Homework 2

Anita

September 22, 2018

1. Load the packages 'tidyverse' and 'dplyr'
   - Note that 'dplyr' gets automatically loaded with 'tidyverse'
2. Load the dataset 'gapminder'. That is what will be explored

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------
--------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------
--------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gapminder)
```

PART I: Smell test the data

Explore the gapminder object:

```
class(gapminder)#class gives information about the class of the
dataset/variable
```

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

```
head(gapminder)#head provides all columns for the first six rows, good to see
what is in a dataset
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
```

```
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

```
str(gapminder)#str explores the structure of the data frame
```

```
## 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 22227415 ...
##  $ gdpPercap: num  779 821 853 836 740 ...
```

```
ncol(gapminder)#tells you how many columns there are in the data frame
```

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

```
nrow(gapminder)#tells us how many rows there are in a data frame
```

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

```
summary(gapminder)#provides the mean, median and Minimum and Maximum value
for each variable in the dataset
```

```
##         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
##
```

```
names(gapminder)#get the names of teh vriables in the dataset
```

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

```
str(gapminder$country)
```

```
##  Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(gapminder$continent)

##  Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...

str(gapminder$year)

##  int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...

str(gapminder$lifeExp)

##  num [1:1704] 28.8 30.3 32 34 36.1 ...

str(gapminder$pop)

##  int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816
13867957 16317921 22227415 ...

str(gapminder$gdpPercap)

##  num [1:1704] 779 821 853 836 740 ...
```

Questions and Responses

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

It is a data frame

2. What is its class?

It's a tibble and a data frame

3. How many variables/columns?

There are 6 variables/columns in the gapminder dataset.

4. How many rows/observations?

There are 1704 rows/observations in the gapminder dataset.

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

To get the number of variables that are in a dataframe, you can use several functions, for example 'ncol', 'summary', 'head', or 'class'. Dpending on what the goal is, different functions may be more useful. For example, to count the number of variables, see their actual nams and also see at a glance that the vectors all contain data points, the 'head' function' is useful. To get an idea of the mean value of the variables contained, the 'summary' variable is useful, and if one is just interested in the number of variables and doesn't want to have to count them (which may be tedious when working with a large dataset) the 'class' or 'ncol' functions are useful.

Similarly, to get the number of rows at just one glance the 'class' and 'nrow' function are useful. While the 'class' function provides additional information, the 'nrow' function spits out only the number of rows. So thta is useful when that is the only variable of interest.

6. What data type is each variable?

Country is a factor 142 levels.

Continent is factor with 5 levels.

Year is an integer.

Life Expectancy is a number.

Population is an integer.

GDP per capita is a number.

PART II Exploring individual variables

Exploring the quantitative variable Life Expectancy

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

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

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

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

```
gapminder%>%
  filter(lifeExp<23.7)
```

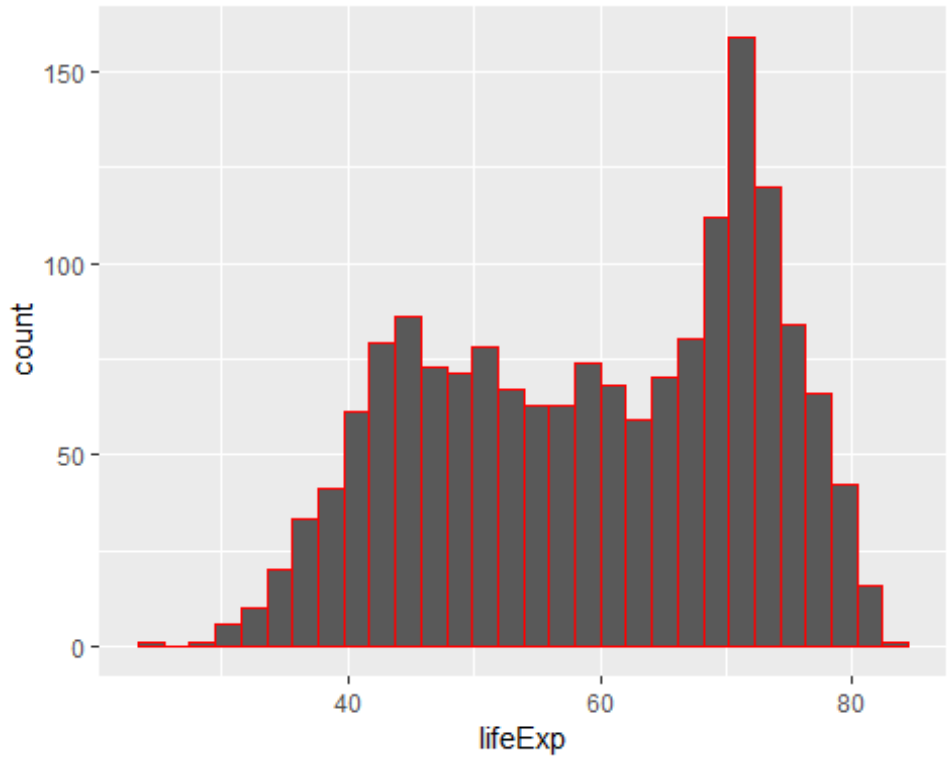
```
## # A tibble: 1 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Rwanda  Africa     1992    23.6 7290203      737.
```

```
gapminder%>%
  filter(lifeExp>82.6)
```

