Hw02

Cody

September 22, 2017

```
library(gapminder)
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

Smell test

- Is it a data.frame, a matrix, a vector, a list?

```
str(gapminder)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1704 obs. of  6 variables:
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
##  $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22
##  $ gdpPercap: num  779 821 853 836 740 ...
```

A: Gapminder is a data.frame.

- What's its class?

```
class(gapminder)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

A: It is actually a tibble, a particular kind of data.frame favorable in the tidyverse.

- How many variables/columns?

```
ncol(gapminder)
```

```
## [1] 6
```

A: 6

- How many rows/observations?

```
nrow(gapminder)
```

```
## [1] 1704
```

A: 1704

- Can you get these facts about "extent" or "size" in more than one way? Can you imagine different functions being useful in different contexts?

```
dim(gapminder)
```

```
## [1] 1704    6
```

A: Yes, there are redundancies in some of these functions and theres good reason for it! Perhaps different inputs are useful for one function over another, the speed of computation could be better for a particular data type in one function or even the outputs of the functions can be used in different manors.

- What data type is each variable?

```
head(gapminder)
```

```
## # A tibble: 6 x 6
##       country continent  year lifeExp      pop gdpPercap
##        <fctr>    <fctr> <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan      Asia  1952  28.801  8425333  779.4453
## 2 Afghanistan      Asia  1957  30.332  9240934  820.8530
## 3 Afghanistan      Asia  1962  31.997 10267083  853.1007
## 4 Afghanistan      Asia  1967  34.020 11537966  836.1971
## 5 Afghanistan      Asia  1972  36.088 13079460  739.9811
## 6 Afghanistan      Asia  1977  38.438 14880372  786.1134
```

A: Country and Continent- Factors, Year and Population- Integers, Life Expectancy and GDP per Capita- Double.

Explore individual Variables

Categorical Variable= Country:

- What are possible values (or range, whichever is appropriate) of each variable?

```
summary(gapminder$country)
```

```
##              Afghanistan                  Albania                  Algeria
##                       12                       12                       12
##                   Angola                Argentina                Australia
##                       12                       12                       12
##                  Austria                  Bahrain               Bangladesh
##                       12                       12                       12
##                  Belgium                    Benin                  Bolivia
##                       12                       12                       12
##   Bosnia and Herzegovina                 Botswana                   Brazil
##                       12                       12                       12
##                 Bulgaria             Burkina Faso                  Burundi
##                       12                       12                       12
##                 Cambodia                 Cameroon                   Canada
##                       12                       12                       12
## Central African Republic                     Chad                    Chile
##                       12                       12                       12
##                    China                 Colombia                  Comoros
##                       12                       12                       12
```

```
##         Congo, Dem. Rep.              Congo, Rep.               Costa Rica
##                       12                       12                       12
##           Cote d'Ivoire                  Croatia                     Cuba
##                       12                       12                       12
##          Czech Republic                  Denmark                 Djibouti
##                       12                       12                       12
##       Dominican Republic                  Ecuador                    Egypt
##                       12                       12                       12
##             El Salvador        Equatorial Guinea                  Eritrea
##                       12                       12                       12
##                Ethiopia                  Finland                   France
##                       12                       12                       12
##                   Gabon                   Gambia                  Germany
##                       12                       12                       12
##                   Ghana                   Greece                Guatemala
##                       12                       12                       12
##                  Guinea            Guinea-Bissau                    Haiti
##                       12                       12                       12
##                Honduras         Hong Kong, China                  Hungary
##                       12                       12                       12
##                 Iceland                    India                Indonesia
##                       12                       12                       12
##                    Iran                     Iraq                  Ireland
##                       12                       12                       12
##                  Israel                    Italy                  Jamaica
##                       12                       12                       12
##                   Japan                   Jordan                    Kenya
##                       12                       12                       12
##         Korea, Dem. Rep.              Korea, Rep.                   Kuwait
##                       12                       12                       12
##                 Lebanon                  Lesotho                  Liberia
##                       12                       12                       12
##                   Libya                Madagascar                   Malawi
##                       12                       12                       12
##                Malaysia                     Mali               Mauritania
##                       12                       12                       12
##                Mauritius                   Mexico                 Mongolia
##                       12                       12                       12
##              Montenegro                  Morocco               Mozambique
##                       12                       12                       12
##                 Myanmar                  Namibia                    Nepal
##                       12                       12                       12
##             Netherlands              New Zealand                Nicaragua
##                       12                       12                       12
##                   Niger                  Nigeria                   Norway
##                       12                       12                       12
##                    Oman                 Pakistan                   Panama
##                       12                       12                       12
##                  (Other)
##                     516
```

A: This variable takes on strings as its possible outputs, from above we see that every country has 12 different entries.

- What values are typical? What's the spread? What's the distribution? Etc., tailored to the variable at

hand.

A: Like above, we see that every country has a uniform number of samples so if we consider the distribution of countries it also would be uniformly distributed. Because this variable is categorical a mean and spread don't make a lot of sense to talk about.

- Feel free to use summary stats, tables, figures. We're NOT expecting high production value (yet).

