AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ INFORMATYKI, ELEKTRONIKI I TELEKOMUNIKACJI

KATEDRA INFORMATYKI

WYKONANO W RAMACH:

Zaawansowane Techniki Integracji Systemów

Temat D: "Analiza przyczyn wypadków drogowych"

WERSJA: **2.58**, 05.06.2015

Dokumentacja projektowa

Łukasz Opioła, Beata Skiba

lopiola@student.agh.edu.pl, bskiba@student.agh.edu.pl

Kraków, 2015

Spis treści

1. Temat i cel projektu

1.1. Temat projektu

Tematem projektu jest integracja różnych źródeł danych opisujących przyczyny wypadków drogowych. Temat obejmuje identyfikację reprezentatywnych źródeł i zebranie z nich danych, sprowadzenie ich do wspólnej reprezentacji a następnie poddanie tak przetworzonych danych wieloprzekrojowej analizie.

1.2. Cel projektu

Głównym celem projektu jest analiza danych na temat wypadków drogowych pod kątem prawdopodobnych przyczyn.

Można wyróżnić kilka pośrednich celów, których realizacja będzie konieczna w ramach projektu. Pierwszym z nich jest zgromadzenie danych na temat wypadków drogowych, pochodzących z reprezentatywnych źródeł. Obszarami, na których będziemy chcieli się skupić są Polska, Europa Zachodnia, USA i Kanada. Celem naszym jest znalezienie danych godnych zaufania, możliwie pełnych i obszernych oraz analiza informacji, jakich te dane nam dostarczają. Aby umożliwić dalszą pracę z danymi pochodzącymi z różnych źródeł, konieczne będzie sprowadzenie ich do wspólnej reprezentacji danych, pozwalającej na dalszą analizę.

Kolejnym celem stawianym w projekcie jest umożliwienie wnioskowania o przyczynach wypadków. Dane o przyczynach mogą być dostępne w danych bezpośrednio, w przeciwnym razie konieczna będzie identyfikacja i selekcja kryteriów, pozwalających na przeprowadzenie wnioskowania w tym zakresie.

Ostatecznie celem projektu będzie przeprowadzenie analiz danych na wielu płaszczyznach z wykorzystaniem określonych wcześniej kryteriów. Pozwoli to zarówno na identyfikację możliwych przyczyn poszczególnych wypadków jak i wyciągnięcie wniosków co do ogólnych przyczyn wypadków i okoliczności zwiększających ryzyko ich wystąpienia.

2. Plan pracy

2.1. Zadania projektowe i plan pracy

Zadanie: Zebranie danych Skrótowa nazwa: Zebranie danych
Czas realizacji: 16.03 - 30.03
Opis: Identyfikacja i ocena dostępnych źródeł danych, wstępna analiza ich przydatności i zawartości.
Produkty: Artefaktem potwierdzającym ukończenie zadania i zawierającym rezultaty przeprowadzonych działań będzie dokument "Analiza źródeł danych"

Zadanie: Wybór atrybutów Skrótowa nazwa: Wybór atrybutów
Czas realizacji: 30.03 - 13.04
Opis: Dogłębna analiza danych zawartych w źródłach i wybór kryteriów przyszłej analizy oraz selekcja tych atrybutów ze źródeł, które będą konieczne do realizacji celów projektowych.
Produkty: Artefaktem z tego zadania będzie dokument "Kryteria analizy"

Zadanie: Opracowanie wspólnego formatu danych i schematu bazy
Skrótowa nazwa: Wspólny format
Czas realizacji: 06.04 - 20.04
Opis: Stworzenie zgodnie z dokonaną wcześniej selekcją atrybutów wspólnego formatu danych. Obejmuje zarówno ujednolicenie formatu danych (np. data, sposób oznaczenia miejsca wypadku) jak i semantyki (np. obranie wspólnych określeń na warunki atmosferyczne panujące w momencie wypadku).
Produkty: W związku z tym zadaniem powstaną następujące artefakty: dokument "Projekt bazy danych" zawierający opis formatu i semantyki danych oraz schemat bazy danych.

Zadanie: Integracja danych i zapis do bazy
Skrótowa nazwa: Integracja
Czas realizacji: 13.04 - 04.05
Opis: Ekstrakcja danych ze źródeł, transformacja do wspólnego formatu i załadowanie do bazy danych. Obejmuje utworzenie skryptów parsujących i konwertujących na wspólny format i wspólną semantykę.
Produkty: Produktami tego zadania będą skrypty oraz uzupełniona baza danych.

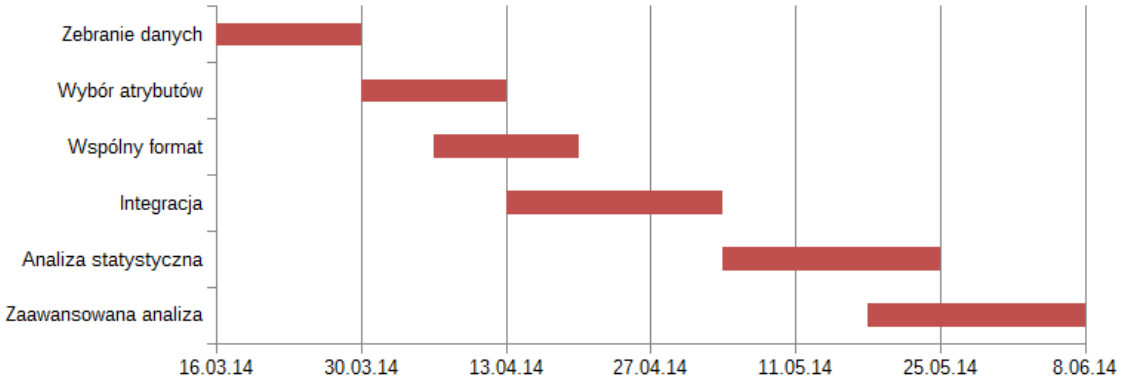
Zadanie: Statystyczna analiza danych
Skrótowa nazwa: Analiza statystyczna

Czas realizacji: 04.05 - 25.05
Opis: Statystyczna analiza danych w celu wyłonienia możliwych przyczyn wypadku, zgodna z określonymi wcześniej kryteriami.
Produkty: Realizacja tego zadania przejawi się w następujących produktach: skrypty przeprowadzające analizy, wykresy obrazujące przeprowadzone analizy, dokument "Raport z analiz", który będzie tworzony stopniowo, być może jako zbiór mniejszych dokumentów wraz z przeprowadzaniem kolejnych analiz oraz dokument "Wnioski z analiz", który będzie opisywał wnioski na temat przyczyn wypadków drogowych wyciągnięte na podstawie przeprowadzonych analiz.

Zadanie: Zaawansowana analiza danych
Skrótowa nazwa: Zaawansowana analiza
Czas realizacji: 18.05 - 8.06
Opis: Zaawansowana analiza danych w celu wyłonienia możliwych przyczyn wypadku, korzystająca z mechanizmów takich jak klasteryzacja czy wyszukiwanie wzorców częstych.
Produkty: Realizacja tego zadania przejawi się w następujących produktach: skrypty przeprowadzające analizy, wykresy obrazujące przeprowadzone analizy, dalsze części dokumentu "Raport z analiz", oraz dalsze części dokumentu "Wnioski z analiz" (patrz zadanie Statystyczna analiza danych).

2.2. Wykres Gantta

Opisane powyżej zadania zostały przedstawione na wykresie Gantta, zgodnie z przewidywanymi terminami ich wykonywania.



Rysunek 2.1: Diagram Gantta

2.3. Ryzyko w projekcie

Prawdopodobieństwo wystąpienia oraz wpływ na projekt podane w skali 1-5

Ryzyko: Problem z uzyskaniem części danych
Opis: Otrzymanie danych z POBR jest uzależnione od fizycznej osoby, która musi przydzielić możli-

wość pełnego dostępu do danych po wcześniejszym otrzymaniu stosownego pisma. Procedura taka może za długo trwać i nie zdążymy uwzględnić tych danych w projekcie.

Prawdopodobieństwo wystąpienia: 4

Wpływ na projekt: 2

Rozwiązanie: Praca z danymi dostępnymi i uwzględnienie niedostępnych danych w postaci statystyk

Ryzyko: Problem z ustaleniem wspólnego formatu danych

Opis: Dane, z których korzystamy, pochodzą z bardzo różnych źródeł, mają różną semantykę i format. Może się okazać niezwykle trudne stworzenie wspólnego formatu danych dla tych źródeł, które jednocześnie będzie maksymalnie wykorzystywał dane w nich dostępne

Prawdopodobieństwo wystąpienia: 4

Wpływ na projekt: 4

Rozwiązanie: W razie konieczności zawężenie wspólnego formatu danych i w konsekwencji utrata części danych z niektórych źródeł

Ryzyko: Duża ilość brakujących danych

Opis: Wiele przydatnych informacji może brakować dla części danych. Część ta może się okazać bardzo duża

Prawdopodobieństwo wystąpienia: 3

Wpływ na projekt: 2

Rozwiązanie: Problem brakujących danych jest trudny do zwalczenia, gdyż nie ma możliwości uzupełnienia ich wszystkich, należy więc w taki sposób zaprojektować analizy aby uwzględnić możliwość braku niektórych danych.

Ryzyko: Duża ilość danych

Opis: Źródła danych, z którymi będziemy pracować w projekcie zawierają dużo danych.

Prawdopodobieństwo wystąpienia: 4

Wpływ na projekt: 2

Rozwiązanie: Duża ilość danych jest z jednej strony dobra gdyż pozwala stwierdzić reprezentatywność analiz i rozważyć wiele przypadków, może być jednak kłopotliwa w przeprowadzaniu analiz. Należy dobrze zaprojektować bazę danych i ograniczyć się do przechowywania tych atrybutów, które będą naprawdę przydatne w analizie. Analizy musza brać pod uwagę ilość danych i np. ograniczać się do analizowania w danym momencie danych z jednego roku.

3. Opis dziedziny problemu

Cel dokumentu

Niniejszy dokument ma na celu podsumowanie informacji na temat źródeł danych, które będziemy mogli wykorzystać w projekcie. Charakterystyka dokumentów obejmuje zakres danych, możliwe problemy z ich otrzymaniem i wykorzystaniem a także zawartość danych - informacje o wypadkach, jakie można znaleźć w opisywanych danych. Dokument zawiera również wstępną definicję kryteriów, w ramach których będziemy analizować przyczyny wypadków drogowych, z uwzględnieniem informacji o źródłach.

3.1. Przegląd źródeł danych

Głównymi obszarami zainteresowania przy szukaniu źródeł danych były: Polska, Europa Zachodnia, USA, Kanada

1. Polska
   - Polskie Obserwatorium Bezpieczeństwa Ruchu Drogowego (POBR) [2]
     - System składa się z dwóch części: hurtowni danych i portalu informacyjnego z interaktywną mapą do przeglądania danych z hurtowni.
     - Dostęp jest ograniczony - bez rejestracji możliwy jedynie dostęp do wersji demo interaktywnej mapy oraz części zbiorczych statystyk.
     - Studenci, dziennikarze, przedstawiciele społeczności lokalnych mogą uzyskać czasowy dostęp do pełnej bazy danych, po uprzednim pisemnym potwierdzeniu, że dane potrzebne będą do przygotowywanych opracowań.
     - Z informacji na stronie wynika, że dane powinny być rzetelne i pełne.
     - Wnioskując po wersji demo mapy interaktywnej, w hurtowni danych dostępne powinny być przynajmniej następujące informacje na temat wypadków:
       * miejsce
       * czas
       * droga (numer, rodzaj)

        * obszar zabudowany/niezabudowany
        * informacje o uczestnikach i ofiarach wypadku
            · wiek
            · płeć
            · pojazd
            · obrażenia
            · "ciężkość" wypadku (śmiertelny, ciężki, lekki)
        * obecność alkoholu u sprawcy
        * nadmierna prędkość

2. Europa Zachodnia
    - CARE [3]
        - Zawiera dane dostarczane przez odpowiednie instytucje ze wszystkich krajów Unii Europejskiej w postaci przetworzonej do wspólnego formatu
        - Dostęp bezpośrednio do bazy niestety nie jest możliwy, dla wiadomości publicznej dostępne są tylko wybrane zbiorcze statystyki
    - Wielka Brytania [5]
        - Dane z wypadków na drogach publicznych, zgłoszonych na policję i odnotowanych przy wykorzystaniu formularza STATS19.
        - Dane dostępne dla lat 1979 - 2013
        - Dane obejmują m.in.:
            * okoliczności wypadku
            * typy pojazdów (marka i model)
            * dane o ofiarach
            * informacje o poziomie alkoholu w wydychanym powietrzu
    - Belgia-Flandria [**?**]
        - Dane zebrane przez National Institute of Statistics (NIS)
        - Obejmują lata 1991 - 2013 i tylko jeden region. Wydaje się, że jest to zbyt mocne ograniczenie.

3. USA i Kanada
    - NHTSA: National Automotive Sampling System (NASS) - Crashworthiness Data System (CDS) [6]
        - Losowe próbki danych na temat wypadków o różnych skutkach (od małych po śmiertelne). Około 5'000 przekrojowo wybranych wypadków rocznie jest dokładnie badanych i dane są publikowane. Dane są zanonimizowane.
        - Dostępne atrybuty (m. in.):

        * powaga wypadku (obrażenia, ilość rannych)
        * pijani kierowcy
        * data
        * udział narkotyków
        * rodzaj wypadku
        * dane nt. pojazdów, wyposażenie
        * brak danych o pogodzie, miejscu zdarzenia
  - NHTSA: Fatality Analysis Reporting System (FARS) [1]
    - ftp://ftp.nhtsa.dot.gov/FARS/
    - Wypadki, w wyniku których odnotowano przynajmniej jedną ofiarę śmiertelną (max. w ciągu 30 dni po wypadku).
    - Od roku 1975 zebrano dane z wypadków w trzech typach dokumentów: Accident, Vehicle, Person.
    - Od roku 2010 dodano wiele nowych dokuemntów: Peperwork, Cevent, Vevent, Vsoe, Distract, Factor, Drimpair, Nmimpair, Maneuver, Nmprior, Nmcrash, Safetyeq, Violatn, Vision, Damage, Vindecode (2013).
    - Dostępne atrybuty (m. in.):
      * szczegółowe dane o biorących udział w wypadku
      * szczegółowe dane o pojazdach
      * miejsce
      * data
      * rodzaj wypadku
      * sposób dojścia do wypadku
      * szczegóły drogi (rodzaj, odl. od skrzyżowania etc)
      * oświetlenie
      * pogoda
      * pijani kierowcy
  - Kanada - nie znaleziono publicznie dostępnych zestawów danych

3.2. Dane z Wielkiej Brytanii

Ogólne informacje

Zbiór danych z Wielkiej Brytanii to dane policyjne z wypadków na drogach publicznych, zgłoszonych i odnotowanych w formularzu ([7]) który jest wypełniany przez funkcjonariuszy przy zgłoszeniu

zdarzenia. Dane są dostępne pod adresem [5]. Geograficznie dane obejmują Anglię, Walię i Szkocję zaś czasowo obejmują okres 1979 - 2013r.

Format danych

Dane są dostępne w dwóch paczkach, osobno dane dla lat 1979 - 2004 i 2005 - 2013, w postaci plików csv. Każda paczka danych zawiera trzy pliki csv zawierające komplet danych opisujących wypadek:

- *Accident* - dane ogólne na temat okoliczności wypadku
- *Casualty* - dane na temat ofiar wypadku i ich obrażeń, połączone logicznie z plikiem accident poprzez atrybut ACC_Index, który jednoznacznie identyfikuje wypadek. Dla jednego wypadku możemy mieć kilka wpisów w pliku casualty, jeżeli była więcej niż jedna osoba poszkodowana.
- *Vehicle* - dane na temat pojazdów, które brały udział w kolizji i ich obrażeń, połączone logicznie z plikiem accident poprzez atrybut ACC_Index. Dla jednego wypadku możemy mieć kilka wpisów w pliku vehicle.
  Pierwsza linijka każdego z plików zawiera nazwy atrybutów.

Nazwy atrybutów oraz ich wartości są przeniesione bezpośrednio z formularza ([7]). Pozwala to na interpretację kodów wartości zawartych w plikach i na ich translację na wartości zrozumiałe dla człowieka. Dodatkową pomoc w interpretacji wartości może stanowić dokument STATS20 ([8]), który zawiera dokładne informacje co do tego, jak wypełniać formularz STATS19. Przedstawiony powyżej podział na trzy pliki odzwierciedla podział formularza na analogiczne trzy części.

Z dodatkowych informacji istotnych dla realizacji projektu należy zaznaczyć, że dane dotyczące części atrybutów mogą być niedostępne (lub nie dotyczyć danego wypadku), wartość takiego atrybutu jest wtedy równa -1. Zgodnie z decyzją projektową ograniczenia się do wypadków śmiertelnych jest istotna również możliwość przefiltrowania danych i wybrania wyłącznie wypadków śmiertelnych. Jest to możliwe dzięki polu Casualty_Severity (ciężkość wypadku) w pliku dotyczącym ofiar wypadku, które może przyjmować jedną z trzech wartości: fatal, serious i slight (śmiertelny, poważny, lekki).

Ilość danych

Paczka danych z lat 1979 - 2004 zawiera następującą ilość danych:

- 6224198 wypadków
- 8264687 ofiar
- 10981968 pojazdów
  Paczka danych z lat 2005 - 2013 zawiera następującą ilość danych:
- 1494275 wypadków
- 2022243 ofiar
- 2735898 pojazdów

Pozostaje do ustalenia jaki procent tych danych stanowią dane o wypadkach śmiertelnych

Atrybuty

Szczegóły dot. wypadku

- data i godzina
- dzień tygodnia
- miejsce wypadku (długość i szerokość geograficzna lub współrzędne OSGR)
- powaga wypadku
- lekki
- poważny
- śmiertelny
- szczegóły dróg (pierwszej i drugiej)
- klasa drogi
- numer drogi
- liczba ofiar
- liczba pojazdów
- typ drogi
- rondo
- droga jednokierunkowa
- jezdnia podwójna
- jezdnia pojedyncza
- droga dojazdowa
- nieznany
- ograniczenie prędkości na drodze
- szczegóły dot. skrzyżowania (np. dalej niż 20 metrów od skrzyżowania, na rondzie, na wyjeździe z drogi prywatnej)
- sposób kierowania ruchem (dla wypadków na skrzyżowaniu)
- szczegóły dot. przejścia dla pieszych i kontroli nad nim (np. dalej niż 50m od najbliższego przejścia, czy na przejściu kontrola osób autoryzowanych)
- rodzaj przejścia dla pieszych
- pogoda
- dobra, bez porywistego wiatru

- deszcz, bez porywistego wiatru
- śnieg, bez porywistego wiatru
- dobra, porywisty wiatr
- deszcz, porywisty wiatr
- śnieg, porywisty wiatr
- mgła
- inne
- nieznane
- stan nawierzchni
- sucha
- mokra/wilgotna
- śnieg
- mróz/lód
- zalana (powyżej 3cm wody)
- światło
- światło dzienne
- ciemność, oświetlenie, zapalone
- ciemność, oświetlenie, nie zapalone
- ciemność, brak oświetlenia
- ciemność, brak danych co do oświetlenia
- warunki nadzwyczajne (np. niedziałające światła, roboty na drodze, uszkodzona nawierzchnia)
- zagrożenia na jezdni (np. obiekty na jezdni, udział w poprzedzającym wypadku, pieszy na jezdni, zwierzę na jezdni)
- Obecność policjanta na miejscu wypadku
- Obszar zabudowany/niezabudowany

Szczegóły dot. pojazdu i kierowcy

- kierownica po lewej stronie
- typ samochodu (podział na 10 kategorii)
- pojemność silnika
- kod napędu (propulsion code)

- pojazdy z naczepami i przegubowe
- wiek kierowcy
- kod pocztowy kierowcy
- ucieczka z miejsca zdarzenia (hit and run)
- test alkoholu w wydychanym powietrzu u kierowcy (?)
- nie dotyczy
- pozytywny
- nie proszony o test
- nie zgodził się na test
- nie było kontaktu z kierowcą w momencie wypadku
- nie podano (powody zdrowotne)
- płeć kierowcy
- umiejscowienie pojazdu w momencie zderzenia (np. na głównej drodze, na pasie dla tramwajów, autobusów lub rowerów)
- umiejscowienie pojazdu na skrzyżowaniu (np. w odległości ponad 20m, wjeżdżający na skrzyżowanie, zjeżdżający, wjeżdżający na rondo, zjeżdżający z głównej drogi lub wjeżdżający na nią, na środku skrzyżowania)
- wykonywany manewr (np. cofanie, zaparkowany, zatrzymany, zatrzymywanie się, skręcanie, zawracanie)
- obiekt uderzony na jezdni (np. poprzedni wypadek, zaparkowany pojazd, most, otwarte drzwi pojazdu, krawężnik)
- miejsce zjazdu z jezdni
- poślizg i dachowanie
- brak poślizgu, dachowania, jack-knifingu (złożenie się samochodu z naczepą na kształt scyzoryka)
- poślizg
- poślizg i dachowanie
- jack-knifing
- jack-knifing i dachowanie
- dachowanie
- pierwszy obiekt uderzony poza jezdnią (np. znak drogowy, latarnia, drzewo)

- pierwsze miejsce uderzenia
- przód
- tył
- lewy bok
- prawy bok
- powód podróży
- w ramach pracy
- do/z pracy
- wiezienie dziecka do/ze szkoły
- inne
- nieznane
- kierunek jazdy pojazdu, 10 możliwości, włączając samochód zaparkowany

Szczegóły dot. poszkodowanych

- w którym pojeździe znajdowała się ofiara
- kod pocztowy
- płeć
- typ poszkodowanego
- kierowca
- pasażer
- pieszy
- wiek poszkodowanego
- powaga obrażeń
- lekkie
- poważne
- śmiertelne
- umiejscowienie pieszego (np. na jezdni, na pasach, na krawężniku, na wysepce centralnej)
- kierunek podążania pieszego, podobnie jak dla pojazdów 10 możliwości, łącznie z nieruchomym pieszym
- ruch pieszego względem pojazdu
- czy pieszy był pracownikiem utrzymania dróg (road maintenance worker)

- czy kierujący rowerem miał na sobie kask
- na którym siedzeniu znajdował się pasażer
- przednie
- tylnie
- szczegóły pasażera autokaru lub autobusu
- poszkodowany nie był pasażerem autobusu
- pasażer wsiadał
- pasażer wysiadał
- pasażer stał
- pasażer siedział
- pasy bezpieczeństwa
- nie dotyczy
- założone, potwierdzone niezależnie
- założone , nie potwierdzone niezależnie
- nie założone
- brak danych

3.3. Dane z USA

Ogólne informacje

Zbiór danych pochodzi z portalu organizacji NHTSA (National Highway Traffic Safety Administration). Jest to jeden z kilku publicznie dostępnych zbiorów, nazywa się FARS (Fatality Analisys Reporting System). Więcej informacji dostępne jest pod linkiem [1]. Zestaw zawiera dane ze wszystkich wypadków śmiertelnych zanotowanych na terenie Stanów Zjednoczonych. Zbiory publikowane są corocznie, w obecnej chwili dostępne są paczki z lat 1975 - 2013.

Format danych

Dane dostępne są w postaci plików bazy danych w jednym z wybranych formatów: **.sas7bdat** lub **.dbf**. Są to standaryzowane formaty i istnieje wiele narzędzi umożliwiających ich konwersję. Każda paczka zawiera następujące pliki:

- ACCIDENT (od 1975) - dane ogólne na temat okoliczności wypadku.
- VEHICLE (od 1975) - informacje na temat pojazdów biorących udział w wypadku, mogą być skojarzone z rekordem ACCIDENT za pomocą pola ST_CASE.

- PERSON (od 1975) - informacje na temat osób (zmotoryzowanych i pieszych) biorących udział w wypadku. Mogą być zkojarzone z rekordami ACCIDENT i VEHICLE.
- PAPERWORK (od 2010) - informacje na temat zaparkowanych pojazdów lub maszyn robót drogowych (biorących udział w wyapdku).
- CEVENT (od 2010) - lista wydarzeń, które doprowadziły do wypadku.
- VEVENT (od 2010) - opisuje sekwencje wydarzeń z CEVENT
- VSOE (od 2010) - uproszczona baza VEVENT
- DAMAGE (od 2010) - lista uszkodzeń pojazdów
- DISTRACT (od 2010) - czynniki odwracające uwgę kierowców
- DRIMPAIR (od 2010) - dane na temat niepełnosprawności kierowców
- FACTOR (od 2010) - okoliczności pojazdów, które mogły doprowadzić do wypadku
- MANEUVER (od 2010) - manewry wykonane przez kierowcę aby uniknąć wypadku
- VIOLATN (od 2010) - wykroczenia kierowców
- VISION (od 2010) - czynniki, które mogły zmniejszać widoczność
- NMCRASH (od 2010) - nieodpowiednie zachowania osób nieporuszających się pojazdami
- NMIMPAIR (od 2010) - dane nt. niepełnosprawości osób nieporuszających się pojazdami
- NMPRIOR (od 2010) - dane nt. czynności wykonywanych przez osoby nieporuszające się pojazdami przed wypadkiem
- SAFETYEQ (od 2010) - dane nt. wyposażenia BHP u osób nieporuszających się pojazdami przed wypadkiem
- VINDECODE (od 2013) - Kody VIN pojazdów biorących udział w wypadku

Atrybuty

Z racji, że większość plików dostępnych jest dopiero od 2010, nie będą one brane pod uwagę przy integracji danych. Bazy ACCIDENT, VEHICLE i PERSON zawierają bardzo szczegółowe dane, które powinny wystarczyć do analiz pod kątem przyczyn wypadków drogowych.

