DOI: https://doi.org/10.36486/978-966-2310-66-5-30

Технології – це всього лише інструмент

Білл Гейтс

РОЗДІЛ 30
Аналіз та операції з великими даними в середовищі R

Микола Карчевський

Про R. R - це мова програмування і програмне середовище (RStudio) для статистичних обчислень, аналізу та візуалізації даних, має широкий набір функцій, які дозволяють швидко і легко опрацьовувати великі масиви даних [1].

R була створена в 1993 році Россом Іхакою та Робертом Джентлменом, працівниками Оклендського Університету в Новій Зеландії. Іхака і Джентлмен були незадоволені наявними мовами програмування для статистичного аналізу. Вони вважали, що ці мови є занадто складними для навчання і використання, а також не пропонують достатнього рівня гнучкості.

> *R була розроблена як спроба створити мову програмування, яка була б простішою в освоєнні, але в той же час пропонувала б потужний набір функцій для статистичного аналізу.*

Назва R походить від перших літер імен її розробників.

R може використовуватися для виконання широкого спектра завдань, включаючи:

збір і підготовку даних
описову статистику
перевірку гіпотез
регресію
кластеризацію
машинне навчання

R має ліцензію GNU General Public License, що дозволяє вільно використовувати, розповсюджувати та модифікувати мову.

R vs. Python. Це обидві мови програмування з відкритим кодом, які широко використовуються для статистичного аналізу. Однак між ними є деякі ключові відмінності. Розробники професійної сертифікаційної програми «Аналіз даних від Google» систематизують ці відмінності наступним чином [2]

Мови	R	Python
Спільні риси	- відкритий вихідний код - дані зберігаються у датафреймах - формули та функції легко використовувати - інтернет-спільнота для розробки коду та консультативної підтримки	- відкритий вихідний код - дані зберігаються у датафреймах - формули та функції легко використовувати - інтернет-спільнота для розробки коду та консультативної підтримки
Переваги	- бібліотеки та пакети для маніпуляцій з даними, їх візуалізації, статистичних операцій *- «скальпельний» підхід до даних: існують спеціалізовані пакети (packages), для різноманітних завдань щодо обробки даних*	- простий синтаксис для потреб машинного навчання - інтегрується з хмарними платформами, такими як Google Cloud, веб-сервісами Amazon і Azure
Виклики	- неповні угоди про найменування ускладнюють для початківців вибір правильних функцій - методи обробки змінних можуть бути складними для розуміння початківцями	- багато різноманітних рішень для початківців щодо введення/виведення даних, структури, змінних, пакетів і об’єктів *- підхід «швейцарського ножа» для даних: необхідно визначати спосіб бажаних дій з даними*

Таким чином, *основна різниця між R та Python полягає в тому, що R - це мова програмування, орієнтована на статистику, тоді як Python - мова програмування загального призначення.*

Корисні посилання

The R Project for Statistical Computing: офіційний сайт для завантаження R та документації. https://www.r-project.org/

R Manuals: посилання на документи та інструкції з використання від команди розробників. https://cran.r-project.org/manuals.html

Coding Club R Tutorials: набір навчальних вправ з R. https://ourcodingclub.github.io/tutorials.html

R for Beginners: інструкція для початкі роботи з даними, аналізом, статистикою та візуалізацією в R. https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

ПРАКТИЧНИЙ КЕЙС

Порівняти кількість облікованих у 2021 та 2022 роках кримінальних проваджень за ознаками крадіжки (ст. 185 КК). Представити у вигляді інтерактивної картограми регіональний розподіл таких проваджень.

Особливості виконання практичного кейсу

Для того, щоб читачі мали можливість працювати над практичним кейсом разом з авторами посібнику, пропонуємо скопіювати вміст створеного нами репозиторію з вхідними даними, скриптами та візуалізаціями [3] у окрему теку, встановити R та RStudio, запустити проєкт R_case.

Також є можливість використовувати хмарний сервіс PositCloud та представлений у відкритому доступі проєкт даного практичного кейсу [4]

Перехід до репозиторію GitHub

Перехід до проєкту у он-лайн середовищі RStudio Posit Cloud

Вхідні дані. Для виконання завдання використаємо базу даних отриману в процесі розробки веб-застосунку «Протидія злочинності в Україні: інфографіка» [5]. Вона міститься у файлі crimes_ukr-REG.RData репозиторію практичного кейсу. Отже, почнемо з завантаження цього файлу (рис. 30.1)

```
# Заванаження необхідної бази даних
load ("crimes_ukr-REG.RData")
```

Рисунок 30.1. Скрипт завантаження БД

Результатом його виконання стане завантаження у середовище R двох датафреймів (data.frame): crimes_Reg та crimes_Reg_f1 (рис. 30.2).

