МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЧЕРНІГІВСЬКА ПОЛІТЕХНІКА»
ННІ ЕЛЕКТРОННИХ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ІНЖЕНЕРІЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Методичні вказівки
до виконання навчально-дослідних завдань
з навчально-технологічної практики
для здобувачів *першого (бакалаврського*) рівня вищої освіти спеціальності 121
– **«Інженерія програмного забезпечення»**

Чернігів НУ «Чернігівська політехніка» 2021

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«ЧЕРНІГІВСЬКА ПОЛІТЕХНІКА»
ННІ ЕЛЕКТРОННИХ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ІНЖЕНЕРІЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Методичні вказівки
до виконання навчально-дослідних завдань
з навчально-технологічної практики
для здобувачів *першого (бакалаврського*) рівня вищої освіти спеціальності 121
– **«Інженерія програмного забезпечення»**

Обговорено і рекомендовано
на засіданні кафедри ІТтаПІ
Протокол №1
від 31.08.21

Чернігів НУ «Чернігівська політехніка» 2021

Інженерія програмного забезпечення. Методичні вказівки до виконання навчально-дослідних завдань з навчально-технологічної практики для здобувачів першого (бакалаврського) рівня вищої освіти спеціальності 121 – «Інженерія програмного забезпечення» / Укл. О.В. Трунова, М.М. Войцеховська – Чернігів: НУ «Чернігівська політехніка», 2021. – 42 с., укр. мовою.

Укладачі: Трунова Олена Василівна, к.пед.н., доцент кафедри інформаційних технологій та програмної інженерії

Войцеховська Марія Михайлівна, доктор філософії, старший викладач кафедри інформаційних технологій та програмної інженерії

Відповідальний за випуск: Білоус І.В., завідувач кафедри інформаційних технологій та програмної інженерії, к.т.н., доцент,

Рецензент: Ткач Юлія Миколаївна, завідувач кафедри кібербезпеки та математичного моделювання, д.пед.н., професор

Зміст

Вступ

1. Знайомство з середовищем розробки RStudio

2. Парсер даних

3. Статистичний аналіз і первинна обробка даних з використанням мови R

3.1 Обчислення заходів центральної тенденції та заходів мінливості

3.2 Побудова частотного розподілу для незгрупованих даних

3.3 Кореляційний аналіз Пірсона

3.4 Обчислення коефіцієнта рангової кореляції

3.5 Регресійний аналіз

3.6 Перевірка гіпотез про рівність середніх і про рівність дисперсій двох груп

3.7 Дисперсійний аналіз

3.8 Дискримінантний аналіз

3.9 Кластерний аналіз

Список використаних джерел

Вступ

Навчально-технологічна практика проводиться згідно «Положення про проведення практики здобувачів вищої освіти Національного університету «Чернігівська політехніка» на базі Навчально-тренувального центру з інформаційної безпеки, а також на кафедрі ІТіПІ НУ «Чернігівська політехніка» або в сторонніх організаціях (підприємствах, НДІ, фірмах). Навчально-технологічна практика проводиться в 4-му семестрі навчання для ЗВО спеціальності 121 – «Інженерія програмного забезпечення».

Тривалість проектно-технологічної практики – 2 тижні.

Метою навчально-технологічної практики є поглиблення і закріплення отриманих у вузі теоретичних і практичних знань і практичних умінь одержаних в процесі засвоєння наступних дисциплін: «Моделювання систем», «Об'єктно-орієнтоване програмування», «Бази даних», а також передумова до подальшого вивчення навчальної дисципліни «Емпіричні методи програмної інженерії». Адаптація до ринку праці за спеціальністю інженерія програмного забезпечення.

Зміст практики

Пропонується наступний орієнтовний план навчально-технологічної практики здобувачам вищої освіти

№ з/п	Назва теми	Кількість годин
1.	Установча конференція. Основи техніки безпеки та охорони праці.	2
2.	Робота за індивідуальним завданням згідно з тематикою та планом	
2.1	Знайомство з середовищем розробки RStudio	12
2.2	Статистичний аналіз і первинна обробка даних з використанням мови R	8
2.3.	Статистична перевірка гіпотез засобами мови R	8
2.4	Кореляційний та регресійний аналіз з використанням мови R	8

2.5	Повний факторний експеримент	12
2.6	Ортогональне центральне композиційне планування	12
3.	Написання звітів про практику.	26
4.	Захист звітів	2
Усього		**90**

Керівник практики дає здобувачу вищої освіти завдання розробити програму на певній мові програмування, що реалізує конкретний алгоритм. За бажанням здобувача вищої освіти завдання може бути підвищеної складності.

Дане видання призначене для виконання навчально-дослідних завдань спрямованих на практичне освоєння студентами середовища розробки RStudio та мови програмування R для аналізу даних.

Відповідно до графіка навчально-технологічної практики студенти перед виконанням завдань повинні ознайомитися з рекомендованою літературою. В методичному виданні містяться основні, базові теоретичні відомості, необхідні для виконання навчально-дослідних завдань. При цьому не варто обмежуватись лише наведеним списком.

Для одержання заліку з кожної роботи студент здає викладачу цілком оформлений звіт, а також демонструє на екрані комп’ютера результати виконання самостійної роботи.

Засобами оцінювання та методами демонстрації результатів навчання з навчально-технологічної практики є:

- залік;
- звіт та щоденник практики;
- презентація результатів виконаних завдань.

Звіт про практику включає наступні розділи:

- титульний аркуш;
- зміст;
- вступ;
- постановка завдання;
- вихідні дані;

- методи й засоби рішення завдання;
- результати рішення завдання;
- висновок;
- перелік посилань;
- додатки.

Розподіл балів, які отримують здобувачі вищої освіти

Вид роботи	Форма контролю	Кількість балів
1. Теоретична та практична частини	1.1 Відповідність умовам завдання	0 -10
	1.2. Відповідність вимогам стандартів	0 -10
2. Звіт з практики	2.1 Обґрунтованість технічних рішень	0 -15
	2.2 Посилання на першоджерела	0 -5
	2.3 Відповідність оформлення вимогам	0 -10
	2.4 Своєчасність виконання	0 -10
3. Захист звіту з практики	3.1 Самостійність виконання (відповіді на запитання)	0 -40
Разом		**100**

1. ЗНАЙОМСТВО З СЕРЕДОВИЩЕМ РОЗРОБКИ RSTUDIO

Мета. Ознайомитися з можливостями середовища розробки RStudio та мови програмування R.

Короткі теоретичні відомості.

RStudio – безкоштовне інтегроване середовище розробки (IDE) для R (рис. 1.1). Програмний продукт має відкритий вихідний код для мови програмування R, призначений для статистичної обробки даних і роботи з графікою, виконувати роботу з R зручною.

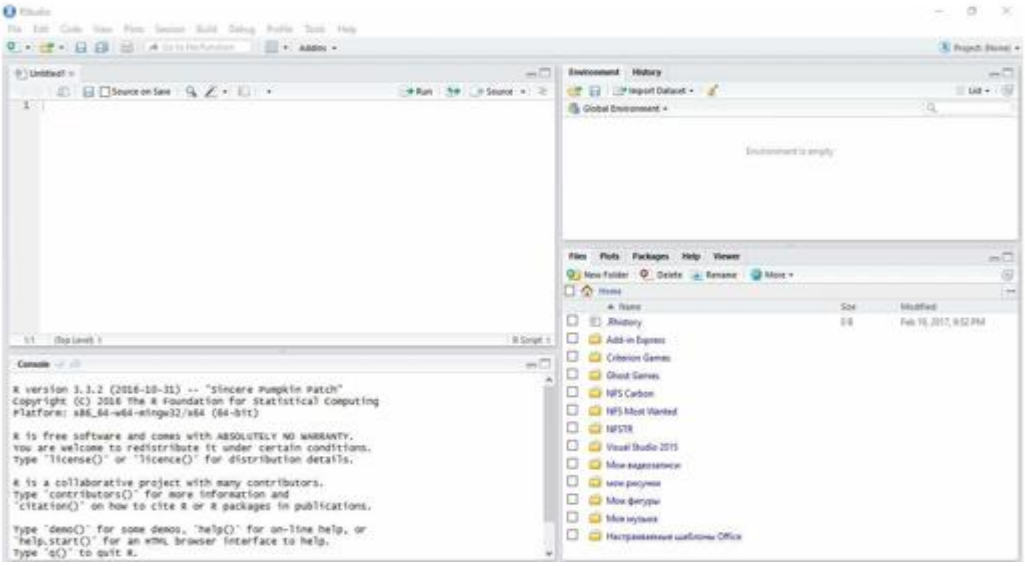


Рисунок 1.1 – Середовище розробки RStudio

Середовище розробки RStudio складається з наступних компонентів:

- панель меню;
- панель інструментів;
- консоль;
- запрошення (командний рядок);
- панель, що містить історію і робочий простір;
- панель з графіками [4].

Консоль RStudio представляє низку опцій, що полегшують роботу з мовою R. На кшталт, розглянемо автоматичне завершення коду: розглядаючи початок команди середовище пропонує користувачу продовження (рис. 1.2).

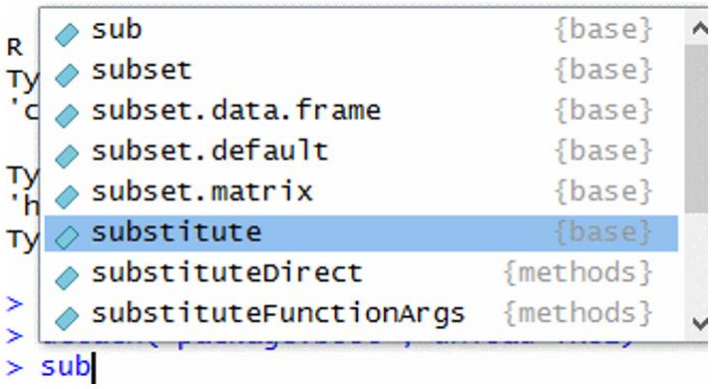


Рисунок 1.2 – Демонстрація автоматичного завершення коду в консолі

Повернення до попередніх команд виконується за допомогою комбінації клавіш Ctrl + ↑ та Ctrl + ↓, при цьому можна переходити до раніше викликаних команд (рис. 1.3).

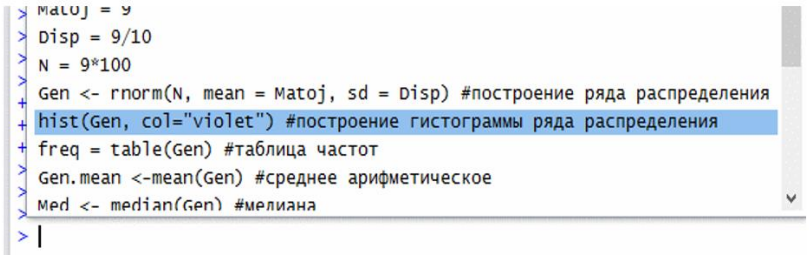


Рисунок 1.3 – Перехід до раніше викликаних команд в консолі

Розглянемо, так звані «Гарячі» клавіши. Їх можна переглянути викликавши HELP – R HELP, або обрати потрібний пункт меню Code, View або інше (рис. 1.4).

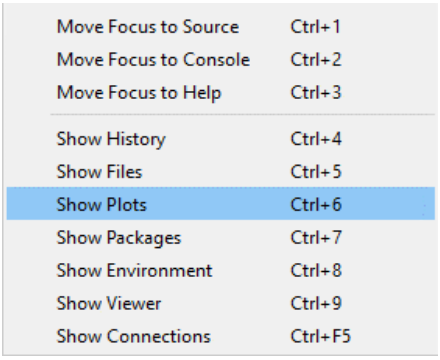


Рисунок 1.4 – Список «Гарячих» клавіш пункту меню View

Для створення нових скриптів необхідно обрати File – New File –> R Script (рис. 1.5). Якщо проект буде мати декілька R файлів, то в першу чергу New Project, а потім New File.

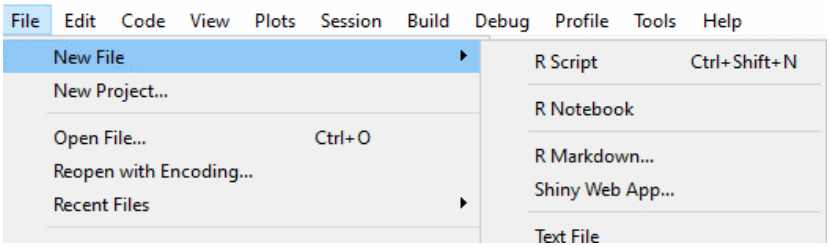


Рисунок 1.5 – Створення нового файлу

Щоб створити функцію треба виділити частину коду та обрати пункт Extract Function (рис. 1.6).

