Strzelaniny w szkołach w USA w latach 1999-2018 - analiza eksploracyjna

Tomasz Owczarek

maj 2018

Contents

Wprowadzenie

Celem projektu jest eksploracja danych o strzelaninach w szołach w USA w latach 1999-2018. Dane na ten temat zostały zebrane przez dziennikarzy *The Washington Post* i udostępnione w tym artykule. Dane skladają się z 217 rekordów i 50 zmiennych. Każdy rekord dotyczy pojedynczej strzelaniny.

Obróbka danych

Dane wczytano bezpośrednio z repozytorium githuba The Washington Post. Po wczytaniu okazało się, że część zmiennych wymaga dodatkowej obróbki, m.in. w niektórych zmiennych liczbowych należało usunąć przecinki. Dodatkowo z kolumny z datami wydobyto dzień i miesiąc.

Ostatecznie okazało się to niepotrzebne, ponieważ w analizie nie użyto tych zmiennych.

Wykorzystane zmienne

W dalszej analizie wykorzystano następujące zmienne:

Zmienna	Opis
year	rok strzelaniny
day_of_week	dzień tygodnia
state	stan
killed	liczba zabitych
injured	liczba rannych
casualties	liczba ofiar (zabici + ranni)
shooting_type	rodzaj ataku (celowy, przypadkowy)
age_shooter1	wiek atakującego
gender_shooter1	płeć atakującego
resource_officer	obecność ochroniarza (0 - brak, 1 - obecny)

Statystyki podsumowujące wybrane zmienne przedstawiają się następująco:

state	killed	injured	casualties	age_shooter1
California : 27	Min. : 0.0000	Min. : 0.000	Min. : 0.000	Min. : 6.00
Florida : 17	1st Qu.: 0.0000	1st Qu.: 0.000	1st Qu.: 1.000	1st Qu.:15.00
Texas : 13	Median : 0.0000	Median : 1.000	Median : 1.000	Median :16.00
North Carolina: 11	Mean : 0.6037	Mean : 1.258	Mean : 1.862	Mean :19.21
Illinois : 10	3rd Qu.: 1.0000	3rd Qu.: 1.000	3rd Qu.: 2.000	3rd Qu.:18.00
Louisiana : 10	Max. :26.0000	Max. :21.000	Max. :34.000	Max. :56.00
(Other) :129	NA	NA	NA	NA's :40

Zmienne, które nie zostały tutaj uwzględnione, będą podsumowane w dalszej części analizy.

Analiza

Przeprowadzona analiza obejmowała następujące zagadnienia:

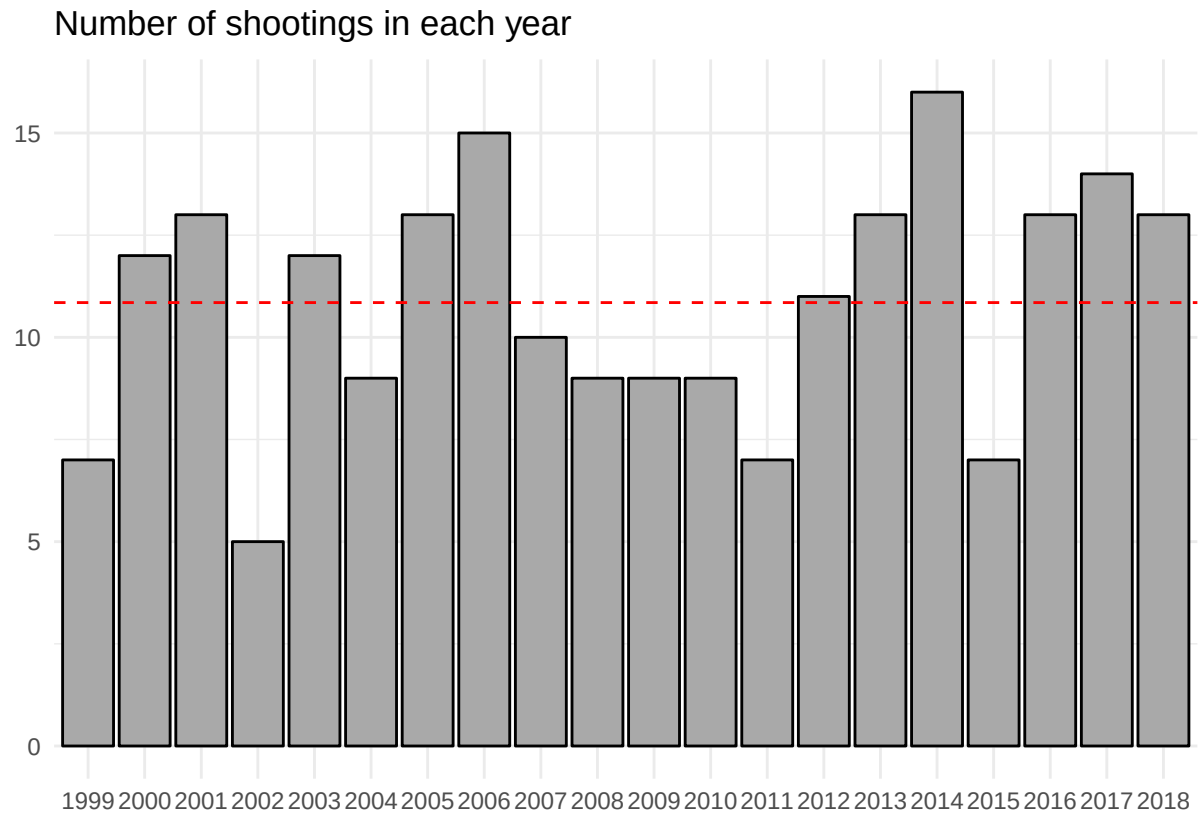
- analizę czasową - m.in. liczbę strzelanin w poszczgólnych latach, liczbę zabitych i rannych,

- analizę przestrzenną - z podziałem na poszczególne stany w USA,
- analizę sprawców - wiek, płeć, rodzaj ataku,
- obecność ochroniarza.

Analiza czasowa

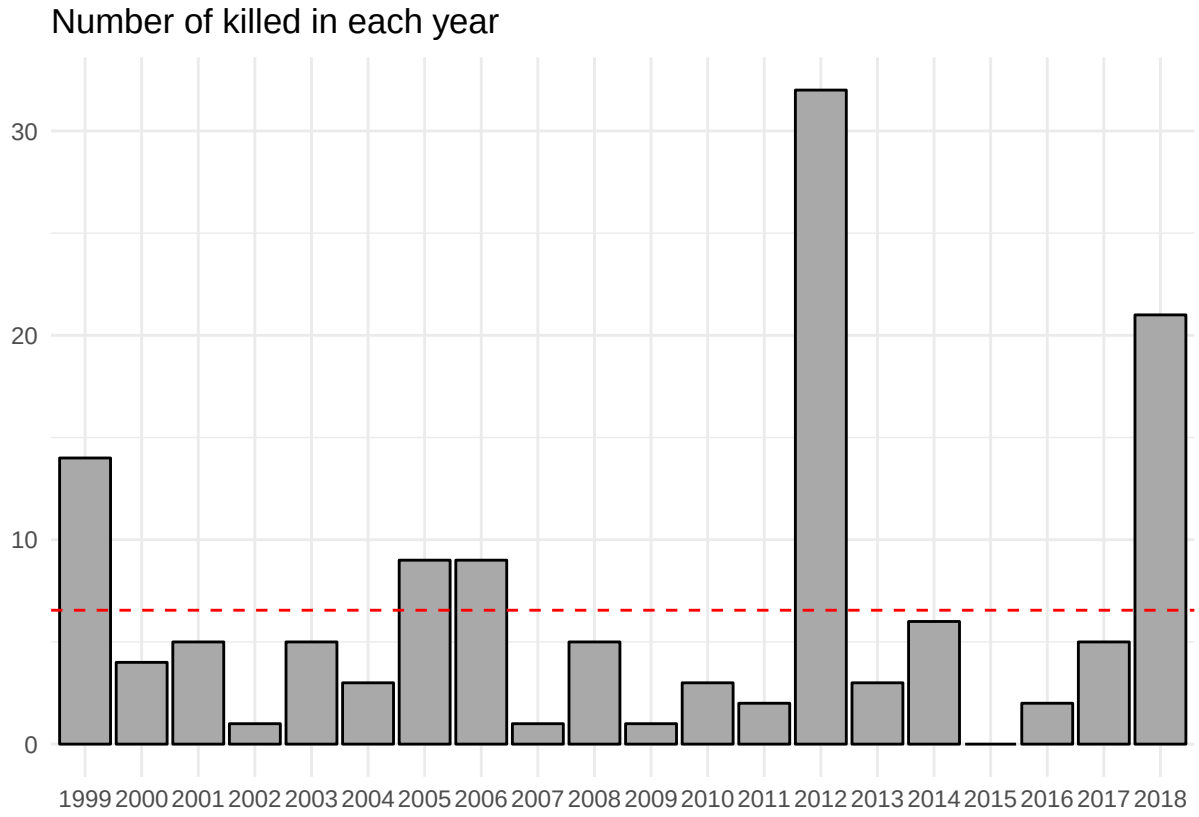
Celem tej analizy jest sprawdzenie, czy można zaobserwować jakieś wzorce związane z czasem.

