Hurtownie danych – Projekt HD

PWr. Wydział Informatyki i Telekomunikacji Data: **15.06.2023**

Student	255356@student.pwr.edu.pl	Ocena
Indeks	255356	
Imię	Jakub	
Nazwisko	Krupiński	

Części dokumentacji, które zostały skorygowane:

- Punkt 2.2 – definicje problemów
- Punkt 4.2, tabela 5
- Punkt 4.3, rysunek 1
- Punkt 5.1, skrypt SQL oraz rysunek 2
- Punkt 5.2 – Data Flow
- Punkt 7.2 - wnioski

1. Tytuł projektu

Rowerowe wypadki w Wielkiej Brytanii z udziałem aut

2. Charakterystyka dziedziny problemowej

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

Analizowane są tutaj rowerowe wypadki drogowe z udziałem aut w celu zrozumienia ich przyczyn oraz skutków. Dane potrzebne do analizy pochodzą z dwóch tabel: „Accidents” oraz „Bikers”. Tabela „Accidents” zawiera informacje na temat wypadków (np. liczba pojazdów biorących udział w zdarzeniu, liczba ofiar, data, czas, warunki itp.), natomiast tabela „Bikers” zawiera informacje dotyczące osób poszkodowanych (płeć, przedział wiekowy oraz informacja o tym jak ciężkie obrażenia zostały odniesione). Analiza danych pozwoli zidentyfikować czynniki ryzyka, które wpływają na występowanie takich wypadków i umożliwi opracowanie odpowiedniej strategii zapobiegania im. Hurtownia danych ułatwi przeprowadzenie tej analizy oraz interpretację jej wyników, dzięki czemu może być pomocna np. dla osób zajmujących się organizacją ruchu.

2.2 Problemy

- Brak informacji dotyczących dokładnej lokalizacji miejsca, w którym doszło do wypadku
- Brak informacji dotyczących infrastruktury drogowej
- Stronniczość danych wynikająca z tego, że znacznie większa ilość rowerzystów jest częścią ruchu drogowego w dni o sprzyjających warunkach pogodowych

2.3 Cel przedsięwzięcia

Analiza danych w celu identyfikacji czynników ryzyka, które wpływaja na występowanie wypadków rowerowych z udziałem aut i poprawienia bezpieczeństwa na drogach.

2.3.1 Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

- Którzy użytkownicy dróg są narażeni na największe ryzyko wypadków i co można zrobić, aby poprawić ich bezpieczeństwo?
- Jakie czynniki zewnętrzne mają największy wpływ na ilość występujących wypadków drogowych?
- Jakie są tendencje dotyczące ilości oraz powagi skutków wypadków drogowych tego typu na przestrzeni ostatnich lat?

2.3.2 Zakres analizy

Badane aspekty to wypadki oraz rowerzyści w kontekstach ilości pojazdów, warunków pogodowych, czasu i miejsca wypadku a także płci, wieku i tego jak poważne obrażenia zostały odniesione przez daną osobę.

Pytania badawcze

1. Którzy użytkownicy dróg są narażeni na największe ryzyko wypadków i co można zrobić, aby poprawić ich bezpieczeństwo?
2. Jakie czynniki zewnętrzne mają największy wpływ na ilość występujących wypadków drogowych?
3. Jakie są tendencje dotyczące ilości oraz powagi skutków wypadków drogowych tego typu na przestrzeni ostatnich lat?

2.3.3 Potencjalni użytkownicy

- Organy decyzyjne zajmujące się organizacją ruchu drogowego
- Służby ratunkowe
- Policja
- Użytkownicy dróg – rowerzyści, kierowcy aut i inni

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.	Accidents	csv	827 861	68.9	Baza zawierająca dane dotyczące wypadków
2.	Bikers	csv	827 871	28	Baza zawierająca dane dotyczące ludzi, którzy brali udział w wypadku

3.2. Lokalizacja, dostępność danych źródłowych

Dane pochodzą ze strony: https://www.kaggle.com/datasets/johnharshith/bicycle-accidents-in-great-britain-1979-to-2018?select=Accidents.csv – źródło publiczne.

3.3. Słownik danych – interpretacja

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

Tabela 2. Słownik atrybutów

Plik: Accidents.csv				
Lp.	**Atrybut**	**Typ danych**	**Znaczenie**	**Uwagi**
1.	Accident_Index	Tekstowy	Unikalny indeks nadawany każdemu wypadkowi	-
2.	Date	Date	Data wystąpienia wypadku	Format: DD-MM-YY
3.	Day	Tekstowy	Dzień tygodnia w którym doszło do wypadku	-
4.	Light_conditions	Tekstowy	Warunki oświetleniowe, które obowiązywały w momencie wystąpienia wypadku	-
5.	Number_of_Casualties	Int	Ilość osób poszkodowanych w wypadku	-
6.	Number_of_Vehicles	Int	Ilość pojazdów, które brały udział w wypadku	-
7.	Road_conditions	Tekstowy	Warunki na drodze w czasie wystąpienia wypadku	-
8.	Road_type	Tekstowy	Typ drogi w miejscu w którym doszło do wypadku	-

9.	Speed_limit	Int	Ograniczenie prędkości, które obowiązywało w miejscu wystapienia wypadku	-
10.	Time	Time	Czas wystąpienia wypadku	Format: HH:MM:SS
11.	Weather_conditions	Tekstowy	Warunki pogodowe w czasie wystąpienia wypadku	-

Tabela 3. Słownik atrybutów

Plik: Bikers.csv				
Lp.	**Atrybut**	**Typ danych**	**Znaczenie**	**Uwagi**
1.	Accident_Index	Tekstowy	Unikalny indeks nadawany każdemu człowiekowi, który brał udział w wypadku	-
2.	Age_Grp	Tekstowy	Grupa wiekowa do której należy człowiek, który brał udział w wypadku	-
3.	Gender	Tekstowy	Płeć osoby, która brała udział w wypadku	-
4.	Severity	Tekstowy	Wyznacznik tego jak poważne obrażenia odniosła osoba biorąca udział w wypadku	-

3.4. Ocena jakościowa danych

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

Tabela 4. Ocena jakościowa danych

Plik: Accidents.csv				
Lp.	**Atrybut**	**Typ danych**	**Zakres wartości**	**Uwagi - ocena jakości danych**
1.	Accident_Index	Tekstowy	-	Poprawne dane

2.	Date	Date	DD/MM/YY	Poprawne dane
3.	Day	Tekstowy	Monday, Tuesday, …	Poprawne dane
4.	Light_conditions	Tekstowy	Darkness lights lit / Darkness no lights / Daylight	Poprawne dane
5.	Number_of_Casualties	Int	0 – oo	Poprawne dane
6.	Number_of_Vehicles	Int	0 – oo	Poprawne dane
7.	Road_conditions	Tekstowy	Dry, Flood, Wet itp.	Poprawne dane
8.	Road_type	Tekstowy	Unknown, Roundabout itp.	Błąd w oznaczeniu kolumny – “One way ‘sreet’” zamiast ‘street’
9.	Speed_limit	Int	0 – oo	Niektóre wartości pozostają niejasne co do tego, jak powinny być interpretowane, bądź też są po prostu błędne (np. wartość ‘660’). Aby przeprowadzić analizę na tych danych utworzone zostanie więc nazwane obliczenie w SSIS grupujące wartości ograniczeń prędkości w przedziały. Dane wymagały również zmiany typu z float na int, jako że wszystkie wartości są liczbami całkowitymi.
10.	Time	Time	HH:MM:SS	Poprawne dane
11.	Weather_conditions	Tekstowy	Missing data, Clear, Rain itp.	Poprawne dane

Tabela 5. Ocena jakościowa danych

Plik: Bikers.csv				
Lp.	**Atrybut**	**Typ danych**	**Zakres wartości**	**Uwagi - ocena jakości danych**
1.	Accident_Index	Tekstowy	-	W tabeli ‘Bikers’ znajduje się 10 rekordów, które nie mają swoich odpowiedników w tabeli ‘Accidents’ – rekordy te są usuwane podczas tworzenia tabel

2.	Age_Grp	Tekstowy	11 to 15, 16 to 20, 21 to 25, itp.	Poprawne dane
3.	Gender	Tekstowy	Male / Female / Other	Poprawne dane
4.	Severity	Tekstowy	Fatal / Serious / Slight	Poprawne dane

4. Analityczne modele wielowymiarowe

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

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

Tabela 6. Fakty podlegające analizie

Lp.	Fakty	Miary	Uwagi
1.	**Accident**	Number_of_Vehicles, Number_of_Casualties	

4.2. Kontekst analizy faktów

Ustalony kontekst analizy faktów został przedstawiony w tab. 6.

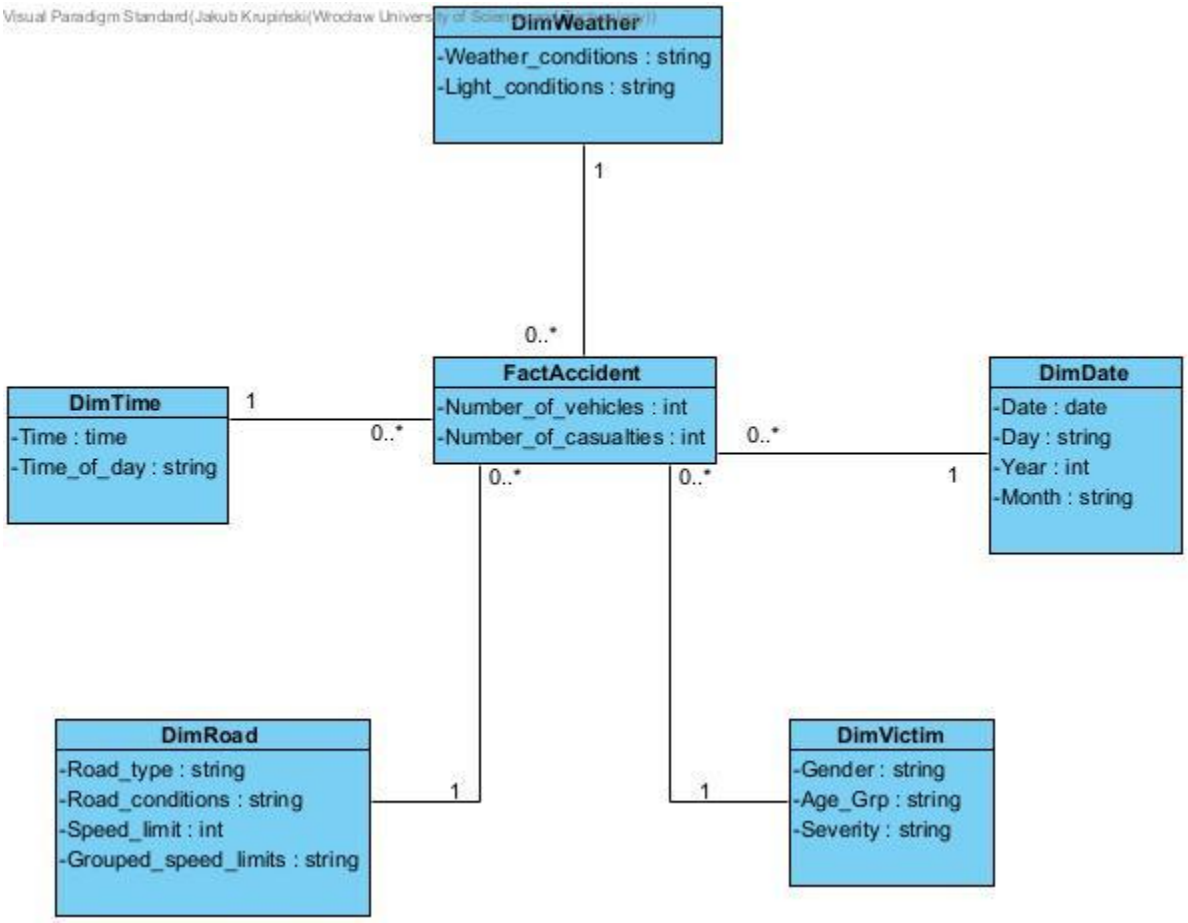
Tabela 3. Wymiary analizy faktów

Lp.	Wymiar	Własności
1.	Weather	- Light_conditions: 3 unikalne wartości - Weather_conditions: 10 unikalnych wartości
2.	Road	- Road_conditions: 6 unikalnych wartosci - Road_type: 6 unikalnych wartosci - Speed_limit: 36 unikalnych wartości -Grouped_speed_limits: 14 unikalnych wartości
3.	Time	- Time: 00:00:00 → 23:59:59 - Time_of_day: 4 unikalne wartości
4.	Date	- Date: 1979-01-01 → 2018-12-31 - Day: 7 unikalnych wartości (dni tygodnia) - Year: 40 unikalnych wartości - Month: 12 unikalnych wartości

5.	Victim	- Age_grp: 9 unikalnych wartości - Gender: 3 unikalne wartości - Severity: 3 unikalne wartości

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: FactAccident wymiarów: DimDate, DimTime, DimRoad, DimWeather oraz DimVictim. Model ten reprezentowany jest w postaci schematu pokazanego poniżej.



