Брошура навчального курсу з журналістики даних
для викладачів журналістики у межах проєкту

ВИКОРИСТАННЯ ДАНИХ У БОРОТЬБІ ПРОТИ МАНІПУЛЯЦІЙ ТА ДЕЗІНФОРМАЦІЇ

TEXTY.ORG.UA

AMO.CZ

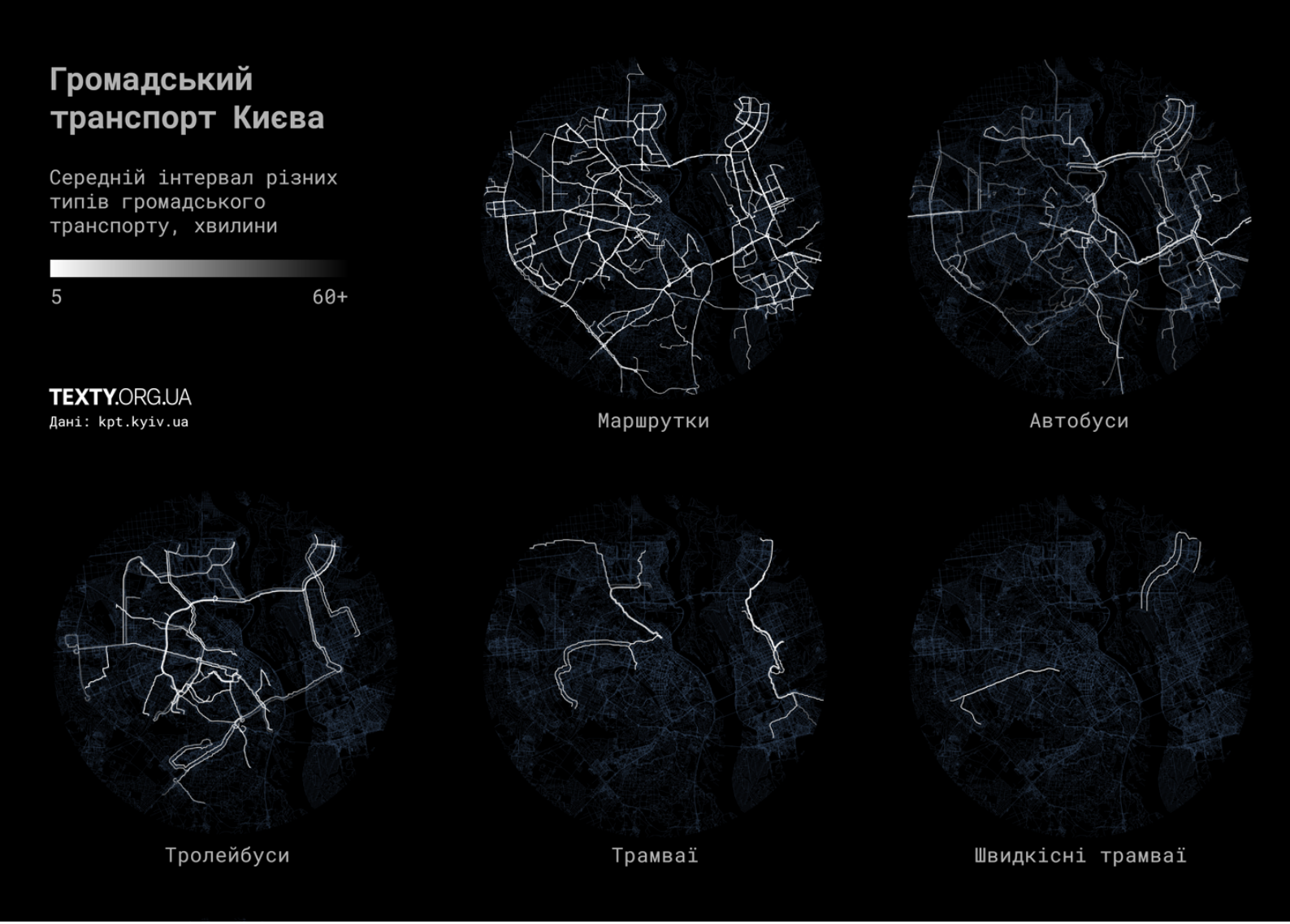
TRANSITION
Ministry of Foreign Affairs of the Czech Republic

ЖУРНАЛІСТИКА ДАНИХ — ЩО ЦЕ?

Це журналістика, яку роблять за допомогою даних*. Останніми можуть бути таблиці, наприклад, державна статистика; зв’язки у соціальних мережах; масиви зображень, як-от супутникові знімки землі; текстові дані, приміром, архіви сайтів новин або стенограми засідань парламенту. Журналістика даних стає можливою завдяки розвитку та демократизації технологій роботи з даними, значною мірою програмуванню.