Quantitative Variable= Life Expectancy:

- What are possible values (or range, whichever is appropriate) of each variable?

```
summary(gapminder$lifeExp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.60   48.20   60.71   59.47   70.85   82.60
```

A: Since this is a numeric variable that is on a continuous scale, everything is quite readily available to find. The range is 82.6-23.6=59 years.

- What values are typical? What's the spread? What's the distribution? Etc., tailored to the variable at hand.

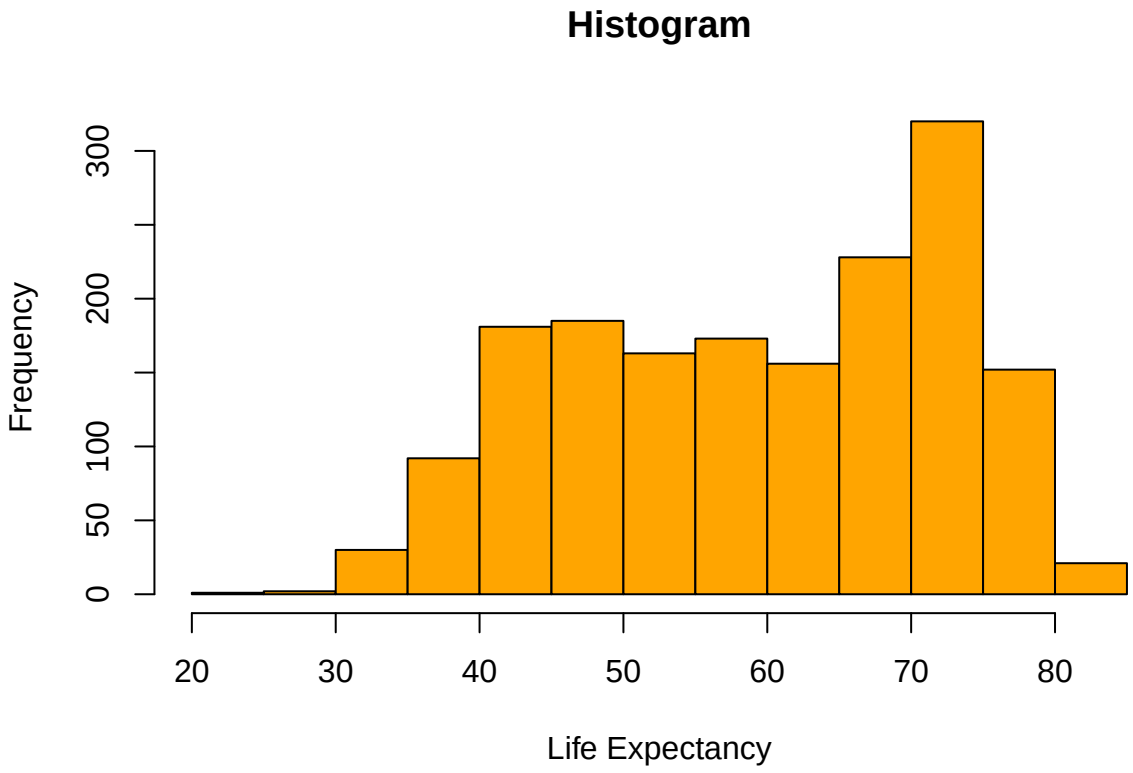
```
mean(gapminder$lifeExp)
```

```
## [1] 59.47444
```

```
sd(gapminder$lifeExp)
```

```
## [1] 12.91711
```

```
hist(gapminder$lifeExp,col="orange",
     main="Histogram",xlab="Life Expectancy")
```



A: The average life expectancy is 60.71 years where its standard deviation is 12.92 years. This tells us that our data is quite spread and we commonly see values in the range of (47,73), one standard deviation away from the mean. From the histogram, we can see a bit of right tailed behavior which indicates that larger values are more probable than smaller. I could also be convinced that this plot indicates a nonsymmetric bimodal structure and thus our mean will not be an accurate measure of center.

