Hurtownie danych – Projekt HD

PWr. Wydział Informatyki i Telekomunikacji Data: 13.06.2022

Student	----------	Ocena
Indeks	256305	
Imię	Grzegorz	
Nazwisko	Dzikowski	

1. Tytuł projektu

Analiza wypadków samochodowych w Wielkiej Brytanii w latach 2005 -2015

2. Charakterystyka dziedziny problemowej

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

Wypadki samochodowe corocznie powodują wiele śmierci oraz kalectw ludzi. Pomimo kampanii społecznych, kontroli drogowych oraz coraz nowszych aut, wypadki dalej występują na drogach.

Faktem, jakim będziemy się zajmować podczas analizy, to wypadek drogowy. Wymiarami są wiek pojazdu, wiek kierowcy, warunki pogodowych, warunki drogowe, jakość drogi, ograniczenie prędkości, liczby poszkodowanych oraz poważność wypadku. Wymiarami są data, czas, lokalizacja, departament policji, kierowca, przyczyna, skutek, przeprowadzona akcja. Miarą faktu jest liczba zatrzymań.

2.2 Problemy

P01 – rosnąca liczba wypadków samochodowych
P02 – wzrost liczby ofiar śmiertelnych
P03 – niszczenie infrastruktury przez wypadki samochodowe
P04 – nieefektywność regulacji na ograniczenie liczby wypadków

2.3 Cel przedsięwzięcia

2.3.1 Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

Analiza udostępni analizę faktów dotyczących wypadków i odpowiedzi na, między innymi, następujące pytania:

1. Jakie jest przekrój wiekowy oraz płciowy ofiar oraz kierowców?
2. Czy starsze auta są bezpieczniejsze?
3. Czy starsi wiekowo kierowcy jeżdżą bezpieczniej?

4. Czy limit prędkości ma wpływ na bezpieczeństwo na drogach?
5. Czy warunki na drodze mają wpływ na bezpieczeństwo?
6. Czy typ drogi ma wpływ na liczbę wypadków?

Właściwa analiza powinna odpowiedzieć na powyższe pytania

2.3.2 Zakres analizy – badane aspekty

Analiza odbędzie się na wielu płaszczyznach. Będzie można dzięki temu podjąć działania ograniczające liczbę wypadków na wielu poziomach, tj. miejsce zdarzenia, warunku pogodowe, profil kierowcy czy typ pojazdu.

2.3.3 Potencjalni użytkownicy

Baza analityczna będzie wspierać ministerstwo transportu w decyzjach dotyczących bezpieczeństwa ruch drogowego, oraz architektów i planistów w decyzjach dotyczących budowy nowych dróg

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	~ 1 780 000	238	Wszystkie wypadki drogowe w latach 2005-2015 w UK
2.	Casualities	.csv	~2 400 000	105	Ofiary w wypadkach drogowych
3.	Vehicles	.csv	~3 200 000	201	Pojazdy uczestniczące w wypadkach
4.	Road-Safety-Open-Dataset-Data-Guide	.xlsx	1580	0.55	Objaśnienie danych w tabelach wyżej

3.2. Lokalizacja, dostępność danych źródłowych

Dane pochodzą z https://www.kaggle.com/datasets/silicon99/dft-accident-data?resource=download, które dla odmiany są zebrane z https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data. Tam też znajduje się słownik pojęć i jego interpretacja

3.3. Słownik danych – interpretacja

Wszystkie dane w tabelach Accidents, Vehicles i Casualities są w formacie numerycznym, które potem jest tłumaczony na odpowiednie wartości przy pomocy pliku RoadSafetyGuide, dlatego w tej tabeli interpretuje i wyjaśniam finalne wartości artybutów

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.	[1st_Road_Class]	Tekstowy	Oznaczenie typu drogi głównej	Typ drogi głównej, na której był wypadek.
2.	[1st_Road_Number]	Numeryczny, całkowity	Numer drogi głównej	
3.	[2nd_Road_Class]	Tekstowy	Oznaczenie typu drogi drugorzędnej	Droga, która znajduje się na skrzyżowaniu 20 metrów od wypadku i głównej drogi
4.	[2nd_Road_Number]	Numeryczny, całkowity	Numer drogi drugorzędnej	Numer drogi na skrzyżowaniu 20 metrów od wypadku.
5.	[Local_Authority_(District)]	Tekstowy	Nazwa dystryktu lokalnych władz	Dystrykt lokalnych władz
6.	[Local_Authority_(Highway)]	Tekstowy	Nazwa jurysdykcji autostrady	Pod jaką jurysdykcją jest autostrada?

7.	[Pedestrian_Crossing-Human_Control]	Tekstowy	Nazwa oznaczająca typ kontroli przejścia dla pieszych	Czy jakiś człowiek kontrolował przejście dla pieszych do 50 metrów od wypadku? Np. szkolny patrol
8.	[Pedestrian_Crossing-Physical_Facilities]	Tekstowy	Nazwa typu przejścia dla pieszych	Charakterystyka fizyczna przejścia dla pieszych na wypadku lub 50 metrów od niego
9.	Accident_Index	Tekstowy	13 znakowy unikalny identyfikator wypadku, klucz naturalny.	Identyfikator wypadku. Używany do łączenia z Casuality i Vehicle
10.	Accident_Severity	Tekstowy	Nazwa oznaczająca poważność wypadku	Oznacza, czy wypadek miał ofiary śmiertelne, poważnie ranne lub tylko lekko ranne.
11.	Carriageway_Hazards	Tekstowy	Nazwa oznaczająca typ zagrożenia na jezdni	Dodatkowe zagrożenia na jezdnii, które znajdowały się na niej w trakcie wypadku
12.	Date	Data	Data w formacie DD/MM/YYYY	Data wystąpienia incydentu
13.	Day_of_Week	Tekstowy	Dzień tygodnia wypadku	
14.	Did_Police_Officer_Attend_Scene_of_Accident	Boolean	Nazwa oznaczająca obecność policjanta przy wypadku	Czy był policjant na miejscu wypadku?
15.	Junction_Control	Tekstowy	Nazwa oznaczająca typ sterowania skrzyżowaniem	Rodzaj sterowania na skrzyżowaniu 20 metrów od wypadku

16.	Junction_Detail	Tekstowy	Nazwa oznaczająca typ skrzyżowania	Typ skrzyżowania w pobliżu 20 metrów od wypadku
17.	Latitude	Numeryczny, Zmiennoprzecinkowy	Szerokość geograficzna, wartość z zakresu od -90 do 90	
18.	Light_Conditions	Tekstowy	Nazwa oznaczająca typ warunków oświetlenia na drodze	Warunki oświetlenia na drodze
19.	Location_Easting_OSGR	Numeryczny, całkowity	Numer siatki wschód – zachód	Lokalizacja wschód zachód na oficjalnej siatce lokalizacji w UK gridreferencefinder.com
20.	Location_Northing_OSGR	Numeryczny, całkowity	Numer siatki północ – południe	Lokalizacja północ południe na oficjalnej siatce lokalizacji w UK gridreferencefinder.com
21.	Longitude	Numeryczny, Zmiennoprzecinkowy	Długość geograficzna, poprawne dane od -180 do 180	
22.	LSOA_of_Accident_Location	Tekstowy	9 znakowy ciąg oznaczający numer geograficzny	Tylko Anglia i Walia – Lower Layer Super Output Areas (LSOA) to geograficzna hierarchia stworzona w celu ulepszenia raportowania statystyk lokalnych w Anglii I Walii Źródło
23.	Number_of_Casualties	Numeryczny, całkowity	Liczba ofiar wypadku	
24.	Number_of_Vehicles	Numeryczny, całkowity	Liczba pojazdów uczestniczących w wypadku	

25.	Police_Force	Tekstowy	Nazwa oznaczająca oddział policji obecny przy wypadku	Oddział policji zajmujący się tym wypadkiem,
26.	Road_Surface_Conditions	Tekstowy	Nazwa oznaczająca warunki na drodze	Warunki na drodze, panujące w momencie wypadku np. wilgoć
27.	Road_Type	Tekstowy	Nazwa oznaczająca typ drogi	Typ drogi, na której odbył się wypadek
28.	Special_Conditions_at_Site	Tekstowy	Nazwa oznaczająca typ warunków na drodze	Specjalne warunki na miejscu wypadku, np. nie działające światła lub prace drogowe
29.	Speed_limit	Numeryczny, całkowity	Ograniczenie prędkości na drodze.	
30.	Time	Czas	Czas wypadku, z dokładnością do minut, format HH:MM	
31.	Urban_or_Rural_Area	Tekstowy	Nazwa oznaczająca typ terenu	Czy teren miejski czy wiejski?
32.	Weather_Conditions	Tekstowy	Typ warunków pogodowych w trakcie wypadku	Warunki pogodowe na drodze, np. deszczowo

Plik: Casualties.CSV				
Lp.	**Atrybut**	**Typ danych**	**Znaczenie**	**Uwagi**
1.	Accident_Index	Tekstow	13 znakowy unikalny identyfikator wypadku, klucz naturalny.	Identyfikator naturalny wypadku. Używany do łączenia z Accident i Vehicle

2.	Age_Band_of_Casualty	Tekstowy	Nazwa oznaczająca przedział wiekowy	Przedział wiekowy ofiary, umożliwia podzielenie ofiar na grupy wiekowe
3.	Age_of_Casualty	Numeryczny, Całkowity	Wiek ofiary	
4.	Bus_or_Coach_Passenger	Tekstowy	Nazwa oznaczająca typ pasażera autobusu	Czy to osoba będąca w autobusie? Jeżeli tak, to gdzie znajdowała się w momencie wypadku
5.	Car_Passenger	Tekstowy	Typ pasażera w aucie	Czy to pasażer auta? Jeżeli tak, to gdzie znajdował się w momencie wypadku?
6.	Casualty_Class	Tekstowy	Klasa ofiary	Klasa ofiary, to jest, czy ofiara była kierowcą czy pasażeremI
7.	Casualty_Home_Area_Type	Tekstowy	Nazwa oznaczająca pochodzenie ofiary	Pochodzenie ofiary, np. Małe miasto
8.	Casualty_Reference	Numeryczny, całkowity	Unikalne ID ofiary w ramach wypadku, klucz obcy	Ten identyfikator wylicza unikalne ofiary w ramach wypadku
9.	Casualty_Severity	Tekstowy	Poważność ofiary wypadku	Czy ofiara była śmiertelna, ciężko ranna czy lekko ranna?
10.	Casualty_Type	Tekstowy	Nazwa oznaczająca typ ofiary	Czy ofiara była np. pieszym?
11.	Pedestrian_Location	Tekstowy	Lokalizacja pieszego w momencie wypadku	Czy pieszy był np. na przejściu?
12.	Pedestrian_Movement	Tekstowy	Jak poruszał się pieszy?	Sposób poruszania się pieszego
13.	Pedestrian_Road_Maintenance_Worker	Tekstowy	Czy pieszy był pracownikiem budowy?	

14.	Sex_of_Casualty	Tekstowy	Płeć kierowcy	
15.	Vehicle_Reference	Numeryczny	Unikalne ID każdego pojazdu w ramach wypadku. Klucz sztuczny	Umożliwia połączenie ofiary z pojazdem

Plik: Vehicles.CSV				
Lp.	**Atrybut**	**Typ danych**	**Znaczenie**	**Uwagi**
1.	[1st_Point_of_Impact]	Tekstowy	W jaką część auta uderzyło auto po raz pierwszy?	Miejsce pierwszego uderzenia
2.	[Engine_Capacity_(CC)]	Numeryczny, całkowity	Liczba oznaczająca pojemność silnika	Pojemność silnika w CC
3.	[Vehicle_Location-Restricted_Lane]	Tekstowy	Miejsce na pasie z ograniczonym roku	Typ pasa awaryjnego, na jakim znajduje się pojazd po wypadku
4.	[Was_Vehicle_Left_Hand_Drive?]	Boolean	Czy pojazd miał kierownicę po lewej stronie?	Na potrzeby finalnych danych, ta dana będzie zamieniona z booleana na nazwy typów „Left”, „Right”. W UK standardem jest kierownica po prawej stronie
5.	Accident_Index	Tekstowy	13 znakowy unikalny	Używany do łączenia z

			identyfikator wypadku, klucz naturalny.	Accident i Vehicle
6.	Age_Band_of_Driver	Tekstowy	Określenie grupy wiekowej	Grupa wiekowa kierowcy
7.	Age_of_Driver	Numeryczny	Wiek kierowcy	
8.	Age_of_Vehicle	Numeryczny	Wiek auta	Wiek auta liczony jest od roku produkcji do dnia wypadku
9.	Driver_Home_Area_Type	Tekstowy	Oznaczenie typu pochodzenia kierowcy	Pochodzenie kierowcy, w znaczeniu czy pochodzi z miasta czy wsi
10.	Driver_IMD_Decile	Tekstowy	Wskaźnik IMD Kierowcy	Wskaźnik IMD kierowcy, wskazujący na poziom miejsca, z którego pochodzi kierowca Źródło
11.	Hit_Object_in_Carriageway	Tekstowy	W jaki obiekt uderzył pojazd na drodze?	Obiekt na drodze, np. inne auto, które bezpośrednio spowodowało wypadek

12.	Hit_Object_off_Carriageway	Tekstowy	W jaki obiekt uderzył pojazd poza drogą	Obiekt poza drogą, np. latarnia, które bezpośrednio spowodowało wpadek
13.	Journey_Purpose_of_Driver	Tekstowy	W jakim celu osoba podróżowała?	Np. Rekreacyjnie lub jako praca
14.	Junction_Location	Tekstowy	Miejsce na skrzyżowaniu	Pozycja na skrzyżowaniu po wypadku
15.	Propulsion_Code	Tekstowy	Typ napędu	Rodzaj napędu pojazdu, zwłaszcza typ paliwa przyjmowanego przez pojazd, np. benzyna
16.	Sex_of_Driver	Tekstowy	Płeć kierowcy,	
17.	Skidding_and_Overturning	Tekstowy	Typ wywrotki lub poślizgu	Czy auto wpadło w poślizg lub wywróciło się?
18.	Towing_and_Articulation	Tekstowy	Typ przyczepy	Czy posiadał Przyczepę?
19.	Vehicle_Leaving_Carriageway	Tekstowy	Sposób opuszczenia jezdni	Czy pojazd opuścił jezdnię i w jaki sposób?
20.	Vehicle_Manoeuvre	Tekstowy	Rodzaj manewru, który przyczynił się do wypadku	Co to był za typ manewru?

21.	Vehicle_Reference	Numeryczny, całkowity	Unikalny numer pojazdu w ramach wypadku	Pozwala na powiązanie ofiary z pojazdem
22.	Vehicle_Type	Tekstowy	Typ pojazdu	Np. Auto, autobus

3.4. Ocena jakościowa danych

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

Dane o wysokiej jakości

Dane o niskiej jakości

Dane nieistotne w analizie

W danych źródłowych bardzo często -1 jest trakowane jako NULL, dlatego ten parament będę interpretował jako null

Tabela 3. Ocena jakościowa danych

Plik: AccidentsCSV					
Lp.	**Atrybut**	**Typ danych**	**Zakres wartości**	**Znaczenie**	**Uwagi - ocena jakości danych**
1.	[1st_Road_Class]	Numeryczny, Całkowity	1 do 6	Numer tłumaczony na rodzaj drogi	0% null/-1 Dane są dobre jakościowo, przydane do analizy
2.	[1st_Road_Number]	Numeryczny, Całkowity	-1 do 9999	Numer drogi w UK	0% null/-1 Dane nie przydatne do analizy, z racji, że nie zajmujemy się w niej lokalizacją
3.	[2nd_Road_Class]	Numeryczny,	-1 do 6	Numer tłumaczony na	41% null/-1. Wartość przydatna do analizy jakości

		Całkowity		rodzaj drogi drugorzędnej	dróg, ale niestety niskiej jakości
4.	[2nd_Road_Number]	Numeryczny, Całkowity	-1 do 9999	Numer drogi drugorzędnej	0% null/-1, natomiast 77% ma wartość 0 - "Unclasified", więc dane bardzo niskiej jakości. Na szczęście nie przydatne do analizy.
5.	[Local_Authority_(District)]	Numeryczny, Całkowity	1 do 941	Numer oznaczający dystrykt lokalnej policji	0% null/-1, Dana nieprzydatna do analizy
6.	[Local_Authority_(Highway)]	Tekstowy	9 znaków	9 znakowy identyfikator lokalnego oddziału policji zajmującego się autostradą	0% null/-1, Dana nieprzydatna do analizy
7.	[Pedestrian_Crossing-Human_Control]	Numeryczny, Całkowity	-1 do 2	Numer oznaczający osobę kontrolującą przejście dla pieszych	0% null/-1, Dana nieprzydatna do analizy
8.	[Pedestrian_Crossing-Physical_Facilities]	Numeryczny, Całkowity	-1 do 8	Numer oznaczający fizyczne ograniczenia na przejściu na pieszych	0% null/-1, Dana nieprzydatna do analizy
9.	Accident_Index	Tekstowy	13 znaków	13 znakowy unikalny identyfikator wypadku, klucz naturalny.	0% null/-1, 100% Key Strength. Ten klucz naturalny jest świetny jako klucz główny do identyfikacji wypadków

10.	Accident_Severity	Numeryczny, Całkowity	1 do 3	Numer oznaczający poważność wypadku	0% null/-1 Atrybut może wydawać się przydatny, ale on jest podmiotem naszej analizy, i przechowywanie i ładowanie go jest redundantne
11.	Carriageway_Hazards	Numeryczny, Całkowity	-1 do 7	Numer oznaczający rodzaj zagrożenia na jezdni	0% null,-1, 98% ma wartość 0 – “None”. Dane nieprzydatne do analizy
12.	Date	Data	01.01.2005 do 31.12.2015, 4017 unikanych dat	Data w formacie DD/MM/YYYY	0% null/-1, Dana bardzo ważna na potrzeby analizy czasowej
13.	Day_of_Week	Numeryczny, Całkowity	1 do 7	Numer oznaczająca dzień tygodnia	0% null/-1, Dana nieprzydatna
14.	Did_Police_Officer_Attend_Scene_of_Accident	Numeryczny, Całkowity	-1 do 3	Numer oznaczający obecność policjanta przy wypadku	0% null/-1, Dana nieprzydatna
15.	Junction_Control	Numeryczny, Całkowity	-1 do 4	Numer oznaczający rodzaj sygnalizacji na skrzyżowaniu	0% null/-1, -1 ma 36% wartości. Dana częściowo niskiej jakości, bo mało jest informacji o rodzaju sterowania na przejściu, ale atrybut nie jest przydatny do analizy

