Pakiety statystyczne - sprawozdanie 1

Wiktoria Fimińska 262283, Julia Grzegorzewska 262314

2022-12-18

Spis treści

1 Wprowadzenie

Niniejsze sprawozdanie jest analizą danych dotyczących wypadków samochodowych w Kanadzie w latach 1999–2014, które zostały pozyskane ze strony kaggle.

Celem pracy jest zbadanie wpływu poszczególnych czynników na liczbę zaistniałych wypadków, sprawdzenie, który z nich w największym stopniu przyczynił się do wystąpienia kolizji, a także przeanalizowanie skutku zdrowotnego wypadku w zależności od innych zmiennych. Wszystkie wykresy, obliczenia oraz przekształcenia danych wykonano przy pomocy języka *R*.

2 Dane

2.1 Opis zmiennych

Pierwotne dane zawierają 22 kolumny, natomiast istotne dla analizy jest 9 z nich. Są to:

(a) **C_YEAR**
Rok kolizji. Wartości od 1999 do 2014.

(b) **C_MNTH**
Miesiąc kolizji. Wartości od 1 do 12.

(c) **C_WDAY**
Dzień kolizji. Wartości od 1 do 7 oznaczające kolejne dni tygodnia.

(d) **C_HOUR**
Godzina kolizji. Wartości od 00 do 23, przy czym np. 1 oznacza przedział od godziny 1^{00} do 1^{59}.

(e) **V_YEAR**
Rok produkcji pojazdu.

(f) **P_SEX**
Płeć osoby biorącej udział w wypadku, przy czym M to mężczyzna, F to kobieta.

(g) **P_AGE**
Wiek osoby biorącej udział w wypadku.

(h) **P_PSN**
Lokalizacja osoby biorącej udział w wypadku, tj. konkretne siedzenie w samochodzie, bądź status pieszego.

(i) **P_ISEV**
Stan zdrowotny po kolizji osoby biorącej w niej udział. Wartości 1-3 oznaczają kolejno uraz, brak urazu oraz wypadek śmiertelny.

2.2 Przygotowanie danych

W celu ułatwienia pracy oraz większej przejrzystości analizy podjęto następujące kroki:

- usunięto nieistotne z punktu widzenia niniejszego sprawozdania kolumny, takie jak np. dokładna konfiguracja kolizji, czy rodzaj skrzyżowania, na którym owa miała miejsce;
- dodano kolumnę **C_DATE** zawierającą datę wydarzenia podaną w typie danych *yearmon*, która jest sformatowana dzięki bibliotece *zoo*;
- zmieniono typ danych w pozostałych kolumnach z *character* na *double*, z wyjątkiem kolumny **P_SEX**, która nie zawiera danych liczbowych;
- wycięto wiersze, dla których wartość **P_PSN** jest inna niż liczba 11, która oznacza kierowcę, aby uniknąć występowania danych z jednego zdarzenia wiele razy (pomijamy pasażerów i uznajemy, że jeden kierowca to jeden wypadek);
- usunięto wiersze zawierające wartości brakujące *NA*, a także pominięto te, dla których **P_SEX** jest inna niż F - female lub M - male;
- ze względu na pewne nieścisłości w danych, usunięto wiersze, w których wiek kierowcy jest mniejszy niż 16, czyli minimalny wiek, od którego można ubiegać się prawo jazdy w Kanadzie.

W ten sposób z 5860405 obserwacji otrzymano 3209732, czyli $54,77\%$, jednak ze względu na liczność zbioru danych oraz skupienie się na jednej grupie osób nie zaburza to wyników analizy. Po przekształceniach pierwsze 10 obserwacji przedstawia się następująco (tabela 1).

C_YEAR	C_MNTH	C_WDAY	C_HOUR	V_YEAR	P_SEX	P_AGE	P_ISEV
1999	1	1	20	1990	M	41	1
1999	1	1	20	1987	M	19	1
1999	1	1	8	1986	M	46	1
1999	1	1	17	1984	M	28	1
1999	1	1	17	1991	M	21	1
1999	1	1	15	1997	M	61	1
1999	1	1	14	1993	F	34	1
1999	1	1	14	1997	F	34	2
1999	1	1	1	1985	M	22	2
1999	1	1	11	1988	F	30	2

Tabela 1: Przygotowane do analizy dane - pierwsze 10 obserwacji.

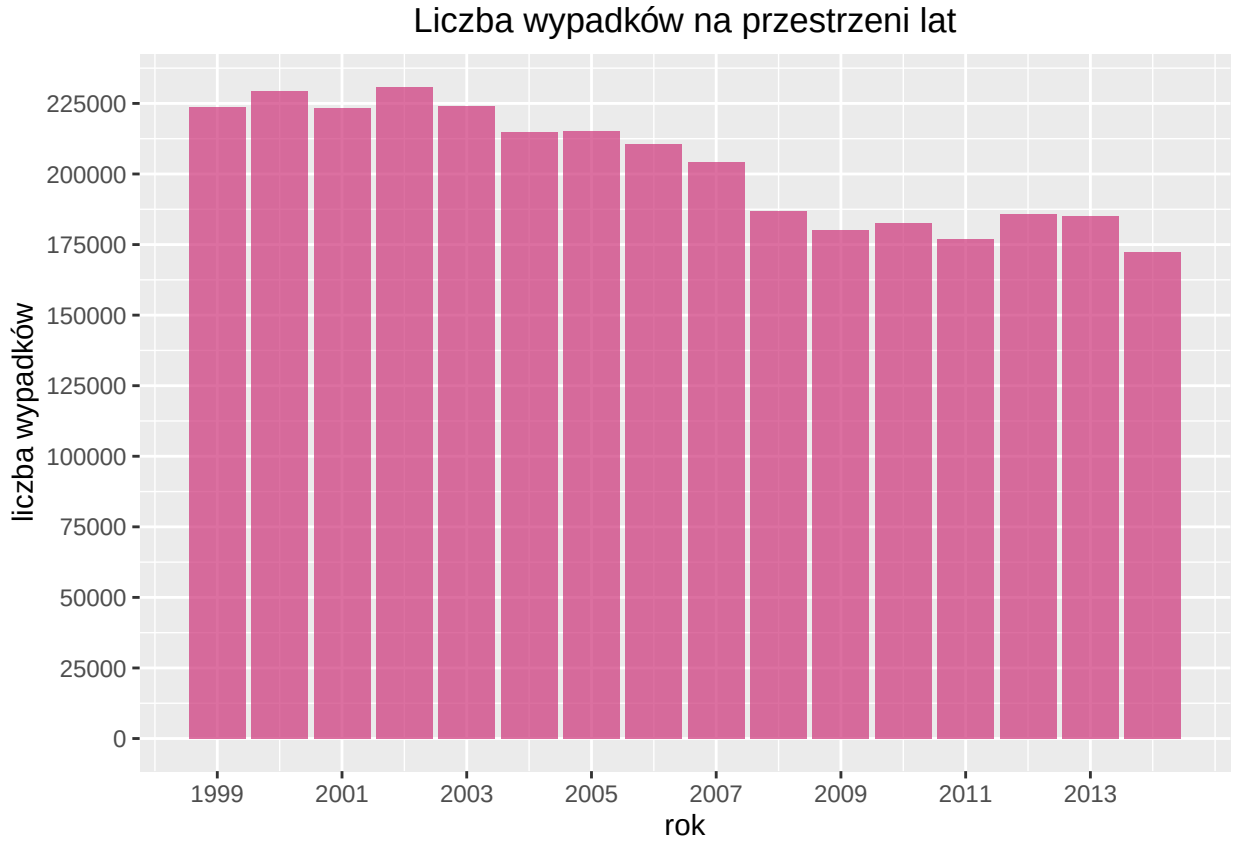
Z uwagi na brak znaczącej liczby zmiennych ciągłych skorzystano z metody *group_by* z biblioteki *dplyr*, aby pogrupować i zliczyć interesujące dane. Zastosowano ją w celu zobrazowania zależności m.in na wykresie typu *heatmap* liczby wypadków od miesięcy. Na wykresie przedstawiającym zależność liczby wypadków od wieku, oprócz wspominanej wcześniej metody, skorzystano z funkcji *discretize* z paczki *arules*, która podzieliła wiek kierowcy na trzy przedziały według liczebności danych, następnie przy pomocy *aggregate* zliczono dla tych przedziałów wartości danych na przestrzeni miesięcy i lat.

