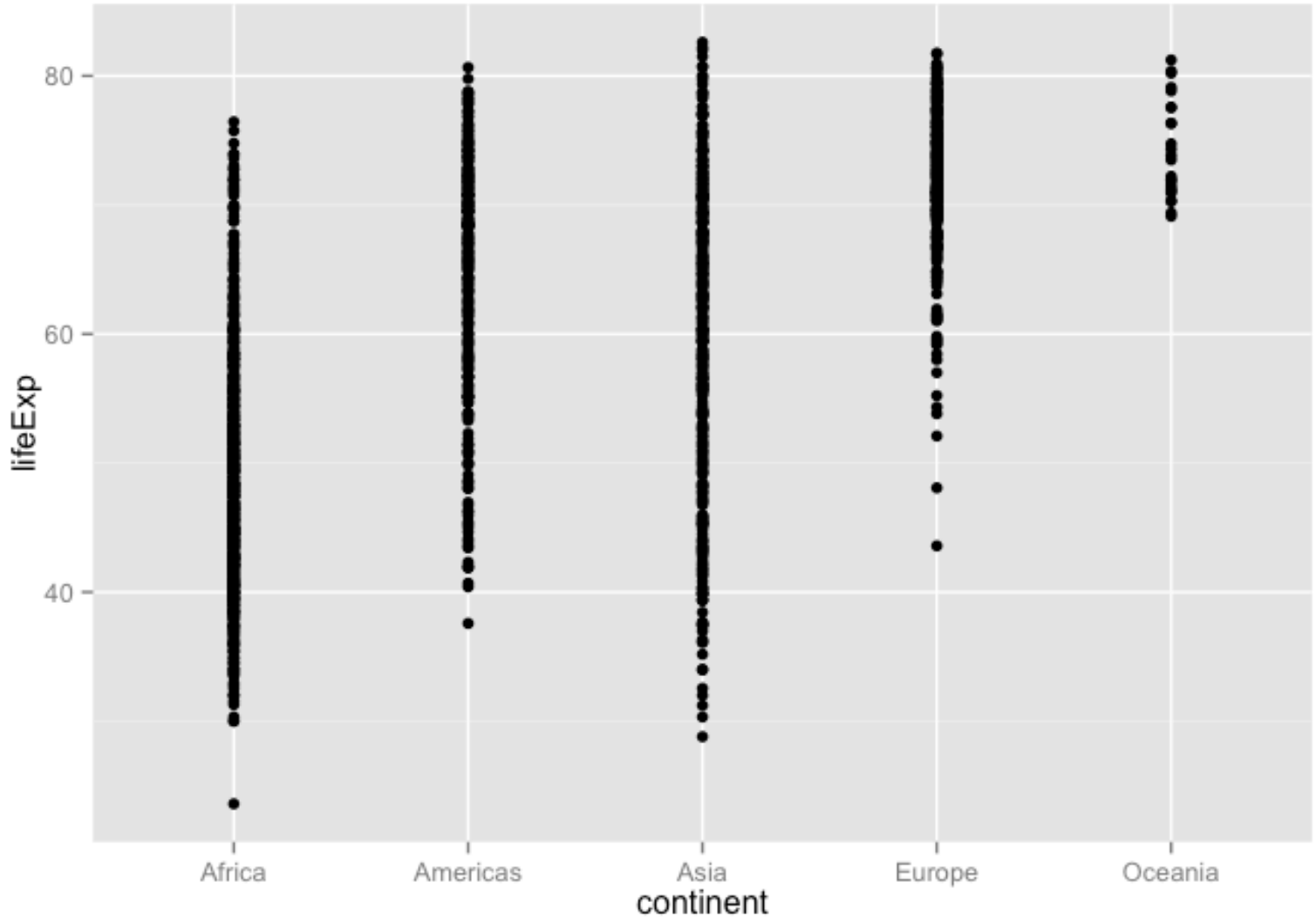



```
p + geom_point() + geom_smooth(lwd = 3, se = FALSE, method
= "lm")
```

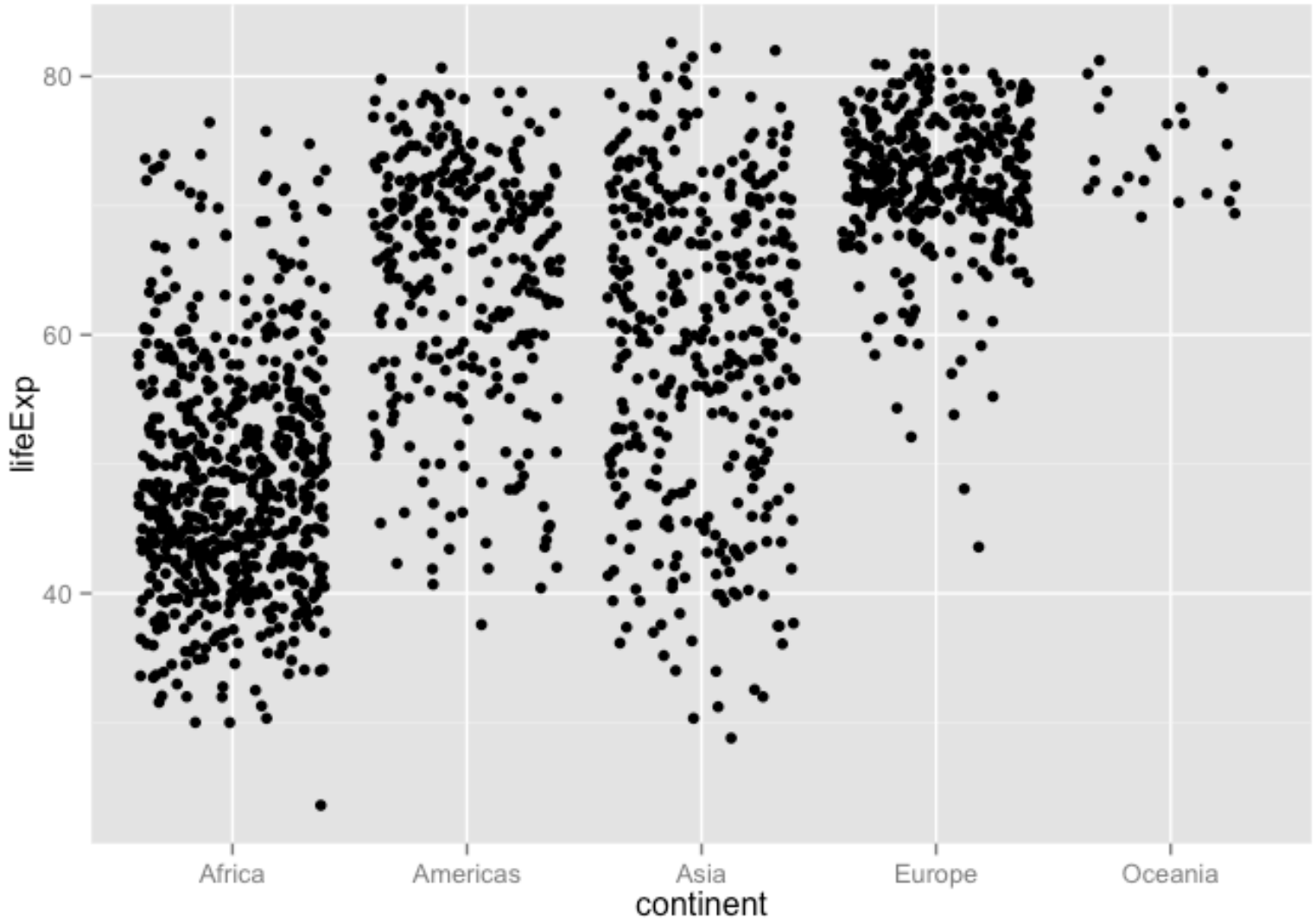

```
ggplot(subset(gapminder, country == "Zimbabwe"),
