MATH630-HW1

Joshua Burkhart

October 13, 2015

HLO Gapminder

```
str(gapminder)
```

```
'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     : num  1952 1957 1962 1967 1972 ...
 $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
 $ pop      : num  8425333 9240934 10267083 11537966 13079460 ...
 $ gdpPercap: num  779 821 853 836 740 ...
```

```
glimpse(gapminder)
```

```
Observations: 1,704
Variables: 6
$ country   (fctr) Afghanistan, Afghanistan, Afghanistan, Afghanistan,...
$ continent (fctr) Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asi...
$ year      (dbl) 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992...
$ lifeExp   (dbl) 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.8...
$ pop       (dbl) 8425333, 9240934, 10267083, 11537966, 13079460, 1488...
$ gdpPercap (dbl) 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 78...
```

```
names(gapminder)
```

```
[1] "country"   "continent" "year"      "lifeExp"   "pop"       "gdpPercap"
```

```
head(gapminder)
```

```
      country continent year lifeExp      pop gdpPercap
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
```

```
nrow(gapminder)
```

```
[1] 1704
```

```r
ncol(gapminder)
```

```
[1] 6
```

```r
unique(is.na(gapminder))
```

```
     country continent  year lifeExp   pop gdpPercap
[1,]   FALSE     FALSE FALSE   FALSE FALSE     FALSE
```

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

> data.frame

What is the unit of analysis in the dataset?

> Excerpt of the Gapminder data on life expectancy, GDP per capita, and population by country, every five years, from 1952 to 2007 from http://www.gapminder.org/data/

How many variables/columns?

> 6

How many rows/observations?

> 1704

Which variables are continuous?

> "lifeExp" "gdpPercap"

Which variables are discrete?

> "country" "continent" "pop" "year"

Which variables are categorical?

> "country" "continent"

How many levels do they have?

> country: 142
>
> continent: 5

What about missing data for any variables?

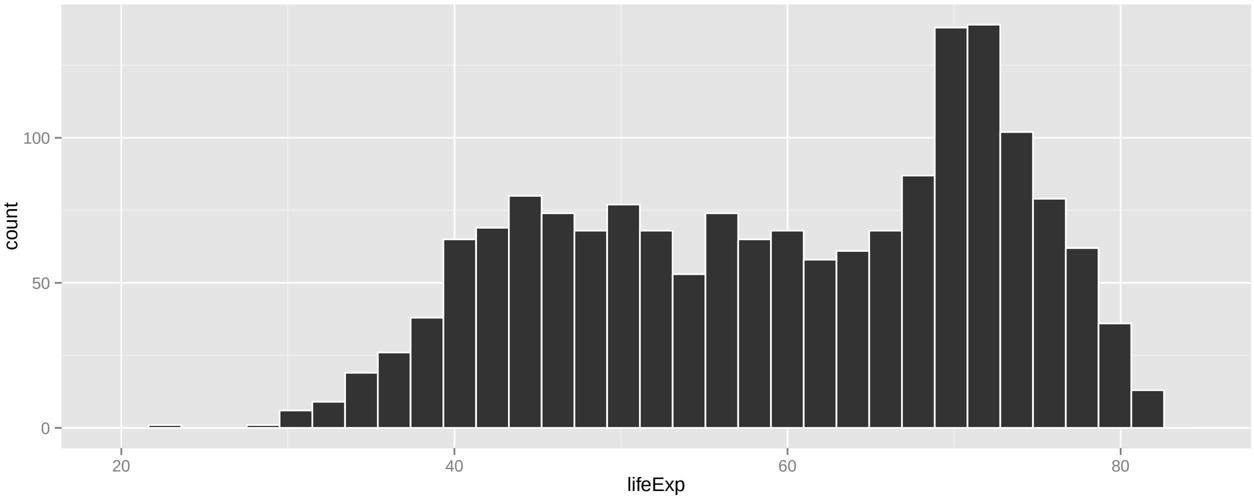
> no missing data reported

Numerical and counting detective work

```
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
```

```
ggplot(gapminder,aes(lifeExp)) +
  geom_histogram(color = "white")
```



Pick one quantitative variable to explore using descriptive statistics as discussed in class.

lifeExp

Characterize the range of possible values, max vs. min, etc.- does it make sense?

Min: 23.6

1st Q: 48.2

Median: 60.71

Mean: 59.47

3rd Q: 70.85

Max: 82.6

These values make sense.

What's the center? What's the spread? What's the shape? Feel free to use summary statistics or tables. You don't need to re-summarise summarised data for us. It is one thing to be able to get R to give you what you ask for. It is another to interpret what R gives you. We are more interested in the latter here, but also that you can do the former without errors.

The distribution looks bimodal, higher peak to the right, skews down

Comment on representativeness of measures of central tendency, given the spread and shape.

IQR / median / mean don't hint at the bimodal distribution but mean < median does hint at skew down

Pick one categorical variable and generate the n's (in whatever the appropriate "unit of analysis" is) and proportions of the sample that contribute to each level of that variable.

continent

Africa :624 = 0.3661972

Americas:300 = 0.1760563

Asia :396 = 0.2323944

Europe :360 = 0.2112676

Oceania : 24 = 0.01408451

```r
filter(gapminder, continent=="Africa") %>%
  nrow / nrow(gapminder)
```

```
[1] 0.3661972
```

```r
filter(gapminder, continent=="Americas") %>%
  nrow / nrow(gapminder)
```

```
[1] 0.1760563
```

```r
filter(gapminder, continent=="Asia") %>%
  nrow / nrow(gapminder)
```

```
[1] 0.2323944
```

```
filter(gapminder, continent=="Europe") %>%
  nrow / nrow(gapminder)
```

```
[1] 0.2112676
```

```
filter(gapminder, continent=="Oceania") %>%
  nrow / nrow(gapminder)
```

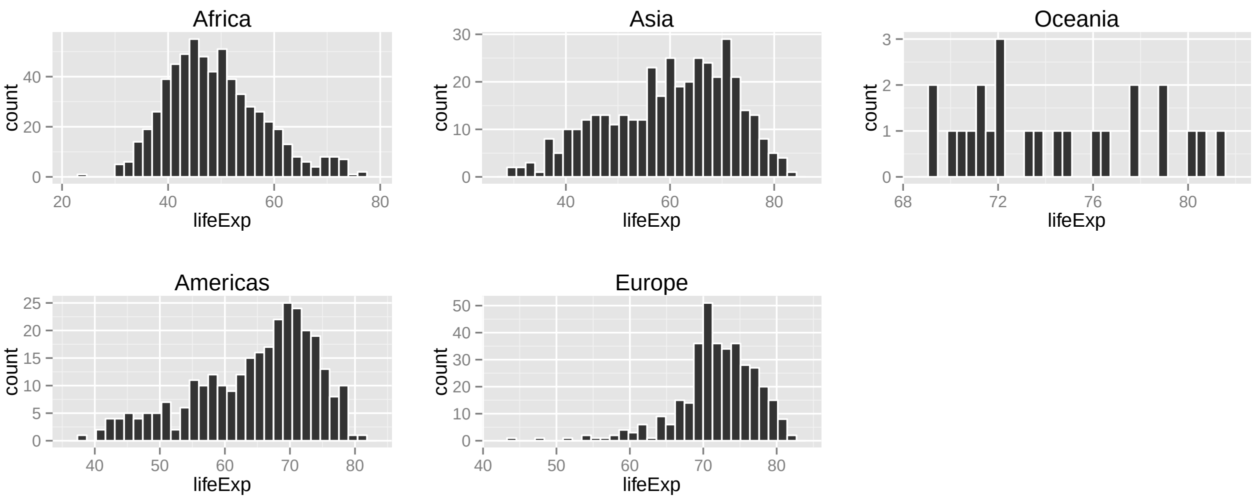
```
[1] 0.01408451
```

Which level contains the smallest number of observations? The largest?