Журналістика даних — це місце зустрічі журналістики з іншими галузями: наукою про дані (data science), графічним дизайном, картографією.

Більшість матеріалів журналістики даних використовують візуалізацію даних, або інфографіку:



На прикладі візуалізовані дані Київпастрансу, які показують покриття маршрутів та інтервали громадського транспорту. Це інфографіка.

Інфографіками часто називають пояснювальні матеріали, в яких дані не є центральним елементом або взагалі не використовують. Хоч у повсякденному словнику постери або інформаційні листівки називають інфографікою, ми будемо надалі визначати інфографіку саме як візуалізацію даних.
Приклад не-інфографіки:

ІСТОРІЇ, РОЗКАЗАНІ ЗА ДОПОМОГОЮ ДАНИХ

Візуалізація даних може бути як основною частиною матеріалу, так і допоміжною. Вона може акцентувати увагу на чомусь, привернути увагу читача, надавати факти та докази. Залежно від мети матеріалу й цільової аудиторії, візуалізація може бути інструментом для читачів (дешборд) або складеною авторами послідовністю (скролітелінг).

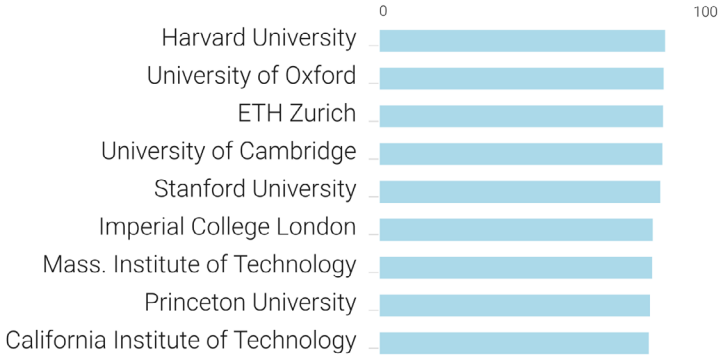
Де шукати історії, розказаних за допомогою даних: Texty.org.ua, Pudding, Propublica, South China Morning Post (інфографіка та постери), Flowingdata (блог), FiveThirtyEight, Федеріка Фрагапане, The Economist, Reuters, datajournalism.com, The Functional Art (блог), Visualising Data

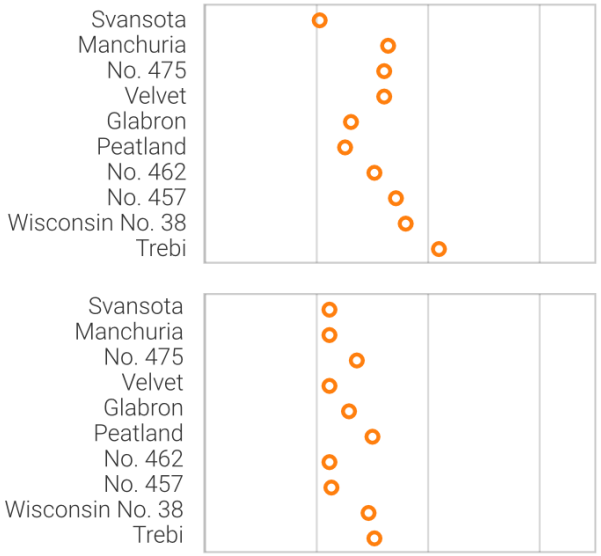
ТИПИ ВІЗУАЛІЗАЦІЙ

Вибір типу візуалізації залежить від типів даних і зв'язків, які ми хочемо показати.

НОМІНАЛЬНЕ ПОРІВНЯННЯ

Коли одна зі змінних є категорією, а інша — числова, наприклад, країни та середня очікувана тривалість життя у кожній з них, підходять стовпчикові або точкові графіки:

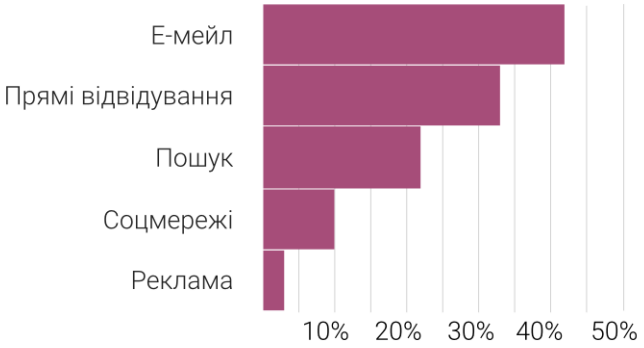




У стовпчикових графіках шкала має брати початок від нуля, а у точкових це необов'язково. Категорійна змінна має бути відсортована, якщо це важливо, наприклад, за найбільшим значенням чи за абеткою.

