Язык R
лекция 5

Артем Артемов
Елена Ставровская
Анастасия Жарикова

30 сентября 2016

reshape2

reshape2

```
>install.packages("reshape2")
>library(reshape2)
```

```
>a=data.frame(name=c('John', 'Mary', 'Peter', 'Susan'),
              sex=c('m','f','m','f'),
              age=c(26,21,19,29),
              weight=c(82, 56, 79, 60),
              height=c(182, 171, 179, 175))
```

name	sex	age	weight	height
John	m	26	82	182
Mary	f	21	56	171
Peter	m	19	79	179
Susan	f	29	60	175

«Расплавление» данных

```
> a_melt -> melt(a, id.vars = c('name','sex'), variable.name = c('a_variable'),
                 value.name = 'a_name')
```

name	sex	age	weight	height
John	m	26	82	182
Mary	f	21	56	171
Peter	m	19	79	179
Susan	f	29	60	175

name	sex	a_variable	a_name
John	m	age	26
Mary	f	age	21
Peter	m	age	19
Susan	f	age	29
John	m	weight	82
Mary	f	weight	56
Peter	m	weight	79
Susan	f	weight	60
John	m	height	182
Mary	f	height	171
Peter	m	height	179
Susan	f	height	175

Формирование данных

> dcast(a_melt,
> name ~ a_variable)

name	sex	a_variable	a_name
John	m	age	26
Mary	f	age	21
Peter	m	age	19
Susan	f	age	29
John	m	weight	82
Mary	f	weight	56
Peter	m	weight	79
Susan	f	weight	60
John	m	height	182
Mary	f	height	171
Peter	m	height	179
Susan	f	height	175

> dcast(a_melt,
> name + sex ~ a_variable)

name	age	weight	height
John	26	82	182
Mary	21	56	171
Peter	19	79	179
Susan	29	60	175

name	sex	age	weight	height
John	m	26	82	182
Mary	f	21	56	171
Peter	m	19	79	179
Susan	f	29	60	175

ggplot2

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 | February 3, 2010
Dobelman Family Junior Chair
Statistics, Rice University | 515 450 8171
6100 Main St MS#138 | hadley@rice.edu
Houston TX 77005-1827 | 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

Установка и загрузка пакета:

```
> install.packages("ggplot2")
> library("ggplot2")
```

data, in data.frame form

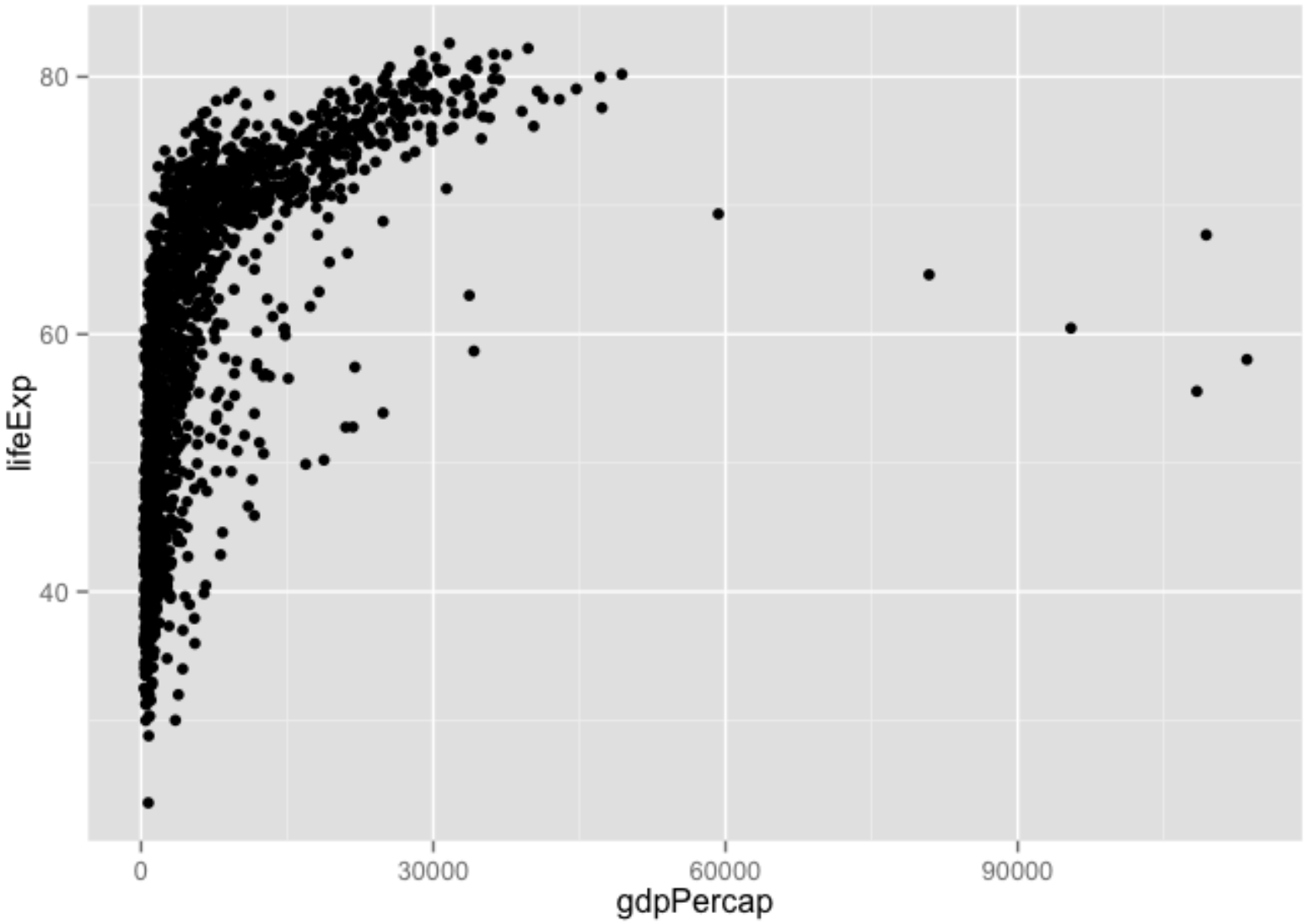
aesthetic: map variables into properties people can perceive visually ... position, color, line type?

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

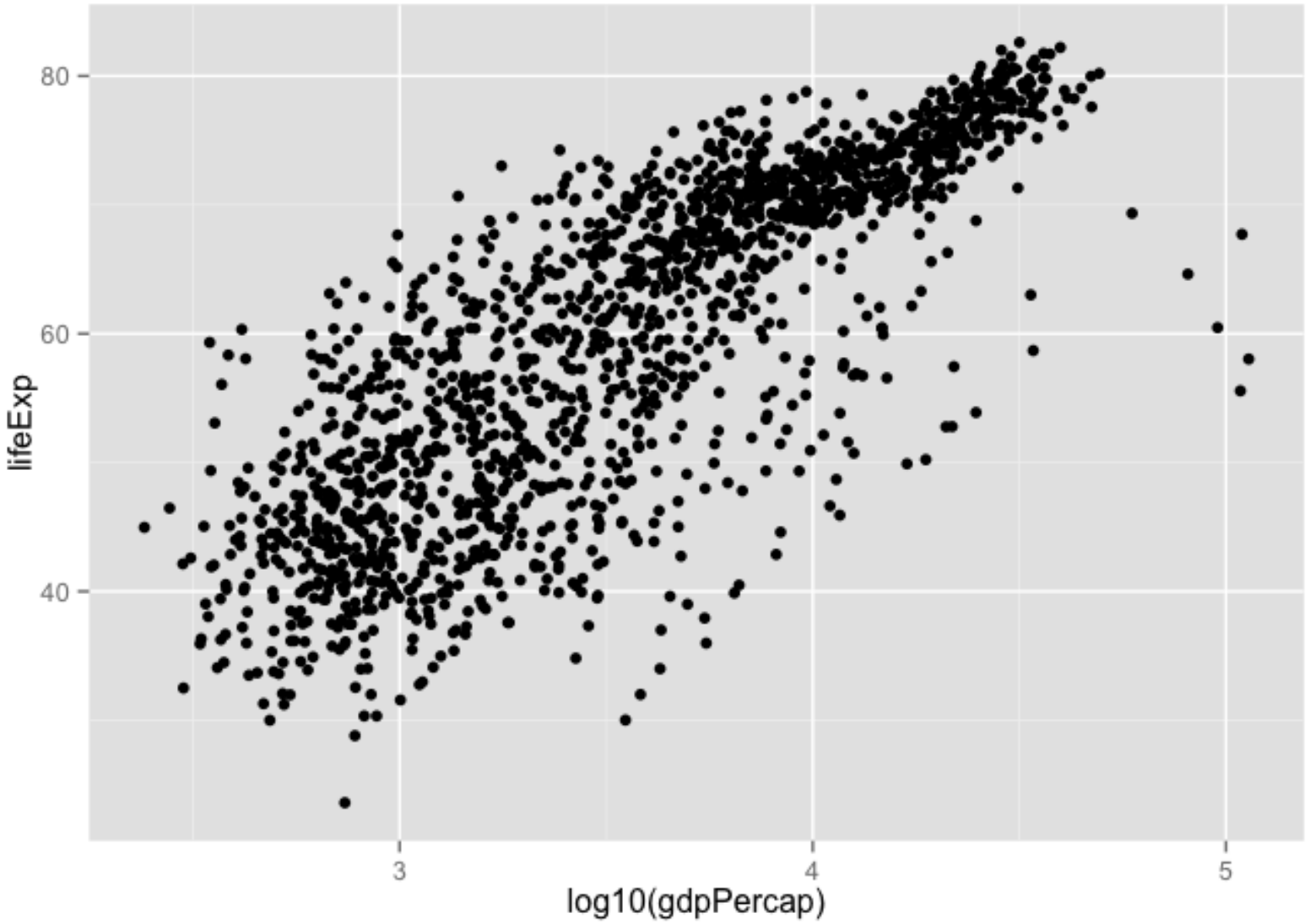
scale: map data values into "computer" values