> smallest: Oceana
>
> largest: Africa

Generate your descriptive statistics again, now stratified by the different levels of your categorical variable.

```
africaPlot <- gapminder %>%
  filter(continent=="Africa") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Africa")
americasPlot <- gapminder %>%
  filter(continent=="Americas") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Americas")
asiaPlot <- gapminder %>%
  filter(continent=="Asia") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Asia")
europePlot <- gapminder %>%
  filter(continent=="Europe") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Europe")
oceaniaPlot <- gapminder %>%
  filter(continent=="Oceania") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Oceania")

#from http://stackoverflow.com/questions/24387376/r-weird-error-could-not-find-function-multiplot
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  require(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)
  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                    ncol = cols, nrow = ceiling(numPlots/cols))
  }

 if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
multiplot(africaPlot,americasPlot,asiaPlot,europePlot,oceaniaPlot, cols=3)
```



```
gapminder %>% filter(continent=="Africa") %>%
  summary
```

```
         country       continent        year         lifeExp
 Algeria     : 12   Africa  :624   Min.   :1952   Min.   :23.60
 Angola      : 12   Americas:  0   1st Qu.:1966   1st Qu.:42.37  
 Benin       : 12   Asia    :  0   Median :1980   Median :47.79  
 Botswana    : 12   Europe  :  0   Mean   :1980   Mean   :48.87  
 Burkina Faso: 12   Oceania :  0   3rd Qu.:1993   3rd Qu.:54.41  
 Burundi     : 12                  Max.   :2007   Max.   :76.44  
 (Other)     :552                                                
      pop              gdpPercap      
 Min.   :    60011   Min.   :  241.2  
 1st Qu.:  1342075   1st Qu.:  761.2  
 Median :  4579311   Median : 1192.1  
 Mean   :  9916003   Mean   : 2193.8  
 3rd Qu.: 10801490   3rd Qu.: 2377.4  
 Max.   :135031164   Max.   :21951.2  
```

```
gapminder %>% filter(continent=="Americas") %>%
  summary
```

```
      country       continent        year         lifeExp     
 Argentina: 12   Africa  :  0   Min.   :1952   Min.   :37.58  
 Bolivia  : 12   Americas:300   1st Qu.:1966   1st Qu.:58.41  
 Brazil   : 12   Asia    :  0   Median :1980   Median :67.05  
 Canada   : 12   Europe  :  0   Mean   :1980   Mean   :64.66  
 Chile    : 12   Oceania :  0   3rd Qu.:1993   3rd Qu.:71.70  
 Colombia : 12                  Max.   :2007   Max.   :80.65  
 (Other)  :228                                                
      pop              gdpPercap    
 Min.   :   662850   Min.   : 1202  
 1st Qu.:  2962359   1st Qu.: 3428  
 Median :  6227510   Median : 5466  
 Mean   : 24504795   Mean   : 7136  
 3rd Qu.: 18340309   3rd Qu.: 7830  
 Max.   :301139947   Max.   :42952  
```

```
gapminder %>% filter(continent=="Asia") %>%
  summary
```

```
             country      continent        year         lifeExp     
 Afghanistan     : 12   Africa  :  0   Min.   :1952   Min.   :28.80  
 Bahrain         : 12   Americas:  0   1st Qu.:1966   1st Qu.:51.43  
 Bangladesh      : 12   Asia    :396   Median :1980   Median :61.79  
 Cambodia        : 12   Europe  :  0   Mean   :1980   Mean   :60.06  
 China           : 12   Oceania :  0   3rd Qu.:1993   3rd Qu.:69.51  
 Hong Kong, China: 12                  Max.   :2007   Max.   :82.60  
 (Other)         :324                                                
      pop              gdpPercap     
 Min.   :1.204e+05   Min.   :   331  
 1st Qu.:3.844e+06   1st Qu.:  1057  
 Median :1.453e+07   Median :  2647  
 Mean   :7.704e+07   Mean   :  7902  
 3rd Qu.:4.630e+07   3rd Qu.:  8549  
 Max.   :1.319e+09   Max.   :113523  
```

```
gapminder %>% filter(continent=="Europe") %>%
  summary
```

```
                   country      continent        year     
 Albania                 : 12   Africa  :  0   Min.   :1952  
 Austria                 : 12   Americas:  0   1st Qu.:1966  
 Belgium                 : 12   Asia    :  0   Median :1980  
 Bosnia and Herzegovina: 12   Europe  :360   Mean   :1980  
 Bulgaria                : 12   Oceania :  0   3rd Qu.:1993  
 Croatia                 : 12                  Max.   :2007  
 (Other)                 :288                                
    lifeExp           pop           gdpPercap    
 Min.   :43.59   Min.   :  147962   Min.   :  973.5  
 1st Qu.:69.57   1st Qu.: 4331500   1st Qu.: 7213.1  
 Median :72.24   Median : 8551125   Median :12081.8  
 Mean   :71.90   Mean   :17169765   Mean   :14469.5  
 3rd Qu.:75.45   3rd Qu.:21802867   3rd Qu.:20461.4  
 Max.   :81.76   Max.   :82400996   Max.   :49357.2  
```

```
gapminder %>% filter(continent=="Oceania") %>%
  summary
```

```
        country      continent        year          lifeExp     
 Australia  :12   Africa  : 0   Min.   :1952   Min.   :69.12  
 New Zealand:12   Americas: 0   1st Qu.:1966   1st Qu.:71.20  
 Afghanistan: 0   Asia    : 0   Median :1980   Median :73.67  
 Albania    : 0   Europe  : 0   Mean   :1980   Mean   :74.33  
 Algeria    : 0   Oceania :24   3rd Qu.:1993   3rd Qu.:77.55  
 Angola     : 0                 Max.   :2007   Max.   :81.23  
 (Other)    : 0                                               
      pop             gdpPercap    
 Min.   : 1994794   Min.   :10040  
 1st Qu.: 3199212   1st Qu.:14142  
 Median : 6403492   Median :17983  
 Mean   : 8874672   Mean   :18622  
 3rd Qu.:14351625   3rd Qu.:22214  
 Max.   :20434176   Max.   :34435  
```

