# Projekt z przedmiotu Eksploracja Danych

Etap I: Zrozumienie problemu i zrozumienie danych

**Temat: Gun Violence Data**

**Hanna Gibus, Igor Józefowicz, Agnieszka Kulesz**

## Ogólny opis zbioru

Zbiór danych zawiera szczegółowe informacje o incydentach z użyciem broni palnej na terenie Stanów Zjednoczonych w latach 2013–2018. Pojedynczy wiersz reprezentuje jeden incydent o unikalnym identyfikatorze. Dla każdego zdarzenia zgromadzono informacje na temat jego lokalizacji (miasto, stan, dokładny adres), daty zajścia, liczby ofiar śmiertelnych i rannych, a także cech charakterystycznych samego incydentu i uczestniczących w nim osób. Zbiór pozwala na analizę najczęściej występujących rodzajów przemocy z użyciem broni w różnych rejonach USA oraz okoliczności, w jakich do nich dochodzi.

## Określenie celu eksploracji i kryteriów sukcesu

Celem eksploracji jest predykcja czy incydent z użyciem broni palnej zakończy się ofiarami śmiertelnymi *(n_killed > 0)*. Docelowo rozwiązywanym problemem będzie klasyfikacja binarna:

- **klasa pozytywna** – wystąpiła co najmniej jedna ofiara śmiertelna
- **klasa negatywna** – brak ofiar śmiertelnych

Dodatkowym celem jest określenie, które atrybuty mają największy wpływ na ryzyko śmiertelności w incydencie – przykładowo typ użytej broni, lokalizacja, liczba zaangażowanych osób lub sztuk broni.

W przypadku danego problemu kluczową metryką będzie czułość, która informuje o tym, jak dobrze model identyfikuje incydenty zakończone śmiercią. Błąd fałszywie negatywny (czyli zaklasyfikowanie śmiertelnego incydentu jako nieskutkującego ofiarami) może oznaczać zignorowanie potencjalnie niebezpiecznego wzorca zachowań lub warunków, dlatego czułość będzie miała wyższy priorytet niż swoistość.

Wzór metryki czułości:

$$CZUŁOŚĆ = \frac{TP}{TP + FN}$$

gdzie:

- ***TP*** *(True Positive)* – liczba prawidłowo rozpoznanych śmiertelnych incydentów
- ***FN*** *(False Negative)* – liczba błędnie sklasyfikowanych śmiertelnych incydentów jako nieszkodliwe

Jednakże, wykorzystanie samej czułości może być niewystarczające, ponieważ ignoruje jakość klasyfikacji incydentów bez ofiar. Dlatego dodatkowo zastosowana zostanie metryka swoistości, która mierzy poprawność klasyfikacji przypadków negatywnych.

$$SWOISTOŚĆ = \frac{TN}{TN + FP}$$

gdzie:

- ***TN*** *(True Negative)* – liczba poprawnie rozpoznanych nieszkodliwych incydentów
- ***FP*** *(False Positive)* – liczba incydentów nieszkodliwych błędnie sklasyfikowanych jako śmiertelne

Ewentualny błąd fałszywie pozytywny (czyli uznanie niegroźnego incydentu za śmiertelny) może skutkować zwiększoną uwagą służb, co nie jest krytyczne, dlatego ważniejszą metryką w analizie będzie czułość.

Sukces eksploracji zostanie osiągnięty, jeśli model uzyska:

- Czułość ≥ 80%
- Swoistość ≥ 60%

# Charakterystyka zbioru danych

Pochodzenie:
https://www.kaggle.com/datasets/jameslko/gun-violence-data

Format:
.csv

Liczba przykładów:
239 677

Ilość zbiorów danych:
1

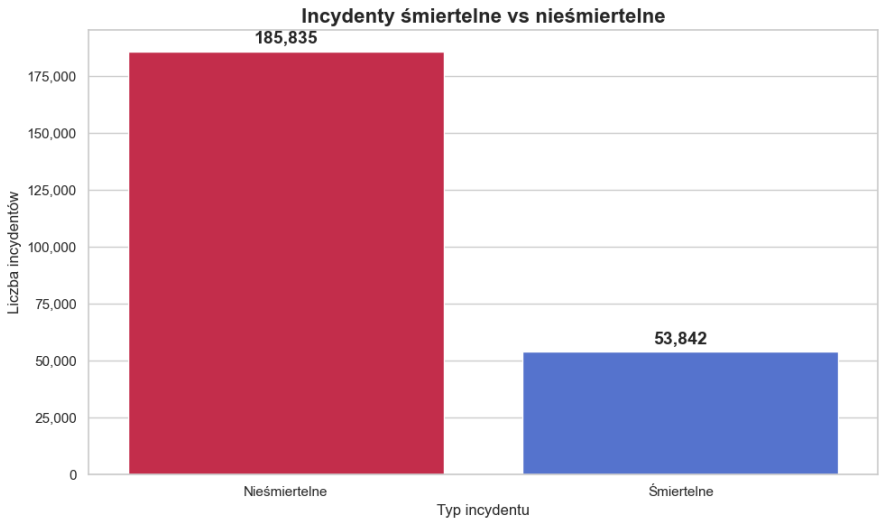
# Opis atrybutów

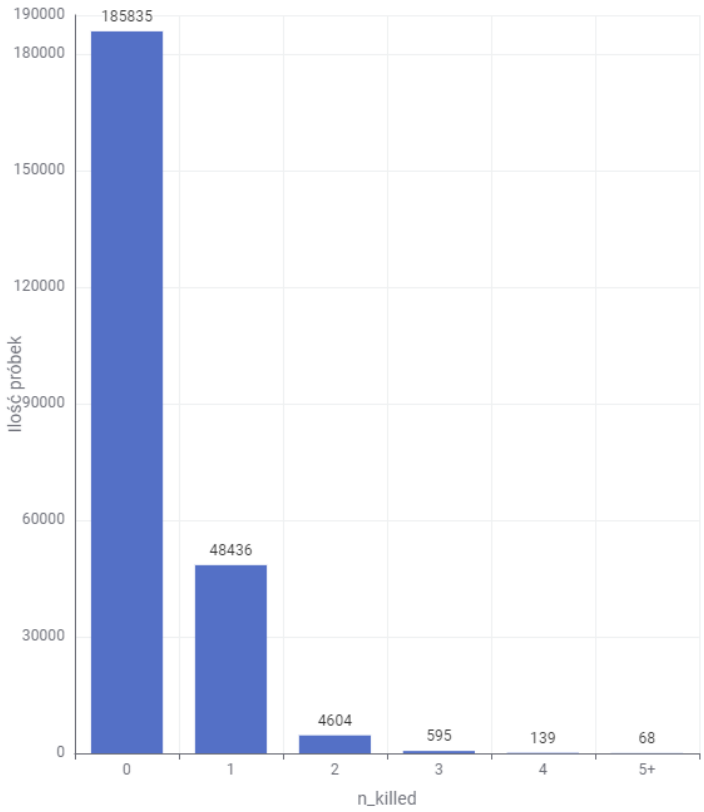
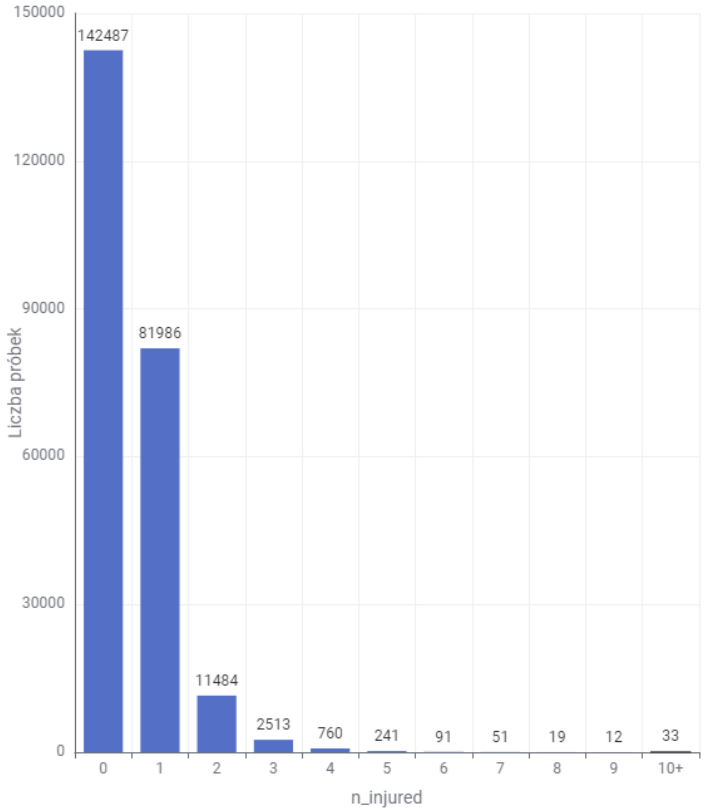
| ***Nazwa*** | ***Typ*** | ***Znaczenie*** |
|---|---|---|
| incident_id | Numeryczny | Unikalny identyfikator incydentu w bazie danych Gun Violence Archive |
| date | Data | Data wystąpienia incydentu w formacie YYYY-MM-DD |
| state | Tekstowy | Stan USA, w którym miał miejsce incydent |
| city_or_county | Tekstowy | Miasto lub hrabstwo, gdzie doszło do incydentu |
| address | Tekstowy | Dokładny adres miejsca incydentu |
| n_killed | Numeryczny | Liczba osób zabitych w wyniku incydentu |
| n_injured | Numeryczny | Liczba osób rannych w wyniku incydentu |
| incident_url | Tekstowy | Link do oryginalnego raportu o incydencie |
| source_url | Tekstowy | Link do źródła informacji (artykuł prasowy, raport policyjny itp.) |
| incident_url_fields_missing | Logiczny | Informacja czy brakuje informacji w polu incident_url (True/False) |
| congressional_district | Numeryczny | Numer okręgu wyborczego do Kongresu USA |
| gun_stolen | Tekstowy | Status broni użytej w incydencie (np. stolen (skradziona), unknown (nieznany)) |
| gun_type | Tekstowy | Typ użytej broni palnej |
| incident_characteristics | Tekstowy | Charakterystyka incydentu |
| latitude | Numeryczny | Szerokość geograficzna miejsca incydentu |
| location_description | Tekstowy | Opis lokalizacji |
| longitude | Numeryczny | Długość geograficzna miejsca incydentu |
| n_guns_involved | Numeryczny | Liczba sztuk broni zaangażowanych w incydent |
| notes | Tekstowy | Dodatkowe notatki i szczegóły dotyczące incydentu |
| participant_age | Tekstowy | Wiek uczestników incydentu |
| participant_age_group | Tekstowy | Grupa wiekowa uczestników |
| participant_gender | Tekstowy | Płeć uczestników incydentu |
| participant_name | Tekstowy | Imiona uczestników (jeśli podane do publicznej wiadomości) |
| participant_relationship | Tekstowy | Relacja między uczestnikami (np. Random, Family, Acquaintance) |
| participant_status | Tekstowy | Zakres szkód wyrządzonych uczestnikowi (np. Arrested, unharmed, injured, killed) |
| participant_type | Tekstowy | Typ uczestnika |

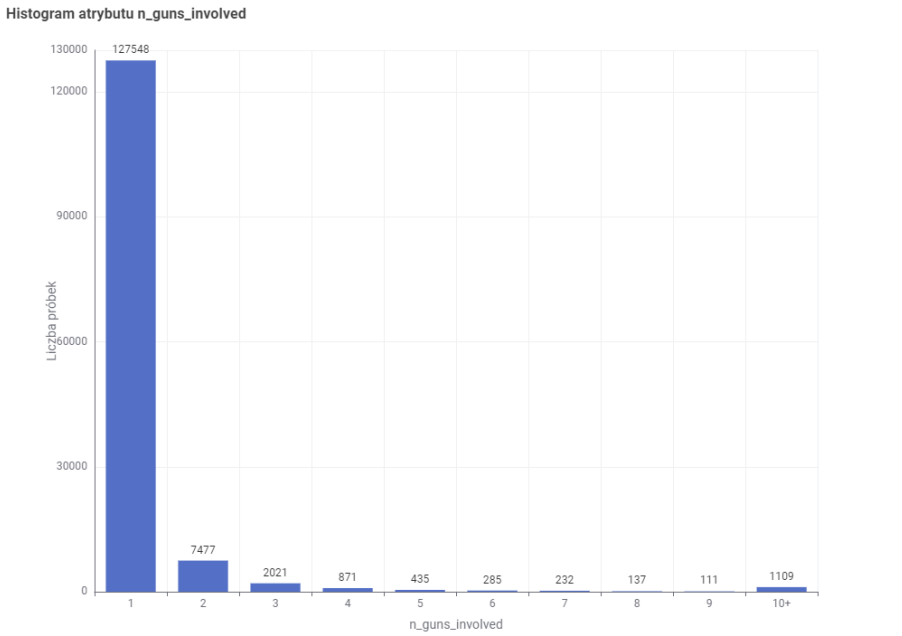
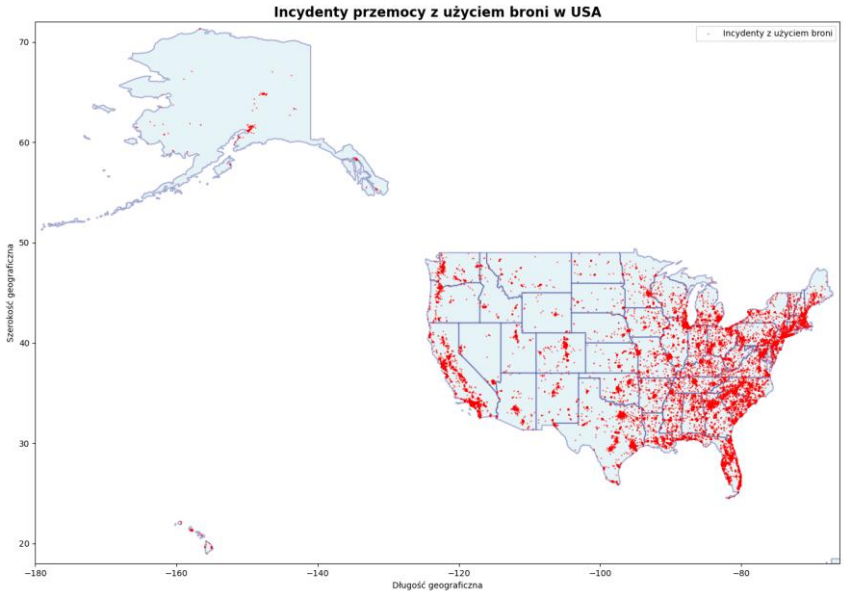
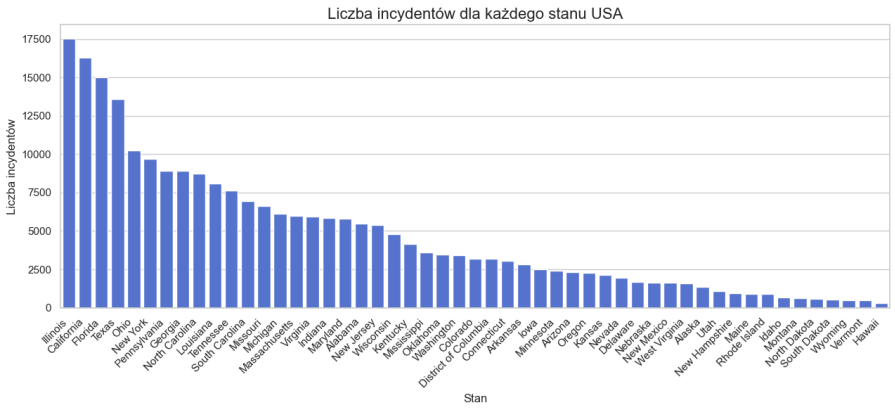
| | | |
|---|---|---|
| sources | Tekstowy | Źródła informacji o uczestnikach |
| state_house_district | Numeryczny | Numer okręgu wyborczego do stanowej izby reprezentantów |
| state_senate_district | Numeryczny | Numer okręgu wyborczego do stanowego senatu |

