УДК 004.4:519.22]-057.68 (043.2)

Левченко М. Р., здобувачка вищої освіти,
Хмелівський Ю. С., асистент кафедри
інформаційних технологій

АНАЛІЗ ВИЖИВАННЯ ПАСАЖИРІВ НА БОРТУ TITANIC ЗА ДОПОМОГОЮ DATASET У МОВІ R

Донецький національний університет імені Василя Стуса, м. Вінниця

Аналіз даних – ключовий інструмент у сучасному світі для розуміння складних систем і прийняття рішень. Titanic Dataset є класичним прикладом реального набору даних, що дає змогу вивчати різноманітні аспекти виживання пасажирів під час катастрофи відомого лайнера. Цей датасет містить інформацію про демографічні, соціальні та економічні характеристики пасажирів, що допомагає досліджувати вплив різних факторів, як-от стать, вік чи соціальний статус, на шанси вижити.

Мова R, як одна з провідних мов для аналізу даних і статистичного моделювання, ідеально підходить для такого дослідження. Завдяки потужному інструментарію для візуалізації, обробки та аналізу даних R дає змогу не лише отримати інсайти, але й побудувати ефективні моделі для прогнозування.

Titanic Dataset часто використовується як навчальний інструмент для опанування методів статистичного аналізу, візуалізації даних та алгоритмів машинного навчання. Він є зручним для початківців, але також пропонує багато можливостей для глибшого аналізу [1].

1. Завантаження датасету та огляд за допомогою матриці діаграми розсіювання.

```
read.csv("/Users/marina/Desktop/stat_navych/titanic.csv")
titanic <- read.csv("/Users/marina/Desktop/stat_navych/titanic.csv")

summary(titanic)
pairs(titanic[, c("Age", "Fare", "Pclass")])
```

Рисунок 1 – Завантаження датасету

За допомогою цієї діаграми можна побачити зв’язок між змінними Age, Fare та Pclass.

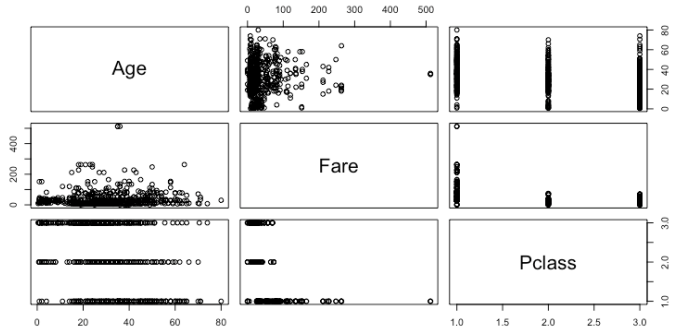

Рисунок 2 – Зв’язок між змінними

```
> summary(Titanic)
  PassengerId       Survived          Pclass          Name
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character
 Mean   :446.0   Mean   :0.3838   Mean   :2.309
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000

     Sex                 Age            SibSp           Parch
 Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
 Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
 Mode  :character   Median :28.00   Median :0.000   Median :0.0000
                    Mean   :29.70   Mean   :0.523   Mean   :0.3816
                    3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
                    Max.   :80.00   Max.   :8.000   Max.   :6.0000
                    NA's   :177
    Ticket              Fare           Cabin             Embarked
 Length:891         Min.   :  0.00   Length:891         Length:891
 Class :character   1st Qu.:  7.91   Class :character   Class :character
 Mode  :character   Median : 14.45   Mode  :character   Mode  :character
                    Mean   : 32.20
                    3rd Qu.: 31.00
                    Max.   :512.33
```

Рисунок 3 – Загальний огляд датасету, його даних

2. Введемо додаткові зміни та візуалізуємо їх [2].

```
Expensive <- rep("Ні", nrow(titanic))
Expensive[titanic$Fare > median(titanic$Fare)] <- "Так"
Expensive <- as.factor(Expensive)
titanic <- data.frame(titanic, Expensive)

plot(titanic$Expensive, titanic$Survived, main = "Виживання залежно від вартості

par(mfrow = c(2, 2))
hist(titanic$Age, breaks = 10, main = "Розподіл віку")
hist(titanic$Fare, breaks = 15, main = "Розподіл вартості квитка")
hist(titanic$Parch, breaks = 6, main = "Розподіл кількості родичів")
hist(titanic$SibSp, breaks = 9, main = "Розподіл кількості братів/сестер")
```

Рисунок 4 – Візуалізація

За допомогою цього коду будуємо графік, який показує, як залежить виживання пасажирів від категорії вартості квитка («Дорогий» чи «Недорогий») [3].

