morales-exam3

Laura Morales

7/8/2021

## Question 1: Clear environment

#clear environment   
rm(list = ls(all=TRUE))

## Question 2: load WDI package and female labor force participation

#load all necessary libraries   
library(rio)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.2 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.1

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

library(googlesheets4)  
library(labelled)  
library(data.table)

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

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

## The following object is masked from 'package:purrr':  
##   
## transpose

library(varhandle)  
library(ggrepel)  
library(geosphere)  
library(rgeos)

## Loading required package: sp

## rgeos version: 0.5-5, (SVN revision 640)  
## GEOS runtime version: 3.8.1-CAPI-1.13.3   
## Linking to sp version: 1.4-2   
## Polygon checking: TRUE

library(viridis)

## Loading required package: viridisLite

library(mapview)  
library(rnaturalearth)  
library(rnaturalearthdata)  
library(devtools)

## Loading required package: usethis

devtools::install\_github("ropensci/rnaturalearthhires")

## Skipping install of 'rnaturalearthhires' from a github remote, the SHA1 (2ed7a937) has not changed since last install.  
## Use `force = TRUE` to force installation

library(rnaturalearthhires)  
library(raster)

##   
## Attaching package: 'raster'

## The following object is masked from 'package:data.table':  
##   
## shift

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

## The following object is masked from 'package:tidyr':  
##   
## extract

library(sp)  
library(sf)

## Linking to GEOS 3.8.1, GDAL 3.2.1, PROJ 7.2.1

devtools::install\_github("yutannihilation/ggsflabel")

## Skipping install of 'ggsflabel' from a github remote, the SHA1 (a489481b) has not changed since last install.  
## Use `force = TRUE` to force installation

library(ggsflabel)

##   
## Attaching package: 'ggsflabel'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_sf\_label, geom\_sf\_text, StatSfCoordinates

library(Imap)

##   
## Attaching package: 'Imap'

## The following object is masked from 'package:purrr':  
##   
## imap

library(ggplot2)  
library(dplyr)  
library(WDI)  
library(countrycode)  
#load the WDi data  
female\_lfp = WDI(country = "all", indicator = c("SL.TLF.CACT.FE.ZS"), start = 2010, end = 2015, cache = NULL)

## Question 3: rename variable

#rename the variable  
female\_lfp <- rename(female\_lfp, "flfp"="SL.TLF.CACT.FE.ZS")

## Question 4: collapse female\_lfpby the mean value forflfp

#collapse female\_lfpby the mean value forflfp   
collapsed\_flfp= female\_lfp %>% group\_by(country, iso2c) %>% summarise(flfp = mean(flfp), na.rm=TRUE)

## `summarise()` has grouped output by 'country'. You can override using the `.groups` argument.

## Question 5: countries which have an average female force participation rates less than 15%

#filter those over 15%  
less\_15 <- filter(collapsed\_flfp, flfp < 15 )

## Question 6: Map of the world of usingcollapsed\_flfp, using the viridis colorschem

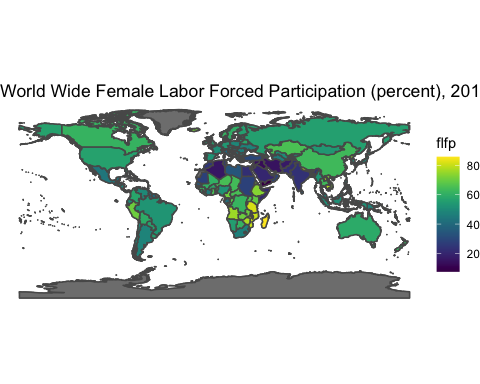
#load world borders  
world\_borders <- st\_read("/Users/lauram./Documents/Gov 355M Data Sci for Soc/Mods/Mod 11/world border shape files/World\_Borders.shp")

## Reading layer `World\_Borders' from data source   
## `/Users/lauram./Documents/Gov 355M Data Sci for Soc/Mods/Mod 11/world border shape files/World\_Borders.shp'   
## using driver `ESRI Shapefile'  
## Simple feature collection with 246 features and 11 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: -180 ymin: -90 xmax: 180 ymax: 83.6236  
## Geodetic CRS: WGS 84

#transform to WGS84 projection  
borders = st\_transform(world\_borders, "+proj=latlong +ellps=WGS84 +datum=WGS84")  
rm(world\_borders)  
#merege the codes and coordinates  
collapsed\_flfp$ISO3= countrycode(sourcevar = collapsed\_flfp$country, origin = "country.name", destination="iso3c", warn=TRUE)