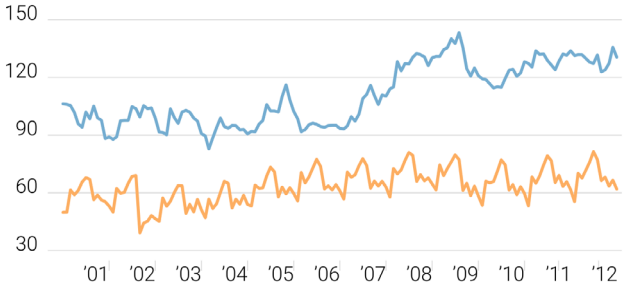
СПІВВІДНОШЕННЯ ЧАСТКИ ТА ЦІЛОГО

Найкращим варіантом є стовпчикові графіки, адже вони дають змогу оцінити розмір частин. Кругові (млинцеві) діаграми краще не використовувати.



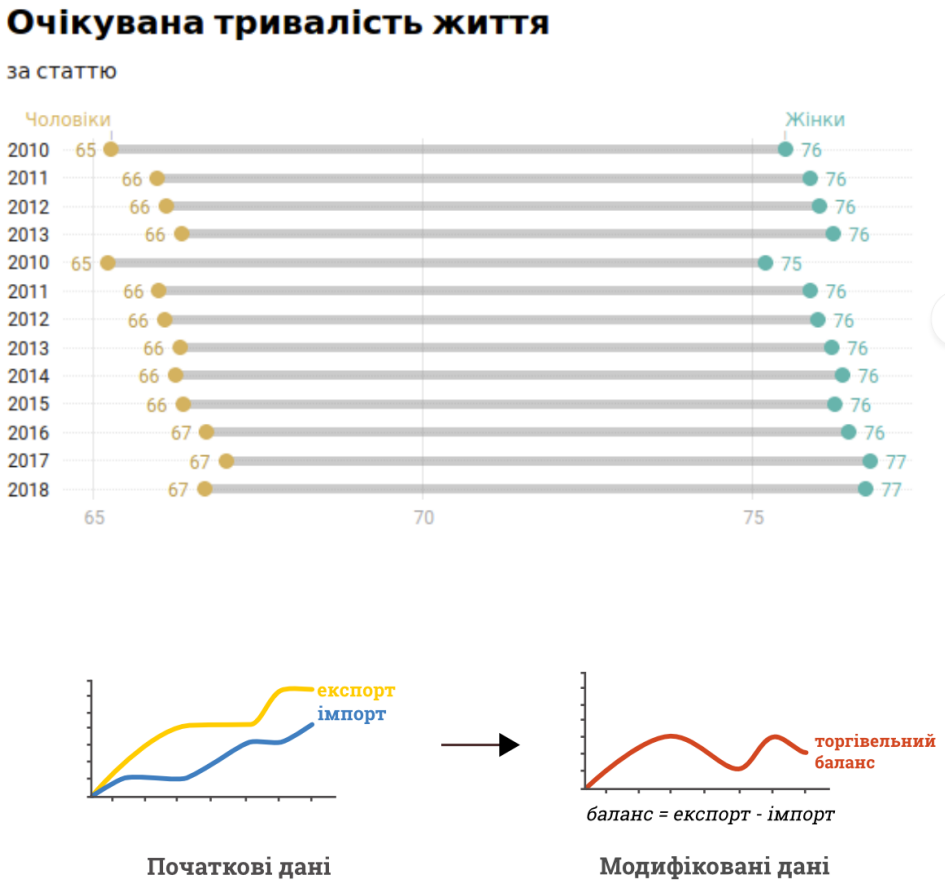
ДИНАМІКА У ЧАСІ

Аби показати, як щось змінювалось у часі, використовують лінійні графіки:



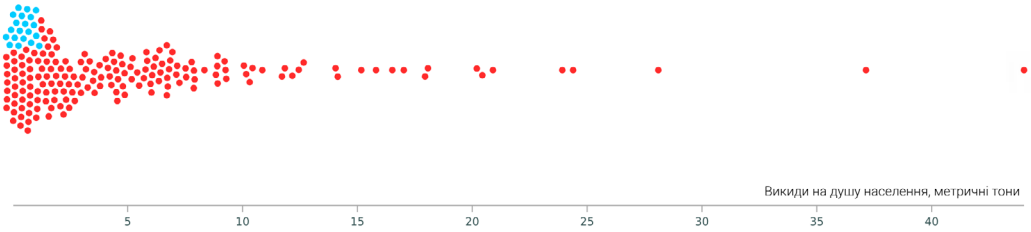
ВІДХИЛЕННЯ

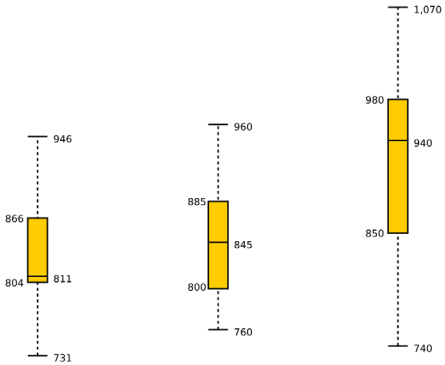
Щоб показати співвідношення між двома величинами, можемо застосувати точкові або лінійні графіки:



РОЗПОДІЛ

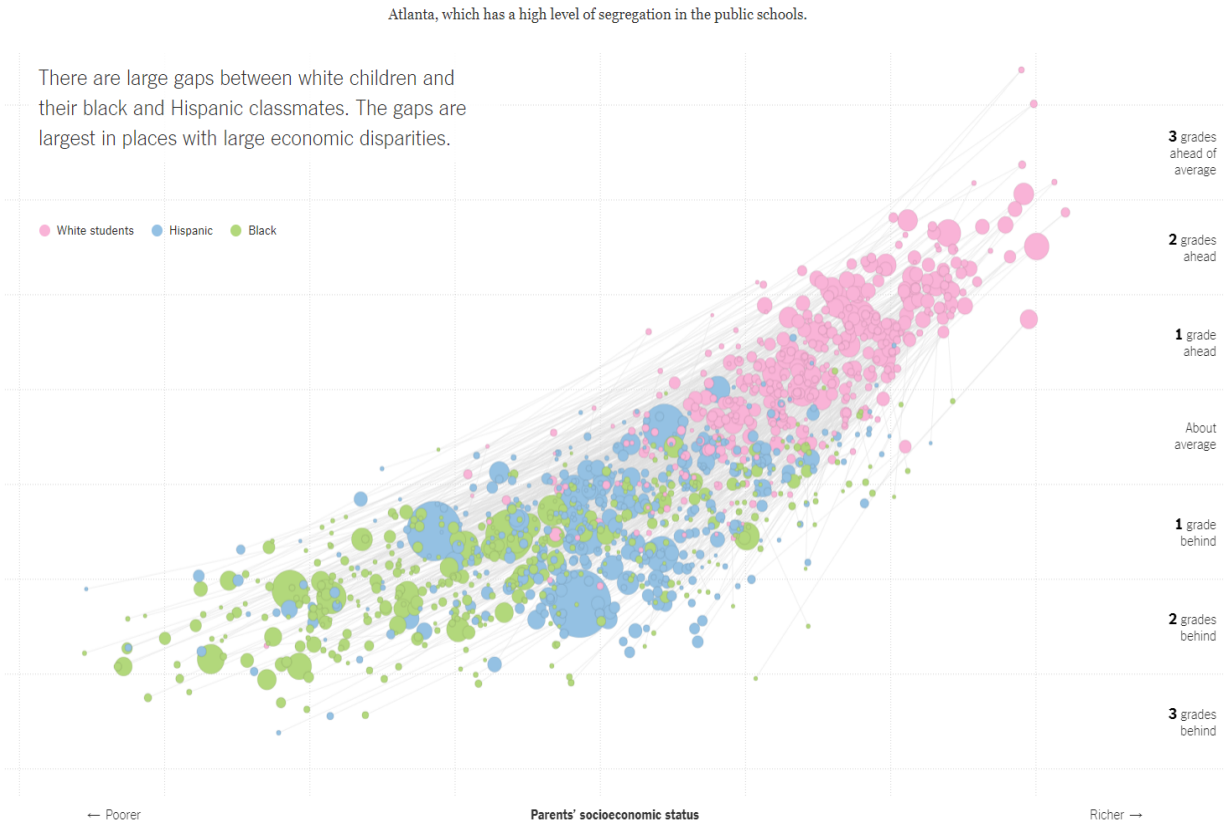
Коли треба показати середнє та розмах однієї величини, використовуємо гістограми, beeswarm або боксплоти:



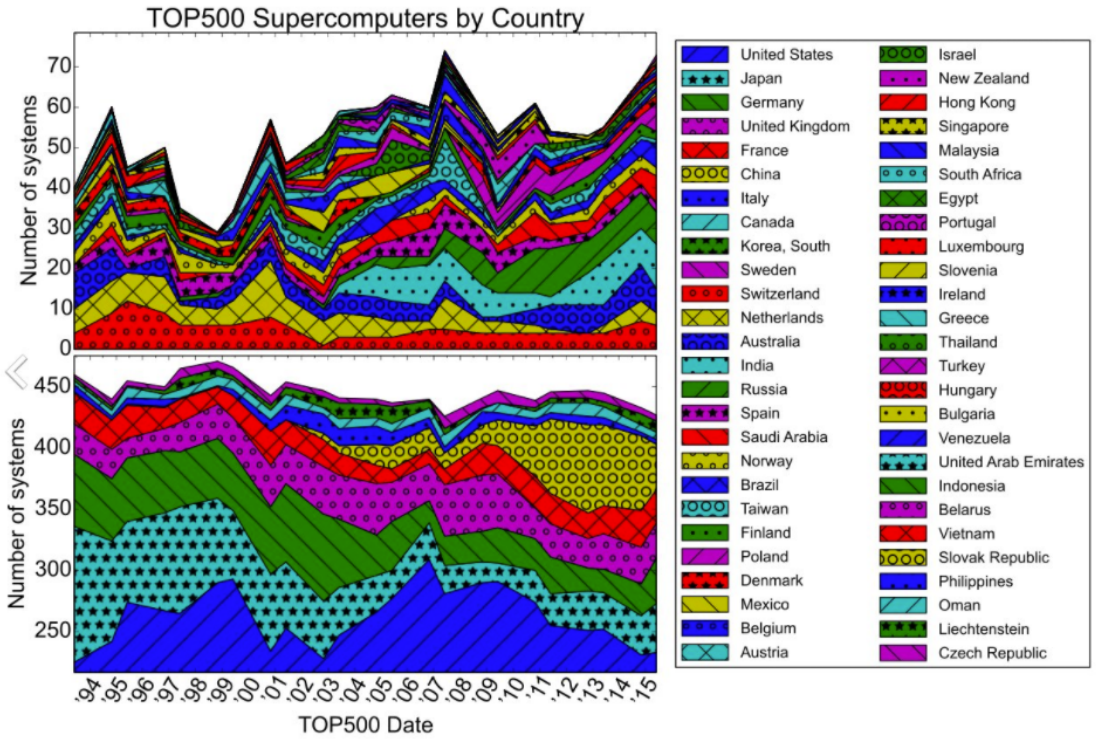


ЗВ’ЯЗОК МІЖ ДВОМА ЗМІННИМИ

Коли маємо дві числові змінні та хочемо показати, як вони (не)пов’язані, використовуємо графік розсіювання, scatterplot. Можемо позначити другорядну третю змінну розміром, якщо вона числова, або кольором, якщо категорійна.



ТИПОВІ ПОМИЛКИ ВІЗУАЛІЗАЦІЇ



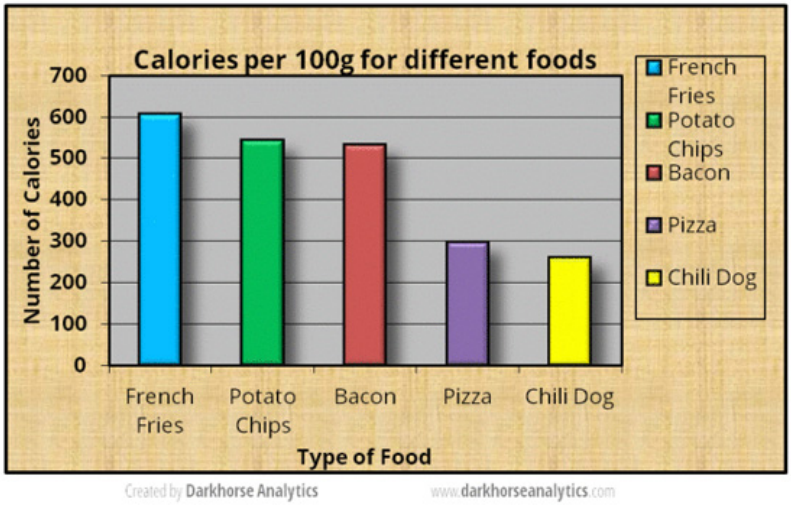
Гарна візуалізація насамперед зрозуміла, а не як на прикладі вище. Серед типових помилок:

1. Відірваність даних від шкали, коли графік створюють «на око», тим паче, коли маніпулюють значеннями;

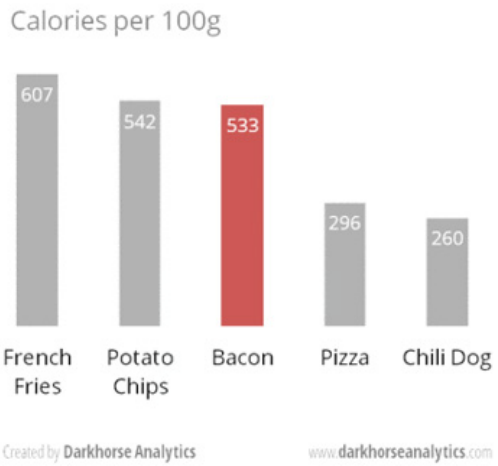
2. Маніпуляції з початком і кінцем шкали, наприклад, на шкалі від 45 до 48 відсотків маленькі зміни видаватимуться дуже великими, хоч і не є такими. Шкала має охоплювати весь діапазон значень, а у стовпчикових графіках — брати початок від нуля;