Rysunek 1. Wielowymiarowy model analityczny przedstawiony na poziomie konceptualnym

5. Projekt procesu ETL

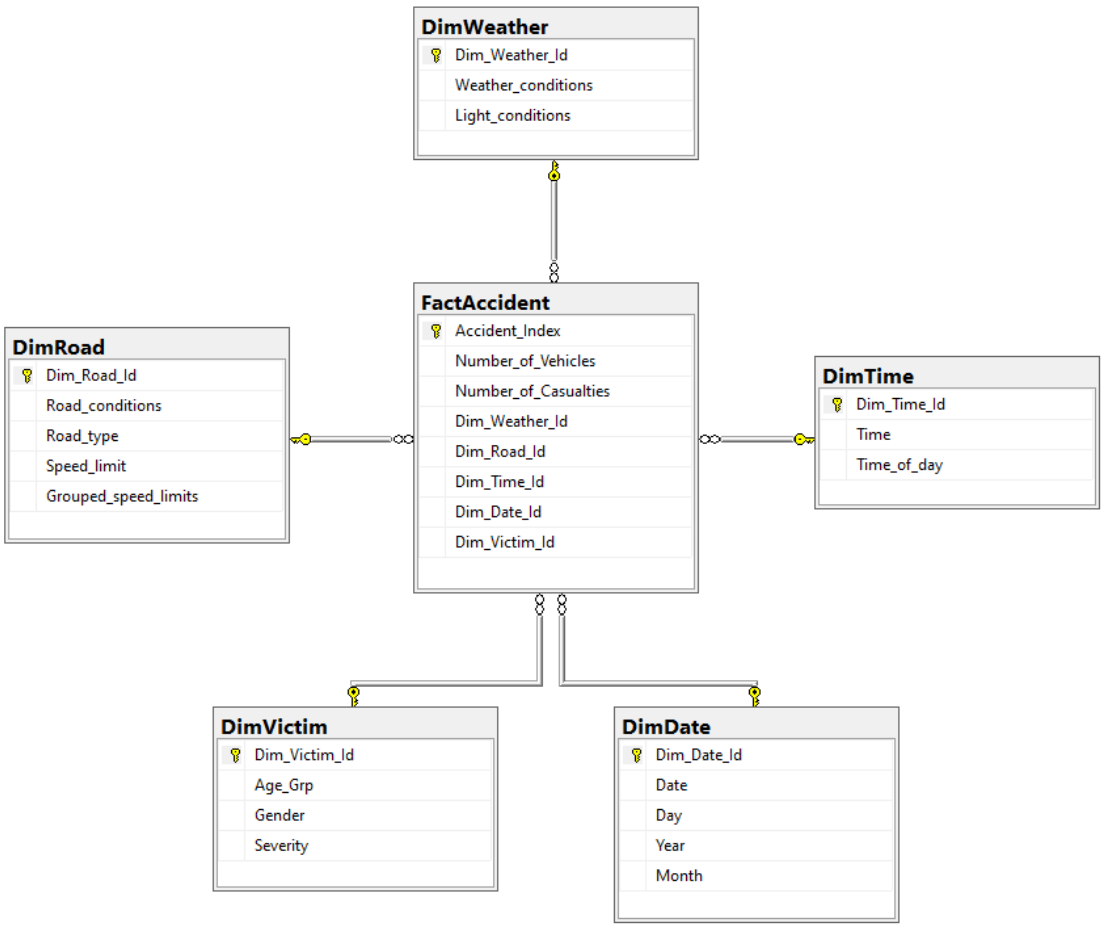
5.1. Schemat bazy danych HD (skrypt SQL)

```sql
ALTER TABLE Accidents
ADD tmp INT;

UPDATE Accidents
SET tmp = CAST(Speed_limit AS INT);

ALTER TABLE Accidents
DROP COLUMN Speed_limit;

EXEC sp_rename 'Accidents.tmp', 'Speed_limit', 'COLUMN';

DELETE FROM Bikers
WHERE Accident_Index NOT IN (
    SELECT Accident_Index FROM Accidents
);

ALTER TABLE Bikers
ADD CONSTRAINT FK_Bikers_Accidents FOREIGN KEY (Accident_Index)
REFERENCES Accidents (Accident_Index);

CREATE TABLE Accidents_Bikers_Combined (
    Accident_Index varchar(50),
    Number_of_Vehicles tinyint,
    Number_of_Casualties tinyint,
    Date DATE,
    Time TIME,
    Speed_limit int,
    Road_conditions varchar(50),
    Weather_conditions varchar(50),
    Day varchar(50),
    Road_type varchar(50),
    Light_conditions varchar(50),
    Biker_Id INT,
    Gender varchar(50),
    Severity varchar(50),
    Age_Grp varchar(50),
    PRIMARY KEY (Accident_Index, Biker_Id)
);

INSERT INTO
    Accidents_Bikers_Combined (
        Accident_Index,
        Biker_Id,
        Number_of_Vehicles,
        Number_of_Casualties,
        Date,
        Time,
        Speed_limit,
        Road_conditions,
        Weather_conditions,
        Day,
        Road_type,
        Light_conditions,
        Gender,
```

```
        Severity,
        Age_Grp
    )
SELECT
    A.Accident_Index,
    B.Biker_Id,
    A.Number_of_Vehicles,
    A.Number_of_Casualties,
    A.Date,
    A.Time,
    A.Speed_limit,
    A.Road_conditions,
    A.Weather_conditions,
    A.Day,
    A.Road_type,
    A.Light_conditions,
    B.Gender,
    B.Severity,
    B.Age_Grp
FROM
    Accidents A
    INNER JOIN Bikers B ON A.Accident_Index = B.Accident_Index;

CREATE TABLE DimWeather (
    Dim_Weather_Id INT IDENTITY(1,1) PRIMARY KEY,
    Weather_conditions varchar(50),
    Light_conditions varchar(50),
);

CREATE TABLE DimRoad (
    Dim_Road_Id INT IDENTITY(1,1) PRIMARY KEY,
    Road_conditions varchar(50),
    Road_type varchar(50),
    Speed_limit float
);

CREATE TABLE DimTime (
    Dim_Time_Id INT IDENTITY(1,1) PRIMARY KEY,
    Date DATE,
    Time TIME(7),
    Day varchar(50)
);

CREATE TABLE DimDate (
    Dim_Date_Id INT IDENTITY(1,1) PRIMARY KEY,
    Date DATE,
    Day varchar(50)
);

CREATE TABLE DimVictim (
    Dim_Victim_Id INT IDENTITY(1,1) PRIMARY KEY,
    Age_Grp varchar(50),
    Gender varchar(50),
    Severity varchar(50)
```

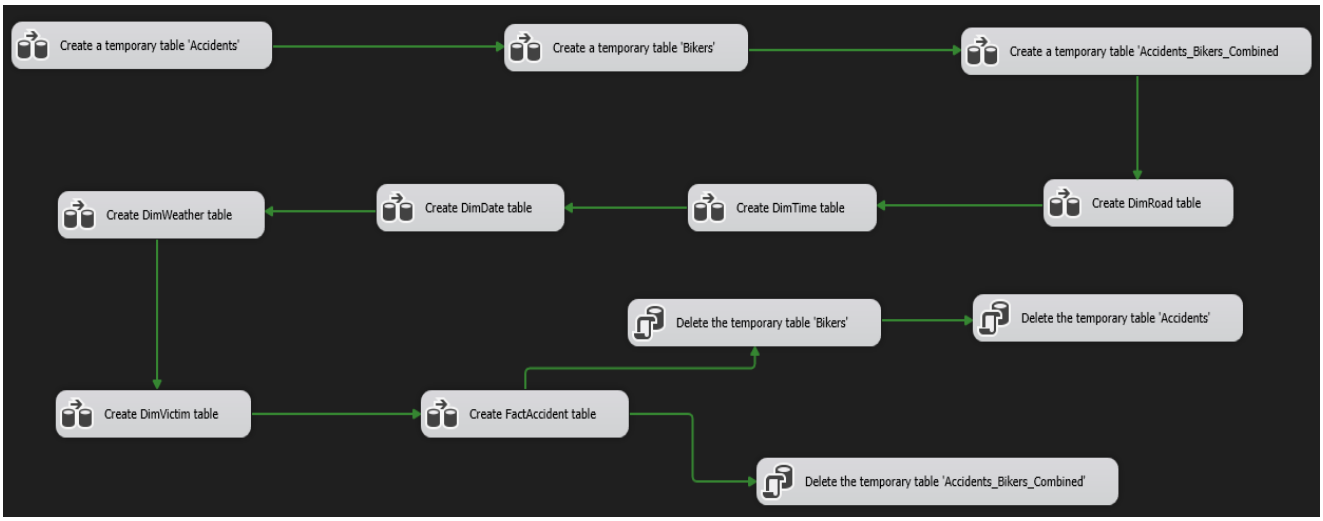
```
);

CREATE TABLE FactAccident (
    Accident_Index varchar(50) PRIMARY KEY,
    Number_of_Vehicles tinyint,
    Number_of_Casualties tinyint,
    Dim_Weather_Id INT,
    Dim_Road_Id INT,
    Dim_Time_Id INT,
    Dim_Date_Id INT,
    Dim_Victim_Id INT,
    FOREIGN KEY (Dim_Weather_Id) REFERENCES DimWeather(Dim_Weather_Id),
    FOREIGN KEY (Dim_Road_Id) REFERENCES DimRoad(Dim_Road_Id),
    FOREIGN KEY (Dim_Time_Id) REFERENCES DimTime(Dim_Time_Id),
    FOREIGN KEY (Dim_Date_Id) REFERENCES DimDate(Dim_Date_Id),
    FOREIGN KEY (Dim_Victim_Id) REFERENCES DimVictim(Dim_Victim_Id),
);

INSERT INTO DimWeather (Weather_conditions, Light_conditions)
SELECT DISTINCT Weather_conditions, Light_conditions
FROM Accidents_Bikers_Combined;

INSERT INTO DimRoad (Road_conditions, Road_type, Speed_limit)
SELECT DISTINCT Road_conditions, Road_type, Speed_limit
FROM Accidents_Bikers_Combined;

INSERT INTO DimTime (Time)
SELECT DISTINCT Time
FROM Accidents_Bikers_Combined;

INSERT INTO DimVictim (Age_Grp, Gender, Severity)
SELECT DISTINCT Age_Grp, Gender, Severity
FROM Accidents_Bikers_Combined;

INSERT INTO DimDate(Date, Day)
SELECT DISTINCT Date, Day
FROM Accidents_Bikers_Combined;

ALTER TABLE DimDate
ADD Year INT NULL;

ALTER TABLE DimDate
ADD Month varchar(50) NULL;

UPDATE DimDate
SET Year = YEAR(Date),
    Month = MONTH(Date);

ALTER TABLE DimTime
ADD Time_of_day varchar(50) NULL;

UPDATE DimTime
SET Time_of_day = CASE
```

```
    WHEN TRY_CAST(Time AS time) >= TRY_CAST('00:00:00' AS time) AND TRY_CAST(Time AS time) <
TRY_CAST('06:00:00' AS time) THEN '00:00-05:59'
    WHEN TRY_CAST(Time AS time) >= TRY_CAST('06:00:00' AS time) AND TRY_CAST(Time AS time) <
TRY_CAST('12:00:00' AS time) THEN '06:00-11:59'
    WHEN TRY_CAST(Time AS time) >= TRY_CAST('12:00:00' AS time) AND TRY_CAST(Time AS time) <
TRY_CAST('18:00:00' AS time) THEN '12:00-17:59'
    WHEN TRY_CAST(Time AS time) >= TRY_CAST('18:00:00' AS time) AND TRY_CAST(Time AS time) <=
TRY_CAST('23:59:59' AS time) THEN '18:00-23:59'
    ELSE 'UNKNOWN'
END;

ALTER TABLE DimRoad
ADD Grouped_speed_limits varchar(50) NULL;

UPDATE DimRoad
SET Grouped_speed_limits =
    CASE
        WHEN Speed_limit <= 10 THEN '0-10'
        WHEN Speed_limit >= 11 AND Speed_limit <= 19 THEN '11-19'
        WHEN Speed_limit = 20 THEN '20'
        WHEN Speed_limit >= 21 AND Speed_limit <= 29 THEN '21-29'
        WHEN Speed_limit = 30 THEN '30'
        WHEN Speed_limit >= 31 AND Speed_limit <= 39 THEN '31-39'
        WHEN Speed_limit = 40 THEN '40'
        WHEN Speed_limit >= 41 AND Speed_limit <= 49 THEN '41-49'
        WHEN Speed_limit = 50 THEN '50'
        WHEN Speed_limit >= 51 AND Speed_limit <= 59 THEN '51-59'
        WHEN Speed_limit = 60 THEN '60'
        WHEN Speed_limit >= 61 AND Speed_limit <= 69 THEN '61-69'
        WHEN Speed_limit = 70 THEN '70'
        ELSE '>70'
    END;

INSERT INTO FactAccident (Accident_Index, Number_of_Vehicles, Number_of_Casualties, Dim_Road_Id,
                          Dim_Time_Id, Dim_Date_Id, Dim_Victim_Id, Dim_Weather_Id)
SELECT DISTINCT Accident_Index, Number_of_Vehicles, Number_of_Casualties, Dim_Road_Id, Dim_Time_Id,
                   Dim_Date_Id, Dim_Victim_Id, Dim_Weather_Id
FROM Accidents_Bikers_Combined
JOIN DimTime ON
Accidents_Bikers_Combined.Time = DimTime.Time
JOIN DimDate ON
Accidents_Bikers_Combined.Date = DimDate.Date AND
Accidents_Bikers_Combined.Day = DimDate.Day
JOIN DimRoad ON
Accidents_Bikers_Combined.Road_conditions = DimRoad.Road_conditions AND
Accidents_Bikers_Combined.Road_type = DimRoad.Road_type AND
Accidents_Bikers_Combined.Speed_limit = DimRoad.Speed_limit
JOIN DimWeather ON
Accidents_Bikers_Combined.Weather_conditions = DimWeather.Weather_conditions AND
Accidents_Bikers_Combined.Light_conditions = DimWeather.Light_conditions
JOIN DimVictim ON
Accidents_Bikers_Combined.Age_Grp = DimVictim.Age_Grp AND
Accidents_Bikers_Combined.Gender = DimVictim.Gender AND
Accidents_Bikers_Combined.Severity = DimVictim.Severity;
```

