Test 3

M. Schonhoeft

7/9/2020

## Git hub link

<https://github.com/mitchelly00/test_3>.

##### Q1 clean the enviroment

rm(list = ls())

##### Q2 tidycensus

In order to save time I did not put the function in rmarkdown to see the variable for the gini index since it takes a long time to load, here it is: v15<- load\_variables(year=2015, “acs5”)

The code is B19083\_001 for gini

# load the library  
library(tidycensus)

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

#load api  
census\_api\_key("279c9b589cc5452c32299f2c9e93deb5c9113c5d",  
 install=TRUE,overwrite = TRUE)

## Your original .Renviron will be backed up and stored in your R HOME directory if needed.

## Your API key has been stored in your .Renviron and can be accessed by Sys.getenv("CENSUS\_API\_KEY").   
## To use now, restart R or run `readRenviron("~/.Renviron")`

## [1] "279c9b589cc5452c32299f2c9e93deb5c9113c5d"

#getting the 2010 and 2015 datasets  
  
US\_2010 <- get\_acs(geography = "state",  
 variables = c(gini = c("B19083\_001")),   
 year = 2010)

## Getting data from the 2006-2010 5-year ACS

US\_2015 <- get\_acs(geography = "state",  
 variables = c(gini = c("B19083\_001")),   
 year = 2015)

## Getting data from the 2011-2015 5-year ACS

###### Q2 now adding them into a single panel dataset

#merging the data together  
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

inequaility\_panel = left\_join(x=US\_2010,  
 y=US\_2015,  
 by=c("GEOID","NAME"))  
  
#taking out extranious variables  
inequaility\_panel$variable.y <- NULL  
inequaility\_panel$variable.x <- NULL  
  
#changing names  
library(data.table)

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

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

setnames(inequaility\_panel,"estimate.x","gini\_2010")  
  
setnames(inequaility\_panel,"estimate.y","gini\_2015")  
setnames(inequaility\_panel,"NAME","state")  
  
#quick peak  
head(inequaility\_panel)

## # A tibble: 6 x 6  
## GEOID state gini\_2010 moe.x gini\_2015 moe.y  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 01 Alabama 0.47 0.003 0.475 0.0023  
## 2 02 Alaska 0.412 0.006 0.418 0.0062  
## 3 04 Arizona 0.453 0.002 0.465 0.0016  
## 4 05 Arkansas 0.459 0.003 0.470 0.0025  
## 5 06 California 0.469 0.001 0.486 0.0008  
## 6 08 Colorado 0.455 0.003 0.459 0.0018

#### Q3 wide data set

When I exported in the data it came as wide so I will keep it and re name it to fit the paramaters of this assignment. You usally use the pivot\_wider() function to make a long data set into a wide one.

inequaility\_wide<-inequaility\_panel  
  
head(inequaility\_panel)

## # A tibble: 6 x 6  
## GEOID state gini\_2010 moe.x gini\_2015 moe.y  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 01 Alabama 0.47 0.003 0.475 0.0023  
## 2 02 Alaska 0.412 0.006 0.418 0.0062  
## 3 04 Arizona 0.453 0.002 0.465 0.0016  
## 4 05 Arkansas 0.459 0.003 0.470 0.0025  
## 5 06 California 0.469 0.001 0.486 0.0008  
## 6 08 Colorado 0.455 0.003 0.459 0.0018

#### Q4 put into long format

library(tidyr)  
  
inequaility\_long<-  
 inequaility\_panel%>%  
 pivot\_longer(cols = starts\_with("gini"), # use columns starting with "year"   
 names\_to ="year", # name of new column   
 names\_prefix = "gini\_", # part of string to drop   
 values\_to = "gini", # where to put numeric values   
 values\_drop\_na = FALSE) %>% # don't drop NAs   
 filter(!(gini==0)) # drop observations with no disbursements

#### Q5 checking that they have the same number of observation

My data came in wide format and not in long, so the number of variables and objects will be different.

summary(inequaility\_long)

