Week 05 - Webscraping

Emil Bæk Mogensen

2023-09-25

After following Adelas script step-by-step on police killings, I decided to choose the following option.

Produce data visualisations that shed light on another interesting aspect of the police killing data.

The purpose of this inquiry is to analyze how race and gender are represented in police killings in the period of 2013-2020 from the outlook of percentages to then compare and reveal statistical over or under-representation in the US population. Lastly I will shortly discuss possible explanations behind the results from the perspective of social sciences.

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)
library(ggplot2)
library(ggridges)
```

First I imported the scraped and cleaned data from the cloned repository

```r
policekillings202210 <- read_csv("data/policekillings202210.csv")
```

```
## Rows: 5430 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (6): Name, Gender, Race, State, Method, Source
## dbl  (2): Age, Year
## date (1): Date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Then calculate the percentage of each race that is represented.

```
data <- policekillings202210
data %>%
group_by(Race) %>%
summarise(percent=100*n()/nrow(data))
```

```
## # A tibble: 11 x 2
##    Race  percent
##    <chr>   <dbl>
##  1 A      1.60
##  2 B     23.2
##  3 H     15.8
##  4 I      0.0552
##  5 L      0.442
##  6 M      0.0184
##  7 N      1.38
##  8 O      0.884
##  9 PI     0.0184
## 10 W     43.8
## 11 <NA>  12.8
```

Running the chunk shows all races in the dataset. I decided to focus on the three most prominent, as the rest were substantially smaller. Secondly, i realized by following Adelas step by step guide to scraping, cleaning and plotting, that the latino variable is missing from 2015, but that the hispanic variable appears from that year up untill 2020. Thus it can be highly assumed that hispanic replaced latino and that the two terms are used synanmously.

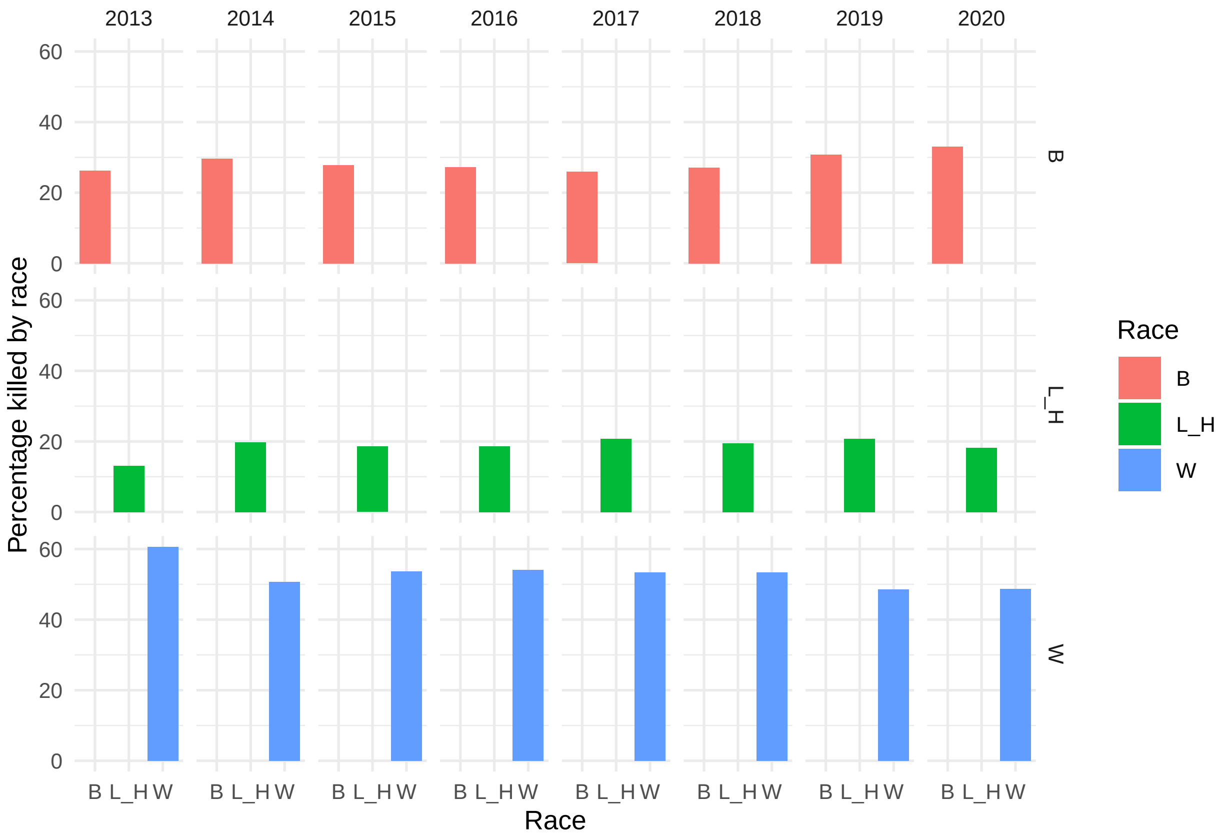
```
data <- policekillings202210
data %>%
filter(Race %in% c("B", "W", "H", "L")) %>%
mutate(Race = ifelse(Race %in% c("L", "H"), "L_H", Race)) %>%
group_by(Race) %>%
summarise(percent = 100 * n() / nrow(data))
```

```
## # A tibble: 3 x 2
##   Race  percent
##   <chr>   <dbl>
## 1 B        23.2
## 2 L_H      16.2
## 3 W        43.8
```