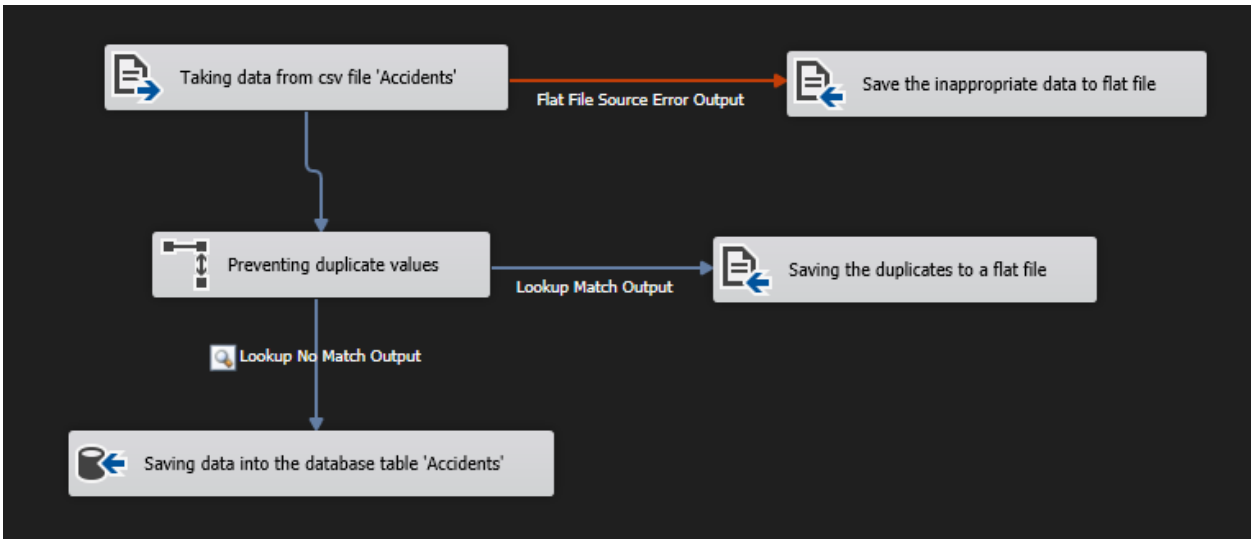


Rysunek 2. Schemat bazy danych utworzonej za pomocą skryptu SQL

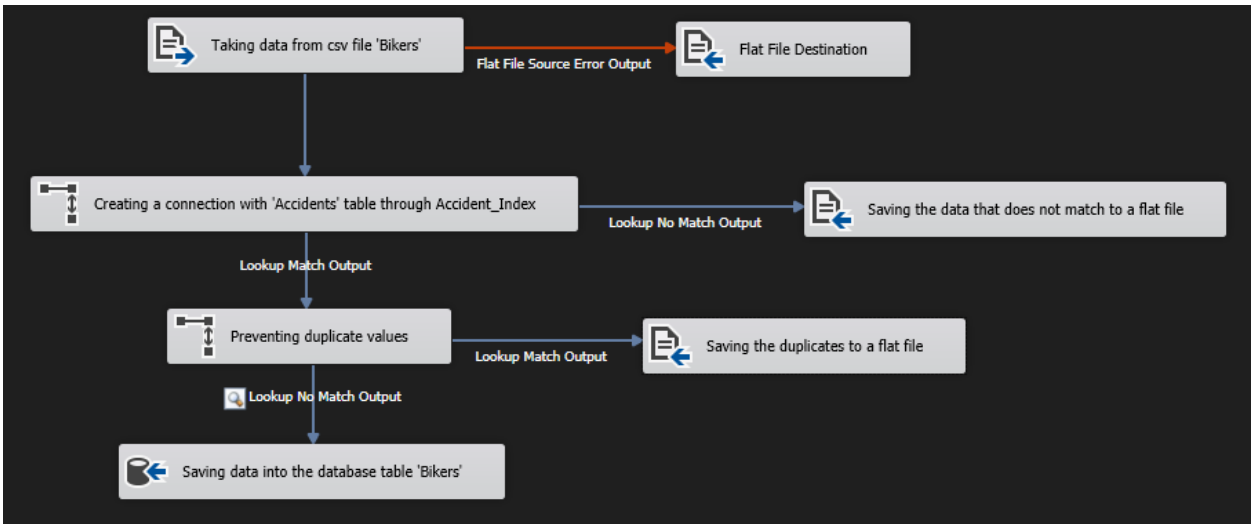
5.2. Specyfikacja procesów ETL (Control Flow + Data Flow)



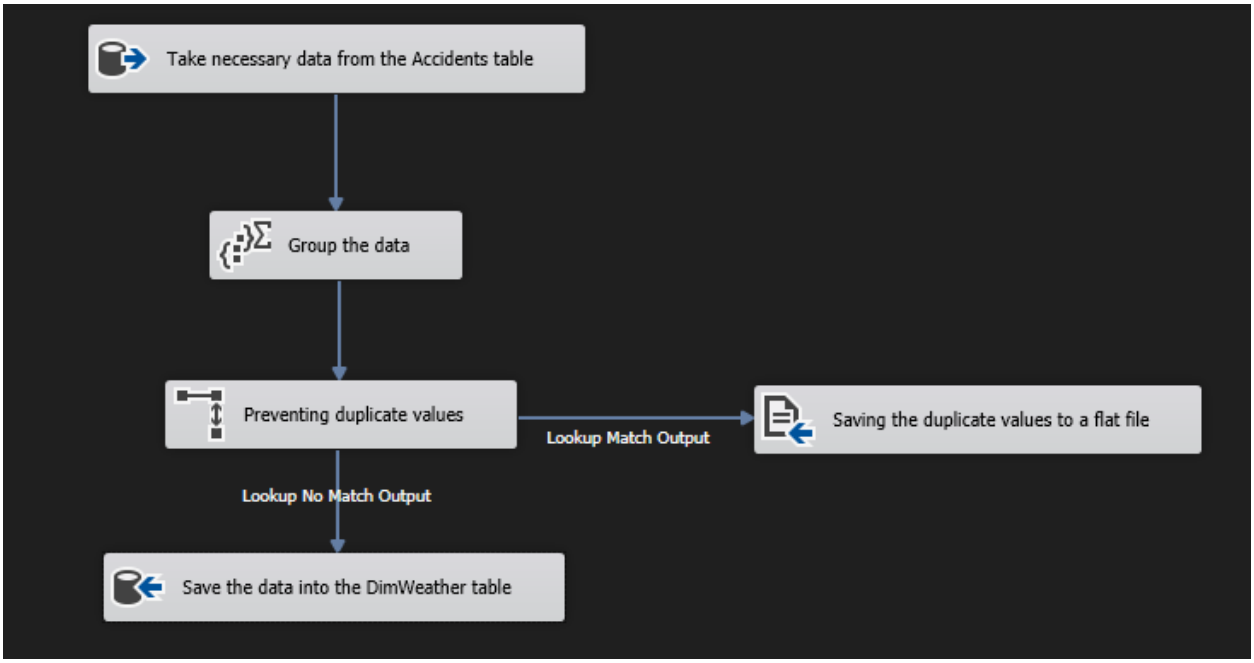
Rysunek 3. Control Flow



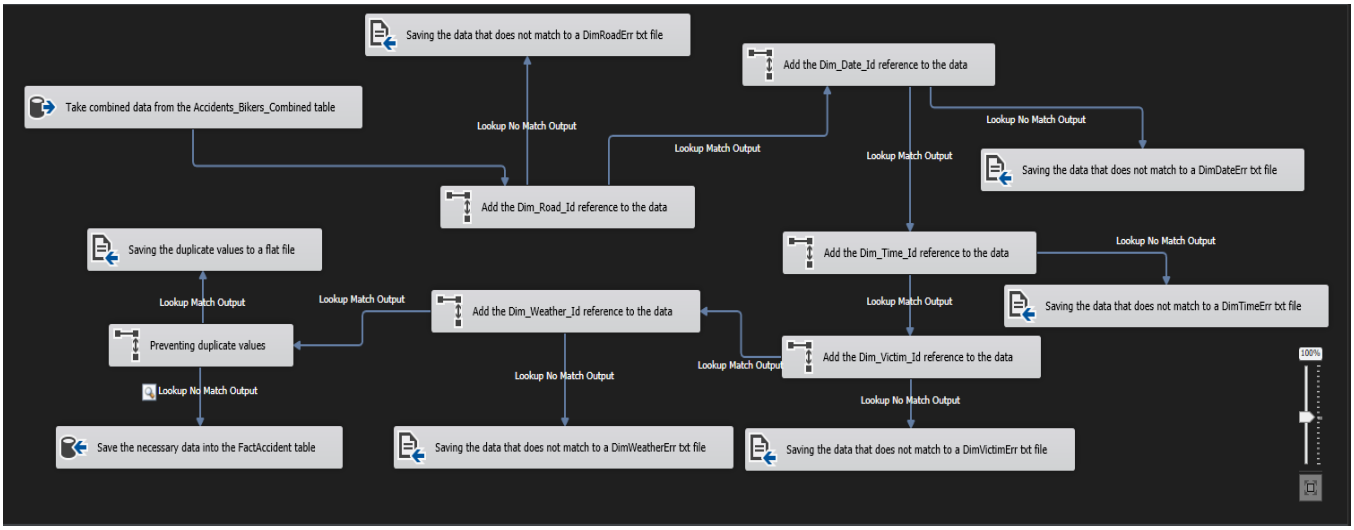
Rysunek 4. Data Flow – Create a temporary table ‘Accidents’