16.	Junction_Detail	Numeryczny, Całkowity	-1 do 9	Numer oznaczający szczegóły skrzyżowania	0% null/-1, Dana nieprzydatna do analizy
17.	Latitude	Numeryczny, Zmienn oprzecinkowy	49.912941 do 60.757544	Pozycja GPS wypadku	< 1% null, Dana nieprzydatna do analizy
18.	Light_Conditions	Numeryczny, Całkowity	1 do 7	Numer oznaczający warunki oświetleniowe na drodze	0% null/-1. Dana przydatna do analizy wpływu światła na wypadki
19.	Location_Easting_OSGR	Numeryczny, Całkowity	64950 do 655540	Numer siatki wschód zachód według OSGR	< 1% null/-1, Dana nieprzydatna do analizy
20.	Location_Northing_OSGR	Numeryczny, Całkowity	10290 do 128800	Numer siatki północ południe według OSGR	< 1% null/-1, Dana nieprzydatna do analizy
21.	Longitude	Numeryczny, Całkowity	-7.516225 do 1.76201	Długość geograficzna	< 1% null/-1, Dana nieprzydatna do analizy
22.	LSOA_of_Accident_Location	Tekstowy	9 znaków	Symbol Lower Layer Super Output Areas (LSOA)	7% null/-1, Dana nieprzydatna do analizy
23.	Number_of_Casualties	Numeryczny, Całkowity	1 do 93	Liczba ofiar uczestnicząca w wypadku	0% null/-1, Dane mają małą szczegółowość. Na podstawie powiązań z ofiarami będę samodzielnie przeliczał te wartości

24.	Number_of_Vehicles	Numeryczny, Całkowity	1 do 67	Liczba pojazdów uczestniczących w wypadku	0% null/-1, dane istotne jako miara
25.	Police_Force	Numeryczny, Całkowity	1 do 98	Numer oznaczający oddział policji zajmujący się wypadkiem	0% null/-1, Dana nieprzydatna
26.	Road_Surface_Conditions	Numeryczny, Całkowity	-1 do 5	Numer oznaczający warunki na drodze	<1% null/-1, Dana przydatna do analizy jakości drogi
27.	Road_Type	Numeryczny, Całkowity	1 do 9	Numer oznaczający typ drogi	0% null/-1, Dana przydatna do analizy jakości drogi
28.	Special_Conditions_at_Site	Numeryczny, Całkowity	-1 do 7	Numer oznaczający dodatkowe warunki na drodze	0% null/-1, 97% ma wartość 0 – “None”. Dana przydatna do analizy, ale bardzo jednolita
29.	Speed_limit	Numeryczny, Całkowity	0 do 70	Ograniczenie prędkości na drodze	0% null/-1, Dana przydatna do analizy wpływu ograniczenia prędkości
30.	Time	Godzina	Od 00:01:00 do 23:59:00 (1 minutowa ziarnistość)	Godzina i Minuta w formacie HH:mm	< 1% null/-1, Dane posiadają godzinę wypadku, co przydatne będzie do analizy pory dnia wypadku
31.	Urban_or_Rural_Area	Numeryczny, Całkowity	1 do 3	Numer oznaczający, czy droga jest miejsca czy wiejska	0% null/-1, Dana przydatna do analizy jakości drogi

32.	Weather_Conditions	Numeryczny, Całkowity	-1 do 9	Numer oznacza warunku pogodowe w momencie wypadku	0% null/-1, Dana przydatna do analizy warunków powstania wypadku

Plik: CasualtyCSV					
Lp.	**Atrybut**	**Typ danych**	**Zakres wartości**	**Znaczenie**	**Uwagi - ocena jakości danych**
1.	Accident_Index	Tekstowy	Tekst o długości 13 znaków, ale 49 wpisów ma długość 1	13 znakowy unikalny identyfikator wypadku, klucz naturalny.	0.002% null
2.	Age_Band_of_Casualty	Numeryczny, Całkowity	-1 do 11	Numer oznaczający grupe wiekową ofiary	2% null/-1
3.	Age_of_Casualty	Numeryczny, Całkowity	-1 do 104	Wiek ofiary	2% null/-1, Dana nieprzydatna ze względu na obecność Age_Band_of_Casualty
4.	Bus_or_Coach_Passenger	Numeryczny, Całkowity	-1 do 4	Numer oznaczający, czy ofiara była w autobusie	0.002% null/-1, Dana przydatna, chociaż większość wartości (97%) to 0 – None
5.	Car_Passenger	Numeryczny, Całkowity	-1 do 2	Numer oznaczający, czy ofiara była pasażerem auta	0.002% null/-1
6.	Casualty_Class	Numeryczny, Całkowity	1 do 3	Numer oznaczający klasę ofiary	0.002% null/-1

7.	Casualty_Home_Area_Type	Numeryczny, Całkowity	-1 do 3	Numer oznaczający typ miejsca, z której pochodzi ofiara	15% null/-1. Dane nieprzydatne analizie
8.	Casualty_Reference	Numeryczny, Całkowity	1 do 852	Numer ofiary w ramach wpadku, klucz sztuczny	0% null/-1. Dana używana do usuwania duplikatów
9.	Casualty_Severity	Numeryczny, Całkowity	1 do 3	Numer oznaczający obrażenia ofiary	0% null
10.	Casualty_Type	Numeryczny, Całkowity	0 do 98	Numer oznaczający typ ofiary	0 % null
11.	Pedestrian_Location	Numeryczny, Całkowity	-1 do 10	Numer oznaczający lokalizację pieszego	0.002% null, atrybut nieistotny dla analizy
12.	Pedestrian_Movement	Numeryczny, Całkowity	-1 do 9	Numer oznaczający sposób poruszania pieszego	0.002% null, atrybut nieistotny dla analizy
13.	Pedestrian_Road_Maintenance_Worker	Numeryczny, Całkowity	-1 do 2	Numer oznaczający, czy pieszy był pracownikiem budowy	0.002% null, atrybut nieistotny dla analizy
14.	Sex_of_Casualty	Numeryczny, Całkowity	-1 do 2	Numer oznaczający płeć ofiary	0.002% null
15.	Vehicle_Reference	Numeryczny, Całkowity	1 do 91	Numer pojazdu w ramach wypadku, w	0.002% null. Dana nieprzydatna dla analizy, ponieważ nie jest istotne

				którym była ofiara	powiązanie ofiary z konkretnym pojazdem

Plik: VehiclesCSV					
Lp.	**Atrybut**	**Typ danych**	**Zakres wartości**	**Znaczenie**	**Uwagi - ocena jakości danych**
1.	[1st_Point_of_Impact]	Numeryczny, Całkowity	-1 do 4	Numer oznaczający miejsce pierwszego uderzenia pojazdu	0.0019% null/-1
2.	[Engine_Capacity_(CC)]	Numeryczny, Całkowity	-1 do 99999	Pojemność silnika	0.0019% null/-1
3.	[Vehicle_Location-Restricted_Lane]	Numeryczny, Całkowity	-1 do 9	Numer oznaczający lokalizację pojazdu po wypadku na pasie awaryjnym	0.0019% null/-1, dana nieistotna dla analizy
4.	[Was_Vehicle_Left_Hand_Drive?]	Prawda/Fałsz, Nieznany (-1), Null	-1 , 1, 2	Numer oznaczający, czy auto ma kierownicę po lewej stronie	0.5% null/-1
5.	Accident_Index	Tekstowy	2 do 13 znaków, 63 wartości -1, klucz sztuczny	13 znakowy unikalny identyfikator wypadku, klucz naturalny.	0 % null/-1, Umożliwia powiązanie pojazdu z wypadkiem

6.	Age_Band_of_Driver	Numeryczny, Całkowity	-1 do 11	Numer oznaczający grupę wiekową kierowcy	11% null/-1
7.	Age_of_Driver	Numeryczny, Całkowity	-1 do 100	Wiek kierowcy	11% null/-1, dana nieistotna ze względu na obecność Age_Band_of_Driver
8.	Age_of_Vehicle	Numeryczny, Całkowity	-1 do 111	Wiek pojazdu	30% null/-1, niestety, dana niskiej jakości, przez co analiza może być nie miarodajna
9.	Driver_Home_Area_Type	Numeryczny, Całkowity	-1 do 3	Numer oznaczający typ terenu z którego pochodzi kierowca	20% null/-1, dana niskiej jakości, ale niepotrzebna w analizie
10.	Driver_IMD_Decile	Numeryczny, Całkowity	-1 do 10	Numer oznaczający wartość IMD kierowcy	33% ma wartość null/-1, dana niskiej jakości, ale niepotrzebna w analizie
11.	Hit_Object_in_Carriageway	Numeryczny, Całkowity	-1 do 12	Numer oznaczający obiekt na jezdni, w który uderzył pojazd	0.0019% null/-1,
12.	Hit_Object_off_Carriageway	Numeryczny, Całkowity	-1 do 11	Numer oznaczający obiekt poza jezdnią, w który uderzył pojazd	0.0019% null/-1
13.	Journey_Purpose_of_Driver	Numeryczny, Całkowity	-1 do 15	Numer oznaczający cel podróży kierowcy	1% null/-1

14.	Junction_Location	Numeryczny, Całkowity	-1 do 8	Numer oznaczający Lokalizacja na skrzyżowaniu po wypadku	0.0019% null/-1, dana nieistotna dla analizy
15.	Propulsion_Code	Numeryczny, Całkowity	-1 do 12	Numer oznaczający typ napędu w pojeździe	26% null/-1, dana słabej jakości, ale nieistotna dla analizy
16.	Sex_of_Driver	Numeryczny, Całkowity	-1 do 3	Numer oznaczający płeć kierowcy	0.0019% null/-1
17.	Skidding_and_Overturning	Numeryczny, Całkowity	-1 do 5	Numer oznaczając typ poślizgu lub wywrotki pojazdu	0.0019% null/-1
18.	Towing_and_Articulation	Numeryczny, Całkowity	-1 do 5	Numer oznaczający typ przyczepy w pojeździe	0.0019% null/-1
19.	Vehicle_Leaving_Carriageway	Numeryczny, Całkowity	-1 do 8	Numer oznaczający sposób opuszczenia jezdni przez pojazd	0.0019% null/-1, dana nieistotna dla analizy
20.	Vehicle_Manoeuvre	Numeryczny, Całkowity	-1 do 9	Numer oznaczający typ manewru wykonywanego przez pojazd przed wypadkiem	0.0019% null/-1

21.	Vehicle_Reference	Numeryczny, Całkowity	1 do 91	Numer pojazdu w wypadku, klucz sztuczny	0 % null/-1, używany do łączenia pojazdu z wypadkiem i do usuwania duplikatów
22.	Vehicle_Type	Numeryczny, Całkowity	-1 do 98	Numer oznaczający typ pojazdu	0.0019% null/-1

4. Analityczne modele wielowymiarowe

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

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

Tabela 4. Fakty podlegające analizie

Lp.	Fakty	Miary	Uwagi
1.	Accident	Severe Casualties, Fatal Casualties, Slight Casualties, Number of Vehicles, Number of Accidents	Miary Severe Casualties, Fatal Casualties I Slight Casualties są kalkulowane na etapie ET, natomiast Numer Of Accidents jest liczbą wypadków z danymi parametrami

4.2. Kontekst analizy faktów

Ustalony kontekst analizy faktów został przedstawiony w tab. 4.

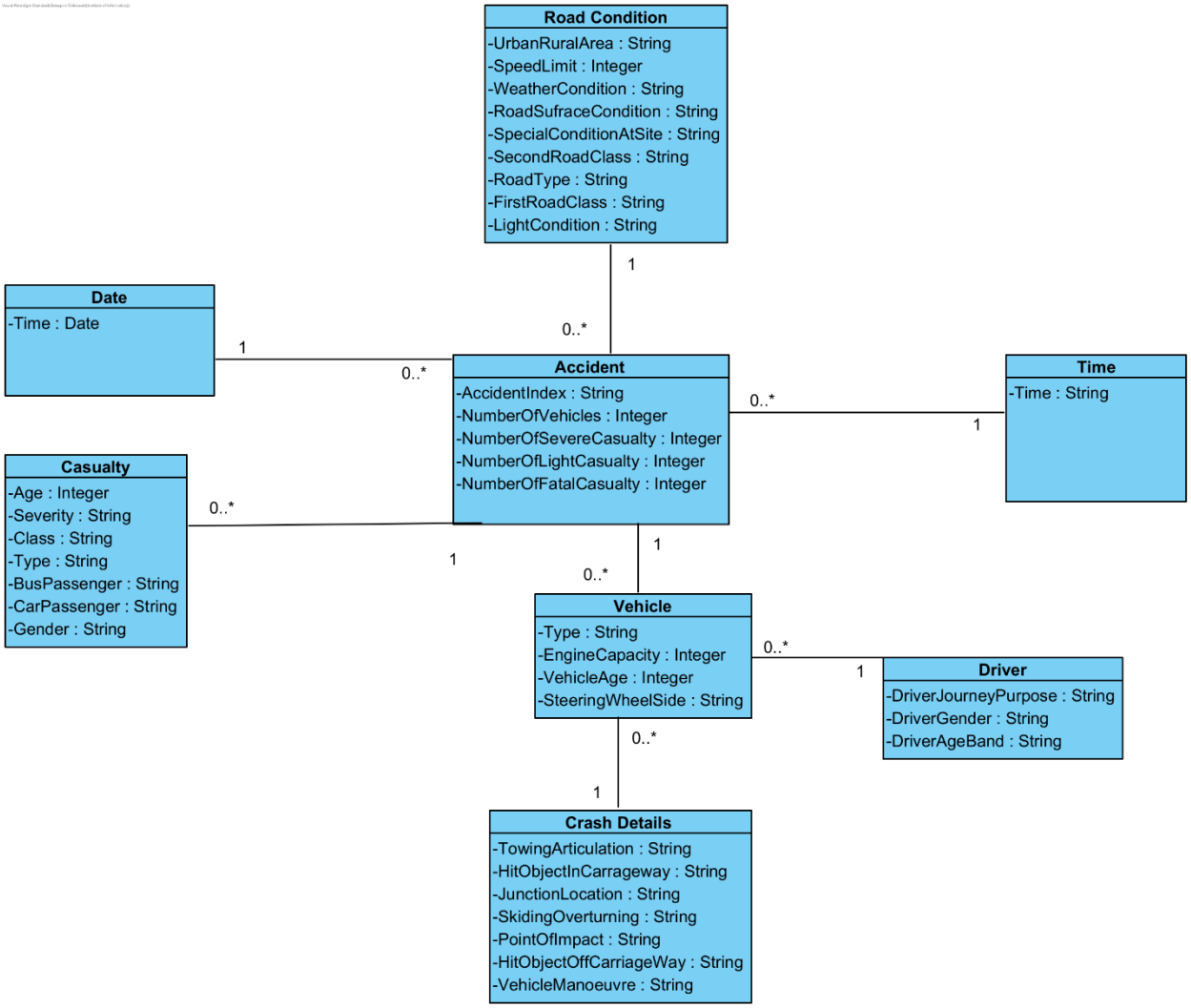
Tabela 5. Wymiary analizy faktów

Lp.	Wymiar	Opis
1.	Casualty	Umożliwia analizę w kontekście informacji na temat ofiary wypadku. Dzięki temu można określić najczęstszy profil ofiary wypadku
2.	Vehicle	Umożliwia analizę w kontekście pojazdu uczestniczącego w wypadku. Dzięki temu można określić typ pojazdu uczestniczącego w wypadku
3.	Driver	Umożliwia analizę biznesową w kontekście kierowców uczestniczących w wypadku. Dzięki temu można określić, jaki typ kierowcy najczęściej uczestniczy w danych wypadkach

4.	Road Condition	Umożliwia analizę biznesową w kontekście warunków pogodowych i drogowych, panujących w trakcie wypadku. Dzięki temu można określić, jakie warunki najczęściej powodują wypadki
5.	Date	Umożliwia analizę czasową, oraz pokazanie zmian wypadków czasie, na przestrzeni lat, miesięcy i dni
6.	Time	Umożliwia analizę godzinową, i pozwala rozłożyć na dzień wypadki
7.	Crash Details	Umożliwia analizę biznesową skutków wypadku dla pojazdu oraz przyczyn wypadku pojazdu

5.1. 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 Accident oraz z 7 wymiarów. Model ten reprezentowany jest w postaci schematu płatku śniegu



Rysunek 1. Wielowymiarowy model analityczny przedstawiony na poziomie konceptualnym

6. Projekt procesu ETL

6.1. Schemat bazy danych HD (skrypt SQL)

Baza danych została utworzona przy pomocy skryptu przedstawionego w Tabela 6 SQL na tworzenie bazy danych

```
--DimAccidentReason

CREATE TABLE [dbo].[DimCrashDetails](
	[TowingArticulation] [nvarchar](100) NOT NULL,
	[HitObjectInCarriageway] [nvarchar](100) NOT NULL,
	[CrashDetailsKey] [bigint] IDENTITY(1,1) NOT NULL,
	[JunctionLocation] [nvarchar](100) NOT NULL,
```

```
	[VehicleLeavingCarriageway] [nvarchar](100) NOT NULL,
	[SkiddingOverturning] [nvarchar](100) NOT NULL,
	[PointOfImpact] [nvarchar](100) NOT NULL,
	[HitObjectOffCarriageway] [nvarchar](100) NOT NULL,
	[VehicleManoeuvre] [nvarchar](100) NOT NULL,
PRIMARY KEY CLUSTERED
(
	[CrashDetailsKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

--DimDriverDetails

CREATE TABLE [dbo].[DimDriverDetails](
	[DriverJourneyPurpose] [nvarchar](100) NOT NULL,
	[DriverGender] [nvarchar](100) NOT NULL,
	[DriverDetailsKey] [bigint] IDENTITY(1,1) NOT NULL,
	[DriverAgeBand] [nvarchar](100) NOT NULL,
 CONSTRAINT [PK_DimDriverDetails] PRIMARY KEY CLUSTERED
(
	[DriverDetailsKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

--DIMVEHICLES
CREATE TABLE [dbo].[DimVehicle](
	[Type] [nvarchar](100) NOT NULL,
	[VehicleIndex] [int] NOT NULL,
	[CrashDetailsKey] [bigint] NOT NULL,
	[EngineCapacity] [nvarchar](100) NOT NULL,
	[VehicleKey] [bigint] IDENTITY(1,1) NOT NULL,
	[VehicleAge] [int] NOT NULL,
	[SteeringWheelSide] [nvarchar](100) NOT NULL,
	[DriverDetailsKey] [bigint] NOT NULL,
 CONSTRAINT [PK_DimVehicle] PRIMARY KEY CLUSTERED
(
	[VehicleKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
```

```
ALTER TABLE [dbo].[DimVehicle] ADD  CONSTRAINT
[FK_DimVehicle_DimCrashDetails] FOREIGN KEY([CrashDetailsKey])
REFERENCES [dbo].[DimCrashDetails] ([CrashDetailsKey])

ALTER TABLE [dbo].[DimVehicle]  ADD  CONSTRAINT
[FK_DimVehicle_DimDriverDetails] FOREIGN KEY([DriverDetailsKey])
REFERENCES [dbo].[DimDriverDetails] ([DriverDetailsKey])


--DIMCasualty
CREATE TABLE [dbo].[DimCasualty](
	[AgeBandOfCasualty] [nvarchar](100) NOT NULL,
	[CasualtyIndex] [int] NOT NULL,
	[CasualtySeverity] [nvarchar](100) NOT NULL,
	[CasualtyClass] [nvarchar](100) NOT NULL,
	[CasualtyType] [nvarchar](100) NOT NULL,
	[CasualtyKey] [bigint] IDENTITY(1,1) NOT NULL,
	[BusPassenger] [nvarchar](100) NOT NULL,
	[Gender] [nvarchar](100) NOT NULL,
	[CarPassenger] [nvarchar](100) NOT NULL,
 CONSTRAINT [PK_DimCasualty] PRIMARY KEY CLUSTERED
(
	[CasualtyKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]


--DIM_ROAD_CONDITION
CREATE TABLE [dbo].[DimRoadCondition](
	[RoadSurfaceKey] [bigint] IDENTITY(1,1) NOT NULL,
	[UrbanRuralArea] [nvarchar](100) NOT NULL,
	[SpeedLimit] [nvarchar](100) NOT NULL,
	[WeatherCondition] [nvarchar](100) NOT NULL,
	[RoadSurfaceCondition] [nvarchar](100) NOT NULL,
	[SpecialConditionAtSite] [nvarchar](100) NOT NULL,
	[SecondRoadClass] [nvarchar](100) NOT NULL,
	[RoadType] [nvarchar](100) NOT NULL,
	[FirstRoadClass] [nvarchar](100) NOT NULL,
	[LightCondition] [nvarchar](100) NOT NULL,
 CONSTRAINT [PK_DimRoadCondition] PRIMARY KEY CLUSTERED
(
	[RoadSurfaceKey] ASC
```

```
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

--Dim Date
CREATE TABLE [dbo].[DimDate](
	[PK_Date] [datetime] NOT NULL,
	[Year] [int] NULL,
	[Half_Year] [int] NULL,
	[Month_Name] [nvarchar](50) NULL,
	[Day_Of_Year] [int] NULL,
	[Month_Of_Year] [int] NULL,
	[Quarter_Of_Year] [int] NULL,
 CONSTRAINT [PK_Time] PRIMARY KEY CLUSTERED
(
	[PK_Date] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]


--DimTime
CREATE TABLE [dbo].[DimTime](
	[Minute] [bigint] NOT NULL,
	[TimeKey] [nvarchar](5) NOT NULL,
	[Hour] [bigint] NOT NULL,
	[AM/PM] [nvarchar](2) NOT NULL,
	[PartOfTheDay] [nvarchar](10) NOT NULL,
 CONSTRAINT [PK_DimTime] PRIMARY KEY CLUSTERED
(
	[TimeKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]


--FACT ACCIDENT
CREATE TABLE [dbo].[FactAccident](
	[TimeKey] [nvarchar](5) NOT NULL,
	[DateKey] [datetime] NOT NULL,
	[RoadSurfaceKey] [bigint] NOT NULL,
	[AccidentIndex] [nvarchar](100) NOT NULL,
	[SevereCasualties] [int] DEFAULT -1 NOT NULL,
```