3 Analiza

3.1 Zależność od daty wypadku

3.1.1 Liczba wypadków na przestrzeni lat

Pierwszy podjęty krok miał na celu zobrazowanie, jak wyglądała sytuacja z kolizjami w kolejnych latach, od roku 1999 do roku 2014. Aby tego dokonać, należało zliczyć liczbę wierszy w zbiorze danych dla każdego roku. Tak otrzymane wartości przedstawiono na wykresie słupkowym (rysunek 1). Można zauważyć, że liczba wypadków była podobna w latach $1999-2003$ (średnio 223 722) oraz w latach $2004-2007$ (średnio 207 932). Następnie zaobserwowano ich lekki spadek i liczba kolizji do roku 2014 utrzymywała się średnio na poziomie 179 913. Może to świadczyć o wzroście świadomości kierowców, ich bezpieczniejszej jeździe lub być skutkiem surowszych kar za łamanie przepisów drogowych. Inną możliwą przyczyną mogła być poprawa stanu nawierzchni dróg.

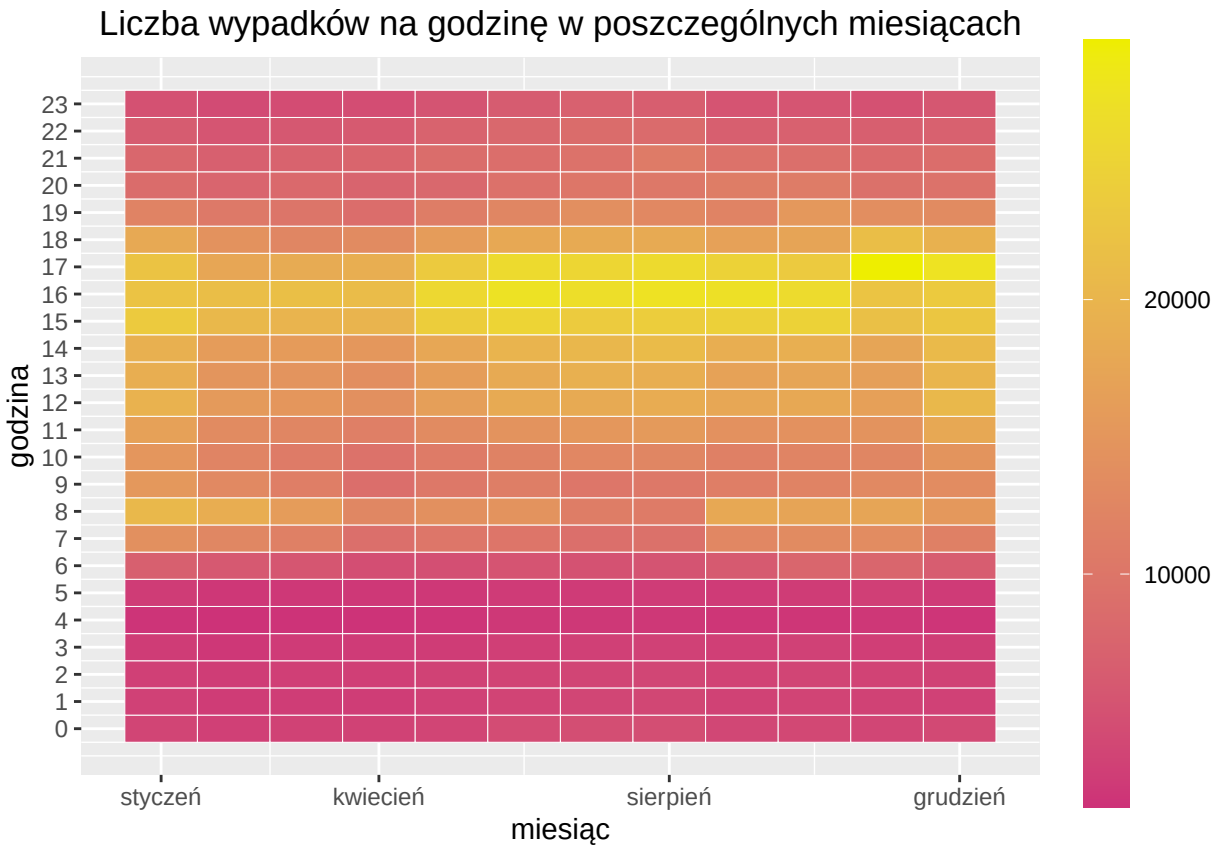


Rysunek 1: Liczba wypadków na przestrzeni lat

3.1.2 Liczba wypadków w zależności od miesiąca oraz godziny

Następnie sprawdzono jak liczba wypadków zmieniała się w poszczególnych godzinach w każdym miesiącu (rysunek 2). W porze nocnej (od godziny 23 do 6), biorąc pod uwagę fakt, że analizujemy tutaj okres 16 lat, jest ona znikoma (poniżej 600 na każdą godzinę), co