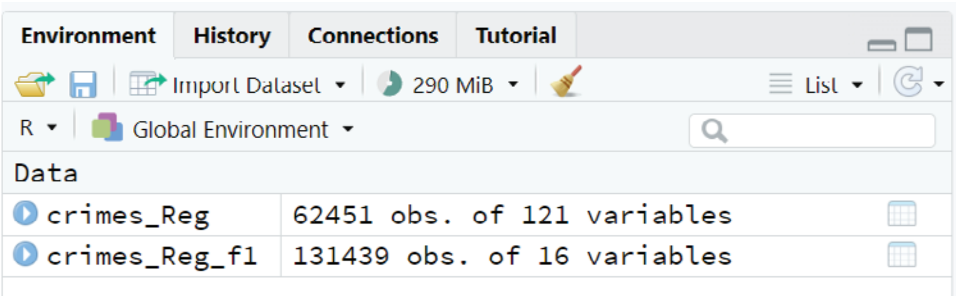

Рисунок 30.2. У середовище R завантажено 2 датафрейми

Перший містить інформацію щодо засуджених осіб, другий – щодо облікованих проваджень. Виходячи з умов кейсу, працюватимемо далі з фреймом crimes_Reg_f1.

Ознайомлювальний аналіз даних. Щоб отримати уявлення про структуру та зміст об'єктів типу data.frame в R, можна використовувати такі команди:

str() - *виводить структуру об'єкта, включаючи назви стовпців, типи даних та кількість спостережень;*
summary() - *виводить статистичний звіт про дані, включаючи середнє, стандартне відхилення, мінімальне та максимальне значення;*
head() та tail() - *виводять перші або останні кілька рядків об'єкта;*
dim() - *виводить кількість стовпців та рядків об'єкта;*
names() - *виводить список назв стовпців об'єкта;*
class() - *виводить тип даних об'єкта.*

Виконання у консолі середовища команди str (crimes_Reg_f1) дозволить отримати уявлення про: кількість спостережень (рядків) та кількість параметрів (стовбців) у датафреймі, який нас цікавить, типи даних у стовбцях, назви стовбців, зміст даних (рис. 30.3).

Отримання даних для візуалізації. Виходячи з умов кейсу, необхідно отримати дані щодо кримінальних проваджень, облікованих за ознаками кримінального правопорушення, передбаченого ст. 185 КК, у 2021 або 2022 році, з урахуванням регіонального розподілу. Отже, з датафрейму crimes_Reg_f1 необхідно отримати данні, що відносяться до 2021 або 2022 року (Year), стосуються статті 185 КК (Article), характеризують обліковані провадження (ACC [6]) та адміністративно-територіальний регіон (Region).

```
Console   Terminal ×   Background Jobs ×
R 4.3.0 · C:/Users/User/Desktop/PractycnyiCase/
> str (crimes_Reg_f1)
'data.frame':   131439 obs. of  16 variables:
 $ Region    : chr  "АР Крим" "АР Крим" "АР Крим" "АР Крим" ...
 $ Year      : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2
013 ...
 $ Article   : chr  "109" "110" "111" "112" ...
 $ nazva     : chr  "Стаття 109. Дії, спрямовані на насильницьку з
міну чи повалення конституційного ладу або на захоплення державної
влади" "Стаття 110. Посягання на територіальну цілісність і недото
рканність України" "Стаття 111. Державна зрада" "Стаття 112. Посяг
ання на життя державного чи громадського діяча" ...
 $ Chapter   : num  1 1 1 1 1 1 1 2 2 2 ...
 $ ChapterTXT: chr  "I.Злочини проти основ національної безпеки Ук
раїни" "I.Злочини проти основ національної безпеки України" "I.Зло
чини проти основ національної безпеки України" "I.Злочини проти ос
нов національної безпеки України" ...
 $ ACC       : int  0 0 0 0 0 0 0 550 0 0 ...
```

Рисунок 30.3. Оцінка структури та змісту датафрейму за допомогою команди str()

Скористаємося для цього функціями select та subset. Дані функції включені до пакету tidyverse [7], тому підключимо його використовуючи команду library () (рис. 30.4).

```
# Підключення бібліотеки
library (tidyverse)
```

Рисунок 30.4. Скрипт підключення бібліотеки

Функція select() використовується для вибору стовпців з об'єкта типу data.frame. Функція приймає як аргументи список назв стовпців, які потрібно вибрати.

Функція subset() використовується для вибору рядків або стовпців з об'єкта типу data.frame за певною умовою. Функція приймає як аргументи список умов, які потрібно задовольнити для вибору рядків або стовпців.

Для того, щоб отримати набір даних, який відповідає умовам практичного кейсу необхідно додати до скрипту такі команди (рис. 30.5):

```
# Створення датафрейму для візуалізації
data<-select(crimes_Reg_f1, c("Year","Region","Article","ACC"))
data<-subset(data,(Year>2020&Article=="185"))

data<-subset(data,!str_detect(Region,"залізниця"))
```

Рисунок 30.5. Скрипт створення data.frame.

