Lecture 6

Vincent

2022-10-11

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(gapminder)
library(rlang)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:rlang':
##
##     set_names
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.6     v purrr   0.3.5
## v tidyr   1.2.1     v stringr 1.4.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x purrr::%@%()         masks rlang::%@%()
## x purrr::as_function() masks rlang::as_function()
## x tidyr::extract()     masks magrittr::extract()
## x dplyr::filter()      masks stats::filter()
## x purrr::flatten()     masks rlang::flatten()
```

```
## x purrr::flatten_chr() masks rlang::flatten_chr()
## x purrr::flatten_dbl() masks rlang::flatten_dbl()
## x purrr::flatten_int() masks rlang::flatten_int()
## x purrr::flatten_lgl() masks rlang::flatten_lgl()
## x purrr::flatten_raw() masks rlang::flatten_raw()
## x purrr::invoke()      masks rlang::invoke()
## x dplyr::lag()         masks stats::lag()
## x purrr::set_names()   masks magrittr::set_names(), rlang::set_names()
## x purrr::splice()      masks rlang::splice()
```

R Markdown

```
data(gapminder)
print(gapminder, n=20)
```

```
## # A tibble: 1,704 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.
##  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.
## 11 Afghanistan Asia       2002    42.1 25268405      727.
## 12 Afghanistan Asia       2007    43.8 31889923      975.
## 13 Albania     Europe     1952    55.2  1282697     1601.
## 14 Albania     Europe     1957    59.3  1476505     1942.
## 15 Albania     Europe     1962    64.8  1728137     2313.
## 16 Albania     Europe     1967    66.2  1984060     2760.
## 17 Albania     Europe     1972    67.7  2263554     3313.
## 18 Albania     Europe     1977    68.9  2509048     3533.
## 19 Albania     Europe     1982    70.4  2780097     3631.
## 20 Albania     Europe     1987    72    3075321     3739.
## # ... with 1,684 more rows
```

```
gapminder %>%
  summarise(
    avg_exp = mean(lifeExp, na.rm = TRUE),
    # find mean
    sd_exp = sd(lifeExp, na.rm=TRUE),
    # standard deviation
    min_exp = min(lifeExp, na.rm = TRUE),
    # find min
    med_exp = median(lifeExp, na.rm = TRUE),
    # find median
    q3_exp = quantile(lifeExp, prob = 0.75, na.rm = TRUE),
```

```
    # find 0.75 quantile
    max_exp = max(lifeExp, na.rm = TRUE),
    # find max
    count = n()
    # samplesize

  )
```

```
## # A tibble: 1 x 7
##   avg_exp sd_exp min_exp med_exp q3_exp max_exp count
##     <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl> <int>
## 1    59.5   12.9    23.6    60.7   70.8    82.6  1704
```

```
# na.rm tells R to ignore the missing value

gapminder %>%
  filter(year == 2002, continent == "Americas") %>%
  summarise(
    avg_exp = mean(lifeExp, na.rm = TRUE),
    sd_exp = sd(lifeExp, na.rm=TRUE),
    min_exp = min(lifeExp, na.rm = TRUE),
    med_exp = median(lifeExp, na.rm = TRUE),
    q3_exp = quantile(lifeExp, prob = 0.75, na.rm = TRUE),
    q3_exp = quantile(lifeExp, prob = 0.25, na.rm = TRUE),
    max_exp = max(lifeExp, na.rm = TRUE),
    count = n()
    # samplesize
  )
```

```
## # A tibble: 1 x 7
##   avg_exp sd_exp min_exp med_exp q3_exp max_exp count
##     <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl> <int>
## 1    72.4   4.80    58.1    72.0   70.7    79.8    25
```

```
gapminder %>%
  filter(continent == "Americas") %>%
  group_by(year) %>%
  summarise(
    avg_exp = mean(lifeExp, na.rm = TRUE),
    sd_exp = sd(lifeExp, na.rm=TRUE),
    min_exp = min(lifeExp, na.rm = TRUE),
    med_exp = median(lifeExp, na.rm = TRUE),
    q1_exp = quantile(lifeExp, prob = 0.75, na.rm = TRUE),
    q3_exp = quantile(lifeExp, prob = 0.25, na.rm = TRUE),
    max_exp = max(lifeExp, na.rm = TRUE),
    count = n()
    # samplesize
  ) %>%
arrange(desc(year))
```

```
## # A tibble: 12 x 9
##     year avg_exp sd_exp min_exp med_exp q1_exp q3_exp max_exp count
```