zapewne jest spowodowane zmniejszonym natężeniem ruchu. W godzinach od 7 do 9, podczas których zazwyczaj ludzie dojeżdżają do pracy lub szkoły, zauważalny jest wzrost liczby kolizji. Najwięcej wypadków w tych godzinach nastąpiło w miesiącach jesienno-zimowych, czyli od września do marca. Kolejne widoczne zwiększenie liczby wypadków można zaobserwować w popołudniowych godzinach szczytu, czyli od 15 do 17. Tym razem rozkładają się one w miarę równomiernie w każdym miesiącu, z wyjątkiem kwietnia, w którym to liczba wypadków była najmniejsza spośród wszystkich miesięcy. Co ciekawe, ich liczba jest większa niż w porannych godzinach o zwiększonym natężeniu ruchu. Generalnie można stwierdzić, że największe znaczenie w tym przypadku miał fakt, iż niektóre godziny są godzinami o wzmożonej intensywności przemieszczania się za pomocą samochodu.



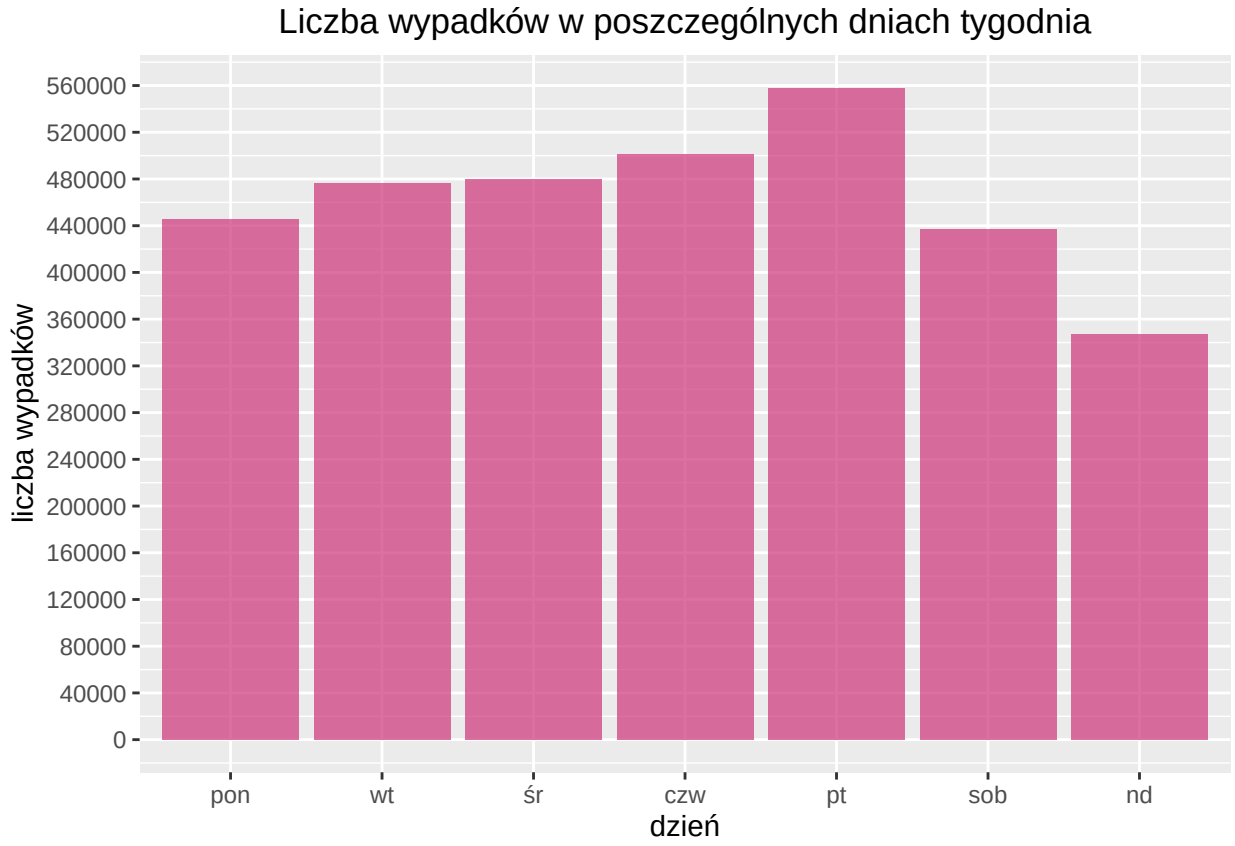
Rysunek 2: Liczba wypadków w zależności od miesiąca i godziny

3.1.3 Liczba wypadków w zależności od dnia tygodnia

Kolejnym czynnikiem mogącym mieć wpływ na liczbę wypadków jest dzień tygodnia, w którym owy miał miejsce. Ponownie, aby zobrazować sytuację, użyto wykresu słupkowego (rysunek 3). Uwzględnione w nim są wszystkie kolizje w latach 1999 – 2014.