У такий спосіб створюється новий датафрейм data. З використанням команди select до нього було включено тільки стовбці з даними про рік обліку, регіон, статтю КК, та кількість облікованих порушень. Далі, використовуючи команду subset у data було залишено тільки ті рядки, що мають відношення до 2021 або 2022 року та стосуються тільки ст. 185 КК.

Якщо переглянути отриманий набір даних, можна побачити, що він містить записи, які відносяться до обліку правопорушень на залізницях, рядки із значенням «Донецька залізниця», «Львівська залізниця», «Одеська залізниця», «Південна залізниця», «Південно-західна залізниця», «Придніпровська залізниця» у полі Region. Для того, щоб вилучити ці записи доповнимо скрипт командою data<-subset(data,!str_detect(Region,»залізниця»))

Вже відома функція subset(), але в якості умови результат функції str_detect(). Остання використовується для виявлення шаблону в рядку, вона повертає логічний вектор із значенням TRUE для кожного елемента рядка, який відповідає шаблону, і FALSE в іншому випадку.

> *У R вектор - це одновимірний об'єкт, який містить ряд значень, які можуть бути одного або різних типів. Вектори можна використовувати для зберігання даних, таких як числа, текст або логічні значення*

Використаємо перед цією функцією оператор !. В R він є логічним, означає заперечення, повертає логічну змінну, яка є протилежною до логічної змінної, яка розташована за ним. Таким чином, отримали умову на підставі якої до data будуть включені тільки записи

```
# Приведення назв регіонів у базі географічних даних
# у відповідність із назвами у базі статистичних даних
shape_data$ADM1_UA <- replace(shape_data$ADM1_UA,
  shape_data$ADM1_UA == "Автономна Республіка Крим", "АР Крим")
```

Рисунок 30.12. Скрипт зміни значень

в яких поле Region не містить слова «залізниця», іншими словами, тільки дані щодо областей, а також міст Києва та Севастополя,.

Перша візуалізація. Побудуємо візуалізацію отриманих даних у вигляді стовпчикової діаграми. Нехай по осі аргументів будуть розташовані адміністративно-територіальні одиниці, вісь результатів характеризуватиме кількість облікованих проваджень, а дані 2021 та 2022 року будуть представлені у двох окремих сюжетах візуалізації (рис. 30.6).

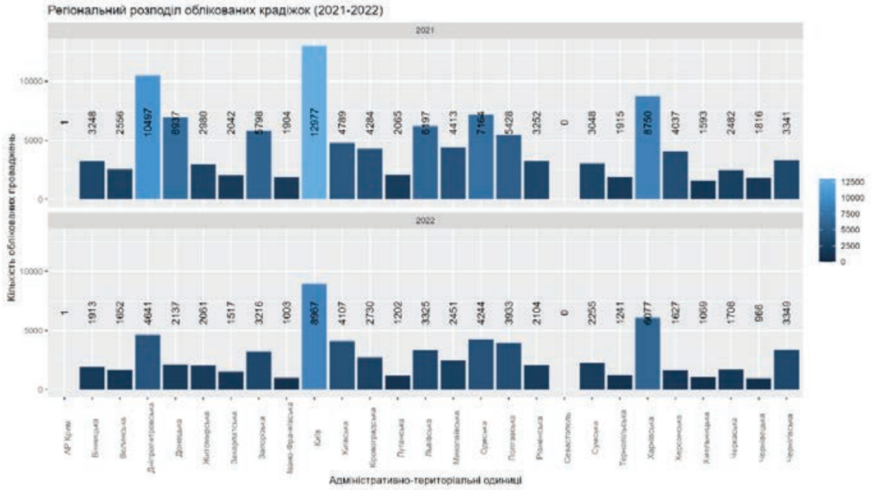

Рисунок 30.6. Візуалізація отриманих даних у вигляді стовпчикової діаграми

Таку візуалізацію було отримано в результаті виконання наступних команд (рис. 30.7):

```
13 # Побудова візуалізації
14 vis1<-ggplot(data,
15              mapping = aes(x=Region, y=ACC, fill=ACC, label=ACC))+
16   geom_bar(stat = "identity")+
17   facet_wrap(.~Year, ncol=1)+
18   theme(legend.title = element_blank(),
19       axis.text.x=element_text(angle = 90))+
20   geom_text(aes(angle=90,y=max(data$ACC/2)))+
21   ggtitle ("Регіональний розподіл облікованих крадіжок (2021-2022)")+
22   ylab ("Кількість облікованих проваджень")+
23   xlab ("Адміністративно-територіальні одиниці")
24
25 # Виведення візуалізації у вікно Viewer
26 vis1
27
28 # Збереження візуалізації у файл
29 ggsave (filename = "diagram.png",vis1,
30         width = 32, height = 18, dpi = 300, units = "cm")
```

Рисунок 30.7. Скрипт візуалізації даних у вигляді стовпчикової діаграми

Прокоментуємо дану частину скрипту. Функція ggplot() є основою пакету tidyverse для візуалізації даних. Вона використовується для створення графіків шляхом поєднання різних компонентів, таких як дані, естетика, геометричні об’єкти та написи.