stat: summarization/transformation of data

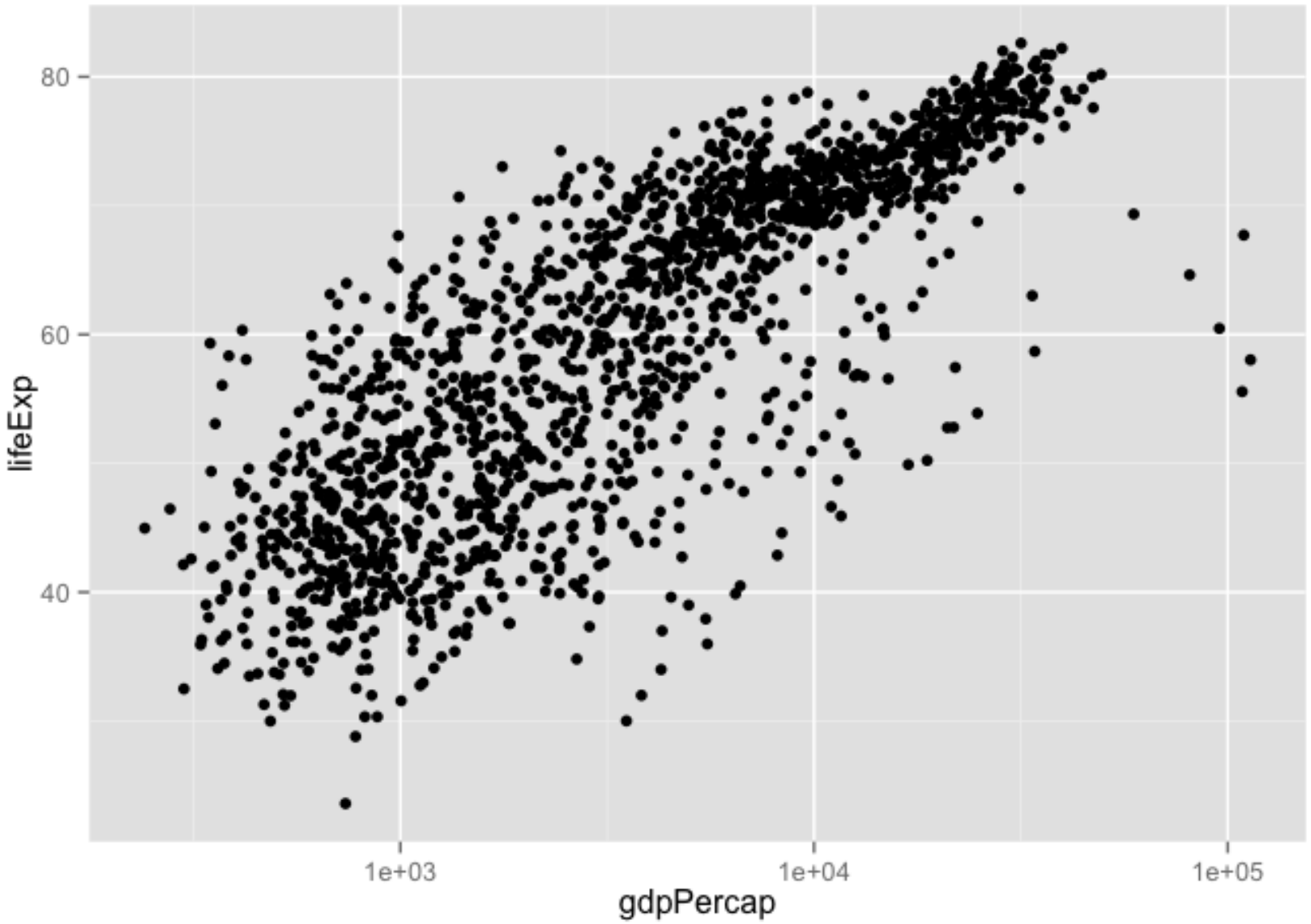
facet: juxtapose related mini-plots of data subsets



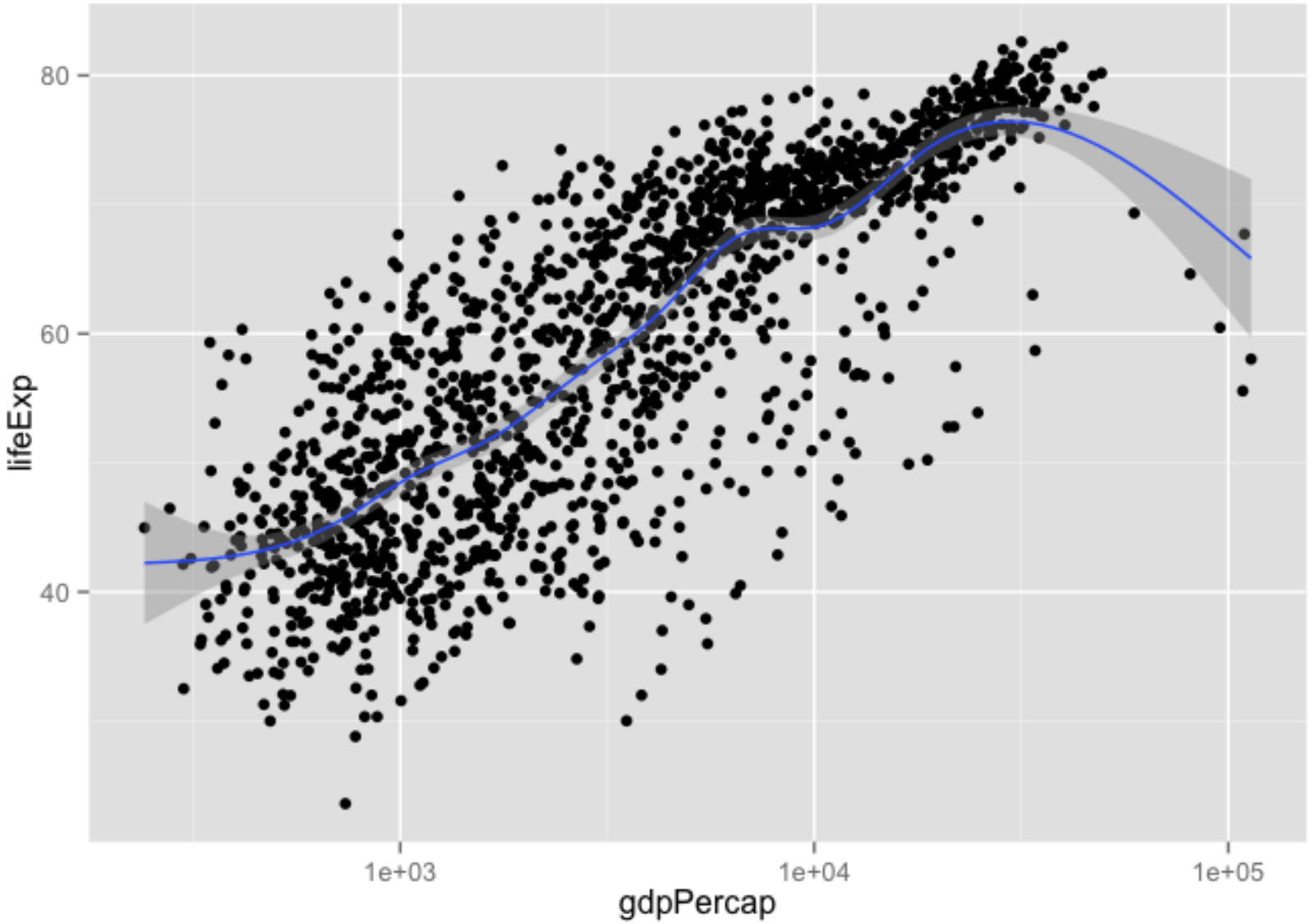

```
p <- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp))
p + geom_point()
```