Mocno zauważalny spadek w obserwowanych danych nastąpił pod koniec tygodnia, czyli w dni wolne od pracy i szkoły.. Wydaje się to być naturalnym zjawiskiem, biorać pod uwagę fakt, że zazwyczaj jest to dzień spędzania czasu z rodziną w domowym zaciszu. Natomiast znaczny

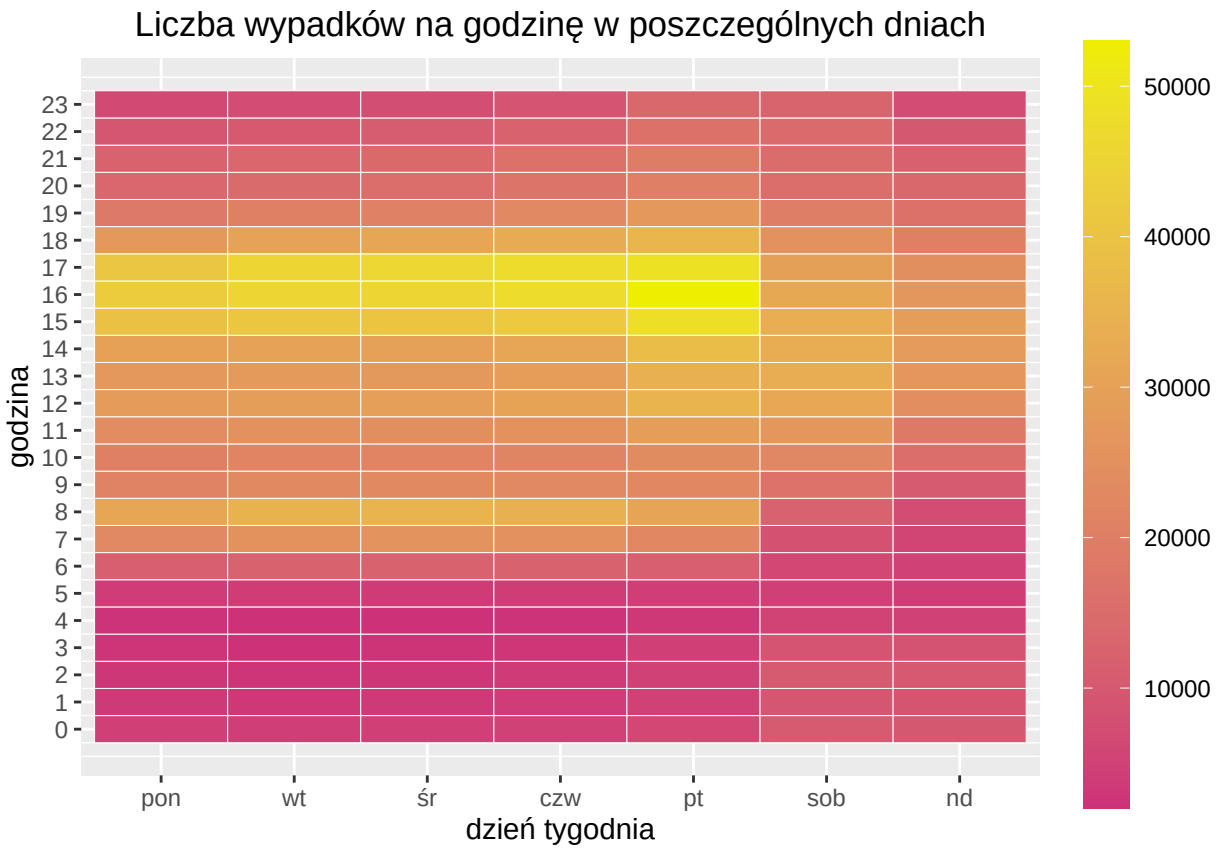
wzrost, bo aż o 56142 kolizji w stosunku do dnia poprzedniego, widoczny jest w piątek, co może być spowodowane wzmożonym ruchem drogowym związanym z powrotem ludzi do rodzinnych stron, weekendowym wyjazdem na urlop, bądź także powrotem z wszelakich imprez, po których to stan kierowcy nie był odpowiedni do prowadzenia pojazdu. Wnioskować więc można o zależności pomiędzy dniami tygodnia, a liczbą zaistniałych wypadków.



Rysunek 3: Liczba wypadków w zależności od dnia tygodnia

3.1.4 Liczba wypadków w zależności od dnia tygodnia i godziny

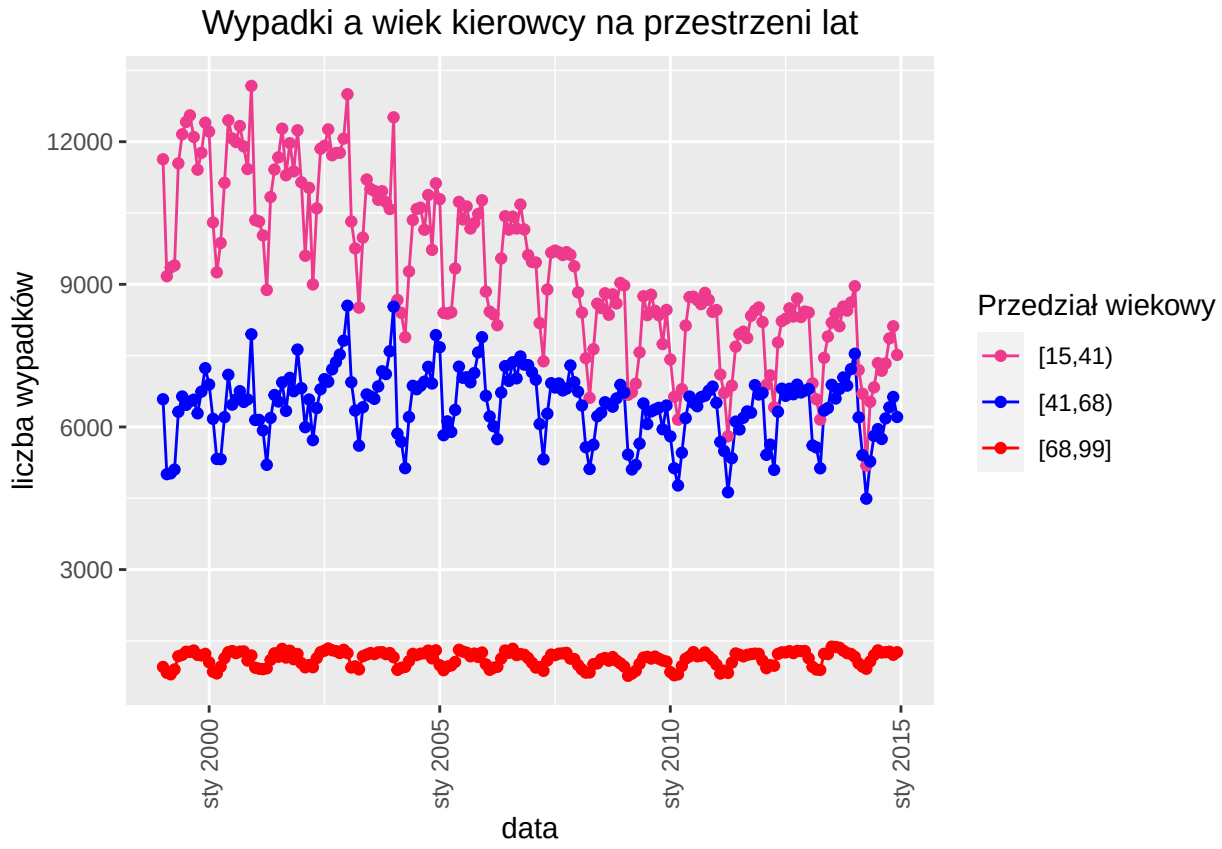
Sprawdzono także, jak rozkłada się liczba wypadków na poszczególne godziny z podziałem na konkretne dni tygodnia. Wyniki ponownie przedstawiono na heatmapie (rysunek 4). W dni robocze (w Kanadzie także obowiązuje pięciodniowy tydzień pracy) występuje już wcześniej zauważona tendencja wzrostowa liczby wypadków w godzinach szczytu. W dni wolne od pracy, czyli soboty i niedziele, zależność ta, jak można było się domyślać, nie występuje. Liczba wypadków w tych dniach jest nieznacznie podwyższona w godzinach popołudniowych, co może być spowodowane wyjazdami w odwiedziny w rodzinne strony, czy też najzwyczajniej w świecie robieniem zakupów. W piątki, czyli dni, w których wystąpiło najwięcej wypadków, liczba tych zdarzeń w poszczególnych godziny rozkłada się proporcjonalnie w stosunku do pozostałych dni tygodnia.



