lab5

2023-02-14

1 Load and check data (5pt)

You first task is to do a very simple data check:

1. (1pt) For solving the problems, and answering the questions, create a new rmarkdown docu-ment with an appropriate title. See https://faculty.washington.edu/otoomet/info201-book/r-markdown.html#r-markdown-rstudio-creating (https://faculty.washington.edu/otoomet/info201-book/r-markdown.html#r-markdown-rstudio-creating).

2. (2pt) Load data. How many rows/columns do we have?

```
library(readr)
gapminder <- read_delim("gapminder.csv")
```

```
## Rows: 13055 Columns: 25
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (6): iso3, name, iso2, region, sub-region, intermediate-region
## dbl (19): time, totalPopulation, fertilityRate, lifeExpectancy, childMortali...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(gapminder)
```

```
## [1] 13055    25
```

Rows: 13055 Columns: 25

3. (2pt) Print a small sample of data. Does it look OK?

```
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
gapminder %>% sample_n(10)
```

```
## # A tibble: 10 x 25
##    iso3  name         iso2  region sub-r~1 inter~2  time total~3 ferti~4 lifeE~5
##    <chr> <chr>        <chr> <chr>  <chr>   <chr>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 GNQ   Equatorial ~ GQ    Africa Sub-Sa~ Middle~  1963  2.66e5    5.69    37.5
##  2 KGZ   Kyrgyzstan   KG    Asia   Centra~ <NA>     1971  3.02e6    5.15    60.6
##  3 JOR   Jordan       JO    Asia   Wester~ <NA>     1989  3.40e6    5.72    69.6
##  4 PRY   Paraguay     PY    Ameri~ Latin ~ South ~  2013  6.51e6    2.58    73.3
##  5 SXM   Sint Maarte~ SX    Ameri~ Latin ~ Caribb~  1965  4.46e3   NA       NA
##  6 MAC   Macao        MO    Asia   Easter~ <NA>     1976  2.39e5    1.44    72.4
##  7 MAR   Morocco      MA    Africa Northe~ <NA>     1992  2.57e7    3.75    65.7
##  8 LKA   Sri Lanka    LK    Asia   Southe~ <NA>     1987  1.66e7    2.70    69.1
##  9 PAN   Panama       PA    Ameri~ Latin ~ Centra~  2011  3.71e6    2.61    77.0
## 10 SOM   Somalia      SO    Africa Sub-Sa~ Easter~  1999  8.55e6    7.66    50.4
## # ... with 15 more variables: childMortality <dbl>, youthFemaleLiteracy <dbl>,
## #   youthMaleLiteracy <dbl>, adultLiteracy <dbl>, GDP_PC <dbl>,
## #   accessElectricity <dbl>, agriculturalLand <dbl>, agricultureTractors <dbl>,
## #   cerealProduction <dbl>, fertilizerHa <dbl>, co2 <dbl>,
## #   greenhouseGases <dbl>, co2_PC <dbl>, pm2.5_35 <dbl>, battleDeaths <dbl>,
## #   and abbreviated variable names 1: `sub-region`, 2: `intermediate-region`,
## #   3: totalPopulation, 4: fertilityRate, 5: lifeExpectancy
```

2 Descriptive statistics (15pt)

1. (3pt) How many countries are there in the dataset? Analyze all three: iso3, iso2 and name.

```
n_iso3 <- gapminder %>% distinct(iso3) %>% nrow()
n_iso3
```

```
## [1] 253
```

```
n_iso2 <- gapminder %>% distinct(iso2) %>% nrow()
n_iso2
```

```
## [1] 249
```

```
n_countries <- gapminder %>% distinct(name) %>% nrow()
n_countries
```

```
## [1] 250
```

There are 253 iso3 codes in the dataset.

There are 249 iso2 codes in the dataset.

There are 250 countries in the dataset.

2. If you did this correctly, you saw that there are more names than iso-2 codes, and there areeven more iso3 -codes. What is going on? Can you find it out?

(a) (5pt) Find how many names are there for each iso-2 code. Are there any iso-2 codes that correspond to more than one name? What are these countries?

```
group_by(gapminder, iso2) %>%
  summarize(n = length(unique(name))) %>%
  filter(n > 1)
```

```
## # A tibble: 1 x 2
##   iso2      n
##   <chr> <int>
## 1 <NA>      2
```

There are two country names that do not have iso2 codes

(b) (5pt) Now repeat the same for name and iso3-code. Are there country names that havemore than one iso3-code? What are these countries?Hint: two of these entitites are CHANISL and NLD CURACAO.

```
group_by(gapminder, name) %>%
  summarize(n = length(unique(iso3))) %>%
  filter(n > 1)
```

```
## # A tibble: 1 x 2
##   name      n
##   <chr> <int>
## 1 <NA>      4
```

```
gapminder[is.na(gapminder$name),] %>%
  group_by(iso3) %>%
  summarize(n = n())
```

```
## # A tibble: 4 x 2
##   iso3            n
##   <chr>       <int>
## 1 CHANISL        60
## 2 GBM            60
## 3 KOS            60
## 4 NLD_CURACAO    60
```

There are 4 unnamed countries with an iso3 code Entities are CHANISL, GBM, KOS, and NLD_CURACAO

