

EDA

Yanyao Gu

Read in Data

```
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    3.5.1      v tibble     3.2.1
v lubridate  1.9.3      v tidyr      1.3.1
v purrr      1.0.2

-- 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
```

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

```
Rows: 3720 Columns: 21
```

```
-- Column specification -----
```

```
Delimiter: ","
```

```
chr  (3): country_name, ISO, region
```

```
dbl (18): year, gdp1000, OECD, OECD2023, popdens, urban, agedep, male_edu, t...
```

```
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.
```

Quick Look at Data

First look at the top lines.

```
data %>% head()
```

```
# A tibble: 6 x 21
  country_name ISO   region   year gdp1000  OECD OECD2023 popdens urban agedep
  <chr>      <chr> <chr>   <dbl>  <dbl> <dbl>   <dbl>   <dbl> <dbl>
1 Afghanistan AFG   Southern~ 2000    NA      0      0    14.1  16.3  108.
2 Afghanistan AFG   Southern~ 2001    NA      0      0    14.2  16.3  109.
3 Afghanistan AFG   Southern~ 2002  0.184    0      0    14.3  16.4  109.
4 Afghanistan AFG   Southern~ 2003  0.200    0      0    14.4  16.6  109.
5 Afghanistan AFG   Southern~ 2004  0.222    0      0    15.2  16.7  109.
6 Afghanistan AFG   Southern~ 2005  0.255    0      0    15.3  16.9  108.
# i 11 more variables: male_edu <dbl>, temp <dbl>, rainfall1000 <dbl>,
#   death <dbl>, conflict <dbl>, maternalMor <dbl>, infantMor <dbl>,
#   neonatalMor <dbl>, under5Mor <dbl>, drought <dbl>, earthquake <dbl>
```

Then the bottom lines

```
data %>% tail()
```

```
# A tibble: 6 x 21
  country_name ISO   region   year gdp1000  OECD OECD2023 popdens urban agedep
  <chr>      <chr> <chr>   <dbl>  <dbl> <dbl>   <dbl>   <dbl> <dbl>
1 Zimbabwe   ZWE   Sub-Saha~ 2014   1.41    0      0    26.5  24.4  85.9
2 Zimbabwe   ZWE   Sub-Saha~ 2015   1.41    0      0    26.5  24.8  85.1
3 Zimbabwe   ZWE   Sub-Saha~ 2016   1.42    0      0    26.5  25.0  84.1
4 Zimbabwe   ZWE   Sub-Saha~ 2017   1.19    0      0    26.5  25.3  83.1
5 Zimbabwe   ZWE   Sub-Saha~ 2018   2.27    0      0    26.5  25.5  82.1
6 Zimbabwe   ZWE   Sub-Saha~ 2019   1.42    0      0    26.5  25.7  81.2
# i 11 more variables: male_edu <dbl>, temp <dbl>, rainfall1000 <dbl>,
#   death <dbl>, conflict <dbl>, maternalMor <dbl>, infantMor <dbl>,
#   neonatalMor <dbl>, under5Mor <dbl>, drought <dbl>, earthquake <dbl>
```

Now randomly select a few lines

```
data %>% slice_sample(n = 6)
```

```
# A tibble: 6 x 21
  country_name ISO region year gdp1000 OECD OECD2023 popdens urban agedep
  <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Somalia SOM Sub-Saha~ 2009 NA 0 0 11.7 26.2 103.
2 Tanzania TZA Sub-Saha~ 2017 0.976 0 0 22.2 17.6 91.1
3 Sri Lanka LKA Southern~ 2012 3.45 0 0 44.1 53.0 50.0
4 Turkey TUR Western ~ 2005 7.37 1 1 30.1 25.2 52.9
5 Lebanon LBN Western ~ 2007 5.16 0 0 56.3 51.6 49.6
6 Panama PAN Latin Am~ 2019 15.8 0 0 40.1 30.3 54.1
# i 11 more variables: male_edu <dbl>, temp <dbl>, rainfall1000 <dbl>,
# death <dbl>, conflict <dbl>, maternalMor <dbl>, infantMor <dbl>,
# neonatalMor <dbl>, under5Mor <dbl>, drought <dbl>, earthquake <dbl>
```