Rysunek 4: Liczba wypadków w zależności od dnia tygodnia i godziny

3.2 Wiek kierowcy biorącego udział w wypadku na przestrzeni lat

Kolejnym podjętym krokiem było zbadanie wpływu wieku kierowcy pojazdu biorącego udział w wypadku na liczbę tych zdarzeń w latach 1999 – 2014. Aby spojrzeć z innej strony na analizowane dane, podzielono je na trzy grupy wiekowe a mianowicie [15, 41), [41,68) oraz [68, 99]. Otrzymane wyniki przedstawiono na wykresie liniowo-punktowym (rysunek 5). Największa liczba wypadków dla najmłodszej grupy wiekowej oraz najmniejsza dla najstarszej wynika przypuszczalnie z różnic w samej liczbie kierowców z poszczególnych grup wiekowych (najwięcej młodych kierowców, najmniej starszych). Zauważalna jest tu pewna sezonowość, a mianowicie co roku w okolicach kwietnia następował gwałtowny spadek liczby kolizji w każdej grupie wiekowej, co jest ciekawą zależnością. Patrząc kolejno na każdą z grup, widać, że w przedziale wiekowym [15, 41) suma wypadków w kolejnych latach maleje. Dla grupy [68, 99] liczba wypadków utrzymuje się w miarę na równym poziomie, za wyjątkiem wspomnianego wcześniej spadku w kwietniu. W środkowym przedziale wiekowym nie widać szczególnych zależności w liczbie wypadków w kolejnych latach.

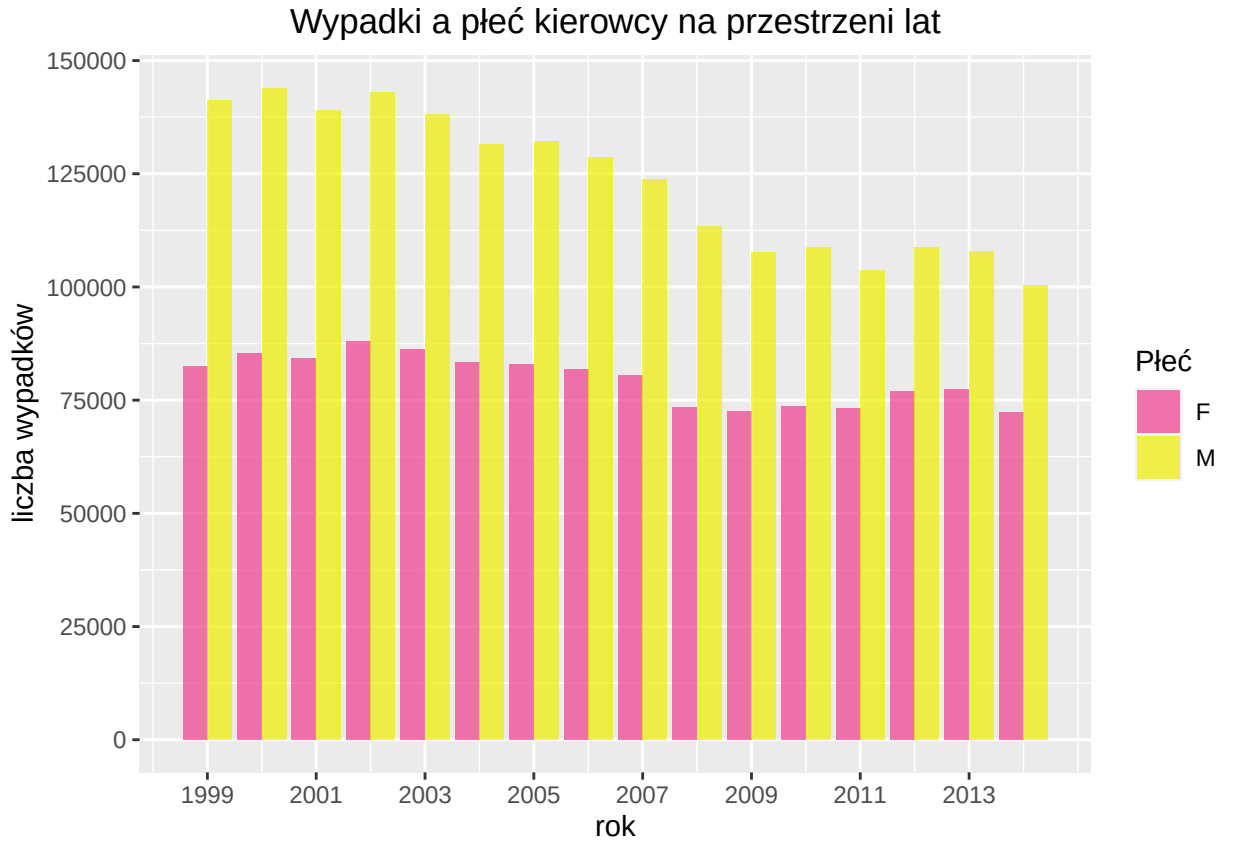


Rysunek 5: Wypadki a wiek kierowcy na przestrzeni lat

3.3 Wypadki a płeć kierowcy

Przeanalizowano także liczbę wypadków spowodowanych przez mężczyzn i kobiety na przestrzeni lat (rysunek 6). Sumarycznie mężczyźni spowodowali 1954792 wypadków, czyli aż

o 699852 więcej od kobiet, co może sugerować, że kobiety jeżdżą bezpieczniej. Jednak ze względu na brak danych dotyczących stosunku ogólnej liczby kierowców płci męskiej do żeńskiej, nie da się jednoznacznie określić przyczyny takiej sytuacji. Można natomiast zauważyć, że w przypadku płci męskiej bardziej zauważalny jest spadek liczby wypadków w kolejnych latach, co może świadczyć o rosnącej świadomości na temat bezpiecznej jazdy w tej grupie osób.



