Practical session 5, SCIENTIFIC PROGRAMMING

Bet Bardají Bofill

2022-11-21

ggplot2

Exercise 1: Show the number of columns and number of observations of the dataset, and print out the names of the variables

```
gapminder
```

```
## # A tibble: 1,704 × 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.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # … with 1,694 more rows
```

We have a data frame with more than a thousand observations(rows)data and 6 variables(columns).

Name of the variables:

```
gapminder%>%
  str()
```

```
## tibble [1,704 × 6] (S3: tbl_df/tbl/data.frame)
##  $ 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 [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 199
2 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460
14880372 12881816 13867957 16317921 22227415 ...
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

```
colnames(gapminder)
```

```
## [1] "country"   "continent" "year"      "lifeExp"   "pop"       "gdpPe
rcap"
```

We have 6 variables named country, continent,year,lifeExp,pop and gdpPercap

Print the names of the Countries in the dataset. How many countries are there?

```
gapminder$country
```

```
##   [1] Afghanistan              Albania                  Algeria
##   [4] Angola                   Argentina                Australia
##   [7] Austria                  Bahrain                  Bangladesh
##  [10] Belgium                  Benin                    Bolivia
##  [13] Bosnia and Herzegovina   Botswana                 Brazil
##  [16] Bulgaria                 Burkina Faso             Burundi
##  [19] Cambodia                 Cameroon                 Canada
##  [22] Central African Republic Chad                     Chile
##  [25] China                    Colombia                 Comoros
##  [28] Congo, Dem. Rep.         Congo, Rep.              Costa Rica
##  [31] Cote d'Ivoire            Croatia                  Cuba
##  [34] Czech Republic           Denmark                  Djibouti
##  [37] Dominican Republic       Ecuador                  Egypt
##  [40] El Salvador              Equatorial Guinea        Eritrea
##  [43] Ethiopia                 Finland                  France
##  [46] Gabon                    Gambia                   Germany
##  [49] Ghana                    Greece                   Guatemala
##  [52] Guinea                   Guinea-Bissau            Haiti
##  [55] Honduras                 Hong Kong, China         Hungary
##  [58] Iceland                  India                    Indonesia
##  [61] Iran                     Iraq                     Ireland
##  [64] Israel                   Italy                    Jamaica
##  [67] Japan                    Jordan                   Kenya
##  [70] Korea, Dem. Rep.         Korea, Rep.              Kuwait
##  [73] Lebanon                  Lesotho                  Liberia
##  [76] Libya                    Madagascar               Malawi
##  [79] Malaysia                 Mali                     Mauritania
##  [82] Mauritius                Mexico                   Mongolia
##  [85] Montenegro               Morocco                  Mozambique
##  [88] Myanmar                  Namibia                  Nepal
##  [91] Netherlands              New Zealand              Nicaragua
##  [94] Niger                    Nigeria                  Norway
##  [97] Oman                     Pakistan                 Panama
```

```
## [100] Paraguay                 Peru                      Philippines
## [103] Poland                   Portugal                  Puerto Rico
## [106] Reunion                  Romania                   Rwanda
## [109] Sao Tome and Principe    Saudi Arabia              Senegal
## [112] Serbia                   Sierra Leone              Singapore
## [115] Slovak Republic          Slovenia                  Somalia
## [118] South Africa             Spain                     Sri Lanka
## [121] Sudan                    Swaziland                 Sweden
## [124] Switzerland              Syria                     Taiwan
## [127] Tanzania                 Thailand                  Togo
## [130] Trinidad and Tobago      Tunisia                   Turkey
## [133] Uganda                   United Kingdom            United States
## [136] Uruguay                  Venezuela                 Vietnam
## [139] West Bank and Gaza       Yemen, Rep.               Zambia
## [142] Zimbabwe
## 142 Levels: Afghanistan Albania Algeria Angola Argentina Australia ...
Zimbabwe

attach(gapminder)
head(country)

## [1] Afghanistan Afghanistan Afghanistan Afghanistan Afghanistan Afghan
istan
## 142 Levels: Afghanistan Albania Algeria Angola Argentina Australia ...
Zimbabwe
```

What is the time span of the observations? How often the observations were taken?

gapminder %>% group_by(year)%>%summarise(n=n()) %>% mutate(Freq=n/sum(n))

```
>  gapminder %>% group_by(year) %>% summarise(n = n()) %>% mutate(Freq = n/
sum(n))
# A tibble: 12 × 3
    year     n   Freq
   <int> <int>  <dbl>
 1  1952   142 0.0833
 2  1957   142 0.0833
 3  1962   142 0.0833
 4  1967   142 0.0833
 5  1972   142 0.0833
 6  1977   142 0.0833
 7  1982   142 0.0833
 8  1987   142 0.0833
 9  1992   142 0.0833
10  1997   142 0.0833
11  2002   142 0.0833
12  2007   142 0.0833
```

Here we can observe the number of observations from each year and their frequency

The result gives us the time span of observations which is from 1952 to 2007, taken every 5 years with a frequency of 8.3%

Compute the mean of the continuous variables in the dataset.

```
gapminder%>%
  summarize(mean_gdppercap = mean(gdpPercap),mean_lifeExp=mean(lifeExp))
```

```
> gapminder%>%
+   summarize(mean_gdppercap = mean(gdpPercap),mean_lifeExp=mean(lifeExp))
# A tibble: 1 × 2
  mean_gdppercap mean_lifeExp
           <dbl>        <dbl>
1          7215.         59.5
```

Build three different datasets with only the data from Europe, Asia and the Americas

```
gapminder%>%
  filter(continent=="Europe"|continent=="Asia"|continent=="Americas")
```

```
> gapminder %>%
+   filter(continent == "Europe")
# A tibble: 360 × 6
   country continent  year lifeExp     pop gdpPercap
   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
 1 Albania Europe     1952    55.2 1282697     1601.
 2 Albania Europe     1957    59.3 1476505     1942.
 3 Albania Europe     1962    64.8 1728137     2313.
 4 Albania Europe     1967    66.2 1984060     2760.
 5 Albania Europe     1972    67.7 2263554     3313.
 6 Albania Europe     1977    68.9 2509048     3533.
 7 Albania Europe     1982    70.4 2780097     3631.
 8 Albania Europe     1987    72   3075321     3739.
 9 Albania Europe     1992    71.6 3326498     2497.
10 Albania Europe     1997    73.0 3428038     3193.
# … with 350 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> gapminder %>%
+   filter(continent == "Asia")
# A tibble: 396 × 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.
 7 Afghanistan Asia       1982    39.9 12881816      978.
 8 Afghanistan Asia       1987    40.8 13867957      852.
 9 Afghanistan Asia       1992    41.7 16317921      649.
10 Afghanistan Asia       1997    41.8 22227415      635.
# … with 386 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
> gapminder %>%
+   filter(continent == "Americas")
# A tibble: 300 × 6
   country   continent  year lifeExp      pop gdpPercap
   <fct>     <fct>     <int>   <dbl>    <int>     <dbl>
 1 Argentina Americas   1952    62.5 17876956      5911.
 2 Argentina Americas   1957    64.4 19610538      6857.
 3 Argentina Americas   1962    65.1 21283783      7133.
 4 Argentina Americas   1967    65.6 22934225      8053.
 5 Argentina Americas   1972    67.1 24779799      9443.
 6 Argentina Americas   1977    68.5 26983828     10079.
 7 Argentina Americas   1982    69.9 29341374      8998.
 8 Argentina Americas   1987    70.8 31620918      9140.
 9 Argentina Americas   1992    71.9 33958947      9308.
10 Argentina Americas   1997    73.3 36203463     10967.
# … with 290 more rows
# ℹ Use `print(n = ...)` to see more rows
```

##Compute the Maximum and minimum of the continuous variables for each continent (irrespectively of the country/year)

