

Язык R

лекция 6

Артем Артемов

Елена Ставровская

Анастасия Жарикова

13 октября 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

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.

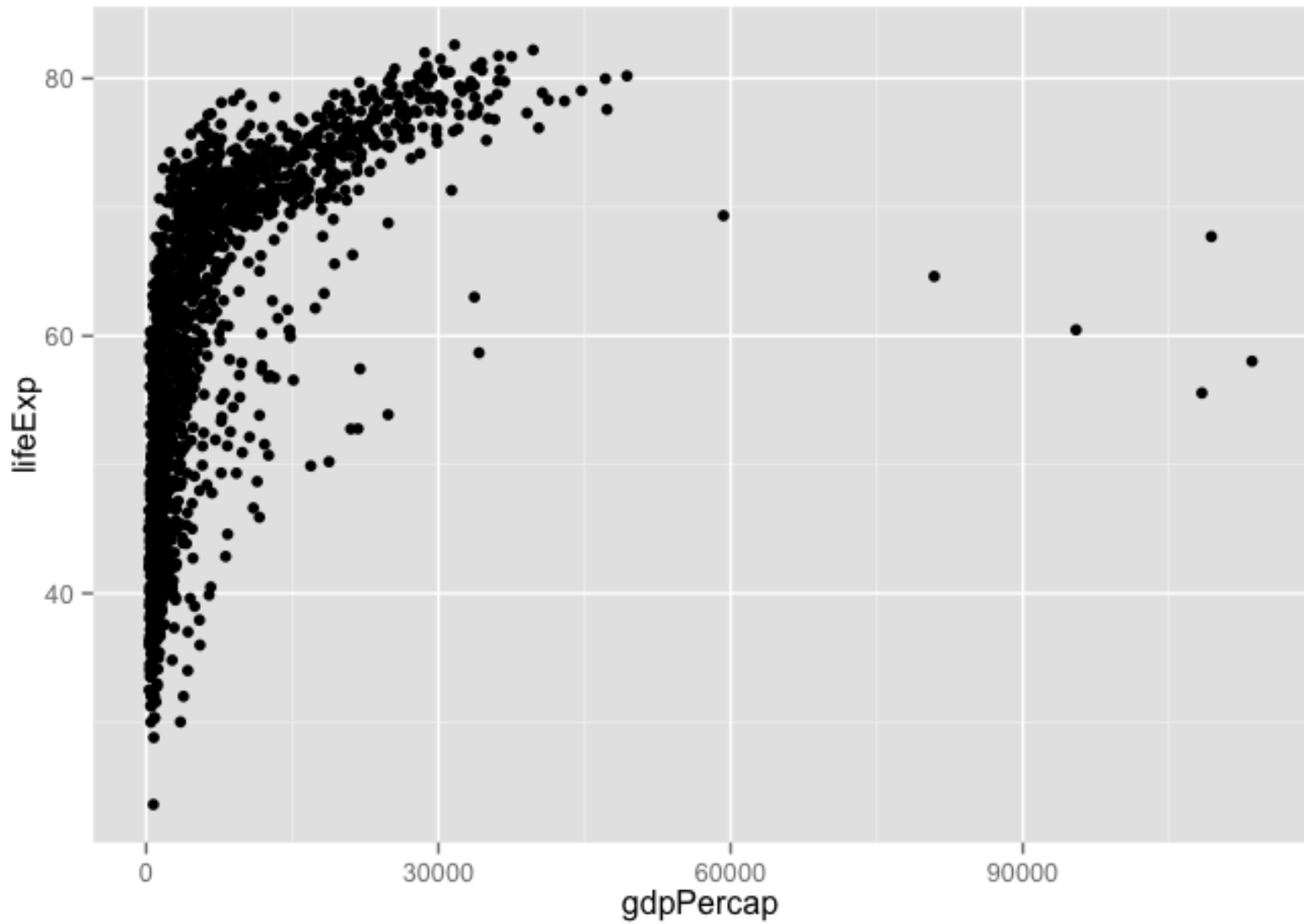
1999 Bachelor of Human Biology, First Class Honours, The University of Auckland, Auckland, New Zealand.

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

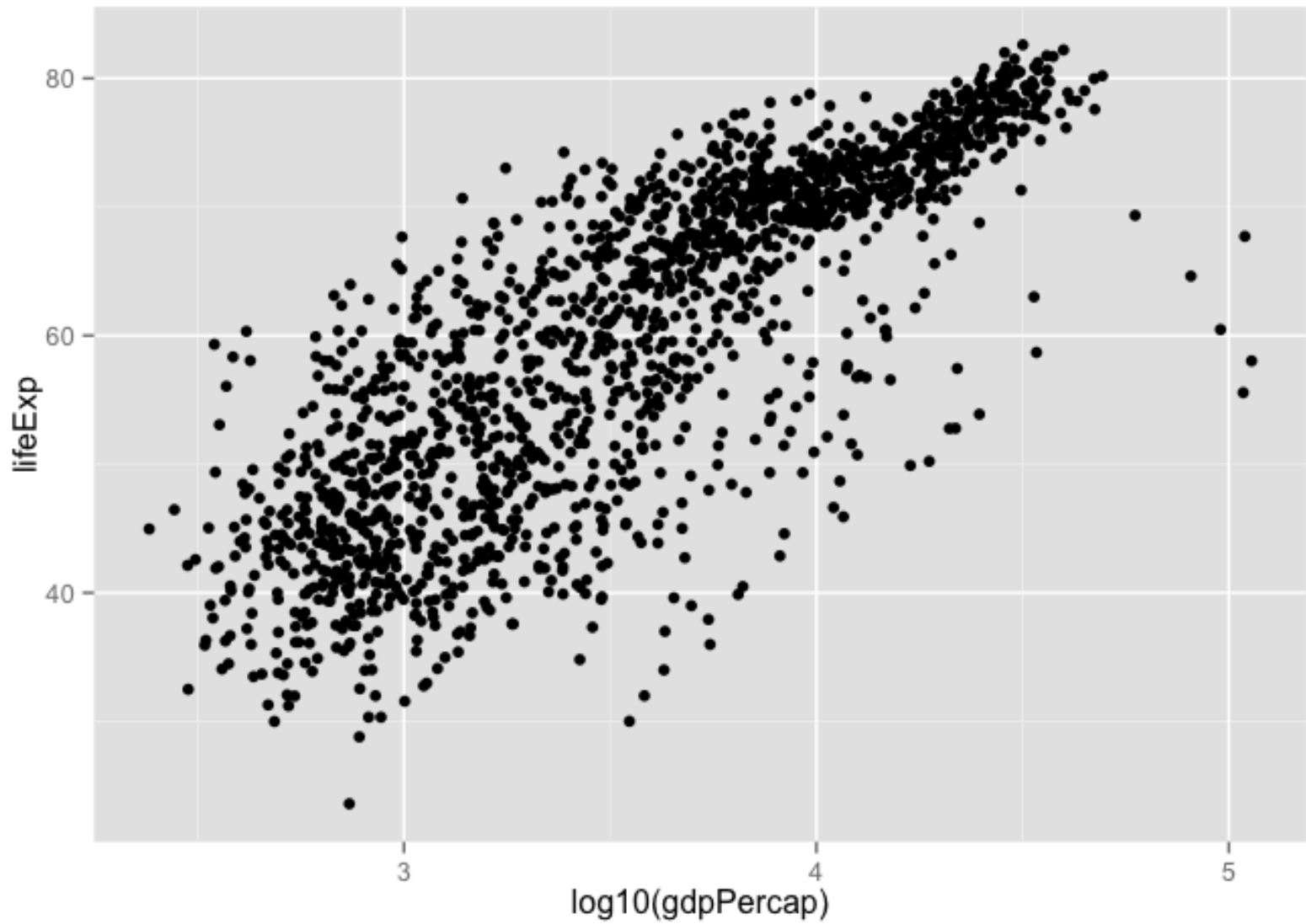
Установка и загрузка пакета:

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

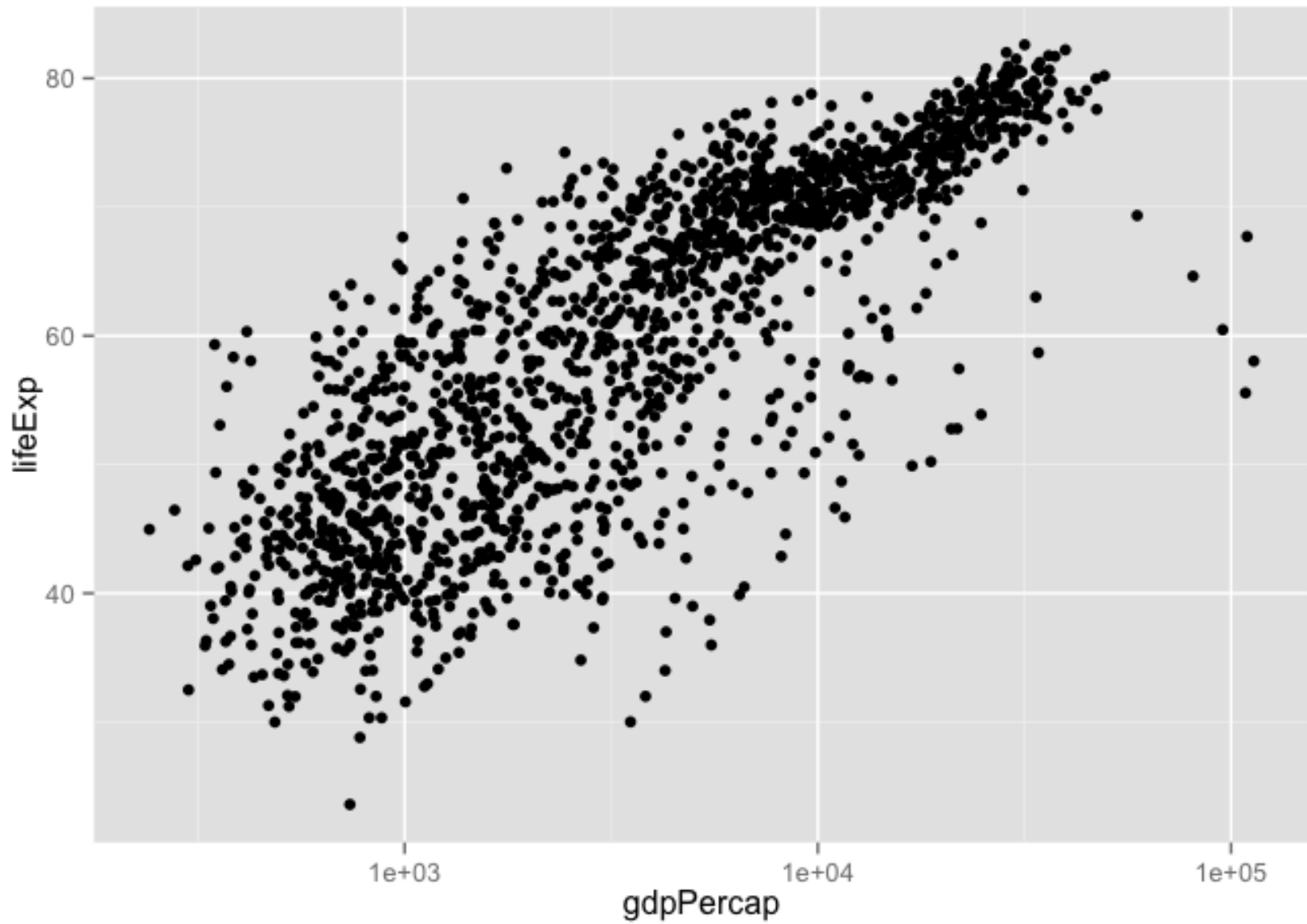
```
> library("ggplot2")
```



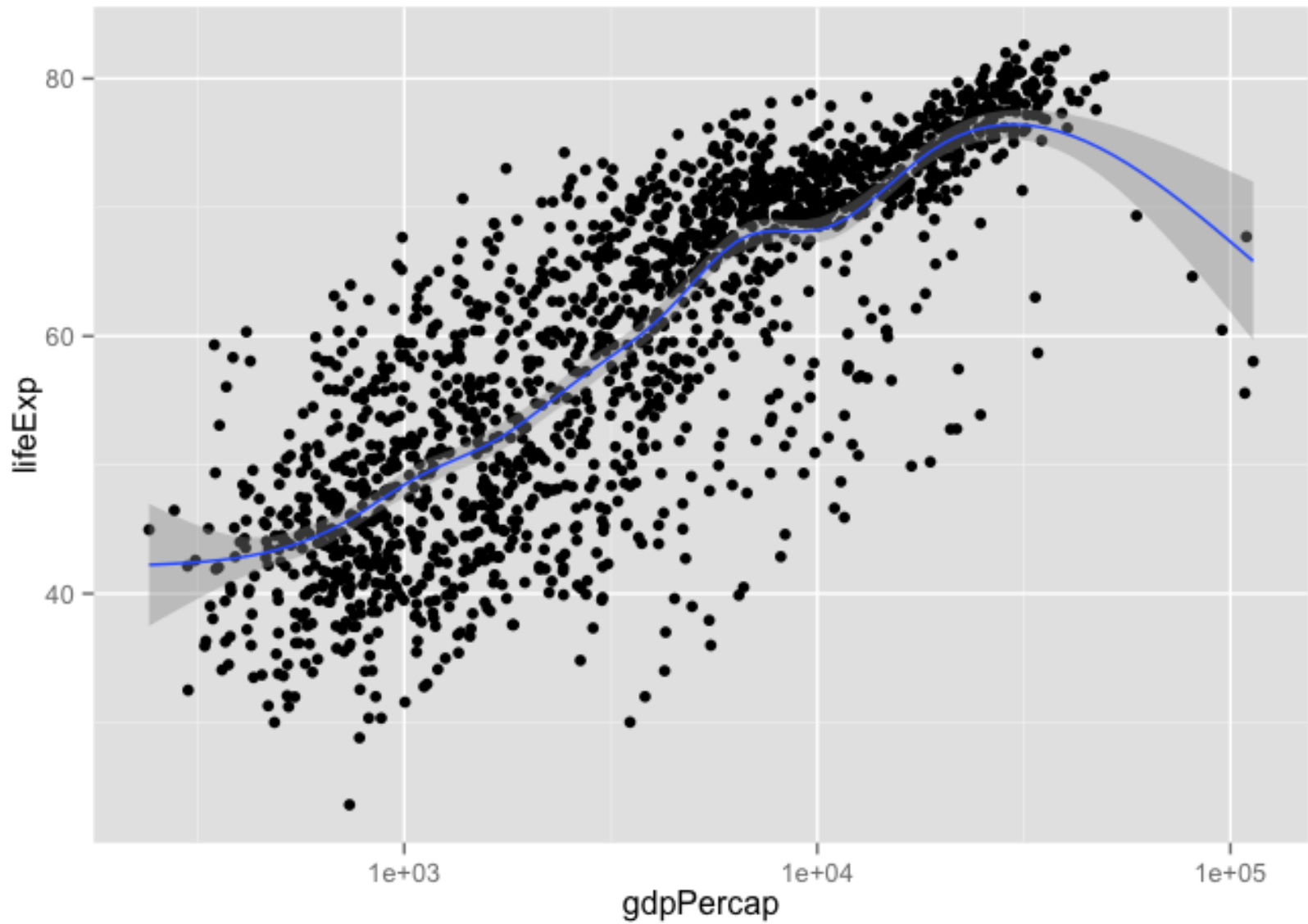
```
p <- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp))  
p + geom_point()
```

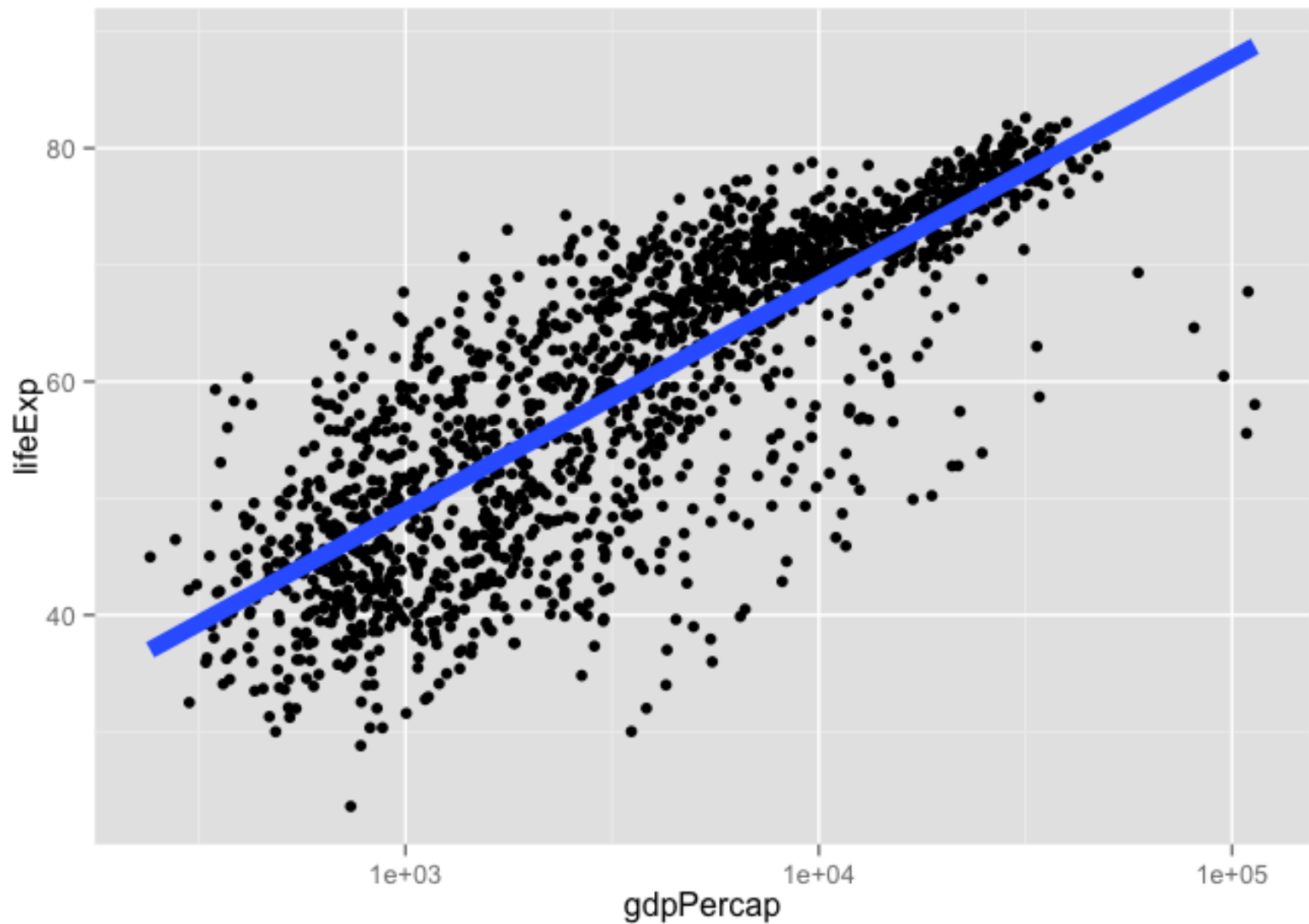
```
ggplot(gapminder, aes(x = log10(gdpPerCap), y = lifeExp)) +  
geom_point()
```