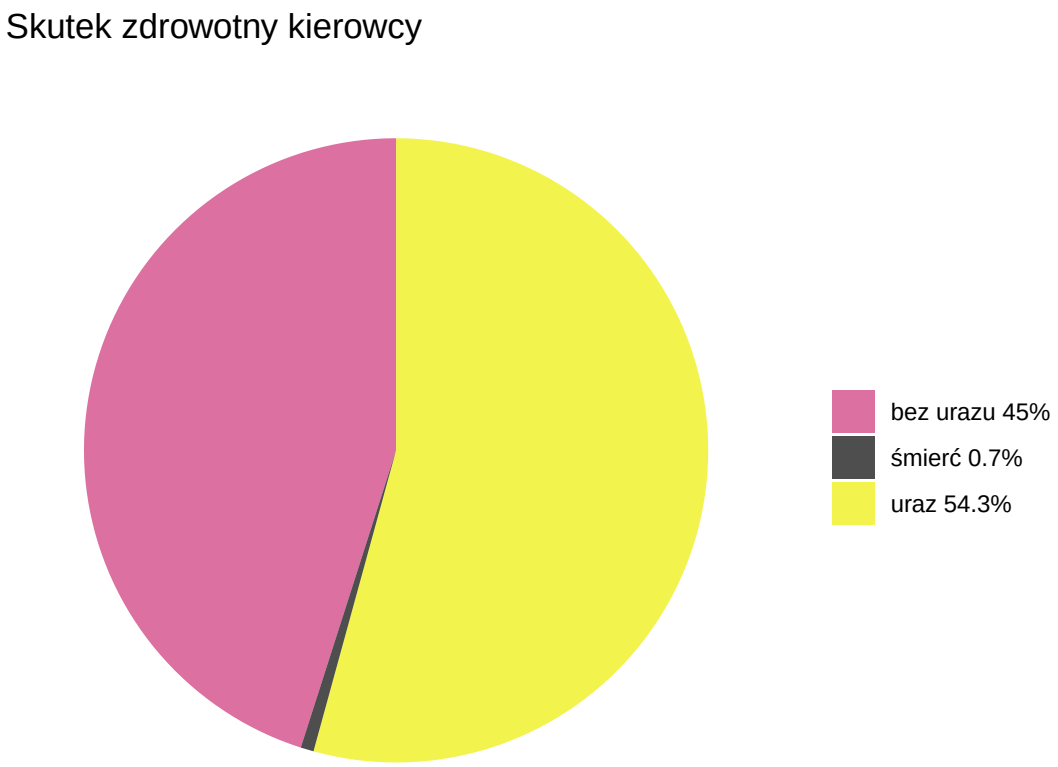
Rysunek 6: Wypadki a płeć kierowcy na przestrzeni lat

3.4 Skutek zdrowotny kierowcy po wypadku

Jeśli chodzi o skutek zdrowotny kierowcy po wypadku, to najwięcej z zaistniałych zdarzeń zakończyło się urazem. Trochę mniej, bo o 9.3 punktów procentowych wypadków nie spowodowało żadnego uszczerbku na zdrowiu, a 0.7 procent z nich niestety było śmiertelnych (rysunek 7).

