Homework 1: Insights on Poverty

February 4, 2016

This homework is due Sunday February 14, 2016 at 11:59 PM. When complete, submit your code in the R Markdown file and the knitted HTML file on Canvas.

Background

This HW is based on Hans Rosling talks New Insights on Poverty and The Best Stats You've Ever Seen.

The assignment uses data to answer specific question about global health and economics. The data contradicts commonly held preconceived notions. For example, Hans Rosling starts his talk by asking: (paraphrased) "for each of the six pairs of countries below, which country do you think had the highest child mortality in 2015?"

1. Sri Lanka or Turkey
2. Poland or South Korea
3. Malaysia or Russia
4. Pakistan or Vietnam
5. Thailand or South Africa

Most people get them wrong. Why is this? In part it is due to our preconceived notion that the world is divided into two groups: the *Western world* versus the *third world*, characterized by "long life,small family" and "short life, large family" respectively. In this homework we will use data visualization to gain insights on this topic.

Problem 1

The first step in our analysis is to download and organize the data. The necessary data to answer these question is available on the gapminder website.

Problem 1.1

We will use the following datasets:

1. Childhood mortality
2. Life expectancy
3. Fertility
4. Population
5. Total GDP

Create five `tbl_df` table objects, one for each of the tables provided in the above files. Hints: Use the `read_csv` function. Because these are only temporary files, give them short names.

```r
library(tidyr)
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(readr)
library(ggplot2)
theme_set(theme_bw()) ##optional

t1 <-read_csv("http://spreadsheets.google.com/pub?key=0ArfEDsV3bBwCcGhBd2NOQVZ1eWowNVpSNjl1c3lRSWc&outpu
t2 <- read_csv("http://spreadsheets.google.com/pub?key=phAwcNAVuyj2tPLxKvvnNPA&output=csv")
t3 <- read_csv("http://spreadsheets.google.com/pub?key=phAwcNAVuyj0TAlJeCEzcGQ&output=csv")
t4 <- read_csv("http://spreadsheets.google.com/pub?key=phAwcNAVuyj0XOoBL_n5tAQ&output=csv")
t5 <- read_csv("http://spreadsheets.google.com/pub?key=pyj6tScZqmEfI4sLVvEQtHw&output=csv")
### To keep track of what these are we will define a vector of values
value_names <- c("child_mortality","life_expectancy","fertility","population","gdp")
```

Problem 1.2

Write a function called my_func that takes a table as an argument and returns the column name. For each of the five tables, what is the name of the column containing the country names? Print out the tables or look at them with View to determine the column.

```
my_func <- function(x){
  names(x)[1]
}

answer <- c( my_func(t1), my_func(t2), my_func(t3), my_func(t4), my_func(t5) )

answer
```

```
## [1] "Under five mortality"
## [2] "Life expectancy with projections. Yellow is IHME"
## [3] "Total fertility rate"
## [4] "Total population"
## [5] "GDP (constant 2000 US$)"
```

Problem 1.3

In the previous problem we noted that gapminder is inconsistent in naming their country column. Fix this by assigning a common name to this column in the various tables.

```
the_name <- "country"
names(t1)[1] <- the_name
names(t2)[1] <- the_name
names(t3)[1] <- the_name
names(t4)[1] <- the_name
names(t5)[1] <- the_name
```

Problem 1.4

Notice that in these tables, years are represented by columns. We want to create a tidy dataset in which each row is a unit or observation and our 5 values of interest, including the year for that unit, are in the columns. The unit here is a country/year pair and each unit gets values:

```
value_names
```

```
## [1] "child_mortality" "life_expectancy" "fertility"       "population"
## [5] "gdp"
```

We call this the *long* format. Use the `gather` function from the `tidyr` package to create a new table for childhood mortality using the long format. Call the new columns `year` and `child_mortality`

```
gather(t1, key=year, value=child_mortality, -country, convert = TRUE)
```

```
## Source: local data frame [59,400 x 3]
##
##                  country  year child_mortality
##                    (chr) (int)           (dbl)
## 1               Abkhazia  1800              NA
## 2            Afghanistan  1800          468.58
## 3  Akrotiri and Dhekelia  1800              NA
## 4                Albania  1800          375.20
## 5                Algeria  1800          460.21
## 6         American Samoa  1800              NA
## 7                Andorra  1800              NA
## 8                 Angola  1800          485.68
## 9               Anguilla  1800              NA
## 10   Antigua and Barbuda  1800          473.60
## ..                   ...   ...             ...
```

Now redefine the remaining tables in this way.

```
t1 <- gather(t1, key=year, value=child_mortality, -country, convert = TRUE)
t2 <- gather(t2, key=year, value=life_expectancy, -country, convert = TRUE)
t3 <- gather(t3, key=year, value=fertility, -country, convert = TRUE)
t4 <- gather(t4, key=year, value=population, -country, convert = TRUE)
t5 <- gather(t5, key=year, value=gdp, -country, convert = TRUE)
```

Problem 1.5

Now we want to join all these files together. Make one consolidated table containing all the columns

```
dat <- t1 %>% full_join(t2) %>% full_join(t3) %>% full_join(t4) %>% full_join(t5)
```

```
## Joining by: c("country", "year")
## Joining by: c("country", "year")
## Joining by: c("country", "year")
## Joining by: c("country", "year")
```