3. (2pt) What is the minimum and maximum year in these data?

```
min_year <- min(gapminder$time, na.rm = TRUE)
max_year <- max(gapminder$time, na.rm = TRUE)
paste("Minimum year:", min_year, "Maximum year:", max_year)
```

```
## [1] "Minimum year: 1960 Maximum year: 2019"
```

3 CO2 emissions (30pt)

1. (2pt) How many missing co2 emissions are there for each year? Analyze both missing CO2and co2_PC. Which years have most missing data?

```
library(dplyr)
missing_co2 <- gapminder %>%
  group_by(time) %>%
  summarise(missing_co2 = sum(is.na(co2)), missing_co2_pc =   sum(is.na(co2_PC))) %>%
  arrange(desc(missing_co2))
missing_co2
```

```
## # A tibble: 61 x 3
##     time missing_co2 missing_co2_pc
##    <dbl>       <int>          <int>
##  1  2017         217            217
##  2  2018         217            217
##  3  2019         217            217
##  4  1960          60             60
##  5  1961          60             60
##  6  1962          58             58
##  7  1963          57             57
##  8  1964          51             51
##  9  1965          51             51
## 10  1966          51             51
## # ... with 51 more rows
```

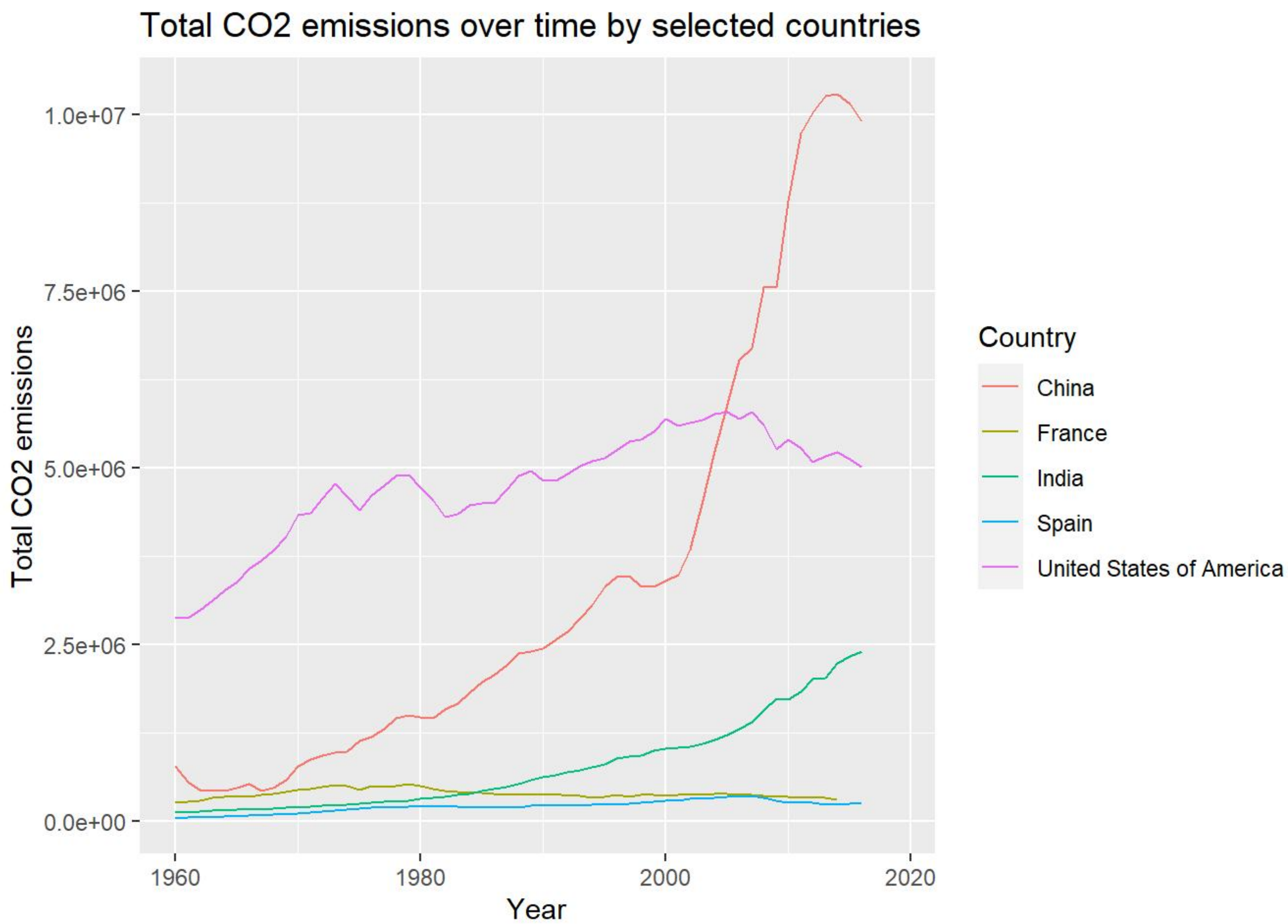
The years 2017, 2018, and 2019 have the most missing CO2 and CO2_pc

2. (5pt) Make a plot of total CO2 emissions over time for the U.S, China, and India. Add a fewmore countries of your choice.