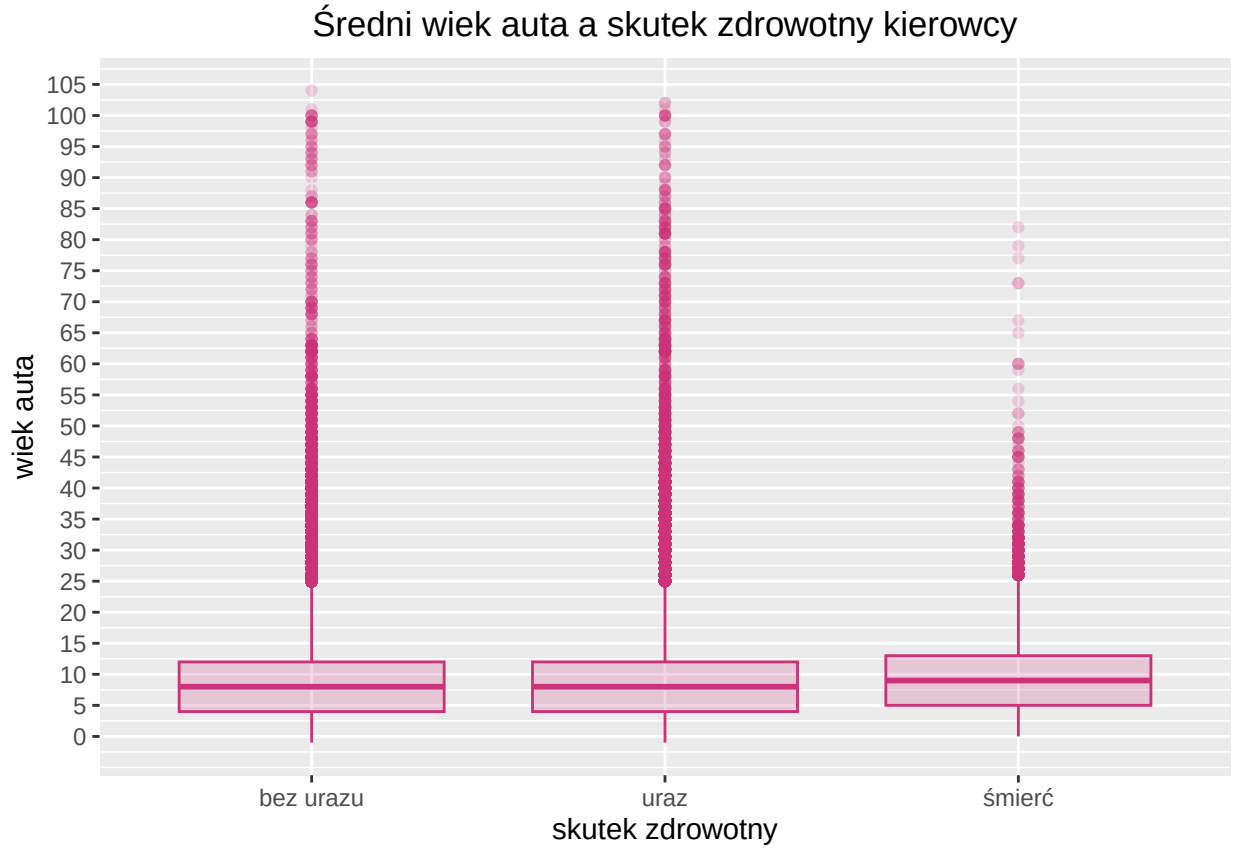
3.4.1 Zależność od wieku auta

Zbadano wpływ wieku auta na stan zdrowotny kierowcy po wypadku (rysunek 8). Najpierw, aby uzyskać wiek auta, od kolumny **C_YEAR** odjęto kolumnę **V_YEAR**. Na każdym wykresie pudełkowym zauważyć można wiele wartości odstających, co wynika z faktu, że auta wieloletnie nie są często spotykane na drogach, a raczej są traktowane, jako zabytki



Rysunek 7: Skutek zdrowotny kierowcy po wypadku

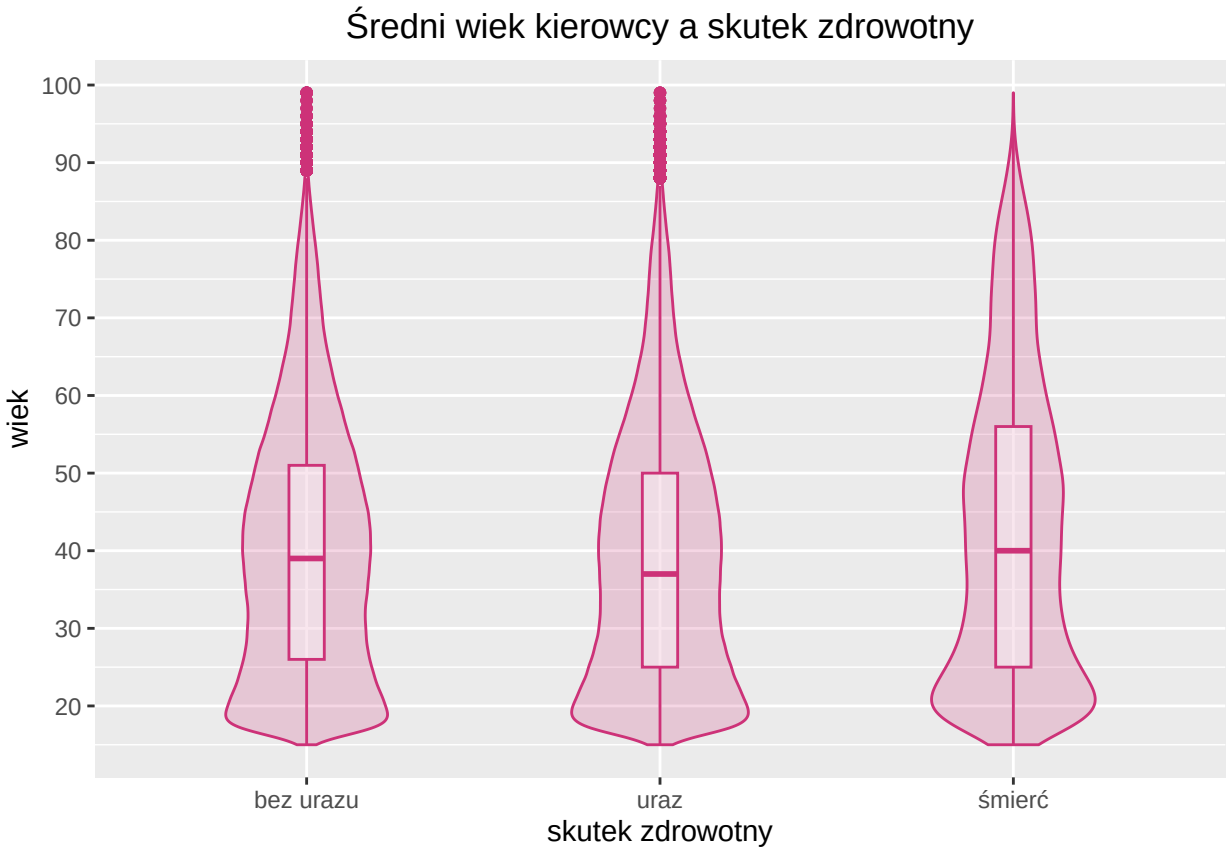
i eksponaty. Mediana wieku auta biorącego udział w wypadku urazowym oraz bezurazowym wynosi około 8, natomiast dla wypadku śmiertelnego około 9. Widoczne jest, że wypadki śmiertelne powodowane były przez nieco starsze auta niż pozostałe.



