Pick four - comparing trends in population over time

## Purpose

The purpose of this report is to compare the population trends for four countries of your choosing.

## Required Libraries

library(ggplot2)

## Data

* Gapminder data [available here](http://www.gapminder.org/data/). [Gapminder data is licensed CC-BY 3.0](https://docs.google.com/document/pub?id=1POd-pBMc5vDXAmxrpGjPLaCSDSWuxX6FLQgq5DhlUhM#h.ul2gu2-uwathz).
* Processed data via [@jennybc](<https://github.com/jennybc>), [R package available here](https://github.com/jennybc/gapminder). The [data-raw](https://github.com/jennybc/gapminder/tree/master/data-raw) sub-directory reveals the journey from Gapminder.org's Excel workbooks to increasingly clean and tidy data.

**Read in data**: To read in the data, make sure this file is in the same directory/folder as the gapminderDataFiveYear.txv file. To set the proper working directory go to Session > Set Working Directory > To Source File Location.

gapMinder <- read.delim("gapminderDataFiveYear.tsv")   
  
##Check data   
head(gapMinder) #First 10 lines of dataset

## country year pop continent lifeExp gdpPercap  
## 1 Afghanistan 1952 8425333 Asia 28.80 779.4  
## 2 Afghanistan 1957 9240934 Asia 30.33 820.9  
## 3 Afghanistan 1962 10267083 Asia 32.00 853.1  
## 4 Afghanistan 1967 11537966 Asia 34.02 836.2  
## 5 Afghanistan 1972 13079460 Asia 36.09 740.0  
## 6 Afghanistan 1977 14880372 Asia 38.44 786.1

dim(gapMinder) #number of rows and columns in data set

## [1] 1704 6

You can see what countries are available by looking at the how many unique categories are in the country column of the gapMinder dataset.

levels(gapMinder$country)

### Pick Four Countries

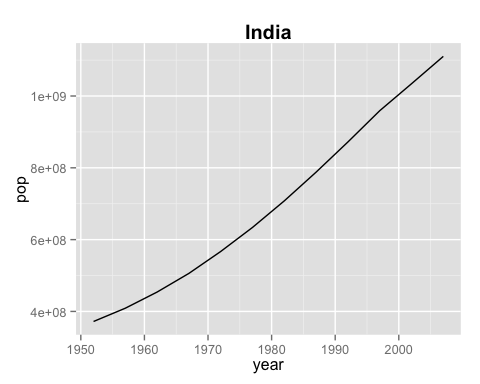
Now pick four countries that you are intrested in. Just replace with the countries name below.

countryName1 <- "India"  
countryName2 <- "United States"  
countryName3 <- "Nigeria"  
countryName4 <- "Germany"

### Country One

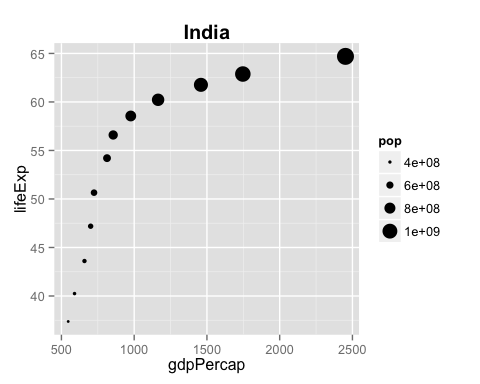
We want to look at how population changes over time for the first country.

country1 <- subset(gapMinder, country == countryName1)  
  
ggplot(country1, aes(year, pop, label = country)) +   
 geom\_path() +  
 ggtitle(countryName1) +  
 theme(plot.title = element\_text(size = 15, face = "bold"))



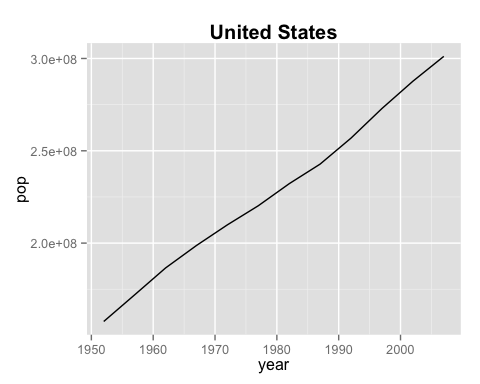
This second graph is looking at the correlation between life expectency (lifeExp) and income per person (gdpPercap). The size of the points on the plot represent.

ggplot(country1, aes(gdpPercap, lifeExp, label = country, size = pop)) +   
 geom\_point() +  
 ggtitle(countryName1) +  
 theme(plot.title = element\_text(size = 15, face = "bold"))

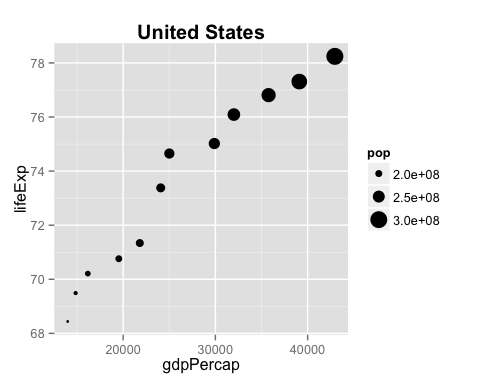


## Country 2

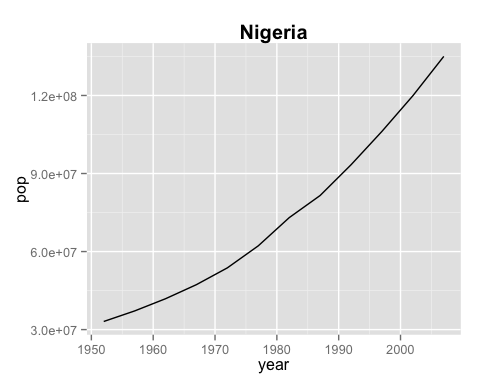
We will do this for each country. Since the code is very similar, we will omit viewing it below.