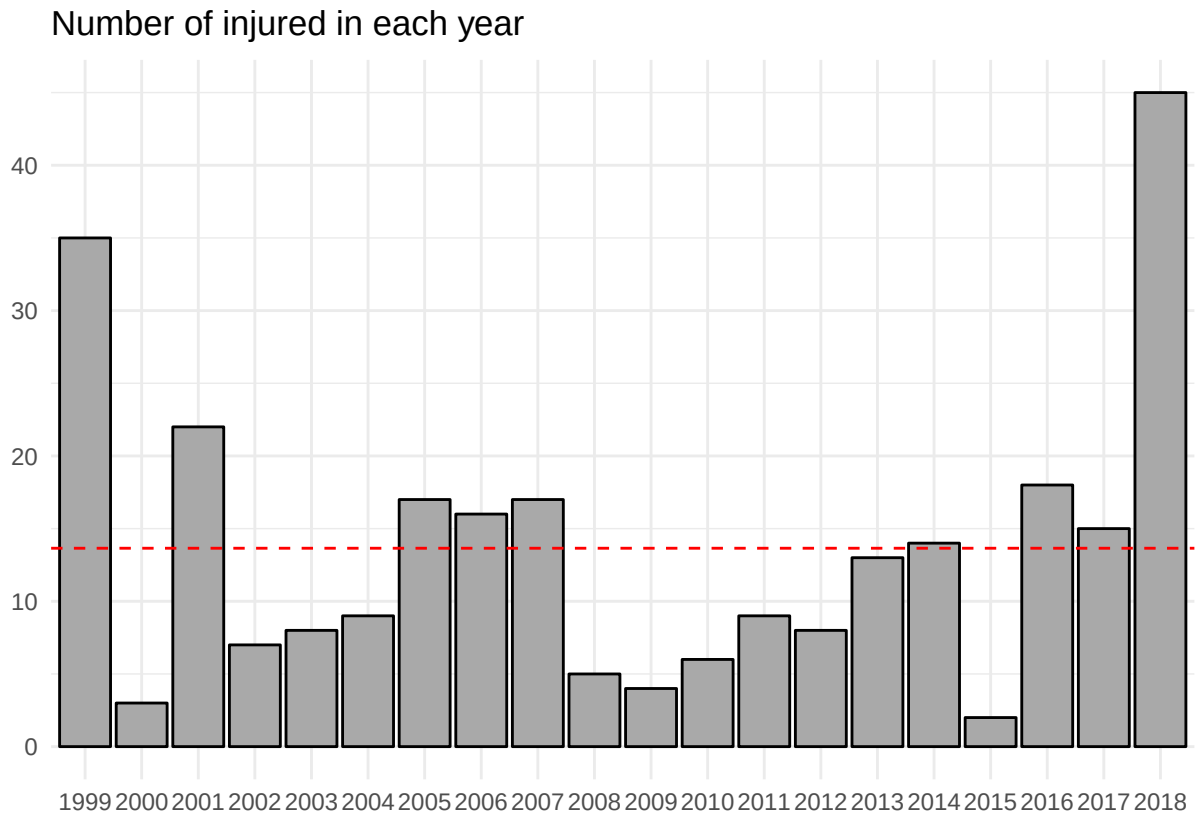
Liczba zdarzeń



Nie widać żadnej tendencji rozwojowej (trendu). Niepokojący jednak jest wynik z roku 2018 - pomimo, że upłynęła dopiero 1/3 roku, to liczba zdarzeń już przekroczyła średnią z ostatnich 20 lat (zanaczoną czerwoną linią).

