Responses to Assignment 5

Contents

Assignment Five: The Week Six Assignment 1

Set up **2**
World Development Indicator - WDI 2
World Inequility Report - WIR2022 2

General Comments **3**
Create a PDF or Word file. 3

Your Work **5**
World Development Indicators - WDI 5
Worldbank 5
OECD Data 6
WIR DAta 6
Toy Data 6

Responses to Questions **6**
Q. How to include value in the graphs, for example, over the columns to know the exact values for each country in forest area? 6

Assignment Five: The Week Six Assignment

- Choose a public data. Clearly state how you obtained the data. Even if you are able to give the URL to download the data, explain the steps you reached and obtained the data.
- Create an R Notebook of a Data Analysis containing the following and submit the rendered HTML file (eg. `a5_123456.nb.html` by replacing 123456 with your ID), and a PDF (or MS Word File).
  1. create an R Notebook using the R Notebook Template in Moodle, save as `a3_123456.Rmd`,
  2. write your name and ID and the contents,
  3. run each code block,
  4. preview to create `a5_123456.nb.html`,
  5. render (or knit) PDF, or Word (and then PDF)
  6. submit `a5_123456.nb.html` and PDF (or Word) to Moodle.

1. Choose a data with at least two numerical variables. One of them can be the year.
   - Information of the data
   - Explain why you chose the data
   - List questions you want to study
2. Explore the data using visualization using `ggplot2`
   - Create various charts, and write observed comments
   - Apply a (linear regression) model, and draw a regression line to at least one chart, and write your conclusion based on the model using the slope value and R squared (and/or adjusted R squared).
3. Observations based on your data visualization, and difficulties and questions encountered if any.

Due: 2023-01-30 23:59:00. Submit your R Notebook file, and a PDF file (or a MS Word file) in Moodle (The Fifth Assignment). Due on Monday!

Set up

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl) # for excel files
library(WDI)
```

World Development Indicator - WDI

The following is useful when you use WDI.

```
wdi_cache <- WDIcache()
```

World Inequility Report - WIR2022

- World Inequality Report: https://wir2022.wid.world/
- Executive Summary: https://wir2022.wid.world/executive-summary/
- Methodology: https://wir2022.wid.world/methodology/
- URL of Executive Summary Data: https://wir2022.wid.world/www-site/uploads/2022/03/WIR2022TablesFigures-Summary.xlsx

Please add `mode="wb"` (web binary). This should work better.

```
url_summary <- "https://wir2022.wid.world/www-site/uploads/2022/03/WIR2022TablesFigures-Summary.xlsx"
download.file(url = url_summary,
              destfile = "./data/WIR2022s.xlsx",
              mode = "wb")
```

If you get an error, download the file directory from the methodology site into your computer, then open it with Excel and save it in the data folder of your R Studio project. Then R studio can recognize it easily as an Excel data.

Generally, a text file such as a CSV file is easy to import, but a binary file is difficult to handle. It is because unless R can recognize its file type, for example, Excel or so, R cannot import the data.

```
excel_sheets("./data/WIR2022s.xlsx")
```

```
##  [1] "Index"     "F1"        "F2"        "F3"        "F4"        "F5."
##  [7] "F6"        "F7"        "F8"        "F9"        "F10"       "F11"
## [13] "F12"       "F13"       "F14"       "F15"       "T1"        "data-F1"
## [19] "data-F2"   "data-F3"   "data-F4"   "data-F5"   "data-F6"   "data-F7"
## [25] "data-F8"   "data-F9"   "data-F10"  "data-F11"  "data-F12"  "data-F13."
## [31] "data-F14." "data-F15"
```

General Comments

Create a PDF or Word file.

A Notebook file is created by pressing the Preview button, and the outputs appear as is. However, making a file with another format, R runs all code chunks from the top. So if the object is not defined above the code used, the knit program stops with an error message. I recommend the following steps.

0. Create a PDF right after you create a new (R Notebook) file (using Template). By this step, you can check your 'Knit to PDF' process by `tinytex` is working well. Please let me know if you fail to create a PDF and cannot solve the problem. I will look at the setting of your PC in class.
1. Run all codes before you preview Notebook. You can use 'Run All', and 'Run All Code Chunks Below' under the 'Run' button if there is an incomplete code chunk.
2. Before you create a PDF or word, you need to correct all errors. But if you could not, add eval = FALSE as an option.

