

Лабораторна робота №2

«Статистичне виведення»

Виконали ст. групи км-11:

Команда “CrAzy_pEnGuIn_cOdErs”

Шушпаннікова Інна

Кракович Борислав

Ягода Євгенія

Лушников Іван

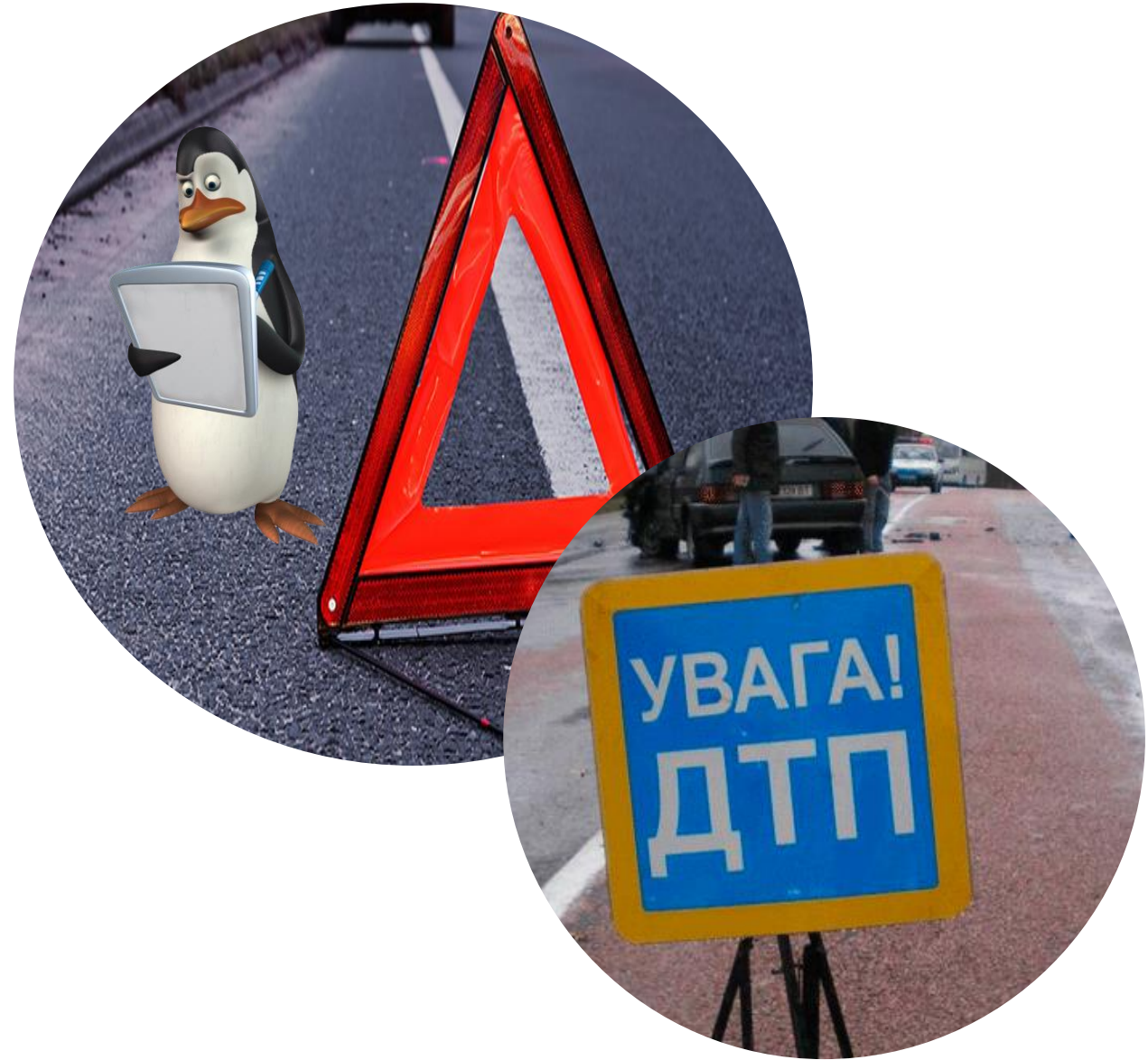
Атаман Юлія



Вступ

Наша команда вирішила зосередитися на дослідженні факторів, що впливають на виникнення та наслідки дорожньо-транспортних пригод.

У даній роботі ми зосетредилися на дослідженні значущості характеристик і закономірностей, які було виявлено у минулій роботі під час розвідкового аналізу даних.



ОПИС ДАНИХ

- Датасет: **UK Road Safety: Traffic Accidents and Vehicles**
- Уряд Великобританії збирає та публікує (зазвичай щорічно) детальну інформацію про дорожньо-транспортні пригоди по всій країні. Ця інформація включає географічне розташування, погодні умови, типи транспортних засобів, кількість постраждалих і маневри транспортних засобів, що робить цей набір дуже цікавим і вичерпним для аналізу та дослідження.
- Датасет демонструє дорожньо-транспортні пригоди та залучені транспортні засоби у Великобританії (2005-2017 рр.). Він містить два файли:
 - Accident_Information.csv(2005-2017): розмір-2047256, 34 змінні. Надає дані про дорожньо-транспортну пригоду.
 - Vehicle_Information.csv(2004-2016): розмір-1488981, 24 змінні. Надає дані транспортного засобу та власника.



Датасет має наступну структуру(обрані колонки)



Accident_Information

1. Accident_Index-Індекс_ДТП
2. Accident_Severity-Серйозність аварії
3. Carriageway Hazards - Небезпека проїжджої частини
4. Date - Дата
5. Day_of_Week – День тижня
6. Junction Detail-перехрестя
7. Light_Conditions - Світлові_умови
8. Number_of_Casualties- Кількість_загиблих
9. Road Surface Conditions-Стан дорожнього покриття
10. Road Type-Тип дороги
11. Speed limit-Обмеження швидкості

12. Time-час
13. Urban or Rural_Area-Міська або сільська місцевість
14. Weather Conditions-Метеорологічні умови
15. Year-Рік
16. Longitude- довгота
17. Latitude-широта
18. Local_Authority_District - райони

Vehicle_Information

1. Accident_Index- Індекс_ДТП
2. Age_Band_of_Driver-Вік автомобіля 1,2
3. Age_of_Vehicle -Вік водія 1,2

4. Sex_of_Driver -Стать водія 1,2
5. make-Марка ТЗ 1,2
6. Vehicle_Manoeuvre -Маневр ТЗ 1,2 під час аварії
7. Vehicle_Type -Тип ТЗ 1,2
8. X1st_Point_of_Impact-Точка зіткнення відносно даного ТЗ 1,2

Опрацювання даних

1. Обираємо 2 транспортні засоби та дані з 2006-2015 роки
2. Видаляємо індекси яких немає у 2 файлі, але є у першому(навпаки)
3. Відбираємо парні та непарні рядки в Vehicle і додаємо в df2 та df1 відповідно
4. Перейменовуємо змінні в df2 та df1 на : Car_Brand1=make, Point_of_Impact1=X1st_Point_of_Impact
5. Об'єднуємо файли за колонкою Accident_Index s_df1, s_df2
6. Об'єднуємо файли c_Accident_Index_filtered , merged_Vehicle за індексом аварії
7. Додаємо дві колонки day та month розділяючи та видаляючи колонку Date
8. Отримуємо датафрейм filtered_data_3 = 349411 записів та 33 змінних.

NA

Age_of_Vehicle1 – 16%
Age_of_Vehicle2 – 18%

Data missing or out of range

Road_Surface_Conditions – 0,13%
Age_Band_of_Driver1 – 5,65%
Age_Band_of_Driver1 – 7,59%



Загальна структура датафрейму після опрацювання

```
'data.frame':  349411 obs. of  33 variables:
 $ Accident_Index      : chr  "200601BS70001" "200601BS70039" "200601BS70041" "200601BS70046" ...
 $ Accident_Severity   : chr  "Slight" "Slight" "Slight" "Slight" ...
 $ Carriageway_Hazards : chr  "None" "None" "None" "None" ...
 $ Day_of_Week         : chr  "Wednesday" "Thursday" "Wednesday" "Wednesday" ...
 $ Junction_Detail     : chr  "T or staggered junction" "Not at junction or within 20 metres" "Roundabout" "Crossroads" ...
 $ Light_Conditions    : chr  "Daylight" "Daylight" "Darkness - lights lit" "Daylight" ...
 $ Number_of_Casualties : int  1 1 1 1 1 1 1 1 2 1 ...
 $ Road_Surface_Conditions : chr  "Dry" "Dry" "Dry" "Dry" ...
 $ Road_Type          : chr  "Single carriageway" "Single carriageway" "Roundabout" "Single carriageway" ...
 $ Speed_limit         : int  30 30 30 30 30 40 30 30 30 30 ...
 $ Area_Type          : chr  "Urban" "Urban" "Urban" "Urban" ...
 $ Weather_Conditions  : chr  "Fine no high winds" "Fine no high winds" "Fine no high winds" "Fine no high winds" ...
 $ Year               : int  2006 2006 2006 2006 2006 2006 2006 2006 2006 2006 ...
 $ Time               : chr  "15:40" "06:30" "23:30" "16:54" ...
 $ Longitude          : num  -0.214 -0.169 -0.216 -0.183 -0.16 ...
 $ Latitude           : num  51.5 51.5 51.5 51.5 51.5 ...
 $ Local_Authority_District: chr  "Kensington and Chelsea" "Kensington and Chelsea" "Kensington and Chelsea" "Kensington and Chelsea" ..
 $ Age_Band_of_Driver1 : chr  "26 - 35" "36 - 45" "26 - 35" "36 - 45" ...
 $ Age_of_Vehicle1     : int  12 8 4 4 13 3 7 10 2 NA ...
 $ Sex_of_Driver1      : chr  "Male" "Male" "Female" "Male" ...
 $ Car_Brand1         : chr  "HONDA" "CITROEN" "MINI" "MAZDA" ...
 $ Vehicle_Manoeuvre1  : chr  "Going ahead other" "U-turn" "Going ahead other" "Waiting to turn right" ...
 $ Vehicle_Type1       : chr  "Motorcycle over 125cc and up to 500cc" "Car" "Car" "Car" ...
 $ Point_of_Impact1    : chr  "Front" "Offside" "Front" "Offside" ...
 $ Age_Band_of_Driver2 : chr  "36 - 45" "26 - 35" "Data missing or out of range" "26 - 35" ...
 $ Age_of_Vehicle2     : int  10 NA 12 1 2 2 3 NA 8 2 ...
 $ Sex_of_Driver2      : chr  "Male" "Male" "Male" "Female" ...
 $ Car_Brand2         : chr  "MITSUBISHI" "YAMAHA" "VAUXHALL" "HONDA" ...
 $ Vehicle_Manoeuvre2  : chr  "Turning right" "Overtaking moving vehicle - offside" "Going ahead other" "Going ahead other" ...
 $ Vehicle_Type2       : chr  "Car" "Car" "Car" "Motorcycle 125cc and under" ...
 $ Point_of_Impact2    : chr  "Front" "Front" "Front" "Front" ...
 $ Day                : int  18 9 1 8 22 27 28 7 13 12 ...
 $ Month              : num  1 2 2 2 2 2 2 3 1 1 ...
```

