Hurtownie danych – Sprawozdanie z etapu 2 projektu.

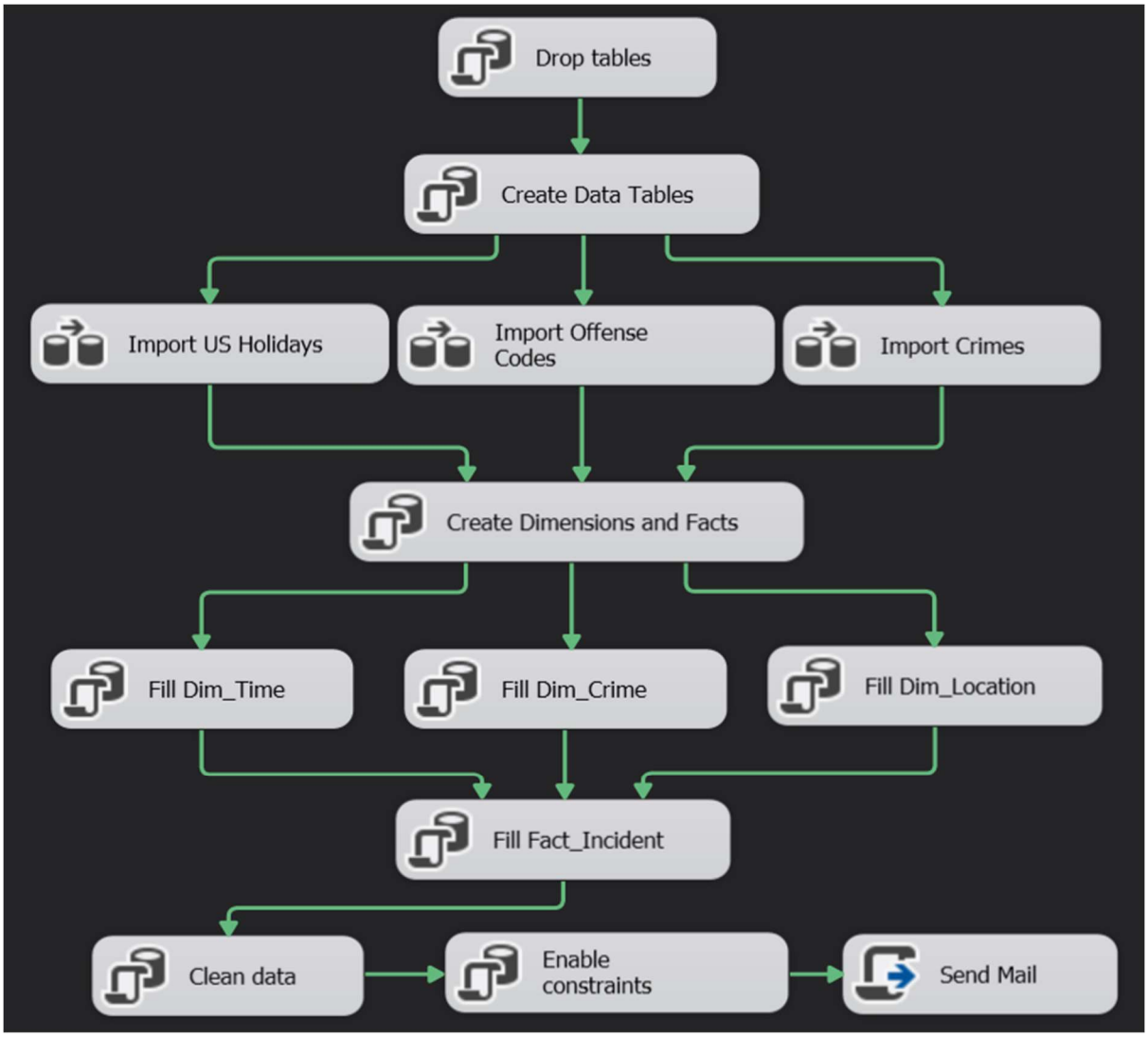
PWr. WIZ, Informatyka, Data: 12.05.2020

Student	Email: 242493@student.pwr.edu.pl	Ocena
Indeks	242493	
Imię	Arkadiusz	
Nazwisko	Rasz	

Spis treści

Proces ETL

Schemat



Opisy poszczególnych kroków

Drop tables

Tabele bazy danych usuwane są (w przypadku kiedy istnieją), aby przygotować środowisko do wprowadzenia nowych danych.

```
DROP TABLE IF EXISTS [Rasz].[Fact_Incident];
DROP TABLE IF EXISTS [Rasz].[Dim_Crime];
DROP TABLE IF EXISTS [Rasz].[Dim_Location];
DROP TABLE IF EXISTS [Rasz].[Dim_Time];

DROP TABLE IF EXISTS [Rasz].[UsHolidays];
DROP TABLE IF EXISTS [Rasz].[OffenseCodes];
DROP TABLE IF EXISTS [Rasz].[Crimes];
```

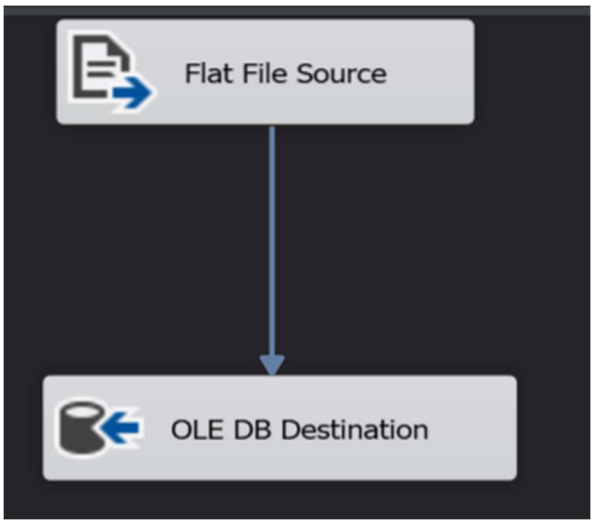
Create Data Tables

Tworzone są tabele, do których zostaną zaimportowane z plików, zawierających dane do utworzenia hurtowni danych.

```
CREATE TABLE [Rasz].[UsHolidays] (
    [Id] numeric(18,0),
    [Date] datetime,
    [Holiday] varchar(50)
);

CREATE TABLE [Rasz].[OffenseCodes] (
    [Code] int,
    [Name] nvarchar(255),
);
CREATE TABLE [Rasz].[Crimes] (
    [INCIDENT_NUMBER] nvarchar(255),
    [OFFENSE_CODE] nvarchar(50),
    [OFFENSE_CODE_GROUP] nvarchar(255),
    [OFFENSE_DESCRIPTION] nvarchar(255),
    [DISTRICT] nvarchar(255),
    [REPORTING_AREA] int,
    [SHOOTING] nvarchar(2),
    [OCCURRED_ON_DATE] datetime,
    [YEAR] int,
    [MONTH] smallint,
    [DAY_OF_WEEK] nvarchar(255),
    [HOUR] smallint,
    [UCR_PART] nvarchar(255),
    [STREET] nvarchar(255),
    [Lat] float,
    [Long] float,
    [Location] nvarchar(255)
);
```