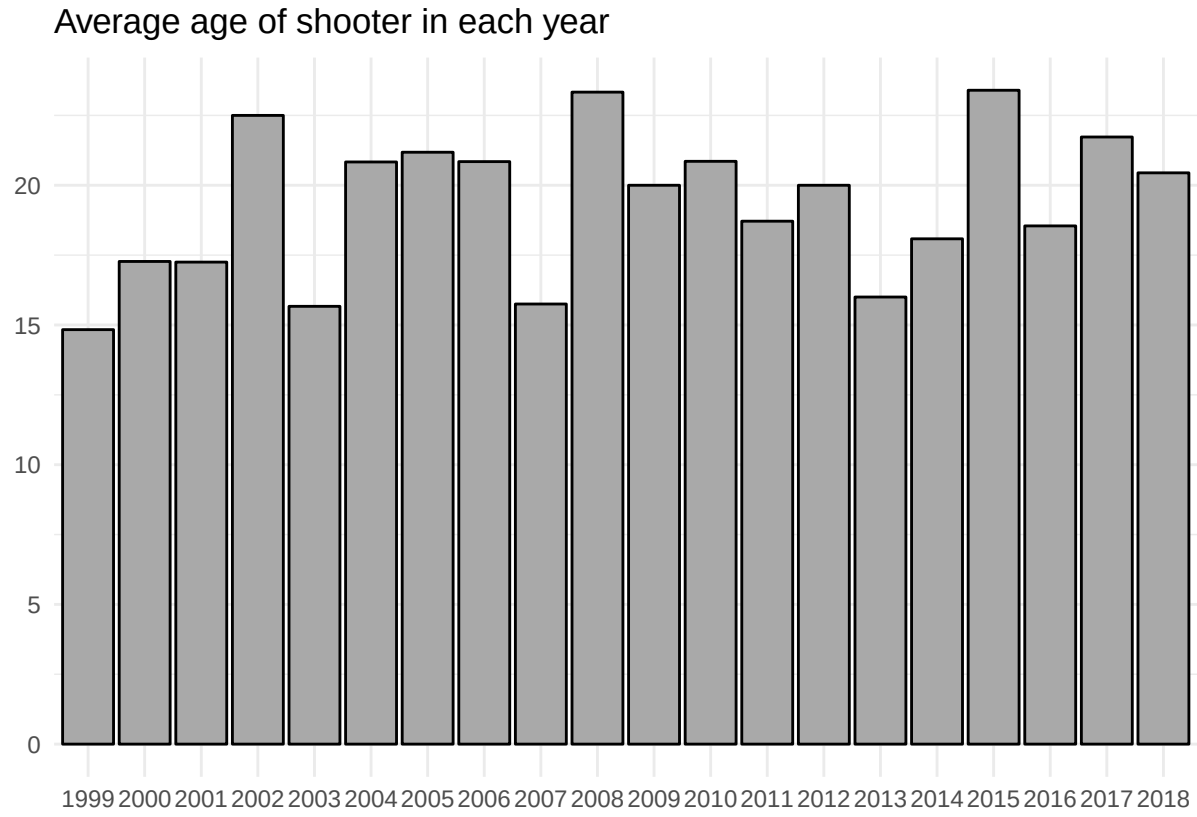
Liczba zabitych i rannych

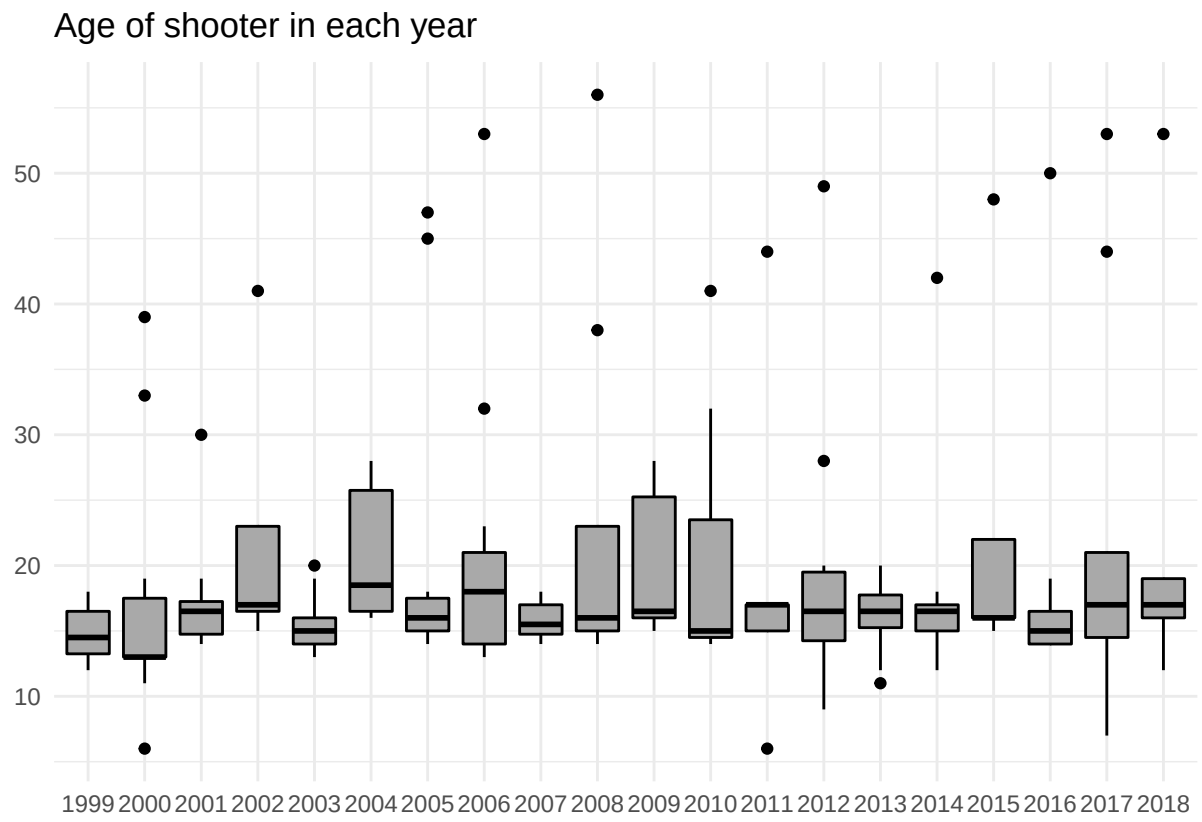




Podobnie jak poprzednio, nie widać tendencji rozwojowych, ale niepokoi liczba ofiar w bieżącym roku, znacznie przewyższająca średnią. Liczba rannych podczas strzelanin w 2018 już jest najwyższa od dwudziestu lat.

Średnia wieku sprawców

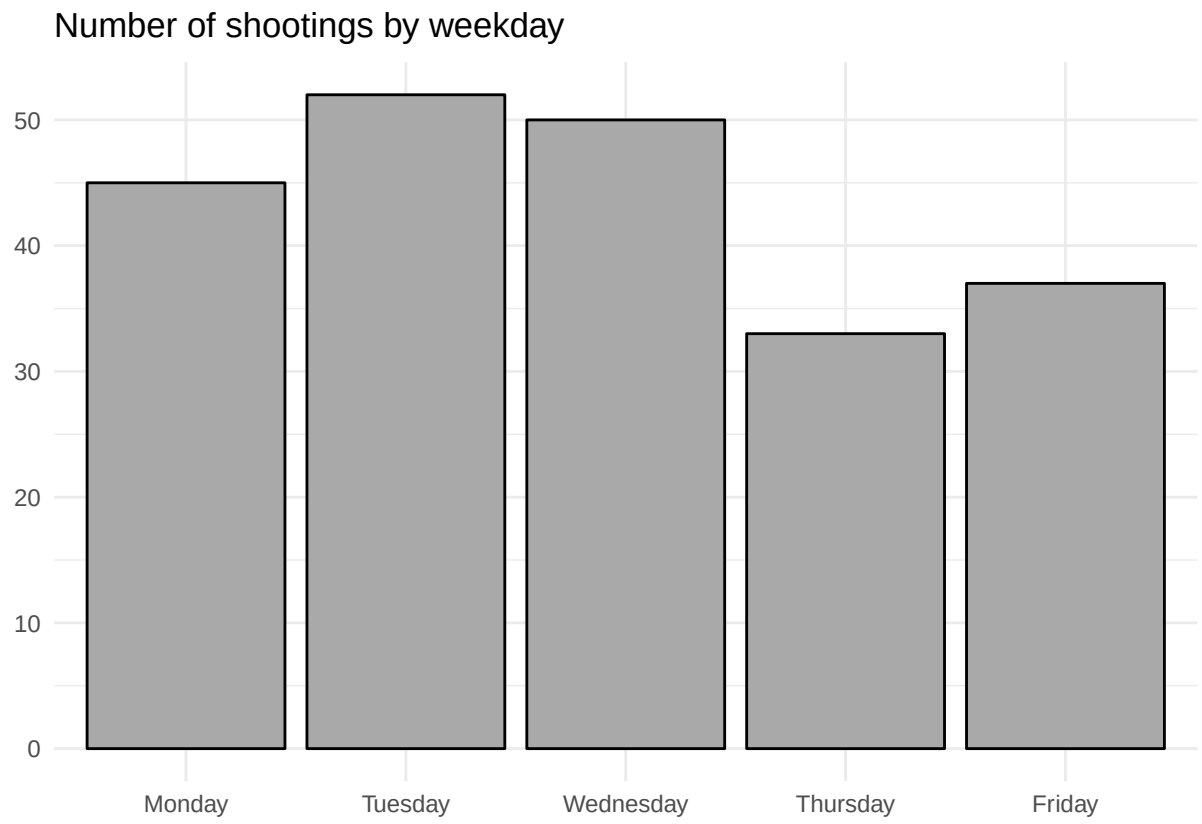




Tutaj także nie widać żadnego trendu, choć rozkłady wieku w poszczególnych latach wydają się w większości prawoskośne (co jest dosyć oczywiste, biorąc pod uwagę analizowane zagadnienie). Ostatni wykres pokazuje też ciekawy fakt - mianowicie to, że sprawcami nie zawsze są uczniowie, ale również ludzie w dojrzałym wieku,

Dzień tygodnia

W tym miejscu sprawdzono liczbę strzelanin z podziałem na poszczególne dni tygodnia.

