exam2

Alex Rigney

6/26/2020

Importing the data.

library(rio)

## Warning: package 'rio' was built under R version 4.0.2

inequality\_data <- import("inequality.xlsx")

Getting to know the data.

str(inequality\_data)

## 'data.frame': 203 obs. of 4 variables:  
## $ iso2c : chr "AL" "AM" "AT" "BY" ...  
## $ country : chr "Albania" "Armenia" "Austria" "Belarús" ...  
## $ inequality\_gini: num 32.9 32.4 30.5 25.6 27.7 NA 47.8 NA NA 46.7 ...  
## $ year : num 2015 2015 2015 2015 2015 ...

head(inequality\_data)

## iso2c country inequality\_gini year  
## 1 AL Albania 32.9 2015  
## 2 AM Armenia 32.4 2015  
## 3 AT Austria 30.5 2015  
## 4 BY Belarús 25.6 2015  
## 5 BE Belgium 27.7 2015  
## 6 BZ Belize NA 2015

tail(inequality\_data)

## iso2c country inequality\_gini year  
## 198 VN Vietnam NA 2015  
## 199 VI Virgin Islands (U.S.) NA 2015  
## 200 PS West Bank and Gaza NA 2015  
## 201 YE Yemen, Rep. NA 2015  
## 202 ZM Zambia 57.1 2015  
## 203 ZW Zimbabwe NA 2015

mean(inequality\_data$year)

## [1] 2015

The data appears to be cross-sectional, as each observation is only sampled once.

Looking at Sweden’s Gini score

se\_gini <- subset(inequality\_data, iso2c == "SE", select = inequality\_gini)  
print(se\_gini)

## inequality\_gini  
## 174 29.2

Looking at Denmark’s score

dk\_gini <- subset(inequality\_data, iso2c == "DK", select = inequality\_gini)  
print(dk\_gini)

## inequality\_gini  
## 40 28.2

Looking at Brazil’s Gini score

br\_gini <- subset(inequality\_data, iso2c == "BR", select = inequality\_gini)  
print(br\_gini)

## inequality\_gini  
## 13 51.9

It is better to have lower Gini Coefficient scores. That means your country’s income is distributed more equally.

Looking at the head of the data.

head(inequality\_data)

## iso2c country inequality\_gini year  
## 1 AL Albania 32.9 2015  
## 2 AM Armenia 32.4 2015  
## 3 AT Austria 30.5 2015  
## 4 BY Belarús 25.6 2015  
## 5 BE Belgium 27.7 2015  
## 6 BZ Belize NA 2015

Creating a function to remove accents from the country variable.

# define the function  
accents.remove <- function(s) {  
# 1 character substitutions  
old <- "ú"  
new <- "u"  
s1 <- chartr(old, new, s)  
  
for(i in seq\_along(old)) s1 <- gsub(old[i], new[i], s1, fixed = TRUE)  
s1  
}  
# finish the accent FIX  
inequality\_data$country = accents.remove(inequality\_data$country)  
  
head(inequality\_data)

## iso2c country inequality\_gini year  
## 1 AL Albania 32.9 2015  
## 2 AM Armenia 32.4 2015  
## 3 AT Austria 30.5 2015  
## 4 BY Belarus 25.6 2015  
## 5 BE Belgium 27.7 2015  
## 6 BZ Belize NA 2015

Sorting the data by Gini score, from lowest to highest.

inequality\_data <- inequality\_data[order(inequality\_data$inequality\_gini),]  
  
head(inequality\_data)

## iso2c country inequality\_gini year  
## 161 SI Slovenia 25.4 2015  
## 190 UA Ukraine 25.5 2015  
## 4 BY Belarus 25.6 2015  
## 39 CZ Czech Republic 25.9 2015  
## 92 XK Kosovo 26.5 2015  
## 160 SK Slovak Republic 26.5 2015

Checking the mean of the Gini Score.

gini\_mean <- mean(inequality\_data$inequality\_gini, na.rm = TRUE)  
gini\_mean

## [1] 36.81375

Trying and failing to create a high and low inequality dummy variable.

inequality\_data$low\_inequality=NA  
inequality\_data$high\_inequality=NA  
  
ifelse(test = inequality\_data$inequality\_gini > 36.81375, yes = inequality\_data$high\_inequality <- 1, no = inequality\_data$low\_inequality <- 1)

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [26] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [51] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [76] 1 1 1 1 1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [101] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [126] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [151] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [176] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [201] NA NA NA

I’ve spent 30 minutes on this, I give up

Printing three names using a for loop.

names <- c("The World Bank", "the African Development Bank", "the Bill and Melinda Gates Foundation")  
for(n in names)  
 print(n)

## [1] "The World Bank"  
## [1] "the African Development Bank"  
## [1] "the Bill and Melinda Gates Foundation"

I chose the literacy rate for adults over the age of 15. I would assume that nations with a low literacy rate would tend to have higher inequality.

