Assignment 1

Ning Fan

2024-06-19

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

```
chn <- fread("hdro_indicators_chn.csv")
irl <- fread("hdro_indicators_irl.csv")
#show the data for first 2 rows with some selected columns
chn[1:2,c(1,7,8)]
```

```
    country_code                  value       year
          <char>                 <char>     <char>
1: #country+code #indicator+value+num #date+year
2:           CHN                 30.986       1990
```

As we see from the output, first row needs to be removed.
Variables *value* and *year* should be converted to numeric class.

```
#remove the first row in both tables
chn <- chn[-1,]
irl <- irl[-1,]

#all variables have class character.
#convert variable value and year to numeric
chn[, value:=as.numeric(value)]
chn[, year:=as.numeric(year)]
irl[, value:=as.numeric(value)]
irl[, year:=as.numeric(year)]
```

	country_code	index_id	value	year
	<char>	<char>	<num>	<num>
1:	CHN	GII	30.986	1990
2:	CHN	GII	27.892	1991
3:	CHN	GII	22.499	1992
4:	CHN	GII	21.184	1993
5:	CHN	GII	15.419	1994

872:	CHN	GII	85.354	2019
873:	CHN	GII	86.366	2020
874:	CHN	GII	86.366	2021
875:	CHN	GII	86.366	2022
876:	CHN	MPI	20.663	2014

Take China data for example, the variable class has been converted succesfully and first row is removed.

2. Merge the data datasets using data.table

```
#combine the data together as they share same column names
dt <- merge(chn,irl,all = TRUE)
#only include selected columns for neater presentation
print(dt[,c("country_code","index_id","value","year")]
      ,topn = 2)
```

```
Key: <country_code>
      country_code index_id  value  year
             <char>   <char>  <num> <num>
   1:          CHN      GII 11.048  2021
   2:          CHN      GII 11.146  2022
  ---
1769:          IRL      GII 86.417  2021
1770:          IRL      GII 86.417  2022
```

3. Do some quick data exploration to know more about your data.

```
max(dt[country_code=="CHN","year"])
```

```
[1] 2022
```

```
min(dt[country_code=="CHN","year"])
```

```
[1] 1990
```

```
max(dt[country_code=="IRL","year"])
```

```
[1] 2022
```

```
min(dt[country_code=="IRL","year"])
```

```
[1] 1990
```

Data recorded in both dataset are ranged between 1990-2022

How many different indicators for each indexX?

```
#filter out the unique rows of indicator
index_indicator <- unique(dt[,
                c("index_id","indicator_id")])
#calculate the number of indicators, by index
index_indicator[,.N, keyby=index_id]
```

```
Key: <index_id>
   index_id     N
     <char> <int>
1:      GDI    11
2:      GII     9
3:      HDI     5
4:     IHDI     5
5:      MPI    10
6:     PHDI     4
```

- Index GDI (Gender Development Index) has the most number of indicators recorded (11)
- Index PHDI (Planetary pressures–adjusted Human Development Index) has least indicators recorded (4)

4.1 Gender Inequality Index

```
dt[(indicator_id=="abr"|indicator_id=="lfpr_f") &
     (year=="2022"|year=="2000"),.(country_code, value)
   , keyby = .(indicator_id, year)]
```

```
Key: <indicator_id, year>
   indicator_id  year country_code  value
          <char> <num>        <char>  <num>
1:          abr  2000           CHN 12.579
2:          abr  2000           IRL 19.615
3:          abr  2022           CHN 11.146
4:          abr  2022           IRL  5.872
5:       lfpr_f  2000           CHN 70.570
6:       lfpr_f  2000           IRL 47.150
7:       lfpr_f  2022           CHN 53.760
8:       lfpr_f  2022           IRL 59.400
```

```
dt[(indicator_id=="pr_f"|indicator_id=="se_f"
    | indicator_id=="gii_rank") &
     (year=="2022"|year=="2000"),.(country_code, value)
   , keyby = .(indicator_id, year)]
```

```
Key: <indicator_id, year>
    indicator_id  year country_code  value
           <char> <num>       <char>  <num>
 1:     gii_rank  2022          CHN 47.000
 2:     gii_rank  2022          IRL 20.000
 3:         pr_f  2000          CHN 21.783
 4:         pr_f  2000          IRL 13.717
 5:         pr_f  2022          CHN 24.941
 6:         pr_f  2022          IRL 27.397
 7:         se_f  2000          CHN 43.377
 8:         se_f  2000          IRL 73.356
 9:         se_f  2022          CHN 79.702
10:         se_f  2022          IRL 88.586
```

