Week 4 in class EDA

Aoqi Xie

2024-09-30

Perfect your GitHub repo

Some of you may still need to organize your GitHub repo. Use this time to do that. When you are confident with your repo, let me know – I will try to reproduce your code.

Your final data should have the following variables (you might have slightly different variable names).

```
finaldata <- read.csv(here("data", "primaryanalysis_data.csv"),
                      header = TRUE)
names(finaldata)
```

```
 [1] "Year"         "InfMor"       "NeoMor"       "UndMor"       "MatMor"
 [6] "ISO"          "totdeath"     "conflict"     "country_name" "region"
[11] "gdp1000"      "OECD"         "OECD2023"     "popdens"      "urban"
[16] "agedep"       "male_edu"     "temp"         "rainfall1000" "drought"
[21] "earthquake"
```

Observations from Canada should look like this...

```
finaldata %>%
  dplyr::filter(country_name == "Canada")
```

```
  Year InfMor NeoMor UndMor MatMor ISO totdeath conflict country_name
1 2000    5.3    3.8    6.2      9 CAN       23        0       Canada
2 2001    5.3    3.8    6.2     10 CAN        1        0       Canada
3 2002    5.3    3.9    6.2     10 CAN        0        0       Canada
4 2003    5.3    3.9    6.2     10 CAN        0        0       Canada
5 2004    5.3    3.9    6.1     10 CAN        0        0       Canada
```

```
6  2005    5.2    3.9    6.1     11 CAN        0        0       Canada
7  2006    5.2    3.9    6.0     11 CAN        0        0       Canada
8  2007    5.1    3.8    6.0     11 CAN        0        0       Canada
9  2008    5.1    3.8    5.9     12 CAN        0        0       Canada
10 2009    5.0    3.8    5.8     12 CAN        0        0       Canada
11 2010    5.0    3.8    5.7     11 CAN        0        0       Canada
12 2011    4.9    3.7    5.7     11 CAN        0        0       Canada
13 2012    4.9    3.7    5.6     11 CAN        0        0       Canada
14 2013    4.8    3.6    5.5     11 CAN        0        0       Canada
15 2014    4.7    3.6    5.4     11 CAN        0        0       Canada
16 2015    4.7    3.6    5.4     11 CAN        0        0       Canada
17 2016    4.6    3.5    5.3     10 CAN        0        0       Canada
18 2017    4.6    3.4    5.2     10 CAN        0        0       Canada
19 2018    4.5    3.3    5.1     NA CAN        0        0       Canada
20 2019    4.4    3.3    5.1     NA CAN        0        0       Canada
             region  gdp1000 OECD OECD2023  popdens    urban   agedep male_edu
1  Northern America 24.27100    1        1 66.19704 56.14335 46.34463 12.30281
2  Northern America 23.82206    1        1 66.45361 56.40270 45.89632 12.35258
3  Northern America 24.25534    1        1 66.71112 56.67093 45.46660 12.40182
4  Northern America 28.30046    1        1 66.96384 56.94365 45.07468 12.45053
5  Northern America 32.14368    1        1 67.21715 57.20020 44.67374 12.49870
6  Northern America 36.38251    1        1 67.47283 57.41671 44.26641 12.54635
7  Northern America 40.50406    1        1 67.73674 57.59143 43.96370 12.59349
8  Northern America 44.65990    1        1 67.99444 57.75691 43.83612 12.64015
9  Northern America 46.71051    1        1 68.25765 57.97905 43.85426 12.68634
10 Northern America 40.87631    1        1 68.53354 58.24228 43.94937 12.73207
11 Northern America 47.56208    1        1 68.80739 58.52809 44.13587 12.77735
12 Northern America 52.22370    1        1 69.04842 58.81437 44.53578 12.82218
13 Northern America 52.66909    1        1 69.27604 59.05573 45.18393 12.86660
14 Northern America 52.63517    1        1 69.50772 59.19713 45.95404 12.91059
15 Northern America 50.95600    1        1 69.76876 59.30361 46.75493 12.95414
16 Northern America 43.59614    1        1 69.98853 59.42627 47.59164 12.99723
17 Northern America 42.31560    1        1 70.21484 59.50521 48.41410 13.03988
18 Northern America 45.12943    1        1 70.40863 59.59325 49.14806 13.08210
19 Northern America 46.54864    1        1 70.63614 59.68433 49.80166 13.12388
20 Northern America 46.32867    1        1 70.83794 59.75984 50.47739 13.16522
       temp rainfall1000 drought earthquake
1  5.486244    0.9971559       0          0
2  6.469105    0.8644873       0          0
3  5.979147    0.9460938       0          0
4  5.416964    1.0189234       0          0
5  5.556961    1.0008237       0          0
6  6.187472    1.0367199       0          0
```

```
7  6.895084    1.0917386       0          0
8  5.900051    1.0134091       0          0
9  5.650118    1.0693435       0          0
10 5.398867    0.9928497       0          0
11 6.781766    1.0379754       0          0
12 6.269133    1.1343442       0          0
13 7.249497    0.9747708       0          0
14 5.954381    1.0282075       0          0
15 5.584650    1.0377695       0          0
16 6.436884    0.9632446       0          0
17 7.184514    0.9677826       0          0
18 6.539669    1.0995322       0          0
19 6.539677    1.0991469       0          0
20 6.539633    1.0987523       0          0
```

