Data Visualization in R

Umair Durrani

2019-06-24

Table of Contents

# Load libraries



# Load data

An excerpt of the data available at Gapminder.org. For each of 142 countries, the package provides values for life expectancy, GDP per capita, and population, every five years, from 1952 to 2007.

# Exploratory Data Analysis

## Data structure

## Observations: 1,704  
## Variables: 6  
## $ country <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan, Af…  
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, …  
## $ year <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, …  
## $ lifeExp <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854…  
## $ pop <int> 8425333, 9240934, 10267083, 11537966, 13079460, 148803…  
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.…

**First 10 rows**

country

continent

year

lifeExp

pop

gdpPercap

Afghanistan

Asia

1952

28.801

8425333

779.4453

Afghanistan

Asia

1957

30.332

9240934

820.8530

Afghanistan

Asia

1962

31.997

10267083

853.1007

Afghanistan

Asia

1967

34.020

11537966

836.1971

Afghanistan

Asia

1972

36.088

13079460

739.9811

Afghanistan

Asia

1977

38.438

14880372

786.1134

Afghanistan

Asia

1982

39.854

12881816

978.0114

Afghanistan

Asia

1987

40.822

13867957

852.3959

Afghanistan

Asia

1992

41.674

16317921

649.3414

Afghanistan

Asia

1997

41.763

22227415

635.3414

## Skim summary statistics  
## n obs: 1704   
## n variables: 6   
##   
## ── Variable type:factor ──────────────────────────────────────────────────────────────────────────────────────────────────  
## variable missing complete n n\_unique  
## continent 0 1704 1704 5  
## country 0 1704 1704 142  
## top\_counts ordered  
## Afr: 624, Asi: 396, Eur: 360, Ame: 300 FALSE  
## Afg: 12, Alb: 12, Alg: 12, Ang: 12 FALSE  
##   
## ── Variable type:integer ─────────────────────────────────────────────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## pop 0 1704 1704 3e+07 1.1e+08 60011 2793664 7e+06   
## year 0 1704 1704 1979.5 17.27 1952 1965.75 1979.5  
## p75 p100 hist  
## 2e+07 1.3e+09 ▇▁▁▁▁▁▁▁  
## 1993.25 2007 ▇▃▇▃▃▇▃▇  
##   
## ── Variable type:numeric ─────────────────────────────────────────────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## gdpPercap 0 1704 1704 7215.33 9857.45 241.17 1202.06 3531.85  
## lifeExp 0 1704 1704 59.47 12.92 23.6 48.2 60.71  
## p75 p100 hist  
## 9325.46 113523.13 ▇▁▁▁▁▁▁▁  
## 70.85 82.6 ▁▂▅▅▅▅▇▃

# Using ggplot2 library for plotting

ggplot2 is the name of a library in R language that is used for plotting. It is a part of the tidyverse library that contains other libraries for *tidy* data analysis.

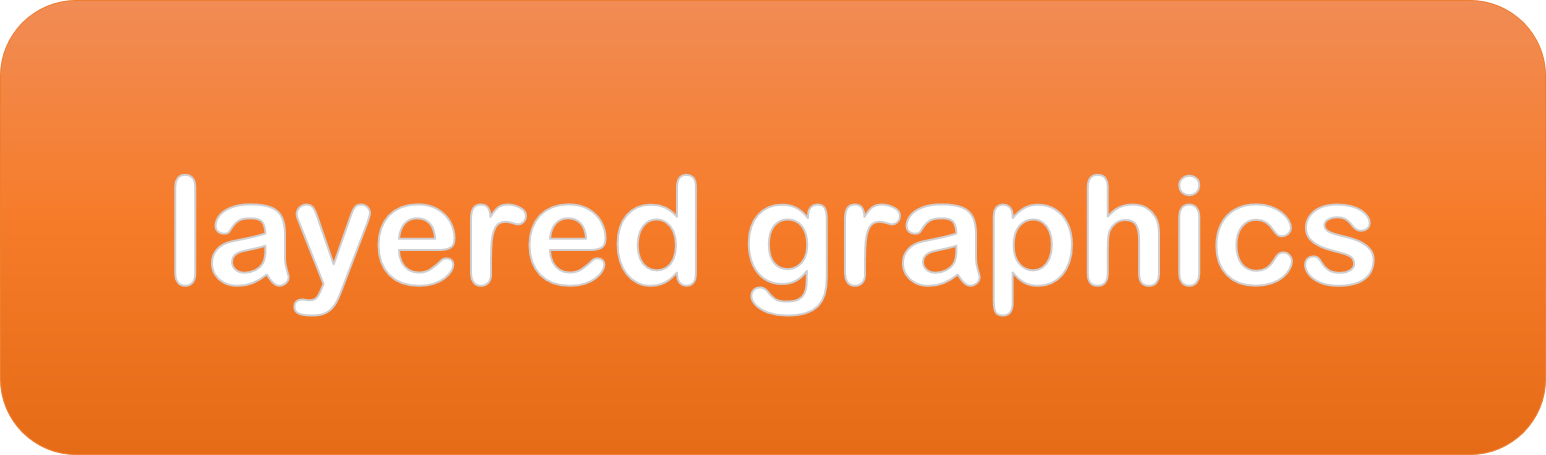
When we ran library(tidyverse), the ggplot2 library was also loaded. If we only want to load ggplot2, we can do so:

