Exploratory analysis of gapminder data

Load the gapminder data. We will also need two tidyverse packages: `dplyr` facilitates exploratory analyses and `ggplot2` allows visualization.

```
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
```

```
library(ggplot2)
```

Take a look at the top and bottom few lines of raw data.

```
head(gapminder)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

```
tail(gapminder)
```

```
## # A tibble: 6 x 6
##   country  continent  year lifeExp      pop gdpPercap
##   <fct>    <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Zimbabwe Africa     1982    60.4  7636524      789.
## 2 Zimbabwe Africa     1987    62.4  9216418      706.
## 3 Zimbabwe Africa     1992    60.4 10704340      693.
## 4 Zimbabwe Africa     1997    46.8 11404948      792.
## 5 Zimbabwe Africa     2002    40.0 11926563      672.
## 6 Zimbabwe Africa     2007    43.5 12311143      470.
```

```
summary(gapminder)
```

```
##         country        continent        year         lifeExp
##  Afghanistan:  12   Africa  :624   Min.   :1952   Min.   :23.60
##  Albania    :  12   Americas:300   1st Qu.:1966   1st Qu.:48.20
##  Algeria    :  12   Asia    :396   Median :1980   Median :60.71
##  Angola     :  12   Europe  :360   Mean   :1980   Mean   :59.47
##  Argentina  :  12   Oceania : 24   3rd Qu.:1993   3rd Qu.:70.85
##  Australia  :  12                  Max.   :2007   Max.   :82.60
##  (Other)    :1632
##       pop              gdpPercap
##  Min.   :6.001e+04   Min.   :   241.2
##  1st Qu.:2.794e+06   1st Qu.:  1202.1
##  Median :7.024e+06   Median :  3531.8
##  Mean   :2.960e+07   Mean   :  7215.3
##  3rd Qu.:1.959e+07   3rd Qu.:  9325.5
##  Max.   :1.319e+09   Max.   :113523.1
##
```

Type `help("gapminder")` in the R console for information about the gapmider dataset.

We will explore the life expectancy variable for the year 2007. First filter the data to just 2007.

```
gapminder07 <- filter(gapminder, year == 2007)
head(gapminder07)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       2007    43.8 31889923      975.
## 2 Albania     Europe     2007    76.4  3600523     5937.
## 3 Algeria     Africa     2007    72.3 33333216     6223.
## 4 Angola      Africa     2007    42.7 12420476     4797.
## 5 Argentina   Americas   2007    75.3 40301927    12779.
## 6 Australia   Oceania    2007    81.2 20434176    34435.
```

In R, the `<-` is the assignment operator that creates new variables/datasets.

Life expectancy by continent

Calculate median life expectancy, first overall, and then by continent.

