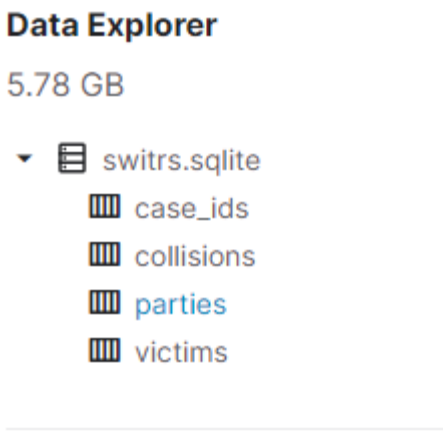
dHurtownie danych – Projekt 2021

Student	----------	Ocena
Indeks	251526	
Imię	Volodymyr	
Nazwisko	Zakhovaiko	

Uzasadnienie wyboru tematu projektu

Projekt	Wypadki drogowe w Kalifornii	Zabójstwa w USA
Plik	Switrs.sqlite	Databae.csv
Rozmiar pliku	5.78 GB	106.63 MB
Liczba kolumn	118	24
Liczba kolumn z wartością NULL	36	0
Zakres czasowy	2001-2020	1980-2014
Fakty	Wypadek drogowy	Zabójstwo
Wymiary	Liczba rannych, Liczba ofiar,	Liczba zabójców, Liczba ofiar
Kontekst	Lokalizacja, warunki drogowe, czas, rodzaj kolizji	Lokalizacja, Data, Agencja, rodzaj zabójstwa

Temat 1:
Plik bazy danych jest w formacie .sqlite i zawiera kilka milinów rekordów. Niestety nie udało mi się przejrzeć tabele w Tableu Prep, bo nie mam dość wolnego miejsca i potężności laptopa. Natomiast, dzięki stronie, z której pobrałem tą bazę możemy przeanalizować dane.



Tabeli, które są w bazie.

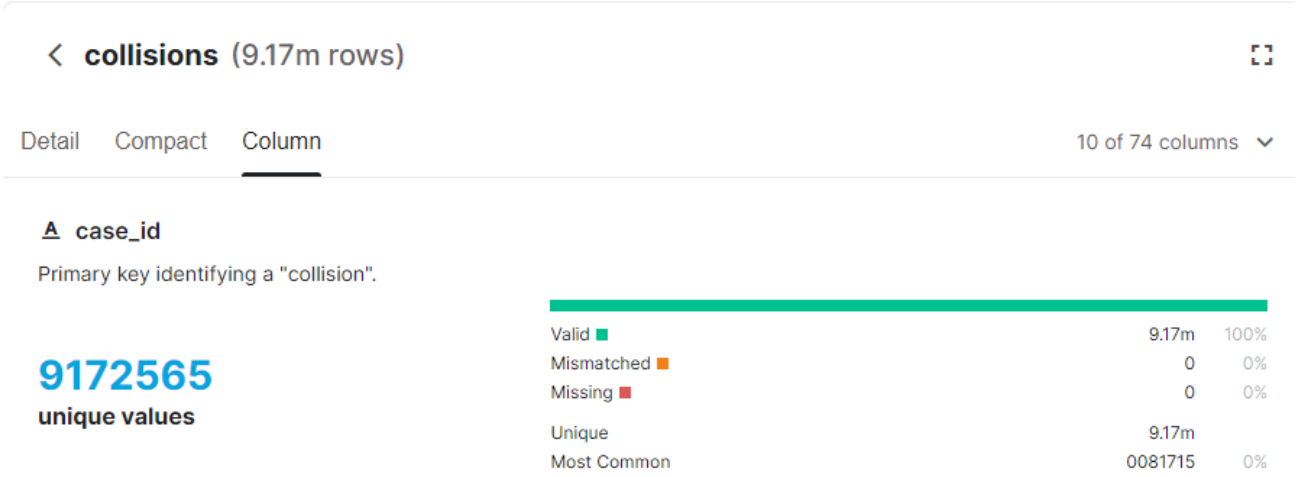
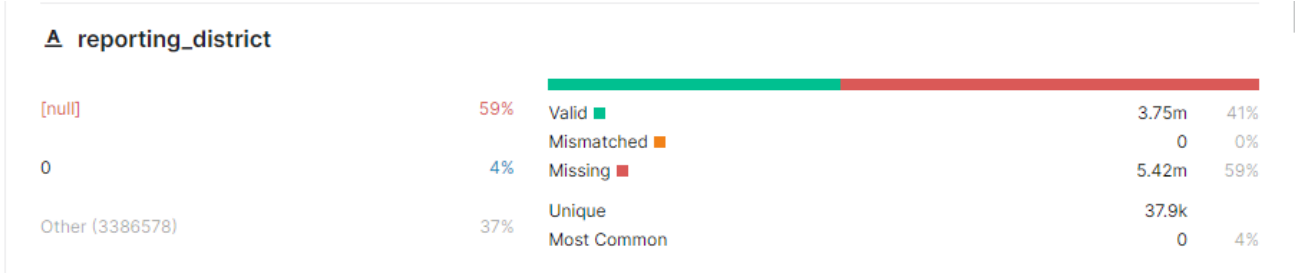
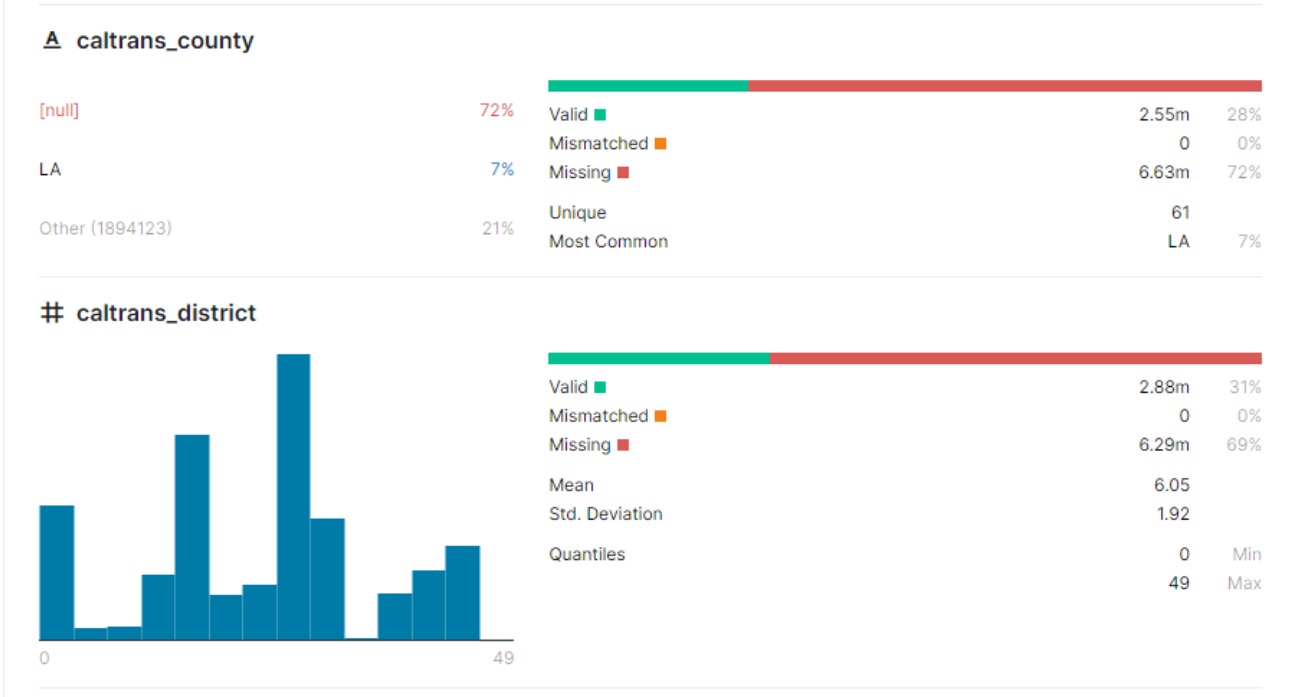


Tabela collisions zawiera ponad 9 mln unikatowych rekordów. I jak widać z rysunku powyżej, tabela zawiera 74 kolumny. Duża część z których nie jest przydatna do analizy.

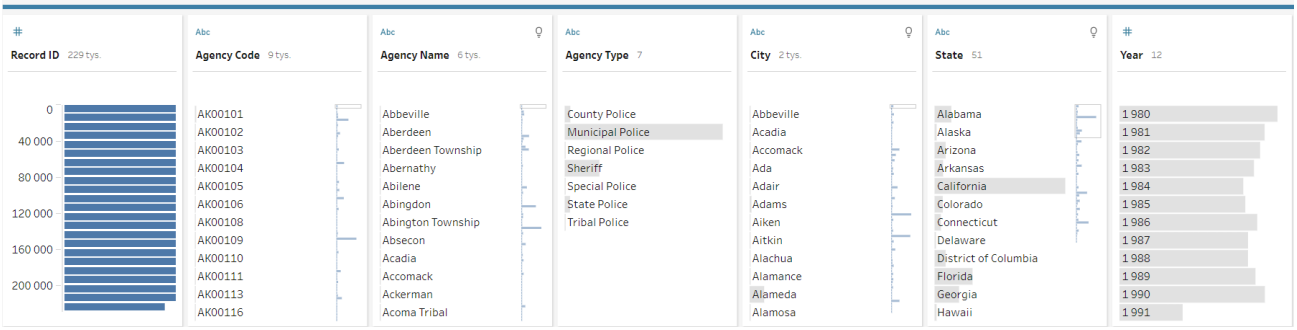


Dla przykładu wezmę kolumnę reporting_district, zawiera ona ponad 5 mln rekordów z wartością NULL.

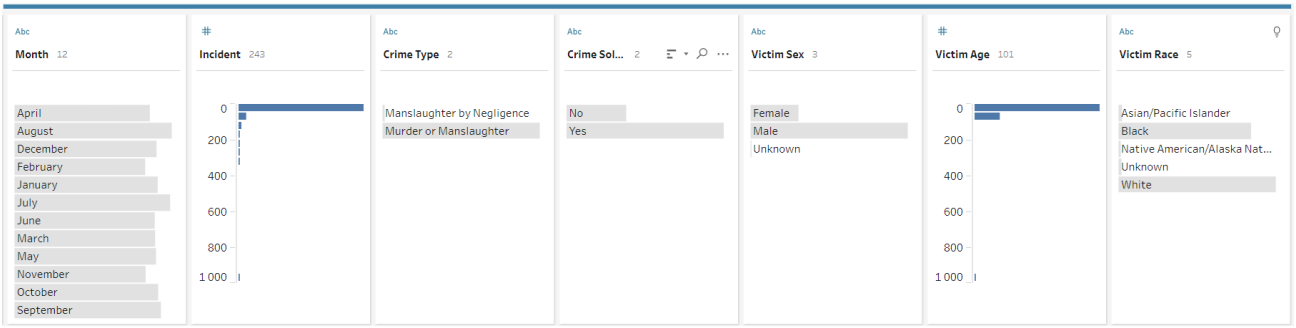


Jeszcze kilka przykładów z pustymi wartościami. Więc, już możemy zrobić wniosek, że więcej połowy danych nie są przydane do robienia jakichkolwiek badań. Poniżej jest wytłumaczenie dlaczego dany temat i baza danych nie zostały wybrane.

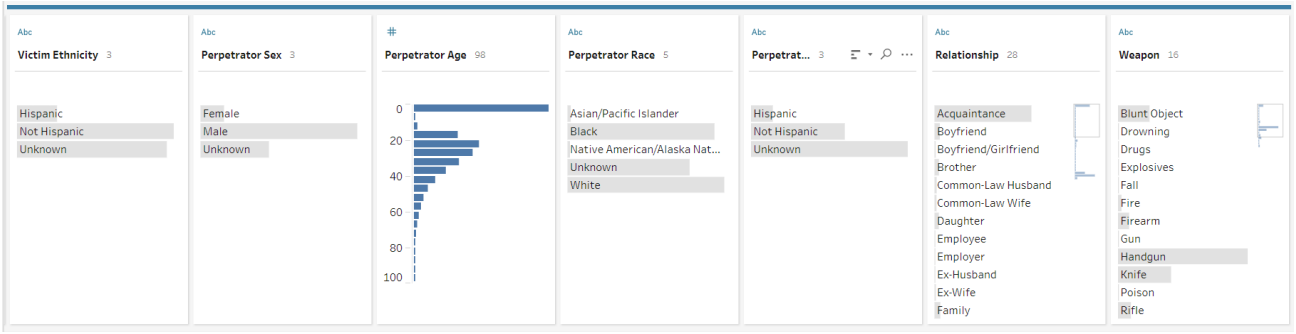
Temat 2:

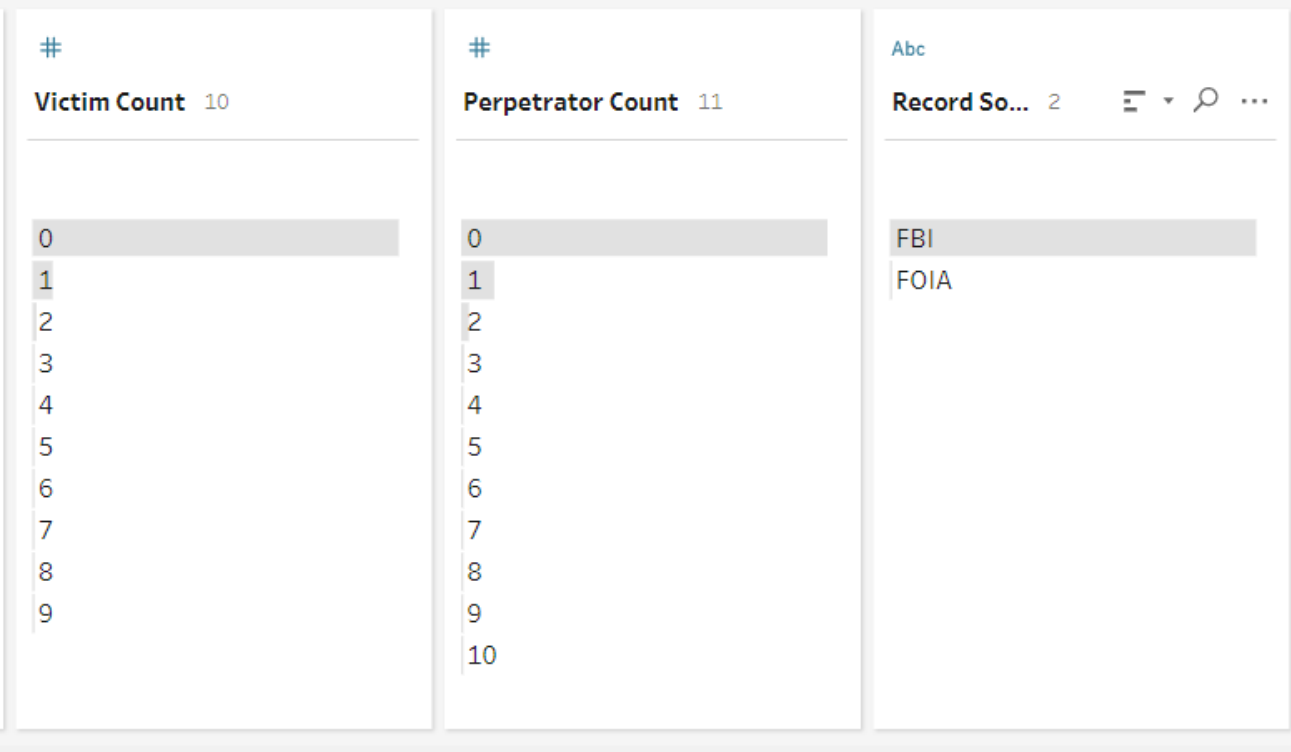


Jak widać z rysunków powyżej, żadna kolumna nie zawiera pustych znaczeń.



Co właśnie możemy zobaczyć i tutaj. Oprócz tego, kolumna Victim Age zawiera kilka pól, gdzie rok ofiary jest większy od 200 (co jest mało prawdopodobnie, a to i nierzeczywiste). Dlatego przed analizą musimy walidować dane takiego typu.





Victim Count 10
0
1
2
3
4
5
6
7
8
9

Perpetrator Count 11
0
1
2
3
4
5
6
7
8
9
10

Abc Record So... 2
FBI
FOIA

Victim i Perpetrator Count zawierają dużo pól z wartością 0, znaczy to, że ofiara jest jedna, zero znaczy – liczbę dodatkowych ofiar oraz przestępców.