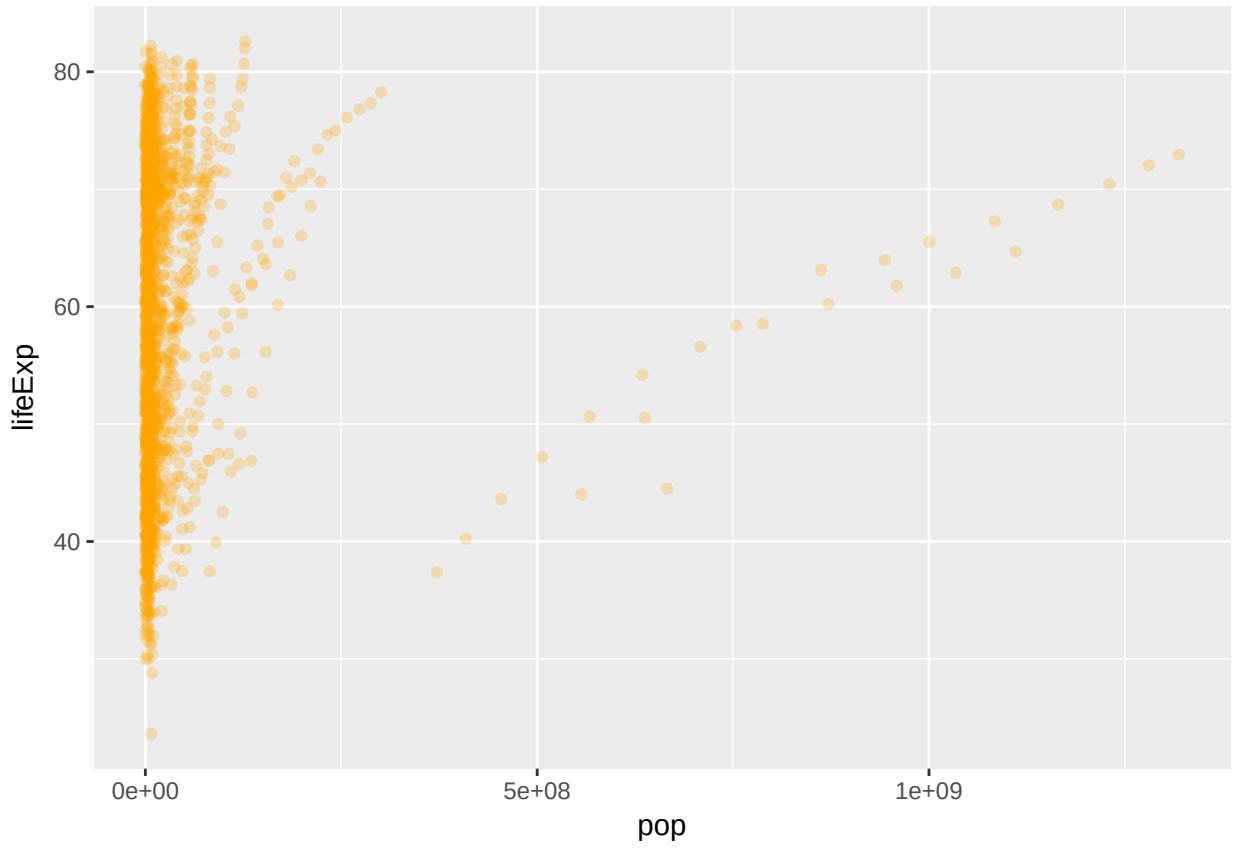
- Feel free to use summary stats, tables, figures. We're NOT expecting high production value (yet).

Explore various plot types

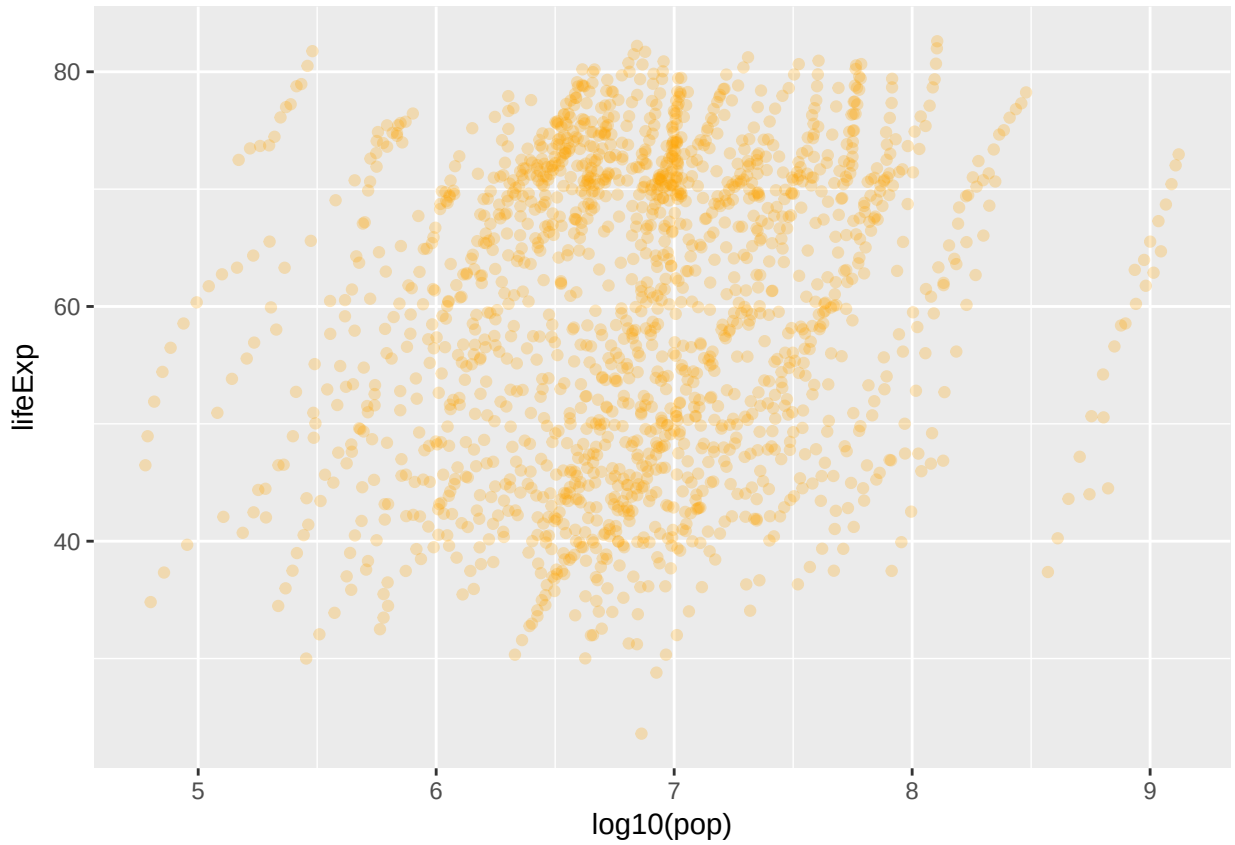
Make a few plots, probably of the same variable you chose to characterize numerically. Try to explore more than one plot type. Just as an example of what I mean:

- A scatterplot of two quantitative variables.

```
q <- ggplot(gapminder,aes(x=pop,y=lifeExp))
q+geom_point(alpha=.25,color="orange")
```

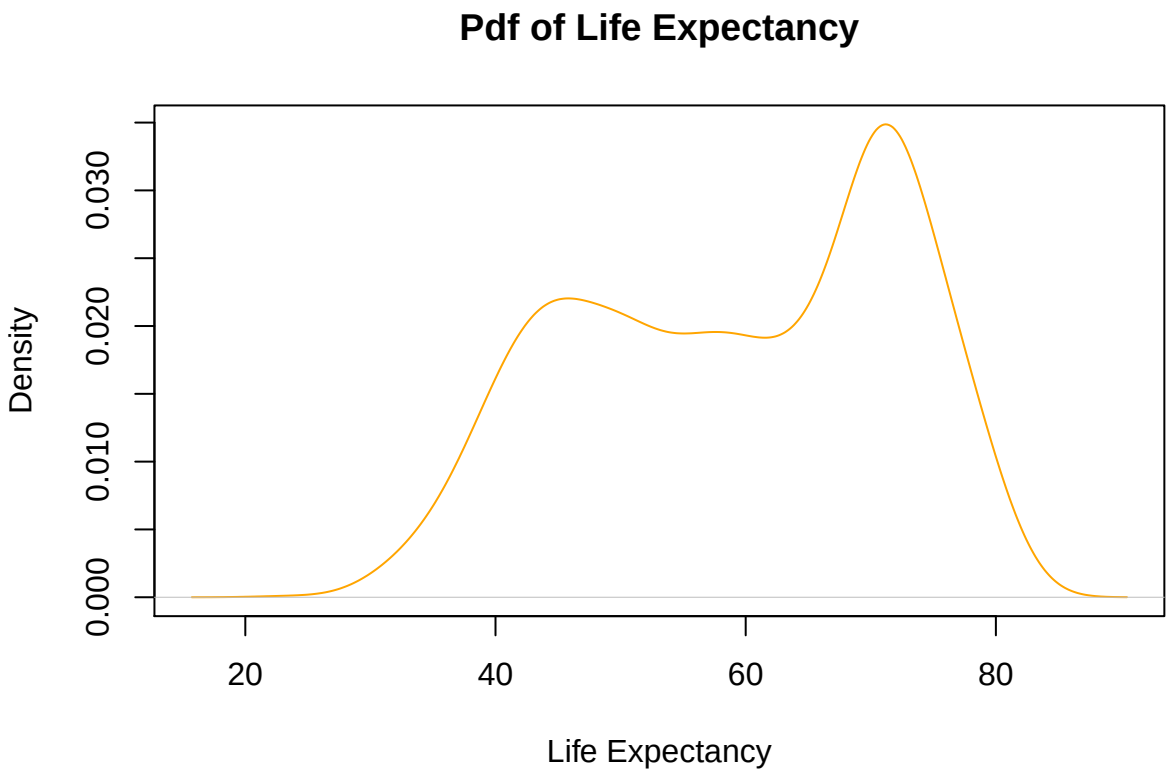



```
p <- ggplot(gapminder,aes(x=log10(pop),y=lifeExp))
p+geom_point(alpha=.25,color="orange")
```



