

Sprawozdanie

Politechnika Wrocławska

Wydział Elektroniki

Przetwarzanie dużych zbiorów danych

Analiza danych dotyczących zabójstw w Stanach Zjednoczonych

Spis treści

1. Wybór danych do analizy	3
2. Przygotowanie danych do analizy	3
2.1. Oczyszczenie wybranych danych – Excel.....	3
2.2. Oczyszczenie wybranych danych – Power BI Desktop	3
2.3. Modyfikacja danych – Power BI Desktop	3
3. Analiza i wizualizacja danych w Power BI.....	6
3.1. Miejsce	6
3.2. Daty	7
3.3. Ofiara	8
3.4. Sprawca	9
3.5. Relacje	10
3.6. Narzędzie	11
4. Wnioski	12

1. Wybór danych do analizy

Dane potrzebne do zrealizowania projektu pochodzą ze strony:

"<https://www.kaggle.com/nevil7/homicide-data-identifying-the-serial-killers?select=database.csv>".

Pobrany plik .csv miał wielkość 106.63 MB oraz posiadał tabelę, która zawierała 24 kolumny oraz 638454 wiersze.

Dane zostały wybrane ze względu na szeroką gamę możliwości, jeśli chodzi o analizę. Wybrany zbiór danych posiada bowiem aż 24 kolumny, spośród których znajdują się informacje takie jak miejsce i data popełnienia przestępstwa, wiek, płeć oraz pochodzenie ofiary i sprawcy. Pozwala to na obszerną analizę oraz uzyskanie ciekawych wniosków.

2. Przygotowanie danych do analizy

2.1. Oczyszczenie wybranych danych – Excel

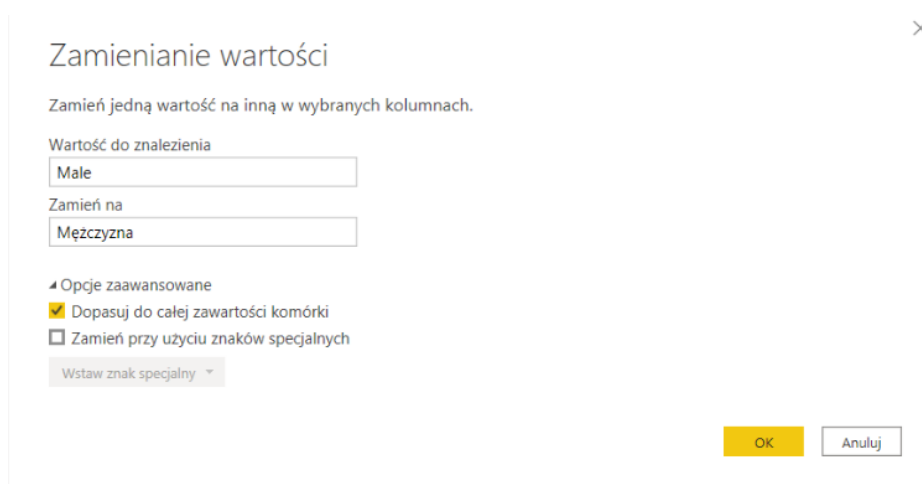
Pierwszym etapem oczyszczania danych było usunięcie wierszy, w których status rozwiązania przestępstwa, czyli „Crime solved” posiadał wartość „yes”, co oznaczało, że sprawa została rozwiązana i posiada wymagane do analizy dane. Pozwoliło to na zmniejszenie ilości wierszy z 638454 na 5434. W Excelu także zostały zmienione nazwy kolumn, przetłumaczono je z języka angielskiego na język polski oraz przetłumaczono także nazwy miesięcy, w których popełniono przestępstwo.

2.2. Oczyszczenie wybranych danych – Power BI Desktop

Z pomocą narzędzia Power BI Desktop zostały usunięte kolumny z Ilością ofiar oraz Ilością sprawców, ponieważ duża część tych kolumn posiadała błędne informacje lub nie posiadała ich w ogóle.

2.3. Modyfikacja danych – Power BI Desktop

Pierwszym krokiem w modyfikacji danych było przetłumaczenie części danych na język polski, żeby były one bardziej przejrzyste w odbiorze, zmieniono wartości w kolumnie „Płeć Ofiary” oraz „Płeć Sprawcy” odpowiednio na Kobieta lub Mężczyzna.



Rysunek 1 Zmiana wartości w kolumnie Płeć z wartości Male na Mężczyzna

×

Zamienianie wartości

Zamień jedną wartość na inną w wybranych kolumnach.

Wartość do znalezienia

Zamień na

> Opcje zaawansowane

OK Anuluj

Rysunek 2 Zmiana wartości w kolumnie Płeć z Female na Kobieta

Dalsza część tłumaczenia polegała na zmianie języka z angielskiego na polski w następujących kolumnach: „Rodzaj zbrodni” oraz „Narzędzie zbrodni”.

×

Zamienianie wartości

Zamień jedną wartość na inną w wybranych kolumnach.

Wartość do znalezienia

Zamień na

⚡ Opcje zaawansowane

☒ Dopasuj do całej zawartości komórki

☐ Zamień przy użyciu znaków specjalnych

Wstaw znak specjalny ▾

OK Anuluj

Rysunek 3 Zmiana wartości w kolumnie Rodzaj Śmierci z Manslaughter by Negligence na Pozbawienie Życia poprzez Zaniedbanie

×

Zamienianie wartości

Zamień jedną wartość na inną w wybranych kolumnach.

