Exploratory data analysis

Exploratory data analysis

Name: Ting Lin

First of all, I would like to perform `summary()` to check some generic information about our data set.

```r
library(here)
```

```
here() starts at /Users/apple/Desktop/armed_conflict
```

```r
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 becom
```

```r
source("R/FinalMergedData.R")
```

```
Warning: Some values were not matched unambiguously: Africa Eastern and Southern, Africa Wes
```

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

```r
head(allfinal,n=10)
```

```
   country_name ISO        region year   gdp1000 OECD OECD2023  popdens
1   Afghanistan AFG Southern Asia 2000        NA    0        0 14.13654
2   Afghanistan AFG Southern Asia 2001        NA    0        0 14.23156
3   Afghanistan AFG Southern Asia 2002 0.1835328    0        0 14.32270
4   Afghanistan AFG Southern Asia 2003 0.2004626    0        0 14.40691
5   Afghanistan AFG Southern Asia 2004 0.2216576    0        0 15.21947
6   Afghanistan AFG Southern Asia 2005 0.2550551    0        0 15.33619
7   Afghanistan AFG Southern Asia 2006 0.2740005    0        0 15.43982
8   Afghanistan AFG Southern Asia 2007 0.3750781    0        0 15.65217
9   Afghanistan AFG Southern Asia 2008 0.3878492    0        0 15.74447
10  Afghanistan AFG Southern Asia 2009 0.4438452    0        0 15.83043
      urban   agedep male_edu     temp rainfall1000 totaldeath armconf1
1  16.25324 108.3466 2.762086 12.69959    0.2763704       5065        1
2  16.25661 108.9899 2.856936 12.85570    0.2793079       5394        1
3  16.42654 109.3472 2.954241 12.71081    0.3805710       5553        1
4  16.60701 109.4475 3.054121 12.16592    0.4288939       1157        1
5  16.71367 109.2868 3.156706 13.04643    0.3754336        944        1
6  16.85096 107.9646 3.262133 12.23141    0.4415680        817        1
7  16.98105 106.3262 3.370551 12.96153    0.4437097       1711        1
8  17.12259 108.3381 3.482112 12.47451    0.4092555       4982        1
9  17.26919 109.2404 3.596977 12.63527    0.3901204       7020        1
10 17.43508 106.8458 3.715306 12.61764    0.4808727       5660        1
   MaternalMortalityRate InfantMortalityRate NeonatalMortalityRate
1                   1450                90.5                  60.9
2                   1390                87.9                  59.7
3                   1300                85.3                  58.5
4                   1240                82.7                  57.2
5                   1180                80.0                  55.9
6                   1140                77.3                  54.6
7                   1120                74.6                  53.2
8                   1090                71.9                  51.7
9                   1030                69.2                  50.3
10                   993                66.7                  48.9
   Under5MortalityRate drought earthquake
1                129.2       1          0
```

```
2                125.2       0          1
3                121.1       0          1
4                116.9       0          1
5                112.6       0          1
6                108.4       0          1
7                104.1       1          1
8                 99.9       0          0
9                 95.7       1          0
10                91.7       0          1
```

```r
tail(allfinal, n=10)
```

```
     country_name ISO             region year   gdp1000 OECD OECD2023  popdens
3711     Zimbabwe ZWE Sub-Saharan Africa 2010 0.9378403    0        0 25.51039
3712     Zimbabwe ZWE Sub-Saharan Africa 2011 1.0826158    0        0 25.53206