Poniżej przedstawiono jakie dane o wypadkach są dostępne w rozważanych bazach:

ACCIDENT

- liczba osób
  - zmotoryzowane
  - niezmotoryzowane
- liczba pojazdów
  - poruszające się

  - zaparkowane
  - pracujące
- miejsce
  - stan
  - miasto
  - okręg
  - rodzaj drogi
  - numer drogi i kamień milowy
  - wsp. GPS
  - jurysdykcja drogi
- data (dzień, miesiąc, rok, godzina, minuta)
- okoliczności
  - pierwsze szkodliwe wydarzenie
  - rodzaj kolizji
  - umiejscowienie względem skrzyżowania
  - udział autobusu szkolnego
  - wypadek przy torach kolejowych
  - czas zgłoszenia
  - czas przyjazdu służb na miejsce
  - czas dotarcia do szpitala
  - **przyczyny wypadku** (np dziurawa droga, ostry zakręt, warunki pogodowe, śliska nawierzchnia)
  - pijani kierowcy
  - ofiary śmiertelne
- oświetlenie
  - dzienne
  - ciemno / ciemno, nie oświetlone
  - ciemno ale oświetlone
  - zmierzch
  - świt
  - ciemno - nieznane oświetlenie

    - inne
    - nieraportowane
    - nieznane
- warunki pogodowe
    - brak niekorzystnych warunków
    - deszcz lub mżawka
    - deszcz ze śniegiem lub grad
    - śnieg lub zamieć
    - mgła, dym lub smog
    - porywisty wiatr
    - wiatr unoszący piach, ziemię lub pył
    - inne
    - zachmurzenie
    - nie raportowane
    - nieznane

VEHICLE

- ilość pasażerów
- rodzaj pojazdu
- ucieczka z miejsca wypadku
- stan w którym pojazd został zarejestrowany
- posiadacz samochodu
- marka
- model
- typo nadwozia
- rok produkcji
- VIN
- ciągnięte przyczepy
- jackknifing
- przedział wagowy
- konfiguracja pojazdu (ciężarowe)

- rodzaj naczepy
- przewóz niebezpiecznych materiałów
- specjalne przeznaczenie pojazdu
- prędkość poruszania
- dachowanie / obrót pojazdu
- miejsce uderzenia
- rozmiar uszkodzeń
- usunięcie pojazdu z miejsca wypadku
- najbardziej szkodliwe wydarzenie
- czynniki zw. ze stanem pojazu, które mogły być **przyczyną wypadku**
- pojawienie się pożaru
- ilość ofiar
- czy kierowca pił
- obecność kierowcy
- stan, gdzie wydano prawo jazdy
- kod pocztowy kierowcy
- wzrost kierowcy
- waga kierowcy
- poprzednie wypadki kierowcy (liczba)
- poprzednie zawieszenia prawa jazdy i skazania
- ilość mandatów za prędkość i innych
- daty pierwszych i ostatnich skazań / zawieszeń
- określenie stopnia prezkroczenia prędkości
- stan kierowcy (zaspany, depresja, chory, blackout, leki, narkotyki) - bardzo dużo możliwych wartości
- ograniczenie prędkości
- nawierzchnia i jej stan
- znaki, jakie napotkał pojazd przez wypadkiem
- sposób poruszania przed wypadkiem
- krytyczne wydarzenie przed wypadkiem

- manewr wymijający
- stabilność pojazdu przed wypadkiem (np. trzyma się drogi, poślizg)
- pozycja na drodze przed wypadkiem
- rodzaj wypadku

PERSON

- wiek
- płeć
- rodzaj (pieszy, zmotoryzowany, rowerzysta etc)
- poziom obrażeń
- zajmowane miejsce w pojeździe
- użycie pasów / hełmu
- czy powyższe były używane w odpowiedni sposób
- czy poduszka się otworzyła
- czy ciało wyleciało z pojazdu, i którędy
- czy osoba musiała być wydobyta przy użyciu sprzętu lub siły
- spożycie alkoholu
- sposób określenia spożycia
- czy test na nietrzeźwość był wykonany
- udział narkotyków
- czy była transportowana do szpitala
- śmierć na miejscu lub w drodze do szpitala
- data śmierci
- **przyczyny wypadku** (dla osób niebędących kierowcą)
- czas między wypadkiem a śmiercią
- rasa (pochodzenie)
- pozycja przed wypadkiem (niezmotoryzowani)

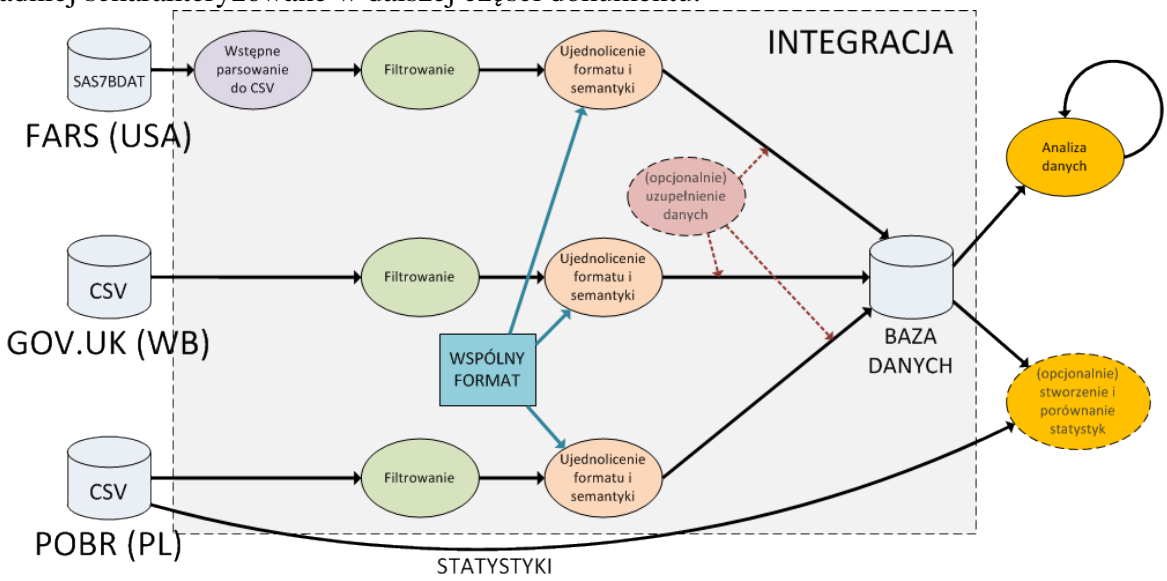
4. Metodyka

4.1. Cel dokumentu

Niniejszy dokument opisuje schemat postępowania w projekcie, poczynając od zebrania danych aż do ich analizy.

4.2. Schemat postępowania:

Poniższy schemat przedstawia zarys postępowania w projekcie, wstępnie określa jego fazy, które są dokładniej scharakteryzowane w dalszej części dokumentu.



4.3. Pobranie danych

W wyniku przeprowadzonej analizy źródeł danych, której rezultaty zostały przedstawione w dokumencie Analiza źródeł danych, wytypowane zostały źródła danych, które zostaną wykorzystane w projekcie. Pierwszą fazą realizacji projektu jest uzyskanie dostępu do danych, oraz pobranie ich ze źródeł. Wybrane dane pochodzą z trzech krajów.

- Polska - dane Polskiego Obserwatorium Ruchu Drogowego
- Wielka Brytania - dane z witryny gov.pl
- Stany Zjednoczone - dane z systemu FARS - Fatality Analysis Reporting System

W przypadku niektórych źródeł dane są dostępne do pobrania bezpośrednio ze strony lub serwera ftp, bez ograniczeń dostępu, jednak w przypadku danych z POBR konieczne jest uzyskanie dodatkowej autoryzacji w celu skorzystania z pełnych danych, więc dla tego źródła danych konieczny jest właśnie ten dodatkowy krok.

4.4. Filtrowanie danych

Dane ze źródeł muszą następnie zostać odfiltrowane. Jest to umotywowane zarówno dużą ilością danych jak i procesem preintegracji danych. Dodatkowym ograniczeniem są podjęte przez nas decyzje projektowe zawężające celowo zakres danych.

Filtrowanie danych będzie przebiegać na trzech płaszczyznach:

- stopień powagi wypadku: podjętą przez nas decyzją projektową chcemy się ograniczyć do wypadków śmiertelnych, będziemy więc gromadzić jedynie dane o takich wypadkach, odrzucając wypadki lekkie i poważne, ale bez skutków śmiertelnych
- czas (wybrane roczniki) w celach interacyjnych: chcemy ograniczyć się do roczników, z których dane są dostępnych we wszystkich źródłach w celu lepszych możliwości porównania
- czas (wybrane roczniki) w celach zapewnienia pełności danych: chcemy odrzucić dane z lat, dla których mamy ich zbyt mało - zakres atrybutów był zbyt wąski lub większość informacji jest niedostępna. Może być to szczególnie widoczne w danych starszych.

4.5. Uwspólnienie formatu i semantyki danych

W celu analizy danych zróżnych źródeł konieczne jest ustalenie wspólnego formatu danych. obejmuje on zarówno kwestie wspólnego formatu (np. wspólny format daty), ale również uwspólnienie semantyki - przykładem może byc opis warunków pogodowych: różne bazy przedstawiają te dane w różny sposób, różną gamą możliwych wartości atrybutu, konieczne jest ustalenie wspólnej listy wartości atrybutów, na którą następnie będzie można zmapować atrybuty z poszczególnych źródeł.

Ważnym zadaniem tego etapu jest selekcja atrybutów, które będą dla nas istotne w przeprowadzanych analizach, w szczególności selekcja atrybutów pozwalających wnioskować o przyczynach wypadków - kryteria analizy. Kryteria te możemy podzielić na następujące ogólne kategorie:

- data

- miejsce
- warunki pogodowe
- warunki środowiskowe
- dane o uczestnikach (np. młodzi / pijani kierowcy)
- dane o pojazdach (np. rocznik, systemy bezp.)

Dokładny opis szczegółowych kryteriów analizy można znaleźć w dokumencie Kryteria analizy.

4.6. Integracja

Integracja danych polegać będzie na sprowadzeniu danych do określonego wcześniej wspólnego formatu. W tym celu konieczne będzie sparsowanie danych przy pomocy skryptów w pythonie a następnie przetłumaczenie i zmianę formatu na wspólny.

W miarę potrzeby i możliwości będziemy również na tym etapie uzupełniać dane z dodatkowych źródeł - np. w razie braku danych o pogodzie dla dużej ilości wypadków, można z innych źródeł pobrać informacje o pogodzie w danym miejscu i czasie.

Tak przetłumaczone i uzupełnione dane zostaną zapisane do relacyjnej bazy danych PostgreSQL, zgodnie z opracowanym schematem bazy danych, który jest dokładnie opisany w dokumencie Projekt bazy danych.

4.7. Analiza danych

Kolejnym, kluczowym krokiem procesu jest analiza danych. Dane zgromadzone w bazie będziemy analizować na kilka sposobów.

Pierwszym z nich jest analiza statystyczna. Przy pomocy narzędzi do wizualizacji oraz narzędzi analizy statystycznej języka R czy platformy Weka będziemy analizować statystyczne własności danych takie jak korelacje czy procentowe udziały atrybutów. Głównymi atrybutami branymi pod uwagę w tej analizie będą atrybuty wskazane w dokumencie Kryteria analizy. W ramach analizy wizualnej będziemy rozważać wykresy generowane z danych a także nanosić dane na mapy w celu dokonania przestrzennej analizy danych.

Z bardziej zaawansowanych metod, chcemy spróbować poddać dane klasteryzacji w celu wskazania podobnych wypadków i podjęcia próby wskazania dla nich wspólnych cech. Dodatkową formą analizy danych będzie wyszukiwanie wzorców częstych - pozwoli ono wskazać często współwystępujące okoliczności wypadków i wysnuć wnioski co do ich wpływu na występowanie wypadków.

4.8. Opcjonalne kierunki analizy

Jeżeli po wykonaniu wymienionych już analiz zostanie nam nieco czasu, możemy podjąć się zrealizowania w projekcie opcjonalnych kierunków analizy. Będą one polegać na stworzeniu zbiorczych statystyk z danych przechowywanych w bazie i porównanie ich ze zbiorczymi danymi z regionów, dla których nie mamy dostępu do szczegółowych danych z wypadków. Chodzi tu szczególnie o dane z Polski, dla których nie mamy pewności, że uda nam się otrzymać do nich dostęp.

4.9. Przegląd narzędzi

W projekcie będą wykorzystywane następujące narzędzia i technologie:

- python [9]
  - skrypty filtrujące / parsujące / konwertujące
  - parser sas7bdat - [10]
- PostgreSQL - [11]
  - relacyjna baza danych do persystencji
- Narzędzia do analizy danych:
  - MS Excel
  - R - [12]
  - Weka - [13]
- Narzędzia do wizualizacji danych
  - Polymaps - [14]
  - Google Charts - [15]
  - OpenLayers - [16]

4.10. Referencje

1. POBR - [2]
2. FARS - [1]
3. Dane z Wielkiej Brytanii [5]

5. Kryteria analizy

5.1. Cel dokumentu

Niniejszy dokument ma za zadanie sprecyzowanie kreteriów analizy przyczyn wypadków. Prezentuje dane i atrybuty, na których skupimy się w naszej analizie, gdyż podejrzewamy, że mogą mieć wpływ na powstawanie wypadków drogowych

5.2. Kryteria analizy

5.2.1. data

– **pora dnia:** Należy dokonać rozróżnienia dzień/noc oraz bardziej szczegółowego - rano/popołudnie/... Przewidujemy, że noc może być czynnikiem powodującym zwiększenie ilości wypadków z powodu gorszej widoczności i gorszego stanu kierowcy(senność), z drugiej strony mniejsza liczba samochodów na drodze obniża ryzyko kolizji. Okolice godzin wieczornych mogą się również okazać okresem, gdzie narażenie na wystąpienie wypadku jest większe - zmrok powoduje znaczne pogorszenie widoczności, może również wystąpić senność a ruch nadal jest wzmożony.

– **pora roku:** Można wnioskować, że większa liczba wypadków będzie występować w zimie, z powodu cięższych warunków atmosferycznych, jednakże mogą one skutkować większą ostrożnością ukierowców.

5.2.2. miejsce

– **umiejscowienie względem skrzyżowania:** Należy zbadać, czy wypadki częściej występują na skrzyżowaniach czy poza nimi oraz w ramach wypadków na skrzyżowaniach jaka zależność występuje od rodzaju skrzyżowania i konkretnego miejsca na skrzyżowaniu. Skłaniamy się do opinii, że więcej wypadków będzie się zdarzało na skrzyżowaniach z powodu podwyższonego ryzyka błędu i przecinania się dróg pojazdów.

– **umiejscowienie względem przejścia dla pieszych:** Należy rozważyć jak częste są wypadki w okolicach przejścia dla pieszych. Podejrzewamy, iż obecność przejścia dla pieszych zwiększa wyzyko wypadku a w szegółności wypadku śmiertelnego z powodu dodatkowego czynnika jakim są piesi, różnicy w prędkości oraz braku zabezpieczeń.

– **współrzędne GPS:** Współrzędne mogą zostać wykorzystane do wizualizacji danych na mapie i przykładowo wnioskowaniu o większej liczbie wypadków w okolicach miast.

5.2.3. warunki pogodowe

– **opady (deszcz, śnieg):** Obecność opadów niewątpliwie wpływa na pogorszenie widoczności a także warunków na drodze co może się przyczyniać do częstszego występowania wypadków. Z drugiej strony, kierowcy przejawiają tendencję do wytężonej uwagi i ostrożniejszej jazdy w takich warunkach.

– **mgła:** Mgła jest czynnikiem zdecydowanie pogarszającym widoczność i sugerującym zwiększenie ryzyka wypadku.

– **wiatr:** Silny wiatr może powodować zwiększenie ryzyka wypadku.

5.2.4. warunki środowiskowe

– **oświetlenie:** może być kluczowym czynnikiem wpływającym na ilość wypadków, spodziewane jest znalezienie częstych wzorców zawierające niekorzystne oświetlenie i pogodę jako okoliczności wypadków.

– **stan nawierzchni:** posłuży do poszukiwania korelacji z danymi na temat poślizgu / braku przyczepności pojazdów oraz do wnioskowania czy stan nawierzchnii może mieć duży wpływ na ilość wypadków

– **rodzaj (klasa) drogi:** spodziewany jest wpływ tego czynnika na ilość wypadków oraz przewidywane jest znalezienie korelacji z takimi danymi jak prędkość poruszania i przekroczenie prędkości

5.2.5. dane o uczestnikach

– **wiek kierowcy:** ta informacja może posłużyć do znalezienia zależności pomiędzy wiekiem kierowcy a powagą wypadku, stopniem przekroczenia prędkości, ilością ofiar etc.

– **poprzednie wypadki kierowcy:** w celu zweryfikowania, czy osoby z historią wypadków i wykroczeń powodują więcej wypadków i czy uczą się na błędach

– **stan kierowcy:** informacje w rodzaju czy kierowca był zaspany / zmęczony / niedołężny etc. mogą posłużyć do bezpośredniego wnioskowania o przyczynach wypadków

- **alkohol:** pozwoli na zobrazowanie jak dużo i jak poważnych wypadków powodowanych jest przez pijanych kierowców
- **narkotyki:** informacje te pojawiają się rzadziej niż związane z alkoholem, ale mogą pozwolić na podobne wnioskowanie
- **typ uczestnika (pieszy, pasażer, . . .):** pozwoli uzyskać statyski, między innymi jak często w wypadkach biorą udział osoby niezmotoryzowane, ile wśród nich jest ofiar, które siedzenia w samochodzie są najmniej bezpieczne
- **użycie pasów bezpieczeństwa:** czy i jak wpływa na stopień obrażeń
- **otwarcie poduszek powietrznych:** czy i jak wpływa na stopień obrażeń
- **użycie kasku:** czy i jak wpływa na stopień obrażeń

5.2.6. dane o pojazdach

- **rok produkcji:** może pozwolić na szukanie wpływu na ilość i powagę wypadków, ilość ofiar
- **marka:** do celów stastycznych, możliwe jest że posiadacze niektórych marek (Samochody sportowe) powodują więcej wypadków
- **model:** do celów stastycznych, możliwe jest że posiadacze niektórych marek (Samochody sportowe) powodują więcej wypadków
- **typ pojazdu (np. ciężarówka):** podobnie jak wyżej, może posłużyć do znalezienia prawidłowości jakie rodzaje samochodów częściej biorą udział w wypadkach
- **prędkość poruszania się:** istotna informacja w kontekście korelacji z powagą wypadku, ilością ofiar
- **przekroczenie prędkości:** istotna informacja w kontekście korelacji z powagą wypadku, ilością ofiar

5.3. Hipotezy do weryfikacji

W ramach analizy, chcemy nie tylko przeanalizować wpływ pojedynczych atrybutów na występowanie wypadków, ale także zweryfikować bardziej zaawansowane hipotezy dotyczące wpływu złożonych czynników na bezpieczeństwo na drodze.

5.3.1. Ograniczenie widoczności

Czynniki ograniczające widoczność powinny mieć duży wpływ na wzrost liczby wypadków. W szczególności groźne są kombinacje takich czynników, przykładowo, niezwykle groźnymi warunkami

na drodze są połączenie noc i mgła, czy noc i mgła i deszcz. Można w tę analizę włączyć jeszcze stan oświetlenia - brak oświetlenia na drodze może dodatkowo pogarszać warunki.

Dodatkowo można rozważyć czy warunki ograniczenia widoczności nie powodują większej liczby wypadków z udziałem pieszych. Piesi są najmniej uprzywilejowanymi uczestnikami ruchu i są też w trudnych warunkach najmniej widoczni, szczególnie w wypadku braku odblasków.

5.3.2. Niesprzyjające warunki atmosferyczne

Deszcz lub śnieg albo ich połączenie są groźnymi warunkami do jazdy. Dodatkowo silny wiatr może sprawić, iż kierowca ma ograniczoną kontrolę nad samochodem. Należy jednak sprawdzić, czy fakt, że w trudnych warunkach kierowcy jeżdżą zdecydowanie ostrożniej i nie decydują się na brawurowe zachowania tak często jak w dobrych warunkach nie sprawia, że wypadków tych nie jest tak dużo więcej jak można by się spodziewać.

5.3.3. Złe warunki a przekraczanie prędkości

Interesującą analizą może być sprawdzenie jak często wypadki są powodowane w niesprzyjających warunkach (atmosferycznych i oświetleniowych) dodatkowo z przekroczeniem prędkości przez kierowcę. Należy to porównać z wypadkami w warunkach sprzyjających.

5.3.4. Rozkład atrybutów w różnych przedziałach czasowych

Ciekawą analizą do wykonania wydaje się rozkład czasowy wartości niektórych atrybutów. Rozważane przedziały czasowe mogą być różne:

- pora dnia
- dzień tygodnia
- pora roku
- rozkład dzienny w ciągu roku

Porównanie można przeprowadzać między innymi pod następującymi względami:

- procent wypadków z przekroczeniem prędkości
- procent wypadków, gdzie kierowca miał alkohol we krwi
- średnia liczba poszkodowanych w wypadku
- procent wypadków z udziałem pieszych
- liczba wypadków

Analiza taka może przynieść kilka bardzo ciekawych wniosków. Przykładowo badanie rozkładu dziennego w ciągu roku może pozwolić wskazać święta w trakcie których zwiększa się ilość wypadków bądź wypadków pod wpływem alkoholu. Przewidujemy, że więcej nieostrożnych i brawurowych kierowców (alkohol i przekraczanie prędkości) może być wieczorami czy w weekendy. Ciekawe może być też porównanie liczby wypadków pomiędzy porami roku, gorsze warunki zimowe powinny sprawić, ale być może jest to zrównoważone przez wzmożoną ostrożność i ograniczenie wyjeżdżania autem do minimum. Kryzysowe mogą się okazać np.okresy przejściowe między jesienią a zimą.

5.4. Zaawansowane formy analizy

W ramach projektu, chcemy zastosować także ekstrakcji wiedzy z danych, bez stawiania wcześniejszych hipotez. Cel ten możemy osiągnąć na dwa sposoby.

Pierwszą możliwością jest przeprowadzenie klasteryzacji danych o wypadkach. Otrzymując wyniki takiej klasteryzacji, możemy przeanalizować podobieństwa pomiędzy wypadkami znajdującymi się w jednym klastrze i próbować wyciągnąć cechy reprezentatywne takich wypadków i dołączyć do analizy liczność klastrów. Można także taką analizę przeprowadzić dla reprezentantów klastrów. Istnieje ryzyko, że taka analiza może okazać się zbyt skomplikowana i dać jedynie ograniczoną liczbę istotnych wniosków.

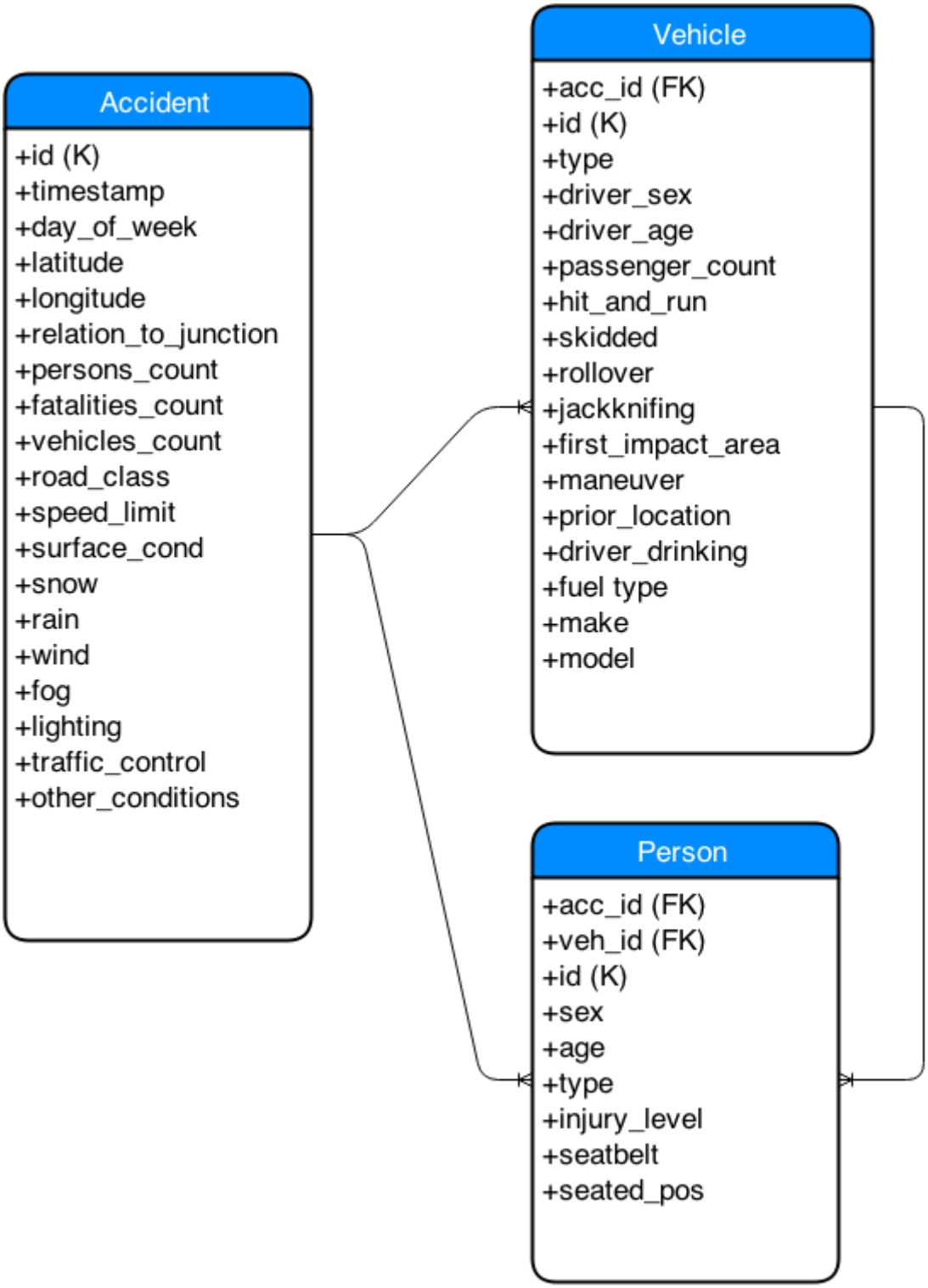
Drugim sposobem jest przeprowadzenie na danych procesu ekstrakcji wzorców częstych. Pozwoli to na wyłonienie wartości atrybutów często występujących wspólnie w czasie wypadków. Może byś to źródłem niezwykle ciekawych wniosków. Należy wziąć pod uwagę, ze taka analiza jest bardzo intensywna obliczeniowo i może być trudna do przeprowadzenia na całości danych, należy wtedy rozważyć możliwość przeprowadzenia kilku analiz na podzbiorze danych i połączenie pośrednich wyników analiz w wynik całościowy.