Wartość do znalezienia

Zamień na

⚡ Opcje zaawansowane

☒ Dopasuj do całej zawartości komórki

☐ Zamień przy użyciu znaków specjalnych

Wstaw znak specjalny ▾

OK Anuluj

Rysunek 4 Zmiana wartości w kolumnie Narzędzie Zbrodni z Blunt Object na Tępy Obiekt

Kolejnym krokiem było dodanie kolumny warunkowej, która miała na celu określić na podstawie podanej relacji pomiędzy sprawcą i ofiarą, czy osoby znały się przed popełnieniem przestępstwa. Utworzona kolumna nosi nazwę „Ofiara znała Sprawcę” i posiada odpowiednio wartość TRUE lub FALSE.

Dodawanie kolumny warunkowej

Dodaj kolumnę warunkową obliczaną na podstawie innych kolumn lub wartości.

Nazwa nowej kolumny

Ofiara znała Sprawcę

Nazwa kolumny	Operator	Wartość ①	Wartość wyjściowa ①
Jeśli Relacja	równa się	ABC 123 Stranger	To ABC 123 false

Dodaj klauzulę

W przeciwnym razie ①

ABC 123 true

OK

Anuluj

Rysunek 5 Proces tworzenia kolumny warunkowej o nazwie Ofiara znała Sprawcę

Po dodaniu kolumny „Ofiara znała Sprawcę” została także utworzona druga, bardziej skomplikowana kolumna warunkowa o nazwie „Ofiara i Sprawca byli biologicznie spokrewnieni”, która miała na celu określić pokrewieństwo między sprawcą i ofiarą.

Dodawanie kolumny warunkowej

Dodaj kolumnę warunkową obliczaną na podstawie innych kolumn lub wartości.

Nazwa nowej kolumny

Ofiara i Sprawca byli biologiczn

Nazwa kolumny	Operator	Wartość ①	Wartość wyjściowa ①
Jeśli Ofiara znała Spra...	równa się	ABC 123 FALSE	To ABC 123 false
Jeśli... Relacja	równa się	ABC 123 Brother	To ABC 123 true
Jeśli... Relacja	równa się	ABC 123 Daughter	To ABC 123 true
Jeśli... Relacja	równa się	ABC 123 Family	To ABC 123 true
Jeśli... Relacja	równa się	ABC 123 Father	To ABC 123 true
Jeśli... Relacja	równa się	ABC 123 Mother	To ABC 123 true

Dodaj klauzulę

W przeciwnym razie ①

ABC 123 false

OK

Anuluj

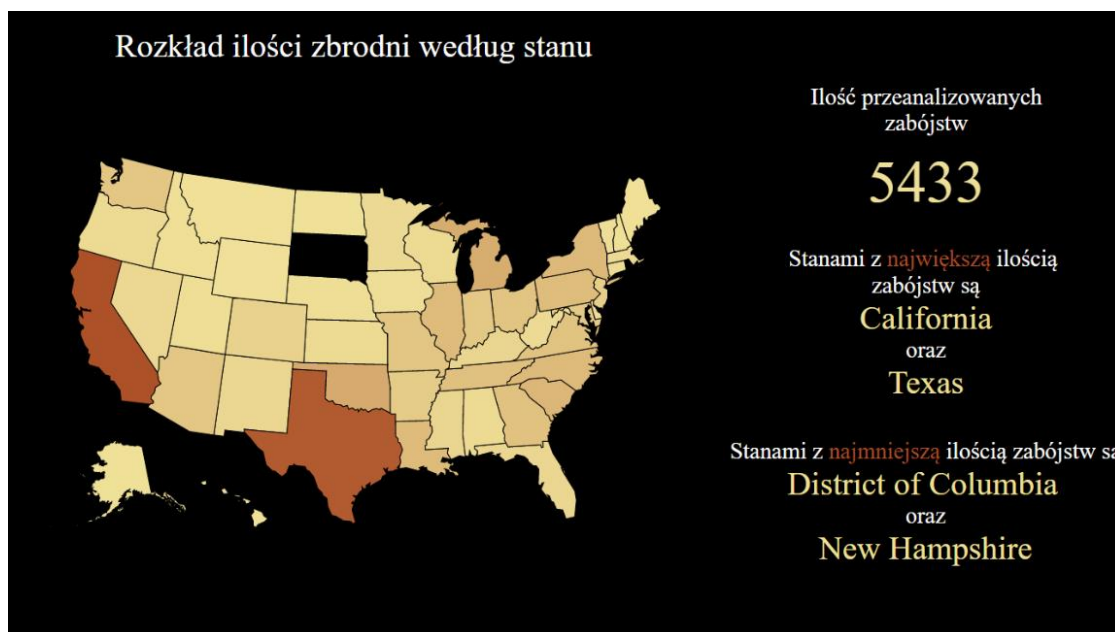
Rysunek 6 Proces tworzenia kolumny warunkowej o nazwie Ofiara i Sprawca byli biologicznie spokrewnieni

3. Analiza i wizualizacja danych w Power BI

Kiedy dane były już oczyszczone i przygotowane do analizy należało przejść do następnego etapu, czyli wizualizacji wybranych danych w narzędziu Power BI Desktop.