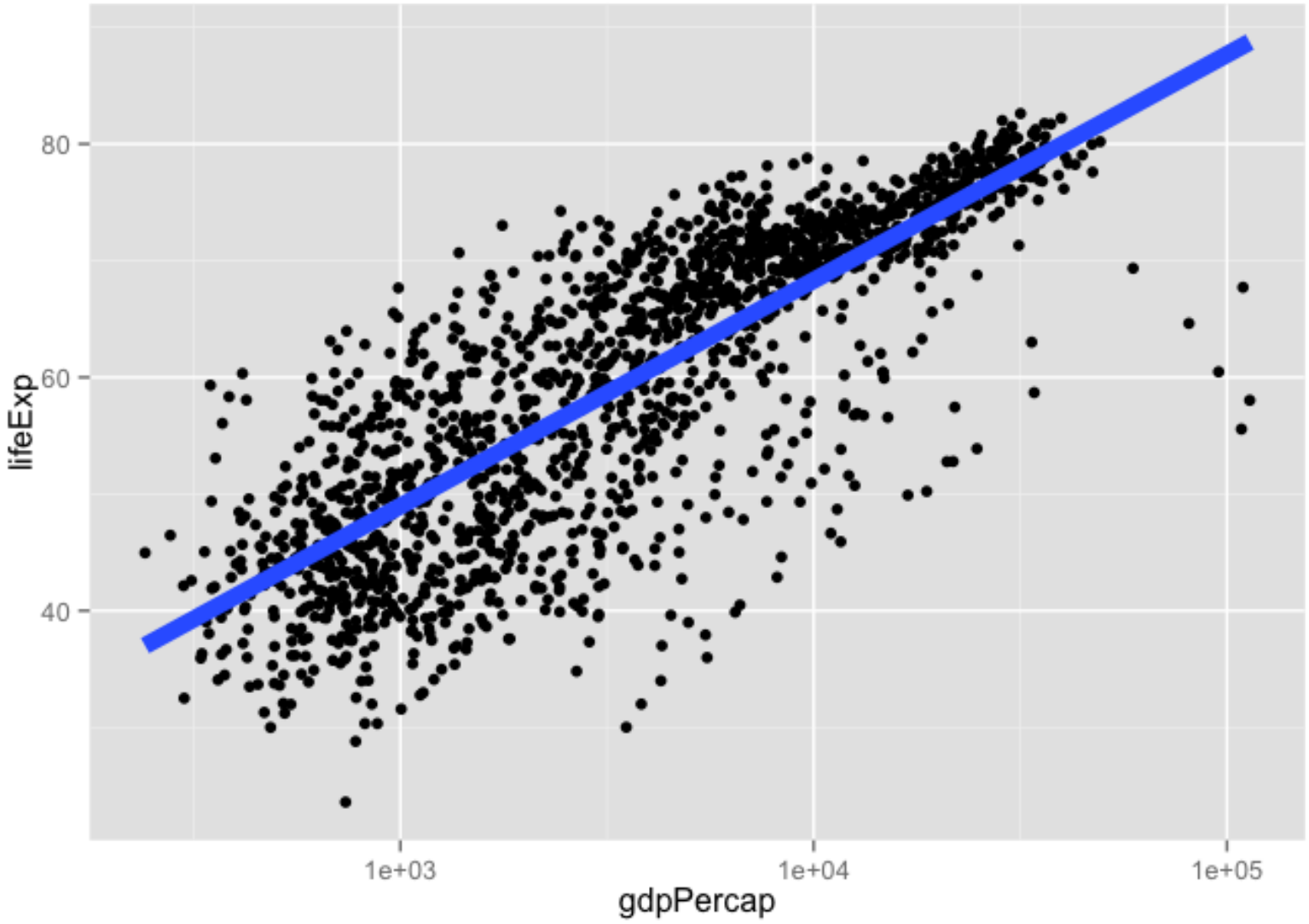

```
ggplot(gapminder, aes(x = log10(gdpPercap), y = lifeExp)) +
geom_point()
```




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

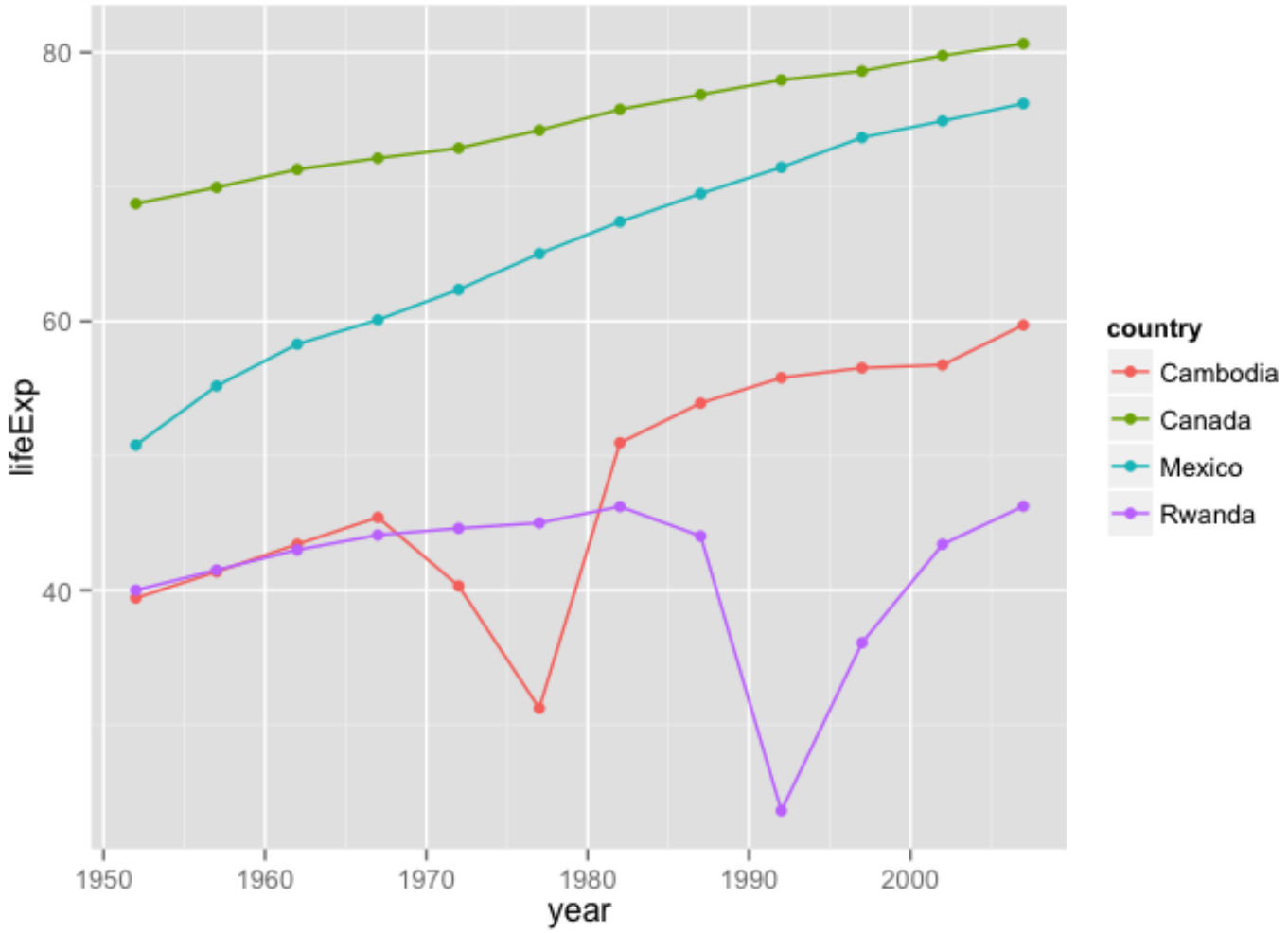



```
p + geom_point() + geom_smooth()
```

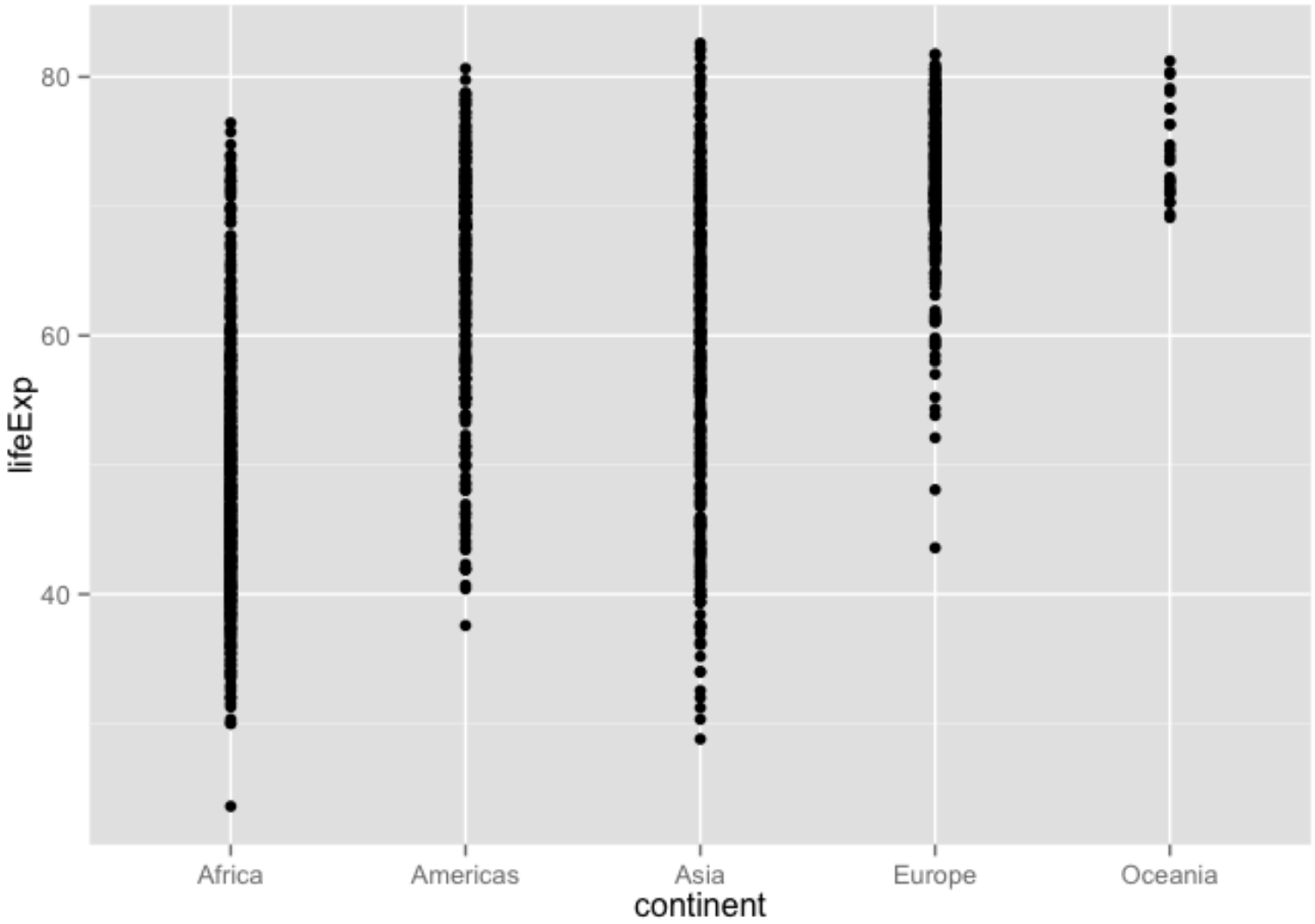



```
p + geom_point() + geom_smooth(lwd = 3, se = FALSE, method
= "lm")
```