```
summarize(gapminder07, median(lifeExp))
```

```
## # A tibble: 1 x 1
##   `median(lifeExp)`
##               <dbl>
## 1              71.9
```

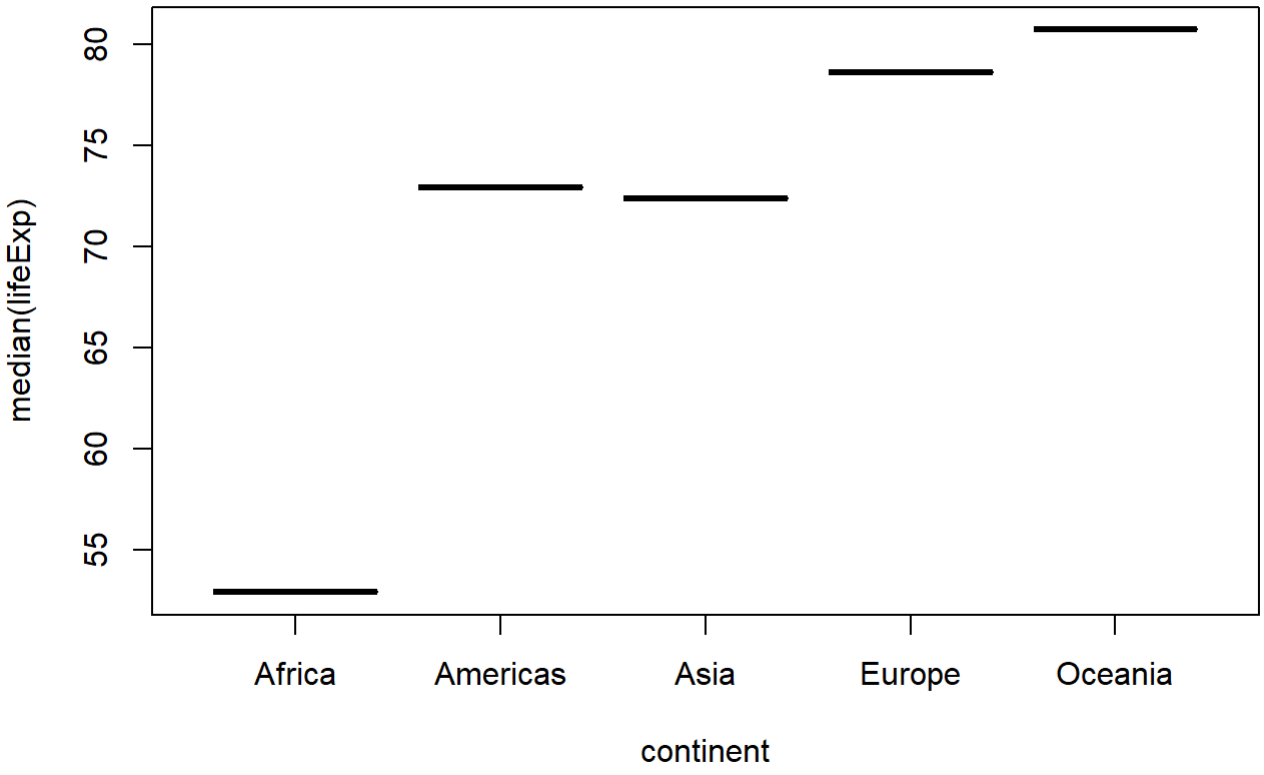
```
by_cont <- group_by(gapminder07, continent)
summarise(by_cont, median(lifeExp))
```

```
## # A tibble: 5 x 2
##   continent `median(lifeExp)`
##   <fct>                 <dbl>
## 1 Africa                 52.9
## 2 Americas               72.9
## 3 Asia                   72.4
## 4 Europe                 78.6
## 5 Oceania                80.7
```

In the above commands, `group_by()` creates a new data set with observations grouped by continent.

We can visualize the median life expectancies.

```
medL <- summarize(by_cont, median(lifeExp))
plot(medL)
```



What is “Oceania”?

```
filter(gapminder07,continent == "Oceania")
```

```
## # A tibble: 2 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Australia   Oceania    2007    81.2 20434176    34435.
## 2 New Zealand Oceania    2007    80.2  4115771    25185.
```

The `dplyr` package allows for us to "chain" the filter, grouping and summary commands. The following is an equivalent way to construct `medL` :

```
medL <- gapminder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarise(medLifeExp = median(lifeExp))
```

Life expectancy over time

First look at African countries

```
medLA <- gapminder %>%
  filter(continent == "Africa") %>%
  group_by(country) %>%
  summarise(medLifeExp = median(lifeExp))
```

Look at a subset of countries with the lowest and highest median life expectancies.