Rysunek 8: Wiek auta a skutek zdrowotny kierowcy po wypadku

3.4.2 Zależność od wieku kierowcy

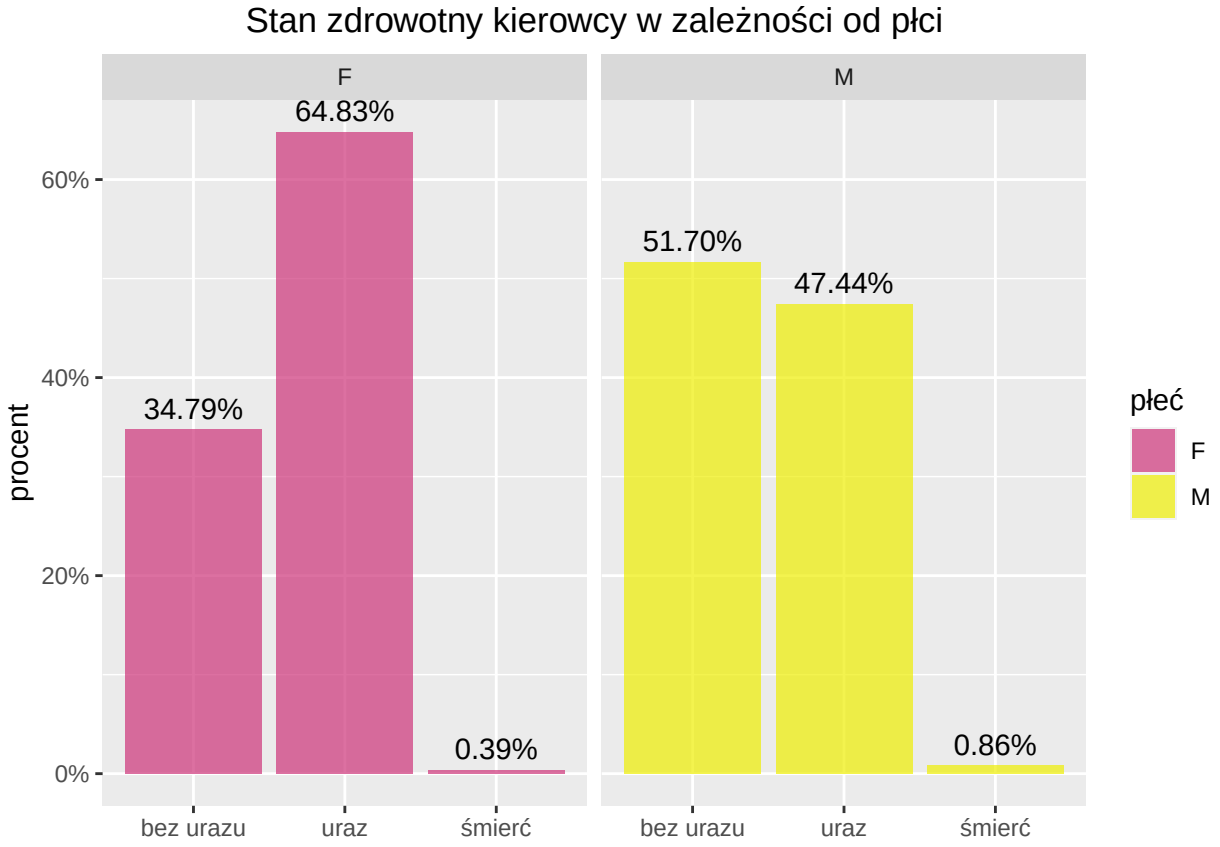
Przeanalizowano również jaki wpływ na stan zdrowotny kierowcy po wypadku miał jego wiek. Dane zobrazowano za pomocą wykresu wiolinowego (rysunek 9), który jest połączeniem wykresu pudełkowego z estymatorem jądrowym gęstośći KDE. Dla wypadków bezurazowych mediana wieku kierowcy to około 40 lat, dla urazowych 37, a dla śmiertelnych nieco ponad 40. Dzięki specyfice wykresu można jednak zauważyć, że najbardziej liczną grupą osób powodujących każdy z rozważanych rodzajów wypadków są młodzi kierowcy, czyli tacy przed 20 rokiem życia. Na pierwszych dwóch wykresach można zauważyć większe podobieństwo kształtu linii KDE, jednak w wypadkach kończących się urazami zauważyć można szersze wybrzuszenie dla młodych kierowców. Ciekawe spostrzeżenia zauważa się także przy ostatnim z wykresów. Mianowicie, ewidentnie największą grupą osób powodującą wypadki śmiertelne są młodzi kierowcy. Widoczny jest w tym przypadku również stosunkowy wzrost fatalnych skutków wypadku dla osób starszych, miedzy 70 a 80 rokiem życia.



Rysunek 9: Wiek kierowcy a skutek zdrowotny

3.4.3 Zależność od płci kierowcy

Na koniec sprawdzono zależność od płci kierowcy (rysunek 10). Zdecydowana większość wypadków spowodowanych przez kobiety zakończyła się dla nich uszczerbkiem na zdrowiu. W przypadku mężczyzn liczba wypadków urazowych i bezurazowych jest bardzo podobna. Jeśli chodzi o wypadki śmiertnelne to procentowo mężczyźni powodują je częściej. Zależności te mogą prowadzić do wniosku, że w sytuacjach kryzysowych na drodze mężczyźni posiadają lepsze umiejętności zachowania się w sposób odpowiedni do zaistniałej sytuacji. Z drugiej strony natomiast mniejsza śmiertelność u kobiet może być spowodowana faktem, iż jeżdżą one wolniej i bezpieczniej, przez co wypadki nie mają tak drastycznych skutków.



Rysunek 10: Stan zdrowotny kierowcy w zależności od jego płci

4 Podsumowanie

Podczas dokonanej analizy zauważono wiele ciekawych obserwacji dotyczących wypadków samochodowych w Kanadzie oraz dokonano ich interpretacji. Zauważono, że liczba wypadków na przestrzeni lat stopniowo malała, co mogło świadczyć o rosnącej świadomości kierowców na temat bezpiecznej jazdy. Nie jest zaskakujące, że zdarzenia te występowały częściej w środku tygodnia niż w dni wolne od pracy. Ciekawą zależnością jest znaczny spadek liczby wypadków w kwietniu. Widoczne jest również, że najwięcej wypadków powodują młodzi kierowcy, jednak

jest to zapewne spowodowane ich największą ilością na drodze w stosunku do innych grup wiekowych. Natomiast zdecydowanie największy wpływ na liczbę wypadków miała pora dnia, co wynika ze wzmożonego natężenia ruchu drogowego w pewnych godzinach. Jeśli chodzi o stan zdrowotny kierowcy po wypadku, to wiek auta minimalnie przyczynił się do większej śmiertelności zaistniałych wypadków. Więcej urazowych wypadków powodują najmłodsi oraz najstarsi kierowcy. U tych drugich widoczna jest także zwiększona ilość wypadków śmiertelnych, co może skłonić do przemyśleń na temat słuszności prowadzenia pojazdów przez osoby starsze, ze względu na ich zazwyczaj gorszy stan zdrowotny. Zależność od płci kierowcy jest również interesująca ze względu na fakt, że wypadki spowodowane przez kobiety o wiele częściej kończą sie urazem, natomiast mężczyźni procentowo powodują więcej wypadków śmiertelnych niż płeć przeciwna.