Explain what do you see.

```
library(ggplot2)
gapminder %>%
  filter(name == "United States of America" |name == "China" | name == "India" |name == "Spai
n"|name == "France") %>%
  select(time, co2, name) %>%
  ggplot(aes(x = time, y = co2, color = name)) +
  geom_line() +
  labs(color = "Country", title = "Total CO2 emissions over time by selected countries", x = "Y
ear", y = "Total CO2 emissions")
```

```
## Warning: Removed 17 rows containing missing values (`geom_line()`).
```



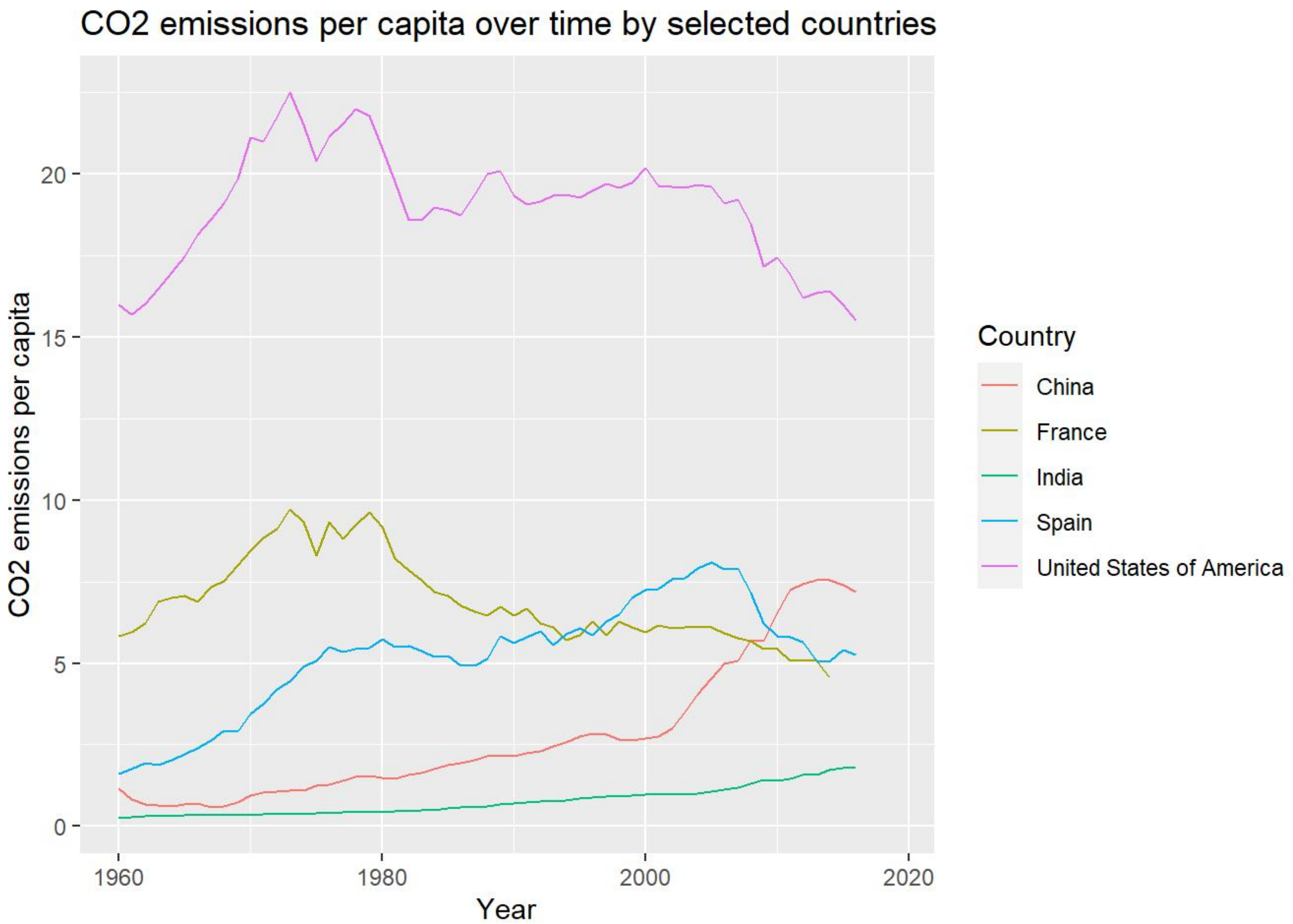
The plot shows that the United States and China continue to be the top two countries with the highest CO2 emissions, with both countries' emissions rising rapidly since the 1990s. India's CO2 emissions have also been increasing steadily over time, though at a slower rate than the United States and China.

3. (5pt) Now let's analyze the CO2 emissions per capita (co2_PC). Make a similar plot of thesame countries. What does

this figure suggest?

```
gapminder %>%
  filter(name == "United States of America" |name == "China" | name == "India" |name == "Spai
n"|name == "France") %>%
  select(time, co2_PC, name) %>%
  ggplot(aes(x = time, y = co2_PC, color = name)) +
  geom_line() +
  labs( color = "Country", title = "CO2 emissions per capita over time by selected countries",
x = "Year", y = "CO2 emissions per capita")
```

```
## Warning: Removed 17 rows containing missing values (`geom_line()`).
```



The plot shows that the United States has consistently had the highest CO2 emissions per capita of the selected countries, with the exception of a brief dip in the 1970s. China and India have had lower CO2 emissions per capita, but both countries have experienced rapid increases in emissions since around 2000.