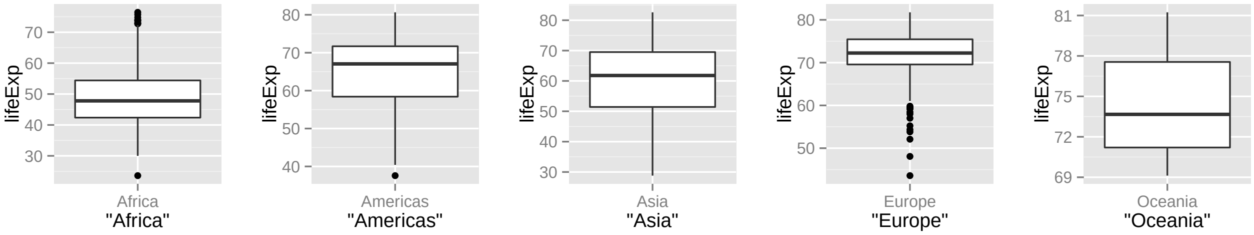
How did any of your initial observations of the quantitative variable change? Foreshadowing: look for differences in both center and spread across categories. Think about what this means in terms of possible comparisons between means across different levels of that factor.

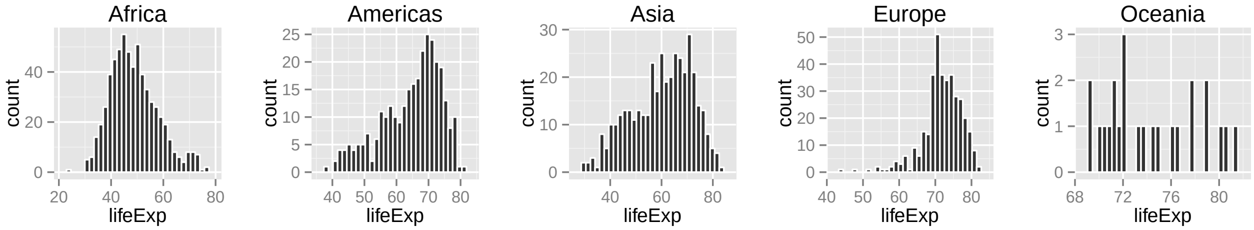
> Asia, Americas, and Europe look similar in that they all skew down. Africa and Oceania skew up. The plots indicate the means of Asia, Americas, and Europe may be lower than their medians, while the plots of Africa and Oceania indicate their means may be higher than their medians.

Graphical detective work

Graphically explore your one quantitative variable using histograms and boxplots. See the exploratory data analysis link for example R code.

```
africaPlot <- gapminder %>%
  filter(continent=="Africa") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Africa")
baf <- gapminder %>%
  filter(continent=="Africa") %>%
  ggplot(aes(x="Africa",y=lifeExp)) +
  geom_boxplot()
americasPlot <- gapminder %>%
  filter(continent=="Americas") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Americas")
bam <- gapminder %>%
  filter(continent=="Americas") %>%
  ggplot(aes(x="Americas",y=lifeExp)) +
  geom_boxplot()
asiaPlot <- gapminder %>%
  filter(continent=="Asia") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Asia")
bas <- gapminder %>%
  filter(continent=="Asia") %>%
  ggplot(aes(x="Asia",y=lifeExp)) +
  geom_boxplot()
europePlot <- gapminder %>%
  filter(continent=="Europe") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") +
  ggtitle("Europe")
beu <- gapminder %>%
  filter(continent=="Europe") %>%
  ggplot(aes(x="Europe",y=lifeExp)) +
  geom_boxplot()
boc <- gapminder %>%
  filter(continent=="Oceania") %>%
  ggplot(aes(x="Oceania",y=lifeExp)) +
  geom_boxplot()
oceaniaPlot <- gapminder %>%
  filter(continent=="Oceania") %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(color = "white") + ggtitle("Oceania")

multiplot(baf,africaPlot,
          bam,americasPlot,
          bas,asiaPlot,
          beu,europePlot,
          boc,oceaniaPlot,
          cols=5)
```





What are you looking for in each plot?

> I'm looking for the IQR (box height) and median (where the box is centered).

Do you notice anything interesting/puzzling/surprising?

> The box plots seem to match the histograms (duh) but Europe's box plot looks like the odd one out as it has such a small Q2-Q3 range and such a high median. Also, Oceania's Q2-Q3 looks unusually large.

Look back at your descriptive statistics for your variable. Comment on the descriptive value of the numbers in light of your visualizations.

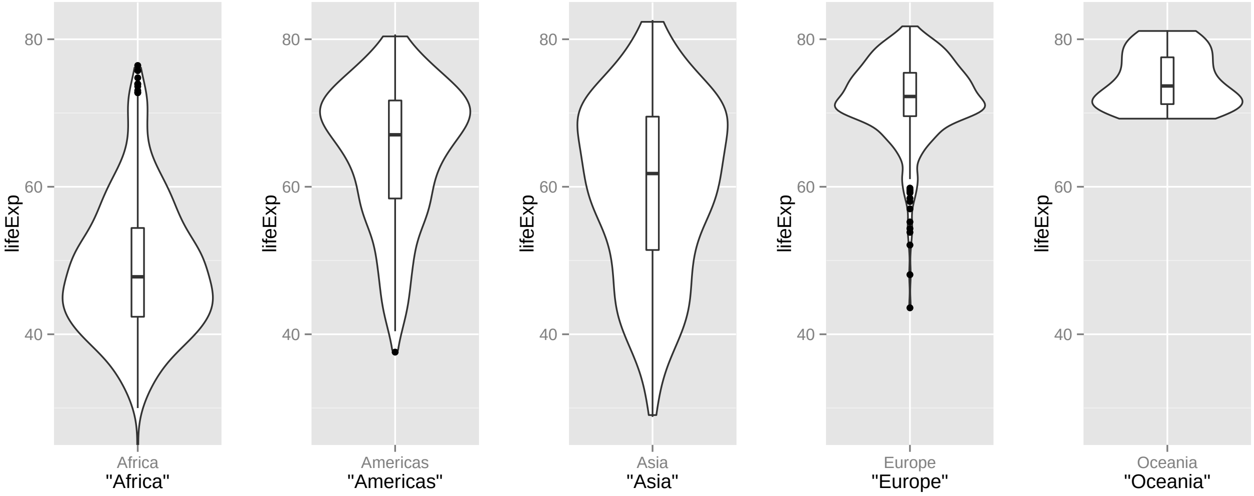
> The numbers aren't wrong, they just don't make the differences as apparent.

Do a quick sanity check- does everything look consistent across numerical and graphical depictions of your data?

