Homework week 48

Assignment 2: Practice Web Scraping

Find a tutorial with instructions on web scraping data on US police killings in this repository at https://github.com/adivea/KilledbyPolice2020.git. Clone it.

I chose the following assigment: *adapt the webscraping example to scrape homicide data from FBI site and produce a meaningful report on how homicide trends evolve around US in relation to this urban unrest.*

Thoughts

I was a bit unsure what the aim for this assignment was and what webpage to scrape data from. I chose this wikipage: https://en.wikipedia.org/wiki/List_of_U.S._states_by_homicide_rate

The webpage contains a list of U.S. states by homicides rate according to FBI Uniform Crime Reports with data from the years 1996, 2000, 2005, 2010, 2014, 2017 and 2018. The states are numbered from 1 to 50 sorted alphabetically.

Answer

I adapted Adela's example to fit my webpage of choice. I managed to scrape and print the date. That resulted in the table shown in figure 1, which shows numbers of homicides committed per 100.000 inhabitants.

Figure 1: A print of the table from the webpage

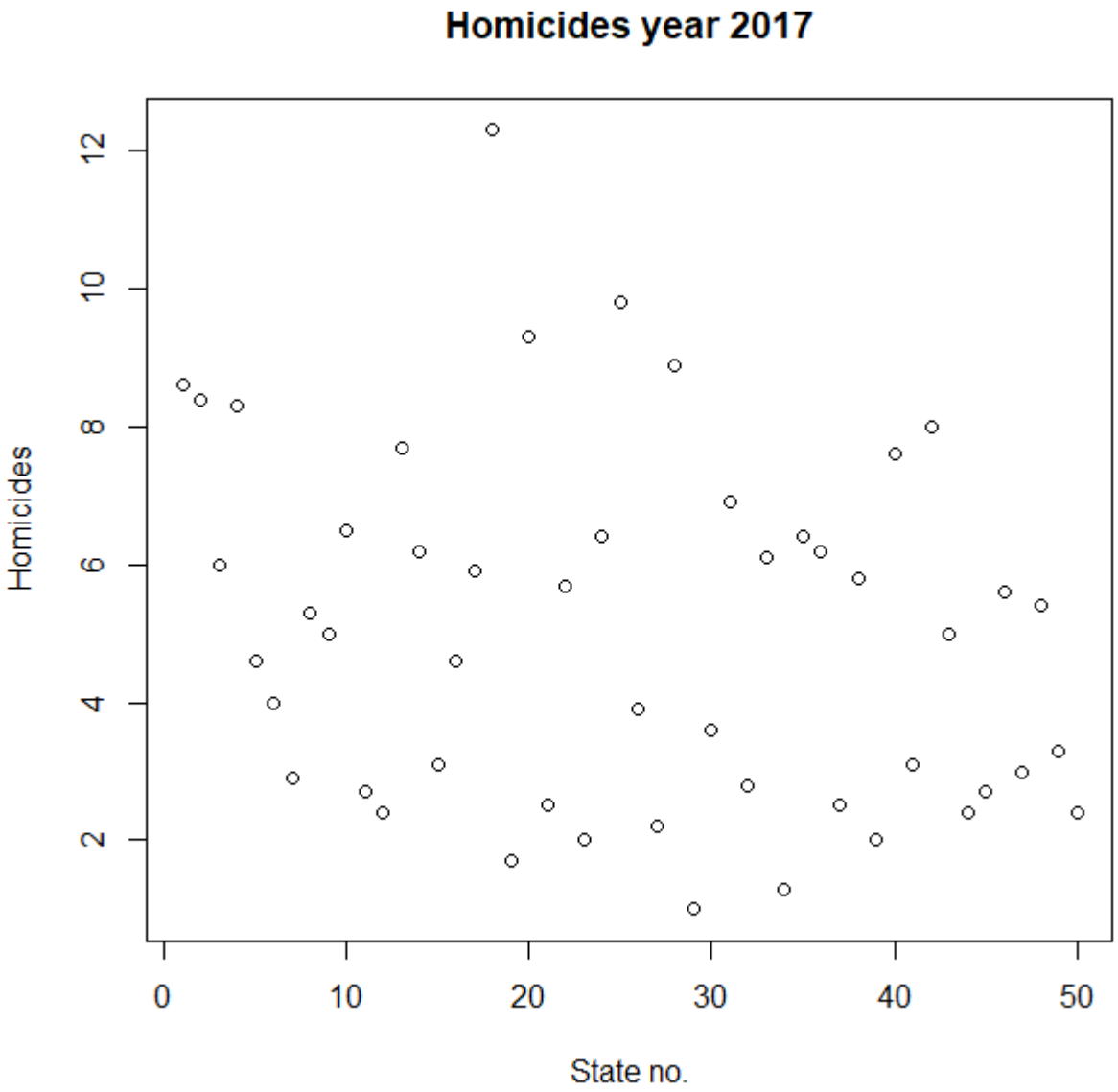
```
> web_table
[[1]]
           State 2018 2017 2014 2010 2005 2000 1996
1        Alabama  7.8  8.6  5.7  5.7  8.2  7.4 10.4
2         Alaska  8.2  8.4  5.6  4.3  4.8  4.3  7.4
3        Arizona  5.1  6.0  4.7  6.4  7.5  7.0  8.5
4       Arkansas  7.2  8.3  5.6  4.6  6.7  6.3  8.7
5     California  4.4  4.6  4.4  4.8  6.9  6.1  9.1
6       Colorado  3.7  4.0  2.8  2.6  3.7  3.1  4.7
7    Connecticut  2.3  2.9  2.4  3.7  2.9  2.9  4.8
8       Delaware  5.0  5.3  5.8  5.7  4.4 12.5  6.4
9        Florida  5.2  5.0  5.8  5.2  5.0  5.6  7.5
10       Georgia  6.1  6.5  5.7  5.7  6.2  8.0  9.5
11        Hawaii  2.5  2.7  1.8  1.8  1.9  2.9  3.4
12         Idaho  2.0  2.4  2.0  1.4  2.4  1.2  3.6
13      Illinois  6.9  7.7  5.3  5.5  6.0  7.2 10.0
