~ АНАЛІЗ, АУДИТ, ОБЛІК ТА ОПОДАТКУВАННЯ ~

УДК311.21; 303.72 **DOI:10.32680/2409-9260-2023-9-310-37-43**

АНАЛІЗ ВЕЛИКИХ ОБ'ЄМІВ ДАНИХ ТА ЇХ ВІЗУАЛІЗАЦІЯ В R

Орлов Є. В., кандидат фізико-математичних наук, доцент кафедри статистики та математичних методів в економіці, Одеський національний економічний університет, м. Одеса, Україна
e-mail:orlov_ev@onu.edu.ua
ORCID ID: 0000-0002-9212-9973

Кривошеїна Є. О., студентка факультету економіки і управління підприємництвом, Одеський національний економічний університет, м. Одеса, Україна
e-mail: yelyzavetakryvosheina@gmail.com

Сіренко А. О. студентка факультету економіки і управління підприємництвом, Одеський національний економічний університет, м. Одеса, Україна
e-mail: sirenko.anna06@gmail.com

Анотація. *Стаття присвячена дослідженню методів та інструментів для аналізу та візуалізації великих об'ємів даних у R. Метою статті є обґрунтування вибору методів аналізу великих обсягів даних із використанням середовища програмування R, а також можливості їх візуалізації. Описано різні підходи до обробки великих наборів даних в R, такі як використання пакетів для роботи з великими об'ємами даних, інструментів візуалізації та аналізу даних. Наведено приклади використання середовища програмування R для комплексного аналізу даних великих компаній. Платформи соціальних медіа, такі як Twitter, Facebook і Instagram, генерують великі обсяги даних, які можна використовувати для аналізу настроїв користувачів, дослідження тенденцій і визначення реакції на новини та події.*
Ключові слова: *великі об'єми даних, середовище програмування R, статистичний аналіз, візуалізація даних.*

ANALYSIS OF LARGE AMOUNT OF DATA AND ITS VISUALIZATION IN R

Orlov Evgeniy, PhD of Physics and Mathematical Science, Associate Professor of the Department of Statistics and Mathematical Methods in Economics, Odessa National Economic University, Odess, Ukraine
e-mail:orlov_ev@onu.edu.ua
ORCID ID: 0000-0002-9212-9973

Krivosheina Elizaveta, student of Faculty of Economics and Enterprise Management, Odesa National Economic University, Odessa, Ukraine
e-mail: yelyzavetakryvosheina@gmail.com

Sirenko Anna, student of Faculty of Economics and Enterprise Management, Odessa National Economic University, Odesa, Ukraine
e-mail: sirenko.anna06@gmail.com

Abstract. *The article is devoted to the study of methods and tools for analysis and visualization of large volumes of data in the R programming environment. One of the unsolved problems in the analysis of large volumes of data and their visualization in R is the need to use specialized tools for working with large volumes of data, which can be quite difficult to master. With large volumes of data, it is important to have effective visualization tools to identify patterns and dependencies. However, automating this process remains a challenge for many developers. There is also the issue of visualizing the results, as visualizing large amounts of data can be a daunting task. Also, some graphs and charts can be too difficult to interpret and understand. Various approaches to handling large data sets in R are described, such as the use of big data packages, visualization tools, and data analysis tools. Examples of using the R programming environment for complex data analysis of large companies are given. An example of analyzing large volumes of data in R is analyzing social media data. Social media platforms such as Twitter, Facebook and Instagram generate large amounts of data that can be used to analyze user sentiment, research trends and determine reactions to news and events. To analyze social media data, you can use the R package "twitteR", which allows you to retrieve and process data from the Twitter API. For example, you can use this package to collect and analyze tweets from specific hashtags or accounts to determine user sentiment and trends. By understanding the characteristics of shopping center visitors, Uber and Twitter users, owners and management companies can make decisions about optimizing the center's operation and improving the level of service.*
Key words: *large volumes of data, R programming environment, statistical analysis, data visualization.*

JEL Classification: C810; C820; C870.