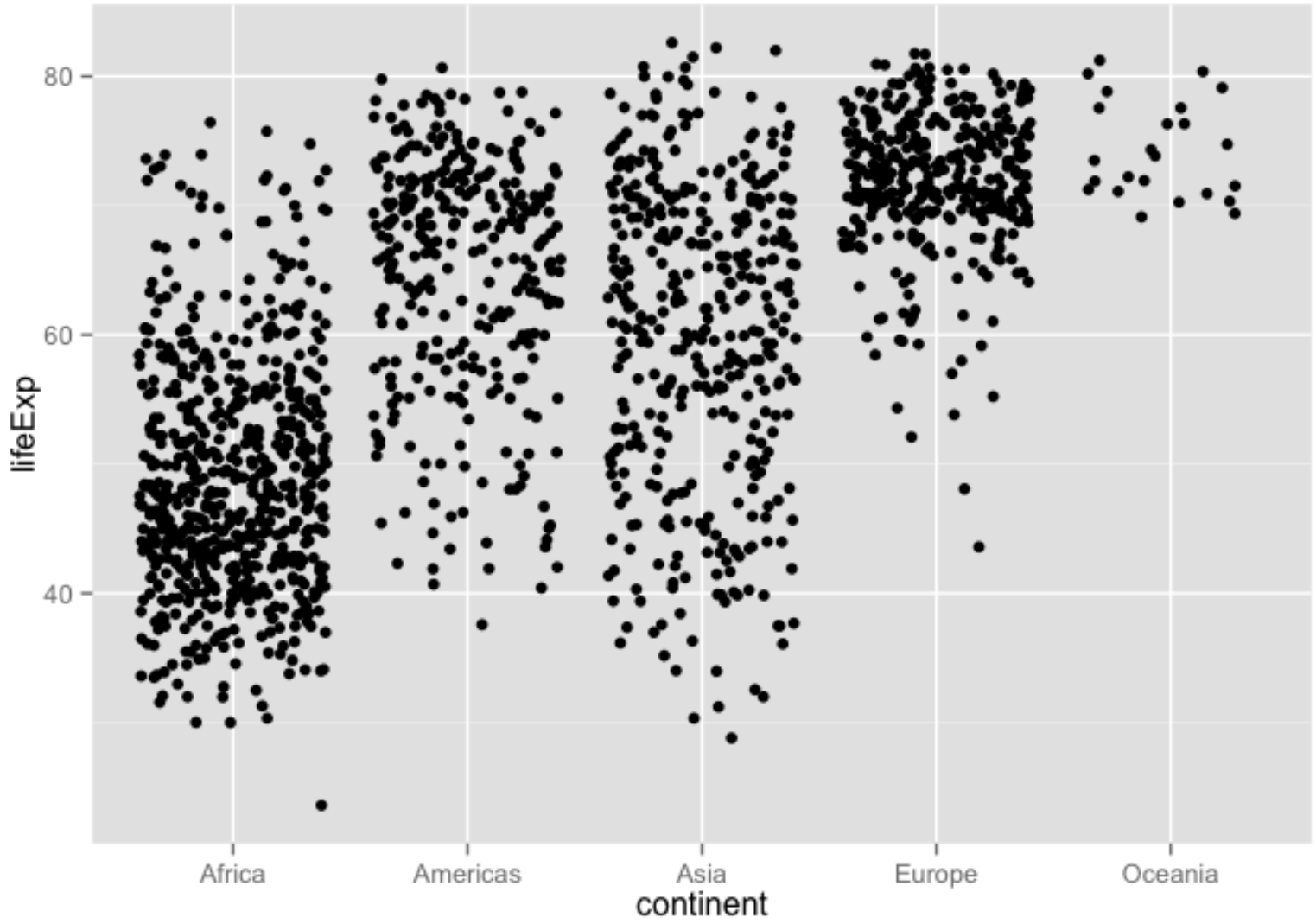

```
ggplot(subset(gapminder, country == "Zimbabwe"),
       aes(x = year, y = lifeExp)) + geom_line() + geom_point()
```



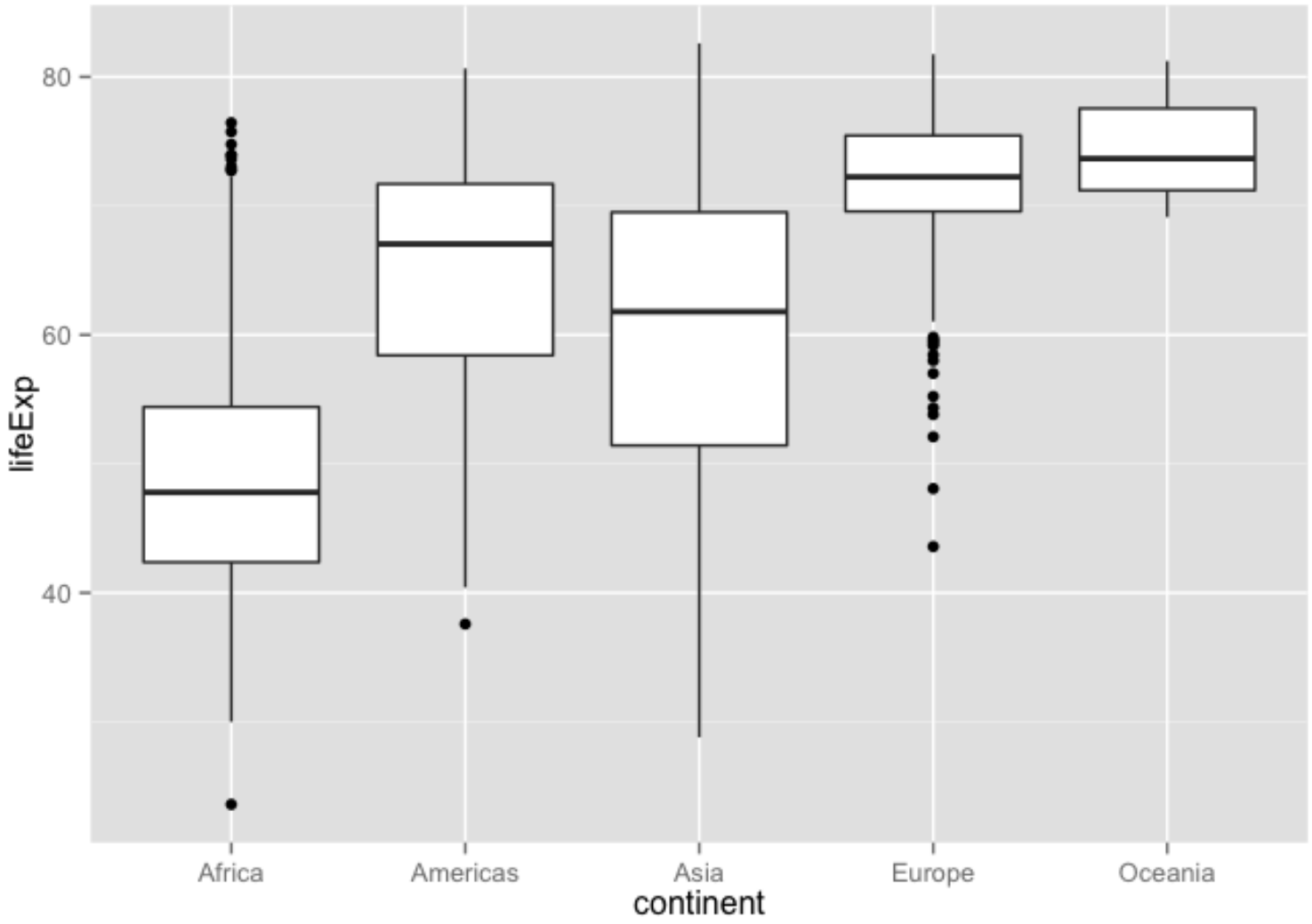

```
jCountries <- c("Canada", "Rwanda", "Cambodia", "Mexico")
ggplot(subset(gapminder, country %in% jCountries),
       aes(x = year, y = lifeExp, color = country)) +
geom_line() + geom_point()
```