Observations from Ecuador should look like this…

```
finaldata %>%
  dplyr::filter(country_name == "Ecuador")
```

```
   Year InfMor NeoMor UndMor MatMor ISO totdeath conflict country_name
1  2000   24.7   14.1   29.5    122 ECU        0        0      Ecuador
2  2001   23.4   13.4   28.0    117 ECU        2        0      Ecuador
3  2002   22.4   12.7   26.6    110 ECU        0        0      Ecuador
4  2003   21.5   12.1   25.4    100 ECU       26        1      Ecuador
5  2004   20.7   11.6   24.4     94 ECU        0        0      Ecuador
6  2005   19.9   11.1   23.5     94 ECU        0        0      Ecuador
7  2006   19.2   10.6   22.6     90 ECU        0        0      Ecuador
8  2007   18.5   10.2   21.7     85 ECU        0        0      Ecuador
9  2008   17.7    9.7   20.8     82 ECU       25        0      Ecuador
10 2009   17.0    9.3   19.9     80 ECU        0        0      Ecuador
11 2010   16.3    8.9   19.0     78 ECU        0        0      Ecuador
12 2011   15.6    8.5   18.1     76 ECU        0        0      Ecuador
13 2012   14.9    8.1   17.3     71 ECU        0        0      Ecuador
14 2013   14.3    7.8   16.6     67 ECU        0        0      Ecuador
15 2014   13.7    7.5   15.9     65 ECU        0        0      Ecuador
16 2015   13.2    7.3   15.4     63 ECU        0        0      Ecuador
17 2016   12.8    7.1   14.8     61 ECU        0        0      Ecuador
18 2017   12.4    6.9   14.4     59 ECU        0        0      Ecuador
19 2018   12.0    6.9   13.9     NA ECU        0        0      Ecuador
20 2019   11.6    6.8   13.4     NA ECU        0        0      Ecuador
                            region  gdp1000 OECD OECD2023  popdens    urban
```

```
1  Latin America and the Caribbean 1.451531    0        0 23.27432 36.19963
2  Latin America and the Caribbean 1.904814    0        0 23.39372 36.67994
3  Latin America and the Caribbean 2.184209    0        0 23.52087 37.08903
4  Latin America and the Caribbean 2.438344    0        0 23.58358 37.23792
5  Latin America and the Caribbean 2.703566    0        0 38.43743 37.39268
6  Latin America and the Caribbean 3.014310    0        0 38.55361 37.36968
7  Latin America and the Caribbean 3.340841    0        0 38.65018 37.47567
8  Latin America and the Caribbean 3.579032    0        0 38.76505 37.68172
9  Latin America and the Caribbean 4.260433    0        0 38.83977 37.67445
10 Latin America and the Caribbean 4.240703    0        0 38.92613 37.39437
11 Latin America and the Caribbean 4.640246    0        0 39.03066 37.26838
12 Latin America and the Caribbean 5.202656    0        0 39.09586 37.61553
13 Latin America and the Caribbean 5.678456    0        0 39.13343 38.00733
14 Latin America and the Caribbean 6.050355    0        0 39.18619 38.22511
15 Latin America and the Caribbean 6.374631    0        0 39.27871 38.12421
16 Latin America and the Caribbean 6.130587    0        0 39.38824 38.15633
17 Latin America and the Caribbean 6.079089    0        0 39.46201 38.45745
18 Latin America and the Caribbean 6.246404    0        0 39.53609 38.65993
19 Latin America and the Caribbean 6.321349    0        0 39.58380 38.87253
20 Latin America and the Caribbean 6.233258    0        0 39.75109 39.05144
     agedep male_edu     temp rainfall1000 drought earthquake
1  67.44216 7.738627 19.54855    1.4201653       0          0
2  66.57356 7.843942 19.66622    1.1667746       0          0
3  65.65488 7.949449 20.24695    1.4577981       0          0
4  64.71472 8.055240 20.05016    1.5781807       0          0
5  63.78049 8.161433 20.10136    1.0683450       0          0
6  62.86530 8.268176 19.88163    0.8555447       0          0
7  61.97042 8.375587 20.07087    1.1114502       0          0
8  61.11422 8.483729 19.49536    1.0899082       0          0
9  60.31015 8.592603 19.85711    1.6184816       0          0
10 59.55262 8.702180 20.39298    1.0870796       1          0
11 58.83793 8.812409 20.11160    1.7045703       0          0
12 58.16553 8.923172 19.86633    1.4518388       0          0
13 57.51051 9.034284 20.19000    1.7520003       0          0
14 56.84804 9.145523 19.85177    1.3735605       1          0
15 56.17001 9.256679 20.42252    1.2572257       0          1
16 55.46511 9.367582 20.95595    1.7284273       0          0
17 54.73369 9.478071 20.77476    1.3168761       0          2
18 53.99096 9.587993 20.53262    1.9544485       0          0
19 53.12249 9.697221 20.53714    1.9573265       0          0
20 52.29278 9.805670 20.54169    1.9602443       0          2
```