Дослідницькі питання

01

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

02

Як впливають характеристики автомобіля на кількість/серйозність аварій?

03

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

04

Вплив фізичних особливостей водія на поведінку на дорозі?

05

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

Гіпотези

01

Гіпотеза про рівність середньої кількості фатальних аварій в теплі та холодні місяці.

02

Гіпотеза про рівність середньої кількості фатальних аварій у робочі та вихідні дні.

03

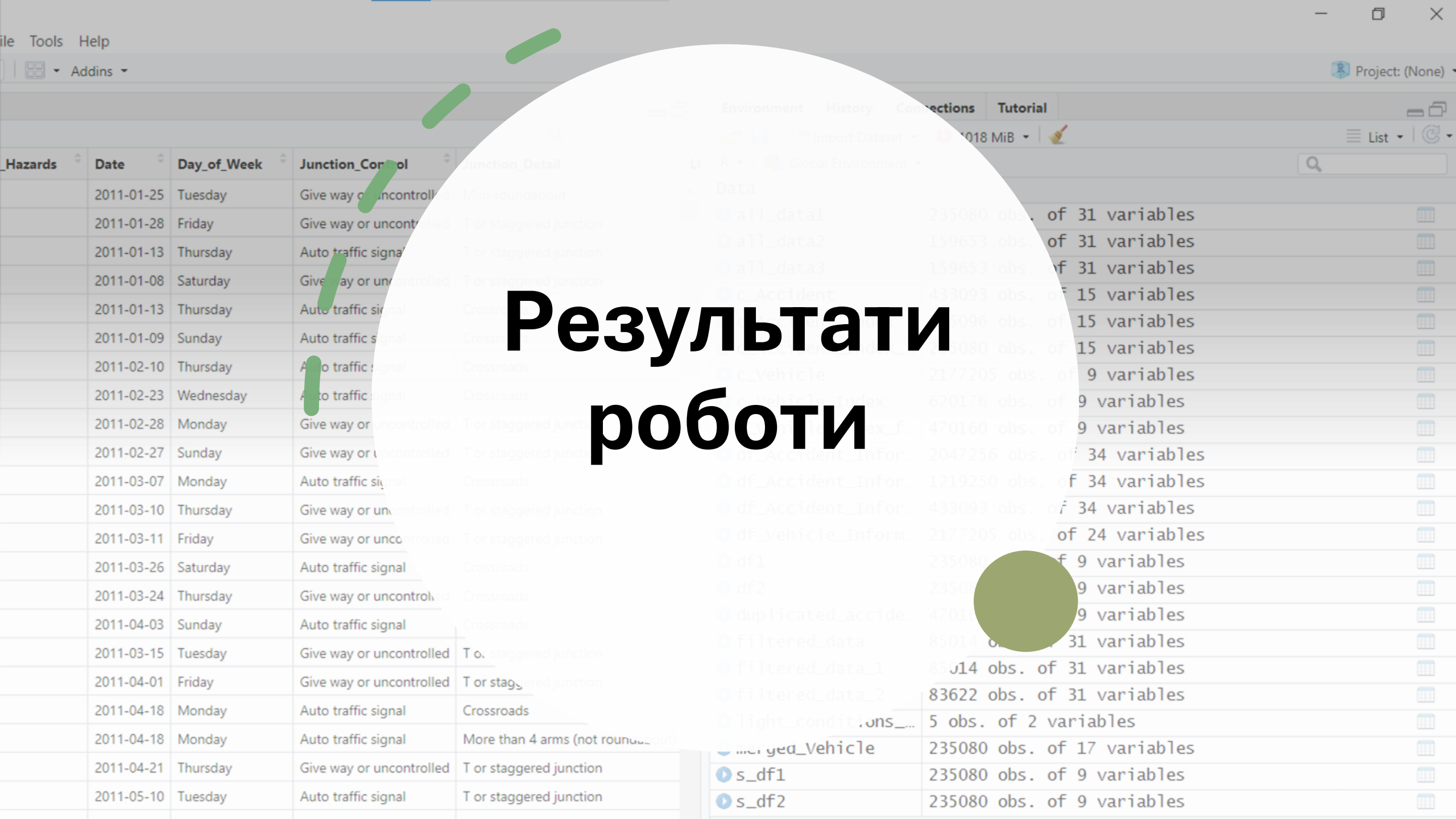
Гіпотеза про рівність кількості фатальних аварій молодших та старших водіїв

04

Гіпотеза про рівність фатальних аварій автомобілей з більшим та меншим віком за роками

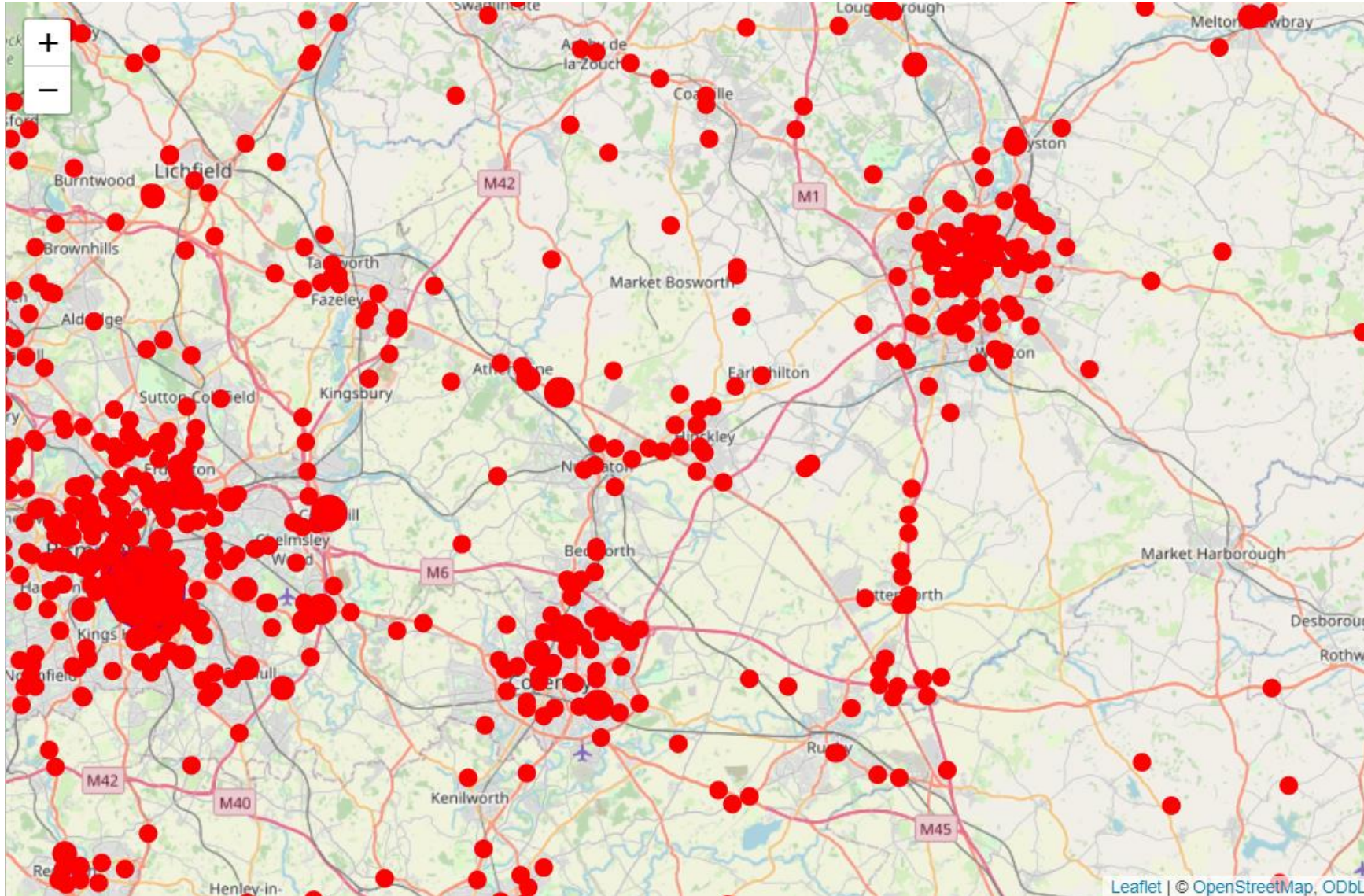
05

Гіпотеза про рівність середньої кількості фатальних аварій за роками у міській та сільській місцевості