Постановка проблеми. У сучасному світі величезна кількість даних збирається щодня у різних сферах діяльності. Однак обробка та аналіз великих обсягів даних може бути складним завданням, особливо без використання спеціалізованих інструментів. У цьому контексті стає все більш актуальним використання засобів аналізу даних та візуалізації, таких як середовище програмування R для обробки великого обсягу даних і вилучення цінної інформації з них.

Аналіз останніх досліджень і публікацій. Ряд досліджень, проведених останніми роками, свідчить про те, що використання R для аналізу даних може призвести до поліпшення якості прийнятих бізнес-рішень та підвищення ефективності роботи у різних сферах. Один із прикладів аналізу великого обсягу даних та їх візуалізації в R пов'язаний із дослідженням, проведеним у 2018 році в Університеті штату Монтана та опублікованим у журналі "PLoS ONE". У цьому дослідженні автори використовували R для аналізу даних, пов'язаних із клінічними випробуваннями з терапії хвороби Паркінсона [1, с. 7-12]. Проте нині є потреба у комплексному дослідженні методів аналізу великих обсягів даних, їх оптимізації та візуалізації у середовищі програмування R, з урахуванням новітніх тенденцій та технологій.

Відокремлення невирішених раніше частин загальної проблеми. Однією з невирішених проблем в аналізі великих об'ємів даних та їх візуалізації в R є необхідність використання спеціалізованих інструментів для роботи з великими об'ємами даних, які можуть бути досить складними для освоєння. При великих обсягах даних важливо мати ефективні інструменти візуалізації для виявлення закономірностей та залежностей. Однак автоматизація цього процесу залишається викликом для деяких розробників. Також виникає проблема візуалізації результатів, оскільки візуалізація великих об’ємів даних може бути непростим завданням. Крім того, деякі графіки та діаграми можуть бути надто складними для інтерпретації та розуміння.

Мета дослідження. Мета статті – розглянути методи аналізу великих обсягів даних із використанням середовища програмування R, а також можливості їх візуалізації. Зокрема, йдеться про можливості використання R для дисперсійного, регресійного, кластерного та інших методів статистичного аналізу та про процес візуалізації даних з використанням різних графічних інструментів доступних в R.

Основний матеріал. R – це інструмент статистичного програмування, який має широкий спектр унікальних можливостей для обробки даних.

R полегшує роботу з даними з різних джерел, від імпорту до аналізу. Крім того, сама система R і бібліотека CRAN пропонують безліч функцій і інструментів для візуалізації даних, що дозволяє професіоналам легко представляти свої дослідження та висновки в ефектному та легкому для читання форматі [2].

Усі дані, зібрані для будь-якого аналізу, є корисними, якщо вони представлені так, щоб вони були легко зрозумілими для всіх і допомагали приймати правильні рішення. Після того, як ми проводимо аналіз даних, ми окреслюємо їх короткий зміст, щоб зрозуміти його набагато краще. Це відомо як узагальнення даних.

Середовище програмування мовою R побудовано навколо стандартного інтерфейсу командного рядка. R Studio підтримує велику кількість пакетів та інструментів, що дозволяє виконувати різноманітні задачі з аналізу даних та машинного навчання.

Користувачі використовують це, щоб читати дані та завантажувати їх у робочу область, вказувати команди та отримувати результати. Команди можуть бути будь-якими: від простих математичних операторів, включаючи +, -, * і /, до більш складних функцій, які виконують лінійну регресію та інші складні обчислення [3]. Однією з найзручніших можливостей R Studio є редактор коду, який дозволяє писати код на мові R зі зручними функціями автодоповнення, корекції помилок та відступів. Крім того, R Studio має вбудовану систему контролю версій git, що дозволяє зберігати та контролювати зміни в коді.

Перед початком роботи з Rstudio потрібно чітко визначитися з тією інформацією, яка буде проаналізована. Однією з головних переваг R Studio є підтримка R Markdown, що дозволяє створювати звіти та документацію з кодом на мові R, які можна легко експортувати в різні формати, такі як HTML, PDF та Word [4]. R має широкий спектр можливостей для роботи з базами даних, включаючи імпорт, експорт, об'єднання та фільтрацію даних. Це дозволяє спростити процес обробки та аналізу.

Наступним кроком після вибору інформації є встановлення пакетів.