6. Projekt bazy danych

6.1. Cel dokumentu

Niniejszy dokument przedstawia schemat bazy danych wykorzystywanej w projekcie. Jest to jednocześnie określenie wspólnego formatu danych wykorzystywanego w procesie integracji. Pola tabel wraz z opisem formatu i ich dopuszczalnych wartości określają wspólny format i semantykę danych.

6.2. Schemat bazy danych



6.3. Opis tabel

6.3.1. Accident

Tabela zawierająca dane i okoliczności dotyczące wypadku

Opis pól:

- id - INT, klucz główny tabeli
- timestamp - TIMESTAMP
- day_of_week - INT, dzień tygodnia
- latitude - DECIMAL, szerokość geograficzna - dla USA dane dopiero od 1999
- longtitude - DECIMAL, długość geograficzna - dla USA dane dopiero od 1999
- relation_to_junction - STRING
  - NON_JUNCTION
  - INTERSECTION
  - DRIVEWAY - podjazd / prywatna droga
  - RAMP - wjazd / zjazd z autostrady
  - UNKNOWN
- persons_count - INT, liczba uczestników w wypadku
- fatalities_count - INT, liczba ofiar śmiertelnych wypadku
- vehicles_count - INT, liczba pojazdów biorących udział w wypadku
- road_class - STRING, rodzaj drogi
  - MOTORWAY - amerykańskie Highway
  - PRINCIPAL - brytyjska klasa A, amerykańskie Principal Arterial
  - MAJOR - brytyjska B
  - MINOR - brytysjka C
  - UNCLASSIFIED
  - UNKNOWN
- speed_limit - INT, wartość ograniczenia prędkości na drodze [km/h]
- surface_cond - STRING, stan nawierzchni
  - DRY
  - WET
  - SNOW
  - ICE
  - FLOOD
  - OTHER (oil, mud, sand, gravel)
  - UNKNOWN
- snow - STRING, śnieg
  - YES - dangerous wind conditions

    - NO
    - UNKNOWN
- rain - STRING, deszcz
    - YES - dangerous wind conditions
    - NO
    - UNKNOWN
- wind - STRING, wiatr
    - YES - dangerous wind conditions
    - NO
    - UNKNOWN
- fog - mgła
    - YES - fog, smoke or smog
    - NO
    - UNKNOWN
- lighting - warunki oświetlenia
    - DAYLIGHT
    - DARK_LIGHTED
    - DARK
    - UNKNOWN
- traffic_control
    - TRAFFIC_SIGNAL
    - SIGNAL_MALF - nie działające światła
    - STOP_SIGN
    - AUTH_PERSON - osoba uprawniona do kierowania ruchem
    - NONE - jeżeli wypadek nie na skrzyżowaniu
    - UNKNOWN

6.3.2. Vehicle

Tabela zawierająca dane na temat pojazdów biorących udział w wypadku i ich kierowców. Jest powiązana relacją n:1 z tabelą *Accident* poprzez klucz obcy *acc*id_ - w jednym wypadku może brać udział wiele pojazdów

Opis pól:

- acc_id - INT, klucz obcy z tabeli *Accident* realizujący relację 1:n, wypadek w jakim brał udział dany pojazd
- id - INT, klucz główny tabeli
- type - STRING
  - CAR
  - MOTORCYCLE
  - BUS
  - CARGO
  - AGRICULTURAL
  - OTHER
  - UNKNOWN
- driver_sex - STRING, płeć kierowcy
  - MALE
  - FEMALE
  - UNKNOWN
- driver_age - INT, wiek kierowcy
- passenger_count - INT, liczba pasażerów
- hit_and_run - STRING, czy kierujący pojazdem uciekł z miejsca wypadku
  - YES
  - NO
  - UNKNOWN
- skidded - STRING, okoliczności dotyczące poślizgu - dla USA dane te występują tylko w latach 2010 - 2013.
  - YES
  - NO
  - UNKNOWN
- rollover - STRING, okoliczności dotyczące dachowania
  - YES
  - NO
  - UNKNOWN
- jackknifing - STRING, okoliczności dotyczące jackknifingu
  - YES

  - NO
  - UNKNOWN
- first_impact_area - STRING, pierwsze miejsce uderzenia pojazdu
  - FRONT
  - BACK
  - LEFT_SIDE
  - RIGHT_SIDE
  - NON_COLLISION
  - UNKNOWN
- maneuver - STRING - dla USA dane te występują tylko w latach 1982 - 2008.
  - STRAIGHT
  - PARKED
  - REVERSING
  - U_TURN
  - LEFT
  - RIGHT
  - CHANGING_LANE
  - OVERTAKING
  - HELD_UP
  - STOPPING
  - STARTING
  - CURVING
  - UNKNOWN
- driver_drinking - STRING, obecność alkoholu we krwi kierowcy
  - YES
  - NO
  - UNKNOWN
- fuel_type - STRING, rodzaj paliwa - dla USA, do roku 2009, te dane zbierane były tylko dla ciężarówek.
  - DIESEL
  - PETROL

  - HYBRID
  - GAS
  - OTHER
  - UNKNOWN

6.3.3. Person

Tabela zawierająca dane na temat uczestników wypadku. Powiązana relacją n:1 z tabelą *Accident* poprzez klucz obcy *acc*id_ - w jednym wypadku może być wiele ofiar. Powiązana relacją n:1..0 z tabelą *Vehicle* poprzez klucz obcy *veh*id_ - w jednym pojeździe może się znajdować wielu uczestników, uczestnik może się znajdować w jednym pojeździe, lub być pieszym i nie znajdować się w żadnym pojeździe.

Opis pól:

- acc_idd - INT, klucz obcy z tabeli *Accident* realizujący relację 1:n, wypadek w jakim brał udział dany uczestnik
- veh_id - INT, klucz obcy z tabeli *Vehicle* realizujący relację 0..1:n, wypadek w jakim brał udział dany uczestnik. Może zawierać wartość NULL.
- id - INT, klucz główny tabeli
- sex - STRING, płeć
  - MALE
  - FEMALE
  - UNKNOWN
- age - INT, wiek
- type - STRING
  - DRIVER
  - PASSENGER
  - PEDESTRIAN
  - UNKNOWN
- injury_level - STRING, powaga obrażeń
  - FATAL
  - SERIOUS
  - SLIGHT
  - NONE
  - UNKNOWN

- seatbelt - STRING, użycie pasów bezpieczeństwa
  - NOT_APPLICABLE
  - WORN_CONFIRMED
  - WORN_NOT_CONFIRMED
  - NOT_WORN
  - UNKNOWN
- seated_pos - STRING, miejsce siedzenia w pojeździe
  - DRIVER
  - PASSENGER
  - BACK
  - NONE
  - UNKNOWN

7. Problemy podczas integracji

Na tej stronie zebrano opis problemów i trudności napotkanych podczas parsowania i integrowania danych.

Problem:
Niewielka cześć wypadków z USA nie posiada określonego czasu wystąpienia.

Rozwiązanie:
Odrzucić wypadki z brakującymi danymi, ponieważ stanowią bardzo małą część rekordów (< 0.1%).

Problem:
Dane z USA publikowane są w różnych formatach w zależności od roku wystąpienia.

Rozwiązanie:
Poszukać bibliotek dla każdego używanego formatu i użyć ich do sprowadzenia danych do jednego formatu (CSV).

Problem:
Znaczenie wartości atrybutów w danych z USA zmienia się w zależności od roku wystąpienia (np. wartość “4” w polu “ATMOSPHERIC CONDITIONS” znaczy “deszcz” w latah 1975 - 1990, natomiast w późniejszych latach “wiatr”). Część atrybutów przestaje być wspierana po jakimś czasie, natomiast część pojawia się dopiero w ostatnich latach. Ponadto, niektóry atrybuty przenoszone są pomiędzy różnymi bazami danych.

Rozwiązanie:
Zaprojektować mechanizm, który pozwoli na wygodne mapowanie wartości atrybutów w zależności od roku wystąpienia i bazy źródłowej. Schemat mapowania mógłby być zapisywany w pliku, co pozwoli na łatwe parsowanie wszystkich danych po stworzeniu odpowiednich schematów.

Problem:
Bardzo duża ilość danych w zestawach z USA i GB.

Rozwiązanie:
Umieścić w bazie część danych z możliwie największym przekrojem jeśli chodzi o czas wystąpienia wypadku. W tym celu można odrzucić część wypadków z każdego roku, ale w taki sposób, aby uzyskać przykłady rekordów z różnych miesięcy.

Problem:
Część rekordów posiada sporo nieokreślonych wartości atrybutów.

Rozwiązanie:
Ustalić limit nieokreślonych wartości, poniżej którego dane będą odrzucane. Należy kierować się tym, aby zostawić wypadki z wystarczającą do wnioskowania ilością informacji. Limit nie może być zbyt wysoki, aby nie spowodować odrzucenia dużej części danych.

Problem:
Część atrybutów posiada bardzo duży zakres możliwych wartości, np. marka i model samochodu.

Rozwiązanie:
Zastosować jakąś metodę przechowywania możliwych wartości, np. w plikach tekstowych albo tabeli bazy danych, aby łatwo było się do nich odnosić podczas parsowania danych.

Problem:
Dane z Wielkiej Brytanii nie mają dokładnego wieku osób, tylko zakwalifikowanie do przedziału wiekowego. Ponadto okazuje się że może danych o wieku nie być w ogóle.

Rozwiązanie:
Zastąpić przedział wiekowy reprezentatywną wartością - np. średnią, albo losować wartość z przedziału wiekowego z rozkładem jednostajnym. Tam gdzie wiek nie jest dostępny, zapisujemy do bazy wiek ujemny, tak aby było wiadomo, iż brak tu danych.

Problem:

Dane z Wielkiej Brytanii nie mają danych o wszystkich pasażerach. Niektóre pojazdy posiadają według dostępnych danych 0 pasażerów.

Rozwiązanie:

Ignorować wartość tego pola dla danych z Wielkiej Brytanii. Niemożliwe jest uzupełnienie tych danych a szkoda stracić wartość badawczą jaką ten atrybut daje dla danych z USA.

8. Wyniki analiz

8.0.4. Cel dokumentu

Niniejszy dokument ma na celu przestawienie wyników przeprowadzonych analiz danych.

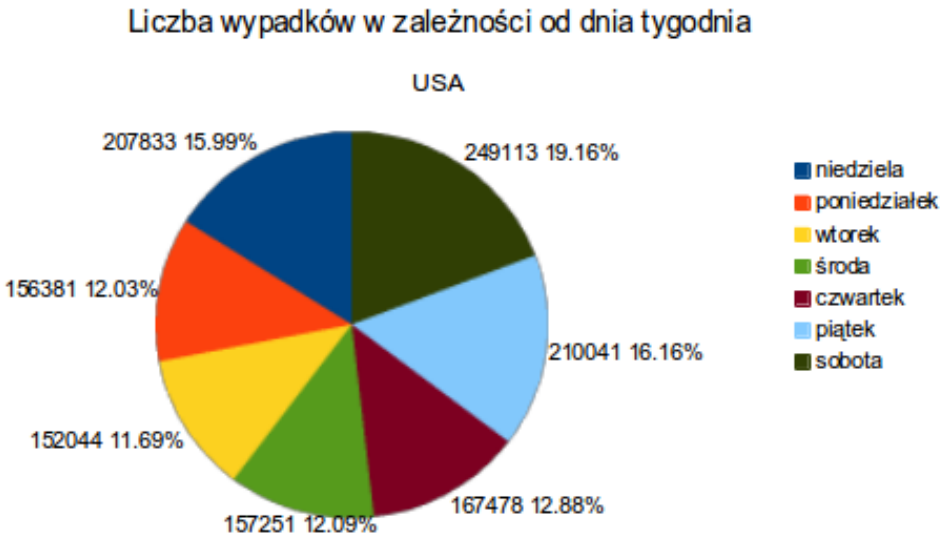
8.0.5. Proste statystyki

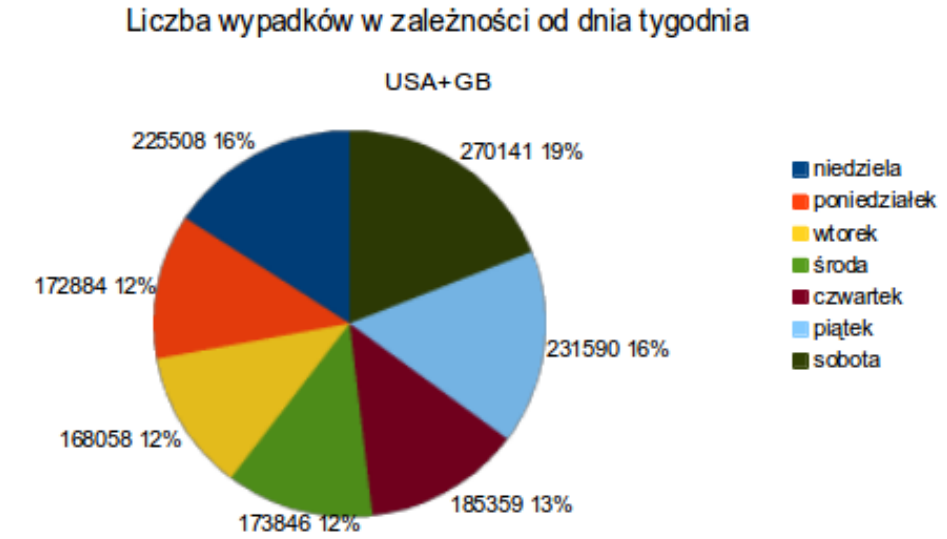
Poniższa tabela przedstawia proste statystyki liczbowe dotyczące zgromadzonych danych. Widzimy dysproporcję pomiędzy ilośćią danych pochodzących z USA a ilością danych z Wielkiej Brytanii. Wynika z niej podejście do analizy, w którym przeprowadzamy badania osobno na dancyh z USA, osobno z Wielkiej Brytanii a następnie na danych połączonych. Należy jednak mieć na uwadze, że dane połączone są mocno zdominowane przez dane z USA.

Ilość danych	Wypadki	Pojazdy	Uczestniczący	Ofiary
USA	1 300 141	1 947 730	3 430 324	1 450 505
GB	127 245	217 595	253 528	138 674
USA + GB	**1 427 386**	**2 183 616**	**3 683 852**	**1 589 179**

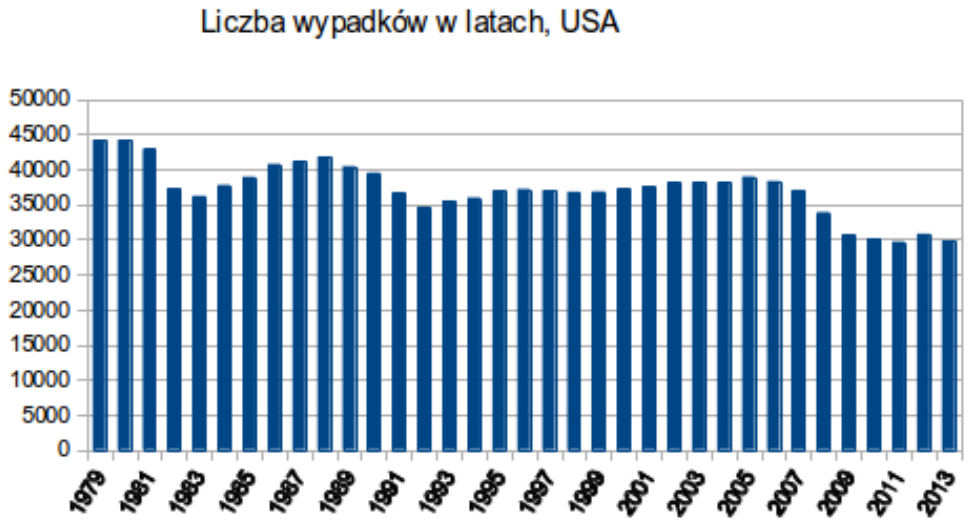
Analiza wypadków w czasie

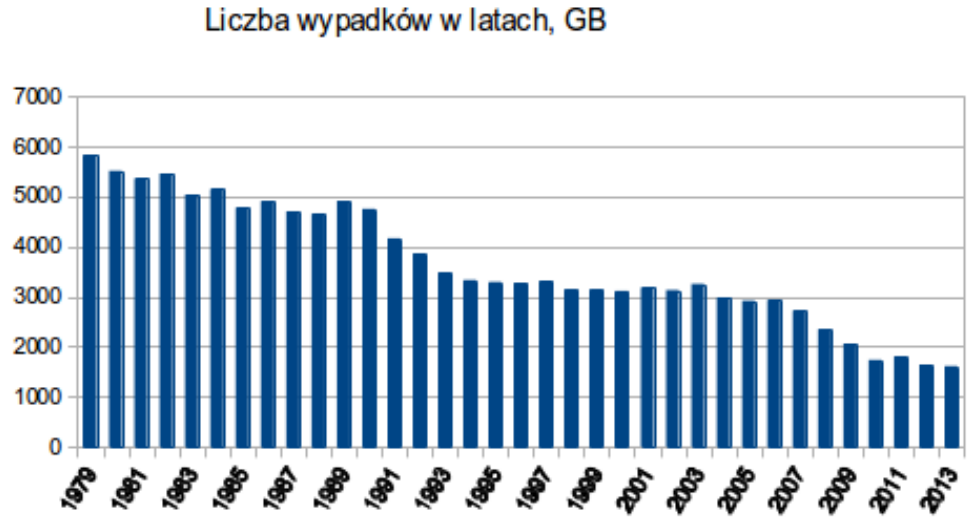
Dzień tygodnia:

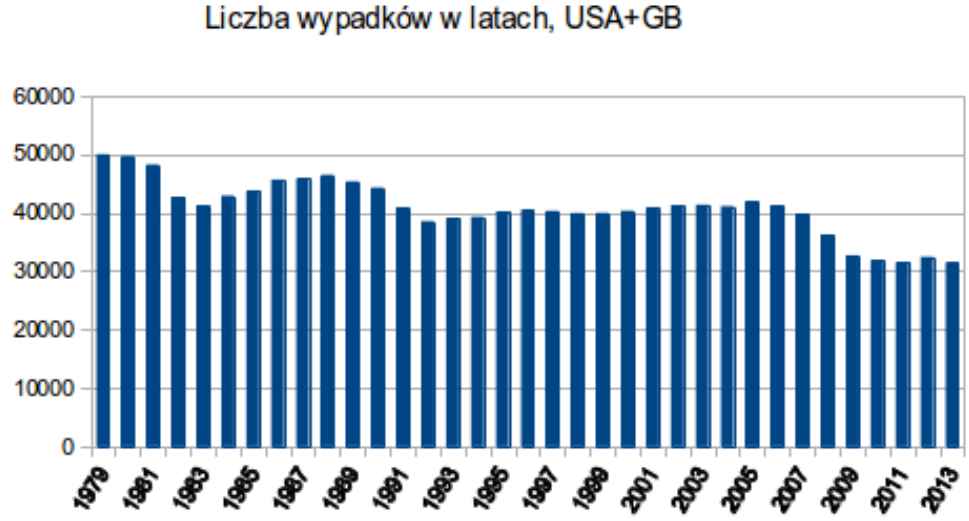




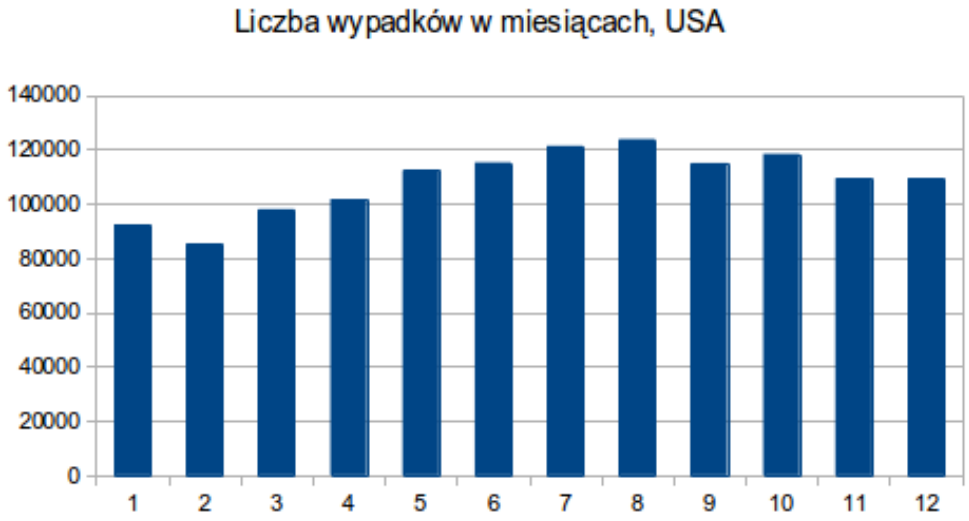
Rok:

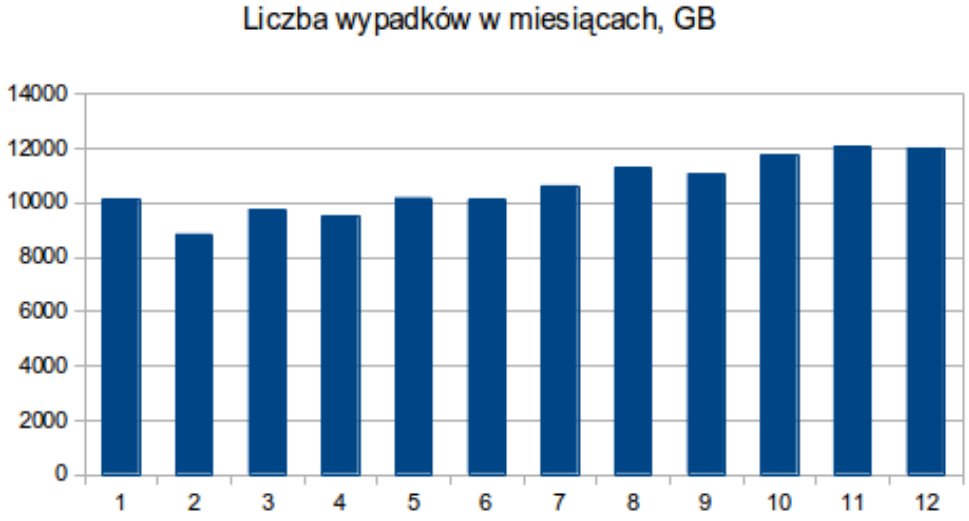

Liczba wypadków w latach, USA
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013

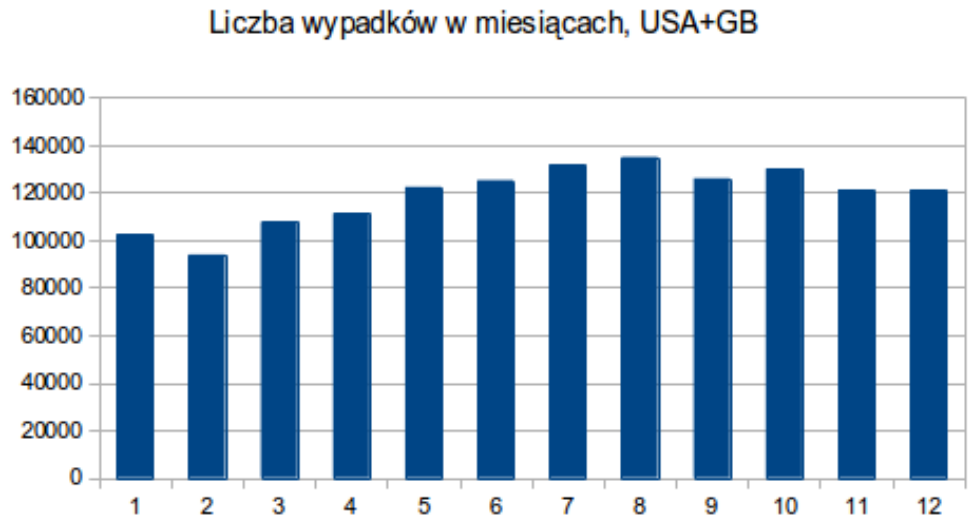

Liczba wypadków w latach, GB
7000
6000
5000
4000
3000
2000
1000
0
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013


Liczba wypadków w latach, USA+GB
60000
50000
40000
30000
20000
10000
0
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013