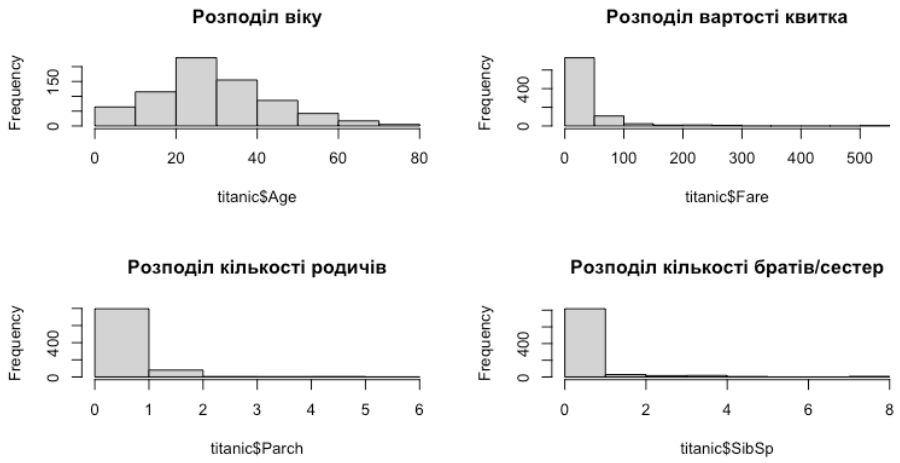

Рисунок 5 – Графіки залежностей

На основі графів можна зробити висновок, що середній вік пасажирів – 20–40 років, більшість квитків були недорогими, пасажири подорожували майже без родичів, вагома частина подорожувала подружжями [4].

3. Дослідження ключових факторів [3].

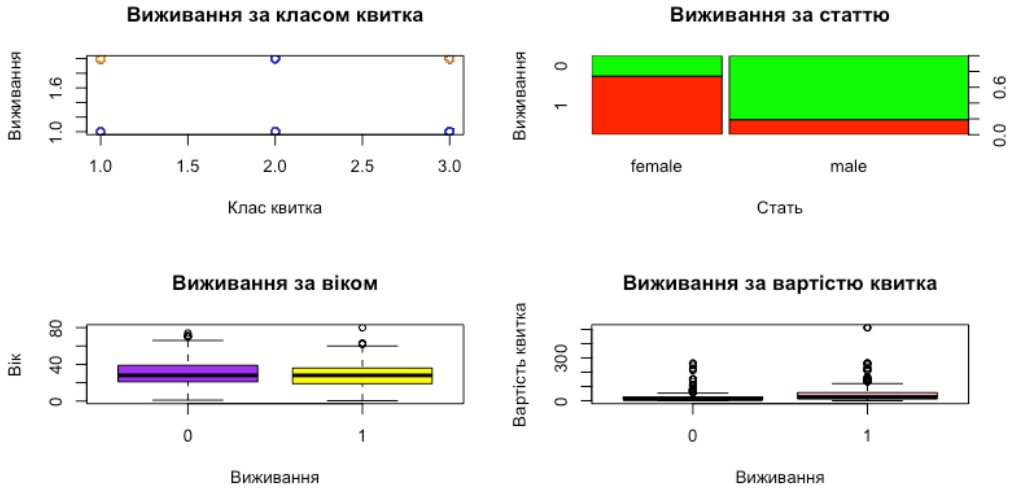

Рисунок 6 – Графік ключових факторів, які впливали на виживання

Уявімо, що ми опинилися на борту Титаніка, і ось наші шанси вижити за цією діаграмою:

Якщо у нас квитки першого класу, ми маємо більше шансів вижити. Також якщо наша стать жіноча, то шанси на виживання виростали ще більше, ніж чоловічої статі. Найбільшим везінням було б, якби ми були дітьми. І також найкраще, коли б квиток мав більшу вартість. За таких умов наші шанси на виживання були б дуже позитивними.

Результати цього аналізу показали, що мова R є потужним інструментом для проведення обробки даних, їх візуалізації та побудови статистичних висновків. Вивчення Titanic Dataset дало змогу не лише виявити закономірності в даних, але й продемонструвати важливість соціальних, економічних та демографічних факторів у виживанні людей під час надзвичайних ситуацій.

Отримані результати підтверджують ефективність використання статистичних методів для аналізу реальних подій, а також підкреслюють значення вивчення історичних даних для розуміння людської поведінки у кризових ситуаціях.

Список використаних джерел

1. DataSet. URL: https://www.kaggle.com/datasets
2. Освоюємо мову R. URL: https://uk.wikibooks.org/wiki/%D0%9E%D1%81%D0%B2%D0%BE%D1%8E%D1%94%D0%BC%D0%BE_R
3. Посібник КПІ з використанням мови R. URL: https://ela.kpi.ua/server/api/core/bitstreams/f78aa74c-7d8d-4c84-87eb-18a4aec57476/content
4. Data Science, лекція. URL: https://www.youtube.com/watch?v=DiKwo_hgFfM