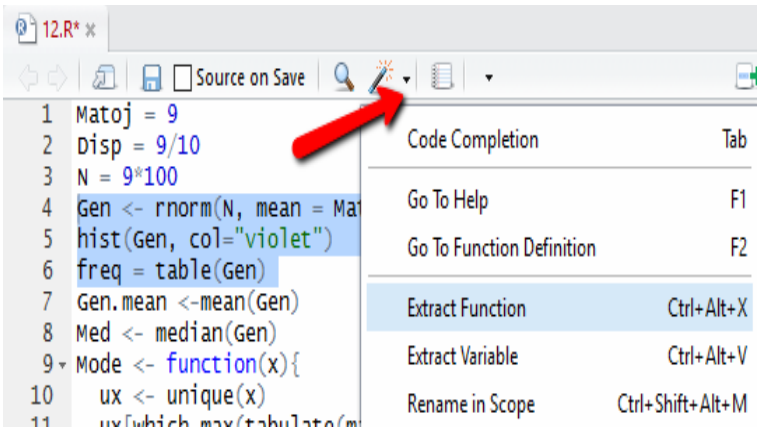


Рисунок 1.6 – Створення функції з виділеного коду

Щоб знайти та/або замінити частину тексту треба викликати вікно пошуку за допомогою меню Edit –> Find and Replace, або Ctrl + F (рис. 1.7).

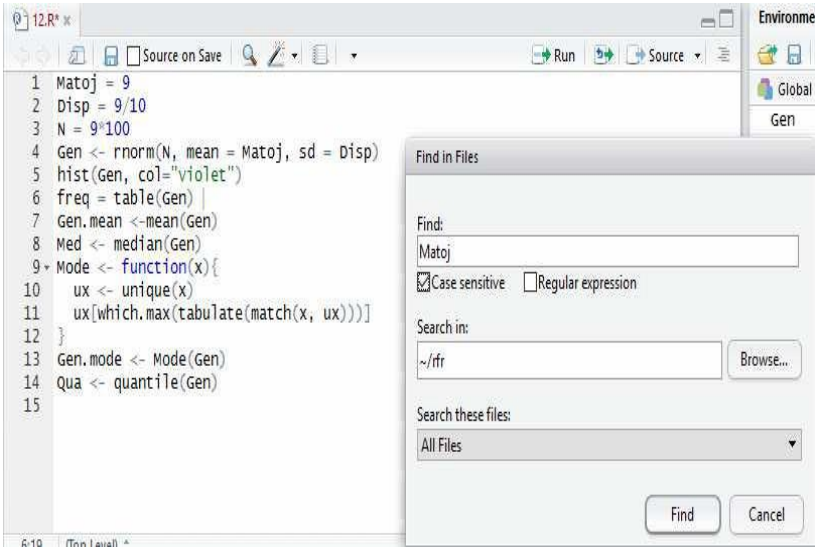


Рисунок 1.7 – Пошук частини коду в тексті

Коментування/зняття з фрагменту коду здійснюється таким чином – виділити фрагмент та натиснути пункт Comment/Uncomment Lines (рис. 1.8).

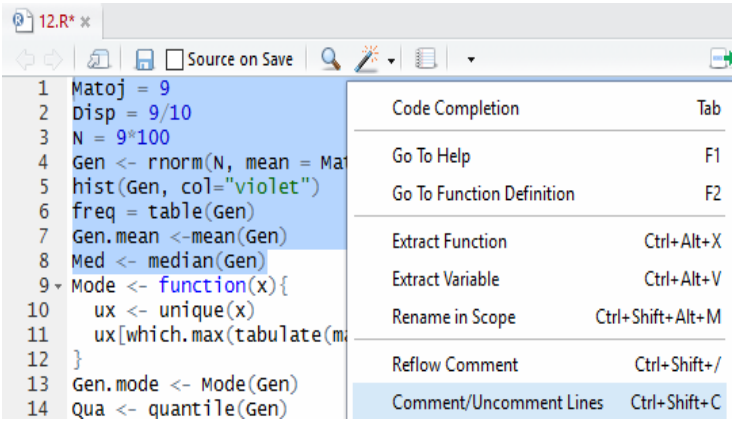


Рисунок 1.8 – Коментування коду

Для виконання коду – виділіть потрібну частину та натисніть Run. Для виконання поточного рядку – натисніть Ctrl + Enter. Після цього редактор автоматично перейде на наступний рядок. Для виконання усіх рядків – натисніть Ctrl + Shift + Enter.

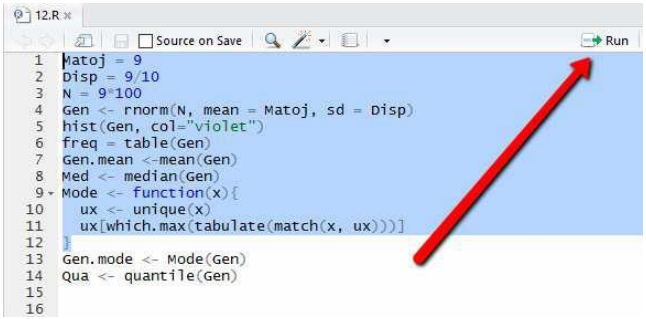


Рисунок 1.9 – Виконання коду

Завдання на самостійну роботу

Скориставшись рекомендованою літературою вивчити можливості середовища розробки RStudio та мови програмування R [4].

2. ПАРСЕР ДАНИХ

Мета. Ознайомитися з можливостями парсеру даних.

Короткі теоретичні відомості.

У теперішній час мережа Інтернет переповнена інформації, тож для її збору потрібні спеціальні програми, що виконують синтаксичний аналіз, – парсери. Синтаксичний аналіз (парсинг) – в інформатиці це процес аналізу вхідної послідовності символів, з метою розбору граматичної структури згідно із заданою формальною граматикою. Синтаксичний аналізатор – це програма або

частина програми, яка виконує синтаксичний аналіз.

Алгоритмом роботи парсерів є: отримання коду веб-сайту, обробка даних та представлення отриманих даних у потрібному користувачеві виді.

Для аналізу будемо брати з веб-ресурсу https://www.drupal.org/project/usage/cdn, який містить інформацію про використання CDN сайтами сервісу Drupal.

Структуру даних сайту зображено на рисунку 2.1.

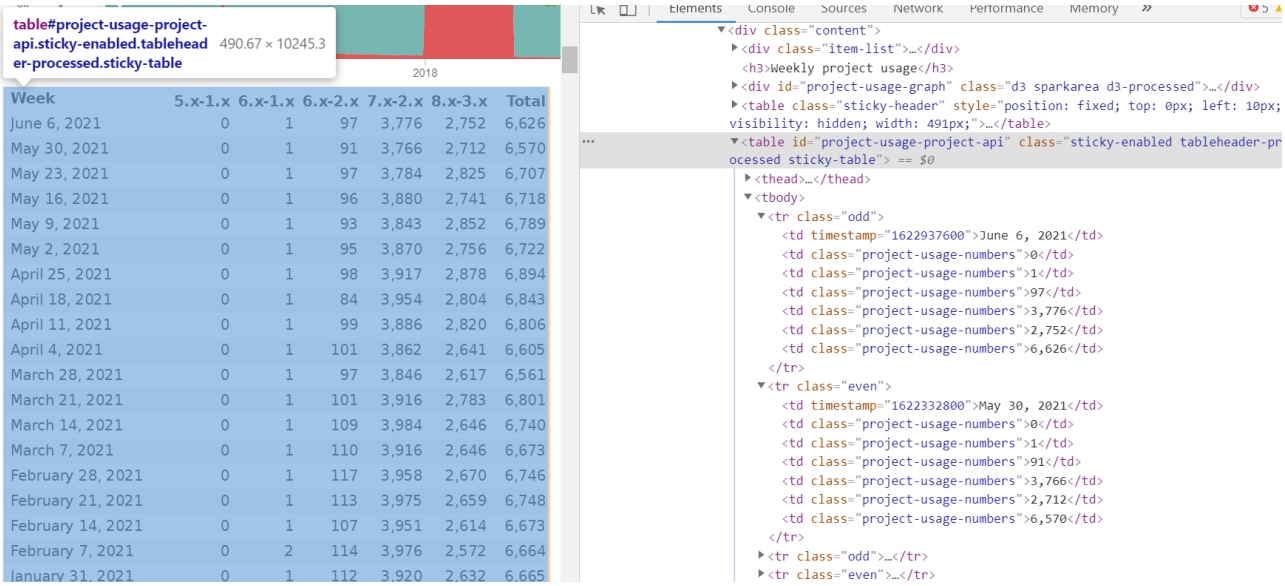


Рисунок 2.1 – Структура даних сайту

З рисунку 3 можна зробити висновки, що потрібні дані зберігаються в таблиці з ідентифікатором «project-usage-project-api».

Текст парсеру для отримання даних про використання CDN з коментарями наведено у лістингу 1.

Лістинг 1 – Текст парсеру для отримання даних про використання CDN з коментарями.

```
#Бібліотека для отримання даних веб-ресурсу
library(rvest)
#Бібліотека для роботи з excel файлами
library(xlsx)
#Бібліотека для роботи з xml файлами
library(xml2)
#Бібліотека для роботи з строками
library(stringr)
#Створення векторів для зберігання даних про використання CND
Number <- c()
```

```
Week <- c()
Version3 <- c()
```

```
Version4 <- c()
Version5 <- c()
Total <- c()
i <- 0
#Отримання коду веб-ресурсу
web <- read_html("https://www.drupal.org/project/usage/cdn")
#Знаходження у коді таблиці з інформацією
urlTable <- web %>% xml_find_all("//table[@id = 'project-usage-project-api']")
#Знаходження блоків, де записана потрібна інформація
urlTableRows <- urlTable %>% html_nodes('tr')
#Проходимо за допомогою циклу масив рядків таблиці з даними
for(row in urlTableRows[-c(1)]){
  #Інкремент змінної, що зберігає номер тижня, та запис у раніше створений вектор
  i = i + 1
  Number[i] <- i
  #Запис даних у рядку в змінну data
  data <- row %>% html_nodes('td')
  #Запис потрібних даних з рядка у створенні раніше вектори
  Week[i] <- data[1] %>% html_text()
  Version3[i] <- as.integer(str_trim(str_replace(data[4] %>% html_text(),",", "")))
  Version4[i] <- as.integer(str_trim(str_replace(data[5] %>% html_text(),",", "")))
  Version5[i] <- as.integer(str_trim(str_replace(data[6] %>% html_text(),",", "")))
  Total[i] <- as.integer(str_trim(str_replace(data[7] %>% html_text(),",", "")))
}
#Збереження отриманих даних у файл з типом xlsx
DataFrame <- data.frame(Number, Week, Version3, Version4, Version5, Total)
write.xlsx2(DataFrame, 'D:/R/statistics.xlsx', sheetName = "CDN", col.names = TRUE, row.names = FALSE, append = FALSE)
```

Частину отриманих парсером даних зображено на рисунку 2.2.

A	B	C	D	E	F
Number	Week	Version3	Version4	Version5	Total
1	June 6, 2021	97	3776	2752	6626
2	May 30, 2021	91	3766	2712	6570
3	May 23, 2021	97	3784	2825	6707
4	May 16, 2021	96	3880	2741	6718
5	May 9, 2021	93	3843	2852	6789
6	May 2, 2021	95	3870	2756	6722
7	April 25, 2021	98	3917	2878	6894
8	April 18, 2021	84	3954	2804	6843
9	April 11, 2021	99	3886	2820	6806
10	April 4, 2021	101	3862	2641	6605
11	March 28, 2021	97	3846	2617	6561
12	March 21, 2021	101	3916	2783	6801
13	March 14, 2021	109	3984	2646	6740
14	March 7, 2021	110	3916	2646	6673
15	February 28, 2021	117	3958	2670	6746
16	February 21, 2021	113	3975	2659	6748
17	February 14, 2021	107	3951	2614	6673
18	February 7, 2021	114	3976	2572	6664

