gapminder-wrangle

guoRandall

2025-06-27

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
gapminder <- read_csv("data/gapminder.csv")
```

```
## Rows: 1704 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): country, continent
## dbl (4): year, pop, lifeExp, gdpPercap
##
## 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.
```

```
View(gapminder)
head(gapminder) # shows first 6
```

```
## # A tibble: 6 x 6
##   country      year      pop continent lifeExp gdpPercap
##   <chr>       <dbl>    <dbl> <chr>       <dbl>     <dbl>
## 1 Afghanistan  1952  8425333 Asia         28.8      779.
## 2 Afghanistan  1957  9240934 Asia         30.3      821.
## 3 Afghanistan  1962 10267083 Asia         32.0      853.
## 4 Afghanistan  1967 11537966 Asia         34.0      836.
## 5 Afghanistan  1972 13079460 Asia         36.1      740.
## 6 Afghanistan  1977 14880372 Asia         38.4      786.
```

```
tail(gapminder) # shows last 6
```

```
## # A tibble: 6 x 6
##   country   year      pop continent lifeExp gdpPercap
##   <chr>    <dbl>    <dbl> <chr>       <dbl>     <dbl>
## 1 Zimbabwe  1982  7636524 Africa       60.4      789.
## 2 Zimbabwe  1987  9216418 Africa       62.4      706.
## 3 Zimbabwe  1992 10704340 Africa       60.4      693.
```

```
## 4 Zimbabwe  1997 11404948 Africa       46.8      792.
## 5 Zimbabwe  2002 11926563 Africa       40.0      672.
## 6 Zimbabwe  2007 12311143 Africa       43.5      470.
```

```
str(gapminder)
```

```
## spc_tbl_ [1,704 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ country  : chr [1:1704] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ year     : num [1:1704] 1952 1957 1962 1967 1972 ...
##  $ pop      : num [1:1704] 8425333 9240934 10267083 11537966 13079460 ...
##  $ continent: chr [1:1704] "Asia" "Asia" "Asia" "Asia" ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   country = col_character(),
##   ..   year = col_double(),
##   ..   pop = col_double(),
##   ..   continent = col_character(),
##   ..   lifeExp = col_double(),
##   ..   gdpPercap = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
filter(gapminder, lifeExp < 29)
```

```
## # A tibble: 2 x 6
##   country      year     pop continent lifeExp gdpPercap
##   <chr>       <dbl>   <dbl> <chr>       <dbl>     <dbl>
## 1 Afghanistan  1952 8425333 Asia         28.8      779.
## 2 Rwanda       1992 7290203 Africa       23.6      737.
```

```
filter(gapminder, country == "Mexico")
```

```
## # A tibble: 12 x 6
##    country  year      pop continent lifeExp gdpPercap
##    <chr>   <dbl>    <dbl> <chr>       <dbl>     <dbl>
##  1 Mexico   1952  30144317 Americas     50.8     3478.
##  2 Mexico   1957  35015548 Americas     55.2     4132.
##  3 Mexico   1962  41121485 Americas     58.3     4582.
##  4 Mexico   1967  47995559 Americas     60.1     5755.
##  5 Mexico   1972  55984294 Americas     62.4     6809.
##  6 Mexico   1977  63759976 Americas     65.0     7675.
##  7 Mexico   1982  71640904 Americas     67.4     9611.
##  8 Mexico   1987  80122492 Americas     69.5     8688.
##  9 Mexico   1992  88111030 Americas     71.5     9472.
## 10 Mexico   1997  95895146 Americas     73.7     9767.
## 11 Mexico   2002 102479927 Americas     74.9    10742.
## 12 Mexico   2007 108700891 Americas     76.2    11978.
```

```
filter(gapminder, country %in% c("Mexico", "Peru"))
```

```
## # A tibble: 24 x 6
##    country  year      pop continent lifeExp gdpPercap
##    <chr>   <dbl>    <dbl> <chr>       <dbl>     <dbl>
##  1 Mexico   1952 30144317 Americas     50.8     3478.
##  2 Mexico   1957 35015548 Americas     55.2     4132.
```