       aes(x = year, y = lifeExp)) + geom_line() + geom_point()
```



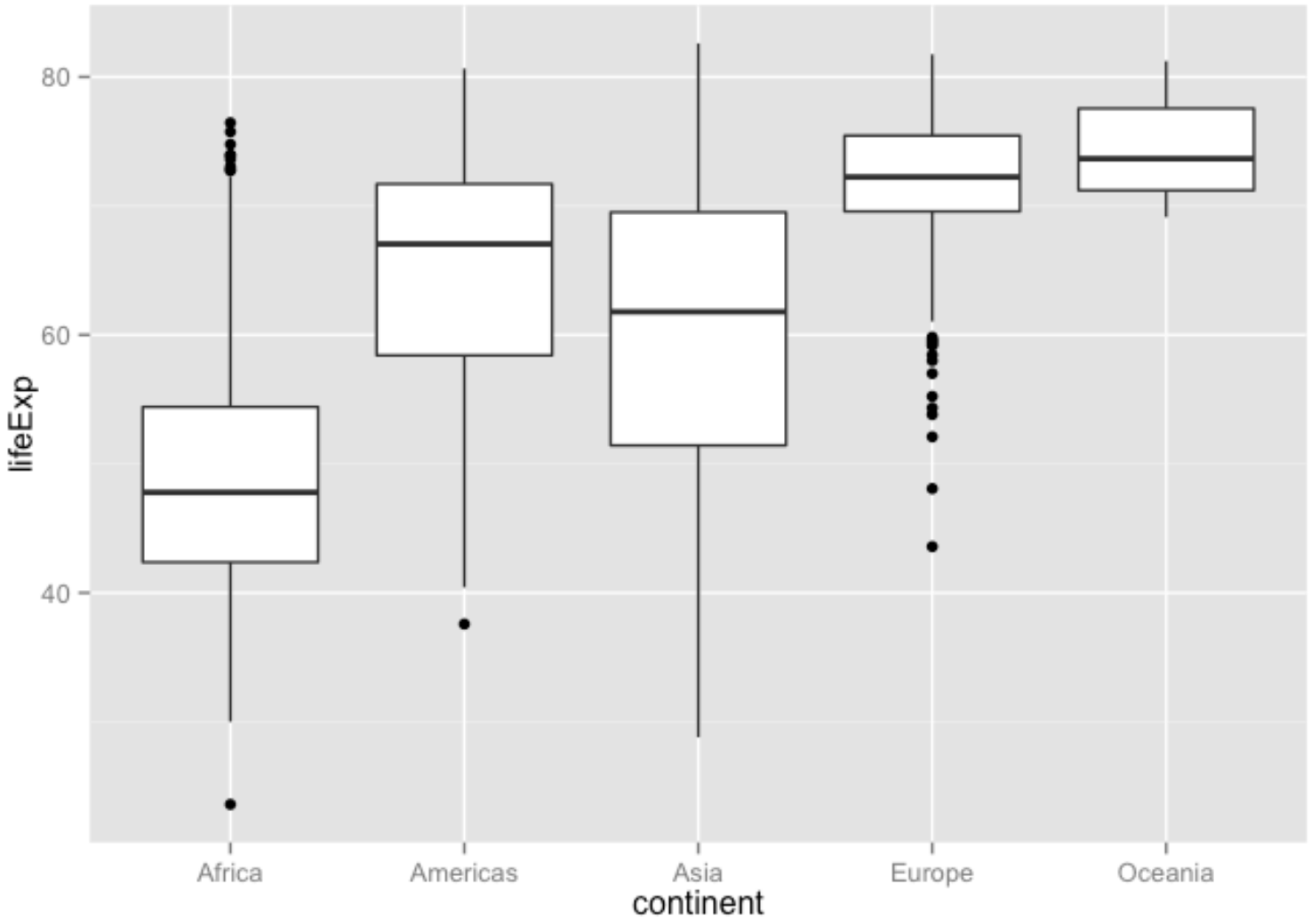

```
jCountries <- c("Canada", "Rwanda", "Cambodia", "Mexico")
ggplot(subset(gapminder, country %in% jCountries),
       aes(x = year, y = lifeExp, color = country)) +