Рисунок 2.2 – Частина отриманих даних

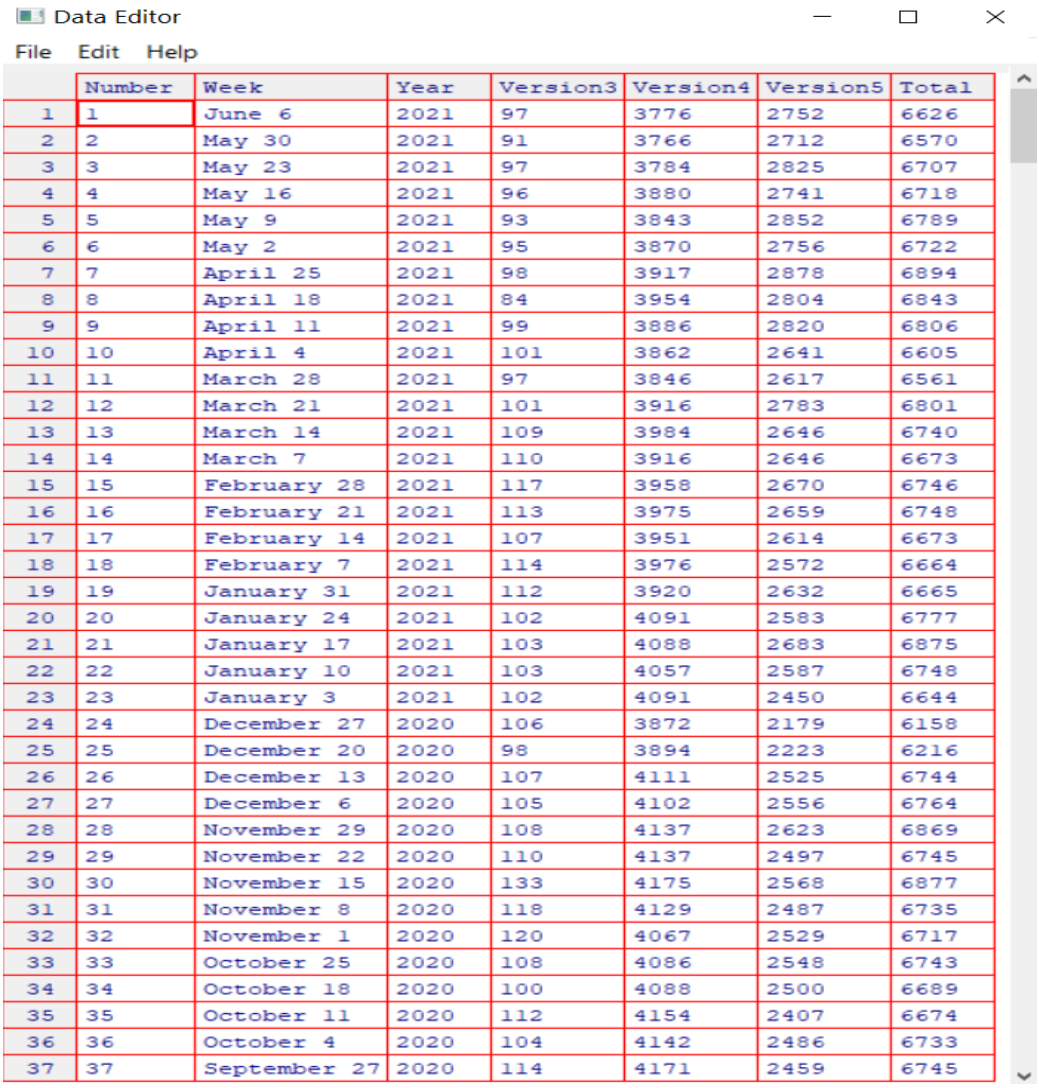
Для імпорту даних отриманих парсером використаємо функцію *read_excel*,

а для їх перегляду – *fix*, що наведено у лістингу 2.

Лістинг 2 – Імпорт та перегляд даних.

```
#Бібліотека для зчитування даних з excel файлів
library(readxl)
#Імпортування отриманих даних
data <- read_excel("D:/R/statistics.xlsx", sheet = "CDN")
#Перегляд вмісту таблиці у вбудованому редакторі
fix(data)
```

Результат перегляду даних у вбудованому редакторі зображено на рисунку 2.3.



	Number	Week	Year	Version3	Version4	Version5	Total
1	1	June 6	2021	97	3776	2752	6626
2	2	May 30	2021	91	3766	2712	6570
3	3	May 23	2021	97	3784	2825	6707
4	4	May 16	2021	96	3880	2741	6718
5	5	May 9	2021	93	3843	2852	6789
6	6	May 2	2021	95	3870	2756	6722
7	7	April 25	2021	98	3917	2878	6894
8	8	April 18	2021	84	3954	2804	6843
9	9	April 11	2021	99	3886	2820	6806
10	10	April 4	2021	101	3862	2641	6605
11	11	March 28	2021	97	3846	2617	6561
12	12	March 21	2021	101	3916	2783	6801
13	13	March 14	2021	109	3984	2646	6740
14	14	March 7	2021	110	3916	2646	6673
15	15	February 28	2021	117	3958	2670	6746
16	16	February 21	2021	113	3975	2659	6748
17	17	February 14	2021	107	3951	2614	6673
18	18	February 7	2021	114	3976	2572	6664
19	19	January 31	2021	112	3920	2632	6665
20	20	January 24	2021	102	4091	2583	6777
21	21	January 17	2021	103	4088	2683	6875
22	22	January 10	2021	103	4057	2587	6748
23	23	January 3	2021	102	4091	2450	6644
24	24	December 27	2020	106	3872	2179	6158
25	25	December 20	2020	98	3894	2223	6216
26	26	December 13	2020	107	4111	2525	6744
27	27	December 6	2020	105	4102	2556	6764
28	28	November 29	2020	108	4137	2623	6869
29	29	November 22	2020	110	4137	2497	6745
30	30	November 15	2020	133	4175	2568	6877
31	31	November 8	2020	118	4129	2487	6735
32	32	November 1	2020	120	4067	2529	6717
33	33	October 25	2020	108	4086	2548	6743
34	34	October 18	2020	100	4088	2500	6689
35	35	October 11	2020	112	4154	2407	6674
36	36	October 4	2020	104	4142	2486	6733
37	37	September 27	2020	114	4171	2459	6745

Рисунок 2.3 – Результат перегляду даних у вбудованому редакторі

Завдання на самостійну роботу

Скориставшись рекомендованою літературою здійснити парсер даних для подальшого дослідження.

3. СТАТИСТИЧНИЙ АНАЛІЗ І ПЕРВИННА ОБРОБКА ДАНИХ З ВИКОРИСТАННЯМ МОВИ R

Мета. Навчитися використовувати середовище розробки RStudio та мову програмування R для первинного аналізу даних. Оволодіти основами мови програмування R для здійснення аналізу даних.

Короткі теоретичні відомості.

Основні функції та команди R наведені у таблиці 3.1.

Таблиця 3.1

Перелік стандартних функцій та команд R

Назва Функції	**Опис**
mean()	Функція для знаходження середнього арифметичного, вона приймає параметр ряду розподілу.
sd()	Функція для пошуку стандартного відхилення, вона приймає параметр ряду розподілу.
rnorm()	Функція слугує для випадкової генерації нормально розподілених чисел. Вона має параметри: *N* – заданий розмір, *mean* – середнє значення, *sd* – стандартне відхилення.
hist()	Функція для створення гістограми. Вона може приймати ті ж параметри, що і у функції для створення графіків *plot()*. Одна з них: *col* – слугує для задання кольору стовпчиків або інших елементів графіків. Слугує оцінкою щільності відповідного розподілу.
table()	Функція визначає таблицю частот відповідних рівнів.
median()	Функція що повертає медіану, та приймає ті ж параметри, що й *mean()*.
which.max()	Функція знаходить порядкові номери елементів з максимальним значенням, а якщо елементів декілька, то буде повернуто номер першого такого елемента.
unique()	Функція повертає вектор, таблицю, або масив, але з однаковими елементами.
tabulate()	Функція бере ціле значення з вектору та підраховує кількість разів яке кожне ціле знаходиться в ньому.
match()	Функція повертає вектор тієї довжини, що і вектор з елементами у місці пошуку.
quantile()	Функція розраховує квантилі, приймає параметр ряду розподілу.
print()	Функція виводить на екран об’єкт, це функція загального призначення – конкретний результат її роботи буде залежати від класу об’єкта.
cat()	Функція більш розширена, ніж *print* і теж служить для виведення інформації на консоль.

Перевірка умов у мові *R* виконується таким чином:

```
if (умова){
    виконується, якщо умова правильна
} else {
    виконується, якщо умова не правильна
}
```

Циклічні вирази із заданою великою кількістю ітерацій виконуються за допомогою конструкції:

```
for (<змінна> in <вираз-1>)
<вираз-2>
```

Результатом <вираз-1> повинен бути вектор, а <змінна> на кожній ітерації циклу приймає значення чергового елемента цього вектора. Кількість ітерацій дорівнює кількості елементів у векторі.

Відповідно, у <вираз-2> може використовуватися <змінна>, яка буде змінною (лічильником) циклу:

```
sum = 0
for (i in 1:20)
  sum = sum + i;
sum
# результат виконання:
[1] 210
```

Аналогічно працює цикл while.

```
while (<умова>) <вираз>
```

Переривання циклу здійснюється командою break. Для переривання поточної ітерації і переходу до наступної служить команда next [4].

3.1 Обчислення заходів центральної тенденції та заходів мінливості

Міра центральної тенденції – це центральне або типове значення для розподілу ймовірностей. Її ще можна назвати центром чи місцем розподілу. Найбільш поширеними мірами центральної тенденції є середнє арифметичне,

медіана та мода. Тому для кожного стовпця визначимо показники опису центральних тенденцій: медіану, моду і середнє арифметичне значення вибірки [6].

Результат розрахунку середнього арифметичного значення, медіани та моди для кожного стовпця за допомогою функцій *mean*, *median* та *which.max*, зображено на рисунку 3.1.

```
> # Розрахунок середнього значення
> mean(version3)
[1] 1893.471
> mean(version4)
[1] 4444.251
> mean(version5)
[1] 760.952
> mean(total)
[1] 7102.148
> # Розрахунок медіани
> median(version3)
[1] 736.5
> median(version4)
[1] 4699
> median(version5)
[1] 144
> median(total)
[1] 6604
> # Розрахунок моди
> sort(unique(version3))[which.max(table(version3))]
[1] 141
> sort(unique(version4))[which.max(table(version4))]
[1] 4541
> sort(unique(version5))[which.max(table(version5))]
[1] 0
> sort(unique(total))[which.max(table(total))]
[1] 6673
```

Рисунок 3.1 – Результати розрахунку показників опису центральних тенденцій

Заходи мінливості – це статистичні показники, що характеризують відмінності між окремими значеннями вибірки. Вони дозволяють судити про ступінь однорідності отриманої вибірки та можливості середнього представляти весь набір даних. Найбільш використовуваними є дисперсія, стандартне відхилення, мінімальне та максимальне значення, розмах, коефіцієнти ексцесу та асиметрії.

Результат знаходження показників мінливості – дисперсії, стандартних відхилень, максимального і мінімального значень, коефіцієнтів ексцесу та асиметрії для кожного стовпця за допомогою функції *var*, *sd*, *range*, *kurtosis* та *skewness* зображено на рисунку 3.2.

```
> # Визначимо дисперсії для кожного стовпця
> var(version3)
[1] 10303763
> var(version4)
[1] 1937445
> var(version5)
[1] 917654.7
> var(total)
[1] 15079760
> #Максимальне і мінімальне значення
> range(version3)
[1]    84 11003
> range(version4)
[1] 1392.545 6437.000
> range(version5)
[1]    0 2878
> range(total)
[1]  2553 17566
> # Визначимо стандартні відхилення для кожного стовпчика
> sd(version3)
[1] 3209.948
> sd(version4)
[1] 1391.921
> sd(version5)
[1] 957.9429
> sd(total)
[1] 3883.267
> # Визначимо коефіцієнт ексцесу
> kurtosis(version3)
[1] 5.821069
> kurtosis(version4)
[1] 2.637175
> kurtosis(version5)
[1] 2.178245
> kurtosis(total)
[1] 4.936753
> # Визначимо коефіцієнт асиметрії
> skewness(version3)
[1] 2.148812
> skewness(version4)
[1] -0.8704316
> skewness(version5)
[1] 0.8659386
> skewness(total)
[1] 1.686331
```

Рисунок 3.2 – Результати розрахунку заходів мінливості

Розмах вибірки, який дає інформацію про ширину інтервалу, в якому зосереджений весь набір числових даних, для Version3, Version4, Version5 і Total дорівнює 10919, 5044.455, 2878 та 15013 відповідно.

Код побудови графіків розкиду даних кожного стовпця за допомогою функції *plot* зображено на рисунку 3.3.

```
> # Побудуємо графіки для кожного стовпця
> plot(version3, col = "red", main = "Version 3", xlab = "Weeks", ylab = "Usage")
> plot(version4, col = "blue", main = "Version 4", xlab = "Weeks", ylab = "Usage")
> plot(version5, col = "green", main = "Version 5", xlab = "Weeks", ylab = "Usage")
> plot(total, col = "purple", main = "Total", xlab = "Weeks", ylab = "Usage")
```

Рисунок 3.3 – Побудова графіків розкиду даних

Результати побудови графіків розкиду даних кожного стовпця за допомогою функції *plot* зображено на рисунку 3.4.