For Gender Inequality Index, in 2022 CHN is ranking after IRL. For the selected 4 indicators, IRL outperformed on all in year 2022. However, back to year 2000, CHN had a better figure on *Adolescent Birth Rate*, *Labour force participation rate (female)*, *Share of seats in parliament (female)*.

4.2 Planetary pressures–adjusted Human Development Index

```
dt[(indicator_id=="co2_prod"|indicator_id=="mf") &
     (year=="2022"|year=="1990"),.(country_code, value)
   , keyby = .(indicator_id,year)]
```

```
Key: <indicator_id, year>
   indicator_id  year country_code  value
          <char> <num>        <char>  <num>
1:     co2_prod  1990           CHN  2.154
2:     co2_prod  1990           IRL  9.452
3:     co2_prod  2022           CHN  7.950
4:     co2_prod  2022           IRL  7.530
5:           mf  1990           CHN  5.229
6:           mf  1990           IRL 22.145
7:           mf  2022           CHN 24.283
8:           mf  2022           IRL 26.347
```

For Planetary pressures–adjusted Human Development Index, CHN has much lower *Material footprint (per capita)* and *Carbon dioxide emissions (per capita)* in year 1990. However, this number increased to almost same level as IRL in 2022, while CHN is higher than IRL on *Carbon dioxide emissions (per capita)*.

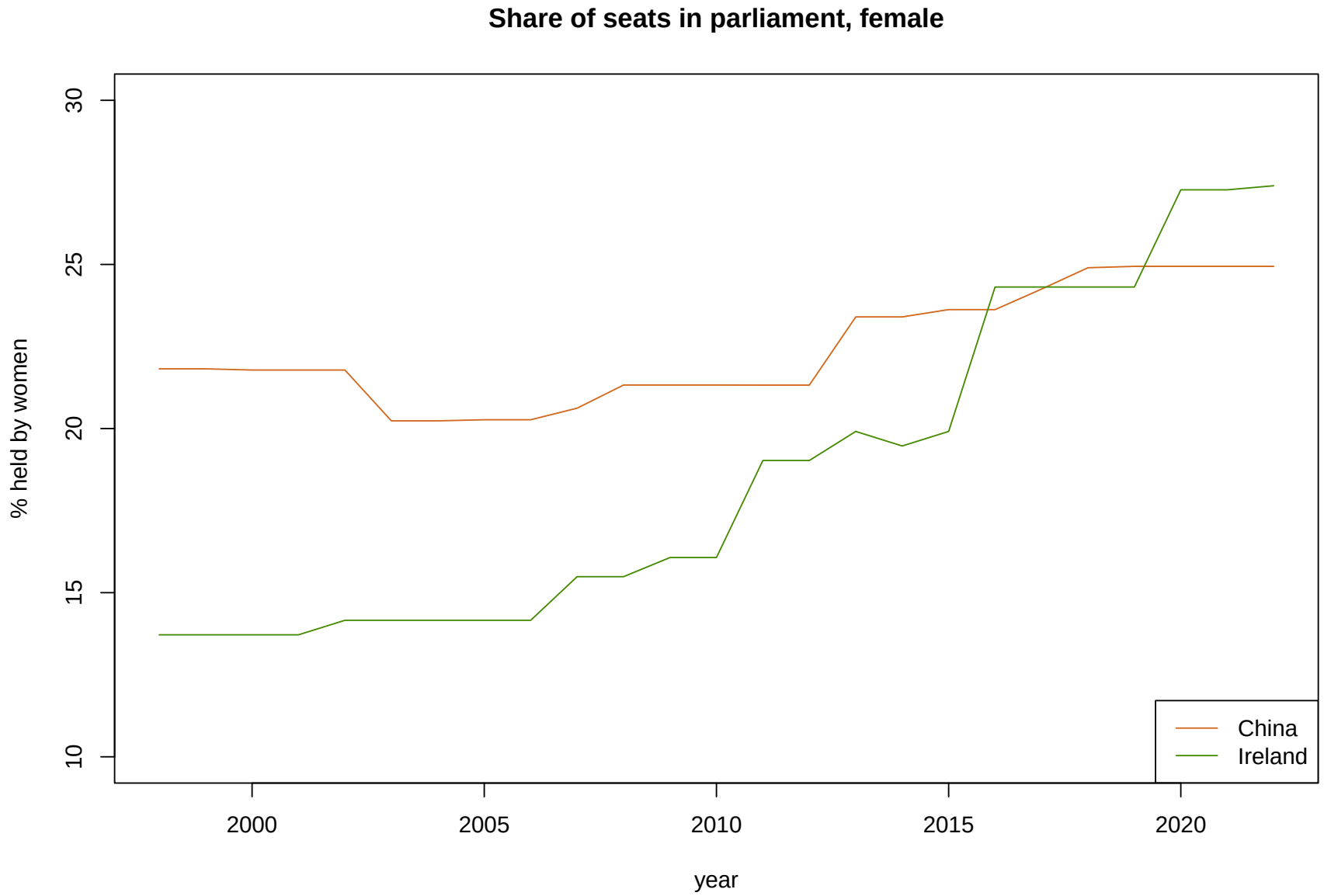
5. Do at least 2 plots using some output from the analysis done in step 4.