14       Indiana  6.5  6.2  5.0  4.1  5.7  5.8  7.2
15          Iowa  1.7  3.1  1.9  1.2  1.3  1.6  1.9
16        Kansas  3.9  4.6  3.1  3.4  3.7  6.3  6.6
17      Kentucky  5.5  5.9  3.6  4.3  4.6  4.8  5.9
18     Louisiana 11.4 12.3 10.3 11.0  9.9 12.5 10.5
19         Maine  1.8  1.7  1.6  1.8  1.4  1.2  2.0
20      Maryland  8.1  9.3  6.1  7.4  9.9  8.1 11.6
21 Massachusetts  2.0  2.5  2.0  3.3  2.7  2.0  2.6
22      Michigan  5.5  5.7  5.4  5.9  6.1  6.7  7.5
23     Minnesota  1.9  2.0  1.6  1.8  2.2  3.1  3.6
24   Mississippi  5.7  6.4  8.6  6.9  7.3  9.0 11.1
25      Missouri  9.9  9.8  6.6  7.0  6.9  6.2  8.1
26       Montana  3.2  3.9  3.6  2.5  1.9  1.8  3.9
27      Nebraska  2.3  2.2  2.9  3.0  2.5  3.7  2.9
28        Nevada  6.7  8.9  6.0  5.8  8.5  6.5 13.7
29 New Hampshire  1.5  1.0  0.9  1.0  1.4  1.8  1.7
30    New Jersey  3.2  3.6  3.9  4.2  4.8  3.4  4.2
31    New Mexico  8.0  6.9  4.8  6.8  7.4  7.4 11.5
32      New York  2.9  2.8  3.1  4.5  4.5  5.0  7.4
33 North Carolina  6.0  6.1  5.1  5.0  6.7  7.0  8.5
34   North Dakota  2.4  1.3  3.0  1.5  1.1  0.6  2.2
35           Ohio  4.8  6.4  4.0  4.2  5.1  3.7  4.8
36       Oklahoma  5.2  6.2  4.5  5.2  5.3  5.3  6.8
37         Oregon  2.0  2.5  2.0  2.5  2.2  2.0  4.0
38   Pennsylvania  6.1  5.8  4.8  5.1  6.1  4.9  5.7
39   Rhode Island  1.5  2.0  2.4  2.8  3.2  4.3  2.5
40 South Carolina  7.7  7.6  6.4  5.7  7.4  5.8  9.0
41   South Dakota  1.4  3.1  2.3  2.8  2.3  0.9  1.2
42      Tennessee  7.4  8.0  5.7  5.6  7.2  7.2  9.5
43          Texas  4.6  5.0  4.4  4.9  6.2  5.9  7.7
44           Utah  1.9  2.4  2.3  1.9  2.3  1.9  3.2
45        Vermont  1.6  2.7  1.6  1.1  1.3  1.5  1.9
46       Virginia  4.6  5.6  7.5  4.7  6.1  5.7  7.5
47     Washington  3.1  3.0  2.5  2.3  3.3  3.3  4.6
48  West Virginia  3.7  5.4  4.0  3.1  4.4  2.5  3.8
49      Wisconsin  3.0  3.3  2.9  2.7  3.5  3.2  4.2
50        Wyoming  2.3  2.4  2.7  1.4  2.7  2.4  3.3
```