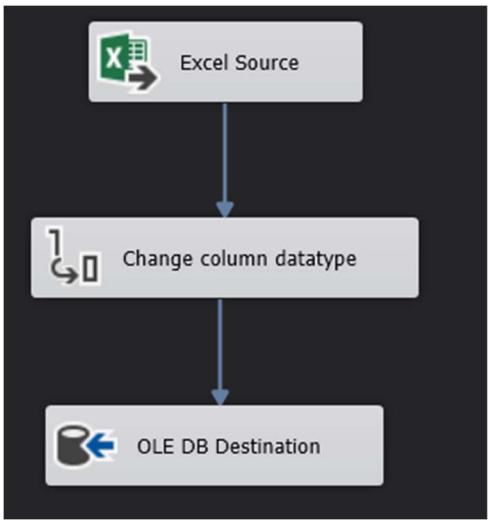
Import US Holidays

Dane o świętach narodowych wczytywane są z pliku .csv z użyciem elementu Flat File Source. Typy kolumn zostały w nim już odpowiednio zmapowane.



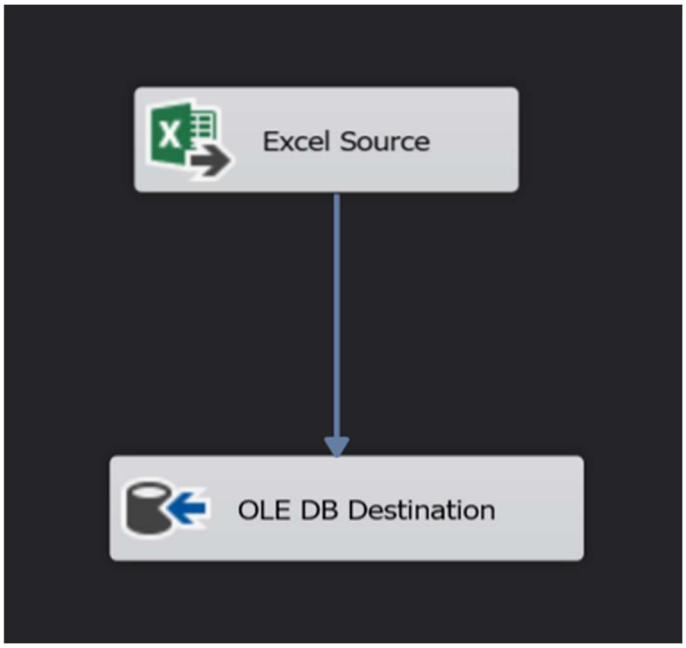
Import Offense Codes

Dane pobierane są z pliku Excel. Przed zapisaniem danych do odpowiedniej tabeli konieczne było przekonwertowanie typu danych columny Code (mogło to być również wykonane w elemencie Excel Source).



Import Crimes

Dane pobierane z pliku Excel, typy kolumn zostały odpowiednio zmodyfikowane.
Jest to najbardziej kosztowny krok w całym procesie, który powoduje dużo problemów programowi Visual Studio oraz dodatków.



Create Dimensions and Facts

Tworzone są tabele wymiarów oraz faktów, bez ograniczeń na tym poziomie.

```
CREATE TABLE [Rasz].[Dim_Time](
	[Time] INT NOT NULL PRIMARY KEY,
	[Year] SMALLINT NOT NULL,
	[Quarter] TINYINT NOT NULL,
	[Month] TINYINT NOT NULL,
	[DayInMonth] TINYINT NOT NULL,
	[Hour] TINYINT NOT NULL,
	[HolidayName] NVARCHAR(50) NULL,
	[MonthName] NVARCHAR(10) NOT NULL,
	[Weekday] NVARCHAR(10) NOT NULL,
);

CREATE TABLE [Rasz].[Dim_Crime](
	[Code] SMALLINT NOT NULL,
	[Category] NVARCHAR(50) NOT NULL,
	[Detail] NVARCHAR(50) NULL
);

CREATE TABLE [Rasz].[Dim_Location](
	[Id] INT NOT NULL,
	[District] NVARCHAR(10),
	[Street] NVARCHAR(50),
);

CREATE TABLE [Rasz].[Fact_Incident](
	[IncidentNumber] NVARCHAR(20) NULL,
	[OffenseCode] SMALLINT NULL,
	[Time] INT NOT NULL,
	[ReportingArea] SMALLINT NULL,
	[Shooting] TINYINT NULL,
	[UCRPart] NVARCHAR(20) NULL,
	[Location] INT NULL,
	[Latitude] DECIMAL(8,2) NULL,
	[Longitude] DECIMAL(8,2) NULL
);
```

Fill Dim_Time

Tabela czasu wypełniana jest na podstawie czasów zawartych w danych o incydentach. Wypełnianie jej jest podobne do tego na listach zadaniowych, lecz jest rozszerzone dodatkowo o godzinę. Klucz główny składa się dodatkowo z dwóch cyfr reprezentujących godzinę.

Na tym etapie dołączone są również nazwy świąt narodowych, dla każdej zgadzającej się daty.

```
with Dates(Date) as (
	select [OCCURRED_ON_DATE]
	from [Rasz].[Crimes]
),
Months(MonthNum, MonthName) as (
	select *
	from (values (1, 'January'), (2, 'February'),
		(3, 'March'), (4, 'April'),
		(5, 'May'), (6, 'June'),
		(7, 'July'), (8, 'August'),
		(9, 'September'), (10, 'October'),
		(11, 'November'), (12, 'December')) AS M(C1, C2)
),
Weekdays(DayNum, Weekday) as (
	select *
	from (values (2, 'Monday'),
		(3, 'Tuesday'),
		(4, 'Wednesday'),
		(5, 'Thursday'),
		(6, 'Friday'),
		(7, 'Saturday'),
		(1, 'Sunday')) as W(C1, C2)
)
INSERT INTO [Rasz].[Dim_Time](Time, Year, Quarter, Month, DayInMonth, Hour,
				HolidayName, MonthName, Weekday)
(
	select distinct
		YEAR(D.Date) * 1000000 +
			MONTH(D.Date) * 10000 +
				DAY(D.Date) * 100 +
					DATEPART(HOUR, D.Date),
		YEAR(D.Date),
		(MONTH(D.Date)-1) / 3 + 1,
		MONTH(D.Date),
		DAY(D.Date),
		DATEPART(HOUR, D.Date),
		H.Holiday,
		MonthName,
		Weekday
	from Dates D
	join Months M
		on MONTH(D.Date) = M.MonthNum
	join Weekdays W
		on DATEPART(WEEKDAY, D.Date) = W.DayNum
		left join [Rasz].[UsHolidays] H
		on YEAR(D.Date) * 1000000 +
			MONTH(D.Date) * 10000 +
				DAY(D.Date) * 100 =
			YEAR(H.Date) * 1000000 +
			MONTH(H.Date) * 10000 +
				DAY(H.Date) * 100
);
```

