Hurtownie danych – Sprawozdanie z zadania 7.

PWr. WIZ, Informatyka, Data: 26.04.2020

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

Projekt – etap I

Spis treści

1. Propozycja tematu

1.1. Tytuł projektu

Analiza popełnionych przestępstw w miescie Boston od roku 2015

1.2. Krótki opis obszaru analizy

Boston jest jednym z największych miast w Stanach Zjednoczonych pod względem populacji oraz gęstości mieszkańców. Jest przez to również niebezpieczniejszym miejscem do życia w porównaniu do większości innych miast państwa. Liczba przestępstw rośnie z roku na rok wraz z rozwojem miasta. W roku 2015 miasto rozpoczęło udostępnianie części danych raportów policyjnych, aby mieszkańcy byli świadomi aktualnej sytuacji oraz aby przyczynili się do zwalczania przestępstw.

1.3. Cel przedsięwzięcia

1.3.1. Oczekiwania

Celem jest analiza wykonanych przestępstw w perspektywie lokalizacji oraz czasu. Zbadane zostanie, które obszary miasta są najbardziej niebezpieczne, w jaki sposób pora roku czy święta wpływają na zdarzenia. Przedstawione wyniki mają pokazać jakie czynniki wpływają na przestępstwa, pokazać porównanie poszczególnych dystryktów oraz zwrócić uwagę, w jaki sposób przestępstwa mogłyby być uniknięte.

1.3.2. Zakres analizy – badane aspekty

Planowane jest przeprowadzenie następujących badań:

- Ogólny wzrost przestępstwa w całym mieście od początku zgromadzonych danych
- Porównanie przestępstw w dzień powszedni z ważniejszym świętem narodowym
- Przedstawienie najczęściej popełnianych przestępstw w zależności od czasu
- Przedstawienie najczęściej popełnianych przestępstw w zależności od lokalizacji
- Liczba przestępstw w zależności od dnia tygodnia
- Liczba przestępstw w zależności od godziny
- Przedstawienie częstotliwości przestępstw w zależności od ich kategorii i podkategorii
- Odnalezienie ulic o największej liczbie wypadków samochodowych
- Analiza użycia broni w zależności od kategorii przestępstwa
- Przedstawienie danych na mapie

1.4. Źródła danych

Dane incydentów będą pobrane z rządowej strony internetowej miasta Boston - https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system. Dane incydentów zawarte są w arkuszach programu MS Excel w formacie xlsx.

Dodatkowo, użyte zostaną użyte daty świąt narodowych dostępne w repozytorium https://www.kaggle.com/gsnehaa21/federal-holidays-usa-19662020

Pobierane pliki mają nierozpoznawalne nazwy z powodu ich ciągłej aktualizacji. Poniżej zostały użyte nazwy plików widoczne na stronie internetowej.

L.p.	Plik	Typ	Liczba rekordów	Rozmiar [MB]	Opis
1.	Crime Incident Reports (August 2015 – To Date) (Source – New System)	Arkusz MS Excel	478618	83	Zbiór inceydentów zareportowanych przez Boston Police Department od roku 2015
2.	RMS_Offense_CodesXLSX	Arkusz MS Excel	576	0.02	Słownik kodów przestępstw
3.	RMS_Crime_Incident_Field _ExplanationXLSX	Arkusz MS Excel	10	0.01	Wyjaśnienie pól w arkuszu głównym, wraz z ich opisami i typami danych
4.	usholidays	Tabularny	485	0.01	Słownik data - święto

2. Profilowanie danych

2.1. Analiza danych

Plik: Crime Incident Reports.xlsx				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	Incident_number	string	-	Unikalna wartość, posłuży jako klucz główny oraz indeks.
2.	Offense_code	number	111-3831	Opisy każdego kodu przestępstwa zawarte są w osobnym pliku.
3.	Offense_code_group	string	-	Wartość podobna do wartości w drugim pliku, zostanie pominięta.
4.	Offense_description	string	-	Krótki opis incydentu, może być zastosowany do przeszukania danych za pomocą Fuzzy Search
5.	District	string	-	Część wartości są puste, zostaną zamienione na „Unknown”
6.	Reporting_area	number	-	Nie podlega analizie
7.	Shooting	string	“”, 0, 1, “Y”	Część wartości jest pusta, część oznaczona jako 0 lub 1 oznaczające PRAWDA/FAŁSZ, część znakiem „Y” od słowa Yes. Wartości zostaną wyczyszczone na „Yes”, „No” oraz „Unknown”
8.	Occurred_on_date	date	15.06.15-25.04.20	Brak
9.	Year	number	2015-2020	Wydobyte z Occurred_on_date, ułatwi wprowadzanie danych do hurtowni
10.	Month	number	1-12	j.w.
11.	Day_of_week	string	Monday-Sunday	j.w.
12.	Hour	number	0-23	j.w.
13.	UCR_Part	string	Part One – Part Tree	Dodatkowo wartości “Other” oraz „”. Puste wartości zostaną zamienione na „Unknown”
14.	Street	string	-	Część wartości jest poprzedona wartością liczbową – zostaną one odcięte z wartości docelowej.
15.	Lat	number	-	Wiele rekordów posiada tutaj wartość -1 – oznacza to brak danych
16.	Lng	number	-	j.w.
17.	Location	Pair of numbers	-	Kolumna zawiera jedynie parę (Lat, Lng), zdefiniowanych już wcześniej. Nie jest więc potrzebna.