## Filtering the gapminder data for **Canada**

**Pseudo code** 1. Take the gapminder data *AND THEN* 2. Filter out all the data except for Canada

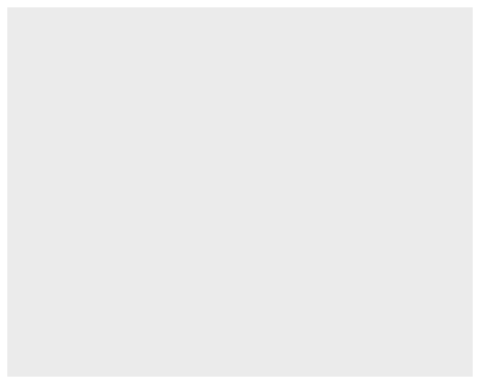
**R code**

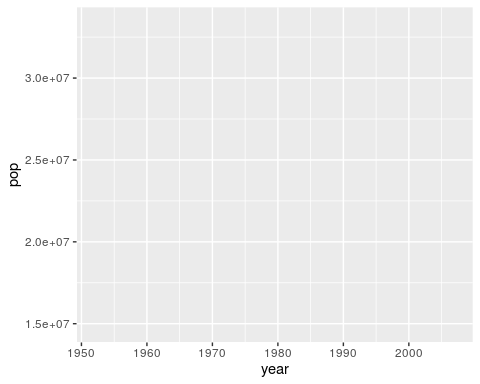
## # A tibble: 12 x 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Canada Americas 1952 68.8 14785584 11367.  
## 2 Canada Americas 1957 70.0 17010154 12490.  
## 3 Canada Americas 1962 71.3 18985849 13462.  
## 4 Canada Americas 1967 72.1 20819767 16077.  
## 5 Canada Americas 1972 72.9 22284500 18971.  
## 6 Canada Americas 1977 74.2 23796400 22091.  
## 7 Canada Americas 1982 75.8 25201900 22899.  
## 8 Canada Americas 1987 76.9 26549700 26627.  
## 9 Canada Americas 1992 78.0 28523502 26343.  
## 10 Canada Americas 1997 78.6 30305843 28955.  
## 11 Canada Americas 2002 79.8 31902268 33329.  
## 12 Canada Americas 2007 80.7 33390141 36319.

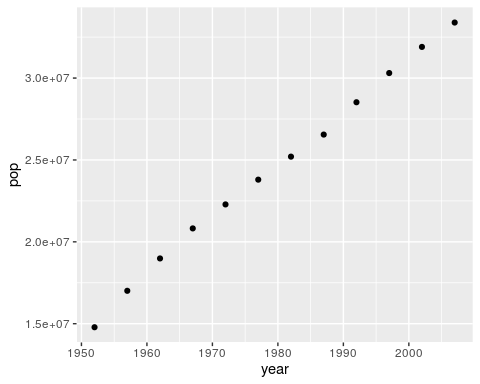




## Plotting population over years

**data** 

**data and aesthetics** 

**data, aesthetics and geometric object** 

YOUR TURN: Create a new data set for your favorite country and plot its population over years

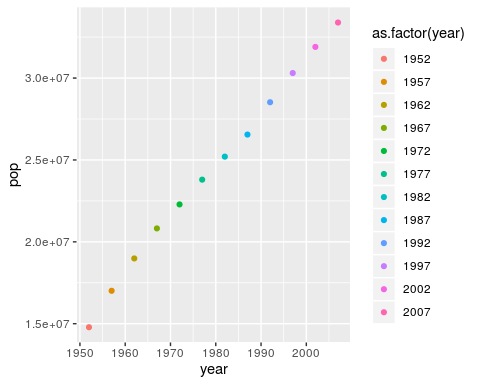


## Color

### Numeric



### Categories



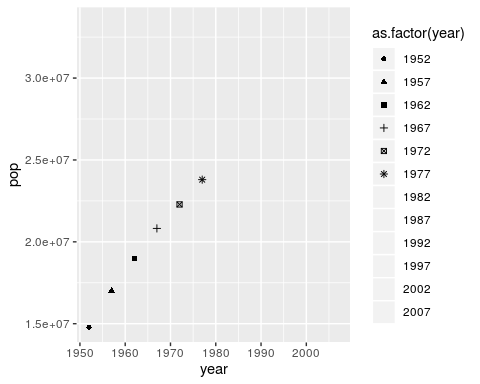
## Size



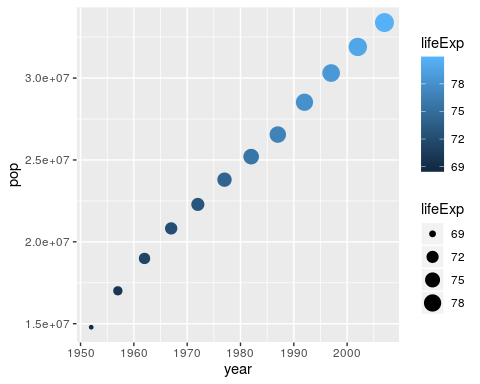
## Shape (Not great for > 3 categories)

## Warning: The shape palette can deal with a maximum of 6 discrete values  
## because more than 6 becomes difficult to discriminate; you have  
## 12. Consider specifying shapes manually if you must have them.

## Warning: Removed 6 rows containing missing values (geom\_point).



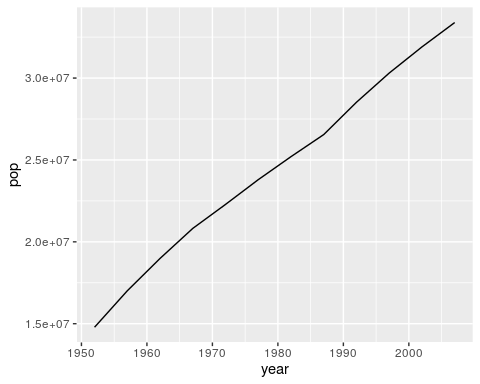
## What if you use all?







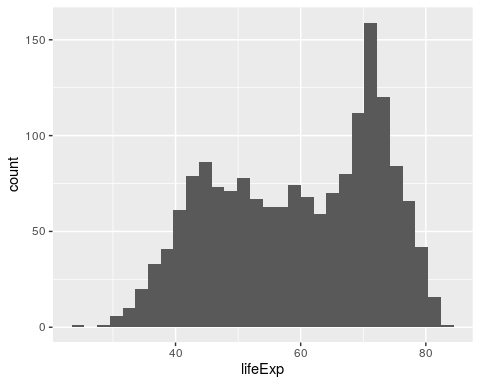
## Line



## Histogram

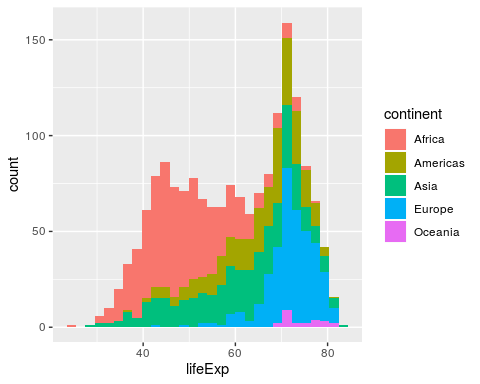
Now using the compete gapminder data:

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



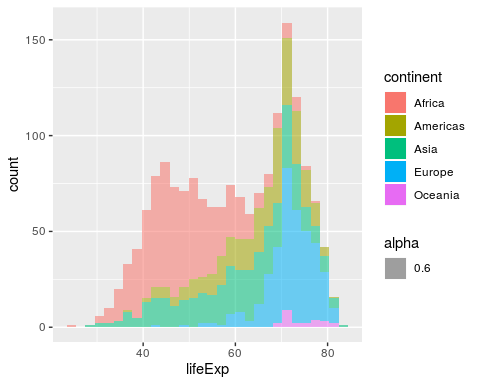
### fill color

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

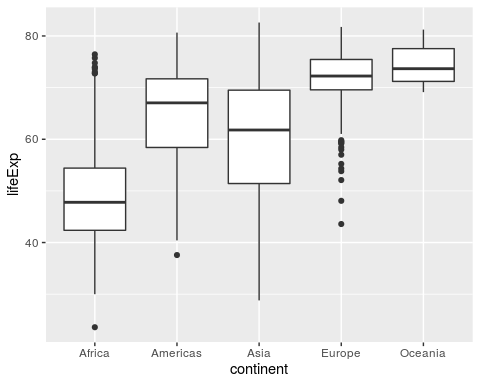


**add transparency**

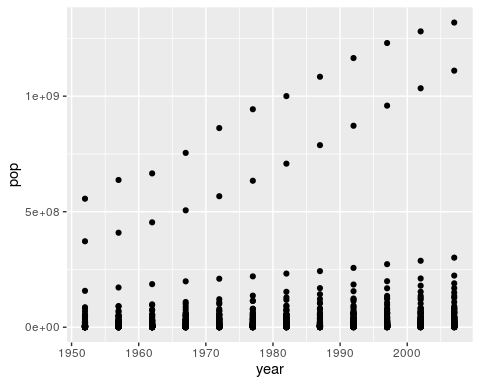
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



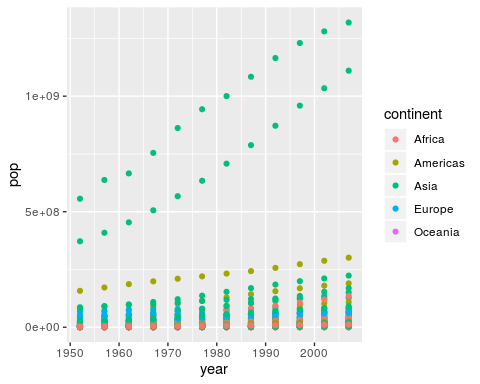
## Boxplot



## Creating a scatter plot for all countries

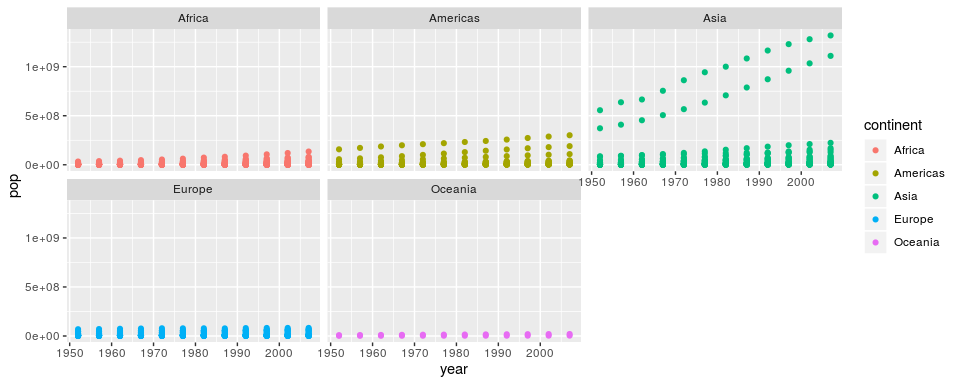
Now we’ll use the complete gapminder data 

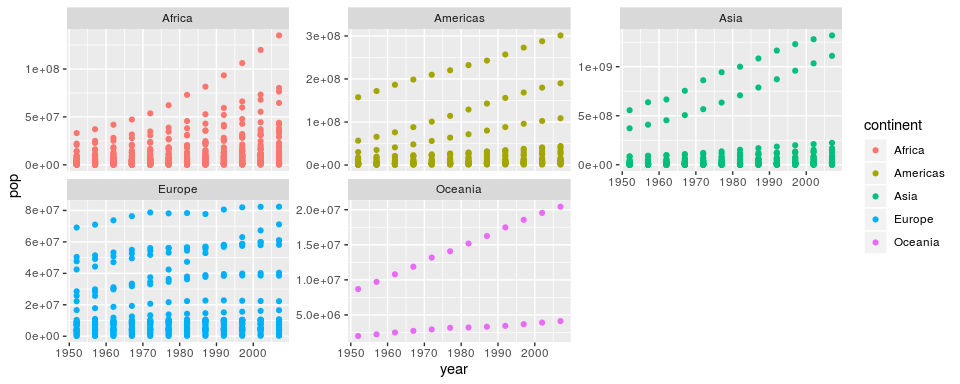
YOUR TURN: Copy the above code and paste it below. Replace year with gdpPercap