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



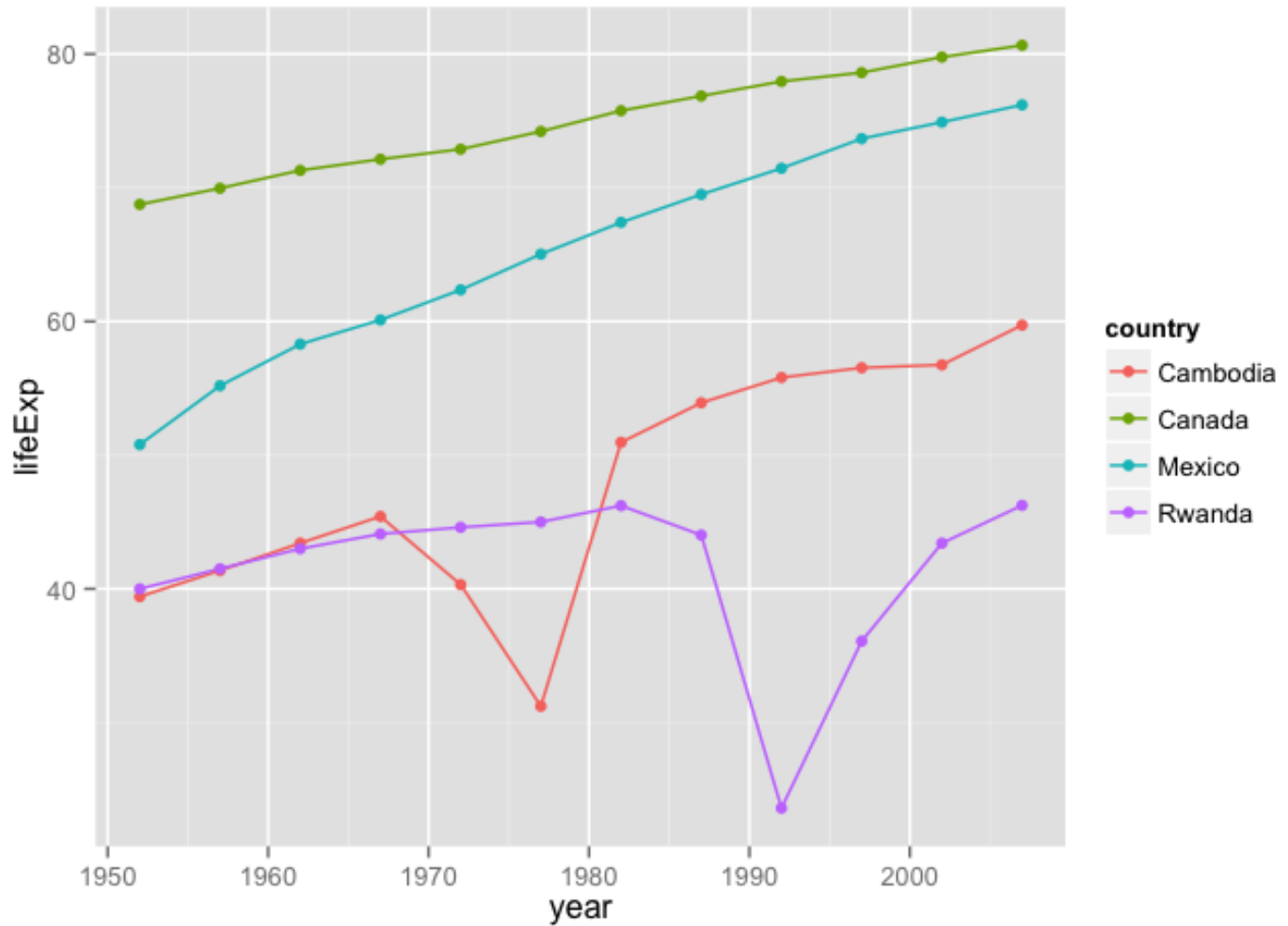
```
p + geom_point() + scale_x_log10() + geom_smooth()
```



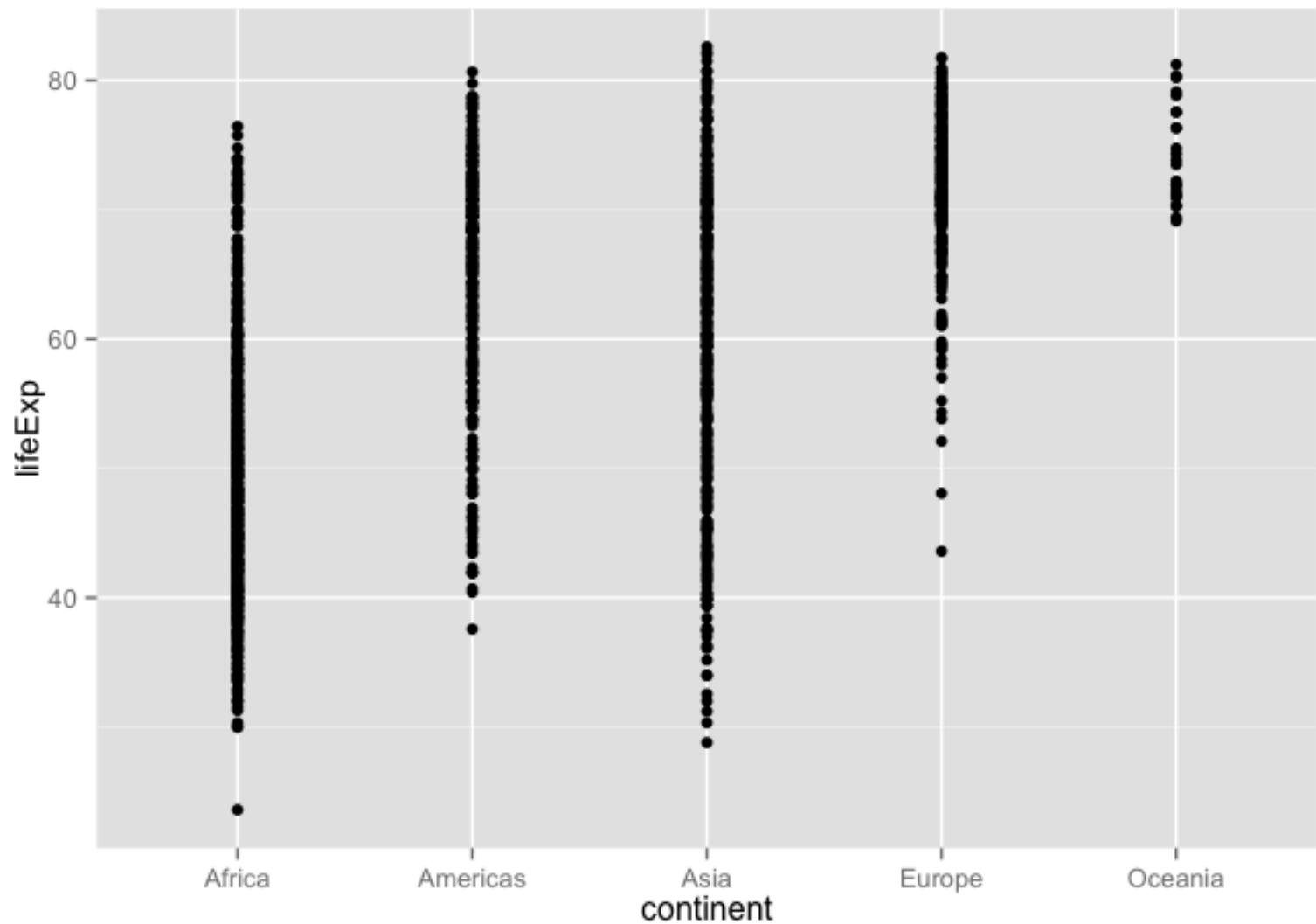
```
p + geom_point() + scale_x_log10() + geom_smooth(lwd = 3,  
se = FALSE, method = "lm")
```



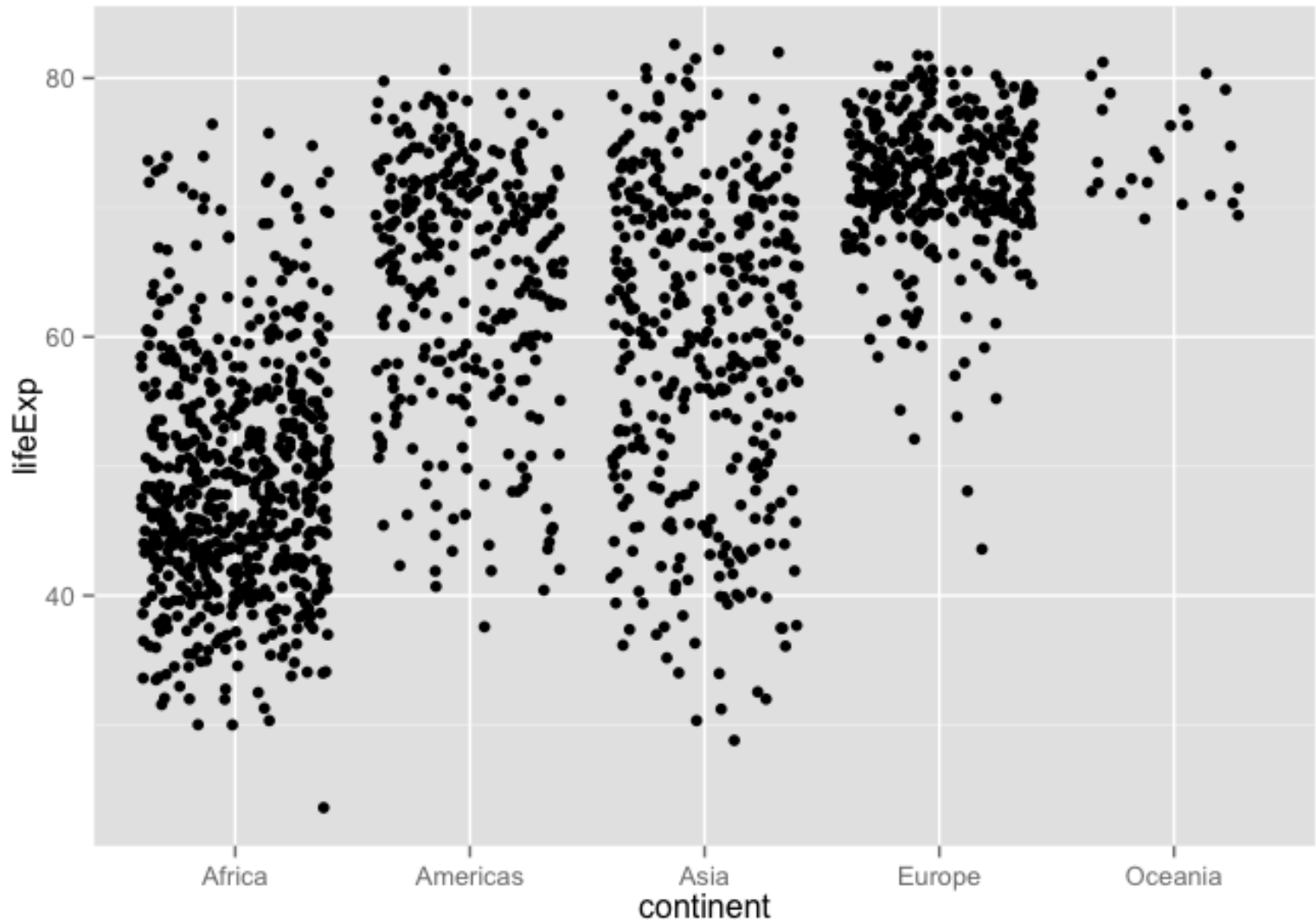
```
ggplot(subset(gapminder, country == "Zimbabwe"),  
  aes(x = year, y = lifeExp)) + geom_line() + geom_point()
```