```
```{r eval=FALSE}
# code chunk with errors
```
```

You can add a similar option from the gear mark at the top right in the code chunk. Select show nothing (don't run code); it adds {r eval = FALSE, include = FALSE}, and the code chunk itself is skipped.

3. Rerun all. If you can reach the end of the file without having an error, 'Knit to PDF' or 'Knit to Word'.

Creating a Word file is similar, and should be more accessible.

If you fail to create a PDF using `Knit to PDF` or `Knit to Word`, the alternative is to open the notebook wile with nb.html at the end in your web browser, such as Google Chrome, Edge, or Safari, and use the functionality of printing to PDF of your browser.

Other Code Chunk Options

Please review EDA5, and try options under the gear mark at the top right of each code chunk. I will add two useful options, I use often

1. `cash = TRUE` option. Downloading data and accessing to the internet takes time, and may cause trouble for the hosting site. With this option, you can avoid it, and shorten the compilation time to render. I always add this option to `WDI()`. As for `WDIsearch()`, if you use `cache = wdi_cache`, you do not need to add this option. It is another benefit to use `cache = wdi_cache`.

```
```{r cash = TRUE}
# download from the internet
```
```

2. `echo = FALSE` option. When you create a PDF with a limit of pages, you do not want to include some code chunks. Then use this option. The output is included, but the code chunk is not. You can select this option by choosing 'Show output only' option.

Reference

- https://yihui.org/knitr/options/
- Cheat Sheet. We distributed in class. You can download the same from Help: Cheatsheet at the top menu of R Stduio.

Long Table

If you do not want to include a long table in your PDF or Word, use the following.

```
wdi_cache$series %>% slice(1:10)
```

```
##               indicator                                      name
## 1    1.0.HCount.1.90usd         Poverty Headcount ($1.90 a day)
## 2     1.0.HCount.2.5usd         Poverty Headcount ($2.50 a day)
## 3  1.0.HCount.Mid10to50   Middle Class ($10-50 a day) Headcount
## 4       1.0.HCount.Ofcl Official Moderate Poverty Rate-National
## 5   1.0.HCount.Poor4uds            Poverty Headcount ($4 a day)
## 6   1.0.HCount.Vul4to10      Vulnerable ($4-10 a day) Headcount
## 7      1.0.PGap.1.90usd               Poverty Gap ($1.90 a day)
## 8       1.0.PGap.2.5usd               Poverty Gap ($2.50 a day)
## 9     1.0.PGap.Poor4uds                  Poverty Gap ($4 a day)
## 10     1.0.PSev.1.90usd          Poverty Severity ($1.90 a day)
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7  The poverty gap captures the mean aggregate income or consumption shortfall relative to the poverty
## 8  The poverty gap captures the mean aggregate income or consumption shortfall relative to the poverty
## 9  The poverty gap captures the mean aggregate income or consumption shortfall relative to the poverty
## 10
##    sourceDatabase
## 1  LAC Equity Lab
## 2  LAC Equity Lab
## 3  LAC Equity Lab
## 4  LAC Equity Lab
## 5  LAC Equity Lab
## 6  LAC Equity Lab
## 7  LAC Equity Lab
## 8  LAC Equity Lab
## 9  LAC Equity Lab
## 10 LAC Equity Lab
##                                                  sourceOrganization
## 1      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 2      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 3      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 4  LAC Equity Lab tabulations of data from National Statistical Offices.
## 5      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 6      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 7      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 8      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 9      LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
## 10     LAC Equity Lab tabulations of SEDLAC (CEDLAS and the World Bank).
```

This will print only the first ten rows. The following R Basic code does almost the same.

