Hurtownie danych – Projekt HD

PWr. WIZ, Data: 8.05.2019

Student	----------	Ocena
Indeks	----------	
Imię	XYZ	
Nazwisko	ZYX	

1. Tytuł projektu

Policyjne kontrole drogowe w stanie Vermont, USA w latach 2010-2015.

2. Charakterystyka dziedziny problemowej

Policja w Stanach Zjednoczonych jest jedną z najważniejszych służb publicznych. Liczy prawie 18 tysięcy jednostek z uwzględnieniem między innymi policji lokalnej, federalnej czy, znanych z popkultury, biur szeryfa. Pomimo szerokiego zakresu działalności w celu wymierzania sprawiedliwości, to zatrzymania drogowe stanowią temat kontrowersyjny, szczególnie ze względów rasowych.

Vermont jest stanem w regionie New England, położonym na północnym-wschodzie Stanów Zjednoczonych Ameryki Północnej. Jest drugim najmniejszym stanem, zamieszkałym przez 626 299 mieszkańców (dane na rok 2018, dla porównania populacja Wrocławia wynosi 632 020 mieszkańców z danych na rok 2016).

Faktem jakim będziemy się zajmować podczas analizy jest zatrzymanie na drodze. Wymiarami są data, czas, lokalizacja, departament policji, kierowca, przyczyna, skutek, przeprowadzona akcja. Miarą faktu jest liczba zatrzymań.

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

Będziemy zajmować się szczegółową analizą zbioru danych dotyczących zatrzymań przez policję w stanie Vermont w latach 2010-2015. Ciekawa może okazać się analiza pod względem miast, w których doszło do zatrzymań, czasu zatrzymania (czy więcej incydentów odbyło się w nocy czy za dnia), rasy i płci zatrzymanego, przyczyny zatrzymania i czy doszło do aresztowania.

2.2 Problemy

Do znaczących problemów należy zaliczyć:

1. Większą częstotliwość zatrzymań osób odmiennej rasy niż biała – powszechna opinia.
2. Eliminacja stereotypów w zakresie oceny płci kierowców
3. Krytyka nieuzasadnionych kontroli samochodów (przeszukiwań) oraz ich skutków - konieczność potwierdzenia w kontekście zarejestrowanych faktów.

2.3 Cel przedsięwzięcia

2.3.1 Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

Analityczna baza danych powinna umożliwić dogłębną analizę faktów dotyczących interwencji policji drogowej i odpowiedź na, między innymi, następujące pytania:

- Czy dochodzi do większej liczby zatrzymań osób czarnoskórych lub ogólnie innej rasy niż biała?
- Czy jest prawdziwe stwierdzenie dotyczące opinii o nieuzasadnionych zatrzymaniach czarnoskórych mężczyzn po zmroku?
- W których miejscach i z jakich powodów dochodzi do dużej liczby wykroczeń?
- Jaka pora dnia sprzyja największej liczby wykroczeń?
- Jakie są najczęstsze przyczyny i skutki zatrzymań?
- W jakim stopniu przeszukania skutkują znalezieniem kontrabandy?

Właściwa (ekspercka) analiza danych historycznych powinna dostarczyć odpowiedzi opartych na faktach na powyższe pytania oraz dostarczyć informacji, z których będzie możliwe wyciągnięcie wniosków stanowiących podstawę podejmowania trafnych decyzji.

2.3.2 Zakres analizy – badane aspekty

Analizie będą podlegać zatrzymania w kontekście płci, rasy i wieku kierowcy, a także z uwzględnieniem miejsca i czasu zatrzymania. Należy również przeanalizować przyczyny oraz skutki zatrzymań, a także działania podjęte przez policję w trakcie zatrzymania, takie jak na przykład przeszukanie pojazdu.

2.3.3 Potencjalni użytkownicy

Baza analityczna będzie wspierać procesy decyzyjne policji oraz udostępniać informacje dziennikarzom zainteresowanym tematyką zatrzymań przez policję w stanie Vermont.

3. Dane źródłowe

3.1. Źródła danych

Charakterystyka pliku zawierający danę źródłowe przeznaczone do stworzenia tematycznej hurtowni danych jest przedstawiona w tab. 1.

Tabela 1. Zbiory danych źródłowych

Lp.	Plik	Typ	Liczba rek.	Rozmiar[MB]	Opis
1.	Data_Vermont.csv	csv	283 285	63.2	Plik zawiera dane dotyczące zatrzymań przez policję w stanie

					Vermont w USA w latach 2010-2015.

3.2. Lokalizacja, dostępność danych źródłowych

Dane pochodzą ze strony
https://openpolicing.stanford.edu/data/
z projektu naukowców z Uniwersytetu Stanforda:
E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, and S. Goel. (2019)
"A large-scale analysis of racial disparities in police stops across the United States".
Są dostępne na mocy licencji Open Data Commons Attribution License.

3.3. Słownik danych – interpretacja

Interpretacja oraz wyjaśnienie znaczeń pojęć dziedzinowych zostały zawarte w tab.2.

Tabela 2. Słownik atrybutów

Plik: Data_Vermont.csv				
Lp.	**Atrybut**	**Typ danych**	**Znaczenie**	**Uwagi**
1.	id	Tekstowe	13-znakowy ciąg w formacie VT-RRRR-NrSeryjny	Unikalny numer nadawany każdemu zatrzymaniu. Składa się z dwóch liter oznaczających nazwę stanu, roku zatrzymania i numeru seryjnego.
2.	state	Tekstowe	2-znakowy skrót oznaczający nazwę stanu	Oznaczenie stanu, w którym doszło do do zatrzymania. W naszym przypadku dane dotyczą Vermont, zatem w każdym rekordzie znajduje się skrót VT.
3.	stop_date	Data	Data zatrzymania w formacie YYYY-MM-DD	Dzień, miesiąc i rok, w którym doszło do zatrzymania.
4.	stop_time	Tekstowe	Czas zatrzymania w formacie 24-godzinnym HH:MM	Godzina i minuta, w której doszło do zatrzymania.
5.	stop_datetime	Datetime	Połączenie daty i czasu zatrzymania	Data oraz czas, w którym doszło do zatrzymania.
6.	location_raw	Tekstowe	Miejsce zatrzymania	Nazwa miejsca, w którym doszło do zatrzymania.

7.	county_name	Tekstowe	Hrabstwo	Nazwa hrabstwa, w którym doszło do zatrzymania.
8.	county_fips	Numeryczne	Kod FIPS	Kod FIPS dokładnie identyfikuje miejsce zatrzymania.
9.	fine_grained_location	Tekstowe	Lokalizacja zatrzymania	Dokładna lokalizacja zdarzenia, z uwzględnieniem ulicy.
10.	police_department	Tekstowe	Departament Policji.	Nazwa departamentu policji, do którego należą policjanci biorący udział w zatrzymaniu.
11.	driver_gender	Tekstowe	Płeć kierowcy	Litera M lub F oznaczająca płeć zatrzymanego kierowcy.
12.	driver_age_raw	Numeryczne	Wiek zatrzymanego kierowcy	Wiek zatrzymanego kierowcy pochodzący z surowych danych.
13.	driver_age	Numeryczne	Wiek zatrzymanego kierowcy	Wiek zatrzymanego kierowcy pochodzący z oczyszczonych danych.
14.	driver_race_raw	Tekstowe	grupa rasowa zatrzymanego kierowcy	Grupa rasowa zatrzymanego kierowcy pochodząca z surowych danych.
15.	driver_race	Tekstowe	grupa rasowa zatrzymanego kierowcy	Grupa rasowa zatrzymanego kierowcy pochodząca z oczyszczonych danych.
16.	violation_raw	Tekstowe	Popełnione wykroczenie	Wykroczenie popełnione przez zatrzymanego kierowcę, dane surowe.
17.	violation	Tekstowe	Popełnione wykroczenie	Wykroczenie popełnione przez zatrzymanego kierowcę, dane oczyszczone.
18.	search_conducted	Bit	Czy zostało wszczęte przeszukanie?	Wartość boolowska oznaczająca czy zostało wszczęte przeszukanie samochodu.
19.	search_type_raw	Tekstowe	Czy zostało wszczęte przeszukiwanie, uwzględniony rodzaj przeszukania.	Wartość boolowska oznaczająca czy zostało wszczęte przeszukanie samochodu. W danych zostały uwzględnione typy przeszukania, jednakże są to dane surowe, zatem są one zanieczyszczone.
20.	search_type	Tekstowe	Typ przeszukania	Rozdzielenie search_type_raw. Określa typ przeszukania.

21.	contraband_found	Bit	Czy została znaleziona kontrabanda?	Wartość boolowska określająca czy po przeprowadzeniu przeszukania została znaleziona kontrabanda.
22.	stop_outcome	Tekstowe	Skutki zatrzymania	Określenie skutków zatrzymania. Według twórców bazy, jeśli zatrzymanie ma kilka skutków, to w bazie znajduje się tylko ten najpoważniejszy z punktu widzenia prawa.
23.	is_arrested	bit	Czy kierowca został aresztowany?	Wartość boolowska określająca czy kierowca został aresztowany po zatrzymaniu.
24.	officer_id	Numeryczne	Numer identyfikujący oficera policji, który doprowadził do zatrzymania pojazdu	Indywidualny numer identyfikujący oficera policji, który uczestniczył w zatrzymaniu.

3.4. Ocena jakościowa danych

Wynik analizy jakościowej przeprowadzonej za pomocą programu Tableau oraz profilu danych SSIS został przedstawiony w tab. 3.

Tabela 3. Ocena jakościowa danych

Plik: Data_Vermont.csv					
Lp.	**Atrybut**	**Typ danych**	**Zakres wartości**	**Znaczenie**	**Uwagi - ocena jakości danych**
25.	id	Tekstowe	VT-2010-00001 - VT-2015-45662	Unikalny numer nadawany każdemu zatrzymaniu. Składa się z dwóch liter oznaczających stan, roku zatrzymania i numeru seryjnego.	Dane są dobre jakościowo, każdy rekord posiada taki numer.
26.	state	Tekstowe	VT	Skrót oznaczający stan, w którym odbyło się zatrzymanie	W naszym przypadku stan w każdym rekordzie jest taki sam -VT oznaczające Vermont.
27.	stop_date	Data		Data zatrzymania w formacie YYYY-MM-DD	Każdy rekord posiada datę.
28.	stop_time	Tekstowe		Czas zatrzymania w formacie 24-	Podobnie jak w przypadku stop_date, czas zamiast w

				godzinnym HH:MM	odpowiednim dla siebie typie jest w typie tekstowym. Każdy rekord posiada czas.
29.	stop_datetime	Datetime		Połączenie daty i czasu zatrzymania	Każdy rekord posiada to pole.
30.	location_raw	Tekstowe		Nazwa miejsca zatrzymania	Miejsce, w którym doszło do zatrzymania, jednakże niedokładne i w formacie niestandardowym dla USA. 0.77% rekordów nie ma wartości w tym polu. Dane oznaczone jako “raw” czyli oryginalne.
31.	county_name	Tekstowe		Nazwa hrabstwa, w którym doszło do zatrzymania	Nazwa hrabstwa(county) jest w standardowym formacie dla USA. 0.25% rekordów nie ma wartości w tym polu.
32.	county_fips	Numeryczne		Kod FIPS oznaczający dokładne miejsce zatrzymania	Kod FIPS dokładnie i jednoznacznie identyfikuje miejsce zdarzenia. 0.25% rekordów nie ma wartości w tym polu.
33.	fine_grained_location	Tekstowe		Dokładna lokalizacja z określeniem ulicy.	Duża ilość rekordów różnie zapisana. Atrybut nieistotny z punktu widzenia hurtowni danych.
34.	police_department	Tekstowe		Nazwa departamentu policji, której patrol dokonał zatrzymania.	0.0004% rekordów nie ma wartości dla tego atrybutu.
35.	driver_gender	Tekstowe	M, F	Litera oznaczająca płeć kierowcy (M-male, F-female)	Tylko 1710 rekordów, czyli około 0.6% wszystkich, nie posiada informacji o płci kierowcy
36.	driver_age_raw	Numeryczne		Wiek zatrzymanego kierowcy	Te dane oznaczone są jako “raw”, co oznacza, że są to dane oryginalne przed obróbką. Znajdują się w nich zarówno liczby takie jak 0 czy 1, ale też 101 oraz słowa, których nie powinno być w wartości tego atrybutu. Dodatkowo 0.4% rekordów nie posiada wartości dla tego atrybutu.

37.	driver_age	Numeryczne		Wiek zatrzymanego kierowcy	W tym przypadku nie ma tylu anomalii jak w poprzednim atrybucie, ponieważ są to dane już oczyszczone. Około 0.45% rekordów nie ma wartości dla tego atrybutu.
38.	driver_race_raw	Tekstowe	Black, White, Asian or Pacific Islander, Unknown, Hispanic	grupa rasowa zatrzymanego kierowcy	Około 1.4% rekordów nie ma wartości dla tego argumentu. 0.29% jest oznaczone jako Unknown. Dane oznaczone jako “raw” czyli przed obróbką.
39.	driver_race	Tekstowe	Hispanic, Asian, Black, White	rasa zatrzymanego kierowcy	1.7% rekordów nie ma wartości dla tego atrybutu.
40.	violation_raw	Tekstowe	Externally Generated	Popełnione wykroczenie	Oryginalne dane (“surowe”) dotyczące wykroczenia popełnionego przez kierowcę
41.	violation	Tekstowe		Popełnione wykroczenie, ustandaryzowane i oczyszczone	Około 0.8% rekordów nie ma wartości dla tego atrybutu.
42.	search_conducted	Bit	1, 0	Wartość boolowska oznaczająca czy przeszukanie zostało wykonane czy nie	Wszystkie rekordy mają wartość w tym polu.
43.	search_type_raw	Tekstowe		Uzasadnienie przeszukania samochodu	Aż 97.65% rekordów ma zaznaczone “No Search Conducted” co oznacza, że pojazd nie został przeszukany. Są to dane “surowe”, dopiero kolejny atrybut pozwoli na ich wyjaśnienie.
44.	search_type	Tekstowe	Consent Search-Probable Cause, Consent Search - Reasonable Suspicion, No Search Conducted	Uzasadnienie przeszukania samochodu ustandaryzowane	98.44% rekordów nie ma wartości dla tego atrybutu. Pozostałe wartości zostały podzielone na 3 kategorie zaznaczone w zakresie wartości. W tym przypadku wartość No Search Conducted ma 0.34% rekordów.