Fill Dim_Crime

Zaimportowane dane o kategoriach incydentów są wprowadzane do tabeli wymiaru. Kolumna posiadająca nazwę kategorii przestępstwa jest dzielona na 2 części na znaku '-', co pozwala utworzyć hierarchię konieczną do późniejszej analizy danych.

```
INSERT INTO [Rasz].[Dim_Crime](Code, Category, Detail) (
   SELECT
      [Code],
         CASE WHEN CHARINDEX('-', [Name]) > 0
            THEN RTRIM(SUBSTRING([Name], 1, CHARINDEX('-', [Name])-1))
            ELSE RTRIM([Name]) end Category,
         CASE WHEN CHARINDEX('-', [Name]) > 0
            THEN LTRIM(SUBSTRING([Name], CHARINDEX('-', [Name])+1, len([Name])))
            ELSE NULL end Detail
   FROM [Rasz].[OffenseCodes]
);
```

Fill Dim_Location

Dystrykty i nazwy ulic z tabeli głównej są importowane do tabeli wymiaru. Dodana jest również kolumna identyfikująca krotkę, wygenerowana przez numer wiersza.

```
INSERT INTO [Rasz].[Dim_Location](Id, District, Street) (
       SELECT ROW_NUMBER() OVER (ORDER BY [DISTRICT], [STREET] ASC) ID,
       [DISTRICT],
       [STREET]
       FROM (
              SELECT DISTINCT [District], [STREET]
              FROM [Rasz].[Crimes]
       ) X
)
```

Fill Fact_Incident

Do tabeli faktów przepisywane są dane z tabeli zaimportowanej z pliku, z odpowiednią już wartością w kolumnie określającą czas oraz lokalizację.

```
INSERT INTO [Rasz].[Fact_Incident]([IncidentNumber], [OffenseCode], [Time],
       [ReportingArea], [Shooting], [UCRPart], [Location], [Latitude], [Longitude]) (
       SELECT
              [Incident_Number],
              [Offense_Code],
              YEAR([Occurred_On_Date]) * 1000000 +
                     MONTH([Occurred_On_Date]) * 10000 +
                            DAY([Occurred_On_Date]) * 100 +
                                   DATEPART(HOUR, [Occurred_On_Date]),
              [Reporting_Area],
              [Shooting],
              [UCR_Part],
              L.[Id],
              [Lat],
              [Long]
       FROM [Rasz].[Crimes] C
       LEFT JOIN [Rasz].[Dim_Location] L
              ON C.[District] = L.[District]
              AND C.[Street] = L.[Street]
)
```

Clean Data

Dodanie wartości do komórek o wartości NULL

```
UPDATE [CrimesInBoston].[Rasz].[Fact_Incident]
SET [UCRPart] = 'Unknown'
WHERE [UCRPart] IS NULL;

UPDATE [CrimesInBoston].[Rasz].[Dim_Location]
SET [District] = 'Other'
WHERE [District] IS NULL;

UPDATE [CrimesInBoston].[Rasz].[Dim_Location]
SET [Street] = 'Other'
WHERE [Street] IS NULL;

UPDATE [CrimesInBoston].[Rasz].[Dim_Time]
SET [HolidayName] = 'None'
WHERE [HolidayName] IS NULL;
```

Usunięcie powtarzających się krotek dotyczących tej samej kategorii przestępstwa. Były one w tej postaci już w importowanym pliku, a powtórzenia różnią się zwykle dodatkowym przecinkiem czy spacją

```
DELETE X FROM (
  SELECT *, rn=row_number() OVER (PARTITION BY [Code] ORDER BY [Category], [Detail])
  FROM [CrimesInBoston].[Rasz].[Dim_Crime]
) X
WHERE rn > 1;
```

Dodanie kategorii przestępstwa reprezentującą brak danych – część kodów podanych dla incydentów nie jest zdefiniowana w kategoriach przestępstw

```
INSERT INTO [CrimesInBoston].[Rasz].[Dim_Crime]
VALUES(0, 'Unknown', 'Unknown');
```

Zastąpienie nieznanych danych referencją na utworzoną krotkę

```
UPDATE [CrimesInBoston].[Rasz].[Fact_Incident]
SET [OffenseCode] = 0
WHERE [OffenseCode] IN (
       SELECT I.[OffenseCode]
       FROM [CrimesInBoston].[Rasz].[Fact_Incident] I
       LEFT JOIN [CrimesInBoston].[Rasz].[Dim_Crime] C
              ON I.OffenseCode = C.Code
       WHERE C.Code IS NULL
);
```

Enable Constraints

Na tym etapie dodawane są już klucze główne oraz obce.

```
ALTER TABLE [CrimesInBoston].[Rasz].[Dim_Crime]
ADD PRIMARY KEY([Code]);

ALTER TABLE [CrimesInBoston].[Rasz].[Dim_Location]
ADD PRIMARY KEY([Id]);


ALTER TABLE [CrimesInBoston].[Rasz].[Fact_Incident]
ADD CONSTRAINT FK_FactIncident_Time_DimTime_Time FOREIGN KEY([Time])
REFERENCES [CrimesInBoston].[Rasz].[Dim_Time]([Time]);

ALTER TABLE [CrimesInBoston].[Rasz].[Fact_Incident]
ADD CONSTRAINT FK_FactIncident_Location_DimLocation_Id FOREIGN KEY([Location])
REFERENCES [CrimesInBoston].[Rasz].[Dim_Location]([Id]);

ALTER TABLE [CrimesInBoston].[Rasz].[Fact_Incident]
ADD CONSTRAINT FK_FactIncident_OffenseCode_DimCrime_Code FOREIGN KEY([OffenseCode])
REFERENCES [CrimesInBoston].[Rasz].[Dim_Crime]([Code]);
```

Send Mail

Do użytkownika wysyłany jest mail z potwierdzeniem zakończenia procesu. Skrypt napisany w c#:

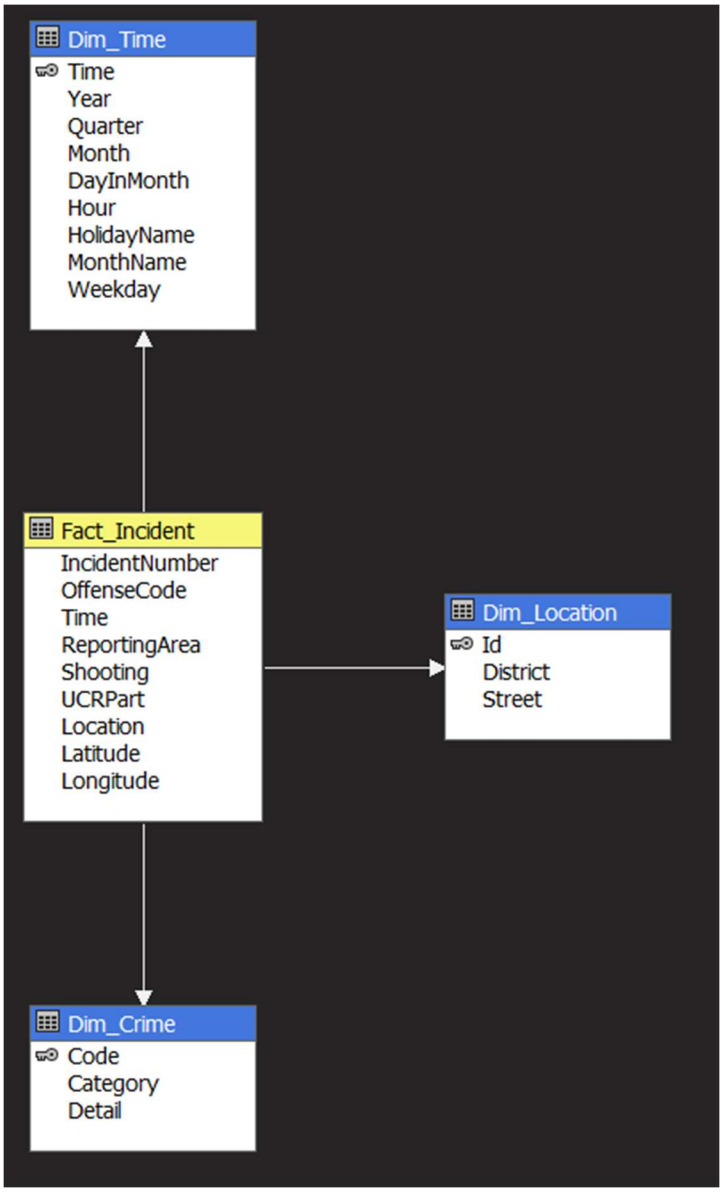
```
var from = new MailAddress("arkadiusz.rasz@gmail.com", "Arkadiusz Rasz");
var to = new MailAddress("arkadiusz.rasz@gmail.com");
var password = "***********";
var smtp = new SmtpClient
{
   Host = "smtp.gmail.com",
   Port = 587,
   EnableSsl = true,
   DeliveryMethod = SmtpDeliveryMethod.Network,
   UseDefaultCredentials = false,
   Credentials = new NetworkCredential(from.Address, password),
   Timeout = 20000
};

var msg = new MailMessage
{
   IsBodyHtml = true,
   Subject = "SSIS: Task Succeeded",
   Body = "Message body",
   From = from
};
msg.To.Add(to);
{
   smtp.Send(msg);
}

TaskResult = (int)ScriptResults.Success;
```

Uwagi

Niestety, po uruchomieniu paczki z procesem ETL, program Visual Studio zamyka się po pewnym czasie bez żadnego błędu, co zatrzymuje postęp procesu. Dzieje się to w kroku importowania danych incydentów. Proces należy powtórzyć kilka razy, aby pomyślnie się zakończył. Podobny problem pojawił się już na listach zadaniowych.

Kostka

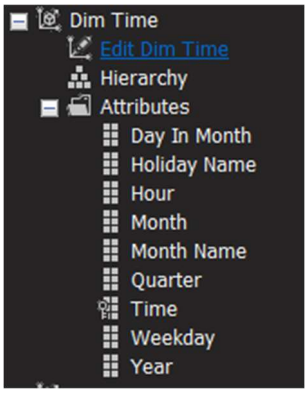
Struktura kostki



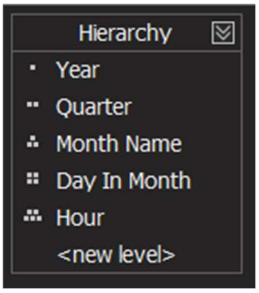
Wymiary

Dim_Time

Jest ro wymiar czasu posiadający hierarchiczną strukturę kolumn, od roku aż do godziny. Dodatkowo dostępne są kolumny o typie ciągu znaków, jak nazwa miesiąca oraz dzień tygodnia. Jest również kolumna Holiday Name, która oznacza nazwę święta narodowego odbywającego się w dany dzień. Jej wartość zwykle wynosi ‘None’.
Time jest kluczem głównym tabeli.

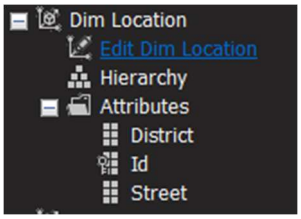


Hierarchia:



Dim_Location

Wymiar zawiera informacje o ulicę zdarzenia oraz dystrykcie, w którym się znajduje. W przypadku braku danych, ich wartość jest równa ‘Unknown’.
Id jest kluczem głównym tabeli.

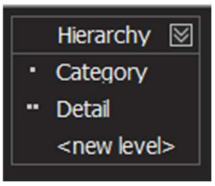


Hierarchia:

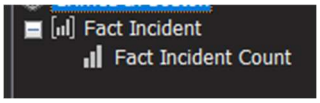
Dim_Crime

Wymiar dotyczy kategorii przestępstwa. Atrybut Category oznacza ogólną kategorię, a Detail uszczegółowienie.
Code jest kluczem głównym tabeli.

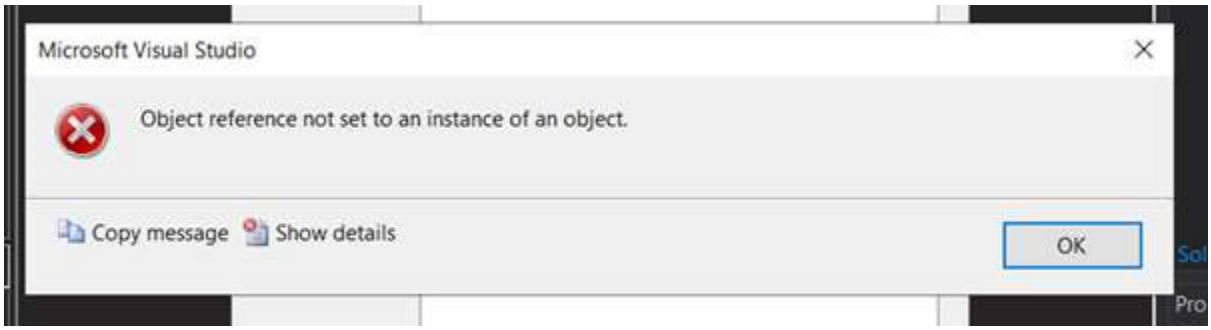
Hierarchia:



Miary



Niestety, główną miarą, która będzie wykorzystywana jest liczba incydentów. Próba dodania innych miar, związanych z np. liczbą strzelanin powodowała niespodziewane błędy, których identyfikacja nie była możliwa. Detale błędu również nie były pomocne.