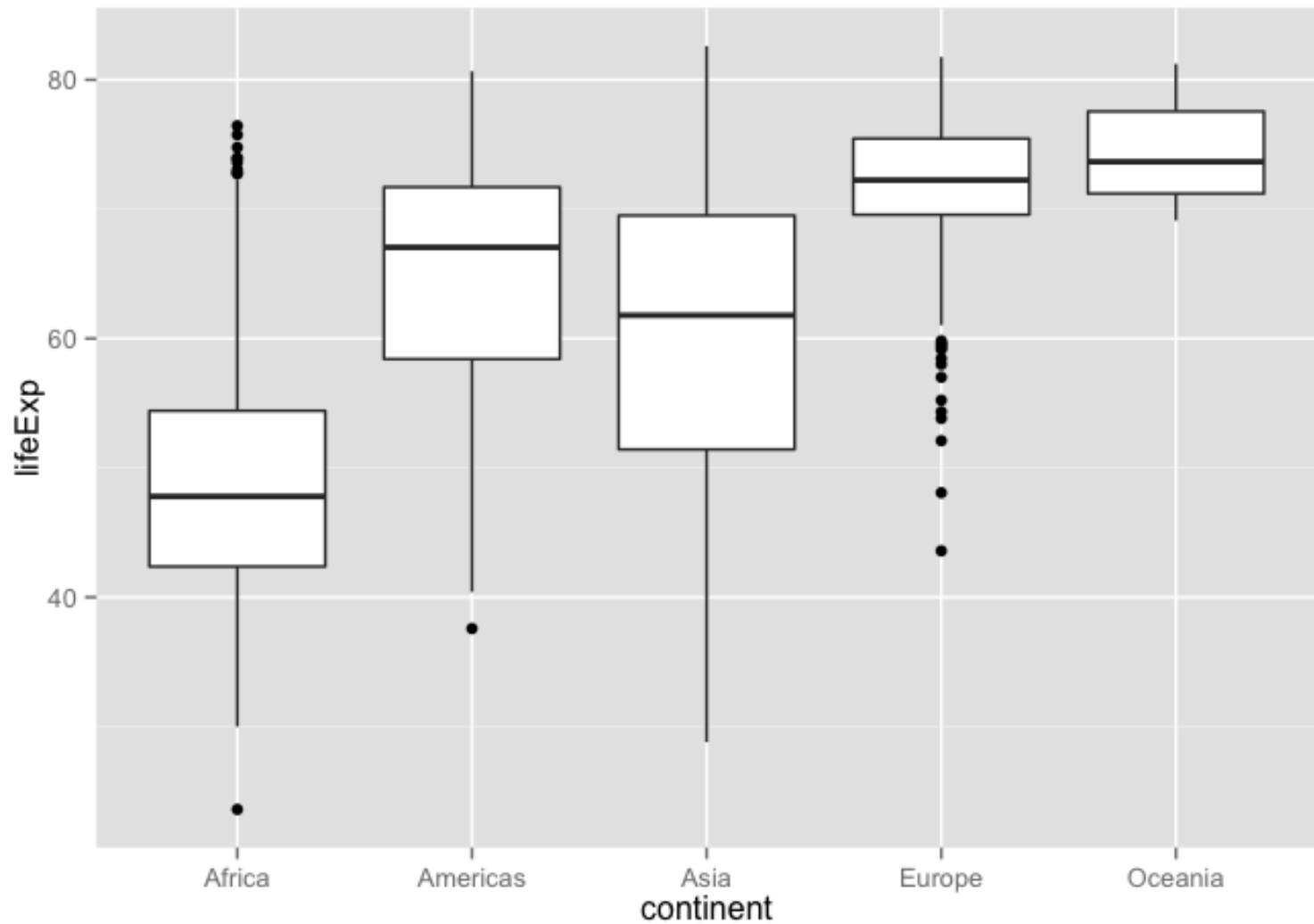
```
jCountries <- c("Canada", "Rwanda", "Cambodia", "Mexico")
ggplot(subset(gapminder, country %in% jCountries),
       aes(x = year, y = lifeExp, color = country)) +
geom_line() + geom_point()
```



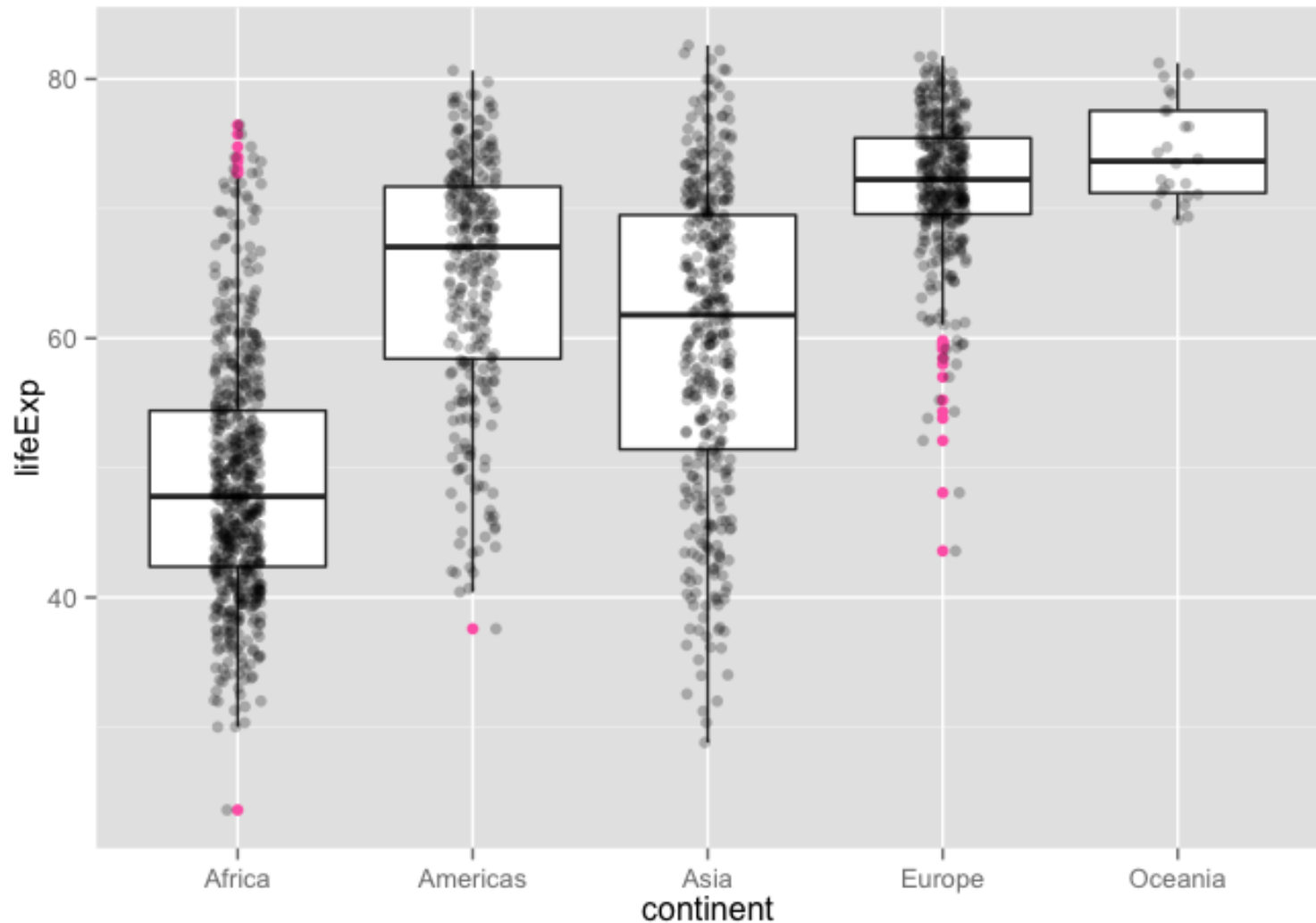
```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +  
geom_point()
```