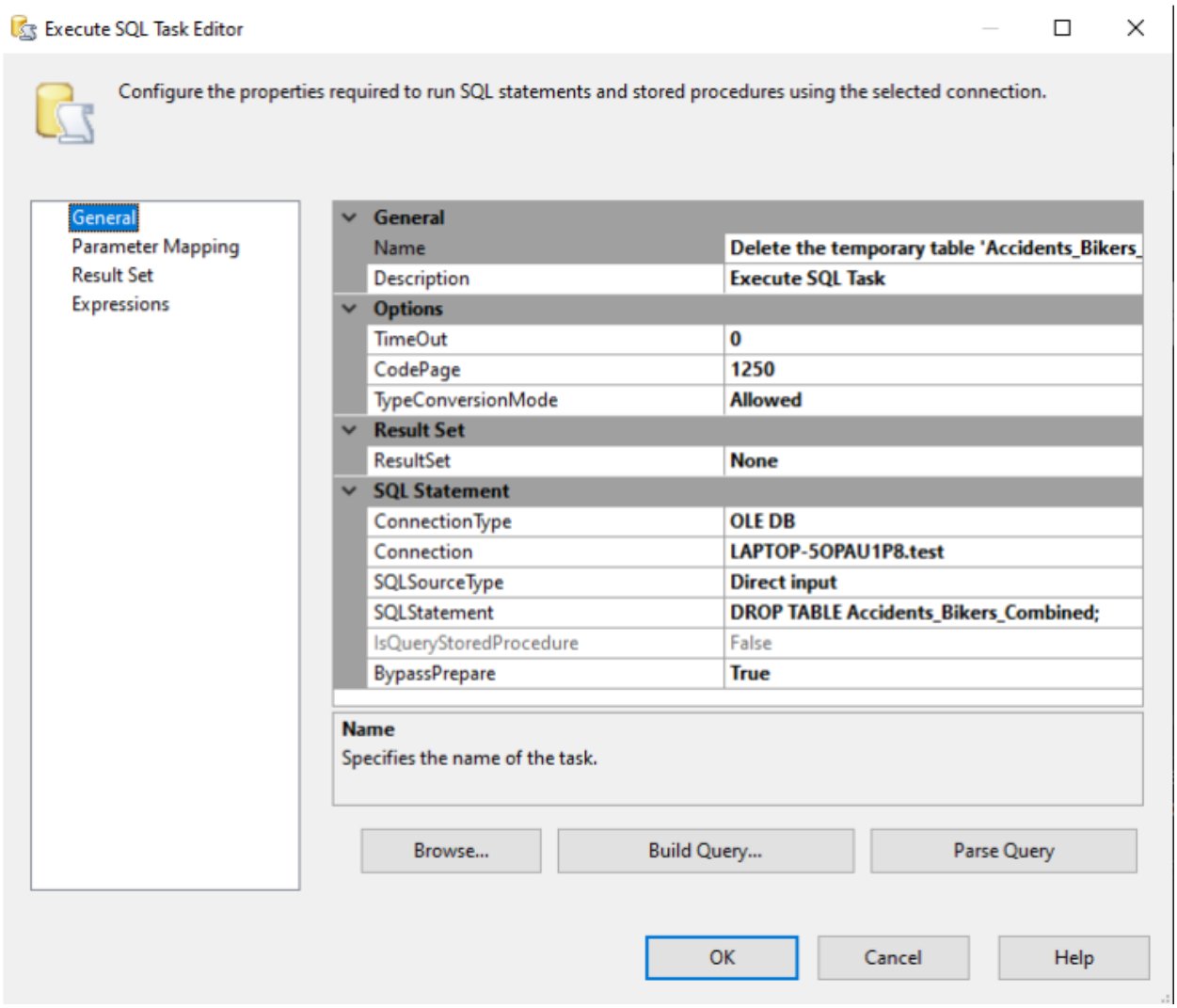
Rysunek 5. Data Flow – Create a temporary table ‘Bikers’



Rysunek 6. Data Flow – Create ‘DimWeather’ table



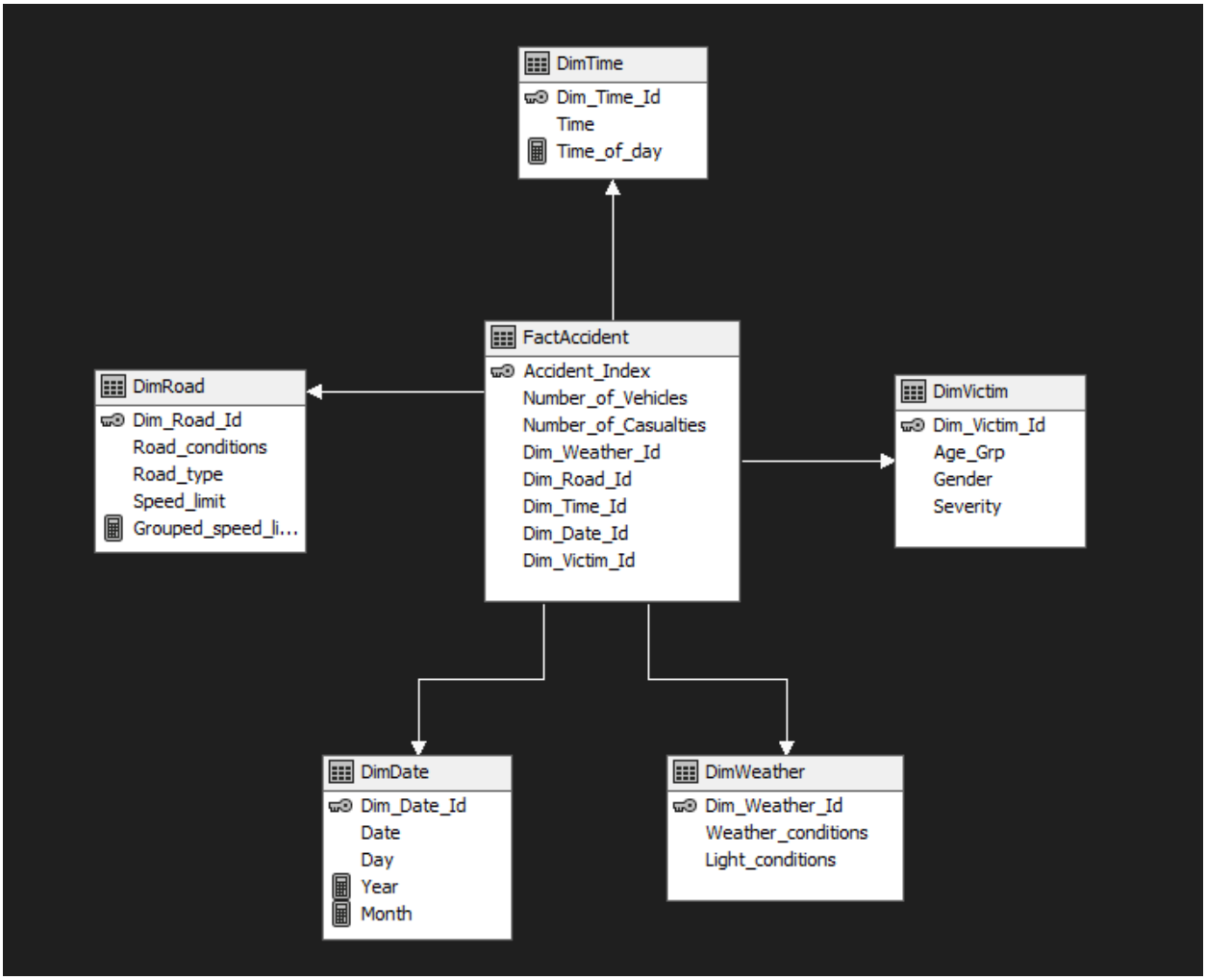
Rysunek 7. Data Flow – Create ‘FactAccident’ table



Rysunek 8. SQL Task – Delete the temporary table ‘Accidents_Bikers_Combined’

6. Implementacja modeli wielowymiarowych

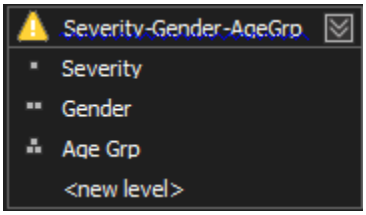
6.1. Widok danych



Rysunek 9. Widok danych

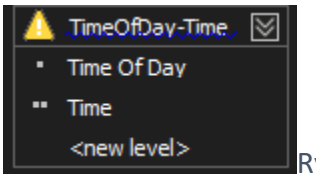
6.2. Wymiary

Rysunek 20. Zdefiniowane wymiary

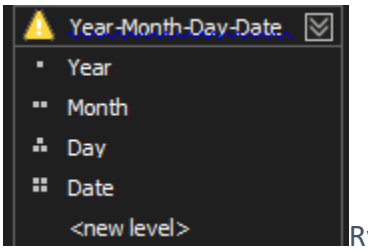


Rysunek 31. Hierarchia dla wymiaru DimVictim

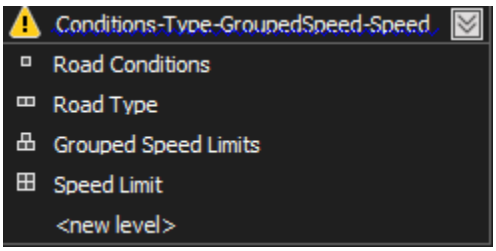
Rysunek 42. Hierarchia dla wymiaru DimWeather