```
##    <int>   <dbl>  <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl> <int>
##  1  2007    73.6   4.44    60.9    72.9   76.4   71.8    80.7    25
##  2  2002    72.4   4.80    58.1    72.0   75.3   70.7    79.8    25
##  3  1997    71.2   4.89    56.7    72.1   74.2   69.4    78.6    25
##  4  1992    69.6   5.17    55.1    69.9   72.8   66.8    78.0    25
##  5  1987    68.1   5.80    53.6    69.5   71.9   64.5    76.9    25
##  6  1982    66.2   6.72    51.5    67.4   70.8   61.4    75.8    25
##  7  1977    64.4   7.07    49.9    66.4   69.5   58.4    74.2    25
##  8  1972    62.4   7.32    46.7    63.4   67.8   58.2    72.9    25
##  9  1967    60.4   7.91    45.0    60.5   65.6   55.9    72.1    25
## 10  1962    58.4   8.50    43.4    58.3   65.1   52.3    71.3    25
## 11  1957    56.0   9.03    40.7    56.1   62.6   48.6    70.0    25
## 12  1952    53.3   9.33    37.6    54.7   59.4   45.3    68.8    25
```

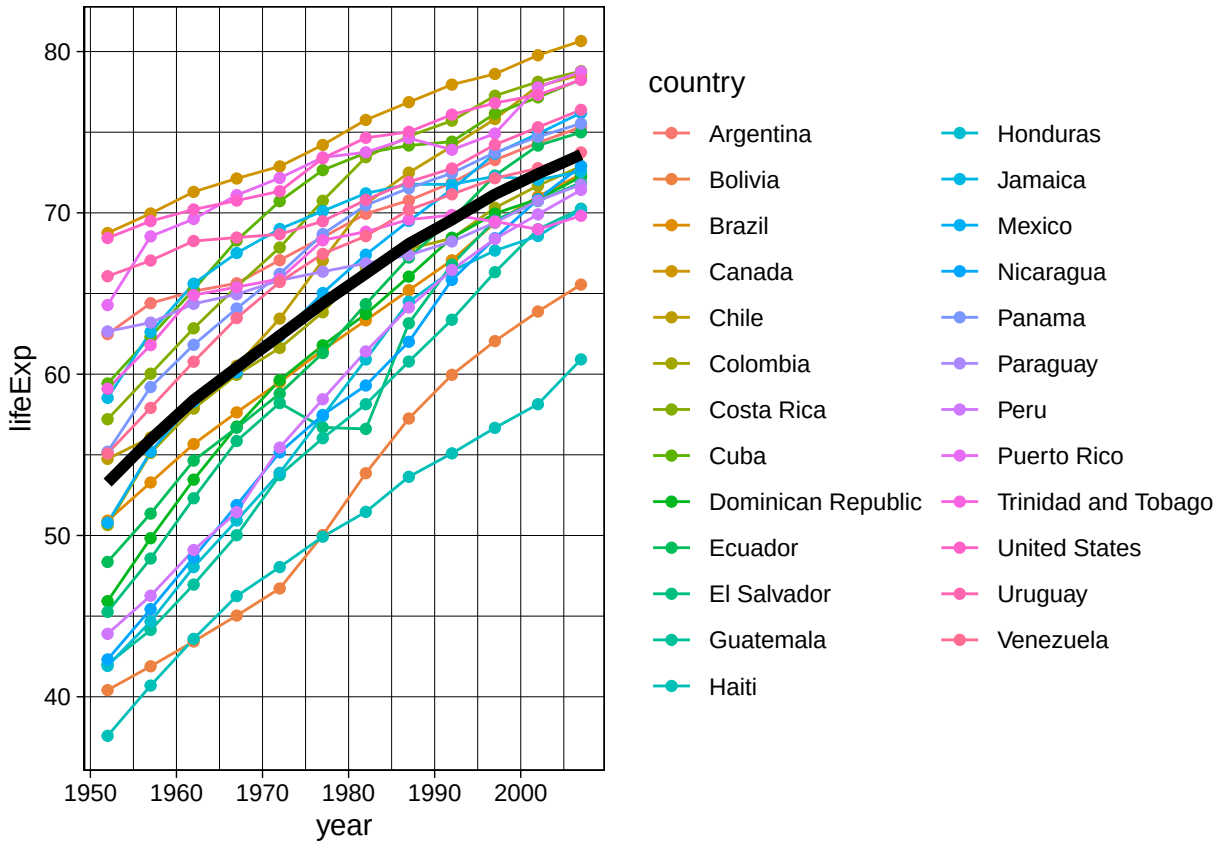
cda

```
americas_summary <- gapminder %>%
  filter(continent == "Americas") %>%
  group_by(year) %>%
  summarise(
    avg_exp = mean(lifeExp, na.rm = TRUE),
    count = n()
  )
```

```
# select() vs filter()

# select() pick out a single appreance in the data set
# filter() pick out all of the data in the column

gapminder  %>%
  filter(continent == "Americas") %>%
  ggplot(aes(x=year,
             y=lifeExp,
             color = country))+
  geom_line()+geom_point()+theme_linedraw()+
  geom_line(data = americas_summary, mapping = aes(x=year , y=avg_exp),
            color = "black", lwd = 2)
```




```
gapminder %>% filter(year == 2007, continent == "Americas") %>%
  select(country, lifeExp) %>% arrange(desc(lifeExp)) %>% head()
```