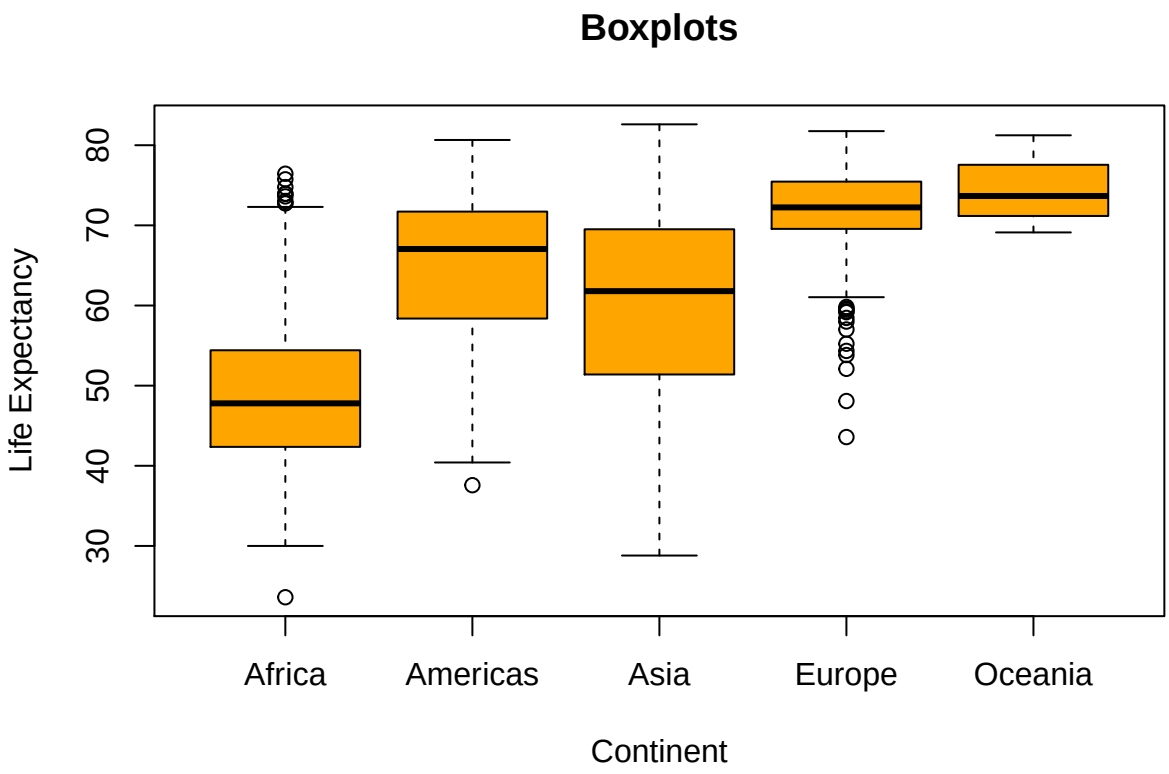
- A plot of one quantitative variable. Maybe a histogram or densityplot or frequency polygon.

```
d <- density(gapminder$lifeExp);
plot(d,col="orange",main="Pdf of Life Expectancy",
     xlab="Life Expectancy")
```



- A plot of one quantitative variable and one categorical. Maybe boxplots for several continents or countries.

```
boxplot(lifeExp~continent,data=gapminder,col="orange",
        xlab="Continent",ylab="Life Expectancy",main="Boxplots")
```



You don't have to use all the data in every plot! It's fine to filter down to one country or small handful of countries.

Use filter(), select() and %>%

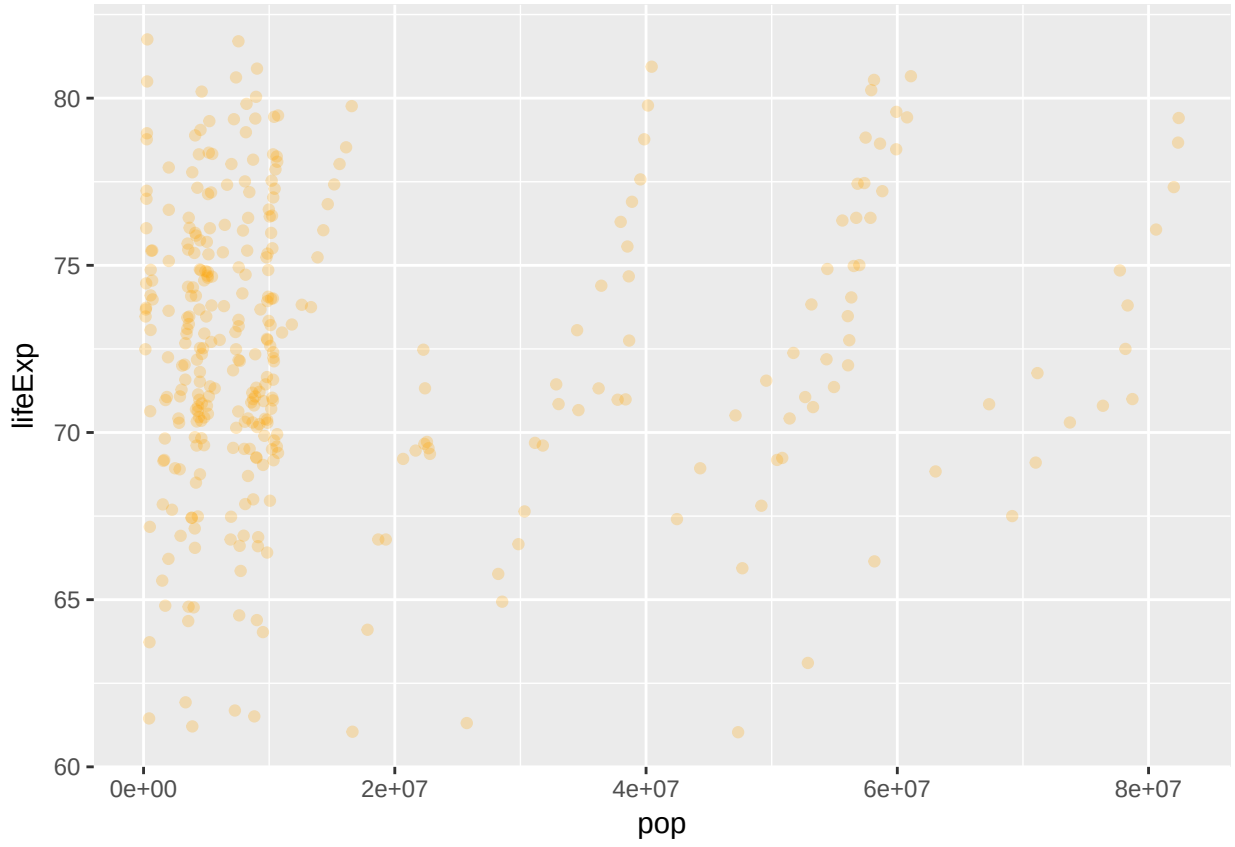
Use filter() to create data subsets that you want to plot.

```
filter(gapminder,continent=="Europe",lifeExp>60)
```