geom_line() + geom_point()
```



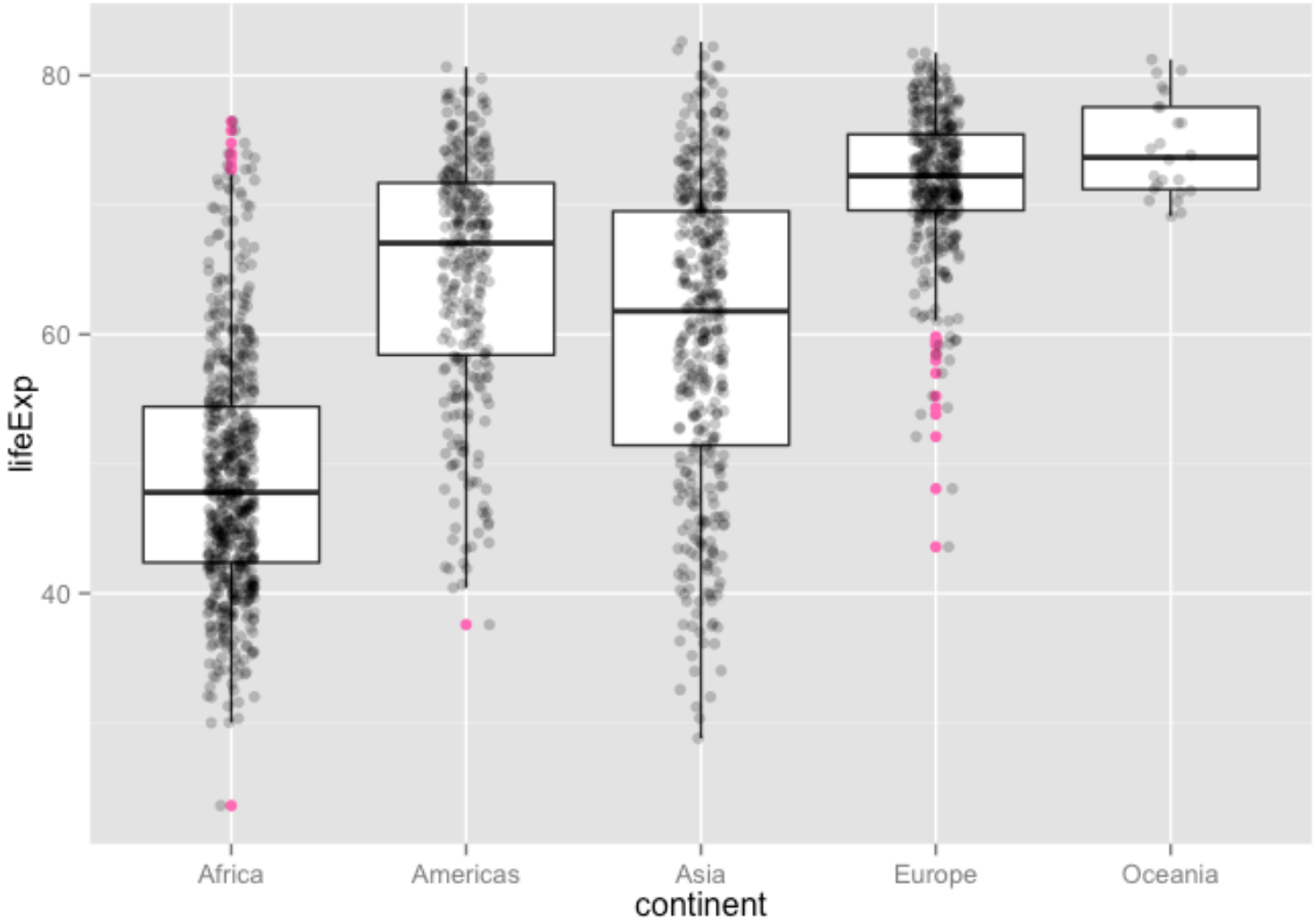

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
geom_point()
```



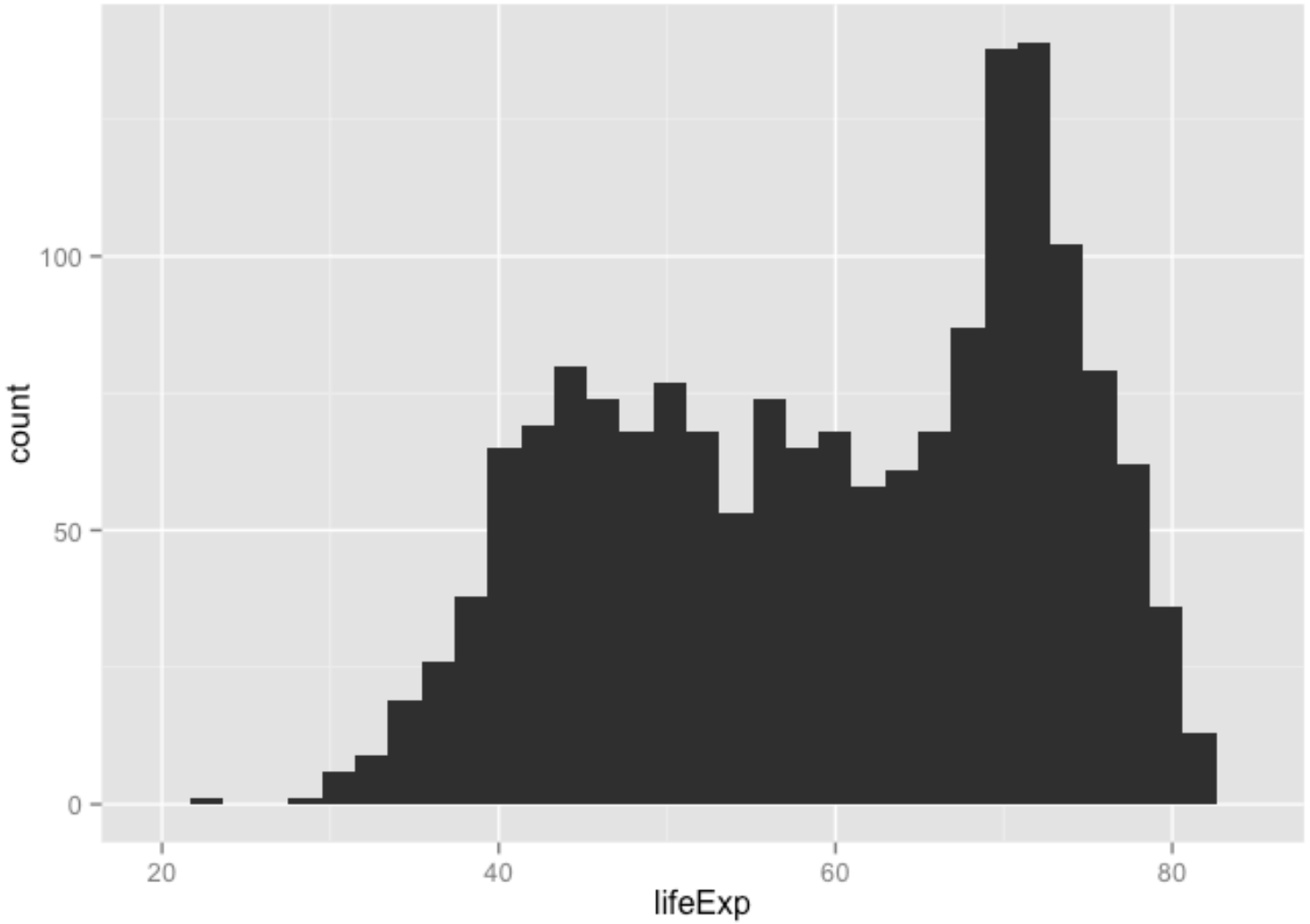

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
