Exam 2

Mary White

6/26/2020

```
#Load inequality dataset
library(rio)
inequality_data <- import("inequality.xlsx")
```

This dataset is cross-sectional because it only has data for one year, 2015.

```
#Show year is only 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
```

```
#Min and Max year are both 2015
```

```
#Using the subset command, provide theinequality_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
```

```
inequality_data_subset <- subset(inequality_data, select = c("country",
                                                             "inequality_gini"))
#filter subset for Denmark
inequality_data_subdenmark <- filter(inequality_data_subset, country == "Denmark")
print(inequality_data_subdenmark)
```

```
##   country inequality_gini
## 1 Denmark            28.2
```

```
#Filter subset for Sweden
inequality_data_subsweden <- filter(inequality_data_subset, country == "Sweden")
print(inequality_data_subsweden)
```

```
##   country inequality_gini
## 1  Sweden            29.2
```

```
#Filter subset for Brazil
inequality_data_subbrazil <- filter(inequality_data_subset, country == "Brazil")
print(inequality_data_subbrazil)
```

```
##   country inequality_gini
## 1  Brazil            51.9
```

It is better to have a lower inequality_gini score, as Brazil's is much higher than Denmark and Sweden's.

```
#Use head command to look at inequality_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
```

```
#Write accent remove function
accent.remove <- function(x)
{
  old1 <- "ú"
  new1 <- "u"
  x1 <- chartr(old1, new1, x)
}
inequality_data$country <- accent.remove(inequality_data$country)
#run head again to show change
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
```

```r
#Sort data by lowest inequality scores
inequality_data <- inequality_data[order(inequality_data$inequality_gini),]
#run head to show top 5 countries with lowest scores
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
```

```r
#Calculate mean inequality_gini score
mean(inequality_data$inequality_gini, na.rm = TRUE)
```

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

```r
#Create dummy variables for high inequality and low inequality
inequality_data$high_inequality = NA
inequality_data$high_inequality <- ifelse(inequality_data$inequality_gini > 36.81375,
                                          "1", "0")
inequality_data$low_inequality = NA
inequality_data$low_inequality <- ifelse(inequality_data$inequality_gini < 36.81375,
                                          "1", "0")

#Create a cross-tabulation btwn high and low inequality and inequality_gini
library(doBy)
```

```
##
## Attaching package: 'doBy'
```

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

```r
summaryBy(inequality_gini ~ (high_inequality & low_inequality), data = inequality_data,
          FUN = c(mean,length))
```

```
##   high_inequality low_inequality inequality_gini.mean inequality_gini.length
## 1               0              1             31.25870                     46
## 2               1              0             44.32941                     34
## 3            <NA>           <NA>                   NA                    123
```

```r
#for loop of institutions
orgs <- c('World Bank', 'African Development Bank',
          'Bill and Melinda Gates Foundation')
for(i in orgs){
  print(i)
}
```

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

For 14, I picked access to electricity as my variable because I thought that in countries with more inequality, smaller parts of the population will have access to electrcitity as a luxury item.

```
#import electricity access data from WDI into R
library(WDI)
electricity_access <- WDI(country = "all",
                          indicator = "EG.ELC.ACCS.ZS",
                          start = 2015, end = 2015, extra = FALSE, cache = NULL)
#rename variable
library(data.table)
```

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

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

```
setnames(electricity_access, "EG.ELC.ACCS.ZS", "electricity_access")
#merge data tables
merged_df <- left_join(x = inequality_data, y = electricity_access,
                       by = c("country", "year", "iso2c"))
#remove NA values on basis of inequality_gini and electricity_access
library(dplyr)
merged_df <-
    merged_df %>%
    dplyr::filter(!(inequality_gini == "NA"))
merged_df <-
    merged_df %>%
    dplyr::filter(!(electricity_access == "NA"))
#use filter to keep inequality_gini scores above 30
data_greater_30 <-
    merged_df %>%
    dplyr::filter((inequality_gini > 30))
#Using data_greater_30, use R to count how many countries have the sequence "ai"in their name.
grep("ai", data_greater_30)
```

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

```
#Use apply to take the sum of inequality_gini in data_greater_30
data_greater_30_a <- sapply(data_greater_30$inequality_gini, sum)
print(data_greater_30_a)
```

```
##  [1] 30.4 30.5 31.1 31.7 31.8 31.8 31.8 32.3 32.4 32.7 32.7 32.8 32.9 33.2 33.5
## [16] 33.8 34.0 34.0 34.2 35.0 35.4 35.5 35.6 35.9 35.9 36.0 36.0 36.2 36.5 37.4
## [31] 37.6 37.7 38.1 38.6 38.6 39.0 39.5 40.1 40.5 40.6 40.8 41.0 41.0 41.5 42.4
## [46] 42.9 43.1 43.4 44.4 44.4 45.2 46.0 46.7 47.6 47.8 48.4 49.6 50.8 51.1 51.9
## [61] 53.3 57.1 59.1
```

```r
#Label variables
library(labelled)
names(merged_df) #see variables u need to rename
```

```
## [1] "iso2c"              "country"            "inequality_gini"
## [4] "year"               "high_inequality"    "low_inequality"
## [7] "electricity_access"
```

```r
var_label(merged_df) <- list(`iso2c` = "ISO-2 Country Code",
                             `country` = "Country",
                             `inequality_gini` = "Inequality Score",
                             `year` = "Year",
                             `high_inequality` = "High Inequality Score",
                             `low_inequality` = "Low Inequality Score",
                             `electricity_access` = "% of Population with Electricity")
```

```r
#Save merged_df
library(rio)
export(merged_df, "final_data.dta")
```