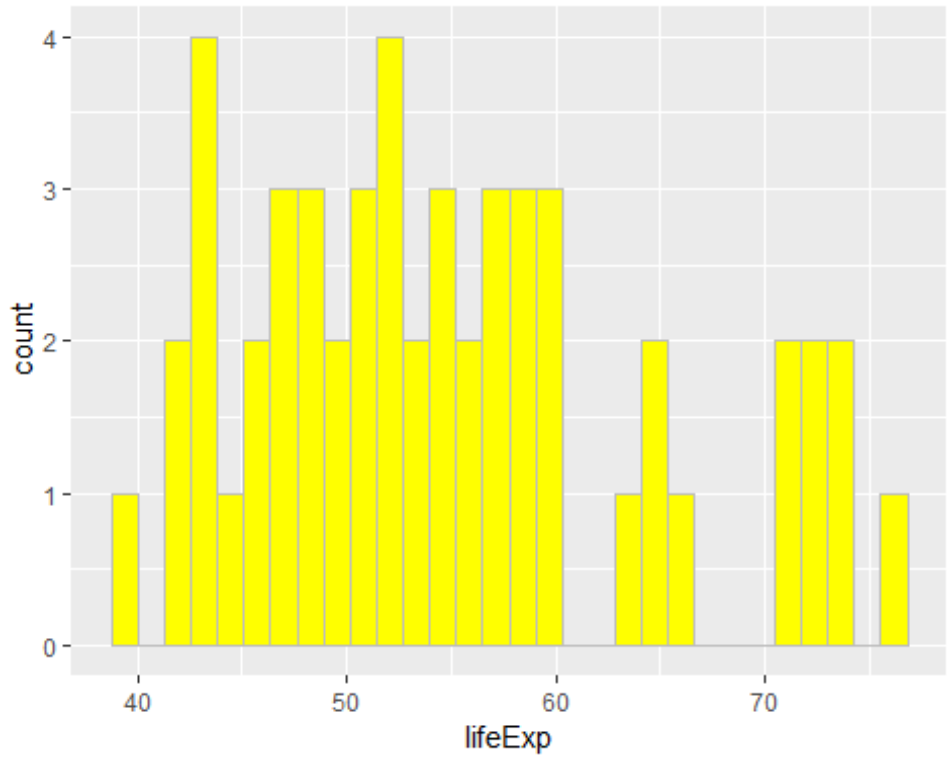
```
## # A tibble: 1 x 6
##   country continent  year lifeExp       pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>     <int>     <dbl>
## 1 Japan   Asia       2007    82.6 127467972    31656.
```

```
ggplot(gapminder, aes(lifeExp)) +
  geom_histogram(colour='red')
```

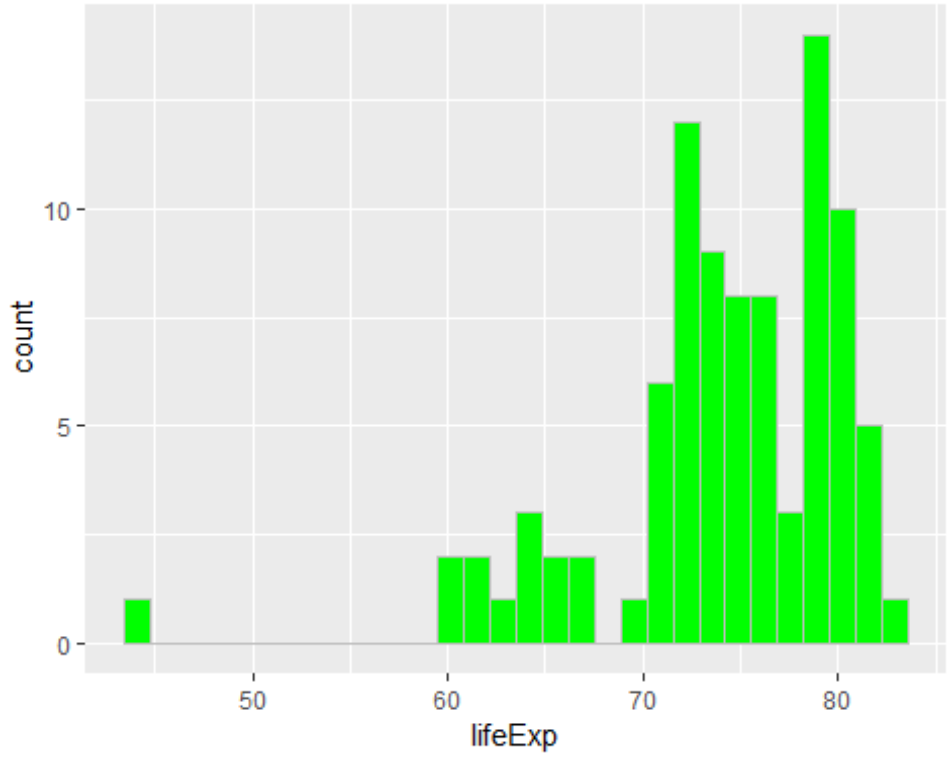
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
a <- gapminder%>%
 filter(continent=="Africa")
  summary(a$lifeExp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.60   42.37   47.79   48.87   54.41   76.44

w <- gapminder%>%
  filter(continent!="Africa")
  summary(w$lifeExp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.80   59.62   68.99   65.60   73.06   82.60

a.recent <- gapminder%>%
 filter((continent=="Africa")&
          year==2007)
  summary(a.recent$lifeExp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39.61   47.83   52.93   54.81   59.44   76.44

ggplot(a.recent, aes(lifeExp)) +
  geom_histogram(color = 'grey', fill = 'yellow')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
w.recent <- gapminder%>%
  filter((continent!="Africa")&
           year==2007)
  summary(w.recent$lifeExp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.83   71.80   74.76   74.06   78.77   82.60

ggplot(w.recent, aes(lifeExp)) +
  geom_histogram(colour='grey', fill='green')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Description of the variable 'life expectancy'

When combined across all 5 continents and years (from 1952 to 2007), then the mean life expecancy is 59.47 years (SD = 12.92) and it ranges from 23.6 years to 82.6 years. When exploring the data further using the filter function, it tirns out that Rwanda had a life expectancy of 23.6 years in 1992 while Japan had a life expectancy of 82.6 in 2007. This enormous difference in life expectancy between these two counties is shocing and shows that aggregate data across so many differnt regions and such a long time span tels us little about the conditions people live in in different places.

