

## Hurtownie danych – Projekt

PWr. Data: 10.06.2022

Student	-----	Ocena
Indeks	<u>256480</u>	
Imię	<u>Ernest</u>	
Nazwisko	<u>Przybył</u>	

### 1. Tytuł projektu

**Cele ataków terrorystycznych w latach 1970 – 1992, 1994 – 2017**

### 2. Charakterystyka dziedziny problemowej

Czym jest terroryzm? Naproście rzecz ujmując: Groźba lub użycie nielegalnych środków przemocy przez podmiot niepaństwowy w celu osiągnięcia celu politycznego, gospodarczego, religijnego lub społecznego, poprzez strach przymus lub zastraszenie.

Każdy atak terrorystyczny wiąże się z poważnymi konsekwencjami społecznymi i materialnymi. Zmniejszają one poczucie bezpieczeństwa wśród obywateli, co przekłada się na zmniejszone zaufanie do rządu i zwiększa starcia pomiędzy grupami społecznymi. Straty materialne są kosztowne i wiążą się z większymi kosztami ubezpieczeń, a śmierć ludzi jest nieocenioną stratą.

#### 2.1 Opis obszaru analizy (wybrany fragment dziedziny, przeznaczony do szczegółowej analizy i opracowania hurtowni danych)

Przedmiotem analizy będą ataki terrorystyczne przeprowadzane na całym świecie w latach 1970 – 1992, 1994 – 2017. Będziemy je analizować pod względem celów, co pozwoli nam lepiej zabezpieczyć szczególnie narażone cele i zrozumieć powody ich wyboru. Z kolei na podstawie analizy narodowości i miejsc w których dochodzi do ataków pozwoli skutecznie określić ryzyko związane z atakami w określonych miejscach oraz określi „grupy podwyższonego” ryzyka dla określonych narodowości. Analiza broni użytej do ataków pomoże określić czym najczęściej posługują się terroryści, co pomoże w badaniu źródeł dostawy tej broni odcięciu jej dopływu.

## 2.2 Problemy

P01 – Rosnący brak zaufania do rządu

P02 – Rosnący brak poczucia bezpieczeństwa obywateli

P03 – Narastanie konfliktów pomiędzy krajami/grupami etnicznymi

P04 – Straty pieniężne i wizerunkowe

## 2.3 Cel przedsięwzięcia

### 2.3.1 Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji (pytania badawcze)

Analiza powinna dostarczyć informacji na temat ataków, takich jak:

- Czy są miejsca/cele bardziej narażone na atak terrorystyczny?
- Jak liczba ataków terrorystycznych zmienia się w czasie?
- Czy są okresy czasu o zwiększonym ryzyku ataku w danym miejscu?
- Jaka broń jest najbardziej śmiertelna?
- Jak aktywność grup terrorystycznych zmieniała się na przestrzeni lat?
- Jakimi atakami kończą się powodzeniem terrorystów?

Analiza odpowiedzi na te pytania pozwoli skuteczniej przygotowywać się na ataki w okresach czasu i miejscach szczególnie narażonych. Pozwoli także określić czynniki które zwiększają skuteczność terrorystów w atakach, a także powodują większe straty wśród ludności.

### 2.3.2 Zakres analizy – badane aspekty

Analizie będą poddawane typ ataku, data i miejsce ataku, cel ataku, liczba ofiar i strat, użyta broń, podłoża ataku oraz inne określone dalej w dokumencie.

### 2.3.3 Potencjalni użytkownicy projektowanej hurtowni danych

- Organizacje rządowe krajów o podwyższonym ryzyku ataków terrorystycznych
- Media telewizyjne i internetowe

### 3. Dane źródłowe

#### 3.1. Źródła danych

Tabela 1. Zbiory danych źródłowych:

4. Lp.	Plik, bazy danych	Typ	Liczba rek.	Rozmiar [MB]	Opis
1	globalterrorismdb_0718dist.csv	.csv	~182_000	159	Baza zawierająca dane o atakach terrorystycznych na świecie. 1970-2017

#### 4.1. Lokalizacja, dostępność danych źródłowych

Dane są dostępne na stronie kaggle(<https://www.kaggle.com/datasets/START-UMD/gtd>) lub na stronie organizacji. Baza jest prowadzona przez naukowców z organizacji START z siedzibą na Uniwersytecie Maryland. Dane są dostępne w formacie csv w jednym pliku. Dodatkowo są nam dostarczone szczegółowe informacje o znaczeniu poszczególnych pól w dokumencie pod linkiem (<https://start.umd.edu/gtd/downloads/Codebook.pdf>)

#### 4.2. Słownik danych – interpretacja

Uwaga z powodu bardzo dużej ilości kolumn (135) ograniczę się tylko do kolumn które przeszły pozytywnie wstępny etap oceny jakości danych.

Dogłębnie wytłumaczone znaczenie atrybutów pod linkiem (<https://start.umd.edu/gtd/downloads/Codebook.pdf>)

Tabela 2. Słownik danych:

Lp.	Nazwa kolumny	Typ	Znaczenie	Uwagi
1.	attacktype1	Number	Określa identyfikator rodzaju ataku	Klucz do pola attacktype1_txt
2.	attacktype1_txt	String	Opisowa wartość rodzaju ataku	
3.	city	String	Nazwa miasta/wioski w której doszło do zdarzenia	
4.	corp1	String	Określa korporację która była celem ataku. Jeśli celem nie	

			byłą korporacja NULL lub nie ma danych..	
5.	country	Number	Kod kraju	Klucz do pola country_txt
6.	country_txt	String	Nazwa kraju o podanym identyfikatorze	
7.	crit1	Number	Określa czy jest to atak o podanym podłożu.	Criterion 1: POLITICAL, ECONOMIC, RELIGIOUS, OR SOCIAL GOAL (CRIT1)
8.	crit2	Number	Określa czy jest to atak o podanym podłożu.	INTENTION TO COERCE, INTIMIDATE OR PUBLICIZE TO LARGER AUDIENCE(S) (CRIT2)
9.	crit3	Number	Określa czy jest to atak o podanym podłożu.	Criterion 3: OUTSIDE INTERNATIONAL HUMANITARIAN LAW (CRIT3)
10.	eventid	Number	12-znakowy identyfikator. 8 pierwszych znaków – zawiera datę wydarzenia w formacie “yyyymmdd”. ostatnie 4 znaki są liczbą porządkową.	
11.	gname	String	Nazwa grupy terrorystycznej odpowiedzialnej za atak.	
12.	guncertain1	Number	Określa czy dane dotyczące grupy terrorystycznej są przypuszczeniem (1) lub są dowiedzione (0)	
13.	iday	Number	Numer dnia miesiąca w którym doszło do zdarzenia	
14.	imonth	Number	Numer miesiąca w którym doszło do zdarzenia	
15.	individual	Number	Określa czy incydent był przeprowadzony przez jedną osobę.	
16.	ishostkid	Number	Określa czy porwano zakładników	1 = "Yes" 0 = "No" -9 = "Unknown"
17.	iyear	Number [1970; 2017]	Rok zdarzenia	
18.	latitude	Number	Szerokość geograficzna	
19.	longitude	Number	Długość geograficzna	
20.	natlty1	Number	Narodowość pierwszorzędnego celu	Klucz do wartości natlty1_txt

21.	natlty1_txt	String	Narodowość pierwszorzędного celu opisowo	
22.	nkill	Number	Całkowita liczba potwierdzonych zgonów w wyniku ataku.	
23.	nkillter	Number	Liczba ofiar w śród sprawców	
24.	nwound	Number	Liczba rannych	
25.	nwoundte	Number	Liczba rannych terrorystów.	
26.	property	Number	Czy uszkodzono czyjąś własność?	
27.	propextent	Number	Określa stopień uszkodzeń własności. Jeśli własność była celem ataku.	
28.	propvalue	Number	Określa wartość uszkodzonej własności w USD. Jeśli taka została uszkodzona.	
29.	provstate	String	Nazwa regionu administracyjnego pierwszego rzędu	
30.	ransom	Number	Określa czy żądano okupu	
31.	region	Number	Numer regionu	Klucz to pola region_txt
32.	region_txt	String	Nazwa regionu geograficznego.	
33.	success	Number	Określa czy zamach osiągnął swoje cele bezpośrednio. Raczej coś w stylu bomba wybuchła, cel został zabity -> sukces. Bomba rozbrojona -> porażka. Niezależnie czy udało się osiągnąć pośrednie cele np. zastraszyć danych ludzi.	
34.	suicide	Number	Określa czy był to zamach samobójczy.	
35.	target1	String	Określa pierwszorzędny cel ataku.	
36.	targsubtype1	Number	Określa typ pierwszorzędного celu ataku	Klucz do wartości opisowej typu targsubtype1_txt
37.	targsubtype1_txt	String	Określa typ pierwszorzędного celu ataku	

38.	targtype1	Number	Określa typ celu ataku.	Klucz do wartości opisowej typu targtype1_txt
39.	targtype1_txt	String	Opisuje typ celu ataku.	
40.	weapsubtype1	Number	Określa pod rodzaj pierwszorzędnej broni użytej w ataku.	Klucz do wartości opisowej typu weapsubtype1_txt
41.	weapsubtype1_txt	String	Opisuje pod rodzaj pierwszorzędnej broni użytej w ataku.	
42.	weaptype1	Number	Określa rodzaj użytej broni	Klucz do pola opisowego weaptype1_txt
43.	weaptype1_txt	String	Określa rodzaj użytej broni	

#### 4.3. Ocena jakościowa danych:

Uwaga z powodu bardzo dużej ilości kolumn (135) ograniczę się tylko do kolumn które przeszły pozytywnie wstępny etap oceny jakości danych.

Większość rekordów posiadała kilka pól o tej samej nazwie w określonej hierarchii np. weaptype1, weaptype2, weaptype3... Określający odpowiednio pierwszorzędą broń terrorystów drugorzędą itd. Ponieważ poza pierwszorzędnymi polami w hierarchii większość z nich ma ilość NULL przekraczającą 70% korzystam wyłącznie z pierwszorzędnych pól w hierarchii.

Nie brałem też pod uwagę pól które posiadały ilość NULL większą od 90%.

Tabela 3. Ocena jakości danych:

Plik: globalterrorismdb_0718dist.csv					
Lp.	Nazwa kolumny	Typ	Zakres danych	Znaczenie	Uwagi - ocena jakości
44.	attacktype1	Number	[1, 9]	Określa identyfikator rodzaju ataku	0% NULL Nieprzydatne. Nie potrzebujemy w hurtowni sztucznego klucza wyznaczającego attacktype1_txt
45.	attacktype1_txt	String	Assassination Hijacking	Opisowa wartość rodzaju ataku	0% NULL Przydatne.

			Kidnapping Barricade Incident Bombing/Explosion Armed Assault Unarmed Assault Facility/Infrastructure Attack Unknown		
46.	city	String		Nazwa miasta/wioski w której doszło do zdarzenia	1% NULL Przydatne
47.	corp1	String		Określa korporację która była celem ataku. Jeśli celem nie była korporacja NULL lub nie ma danych.	23% NULL Nieprzydatne. Większość korporacji była celem tylko kilku ataków, a nieliczne był celem ponad 100 razy.
48.	country	Number		Kod kraju	0% NULL Nieprzydatne. Nie potrzebujemy w hurtowni sztucznego klucza wyznaczającego country_txt
49.	country_txt	String		Nazwa kraju o podanym identyfikatorze	0% NULL Przydatne
50.	crit1	Number	[0; 1] (TAK/NIE)	Określa czy jest to atak o podanym podłożu. Criterion 1: POLITICAL, ECONOMIC, RELIGIOUS, OR SOCIAL GOAL (CRIT1)	0% NULL Nieprzydatne. Tylko 560 rekordów (1%<) nie spełnia tego kryterium.
51.	crit2	Number	[0; 1] (TAK/NIE)	Określa czy jest to atak o podanym podłożu. INTENTION TO COERCE, INTIMIDATE OR PUBLICIZE TO LARGER AUDIENCE(S) (CRIT2)	0% NULL Nieprzydatne. Tylko 495 rekordów (1%<) nie spełnia tego kryterium.

52.	crit3	Number	[0; 1] (TAK/NIE)	Określa czy jest to atak o podanym podłożu. criterion 3: OUTSIDE INTERNATIONAL HUMANITARIAN LAW (CRIT3)	0% NULL Nieprzydatne w kontekście naszej analizy.
53.	eventid	Number		12-znakowy identyfikator. 8 pierwszych znaków – zawiera datę wydarzenia w formacie “yyyymmdd”. ostatnie 4 znaki są liczbą porządkową.	0% NULL Przydatne. Pomoże nam w unikaniu wstawiania duplikatów zdarzeń do hurtowni danych.
54.	gname	String		Nazwa grupy terrorystycznej odpowiedzialnej za atak.	0% NULL Przydatne.
55.	guncertain1	Number	[0; 1] (TAK/NIE)	Określa czy dane dotyczące grupy terrorystycznej są przypuszczeniem (1) lub są dowiedzione (0)	1%< NULL Przydatne. Kilka rekordów nie posiada wartości możemy je bezpiecznie usunąć.
56.	iday	Number	[0; 31]	Numer dnia miesiąca w którym doszło do zdarzenia	0% NULL Przydatne. Nie trzeba ręcznie generować pola.
57.	imonth	Number	[0; 12]	Numer miesiąca w którym doszło do zdarzenia	0% NULL Przydatne. Nie trzeba ręcznie generować pola.
58.	individual	Number	[0; 1] (TAK/NIE)	Określa czy incydent był przeprowadzony przez jedną osobę.	0% NULL Przydatne.
59.	ishostkid	Number	[0; 1] (TAK/NIE)	Określa czy porwano zakładników	0% NULL Przydatne.



60.	iyear	Number [1970; 2017]		Rok zdarzenia	0% NULL Przydatne. Nie trzeba ręcznie generować pola.
61.	latitude	Number	[-53,15; 74,63]	Szerokość geograficzna	2.5% NULL Przydatne. Nie występują, żadne anomalie, a rekordy z brakującymi danymi można zignorować.
62.	longitude	Number	[-86185896; 179,37]	Długość geograficzna	2.5% NULL Przydatne. Występują błędy wartości. Najwyraźniej ktoś zapomniał dodać przecinka. Rekordy z brakującymi danymi można zignorować.
63.	natlty1	Number	[4; 1004]	Narodowość pierwszorzędnego celu. Klucz do wartości natlty1_txt	1%< NULL Nieprzydatne. Niepotrzebny sztuczny klucz.
64.	natlty1_txt	String		Narodowość pierwszorzędnego celu opisowo	1%< NULL Przydatne.
65.	nkill	Number	[0; 1570]	Całkowita liczba potwierdzonych zgonów w wyniku ataku.	6% NULL Przydatne. Ponieważ ilość zabitych będzie przedmiotem analizy możemy zignorować wartości null.
66.	nkillter	Number	[0; 500]	Liczba ofiar w śród sprawców	36% NULL Przydatne. Może posłużyć jako dodatkowa informacja do wartości nkill. Duże braki

					ograniczą zbiór danych dla analiz.
67.	nwound	Number	[0; 8191]	Liczba rannych	9% NULL Przydatne. Ponieważ ilość rannych będzie przedmiotem analizy możemy zignorować wartości null, choć stracimy prawie 10% danych.
68.	nwoundte	Number	[0; 200]	Liczba rannych terrorystów.	28% NULL Przydatne. Może posłużyć jako dodatkowa informacja do wartości nwound. Duże braki ograniczą zbiór danych dla analiz.
69.	property	Number	0 = TAK 1= NIE - 9=brak danych	Czy uszkodzono czyjaś własność?	0% NULL Niepotrzebne w analizie.
70.	propextent	Number	1="katastrofa" 2=znaczące uszkodzenia 3=niewielkie uszkodzenia 4=nieznane	Określa stopień uszkodzeń własności. Jeśli własność była celem ataku.	65% NULL Nieprzydatne. Relatywna skala uszkodzeń nie będzie nas interesowała.
71.	propvalue	Number	-99=brak danych [0; 2700000000]	Określa wartość uszkodzonej własności w USD. Jeśli taka została uszkodzona.	78% NULL Duże braki w danych. Jednak pole może być przydatne w analizie.
72.	provstate	String		Nazwa regionu administracyjnego o pierwszego rzędu	1%< NULL Przydatne.
73.	ransom	Number	0 = TAK 1= NIE - 9=brak danych NULL=nie dotyczy	Określa czy żądano okupu	57% NULL Przydatne. Duże braki w danych. Jednak może pomóc w analizie danych.