## GEOID state moe.x moe.y   
## Length:104 Length:104 Min. :0.001000 Min. :0.000800   
## Class :character Class :character 1st Qu.:0.002000 1st Qu.:0.001575   
## Mode :character Mode :character Median :0.003000 Median :0.002200   
## Mean :0.002923 Mean :0.002487   
## 3rd Qu.:0.003000 3rd Qu.:0.003175   
## Max. :0.007000 Max. :0.006200   
## year gini   
## Length:104 Min. :0.4120   
## Class :character 1st Qu.:0.4410   
## Mode :character Median :0.4585   
## Mean :0.4586   
## 3rd Qu.:0.4712   
## Max. :0.5432

summary(inequaility\_panel)

## GEOID state gini\_2010 moe.x   
## Length:52 Length:52 Min. :0.4120 Min. :0.001000   
## Class :character Class :character 1st Qu.:0.4368 1st Qu.:0.002000   
## Mode :character Mode :character Median :0.4535 Median :0.003000   
## Mean :0.4535 Mean :0.002923   
## 3rd Qu.:0.4650 3rd Qu.:0.003000   
## Max. :0.5380 Max. :0.007000   
## gini\_2015 moe.y   
## Min. :0.4181 Min. :0.000800   
## 1st Qu.:0.4482 1st Qu.:0.001575   
## Median :0.4626 Median :0.002200   
## Mean :0.4637 Mean :0.002487   
## 3rd Qu.:0.4762 3rd Qu.:0.003175   
## Max. :0.5432 Max. :0.006200

#### Q6 collapse data

collapsed\_collapsed<-  
 inequaility\_long %>%   
 group\_by(state,GEOID) %>%   
 summarise(across(where(is.numeric),mean))

## `summarise()` regrouping output by 'state' (override with `.groups` argument)

#### Q7 shape data

library(sf)

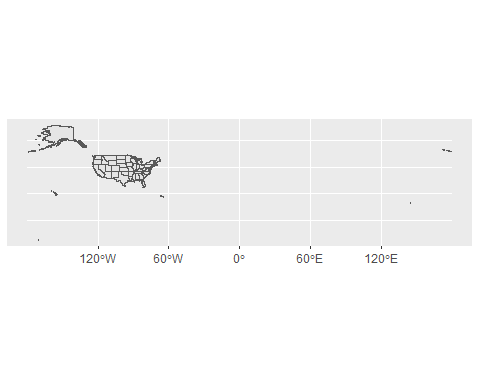
## Warning: package 'sf' was built under R version 4.0.2

## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1

library(ggplot2)  
us\_borders <- st\_read("C:/Users/schon/Downloads/Gov 355/test\_3/cb\_2018\_us\_state\_500k")

## Reading layer `cb\_2018\_us\_state\_500k' from data source `C:\Users\schon\Downloads\Gov 355\test\_3\cb\_2018\_us\_state\_500k' using driver `ESRI Shapefile'  
## Simple feature collection with 56 features and 9 fields  
## geometry type: MULTIPOLYGON  
## dimension: XY  
## bbox: xmin: -179.1489 ymin: -14.5487 xmax: 179.7785 ymax: 71.36516  
## geographic CRS: NAD83

state\_borders <- st\_transform(us\_borders,"+proj=latlong +ellps=WGS84 +datum=WGS84")  
rm(us\_borders)  
  
us\_basic = ggplot()+  
 geom\_sf(data=state\_borders)  
  
us\_basic



####q8 GDP

library(WDI)  
gdp<- WDI(country="all",indicator='NY.GDP.MKTP.CD',   
 start=2006, end=2007)  
library(data.table)   
setnames(gdp,"NY.GDP.MKTP.CD", "gdp\_current")

#### Q9 deflator

library(WDI)  
deflator\_data = WDI(country = "all", indicator = c("NY.GDP.DEFL.ZS"),  
 start = 2001,   
 end = 2017,   
 extra = FALSE, cache = NULL)

#rename variables  
library(data.table)   
setnames(deflator\_data,"NY.GDP.DEFL.ZS", "deflator")  
  