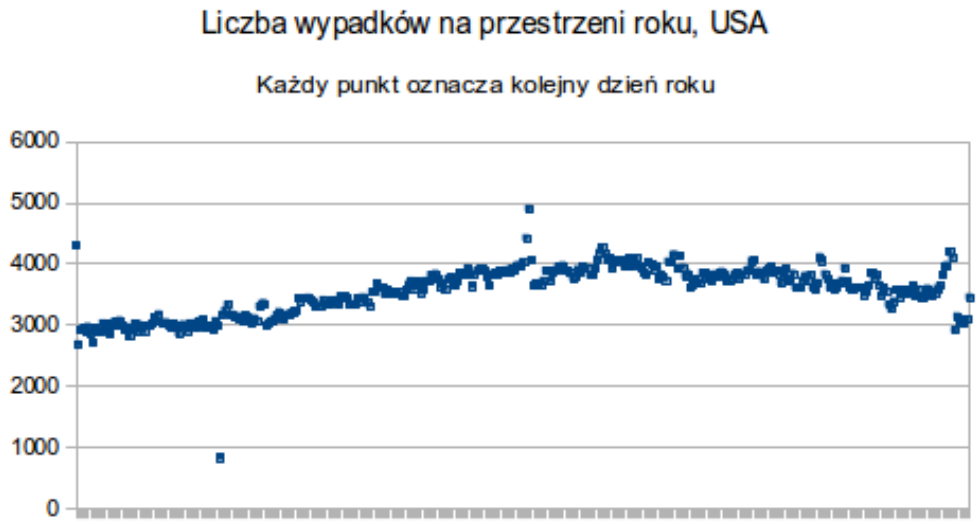
Miesiąc:

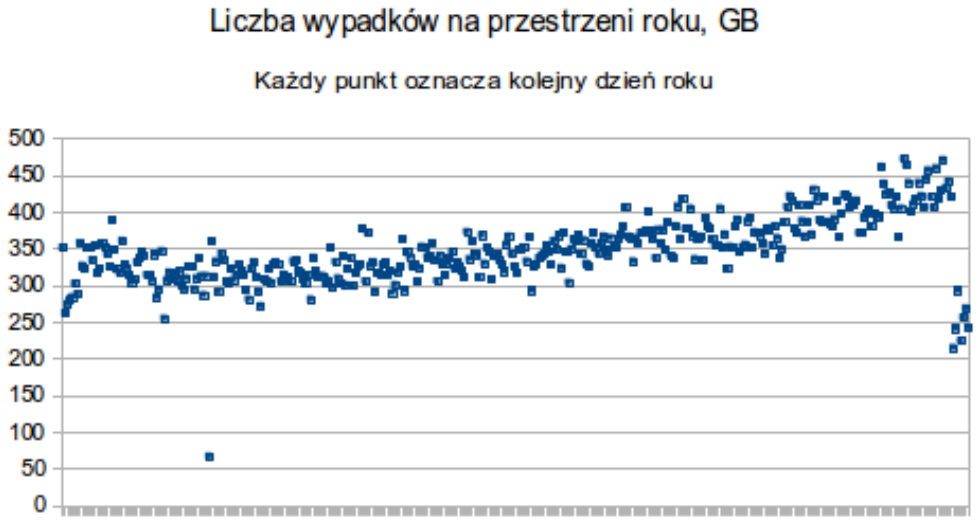

Liczba wypadków w miesiącach, USA
140000
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
11
12

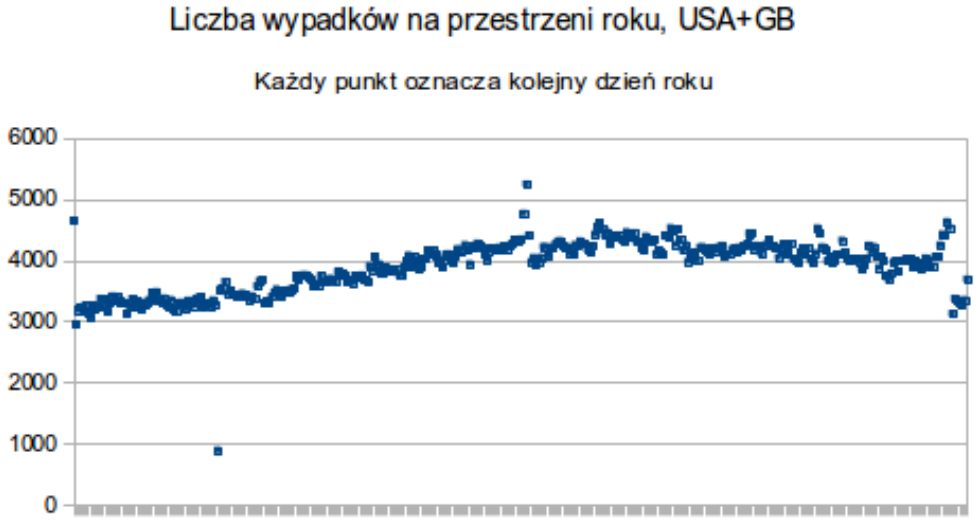

Liczba wypadków w miesiącach, GB
14000
12000
10000
8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
9
10
11
12


Liczba wypadków w miesiącach, USA+GB
160000
140000
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
11
12

Dzień:







Bardzo niska wartość jest dla 29.02 - występuje tylko w latach przestępnych i istotnie jest to wartość średnio 4 razy mniejsza.

Lokalne “piki” - **USA**:

- 01.01
- 03.07
- 04.07
- 02.08
- 03.08
- 31.10
- 01.11
- 20.12
- 21.12
- 22.12
- 23.12
- 24.12
- 31.12

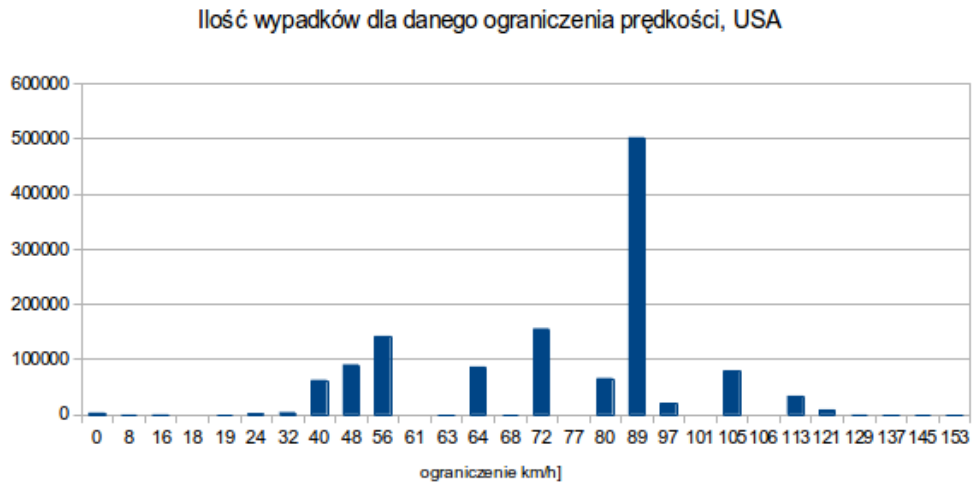
Lokalne “piki” - **GB**:

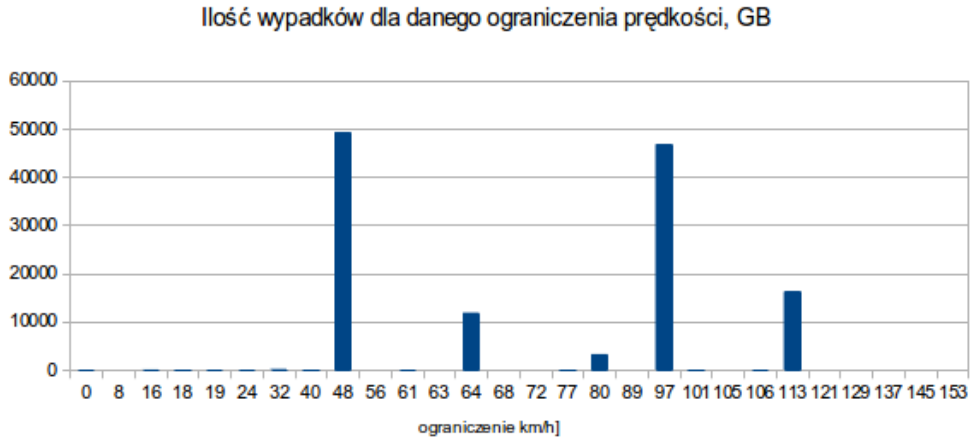
- 21.01
- 01.03
- 18.04
- 01.05
- 15.08
- 07.09
- 20.10
- 30.10
- 05.12
- 21.12

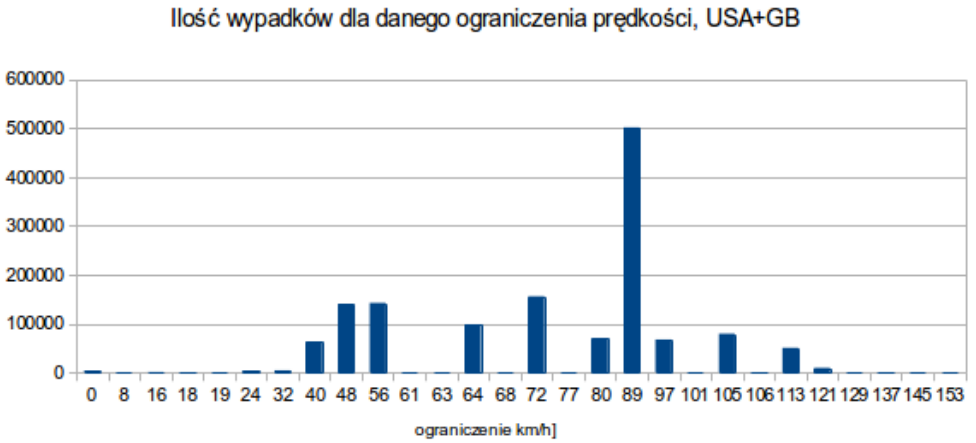
Ciekawy jest spadek liczby wypadków w ostatnich dniach roku - spowodowany prawdopodobnie faktem, że w tym okresie znacznie mniej osób jeździ do pracy i spędza więcej czasu w domu z rodziną.

Godzina:

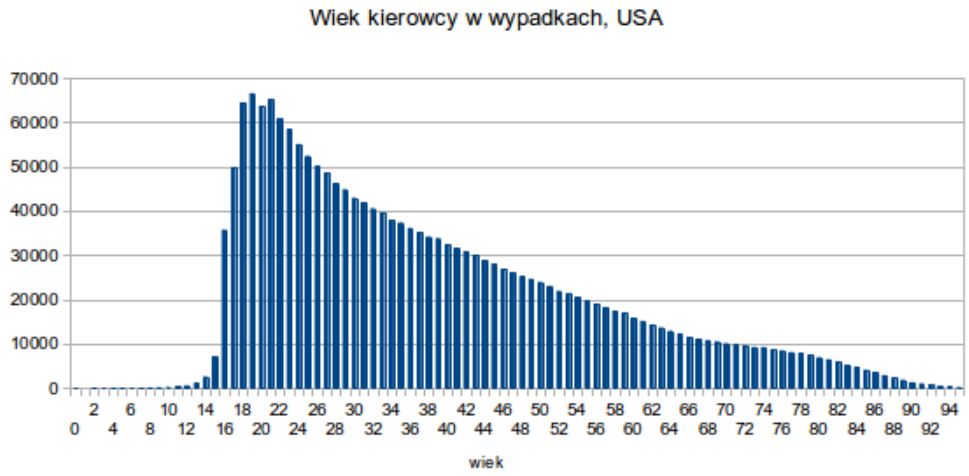
Ograniczenie prędkości

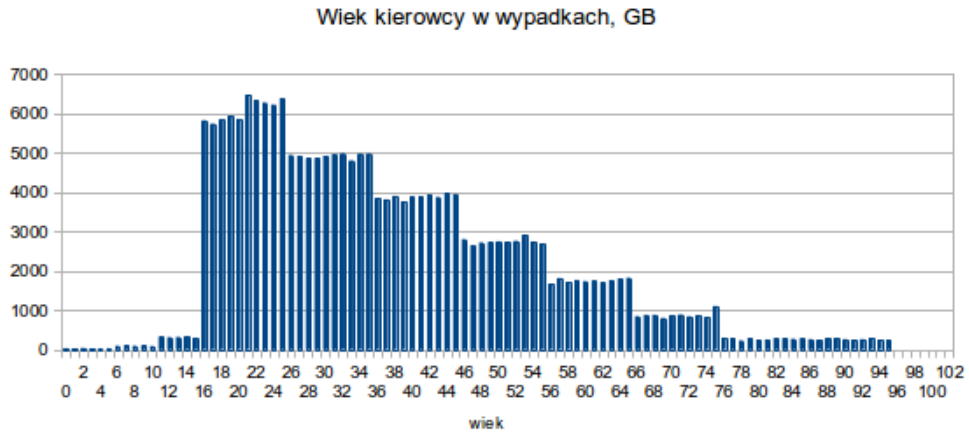


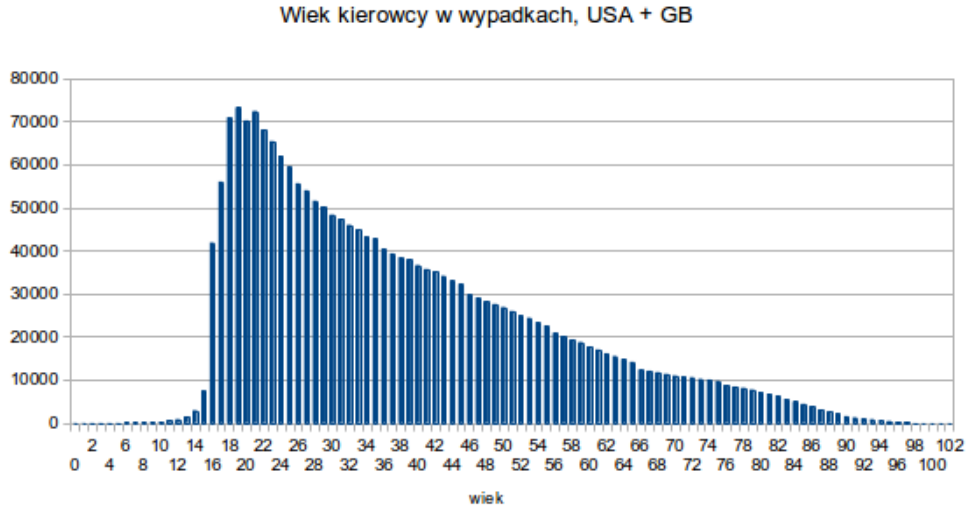




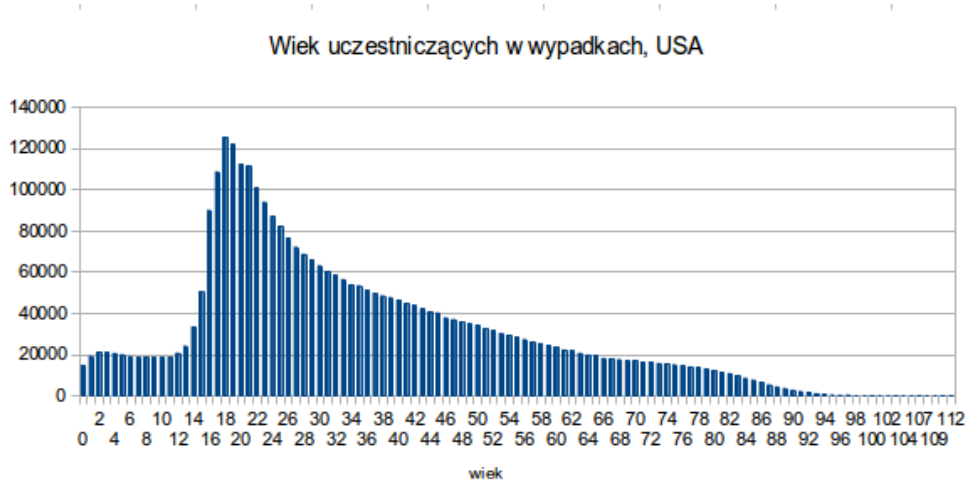
Wiek kierowcy

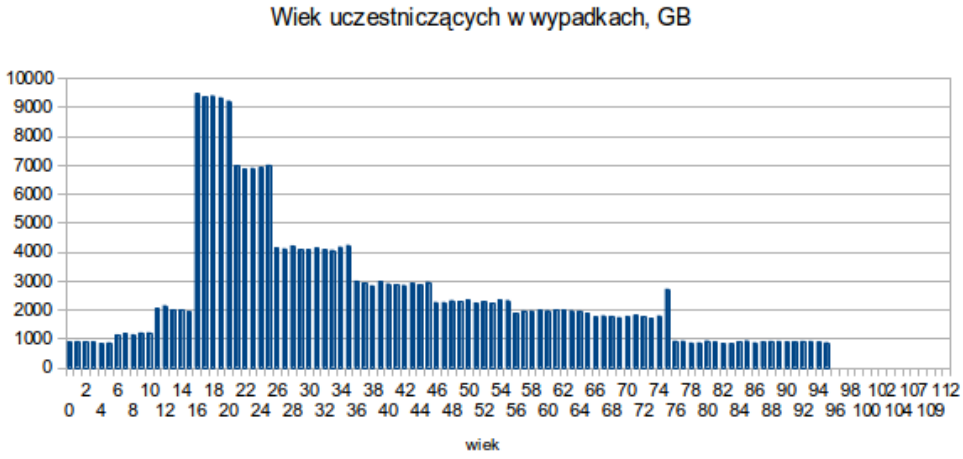


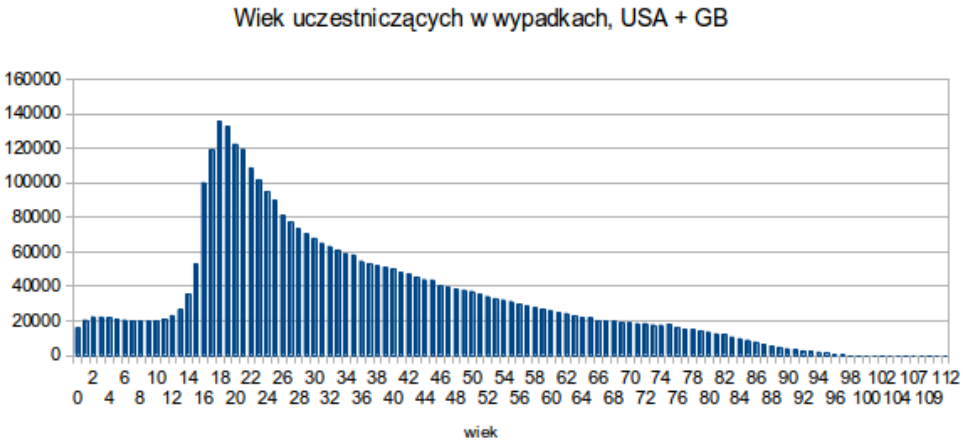




Wiek uczestników wypadku

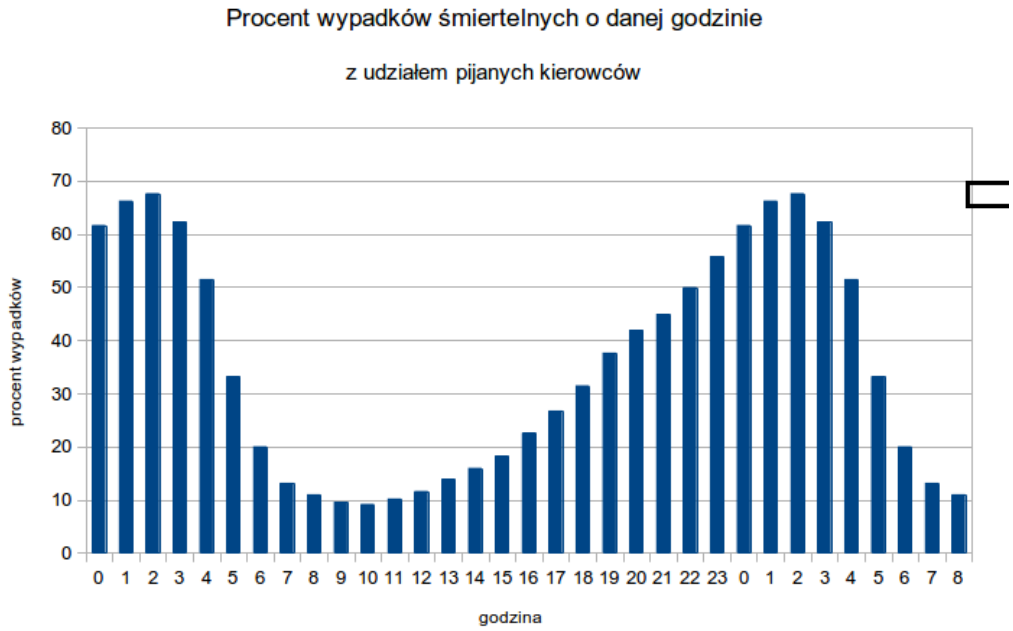


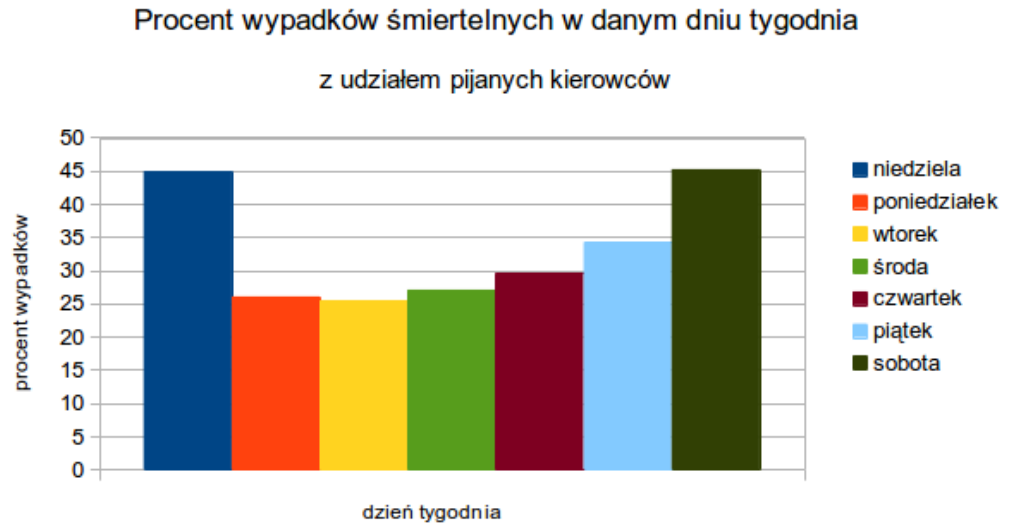


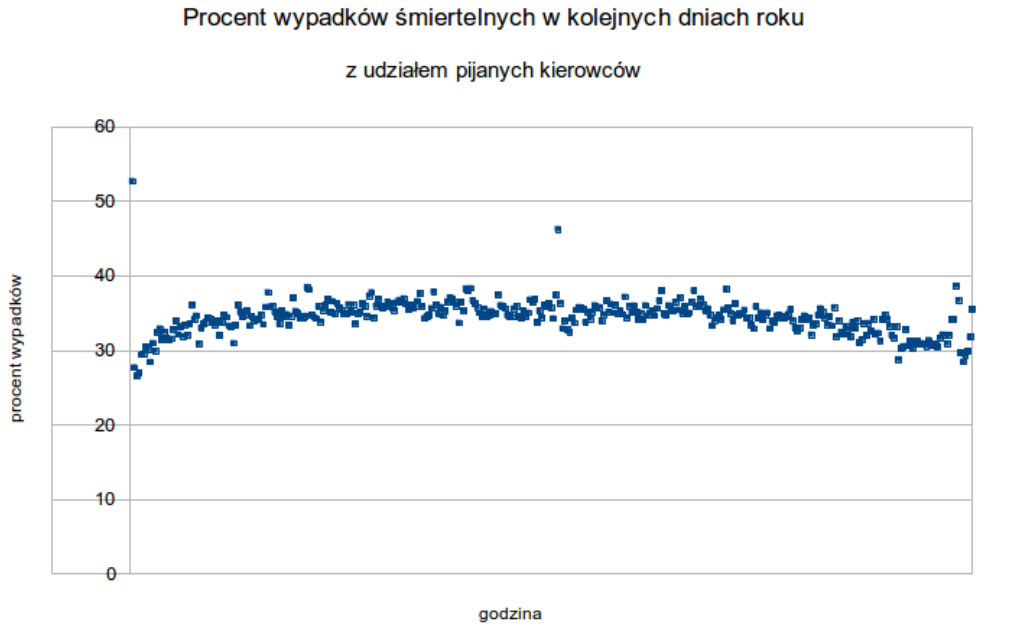


Kierowcy pod wpływem alkoholu

Dane o obecności alkoholu we krwi kierowcy są dostępne jedynie dla danych z USA.