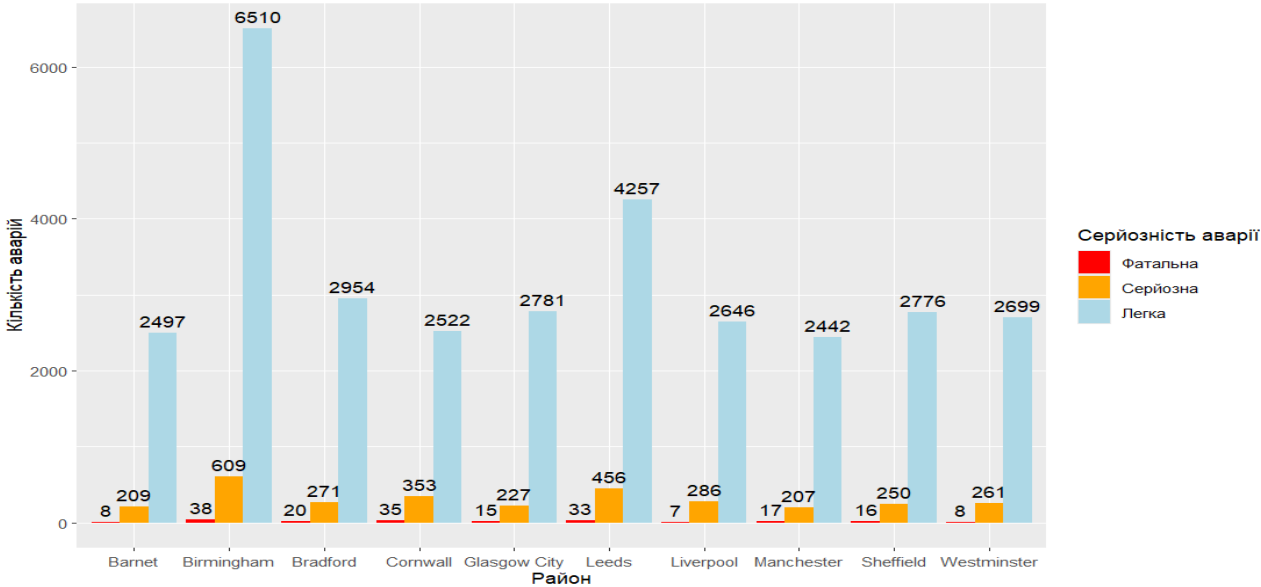
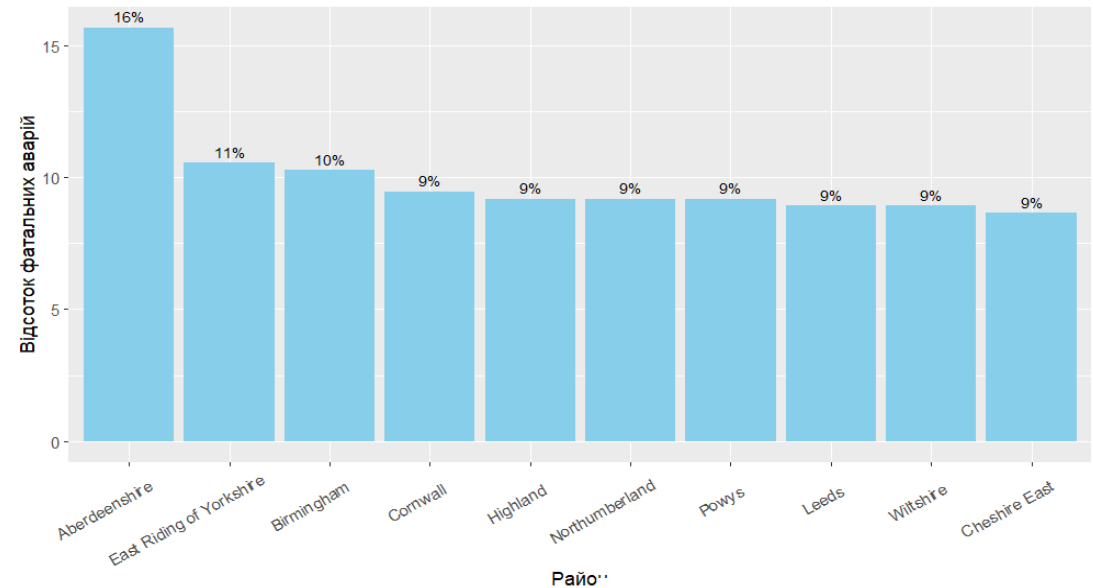
Результати роботи

Мапа аварій

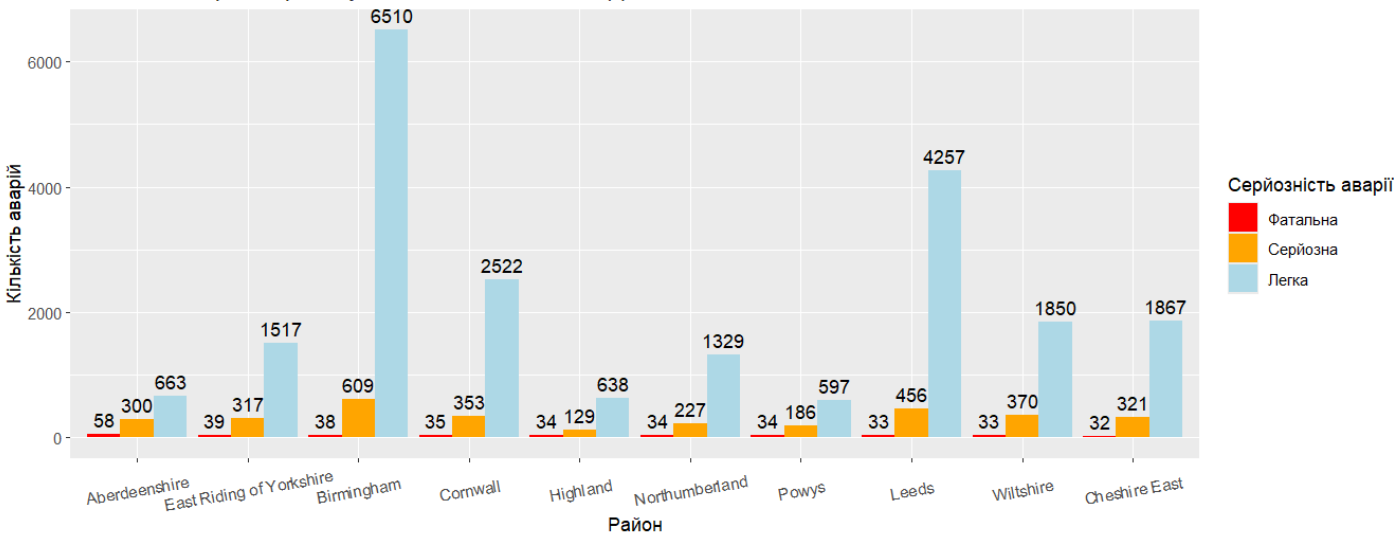


Відображення кількості аварій за районами

Топ-10 фатальних аварій за районами (у відсотках)



Розподіл аварій по району за тяжкістю і кількістю ДТП



1. Гіпотеза про рівність середньої кількості фатальних аварій в теплі та холодні місяці

Обчислимо довірчі інтервали:

Холодних місяців(1,2,3,4,10,11,12):