4.(6pt) Compute average CO2 emissions per capita across the continents (assume region is thesame as continent). Comment what do you see.

Note: just compute averages over countries and ignore the fact

that countries are of different size.

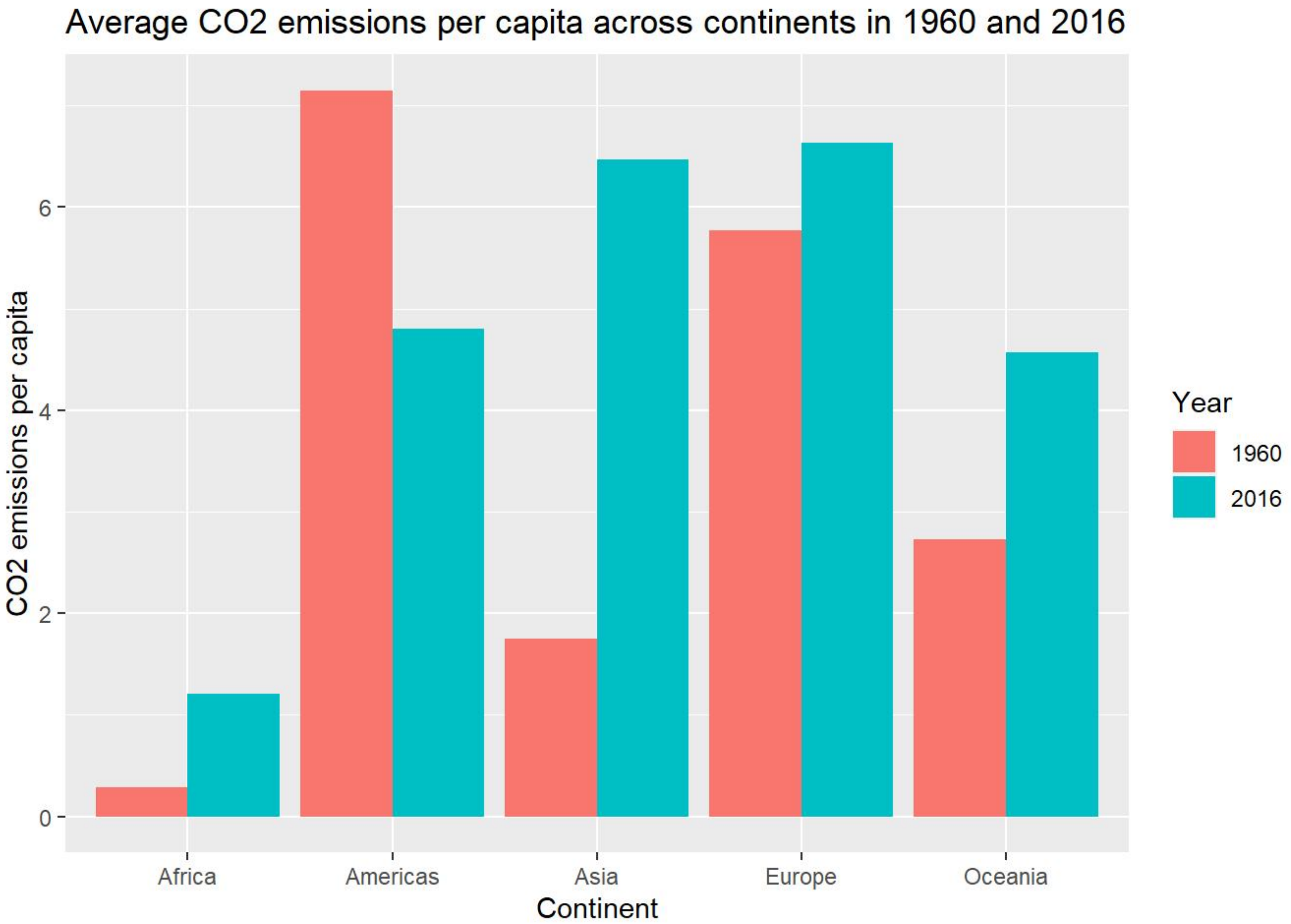
```
gapminder %>%
  filter(time == 1960 | time == 2016, !is.na(co2_PC), !is.na(region)) %>%
  group_by(time, region) %>%
  summarize(average_CO2_PC = mean(co2_PC, na.rm = TRUE), .groups = 'drop')
```

```
## # A tibble: 10 x 3
##     time region   average_CO2_PC
##    <dbl> <chr>             <dbl>
##  1  1960 Africa            0.291
##  2  1960 Americas          7.15
##  3  1960 Asia              1.74
##  4  1960 Europe            5.77
##  5  1960 Oceania           2.73
##  6  2016 Africa            1.20
##  7  2016 Americas          4.80
##  8  2016 Asia              6.47
##  9  2016 Europe            6.64
## 10  2016 Oceania           4.57
```

The results show that there is a large variation in average CO2 emissions per capita across the continents over time. In 1960, the Americas had the highest average CO2 emissions per capita, while Africa had the lowest. By 2016, Asia had the highest average CO2 emissions per capita, while Africa still had the lowest.