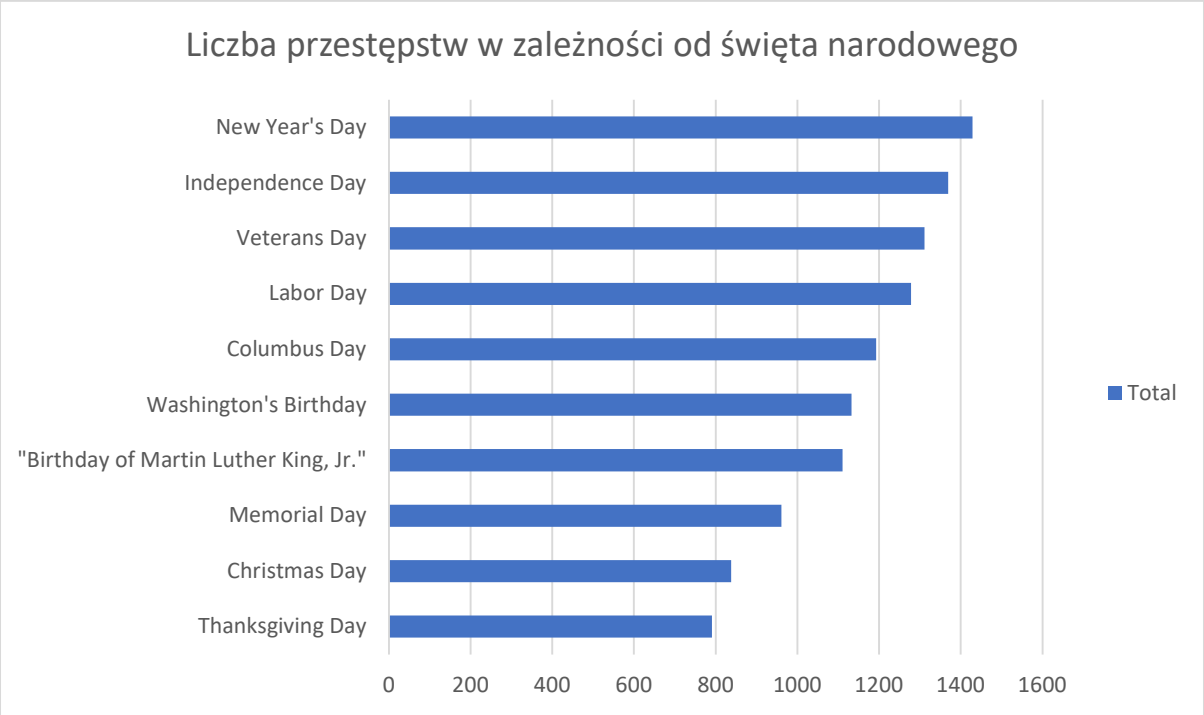


Analiza danych

Liczba zanotowanych przestępstw w mieście w zależności od roku i kwartału

Liczba przestępstw zdaje się utrzymywać na stałym poziomie przez ostatnie 5 lat. Jest to zgodne z rozwojem miasta – populacja Bostonu w przeciągu tych 5 lata powiększyła się o zaledwie 2%.

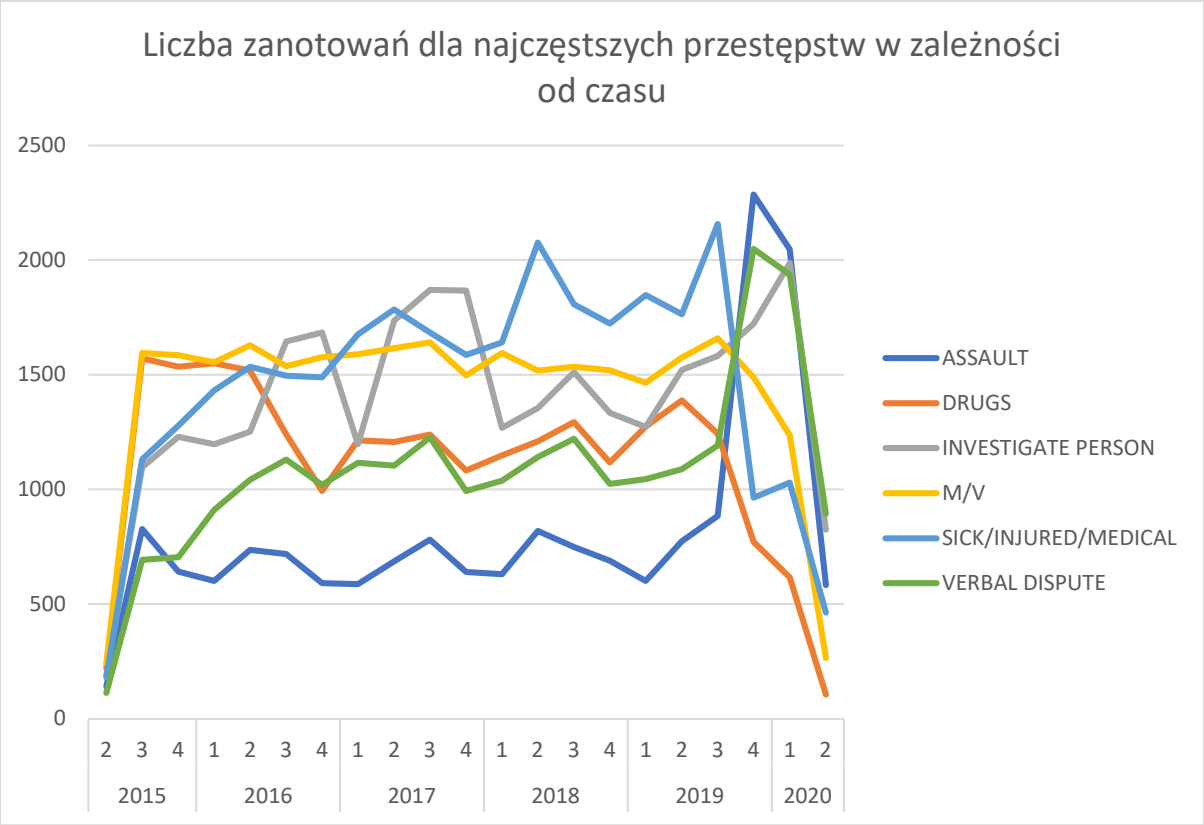
Liczba przestępstw w zależności od święta narodowego



Widoczna jest zależność przestępstw od świąt narodowych. Najwięcej z nich wydarzyło się w Nowy Rok oraz święto niepodległości – można przypuszczać, że jest to związane z niezachowaniem ostrożności podczas puszczania fajerwerków jak i w uczestnictwie w pochodach.