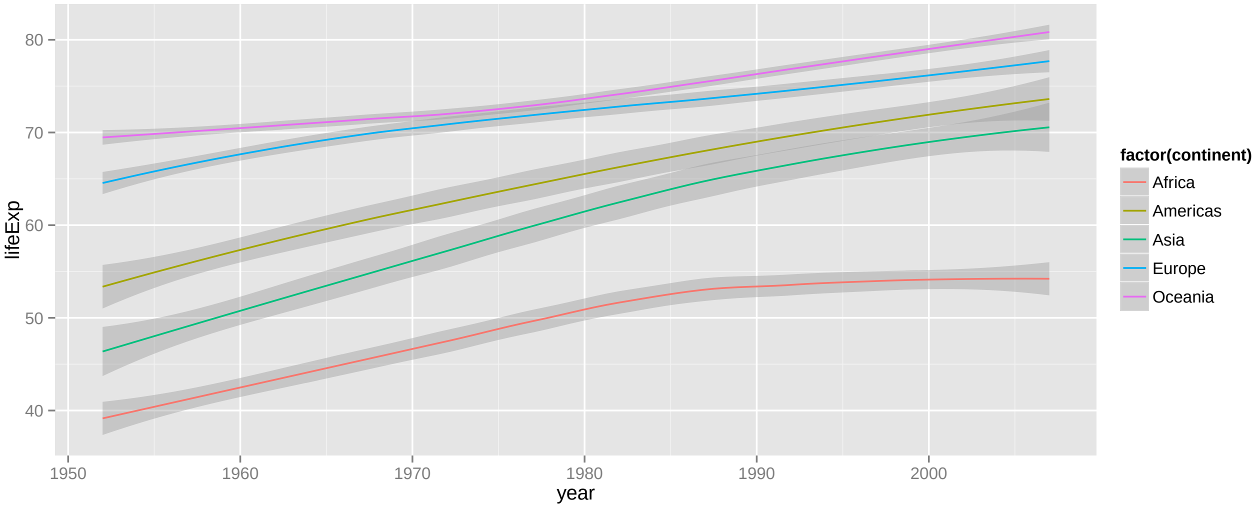
> yes

Add your one categorical variable to the mix and graphically explore your quantitative variable using any of the combination plots discussed in class. Your new plot must account for the categorical variable, either by facetting by levels of that variable, setting an aesthetic (color, shape, etc.) to differ across levels, or stratifying the x-axis by the different levels of your categorical variable. See the exploratory data analysis link for ideas. We want to see you exploring multiple types of plots, and each plot should include at least 2 "layers" of information. Sampling 100 random rows from the dataset is a valid strategy here (reference last slide from Class 2 EDA class) if you want to compare big n/small n types of plots.

```
vaf <- gapminder %>%
  filter(continent=="Africa") %>%
  ggplot(aes(x="Africa",y=lifeExp)) +
  geom_violin() +
  geom_boxplot(width=0.1) +
  coord_cartesian(ylim = c(25,85))
vam <- gapminder %>%
  filter(continent=="Americas") %>%
  ggplot(aes(x="Americas",y=lifeExp)) +
  geom_violin() +
  geom_boxplot(width=0.1) +
  coord_cartesian(ylim = c(25,85))
vas <- gapminder %>%
  filter(continent=="Asia") %>%
  ggplot(aes(x="Asia",y=lifeExp)) +
  geom_violin() +
  geom_boxplot(width=0.1) +
  coord_cartesian(ylim = c(25,85))
veu <- gapminder %>%
  filter(continent=="Europe") %>%
  ggplot(aes(x="Europe",y=lifeExp)) +
  geom_violin() +
  geom_boxplot(width=0.1) +
  coord_cartesian(ylim = c(25,85))
voc <- gapminder %>%
  filter(continent=="Oceania") %>%
  ggplot(aes(x="Oceania",y=lifeExp)) +
  geom_violin() +
  geom_boxplot(width=0.1) +
  coord_cartesian(ylim = c(25,85))

multiplot(vaf,
          vam,
          vas,
          veu,
          voc,
          cols=5)
```



```
gapminder %>%
  ggplot(aes(colour=factor(continent),x=year,y=lifeExp)) +
  geom_smooth()
```



In-depth detective work

Manipulate and further explore the gapminder dataset with the dplyr package, complemented by visualizations made with ggplot2. Pick at least two of the tasks below from the task menu and approach each with a table and figure.

-dplyr should be your main data manipulation tool

-ggplot2 should be your main visualization tool

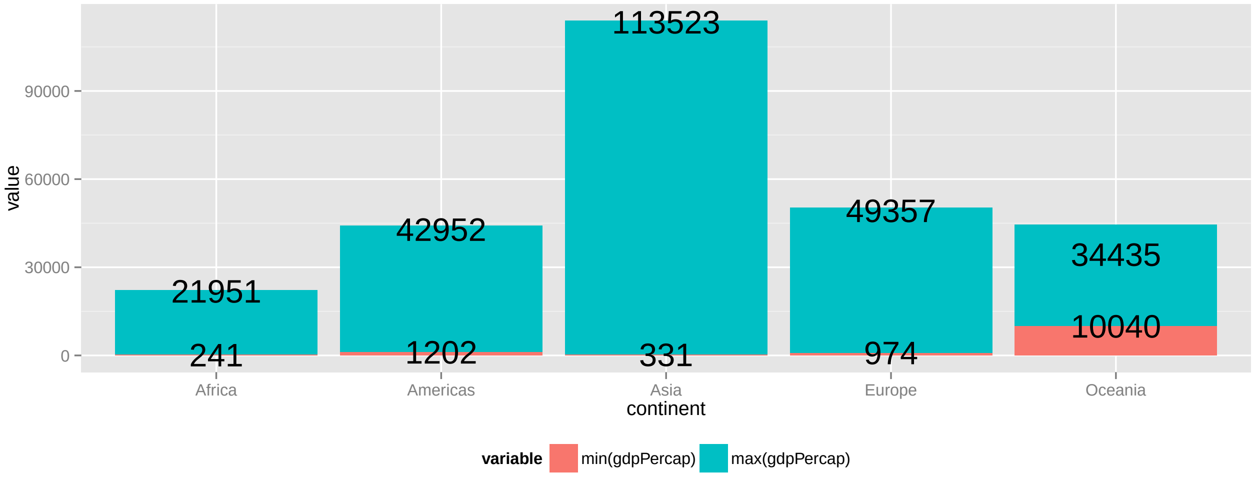
Make observations about what your tables/figures show and about the process. If you want to do something comparable but different, i.e. swap one quantitative variable for another- go for it!

You do not have to use tidyr or otherwise worry about reshaping your tables. Many of your tables may not be formatted perfectly in the report. Simply printing dplyr tabular output is fine. For all things, graphical and tabular, if you're dissatisfied with a result, discuss the problem, what you tried to do to fix it, and move on.