Problem 1.6

Add a column to the consolidated table containing the continent for each country. Hint: We have created a file that maps countries to continents here. Hint: Learn to use the `left_join` function.

```
map <- read_delim("https://raw.githubusercontent.com/datasciencelabs/data/master/homework_data/continen
dat <- left_join(dat, map, continent = continents)
```

```
## Joining by: "country"
```

Advanced solution:

```
t1 <-read_csv("http://spreadsheets.google.com/pub?key=0ArfEDsV3bBwCcGhBd2NOQVZ1eWowNVpSNjl1c3lRSWc&outpu
t2  <- read_csv("http://spreadsheets.google.com/pub?key=phAwcNAVuyj2tPLxKvvnNPA&output=csv")
t3  <- read_csv("http://spreadsheets.google.com/pub?key=phAwcNAVuyj0TAlJeCEzcGQ&output=csv")
t4 <- read_csv("http://spreadsheets.google.com/pub?key=phAwcNAVuyj0XOoBL_n5tAQ&output=csv")
t5 <- read_csv("http://spreadsheets.google.com/pub?key=pyj6tScZqmEfI4sLVvEQtHw&output=csv")
### To keep track of what these are we will define a vector of values
value_names <- c("child_mortality","life_expectancy","fertility","population","gdp")
l <- list(t1,t2,t3,t4,t5)
rm(t1,t2,t3,t4,t5)
gc()
fix_table <- function(tab, val_name){
  names(tab)[1] <- "country"
  tab <- gather(tab, key=year, value=y, -country, convert = TRUE)
  names(tab)[which(names(tab)=="y")] <- val_name
  return(tab)
}
for(i in seq_along(l) ) l[[i]] <-fix_table(l[[i]], value_names[i])
dat <- Reduce(full_join, l)

dat <- left_join(dat, map, continent = continents)
```

Problem 2

Report the child mortalilty rate in 2015 for these 5 pairs:

1. Sri Lanka or Turkey
2. Poland or South Korea
3. Malaysia or Russia
4. Pakistan or Vietnam
5. Thailand or South Africa

```r
countries <- c("Sri Lanka","Turkey", "Poland", "South Korea", "Malaysia", "Russia","Pakistan","Vietnam"
answer <- filter(dat, country%in%countries & year==2015) %>% select(country, child_mortality)
### optionally we can plot them next to each other
answer <- bind_cols( slice( answer, match(countries[ seq(1,9,2)], country)),
           slice( answer, match(countries[ seq(2,10,2)], country)))
answer
```

```
## Source: local data frame [5 x 4]
##
##      country child_mortality      country child_mortality
##        (chr)           (dbl)        (chr)           (dbl)
## 1 Sri Lanka             8.7       Turkey            13.5
## 2    Poland             5.2  South Korea             3.5
## 3  Malaysia             8.2       Russia             9.6
## 4  Pakistan            81.1      Vietnam            21.7
## 5  Thailand            12.3 South Africa            42.1
```

Problem 3

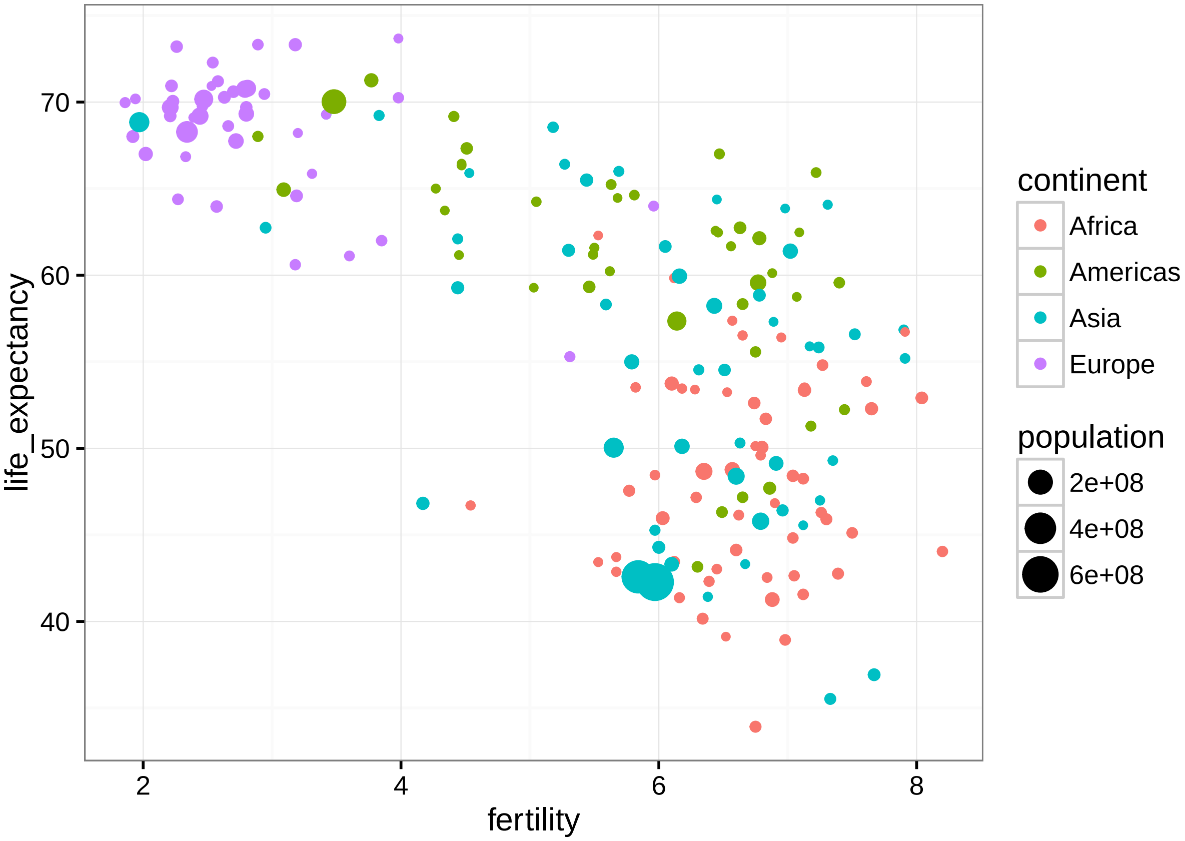
To examine if in fact there was a long-life-in-a-small-family and short-life-in-a-large-family dichotomy, we will visualize the average number of children per family (fertility) and the life expectancy for each country.