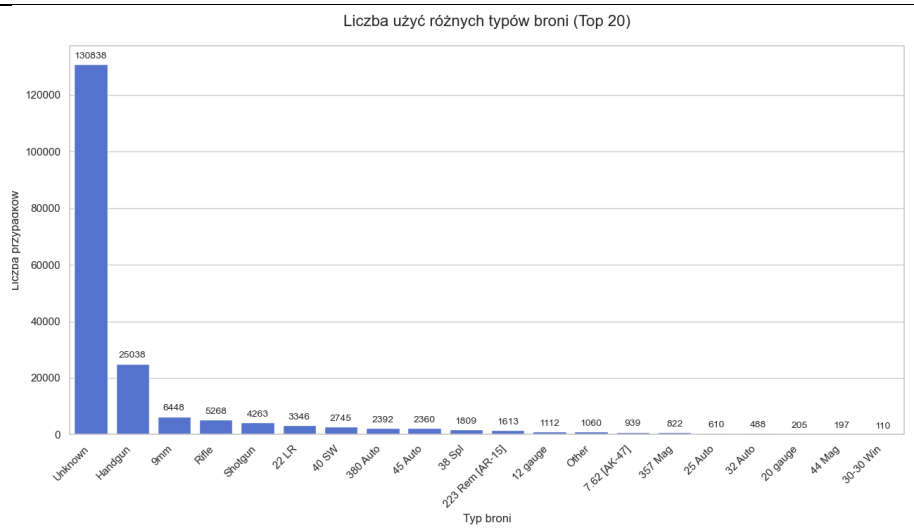
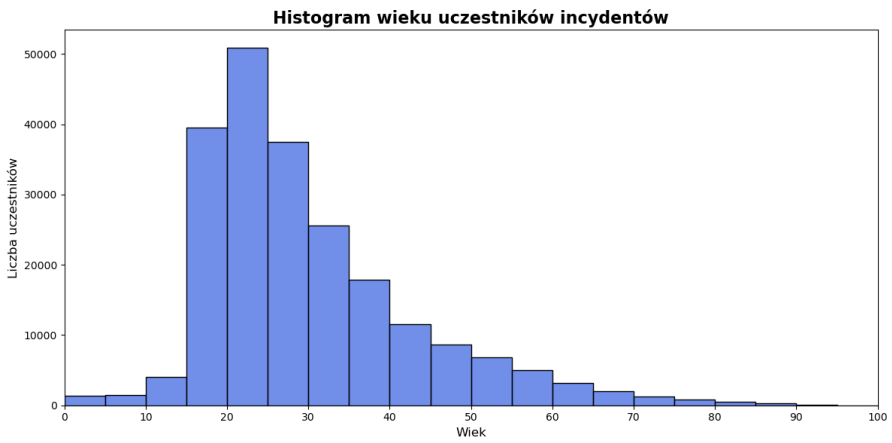
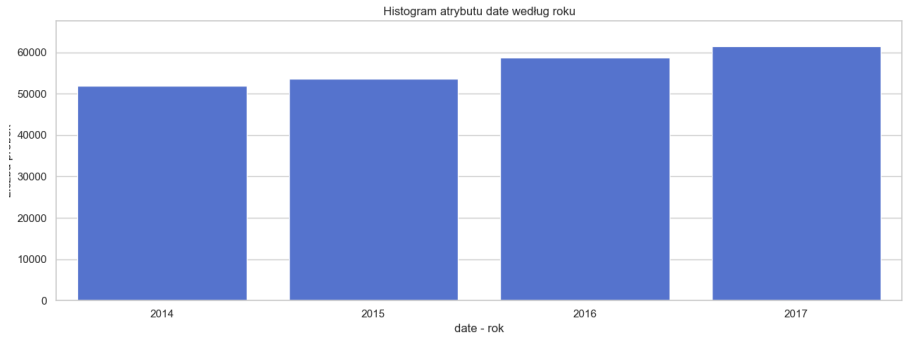
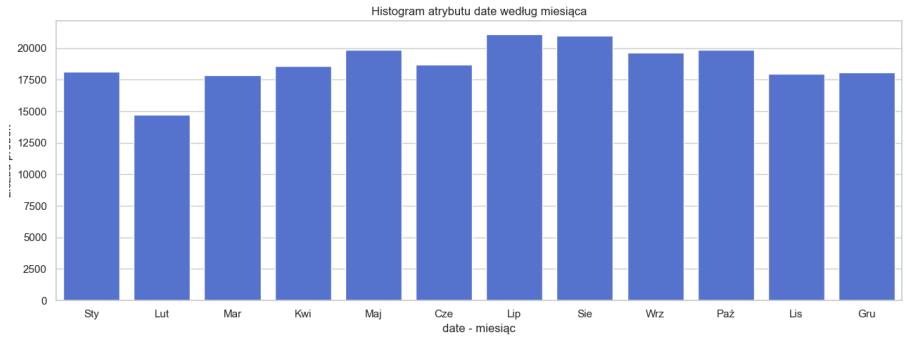
# Wyniki eksploracyjnej analizy danych