```
gapminder%>%group_by(continent)%>%summarize(max_lifeExp = max(lifeExp),mi
n_lifeExo=min(lifeEx p))
```

```
> gapminder%>%
+   group_by(continent)%>%
+   summarize(max_lifeExp = max(lifeExp),min_lifeExo=min(lifeExp))
# A tibble: 5 x 3
  continent max_lifeExp min_lifeExo
  <fct>           <dbl>       <dbl>
1 Africa           76.4        23.6
2 Americas         80.7        37.6
3 Asia             82.6        28.8
4 Europe           81.8        43.6
5 Oceania          81.2        69.1
>
```

```
gapminder%>%
   group_by(continent)%>%
summarize(max_gdppercap = max(gdpPercap),min_gdppercap=min( gdpPercap))
```

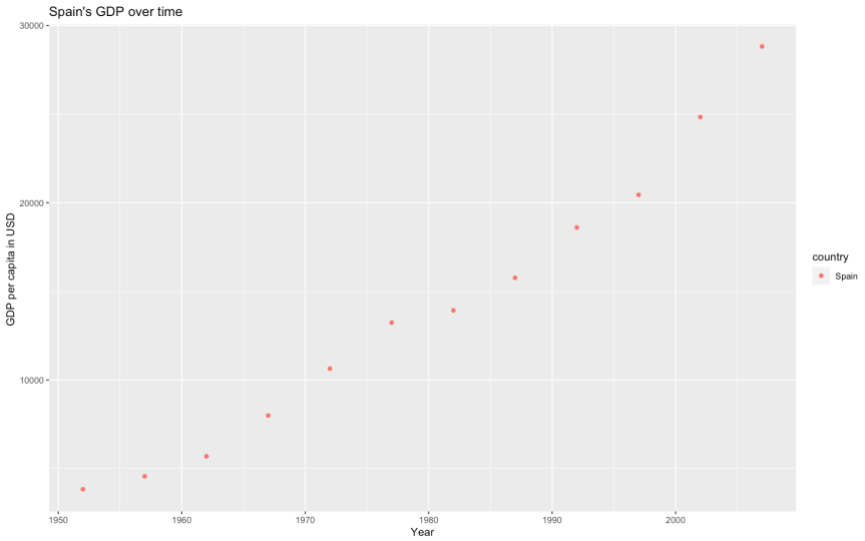
```
> gapminder%>%
+   group_by(continent)%>%
+   summarize(max_gdppercap = max(gdpPercap),min_gdppercap=min(gdpPerca
p))
# A tibble: 5 x 3
  continent max_gdppercap min_gdppercap
  <fct>             <dbl>         <dbl>
1 Africa            21951.          241.
2 Americas          42952.         1202.
3 Asia             113523.          331
4 Europe            49357.          974.
5 Oceania           34435.        10040.
>
```

PLOTTING:

##Investigate on how to add titles and labels to the x and y axis to the plots.
- With xlab and ylab

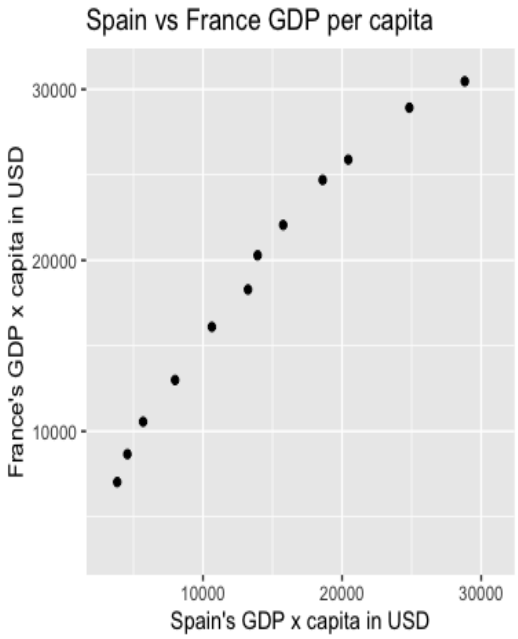
##Plot Spanish GDP vs time

```
spain <- gapminder%>% filter(country=="Spain")
ggplot()+ aes(x=year, y=gdpPercap, col=country)) + geom_point() + xlab("Y
ear") + ylab("GDP per capita in USD ") + ggtitle("Spain's GDP over time")
```



##Obtain a scatter plot with the GDP of France vs. the GDP of Spain (x GDP for Spain, y GDP for France).

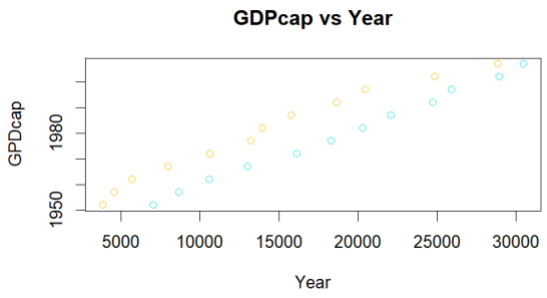
```
france <- gapminder[gapminder$country == "France",]
spainfrance <- data.frame("spain_gdp" = spain$gdpPercap, "france_gdp" = france$g
dpPercap)