3713     Zimbabwe ZWE Sub-Saharan Africa 2012 1.2901940    0        0 25.55349
3714     Zimbabwe ZWE Sub-Saharan Africa 2013 1.4083678    0        0 25.53286
3715     Zimbabwe ZWE Sub-Saharan Africa 2014 1.4070343    0        0 26.52884
3716     Zimbabwe ZWE Sub-Saharan Africa 2015 1.4103292    0        0 26.54454
3717     Zimbabwe ZWE Sub-Saharan Africa 2016 1.4217878    0        0 26.53811
3718     Zimbabwe ZWE Sub-Saharan Africa 2017 1.1921070    0        0 26.49281
3719     Zimbabwe ZWE Sub-Saharan Africa 2018 2.2691770    0        0 26.47943
3720     Zimbabwe ZWE Sub-Saharan Africa 2019 1.4218686    0        0 26.46341
        urban   agedep male_edu     temp rainfall1000 totaldeath armconf1
3711 23.28851 85.56457 8.250225 21.53473    0.7290925          0        0
3712 23.43075 86.40049 8.358820 20.87452    0.8582386          0        0
3713 23.70160 86.71712 8.466529 20.98071    0.6259767          1        0
3714 24.04603 86.44543 8.573429 20.77221    0.6717220          1        0
3715 24.40427 85.87550 8.679591 20.87651    0.6777257          0        0
3716 24.75233 85.08337 8.785078 21.45470    0.4490721          0        0
3717 25.02842 84.11222 8.889947 21.39290    0.4939246          0        0
3718 25.29333 83.10129 8.994252 20.85962    0.9533149          0        0
3719 25.53759 82.12335 9.098048 20.86041    0.9535655          0        0
3720 25.70572 81.20786 9.201384 20.86120    0.9538138          4        0
     MaternalMortalityRate InfantMortalityRate NeonatalMortalityRate
3711                   598                52.1                  30.8
3712                   557                50.8                  30.1
3713                   528                46.5                  29.4
3714                   509                44.8                  28.7
3715                   494                42.9                  28.2
3716                   480                42.1                  27.8
3717                   468                40.8                  27.4
```

```
3718                   458                39.9                  27.0
3719                    NA                38.8                  26.6
3720                    NA                38.1                  26.2
     Under5MortalityRate drought earthquake
3711                86.4       1          0
3712                80.8       0          0
3713                72.2       0          0
3714                66.3       1          0
3715                62.7       0          0
3716                61.3       0          0
3717                58.7       0          0
3718                57.0       1          0
3719                54.8       0          0
3720                54.2       0          0
```

```r
summary(allfinal)
```

```
 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       totaldeath         armconf1      MaternalMortalityRate
 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
 InfantMortalityRate NeonatalMortalityRate Under5MortalityRate
 Min.   :  1.60      Min.   : 0.80         Min.   :  2.00
 1st Qu.:  7.60      1st Qu.: 4.90         1st Qu.:  9.00
 Median : 18.90      Median :12.10         Median : 22.20
 Mean   : 28.90      Mean   :16.18         Mean   : 40.50
 3rd Qu.: 44.52      3rd Qu.:25.32         3rd Qu.: 61.33
 Max.   :138.10      Max.   :60.90         Max.   :224.90
 NA's   :20          NA's   :20            NA's   :20
    drought          earthquake
 Min.   :0.00000   Min.   :0.00000
 1st Qu.:0.00000   1st Qu.:0.00000
 Median :0.00000   Median :0.00000
 Mean   :0.08737   Mean   :0.08333
 3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :1.00000   Max.   :1.00000
```