Check the class of all variables

```
data %>% glimpse()
```

```
Rows: 3,720
Columns: 21
$ country_name <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan~
$ ISO <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "~
$ region <chr> "Southern Asia", "Southern Asia", "Southern Asia", "South~
$ year <dbl> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 200~
$ gdp1000 <dbl> NA, NA, 0.1835328, 0.2004626, 0.2216576, 0.2550551, 0.274~
$ OECD <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ OECD2023 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ popdens <dbl> 14.13654, 14.23156, 14.32270, 14.40691, 15.21947, 15.3361~
$ urban <dbl> 16.25324, 16.25661, 16.42654, 16.60701, 16.71367, 16.8509~
$ agedep <dbl> 108.34663, 108.98989, 109.34716, 109.44753, 109.28682, 10~
$ male_edu <dbl> 2.762086, 2.856936, 2.954241, 3.054121, 3.156706, 3.26213~
$ temp <dbl> 12.69959, 12.85570, 12.71081, 12.16592, 13.04643, 12.2314~
$ rainfall1000 <dbl> 0.2763704, 0.2793079, 0.3805710, 0.4288939, 0.3754336, 0.~
$ death <dbl> 5065, 5394, 5553, 1157, 944, 817, 1711, 4982, 7020, 5660,~
$ conflict <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ maternalMor <dbl> 1450, 1390, 1300, 1240, 1180, 1140, 1120, 1090, 1030, 993~
$ infantMor <dbl> 90.5, 87.9, 85.3, 82.7, 80.0, 77.3, 74.6, 71.9, 69.2, 66.~
$ neonatalMor <dbl> 60.9, 59.7, 58.5, 57.2, 55.9, 54.6, 53.2, 51.7, 50.3, 48.~
$ under5Mor <dbl> 129.2, 125.2, 121.1, 116.9, 112.6, 108.4, 104.1, 99.9, 95~
$ drought <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, ~
$ earthquake <dbl> 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, ~
```

And let's see some basic statistics of the variables

```
data %>% summary()
```

country_name	ISO	region	year
Length:3720	Length:3720	Length:3720	Min. :2000
Class :character	Class :character	Class :character	1st Qu.:2005
Mode :character	Mode :character	Mode :character	Median :2010
			Mean :2010
			3rd Qu.:2014
			Max. :2019

gdp1000	OECD	OECD2023	popdens
Min. : 0.1105	Min. :0.000	Min. :0.0000	Min. : 0.00
1st Qu.: 1.2383	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:14.79
Median : 4.0719	Median :0.000	Median :0.0000	Median :27.52
Mean : 11.4917	Mean :0.171	Mean :0.1882	Mean :30.57
3rd Qu.: 13.1531	3rd Qu.:0.000	3rd Qu.:0.0000	3rd Qu.:40.72
Max. :123.6787	Max. :1.000	Max. :1.0000	Max. :99.86
NA's :62			NA's :20

urban	agedep	male_edu	temp
Min. : 0.1025	Min. : 16.17	Min. : 1.067	Min. : -2.405
1st Qu.:17.2872	1st Qu.: 47.94	1st Qu.: 5.904	1st Qu.:12.928
Median :30.2535	Median : 55.51	Median : 8.368	Median :21.958
Mean :30.6948	Mean : 61.94	Mean : 8.258	Mean :19.625
3rd Qu.:41.6558	3rd Qu.: 77.11	3rd Qu.:10.849	3rd Qu.:25.869
Max. :93.4135	Max. :111.48	Max. :14.441	Max. :29.676
NA's :20		NA's :20	NA's :20

rainfall1000	death	conflict	maternalMor
Min. :0.01993	Min. : 0.0	Min. :0.0000	Min. : 2.0
1st Qu.:0.59146	1st Qu.: 0.0	1st Qu.:0.0000	1st Qu.: 17.0
Median :1.01288	Median : 0.0	Median :0.0000	Median : 66.0
Mean :1.20216	Mean : 361.1	Mean :0.1892	Mean : 210.6
3rd Qu.:1.68706	3rd Qu.: 2.0	3rd Qu.:0.0000	3rd Qu.: 299.8
Max. :4.71081	Max. :78644.0	Max. :1.0000	Max. :2480.0
NA's :20			NA's :426

infantMor	neonatalMor	under5Mor	drought
Min. : 1.60	Min. : 0.80	Min. : 2.00	Min. :0.00000
1st Qu.: 7.60	1st Qu.: 4.90	1st Qu.: 9.00	1st Qu.:0.00000
Median : 18.90	Median :12.10	Median : 22.20	Median :0.00000
Mean : 28.90	Mean :16.18	Mean : 40.50	Mean :0.08737
3rd Qu.: 44.52	3rd Qu.:25.32	3rd Qu.: 61.33	3rd Qu.:0.00000
Max. :138.10	Max. :60.90	Max. :224.90	Max. :1.00000
NA's :20	NA's :20	NA's :20	

```
    earthquake
Min.      :0.00000
1st Qu.   :0.00000
Median    :0.00000
Mean      :0.08333
3rd Qu.   :0.00000
Max.      :1.00000
```