HW1_RJS_17208987

Liang Tong

2024-06-18

1.Use `data.table` to read in the data and assign the correct class to the variables.

The data I choose is:

1. ***Human Development Indicators for Ireland***
2. ***Human Development Indicators for China***

In this step I'll load required libraries and datasets

```
# load library
library(data.table)
library(ggplot2)

# load datasets, skip first row
country1 <- fread("Human Development Indicators for Ireland2024.csv",skip=1)
country2 <- fread("Human Development Indicators for China2024.csv",skip=1)


# convert to data table format, good for large size
country1_DT <- as.data.table(country1)
country2_DT <- as.data.table(country2)
class(country1_DT$'#indicator+value+num')
```

```
[1] "numeric"
```

1.1

Assign the correct class to the variables, here the variable unit is correct: *numeric* for **'Value'** and *int* for **'Year'**

```
#reset names and check unit
 setnames(country1_DT, c("country_code", "country_name", "indicator_id", "indicator_name", "index_id", "index_r
  setnames(country2_DT, c("country_code", "country_name", "indicator_id", "indicator_name", "index_id", "index_
  country1_DT
```

```
    country_code country_name indicator_id
          <char>       <char>       <char>
  1:          IRL      Ireland          abr
  2:          IRL      Ireland          abr
  3:          IRL      Ireland          abr
  4:          IRL      Ireland          abr
  5:          IRL      Ireland          abr
 ---
890:          IRL      Ireland         se_m
891:          IRL      Ireland         se_m
892:          IRL      Ireland         se_m
893:          IRL      Ireland         se_m
894:          IRL      Ireland         se_m
                                                                indicator_name
                                                                        <char>
  1:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
  2:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
  3:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
  4:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
  5:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
 ---
890: Population with at least some secondary education, male (% ages 25 and older)
891: Population with at least some secondary education, male (% ages 25 and older)
892: Population with at least some secondary education, male (% ages 25 and older)
893: Population with at least some secondary education, male (% ages 25 and older)
```

2. Merge the data data sets using `data.table`

```
# base R example,  by=(); for data.table , .on(...)
#print(colnames(country2_DT))

# useing data table to joining tables
merged_DT <- rbind(country1_DT, country2_DT)
#merged_DT <- country1_DT[country1_DTT, on = .(country_code,country_name, #indicator_id,indicator_name, index_i
merged_DT
```

```
      country_code country_name       indicator_id
             <char>       <char>             <char>
   1:          IRL      Ireland                abr
   2:          IRL      Ireland                abr
   3:          IRL      Ireland                abr
   4:          IRL      Ireland                abr
   5:          IRL      Ireland                abr
  ---
1766:          CHN        China               se_m
1767:          CHN        China               se_m
1768:          CHN        China               se_m
1769:          CHN        China               se_m
1770:          CHN        China years_of_schooling
                                                                    indicator_name
                                                                            <char>
   1:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
   2:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
   3:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
   4:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
   5:                    Adolescent Birth Rate (births per 1,000 women ages 15-19)
  ---
1766: Population with at least some secondary education, male (% ages 25 and older)
1767: Population with at least some secondary education, male (% ages 25 and older)
1768: Population with at least some secondary education, male (% ages 25 and older)
1769: Population with at least some secondary education, male (% ages 25 and older)
```

3. Quick data exploration

here show the variable names of the merged data sets:

- we quickly explore the difference between Ireland and China
- The index I’m looking intro is

	Indicator_name	Indicator_id
1	Expected Years of Schooling (years)	eys
2	Life Expectancy at Birth (years)	le
3	Material footprint per capita (tonnes)	mf

```
#show the variable name
print(colnames(merged_DT))
```

```
[1] "country_code"   "country_name"   "indicator_id"   "indicator_name"
[5] "index_id"       "index_name"     "value"          "year"
```

3.1

Here we can see The mean,min and max for Ireland and China's **Expected Years of Schooling (years)** for past 33 years. Ireland has a much higher eys for the quick check.

```
#Ireland and China's gender mean In equality Index
merged_DT[indicator_id=="eys",
          .(eys_mean = mean(value),eys_min = min(value),eys_max = max(value), .N),
          by= country_name]
```

```
   country_name eys_mean eys_min eys_max     N
         <char>    <num>   <num>   <num> <int>
1:      Ireland 16.70494  12.679  19.756    33
2:        China 11.73403   8.581  15.218    33
```

