НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет прикладної математики

Кафедра прикладної математики

Лабораторна робота №1

із дисципліни “Аналіз даних”

на тему:

“Розвідковий аналіз даних”

| Виконали: | Керівник: |
| --- | --- |
| команда | старший викладач |
| “CrAzy\_pEnGuIn\_cOdeRs”  Атаман Юлія, КМ-11  Кракович Борислав, КМ-11  Лушников Іван, КМ-11  Шушпаннікова Інна, КМ-11  Ягода Євгенія, КМ-11 | Тавров Д.Ю. |

Київ — 2024

Зміст

[ВСТУП](#_n1qh3whej5tf) 3

[Підготовка даних до роботи](#_b6bjx8nitgbr) 5

[1. Як особливості місцевості та дорожнього покриття (стан доріг, дорожні роботи тощо) впливають на ймовірність виникнення аварій та тяжкість?](#_5sj5b9iqyprm) 27

[2. Вплив характеристик автомобіля](#_meukyafmyhnp) 30

[3. Час/день/рік. Вплив на кількість аварій та їх серйозність](#_s21g8kyv7kcp) 31

[4. Вплив фізичних особливостей (віку та статі) водія на поведінку на дорозі?](#_ya2bn78a4exl) 35

[5. Як фактори на які не може впливати водій впливають на кількість/серйозність аварій?](#_iksl0jkt1vcv) 37

# 

# ВСТУП

Мотивація дослідження

Мотивація роботи полягає у дослідженні інформації з бази даних про дорожньо-транспортні пригоди у Великобританії в період з 2006 по 2015 роки з метою їх аналізу та встановлення зв’язків між характеристиками транспортних засобів, водіїв, дорожніх покриттів тощо та власне ДТП для виявлення факторів, що можуть бути причинами ДТП. Результати роботи можна використати як інформаційні матеріали, що можуть:

* підвищити обізнаність водіїв і пішоходів, що допоможе запобігти аваріям
* бути передані і використані відповідними організаціями для покращення ситуацій на дорогах
* дати інформацію про недоліки або проблеми деяких транспортних засобів, що можуть призвести до ДТП.

Дослідницькі питання

1. Як особливості місцевості та дорожнього покриття (стан доріг, дорожні роботи тощо) впливають на ймовірність виникнення аварій та тяжкість?

2. Як впливають характеристики автомобіля на кількість/серйозність аварій?

3. Як впливає час/день/рік на кількість/серйозність аварій?

4. Залежність фізичних особливостей водія на поведінку на дорозі?

5. Як фактори на які не може впливати водій впливають на кількість/серйозність аварій?

Очікування і гіпотези

1. Молоді водії та підлітки мають більшу ймовірність потрапляння у ДТП, але старші водії можуть мати більш важкі наслідки внаслідок своїх фізичних обмежень.

2. Несприятливі погодні умови мають значний вплив на серйозність ДТП.

3. Стан дорожнього покриття і освітленість полотна не повинні мати значного впливу на кількість ДТП чи їхню важкість, оскільки поведінка водіїв на дорозі має змінюватись з огляду на ці фактори (увімкнення фар, зменшення швидкості під час руху на пошкодженому чи мокрому полотні).

4. Є відмінності між кількістю та важкістю ДТП залежно від статі від статі та віку водія.

5. Напередодні свят і вихідних збільшується кількість легких ДТП.

6. Автомобілі дешевих брендів і старі авто (>5 років) з більшою вірогідністю потрапляють у ДТП.

# Підготовка даних до роботи

Даний набір даних про дорожньо-транспортні пригоди та залучені транспортні засоби у Великобританії (2005-2017 рр.).

Даний датасет містить два файли:  
 Accident\_Information.csv: кожен рядок у файлі представляє унікальну дорожньо-транспортну пригоду (ідентифіковану стовпцем Accident\_Index), що містить різні властивості, пов’язані з аварією. Період: 2005-2017 рр

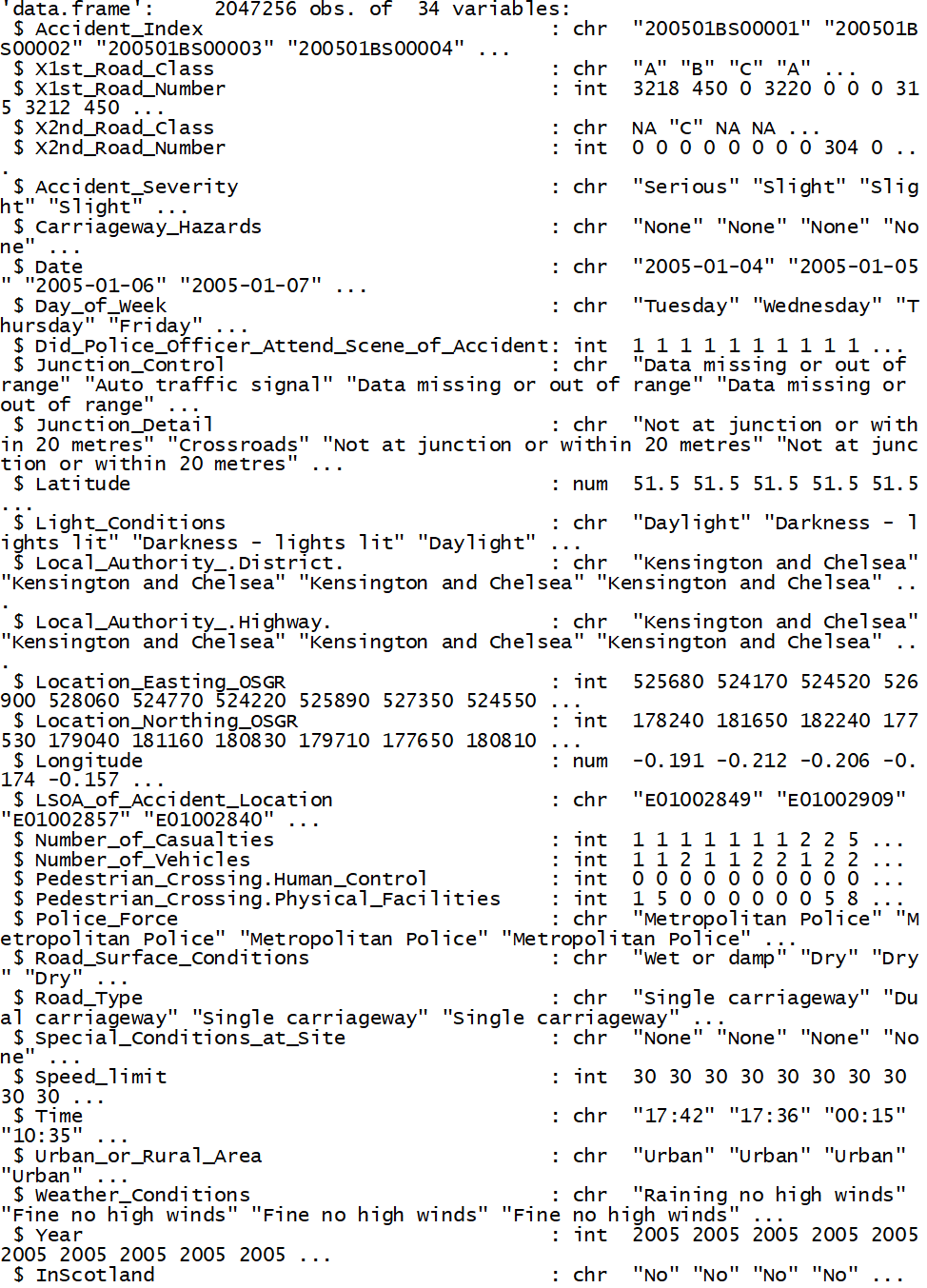
Датасет перейменовується на df\_Accident\_Information

Vehicle\_Information.csv: кожен рядок у файлі відображає участь унікального транспортного засобу в унікальній дорожньо-транспортній пригоді, показуючи різні властивості автомобіля та пасажирів. Період: 2004-2016 рр

Датасет перейменовується на df\_Vehicle\_Information

Інформація про поля:  
df\_Accident\_Information:

| № | Назва поля | Значення |
| --- | --- | --- |
| 1 | Accident\_Index | Неперервні значення |
| 2 | X1st\_Road\_Class | A,B,C, Unclassified, Motorway, A(M) |
| 3 | X1st\_Road\_Number | Цілі числа |
| 4 | X2nd\_Road\_Class | A,B,C, Unclassified, Motorway, A(M), NA |
| 5 | X2nd\_Road\_Number | Цілі числа |
| 6 | Accident\_Severity | Slight, Serious, Fatal |
| 7 | Carriageway\_Hazards | None, Other object on road, Pedestrian in carriageway - not injured, Vehicle load on road, Previous accident, Any animal in carriageway (except ridden horse), Data missing or out of range |
| 8 | Date | Неперервні значення |
| 9 | Day\_of\_Week | Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday |
| 10 | Did\_Police\_Officer\_Attend\_Scene\_of\_Accident | 1,2,3,NA |
| 11 | Junction\_Control | Auto traffic signal, Give way or uncontrolled, Stop sign, Authorised person, Not at junction or within 20 metres, Data missing or out of range |
| 12 | Junction\_Detail | Not at junction or within 20 metres, Crossroads T or staggered junction, Mini-roundabout, Other junction, More than 4 arms (not roundabout), Private drive or entrance, Roundabout, Slip road, Data missing or out of range |
| 13 | Latitude | Цілі числа |
| 14 | Light\_Conditions | Daylight, Darkness - lights lit, Darkness - lighting unknown, Darkness - lights unlit, Darkness - no lighting, Data missing or out of range |
| 15 | Local\_Authority\_.District | Текстові значення |
| 16 | Local\_Authority\_.Highway | Текстові значення |
| 17 | Location\_Easting\_OSGR | Цілі числа |
| 18 | Location\_Northing\_OSGR | Цілі числа |
| 19 | Longitude | Цілі числа |
| 20 | LSOA\_of\_Accident\_Location | Неперервні значення |
| 21 | Number\_of\_Casualties | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 32, 33, 34, 35, 36, 38, 40, 41, 42, 43, 45, 46, 47, 48, 51, 54, 58, 62, 63, 68, 70, 87, 93 |
| 22 | Number\_of\_Vehicles | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 28, 29, 32, 34, 37, 67 |
| 23 | Pedestrian\_Crossing.Human\_Control | 0, 1, 2, NA |
| 24 | Pedestrian\_Crossing.Physical\_Facilities | 0, 1, 4, 5, 7, 8, NA |
| 25 | Police\_Force | Metropolitan Police, City of London, Cumbria. Lancashire, Merseyside Greater, Manchester, Cheshire, Northumbria, Durham, North Yorkshire, West Yorkshire, South Yorkshire, Humberside, Cleveland, West Midlands, Staffordshire, West Mercia, Warwickshire, Derbyshire, Nottinghamshire, Lincolnshire, Leicestershire, Northamptonshire, Cambridgeshire, Norfolk, Suffolk, Bedfordshire, Hertfordshire, Essex, Thames, Valley Hampshire, Surrey Kent, Sussex, Devon and Cornwall, Avon and Somerset, Gloucestershire, Wiltshire Dorset, North Wales, Gwent, South Wales, Dyfed-Powys, Northern Grampian, Tayside Fife, Lothian and Borders, Central Strathclyde, Dumfries and Galloway |
| 26 | Road\_Surface\_Conditions | Wet or damp, Dry, Frost or ice, Snow, Flood over 3cm. deep, Data missing or out of range |
| 27 | Road\_Type | Single carriageway, Dual carriageway, One way street, Roundabout, Slip road, Unknown, Data missing or out of range |
| 28 | Special\_Conditions\_at\_Site | None, Oil or diesel, Roadworks, Auto signal part defective, Road surface defective, Auto traffic signal – out, Road sign or marking defective or obscured, Mud, Data missing or out of range |
| 29 | Speed\_limit | 0, 10, 15, 20, 30, 40, 50, 60, 70, NA |
| 30 | Time | HH:MM |
| 31 | Urban\_or\_Rural\_Area | Urban, Rural, Unallocated |
| 32 | Weather\_Conditions | Raining no high winds, Fine no high winds, Unknown, Snowing no high winds, Other, Fine + high winds, Raining + high winds, Fog or mist, Snowing + high winds, Data missing or out of range |
| 33 | Year | 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017 |
| 34 | InScotland | Yes, No, NA |