```
filter(medLA,medLifeExp<40)
```

```
## # A tibble: 3 x 2
##   country       medLifeExp
##   <fct>              <dbl>
## 1 Angola              39.7
## 2 Guinea-Bissau       38.4
## 3 Sierra Leone        37.6
```

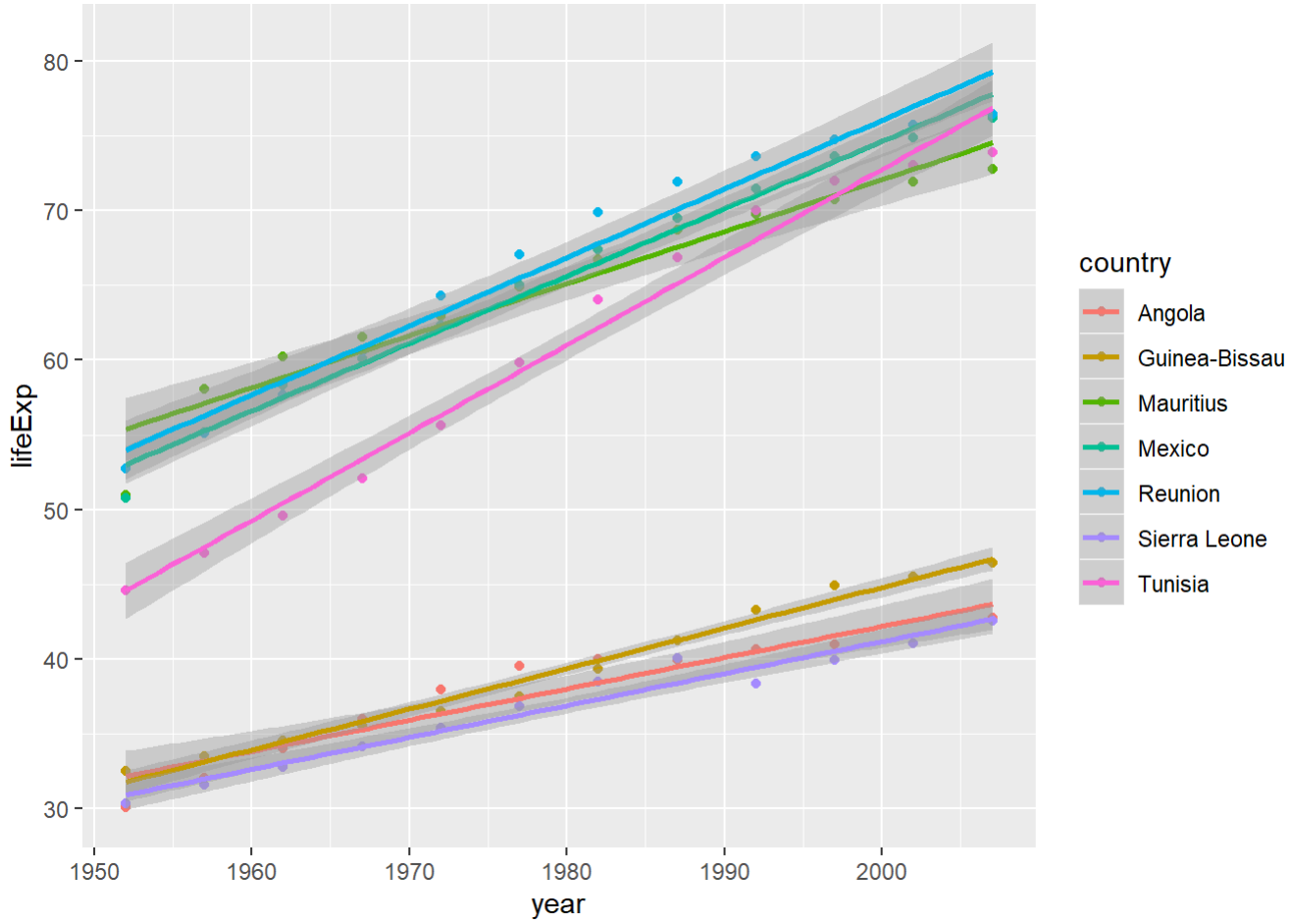
```
filter(medLA,medLifeExp>60)
```

```
## # A tibble: 3 x 2
##   country   medLifeExp
##   <fct>          <dbl>
## 1 Mauritius       65.8
## 2 Reunion         68.5
## 3 Tunisia         61.9
```

```
cc = c("Angola","Guinea-Bissau","Sierra Leone",
       "Mauritius","Reunion","Tunisia",
       "Mexico") # Mexico for comparison
```

Plot life expectancy over time. Illustrate chaining of filtering (on country) and ggplot.

