assignment1

Zixuan Gao23206703

2024-06-01

Contents

Dataloading

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

Merge the data datasets using data.table.

read in the data&assign the correct class to the variables.

```r
library(data.table)

hdro_indicators_irl <- fread('hdro_indicators_irl.csv')
hdro_indicators_irl <- hdro_indicators_irl[-1]
hdro_indicators_irl[, year := as.integer(year)]
hdro_indicators_irl[, value := as.numeric(value)]
hdro_indicators_irl[, country_code := as.factor(country_code)]
hdro_indicators_irl[, country_name := as.factor(country_name)]
hdro_indicators_irl[, indicator_id := as.factor(indicator_id)]
hdro_indicators_irl[, indicator_name := as.factor(indicator_name)]
hdro_indicators_irl[, index_id := as.factor(index_id)]
hdro_indicators_irl[, index_name := as.factor(index_name)]
```

read in the data&assign the correct class to the variables.

```r
hdro_indicators_jpn <- fread('hdro_indicators_jpn.csv')
hdro_indicators_jpn <- hdro_indicators_jpn[-1]
hdro_indicators_jpn[, year := as.integer(year)]
hdro_indicators_jpn[, value := as.numeric(value)]
hdro_indicators_jpn[, country_code := as.factor(country_code)]
hdro_indicators_jpn[, country_name := as.factor(country_name)]
hdro_indicators_jpn[, indicator_id := as.factor(indicator_id)]
hdro_indicators_jpn[, indicator_name := as.factor(indicator_name)]
hdro_indicators_jpn[, index_id := as.factor(index_id)]
hdro_indicators_jpn[, index_name := as.factor(index_name)]
```

read in the data&assign the correct class to the variables.

```
hdro_indicators_chn <- fread('hdro_indicators_chn.csv')
hdro_indicators_chn <- hdro_indicators_chn[-1]
hdro_indicators_chn[, year := as.integer(year)]
hdro_indicators_chn[, value := as.numeric(value)]
hdro_indicators_chn[, country_code := as.factor(country_code)]
hdro_indicators_chn[, country_name := as.factor(country_name)]
hdro_indicators_chn[, indicator_id := as.factor(indicator_id)]
hdro_indicators_chn[, indicator_name := as.factor(indicator_name)]
hdro_indicators_chn[, index_id := as.factor(index_id)]
hdro_indicators_chn[, index_name := as.factor(index_name)]
```

Merge the data datasets

```
data <- list(hdro_indicators_irl,
             hdro_indicators_jpn,
             hdro_indicators_chn)
hdro_data <- rbindlist(data)
```

Exploratory Data Analysis(EDA)

part1 quick data exploration

```
library(dplyr)
str(hdro_data) # structure
```

```
Classes 'data.table' and 'data.frame':  2664 obs. of  8 variables:
 $ country_code  : Factor w/ 3 levels "IRL","JPN","CHN": 1 1 1 1 1 1 1 1 1 1 ...
 $ country_name  : Factor w/ 3 levels "Ireland","Japan",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ indicator_id  : Factor w/ 44 levels "abr","co2_prod",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ indicator_name: Factor w/ 44 levels "Adolescent Birth Rate (births per 1,000 women ages 15
 $ index_id      : Factor w/ 6 levels "GDI","GII","HDI",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ index_name    : Factor w/ 6 levels "Gender Development Index",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ value         : num  15.8 16.6 16.5 15.5 14.4 ...
 $ year          : int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
```

```
- attr(*, ".internal.selfref")=<externalptr>
```

From the dataset structure, there are 2 numerical variables and 6 factor variables.