3. Перевантаженість візуалізації кольорами, зайвими елементами й «прикрашанням», відсутність «білого простору»:

Before



After



data-ink ratio

ТАБЛИЧНІ ДАНІ ДЛЯ ВІЗУАЛІЗАЦІЇ

Це дані, з якими зручно проводити аналіз. Вони мають відповідати таким вимогам:

- Один рядок = 1 спостереження. Це можуть бути дані однієї особи, однієї державної закупівлі, однієї області чи країни.
- Один стовпчик = 1 змінна. Серед помилок часто трапляються об’єднання змінних в одному стовпці (назва та форма юридичної особи) або використання значень категорійної змінної як стовпця (наприклад, вікових груп, років чи форм юридичної особи).
- Не використовуємо об’єднаних комірок.
- Не використовуємо перенос рядка в текстовій комірці.
- Звертаємо увагу та правильно обираємо формат комірок (текст, число, дата).
- Не поєднуємо кілька одиниць даних в одній комірці.
- Не використовуємо латинські літери замість кириличних, і навпаки.
- На початку і в кінці комірки з даними не має бути пробілів. Самі пробіли не можуть повторюватись.

- Використовується один тип лапок, бажано французькі («»).
- Один тип апострофа (’).
- Пробіли після розділових знаків, але не перед ними.
- Однаковий запис дати та часу, найкраще — за стандартом ISO8601 (2020-10-01 15:40).
- Використовуйте простий формат для чисел, без зайвих пробілів між розрядами. Десятковий роздільник має бути крапкою (1.999). Формат десяткового роздільника можна змінити в налаштуваннях гугл-таблиць.
- Якщо потрібно зберегти провідні нулі, форматуйте комірки з числами, як текст. Якщо число не є по суті числом, наприклад, код ЄДРПОУ або ідентифікаційний номер, теж зробіть тип змінної текстовим.
- Не треба писати текстові значення чи назви колонок ВЕЛИКИМИ ЛІТЕРАМИ (крім абревіатур).

- Ці правила стосуються таблиць саме для аналізу даних в електронних таблицях або інших інструментах. Таблиці для друку та представлення даних можуть містити й об’єднані комірки, і значення категорійної змінної як стовпчиків.
  Приклад форматування таблиці.

ТИПИ ЗМІННИХ

Тип	Опис	Операції
Категорійні (номінальні)	Позначають унікальну категорію, до якої належить або не належить спостереження. Наприклад, стать, країна, форма юридичної особи.	Дорівнює та не дорівнює
Порядкові	Шкала з впорядкованими категоріями, відстань між якими невідома. Наприклад, самооцінка задоволеності чимось: «повністю задоволений», «частково задоволений», «і так, і ні», «радше незадоволений», «повністю незадоволений»; або рейтинги від 1 до 5 зірок; або дні тижня.	Порівняння: більше або менше
Числові (інтервальні, метричні)	Дані, які можна виразити числом: кількість чогось, сума коштів, температура, ВВП, вік, час...	Арифметичні операції: сума, віднімання, середнє, варіація

ДЕ ШУКАТИ ДАНІ

- data.gov.ua/ — Державний портал відкритих даних, спробуйте сортувати за популярністю наборів даних.

- Регіональні портали відкритих даних.

- Statista або Kaggle — багато даних на різні теми.

- Сайти державних органів, міністерств, місцевого самоврядування — публікують офіційну статистику, часто у недостатньому обсязі й мало.

- Держстат — мікродані (дані для кожного учасника опитування) обстежень Держстату або інші дані Держстату (не у формі відкритих даних, можуть бути в збірниках у розділі «Публікації»).

- Ring — пошук відкритими даними та проєктами Канцелярської сотні. Взагалі зверніть увагу на її дані. Є API й машиночитані дані.

- declarations.com.ua — дані декларацій чиновників та депутатів, їх багато, треба шукати щось конкретне.

- Прозорро — дані публічних закупівель.
- Spending — використання публічних коштів.
- Дані міжнародних організацій — OECD (зокрема, опитування PISA), ООН, Світового Банку, МВФ, Євростату.
- Центр суспільних даних — мають гарний набір даних із причин смертності, електоральні дані, аналізи даних й оброблені дані Судового реєстру.
- УЦОЯО — дані ЗНО (їх можна поєднати з даними Мінфіну про фінансування шкіл, наприклад).
- Верховна Рада — дані голосувань, хто з ким разом подавав законопроєкти тощо.
- Медреформа — дані про декларації з лікарями.
- datasetsearch.research.google.com — пошук за наборами даних від Гугл.