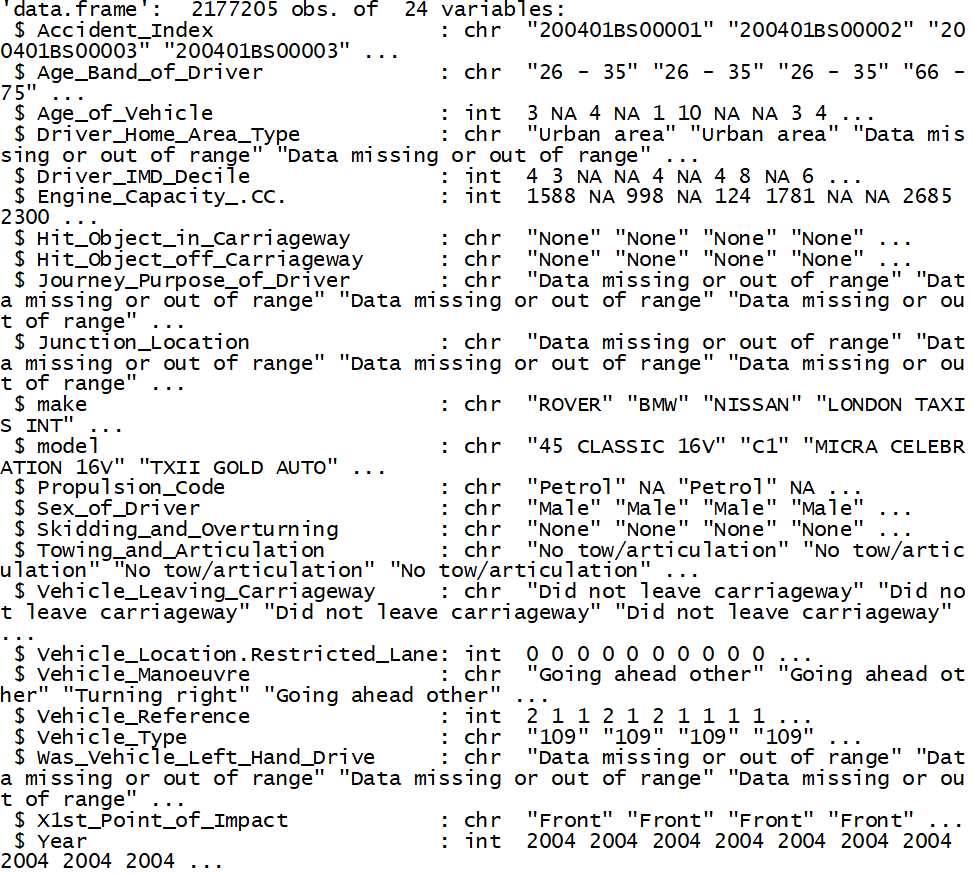


Набір даних має 34 поля та 2047256 спостережень

Категорійні змінні: X1st\_Road\_Class, X2nd\_Road\_Class, Accident\_Severity, Carriageway\_Hazards, Day\_of\_Week, Junction\_Control, Junction\_Detail, Light\_Conditions, Local\_Authority\_.District, Police\_Force, Road\_Surface\_Conditions, Road\_Type, Special\_Conditions\_at\_Site, Urban\_or\_Rural\_Area, Weather\_Conditions, Year, InScotland

df\_Vehicle\_Information:

| № | Назва поля | Значення |
| --- | --- | --- |
| 1 | Accident\_Index | Неперервні значення |
| 2 | Age\_Band\_of\_Driver | 0 - 5, 6 - 10, 11 - 15, 16 - 20, 21 - 25, 26 - 35, 36 - 45, 46 - 55, 56 - 65, 66 - 75, Over 75,  Data missing or out of range |
| 3 | Age\_of\_Vehicle | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 83, 84, 85, 90, 91, 95, 104, 105, 111, NA |
| 4 | Driver\_Home\_Area\_Type | Urban area, Rural, Small town, Data missing or out of range |
| 5 | Driver\_IMD\_Decile | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, NA |
| 6 | Engine\_Capacity\_.CC. | Цілі числа |
| 7 | Hit\_Object\_in\_Carriageway | None, Parked vehicle, Kerb, Previous accident, Open door of vehicle, Bollard or refuge, Road works, Other object, Bridge (roof), Central island of roundabout, Bridge (side), Any animal (except ridden horse), Data missing or out of range, |
| 8 | Hit\_Object\_off\_Carriageway | None, Other permanent object, Road sign or traffic signal, Lamp post, Near/Offside crash barrier, Tree Bus stop or bus shelter, Central crash barrier, Entered ditch, Telegraph or electricity pole, Submerged in water, Wall or fence, Data missing or out of range |
| 9 | Journey\_Purpose\_of\_Driver | Journey as part of work, Other/Not known (2005-10), Commuting to/from work, Pupil riding to/from school, Taking pupil to/from school, Not known, Other, Data missing or out of range |
| 10 | Junction\_Location | Leaving roundabout, Not at or within, 20 metres of junction, Approaching junction or waiting/parked at junction approach, Mid, Junction - on roundabout or on main road, Cleared junction or waiting/parked at junction exit, Entering main road, Entering from slip road, Leaving main road, Entering roundabout, Data missing or out of range |
| 11 | make | Текстові значення |
| 12 | model | Текстові значення |
| 13 | Propulsion\_Code | Petrol, Heavy oil, Gas/Bi-fuel, Electric Petrol/Gas (LPG), Gas, Hybrid electric, New fuel technology, Gas Diesel, Fuel cells, Electric diesel, Steam, NA |
| 14 | Sex\_of\_Driver | Male, Female, Not known, Data missing or out of range |
| 15 | Skidding\_and\_Overturning | None, Skidded, Skidded and overturned , Overturned, Jackknifed, Jackknifed and overturned, Data missing or out of range |
| 16 | Towing\_and\_Articulation | No tow/articulation, Articulated vehicle , Single trailer, Other tow, Double or multiple trailer, Caravan, Data missing or out of range |
| 17 | Vehicle\_Leaving\_Carriageway | Did not leave carriageway, Offside Nearside, Offside on to central reservation, Nearside and rebounded, Offside on to centrl res + rebounded, Straight ahead at junction, Offside - crossed central reservation, Offside and rebounded, Data missing or out of range |
| 18 | Vehicle\_Location.Restricted\_Lane | 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, NA |
| 19 | Vehicle\_Manoeuvre | Going ahead other, Turning right, U-turn, Slowing or stopping, Going ahead right-hand bend, Moving off, Waiting to go - held up, Parked, Going ahead left-hand bend , Overtaking moving vehicle - offside, Changing lane to right, Waiting to turn right, Turning left, Overtaking – nearside, Overtaking static vehicle – offside, Reversing, Changing lane to left, Waiting to turn left, Data missing or out of range |
| 20 | Vehicle\_Reference | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 63, 64, 66, 91 |
| 21 | Vehicle\_Type | 109 Motorcycle, 125cc and under Van / Goods, 3.5 tonnes mgw or under Bus or coach (17 or more pass seats) Goods, 7.5 tonnes mgw and over, 108 Motorcycle, 50cc and under, 106 Other vehicle, Goods over 3.5t. and under 7.5t Minibus (8 - 16 passenger seats), Agricultural vehicle, Pedal cycle, Tram, Car, Motorcycle over 500cc, Motorcycle over 125cc and up to 500cc, Taxi/Private hire car, Ridden horse, Motorcycle - unknown cc, Mobility scooter, Goods vehicle - unknown weight, Electric motorcycle, Data missing or out of range |
| 22 | Was\_Vehicle\_Left\_Hand\_Drive | Yes, No, Data missing or out of range |
| 23 | X1st\_Point\_of\_Impact | Front, Offside, Did not impact, Back, Nearside, Data missing or out of range |
| 24 | Year | 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016 |



Набір даних має 24 поля та 2177205 спостережень

Категорійні змінні: Age\_Band\_of\_Driver, Driver\_Home\_Area\_Type, Driver\_IMD\_Decile, Engine\_Capacity\_.CC, Hit\_Object\_in\_Carriageway, Hit\_Object\_off\_Carriageway, Journey\_Purpose\_of\_Driver, Junction\_Location, make, model, Propulsion\_Code, Sex\_of\_Driver, Skidding\_and\_Overturning, Towing\_and\_Articulation, Vehicle\_Leaving\_Carriageway, Vehicle\_Manoeuvre, Vehicle\_Type, Was\_Vehicle\_Left\_Hand\_Drive, X1st\_Point\_of\_Impact,Year

Ми будемо розглядати ДТП, у яких примйають участь 2 транспортних засоби у проміжку від 2006 до 2015 року

У датасеті df\_Accident\_Information відбираються лише ті рядки, де тільки 2 транспортні засоби беруть участь у ДТП. Оновлений датасет перейменовується на df\_Accident\_Information\_2:

df\_Accident\_Information\_2 <- filter(df\_Accident\_Information, Number\_of\_Vehicles == 2 )

У датасеті df\_Accident\_Information\_2 відбираються лише ті рядки, де рік у проміжку від 2006 до 2015 Оновлений датасет перейменовується на df\_Accident\_Information\_3:

df\_Accident\_Information\_3 <- filter(df\_Accident\_Information\_2, Year %in% c(2006,2007,2008,2009,2010,2011, 2012, 2013, 2014, 2015))

Далі з обох датасетів (df\_Accident\_Information\_3, df\_Vehicle\_Information) відбираються ті поля, які нам цікаві для проведення EDA. Датасети перейменовуються на c\_Accident і c\_Vehicle відповідно:

df\_Accident\_Information\_3: Accident\_Index, Accident\_Severity, Carriageway\_Hazards, Date, Day\_of\_Week, Junction\_Detail, Light\_Conditions, Number\_of\_Casualties, Road\_Surface\_Conditions, Road\_Type, Speed\_limit, Urban\_or\_Rural\_Area, Weather\_Conditions, Year

c\_Accident <- select(df\_Accident\_Information\_3,

Accident\_Index, Accident\_Severity, Carriageway\_Hazards,

Date, Day\_of\_Week,

Junction\_Detail,

Light\_Conditions, Number\_of\_Casualties,

Road\_Surface\_Conditions, Road\_Type, Speed\_limit,

Urban\_or\_Rural\_Area, Weather\_Conditions, Year, Time

)

df\_Vehicle\_Information: Accident\_Index, Age\_Band\_of\_Driver, Age\_of\_Vehicle, Sex\_of\_Driver, make,

Vehicle\_Manoeuvre, Vehicle\_Type, X1st\_Point\_of\_Impact

c\_Vehicle <- select(df\_Vehicle\_Information,

Accident\_Index,

Age\_Band\_of\_Driver,

Age\_of\_Vehicle,

Sex\_of\_Driver,

make,

Vehicle\_Manoeuvre,

Vehicle\_Type,

X1st\_Point\_of\_Impact

)

Далі для приведення датасетів до належного вигляду перед їх об’єднанням з датаесту c\_Vehicle видаляємо індекси яких немає в c\_Accident і перейменовуємо в c\_Vehicle\_Index, а потім з датасету c\_Accident видаляємо індекси яких немає в датасеті c\_Vehicle\_Index:

c\_Vehicle\_Index <- c\_Vehicle %>%

filter(Accident\_Index %in% c\_Accident$Accident\_Index)

c\_Accident\_Index <- c\_Accident %>%

filter(Accident\_Index %in% c\_Vehicle\_Index$Accident\_Index)

Робимо відбір рядків з датасету c\_Vehicle\_Index, де кількість водіїв до одного ДТП, яке зафіксовано в полі Accident\_Index дорівнювало двом. Оновлений датасет перейменовуємо в c\_Vehicle\_Index\_filtered:

duplicated\_accidents <- c\_Vehicle\_Index %>%

group\_by(Accident\_Index) %>%

filter(n() == 2)

c\_Vehicle\_Index\_filtered <- c\_Vehicle\_Index %>%

filter(Accident\_Index %in% duplicated\_accidents$Accident\_Index)

З датасету c\_Accident\_Index знову видаляємо індекси яких немає в c\_Vehicle\_Index\_filtered і відбираємо парні та непарні спостереження (Учасник ДТП 1 і Учасник ДТП 2). Оновлені датасети називаємо df\_1 і df\_2:

c\_Accident\_Index\_filtered <- c\_Accident\_Index %>%

filter(Accident\_Index %in% c\_Vehicle\_Index\_filtered$Accident\_Index)

df1 <- c\_Vehicle\_Index\_filtered %>%

filter(row\_number() %% 2 == 1)

df2 <- c\_Vehicle\_Index\_filtered %>%

filter(row\_number() %% 2 == 0)

Після цього, перейменовуємо деякі поля датасетів:

s\_df1 <- df1 %>%

rename(

Age\_Band\_of\_Driver1=Age\_Band\_of\_Driver,

Age\_of\_Vehicle1=Age\_of\_Vehicle,

Sex\_of\_Driver1=Sex\_of\_Driver,

Car\_Brand1=make,

Vehicle\_Manoeuvre1=Vehicle\_Manoeuvre,

Vehicle\_Type1=Vehicle\_Type,

Point\_of\_Impact1=X1st\_Point\_of\_Impact

)

s\_df2 <- df2 %>%

rename(

Age\_Band\_of\_Driver2=Age\_Band\_of\_Driver,

Age\_of\_Vehicle2=Age\_of\_Vehicle,

Sex\_of\_Driver2=Sex\_of\_Driver,

Car\_Brand2=make,

Vehicle\_Manoeuvre2=Vehicle\_Manoeuvre,

Vehicle\_Type2=Vehicle\_Type,

Point\_of\_Impact2=X1st\_Point\_of\_Impact

)