Problem 3.1

Use `ggplot2` to create a plot of life expectancy versus fertility for 1962 for Africa, Asia, Europe, and the Americas. Use color to denote continent and point size to denote population size:

```r
p <- ggplot( filter(dat, year==1962 & continent != "Oceania"), aes(x=fertility, y=life_expectancy))
p + geom_point( aes( color=continent, size=population))
```

```
## Warning: Removed 44 rows containing missing values (geom_point).
```



Do you see a dichotomy? Explain.

Problem 3.2

Now we will annotate the plot to show different types of countries.

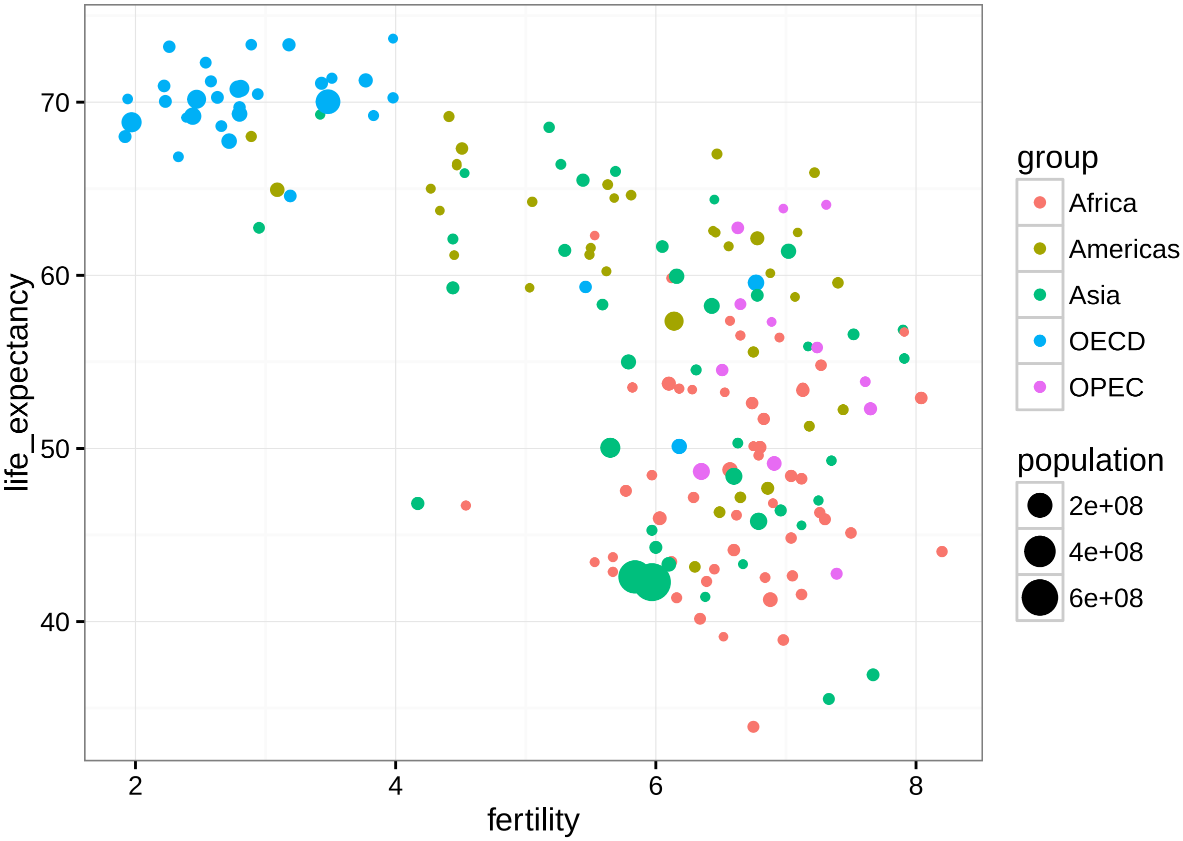
Learn about OECD and OPEC. Add a couple of columns to your consolidated tables containing a logical vector that tells if a country is OECD and OPEC respectively. It is ok to base membership on 2015.

```
oecd <- c("Australia","Austria","Belgium","Canada","Chile","Country","Czech Republic","Denmark","Estoni
          "Iceland","Ireland","Israel","Italy","Japan","Korea","Luxembourg","Mexico","Netherlands","New
          "Slovak Republic","Slovenia","Spain","Sweden","Switzerland","Turkey","United Kingdom","United

opec <- c("Algeria", "Angola", "Ecuador", "Iran", "Iraq", "Kuwait", "Libya","Nigeria",
        "Qatar", "Saudi Arabia", "United Arab Emirates", "Venezuela")

dat <- mutate(dat, oecd = country %in% oecd, opec = country %in% opec)
```