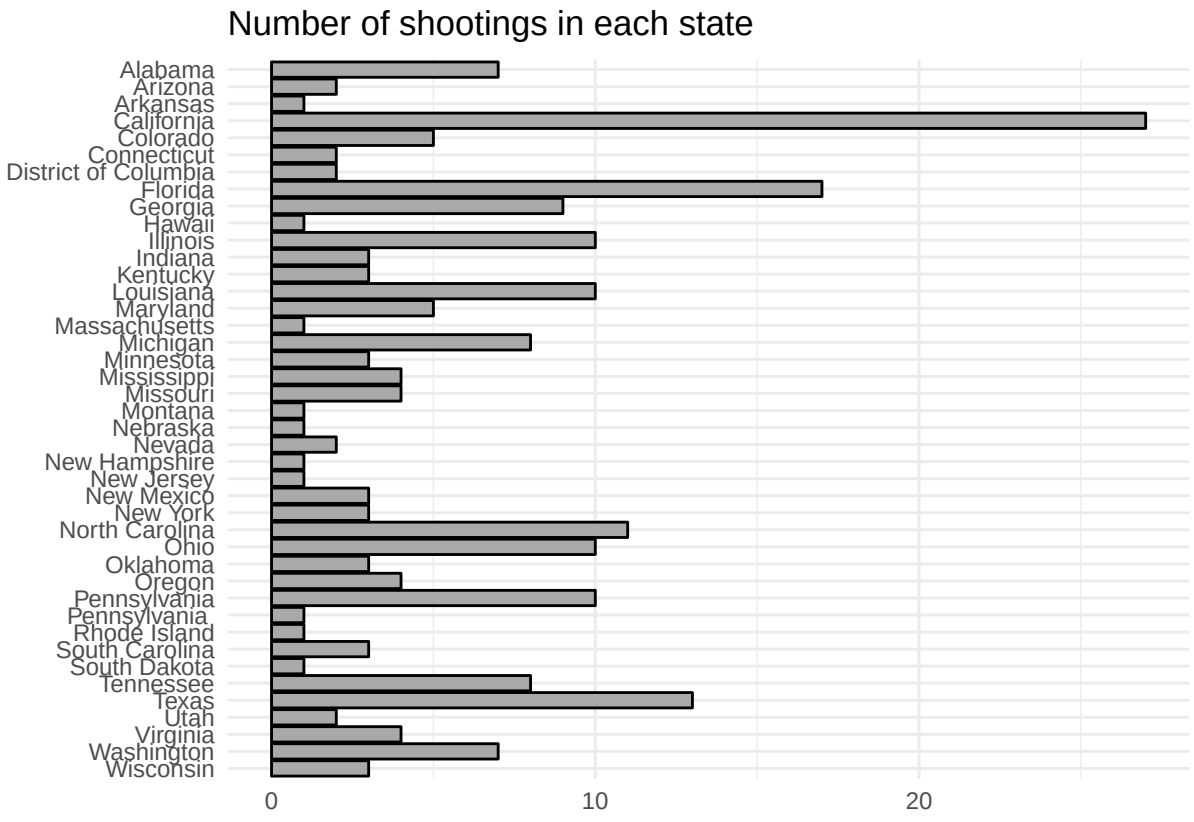


Wykres pokazuje, że bardziej "niebezpieczna" jest pierwsza połowa tygodnia, choć różnice nie są znaczące.

Analiza przestrzenna

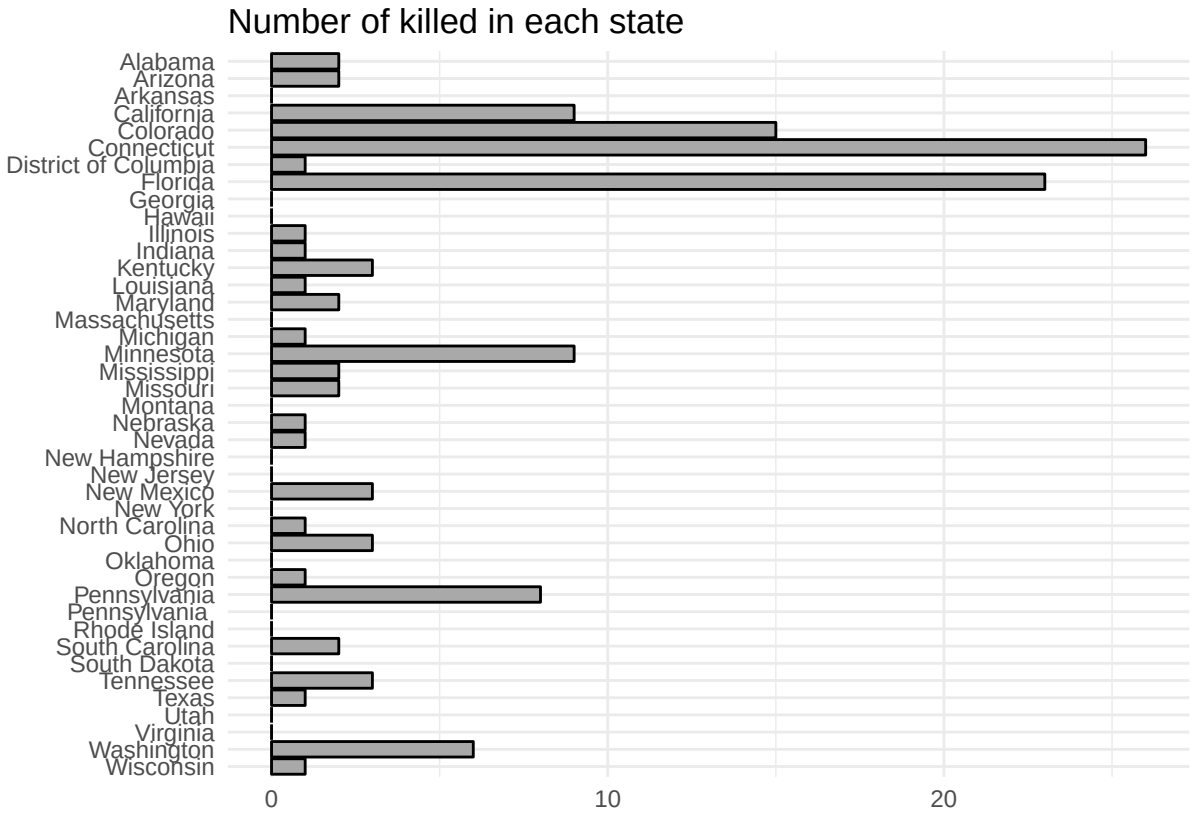
W tej części opracowania została przeprowadzona analiza z podziałem na stany.

