

Eksploracja danych

Agnieszka Kulesz, Hania Gibus, Igor Józefowicz





Opis zbioru

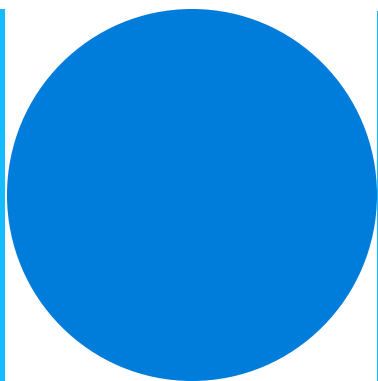
Gun violence data

Szczegółowe informacje o incydentach z użyciem broni palnej na terenie Stanów Zjednoczonych w latach 2013–2018



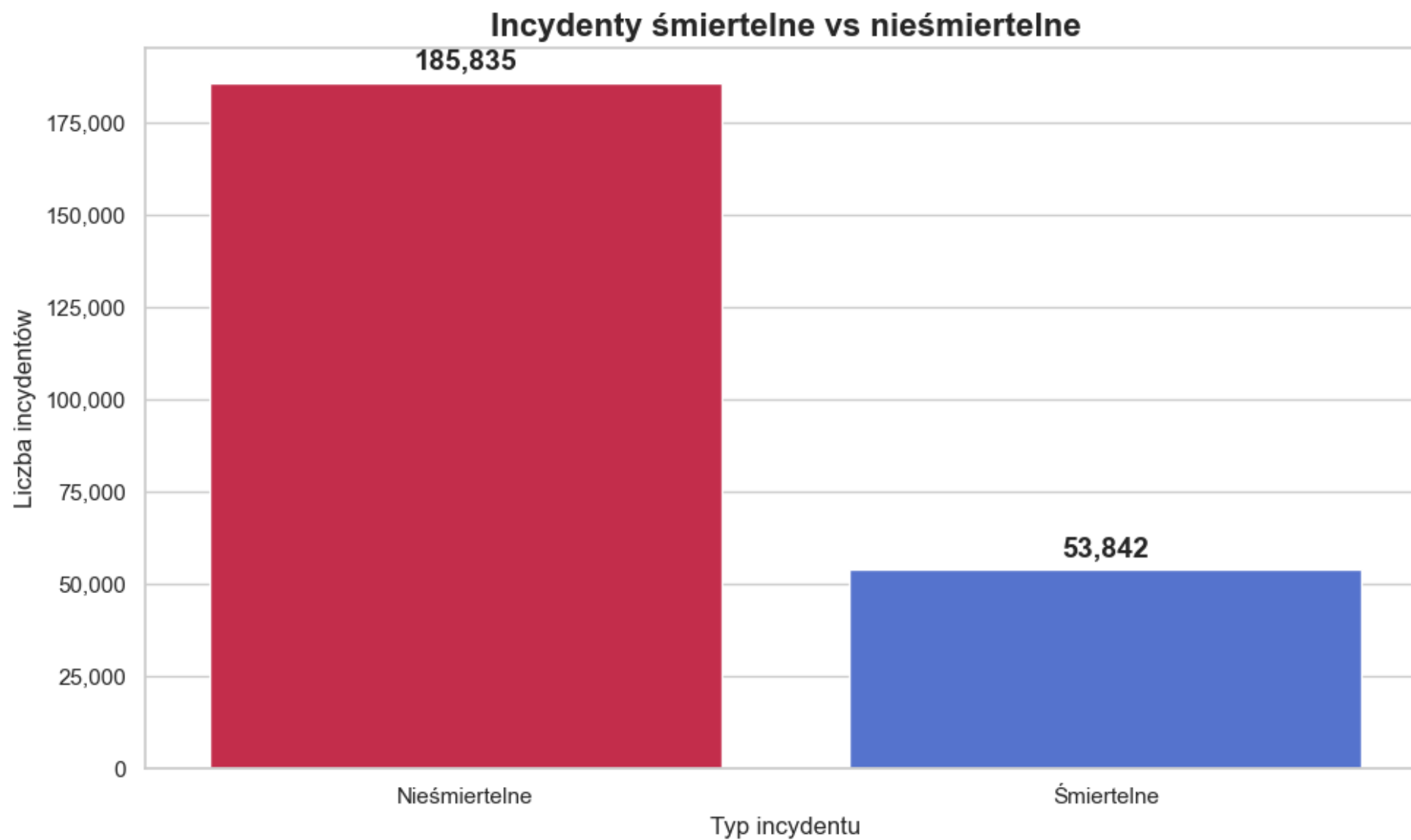
Cel eksploracji

- Predykcja, czy incydent z użyciem broni palnej zakończy się ofiarami śmiertelnymi
- Identyfikacja kluczowych czynników wpływających na ryzyko śmiertelności
- Maksymalizacja skuteczności modelu w wykrywaniu śmiertelnych incydentów