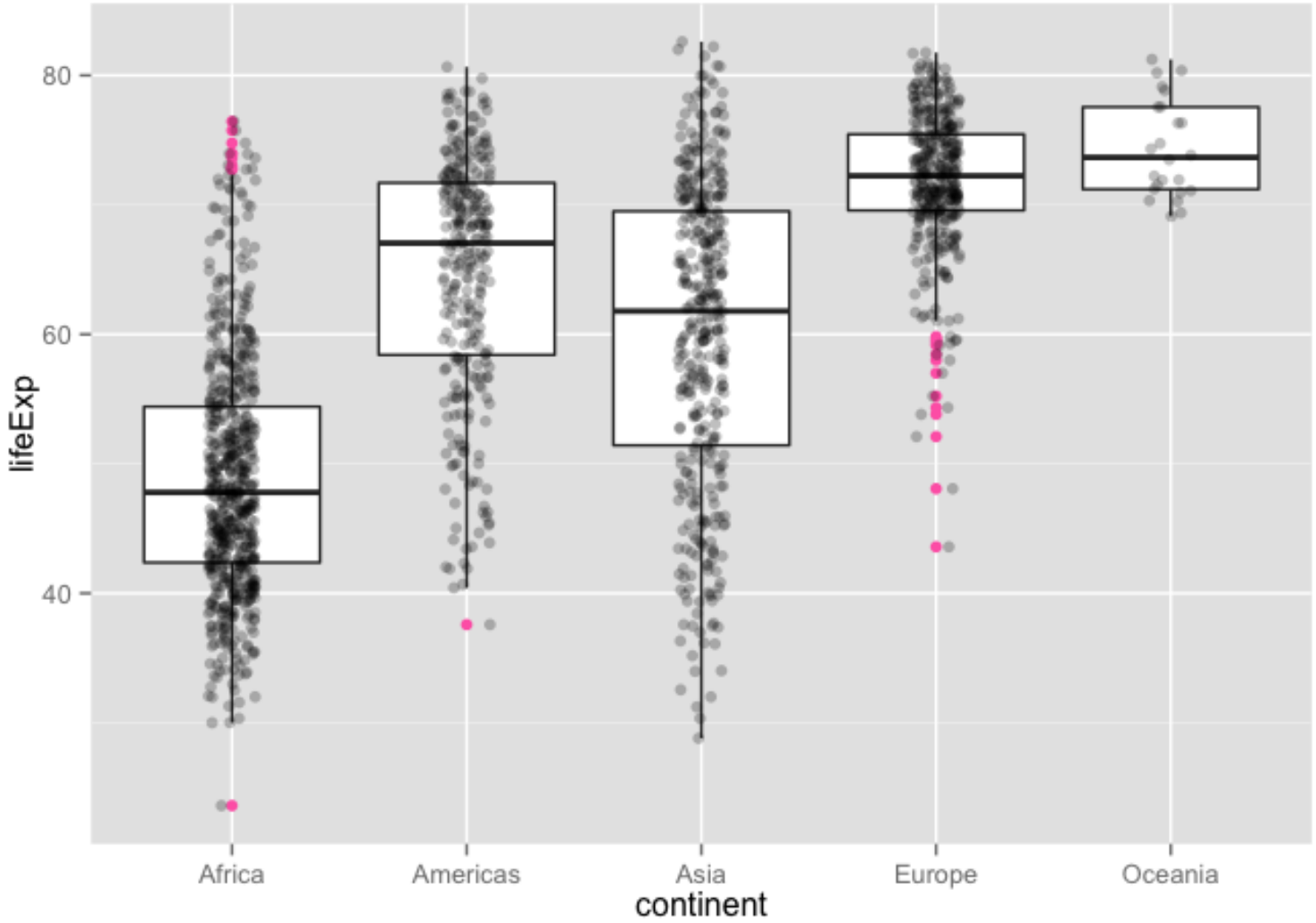

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
geom_point()
```



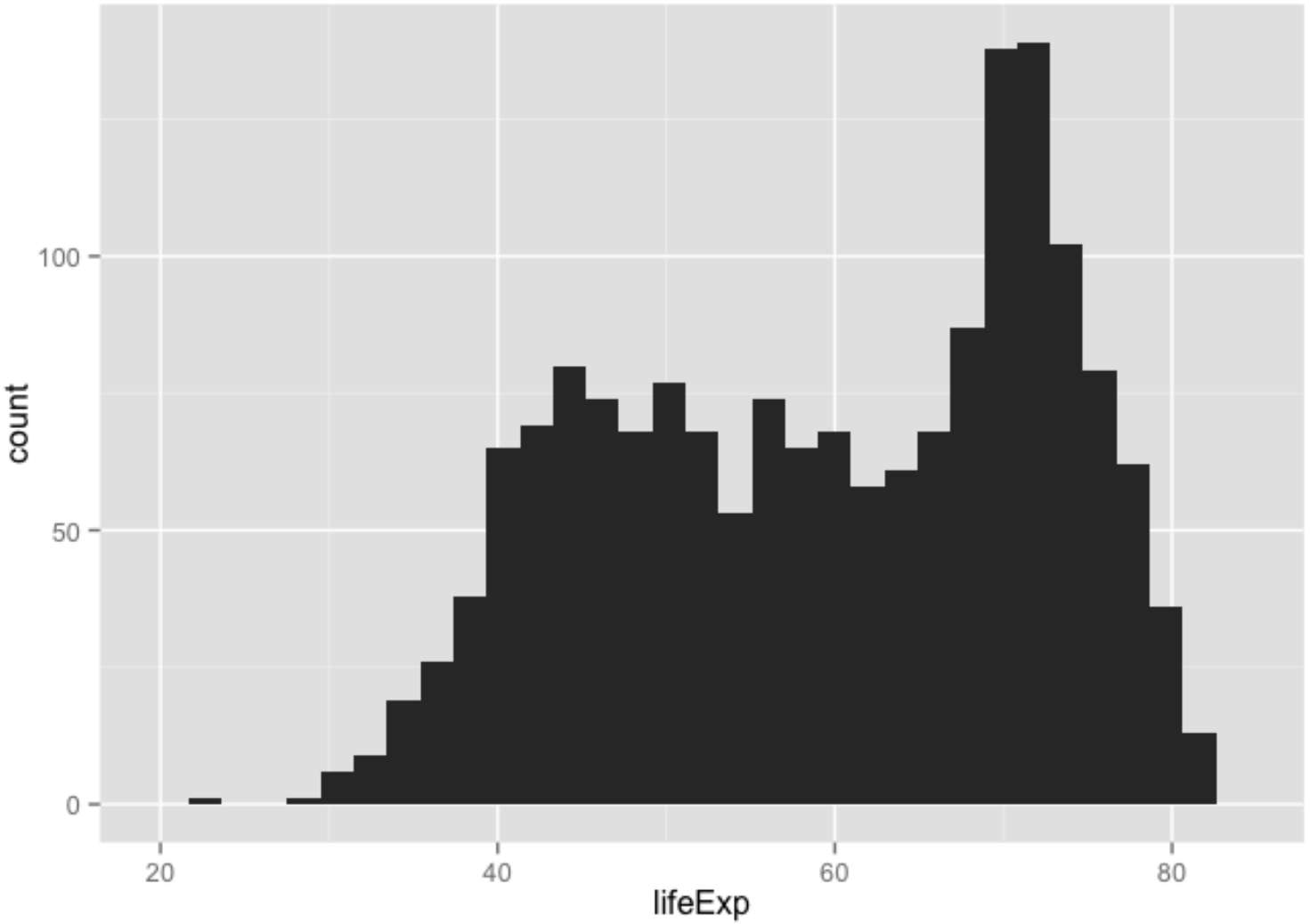

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
geom_jitter()
```