Warunki pogodowe

Udział wypadków z różnymi kombinacjami warunków pogodowych

S - Snow

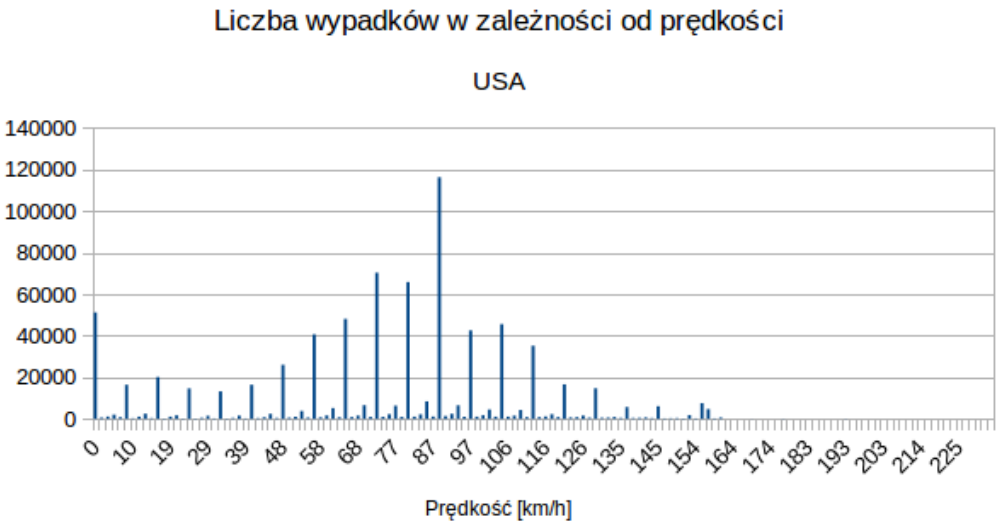
W - Wind

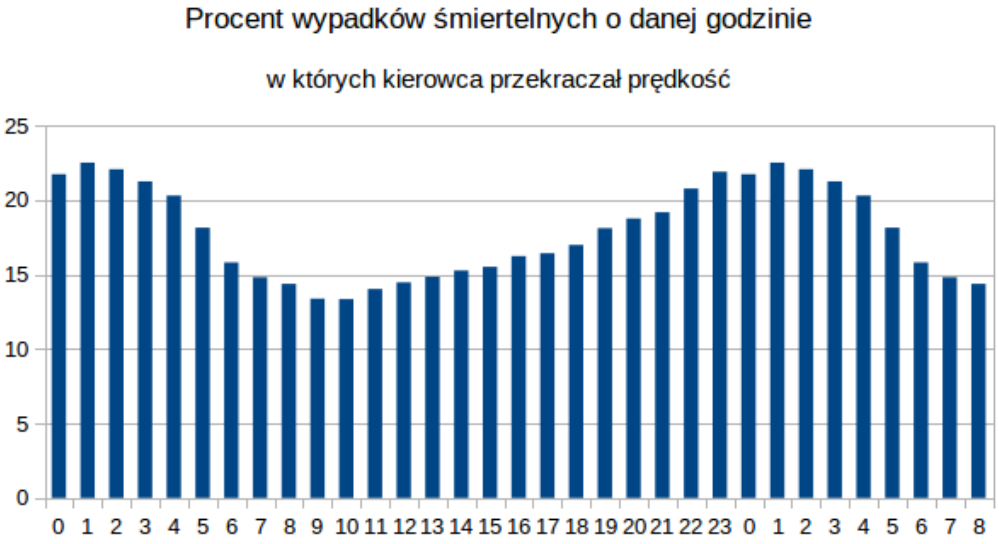
R - Rain

F - Fog

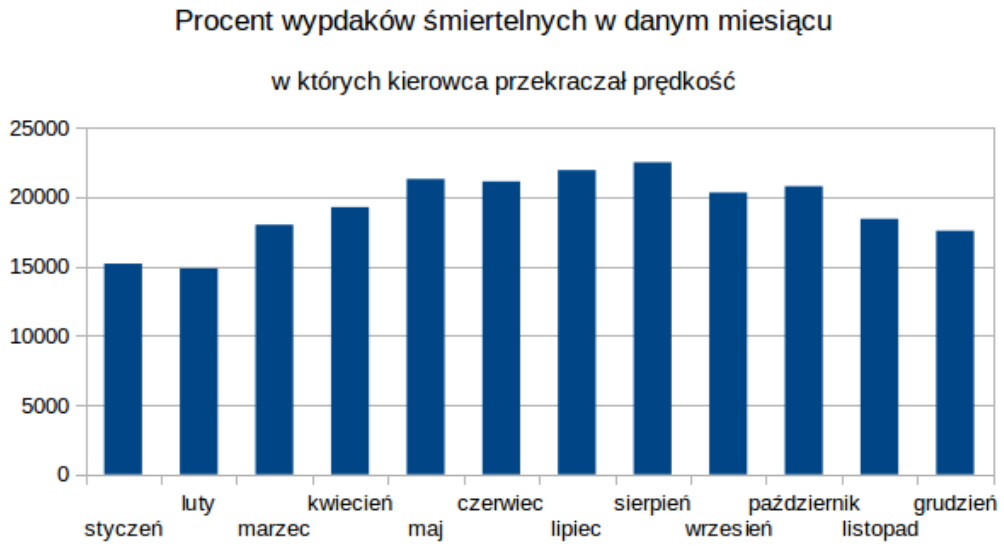
CONDS	USA	USA %	GB	GB %	USA+GB	USA+GB %
NONE	1142653	88,17%	104834	82,98%	1247487	87,71%
F	17222	1,33%	1306	1,03%	18528	1,30%
W	445	0,03%	2798	2,21%	3243	0,23%
WF	2	0,00%	0	0,00%	2	0,00%
R	108652	8,38%	14463	11,45%	123115	8,66%
RF	1444	0,11%	0	0,00%	1444	0,10%
RW	65	0,01%	2177	1,72%	2242	0,16%
RWF	0	0,00%	0	0,00%	0	0,00%
S	19108	1,47%	565	0,45%	19673	1,38%
SF	7	0,00%	0	0,00%	7	0,00%
SW	2005	0,15%	191	0,15%	2196	0,15%
SWF	6	0,00%	0	0,00%	6	0,00%
SR	4283	0,33%	0	0,00%	4283	0,30%
SRF	12	0,00%	0	0,00%	12	0,00%
SRW	78	0,01%	0	0,00%	78	0,01%
SRWF	0	0,00%	0	0,00%	0	0,00%

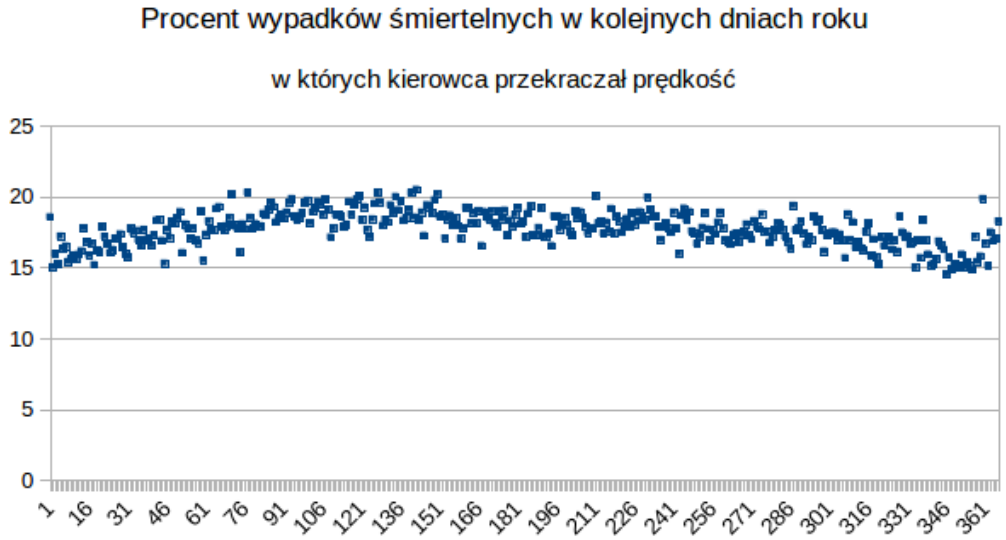
Prędkość


Liczba wypadków w zależności od prędkości
USA
140000
120000
100000
80000
60000
40000
20000
0
0 10 19 29 39 48 58 68 77 87 97 106 116 126 135 145 154 164 174 183 193 203 214 225
Prędkość [km/h]

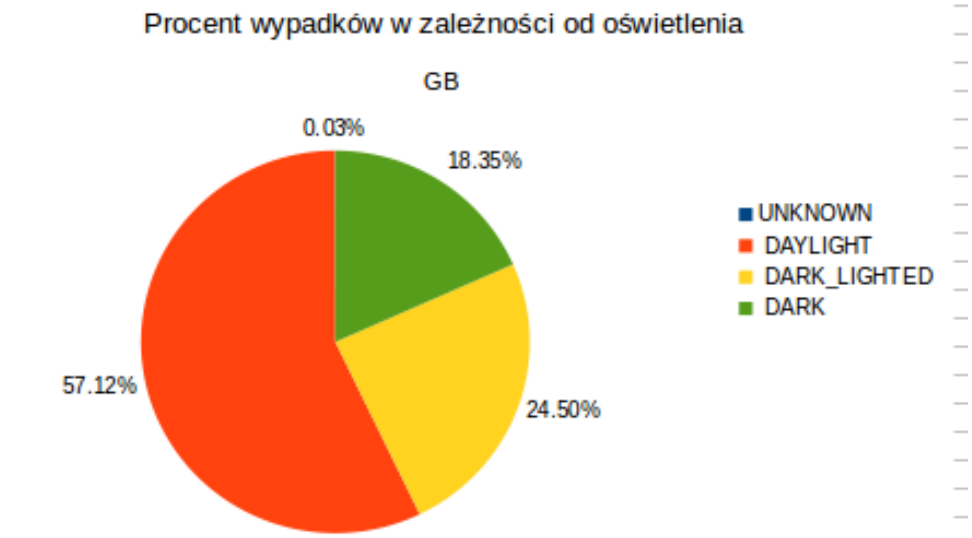

Procent wypadków śmiertelnych o danej godzinie
w których kierowca przekraczał prędkość
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 0 1 2 3 4 5 6 7 8

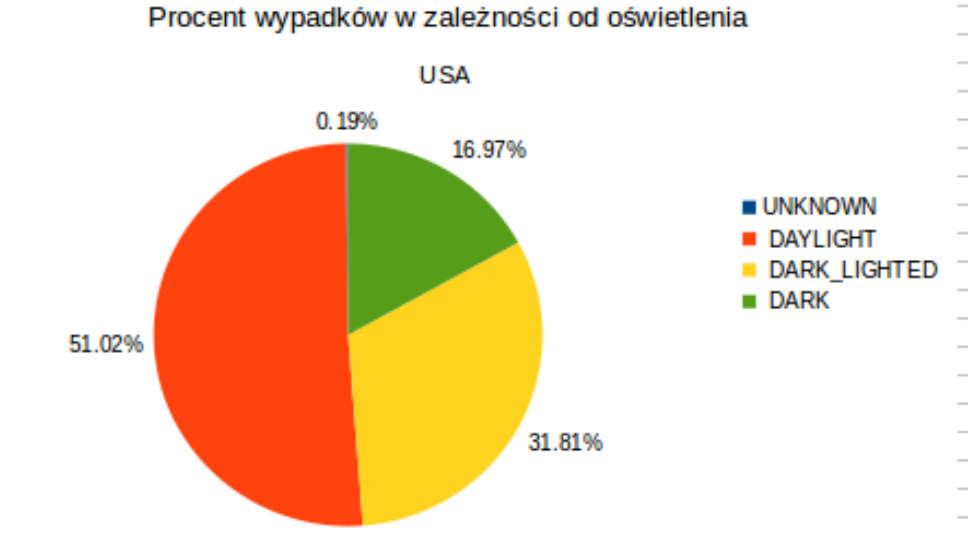

Procent wypadków śmiertelnych w danym dniu tygodnia
w których kierowca przekraczał prędkość
25
20
15
10
5
0
niedziela
poniedziałek
wtorek
środa
czwartek
piątek
sobota

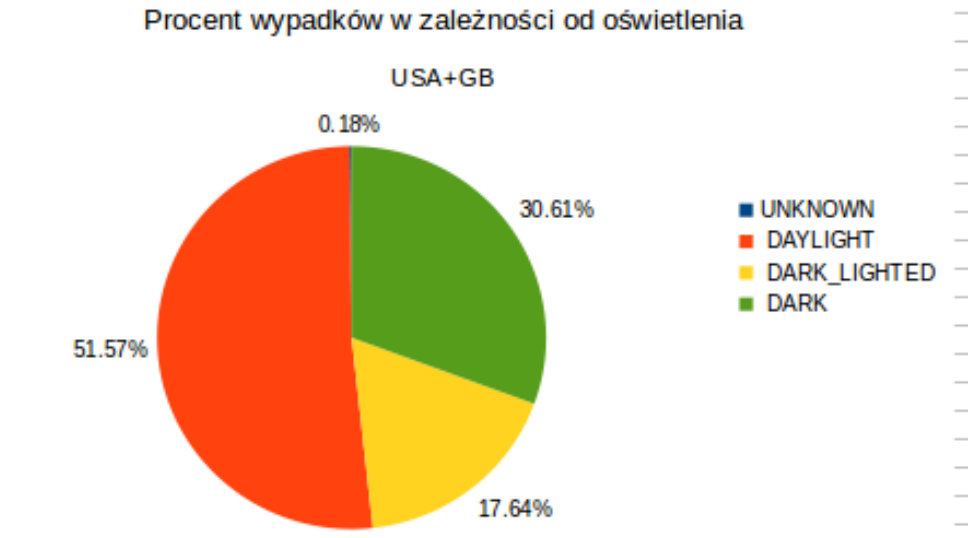




Oświetlenie

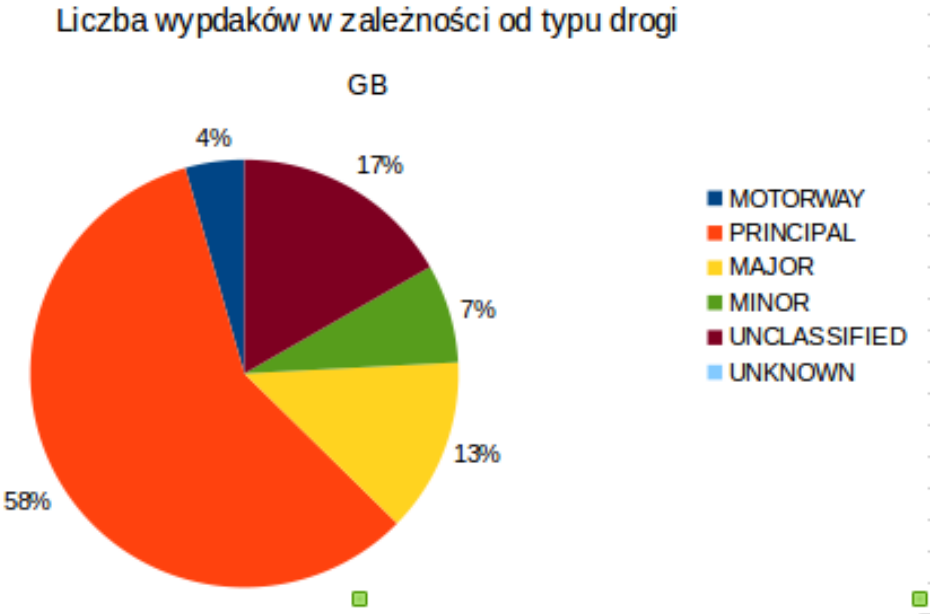


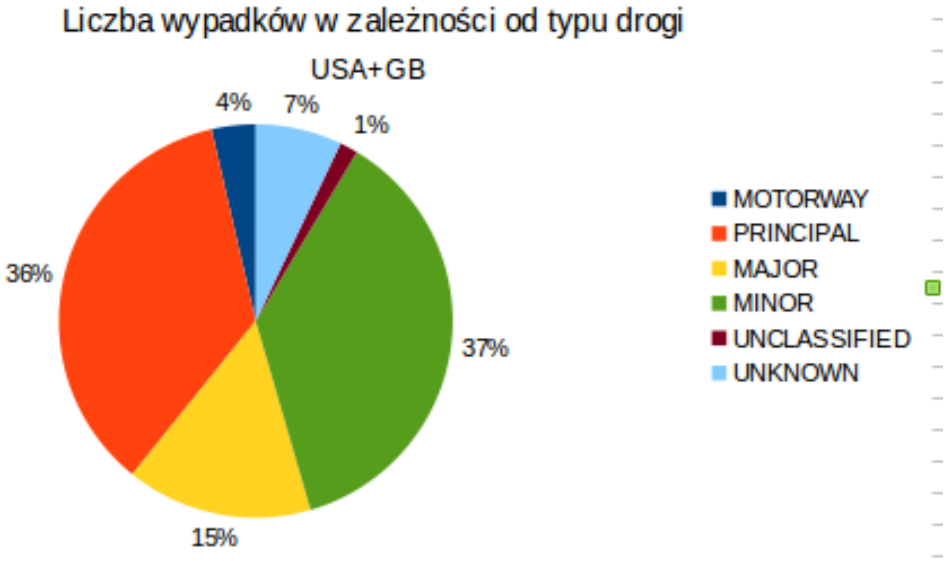




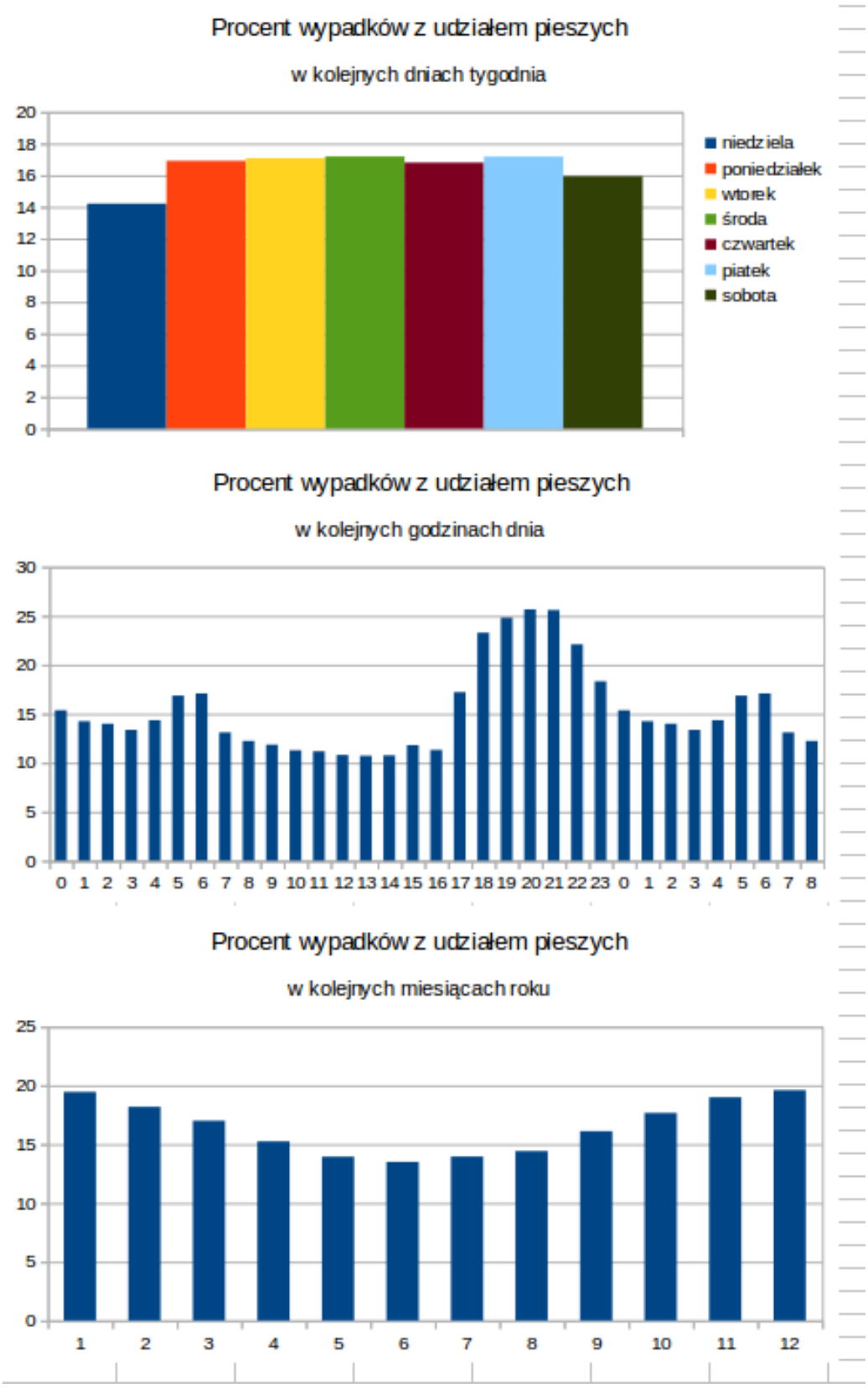
Rodzaj drogi


Liczba wypadków w zależności od typu drogi
USA
4%
8%
33%
40%
15%
MOTORWAY
PRINCIPAL
MAJOR
MINOR
UNCLASSIFIED
UNKNOWN


Liczba wypdaków w zależności od typu drogi
GB
4%
17%
7%
13%
58%
MOTORWAY
PRINCIPAL
MAJOR
MINOR
UNCLASSIFIED
UNKNOWN


Liczba wypadków w zależności od typu drogi
USA+GB
4%
7%
1%
36%
37%
15%
MOTORWAY
PRINCIPAL
MAJOR
MINOR
UNCLASSIFIED
UNKNOWN

Piesi



Wypadki z udziałem pieszych w warunkach ograniczających widoczność.

warunki	procent
wszystkie	16.36
deszcz	17.10

warunki	procent
śnieg	10.37
mgła	13.72
ciemność	21.51

9. Hipotezy analityczne

9.1. Cel dokumentu

Niniejszy dokument ma na celu przedstawienie hipotez, które stawiamy sobie w projekcie przed preprowadzeniem analizy danych. Pod kątem tych własnie hipotez będziemy przeprowadzać analizy a następnie opisane zostaną wyniki i wnioski wraz z komentarzem, czy hipoteza została potwierdzona czy obalona i z czego wynika taki a nie inny rezultat.

9.2. Hipotezy

W ramach analizy, chcemy nie tylko przeanalizować wpływ pojedynczych atrybutów na występowanie wypadków, ale także zweryfikować bardziej zaawansowane hipotezy dotyczące wpływu złożonych czynników na bezpieczeństwo na drodze.

Numer: 1
Nazwa: Ograniczenie widoczności
Treść: Czynniki ograniczające widoczność powinny mieć duży wpływ na wzrost liczby wypadków. W szczególności groźne są kombinacje takich czynników, przykładowo, niezwykle groźnymi warunkami na drodze są połączenie noc i mgła, czy noc i mgła i deszcz. Można w tę analizę włączyć jeszcze stan oświetlenia - brak oświetlenia na drodze może dodatkowo pogarszać warunki.

Numer: 2
Nazwa: Ograniczenie widoczności i piesi
Treść: Warunki ograniczenia widoczności mogą powodować większą liczbę wypadków z udziałem pieszych. Piesi są najmniej uprzywilejowanymi uczestnikami ruchu i są też w trudnych warunkach najmniej widoczni, szczególnie w wypadku braku odblasków.

Numer: 3
Nazwa: Niesprzyjające warunki atmosferyczne a ostrożność kierowców
Treść: Deszcz lub śnieg albo ich połączenie są groźnymi warunkami do jazdy. Dodatkowo silny wiatr może sprawić, iż kierowca ma ograniczoną kontrolę nad samochodem. Należy jednak sprawdzić, czy fakt, że w trudnych warunkach kierowcy jeżdżą zdecydowanie ostrożniej i nie decydują się na

brawurowe zachowania tak często jak w dobrych warunkach nie sprawia, że wypadków tych nie jest tak dużo więcej jak można by się spodziewać.

Numer: 4
Nazwa: Złe warunki a przekraczanie prędkości
Treść: W niesprzyjających warunkach (atmosferycznych i oświetleniowych) kierowcy będą rzadziej przekraczać prędkość niż w warunkach sprzyjających, stąd większy procent wypadków przy dodatkowym przekroczeniu prędkości będzie w warunkach sprzyjających.

Numer: 5
Nazwa: Alkohol we krwi kierowcy a czas
Treść: Częściej wypadki spowodowane obecnością alkoholu we krwi kierowcy będą zdarzać się w okolicach świąt, wieczorami i w weekendy.

Numer: 6
Nazwa: Przekraczanie prędkości a czas
Treść: Rzadziej wypadki spowodowane przekroczeniem prędkości przez kierowcę będą zdarzać się w zimie i na jesień niż w pozostałych porach roku, gdyż kierowcy są ostrożniejsi w trudniejszych warunkach.

Numer: 7
Nazwa: Wypadki z udziałem pieszych a czas
Treść: Wypadki z udziałem pieszych mogą być częstsze w weekendy oraz na wiosnę i w lecie, wtedy pieszych na drogach jest więcej.

Numer: 8
Nazwa: Liczba wypadków a czas
Treść: W ciągu dnia wypadków może być więcej w godzinach szczytu, wieczorem w okolicach zmroku, kiedy widoczność jest najgorsza. W ciągu roku mogłoby być ich więcej w zimie, z powodu gorszych warunków. W skali roku na poziomie dni może ich być njawięcej w okolicach świąt, gdyż jest wtedy wzmożony ruch i więcej pijanych kierowców (patrz hipoteza 5).

Numer: 9
Nazwa: Wypadki a wiek kierowców
Treść: Najwięcej wypadków będzie powodowanych przez kierowców młodych ze względu na brawurę i brak doświadczenia.

Numer: 10
Nazwa: Wypadki a rodzaj drogi
Treść: Na autostradach będzie mało wypadków, ponieważ są one bezpiecznymi drogami - obowiązuje na nich nieskomplikowany sposób poruszania się, nie ma skrzyżowań, oraz mieszkańcy USA i WB są przyzwyczajeni do częstego korzystania z nich. Ponadto nie poruszają się po nich piesi. Natomiast więcej wypadków może pojawić się na drogach o randze porównywalnej z polskimi krajowymi oraz wojewódzkimi (Principal, Major), ponieważ z jednej strony mogą mieć skomplikowana infrastruk-

turę co może wymagać więcej umiejętności od kierowców, a z drugiej pozwalają osiągnąć znaczne prędkości. Możliwe jest też zaobserwowanie znacznej ilości wypadków śmiertelnych na drogach drugorzędnych (Minor), ponieważ często na terenach wiejskich ludzie dopuszczają się bardziej brawurowej jazdy i kierowania po spożyciu, gdyż rzadziej można tam spotkać funkcjonariuszy drogówki.

10. Weryfikacja hipotez

10.1. Hipoteza 1

10.1.1. Opis hipotezy

Numer: 1
Nazwa: Ograniczenie widoczności
Treść: Czynniki ograniczające widoczność powinny mieć duży wpływ na wzrost liczby wypadków. W szczególności groźne są kombinacje takich czynników, przykładowo, niezwykle groźnymi warunkami na drodze są połączenie noc i mgła, czy noc i mgła i deszcz. Można w tę analizę włączyć jeszcze stan oświetlenia - brak oświetlenia na drodze może dodatkowo pogarszać warunki.