Task menu

Get the maximum and minimum of GDP per capita for all continents.

```
max_min_gdp <- gapminder %>%
  group_by(continent) %>%
  summarize(min(gdpPercap), max(gdpPercap))
max_min_gdp_stack <- max_min_gdp %>%
  melt(id="continent")
max_min_gdp_stack %>%
  ggplot(aes(x=continent, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = round(value)), size = 7) +
  theme(legend.position="bottom")
```



```
max_min_gdp
```

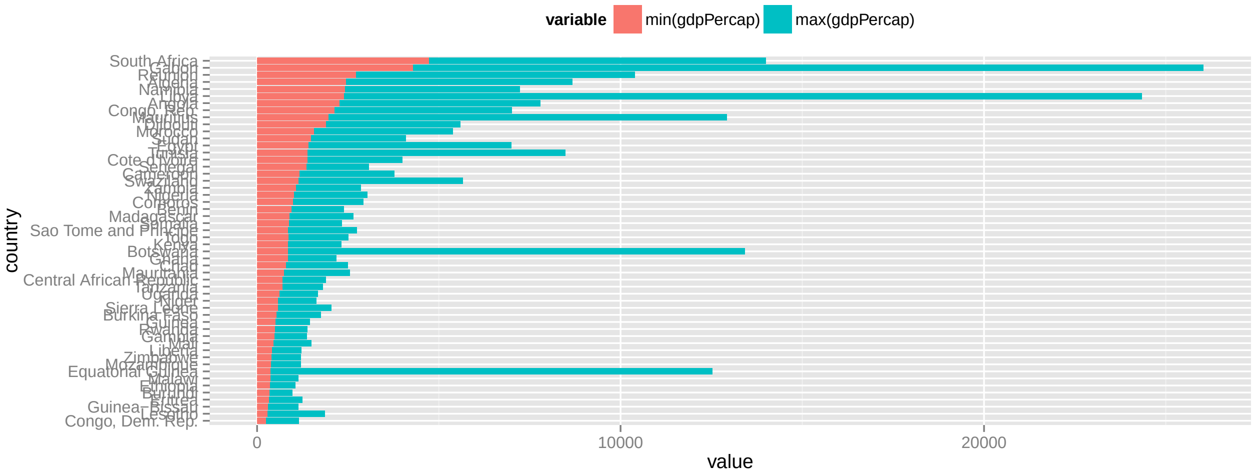
```
Source: local data frame [5 x 3]

  continent min(gdpPercap) max(gdpPercap)
     (fctr)          (dbl)          (dbl)
1    Africa       241.1659       21951.21
2  Americas      1201.6372       42951.65
3      Asia       331.0000      113523.13
4    Europe       973.5332       49357.19
5   Oceania     10039.5956       34435.37
```

Look at the spread of GDP per capita across countries within the continents.

```
max_min_sum <- gapminder %>%
  filter(continent=="Africa") %>%
  group_by(country) %>%
  summarize(min(gdpPercap), max(gdpPercap))
max_min_sum
```

```
Source: local data frame [52 x 3]

                    country min(gdpPercap) max(gdpPercap)
                     (fctr)          (dbl)          (dbl)
1                   Algeria      2449.0082      6223.3675
2                    Angola      2277.1409      5522.7764
3                     Benin       949.4991      1441.2849
4                  Botswana       851.2411     12569.8518
5              Burkina Faso       543.2552      1217.0330
6                   Burundi       339.2965       631.6999
7                  Cameroon      1172.6677      2602.6642
8  Central African Republic       706.0165      1193.0688
9                      Chad       797.9081      1704.0637
10                  Comoros       986.1479      1937.5777
..                      ...            ...            ...
```

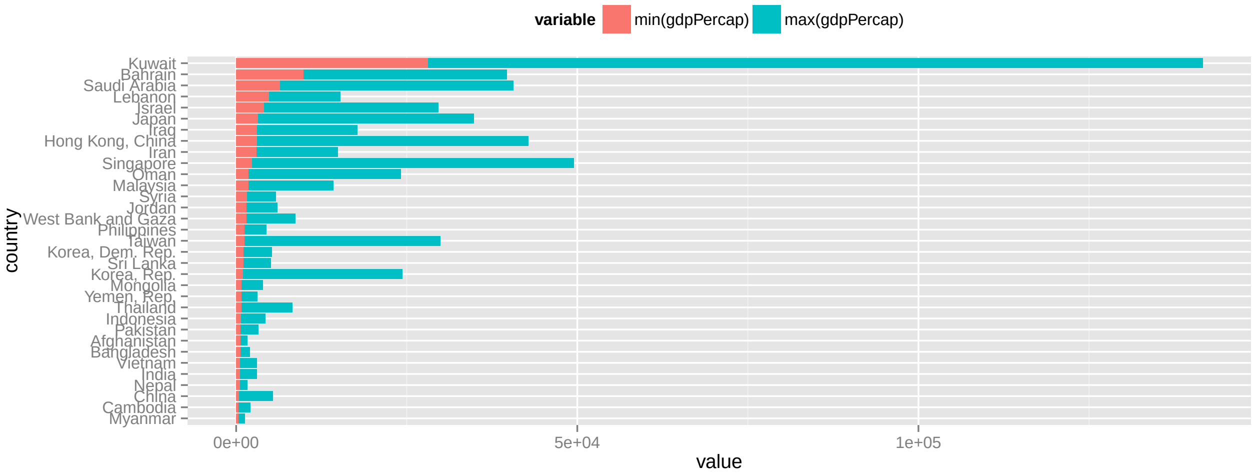
```
country_stack <- max_min_sum %>%
  melt(id=c("country"))
ordered_stack <- country_stack
ordered_stack$country <-
  factor(country_stack$country,
         levels=country_stack[order(country_stack$value),"country"])
oaf <- ordered_stack %>%
  ggplot(aes(x=country, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  coord_flip() +
  theme(legend.position="top")
oaf
```



```
max_min_sum <- gapminder %>%
  filter(continent=="Asia") %>%
  group_by(country) %>%
  summarize(min(gdpPercap), max(gdpPercap))
max_min_sum
```

```
Source: local data frame [33 x 3]

             country min(gdpPercap) max(gdpPercap)
              (fctr)          (dbl)          (dbl)
1        Afghanistan       635.3414       978.0114
2            Bahrain      9867.0848     29796.0483
3         Bangladesh       630.2336      1391.2538
4           Cambodia       368.4693      1713.7787
5              China       400.4486      4959.1149
6   Hong Kong, China      3054.4212     39724.9787
7              India       546.5657      2452.2104
8          Indonesia       749.6817      3540.6516
9               Iran      3035.3260     11888.5951
10              Iraq      3076.2398     14688.2351
..               ...            ...            ...
```