Importing the data and changing the name to something sensable.

library(WDI)  
lit\_data = WDI(country = "all",  
indicator = c("SE.ADT.LITR.ZS"), # indicator from web  
start = 2015, end = 2015, extra = FALSE, cache = NULL)  
  
library(data.table)  
setnames(lit\_data, "SE.ADT.LITR.ZS", "lit\_rate")

Merging the two data frames into one, left side join.

library(tidyverse)

## -- Attaching packages ----------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.1.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.2

## -- Conflicts -------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::between() masks data.table::between()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::first() masks data.table::first()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::last() masks data.table::last()  
## x purrr::transpose() masks data.table::transpose()

merged\_df <- left\_join(x=inequality\_data,  
 y=lit\_data,  
 by = c("iso2c", "country", "year"))  
  
str(merged\_df)

## 'data.frame': 203 obs. of 7 variables:  
## $ iso2c : chr "SI" "UA" "BY" "CZ" ...  
## $ country : chr "Slovenia" "Ukraine" "Belarus" "Czech Republic" ...  
## $ inequality\_gini: num 25.4 25.5 25.6 25.9 26.5 26.5 26.8 26.8 27 27.1 ...  
## $ year : num 2015 2015 2015 2015 2015 ...  
## $ low\_inequality : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ high\_inequality: num 1 1 1 1 1 1 1 1 1 1 ...  
## $ lit\_rate : num NA NA NA NA NA NA NA NA NA NA ...  
## ..- attr(\*, "label")= chr "Literacy rate, adult total (% of people ages 15 and above)"

names(merged\_df)

## [1] "iso2c" "country" "inequality\_gini" "year"   
## [5] "low\_inequality" "high\_inequality" "lit\_rate"

head(merged\_df)

## iso2c country inequality\_gini year low\_inequality high\_inequality  
## 1 SI Slovenia 25.4 2015 1 1  
## 2 UA Ukraine 25.5 2015 1 1  
## 3 BY Belarus 25.6 2015 1 1  
## 4 CZ Czech Republic 25.9 2015 1 1  
## 5 XK Kosovo 26.5 2015 1 1  
## 6 SK Slovak Republic 26.5 2015 1 1  
## lit\_rate  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

Dropping all the countries with NAs in either the gini scores or the literacy rate.

merged\_df <-  
merged\_df %>%  
 filter(!(inequality\_gini=="NA")) %>%   
 filter(!(lit\_rate=="NA"))

Creating a data frame of only those countries with gini scores above 30.

data\_greater\_30 <-  
merged\_df %>%  
 filter(inequality\_gini > 30)  
data\_greater\_30

## iso2c country inequality\_gini year low\_inequality high\_inequality  
## 1 GM Gambia, The 35.9 2015 1 1  
## 2 TH Thailand 36.0 2015 1 1  
## 3 ES Spain 36.2 2015 1 1  
## 4 UY Uruguay 40.1 2015 1 1  
## 5 SV El Salvador 40.6 2015 1 1  
## 6 ID Indonesia 41.0 2015 1 1  
## 7 CV Cabo Verde 42.4 2015 1 1  
## 8 TR Turkey 42.9 2015 1 1  
## 9 TG Togo 43.1 2015 1 1  
## 10 PE Peru 43.4 2015 1 1  
## 11 CL Chile 44.4 2015 1 1  
## 12 PH Philippines 44.4 2015 1 1  
## 13 DO Dominican Republic 45.2 2015 1 1  
## 14 EC Ecuador 46.0 2015 1 1  
## 15 BO Bolivia 46.7 2015 1 1  
## 16 PY Paraguay 47.6 2015 1 1  
## 17 HN Honduras 49.6 2015 1 1  
## 18 CO Colombia 51.1 2015 1 1  
## 19 BR Brazil 51.9 2015 1 1  
## lit\_rate  
## 1 50.77797  
## 2 92.86831  
## 3 98.14326  
## 4 98.52387  
## 5 87.96985  
## 6 95.21793  
## 7 86.79029  
## 8 95.60142  
## 9 63.74562  
## 10 94.16234  
## 11 96.87413  
## 12 98.18255  
## 13 91.99121  
## 14 94.45566  
## 15 92.45508  
## 16 95.55484  
## 17 87.90683  
## 18 94.24505  
## 19 92.04790

summing up the number of countries that have “ai” in their name.

sum(grep("ai", data\_greater\_30))

## [1] 2

Summing up the gini scores, or at least trying to.

sapply(data\_greater\_30$inequality\_gini, FUN = sum)

## [1] 35.9 36.0 36.2 40.1 40.6 41.0 42.4 42.9 43.1 43.4 44.4 44.4 45.2 46.0 46.7  
## [16] 47.6 49.6 51.1 51.9

Labeling and exporting the data frame so future researchers will be able to continue my great work.

library(labelled)

## Warning: package 'labelled' was built under R version 4.0.2

var\_label(data\_greater\_30) <- list(  
 `iso2c` = "boop beep boop",   
 `country` = "different thing",  
 `inequality\_gini` = "another thing",  
 `year` = "How many are there?",  
 `low\_inequality` = "it can't go on any longer",  
 `high\_inequality` = "but it is!",  
 `lit\_rate` = "thanks for grading the test!")  
  
# save the dataset in Stata format with the labels  
export(data\_greater\_30, file = "final\_data.dta")

<https://github.com/adrr1/rwork>

Thank you!