This shows that 43,8 of police killings in the period were of whites, 23,8 of blacks and 16,20% of latinos and hispanics.

Now I want to create a graph that visualises the percentages. For this I gathered inspiration from Sobotkova's script and implemented my own variables.

```
data <- policekillings202210
data %>%
  filter(Race %in% c("B", "W", "L", "H")) %>%  # Filter for "B," "W," "L," and "H" races
  mutate(Race = ifelse(Race %in% c("L", "H"), "L_H", Race)) %>%
  group_by(Year, Race) %>%
  tally() %>%
  mutate(perc = n / sum(n) * 100) %>%
  ggplot(aes(Race, perc, fill = Race)) +
  geom_col() +
  facet_grid(Race ~ Year) +
  theme_minimal(base_size = 10) +
  xlab("Race") +
  ylab("Percentage killed by race")
```



Finally I want to calculate whether there is talk of over- or under-representation in regards to the total US population in the year of 2020, but the process is applicable to the total period of 2013-2020. I got the ‘race data’ and the ‘total us population data’ from the US Census Bureau.

According to the source, of the total population of 331,449 million in 2020 75,5% were white, 13,6% black and 19,1% hispanic or latino.

First I need to isolate and calculate the amount of people killed by police in 2020.

```
data <- policekillings202210
total_killed_2020 <- data %>%
filter(Year == 2020) %>%
nrow()
print(total_killed_2020)
```

```
## [1] 326
```

Then narrow down my inquiry to focus only on the three most prominent races in police killings. Running the code without.

```
total_killed_2020 <- data %>%
  filter(Year == 2020) %>%
  nrow()
race_killed_2020 <- data %>%
  filter(Year == 2020) %>%
   filter(Race %in% c("B", "W", "L", "H")) %>%  # Filter for "B," "W," "L," and "H" races
  mutate(Race = ifelse(Race %in% c("L", "H"), "L_H", Race)) %>%
  group_by(Race) %>%
  summarise(Count = n())
race_killed_2020 <- race_killed_2020 %>%
  mutate(Percentage = (Count / total_killed_2020) * 100)
print(race_killed_2020)
```

```
## # A tibble: 3 x 3
##   Race  Count Percentage
##   <chr> <int>      <dbl>
## 1 B        40      12.3
## 2 L_H      22       6.75
## 3 W        59      18.1
```