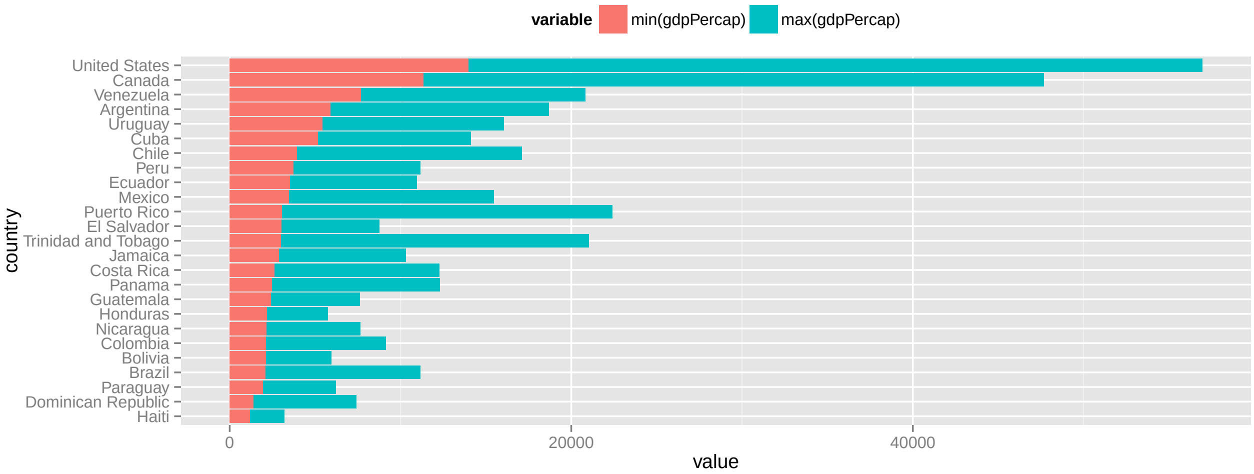
```r
country_stack <- max_min_sum %>%
  melt(id=c("country"))
ordered_stack <- country_stack
ordered_stack$country <-
  factor(country_stack$country,
         levels=country_stack[order(country_stack$value),"country"])
oas <- ordered_stack %>%
  ggplot(aes(x=country, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  coord_flip() +
  theme(legend.position="top")
oas
```



```r
max_min_sum <- gapminder %>%
  filter(continent=="Americas") %>%
  group_by(country) %>%
  summarize(min(gdpPercap), max(gdpPercap))
max_min_sum
```

```
Source: local data frame [25 x 3]

              country min(gdpPercap) max(gdpPercap)
               (fctr)          (dbl)          (dbl)
1           Argentina       5911.315      12779.380
2             Bolivia       2127.686       3822.137
3              Brazil       2108.944       9065.801
4              Canada      11367.161      36319.235
5               Chile       3939.979      13171.639
6            Colombia       2144.115       7006.580
7          Costa Rica       2627.009       9645.061
8                Cuba       5180.756       8948.103
9  Dominican Republic       1397.717       6025.375
10            Ecuador       3522.111       7429.456
..                ...            ...            ...
```

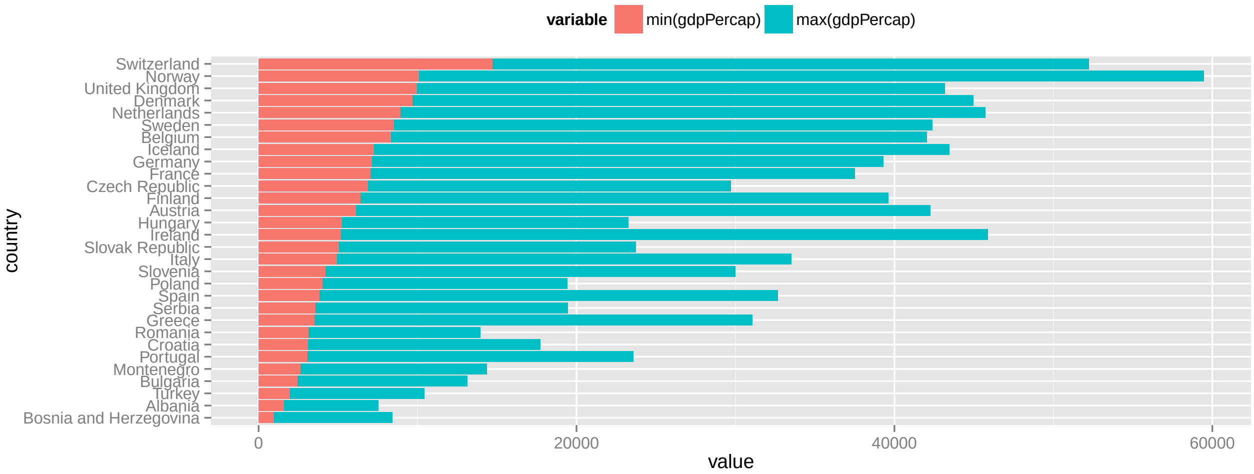
```
country_stack <- max_min_sum %>%
  melt(id=c("country"))
ordered_stack <- country_stack
ordered_stack$country <-
  factor(country_stack$country,
         levels=country_stack[order(country_stack$value),"country"])
oam <- ordered_stack %>%
  ggplot(aes(x=country, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  coord_flip() +
  theme(legend.position="top")
oam
```



```
max_min_sum <- gapminder %>%
  filter(continent=="Europe") %>%
  group_by(country) %>%
  summarize(min(gdpPercap), max(gdpPercap))
max_min_sum
```

```
Source: local data frame [30 x 3]

                  country min(gdpPercap) max(gdpPercap)
                   (fctr)          (dbl)          (dbl)
1                 Albania      1601.0561       5937.030
2                 Austria      6137.0765      36126.493
3                 Belgium      8343.1051      33692.605
4  Bosnia and Herzegovina       973.5332       7446.299
5                Bulgaria      2444.2866      10680.793
6                 Croatia      3119.2365      14619.223
7          Czech Republic      6876.1403      22833.309
8                 Denmark      9692.3852      35278.419
9                 Finland      6424.5191      33207.084
10                 France      7029.8093      30470.017
..                    ...            ...            ...
```

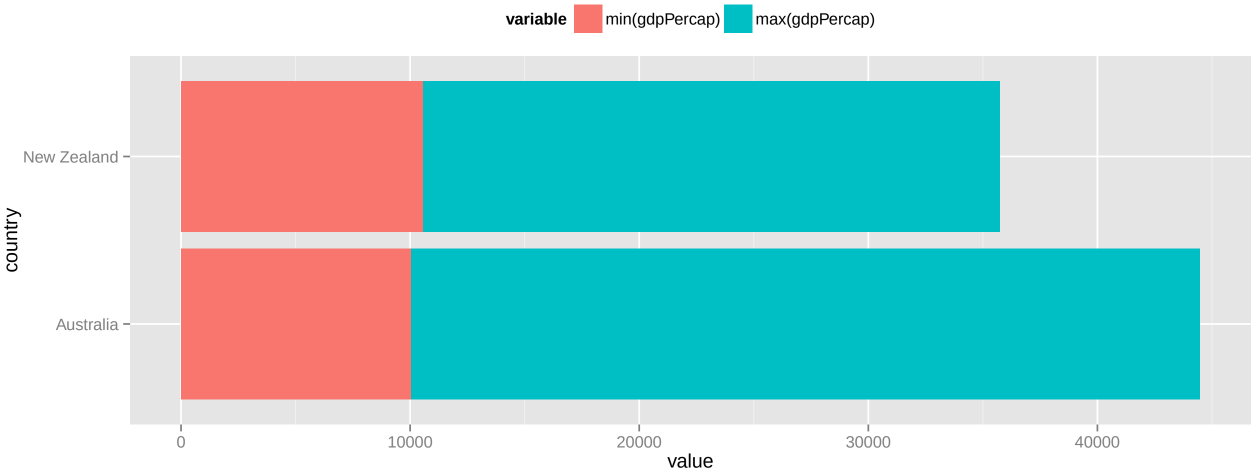
```
country_stack <- max_min_sum %>%
  melt(id=c("country"))
ordered_stack <- country_stack
ordered_stack$country <-
  factor(country_stack$country,
         levels=country_stack[order(country_stack$value),"country"])
oeu <- ordered_stack %>%
  ggplot(aes(x=country, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  coord_flip() +
  theme(legend.position="top")
oeu
```