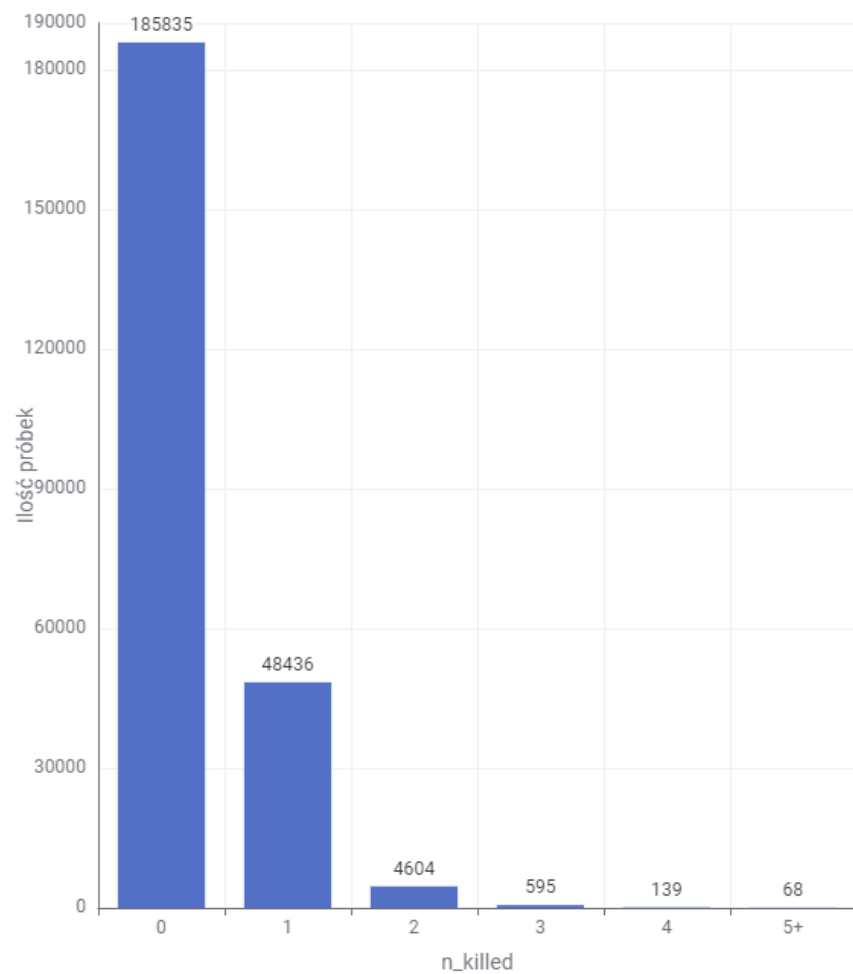
Rozkłady wartości atributów



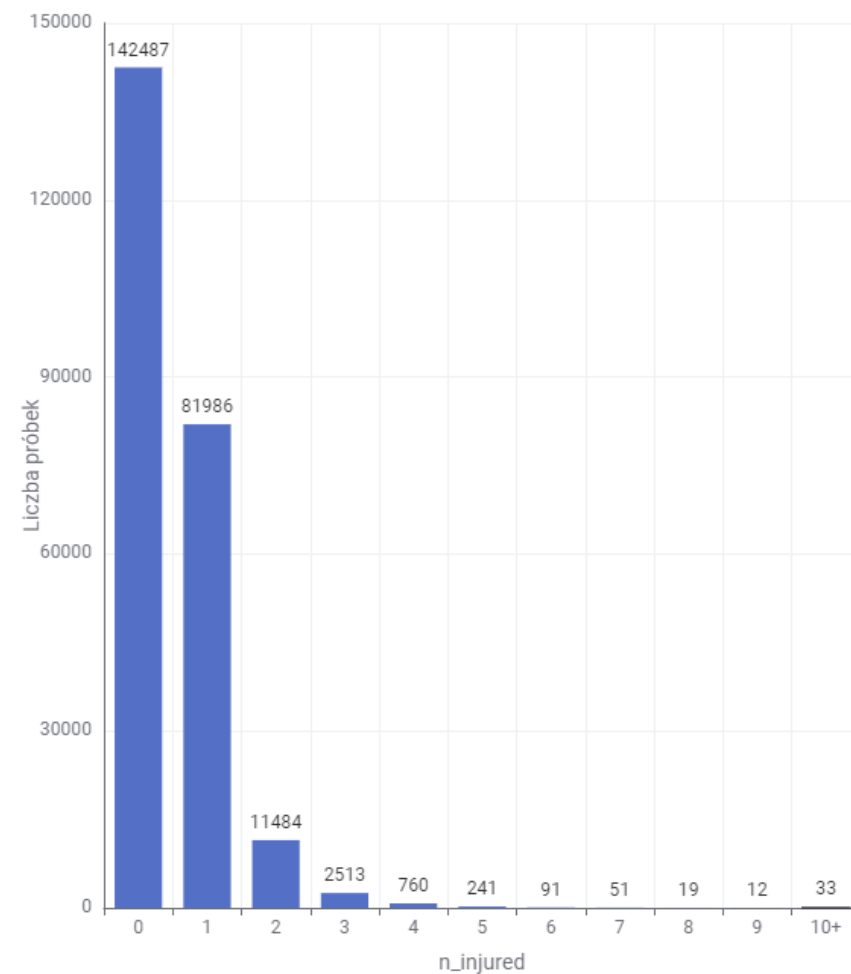