part1 quick data exploration

```r
summary(hdro_data)
```

```
country_code  country_name        indicator_id
IRL:894      Ireland:894   abr           :  99
JPN:894      Japan  :894   co2_prod      :  99
CHN:876      China  :876   diff_hdi_phdi:  99
                           eys           :  99
                           eys_f         :  99
                           eys_m         :  99
                           (Other)       :2070
                                                  indicator_name index_id
Adolescent Birth Rate (births per 1,000 women ages 15-19):  99   GDI :993
Carbon dioxide emissions per capita (production) (tonnes):  99   GII :779
Difference from HDI value (%)                            :  99   HDI :399
Expected Years of Schooling (years)                      :  99   IHDI:183
Expected Years of Schooling, female (years)              :  99   PHDI:300
Expected Years of Schooling, male (years)                :  99   MPI : 10
(Other)                                                  :2070
                                               index_name      value
Gender Development Index                            :993   Min.   :   -22.00
Gender Inequality Index                             :779   1st Qu.:     9.39
Human Development Index                             :399   Median :    17.12
Inequality-adjusted Human Development Index         :183   Mean   :  3500.71
Planetary pressures-adjusted Human Development Index:300   3rd Qu.:    78.16
Multidimensional Poverty Index                      : 10   Max.   :108423.61

     year
Min.   :1990
1st Qu.:1999
Median :2007
Mean   :2007
3rd Qu.:2015
Max.   :2022
```

Explanation

- The `value` field ranges widely from -22 (possibly indicating some form of deficit or decrease) to 108,423.61, reflecting substantial variation possibly due to different types of indicators included (such as monetary values, rates, or counts).
- Data collection spans from 1990 to 2022, allowing for longitudinal studies and trend analysis over a significant period.

part1 quick data exploration

```r
# Calculate the number of data rows for each combination
# of country and indicator using data.table syntax
result <- hdro_data[, .N, by = .(country_name,
                                 indicator_name)]
result[order(-N)]
```

```
     country_name                                           indicator_name
           <fctr>                                                   <fctr>
  1:      Ireland Adolescent Birth Rate (births per 1,000 women ages 15-19)
  2:      Ireland Carbon dioxide emissions per capita (production) (tonnes)
  3:      Ireland                            Difference from HDI value (%)
  4:      Ireland                      Expected Years of Schooling (years)
  5:      Ireland              Expected Years of Schooling, female (years)
 ---                                                                      
108:        China                                            Nutrition (%)
109:        China                                 Difference from HDI rank
110:        China                                           Sanitation (%)
111:        China                                    School attendance (%)
112:        China                                   Years of schooling (%)
         N
     <int>
  1:    33
  2:    33
  3:    33
  4:    33
  5:    33
 ---      
108:     1
```

```
109:     1
110:     1
111:     1
112:     1
```

```
# Use dplyr syntax to group data by country name and
# calculate the mean value and count of records for each group
hdro_data %>%
  group_by(country_name) %>%
  summarise(mean_value = mean(value), n = n())
```

```
# A tibble: 3 x 3
  country_name mean_value     n
  <fct>             <dbl> <int>
1 Ireland           5234.   894
2 Japan             4331.   894
3 China              884.   876
```

Explanation

Above is the number of data rows for each combination of country and indicator using data.table syntax and then use dplyr syntax to group data by country name and calculate the mean value and count of records for each group. The mean value of Ireland, Japan, China are 5234.4765, 4331.2698, 883.6944.