geom_jitter()
```



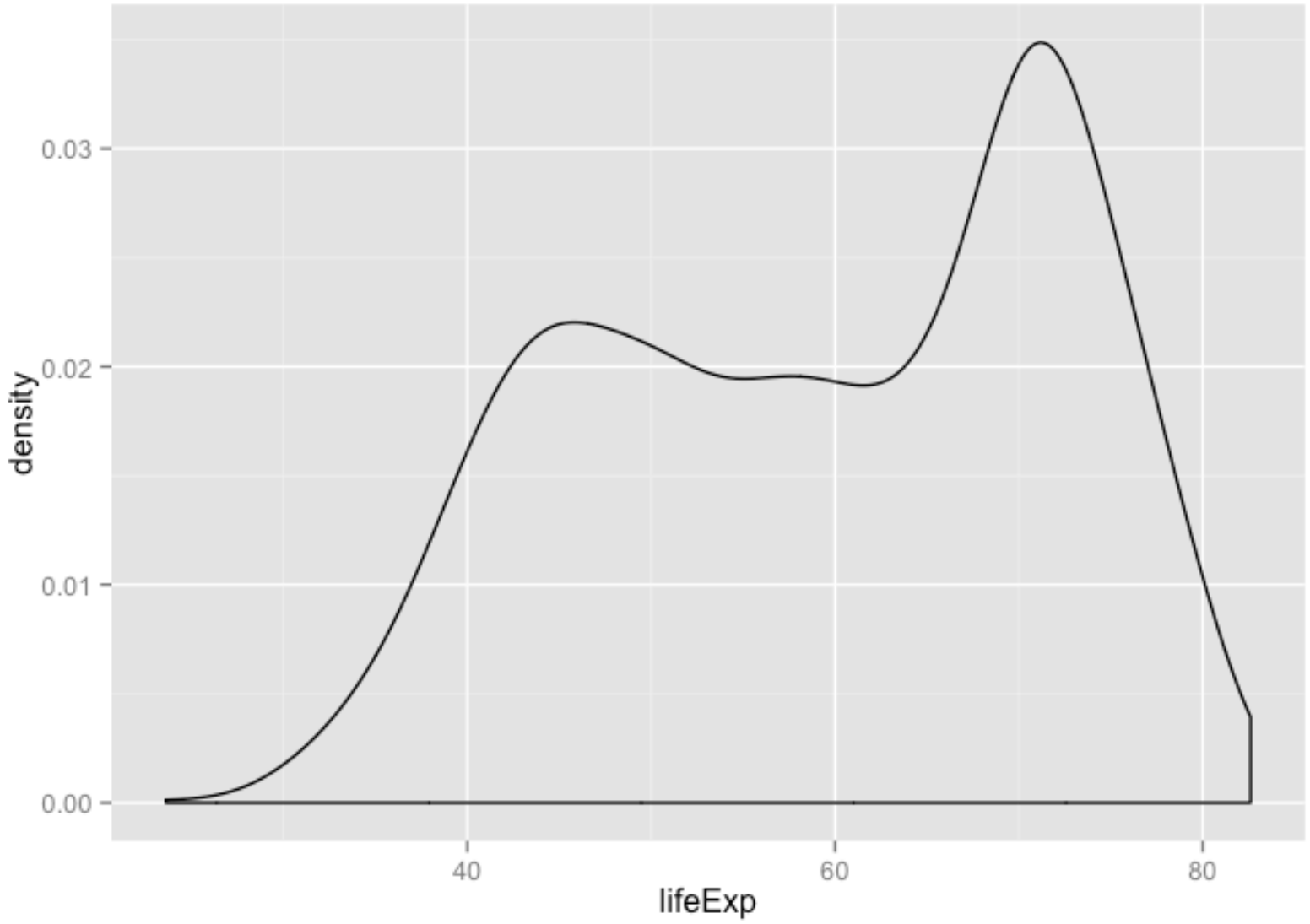

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
geom_boxplot()
```



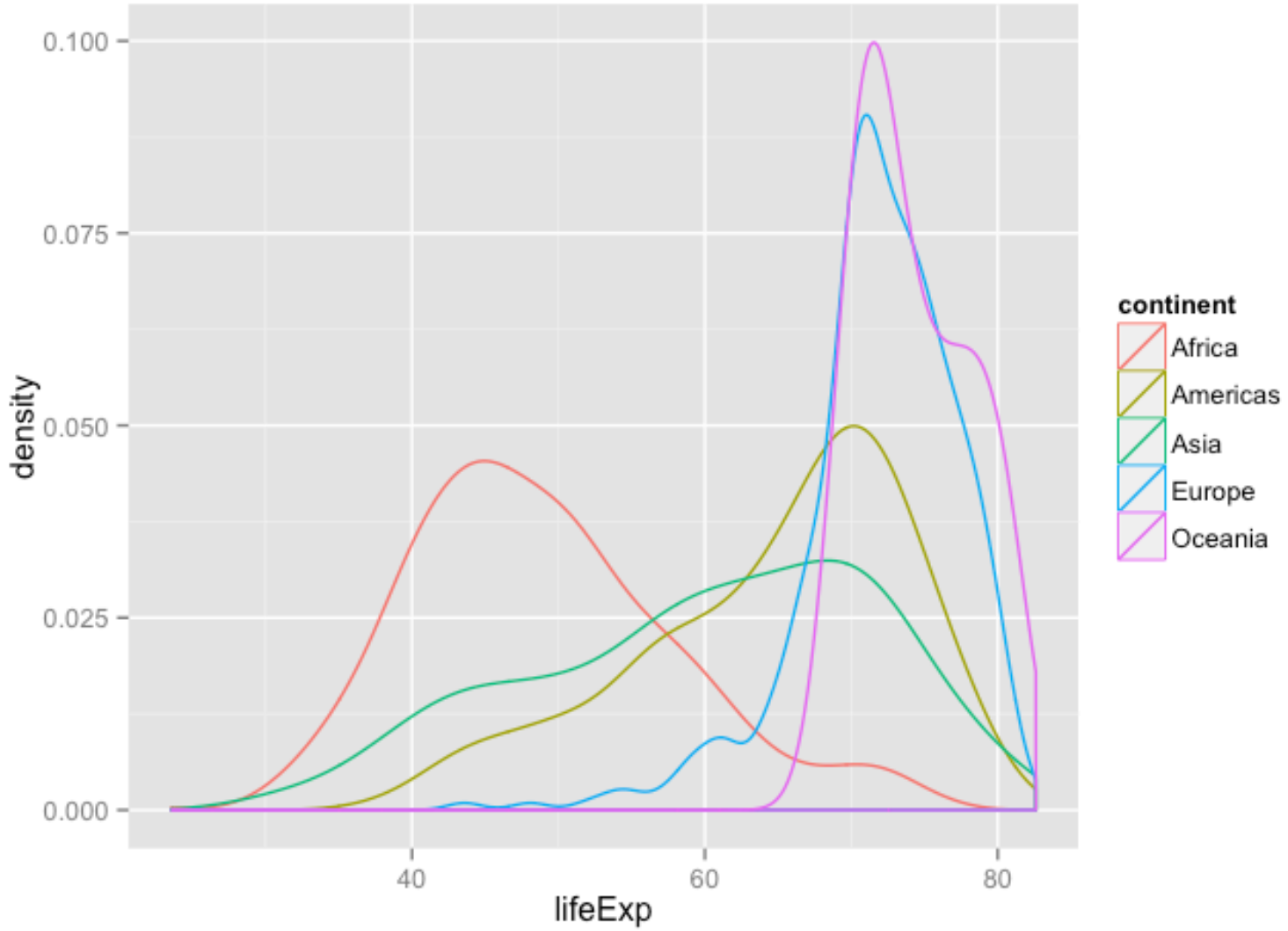

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