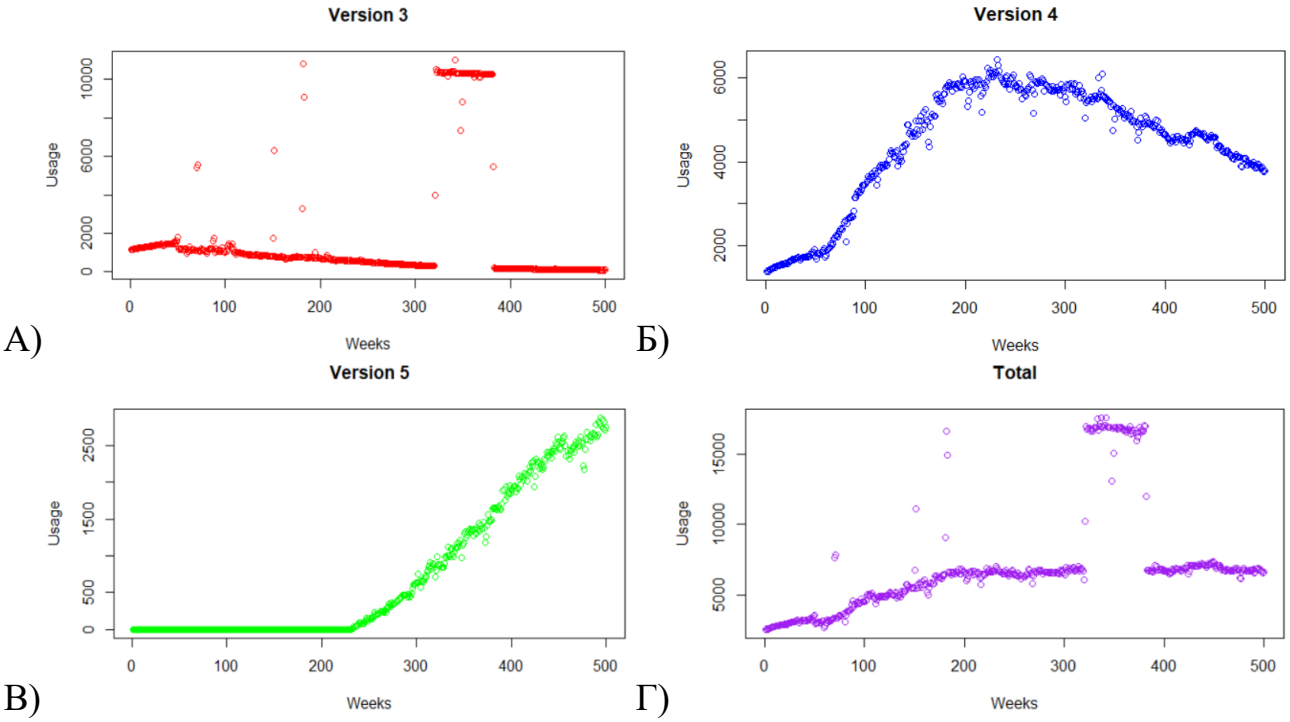


Рисунок 3.4 – Графіки розкиду даних

Графік А – графік розкиду даних використання CDN сайтів третьої версії Drupal, що зберігаються у стовпці Version3

Графік Б – графік розкиду даних використання CDN сайтів четвертої версії Drupal, що зберігаються у стовпці Version4

Графік В – графік розкиду даних використання CDN сайтів п’ятої версії Drupal, що зберігаються у стовпці Version5

Графік Г – графік розкиду даних використання CDN сайтів усіх версій

Drupal, що зберігаються у стовпці Total

Код побудови ще чотирьох графіків для зображення можливостей функції *plot* з використанням різних типів побудови: 1) гістограмні вертикальні лінії; 2) лінії; 3) точки з лініями; 4) точки поверх ліній; зображено на рисунку 3.5.

```
> # Побудуємо графіки для зображення можливостей
> plot(version3, type = "h", col = "red", main = "Version 3", xlab = "Weeks", ylab = "Usage")
> plot(version4, type = "l", col = "blue", main = "Version 4", xlab = "Weeks", ylab = "Usage")
> plot(version5, type = "b", col = "green", main = "Version 5", xlab = "Weeks", ylab = "Usage")
> plot(total, type = "o", col = "purple", main = "Total", xlab = "Weeks", ylab = "Usage")
```

Рисунок 3.5 – Побудова графіків розкиду даних

Результати побудови чотирьох графіків для зображення можливостей функції *plot* з використанням різних типів побудови зображено на рисунку 3.6.

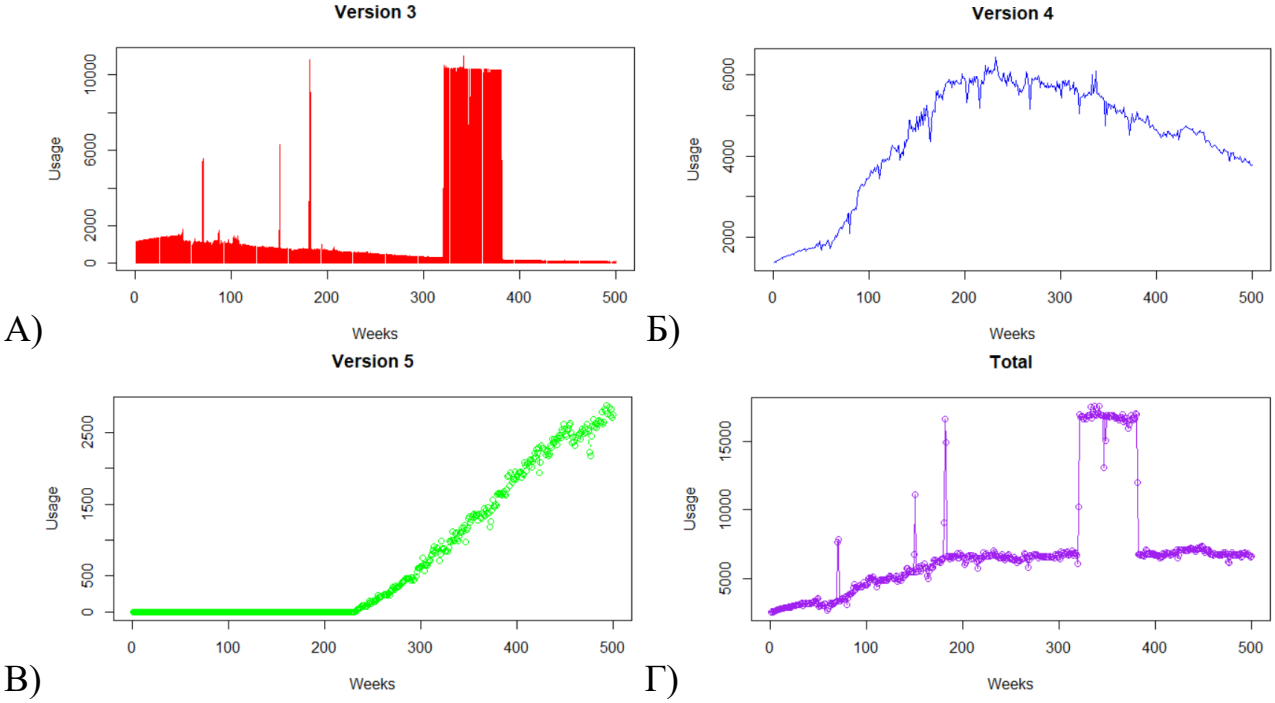


Рисунок 3.6 – Побудова графіків розкиду даних для зображення можливостей функції *plot*

Графік А – графік розкиду даних використання CDN сайтів третьої версії Drupal, що побудований з використанням гістограмних вертикальних ліній

Графік Б – графік розкиду даних використання CDN сайтів четвертої версії Drupal, що побудований з використанням простих ліній

Графік В – графік розкиду даних використання CDN сайтів п’ятої версії Drupal, що побудований з використанням точок з лініями

Графік Г – графік розкиду даних використання CDN сайтів усіх версій Drupal, що побудований з використанням точок поверх ліній

Завдання на самостійну роботу

Розрахувати міри центральної тенденції для знаходження типових значень вибірки. Визначити середнє арифметичне, що використовується для визначення середніх величин вибірки; медіану, моду. Міри мінливості, які дозволяють судити про ступінь однорідності отриманої вибірки та можливості середнього представляти весь набір даних. Визначити стандартну похибку, яка є одним з показників розкиду значень вибірки відносно її центру розподілу; дисперсію, яка дозволяє виміряти наскільки далеко значення вибірки розподілені від їх середнього значення; максимальне і мінімальне, за допомогою яких можна розрахувати розмах вибірки, який дає інформацію про ширину інтервалу, в якому зосереджений весь набір числових даних; коефіцієнт ексцесу, який використовується для попередньої перевірки на нормальність, та спростувати або підтвердити гіпотезу про нормальний розподіл, коефіцієнт асиметрії, щоб перевірити розподіл на симетричність.

3.2 Побудова частотного розподілу для незгрупованих даних

Частотний розподіл – метод статистичного опису даних (виміряних значень, характерних значень). Математично розподіл частот, є функцією, яка в першу чергу визначає для кожного показника ідеальне значення, так як ця величина зазвичай вже виміряна [6].

Результат структури об’єкта *data* за допомогою функції *str* зображено на рисунку 3.7.

```
> # Перегляд структури data
> str(data)
'data.frame':   500 obs. of  7 variables:
 $ Number  : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Week    : chr  "June 6" "May 30" "May 23" "May 16" ...
 $ Year    : num  2021 2021 2021 2021 2021 ...
 $ Version3: num  97 91 97 96 93 95 98 84 99 101 ...
 $ Version4: num  3776 3766 3784 3880 3843 ...
 $ Version5: num  2752 2712 2825 2741 2852 ...
 $ Total   : num  6626 6570 6707 6718 6789 ...
```

Рисунок 3.7 – Структура об’єкта *data*

У результаті бачимо, що кількість записів (рядків) таблиці – 500, та кількість стовпчиків таблиці – 7. Далі йде перелік стовпчиків, тип даних, які вони

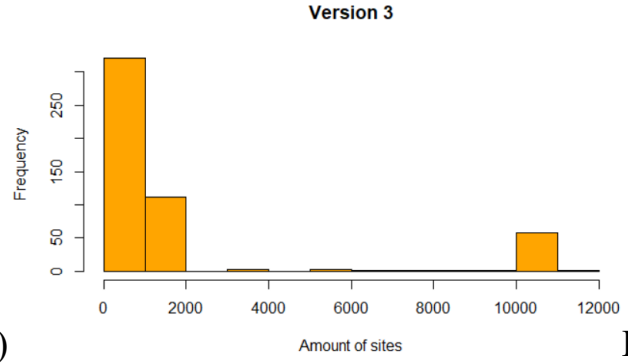
містять, та перелік цих самих даних.

Код побудови гістограм даних для числових стовпців за допомогою функції *hist* зображено на рисунку 3.8.

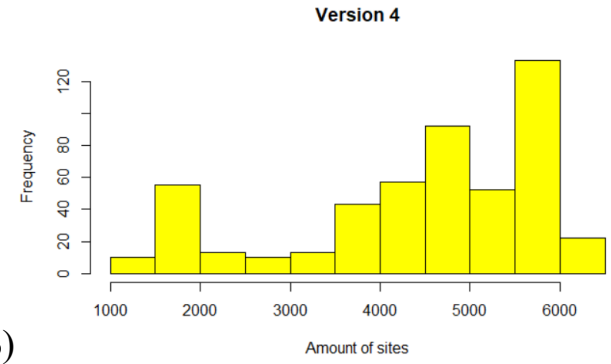
```
> # Побудова гістограм даних
> hist(version3, col = "orange", main = "Version 3", xlab = "Amount of sites", ylab = "Frequency")
> hist(version4, col = "yellow", main = "Version 4", xlab = "Amount of sites", ylab = "Frequency")
> hist(version5, col = "pink", main = "Version 5", xlab = "Amount of sites", ylab = "Frequency")
> hist(total, col = "magenta", main = "Total", xlab = "Amount of sites", ylab = "Frequency")
```

Рисунок 3.8 – Побудова гістограм даних

Результати побудови гістограм даних для числових стовпців зображено на рисунках 3.9 та 3.10.



А)



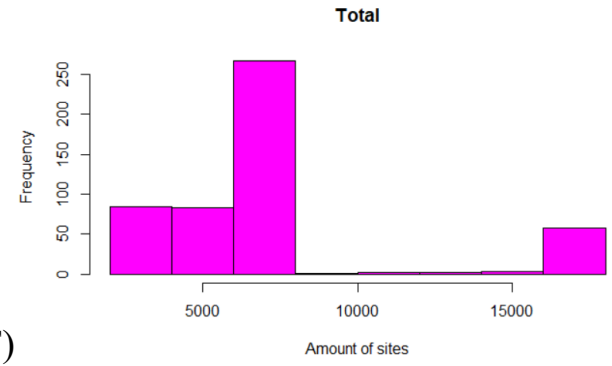
Б)

Рисунок 3.9 – Гістограми даних

Графік А – гістограма даних з використання CDN сайтів третьої версії

Графік Б – гістограма даних з використання CDN сайтів четвертої версії

В)



Г)

Рисунок 3.10 – Гістограми даних

Графік В – гістограма даних з використання CDN сайтів п’ятої версії

Графік Г – гістограма даних з використання CDN сайтів усіх версій

Результат розрахунку показників описової статистики для кожного стовпця за допомогою функції *summary* зображено на рисунку 3.11.

```
> # Розрахунок показників описової статистики для кожного стовпця
> summary(data$Version3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   84.0   322.8   736.5  1893.5  1217.1 11003.0
> summary(data$Version4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1393    3888    4699    4444    5655    6437
> summary(data$Version5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0     144     761    1472    2878
> summary(data$Total)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2553    5009    6604    7102    6835   17566
```

Рисунок 3.11 – Результати розрахунку показників описової статистики

Результат знаходження показників мінливості – дисперсії та стандартного відхилення для кожного стовпця за допомогою функцій *var* та *sd* зображено на рисунку 3.12.

```
> # Визначимо дисперсії для кожного стовпця
> var(data$Version3)
[1] 10303763
> var(data$Version4)
[1] 1937445
> var(data$Version5)
[1] 917654.7
> var(data$Total)
[1] 15079760
> # Визначимо стандартні відхилення для кожного стовпчика
> sd(data$Version3)
[1] 3209.948
> sd(data$Version4)
[1] 1391.921
> sd(data$Version5)
[1] 957.9429
> sd(data$Total)
[1] 3883.267
```

Рисунок 3.12 – Результати розрахунку дисперсії та стандартних відхилень

Результат розрахунку квантилів для кожного стовпця за допомогою функції *quantile* зображено на рисунку 3.13.

```
> # Визначимо квантилі для кожного стовпчика
> quantile(data$Version3)
       0%       25%       50%       75%      100%
   84.000   322.750   736.500  1217.114 11003.000
> quantile(data$Version4)
      0%      25%      50%      75%     100%
1392.545 3888.250 4699.000 5655.000 6437.000
> quantile(data$Version5)
  0%  25%  50%  75% 100%
   0    0  144 1472 2878
> quantile(data$Total)
      0%      25%      50%      75%     100%
 2553.00  5009.25  6604.00  6835.25 17566.00
```

Рисунок 3.13 – Результати розрахунку квантилів

Результат розрахунку 10-го та 90-го процентилів для кожного стовпця за допомогою функції *quantile* з аргументом *probs* – числовим вектор імовірностей – зображено на рисунку 3.14.

```
> # Визначимо 10 та 90 процентилі
> quantile(data$Version3, probs = c(0.1, 0.9))
    10%     90%
  138.0 10261.2
> quantile(data$Version4, probs = c(0.1, 0.9))
     10%      90%
1796.809 5859.200
> quantile(data$Version5, probs = c(0.1, 0.9))
   10%    90%
   0.0 2416.8
> quantile(data$Total, probs = c(0.1, 0.9))
    10%     90%
 3162.8 16637.2
```

Рисунок 3.14 – Результат розрахунку 10-го та 90-го процентилів

Завдання на самостійну роботу

Побудувати частотний розподіл для незгрупованих даних. Розрахувати такі показники описової статистики як: мінімальне та максимальне значення; середнє арифметичне для визначення середніх величин вибірки; медіану, моду квантилі, які відсікають в межах ряду певну частину його членів, у тому числі перший та третій квартилі, а також розрахували десятий та дев'яностий процентилі, які ділять всю вибірку на певні частини, виражені у процентах.