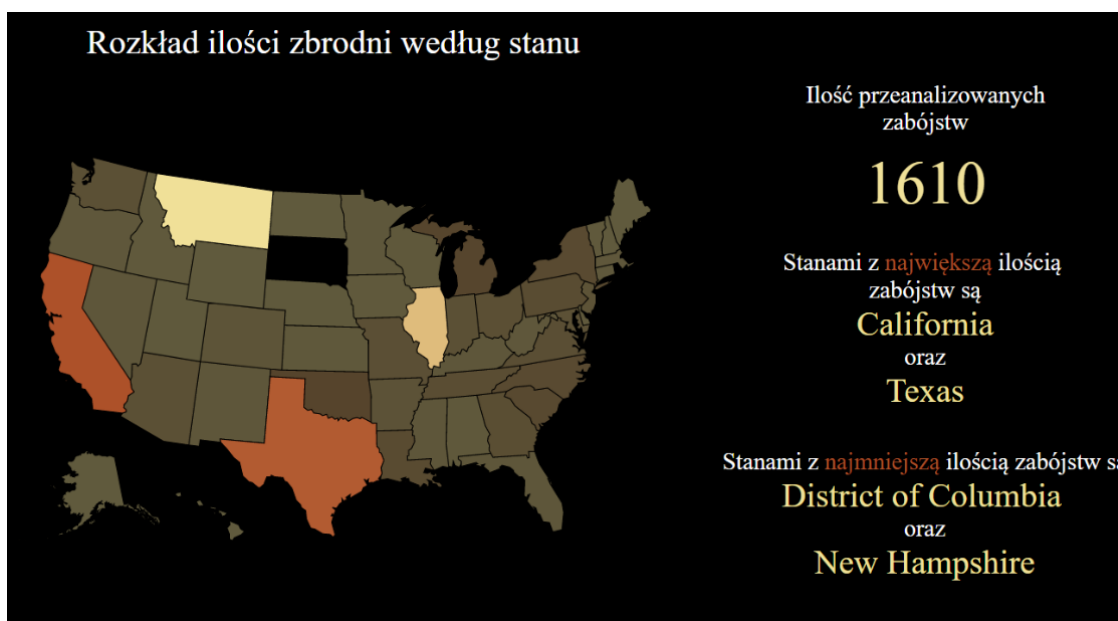
3.1. Miejsce

Pierwsza strona raportu zawiera informacje o rozkładzie ilości zbrodni według stanów znajdujących się w USA. Obok interaktywnej mapy znajdują się karta wyświetlająca ilość zbrodni, w zależności od zaznaczonych przez użytkownika stanów oraz wnioski wynikające z obserwacji.



Rysunek 7 Widok pierwszej strony raportu

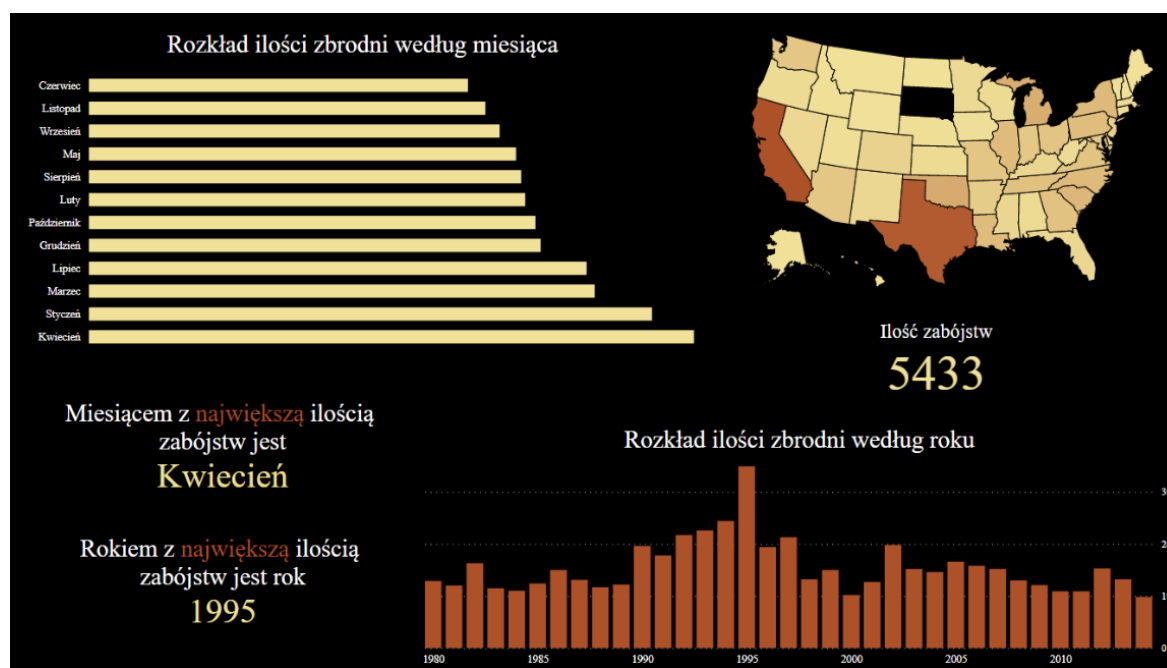
Interaktywna mapa umożliwia wyświetlenie ilości popełnionych zbrodni po kliknięciu w wybrany stan. Wraz z tym zostaje zmieniona ilość przeanalizowanych zabójstw znajdująca się przy mapie.