- Гугл-тренди — популярність різних пошукових запитів.

- World Values Survey, European Social Survey — анонімні мікродані соціологічних моніторингових досліджень.

- Дані компаній: Uber викладає дані швидкості водіїв, Airbnb дає дані пропозиції у деяких містах тощо.

- Sentinel — супутникові знімки.

ФУНКЦІЇ ДЛЯ ЧИСТКИ ДАНИХ

Щоб скористатися функцією, треба поставити курсор у комірку, вгорі в рядку функцій написати «=» та назву функції, у дужках дати потрібні аргументи, як вказано в інструкції, що спливає. До кожної функції є довідка, можна знайти приклади. Виконавши функцію для одного рядка, її можна «протягнути» на інші рядки, за потреби поставивши знак «$» між номером рядка та літерою стовпчика, можемо зафіксувати значення.

Функція	Дія
TRIM	Видаляє з тексту всі пробіли, за винятком поодиноких пробілів між словами
LEFT / RIGHT	Повертає перший символ або кілька перших символів текстового рядка на основі заданого числа символів. Right повертає останні символи
CONCAT	Об’єднує текст із кількох діапазонів і (або) рядків. Замість функції можна використовувати символ амперсанду “&”, у такому випадку вираз матиме такий вигляд: =(“Слава”&” ”&”Україні”&”!”) і видасть у результаті фразу: “Слава Україні!”
PROPER	Перша буква в рядку тексту і всі перші літери, наступні за знаками, відмінними від букв, перетворюються на прописні (великі літери). Всі інші букви в тексті перетворюються на малі (нижній регістр)

Функція	Дія
MID	Повертає задану кількість символів із рядка тексту, починаючи зі вказаної позиції, на основі заданої кількості символів
Ctrl + H	Вбудована можливість знайти та замінити. Можна використовувати регулярні вирази
Текст по стовпцях (Data ➜ Split text to columns)	Вбудована функція розділення тексту по стовпцях, наприклад, за пробілами
VLOOKUP	Шукає значення в першому стовпці таблиці та повертає значення з того ж рядка з іншого стовпця таблиці. З її використанням можна з’єднати дві таблиці

ОСНОВНІ ОПЕРАЦІЇ З ДАНИМИ

ФІЛЬТРУВАННЯ

Це відбір потрібних даних за значеннями певної змінної. Наприклад, вибрати з таблиці лише рядки, де змінна дорівнює якомусь значенню, або більше чи менше певного числа.

Щоб відфільтрувати, шукайте на панелі інструментів позначку фільтра. Клікнувши на неї, з’являється меню фільтра у кожному стовпці. Натисніть на цю позначку ще раз, аби вимкнути фільтри. Якщо меню не з’являються або з’являються некоректно, виділіть усю таблицю та знову натисніть на фільтр. Якщо ви не знаходите таку позначку на панелі інструментів, оберіть «Фільтр» у меню «Дані» (Data ➜ Create a filter).

СОРТУВАННЯ

Операція переміщення рядків за спаданням або зростанням певної змінної. Можна сортувати за абеткою, за числом, за датою.
Щоб відсортувати дані, виділіть усі колонки з даними та знайдіть у меню «Дані» пункт «Сортувати діапазон» (Data ➜ Sort range). Вкажіть, чи є у ваших даних заголовок, за якими колонками сортувати, та виберіть порядок сортування — за спаданням чи за зростанням змінної.

АГРЕГАЦІЯ ДАНИХ

Часто нам потрібні не всі дані, а середні значення чи суми за категорійними змінними. Для цього можемо скористатися такими функціями для числових даних:

Функція	Дія
MIN	Знайти мінімальне значення діапазону
MAX	Знайти максимальне значення діапазону
AVERAGE	Знаходить середнє значення діапазону. Це сума всіх значень, поділена на їхню кількість
SUM	Підраховує суму в діапазоні
MEDIAN	Підраховує медіану діапазону. Медіана — це центральне значення впорядкованого ряду чисел, наприклад, якщо в діапазоні є числа 1, 2, 2, 2, 4, 5, 7 — медіана буде «2», це четверте значення з семи чисел, відсортованих від більшого до меншого

Функція	Дія
MODE	Повертає значення, яке найчастіше трапляється в розподілі. У ряді чисел із попереднього прикладу мода дорівнює двом
FREQUENCY	Частота знаходження певних значень у діапазоні
COUNTA	Підрахунок кількості значень
COUNTIF	Підрахунок кількості значень, які відповідають умові
UNIQUE	Пошук унікальних значень із діапазону

Потужний інструмент для агрегації даних — зведені таблиці (pivot table). За допомогою останніх можна порахувати статистики за категорійними змінними або подивитися на зв'язок категорійних змінних. Щоб зробити зведену таблицю, знайдіть у меню «Дані» відповідний пункт (Data ➜ Pivot table). У листі зі зведеною таблицею на бічному меню можна вибрати змінну в рядках — це та, за якою здійснюється групування даних, і вибрати значення (values), які давати за вибраними категоріями. Більше про зведені таблиці тут.