3.3 Кореляційний аналіз Пірсона

Коефіцієнт кореляції Пірсона — показник кореляції (лінійної залежності) між двома змінними, що набуває значень [-1; 1]. Він широко використовується в науці для вимірювання ступеня лінійної залежності між двома змінними [6].

Результат кореляційного аналізу для всієї матриці даних за допомогою функції *cor* зображено на рисунку 3.15.

```
> #Кореляційний аналіз таблиці
> cor(data[3:7])
              Year   Version3  Version4   Version5     Total
Year     1.0000000 0.12787484 0.5211214 0.89738735 0.5134229
Version3 0.1278748 1.00000000 0.1400789 0.05743908 0.8908602
Version4 0.5211214 0.14007888 1.0000000 0.12134558 0.5040759
Version5 0.8973873 0.05743908 0.1213456 1.00000000 0.3371883
Total    0.5134229 0.89086015 0.5040759 0.33718829 1.0000000
```

Рисунок 3.15 – Результат кореляційного аналізу для всієї матриці даних

Код побудови графіку, який ілюструє результати кореляційного аналізу, у вигляді еліпсів за допомогою функції *plotcorr* зображено на рисунку 3.16.

```
> #Відображення кореляційного аналізу
> plotcorr(cor(data[3:7]), type = "full")
```

Рисунок 3.16 – Побудова графіку, який ілюструє результати аналізу

Результати побудови графіку, який ілюструє результати кореляційного аналізу, зображено на рисунку 3.17.

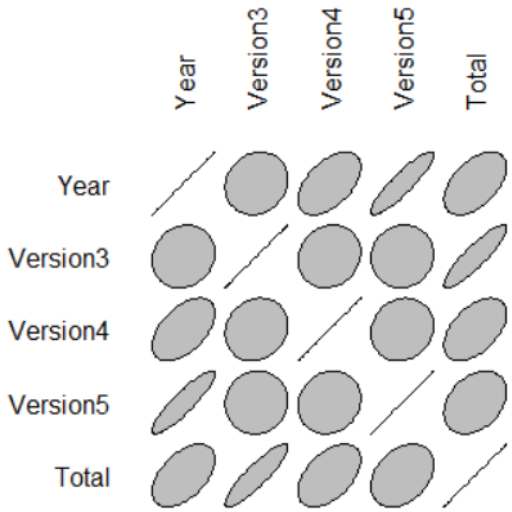


Рисунок 3.17 – Графік, який ілюструє результати кореляційного аналізу

Діаграма розсіювання – один із типів математичних діаграм. Дані показані у вигляді набору точок, кожен з яких має значення однієї змінної, тобто визначає її положення на горизонтальній осі та значення іншої змінної — її положення на вертикальній осі. На такій діаграмі можна простежити, як від значення незалежної змінної змінюється значення досліджуваної.

Код побудови діаграми розсіювання з незалежною змінною Version3 і досліджуваним параметром Total за допомогою функції *plot* зображено на рисунку 3.18.

```
> #Діаграма розсіювання
> plot(data$Version3, data$Total, main="Correlation", xlab="Version3 sites", ylab="Total", col="orange")
```

Рисунок 3.18 – Побудова діаграми розсіювання з незалежною змінною Version3 і досліджуваним параметром Total

Результати побудови діаграми розсіювання з незалежною змінною Version3 і досліджуваним параметром Total зображено на рисунку 3.19.

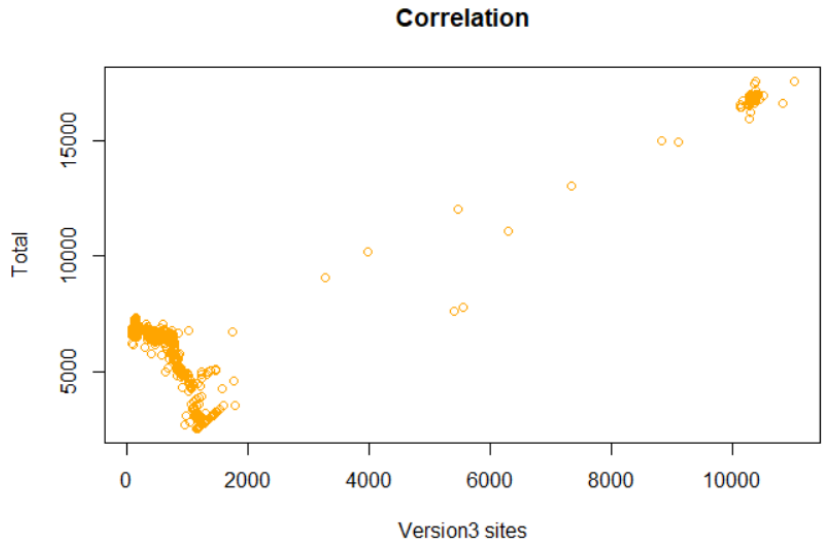


Рисунок 3.19 – Діаграми розсіювання зі змінними Version3 і Total

Результат обширного тесту кореляції між Version3 та Total з статистикою про t-критерій Стьюдента, степінь свободи, рівень значимості t-критерія, довірчі інтервали з надійністю 0.95 і значення кореляції за допомогою функції **cor.test** зображено на рисунку 3.20.

```
> #Кореляційний аналіз параметрів
> cor.test(data$Version3, dataset$Total)

        Pearson's product-moment correlation

data:  data$Version3 and dataset$Total
t = 43.763, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8712301 0.9076454
sample estimates:
      cor
0.8908602
```

Рисунок 3.20 – Результат тесту кореляції між Version3 та Total

Завдання на самостійну роботу

Знайшли коефіцієнт кореляції Пірсона для вимірювання ступеня лінійної залежності між двома змінними. Побудувати графік, який ілюструє результати кореляційного аналізу, та діаграму розсіювання, за допомогою якої можна простежити, як від значення незалежної змінної змінюється значення досліджуваного параметру. Визначити рівень значимості t-критерія Стьюдента та коефіцієнт кореляції, довірчий інтервал, до якого належать значення оцінюваної величини із довірчою імовірністю 0.95.

3.4 Обчислення коефіцієнта рангової кореляції

Рангова кореляція – метод кореляційного аналізу, який використовується для сукупностей невеликого обсягу і для кількісних ознак, якщо їхня сукупність не має нормального розподілу.

Коефіцієнт кореляції рангу Спірмена – непараметрична міра статистичної залежності між двома змінними, що оцінює наскільки добре можна описати відношення між двома змінними за допомогою монотонної функції. Якщо немає повторних значень, то коефіцієнт Спірмена дорівнює 1 чи -1, це відбувається коли кожна змінна є монотонною функцією від іншої.

Результат побудови кореляційної матриці методом Спірмена за допомогою функції *cor* зображено на рисунку 3.21.

```
> #Кореляційний аналіз Спірмена
> cor(data[3:7], method = "spearman")
              Year   Version3   Version4   Version5      Total
Year      1.0000000 -0.6112720  0.3426245  0.9447423  0.8291677
Version3 -0.6112720  1.0000000 -0.1786202 -0.5658967 -0.2094038
Version4  0.3426245 -0.1786202  1.0000000  0.1218546  0.5375448
Version5  0.9447423 -0.5658967  0.1218546  1.0000000  0.7480956
Total     0.8291677 -0.2094038  0.5375448  0.7480956  1.0000000
```

Рисунок 3.21 – Результат побудови кореляційної матриці методом Спірмена

За методом Спірмена найбільший вплив на Total здійснює Year (82.92%).

Результат обширного тесту кореляції між Year та Total з статистикою про значення оціночної статистики, рівень значимості, значення кореляції за Спірменом за допомогою функції *cor.test* зображено на рисунку 3.22.

```
> #Кореляційний аналіз Спірмена
> cor.test(data$Year, data$Total, method = "spearman", exact = FALSE)

        Spearman's rank correlation rho

data:  data$Year and data$Total
S = 3558991, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8291677
```

Рисунок 3.22 – Результат тесту кореляції між Year та Total

У статистиці коефіцієнт кореляції рангу Кендала, як правило, називають τ коефіцієнт (тау-коефіцієнт) Кендала. Він використовується у статистиці для вимірювання зв'язку між двома величинами. τ-тест – це непараметричний тест статистичних гіпотез залежності на основі -коефіцієнта. Зокрема, він є мірою рангової кореляції, тобто подібності упорядкування даних, коли вони

упорядкуванні за своєю величиною.

Результат побудови кореляційної матриці методом Кендала за допомогою функції *cor* зображено на рисунку 3.23.

```
> #Кореляційний аналіз Кендала
> cor(data[3:7], method = "kendall")
              Year    Version3    Version4    Version5      Total
Year      1.0000000 -0.57951832  0.18028717  0.86754069  0.6750003
Version3 -0.5795183  1.00000000 -0.08382864 -0.49964633 -0.2684340
Version4  0.1802872 -0.08382864  1.00000000 -0.02226041  0.4367455
Version5  0.8675407 -0.49964633 -0.02226041  1.00000000  0.5480572
Total     0.6750003 -0.26843405  0.43674548  0.54805718  1.0000000
```

Рисунок 3.23 – Результат побудови кореляційної матриці методом Кендала

За даним тестом параметр Total проявляє найбільшу залежність від незалежної змінної Year (67.5%).

Результат обширного тесту кореляції між Year та Total з статистикою про емпіричне значення рівня значимості, рівень значимості і значення кореляції за Кендалом за допомогою функції *cor.test* зображено на рисунку 3.24 .

```
> #Кореляційний аналіз Кендала
> cor.test(data$Year, data$Total, method = "kendall")

        Kendall's rank correlation tau

data:  data$Year and data$Total
z = 21.533, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.6750003
```

Рисунок 3.24 – Результат тесту кореляції між Year та Total

Завдання на самостійну роботу

Знайти коефіцієнт кореляції рангу Спірмена, що оцінює наскільки добре можна описати відношення між двома змінними за допомогою монотонної функції.

3.5 Регресійний аналіз

Регресійний аналіз – розділ математичної статистики, присвячений методам аналізу залежності однієї величини від іншої. На відміну від кореляційного аналізу, не з'ясовує чи істотний зв'язок, а займається пошуком його моделі, вираженої у функції регресії. Регресійний аналіз використовується тоді, коли відношення між змінними можуть бути виражені кількісно у виді

деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, що приймає цільова змінна. У простому випадку використовуються стандартні статистичні методи, як лінійна регресія [6].

