Exam 2

Brianna Baker

6/26/2020

Exam 2

Load the library and save the data frame as inequality_data.

```
library(rio)
education_data = import("~/Desktop/inequality.xlsx", which =1)
#saving data frame
inequality_data <- education_data
#removing education_data from environment
rm(education_data)
```

Question 3

This is a cross-sectional dataset because it provides a snapshot of data from the same time and not change over time. We can see this is the code below.

```
summary(inequality_data$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2015    2015    2015    2015    2015    2015
```

##Using the subset command to show the inequality_gini scores for Denmark and Sweden

```
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```r
subset(inequality_data, country == "Denmark")
```

```
##    iso2c country inequality_gini year
## 40    DK Denmark            28.2 2015
```

```r
subset(inequality_data, country == "Sweden")
```

```
##     iso2c country inequality_gini year
## 174    SE  Sweden            29.2 2015
```

##Inequality score for Brazil.

```r
subset(inequality_data, country == "Brazil")
```

```
##    iso2c country inequality_gini year
## 13    BR  Brazil            51.9 2015
```

##Question 6 Since Denamrk and Sweden were described as having “optimal Gini index scores,” and they have much lower scores than Brazil, it appears that it is better to have a low inequality gini score.

###Quick peak at data frame:

```r
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
```

###Removing accent with new function “accent.remove.”

```r
#change default text encoding to UTF-8
#define a function
remove.accents <- function(s) {
  #1 character substiutions
  old1 <- "ú"
  new1 <- "u"
  s1 <- chartr(old1, new1, s)
}
#remove accents
inequality_data$country <- remove.accents(inequality_data$country)
```

###Quick peak to show accent removal

```r
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 data by countries with lowest inequality_gini scores.

```
inequality_data <- inequality_data[order(inequality_data$inequality_gini),]
#top 5 countries
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
```

###Mean of inequality gini scores

```
mean(inequality_data$inequality_gini, na.rm = TRUE)
```

```
## [1] 36.81375
```

###Using if else to recode variables and assign values based on relation to mean. for (r in 1:nrow(inequality_data)){ for(c in 1:ncol(inequality_data)) { if(inequality_data$inequality_{g}ini[r, c] > 36.81375)inequality_{d}ata[r, c] = ”high_{i}nequality”elseif(inequality_{d}ata$inequality_gini[r,c] < 36.81375) { inequality_data[r,c] = “low_inequality” } else{ } } }

Question 13

```
#create vector
actors <- c('World Bank', 'African Development Bank', 'Bill and Melinda Gates Foundation')

#create for statement
for (i in actors) {
  print (i)
}
```

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

###Question 14 I chose the variable “Employment to population ratio” to demonstrate inequality because the comparison could show whether or not the majority of a country has employed inhabitants or not. The more employed, the lower the inquality- is my prediction.

```r
library(devtools)
```

```
## Loading required package: usethis
```

```r
library(remote)
```

```
## Loading required package: Rcpp

## Loading required package: raster

## Loading required package: sp

##
## Attaching package: 'raster'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
#add some data from the World Development Indicators (WDI)
library(WDI)
employment_ratio = WDI(country = "all",
                      indicator = "SL.EMP.TOTL.SP.ZS",
                      start = 2015, end = 2015, extra = FALSE, cache = NULL)
#quick peak
summary(employment_ratio)
```

```
##     iso2c              country          SL.EMP.TOTL.SP.ZS      year     
##  Length:264         Length:264         Min.   :32.22     Min.   :2015  
##  Class :character   Class :character   1st Qu.:51.14     1st Qu.:2015  
##  Mode  :character   Mode  :character   Median :58.09     Median :2015  
##                                        Mean   :57.59     Mean   :2015  
##                                        3rd Qu.:63.70     3rd Qu.:2015  
##                                        Max.   :87.75     Max.   :2015  
##                                        NA's   :31
```

```r
#changing name of variable
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following object is masked from 'package:raster':
##
##     shift

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
#changing name of column to something easier to interpret
setnames(employment_ratio, "SL.EMP.TOTL.SP.ZS", "employment_ratio")
```

###Merge new variable into other dataset

```
merged_df = left_join(inequality_data,
                        employment_ratio,
                        by = c("iso2c", "year"))
```

#drop country.y and rename country.x as country library(tidyverse) merged_df <- merged_df %>% select(-c(“country.x”)) %>% rename(“country” = “country.y”)

###Removing NAs