Exploratory data analysis

Use the rest of the class time to explore the final data that will be used for analysis starting next week. At the end of the class, write a summary of your findings and push your **Quarto document (pdf)** to your repo.

1. Understand the structure of the data From the results below, there are a total of 5320 observations with 21 variables (8 integer types, 10 numerical types, and 3 character types). We can also see some missingness in the data set, which should be carefully taken care of.

```
str(finaldata)
```

```
'data.frame':   5320 obs. of  21 variables:
 $ Year        : int  2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 ...
 $ InfMor      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ NeoMor      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ UndMor      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ MatMor      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ ISO         : chr  "ABW" "ABW" "ABW" "ABW" ...
 $ totdeath    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ conflict    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ country_name: chr  NA NA NA NA ...
 $ region      : chr  NA NA NA NA ...
 $ gdp1000     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ OECD        : int  NA NA NA NA NA NA NA NA NA NA ...
 $ OECD2023    : int  NA NA NA NA NA NA NA NA NA NA ...
 $ popdens     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ urban       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ agedep      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ male_edu    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ temp        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ rainfall1000: num  NA NA NA NA NA NA NA NA NA NA ...
 $ drought     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ earthquake  : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
summary(finaldata)
```

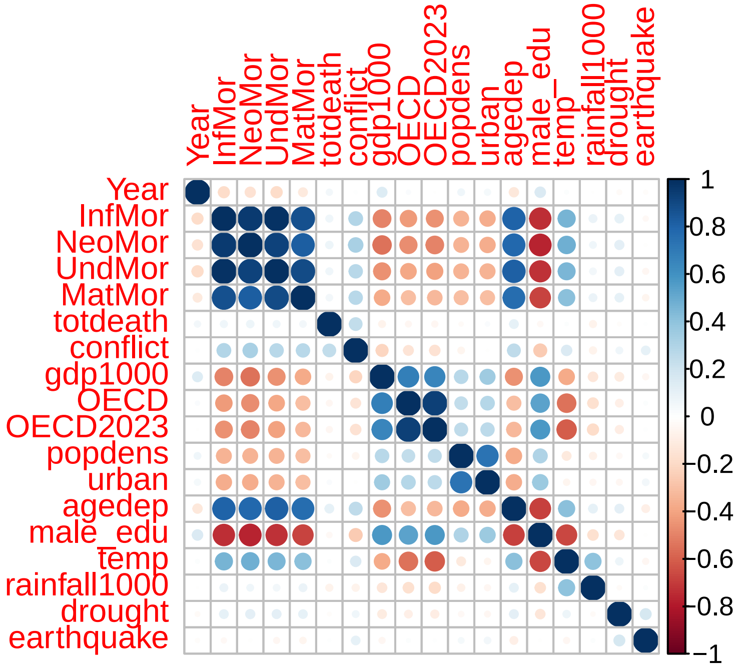
```
      Year          InfMor           NeoMor          UndMor
 Min.   :2000   Min.   :  1.60   Min.   : 0.80   Min.   :  1.800
 1st Qu.:2005   1st Qu.:  8.41   1st Qu.: 5.30   1st Qu.:  9.992
```

```
Median :2010   Median : 21.10   Median :13.18   Median : 24.900
Mean   :2010   Mean   : 29.84   Mean   :16.83   Mean   : 41.851
3rd Qu.:2014   3rd Qu.: 46.40   3rd Qu.:26.45   3rd Qu.: 64.600
Max.   :2019   Max.   :138.10   Max.   :60.90   Max.   :224.900
               NA's   :500      NA's   :500     NA's   :500
    MatMor           ISO               totdeath          conflict
Min.   :   2.0   Length:5320        Min.   :    0.0   Min.   :0.0000