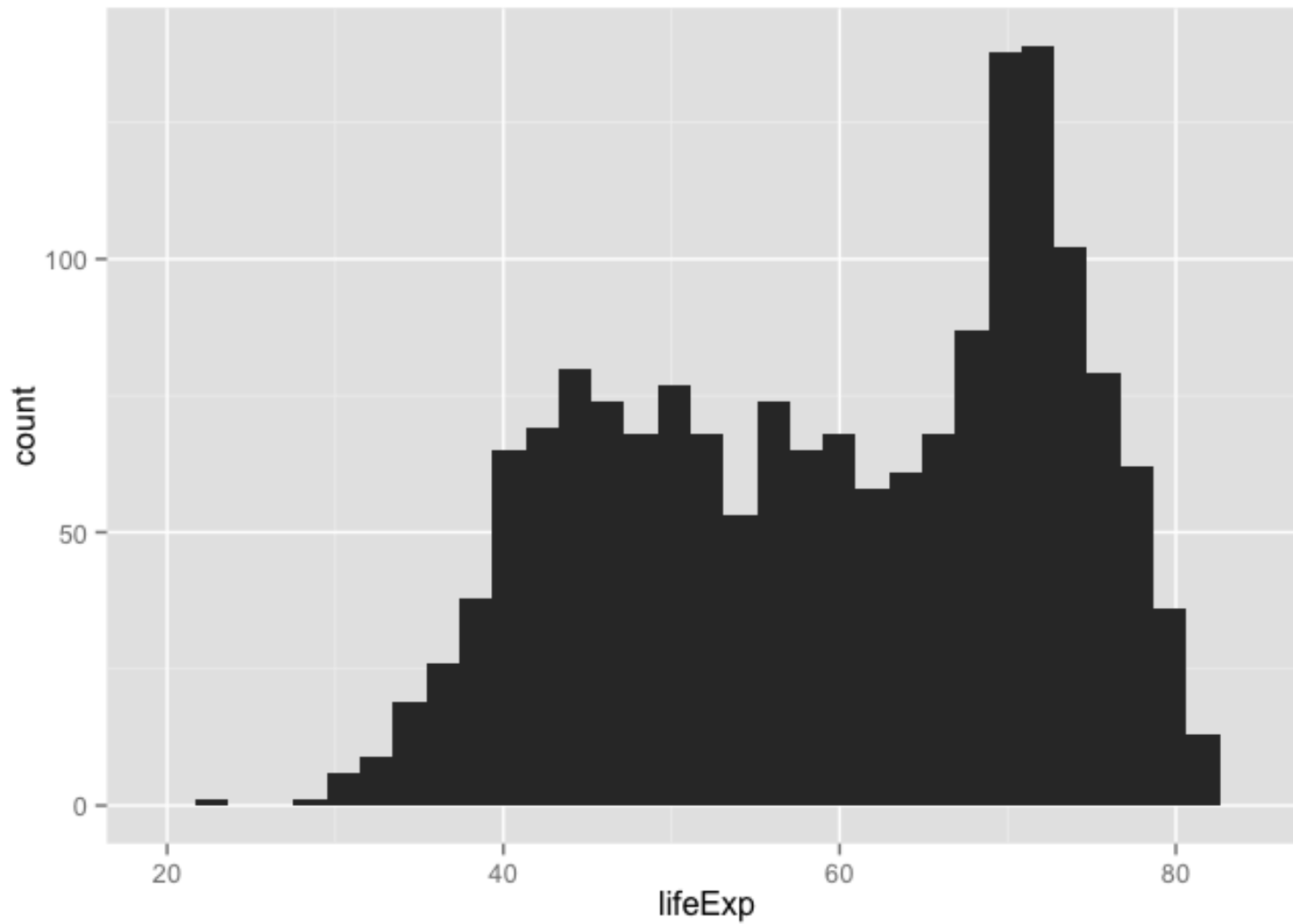
```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +  
geom_jitter()
```

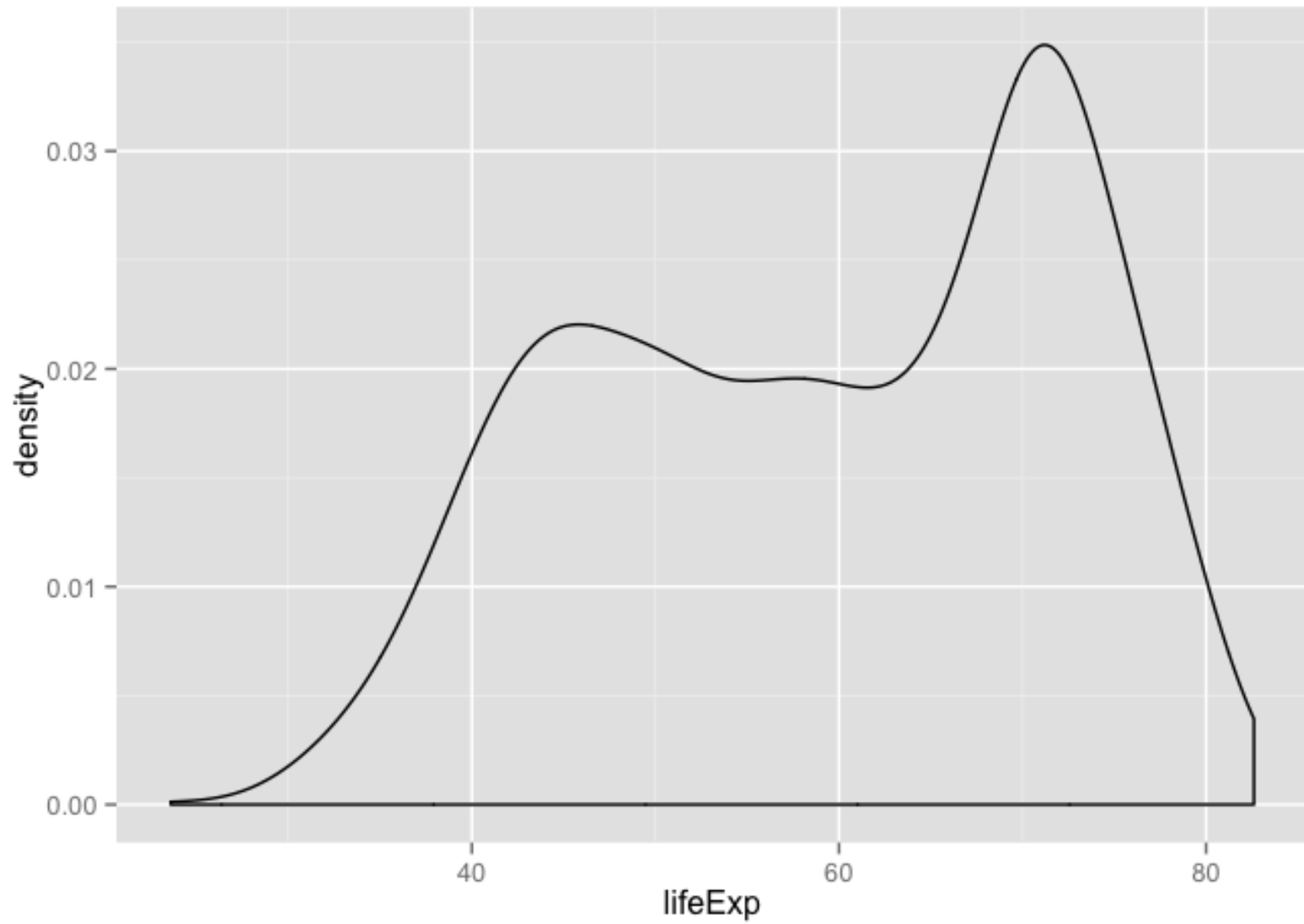
```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +  
geom_boxplot()
```



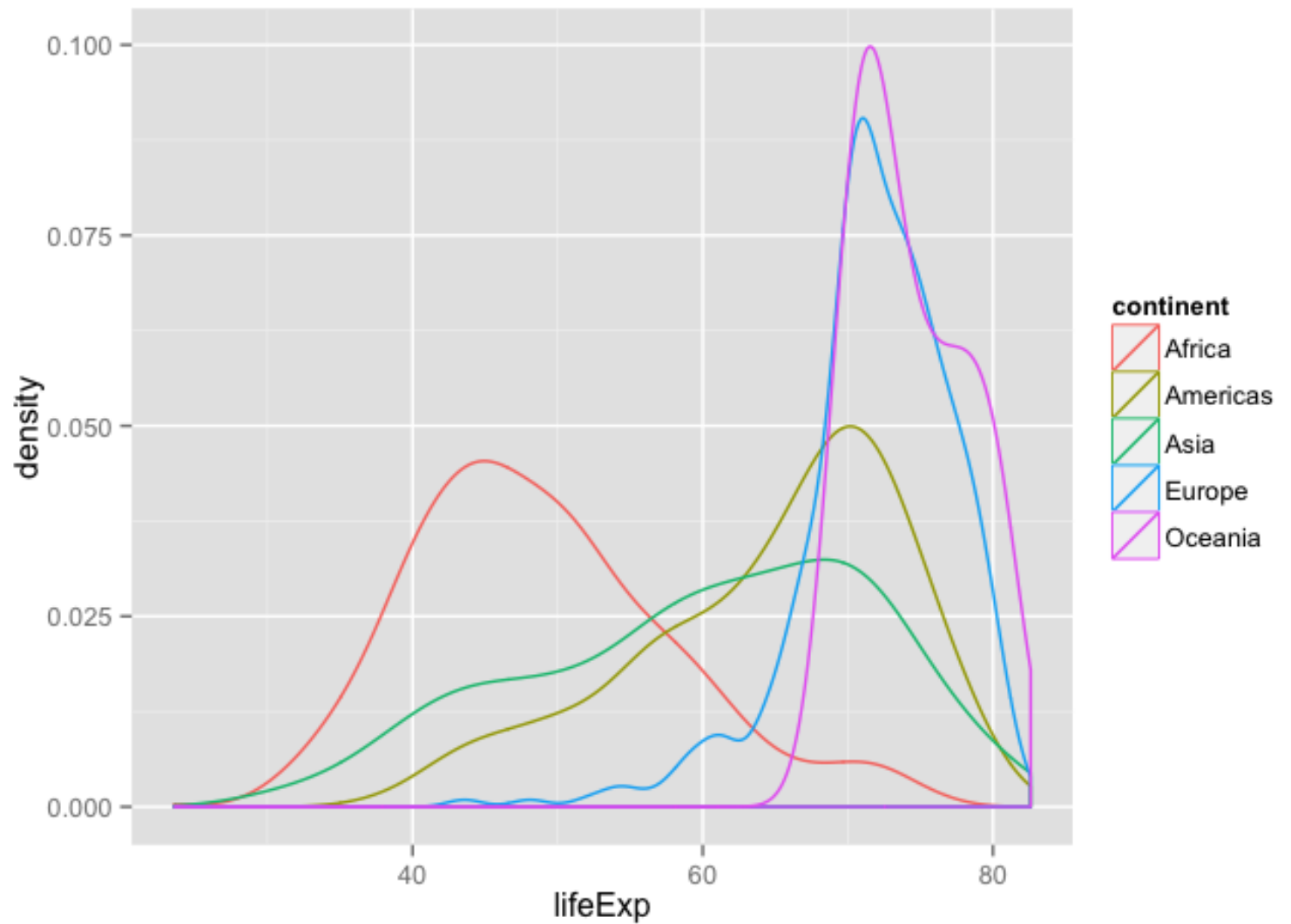
```
ggplot(gapminder, aes(x = continent, y = lifeExp)) +  
  geom_boxplot(outlier.colour = "hotpink") +  
  geom_jitter(position = position_jitter(width = 0.1, height =  
0), alpha = 1/4)
```