Nieprzydatne kolumny do analizy:

- Record Source

Uzasadnienie:

1. Wypadki drogowe w Kalifornii są bardzo poważnym tematem, bo codziennie dużo ludzi giną od takich zdarzeń. Przeanalizowanie danych może zapobiegnąć większą część tych wypadków przez ustawienie i znalezienie słabych miejsc w przepisach drogowych. Podana baza danych daje nam szeroki wybór badań. Zawiera ona ponad 6 milinów rekordów, co jest zbyt dużo ze względu wydajności sprzętu, na którym będą przeprowadzone badania. Oprócz tego, jak widać z tabeli, ta baza zawiera duża część niepotrzebnych kolumn, gdzie więcej połowy wartości są NULL. Jest tu zbyt dużo opuszczonych danych.

2. Zabójstwa w USA są nie mniej ważnym tematem. Analiza takich danych pozwala zapobiegać dużą liczbę zabójstw. Jest to ważny punkt ze względu ratowania życia innych ludzi. Także baza danych zawiera 100% prawidłowych (bez NULL) rekordów i wiele różnych pól (a nawet różnych typów), co daje nam szeroki wybór prowadzania różnych badań, takie jak miasta z największym wskaźnikiem zabójstw, średni wiek zabójców i inne.

Jak wynik, wybieram drugi temat, bo wydaje mi się on bardziej sensowny ze względu poprawności danych oraz ich kompletności. Baza nie zawiera pustych danych, czyli rekordów ze znaczeniem NULL, nie zawiera niepotrzebnych kolumn.

Zakres opracowania projektu HD

1. **Tytuł projektu**

   Zabójstwa w USA w latach 1980-2014

2. **Charakterystyka dziedziny problemowej**

   2.1. Opis obszaru analizy wraz z uzasadnieniem (wybrany fragment dziedziny, przeznaczony do szczegółowej analizy i opracowania hurtowni danych)

   Dana baza jest najpełniejszą dostępną obecnie bazą danych o zabójstwach w Stanach Zjednoczonych. Ten zbiór danych obejmuje morderstwa z Dodatkowego Raportu Zabójstw FBI od 1976 do chwili obecnej oraz dane z Freedom of Information Act dotyczące ponad 22 000 zabójstw, które nie zostały zgłoszone do Justice Department. Ten zbiór danych obejmuje wiek, rasę, płeć, pochodzenie etniczne ofiar i sprawców, a także związek między ofiarą a sprawcą i używaną bronią. Wybranym fragmentem zostanie analiza najbardziej niebezpiecznych miast oraz województw, a więc czy ma jakiś wpływ wiek zabójców oraz ich dane etniczne.

   2.2. Problemy

   Problemem jest dość wysoki wskaźnik zabójstw w USA, jak kraju wysoko rozwiniętym. Także: które miasto oraz województwo mają najwięcej zabójstw, średni wiek zabójcy w każdym mieście, miesiące z największą liczbą zabójstw, średni wiek ofiar.

   2.3. Cel i zakres przedsięwzięcia

   Celem jest analiza utworzonej bazy danych z rekordami zabójstw i poznanie głównych wskaźników problemu. Także celem analizy jest redukcja liczby zabójstw w danym kraju przez odpowiednie reagowanie służb ratowniczych oraz policji.

   2.4. Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

   - Sumaryczna liczba zabójstw każdego roku
   - Sumaryczna liczba zabójców każdego roku
   - Miasta z największym wskaźnikiem przestępstw
   - Rok z największą liczbą zabójstw (coś miało wpływ na to)
   - Miesiące z największą liczbą zabójstw
   - Średnia liczba ofiar według każdego miasta

   2.5. Zakres analizy – badane aspekty

   2.5.1. Wpływ miasta na ilość przestępstw

2.5.2. Wpływ miesiąca na ilość przestępstw
2.5.3. Typ związku między ofiarą a przestępcą
2.5.4. Analiza liczby zabójstw każdego roku (czy jest tendencja spadkowa, czy nie)

3. **Źródła danych**

3.1. Charakterystyka źródeł danych

Lp.	Plik, nazwa bazy danych	Typ	Liczba rek.	Rozmiar[MB]	Opis
1.	Zabów.csv	CSV	638454	106.63	Raporty o zabójstwach z wniosków FBI i FOIA

3.2. Lokalizacja, dostępność

Plik bazy można pobrać ze strony https://www.kaggle.com/murderaccountability/homicide-reports poprzednio utworzywszy konto. Jest on w wolnym dostępie. Licencja: CC BY-SA 4.0

3.3. Słownik danych

- Record ID – numer identyfikiujący rekordu
- Agency Code – kod agencji
- Agency Name – nazwa agencji
- Agency Type – typ agencji
- City – miasto
- State – województwo, gdzie było zabójstwo
- Year – rok zabójstwa
- Month – jego miesiąc
- Incident - wydarzenie
- Crime Type - typ przestępstwa
- Crime Solved – czy przestępstwo zostało rozwiązane
- Victim Sex – płeć ofiary
- Victim Age – wiek ofiary
- Victim Race – rasa ofiary
- Victim Ethnicity – pochodzenie etniczne ofiary
- Perpetrator Age - wiek przestępcy
- Perpetrator Race - rasa przestępcy
- Perpetrator Ethnicity - pochodzenie etniczne przestępcy
- Relationship - związek
- Weapon - broń
- Victim Count – liczba ofiar
- Perpetrator Count – liczba przestępców

- Record Source – źródło danych

3.4. Ocena jakościowa źródeł danych

Plik: zabój.csv				
Lp.	**Atrybut**	**Typ wartości**	**Zakres wartości**	**Ocena jakości danych**
1.	Record ID	Numeryczne	0 – 230 000	0% null, 229 007 unique
2.	Agency Code	Tekstowe	Długość: 7	0% null, 8 965 unique
3	Agency Name	Tekstowe	Długość: 4-12	0% null, 6 312 unique
4	Agency Type	Tekstowe	Długość: 7-20	0% null, 7 unique
5	City	Tekstowe	Długość: 4-32	0% null, 1682 unique
6	State	Tekstowe	Długość: 4-32	0% null, 51 unique
7	Year	Numeryczne	1980 - 2014	0% null, 12 unique
8	Month	Tekstowe	Długość: 3-12	0% null, 12 unique
9	Incident	Numeryczne	0 - 1000	0% null 243 unique
10	Crime Type	Tekstowe	Długość: 24-30	0% null, 2 unique
11	Crime Solved	Tekstowe	„No” (28%) lub „Yes” (72%)	0% null, 2 unique
12	Victim Sex	Tekstowe	Długość: 4-7	0% null, 3 unique
13	Victim Age	Numeryczne	0 - 100	0% null, 1 value = 1000
14	Victim Race	Tekstowe	Długość: 5-25	0% null, 5 unique
15	Victim Ethnicity	Tekstowe	Długość: 8-12 „Unknown” 43%	0% null, 3 unique
16	Perpetrator Sex	Tekstowe	Długość: 4-7 „Unknown” 28%	0% null, 3 unique
17	Perpetrator Age	Numeryczne	0 - 100	0% null, 98 unique
18	Perpetrator Race	Tekstowe	Długość: 5-28 ”Unknown” 28%	0% null, 5 unique
19	Perpetrator Ethnicity	Tekstowe	Długość: 8-12 „Unknown” 58%	0% null, 3 unique
20	Relationship	Tekstowe	Długość: 3-22	0% null, 28 unique
21	Weapon	Tekstowe	Długość: 4-12	0% null, 16 unique
22	Victim Count	Numeryczne	0 - 9	0% null, 10 unique
23	Perpetrator Count	Numeryczne	0 - 10	11 unique, 0% null
24	Record Source	Tekstowe	„FBI” (>99%) lub „FOIA”	0% null

4. Analityczne modele wielowymiarowe

4.1. Fakty podlegające analizie oraz miary

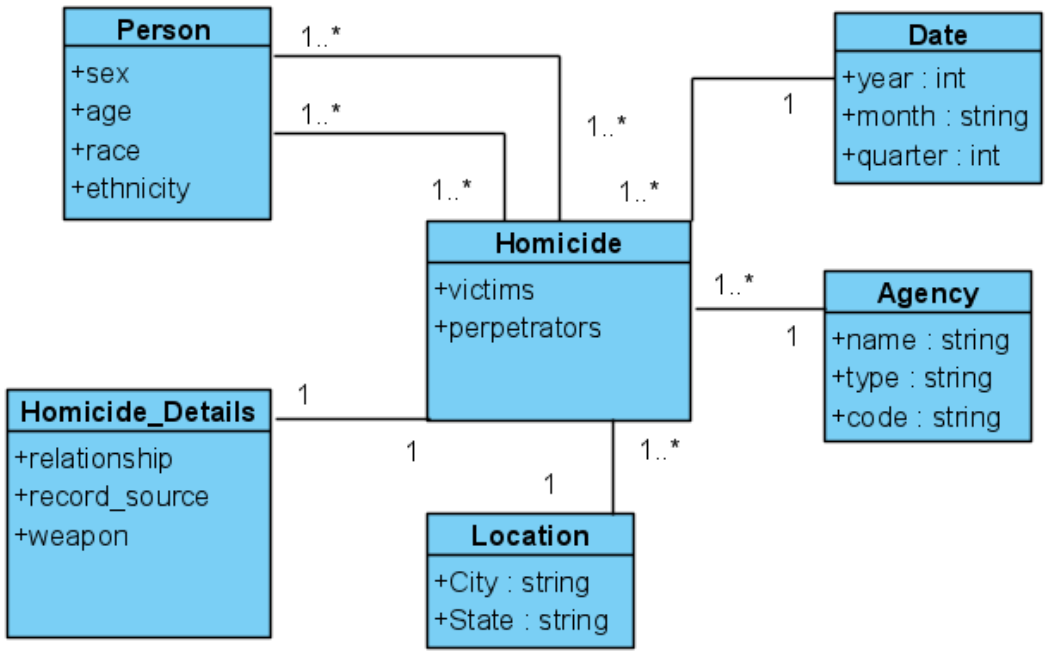
Lp.	Fakt	Miara	Uwagi
1.	Zabójstwo	Liczba ofiar	
		Liczba przestępców	

4.2. Kontekst analizy faktów np. czas (ziarnistość), lokalizacja, warunki pogodowe, itd.

Lp.	Wymiary	Atrybuty	Uwagi

1.	Lokalizacja	Miasto, Województwo	
2.	Agencja	Nazwa, typ, kod agencji	
3.	Czas	Miesiąc, rok	
4.	Ofiara	Rasa, pochodzenie etniczne, wiek, płeć	

4.3. Wielowymiarowe modele analityczne – poziom konceptualny (diagram klas UML)



5. Projekt procesu ETL

5.1. Schemat bazy danych HD (skrypt SQL)

```
CREATE DATABASE Homicides;
CREATE TABLE [dimAgency](
      [PK_dimAgency] [int] IDENTITY(1,1) NOT NULL,
      [Name] [nvarchar](50) NULL,
      [Type] [nvarchar](50) NULL,
      [Code] [nvarchar](50) NULL,
 CONSTRAINT [PK_dimAgency] PRIMARY KEY CLUSTERED
(
      [PK_dimAgency] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY =
OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dimDate](
      [PK_Date] [datetime] NOT NULL,
      [Date_Name] [nvarchar](50) NULL,
      [Year] [datetime] NULL,
      [Year_Name] [nvarchar](50) NULL,
      [Quarter] [datetime] NULL,
```

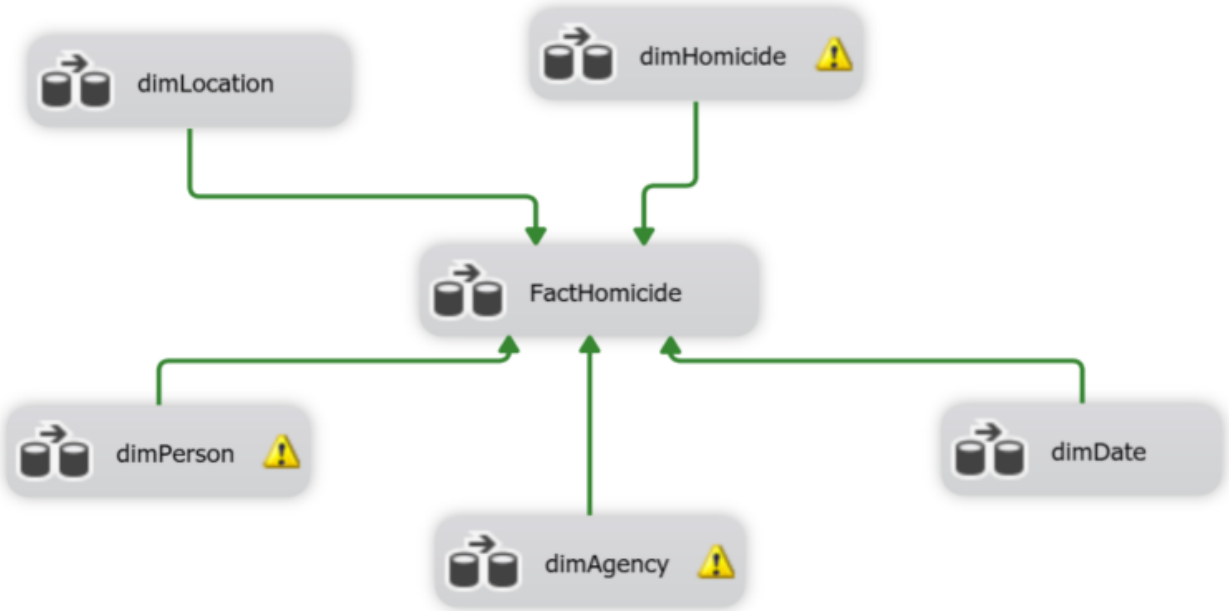
```
	[Quarter_Name] [nvarchar](50) NULL,
	[Month] [datetime] NULL,
	[Month_Name] [nvarchar](50) NULL,
	[Month_Of_Year] [int] NULL,
	[Month_Of_Year_Name] [nvarchar](50) NULL,
	[Month_Of_Quarter] [int] NULL,
	[Month_Of_Quarter_Name] [nvarchar](50) NULL,
	[Quarter_Of_Year] [int] NULL,
	[Quarter_Of_Year_Name] [nvarchar](50) NULL,
 CONSTRAINT [PK_dimDate] PRIMARY KEY CLUSTERED
(
	[PK_Date] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY =
OFF) ON [PRIMARY]
) ON [PRIMARY]


CREATE TABLE [dimHomicide](
	[PK_dimHomicide] [int] IDENTITY(1,1) NOT NULL,
	[Relationship] [nvarchar](50) NULL,
	[RecordSource] [nvarchar](50) NULL,
	[Weapon] [nvarchar](50) NULL,
	[FK_dimPerson_Victim] [int] NULL,
	[FK_dimPerson_Perpetrator] [int] NULL,
 CONSTRAINT [PK_dimHomicide_1] PRIMARY KEY CLUSTERED
(
	[PK_dimHomicide] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY =
OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE TABLE [dimLocation](
	[PK_dimLocation] [int] IDENTITY(1,1) NOT NULL,
	[State] [nvarchar](50) NULL,
	[City] [nvarchar](50) NULL,
 CONSTRAINT [PK_dimLocation] PRIMARY KEY CLUSTERED
(
	[PK_dimLocation] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY =
OFF) ON [PRIMARY]
) ON [PRIMARY]


CREATE TABLE [dimPerson](
	[PK_dimPerson] [int] IDENTITY(1,1) NOT NULL,
	[Sex] [nvarchar](50) NULL,
	[Age] [smallint] NULL,
	[Race] [nvarchar](50) NULL,
	[Ethnicity] [nvarchar](50) NULL,
 CONSTRAINT [PK_dimPerson] PRIMARY KEY CLUSTERED
(
	[PK_dimPerson] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY =
OFF) ON [PRIMARY]
) ON [PRIMARY]


CREATE TABLE [FactHomicide](
	[FK_dimDate] [datetime] NULL,
```

```
	[FK_dimHomicide] [int] NULL,
	[FK_dimLocation] [int] NULL,
	[FK_dimAgency] [int] NULL,
	[FK_dimPerson] [int] NULL,
	[Victims] [int] NULL,
	[Perpetrators] [int] NULL,
	[PK_factHomicide] [int] IDENTITY(1,1) NOT NULL,
 CONSTRAINT [PK_FactHomicide] PRIMARY KEY CLUSTERED
(
	[PK_factHomicide] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON, OPTIMIZE_FOR_SEQUENTIAL_KEY =
OFF) ON [PRIMARY]
) ON [PRIMARY]

CREATE NONCLUSTERED INDEX [IX_FactHomicide] ON [FactHomicide]
(
	[FK_dimDate] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, SORT_IN_TEMPDB = OFF,
DROP_EXISTING = OFF, ONLINE = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]

CREATE NONCLUSTERED INDEX [IX_FactHomicide1] ON [FactHomicide]
(
	[FK_dimHomicide] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, SORT_IN_TEMPDB = OFF,
DROP_EXISTING = OFF, ONLINE = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]

CREATE NONCLUSTERED INDEX [IX_FactHomicide2] ON [FactHomicide]
(
	[FK_dimLocation] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, SORT_IN_TEMPDB = OFF,
DROP_EXISTING = OFF, ONLINE = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]

CREATE NONCLUSTERED INDEX [IX_FactHomicide3] ON [FactHomicide]
(
	[FK_dimAgency] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, SORT_IN_TEMPDB = OFF,
DROP_EXISTING = OFF, ONLINE = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]

CREATE NONCLUSTERED INDEX [IX_FactHomicide4] ON [FactHomicide]
(
	[FK_dimPerson] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, SORT_IN_TEMPDB = OFF,
DROP_EXISTING = OFF, ONLINE = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]

ALTER TABLE [dimHomicide]  WITH CHECK ADD  CONSTRAINT
[FK_dimHomicide_Perpetrator] FOREIGN KEY([FK_dimPerson_Perpetrator])
REFERENCES [dimPerson] ([PK_dimPerson])

ALTER TABLE [dimHomicide] CHECK CONSTRAINT [FK_dimHomicide_Perpetrator]

ALTER TABLE [dimHomicide]  WITH CHECK ADD  CONSTRAINT
[FK_dimHomicide_Victim] FOREIGN KEY([FK_dimPerson_Victim])
REFERENCES [dimPerson] ([PK_dimPerson])

ALTER TABLE [dimHomicide] CHECK CONSTRAINT [FK_dimHomicide_Victim]
```

```
ALTER TABLE [FactHomicide]  WITH CHECK ADD  CONSTRAINT [FactHomicide-
dimAgency] FOREIGN KEY([FK_dimAgency])
REFERENCES [dimAgency] ([PK_dimAgency])

ALTER TABLE [FactHomicide] CHECK CONSTRAINT [FactHomicide-dimAgency]

ALTER TABLE [FactHomicide]  WITH CHECK ADD  CONSTRAINT [FactHomicide-
dimDate] FOREIGN KEY([FK_dimDate])
REFERENCES [dimDate] ([PK_Date])

ALTER TABLE [FactHomicide] CHECK CONSTRAINT [FactHomicide-dimDate]

ALTER TABLE [FactHomicide]  WITH CHECK ADD  CONSTRAINT [FactHomicide-
dimHomicide] FOREIGN KEY([FK_dimHomicide])
REFERENCES [dimHomicide] ([PK_dimHomicide])

ALTER TABLE [FactHomicide] CHECK CONSTRAINT [FactHomicide-dimHomicide]

ALTER TABLE [FactHomicide]  WITH CHECK ADD  CONSTRAINT [FactHomicide-
dimLocation] FOREIGN KEY([FK_dimLocation])
REFERENCES [dimLocation] ([PK_dimLocation])

ALTER TABLE [FactHomicide] CHECK CONSTRAINT [FactHomicide-dimLocation]

ALTER TABLE [FactHomicide]  WITH CHECK ADD  CONSTRAINT [FactHomicide-
dimPerson] FOREIGN KEY([FK_dimPerson])
REFERENCES [dimPerson] ([PK_dimPerson])

ALTER TABLE [FactHomicide] CHECK CONSTRAINT [FactHomicide-dimPerson]
```