ggplot(spainfrance, aes(x = spain_gdp, y = france_gdp)) + geom_point() + xlim(30
00,31000) + ylim(3000,31000) + theme(aspect.ratio = 1) + xlab("Spain's GDP x cap
ita in USD") +
  ylab("France's GDP x capita in USD") +
  ggtitle("Spain vs France GDP per capita")
```



Additional: sctatterplot of spain and France gdpx cspita over the years

Code:

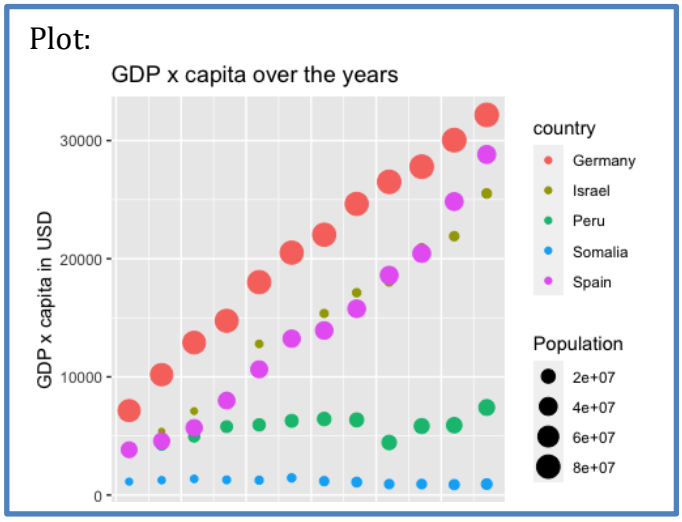
```
spafran <- gapminder [gapminder$country
%in% c("Spain", "France"), ]

spain <- spafran$gdpPercap

frann <- spafran$gdpPercap


plot(gdpPercap, year, main = "GDPcap vs
Year",

   xlab = "Year", ylab = "GPDcap",col=
country,

   data=spafran)
```



Plot the GDP vs. time for Germany, Israel, Peru, Somalia and Spain on a single plot. Set each country with a different color an the size of each point in the trend indicating the population size.

```
gapmind2 <- gapminder [gapminder$country %in% c("Spain", "Germany", "Soma
lia","Israel","Peru"), ]
 head(gapmind2)

## # A tibble: 6 × 6
##   country continent  year lifeExp      pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Germany Europe     1952    67.5 69145952     7144.
## 2 Germany Europe     1957    69.1 71019069    10188.
## 3 Germany Europe     1962    70.3 73739117    12902.
## 4 Germany Europe     1967    70.8 76368453    14746.
## 5 Germany Europe     1972    71   78717088    18016.
## 6 Germany Europe     1977    72.5 78160773    20513.

gapmind2 %>%    ggplot() +  aes(x = year, y = gdpPercap, col=country,size
=pop) +    geom_point() +xlab("year") + ylab("GDP x capita in USD") + ggt
itle("GDP x capita over the years") + scale_size_continuous(name = "Popul
ation")
```

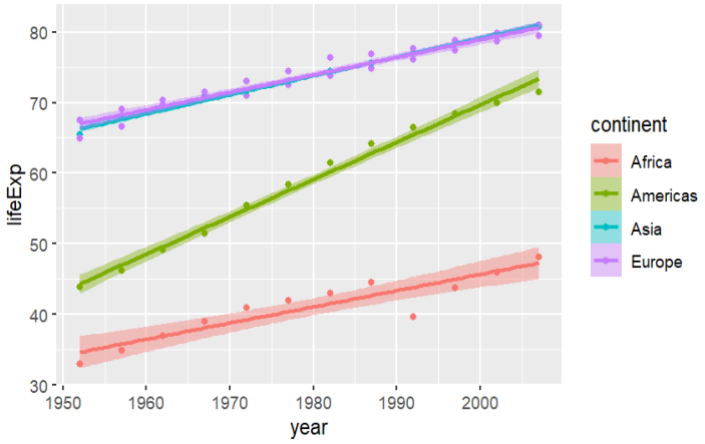
Plot:



Same plot with lines to show the tendency:

```
ggplot(gapmind2, aes(x=year, y=lifeExp, color=continent)) +
geom_point(size=1.5) +  geom_smooth(aes(fill=continent), method="lm")
```

Plot with lines to show the tendency:



Exploratory analysis.

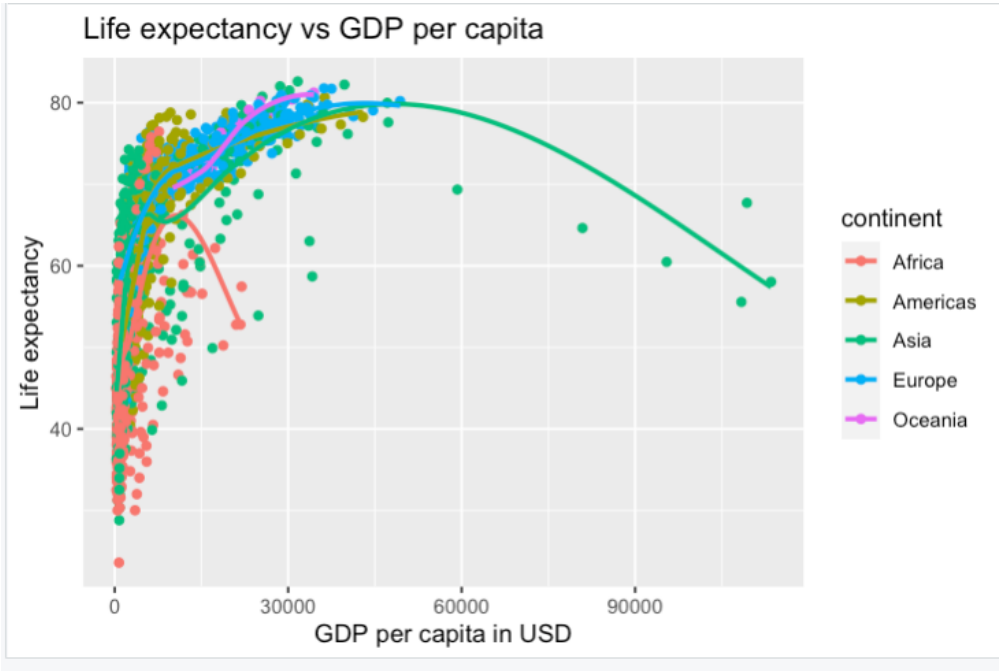
Exercise 1:

Is there a relationship between GDP per capita and life expectancy. Are there differences between continents? Provide your answer with plots supporting your observations.

```
cor(gapminder$gdpPercap,gapminder$lifeExp)
```

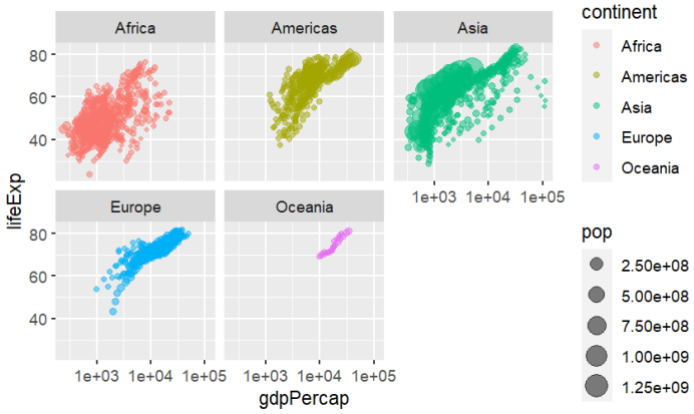
```
> cor(a$gdpPercap, a$lifeExp)
[1] 0.5837062
> attach(gapminder)
> ggplot(a) +
+   aes(x = gdpPercap, y = lifeExp) +
+   geom_point(colour = "#0c4c8a") +
+   theme_minimal()
```