Із попередніх досліджень наших було з'ясовано, що на значення Total найбільше впливають незалежні змінні Version3 та Year. Результати створення лінійних регресійних моделей для залежності Total-Version3 і Total-Year та їх дослідження функціями *lm* та *summary* відповідно зображено на рисунку 3.25.

```
> # Гіпотеза про зв'язок лінійного характеру
> lin_model = lm(formula = Total ~ Version3, data=data)
> summary(lin_model)

Call:
lm(formula = Total ~ Version3, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -3759  -1077    714   1344   2142

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.062e+03  9.171e+01   55.19   <2e-16 ***
Version3    1.078e+00  2.463e-02   43.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1766 on 498 degrees of freedom
Multiple R-squared:  0.7936,    Adjusted R-squared:  0.7932
F-statistic:  1915 on 1 and 498 DF,  p-value: < 2.2e-16
> lin_model = lm(formula = Total ~ Year, data=data)
> summary(lin_model)

Call:
lm(formula = Total ~ Year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4005.5 -1632.5  -758.8   -23.8 10362.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.434e+06  1.079e+05  -13.29   <2e-16 ***
Year         7.146e+02  5.352e+01   13.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3336 on 498 degrees of freedom
Multiple R-squared:  0.2636,    Adjusted R-squared:  0.2621
F-statistic: 178.3 on 1 and 498 DF,  p-value: < 2.2e-16
```

Рисунок 3.25 – Результати дослідження лінійних регресійних моделі

Результат прогнозу значень функцією *predict* зображено на рисунку 3.26.

```
> # Передбачення значень
> pred.data = data.frame(Version3 = c(500, 1000, 1200))
> predict(lin_model, newdata = pred.data, interval = "confidence")
       fit      lwr      upr
1 5600.367 5431.193 5769.541
2 6139.230 5978.162 6300.298
3 6354.776 6196.031 6513.520
> pred.data = data.frame(Year = c(2022, 2023, 2024))
> predict(lin_model2, newdata = pred.data, interval = "confidence")
       fit      lwr      upr
1 11281.08 10599.86 11962.30
2 11995.67 11218.21 12773.14
3 12710.26 11834.51 13586.02
```

Рисунок 3.26 – Результат прогнозу значень Total

Побудова графіків з регресійною лінією, довірчими інтервалами та інтервалами передбачення лінійної регресійної моделі з незалежною змінною Version3 та досліджуваним параметром Total зображено на рисунках 3.27.

```
> # Створюємо лінійну модель
> lin_model = lm(formula = Total ~ Version3, data = data)
> # Додаємо передбачення
> pred.int <- predict(lin_model, interval = "prediction")
> mydata <- cbind(data, pred.int)
> # Додаємо синю регресійну лінію та сірі довірчі інтервали
> p <- ggplot(mydata, aes(Version3, Total)) + geom_point() + stat_smooth(method = lm)
> # Додаємо червоні інтервали передбачення
> p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
+   geom_line(aes(y = upr), color = "red", linetype = "dashed")
> lin_model = lm(formula = Total ~ Year, data = data)
> # Додаємо передбачення
> pred.int <- predict(lin_model, interval = "prediction")
> mydata <- cbind(data, pred.int)
> # Додаємо синю регресійну лінію та сірі довірчі інтервали
> p <- ggplot(mydata, aes(Year, Total)) + geom_point() + stat_smooth(method = lm)
> # Додаємо червоні інтервали передбачення
> p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
+   geom_line(aes(y = upr), color = "red", linetype = "dashed")
```

Рисунок 3.27 – Побудова графіків моделей Total-Version3 і Total-Year

Результат побудови графіку лінійних регресійних моделей з незалежними змінними Version3 та Year і досліджуваним параметром Total зображено на рисунках 3.28 і 3.29.

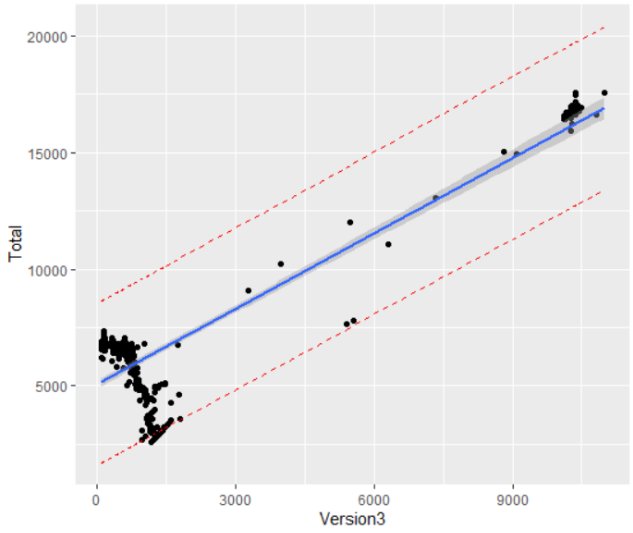


Рисунок 3.28 – Результат побудови лінійної регресійної моделі Total-Version3

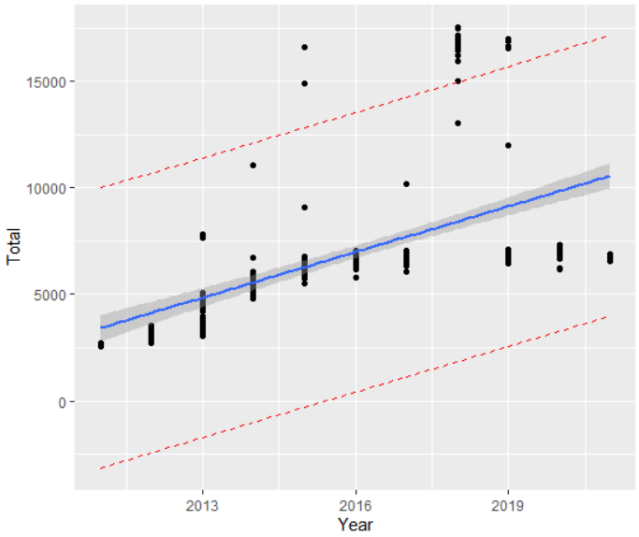


Рисунок 3.29 – Результат побудови лінійної регресійної моделі Total-Year

Завдання на самостійну роботу

Провести регресійний аналіз, що займається пошуком моделі зв’язку, вираженої у функції регресії. Створити лінійні регресійні моделі для залежностей. З’ясувати коефіцієнти регресійного аналізу, що наведені під заголовком *Coefficients*, де наведені оцінки коефіцієнтів побудованих моделей. За допомогою F-критерію перевірити гіпотезу про відсутність зв'язку між залежною змінною і предиктором. Виходячи з коефіцієнта детермінації R2, статистичної міри того, наскільки дані близькі до встановленої лінії регресії, можемо переконатися у точності моделей.

3.6 Перевірка гіпотез про рівність середніх і про рівність дисперсій двох груп

Для перевірки гіпотези про рівність середніх двох груп найчастіше застосовують так званий t-критерій Стьюдента, а для перевірки гіпотези про рівність дисперсій двох груп – F-тест (критерій Фішера F). Основною вимогою для використання даних критеріїв є нормальний розподіл кожної з двох. У мові програмування R існує цілий ряд статистичних тестів, спеціально розроблених для перевірки нормальності розподілу даних. У загальному вигляді нульова гіпотеза, що перевіряється за допомогою цих тестів, може бути сформульована наступним чином: «Вибірка, що аналізується, виходить з генеральної сукупності,

що має нормальний розподіл». Якщо значення p-value того чи іншого тесту є меншим за деякий, попередньо прийнятий рівень значимості, то нульова гіпотеза відхиляється [6].

Результат перевірки нормальності розподілу даних за критерієм хі-квадрат за допомогою функції *pearson.test()* зображено на рисунку 3.30.

```
> # Перевірка нормальності розподілу
> pearson.test(Version3)

        Pearson chi-square normality test

data:  Version3
P = 2336.2, p-value < 2.2e-16

> pearson.test(Version4)

        Pearson chi-square normality test

data:  Version4
P = 238.83, p-value < 2.2e-16

> pearson.test(Version5)

        Pearson chi-square normality test

data:  Version5
P = 1696.9, p-value < 2.2e-16

> pearson.test(Total)

        Pearson chi-square normality test

data:  Total
P = 1544.7, p-value < 2.2e-16
```

Рисунок 3.30 – Результат перевірки нормальності розподілу

Завдання на самостійну роботу

Провести перевірку нормальності, знайшти статистичний критерій та значення p-value.

3.7 Дисперсійний аналіз

Дисперсійний аналіз являє собою статистичний метод аналізу результатів, які залежать від якісних ознак. Кожен фактор може бути дискретною чи неперервною випадковою змінною, яку розділяють на декілька сталих рівнів (градацій, інтервалів). У дисперсійному аналізі перевірка статистичної значущості відмінності між середніми декількох груп здійснюється на основі вибіркових дисперсій. Ця перевірка проводиться за допомогою розбиття загальної дисперсії (варіації) на частини, одна з яких обумовлена випадковою помилкою (тобто внутрішньогруповою мінливістю), а друга пов'язана з відмінністю середніх значень. Якщо ця відмінність значуща, нульова гіпотеза щодо існування відмінності між середніми значеннями відкидається на певному рівні значущості.

Результат двохфакторного дисперсійного аналізу залежності Total від Year і Version3 за допомогою функцій *summary* і *aov* зображено на рисунку 3.31.

```
> #Дисперсійний аналіз
> summary(aov(Total ~ Year * Version3, data = data))
               Df    Sum Sq   Mean Sq F value   Pr(>F)
Year            1 1.984e+09 1.984e+09 3234.32  < 2e-16 ***
Version3        1 5.209e+09 5.209e+09 8494.11  < 2e-16 ***
Year:Version3   1 2.774e+07 2.774e+07   45.23 4.83e-11 ***
Residuals     496 3.042e+08 6.133e+05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рисунок 3.31 – Результат двохфакторного дисперсійного аналізу залежності Total від Year і Version3

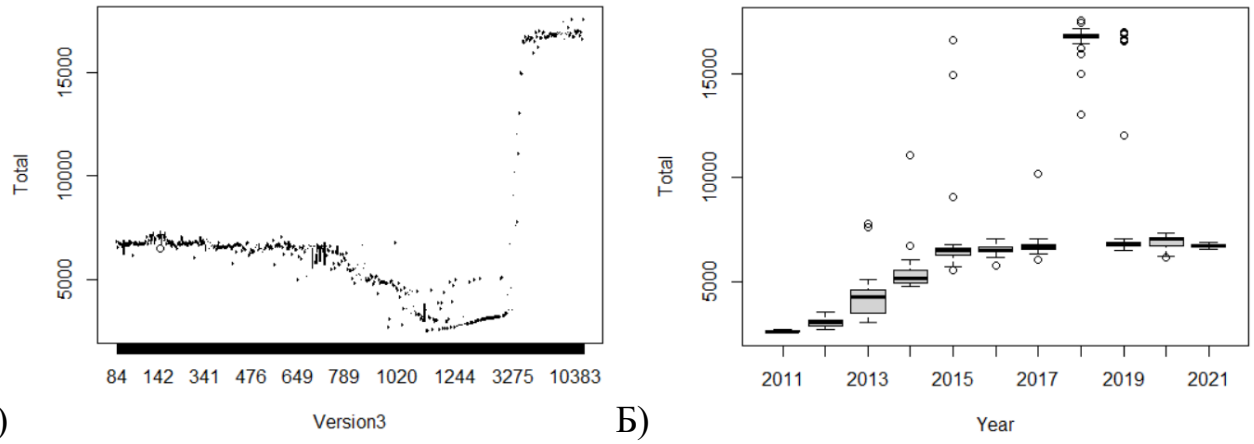
З результату можна зробити висновок про присутність статистично значущого зв'язку Total-Year і Total-Version3, а також між Year і Version3, бо значення *Pr(>F)* є суттєво меншим за рівень значимості 0.05.

Побудова графіків з отриманими даними про зв’язки між Year, Version3 і Total за допомогою функції *boxplot* зображено на рисунку 3.44.

```
> boxplot(Total ~ Year, data = data, ylab="Total", xlab="Year")
> boxplot(Total ~ Version3, data = data, ylab="Total", xlab="Version3")
```

Рисунок 3.32 – Побудова графіків з отриманими даними

Результат побудови графіків з отриманими даними про зв’язки між Year, Version3 і Total за допомогою функції *boxplot* зображено на рисунку 3.33.



А) Б)

Рисунок 3.33 – Графіки отриманих даних про зв’язки між Year, Version3 і Total

Графік А – діаграма розмаху даних про зв’язок між Version3 і Total

Графік Б – діаграма розмаху даних про зв’язок між Year і Total

Завдання на самостійну роботу

Провели двохфакторний дисперсійний аналіз, що досліджує вплив факторів на мінливість середніх значень вибірки Дізнатися розкид даних і дисперсію всередині груп та між ними.

3.8 Дискримінантний аналіз

Дискримінантний аналіз — різновид багатовимірного аналізу, призначеного для вирішення задач розпізнавання образів. Використовується для прийняття рішення про те, які змінні розділюють (тобто «дискримінують») певні масиви даних (так звані «групи»).

Результат формування груп з Year та побудови моделі дискримінантного аналізу з предикаторами Version3, Version4 і Version5, на основі яких формується група, за допомогою функції *lda***,** яка створює передбачені групи для кожного тестованого значення, зображено на рисунках 3.34 та 3.35.