3.2

do the same for **Life Expectancy at Birth (years)** and **Material footprint per capita (tonnes)**. This time Ireland has a slightly higher le and a much higher mf in the past 33 years

```
merged_DT[indicator_id=="le",
          .(le_mean = mean(value),le_min = min(value),le_max = max(value), .N),
          by= country_name]
```

```
   country_name  le_mean le_min le_max     N
         <char>    <num>  <num>  <num> <int>
1:      Ireland 78.84452 74.842 82.716    33
2:        China 73.88658 68.005 78.587    33
```

```
merged_DT[indicator_id=="mf",
          .(Gmf_mean = mean(value),mf_min = min(value),mf_max = max(value), .N),
          by= country_name]
```

```
   country_name Gmf_mean mf_min mf_max     N
         <char>    <num>  <num>  <num> <int>
1:      Ireland 30.78648 17.482 61.137    33
2:        China 14.44979  5.229 24.283    33
```

4. Analysis using data.table - keyby() used

compare **Expected Years of Schooling (years)** and rank it by year and country. we could find that both countries's year is increasing over the past 33 years

```
eys_IRL <- merged_DT[ country_name=="Ireland" & indicator_id=="eys",
                      .SD,
                      keyby=.(year,value)]
eys_CHN <- merged_DT[ country_name=="China" & indicator_id=="eys",
                      .SD,
                      keyby=.(year,value)]
eys_IRL
```

```
Key: <year, value>
     year  value country_code country_name indicator_id
    <int>  <num>       <char>       <char>       <char>
 1:  1990 12.679          IRL      Ireland          eys
 2:  1991 12.741          IRL      Ireland          eys
 3:  1992 12.938          IRL      Ireland          eys
 4:  1993 13.409          IRL      Ireland          eys
 5:  1994 13.765          IRL      Ireland          eys
 6:  1995 13.861          IRL      Ireland          eys
 7:  1996 13.907          IRL      Ireland          eys
 8:  1997 13.989          IRL      Ireland          eys
 9:  1998 16.495          IRL      Ireland          eys
10:  1999 16.325          IRL      Ireland          eys
11:  2000 16.430          IRL      Ireland          eys
12:  2001 16.563          IRL      Ireland          eys
13:  2002 16.806          IRL      Ireland          eys
14:  2003 16.982          IRL      Ireland          eys
15:  2004 17.240          IRL      Ireland          eys
16:  2005 17.239          IRL      Ireland          eys
17:  2006 16.971          IRL      Ireland          eys
18:  2007 17.014          IRL      Ireland          eys
19:  2008 17.262          IRL      Ireland          eys
20:  2009 17.037          IRL      Ireland          eys
21:  2010 17.463          IRL      Ireland          eys
22:  2011 17.938          IRL      Ireland          eys
```

```
#eys_CHN
```

4. 1

```
1 eys_CHN
```

```
Key: <year, value>
     year  value country_code country_name indicator_id
    <int>  <num>       <char>       <char>       <char>
 1:  1990  8.606          CHN        China          eys
 2:  1991  8.596          CHN        China          eys
 3:  1992  8.600          CHN        China          eys
 4:  1993  8.606          CHN        China          eys
 5:  1994  8.581          CHN        China          eys
 6:  1995  8.725          CHN        China          eys
 7:  1996  8.975          CHN        China          eys
 8:  1997  9.137          CHN        China          eys
 9:  1998  9.393          CHN        China          eys
10:  1999  9.649          CHN        China          eys
11:  2000  9.905          CHN        China          eys
12:  2001 10.160          CHN        China          eys
13:  2002 10.506          CHN        China          eys
14:  2003 10.852          CHN        China          eys
15:  2004 11.199          CHN        China          eys
16:  2005 11.545          CHN        China          eys
17:  2006 11.891          CHN        China          eys
18:  2007 12.203          CHN        China          eys
19:  2008 12.519          CHN        China          eys
20:  2009 12.853          CHN        China          eys
21:  2010 13.043          CHN        China          eys
22:  2011 13.158          CHN        China          eys
```