Результат роботи функції ggplot присвоєно змінній vis1 (рис. 30.7, рядок 14, vis1 <- ggplot…), в нашому випадку це зроблено для того, щоб зберегти побудовану візуалізацію у вигляді графічного файлу.

Джерело даних, осі аргументів та результатів, колір діаграми та зміст підписів визначено у наступній частині: «ggplot(data, mapping = aes(x=Region, y=ACC, fill=ACC, label=ACC))» (рис. 30.7, рядки 14 та 15). Як можна побачити, джерелом даних визначається створений раніше датафрейм data. Естетика (**aes**) діаграми характеризується: 1) змінними для осі x, які містить поле Region (назви адміністративно-територіальних одиниць); 2) змінними для осі y, які містить поле ACC (кількість облікованих проваджень); 3) кольором заповнення елементів діаграми та підписами, які визначаються кількістю облікованих проваджень (fill=ACC, label=ACC).

Команда «geom_bar(stat = "identity")» (рис. 30.7, рядок 16) визначає форму візуалізації як стовпчикову діаграму.

Визначення двох сюжетів за роками спостереження забезпечується командою «facet_wrap(.~Year, ncol=1)» (рис. 30.7, рядок 17).

Налаштування «theme(legend.title = element_blank(), axis.text.x=element_text(angle = 90))» (рис. 30.7, рядки 18 та 19) встановлює, що підпис легенди відсутній (в даному випадку він був би збитковою інформацією), а підписи по горизонтальній осі слід повернути на 90 градусів.

Команда «geom_text(aes(angle=90,y=max(data$ACC/2)))» (рис. 30.7, рядок 20) забезпечує вертикальне розміщення підписів даних на рівні середини максимального значення досліджуваного показника (ACC).

Блок команд «ggtitle («Регіональний розподіл облікованих крадіжок (2021-2022)») + ylab («Кількість облікованих проваджень») + xlab («Адміністративно-територіальні одиниці»)» (рис. 30.7, рядки 21-23) визначає заголовок діаграми, підписи вертикальної та горизонтальної осей.

Команда vis1 забезпечує виведення візуалізації у вікні Plots (рис. 30.7, рядок 26).

Нарешті, збереження отриманої візуалізації у файл «diagram1.png» забезпечується командою «ggsave (filename = "diagram1.png",vis1, width = 32, height = 18, dpi = 300, units = "cm")» (рис. 30.7, рядки 29 та 30). Як можна побачити, аргументами функції ggsave виступають ім'я файлу, змінна, яка містить створену візуалізацію (vis1), розміри візуалізації, одиниці вимірювання, якість графіки.

Повний скрипт для розв'язання цієї частини кейсу представлено у файлі vis1.R репозиторію R_case.

Картограма. Для побудови картограми будемо використовувати бібліотеки leaflet [8], sf [9] та htmlwidgets [10] (рис. 30.8).

```
# Підключення бібліотек
library (leaflet)
library (sf)
library (htmlwidgets)
```

Рисунок 30.8. Застосовувані біліотеки

Отримання даних. Географічні дані про межі областей, міст Києва та Севастополя отримаємо з сайту The Humanitarian Data Exchange, що є ресурсом відкритих даних Офісу ООН з координації гуманітарних питань [11]. Копії отриманих з цього ресурсу файлів розміщено нами у теці Shape_files репозиторію практичного кейсу. У датафреймі shape_data збережемо отримані географічні дані використовуючи спеціальну функцію read_sf() (рис. 30.9, рядок 7).

```
6  # Отримання геоданих
7  shape_data <- st_read("Shape_files/ukr_admbnda_adm1_sspe_20230201.shp")
```

Рисунок 30.9. Скрипт збереження географічних даних

Цілком очевидно, що для побудови картограми необхідно буде об'єднати географічні (shape_data) та статистичні дані (data). Спільними полями використовуваних у проєкті датафреймів є поле Region датафрейму data та поле ADM1_UA датафрейму shape_data. Кожне з цих полей містить назви адміністративно-територіальних одиниць. Отже, для коректного об'єднання географічних та статистичних даних необхідно, щоб поля data$Region та shape_data$ADM1_UA містили однакові назви адміністративно-територіальних одиниць.