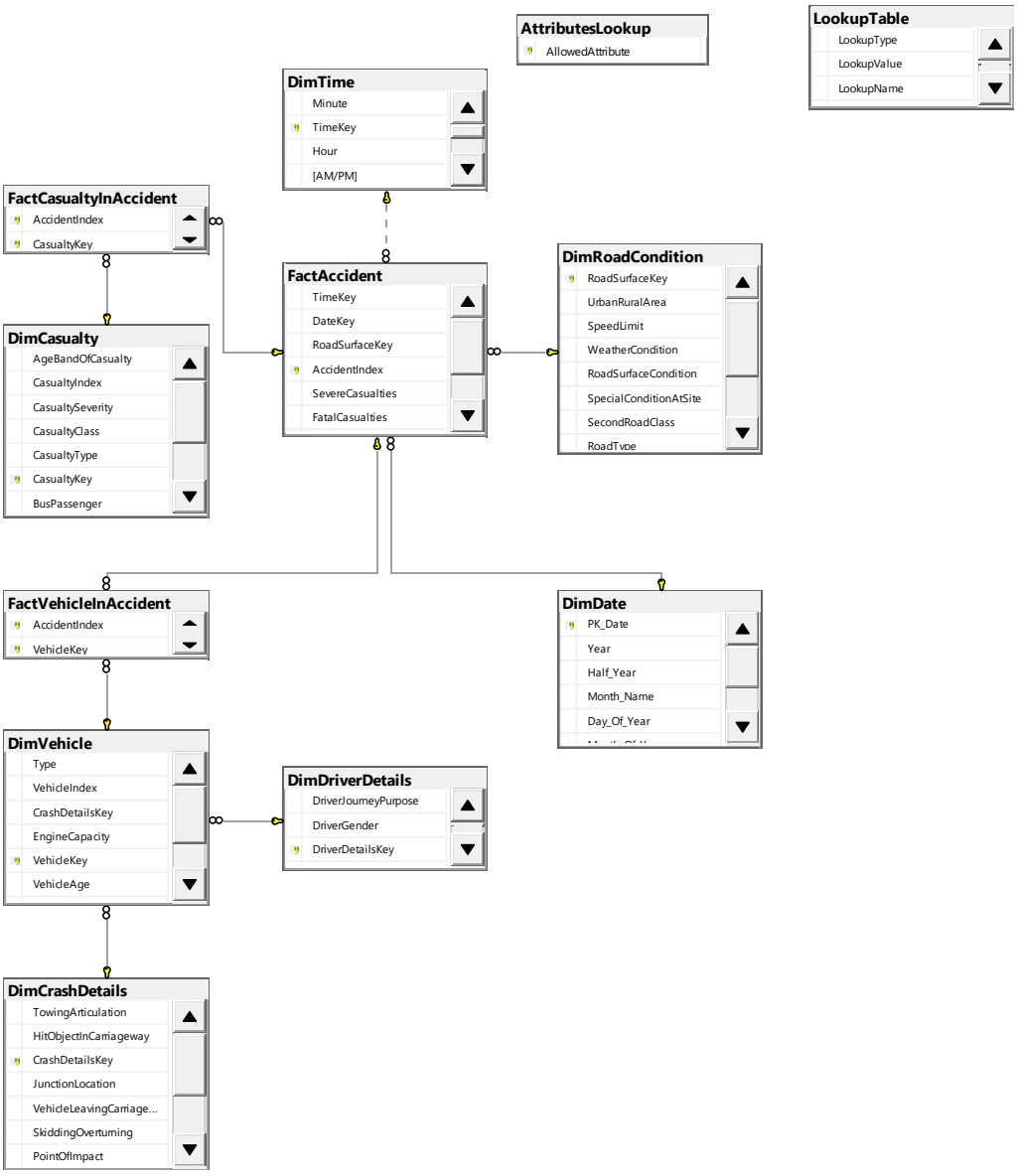
```
	[FatalCasualties] [int] DEFAULT -1 NOT NULL,
	[LightCasualties] [int] DEFAULT -1 NOT NULL,
	[VehiclesNumber] [int] DEFAULT -1 NOT NULL
 CONSTRAINT [PK__FactAcci__C031595BD633AA97] PRIMARY KEY CLUSTERED
(
	[AccidentIndex] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]


ALTER TABLE [dbo].[FactAccident]  ADD  CONSTRAINT [FK_FactAccident_DimDate]
FOREIGN KEY([DateKey])
REFERENCES [dbo].[DimDate] ([PK_Date])

ALTER TABLE [dbo].[FactAccident] ADD  CONSTRAINT
[FK_FactAccident_DimRoadCondition] FOREIGN KEY([RoadSurfaceKey])
REFERENCES [dbo].[DimRoadCondition] ([RoadSurfaceKey])

ALTER TABLE [dbo].[FactAccident] ADD CONSTRAINT [FK_FactAccident_DimTime]
FOREIGN KEY([TimeKey])
REFERENCES [dbo].[DimTime] ([TimeKey])

ALTER TABLE [dbo].[FactAccident] NOCHECK CONSTRAINT
[FK_FactAccident_DimTime];


--Fact Casualty in an accident

CREATE TABLE [dbo].[FactCasualtyInAccident](
	[AccidentIndex] [nvarchar](100) NOT NULL,
	[CasualtyKey] [bigint] NOT NULL,
 CONSTRAINT [PK_FactCasualtyInAccident] PRIMARY KEY CLUSTERED
(
	[AccidentIndex] ASC,
	[CasualtyKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

ALTER TABLE [dbo].[FactCasualtyInAccident] ADD  CONSTRAINT
[FK_FactCasualtyInAccident_DimCasualty] FOREIGN KEY([CasualtyKey])
REFERENCES [dbo].[DimCasualty] ([CasualtyKey])
```

```
ALTER TABLE [dbo].[FactCasualtyInAccident] ADD  CONSTRAINT
[FK_FactCasualtyInAccident_FactAccident] FOREIGN KEY([AccidentIndex])
REFERENCES [dbo].[FactAccident] ([AccidentIndex])


--Fact vehicle in an accident
CREATE TABLE [dbo].[FactVehicleInAccident](
	[AccidentIndex] [nvarchar](100) NOT NULL,
	[VehicleKey] [bigint] NOT NULL,
 CONSTRAINT [PK_FactVehicleInAccident] PRIMARY KEY CLUSTERED
(
	[AccidentIndex] ASC,
	[VehicleKey] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

ALTER TABLE [dbo].[FactVehicleInAccident] ADD  CONSTRAINT
[FK_FactVehicleInAccident_DimVehicle] FOREIGN KEY([VehicleKey])
REFERENCES [dbo].[DimVehicle] ([VehicleKey])

ALTER TABLE [dbo].[FactVehicleInAccident]  ADD  CONSTRAINT
[FK_FactVehicleInAccident_FactAccident] FOREIGN KEY([AccidentIndex])
REFERENCES [dbo].[FactAccident] ([AccidentIndex])


CREATE TABLE [dbo].[AttributesLookup](
	[AllowedAttribute] [nvarchar](50) NOT NULL,
 CONSTRAINT [PK_AttributesLookup] PRIMARY KEY CLUSTERED
(
	[AllowedAttribute] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]

INSERT INTO [dbo].[AttributesLookup]
	([AllowedAttribute])
    VALUES
	('AgeBand'), ('BusPassenger'), ('CarPassenger'), ('CasualtyClass'),('CasualtyType'),

('Gender'),('HitObjectInCarriageway'),('HitObjectOffCarriageway'),('JourneyPurpose'),

('JunctionLocation'),('LightCondition'),('PointOfImpact'),('RoadClass'),('RoadSurfaceConditio
n'),
```

```
('RoadType'),('Severity'),('SkiddingOverturning'),('SpecialConditionAtSite'),('SteeringWheelSide'),

('TowingArticulation'),('UrbanRuralArea'),('VehicleLeavingCarriageway'),('VehicleManoeuvre'),
                    ('VehicleType'),('WeatherCondition')

CREATE TABLE [dbo].[LookupTable](
        [LookupType] [nvarchar](50) NOT NULL,
        [LookupValue] [nvarchar](100) NULL,
        [LookupName] [nvarchar](100) NOT NULL,
        [LookupID] [bigint] IDENTITY(1,1) NOT NULL,
 CONSTRAINT [PK_LookupTable] PRIMARY KEY CLUSTERED
(
        [LookupID] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON, OPTIMIZE_FOR_SEQUENTIAL_KEY = OFF) ON [PRIMARY]
) ON [PRIMARY]
```

Tabela 6 SQL na tworzenie bazy danych

Diagram tabel w MS SQL Server Management Studio prezentuje się na Rysunek 2



Rysunek 2 Diagram bazy danych

Jak widać na Rysunek 2, oprócz tabel potrzebnych do analizy dodane zostały 2 tabele potrzebne dla procesu ETL na rzecz słownika. AttributesLookup przechowuje listę dozwolonych atrybutów do słownika, natomiast LookupTable przechowuje słownik wartości

Dodatkowo, utworzyłem 2 fakty pomostowe, ze względu na to, że w 1 wypadku może być wiele ofiar, oraz w 1 wypadku może być wiele pojazdów

Na tym etapie tworzone są też tabele tymczasowe, używane przy ładowaniu danych oraz pomostów

```
--TEMP TABLES

CREATE TABLE [dbo].[Temp_Accidents](
        [Accident_Index] [nvarchar](100) NULL,
        [Date] [nvarchar](100) NULL,
        [Time] [nvarchar](5) NULL,
        [Speed_limit] [nvarchar](100) NULL,
        [Road_Type] [nvarchar](100) NULL,
        [2nd_Road_Class] [nvarchar](100) NULL,
        [Light_Conditions] [nvarchar](100) NULL,
        [Weather_Conditions] [nvarchar](100) NULL,
        [Road_Surface_Conditions] [nvarchar](100) NULL,
        [Special_Conditions_at_Site] [nvarchar](100) NULL,
        [Urban_or_Rural_Area] [nvarchar](100) NULL,
        [1st_Road_Class] [nvarchar](100) NULL,
   [Number_Of_Vehicles] [int] NULL,
   [Number_Of_Casualties] [int] NULL
) ON [PRIMARY]


CREATE TABLE [Temp_Casualty] (
   [Accident_Index] nvarchar(100),
   [Casualty_Index] [int],
   [Casualty_Class] [nvarchar](100),
   [Sex_of_Casualty] [nvarchar](100),
   [Age_of_Casualty] [nvarchar](100),
   [Casualty_Severity] [nvarchar](100),
   [Car_Passenger] [nvarchar](100),
   [Bus_or_Coach_Passenger] [nvarchar](100),
   [Casualty_Type] [nvarchar](100)
)


CREATE TABLE [Temp_Vehicle] (
   [Accident_Index] nvarchar(100),
   [Vehicle_Type] [nvarchar](100),
   [Vehicle_Index] [int],
   [Towing_and_Articulation] [nvarchar](100),
   [Vehicle_Manoeuvre] [nvarchar](100),
   [Junction_Location] [nvarchar](100),
   [Skidding_and_Overturning] [nvarchar](100),
   [Hit_Object_in_Carriageway] [nvarchar](100),
   [Vehicle_Leaving_Carriageway] [nvarchar](100),
```

```
    [Hit_Object_off_Carriageway] [nvarchar](100),
    [1st_Point_of_Impact] [nvarchar](100),
    [Was_Vehicle_Left_Hand_Drive?] [nvarchar](100),
    [Journey_Purpose_of_Driver] [nvarchar](100),
    [Sex_of_Driver] [nvarchar](100),
    [Age_Band_of_Driver] [nvarchar](100),
    [Engine_Capacity_(CC)] [nvarchar](100),
    [Age_of_Vehicle] [int]
)

CREATE TABLE [Temp_FactCasualtyInAccident](
    [Accident_Index] nvarchar(100),
    [CasualtyID] numeric(20,0),
)


CREATE TABLE [Temp_FactVehicleInAccident] (
    [Accident_Index] nvarchar(100),
    [VehicleID] numeric(20,0)
)
```

Rysunek 3 Skrypt tworzący tabele tymczasowe

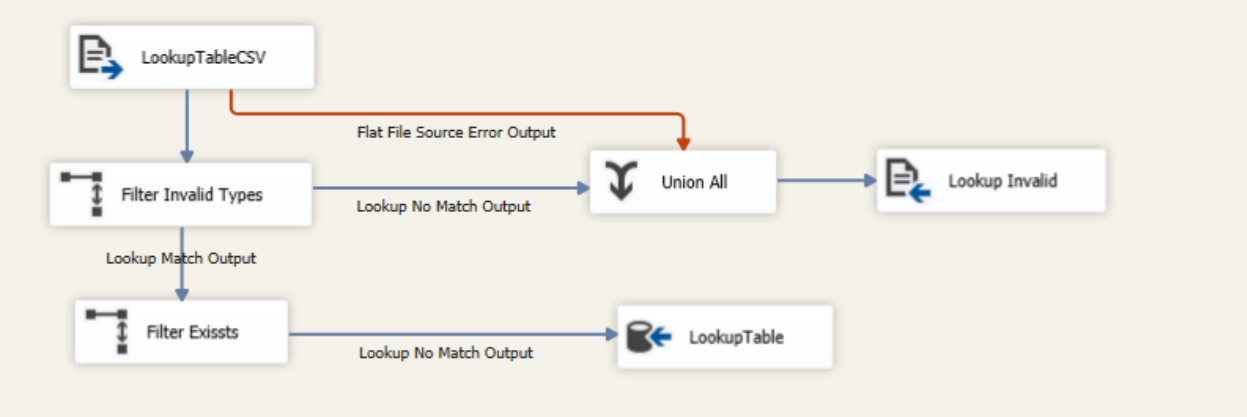
Tabele tymczasowe są czyszczone na początku i po zakończeniu całego procesu ETL, więc nie trzeba ich tworzyć od nowa za każdym wywołaniem procesu

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

Cały proces ETL został przeze mnie podzielony na drzewo, w którym:

1. Na samym początku ładowane są dane do tabeli słownikowej
2. Oddzielne gałęzie ładują dane do tabel tymczasowych z tłumaczeniem atrybutów
3. Generowane są wymiary charakterystyczne dla konkretnego pliku
4. Dane są przenoszone do bazy
5. Dane są wiązane faktami, oraz przeliczane są miary.

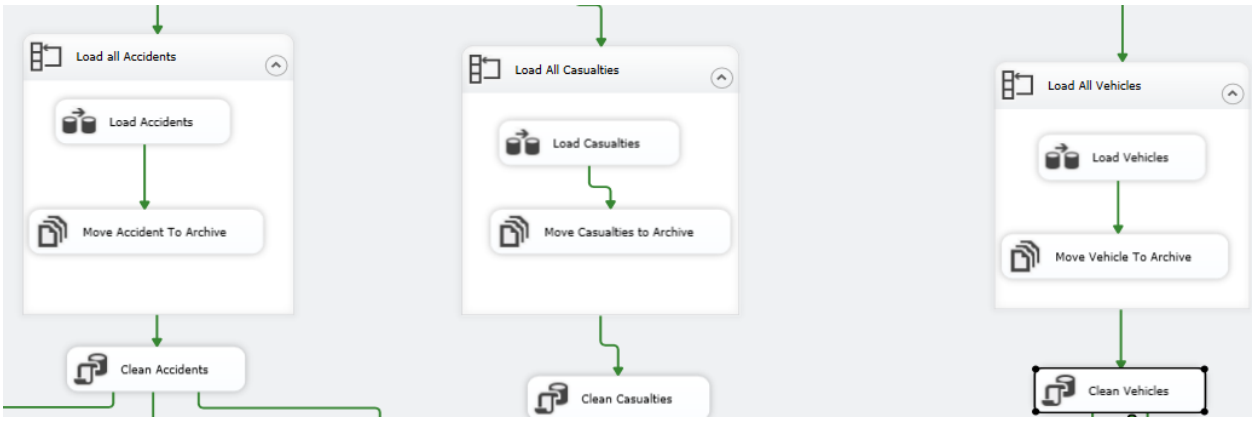
1. Ładowanie danych do tabeli słownikowej



Rysunek 4 Ładowanie danych do słownika

Rysunek 4 przedstawia proces przepływu danych do tabeli słownikowej. ETL importuje dane z pliku oraz filtruje dozwolone parametry. Następnie filtruje już istniejące wpisy i dodaje je do bazy danych.