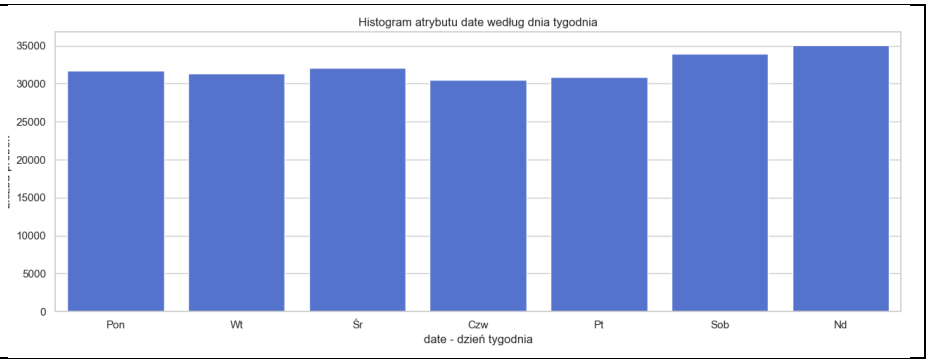
## Rozkłady wartości atrybutów

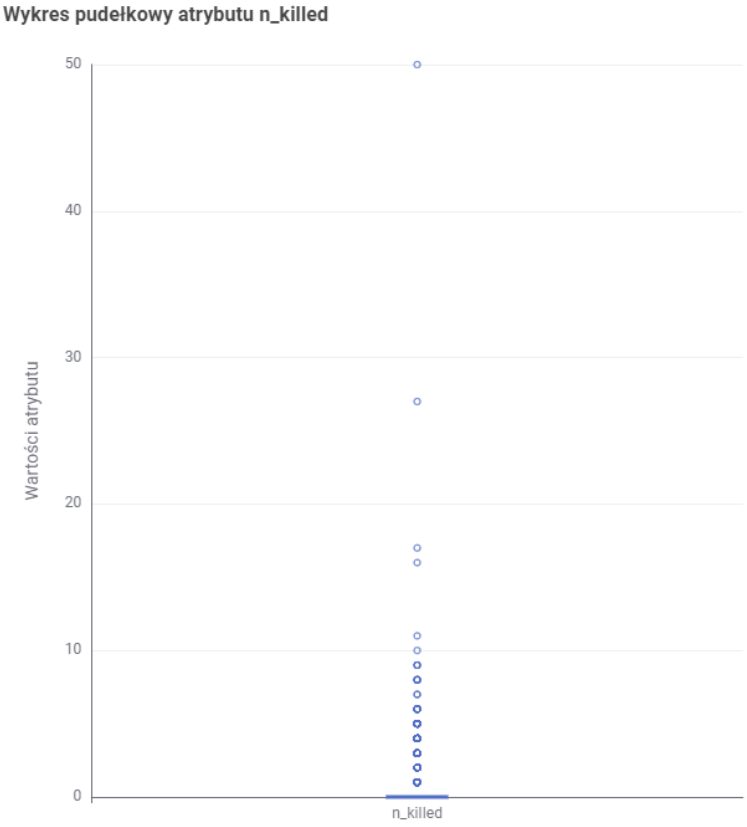
| Atrybut | Histogram |
|---|---|
| mortality<br>*(n_killed > 0)* |  |
| n_killed | |

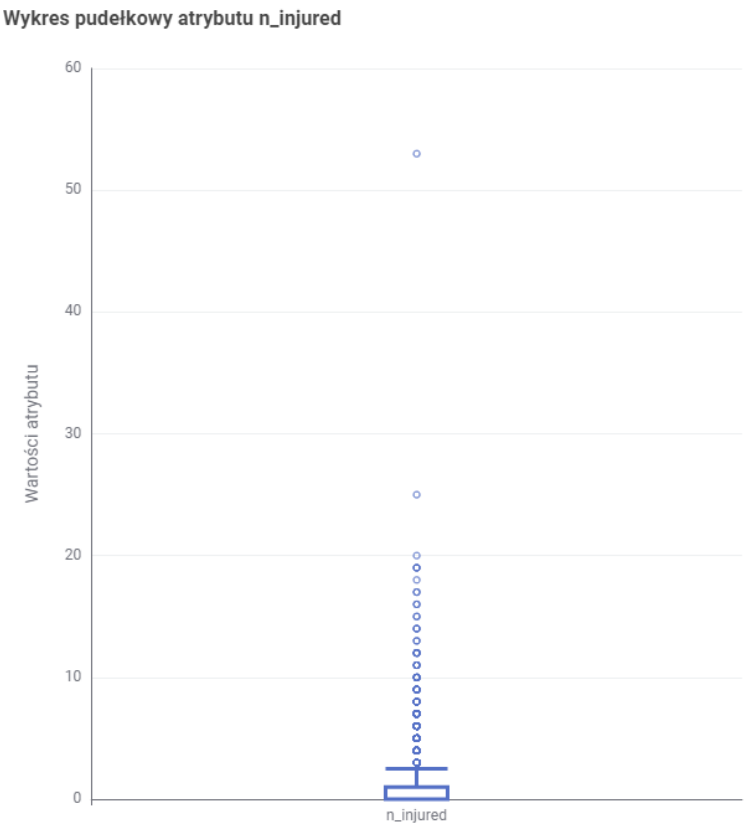
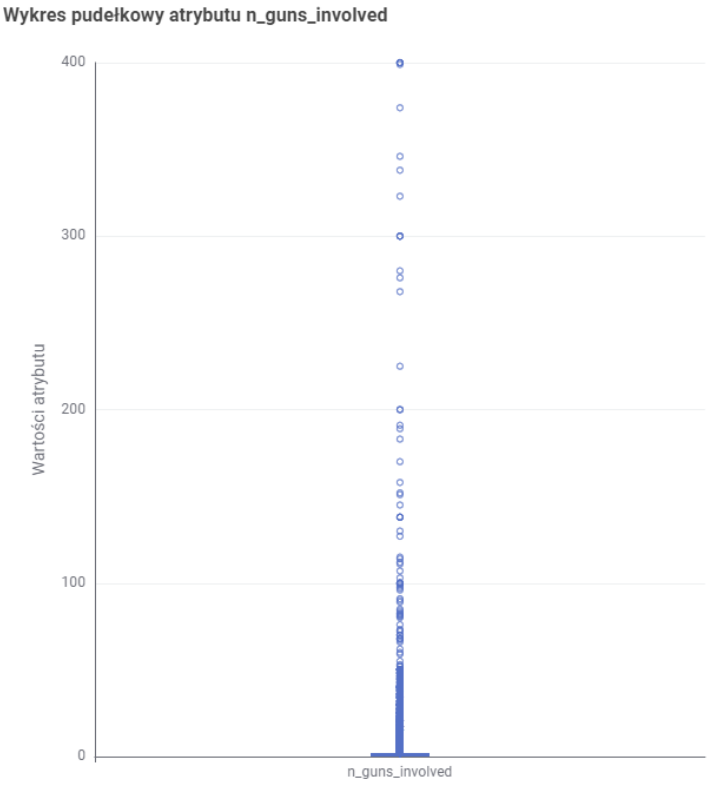
| | |
|---|---|
| | **Histogram atrybutu n_killed**<br>  |
| n_injured | **Histogram atrybutu n_injured**<br>  |

| | |
|---|---|
| n_guns_involved |  |
| longitude, latitude |  |
| state |  |

| | |
|---|---|
| gun_type |  |
| participant_age |  |
| date<br>*(year)* |  |
| date<br>*(month)* |  |

| | |
|---|---|
| date<br>*(day)* |  |

| Atrybut | Wykres pudełkowy |
|---|---|
| n_killed |  |

| | |
|---|---|
| n_injured |  |
| n_guns_involved |  |

## Wnioski z analizy rozkładów:

### Atrybut n_killed

- **Silnie prawoskośny rozkład** z medianą bliską 0
- **Dominacja incydentów bez ofiar śmiertelnych,** większość przypadków nie kończy się śmiercią
- **Długi prawy ogon** wskazujący na rzadkie, ale ekstremalne przypadki masowych strzelanin
- **Implikacje -** Nieliczne, ale bardzo poważne incydenty znacząco wpływają na statystyki śmiertelności

### Atrybut n_injured

- **Podobny wzorzec prawoskośności** jak w przypadku ofiar śmiertelnych
- **Koncentracja w niskich wartościach,** większość incydentów powoduje niewiele obrażeń
- **Obecność wartości ekstremalnych** sugeruje incydenty w miejscach o dużym zagęszczeniu ludzi
- **Korelacja z lokalizacją -** Przypadki z większą liczbą rannych prawdopodobnie występują w przestrzeniach publicznych

### Atrybut n_guns_involved

- **Zdecydowana dominacja pojedynczej broni** w incydentach
- **Rzadkość przypadków wielobronnych** może wskazywać na bardziej zorganizowane działania przestępcze
- **Prostota większości incydentów** - typowy przypadek to jedna osoba z jedną bronią

### Atrybuty longitude, latitude

- **Wschodnie wybrzeże**: Najwyższa gęstość incydentów, szczególnie w stanach gęsto zaludnionych
- **Kalifornia**: Druga po wybrzeżu wschodnim koncentracja przypadków
- **Środkowy zachód i stany górskie**: Znacznie niższa częstotliwość incydentów
- **Alaska**: Praktycznie brak zarejestrowanych przypadków w zbiorze

### Atrybut state

- **Duże zróżnicowanie między stanami**: Wyraźne różnice w liczbie incydentów między poszczególnymi stanami
- **Liderzy negatywni**: Stany takie jak Illinois, California, Florida, Texas, Floryda i Ohio dominują w statystykach
- **Stany o niskim ryzyku**: Część stanów wykazuje znacznie mniejszą liczbę incydentów, np. Hawaii, Vermont, Wyoming i South Dakota.
- **Potencjał predykcyjny**: Stan może być silnym predyktorem zarówno częstotliwości jak i śmiertelności incydentów

