МІНІСТЕРСТВО НАУКИ І ОСВІТИ УКРАЇНИ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ № 1 ЗА ТЕМОЮ:

**Попередня підготовка даних**

Група \_\_\_\_\_11\_\_\_\_\_\_

Курс \_\_\_\_\_1\_\_\_\_\_\_\_\_\_

Студент (ка) \_\_\_\_Супруненко М. І. \_\_\_\_

Дата оформлення \_\_\_\_\_\_16.02.2025\_\_\_\_\_\_\_

Перевірив \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

МЕТА РОБОТИ: Метою лабораторної роботи є формування професійних вмінь та навичок щодо використання інструментів та методів програмного середовища R в задачах виявлення та обробки викидів в наборах даних, вміння застосовувати отримані знання на практиці в практичних задачах інтелектуального аналізу даних.

ТЕОРІЯ: Основний теоретичний матеріал висвітлений в лекційних матеріалах за тематикою «Попередня обробка даних у інтелектуальному аналізі даних».

Викид — це значення або спостереження, яке є далеким від інших спостережень, тобто точка даних, яка суттєво відрізняється від інших точок даних. Деякі автори йдуть ще далі, оскільки розглядають викиди як значення, які настільки сильно відрізняються від інших спостережень, що можна припустити інший основний механізм вибірки.

У цій роботі ми дослідимо кілька підходів для виявлення викидів у R, від простих методів, таких як описова статистика (включно з мінімумом, максимумом, гістограмою, boxplot діаграмою та процентилями), до більш формальних методів, таких як фільтр Гампеля, Граббса, Діксона та Тести Рознера на викиди.

Усі спостереження за межами наступного інтервалу вважатимуться потенційними викидами:

,

Де -  interquartile range (внутрішній квартильний діапазон).

За допомогою методу процентилів усі спостереження, що знаходяться за межами інтервалу, утвореного 2,5 та 97,5 процентилями, розглядатимуться як потенційні викиди.

Якщо ваші дані походять із нормального розподілу, ви можете використовувати z-показники. Відповідно до цього методу будь-який z-показник:

< -2 або > 2 вважаються рідкісними.

< -3 або > 3 вважаються надзвичайно рідкісними.

Інший метод, відомий як фільтр Гампеля, полягає в розгляді як викидів значень за межами інтервалу (I), утвореного медіаною плюс-мінус 3 медіанних абсолютних відхилень (MAD):

– медіана.

Для цього методу ми спочатку встановлюємо межі інтервалів завдяки функціям median() і mad().

У роботі ми розглянемо 3 перевірки гіпотез для виявлення викидів: тест Граббса, тест Діксона, тест Рознера (Grubbs’s test, Dixon’s test, Rosner’s test)

Ці 3 статистичні тести є частиною більш формальних методів виявлення викидів, оскільки всі вони включають обчислення тестової статистики, яка порівнюється з табличними критичними значеннями (які базуються на розмірі вибірки та бажаному рівні достовірності).

Зауважте, що 3 тести підходять лише тоді, коли дані без будь-яких викидів розподіляються приблизно нормально. Рекомендується перевіряти нормальність візуально, наприклад, за допомогою графіка QQ-гістограми та/або boxplot графіка. Хоча його також можна перевірити за допомогою формального тесту на нормальність (наприклад, тесту Шапіро-Вілка), наявність одного або кількох викидів може призвести до того, що тест нормальності відхилить нормальність, коли це насправді є розумним припущенням для застосування одного із 3 тестів на викиди, згаданих вище.

Тест Граббса дозволяє визначити, чи є найвище чи найнижче значення в наборі даних викидом. Тест Граббса виявляє один викид за раз (найвище або найнижче значення), тому нульова та альтернативна гіпотези є такими:

H0 : найвище значення не є викидом;

H1: найвище значення є викидом;

якщо ми хочемо перевірити найвище значення, або:

H0: Найнижче значення не є викидом;

H1: Найнижче значення є викидом;

якщо ми хочемо перевірити найменше значення.

Як і для будь-якого статистичного тесту, якщо p-значення менше вибраного порогу значущості (зазвичай α=0,05), тоді нульова гіпотеза відхиляється, і ми робимо висновок, що найнижче/найвище значення є викидом.

Навпаки, якщо p-значення більше або дорівнює рівню значущості, нульова гіпотеза не відхиляється, і ми робимо висновок, що на основі даних ми не відхиляємо гіпотезу про те, що найнижче/найвище значення не є викид.