R Studio має вбудовані та додаткові пакети, які можна легко встановлювати з використанням командного рядку, який вже вбудовано в R Studio, тобто не потрібно заходити в браузер,

шукати ці пакети та вказувати шлях встановлення. Завантажені пакети можуть містити різноманітні функції та методи для обробки та аналізу даних, графічних інтерфейсів користувача для відображення та візуалізації даних, а також приклади та документацію для використання цих інструментів. Для їх установки в R Studio використовується функція install.packages(), яка завантажує пакети з відповідних репозиторіїв. Після встановлення пакетів їх можна використовувати в своїх програмах та скриптах.

Розробники мають змогу створювати власні пакети для вирішення конкретних завдань з аналізу даних, що дозволяє розширити функціональність R Studio та виконувати складніші завдання, у тому числі й у сфері бізнесу.

Зараз, коли ми маємо доступ до великої кількості даних, аналіз та візуалізація цих даних стають важливими інструментами для розвитку бізнесу. Наприклад, у сфері сервісів таксі, таких як Uber, дані про поїздки можуть бути використані для належного використання ресурсів, таких як транспортні засоби та водії. Це допомагає Uber управляти своїми резервами та планувати маршрути, що зменшує час очікування та підвищує задоволеність клієнтів.

Також R може використовуватися для прогнозування попиту на послуги Uber, що допомагає компанії забезпечити належну кількість транспортних засобів у певні періоди часу. Загалом це покращує ефективність сервісу та підвищує рівень задоволеності користувачів.

Для аналізу та візуалізації даних про поїздки в Uber ми можемо використовувати мову програмування R, а саме пакети ggplot2 і dplyr та функцію ggmap. (рис. 1).

Використання функцій ggplot2 та ggmap дозволяє побудувати карту Нью-Йорка з позначенням місць, звідки стартували поїздки Uber протягом 2014 року (квітень-вересень). Крім того, використання функції dplyr дозволяє згрупувати дані за годинами дня, а також за днями тижня, що дає змогу проаналізувати паттерни використання Uber у різний час. Внаслідок такої візуалізації стає можливим зрозуміти, як змінюється популярність Uber у різних районах міста у різний час, а також визначити піки популярності та найбільш завантажені години для поїздок.