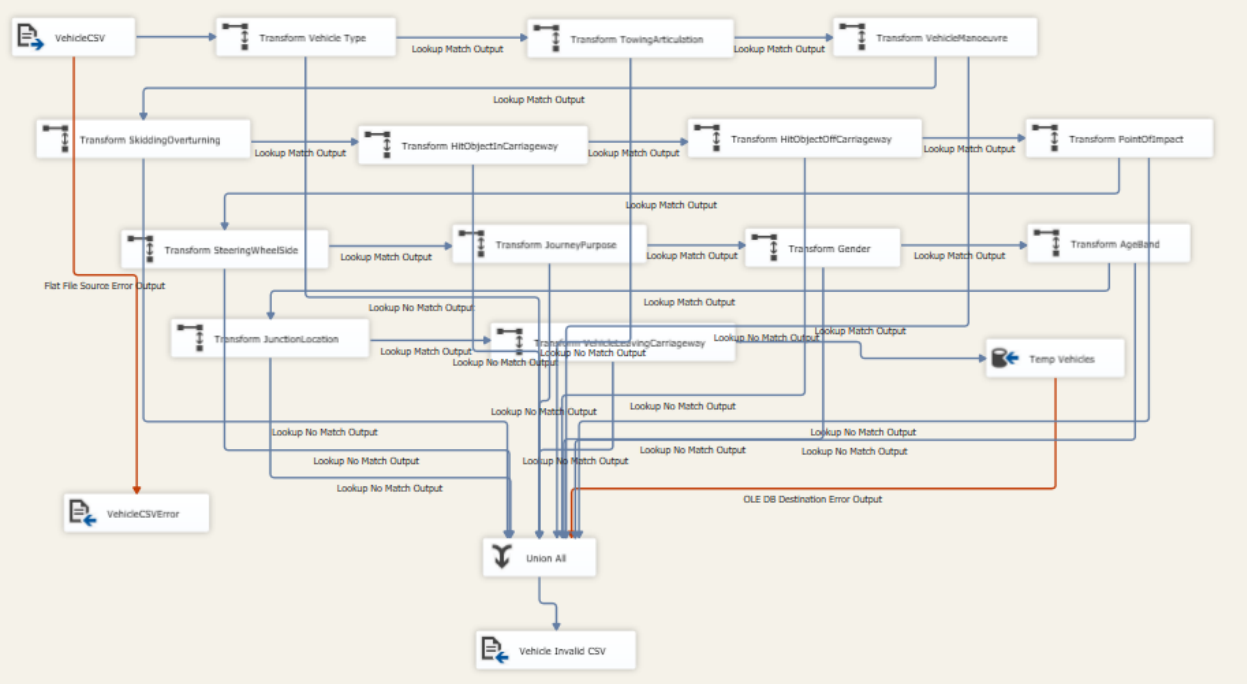
2. Ładowanie danych do tabeli tymczasowych oraz tłumaczenie atrybutów



Rysunek 5 Ładowanie danych to tabel tymczasowych i translacja słownika

Na tym etapie drzewo rozgałęzia się na 3 równoległe procesy, które ładują wszystkie pliki z folderu odpowiadającemu typowi danych. Następnie dane są ładowane, tłumaczone oraz dodawane do tabel tymczasowych. Po zakończeniu ładowania z plików tabele tymczasowe zostają wyczyszczone z danych, które zostały już dodane do baz docelowych.

Na przykładzie Load Vehicles pokaże proces ładowania danych do tabel tymczasowych



Rysunek 6 Ładowanie i translacja danych Vehicle

Na rysunku widać, że dane są ładowane z pliku oraz przechodzą szereg tłumaczeń przy pomocy lookupa z tabeli słownika. Gdziekolwiek system nie znajdzie powiązania, tam zwraca dany wiersz jako błąd i przekierowuje go do pliku z błędnymi danymi. Finalnie, dane są dodawane do tymczasowej tabeli Temp_Vehicles. Proces wygląda analogicznie dla Accidents i dla Casualities

Czyszczenie robione jest skryptem SQL

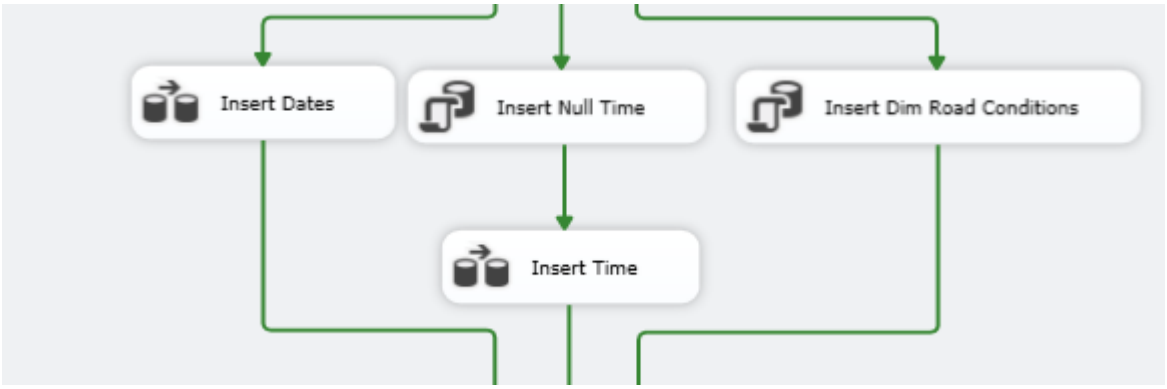
```
WITH JoinedFactVehicle AS (


        SELECT

                fact.AccidentIndex,

                VehicleIndex

        FROM dbo.FactVehicleInAccident fact

        JOIN dbo.DimVehicle dim ON dim.VehicleKey = fact.VehicleKey


)
```

```
DELETE dbo.[Temp_Vehicle]
FROM dbo.[Temp_Vehicle] temp
INNER JOIN JoinedFactVehicle fact
        ON temp.Accident_Index = fact.AccidentIndex
        AND temp.Vehicle_Index = fact.VehicleIndex
```

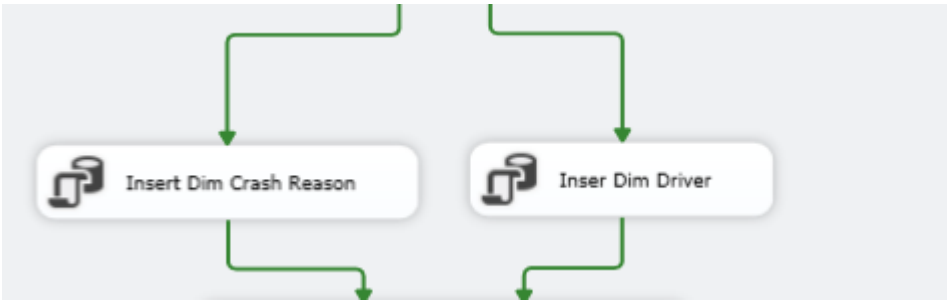
Rysunek 7 Czyszczenie tabeli tymczasowej z duplikatów

Z racji, że dane są animizowane, jedyne atrybuty, po którym można określić duplikat, to Accident_Index i Vehicle_Index, dlatego na podstawie tych dwóch atrybutów czyszczę tabele

3. Generowanie wymiarów



Rysunek 8 Generowanie wymiarów z pliku Accidents



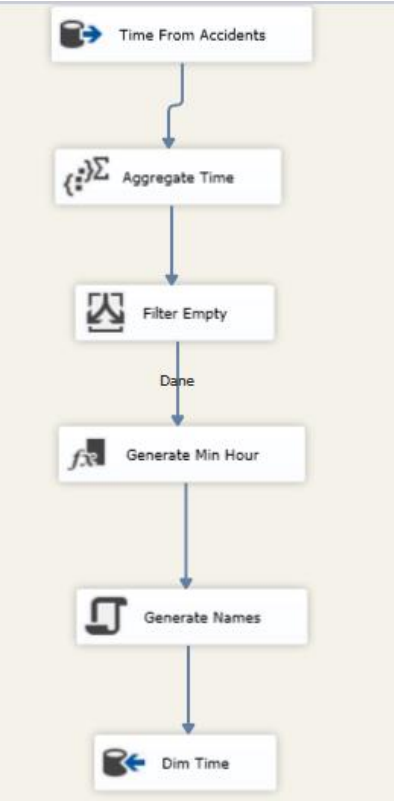
Rysunek 9 Generowanie wymiarów z pliku Vehicle

Generowanie prostych wymiarów polega na wywołaniu zapytania SQL, które grupuje potrzebne atrybuty z tabeli tymczasowej, następnie sprawdza, czy dany wymiar już istnieje. Jeżeli nie, to tworzy nowy.

```
INSERT INTO dbo.DimDriverDetails(DriverAgeBand, DriverGender, DriverJourneyPurpose)
SELECT Age_Band_of_Driver, Sex_of_Driver, Journey_Purpose_of_Driver
FROM dbo.Temp_Vehicle veh
LEFT JOIN [dbo].DimDriverDetails driver
        ON driver.DriverAgeBand = veh.Age_Band_of_Driver
        AND driver.DriverGender = veh.Sex_of_Driver
        AND driver.DriverJourneyPurpose = veh.Journey_Purpose_of_Driver
WHERE driver.DriverDetailsKey is Null
GROUP BY
                [Sex_of_Driver]
                ,[Journey_Purpose_of_Driver]
                ,[Age_Band_of_Driver]
```

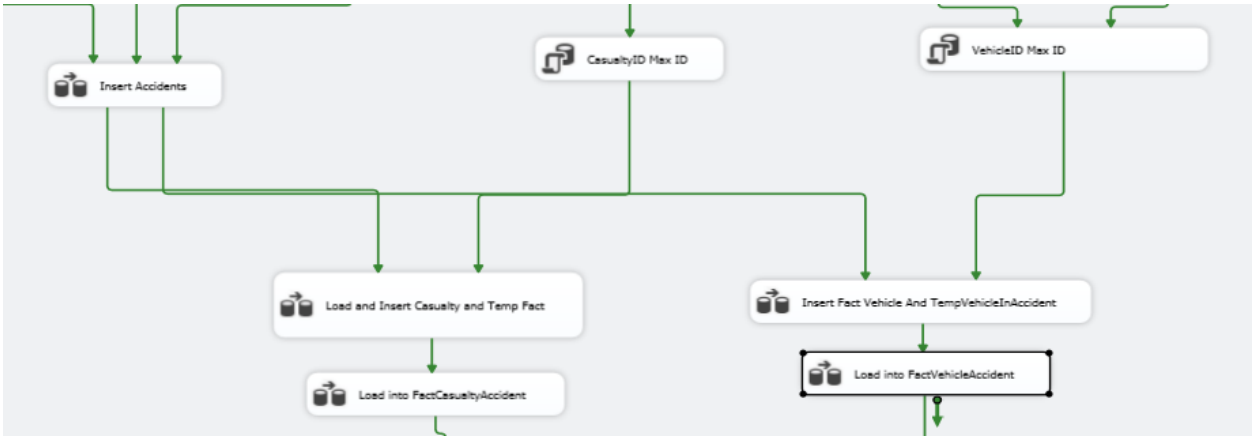
Rysunek 10 Tworzenie wymiaru DimDriverDetails

Natomiast niektóre wymiary wymagają dodatkowego wygenerowania danych lub przeliczania. W tym celu dodaje dodatkowe kroki w data flow, które grupują, filtrują, Generują atrybuty i dodają nazwy



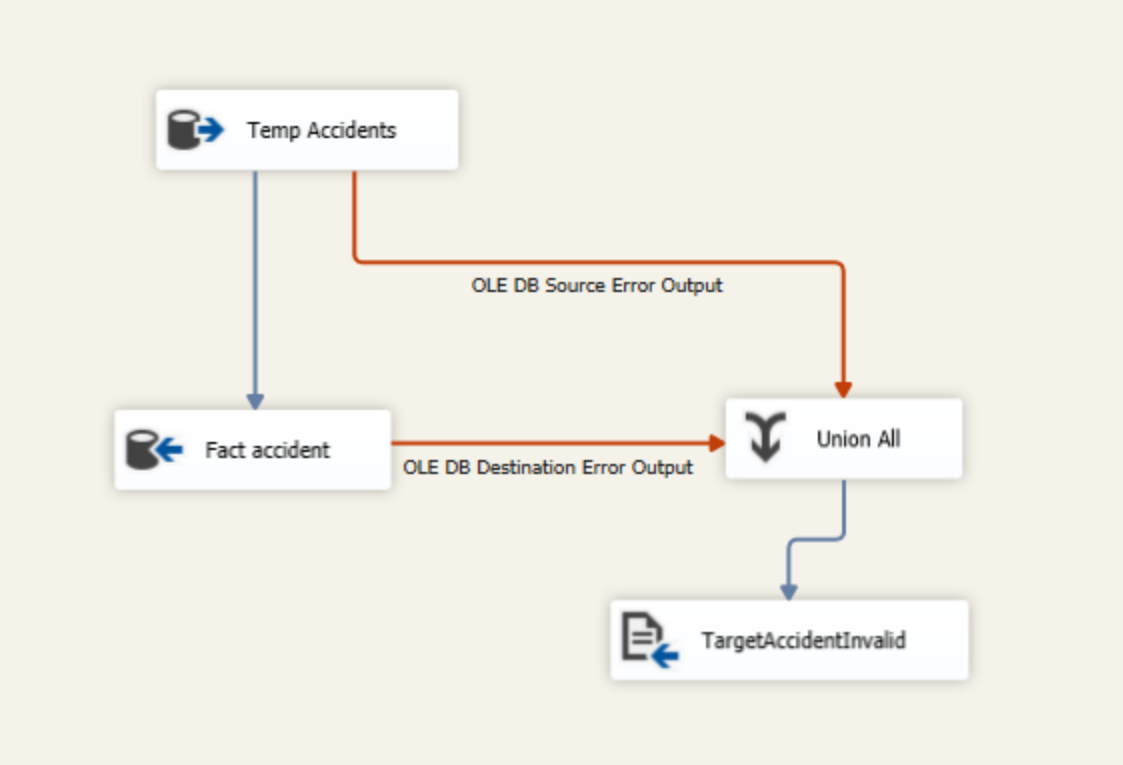
Rysunek 11 Tworzenie wymiaru DimTime

4. **Przenoszenie danych do tabel końcowych oraz ich wiązanie**



Rysunek 12 Ładowanie danych z tabel tymczasowych do tabel docelowych

Accidents, po wygenerowaniu wymiarów, jest gotowy do załadowania, dlatego przy pomocy zapytania SQL pobierane są z tabeli tymczasowej przygotowane dane, zawierające już referencje na odpowiednie wymiary.



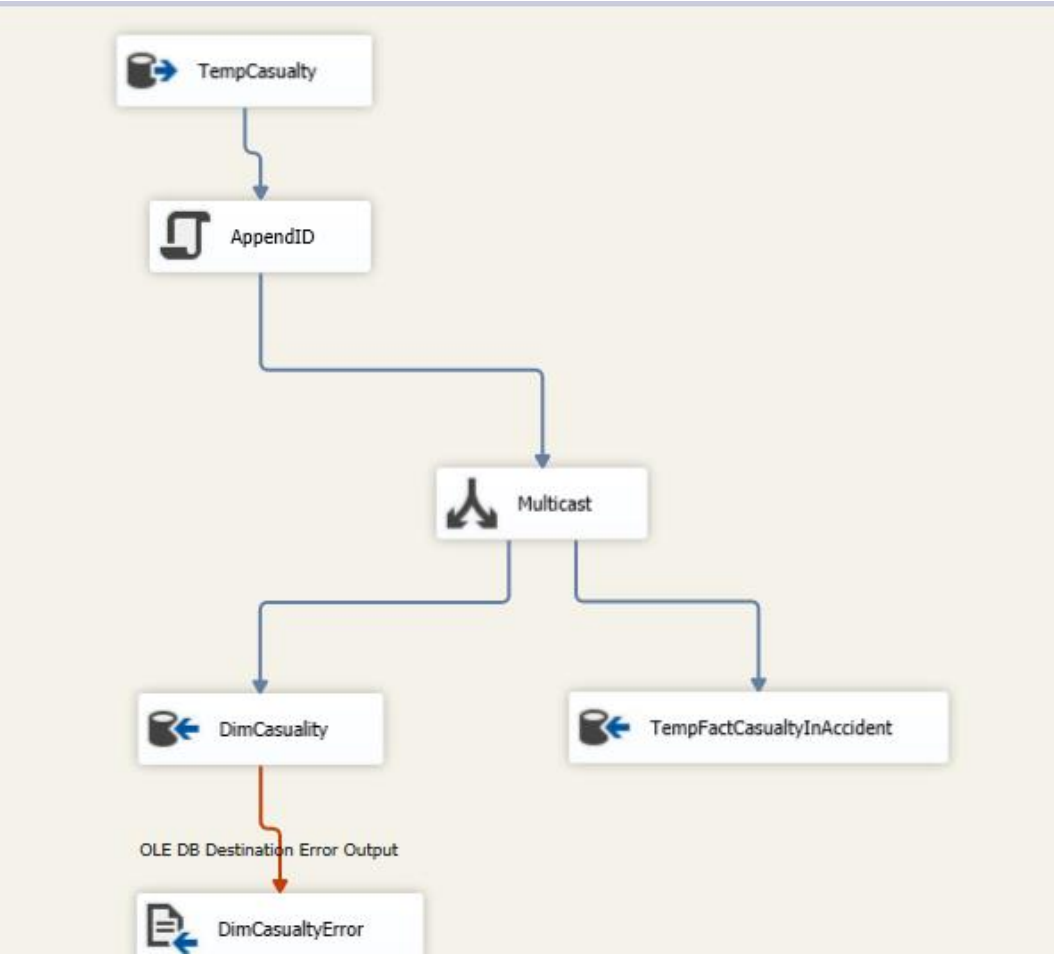
Rysunek 13 Insert Accidents Data Flow

```
SELECT Accident_Index, [Date], [Time], [dim].[RoadSurfaceKey],
temp.[Number_Of_Vehicles], temp.[Number_Of_Casualties]

FROM dbo.Temp_Accidents temp

JOIN dbo.DimRoadCondition dim

        ON dim.FirstRoadClass = temp.[1st_Road_Class]

        AND dim.SecondRoadClass = temp.[2nd_Road_Class]

        AND dim.LightCondition = temp.[Light_Conditions]

        AND dim.RoadSurfaceCondition = temp.[Road_Surface_Conditions]

        AND dim.RoadType = temp.[Road_Type]

        AND dim.SpecialConditionAtSite = temp.Special_Conditions_at_Site

        AND dim.SpeedLimit = temp.Speed_limit

        AND dim.UrbanRuralArea = temp.Urban_or_Rural_Area
```

```
        AND dim.WeatherCondition = temp.Weather_Conditions
```