45.	contraband_found	Bit	1, 0	Określenie czy została znaleziona kontrabanda	TRUE oznacza znalezioną kontrabandę po przeszukaniu. FALSE oznacza, że nie było przeszukania. Około 0.4% rekordów nie ma wartości dla tego atrybutu.
46.	stop_outcome	Tekstowe	Written Warning, Arrest for Violation, Citation	Określenie skutków zatrzymania	Według twórców bazy, jeśli zatrzymanie ma kilka skutków, to w bazie znajduje się tylko ten najpoważniejszy z punktu widzenia prawa. Około 0.8% rekordów nie ma wartości dla tego atrybutu.
47.	is_arrested	bit	1, 0	Określenie czy doszło do aresztowania	Aż w około 98% rekordów znajduje się wartość FALSE.
48.	officer_id	Numeryczne		Numer identyfikujący oficera policji, który doprowadził do zatrzymania pojazdu	Atrybut nieistotny z punktu widzenia hurtowni i pozostałych danych.

4. Analityczne modele wielowymiarowe

4.1. Fakty **podlegające analizie oraz ich miary**

Analizie będzie podlegał zbiór zarejestrowanych zdarzeń zatrzymania pojazdów (tab. 4.)

Tabela 4. Fakty podlegające analizie

Lp.	Fakty	Miary	Uwagi
1.	Fact Stop	Stop Quantity	Liczba zatrzymań

4.2. Kontekst analizy faktów

Ustalony kontekst analizy faktów został przedstawiony w tab. 4.

Tabela 5. Wymiary analizy faktów

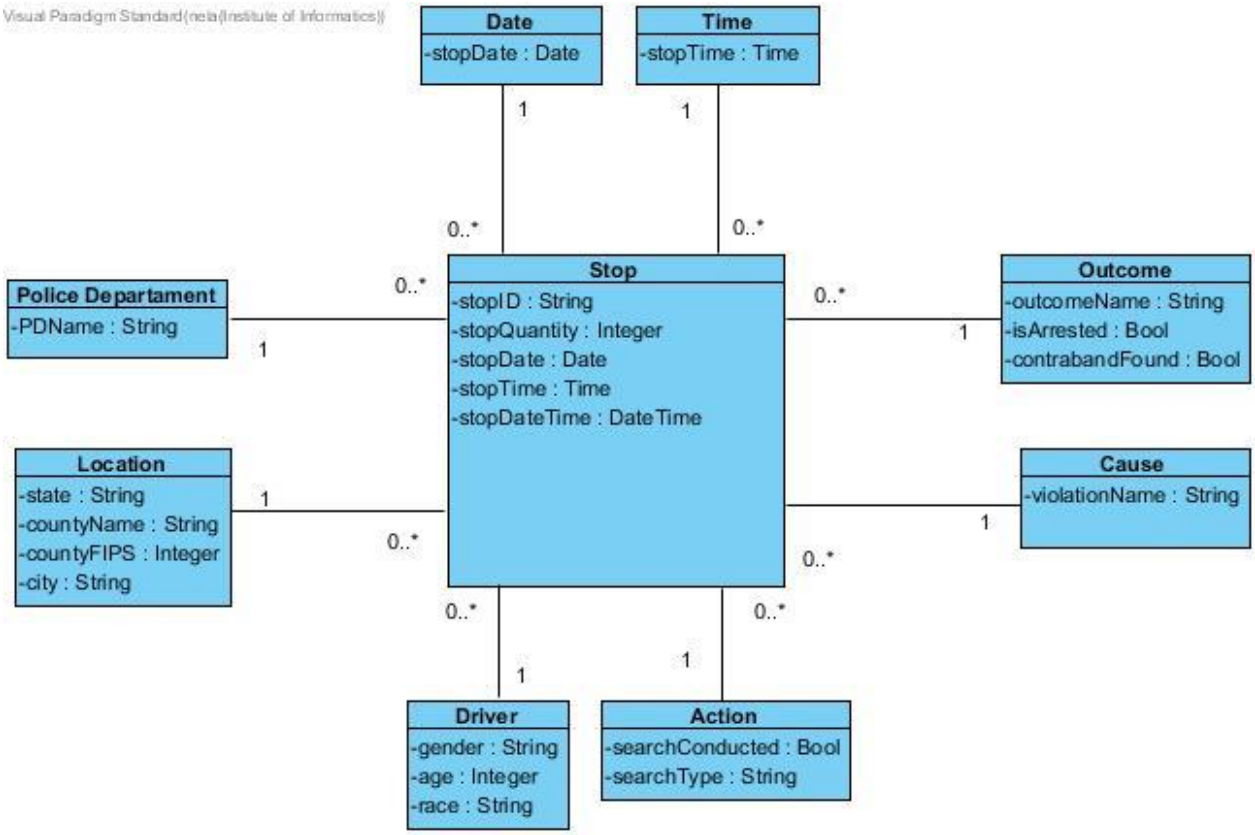
Lp.	Wymiar	Opis
1.	Time	Umożliwia analizę biznesową w kontekście historii zdarzenia (faktu). Pozwoli na ocenę rozłożenia w trakcie dnia ilości wykroczeń na drodze.
2.	Date	Umożliwia analizę biznesową w kontekście historii zdarzenia. Pozwoli na

		ocenę ilości wykroczeń na drodze na przestrzeni lat i miesięcy.
3.	Location	Umożliwia analizę biznesową w kontekście miejsca zdarzenia (faktu). Ponieważ dane dostarczają dość dokładne narzędzia pozwalające na określenie w jakim miejscu doszło do wykroczenia, będzie można stwierdzić jakie części stanu Vermont należy częściej patrolować.
4.	Police Departament	Umożliwia analizę biznesową w kontekście zaangażowanego departamentu policji. Dzięki temu wymiarowi będzie można stwierdzić, które departamenty policji w stanie Vermont są najbardziej zaangażowane w zwalczanie wykroczeń na drodze.
5.	Driver	Umożliwia analizę biznesową w kontekście informacji o zatrzymanym kierowcy. Dzięki niej będzie można określić przedział wiekowy kierowcy, płeć oraz rasę.
6.	Cause	Umożliwia analizę biznesową w kontekście przyczyny zatrzymania. Dzięki temu będzie można określić jakie są najczęstsze przyczyny zatrzymań na drodze, a także jak dobrze uzasadnione jest zatrzymanie.
7.	Outcome	Umożliwia analizę biznesową w kontekście skutków zatrzymania. Pozwoli na sprawdzenie, czy skutek był uzasadniony (po połączeniu z przyczyną).
8.	Action	Umożliwia analizę biznesową w kontekście akcji, takich jak przeszukanie, podjętych przez policję w trakcie zatrzymania.

4.3. Modele wielowymiarowe (UML)

Po przeanalizowaniu atrybutów źródła danych oraz ustalonego faktu i kontekstu analizy zaproponowano wielowymiarowy model konceptualny (rys. 1.). Składa się on z faktu **Stop** i ośmiu wymiarów. Model ten reprezentowany jest w postaci schematu gwiazdy.

Uwaga: Warto zaznaczyć, że StopID jest kluczem pochodzącym z danych źródłowych.



Rysunek 1. Wielowymiarowy model analityczny przedstawiony na poziomie konceptualnym

5. Projekt procesu ETL

5.1. Schemat bazy danych HD (skrypt SQL)

Baza danych została utworzona za pomocą skryptu przedstawionego w tabeli 6.

Tabela 6. Skrypt SQL tworzenia bazy hurtowni danych

```
CREATE TABLE DimDate (

        DateKey int IDENTITY(1,1) NOT NULL,

        FullStopDate DATE NOT NULL,
```

```
        DayNumberOfMonth tinyint NOT NULL,
        MonthNumberOfYear tinyint NOT NULL,
        CalendarYear smallint NOT NULL,
        CalendarQuarter tinyint NOT NULL,
        CalendarSemester tinyint NOT NULL
        CONSTRAINT PK_DateKey PRIMARY KEY (DateKey)
);
CREATE TABLE DimTime (
        TimeKey int IDENTITY(1,1) NOT NULL,
        FullStopTime TIME NOT NULL,
        CONSTRAINT PK_TimeKey PRIMARY KEY (TimeKey)
);
CREATE TABLE DimPoliceDepartament (
        PDKey int IDENTITY(1,1) NOT NULL,
        PDName nvarchar(50) NULL
        CONSTRAINT PK_PDKey PRIMARY KEY (PDKey)
);

CREATE TABLE DimLocation (
        LocationKey int IDENTITY(1,1) NOT NULL,
        State nvarchar(2) NULL,
        CountyName nvarchar(50) NULL,
        CountyFIPS int NULL,
        City nvarchar(50) NULL,
        CONSTRAINT PK_LocationKey PRIMARY KEY (LocationKey)
);
CREATE TABLE DimDriver (
        DriverKey int IDENTITY(1,1) NOT NULL,
```

```
        Gender nvarchar(1) NULL,
        Age int NULL,
        Race nvarchar(50) NULL,
        CONSTRAINT PK_DriverKey PRIMARY KEY (DriverKey)
    );
    CREATE TABLE DimAction (
        ActionKey int IDENTITY(1,1) NOT NULL,
        SearchConducted bit NULL,
        SearchType nvarchar(50) NULL
        CONSTRAINT PK_ActionKey PRIMARY KEY (ActionKey)
    );
    CREATE TABLE DimCause (
        CauseKey int IDENTITY(1,1) NOT NULL,
        ViolationName nvarchar(50) NULL
        CONSTRAINT PK_CauseKey PRIMARY KEY (CauseKey)
    );
    CREATE TABLE DimOutcome (
        OutcomeKey int IDENTITY(1,1) NOT NULL,
        OutcomeName nvarchar(50) NULL,
        IsArrested bit NULL,
        ContrabandFound bit NULL,
        CONSTRAINT PK_OutcomeKey PRIMARY KEY (OutcomeKey)
    );
    CREATE TABLE FactStop (
        StopKey nvarchar(13) NOT NULL,
        StopDateKey int NOT NULL,
        StopTimeKey int NOT NULL,
        PDKey int NOT NULL,
```

```
            StopLocationKey int NOT NULL,
            DriverKey int NOT NULL,
            ActionKey int NOT NULL,
            CauseKey int NOT NULL,
            OutcomeKey int NOT NULL,
            StopDateTime DATETIME NOT NULL,
            StopDate DATE NOT NULL,
            StopTime TIME NOT NULL,
            CONSTRAINT PK_StopKey PRIMARY KEY (StopKey),
            CONSTRAINT FK_StopDate FOREIGN KEY (StopDateKey)
                  REFERENCES DimDate (DateKey),
            CONSTRAINT FK_StopTime FOREIGN KEY (StopTimeKey)
                  REFERENCES DimTime (TimeKey),
            CONSTRAINT FK_PDKey FOREIGN KEY (PDKey)
                  REFERENCES DimPoliceDepartament (PDKey),
            CONSTRAINT FK_LocationKey FOREIGN KEY (StopLocationKey)
                  REFERENCES DimLocation (LocationKey),
            CONSTRAINT FK_DriverKey FOREIGN KEY (DriverKey)
                  REFERENCES DimDriver (DriverKey),
            CONSTRAINT FK_ActionKey FOREIGN KEY (ActionKey)
                  REFERENCES DimAction (ActionKey),
            CONSTRAINT FK_CauseKey FOREIGN KEY (CauseKey)
                  REFERENCES DimCause (CauseKey),
            CONSTRAINT FK_OutcomeKey FOREIGN KEY (OutcomeKey)
                  REFERENCES DimOutcome (OutcomeKey)
      );
```

Wygenerowany przez Microsoft SQL Server Management Studio schemat bazy został przedstawiony na rys. 2.

Rysunek 2. Schemat bazy danych utworzonej za pomocą skryptu (tab. 6.)

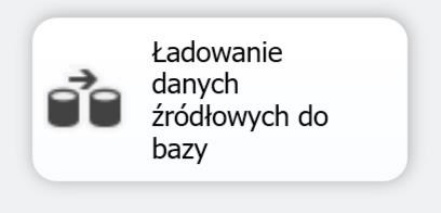
5.2. Specyfikacja procesów ETL (Control Flow + Data Flow)

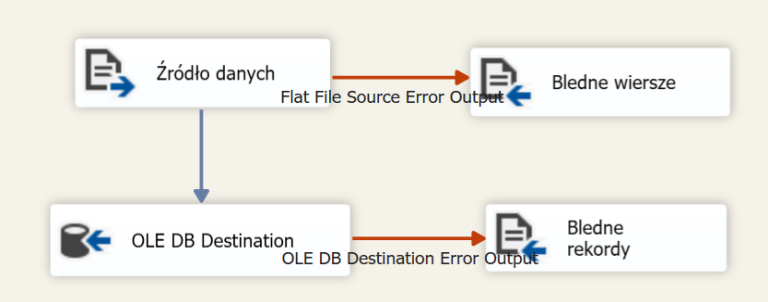
Zostały zdefiniowane następujące pakiety:

1. Ładowanie danych źródłowych do bazy z uwzględnieniem odpowiednich typów.
2. Tworzenie tabel tymczasowych.
3. Aktualizacja danych - proces ETL.

Pakiet 1: Ładowanie danych źródłowych do bazy

Pakiet został stworzony z powodu problemów, jakie pojawiały się w późniejszym importowaniu danych do tabel tymczasowych. Przepływ danych realizujący import danych źródłowych do bazy danych został przedstawiony na rys. 3.

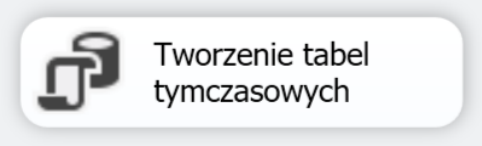




Rysunek 3. Przepływ danych (ang. Data flow) realizujący import danych źródłowyxh

Pakiet 2: Tworzenie tabel tymczasowych

W tym etapie za pomocą Execute SQL Task zostaje wykonany skrypt SQL tworzący tabele tymczasowe potrzebne w realizacji kolejnego etapu.



Skrypt SQL wykonywany przez ten element SSIS umieszczony jest w tabeli 7.