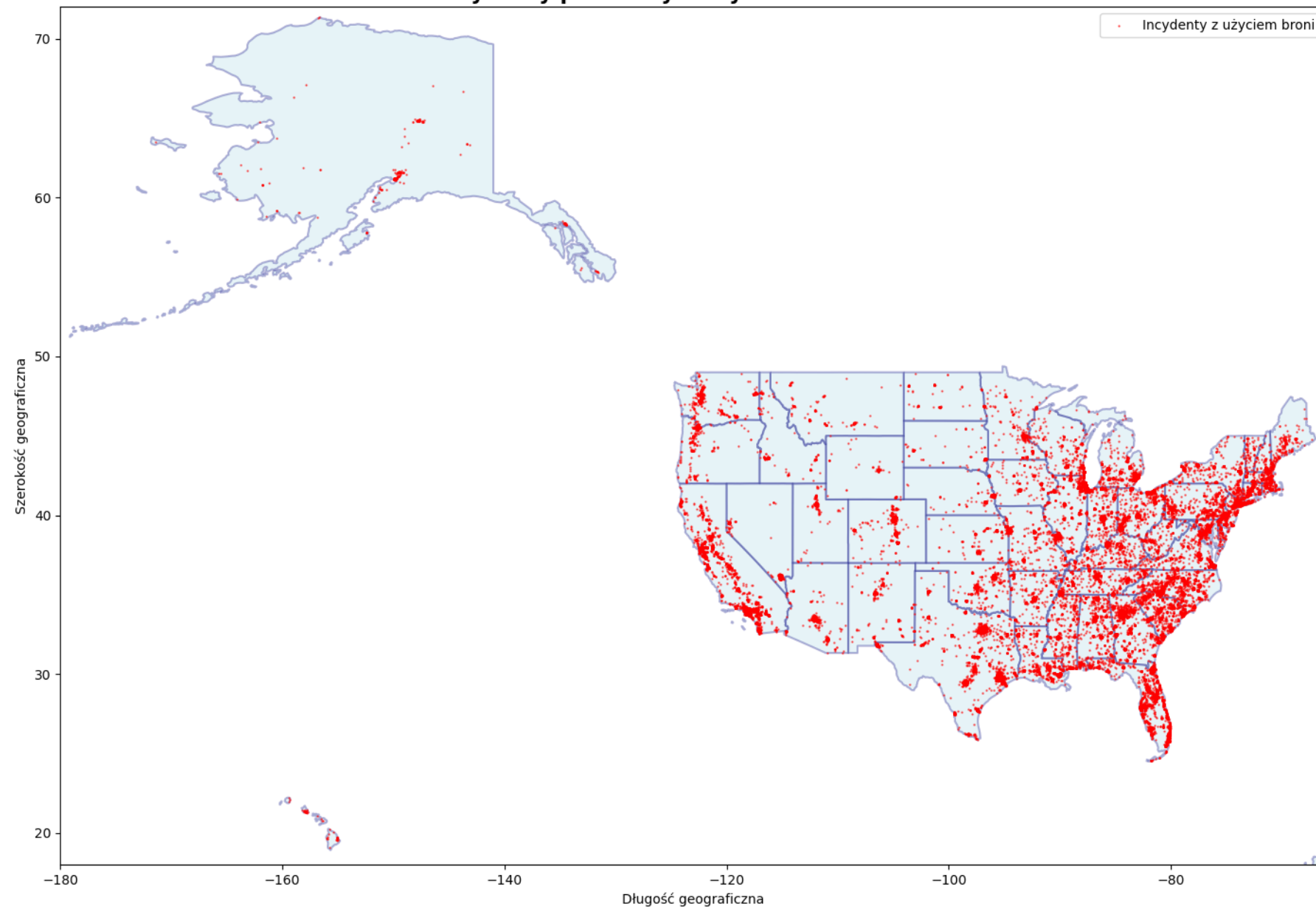
Histogram atrybutu n_killed

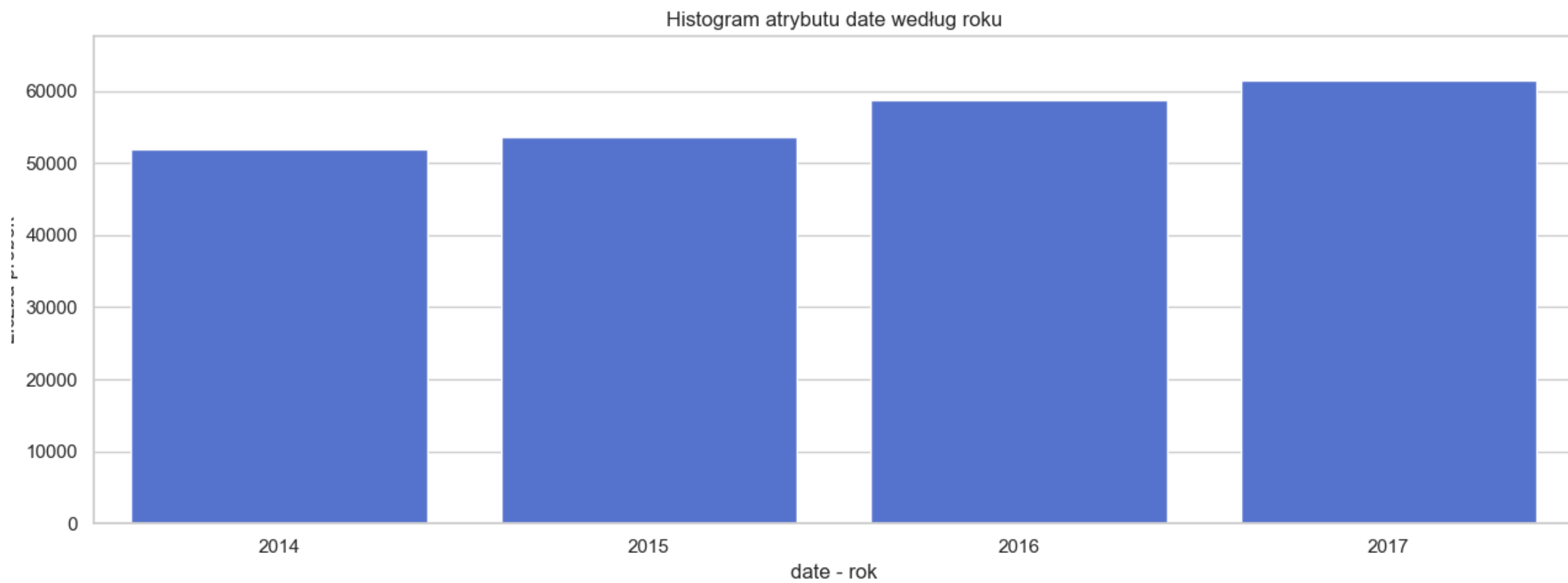