  geom_boxplot(outlier.colour = "hotpink") +
  geom_jitter(position = position_jitter(width = 0.1, height =
0), alpha = 1/4)
```



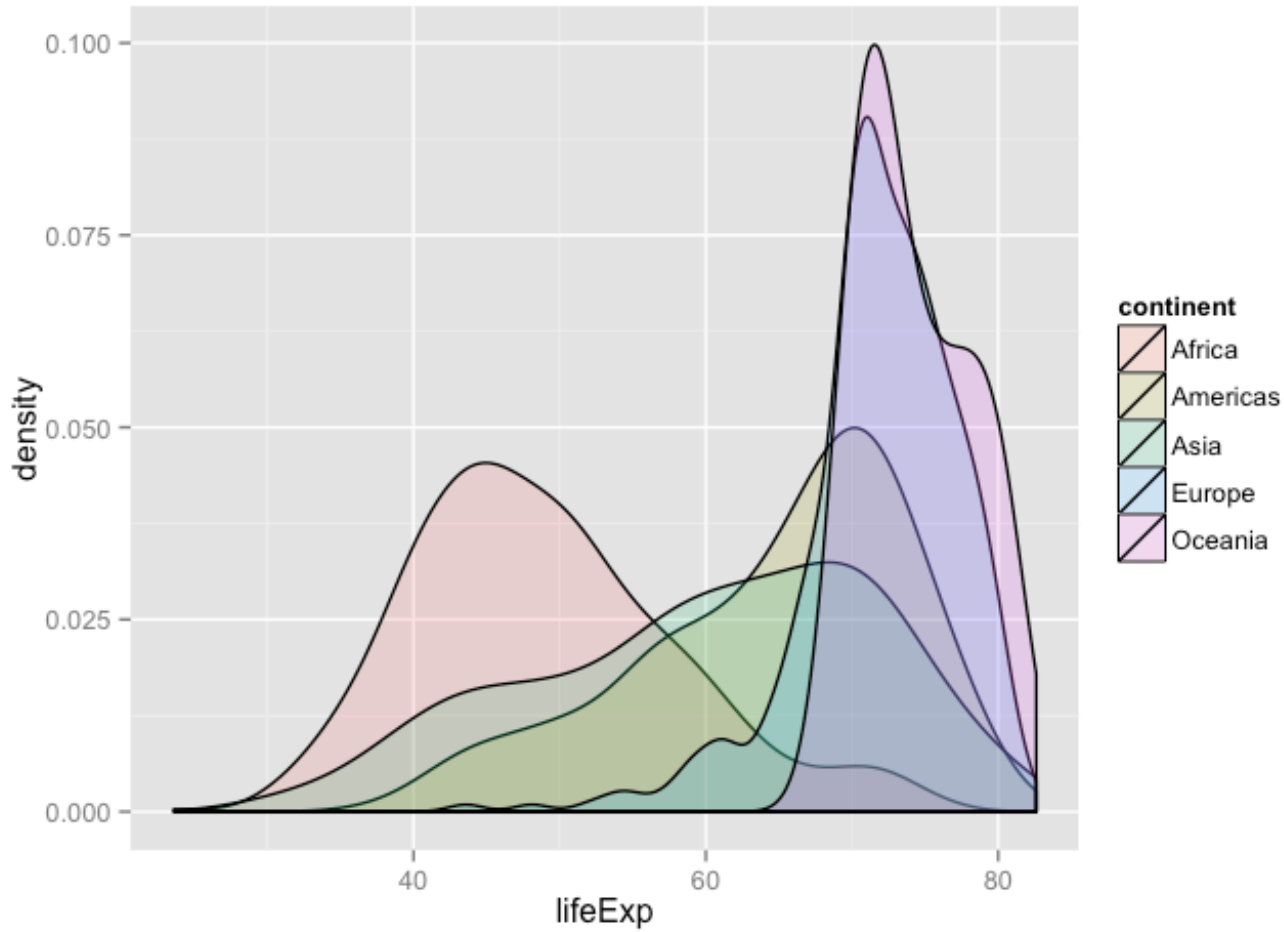

```
ggplot(gapminder, aes(x = lifeExp)) + geom_histogram()