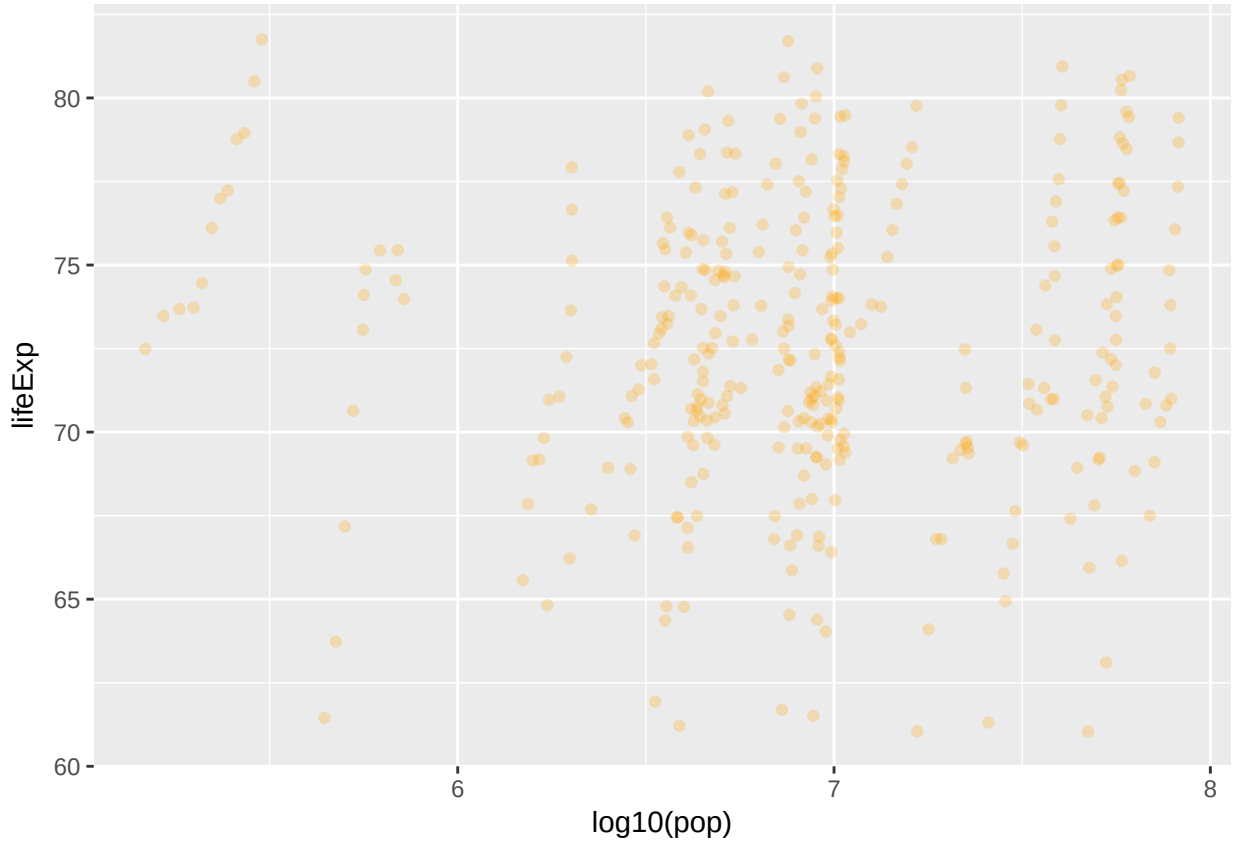
```
## # A tibble: 346 x 6
##    country continent  year lifeExp     pop gdpPercap
##     <fctr>    <fctr> <int>   <dbl>   <int>     <dbl>
##  1 Albania    Europe  1962  64.820 1728137  2312.889
##  2 Albania    Europe  1967  66.220 1984060  2760.197
##  3 Albania    Europe  1972  67.690 2263554  3313.422
##  4 Albania    Europe  1977  68.930 2509048  3533.004
##  5 Albania    Europe  1982  70.420 2780097  3630.881
##  6 Albania    Europe  1987  72.000 3075321  3738.933
##  7 Albania    Europe  1992  71.581 3326498  2497.438
##  8 Albania    Europe  1997  72.950 3428038  3193.055
##  9 Albania    Europe  2002  75.651 3508512  4604.212
## 10 Albania    Europe  2007  76.423 3600523  5937.030
## # ... with 336 more rows
```

Practice piping together filter() and select(). Possibly even piping into ggplot().

```
gapminder %>%
  filter(continent=="Europe",lifeExp>60) %>%
  select(pop,lifeExp) %>%
  ggplot(aes(x=pop,y=lifeExp))+geom_point(alpha=.25,color="orange")
```




```
gapminder %>%
  filter(continent=="Europe",lifeExp>60) %>%
  select(pop,lifeExp) %>%
  ggplot(aes(x=log10(pop),y=lifeExp))+geom_point(alpha=.25,color="orange")
```



Note: Because population is exponential, it makes more sense to represent it by scaling it logarithmically.