Histogram atrybutu n_injured



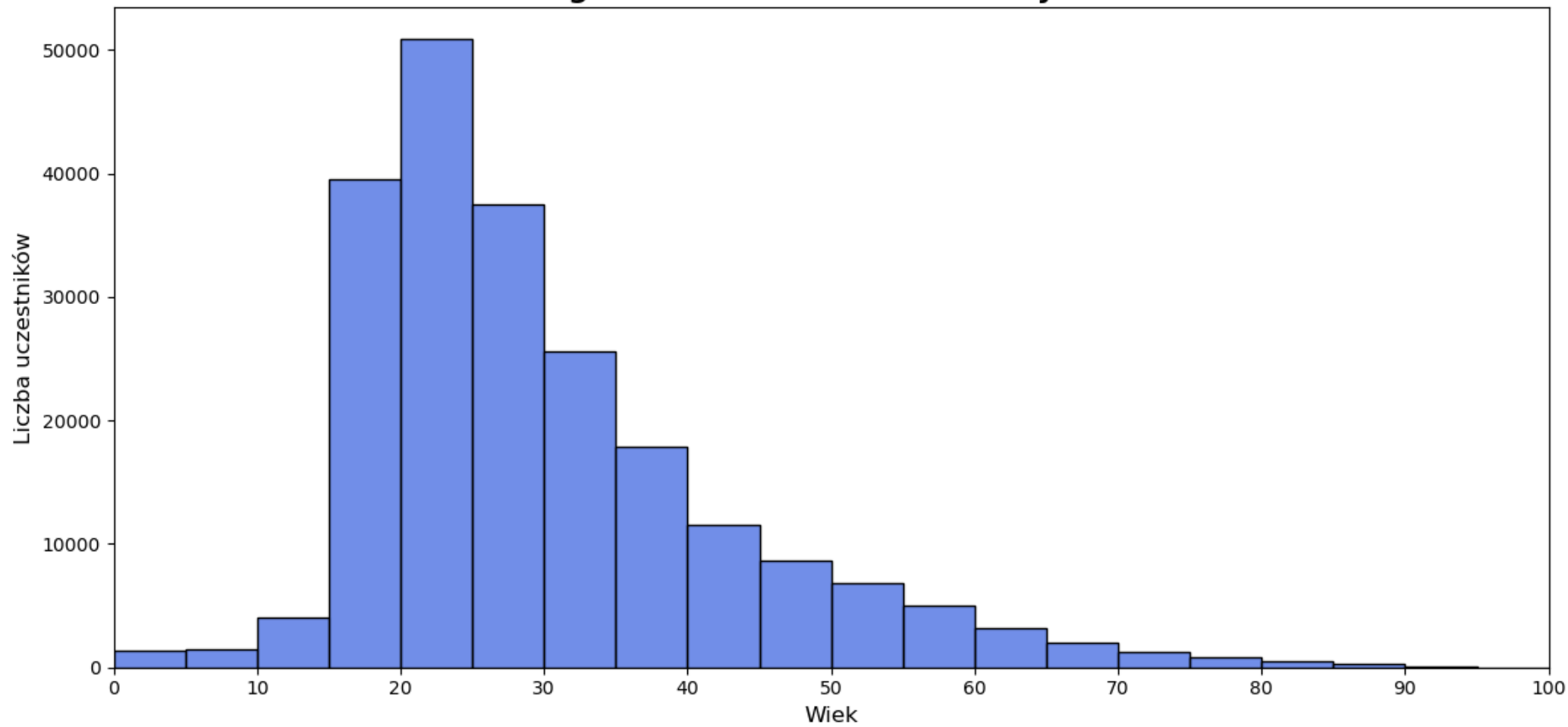
Incydenty przemocy z użyciem broni w USA