Tabela 7. Skrypt tworzący tabele tymczasowe w hurtowni danych

```
CREATE TABLE TMPDimDate (
        DateKey int IDENTITY(1,1) NOT NULL,
        FullStopDate DATE NOT NULL,
        DayNumberOfMonth tinyint NOT NULL,
        MonthNumberOfYear tinyint NOT NULL,
        CalendarYear smallint NOT NULL,
        CalendarQuarter tinyint NOT NULL,
        CalendarSemester tinyint NOT NULL
        CONSTRAINT PK_TMPDateKey PRIMARY KEY (DateKey)
```

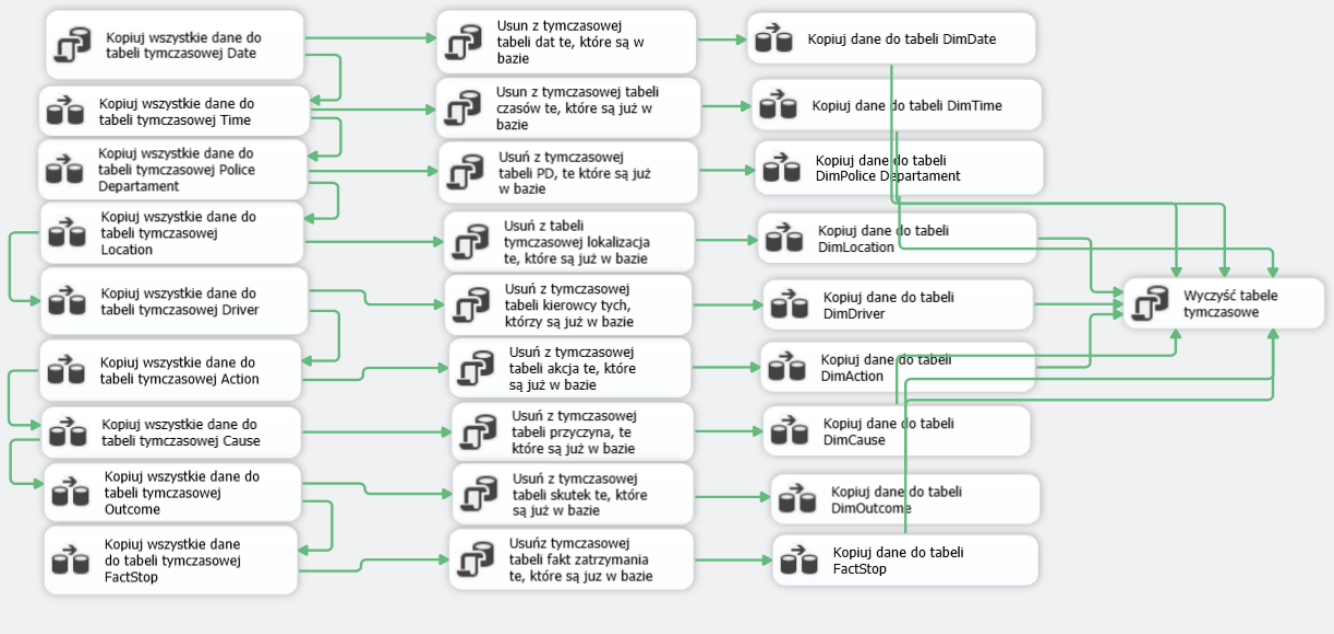
```
);


CREATE TABLE TMPDimTime (
	TimeKey int IDENTITY(1,1) NOT NULL,
	FullStopTime TIME NOT NULL,
	CONSTRAINT PK_TMPTimeKey PRIMARY KEY (TimeKey)
);
CREATE TABLE TMPDimPoliceDepartament (
	PDKey int IDENTITY(1,1) NOT NULL,
	PDName nvarchar(50) NULL
	CONSTRAINT PK_TMPPDKey PRIMARY KEY (PDKey)
);
CREATE TABLE TMPDimLocation (
	LocationKey int IDENTITY(1,1) NOT NULL,
	State nvarchar(2) NULL,
	CountyName nvarchar(50) NULL,
	CountyFIPS int NULL,
	City nvarchar(50) NULL,


	CONSTRAINT PK_TMPLocationKey PRIMARY KEY (LocationKey)
);
CREATE TABLE TMPDimDriver (
	DriverKey int IDENTITY(1,1) NOT NULL,
	Gender nvarchar(1) NULL,
	Age int NULL,
	Race nvarchar(50) NULL,
	CONSTRAINT PK_TMPDriverKey PRIMARY KEY (DriverKey)
);
```

```
CREATE TABLE TMPDimAction (
	ActionKey int IDENTITY(1,1) NOT NULL,
	SearchConducted bit NULL,
	SearchType nvarchar(50) NULL
	CONSTRAINT PK_TMPActionKey PRIMARY KEY (ActionKey)
);
CREATE TABLE TMPDimCause (
	CauseKey int IDENTITY(1,1) NOT NULL,
	ViolationName nvarchar(50) NULL
	CONSTRAINT PK_TMPCauseKey PRIMARY KEY (CauseKey)
);
CREATE TABLE TMPDimOutcome (
	OutcomeKey int IDENTITY(1,1) NOT NULL,
	OutcomeName nvarchar(50) NULL,
	IsArrested bit NULL,
	ContrabandFound bit NULL,
	CONSTRAINT PK_TMPOutcomeKey PRIMARY KEY (OutcomeKey)
);
CREATE TABLE TMPFactStop (
	StopKey nvarchar(13) NOT NULL,
	StopDateKey int NOT NULL,
	StopTimeKey int NOT NULL,
	PDKey int NOT NULL,
	StopLocationKey int NOT NULL,
	DriverKey int NOT NULL,
	ActionKey int NOT NULL,
	CauseKey int NOT NULL,
	OutcomeKey int NOT NULL,
```

```
            StopDateTime DATETIME NOT NULL,
            StopDate DATE NOT NULL,
            StopTime TIME NOT NULL,
            CONSTRAINT PK_TMPStopKey PRIMARY KEY (StopKey),
            CONSTRAINT FK_TMPStopDate FOREIGN KEY (StopDateKey)
                  REFERENCES TMPDimDate (DateKey),
            CONSTRAINT FK_TMPStopTime FOREIGN KEY (StopTimeKey)
                  REFERENCES TMPDimTime (TimeKey),
            CONSTRAINT FK_TMPPDKey FOREIGN KEY (PDKey)
                  REFERENCES TMPDimPoliceDepartament (PDKey),
            CONSTRAINT FK_TMPLocationKey FOREIGN KEY (StopLocationKey)
                  REFERENCES TMPDimLocation (LocationKey),
            CONSTRAINT FK_TMPDriverKey FOREIGN KEY (DriverKey)
                  REFERENCES TMPDimDriver (DriverKey),
            CONSTRAINT FK_TMPActionKey FOREIGN KEY (ActionKey)
                  REFERENCES TMPDimAction (ActionKey),
            CONSTRAINT FK_TMPCauseKey FOREIGN KEY (CauseKey)
                  REFERENCES TMPDimCause (CauseKey),
            CONSTRAINT FK_TMPOutcomeKey FOREIGN KEY (OutcomeKey)
                  REFERENCES TMPDimOutcome (OutcomeKey)
      );
```

Pakiet 3: Aktualizacja danych - proces ETL

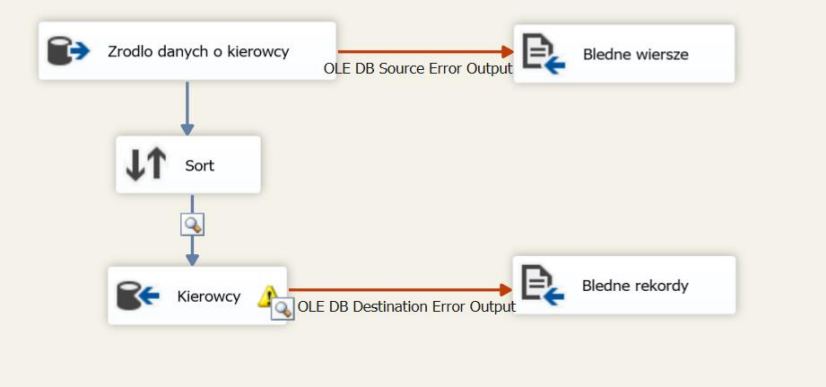
Control Flow:



Rysunek 4. ???

Control Flow składa się z 3 etapów:

- **Extract** - czyli pierwsza ‘kolumna’ polega na kopiowaniu do tabel tymczasowych odpowiednich danych z otrzymanego źródła. W zasadzie wszystkie Data Flow Task wyglądają podobnie. Poniżej pokażę przykład kopiowania danych do tabeli tymczasowej Kierowcy:



Rysunek 5. ???

Sortowanie w SSIS posiada opcję usuwania duplikatów, którą zaznaczamy. Błędne rekordy oraz błędne wiersze trafiają do odpowiedniego źródła Flat File.

Jedynym wyjątkiem w module Extract jest ładowanie danych do tabeli tymczasowej Data (TMPDimDate), ponieważ odbywa się ono z wykorzystaniem SQL Task i wykonaniem poniższego skryptu SQL:

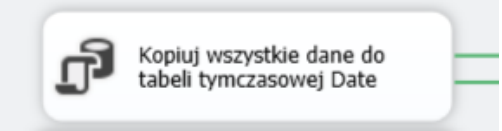


Tabela 8. ???

```
INSERT INTO [dbo].[TMPDimDate] ([FullStopDate], [DayNumberOfMonth],
[MonthNumberOfYear], [CalendarYear], [CalendarQuarter],
[CalendarSemester])
SELECT DISTINCT
      [stop_date],
      DATEPART(DAY, [stop_date]),
      DATEPART(MONTH, [stop_date]),
      DATEPART(YEAR, [stop_date]),
      DATEPART(QUARTER, [stop_date]),
      CASE
            WHEN DATEPART(quarter,[stop_date]) >= 3 THEN 2 ELSE 1
      END
FROM [dbo].[Data_Vermont];
```

Uwaga: Uzasadnienie wykorzystania skryptu , ponieważ chciałam równocześnie załadować odpowiedni dzień, miesiąc, rok, półrocze i kwartał powstałe za pomocą funkcji DATEPART, przyjmującej odpowiedni argument oraz [stop_date].

- **Transform** - w tym etapie poprzez odpowiednie skrypty SQL usuwamy powtórzone dane z tabel tymczasowych. Oznacza to, że dostając nowe dane nie wgrywamy drugi raz tych samych rekordów, tylko sprawdzamy, czy są one już w bazie (w tabelach docelowych, nie tymczasowych) i jeśli tak, to usuwamy je, co skutkuje dopisywaniu nowych rekordów, bez powtórzeń. Przykładowy skrypt dla tabeli TMPDriver wygląda następująco:

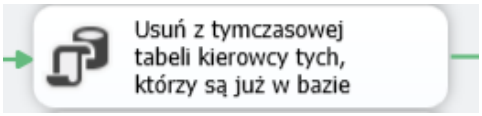
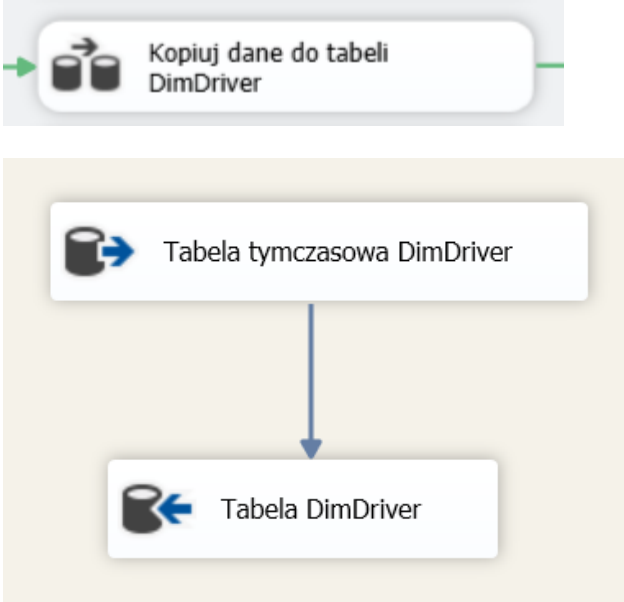


Tabela 9. ???

```
DELETE [dbo].[TMPDimDriver]
FROM [dbo].[TMPDimDriver]
      INNER JOIN [dbo].[DimDriver] ON
            [dbo].[DimDriver].Age = [dbo].[TMPDimDriver].Age
            AND [dbo].[DimDriver].Gender = [dbo].[TMPDimDriver].Gender
            AND [dbo].[DimDriver].Race = [dbo].[TMPDimDriver].Race;
```

Przygotowałam skrypt przedstawiony w tab. 9 z uwagi na występujące w hurtowni klucze sztuczne.

- **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 Data Flow dla Data Flow Task kopującego dane do tabeli docelowej DimDriver wygląda następująco:

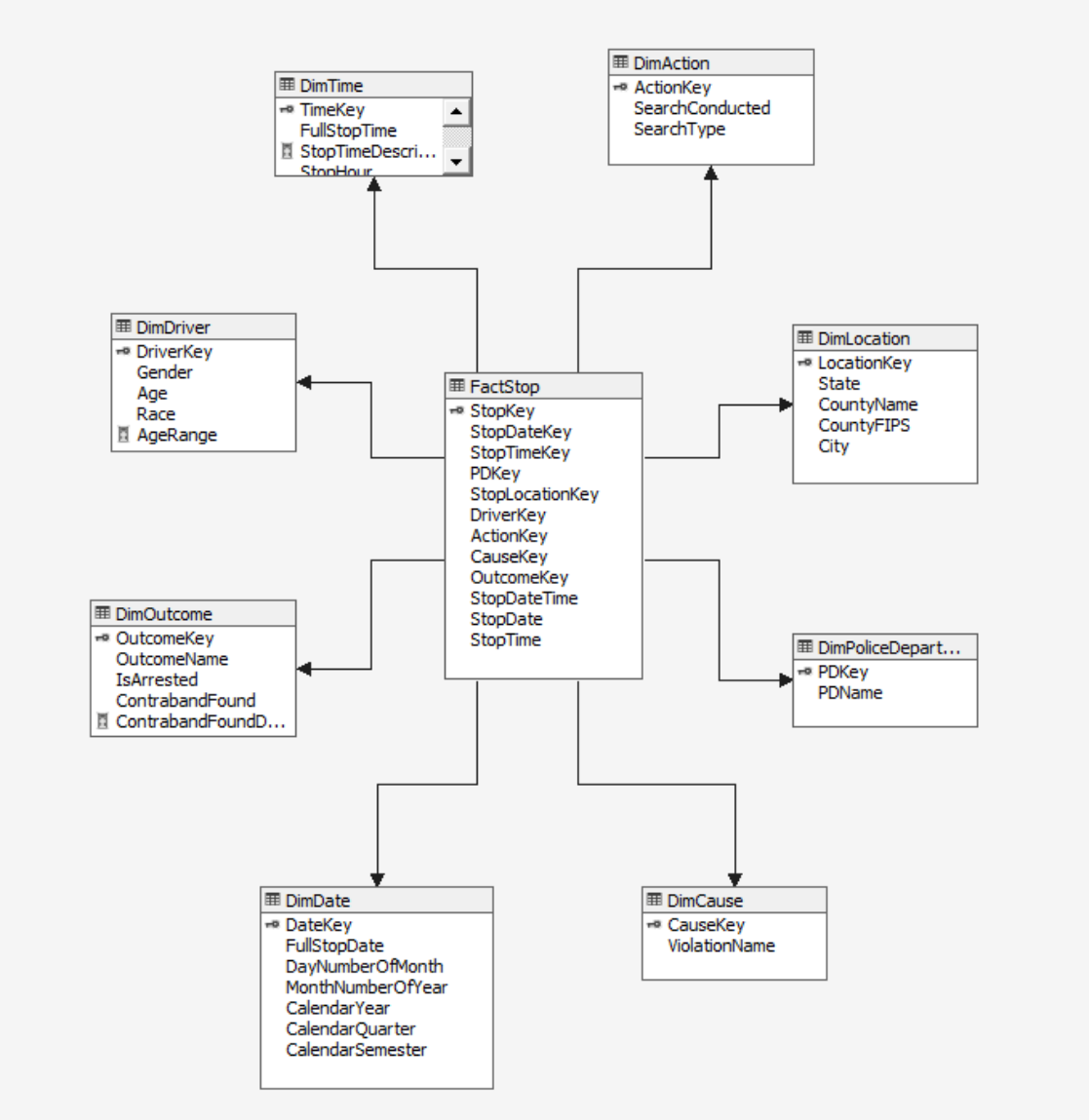


Rysunek 6. ???