Najmniej przypadków zdarzyło się w wigilię oraz święto dziękczynienia.

Liczba zanotowań dla najczęstszych przestępstw w zależności od czasu

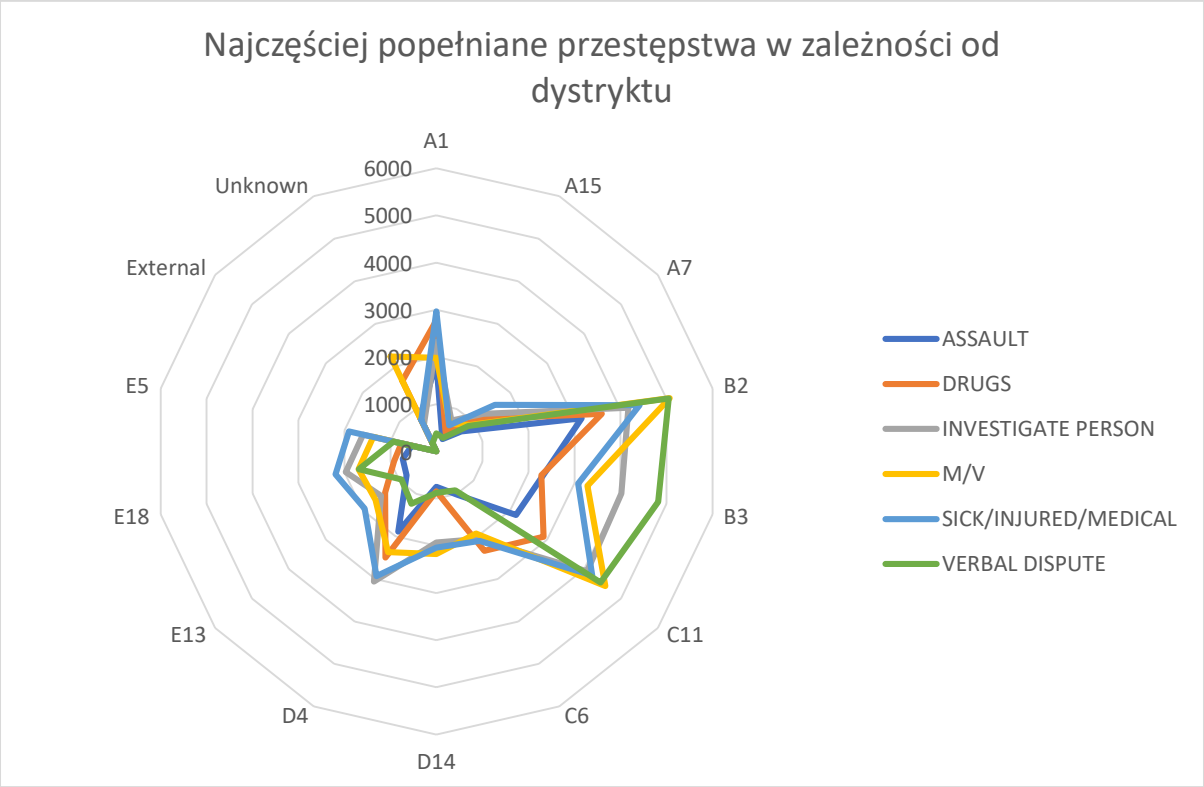


Wykres przedstawia 6 najczęstszych przestępstw zanotowanych w bazie danych. Widoczny jest wyraźny skok w napadach oraz sporach ustnych. Może być na to wiele powodów, jeden prawdopodobny to zbliżające się wybory prezydenckie.

Widać również spadek w przestępstwach związanych z użyciem narkotyków. Trudno wywnioskować co jest tego powodem, lecz nie jest to związane z legalizacją marihuany. Stało się to bowiem już w roku 2016.

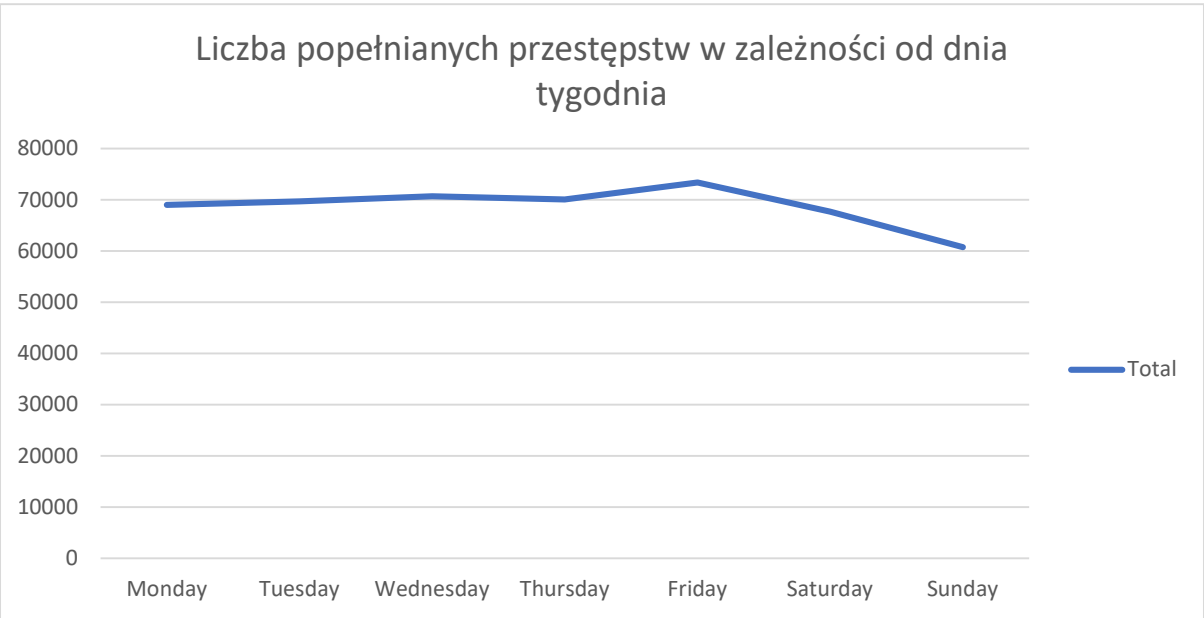
Pozostałymi najczęstszymi incydentami są sprawdzanie osób, zdarzenia drogowe oraz zdarzenia związane ze zdrowiem.