Furthermore, the mean life expectancy of Africa across the entire time (from 1952 to 2007) was 48.87 years, and hence 10 years lower than that of the entire world combined (including Africa). The life expectancy of the rest of the world (excluding Africa) was 65.60 years. One might think that the stark discrepancy in life expectancy is driven by past data. However, a glance at the most recent year for which data is available shows that difference in life expectancy has further increased. While the life expectancy in Africa has increased by about 6 years to 54.81 that of the rest f the world has increased even more steeply and is now at 74.06 years.

Exploring the categorical variable Continent

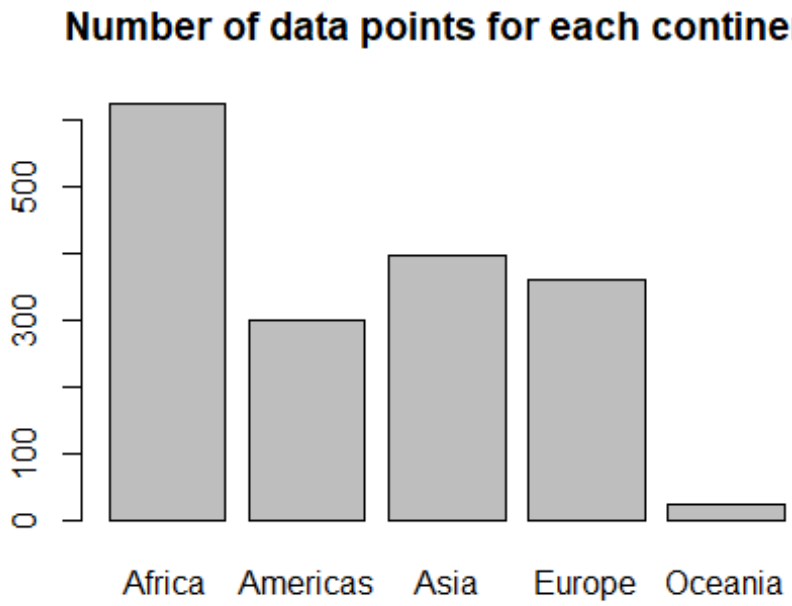
```
levels(gapminder$continent)
```

```
## [1] "Africa"   "Americas" "Asia"     "Europe"   "Oceania"
```

```
table(gapminder$continent)
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##      624      300      396      360       24
```

```
plot(gapminder$continent, main='Number of data points for each continent')
```



Description of the variable 'continent'

There are 5 levels of the variable 'continent'. In other words, the dataset includs information from 5 continents: Africa, Americas, Asia, Europe, and Oceania.

As can be seen using the tabe function (or from the plot), Africa has the most data points, namely 624. The Amerias, Asia, and Europe fall in the middle and have each between 300 and 396 data points. And Oceania has only 24 data points.

PART III Exploring various plot types (for the variable Life Expectancy)

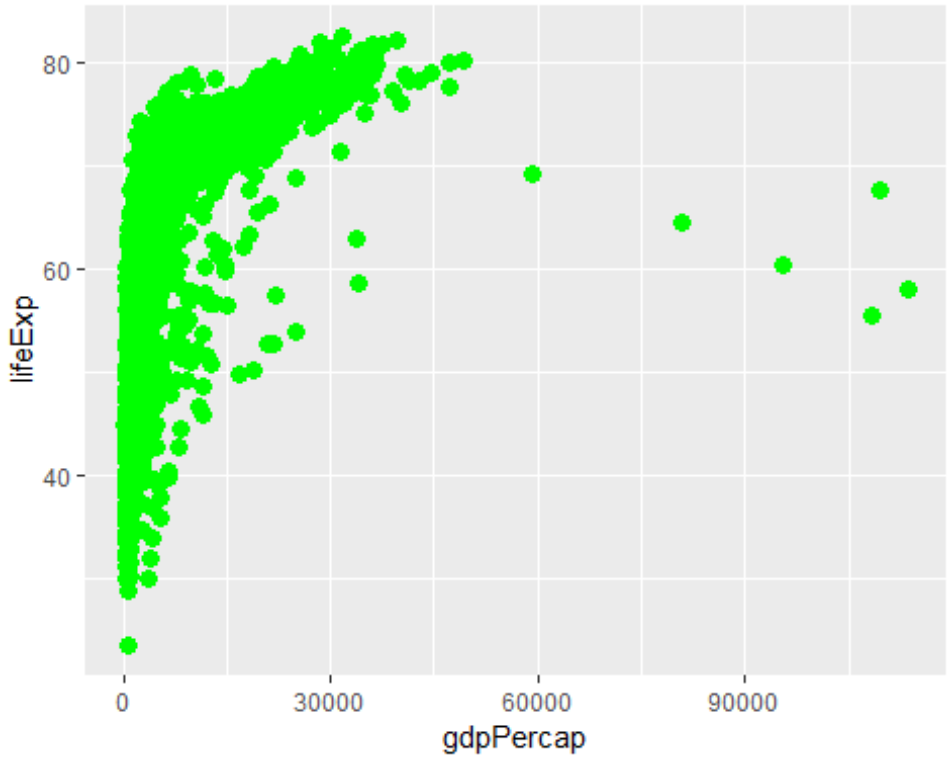
Exploring the quantitative variable Life Expectancy

Explore various plot types Make a few plots, probably of the same variable you chose to characterize numerically. You can use the plot types we went over in class (cm006) to get an idea of what you'd like to make. Try to explore more than one plot type. Just as an example of what I mean:

A scatterplot of two quantitative variables. A plot of one quantitative variable. Maybe a histogram or densityplot or frequency polygon. A plot of one quantitative variable and one categorical. Maybe boxplots for several continents or countries. 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.