Plik: rmsoffensecodes.xlsx				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	CODE	number	111-3831	Unikalny kod identyfikujący rodzaj przestępstwa
2.	NAME	string	-	Opis przestępstwa w postaci hierarchicznej (np. „MANSLAUGHTER - VEHICLE - NEGLIGENCE"). Dane zostaną podzielone na odpowiednie kolumy.

Plik: usholidays.csv				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	Idx	number	1-484	Wartość porządkowa, zbędna do analizy
2.	Date	date	25.12.66-25.12.20	Potrzebne są wartości jedynie od roku 2015.
3.	Holiday	string	-	Opis narodowego święta

2.2. Ocena przydatności danych w pliku do tworzenia hurtowni danych

Lp.	Plik	Ocena jakości danych
1.	Crime Incident Reports (August 2015 – To Date) (Source – New System)	Plik zawiera większość danych, na podstawie których zostaną utworzone tabele wymiarów oraz tabela faktów w hurtowni danych. Część komórek nie posiada wartości – będą one ustawione na wartość znakową. Część danych jest już przetworzona na wartości elementarne (jak data na rok, miesiąc, dzień), co ułatwi wprowadzanie danych. Jedna kolumna zawiera klucz obcy w postaci kodu przestępstwa – będzie to jeden z punktów denormalizacji.
2.	RMS_Offense_CodesXLSX	Dane w pliku są w postaci słownika kod przestępstwa – wartość znakowa. Przestępstwo w tym przypadku jest zapisane w postaci hierarchicznej, np. „ROBBERY – STREET" oraz „ROBBERY – BANK". Dane te zostaną w miarę możliwości przetworzone na hierarchię przedstawioną za pomocą osobnych kolumn. Utrudnieniem są wartości, które posiadają jedynie jedną część.
3.	RMS_Crime_Incident_Field_ ExplanationXLSX	Dane zawarte w pliku posiadają jedynie wytłumaczenie struktury danych w pliku głównym. Będzie to przydatne podczas tworzenia hurtowni danych, lecz dane nie trafią do tabel wymiarów ani faktów.
4.	usholidays	Plik posiada rekordy tylko dla dat, w których odbywa się święto. Dane będą zdenormalizowane wraz z plikiem głównym na połączeniu na kolumnie data.

2.3. Definicja typów encji/klas oraz związków pomiędzy nimi

Przestępstwo

Atrybut	Opis	Typ
Code	Kod przestępstwa, klucz główny	Liczba
Category	Ogólny opis przestępstwa, kategoria	Ciąg znaków
Detail	Dodatkowe informacje o przestępstwie	Ciąg znaków

Lokalizacja

Atrybut	Opis	Typ
Id	Unikalny identyfikator, klucz główny	Liczba
District	Kod dystryktu	Ciąg znaków
Street	Nazwa ulicy	Ciąg znaków

UCR

Atrybut	Opis	Typ
Id	Identyfikator, klucz główny	Liczba
Name	Nazwa kategorii	Ciąg znaków