10.1.2. Wyniki związane z hipotezą

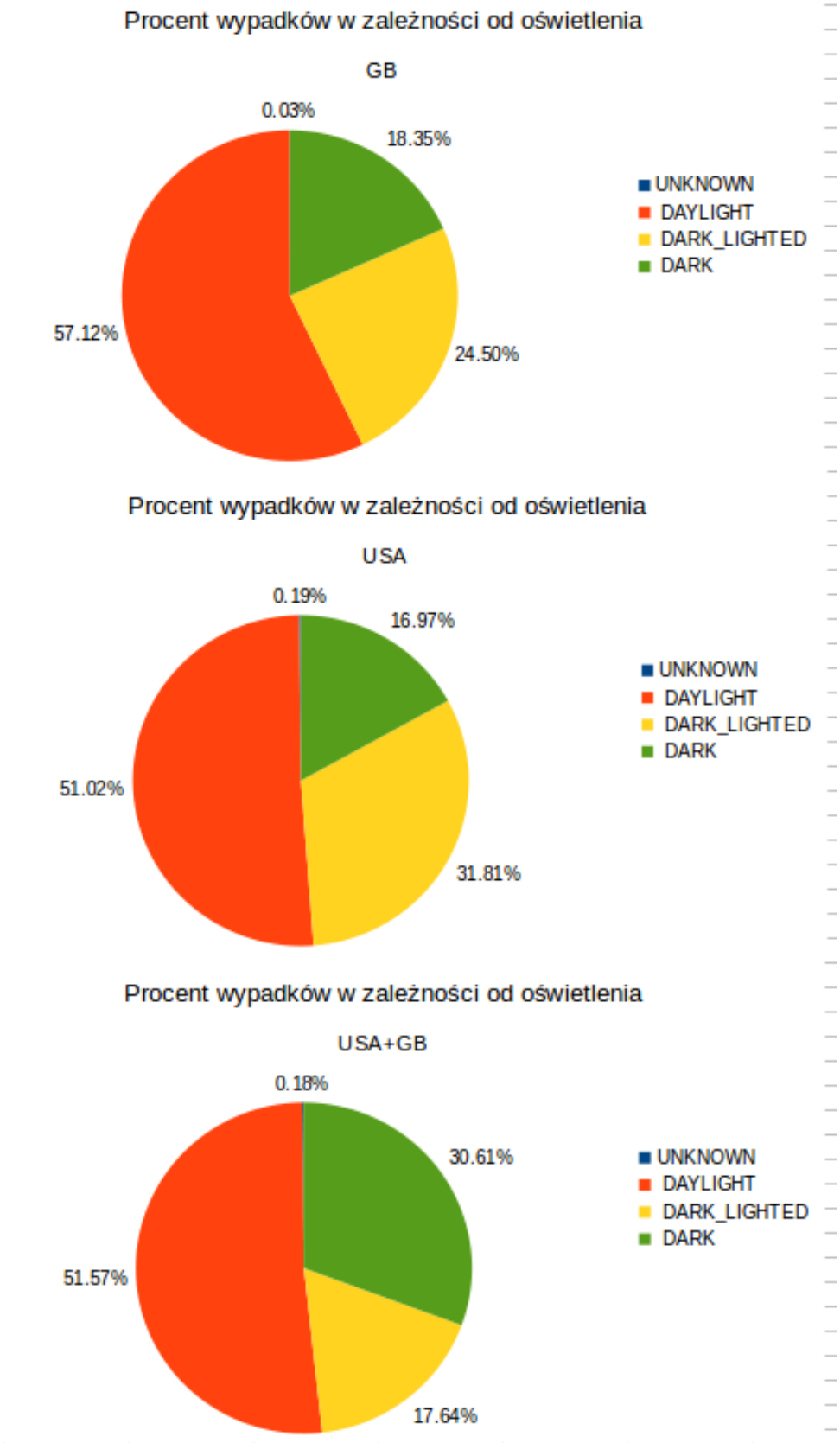
Warunki pogodowe

Udział wypadków z różnymi kombinacjami warunków pogodowych
S - Snow
W - Wind
R - Rain
F - Fog

CONDS	USA	USA %	GB	GB %	USA+GB	USA+GB %
NONE	1142653	88,17%	104834	82,98%	1247487	87,71%
F	17222	1,33%	1306	1,03%	18528	1,30%
W	445	0,03%	2798	2,21%	3243	0,23%
WF	2	0,00%	0	0,00%	2	0,00%
R	108652	8,38%	14463	11,45%	123115	8,66%
RF	1444	0,11%	0	0,00%	1444	0,10%

CONDS	USA	USA %	GB	GB %	USA+GB	USA+GB %
RW	65	0,01%	2177	1,72%	2242	0,16%
RWF	0	0,00%	0	0,00%	0	0,00%
S	19108	1,47%	565	0,45%	19673	1,38%
SF	7	0,00%	0	0,00%	7	0,00%
SW	2005	0,15%	191	0,15%	2196	0,15%
SWF	6	0,00%	0	0,00%	6	0,00%
SR	4283	0,33%	0	0,00%	4283	0,30%
SRF	12	0,00%	0	0,00%	12	0,00%
SRW	78	0,01%	0	0,00%	78	0,01%
SRWF	0	0,00%	0	0,00%	0	0,00%

Oświetlenie



10.1.3. Weryfikacja i wnioski

Analiza tabeli dotyczącej wpływu warunków pogodowych na liczbę wypadków pokazuje, że pod tym względem hipoteza sprawdza się w ograniczonym zakresie, prawie 80 - 90% wypadków występuje

przy pogodzie nieutrudniającej jazdy. Warto zauważyć wyższą wartość procentową wypadków w trakcie deszczu dla Wielkiej Brytanii - związane jest to zapewne z faktem, że pogoda Wielkiej Brytanii jest stosunkowo często deszczowa.

Wydaje się, że zła pogoda podwyższa nieco liczbę wypadków. W deszczowej Wielkiej Brytanii możemy się spodziewać, że deszcz pada przez przynajmniej 10% czasu. Oznacza to, że udział wypadków przy pogodzie deszczowej jest trochę wyższy niż statystycznie uzadaniony i deszcz przyczynia się do zwiększenia liczby wypadków. Z drugiej strony analiza hipotezy 8 pokazuje, że rozkład wypadków w miesiącach różni się między USA i Wielką Brytanią. W miesiącach najbardziej deszczowych (październik, listopad, grudzień) mamy do czynienia z podwyższeniem liczby wypadków w Wielkiej Brytanii, w USA za to największa jest liczba wypadków w miesiącach wakacyjnych (lipiec, sierpień), kiedy warunki do jazdy powinny być nalepsze.

Liczba wypadków przy opadach śniegu jest procentowo większy w USA. Ma to sens, jako że w łagodnym klimacie Wielkiej Brytanii śnieg pada niezwykle rzadko. Wartości te sugerują, że śnieg również wpływa na zwiększenie prawdopodobieństwa wypadków. Wartości te wydają się być wyższe niż te statystycznie uzasadnione.

Wykresy dotyczące oświetlenia pokazują, że ciemność wpływa na wzrost liczby wypadków. Wniosek ten płynie stąd, że niewątpliwie ruch w nocy jest mniejszy, więc prawdopodobieństwo wypadku w nocy powinno być znacznie mniejsze. Wykresy pokazują jednak, że prawie połowa wypadków zdarza się w ciemności, co implikuje wzrost prawdopodobieństwa wypadku w takich warunkach. Szczególnie widoczne jest to w Stanach Zjednoczonych, nieco mniejszy procent w Wielkiej Brytanii.

Podsumowując, trudne warunki nie wpływają aż tak bardzo na wzrostliczby wypadków jak można by się spodziewać. Nadal znacząca większość wypadków zdarza się w warunkach sprzyjających. Działają tu prawdopodobnie czynniki psychologiczne - w warunkach gorszych jesteśmy ostrożniejsi i bardziej uważamy na to co dzieje się na drodze, niwelując niekorzystny wpływ warunków.

10.2. Hipoteza 2

10.2.1. Opis hipotezy

Numer: 2
Nazwa: Ograniczenie widoczności i piesi
Treść: Warunki ograniczenia widoczności mogą powodować większą liczbę wypadków z udziałem pieszych. Piesi są najmniej uprzywilejowanymi uczestnikami ruchu i są też w trudnych warunkach najmniej widoczni, szczególnie w wypadku braku odblasków.

10.2.2. Wyniki związane z hipotezą

warunki	procent
wszystkie	16.36
deszcz	17.10
śnieg	10.37
mgła	13.72
ciemność	21.51

10.2.3. Weryfikacja i wnioski

Można powiedzieć iż hipoteza potwierdziła się częściowo. Widzimy wyższy procent wypadków z pieszymi w przypadku deszczu oraz ciemności. Szczególnie ta druga pokazuje konieczność edukowania ludzi co do konieczności noszenia na odzieży elementów odblaskowych, aby kierowca miał szansę zauważyć pieszego na ulicy.

Ciekawa jest dużo niższa wartość dla śniegu i mgły. Może być to związane z wzmożoną ostrożnością kierowców w takich warunkach oraz rzadszym wychodzeniem pieszych na drogę, szczególnie kiedy pada śnieg. W deszczu nie obserwujemy jednak tego efektu, a piesi są mniej widoczni, stąd już wzrost w warunkach deszczowych.

10.3. Hipoteza 3

10.3.1. Opis hipotezy

Numer: 3
Nazwa: Niesprzyjające warunki atmosferyczne a ostrożność kierowców
Treść: Deszcz lub śnieg albo ich połączenie są groźnymi warunkami do jazdy. Dodatkowo silny wiatr może sprawić, iż kierowca ma ograniczoną kontrolę nad samochodem. Należy jednak sprawdzić, czy fakt, że w trudnych warunkach kierowcy jeżdżą zdecydowanie ostrożniej i nie decydują się na brawurowe zachowania tak często jak w dobrych warunkach nie sprawia, że wypadków tych nie jest tak dużo więcej jak można by się spodziewać.

10.3.2. Wyniki i weryfikacja

Wyniki i wnioski dotyczące tej hipotezy zostały zawarte w dokumencie dotyczącym hipotezy 1.

10.4. Hipoteza 4

10.4.1. Opis hipotezy

Numer: 4
Nazwa: Złe warunki a przekraczanie prędkości
Treść: W niesprzyjających warunkach (atmosferycznych i oświetleniowych) kierowcy będą rzadziej przekraczać prędkość niż w warunkach sprzyjających, stąd większy procent wypadków przy dodatkowym przekroczeniu prędkości będzie w warunkach sprzyjających.

10.4.2. Wyniki związane z hipotezą

Procent wypadków, w których przekroczono prędkość w różnych warunkach

warunki	procent
wszystkie	17.79
deszcz	12.60
śnieg	6.50
mgła	14.76
ciemność	19.83

10.4.3. Weryfikacja i wnioski

Hipoteza znajduje zdecydowane potwierdzenie w danych. Najbardziej drastyczny (trzykrotny) spadek udziału wypadków z przekroczeniem prędkości obserwujemy, gdy pada śnieg. Trudne warunki na drodze wymuszają często jazdę dużo poniżej ograniczenia i powodują zwiększenie ostrożności u kierowców.

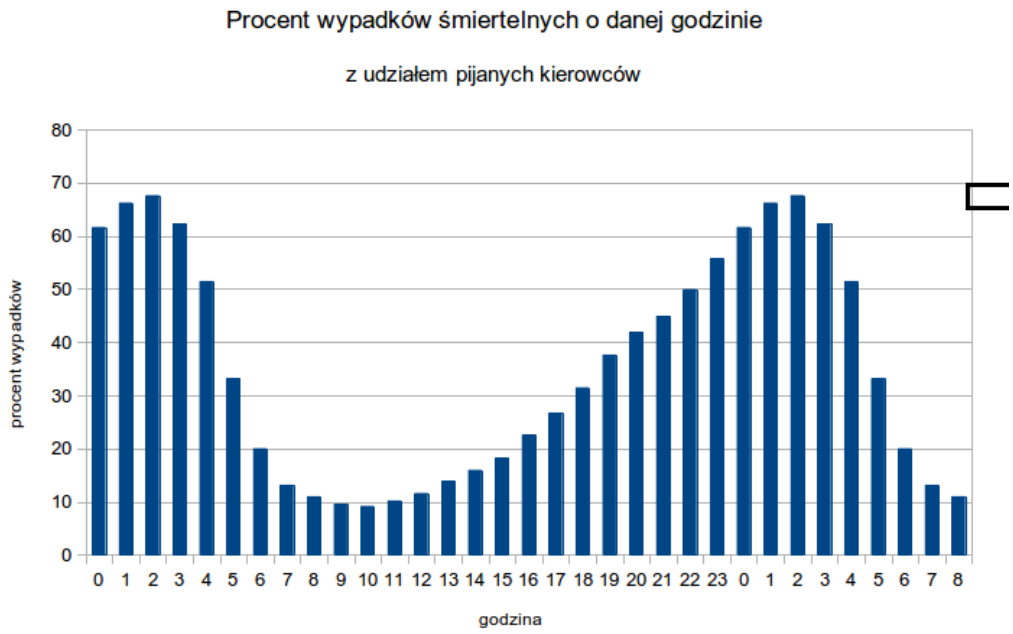
Ciekawy jest pzrost udziału wypadków z przekroczeniem prędkości w nocy. Jest to prawdopodobnie spowodowane faktem, że w nocy na drogach jest pusto. Jeżeli nie ma innych czynników ograniczających widoczność, kierowcy czują się na drodze pewniej i jadą szybciej.

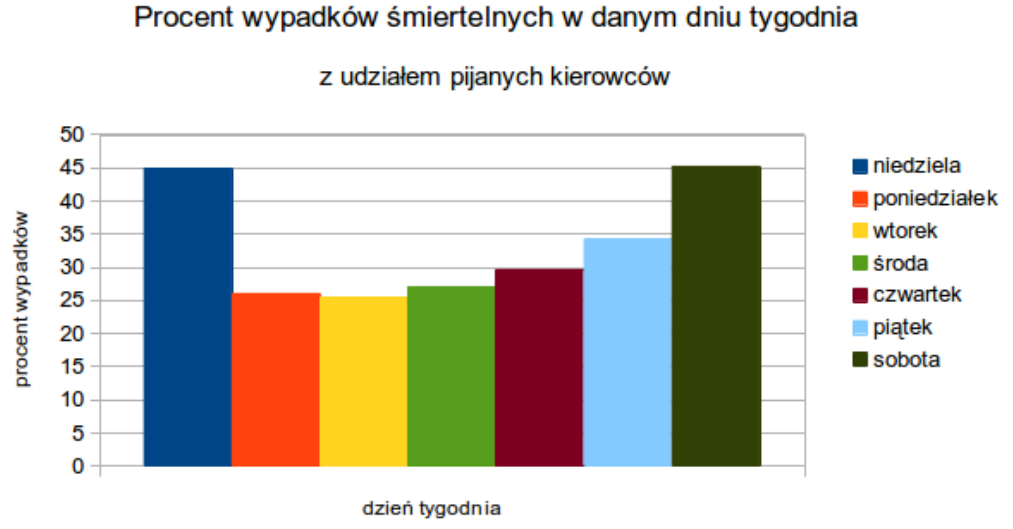
10.5. Hipoteza 5

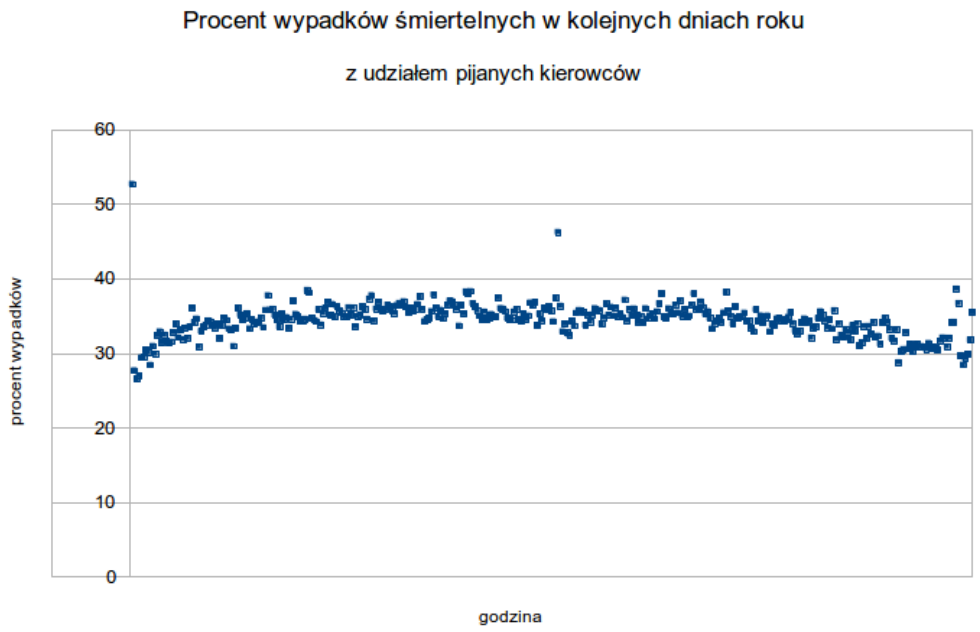
10.5.1. Opis hipotezy

Numer: 5
Nazwa: Alkohol we krwi kierowcy a czas
Treść: Częściej wypadki spowodowane obecnością alkoholu we krwi kierowcy będą zdarzać się w okolicach świąt, wieczorami i w weekendy.

10.5.2. Wyniki związane z hipotezą







10.5.3. Weryfikacja i wnioski

Postawiona hipoteza sprawdziła się. Wyraźny wzrost procentowego udziału wypadków spowodowanych przez pijanych kierowców obserwujemy w weekend - zarówno w sobotę jak i w niedzielę jest to ok. 45% a więc bardzo wysoka wartość.

Patrząc na rozkład godzinny, widzimy wyraźny stały wzrost aż do godziny 2 gdzie wartość osiąga maksimum w okolicach bardzo wysokiej wartości 68%, który następnie gwałtownie spada aż do minimum w godzinach porannych (8 - 10).

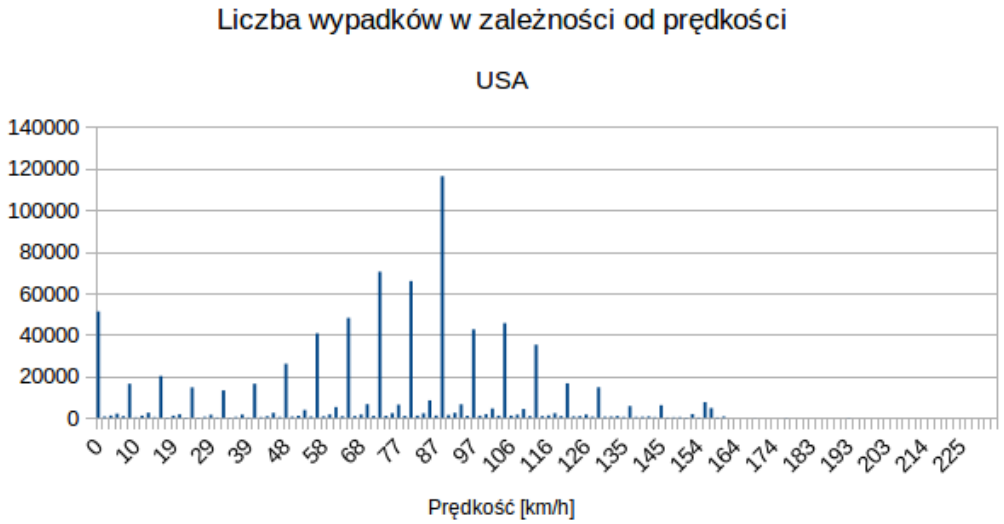
W ciągu roku najbardziej wybija się 1 stycznia z wartością ponad 50% oraz święto niepodległości 4 lipca (47%). Nieznaczny wzrost notujemy również pod koniec roku, w okolicach świąt Bożego Narodzenia.

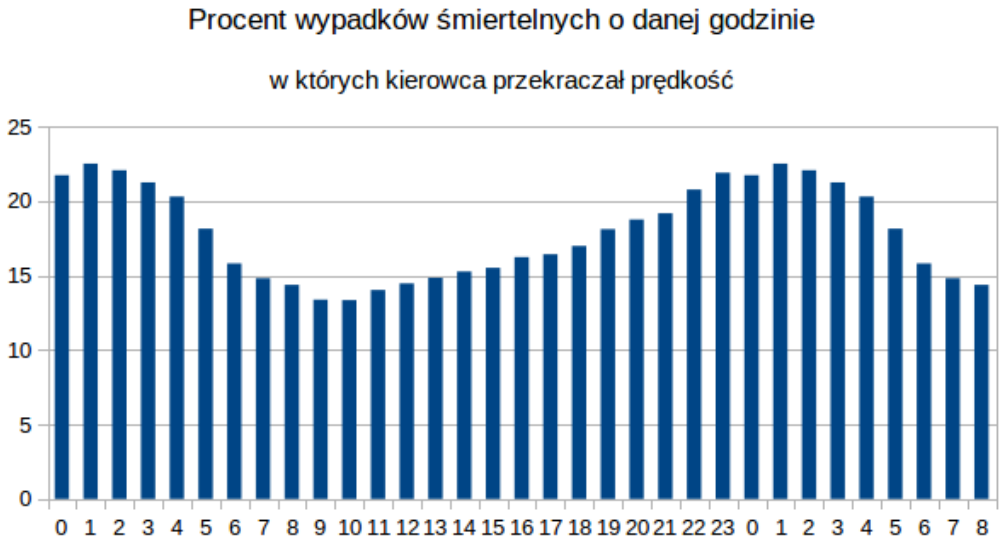
10.6. Hipoteza 6

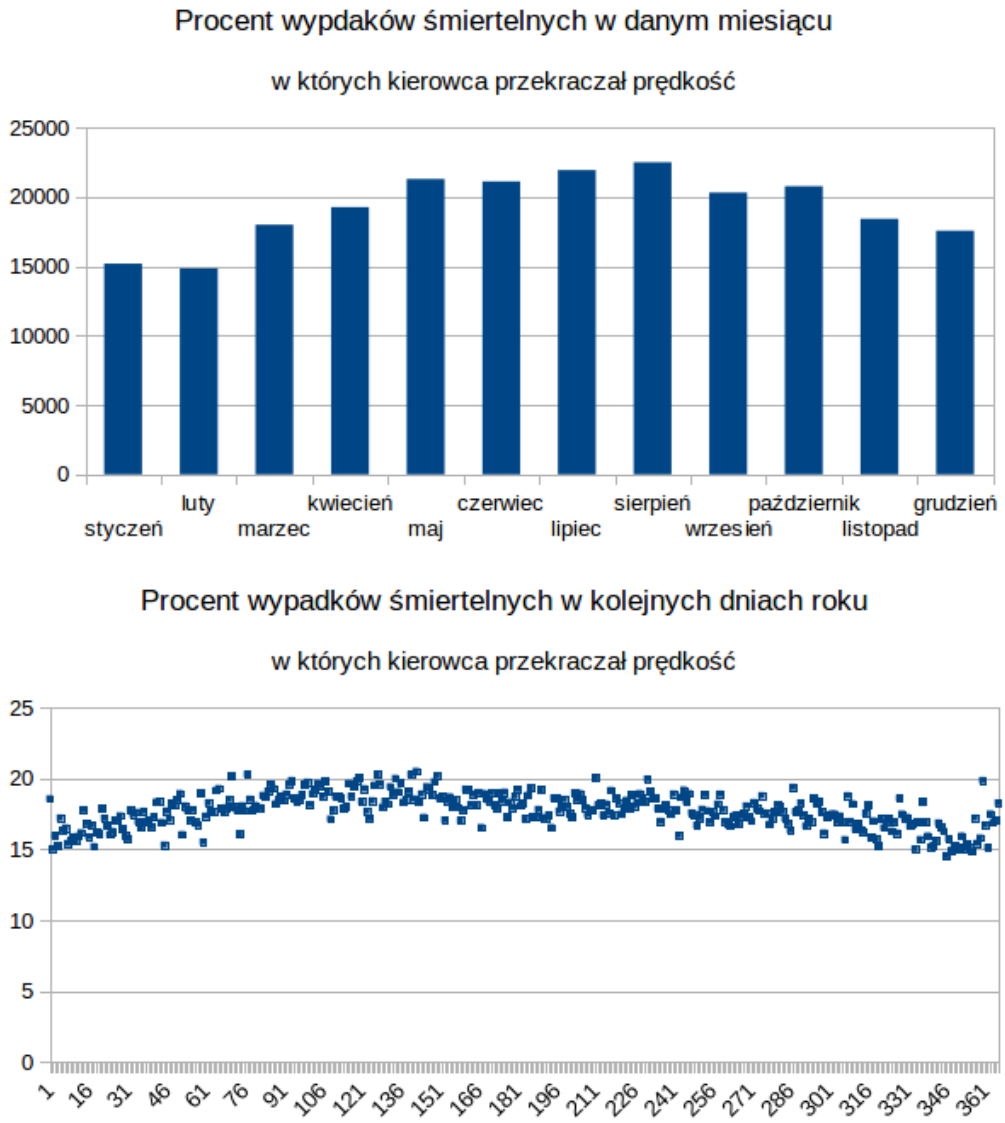
10.6.1. Opis hipotezy

Numer: 6
Nazwa: Przekraczanie prędkości a czas
Treść: Rzadziej wypadki spowodowane przekroczeniem prędkości przez kierowcę będą zdarzać się w zimie i na jesień niż w pozostałych porach roku, gdyż kierowcy są ostrożniejsi w trudniejszych warunkach.

10.6.2. Wyniki związane z hipotezą







10.6.3. Weryfikacja i wnioski

Hipoteza potwierdziła się. Patrząc na wykres w zależności od miesiąca oraz dnia w roku, najmniej wypadków związanych z przekroczeniem prędkości obserwujemy w miesiącach zimowych (styczeń, luty, grudzień). Kierowcy z mniejszą brawurą podchodzą do jazdy w trudniejszych warunkach i rzadziej łamią wtedy przepisy.