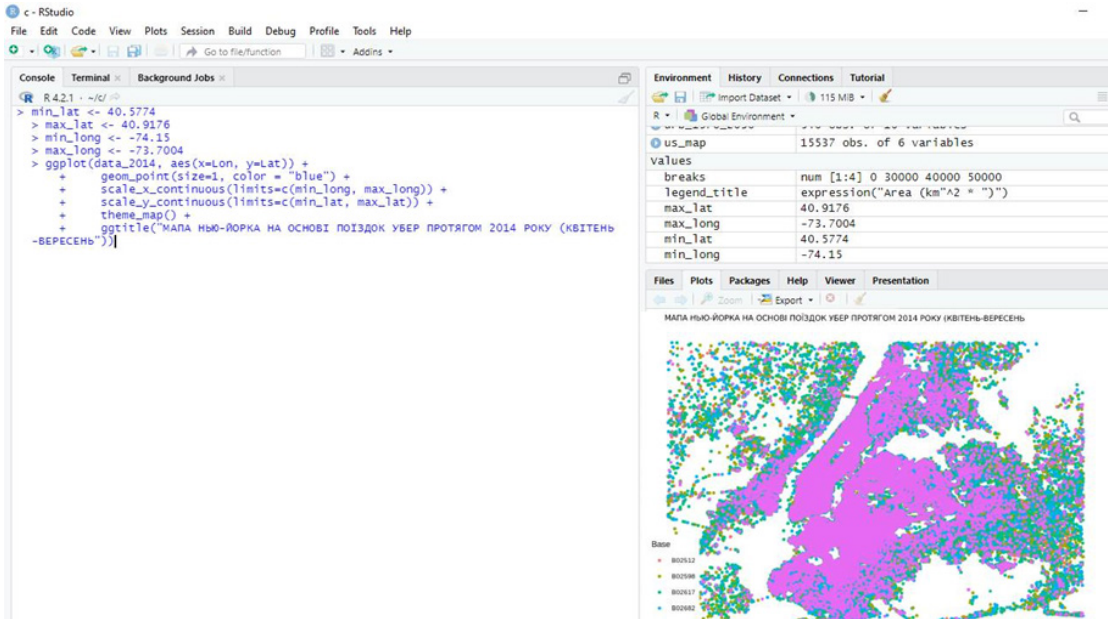

Рис. 1. Графік у вигляді мапи ggplot2

Джерело: побудовано авторами

Досить простим, але водночас доступним та зрозумілим способом візуалізації даних є секторна діаграма. У R для побудови секторної діаграми використовують функцію pie(). Наприклад, якщо маємо вектор x зі значеннями кожного сектора, то використовуючи pie(x), ми отримаємо секторну діаграму з відповідними розмірами секторів. Для кращої візуалізації можна використовувати додаткові параметри, такі як кольори, назви секторів, відстані між ними тощо.

Однією з основних особливостей секторної діаграми є її здатність показати співвідношення

між категоріями відносно всього. Кругла форма діаграми дозволяє легко порівнювати розміри секторів і зрозуміти, яка частина від загальної кількості даних представлена кожним сектором. Використання такої діаграми може стати незамінним інструментом при аналізі даних відвідувачів торгівельного центру, а саме для зображення розподілу даних відвідувачів за критерієм «стать» (рис. 2).

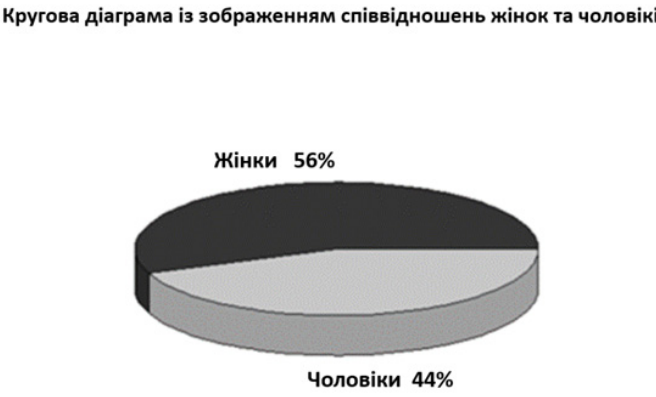

Рис. 2. Секторна діаграма pie()

Джерело: побудовано авторами

Секторна діаграма дозволяє виділити окремі сегменти даних, що є особливо важливим для візуалізації даних, які мають значні відмінності в розмірах. Вона також може використовуватися для відображення змін у часі або розподілу даних на різних рівнях групування.

Одним з найбільш поширених способів візуалізації даних є зображення даних у вигляді стовпчастої гістограми. Для цього використовують функцію barplot() (рис. 3), яка дозволяє легко відобразити велику кількість даних у зручному для сприйняття форматі та зробити висновки про тенденції та залежності між ними. Функція barplot() дозволяє відобразити кількість спостережень на вісі Y та їх категорії на вісі X. Кожен стовпчик на гістограмі представляє окрему категорію, а його висота відображає кількість спостережень, що відповідає цій категорії. За допомогою різних параметрів функції barplot() можна налаштувати колір, ширину та інші візуальні характеристики гістограми, що дозволяє зробити її більш зрозумілою та привабливою для аудиторії.

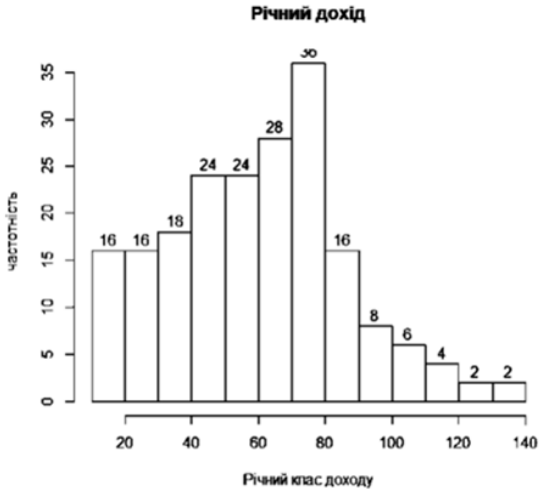

Рис. 3. Гістограма barplot()

