# Assignment1

1. ***Step 1: Load the necessary library and read in the data***

```r
# Load the data.table package
library(data.table)

# Read the CSV files into data tables
nga_data <- fread("hdro_indicators_nga.csv")
irl_data <- fread("hdro_indicators_irl.csv")

# Delete the first row
nga_data <- nga_data[-1,]
irl_data <- irl_data[-1,]

# Display the first few rows of each dataset to understand their structure
head(nga_data)
```

```
   country_code country_name indicator_id
         <char>       <char>       <char>
1:          NGA      Nigeria          abr
2:          NGA      Nigeria          abr
3:          NGA      Nigeria          abr
4:          NGA      Nigeria          abr
5:          NGA      Nigeria          abr
6:          NGA      Nigeria          abr
                                                indicator_name index_id
                                                        <char>   <char>
1: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
2: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
3: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
4: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
5: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
6: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
```

```
                index_name   value   year
                    <char>  <char> <char>
1: Gender Inequality Index 140.866   1990
2: Gender Inequality Index 138.452   1991
3: Gender Inequality Index 138.666   1992
4: Gender Inequality Index 140.754   1993
5: Gender Inequality Index 140.344   1994
6: Gender Inequality Index  136.68   1995
```

```
head(irl_data)
```

```
   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
6:          IRL      Ireland          abr
                                              indicator_name index_id
                                                      <char>   <char>
1: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
2: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
3: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
4: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
5: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
6: Adolescent Birth Rate (births per 1,000 women ages 15-19)      GII
                index_name  value   year
                    <char> <char> <char>
1: Gender Inequality Index 15.814   1990
2: Gender Inequality Index 16.584   1991
3: Gender Inequality Index 16.457   1992
4: Gender Inequality Index 15.454   1993
5: Gender Inequality Index 14.375   1994
6: Gender Inequality Index 14.338   1995
```

***Step 2: Assign the correct class to the variables***

```
# Check the structure of the data
str(nga_data)
```

```
Classes 'data.table' and 'data.frame':  777 obs. of  8 variables:
```

```
 $ country_code  : chr  "NGA" "NGA" "NGA" "NGA" ...
 $ country_name  : chr  "Nigeria" "Nigeria" "Nigeria" "Nigeria" ...
 $ indicator_id  : chr  "abr" "abr" "abr" "abr" ...
 $ indicator_name: chr  "Adolescent Birth Rate (births per 1,000 women ages 15-19)" "Adolesc
 $ index_id      : chr  "GII" "GII" "GII" "GII" ...
 $ index_name    : chr  "Gender Inequality Index" "Gender Inequality Index" "Gender Inequali
 $ value         : chr  "140.866" "138.452" "138.666" "140.754" ...
 $ year          : chr  "1990" "1991" "1992" "1993" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```
str(irl_data)
```

```
Classes 'data.table' and 'data.frame':  894 obs. of  8 variables:
 $ country_code  : chr  "IRL" "IRL" "IRL" "IRL" ...
 $ country_name  : chr  "Ireland" "Ireland" "Ireland" "Ireland" ...
 $ indicator_id  : chr  "abr" "abr" "abr" "abr" ...
 $ indicator_name: chr  "Adolescent Birth Rate (births per 1,000 women ages 15-19)" "Adolesc
 $ index_id      : chr  "GII" "GII" "GII" "GII" ...
 $ index_name    : chr  "Gender Inequality Index" "Gender Inequality Index" "Gender Inequali
 $ value         : chr  "15.814" "16.584" "16.457" "15.454" ...
 $ year          : chr  "1990" "1991" "1992" "1993" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

From above we know that we should change the class of "value" and "year"

```
#'year' should be integer and 'value' should be numeric:
nga_data[, year := as.integer(year)]
nga_data[, value := as.numeric(value)]

irl_data[, year := as.integer(year)]
irl_data[, value := as.numeric(value)]

# Verify the changes
str(nga_data)
```

```
Classes 'data.table' and 'data.frame':  777 obs. of  8 variables:
 $ country_code  : chr  "NGA" "NGA" "NGA" "NGA" ...
 $ country_name  : chr  "Nigeria" "Nigeria" "Nigeria" "Nigeria" ...
 $ indicator_id  : chr  "abr" "abr" "abr" "abr" ...
 $ indicator_name: chr  "Adolescent Birth Rate (births per 1,000 women ages 15-19)" "Adolesc
```