```



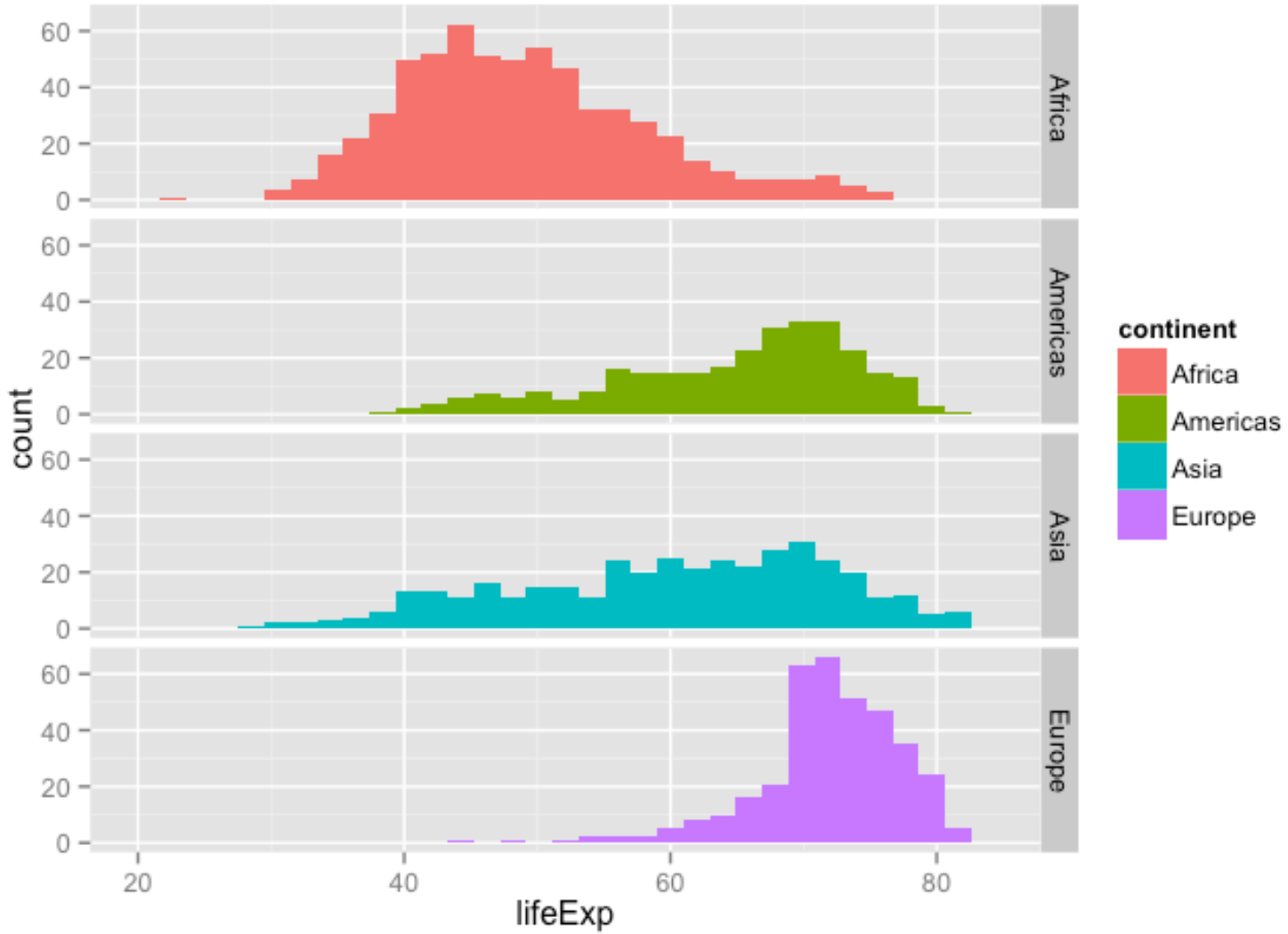

```
ggplot(gapminder, aes(x = lifeExp)) + geom_density()
```



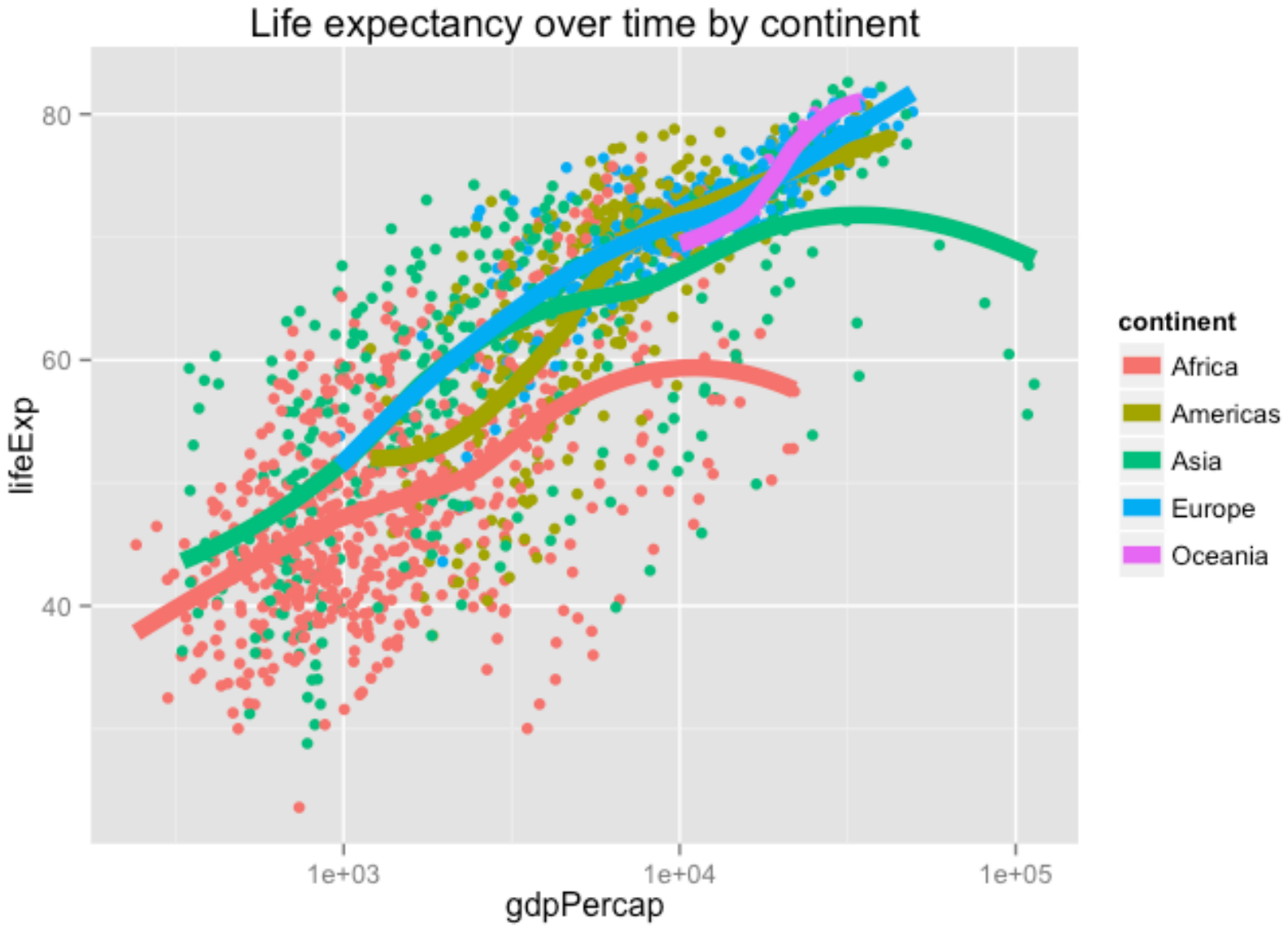

```
ggplot(gapminder, aes(x = lifeExp, color = continent)) +
geom_density()