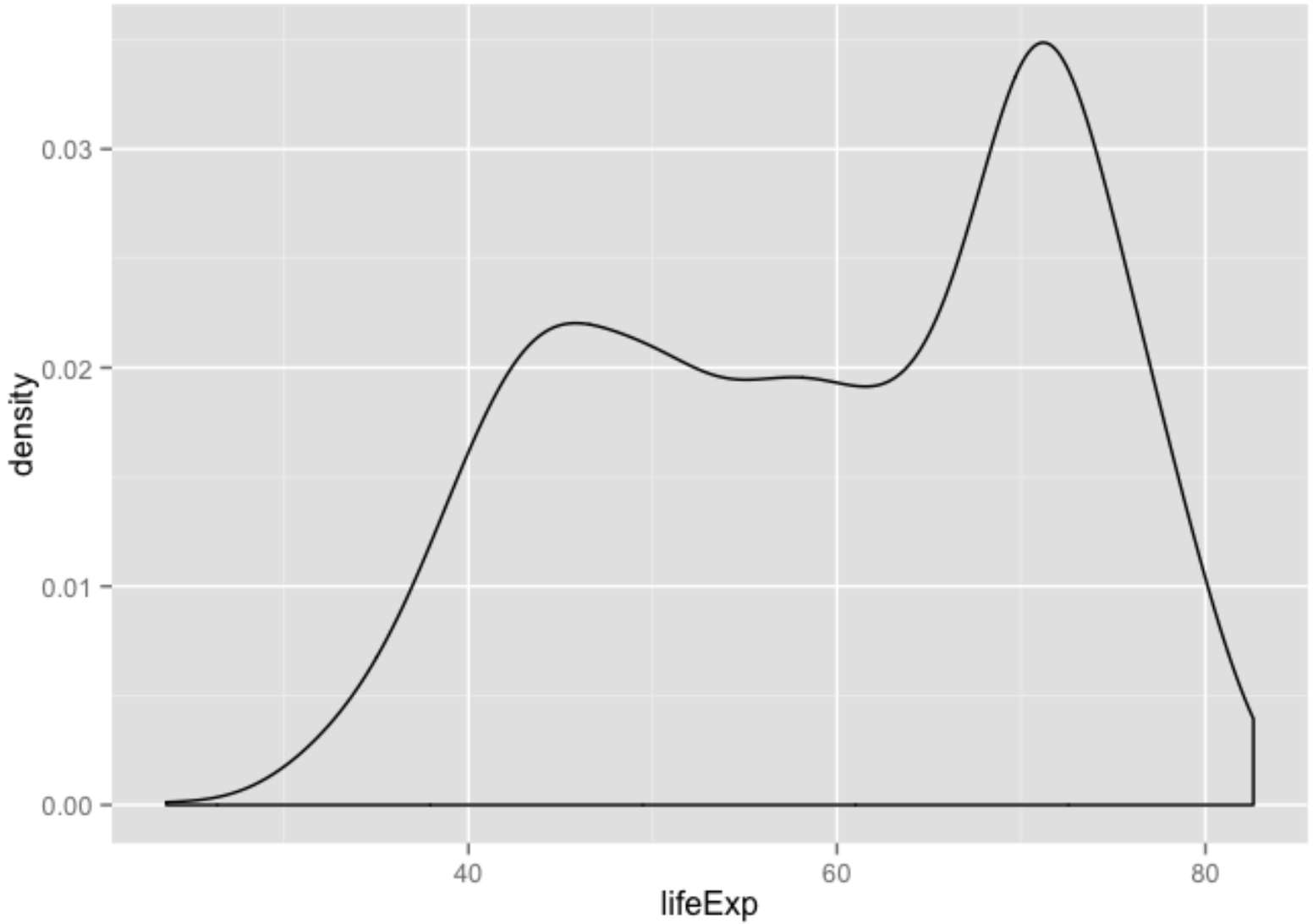

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
geom_boxplot()
```



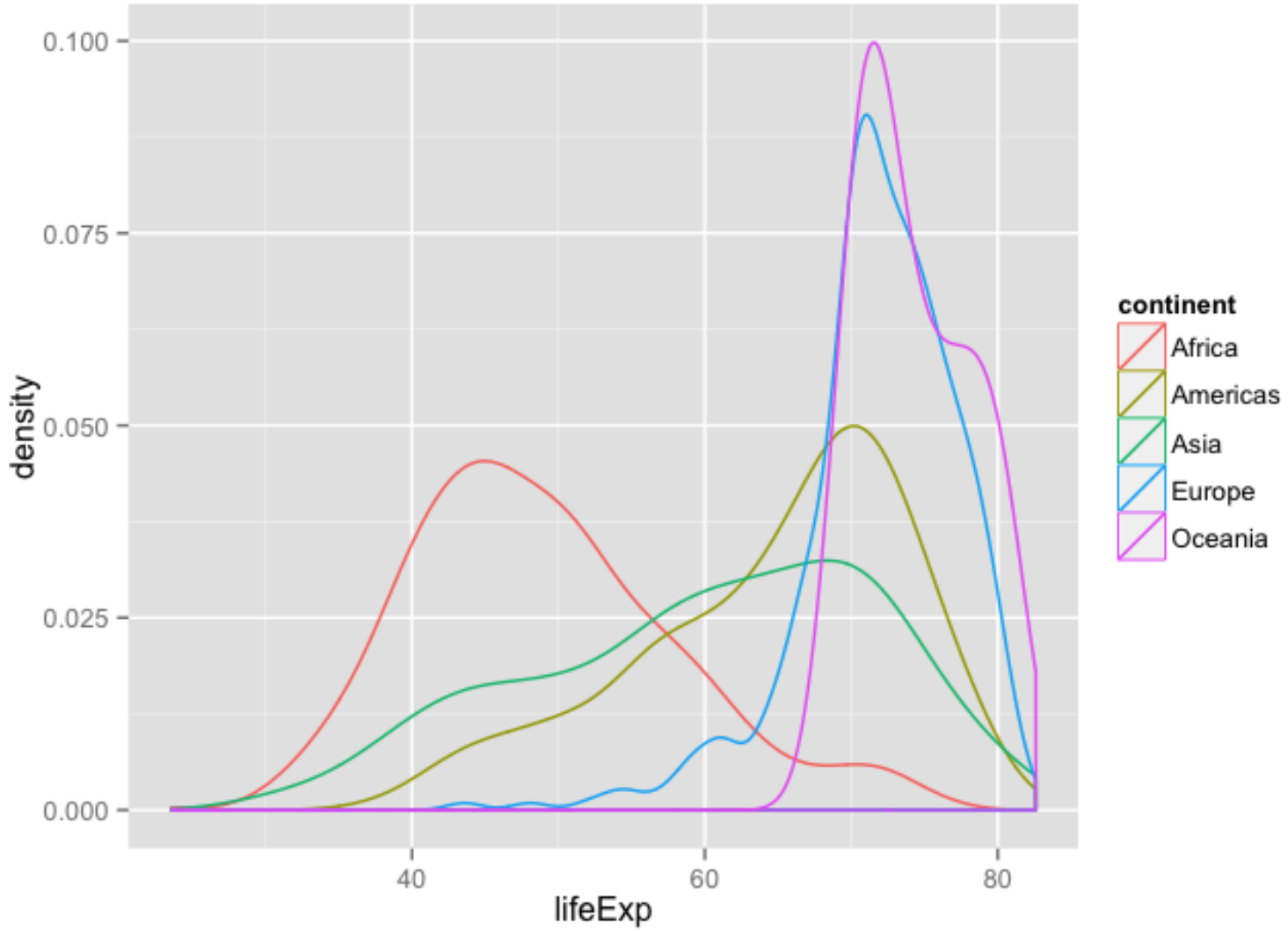

```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +
  geom_boxplot(outlier.colour = "hotpink") +
  geom_jitter(position = position_jitter(width = 0.1, height =
0), alpha = 1/4)
```



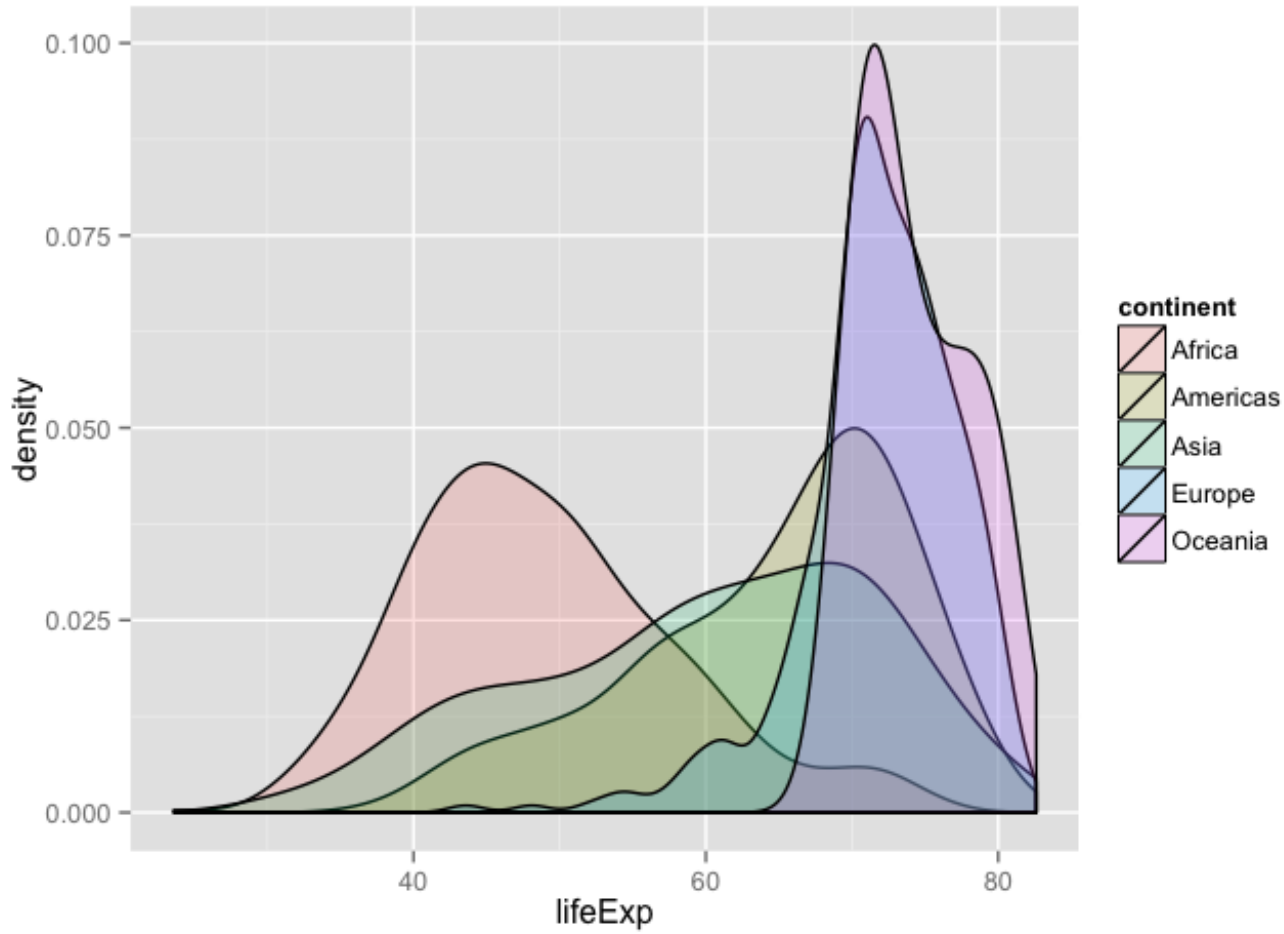

```
ggplot(gapminder, aes(x = lifeExp)) + geom_histogram()
```



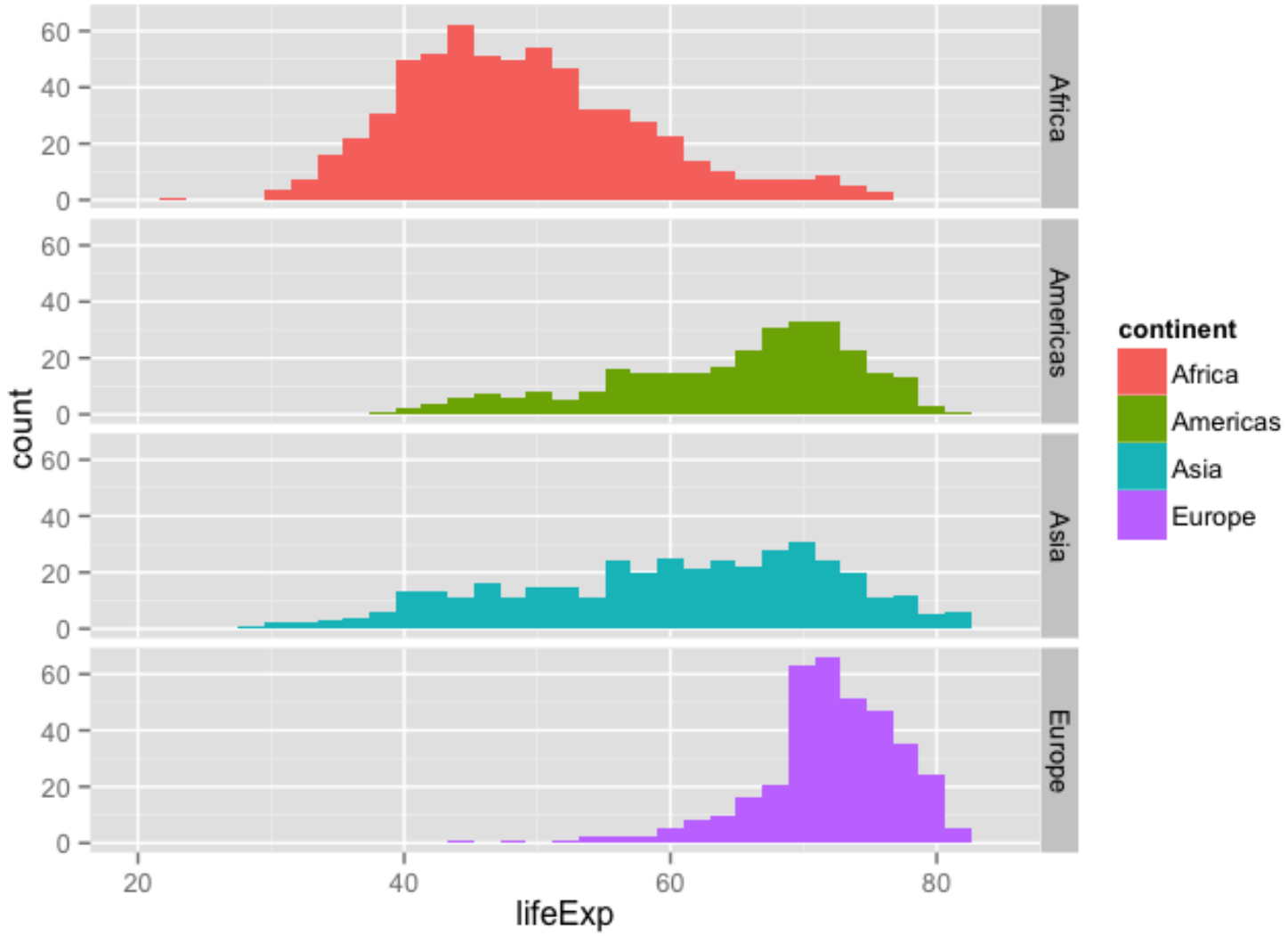