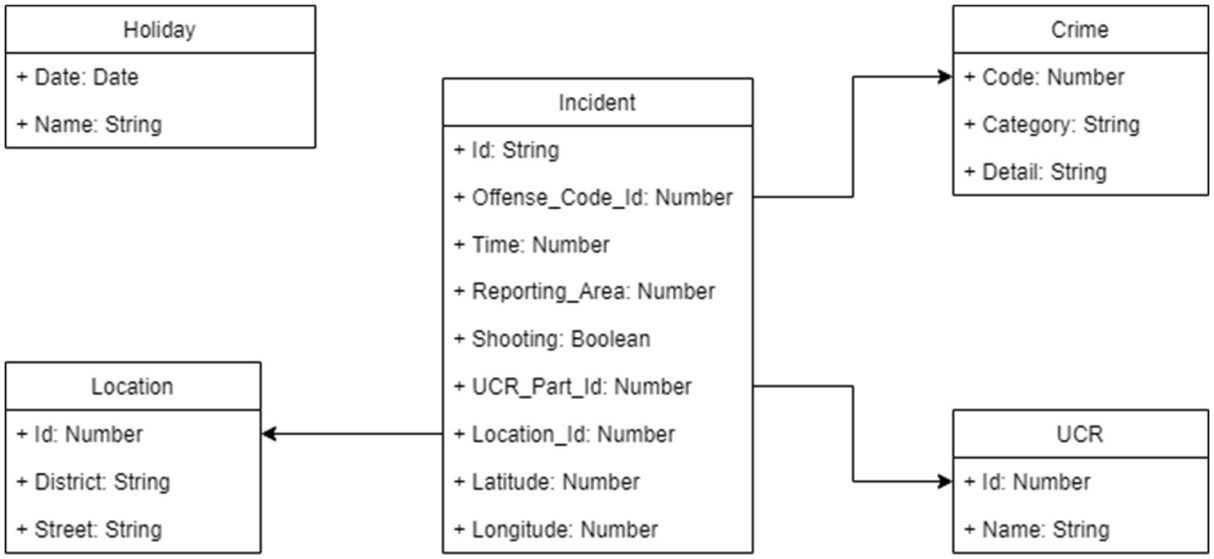
Incydent

Atrybut	Opis	Typ
Id	Identyfikator incydentu, klucz główny	Ciąg znaków
Offense_Code	Kod kategorii przestępstwa, klucz obcy	Liczba
Date	Data i czas zajścia incydentu	Data
Reporting_Area	Kod strefy, z której incydent został zaraportowany	Liczba
Shooting	Oznacza, czy w incydencie doszło do strzelaniny	Wartość logiczna
UCR_Part_Id	Oznacza srogość przestępstwa, klucz obcy	Liczba
Location_Id	Lokalizacja incydentu	Liczba
Latitude	Dokładna szerokość geograficzna incydentu	Liczba
Longitude	Dokładna wysokość geograficzna incydentu	Liczba

Święto narodowe

Atrybut	Opis	Typ
Date	Data danego święta. Klucz główny.	Data
Name	Nazwa święta.	Ciąg znaków

2.4. Diagram klas



2.5. Propozycja wymiarów, faktów oraz hierarchii

Wymiary

DIM_TIME

Atrybut	**Opis**	**Typ**
Time	Data w postaci liczbowej składająca się z roku, miesiąca, dnia oraz godziny. Klucz główny	Liczba
Year	Rok	Liczba
Quarter	Numer kwartału w roku	Liczba
Month	Numer miesiąca w roku	Liczba
Month_Name	Nazwa miesiąca w języku angielskim	Ciąg znaków
Day	Numer dnia tygodnia w miesiącu	Liczba
Weekday	Dzień tygodnia, słownie w j. angielskim	Ciąg znaków
Hour	Godzina, w formacie 24-godzinnym	Liczba
Holiday_Name	Nazwa święta narodowego odbywającego się w danym dniu.	Ciąg znaków

DIM_CRIME

Atrybut	Opis	Typ
Code	Kod przestępstwa, klucz główny	Liczba
Category	Ogólny opis przestępstwa, kategoria	Ciąg znaków
Detail	Dodatkowe informacje o przestępstwie	Ciąg znaków

DIM_LOCATION

Atrybut	Opis	Typ
Id	Unikalny identyfikator, klucz główny	Liczba
District	Kod dystryktu	Ciąg znaków
Street	Nazwa ulicy	Ciąg znaków

Fakty

FACT_INCIDENTS

Atrybut	Opis	Typ
Offense_Code	Kod kategorii przestępstwa, klucz obcy	Liczba
Time	Data i czas zajścia incydentu. Klucz obcy	Liczba
Reporting_Area	Kod strefy, z której incydent został zaraportowany	Liczba
Shooting	Oznacza, czy w incydencie doszło do strzelaniny	Ciąg znaków
UCR_Part	Oznacza srogość przestępstwa	Ciąg znaków
Location	Kod lokalizacji incydentu	Liczba
Latitude	Dokładna szerokość geograficzna incydentu	Liczba
Longitude	Dokładna wysokość geograficzna incydentu	Liczba

Hierarchie

Dim_Time: Year -> Quarter -> Month -> Day -> Hour

Dim_Crime: Category -> Detail

Dim_Location: District -> Street

3. Utworzenie tabel w bazie danych

```
CREATE SCHEMA Rasz;

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 PRIMARY KEY,
	[Category] NVARCHAR(50) NOT NULL,
	[Detail] NVARCHAR(50) NULL
);

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

CREATE TABLE [Rasz].[Fact_Incident](
	[OffenseCode] SMALLINT NULL,
	[Time] INT NOT NULL,
	[ReportingArea] SMALLINT NULL,
	[Shooting] NVARCHAR(7) NOT NULL,
	[UCRPart] NVARCHAR(20) NOT NULL,
	[Location] INT NULL,
	[Latitude] DECIMAL(8,2) NULL,
	[Longitude] DECIMAL(8,2) NULL
);
```