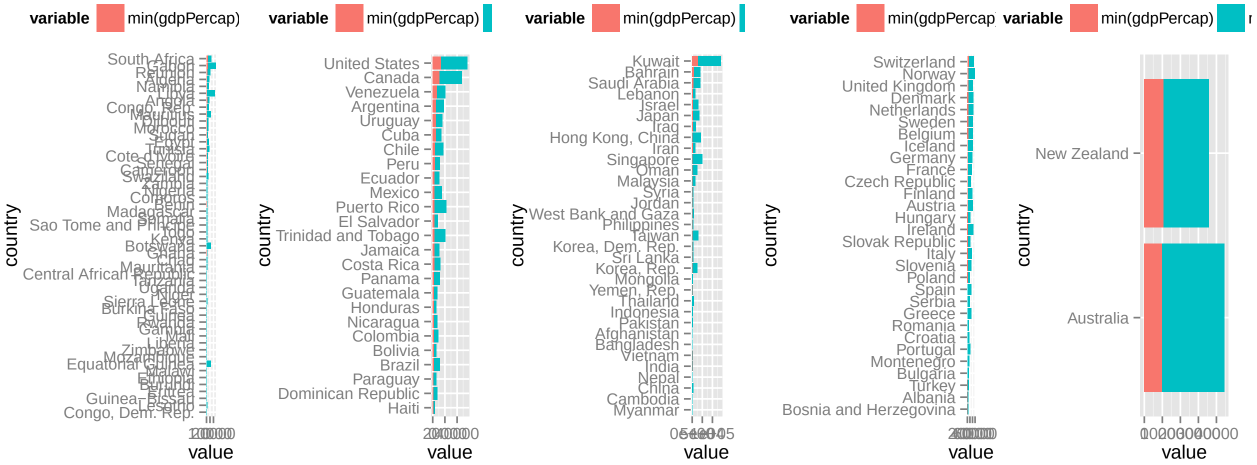
```
max_min_sum <- gapminder %>%
  filter(continent=="Oceania") %>%
  group_by(country) %>%
  summarize(min(gdpPercap), max(gdpPercap))
max_min_sum
```

```
Source: local data frame [2 x 3]

      country min(gdpPercap) max(gdpPercap)
       (fctr)          (dbl)          (dbl)
1   Australia       10039.60       34435.37
2 New Zealand       10556.58       25185.01
```

```
country_stack <- max_min_sum %>%
  melt(id=c("country"))
ordered_stack <- country_stack
ordered_stack$country <-
  factor(country_stack$country,
         levels=country_stack[order(country_stack$value),"country"])
ooc <- ordered_stack %>%
  ggplot(aes(x=country, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  coord_flip() +
  theme(legend.position="top")
ooc
```

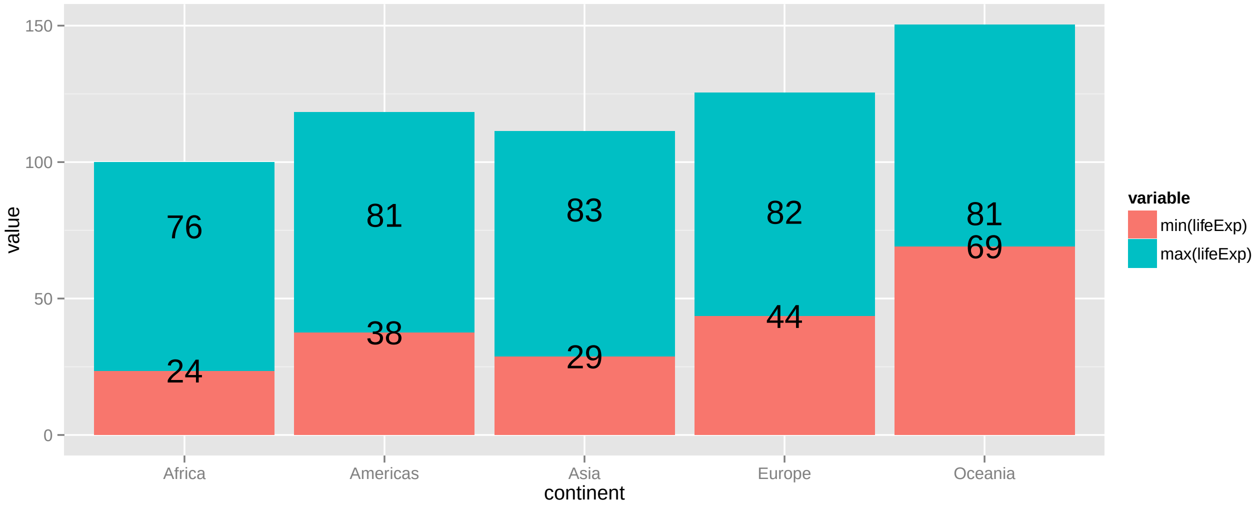


```
multiplot(oaf,
          oam,
          oas,
          oeu,
          ooc,
          cols=5)
```



How does life expectancy vary across different continents?

```
max_min_sum <- gapminder %>%
  group_by(continent) %>%
  summarize(min(lifeExp), max(lifeExp))
continent_stack <- max_min_sum %>%
  melt(id="continent")
continent_stack %>%
  ggplot(aes(x=continent, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = round(value)), size = 7)
```

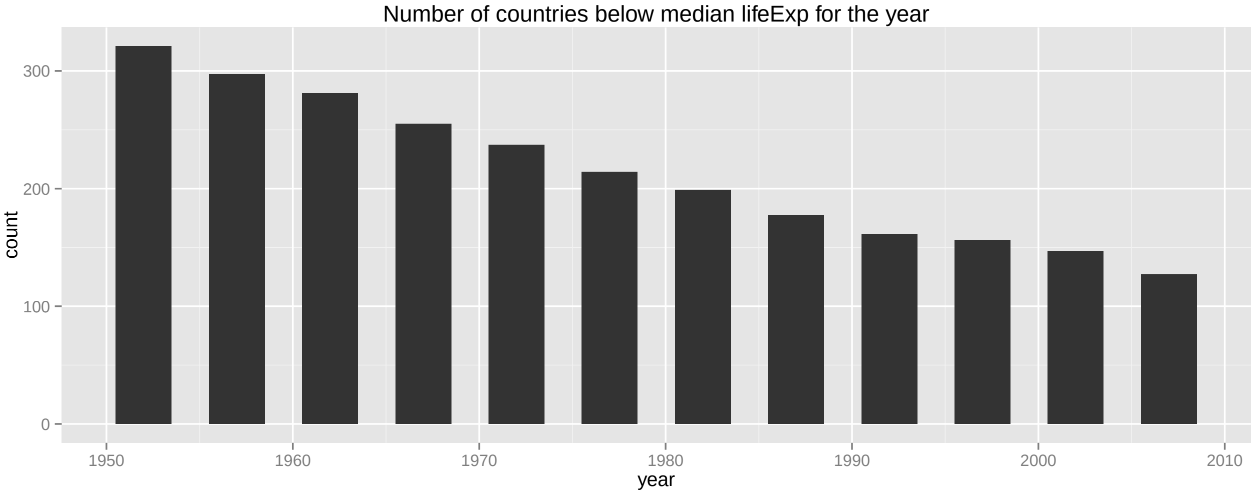


```
max_min_sum
```

```
Source: local data frame [5 x 3]

  continent min(lifeExp) max(lifeExp)
     (fctr)        (dbl)        (dbl)
1    Africa       23.599       76.442
2  Americas       37.579       80.653
3      Asia       28.801       82.603
4    Europe       43.585       81.757
5   Oceania       69.120       81.235
```