But I want to do more!

Evaluate this code and describe the result. Presumably the analyst's intent was to get the data for Rwanda and Afghanistan. Did they succeed? Why or why not? If not, what is the correct way to do this?

```
filter(gapminder, country == c("Rwanda", "Afghanistan"))
```

```
## # A tibble: 12 x 6
##        country continent  year lifeExp      pop gdpPercap
##         <fctr>    <fctr> <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan      Asia  1957  30.332  9240934  820.8530
##  2 Afghanistan      Asia  1967  34.020 11537966  836.1971
##  3 Afghanistan      Asia  1977  38.438 14880372  786.1134
##  4 Afghanistan      Asia  1987  40.822 13867957  852.3959
##  5 Afghanistan      Asia  1997  41.763 22227415  635.3414
##  6 Afghanistan      Asia  2007  43.828 31889923  974.5803
##  7      Rwanda    Africa  1952  40.000  2534927  493.3239
##  8      Rwanda    Africa  1962  43.000  3051242  597.4731
##  9      Rwanda    Africa  1972  44.600  3992121  590.5807
## 10      Rwanda    Africa  1982  46.218  5507565  881.5706
## 11      Rwanda    Africa  1992  23.599  7290203  737.0686
## 12      Rwanda    Africa  2002  43.413  7852401  785.6538
```

A: No, not all of the required data is here. The reason this fails is that we are filtering by a vector of factors.

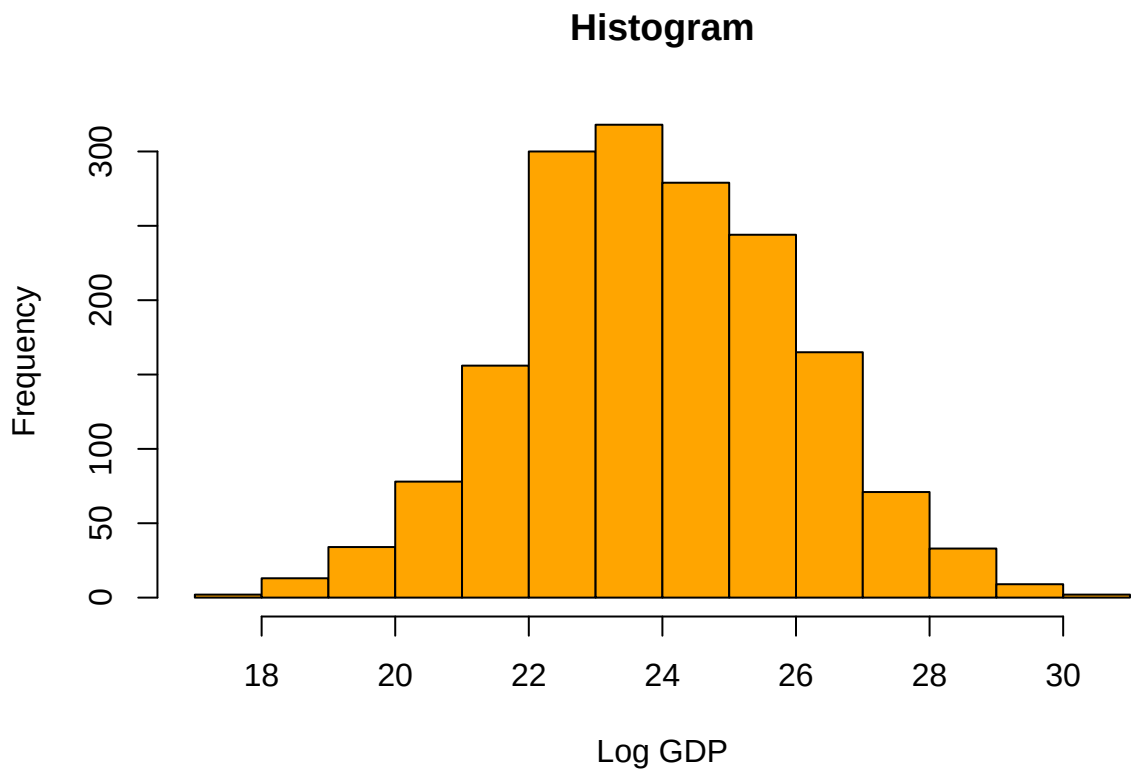
What we really want to filter by is an or statement that evaluates true for both countries.

```
filter(gapminder,country=="Afghanistan"|country=="Rwanda")
```

```
## # A tibble: 24 x 6
##        country continent  year lifeExp      pop gdpPercap
##         <fctr>    <fctr> <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan      Asia  1952  28.801  8425333  779.4453
##  2 Afghanistan      Asia  1957  30.332  9240934  820.8530
##  3 Afghanistan      Asia  1962  31.997 10267083  853.1007
##  4 Afghanistan      Asia  1967  34.020 11537966  836.1971
##  5 Afghanistan      Asia  1972  36.088 13079460  739.9811
##  6 Afghanistan      Asia  1977  38.438 14880372  786.1134
##  7 Afghanistan      Asia  1982  39.854 12881816  978.0114
##  8 Afghanistan      Asia  1987  40.822 13867957  852.3959
##  9 Afghanistan      Asia  1992  41.674 16317921  649.3414
## 10 Afghanistan      Asia  1997  41.763 22227415  635.3414
## # ... with 14 more rows
```