Liczba strzelanin

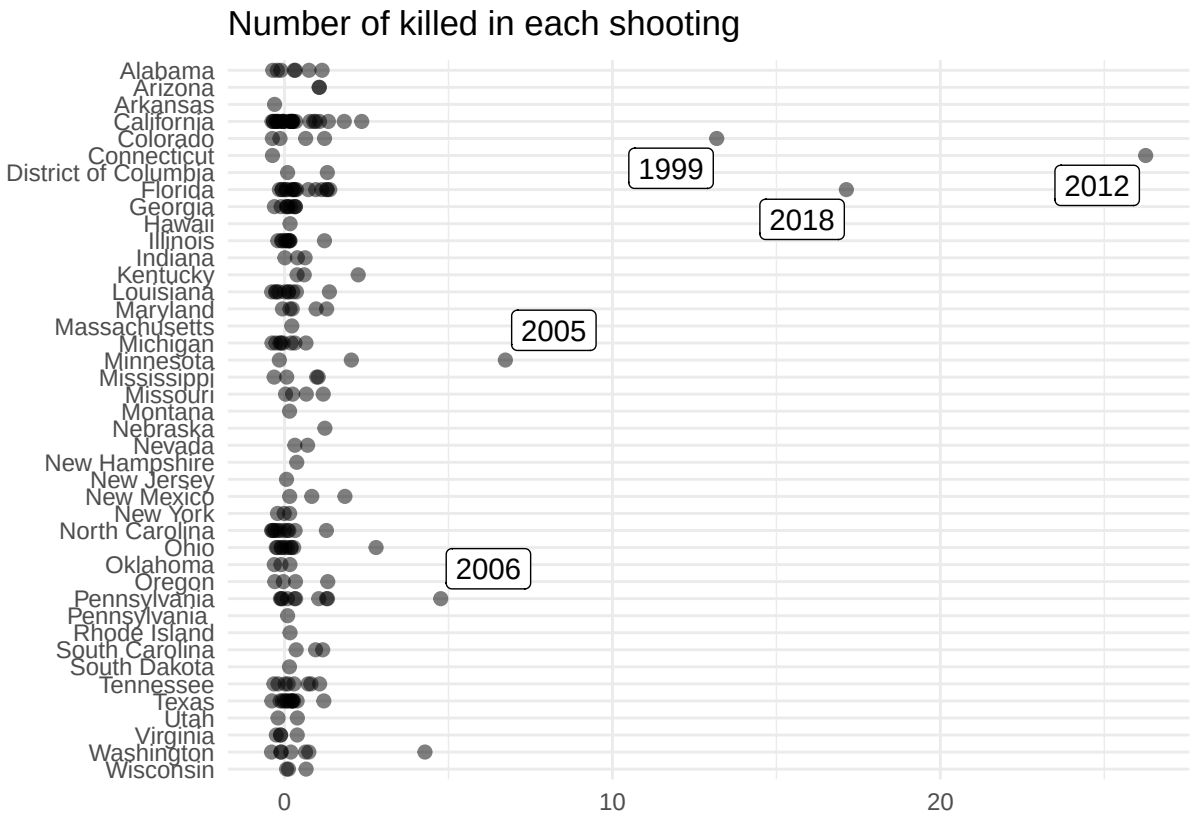


Z wykresu wynika, że w liczbie strzelanin zdecydowanie przodują California, Floryda i Texas, ale też to właśnie te stany mają największą populację (źródło), więc być może nie ma się czemu dziwić. Chociaż już czwarty pod względem populacji stan Nowy York odnotowuje niską liczbę incydentów. Z wykresu można też w jednym rekordzie stan Pennsylvania zostal wprowadzony ze spacją na końcu.

Liczba zabitych



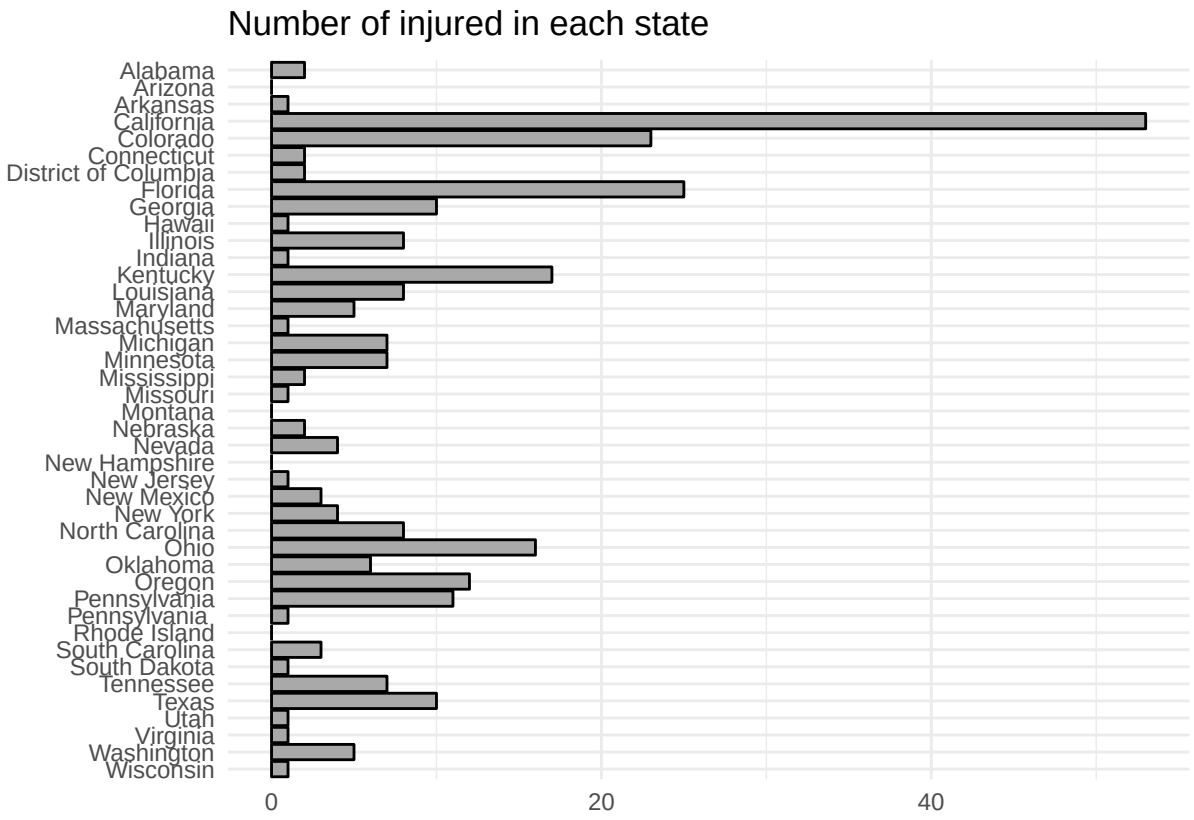
Tutaj z kolei przodują stany Connecticut i Floryda. Najprawdopodobniej to właśnie tam miały miejsce najtragiczniejsze w skutkach strzelaniny. Żeby to sprawdzić, poniżej przedstawiono dane o liczbie zabitych z podziałem na poszczególne incydenty (dla czytelności dołożono lekki szum losowy, żeby punkty nie nakładały się bezpośrednio na siebie).