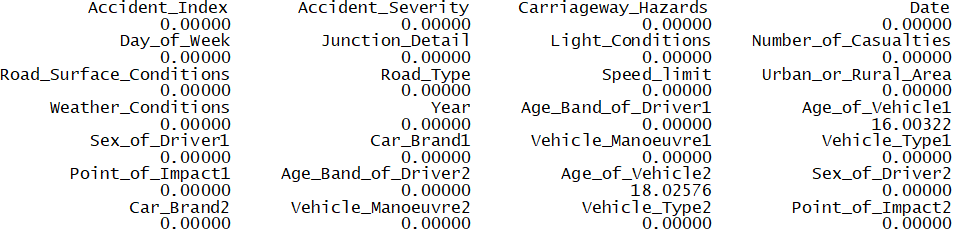
Ну і нарешті, об’єднуємо ці датасети, спочатку об’єднаємо датасети s\_df1 і s\_df2 за полем Accident\_Index:

merged\_Vehicle <- merge(s\_df1, s\_df2, by = "Accident\_Index", all = TRUE)

Далі об’єднаємо датасети c\_Accident\_Index\_filtered і merged\_Vehicle:

all\_data1 <- merge(c\_Accident\_Index\_filtered , merged\_Vehicle, by = "Accident\_Index", all = TRUE)

Перевіряємо відсоток NA даних в кожному стовпці:



Як бачимо, відсоток NA даних не перевищує 20%, тому наступним кроком буде видалення всіх рядків, де хоча б одне поле має NA значення

all\_data2 <- na.omit(all\_data1)

selected\_columns <- c(

"Accident\_Index" ,

"Accident\_Severity","Carriageway\_Hazards"

,"Date"

,"Day\_of\_Week"

,"Junction\_Detail"

,"Light\_Conditions"

, "Number\_of\_Casualties"

,"Road\_Surface\_Conditions"

, "Road\_Type"

, "Speed\_limit"

, "Urban\_or\_Rural\_Area"

, "Weather\_Conditions"

, "Year"

, "Time"

, "Age\_Band\_of\_Driver1"

, "Age\_of\_Vehicle1"

, "Sex\_of\_Driver1"

, "Car\_Brand1"

,"Vehicle\_Manoeuvre1"

, "Vehicle\_Type1"

, "Point\_of\_Impact1"

, "Age\_Band\_of\_Driver2"

, "Age\_of\_Vehicle2"

, "Sex\_of\_Driver2"

, "Car\_Brand2"

, "Vehicle\_Manoeuvre2"

, "Vehicle\_Type2"

, "Point\_of\_Impact2"

)

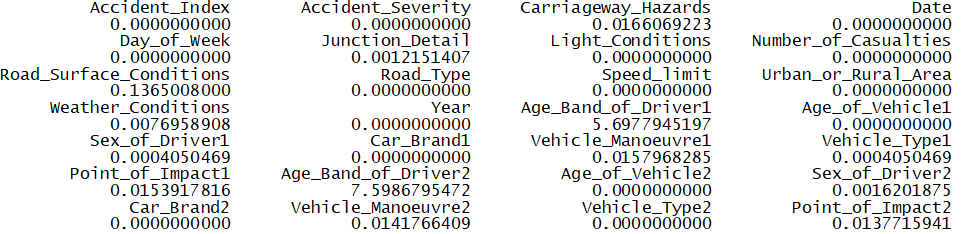
all\_data3 <- all\_data2[, selected\_columns]

Як бачимо, відсоток NA даних не перевищує 20%, тому наступним кроком буде видалення всіх рядків, де хоча б одне поле має NA значення

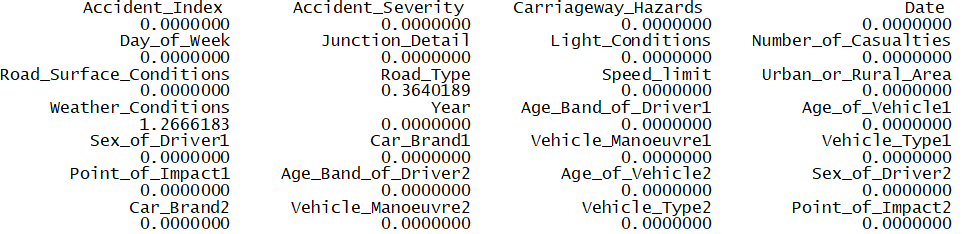
У датасеті також є специфічно закодовані пропущені дані: Data missing or out of range, Unknown, Not known, Unallocated

Перевіряємо відсоток цих даних:

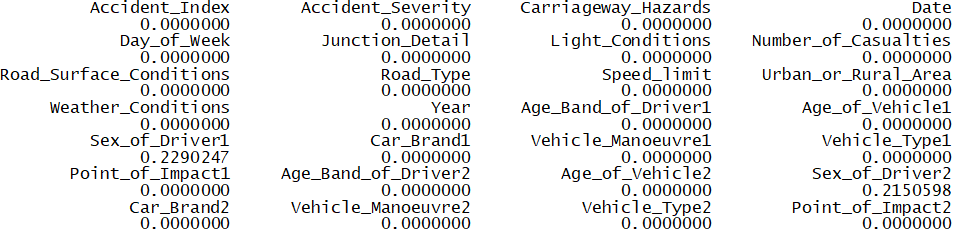
Data missing or out of range:



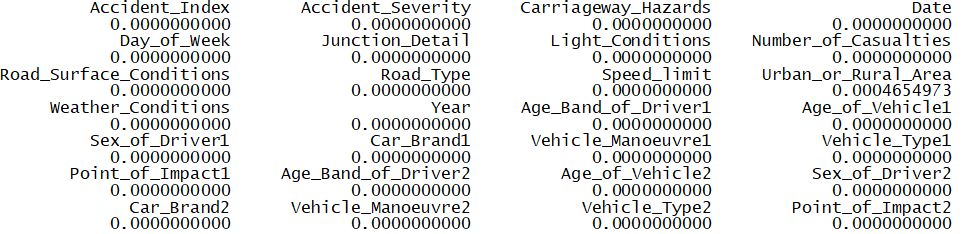
Unknown:



Not known:



Unallocated:



Оскільки не має перевищення 20%, то також видаляємо рядки, де хоча б одне поле містить одне з вище наведених значень:

filtered\_data <- all\_data3 %>%

filter\_all(all\_vars(. != "Data missing or out of range"))

filtered\_data\_1 <- filtered\_data[, selected\_columns]

filtered\_data\_2 <- filtered\_data\_1 %>%

filter\_all(all\_vars(!(. %in% c("Unknown", "Not known", "Unallocated"))))

Також перейменовується ще один рядок датасету:

filtered\_data\_3 <- filtered\_data\_2 %>%

rename(

Area\_Type=Urban\_or\_Rural\_Area,

)

І останній крок: ми модифікуємо поле Date, перевівши його в тип Date, і поділивши його на дві колонки для дня та місця, для того, щоб побачити статистику ДТП в різні місяці та дні, після цього ми видаляємо оригінальне поле Date:

library(lubridate)

filtered\_data\_3$Date <- as.Date(filtered\_data\_3$Date)

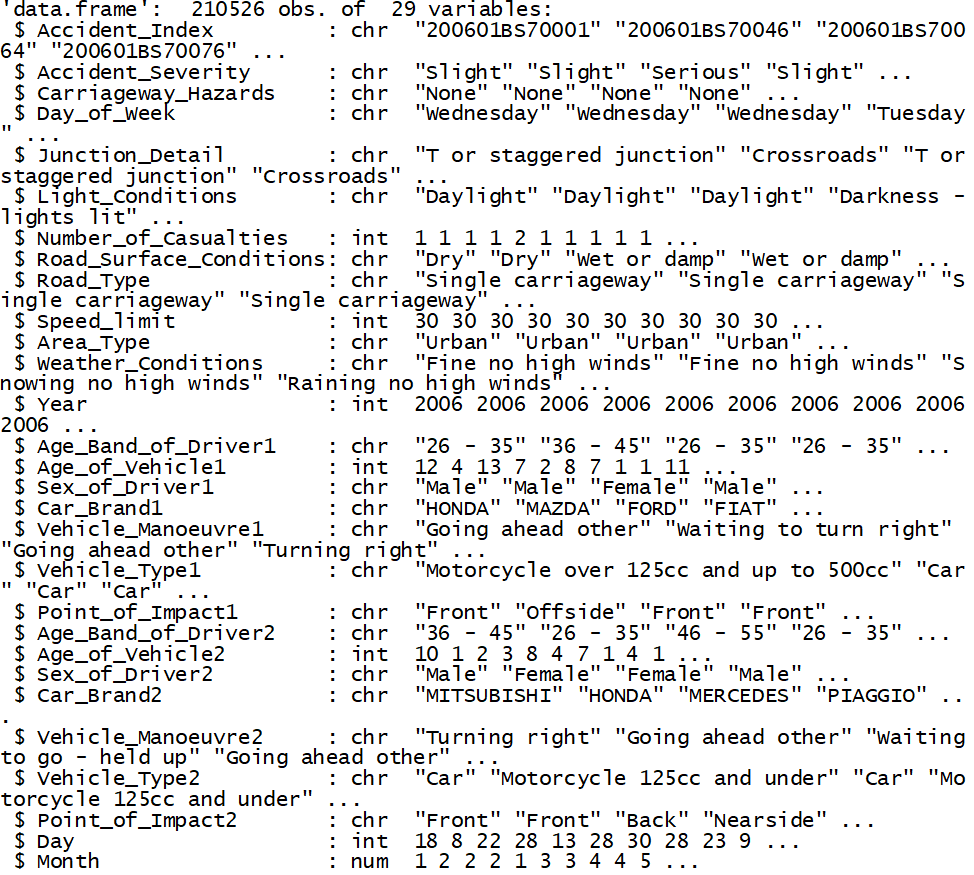
filtered\_data\_3$Day <- day(filtered\_data\_3$Date)

filtered\_data\_3$Month <- month(filtered\_data\_3$Date)

filtered\_data\_3 <- subset(filtered\_data\_3, select = -Date)

Отже, структура основного датасету filtered\_data\_3:

| № | Назва поля | Значення |
| --- | --- | --- |
| 1 | Accident\_Index | Неперервні значення |
| 2 | Accident\_Severity | Slight, Serious, Fatal |
| 3 | Carriageway\_Hazards | None, Other object on road, Pedestrian in carriageway - not injured, Vehicle load on road, Previous accident, Any animal in carriageway (except ridden horse) |
| 4 | Day\_of\_Week | Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday |
| 5 | Junction\_Detail | Auto traffic signal, Give way or uncontrolled, Stop sign, Authorised person, Not at junction or within 20 metres |
| 6 | Light\_Conditions | Daylight, Darkness - lights lit, Darkness - lighting unknown, Darkness - lights unlit, Darkness - no lighting |
| 7 | Number\_of\_Casualties | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 32, 33, 34, 35, 36, 38, 40, 41, 42, 43, 45, 46, 47, 48, 51, 54, 58, 62, 63, 68, 70, 87, 93 |
| 8 | Road\_Surface\_Conditions | Wet or damp, Dry, Frost or ice, Snow, Flood over 3cm. deep |
| 9 | Road\_Type | Single carriageway, Dual carriageway, One way street, Roundabout, Slip road |
| 10 | Speed\_limit | 0, 10, 15, 20, 30, 40, 50, 60, 70 |
| 11 | Area\_Type | Urban, Rural |
| 12 | Weather\_Conditions | Raining no high winds, Fine no high winds, Snowing no high winds, Other, Fine + high winds, Raining + high winds, Fog or mist, Snowing + high winds |
| 13 | Year | 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015 |
| 14 | Age\_Band\_of\_Driver1 | 0 - 5, 6 - 10, 11 - 15, 16 - 20, 21 - 25, 26 - 35, 36 - 45, 46 - 55, 56 - 65, 66 - 75, Over 75 |
| 15 | Age\_of\_Vehicle1 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 83, 84, 85, 90, 91, 95, 104, 105, 111 |
| 16 | Sex\_of\_Driver1 | Male, Female |
| 17 | Car\_Brand1 | Текстові значення |
| 18 | Vehicle\_Manoeuvre1 | Going ahead other, Turning right, U-turn, Slowing or stopping, Going ahead right-hand bend, Moving off, Waiting to go - held up, Parked, Going ahead left-hand bend , Overtaking moving vehicle - offside, Changing lane to right, Waiting to turn right, Turning left, Overtaking – nearside, Overtaking static vehicle – offside, Reversing, Changing lane to left, Waiting to turn left |
| 19 | Vehicle\_Type1 | 109 Motorcycle, 125cc and under Van / Goods, 3.5 tonnes mgw or under Bus or coach (17 or more pass seats) Goods, 7.5 tonnes mgw and over, 108 Motorcycle, 50cc and under, 106 Other vehicle, Goods over 3.5t. and under 7.5t Minibus (8 - 16 passenger seats), Agricultural vehicle, Pedal cycle, Tram, Car, Motorcycle over 500cc, Motorcycle over 125cc and up to 500cc, Taxi/Private hire car, Ridden horse, Motorcycle - unknown cc, Mobility scooter, Goods vehicle - unknown weight, Electric motorcycle |
| 20 | Point\_of\_Impact1 | Front, Offside, Did not impact, Back, Nearside |
| 21 | Age\_Band\_of\_Driver2 | 0 - 5, 6 - 10, 11 - 15, 16 - 20, 21 - 25, 26 - 35, 36 - 45, 46 - 55, 56 - 65, 66 - 75, Over 75 |
| 22 | Age\_of\_Vehicle2 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 83, 84, 85, 90, 91, 95, 104, 105, 111 |
| 23 | Sex\_of\_Driver2 | Male, Female |
| 24 | Car\_Brand2 | Текстові значення |
| 25 | Vehicle\_Manoeuvre2 | Going ahead other, Turning right, U-turn, Slowing or stopping, Going ahead right-hand bend, Moving off, Waiting to go - held up, Parked, Going ahead left-hand bend , Overtaking moving vehicle - offside, Changing lane to right, Waiting to turn right, Turning left, Overtaking – nearside, Overtaking static vehicle – offside, Reversing, Changing lane to left, Waiting to turn left |
| 26 | Vehicle\_Type2 | 109 Motorcycle, 125cc and under Van / Goods, 3.5 tonnes mgw or under Bus or coach (17 or more pass seats) Goods, 7.5 tonnes mgw and over, 108 Motorcycle, 50cc and under, 106 Other vehicle, Goods over 3.5t. and under 7.5t Minibus (8 - 16 passenger seats), Agricultural vehicle, Pedal cycle, Tram, Car, Motorcycle over 500cc, Motorcycle over 125cc and up to 500cc, Taxi/Private hire car, Ridden horse, Motorcycle - unknown cc, Mobility scooter, Goods vehicle - unknown weight, Electric motorcycle |
| 27 | Point\_of\_Impact2 | Front, Offside, Did not impact, Back, Nearside |
| 28 | Day | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31 |
| 29 | Month | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 |
| 30 | Time | HH:MM |