part2 More data exploration analysis

```
setDT(hdro_data)

# Calculate the average Human Development Index (HDI)
# for each country and year
average_HDI_by_country <- hdro_data[index_name ==
                                      "Human Development Index",
                                    .(average_HDI = mean(value,
                                                         na.rm = TRUE)),
                                    keyby = .(country_name, year)]

# Identify the latest year of data for each country
```

```r
latest_year_idx <- average_HDI_by_country[, .I[year == max(year)],
                                            by = country_name]$V1

latest_HDI_by_country <- average_HDI_by_country[latest_year_idx][
  order(-average_HDI)]

average_HDI_by_country
```

```
Key: <country_name, year>
    country_name  year average_HDI
          <fctr> <int>       <num>
 1:      Ireland  1990   6175.0917
 2:      Ireland  1991   6284.7792
 3:      Ireland  1992   6369.4932
 4:      Ireland  1993   6537.3030
 5:      Ireland  1994   6917.5000
 6:      Ireland  1995   7490.5770
 7:      Ireland  1996   8023.0878
 8:      Ireland  1997   8691.4825
 9:      Ireland  1998   9290.3662
10:      Ireland  1999   9900.1390
11:      Ireland  2000  10669.5450
12:      Ireland  2001  10826.5572
13:      Ireland  2002  11074.6342
14:      Ireland  2003  11537.6292
15:      Ireland  2004  12169.5342
16:      Ireland  2005  12622.5552
17:      Ireland  2006  13055.0710
18:      Ireland  2007  13189.1900
19:      Ireland  2008  12354.0122
20:      Ireland  2009  11220.5382
21:      Ireland  2010  11426.1793
22:      Ireland  2011  11134.8978
23:      Ireland  2012  10994.0742
24:      Ireland  2013  11515.0302
25:      Ireland  2014  12443.3267
26:      Ireland  2015  13787.0850
27:      Ireland  2016  14735.1433
28:      Ireland  2017  15529.1108
29:      Ireland  2018  16269.8488
30:      Ireland  2019  16941.0940
31:      Ireland  2020  17305.4448
```

```
32:      Ireland  2021  19766.3380
33:      Ireland  2022  17517.6026
34:        Japan  1990   8299.8805
35:        Japan  1991   8559.9460
36:        Japan  1992   8617.7405
37:        Japan  1993   8552.8302
38:        Japan  1994   8613.9682
39:        Japan  1995   8820.8795
40:        Japan  1996   9109.4390
41:        Japan  1997   9186.5217
42:        Japan  1998   9038.8560
43:        Japan  1999   8991.5300
44:        Japan  2000   9244.2620
45:        Japan  2001   9267.9825
46:        Japan  2002   9241.3002
47:        Japan  2003   9377.2375
48:        Japan  2004   9608.4237
49:        Japan  2005   9807.1517
50:        Japan  2006   9978.0212
51:        Japan  2007  10151.6770
52:        Japan  2008   9981.5175
53:        Japan  2009   9403.8337
54:        Japan  2010   9798.8288
55:        Japan  2011   9839.3405
56:        Japan  2012   9975.6358
57:        Japan  2013  10254.9517
58:        Japan  2014  10325.4860
59:        Japan  2015  10524.8742
60:        Japan  2016  10566.3477
61:        Japan  2017  10772.4095
62:        Japan  2018  10865.3865
63:        Japan  2019  10847.2157
64:        Japan  2020  10389.6390
65:        Japan  2021  10780.2738
66:        Japan  2022   8756.1618
67:        China  1990    376.5030
68:        China  1991    403.7172
69:        China  1992    452.9095
70:        China  1993    506.9002
71:        China  1994    565.3340
72:        China  1995    611.2890
73:        China  1996    663.9145
74:        China  1997    717.9883
```

```
75:        China  1998    762.6985
76:        China  1999    814.5865
77:        China  2000    876.0490
78:        China  2001    938.7942
79:        China  2002   1019.4932
80:        China  2003   1116.7580
81:        China  2004   1224.2115
82:        China  2005   1347.8358
83:        China  2006   1515.4725
84:        China  2007   1725.9025
85:        China  2008   1887.9455
86:        China  2009   2037.7178
87:        China  2010   2235.4358
88:        China  2011   2421.4490
89:        China  2012   2610.8107
90:        China  2013   2777.3675
91:        China  2014   2991.0213
92:        China  2015   3162.6370
93:        China  2016   3357.8415
94:        China  2017   3581.0648
95:        China  2018   3792.0033
96:        China  2019   4008.6277
97:        China  2020   4067.3083
98:        China  2021   4400.1743
99:        China  2022   3640.3598
    country_name  year average_HDI
```

```
latest_HDI_by_country
```

```
Key: <country_name, year>
   country_name  year average_HDI
         <fctr> <int>       <num>
1:      Ireland  2022   17517.603
2:        Japan  2022    8756.162
3:        China  2022    3640.360
```