Next, I converted the data to a data frame and plotted the data frame, see figure 2 and 3.

Figure 2: Script from Rstudio

```r
library(tidyverse)
library(rvest)
library(janitor)
library(stringr)
library(lubridate)
library(ggplot2)
library(ggridges)

url <- "https://en.wikipedia.org/wiki/List_of_U.S._states_by_homicide_rate"

#Scrape the website using the command read_html
url_html <- read_html(url)

#I extract the html table through the tag using the command html_nodes and html_table.
web_table <- url_html %>%
  html_nodes("table") %>% #Finds the first "table"
  html_table #Parse the html table found from html_node into a dataframe

#print table
web_table

#takes the downloaded html table, unlists it and and then combine the individual elements as columns.
new_table <- do.call(cbind,unlist(web_table, recursive = FALSE))
head(new_table) #print the first 6 lines.

#converts table into data frame
new_data_frame <- as.data.frame.matrix(new_table)

#select column
select(new_data_frame,"2017")

#plots number of homicides as a function of state for the year 2018
plot(new_data_frame[,"2017"], main="Homicides year 2017",
     xlab = "State no.",
     ylab = "Homicides")
```

Figure 3: Plot from Rstudio



Interpretation:

I struggled to manage the data inside Rstudio, so I exported all the data to a csv file using the command write_excel_csv2(new_data_frame, "excel_homicide.csv"). Now I can manage the data inside Excel, which I am more familiar with.

It's hard to interpret the visualization. I need the names of the states or a more fine divided x-axis to get anything from this. In figure 3 I see an outlier at 12-point-something and by comparing that to figure 1 I see that it must be Louisiana. So: In 2017 Louisiana was the state with the highest number of homicides per 100.000 inhabitants. In figure 1 I see the same trend: Louisiana has far more incidents than any other state, but is followed by states as Alabama, Arkansas, Maryland, Missouri,

Nevada, New Mexico, South Carolina, Tennessee, which are all states in the deep south (except from Maryland). I measure the mean rate of incidents as 5. All the mentioned states have a rate of incidence above 5. This trend suggests that there are more homicides in the southern states than in the northern states – in 2017. You have the investigate that for other years to state anything.

Bugs and suggestions:

I struggled to extract state names instead of the state numbers. I believe that could have been fixed spending more time on data cleaning.

The states are sorted alphabetical by name. I could have sorted them either by size or by number of homicides using the mutate function. I could also have sorted them by region (north, south, east west), depending of what story I would like to tell.