Report the absolute and/or relative abundance of countries with low life expectancy over time by continent: Compute some measure of worldwide life expectancy - you decide - a mean or median or some other quantile or perhaps your current age. Then determine how many countries on each continent have a life expectancy less than this benchmark, for each year.

```
#group by year
g_by_year <- gapminder %>%
  melt(id=c("year","lifeExp","country"))

#low lifeExp = lifeExp < median for year
median_by_year <- g_by_year %>%
  group_by(year) %>%
  summarize(median(lifeExp))

#num countries where lifeExp < low lifeExp
num_c_by_year <- g_by_year %>%
  group_by(year) %>%
  summarise(count = length(country[lifeExp < median_by_year$`median(lifeExp)`]))

#x=year,y=num_countries
num_c_by_year %>%
  ggplot(aes(x=year,y=count)) +
  geom_bar(stat="identity",width=3) +
  ggtitle("Number of countries below median lifeExp for the year")
```



```
num_c_by_year
```

```
Source: local data frame [12 x 2]

    year count
   (dbl) (int)
1   1952   321
2   1957   297
3   1962   281
4   1967   255
5   1972   237
6   1977   214
7   1982   199
8   1987   177
9   1992   161
10  1997   156
11  2002   147
12  2007   127
```

Make up your own! Look back at our Class 2 slides for dplyr example ideas with the diamonds dataset, and the package vignettes for other ideas.

Further examining distribution of lifeExp accross continents…

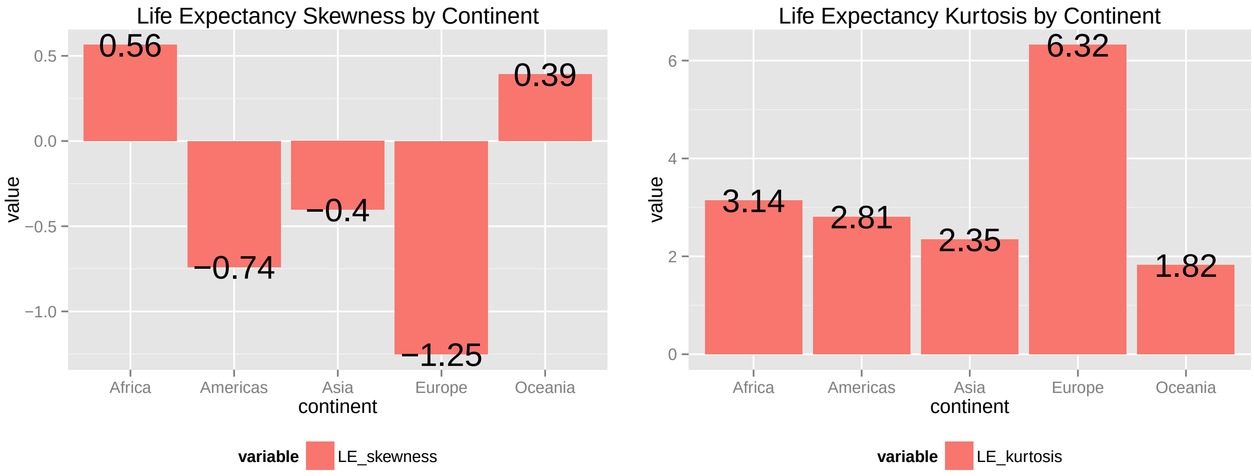
```
max_min_sum <- gapminder %>%
  group_by(continent) %>%
  summarize(LE_skewness=moments::skewness(lifeExp))
continent_stack <- max_min_sum %>%
  melt(id="continent")
skew <- continent_stack %>%
  ggplot(aes(x=continent, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = round(value, digits=2)), size = 7) +
  ggtitle("Life Expectancy Skewness by Continent") +
  theme(legend.position="bottom")
max_min_sum
```

```
Source: local data frame [5 x 2]

  continent LE_skewness
     (fctr)       (dbl)
1    Africa   0.5645229
2  Americas  -0.7386398
3      Asia  -0.4025926
4    Europe  -1.2513139
5   Oceania   0.3921753
```

```
max_min_sum <- gapminder %>%
  group_by(continent) %>%
  summarize(LE_kurtosis=moments::kurtosis(lifeExp))
continent_stack <- max_min_sum %>%
  melt(id="continent")
kurt <- continent_stack %>%
  ggplot(aes(x=continent, y=value, fill=variable)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = round(value, digits=2)), size = 7) +
  ggtitle("Life Expectancy Kurtosis by Continent") +
  theme(legend.position="bottom")
max_min_sum
```

```
Source: local data frame [5 x 2]

  continent LE_kurtosis
     (fctr)       (dbl)
1    Africa    3.143660
2  Americas    2.811413
3      Asia    2.345403
4    Europe    6.320830
5   Oceania    1.820828
```

```
multiplot(skew,kurt,cols=2)
```



Companion graphs

For each table, make sure to include a relevant figure. One tip for starting is to draw out on paper what you want your x- and y-axis to be first and what your geom is; that is, start by drawing the plot you want ggplot to give you. Your figure does not have to depict every single number present in the table. Use your judgement. It just needs to complement the table, add context, and allow for some sanity checking.

Notice which figures are easy/hard to make, and whether the visualization adds clarity, detracts from, or is completely redundant (and therefore probably unnecessary) with respect to the tabular display.

The two most time-consuming plots / tables to generate involved

-ordering bars by a value other than their label (ordering countries by min life exp)

-grouping items into a variable by condition (counting countries w < median life exp)

Report your process

1. consider how a table or plot should look
2. review documentation on https://rpubs.com/bradleyboehmke/data_wrangling
3. guess at what to do based on prior knowledge
4. google frantically (typically landing somewhere on stackoverflow.com or rpubs.com)
5. apply lessons learned from forums to my current problem
6. wrangle my data into a sufficient-looking table
7. repeat steps 3-5 to produce plot

You're encouraged to reflect on what was hard/easy, problems you solved, helpful tutorials you read, etc.