5. (7pt) Make a barplot where you show the previous results–average CO2 emissions per capitaacross continents in 1960 and 2016.Hint: it should look something along these lines:

```
gapminder %>%
  filter(time %in% c(1960, 2016), !is.na(region), !is.na(co2_PC)) %>%
  group_by(time, region) %>%
  summarise(avg_co2PC = mean(co2_PC),.groups = 'drop') %>%
  ggplot(aes(x = region, y = avg_co2PC, fill = as.factor(time))) +
  geom_col(position = "dodge") +
  labs(title = "Average CO2 emissions per capita across continents in 1960 and 2016", x = "Cont
inent", y = "CO2 emissions per capita") +
  scale_fill_discrete(name = "Year")
```



6.Which countries are the three largest, and three smallest CO2 emitters (in terms of CO2 percapita) in 2019 for each continent? (Assume region is continent).

```
gapminder %>%
  filter(time == 2016, !is.na(co2_PC), !is.na(region)) %>%
  filter(name != "") %>%
  group_by(region, name) %>%
  summarize(co2_data = co2_PC, .groups = 'drop') %>%
  arrange(region, desc(co2_data)) %>%
  mutate(ranking = row_number()) %>%
  filter(ranking <= 3 | ranking >= n() - 2) %>%
  ungroup() %>%
  arrange(region, ranking)
```

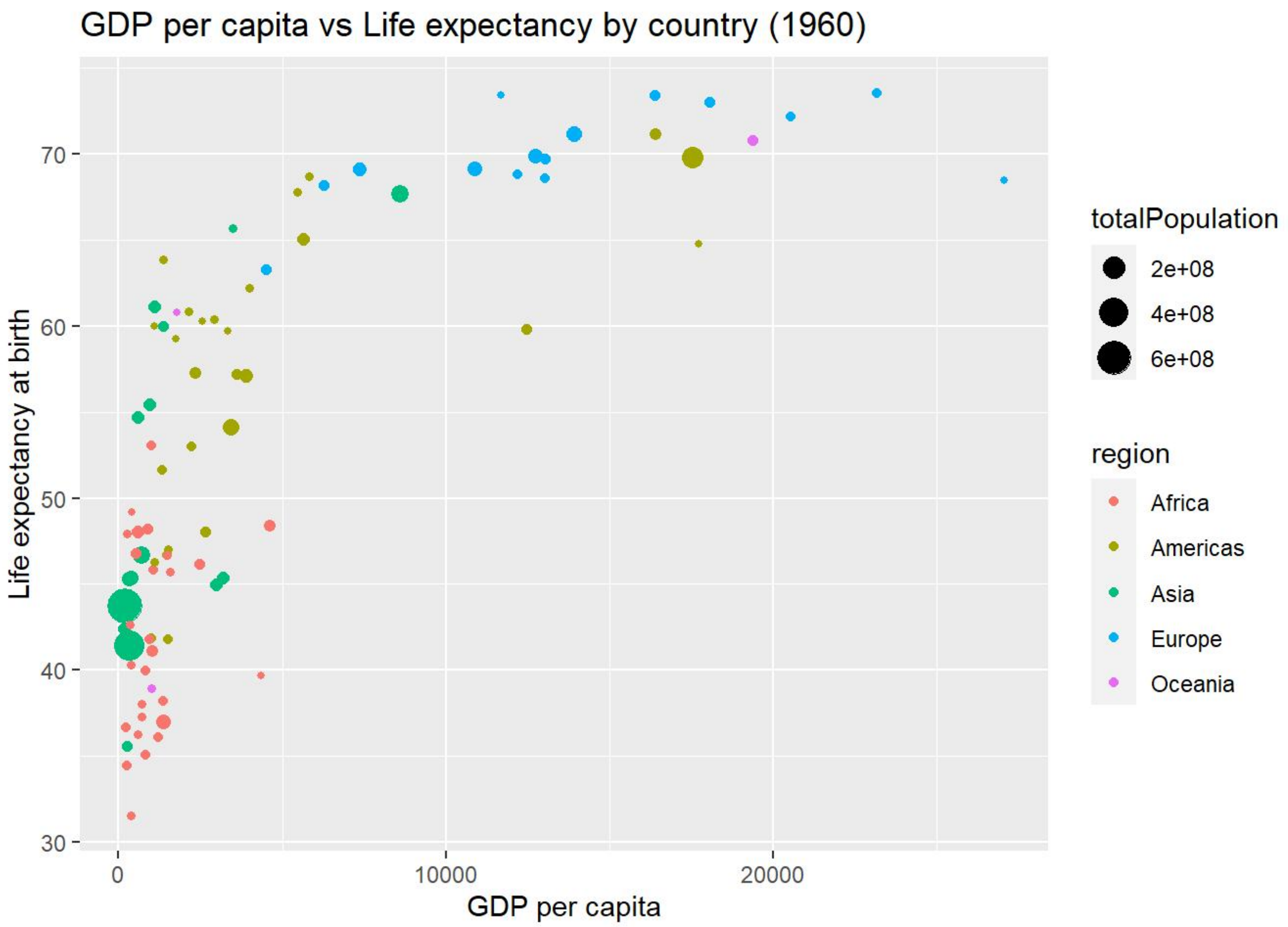
```
## # A tibble: 6 x 4
##   region  name            co2_data ranking
##   <chr>   <chr>              <dbl>   <int>
## 1 Africa  South Africa       8.48        1
## 2 Africa  Libya              7.79        2
## 3 Africa  Seychelles         6.39        3
## 4 Oceania Kiribati           0.587     200
## 5 Oceania Vanuatu            0.527     201
## 6 Oceania Solomon Islands    0.272     202
```

4 GDP per capita (50pt)

Let’s look at GDP per capita (GDP_PC).

