Homework 04 - Import + clean data

Jemimah Osei (jko35)

2025-03-06

Setup

Load packages and data:

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(googlesheets4)
```

Exercises

Exercise 1

```
mass_shootings <- read_sheet("https://docs.google.com/spreadsheets/d/1b9o6uDO18
sLxBqPwl_Gh9bnhW-ev_dABH83M5Vb5L8o/edit?gid=0#gid=0")
```

```
! Using an auto-discovered, cached token.
```

```
  To suppress this message, modify your code or options to clearly consent to
  the use of a cached token.
```

```
  See gargle's "Non-interactive auth" vignette for more details:
```

```
  <https://gargle.r-lib.org/articles/non-interactive-auth.html>
```

```
ℹ The googlesheets4 package is using a cached token for 'jko35@cornell.edu'.
```

```
✔ Reading from "Mother Jones - Mass Shootings Database, 1982 - 2024".
```

```
✔ Range 'Sheet1'.
```

```
New names:
• `location` -> `location...2`
• `location` -> `location...8`
```

```
mass_shootings
```

```
# A tibble: 151 × 24
   case              location...2 date                summary fatalities injured
   <chr>             <chr>        <dttm>              <chr>        <dbl>   <dbl>
 1 Apalachee High S… Winder, Geo… 2024-09-04 00:00:00 "Colt …          4       9
 2 Arkansas grocery… Fordyce, Ar… 2024-06-21 00:00:00 "Travi…          4      10
 3 UNLV shooting     Las Vegas, … 2023-12-06 00:00:00 "Antho…          3       1
 4 Maine bowling al… Lewiston, M… 2023-10-25 00:00:00 "Rober…         18      13
 5 Jacksonville Dol… Jacksonvill… 2023-08-26 00:00:00 "Ryan …          3       0
 6 Orange County bi… Trabuco Can… 2023-08-23 00:00:00 "John …          3       6
 7 Philidelphia nei… Philadelphi… 2023-07-03 00:00:00 "Kimbr…          5       2
 8 New Mexico neigh… Farmington,… 2023-05-15 00:00:00 "Beau …          3       6
 9 Texas outlet mal… Allen, Texas 2023-05-06 00:00:00 "Mauri…          8       7
10 Louisville bank … Louisville,… 2023-04-10 00:00:00 "Conno…          5       8
# ℹ 141 more rows
# ℹ 18 more variables: total_victims <dbl>, location...8 <chr>,
#   age_of_shooter <list>, prior_signs_mental_health_issues <chr>,
#   mental_health_details <chr>, weapons_obtained_legally <chr>,
#   where_obtained <chr>, weapon_type <chr>, weapon_details <chr>, race <chr>,
#   gender <chr>, sources <chr>, mental_health_sources <chr>,
#   sources_additional_age <chr>, latitude <list>, longitude <list>, …
```

Exercise 2

```
columns_to_remove        <-        c("sources",        "mental_health_sources",
"sources_additional_age",
                "weapon_details", "where_obtained", "weapons_obtained_legally")
mass_shootings_new <- mass_shootings |> select(-any_of(columns_to_remove))

mass_shootings_new <- mass_shootings_new |>
  rename(
    shooting_case_name = case,
    shooting_location = location...2,
    location_type = location...8,
    shooting_date = date,
    age_of_shooter = age_of_shooter,
    incident_type = type
  )


mass_shootings_new <- mass_shootings_new |>
  mutate(
     shooting_date  =  parse_date_time(shooting_date,  orders  =  c("mdy",  "ymd",
"dmy")),
    shooting_date = as.Date(shooting_date),


      across(c(fatalities,  injured,  total_victims,  age_of_shooter,  latitude,
longitude),
           ~ as.numeric(unlist(.x))),

    year = year(shooting_date),
    month = month(shooting_date, label = TRUE),

    race = as.factor(str_to_title(race)),
    gender = as.character(gender),
    incident_type = as.factor(incident_type),

    race = replace(race, race %in% c("Unknown", "TBD", "-"), NA),
    gender = replace(gender, gender %in% c("Unknown", "TBD", "-"), NA),

    gender = case_when(
      str_detect(gender, "M") ~ "Male",
      str_detect(gender, "F") ~ "Female",
      str_detect(gender, "Transgender") ~ "Transgender",
      TRUE ~ NA_character_
    )
  )
```