Rysunek 8 Widok strony raportu z wybranymi stanami Texas, California, Montana i Illinois

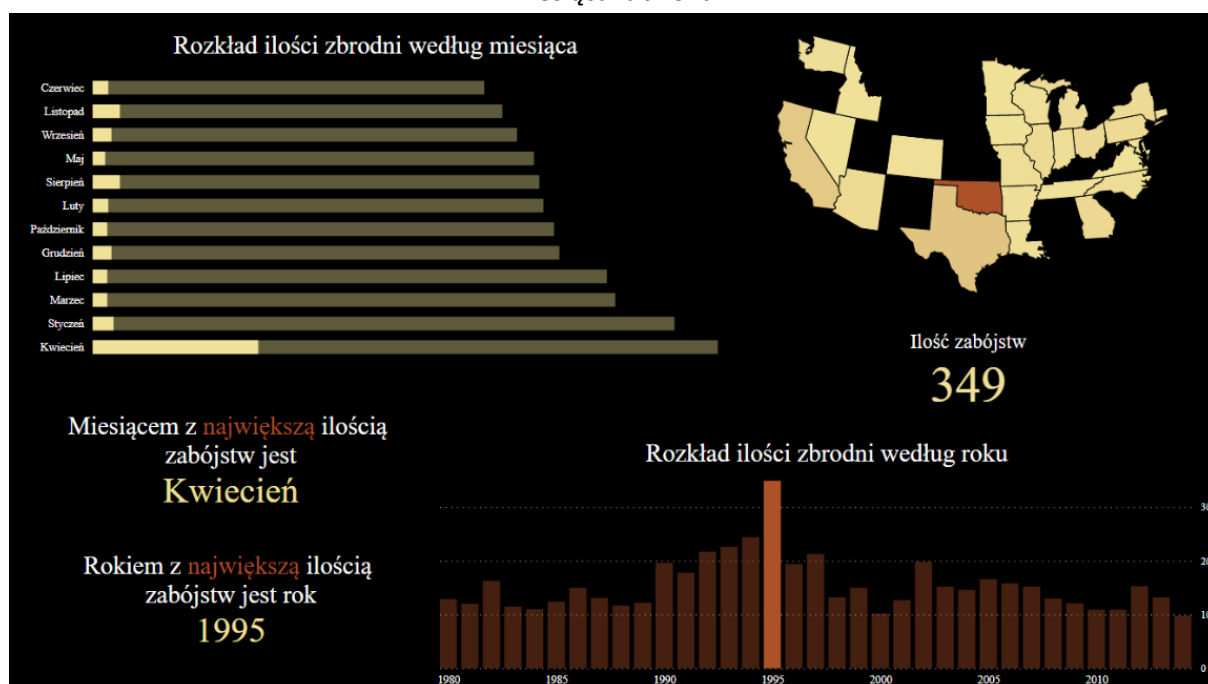
3.2. Daty

Druga strona raportu zawiera informacje na temat czasu, w którym zbrodnie zostały popełnione, czyli miesiąc i rok. Dane zostały przedstawione za pomocą dwóch interaktywnych wykresów słupkowego, przedstawiającego miesiąc oraz kolumnowego, przedstawiającego rok. Podobnie jak na stronie z miejscem umieszczony został licznik zbrodni. W raporcie pojawiła się także mapa, która pozwala sprawdzić požądane dane dla konkretnego stanu.



Rysunek 9 Widok drugiej strony raportu

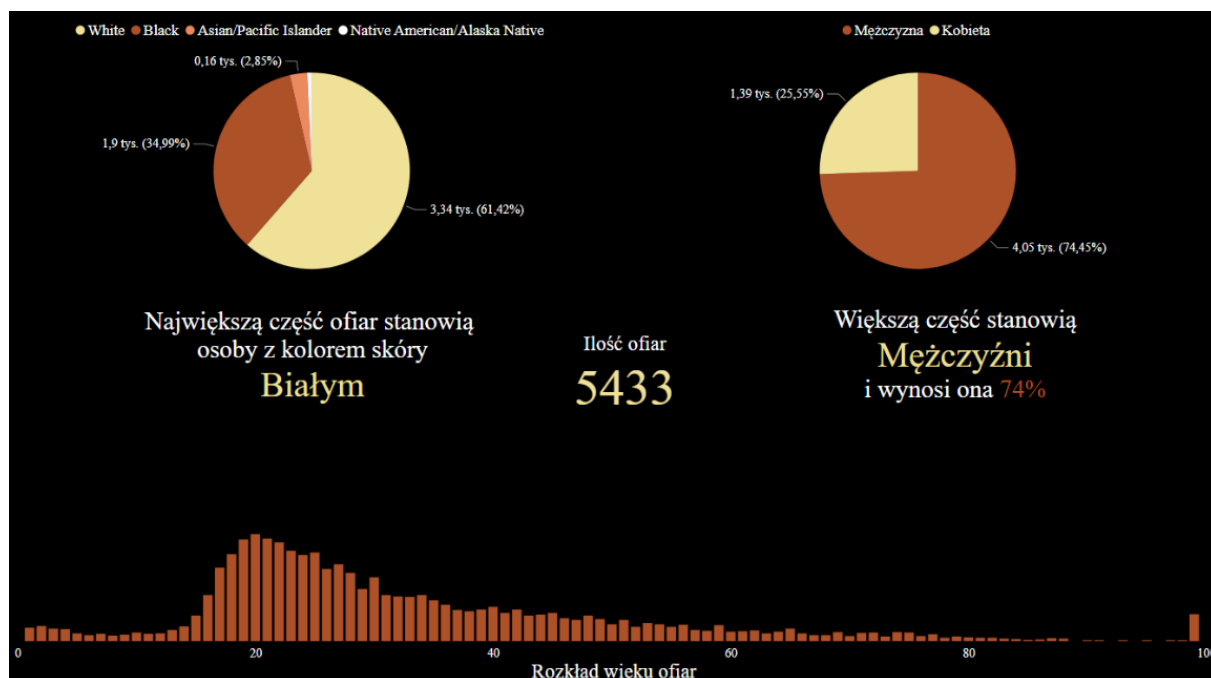
Dzięki klikanym wykresom istnieje możliwość zaznaczenia wybranego interesującego użytkownika miesiąca lub roku.