74.	region	Number		Numer regionu. Klucz to pola region_txt	0% NULL Nieprzydatne nie potrzeba dodatkowego sztucznego klucza.
75.	region_txt	String		Nazwa regionu geograficznego.	0% NULL Przydatne
76.	success	Number	[0; 1] (TAK/NIE)	Określa czy zamach osiągnął swoje cele bezpośrednie. Raczej coś w stylu bomba wybuchła, cel został zabity -> sukces. Bomba rozbrojona -> porażka. Niezależnie czy udało się osiągnąć pośrednie cele np. zastraszyć danych ludzi.	0% NULL Przydatne
77.	suicide	Number	[0; 1] (TAK/NIE)	Określa czy był to zamach samobójczy.	0% NULL Przydatne
78.	target1	String		Określa pierwszorzędny cel ataku.	1< % NULL Nieprzydatne. Większość nazw jest unikalna i nie posiada potencjału do agregacji.
79.	targsubtype1	Number	[1; 113]	Określa typ pierwszorzędnego celu ataku. Klucz do wartości opisowej typu targsubtype1_txt	1< % NULL Nieprzydatne nie potrzeba dodatkowego sztucznego klucza.
80.	targsubtype1_txt	String		Określa typ pierwszorzędnego celu ataku	1< % NULL Przydatne. Pojedyncze rekordy z brakami. Można zignorować błędne rekordy.

81.	targtype1	Number	[1; 22]	Określa typ celu ataku. Klucz do wartości opisowej typu targtype1_txt	0 % NULL Nieprzydatne nie potrzeba dodatkowego sztucznego klucza.
82.	targtype1_txt	String		Opisuje typ celu ataku.	0 % NULL Przydatne.
83.	weapsubtype1	Number	[1; 31]	Określa pod rodzaj pierwszorzędnej broni użytej w ataku. Klucz do wartości opisowej typu weapsubtype1_txt	11 % NULL Nieprzydatne nie potrzeba dodatkowego sztucznego klucza.
84.	weapsubtype1_txt	String		Opisuje pod rodzaj pierwszorzędnej broni użytej w ataku.	11 % NULL Przydatne. Nie wszystkie typy bron mają określony podtyp.
85.	weaptype1	Number	[1; 13]	Określa rodzaj użytej broni. Klucz do pola opisowego weaptype1_txt	0 % NULL Nieprzydatne nie potrzeba dodatkowego sztucznego klucza.
86.	weaptype1_txt	String		Określa rodzaj użytej broni	0 % NULL Przydatne.

## 5. Analityczne modele wielowymiarowe

### 5.1. Fakty podlegające analizie oraz ich miary

Tabela 5. Fakty oraz ich miary opracowywanych modeli analitycznych

Lp.	Fakt	Miary
1.	Atak terrorystyczny	liczba ofiar i rannych, czy terroryści osiągnęli cel/e, wartość uszkodzeń,

## 5.2. Kontekst analizy faktów

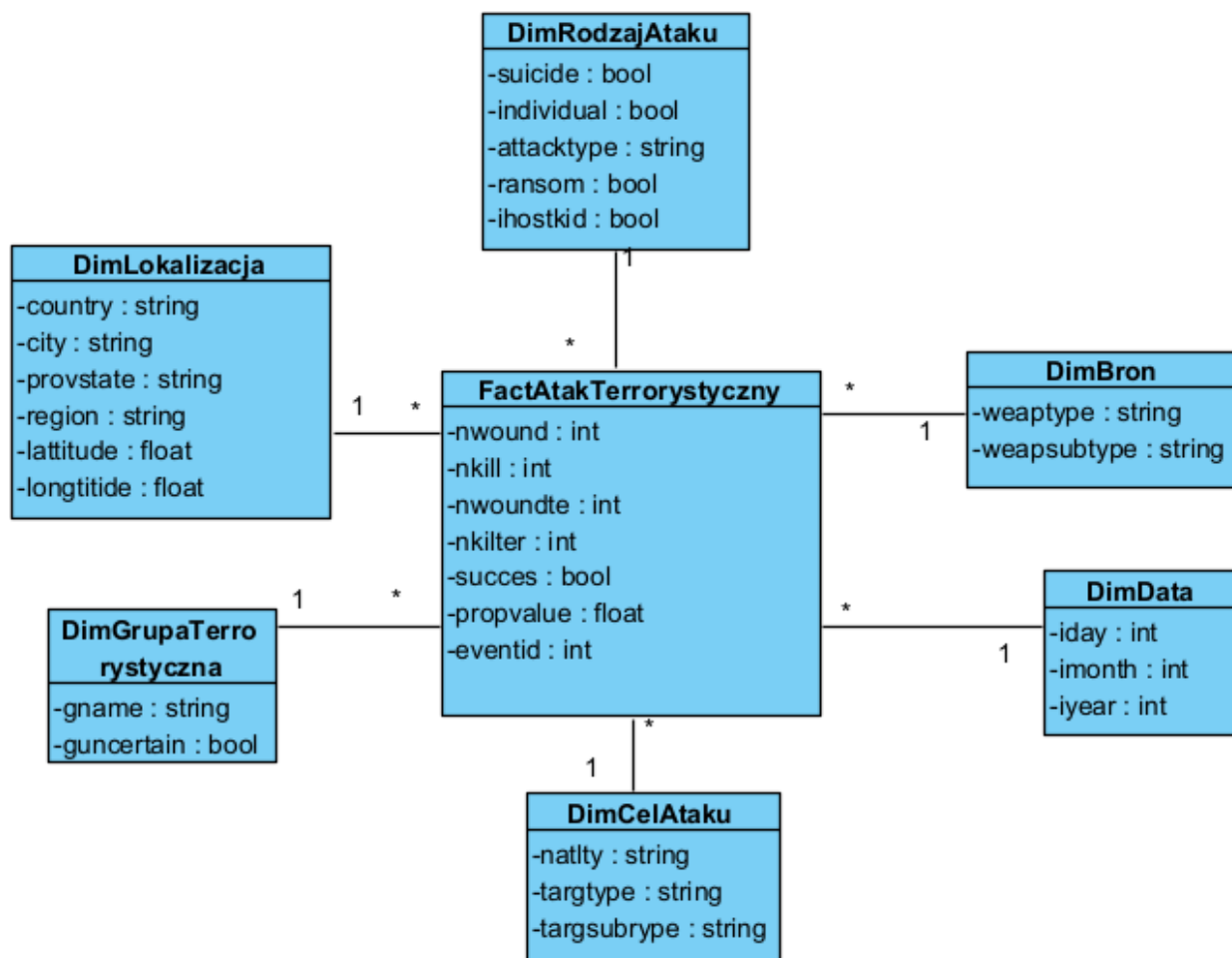
Tabela 4. Zidentyfikowane wymiary

Lp.	Kontekst analizy - wymiary	Opis
1	Lokalizacja	Umożliwia analizę biznesową w kontekście miejsca zdarzenia (faktu). Dostajemy dane o wysokiej precyzji lokalizacji co umożliwi nam tworzenie łatwych wizualizacji na mapach i łatwe grupowanie względem różnych regionów administracyjnych.
2	Grupa terrorystyczna	Pozwoli wytypować najbardziej niebezpieczne grupy terrorystyczne oraz określić taktykę ich działania.
3	Cel ataku	Umożliwi analizę względem typu i narodowości danego celu ataku. Co ułatwi wytypowanie najbardziej narażonych celów.
4	Data	Pozwoli umiejscowić w czasie zdarzenia rejestrowane w fakcie ataku i podjąć analizę na przestrzeni dni, miesięcy oraz lat.
5	Bron	Pozwoli określić typy broni w które zaopatrują się terroryści. Pozwoli także na określenie najbardziej śmiertelnych i generujących największe straty broni, aby opracować sposoby przeciwdziałania ich dostarczania do terrorystów.
6	Rodzaj ataku	Pozwoli określić jakiego rodzaju był atak. Czy porwano zakładników, żądano okupu, czy organizowała go grupa, czy pojedyncza osoba oraz wiele innych.

### 5.3. Modele wielowymiarowe (UML)

Model UML (Rys 1.) zawiera w sobie dane pola o nazwach odpowiadających polom w danych źródłowych, aby ułatwić sobie pracę z nimi. Na tym etapie nie wprowadzam kluczy sztucznych.

**Uwaga:** eventid jest kluczem sztucznym z danych źródłowych.



Rys 1. Wielowymiarowy model analityczny przedstawiony na poziomie konceptualnym

## 6. Projekt procesu ETL

### 6.1. Schemat bazy danych HD (skrypt SQL):

Baza danych utworzona za pomocą poniższego skryptu:

Skrypt 1 - tworzenie hurtowni

```
USE GlobalTerrorism;

CREATE TABLE DimData(
    DataId int IDENTITY(1,1) NOT NULL,
    Dzień tinyint NOT NULL,
    Miesiąc tinyint NOT NULL,
    Rok smallint NOT NULL,

    CONSTRAINT PK_DimData PRIMARY KEY (DataId)
);

CREATE TABLE DimBron(
    BronId int IDENTITY(1,1) NOT NULL,
    TypBroni nvarchar(128) NOT NULL,
    PodTypBroni nvarchar(64) NULL, -- 11% NULL

    CONSTRAINT PK_DimBron PRIMARY KEY (BronId)
);

CREATE TABLE DimCelAtaku (
    CelAtakuId int IDENTITY(1,1) NOT NULL,
    Natality nvarchar(64) NOT NULL,
    TargetType nvarchar(64) NOT NULL, --
    TargetSubType nvarchar(128) NULL, -- 5% NULL

    CONSTRAINT PK_DimCelAtaku PRIMARY KEY (CelAtakuId)
);

CREATE TABLE DimGrupaTerrorystyczna (
    GrupaTerrorystycznaId int IDENTITY(1,1) NOT NULL,
    [Name] nvarchar(128) NOT NULL,
    GroupNameCertain bit NOT NULL, -- 1%< NULL

    CONSTRAINT PK_DimGrupaTerrorystyczna PRIMARY KEY (GrupaTerrorystycznaId)
);

CREATE TABLE DimLokalizacja (
    LokalizacjaId int IDENTITY(1,1) NOT NULL,
    Country nvarchar(64) NOT NULL,
    City nvarchar(128) NOT NULL, -- 1% < NULL
    Provstate nvarchar(64) NOT NULL, -- 1% < NULL
    Region nvarchar(32) NOT NULL,
    DlugoscGeo float NOT NULL, -- 2.5% NULL
    SzerokoscGeo float NOT NULL, -- 2.5% NULL

    CONSTRAINT PK_DimLokalizacja PRIMARY KEY (LokalizacjaId)
);

CREATE TABLE DimRodzajAtaku(
```

```

RodzajAtakuId int IDENTITY(1,1) NOT NULL,
Samobojstwo bit NOT NULL,
OrganizowanyPrzezJednaOsobe bit NOT NULL,
ZadanyOkup bit NULL,
Zakladnicy bit NOT NULL,

CONSTRAINT PK_DimRodzajAtaku PRIMARY KEY (RodzajAtakuId)
);

CREATE TABLE FactAtakTerrorystyczny(
FactAtakTerrorystycznyId int IDENTITY(1,1) NOT NULL,
EventId bigint NOT NULL,
Rannych int NOT NULL, -- 9% NULL
Zabitych int NOT NULL, -- 5% NULL
RannychTerrorystow int NULL, -- 38% NULL
ZabitychTerrorystow int NULL, -- 36% NULL
SukcessAtaku bit NULL, -- 57% NULL
WartoscUszkodzonegoMienia float NULL, -- 78 % NULL

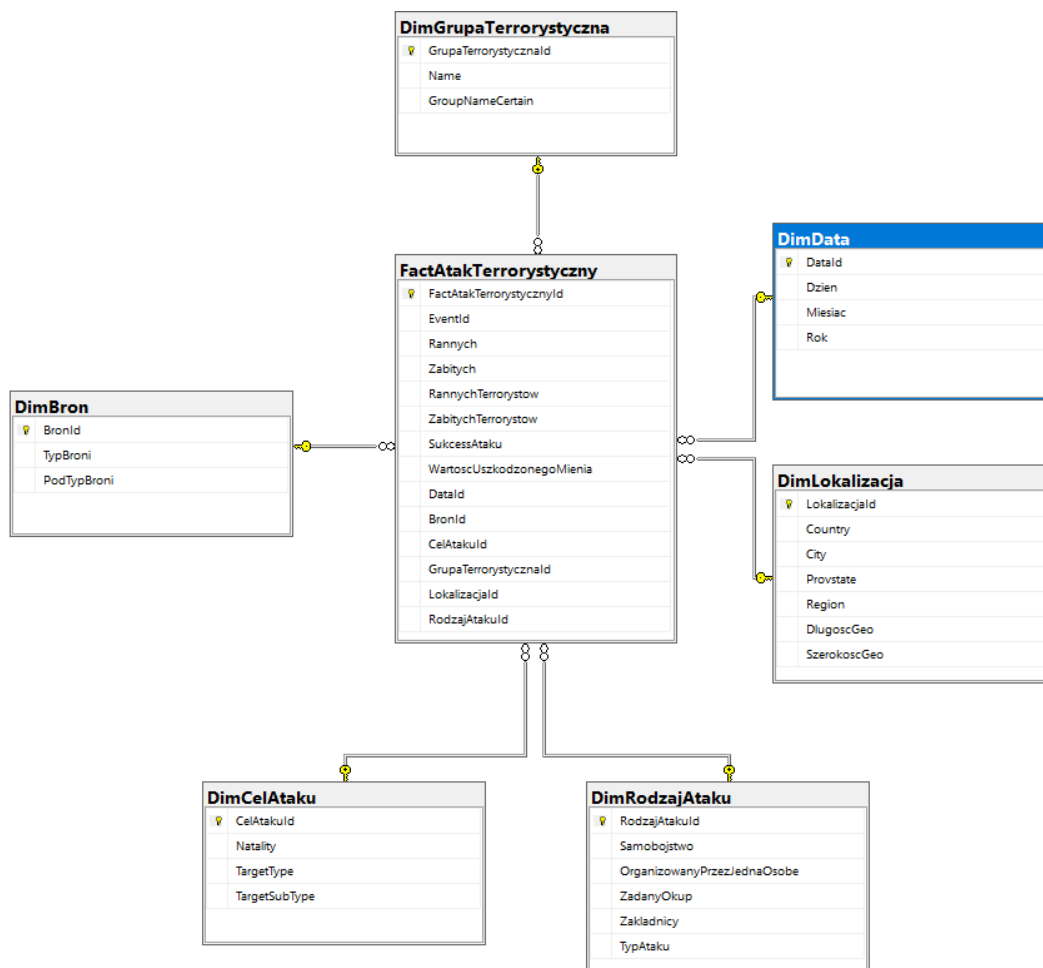
DataId int NOT NULL,
BronId int NOT NULL,
CelAtakuId int NOT NULL,
GrupaTerrorystycznaId int NOT NULL,
LokalizacjaId int NOT NULL,
RodzajAtakuId int NOT NULL,

CONSTRAINT PK_FactAtakTerrorystyczny PRIMARY KEY (FactAtakTerrorystycznyId),
CONSTRAINT FK_FactAtakTerrorystyczny_DimData FOREIGN KEY (DataId)
REFERENCES DimData (DataId),
CONSTRAINT FK_FactAtakTerrorystyczny_DimBron FOREIGN KEY (BronId)
REFERENCES DimBron (BronId),
CONSTRAINT FK_FactAtakTerrorystyczny_DimCelAtaku FOREIGN KEY (CelAtakuId)
REFERENCES DimCelAtaku (CelAtakuId),
CONSTRAINT FK_FactAtakTerrorystyczny_DimGrupaTerrorystyczna FOREIGN KEY
(GrupaTerrorystycznaId)
REFERENCES DimGrupaTerrorystyczna (GrupaTerrorystycznaId),
CONSTRAINT FK_FactAtakTerrorystyczny_DimLokalizacja FOREIGN KEY (LokalizacjaId)
REFERENCES DimLokalizacja (LokalizacjaId),
CONSTRAINT FK_FactAtakTerrorystyczny_DimRodzajAtaku FOREIGN KEY (RodzajAtakuId)
REFERENCES DimRodzajAtaku (RodzajAtakuId),
);

```



Schemat bazy danych wygenerowany przez Microsoft SQL Server Management Studio (Rys. 2):



Rys. 2 - schemat bazy hurtownianej

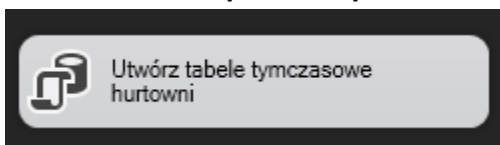
## 6.2. Specyfikacja procesów ETL (Control Flow + Data Flow)

Etapy:

### 1. Ładowanie surowych danych do bazy danych:

Z powodu problemów z ładowaniem danych bezpośrednio z pliku za pomocą MS SISS, zdecydowałem się na surowych danych do tabeli wygenerowane przez Tabelau (zawierające wszystkie kolumny pliku .csv z danymi). Na jej podstawie można było przeprowadzić profilowanie danych oraz dalszy proces ETL.

