Test 2 Final

Brandon Taylor

6/26/2020

1. Please clear the environment in R.

remove(list = ls())

1. Load the “inequality” dataset into R, and save the data frame as ‘inequality\_data’.

library(rio)  
inequality\_data <- import("inequality.xlsx")  
attach(inequality\_data)

1. Is this dataset a cross-sectional or panel dataset? Explain why in words and provide some R code to prove that your answer is correct.

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

summary(inequality\_data)

## iso2c country inequality\_gini year   
## Length:203 Length:203 Min. :25.40 Min. :2015   
## Class :character Class :character 1st Qu.:31.55 1st Qu.:2015   
## Mode :character Mode :character Median :35.75 Median :2015   
## Mean :36.81 Mean :2015   
## 3rd Qu.:41.12 3rd Qu.:2015   
## Max. :59.10 Max. :2015   
## NA's :123

1. Using the subset command, provide the inequality\_gini scores for Denmark and Sweden. library(dplyr) inequality\_datasub <- subset(inequality\_gini, select = c(“Denmark”, “Sweden”))
2. Since Brazil started the Bolsa Familia conditional cash transfer program in 1990s, inequality in Brazil has decreased significantly. Just the same, inequality in Brazil is very high comparatively. Using the subset command, please show the inequality\_gini score for Brazil.

inequality\_data <- subset(inequality\_gini, “Brazil”)

1. Given your answers to the previous questions, is it better to have a high or low inequality\_gini scores?

It is better to have a lower score.

1. Use the head command to get a quick peak at the data frame.

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

1. Write a function called “accent.remove” to remove the accent on Belarus, apply that function, and run the head command again to show that you removed the accent.
2. Sort the data by the countries with the lowest inequality\_gini scores and then run the head command again to show what the top 5 countries are.
3. What is the mean inequality\_gini score? Provide the relevant R code.

mean(inequality\_gini)

## [1] NA

1. Using the ifelse command, create two new dummy variables, high\_inequality and low\_inequality, which takes values of either zero or one. The low\_inequality variable should correspond to countries with inequality\_gini scores below the mean. The high\_inequality variable should correspond to countries with inequality\_gini scores above the mean. (Note: we will not accept answers that do not use the ifelse command to create the variables.)
2. Runacross-tabusingthehigh\_inequalityandlow\_inequalityvariablesthatyoucre- ated in the previous question. The cross-tab should provide the mean inequality\_gini score and number of observations for each category of inequality. (Note: if you had trouble using the ifelse command, we couldn’t provide points for the previous question. However, you can create the variables using the indexing method)
3. The World Bank, the African Development Bank, and the Bill and Melinda Gates Foundation are all working on reducing inequality in Africa. Write a for loop that prints the names of these three actors. (Note: we will not accept answers that do not provide a for loop.)
4. Use this website to find a variable from the World Development Indicators that you think is correlated with inequality. Tell us what variable you picked and why you picked it. (Don’t worry if your prediction is not correct. We just want you to provide some rationale.)

Mobile cellular subscriptions (per 100 people), is mobile phone susbscriptions corllated with the inequlity level.

1. Import that variable directly into R. (Note: if you are having trouble, read Mike Denly’s Canvas announcement from the other day.)

mobphone <- import("API\_IT.CEL.SETS.P2\_DS2\_en\_csv\_v2\_1121611.csv")  
attach(mobphone)

1. Rename the variable that you imported into something that we can actually understand.

mobphone

1. Merge the new variable into the other dataset, using inequality\_data as the x and and your new data frame as the y. When merging use the command that only keeps the rows in your x data frame. Call your new data frame merged\_df. Ensure that you have no variables with .x or .y at the end of them in your new merged\_df, while at the same time ensuring there are still variables like country and year.
2. In merged\_df, remove the missing data on the basis of inequality\_gini and your new variable that you took from the World Development Indicators.
3. Using the filter command and piping method, only keep the data with inequality\_gini scores greater than 30. Save the new data frame as data\_greater\_30. (Note: we will not accept answers using subset.)
4. Using data\_greater\_30, use to R to count how many countries have the sequence “ai”in their name.
5. Use any command from the apply family to take the sum of inequality\_gini in data\_greater\_30.
6. Label your variables in merged\_df. Any labels will suffice. library(labelled)
7. Save the labeled data frame as a Stata dataset called final\_data.

export(merged\_df, file = “final\_dtat.dta”)

1. Save all of the files (i.e. .Rmd, .dta, .xlsx, .pdf/Word Doc), push them to your GitHub repo, and provide us with the link to that repo.

<https://github.com/brantaylor1595/datascienceclass2>