```
##  3 Mexico   1962 41121485 Americas     58.3     4582.
##  4 Mexico   1967 47995559 Americas     60.1     5755.
##  5 Mexico   1972 55984294 Americas     62.4     6809.
##  6 Mexico   1977 63759976 Americas     65.0     7675.
##  7 Mexico   1982 71640904 Americas     67.4     9611.
##  8 Mexico   1987 80122492 Americas     69.5     8688.
##  9 Mexico   1992 88111030 Americas     71.5     9472.
## 10 Mexico   1997 95895146 Americas     73.7     9767.
## # i 14 more rows
```

```
filter(gapminder, country == "Mexico", year == 2002)
```

```
## # A tibble: 1 x 6
##   country  year       pop continent lifeExp gdpPercap
##   <chr>   <dbl>     <dbl> <chr>       <dbl>     <dbl>
## 1 Mexico   2002 102479927 Americas     74.9    10742.
```

```
gap1 <- dplyr::select(gapminder, year, country, lifeExp)# choose column
gap2 <- dplyr::select(gapminder, year:lifeExp)
gap3 <- dplyr::select(gapminder, 1, 2, 4)# We can select columns with indices
gap4 <- dplyr::select(gapminder, -continent, -lifeExp)# don't want some column
```

```
gap_cambodia <- filter(gapminder, country == "Cambodia")
gap_cambodia2 <- dplyr::select(gap_cambodia, -continent, -lifeExp)# easy to make mistake
#need new method
```

```
gapminder |> head(3) #|> #cmd+shift+M #=head(gapminder, 3).
```

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

```
#"and then":take the gapminder data, and then give me the first three entries
```

```
## instead of this...
gap_cambodia <- filter(gapminder, country == "Cambodia")
gap_cambodia2 <- dplyr::select(gap_cambodia, -continent, -lifeExp)
## ...we can do this
gap_cambodia <- gapminder |> filter(country == "Cambodia")
gap_cambodia2 <- gap_cambodia |> dplyr::select(-continent, -lifeExp)
## We can use the pipe to chain those two operations together:
gap_cambodia <- gapminder |>
filter(country == "Cambodia") |>
dplyr::select(-continent, -lifeExp)
```

```
gapminder |>
  mutate(gdp = pop * gdpPercap)#create a new column named gdp.
```

```
## # A tibble: 1,704 x 7
##    country      year      pop continent lifeExp gdpPercap          gdp
##    <chr>       <dbl>    <dbl> <chr>       <dbl>     <dbl>        <dbl>
##  1 Afghanistan  1952  8425333 Asia         28.8      779.  6567086330.
##  2 Afghanistan  1957  9240934 Asia         30.3      821.  7585448670.
##  3 Afghanistan  1962 10267083 Asia         32.0      853.  8758855797.
```

```
##  4 Afghanistan  1967 11537966 Asia         34.0      836.  9648014150.
##  5 Afghanistan  1972 13079460 Asia         36.1      740.  9678553274.
##  6 Afghanistan  1977 14880372 Asia         38.4      786. 11697659231.
##  7 Afghanistan  1982 12881816 Asia         39.9      978. 12598563401.
##  8 Afghanistan  1987 13867957 Asia         40.8      852. 11820990309.
##  9 Afghanistan  1992 16317921 Asia         41.7      649. 10595901589.
## 10 Afghanistan  1997 22227415 Asia         41.8      635. 14121995875.
## # i 1,694 more rows
```

```
gapminder |>
  filter(year == 2002) |>
  group_by(continent) |>
  mutate(cont_pop = sum(pop))
```

```
## # A tibble: 142 x 7
## # Groups:   continent [5]
##    country      year       pop continent lifeExp gdpPercap   cont_pop
##    <chr>       <dbl>     <dbl> <chr>       <dbl>     <dbl>      <dbl>
##  1 Afghanistan  2002  25268405 Asia         42.1      727. 3601802203
##  2 Albania      2002   3508512 Europe       75.7     4604.  578223869
##  3 Algeria      2002  31287142 Africa       71.0     5288.  833723916
##  4 Angola       2002  10866106 Africa       41.0     2773.  833723916
##  5 Argentina    2002  38331121 Americas     74.3     8798.  849772762
##  6 Australia    2002  19546792 Oceania      80.4    30688.   23454829
##  7 Austria      2002   8148312 Europe       79.0    32418.  578223869
##  8 Bahrain      2002    656397 Asia         74.8    23404. 3601802203
##  9 Bangladesh   2002 135656790 Asia         62.0     1136. 3601802203
## 10 Belgium      2002  10311970 Europe       78.3    30486.  578223869
## # i 132 more rows
```