Зауважте, що тест Граббса не підходить для розміру вибірки 6 або менше (n≤6).

Подібно до тесту Граббса, тест Діксона використовується для перевірки того, чи є окреме низьке чи високе значення викидом. Таким чином, якщо підозрюється більше ніж один викид, тест потрібно проводити окремо для цих підозрюваних викидів. Зауважте, що тест Діксона найбільш корисний для невеликої вибірки (зазвичай n≤25).

Якщо вам потрібно знову виконати тест без найвищого чи найнижчого значення, це можна зробити, знайшовши номер рядка максимального чи мінімального значення, виключивши цей номер рядка з набору даних і нарешті застосувавши тест Діксона до цього нового набору даних.

Тест Рознера на викиди має такі переваги, як:

* він використовується для виявлення кількох викидів одночасно (на відміну від тесту Ґраббса та Діксона, який потрібно виконувати ітераційно для відсіву кількох викидів), і
* його розроблено, щоб уникнути проблеми маскування, коли викид, близький за значенням до іншого викиду, може залишитися непоміченим.
* на відміну від тесту Діксона, зауважте, що тест Рознера є найбільш доцільним, коли розмір вибірки великий (n≥20).

Ви знайдете багато інших методів виявлення викидів:

1. у пакетах {outliers},
2. за допомогою функції lofactor() із пакету {DMwR}: Локальний фактор викиду (LOF) — це алгоритм, який використовується для визначення викидів шляхом порівняння локальної щільності точки з щільністю її сусідів,
3. outlierTest() із пакета {car} дає найбільш екстремальне спостереження на основі заданої моделі та дозволяє перевірити, чи є це викидом,
4. у пакеті {OutlierDetection} та
5. за допомогою функції aq.plot() із {mvoutlier } пакету.

# **Лабораторна робота: Аналіз та виявлення викидів у даних**

## **Хід роботи**

### 1. Завантаження та підготовка даних

Для аналізу використаємо набір даних mpg. Однак для досліджень можна використовувати інший датасет.

library(ggplot2)

library(dplyr)

library(outliers)

library(rstatix)

library(EnvStats)

# Завантаження даних

data(mpg)

df <- mpg

### 2. Описові статистики змінної hwy

summary(df$hwy)

range(df$hwy)

Аналізуючи мінімальне, максимальне значення та діапазон, можна зробити висновки щодо потенційних викидів.



### 3. Побудова гістограми

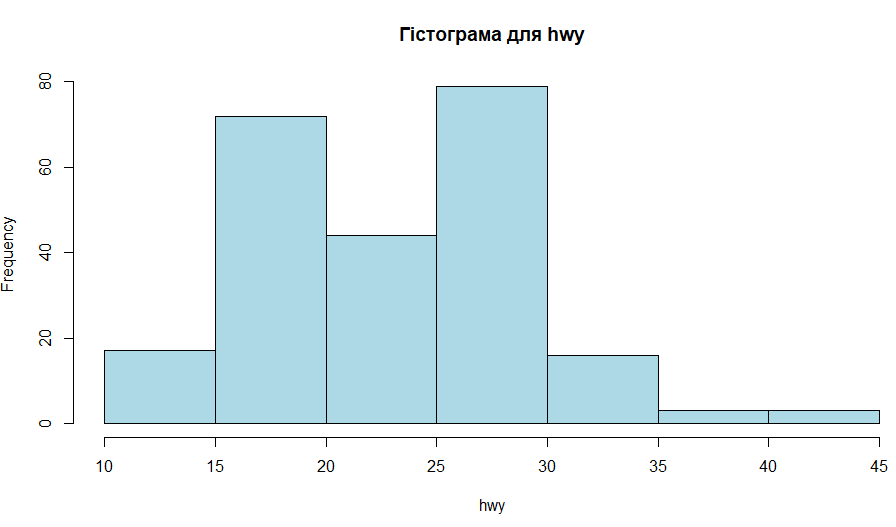
hist(df$hwy, main="Гістограма для hwy", xlab="hwy", col="lightblue", border="black")