## Warning in countrycode(sourcevar = collapsed\_flfp$country, origin = "country.name", : Some values were not matched unambiguously: Africa Eastern and Southern, Africa Western and Central, Arab World, Caribbean small states, Central Europe and the Baltics, Channel Islands, Early-demographic dividend, East Asia & Pacific, East Asia & Pacific (excluding high income), East Asia & Pacific (IDA & IBRD countries), Euro area, Europe & Central Asia, Europe & Central Asia (excluding high income), Europe & Central Asia (IDA & IBRD countries), European Union, Fragile and conflict affected situations, Heavily indebted poor countries (HIPC), High income, IBRD only, IDA & IBRD total, IDA blend, IDA only, IDA total, Kosovo, Late-demographic dividend, Latin America & Caribbean, Latin America & Caribbean (excluding high income), Latin America & the Caribbean (IDA & IBRD countries), Least developed countries: UN classification, Low & middle income, Low income, Lower middle income, Middle East & North Africa, Middle East & North Africa (excluding high income), Middle East & North Africa (IDA & IBRD countries), Middle income, North America, Not classified, OECD members, Other small states, Pacific island small states, Post-demographic dividend, Pre-demographic dividend, Small states, South Asia, South Asia (IDA & IBRD), Sub-Saharan Africa, Sub-Saharan Africa (excluding high income), Sub-Saharan Africa (IDA & IBRD countries), Upper middle income, World

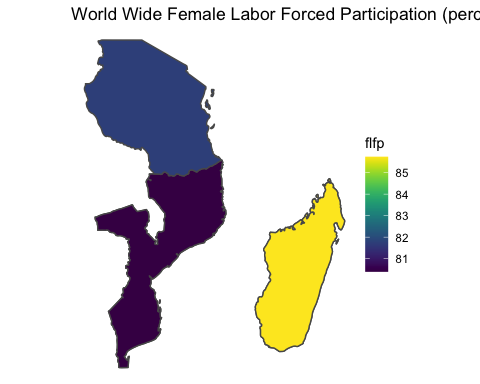
collapsed\_flfp\_small = na.omit(collapsed\_flfp, select("country", "flfp", "ISO3"))  
merged\_data= left\_join(borders, collapsed\_flfp\_small, by="ISO3")  
#create the map  
flfp\_map= ggplot()+  
 geom\_sf(data=borders)+  
 geom\_sf(data=merged\_data, aes(fill=flfp))+  
 scale\_fill\_viridis(option="viridis")+  
 ggtitle("World Wide Female Labor Forced Participation (percent), 2010-2015")+  
 theme(plot.title= element\_text(hjust = 0.5))+  
 theme\_void()  
print(flfp\_map)

 ## Question 7: Based on the map, which area of the world has, perhaps surprisingly, a cluster of yellow-colored average female labor force participation rate states, indicating the highest onthe scale?

I thought it was surprising that south easter africa has a very high female labor participation.

## Question 8: Use R to show the same cluster of states referenced in the previous question.

#subset data to only referenced countries  
sa\_fl <- subset(merged\_data, country=="Madagascar"| country=="Mozambique"| country=="Tanzania")  
sa\_fl\_map <- subset(borders, NAME=="Madagascar"| NAME=="Mozambique"| NAME=="Tanzania")  
#plot the data  
flfp\_map= ggplot()+  
 geom\_sf(data=sa\_fl\_map )+  
 geom\_sf(data=sa\_fl, aes(fill=flfp))+  
 scale\_fill\_viridis(option="viridis")+  
 ggtitle("World Wide Female Labor Forced Participation (percent), 2010-2015")+  
 theme(plot.title= element\_text(hjust = 0.3))+  
 theme\_void()  
print(flfp\_map)

 ## Question 9: In a Shiny app, what are the three main components and their subcomponents? The app has a user interface object, a server function, and a call to the shinyApp function.

## Question 10: Pull this.pdffile from Mike Denly’s webpage. It is a report that Mike Denly and MikeFindley prepared for the US Agency for International Development (USAID).

#load libraries  
library(pdftools)

## Using poppler version 20.12.1

library(tidyr)  
library(tidytext)  
library(dplyr)  
library(stringr)  
#retrive pdf  
armeniatext= pdf\_text(pdf="https://pdf.usaid.gov/pdf\_docs/PA00TNMJ.pdf")

## Question 11: Convert the text pulled from this.pdffile to a data frame, using the, stringsAsFactors=FALSE option. Call the data frame armeniatext

#convert text to data frame  
armeniatext = as.data.frame(armeniatext, stringsAsFactors=FALSE)  
armeniatext$page=c(1:59)  
colnames(armeniatext)[which(names(armeniatext)=="armeniatext")] <- "text"

## Question 12: Tokenize the data by word and then remove stop words.

#tokenize text and remove stop words  
armeniatext <- armeniatext %>% unnest\_tokens(word, text)  
data(stop\_words)  
armeniatext <- armeniatext %>% anti\_join(stop\_words)

## Joining, by = "word"

## Question 13: Figure out the top 5 most used word in the report

#count frequency of words  
hpfreq <- armeniatext %>% count(word, sort= TRUE)  
head(hpfreq)

## word n  
## 1 law 276  
## 2 corruption 242  
## 3 rule 206  
## 4 armenia 195  
## 5 european 105  
## 6 political 102

## Question 14: Load the Billboard Hot 100 webpage, which we explored in the course modules. Namethe list object:hot100exam

