Week 4 Supplement

BL Mosby

2/6/2021

r = getOption("repos")  
r["CRAN"] = "http://cran.us.r-project.org"  
options(repos = r)  
  
library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(haven)  
library(readxl)

## Correction to the Excel download from Async

Here are a few tweaks to the async code for the Excel file download that may run a bit smoother:

## Web page:   
##http://nces.ed.gov/programs/digest/d14/tables/dt14\_204.10.asp  
  
#If you have been getting an error with the download, add the mode argument (which directs RStudio to download as a binary file).  
if(file.exists("free.xls")==FALSE){  
 download.file("http://nces.ed.gov/programs/digest/d14/tables/xls/tabn204.10.xls",destfile="free.xls", mode="wb")  
free<-read\_excel("free.xls",skip=4,col\_names=FALSE)   
}else{  
 free<-read\_excel("free.xls",skip=4,col\_names=FALSE)  
}

## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## \* `` -> ...3  
## \* `` -> ...4  
## \* `` -> ...5  
## \* ...

head(free)

## # A tibble: 6 x 19  
## ...1 ...2 ...3 ...4 ...5 ...6 ...7 ...8 ...9 ...10 ...11 ...12  
## <chr> <dbl> <chr> <dbl> <dbl> <chr> <dbl> <dbl> <chr> <dbl> <dbl> <chr>  
## 1 Unit~ 4.66e7 "\\1~ 4.89e7 4.90e7 "\\1~ 4.91e7 1.78e7 "\\1~ 2.35e7 2.43e7 "\\1~  
## 2 Alab~ 7.28e5 <NA> 7.30e5 7.32e5 <NA> 7.40e5 3.35e5 <NA> 4.02e5 4.20e5 <NA>  
## 3 Alas~ 1.05e5 <NA> 1.32e5 1.31e5 <NA> 1.31e5 3.25e4 <NA> 5.07e4 5.32e4 <NA>  
## 4 Ariz~ 8.78e5 "\\2~ 1.07e6 1.02e6 <NA> 9.90e5 2.74e5 "\\2~ 4.82e5 5.12e5 <NA>  
## 5 Arka~ 4.50e5 <NA> 4.82e5 4.83e5 <NA> 4.86e5 2.05e5 <NA> 2.92e5 2.94e5 <NA>  
## 6 Cali~ 6.05e6 <NA> 6.17e6 6.20e6 "\\2~ 6.18e6 2.82e6 <NA> 3.34e6 3.35e6 "\\2~  
## # ... with 7 more variables: ...13 <dbl>, ...14 <dbl>, ...15 <chr>,  
## # ...16 <dbl>, ...17 <dbl>, ...18 <chr>, ...19 <dbl>

Next, to get rid of the empty rows and columns, you may need to adjust the column names to match what’s in your dataframe.

# Now need to clean up   
#Get rid of unwanted columns  
  
free2<-free[ ,-(c(3,6,9,12,15,18))]  
  
#Get rid of unwanted rows-- My column titles are "...1" so I will change the is.na argument to reflect that:  
free2<-free2%>%filter(is.na(...1)==FALSE)

The rest of the code chunks for the Excel download should work as is.

## Zipped Files

Some data files that are very large are compressed into a zipped file (along with codebooks or other metadata files). These take a two step process, where you must also know the name of the data file within the zipped folder.

#creating a temporary file to hold the zipped folder  
temp <- tempfile()  
  
#downloading the zipped folder from the data source webpage, and unzipping the SPSS data file  
download.file("https://gss.norc.org/Documents/spss/GSS\_2014\_Merged\_SPSS.zip", temp)  
unzip(temp, "gss2014merged\_r10 - spss.sav")  
  
#reading in the spss command and closing the connection  
gss14<-read\_spss("gss2014merged\_r10 - spss.sav")  
unlink(temp)  
  
head(gss14)

