# Uzasadnienie wyboru tematu projektu

Propozycja 1. **Naruszenia drogowe w hrabstwie Maryland**

Tabela 1. Źródło danych dla tematu „**Naruszenia drogowe w hrabstwie Maryland”**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lp.** | **Plik, bazy danych** | **Typ** | **Liczba rek.** | **Rozmiar**  **[MB]** | **Opis** |
| 1. | Traffic\_violations | .csv | ~1 300 000 | 448 | Baza zawierająca dane dotyczące naruszeń drogowych w hrabstwie Maryland w okresie od ??? do ??? |

Tabela 2. Ocena jakości danych w pliku „**Traffic\_violations.csv”**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| L**p.** | **Kolumna** | **Typ** | **Zakres** | **Ocena jakości** |
| **1.** | Date of stop | Data | 01.01.2012 – 01.01.2018 | Data wydarzenia, format DD.MM.RRRR |
| **2.** | Time of stop | Godzina | 00:00:00 – 23:59:59 | Czas zdarzenia, format GG:MM:SS |
| **3.** | Agency | Tekstowy | MCP | Nazwa agencji |
| **4.** | SubAgency | Tekstowy |  | Kod sądu określający okrąg funkcjonariusza |
| **…** | … |  |  |  |
| **7.** | Latitude | Numeryczny | [-100; 40] | Szerokość geograficzna  10% NULL |
| **8.** | Longitude | Numeryczny | [-80; 40] | Długość geograficzna  10% NULL |
| **9.** | Accident | Tekstowy | Yes, No | Informacja czy wykroczenie obejmowało wypadek |
| **10.** | Belts | Tekstowy | Yes, No | Informacja czy zapięty był pas bezpieczeństwa |
| **11.** | Personal injury | Tekstowy | Yes, No | Informacja czy wykroczenie obejmowało obrażenia ludzi |
| **12.** | Property damage | Tekstowy | Yes, No | Informacja czy wykroczenie obejmowało uszkodzenia mienia |
| **13.** | Fatal | Tekstowy | Yes, No | Informacja czy wykroczenie spowodowało śmierć |
| **14.** | Commercial license | Tekstowy | Yes, No | Informacja czy kierowca samochód jest na licencji służbowej |
| **15.** | HAZMAT | Tekstowy | Yes, No | Informacja czy wykroczenie dotyczy materiałów niebezpiecznych |
| **16.** | Commercial vehicle | Tekstowy | Yes, No | Informacja czy pojazd jest … |
| … | … | … | … | … |

Propozycja 2. **Zmiany klimatyczne**

Tabela 3. Źródła danych dla tematu „**Zmiany klimatyczne**”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lp.** | **Plik, nazwa bazy danych** | **Typ** | **Liczba rek.** | **Rozmiar**  **[MB]** | **Opis** |
| 1. | GlobalLandTemperaturesByCity | .csv | 1 048 576 | 508 | Temperatury w miastach |
| 2. | GlobalLandTemperaturesByCountry | .csv | 577 463 | 22 | Temperatury państwach |
| 3. | GlobalLandTemperaturesByMajorCity | .csv | 239 178 | 13 | Temperatury w dużych miastach |
| 4. | GlobalLandTemperaturesByState | .csv | 645 676 | 29 | Temperatury w stanach |
| 5. | GlobalTemperatures | .csv | 3 193 | 0,2 | Globalne temperatury |

Tabela:

Tabela 4. Ocena jakości danych w pliku „**GlobalLandTempteraturesByCity.csv”**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lp.** | **Kolumna** | **Typ** | **Zakres** | **Ocena jakości** |
| **1.** | dt | Data | 01.11.1743 – 01.09.2013 | Data pomiaru |
| **2.** | Average Temperature | Numeryczny | -30,182 – 34,847 | Średnia temperatura 4% NULL |
| **3.** | Average Temperature Uncertainty | Numeryczny | 0,036 - 0,312 | Niepewność pomiary 4% NULL |
| **4.** | City | Tekstowy |  | Miasto, w którym dokonano pomiary |
| **5.** | Country | Tekstowy |  | Kraj, w którym dokonano pomiary |
| **6.** | Latitude | Tekstowy |  | Szerokość geograficzna, na której dokonano pomiary |
| … | … | … |  |  |