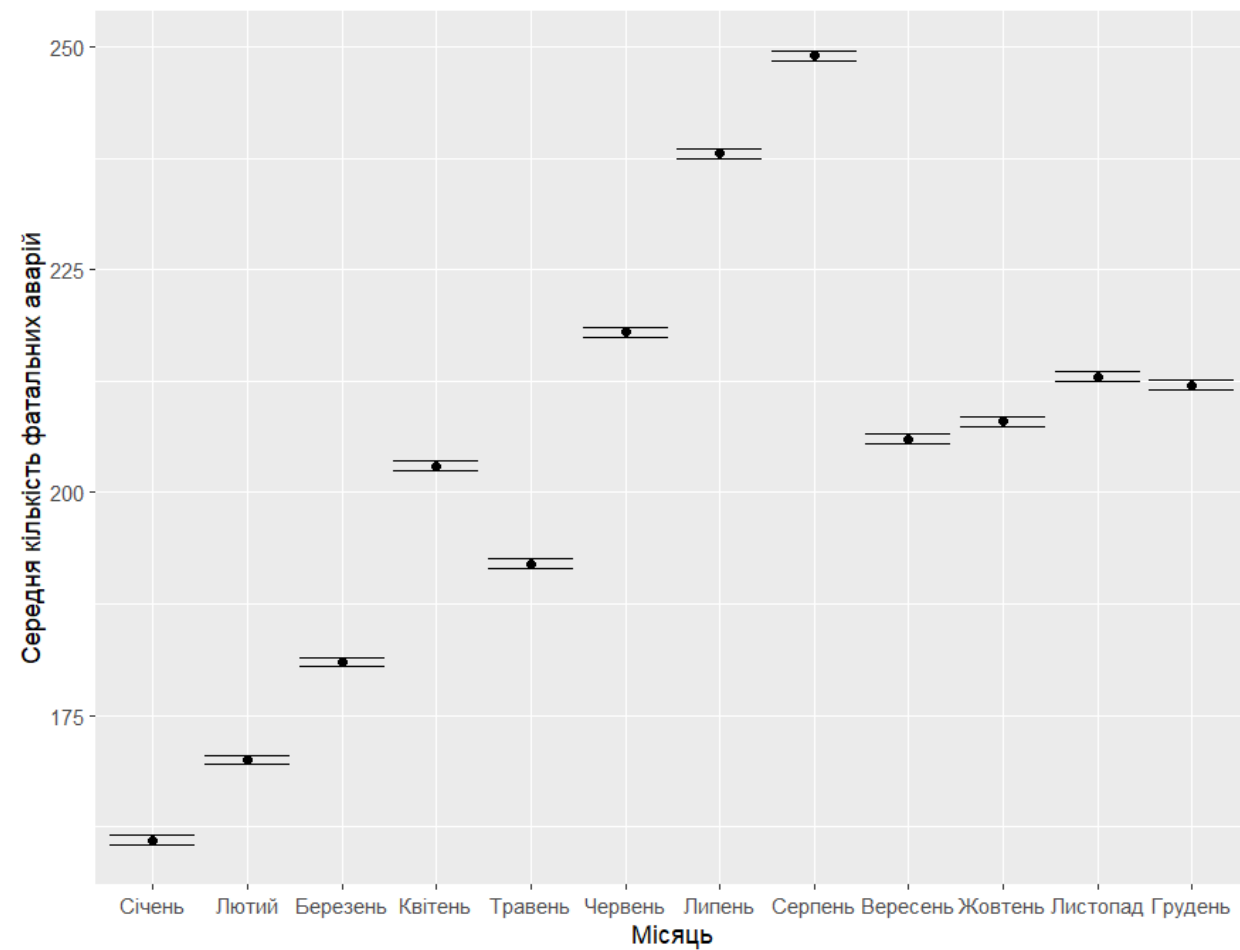
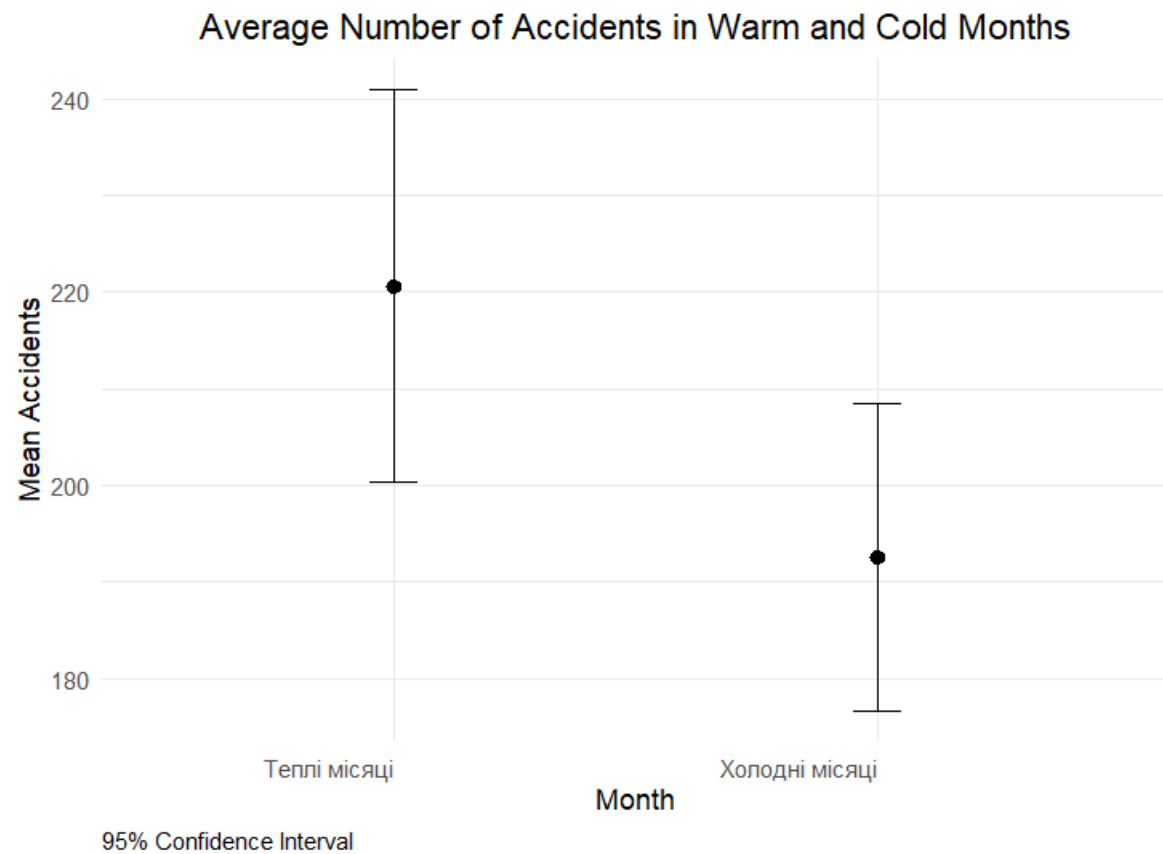
mean	sd	n	a	b
<dbl>	<dbl>	<int>	<dbl>	<dbl>
193.	21.5	7	177.	209.

Теплих місяців(5,6,7,8,9):

mean	sd	n	a	b
<dbl>	<dbl>	<int>	<dbl>	<dbl>
221.	23.2	5	200.	241.

```
[1] "Довірчий інтервал для 1"
  mean      sd      n      a      b
1  161 12.26476 2451 160.5144 161.4856
[1] "Довірчий інтервал для 2"
  mean      sd      n      a      b
1  170 12.57811 2451 169.502 170.498
[1] "Довірчий інтервал для 3"
  mean      sd      n      a      b
1  181 12.94734 2451 180.4874 181.5126
[1] "Довірчий інтервал для 4"
  mean      sd      n      a      b
1  203 13.64503 2451 202.4598 203.5402
[1] "Довірчий інтервал для 5"
  mean      sd      n      a      b
1  192 13.30262 2451 191.4734 192.5266
[1] "Довірчий інтервал для 6"
  mean      sd      n      a      b
1  218 14.09292 2451 217.4421 218.5579
[1] "Довірчий інтервал для 7"
  mean      sd      n      a      b
1  238 14.65911 2451 237.4197 238.5803
[1] "Довірчий інтервал для 8"
  mean      sd      n      a      b
1  249 14.95673 2451 248.4079 249.5921
[1] "Довірчий інтервал для 9"
  mean      sd      n      a      b
1  206 13.73631 2451 205.4562 206.5438
[1] "Довірчий інтервал для 10"
  mean      sd      n      a      b
1  208 13.79668 2451 207.4538 208.5462
[1] "Довірчий інтервал для 11"
  mean      sd      n      a      b
1  213 13.94595 2451 212.4479 213.5521
[1] "Довірчий інтервал для 12"
  mean      sd      n      a      b
1  212 13.91629 2451 211.4491 212.5509
```

← Для кожного місяця



Проведемо двосторонній тест Волда та t-test Велча на рівні значущості 0.05, а також критерій (Фішера)

Результати тесту Волда:

```
$mean_x  
[1] 220.6  
  
$mean_y  
[1] 192.5714  
  
$p_value  
[1] 0.06467324  
  
$conf_int  
[1] -2.138961 58.196104
```



За результатами тесту Волда отримано наступні значення:
Довірчий інтервал для різниці середніх між теплими та холодними місяцями: від -2.138961 до 58.196104
Р-значення = 0.06467324 перевищує зазначений рівень значущості(0.05), що означає відсутність достовірних даних для відхилення нульової гіпотези про рівність середніх.

Welch Two Sample t-test

```
data: ci11$mean_accidents and ci22$mean_accidents  
t = -2.1273, df = 8.3413, p-value = 0.06467  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -58.196104  2.138961  
sample estimates:  
mean of x mean of y  
 192.5714  220.6000
```



Результат Welch Two Sample t-test показує, що p-value = 0.06467, що більше типового рівня значимості 0.05. Це означає, що немає статистично значущих доказів на користь того, що середні значення кількості фатальних аварій в холодні та теплі місяці відрізняються.