```
Graphical representation:
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, color= continent)) +
geom_point() +
  stat_smooth(se=FALSE) + xlab("GDP per capita in USD") + ylab("Life
expectancy") +
  ggtitle("Life expectancy vs GDP per capita")
```

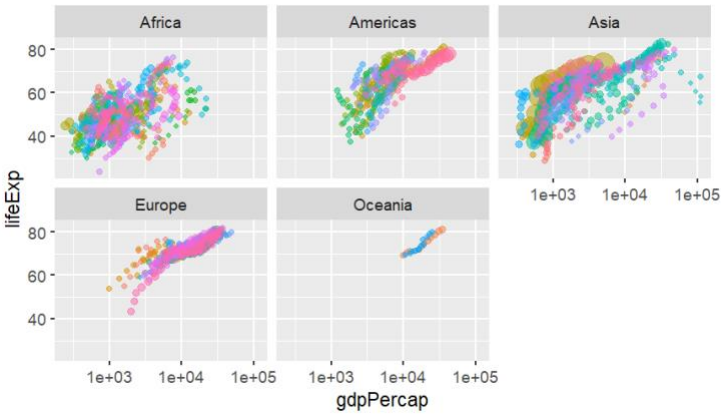


Code:

```
gapminder %>%
  ggplot() +
  aes(x = gdpPercap, y =
lifeExp, color =
continent,size=pop) +
  geom_point() +
  facet_grid( continent)
```



```
Code
gapminder %>%

  ggplot() +
  aes(x = gdpPercap, y =
lifeExp, color = country) +
  geom_point() +
  facet_grid(continent)
```

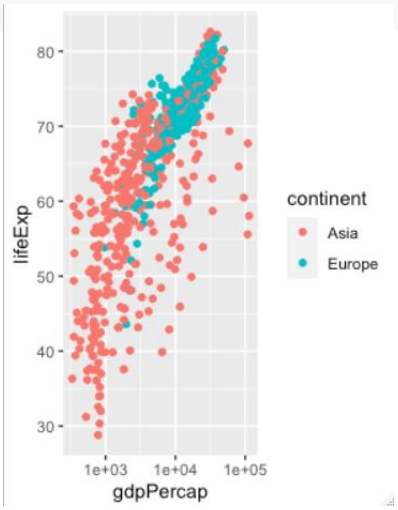


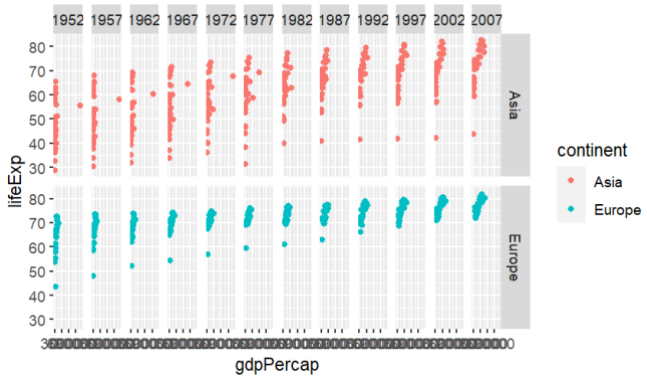
The correlation coefficient between gdpPercap and lifeExp is 58%. This mens two things. - 1. The have a positive correlation - 2. The have some correlation but it is not really strongh. In the graph, it seems that there's a strong connection until 10000 USD per capita is reached. Beyond that,the coefficient goes down. Lower GDP is related to lower life expectancy due to poverty, meanwhile, with higher GDP life expectancy tend to grow. In Asia we can see some visible outliers, especially.

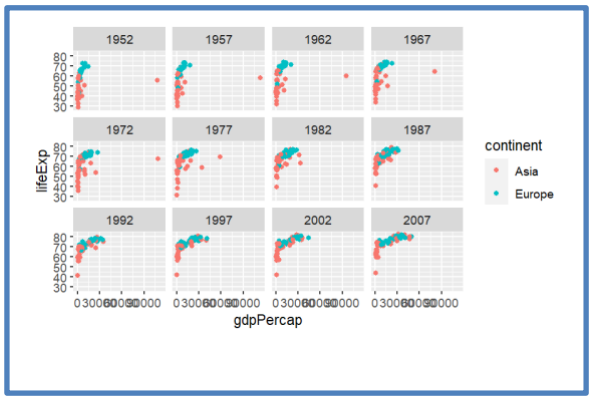
Exercise 2:

Take a closer look to the behavior of variables in different continents. For example, take Europe and Asia. Is life expectancy directly related to the GDP per capita (that, is, life expectancy is higher for countries with higher GDP?). Does this statement holds for all countries on a given continent?

```
gapminderx<-gapminder%>%
filter(continent=='Europe'|continent
=='Asia')

ggplot(data=gapminderx,
aes(x=gdpPercap, y=lifeExp,
col=continent)) +

 geom_point() + scale_x_log10()
```





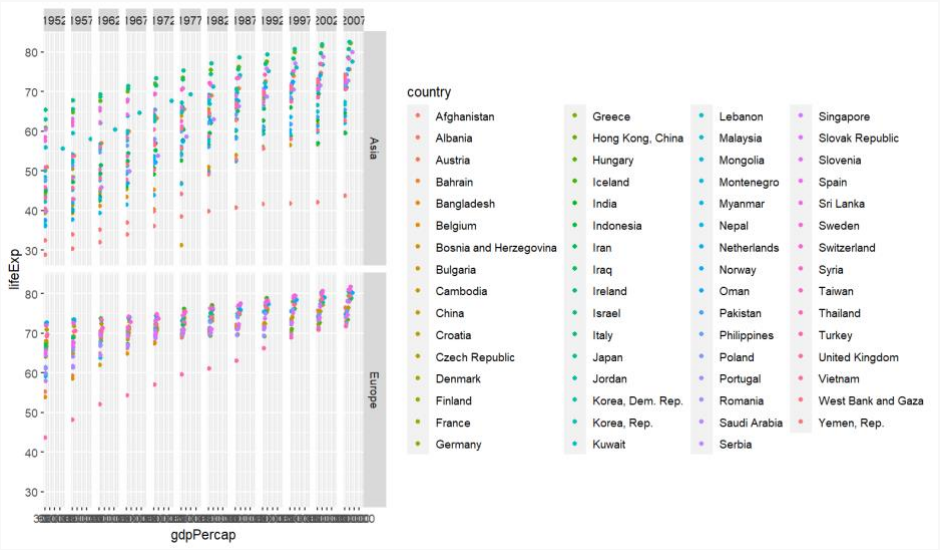



```
vector_continents <- c("Europe", "Asia")

gapminder %>%
  filter(%in% vector_continents) %>%
  ggplot() +
  aes(x = gdpPercap, y = lifeExp, color = continent) +
  geom_point() +
  facet_grid(continent ~ year)
```

I decided to include ethe years to obtain stronger conclusions. As we can appreciate both continents have increased their economy over the years, although Europe has always presented higher life expectancy even though when the gdp were similar. I believe that one of the reasons is that Europe is a more homogeneous continent in terms of lifestyle and diet compared to Asia.

1. Yes it seems to be true that life expectancy and gdp are positive correlative. Although we should apply a correlation test to prove it.