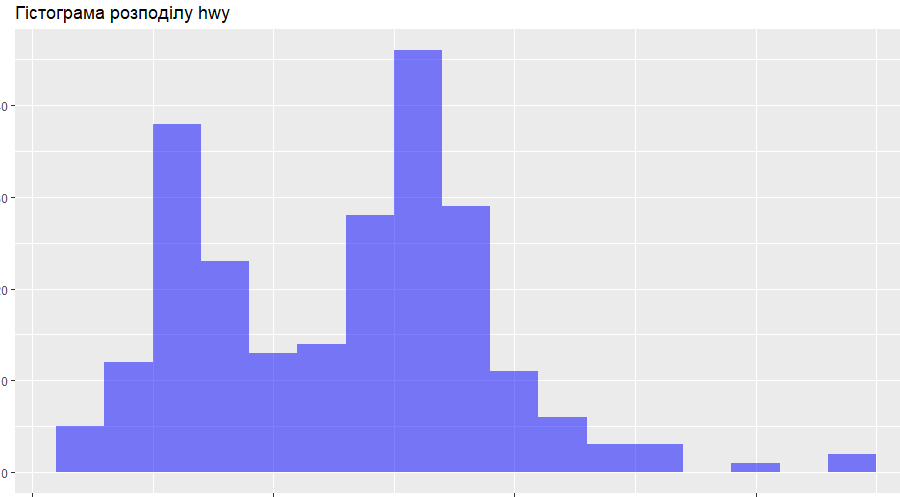
# Використання ggplot2

ggplot(df, aes(x=hwy)) +

geom\_histogram(binwidth=2, fill="blue", alpha=0.5) +

labs(title="Гістограма розподілу hwy")





### 4. Побудова boxplot діаграми

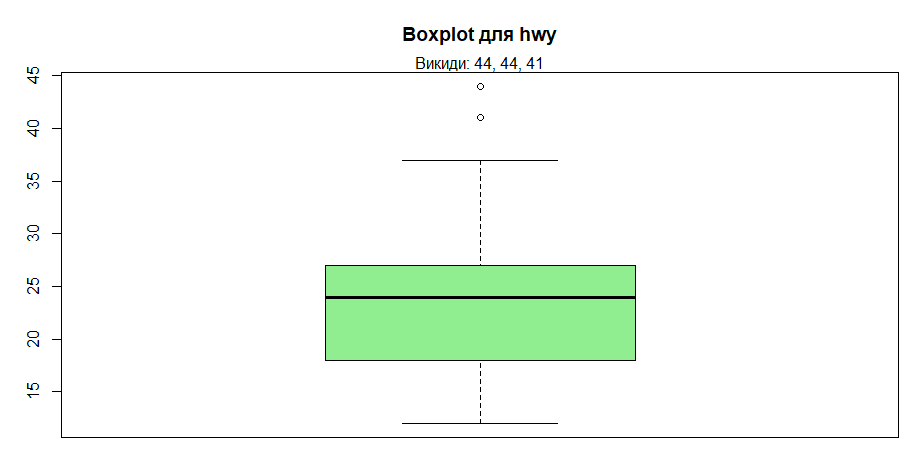
boxplot(df$hwy, main="Boxplot для hwy", col="lightgreen")

Використовуємо boxplot.stats() для виявлення викидів:

outliers <- boxplot.stats(df$hwy)$out

Outliers





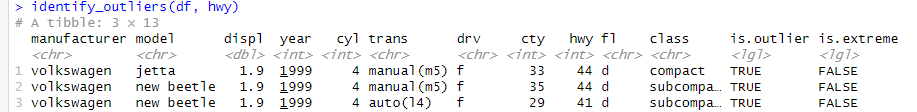
Знаходимо індекси викидів:

which(df$hwy %in% outliers)



### 5. Використання функції identify\_outliers()

identify\_outliers(df, hwy)



### 6. Додавання викидів на boxplot

mtext(paste("Викиди:", paste(outliers, collapse=", ")))

### 7. Визначення викидів за допомогою процентилів

Q1 <- quantile(df$hwy, 0.25)

Q3 <- quantile(df$hwy, 0.75)