```
head(wdi_cache$country, 10)
```

```
##    iso3c iso2c                    country                     region
## 1    ABW    AW                      Aruba  Latin America & Caribbean
## 2    AFE    ZH Africa Eastern and Southern                 Aggregates
## 3    AFG    AF                Afghanistan                 South Asia
```

```
## 4    AFR    A9                      Africa                Aggregates
## 5    AFW    ZI  Africa Western and Central                Aggregates
## 6    AGO    AO                      Angola         Sub-Saharan Africa
## 7    ALB    AL                     Albania      Europe & Central Asia
## 8    AND    AD                     Andorra      Europe & Central Asia
## 9    ARB    1A                  Arab World                Aggregates
## 10   ARE    AE        United Arab Emirates Middle East & North Africa
##             capital longitude latitude              income        lending
## 1        Oranjestad  -70.0167  12.5167         High income Not classified
## 2                                            Aggregates     Aggregates
## 3             Kabul   69.1761  34.5228          Low income            IDA
## 4                                            Aggregates     Aggregates
## 5                                            Aggregates     Aggregates
## 6            Luanda    13.242 -8.81155 Lower middle income           IBRD
## 7            Tirane   19.8172  41.3317 Upper middle income           IBRD
## 8  Andorra la Vella    1.5218  42.5075         High income Not classified
## 9                                            Aggregates     Aggregates
## 10        Abu Dhabi   54.3705  24.4764         High income Not classified
```

Your Work

Here is a list of data your classmates used for Assignment Five.

World Development Indicators - WDI

- SP.POP.TOTL: Population, total
- NY.GDP.MKTP.KD.ZG: GDP annual growth
- NY.GDP.MKTP.CD: GDP (current US$)
- NY.GDP.MKTP.KD.ZG: GDP growth (annual %)
- NY.GDP.PCAP.KD: GDP per capita (constant 2015 US$)
- NY.GNS.ICTR.ZS: Gross savings (% of GDP)
- BX.TRF.PWKR.CD.DT: Personal remittances, received (current US$)
- SI.POV.GINI: Gini index
- SL.TLF.TOTL.FE.ZS: Labor force, female (% of total labor force)
- SI.DST.10TH.10: Income share held by highest 10%
- SL.UEM.TOTL.ZS: Unemployment, total (% of total labor force) (modeled ILO estimate)
- BX.KLT.DINV.CD.WD: Foreign Direct Investment (FDI)
- AG.LND.FRST.K2: Forest area
- EN.ATM.CO2E.KT: CO2 emissions (kt)
- EG.USE.ELEC.KH.PC:Electric power consumption (kWh per capita)
- FB.ATM.TOTL.P5: Automated teller machines (ATMs) (per 100,000 adults)
- SM.POP.REFG.OR: Refugee population by country or territory of origin
- SG.GEN.PARL.ZS: Proportion of seats held by women in national parliaments (%)
- SE.XPD.TOTL.GD.ZS: Government expenditure on education, total (% of GDP)
- GB.XPD.RSDV.GD.ZS: Research and development expenditure (% of GDP)
- SE.SEC.ENRR: School enrollment rate, secondary (% gross)
- IP.PAT.RESD: Patent applications, residents
- IP.IDS.RSCT: Industrial design applications, resident, by count
- IP.JRN.ARTC.SC: Scientific and technical journal articles
- BM.GSR.ROYL.CD: Intellectual Property Payments (BOP, Current US$)

Worldbank

- Climate Change Knowledge Portal: https://climateknowledgeportal.worldbank.org

- country summary

OECD Data

- Public spending on education: https://data.oecd.org/eduresource/public-spending-on-education.htm
- Private spending on education: https://data.oecd.org/eduresource/private-spending-on-education.htm

WIR DAta

- Executive Summary: https://wir2022.wid.world/executive-summary/

Toy Data

- `datasets::mtcars`: Motor Trend Car Road Tests

Responses to Questions

Q. How to include value in the graphs, for example, over the columns to know the exact values for each country in forest area?

A. Use `geom_text()`. Sometimes `geom_label()` works better.

