ПВНЗ «Буковинський університет»

Факультет комп’ютерних систем і технологій

ЗВІТ

Про виконання лабораторної роботи №3

з дисципліни  “Інтелектуальний аналіз даних”

Виконав студент групи К-300  
Крячун Б.С.  
Прийняв викладач  
Артеменко О.І.

Лабораторна робота №3

Тема: Аналіз та візуалізація числових даних. Описова статистика. Графічний аналіз числових наборів даних.

Мета: Ознайомлення з основами візуального аналізу числових наборів даних, засобами їх побудови в середовищі RStudio, формування навичок виявлення закономірностей та викидів за допомогою графіків.

Використаний набір даних: lap\_times.csv (часи проходження кіл гонщиками у Формулі-1)

Посилання на GitHub: <https://github.com/kbs281103/IAD_KBS_K-300>

**Опис датасету**

Дані містять понад 580 000 записів, кожен з яких відповідає одному колу гонки, пройденому конкретним гонщиком. Основна змінна для аналізу:

milliseconds — час проходження кола у мілісекундах.

**Побудова графіків**

Boxplot

Графік відображає розподіл часу проходження кола. За допомогою цього графіка легко побачити викиди та діапазон основної частини даних.

Violin Plot

Комбінує функціональність boxplot і діаграми щільності. Дає змогу візуально оцінити форму розподілу.

Гістограма

Дає змогу побачити частотний розподіл часу проходження кола. Показує, які значення зустрічаються найчастіше.

Density Plot

Показує гладку оцінку розподілу. Дозволяє краще зрозуміти форму кривої розподілу та наявність мод (піків).

**Очищення даних**

Під час аналізу було виявлено значну кількість аномально високих значень у полі milliseconds, деякі з яких перевищували 7 000 000 мс (понад 2 години). Такі значення викликали спотворення графіків та не мають сенсу у контексті перегонів.

Тому було застосовано фільтрацію, яка залишила лише значення менше 200 000 мс (до 3 хвилин на коло). Це відповідає реалістичним гоночним даним.

Код, що виконує завдання повністю:

# Підключення бібліотек

library(ggplot2)

library(vioplot)

# Вибір CSV-файлу вручну

file\_path <- file.choose()

lap\_data <- read.csv(file\_path)

# Фільтрація: прибираємо аномалії > 200000 мс

lap\_data\_clean <- subset(lap\_data, milliseconds < 200000)

# Отримання шляху до папки

output\_dir <- dirname(file\_path)

# Побудова та збереження графіків

# Boxplot

boxplot\_plot <- ggplot(lap\_data\_clean, aes(y = milliseconds)) +

geom\_boxplot(fill = "lightblue") +

ggtitle("Boxplot часу проходження кола (мс) без викидів")

ggsave(file.path(output\_dir, "boxplot\_clean.png"), boxplot\_plot, width=6, height=4)

# Violin Plot

violin\_plot <- ggplot(lap\_data\_clean, aes(x = "", y = milliseconds)) +

geom\_violin(fill = "grey") +

geom\_boxplot(width = 0.1, alpha = 0.4) +

ggtitle("Violin Plot часу проходження кола (без викидів)")

ggsave(file.path(output\_dir, "violin\_plot\_clean.png"), violin\_plot, width=6, height=4)

# Гістограма

histogram\_plot <- ggplot(lap\_data\_clean, aes(x = milliseconds)) +

geom\_histogram(fill = "orange", color = "black", binwidth = 1000) +

ggtitle("Гістограма часу проходження кола (без викидів)")

ggsave(file.path(output\_dir, "histogram\_clean.png"), histogram\_plot, width=6, height=4)

# Density Plot

density\_plot <- ggplot(lap\_data\_clean, aes(x = milliseconds)) +

geom\_density(fill = "green", alpha = 0.4) +

ggtitle("Density Plot часу проходження кола (без викидів)")

ggsave(file.path(output\_dir, "density\_plot\_clean.png"), density\_plot, width=6, height=4)

Результати візуалізації після очищення

