Homerwork 2

Eugene Nesterenko

2023-05-21

# Data Visualisation - Exploration

Now that you’ve demonstrated your software is setup, and you have the basics of data manipulation, the goal of this assignment is to practice transforming, visualising, and exploring data.

# Mass shootings in the US

In July 2012, in the aftermath of a mass shooting in a movie theater in Aurora, Colorado, [Mother Jones](https://www.motherjones.com/politics/2012/07/mass-shootings-map/) published a report on mass shootings in the United States since 1982. Importantly, they provided the underlying data set as [an open-source database](https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/) for anyone interested in studying and understanding this criminal behavior.

## Obtain the data

Rows: 125  
Columns: 14  
$ case <chr> "Oxford High School shooting", "San Jose VTA shoo…  
$ year <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2020, 2020, 2…  
$ month <chr> "Nov", "May", "Apr", "Mar", "Mar", "Mar", "Mar", …  
$ day <dbl> 30, 26, 15, 31, 22, 16, 16, 26, 10, 6, 31, 4, 3, …  
$ location <chr> "Oxford, Michigan", "San Jose, California", "Indi…  
$ summary <chr> "Ethan Crumbley, a 15-year-old student at Oxford …  
$ fatalities <dbl> 4, 9, 8, 4, 10, 8, 4, 5, 4, 3, 7, 9, 22, 3, 12, 5…  
$ injured <dbl> 7, 0, 7, 1, 0, 1, 0, 0, 3, 8, 25, 27, 26, 12, 4, …  
$ total\_victims <dbl> 11, 9, 15, 5, 10, 9, 4, 5, 7, 11, 32, 36, 48, 15,…  
$ location\_type <chr> "School", "Workplace", "Workplace", "Workplace", …  
$ male <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, T…  
$ age\_of\_shooter <dbl> 15, 57, 19, NA, 21, 21, 31, 51, NA, NA, 36, 24, 2…  
$ race <chr> NA, NA, "White", NA, NA, "White", NA, "Black", "B…  
$ prior\_mental\_illness <chr> NA, "Yes", "Yes", NA, "Yes", NA, NA, NA, NA, NA, …

| column(variable) | description |
| --- | --- |
| case | short name of incident |
| year, month, day | year, month, day in which the shooting occurred |
| location | city and state where the shooting occcurred |
| summary | brief description of the incident |
| fatalities | Number of fatalities in the incident, excluding the shooter |
| injured | Number of injured, non-fatal victims in the incident, excluding the shooter |
| total\_victims | number of total victims in the incident, excluding the shooter |
| location\_type | generic location in which the shooting took place |
| male | logical value, indicating whether the shooter was male |
| age\_of\_shooter | age of the shooter when the incident occured |
| race | race of the shooter |
| prior\_mental\_illness | did the shooter show evidence of mental illness prior to the incident? |

## Explore the data

### Specific questions

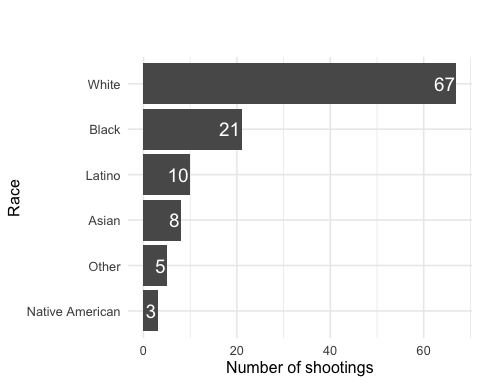
* Generate a data frame that summarizes the number of mass shootings per year.

# Count the number of mass shootings per year  
mass\_shootings %>%   
 group\_by(year) %>%   
 summarise(number = n())

# A tibble: 37 × 2  
 year number  
 <dbl> <int>  
 1 1982 1  
 2 1984 2  
 3 1986 1  
 4 1987 1  
 5 1988 1  
 6 1989 2  
 7 1990 1  
 8 1991 3  
 9 1992 2  
10 1993 4  
# ℹ 27 more rows

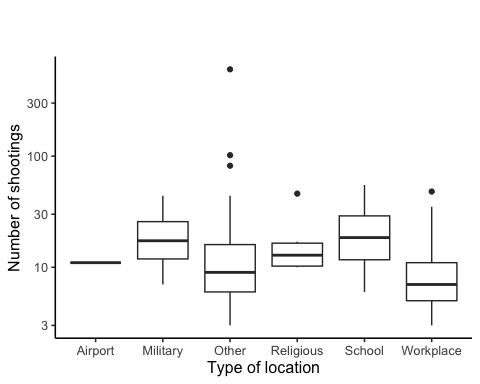
* Generate a bar chart that identifies the number of mass shooters associated with each race category. The bars should be sorted from highest to lowest and each bar should show its number.