5.2. Specyfikacja procesów ETL (Control Flow + Data Flow)

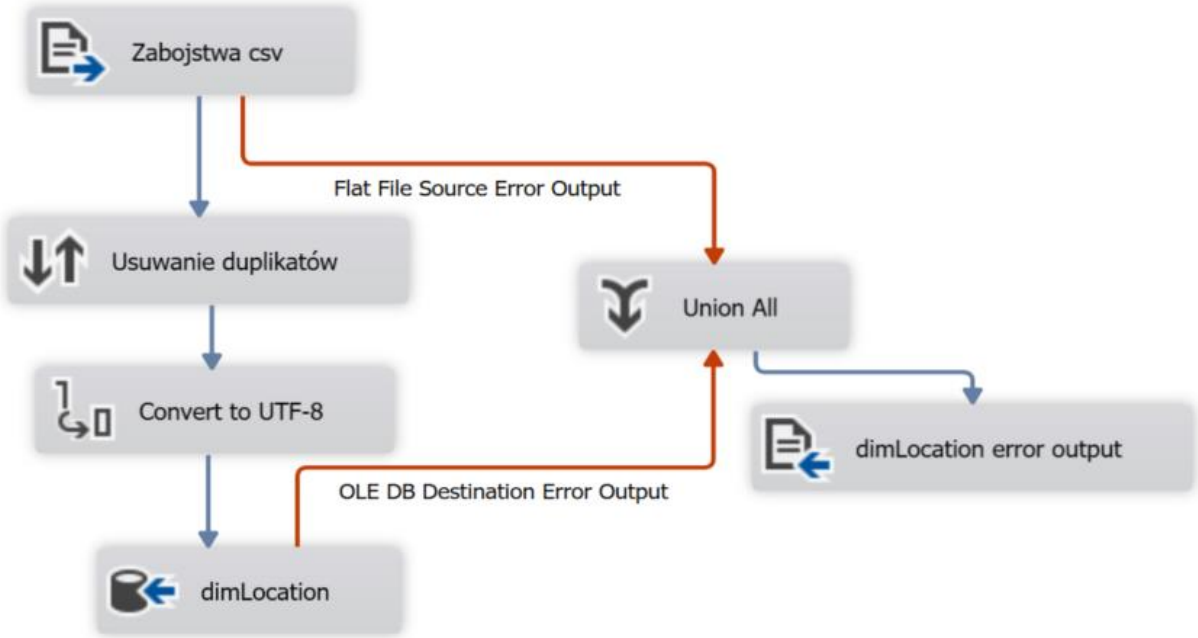
Control Flow:



Rys 1.

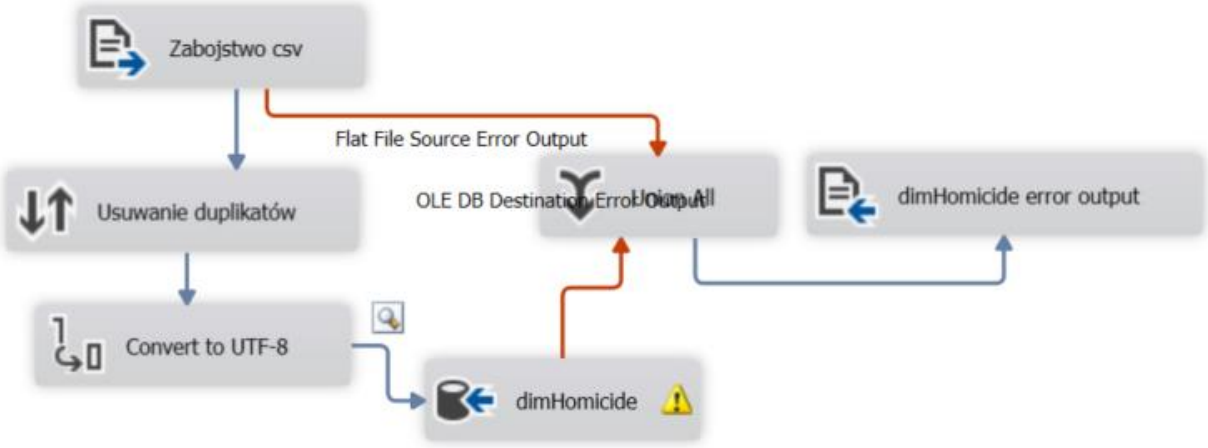
Control flow składa się z trzech etapów, czyli: **extract** danych z pliku źródłowego, transform - **transformacja** danych w potrzebną mi postać oraz **load** – dodanie otrzymanych danych do punktu końcowego (bazy danych).

a. **Wymiar lokalizacji:**



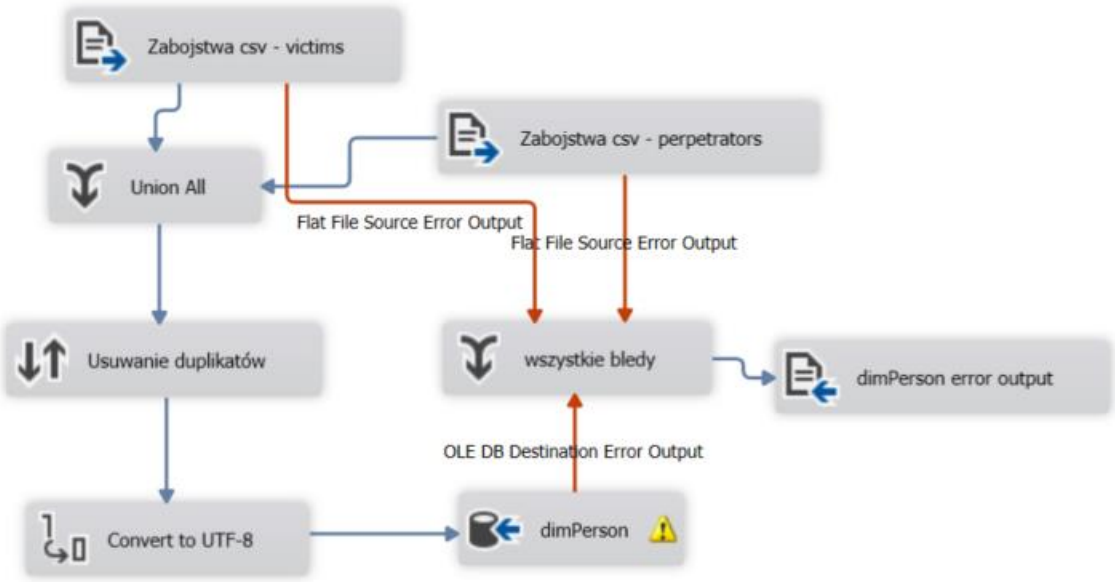
Składa się z kilku różnych etapów, główne z których są: usuwanie duplikatów przez sortowanie danych wejściowych, oraz konwersja danych do potrzebnego kodowania (UTF-8).

b. **Wymiar Homicide:**



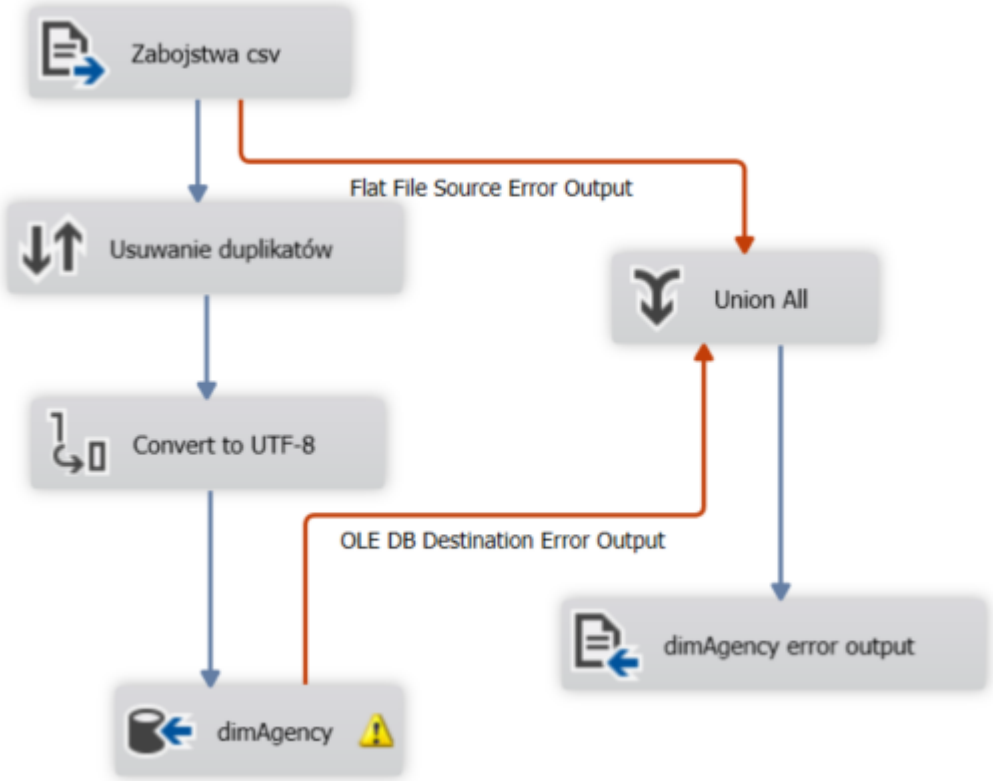
Podobnie do poprzedniego jest zrobiony ten wymiar. Też on zawiera usuwania duplikatów i konwersję danych do potrzebnego formatu.

c. **Wymiar Person:**



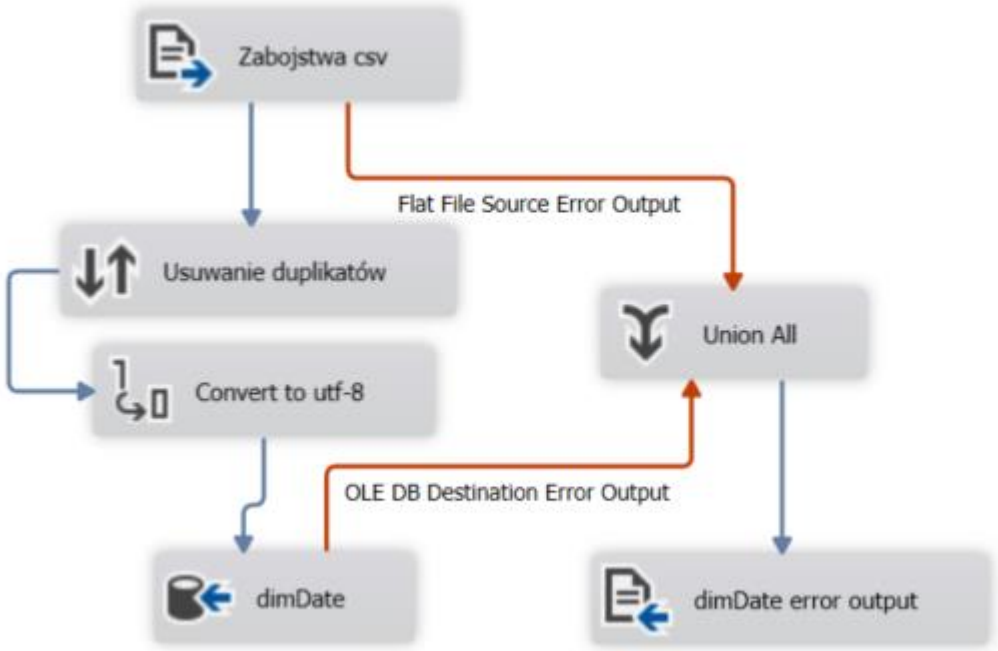
ETL zawiera unie dwóch rodzajów ludzi. Czyli ofiar oraz przestępców. Przed usuwaniem duplikatów łączymy ze sobą te dwie grupy i kontynuujemy proces ETLowy.

d. **Wymiar Agencji:**



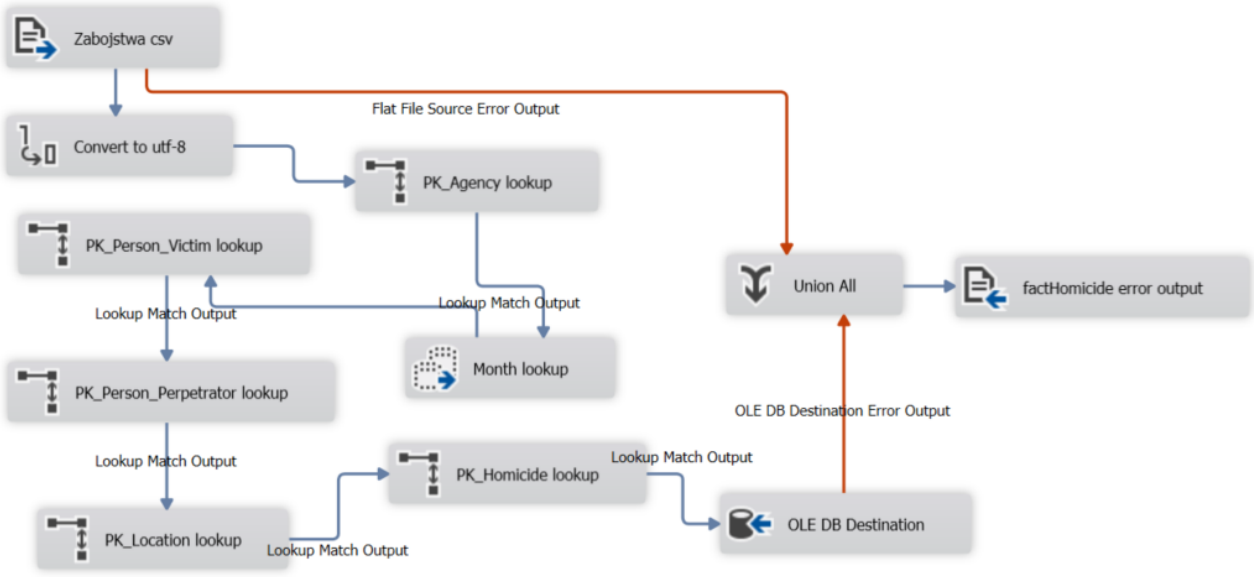
Podobnie do poprzednich przykładów robimy ten.

e. Wymiar Daty:



f. Fact Homicide

I na koniec mamy ETL faktu, gdzie łączymy wszystkie wymiary w jednym miejscu. Użyjmy Fuzzing lookup dla wyszukiwania daty po roku oraz miesiącu.