```
## # A tibble: 6 x 2
##   country       lifeExp
##   <fct>           <dbl>
## 1 Canada           80.7
## 2 Costa Rica       78.8
## 3 Puerto Rico      78.7
## 4 Chile            78.6
## 5 Cuba             78.3
## 6 United States    78.2
```

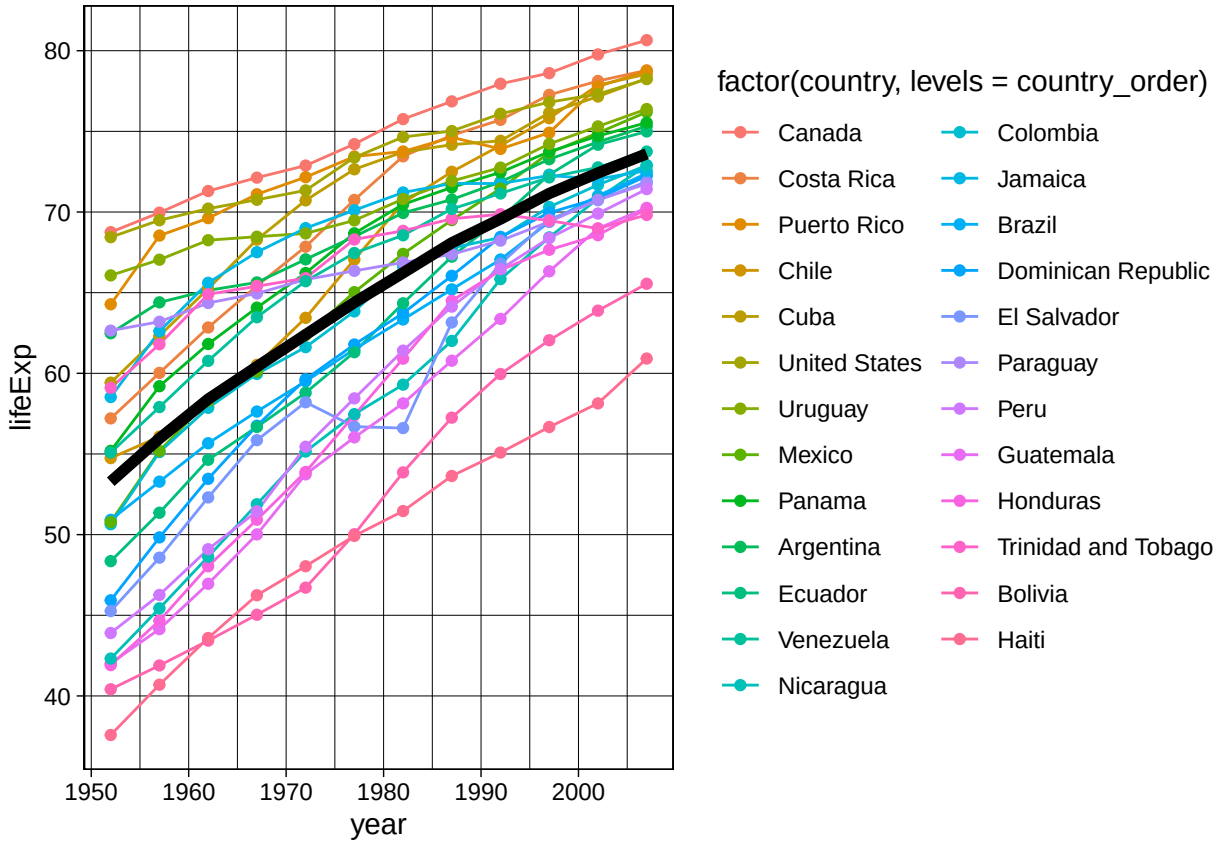
```
country_order <- gapminder %>%
  filter(year == 2007, continent == "Americas") %>%
  select(country,lifeExp) %>% arrange(desc(lifeExp)) %>% pull(country)
country_order
```

```
##  [1] Canada              Costa Rica          Puerto Rico
##  [4] Chile               Cuba                United States
##  [7] Uruguay             Mexico              Panama
## [10] Argentina           Ecuador             Venezuela
## [13] Nicaragua           Colombia            Jamaica
## [16] Brazil              Dominican Republic  El Salvador
## [19] Paraguay            Peru                Guatemala
```

```
## [22] Honduras            Trinidad and Tobago Bolivia
## [25] Haiti
## 142 Levels: Afghanistan Albania Algeria Angola Argentina Australia ... Zimbabwe
```

```
# place all of the country in order based on their lifeExp
## NOTICE, this is not base on the int but the name of the country


gapminder %>%
  filter(continent == "Americas")%>%
  ggplot(aes(x=year,
             y=lifeExp,
             color = factor(country, levels = country_order)))+
  geom_line()+geom_point()+theme_linedraw()+  geom_line(data = americas_summary, mapping = aes(x=year ,
            color = "black", lwd = 2)
```




```
# "levels" will determine the order saved in country_order and arrange the color base on the order.
```

```
gapminder %>%
  filter(year==2007) %>%
  summarise(
    total_pop = sum(pop, na.rm =TRUE),
    count = n()
  )
```

```
## # A tibble: 1 x 2
```

```
##    total_pop count
##        <dbl> <int>
## 1 6251013179   142
```