Для здійснення перевірки використаємо функцію setdiff(), яка порівнює два вектори по елементах. Для кожного елемента першого вектора перевіряється, чи він є в другому векторі. Якщо елемент відсутній в другому векторі, він додається до вектору, який буде повернуто як результат. Виконання в консолі відповідної команди показало, що найменування «АР Крим» присутнє тільки у полі data$Region (рис. 30. 10).

```
> setdiff(data$Region,shape_data$ADM1_UA)
[1] "АР Крим"
```

Рисунок 30.10. Перевірка відповідності даних

Далі дізнаємося, які унікальні значення містить поле найменувань датафрейму з географічними даними. Використаємо для цього функцію unique(). Так само виконаємо її в консолі (рис. 30.11)

```
> setdiff(data$Region,shape_data$ADM1_UA)
[1] "АР Крим"
> unique(shape_data$ADM1_UA)
 [1] "Автономна Республіка Крим" "Черкаська"
 [3] "Чернігівська"              "Чернівецька"
 [5] "Дніпропетровська"          "Донецька"
 [7] "Івано-Франківська"         "Харківська"
 [9] "Херсонська"                "Хмельницька"
[11] "Кіровоградська"            "Київ"
[13] "Київська"                  "Луганська"
[15] "Львівська"                 "Миколаївська"
[17] "Одеська"                   "Полтавська"
[19] "Рівненська"                "Севастополь"
[21] "Сумська"                   "Тернопільська"
[23] "Вінницька"                 "Волинська"
[25] "Закарпатська"              "Запорізька"
[27] "Житомирська"
>
```

Рисунок 30.11. Скрипт перевірки унікальних значень, що містить поле найменувань датафрейму з географічними даними

Як можна побачити, якщо у датафреймі географічних даних використовується значення «Автономна Республіка Крим», то у датафреймі статистичних даних використовується значення «АР Крим». Використовуючи функцію replace() здійснимо необхідні зміни у shape_data (рис. 30.12, рядки 11 та 12).

Функція replace() призначена для заміни значень в векторі. В нашому випадку вона використовується з трьома аргументами:

1) shape_data$ADM1_UA - вектор, в якому потрібно замінити значення;

2) shape_data$ADM1_UA == "Автономна Республіка Крим" - значення, яке потрібно замінити;

3) "АР Крим" - значення, на яке потрібно замінити.

Виконавши необхідні підготовчі дії, об'єднаємо статистичний та географічний датафрейми за полями з назвами адміністративно-територіальних одиниць. Використаємо для цього пайп-оператор %>% та функцію left_join() (рис. 30.13, рядок 15).

```
14 # Об'єднання географічної та статистичної баз даних за назвою регіону
15 shape_data<-shape_data %>% left_join(data, by = c( "ADM1_UA"="Region" ))
```

Рисунок 30.13. Скрипт обєднання датафреймів

Пайп-оператор %>% використовується для з'єднання кількох функцій в одну послідовність.

Функція left_join() використовується для об'єднання двох наборів даних, використовуючи значення в першому наборі даних як ключі.

> *Ключі наборів даних - це значення, які використовуються для об'єднання двох наборів даних. Ключі повинні бути унікальними в кожному наборі даних, щоб забезпечити правильне об'єднання.*

Функція повертає новий набір даних, який містить всі записи з першого набору даних, а також записи з другого набору даних, у яких відповідні ключі збігаються.

В нашому випадку, виконання даної частини скрипту приведе до формування датафрейму, де кожній адміністративно-територіальній одиниці буде відповідати два записи (2021 та 2022 рік).

Наступним кроком, необхідно сформувати окремі набори даних для шарів майбутньої картограми. Виходячи з умов кейсу такими шарами будуть дані за 2021 рік та дані за 2022 рік (рис. 30.14).

```
17 # Отримання даних для шарів картограми (по роках)
18 shape_data_2022 <- shape_data %>% filter(Year == 2022)
19 shape_data_2021 <- shape_data %>% filter(Year == 2021)
```

Рисунок 30.14. Скрипт формування даних для шарів картограми

Нарешті для побудови картограми знадобіться кольорова шкала. Побудуємо її за допомогою функції colorBin() та призначимо для змінної pal_d (рядки 23 та 24, рис. 30.15).

```
21 # Функція для визначення кольорів регіональних полігонів
22 # залежно від кількості облікованих проваджень
23 pal_d<-colorBin(palette =  c('#006837','#a6d96a','#ffffbf','#fdae61','#a50026'),
24                domain = shape_data$ACC, bins = 5)
```

Рисунок 30.15. Скрипт для визначення кольорів за регіонами

Функція colorBin() використовується з метою створення шкали кольорів для відображення даних на карті. Вона приймає три аргументи: 1) palette - палітра кольорів, яка буде використовуватися для відображення даних; 2) domain - діапазон даних, які потрібно відобразити; 3) bins - кількість відтінків кольору, які потрібно використовувати. У випадку нашого кейсу палітру складають кольори, що традиційно використовуються для візуалізації рельєфу на фізичних мапах (рядок 23, рис. 15), діапазон даних – всі можливі значення кількості облікованих проваджень за ознаками крадіжки у 2021 або 2022 році, кількість відтінків дорівнює п'яти.