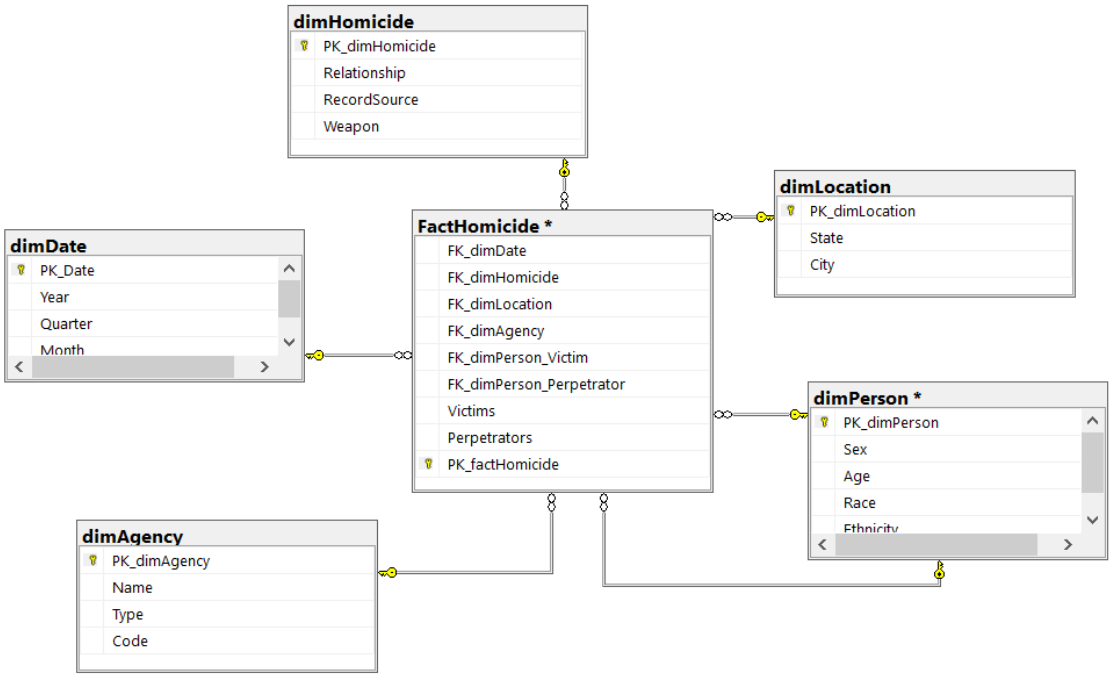
Aby policzyć do którego kwartału należy miesiąc użyjmy takiej klauzy. Będzie to dodatkowa informacja dla analizy.

```
UPDATE dbo.dimdate SET [Quarter] =
CASE dbo.dimdate.[month]
    WHEN 'January' THEN 1
    WHEN 'February' THEN 1
```

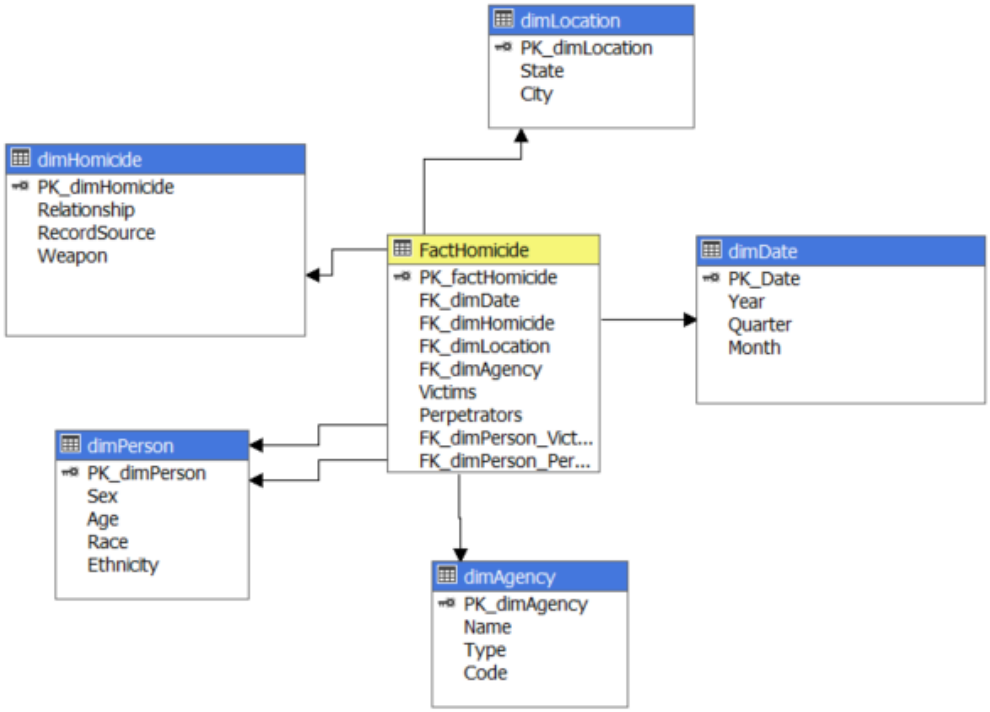
```
    WHEN 'March' THEN 1
    WHEN 'April' THEN 2
    WHEN 'May' THEN 2
    WHEN 'June' THEN 2
    WHEN 'July' THEN 3
    WHEN 'August' THEN 3
    WHEN 'September' THEN 3
    WHEN 'October' THEN 4
    WHEN 'November' THEN 4
    WHEN 'December' THEN 4
END
FROM dbo.dimdate;
```

6. Implementacja modeli wielowymiarowych

6.1. Widok danych

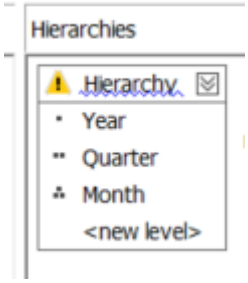


6.2. Kostki (modele wielowymiarowe)



Lista wymiarów użytych w projekcie:

Dla daty zdecydowałem zostawić tylko rok i miesiąc, według tego, że nie potrzebujemy dnia, w którym był wypadek (bo w tabeli źródłowej nie ma dni).

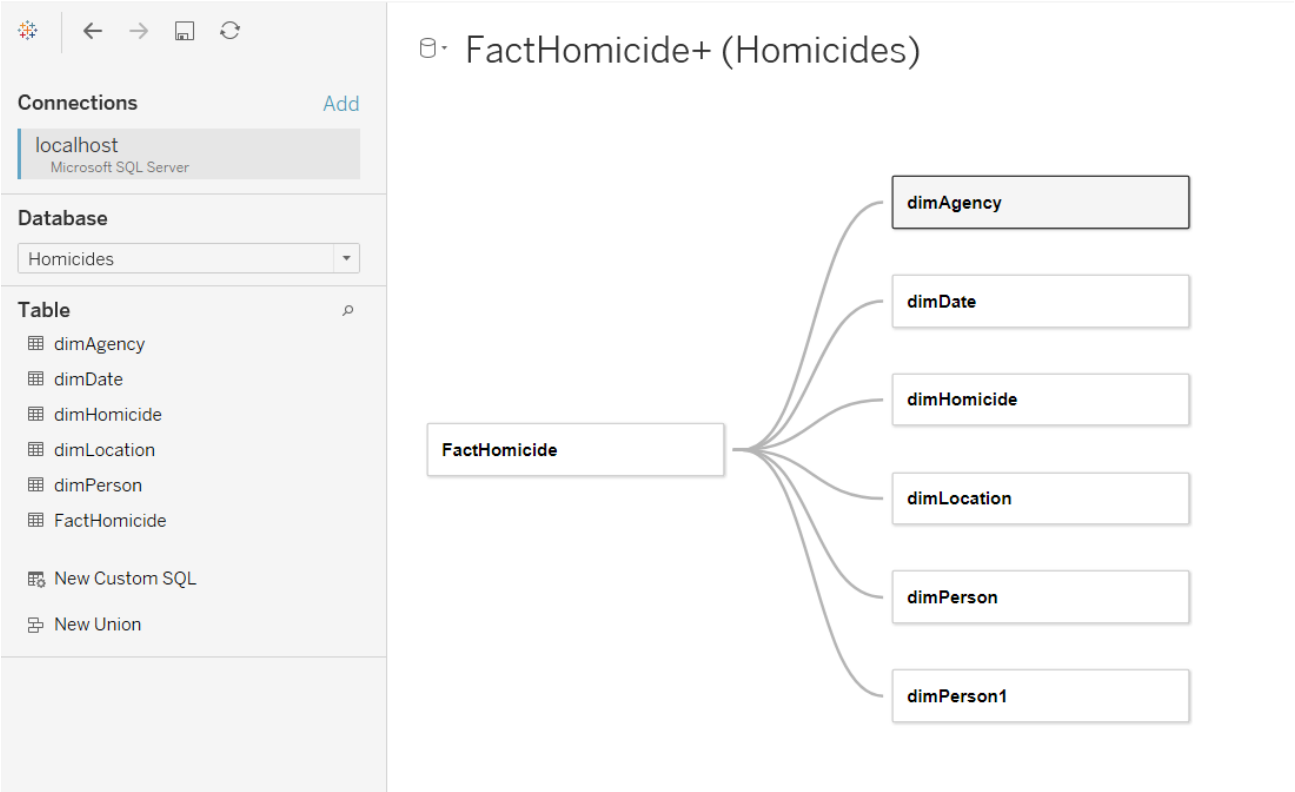


7. **Analiza danych**

7.1. Prezentacja procesu analitycznego

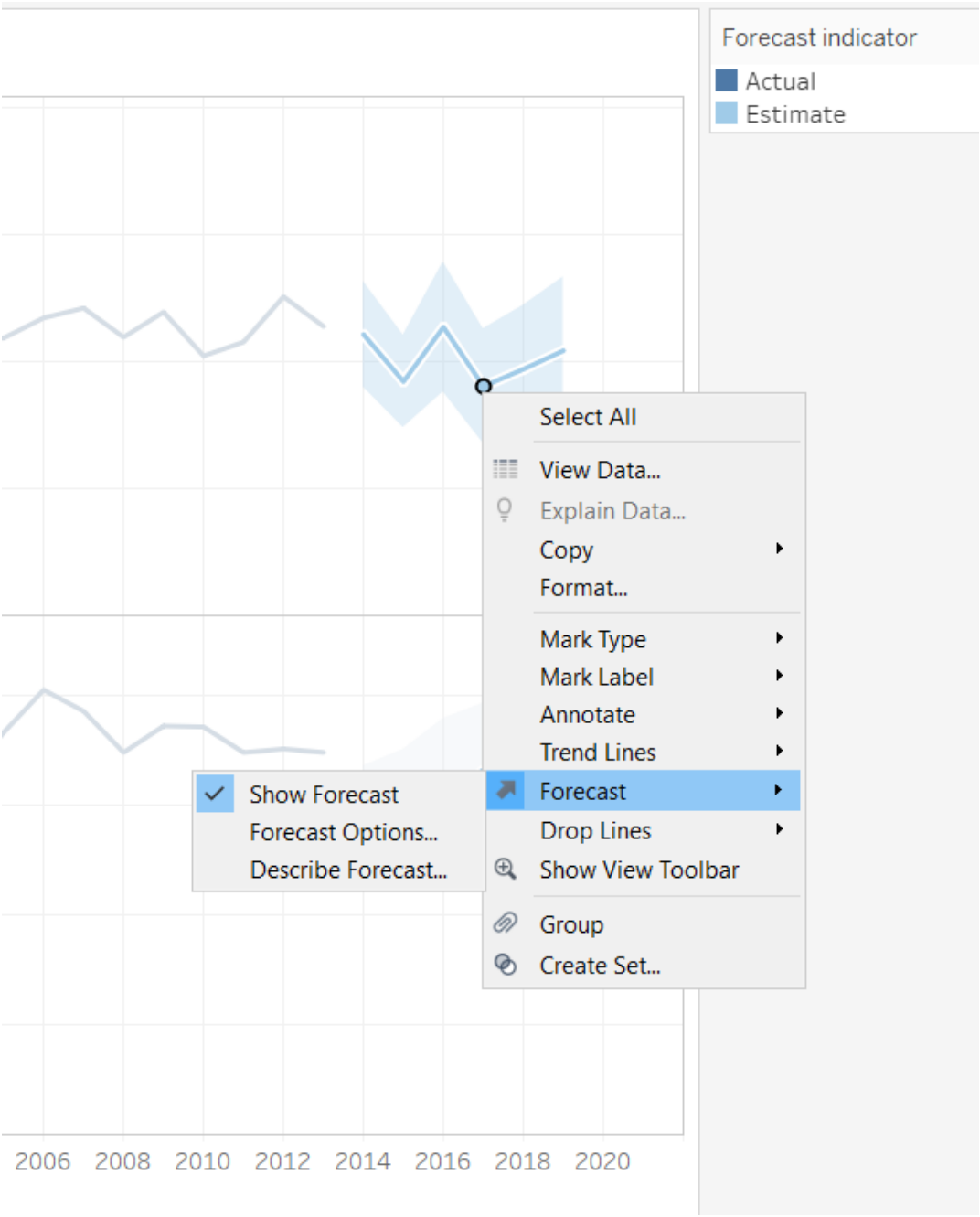
Procesy analityczne wykonałem za pomocą Tableau Desktop.

Po podłączeniu do MSSQL serweru możemy wybrać tabeli do analizy, Wybieramy wszystkie.

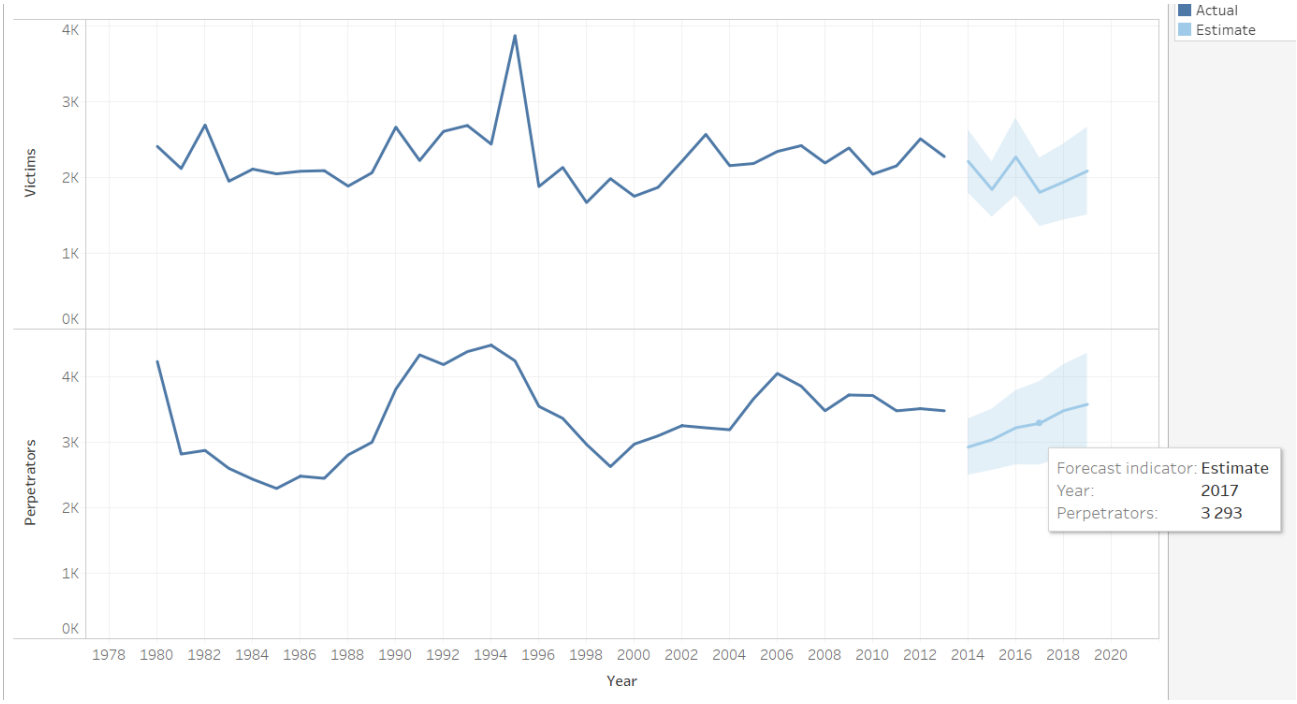


Po dodaniu przechodzimy do wizualizacji etapów badań.