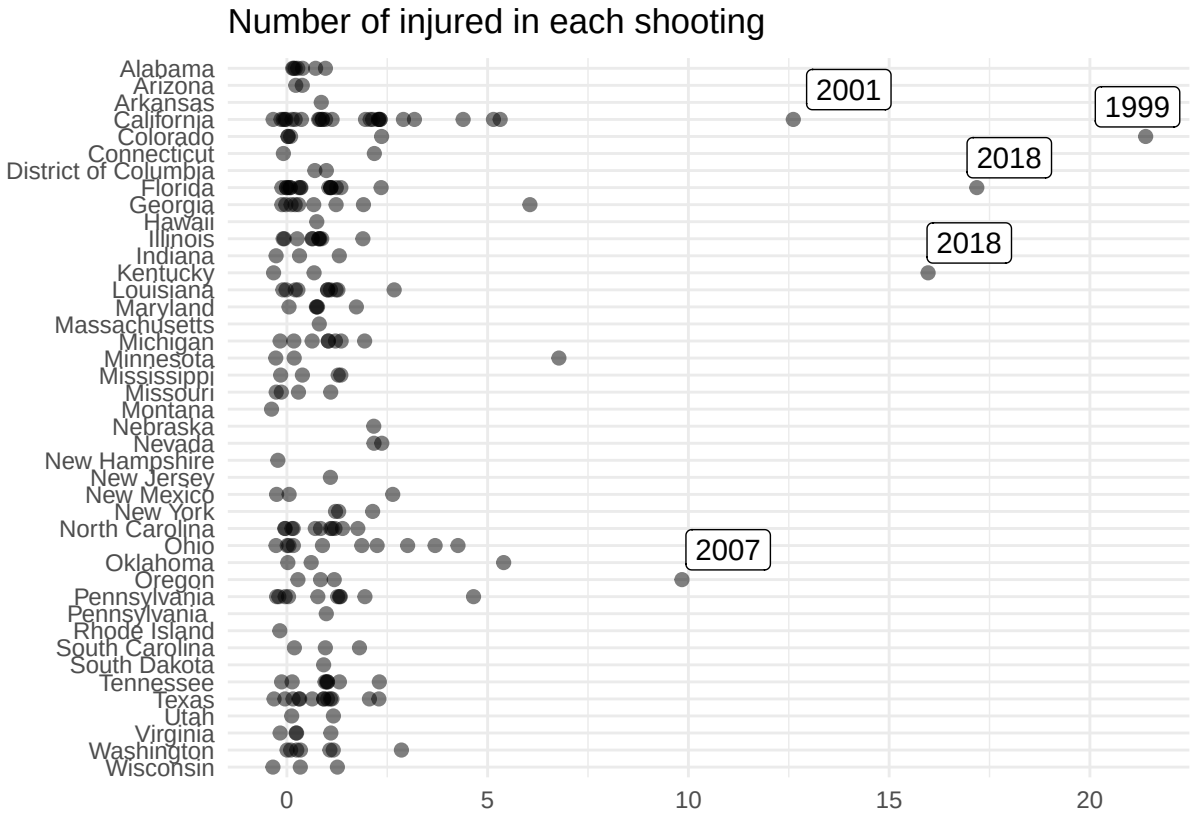
Z wykresu wynika, że rzeczywiście w Connecticut i na Florydzie miały miejsce strzelaniny z największą liczbą śmiertelnych ofiar.

Liczba rannych

W podobny sposób sprawdzono liczbę rannych w poszczególnych stanach.



Tutaj zdecydowanie przoduje California (zapewne ze względu na liczbę incydentów). Poniżej poszczególne strzelaniny z podziałem na stany.



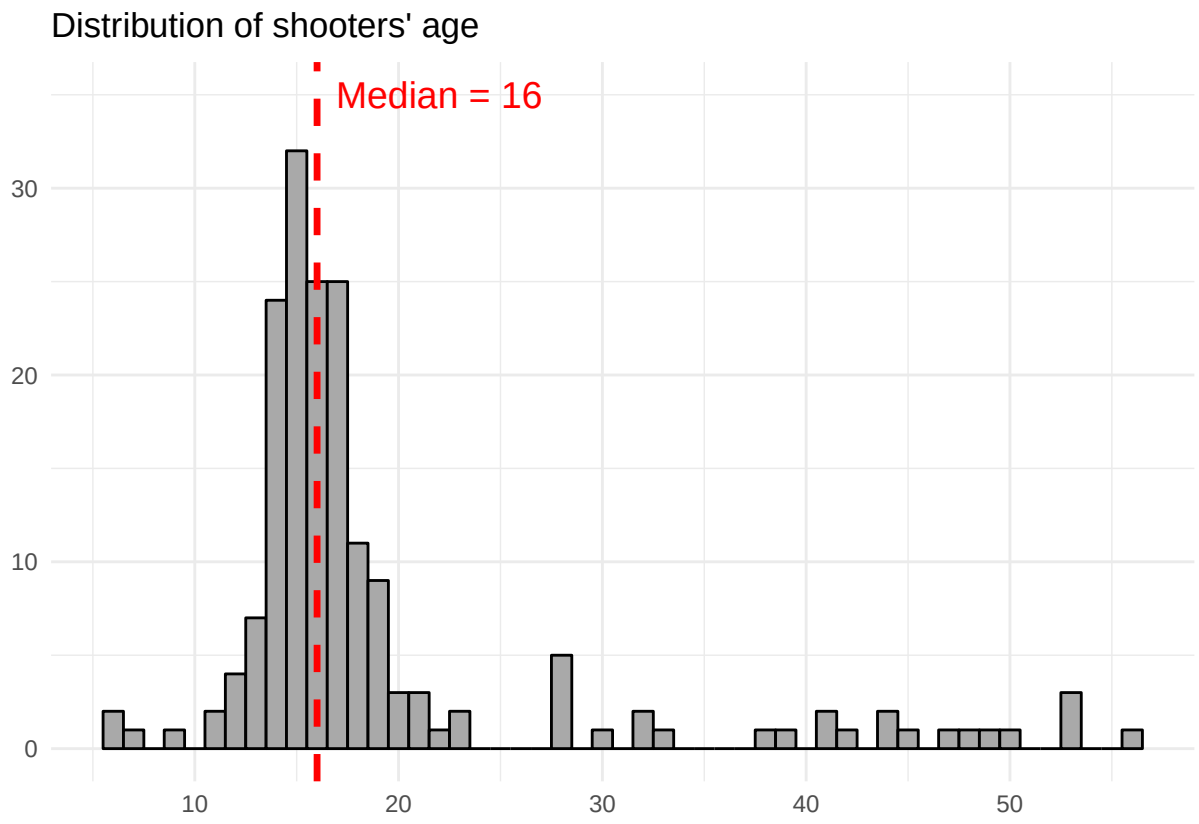
Okazuje się, że najwięcej rannych zostało w trakcie strzelaniny w Colorado w 1999 roku. Na drugim miejscu natomiast jest niedawna strzelanina na Florydzie.

Analiza sprawców

W tym miejscu analizę eksploracyjną skoncentrowano na sprawcach i typach ataków. Sprawdzono wiek oraz płeć sprawców, a także liczbę ofiar biorąc pod uwagę charakter ataku.

Wiek

W poniższym wykresie wykluczono sprawców, których wiek nie został określony.



Histogram wieku pokazuje prawostronną skośność jego rozkładu. Mediana wieku sprawców wynosi 16 lat (co jest zgodne z artykułem z The Washington Post).

Płeć

płeć	liczba ataków
nieznana	14
K	10
M	193

Jak pokazuje powyższe podsumowanie, zdecydowana liczba atakóW przeprowadzona była przez mężczyzn.

