Exam 2

Connor_Hanna

6/26/2020

Question 1

"Please clear the environment in R"

```
clear
```

Question 2

"Load the "inequality" dataset into R, and save the data frame as 'inequality_data'"

```
inequality_data <- read_excel("~/School/2020 SS/Data Science/Exam 2/inequality.xlsx")
View(inequality)
```

Question 3

"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"

Inequality_data contains cross-sectional data, but could potentially be mended with other datasets from other years to create panel data.

Question 4

"The data frame contains a variable called inequality_gini… Using the subset command, provide the inequality_gini scores for Denmark and Sweden."

```
inequality_data_scandi <- subset(inequality_data, country == "Denmark" | country == "Sweden", select = c
View(inequality_data_scandi)
```

Question 5

"Since Brazil … Using the subset command, please show the inequality_gini score for Brazil."

```
inequality_data_brazi <- subset(inequality_data, country == "Brazil", select = c(country, inequality_gin
View(inequality_data_brazi)
```

Question 6

"Given your answers to the previous questions, is it better to have a high or low inequality_gini scores?"

Given the answers to the above, a lowere gini score is better.

Question 7

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

```
head(inequality_data)
```

Question 8

"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"

```
accent.remove <- iconv(inequality_data$country, from = "ú", to = "u", sub = NA)
sapply(inequality_data$country, accent.remove)
head(inequality_data)
```

This function is saying that the character conversion is unsupported. I'll revisit this later.

Question 9

"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."

```
sort_by_gini <- order(inequality_data$inequality_gini, decreasing = TRUE)
inequality_data[sort_by_gini, ]
head(inequality_data)
```

Question 10

"What is the mean inequality_gini score? Provide the relevant R code"

```
mean(inequality_data$inequality_gini)
```

Question 11

"Using the ifelse command, create two new dummy variables, high_inequality and low_inequality, which takes values of either zero or one. . . "

```
mean_gini <- mean(inequality_data$inequality_gini)
inequality_data$high_inequality <- ifelse(inequality_data$inequality_gini >= mean_gini, 1, 0)
inequality_data$low_inequality <- ifelse(inequality_data$inequality_gini < mean_gini, 1, 0)
```

##Question 12

"Run a cross-tab using the high_inequality and low_inequality variables that you created in the previous question."

```
table(inequality_data$high_inequality, inequality_data$low_inequality)
```

##Question 13

"Write a for loop that prints the names of these three actors"

```
names_list <- c("Bill and Melinda Gates Foundation", "The World Bank", "The African Development Bank")
innovating_for_development <- c(1, 1, 1)
looper_frame <- data.frame(names_list, innovating_for_development)
for (val in looper_frame) {
  if(innovating_for_development == 1)
  print(names_list)
}
```

Question 14

"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."

I'm choosing cereal yield (AG.YLD.CREL.KG) and predict that cereal yield will positively correlate with inequality as Cheerios release fluoride into the water.

Question 15

“Import that variable directly into R”

```
library(WDI)
WDI_cereal <- WDI(country = "all", indicator = "AG.YLD.CREL.KG", start = 2015, end = 2015, extra = FALSE
```

Question 16

“Rename the variable that you imported into something that we can actually understand”

```
library(tidyverse)
rename(AG.YLD.CREL.KG = cereal_yield_kg_per_hectare)
```

Question 17

“Merge the new variable into the other dataset, using inequality_data as the x and and your new data frame as the y. . . ”

```
merged_df <- merge(inequality_data, WDI_cereal, by = "country")
```

Question 18

“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”

```
na.omit(merged_df)
```

Question 19

“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”

```
data_greater_30 <- filter(merged_df, inequality_gini > 30)
```

Question 20

“Using data_greater_30, use R to count how many countries have the sequence “ai” in their name"

```
install.packages(stringr)
library(stringr)
str_count(data_greater_30$country, "ai")
```

Question 21

“Use any command from the apply family to take the sum of inequality_gini in data_greater_30”

```
sapply(data_greater_30$inequality_gini, sum)
```

Question 22

“Label your variables in merged_df. Any labels will suffice”

Sadly my WDI package stopped working and I didn’t have time to fix it.

Question 23

“Save the labeled data frame as a Stata dataset called final_data”

```
save(merged_df, file="final_data.dta")
```

Question 24

Will be submitted via email