```
na.omit(merged_df, select =c("employment_ratio", "inequality_gini"))
```

```
##    iso2c          country.x inequality_gini year          country.y
## 1     SI           Slovenia            25.4 2015           Slovenia
## 2     UA            Ukraine            25.5 2015            Ukraine
## 3     BY            Belarus            25.6 2015            Belarus
## 4     CZ     Czech Republic            25.9 2015     Czech Republic
## 6     SK    Slovak Republic            26.5 2015    Slovak Republic
## 7     IS            Iceland            26.8 2015            Iceland
## 8     KZ         Kazakhstan            26.8 2015         Kazakhstan
## 9     MD            Moldova            27.0 2015            Moldova
## 10    FI            Finland            27.1 2015            Finland
## 11    NO             Norway            27.5 2015             Norway
## 12    BE            Belgium            27.7 2015            Belgium
## 13    DK            Denmark            28.2 2015            Denmark
## 14    NL        Netherlands            28.2 2015        Netherlands
## 15    KG    Kyrgyz Republic            29.0 2015    Kyrgyz Republic
## 16    SE             Sweden            29.2 2015             Sweden
## 17    MT              Malta            29.4 2015              Malta
## 18    HU            Hungary            30.4 2015            Hungary
## 19    AT            Austria            30.5 2015            Austria
## 20    HR            Croatia            31.1 2015            Croatia
## 21    DE            Germany            31.7 2015            Germany
## 22    EG   Egypt, Arab Rep.            31.8 2015   Egypt, Arab Rep.
## 23    IE            Ireland            31.8 2015            Ireland
## 24    PL             Poland            31.8 2015             Poland
## 25    CH        Switzerland            32.3 2015        Switzerland
## 26    AM            Armenia            32.4 2015            Armenia
## 27    EE            Estonia            32.7 2015            Estonia
## 28    FR             France            32.7 2015             France
## 29    TN            Tunisia            32.8 2015            Tunisia
## 30    AL            Albania            32.9 2015            Albania
## 31    GB     United Kingdom            33.2 2015     United Kingdom
## 32    PK           Pakistan            33.5 2015           Pakistan
## 33    LU         Luxembourg            33.8 2015         Luxembourg
## 34    CY             Cyprus            34.0 2015             Cyprus
## 35    TJ         Tajikistan            34.0 2015         Tajikistan
## 36    LV             Latvia            34.2 2015             Latvia
## 37    ET           Ethiopia            35.0 2015           Ethiopia
## 38    IT              Italy            35.4 2015              Italy
## 39    PT           Portugal            35.5 2015           Portugal
## 40    MK    North Macedonia            35.6 2015    North Macedonia
## 41    GM        Gambia, The            35.9 2015        Gambia, The
## 42    RO            Romania            35.9 2015            Romania
## 43    GR             Greece            36.0 2015             Greece
## 44    TH           Thailand            36.0 2015           Thailand
## 45    ES              Spain            36.2 2015              Spain
## 46    GE            Georgia            36.5 2015            Georgia
## 47    LT          Lithuania            37.4 2015          Lithuania
## 48    TO              Tonga            37.6 2015              Tonga
## 49    RU Russian Federation            37.7 2015 Russian Federation
## 50    MM            Myanmar            38.1 2015            Myanmar
## 51    BG           Bulgaria            38.6 2015           Bulgaria
## 52    CN              China            38.6 2015              China
## 53    ME         Montenegro            39.0 2015         Montenegro
## 54    IR Iran, Islamic Rep.            39.5 2015 Iran, Islamic Rep.
## 55    UY            Uruguay            40.1 2015            Uruguay
## 56    RS             Serbia            40.5 2015             Serbia
## 57    SV        El Salvador            40.6 2015        El Salvador
## 58    KE              Kenya            40.8 2015              Kenya
## 59    ID          Indonesia            41.0 2015          Indonesia
## 60    MY           Malaysia            41.0 2015           Malaysia
## 61    CI      Cote d’Ivoire            41.5 2015      Cote d’Ivoire
## 62    CV         Cabo Verde            42.4 2015         Cabo Verde
## 63    TR             Turkey            42.9 2015             Turkey
## 64    TG               Togo            43.1 2015               Togo
## 65    PE               Peru            43.4 2015               Peru
## 66    CL              Chile            44.4 2015              Chile
## 67    PH        Philippines            44.4 2015        Philippines
## 68    DO Dominican Republic            45.2 2015 Dominican Republic
## 69    EC            Ecuador            46.0 2015            Ecuador
## 70    BO            Bolivia            46.7 2015            Bolivia
## 71    PY           Paraguay            47.6 2015           Paraguay
## 72    BJ              Benin            47.8 2015              Benin
## 73    CR         Costa Rica            48.4 2015         Costa Rica
## 74    HN           Honduras            49.6 2015           Honduras
## 75    PA             Panama            50.8 2015             Panama
## 76    CO           Colombia            51.1 2015           Colombia
## 77    BR             Brazil            51.9 2015             Brazil
## 78    BW           Botswana            53.3 2015           Botswana
## 79    ZM             Zambia            57.1 2015             Zambia
## 80    NA            Namibia            59.1 2015            Namibia
##    employment_ratio
## 1            52.266
## 2            49.738
## 3            60.461
## 4            56.613
## 6            52.768
## 7            73.954
## 8            67.399
## 9            42.477
## 10           53.325
## 11           62.102
## 12            48.912
## 13            58.197
## 14            59.667
## 15            57.668
## 16            59.235
## 17            51.615
## 18            51.094
## 19            56.661
## 20            44.044
## 21            57.304
## 22            41.666
## 23            55.996
## 24            52.374
## 25            64.938
## 26            46.663
## 27            58.349
## 28            49.721
## 29            39.737
## 30            45.640
## 31            59.061
## 32            51.142
## 33            55.404
## 34            53.296
## 35            37.298
## 36            54.129
## 37            78.365
## 38            42.945
## 39            51.323
## 40            41.047
## 41            53.516
## 42            50.742
## 43            39.199
## 44            68.634
## 45            45.587
## 46            57.827
## 47            53.859
## 48            59.270
## 49            59.140
## 50            65.047
## 51            49.224
## 52            66.593
## 53            43.588
## 54            37.750
## 55            60.100
## 56            42.866
## 57            56.485
## 58            72.312
## 59            63.494
## 60            62.441
## 61            55.906
## 62            52.798
## 63            45.753
## 64            76.523
## 65            73.405
## 66           58.043
## 67           60.302
## 68           58.878
## 69           64.074
## 70           64.988
## 71           66.469
## 72           68.964
## 73           56.375
## 74           62.474
## 75           62.942
## 76           64.063
## 77           58.652
## 78           58.311
## 79           67.572
## 80           47.981
```

##Filtering out data with inequality gini scores greater than 30