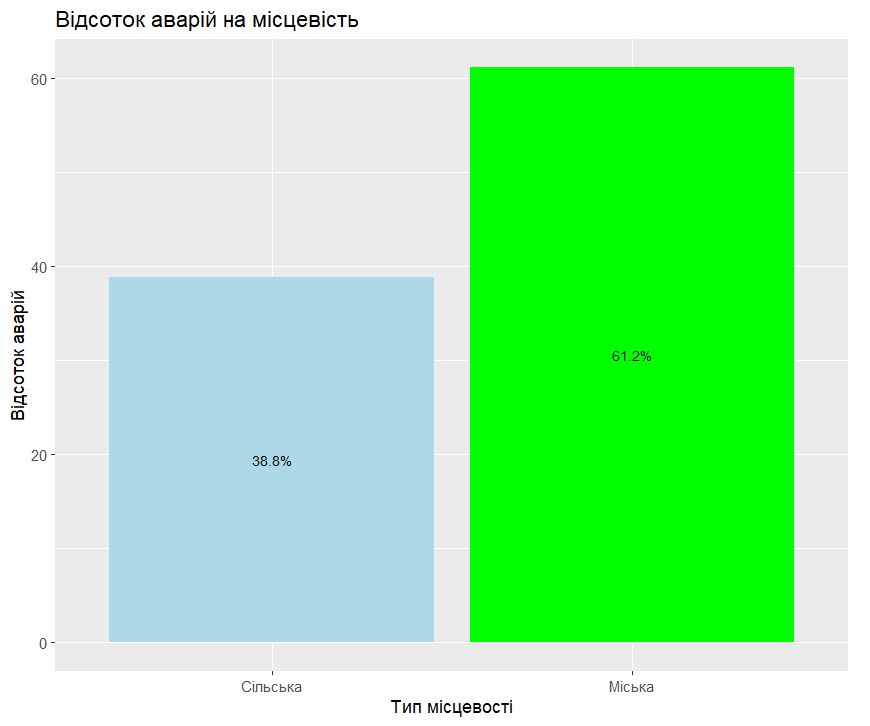


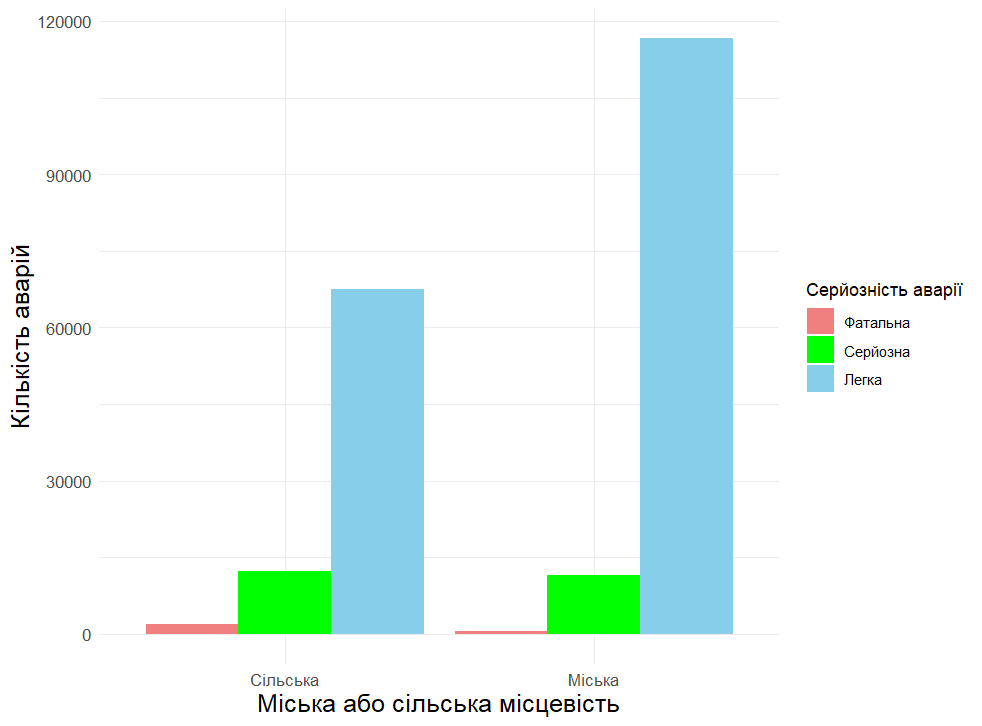
Набір даних має 30 полів та 210526 спостережень

Категорійні змінні: Accident\_Severity, Carriageway\_Hazards, Day\_of\_Week, Junction\_Detail, Light\_Conditions, Road\_Surface\_Conditions, Road\_Type, Area\_Type, Weather\_Conditions, Age\_Band\_of\_Driver1, Sex\_of\_Driver1, Car\_Brand1, Vehicle\_Manoeuvre1, Vehicle\_Type1, Point\_of\_Impact1, Age\_Band\_of\_Driver2, Sex\_of\_Driver2, Car\_Brand2, Vehicle\_Manoeuvre2, Vehicle\_Type2, Point\_of\_Impact2

У полі Carriageway\_Hazards залишається змінна None, так як вона буде мати вплив під час EDA. В даному випадку вона означає, що небезпеки на проїжджій частині під час ДТП не було.

# 1. Як особливості місцевості та дорожнього покриття впливають на ймовірність виникнення аварій та тяжкість?

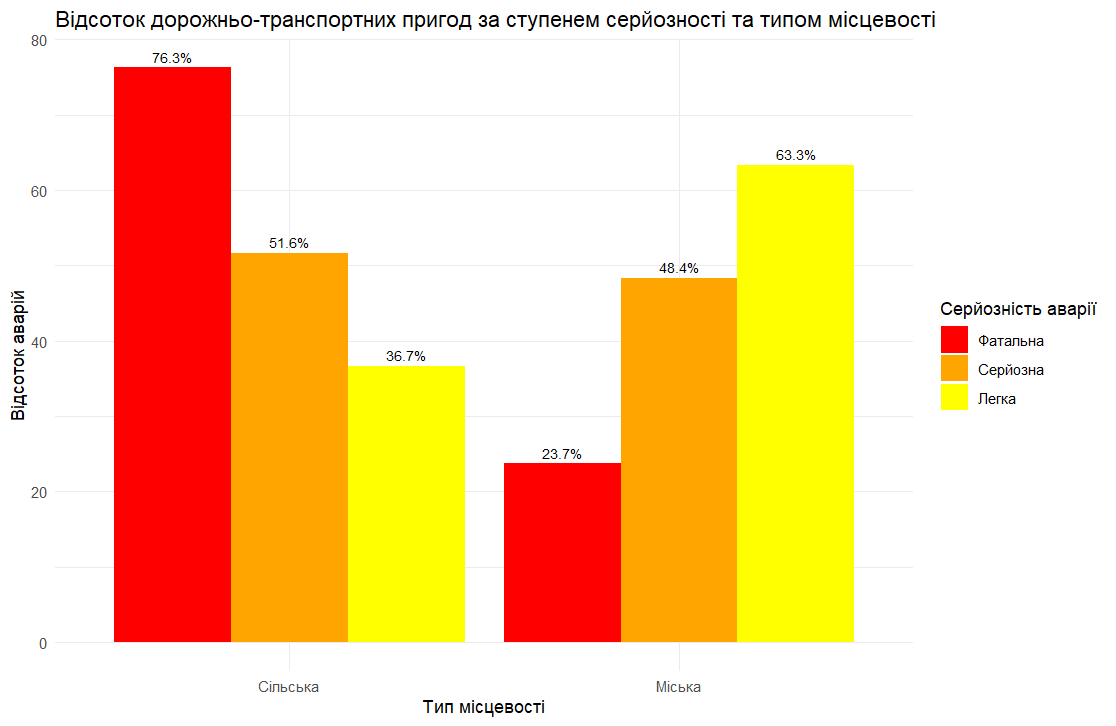
Кількість аварій і їх серйозність у міській і сільській місцевості:



Як бачимо, в сільській місцевості аварії трапляються не так часто, проте частка фатальних ДТП набагато вища, ніж у міській місцевості.

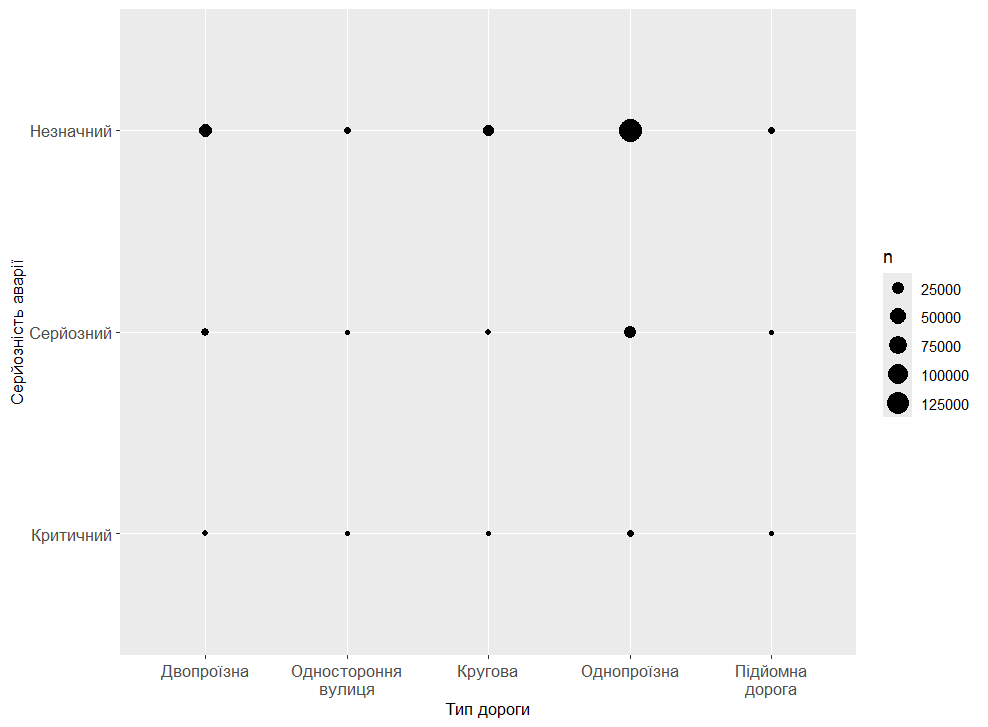
Це може бути пов’язано з багатьма факторами, але найбільш імовірним з яких є перевищення допустимої швидкості під час їзди в сільській місцевості та гіршим станом доріг. За наявності додаткових даних, наприклад тієї ж швидкості автомобіля або довжині гальмівного шляху можна було б встановити істинність або хибність даної гіпотези.