### Atrybut gun_type

- **Wybrakowane dane:** Kategoria “Unknown” stanowi problem, jest najliczniejszą kategorią w zbiorze (~130,000), wskazuje na braki w danych

- **Dominuje broń krótka:** Handgun jest drugą najczęstszą kategorią (~25,000 przypadków)
- **Pozostałe typy broni są marginalne:** Każdy poniżej 10,000 przypadków

### Atrybut participant_age

- **Rozkład prawoskośny z dominacją młodych dorosłych**: Wyraźny szczyt w przedziale 20-30 lat (~50,000 przypadków)
- **Koncentracja w grupie 18-40 lat**: Około 70% wszystkich incydentów dotyczy tej grupy wiekowej
- **Stopniowy spadek z wiekiem**: Liczba incydentów systematycznie maleje po 30. roku życia
- **Niska reprezentacja skrajnych grup**: Bardzo niewiele przypadków poniżej 18 lat i powyżej 60 lat
- **Długi prawy ogon**: Pojedyncze przypadki w bardzo wysokich grupach wiekowych (80+ lat)

### Atrybut date - *year*

- **Wyraźny trend wzrostowy**: Wzrost z ~51,500 do ponad 61,000 incydentów (+18% w 4 lata)
- **Największy skok**: Między 2016 a 2017 rokiem
- **Niepokojąca tendencja**: Systematyczny wzrost problemu przemocy z bronią

### Atrybut date - *month*

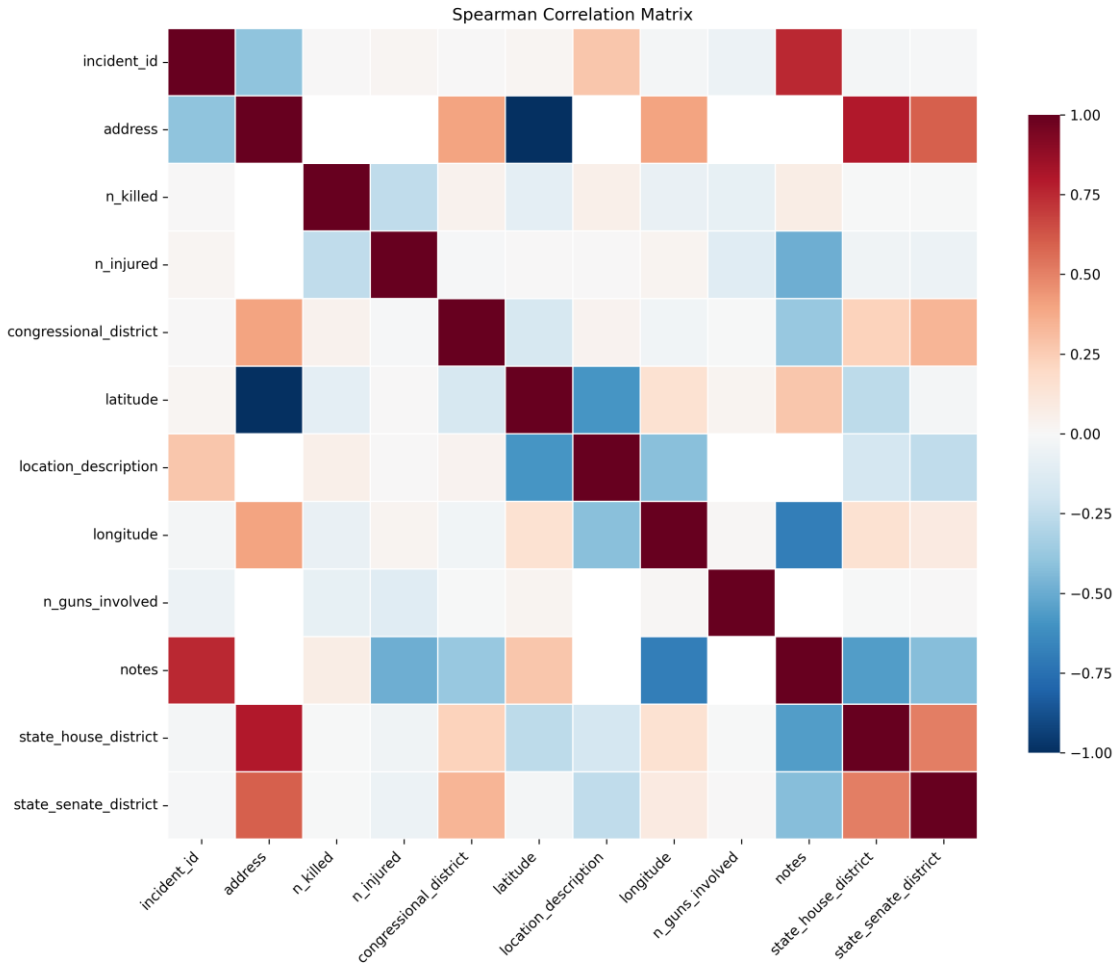
- **Wyraźna cykliczność sezonowa**:
  - **Minimum**: Luty (~14,500 incydentów)
  - **Maksimum**: Lipiec i sierpień (powyżej 21,000 każdy)
- **Miesiące letnie (maj-sierpień)**: Konsekwentnie wyższe wartości
- **Miesiące zimowe (grudzień-marzec)**: Najniższe liczby incydentów
- **Różnica sezonowa**: ~46% więcej incydentów w szczycie lata niż zimą

### Atrybut date - *day*

- **Względna równomierność** w ciągu tygodnia
- **Lekka tendencja weekendowa**: Sobota i niedziela z najwyższymi wartościami (~34,500-35,000)
- **Minimum w piątek**: ~30,500 incydentów
- **Różnica tygodniowa**: ~15% więcej incydentów w weekendy