Tableau Desktop da możliwość przewidywania danych w przyszłości:

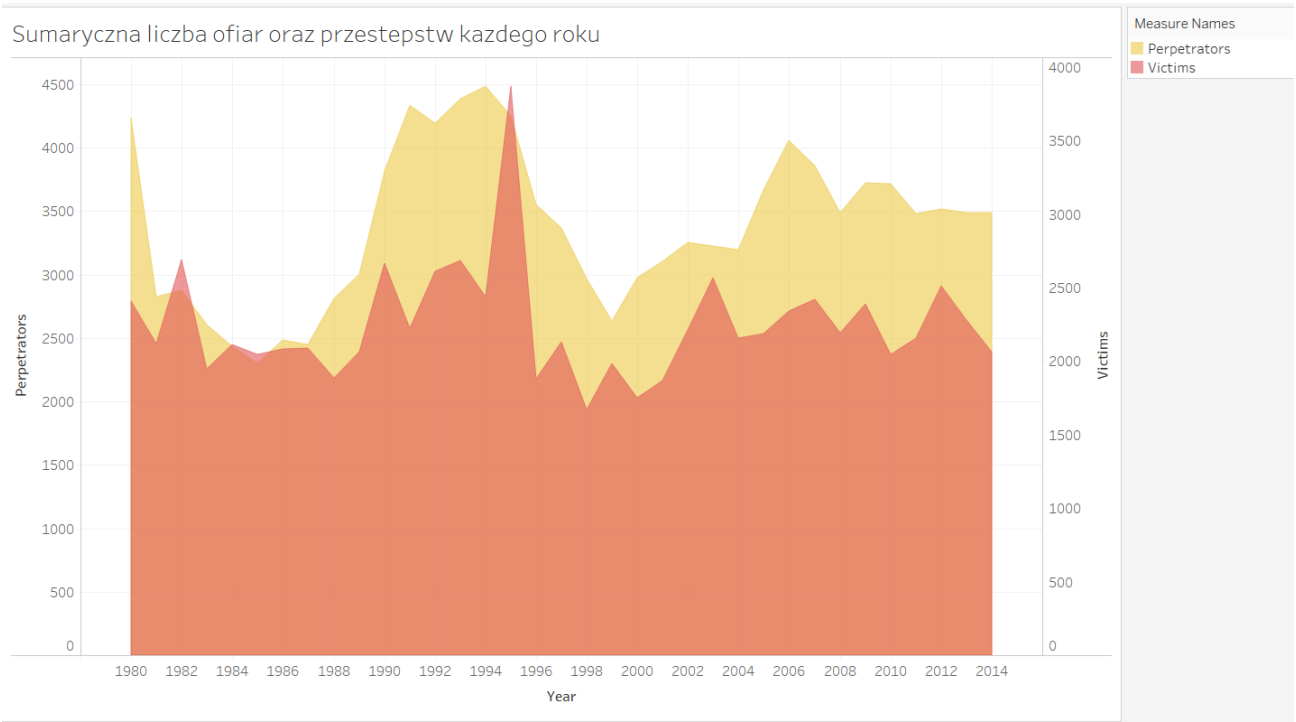


Na przykład otrzymaliśmy takie znaczenia:



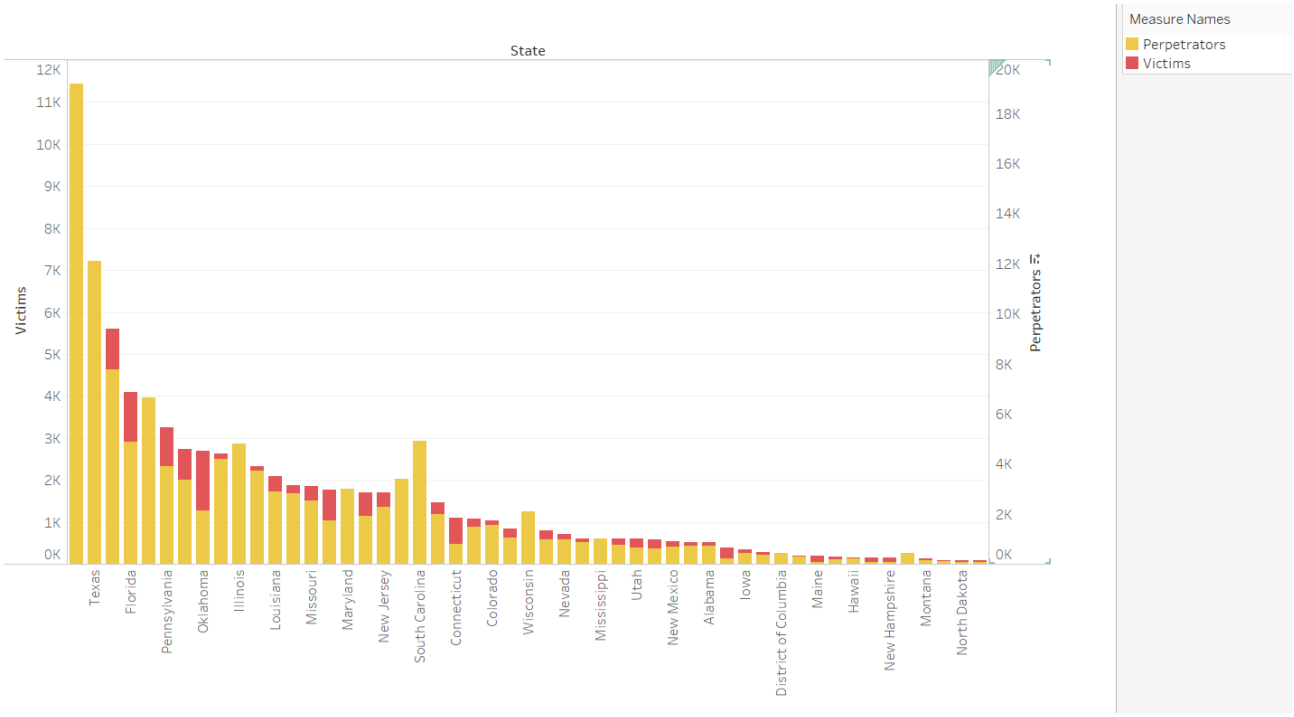
Możemy sprawdzić czy jest to dobra prognoza czy nie.

7.2. Podsumowanie - wnioski z analizy



Rys. 1

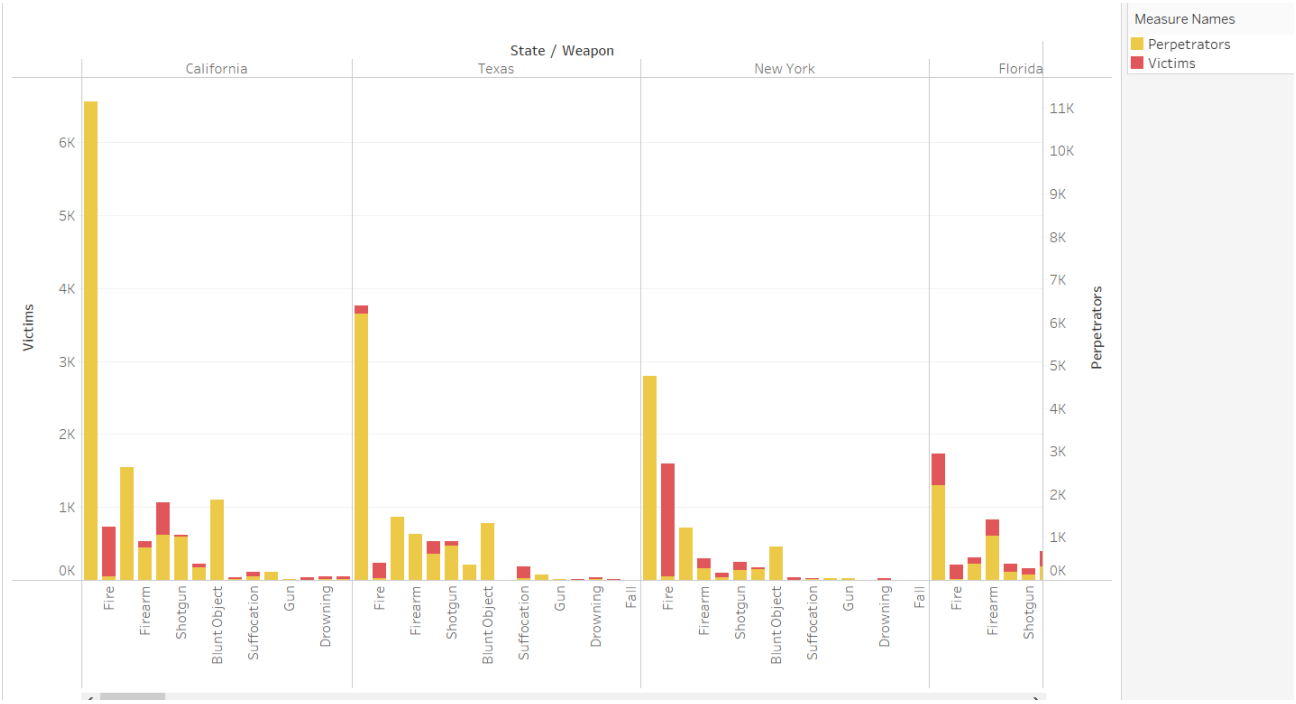
Ilość przestępców oraz ofiar wzrosła w 1990 i miała maksymalne znaczenie w 1994-1996 latach. Natomiast, najmniejsza ilość ofiar była w 1998, a przestępców w 1985 roku. Białoniebieskim kolorem jest oznaczona predykcja.



Rys 2.

Wykres przestępców według województwa. Jak widać, najbardziej przestępcze województwo jest California oraz Texas jako drugie.

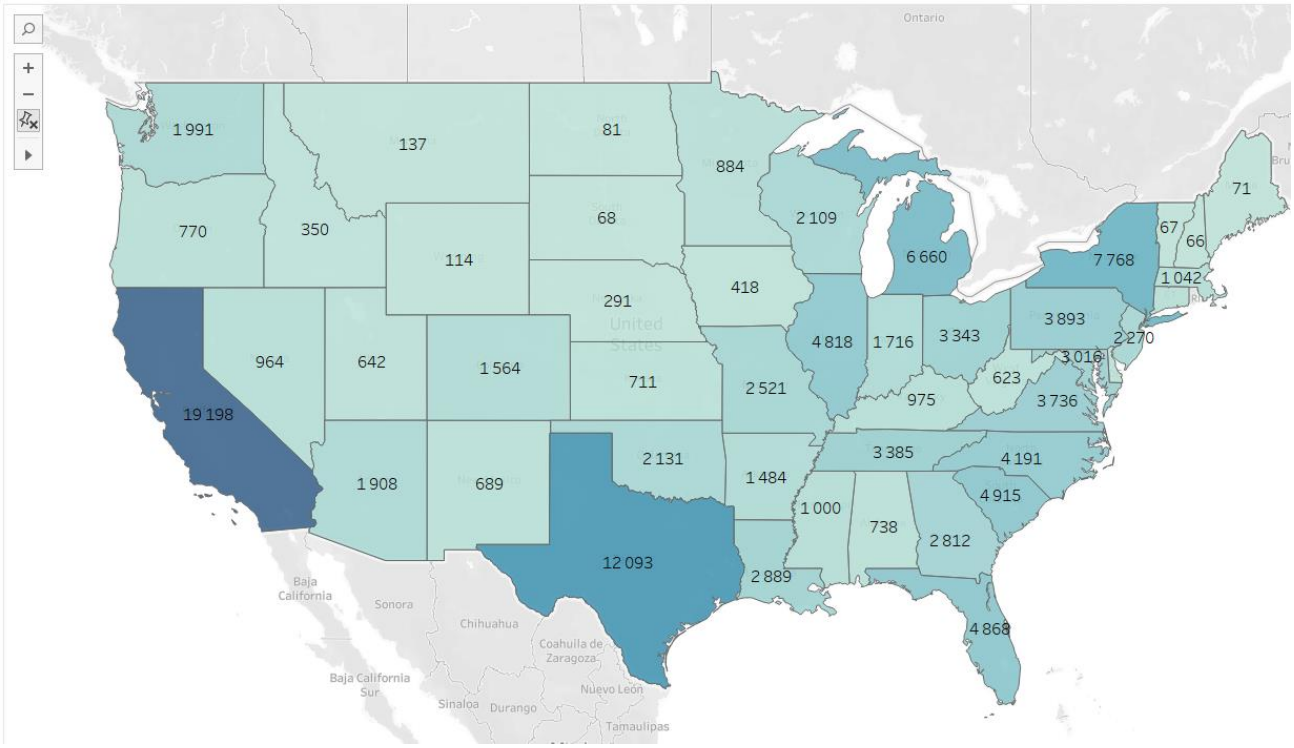
Broń według stanów:



Rys. 3

Grupowanie po województwu oraz broni. Z wykresu widać, że najczęściej używana broń jest pistolet. Niestety także, dużo ludzi ginie przez pożar.

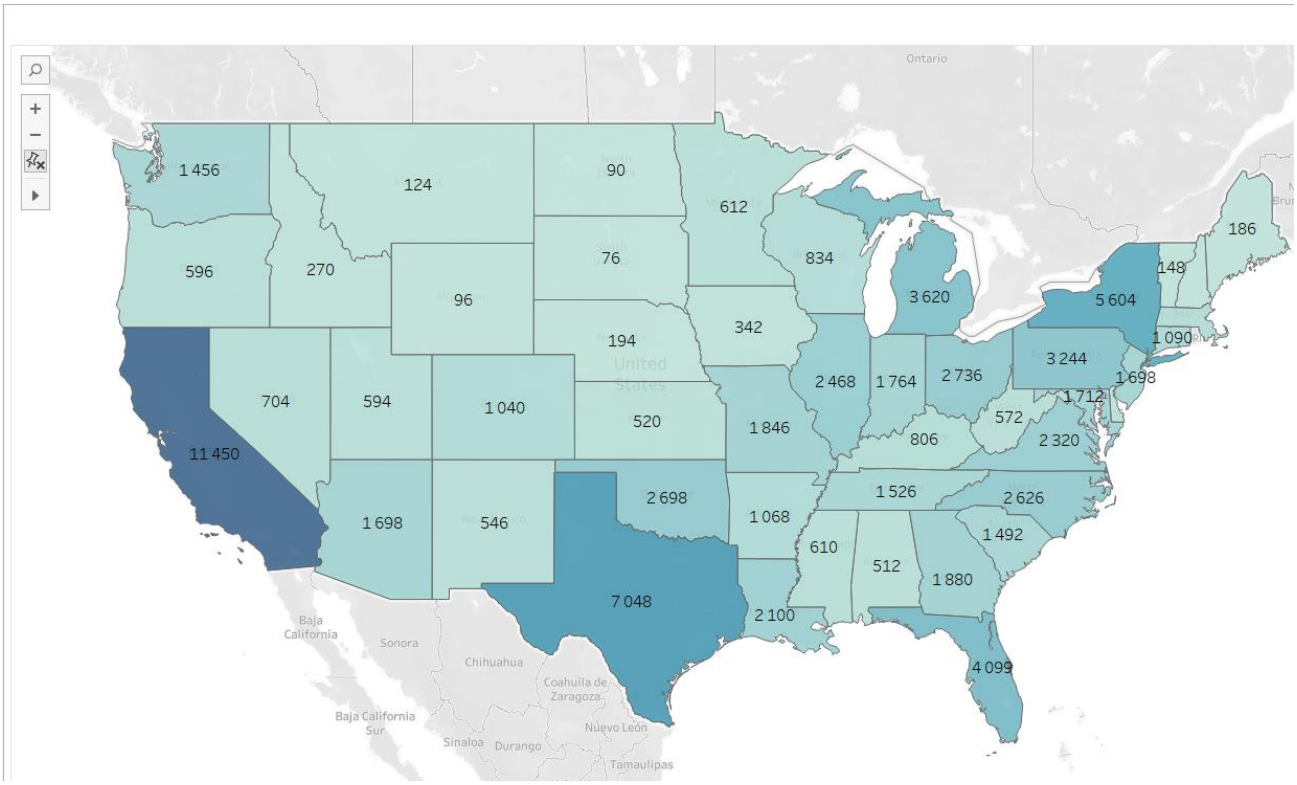
Liczba przestępców:



Rys. 4

Mapa USA według ilości przestępców. Jak już było omówiono, najbardziej przestępczym województwem jest Kalifornia a za nim Techas.