```
gapminder |>
  group_by(continent) |>
  summarize(cont_pop = sum(pop)) |>
  ungroup() # summarize() will actually only keep the columns that are grouped_by or summarized. So if
```

```
## # A tibble: 5 x 2
##   continent    cont_pop
##   <chr>           <dbl>
## 1 Africa     6187585961
## 2 Americas   7351438499
## 3 Asia      30507333902
## 4 Europe     6181115304
## 5 Oceania     212992136
```

```
gapminder |>
  group_by(continent, year) |>
  summarize(cont_pop = sum(pop))
```

```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 60 x 3
## # Groups:   continent [5]
##    continent  year  cont_pop
##    <chr>     <dbl>     <dbl>
##  1 Africa     1952 237640501
##  2 Africa     1957 264837738
```

```
##  3 Africa     1962 296516865
##  4 Africa     1967 335289489
##  5 Africa     1972 379879541
##  6 Africa     1977 433061021
##  7 Africa     1982 499348587
##  8 Africa     1987 574834110
##  9 Africa     1992 659081517
## 10 Africa     1997 743832984
## # i 50 more rows
```

```
gapminder |>
  group_by(continent, year) |>
  summarize(cont_pop = sum(pop)) |>
  arrange(year)
```

```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 60 x 3
## # Groups:   continent [5]
##    continent  year    cont_pop
##    <chr>     <dbl>       <dbl>
##  1 Africa     1952  237640501
##  2 Americas   1952  345152446
##  3 Asia       1952 1395357352.
##  4 Europe     1952  418120846
##  5 Oceania    1952   10686006
##  6 Africa     1957  264837738
##  7 Americas   1957  386953916
##  8 Asia       1957 1562780599
##  9 Europe     1957  437890351
## 10 Oceania    1957   11941976
## # i 50 more rows
```

How to make data tidy?

```
## wide format
gap_wide <- readr::read_csv('data/gapminder_wide.csv')
```

```
## Rows: 142 Columns: 38
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): continent, country
## dbl (36): gdpPercap_1952, gdpPercap_1957, gdpPercap_1962, gdpPercap_1967, gd...
##
## 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.
```

```
gapminder <- readr::read_csv('data/gapminder.csv')
```

```
## Rows: 1704 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): country, continent
## dbl (4): year, pop, lifeExp, gdpPercap
##
```

```
## 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.
```

```
head(gap_wide)
```

```
## # A tibble: 6 x 38
##   continent country  gdpPercap_1952 gdpPercap_1957 gdpPercap_1962 gdpPercap_1967
##   <chr>     <chr>             <dbl>          <dbl>          <dbl>          <dbl>
## 1 Africa    Algeria           2449.          3014.          2551.          3247.
## 2 Africa    Angola            3521.          3828.          4269.          5523.
## 3 Africa    Benin             1063.           960.           949.          1036.
## 4 Africa    Botswana           851.           918.           984.          1215.
## 5 Africa    Burkina~           543.           617.           723.           795.
## 6 Africa    Burundi            339.           380.           355.           413.
## # i 32 more variables: gdpPercap_1972 <dbl>, gdpPercap_1977 <dbl>,
## #   gdpPercap_1982 <dbl>, gdpPercap_1987 <dbl>, gdpPercap_1992 <dbl>,
## #   gdpPercap_1997 <dbl>, gdpPercap_2002 <dbl>, gdpPercap_2007 <dbl>,
## #   lifeExp_1952 <dbl>, lifeExp_1957 <dbl>, lifeExp_1962 <dbl>,
## #   lifeExp_1967 <dbl>, lifeExp_1972 <dbl>, lifeExp_1977 <dbl>,
## #   lifeExp_1982 <dbl>, lifeExp_1987 <dbl>, lifeExp_1992 <dbl>,
## #   lifeExp_1997 <dbl>, lifeExp_2002 <dbl>, lifeExp_2007 <dbl>, ...
```

```
#str(gap_wide)
```