```
 $ index_id      : chr  "GII" "GII" "GII" "GII" ...
 $ index_name    : chr  "Gender Inequality Index" "Gender Inequality Index" "Gender Inequalit
 $ value         : num  141 138 139 141 140 ...
 $ year          : int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```r
str(irl_data)
```

```
Classes 'data.table' and 'data.frame':  894 obs. of  8 variables:
 $ country_code  : chr  "IRL" "IRL" "IRL" "IRL" ...
 $ country_name  : chr  "Ireland" "Ireland" "Ireland" "Ireland" ...
 $ indicator_id  : chr  "abr" "abr" "abr" "abr" ...
 $ indicator_name: chr  "Adolescent Birth Rate (births per 1,000 women ages 15-19)" "Adolesc
 $ index_id      : chr  "GII" "GII" "GII" "GII" ...
 $ index_name    : chr  "Gender Inequality Index" "Gender Inequality Index" "Gender Inequalit
 $ 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>
```

2. ***Merge the data datasets using `data.table`.***

```r
# Merge the datasets using rbind
merged_data <- rbind(nga_data, irl_data)

# Check the structure of merged_data
str(merged_data)
```

```
Classes 'data.table' and 'data.frame':  1671 obs. of  8 variables:
 $ country_code  : chr  "NGA" "NGA" "NGA" "NGA" ...
 $ country_name  : chr  "Nigeria" "Nigeria" "Nigeria" "Nigeria" ...
 $ indicator_id  : chr  "abr" "abr" "abr" "abr" ...
 $ indicator_name: chr  "Adolescent Birth Rate (births per 1,000 women ages 15-19)" "Ado
 $ index_id      : chr  "GII" "GII" "GII" "GII" ...
 $ index_name    : chr  "Gender Inequality Index" "Gender Inequality Index" "Gender Ineq
 $ value         : num  141 138 139 141 140 ...
 $ year          : int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

3. In addition to the above I have repeatedly used str to check the structure of the data. Next, I will ***compare the average of the same indicator in the two countries over years.***

```r
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:data.table':

    between, first, last

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# Filter data for Nigeria and Ireland
nigeria_data <- merged_data %>% filter(country_name == "Nigeria")
ireland_data <- merged_data %>% filter(country_name == "Ireland")

# Calculate mean of 'value' for each 'indicator_name' for Nigeria
nigeria_mean <- nigeria_data %>%
  group_by(indicator_name) %>%
  summarise(
    mean_value_Nigeria = mean(value, na.rm = TRUE),
    .groups = 'drop'
  )

# Calculate mean of 'value' for each 'indicator_name' for Ireland
ireland_mean <- ireland_data %>%
  group_by(indicator_name) %>%
  summarise(
    mean_value_Ireland = mean(value, na.rm = TRUE),
    .groups = 'drop'
  )

# Merge the mean values for Nigeria and Ireland into one table
mean_values_combined <- merge(nigeria_mean, ireland_mean, by = "indicator_name",
                              all = TRUE)
```

```r
# Print the combined mean values
print(mean_values_combined)
```

```
                                                                indicator_name
1                        Adolescent Birth Rate (births per 1,000 women ages 15-19)
2                                                                     Assets (%)
3                       Carbon dioxide emissions per capita (production) (tonnes)
4                                                           Child mortality (%)
5                                          Coefficient of human inequality
6                                                              Cooking fuel (%)
7                                                    Difference from HDI rank
8                                             Difference from HDI value (%)
9                                                          Drinking water (%)
10                                                            Electricity (%)
11                                        Expected Years of Schooling (years)
12                                  Expected Years of Schooling, female (years)
13                                    Expected Years of Schooling, male (years)
14                                                                  GDI Group
15                                                                   GII Rank
16                                 Gross National Income Per Capita (2017 PPP$)
17                            Gross National Income Per Capita, female (2017 PPP$)
18                              Gross National Income Per Capita, male (2017 PPP$)
19                                                                 HDI female
20                                                                   HDI male
21                                                                   HDI Rank
22                                                                Housing (%)
23                                                  Inequality in eduation
24                                                    Inequality in income
25                                          Inequality in life expectancy
26                  Labour force participation rate, female (% ages 15 and older)
27                    Labour force participation rate, male (% ages 15 and older)
28                                       Life Expectancy at Birth (years)
29                                 Life Expectancy at Birth, female (years)
30                                   Life Expectancy at Birth, male (years)
31                                   Material footprint per capita (tonnes)
32                      Maternal Mortality Ratio (deaths per 100,000 live births)
33                                       Mean Years of Schooling (years)
34                                 Mean Years of Schooling, female (years)
35                                   Mean Years of Schooling, male (years)
36                                               MPI Value (Range: 0 to 1)
37                                                        Overall loss (%)
```