Відсоток аварій за ступенем серйозності та типом місцевості:

****

Як бачимо, в сільській місцевості аварії трапляються не так часто, проте частка фатальних ДТП набагато вища, ніж у міській місцевості.

Залежність типу дороги від серйозності аварії:

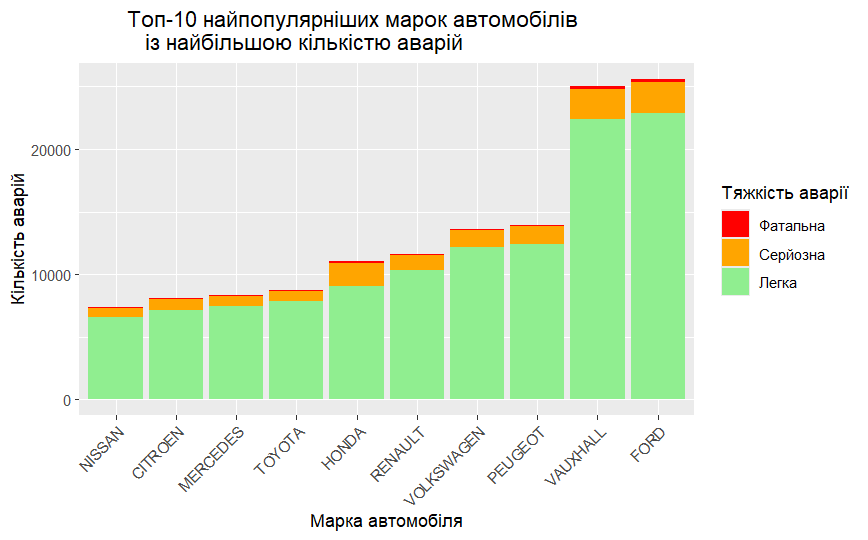


Найбільша кількість незначних аварій відбувається на одно проїзному та двопроїзному типі дороги.

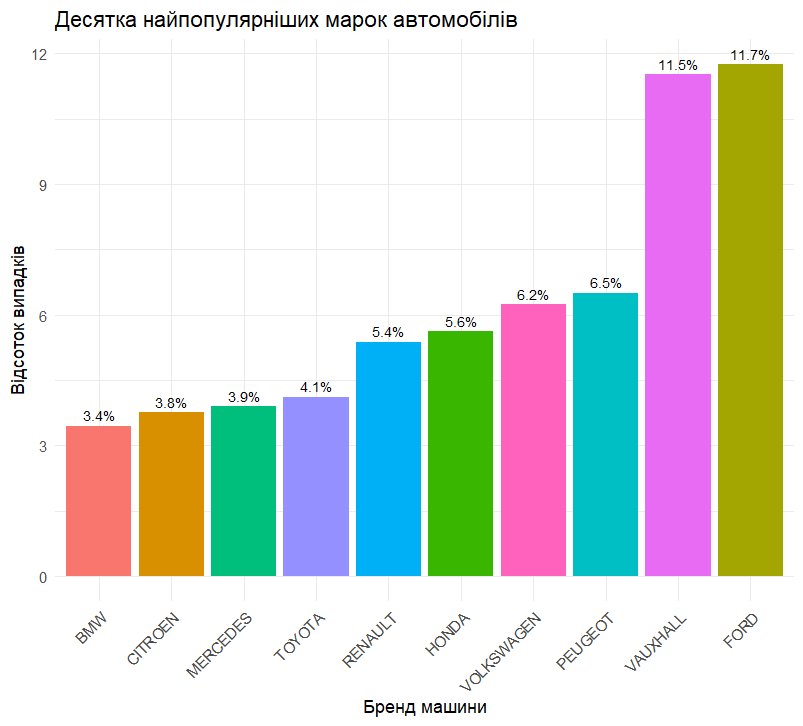
Також досить серйозні та критичні аварії за ступенем тяжкості трапляються частіше.

# 2. Як впливають характеристики автомобіля на кількість/серйозність аварій?

Топ-10 марок автомобілів із найбільшою кількістю аварій:



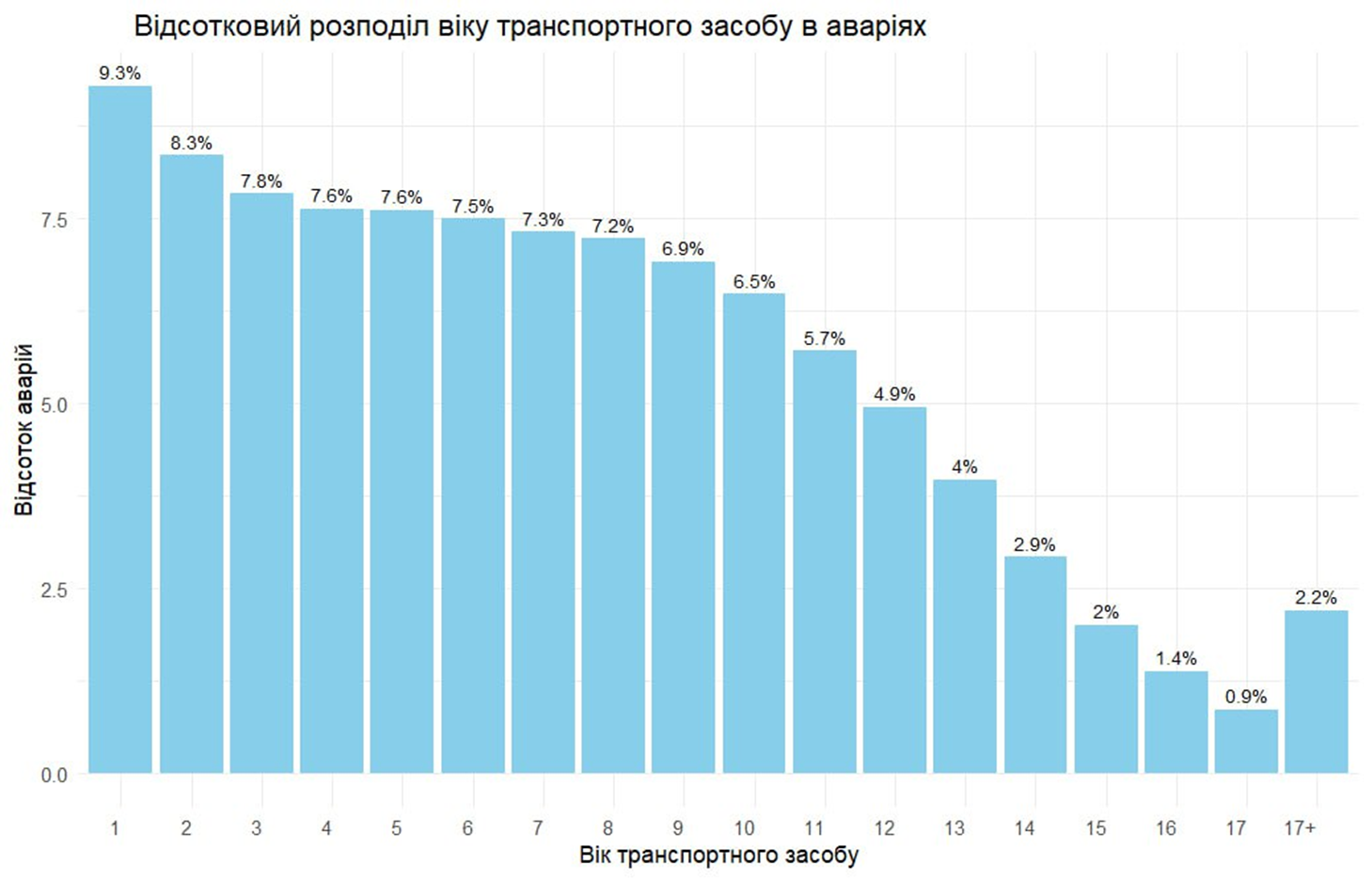
Топ-10 найпопулярніших марок автомобілів:



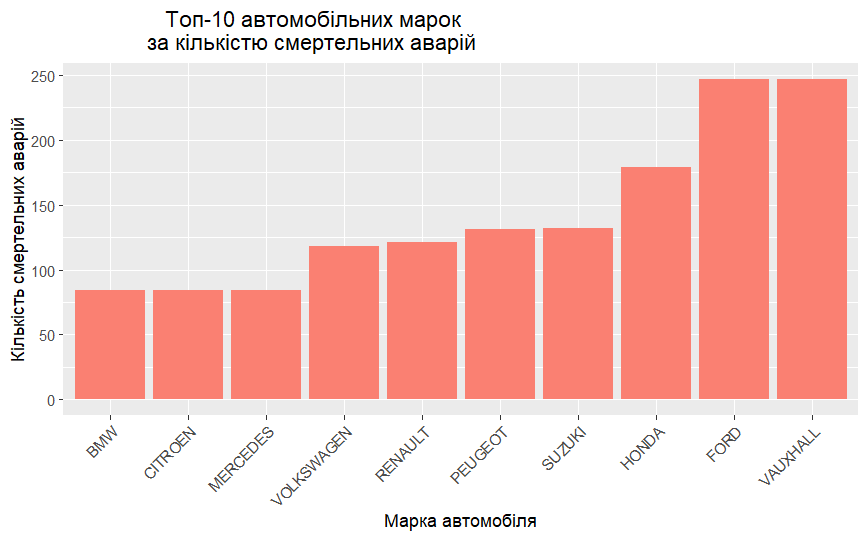
Можна побачити, що різниця в кількості ДТП для різних марок авто обумовлена їхньою популярністю, і,отже, марка машини на ймовірність потрапляння в ДТП безпосередньо не впливає, або впливає дуже незначно.

Натомість, власники автомобілів Honda і Suzuki потрапляли в смертельні ДТП частіше, ніж власники всіх інших машин, якщо брати в урахування співвідношення популярності цих марок.

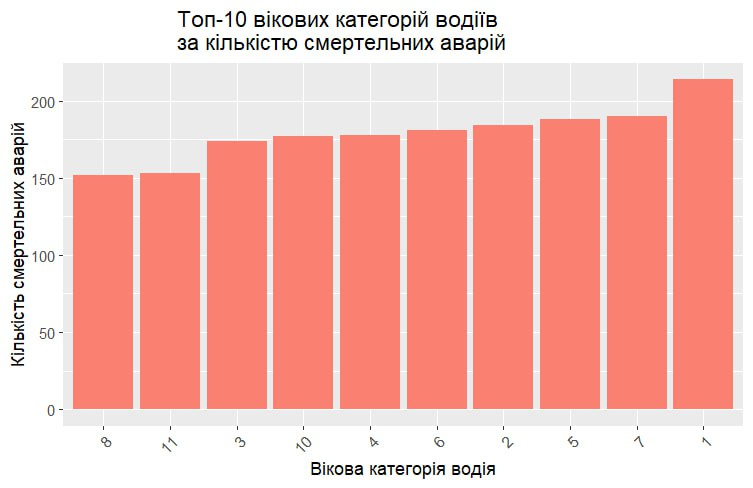
Відсотковий розподіл віку транспортного засобу в аваріях:



Топ-10 автомобільних марок за кількістю смертельних аварій:



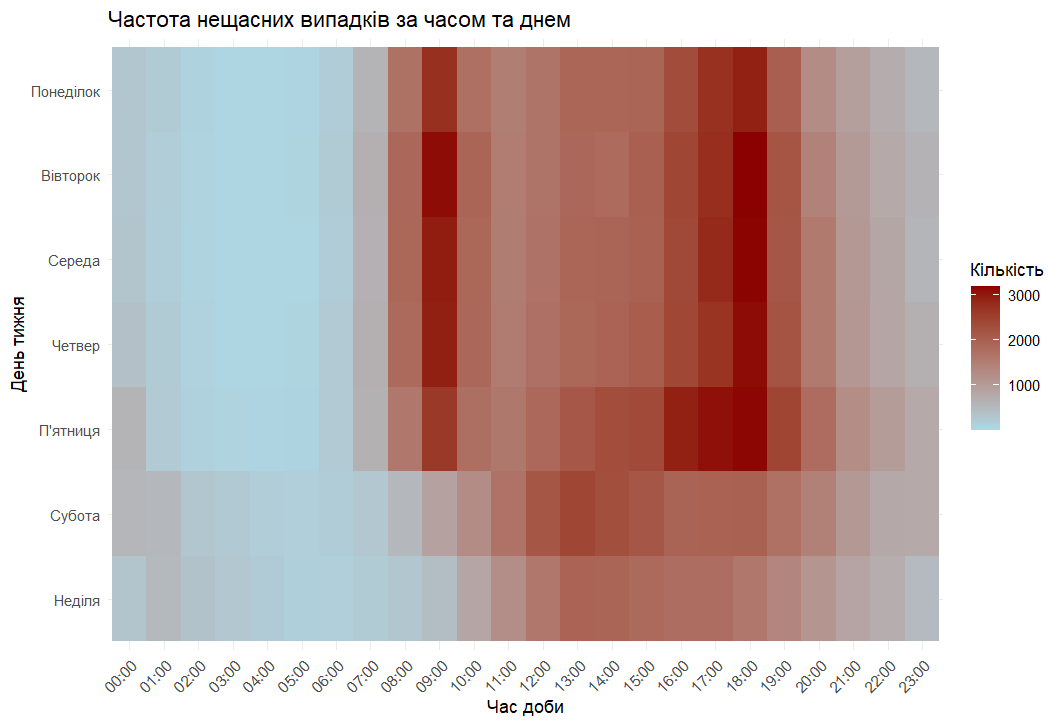
Топ-10 вікових категорій водіїв за кількістю смертельних аварій:



Також цікавим є те, що в смертельні ДТП найбільше потрапляли водії, машина яких була в них лише рік, і це пов’язано не лише з тим, що машин віком 1 рік найбільше, оскільки тоді машини віку 2, 3 і т.д. теж були б “лідерами”.