Scatterplot between GDP per capita and Life Expectancy

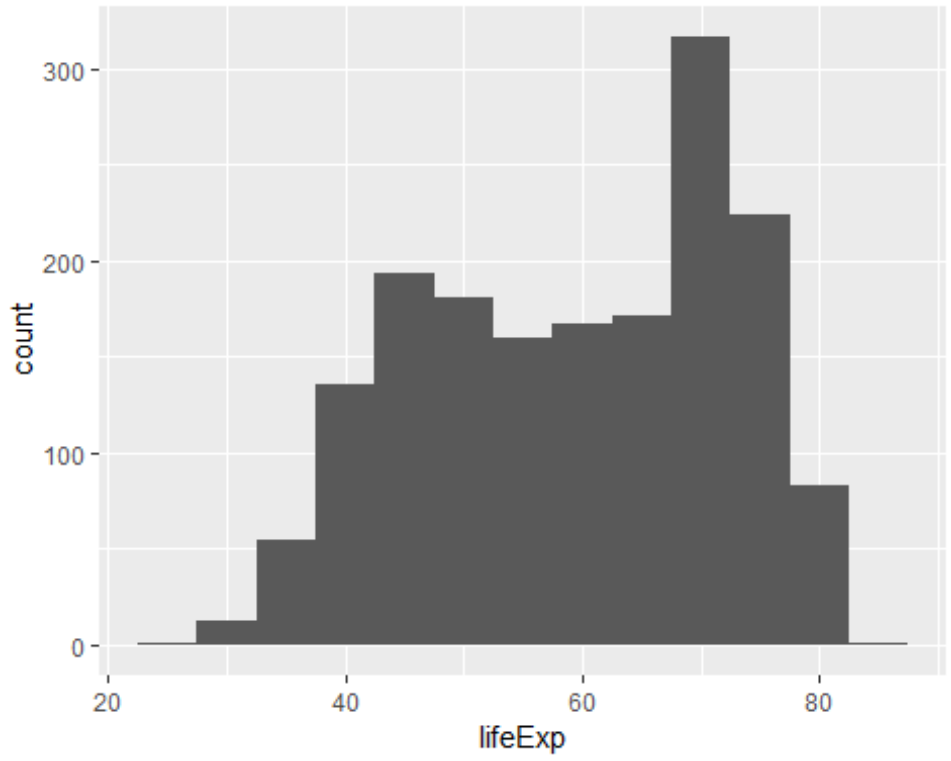
```
ggplot(gapminder, aes(x=gdpPercap, y=lifeExp))+
  geom_point(colour='green', size=3)
```



Various histograms for the variable Life Expectancy

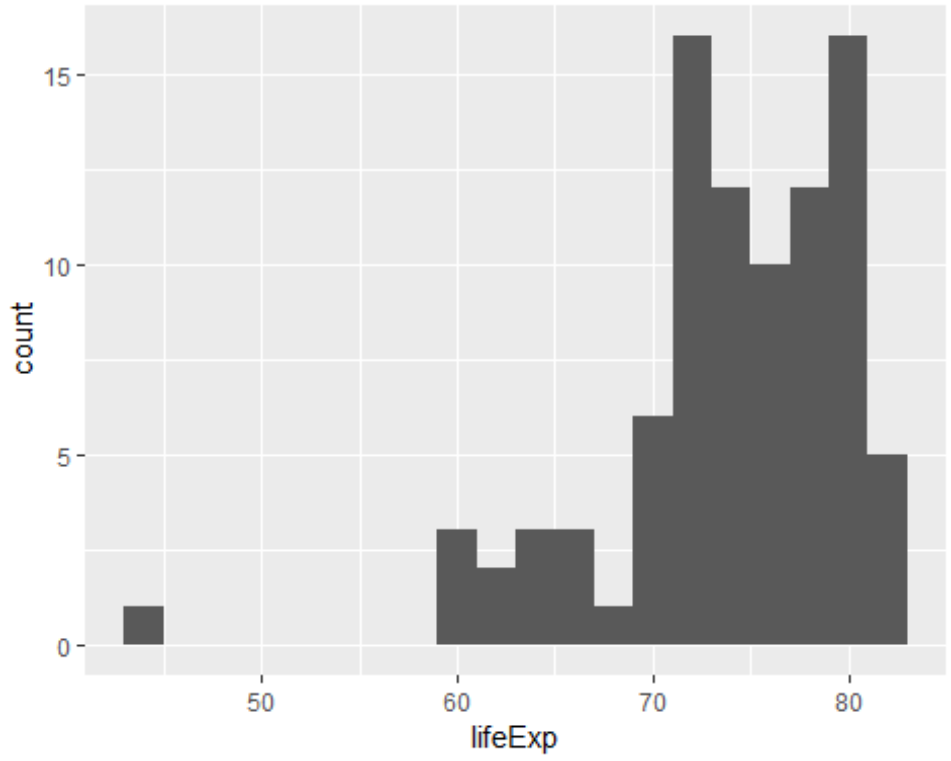
Histogram of life expectancy including all countries and years

```
ggplot(gapminder, aes(lifeExp)) +
  geom_histogram(binwidth = 5)
```