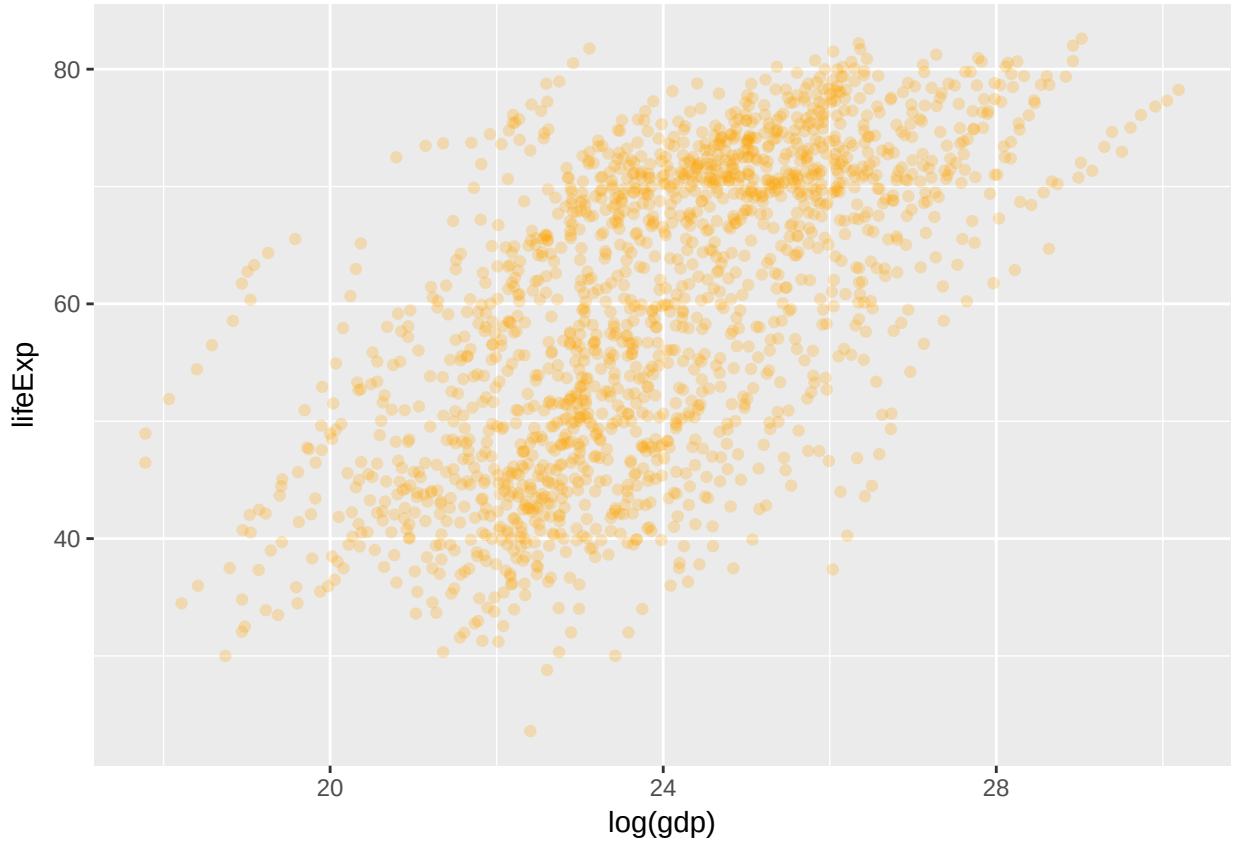
Use more of the dplyr functions in the data set.

```
newData=mutate(gapminder,
       gdp=pop*gdpPercap)
hist(log(newData$gdp),col="orange",
     xlab="Log GDP",main="Histogram")
```



Note: A very normal histogram!

```
p <- ggplot(newData,aes(x=log(gdp),y=lifeExp))
p+geom_point(alpha=.25,color="orange")
```



Note: The relationship here is becoming pretty linear! A regression analysis would prove to show some cool trends in the data of GDP and Life Expectancy!

Reflection

Reflect on what was hard/easy, problems you solved, helpful tutorials you read, etc. What things were hard, even though you saw them in class? What was easy(-ish) even though we haven't done it in class?

A lot of the functionality of the exercises here were very self explanatory! I rarely had to refer to the in class exercises but when I did, everything I needed was there. I did have to ?____ a few commands and once I did need to pull up an example of multiple box plots by factor, but none of this ever took long!

I wouldn't say there were too many challenges here as this homework proved to be more practice which is always needed when programming in a new language.

As for the data set, I had fun comparing variables in different styles of plots and I can easily spend a lot of time making 100 different plots to see different relationships. Also adding features to the plot an exploring options is always a ton of fun!