# Macierze korelacji

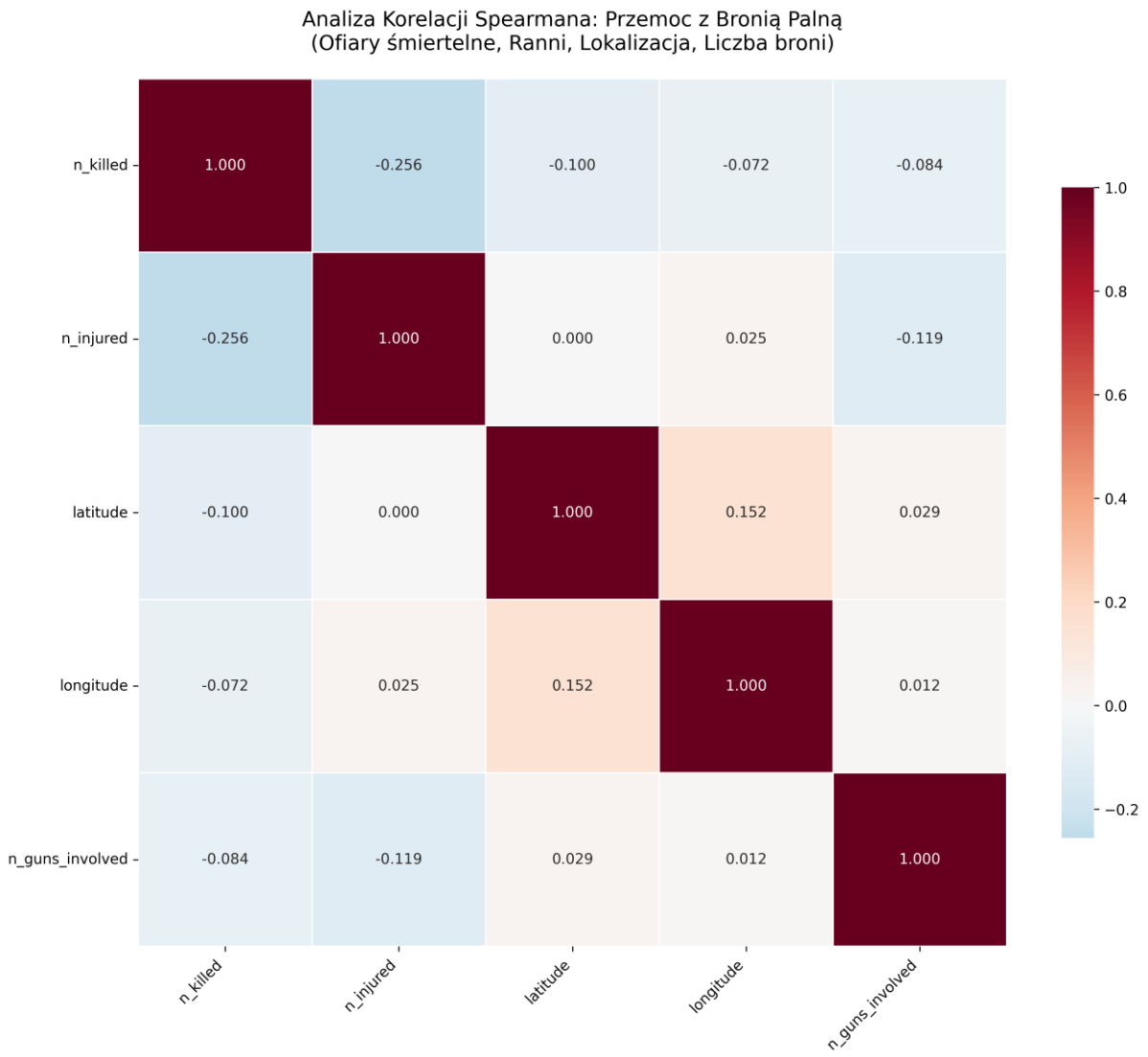
## Macierz korelacji bez żadnych dodatkowych warunków



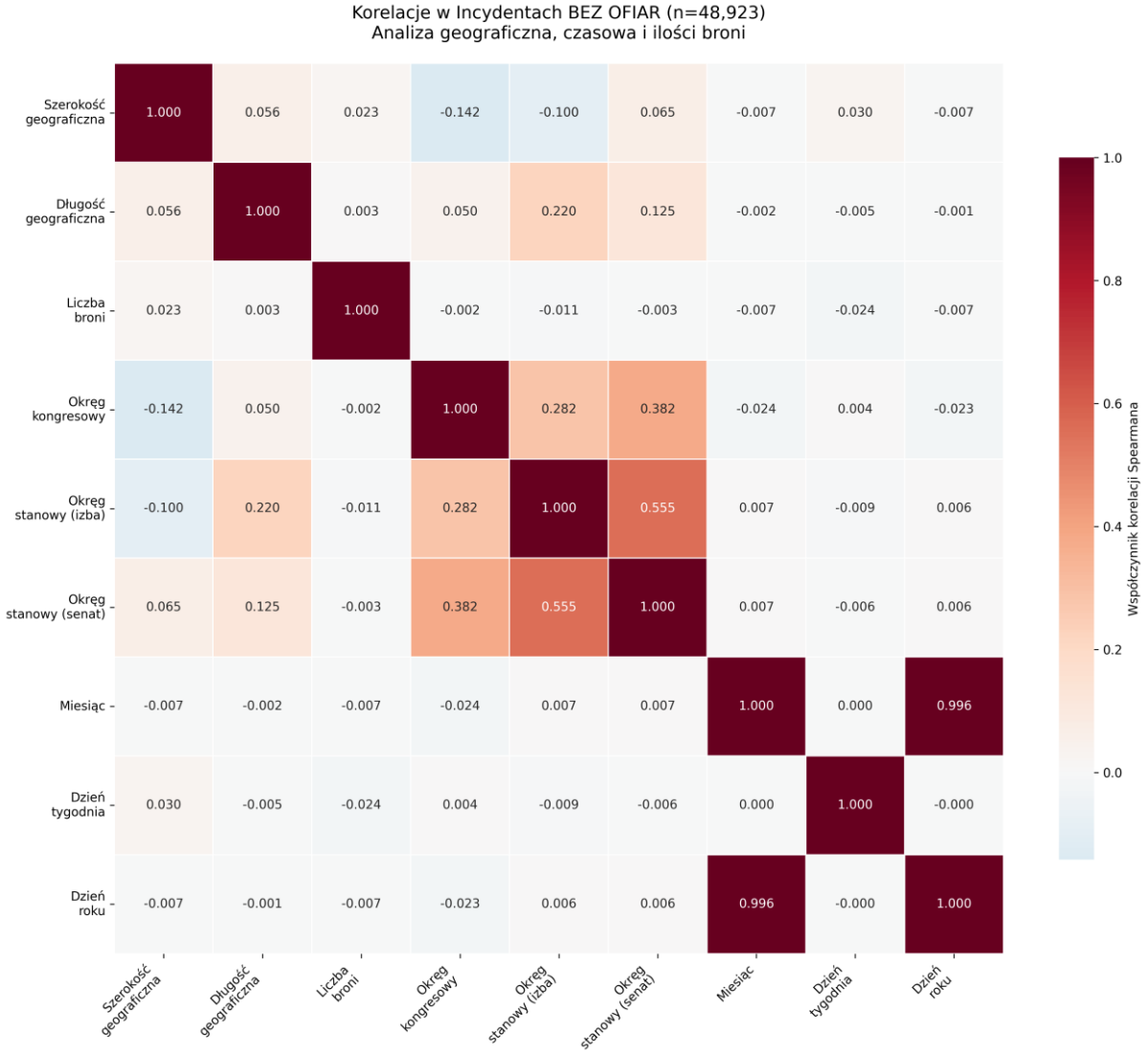

Wnioski:

- „Zabici” vs „Ranni” – Spearman ρ jest lekko ujemny (~–0.25), co potwierdza poprzednie obserwacje – w incydentach o bardzo dużej liczbie zabitych proporcja rannych spada.
- Liczba broni (“n_guns_involved”) – Korelacje z ofiarami ($<|0.05|$) są praktycznie zerowe – więcej sztuk broni nie idzie w parze ani z większą śmiertelnością, ani z liczbą rannych.
- Zmienne tekstowe/identyfikatory – Bardzo wysokie ρ między „incident_id”, „address”, „notes” czy „location_description” a różnymi kolumnami to artefakty numerycznego kodowania tekstu – te pola należy pominąć lub poprawnie zakodować przed dalszą analizą.
- Geografia – Latitude vs longitude prawie nie korelują z ofiarami ($|\rho|<0.1$). Nie ma monotonicznego trendu „gdzie więcej ofiar” – przestrzeń pewnie będzie lepiej badać przez mapy/klastry niż proste korelacje.

