And Advanced Technology (IJERAT). 2016. Vol. 02. Issue. 12. URL: https://ijerat.com/uploads/2/3222_pdf.pdf (дата звернення: 29.10.2019).

6. Sucharitha V., Subash S.R., Prakash P. Visualization of Big Data: Its Tools and Challenges. *International Journal of Applied Engineering Research.* 2014. Vol. 9(18). P. 5277-5290.

ВИКОРИСТАННЯ БІБЛІТЕКИ GGPLOT2 ДЛЯ ВІЗУАЛІЗАЦІЇ ДАНИХ

Марець Оксана Романівна,
кандидат економічних наук, доцент кафедри статистики,
Львівський національний університет імені Івана Франка

Одним із популярних нині інструментів для візуалізації та аналізу статистичних даних є середовище та мова програмування R. Найбільш універсальна сфера їх застосування – це аналітика. R дає можливість проводити статистичні тести, перевіряти гіпотези, будувати графіки та робити прогнози. Перевагами використання мови програмування R для аналізу є такі:

– це програмне забезпечення з відкритим кодом;

– мова програмування з відносно інтуїтивним кодом, оскільки створена для аналізу даних;

– відтворюваність результатів аналізу, тобто будь-хто може подивитися на код і з’ясувати, як був зроблений аналіз;

– повторюваність результатів аналізу: якщо з’являються нові дані, можна легко запустити вже написаний код;

– дружня спільнота з блогами, статтями, форумами, де дадуть відповідь на питання різного рівня складності;

– наявність великої кількості бібліотек, призначених для виконання найрізноманітніших специфічних задач.

Одна з цих бібліотек – ggplot2 – представляє засоби для візуалізації даних і нині її вважають серед найкращих у цій сфері. Вона є частиною системи tidyverse – комплексу бібліотек, призначених для завантаження, обробки, очищення, візуалізації та аналізу даних. Основними бібліотеками tidyverse є ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr та forcats, які надають інструменти для моделювання, перетворення та візуалізації даних.

Хочемо зазначити, що безпосередній візуалізації передує важливий і часто трудомісткий етап роботи з завантаження та підготовки даних. Саме для легкого виконання цих процесів призначені бібліотеки системи tidyverse. Ця система базується на принципах чистих даних (tidydata) та граматики графіки. *Чисті дані* (tidydata) – набори даних, де кожна змінна (показник) – це стовпець, а кожне спостереження (одиниця сукупності) – рядок. Термін *граматика графіки* запропонував Л.Уілкінсон у 1999 р. Вона охоплює два