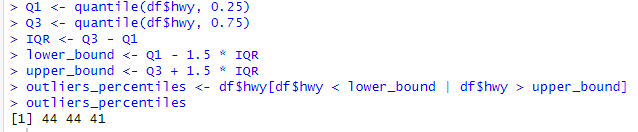
IQR <- Q3 - Q1

lower\_bound <- Q1 - 1.5 \* IQR

upper\_bound <- Q3 + 1.5 \* IQR

outliers\_percentiles <- df$hwy[df$hwy < lower\_bound | df$hwy > upper\_bound]

outliers\_percentiles

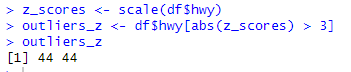


### 8. Нормалізація Z-score

z\_scores <- scale(df$hwy)

outliers\_z <- df$hwy[abs(z\_scores) > 3]

outliers\_z



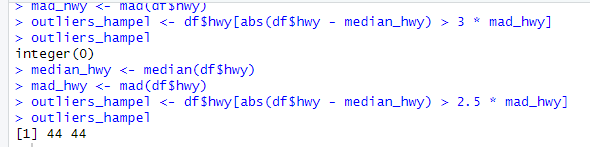
### 9. Метод Гампеля (Hampel filter)

median\_hwy <- median(df$hwy)

mad\_hwy <- mad(df$hwy)

outliers\_hampel <- df$hwy[abs(df$hwy - median\_hwy) > 3 \* mad\_hwy]

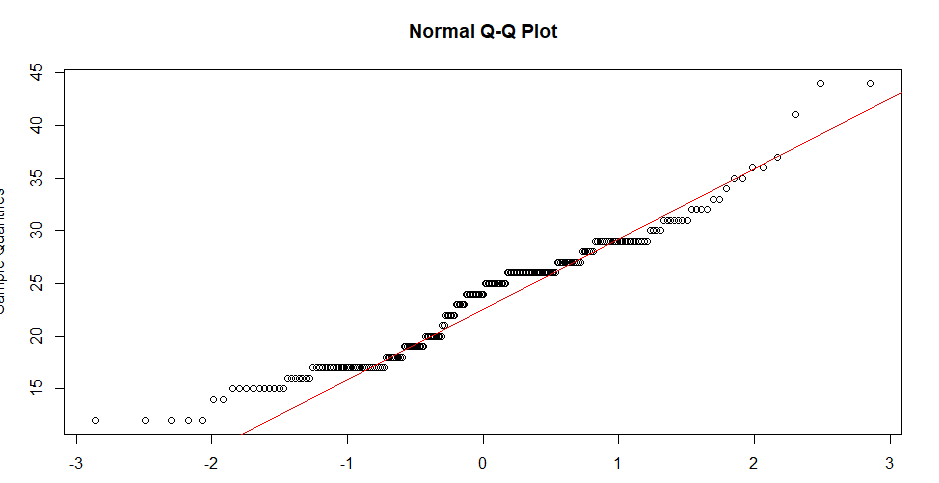
outliers\_hampel



### 10. Перевірка нормальності за допомогою QQ-plot

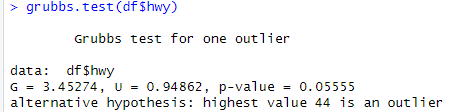
qqnorm(df$hwy)

qqline(df$hwy, col="red")



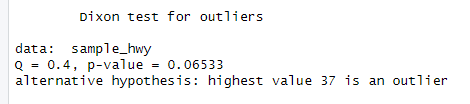
### 11. Grubbs’s test

grubbs.test(df$hwy)



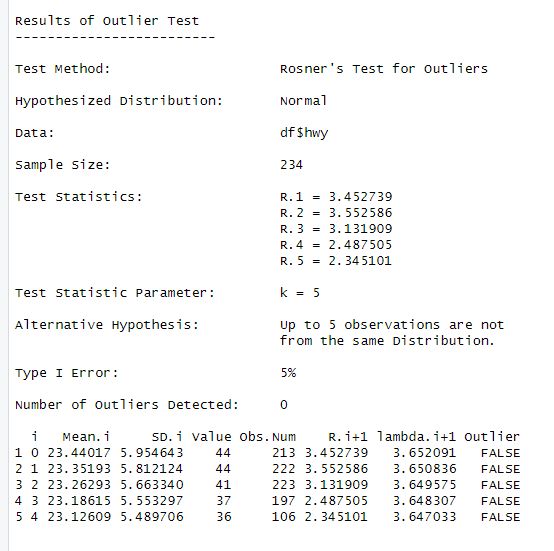
### 12. Dixon test

dixon.test(df$hwy)



### 13. Rosner’s test

rosnerTest(df$hwy, k=5)



ВИСНОВКИ:

На основі проведених тестів та візуалізацій були визначені можливі викиди в наборі даних. Було використано різні методи, зокрема boxplot, Z-score, процентильний метод та статистичні тести. Результати тестів слід аналізувати в комплексі, щоб отримати точні висновки про викиди.