Дисперсія для холодних місяців: 463.619047619048
Дисперсія для теплих місяців: 536.8

Перевірка гіпотези про рівність генеральних дисперсій. F-критерій (Фішера)

F = 0.86367, num df = 6, denom df = 4, p-value = 0.8294

2. Гіпотеза про рівність середньої кількості фатальних аварій у робочі та вихідні дні

Обчислимо довірчі інтервали для легких та серйозних аварій:

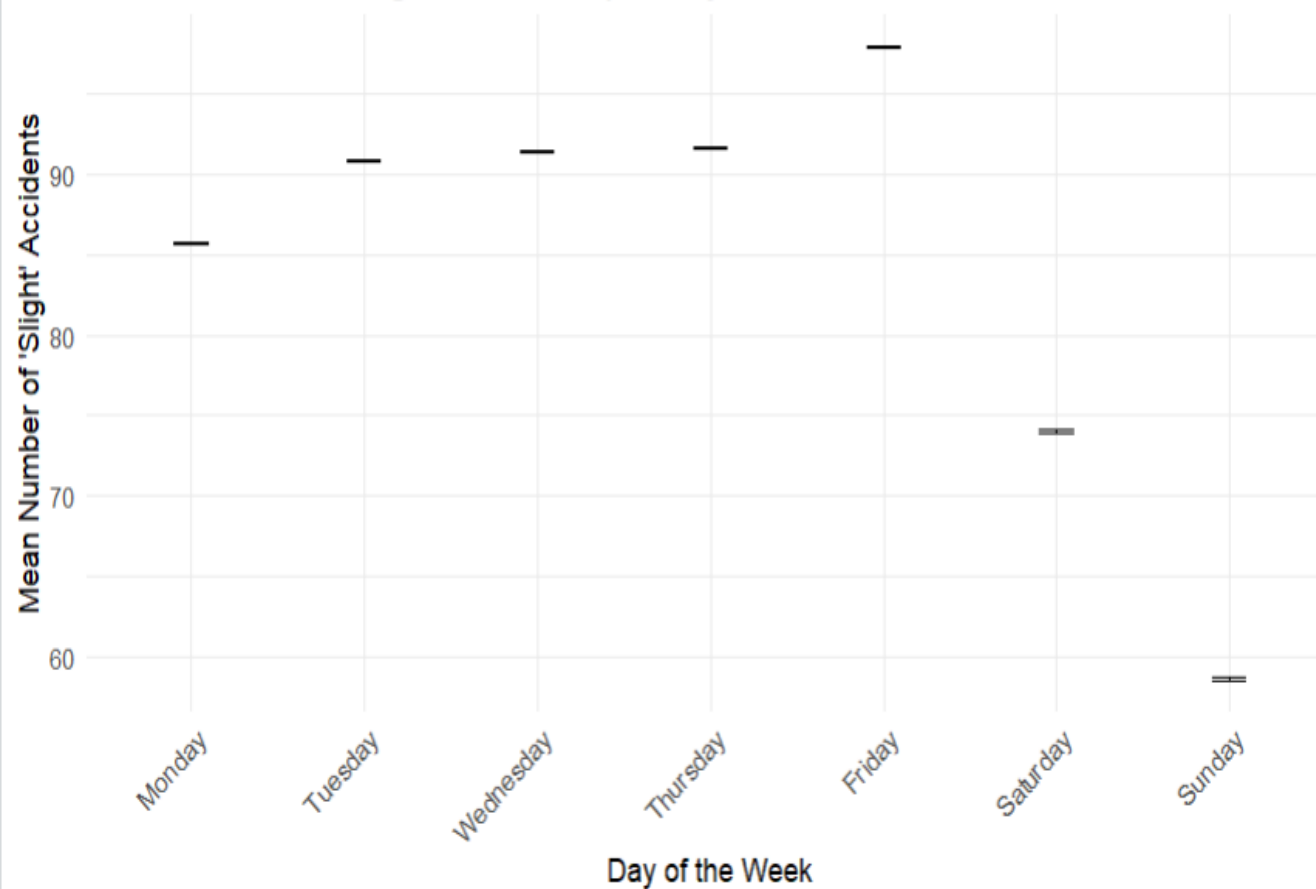
```
Day_of_Week mean_slight sd_slight      n      a      b
<chr>      <dbl>      <dbl> <int>  <dbl>  <dbl>
1 Monday      85.7356      8.46487 44754 85.6572 85.8141
2 Tuesday     90.8317      8.66352 47505 90.7538 90.9096
3 Wednesday   91.4195      8.68383 47721 91.3416 91.4975
4 Thursday    91.6628      8.69292 47848 91.5849 91.7407
5 Friday      97.9098      8.91683 51011 97.8324 97.9872
6 Saturday    73.9981      7.96794 38553 73.9185 74.0776
7 Sunday      58.6207      7.21371 30600 58.5399 58.7015
```

>

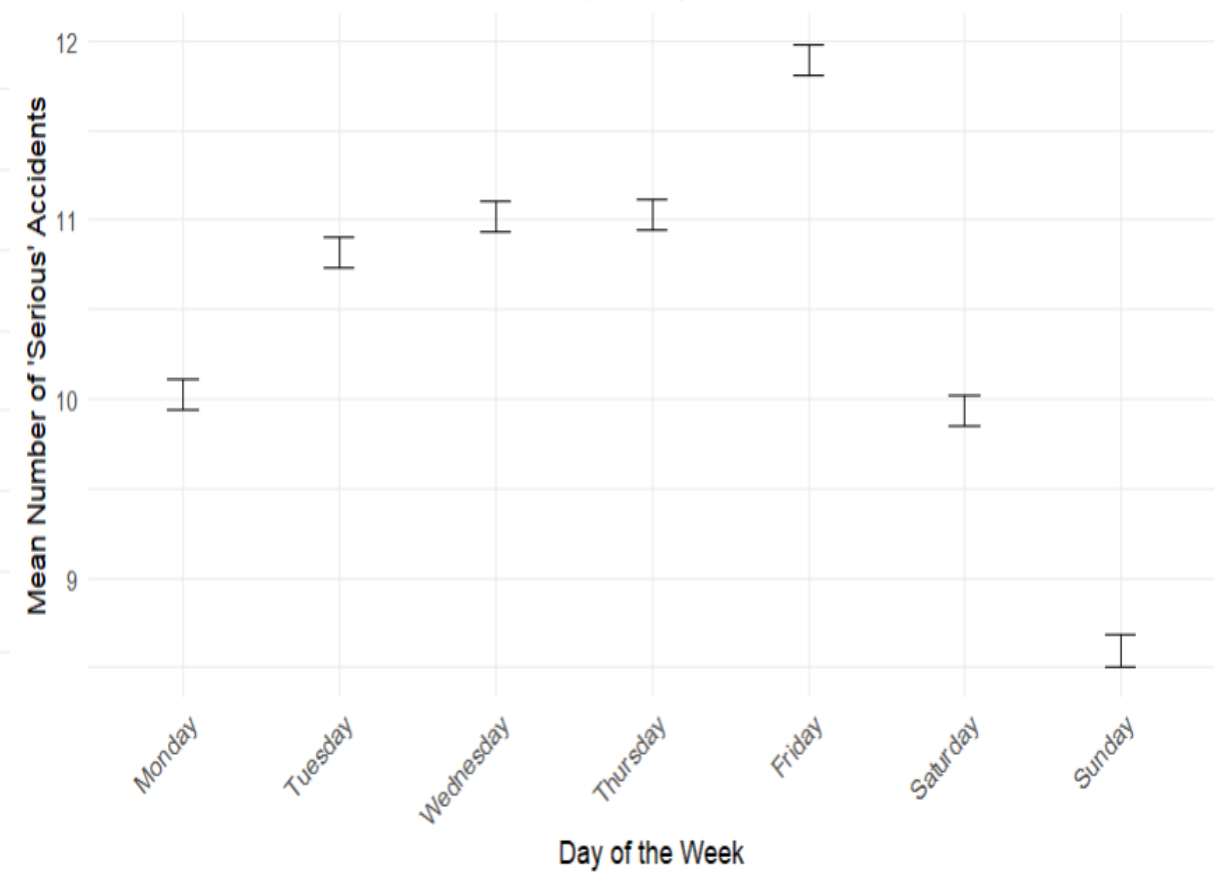
```
Day_of_Week mean_serious sd_serious      n      a      b
<chr>      <dbl>      <dbl> <int>  <dbl>  <dbl>
1 Monday     10.0211      3.13502  5221  9.93608 10.1062
2 Tuesday     10.8205      3.25491  5605 10.7353 10.9057
3 Wednesday   11.0173      3.28382  5718 10.9322 11.1025
4 Thursday    11.0289      3.28550  5724 10.9438 11.1140
5 Friday      11.8942      3.40913  6185 11.8093 11.9792
6 Saturday     9.93629      3.12181  5147  9.85101 10.0216
7 Sunday       8.59332      2.90659  4374  8.50718  8.67946
```