```
> #Дискримінантний аналіз
> fit <- lda(Year ~ Version3 + Version4 + Version5, data = data,na.action="na.omit", CV=TRUE)
> fit
$class
  [1] 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 2020 2021
 [22] 2021 2020 2020 2020 2020 2020 2021 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020
 [43] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020
 [64] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 2019 2019 2020 2020 2020 2020 2020 2019 2019
 [85] 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019
[106] 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 2018 2018 2018 2018 2018 2018 2018 2018
[127] 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018
[148] 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018
[169] 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017
[190] 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017
[211] 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 2016 2016 2017 2016
[232] 2016 2015 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016
[253] 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016
[274] 2016 2016 2016 2016 2016 2016 2016 2015 2015 2015 2015 2015 2015 2015 2015 2016 2016 2015 2016 2015 2016 2015
[295] 2015 2016 2015 2015 2015 2015 2016 2016 2016 2016 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015
[316] 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 2014
[337] 2014 2014 2014 2014 2015 2014 2015 2014 2014 2015 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
[358] 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
[379] 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 2013 2013 2014 2014 2014 2013 2014 2014 2013 2013 2014 2013
[400] 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013
[421] 2012 2013 2013 2013 2013 2013 2013 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012
[442] 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012
[463] 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012
[484] 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012
Levels: 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
```

Рисунок 3.34 – Передбачені групи для кожного року

```
$posterior
             2011          2012          2013          2014          2015          2016          2017
1   8.935005e-161 2.786522e-154 1.087354e-134 1.041883e-119 2.817502e-115 2.618649e-108 2.149863e-74
2   1.210775e-156 3.351659e-150 8.186332e-131 4.303965e-116 6.856347e-112 4.381203e-105 8.731677e-72
3   2.719246e-168 1.033731e-161 8.661158e-142 2.186210e-126 1.397525e-121 2.564883e-114 3.020666e-79
4   5.725464e-162 2.375971e-155 3.368410e-135 1.832024e-119 2.002501e-114 2.436853e-107 1.635258e-73
5   1.974352e-172 9.682231e-166 2.406142e-145 2.538990e-129 5.300788e-124 1.553986e-116 5.623090e-81
6   3.090764e-163 1.290778e-156 1.839236e-136 9.961209e-121 1.100842e-115 1.475503e-108 1.648964e-74
7   6.287530e-177 4.211729e-170 3.971924e-149 2.408250e-132 2.112085e-126 1.046119e-118 1.191171e-82
8   5.719404e-170 3.507870e-163 2.556518e-142 1.172536e-125 7.571592e-120 2.117251e-112 1.544416e-77
9   4.207342e-170 2.174370e-163 7.179589e-143 1.133223e-126 3.169377e-121 8.033162e-114 9.487212e-79
10  5.871877e-152 1.835555e-145 8.531642e-126 1.117686e-110 3.514848e-106  1.739211e-99 3.847278e-67
11  2.199198e-149 6.239554e-143 1.948166e-123 1.519787e-108 3.064926e-104  1.188267e-97 1.169782e-65
12  5.397813e-167 2.781879e-160 9.635249e-140 1.667422e-123 4.838049e-118 9.712382e-111 3.121747e-76
13  4.800467e-155 2.159175e-148 5.026150e-128 5.653717e-112 1.019820e-106 7.918869e-100 2.573755e-67
14  1.361438e-153 5.026669e-147 4.857859e-127 1.689487e-111 1.181044e-106 7.290931e-100 2.105417e-67
15  8.673215e-157 3.830971e-150 7.917096e-130 7.462671e-114 1.195448e-108 1.035962e-101 7.171542e-69
16  4.284767e-156 1.937133e-149 4.519345e-129 5.049327e-113 9.174686e-108 7.690246e-101 3.791003e-68
17  2.406441e-151 9.131055e-145 1.044500e-124 4.653860e-109 3.861299e-104  2.071955e-97 2.245231e-65
18  4.553537e-148 1.688391e-141 1.842084e-121 7.909887e-106 6.153224e-101  2.574614e-94 7.750058e-63
19  2.279781e-152 8.249428e-146 7.429636e-126 2.372874e-110 1.529636e-105  8.543107e-99 1.575500e-66
20  1.724337e-151 9.090078e-145 4.697775e-124 1.593330e-107 6.592755e-102  4.393716e-95 2.252615e-63
21  7.552409e-161 5.086154e-154 6.732440e-133 7.409102e-116 8.604971e-110 1.300102e-102 1.728759e-69
22  4.000373e-151 1.930572e-144 6.699067e-124 1.335074e-107 3.617826e-102  2.225296e-95 1.222045e-63
23  5.344047e-140 2.079608e-133 3.315147e-113  2.553210e-97  2.877200e-92  7.000587e-86 6.489078e-56
24  5.189870e-112 7.416172e-106  1.806089e-87  4.797285e-74  3.422399e-71  4.534754e-66 2.594530e-40
```

Рисунок 3.35 – Імовірності віднесення років до кожної із груп

Результат отримання коефіцієнтів дискримінантної функції за допомогою функції *lda* зображено на рисунку 3.36.

```
> #Коефіцієнти дискримінантної функції
> fit <- lda(Year ~ Version3 + Version4 + Version5, data = data)
> fit
Call:
lda(Year ~ Version3 + Version4 + Version5, data = data)

Prior probabilities of groups:
 2011  2012  2013  2014  2015  2016  2017  2018  2019  2020  2021
0.016 0.104 0.104 0.104 0.104 0.104 0.106 0.104 0.104 0.104 0.046

Group means:
       Version3 Version4  Version5
2011  1192.7273 1431.273    0.0000
2012  1348.9528 1693.701    0.0000
2013  1333.9615 2858.462    0.0000
2014   977.2692 4411.308    0.0000
2015  1130.0962 5666.596    0.0000
2016   541.4038 5896.269  102.2692
2017   438.6415 5730.491  562.6981
2018 10243.5385 5301.808 1166.1538
2019  1824.6731 4701.058 1868.0385
2020   130.9615 4406.423 2416.0385
2021   101.7826 3926.391 2692.3043

Coefficients of linear discriminants:
                    LD1           LD2           LD3
Version3 -0.0002532478  0.0001045888  0.0007039184
Version4 -0.0014598812 -0.0030327399 -0.0001315028
Version5 -0.0086327660  0.0015563810 -0.0003628984

Proportion of trace:
   LD1    LD2    LD3
0.7837 0.1750 0.0413
```

Рисунок 3.36 – Коефіцієнти дискримінантної функції

Результат побудови таблиці класифікацій, у якій на головній діагоналі вказано число правильно прогнозованих випадків по кожній групі, а інші

значення – невірно, за допомогою функції table зображено на рисунку 3.37.

```
> #Таблиця класифікацій
> ct <- table(data$Year, fit$class)
> ct

     2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
2011    0    8    0    0    0    0    0    0    0    0    0
2012    0   52    0    0    0    0    0    0    0    0    0
2013    0   14   32    6    0    0    0    0    0    0    0
2014    0    0    0   49    3    0    0    0    0    0    0
2015    0    0    0    1   42    9    0    0    0    0    0
2016    0    0    0    0    5   47    0    0    0    0    0
2017    0    0    0    0    0    4   49    0    0    0    0
2018    0    0    0    0    0    0    0   52    0    0    0
2019    0    0    0    0    0    0    0    8   39    5    0
2020    0    0    0    0    0    0    0    0    0   51    1
2021    0    0    0    0    0    0    0    0    0    2   21
```

Рисунок 3.37 – Таблиця класифікацій

Отже, правильно було прогнозовано 434 випадків.

Результат підрахунку проценту вірно прогнозованих випадків за допомогою функцій *prop.table*, *diag* і *sum* зображено на рисунку 3.38.

```
> #Підрахунок проценту вірно прогнозованих випадків
> sum(diag(prop.table(ct)))
[1] 0.868
```

Рисунок 3.38 – Процент правильно прогнозованих випадків

Результат виведення інформації про кількість груп та об’єктів у них, середні значення предикатів для груп, коефіцієнти дискримінантного рівняння (без константи) за допомогою функцій *counts*, *means* і *scaling* зображено на рисунку 3.39.

```
> #Виведення кількості груп та об'єктів у них
> fit$counts
2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
   8   52   52   52   52   52   53   52   52   52   23
> #Інформація про середні значення предиктів для груп
> fit$means
      Version3 Version4  Version5
2011  1192.7273 1431.273    0.0000
2012  1348.9528 1693.701    0.0000
2013  1333.9615 2858.462    0.0000
2014   977.2692 4411.308    0.0000
2015  1130.0962 5666.596    0.0000
2016   541.4038 5896.269  102.2692
2017   438.6415 5730.491  562.6981
2018 10243.5385 5301.808 1166.1538
2019  1824.6731 4701.058 1868.0385
2020   130.9615 4406.423 2416.0385
2021   101.7826 3926.391 2692.3043
> #Коефіцієнти дискримінантного рівняння (без константи)
> fit$scaling
                   LD1           LD2           LD3
Version3 -0.0002532478  0.0001045888  0.0007039184
Version4 -0.0014598812 -0.0030327399 -0.0001315028
Version5 -0.0086327660  0.0015563810 -0.0003628984
```

Рисунок 3.39 значення предикатів для груп, коефіцієнти дискримінантного рівняння (без константи)

Завдання на самостійну роботу

Провести дискримінантний аналіз, що використовується для прийняття рішення про те, які змінні розділюють певні масиви даних.

3.9 Кластерний аналіз

Кластерний аналіз — задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. На нього не накладаються обмеження інших статистичних методів, такі як однорідність і обсяг вибірки. Можна аналізувати будь-які дані про будь-які об’єкти, і якщо між ними є зв’язок, то даний аналіз обов’язково їх знайде.

Результат стандартизації даних, які використовуватимемо при проведенні кластерного ієрархічного аналізу, для перших 50 значень вибірки за допомогою функції *scale*, яка з кожного значення змінної віднімає її середнє значення і ділить результат на її стандартне відхилення, зображено на рисунку 3.40.

```
> #Стандартизація даних
> scale(data[1:50,4:7])
      Version3    Version4    Version5       Total
1  -0.99321402 -1.63668709  1.16515633 -0.74904130
2  -1.37718852 -1.69201796  0.90712149 -1.06987893
3  -0.99321402 -1.59242240  1.63606991 -0.28497257
4  -1.05720977 -1.06124606  1.09419675 -0.22195089
5  -1.24919702 -1.26597027  1.81024343  0.18482540
6  -1.12120552 -1.11657693  1.19095981 -0.19903392
7  -0.92921827 -0.85652184  1.97796608  0.78639597
8  -1.82515876 -0.65179763  1.50060162  0.49420455
9  -0.86522253 -1.02804754  1.60381556  0.28222254
10 -0.73723103 -1.16084162  0.44910964 -0.86935541
11 -0.99321402 -1.24937101  0.29428874 -1.12144213
12 -0.73723103 -0.86205493  1.36513333  0.25357632
13 -0.22526504 -0.48580503  0.48136400 -0.09590754
14 -0.16126929 -0.86205493  0.48136400 -0.47976685
15  0.28670096 -0.62966528  0.63618490 -0.06153208
16  0.03071796 -0.53560281  0.56522532 -0.05007359
17 -0.35325653 -0.66839689  0.27493612 -0.47976685
18  0.09471371 -0.53006972  0.00399954 -0.53133004
19 -0.03327779 -0.83992258  0.39105180 -0.52560080
20 -0.67323528  0.10623527  0.07495912  0.11607447
21 -0.60923953  0.08963601  0.72004623  0.67754034
22 -0.60923953 -0.08188969  0.10076261 -0.05007359
23 -0.67323528  0.10623527 -0.78300673 -0.64591491
24 -0.41725228 -1.10551075 -2.53119278 -3.43032726
25 -0.92921827 -0.98378284 -2.24735445 -3.09803114
```

```
26 -0.35325653  0.21689700 -0.29919140 -0.07299056
27 -0.48124803  0.16709922 -0.09921440  0.04159431
28 -0.28926079  0.36075726  0.33299396  0.64316488
29 -0.16126929  0.36075726 -0.47981579 -0.06726132
30  1.31063294  0.57101456 -0.02180394  0.68899883
31  0.35069670  0.31649257 -0.54432450 -0.12455375
32  0.47868820 -0.02655882 -0.27338791 -0.22768014
33 -0.28926079  0.07856983 -0.15082137 -0.07871981
34 -0.80122678  0.08963601 -0.46046318 -0.38809896
35 -0.03327779  0.45481974 -1.06039418 -0.47403761
36 -0.54524378  0.38842270 -0.55077537 -0.13601224
37  0.09471371  0.54888222 -0.72494889 -0.06726132
38  0.03071796  0.88640051 -0.95072937  0.08169901
39  0.79866695  1.12432325 -0.88622066  0.44837060
40  1.43862444  1.04132695 -1.60226735 -0.21622165
41  1.63061168  0.96939682 -1.33133077 -0.03288586
42  1.63061168  0.79233804 -0.90557328  0.16190842
43  1.24663719  0.94726447 -1.42164296 -0.17038770
44  1.88659468  1.15752177 -0.44111056  0.97546100
45  1.43862444  1.57250328  0.35879745  2.07547576
46  1.56661593  1.72742972  0.17817306  2.08693424
47  1.63061168  1.65549959 -0.11856701  1.75463812
48  1.82259893  1.66656576 -0.29919140  1.62286552
49  1.69460743  1.45630846 -0.79590847  0.95254403
50  1.05464994  2.07048110 -0.61528408  1.69161644
```

Рисунок 3.40 – Стандартизовані дані

Результат побудови матриці відстаней за допомогою функції *daisy* з пакету *cluster* зображено на рисунку 3.41.