# Macierz korelacji ograniczona tylko do najbardziej przydatnych kolumn



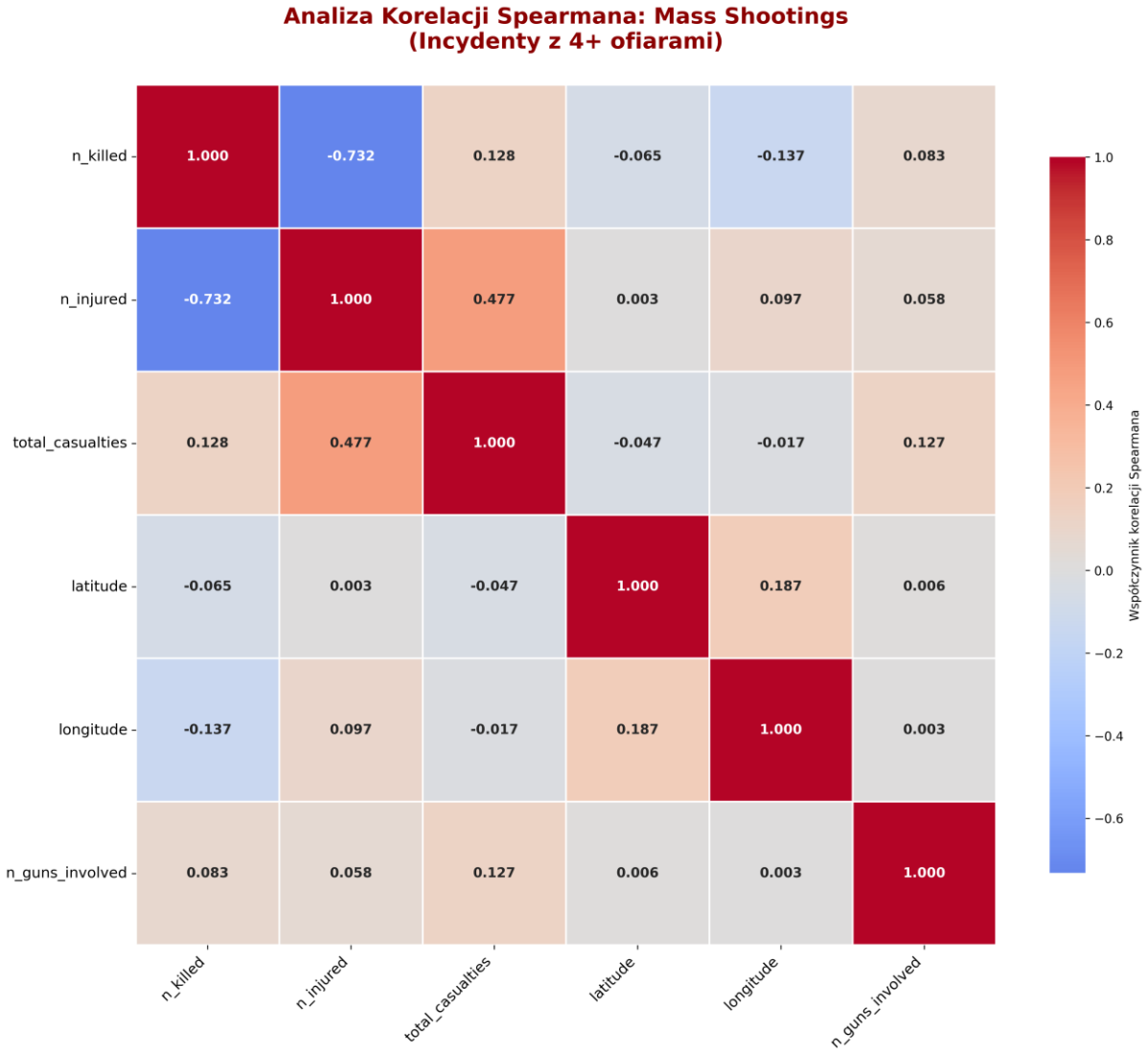
## Macierz korelacji dla incydentów bez ofiar




Wnioski:

- **Numery okręgów politycznych**
  • Izba ⇄ Senat (ρ≈0,56) – silna dodatnia: wspólna numeracja w stanach.
  • Kongres ⇄ Izba (ρ≈0,28) i Kongres ⇄ Senat (ρ≈0,38) – umiarkowane, bo okręgi się częściowo pokrywają.
- **Czas**
  • Miesiąc ⇄ Dzień roku (ρ≈0,996) – niemal idealna korelacja, ale nie daje żadnej przydatnej informacji; to powinno być traktowane jako cecha cykliczna.
  • Dzień tygodnia ≈ 0 ze wszystkimi – brak monotonicznych zależności.
- **Liczba broni (“n_guns_involved”)**
  • |ρ|<0,03 ze wszystkimi – praktycznie zerowa monotoniczna relacja.

## Macierz korelacji dla incydentów z 4+ ofiarami

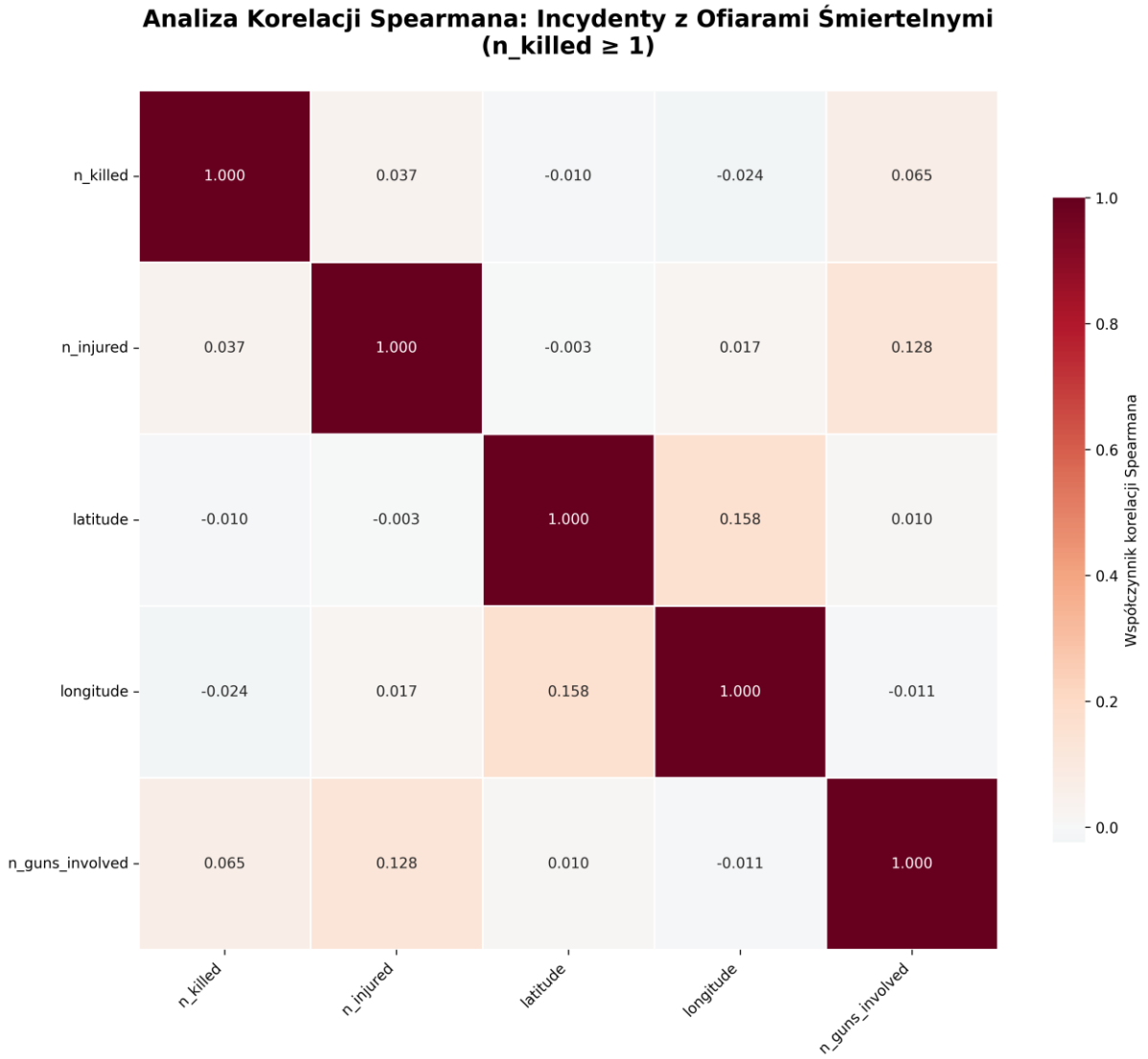



Wnioski:

- **n_killed vs n_injured (ρ≈–0.732)**
  Bardzo silna ujemna korelacja - im więcej ofiar śmiertelnych, tym proporcjonalnie mniej rannych (i odwrotnie). W masowych strzelaninach często jest albo wielu rannych, albo wielu zabitych.
- **n_injured vs total_casualties (ρ≈0.477)**
  Umiarkowana dodatnia - liczba rannych jest istotnym składnikiem całkowitych ofiar. Dla zabitych ↔ total_casualties korelacja jest dużo słabsza (ρ≈0.128), bo przy bardzo wysokich liczbach zabitych rangi „total” mogą być wyrównane przez różnice w rannych.
- **n_guns_involved ↔ ofiary (ρ<0.13)**
  Bardzo słabe, dodatnie związki z zabitymi (0.083), rannymi (0.058) i łącznymi ofiarami (0.127). Widać, że sama liczba użytych sztuk broni nie determinuje monotonicznie skali tragedii.