Histogram of life expectancy including data points from all continents *EXCEPT Africa* for the year 2007

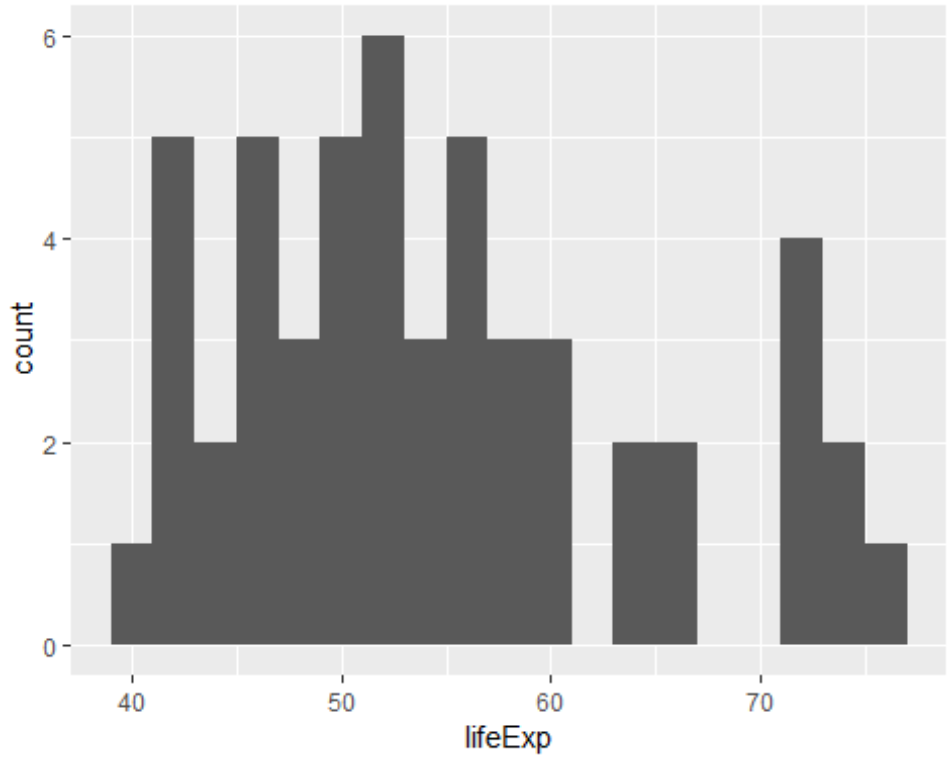
```
ggplot(w.recent, aes(lifeExp)) +
  geom_histogram(binwidth = 2)
```



As we can see in the histogram, there is an outlier where life expectancy seems to be much lower than in the other countries.

Histogram of life expectancy including data points from *Africa ONLY* for the year 2007

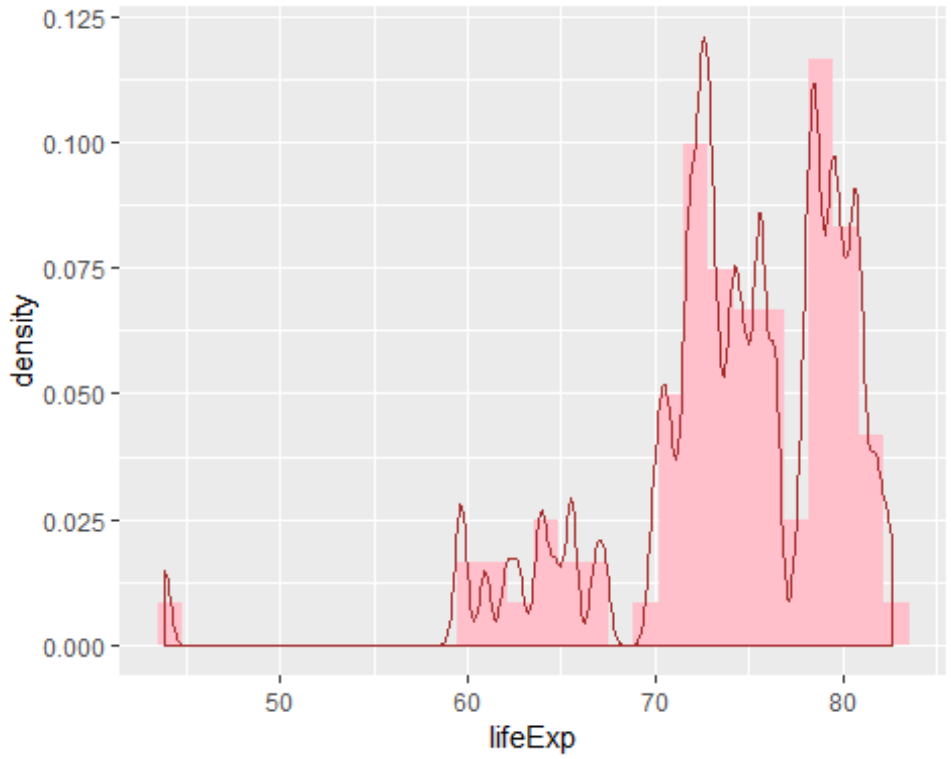
```
ggplot(a.recent, aes(lifeExp)) +
  geom_histogram(binwidth = 2)
```



We can see in this histogram that the variability in life expecatncy within Africa is greater than that of the rest of the world. Most countries in 2007 in Africa had a life expectancy somewhere between 39 and 61 years, but there are also quite a few countries with a much higher ife expectancy, some have even a life expectancy of well over 70.