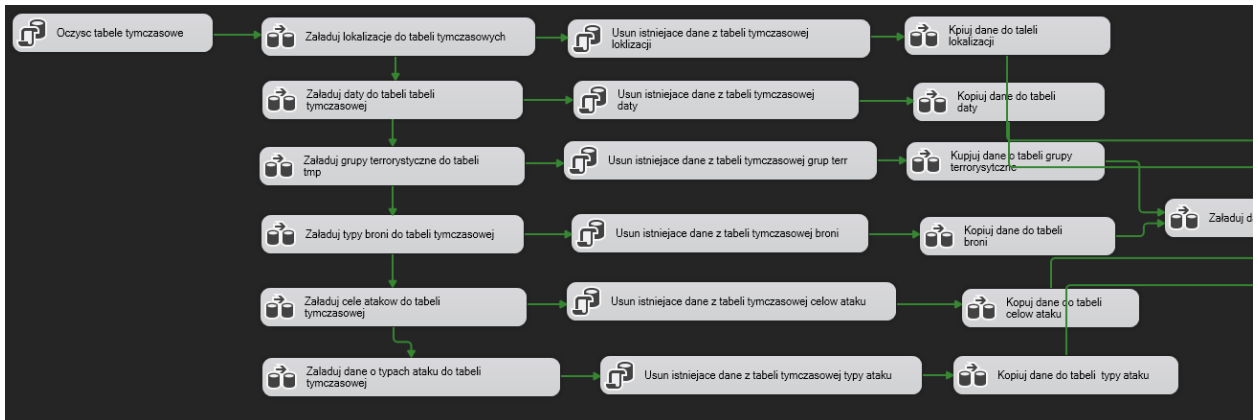
### 2. Tworzenie tabel tymczasowych



Rys. 3

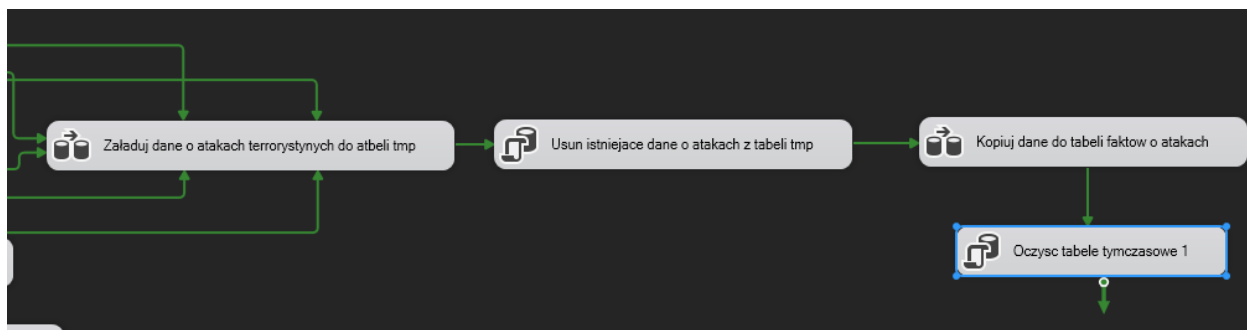
Tabele tymczasowe są tworzone w analogiczny sposób do tabel hurtowni ([Skrypt 1](#)), z tą różnicą, że przy nazwie każdej z nich występuje prefix „Tmp”.

### 3. Aktualizacja danych w hurtowni – proces ETL:



Rys 4.1

Ciąg dalszy...



Rys. 4.2

- **Extract** – czyli druga kolumna (Rys 4.1), polega na ładowaniu do tabel tymczasowych danych ze zdefiniowanego źródła oraz ich odpowiednim oczyszczeniu.

Wszystkie data flow z tego etapu wyglądają podobnie.

Źródłem danych jest dla nas zawsze kwerenda wybierająca dla nas interesujące nas dane.

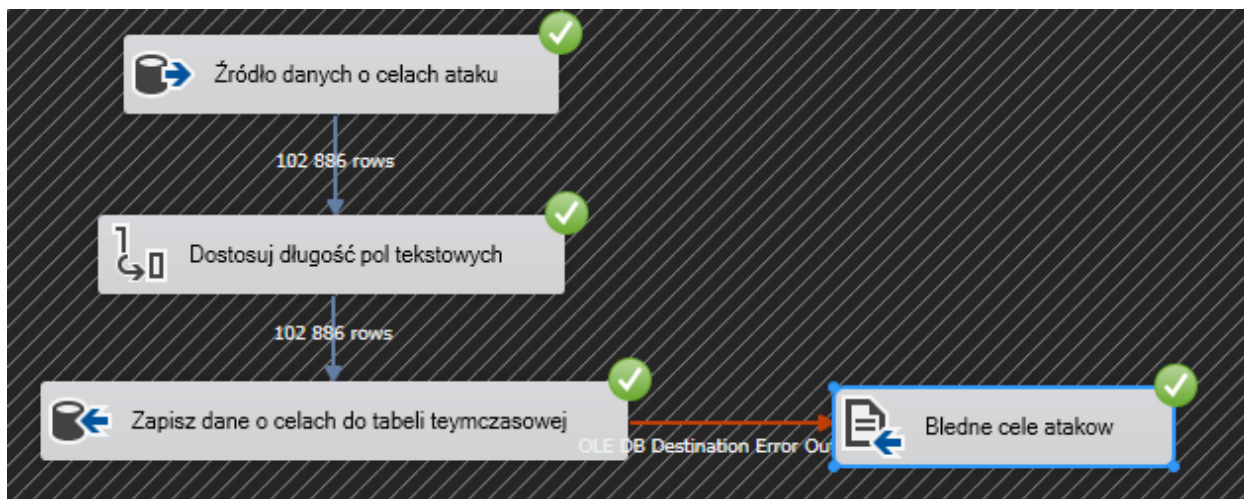
Np. skrypt dla „źródła danych o celach ataku” (Rys. 7):

```
SELECT DISTINCT
    LOWER(targsubtype1_txt),
    LOWER(targtype1_txt),
    LOWER([natlty1_txt])
FROM [dbo].RawData
WHERE targtype1_txt is NOT NULL
    AND [natlty1_txt] is NOT NULL;
```

### Skrypt 2

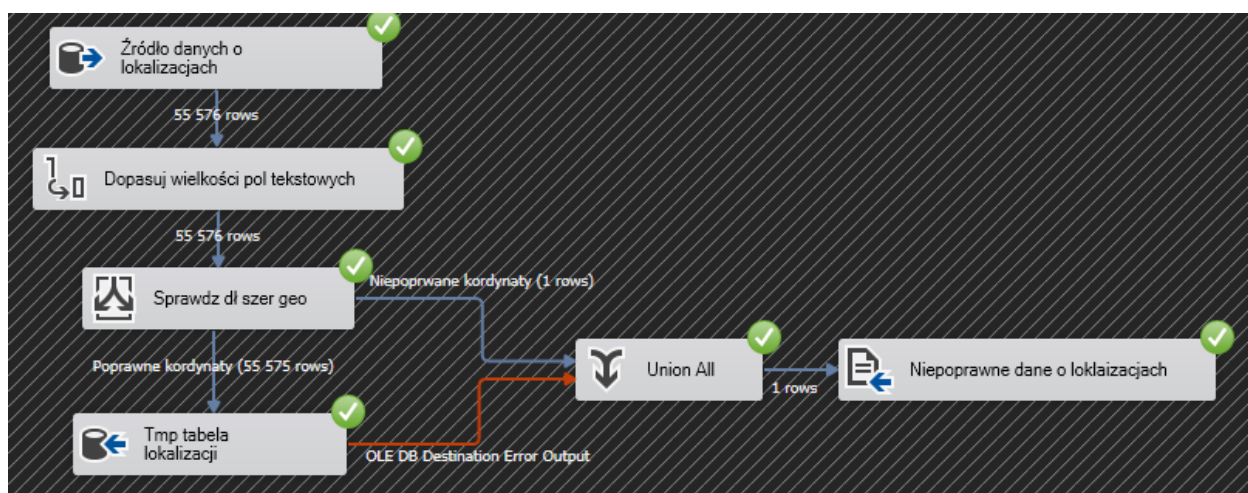
Wybranie skryptu z klauzulą **DISTINCT** pozwoliło na zaoszczędzenie sporego czasu porównując do funkcji „sort” z usuwaniem duplikatów z programu MS SISS. Umożliwiło to testowanie dataflow w realnym dla śmiertelników czasie i realnych dla mojego laptopa zasobach.

We wszystkich polach tekstowych występujących w danych źródłowych stosuję funkcję LOWER(), aby uniknąć zbędnych duplikatów.



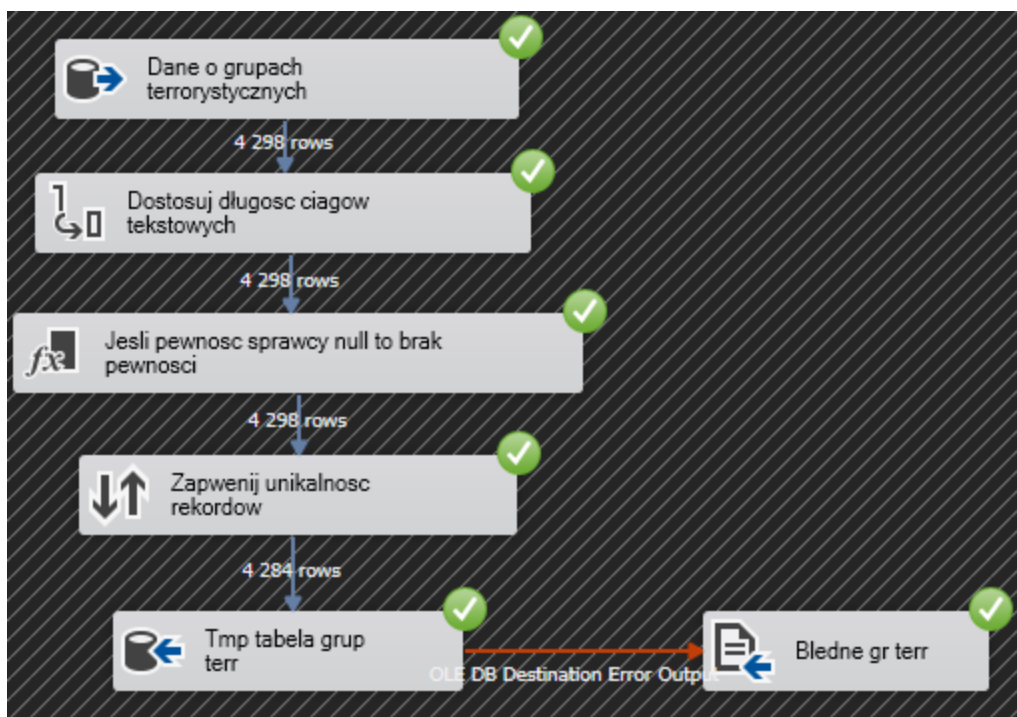
Rys. 7 – ładowanie danych o celach ataku

Następnie dokonujemy rzutowania na odpowiednie typy. W razie wyjątku rzucamy błędem ponieważ nie spodziewamy się go tutaj ze względu na wcześniej przeprowadzone profilowanie danych i dostosowanie do typów w tabelach docelowych.



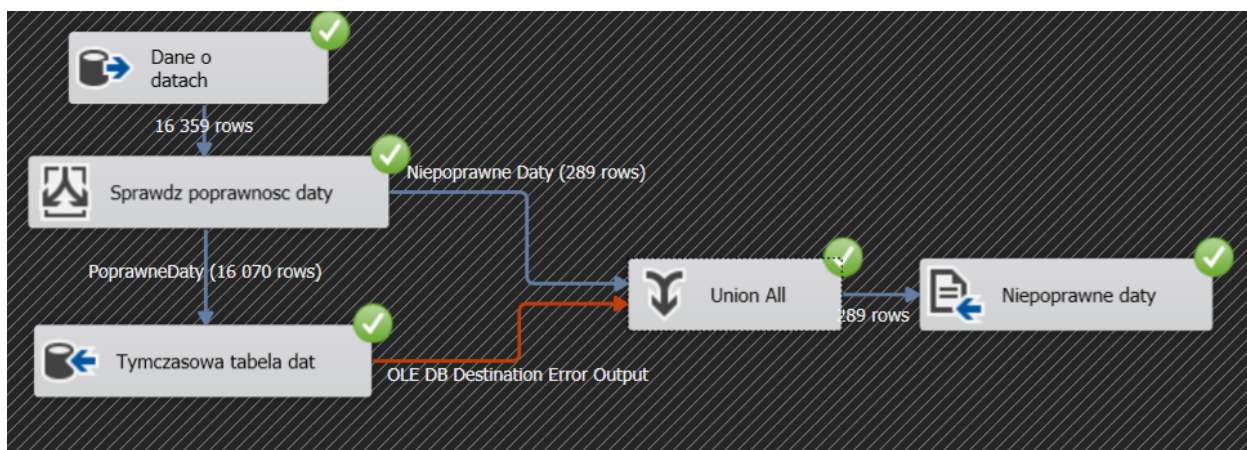
Rys. 5 – ładowanie danych o lokalizacjach

To co odróżni ten data flow (Rys. 5) od wszystkich pozostałych jest element sprawdzający poprawność dł. i szer. Geograficznej. Jak widzimy jest tylko 1 niepoprawny rekord który wygląda na brak przecinka w wartości.



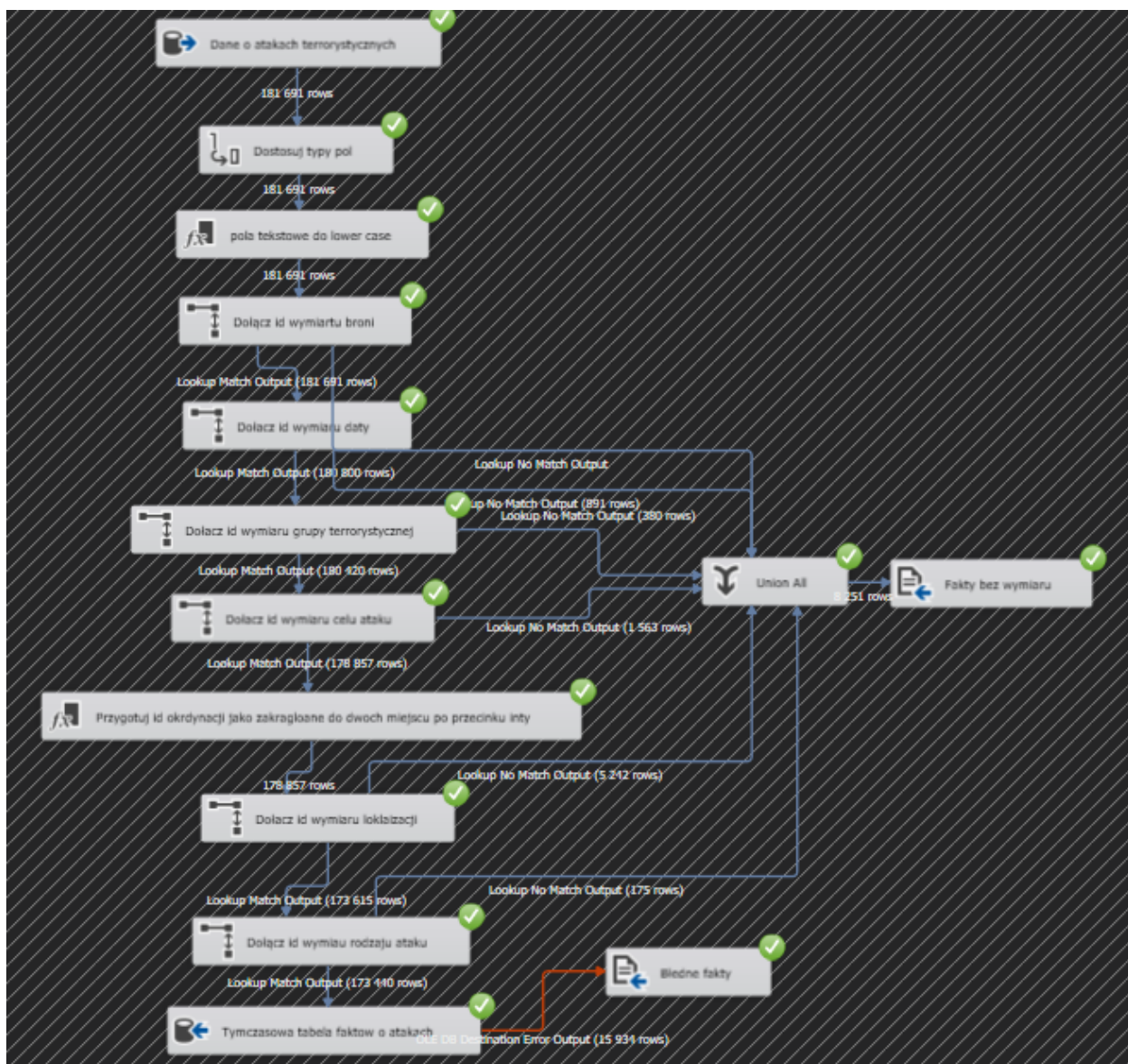
Rys. 6 – ładowanie danych o grupach terrorystycznych

Dataflow z Rys. 6 posiada dodatkowy element zamieniający wartości null pola gncertain (określającego czy jesteśmy pewni co do sprawcy zamachu) na wartość false.