Inspiracją dla procesu ETL przedstawionego w projekcie był slajd 4 z wykładu Tomasza Golańskiego dostępnego na BOARDzie dr Tuzinkiewicza, a także przykładowy projekt "Biuro Podróży" również dostępny w tym samym miejscu.

6. Implementacja modeli wielowymiarowych

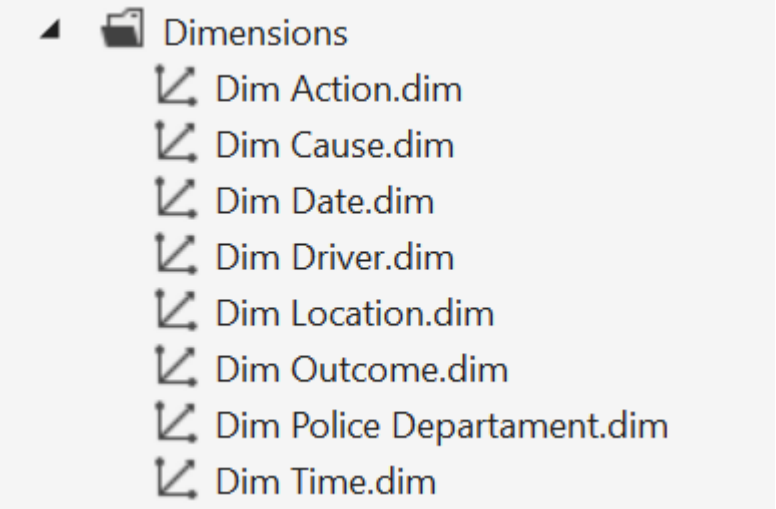
6.1. Widok danych



Rysunek 7. ???

6.2. Wymiary

Zdefiniowałam następujące wymiary:



Rysunek 8. ???

Dodatkowo, dodałam atrybuty pozwalające na inny sposób wyświetlania danych (Named Calculation) dla następujących wymiarów:

- DimTime:
  - StopTimeDescriptive - atrybut pozwalający na określenie pory dnia.

    Tabela 10. ???

```
CASE
      WHEN DATEPART(HOUR, [FullStopTime]) >= 6 AND DATEPART(HOUR,
[FullStopTime]) < 12 THEN 'Morning'
      WHEN DATEPART(HOUR, [FullStopTime]) >= 12 AND DATEPART(HOUR,
[FullStopTime]) < 18 THEN 'Afternoon'
      WHEN DATEPART(HOUR, [FullStopTime]) >= 18 AND DATEPART(HOUR,
[FullStopTime]) < 24 THEN 'Evening'
      WHEN DATEPART(HOUR, [FullStopTime]) >= 00 AND DATEPART(HOUR,
[FullStopTime]) < 6 THEN 'Night'
      ELSE 'No Data Provided'
 END
```

  - StopHour - atrybut pozwalający na określenie godziny.

    Tabela 11. ???

```
DATEPART(HOUR, [FullStopTime])
```

- DimOutcome:
    - ContrabandFoundDescriptive - zamienia wartości boolowskie (w naszym przypadku dodatkowo będące 0, 1) na wartości opisowe:

      Tabela 12. ???

```
CASE
      WHEN [ContrabandFound] = 0 THEN 'Not Found'
      WHEN [ContrabandFound] = 1 THEN 'Found'
      ELSE 'No Data Provided'
END
```

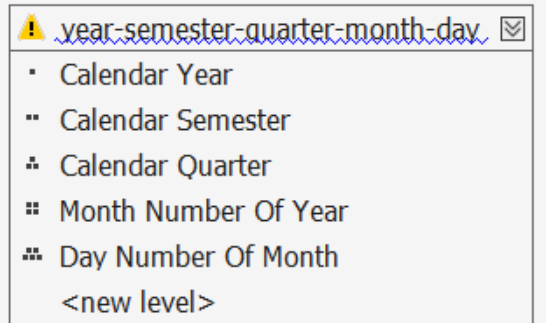
      Wprowadzony z powodu problemów z konwersją danych przy procesowaniu kostki.

- DimDriver:
    - AgeRange - zamienia wiek kierowcy na przedział wiekowy, bardziej przydatny z punktu widzenia późniejszej analizy

      Tabela 13. ???

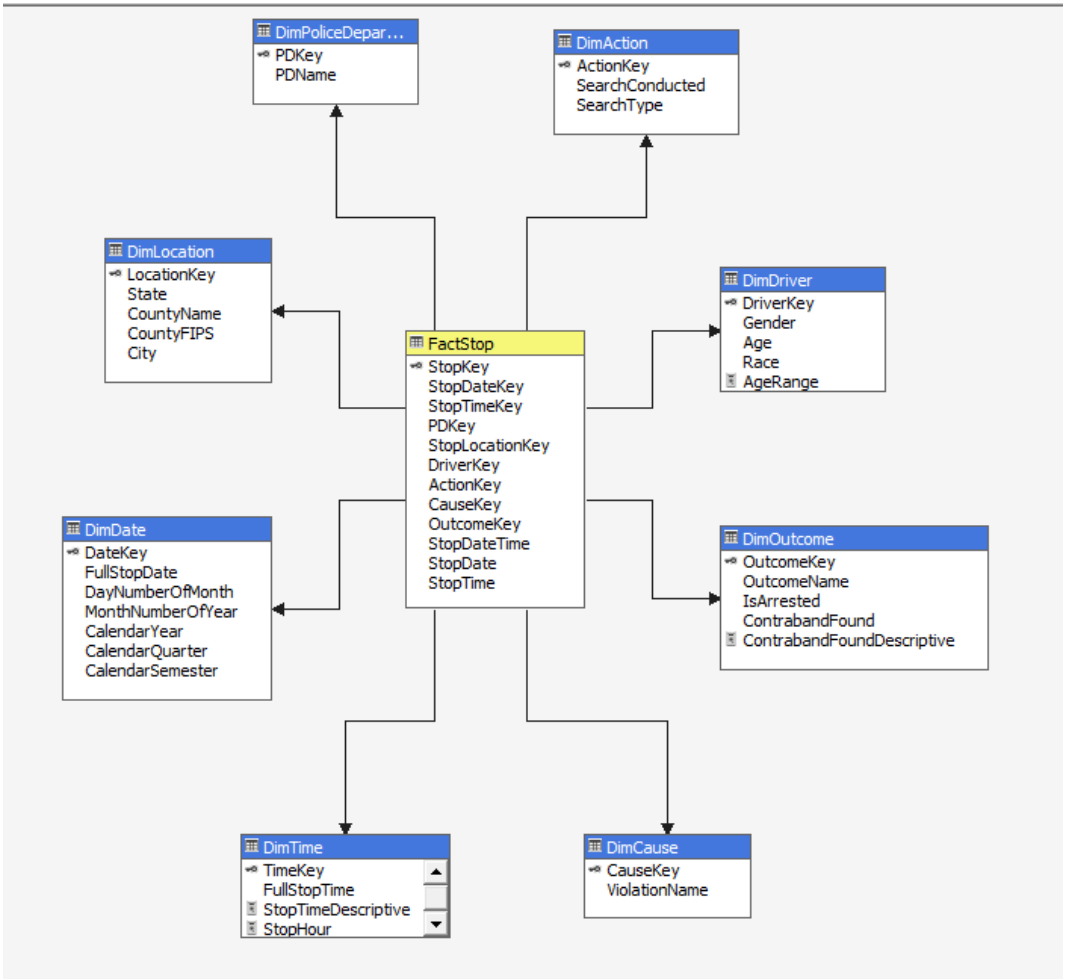
```
CASE
      WHEN [Age] < 16 THEN 'Under 16'
      WHEN [Age] >= 16 AND [Age] < 20 THEN '16 - 19'
      WHEN [Age] >= 20 AND [Age] < 30 THEN '20 - 29'
      WHEN [Age] >= 30 AND [Age] < 40 THEN '30 - 39'
      WHEN [Age] >= 40 AND [Age] < 50 THEN '40 - 49'
      WHEN [Age] >= 50 AND [Age] < 60 THEN '50 - 59'
      WHEN [Age] >= 60 AND [Age] < 70 THEN '60 - 69'
      WHEN [Age] >= 70 AND [Age] < 80 THEN '70 - 79'
      WHEN [Age] >= 80 THEN 'Over 80'
      ELSE 'No Data Provided'
END
```

Zdefiniowałam także następującą hierarchię dla wymiaru DimDate:



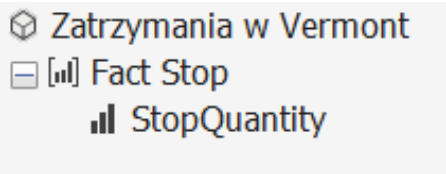
Rysunek 9. ???

6.3. Modele wielowymiarowe - Kostki



Rysunek 10. ???

Miara została zdefiniowana na etapie tworzenia kostki:



Rysunek 11. ???

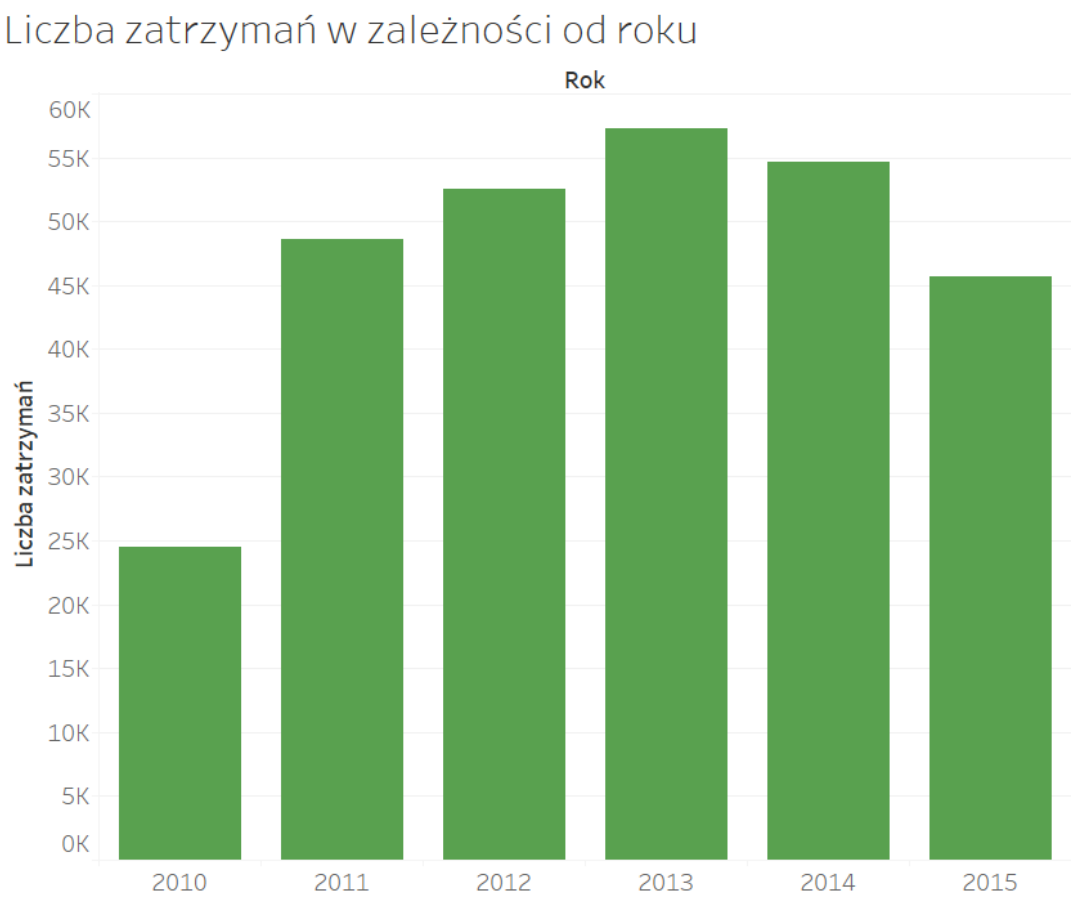
7. Analiza danych

Procesy analityczne wykonałam za pomocą Tableau Desktop i zrealizowałam według następującego scenariusza:

Liczba zatrzymań w zależności od:

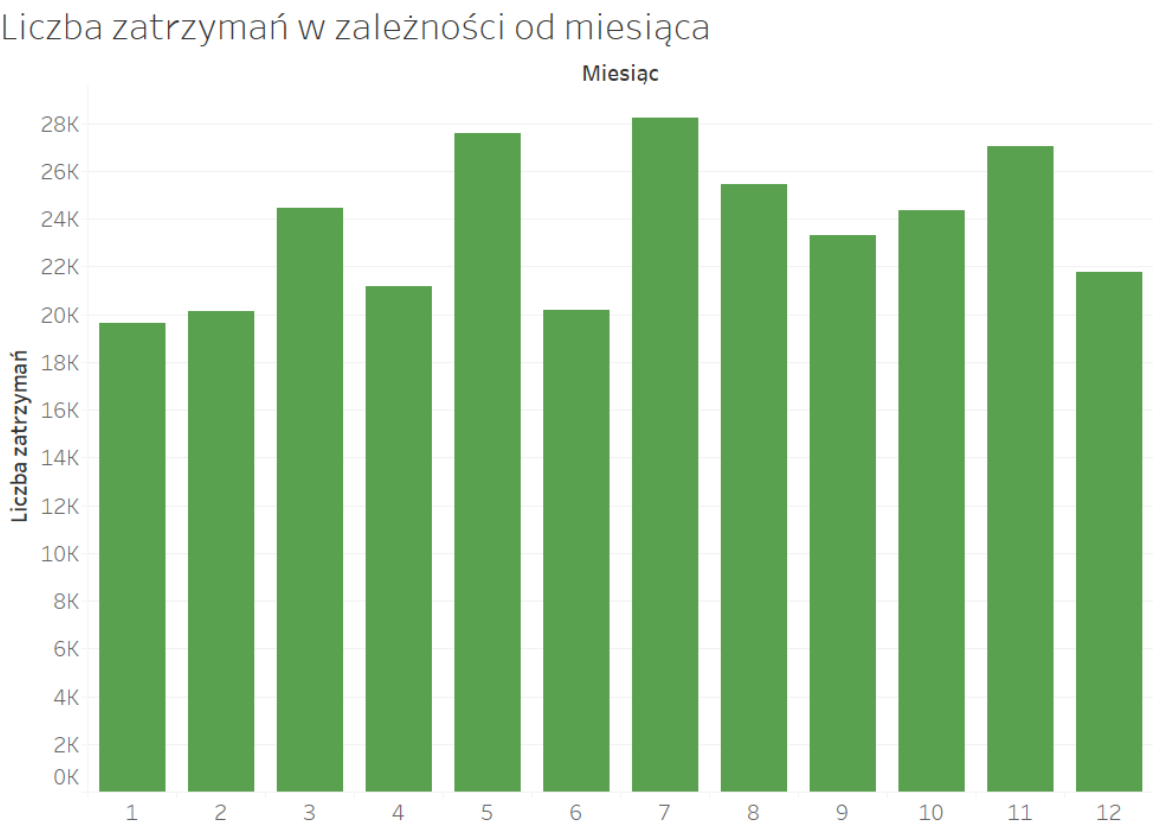
- Ogólne: rok, miesiąc, przynależność rasowa, lokalizacja, departament policji, wiek kierowcy.
- Szczegółowe: wiek + płeć, wiek + płeć + przynależność rasowa (bez uwzględnienia rasy białej), pora dnia + godzina, wykroczenie + pora dnia, pora dnia + przynależność rasowa, przeszukanie + znalezienie kontrabandy, przeszukanie + znalezienie kontrabandy + przynależność rasowa, skutki zatrzymania + wykroczenie, aresztowanie + najczęstsze wykroczenie + przynależność rasowa + płeć.

7.1. Realizacja procesów analitycznych



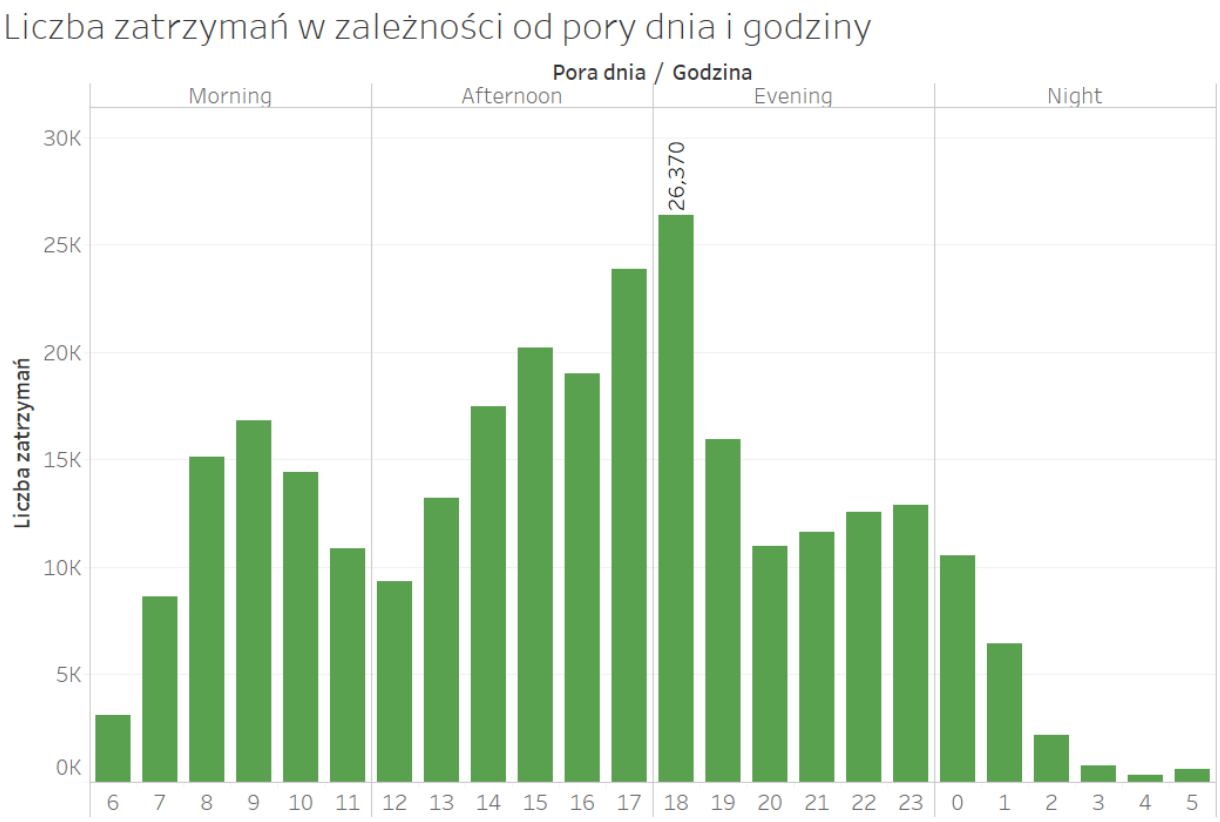
Rysunek 12. ???

Analizę rozpoczynam od sprawdzenia jak zmieniła się liczba zatrzymań na przestrzeni lat. Od roku 2010 do roku 2011 liczba zatrzymań wzrosła dwukrotnie i miała tendencję rosnącą aż do 2014 roku, kiedy to zaczęła spadać. Wzrost liczby zatrzymań a także jej spadek może być związany ze zmianami demograficznymi w badanym stanie. W roku 2010 populacja Vermont wynosiła 622 433 mieszkańców, w roku 2011- 626 979, 2012- 626 063, 2013 - 626 212, 2014 - 625 218 i 2015 - 625 197 ([1]), zatem gdy tylko następowała duża, czyli liczona w tysiącach zmiana liczebności populacji, zmieniała się także liczba zatrzymań.



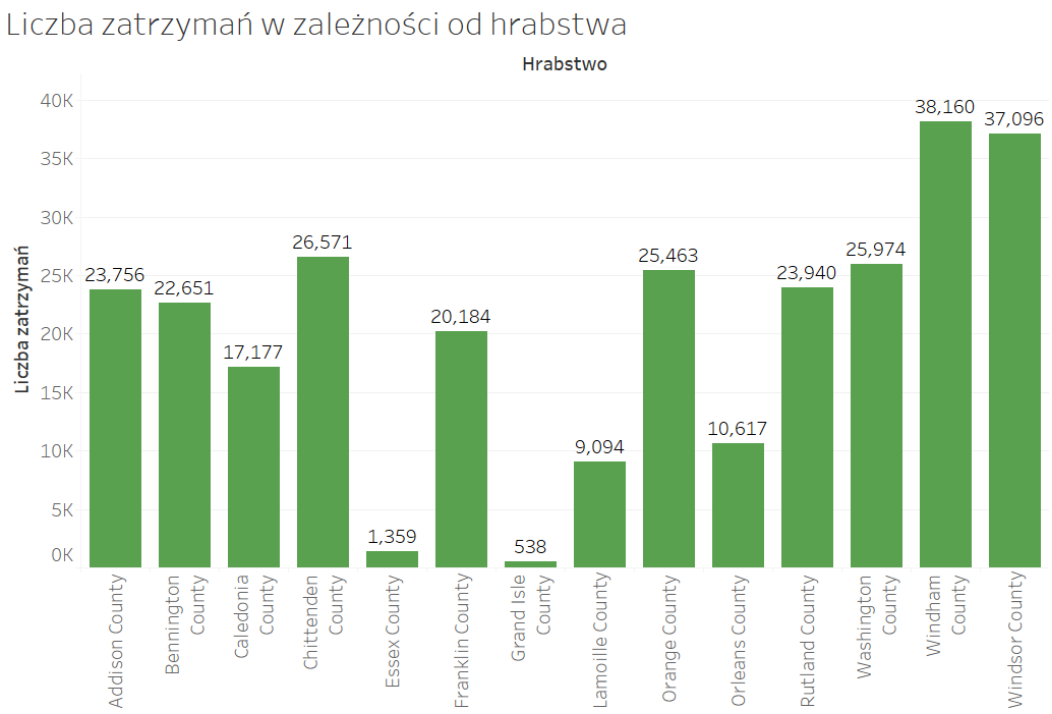
Rysunek 13. ???

Jeśli chodzi o rozkład liczby zatrzymań z uwzględnieniem miesięcy, możemy zauważyć, że jest ich najwięcej w miesiącach takich jak maj, lipiec, sierpień i listopad. Niewielka liczba zatrzymań w okresie zimowym może być związana z trudnymi warunkami atmosferycznymi. Zima w Vermont jest zazwyczaj śnieżna ([2]), co nie sprzyja rozwijaniu prędkości. Ciekawa jest liczba osób zatrzymanych w czerwcu, która jest znacząco mniejsza od liczby zatrzymań w maju i lipcu. Niestety nie udało mi się znaleźć informacji wyjaśniających to zjawisko.



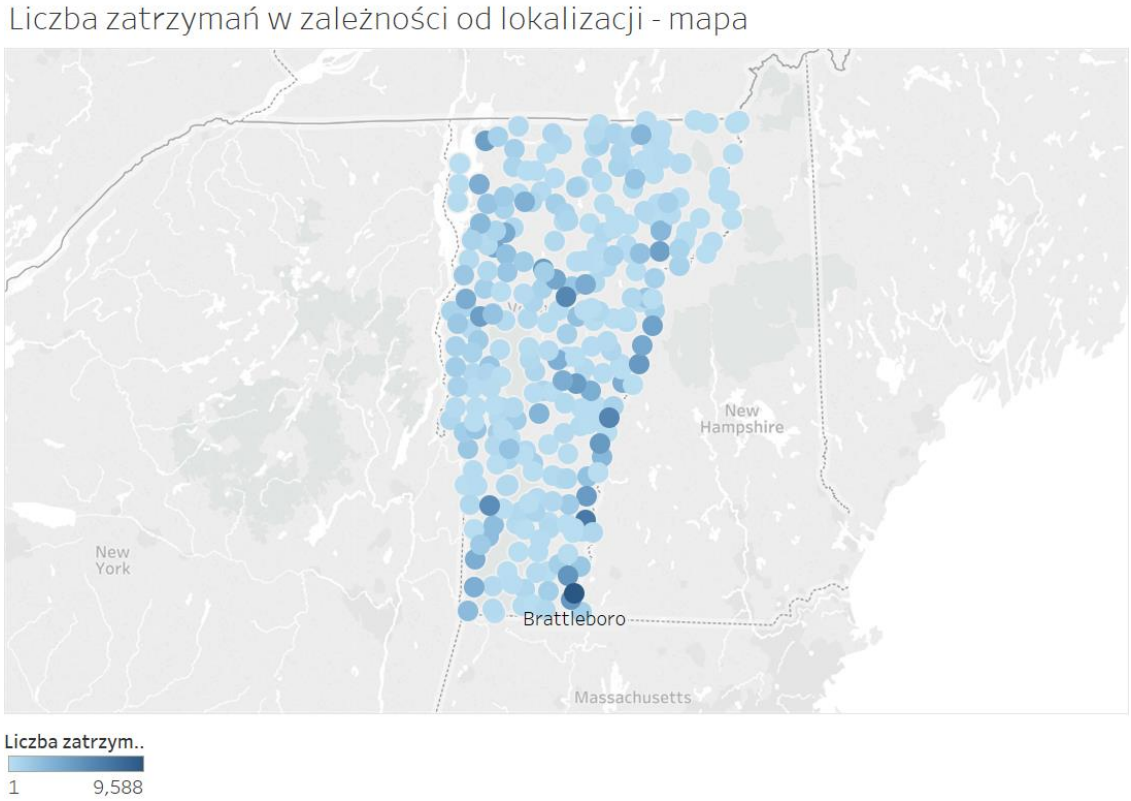
Rysunek 14. ???

Jeśli chodzi o porę dnia i godzinę, największa liczba zatrzymań jest o godzinie 18. Uważam, że wynika to z faktu popularnego trybu pracy w Stanach Zjednoczonych zwanego “Nine To Five”, co oznacza, że większość osób kończy pracę o godzinie 17, zatem największa liczba zatrzymań o godzinie 18 jest uzasadniona.