Rysunek 53. Hierarchia dla wymiaru DimTime

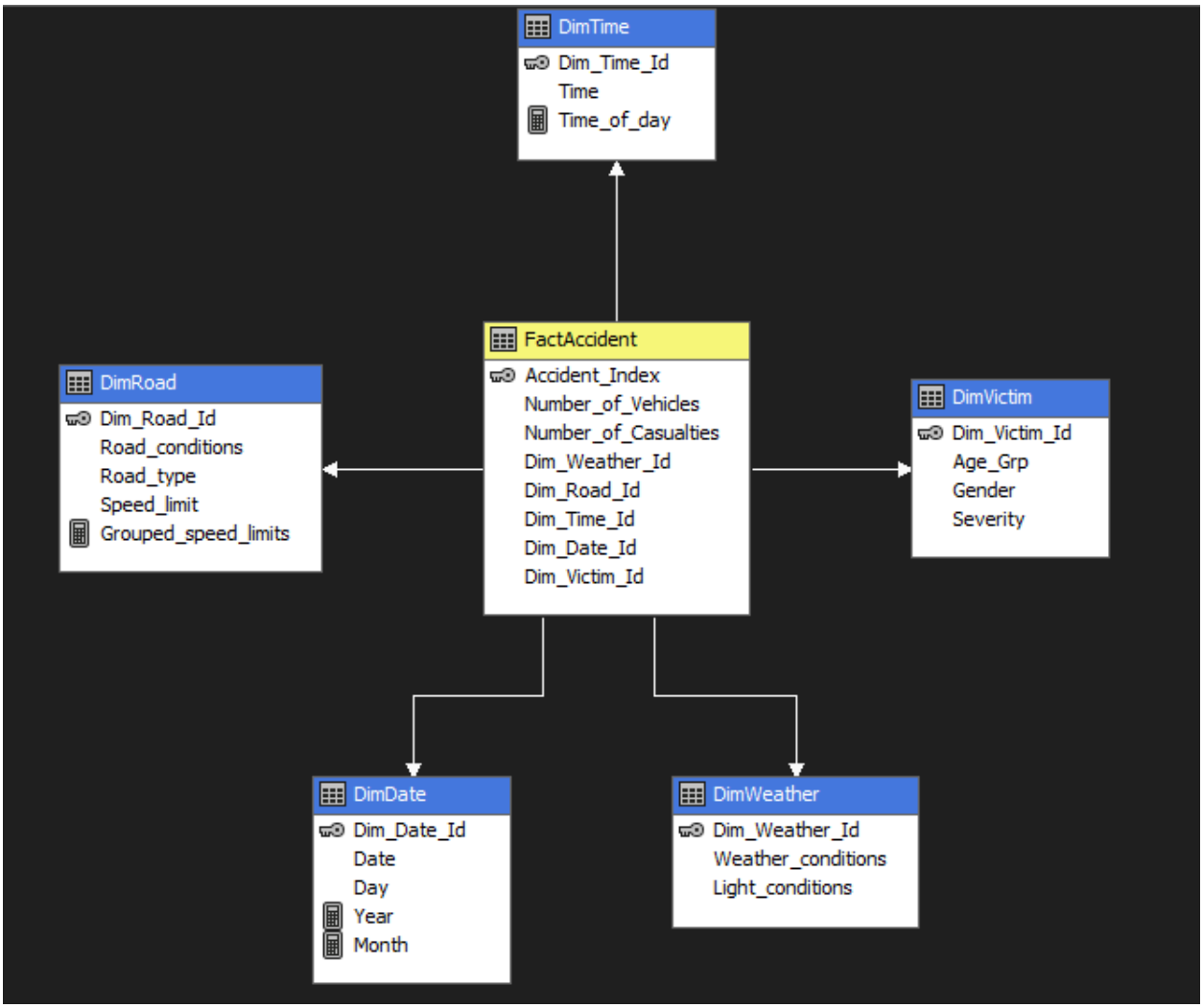


Rysunek 64. Hierarchia dla wymiaru DimDate



Rysunek 75. Hierarchia dla wymiaru DimRoad

6.3. Modele wielowymiarowe – Kostki

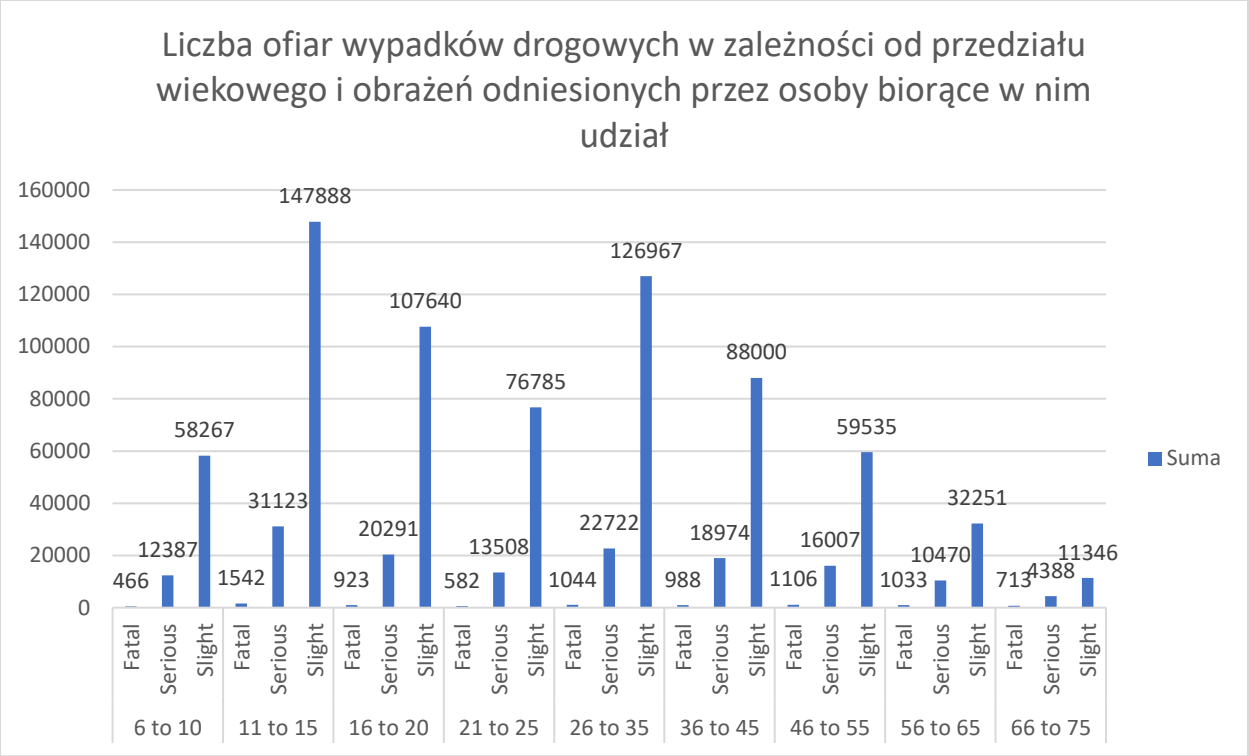


Rysunek 86. Widok wielowymiarowego modelu danych wygenerowany przez SSIS

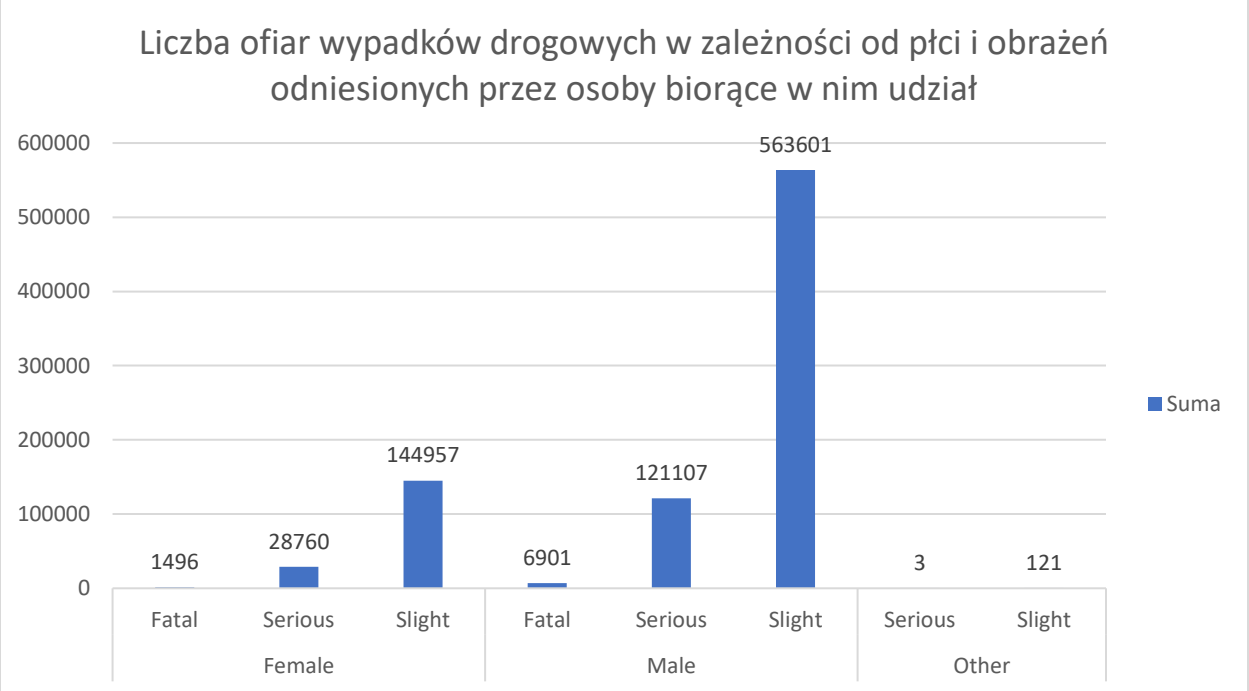
7. Analiza danych

7.1. Realizacja procesów analitycznych

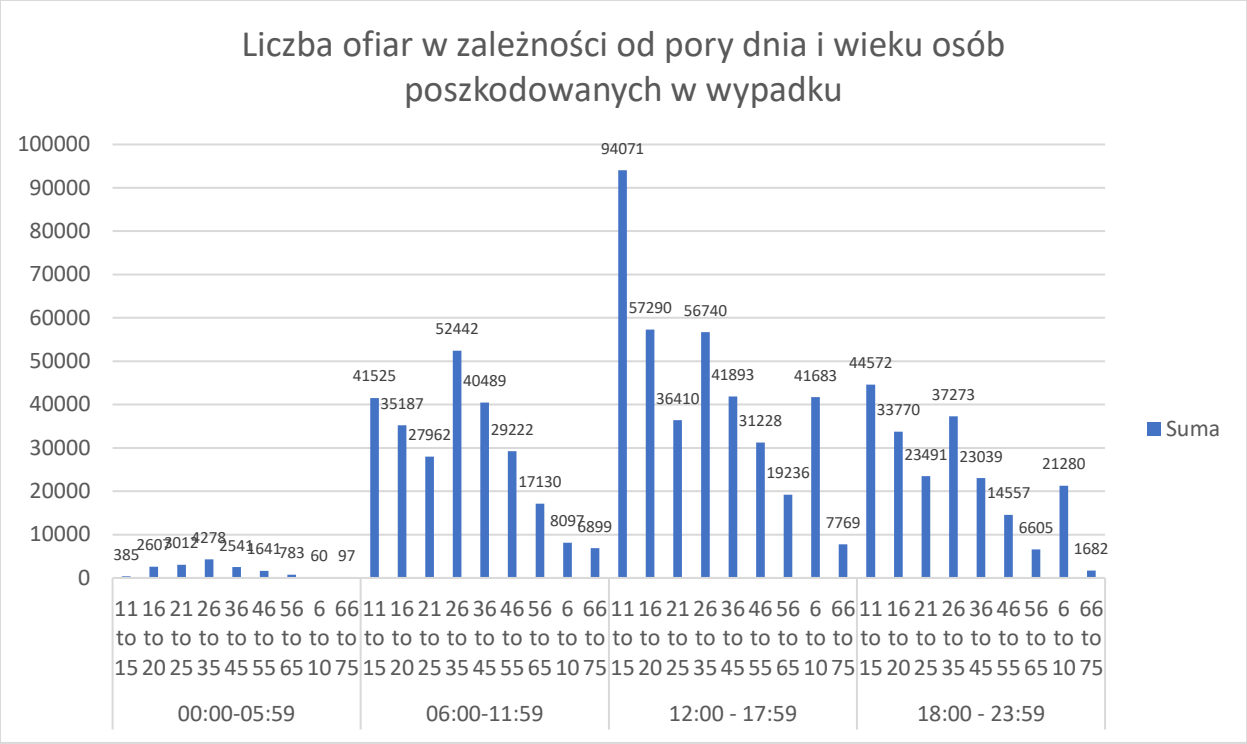
- Którzy użytkownicy dróg są narażeni na największe ryzyko wypadków?



Rysunek 97. Liczba ofiar wypadków drogowych w zależności od przedziału wiekowego i obrażeń odniesionych przez osoby biorące w nim udział



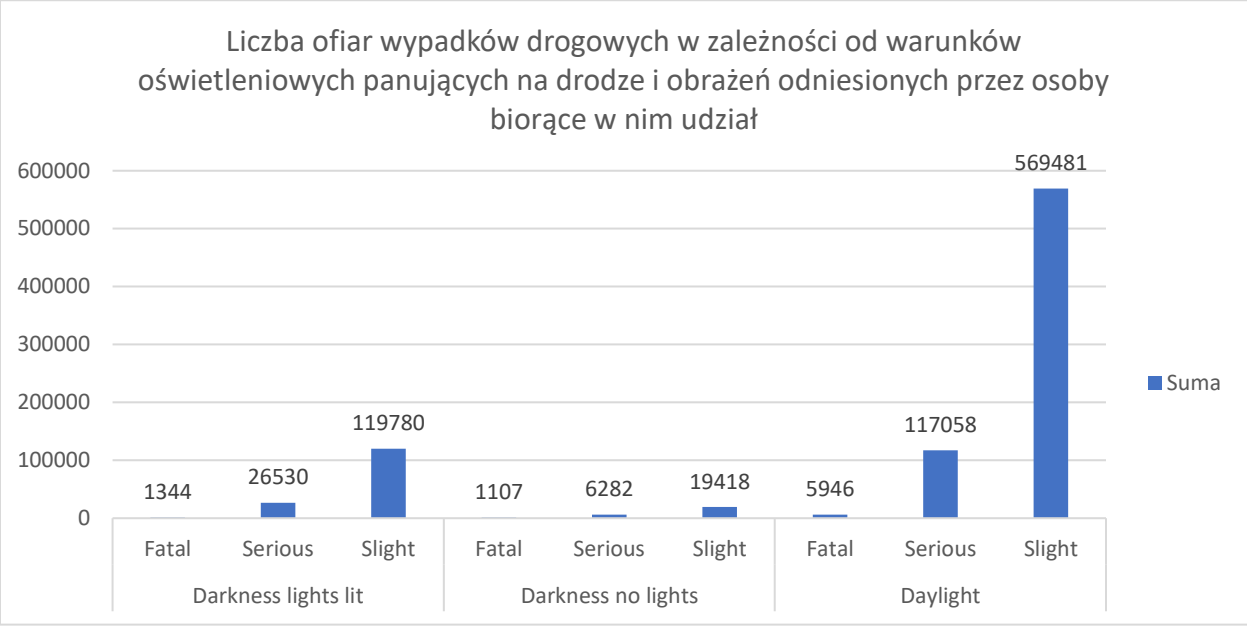
Rysunek 108. Liczba ofiar wypadków drogowych w zależności od płci i obrażeń odniesionych przez osoby biorące w nim udział