Explanation

Above is the average Human Development Index (HDI) for each country and year and the latest year of data for each country.

part2 More data exploration analysis

```
gender_inequality_index <- hdro_data[index_name ==
                                       "Gender Inequality Index",
                                     .(mean_value = mean(
                                       value, na.rm = TRUE)),
                                     keyby = .(country_name, year)]


gender_inequality_index[, prev_value := shift(mean_value),
                        by = country_name]

gender_inequality_index[, change := mean_value - prev_value]
print(gender_inequality_index)
```

```
Key: <country_name, year>
    country_name  year mean_value prev_value     change
          <fctr> <int>      <num>      <num>      <num>
 1:      Ireland  1990   43.91563         NA         NA
 2:      Ireland  1991   44.42612   43.91563  0.5105000
 3:      Ireland  1992   44.70575   44.42612  0.2796250
 4:      Ireland  1993   44.93212   44.70575  0.2263750
 5:      Ireland  1994   45.37050   44.93212  0.4383750
 6:      Ireland  1995   45.82850   45.37050  0.4580000
 7:      Ireland  1996   46.37788   45.82850  0.5493750
 8:      Ireland  1997   47.21425   46.37788  0.8363750
 9:      Ireland  1998   47.93337   47.21425  0.7191250
10:      Ireland  1999   48.56712   47.93337  0.6337500
11:      Ireland  2000   49.36538   48.56712  0.7982500
12:      Ireland  2001   49.69513   49.36538  0.3297500
13:      Ireland  2002   49.71637   49.69513  0.0212500
14:      Ireland  2003   50.26725   49.71637  0.5508750
15:      Ireland  2004   50.67225   50.26725  0.4050000
16:      Ireland  2005   51.51250   50.67225  0.8402500
17:      Ireland  2006   51.76313   51.51250  0.2506250
18:      Ireland  2007   52.63113   51.76313  0.8680000
19:      Ireland  2008   52.42200   52.63113 -0.2091250
20:      Ireland  2009   51.76387   52.42200 -0.6581250
21:      Ireland  2010   51.09350   51.76387 -0.6703750
22:      Ireland  2011   51.35788   51.09350  0.2643750
23:      Ireland  2012   51.21125   51.35788 -0.1466250
```