Побудова картограми. Побудова картограми відбувається шляхом призначення певній змінній, в нашому випадку map, результату виконання функції leaflet() (рис. 30.16):

```
26 #Побудова картограми
27 map<-leaflet()%>%setMaxBounds(20,45,41,53)%>%addProviderTiles("Esri.WorldGrayCanvas")%>%
28   addPolygons(data=shape_data_2021,group = "2021",
29               weight=1,fillOpacity = 0.6,
30               color = pal_d (shape_data_2021$ACC),
31               popup = ~paste0("<b>", ADM1_UA, "</b>",
32                   "<br/>","Кількість облікованих проваджень у 2021 році:", ACC),
33               highlight = highlightOptions(weight = 3, color = "white"))%>%
34   addPolygons(data=shape_data_2022,group = "2022",
35               weight=1,fillOpacity = 0.6,
36               color = pal_d (shape_data_2022$ACC),
37               popup = ~paste0("<b>", ADM1_UA, "</b>",
38                   "<br/>","Кількість облікованих проваджень у 2022 році: ", ACC),
39               highlight = highlightOptions(weight = 3, color = "white"))%>%
40   addLayersControl(baseGroups = c("2021","2022"), )%>%
41   addControl(html = "Географічний розподіл проваджень,
42 облікованих у 2021 та 2022 роках за ознаками крадіжки",position = "bottomright")
```

Рисунок 30.16. Скрипт побудови картограми

Прокоментуємо дану частину скрипту. Команди у рядку 27, map<-leaflet()%>%setMaxBounds(20,45,41,53)%>%addProviderTiles("Esri.WorldGrayCanvas"), ініціюють створення віджету з фізичною мапою, обмежують поле відображення. Зазначимо, що в даному випадку для фонової мапи ми використали представлений у вільному доступі набір даних компанії Esri. Команди у рядках 28-33 забезпечують створення шару картограми з даними 2021 року. Команди у рядках 34-39 забезпечують створення шару картограми з даними 2022 року. Команда addLayersControl (рядок 40) забезпечує включення до віджету елементу для перемикання шарів. Команда addControl (рядки 41 та 42) забезпечує додавання до віджету текстової анотації.

Створення річних шарів забезпечується командою addPolygons, яка додає до віджету полігони областей, міст Києва та Севастополя. Охарактеризуємо параметри даної функції, використані для розв'язання кейсу:

- data – *набір даних для шару полігонів;*
- group – *назва шару, яка буде використовуватися у меню керування шарами, які відображує віджет;*
- weight – *товщина межі полігону;*
- fillOpacity – *прозорість заповнення полігону;*
- color – *колір полігону, що залежить від раніше визначеної функції pal_d;*
- popup – *зміст повідомлення, яке буде з'являтися при виборі полігону;*
- highlight – *реакція полігону при появі вказівника над ним (в нашому випадку, зміна товщини та кольору меж полігону).*

Виведення та збереження картограми. Команда map (рис. 30.17, рядок 46) забезпечує виведення отриманої картограми у вікно Viewer використовуваного нами програмного середовища.

```
45  # Виведення картограми у вікно Viewer
46  map
47
48  # Збереження картограми у файл *.html
49  saveWidget (map, "map.html")
```

Рисунок 30.17. Скрипт виведення картограми

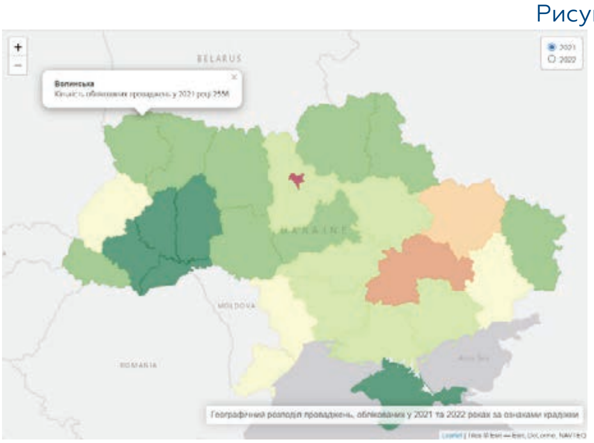

Рисунок 30.18. Картограмма за кількістю злочинів у регіонах

Використання функції saveWidget () приведе до створення файлу map.html, в якому буде збережено створений віджет.

Отримана картограма (рис. 33.18) відповідає завданням практичного кейсу:

- *візуалізує необхідні дані, кримінальних проваджень, облікованих у 2021 та 2022 році за ознаками крадіжки;*
- *є інтерактивною, вибір полігону приводить до появи інформативного вікна, а при появі вказівника над полігоном виділяються його межі;*
- *має два шари, дані 2021 року та дані 2022 року, їх перемикання забезпечується елементом у правому верхньому куті;*
- *зберігається в окремому файлі *.html.*

Повний скрипт для розв'язання цієї частини кейсу представлено у файлі map.R репозиторію R_case.