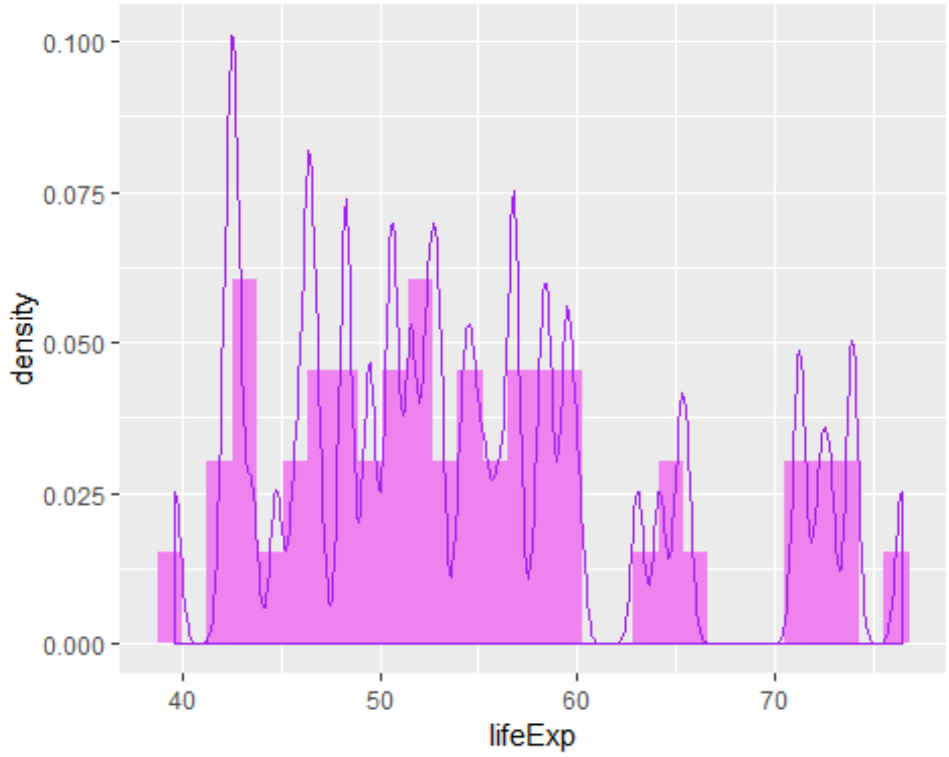
Histogram including kernel density plot of life expectancy including data points from all continents *EXCEPT Africa* for the year 2007

```
ggplot(w.recent, aes(lifeExp)) +
  geom_histogram(aes(y=..density..), fill='pink', bw=5) +#this .. part in
here is to map the histogram scale   #onto the density plot
  geom_density(bw=0.3, colour='brown')

## Warning: Ignoring unknown parameters: bw

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



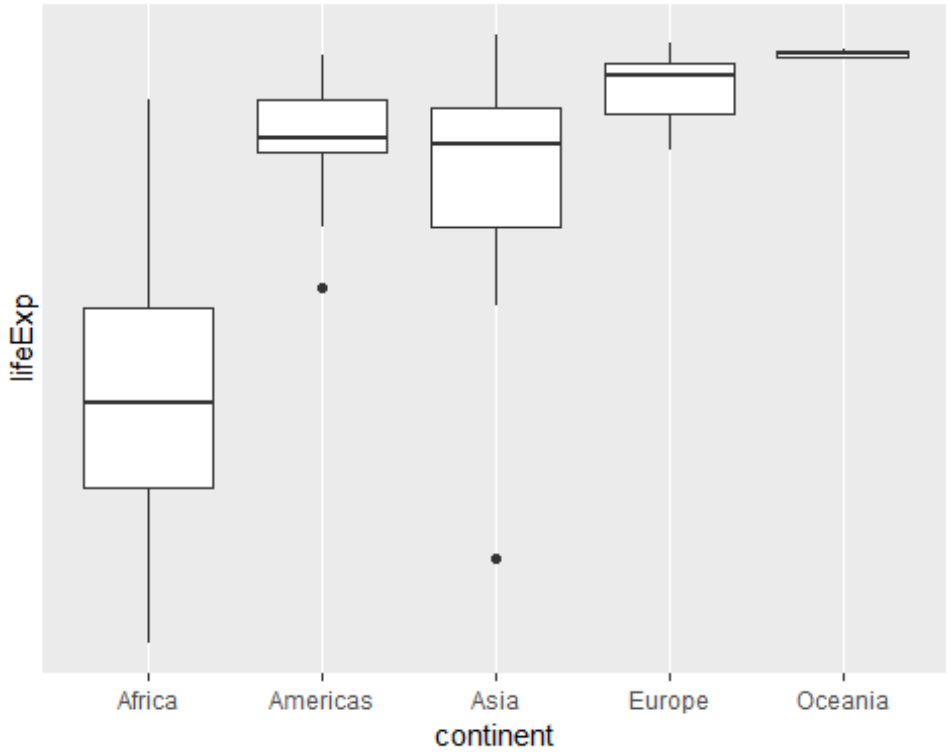
Histogram including kernel density plot of life expectancy including data points from *Africa ONLY* for the year 2007