```



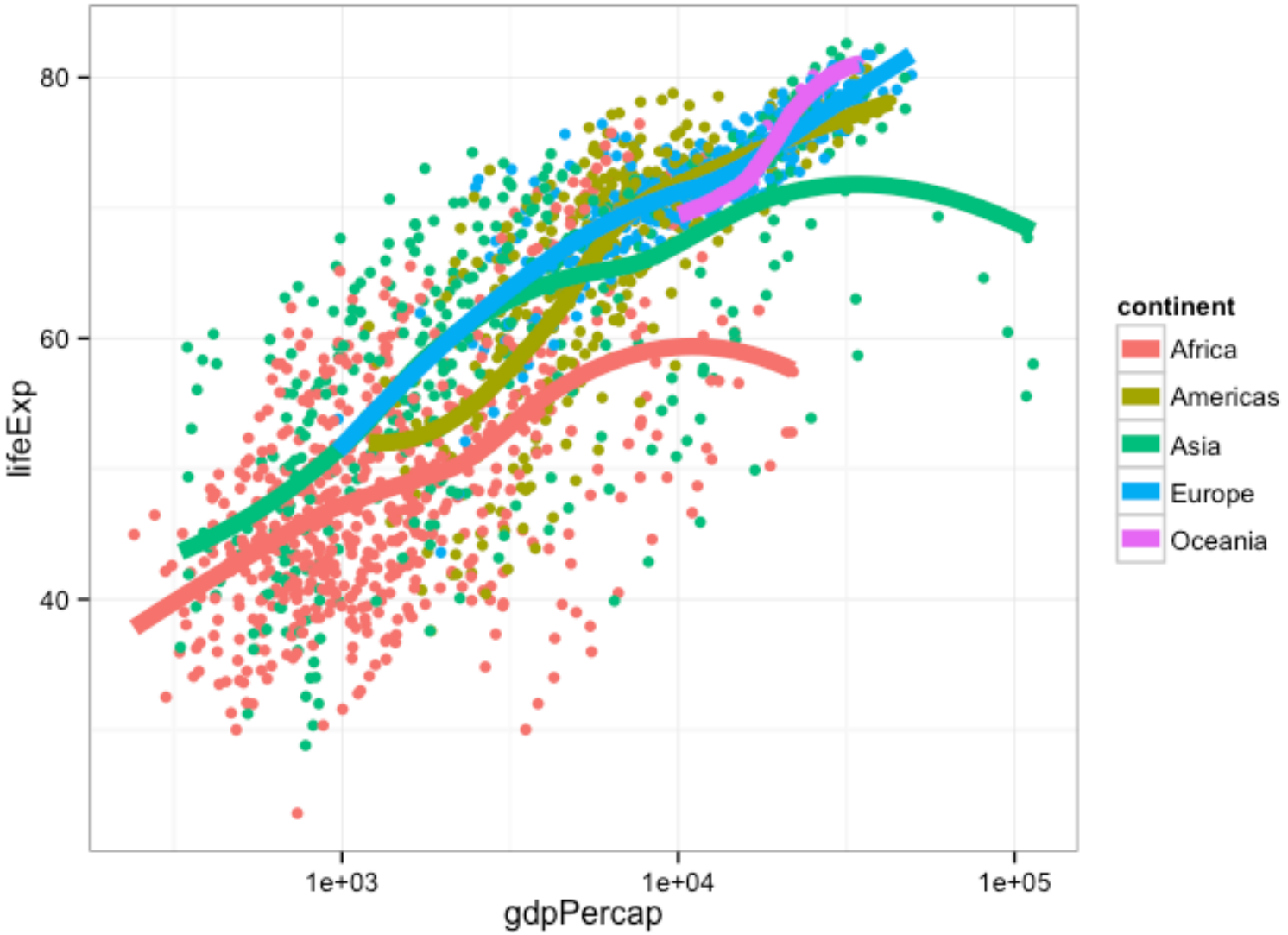

```
ggplot(gapminder, aes(x = lifeExp, fill = continent)) +
  geom_density(alpha = 0.2)
```




```
ggplot(subset(gapminder, continent != "Oceania"),
       aes(x = lifeExp, fill = continent)) +
geom_histogram() +
  facet_grid(continent ~ .)
```




```
p + ggtitle("Life expectancy over time by continent")
```


80
60
40
lifeExp
1e+03
1e+04
1e+05
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania

```
p + theme_bw()
```