Assignment 2

Isabel

9/19/2019

First, load the `gapminder` and `tidyverse` packages. The `dplyr` package will be loaded via the `tidyverse` package.

```
suppressPackageStartupMessages(library(gapminder))
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(DT))
```

Exercise 1

Let's first have an overview of the data:

```
DT::datatable(as_tibble(gapminder))
```

Let's now focus on the following three countries: Singapore, Malaysia and Indonesia:

```
filtered<-gapminder%>%
  filter(year>1969 & year<1980)%>%
  filter(country %in% c('Singapore', 'Malaysia', 'Indonesia'))

DT::datatable(filtered)
```

We now only want the columns 'country' and 'gdpPercap' from the above dataset:

```
filtered%>%
  select(country, gdpPercap)
```

```
## # A tibble: 6 x 2
##   country   gdpPercap
##   <fct>         <dbl>
## 1 Indonesia     1111.
## 2 Indonesia     1383.
## 3 Malaysia      2849.
## 4 Malaysia      3828.
## 5 Singapore     8598.
## 6 Singapore    11210.
```

We want to see which countries have experienced a drop in life expectancy.

Here, in my code, I filter out the rows for year=1952 because these reflect the difference between *two* countries' life expectancies. However, we only want to compare within-country differences.

From the data, we can see that the biggest drop in life expectancy occured in Rwanda between 1987 and 1992. On the converse, we can see that the biggest increase in life expectancy occured in Cambodia between 1977 and 1982. These results are not surprising, given that both countries experienced devastating genocides during those time periods.

```
gapminder_mutated<-
  gapminder%>%
  mutate(difference=lifeExp-lag(lifeExp, 1))

gapminder_mutated%>%
  filter(difference<0)%>%
  filter(year!=1952)
```

```
## # A tibble: 102 x 7
##    country  continent  year lifeExp     pop gdpPercap difference
##    <fct>    <fct>     <int>   <dbl>   <int>     <dbl>      <dbl>
##  1 Albania  Europe     1992    71.6 3326498     2497.     -0.419
##  2 Angola   Africa     1987    39.9 7874230     2430.     -0.036
##  3 Benin    Africa     2002    54.4 7026113     1373.     -0.371
##  4 Botswana Africa     1992    62.7 1342614     7954.     -0.877
##  5 Botswana Africa     1997    52.6 1536536     8647.    -10.2
##  6 Botswana Africa     2002    46.6 1630347    11004.     -5.92
##  7 Bulgaria Europe     1977    70.8 8797022     7612.     -0.09
##  8 Bulgaria Europe     1992    71.2 8658506     6303.     -0.15
##  9 Bulgaria Europe     1997    70.3 8066057     5970.     -0.87
## 10 Burundi  Africa     1992    44.7 5809236      632.     -3.48
## # ... with 92 more rows
```

```
gapminder_mutated%>%
    filter(year!=1952)%>%
  filter(difference==min(difference))
```

```
## # A tibble: 1 x 7
##   country continent  year lifeExp     pop gdpPercap difference
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>      <dbl>
## 1 Rwanda  Africa     1992    23.6 7290203      737.      -20.4
```

```
gapminder_mutated%>%
    filter(year!=1952)%>%
  filter(difference==max(difference))
```

```
## # A tibble: 1 x 7
##   country  continent  year lifeExp     pop gdpPercap difference
##   <fct>    <fct>     <int>   <dbl>   <int>     <dbl>      <dbl>
## 1 Cambodia Asia       1982    51.0 7272485      624.       19.7
```