ВІД ДАНИХ ДО ГРАФІКА — ВІЗУАЛЬНЕ КОДУВАННЯ

Візуальне кодування — це поєднання змінних із каналами візуального представлення та мітками. Останніми ми називаємо прості форми, які можуть позначати елементи даних: точки, лінії, площини. Канали — це спосіб, у який ми показуємо позначки: позиція, розмір, форма, орієнтація, насиченість чи відтінок кольору.

Вибір каналу залежить від типу даних (категорійні, порядкові чи метричні). Важливо пам’ятати, що різні канали мають різну силу, тож найважливіші дані треба кодувати найсильнішими каналами, а другорядні дані — слабшими.

Ефективність візуальних каналів

Величинні канали: порядкові(кількісні) дані

Позиція, спільна база

Позиція, різна база

Довжина

Кут/Нахил

Площа

Глибина

Освітленість

Насиченість

Кривизна

Об"єм

Ефективність: Найбільш — Найменш

Канали **ідентичності** : категорійні дані

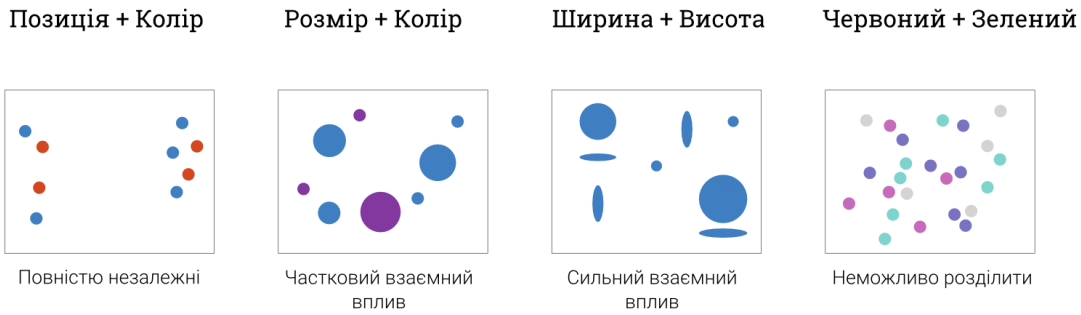
Позиція в просторі

Відтінок кольору

Рух

Форма

Крім того, вплив каналів може змішуватись:



У цьому прикладі розмір впливає на відчуття кольору (2-ге зображення), Горизонтальний розмір + вертикальний розмір змішуються, адже ми сприймаємо площу (3). На останньому зображенні спроба закодувати одну змінну зеленим кольором, а іншу — червоним, призводить до того, що ми просто сприймаємо різні кольори, а не відсотки червоного й зеленого.

Більше про візуальне кодування та особливості сприйняття візуалізацій:
lections/01_03_vis_language,
lections/02_03_vis_language

ДЖЕРЕЛА

Каїро, А. (2017). *Функціональне мистецтво: вступ до інфографіки та візуалізації.* Видавництво Українського Католицького Університету.

Бондаренко, А., & Романенко, Н. (2020, Лютий). Програма курсу «Візуалізація даних» магістерської програми з медіакомунікацій УКУ.

Ґрей, Д., Бонегру, Л., & Чемберс, Л. (Eds.). (2012). Журналістика даних: Посібник (С. Лук'янчук, Trans.). texty.org.ua за підтримки фонду «Відродження».

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association.

From Data to Viz. (2018).

Munzner, T. (2017, June 23). Data Visualization Pitfalls to Avoid.

Wilke, C. O. (2019). Fundamentals of Data Visualization (1st ed.). O'Reilly Media, Inc.

РЕСУРСИ ДЛЯ ОНЛАЙН-ГРАФІКІВ

Tableau
RawGraphs
Flourish
Datawrapper

ДЛЯ МАП

QGis
Keppler.gl
OpenStreetMap
Overpass

ДЛЯ ПІДБОРУ КОЛЬОРІВ

ColorBrewer
Chroma.js

КУРСИ УКРАЇНСЬКОЮ

Візуалізація даних
Аналіз даних

Матеріал створено в межах проєкту «Використання даних у боротьбі проти маніпуляцій та дезінформації», який здійснює Асоціація з міжнародних питань (АМО) у співпраці з Texty.org.ua та «Інтерньюз-Україна», за фінансової підтримки Програми трансформаційної співпраці Міністерства закордонних справ Чехії.