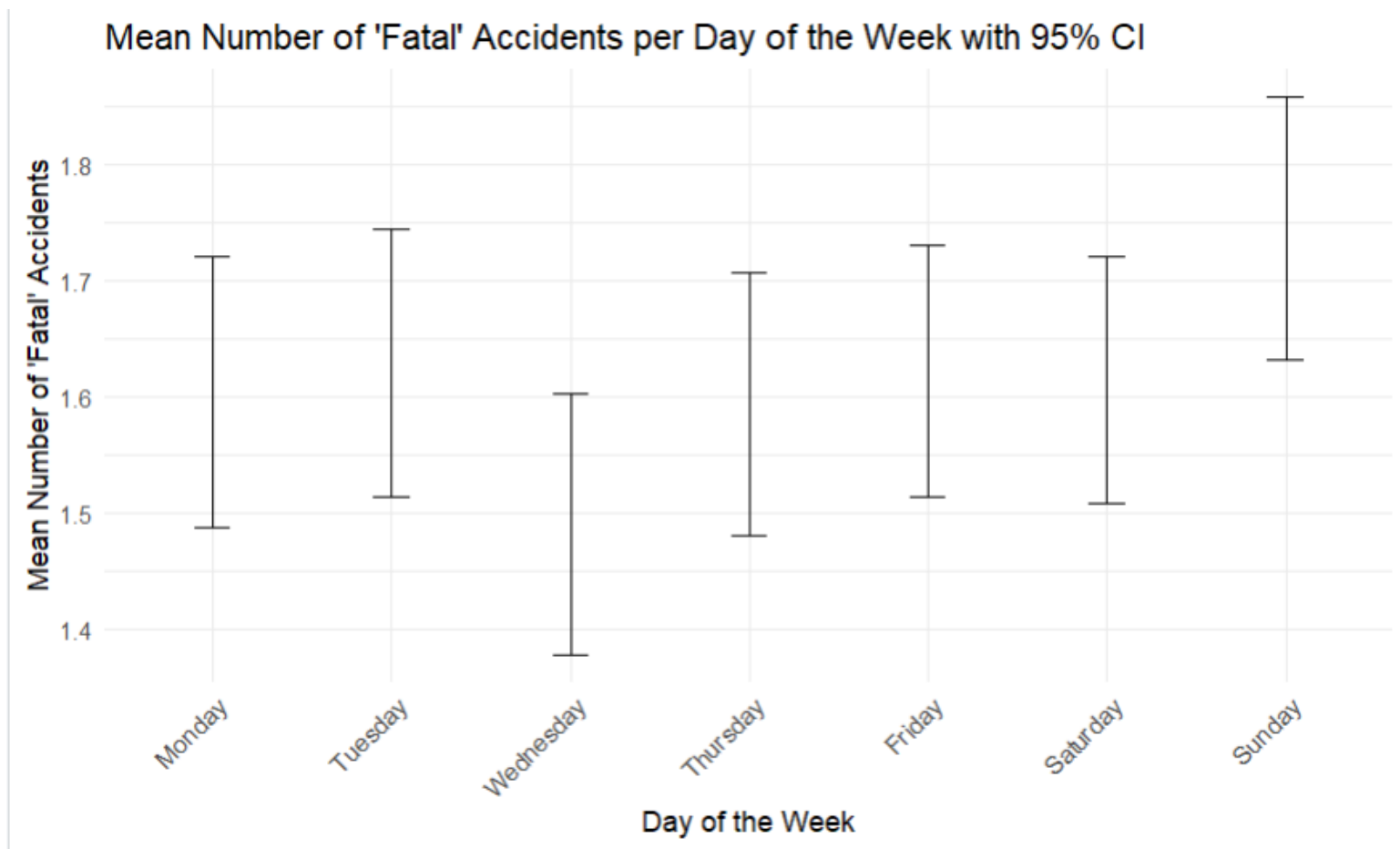
Mean Number of 'Slight' Accidents per Day of the Week with 95% CI



Mean Number of 'Serious' Accidents per Day of the Week with 95% CI



Day_of_Week	mean_fatal	sd_fatal	n	a	b
<chr>	<dbl>	<dbl>	<int>	<dbl>	<dbl>
1 Monday	1.60357	1.26269	449	1.48678	1.72037
2 Tuesday	1.62847	1.27250	469	1.51331	1.74364
3 Wednesday	1.49013	1.21771	453	1.37800	1.60227
4 Thursday	1.59333	1.25892	478	1.48048	1.70619
5 Friday	1.62154	1.27022	527	1.51309	1.72999
6 Saturday	1.61471	1.26769	549	1.50866	1.72075
7 Sunday	1.74497	1.31710	520	1.63176	1.85817



Проведемо двосторонній тест Волда та t-test Велча на рівні значущості 0.05, а також порівняємо результати

Для перевірки гіпотези було обрано дні: четвер (що має “середні” значення відносно інших будніх днів і неділя.

Результати тестів Волда і Велча:

```
Wald Test Statistic: 1.85921  
P-value (Wald Test): 0.0314988
```

```
Welch Test Statistic: 1.85921  
P-value (Welch Test): 0.0316464
```

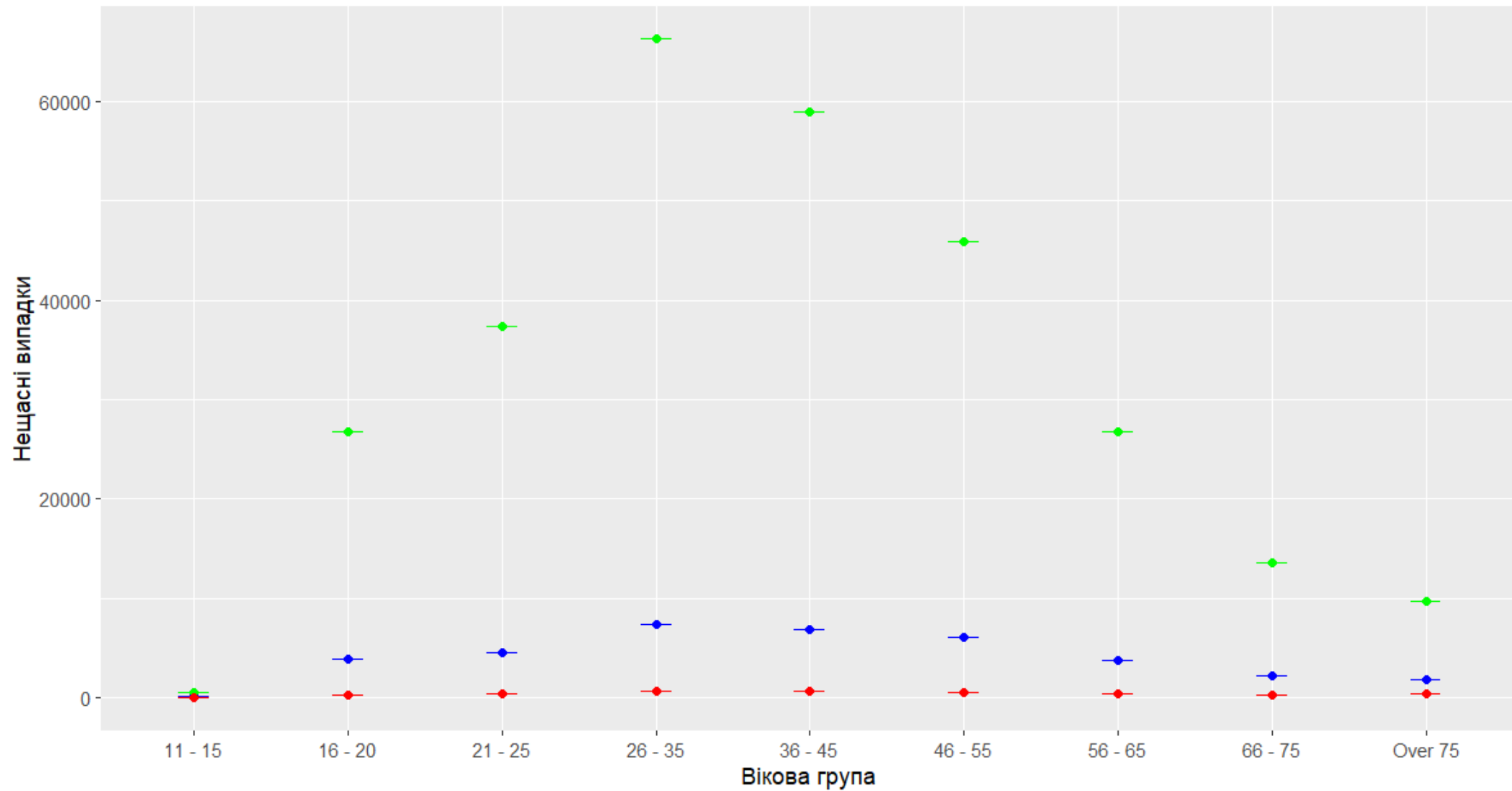
Результати тестів вказують на статистично значущу різницю у середній кількості смертельних аварій між цими двома днями. Відхиливши нульову гіпотезу, ми можемо зробити висновок, що існує статистично значуща різниця між кількістю смертельних аварій у четвер та неділю.

3. Гіпотеза про рівність кількості фатальних аварій молодших та старших водіїв

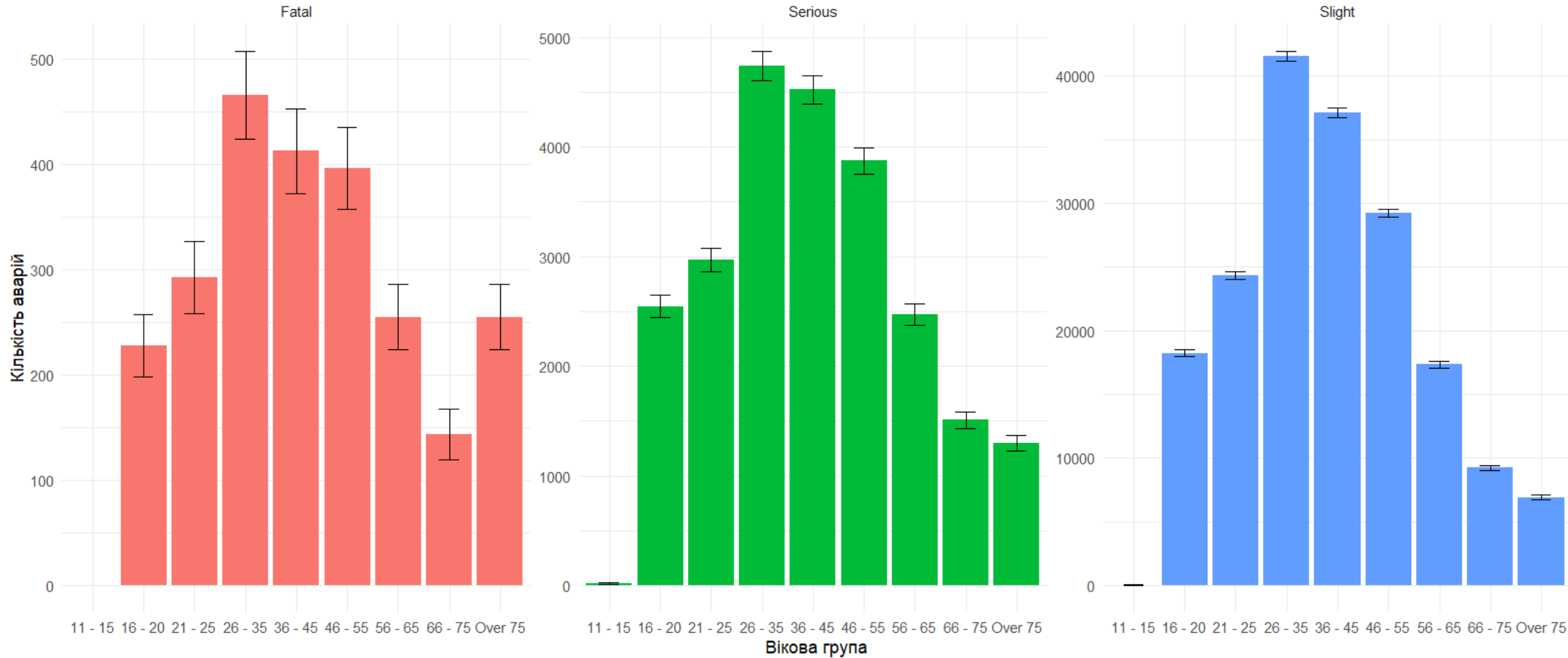
Обчислимо довірчі інтервали:

	Accident_Severity <chr>	Age_Band_of_Driver1 <chr>	Count <int>	Lower_CI <dbl>	Upper_CI <dbl>
1	Fatal	16 - 20	228	198	258
2	Fatal	21 - 25	293	259	327
3	Fatal	26 - 35	466	424	508
4	Fatal	36 - 45	413	373	453
5	Fatal	46 - 55	397	358	436
6	Fatal	56 - 65	255	224	286
7	Fatal	66 - 75	144	120	168
8	Fatal	Over 75	255	224	286
9	Serious	11 - 15	17	9	25
10	Serious	16 - 20	2547	2448	2646
11	Serious	21 - 25	2968	2861	3075
12	Serious	26 - 35	4739	4604	4874
13	Serious	36 - 45	4522	4390	4654
14	Serious	46 - 55	3874	3752	3996
15	Serious	56 - 65	2472	2375	2569
16	Serious	66 - 75	1511	1435	1587
17	Serious	Over 75	1301	1230	1372
18	Slight	11 - 15	39	27	51
19	Slight	16 - 20	18241	17976	18506
20	Slight	21 - 25	24365	24059	24671
21	Slight	26 - 35	41558	41158	41958
22	Slight	36 - 45	37127	36749	37505
23	Slight	46 - 55	29242	28907	29577
24	Slight	56 - 65	17361	17103	17619
25	Slight	66 - 75	9262	9073	9451
26	Slight	Over 75	6929	6766	7092

Середня кількість ДТП за віковими групами водіїв та рівнем тяжкості



Кількість аварій за віковими групами та тяжкістю з довірчими інтервалами

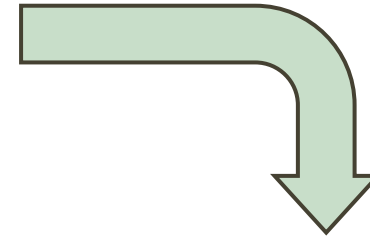


Проведемо Хі-квадрат тест

Нульова гіпотеза (H_0): Середня кількість фатальних аварій на дорогах серед молодими та старшими водіями не відрізняється.

Альтернативна гіпотеза (H_1): Середня кількість фатальних аварій на дорогах серед молодими та старшими водіями відрізняється.

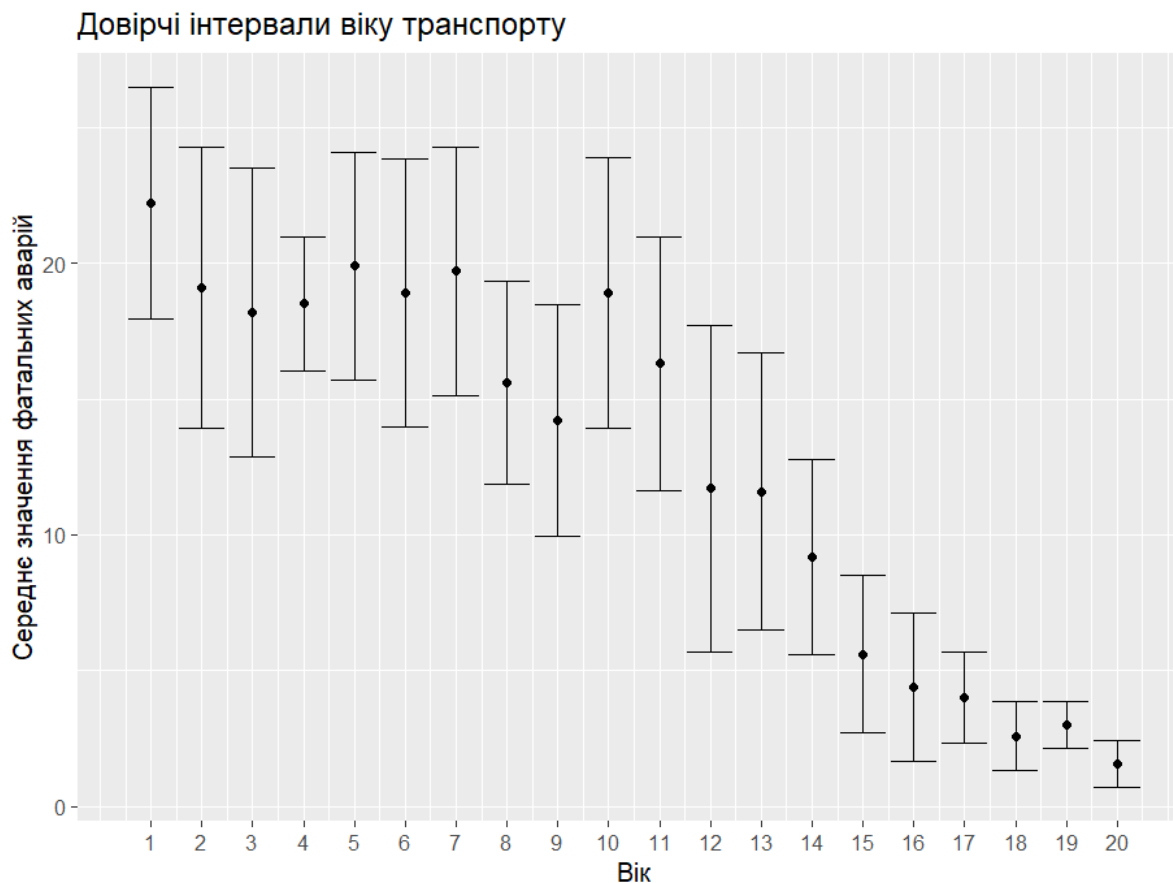
```
[1] "Chi-Squared Test Results:"  
[1] "Chi-Squared Statistic: 46.7456713145021"  
[1] "Degrees of Freedom: 1"  
[1] "p-value: 8.08232661963512e-12"
```



Можна зробити висновок, що молоді водії, ймовірно, демонструють вищу схильність до смертельних аварій, можливо, через їхню більш ризиковану поведінку та/або брак досвіду за кермом порівняно зі старшими водіями.

4. Гіпотеза про рівність фатальних аварій автомобілей з більшим та меншим віком за роками

	mean	sd	n	a	b
	<dbl>	<dbl>	<int>	<dbl>	<dbl>
1	22.2	6.86	10	17.9	26.5
2	19.1	8.37	10	13.9	24.3
3	18.2	8.57	10	12.9	23.5
4	18.5	3.98	10	16.0	21.0
5	19.9	6.77	10	15.7	24.1
6	18.9	7.98	10	14.0	23.8
7	19.7	7.38	10	15.1	24.3
8	15.6	6.04	10	11.9	19.3
9	14.2	6.86	10	9.95	18.5
10	18.9	8.06	10	13.9	23.9
11	16.3	7.54	10	11.6	21.0
12	11.7	9.72	10	5.68	17.7
13	11.6	8.21	10	6.51	16.7
14	9.2	5.79	10	5.61	12.8
15	5.6	4.67	10	2.70	8.50
16	4.4	4.43	10	1.66	7.14
17	4	2.71	10	2.32	5.68
18	2.6	2.01	10	1.35	3.85
19	3	1	5	2.12	3.88
20	1.57	1.13	7	0.731	2.41



Медіана віку автомобілів у фатальних аваріях: 7
95% довірчий інтервал медіани: 6.733564 - 7.266436

Проведемо двосторонній тест Волда та t-test Велча на рівні значущості 0.05, а також обч. Коеф. Спірмана

Для визначення зв'язку між віком автомобіля та кількістю фатальних аварій обчислимо коефіцієнт Спірмана

```
[1] "Коефіцієнт кореляції Спірмана між віком автомобіля і кількістю фатальними аваріями  
: -0.831011184486773"
```

Коефіцієнт кореляції Спірмана -0.831011184486773 вказує на те, що існує досить сильний зворотній взаємозв'язок між віком автомобіля та кількістю фатальних аварій. Якщо поділити вік авто на дві групи 1-10 та >10, то

```
"Коефіцієнт кореляції Спірмана між віком автомобіля і кількістю фатальними аваріями : -1"
```

Це означає дуже сильну негативну лінійну залежність між цими віковими групами та кількістю фатальних аварій.