4. 2

now analysis the **Life Expectancy at Birth (years).** This time get the same result but using different method-'keyby = TRUE'. sort using country list (China 33 then Ireland 33). we found that the life time is increasing , but Ireland is slightly higher than China

```
le_both <- merged_DT[ indicator_id=="le",
                      .SD,
                      by=.(country_name,year, value),
                      keyby = TRUE]#same result
le_both
```

```
Key: <country_name, year, value>
    country_name  year  value country_code indicator_id
          <char> <int>  <num>       <char>       <char>
 1:        China  1990 68.005          CHN           le
 2:        China  1991 68.169          CHN           le
 3:        China  1992 68.734          CHN           le
 4:        China  1993 69.216          CHN           le
 5:        China  1994 69.520          CHN           le
 6:        China  1995 70.008          CHN           le
 7:        China  1996 70.266          CHN           le
 8:        China  1997 70.672          CHN           le
 9:        China  1998 71.172          CHN           le
10:        China  1999 71.419          CHN           le
11:        China  2000 71.881          CHN           le
12:        China  2001 72.606          CHN           le
13:        China  2002 72.985          CHN           le
14:        China  2003 73.371          CHN           le
15:        China  2004 73.748          CHN           le
16:        China  2005 74.111          CHN           le
17:        China  2006 74.504          CHN           le
18:        China  2007 74.762          CHN           le
19:        China  2008 74.872          CHN           le
20:        China  2009 75.343          CHN           le
21:        China  2010 75.599          CHN           le
22:        China  2011 75.903          CHN           le
```

4. 3

For **Material footprint per capita (tonnes)**, I want to check data for **recent 24 years.** We can found that Ireland has a overall higher footprint. China 's is increase and Ireland's is increase then decreasing in recent 24 years

```
mf_both <- merged_DT[ indicator_id=="mf" & year>2000,
                      .SD,
                      by=.(country_name,year, value),
                      keyby = TRUE]
mf_both
```

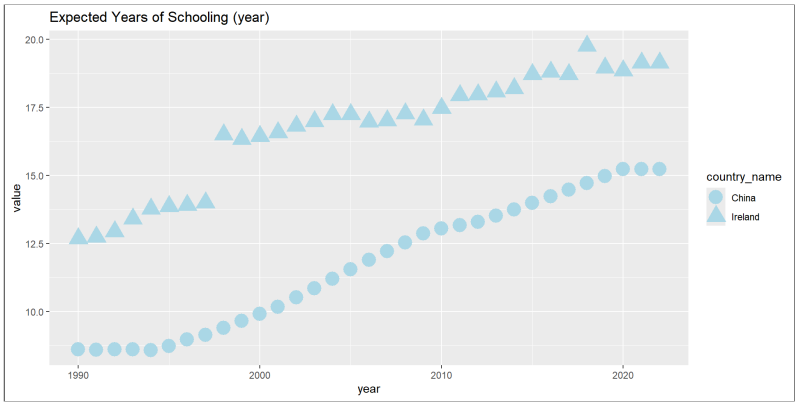
```
Key: <country_name, year, value>
    country_name  year  value country_code indicator_id
          <char> <int>  <num>       <char>       <char>
 1:        China  2001  9.023          CHN           mf
 2:        China  2002  9.510          CHN           mf
 3:        China  2003  9.931          CHN           mf
 4:        China  2004 10.875          CHN           mf
 5:        China  2005 11.555          CHN           mf
 6:        China  2006 12.765          CHN           mf
 7:        China  2007 13.690          CHN           mf
 8:        China  2008 14.400          CHN           mf
 9:        China  2009 16.532          CHN           mf
10:        China  2010 18.432          CHN           mf
11:        China  2011 20.296          CHN           mf
12:        China  2012 21.295          CHN           mf
13:        China  2013 22.860          CHN           mf
14:        China  2014 23.329          CHN           mf
15:        China  2015 22.859          CHN           mf
16:        China  2016 22.683          CHN           mf
17:        China  2017 22.697          CHN           mf
18:        China  2018 22.580          CHN           mf
19:        China  2019 21.944          CHN           mf
20:        China  2020 23.089          CHN           mf
21:        China  2021 23.947          CHN           mf
22:        China  2022 24.283          CHN           mf
```