Rysunek 10 Widok strony raportu z wybranym rokiem z największą ilością zbrodni 1995

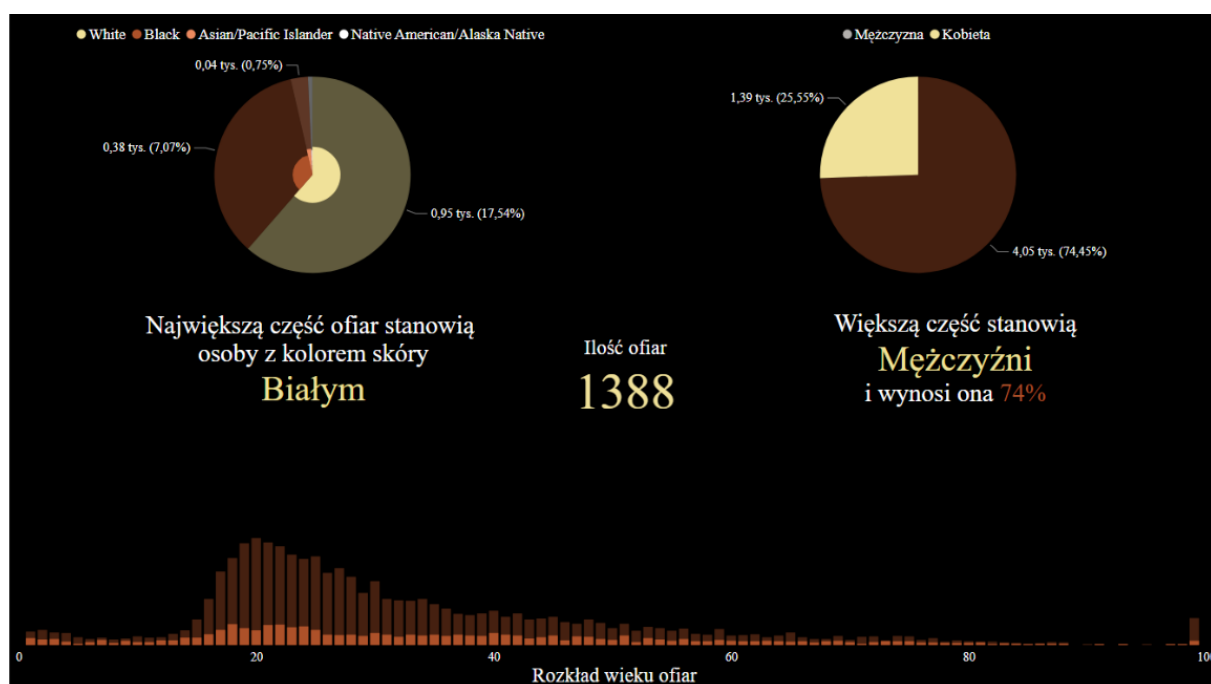
3.3. Ofiara

Następna strona raportu, czyli trzecia zawiera analizę danych na temat ofiar przestępstw. Znajdują się tam dwa wykresy kołowe. Pierwszy wykres kołowy zawiera informacje o kolorze skóry ofiary, a drugi wskazuje na płeć.



Rysunek 11 Widok trzeciej strony raportu

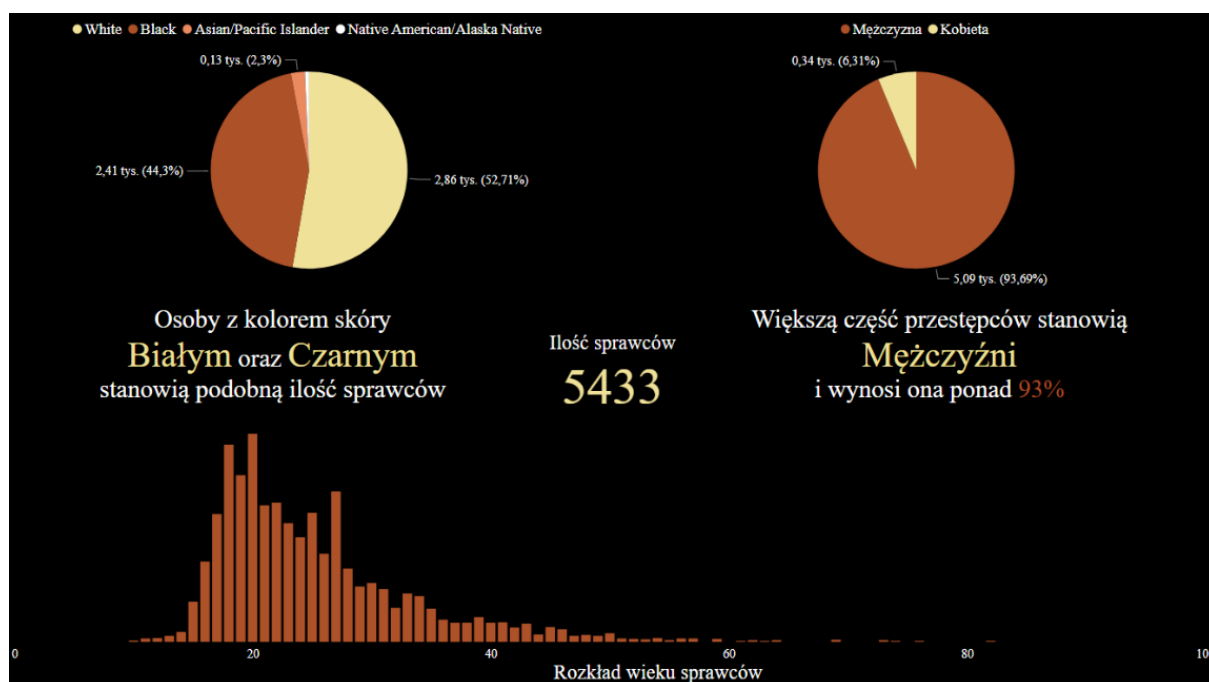
Dzięki dodaniu etykiet do wykresów są one przejrzyste w odbiorze. Podobnie jak na poprzednich stronach umieszczone zostały obserwacje wynikające z analizy wykresów oraz karta zliczająca ilość ofiar. Na dole strony raportu umieszczony został wykres kolumnowy zawierający rozkład wieku ofiar.