```
data_greater_30 <-
  merged_df %>%
  dplyr::filter(inequality_gini > 30)
```

###Count how many countries contain "ai"

```
grep("ai", data_greater_30)
```

```
## [1] 2 5
```

###Using lapply to take sum of inequality gini

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

```
## [[1]]
## [1] 30.4
##
## [[2]]
## [1] 30.5
##
## [[3]]
## [1] 31.1
##
## [[4]]
## [1] 31.7
##
## [[5]]
## [1] 31.8
##
## [[6]]
## [1] 31.8
##
## [[7]]
## [1] 31.8
```

```
##
## [[8]]
## [1] 32.3
##
## [[9]]
## [1] 32.4
##
## [[10]]
## [1] 32.7
##
## [[11]]
## [1] 32.7
##
## [[12]]
## [1] 32.8
##
## [[13]]
## [1] 32.9
##
## [[14]]
## [1] 33.2
##
## [[15]]
## [1] 33.5
##
## [[16]]
## [1] 33.8
##
## [[17]]
## [1] 34
##
## [[18]]
## [1] 34
##
## [[19]]
## [1] 34.2
##
## [[20]]
## [1] 35
##
## [[21]]
## [1] 35.4
##
## [[22]]
## [1] 35.5
##
## [[23]]
## [1] 35.6
##
## [[24]]
## [1] 35.9
##
## [[25]]
## [1] 35.9
```

```
##
## [[26]]
## [1] 36
##
## [[27]]
## [1] 36
##
## [[28]]
## [1] 36.2
##
## [[29]]
## [1] 36.5
##
## [[30]]
## [1] 37.4
##
## [[31]]
## [1] 37.6
##
## [[32]]
## [1] 37.7
##
## [[33]]
## [1] 38.1
##
## [[34]]
## [1] 38.6
##
## [[35]]
## [1] 38.6
##
## [[36]]
## [1] 39
##
## [[37]]
## [1] 39.5
##
## [[38]]
## [1] 40.1
##
## [[39]]
## [1] 40.5
##
## [[40]]
## [1] 40.6
##
## [[41]]
## [1] 40.8
##
## [[42]]
## [1] 41
##
## [[43]]
## [1] 41
```

```
##
## [[44]]
## [1] 41.5
##
## [[45]]
## [1] 42.4
##
## [[46]]
## [1] 42.9
##
## [[47]]
## [1] 43.1
##
## [[48]]
## [1] 43.4
##
## [[49]]
## [1] 44.4
##
## [[50]]
## [1] 44.4
##
## [[51]]
## [1] 45.2
##
## [[52]]
## [1] 46
##
## [[53]]
## [1] 46.7
##
## [[54]]
## [1] 47.6
##
## [[55]]
## [1] 47.8
##
## [[56]]
## [1] 48.4
##
## [[57]]
## [1] 49.6
##
## [[58]]
## [1] 50.8
##
## [[59]]
## [1] 51.1
##
## [[60]]
## [1] 51.9
##
## [[61]]
## [1] 53.3
```

```
##
## [[62]]
## [1] 57.1
##
## [[63]]
## [1] 59.1
```

###Labeling variables and save as Stata library(labelled) #use ‘for variable names and "" for labels var_label(merged_df) <- list(country= "country"year= "year",inequality_gini= "inequality gini score",population= "population (inhabitants)",iso2c= "ISO-2 country code",employment_ratio‘ = “ratio of employment to population”)

```
#save the data frame as a Stata dataset
library(rio)
#export(merged_df, "final_data.dta")
```