```
ggplot(gapminder, aes(x = lifeExp)) + geom_density()
```



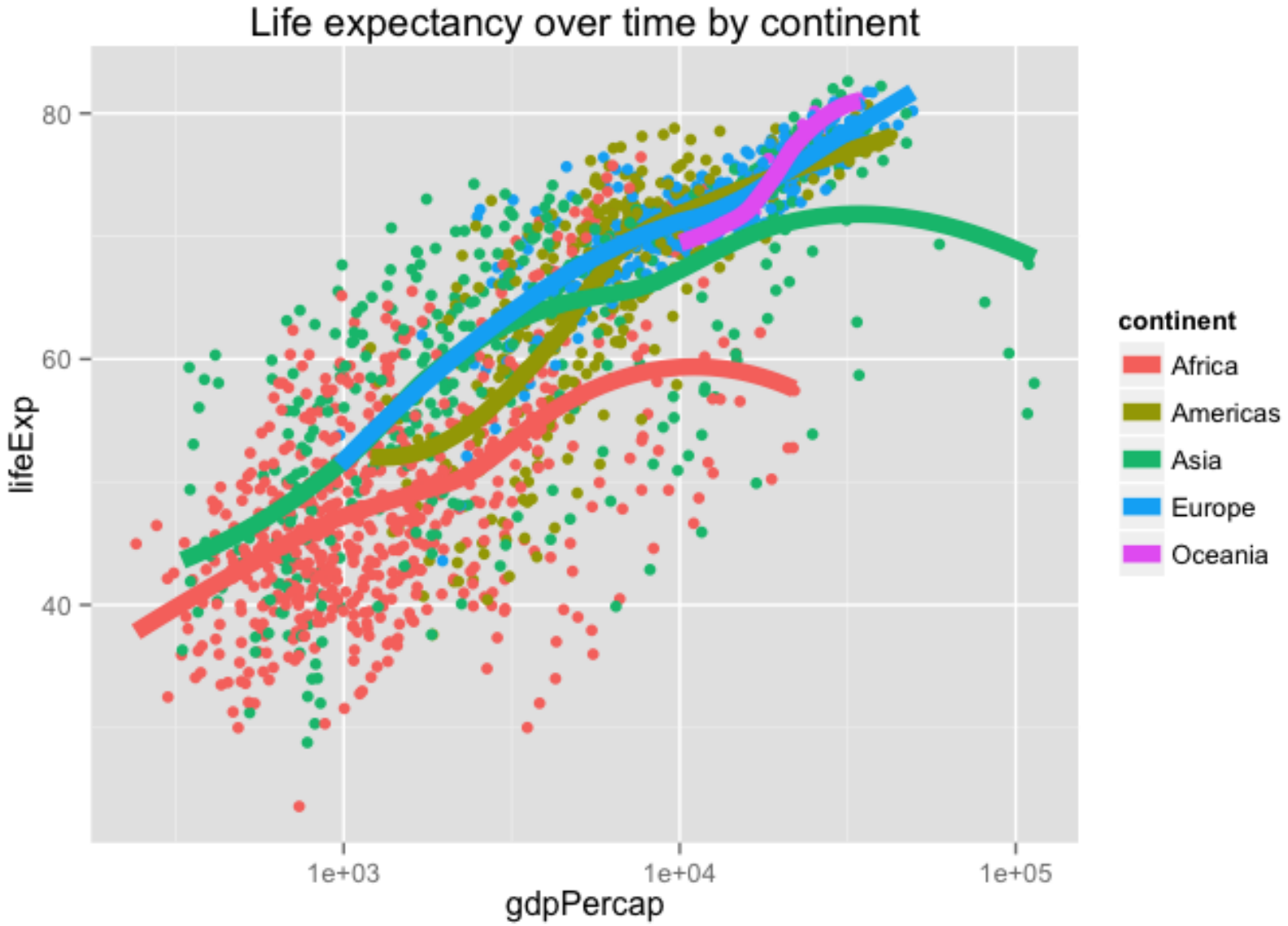

```
ggplot(gapminder, aes(x = lifeExp, color = continent)) +
geom_density()
```



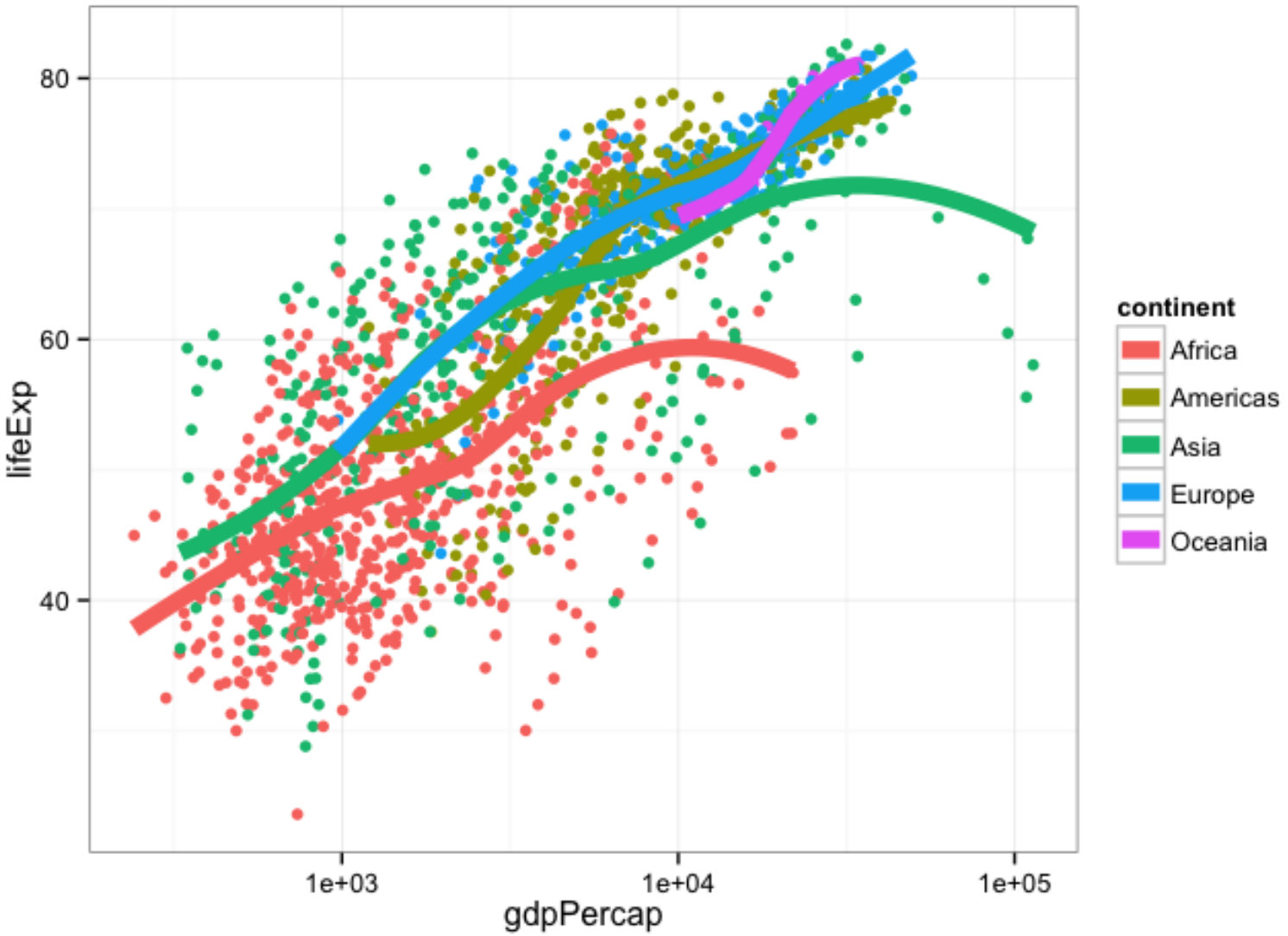