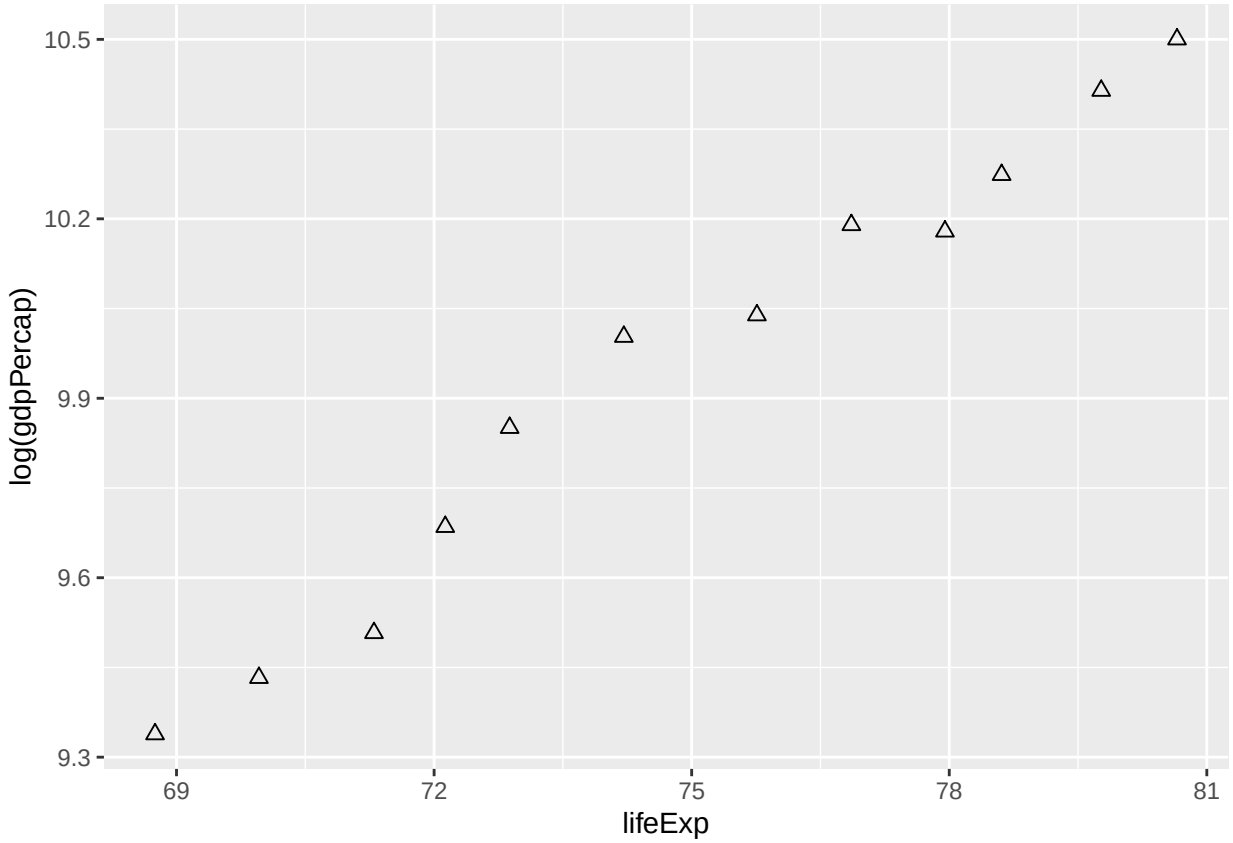
The following shows the maximum GDP per capita experienced by each country:

```
gapminder%>%
  group_by(country)%>%
  filter(gdpPercap==max(gdpPercap))
```

```
## # A tibble: 142 x 6
## # Groups:   country [142]
##    country     continent  year lifeExp       pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>     <int>     <dbl>
##  1 Afghanistan Asia       1982    39.9  12881816      978.
##  2 Albania     Europe     2007    76.4   3600523     5937.
##  3 Algeria     Africa     2007    72.3  33333216     6223.
##  4 Angola      Africa     1967    36.0   5247469     5523.
##  5 Argentina   Americas   2007    75.3  40301927    12779.
##  6 Australia   Oceania    2007    81.2  20434176    34435.
##  7 Austria     Europe     2007    79.8   8199783    36126.
##  8 Bahrain     Asia       2007    75.6    708573    29796.
##  9 Bangladesh  Asia       2007    64.1 150448339     1391.
## 10 Belgium     Europe     2007    79.4  10392226    33693.
## # ... with 132 more rows
```