Rysunek 12 Widok strony raportu po wybraniu płci ofiary

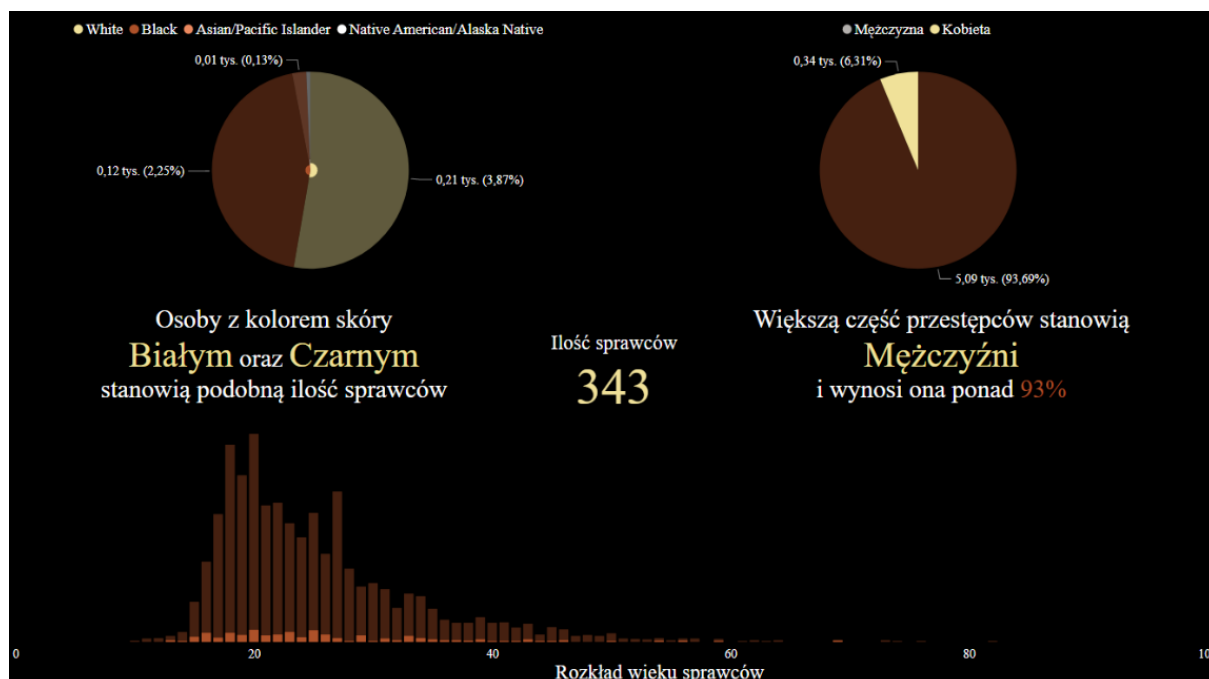
3.4. Sprawca

Czwarta strona raportu została poświęcona przedstawieniu danych na temat sprawców zbrodni. Tak samo jak w przypadku ofiar analizie poddane zostały kolor skóry, płeć oraz wiek zabójców.



Rysunek 13 Widok czwartej strony raportu

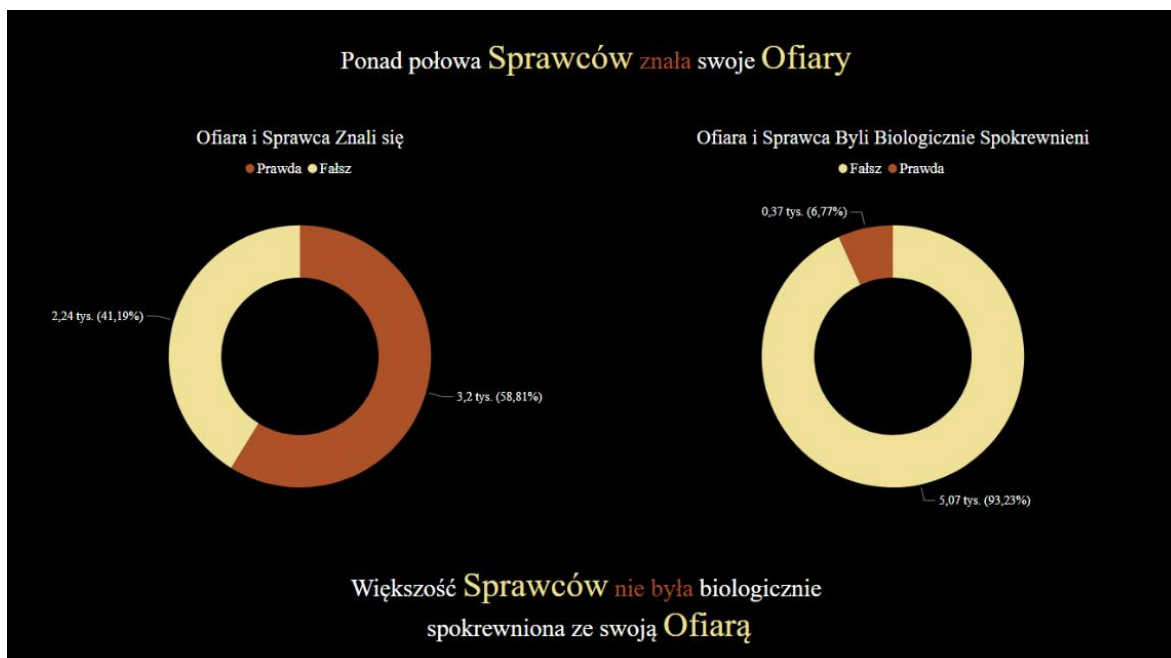
Interaktywne wykresy pozwalają na szybkie wyciągnięcie potrzebnych użytkownikowi informacji, przykładowo tak jak pokazano na Rysunku 13, gdzie dzięki zaznaczeniu konkretnej płci możliwe było poznanie szczegółów o zbrodniach popełnionych przez kobiety.