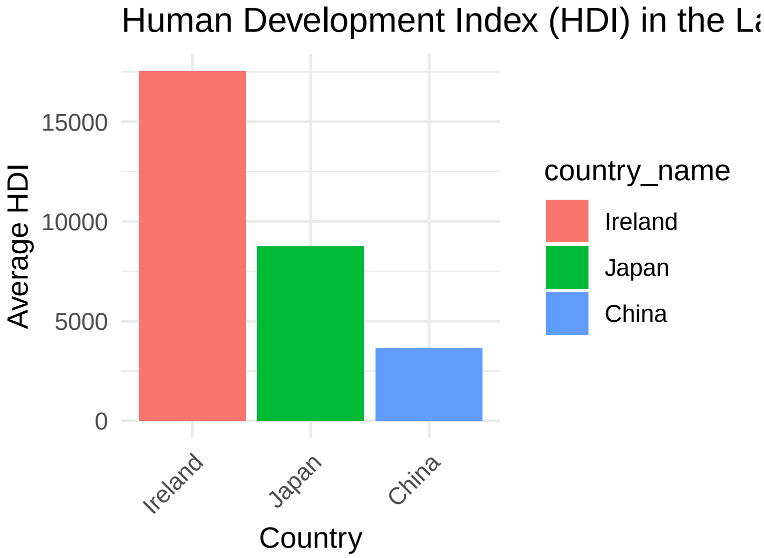
```
24:      Ireland  2013   51.12050   51.21125 -0.0907500
25:      Ireland  2014   51.01650   51.12050 -0.1040000
26:      Ireland  2015   51.09887   51.01650  0.0823750
27:      Ireland  2016   51.06700   51.09887 -0.0318750
28:      Ireland  2017   51.07975   51.06700  0.0127500
29:      Ireland  2018   51.19013   51.07975  0.1103750
30:      Ireland  2019   51.28800   51.19013  0.0978750
31:      Ireland  2020   51.04050   51.28800 -0.2475000
32:      Ireland  2021   51.59225   51.04050  0.5517500
33:      Ireland  2022   48.41878   51.59225 -3.1734722
34:        Japan  1990   47.60575         NA         NA
35:        Japan  1991   48.04800   47.60575  0.4422500
36:        Japan  1992   48.45225   48.04800  0.4042500
37:        Japan  1993   48.72575   48.45225  0.2735000
38:        Japan  1994   48.92912   48.72575  0.2033750
39:        Japan  1995   49.36325   48.92912  0.4341250
40:        Japan  1996   49.52600   49.36325  0.1627500
41:        Japan  1997   49.86725   49.52600  0.3412500
42:        Japan  1998   50.06625   49.86725  0.1990000
43:        Japan  1999   50.21063   50.06625  0.1443750
44:        Japan  2000   50.37300   50.21063  0.1623750
45:        Japan  2001   50.46462   50.37300  0.0916250
46:        Japan  2002   50.45263   50.46462 -0.0120000
47:        Japan  2003   50.49225   50.45263  0.0396250
48:        Japan  2004   50.56387   50.49225  0.0716250
49:        Japan  2005   50.71300   50.56387  0.1491250
50:        Japan  2006   50.83700   50.71300  0.1240000
51:        Japan  2007   50.97187   50.83700  0.1348750
52:        Japan  2008   51.11175   50.97187  0.1398750
53:        Japan  2009   51.20237   51.11175  0.0906250
54:        Japan  2010   51.27213   51.20237  0.0697500
55:        Japan  2011   51.50075   51.27213  0.2286250
56:        Japan  2012   51.76112   51.50075  0.2603750
57:        Japan  2013   52.15088   51.76112  0.3897500
58:        Japan  2014   52.54125   52.15088  0.3903750
59:        Japan  2015   52.83250   52.54125  0.2912500
60:        Japan  2016   53.00037   52.83250  0.1678750
61:        Japan  2017   53.15713   53.00037  0.1567500
62:        Japan  2018   53.46475   53.15713  0.3076250
63:        Japan  2019   53.53925   53.46475  0.0745000
64:        Japan  2020   53.62438   53.53925  0.0851250
65:        Japan  2021   53.66887   53.62438  0.0445000
66:        Japan  2022   50.22233   53.66887 -3.4465417
```

```
67:        China 1990   57.86683         NA         NA
68:        China 1991   54.43017   57.86683 -3.4366667
69:        China 1992   54.70650   54.43017  0.2763333
70:        China 1993   52.82500   54.70650 -1.8815000
71:        China 1994   52.49450   52.82500 -0.3305000
72:        China 1995   52.62933   52.49450  0.1348333
73:        China 1996   51.36200   52.62933 -1.2673333
74:        China 1997   51.29050   51.36200 -0.0715000
75:        China 1998   51.10675   51.29050 -0.1837500
76:        China 1999   51.82850   51.10675  0.7217500
77:        China 2000   53.53887   51.82850  1.7103750
78:        China 2001   51.43250   53.53887 -2.1063750
79:        China 2002   51.51900   51.43250  0.0865000
80:        China 2003   50.40212   51.51900 -1.1168750
81:        China 2004   51.40925   50.40212  1.0071250
82:        China 2005   51.45062   51.40925  0.0413750
83:        China 2006   51.44950   51.45062 -0.0011250
84:        China 2007   51.56600   51.44950  0.1165000
85:        China 2008   51.89600   51.56600  0.3300000
86:        China 2009   51.85800   51.89600 -0.0380000
87:        China 2010   52.69138   51.85800  0.8333750
88:        China 2011   51.97513   52.69138 -0.7162500
89:        China 2012   52.17012   51.97513  0.1950000
90:        China 2013   52.67638   52.17012  0.5062500
91:        China 2014   52.95038   52.67638  0.2740000
92:        China 2015   53.45613   52.95038  0.5057500
93:        China 2016   53.39462   53.45613 -0.0615000
94:        China 2017   53.37588   53.39462 -0.0187500
95:        China 2018   53.33450   53.37588 -0.0413750
96:        China 2019   53.59150   53.33450  0.2570000
97:        China 2020   53.50700   53.59150 -0.0845000
98:        China 2021   53.55800   53.50700  0.0510000
99:        China 2022   52.84022   53.55800 -0.7177778
    country_name year mean_value prev_value     change
```