# select only the United States data   
usd\_deflator = subset(deflator\_data, country=="United States")  
subset(usd\_deflator, deflator==100)

## iso2c country deflator year  
## 4304 US United States 100 2015

#remove deflator  
rm(deflator\_data)  
  
#drop unecisary variables  
usd\_deflator$country <- NULL   
usd\_deflator$iso2c <- NULL  
  
#merging the data  
gdp\_deflated= left\_join(gdp, usd\_deflator, by=c("year"))  
  
#gdp\_deflated$deflated\_amount =   
 #gdp\_deflated$current\_amount/ (gdp\_deflated$deflator/100)   
  
#head(gdp\_deflated)

####Q10 Shiny app

####Q11 pull a pdf

library(pdftools)

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

## Using poppler version 0.73.0

armeniatext=pdf\_text(pdf = "https://pdf.usaid.gov/pdf\_docs/PA00TNMG.pdf")

####Q12 dataframe text

armeniatext=as.data.frame(armeniatext)   
armeniatext$page=c(1:65)   
colnames(armeniatext)[which(names(armeniatext) == "armeniatext")] <- "text" #change

####Q13 tokenize and get rid of stop words

library(tidytext)

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

armeniatext=armeniatext %>% unnest\_tokens(word, text)  
  
#in order to get rid of stop words:   
data(stop\_words)  
armeniatext <- armeniatext %>% anti\_join(stop\_words)

## Joining, by = "word"

####Q14 finding most common words

hpfreq <- armeniatext %>% count(word, sort = TRUE)   
head(hpfreq,5)

## word n  
## 1 armenia 252  
## 2 political 207  
## 3 corruption 186  
## 4 governance 185  
## 5 democracy 132

#### Q15 Billboard top 100 webpage

library(rvest)

## Loading required package: xml2

library(dplyr)  
hot100page <- "https://www.billboard.com/charts/hot-100"   
hot100 <- read\_html(hot100page)

#### Q16 nodes on page

body\_nodes <- hot100 %>%  
 html\_node("body") %>%  
 html\_children()   
body\_nodes

## {xml\_nodeset (36)}  
## [1] <div class="header-wrapper ">\n<header id="site-header" class="site-head ...  
## [2] <div class="site-header\_\_placeholder"></div>  
## [3] <script>\n var PGM = window.PGM || {};\n PGM.config = PGM. ...  
## [4] <div class="chart-piano-overlay\_\_attachment-point"></div>  
## [5] <main id="main" class="page-content"><div id="charts" data-page-title="T ...  
## [6] <div class="ad\_desktop dfp-ad dfp-ad-promo " data-position="promo" data- ...  
## [7] <div class="ad-container footerboard footerboard--bottom">\n <div cla ...  
## [8] <footer id="site-footer" class="site-footer"><div class="container foote ...  
## [9] <div class="biz-modal">\n <div class="biz-modal\_\_content">\n < ...  
## [10] <script>\n window.CLARITY = window.CLARITY || [];\n</script>  
## [11] <div class="ad\_clarity" data-out-of-page="true" style="display: none;">< ...  
## [12] <script>\n var darkMatterCMD = function() {\n this.darkMatterC ...  
## [13] <script src="https://www.billboard.com/assets/1593527595/js/vendors\_/art ...  
## [14] <script src="https://www.billboard.com/assets/1593527595/js/vendors\_/clo ...  
## [15] <script src="https://www.billboard.com/assets/1593527595/js/vendors\_/rea ...  
## [16] <script src="https://www.billboard.com/assets/1593527595/js/vendors\_/rea ...  
## [17] <script src="https://www.billboard.com/assets/1593527595/js/vendors\_/rea ...  
## [18] <script src="https://www.billboard.com/assets/1593527595/js/vendors\_/rea ...  
## [19] <script src="https://www.billboard.com/assets/1593527595/js/default\_/art ...  
## [20] <script src="https://www.billboard.com/assets/1593527595/js/default\_/rea ...  
## ...