```r
vector_continents <- c("Europe", "Asia")

gapminder %>%
  filter(year & continent %in% vector_continents) %>%
  ggplot() +
  aes(x = gdpPercap, y = lifeExp, color = country) +
  geom_point() +
  facet_grid(continent ~ year)
```

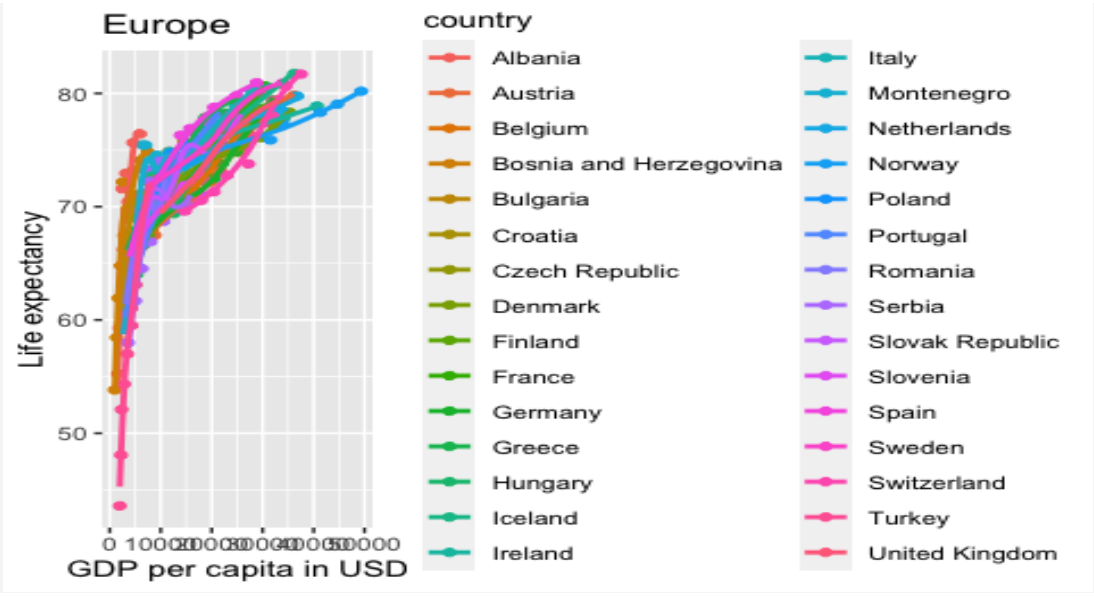


2. Yes.As we can appreciate, it does seem like countries in a continent with lower gdp present also lower life expectancy.

More visualization:

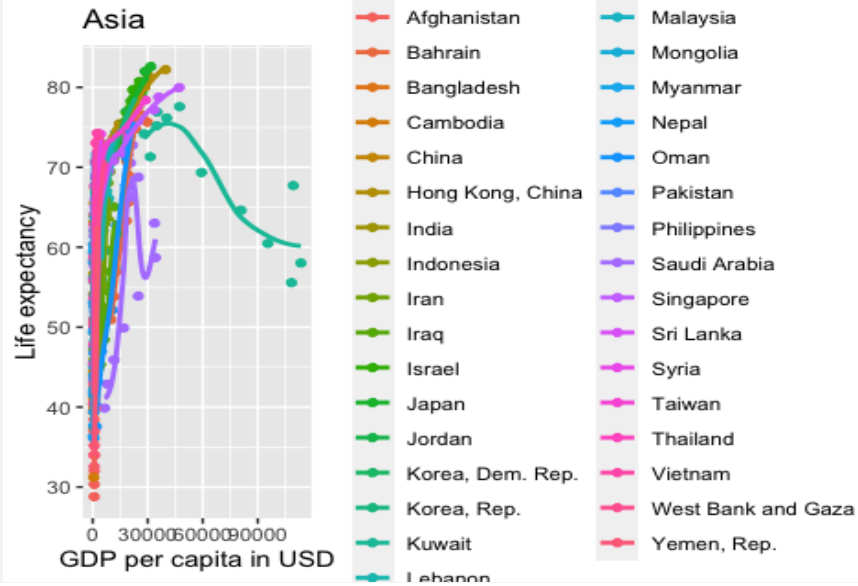
```r
europe <- gapminder%>%filter(continent=="Europe")
ggplot(europe, aes(x = gdpPercap, y = lifeExp, color= country)) + geom_po
int() +
  geom_smooth(se=FALSE) + ggtitle("Europe") + xlab("GDP per capita in USD
") +
  ylab("Life expectancy")
```

Plot:




```
asia <- gapminder%>% filter(continent=="Asia")
ggplot(asia, aes(x = gdpPercap, y = lifeExp, color= country)) + geom_poin
t() + geom_smooth(se=FALSE) + ggtitle("Asia") + xlab("GDP per capita in U
SD") + ylab("Life expectancy")
```

Plot:

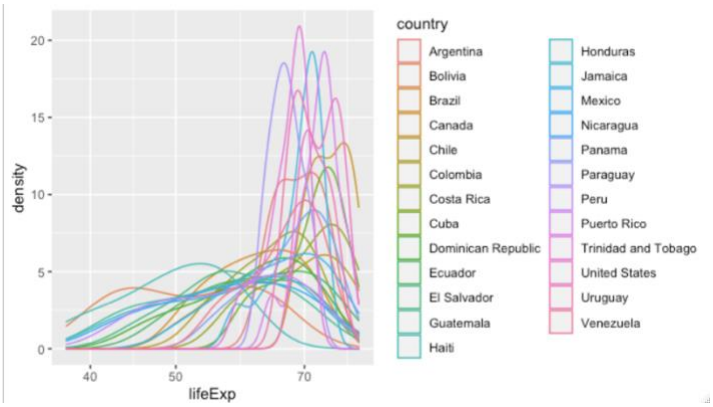


To conclude, life expectancy is significantly tied to GDP per capita, particularly in Europe. Nonetheless, there are several outliers in Asia

Exercise 3: Observe the density estimates of life expectancy in the Americas for the different countries in the dataset. You can use geom_density() for this purpose. Is the distribution of life expectancy similar for different countries? If not, try to explain this difference by inspecting other variables in the dataset. Remember to support your observations by providing the necessary plots.

Code:

```
americ <- gapminder %>%
 filter(continent =="Americas") %>%
 select(country,year,lifeExp,pop,gdpPercap)
ggplot(data=americ,aes(x=lifeExp,col=country))
plt + geom_density(alpha=0.3) + scale_x_log10()
```



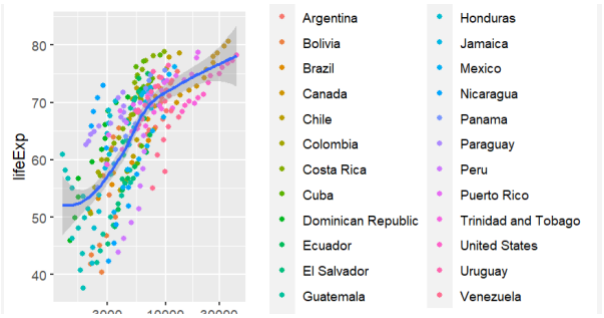
As we can see, the distribution is not equitable within the countries in America. The countries that present higher concentrated life expectancy are the United States and Puerto Rico and the lower is concentrated in Honduras. In addition, the most common life expectancy seems to be around 70 years, with a trend towards 80..

Below we find a plot comparing the life expectancy with the gdp to understand this difference. Again, the higher gdp the higher life expectancy

Code:

```
x <- gapminder %>%
  filter(continent == "Americas") %>%
  select(country, year, lifeExp, pop, gdpPercap)

> plt <- ggplot(data=x,
+               aes(x=gdpPercap, y=lifeExp))
> plt + geom_point(aes(color=country)) +
+     geom_smooth(method="loess") +
+     scale_x_log10()
```

Plot:



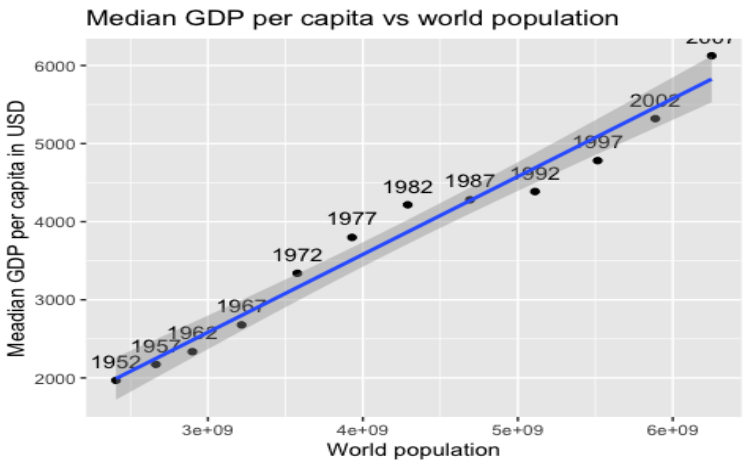
Exercise 1: Obtain the global population, the expected (median) GDP per capita and the mean life expectancy for each year. Plot the global population vs. the expected GDP per capita for each year. To make your plot clearer, use labels for each point displaying the year. Is the relationship between population size and GDP linear?

```
gm_by_year <- group_by(gapminder, year)
globalpop <- summarise(gm_by_year,
                       pop_num = sum(pop),
                       medianGdp = median(gdpPercap),
                       meanLifeExp = mean(lifeExp))
globalpop
```

```
## # A tibble: 12 × 4
##     year    pop_num medianGdp meanLifeExp
##    <int>      <dbl>     <dbl>       <dbl>
##  1  1952 2406957150     1969.        49.1
##  2  1957 2664404580     2173.        51.5
##  3  1962 2899782974     2335.        53.6
##  4  1967 3217478384     2678.        55.7
##  5  1972 3576977158     3339.        57.6
##  6  1977 3930045807     3799.        59.6
##  7  1982 4289436840     4216.        61.5
##  8  1987 4691477418     4280.        63.2
##  9  1992 5110710260     4386.        64.2
## 10  1997 5515204472     4782.        65.0
## 11  2002 5886977579     5320.        65.7
## 12  2007 6251013179     6124.        67.0
```

Graphical representation:

```
ggplot(globalpop, aes(x=pop_num, y=medianGdp, label=year)) + geom_point()
+
  geom_text(vjust=-1) + geom_smooth(method="lm") +
  ggtitle("Median GDP per capita vs world population") +
  ylab("Meadian GDP per capita in USD") + xlab("World population")
```



The graph shows us a linear relationship, although some of the data points fall outside of the expected confidence interval(CI)

Exercise 2. What is the difference between filter() and select()?

```
help("filter")

## Help on topic 'filter' was found in the following packages:
##
##   Package               Library
##   stats                 /Library/Frameworks/R.framework/Versions/4.2/R
esources/library
##   dplyr                 /Library/Frameworks/R.framework/Versions/4.2/R
esources/library
##
##
## Using the first match ...

help("select")
```

- The main difference is that filter() operates on rows, whereas select() operates on columns

Exercise 3. Using pipes, write code to answer the following question: how many countries per continent are there in the dataset?

```
countrcont <- gapminder %>%
  group_by(continent) %>%
  summarise(n_obs= n(), n_countries =n_distinc(country))
countrcont

## # A tibble: 5 × 3
##   continent n_obs  n_countries
```

```
##   <fct>     <int>  <int>
## 1 Africa     624     52
## 2 Americas   300     25
## 3 Asia       396     33
## 4 Europe     360     30
## 5 Oceania    24       2
```

Here we can see that Africa is the continent with the highest number of countries, followed by Asia and Europe.

Exercise 4. Which are the top 3 countries with the highest GDP per capita for each continent in the year 2007 and which percentage of the total GDP of the continent do they represent? *Hint: use the verbs* mutate() *and* slice_max(). *If you are doing it correctly you should obtain a 100% when summing all countries in Oceania*.

```
top3 <- gapminder %>%
  group_by(continent) %>%
  filter(year==2007) %>%
  mutate(gdpratio = 100 * (gdpPercap*pop) /sum(gdpPercap *pop))  %>%
  top_n(3,gdpPercap)
select(country, continent, gdpPercap, gdpratio))