```
Warning: There were 3 warnings in `mutate()`.
The first warning was:
```

```
ℹ In argument: `across(...)`.
Caused by warning:
! NAs introduced by coercion
ℹ Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
```

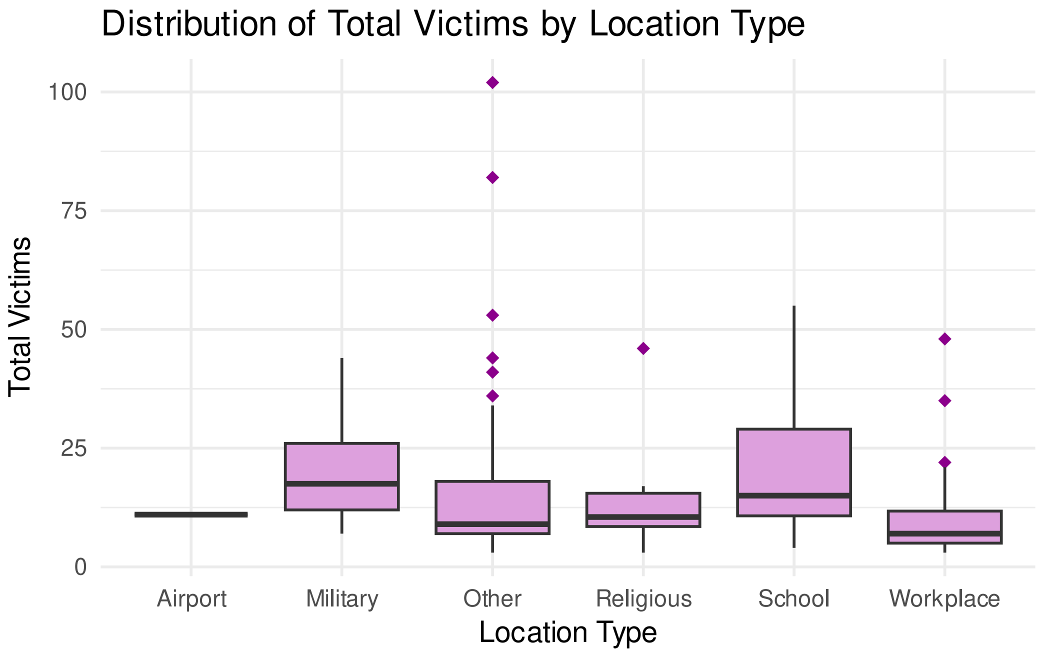
```r
glimpse(mass_shootings_new)
```

```
Rows: 151
Columns: 19
$ shooting_case_name                <chr> "Apalachee High School shooting", "Ar…
$ shooting_location                 <chr> "Winder, Georgia", "Fordyce, Arkansas…
$ shooting_date                     <date> 2024-09-04, 2024-06-21, 2023-12-06, …
$ summary                           <chr> "Colt Gray, 14, was apprehended by re…
$ fatalities                        <dbl> 4, 4, 3, 18, 3, 3, 5, 3, 8, 5, 6, 3, …
$ injured                           <dbl> 9, 10, 1, 13, 0, 6, 2, 6, 7, 8, 6, 5,…
$ total_victims                     <dbl> 13, 14, 4, 31, 3, 9, 7, 9, 15, 13, 12…
$ location_type                     <chr> "School", "workplace", "School", "Oth…
$ age_of_shooter                    <dbl> 14, 44, 67, 40, 21, 59, 40, 18, 33, 2…
$ prior_signs_mental_health_issues <chr> "yes", "-", "-", "yes", "yes", "-", "…
$ mental_health_details             <chr> "-", "-", "-", "Card reportedly spoke…
$ weapon_type                       <chr> "semiautomatic rifle", "shotgun; semi…
$ race                              <fct> White, White, White, White, White, Wh…
$ gender                            <chr> "Male", "Male", "Male", "Male", "Male…
$ latitude                          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ longitude                         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ incident_type                     <fct> mass, mass, mass, Spree, mass, mass, …
$ year                              <dbl> 2024, 2024, 2023, 2023, 2023, 2023, 2…
$ month                             <ord> Sep, Jun, Dec, Oct, Aug, Aug, Jul, Ma…
```