I'd like to examine the trend of two indicators for each country.

- Share of seats in parliament, female (% held by women)
- Carbon dioxide emissions per capita (production) (tonnes)

- Share of seats in parliament, female (% held by women)

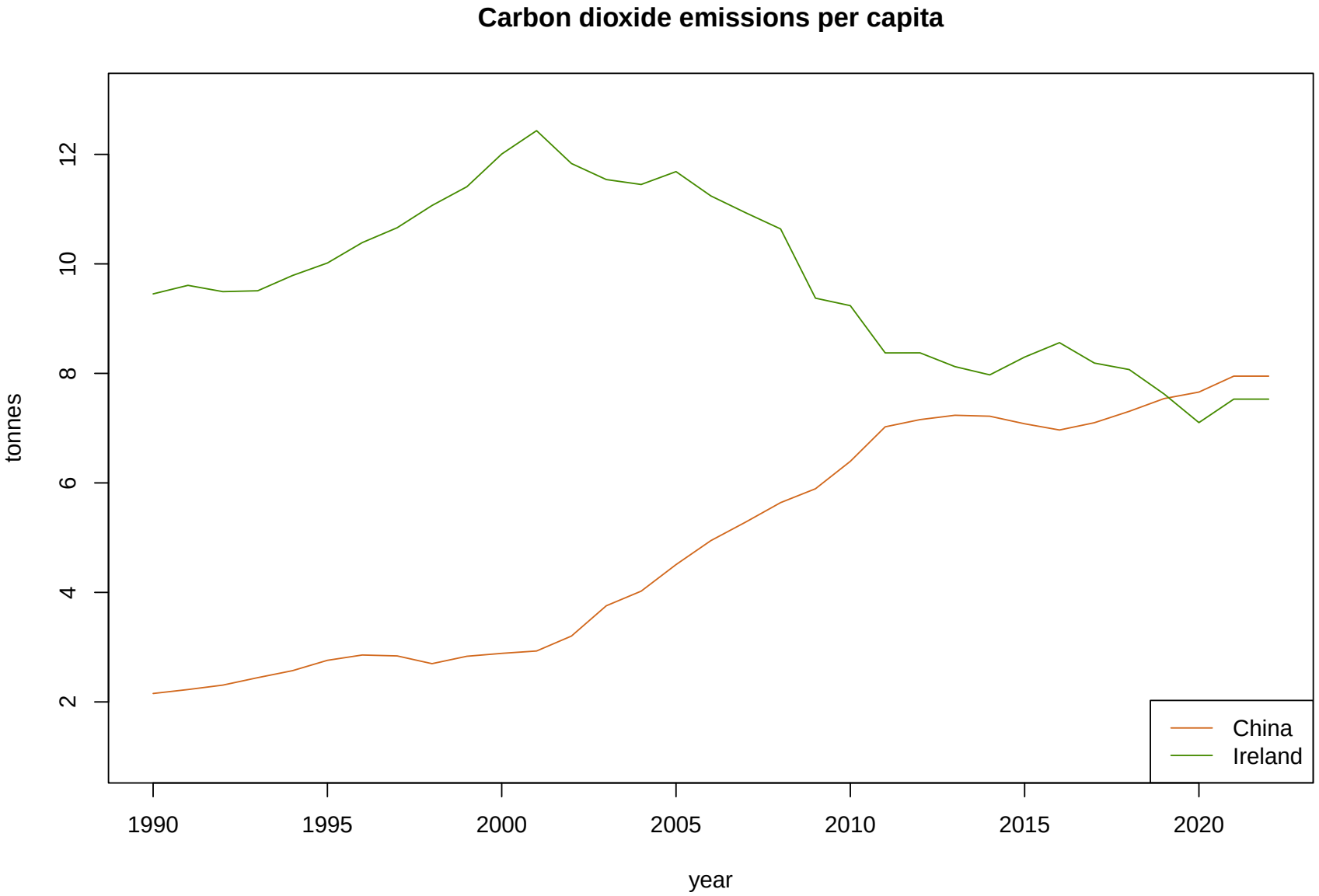
```
#get the data for each country
dt_chn_plot1 <- dt[indicator_id=="pr_f"
      & country_code=="CHN",.(value), keyby = .(year)]
dt_irl_plot1 <-dt[indicator_id=="pr_f" & year>1997
      & country_code=="IRL",.(value), keyby = .(year)]
#select year 1998-2022 as data for CHN is N/A before 1998
x<-c(1998:2022)
#plot
plot(x,dt_chn_plot1$value, type = "l", col="chocolate",
     ylim =c(10,30),ylab = "% held by women", xlab ="year")
lines(x,dt_irl_plot1$value, type = "l", col="chartreuse4")
#add legend and title
legend(x="bottomright", legend = c("China","Ireland")
       , col = c("chocolate","chartreuse4"),lty = c(1,1))
title("Share of seats in parliament, female")
```


Share of seats in parliament, female
% held by women
10
15
20
25
30
2000
2005
2010
2015
2020
year
China
Ireland

There's a gradual increase in female share of seats in parliament for Ireland during the 25 years. China only shows a slight increase in the female shares over the time period. Ireland performed poorly in the early time however it overtook China in year 2016 and after year 2020.

- Carbon dioxide emissions per capita (production) (tonnes)

```
#get the data for each country
dt_chn_plot2 <- dt[indicator_id=="co2_prod"
       & country_code=="CHN",.(value), keyby = .(year)]
dt_irl_plot2 <-dt[indicator_id=="co2_prod"
       & country_code=="IRL",.(value), keyby = .(year)]
x <- c(1990:2022)
#plot
plot(x,dt_chn_plot2$value, type = "l",col="chocolate",
     ylim =c(1,13),ylab = "tonnes", xlab ="year")
lines(x,dt_irl_plot2$value, type = "l", col="chartreuse4")
#add legend and title
legend(x="bottomright", legend = c("China","Ireland")
       , col = c("chocolate","chartreuse4"),lty = c(1,1))
title("Carbon dioxide emissions per capita")
```


Carbon dioxide emissions per capita
tonnes
2
4
6
8
10
12
1990
1995
2000
2005
2010
2015
2020
year
China
Ireland

China had a very low level of carbon dioxide emissions in 1990, while the level for Ireland is almost 4 times higher. During the decades, CO2 emissions for China kept increasing, while for Ireland it reached the peak in year 2000 then shows a significant decrease. In year 2022, Ireland has lower carbon dioxide emissions (per capita) than China.