top3
```

```
## # A tibble: 14 × 4
## # Groups:   continent [5]
##    country           continent gdpPercap gdpratio
##    <fct>             <fct>         <dbl>    <dbl>
##  1 Gabon             Africa       13206.    0.807
##  2 Botswana          Africa       12570.    0.866
##  3 Equatorial Guinea Africa       12154.    0.281
##  4 United States     Americas     42952.   66.6
##  5 Canada            Americas     36319.    6.25
##  6 Puerto Rico       Americas     19329.    0.392
##  7 Kuwait            Asia         47307.    0.572
##  8 Singapore         Asia         47143.    1.04
##  9 Hong Kong, China  Asia         39725.    1.34
## 10 Norway            Europe       49357.    1.54
## 11 Ireland           Europe       40676.    1.13
## 12 Switzerland       Europe       37506.    1.92
## 13 Australia         Oceania      34435.   87.2
## 14 New Zealand       Oceania      25185.   12.8
```

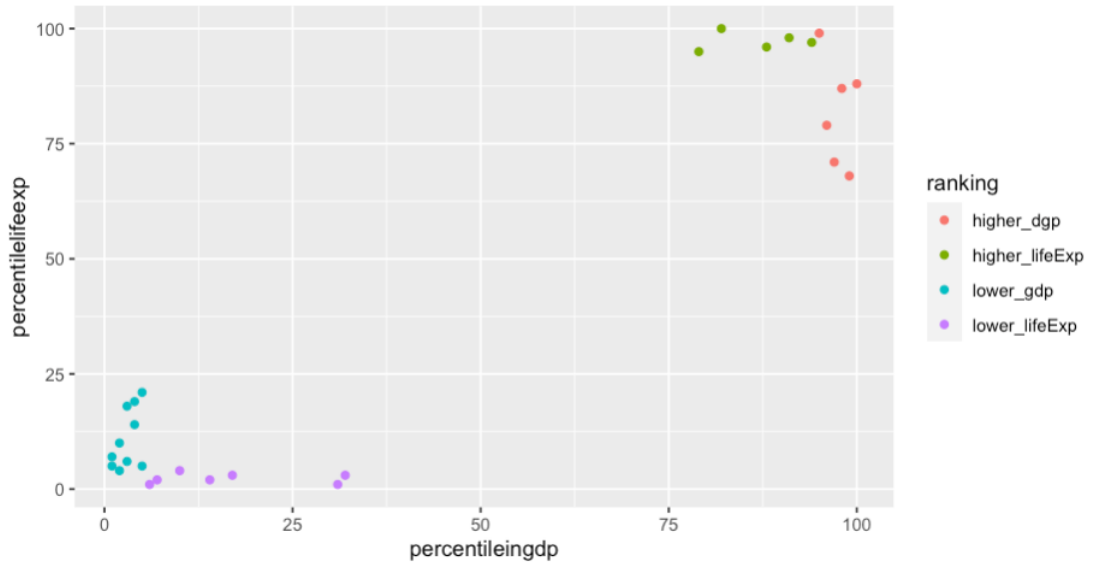
Exercise 5: Is life expectancy higher in countries with high GDP per capita? To solve this question, you can calculate percentiles on the variables lifeExp and gdpPercap for the year 2007. Obtain the countries with the highest and lowest percentiles and comapre them. *Hint: use ntile() for calculating the percentiles*

```
exercici5 <- gapminder %>%
  filter(year==2007) %>%
  mutate(percentileingdp=ntile(gdpPercap, 100)) %>%
  mutate(percentilelifeexp=ntile(lifeExp, 100)) %>%
  mutate(rating = case_when(percentileingdp >= 95 ~ 'higher_gdp',
                            percentilelifeexp >= 95 ~ 'higher_lifeExp',
                            percentileingdp <= 5 ~ 'lower_gdp',
                            percentilelifeexp <= 5 ~ 'lower_lifeExp')) %>
%
  subset(!is.na(rating))
```

```
> exercici5
# A tibble: 28 × 9
   country                 contin…¹  year lifeExp    pop gdpPe…² perce…³ perce…⁴ ranking
   <fct>                   <fct>    <int>   <dbl>  <int>   <dbl>   <int>   <int> <chr>
 1 Afghanistan             Asia      2007    43.8 3.19e7    975.      10       4 lower_…
 2 Angola                  Africa    2007    42.7 1.24e7   4797.      32       3 lower_…
 3 Australia               Oceania   2007    81.2 2.04e7  34435.      88      96 higher…
 4 Burundi                 Africa    2007    49.6 8.39e6    430.       2      10 lower_…
 5 Central African Republic Africa   2007    44.7 4.37e6    706.       5       5 lower_…
 6 Congo, Dem. Rep.        Africa    2007    46.5 6.46e7    278.       1       7 lower_…
 7 Eritrea                 Africa    2007    58.0 4.91e6    641.       4      19 lower_…
 8 Ethiopia                Africa    2007    52.9 7.65e7    691.       4      14 lower_…
 9 Gambia                  Africa    2007    59.4 1.69e6    753.       5      21 lower_…
10 Guinea-Bissau           Africa    2007    46.4 1.47e6    579.       3       6 lower_…
# … with 18 more rows, and abbreviated variable names ¹continent, ²gdpPercap,
#   ³percentileingdp, ⁴percentilelifeexp
# ℹ Use `print(n = ...)` to see more rows
```

Graphical representation:

```
ggplot(gdpPer, aes(x = gdpPercentile, y = lifePercentile, color=rating))
+ geom_point()
```



It

We can observe that GDP x capita and life expectancy appear to be connected. Lower GDP and life expectancy are clustered together, while greater GDP and life expectancy are grouped together.

Exercise 6: Which continent had the fastest rate of change in population in the period between 1952 and 2007. *Hint: take derivatives*

```
popgrowth <- gapminder %>%
  group_by(continent, year) %>%
  summarise(totalpop = sum(pop)) %>%
  mutate(absolutegrowth=pop_total - lag(pop_total)) %>%
  mutate(relative_growth=absolute_growth/lag(pop_total) * 100)

popgrowth
```

```
# A tibble: 60 × 5
# Groups:   continent [5]
   continent  year  totalpop absolutegr…¹ relat…²
   <fct>     <int>     <dbl>         <dbl>   <dbl>
 1 Africa     1952 237640501            NA      NA
 2 Africa     1957 264837738      27197237    11.4
 3 Africa     1962 296516865      31679127    12.0
 4 Africa     1967 335289489      38772624    13.1
 5 Africa     1972 379879541      44590052    13.3
 6 Africa     1977 433061021      53181480    14.0
 7 Africa     1982 499348587      66287566    15.3
 8 Africa     1987 574834110      75485523    15.1
 9 Africa     1992 659081517      84247407    14.7
10 Africa     1997 743832984      84751467    12.9
# … with 50 more rows, and abbreviated variable
#   names ¹absolutegrowth, ²relativegrowth
# ℹ Use `print(n = ...)` to see more rows
```

```
slice_max(ungroup(pop_cont), n=1, order_by = absolute_growth)
```

```
## A tibble: 1 × 5
##   continent  year  totalpop absolute_growth relative_growth
##   <fct>     <int>     <dbl>           <dbl>           <dbl>
## 1 Asia       1992 3133292191       262071429            9.13
```

```
slice_max(ungroup(pop_cont), n=1, order_by = relative_growth)
```

```
## # A tibble: 1 × 5
##   continent  year pop_total absolute_growth relative_growth
##   <fct>     <int>     <dbl>           <dbl>           <dbl>
## 1 Africa     1982 499348587        66287566            15.3
```

Absolute growth: As we can see on the first table, Asia presents the most growth in terms of population and Africa had the highest relative population growth from 1977 to 1982.

Exercise 7: Compare the growth between developing and developed countries. Take for example the Americas. You can divide them into two regions: Northern America (developed) and Latin America and the Caribbean (developing economies). Is the mean GDP per year increasing at the same rate in both regions? What about the total population and the mean life expectancy? For solving this exercise you will need to filter the Americas and create an additional variable: *region*. For creating this variable, you can use the following vectors