**Notes**: In a real report you can add in information about the results of the analysis you are performing. That way your code, analysis, questions, and results are all in one place.

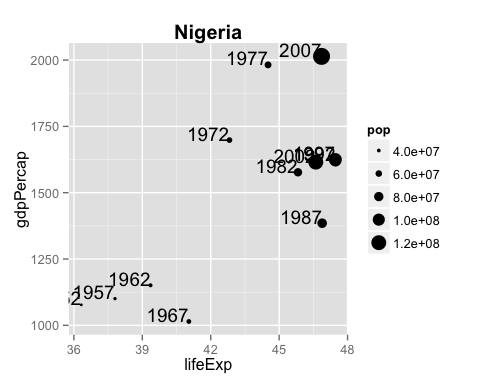


## Country 3

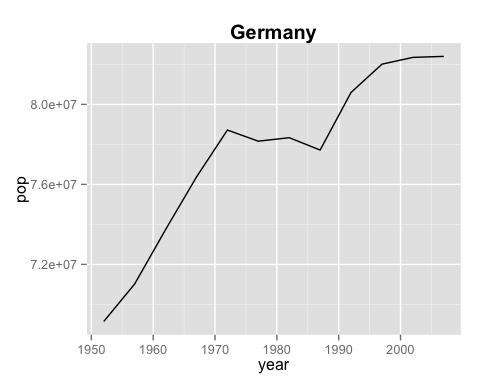


**Notes** Maybe a country has an unusual distribution and we want to label the graph with the year. So we add a bit of code and label with year.

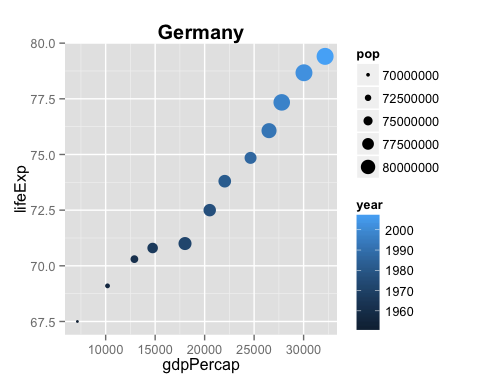
ggplot(country3, aes(lifeExp, gdpPercap, size = pop, label = year)) +   
 geom\_point() +  
 geom\_text(hjust = 1, vjust = 0, size = 5) +  
 ggtitle(countryName3) +  
 theme(plot.title = element\_text(size = 15, face = "bold"))



## Country 4



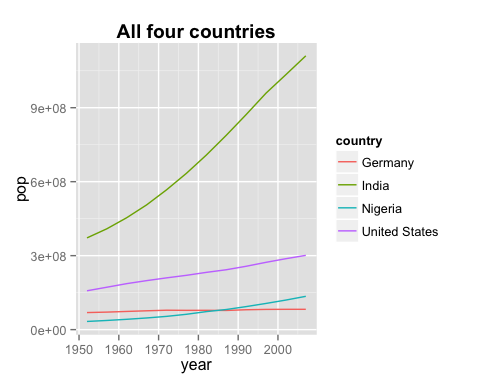
**Notes**: Or maybe try out labeling the year by adding color.



# All four countries

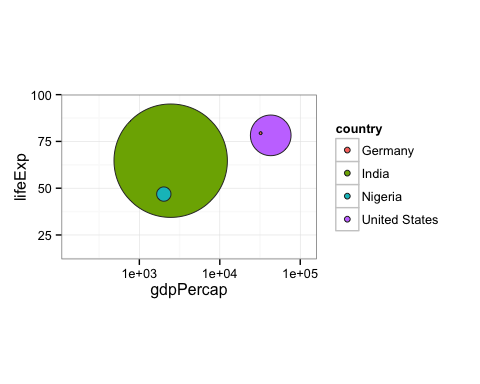
Let's add all four countries together and to see how they compare.

#Add subsetted data together  
allCountries <- rbind(country1, country2, country3, country4)  
  
#Notice the code for this is similar to when   
#we are just looking at one country just   
#with the added the color option  
ggplot(allCountries, aes(year, pop, color=country)) +   
 geom\_path() +  
 ggtitle("All four countries") +  
 theme(plot.title = element\_text(lineheight=.8, face = "bold"))



What about what is occuring in a particular year? You can change the year by changing the code in the year == 2007 section. To look at what years are possible use allCountries$year.

ggplot(subset(allCountries, year == 2007),  
 aes(x = gdpPercap, y = lifeExp, fill = country, size=sqrt(pop/pi))) +  
 scale\_x\_log10(limits = c(150, 115000)) +   
 ylim(c(16, 96)) +  
 geom\_point(pch = 21, color = 'grey20') +   
 scale\_size\_continuous(range=c(1,40)) +  
 coord\_fixed(ratio = 1/43) +  
 theme\_bw() +  
 guides(size=FALSE) +  
 theme(strip.text = element\_text(size = rel(1.1)))



You can plot all the years at once also!

ggplot(allCountries,  
 aes(x = gdpPercap, y = lifeExp, fill = country, size = sqrt(pop/pi))) +  
 scale\_x\_log10(limits = c(150, 115000)) +   
 ylim(c(16, 96)) +  
 geom\_point(pch = 21, color = 'grey20', alpha = 0.8) +   
 scale\_size\_continuous(range = c(1,40)) +  
 coord\_fixed(ratio = 1/43) +  
 theme\_bw() +  
 guides(size = FALSE) +  
 theme(strip.text = element\_text(size = rel(1.1)))