1. (8pt) Make a scatterplot of GDP per capita versus life expectancy by country, using data for1960. Make the point size dependent on the country size, and color those according to thecontinent. Feel free to adjust the plot in other ways to make it better.Comment what do you see there.

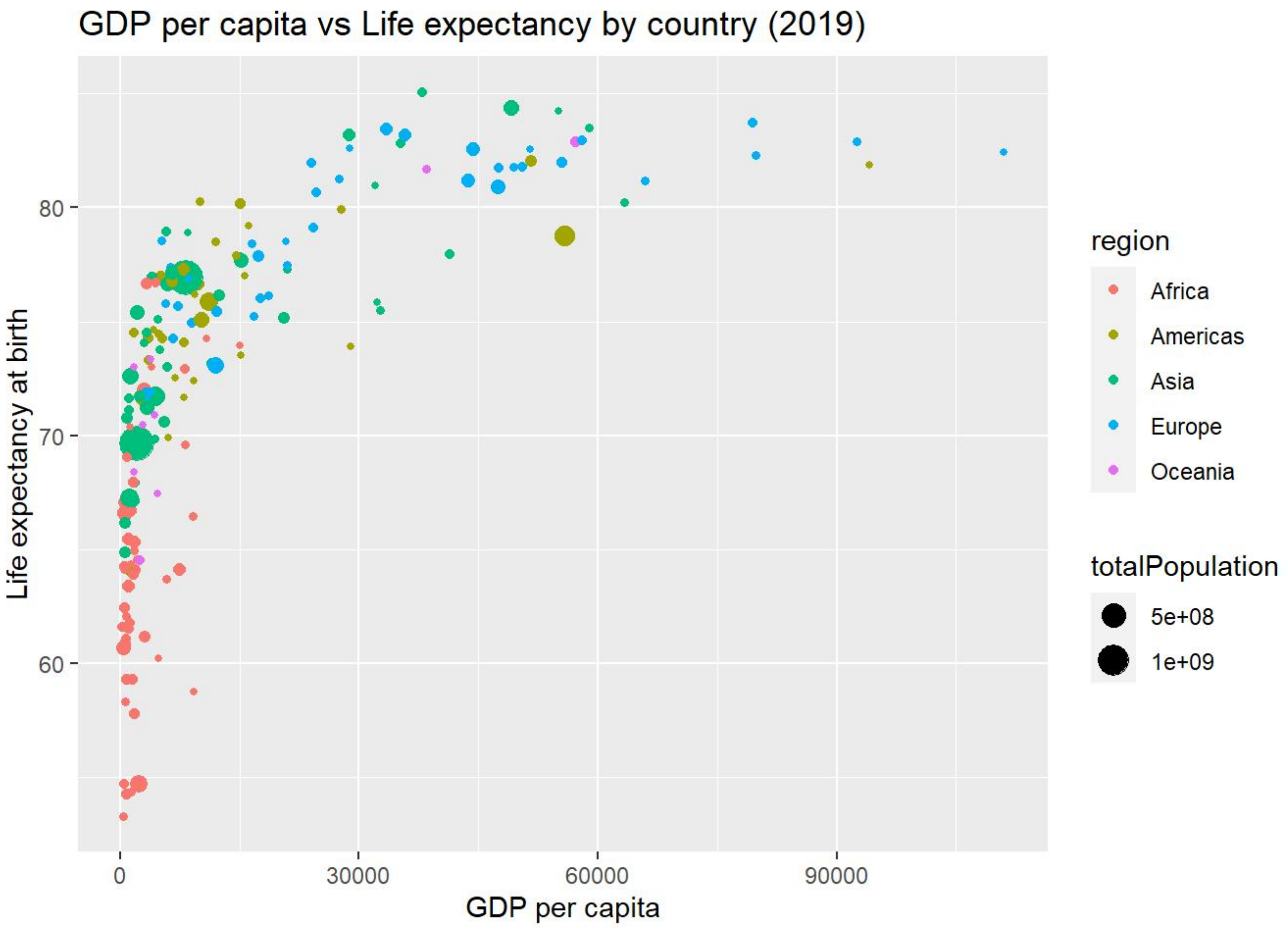
```
gapminder %>%
  filter(time == 1960 & !is.na(region),
         !is.na(GDP_PC),
         !is.na(lifeExpectancy)) %>%
  ggplot(aes(GDP_PC, lifeExpectancy, size = totalPopulation, col = region)) +
    geom_point() +
  labs(title = "GDP per capita vs Life expectancy by country (1960)", x = "GDP per capita", y =
"Life expectancy at birth")
```



The plot shows that there is a positive relationship between GDP per capita and life expectancy, with countries with higher GDP per capita generally having higher life expectancies. However, there is a wide variation in life expectancy within each level of GDP per capita, suggesting that other factors besides economic development also play a role in determining life expectancy.