Problem 3.3

Make the same plot as in Problem 3.1, but this time use color to annotate the OECD countries and OPEC countries. For countries that are not part of these two organization annotate if they are from Africa, Asia, or the Americas.

```
group <- ifelse( dat$country%in%oecd , "OECD", dat$continent  )
group <- ifelse( dat$country%in%opec , "OPEC", group  )
group <- ifelse( group%in%c("Europe","Oceania"), NA, group)

##add group and avoid NAs
dat2 <- mutate(dat, group=group) %>%
  filter(!is.na(group) & !is.na(population) & !is.na(fertility) & !is.na(life_expectancy))

p <- ggplot( mutate(dat2, group = group) %>% filter(year==1962),
            aes(x=fertility, y=life_expectancy))
p + geom_point( aes( color=group, size=population))
```

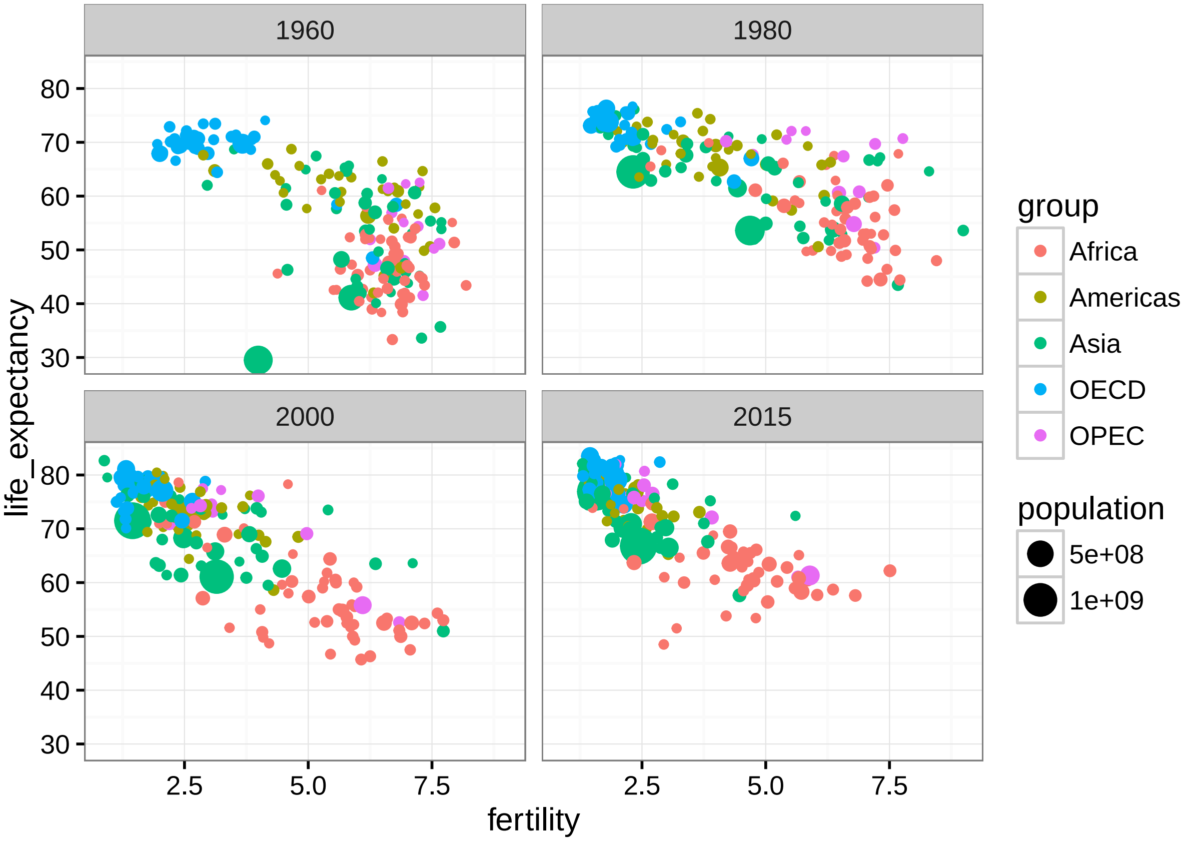


How would you describe the dichotomy?