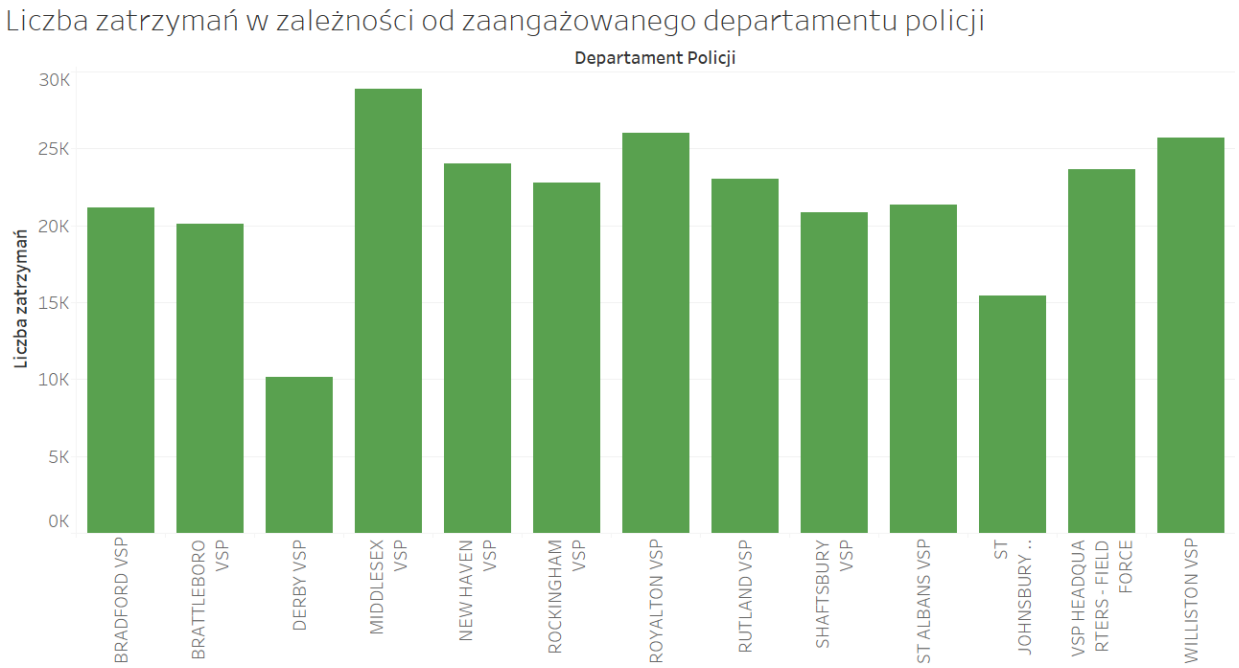


Rysunek 15. ???

Do największej liczby zatrzymań dochodzi w hrabstwach Windsor i Windham, które co prawda nie są najbardziej zaludnione, ale leżą na południowym - wschodzie, gdzie prawdopodobnie z powodu sąsiedztwa z New Hampshire dochodzi do największej liczby zatrzymań, co ilustruje poniższa mapka.

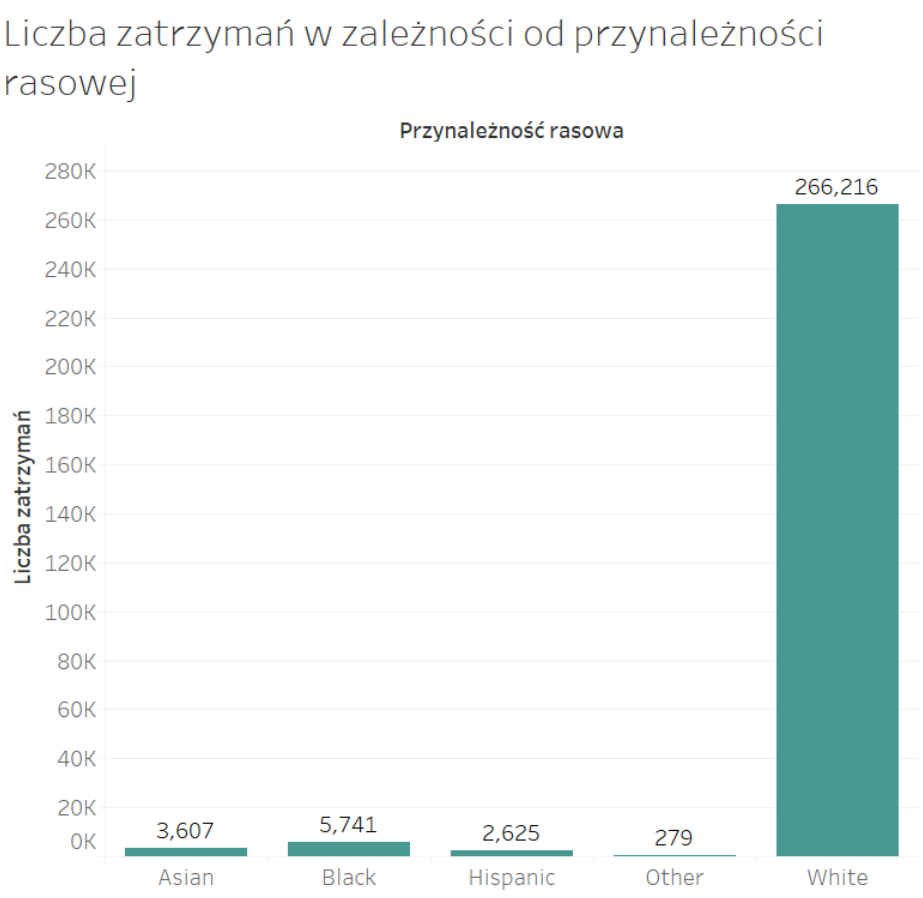


Rysunek 16. ???



Rysunek 17. ???

Najbardziej zaangażowanym w przestrzeganie prawa jest Middlesex VSP (Vermont State Police), który w ciągu badanych 5 lat zatrzymał 28 861 kierowców. Departament ten znajduje się w hrabstwie Washington i obejmuje środkową część stanu.

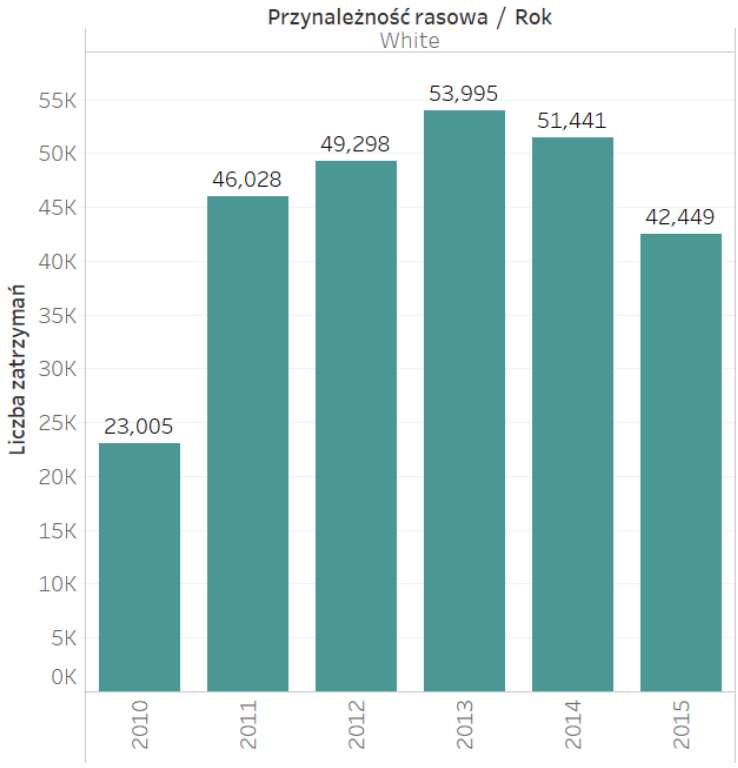


Rysunek 18. ???

Jeśli chodzi o liczbę zatrzymań w zależności od przynależności rasowej, warto zaznaczyć, że w zasadzie na przestrzeni 5 badanych lat, jak i w dzisiejszych czasach, około 90% ludności stanowią osoby o białej przynależności rasowej. Również widać to w liczbie zatrzymań - prawie 94% zatrzymanych osób jest rasy białej. Następnie najczęściej zatrzymywane są osoby czarnoskóre, a następnie kierowcy pochodzenia azjatyckiego i latynoskiego. Co ciekawe, większy procent populacji Vermont na przestrzeni lat stanowią osoby pochodzenia azjatyckiego niż osoby o czarnej przynależności rasowej, a to ta druga grupa jest zatrzymywana częściej. Może to jednak wynikać z różnic kulturowych.

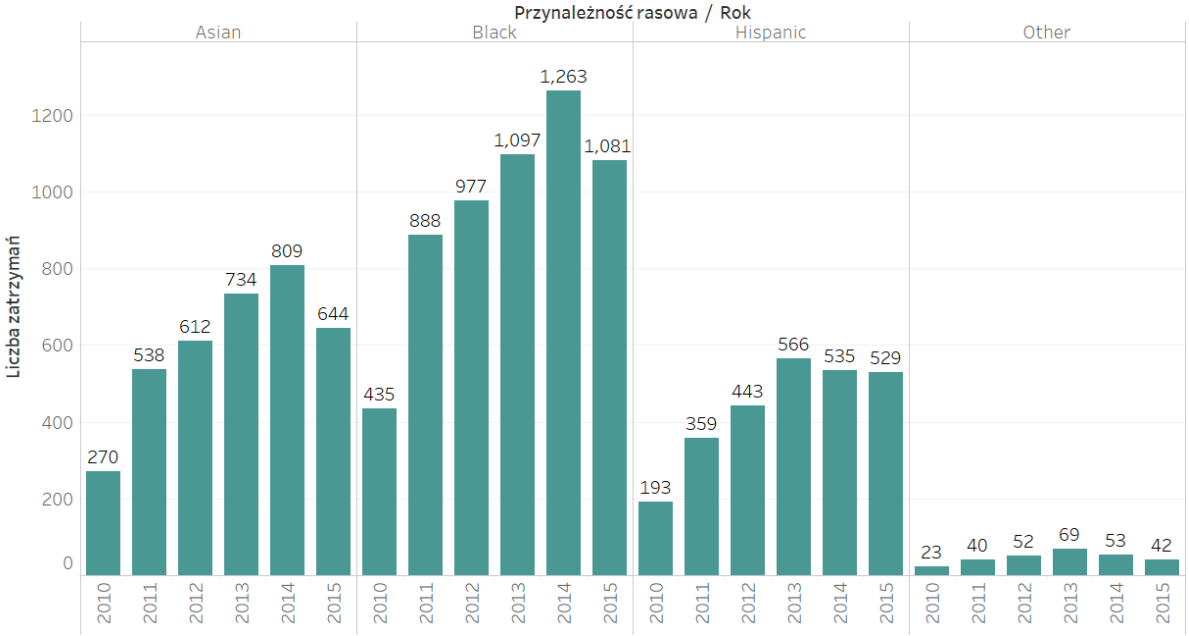
Dwa kolejne wykresy pokazują rozłożenie liczby zatrzymań w zależności od przynależności rasowej osobno dla rasy białej i pozostałych. Interesujący okazuje się fakt, że gdy w roku 2014 dla wszystkich grup liczba zatrzymań spadła, dla grupy osób czarnoskórych wzrosła.

Liczba zatrzymań w zależności od lat dla rasy białej

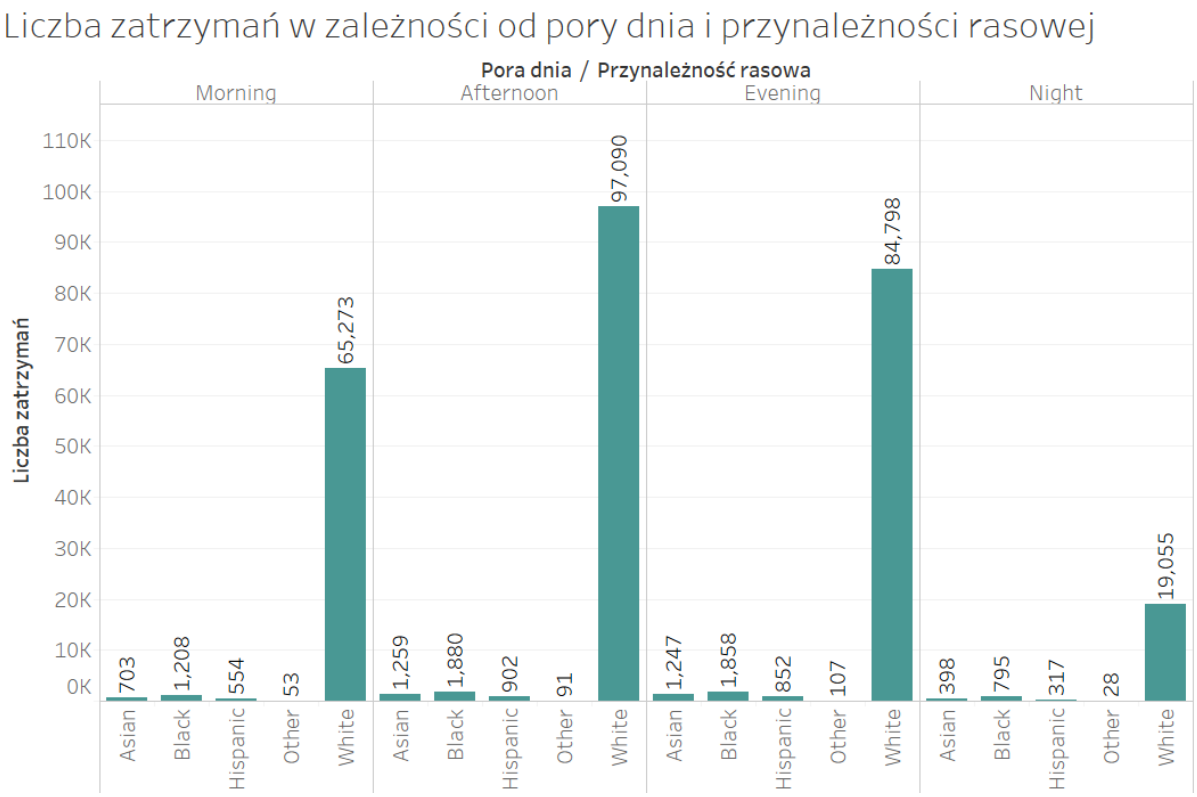


Rysunek 19. ???