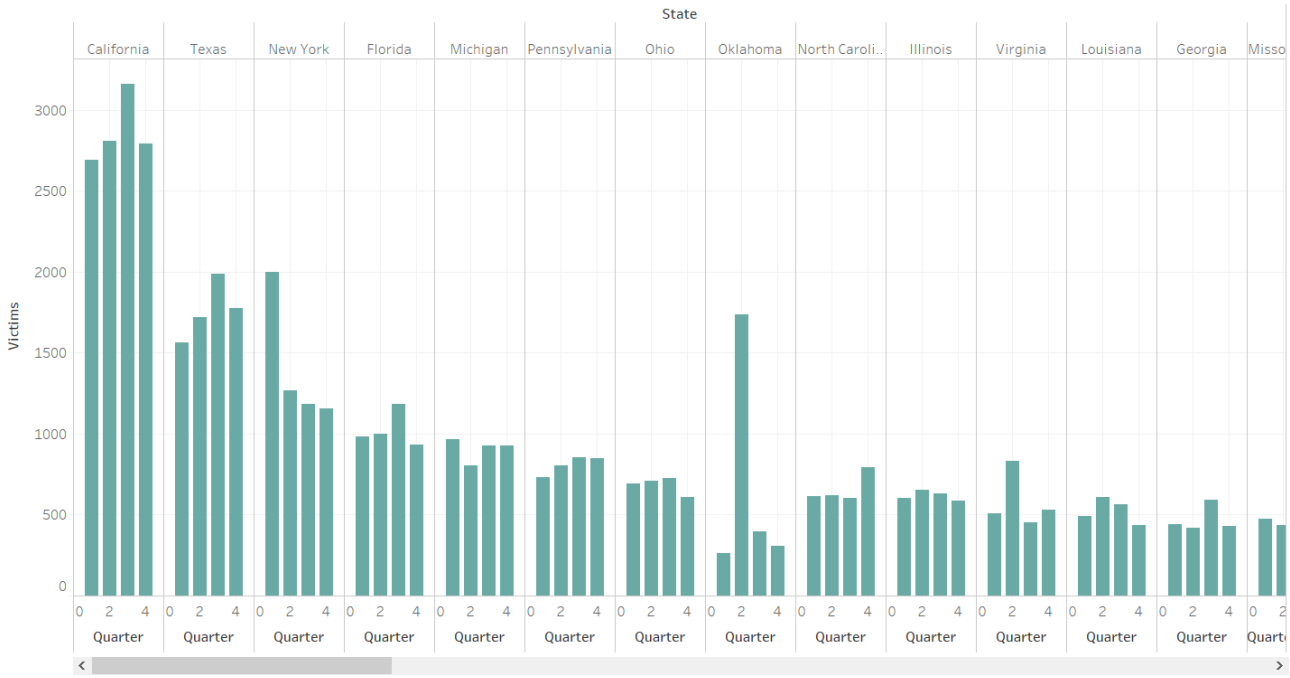
Liczba ofiar:



Rys. 5

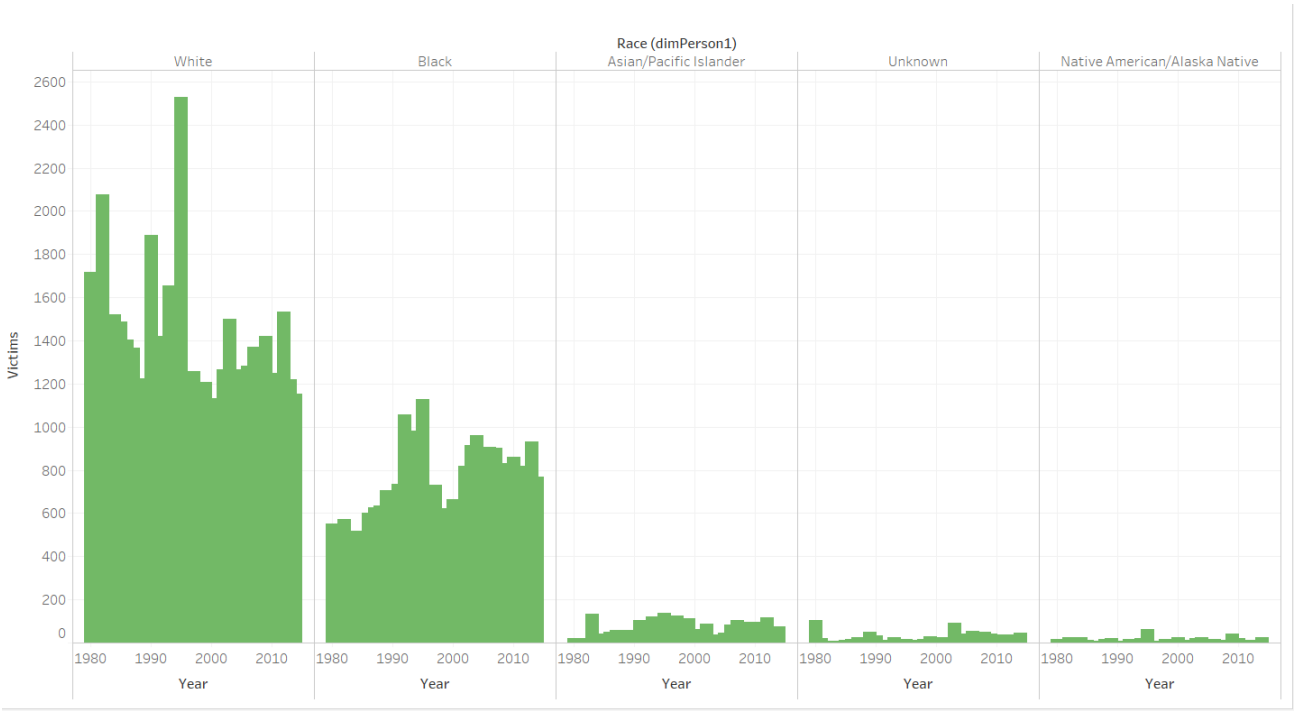
Odpowiednio do ilości przestępców, liczba ofiar jest w niektórym sensie zależna od tego. Więc, mamy takie wyniki, że najwięcej ofiar są w Kalifornii.

Ilość zabójstwa według kwartału oraz stanu:



Każdy stan ma swoją liczbę zabójstw według pory roku. Bardziej południowe stany informują, że latem jest więcej zabójstw, niż w jakąkolwiek inną porę roku. Jest to związane z lokalizacją geograficzną tych stanów. Widząc, że ilość zabójstw w stanie New York (który znajduje się na północy) jest maksymalna wiosną, co znaczy, że w większości przypadków przestępcom jest ta pora wygodniejsza dla ich akcji.

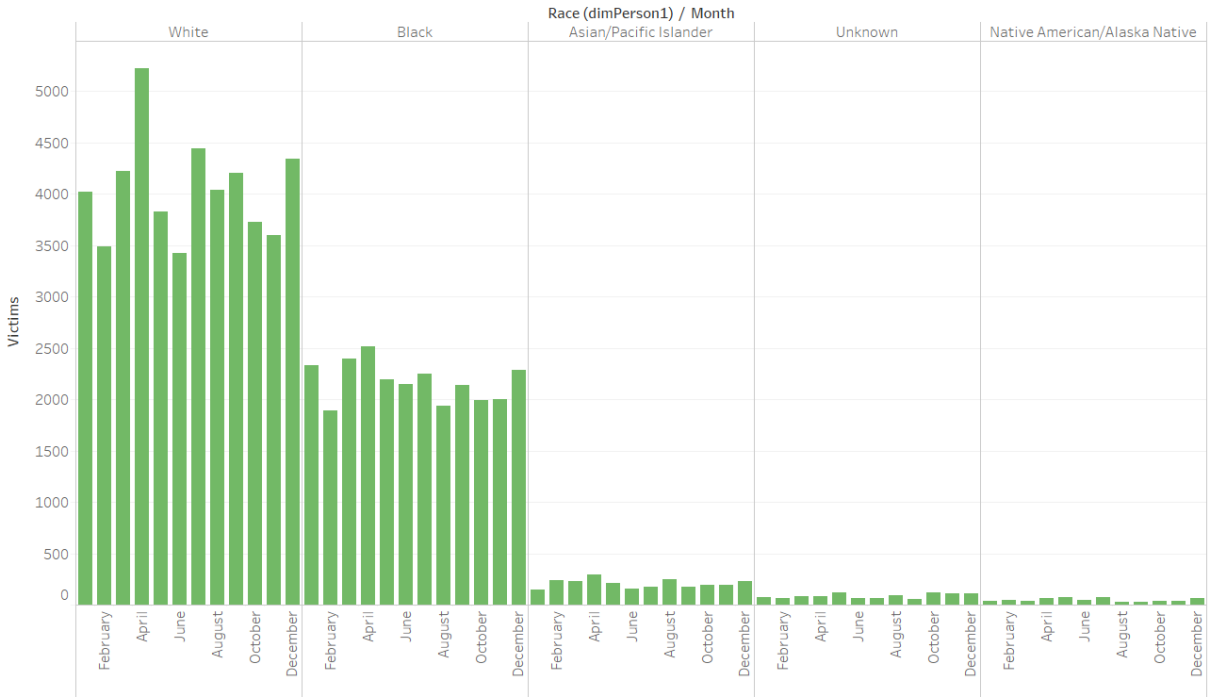
Rasa ofiary według każdego roku:



Rys. 6

Z wykresu robimy wniosek, że najczęściej są ofiarami ludzi białej rasy, a niż czarnoskóre.

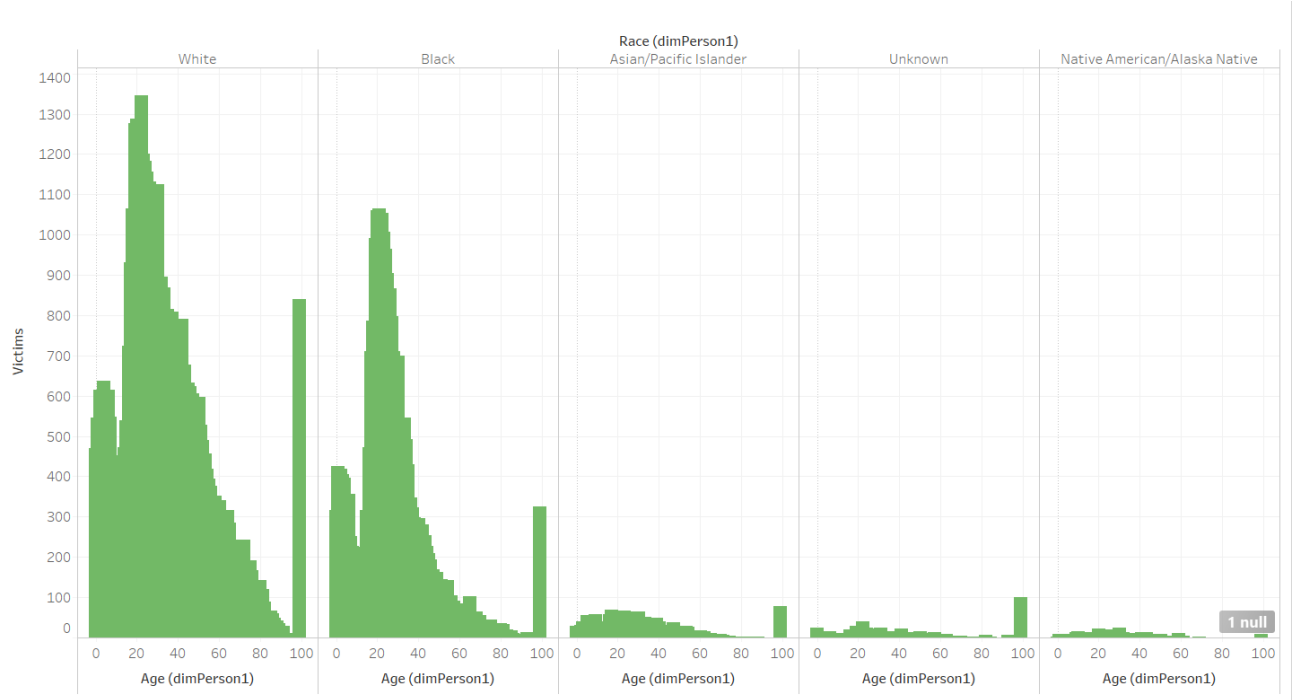
Rasa ofiary według miesiąca:



Rys. 7

Jak widać z wykresu, liczba zabójstw ludzi różnych ras nie bardzo się liczy. Dlatego można zrobić wniosek, że rasa człowieka nie bardzo wpływa na ilość zabójstw.

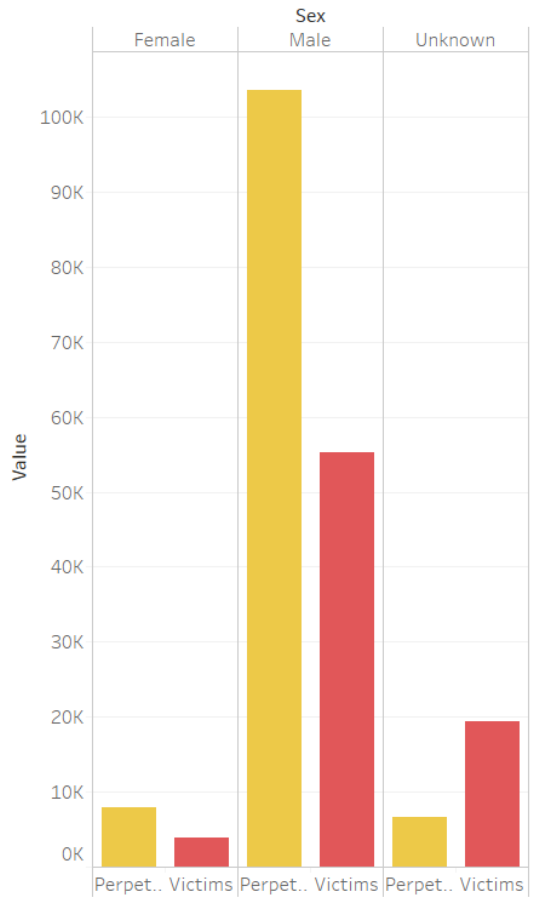
Wiek ofiary według rasy:



Rys. 8

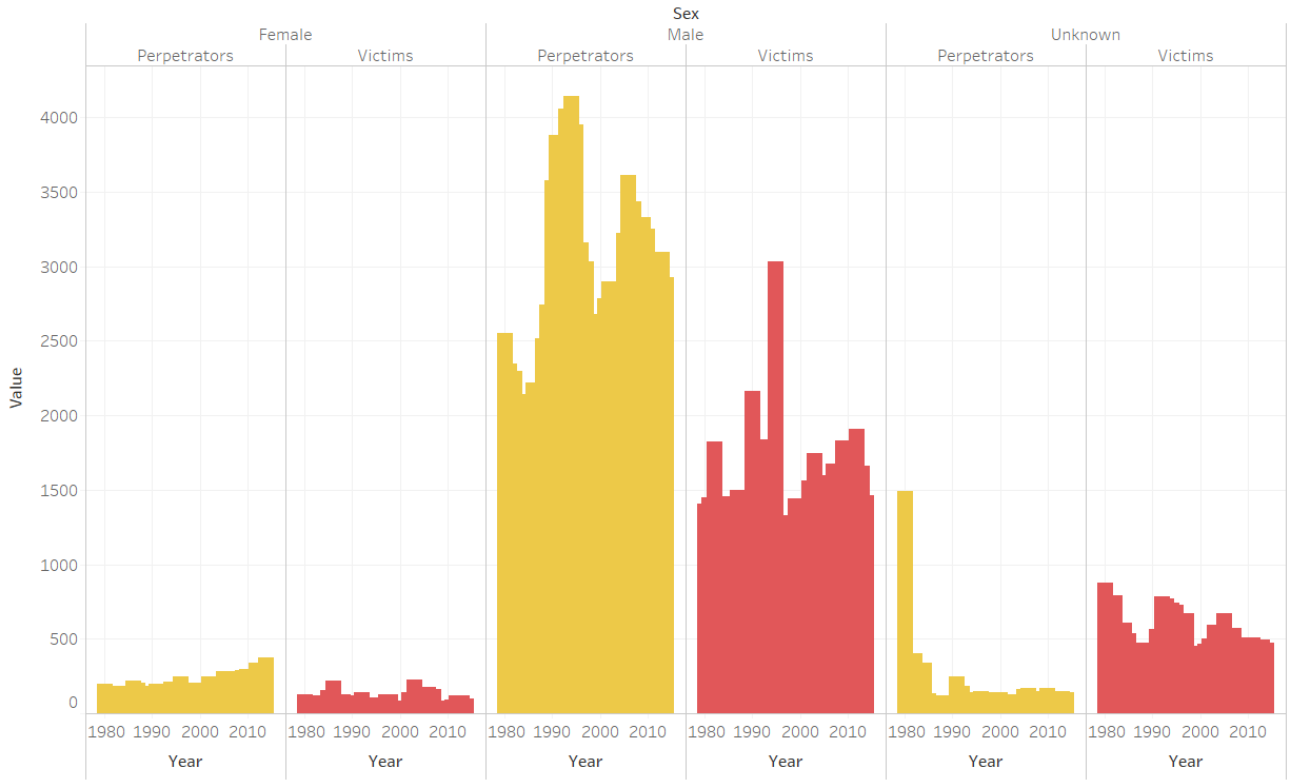
Najwięcej ofiar są z wiekiem 22 lata. Najbardziej zabójstw przypada na wiek 19-24 lata. Także duża cześć zabójstw przypada na wiek 99 lat niezależnie od rasy człowieka.