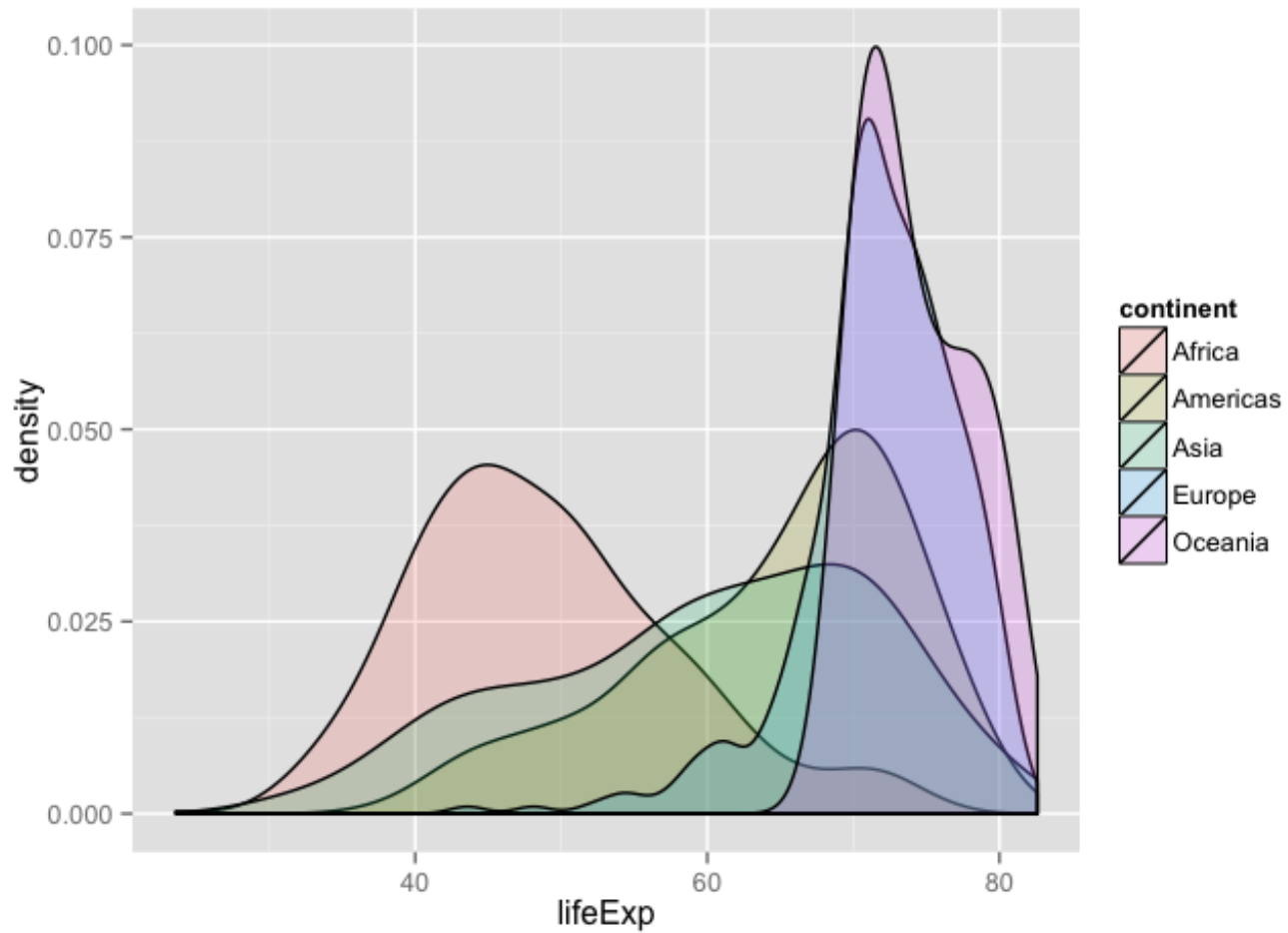
```
ggplot(gapminder, aes(x = lifeExp)) + geom_histogram()
```



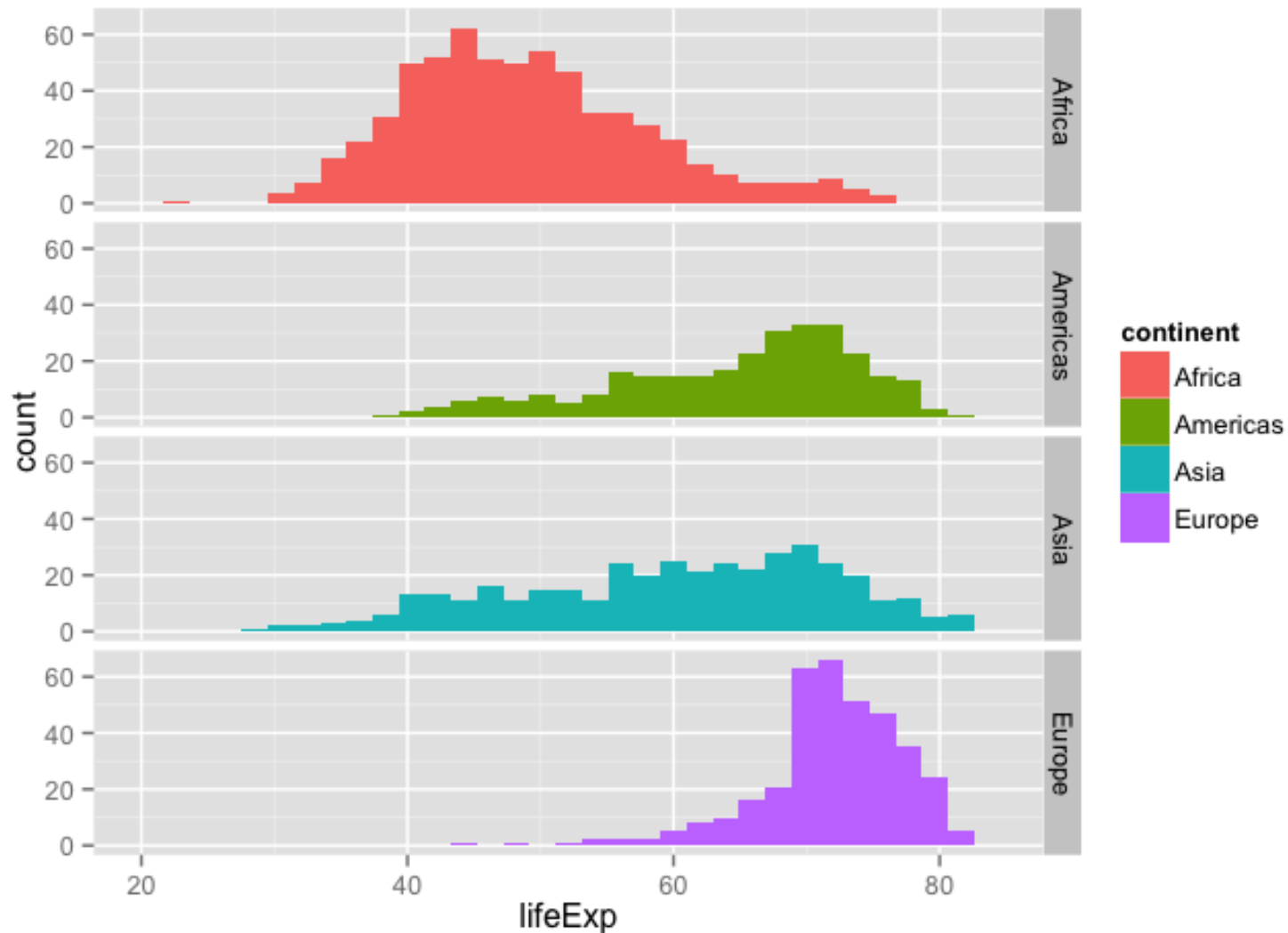
```
ggplot(gapminder, aes(x = lifeExp)) + geom_density()
```



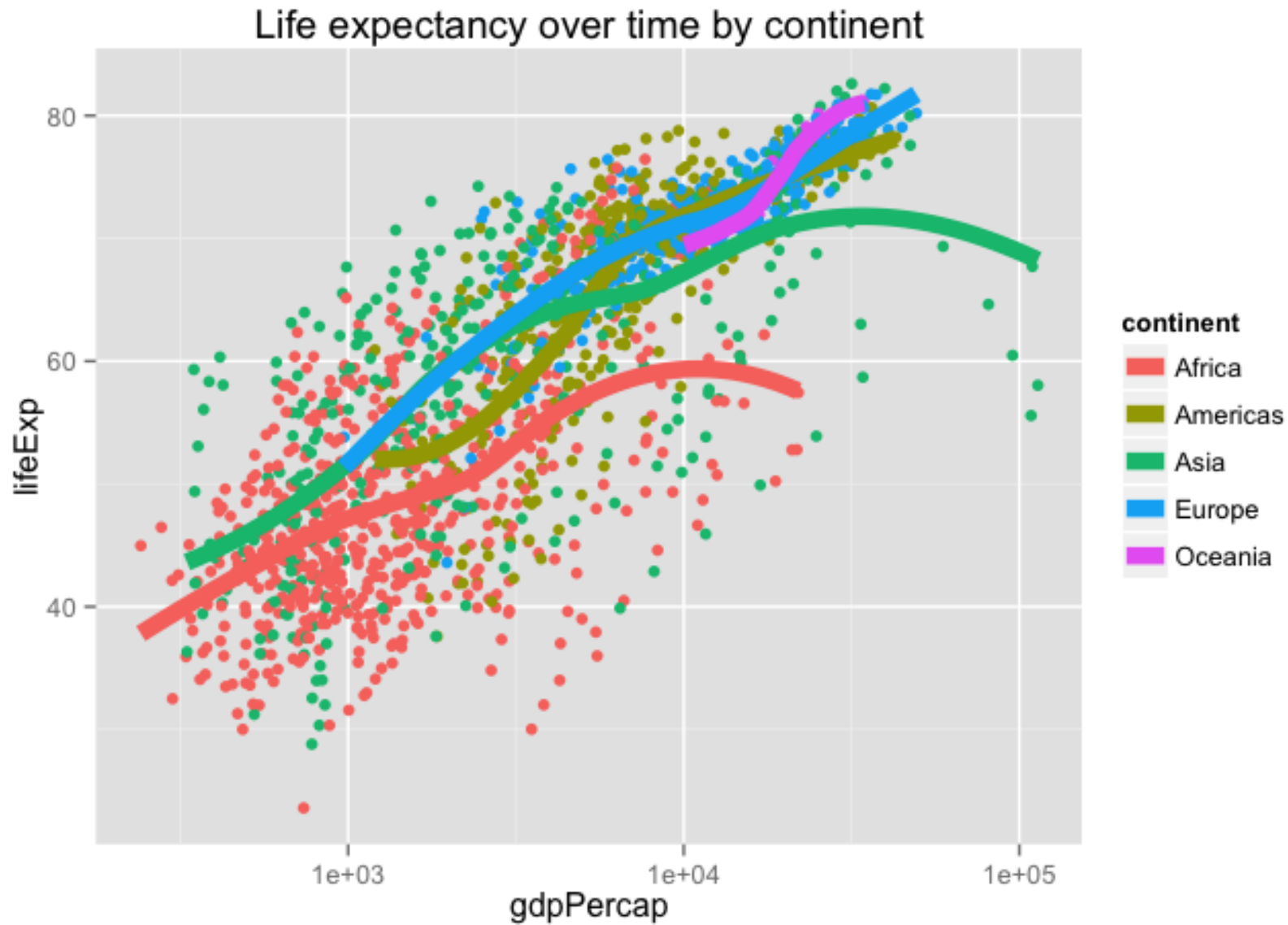
```
ggplot(gapminder, aes(x = lifeExp, color = continent)) +  
geom_density()
```