```
latin_america_caribbean <- c('Brazil','Colombia','Argentina','Peru','Vene
zuela','Chile',
```

```
'Ecuador','Bolivia','Paraguay','Uruguay', 'Guyana','Suriname',
'French Guiana','Falkland Islands','Mexico','Guatemala',
'Dominican Republic','Honduras','Nicaragua','El Salvador',
'Costa Rica', 'Panama','Belize','Cuba','Haiti','Puerto Rico',
'Trinidad and Tobago','Guadeloupe','Bahamas','Barbados',
'Saint Lucia','Curaçao','Grenada','Dominican Republic',
'United States Virgin Islands','Aruba','Antigua and Barbuda',
'Dominica','Cayman Islands','Jamaica','Saint-Barthélemy',
'Sint Maarten','Saint Martin','Turks and Caicos Islands',
'British Virgin Islands','Caribbean Netherlands',
'Anguilla','Montserrat','Saint Kitts and Nevis')

northern_america <- c('United States','Canada','Bermuda','Greenland',
'Saint Pierre and Miquelon')

exercici7 <- gapminder %>%
  filter(continent=="Americas") %>%
  mutate(region = case_when(country %in% latin_america_caribbean ~ 'Latin
America',country %in% northern_america ~ 'North America')) %>%
  mutate(gdp=gdpPercap * pop) %>%
  group_by(region, year) %>%
  summarise(mean_gdp = mean(gdp), totalpop = sum(pop), mean_lifeExp =

 mean(lifeExp)) %>%
  mutate(gdpGrowth = (mean_gdp - lag(mean_gdp))/lag(mean_gdp) * 100) %>%
  mutate(popGrowth = (totalpop - lag(totalpop)) / lag(totalpop) * 100) %>
%
  mutate(lifeGrowth = (mean_lifeExp - lag(mean_lifeExp)))
```

```
> exercici7
# A tibble: 24 × 8
# Groups:   region [2]
   region        year      mean_gdp  totalpop mean_lifeExp gdpgrowth popgrowth lifegrowth
   <chr>        <int>         <dbl>     <int>        <dbl>     <dbl>     <dbl>      <dbl>
 1 Latin america  1952  24833147414. 172813862         51.9     NA        NA         NA
 2 Latin america  1957  32804445523. 197959762         54.8     32.1      14.6        2.82
 3 Latin america  1962  41579292861. 227746405         57.3     26.7      15.0        2.56
 4 Latin america  1967  53524323957. 261214856         59.5     28.7      14.7        2.13
 5 Latin america  1972  74096959319. 297203710         61.6     38.4      13.8        2.10
 6 Latin america  1977  98900817749. 334032299         63.6     33.5      12.4        2.02
 7 Latin america  1982 117340945569. 372901185         65.4     18.6      11.6        1.87
 8 Latin america  1987 131445343994. 413400738         67.4     12.0      10.9        1.96
 9 Latin america  1992 142368695048. 453856413         68.9      8.31      9.79       1.51
10 Latin america  1997 170803310507. 493682807         70.6     20.0       8.78       1.66
# … with 14 more rows
# ℹ Use `print(n = ...)` to see more rows
```

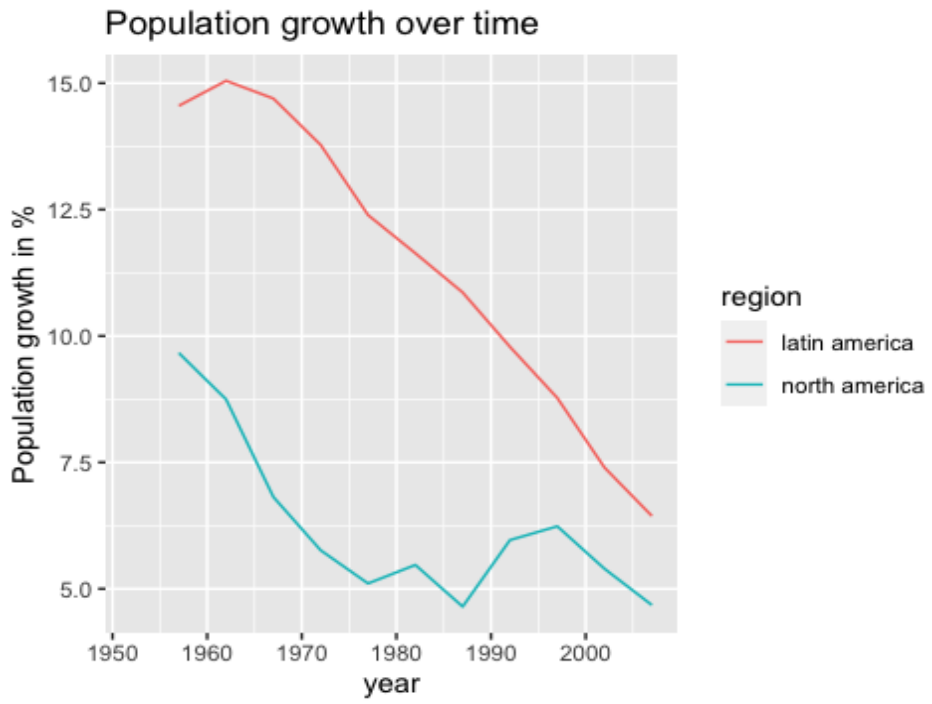
```
ggplot(exercici7, aes(x = year, y = gdpGrowth, color = region)) + geom_li
ne() + ggtitle("GDP growth over time") + ylab("Growth of mean GDP in %")

## Warning: Removed 2 row(s) containing missing values (geom_path).
```

MEAN GROWTH

The mean GDP of Latin and North America, presents a different growth. Latin America seems to present a higher growth in terms of GDP rather than north America. Additionally, though Latin America presents stronger fluctuations than North America.

POPULATION GROWTH

```
ggplot(exercici7, aes(x = year, y = popGrowth, color = region)) + geom_li
ne() +
  ggtitle("Population growth over time") + ylab("Population growth in %")
```



Latin America's population growth rate is substantially larger than that of North America, with lower variations. Both follow similar paths and finish up with comparable growth in the year 2000.

```
ggplot(exercici7, aes(x = year, y = lifeGrowth, color = region)) + geom_l
ine() + ggtitle("Life expectancy over time") + ylab("Increase in life exp
ectancy in years")
```

Life expectancy tends to follow a similar pattern to the other two variables mentioned above. As we can see, Latin America has a higher life expectancy than North America, although they both have a similar level.