38 Population with at least some secondary education, female (% ages 25 and older)
39 Population with at least some secondary education, male (% ages 25 and older)
40 Sanitation (%)
41 School attendance (%)
42 Share of seats in parliament, female (% held by women)
43 Share of seats in parliament, male (% held by men)
44 Years of schooling (%)

| | mean_value_Nigeria | mean_value_Ireland |
|---|---|---|
| 1 | 124.6287879 | 13.6659394 |
| 2 | 4.9450000 | NA |
| 3 | 0.6472121 | 9.6820909 |
| 4 | 19.4730000 | NA |
| 5 | 33.9498333 | 7.5416154 |
| 6 | 10.1140000 | NA |
| 7 | -1.0000000 | -4.0000000 |
| 8 | 1.7835500 | 17.2522424 |
| 9 | 5.7840000 | NA |
| 10 | 7.9000000 | NA |
| 11 | 8.3309394 | 16.7049394 |
| 12 | 7.6919394 | 16.9556061 |
| 13 | 8.9829394 | 16.4614242 |
| 14 | 5.0000000 | 1.0000000 |
| 15 | 165.0000000 | 20.0000000 |
| 16 | 3865.5803636 | 46941.7705455 |
| 17 | 2951.4644545 | 32779.2894545 |
| 18 | 4767.5018485 | 61296.9152727 |
| 19 | 0.4627000 | 0.8559697 |
| 20 | 0.5426500 | 0.8804848 |
| 21 | 161.0000000 | 7.0000000 |
| 22 | 7.8490000 | NA |
| 23 | 41.4397692 | 3.7830769 |
| 24 | 18.9270000 | 15.5516154 |
| 25 | 41.9703846 | 3.2901538 |
| 26 | 52.5848485 | 49.7357576 |
| 27 | 69.6445455 | 70.0693939 |
| 28 | 49.2813333 | 78.8445152 |
| 29 | 50.1473636 | 81.1965455 |
| 30 | 48.4438485 | 76.5158788 |
| 31 | 3.5547576 | 30.7864848 |
| 32 | 1128.8911212 | 7.7716667 |
| 33 | 6.2210500 | 10.1845455 |
| 34 | 5.0538500 | 10.3461212 |
| 35 | 7.4498000 | 10.0172727 |

```
36          0.1750000                NA
37         34.7269167         7.7268462
38         39.0253500        78.0411212
39         55.2764000        77.1300606
40          8.4050000                NA
41         19.6540000                NA
42          5.4762609        17.4258485
43         94.5237391        82.5741515
44         15.8760000                NA
```

From the chart, we can clearly see the huge gap between the two countries for the same indicator, such as the Adolescent Birth Rate (births per 1,000 women ages 15-19), which is 124.6 in Nigeria and 13.7 in Ireland. A lot of useful information can be obtained intuitively, which is convenient for follow-up research.

