assignment1

Zixuan Gao23206703

2024-06-19

Data loading

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.

```
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_coc
hdro_indicators_irl[, country_name := as.factor(country_nar
hdro_indicators_irl[, indicator_id := as.factor(indicator_i
hdro_indicators_irl[, indicator_name := as.factor(indicator
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.

```
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_co
hdro_indicators_jpn[, country_name := as.factor(country_na
hdro_indicators_jpn[, indicator_id := as.factor(indicator_
hdro_indicators_jpn[, indicator_name := as.factor(indicato
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_coc
hdro_indicators_chn[, country_name := as.factor(country_nam
hdro_indicators_chn[, indicator_id := as.factor(indicator_i
hdro_indicators_chn[, indicator_name := as.factor(indicator
hdro_indicators_chn[, index_id := as.factor(index_id)]
hdro_indicators_chn[, index_name := as.factor(index_name)]
```

Merge the data datasets

```r
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 var
 $ country_code  : Factor w/ 3 levels "IRL","JPN","CHN": 1
 $ country_name  : Factor w/ 3 levels "Ireland","Japan",..:
 $ indicator_id  : Factor w/ 44 levels "abr","co2_prod",..:
 $ indicator_name: Factor w/ 44 levels "Adolescent Birth Ra
 $ index_id      : Factor w/ 6 levels "GDI","GII","HDI",..:
 $ index_name    : Factor w/ 6 levels "Gender Development I
 $ value         : num  15.8 16.6 16.5 15.5 14.4 ...
 $ year          : int  1990 1991 1992 1993 1994 1995 1996
 - 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
                                                         indicato
 Adolescent Birth Rate (births per 1,000 women ages 15-19):
 Carbon dioxide emissions per capita (production) (tonnes):
 Difference from HDI value (%)                            :
 Expected Years of Schooling (years)                      :
 Expected Years of Schooling, female (years)              :
 Expected Years of Schooling, male (years)
```

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

```
# 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
           <fctr>
  1:      Ireland Adolescent Birth Rate (births per 1,000 w
  2:      Ireland Carbon dioxide emissions per capita (prod
  3:      Ireland                                  Difference fr
  4:      Ireland                            Expected Years of S
  5:      Ireland                Expected Years of Schooling
 ---
108:        China
109:        China                                         Differen
110:        China
```

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

```r
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 In
                                    .(average_HDI = mean(va
                                                         na
                                    keyby = .(country_name,

# Identify the latest year of data for each country
latest_year_idx <- average_HDI_by_country[, .I[year == max(
                                          by = country_name

latest_HDI_by_country <- average_HDI_by_country[latest_year
  order(-average_HDI)]
```

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 I
                                     .(mean_value = mean(
                                       value, na.rm = TRUE)
                                     keyby = .(country_name


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
```

part3 Plot of HDI

```r
library(ggplot2)

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

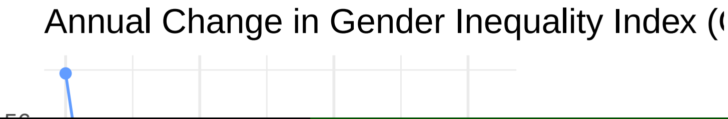
Explanation

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 I
ggplot(gender_inequality_index, aes(x = year, y = mean_valu
  geom_line() +
  geom_point() +
  labs(title = "Annual Change in Gender Inequality Index (
       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.