Here is a scatterplot showing Canada's life expectancy versus GDP per capita (logged):

```
gapminder%>%
  filter(country=="Canada")%>%
  ggplot(aes(lifeExp, log(gdpPercap)))+
  geom_point(size=2, shape=2)
```



Exercise 2

Exploring countries

There are 142 distinct countries represented in the gapminder dataset.

```
gapminder%>%
  distinct(country)
```

```
## # A tibble: 142 x 1
##    country
##    <fct>
##  1 Afghanistan
##  2 Albania
##  3 Algeria
##  4 Angola
##  5 Argentina
##  6 Australia
##  7 Austria
##  8 Bahrain
##  9 Bangladesh
## 10 Belgium
## # ... with 132 more rows
```

We can randomly select 10 distinct countries to have a feel of the possible values.

```
gapminder%>%
  sample_n(10)%>%
  distinct()%>%
  select(country)
```

```
## # A tibble: 10 x 1
##    country
##    <fct>
##  1 Serbia
##  2 Mozambique
##  3 Czech Republic
##  4 Tanzania
##  5 Nicaragua
##  6 Guatemala
##  7 Mongolia
##  8 Mongolia
##  9 Cuba
## 10 Sri Lanka
```

We can find out how many countries there are in each continent, with Africa having the highest number of distinct countries (52) and Oceania having the least number of distinct countries (2).

```
gapminder%>%
  group_by(continent)%>%
  mutate(no_of_countries=n()/12)%>%
  select(continent, no_of_countries)%>%
  distinct()
```

```
## # A tibble: 5 x 2
## # Groups:   continent [5]
##   continent no_of_countries
##   <fct>               <dbl>
## 1 Asia                   33
## 2 Europe                 30
## 3 Africa                 52
## 4 Americas               25
## 5 Oceania                 2
```

Exploring life expectancy

We can obtain summary statistics for life expectancy, including the mininum value, 1st quartile, median, mean, 3rd quantile and maximum value.

The *range* for life expectancy is (23.60, 82.60), and its *IQR* is 22.65.

The *mean* life expectancy is 59.47 and the *median* life expectancy is 60.71.

```
gapminder%>%
  select(lifeExp)%>%
  summary()
```

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

The country with the lowest life expectancy is Rwanda in 1992 and the country with the highest is Japan in 2007.

```
gapminder%>%
  filter(lifeExp==min(lifeExp))%>%
  select(country, year)
```

```
## # A tibble: 1 x 2
##   country  year
##   <fct>   <int>
## 1 Rwanda   1992
```

```r
gapminder%>%
  filter(lifeExp==max(lifeExp))%>%
  select(country, year)
```

```
## # A tibble: 1 x 2
##   country  year
##   <fct>   <int>
## 1 Japan    2007
```

We can also look at which continents have the highest and lowest average life expectancies in the world. Africa has the lowest average life expectancy at 49 years, while Oceania had the highest average life expectancy at 74 years.