Problem 3.4

Explore how this figure changes across time. Show us 4 figures that demonstrate how this figure changes through time.

```
years <-  c(1960, 1980, 2000, 2015)
p <- ggplot( mutate(dat2, group = group) %>% filter(year%in%years),
            aes(x=fertility, y=life_expectancy))
p + geom_point( aes( color=group, size=population)) + facet_wrap(~year)
```



Would you say that the same dichotomy exists today? Explain:

Answer: The dichotomy today is not as clear. If there is a dichotomy it is mainly between sub-saharan Africa and the rest of the world. The transition has been relatively smooth. Extra points: You can see the effects of the AIDS epidemic which may explain why life expectancy is down for some countries in 2000 compared to 1980.

Problem 3.5 (Optional)

Make an animation with the `gganimate` package.

```
library(gganimate)
p <- ggplot( filter(dat2, year>1950),
             aes(x=fertility, y=life_expectancy)) + coord_cartesian(ylim = c(30, 85)) +
  geom_point( aes( color=group, size=population, frame=year))
gg_animate(p, "output.mp4")
```

Problem 4

Having time as a third dimension made it somewhat difficult to see specific country trends. Let's now focus on specific countries.

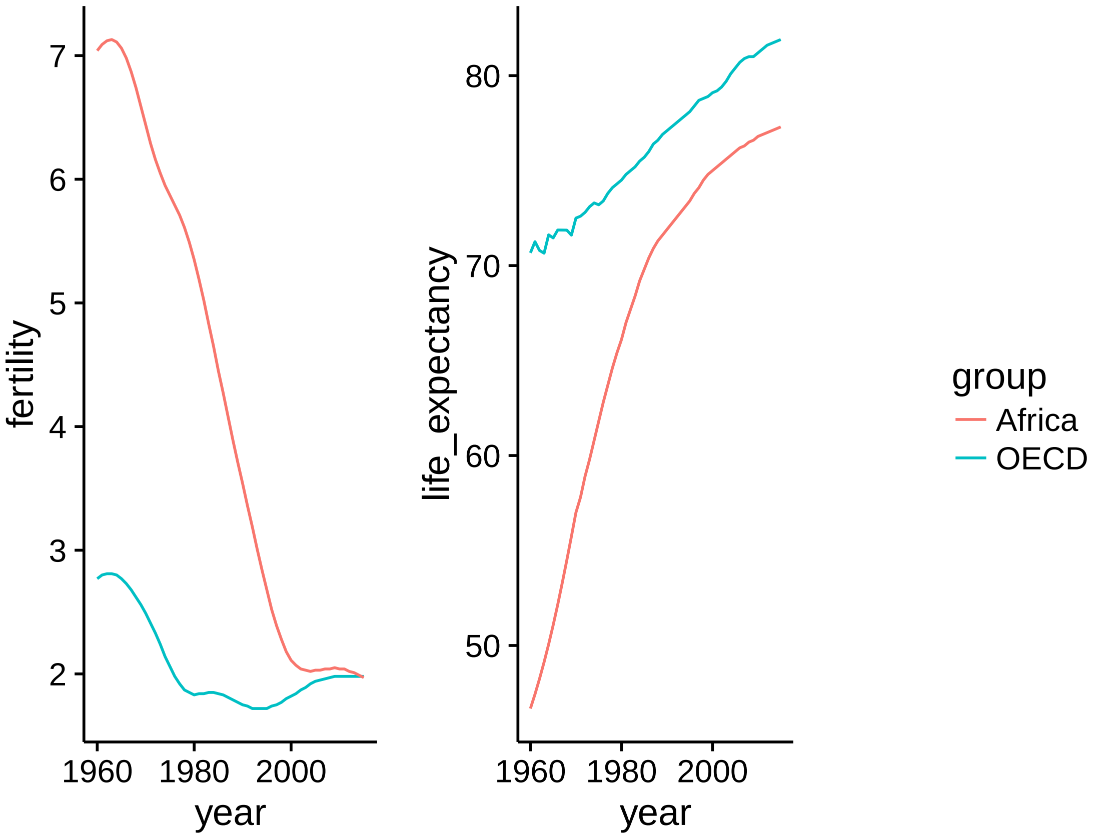
Problem 4.1

Let's compare France and its former colony Tunisia. Make a plot of fertility versus year with color denoting the country. Do the same for life expecancy. How would you compare Tunisia's improvement compared to France's in the past 60 years? Hint: use `geom_line`

```
countries <- c("Tunisia","France")
p1 <- ggplot(filter(dat2, country %in% countries & year > 1959),
             aes(year,fertility,group=country,color=group)) + geom_line() + theme(legend.position="none
p2 <- ggplot(filter(dat2, country %in% countries & year > 1959),
             aes(year,life_expectancy,group=country,color=group)) + geom_line() + theme(legend.position

library(cowplot) ##this is optional
```