## # A tibble: 6 x 1,001  
## year id wrkstat hrs1 hrs2 evwork wrkslf wrkgovt OCC10 prestg10  
## <dbl> <dbl> <dbl+l> <dbl> <dbl> <dbl+lb> <dbl+l> <dbl+l> <dbl+lbl> <dbl+lb>  
## 1 2014 1 1 [WOR~ 60 NA NA 1 [SEL~ 2 [PRI~ 220 [Con~ 53  
## 2 2014 2 1 [WOR~ 40 NA NA 2 [SOM~ 2 [PRI~ 700 [Log~ 51  
## 3 2014 3 4 [UNE~ NA NA NA 2 [SOM~ 2 [PRI~ 8740 [Ins~ 39  
## 4 2014 4 2 [WOR~ 20 NA NA 2 [SOM~ 2 [PRI~ 50 [Mar~ 53  
## 5 2014 5 5 [RET~ NA NA 1 [YES] 2 [SOM~ 1 [GOV~ 800 [Acc~ 60  
## 6 2014 6 1 [WOR~ 60 NA NA 1 [SEL~ 2 [PRI~ 5000 [Fir~ 46  
## # ... with 991 more variables: prestg105plus <dbl+lbl>, INDUS10 <dbl+lbl>,  
## # marital <dbl+lbl>, martype <dbl+lbl>, divorce <dbl+lbl>, widowed <dbl+lbl>,  
## # spwrksta <dbl+lbl>, sphrs1 <dbl+lbl>, sphrs2 <dbl+lbl>, spevwork <dbl+lbl>,  
## # spwrkslf <dbl+lbl>, SPOCC10 <dbl+lbl>, sppres10 <dbl+lbl>,  
## # sppres105plus <dbl+lbl>, SPIND10 <dbl+lbl>, pawrkslf <dbl+lbl>,  
## # PAOCC10 <dbl+lbl>, papres10 <dbl+lbl>, papres105plus <dbl+lbl>,  
## # PAIND10 <dbl+lbl>, mawrkslf <dbl+lbl>, MAOCC10 <dbl+lbl>,  
## # mapres10 <dbl+lbl>, mapres105plus <dbl+lbl>, MAIND10 <dbl+lbl>,  
## # sibs <dbl+lbl>, childs <dbl+lbl>, age <dbl+lbl>, agekdbrn <dbl+lbl>,  
## # educ <dbl+lbl>, paeduc <dbl+lbl>, maeduc <dbl+lbl>, speduc <dbl+lbl>,  
## # degree <dbl+lbl>, padeg <dbl+lbl>, madeg <dbl+lbl>, spdeg <dbl+lbl>,  
## # major1 <dbl+lbl>, major2 <dbl+lbl>, DIPGED <dbl+lbl>, SPDIPGED <dbl+lbl>,  
## # whenhs <dbl+lbl>, whencol <dbl+lbl>, sector <dbl+lbl>, eftotlt <dbl+lbl>,  
## # barate <dbl+lbl>, gradtounder <dbl+lbl>, VOEDCOL <dbl+lbl>,  
## # VOEDNME1 <dbl+lbl>, VOEDNCOL <dbl+lbl>, VOEDNME2 <dbl+lbl>,  
## # spsector <dbl+lbl>, speftotlt <dbl+lbl>, spbarate <dbl+lbl>,  
## # spgradtounder <dbl+lbl>, sex <dbl+lbl>, race <dbl+lbl>, res16 <dbl+lbl>,  
## # reg16 <dbl+lbl>, mobile16 <dbl+lbl>, family16 <dbl+lbl>,  
## # famdif16 <dbl+lbl>, mawrkgrw <dbl+lbl>, incom16 <dbl+lbl>, born <dbl+lbl>,  
## # parborn <dbl+lbl>, granborn <dbl+lbl>, hompop <dbl+lbl>, babies <dbl+lbl>,  
## # preteen <dbl+lbl>, teens <dbl+lbl>, adults <dbl+lbl>, unrelat <dbl+lbl>,  
## # earnrs <dbl+lbl>, income <dbl+lbl>, rincome <dbl+lbl>, income06 <dbl+lbl>,  
## # rincom06 <dbl+lbl>, region <dbl+lbl>, xnorcsiz <dbl+lbl>,  
## # srcbelt <dbl+lbl>, size <dbl+lbl>, partyid <dbl+lbl>, VOTE08 <dbl+lbl>,  
## # PRES08 <dbl+lbl>, IF08WHO <dbl+lbl>, VOTE12 <dbl+lbl>, PRES12 <dbl+lbl>,  
## # IF12WHO <dbl+lbl>, polviews <dbl+lbl>, natspac <dbl+lbl>,  
## # natenvir <dbl+lbl>, natheal <dbl+lbl>, natcity <dbl+lbl>,  
## # natcrime <dbl+lbl>, natdrug <dbl+lbl>, nateduc <dbl+lbl>,  
## # natrace <dbl+lbl>, natarms <dbl+lbl>, nataid <dbl+lbl>, ...