```
gapminder %>%
  filter(country %in% cc) %>%
  ggplot(aes(x=year,y=lifeExp,color=country)) +
   geom_point() +
   geom_smooth(method = "lm")
```



Here's another interesting plot of life expectancy over time:

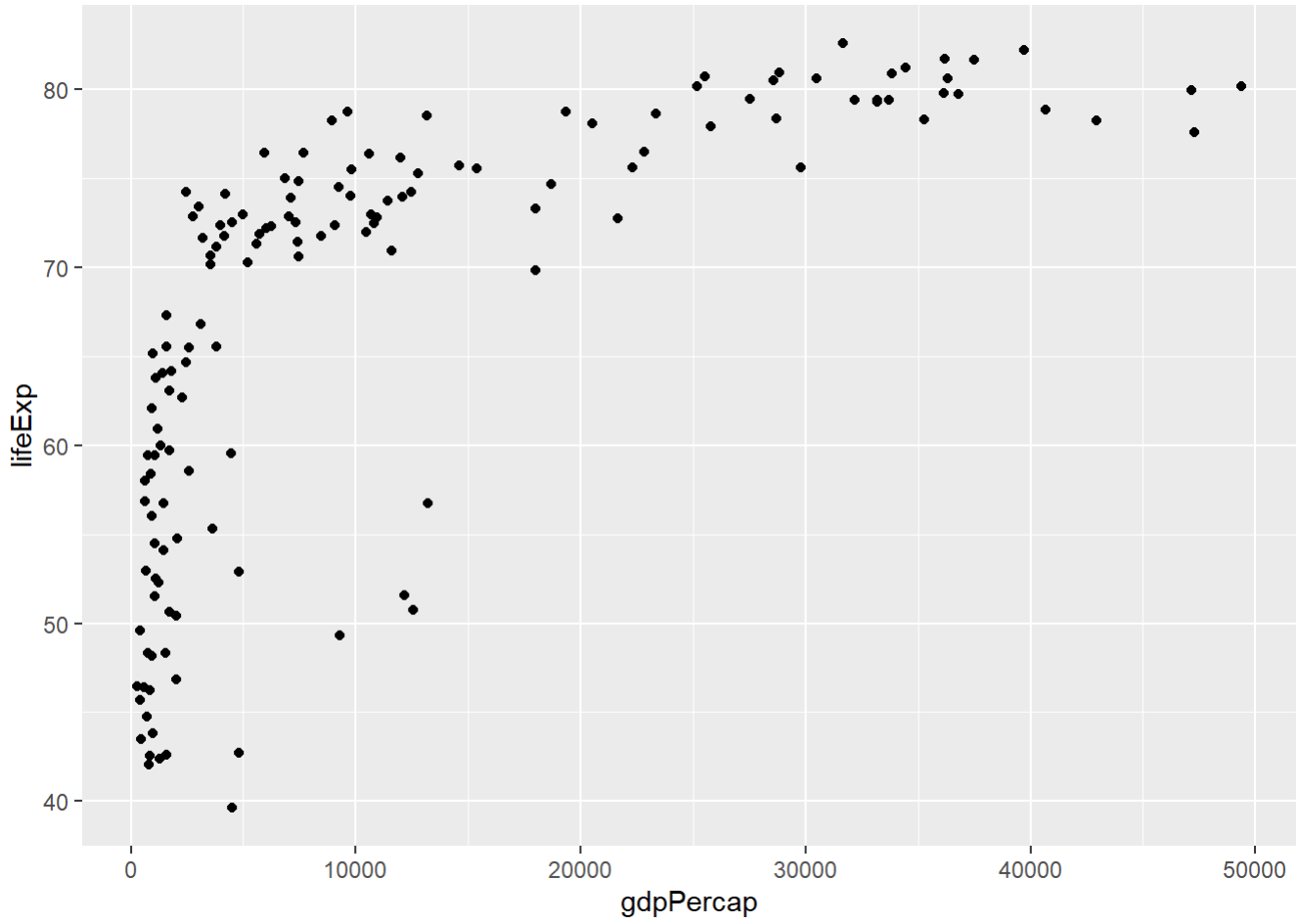
```
gapminder %>%
  filter(continent == "Oceania") %>%