#load libraries  
library(rvest)

##   
## Attaching package: 'rvest'

## The following object is masked from 'package:readr':  
##   
## guess\_encoding

library(dplyr)  
library(ggplot2)  
#load webpage  
hot100exam <- "https://www.billboard.com/charts/hot-100"  
hot100 <- read\_html(hot100exam)

## Question : Use rvest to obtain identify all of the nodes in the webpage.

#declare full (enough) struture to r  
body\_nodes <- hot100 %>%   
 html\_node('body') %>%   
 html\_children()  
body\_nodes

## {xml\_nodeset (37)}  
## [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] <main id="main" class="page-content"><div id="charts" data-page-title="T ...  
## [5] <div class="ad\_desktop dfp-ad dfp-ad-promo " data-position="promo" data- ...  
## [6] <div class="ad-container footerboard footerboard--bottom">\n <div cla ...  
## [7] <footer id="site-footer" class="site-footer"><div class="container foote ...  
## [8] <div class="biz-modal">\n <div class="biz-modal\_\_content">\n < ...  
## [9] <script>\n window.CLARITY = window.CLARITY || [];\n</script>  
## [10] <div class="ad\_clarity" data-out-of-page="true" style="display: none;">< ...  
## [11] <script>\n\n window.top.pageLevelKeys = {};\n window.top.pageAdZon ...  
## [12] <script type="text/javascript" async="async" data-cfasync="false" src="h ...  
## [13] <script type="text/javascript">\n let detectDevice = function() {\n ...  
## [14] <script src="https://cdn.cookielaw.org/opt-out/otCCPAiab.js" type="text/ ...  
## [15] <script>\n\n function loadEUScript(source, attributes = {}) {\n\n ...  
## [16] <script src="https://geolocation.onetrust.com/cookieconsentpub/v1/geo/lo ...  
## [17] <script src="https://www.billboard.com/assets/1624920239/js/vendors\_/art ...  
## [18] <script src="https://www.billboard.com/assets/1624920239/js/vendors\_/clo ...  
## [19] <script src="https://www.billboard.com/assets/1624920239/js/vendors\_/rea ...  
## [20] <script src="https://www.billboard.com/assets/1624920239/js/vendors\_/rea ...  
## ...

body\_nodes %>%   
 html\_children()

## {xml\_nodeset (9)}  
## [1] <header id="site-header" class="site-header " role="banner"><div class="s ...  
## [2] <div class="header-wrapper\_\_secondary-header">\n<nav class="site-header-l ...  
## [3] <div id="charts" data-page-title="THE HOT 100" data-chart-code="HSI" data ...  
## [4] <div class="footerboard-wrapper">\n <div class="ad\_desktop\_placeho ...  
## [5] <div class="container footer-content">\n\t\t\t\t\t<div class="cover-image ...  
## [6] <div class="container">\n\t\t<p class="copyright\_\_paragraph">© 2021 Billb ...  
## [7] <div class="container">\n\t\t<p class="station-identification">\n\t\t\tBI ...  
## [8] <div class="container">\n\t\t\n\n\n <div class="ad\_desktop dfp-ad dfp- ...  
## [9] <div class="biz-modal\_\_content">\n <button class="biz-modal\_\_close ...

## Question 16: Use Google Chrome developer to identify the necessary tags and pull the data onRank,Artist,Title, andLast Week. HINT 1: In class we showed you how to get the first threeof these. You simply need to add theLast Weekranking. HINT 2: You can navigatetwo ways. Hovering to find what you need or by doingCmd+F / Ctrl+Fand usingactual data to find the location. HINT 3: You’re looking to update the code based ontheway the information is in referenced. Try out some different options and see whatshows up in the environment. Keep trying until you see that you have achr [1:100]with values that correspond to what is in the web page.

#pull out specific data form the webpage  
#rank, artists, title, Last week  
rank <- hot100 %>%   
 rvest::html\_nodes('body') %>%   
 xml2::xml\_find\_all("//span[contains(@class,  
 'chart-element\_\_rank\_\_number')]") %>%   
 rvest::html\_text()  
  
artist <- hot100 %>%   
 rvest::html\_nodes('body') %>%   
 xml2::xml\_find\_all("//span[contains(@class, 'chart-element\_\_information\_\_artist')]") %>%  
 rvest::html\_text()  
  
title <- hot100 %>%   
 rvest::html\_nodes('body') %>%   
 xml2::xml\_find\_all("//span[contains(@class,  
 'chart-element\_\_information\_\_song')]") %>%   
 rvest::html\_text()  
  
 last\_week <- hot100 %>%   
 rvest::html\_nodes('body') %>%   
 xml2::xml\_find\_all("//span[contains(@class,  
 'chart-element\_\_meta text--center color--secondary text--last')]") %>%   
 rvest::html\_text()

## Question 17: Save all of the files (i.e..Rmd,.dta,.pdf/Word Doc), push them to your GitHub repo,and provide us with the link to that repo.