## Homework 4 Exercises

For Assignment 4, replace #2 and #3 with the following:

1. U.S. National Election Survey data from 2004. Save as file name nes.RData. [https://osf.io/zpuhd/download](%22https://osf.io/zpuhd/download%22)
2. General Social Survey (GSS), pick any of the individual year datasets (1972 - 2018). Save as file name gss<year>.Rdata. Note this will be an SPSS file within a zipped file, and you will need to use the code above. <https://gss.norc.org/get-the-data/spss>.

## Extra Excel data

My code for the additional Excel dataset from async. (Note, there is more than one correct way to go about cleaning the data.)

## Web page:   
##http://nces.ed.gov/programs/digest/d14/tables/dt14\_204.10.asp  
  
download.file("http://nces.ed.gov/programs/digest/d14/tables/xls/tabn302.10.xls",destfile="newxl.xls", mode="wb")  
  
newxl<-read\_excel("newxl.xls",skip=6,col\_names=FALSE)

## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## \* `` -> ...3  
## \* `` -> ...4  
## \* `` -> ...5  
## \* ...

#Get rid of empty rows and unneeded columns.  
newxl<-newxl%>%filter(is.na(...1)==FALSE)  
  
newxl<-newxl[1:54,]  
newxl<-newxl[ ,-(c(3,5,7,9,11,13,15,17,19,21,23,25))]  
  
  
#Add useful names to the columns.  
names(newxl)<-c("year", "hscomptotal", "hscompmales","hscompfemales","incollegetotal","total2y","total4y", "collegetotalmales", "males2y", "males4y", "collegetotalfemales", "females2y", "females4y")  
  
  
#Change the year variable into a format that R can understand and drop the old one.  
yeardate = seq(from = as.Date("1960-01-01"), to = as.Date("2013-01-01"), by = 'year')  
  
newxl2<-cbind(yeardate, newxl)  
newxl2<-newxl2[,-2]  
  
head(newxl2)

## yeardate hscomptotal hscompmales hscompfemales incollegetotal total2y  
## 1 1960-01-01 1679 756 923 45.14592 ---  
## 2 1961-01-01 1763 790 973 48.04311 ---  
## 3 1962-01-01 1838 872 966 48.96627 ---  
## 4 1963-01-01 1741 794 947 45.03159 ---  
## 5 1964-01-01 2145 997 1148 48.34499 ---  
## 6 1965-01-01 2659 1254 1405 50.92140 ---  
## total4y collegetotalmales males2y males4y collegetotalfemales females2y  
## 1 --- 53.96825 --- --- 37.91983 ---  
## 2 --- 56.32911 --- --- 41.31552 ---  
## 3 --- 55.04587 --- --- 43.47826 ---  
## 4 --- 52.26700 --- --- 38.96515 ---  
## 5 --- 57.17151 --- --- 40.67944 ---  
## 6 --- 57.25678 --- --- 45.26690 ---  
## females4y  
## 1 ---  
## 2 ---  
## 3 ---  
## 4 ---  
## 5 ---  
## 6 ---

This was just a graph I did for fun, to visualize one of the variables.

gg<-ggplot(newxl2, aes(x=yeardate, y=incollegetotal))  
gg<-gg+geom\_point()  
gg