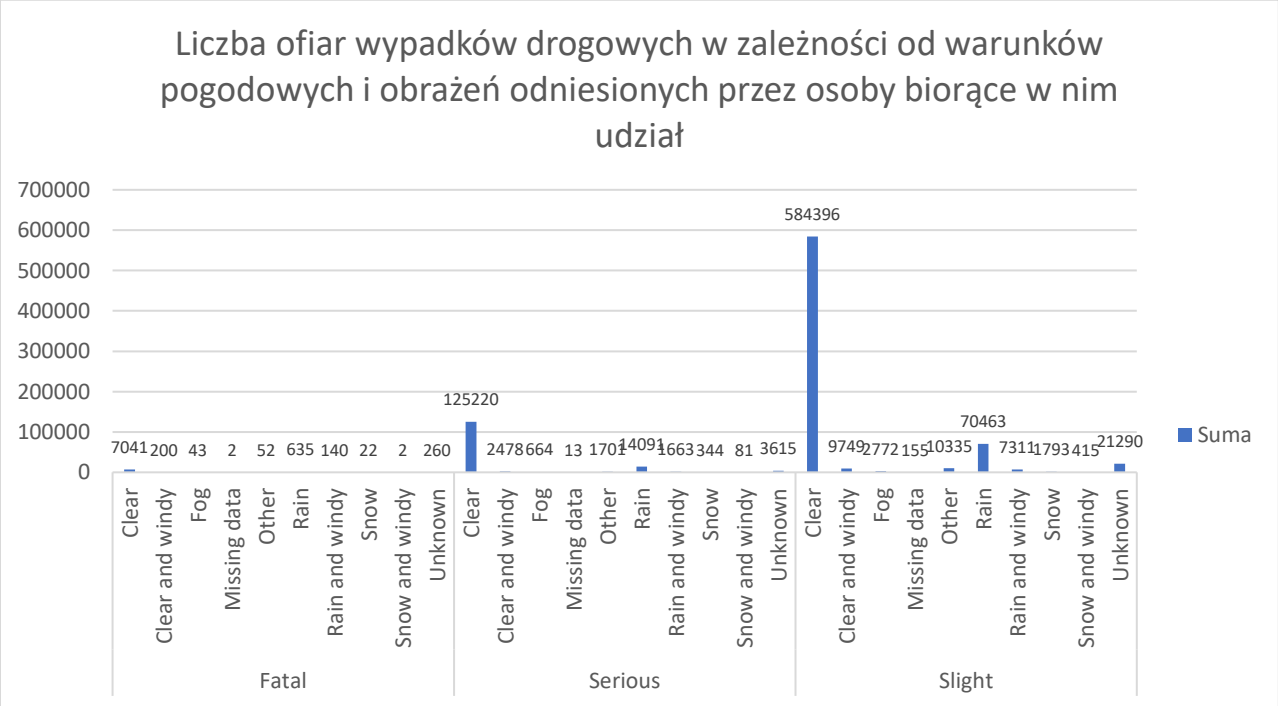
Rysunek 11*9*. Liczba ofiar w zależności od pory dnia i wieku osób poszkodowanych w wypadku

Na podstawie pierwszych trzech sporządzonych wykresów można zauważyć, że zdecydowanie przeważające są wypadki, których uczestnicy odnoszą jedynie drobne obrażenia – najbardziej zagrożeni są wg tych danych mężczyźni, jako że suma przypadków, kiedy to właśnie oni są poszkodowani w wypadku jest niemal cztery razy większa niż w przypadku kobiet. Można z tego powodu wywnioskować więc, że najczęściej rowerem poruszają się właśnie mężczyźni. Liczby ofiar wypadków w zależności od grup wiekowych są dość zróżnicowanea tendencja wydaje się być spadkowa względem wieku – osoby młodsze ulegają wypadkom znacznie częściej. Najbardziej zagrożona na drogach kategoria wg sporządzonych przeze mnie wykresów to grupa 11-15 lat – są to lata, kiedy rowerzyści są młodzi, niedoświadczeni, co bez wątpienia prowadzi do tego, że podejmują nieodpowiedzialne decyzje w ruchu drogowym i nie wykazują się odpowiednią ostrożnością. Osoby młode w wieku 16-20 , 21-25, 26-35 i 36-45 również ulegają wielu wypadkom, jednak poza wzrostem tych statystyk w latach 26-35 już tutaj widoczna jest tendencja spadkowa. Najbezpieczniejsze są osoby starsze – 56-65 i 66-75, jako że z racji wieku są to osoby, które bez wątpienia najrzadziej przemieszczają się rowerami. Rysunek pokazujący liczbę ofiar wypadków w zależności od pory dnia pozwala zauważyć m.in., że wyjątkowo duża liczba wypadków przypada na grupę wiekową 11 – 15 w godzinach popołudniowych, można więc podejrzewać, że są to osoby młode, które w tych godzinach mogą wracać z zajęć szkolnych, bądź spotykać się z przyjaciółmi.

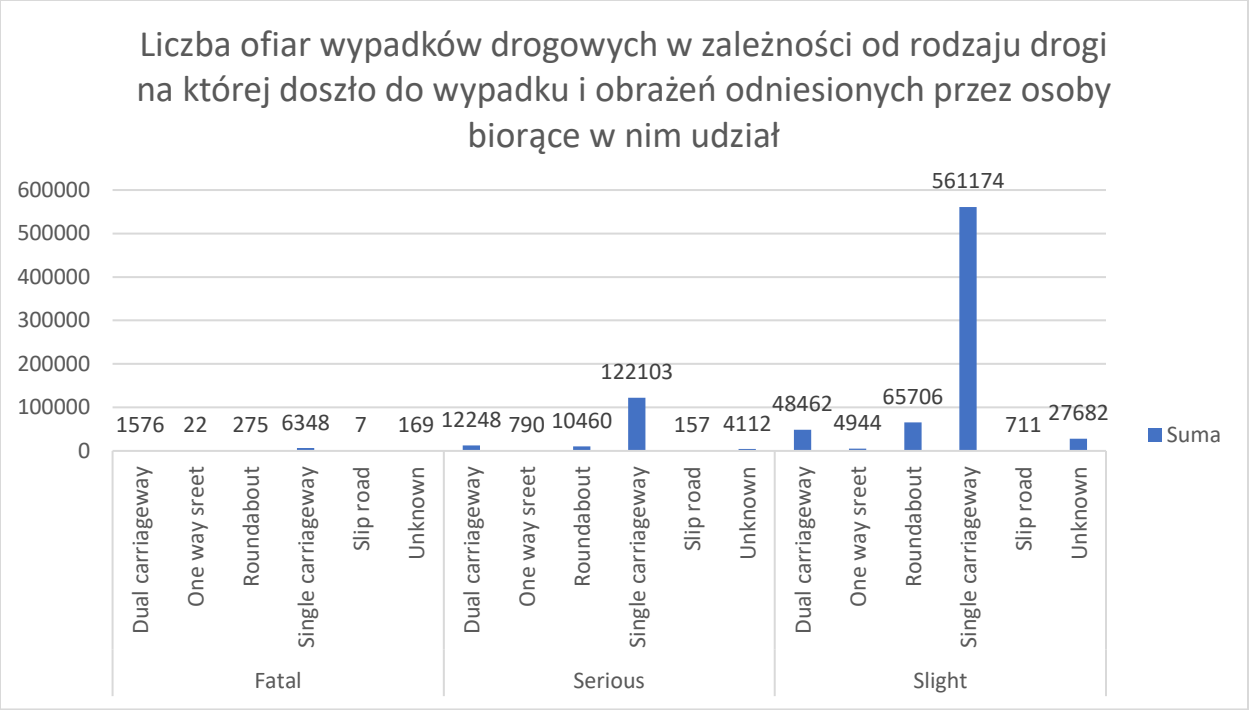
- Jakie czynniki zewnętrzne mają największy wpływ na ilość występujących wypadków drogowych?



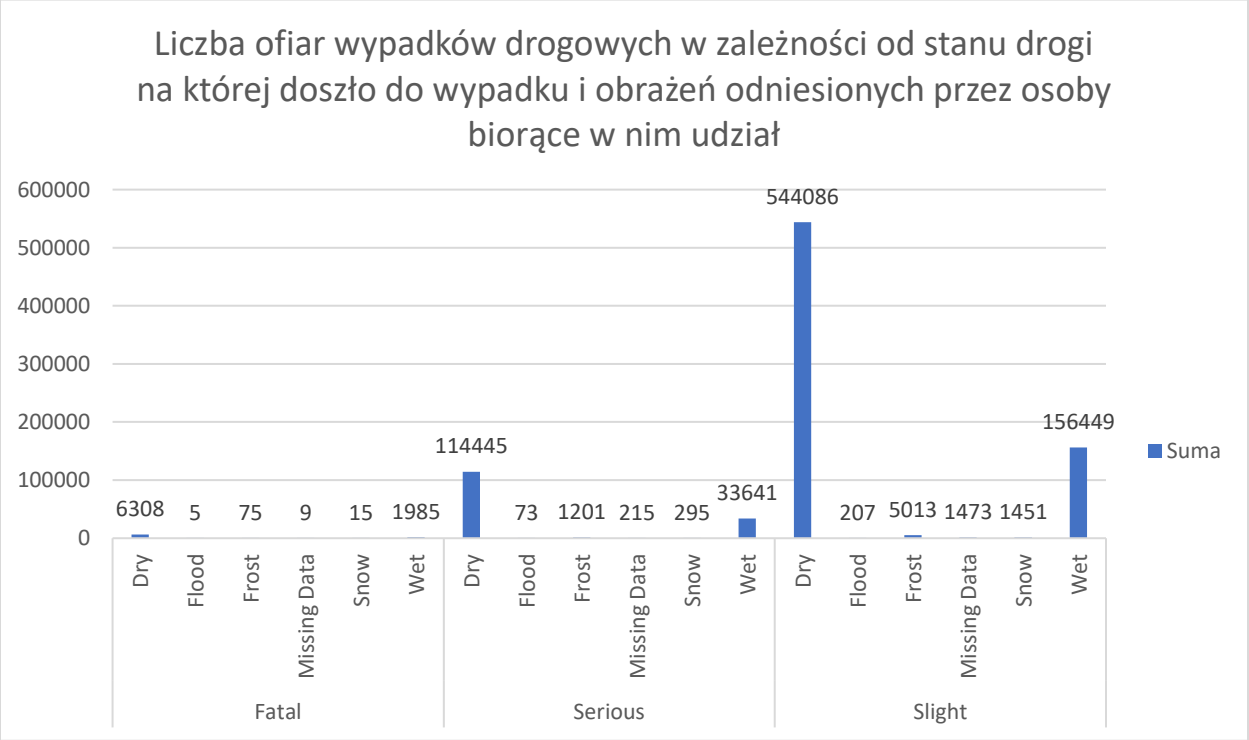
Rysunek *20*. Liczba ofiar wypadków drogowych w zależności od warunków oświetleniowych panujących na drodze i obrażeń odniesionych przez osoby biorące w nim udział