У даному практичному кейсі було продемонстровано основні можливості R щодо збирання даних, здійснення операцій з ними та візуалізації. Далі наводиться описання реалізованого нами проєкту інтерактивного довідника даних офіційної статистики щодо протидії злочинності в Україні.

РЕАЛІЗОВАНИЙ В R ПРОЄКТ З АНАЛІЗУ ДАНИХ ОФІЦІЙНОЇ СТАТИСТИКИ ЩОДО ПРОТИДІЇ ЗЛОЧИННОСТІ В УКРАЇНІ

Проблема. В Україні недостатньо інструментів аналізу публічних даних щодо протидії злочинності. Це унеможливлює проведення якісних наукових досліджень, які б ґрунтувалися на даних, обмежує можливості оцінки ефективності правоохоронної діяльності, призводить до незбалансованих змін кримінального законодавства, ускладнює моніторинг реформи кримінальної юстиції.

Базові показники кримінально-правового регулювання представлено у вигляді періодичних звітів у форматі електронних таблиць доступних для завантаження з сайтів ОГП та ДСА України. Такий формат даних істотно ускладнює роботу з ними. Отримання знань щодо ефективності протидії злочинності на підставі аналізу відкритих статистичних даних обмежено необхідністю проведення складних технічних операцій.

Відповідно, результати роботи правоохоронних та судових органів не є прозорими, механізм демократичного нагляду за їх діяльністю не є достатньо ефективним.

Запропоноване розв'язання. Автором [5] розроблено та представлено у відкритому доступі [12] веб застосунок для аналізу значної частини цього масиву інформації.

Методологію розробки застосунку характеризують юридичний, алгоритмічний та технічний складники.

Юридичний полягає в тому, що вхідні дані ми розглядаємо з усталених і найбільш поширених позицій правової науки. Зокрема, це стосується змісту використовуваних у дослідженні понять «кримінальне правопорушення», «обліковане правопорушення», «вирок, що набрав чинності» тощо. Розгляд процесу протидії злочинності зумовлений змістом аналізованих статистичних звітів.

У загальних рисах процес протидії злочинності, представлений у статистичних звітах, полягає у наступному:

1) певні події правоохоронні органи обліковують як такі, що потребують дослідження в межах кримінальних проваджень;

2) у результаті досудового розслідування таких подій може бути прийнято рішення про скерування матеріалів провадження до суду;

3) у такому разі можливі чотири варіанти: скерування матеріалів з обвинувальним актом, скерування клопотання про звільнення від кримінальної відповідальності, скерування клопотання про застосування примусових заходів медичного характеру, скерування клопотання про застосування примусових заходів виховного характеру;

4) розгляд матеріалів провадження в суді закінчується набранням чинності певним судовим рішенням;

5) такі рішення також можуть бути чотирьох різновидів: про засудження особи, про виправдання особи, про застосування примусових заходів медичного характеру та про закриття справи;

6) якщо особа визнається винною, реалізація кримінальної відповідальності може відбуватися в репресивній (призначення покарання) або ліберальній (звільнення від покарання) формах.

Представлені у відкритому доступі та використані як вхідні дані річні звіти Офісу Генерального Прокурора України та Державної судової адміністрації України є результатом покладених на ці органи обов'язків, передбачених відповідними нормативними документами [13;14].

Алгоритмічний складник стосується особливостей проведення відтворюваного дослідження (reproducible research). Збір даних відбувається в спосіб завантаження звітів, представлених на офіційних інтернет-представництвах Офісу Генерального прокурора України та Державної судової адміністрації України. Посилання на звіти додано до відповідних програмних скриптів. Виконання скриптів дає змогу отримати локальні копії вхідних даних. Трансформація даних у цьому дослідженні полягала в перетворенні звітів на дані, придатні для автоматизованого опрацювання («чисті» дані). На підставі аналізу особливостей структури вхідних звітів було розроблено скрипти очищення даних. Виконання цих скриптів дало змогу отримати файли, що містять тільки показники кримінально-правового регулювання, властиві певним статтям Особливої частини КК. У цих файлах відсутні елементи, які утруднюють автоматизовану обробку (розмітка таблиць, підсумкові рядки, заголовки тощо). Скрипти автоматизованої обробки створюють з отриманих на попередньому етапі «чистих» даних добірки та узагальнення для дальших візуалізацій. Візуалізація здійснюється знову ж таки через виконання скриптів, що використовують добірки, отримані на попередньому етапі, і стандартні бібліотеки графічного представлення даних.

Технічний складник полягає в характеристиці застосованого програмного забезпечення. Для виконання завдань дослідження використовувалася мова програмування R середа розробки програмного забезпечення з відкритим вихідним кодом RSudio. Розробку оболонки веб-застосунку здійснено з використанням Shiny R - програмного середовища для створення інтерактивних програмних продуктів з графічним інтерфейсом на основі мови програмування R. Для розробки застосунку використано такі бібліотеки (libraries) R:

Shiny [15],
readxl [16],
readr [17],
DT [18],
xlsx [19],
tidyverse [7],
textreadr [20],
stringr[21],
leaflet [22].