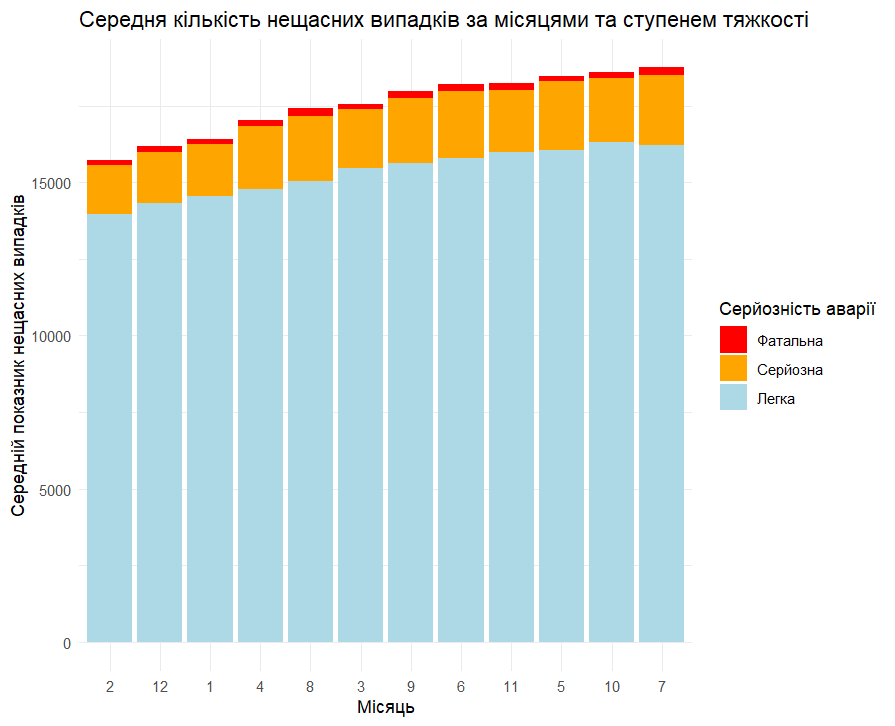
3. Як впливає час/день/рік на кількість/серйозність аварій?

Частота нещасних випадків (аварій) за часом та днем



Як можна побачити, найбільше ДТП трапляється в часи пік в будні дні з 9 до 10 ранку та з 16 по 19 увечорі. Натомість у вихідні дні, більше аварій трапляється після 10-11 ранку.

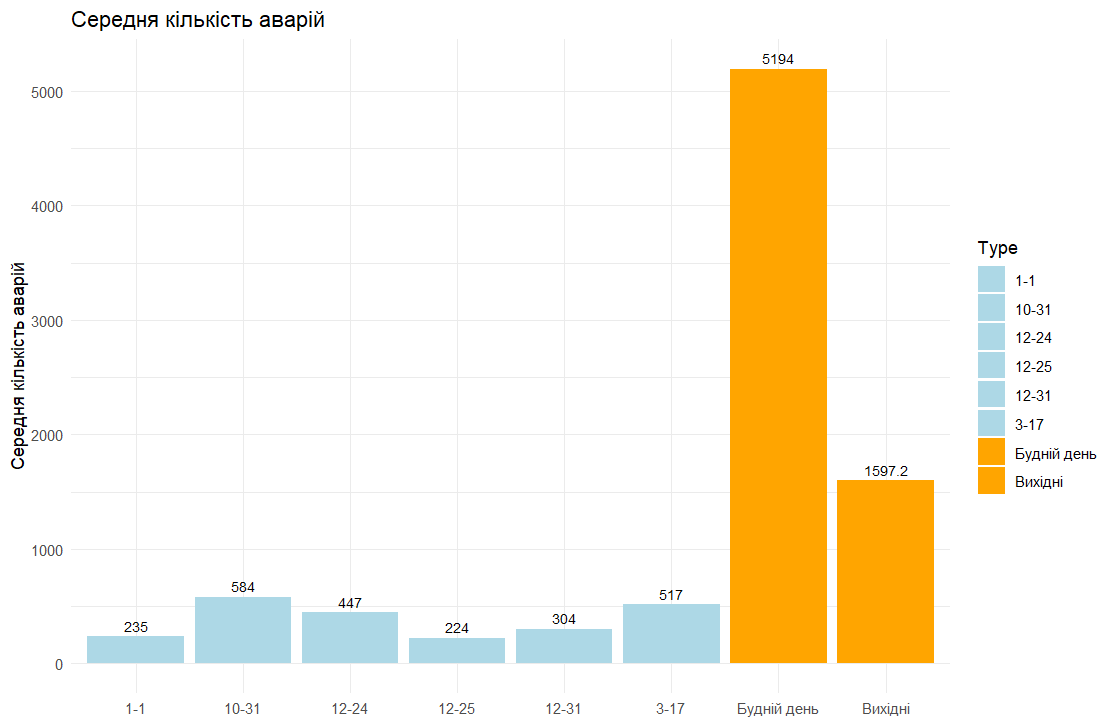
Середня кількість нещасних випадків за місяцями та ступенями тяжкості:



Найменше ДТП трапляється у зимові місяці, що є доволі неочікуваним результатом. Загалом частка серйозних і смертельних ДТП на кожен місяць пропорційна власне кількості ДТП за цей місяць, де винятками знову таки є зимові місяці.

Середня кількість аварій в різні дні:

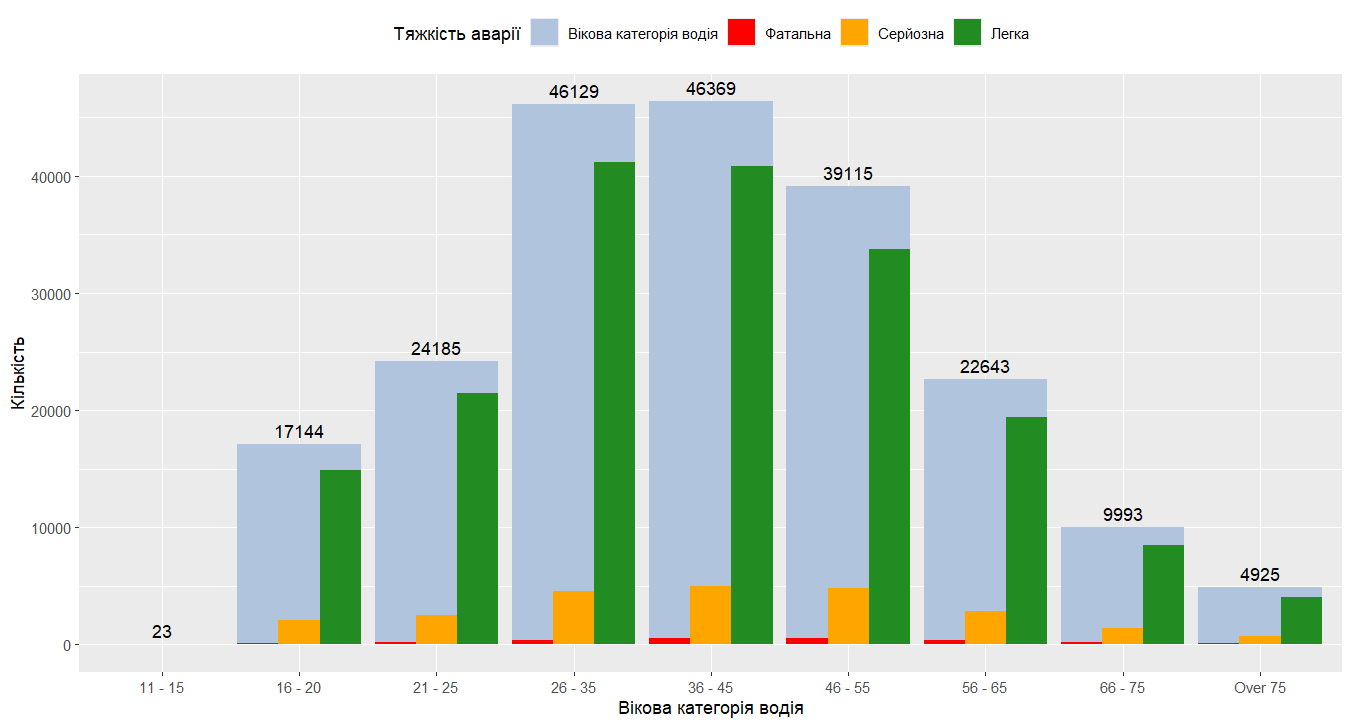
# 



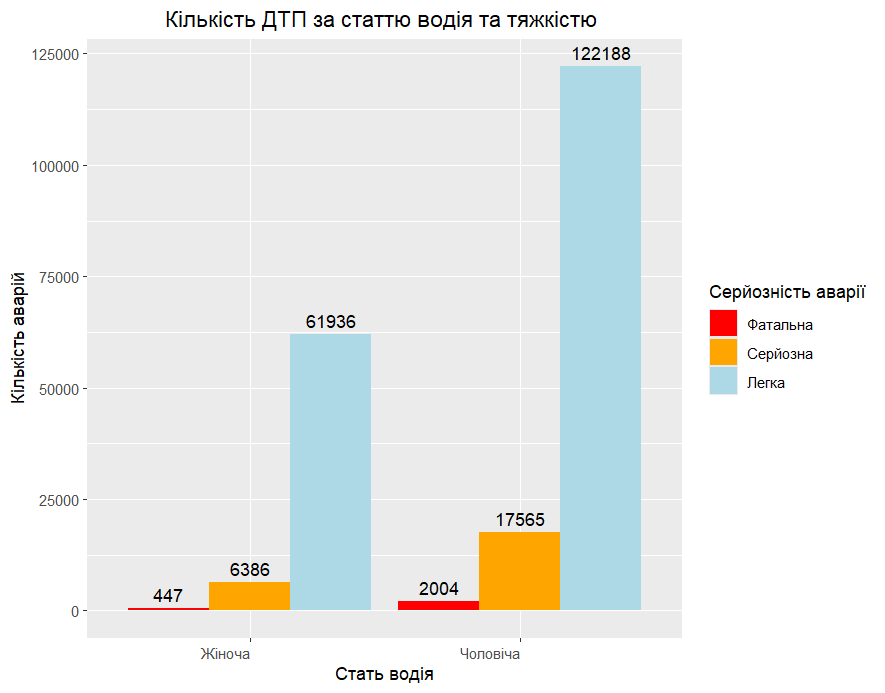
Як можна побачити, у вихідні дні середня кількість ДТП суттєво менша, ніж у будні дні. А у свята ця кількість іще менша, при цьому найменше ДТП трапляється у різдвяні свята, а найбільша - напередодні Нового року.