1st Qu.:  20.0   Class :character   1st Qu.:    0.0   1st Qu.:0.0000
Median :  78.0   Mode  :character   Median :    0.0   Median :0.0000
Mean   : 220.5                      Mean   :  233.9   Mean   :0.1233
3rd Qu.: 331.8                      3rd Qu.:    0.0   3rd Qu.:0.0000
Max.   :2480.0                      Max.   :78644.0   Max.   :1.0000
NA's   :1126
country_name          region             gdp1000              OECD
Length:5320        Length:5320        Min.   :  0.1105   Min.   :0.000
Class :character   Class :character   1st Qu.:  1.2383   1st Qu.:0.000
Mode  :character   Mode  :character   Median :  4.0719   Median :0.000
                                      Mean   : 11.4917   Mean   :0.171
                                      3rd Qu.: 13.1531   3rd Qu.:0.000
                                      Max.   :123.6787   Max.   :1.000
                                      NA's   :1662       NA's   :1600
   OECD2023         popdens          urban             agedep
Min.   :0.0000   Min.   : 0.00   Min.   : 0.1025   Min.   : 16.17
1st Qu.:0.0000   1st Qu.:14.79   1st Qu.:17.2872   1st Qu.: 47.94
Median :0.0000   Median :27.52   Median :30.2535   Median : 55.51
Mean   :0.1882   Mean   :30.57   Mean   :30.6948   Mean   : 61.94
3rd Qu.:0.0000   3rd Qu.:40.72   3rd Qu.:41.6558   3rd Qu.: 77.11
Max.   :1.0000   Max.   :99.86   Max.   :93.4135   Max.   :111.48
NA's   :1600     NA's   :1620    NA's   :1620      NA's   :1600
   male_edu           temp          rainfall1000       drought
Min.   : 1.067   Min.   :-2.405   Min.   :0.0199   Min.   :0.00000
1st Qu.: 5.904   1st Qu.:12.928   1st Qu.:0.5915   1st Qu.:0.00000
Median : 8.368   Median :21.958   Median :1.0129   Median :0.00000
Mean   : 8.258   Mean   :19.625   Mean   :1.2022   Mean   :0.06372
3rd Qu.:10.849   3rd Qu.:25.869   3rd Qu.:1.6871   3rd Qu.:0.00000
Max.   :14.441   Max.   :29.676   Max.   :4.7108   Max.   :3.00000
NA's   :1620     NA's   :1620     NA's   :1620
  earthquake
Min.   : 0.0000
1st Qu.: 0.0000
Median : 0.0000
Mean   : 0.1017
3rd Qu.: 0.0000
```

```
Max.   :11.0000
```

2. Understand the correlation among the numeric variables. We can see that the mortality rates are highly correlated, and moderate correlation observed among `gdp1000`, `OECD`, and `popdens`.

```
correlation_matrix <- cor(finaldata %>% select_if(is.numeric), use = "complete.obs")  # Sele
corrplot(correlation_matrix, method = "circle")  # Visualize correlation matrix
```


Year
InfMor
NeoMor
UndMor
MatMor
totdeath
conflict
gdp1000
OECD
OECD2023
popdens
urban
agedep
male_edu
temp
rainfall1000
drought
earthquake