- **Geografia ↔ ofiary (|ρ|<0.14)**
  Latitude i longitude niemal nie korelują z liczbą zabitych/rannych – nie ma prostego monotonicznego wzoru „gdzie więcej ofiar”.
- **Latitude vs Longitude (ρ≈0.187)**
  Słaba dodatnia relacja wewnątrz tego zbioru, wynikająca z rozmieszczenia masowych strzelanin w konkretnych regionach, ale i ona jest niewielka.

## Macierz korelacji dla incydentów z przynajmniej 1 ofiarą śmiertelną




Wnioski:

- **Zabici ↔ Ranni (ρ≈0.04)**
  Niezwykle słaba, dodatnia korelacja – w incydentach, gdzie ktoś umarł, liczba rannych nie rośnie już kosztem zabitych (był silnie ujemny efekt przy uwzględnieniu zer), teraz jest praktycznie niezależna.

- **Broń ↔ ofiary (ρ≈0.07 z zabitymi, ρ≈0.13 z rannymi)**
  Również słabe, dodatnie związki: więcej sztuk broni daje minimalnie wyższą liczbę rannych, ale efekt jest bardzo nikły.
- **Geografia ↔ ofiary (|ρ| <0.03)**
  Latitude i longitude niemal nie korelują z żadnym z rodzajów ofiar – w fatalnych strzelaninach lokalizacja nie determinowała monotonicznie, czy będzie więcej zabitych czy rannych.
- **Latitude ↔ Longitude (ρ≈0.16)**
  Słaba dodatnia korelacja wewnątrz tego podzbioru – wskazuje, że masowe strzelaniny z ofiarami mają tendencję do występowania w pewnym pasie geograficznym (np. bardziej na wschód, im dalej na północ).

# Uwagi na temat jakości danych

## Dane brakujące

Mimo bogatego zakresu informacji, w zbiorze brakuje niektórych potencjalnie istotnych zmiennych. Na przykład nie jest rejestrowana godzina zdarzenia – dostępna jest tylko data, co uniemożliwia analizę wzorców godzinowych (pora dnia). Zbiór nie zawiera też bezpośrednich danych o kontekście zdarzenia (np. motyw, okoliczności zajścia) ani szczegółowych danych demograficznych ofiar/sprawców (poza wiekiem i płcią). Jeśli eksploracja wymaga analizy czynników zewnętrznych (np. gęstości zaludnienia, lokalnych przepisów dot. broni), konieczne będzie połączenie danych z dodatkowymi źródłami zewnętrznymi.

## Dane niespójne

Ogólnie dane są spójne wewnętrznie – każdy incydent ma unikalny identyfikator (incident_id), więc nie występują duplikaty, a wartości w kluczowych polach zachowują logiczne zakresy (np. brak ujemnej liczby ofiar). Występują jednak drobne niespójności w formacie niektórych pól. Na przykład atrybut city_or_county bywa niejednoznaczny – przechowuje nazwę miasta lub hrabstwa w jednym polu, co może utrudniać automatyczne grupowanie danych po lokalizacji. Dodatkowo brakujące informacje w polach słownikowych są oznaczane wartością "Unknown" zamiast pustej (np. status broni gun_stolen przyjmuje tylko wartości "Stolen" lub "Unknown"), co należy uwzględnić podczas czyszczenia danych.

## Dane niezrozumiałe

Niektóre elementy danych są trudne do zrozumienia bez uprzedniej obróbki ze względu na sposób ich zapisu. W szczególności pola zawierające listy lub słowniki (dotyczące broni i uczestników) są zakodowane jako tekst z separatorami || oraz ::, przez co pojedynczy rekord może zawierać wiele wartości w jednej komórce. Bez odpowiedniego przetworzenia (np. rozdzielenia takich pól na osobne wiersze lub kolumny) ich zawartość jest mało czytelna i podatna na błędną interpretację. Również

niektóre opisy tekstowe (np. w polu incident_characteristics) mogą być niejasne i wymagają znajomości kontekstu, aby je właściwie zinterpretować.

### Punkty oddalone (outliers)

Analiza rozkładu ujawnia obecność wartości odstających, zwłaszcza w liczbie ofiar w niektórych incydentach. Większość zdarzeń to pojedyncze ofiary lub ich brak, jednak występują skrajne przypadki o bardzo dużej liczbie poszkodowanych – przykładem jest atak w klubie nocnym w Orlando w 2016 roku, gdzie zginęło 49 osób, a 53 zostały ranne. Największy masowy incydent z tego okresu (strzelanina w Las Vegas, 2017) nie został ujęty w zbiorze z powodu problemów z ekstrakcją danych, dlatego to Orlando pozostaje najbardziej odstającym punktem w danych. Obecność takich ekstremów może silnie wpływać na statystyki (np. zawyżenie średniej liczby ofiar), więc w analizie warto rozważyć ich odrębne traktowanie lub użycie miar odpornych na wartości skrajne.

## Czy da się wykorzystać te dane do zrealizowania celu eksploracji?

Zbiór danych można wykorzystać do realizacji celu eksploracji, pod warunkiem odpowiedniego przygotowania. Pomimo pewnych braków i wad, dane obejmują wystarczająco duży zakres przypadków oraz cech, aby wyciągać istotne wnioski – ubytki informacji dotyczą głównie mniej istotnych pól i nie powinny znacząco zniekształcić wyników przy tak dużej liczbie obserwacji. Kluczowe jest jednak przeprowadzenie wstępnego czyszczenia danych: uzupełnienie lub oznaczenie brakujących wartości, rozdzielenie złożonych pól na prostsze elementy oraz uwzględnienie wpływu punktów odstających na analizy. Po tych zabiegach jakość i szczegółowość danych w pełni umożliwiają przeprowadzenie rzetelnej eksploracji oraz pozwalają na formułowanie wiarygodnych wniosków na temat zjawiska przemocy z użyciem broni.

## Podsumowanie

Ogólnie rzecz biorąc, jakość danych **Gun Violence Data** można ocenić jako dobrą, a zidentyfikowane problemy dają się rozwiązać na etapie przygotowania danych. Zbiór jest obszerny, reprezentatywny dla analizowanego zjawiska i zawiera wiele szczegółowych informacji, co stanowi solidną podstawę eksploracji. Wykryte mankamenty – brakujące lub niejednoznaczne dane oraz sporadyczne wartości odstające – wymagają uwagi, lecz nie uniemożliwiają efektywnego wykorzystania zbioru. Po wyczyszczeniu i ujednoliceniu danych, zestaw ten pozwoli na uzyskanie wiarygodnych rezultatów i odkrycie istotnych prawidłowości dotyczących przemocy z użyciem broni w USA.