  ggplot(aes(x=year,y=lifeExp,color=country)) +
  geom_point() +
  geom_smooth(method = "loess", span=3/4)
```

happend in the mid-1970s in Australia?

Life expectancy versus per capita GDP

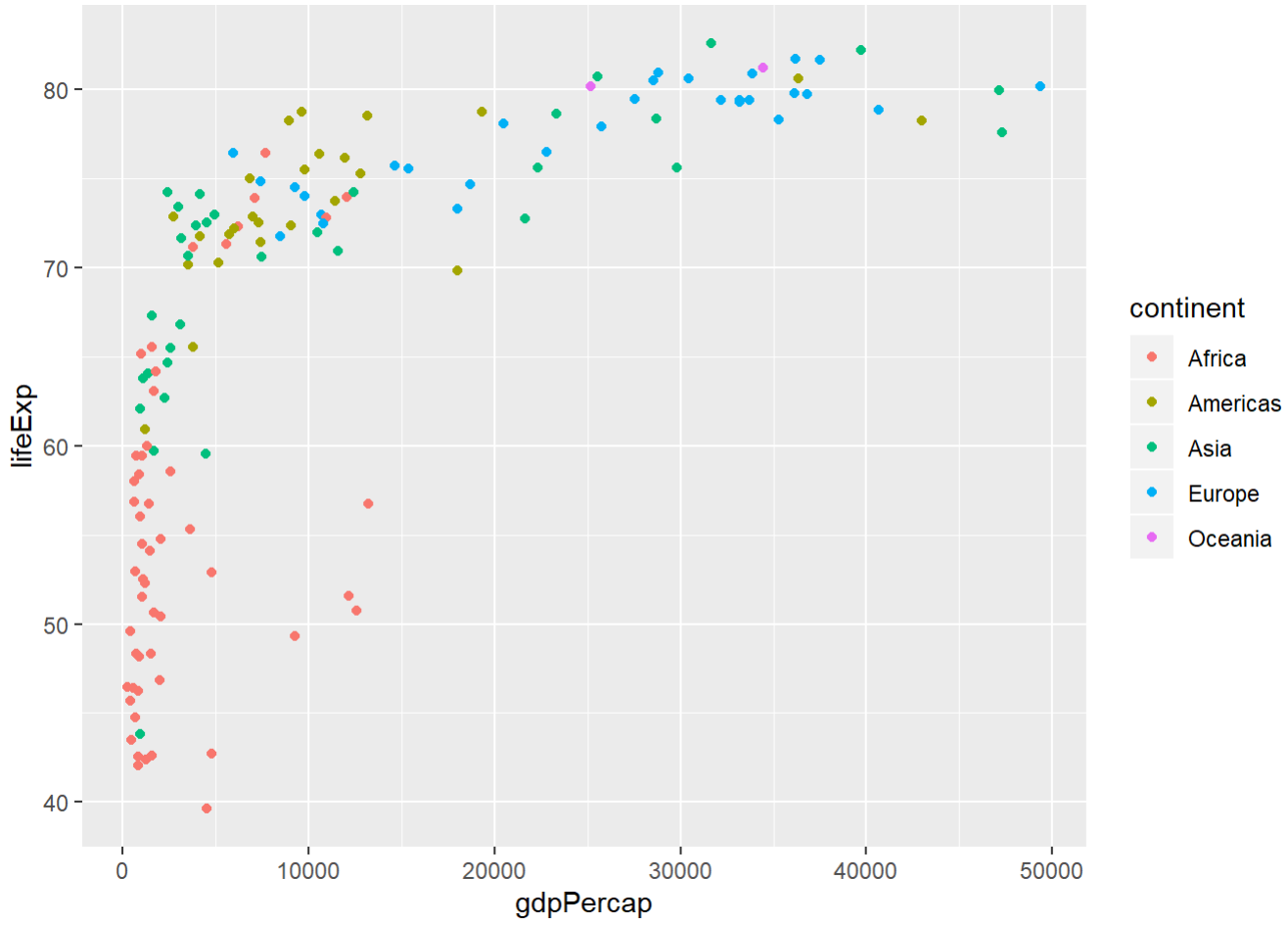
First try a simple scatterplot of `lifeExp` versus `gdpPercap` .