```
df_q1 <- WDI(country = "all", indicator=c(parl = "SG.GEN.PARL.ZS", gini = "SI.POV.GINI"),
             start=1999, extra=TRUE, cache=wdi_cache)
# df_q1
```

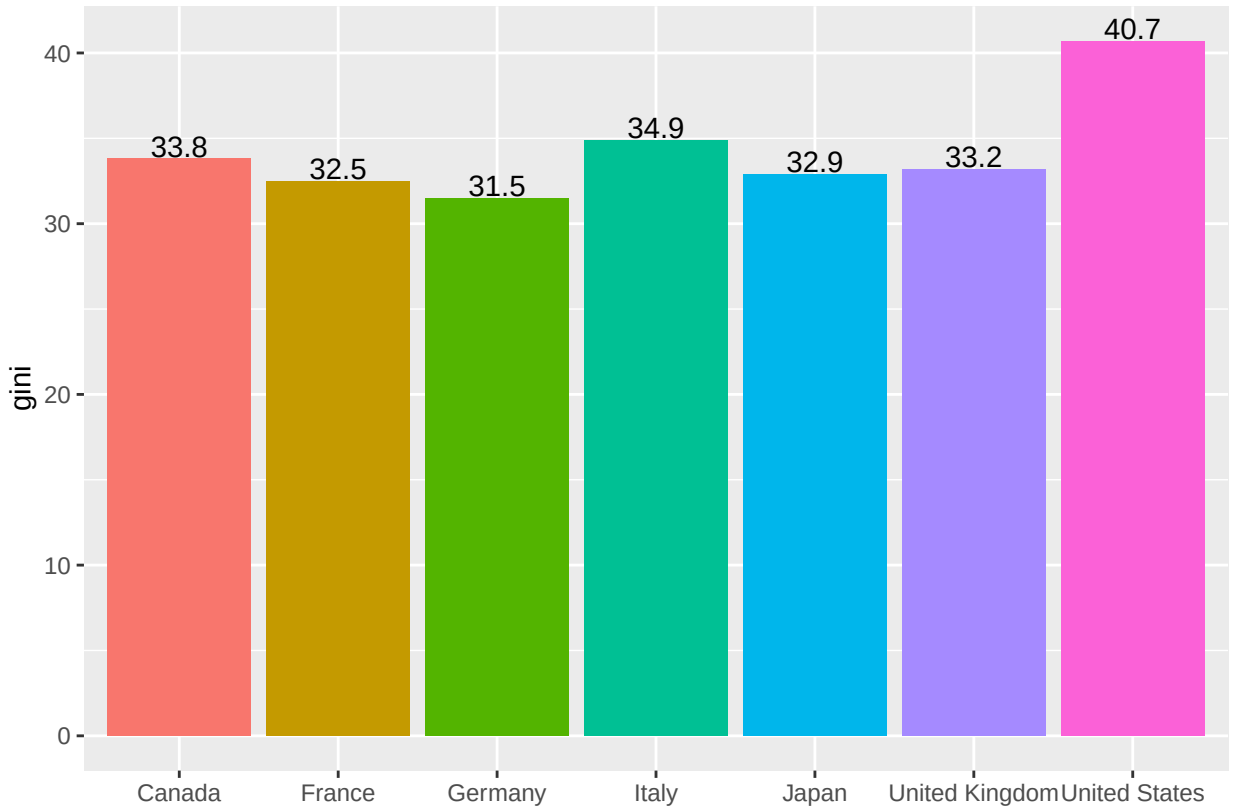
```
asean <- c("Brunei Darussalam", "Cambodia", "Lao PDR", "Myanmar",
           "Philippines", "Indonesia", "Malaysia", "Singapore")
brics <- c("Brazil", "Russian Federation", "India", "China")
g7 <- c("Canada", "France", "Italy", "Japan", "Germany", "United Kingdom", "United States")
```