Najczęściej popełniane przestępstwa w zależności od dystryktu



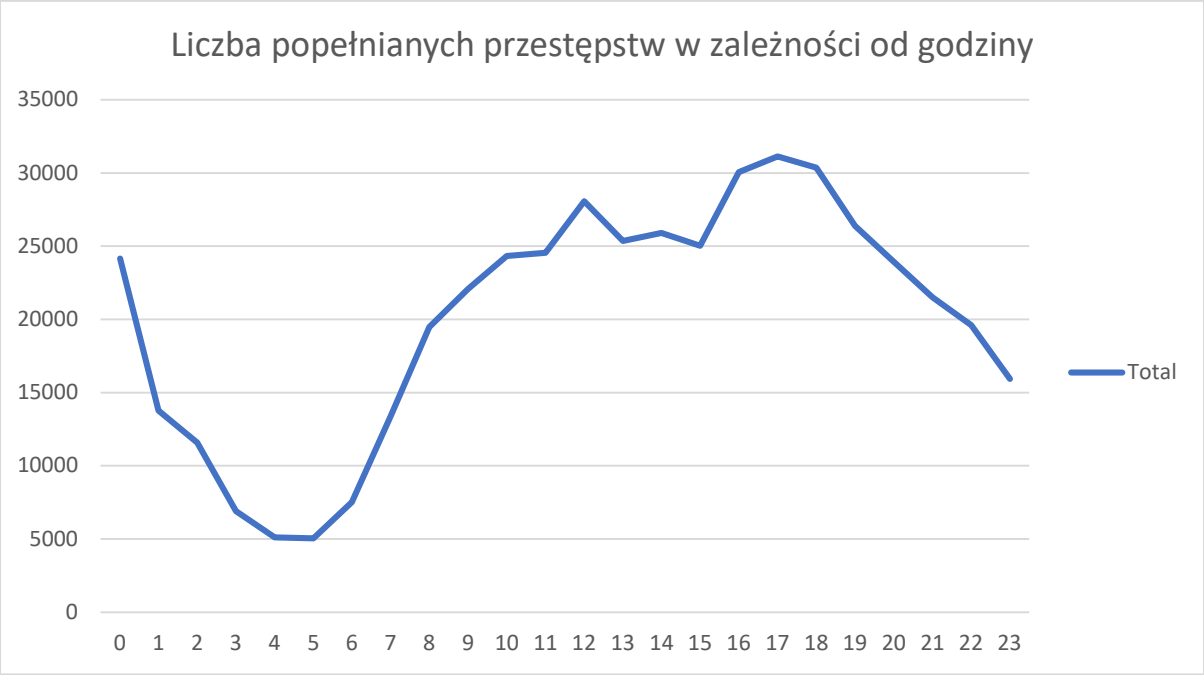
Wykres przedstawia podane wyżej przestępstwa w zależności od dystryktu. Widoczne jest, że dystrykty B2, B3, C11 mają najwięcej przypadków, co może oznaczać, że są najniebezpieczniejsze. Pomocne do ocenienia danych byłyby dane o populacji dystryktów, lecz nie udało się takich odnaleźć.

Liczba popełnianych przestępstw w zależności od dnia tygodnia



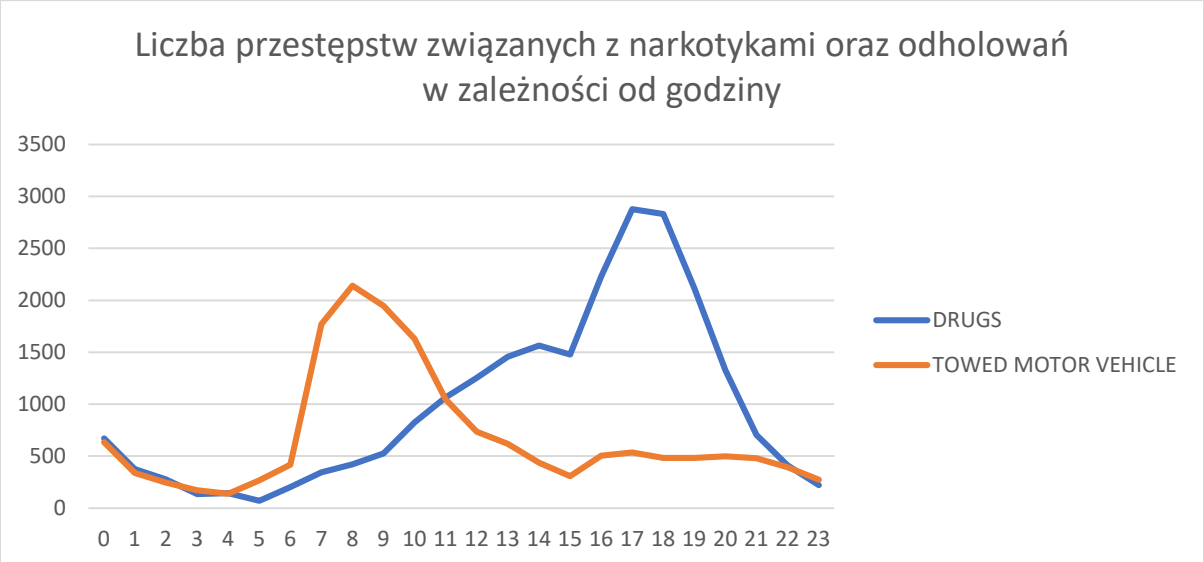
Liczba przestępstw jest największa w środy, po czym spada w weekend, gdzie jest najmniejsza w całym tygodniu.

Liczba popełnianych przestępstw w zależności od godziny



Najmniej przypadków raportowanych jest w okolicach 4 i 5 nad ranem, a najwięcej – 12 godzin później. Wykres nie przedstawia nic nowego – liczba przypadków jest ściśle powiązana z aktywnością mieszkańców, po której można spodziewać się podobnego wykresu. Ciekawe jest jednak jego podobieństwo do sinusoidy.

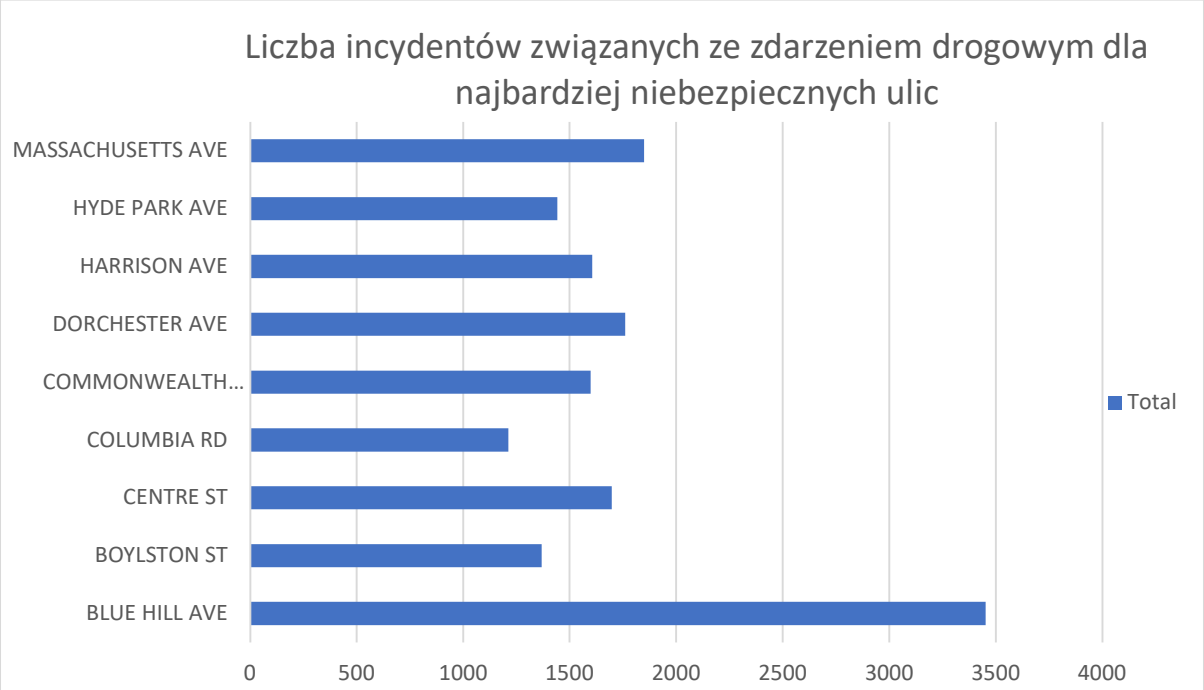
Liczba przestępstw związanych z narkotykami oraz odholowań w zależności od godziny