```
ggplot(gapminder, aes(x = lifeExp, fill = continent)) +
  geom_density(alpha = 0.2)
```




```
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()
```

Очистка данных

Источники проблем в данных

- Особенности формата (лишние строки в начале файла, наличие/отсутствие заголовка, нетрадиционные разделители, etc.)
- Отсутствие некоторых данных (na)
- Типы данных (перевод строк в числа и т.п.)
- Выбросы, которые искажают общий тренд

Данные про жилье

```
> install.packages("gdata")
> require(gdata)

>bk <-
read.xls("rollingsales_brooklyn.xls",pattern="BOROUGH")
#все что до строки, содержащей , "BOROUGH", не читаем


head(bk) #смотрим на данные
summary(bk) #сводная статистика, чего сколько
```

Чистим данные

```
head(bk$SALE.PRICE)
[1] $403,572 $218,010 $952,311 $842,692 $815,288 $815,288
3318
Levels: $0 $1 $10 $100 $1,000 $10,000 $100,000 $1,000,000 ...
$999,999
```

Переводим цены в числовой формат

```
>bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", bk
$SALE.PRICE))
# убираем все кроме цифр, т.е. заменяем все кроме цифр
на “”
```

Чистим данные

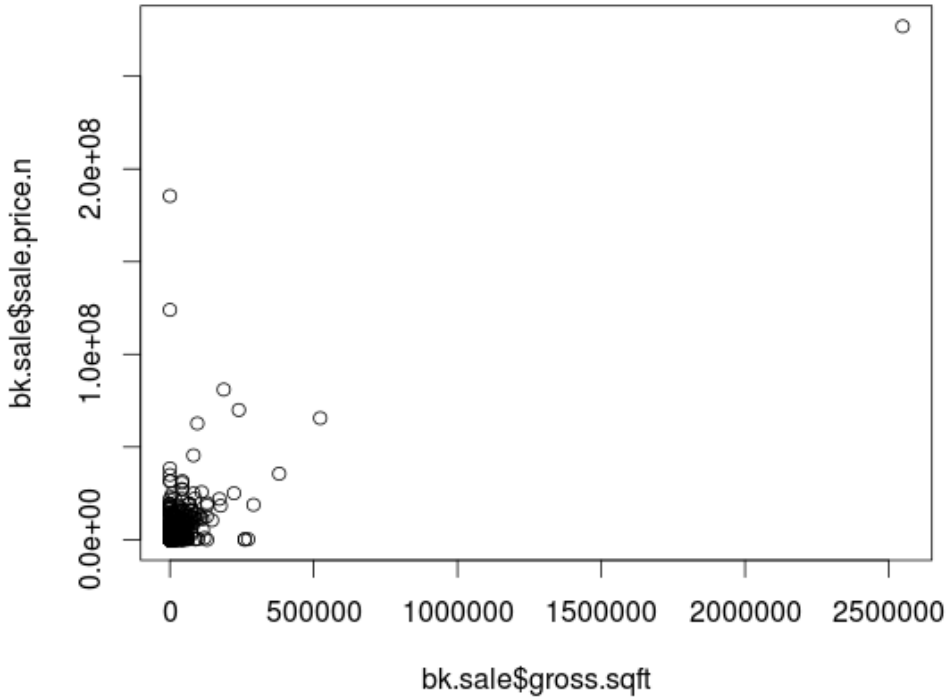
Смотрим, для скольки объектов у нас нет данных про цены
>count(is.na(bk$SALE.PRICE.N))#sum
Сделаем все имена столбцов маленькими буквами
>names(bk) <- tolower(names(bk))
Приведем в порядок площади
>bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", bk $gross.square.feet))
>bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","", bk $land.square.feet))
Приведем в порядок даты
>bk$sale.date <- as.Date(bk$sale.date)
>bk$year.built <- as.numeric(as.character(bk$year.built))

Надоело писать длинные имена?
Работаем с одной таблицей?
Нет проблем!

```
>attach(bk)#теперь по-умолчанию работаем только с bk
>hist(sale.price.n)#обращаемся прямо по имени поля
>hist(sale.price.n[sale.price.n>0])
>hist(gross.sqft[sale.price.n==0])
>detach(bk)#закончили работать, открепляемся!
```

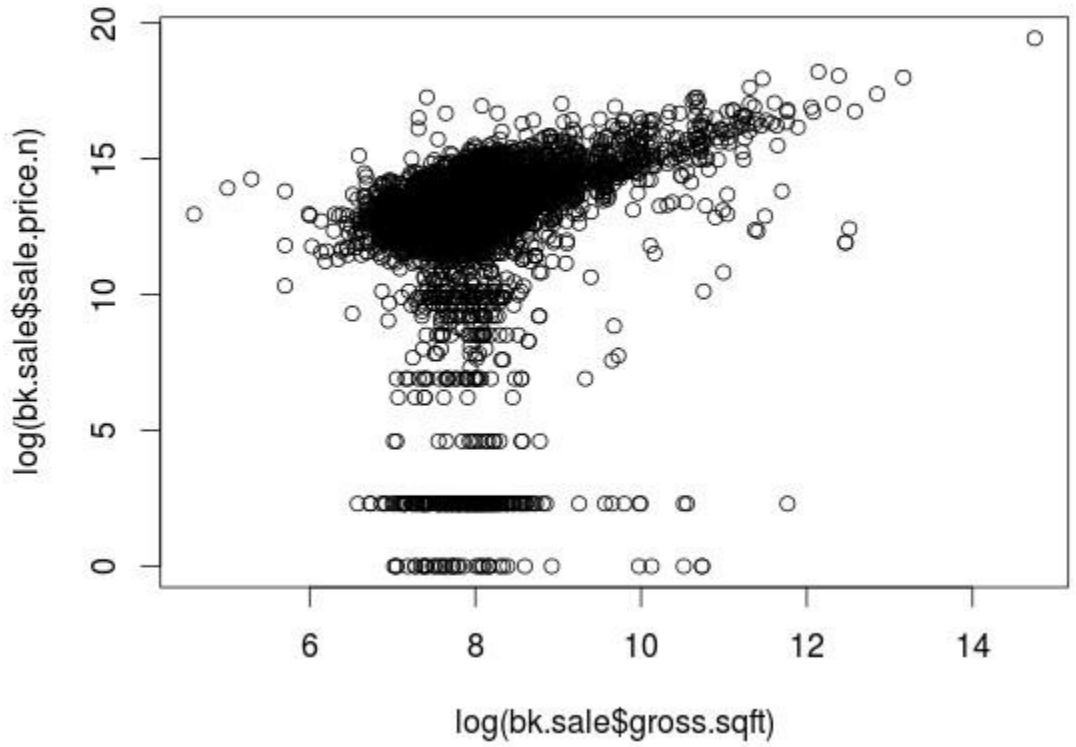
Теперь беглый анализ, как устроены данные

```
>bk.sale <- bk[bk$sale.price.n!=0,]
>plot(bk.sale$gross.sqft,bk.sale
$sale.price.n)
```



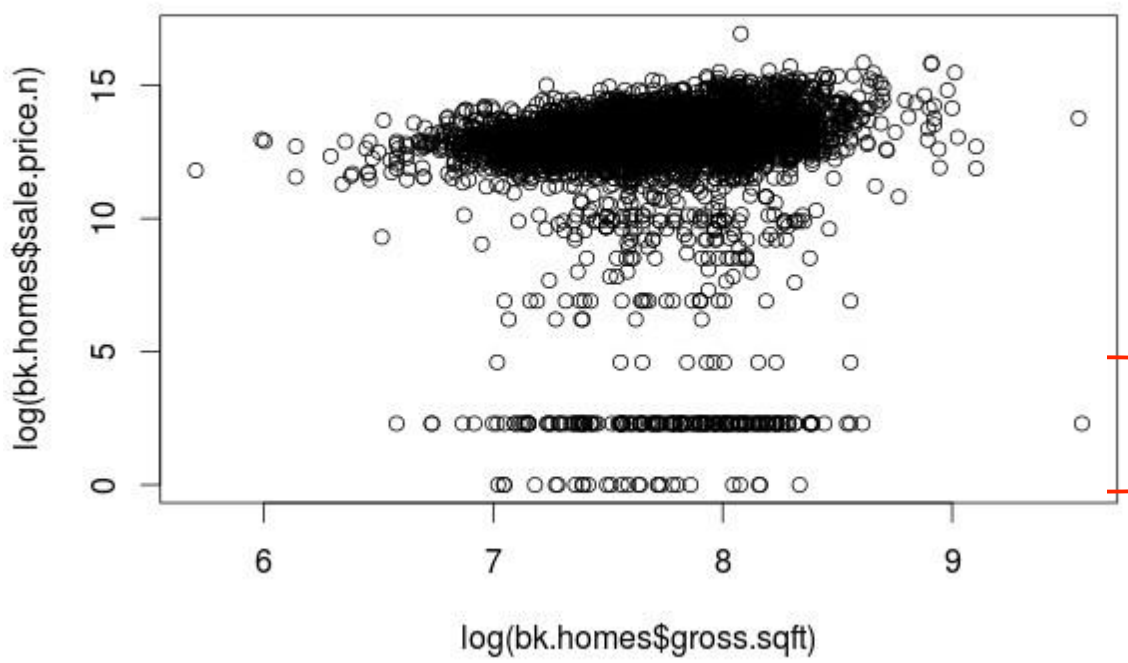
“Вынем” данные из нуля

- plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n))



Выберем для анализа только дома (категория содержит в названии “FAMILY”)

```
>bk.homes <- bk.sale[which(grepl("FAMILY", bk.sale
$building.class.category)),]
>plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
```



Похоже, эти дома какие-то особенные

Уберем “особенные” дома

```
>bk.homes$outliers <- (log(bk.homes$sale.price.n) <=5) + 0
>bk.homes <- bk.homes[which(bk.homes$outliers==0),]
>plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
```