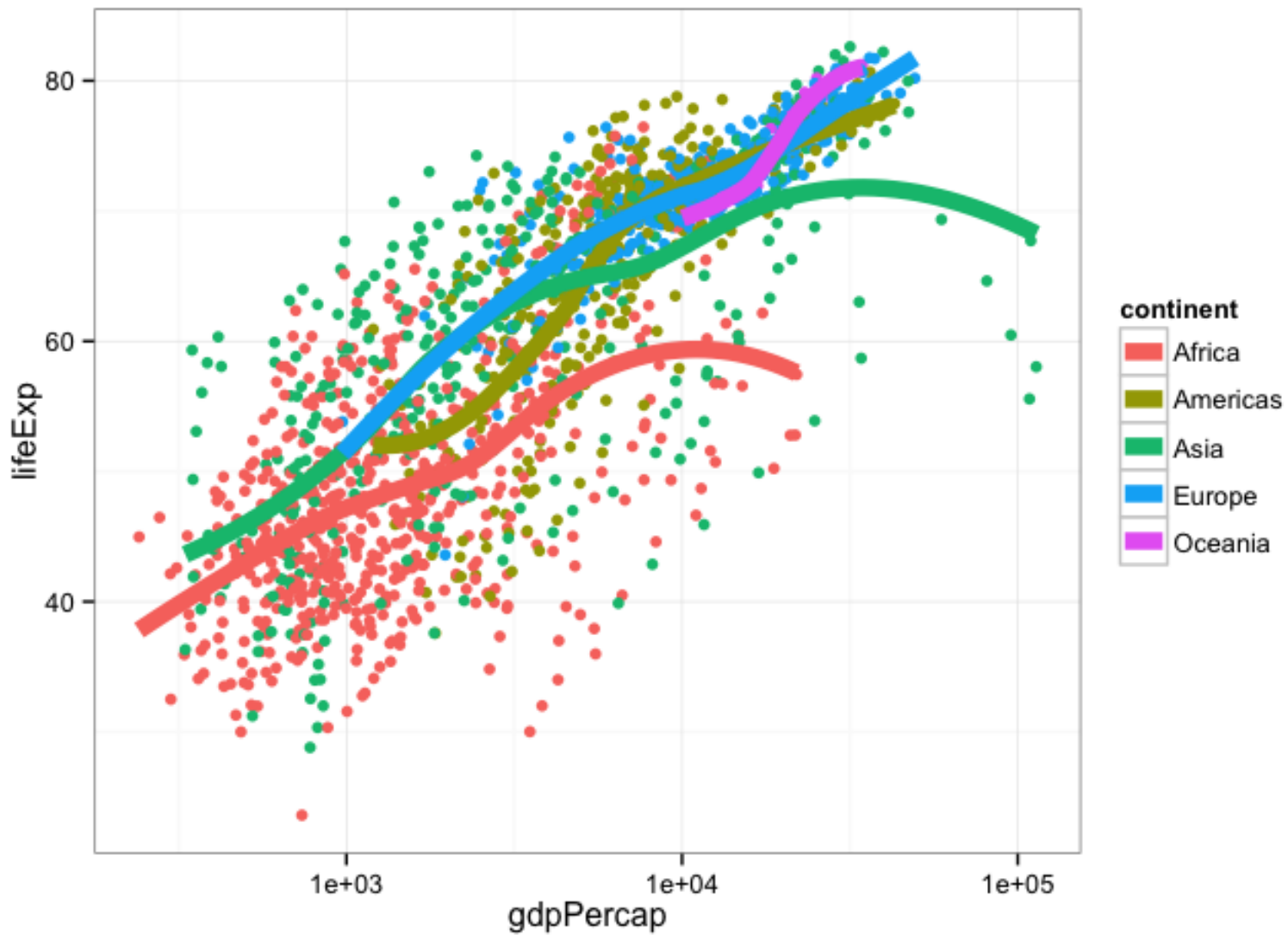
```
ggplot(gapminder, aes(x = lifeExp, fill = continent)) +  
  geom_density(alpha = 0.2)
```



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



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

```
p + theme_bw()
```

Пример с melt

```
>grades<-read.csv("grades.csv")
```

```
>head(grades)
```

	id	write	math	science	socst
1	70	52	41	47	57
2	121	59	53	63	61
3	86	33	54	58	31
4	141	44	47	53	56
5	172	52	57	53	61
6	113	52	51	63	61

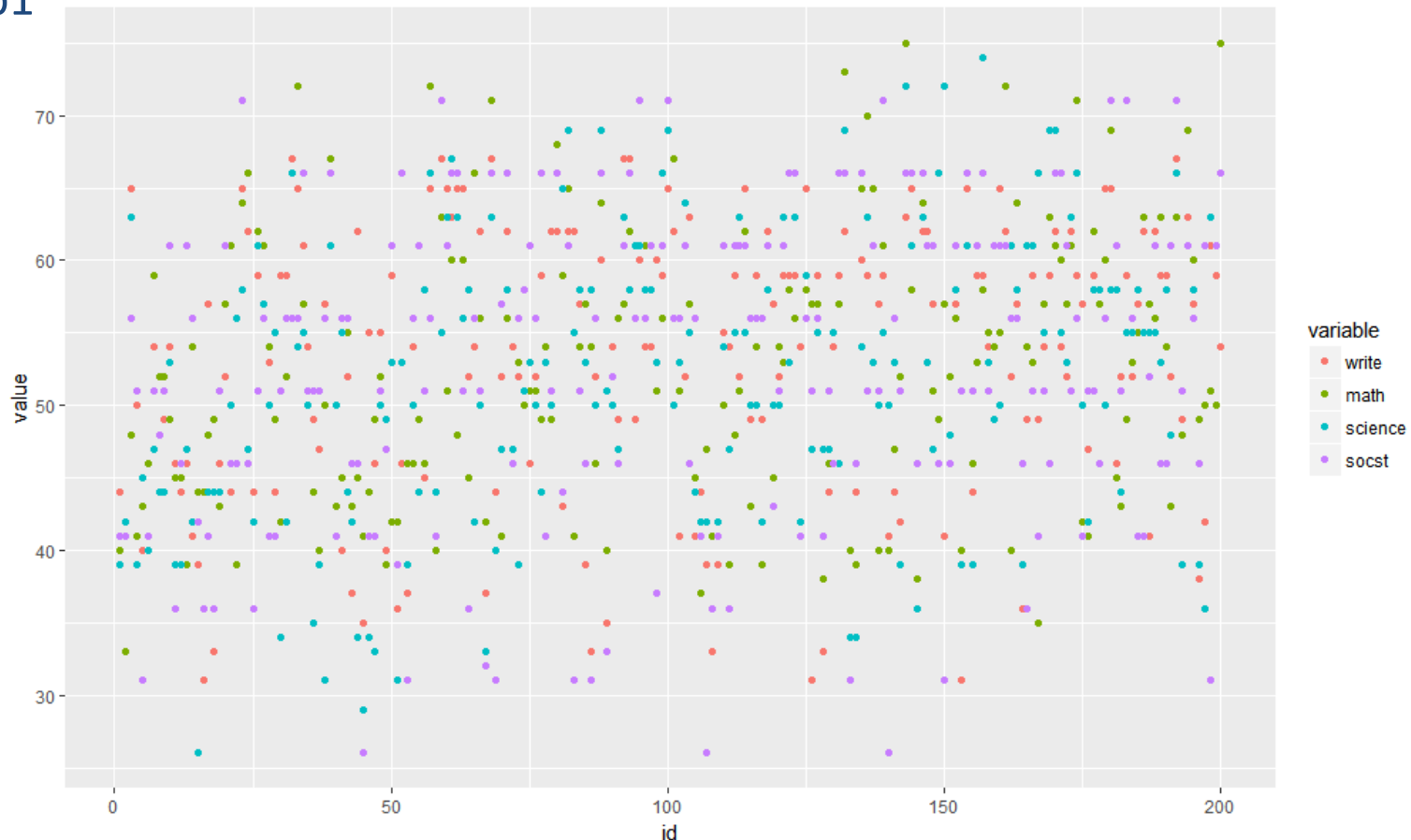
Хотим построить точечный график оценок для
каждого студента, каждый предмет своим цветом

Пример с melt

```
>grades_m<-melt(grades, id.vars = c("id", "X"))
```


```
>p1<-ggplot(grades_m, aes(x = id, y = value, color=variable)) + geom_point()
```

```
>p1
```



Больше графиков тут:

<http://www.r-graph-gallery.com/portfolio/ggplot2-package/>



The screenshot shows a web browser window displaying the 'GGPLOT2' page on 'THE R GRAPH GALLERY' website. The browser's address bar shows the URL 'www.r-graph-gallery.com/portfolio/ggplot2-package/'. The website header includes the logo (an orange circle with a white line graph) and the text 'THE R GRAPH GALLERY'. A navigation menu contains links: HOME, GGPLOT2, ALL GRAPHS, BLOG, ABOUT, and PYTHON. The main content area is titled 'GGPLOT2' in orange. Below the title, there is a paragraph of text about the GGplot2 package, followed by three columns of text. The first column describes the package and its creator, Hadley Wickham. The second column mentions a 'quick start page' and a 'cheat sheet'. The third column discusses finding graphs and contributing. At the bottom of the page, there are four promotional banners: a black one with white line graphs, an orange one for 'The fastest way to learn R!', a dark blue one for 'EARL CONFERENCE', and a light blue one for a 'Beginner's Guide to Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-INLA'.