```
df_q1 %>% filter(country %in% g7) %>% drop_na() %>% group_by(year) %>% summarize(n = n()) %>% arrange(d
```

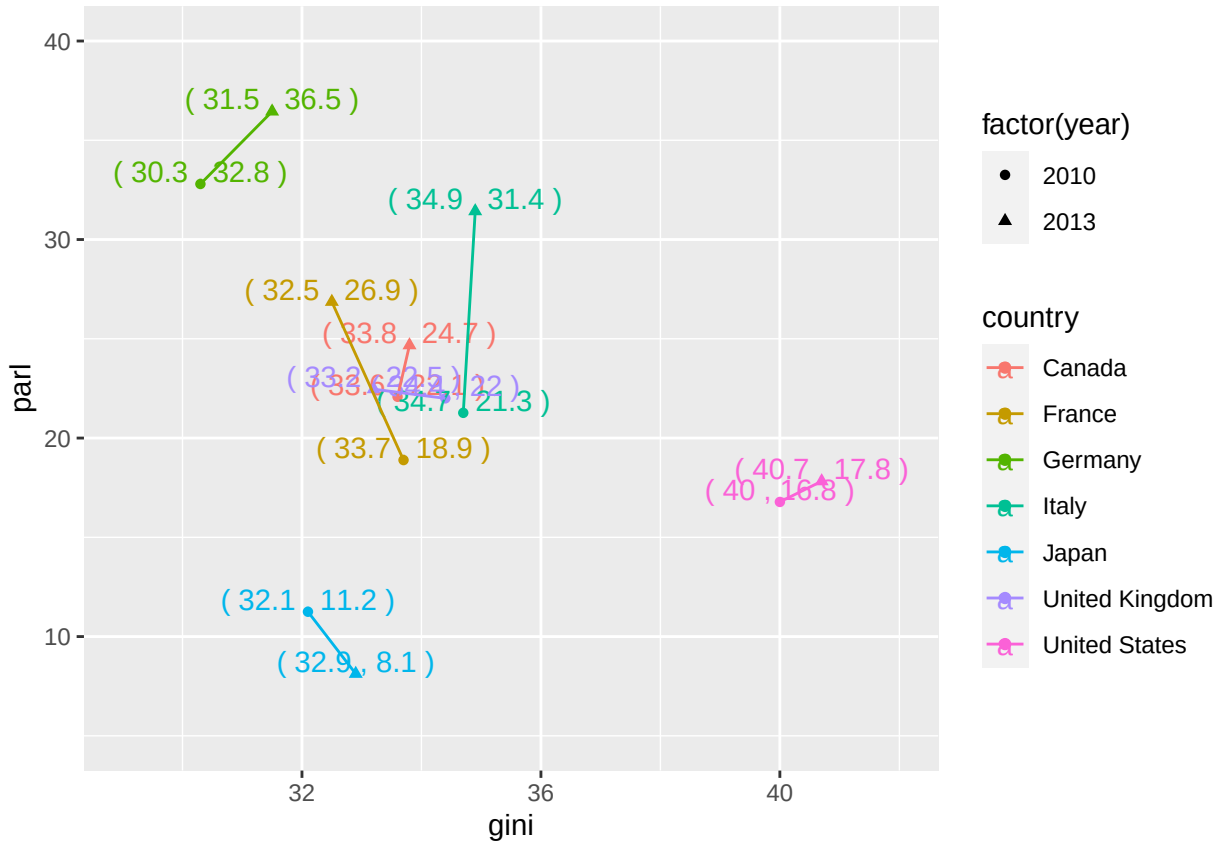
```
## Selecting by n
```