2. (4pt) Make a similar plot, but this time use 2019 data only.

```
gapminder %>%
  filter(time == 2019 & !is.na(region),
         !is.na(GDP_PC),
         !is.na(lifeExpectancy)) %>%
  ggplot(aes(GDP_PC, lifeExpectancy, size = totalPopulation, col = region)) +
    geom_point()+
  labs(title = "GDP per capita vs Life expectancy by country (2019)", x = "GDP per capita", y =
"Life expectancy at birth")
```



3. (6pt) Compare these two plots and comment what do you see. How has world developed through the last 60 years?

The overall relationship between GDP per capita and life expectancy remains strong for both years, with countries with higher GDP per capita tending to have higher life expectancy. However, the relationship appears to be stronger in 2019 than in 1960, with the points in the upper right quadrant of the graph clustered more closely, and the range of GDP per capita values and life expectancy increasing substantially over the past 60 years. In 1960, most countries had per capita GDP values below $10,000, and life expectancy in many countries was below 50 years. By 2019, many countries had per capita GDP values exceeding $10,000 and minimum life expectancy values exceeding 50 years. These last two graphs show that the world has experienced significant economic and social development over the past 60 years, with many countries experiencing significant improvements in GDP per capita and life expectancy.

4. (6pt) Compute the average life expectancy for each continent in 1960 and 2019. Do the resultsfit with what do you see on the figures?Note: here as average I mean just average over

countries, ignore the fact that countries are ofdifferent size.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v stringr 1.5.0
## v tidyr   1.3.0     v forcats 1.0.0
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
lifeexp_continents <- gapminder %>%
  filter(time %in% c(1960, 2019), !is.na(region)) %>%
  group_by(region, time) %>%
  summarize(avg_life_expectancy = mean(lifeExpectancy, na.rm = TRUE), .groups = 'drop')

lifeexp_continents_wide <- spread(lifeexp_continents, time, avg_life_expectancy)

knitr::kable(lifeexp_continents_wide, caption = "Average life expectancy by continent and yea
r")
```

Average life expectancy by continent and year

region	1960	2019
Africa	41.46600	64.11014
Americas	58.64651	75.83206
Asia	51.64931	74.61739
Europe	68.28254	79.35714
Oceania	56.39613	73.52827

The results are consistent with what we saw in the plots.

5. (8pt) Compute the average LE growth from 1960-2019 across the continents. Show the results in the order of growth. Explain what do you see. Hint: these data (data in long form) is not the simplest to compute growth. But you may want to check out the lag() function. And do not forget to group data by continent when using lag(), otherwise your results will be messed up! See https://faculty.washington.edu/otoomet/info201-

book/dplyr.html#dplyr-helpers-compute (https://faculty.washington.edu/otoomet/info201-book/dplyr.html#dplyr-helpers-compute).

```
lifeexp_growth <- gapminder %>%
  filter(time %in% c(1960, 2019),!is.na(region)) %>%
  group_by(region) %>%
  summarize(avg_life_expectancy_1960 = mean(lifeExpectancy[time == 1960], na.rm = TRUE),
            avg_life_expectancy_2019 = mean(lifeExpectancy[time == 2019], na.rm = TRUE),
            life_expectancy_growth = avg_life_expectancy_2019 - avg_life_expectancy_1960,
            growth_rate = ((avg_life_expectancy_2019 / avg_life_expectancy_1960) ^ (1 / 59)) -
1) %>%
  arrange(life_expectancy_growth)