Poprzedni wykres przedstawiono dla dwóch, ciekawych kategorii. Widać, że liczba odholowań samochodów największa jest w godzinach porannych, co może wskazywać na liczbę patrolów w tych godzinach.

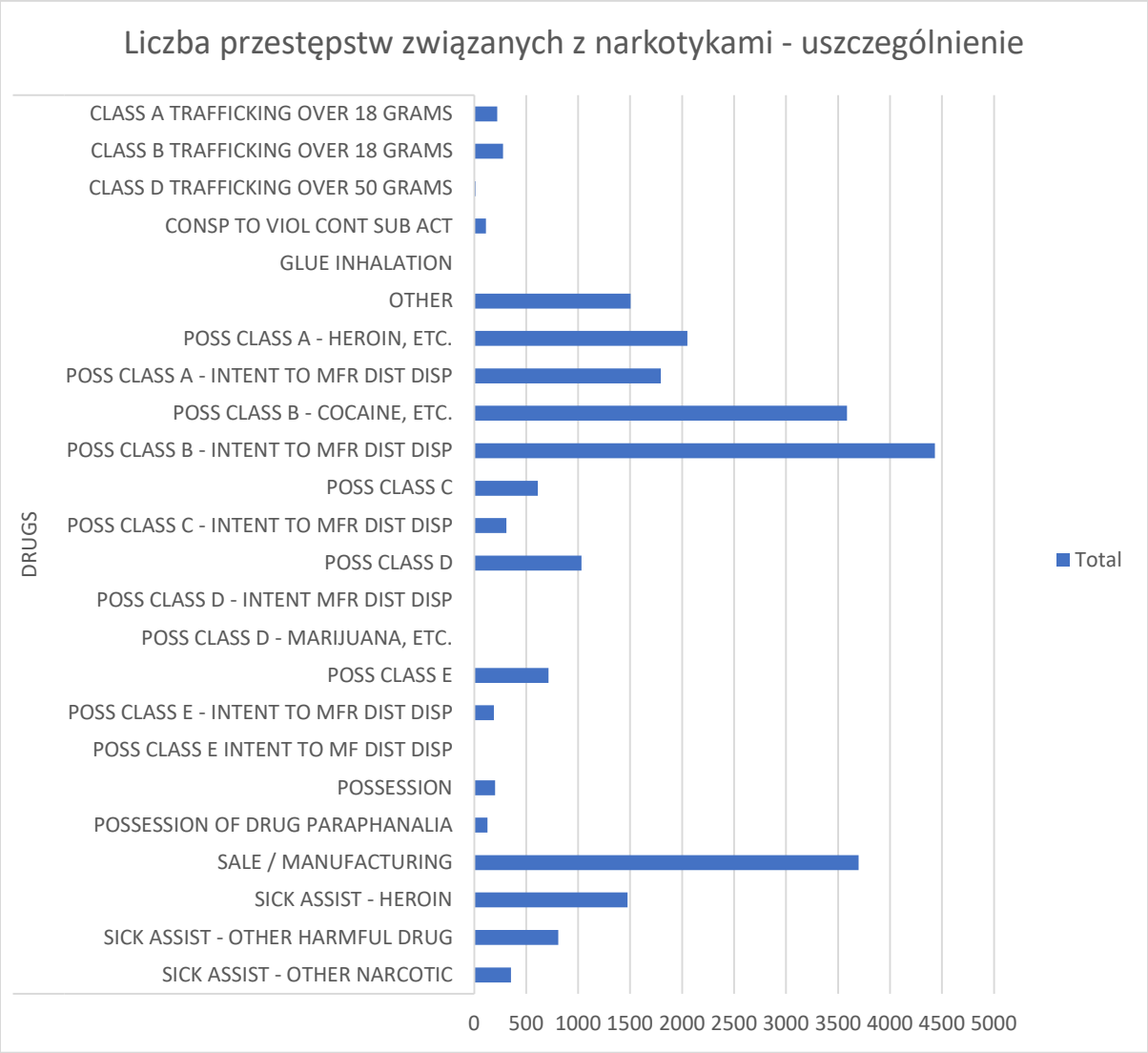
Użycie narkotyków jest największe w godzinach popołudniowych, a w nocy jest niemal nieistniejące.

Liczba incydentów związanych ze zdarzeniem drogowym dla najbardziej niebezpiecznych ulic



Wykres przedstawia ulice, na których zanotowano najwięcej przypadków drogowych. W rankingu prowadzi Blue Hill Ave – ulica o zaledwie 4 milach. Jest to jednak obszar uważany za jeden z najniebezpieczniejszych obszarów w mieście.

Liczba przestępstw związanych z narkotykami - uszczególnienie



Wykres przedstawia szczegóły przestępstw związanych z narkotykami. Najwięcej przypadków dotyczy ich posiadania, rozprowadzania oraz wytwarzania. Najczęstszą kategorią narkotyków jest kategoria B, do której zalicza się np. kokaina.

Wnioski

W planach były do przygotowania dodatkowe wykresy, związane z danymi o strzelaninach oraz przedstawienia danych na mapie. Z powodu problemów związanych z oprogramowaniem, jak i brakiem czasu, nie udało się tych danych przedstawić.

Niemniej jednak, przeanalizowane dane wydają cię ciekawe oraz przydatne, początkowe założenie zostało spełnione.

Dane nadają się również do przeanalizowania w aspekcie wpływu obecnej epidemii na incydenty, lecz wprowadzonych danych z ostatnich miesięcy jest jeszcze zbyt mało, aby wykonać poprawne wnioski.