From the `summery()`, there are 426 missing values for Maternal Mortality Rate, 62 missing values for gpd100, 20 missing values for popdens, urban, male_edu, temp, rainfall1000, Infant Mortality Rate, Neonatal Mortality Rate and Under5 Mortality Rate.

Then I would like to look at Total Death in details.

The Minimum value for total death is 20, and Maximum is 78644. The range is pretty wide.

I would like to locate which country has the maximum total death and visualize it by year to see the patterns for this specific country.

```
# visualize Maternal Mortality Rate by year.
max_death_country <- allfinal$country_name[which.max(allfinal$totaldeath)]
print(max_death_country)
```

```
[1] "Syria"
```

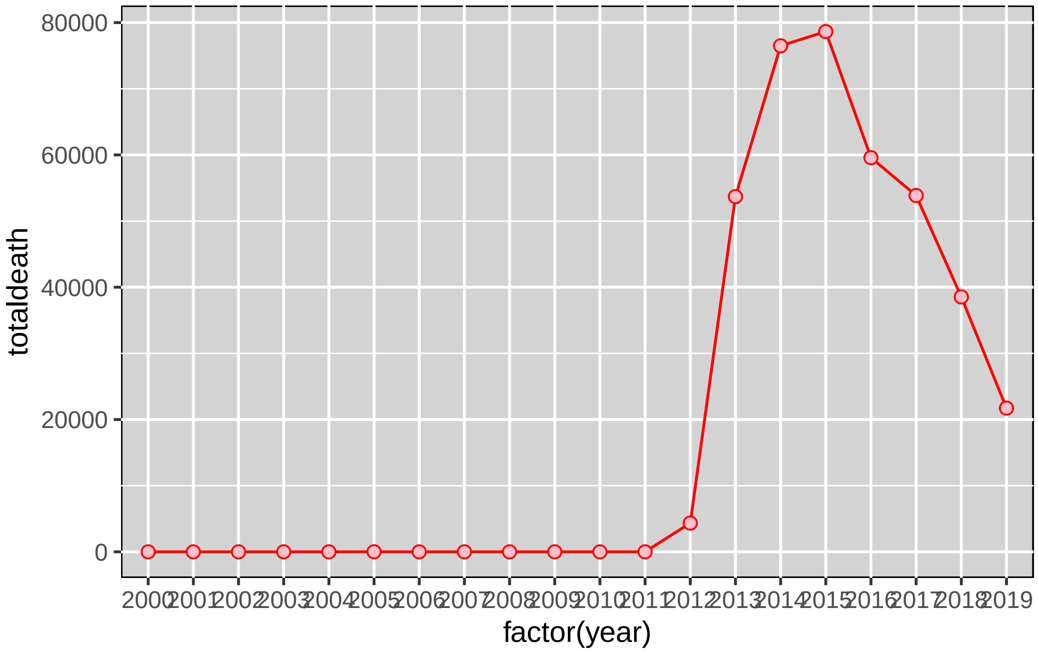
```r
library(ggplot2)
```

And the answer is Syria.