```r
#mass_shootings_precleaned <- read_rds("data/mass-shootings.rds")
#waldo::compare(
  #x = mass_shootings_precleaned,
  #y = mass_shootings_new,
  #tolerance = 1e-4
#)

#write_rds(mass_shootings_new, "data/mass_shootings_cleaned.rds")
```

Exercise 3

```r
mass_shootings_filtered <- mass_shootings_new |>
  filter(shooting_case_name != "Las Vegas Strip massacre") |>
  mutate(location_type = str_to_title(location_type))

ggplot(data = mass_shootings_filtered, mapping = aes(x = location_type, y =
total_victims)) +
  geom_boxplot(fill = "plum", outlier.color = "darkmagenta", outlier.shape = 18,
outlier.size = 2) +
  theme_minimal() +
  labs(
    title = "Distribution of Total Victims by Location Type",
    x = "Location Type",
    y = "Total Victims",

  )
```



Interpretation:

From my box plot visualization of the total victims by location type, it is observable that the military and school have the highest medians hence the highest number of victims while the workplace and religious locations are on the lower end of the medians with more outliers recorded in their case

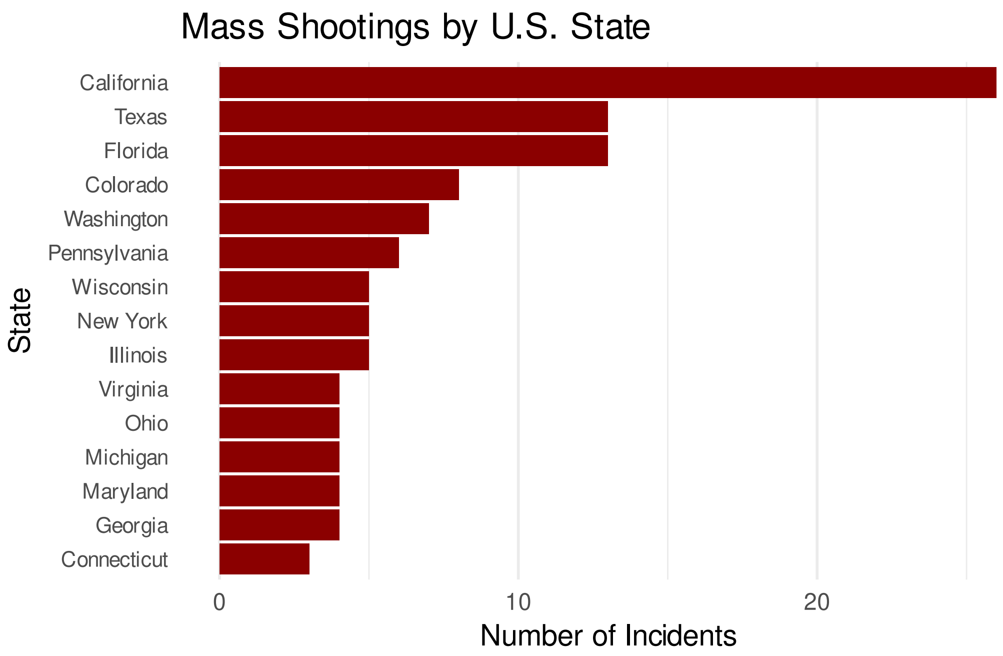
Exercise 4