Rys. 11 – ładowanie dat zdarzeń

Przy ładowaniu dat (Rys. 11) należało przeprowadzić dodatkowy etap weryfikacji poprawności daty ponieważ zdarzało się trafiać na zerowy dzień miesiąca.



Rys. 8 – ładowanie faktów ataku

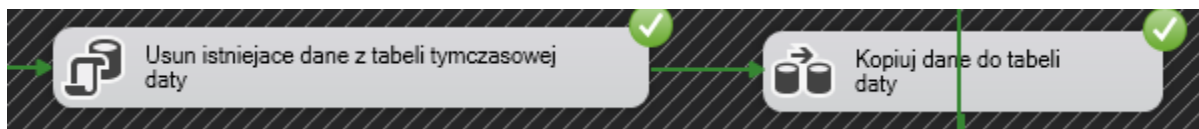
Ten dataflow (Rys. 8) jest nieco bardziej skomplikowany ponieważ teraz musimy dołączyć do każdego faktu odpowiadając mu wymiary oraz zapisać fakty które nie miały wystarczających danych aby utworzyć wymiar lub odpowiednich miar.

To co sprawiło mi największy problem to lookup dla Wymiaru lokalizacji ponieważ nie znalazłem w MS SISS, żadnego element który pozwolił zrobić join na typie float (dł. Szer. geo), dlatego postanowiłem, że przekształcę te wartości na integer z pewną dokładnością po przecinku i wtedy dokonał złączenia.

Oczywiście wszystkie pola tekstowe przed lookup-em musieliśmy doprowadzić do postaci lower case jak to zrobiliśmy wcześniej z wymiarami.

Z danych wejściowych po całym procesie ETL odrzuciliśmy około 24 tys. rekordów.

- **Transform** – w kolejnym z etapów po skopiowaniu danych do tabel tymczasowych wykonujemy dla każdego wymiaru i faktu SQL task który usuwa nam istniejące już w HD dane, aby unikać duplikatów w bazie. Przykład dla TmpDimData poniżej.

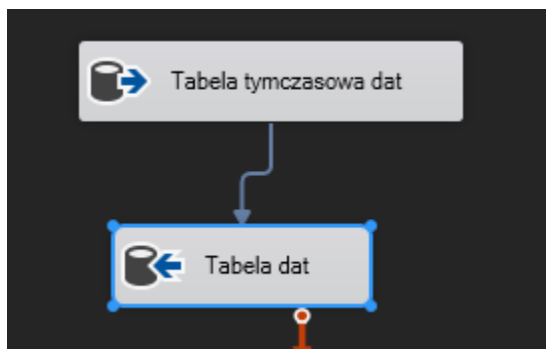


Rys. 9 – Fragment Rys. 4

```
DELETE TmpDimData
FROM TmpDimData td
    INNER JOIN DimData d
    ON td.Dzien = d.Dzien
    AND td.Miesiac = d.Miesiac
    AND td.Rok = d.Rok;
```

Skrypt 3 – przykład usuwania duplikatów z tabeli tymczasowej

- **Load** - w ostatnim z etapów kopiujemy dane do tabel docelowych. Następnie przy pomocy SQL Task czyścimy tabele tymczasowe, usuwając z nich dane, aby przygotować je na proces od nowa, gdy dostaniemy nowe dane. Przykładowy dataflow dla tabeli DimData i TmpDimData.



Rys. 10 – kopiowanie danych z tabel tymczasowych do właściwych

```
DELETE FROM TmpFactAtakTerrorystyczny WHERE 1=1;
```

```
DELETE FROM TmpDimData WHERE 1=1;
```

```
DELETE FROM TmpDimBron WHERE 1=1;
```

```
DELETE FROM TmpDimCelAtaku WHERE 1=1;
```

```
DELETE FROM TmpDimGrupaTerrorystyczna WHERE 1=1;
```

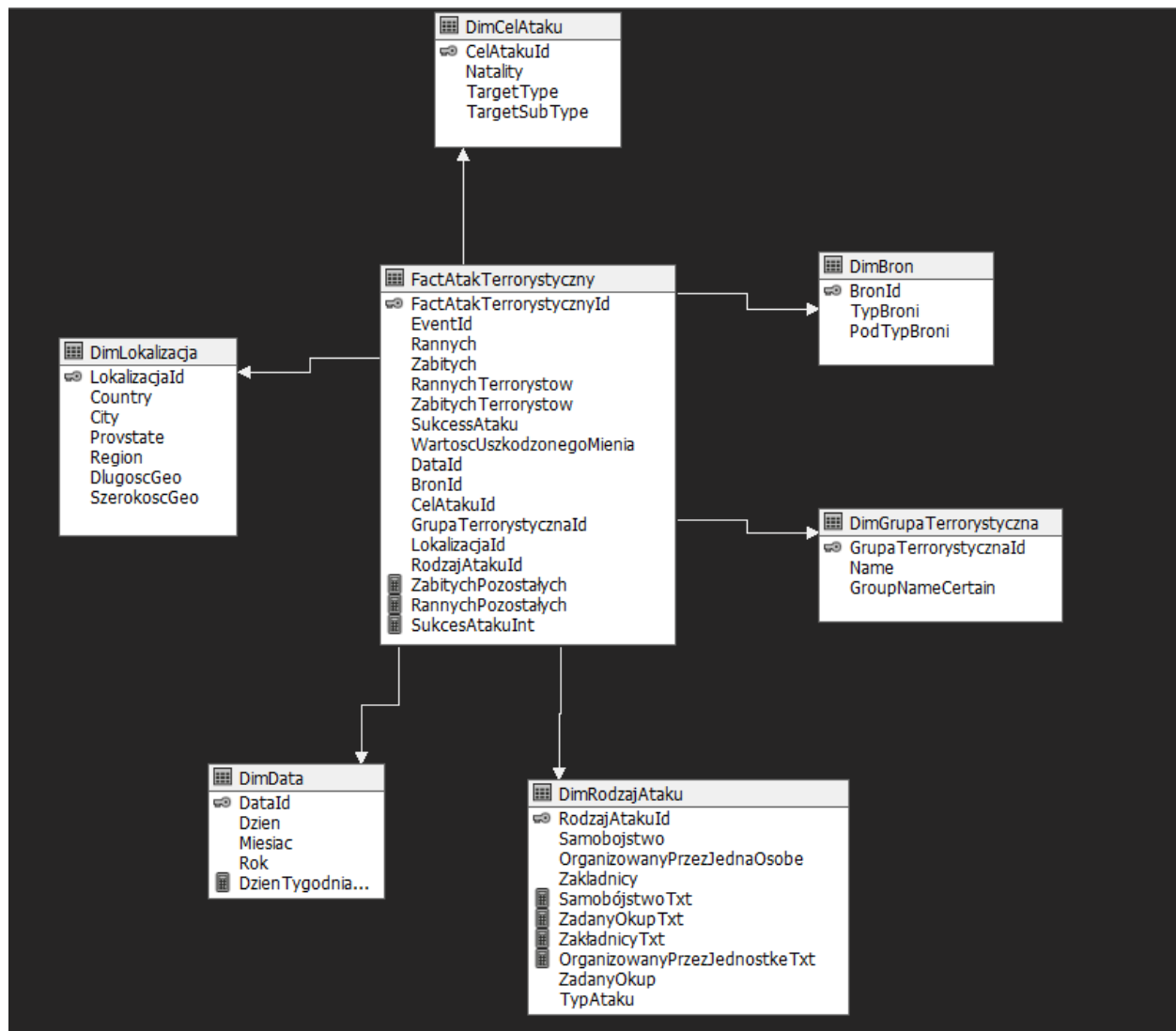
```
DELETE FROM TmpDimLokalizacja WHERE 1=1;
```

```
DELETE FROM TmpDimRodzajAtaku WHERE 1=1;
```

Skrypt 4 – czyszczący tabele tymczasowe.

## 7. Implementacja modeli wielowymiarowych

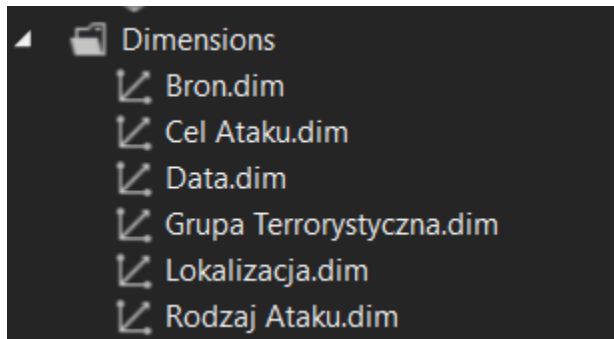
### 7.1. Widok danych



Rys. 11 – schemat HD



## 7.2. Wymiary:



Rys. 12 – zdefiniowane wymiary

Dodatkowo dodałem atrybuty wyliczane w celu ułatwienia analizy danych:

- DimData:
  - DzieńTygodniaAng – Pozwala badać występowanie ataków w zależności od dnia tygodnia

```
SELECT DATENAME(WEEKDAY, CONCAT(  
    REPLACE(STR([Rok],4), ' ', '0'), '-',  
    REPLACE(STR([Miesiac],2), ' ', '0'), '-',  
    REPLACE(STR([Dzien],2), ' ', '0'))  
FROM [dbo].[DimData];
```

Skrypt 5 – wyznaczania pola DzieńTygodniaAng

- DimRodzajAtaku:
  - SamobójstwoTxt, ZadanyOkupTxt, ZakładnicyTxt, OrganizowanyPrzezEdnostkeTxt – zamieniamy wartości boolowskie na łatwiejsze do interpretacji dla użytkownika

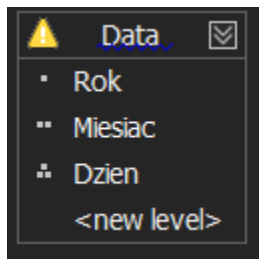
```
CASE  
WHEN [Samobojstwo]= 1 THEN 'Samobójczy'  
WHEN [Samobojstwo] = 0 THEN 'Inny'  
ELSE 'Brak danych'  
END
```

Skrypt 6

Analogicznie postąpiłem z pozostałymi wartościami bookowskimi: ZadanyOkupTxt, ZakładnicyTxt oraz OrganizowanyPrzezEdnostkeTxt.

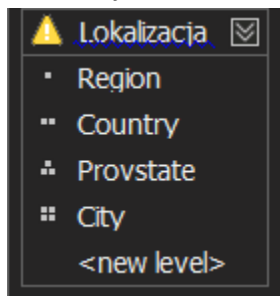
Zdefiniowałem także następujące hierarchie:

- DimData:



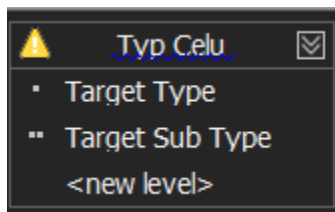
Rys. 13

- DimLokalizacja:



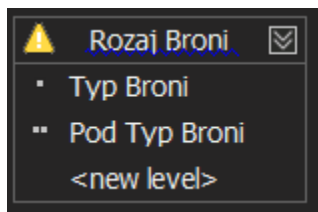
Rys. 14

- DimCelAtaku



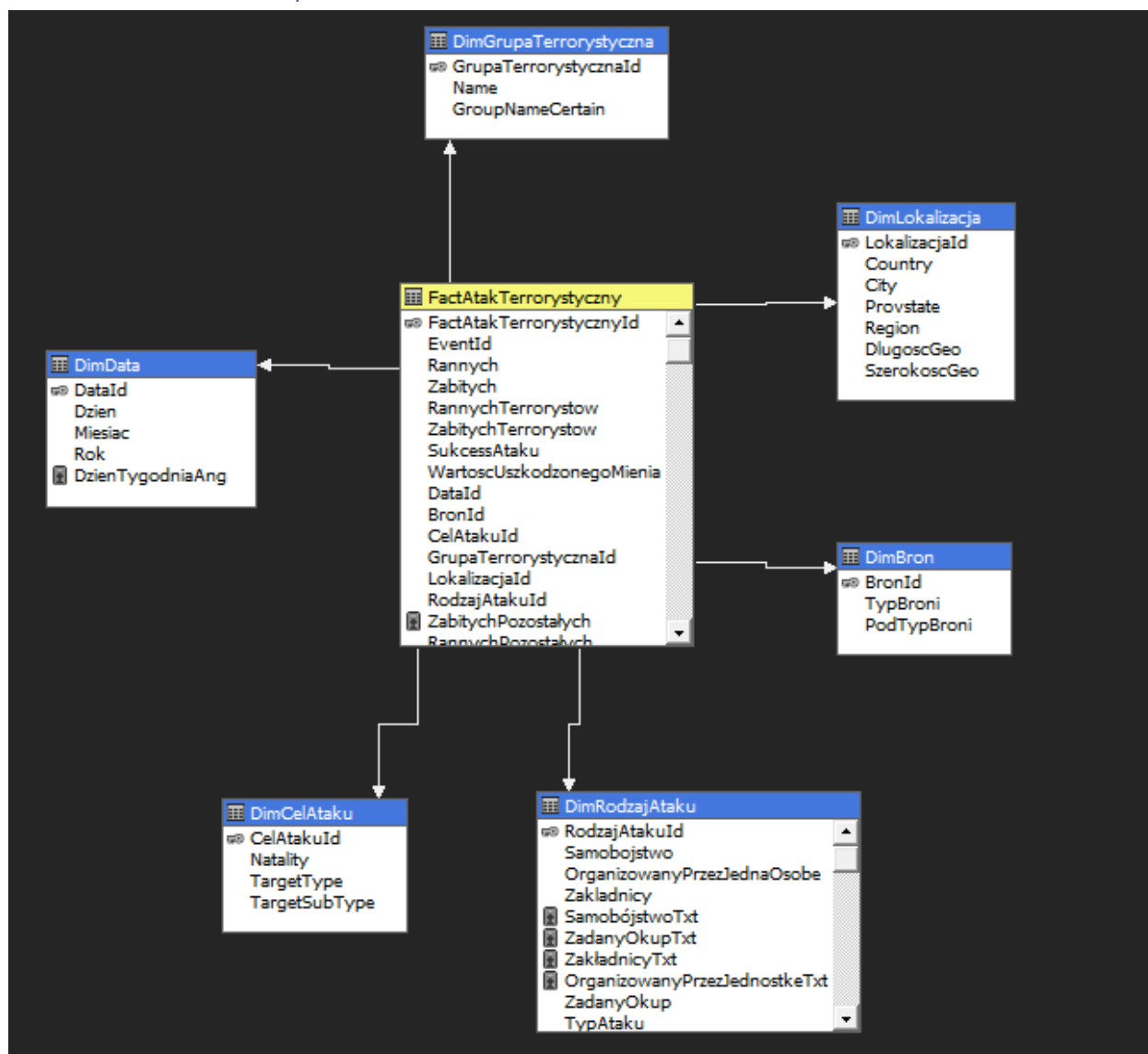
Rys. 15

- DimBron:



Rys. 16

### 7.3. Modele wielowymiarowe – Kostki

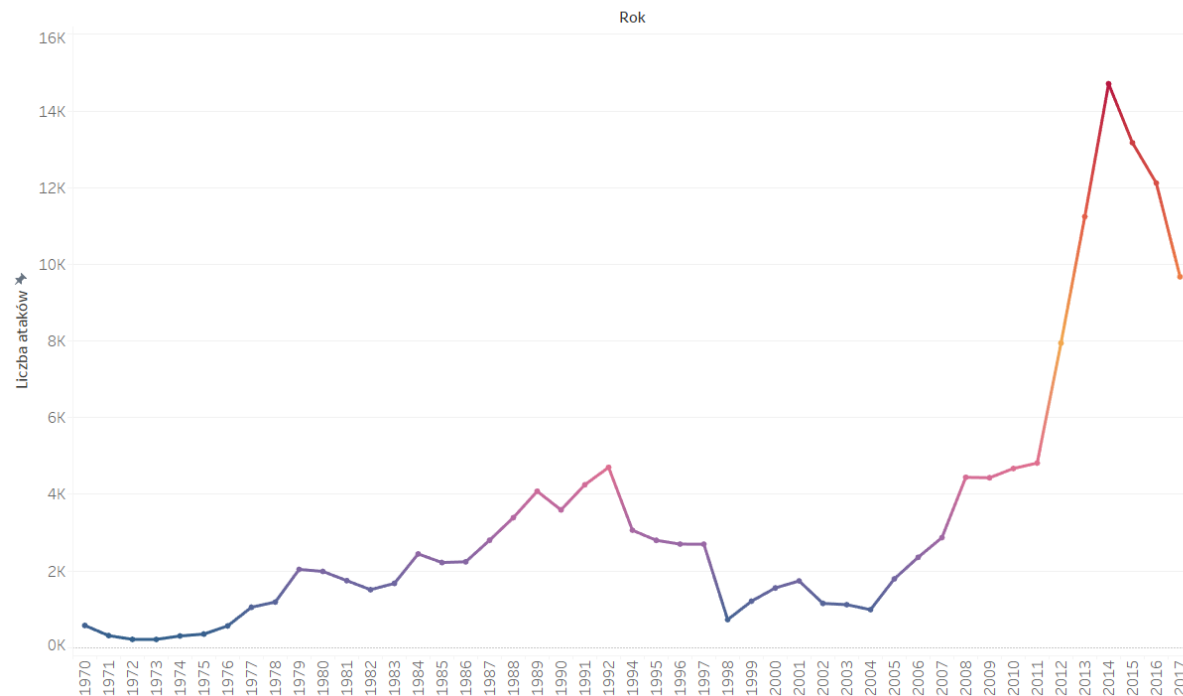


Rys. 17

## 8. Analiza danych

### 8.1. Realizacja procesów analitycznych

Liczba ataków w zależności od roku

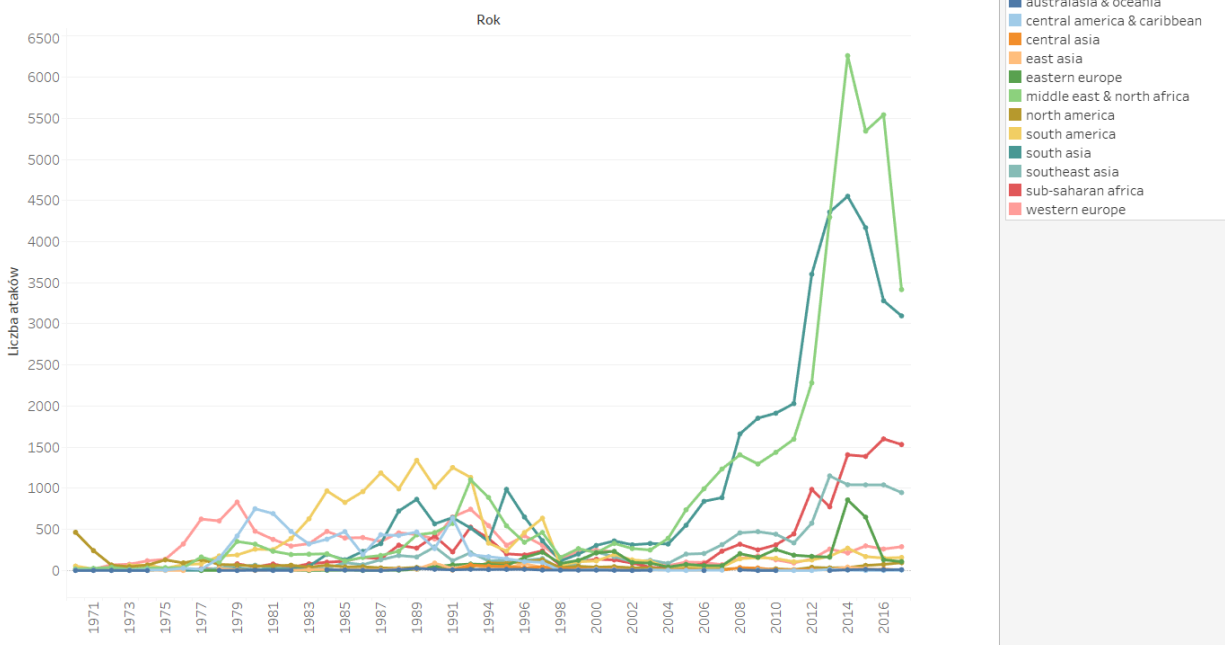


Wykr. 1

Analizę rozpoczynam od analizy rozkładu ataków terrorystycznych na przestrzeni lat. Z łatwością można zaobserwować dwa maksima pojawiające się na wykresie przypadające na **1992** oraz **2013**, po których zamiast wzrostu ilości ataków obserwowaliśmy ich spadek.

Szczególnie wyróżnia tu się okres **2011-2014**, gdzie obserwujemy bardzo intensywny wzrost ilości ataków. Powinniśmy sprawdzić czy ten wzrost dotyczy konkretnych regionów czy jest Trądem światowym

Liczba ataków w regionach w zależności od roku

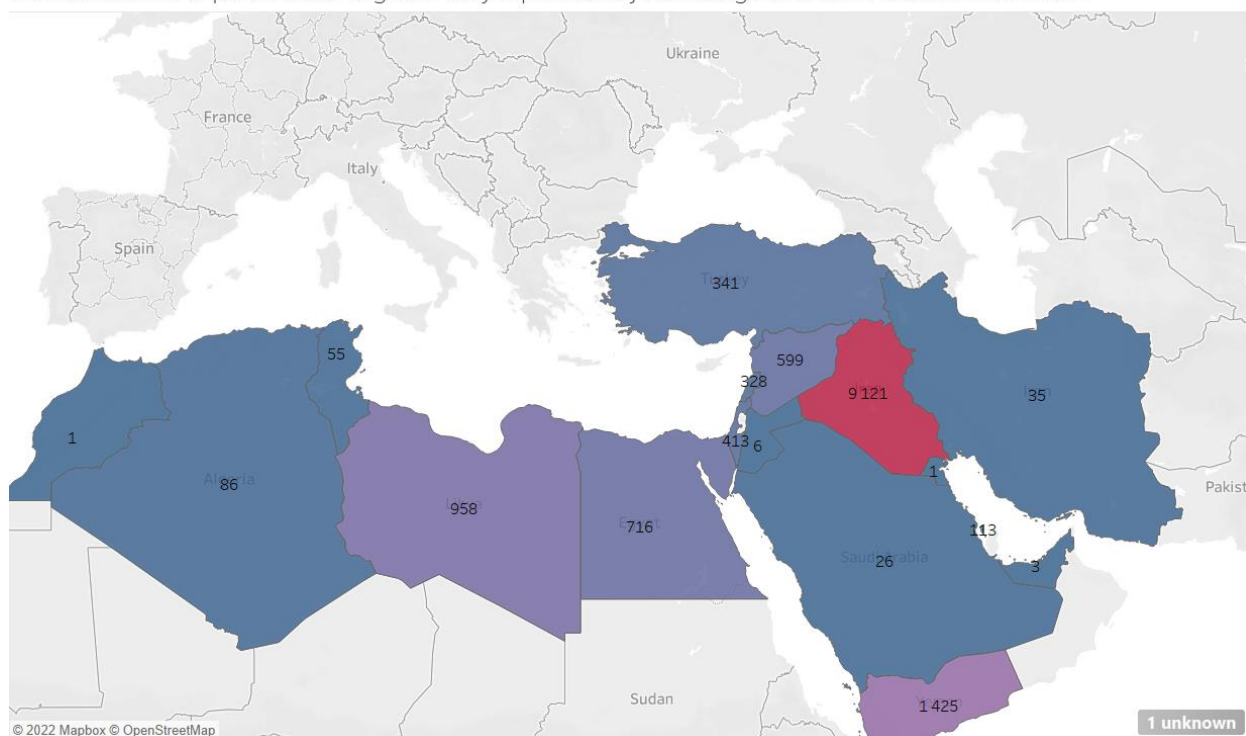


Wykr. 2

Na tym wykresie udało mi się zauważyć, że bardzo intensywny wzrost liczby ataków terrorystycznych w latach **2011-2014** jest związany z konkretnymi regionami. Mianowicie **Południowa Azja** oraz **Bliski Wschód & Północna Afryka**. Jednocześnie kolejne największe wartości przyjmują regiony przylegające do nich.

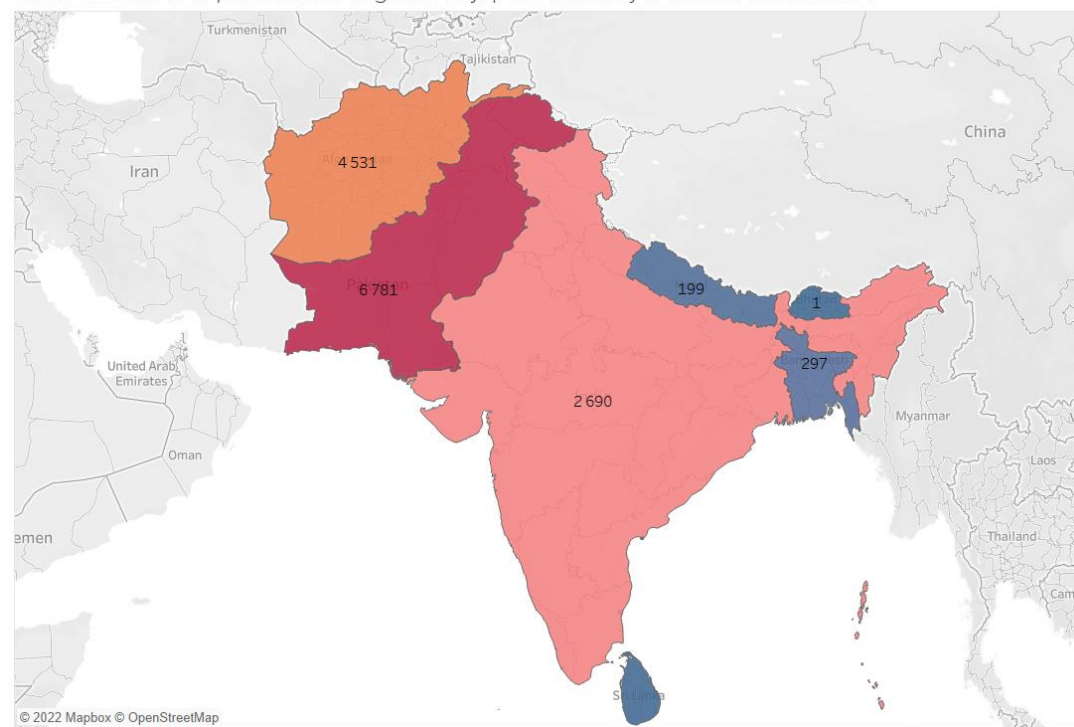
Drugi obserwowalny wzrost przypadający na okres **1983-1994** dotyczy najbardziej **South America**.

Liczba ataków w państwach regionu afryki północnej i bliskiego wschodu latach 2011-2014



Wykr. 3

Liczba ataków w państwach regionu azji południowej w latach 2011-2014

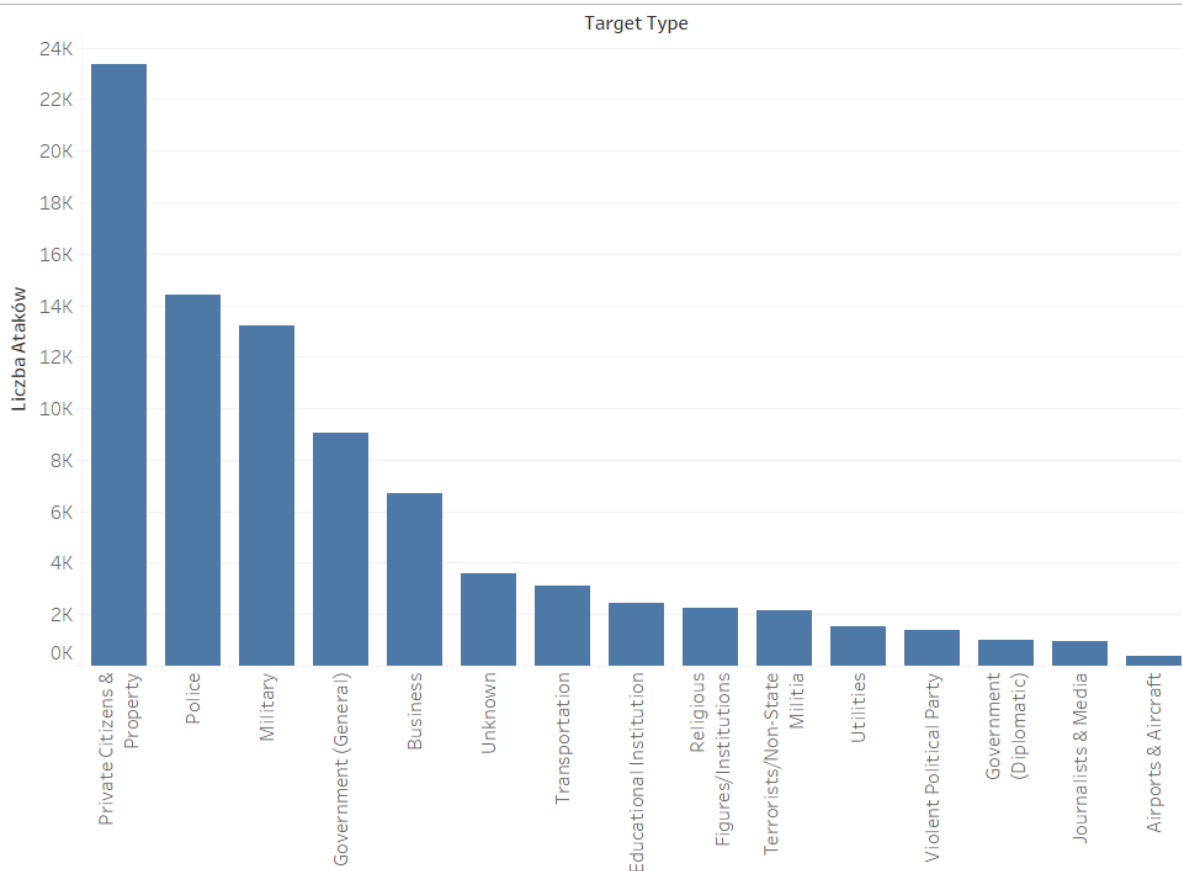


Wykr. 4

Na podstawie tych wykresów możemy stwierdzić, że to **Irak, Pakistan i Afganistan**. Były regionami w których doszło do największego przyrostu ataków terrorystycznych w tych latach i jednocześnie są to państwa o największym zagrożeniu atakiem terrorystycznym.

Dla Iraku powodem dużej aktywności terroryzmu jest destabilizacja regionu przez Amerykanów oraz duże różnice ideowe powstałe w zamieszkującym ten kraj ludzie. W 2011 roku amerykanie opuścili teren Iraku, aby powrócić w 2014. To był z pewnością jeden z głównych czynników wpływających na ilość ataków.