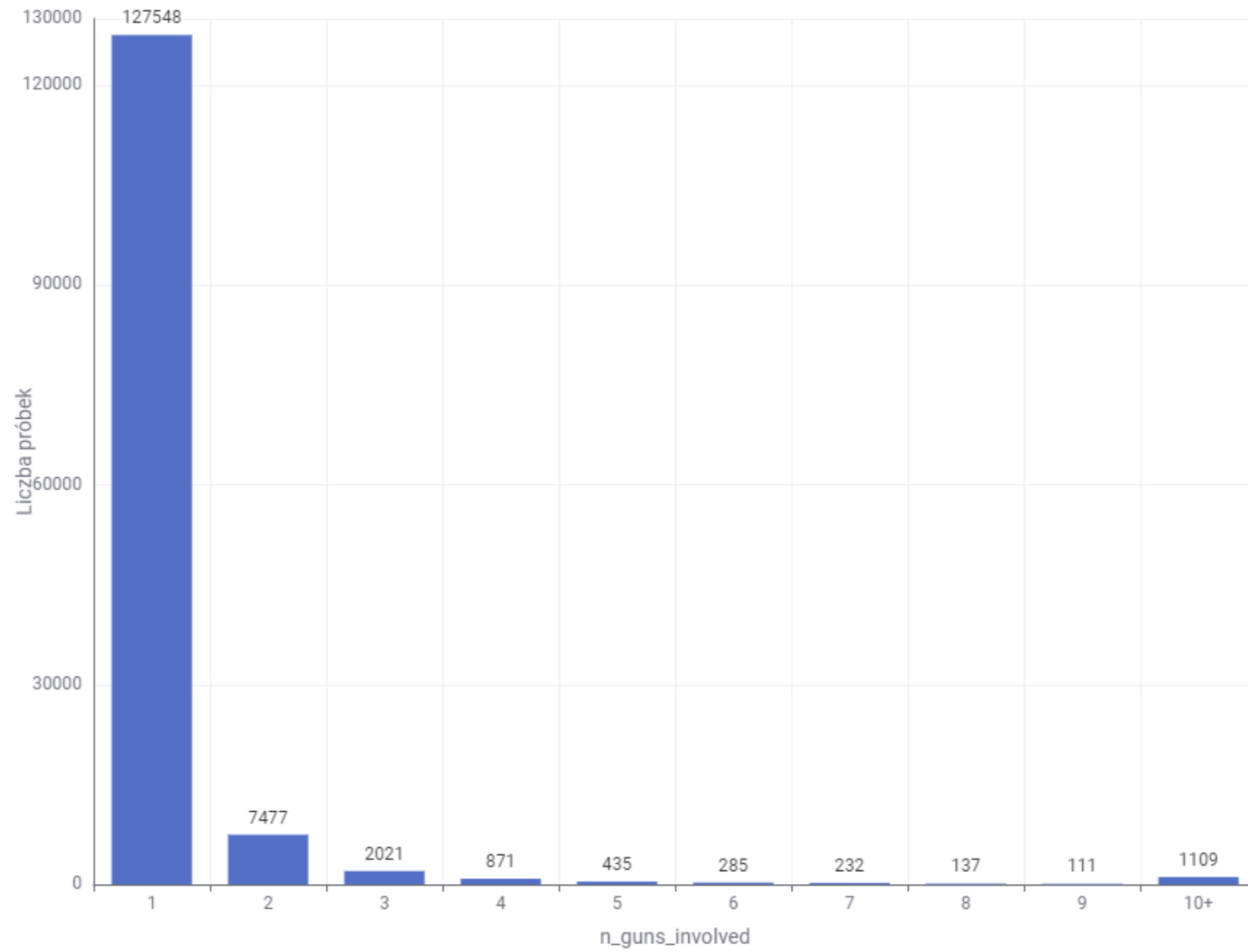




Histogram wieku uczestników incydentów

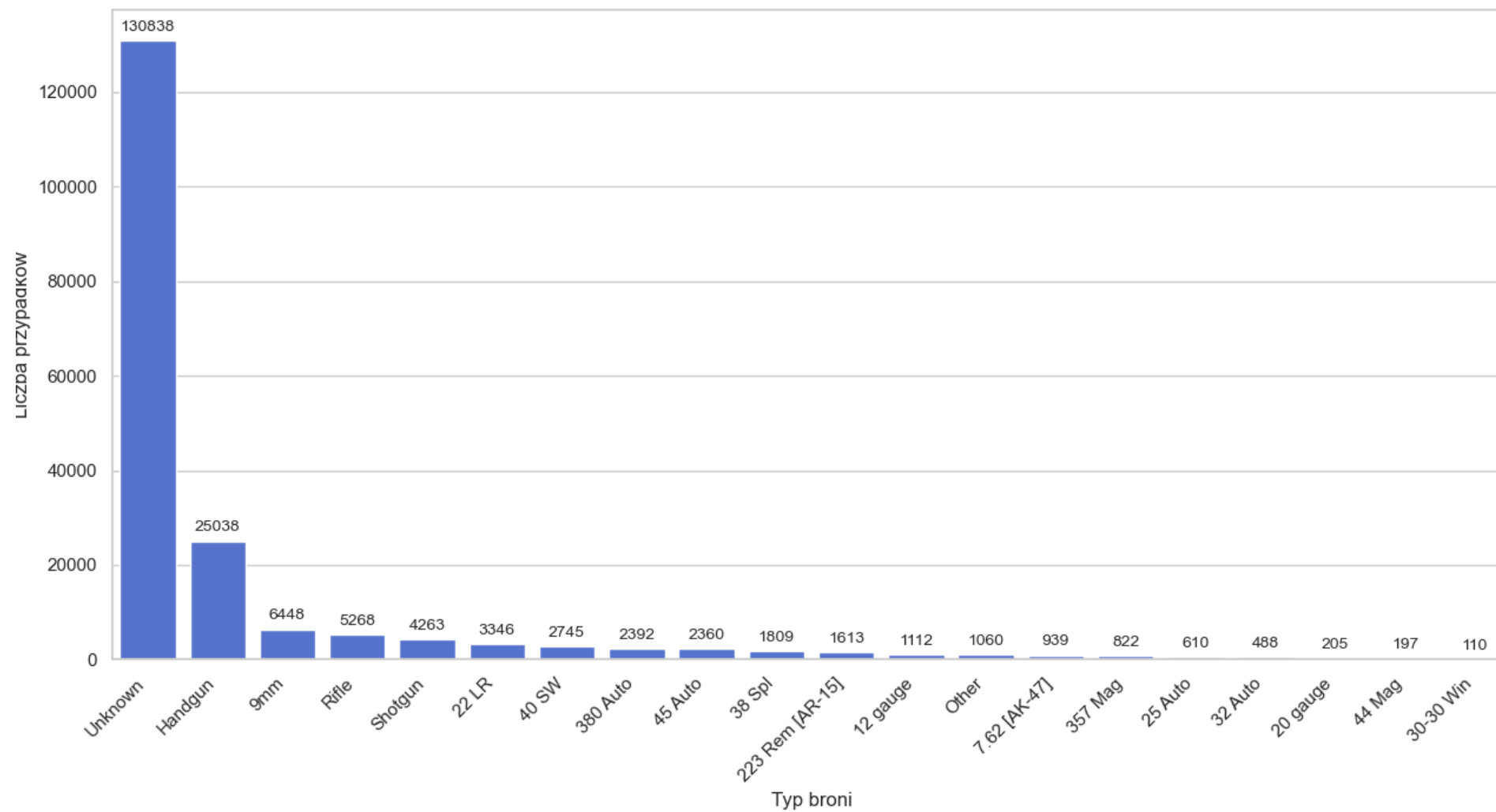