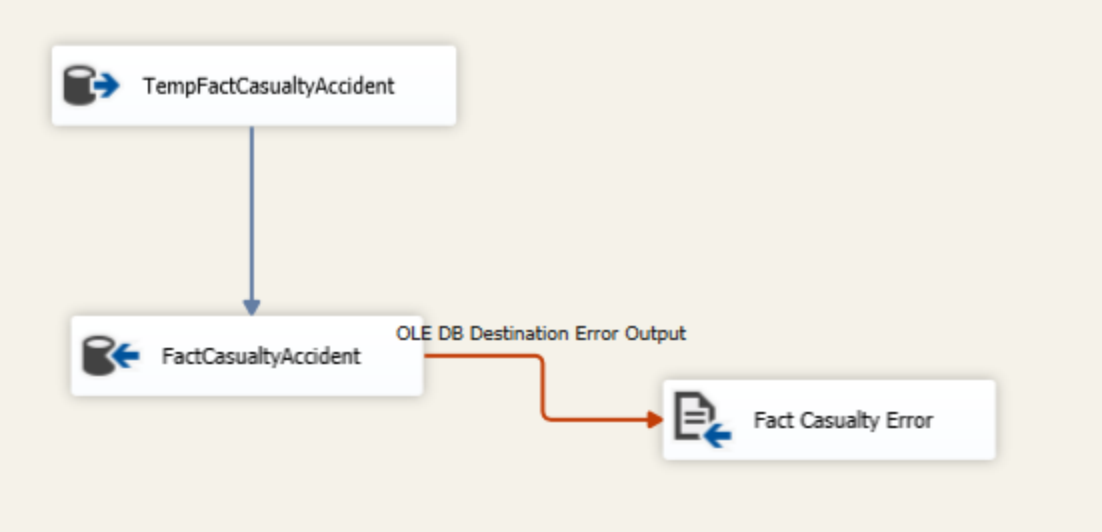
Rysunek 14 Zapytanie pobierające dane dane z tabeli tymczasoej oraz łączący je z wymiarami

Natomiast Casualty i Vehicle potrzebuje jeszcze jednego kroku – każdy pojazd i ofiara musi mieć wygenerowane ID, na podstawie którego zostają powiązane przy pomocy mostu z wypadkiem. W tym celu przy pomocy ExecuteSQL pobierane są maksymalne wartości identyfikatorów ofiary i wypadku, oraz przy wstawianiu danych do tabeli generowany jest identyfikator, którego powiązanie z wypadkiem wstawiane jest do tabeli tymczasowej. Użycie tabeli tymczasowej w tym miejscu spowodowane jest tym, że zrównolegnione zadania wprowadzania danych to tabel może wywołać błąd powiązania w tabeli bridge, więc dlatego jest zrobiony bufor, który ma za zadania te powiązania zapisać



Rysunek 15 Ładowanie danych do DimCasualty

Na samym końcu dane są ładowane to tabeli mostu



Rysunek 16 Ładowanie danych do tabeli mostów FactCasualtyAccident

5. Generowanie miar

```
WITH NullAccidents as
(
	SELECT AccidentIndex
	FROM dbo.FactAccident
	WHERE [SevereCasualties] = -1 OR [FatalCasualties] = -1 OR [LightCasualties] = -1
),
JoinedNumbers(AccidentIndex, FatalCasualties, SevereCasualties, LightCasualties) AS
(
	SELECT
		nullable.AccidentIndex,
		SUM(IIF(dim.CasualtySeverity = 'Fatal', 1,  0)),
		SUM(IIF(dim.CasualtySeverity = 'Serious', 1, 0)),
		SUM(IIF(dim.CasualtySeverity = 'Slight', 1, 0))
	FROM NullAccidents nullable
```

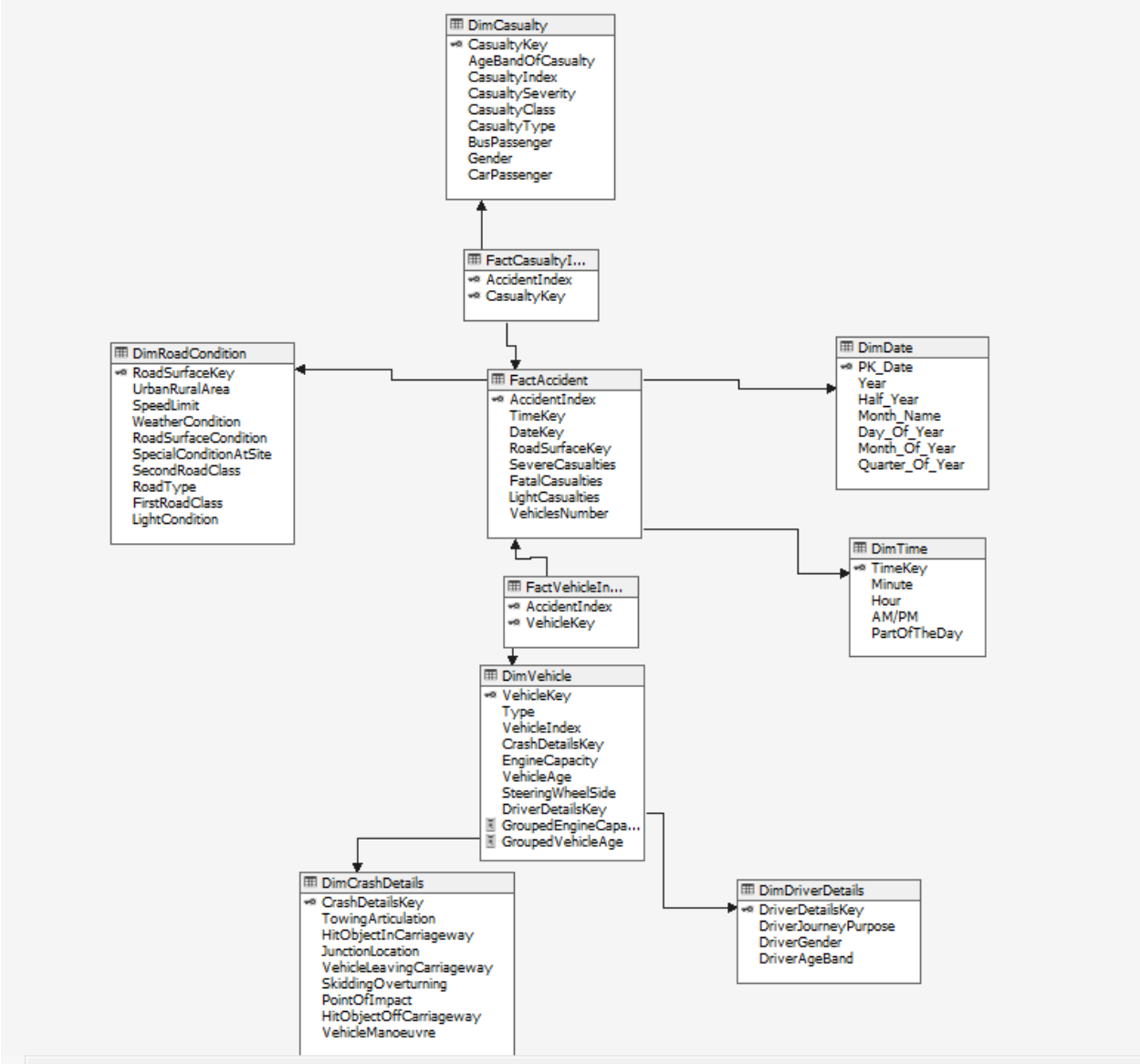
```
        FULL JOIN dbo.FactCasualtyInAccident fact
        ON fact.AccidentIndex = nullable.AccidentIndex
        FULL JOIN dbo.DimCasualty dim
        ON fact.CasualtyKey = dim.CasualtyKey
        GROUP BY nullable.AccidentIndex, dim.CasualtySeverity


)

UPDATE
  dbo.FactAccident
SET
        [FatalCasualties] = t2.FatalCasualties,
        [SeverecCasualties] = t2.SevereCasualties,
        [LightCasualties] = t2.LightCasualties
FROM
   dbo.FactAccident acc
   JOIN
   JoinedNumbers t2 ON t2.AccidentIndex = acc.AccidentIndex;
```

Rysunek 17 Generowanie miar faktów

Na samym końcu generowane są miary dla fakty wypadku, przy pomocy SQLa, oraz czyszczone są tabele tymczasowe.

7. Implementacja modeli wielowymiarowych

7.1. Widok danych



Rysunek 18 Widok danych Accidents View

Do tabeli DimVehicle zostały dodane atrybuty przeliczalne. Jeden zaokrągla pojemności silników do okrągłych wartości, drugi grupuje wiek pojazdów

```
IIF(
        [EngineCapacity] < 10,
        [EngineCapacity],
        IIF(
            [EngineCapacity]<1000,
```

```
                [EngineCapacity]/10*10,
                IIF(
                        [EngineCapacity] < 10000,
                        [EngineCapacity]/100*100,
                        [EngineCapacity]/1000*1000
                        )
                )
        )
```

Rysunek 19 Wyrażenie zaokrąglające pojemność silnika

```
(IIF(VehicleAge = -1,
        '-1',
        IIF(VehicleAge < 3,
                '0-2',
                IIF(VehicleAge < 7,
                        '3-6',
                        IIF(VehicleAge < 11,
                                '7-10',
                                IIF(VehicleAge < 15,
                                        '11-14',
                                        IIF(VehicleAge < 20,
                                                '15-19',
                                                '20+'
                                                )
                                        )
```

```
                                        )

                                    )

                                )

                            ))
```

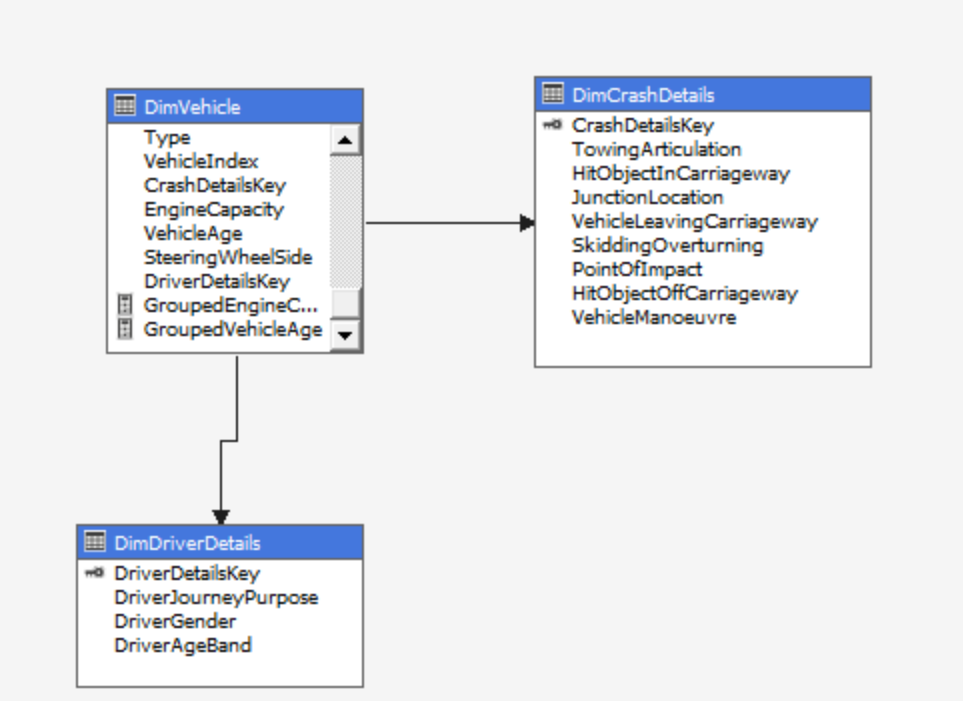
Rysunek 20 Wyrażenie grupujące wiek pojazdu

7.2. Wymiary

Rysunek 21 Wymiary zdefiniowane w modelu

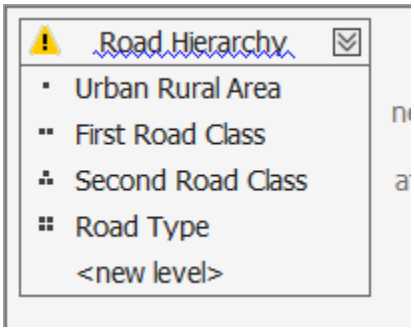
Wymiar DimFact Accident jest wymiarem potrzebnym na potrzeby mostu łączącego Ofiarę wypadku i Pojazd w wypadku z wypadkiem

Vehicle zawiera w sobie 3 tabele



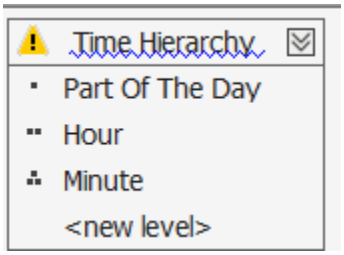
Rysunek 22 Tabele w wymiarze DimVehicle

Zdefiniowano następujące hierarchię:



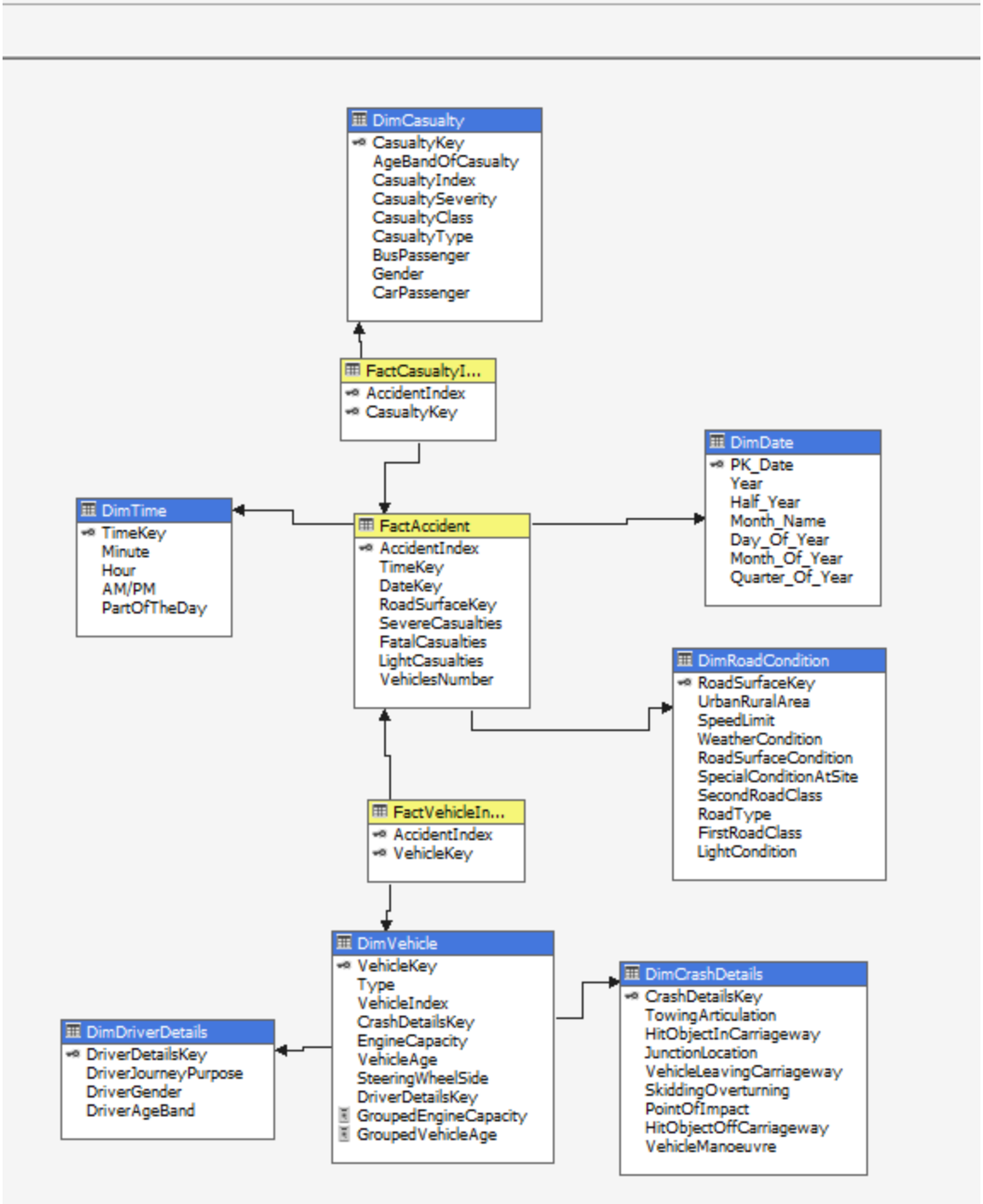
Rysunek 23 Hierarchia typów drogi w wymiarze Road Conditions

Rysunek 24 Hierarchia dat w wymiarze Date

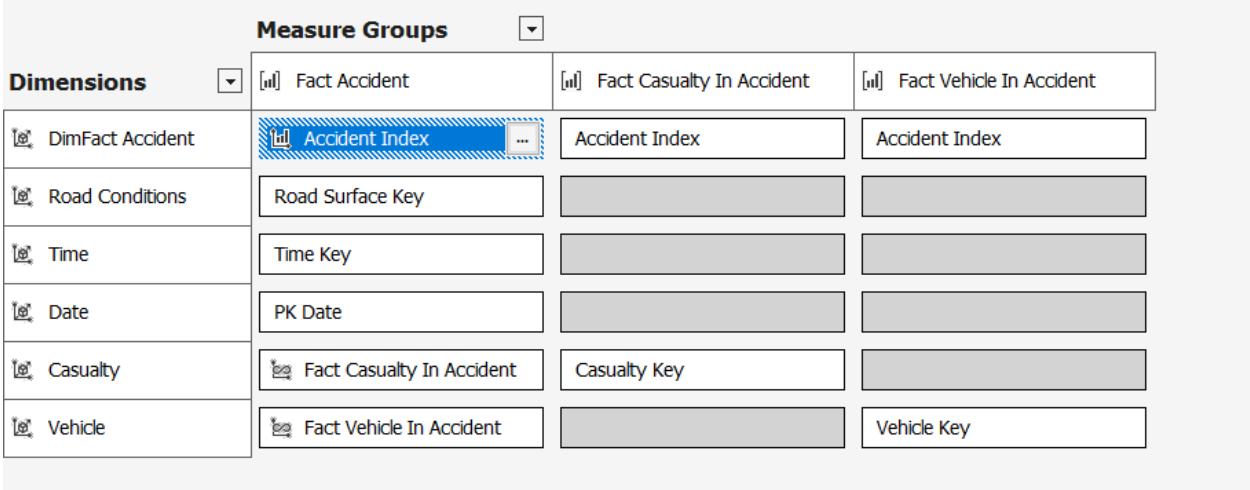


Rysunek 25 Hierarchia czasu w wymiarze Time

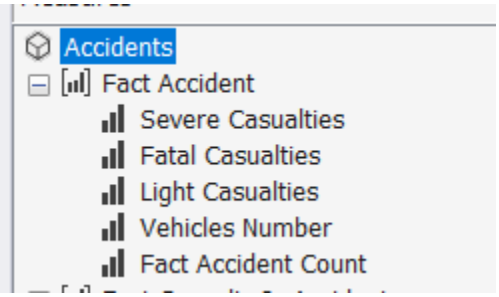
7.3. Modele wielowymiarowe – Kostki



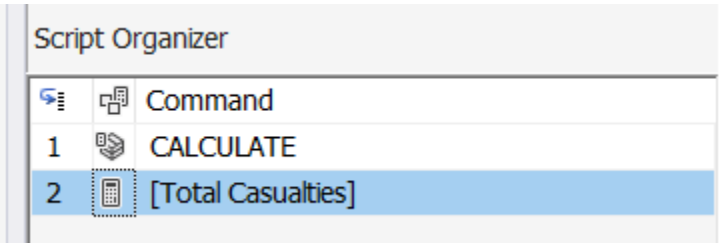
Rysunek 26 Widok kostki Accidents



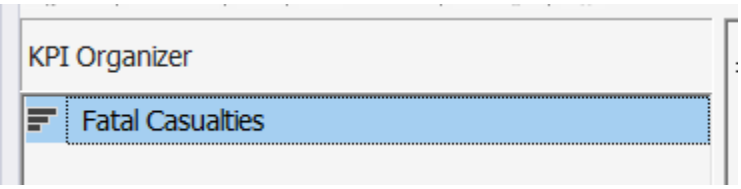
Rysunek 27 Powiązania wymiarów oraz pomostów