Tabela 5. Ocena jakości danych w pliku „**GlobalLandTempteraturesByMajorCity.csv**”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lp.** | **Kolumna** | **Typ** | **Zakres** | **Ocena jakości** |
| **1.** | dt | Data | 01.11.1743 – 01.09.2013 | Data pomiaru |
| **2.** | Average Temperature | Numeryczny | -26,722 – 23,151 | Średnia temperatura 5% NULL |
| **3.** | Average Temperature Uncertainty | Numeryczny | 0,04 - 0,256 | Niepewność pomiary 5% NULL |
| **5.** | Country | Tekstowy |  | Kraj, w którym dokonano pomiary |
| **6.** | Latitude | Tekstowy |  | Szerokość geograficzna, na której dokonano pomiary |

…

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lp.** | **Kolumna** | **Typ** | **Zakres** | **Ocena jakości** |
| **1.** | dt | Data | 01.01.1750 – 01.12.2015 | Data pomiaru |
| **2.** | LandAverageTemperature | Numeryczny | -2,08 – 19,021 | Średnia temperatura 1% < NULL |
| **3.** | Average Temperature Uncertainty | Numeryczny | 0,034 - 0,087 | Niepewność pomiaru średniej temperatury 1% < NULL |
| **4.** | Land Max Temperature | Numeryczny | ? | Maksymalna temperatura 1% < NULL |
| **5.** | Land Max Temperature Uncertainty | Numeryczny | ? | Niepewność pomiaru maksymalnej temperatury 38% NULL |
| **6.** | Land Min Temperature | Numeryczny | ? | Minimalna temperatura 38% NULL |
| **7.** | Land And Ocean Average Temperature | Numeryczny | ? | Średnia temperatura lądu i oceanu 38% NULL |
| **8.** | Land And Ocean Average  Temperature Uncertainty | Numaryczny | ? | Niepewność pomiaru średniej temperatury lądu i |

# Podsumowane i wnioski

Zmiany klimatyczne dostarczone są w 5 plikach .csv, z czego każdy z nich ma liczbę kolumn między 4 a 9. Naruszenia drogowe natomiast są zamieszczone w jednym pliku, który zawiera 35 kolumn. Łączna liczba kolumn jest zatem podobna.

Dane:

Zakres czasowy danych temperaturowych jest zdecydowanie większy – ponad 250 lat kontra 5 lat.

Dane temperaturowe są kompletne – w większości osiągają maksymalnie 6% nulli (nie licząc kilku kolumn w pliku GlobalTemperatures, które osiągają 38% nulli). Jednakże największym minusem danych temperaturowych jest fakt, że zawierają one tylko dane temperatur – mocno ogranicza to analizę.

Dane wykroczeń również są kompletne i mają niski poziom wartości pustych.

Największym minusem jeżeli chodzi o dane naruszeń drogowych jest fakt, że nie są one podzielone wg kategorii. Tzn. jedynym sposobem na znalezienie informacji, jakie naruszenie miało miejsce, jest analiza pola DESCRIPTION, które jest tekstowym opisem naruszenia. Tutaj zdecydowanie wygodniejsze byłoby zastosowanie odnośnika do osobnej tabeli, która zawierałaby typ wykroczenia i w osobnym polu opis.

Jednakże nawet po odrzuceniu kolumn, które nie będą nadawały się do analizy, baza wykroczeń daje dużo więcej możliwości na analizę niż baza temperatur.

Temat:

Ostatnią kwestią jest sam temat, którego dotyczą dane. Temat zmian klimatycznych jest w obecnych czasach często poruszany i jest on wg mnie ważny, a dane z bardzo dużego zakresu czasowego mogłyby pozwolić na ciekawą analizę. Jednakże wydaje mi się, że same dane temperaturowe nie pozwolą na wystarczającą analizę.

Temat wykroczeń drogowych jest również ważnym tematem, dotyczącym nas wszystkich na co dzień. Dane obejmują wiele cech, co może skutkować bardzo obszerną i ciekawą analizą.

Podsumowując, głównym czynnikiem wpływającym na wybór tematu wykroczeń jest fakt, że baza ta pozwala na zdecydowanie szerszą analizę niż tylko dane temperaturowe. Gdyby dane temperaturowe połączone były z innymi czynnikami takimi jak np. ilość opadów itd. to byłby to zdecydowanie lepszy temat do analizy. Dane wykroczeń wydają się być więc lepiej przystosowane do analizy w postaci hurtowni danych.

# Decyzja

Wybrany temat: **Naruszenia drogowe w hrabstwie Maryland**