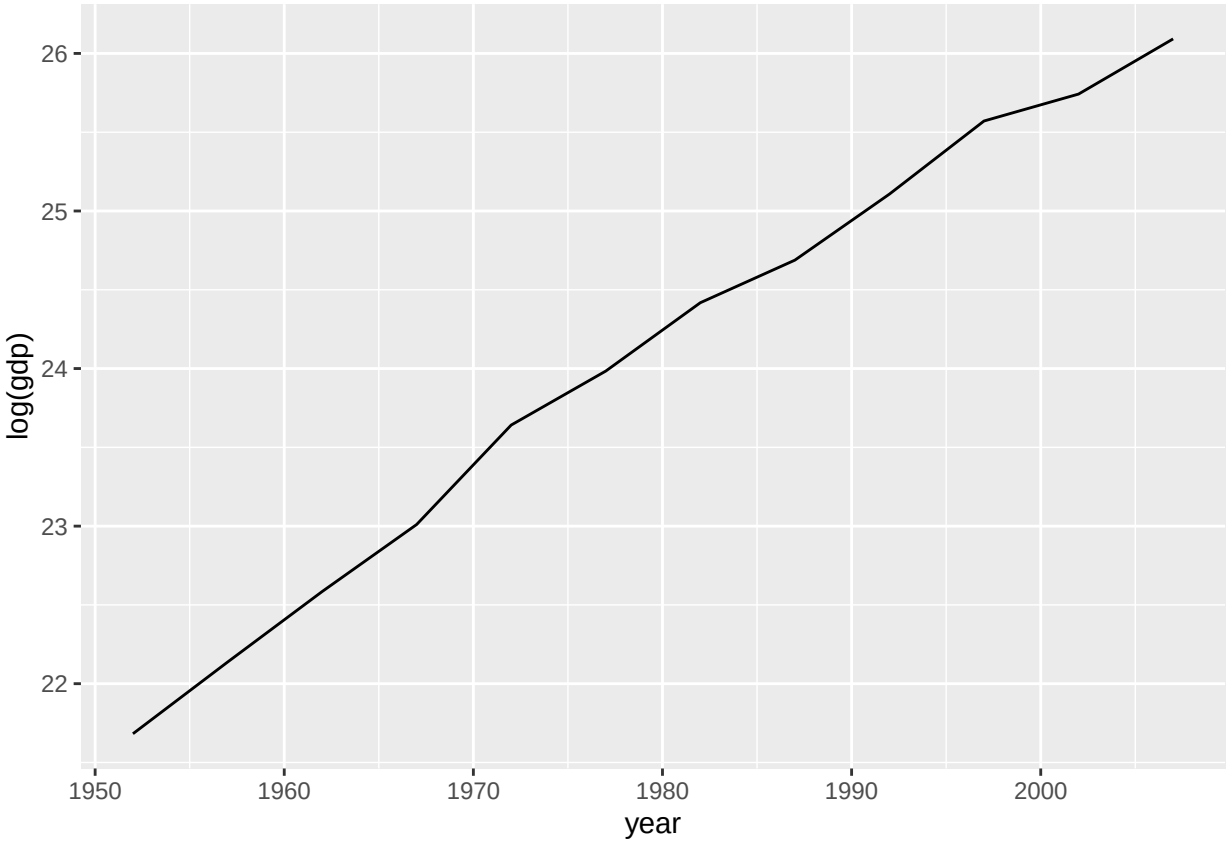
```r
gapminder%>%
  group_by(continent)%>%
  summarise(mean(lifeExp))
```

```
## # A tibble: 5 x 2
##   continent `mean(lifeExp)`
##   <fct>                <dbl>
## 1 Africa                48.9
## 2 Americas              64.7
## 3 Asia                  60.1
## 4 Europe                71.9
## 5 Oceania               74.3
```