Rysunek 28 Miary wygenerowane na etapie tworzenia kostki



Rysunek 29 Miary kalkulowane, wygenerowane w kostce

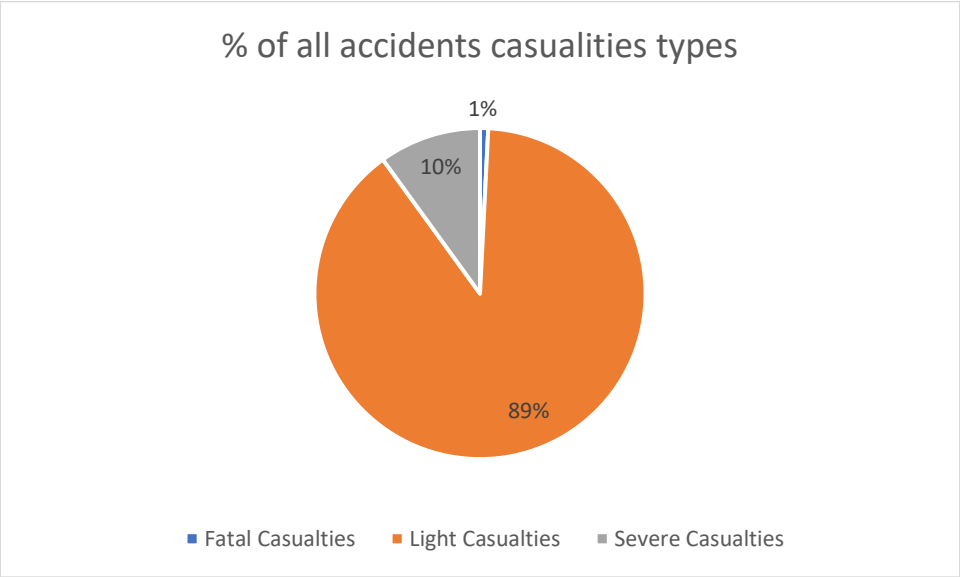


Rysunek 30 Utworzone KPI, analizujący liczbę śmiertelnych wypadków

8. Analiza danych

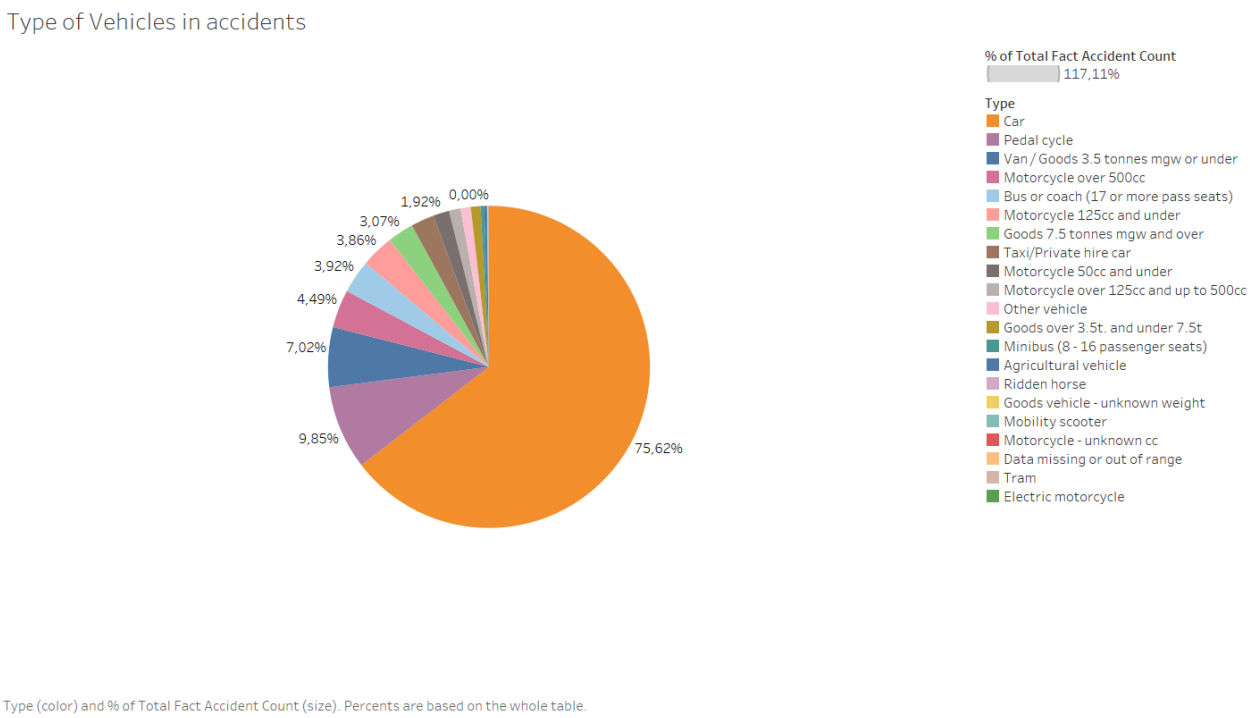
8.1. Realizacja procesów analitycznych

Proces analityczny wykonuje za pomocą Excela oraz Tablau Desktop



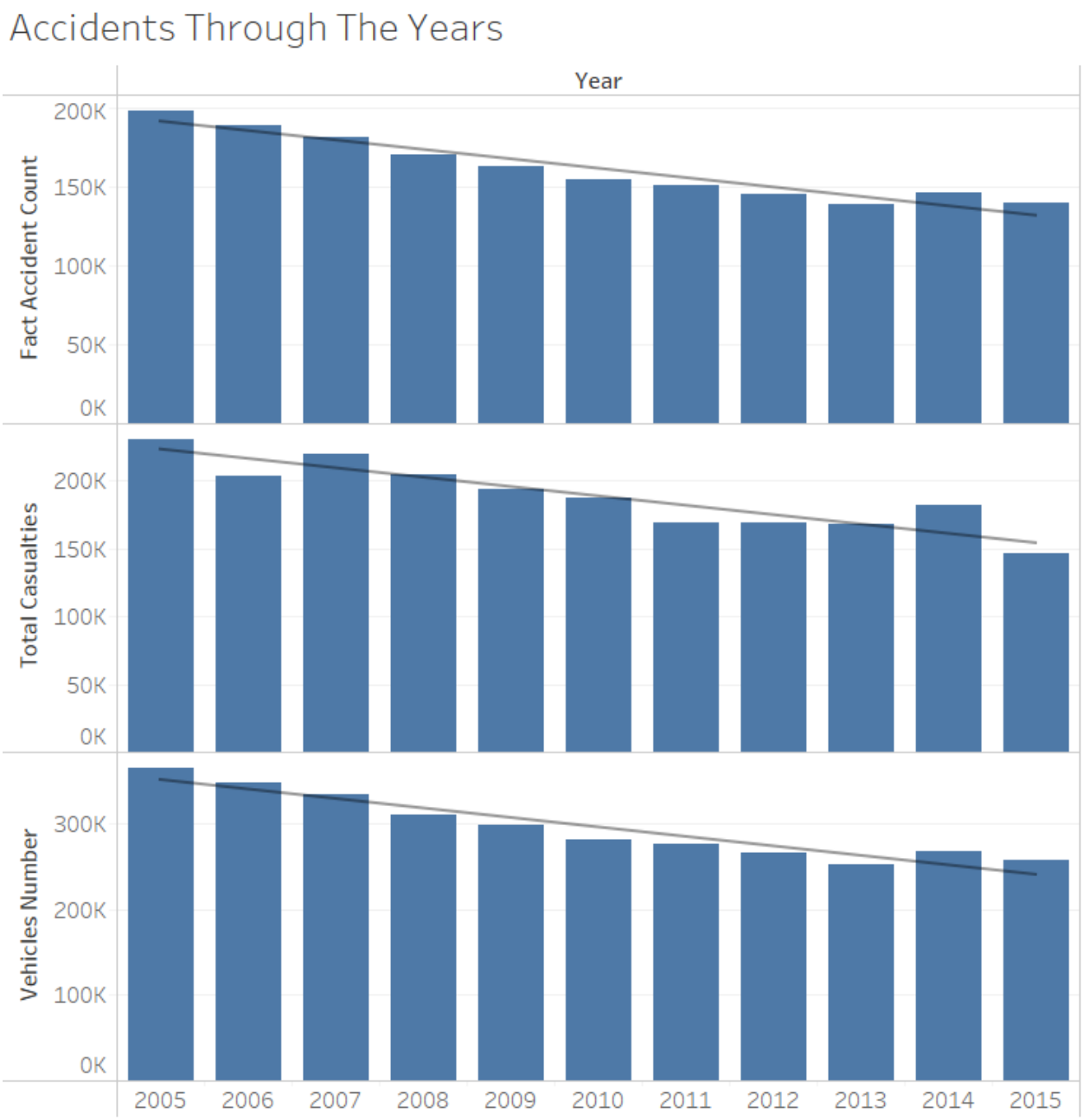
Rysunek 31 Udział typów ofiar w danych

Na wstępnie trzeba zaznaczyć, że zalewie 1% wszystkich wypadków w danych stanowią wypadki śmiertelne. 10% stanowią wypadki o poważnych konsekwencjach, natomiast większość, bo aż 90% wypadków, stanowią wypadki o lekkich obrażeniach.



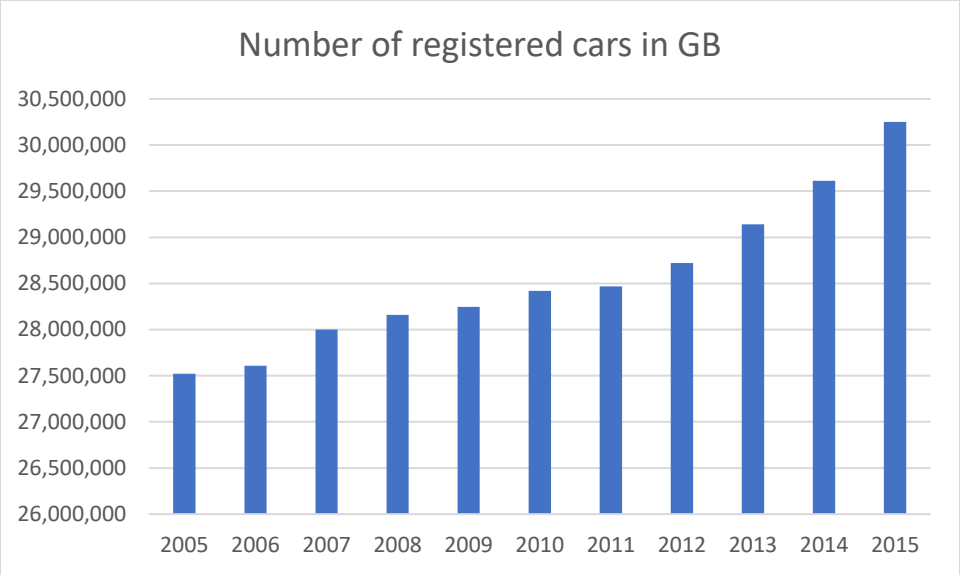
Rysunek 32 Typy pojazdów uczestniczące w wypadku

Ogólnie, 75% pojazdów uczestniczących w wypadku to samochody, 10% to rowery, 7% to ciężarówki. Zgadza się to ze źródłem, które mówi, że najwięcej zarejestrowanych jest aut osobowych.

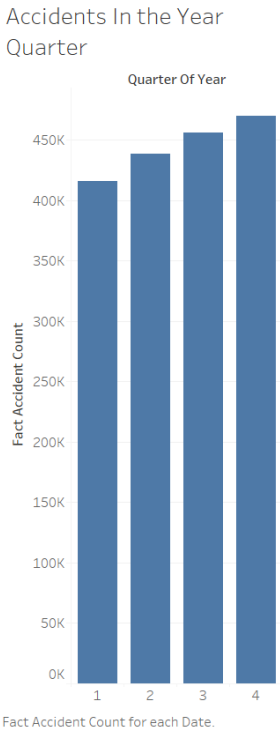


Rysunek 33 Accidents Through The Years

Na wykresie przedstawiłem podstawową metrykę – jak zmieniała się liczba wypadków, liczba ofiar śmiertelnych oraz liczba pojazdów uczestniczących w wypadku. Na podstawie tych danych można określić, że trend jest zasadniczo malejący i z roku na rok jest coraz mniej wypadków w UK. Na podstawie źródła wynika, że liczba aut w latach 2005 – 2015 wzrosła, co przestawia Rysunek 34. Można dzięki temu założyć, że rzeczywiście, trend jest malejący



Rysunek 34 Liczba aut zarejestrowanych w UK w latach 2005 -2015



Rysunek 35 Wypadki w kwartałach lat

Jeżeli chodzi o podział roku na kwartały, definitywnie najwięcej wypadków jest w ostatnim kwartale roku. Mogą być 2 przyczyny takiego trendu. Pierwsze to zwiększony ruch w okresie

zimowym ze względu na święta, drugi to pogorszone warunku atmosferyczne w tym okresie. Skupimy się na drugiej przyczynie

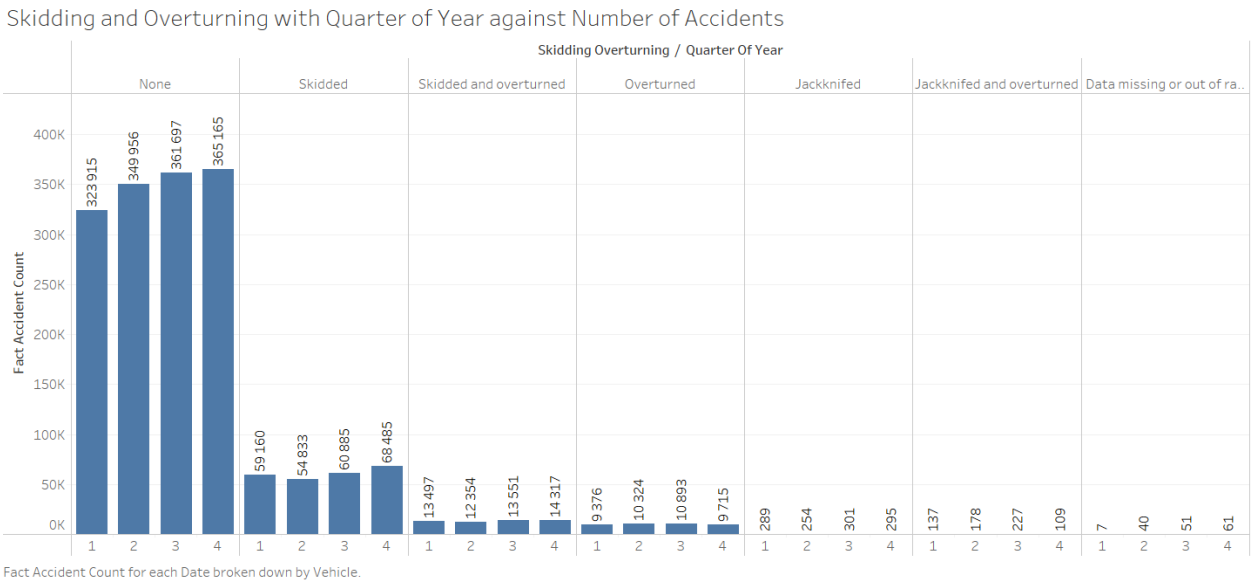
Weather Condition and Road Surface Condition with Accidents

Quarter ..	Weather Condition	Road Surface Condition: Data mis..	Dry	Flood ove..	Frost or i..	Snow	Wet or da..
1	Data missing or out ..	2	23		3		5
	Fine + high winds	3	5 218	16	203	39	3 745
	Fine no high winds	124	235 506	84	11 074	1 003	71 221
	Fog or mist	3	465		561	21	1 955
	Other	19	1 914	6	4 226	328	7 085
	Raining + high winds		47	206	63	25	7 672
	Raining no high winds	3	297	232	414	97	43 737
	Snowing + high winds		11	5	264	1 088	294
	Snowing no high win..	1	62	6	1 322	4 843	2 054
	Unknown	442	6 291	5	234	60	1 652
2	Data missing or out ..	3	39				
	Fine + high winds	2	2 774		15	1	653
	Fine no high winds	111	352 495	51	143	28	24 493
	Fog or mist		234	2	6	1	482
	Other	8	1 498	8	66	14	2 679
	Raining + high winds	1	32	114	8	3	2 618
	Raining no high winds		359	261	18	26	41 398
	Snowing + high winds		6		8	43	32
	Snowing no high win..		30		24	137	176
	Unknown	370	6 431	7	5	4	768
3	Data missing or out ..	5	37				4
	Fine + high winds	4	1 847	3	7		614
	Fine no high winds	114	355 603	59	30	19	29 666
	Fog or mist		270	2	1		525
	Other	17	1 460	4	3	1	3 393
	Raining + high winds		40	117	1	3	2 850
	Raining no high winds	4	392	395	11	32	50 680
	Snowing + high winds		1		2	2	13
	Snowing no high win..		21	1	3	2	55
	Unknown	410	6 231	7	1	2	787
4	Data missing or out ..	4	25				11
	Fine + high winds	4	3 902	21	172	17	4 051
	Fine no high winds	147	232 546	105	10 261	765	97 384
	Fog or mist	3	507	3	791	22	3 843
	Other	38	1 890	25	4 222	281	9 974
	Raining + high winds	2	72	372	70	15	11 521
	Raining no high winds	9	375	446	479	94	70 716
	Snowing + high winds		7	3	119	320	90
	Snowing no high win..		23	7	948	2 095	590
	Unknown	582	7 307	11	211	44	2 245

Fact Accident Count: 1 – 355 603

Fact Accident Count broken down by Road Conditions vs. Date and Road Conditions. Color shows Fact Accident Count. The marks are labeled by Fact Accident Count.