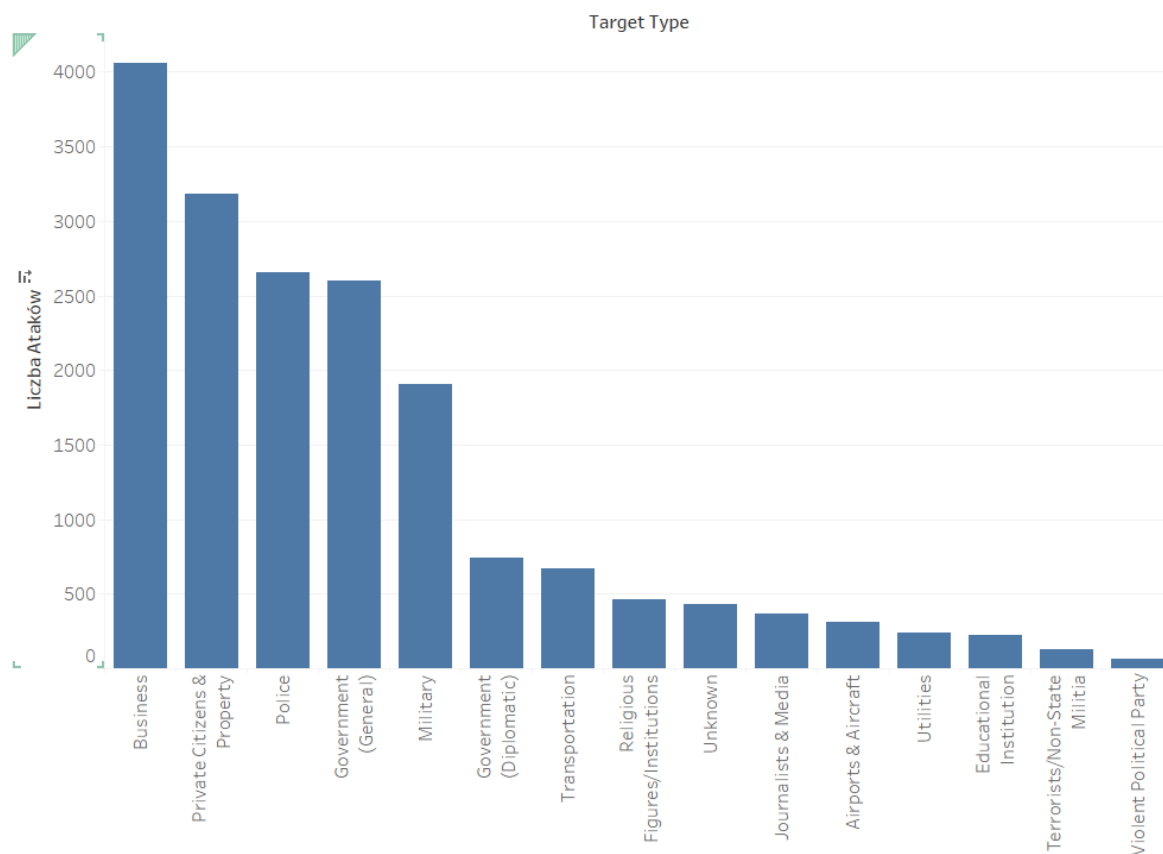
Liczba ataków na określonego typu cele w karajach islamskich



Wykr. 5

Jak widać na wykresie najbardziej narażoną grupą na atak w państwach islamskich są cywile. To potwierdza, że największym problemem tych ludów jest brak tolerancji religijnej i ideowej oraz wpajana w kulturze nienawiść do drugiego człowieka z powodu tego kim jest.

## Liczba ataków na określonego typu cele w europie

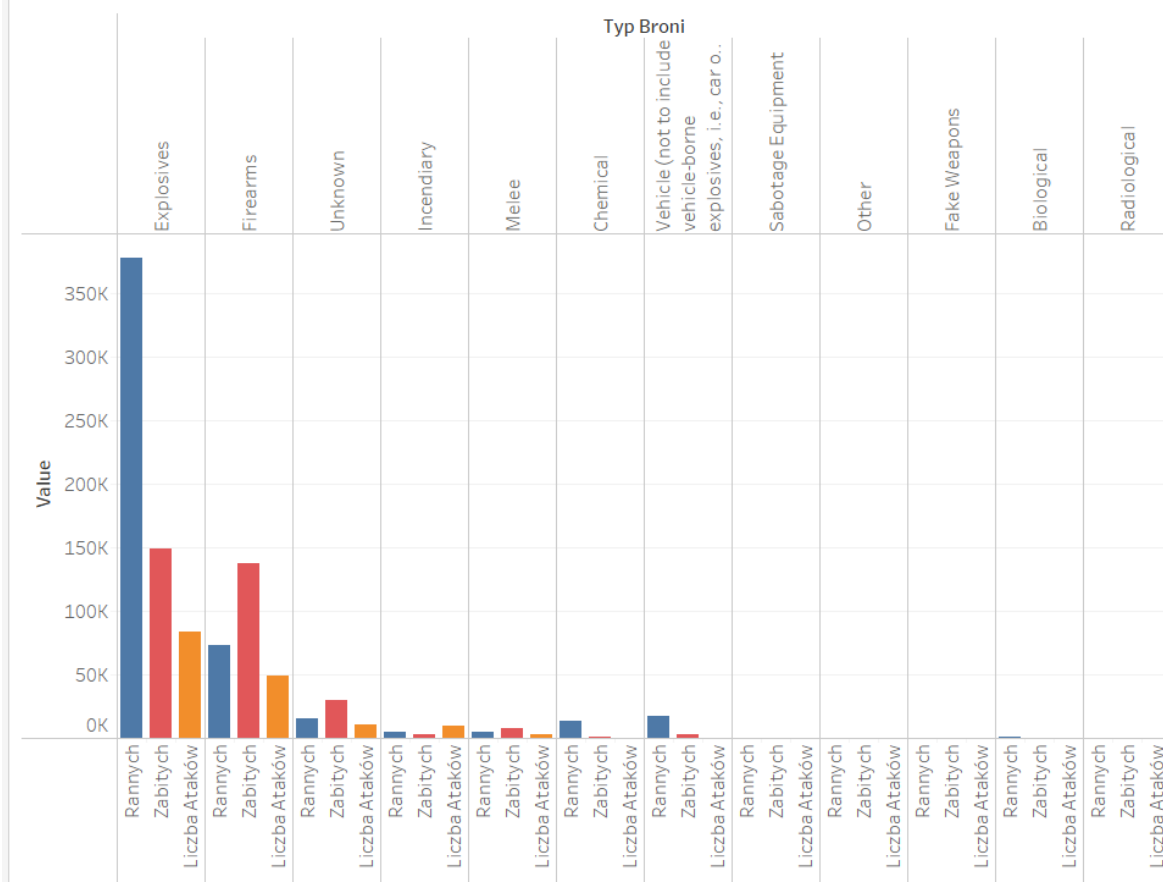


Wykr. 6

Dla odmiany na terenie Europy częściej celem ataku padają duże korporacje, co może wynikać z tego w jaki sposób traktują one ludzi i środowisko. Ale terroryzm skierowany na cywili jest również obecny i to na drugim miejscu. Możemy zauważyć też, że niezależnie od regionu policja i organizacje próbujące pacyfikować/"utrzymywać porządek" w danym regionie zawsze padają dość często celem ataków. Widocznie ludzie zawsze nie lubią kiedy ktoś im mówi co mają robić.

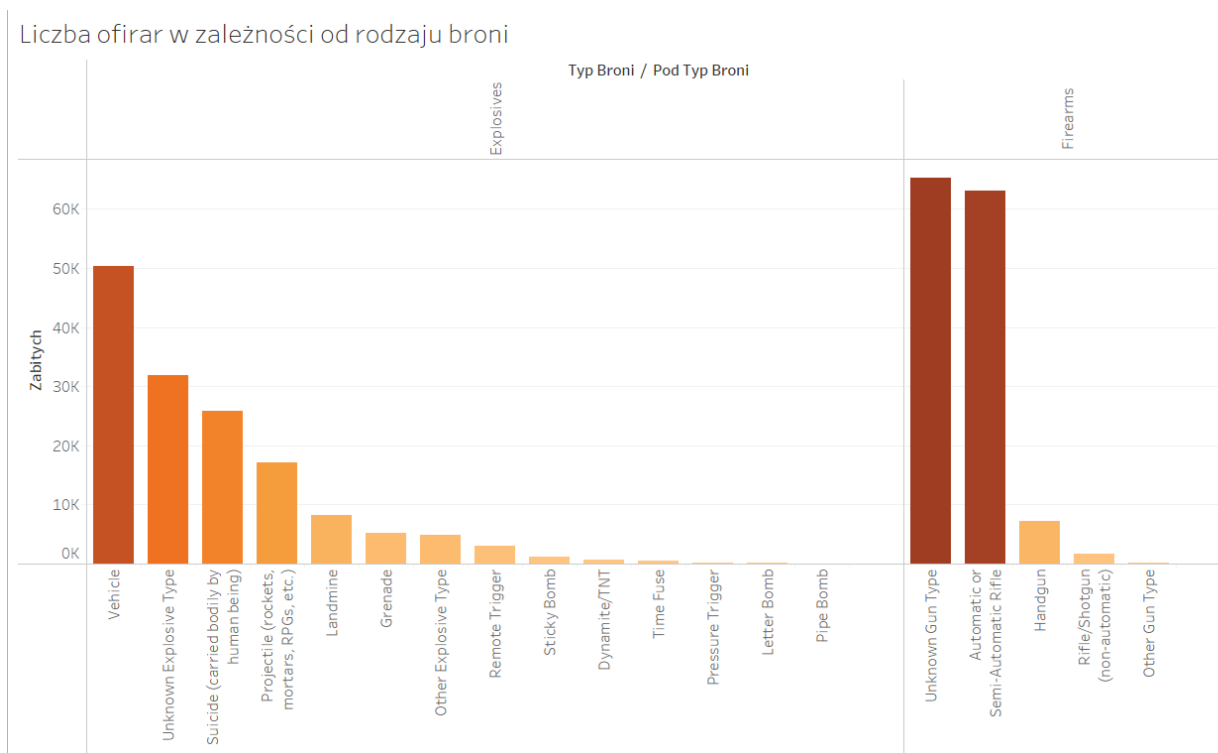


## Tyby broni używane do ataków



Wykr. 7

Możemy zauważyć, że wśród terrorystów królują **ładunki wybuchowe** oraz **Broń palna**. Możemy zauważyć, że to właśnie broń palna jest najbardziej śmiertelną bronią powodując największą liczbę zgonów w jednym ataku i relatywnie niewiele rannych. Z kolei ładunki wybuchowe powodują obrażenia u znacznie większej liczby osób jednak często są to rany z których udaje się je uratować.



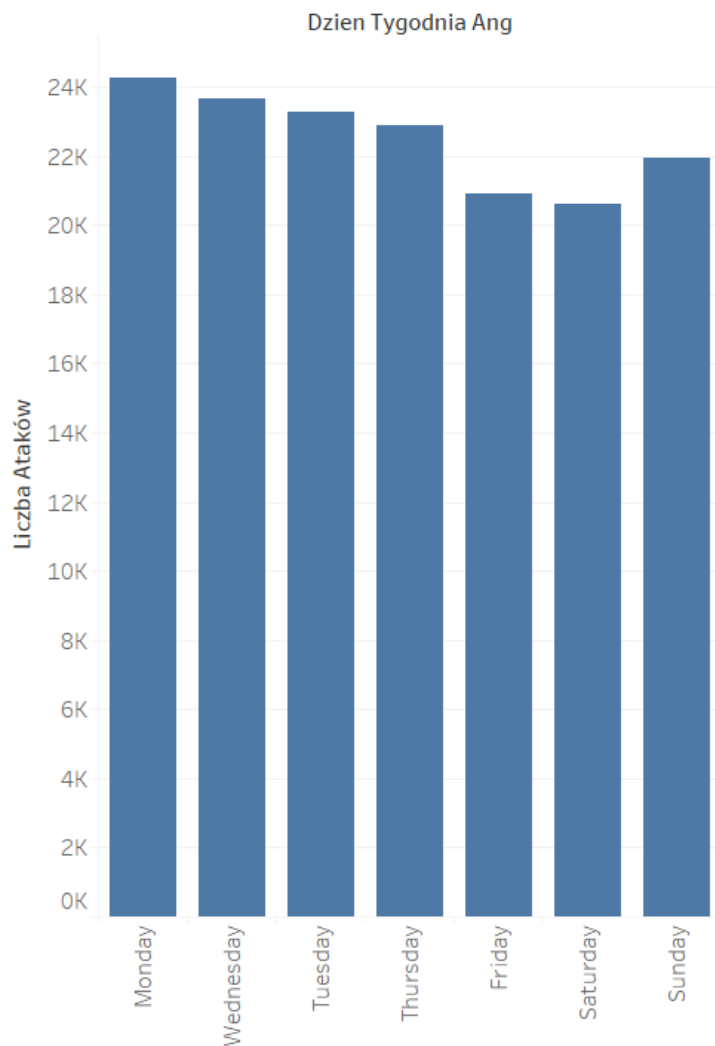
Wykr. 8

Zawężając naszą analizę. Poszukujemy broni które są największym zagrożeniem dla ludzkiego życia.

Wśród ładunków wybuchowych królują „**Samochody bomby**”, za którym w czołówce plasują się ataki **samobójcze** z ładunkami przy ciele. Dlatego powinniśmy skupić się na szukaniu sposobów wczesnego wykrywania tego typu ataków.

Zauważamy, że wśród broni palnych największe zagrożenie stanowią różnego rodzaju **bronie automatyczne i powtarzalne**. Nie bez powodu w wielu krajach zabrania się sprzedaży broni automatycznej i powtarzalnej długiej (którą łatwo przerobić na automatyczną). Ma to ograniczyć dostęp do tego typu broni niepowołanym osobą. Jednak moim zdaniem nie sprawdza to się w przypadku przestępczości zorganizowanej. Wtedy grupa przestępcza i tak znajdzie sposób na zdobycie broni.

## Liczba ataków w poszczególnych dniach tygodnia:

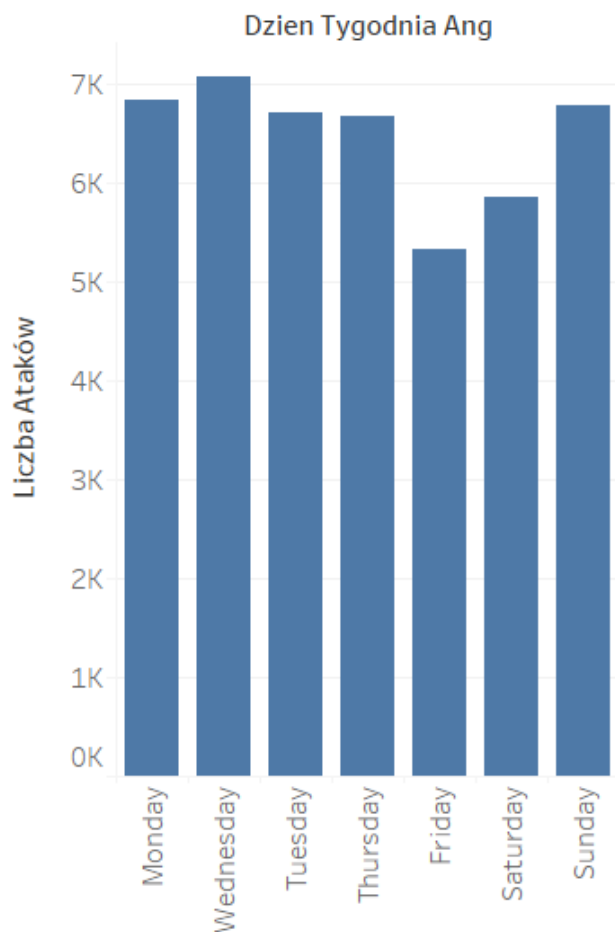


Wykr. 9

Nie obserwujemy szczególnie znaczących wzorców w wyborze dnia ataku. Poza tym, że w poniedziałek występują największa liczba ataków i spada łagodnie do piątku, by znowu wzrastać.

Być może wpływ na to ma kultura/religia krajów o największej liczbie aktów terrorystycznych. W końcu piątek jest dniem wspólnej modlitwy dla religii islamu.