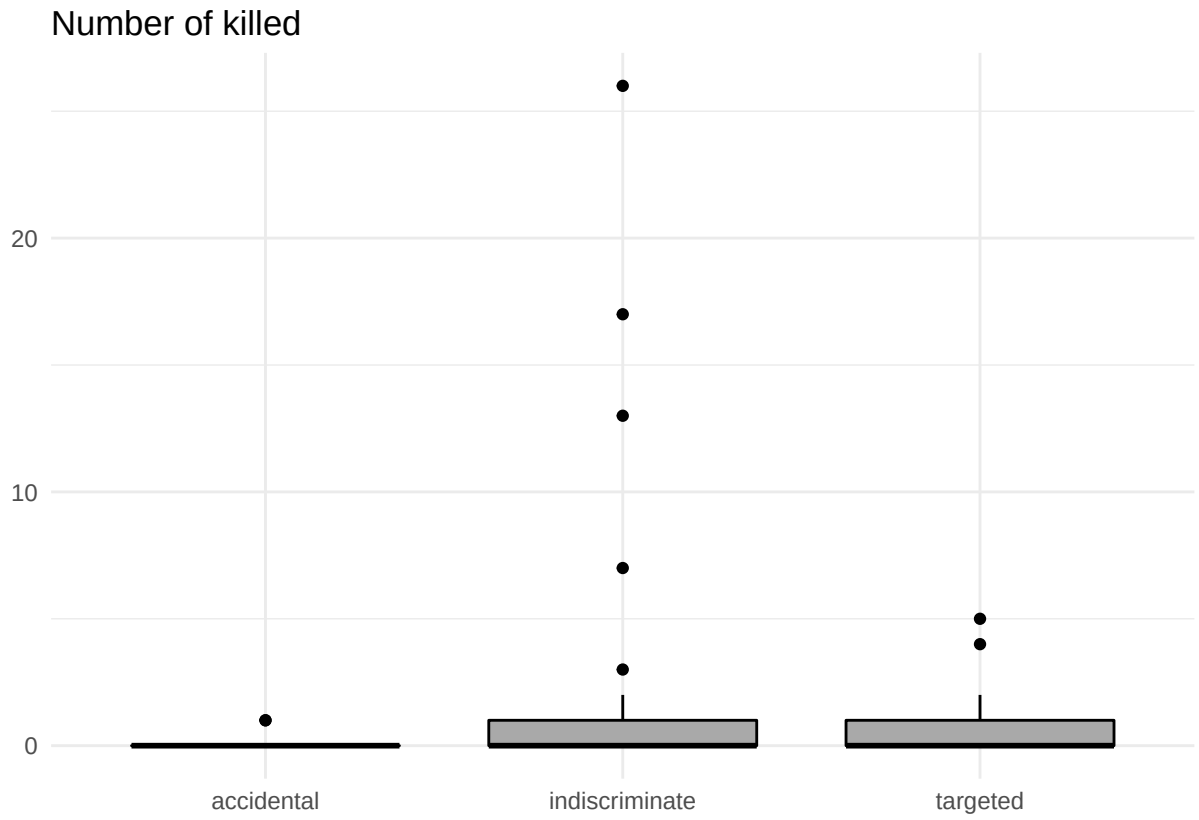
Rodzaj ataku

rodzaj ataku	liczebność
targeted	130
indiscriminate	42
accidental	26
targeted and indiscriminate	5
public suicide	4
unclear	4
accidental or targeted	2

rodzaj ataku	liczebność
hostage suicide	2
	1
public suicide (attempted)	1

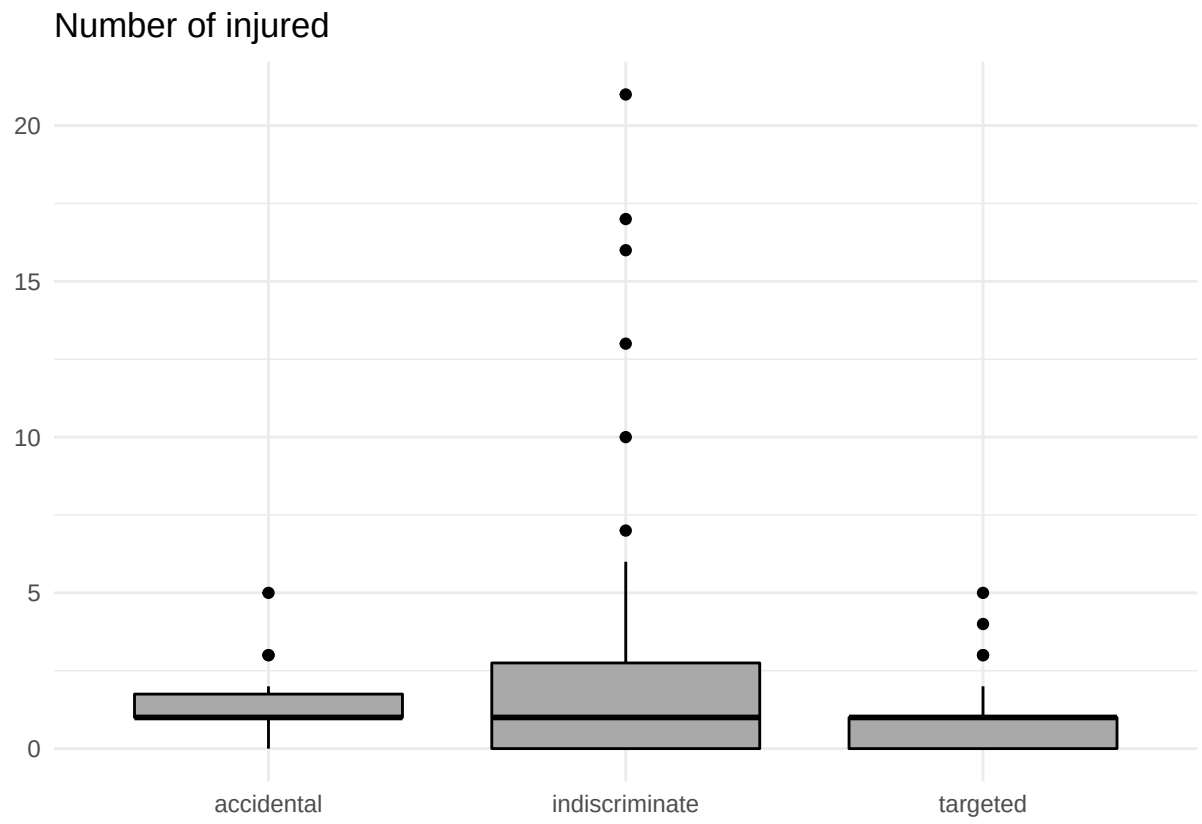
Zdecydowanie przważają ataki celowe (*targeted*, czyli z checią zabicia konkretnych osób), oprócz tego jeszcze często zdarzały się przypadki określone jako *indiscriminate* (strzelanie bez konkretnych celów) i strzały przypadkowe (*accidental*). Te trzy typy strzelanin zostaną wzięte do dalszej analizy.

Liczba zabitych



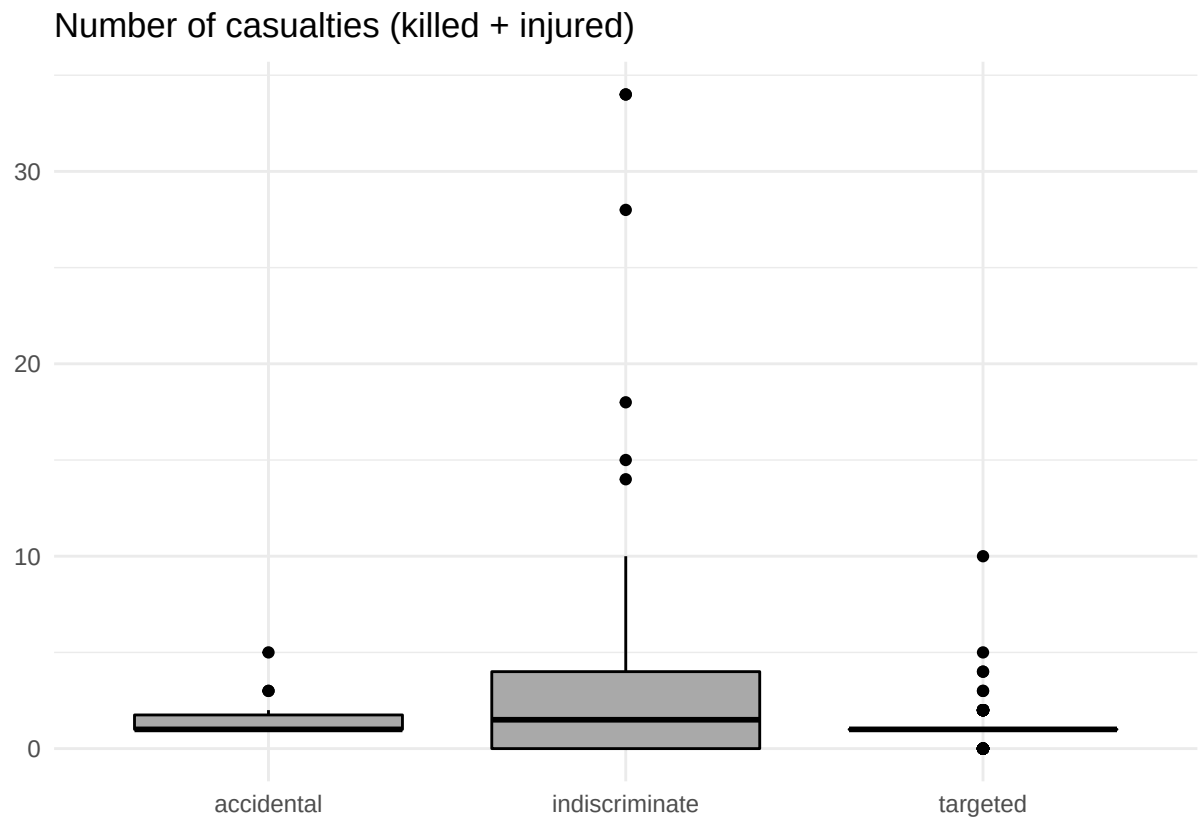
Przypadkowe strzelaniny (albo raczej wystrzały) rzadko są przyczyną śmierci. Najbardziej zabójcze wydają się ataki określone jako *indiscriminate*.