Większa brawura kierowców wychodzi w weekendy, wartości procentowe w sobotę i niedzielę przekraczają o blisko 5 procent pozostałe dni.

Ciekawa jest także zależność liczby wypadków od prędkości. Widać, że największa wartość przypada na 90 km/h.

10.7. Hipoteza 7

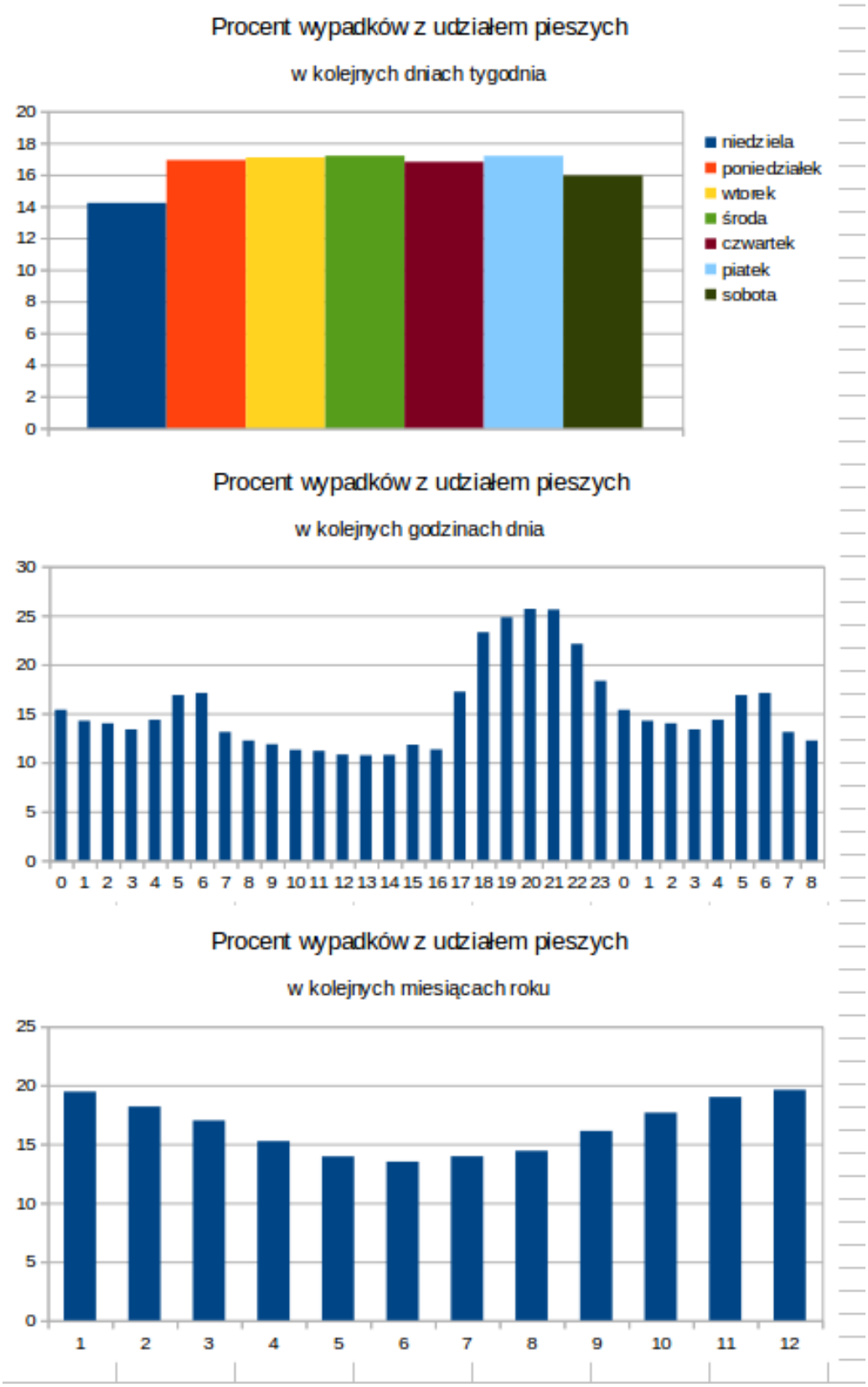
10.7.1. Opis hipotezy

Numer: 7

Nazwa: Wypadki z udziałem pieszych a czas

Treść: Wypadki z udziałem pieszych mogą być częstsze w weekendy oraz na wiosnę i w lecie, wtedy pieszych na drogach jest więcej.

10.7.2. Wyniki związane z hipotezą



10.7.3. Weryfikacja i wnioski

Hipoteza nie potwierdziła się. Większy udział procentowy wypadków z udziałem pieszych obserwujemy w ciągu tygodnia od poniedziałku do piątku. W ciągu roku najmniej wypadków z udziałem

pieszych jest właśnie w miesiącach wakacyjnych (czerwiec, lipiec, sierpień).

Przyczyn takiego stanu rzeczy można szukać np. w fakcie, że jednak ruch pieszych jest większy poza weekendem, gdyż większość ruchu pieszych to jednak ludzie udający się do pracy/szkoły. Większy ruch pieszych sprawia, że i wypadków z ich udziałem jest więcej.

Trudniej wytłumaczyć rozkład wypadków w ciągu roku. Można by się pokusić o stwierdzenie, że trudniejsze warunki powodują, że wypadków z pieszymi jest więcej i zgodnie z hipotezą 2 czasem faktycznie tak jest, chociaż np. śnieg sprawia, że procentowy udział wypadków z pieszymi jest mniejszy. Można wskazać ciemność jako jeden z czynników decydujących - w zimie czy na jesień wcześniej robi się ciemno a w ciemności obserwujemy wyraźny wzrost procentowego udziału wypadków z pieszymi.

Analiza wykresu godzinowego może być poparciem tezy o dużym wpływie ruchu do/z pracy na liczbę wypadków z udziałem pieszych. Największy procent obserwujemy w godzinach 17-22. Częściowo jest to ruch powrotny z pracy/szkoły a później ruch związany z wieczornymi wyjściami.

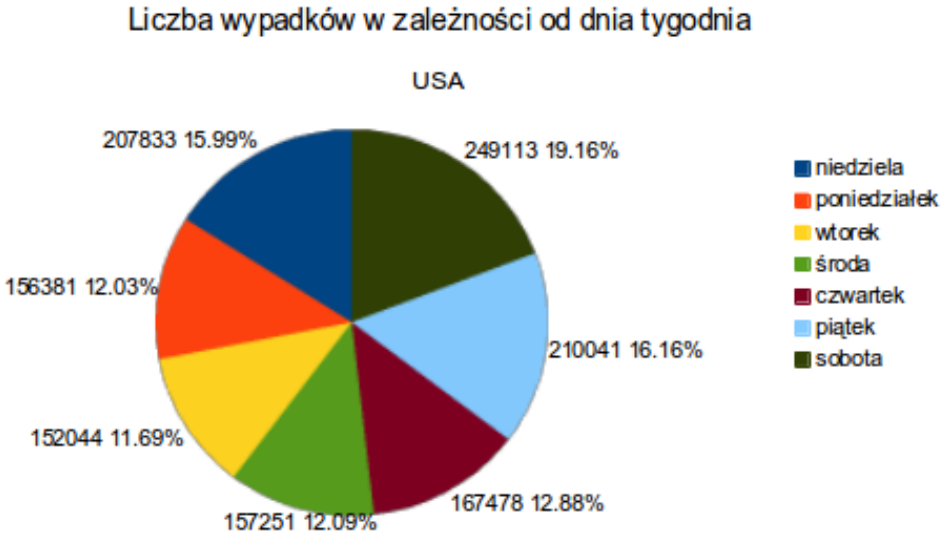
10.8. Hipoteza 8

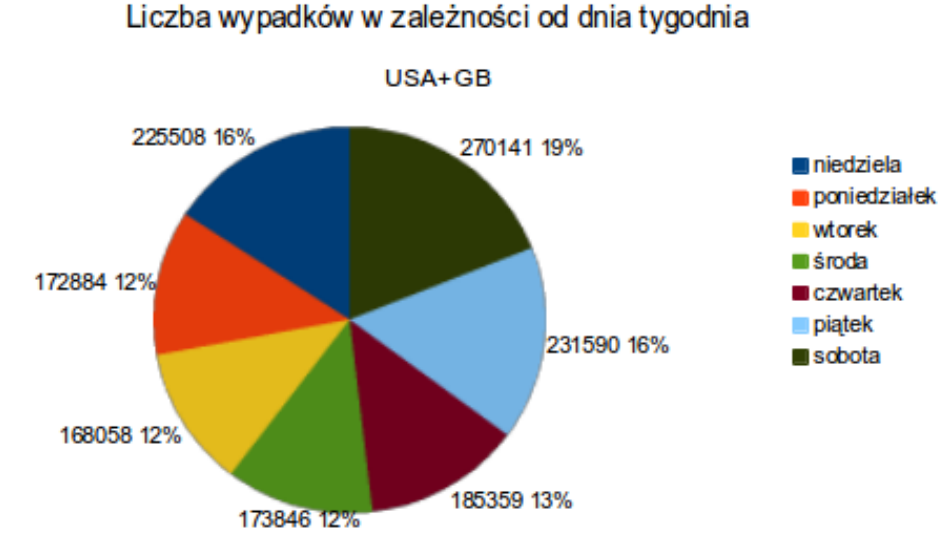
10.8.1. Opis hipotezy

Numer: 8
Nazwa: Liczba wypadków a czas
Treść: W ciągu dnia wypadków może być więcej w godzinach szczytu, wieczorem w okolicach zmroku, kiedy widoczność jest najgorsza. W ciągu roku mogłoby być ich więcej w zimie, z powodu gorszych warunków. W skali roku na poziomie dni może ich być njawięcej w okolicach świąt, gdyż jest wtedy wzmożony ruch i więcej pijanych kierowców (patrz hipoteza 5).

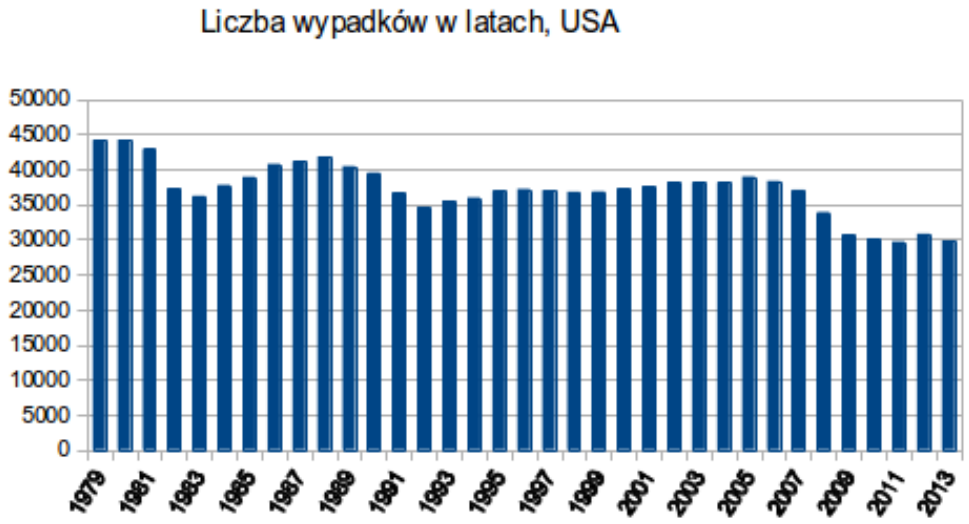
10.8.2. Wyniki związane z hipotezą

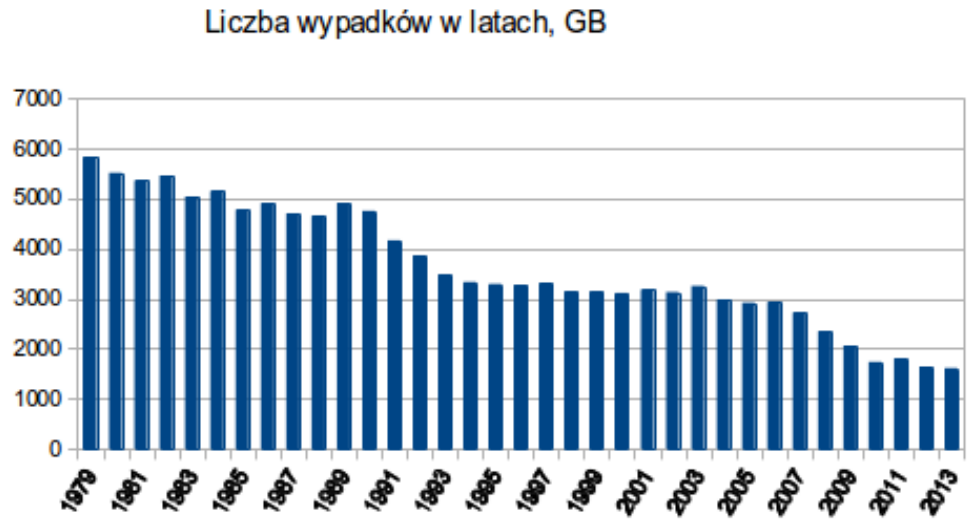
Dzień tygodnia:

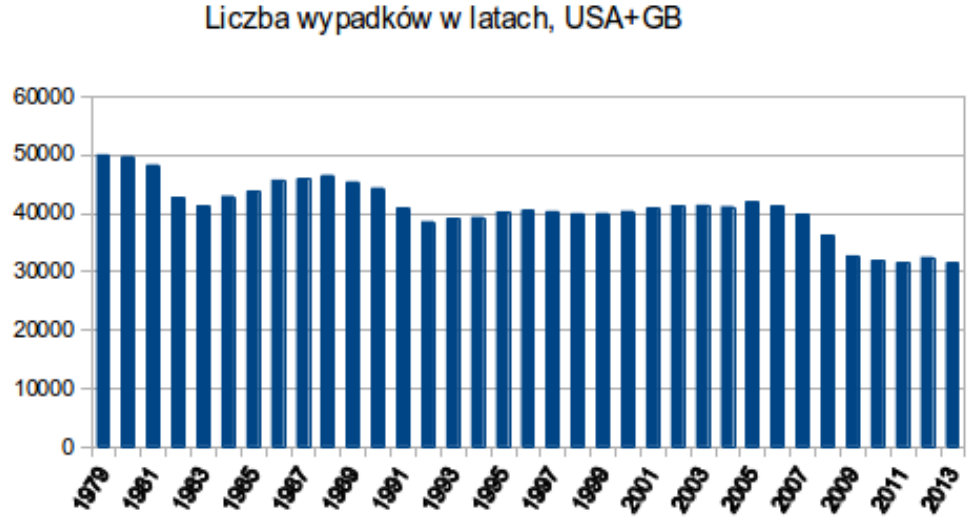




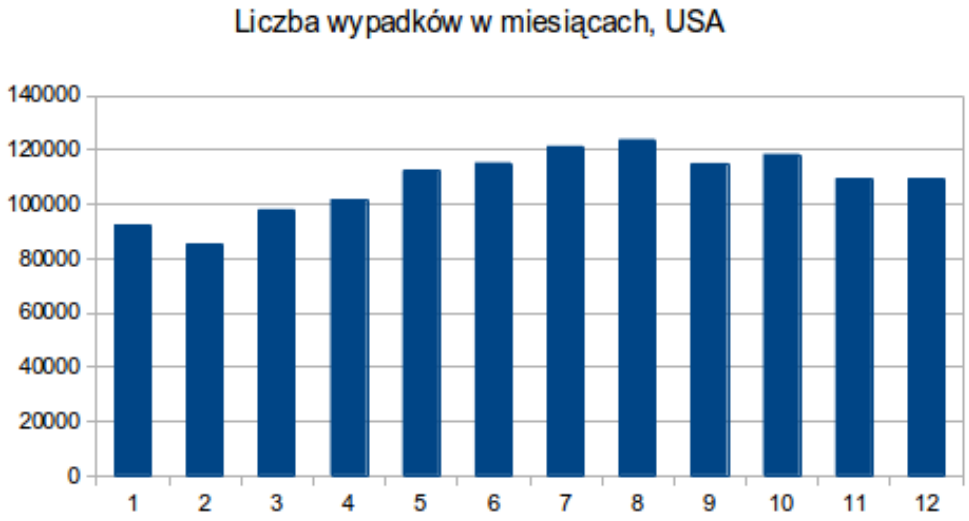
Rok:

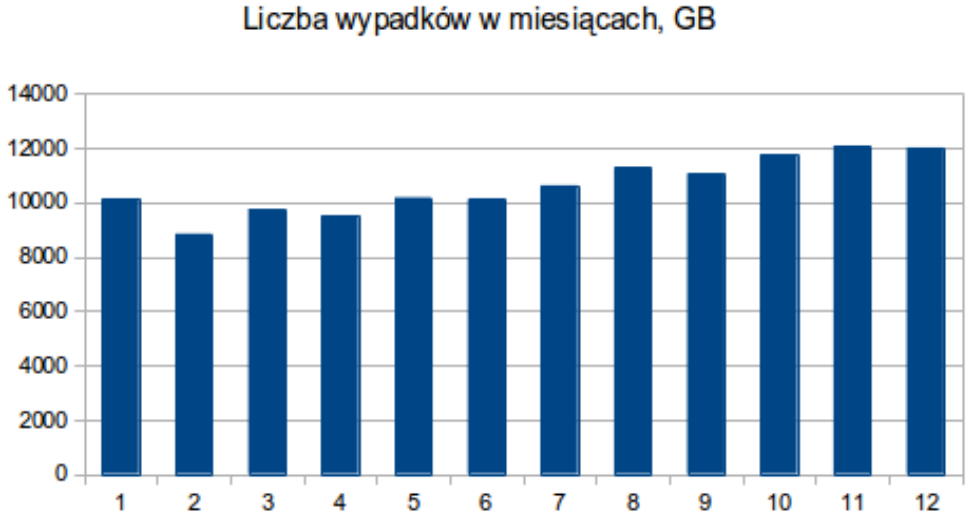

Liczba wypadków w latach, USA
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013

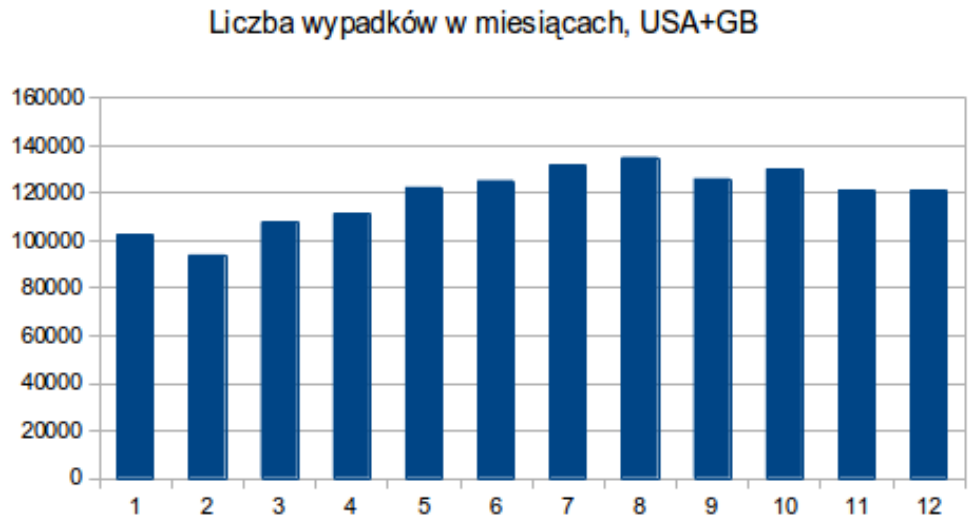

Liczba wypadków w latach, GB
7000
6000
5000
4000
3000
2000
1000
0
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013


Liczba wypadków w latach, USA+GB
60000
50000
40000
30000
20000
10000
0
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013

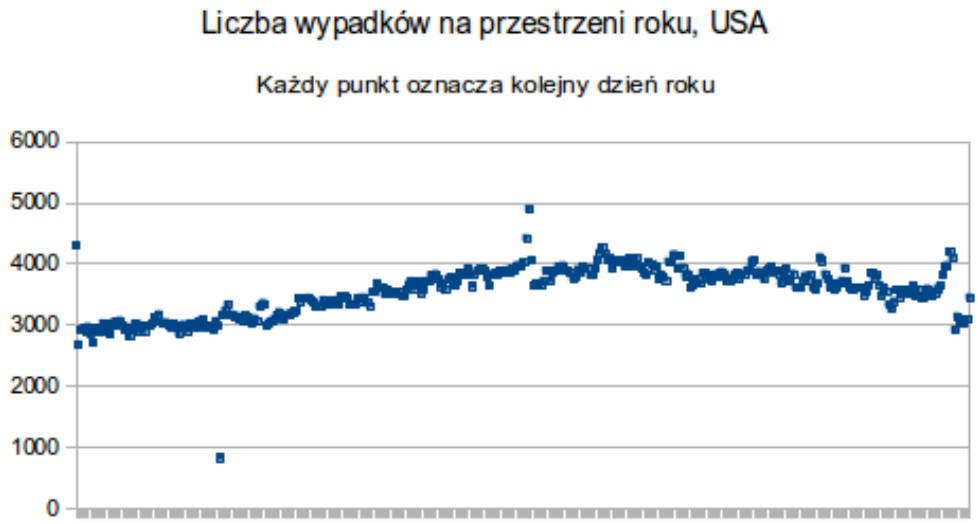
Miesiąc:

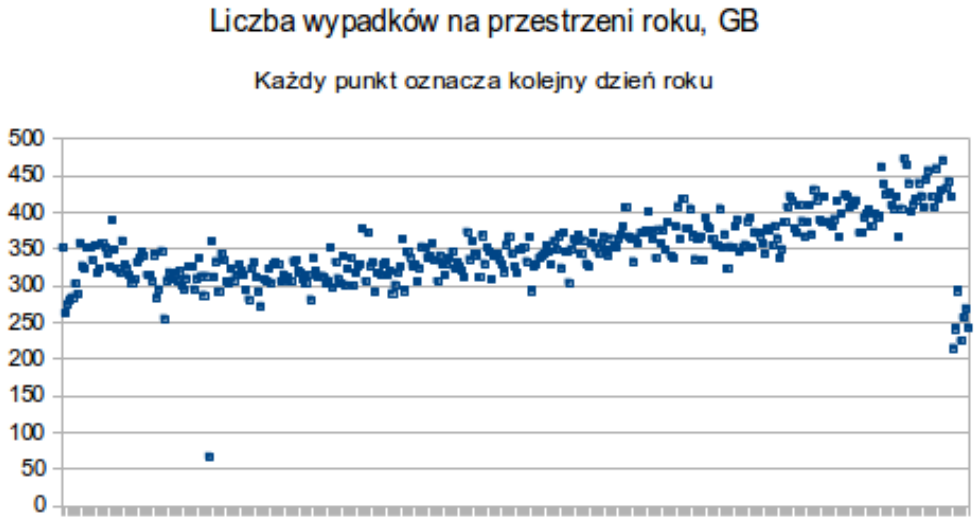

Liczba wypadków w miesiącach, USA
140000
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
11
12

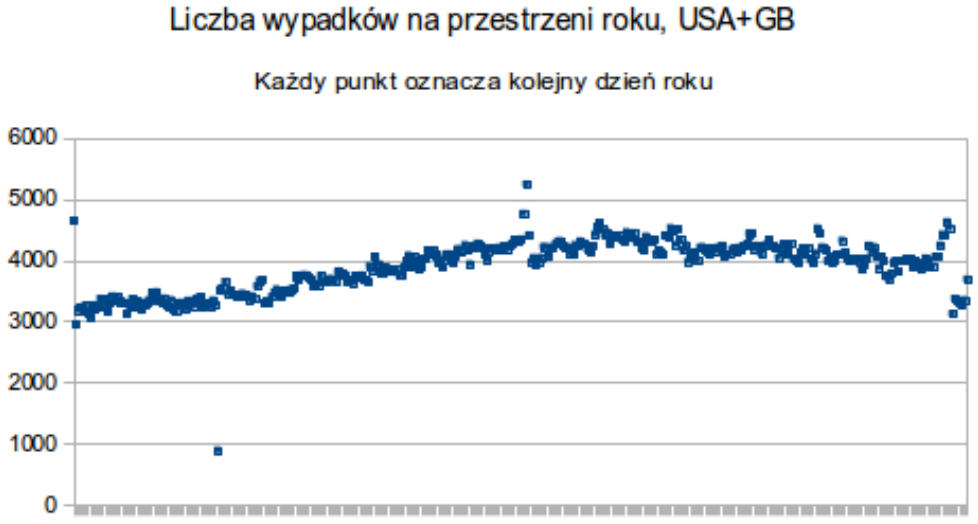

Liczba wypadków w miesiącach, GB
14000
12000
10000
8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
9
10
11
12


Liczba wypadków w miesiącach, USA+GB
160000
140000
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
11
12

Dzień:







Bardzo niska wartość jest dla 29.02 - występuje tylko w latach przestępnych i istotnie jest to wartość średnio 4 razy mniejsza.