```
ggplot(a.recent, aes(lifeExp)) +
  geom_histogram(aes(y=..density..), fill='violet', bw=5) +#this .. part in
here is to map the histogram scale   #onto the density plot
  geom_density(bw=0.3, colour='purple')

## Warning: Ignoring unknown parameters: bw

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



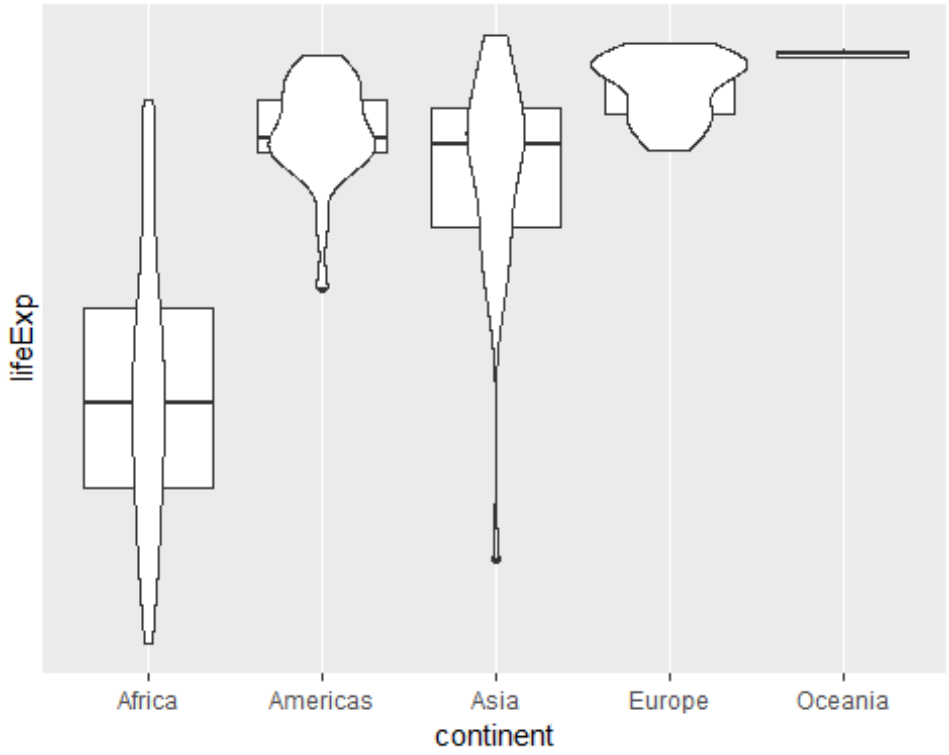
Box plots for the life expectancy of *2007 ONLY* broken down by continents

```
y<-filter(gapminder, year==2007)%>%
ggplot(aes(continent, lifeExp)) +
  scale_y_log10()+
  geom_boxplot()
y
```



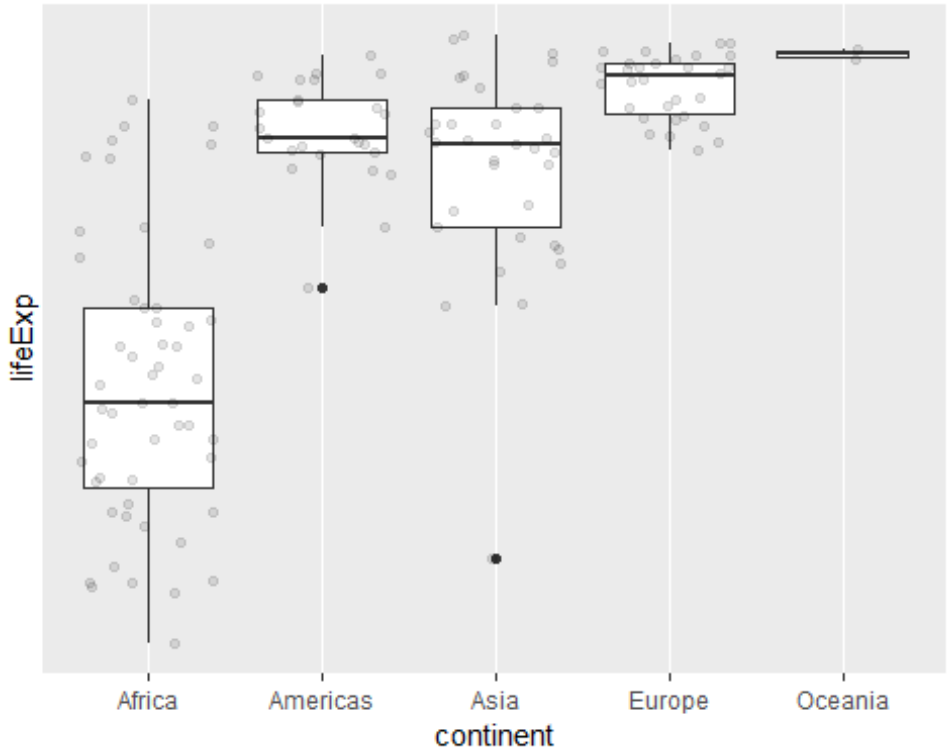
Box plots in conjunction with violin plots for the life expectancy of *2007 ONLY* broken down by continents

```
y + geom_violin() #Note, I stored the box plot in y above, so that I don't
have to retype everything
```



Box plots in conjunction with violin plotswith data points (jitter) for the life expectancy of *2007 ONLY* broken down by continents

```
y + geom_jitter(alpha=0.1)
```