Rysunek 14 Widok strony raportu po wybraniu płci sprawcy

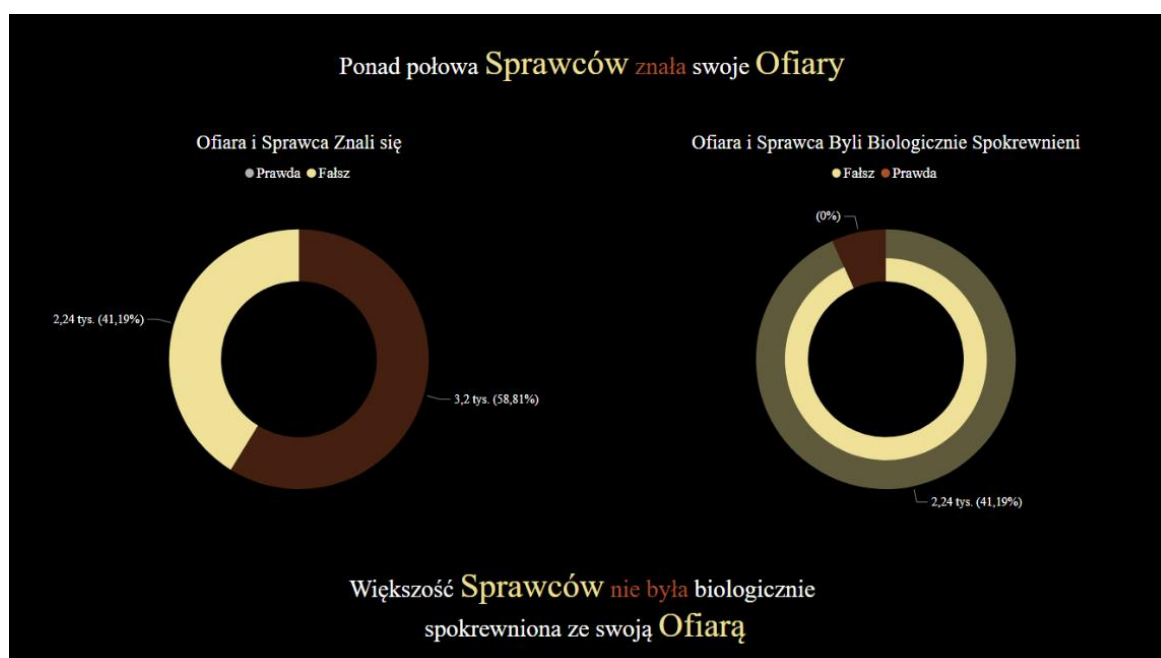
3.5. Relacje

Następna strona raportu przedstawia relacje między sprawcą a ofiarą. Dane te zostały uzyskane poprzez stworzenie tabel warunkowych. Ze względu na dużą liczbę rodzajów relacji wykresy słupkowy został pominięty. Na stronie raportu znajdują się dwa wykresy opisujące znajomość oraz spokrewnienie.



Rysunek 15 Widok piątej strony raportu

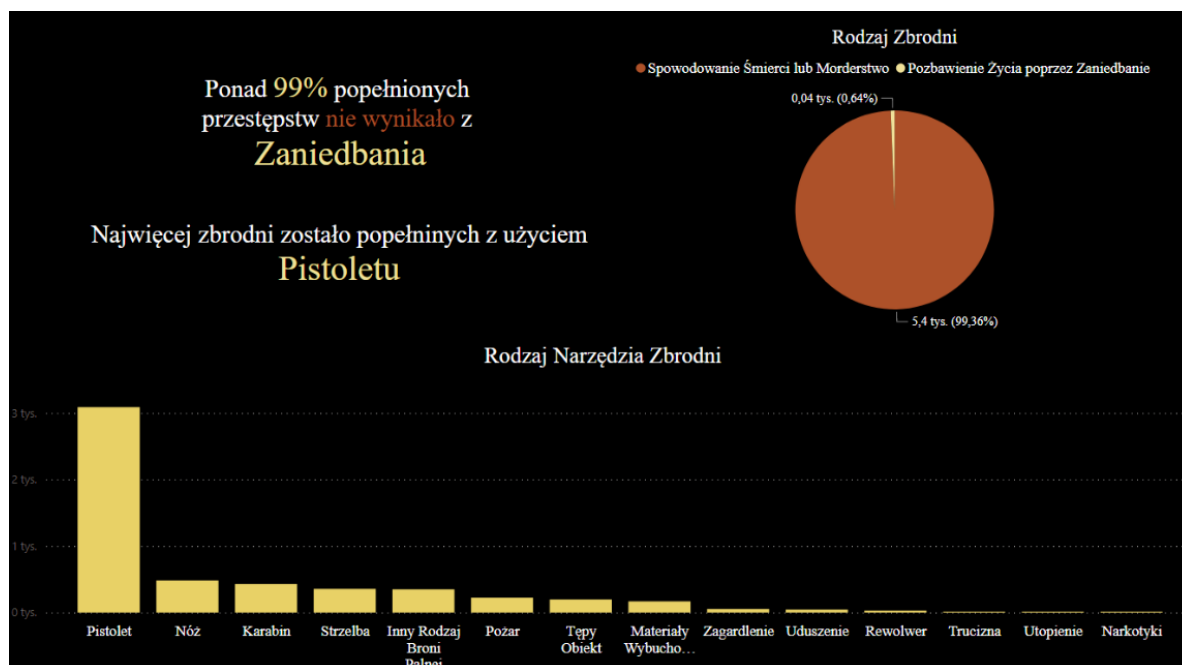
Dzięki dwóm wykresom pierścieniowym istnieje możliwość uzyskania dokładnych danych jaki procent osób, które się znali był ze sobą także spokrewniony biologicznie. Jak można zauważyć na Rysunku 16 41% ofiar było spokrewnionych ze swoim zabójcą.