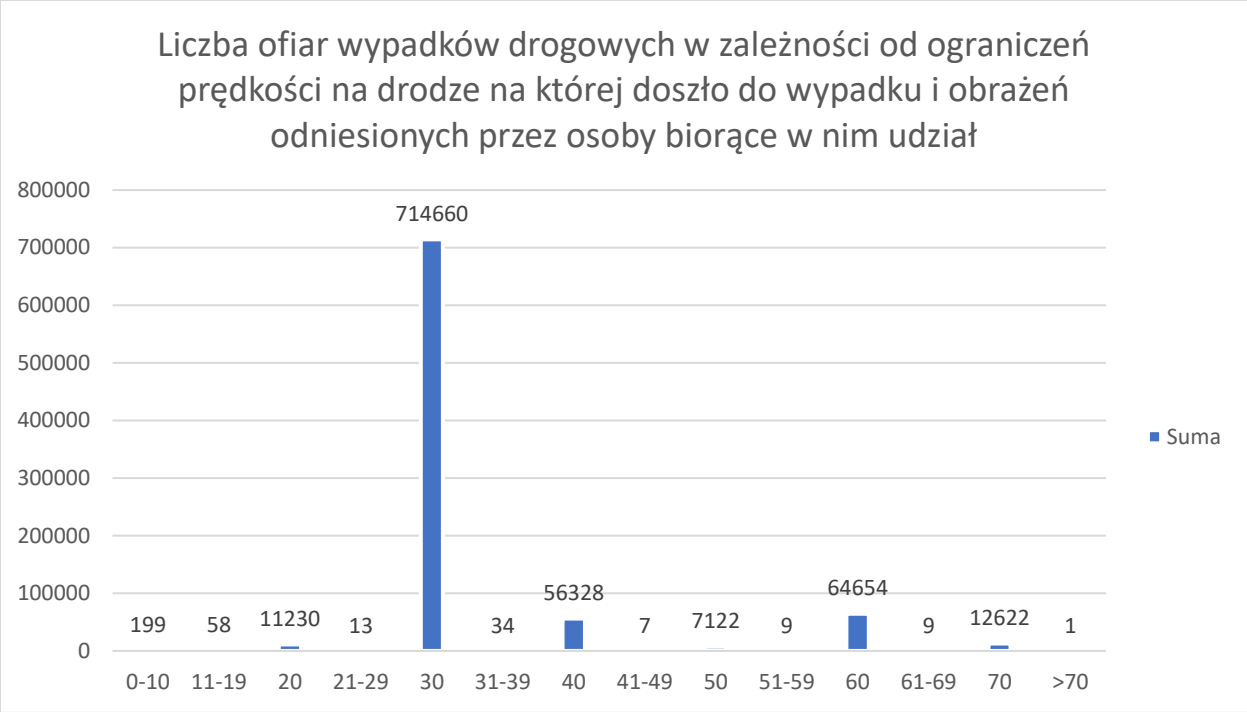
Rysunek 212. Liczba ofiar wypadków drogowych w zależności od warunków pogodowych i obrażeń odniesionych przez osoby biorące w nim udział



Rysunek *22*. Liczba ofiar wypadków drogowych w zależności od rodzaju drogi na której doszło do wypadku i obrażeń odniesionych przez osoby biorące w nim udział



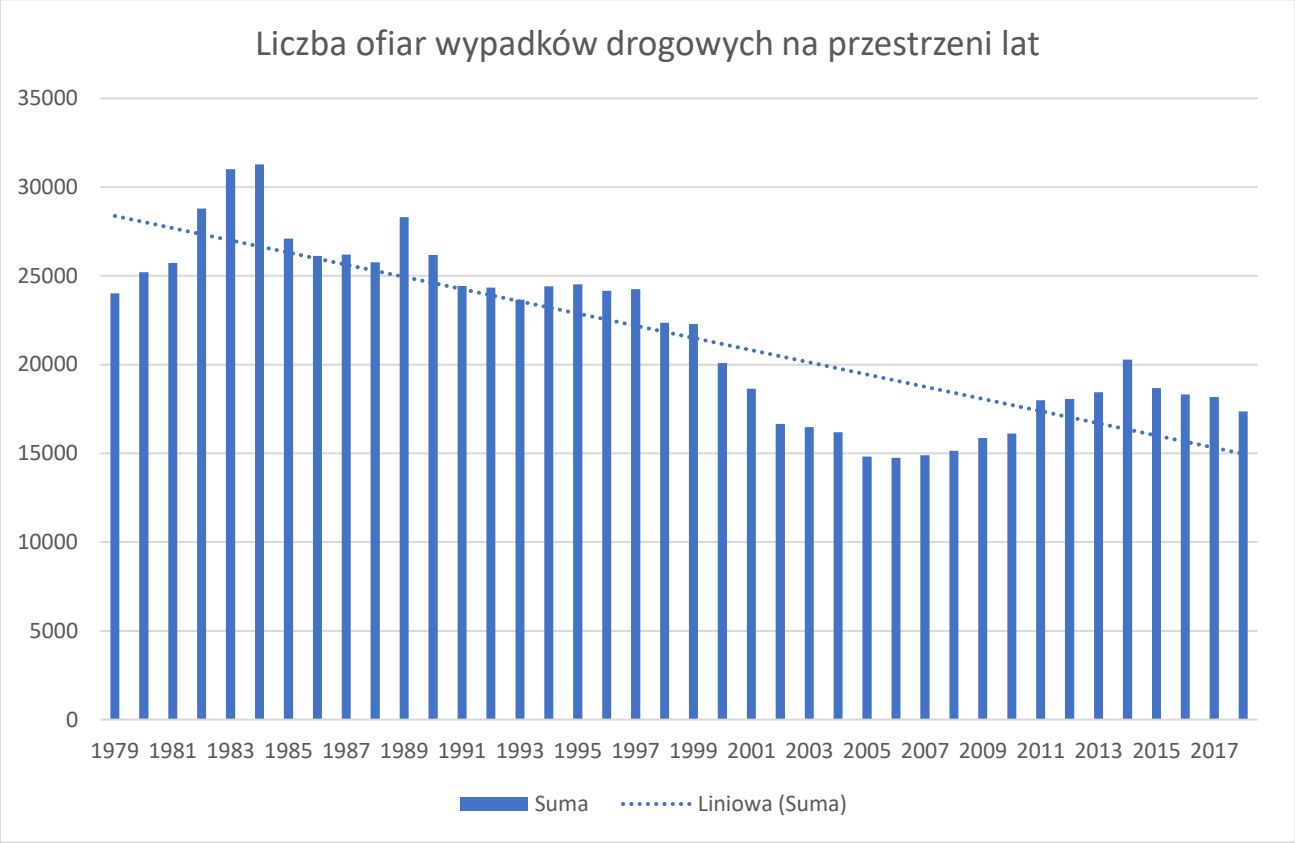
Rysunek *23*. Liczba ofiar wypadków drogowych w zależności od stanu drogi na której doszło do wypadku i obrażeń odniesionych przez osoby biorące w nim udział



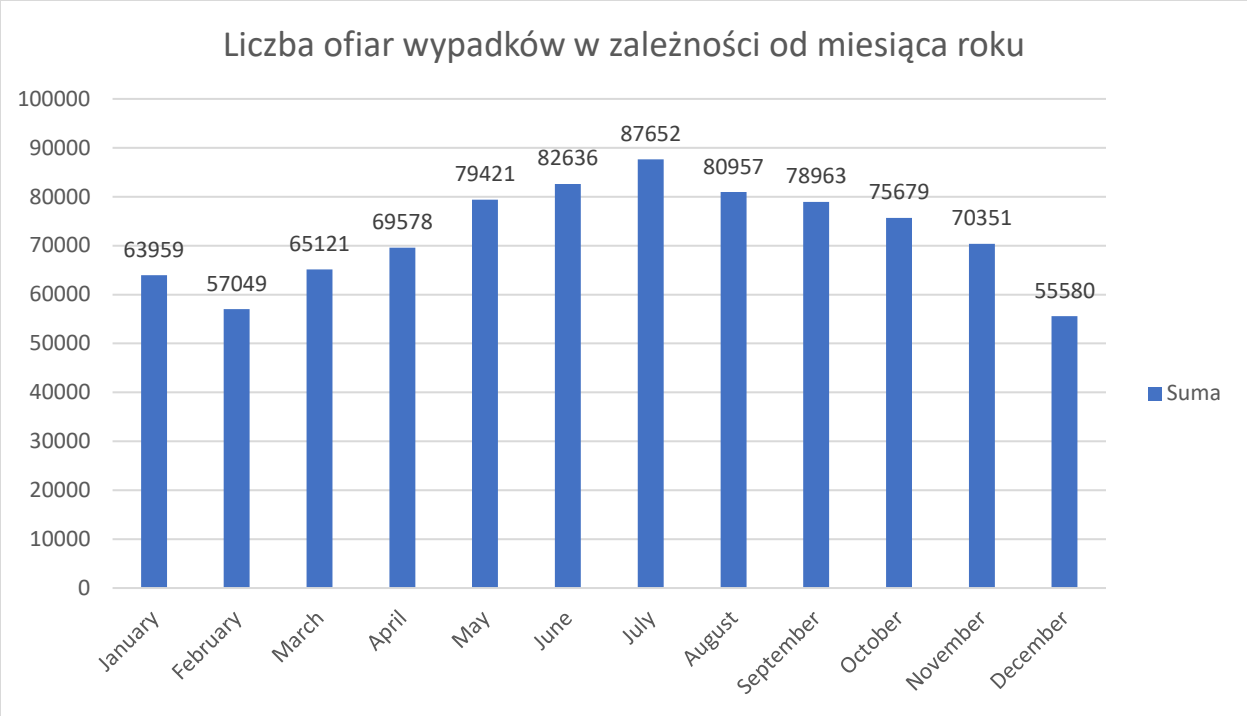
Rysunek *24*. Liczba ofiar wypadków drogowych w zależności od ograniczeń prędkości na drodze na której doszło do wypadku i obrażeń odniesionych przez osoby biorące w nim udział

Kwestia określenia tego, które czynniki zewnętrzne mają największy wpływ na ilość występujących wypadków drogowych jest z natury myląca. Na podstawie rysunków 22, 23 i 25 można bowiem wywnioskować, że najbardziej niebezpieczna pora na jazdę na rowerze to słoneczny dzień oraz sucha nawierzchnia – i faktycznie, statystyki pokazują, że jest to prawda, nie jest to jednak powód tego, że jazda po mokrej nawierzchni jest bezpieczniejsza niż jazda po suchej nawierzchni – wręcz przeciwnie – takie wyniki są bowiem spowodowane tym, że ludzie zdecydowanie najczęściej wsiadają za kierownicę roweru w ciepłą, przejrzystą pogodę i z tego wynika stronniczość danych, które można ujrzeć na wykresach. Można tu dostrzec też takie dane, które wskazują na to, że najbardziej niebezpieczne są drogi o ograniczeniu prędkości 30 mph – wynika to z tego, że jest to najczęściej spotykane ograniczenie w obszarach zabudowanych (jest to odpowiednik ograniczenia prędkości 50 km/h w Polsce w obszarze zabudowanym) – jako że rowerzyści najczęściej poruszają się właśnie po obszarach zabudowanych, to tam dochodzi do największej ilości wypadków. Podobnie wygląda analiza typów dróg – najwięcej wypadków odbywa się na drogach ‘single carriageway’,‘dual carriageway’ oraz ‘roundabout’ (rondo), czyli na drogach najczęściej spotykanych właśnie w miastach. Różnice pomiędzy innymi typami dróg są nieznaczne, tak samo jak nieznaczne są różnice poza innymi warunkami pogodowymi niż ‘Clear’ i ‘Rain’ (ale przy zdecydowanej przewadze przypadków z ‘Clear’), innymi stanami dróg niż ‘Dry’ i ‘Wet’ (przy zdecydowanej przewadze przypadków z ‘Dry’), innym oświetleniu niż ‘Daylight’ (jazda w pełnym świetle, w ciągu dnia) i przy ograniczeniach innych niż 30, 40 i 60 (przy zdecydowanej przewadze ograniczenia 30mph).