4. Next, I will ***explore the Adolescent Birth Rate (births per 1,000 women ages 15-19) for Ireland and Nigeria from 1990 to 2022.***

```
# Filter data for the specific indicator_name
filtered_data <- merged_data[indicator_name == "Adolescent Birth Rate (births per 1,000 wo

# Set keys and calculate mean 'value' by 'country_name' and 'year'
result <- filtered_data[, .(mean_value = mean(value, na.rm = TRUE)), keyby =
                          .(country_name, year)]

# Print the result
print(result)
```

```
Key: <country_name, year>
    country_name  year mean_value
          <char> <int>      <num>
 1:      Ireland  1990     15.814
 2:      Ireland  1991     16.584
 3:      Ireland  1992     16.457
 4:      Ireland  1993     15.454
 5:      Ireland  1994     14.375
 6:      Ireland  1995     14.338
 7:      Ireland  1996     15.842
 8:      Ireland  1997     17.063
 9:      Ireland  1998     19.050
10:      Ireland  1999     20.021
11:      Ireland  2000     19.615
```

```
12:      Ireland  2001     19.879
13:      Ireland  2002     19.449
14:      Ireland  2003     18.854
15:      Ireland  2004     17.568
16:      Ireland  2005     16.689
17:      Ireland  2006     16.387
18:      Ireland  2007     16.909
19:      Ireland  2008     16.533
20:      Ireland  2009     15.939
21:      Ireland  2010     14.665
22:      Ireland  2011     12.403
23:      Ireland  2012     11.332
24:      Ireland  2013      9.655
25:      Ireland  2014      8.511
26:      Ireland  2015      7.924
27:      Ireland  2016      7.217
28:      Ireland  2017      6.603
29:      Ireland  2018      6.370
30:      Ireland  2019      5.796
31:      Ireland  2020      5.867
32:      Ireland  2021      5.941
33:      Ireland  2022      5.872
34:      Nigeria  1990    140.866
35:      Nigeria  1991    138.452
36:      Nigeria  1992    138.666
37:      Nigeria  1993    140.754
38:      Nigeria  1994    140.344
39:      Nigeria  1995    136.680
40:      Nigeria  1996    132.575
41:      Nigeria  1997    129.218
42:      Nigeria  1998    128.220
43:      Nigeria  1999    131.906
44:      Nigeria  2000    132.063
45:      Nigeria  2001    131.315
46:      Nigeria  2002    133.101
47:      Nigeria  2003    134.751
48:      Nigeria  2004    130.599
49:      Nigeria  2005    131.024
50:      Nigeria  2006    126.655
51:      Nigeria  2007    123.387
52:      Nigeria  2008    123.818
53:      Nigeria  2009    123.081
54:      Nigeria  2010    123.402
```

```
55:      Nigeria  2011    125.756
56:      Nigeria  2012    127.692
57:      Nigeria  2013    128.112
58:      Nigeria  2014    124.209
59:      Nigeria  2015    114.982
60:      Nigeria  2016    107.474
61:      Nigeria  2017    103.925
62:      Nigeria  2018    103.466
63:      Nigeria  2019    102.807
64:      Nigeria  2020    102.215
65:      Nigeria  2021    101.675
66:      Nigeria  2022     99.560
    country_name  year mean_value
```

Now we have successfully obtained this table, from which we can intuitively see the changes in the values. However, in order to more directly compare the differences in the values of the two countries with the changes in recent years, we will draw some graphs.

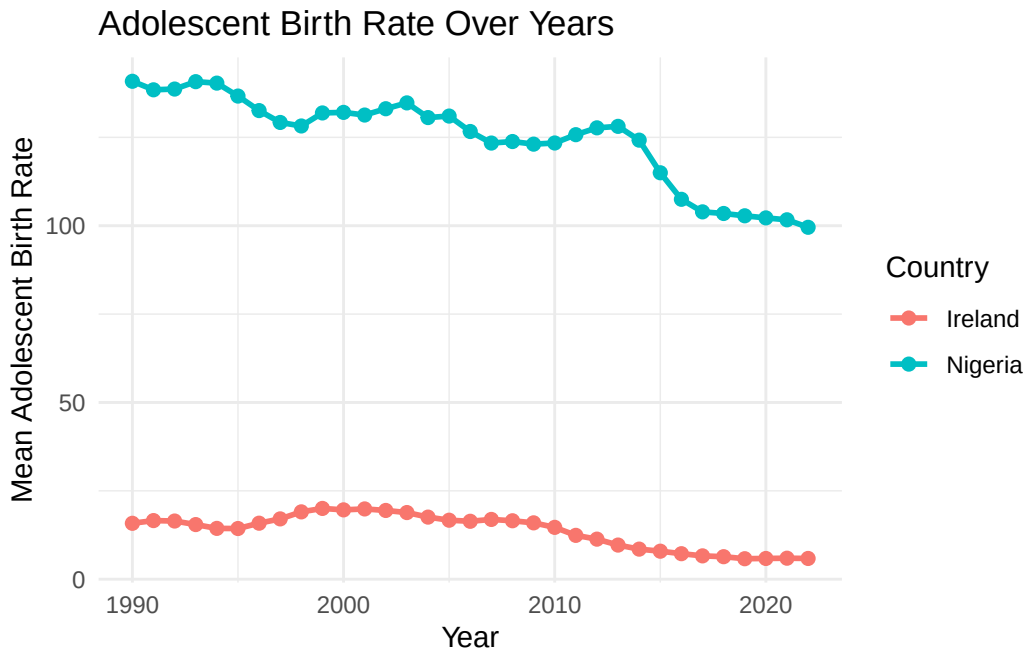
5. Next, I will ***use different line colors to draw line charts of Adolescent Birth Rate changes in these two countries according to years.***

```
# Load necessary packages
library(ggplot2)

# Create a line plot for mean_value over the years for each country
line_plot <- ggplot(result, aes(x = year, y = mean_value, color = country_name,
                                group = country_name)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "Adolescent Birth Rate Over Years",
       x = "Year",
       y = "Mean Adolescent Birth Rate",
       color = "Country") +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
# Print the line plot
print(line_plot)
```