Джерело: побудовано авторами

Важливим елементом під час використання функції barplot() є функція table(). Функція table() використовується для підрахунку частотних таблиць, тобто кількості входжень кожного унікального значення у векторі або списку даних. Функція повертає таблицю, яка містить кількість входжень кожного унікального значення у вхідний вектор або список. Отже, за допомогою функції table() можна швидко та легко отримати інформацію про розподіл даних та використовувати її для подальшого аналізу.

Дані функції можна використати для оцінки та аналізу відвідувачів торгівельного центру. Ми можемо використати функцію table(), щоб підрахувати кількість відвідувачів кожної статі, а потім передати ці дані у функцію barplot(), щоб побудувати графік, який показує кількість відвідувачів за статтю. Завдяки цій стовпчастій гістограмі можна легко оцінити статеву приналежність відвідувачів. Це дозволяє нам зробити висновки про тенденції та залежності між статевим складом відвідувачів та їх кількістю.

Іншим зручним інструментом є функція boxplot() (рис. 4), завдяки їй можна побудувати коробковий графік або графік розмаху.

Рис 4. Коробковий графік boxplot()

Джерело: побудовано авторами

«Коробковий графік» – це швидкий спосіб вивчення одного або декількох наборів даних у графічному вигляді. «Коробковий графік» може здатися примітивнішим за гістограму, але цей метод має деякі переваги. По-перше, коробковий графік дозволяє зобразити розподіл даних на основі п'яти числових характеристик (мінімум, перший квартиль, медіана, третій квартиль та максимум). По-друге, коробкові графіки дозволяють порівняти розподіли даних між різними категоріями, що дозволяє виявити відмінності та подібності між ними. По-третє, коробковий графік займає менше місця і тому особливо корисний для порівняння розподілу між кількома групами або наборами даних [5, с. 52-55].

Коробковий графік є корисним інструментом для аналізу різних типів даних, у тому числі даних про відвідувачів торгівельного центру [6,7]. Коробкові графіки зручні для порівняння розподілу даних між декількома групами, що допомагає виявляти взаємозв'язки між змінними та проводити статистичні тестування.

За допомогою коробкового графіка можна проаналізувати розподіл числових даних (наприклад, витрати відвідувачів на покупки), виявити наявність викидів або незвичайних значень, порівняти розподіл між різними групами відвідувачів (наприклад, чоловіки та жінки), а також зробити висновки про типові значення та варіативність витрат [8].

Наприклад, якщо ми використовуємо коробковий графік для аналізу витрат відвідувачів за різними днями тижня, то ми можемо порівняти медіанні та середні витрати, виявити найбільші та найменші витрати, а також визначити, чи є значні відмінності між різними днями [9, 10].

Іншим прикладом аналізу великого обсягу даних в R є аналіз соціальних медіа-даних. Соціальні медіа-платформи, такі як Twitter, Facebook та Instagram, генерують велику кількість даних, які можна використовувати для аналізу настроїв користувачів, дослідження трендів та визначення реакції на новини та події. Для аналізу соціальних медіа-даних можна

використовувати пакет R "twitteR", який дозволяє отримувати та оброблювати дані з Twitter API. Наприклад, можна використовувати цей пакет для збору та аналізу твітів з певних хештегів або облікових записів, щоб визначити настрої користувачів та тренди.

Зрозумівши характеристики відвідувачів торгівельного центру, користувачів сервісу Uber та Twitter, власники та управляючі компанії можуть приймати рішення щодо оптимізації роботи центру та покращення рівня обслуговування.

Висновки. Різні мови програмування сумісні з процесом обробки даних, але R робить оцінку та збір даних захоплюючим і унікальним. R – це вдосконалена мова, яка виконує різні складні статистичні обчислення. Тому саме він широко використовується фахівцями з обробки даних і бізнес-лідерами в багатьох галузях.