Rysunek 16 Rysunek 14 Widok strony raportu, gdzie ofiara i zabójca się znali

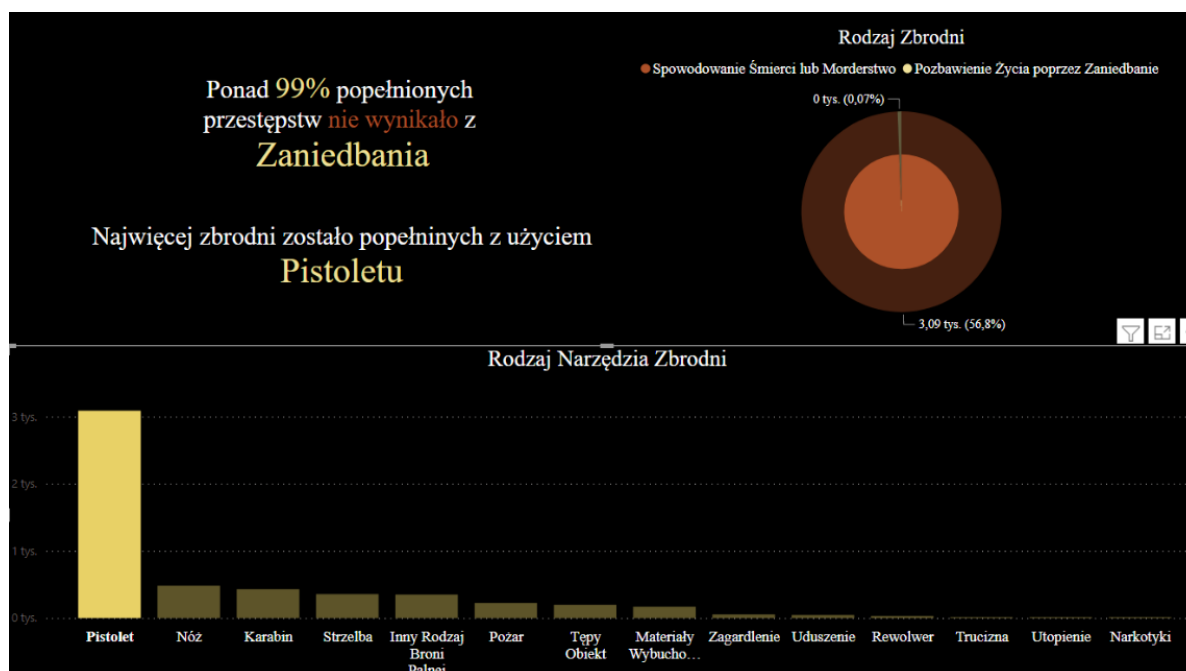
3.6. Narzędzie

Szóstą i zarazem ostatnią stroną raportu poświęcono narzędziu zbrodni. Umieszczony został wykres kołowy zawierający informacje o rodzaju popełnionej zbrodni oraz wykres słupkowy, przedstawiający ilość występowania narzędzi zbrodni.



Rysunek 17 Widok szóstej strony raportu

Zaskakującym wynikiem była ilość popełnionych zbrodni z użyciem jednej broni, mianowicie pistoletu, która znacząco odbiegała od reszty.



Rysunek 18 Rysunek 16 Widok strony raportu z wybranym pistoletem jako narzędziem zbrodni

4. Wnioski

W ramach projektu „Analiza danych dotyczących zabójstw w Stanach Zjednoczonych z użyciem Power BI Desktop” został przygotowany raport oraz sprawozdanie, zawierające opis poszczególnych czynności wykonywanych w ramach projektu.

Proces tworzenia projektu przebiegł bez większych trudności. Wbrew początkowym założeniom trudniejszą częścią projektu było znalezienie odpowiedniego zbioru danych, natomiast oczyszczenie danych oraz praca w narzędziu Power BI Desktop okazała się nie sprawiać problemów.

Narzędzie Power BI Desktop jest proste w obsłudze i intuicyjnym narzędziem. Wykresy utworzone w programie są czytelne i do tego interaktywne, w łatwy sposób możemy zaznaczyć interesujący nas fragment wykresu i szczegółowo przyjrzeć się interesującej nas części. Opcja, która jest szczególnie przydatna to wbudowane już w program mapy i możliwość przedstawiania na nich danych.

Projekt pozwolił na zdobycie wielu nowych umiejętności, w tym obsługi programu Power BI oraz poszerzył wiedzę na temat przetwarzania danych.