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

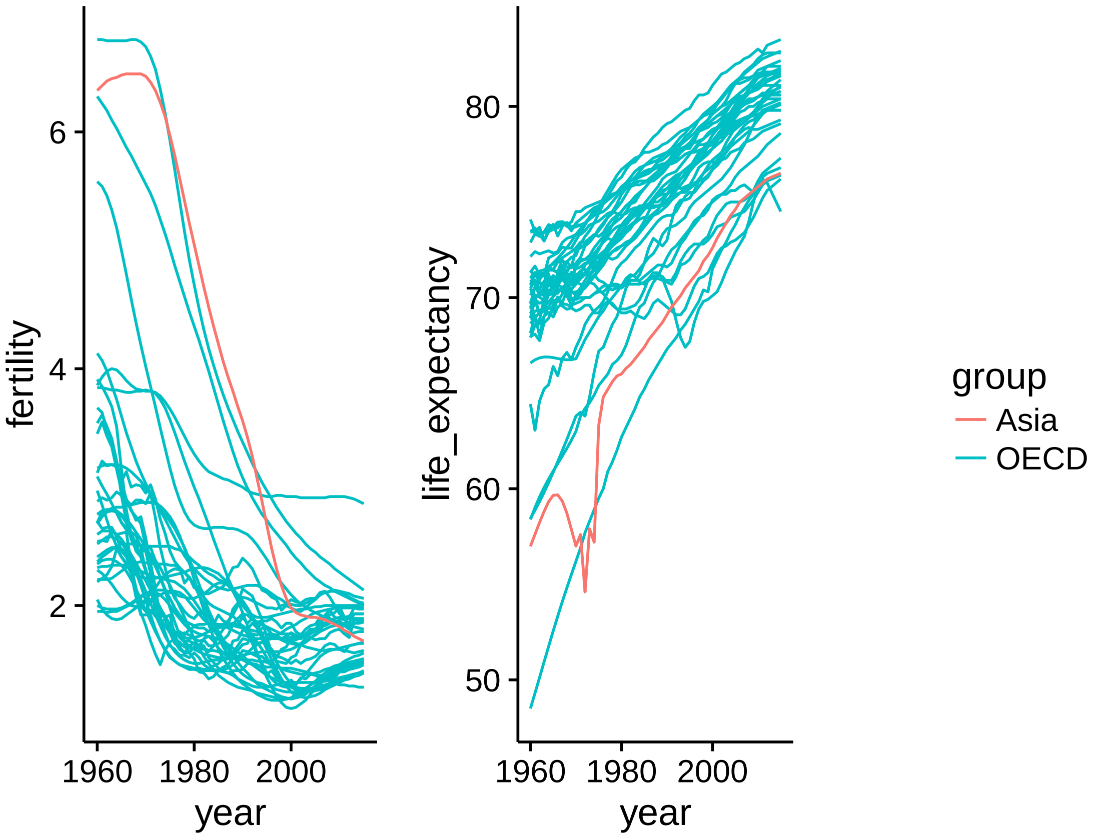
```
g_legend<-function(a.gplot){
  tmp <- ggplot_gtable(ggplot_build(a.gplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)
}
pleg <- ggplot( filter(dat2,country%in%countries & year > 1959), aes(year,fertility,group=country,color
            geom_line()
pleg = g_legend(pleg)

plot_grid(p1, p2, pleg, ncol = 3)
```



Problem 4.2

Do the same, but this time compare Vietnam to the OECD countries.

```
countries <- c("Vietnam",oecd)
p1 <- ggplot( filter(dat2, country%in%countries & year > 1959),
             aes(year,fertility,group=country,color=group)) + geom_line() + theme(legend.position="none
p2 <- ggplot( filter(dat2, country%in%countries & year > 1959),
             aes(year,life_expectancy,group=country,color=group)) + geom_line() + theme(legend.position=
library(cowplot) ##this is optional
g_legend<-function(a.gplot){
  tmp <- ggplot_gtable(ggplot_build(a.gplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)
}
pleg <- ggplot( filter(dat2,country%in%countries & year > 1959), aes(year,fertility,group=country,color=
pleg = g_legend(pleg)

plot_grid(p1, p2, pleg,ncol = 3)
```