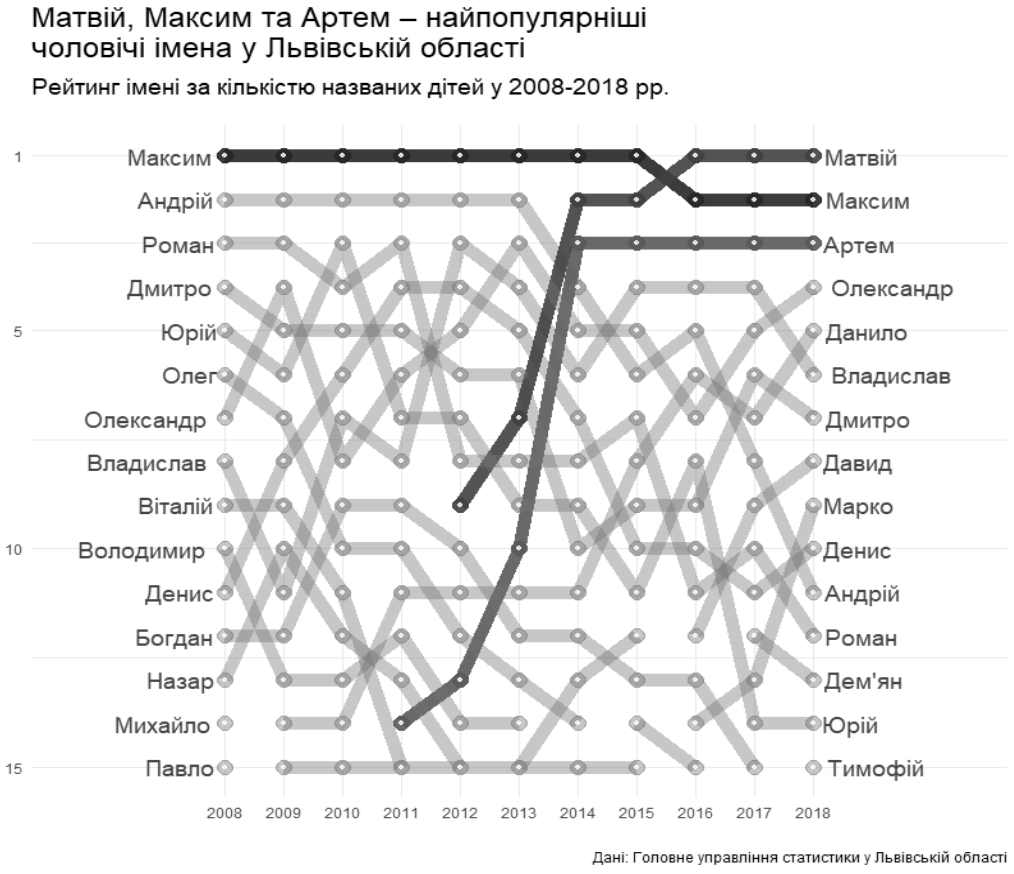
принципи: 1) нашарування граматичних елементів; 2) створення візуалізацій через естетичні відображення (aesthetic mappings).

Отже, створення діаграми ggplot2 – це робота з шарами графічних елементів. У бібліотеку ggplot2 вбудовано такі графічні елементи:

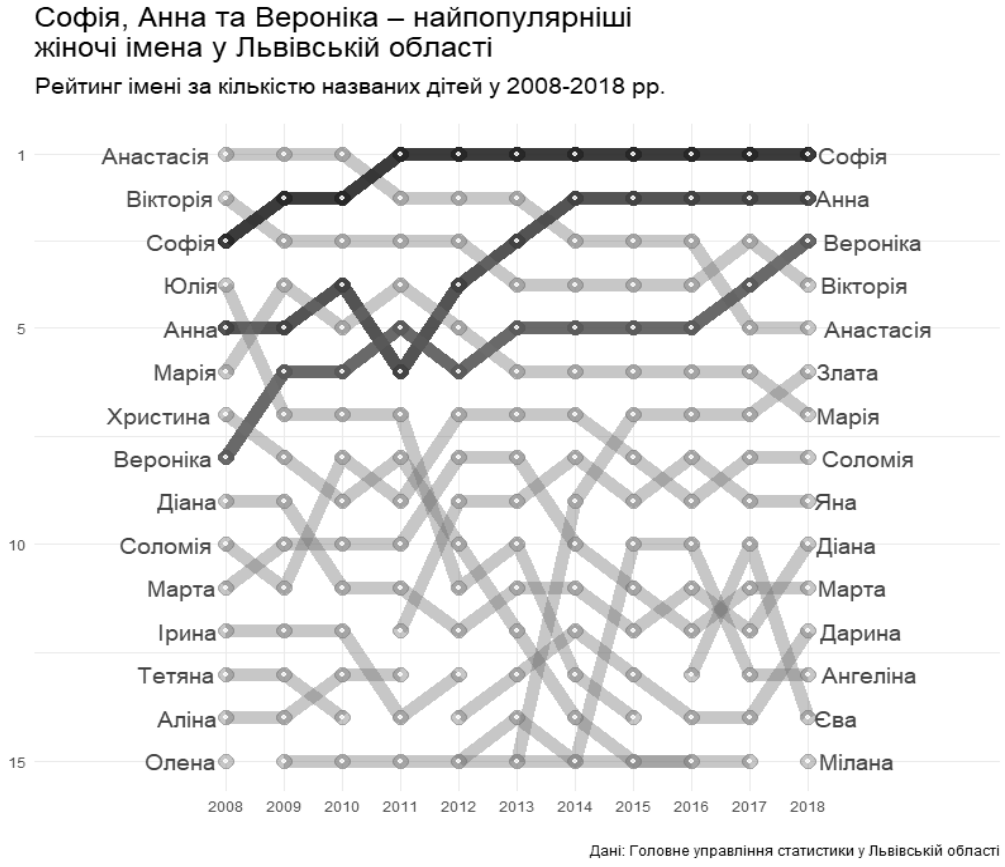
- дані (data) – набір даних, на основі якого створюється візуалізація;
- естетики (aesthetics) – узгодження різних розмірностей даних та їх адаптація під площу графіка. Тут ми вказуємо, наприклад, що відкладаємо на осях;
- геометрія (geometries) – візуальні елементи. Для лінійної діаграми вказуємо geom_line, для гістограми – geom_histogram тощо;
- стиль (themes) – загальний вигляд діаграми;
- статистика (statistics) – використання статистичних інструментів для покращення читабельності даних;
- координати (coordinates) – задання простору для розміщення геометричних елементів
- панелі (facets) – інструмент для побудови мультиграфіків.