```
> #Побудови матриці відстаней
> to_clust = data[1:50,4:7]
> dm = daisy(scale(to_clust))
> dm
Dissimilarities :
          1         2         3         4         5         6         7         8         9        10
2  0.5657002
3  0.6626302 1.1422782
4  0.7861853 1.1199280 0.7640956
5  1.2211481 1.6086674 0.6504958 0.8700350
6  0.7681604 1.1115732 0.6695664 0.1305574 0.7546922
7  1.9055132 2.3433301 1.3454994 1.3623896 0.8124361 1.3018981
8  1.8221754 2.0202256 1.4840403 1.1980973 0.9489644 1.1349130 1.0760455
9  1.2816985 1.7368152 0.8109571 0.7428751 0.5061050 0.6895088 0.6539847 1.0576536
10 0.9050716 0.9704012 1.4149796 0.9734350 1.7992064 1.0719461 2.2821885 2.0994389 1.6411866
11 1.0232822 0.8494678 1.6179438 1.2200097 2.0174876 1.2995661 2.5754692 2.2615771 1.9366210 0.4011009
12 1.3079813 1.7491978 0.9810227 0.6645271 0.7925317 0.6688874 0.8344762 1.1419494 0.3189439 1.4796447
13 1.6758374 1.9778188 1.7842194 1.1894181 1.8713167 1.3094823 1.9108266 1.9935624 1.4513623 1.1476260
14 1.3536128 1.6421936 1.6114773 1.1333223 1.8852951 1.2524361 2.1054126 2.1909363 1.5374098 0.7575124
15 1.8451978 2.2332244 1.9012573 1.4926041 2.0501151 1.5956082 2.0122828 2.3487232 1.5937730 1.4206198
16 1.7633197 2.1157487 1.8349905 1.3301279 1.9434162 1.4415934 1.9287261 2.1514988 1.4947580 1.2904940
17 1.4873239 1.6864541 1.7759501 1.1779379 1.9896368 1.3071042 2.2069441 2.1488713 1.6546922 0.7563424
18 1.9503518 2.1499498 2.2442205 1.7009766 2.4744780 1.8280442 2.6053701 2.6442895 2.0954089 1.1842252
19 1.4850841 1.7591842 1.7594617 1.2977045 2.0442080 1.4165533 2.2455773 2.3489002 1.6884601 0.8485736
20 2.2532465 2.4141757 2.3634149 1.6320403 2.2870545 1.7437597 2.2501544 2.0191534 1.9205374 1.6494363
21 2.3153587 2.6177295 2.1776794 1.5729844 1.9179447 1.6453574 1.6099109 1.6343384 1.5006892 2.0115589
          11        12        13        14        15        16        17        18        19        20
2
3
4
5
6
7
8
9
10
11
12 1.8035867
13 1.5031623 1.1431799
14 1.1352930 1.2847460 0.5413016
15 1.8062447 1.3164729 0.5549380 0.6734776
16 1.6670750 1.1951455 0.2777419 0.5788775 0.2820312
17 1.0766607 1.3824840 0.4895733 0.3420161 0.8464436 0.6587676
18 1.4606441 1.8086209 0.7223679 0.6373987 0.8167906 0.7420972 0.5439459
19 1.2056183 1.4324775 0.5958615 0.1647079 0.6496496 0.5942798 0.3839154 0.5120868
20 1.8760617 1.6202276 0.8725195 1.3114446 1.3451935 1.0841920 1.0475968 1.1911329 1.3477470
21 2.3147319 1.2320634 1.0647815 1.5819994 1.3687090 1.1635552 1.4756775 1.6892826 1.6587938 0.8577600
```

Рисунок 3.41 – Матриця відстаней

Проведення кластерного аналізу, використовуючи метод Уорда в якості методу агрегування, за правилами якого необхідна матриця з квадратами евклідових відстаней, для отримання якої використаємо метод *ward.D2*, який сам підносить значення відстаней із матриці відстаней до квадрату, за допомогою функції *hclust*, та побудова дендограми зображено на рисунку 3.42.

```
> #Побудова дендограми
> hc = hclust(dm, method = "ward.D2")
> plot(hc)
```

Рисунок 3.42 – Проведення кластерного аналізу та побудова дендограми

Результат проведення кластерного аналізу та побудови дендограми за допомогою функції *plot* зображено на рисунку 3.43.

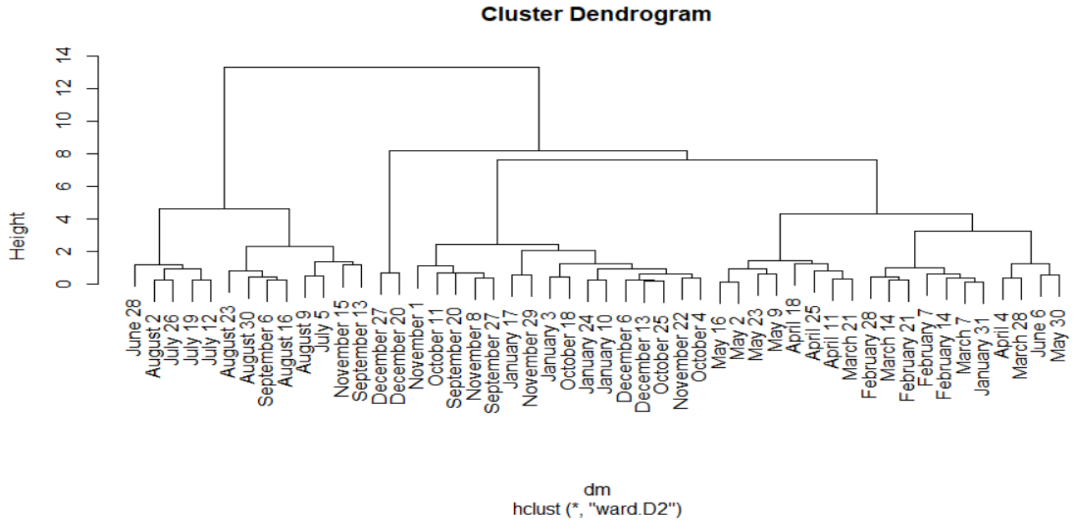


Рисунок 3.43 – Результат проведення кластерного аналізу і побудови дендограми

Виділення п’яти кластерів за допомогою функції *rect.hclust* з параметрами *hc* та *k*, що дорівнює 5, зображено на рисунку 3.44.

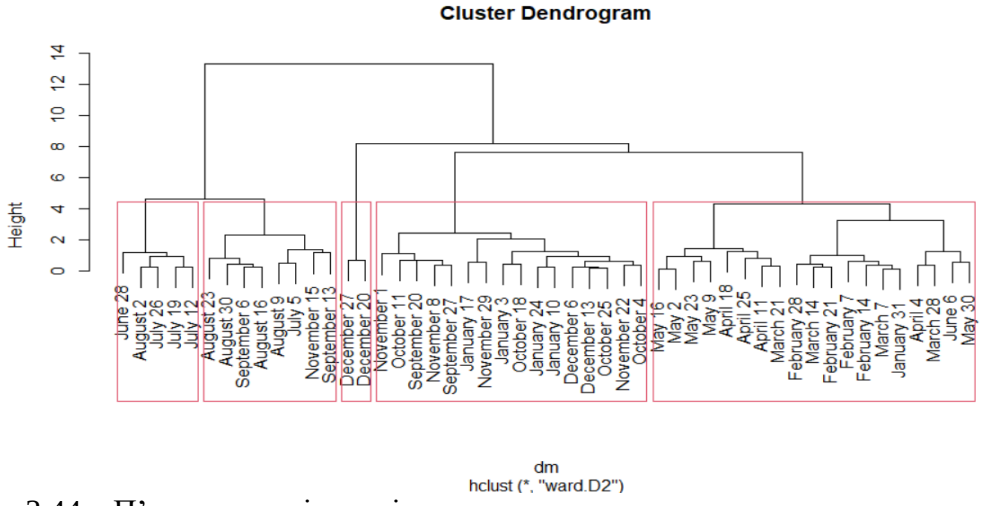


Рисунок 3.44 – П’ять кластерів, виділених червоними прямокутниками на дендрограмі

Завдання на самостійну роботу

Провести кластерний аналіз даних.

Контрольні запитання

1. Основні характеристики розподілу ймовірностей. Записати аналітичні вирази.
2. Що таке квантиль, мода, медіана?
3. Як визначити знак коефіцієнта асиметрії з вигляду графіка щільності ймовірності?
4. Як визначити знак коефіцієнта ексцесу з вигляду графіка щільності ймовірності?
5. Що таке ряди розподілу?
6. Які характеристики розподілу ви знаєте?
7. Як будується гістограма?
8. Що таке таблиця частот?
9. На що впливає ширина інтервалу?
10. Які ви знаєте параметри нормального закону розподілу випадкових величин.
11. Наведіть аналітичний вираз щільності ймовірності для нормального закону розподілу.
12. Накресліть функцію щільності ймовірності для нормального закону.
13. Чому дорівнюють коефіцієнт асиметрії та коефіцієнт ексцесу для нормального закону?
14. Яким критерієм визначається закон розподілу випадкової величини? Записати аналітичний вираз.
15. Як змінюється вигляд гістограми при зміні величини інтервалу?
16. Як використовується правило трьох сігм для визначення закону розподілу випадкової величини?
17. Необхідність застосування статистичних методів обробки результатів спостережень.
18. У чому полягає основний принцип статистичних гіпотез?
19. Які гіпотези можна перевірити за допомогою критерія Пірсона, Колмогорова, Фішера, Стьюдента, Кохрена?
20. Як визначити довірчі інтервали для математичного сподівання?
21. Що значать поняття статистична залежність, функціональна залежність, кореляційна залежність?
22. Коли використовується кореляційний аналіз?
23. Як визначається індекс кореляції і що він показує?
24. Як визначається коваріація і коефіцієнт кореляції?
25. Як визначити вибіркове значення коефіцієнта кореляції?
26. Як визначається кореляційне відношення? Який його фізичний зміст?
27. Яким чином перевіряється гіпотеза про відсутність кореляційного зв’язку?
28. Для чого призначений узагальнений засіб обчислення парних кореляційних характеристик?
29. Як розраховується узагальнений коефіцієнт кореляції?

30. В чому полягає мета регресійного аналізу?
31. Для чого використовується метод найменших квадратів?
32. Наведіть передумови регресійного аналізу.
33. Як виконується перевірка значущості оцінок коефіцієнтів рівняння регресії?
34. За допомогою якого критерію виконується перевірка адекватності рівняння регресії?
35. Як виконується вибір найкращої моделі рівняння регресії?
36. Дайте визначення теорії планування експерименту.
37. Дайте визначення активного та пасивного експерименту.
38. Вимоги для вибору відгуків, факторів, координат базової точки та ступенів варіювання.
39. Призначення ПФЕ та етапи його проведення.
40. Що таке матриця планування. Як визначається кількість можливих комбінацій рівнів варіювання.
41. Властивості МП. Наведіть приклади кодування факторів.
42. З якою метою необхідна рандомізація дослідів.
43. Як впливає величина ступенів варіювання на значущість коефіцієнтів і на адекватність рівняння регресії.
44. Коли використовується ОЦКП?
45. З яких частин складається ЦКП?
46. Назвіть види ЦКП. Чим відрізняються види ЦКП?
47. Розкажіть о зоряних точках при ЦКП.
48. Що є критерієм оптимальності для ОЦКП?
49. Що є критерієм оптимальності для РЦКП?
50. Яким чином в ОЦКП забезпечується ортогональність векторів-стовпців, включаючи квадратичні вектори-стовпці?
51. Яким чином, знаючи величину зоряного плеча в нормованому масштабі, можна перейти до звичайного масштабу?
52. Як будується МП при ОЦКП?
53. Методика обчислювання формул ОЦКП.
54. Як виконується аналіз моделей другого порядку?
55. Поняття стаціонарної точки.

Список використаних джерел

1. Згуровський М.З. Основи системного аналізу: Підручник / М.З. Згуровський, Н.Д.Панкратова; За заг.ред. М.З. Згуровського. – К.: Видавнича група BHV, 2007. – 544 с., іл. (Інформатика).
2. Катренко А.В. Системний аналіз об'єктів та процесів комп'ютеризації: Навчальний посібник. – Львів: Новий світ, 2003. – 424 с.
3. Катренко А.В. Системний аналіз: Підручник/ За наук. ред. В.В. Пасічника. – Львів: Новий світ-2000, 2011. – 396 с. (Комп'ютинг).
4. Методичні вказівки до самостійних робіт з дисципліни «Емпіричні методи програмної інженерії» для студентів напряму підготовки 121 «Інженерія програмного забезпечення» (всіх форм навчання) / А. В. Притула, Н.О. Миронова. – Запоріжжя: ЗНТУ, 2017. – 36 с.
5. Проектування інформаційних систем: Посібник/ За ред. В.С. Пономаренка. – К.: Академія, 2002. – 488с. (Альма-матер).
6. Томашевський О. В. Комп'ютерні технології статистичної обробки даних: навч. посібник для студ. вищ. навч. закл. / О. В. Томашевський, В. П. Рисіков. – Запоріжжя: ЗНТУ, 2006. – 174 с.
7. ГОСТ 19.701-90 – ЕСПД. Схеми алгоритмів, програм, даних і систем. Умовні позначки і правила виконання. [Чинний від 1992-01-01]. – (Національний стандарт України).
8. ГОСТ 2.105-95 – ЕСКД. Загальні вимоги до текстових документів. [Чинний від 01.07.1996]. 30 с. – (Національний стандарт України).