Lokalne “piki” - **USA**:

- 01.01
- 03.07
- 04.07
- 02.08
- 03.08
- 31.10
- 01.11
- 20.12
- 21.12
- 22.12
- 23.12
- 24.12
- 31.12

Lokalne “piki” - **GB**:

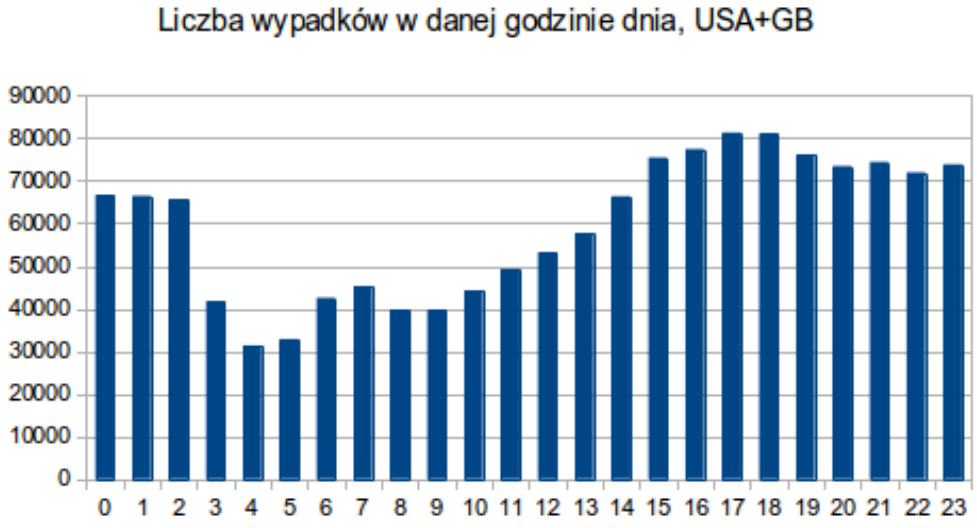
- 21.01
- 01.03
- 18.04
- 01.05
- 15.08
- 07.09
- 20.10
- 30.10
- 05.12
- 21.12

Ciekawy jest spadek liczby wypadków w ostatnich dniach roku - spowodowany prawdopodobnie faktem, że w tym okresie znacznie mniej osób jeździ do pracy i spędza więcej czasu w domu z rodziną.

Godzina:


Liczba wypadków w danej godzinie dnia, USA
80000
70000
60000
50000
40000
30000
20000
10000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


Liczba wypadków w danej godzinie dnia, GB
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


Liczba wypadków w danej godzinie dnia, USA+GB
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

10.8.3. Weryfikacja i wnioski

Hipoteza analityczna dotycząca liczby wypadków w czasie w znacznej mierze się potwierdziła. Podzielimy wyniki tej analizy w zależności od rozważanych wymiarów czasowych

Godzina

Bardzo wyraźnie widać wrost liczby wypadków w godzinach szczytu - w trakcie powrotów z pracy. W Wielkiej Brytanii są to przede wszystkim okolice godziny 16 i 17, w USA godzina 17-18. Widać również kiedy rano ruch się zaczyna wzmagać i jest to w okolicach godziny 7, jednak nie ma tak dużego skoku ilości wypadków dla porannej pory dojazdu do pracy. Może to być związane z faktem że w drodze powrotnej kierowcy są bardziej rozkojarzeni i zmęczeni niż rano. Porównanie Wielkiej Brytani i USA pokazuje również, że w USA znacznie więcej wypadków zdarza się we wczesnych godzinach nocnych (23 - 2). Powodem takiego stanu rzeczy może być na przykład większy ruch tranzytowy w tych godzinach lub częstszy wybór samochodu jako środka transportu w przypadku późnego powrotu do domu.

Miesiąc

Analizując liczbę wypadków w kolejnych miesiącach widzimy znaczne różnice między rozkładem w Wielkiej Brytanii i USA. W Stanach, wyraźne nasilenie ilości wypadków jest w miesiącach wakacyjnych, mimo że można się spodziewać, że warunki pogodowe będą wtedy lepsze. Nasilenie ruchu musi byc znacznie większe w tym czasie, może być też większa brawura kierowców. W Wielkiej Brytanii można bardziej wnioskować o korelacji między warunkami pogodowymi a liczbą wypadków. Okazuje się że najniebezpieczniejsze są październik, listopad i grudzień, kiedy warunki nie są aż tak wymagające jak w styczniu czy w lutym, jednak ludzie nie unikają jazdy samochodem i nie uważają tak bardzo jak w miesiącach ściśle zimowych.

Dzień tygodnia

Potwierdziła się hipoteza o większej liczbie wypadków w weekend. Jeżeli dodamy wartości liczby wypadków dla piątku, soboty i niedzieli, zarówno dla Wielkiej Brytanii jak i dla USA stanowią one około połowy całkowitej liczby wypadków.

Dzień

Z analizy rozkładu liczby wypadków na dni w ciągu roku można zidentyfikować niektóre dni wolne. W USA wybija się 4 lipca - dzień niepodległości. Można zauważyć wzrost np przed Halloween czy przed 6.12. Wzrost można obserwować w okolicach Obserwujemy spadek liczby wypadków w okresie świątecznym, kiedy ludzie mają wolne w pracy i przeważnie spędzają ten czas w domu z rodziną.

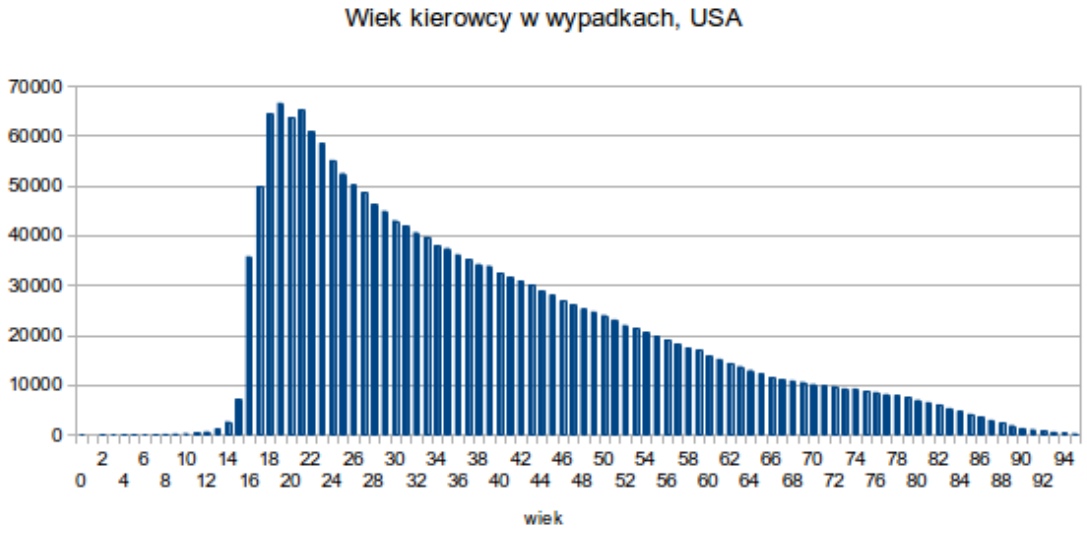
10.9. Hipoteza 9

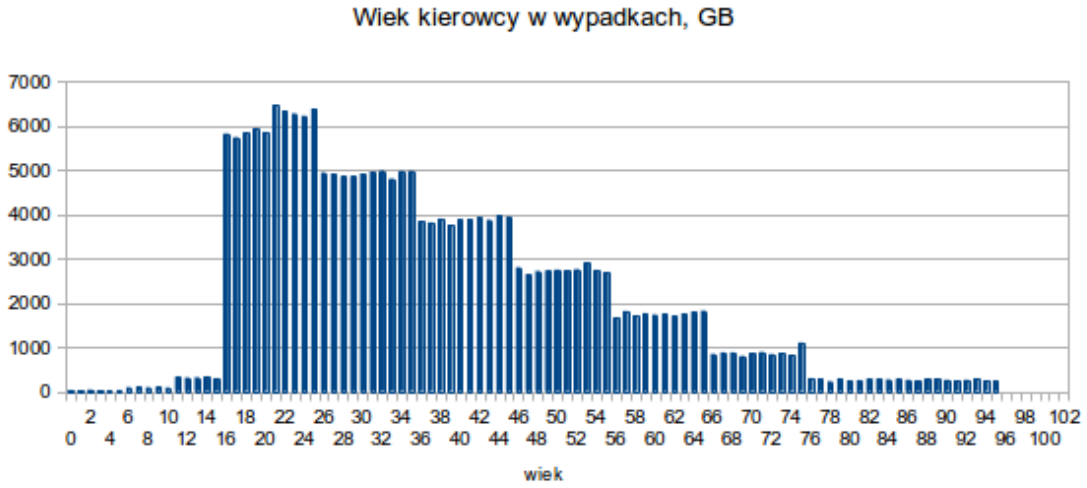
10.9.1. Opis hipotezy

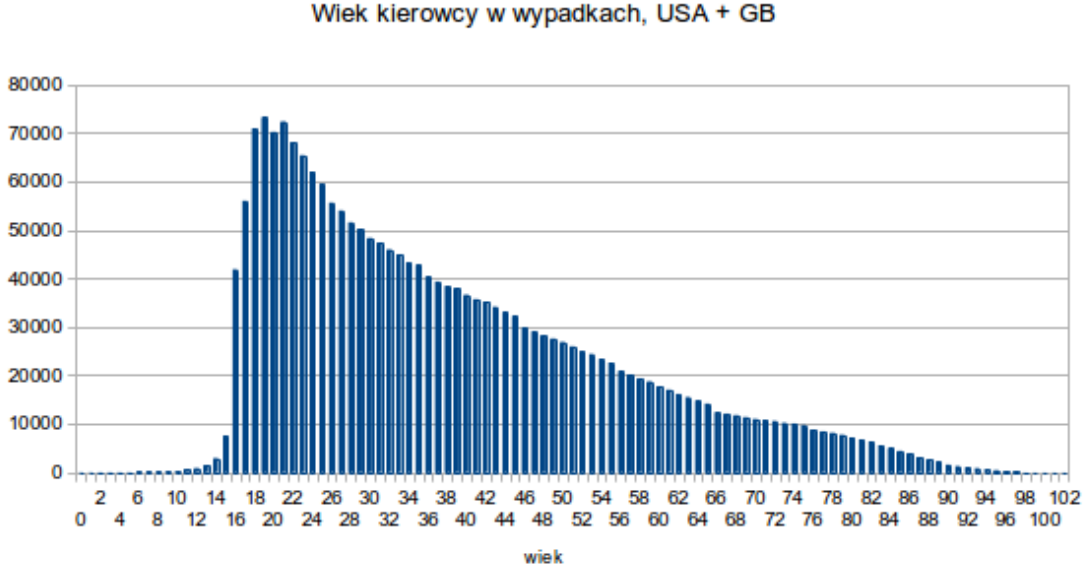
Numer: 9
Nazwa: Wypadki a wiek kierowców
Treść: Najwięcej wypadków będzie powodowanych przez kierowców młodych ze względu na brawurę i brak doświadczenia.

10.9.2. Wyniki związane z hipotezą

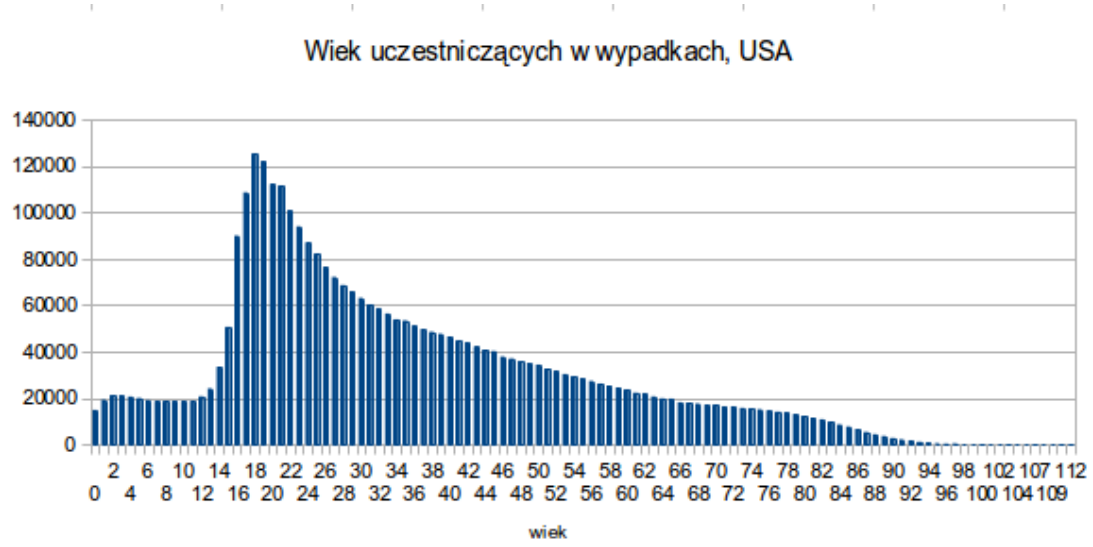
Wiek kierowcy

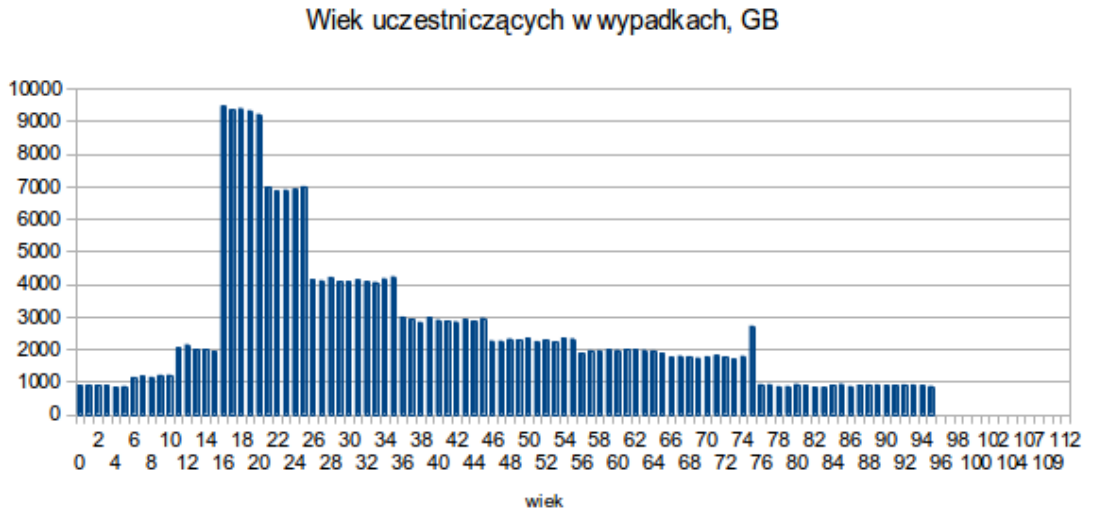


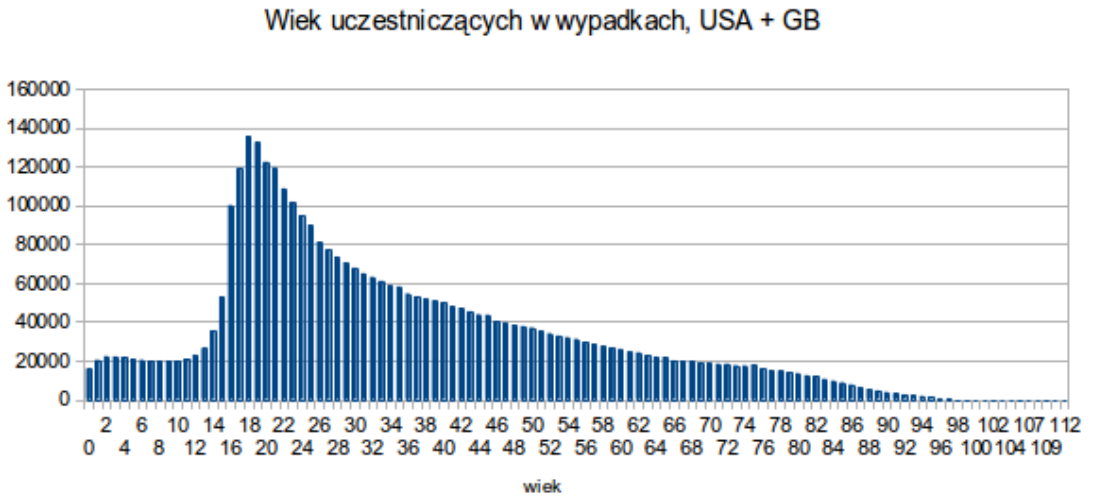




Wiek uczestników wypadku







10.9.3. Weryfikacja i wnioski

Hipoteza potweirdziła się, największa liczba wypadków jest wśród młodych kierowców. Jest to związane przede wszystkim z brakiem doświadczenia i umiejętności a także często z brawurą i zbytnią pewnością siebie.

Niestety z powodu braku dokładnych danych o wieku w danych z Wielkiej Brytanii a jedynie przedziałów wiekowych, nie widać różnicy między wiekiem, w którym kierowcy mogą rozpocząc kierowanie pojazdami. Widać ze w USA jest to 16 - 18 lat (zależnie od satnu), jednak dla danych z Wielkiej Brytanii mozemy stwierdzić jedynie, że jest to gdzieś w przedziale 16 - 21.

Analiza wieku ofiar pokazuje, poprzez swoje podobieństwo do wykresu dla kierowców, że często kierowca jest jedyną osobą w pojeździe bądź też podróżuje ze swoimi rówieśnikami.

10.10. Hipoteza 10

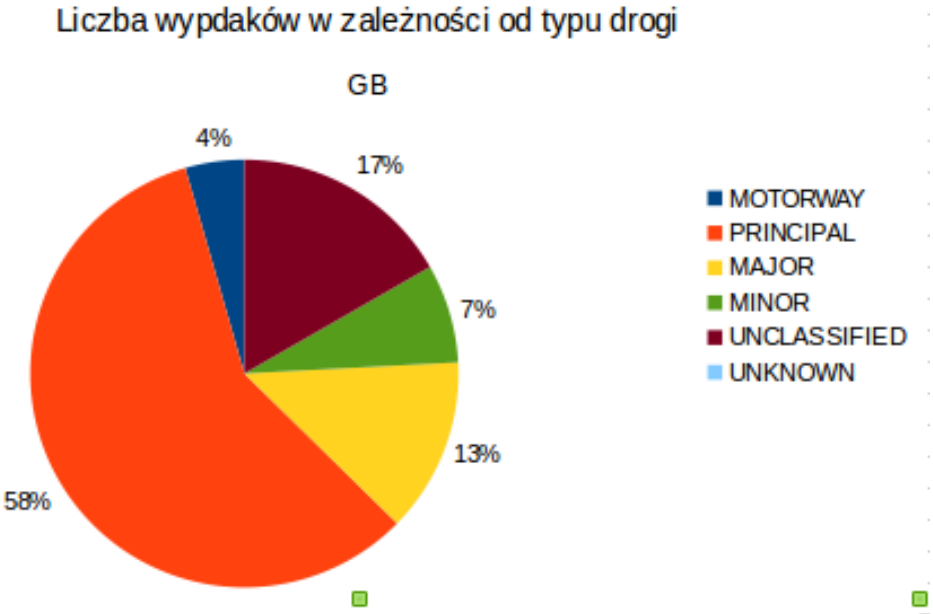
10.10.1. Opis hipotezy

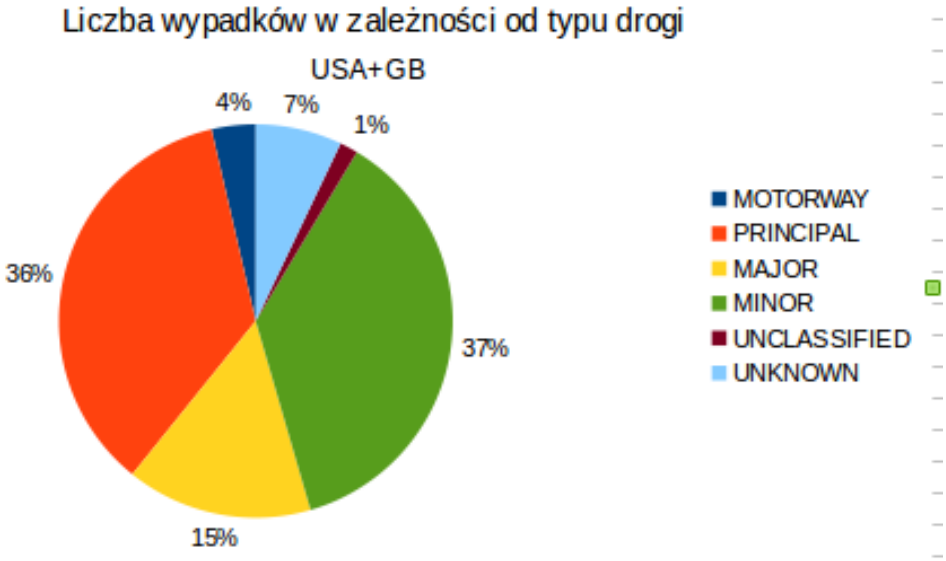
Numer: 10
Nazwa: Wypadki a rodzaj drogi
Treść: Na autostradach będzie mało wypadków, ponieważ są one bezpiecznymi drogami - obowiązuje na nich nieskomplikowany sposób poruszania się, nie ma skrzyżowań, oraz mieszkańcy USA i WB są przyzwyczajeni do częstego korzystania z nich. Ponadto nie poruszają się po nich piesi. Natomiast więcej wypadków może pojawić się na drogach o randze porównywalnej z polskimi krajowymi oraz wojewódzkimi (Principal, Major), ponieważ z jednej strony mogą mieć skomplikowana infrastrukturę co może wymagać więcej umiejętności od kierowców, a z drugiej pozwalają osiągnąć znaczne prędkości. Możliwe jest też zaobserwowanie znacznej ilości wypadków śmiertelnych na drogach drugorzędnych (Minor), ponieważ często na terenach wiejskich ludzie dopuszczają się bardziej brawurowej jazdy i kierowania po spożyciu, gdyż rzadziej można tam spotkać funkcjonariuszy drogówki.

10.10.2. Wyniki związane z hipotezą

Rodzaj drogi


Liczba wypadków w zależności od typu drogi
USA
4%
8%
33%
40%
15%
MOTORWAY
PRINCIPAL
MAJOR
MINOR
UNCLASSIFIED
UNKNOWN


Liczba wypdaków w zależności od typu drogi
GB
4%
17%
7%
13%
58%
MOTORWAY
PRINCIPAL
MAJOR
MINOR
UNCLASSIFIED
UNKNOWN


Liczba wypadków w zależności od typu drogi
USA+GB
4%
7%
1%
36%
37%
15%
MOTORWAY
PRINCIPAL
MAJOR
MINOR
UNCLASSIFIED
UNKNOWN

Hipoteza potwierdziła się, na autostradach obserwuje się niewiele wypadków śmiertelnych w porównaniu do innych typów drogi. Tutaj instnieje spore podobieństwo pomiędzy USA a WB. Natomiast w przypadku innych rodzajów dróg pojawiają się duże rozbieżności, choć wyniki te potwierdzają prawdziwość hipotezy. W WB wypadki śmiertelne częściej spotykane są na drogach krajowych, podczas gdy w USA spora część tego udziału przypada drogom drugorzędnym. Może to wynikać z udziału poszczególnych typów dróg w infrastrukturze danego kraju, lub rozbieżności w ich klasyfikacjach.

11. Podsumowanie

W ramach projektu zintegrowano dane o wypadkach drogowych w Stanach Zjednoczonych Ameryki Północnej oraz Wielkiej Brytanii, w latach 1979 - 2013. Dostępne były szczegółowe informacje o miejscu zdarzenia i okolicznościach wypadku, osobach uczestniczących oraz pojazdach. Po sprowadzeniu do wspólnego formatu, dane zostały umieszczone w bazie danych i poddane analizie. W ramach analizy zbadano różne aspekty zainstniałych wypadków, które mogły bezpośrednio lub pośrednio być ich przyczynami. Zweryfikowano 10 hipotez analitycznych, które postawione były jeszcze przed zebraniem danych. Hipotezy dotyczyły wpływu czynników takich jak pogoda, widoczność, udział pijanych kierowców, wiek kierowców, pora dnia, roku i wielu innych, na występowanie wypadków drogowych. Część hipotez potwierdziła się, część tylko w pewnym stopniu. Natomiast każda z analiz pozwoliła wyciągnąć interesujące wnioski na temat przyczyn wypadków drogowych, nawet jeśli hipoteza nie potwierdzały się.

Bibliografia

[1] Fatality Analysis Reporting System (FARS): Online, accessed 05.06.2015. http://www.nhtsa.gov/FARS

[2] Polskie Obserwatorium Bezpieczeństwa Ruchu Drogowego: Online, accessed 05.06.2015. http://www.obserwatoriumbrd.pl/pl/

[3] Community Road Accident Database: Online, accessed 05.06.2015. http://ec.europa.eu/idabc/en/document/2281/5926.html, http://ec.europa.eu/transport/road_safety/specialist/statistics/index_en.htm

[4] DATA.GOV.UK, Road safety data: Online, accessed 05.06.2015. http://data.gov.uk/dataset/road-accidents-safety-data

[5] Frequent Itemset Mining Dataset Repository: Online, accessed 05.06.2015. http://fimi.ua.ac.be/data/

[6] NHTSA, National Automotive Sampling System: Online, accessed 05.06.2015. http://www.nhtsa.gov/NASS

[7] DATA.GOV.UK, formularz STATS19: Online, accessed 05.06.2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230590/stats19.pdf

[8] DATA.GOV.UK, formularz STATS20: Online, accessed 05.06.2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230596/stats20-2011.pdf

[9] Python programming language: Online, accessed 05.06.2015. https://www.python.org/

[10] sas7bdat conversion library: Online, accessed 05.06.2015. http://github.com/openfisca/sas7bdat

[11] PostgreSQL database: Online, accessed 05.06.2015. http://www.postgresql.org/

[12] R programming language: Online, accessed 05.06.2015. http://www.r-project.org/

[13] WEKA: Online, accessed 05.06.2015. http://www.cs.waikato.ac.nz/ml/weka/

[14] Polymaps: Online, accessed 05.06.2015. http://polymaps.org/

[15] google charts: Online, accessed 05.06.2015. https://developers.google.com/chart/

[16] OpenLayers: Online, accessed 05.06.2015. http://openlayers.org