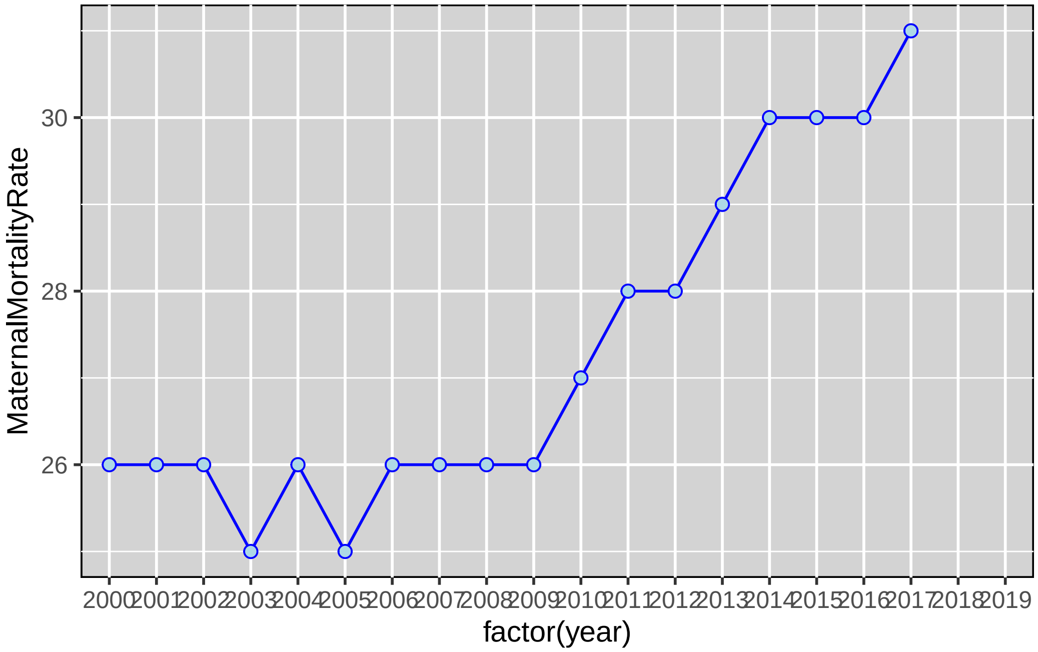
```r
library(ggplot2)
library(gridExtra)
```

```

Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

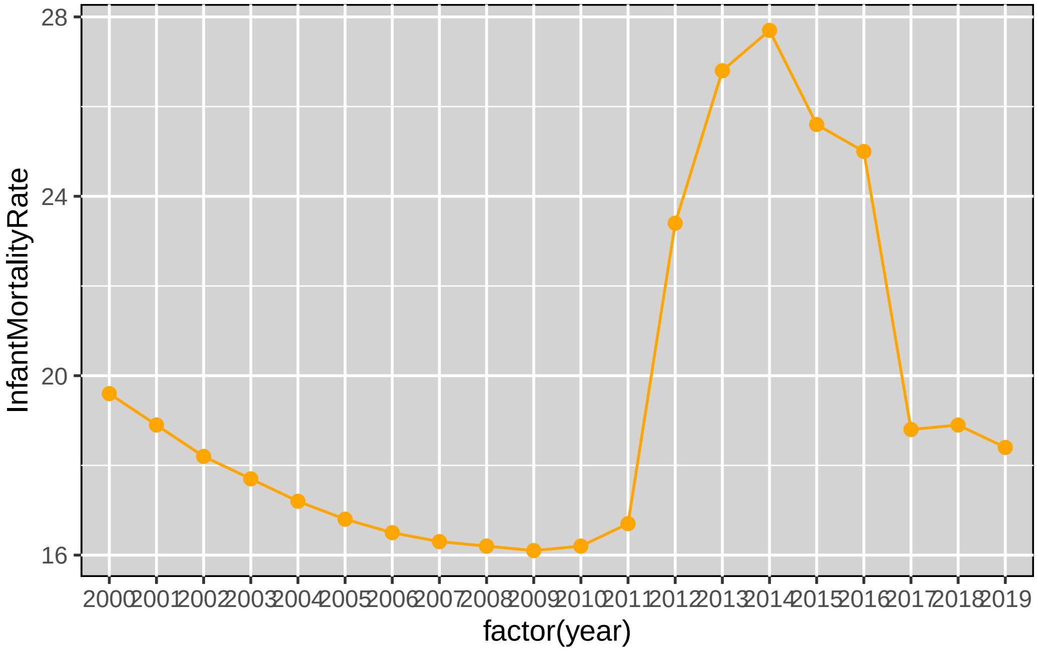
```r
Syria <- allfinal[allfinal$country_name == "Syria",]
plot1 <- Syria %>%
  ggplot(mapping = aes(x = factor(year), y = totaldeath)) +
  geom_line(na.rm = TRUE, group = 1, color="red") +
  geom_point(na.rm = TRUE,shape=21, color="red", fill="pink", size=2) +
  scale_x_discrete(breaks = unique(Syria$year)) +
  theme(panel.background = element_rect(fill = "lightgrey", color = "black"))
plot2 <- Syria %>%
  ggplot(mapping = aes(x = factor(year), y = MaternalMortalityRate)) +
  geom_line(na.rm = TRUE, group = 1, color = "blue") +
  geom_point(na.rm = TRUE, shape = 21, color = "blue", fill = "lightblue", size = 2) +
  scale_x_discrete(breaks = unique(Syria$year)) +
  theme(panel.background = element_rect(fill = "lightgrey", color = "black"))
plot3 <- Syria %>%
  ggplot(mapping = aes(x = factor(year), y = InfantMortalityRate)) +
  geom_line(na.rm = TRUE, group = 1, color = "orange") +
  geom_point(na.rm = TRUE, shape = 21, color = "orange", fill = "orange", size = 2) +
  scale_x_discrete(breaks = unique(Syria$year)) +
  theme(panel.background = element_rect(fill = "lightgrey", color = "black"))
plot4 <- Syria %>%
  ggplot(mapping = aes(x = factor(year), y = NeonatalMortalityRate)) +
  geom_line(na.rm = TRUE, group = 1, color = "purple") +
  geom_point(na.rm = TRUE, shape = 21, color = "purple", fill = "purple", size = 2) +
  scale_x_discrete(breaks = unique(Syria$year)) +
  theme(panel.background = element_rect(fill = "lightgrey", color = "black"))
plot5 <- Syria %>%
  ggplot(mapping = aes(x = factor(year), y = Under5MortalityRate)) +
  geom_line(na.rm = TRUE, group = 1, color = "green") +
  geom_point(na.rm = TRUE, shape = 21, color = "green", fill = "green", size = 2) +
  scale_x_discrete(breaks = unique(Syria$year)) +
  theme(panel.background = element_rect(fill = "lightgrey", color = "black"))