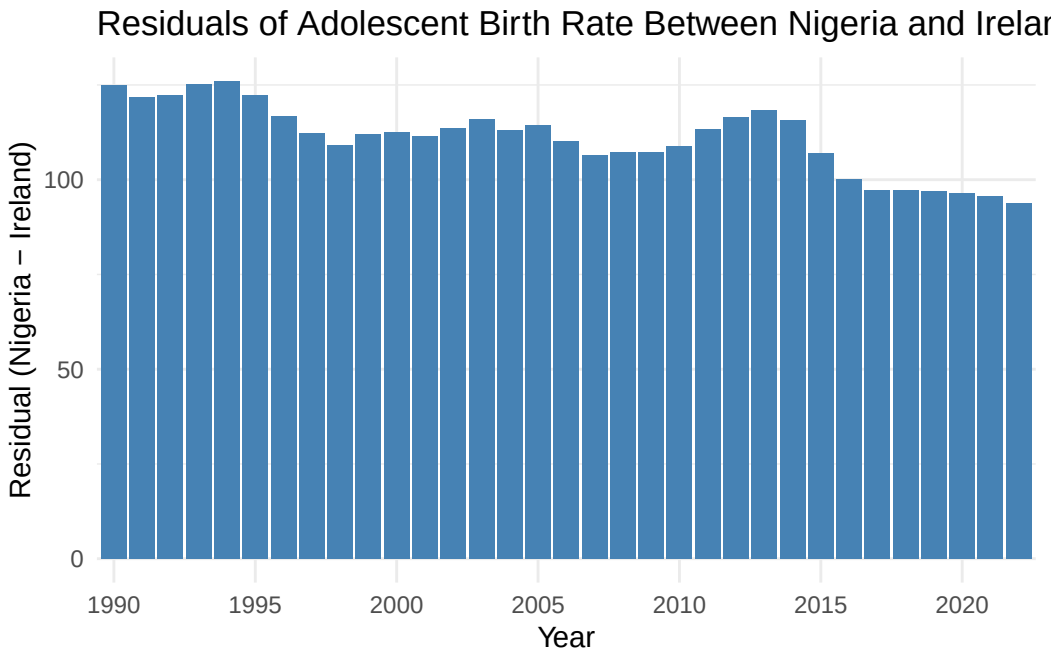



From the figure, we can see a huge gap between this two countris with Adolescent Birth Rate, the values in Ireland are far lower than in Nigeria. And we can see that the Adolescent Birth Rate in Nigeria shows a decreasing trend with the increase of years. Ireland is also on the decline, which indicates that with the development of The Times and the progress of society, the Adolescent Birth Rate is in a downward trend. Now we want to explore whether the difference in Adolescent Birth rates between the two countries changes over the years, so we next plot a bar chart of their differences.

```
# Calculate the difference (residuals) between mean_value of Nigeria and Ireland for each
residuals <- result[, .(residual = mean_value[country_name == "Nigeria"] - mean_value[coun

# Plot a barplot of residuals with adjusted x-axis labels
barplot <- ggplot(residuals, aes(x = factor(year), y = residual)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Residuals of Adolescent Birth Rate Between Nigeria and Ireland",
       x = "Year",
       y = "Residual (Nigeria - Ireland)") +
  scale_x_discrete(breaks = seq(min(residuals$year), max(residuals$year), by = 5)) +  # Ad
  theme_minimal()

# Print the barplot
print(barplot)
```




From the figure, we can see that the Adolescent Birth Rate gap between the two countries also shows a downward trend, which should be due to the large room for decline in Nigeria, while the value of Ireland is relatively low, so it is expected that the gap will continue to narrow in the future.