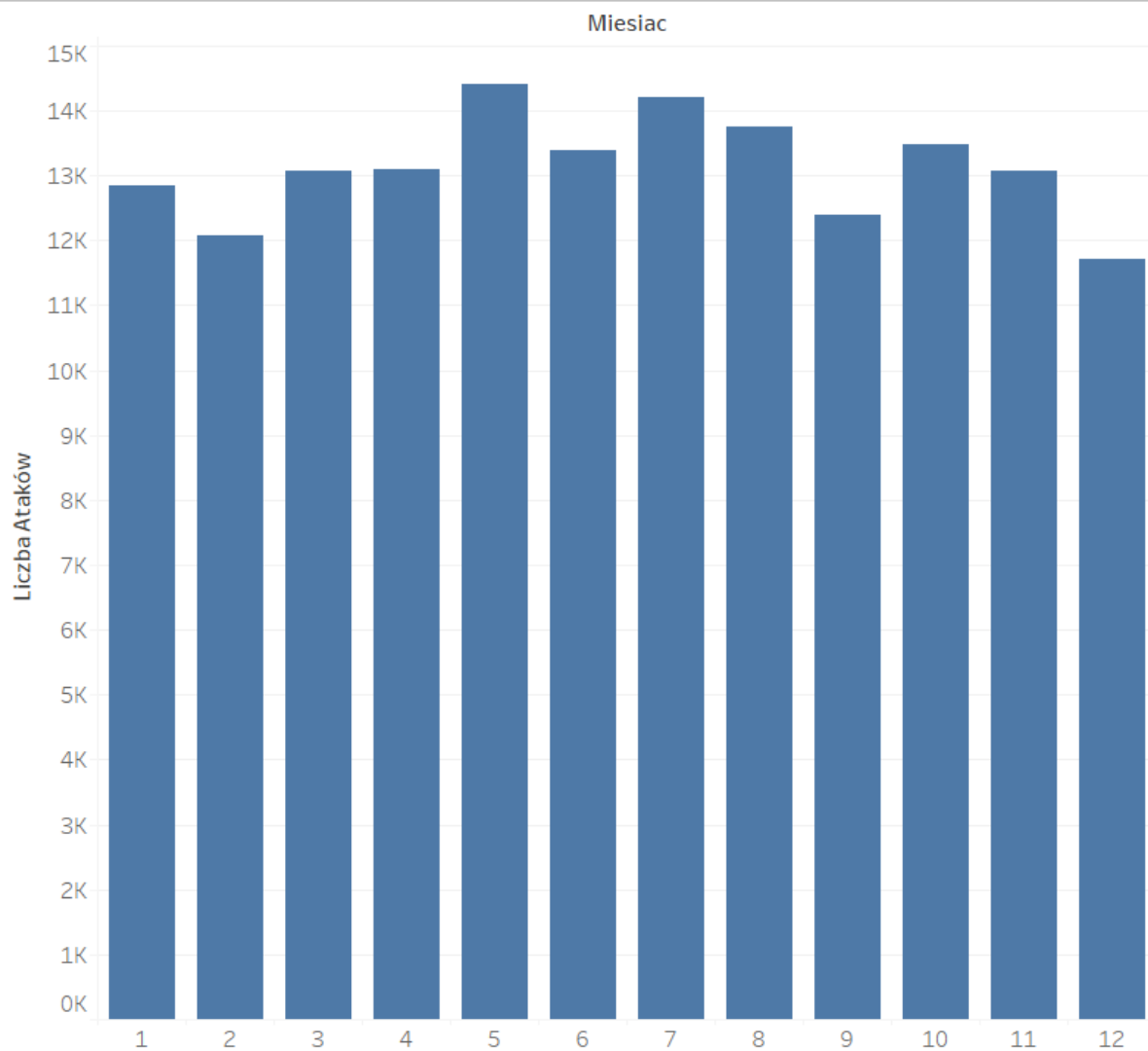
## Liczba ataków w poszczególnych dniach tygodnia w regionach krajów arabskich



Wykr. 10

Wykresy potwierdzają. W każdym z regionów w którym dominująca religią jest islam obserwujemy spadek aktywności terrorystycznej o prawie 20% w stosunku do pozostałych dni. Nie jest to jednak na tyle znaczna różnica, aby czuć się bezpiecznie w piątek będąc potencjalnym celem ataku w krajach islamskich.

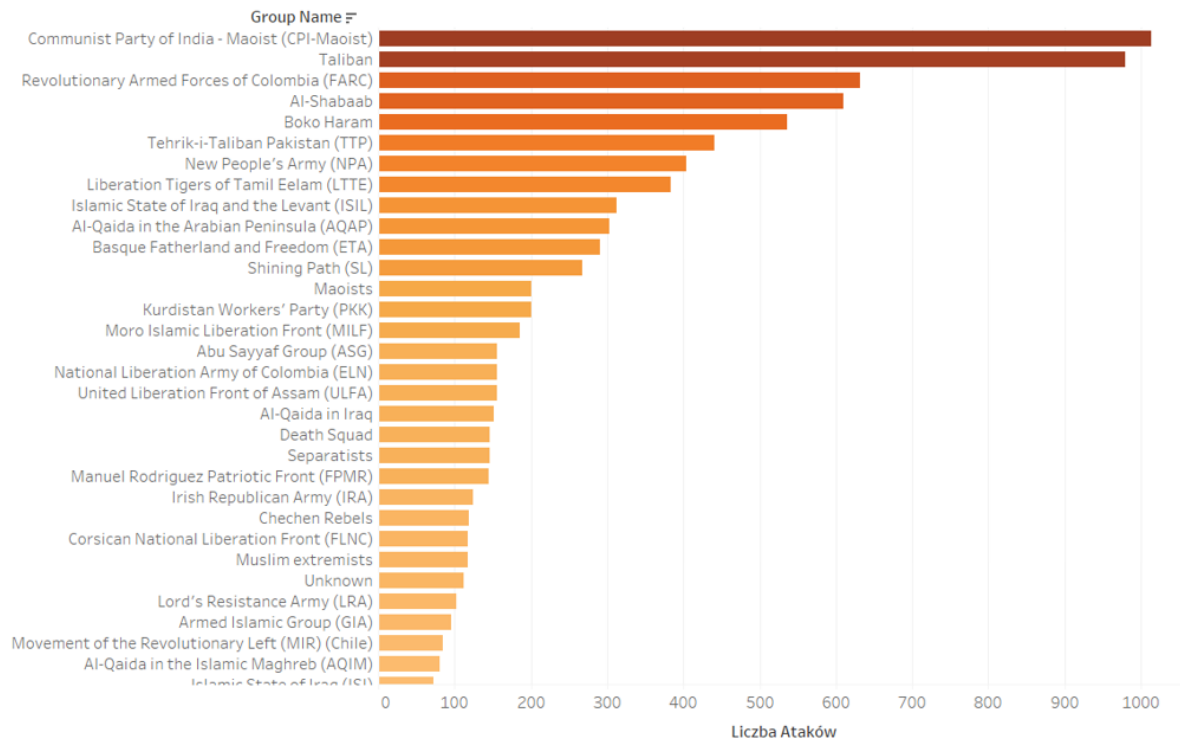
## Liczba ataków w poszczególnych miesiącach



Wykr. 11

Nie obserwujemy szczególnej różnicy w liczbie ataków w poszczególnych miesiącach. Podobnie wyniki uzyskiwałem sprawdzając te dane w kontekście poszczególnych regionów. Ciekawe mogło by być zestawienie świąt obchodzonych przez różne kultury, ale obecnie nie posiadam tego typu danych w hurtowni.

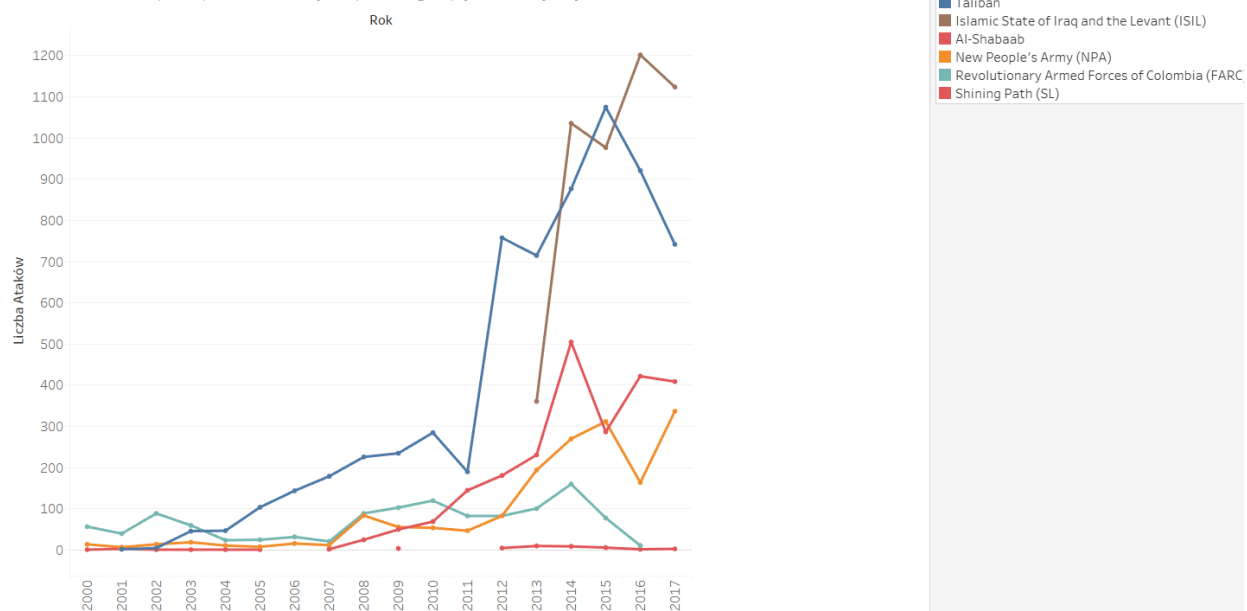
## Liczba potwierdzonych ataków przeprowadzonych przez grupy terrorystyczne



Wykr. 12

Na przestrzeni lat najaktywniejszymi grupami terrorystycznymi były **CPI-Maoist** oraz **Talibowie**. Jednak ten wykres nie pozwoli na określenie czy są one nadal aktywnym zagrożeniem. Dlatego dla 7 najbardziej aktywnych grup należy przygotować wykres w zależności od czasu.

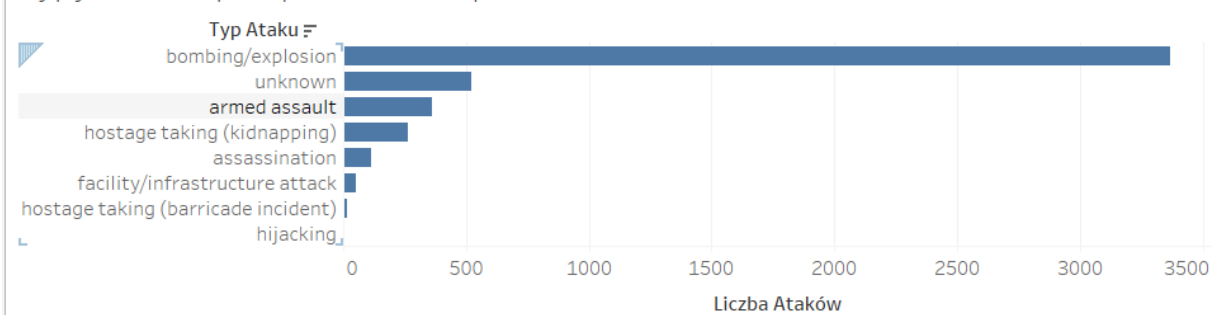
Liczba ataków przeprowadzonych przez grupy terrorystyczne



Wykr. 13

Z tego wykresu jednoznacznie wynika, że grupa **CPI-Maoist**, nie jest już zagrożeniem lub zmieniła nazwę. Największą grupą obok **Talibów** grupą pod względem liczby akcji jest grupa **ISIL**, która powstała stosunkowo niedawno i posiada trend wzrostowy. Należało by się baczenie przyrzeć aktywności tej grupy terrorystycznej, aby określić obszary jej działania główne cele i metody działania.

Typy ataków przeprowadzane przez ISIL



Wykr. 13

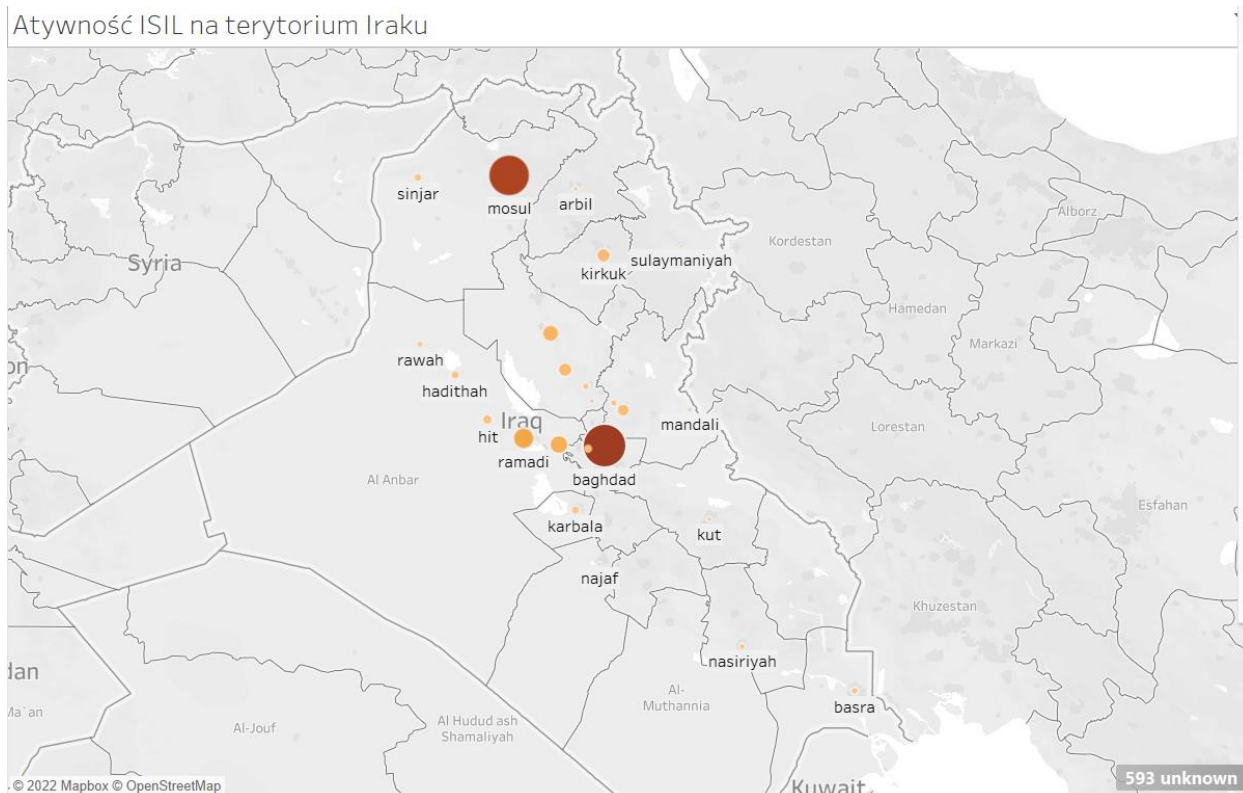
Z wykresu wynika, że ISIL specjalizuje się przede wszystkim w atakach z użyciem materiałów wybuchowych.

© 2022 Mapbox © OpenStreetMap

3 unknown

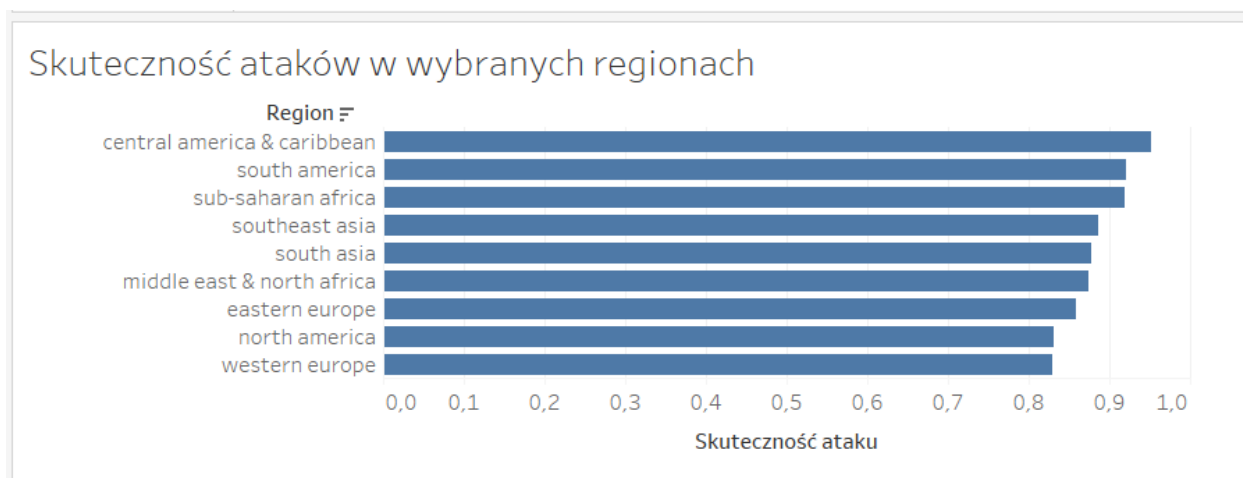
Pomimo tak dużej liczby ataków przeprowadzonych przez tą grupę ogranicza się ona tylko do swojego lokalnego regionu i nie jest zbyt zainteresowana atakami poza nim.