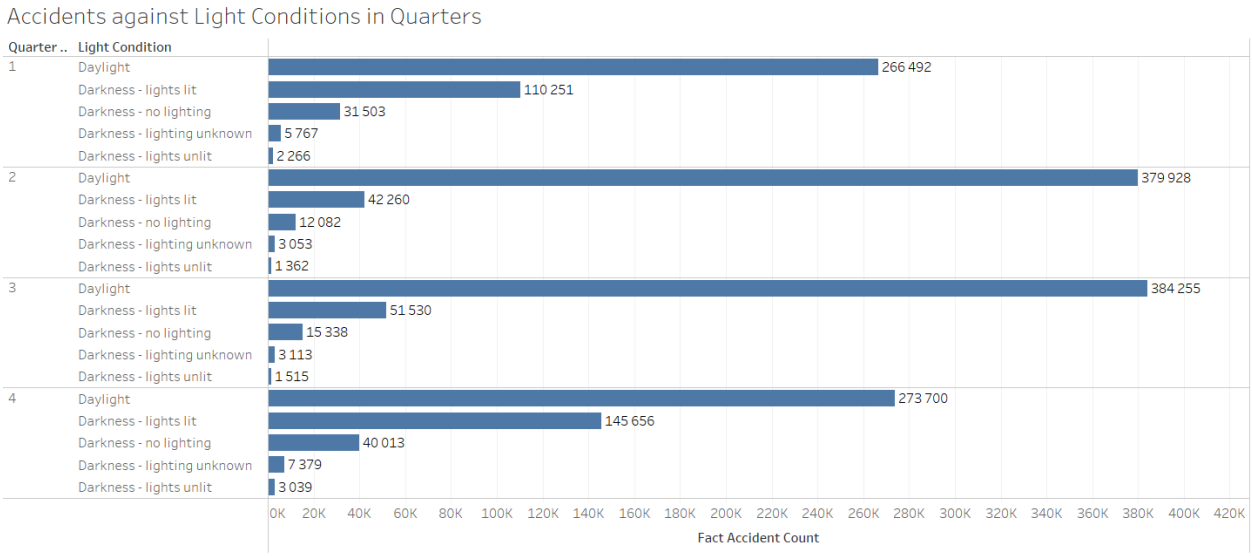
Rysunek 36 Liczba wypadków w zależności od pogody i warunków na drodze, podzielone na kwartały



Fact Accident Count for each Date broken down by Vehicle.

Rysunek 37 Poślizg i wywrócenie się podzielone na kwartały

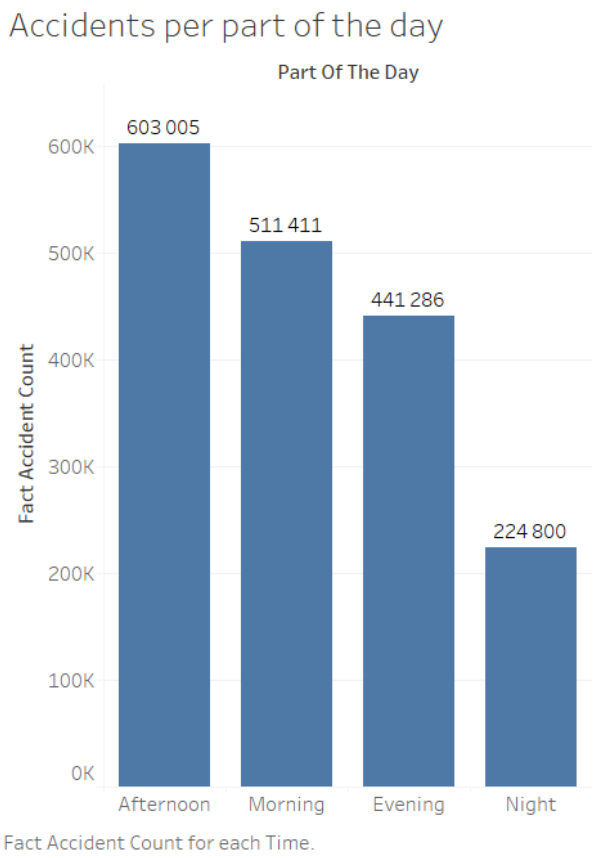
Dwa powyższe rysunki pokazują, że większość wypadków w tych okresach nie odbywa się poprzez wpadnięcie w poślizg lub wywrócenie się, natomiast widać, że wartość ta jest większa w okresach zimowych. Zimowa pogoda sprzyja mokrej lub zamrożonej powierzchni siłą rzeczy te wartości będą wyższe



Fact Accident Count for each Road Conditions broken down by Date.

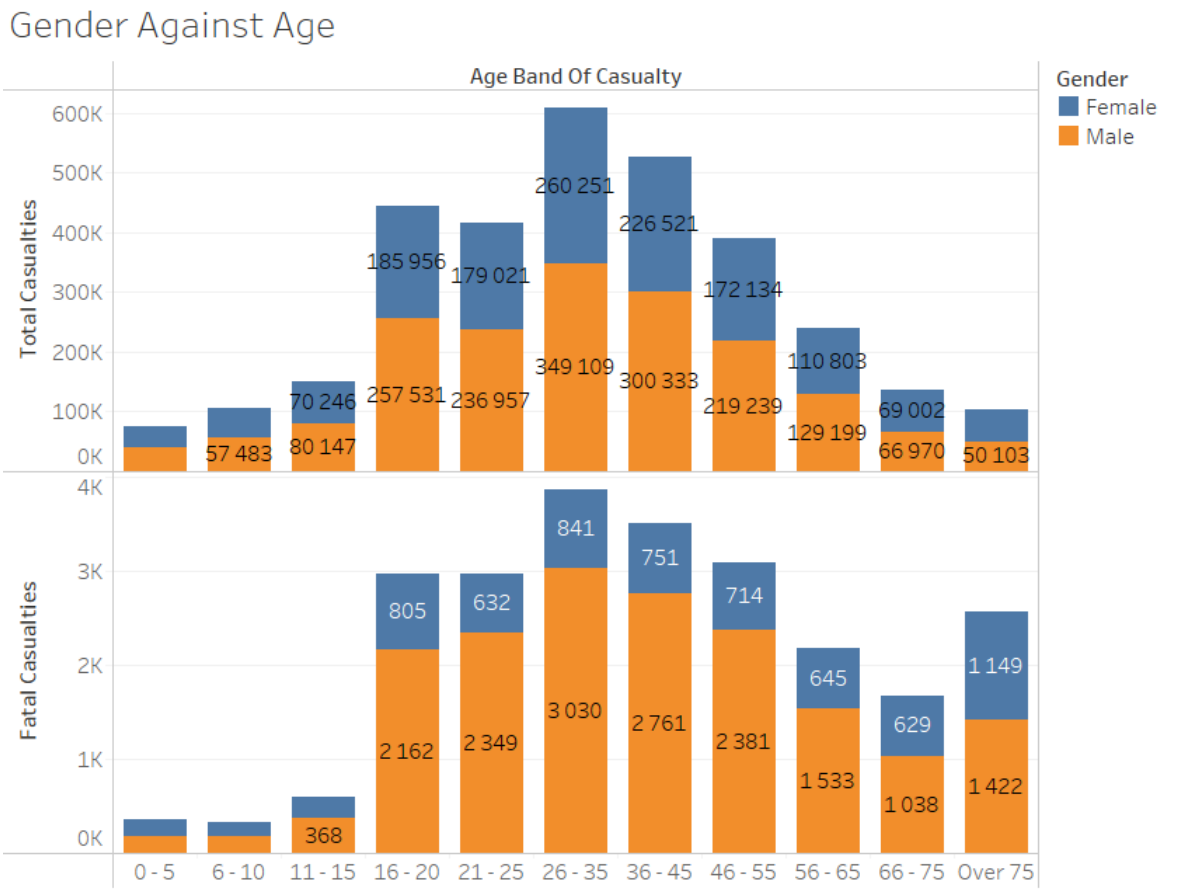
Rysunek 38 Liczba wypadków w kwartałach w zależności od warunków oświetleniowych

Natomiast z drugiej strony zima przynosi krótsze dnie i dłuższe noce, dlatego w okresach, gdzie jest większość dnia ciemno widać drastyczny wzrost liczby wypadków w ciemnościach. Natomiast do tego trzeba zaznaczyć, że większość wypadków i tak dzieje się w trakcie dnia, co widać na Rysunek 39



Rysunek 39 Wypadki w danych porach dnia

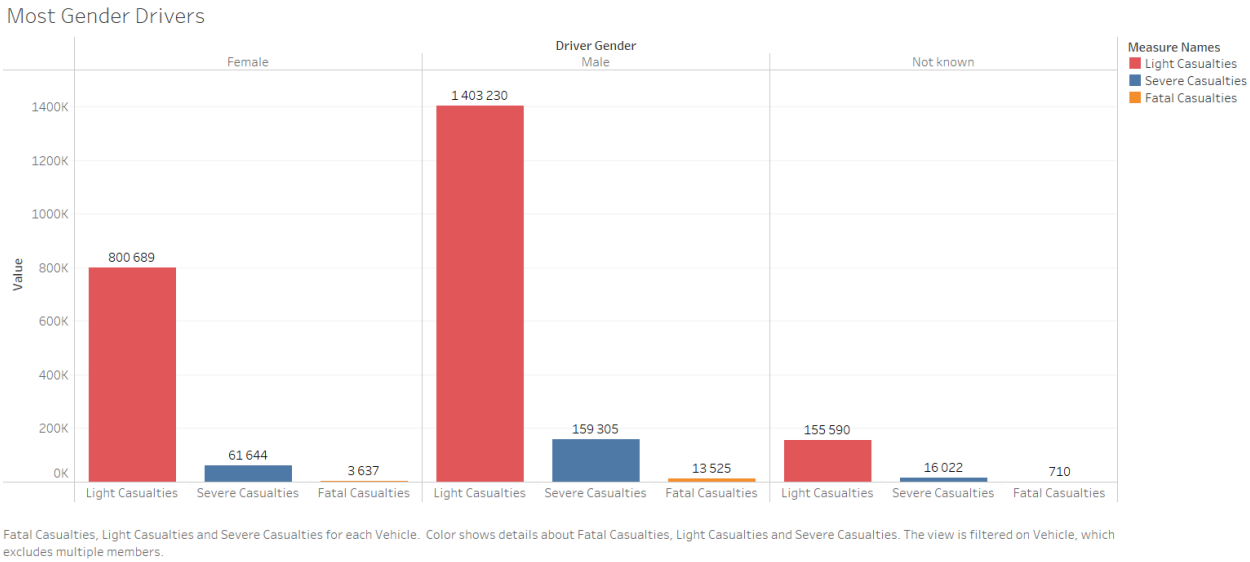
Przejdźmy teraz do szczegółowej analizy ofiar i kierowców w wypadkach,



Rysunek 40 Podział ofiar wszystkich i ofiar śmiertelnych ze względu na grupę wiekową i płeć

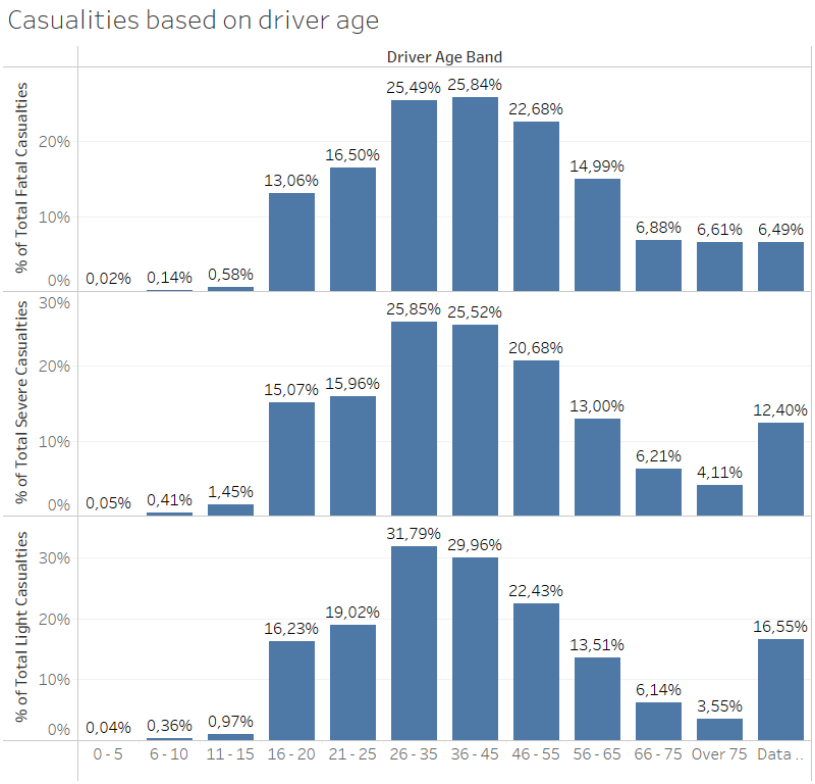
Na powyższym obrazku, widać, że większość ofiar śmiertelnych to mężczyźni, w wieku 16-65. Jest to grupa wiekowa, w której jest najwięcej kierowców samochodów. Natomiast w ogólnym rozrachunku ofiar mniej więcej po równo, z przewagą nadal mężczyzn. Bardzo dużo ofiar śmiertelnych jest w osobach powyżej 75 roku życia, natomiast to spowodowane jest tym, że w tym wieku każdy lekki wypadek może być śmiertelny.

Gdy spojrzymy na kierowców, tutaj także pokazuje się podobny obraz

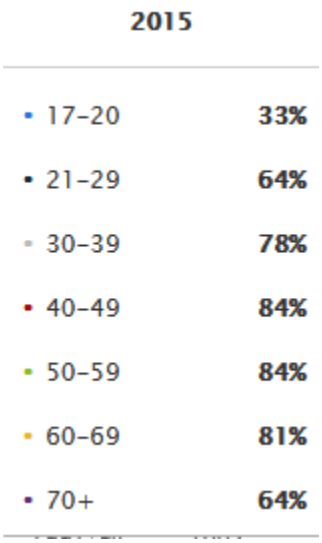


Rysunek 41 Kierowcy pojazdów uczestniczących w wypadku

Na podstawie tych danych można określić, że dużo więcej mężczyzn jest kierowcami w wypadkach śmiertelnych od kobiet. Natomiast według źródła więcej mężczyzn posiada prawo jazdy od kobiet, więc z tego też powodu więcej kierowców może być pośród mężczyzn niż pośród kobiet.

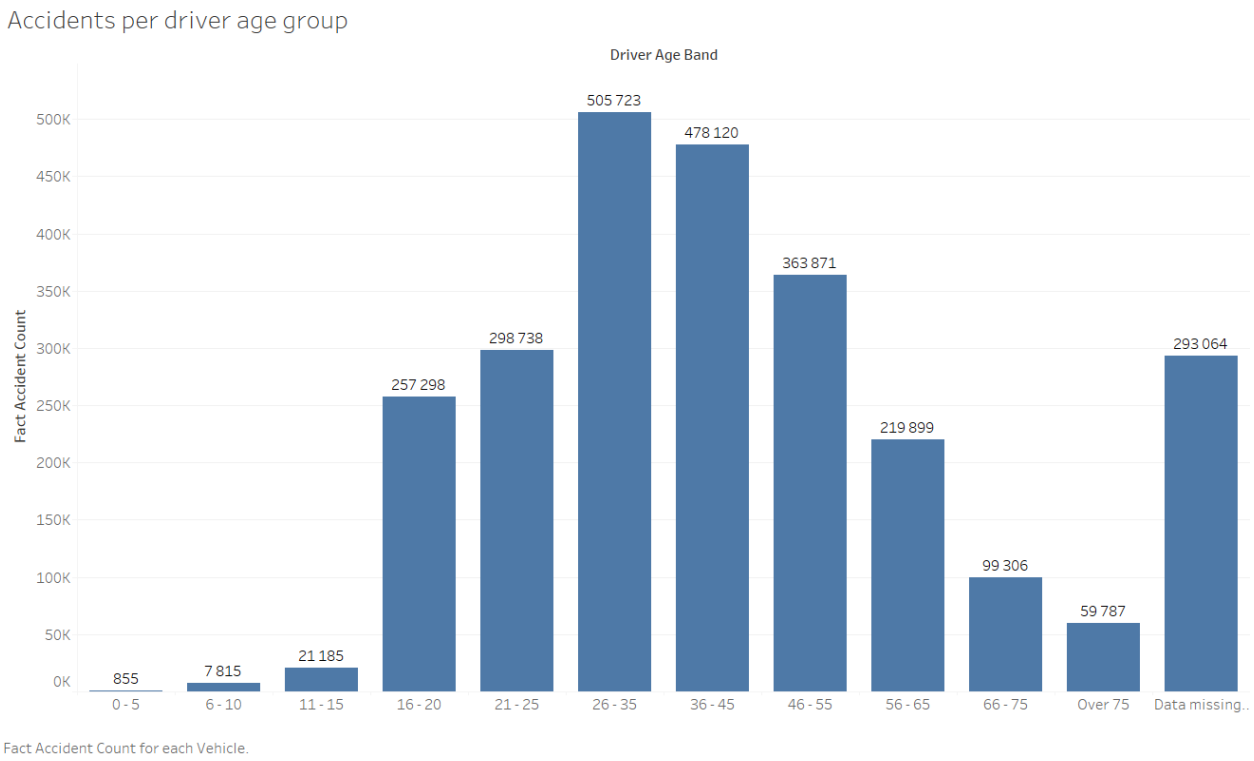


Rysunek 42 Ofiary w zależności od wieku kierowcy



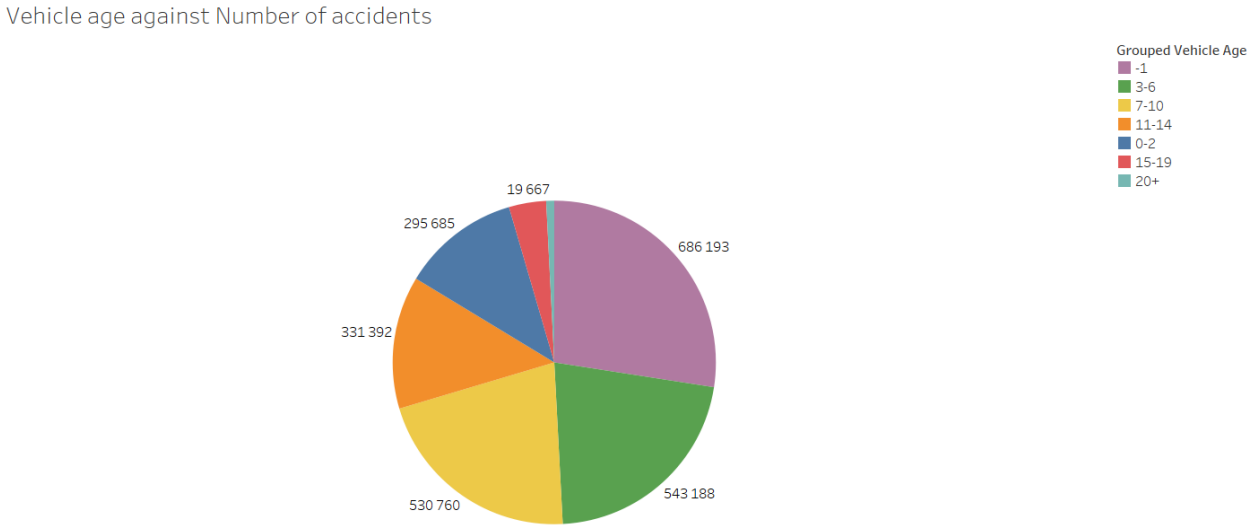
Rysunek 43 Udział praw jazdy w 2015 roku