Histogram atrybutu n_guns_involved

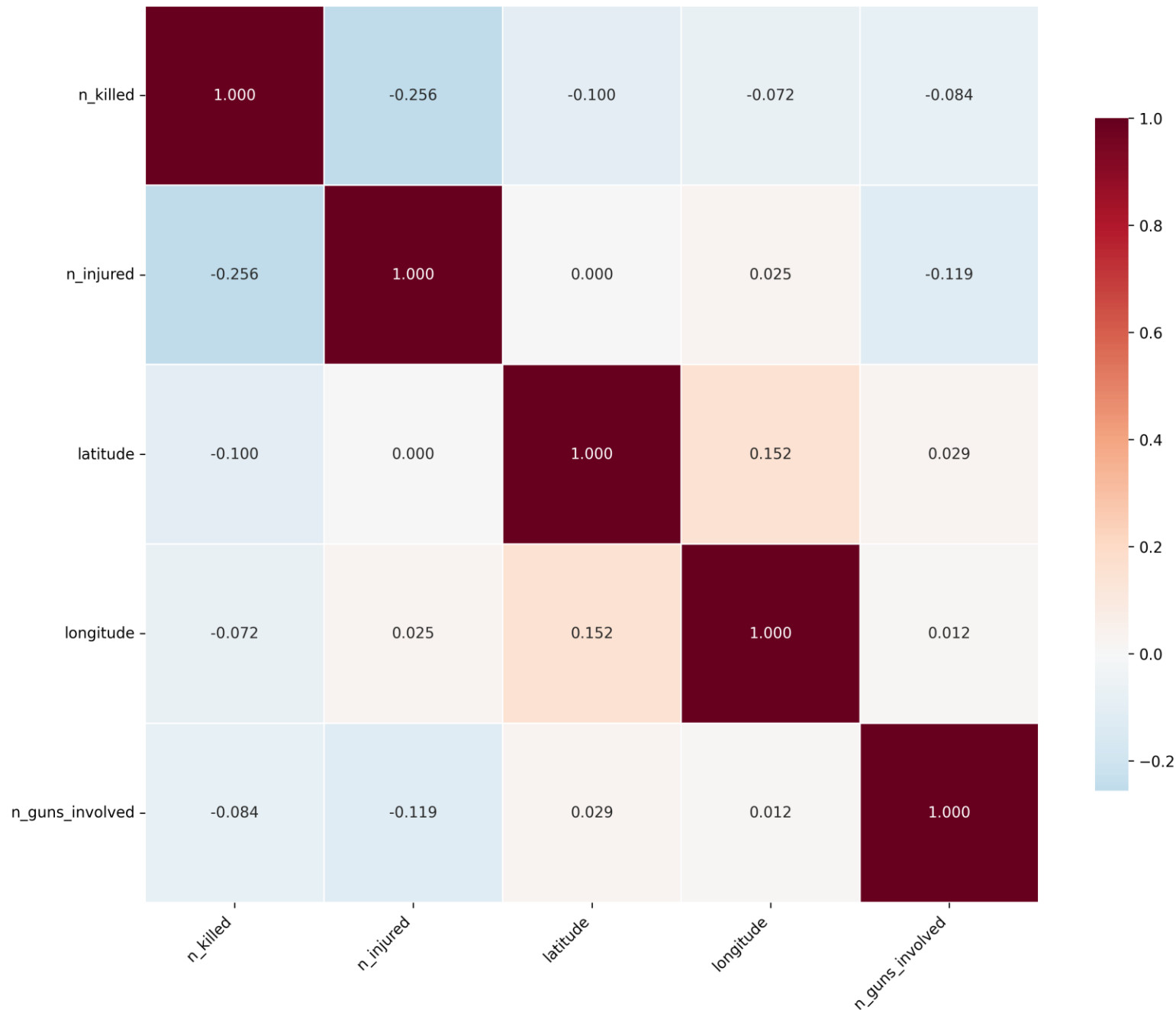




Liczba użyć różnych typów broni (Top 20)