THE R GRAPH GALLERY

HOME GGPLOT2 ALL GRAPHS BLOG ABOUT PYTHON

GGPLOT2

GGplot2 is an R package created by Hadley Wickham in 2005. It can highly improve the quality and aesthetic of your graphs ! If you speak french, check

this really clear [quick start page](#). More over R studio provides a really helpfull [cheat sheet](#) summarizing the main ggplot2 functions. I hope you will find

the graph you are looking for in the examples below !If it is not the case, don't forget to contribute by [requesting](#) or [proposing](#) a graph !

The fastest way to learn R!
R Courses for Professionals

EARL
CONFERENCE

Beginner's Guide to
Spatial, Temporal and
Spatial-Temporal Ecological
Data Analysis with R-INLA

Beginner's Guide to
Spatial, Temporal
and Spatial-
Temporal Ecological

Чистка 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 сколько объектов у нас нет данных про цены

```
>sum(is.na(bk$SALE.PRICE.N))
```

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))
```

Чистим данные

Приводим в порядок даты

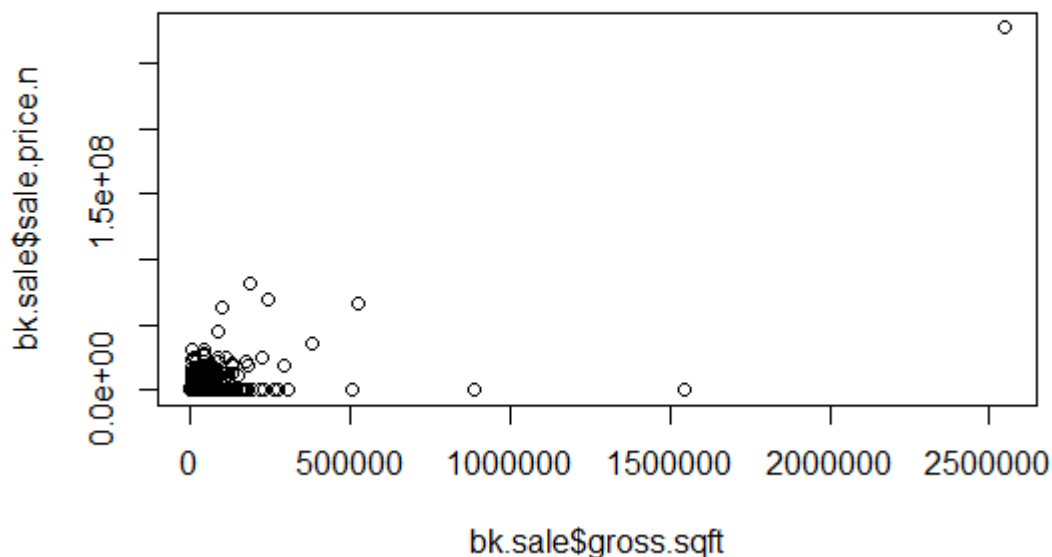
```
>bk$sale.date <- as.Date(bk$sale.date)  
>bk$year.built <- as.numeric(as.character(bk$year.built))
```

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!
```

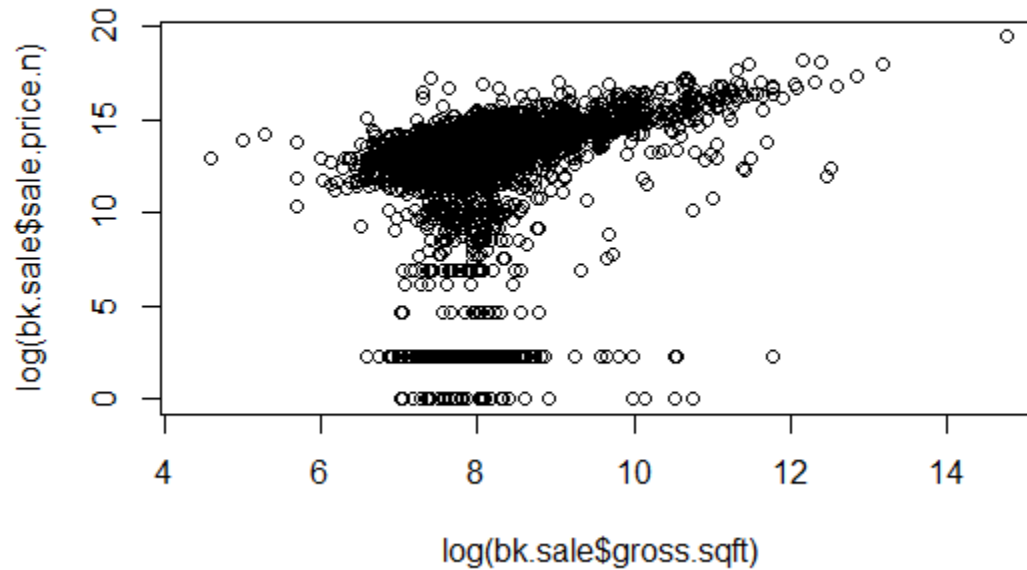
Теперь беглый анализ, как устроены данные

```
>bk.sale <- filter(bk, bk$sale.price!=0 & bk$gross.sqft!=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))
```

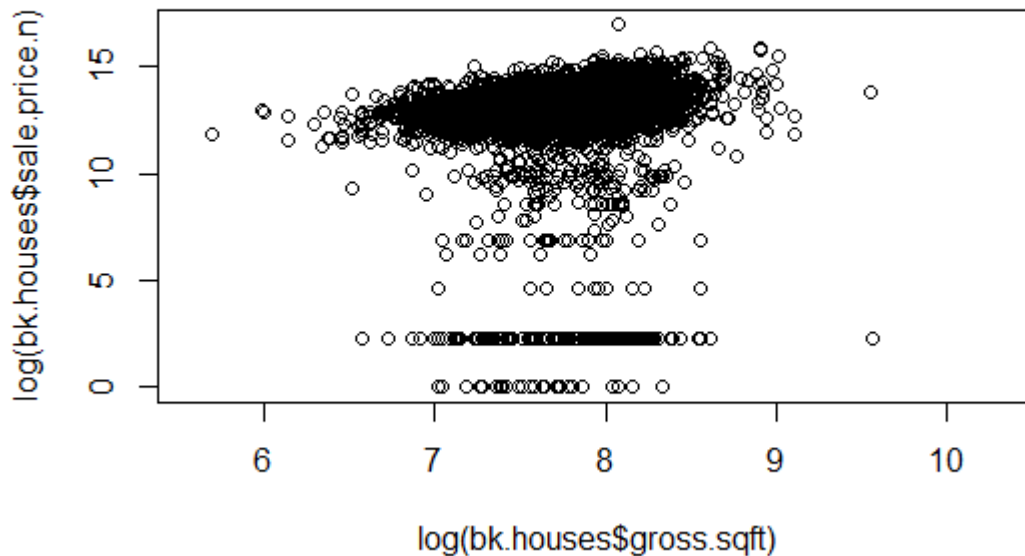


Выберем для анализа только дома (категория содержит в названии "FAMILY")

```
>library(data.table)
```

```
> bk.houses <- bk.sale %>% filter(building.class.category %like%  
"FAMILY")
```

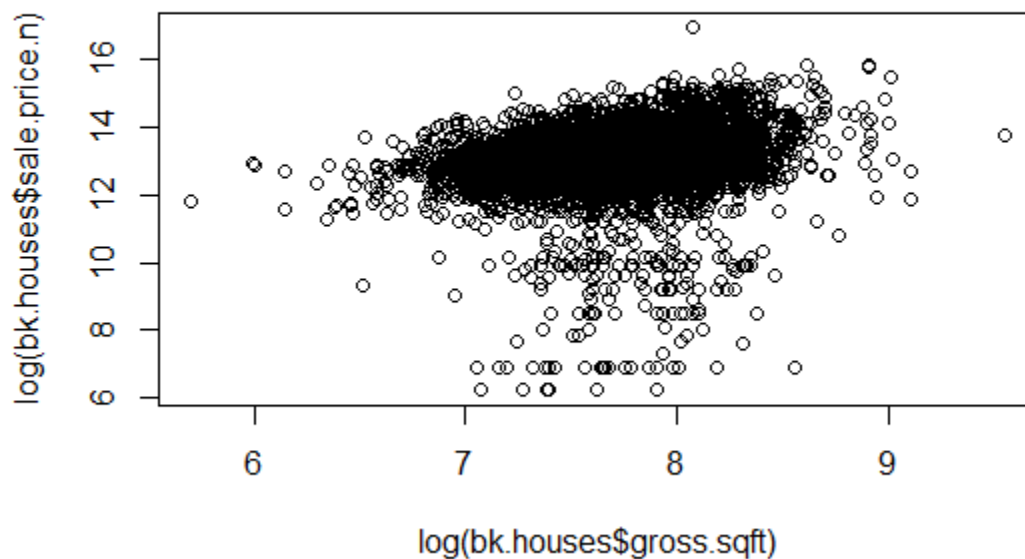
```
> plot(log(bk.houses$gross.sqft),log(bk.houses$sale.price.n))
```



Похоже,
эти дома
какие-то
особенные

Уберем “особенные” дома

```
> bk.houses <- bk.houses %>% filter(log(sale.price.n) > 5)  
> plot(log(bk.houses$gross.sqft), log(bk.houses$sale.price.n))
```



Добавим линию регрессии

```
>abline(lm(log(bk.houses$sale.price.n)~log(bk.h  
ouses$gross.sqft)), col="red")
```

