Web Scraping Using API HW

Isabella Natke

4/5/2021

library(tidycensus)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

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

library(dplyr)  
library(ggplot2)  
  
# 1) Show and use a census API key that gives you access to the ACS data. Do not use my API key, use and show your own key.  
census\_api\_key("aee839cc2dc7c267c7b9f1ea763e68333400686b")

## To install your API key for use in future sessions, run this function with `install = TRUE`.

install = TRUE  
  
# 2) Using ACS census data from 2015, show and use R code to do the following:  
# a) Produce a tibble that shows the median income estimates and the margin of errors for white males who are 35-44 in the counties of California.   
# The required variable code starts with the characters BO1OO1. Use the table to find the other characters.  
v2 <- load\_variables(2015, "acs5", cache = TRUE)  
  
ca\_35to44 <- get\_acs(geography = "county",   
 variables = c(medincome = "B01001A\_011"),   
 state = "CA",   
 year = 2015)

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

ca\_35to44

## # A tibble: 58 x 5  
## GEOID NAME variable estimate moe  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 06001 Alameda County, California medincome 51644 667  
## 2 06003 Alpine County, California medincome 50 26  
## 3 06005 Amador County, California medincome 1809 72  
## 4 06007 Butte County, California medincome 9962 128  
## 5 06009 Calaveras County, California medincome 1927 74  
## 6 06011 Colusa County, California medincome 1147 79  
## 7 06013 Contra Costa County, California medincome 42756 605  
## 8 06015 Del Norte County, California medincome 1629 90  
## 9 06017 El Dorado County, California medincome 8609 141  
## 10 06019 Fresno County, California medincome 34979 714  
## # … with 48 more rows

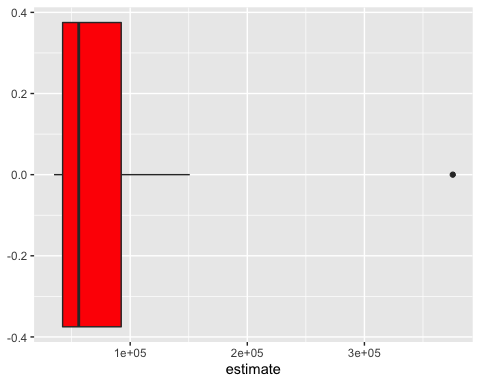
# b) Use a dplyr functions to change your table of part a so that it reflects estimates that are greater than $30,00 dollars and list the estimates in descending order.  
  
ca\_35to44%>%  
 filter(estimate > 30000)%>%  
 arrange(desc(estimate)) -> ca2  
ca2

## # A tibble: 13 x 5  
## GEOID NAME variable estimate moe  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 06037 Los Angeles County, California medincome 375435 2332  
## 2 06073 San Diego County, California medincome 150891 1008  
## 3 06059 Orange County, California medincome 126819 1152  
## 4 06065 Riverside County, California medincome 92346 1004  
## 5 06071 San Bernardino County, California medincome 80925 1160  
## 6 06085 Santa Clara County, California medincome 63036 879  
## 7 06067 Sacramento County, California medincome 56066 553  
## 8 06001 Alameda County, California medincome 51644 667  
## 9 06013 Contra Costa County, California medincome 42756 605  
## 10 06075 San Francisco County, California medincome 42307 542  
## 11 06029 Kern County, California medincome 42121 575  
## 12 06111 Ventura County, California medincome 41155 557  
## 13 06019 Fresno County, California medincome 34979 714

# c) Using the tibble that you produced in part b, use and show R code that will show the county that has a median income estimate of 51644 and a margin of error of 667.  
   
ca2%>%  
 filter(estimate == 51644, moe == 667)

## # A tibble: 1 x 5  
## GEOID NAME variable estimate moe  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 06001 Alameda County, California medincome 51644 667

# Alameda County  
  
# d) Use and show ggplot coding that will produce the following boxplot for the data that you generated for part b.  
  
ggplot(data = ca2) +  
 geom\_boxplot(mapping = aes(y=estimate), fill = "red") +  
 coord\_flip()



# e) Use and show R code that will produce the following graph for the data generated in part b  
ca2%>%  
 mutate(NAME = gsub("County, California", "", NAME)) %>%  
 ggplot(aes(x = estimate, y = reorder(NAME, estimate))) +  
 geom\_errorbarh(aes(xmin = estimate - moe, xmax = estimate + moe)) +  
 geom\_point(color = "blue", size = 3) +  
 labs(title = "Median Income for White Males by County",  
 subtitle = "2014-2018 American Community Survey",  
 y = "",  
 x = "ACS estimate (bars represent margin of error)")