Liczba rannych



Pomimo takiej samej mediany, wyraźnie widać, że ataki typu *indiscriminate* skutkują największą liczbą rannych. Co ciekawe, liczba rannych w strzelaninach celowych i przypadkowych jest raczej podobna. Poniżej sprawdzono jeszcze liczbę wszystkich ofiar z podziałem na typy strzelanin.

Liczba ofiar ogółem

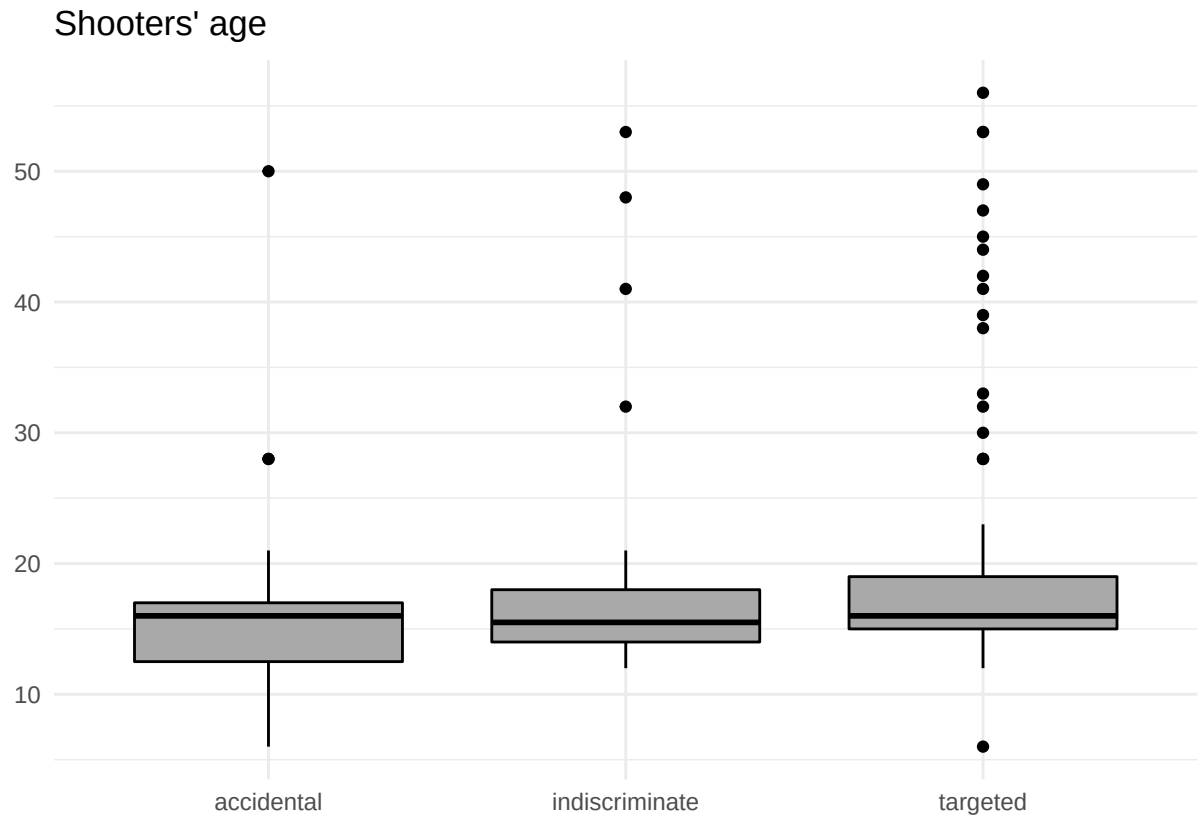


Potwierdza się wcześniejszy wniosek o tym, że najbardziej fatalne w skutkach są ataki typu *indiscriminate*. Za pomocą testu *Kruskala-Wallisa* sprawdzono statyczną istotność różnicy w rozkładach ofiar poszczególnych typów strzelanin (H_0 - rozkłady (a dokładniej mediany rozkładów) są takie same, H_1 - rozkłady są różne).

```
##
##  Kruskal-Wallis rank sum test
##
## data:  casualties by shooting_type
## Kruskal-Wallis chi-squared = 7.2317, df = 2, p-value = 0.02689
```

Ponieważ *p-value* < 0.05 można odrzucić hipotezę H_0 i przyjąć hipotezę, że rozkłady liczby ofiar tych ataków są rzeczywiście różne.

Wiek sprawców



Wydaje się, ze istnieje związek między wiekiem a typem strzelaniny - te celowe są przeprowadzane przez starszych, a te przypadkowe - przez młodszych (choć mediany tego nie pokazują, ale wykresy pudełkowe już tak).

Sprawdzono średnią wieku sprawców ze względu na typ strzelaniny:

rodzaj ataku	średni wiek
accidental	16.95
indiscriminate	18.50
targeted	20.18

Różnice w średnim wieku sprawców są wyraźne, ale ze względu na stosunkowo niewielką liczbę ataków innych niż *targeted* mogą okazać się nieistotne statystycznie. Podobnie jak poprzednio, wykorzystamy test *Kruskala-Wallisa* (anova zakłada normalność rozkładów, której nie sprawdzono, natomiast test *t-studenta* - w tym przypadku należałoby zastosować wielokrotny test - możliwy jest tylko dla dużych prób, a ataków typu *accidental* jest mniej niż 30).

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age_shooter1 by shooting_type
## Kruskal-Wallis chi-squared = 3.6636, df = 2, p-value = 0.1601
```

Uzyskana wartość *p-value* (0.16) nie pozwala na odrzucenie hipotezy H_0 (czyli nie można stwierdzić, że wiek sprawców poszczególnych typów ataków jest statystycznie różny).

Obecność ochroniarza

Ostatnim rodzajem analizy jest sprawdzenie, czy można zaobserwować związek między obecnością ochroniarza a liczbą ataków i ofiar.

ochroniarz obecny	liczba strzelanin	liczba zabitych	liczba rannych
0	143	68	130
1	74	63	143

I to dość ciekawe - dwa razy mniej strzelanin, gdy ochroniarz jest obecny, ale liczba zabitych i rannych jest podobna. Może to oznaczać (ale to tylko przypuszczenie), że obecność ochroniarza odstrasza potencjalnych sprawców, ale jeśli pomimo jego obecności zdecydują się na atak, to są lepiej do niego przygotowani.

Wnioski

(wnioski z przeprowadzonej analizy - krótkie podsumowanie najważniejszych obserwacji)

Co dalej?

Dalsze analizy mogłyby uwzględnić:

- strukturę uczniów w tych szkołach,
- rodzaj i pochodzenie broni (choć tutaj dane są trudniejsze do obróbki)