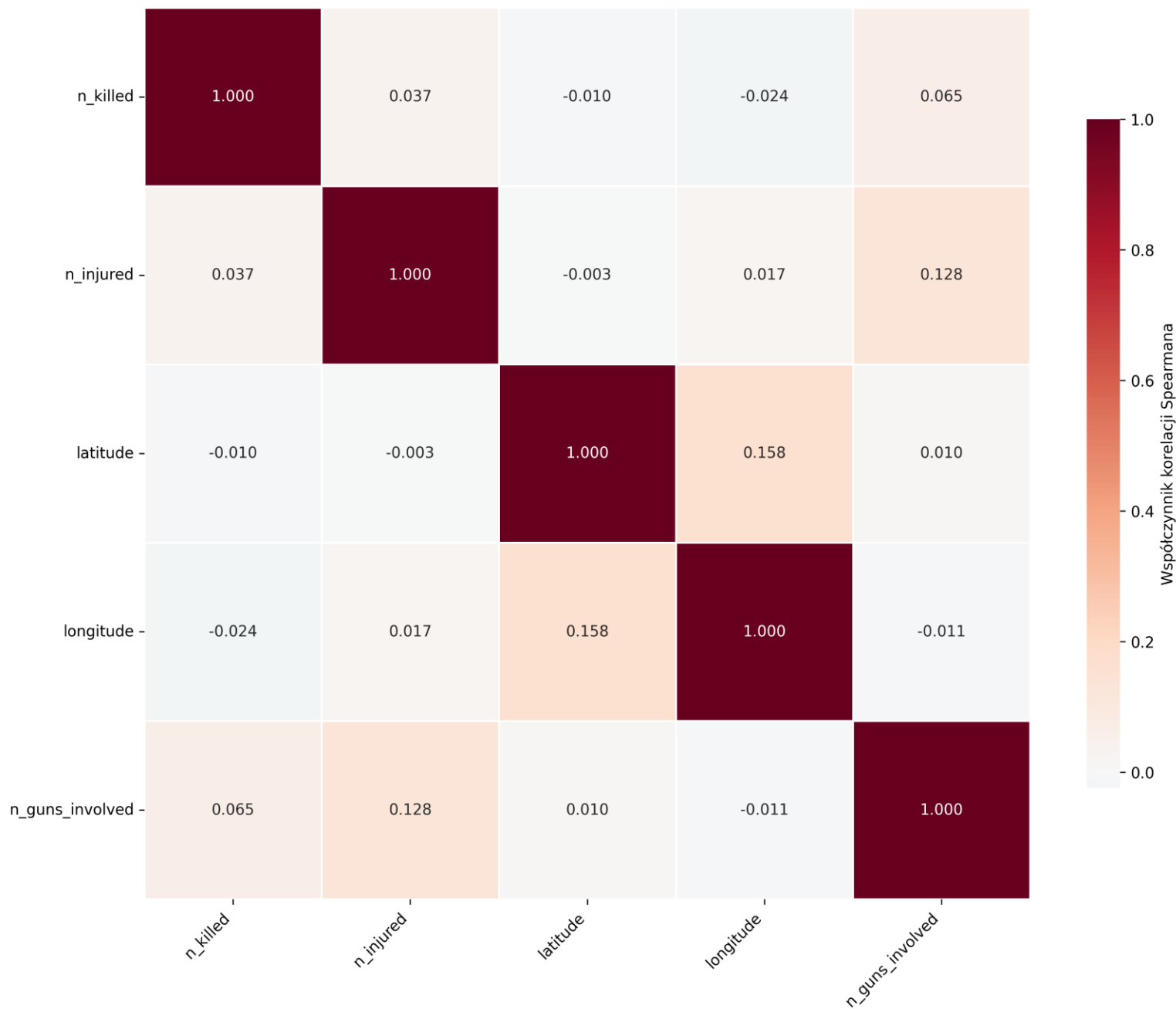
Analiza Korelacji Spearmana: Przemoc z Bronią Palną
(Ofiary śmiertelne, Ranni, Lokalizacja, Liczba broni)



Wszystkie przypadki

- zabici \rightleftharpoons ranni
 - słaba, ujemna
- broń \rightleftharpoons ofiary
 - zerowa korelacja
- geografia
 - słabe związki

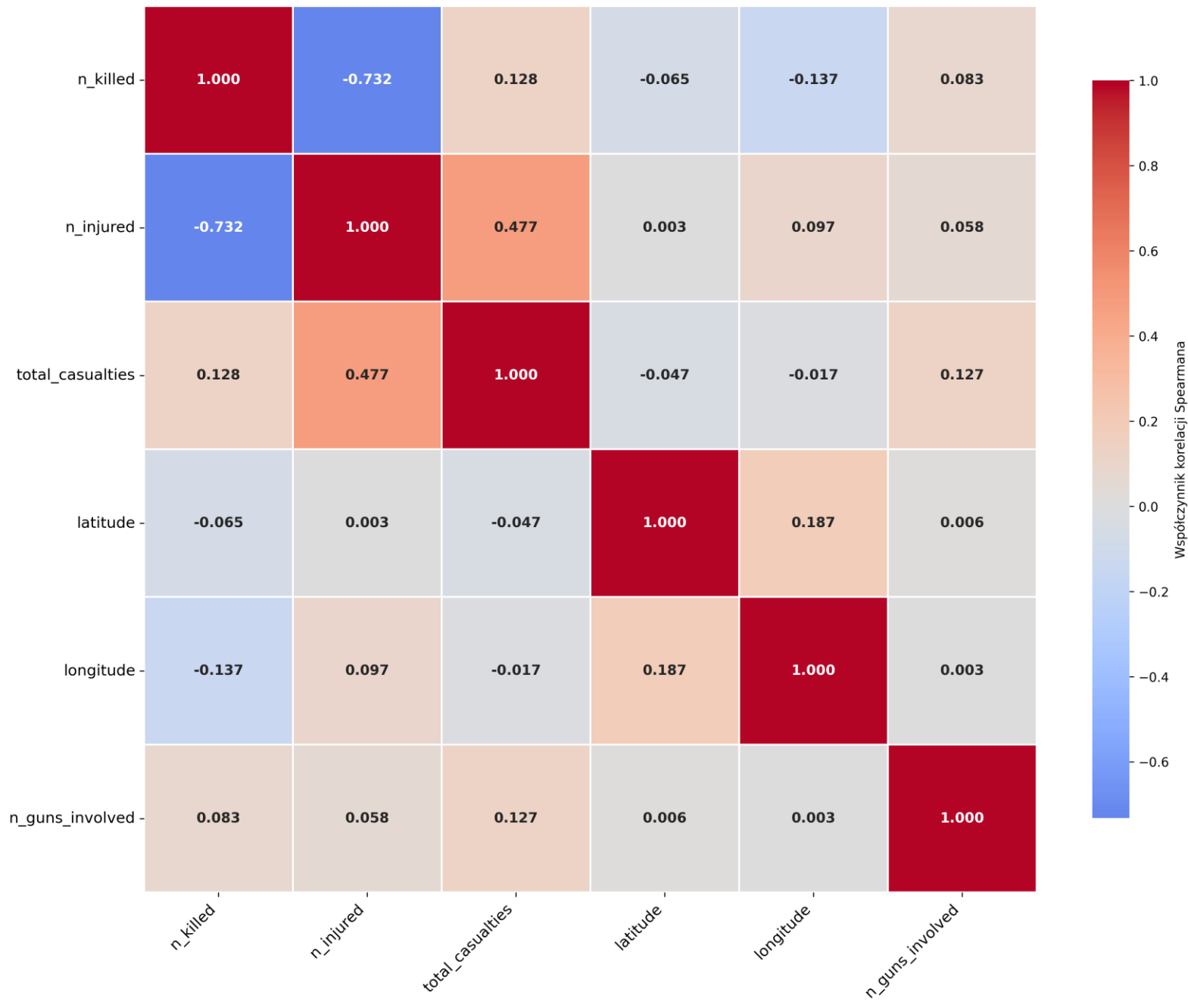
Analiza Korelacji Spearmana: Incydenty z Ofiarami Śmiertelnymi ($n_{\text{killed}} \geq 1$)