Wykr. 15

To co się wyróżnia na tej mapce to aktywność ISIL na terenie Mosulu. Aktywność terrorystów na tym terenie jest równie wielka co w stolicy (Bagdadzie) jednak jest to prawie 5-cio krotnie mniejsze miasto. Coś powoduje, że to właśnie tam znajduje się więcej celów dla ISIL.

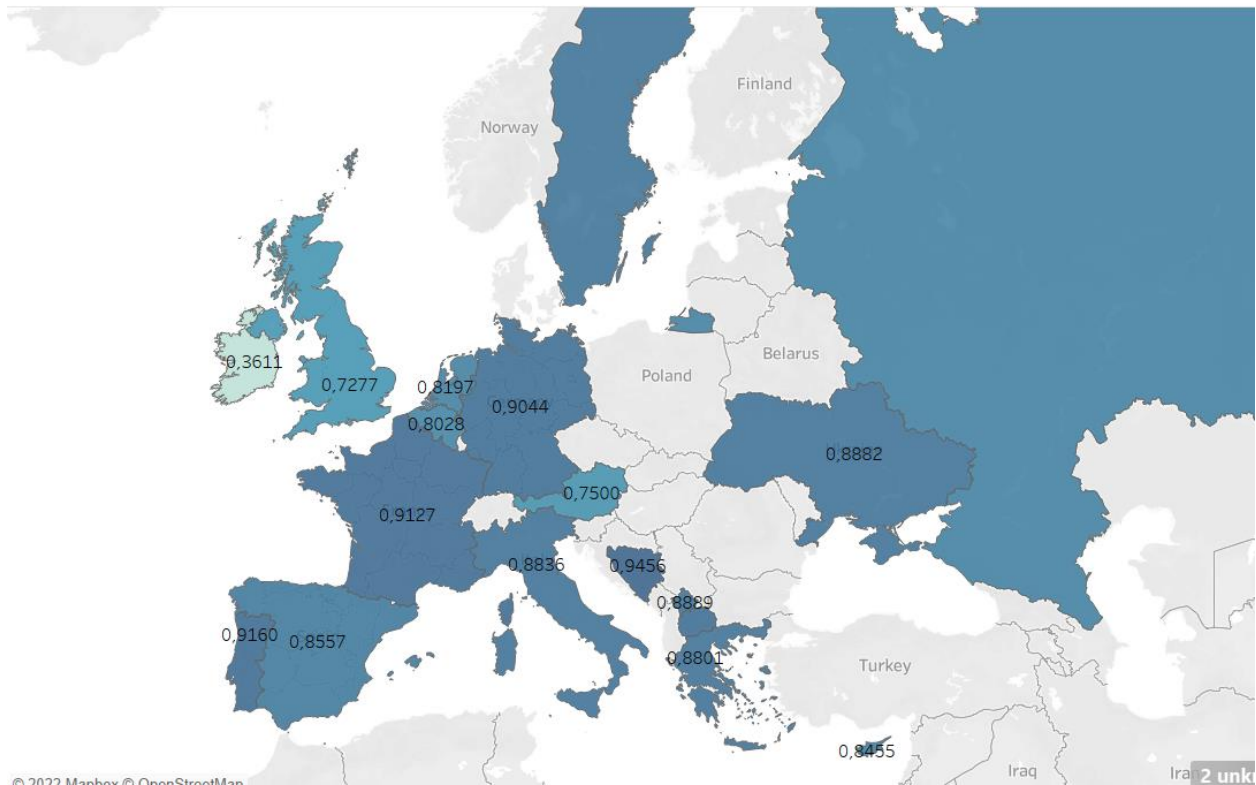


Wykr. 16

Na wykresie uwzględniam tylko regiony w których było ponad 1500 zarejestrowanych ataków.

Widzimy, że Europa zachodnia najlepiej radzi sobie z powstrzymywaniem ataków terrorystycznych, gdyż ponad 17% ataków zakończyło się niepowodzeniem. Mimo to 83% ataków zostało pomyślnie zrealizowanych. Powinniśmy się przyjrzeć jakiego typu czynnik mogą zwiększać skuteczność ataków na przykładzie Europy.

Skuteczność ataków w wybranych państwach europy



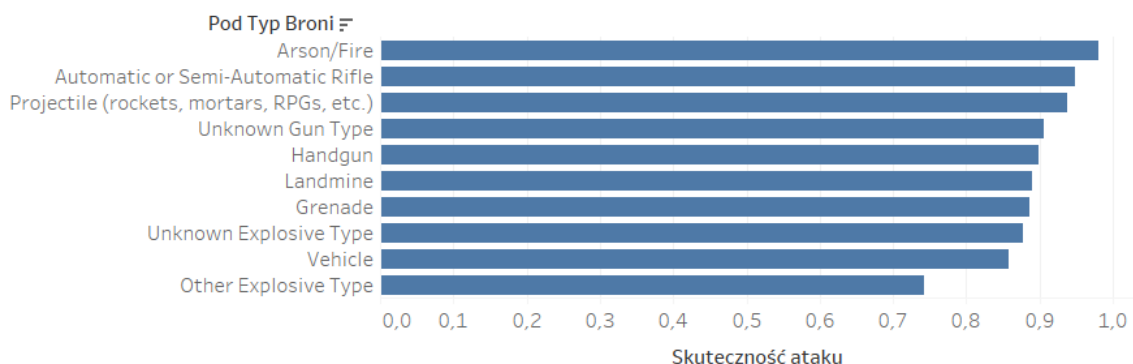
Wykr. 16

Uwzględniłam tylko kraje w których liczba ataków przekracza 100.

Na tle wszystkich krajów wyróżnia się wielka Brytania. Irlandia i Anglia posiadają wyjątkowo niską skuteczność przeprowadzanych tam ataków.

Choć w przypadku Irlandii wynik ten niepokojąco odstaje od pozostałych co może oznaczać, że wynik jest zaniżany przez jakiś czynnik.

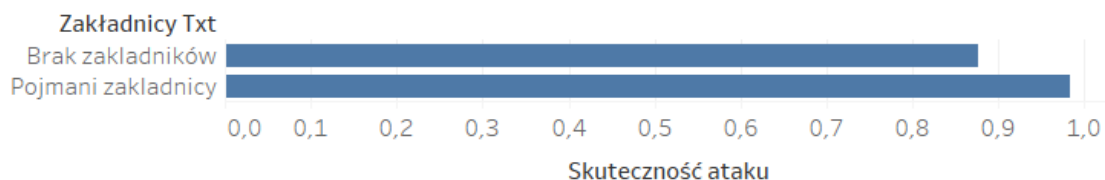
## Skuteczność ataków wybranymi brońmi w europie



Wykr. 17

Okazuje się, że najskuteczniejszym sposobem na akt terroru w Europie są **Podpalenia** oraz **Bronie ręczne**. Dają one niemal gwarancję sukcesu. Dużo niższą skuteczność ataku posiadają różnego rodzaju ładunki wybuchowe, które w końcu można wykryć i rozbroić.

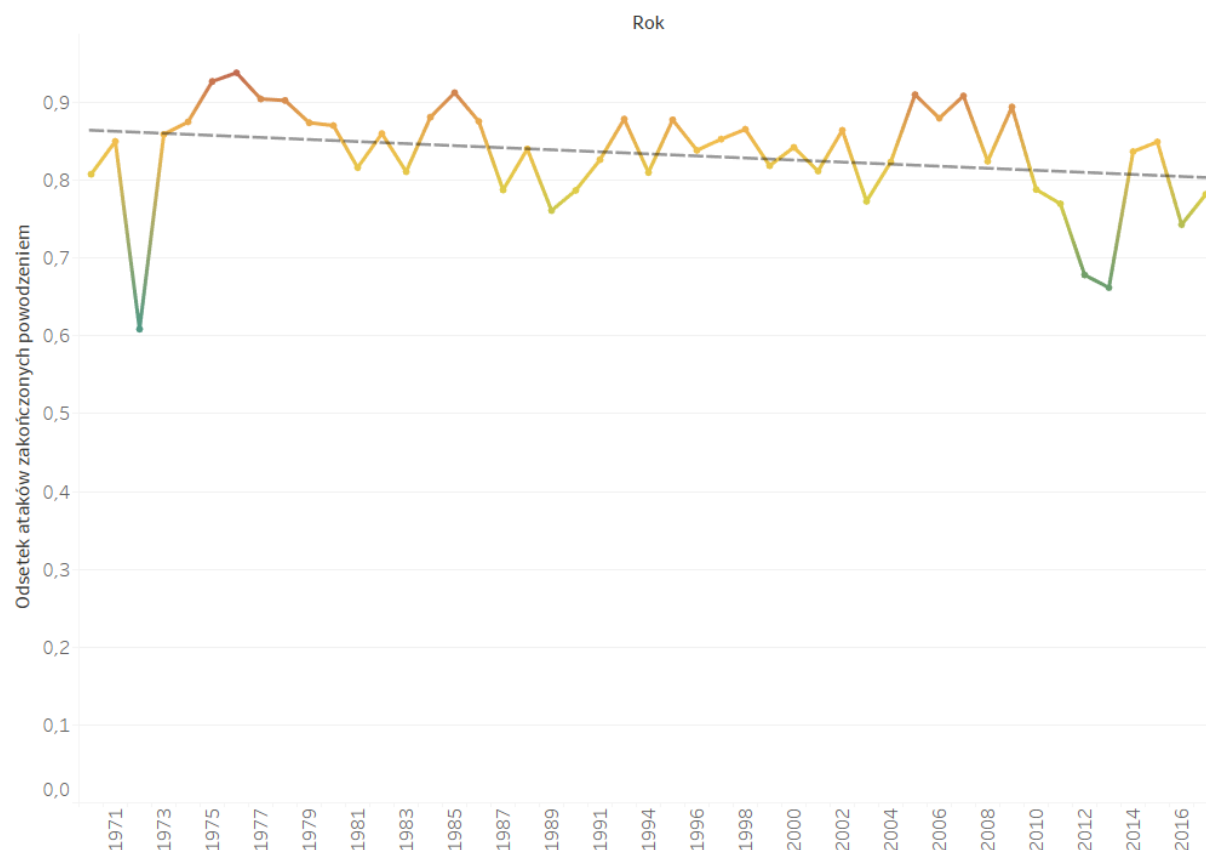
## Skuteczność ataków w zależności od posiadania zakładników



Wykr. 18

Posiadanie zakładników jest czynnikiem który zwiększa skuteczność ataków blisko o 10%. Opracowanie metod skutecznego odzyskiwania zakładników pozwoli na zmniejszenie skuteczności ataków.

## Skuteczność zamachów na przestrzeni lat



Wykr. 19

Na przestrzeni lat skuteczność zamachów powoli lecz sukcesywnie spada. Co sugeruje, że metody zapobiegania atakom rozwijają się szybciej niż metody ataku.

## 8.2. Podsumowanie – wnioski z analizy

Odpowiedzi na pytania zadane na początku dokumentu:

### • Czy są miejsca/cele bardziej narażone na atak terrorystyczny?

Tak. Zdecydowanie najbardziej narażonymi regionami są Afryka północna, „Bliski wschód” oraz południowa Azja. Pośród nich wyróżniają się kraje takie jak **Irak, Pakistan i Afganistan**. Prawdopodobnie z powodu świeżych i ciągle podsycanych przez religię nienawiści.

Niezależnie od regionu służby porządkowe tj. **policja** i **wojsko** są grupą podwyższonego ryzyka jeśli chodzi o ataki.

W krajach **zachodnich** obserwujemy, że częściej celami ataku stają się duże „złe” **korporacje** / „organizacje biznesowe”. Najprawdopodobniej za postawę wobec ludzi i środowiska.

Z kolei państwa **islam**u obierają za cel przede wszystkim **cywili**. Najprawdopodobniej za przynależność do innej grupy etnicznej, religijnej itp.

- **Jak liczba ataków terrorystycznych zmienia się w czasie?**

Na przestrzeni lat obserwujemy wzrost liczby akcji terrorystycznych. Choć kilku ostatnich obserwujemy spadek aktywności terrorystycznej, co napawa optymizmem.

Większa liczba rejestrowanych aktów terroryzmu może być związana z po prostu skuteczniejszą ich rejestracją, w związku z postępującą cyfryzacją. (Ktoś może zrobić zdjęcie i łatwo to udokumentować).

- **Czy są okresy czasu o zwiększonym ryzyku ataku w danym miejscu?**

W ogólności nie obserwujemy, żadnych znaczących relacji, choć najczęściej ataki przeprowadzane są na początku tygodnia. Możemy natomiast zaobserwować, że w regionach państw islamskich liczba ataków w **piątki** i po nich spada co może być spowodowane dniem modlitwy w religii islamu.

- **Jaka broń jest najbardziej śmiertelna?**

Zdecydowanie broń **automatyczna i powtarzalna(długa)**. Mimo to, że rejestrujemy ponad 5 razy więcej ataków z użyciem materiałów wybuchowych, to broń **automatyczna i powtarzalna(długa)** zbiera największe żniwo ofiar w stosunku do liczby ataków.

Największą ilość poszkodowanych generują różnego rodzaju materiały wybuchowe. Szczególnie dużo rannych powodują „**Samochody bomby**”. Prawdopodobnie ze względu na to, że w samochodzie łatwo ukryć bombę w porównaniu do montażu bomby bezpośrednio na miejscu.

- **Jak aktywność grup terrorystycznych zmieniała się na przestrzeni lat?**

Stosunkowo niedawno (2012) zaczęto rejestrować ataki organizowane przez **ISIL**. Ta organizacja w roku 2016 wyprzedziła w liczbie ataków nawet **Talibów**. A trend liczby ataków jest nadal wzrostowy.

Jednocześnie **FARC** w roku 2016 przestała funkcjonować.

Z kolei liczba ataków ze strony Talibów od 2015 ma dostać mocną tendencję spadkową.

- **Jakie ataki kończą się powodzeniem terrorystów?**

Powodzeniem kończą się najczęściej ataki w których pojmano zakładników.

Kolejnym czynnikiem jest dobór broni. Zwykle broń rażąca cele bezpośrednio (broń ręczna) zapewnia największą skuteczność terrorystom. Z kolei ładunki wybuchowe są znaczenie łatwiejsze „do powstrzymania” od uzbrojonych terrorystów.

Szczególnie skutecznym typem ataku w Europie okazują się być podpalenia. Pożar zawsze zdoła dokonać strat zanim uda się go opanować.

## 9. Wnioski końcowe z realizacji projektu

### 9.1. Problemy

W czasie kursu problemem było to, że najpierw zajmowaliśmy się MS SSAS, a dopiero na końcowych zajęciach procesem ETL w MS SSIS. Moim zdaniem powinniśmy wcześniej zająć się procesem ETL na zajęciach. Dało by to więcej czasu na zapoznanie się z narzędziem MS SSIS i pozwoliło na pracę przyrostowo nad projektem.

Podczas realizacji projektu następowało wiele pomyłek związanych z dużą liczbą kolumn. Nieprzyjazny dla użytkownika system wyboru wartości w produktach Microsoftu bez możliwości sortowania i filtrowania utrudniał pracę.

Podczas analizy wyników okazywało się, że pewne wartości i miary są zbędne, a pewnych brakuje. Zmiana modelu koncepcyjnego wymagała ponownej edycji wszystkich następujących procesów i ich ponownego wykonania.

Tabela nie pozwoliła mi na użycie długości i szerokości geograficznej na mapach byłem zmuszony do używania miast, państw i regionów.

## 9.2. Pozyskana wiedza i doświadczenie

Podczas realizacji projektu nauczyłem się przede wszystkim, czego wymagać od danych i które z nich mogą okazać się przydatne podczas realizacji projektów.

W czasie zajęć nabrałem także doświadczenie w używaniu MS SSIS oraz SSAS. Mam wrażenie, że przez słabą dokumentację i brak tutoriali dobrej jakości w Internecie nie bym był w stanie sam zacząć nauki z tymi narzędziami sam.

Najprzyjemniejszym i najbardziej bezproblemowym z narzędzi w całym projekcie był Tabelau. W łatwy sposób można było znaleźć jak projektować określonego typu wykresy, dużo bardziej atrakcyjne wizualnie od tych Excellowskich.

## 10. Źródła informacji użyte w etapie analizy danych

Podczas Analizy Danych wspomagałem się następującymi źródłami:

1. Piątek dzień święty islamu: <https://pl.wikipedia.org/wiki/Pi%C4%85tek>
2. Obecność amerykańców w Iraku na przestrzeni lat:  
<https://www.defensepriorities.org/explainers/leaving-iraq-serves-us-interests>