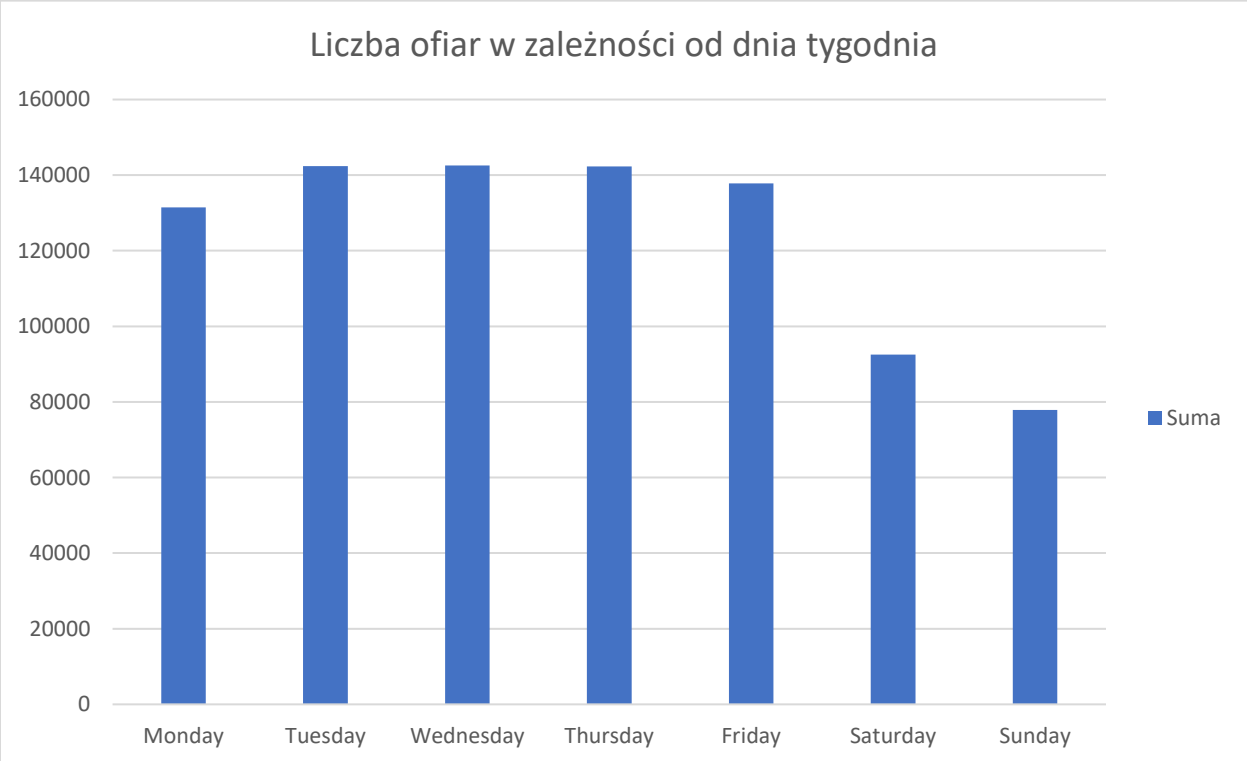
- Jakie są tendencje dotyczące ilości oraz powagi skutków wypadków drogowych tego typu na przestrzeni ostatnich lat?



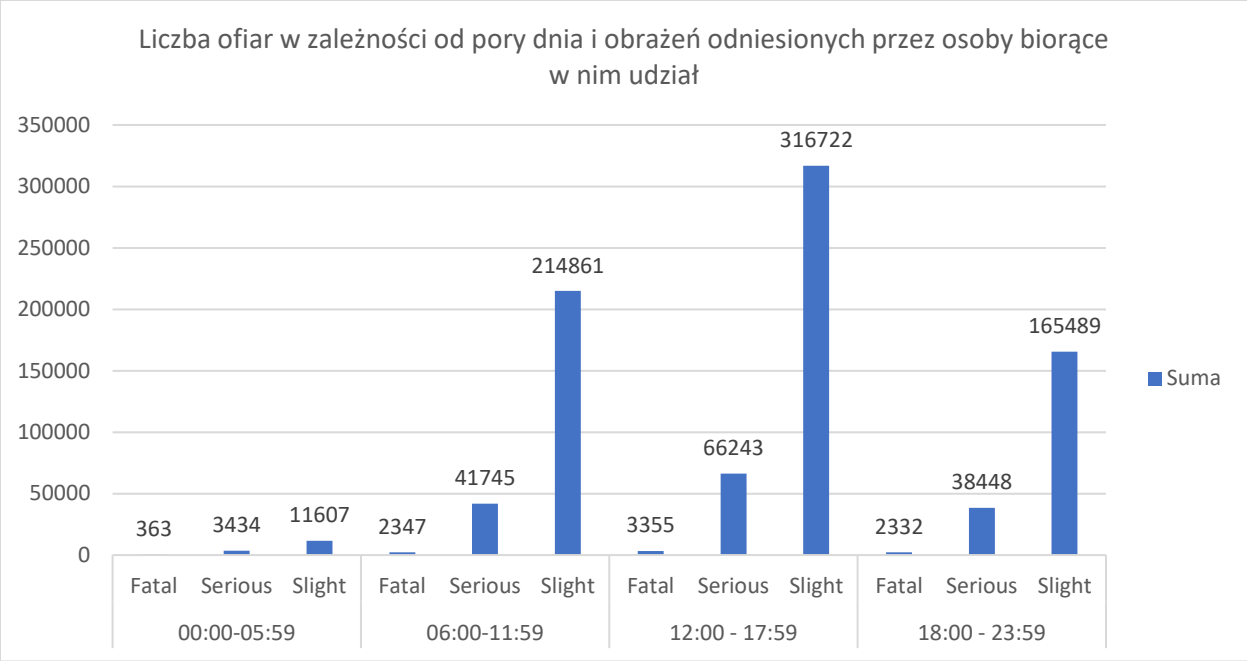
Rysunek *25*. Liczba ofiar wypadków drogowych na przestrzeni lat



Rysunek *26*. Liczba ofiar wypadków w zależności od miesiąca roku



Rysunek *27*. Liczba ofiar wypadków w zależności od dnia tygodnia



Rysunek *28*. Liczba ofiar w zależności od pory dnia i obrażeń odniesionych przez osoby biorące w nim udział

Określenia tego jakie są tendencje na przestrzeni dnia, miesiąca, tygodnia podjąłem się zaczynając od skali roku – po sporządzonych przeze mnie wykresach widocznych na rysunkach 27 i 28 widać, że tendencja jest spadkowa według mojego zestawu danych, dane te nie są jednak liniowe – widoczne są wyraźne fluktuacje – największe wartości przypadają na rok 1984, a najmniejsze 2006. Ogólny spadek ilości wypadków można przypisać temu, że z roku na rok rozbudowywana jest infrastruktura dla transportu rowerowego, np. drogi dla rowerów, co zwiększa bezpieczeństwo poruszania się w ten sposób. Nie jest jasne, czemu w roku 2006 wypadków było najmniej i czemu od tamtego czasu ta liczba wypadków wzrosła, można jednak domyślać się, że może to być spowodowane rosnącą popularnością poruszania się rowerem – w ostatnich latach można było zaobserwować chociażby kampanie reklamowe zachęcające do dostawania się do pracy rowerem, w Anglii sytuacja wygląda więc zapewne bardzo podobnie. W skali miesięcznej najwięcej wypadków przypada na najcieplejszy okres – tendencja jest wzrostowa od lutego aż do lipca, a następnie spadkowa przez resztę roku. Skala tygodniowa pokazuje, że dane nie różnią się zbytnio i ilości wypadków każdego dnia są bardzo podobne – jedyne różnice, które są godne zauważenia, to niewielka liczba incydentów w okresie weekendu (sobota i niedziela), co bez wątpienia wynika z tego, że ludzie nie muszą w te dni jeździć do pracy, a osoby młodsze nie chodzą wtedy do szkół. Skala dobowa również nie prezentuje nic zaskakującego – największa liczba wypadków przypada na okres powrotu ze szkoły oraz pracy, czyli godziny z zakresu 12:00 do 17:59. Zakresy godzinowe 06:00-11:59 i 18:00-23:59, choć bez wątpienia różne (więcej wypadków w godzinach 06:00-11:59), nie są żadnym zaskoczeniem, tak samo jak fakt, że w godzinach nocnych (00:00-05:59) liczba wypadków spada drastycznie względem pozostałych przedziałów w ciągu doby.

7.2. Podsumowanie - wnioski z analizy

Na podstawie przeprowadzonej analizy danych w kontekście postawionych z początku pytań badawczych można wyciągnąć następujące wnioski:

- Najbardziej narażeni na wypadki rowerowe są mężczyźni, najbardziej zagrożona jest grupa wiekowa 11 do 15 lat, a najmniej – osoby w wieku od 66 do 75 lat
- Czynniki zewnętrzne takie jak warunki pogodowe, oświetleniowe, rodzaj i stan dróg oraz ograniczenia prędkości, choć bez wątpienia są wpływowe, na podstawie posiadanych przeze mnie danych nie jest możliwe wyciągnięcie żadnych rzetelnych wniosków – najbardziej decydującym czynnikiem jest to, jakie warunki zachęcają ludzi do wyjechania rowerem na drogę i stąd przede wszystkim wynika stronniczość danych według badanego przeze mnie zestawu
- Inwestycje w infrastruktury drogowe na przestrzeni lat przyczyniły się do tego, że liczba wypadków spada – jest ich dziś znacznie mniej niż w latach 1980-tych, dokąd sięga ten zestaw danych, a obecna tendencja wzrostowa wskazuje na to, że rowerzystów w miastach jest dziś coraz więcej dzięki rosnącej świadomości klimatycznej społeczeństwa. Do największej ilości wypadków dochodzi w miesiącach letnich, w ciągu tygodnia (poniedziałek – piątek) oraz w ciągu dnia, kiedy ulice są najbardziej zatłoczone. Dalsza pomogłaby lepiej zrozumieć przyczyny obecnej tendencji wzrostowej, aby możliwe było lepsze zapobieganie kolejnym wypadkom.

Niestety nie udało się osiągnąć w pełni postawionego na początku projektu celu – stronniczość zestawu danych uniemożliwiła jednoznaczną identyfikację czynników ryzyka i tego jaki jest ich wpływ na ilość występujących wypadków drogowych, przez co udzielenie dokładnych sugestii dotyczących działań, które zwiększyłyby bezpieczeństwo na drogach pozostaje niemożliwe. Analiza czynników zewnętrznych nie dostarcza jednoznacznych wniosków o ich wpływie na ilość wypadków rowerowych. W celu uzyskania bardziej kompleksowej analizy i wyciągnięcia bardziej precyzyjnych wniosków zalecane byłoby kontynuowanie badań i uwzględnienie dodatkowych czynników, taich jak dokładne dane lokalizacyjne oraz dotyczące infrastruktury rowerowych, co umożliwiłoby dokładną analizę przyczyn wypadków oraz potencjalnych działań zapobiegawczych.

8. Wnioski końcowe z realizacji projektu

8.1. Problemy

Podczas realizacji projektu największy kłopot sprawiła mi analiza danych – uważam, że choć udało mi się ją sporządzić, mogłaby ona być na wyższym poziomie oraz że zestaw danych brakował kilku dodatkowych rekordów nowych informacji, które umożliwiłyby na przeprowadzenie dokładniejszej analizy i dokładniejsze odseparowanie przyczyn występowania wypadków rowerowych innych niż to, że najważniejszym czynnikiem ilości wypadków jest ruch drogowy i ilość rowerzystów na drogach. Problem stanowiło również dla mnie sporządzenie odpowiedniego schematu procesów ETL, przede wszystkim danych do tabeli faktu FactAccident – wymagało to ode mnie zmiany podejścia i dodania dodatkowej tabeli tymczasowej ‘Accidents_Bikers_Combined’, aby uniknąć utraty części rekordów. Uważam, że

bardzo pomocne byłoby wprowadzenie przynajmniej jednych zajęć więcej z ETL'a w semestrze, jako że jestem pewny, że przy nieco większej wiedzy praktycznej z zajęć rozwiązanie tego problemu nie sprawiłoby mi takich kłopotów.

8.2. Pozyskana wiedza i doświadczenie

Uważam, że zdobyta przeze mnie wiedza jest całkiem duża – chociażby zapoznanie się z procesem ETL. Bez wątpienia wzrosły też moje umiejętności w języku SQL – co prawda nie jest on powszechnie używany w projekcie, czasem był on konieczny i stanowiło to dla mnie dobry trening. Ciekawe było też zapoznanie się z narzędziami Tableau oraz Tableau Prep – co prawda ostatecznie wykresy sporządziłem za pomocą Microsoft Excel, bez wątpienia polubiłem także nowe oprogramowanie od Tableau, które stanowi ciekawą alternatywę. Sama analiza również pomogła mi na rozwinięcie pewnych umiejętności, jako że zmusiła mnie do szukania ciekawych zależności między danymi w sporządzonej przeze mnie hurtowni danych oraz do ciągłych przemyśleń na temat otrzymanych wyników w szerszym kontekście.

9. Źródła informacji użyte w etapie analizy danych

Dane klimatyczne dotyczące Wielkiej Brytanii: https://pl.weatherspark.com/y/45062/%C5%9Arednie warunki-pogodowe-w:-Londyn-Wielka-Brytania-w-ci%C4%85gu-roku

Ogranicznia prędkości w Wielkiej Brytanii: https://www.autoeurope.pl/informacje-drogowe wielka-brytania