```
num_incidents <- mass_shootings_new |>
  filter(
    str_detect(str_to_lower(race), "white"),
    str_detect(str_to_lower(prior_signs_mental_health_issues), "yes"),
    year > 2000,
    str_detect(str_to_lower(incident_type), "mass")
  ) |>
  nrow()

num_incidents
```

```
[1] 26
```

Exercise 5

```
mass_shootings_extract <- mass_shootings_new |>
  mutate(state = word(shooting_location, -1, sep = ", "))

state_counts <- mass_shootings_extract |>
  count(state, sort = TRUE)|>
  slice_head(n = -25)


ggplot(data = state_counts, mapping = aes(x = fct_reorder(state, n), y = n)) +
  geom_col(fill = "red4") +
  coord_flip() +
  labs(
    title = "Mass Shootings by U.S. State",
    x = "State",
    y = "Number of Incidents",
  ) +
  theme_minimal( )+
  theme(
    axis.text.y = element_text(size = 8),
    panel.grid.major.y = element_blank(),
  )
```



Interpretation:

From my graph, I observed that the states of California, Texas, and Florida had the highest mass shootings of the group with California leading at a very high point of over 25 number of incidents. One common thread I see among the highest shootings is that the top states almost all have really big metropolitan areas/cities.

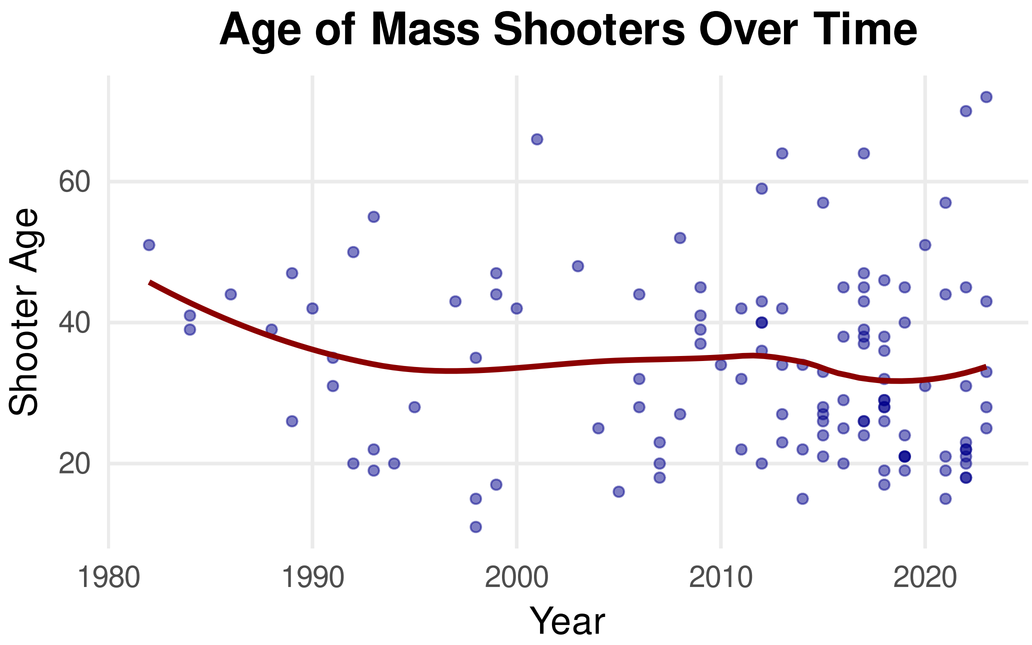
Exercise 6

```r
mass_shootings_filtered <- mass_shootings_new |>
  filter(incident_type == "Mass")
ggplot(mass_shootings_filtered, aes(x = year, y = age_of_shooter)) +
  geom_point(alpha = 0.5, color = "blue4") +
  geom_smooth(se = FALSE, color = "red4", linetype = "solid") +
  labs(
    title = "Age of Mass Shooters Over Time",
    x = "Year",
    y = "Shooter Age",
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    panel.grid.minor = element_blank()
  )
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale
range
(`geom_point()`).
```



Interpretation:

There's a general downward trend that I observe which indicates that the average shooter age declined at around the 1980's to 2000's. From 2000 to 2010 the shooter age remained relatively stable. Then later, an observable mix between majority younger perpetrators with some older ones included.