Najwięcej ofiar jest przy kierowcach w wieku 26-55. Jednak koreluje ze źródłem, które mówi, że najwięcej jest kierowców w tych grupach wiekowych. Z tego powodu nie widać zależności, że im starszy kierowca, tym bezpieczniej jeździ

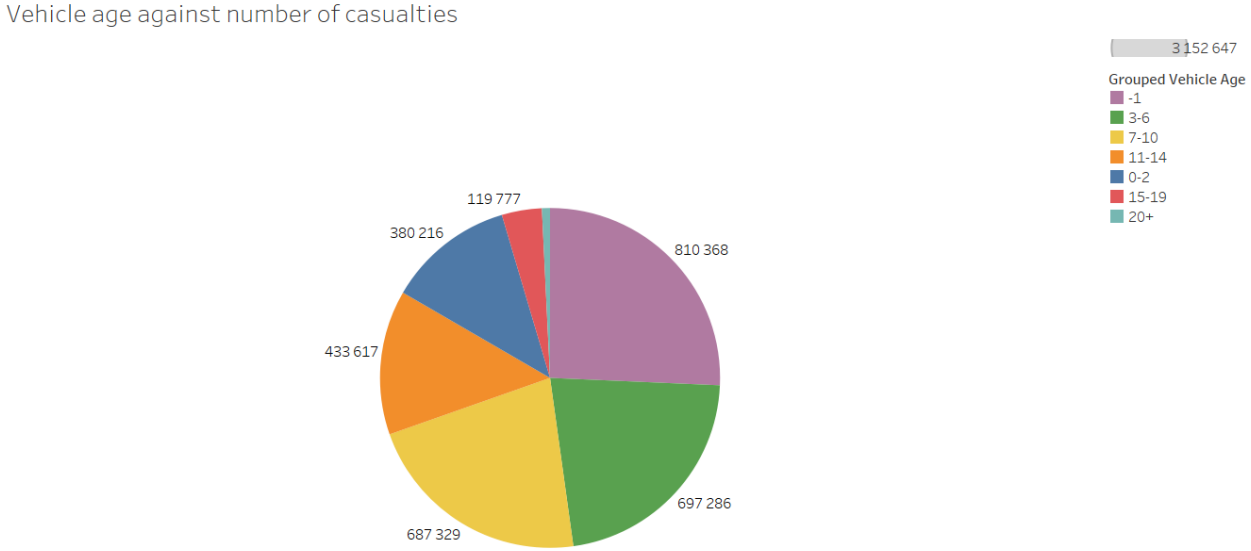


Rysunek 44 Liczba wypadków spowodowana przez konkretne grupy kierowców

Widać tutaj też, że najwięcej wypadków powodowane jest przez osoby w wieku 26-45.

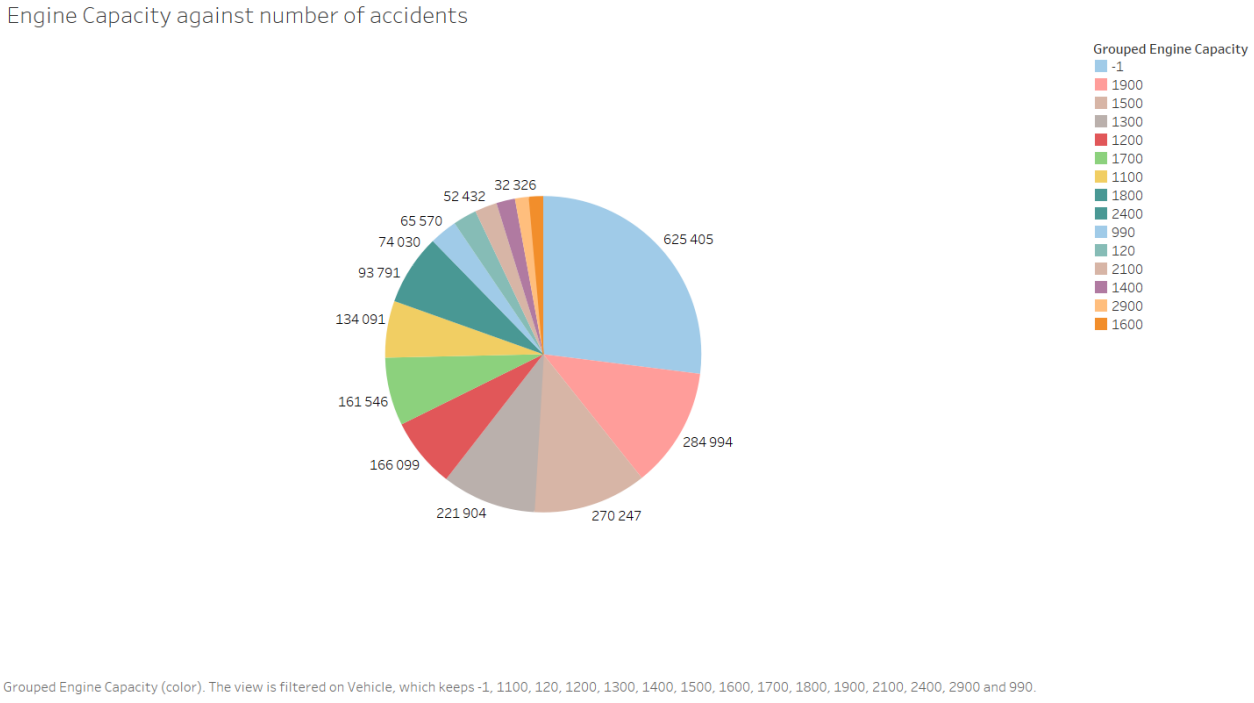


Rysunek 45 Wiek aut w stosunku do liczby wypadków



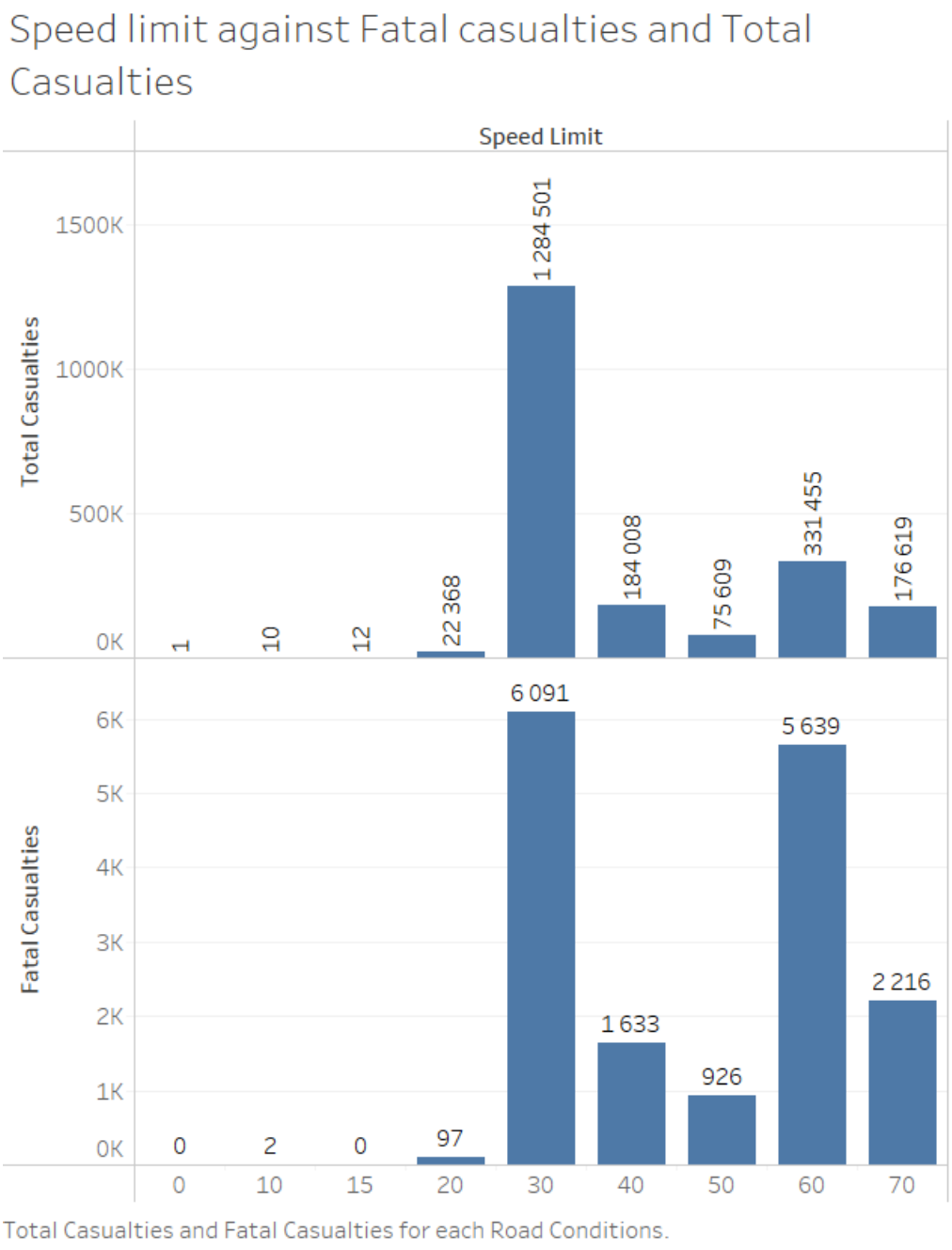
Rysunek 46 Wiek auta w stosunku do liczby ofiar

Niestety, ¼ pojazdów nie ma danych o wieku, dlatego analiza jest częściowa, natomiast większość aut uczestniczących w wypadku mają między 3 a 10 lat. Niestety, może to wynikać z tego, że średni wiek aut w UK to 8.4 roku[źródło], więc w tych zakresach jest po prostu najwięcej aut.



Rysunek 47 Pojemność silnika w stosunku do liczby wypadków

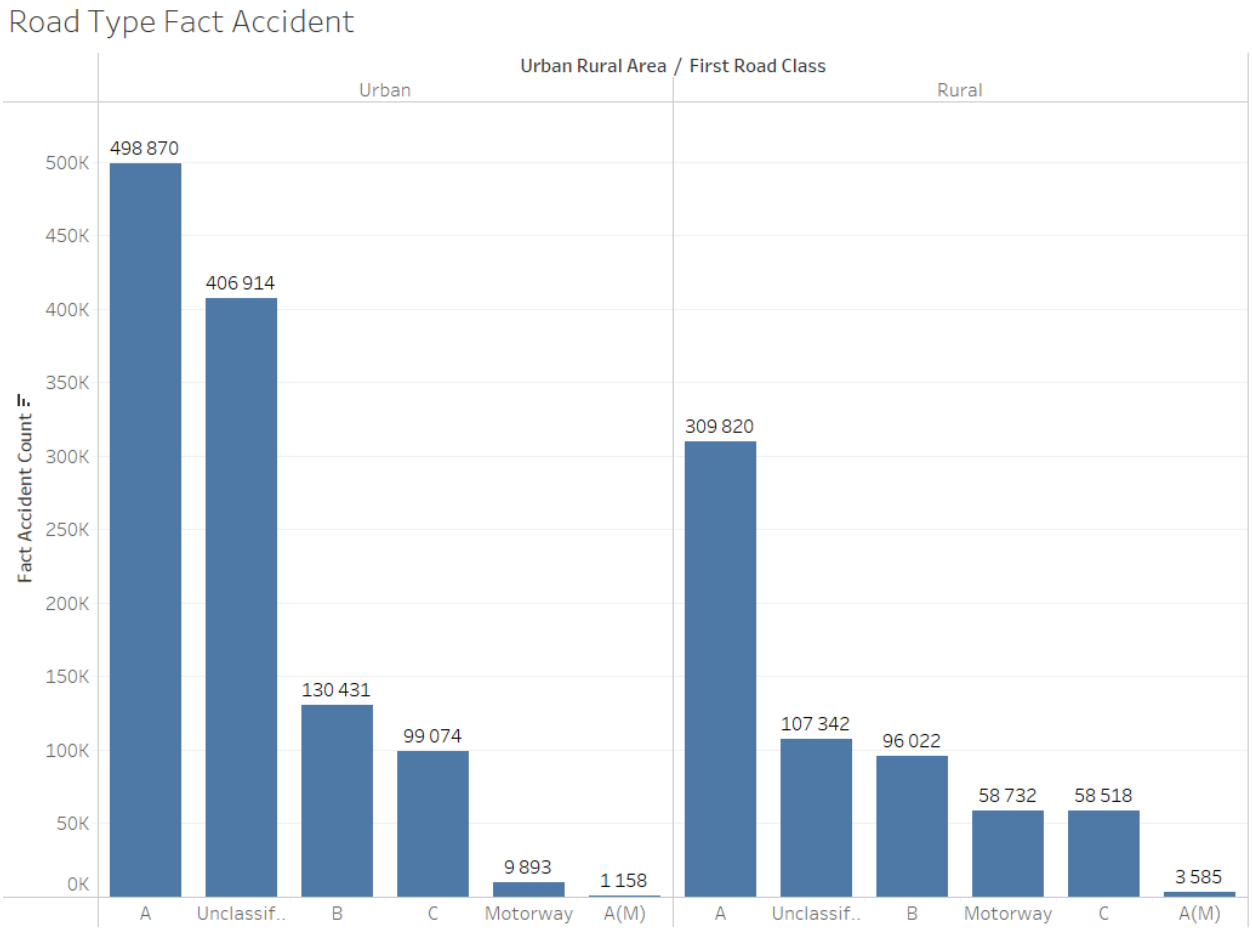
Z tych danych wynika, że, znowu, niestety większość pojazdów nie ma podanej pojemności silnika. Z tego powodu na podstawie częściowych danych wynika, że większość wypadków jest spowodowane przez pojazdu o pojemności silnika między 1000 a 2000 CC. Koreluje to z tym, że średnia moc silnika w UK w roku 2012 było 1735cc, a w 2009 1750cc (źródło), więc tych aut jest najwięcej na rynku



Rysunek 48 Limity prędkości w stosunku do całkowitej liczby i śmiertelnej liczby ofiar

Na podstawie tych danych widać, że najwięcej ofiar jest na drogach o prędkości o limicie 30mph, natomiast śmiertelnych ofiar widać po drugim wykresie, że ich liczba rośnie na drogach o limicie prędkości 60mph lub więcej

Z prędkością związana jest też jakość drógi



Rysunek 49 Typ drogi i liczba wypadków

Na podstawie tej analizy widać, że najwięcej wypadków dzieje się w miastach na drogach typu A, oraz na drogach nieklasyfikowanych. Dodatkowo, także w wiejskich warunkach klasa A jest najpopularniejszym miejscem wypadku

8.2. Podsumowanie - wnioski z analizy

Analiza pomogła znaleźć odpowiedzi na następujące pytania:

1. **Jakie jest przekrój wiekowy oraz płciowy ofiar oraz kierowców?**

   W przypadku ofiar, najczęstszą ofiarą jest mężczyzna w wieku 26-35 lat.

2. **Czy starsze auta są bezpieczniejsze?**

   Na podstawie danych nie stwierdzono, żeby starsze auta były bezpieczniejsze

3. **Czy starsi wiekowo kierowcy jeżdżą bezpieczniej?**

Na podstawie danych wynika, że kierowcy w wieku średnim jeżdżą najniebezpieczniej

4. **Czy limit prędkości ma wpływ na bezpieczeństwo na drogach?**

Limit prędkości nie pomaga w bezpieczeństwie na drogach, jednak widać niewielki trend, że przy większych limitach prędkości jest więcej śmiertelnych wypadków

5. **Czy warunki na drodze mają wpływ na bezpieczeństwo?**

Tak, warunki na drodze mają wpływ na bezpieczeństwo

6. **Czy typ drogi ma wpływ na liczbę wypadków?**

Najwięcej wypadków jest na typie dróg A, w mieście

Przedstawiona analiza odpowiedziała na zadane pytania, a także wiele innych.

9. Wnioski końcowe z realizacji projektu

9.1. Problemy

Podczas realizacji projektu spotkałem się z wieloma problemami, które bardzo często były blokadą w projekcie. Najpierw, przetworzenie danych – zrozumienie danych i ich analiza zajęło mi dużo czasu, przetwarzania i sprawdzania przy pomocy SQL. Następnie ETL, którego robiłem 2 razy, ponieważ za pierwszym razem nie zrozumiałem, o co chodzi w procesie ETL, więc zrobiłem bazę danych generowaną jednorazowo zamiast przyrostowo. Po stworzeniu ETLa wiele razy poprawiałem go, bo wychodziły małe szczegóły blokujące proces, tak jak na przykład złe atrybuty lub literówka. Przy przejściu do analizy, nie umiałem zrozumieć, co mam zrobić ze wskaźnikiem KPI, tak więc chociaż wygenerowałem go, tak nie do końca rozumiem, w jaki sposób go zaprezentować. Także kostka przysporzyła mi wiele problemów, głównie wynikających z tego, że mam 2 fakty pomostowe i SSIS nie umiał tego przetworzyć początkowo. Na szczęście finalnie udało się zmusić go do współpracy.

9.2. Pozyskana wiedza i doświadczenie

Generalnie, chociaż sam projekt był trudny i czasochłonny, podoba mi się analiza danych i polubiłem przetwarzanie danych. Gdybym miał lepsze, bardziej ciekawsze, dla mnie, dane, to bardzo chętnie podszedłbym do tego jeszcze raz, i zrobił analizę na

innych danych. Myślę, że po kursie we własnym czasie nauczę się alternatywnych metod tworzenia procesu ETL oraz analizy danych, bo chociaż podoba mi się graficzny aspekt tworzenia ETL i Kostki przy pomocy Visual Studio, tak sam program przysporzył mi koszmary w obsłudze, wydajności działania oraz ogólnego doświadczenia. Na pewno spróbuję nauczyć się Pandas oraz różnych innych bibliotek Pythona, które pozwalałby na tworzenie procesów ETL oraz analizę wielowymiarową, bo temat bardzo mi się podoba. Natomiast Tablau jest świetnym programem do tworzenia wykresów, według mnie lepszym od Excela, którego na pewno będę korzystał, dopóki będę miał licencję na niego, a potem nawet może wykupie dostęp do niego.

10. Źródła informacji użyte w etapie analizy danych

1. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/198753/vls-2012.pdf
2. https://www.autoexpress.co.uk/news/59950/average-age-uk-cars-reaches-record-high
3. https://www.statista.com/statistics/314898/share-driving-licence-holders-by-age-england/
4. https://www.statista.com/statistics/314886/percentage-of-adults-holding-driving-licences-england/
5. https://www.nimblefins.co.uk/cheap-car-insurance/number-cars-great-britain#nogo

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!