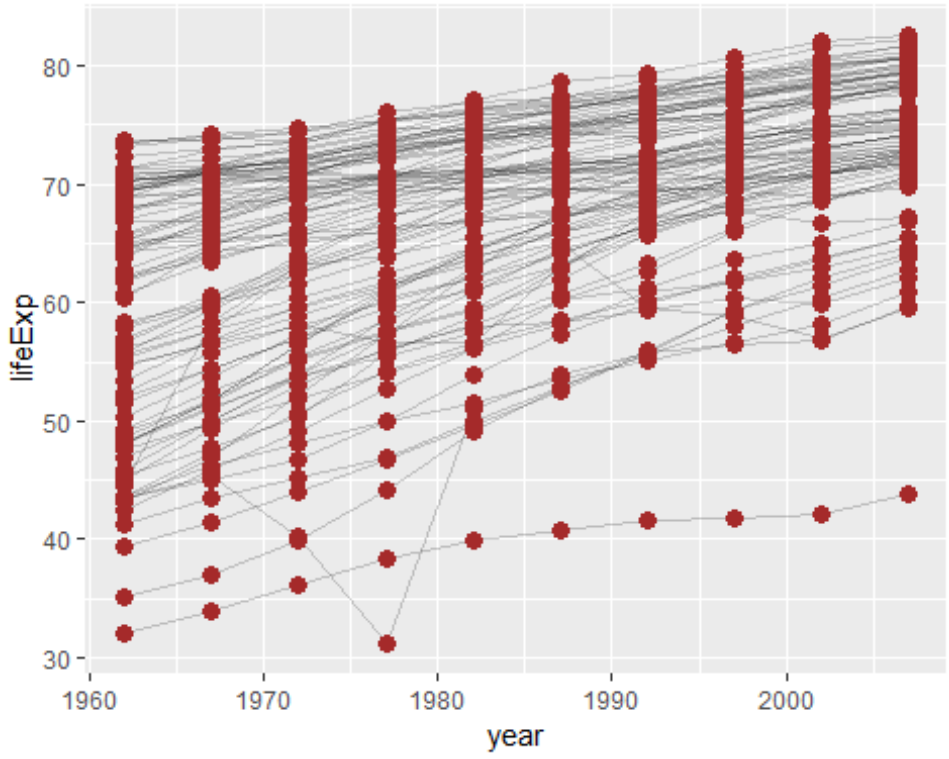


As can be seen from this plot, there are realtively few data points for each continent for 2007. In particular, in Oceania there are only 2 data points.

Time/Line Plots

Create a plot for countries from all continents except Africa showing the average life expectancy as a function of the year (ranging from 1961 to 2007).

```
gapminder%>%
  select(continent, country,year, lifeExp)%>%
  filter(!(continent=='Africa')
         & year > 1960)%>%
  ggplot(aes(year, lifeExp))+
  geom_line(aes(group=country), alpha=0.2) +
  geom_point(colour='brown', size=3)
```



Additional Exploration

Table showing data for African countries in 2007

```
t<-gapminder%>%
  filter(continent=='Africa' & year > 2006)%>%
  select(country, lifeExp, pop, gdpPercap )
knitr::kable(t)
```

country	lifeExp	pop	gdpPercap
Algeria	72.301	33333216	6223.3675
Angola	42.731	12420476	4797.2313
Benin	56.728	8078314	1441.2849
Botswana	50.728	1639131	12569.8518
Burkina Faso	52.295	14326203	1217.0330
Burundi	49.580	8390505	430.0707
Cameroon	50.430	17696293	2042.0952
Central African Republic	44.741	4369038	706.0165
Chad	50.651	10238807	1704.0637
Comoros	65.152	710960	986.1479
Congo, Dem. Rep.	46.462	64606759	277.5519

Congo, Rep.	55.322	3800610	3632.5578
Cote d'Ivoire	48.328	18013409	1544.7501
Djibouti	54.791	496374	2082.4816
Egypt	71.338	80264543	5581.1810
Equatorial Guinea	51.579	551201	12154.0897
Eritrea	58.040	4906585	641.3695
Ethiopia	52.947	76511887	690.8056
Gabon	56.735	1454867	13206.4845
Gambia	59.448	1688359	752.7497
Ghana	60.022	22873338	1327.6089
Guinea	56.007	9947814	942.6542
Guinea-Bissau	46.388	1472041	579.2317
Kenya	54.110	35610177	1463.2493
Lesotho	42.592	2012649	1569.3314
Liberia	45.678	3193942	414.5073
Libya	73.952	6036914	12057.4993
Madagascar	59.443	19167654	1044.7701
Malawi	48.303	13327079	759.3499
Mali	54.467	12031795	1042.5816
Mauritania	64.164	3270065	1803.1515
Mauritius	72.801	1250882	10956.9911
Morocco	71.164	33757175	3820.1752
Mozambique	42.082	19951656	823.6856
Namibia	52.906	2055080	4811.0604
Niger	56.867	12894865	619.6769
Nigeria	46.859	135031164	2013.9773
Reunion	76.442	798094	7670.1226
Rwanda	46.242	8860588	863.0885
Sao Tome and Principe	65.528	199579	1598.4351
Senegal	63.062	12267493	1712.4721
Sierra Leone	42.568	6144562	862.5408
Somalia	48.159	9118773	926.1411
South Africa	49.339	43997828	9269.6578
Sudan	58.556	42292929	2602.3950
Swaziland	39.613	1133066	4513.4806
Tanzania	52.517	38139640	1107.4822

Togo	58.420	5701579	882.9699
Tunisia	73.923	10276158	7092.9230
Uganda	51.542	29170398	1056.3801
Zambia	42.384	11746035	1271.2116
Zimbabwe	43.487	12311143	469.7093