5. Plotting

plot below comparing the **Expected Years of Schooling (in years)** between Ireland and China. The data shows that Ireland's overall schooling time is higher than China's. Notably, there is a significant jump around 1998 in Ireland. Additionally, both countries have shown an increasing trend in schooling time over the years.

```
#x <- eys_IRL$year
#y1 <- eys_IRL$value
#y2 <- eys_CHN$value
#plot(x,y1,x,y2)
library(hrbrthemes)
ggplot(merged_DT[indicator_id=="eys"], aes(x=year, y=value, shape=country_name)) +
    geom_point(size=6, color="lightblue") +
    ggtitle("Expected Years of Schooling (year)") #+
```

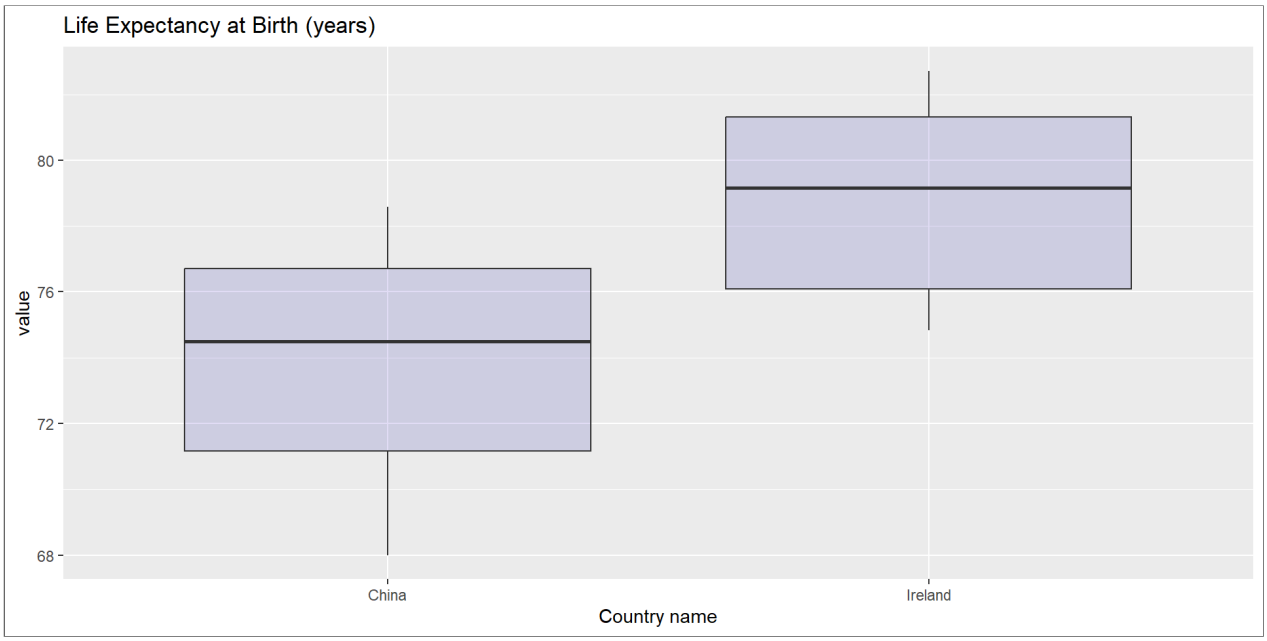
```
    #theme_ipsum()
```



5.1

For **Life Expectancy at Birth (years)** I want to plot the box plot. We found that Ireland has a higher mean and median

```
ggplot(merged_DT[indicator_id=="le"], aes(x=country_name, y=value)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("Country name")+
    ggtitle("Life Expectancy at Birth (years)")
```



5.2

For **Material footprint per capita (tonnes)**, I want to plot the line chat

```
ggplot(merged_DT[indicator_id=="le"& year>2000], aes(x=year, y=value, group=country_name, color=country_name))
    geom_line()+
  ggtitle("Material footprint per capita (tonnes)-recent 24 years")
```