Тест Волда

```
$mean_x
[1] 18.52
$mean_y
[1] 6.997143
$p_value
[1] 1.61167e-05
$conf_int
[1] 7.824452 15.221263
```



На підставі отриманих результатів можна відкинути нульову гіпотезу про рівність середніх значень двох груп даних. Значення p-value ($1.61167e-05$) менше за зазначений рівень значимості (наприклад, 0.05), що свідчить про статистичну значущість різниці між групами. Таким чином, ми маємо достатні підстави вважати, що середні значення цих двох груп статистично відмінні одне від одного.

Тест Велча

Welch Two Sample t-test

```
data: first_10_means and second_10_means
```

```
t = 6.755, df = 12.559, p-value = 1.612e-05
```

```
alternative hypothesis: true difference in means is not equal to 0
```

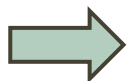
```
95 percent confidence interval:
```

```
7.824452 15.221263
```

```
sample estimates:
```

```
mean of x mean of y
```

```
18.520000 6.997143
```

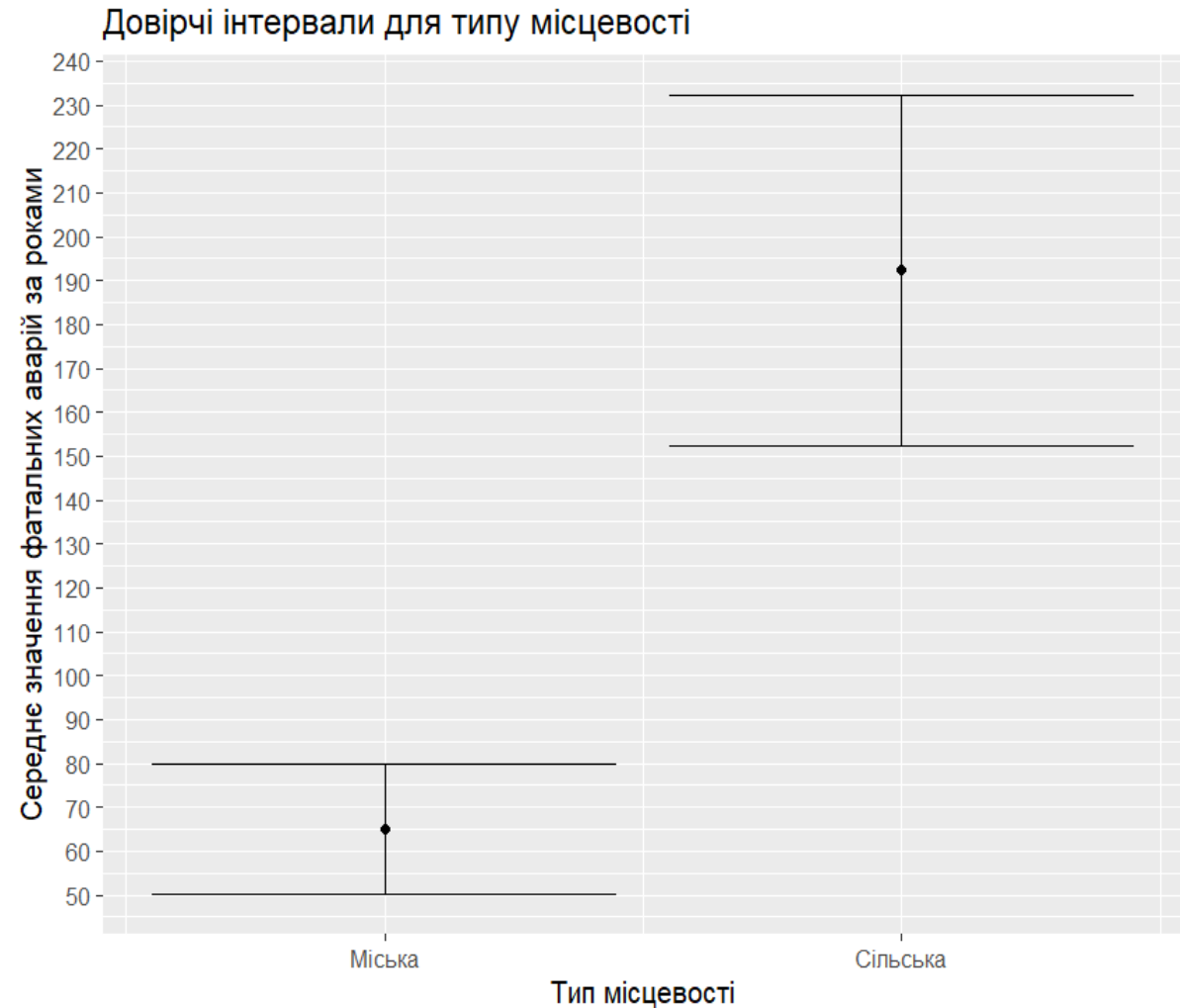


З результатів тесту Велча можна зробити наступні висновки: значення p-рівня статистичної значущості груп (перші 10 років та другі 10) дорівнює $1.612e-05$, що є дуже низьким значенням. Це свідчить про те, що різниця між середніми значеннями груп є статистично значущою

5. Гіпотеза про рівність середньої кількості фатальних аварій за роками у міській та сільській місцевості

Довірчі інтервали для міської і сільської місцевості відповідно:

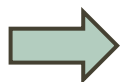
	mean	sd	n	a	b
	<dbl>	<dbl>	<int>	<dbl>	<dbl>
1	65	24.1	10	50.1	79.9
2	192.	64.4	10	152.	232.



Проведемо двосторонній тест Волда та t-test Велча на рівні значущості 0.05, а також порівняємо результати

Тест Волда:

```
$mean_x  
[1] 192.4  
$mean_y  
[1] 65  
$p_value  
[1] 9.317599e-05  
$conf_int  
[1] 79.74555 175.05445
```



На підставі результатів тесту Волда, який показав дуже мале значення р-рівня ($9.317599e-05$), можна відхилити нульову гіпотезу про рівність середньої кількості фатальних аварій у сільській та міській місцевості.

Тест Велча:

Welch Two Sample t-test

```
data: mean_fatal_rural and mean_fatal_urban  
t = 5.8548, df = 11.471, p-value = 9.318e-05  
alternative hypothesis: true difference in means is not equal  
to 0  
95 percent confidence interval:  
 79.74555 175.05445  
sample estimates:  
mean of x mean of y  
 192.4      65.0
```



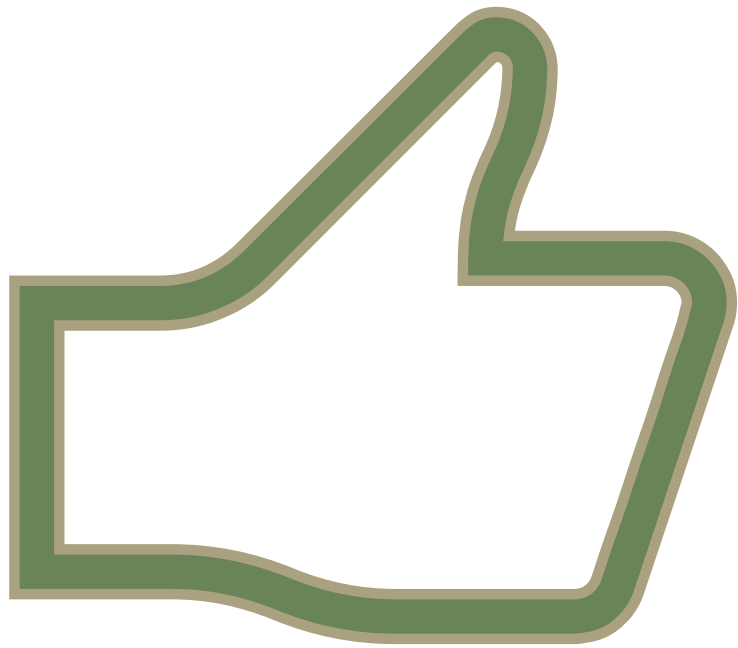
Отримане значення р-рівня ($9.318e-05$) є набагато меншим, ніж зазначений рівень значущості, що свідчить про статистичну значимість різниці у середніх кількостях фатальних аварій у сільській та міській місцевості. Отже, ми можемо відхилити нульову гіпотезу про рівність середніх.

ВИСНОВКИ

В ході цієї роботи ми провели статистичне виведення для різних гіпотез, що стосувалися факторів, які можуть впливати на серйозність і частоту дорожньо-транспортних пригод. Результати нашого аналізу дозволили нам встановити декілька ключових зв'язків:

1. Сезонність аварій: Виявлено статистично значущу різницю в кількості аварій між теплими та холодними місяцями.
2. Вплив дня тижня на прикладі четверга і неділі: Фатальні аварії частіше відбуваються у вихідні, що може бути пов'язано з характером поїздок .
3. Вік водіїв: Молодші водії схильні до більшої кількості аварій, що свідчить про важливість освітніх програм для цієї категорії.
4. Вік транспортного засобу: Авто з меншим віком потрапляють частіше в ДТП.
5. Вплив місцевості: Виявлено, що фатальні аварії частіше відбуваються у сільській місцевості, ніж у міській.

Ці відкриття можуть бути використані для розробки цілеспрямованих заходів з підвищення дорожньої безпеки, включаючи інформаційні кампанії, поліпшення інфраструктури та зміну дорожнього законодавства, щоб зменшити частоту та серйозність аварій.



Дякуємо за увагу!