# 4. Залежність фізичних особливостей (віку та статі) водія на поведінку на дорозі?

Вплив вікової категорії водія на серйозність аварії:

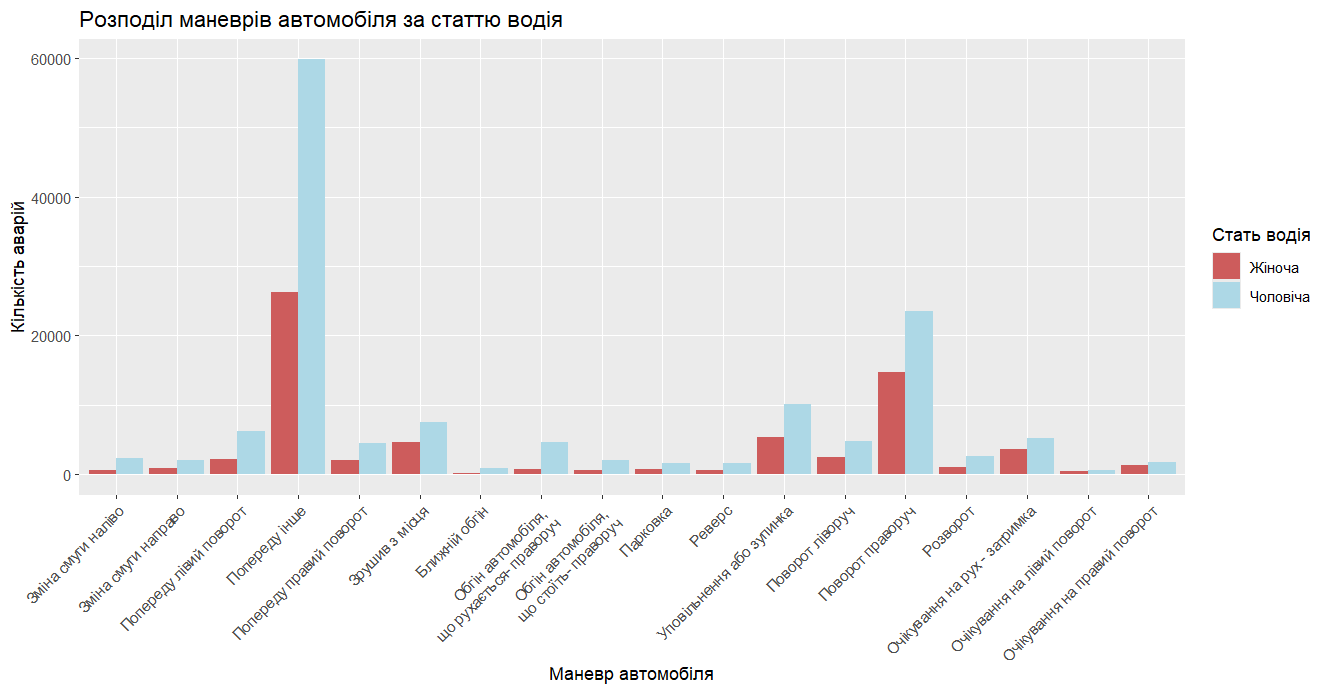


Вплив статті водія на тяжкість і кількість ДТП:



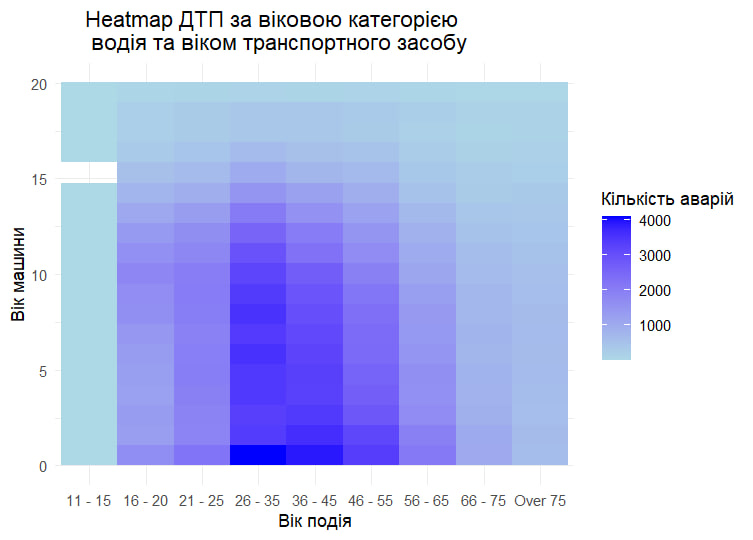
На жаль ми не маємо точної інформації про загальну кількість водіїв жінок та чоловіків, однак з офіційних ресурсів ми можемо дізнатись про те що кількість виданих ліцензій на водіння за весь час для чоловіків 25,956,385 а для жінок 23,438,355. Отже, чоловіки складають приблизно 52.49% від всіх виданих ліцензій на водіння, а жінки - приблизно 47.51%. Так що можна сказати що їх приблизно однакова кількість. Тим не менше, чоловіки частіше потрапляли в аварії, причому частка смертельних ДТП у них майже в 4.5 рази більша ніж у жінок, а частка серйозних ДТП майже в 3 рази більша.

Розподіл маневрів автомобіля за статтю водія:



Більшість ДТП за участю двох водіїв трапилася під час обгону або повороту направо. Причому якщо кількість ДТП під час більшості маневрів між чоловіками і жінками майже однакова (врахуємо, що чоловіки потрапляли в ДТП близько в два рази більше), то чоловіки частіше потрапляли в ДТП під час обгону (лівої чи правої сторони), а жінки - під час повороту.

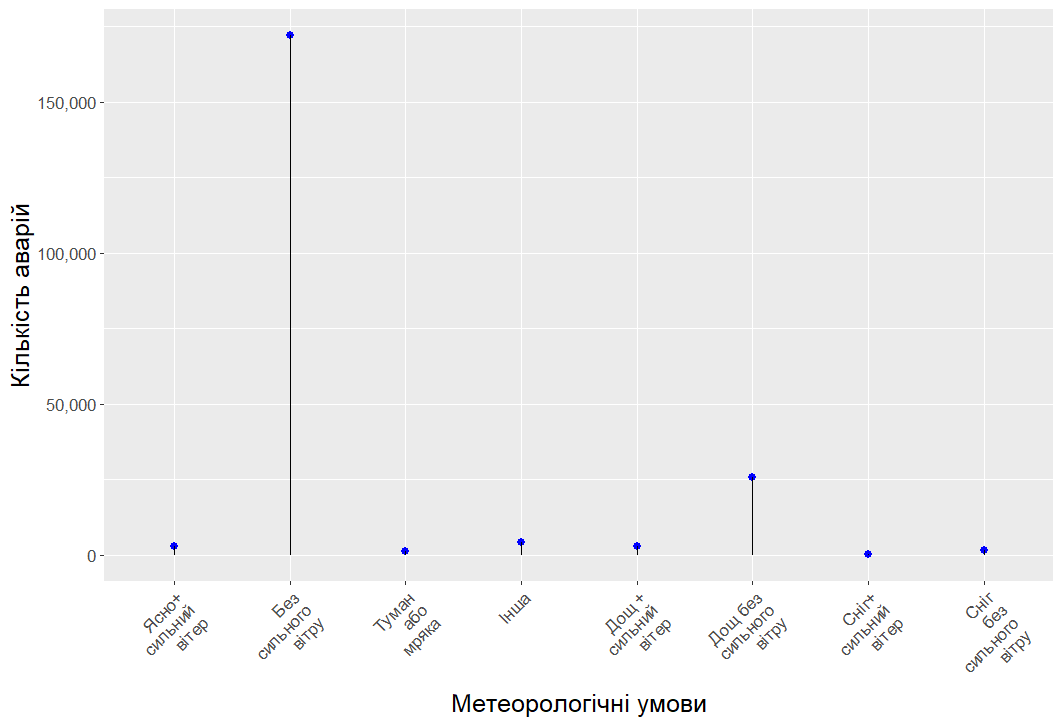
Heatmap ДТП за віковою категорією водія та віком транспортного засобу:



Можна побачити що водії віком від 26 до 55 частіше попадають в аварії ніж інші вікові групи. Це може бути з тим що вони набагато частіше користуються автомобілем. Також можна побачити що вік автомобіля майже не впливає на кількість аварій.

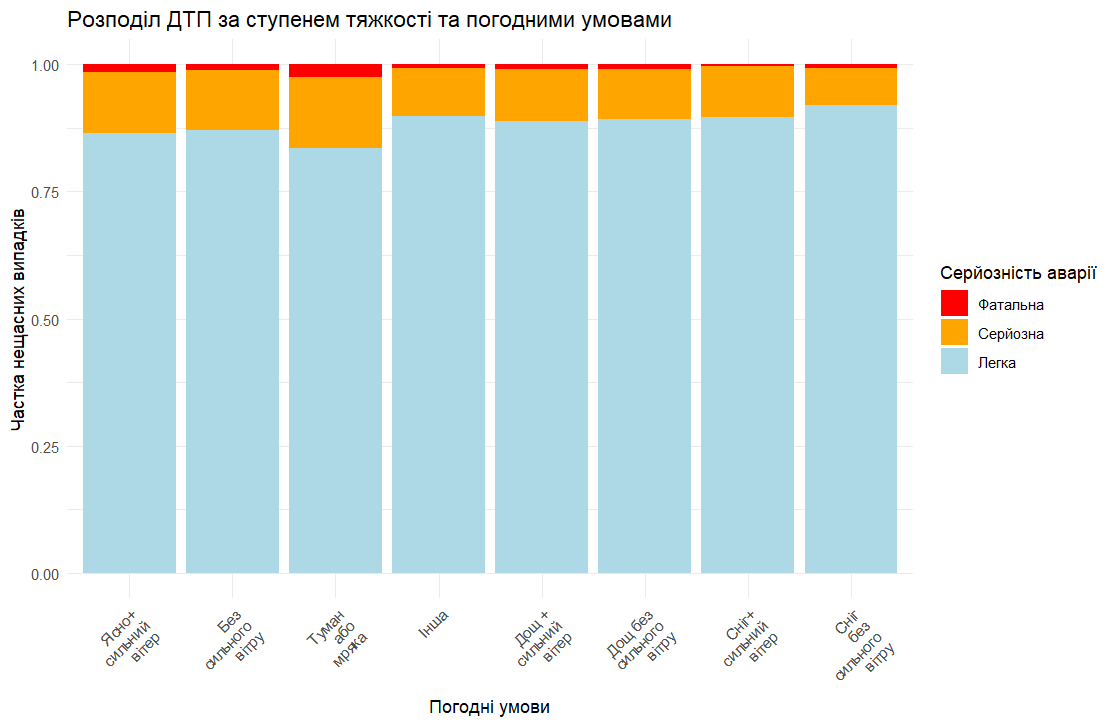
5. Як фактори на які не може впливати водій впливають на кількість /серйозність аварій?

Lolypop Погодні умови під час ДТП:



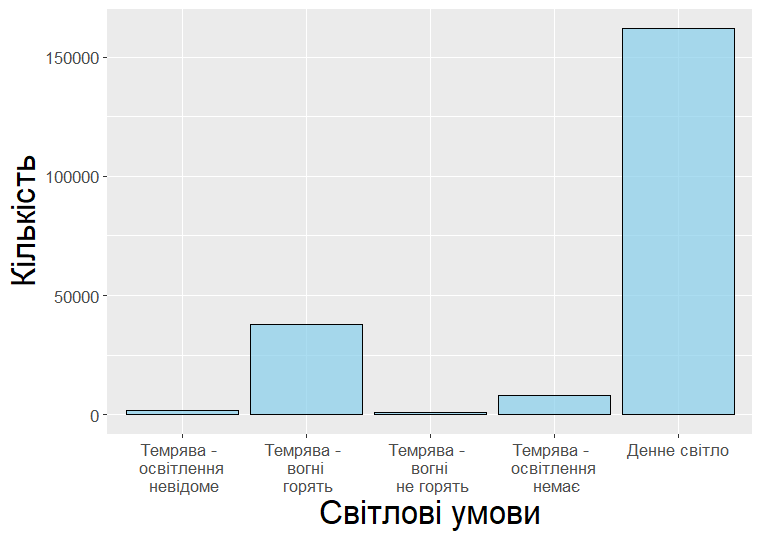
Найбільше аварій трапляється коли немає сильного вітру або дощ без сильного вітру.

Розподіл ДТП за ступенем тяжкості та погодними умовами:

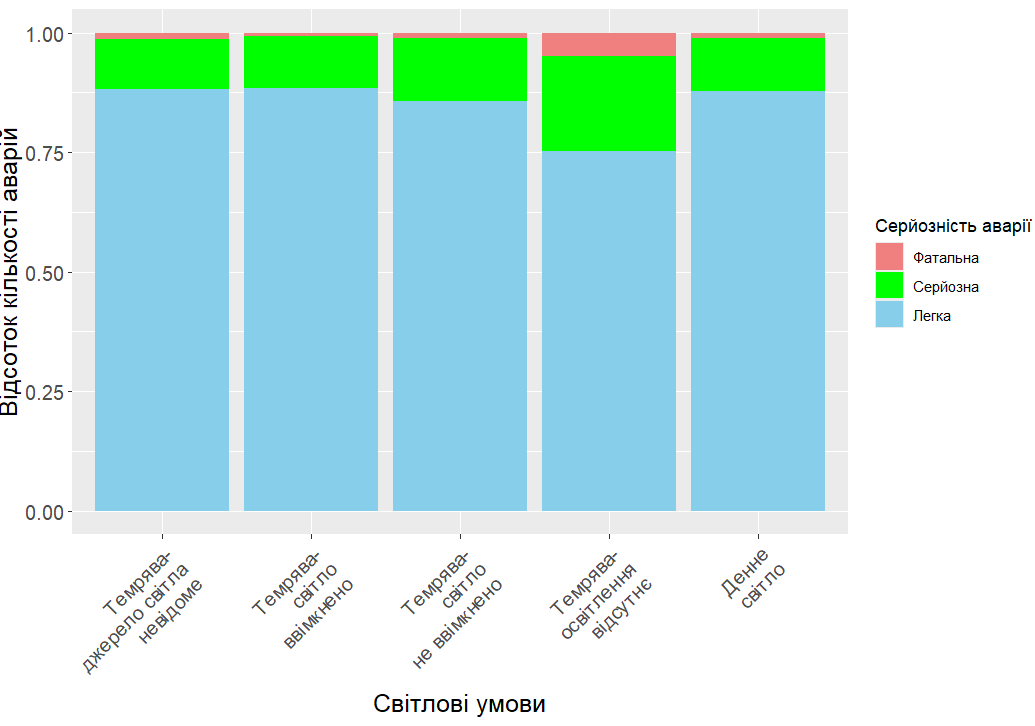


Метеорологічні умови вказують на те, що відсутність сильного вітру пов'язана з найбільшою кількістю аварій, тоді як при тумані та мряці найчастіше відбуваються аварії з серйозними та фатальними наслідками.