Liczba zatrzymań w zależności od przynależności rasowej i lat (bez uwzględnienia rasy białej)

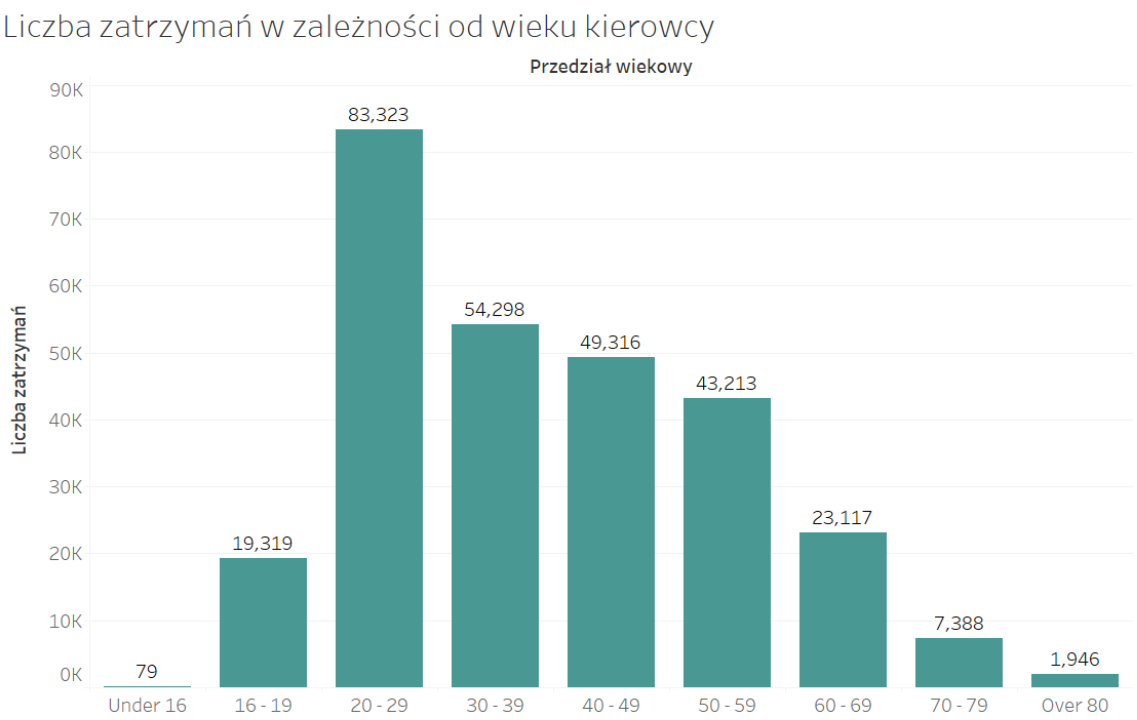


Rysunek 20. ???



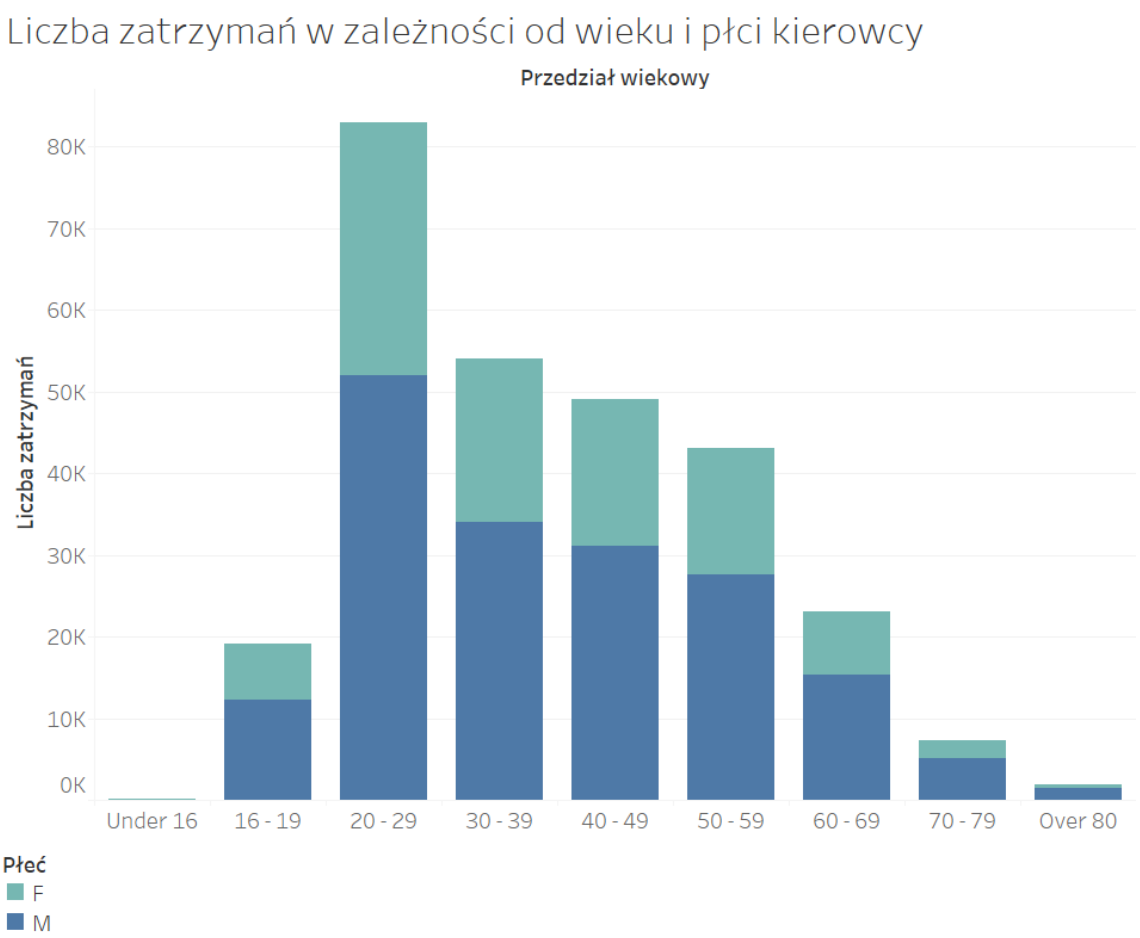
Rysunek 21. ???

Badanie zatrzymań pod względem pory dnia i przynależności rasowej nie pokazuje zaskakujących faktów. Podobnie jak dla wykresu liczby zatrzymań dla pory dnia i godziny, najwięcej zatrzymań ma miejsce po południu, czyli od godziny 12 do 18. Nie ma również anomalii pod względem rasowym.



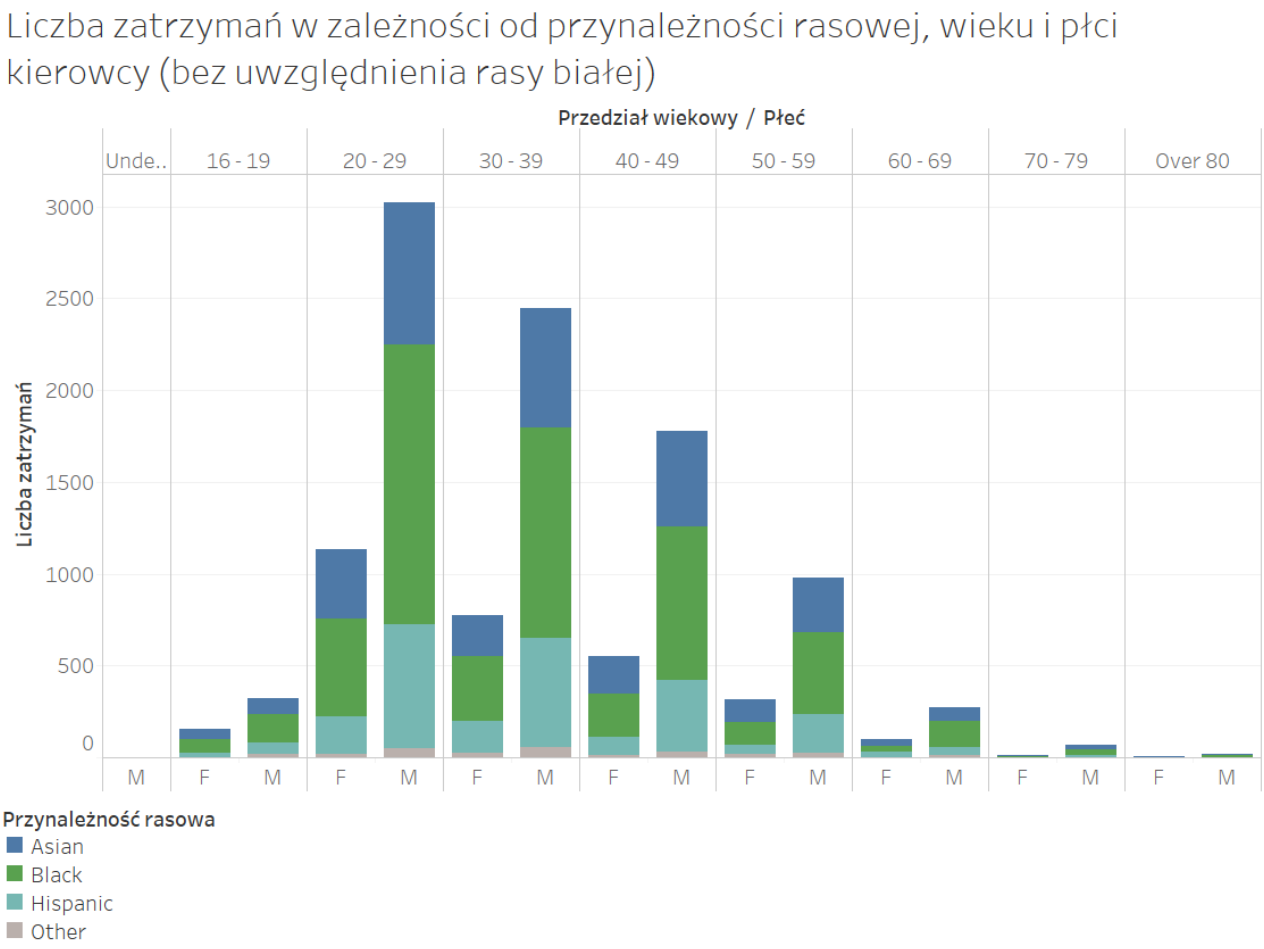
Rysunek 22. ???

Największa liczba zatrzymanych kierowców to osoby w wieku 20-29 lat. W Stanach Zjednoczonych prawo jazdy można uzyskać w wieku lat 16, jednakże to grupa osób z przedziału wiekowego 20-29 jest najliczniejsza w stanie Vermont, poza osobami starszymi, ale one z racji wieku stosunkowo rzadko popełniają wykroczenia.



Rysunek 23. ???

Wykres wieku i płci pokazuje, że nawet w najliczniejszych grupach wiekowych rzadziej dochodzi do zatrzymania kobiety niż mężczyzny. Może to wynikać z faktu, że kobiety są ostrożniejsze na drodze lub rzadziej jeżdżą samochodem.



Rysunek 24. ???

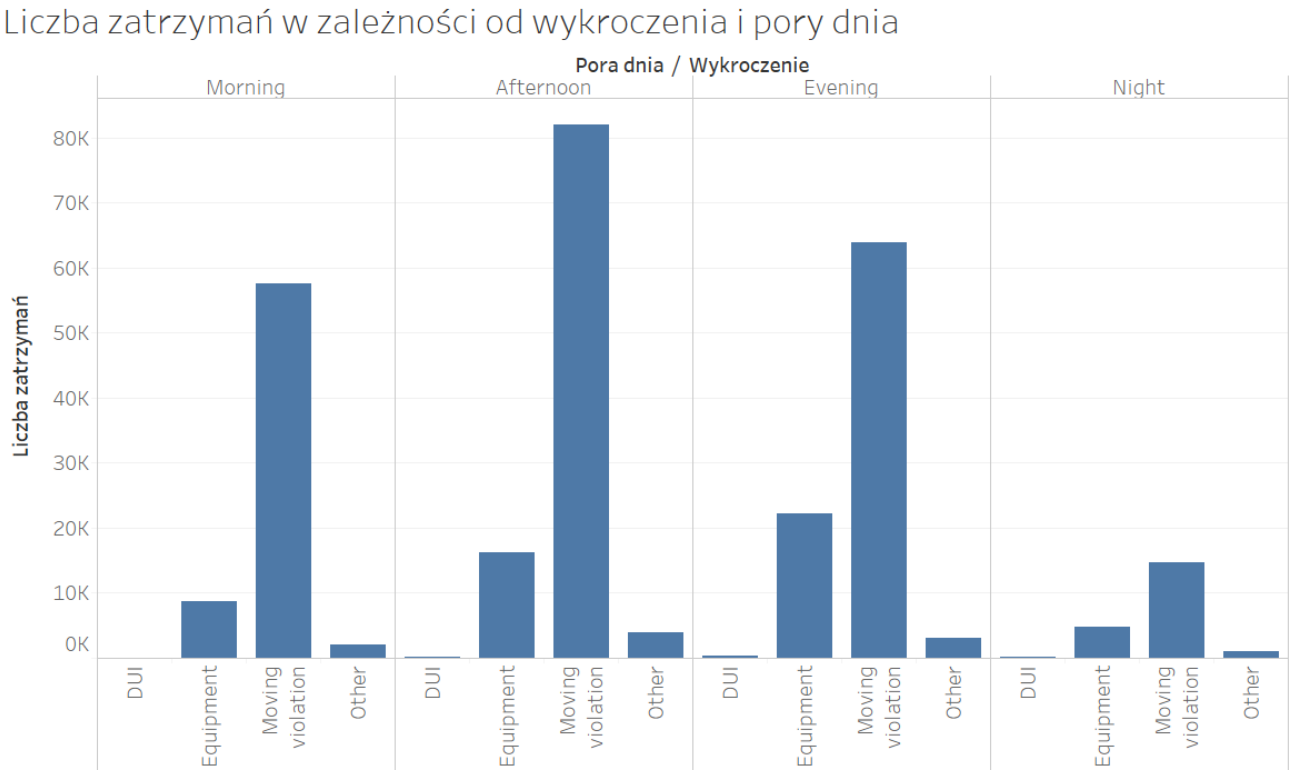
Powyższy wykres ilustruje rozkład zatrzymań z uwzględnieniem kobiet i mężczyzn a także przynależności rasowej bez uwzględnienia najliczniejszej rasy białej. Procentowe porównanie liczby zatrzymanych mężczyzn i kobiet dla najliczniejszej grupy 20-29 przedstawię w tabeli:

Tabela 14. ???