В онлайн-додатках представлено вихідні дані, дані, придатні до автоматизованої обробки, програмні коди для отримання первинних даних, їх обробки й візуалізації [23]. Всі використані програмні продукти та дані представлено у вільному доступі, що є додатковою гарантією можливості відтворення здійсненого дослідження

Основні характеристики веб-застосунку «Протидія злочинності в Україні: інфографіка».

Довідник представлено у вигляді веб-застосунку, який інтегрує та дає можливість аналізувати дані, що містяться у звітах Офісу Генерального прокурора України та Державної судової адміністрації:

- *період спостережень охоплює десять років (2013-2022);*
- *більше ніж 100 параметрів аналізу застосування кримінального законодавства;*
- *набір даних, який налічує більше 10 мільйонів показників функціонування національної системи кримінальної юстиції;*
- *режим створення власних візуалізацій та наборів даних, можливість комбінувати параметри, визначати період спостережень, рівень узагальнення, та необхідність обрахунку відносних значень;*
- *режим побудови фонових картограм на основі регіональних даних.*

ВИКОРИСТАНІ ДЖЕРЕЛА

1. R Core Team (2022). R: A language and environment for statistical computing. R. Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/
2. Data Analysis with R Programming. Google Data Analytics professional certification program. Retrieved from https://www.coursera.org/learn/data-analysis-r
3. Карчевський М.В. Аналіз та операції з великими даними в середовищі R (практичний кейс). Репозиторій GitHub. URL: https://github.com/Nickolay78/R_case
4. Karchevskyi M. R_case - project. Part of practical case «Big data analysis and operations in the R environment». URL: https://posit.cloud/content/6972036
5. Карчевський М.В. Протидія злочинності в Україні у форматі DATA SCIENCE. *Вісник Луганського державного університету внутрішніх справ імені Е.О. Дідоренка.* 2022. № 2(98). С. 202-227.
6. Протидія злочинності в Україні: інфографіка. Репозиторій. URL: field_n.xlsx
7. Wickham et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.
8. Cheng, J., Schloerke, B., Karambelkar, B. & Xie, Y. (2023). _leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.2.0. Retrieved from https://CRAN.R-project.org/package=leaflet
9. Pebesma, E. & Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016; Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446. Retrieved from https://doi.org/10.32614/RJ-2018-009
10. Vaidyanathan R., Xie Y., Allaire J., Cheng J., Sievert C. & Russell., K. (2023). _htmlwidgets: HTML Widgets for R. R package version 1.6.2. Retrieved from https://CRAN.R-project.org/package=htmlwidgets
11. Ukraine - Subnational Administrative Boundaries. - The Humanitarian Data Exchange. Retrieved from https://data.humdata.org/dataset/cod-ab-ukr
12. Карчевський М. Протидія злочинності в Україні: інфорграфіка: інтерактивний довідник. Версія 3.0. URL : https://karchevskiy.com/i-dovidnyk/
13. Про затвердження річних форм звітів щодо здійснення правосуддя місцевими та апеляційними судами. Наказ Державної судової адміністрації України від 23.06.2018 № 325.
14. Про затвердження форм єдиної звітності про стан кримінальної протиправності. Наказ Офіса Генерального Прокурора України від 30.06.2020 № 299.
15. Winston Chang, Joe Cheng, JJ. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert & Barbara Borges (2021). Shiny: Web Application Framework for R. R package version 1.7.1. https://CRAN.R-project.org/package=shiny
16. Hadley, Wickham & Jennifer, Bryan (2022). Readxl: Read Excel Files. R package version 1.4.0. Retrieved from https://CRAN.R-project.org/package=readxl
17. Hadley, Wickham, Jim, Hester & Jennifer, Bryan (2022). Readr: Read Rectangular Text Data. R package version 2.1.2. Retrieved from https://CRAN.R-project.org/package=readr
18. Yihui, Xie, Joe, Cheng & Xianying, Tan (2022). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.22. Retrieved from https://CRAN.R-project.org/package=DT
19. Adrian, Dragulescu & Cole, Arendt (2020). Xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.5. Retrieved from https://CRAN.R-project.org/package=xlsx
20. Rinker, T. W. (2021). Textreadr: Read Text Documents into R version 1.2.0. Buffalo, New York. Retrieved from https://github.com/trinker/textreadr
21. Hadley, Wickham (2019). Stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. Retrieved from https://CRAN.R-project.org/package=stringr
22. Cheng, J., Schloerke, B., Karambelkar, B. & Xie, Y. (2023). _leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.2.0. Retrieved from https://CRAN.R-project.org/package=leaflet .
23. Input data, scripts for collecting, cleaning, processing and building visualizations for web-application by M. Karchvskyi "Interactive reference book "Combating Crime in Ukraine" (2013-2022). Retrieved from https://github.com/Nickolay78/Ukraine_2013-2022