DS 613 Homework 9 Web Scraping using an API

Xuehao Wang

04/07/2021

Instructions: Generate an Rmardown file and a Word file that shows all required code and output. Push your files to GitHub Repository. 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.

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.3

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

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.3  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.3

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

library(tidycensus)

## Warning: package 'tidycensus' was built under R version 4.0.3

library(keyring)

## Warning: package 'keyring' was built under R version 4.0.3

#key\_set("CENSUS\_API\_KEY\_SECURE")  
#my\_census\_key <- key\_get("CENSUS\_API\_KEY\_SECURE")

1. Using ACS census data from 2015, show and use R code to do the following:
2. Produce a tibble that shows the median income estimates and the margin of errors for white males who are 35 to 44 years old in the counties of California. The required variable code starts with the characters BO1OO1. Use the table to find the other characters. The first five rows of your tibble are provided below:

v15 <- load\_variables(2015, "acs5", cache = TRUE)  
  
median\_income\_2015 <- get\_acs(geography = "county",   
 variables = c(medincome = "B01001A\_011"),   
 state = "CA",   
 year = 2015,  
 key = Sys.getenv("CENSUS\_API\_KEY\_SECURE"))

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

head(median\_income\_2015, n=5)

## # A tibble: 5 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

1. 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.

median\_income\_2015 %>%   
 filter(estimate > 3000) %>%   
 arrange(desc(estimate)) -> median\_income\_above3000

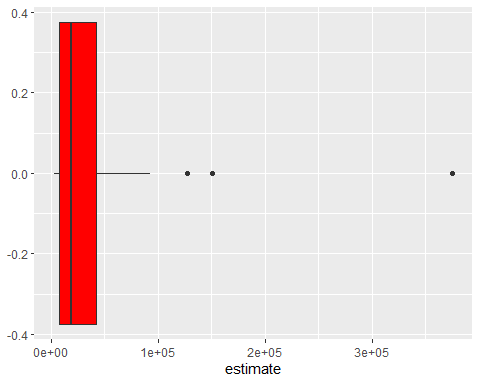
1. Using the tibble that you produced in part b, use and show R code that will show the county that has a median income estimate of 51644 and a margin of error of 667.

median\_income\_above3000 %>%   
 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

1. Use and show ggplot coding that will produce the following boxplot for the data that you generated for part b.

median\_income\_above3000 %>%   
 ggplot(aes(x = estimate))+  
 geom\_boxplot(fill = "red")



1. Use and show R code that will produce the following graph for the data generated in part b

median\_income\_above3000 %>%  
 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)")