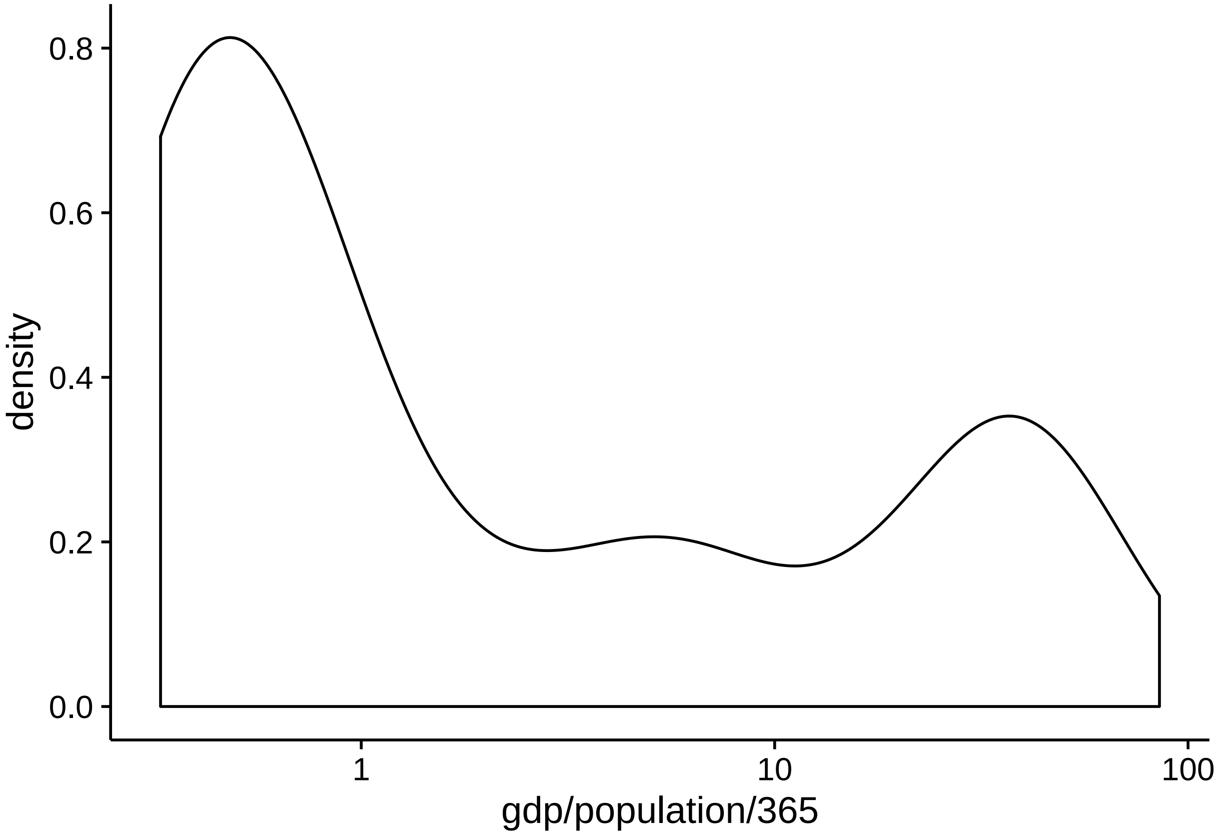
Problem 5

We are now going to examine GDP per capita per day.

Problem 5.1

Create a smooth density estimate of the distribution of GDP per capita per day across countries in 1970. Include OECD, OPEC, Asia, Africa, and the Americas in the computation. When doing this we want to weigh countries with larger populations more. We can do this using the "weight" argument in `geom_density`.

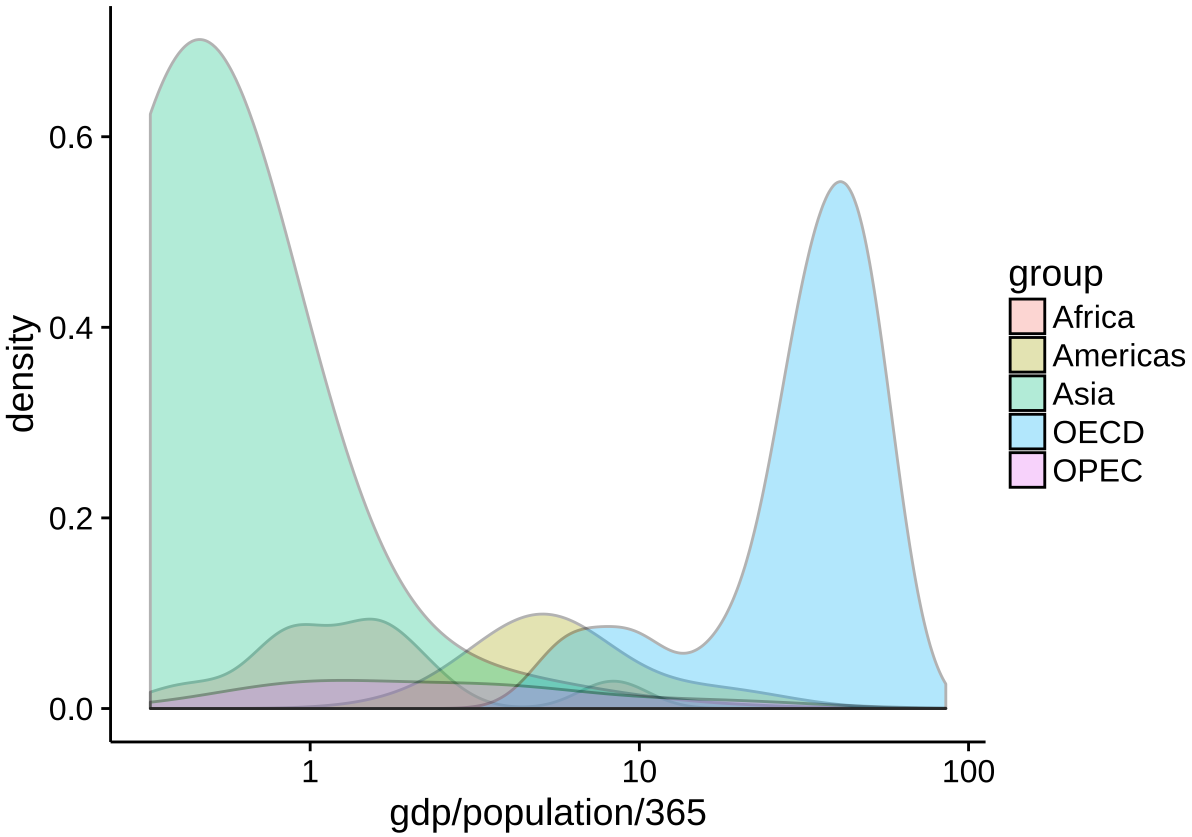
```
dat2 <- mutate(dat, group=group) %>%
  filter(!is.na(group) & !is.na(population) & !is.na(gdp))

ggplot(filter(dat2, year==1970 & group%in%c("OECD","OPEC","Asia","Africa","Americas")),
       aes(x = gdp/population/365)) +
  geom_density(aes(weight=population/sum(population))) + scale_x_log10()
```