Przynależność rasowa	% zatrzymanych kobiet	% zatrzymanych mężczyzn
White	37%	63%
Asian	32%	68%
Black	25%	75%
Hispanic	23%	77%
Other	28%	62%

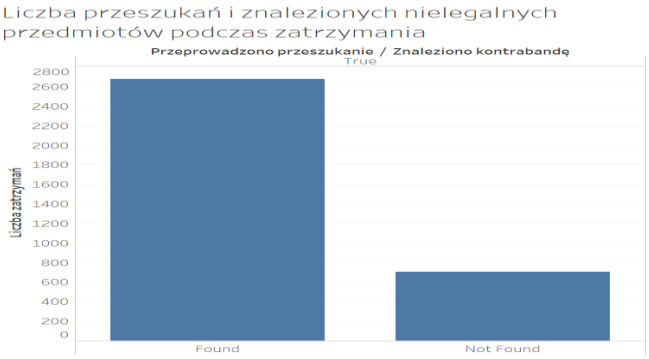
Największy procent zatrzymanych kobiet jest rasy białej, a najmniejszy dla grupy osób pochodzenia hiszpańskiego, latynoskiego. Prawdopodobnie wynika to z różnic kulturowych, a także przychodów. Najbogatszymi grupami etnicznymi w Stanach Zjednoczonych są osoby pochodzenia azjatyckiego

jak i ludzie biali ([3]). To oznacza, że więcej osób z tych grup społecznych może pozwolić sobie na samochód.



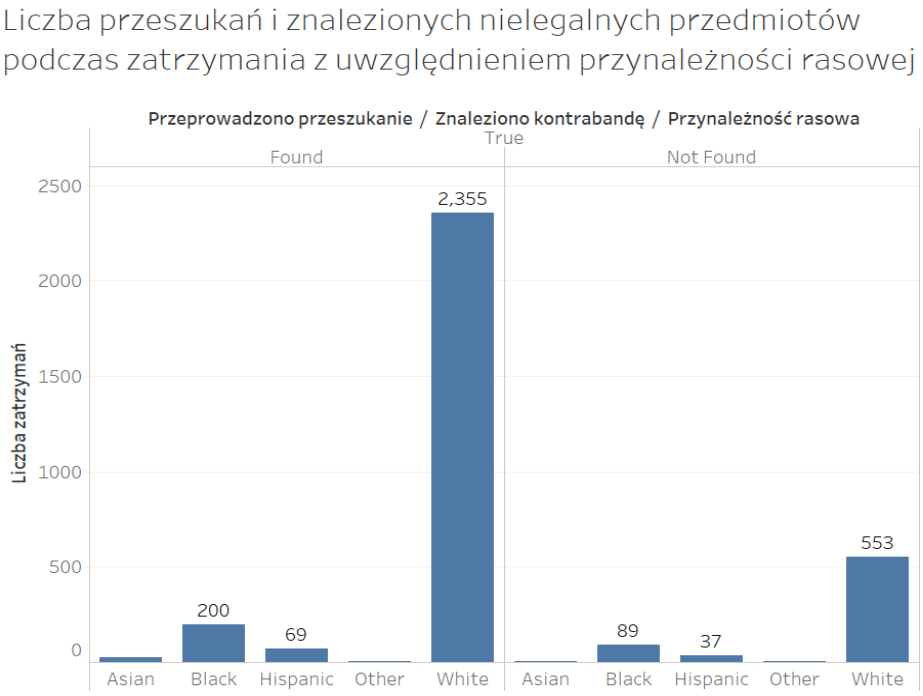
Rysunek 25. ???

Jeśli chodzi o badanie wykroczeń, najczęściej popełnianym wykroczeniem jest tak zwane “Moving Violation”, czyli szeroko pojęte naruszenie zasad ruchu drogowego. Następnym popularnym wykroczeniem jest “Equipment”, czyli przykładowo brak tablicy rejestracyjnej. “DUI” oznacza “Driving under the influence”, czyli jazdę pod wpływem substancji psychoaktywnych. Wykroczeń z tej trzeciej kategorii jest bardzo niewiele, jednakże zdarzają się, szczególnie w godzinach wieczornych, czyli od godziny 18 do 23.



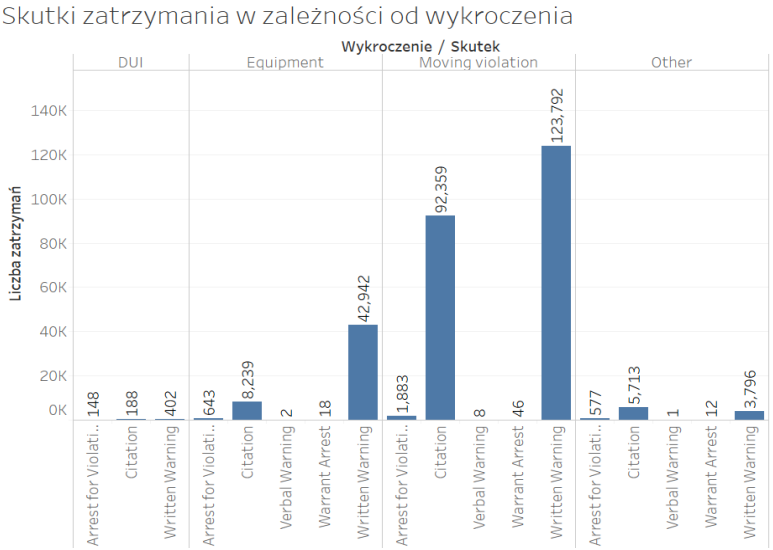
Rysunek 26. ???

Liczba przeszukań pojazdu uwzględniona w źródle danych jest stosunkowo niewielka, stanowi tylko 3385 przypadków, z czego 2681 to przeszukania zakończone znalezieniem nielegalnych przedmiotów. Daje to możliwość postawienia tezy, że przeszukania są uzasadnione i skuteczne.



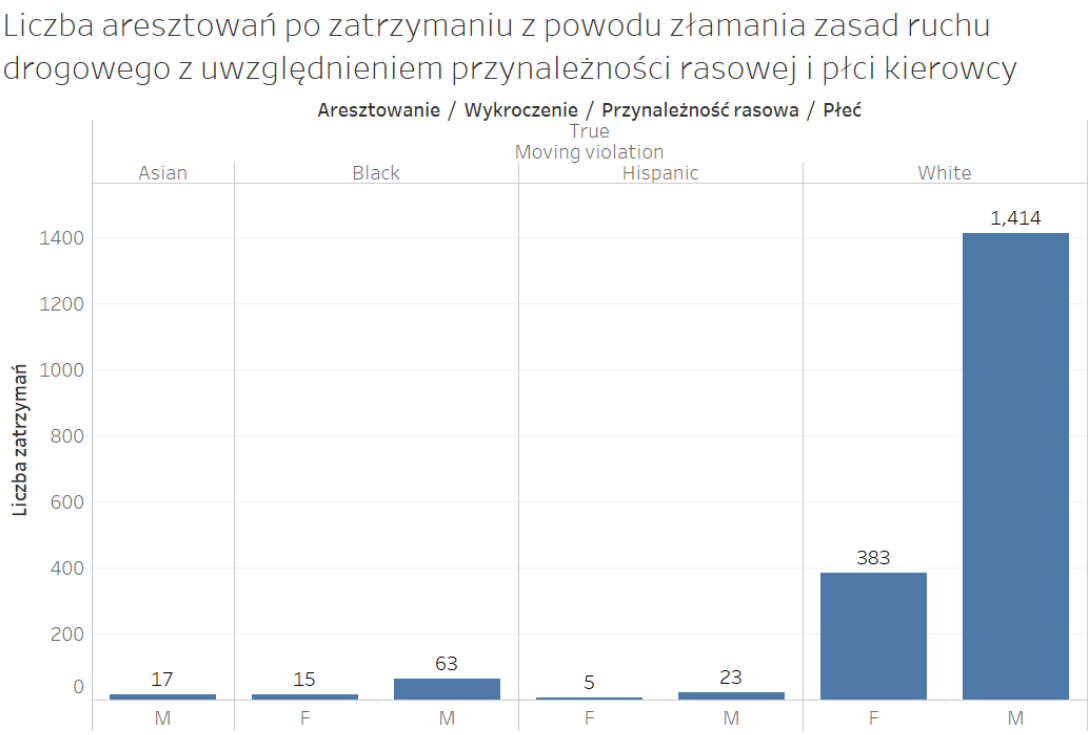
Rysunek 27. ???

Jeśli chodzi o liczbę przeszukań z uwzględnieniem przynależności rasowej, około 81% przeszukań pojazdu osób rasy białej zakończone zostało znalezieniem nielegalnych przedmiotów, natomiast 69% wszystkich przeszukań pojazdu osoby czarnej skutkowało znalezieniem nielegalnych przedmiotów.



Rysunek 28. ???

Najczęstszym skutkiem zatrzymania, w zasadzie niezależnie od wykroczenia jest tak zwane "Written Warning", czyli ostrzeżenie nie pociągające za sobą skutków prawnych czy finansowych. Drugim częstym skutkiem jest "Citation", czyli mandat. Aresztowanie zdarza się najczęściej dla naruszenia zasad ruchu drogowego, co zostanie pokazane na kolejnym wykresie.



Rysunek 29. ???

Podobnie jak w przypadku poprzednich wykresów, największą grupę zatrzymanych i aresztowanych stanowią biali mężczyźni, około 21% wszystkich aresztowań osób rasy białej stanowią kobiety. Co ciekawe, jeśli chodzi o kierowców pochodzenia azjatyckiego, w ciągu 5 lat zarejestrowano tylko 17 aresztowań samych mężczyzn. Wśród osób rasy czarnej doszło do 78 aresztowań, natomiast wśród grupy pochodzenia hiszpańskiego odnotowano ich 28. Warto zaznaczyć, że w Stanach Zjednoczonych, w większości stanów można zostać aresztowanym łamiąc każde prawo drogowe, poza przekroczeniem prędkości, za które karą jest mandat lub ostrzeżenie. ([4])

7.2. Podsumowanie - wnioski z analizy

Analiza danych miała pomóc w znalezieniu odpowiedzi na następujące pytania:

- **Czy dochodzi do większej liczby zatrzymań osób czarnoskórych lub ogólnie innej rasy niż biała?**
  Nie, w stanie Vermont dominują zatrzymania osób rasy białej.

- **Czy potwierdzony zostaje stereotyp zatrzymania czarnoskórego mężczyzny po zmroku?**
  Nie, tak samo jak po zmroku spada liczba zatrzymanych osób rasy białej, tak samo spada liczba zatrzymanych osób rasy czarnej, więc nie zostaje potwierdzony wspomniany stereotyp.
- **W jakich miejscach dochodzi do największej liczby wykroczeń?**
  Do największej liczby wykroczeń dochodzi w szczególności na granicy Vermont z New Hampshire, czyli w takich hrabstwach jak Windsor i Windham.

- **O jakiej porze dnia dochodzi do największej liczby wykroczeń?**
  Do największej liczby wykroczeń dochodzi po południu, czyli między 12 a 17, a najwięcej zatrzymań odbywa się o godzinie 18. Z czego to może wynikać?
- **Jakie są najczęstsze przyczyny i skutki zatrzymań?**
  Najczęstszą przyczyną zatrzymania jest naruszenie przepisów ruchu drogowego, natomiast najczęstszymi skutkami są: pisemne ostrzeżenie i mandat. Niestety zbiór danych nie dostarcza nam dokładnych informacji dotyczących rozpoznania naruszonych przepisów.
- **W jakim stopniu przeszukania skutkują znalezieniem kontrabandy?**
  W znacznym zakresie, bowiem około 79% wszystkich przeszukań skutkuje znalezieniem kontrabandy, czyli nielegalnych przedmiotów.

Przeprowadzona analiza pozwoliła odpowiedzieć na postawione pytania, a także na wiele innych. Jednakże, należy zaznaczyć, że Vermont nie stanowi odpowiedniego odnośnika do całych Stanów Zjednoczonych ze względu na ubogie zróżnicowanie demograficzne. Aby odpowiedzieć na postawione pytania, ale przenosząc je na skalę całego kraju, należałoby przeanalizować zatrzymania również w pozostałych stanach, szczególnie tych większych i bardziej zróżnicowanych.

8. Wnioski końcowe z realizacji projektu

8.1. Problemy

Podczas realizacji projektu spotkałam się z dwoma większymi problemami i oczywiście całą masą tych mniejszych. Do większych problemów zaliczam model konceptualny oraz proces ETL. Model konceptualny, jak się później przekonałam, był jednym z najważniejszych etapów projektu, ale dość sporo czasu zajęło mi rozdzielenie atrybutów do odpowiednich tabel, tak aby przyszła hurtownia miała sens. Jednakże, nie mogę porównywać poziomu trudności modelu konceptualnego z poziomem trudności procesu ETL. Sama koncepcja zajęła mi 2 dni, a implementacja kolejne 2. Pomimo tego, że mieliśmy zajęcia z procesu ETL, dość sporą abstrakcję stanowiło dla mnie połączenie kolejnych procesów w całość. Wiem, że w semestrze nie mamy za dużo czasu, ale myślę, że jeszcze jedna godzina zajęć z ETL mogłaby przyspieszyć moją pracę.

8.2. Pozyskana wiedza i doświadczenie

Muszę przyznać, że realizując projekt naprawdę dużo się nauczyłam! Pomimo tego, że ETL zajął mi 4 dni, to dzięki temu, że każdy problem rozwiązywałam sama, lepiej rozumiem ten proces. Poza tym …

Dodatkowo, bardzo podobała mi się praca z Tableau, dzięki któremu, pomimo braku pakietu Office na komputerze, byłam w stanie stworzyć czytelne i estetyczne wykresy, więc cieszę się, że to narzędzie zostało wprowadzone na zajęciach.

9. Źródła informacji użyte w etapie analizy danych

Podczas analizy danych, aby lepiej zrozumieć analizowane wykresy, a także żeby wyciągnąć odpowiednie wnioski korzystałam z następujących źródeł:

1. Źródło informacji demograficznych: United States Census Bureau
2. Źródło informacji dotyczących klimatu w Vermont: www.usclimatedata.com
3. Źródło informacji dotyczących przychodów w USA: https://en.wikipedia.org/wiki/List_of_ethnic_groups_in_the_United_States_by_household_income
4. Źródło informaci dotyczących prawnych podstaw do aresztowania kierowcy : https://www.defensivedriving.com/blog/traffic-violations-arrest/

Źródła te są oznaczone w tekście jako **([k])**, gdzie k należy do zbioru {1, 2, 3, 4}.

Uwaga:

- Niekompletny projekt nie będzie sprawdzany i tym samym ocena będzie negatywna!
- Kompletna dokumentacja musi być przesłana do sprawdzenia w formie pliku pdf nie później niż trzy dni przed terminem odbioru i prezentacji opracowanej hurtowni danych!