**Colour the continents** 

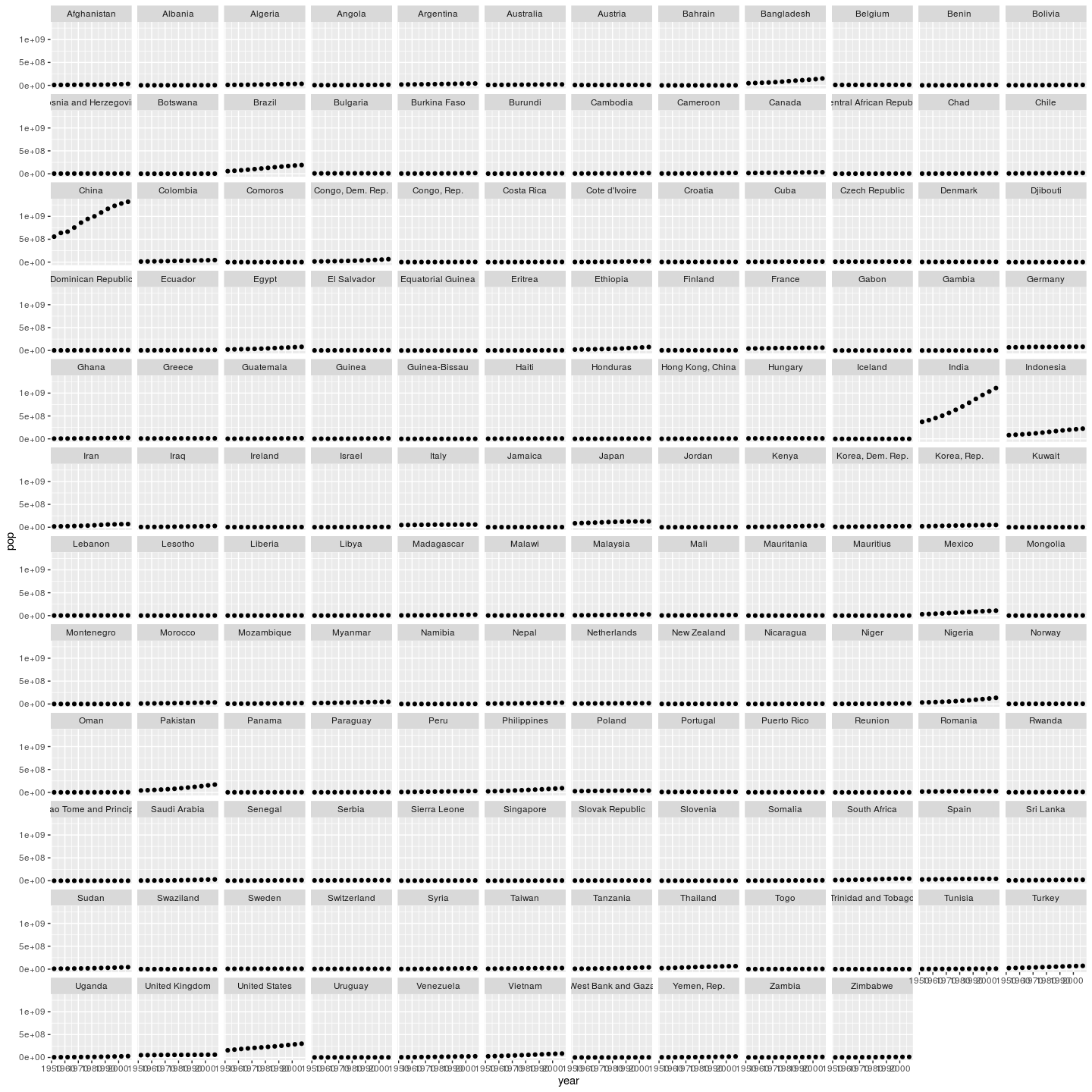
YOUR TURN: Copy the above code and paste in the R chunk below. Then change colorto size and run it.

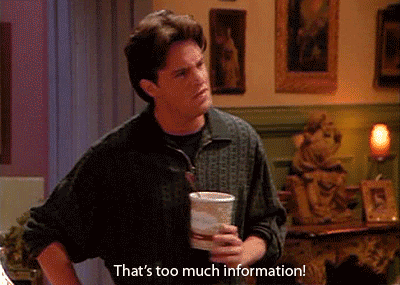


**Separate the continents using facets** 

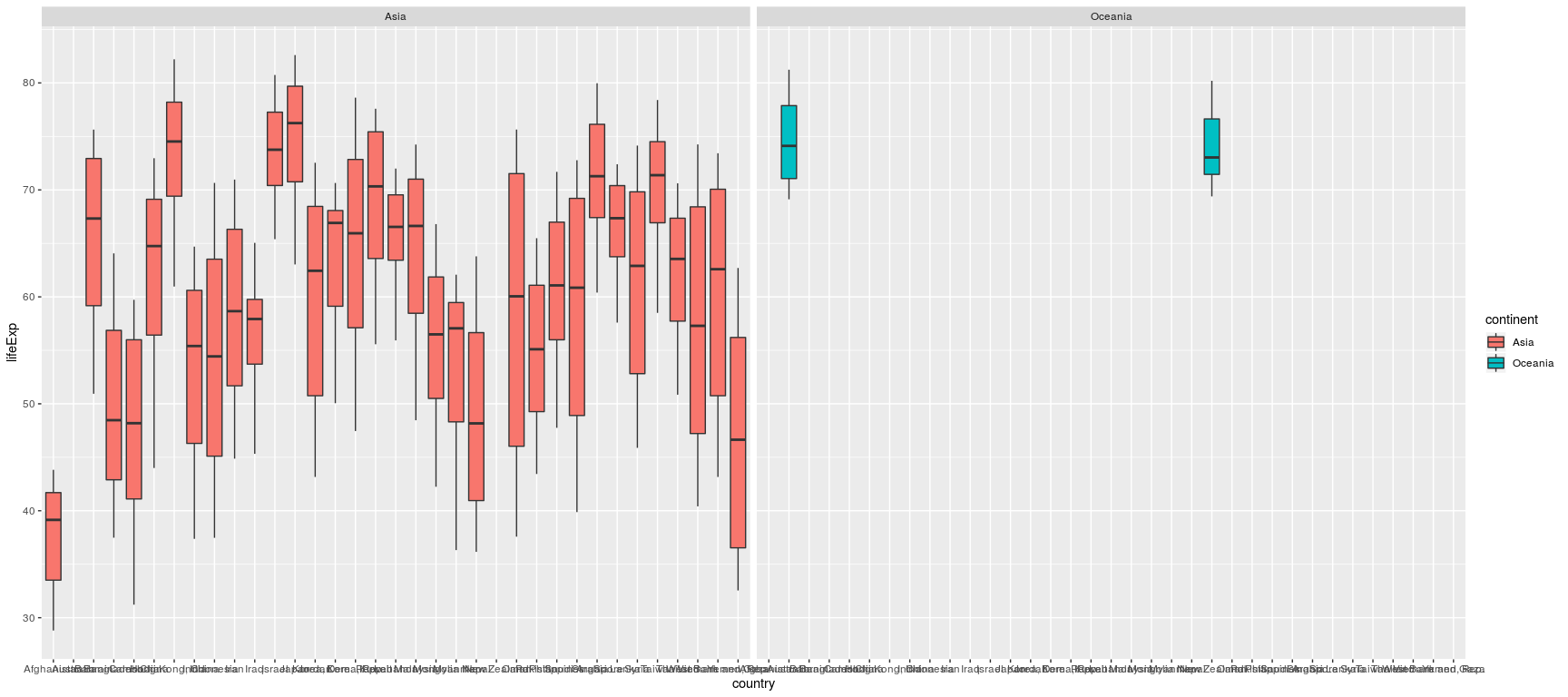
**Change scales** 

**Can we create a facet for each country?** > YES





YOUR TURN: Create a facted plot for each year. Use x = gdpPercap, y = lifeExp and color = continent. You can use the above code to start.