# Counting number of shootings per race,excluding cases with unidentified race (NA)  
mass\_shootings %>%  
 filter(!is.na(race)) %>%   
 group\_by(race) %>%   
 summarise(number = n()) %>%   
   
 # Rearranging in descending order for total number  
 mutate(race = fct\_reorder(race, number)) %>%  
   
 # Drawing a horizontal bar chart with labels  
 ggplot(aes(x = number, y = race)) +  
 geom\_col() +  
 geom\_text(  
 aes(label = number, x = number - .25),  
 colour = "white",  
 size = 5,  
 hjust = 1) +  
   
 # Adding theme and axis titles  
 theme\_minimal(base\_size = 12)+  
 labs(  
 title = "",  
 subtitle = "",  
 x = "Number of shootings",  
 y = "Race"  
 ) +  
 NULL



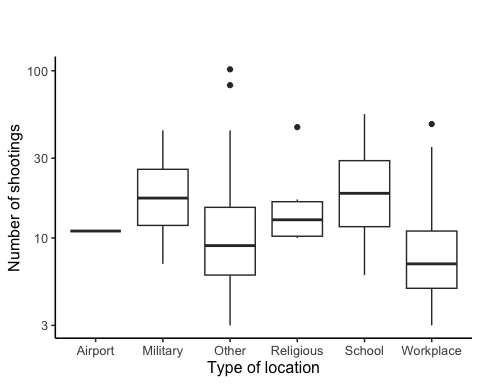
* Generate a boxplot visualizing the number of total victims, by type of location.

# Grouping shootings by location type, excluding those with unidentified location (NA)  
mass\_shootings %>%  
 filter(!is.na(location\_type)) %>%   
 group\_by(location\_type) %>%   
   
 # Drawing a boxplot  
 ggplot(aes(y = total\_victims, x = location\_type)) +  
 geom\_boxplot() +  
   
 # Drawing y-axis (number of shootings) in a log scale to interpret graph more easily  
 scale\_y\_log10() +  
   
 # Adding theme and and axis titles  
 theme\_classic(base\_size = 12) +  
 labs(  
 title = "",  
 subtitle = "",  
 x = "Type of location",  
 y = "Number of shootings"  
 ) +  
 NULL



* Redraw the same plot, but remove the Las Vegas Strip massacre from the dataset.

# Grouping shootings by location type, excluding those with unidentified location (NA)  
mass\_shootings %>%  
 filter(!is.na(location\_type)) %>%  
   
 # Excluding the Las Vegas Strip massacre  
 filter(case != "Las Vegas Strip massacre") %>%   
 group\_by(location\_type) %>%   
   
 # Drawing a boxplot  
 ggplot(aes(y = total\_victims, x = location\_type)) +  
 geom\_boxplot() +  
   
 # Drawing y-axis (number of shootings) in a log scale to interpret graph more easily  
 scale\_y\_log10() +  
   
 # Adding theme and and axis titles  
 theme\_classic(base\_size = 12) +  
 labs(  
 title = "",  
 subtitle = "",  
 x = "Type of location",  
 y = "Number of shootings"  
 ) +  
 NULL



### More open-ended questions

Address the following questions. Generate appropriate figures/tables to support your conclusions.

* How many white males with prior signs of mental illness initiated a mass shooting after 2000?

# Selecting only cases after 2000, where the shooter was white male with prior signs of mental illness  
mass\_shootings %>%  
 filter(race == "White") %>%  
 filter(male == TRUE) %>%   
 filter(prior\_mental\_illness == "Yes") %>%   
 filter(year > 2000) %>%   
   
 # Counting total number of cases after applyied filters  
 summarise(count = n())

# A tibble: 1 × 1  
 count  
 <int>  
1 22

* Which month of the year has the most mass shootings? Generate a bar chart sorted in chronological (natural) order (Jan-Feb-Mar- etc) to provide evidence of your answer.

# We have not found a function in lubridate library, so we will create month index manually  
mass\_shootings\_month <- mass\_shootings %>%   
 mutate(month\_index = case\_when(  
 month == "Jan" ~1,  
 month == "Feb" ~2,  
 month == "Mar" ~3,  
 month == "Apr" ~4,  
 month == "May" ~5,  
 month == "Jun" ~6,  
 month == "Jul" ~7,  
 month == "Aug" ~8,  
 month == "Sep" ~9,  
 month == "Oct" ~10,  
 month == "Nov" ~11,  
 month == "Dec" ~12,  
 TRUE ~0  
 )  
 )  
  
# Count cases of mass shooting by month and save as new data frame  
mass\_shootings\_month\_chronological\_summary <- mass\_shootings\_month %>%   
 group\_by(month, month\_index) %>%  
 summarise(count = n()) %>%   
   
 # We arrange in chronological order using month index (Jan = 1, Dec = 12)  
 arrange(month\_index)

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

# Attempt to arrange bar chart in chronologocal order usung month index (unsuccessful)  
mass\_shootings\_month\_chronological\_summary %>%   
 slice\_max(order\_by = month\_index, n=12) %>%   
  
   
# Drawing a vertical bar chart with labels  
 ggplot(aes(x = month, y = count)) +  
 geom\_col() +  
   
 # Adding theme and axis titles  
 theme\_minimal(base\_size = 12) +  
 labs(  
 title = "",  
 subtitle = "",  
 x = "Number of shootings",  
 y = "Month"  
 ) +  
 NULL