Płeć osoby (ofiary oraz przestępcy):



Rys. 9

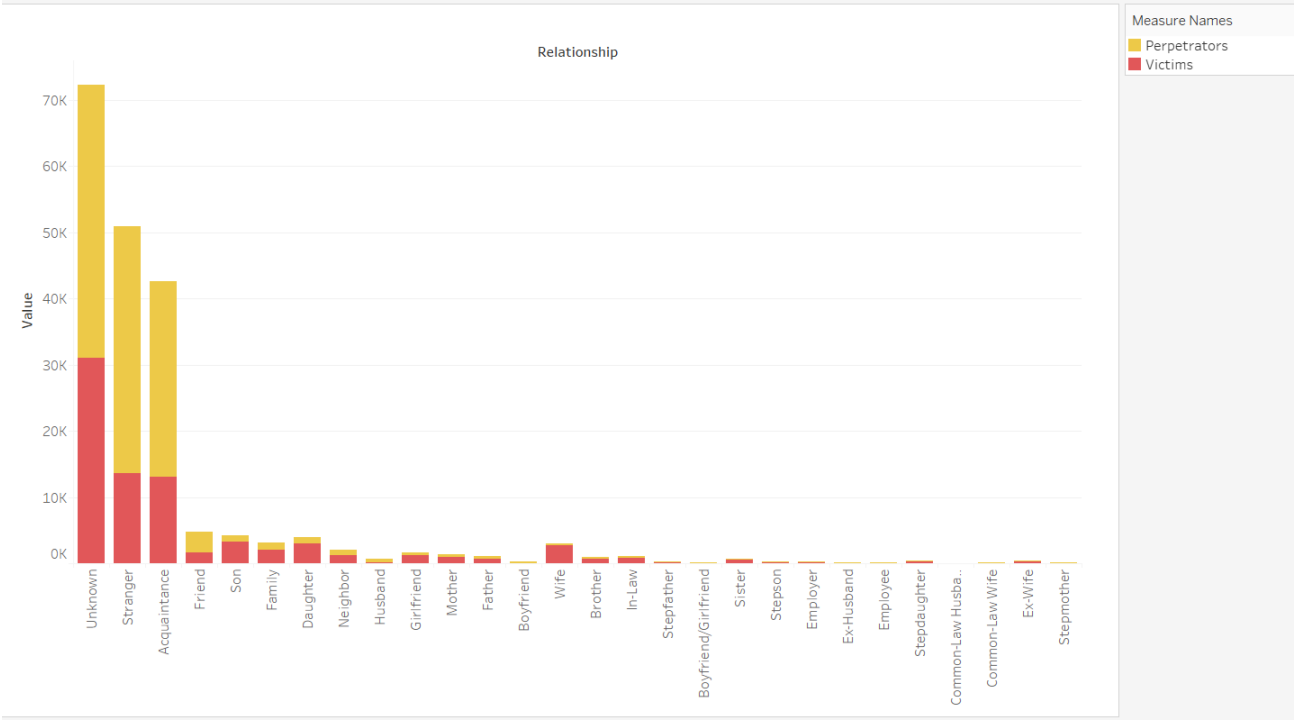
Wnioskujemy, że przestępców mężczyźni jest więcej niż kobiet. Zobaczmy co się dzieje dla każdego roku:



Rys. 10

Widzimy, że ilość kobiet ofiar rośnie od 1990 roku.

Relacja między ofiarą a przestępcą:

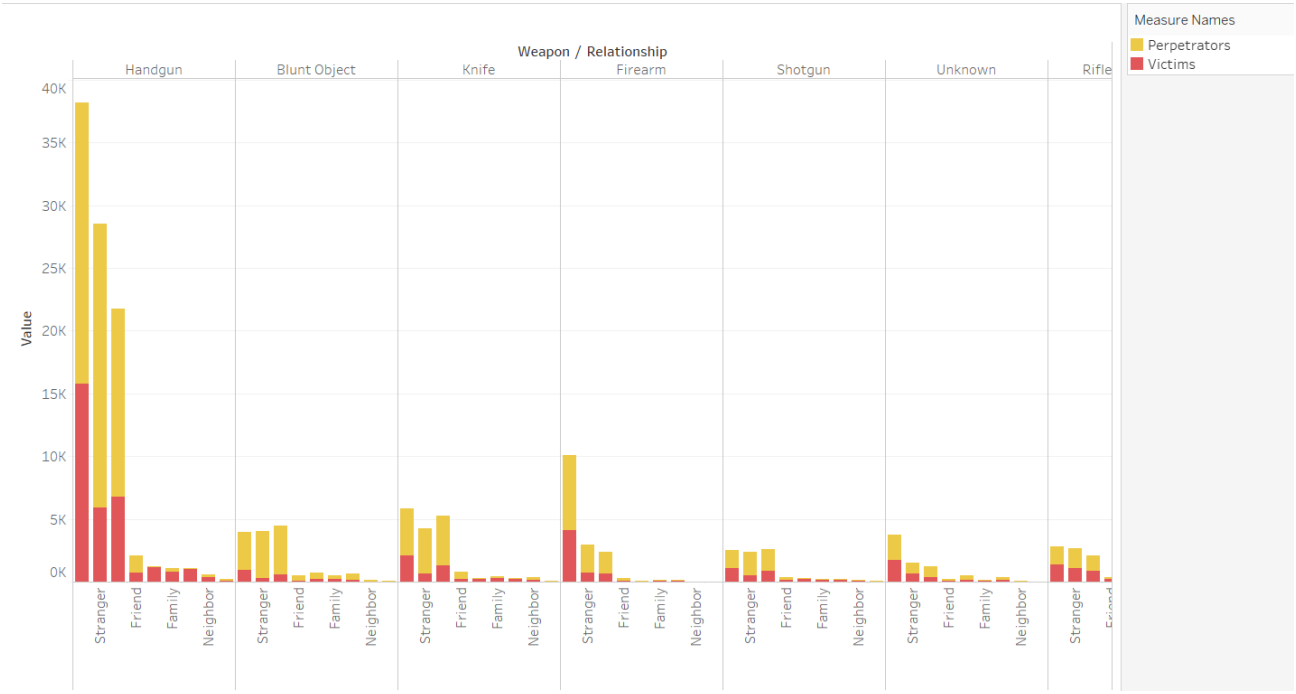


Rys. 11

Niestety duża cześć danych jest stracona, przez to, że nie wszystkie wypadki da się zapisać. I nie wszystkich przestępców da się znaleźć policjantom. Ale uwzględniając te dane, którymi dysponujemy możemy zrobić wniosek, że duża cześć (a nawet połowa) to są przestępcy które mają jakikolwiek związek z ofiarą. To znaczy, że ofiara znała przestępcę.

Także widać, że dość dużą częścią przestępców są bliska rodzina (syn, córka, żona i tak dalej). Jest to bardzo ważna informacja dla policji stanów.

Relacja między osobami oraz użyta broń:



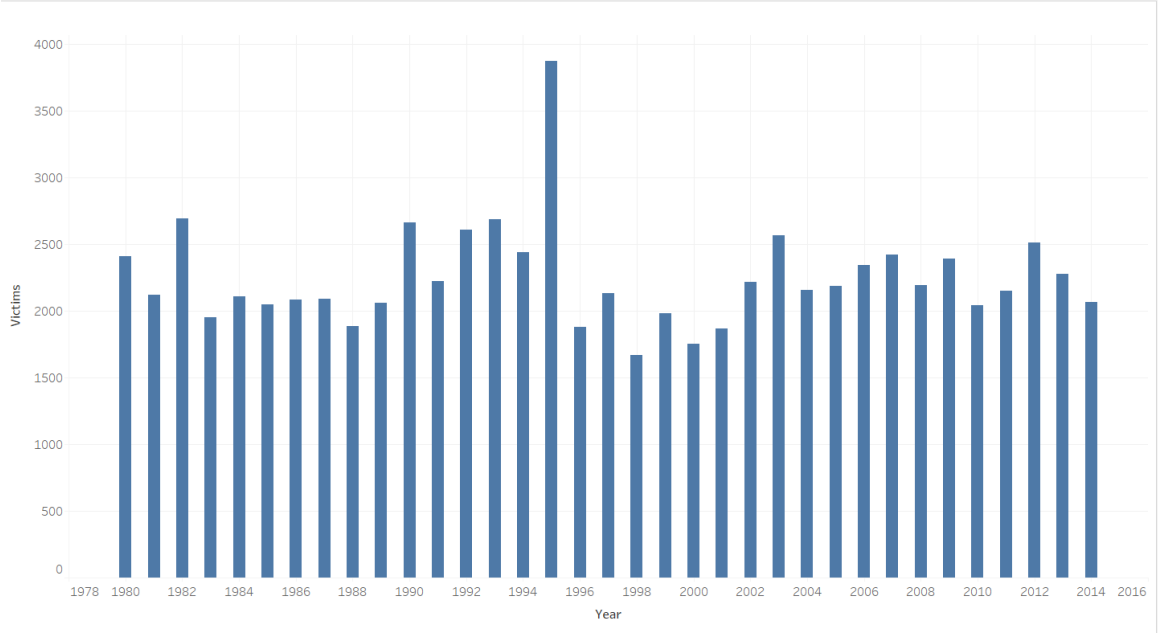
Rys. 12

Najwięcej zabójstw z relacja rodzinną jest z użyciem pistoletu oraz tępego przedmiotu.

Podsumowanie

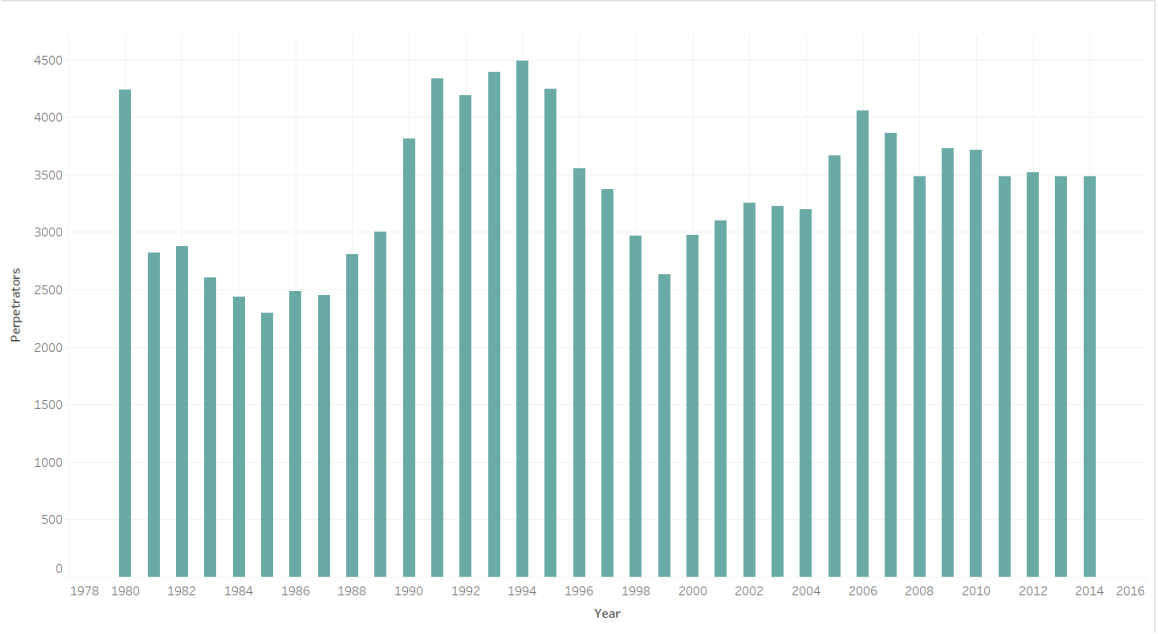
Otóż tak. Przejdziemy po każdemu punktu, który mieliśmy zbadać:

1. Sumaryczna liczba zabójstw każdego roku



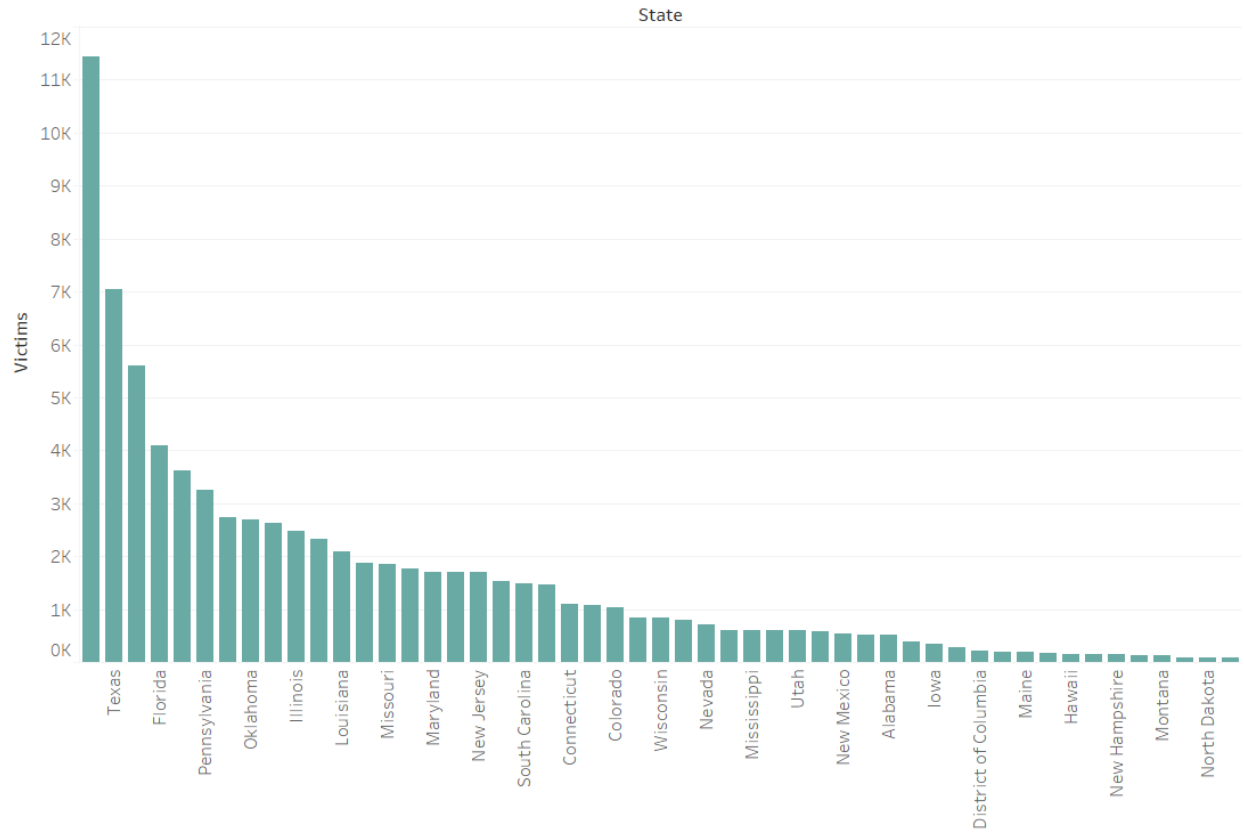
Najwięcej zabójstw było w 1995 roku.