part3 Plot of HDI

```
library(ggplot2)

# barplot
ggplot(latest_HDI_by_country, aes(x = reorder(country_name, -average_HDI), y = average_HDI
  geom_bar(stat = "identity") +
  labs(title = "Human Development Index (HDI) in the Latest Recorded Year by Country",
       x = "Country",
       y = "Average HDI") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



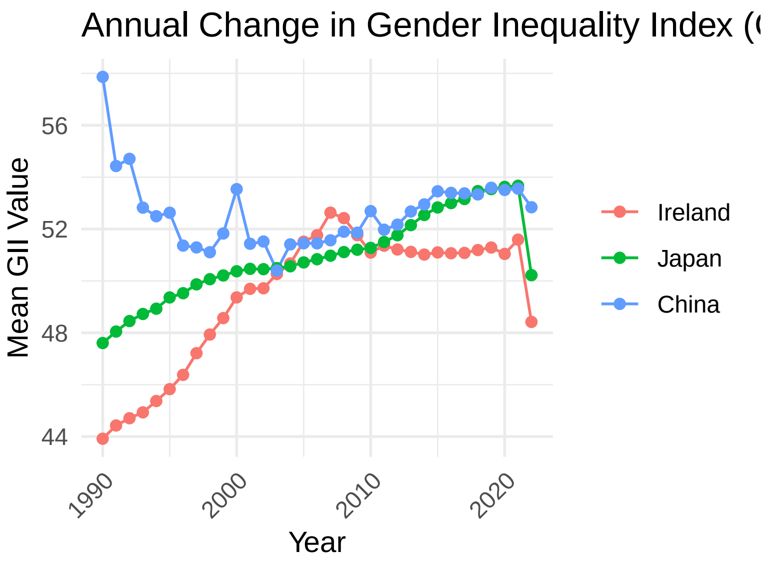
Explanation

The bar graph titled “Human Development Index (HDI) in the Latest Recorded Year by Country” compares the average HDI for Ireland, Japan, and China. It shows Ireland with an anomalously high HDI around 15,000, followed by Japan at about 7,500, and China at around 5,000. These values are unusually high for HDI, which typically ranges between 0 and 1, suggesting a potential error in data scaling or representation. The graph uses distinct colors for each country, facilitating easy visual comparison, but caution is advised in interpreting these results due to the likely data error.

part3 Plot of GII

```
library(ggplot2)

# Plot a line graph to display the changes in the Gender Inequality Index (GII) for the th
ggplot(gender_inequality_index, aes(x = year, y = mean_value, color = country_name, group
  geom_line() +
  geom_point() +
  labs(title = "Annual Change in Gender Inequality Index (GII) for JPN, CHN, and IRL",
       x = "Year",
       y = "Mean GII Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.title = element_blank())
```



Explanation

The line graph titled “Annual Change in Gender Inequality Index (GII) for JPN, CHN, and IRL” shows the trends in GII values for Japan, China, and Ireland from 1990 to around 2020. Japan exhibits a generally upward trend with some volatility, indicating a slow increase in gender inequality over the years. China’s GII also shows a steady increase, suggesting worsening gender inequality. In contrast, Ireland’s GII initially increases but shows a dramatic drop around 2020, suggesting a significant improvement in gender equality in that year.