Використання мови програмування R є важливим інструментом для бізнесу, що дозволяє ефективно управляти торговою діяльністю, а також прогнозувати та планувати розвиток бізнесу. Завдяки точному аналізу великої кількості даних можна оптимізувати розміщення магазинів та рекламних матеріалів, аналізувати ефективність маркетингових акцій, встановлювати оптимальні режими роботи, залежно від того, коли відвідувачі найбільш активні, та багато іншого.

Використання даних функцій та графіків у R дозволяє проводити статистичний аналіз даних, виявляти залежності та тенденції, знаходити випадки аномалій та визначати причини їх виникнення. Це може неабияк допомогти у розв'язанні проблем та виявленні нових можливостей для розвитку компанії.

R є незамінним інструментом, завдяки якому можна робити все, що завгодно, коли справа стосується даних або науки про дані. Майбутнє мови програмування R є яскравим, адже великі бізнес-організації віддають перевагу платформам, інструментам і технологіям саме з відкритим кодом для аналізу масивних критичних даних.

Список літератури

1. Evaluation of cerebrospinal fluid proteins as potential biomarkers for early stage Parkinson's disease diagnosis / D. Scheller et al. PLoS ONE. 2018. С. 7–12.
2. The Comprehensive R Archive Network. URL: https://cran.r-project.org/ (дата звернення 28.04.2023).
3. Paradis E. R for beginners. The Comprehensive R Archive Network. URL: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf / (дата звернення 28.04.2023).
4. R Markdown. URL: https://rmarkdown.rstudio.com/lesson-1.html (date of access: 29.04.2023).
5. Dalpiaz D. Applied Statistics with R. Illinois, 2016. 417 p. URL: https://dokumen.tips/documents/applied-statistics-with-r-david-dalpiaz-collinearity-287-welcome-to-applied.html?page=1 С. 52–55. (дата звернення 28.04.2023).
6. Hodeghatta, U. R. Practical business analytics using R and Python: solve business problems using a data-driven approach. Apress Berkeley. 2023. 706 p. https://doi.org/10.1007/978-1-4842-8754-5
7. de Micheaux P. L., Drouilhet R., Liquet B. The R Software. Fundamentals of Programming and Statistical Analysis. Springer New York. 2016. 628 p. DOI https://doi.org/10.1007/978-1-4614-9020-3
8. Fox J., Weisberg S. An R Companion to Applied Regression. SAGE Publications. 2018. 608 p.
9. Tattar P. N., Ramaiah S., Manjunath B. G. A Course in Statistics with R. Wiley. 2016. 696 p.
10. Chambers J. M. Software for data analysis. Springer. 2008. 500 p. DOI https://doi.org/10.1007/978-0-387-75936-4

References

1. Evaluation of cerebrospinal fluid proteins as potential biomarkers for early stage Parkinson's disease diagnosis (2018). D. Scheller et al. PLoS ONE.
2. The Comprehensive R Archive Network. Retried from https://cran.r-project.org/.
3. Paradis, E. R for beginners. The Comprehensive R Archive Network. Retried from https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

4. R Markdown. Retried from https://rmarkdown.rstudio.com/lesson-1.html.
5. Dalpiaz, D. (2016). Applied Statistics with R. Illinois. Retried from https://dokumen.tips/documents/applied-statistics-with-r-david-dalpiaz-collinearity-287-welcome-to-applied.html?page=1/
6. Hodeghatta , U. R. (2023). Practical business analytics using R and Python: solve business problems using a data-driven approach. Apress Berkeley. Retried from https://doi.org/10.1007/978-1-4842-8754-5
7. de Micheaux P. L., Drouilhet R., Liquet B. (2016). The R Software. Fundamentals of Programming and Statistical Analysis. Springer New York. Retried from https://doi.org/10.1007/978-1-4614-9020-3
8. Fox, J., Weisberg, S. (2018). An R Companion to Applied Regression. SAGE Publications.
9. Tattar, P. N., Ramaiah, S., Manjunath, B. G. (2016). A Course in Statistics with R. Wiley.
10. Chambers, J. M. (2008). Software for data analysis. Springer. Retried from https://doi.org/10.1007/978-0-387-75936-4

Стаття надійшла до редакції 13.09.2023
Прийнята до публікації 19.09.2023