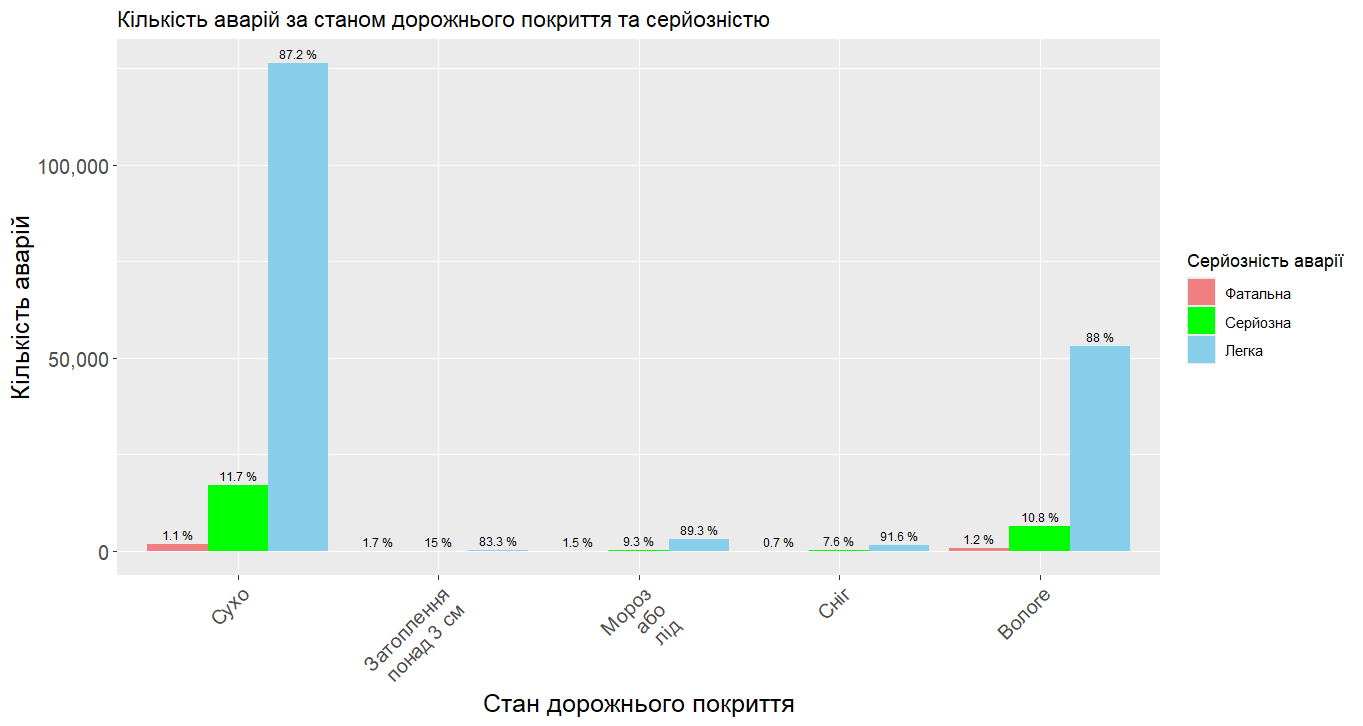
Залежність кількості аварій від світлових умов:



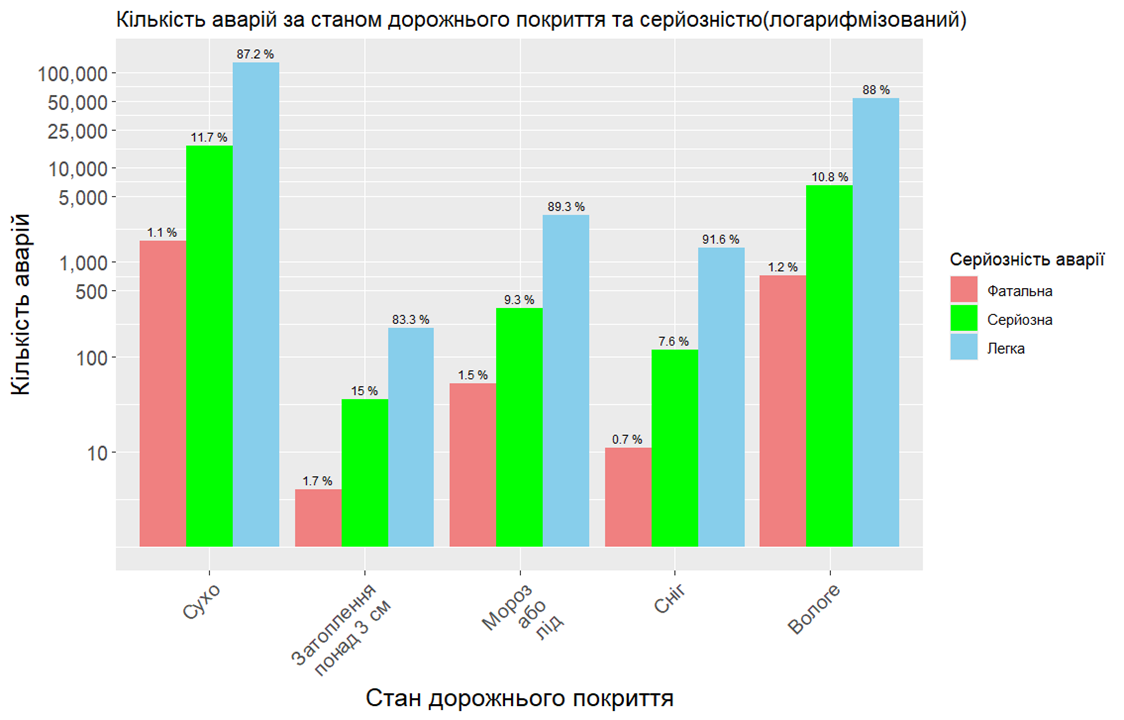
Залежність відсотка кількості аварій від світлових умов:



Кількість аварій за станом дорожнього покриття та серйозністю:

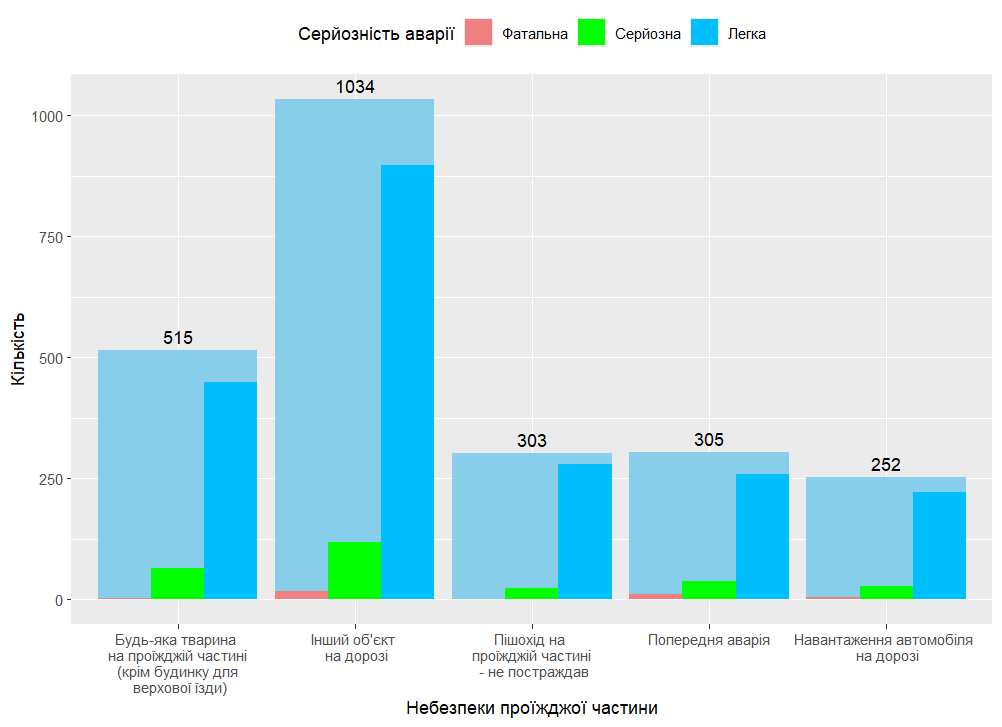


Можемо бачити що, стан покриття "Сухе" становить найбільшу кількість аварій. Умови, коли на дорозі затоплення понад 3 см, найменшу кількість аварій , але високий відсоток серйозних аварій (15%) та найбільший серед фатальних (1.7%). Стан покриття при снігові має найменший відсоток фатальних наслідків та найбільший легких(0.7% та 91.6% відповідно). Стан покриття при умовах "Мороз" та "Вологість" також мають високі відсотки серйозних аварій (9.26% та 10.8% відповідно).



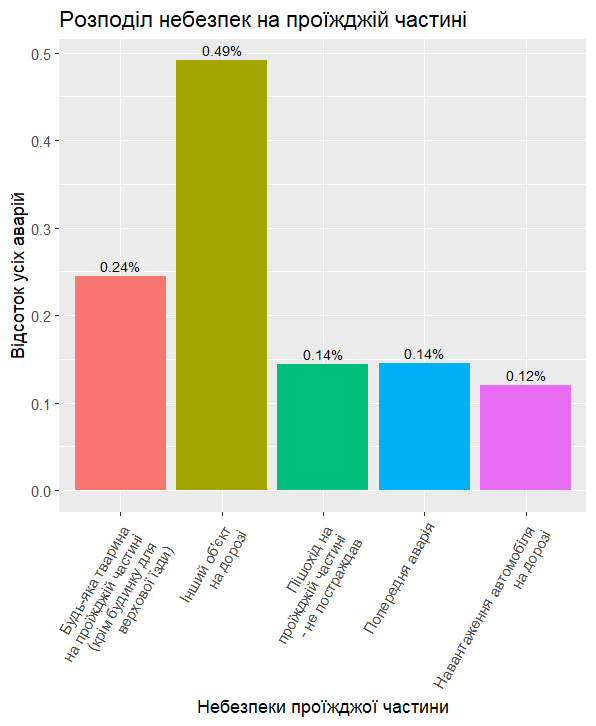
Отже, метеорологічні, світлові умови та умови дорожнього покриття мають значний вплив на серйозність та кількість аварій, що відображає важливість уважності та обережності водіїв при різних факторах на які вони не можуть вплинути.

Залежність кількості і тяжкості аварії від небезпеки проїжджої частини:



ДТП, причинами (неопосередкованими чи побічними) яких були наявність певних перешкод на дорожньому шляху є дуже малою, порівняно із загальною кількістю задокументованих ДТП, де найбільше ДТП спричинили некласифіковані об’єкти на дорозі (наприклад повалене дерево абощо), та наявність тварин на шляху. Однак найбільш смертельно небезпечним були некласифіковані об’єкти та минулі ДТП.

Діаграма наявності перепони на дорожньому шляху, лише з тих випадків, де перепона була:



## ВИСНОВКИ

Для очищення даних було видалено поля, що не мають важливості для EDA в контексті питань, що були розглянуті (наприклад поля, що надає інформацію про присутність представника правоохоронних органів на місці ДТП або широта і довгота місця ДТП), а також полів, інформація в яких є недостовірною або не повною (наприклад поле Journey\_Purpose\_of\_Drive).

У датасеті присутні записи ДТП, в яких Age\_of\_Vehicle дорівнює 63 або 86, однак прямої залежності між віком транспортного засобу і ймовірністю потрапляння на ньому в ДТП не було виявлено.

Не можна впевнено казати про фактори, що спричинили ДТП, оскільки вони не є незалежними. Фатальна аварія могла статись вночі, без освітлення в дощ на поганій дорозі, але казати, що всі три фактори рівнозначно вплинули на ДТП не можна. За деякими виключеннями, всі фактори розглядалися окремо.

Загалом, найбільший вплив на тяжкість і ймовірність потрапляння ДТП є:

* місцевість (в сільській місцевості кількість фатальних ДТП є значно вищою, ніж в міській, і навпаки, кількість легких ДТП в місті перевищує такі в сількій місцевості).
* погодні умови (ДТП під час туману мають найбільше співвідношення фатальних ДТП до легких і серйозних, далі йдуть ДТП під час дощу і снігу)
* світлові умови (ДТП під час темряви, де транспортний засіб не увімкнув фари складають найбільшу частку фатальних і серйозних)