Зазначимо, що перші три з перелічених елементів є необхідними шарами для побудови діаграм у ggplot2.

Покажемо можливості використання бібліотеки ggplot2 на прикладі побудови двох нестандартних діаграм: графік-зигзаг (bump-chart) та графік-льодяник (diverging lollipop). Перший призначений для візуалізації зміни рангів у часі (рис. 1, а), б)).



а)



б)

Рис. 1. Графік-зигзаг (bump-chart), виконаний у середовищі R за допомогою бібліотеки ggplot2

З його допомогою можна візуалізувати досить велику кількість даних. Зокрема, основою однієї з діаграм на рис. 1 є таблиця розмірністю 165x4. Такою діаграмою легко показати зміну рангів областей, районів країн.

Графік-льодяник – це альтернатива стовпчиковим діаграмам для візуалізації якісної та кількісної змінних, де прямокутник перетворений в лінію та точку. Льодяник виглядає більш сучасно та привабливо, порівняно зі стовпчиковою діаграмою він є легшим для сприйняття. Наприклад, коли дані приблизно однакові, стовпці також будуть приблизно однакової висоти, що може створити ефект муару. Натомість лінія з точкою буде значно менше захаращувати простір. Крім того, цю діаграму можна трансформувати, щоб показати різні акценти в даних. Наприклад, на рис. 2 представлений графік-льодяник для зображення відхилень від середнього (diverging lollipop). Анотації на діаграмі ілюструють її основні акценти.

Зосередимо увагу на питанні перекладу англомовних термінів. З одного боку, при пошуку інформації про ці діаграми зручно знати саме англійські варіанти їх назв. Проте, щоб не засмічувати мову, варто підбирати й і українські відповідники. У нашому випадку ми не дослівно переклали *bumpchart* як *зигзаг*, бо діаграма за формою нагадує саме цю ламану лінію. Назву діаграми *lollipop chart* ми переклали дослівно – *льодяник*.

Рис. 2. Графік-льодяник для відхилень від середнього (diverging lollipop), виконаний в середовищі R за допомогою бібліотеки ggplot2

Отже, ми перелічили переваги використання програмного середовища та мови програмування R та представили можливості бібліотеки ggplot2 на прикладі побудови двох нетрадиційних графіків. Перевагою використання таких графіків є їх розширені можливості порівняно з традиційними діаграмами. Але все ж варто зважати на потреби користувачів інформації, поданої у такому вигляді, оскільки не завжди нестандартні рішення сприймаються добре.