2. Sumaryczna liczba zabójców każdego roku



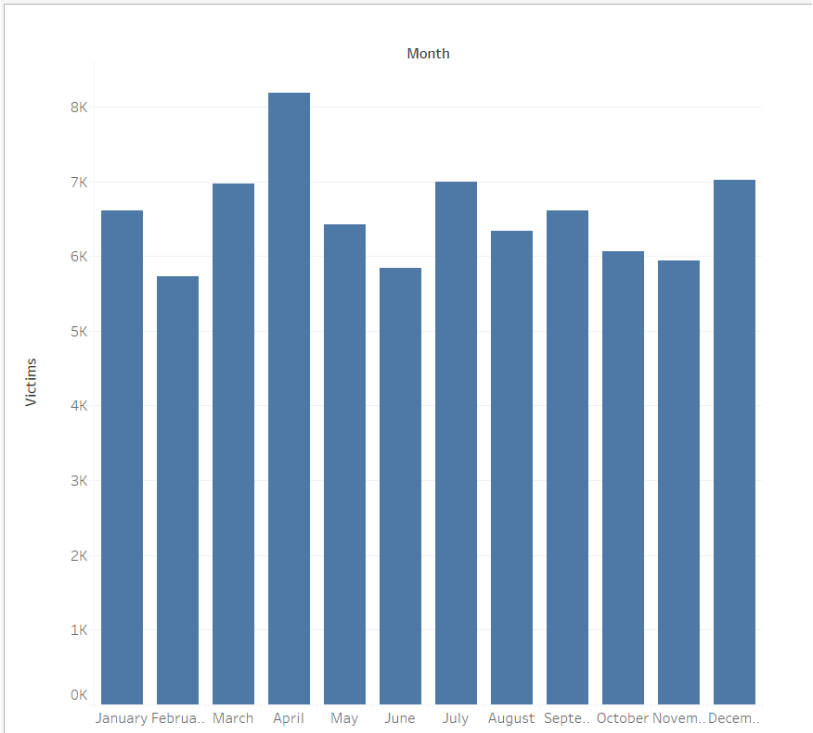
Najwięcej zabójców było w 1994 roku (4 488). Natomiast najmniej 1985 (2 297).

3. Miasta oraz stany z największym wskaźnikiem przestępstw



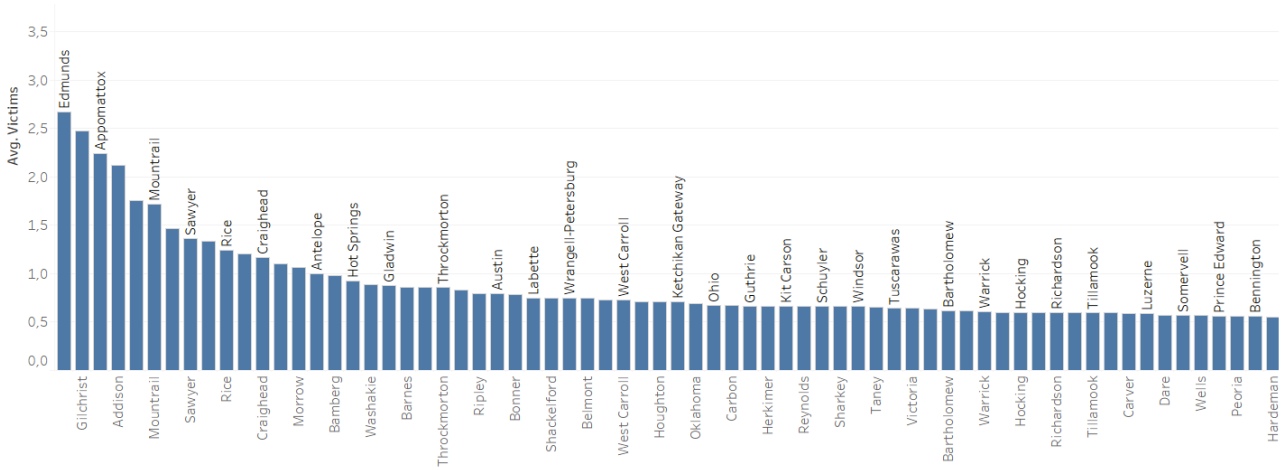
Robię wniosek, że bardziej południowe stany mają większe wskaźniki prześtepstwa.

4. Miesiące z największą liczbą zabójstw

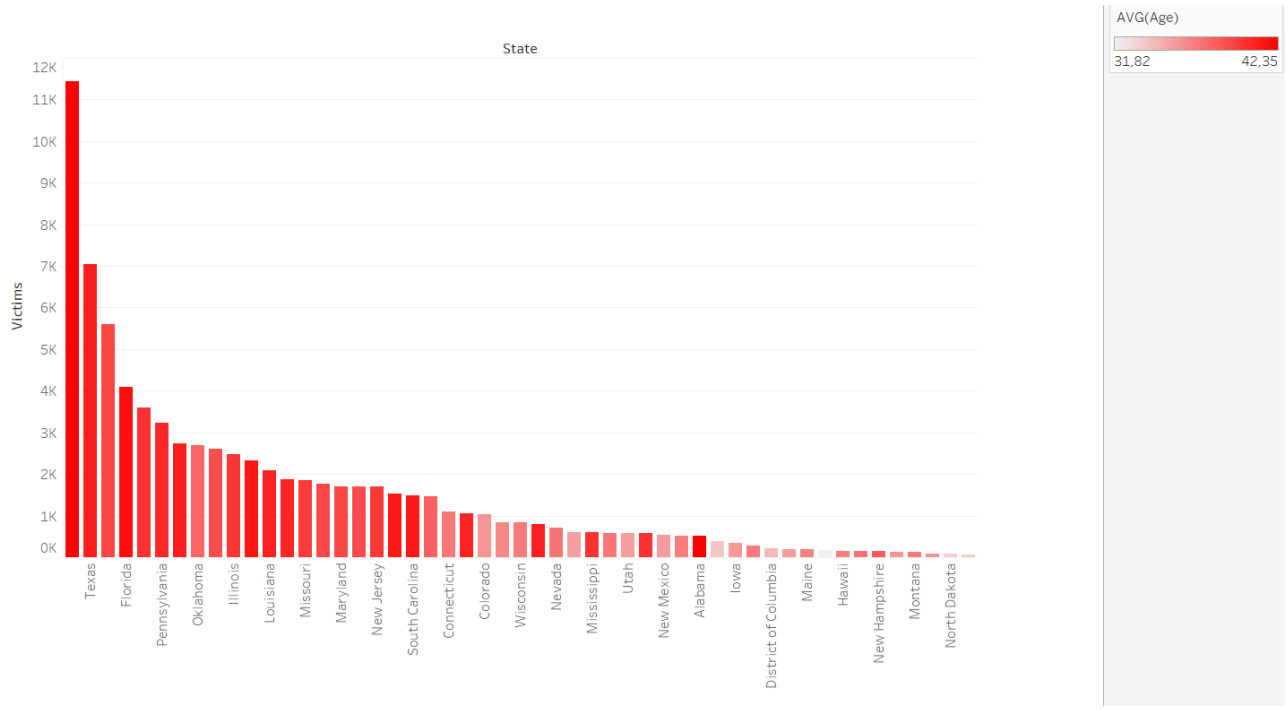


Jest to Kwiecień z ilością ofiar 8184 osoby.

5. Średnia liczba ofiar według każdego miasta



Na rysunku pokazałem kilka miast, z największą wartością średnią.



Rys. 14

Średni wiek według stanów

Wychodząc z otrzymanych informacji możemy wnioskować, że:

- w latach 1980 – 2014 rasa człowieka nie miała wielkiego znaczenia
- liczba zabójstwa zależy od lokalizacji geograficznej stanu
- miasta z większą populacją będą rzeczywiście mieli większą liczbę zabójstw
- najczęściej używana broń to pistolet

- zabójstwa z relacją rodzinną najczęściej są wykonane z użyciem tępego przedmiotu, co znaczy, że dużo rodzin potrzebują pomocy psychologicznej
- kobiety są częściej ofiarami niż mężczyźni
- mężczyźni są w znacznej części przestępcami niż kobiety
- średni wiek według stanów jest 39-42 lata, natomiast najwięcej zabójstw przypada na wiek 19-22 lata

8. Wnioski końcowe z realizacji projektu

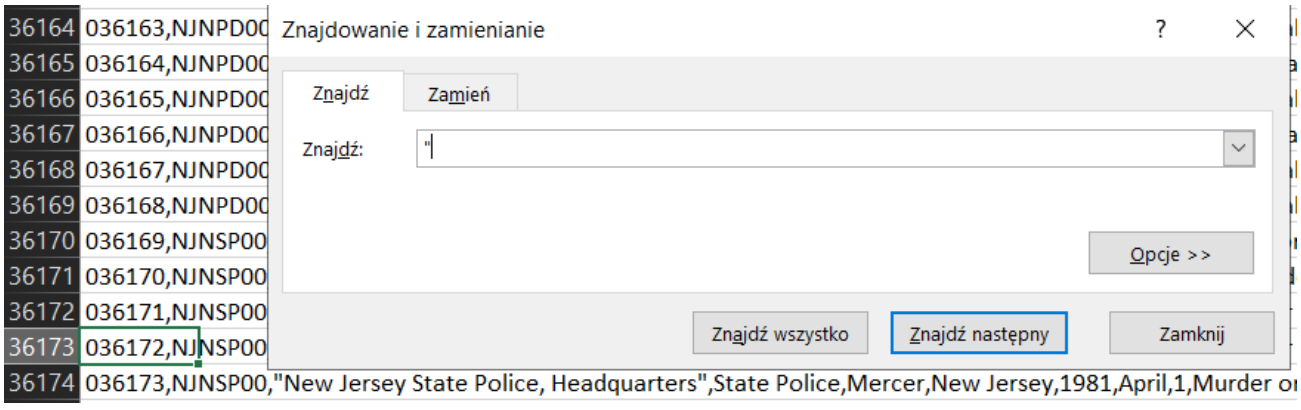
8.1. Napotkane problemy i sposoby ich rozwiązania

W samym początku, gdy tylko zacząłem pracować z plikiem źródłowym, który jest w formacie CSV, napotkałem taki problem, że niektóre rekordy są „zepsute” (rys. 1)

Rys. 1

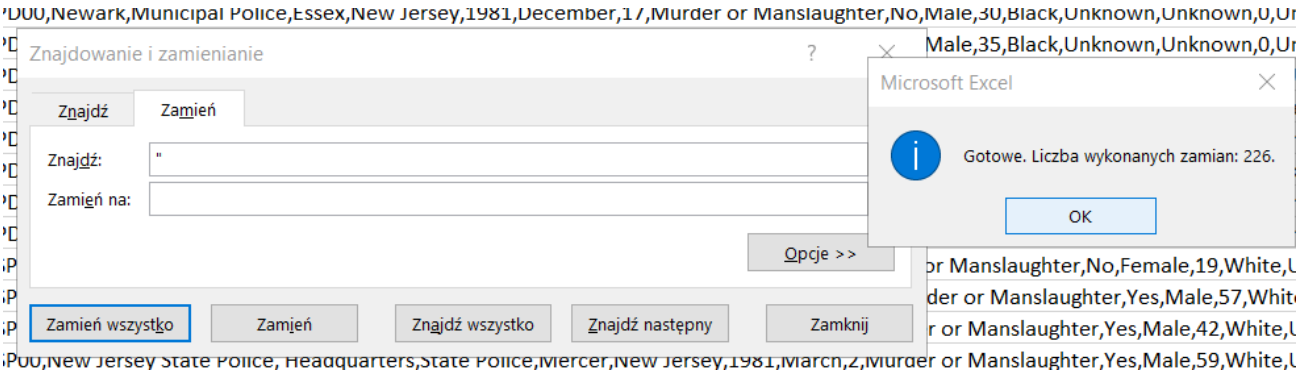
Czyli, jak widzimy, nazwa agencji jest zapisana w cudzysłowach, a nie jak zwykle bez nich. Spowodowało to problem, że SSIS odczytywał nie w prawidłowy sposób i wszystkie kolejne kolumny zostały zmieszczone (np. w takim przypadku kolumna „Wiek” mogła zawierać wartość nie podobną na liczbę, co jest niedopuszczalne).

Zdecydowałem usunąć wszystkie cudzysłowy, najpierw sprawdzając czy jakaś kolumna nie zawiera rekordów, gdzie są one konieczne (wyszukiwałem przez funkcję „Znajdź” w Excel rys. 2).



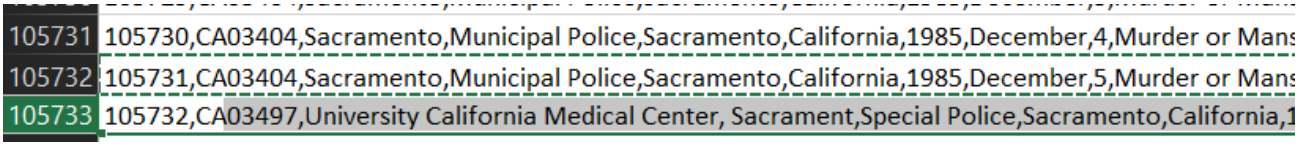
Rys. 2

Następnie przez funkcję „Zamień” znalazłem wszystkie takie miejsca i usunąłem ich Rys. 3.



Rys. 3

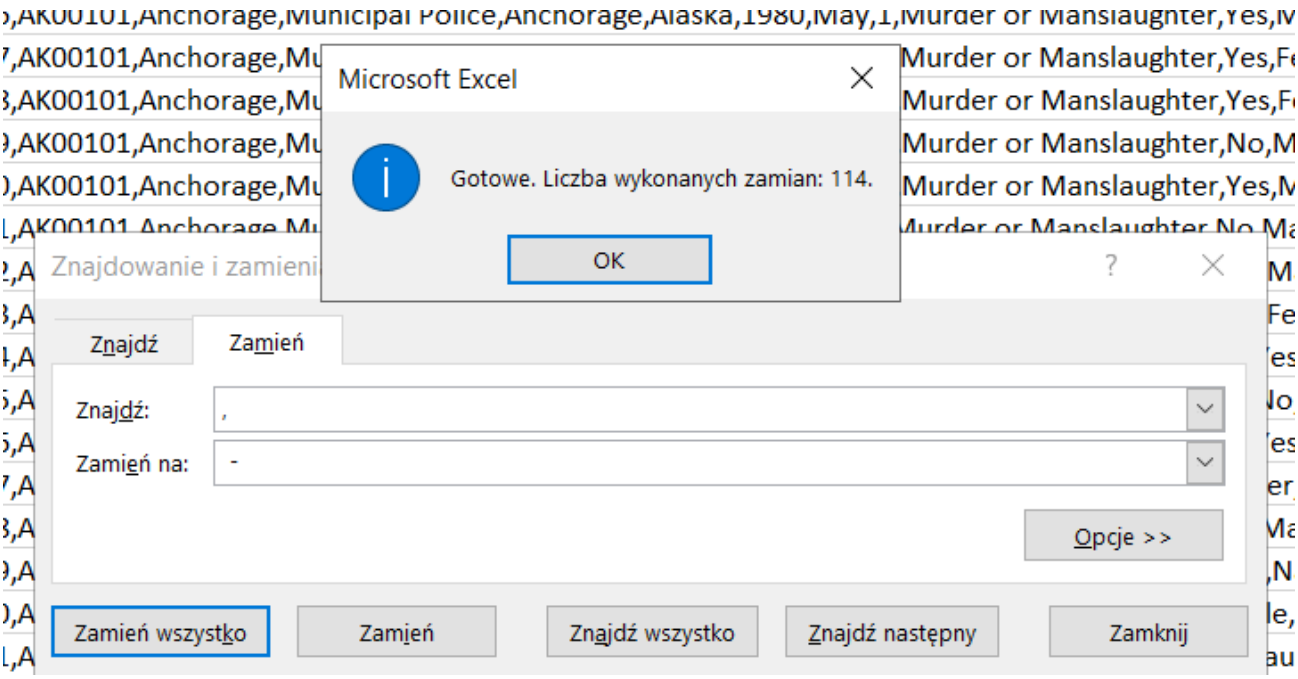
Następnym problemem, który napotkałem, były wartości zawierające przecinki (rys. 4)



Rys. 4

Jak widać, rzeczywista nazwa agencji jest taka: „University California Medical Center, Sacrament”, natomiast SSIS czyta to jak oddzielną kolumnę, bo używamy plik formatu CSV (separacja po przecinkach).

Dlatego, żeby poprawić to, możemy zamienić „, „ na „ – „ (rys. 5)



Rys 5.

8.2. Pozyskana wiedza i doświadczenie

Osobiście dla siebie pozyskałem wiedzę o jeszcze jednym dobrym sposobie przetwarzania i hurtowania danych, które pomogą mi lepiej analizować i przewidywać dane w przyszłości. Jest to niezbędna umiejętność dla Machine Learningu, którym marzę zająć się.

Także dobrym narzędziem jest program Tableau, który pomógł mi z wizualizacją analizy mojej bazy danych.