Śmiertelne incydenty


- zabici \rightleftharpoons ranni
 - $\rho \approx 0$
- broń \rightleftharpoons ofary
 - słabe dodatnie
- lokalizacja
 - brak monotonicznego trendu

Analiza Korelacji Spearmana: Mass Shootings (Incydenty z 4+ ofiarami)



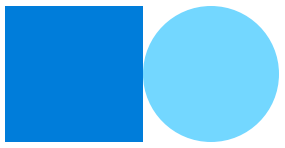
Masowe strzelaniny

- zabici \rightleftharpoons ranni
 - $\rho \approx -0.732$
- ranni \rightleftharpoons ofiary
 - $\rho \approx 0.477$
- broń \rightleftharpoons ofiary
 - słabe dodatnie
- geografia \rightleftharpoons ofiary
 - $|\rho| < 0.14$



Uwagi na temat jakości danych

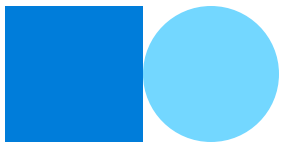
Brakujące dane, niespójności, niezrozumiałe formaty i wartości odstające



Brakujące dane

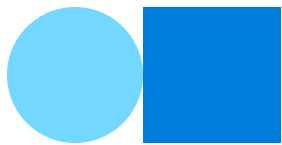
- Brak godziny zdarzenia
 - Brak analizy pór dnia
- Pola kategoryczne: „Unknown” zamiast NULL
 - Utrudnione filtrowanie

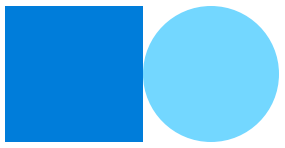




Niespójności

- city_or_county: miasto + hrabstwo
- Różne formaty nazw miejsc
- Wymaga rozdzielenia kolumn
- Problemy z łączeniem źródeł

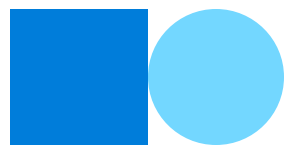




Formaty i normalizacja

- Separatory || / :: → rozbicie
- Pola listowe → JSON / tabele
- incident_characteristics: słownik normalizacji
- NULL vs kategorie: spójna konwencja





Outliers

- Orlando 2016, Las Vegas 2017
- Skrajne masowe zdarzenia
- Wpływ na średnie/odchylenie
- Oddzielna analiza

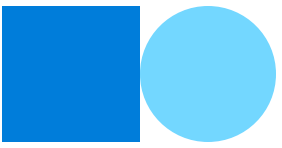




Przygotowanie danych

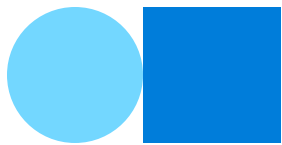
- Miary odporne: mediana, IQR
- Kodowanie kategoriycznych: one-hot, embedding
- Priorytet: czyszczenie braków
- Finalny zbiór gotowy do modelowania

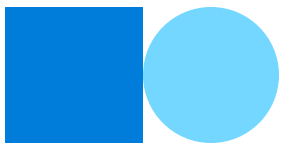




Wykorzystanie danych do eksploracji

- Dane obejmują szeroki zakres zmiennych i przypadków
- Braki dotyczą głównie mniej istotnych pól
- Wstępne czyszczenie kluczowe:
 - Uzupełnienie/oznaczenie braków
 - Rozdzielenie złożonych pól
 - Uwzględnienie outlierów
- Po przygotowaniu – eksploracja i wnioski





Podsumowanie

- Jakość danych: dobra, obszerny i reprezentatywny zbiór
- Problemy do rozwiązania:
 - Brakujące/niejednoznaczne wartości
 - Sporadyczne wartości odstające
- Po ujednoliceniu – solidna podstawa do odkrywania prawidłowości przemocy z użyciem broni





Dziękujemy za uwagę