```
## # A tibble: 2 x 2
##    year     n
##   <int> <int>
## 1  2013     7
## 2  2010     7
```

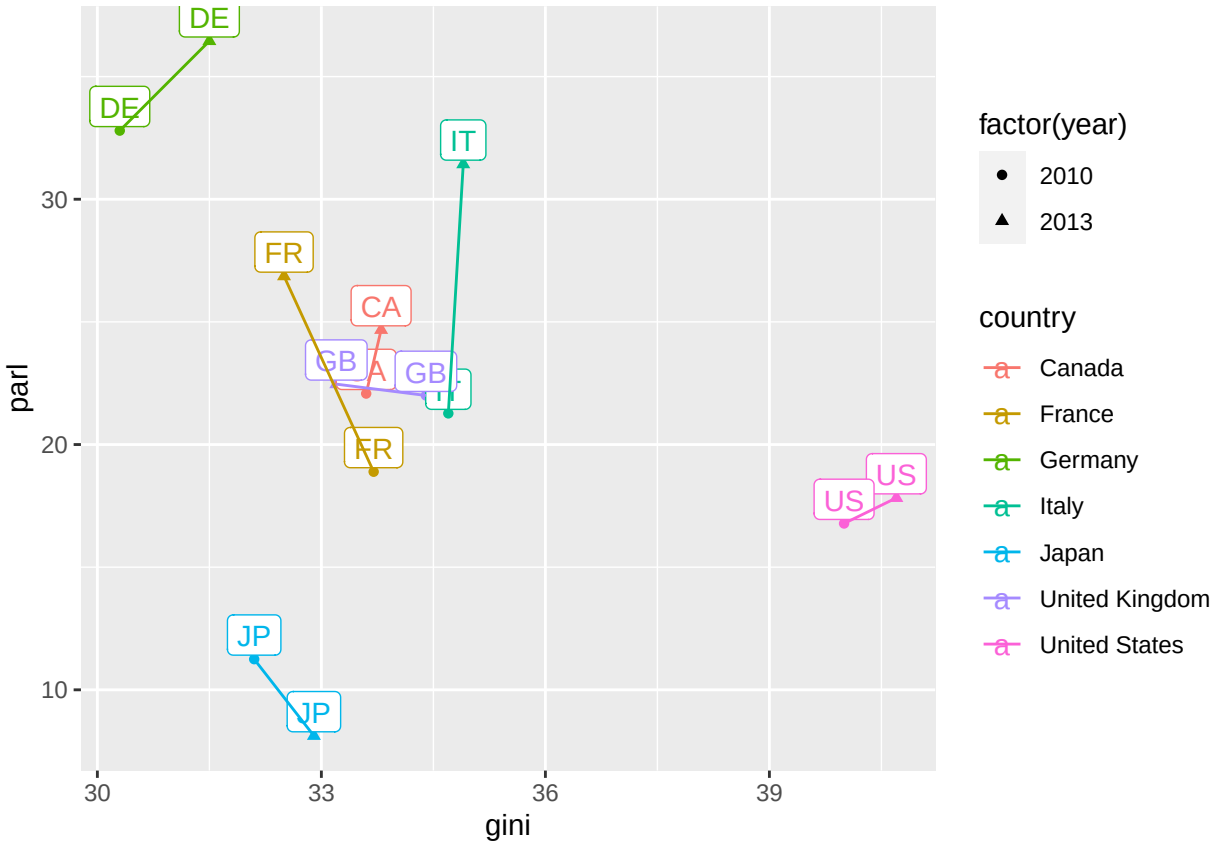
```
df_q1 %>% filter(country %in% g7, year == 2013) %>%
  ggplot(aes(country, gini, fill = country)) + geom_col() +
  geom_text(aes(country, gini, label = gini), vjust = -0.1) + labs(x = "") +
  theme(legend.position = "none")
```




```
df_q1 %>% filter(country %in% g7, year %in% c(2010, 2013)) %>%
  ggplot(aes(gini, parl, color = country)) + geom_point(aes(shape = factor(year))) +
  geom_text(aes(gini, parl, label = paste("(",gini,",", round(parl,1),")")), vjust = -0.1) +
  geom_path(aes(gini, parl, color = country)) + ylim(5,40) + xlim(29,42)
```




```
df_q1 %>% filter(country %in% g7, year %in% c(2010, 2013)) %>%
  ggplot(aes(gini, parl, color = country)) + geom_point(aes(shape = factor(year))) +
  geom_label(aes(gini, parl, label = iso2c), vjust = -0.1) +
  geom_path(aes(gini, parl, color = country))
```



See Responses to Assignment Four, See 4.4.

https://icu-hsuzuki.github.io/da4r2022_note/a4_resp.nb.html

See also:

https://ds-sl.github.io/data-analysis/wir2022.nb.html

Explanation of F1.