plot1
```



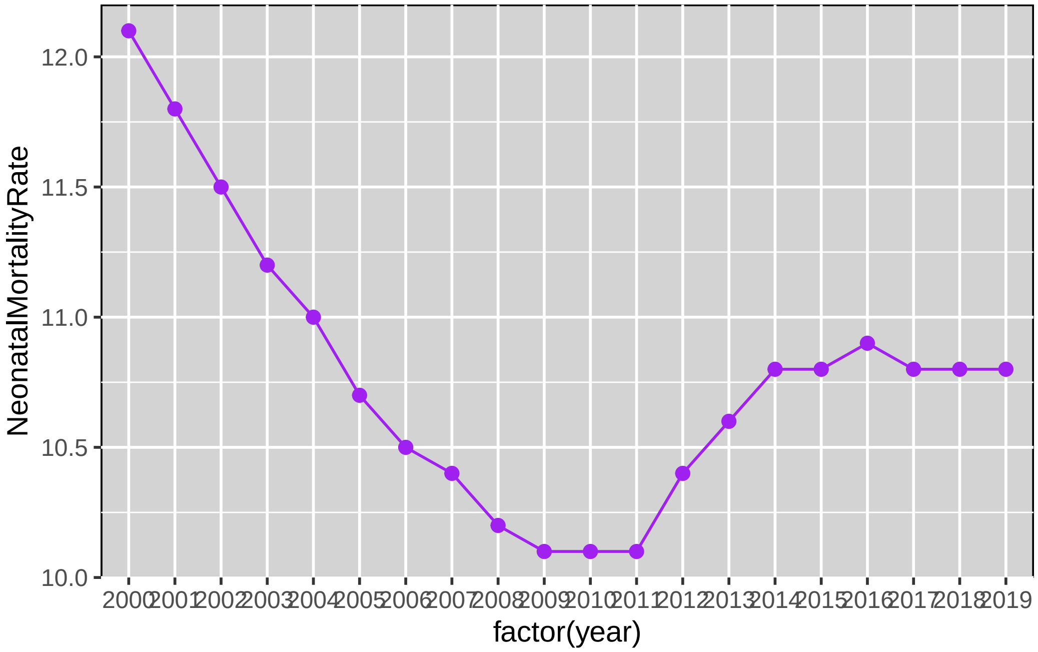
```
plot2
```


MaternalMortalityRate
30
28
26
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
factor(year)

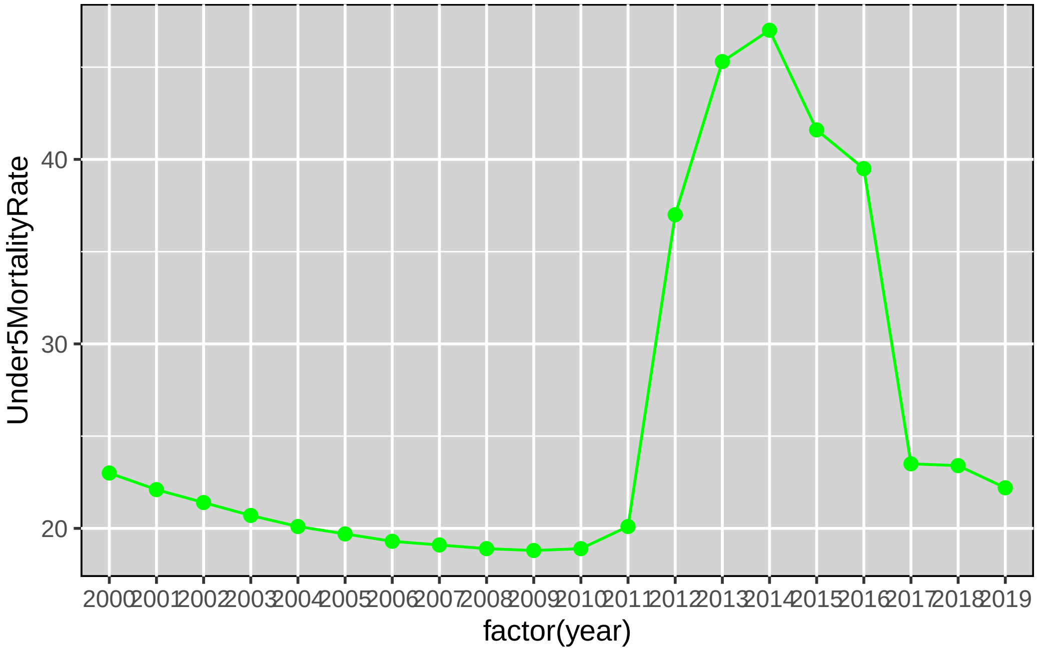
```
plot3
```


InfantMortalityRate
28
24
20
16
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
factor(year)

```
plot4
```



```
plot5
```



```
#plotall4 <- grid.arrange(plot2, plot3, plot4, plot5, ncol = 2)
#plotall4
```