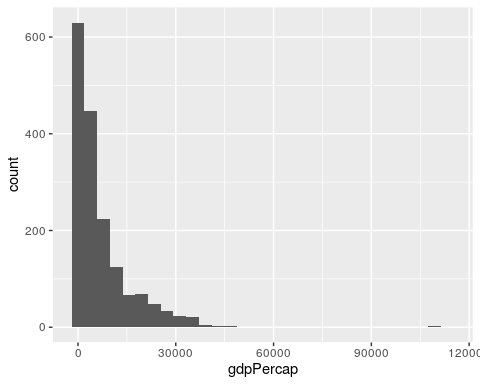




# Transforming a distribution

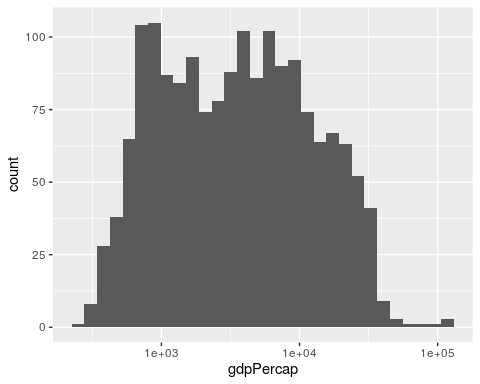
## Original

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

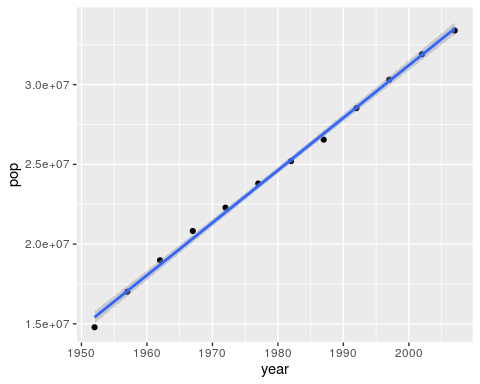


## Transformed

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



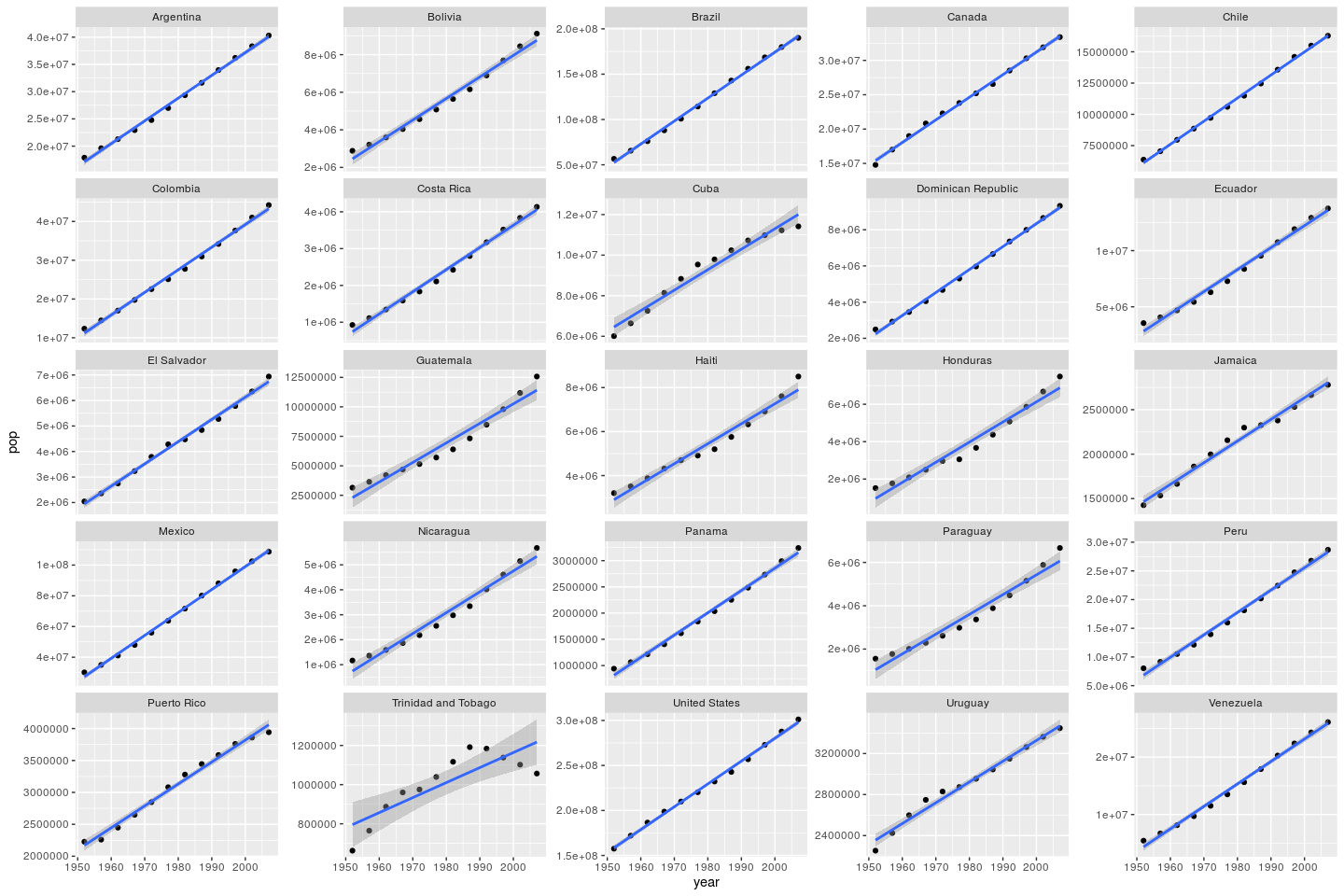
# Plotting a linear model for Canada



### Let’s try 1 continent (Americas)

## [1] Asia Europe Africa Americas Oceania   
## Levels: Africa Americas Asia Europe Oceania

**Create a dataframe for Americas**

**Plot** 

YOUR TURN: Do a similar plot like above for Oceania

YOUR TURN: Using the 2007 data set (created below), plot the life expectancy as a function of GDP. Color each continent and also use size = pop.