Problem 5.2

Now do the same but show each of the five groups separately.

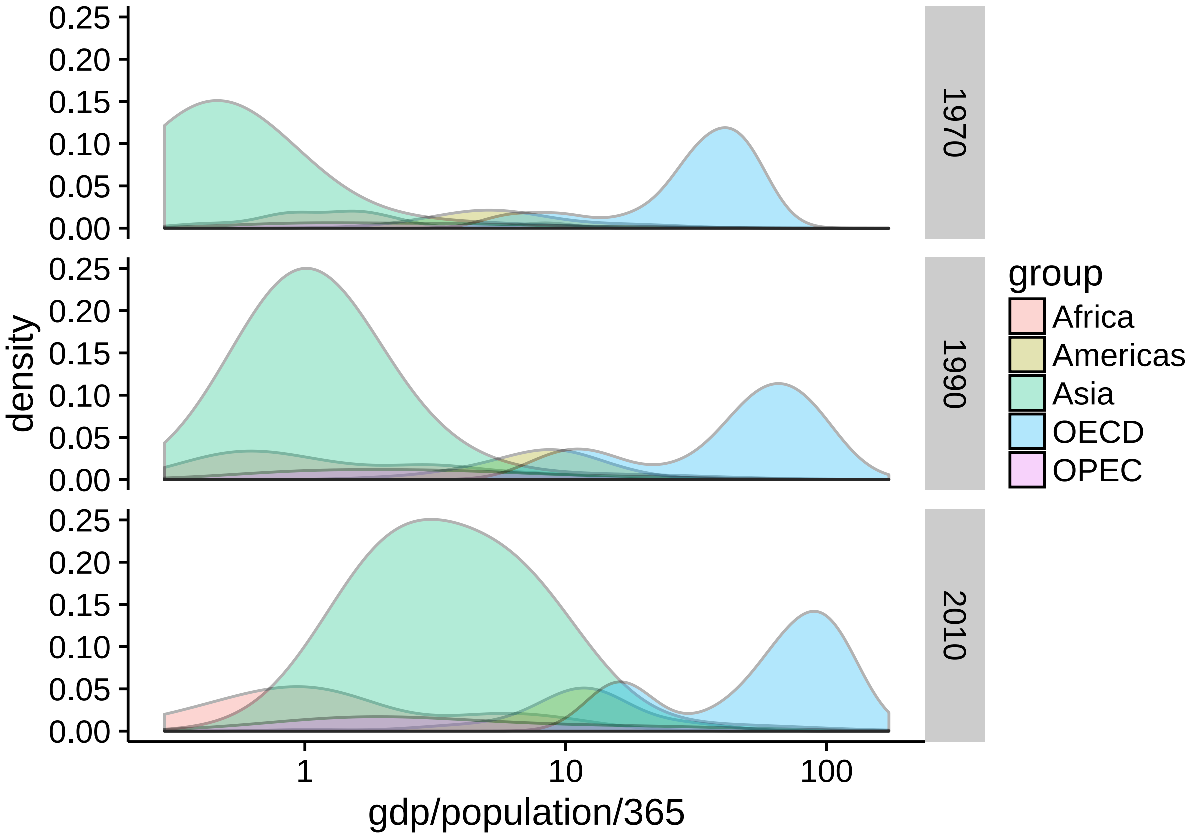
```
dat2 <- mutate(dat, group=group) %>%
  filter(!is.na(group) & !is.na(population) & !is.na(gdp))

filter(dat2, year==1970 & group%in%c("OECD","OPEC","Asia","Africa","Americas")) %>%
  ggplot(aes(x = gdp/population/365)) +
  geom_density(aes(weight=population/sum(population), fill=group), alpha=0.3) + scale_x_log10()
```



Problem 5.3

Visualize these densities for several years. Show a couple of of them. Summarize how the distribution has changed through the years.

```
ggplot(filter(dat2, year%in%c(1970,1990,2010) & group%in%c("OECD","OPEC","Asia","Africa","Americas")),
       aes(x = gdp/population/365)) +
  geom_density(aes(weight=population/sum(population), fill=group), alpha=0.3) + scale_x_log10() +
  facet_grid(year~.)
```



0.25
0.20
0.15
0.10
0.05
0.00
density
1970
1990
2010
group
Africa
Americas
Asia
OECD
OPEC
1
10
100
gdp/population/365