knitr::kable(lifeexp_growth, caption = "Average LE growth from 1960 to 2019 by continent")
```

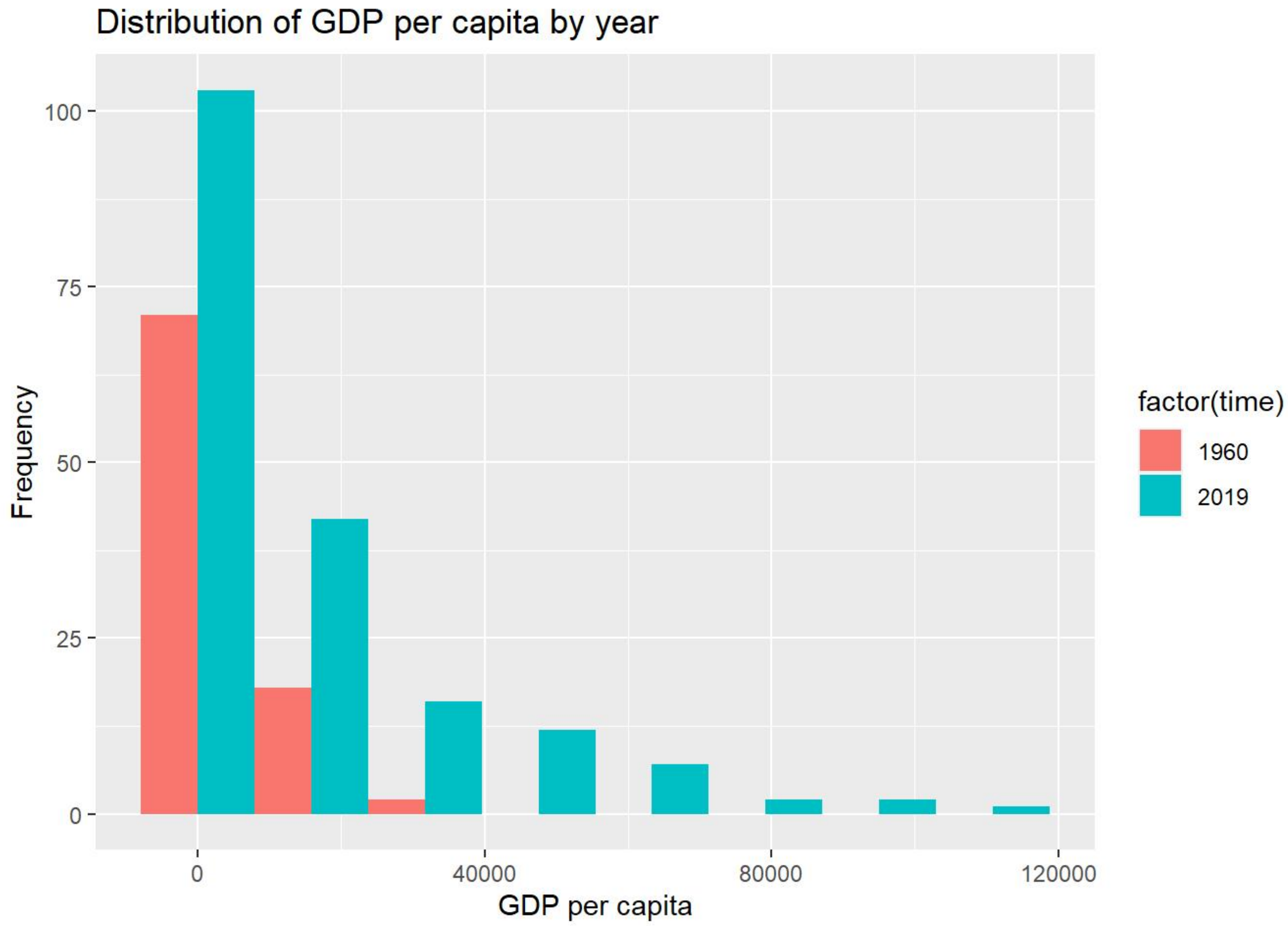
Average LE growth from 1960 to 2019 by continent

region	avg_life_expectancy_1960	avg_life_expectancy_2019	life_expectancy_growth	growth_rate
Europe	68.28254	79.35714	11.07460	0.0025508
Oceania	56.39613	73.52827	17.13214	0.0045062
Americas	58.64651	75.83206	17.18555	0.0043653
Africa	41.46600	64.11014	22.64414	0.0074126
Asia	51.64931	74.61739	22.96808	0.0062550

The average life expectancy has increased for all continents from 1960 to 2019. The average growth rate for life expectancy is highest for Africa, followed by Oceania, the Americas, Asia, and Europe.

6. (6pt) Show the histogram of GDP per capita for years of 1960 and 2019. Try to put both histograms on the same graph, see how well you can do it!

```
gapminder %>%
  filter(time == 1960 | time == 2019,
         !is.na(GDP_PC)) %>%
  ggplot(aes(GDP_PC, fill = factor(time))) +
  geom_histogram(position = "dodge", bins = 8) +
  labs(title = "Distribution of GDP per capita by year", x = "GDP per capita", y = "Frequency")
```



7. (6pt) What was the ranking of US in terms of life expectancy in 1960 and in 2019? (When counting from top.) Hint: check out the function rank()! Hint2: 17 for 1960.

```
gapminder %>%
  filter(time == 1960 | time == 2019,
         !is.na(name)) %>%
  group_by(time) %>%
  mutate(rank = rank(desc(lifeExpectancy))) %>%
  filter(iso3 == "USA") %>%
  select(name, time, lifeExpectancy, rank) %>%
  print()
```

```
## # A tibble: 2 x 4
## # Groups:   time [2]
##   name                      time lifeExpectancy  rank
##   <chr>                    <dbl>          <dbl> <dbl>
## 1 United States of America  1960           69.8    17
## 2 United States of America  2019           78.8    46
```

8. (6pt) If you did this correctly, then you noticed that US ranking has been falling quite a bit. But we also have more countries in 2019–what about the relative rank divided by the corresponding number of countries that have LE data in the corresponding

year? Hint: 0.0904 for 1960.

```
gapminder %>%
  filter(time == 1960 | time == 2019,
         !is.na(name) & !is.na(lifeExpectancy)) %>%
  group_by(time) %>%
  mutate(ranking = rank(desc(lifeExpectancy))) %>%
  mutate(perc = ranking / n()) %>%
  filter(iso3 == "USA") %>%
  select(name, time, ranking, perc)
```

```
## # A tibble: 2 x 4
## # Groups:   time [2]
##   name                      time ranking   perc
##   <chr>                    <dbl>   <dbl>  <dbl>
## 1 United States of America  1960      17 0.0904
## 2 United States of America  2019      46 0.235
```

Finally tell us how many hours did you spend on this PS.

10hours