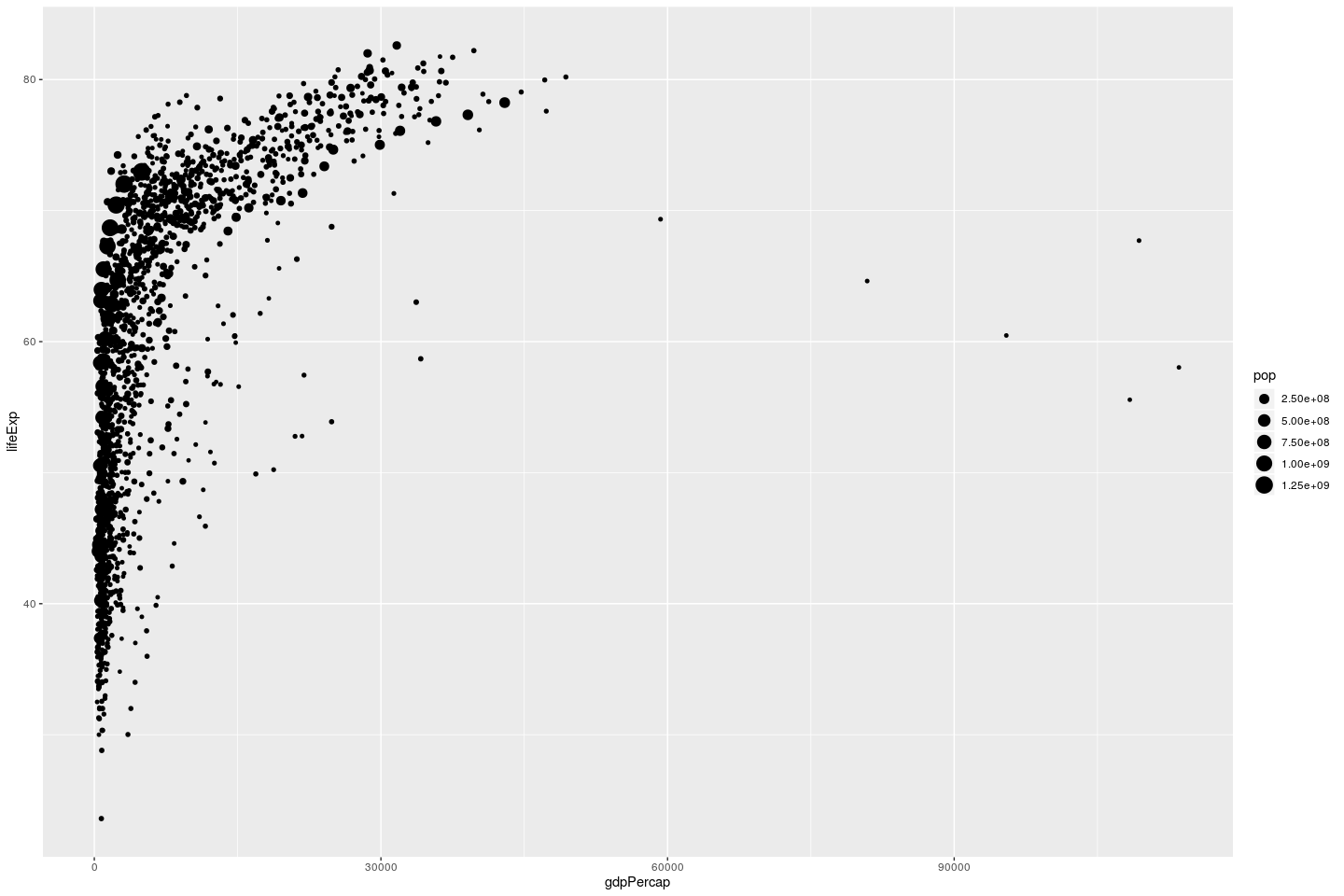
**Data**

**Plot**

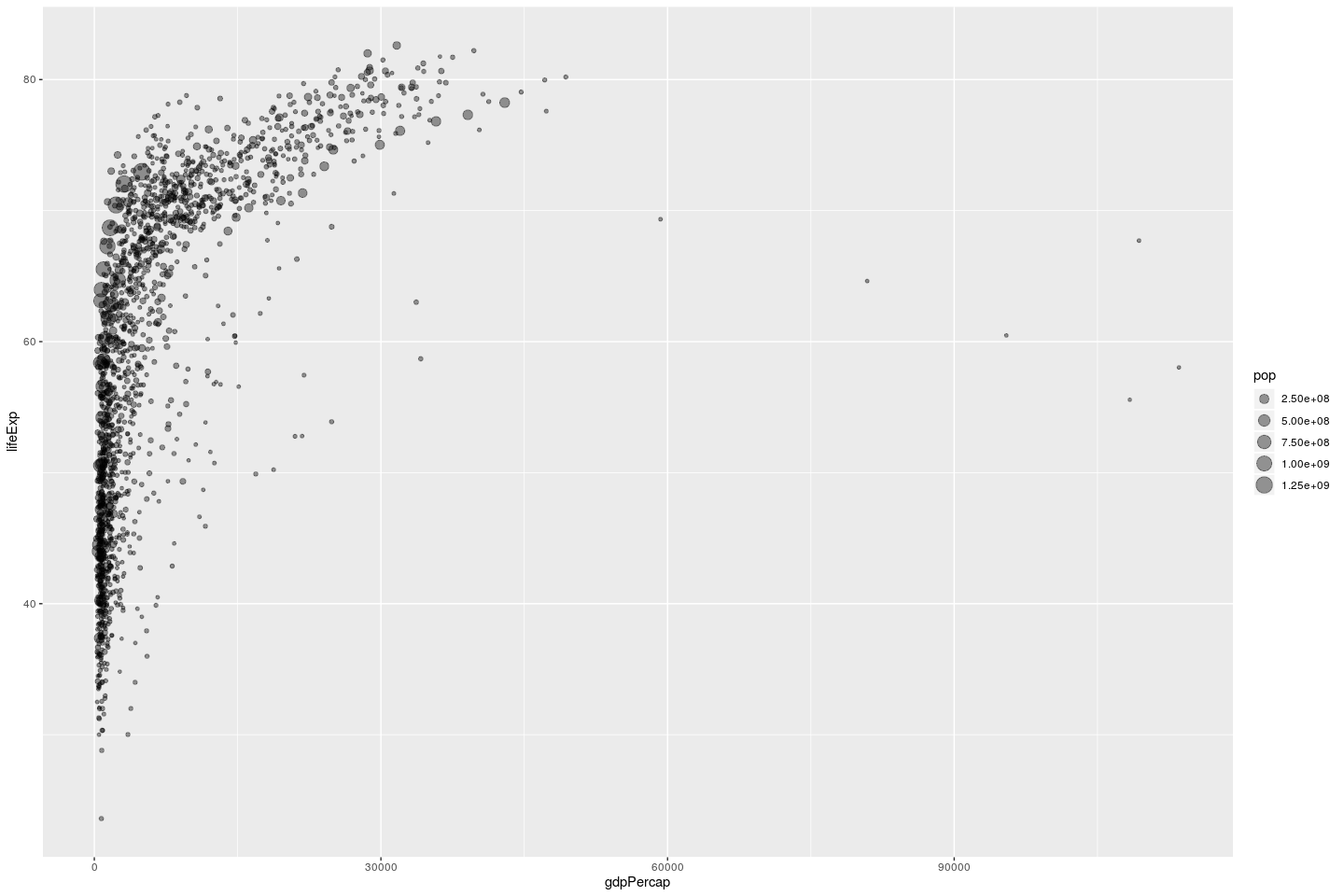


# Improving plot step-by-step

## plot everything



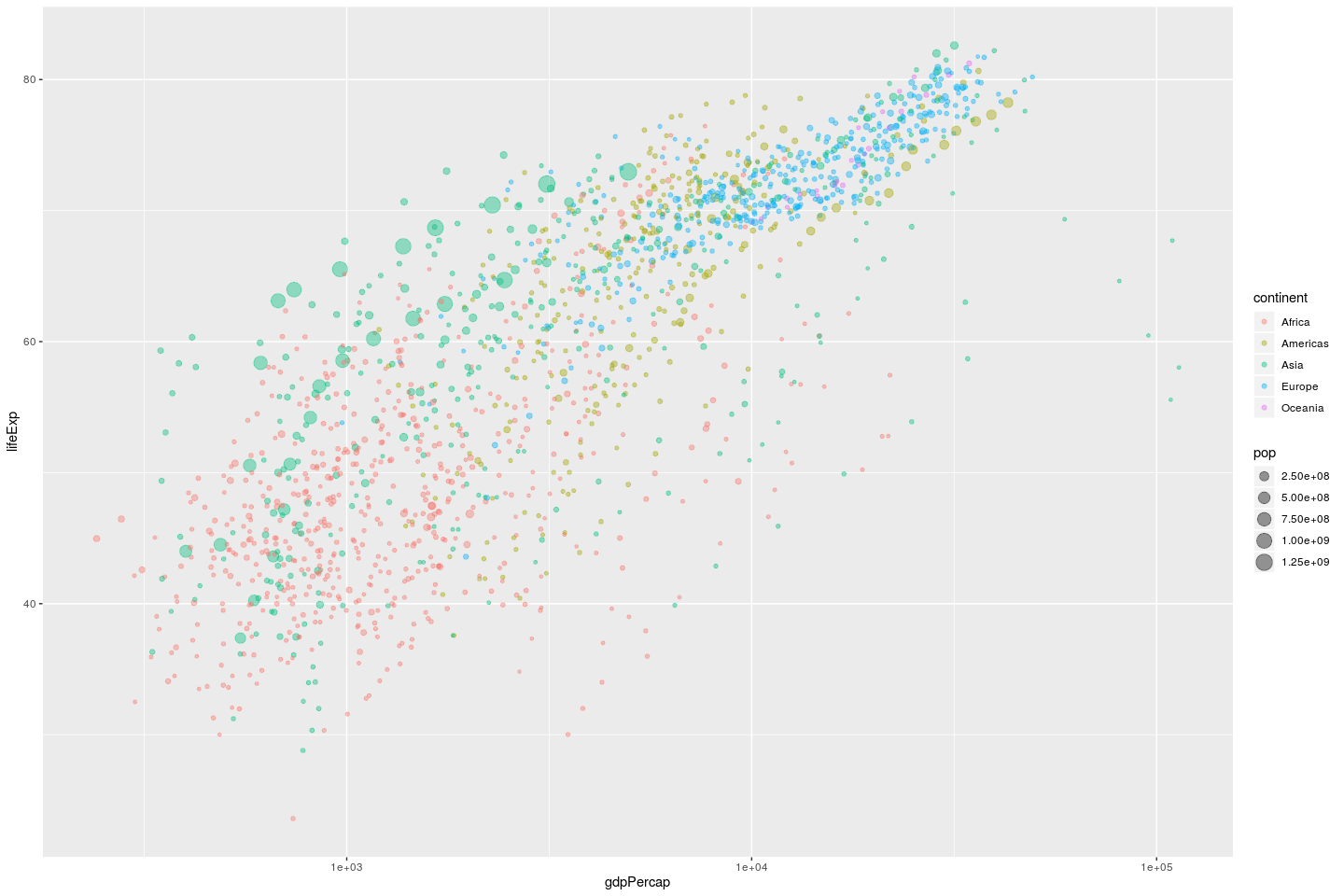
## increase point transparency



## add color



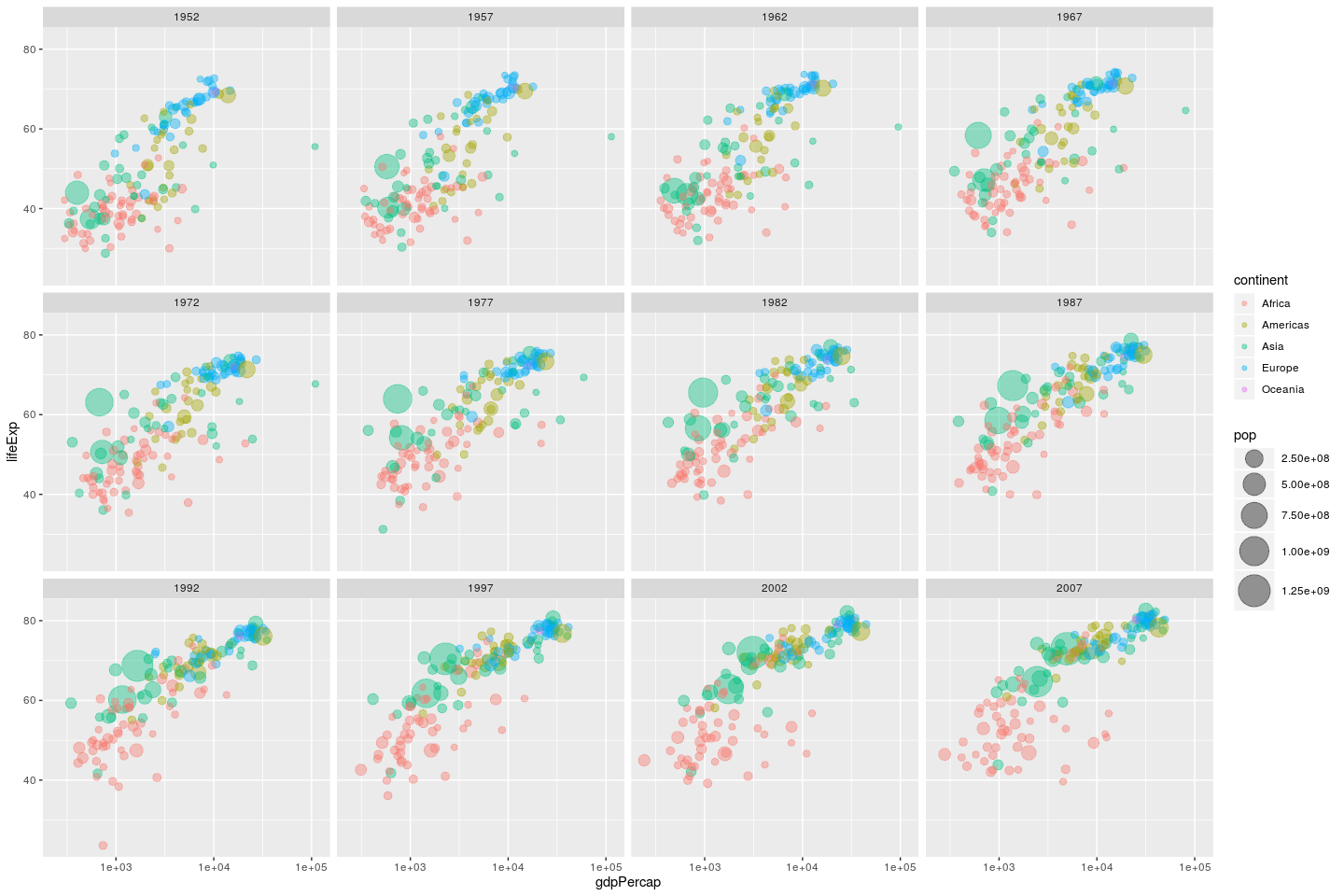
## transform



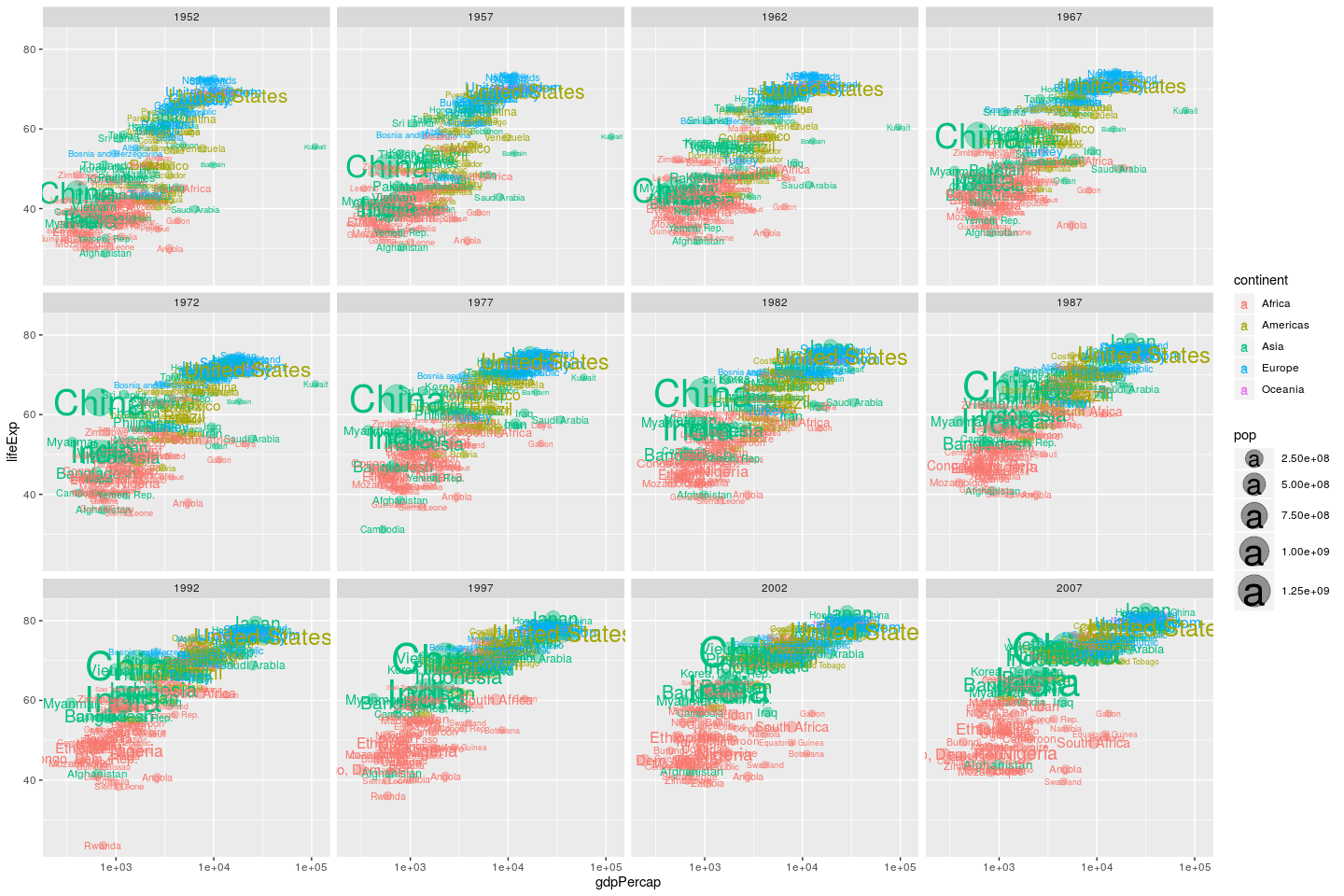
## facet by year



## improve point size



## labels





## # A tibble: 142 x 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Afghanistan Asia 1952 28.8 8425333 779.  
## 2 Albania Europe 1952 55.2 1282697 1601.  
## 3 Algeria Africa 1952 43.1 9279525 2449.  
## 4 Angola Africa 1952 30.0 4232095 3521.  
## 5 Argentina Americas 1952 62.5 17876956 5911.  
## 6 Australia Oceania 1952 69.1 8691212 10040.  
## 7 Austria Europe 1952 66.8 6927772 6137.  
## 8 Bahrain Asia 1952 50.9 120447 9867.  
## 9 Bangladesh Asia 1952 37.5 46886859 684.  
## 10 Belgium Europe 1952 68 8730405 8343.  
## # … with 132 more rows

## create data for labels

### plot



### plot (with ggrepel)





# Cleaning up plot

## add axis labels and title



## default themes





## save plot

# Bonus: Animation