Exercise 3

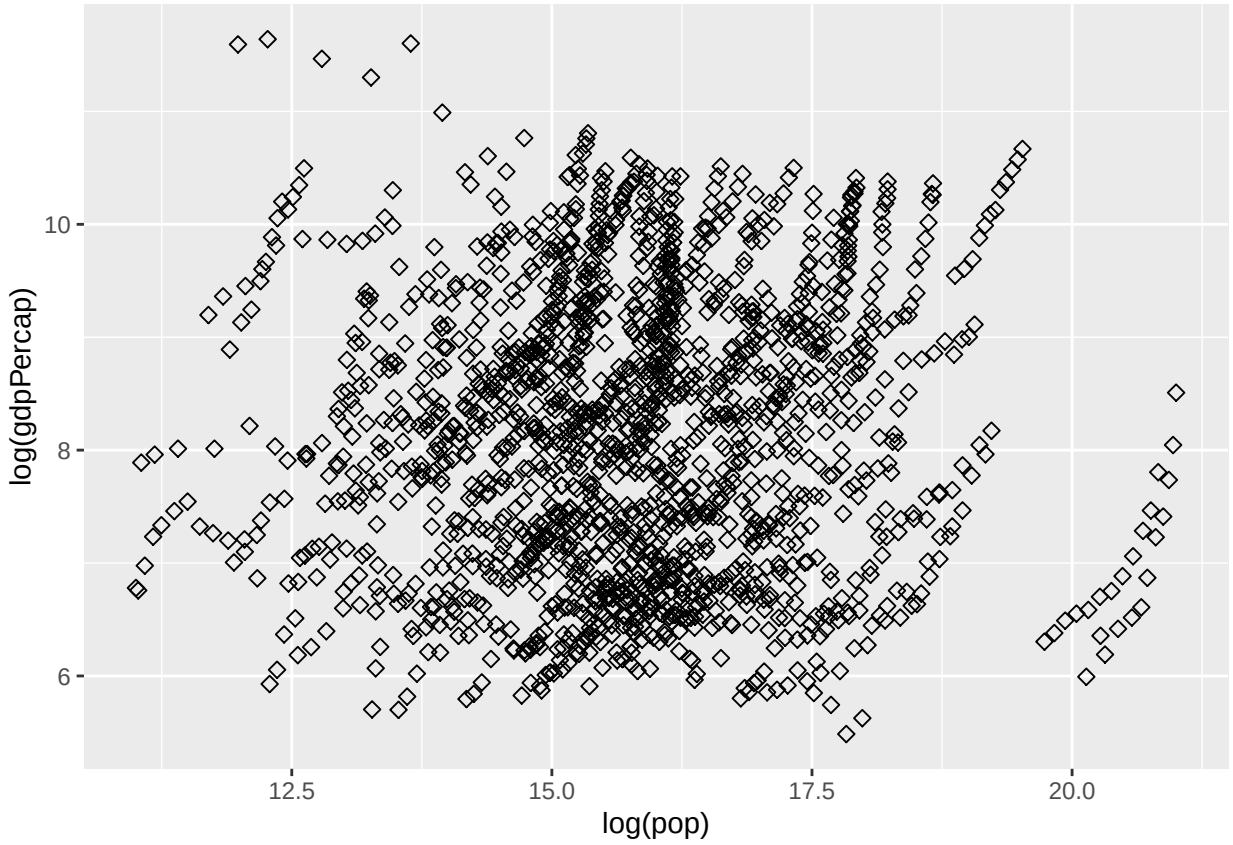
Let's look at a time series of GDP across time for Singapore. We can notice a positive trend in which GDP is increasing across time.

```
gapminder%>%
  mutate(gdp=gdpPercap*pop)%>%
  filter(country=="Singapore")%>%
  ggplot(aes(year, log(gdp)))+
  geom_line()
```



Let's now look at the relationship between population size and GDP per capita to see if larger countries have an economic advantage. From the scatterplot below though, it seems like this is not the case. Conversely, small countries seem to have an economic advantage.

```
gapminder%>%
  ggplot(aes(log(pop), log(gdpPercap)))+
  geom_point(size=2, shape=23)
```



Let's now look at the average GDP per capita for each continent.The boxplots have been arranged in order of increasing magnitude to better reflect the differences between continents. Again, we can see that despite having the most number of countries, Africa has the lowest median GDP per capita. On the other hand, despite having the least number of countries, Oceania has the highest median GDP per capita.

It would be good to label the outliers in the boxplots as well, and in the subsequent weeks, I hope I can figure out how to do that.

```
gapminder%>%
  ggplot(aes(x=reorder(continent, log(gdpPercap), FUN=median), log(gdpPercap)))+
  geom_boxplot(outlier.colour="red")+
  xlab("Continent")
```