```
qplot(gdpPercap,lifeExp,data=gapminder07)
```



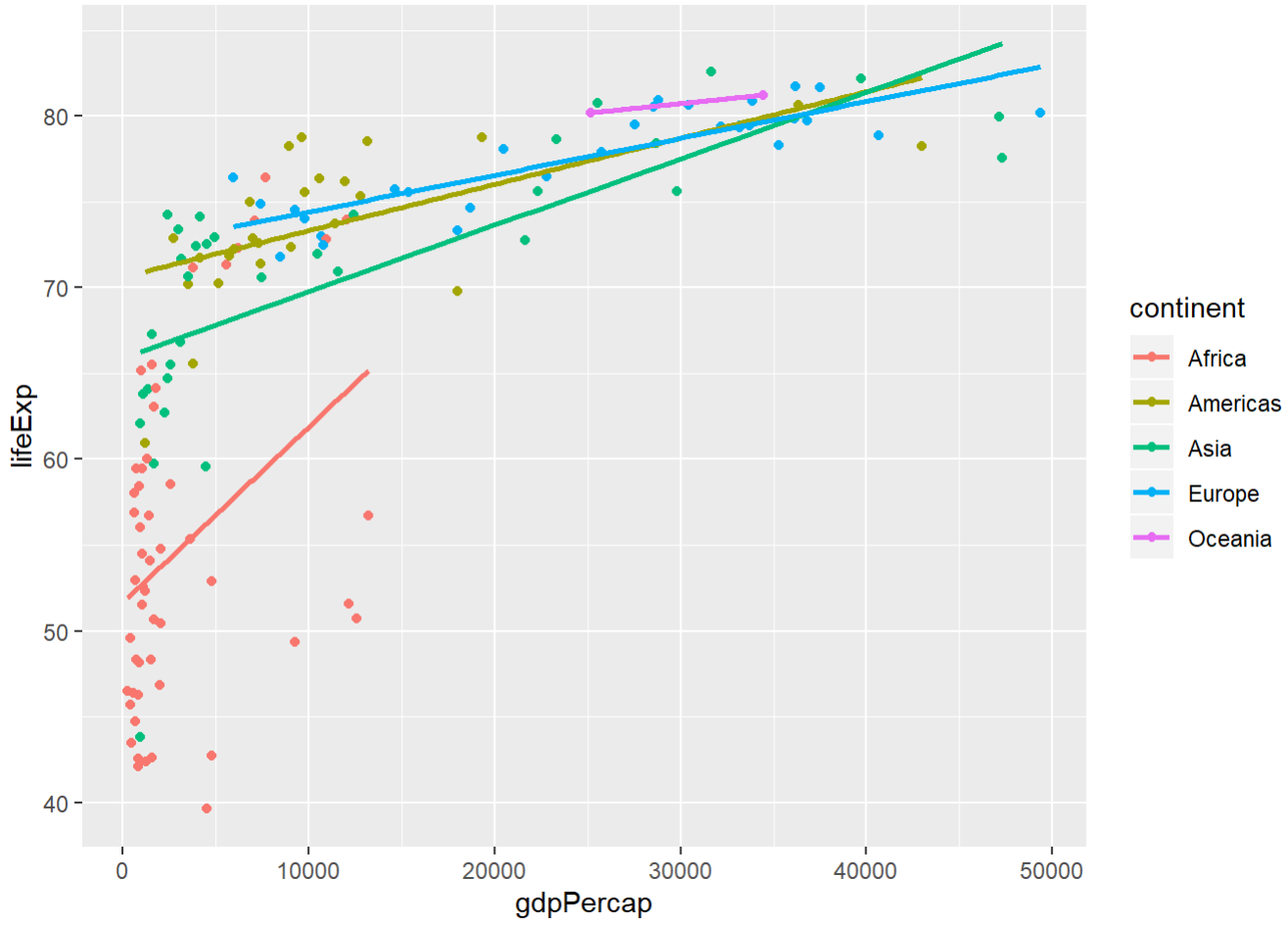
It is hard to make sense of the pattern in `lifeExp` versus `gdpPercap`. Try grouping the data by continent. (Note: This does not use our `by_cont` data set. We'll talk about why later.)

```
qplot(gdpPercap,lifeExp,data=gapminder07,color = continent)
```



Add regression lines for each continent. Doing so uses a more complicated graphing function from `ggplot2`.

```
ggplot(gapminder07, aes(x=gdpPercap,y=lifeExp,color=continent)) +
 geom_point() +
 geom_smooth(method = "lm", se=FALSE)
```


lifeExp
80
70
60
50
40
0
10000
20000
30000
40000
50000
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania