Візуалізація даних графіками

План

1. Критерії вибору графіку
2. Рейтинг візуальних каналів
3. Знаходження міток

Типи графіків. Як вибрати вірний графік для різних задач

Графіки дозволяють ефективно показувати різноманітні зв’язки, відношення між різними атрибутами (змінними) у наших даних. Вони надають характерну візуальну форму для кожного типу зв’язку. Корисно розуміти, які типи графіків можуть бути застосовані для різних типів зв’язків.

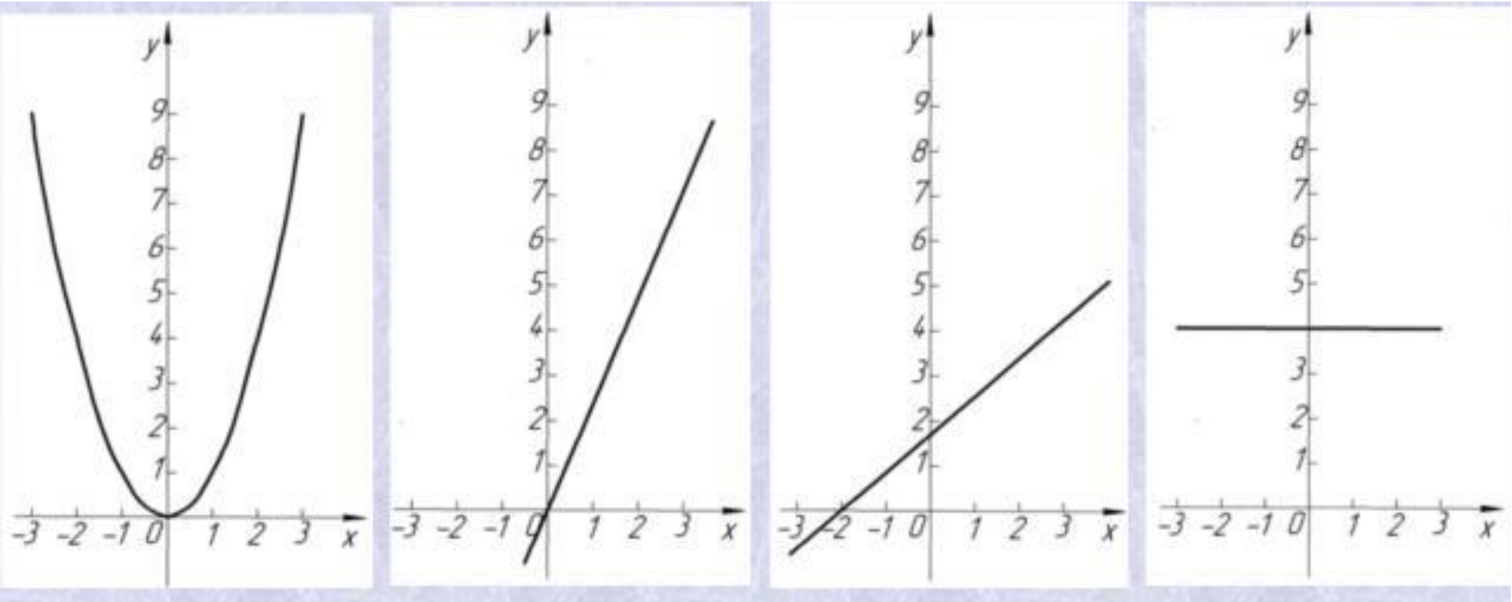
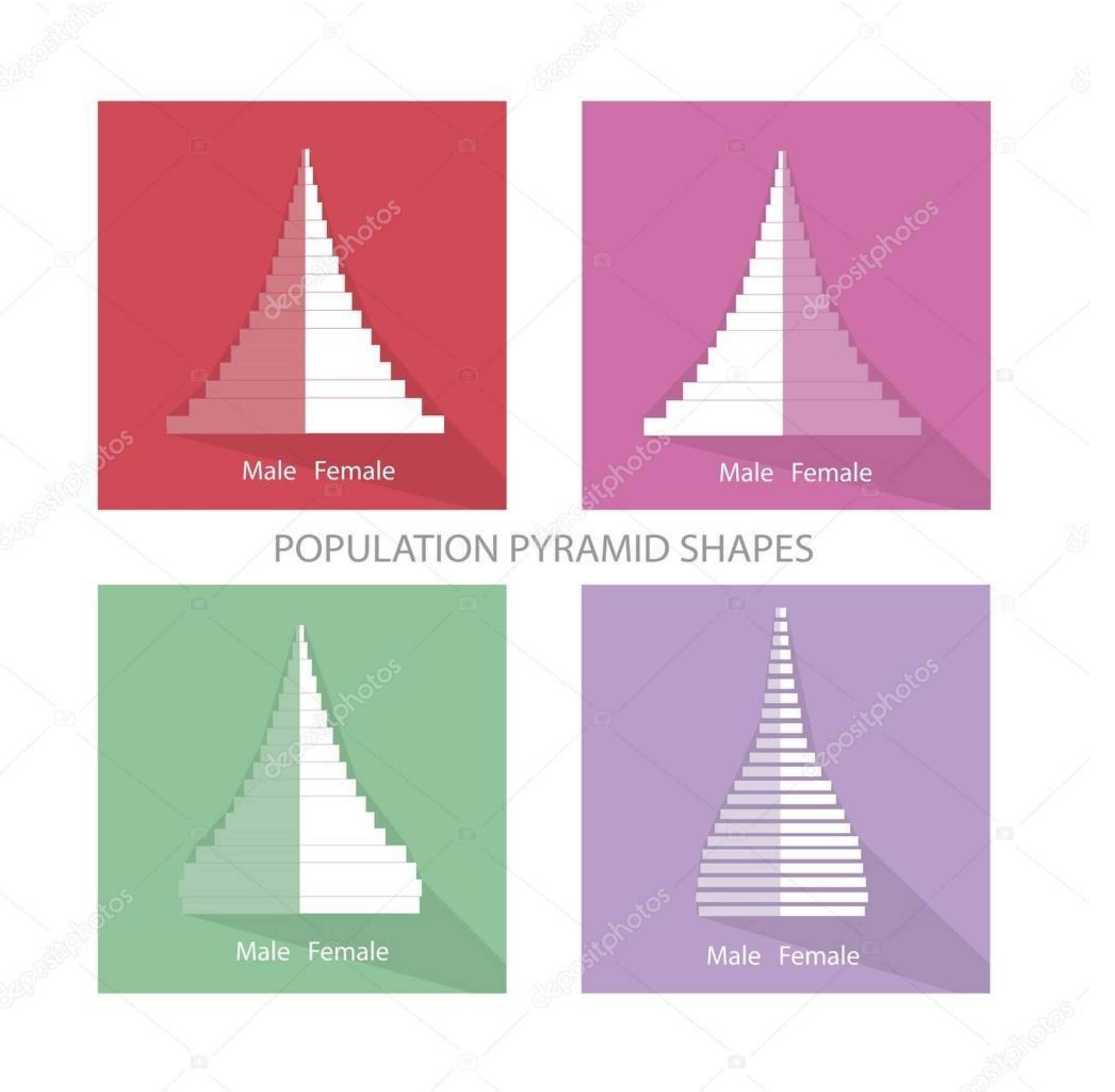


Рисунок 1 Рисунок 2 Рисунок 3 Рисунок 4

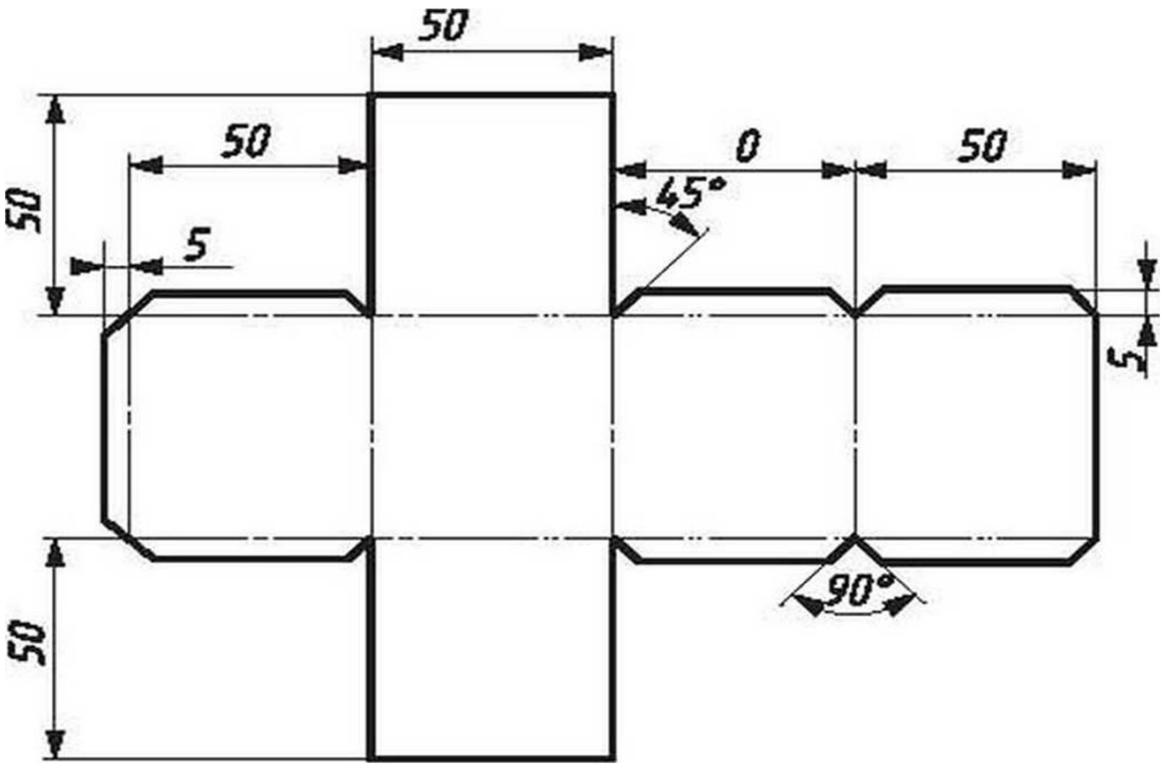
Є декілька таких типів:

- Еволюція в часі
- Ранжування
- Співвідношення частки і цілого
- Відхилення
- Розподіл
- Кореляція
- Географічні дані
- Номінальне порівняння



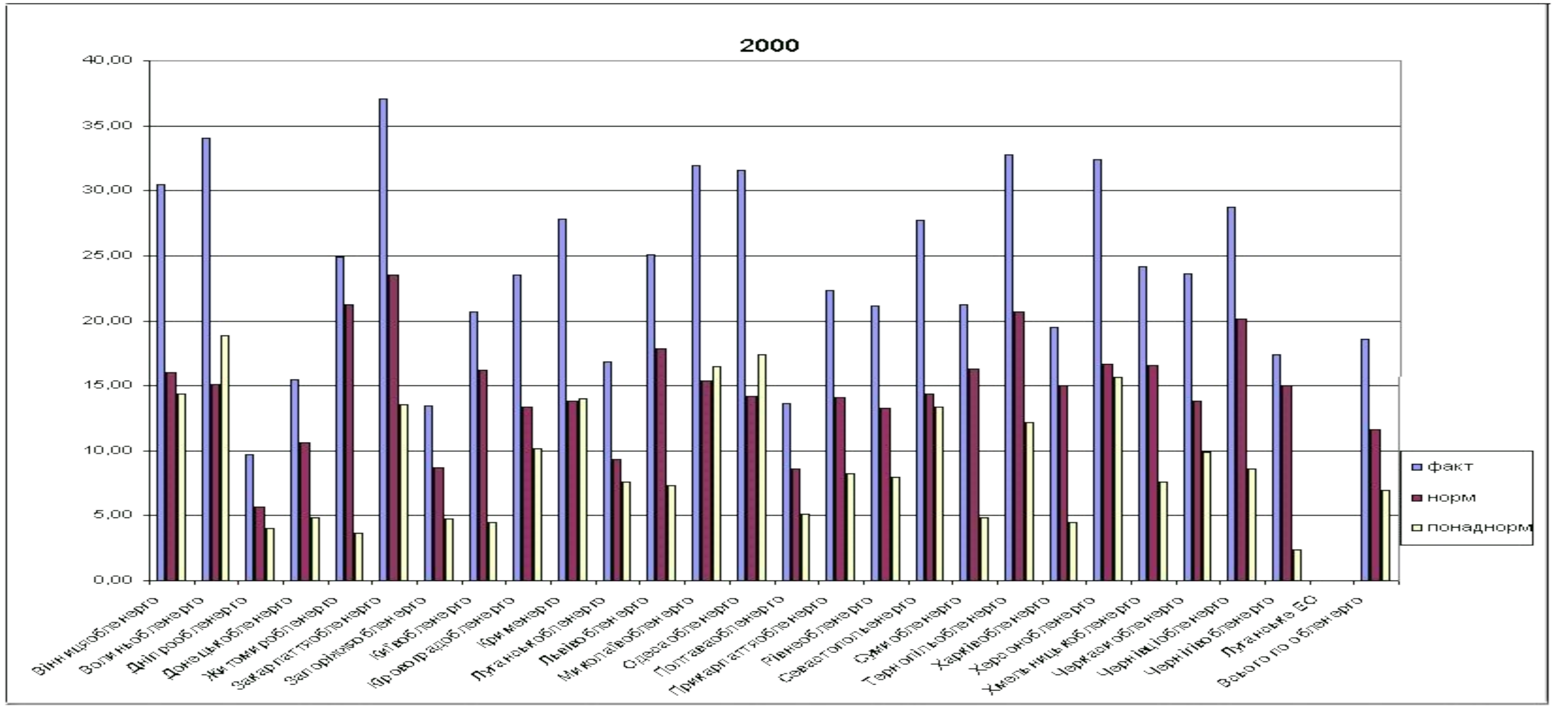
Згадаємо, які візуальні мітки використовуються для кодування даних на графіках:

- точки; лінії; горизонтальні та вертикальні стовпці;
- горизонтальні та вертикальні бокси.

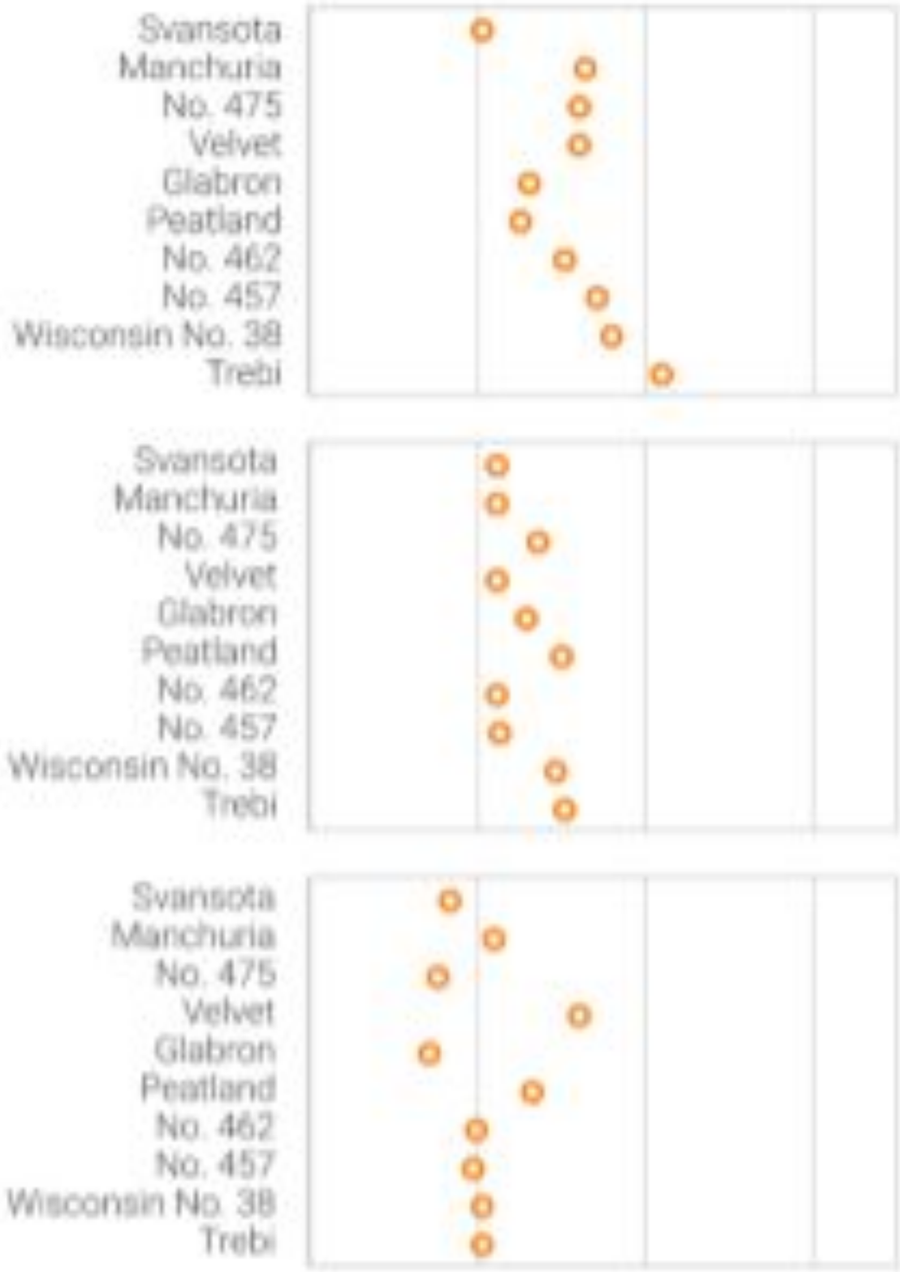


Для визначення, який саме тип нам потрібно показати (тобто який графік вибрати), потрібно пошукати в описі задачі задані ключові слова, за якими можна визначити тип зв’язку:

1. Номінальне порівняння - серія невпорядкованих дискретних кількісних значень - найпростіший тип зв’язку.



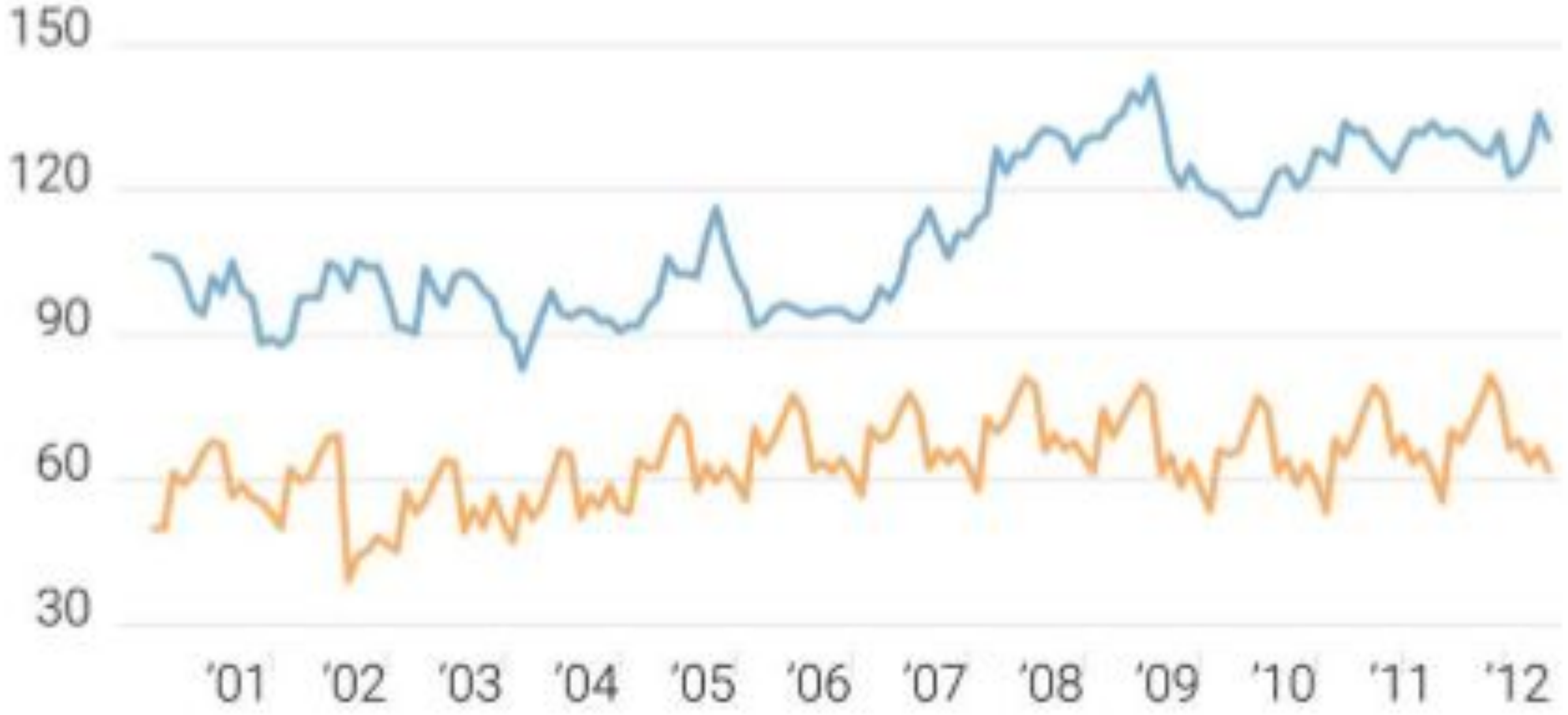
Ми просто маємо показати серію дискретних кількісних значень - кожна з яких відноситься до своєї категорії, щоб порівняти їх відносний розмір. Наші змінні - категорійна і кількісна, кодуємо їх як позицію. Хорошими варіантами мають бути стовпчикові графіки (вертикальні або горизонтальні) або точкові графіки. Особливість - точкові графіки можуть починатися не від нульового значення, а стовпчикові - лище з нуля


Svansota
Manchuria
No. 475
Velvet
Glabron
Peatland
No. 462
No. 457
Wisconsin No. 38
Trebi
Svansota
Manchuria
No. 475
Velvet
Glabron
Peatland
No. 462
No. 457
Wisconsin No. 38
Trebi
Svansota
Manchuria
No. 475
Velvet
Glabron
Peatland
No. 462
No. 457
Wisconsin No. 38
Trebi

2. Еволюція в часі

Ключові слова: тренд, зміна, зростання (падіння), збільшення (зменшення), підвищення (пониження), коливання (флуктуація).

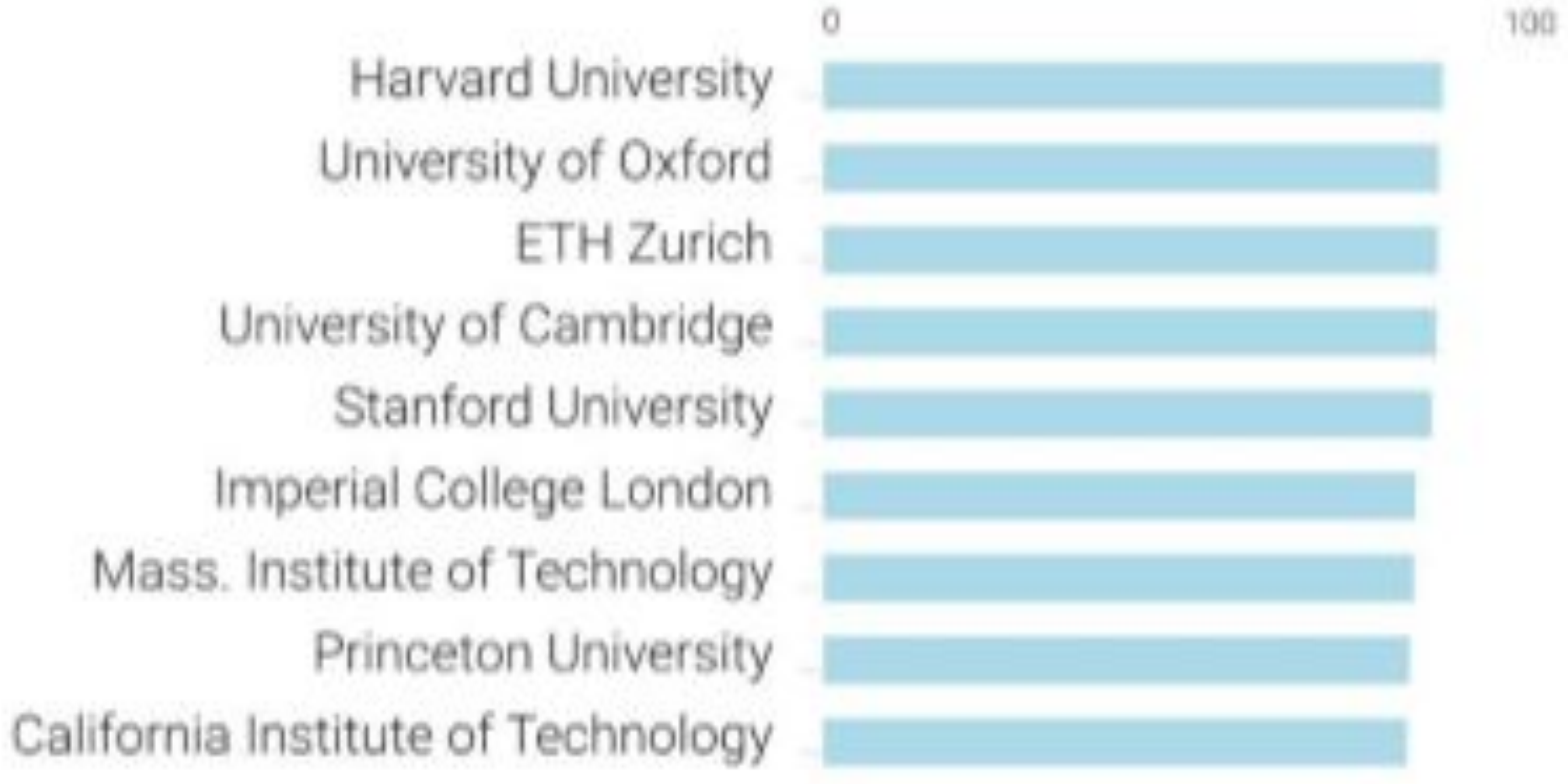
Змінні – впорядкована (час) та кількісна. Кодуємо позицією. Лінійний графік - перший вибір. Також, вертикальні стовпчики - не горизонтальні, в яких час йде по вертикалі. Чому? Тому що показуємо значення продовж (довжина, не висота!) якогось часу - сильна культурна традиція зліва направо. Такі графіки не показуємо по вертикалі. Точковий графік погано піходить, тому що точки гірше показують зв’язок між сусідніми часовими інтервалами.


150
120
90
60
30
'01
'02
'03
'04
'05
'06
'07
'08
'09
'10
'11
'12

3. Ранжування

Ключові слова: більше (менше) ніж, дорівнює.

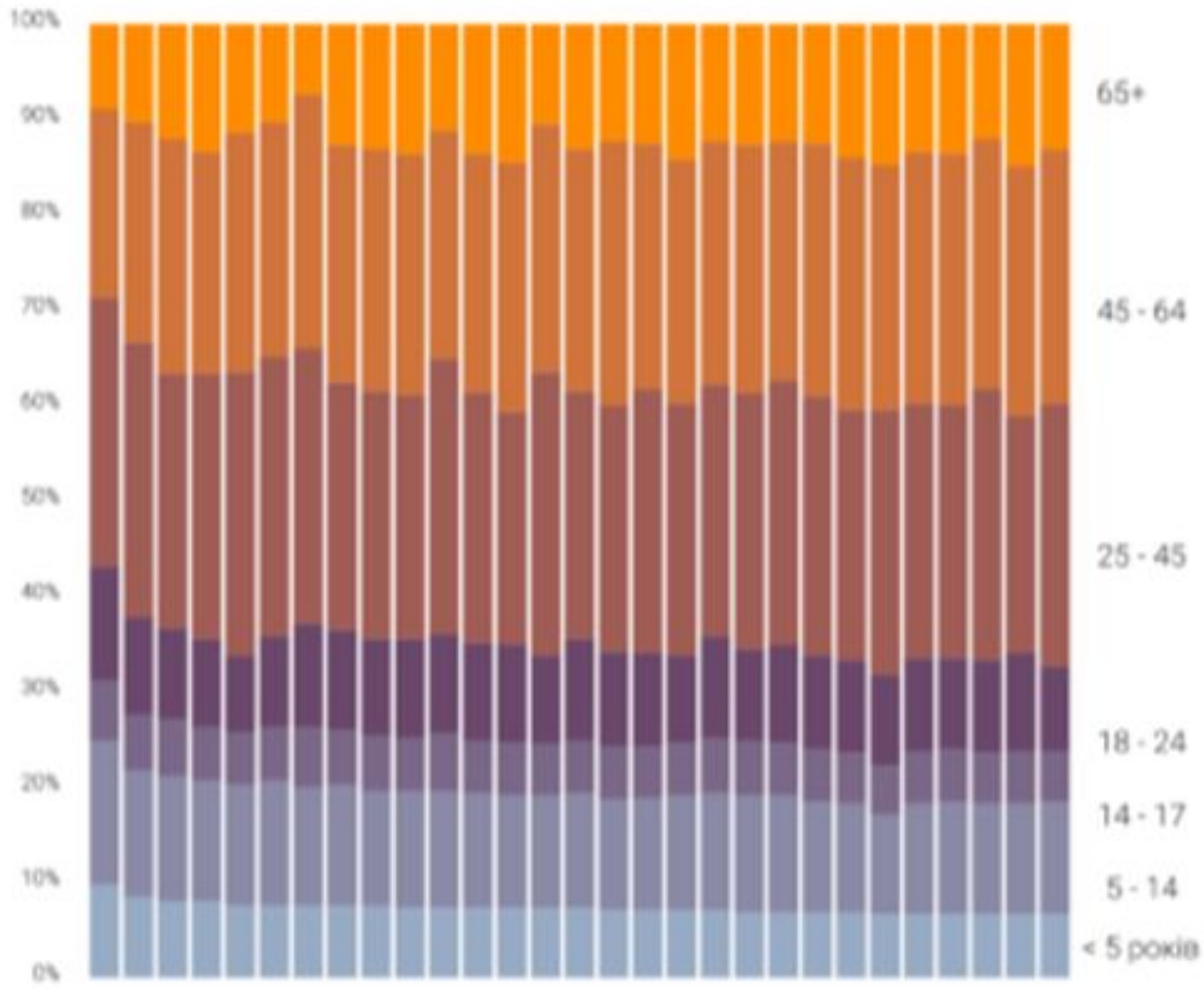
Те ж саме, що номінальне порівняння, однак обов’язково використовуємо сортування! У порядку зменщення або навпаки - в залежності, що саме хочимо показати.


0
100
Harvard University
University of Oxford
ETH Zurich
University of Cambridge
Stanford University
Imperial College London
Mass. Institute of Technology
Princeton University
California Institute of Technology

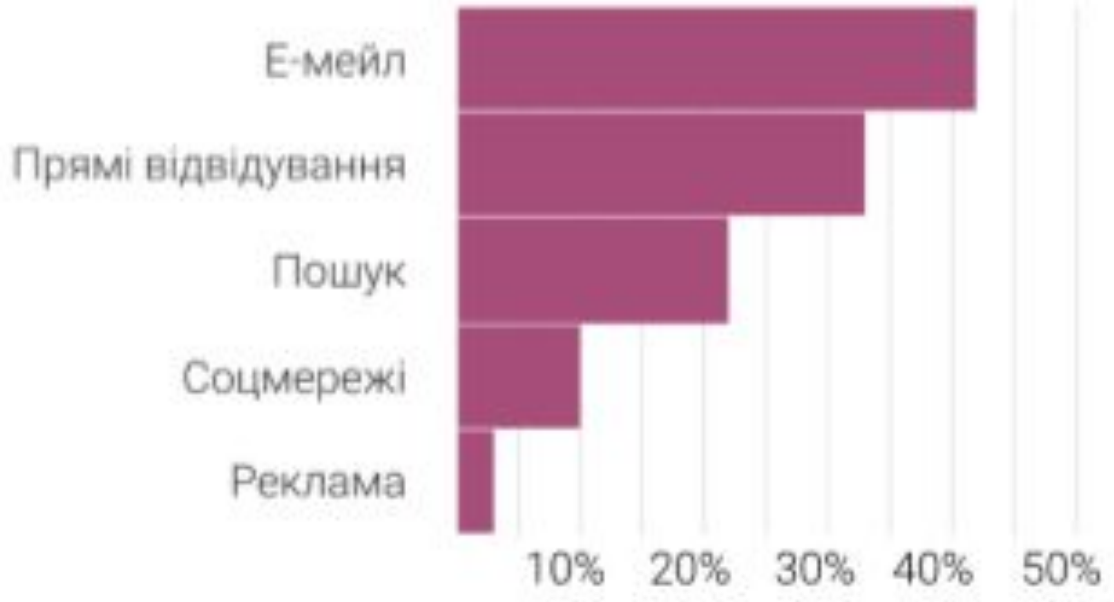
4. Співвідношення частки і цілого

Ключові слова: відношення, відсоток, частка.

Ми натреновані розуміти частку як відсоток. Наші змінні - це категорії (частки цілого) та їх внесок у ціле. Кодування може бути - колір для категорій та довжина для значень часток, виходить складена стовпчикова діаграма. Це лише трохи краще ніж млинці.


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
65+
45 - 64
25 - 45
18 - 24
14 - 17
5 - 14
< 5 років

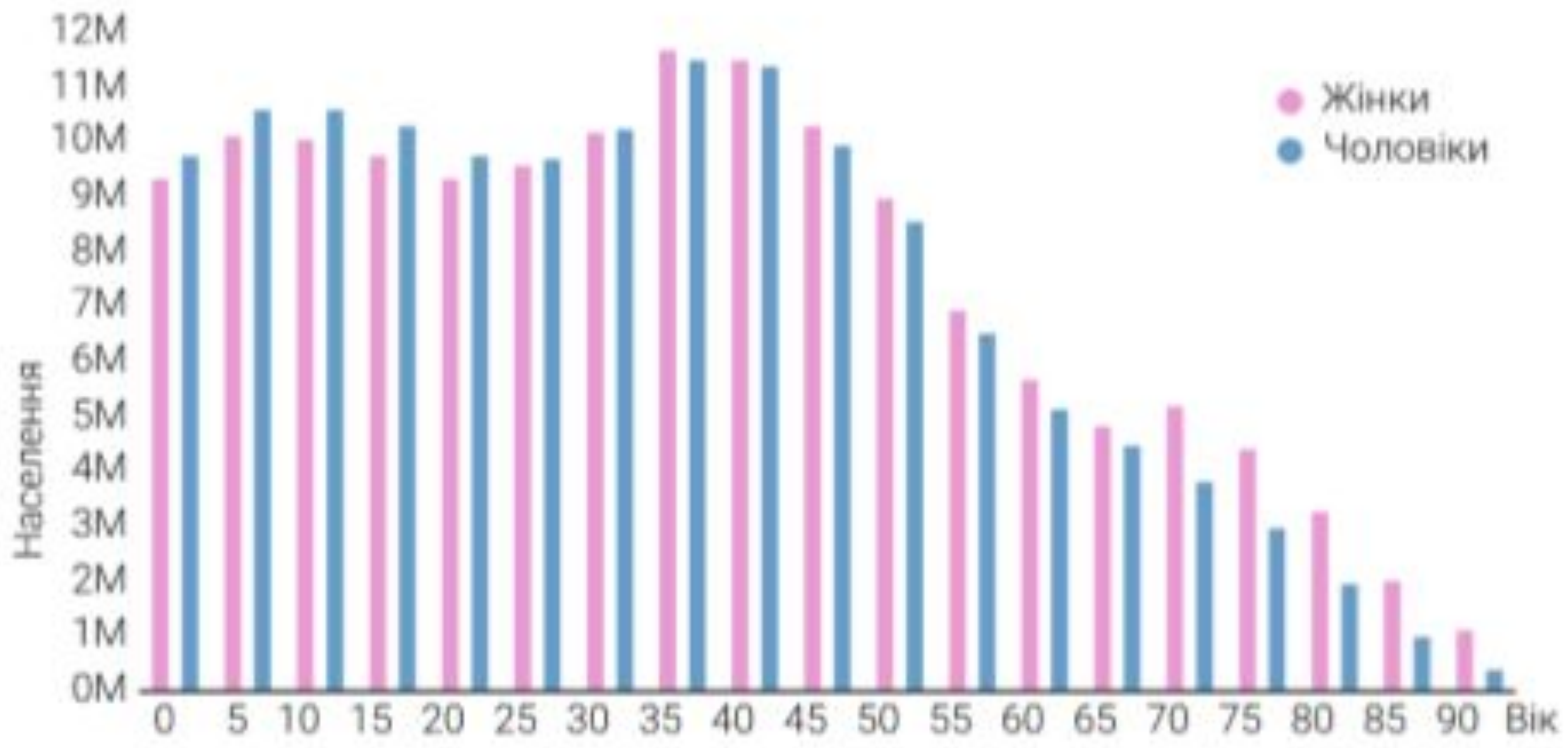
Кращим варіантом буде позиція-позиція чи позиція довжина - коли ми розберем складені фрагменти і відсортуємо їх. Зверніть увагу, що стовпчики поєднані і в заголовку точно вказано що це - частки цілого.



5. Відхилення

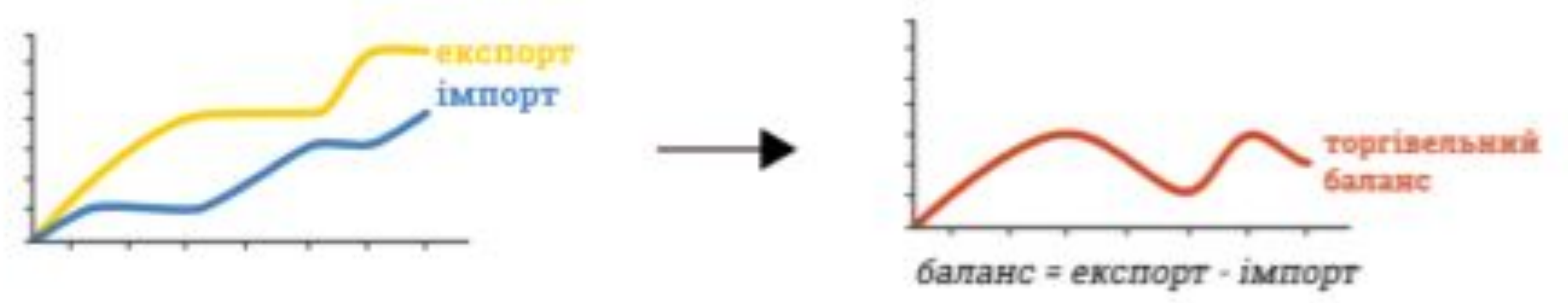
Ключові слова: плюс або мінус, варіація (відхилення), різниця, порівнюючи з.

Знов маємо кількість і категорію в даних, знову використовуємо позицію для кодування. Перший варіант - парні стовпчики, однак якщо цікаво лише різниця, то варто її і показувати, і підкреслити кольором.



12M
11M
10M
9M
8M
7M
6M
5M
4M
3M
2M
1M
0M
Населення
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 Вік
Жінки
Чоловіки

Інший варіант - використати різницю у відсотках, для того щоб об’єктивно оцінити відхидення для різних категорій. Для відхилення у часі використовують лінійний графік з різницею м



Початкові дані

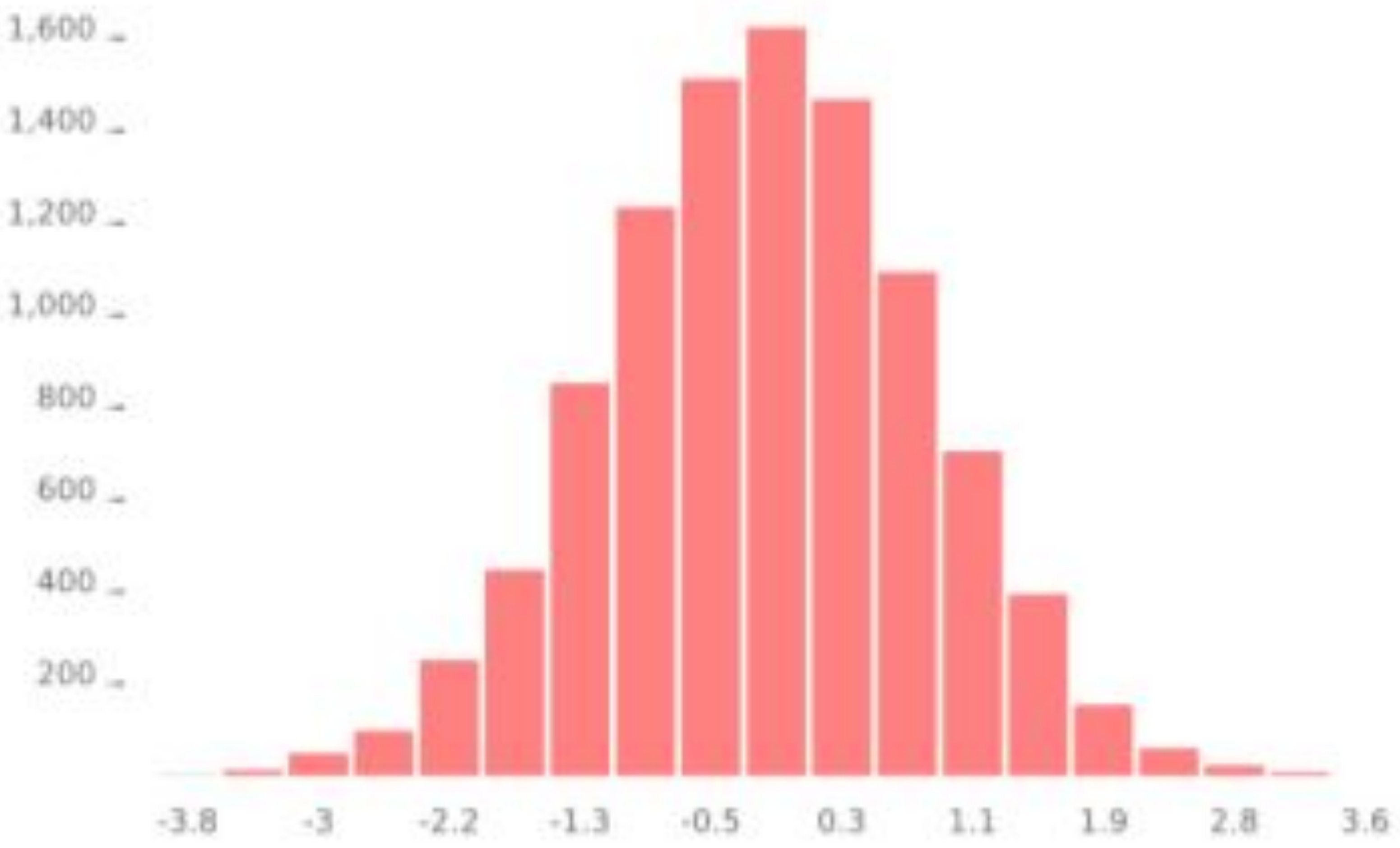
Модифіковані дані

6. Розподіл

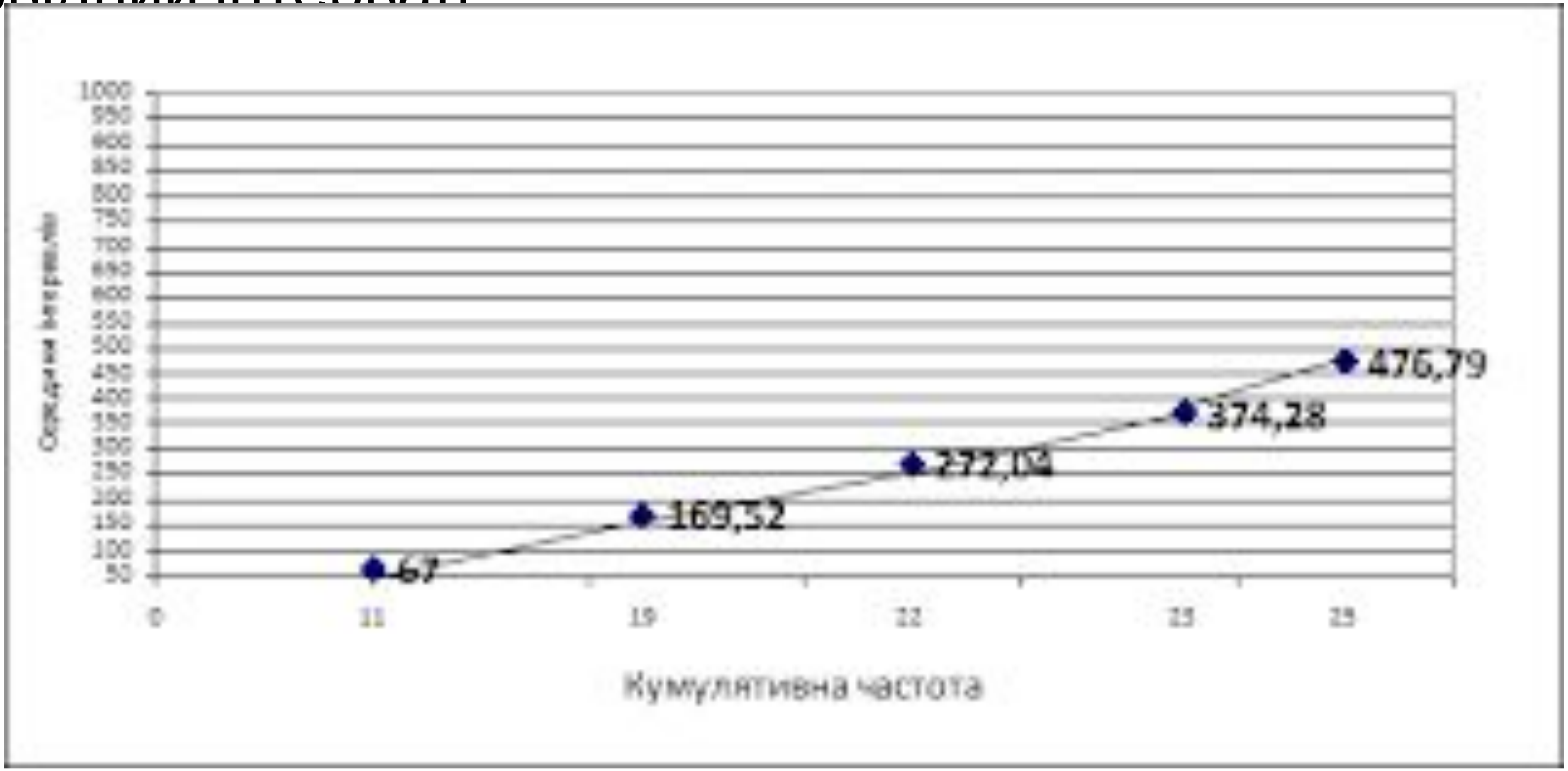
Ключові слова: частота, розподіл, концентрація, нормальний розподіл (крива Гауса, крива Белла).

Графік розподілу показує, наскільки часто значення кількісної змінної зустрічаються вздовж всього діапазону своїх значень, від найменшого до набільшого. Зазвичай, весь цей діапазон розбивається на рівні інтервали (номер такого інтервала - це змінна впорядкованого типу даних), і для кожного інтервалу рахується скільки разів або який відсоток кількісна змінна потрапила в цей інтервал.

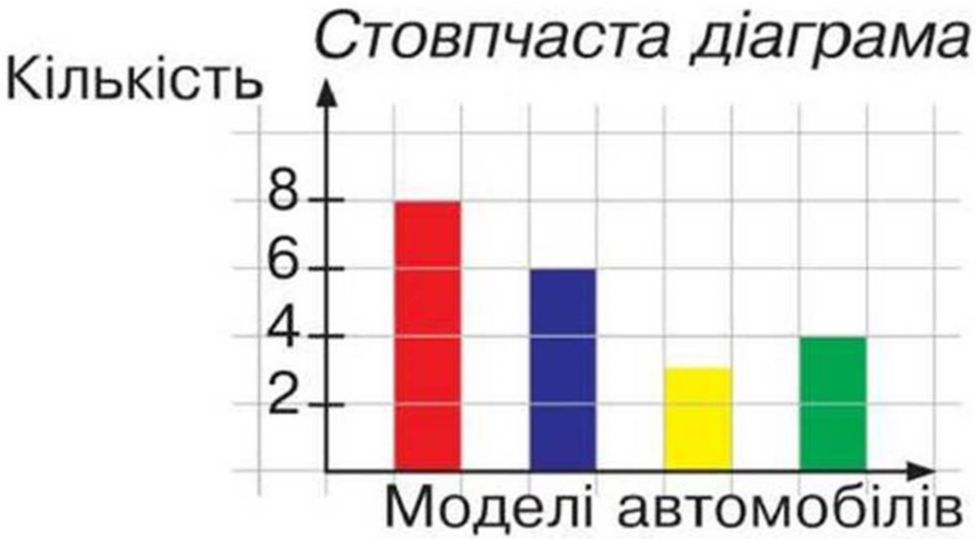
Для такої задачі найчастіше використовується гістограма.

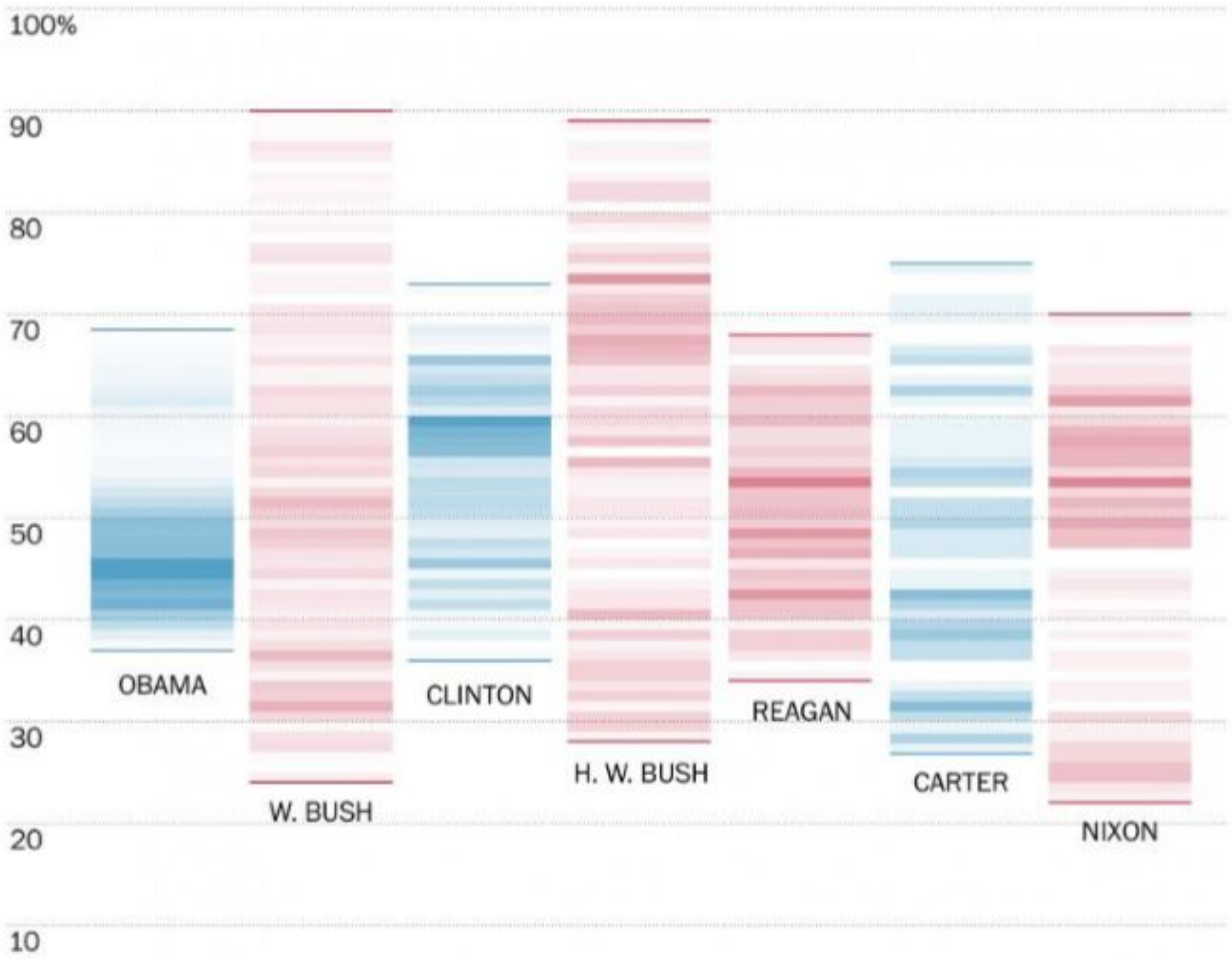

1,600
1,400
1,200
1,000
800
600
400
200
-3.8
-3
-2.2
-1.3
-0.5
0.3
1.1
1.9
2.8
3.6

Кумулятивна частота - теж корисний графік, для того, щоб швидко рахувати внесок декількох діапазонів. Ефективним способом показати всі значення є «точкова гістограма», в якій кожна точка є одним значенням змінної, що потрапила у відповідний інтервал.

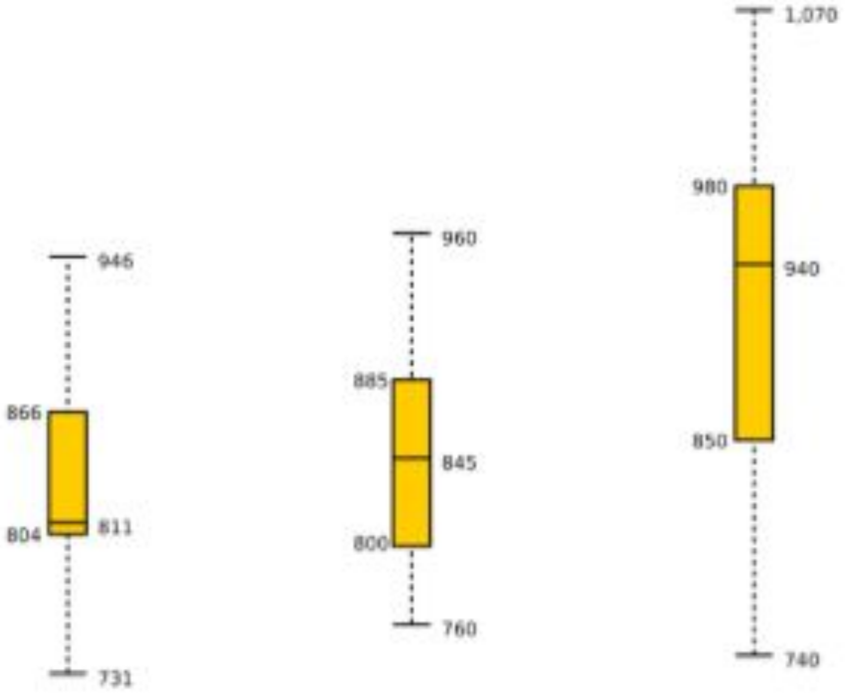


Смужкова діаграма використовує прозорість, щоб показати скупчення в тому чи іншому інтервалі. Цей трюк з прозорістю часто використовують на точкових ‘x y’ - графіках та на картах.



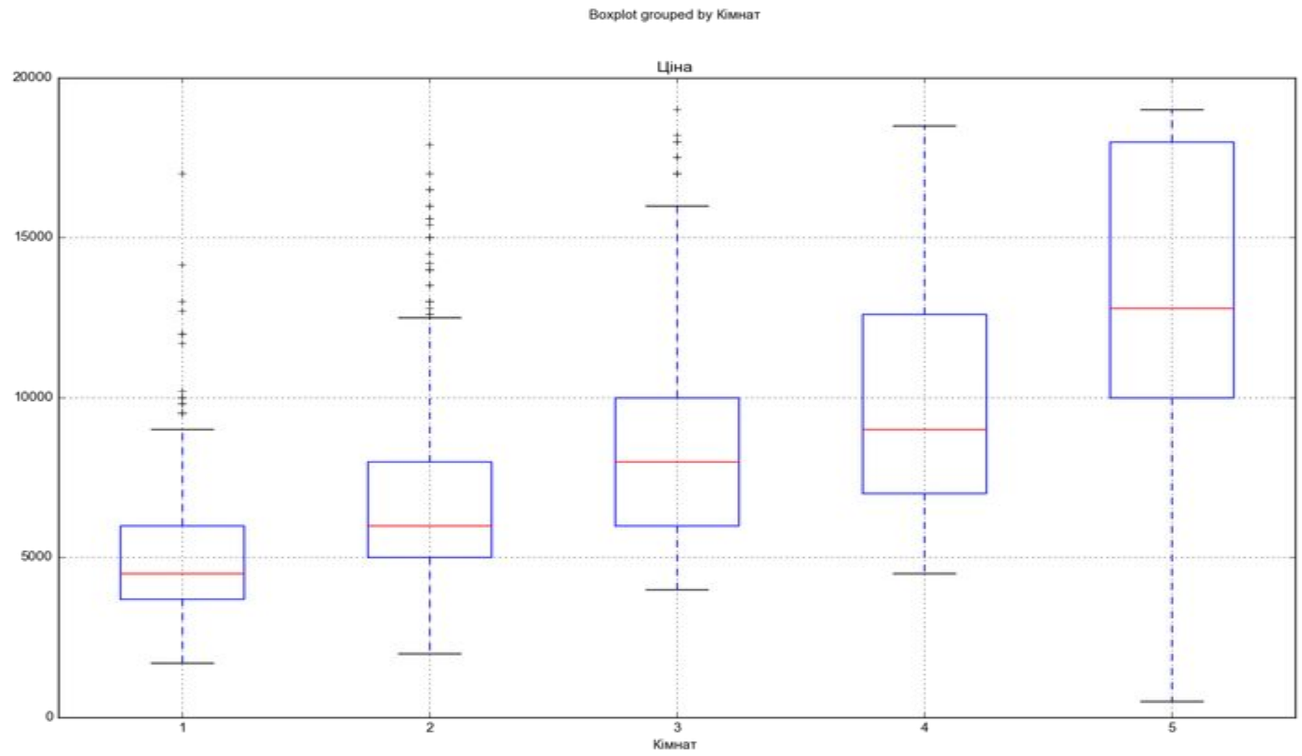

100%
90
80
70
60
50
40
30
20
10
OBAMA
W. BUSH
CLINTON
H. W. BUSH
REAGAN
CARTER
NIXON

Багато розподілів показують за допомогою найбільш популярного способу - лінійного частотного полігона або бокс - графіка



Бокс дозволяє показати одразу розмах значень, найбільше і найменше значення конкретного розподілу, медіану та діапазон, куди потрапляє 50 відсотків усіх значень.

Бокс - графік напевне   один із найбільш ефективних способів показати одразу багато розподілів різних величин на одному листку.



7. Кореляція

Ключові слова: зростає разом з, падає разом з, змінюється разом з, викликане, причина якого, слідує за.

Щоб показати зв’язок між парами кількісних змінних: позиція-позиція = x y - використовуємо точковий графік).

Якщо позиція (для номінальних категорії) + довжина(значення 1) + довжина (значення 2) = парний стовпчиковий графік

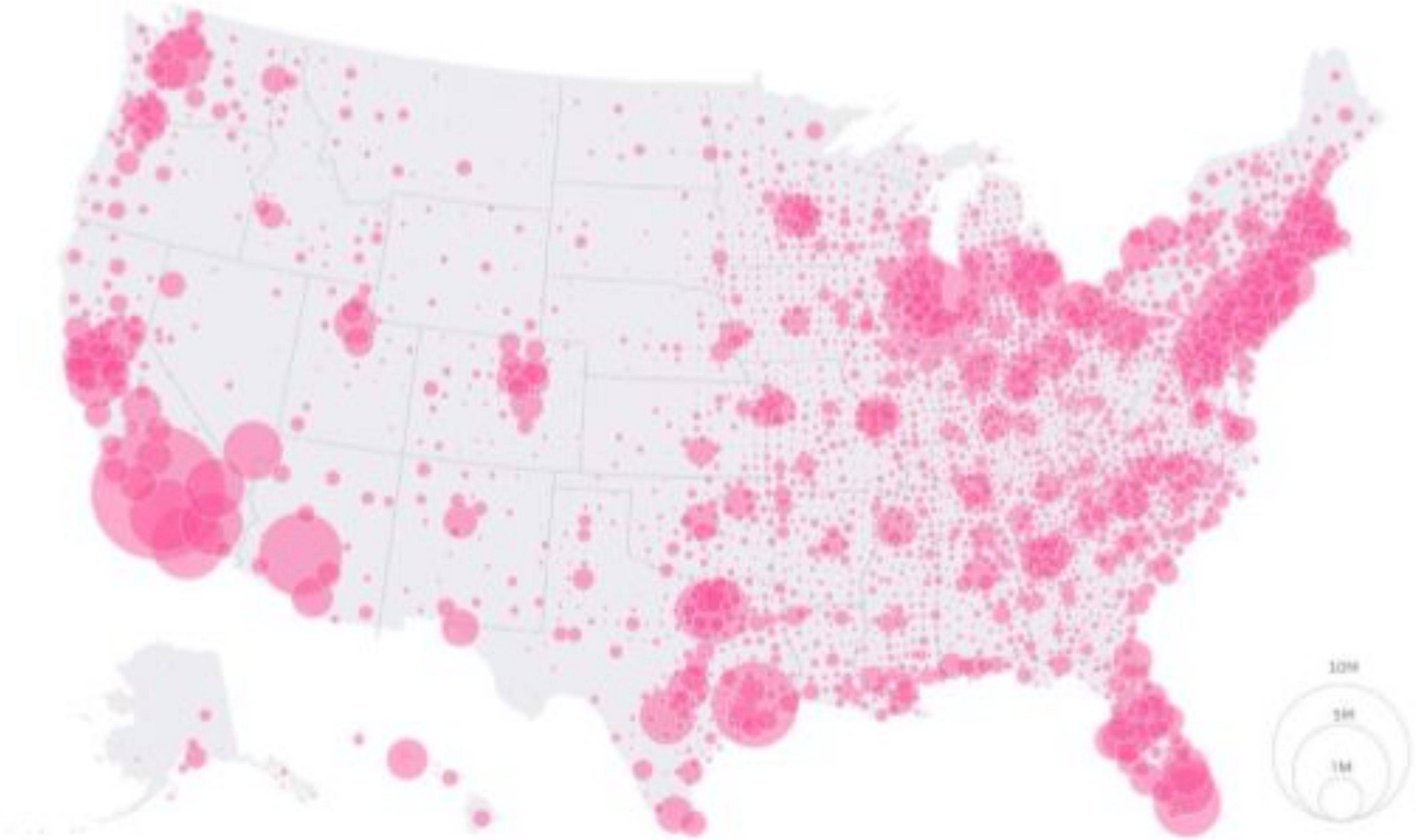
75000
50000
25000
0
Porsche Cayman
0
25,000
50,000
75,000
100,000

8. Географічні дані

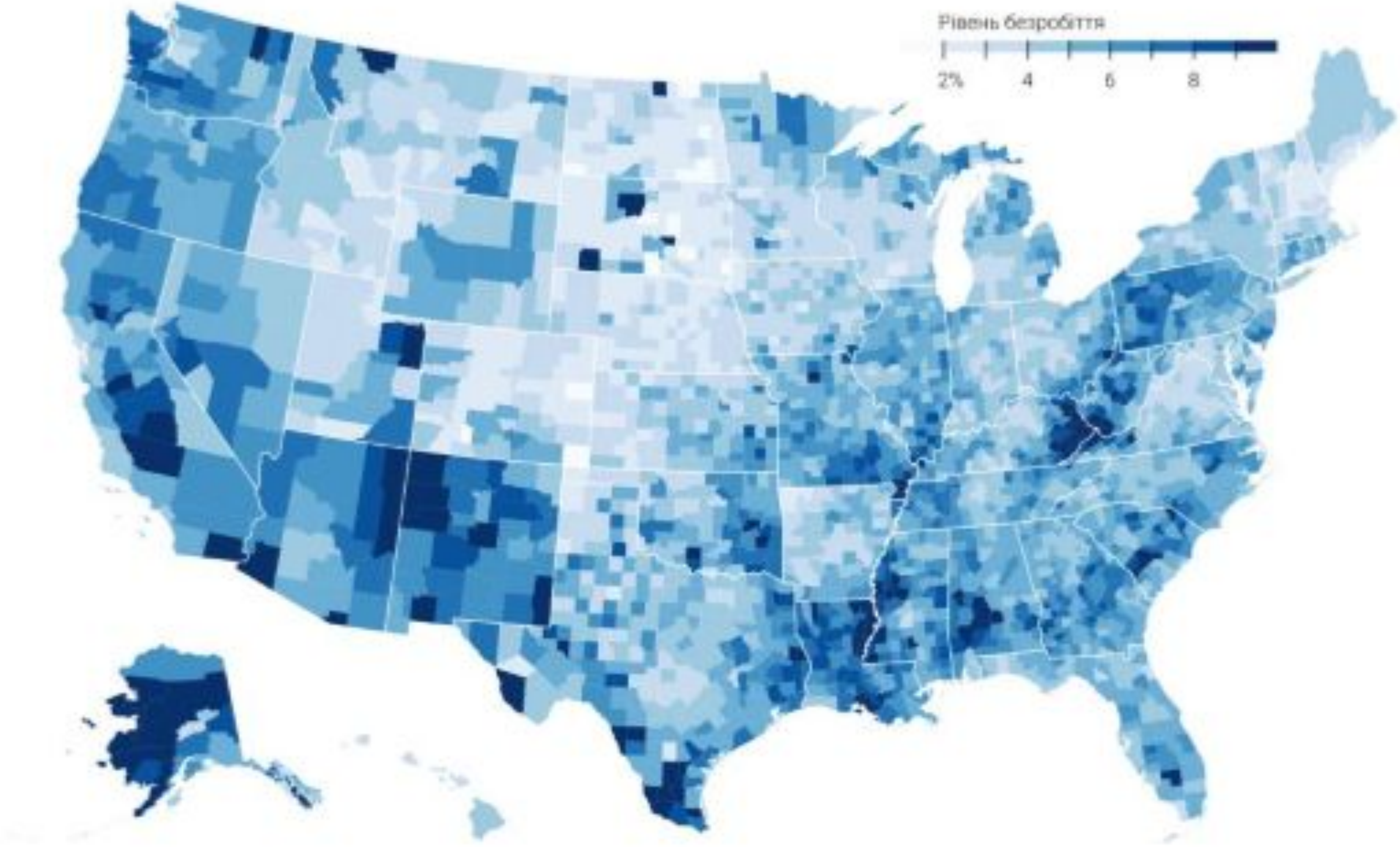
Ключові слова: географія, локація (позиція), де розташоване, регіон, територія, країна, місто, область тощо.

Є проблема - два найкращих способи кодування, які ми використовували весь час, зайняті для карти (x y та довжина із спільною базою). Залишається розмір, інтенсивність кольору та ширина для кількісних даних, хоча вони не такі точні:

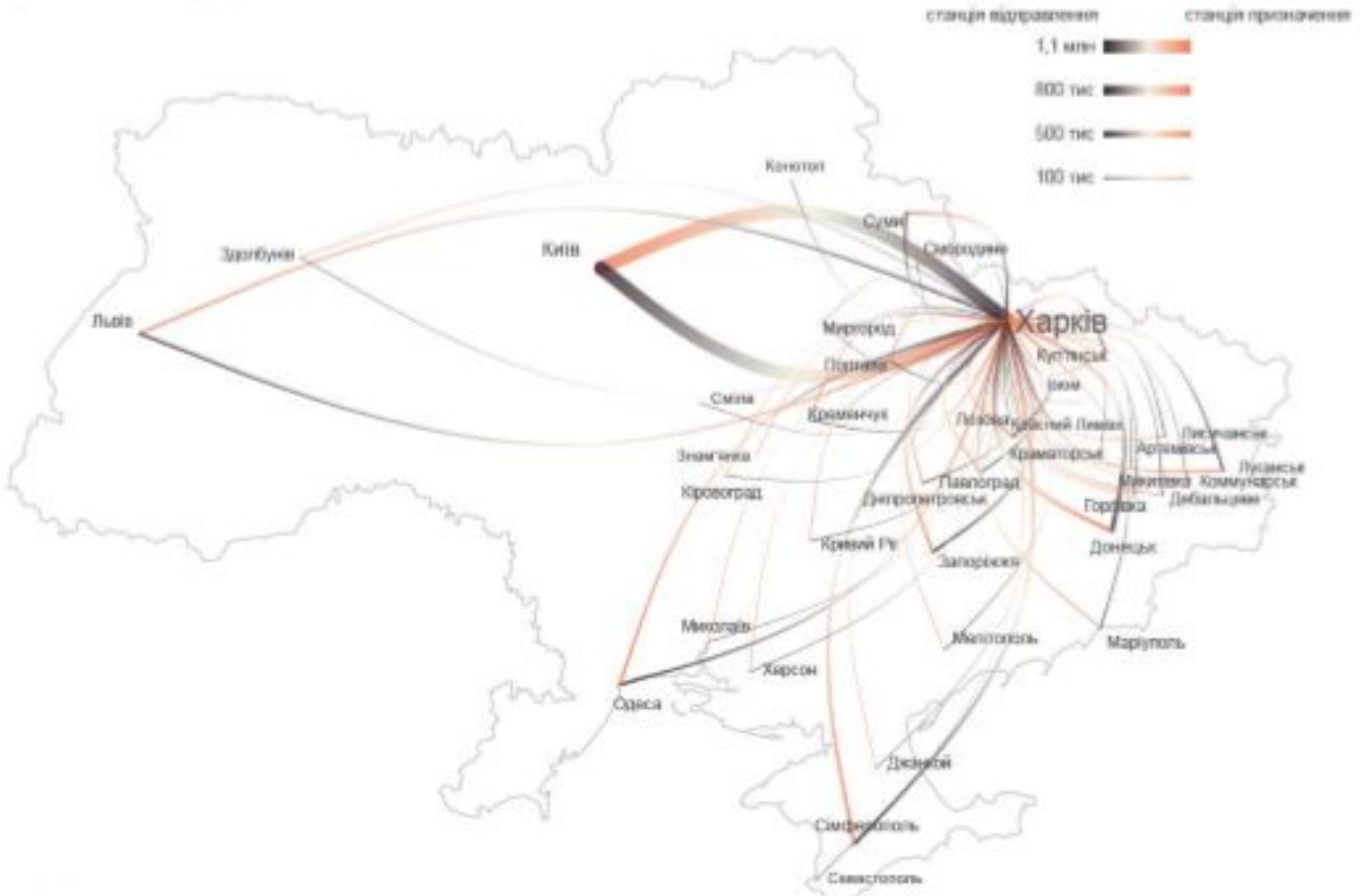
- точки різного розміру;
- точки або інші форми з різною інтенсивністю кольору;


10M
5M
1M

- інтенсивніть кольору для різних гео-регіонів;

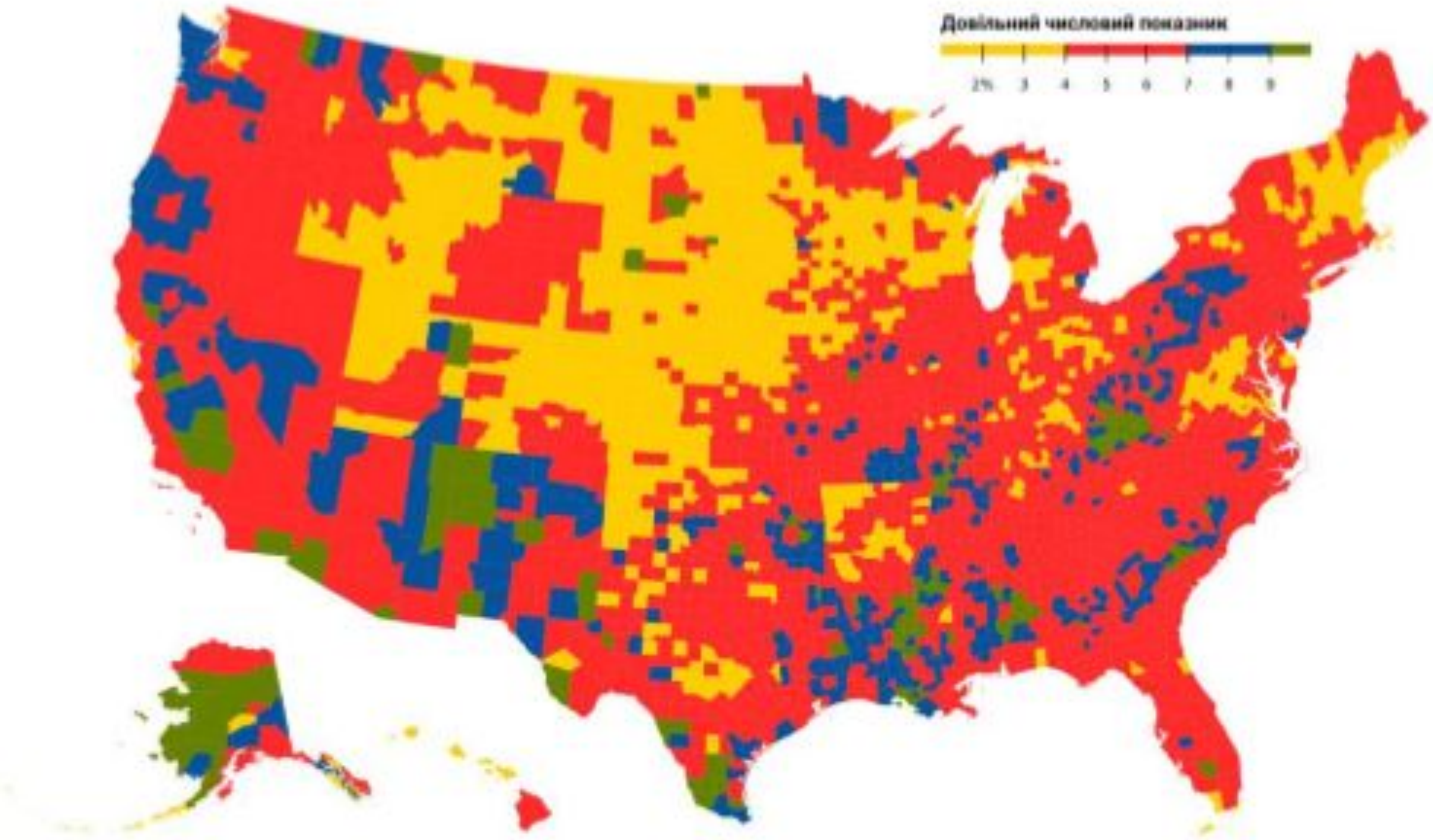


- лінії різної ширини або інтенсивність кольору

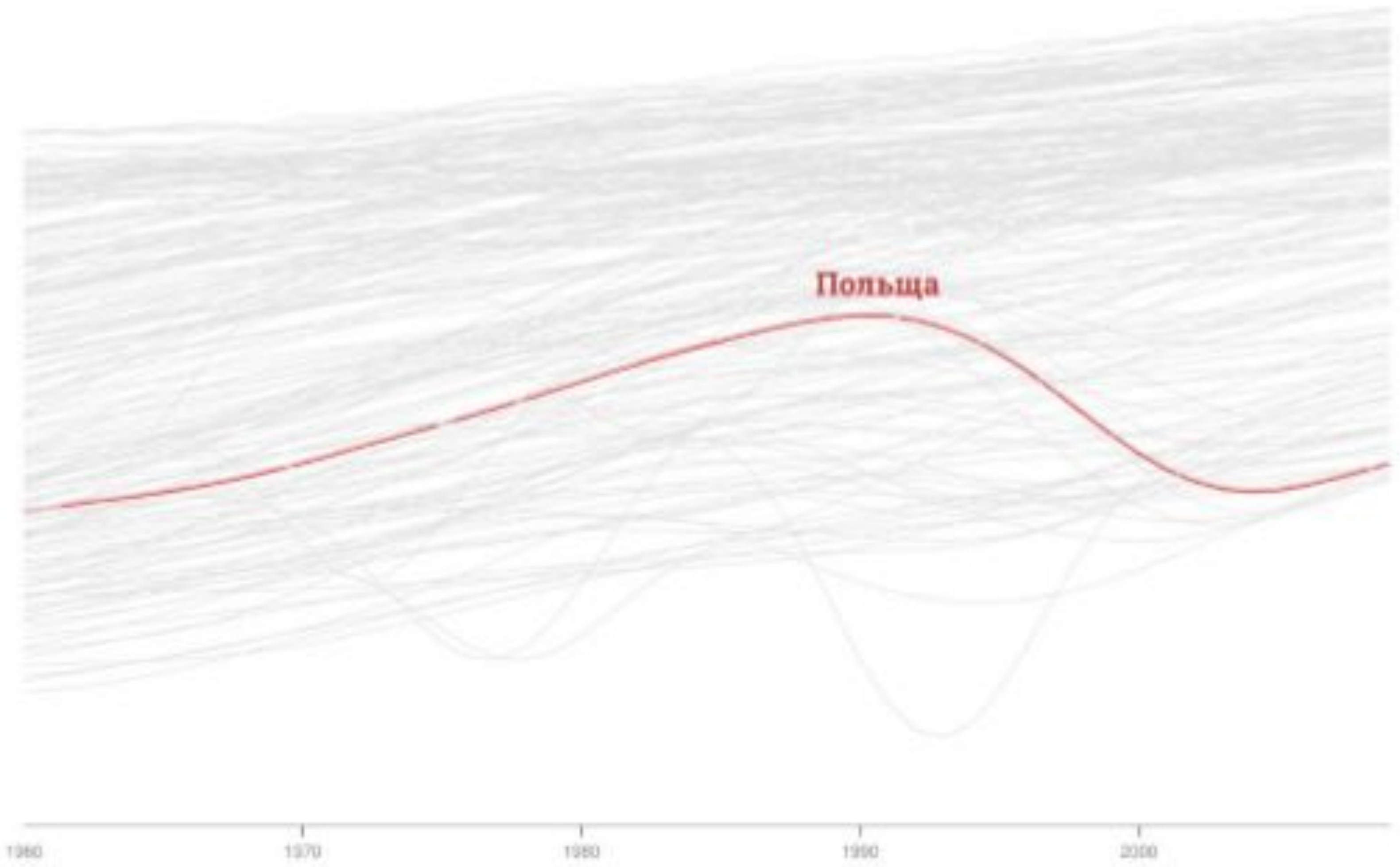


Ефективність візуального кодування

Нагадаю, що раніше ми говорили про мову візуалізації: мітки, візуальні канали, а також, що для різних типів даних потрібно використовувати відповідні візуальні канали. Важливо запам’ятати, що для першого типу даних - категорійних, невпорядкованих потрібно показувати за допомогою такого візуального кодування, що зберігає відмінність та ідентичність - наприклад, різні кольори або різна геометрична форма. Другий тип даних - впорядковані (та кількісні) дані потрібно показувати так, щоб наша візуальна система сприймала порядок - наприклад, якщо у якості каналу задіяний колір, тоді це має бути або перехід від кольору до сірого (десатурація), або однаковий колір з різними інтенсивностями, а не різні кольори.


Довільний числовий показник
2%
3
4
5
6
7
8
9

З іншого боку, різні канали навіть якщо придатні - не однаково ефективні. Тому необхідно використовувати принцип - ''за допомогою найбільш видимого (сильного) каналу порібно кодувати найголовнішу інформацію'' (атрибут в даних). Як визначити найголовнішу інформацію ми говорили на початку курсу. Іншими словами - більш важливі змінні (атрибути) у ваших даних повинні кодуватися більш ефективними, найбільш помітними візуальними каналами. Менш важливі - менш ефективними. Напевне, це найбільш важливий принцип інформаційного дизайну.

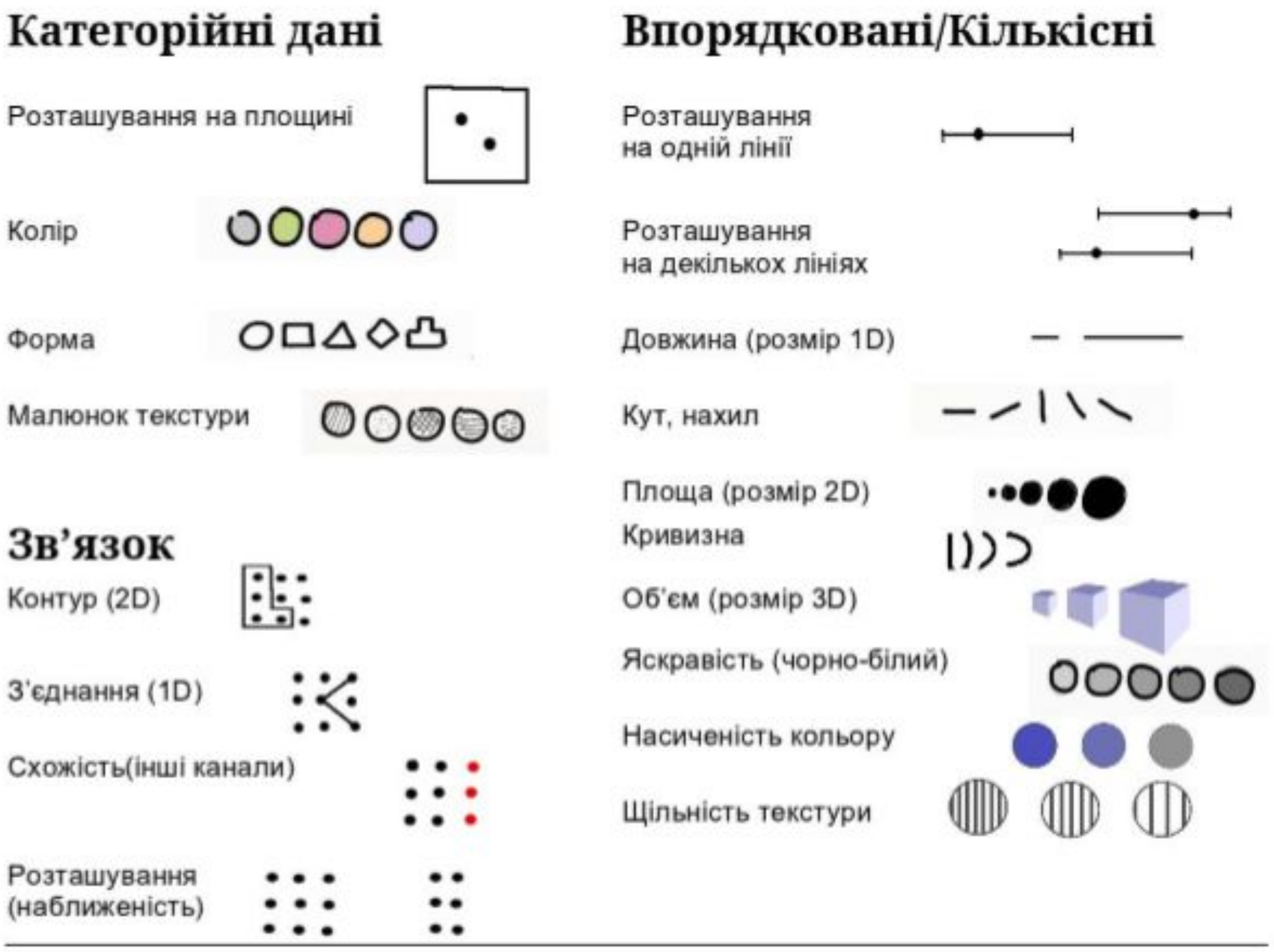

Польща
1960
1970
1980
1990
2000

Рейтинг (візуальних) каналів

За допомогою таблиці зображеної на наступному слайді ви зможете швидко обрати той візуальний канал для конкретних даних, який більш підходить.

Кодувати дані за допомогою різних величин площі – це не самий ефективний спосіб. Млинцеві діаграми – не є ефективними с точки зору показу кількісної інформації.

Рейтинг (візуальних) каналів



1. Акуратність представлення

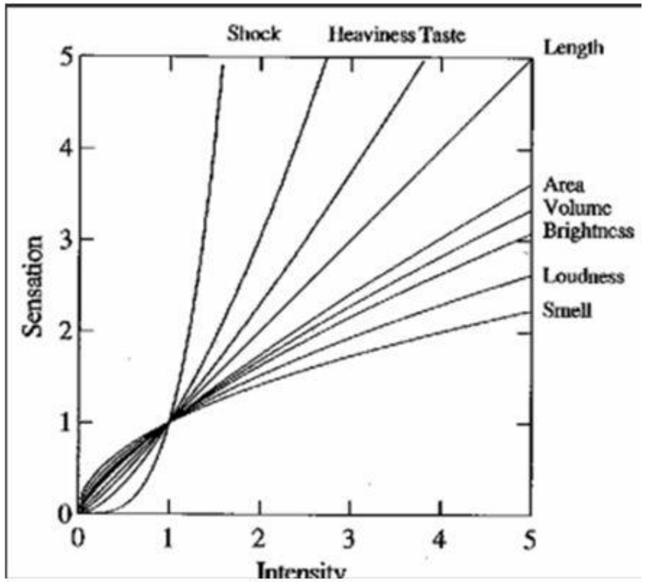
Візуальні судження після візуального стимулу - Психологічний степеневий закон Стівенса.

Stevens' Power Law

$$s(x) = ax^b$$

s is the sensation
x is the intensity of the attribute
a is a multiplicative constant
b is the power

$b > 1$: overestimate
$b < 1$: underestimate



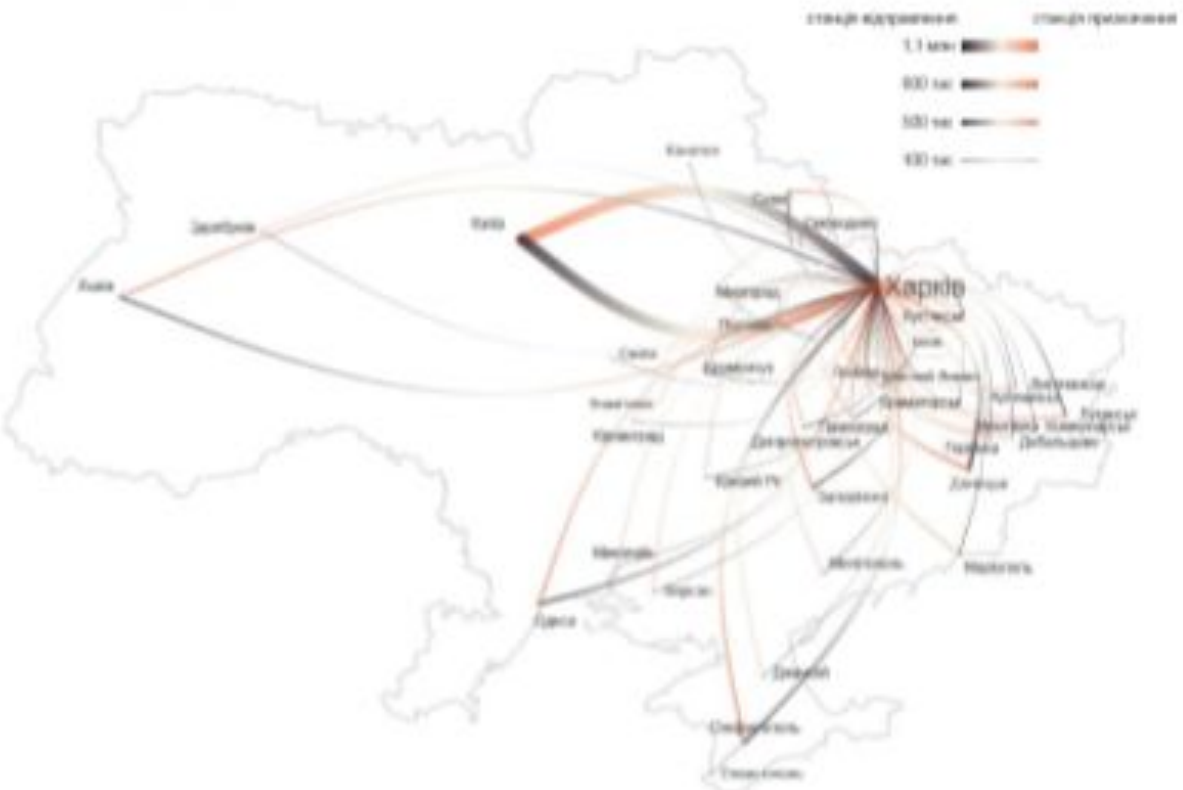
[graph from Wilkinson 99]

Стимул	Експонента	Умови
Звук	0.67	Тиск від 3000-Hz сигналу
Вібрація	0.95	частота 60 Hz на палець
Вібрація	0.6	частота 250 Hz на палець
Яскравість	0.33	5° мішень в темряві
Яскравість	0.5	Точкове джерело
Яскравість	5	Короткий спалах
Яскравість	1	Точкове джерело, короткий спалах
Освітленість	1.2	Альбедо сірого паперу
Довжина	1	Лінія
Площа	0.7	Квадрат
Насиченість	1.7	Перехід червоне-сіре
Смак	1.3	Сахароза
Електрошок	3.5	Ток через палець

2. Здатнсть розрізнити варіанти

Чи може людина побачити відмінності атрибутів візуального какналу?

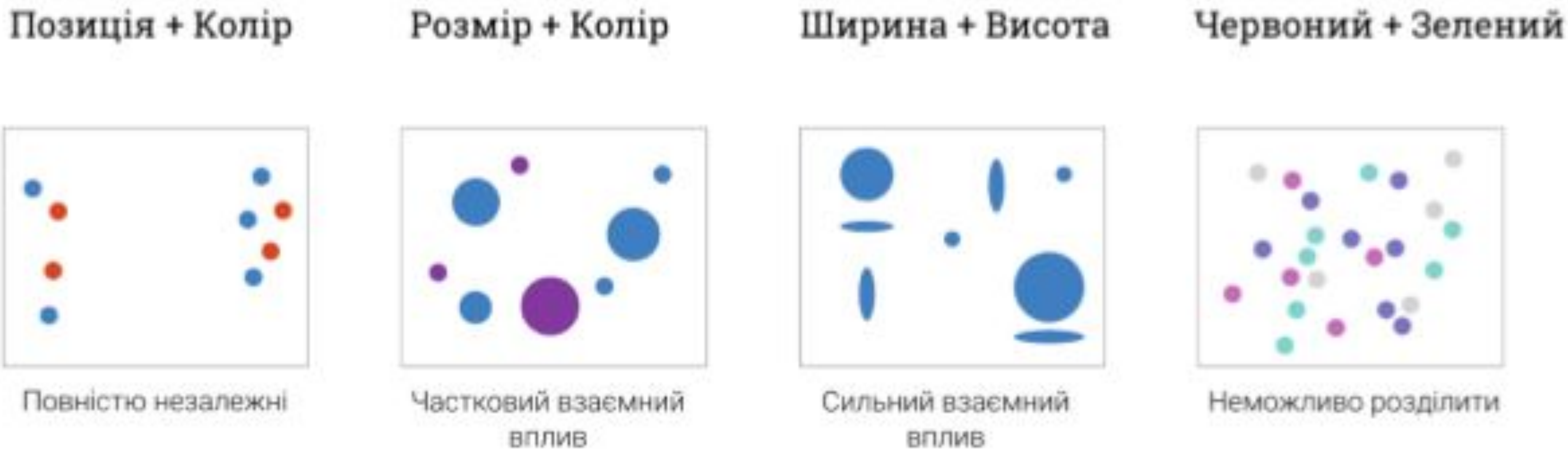
Наприклад, товщина лінії – скільки варіантів товщини повинно бути, щоб ми все ще їх розрізняли між собою?



3. Ортогональність (незалежність) каналів

Пари каналів - не є повністю незалежними і впливають один на одного.

- Позиція + колір - незалежні канали
- Розмір + колір - розмір впливає на відчуття кольору (на маленьких об’єктах його не видно)
- Горизонтальний розмір + вертикальний розмір (насправді, ми сприймаємо площу)
- Червоний + зелений - повністю нероздільні канали, ми не можемо сказати скільки в кольорі точок червоного, а скільки - зеленого (у нас по-іншому працює обробка кольору).


Позиція + Колір
Розмір + Колір
Ширина + Висота
Червоний + Зелений
Повністю незалежні
Частковий взаємний вплив
Сильний взаємний вплив
Неможливо розділити

Знаходження міток

4. Поп-аут - атрибут стрибає вам в очі (моментальне виявлення, паралельна обробка)

Багато каналів можуть забезпечити моментальне знаходження мітки серед інших міток (незалежно від їх кількості).

- колір
- орієнтація
- розмір
- форма
- близкість

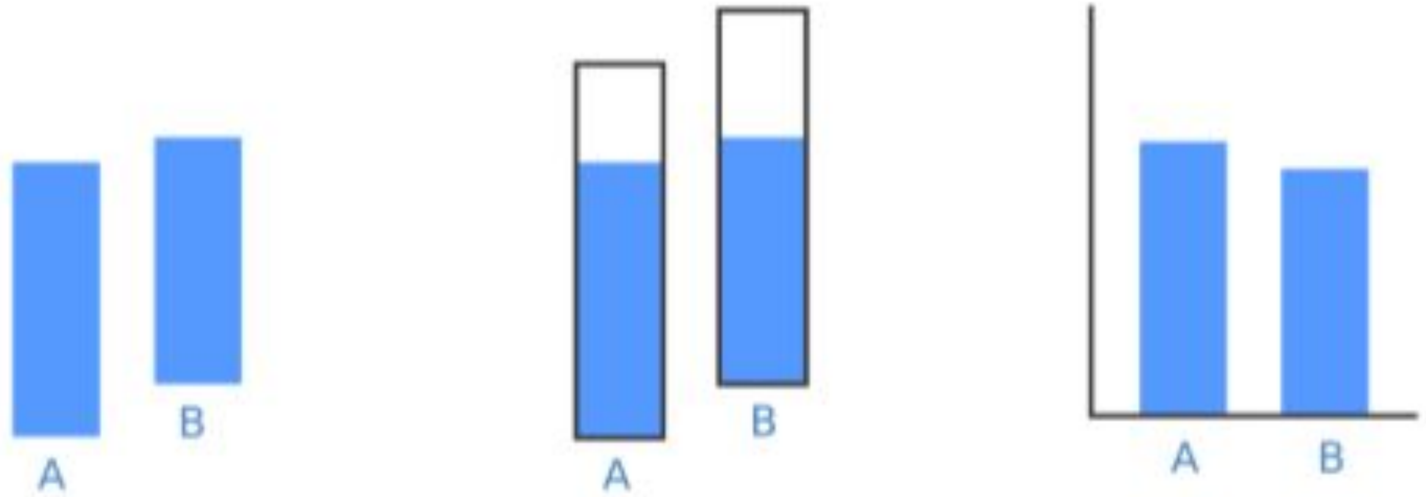
Якщо говорити про пари - більшість комбінацій каналів не дозволяють створювати поп-аут. Однак простір та колір і рух та форма - дозволяють. Інші - перетворюють паралельний пошук в послідовний, тобто значно повільніший (((Форма + колір))).

Три і чотири канали - немає поп-ауту взагалі.

Правило - для поп-ауту розраховуйте лише на один канал.

Поп-аут найчастіше використовують для показу найбільш важливої із важливих атрибутів.

Необхідно пам’ятати: всі наші оцінки - відносні якщо намалювати не вирівняні стовпчики, то їх важко порівняти якщо додати фрейм (рамку) одного розміру - одразу легше (міряємо пусту частину стовпчика) краще вирівняти по лінії



Без вирівнювання
Без рамки

Без вирівнювання
У рамці

Вирівнювання
Без рамки

Особливо це видно по кольору та по інтенсивності.

Сьогодні ми детально поговорили про ефективність візуального кодування.

Мені хотілося би, щоб ви запам’ятали - найголовнішу інформацію потрібно кодувати за допомогою найбільш сильного візуального каналу, у якому можна досягнути ефекту моментального виявлення ваших даних.

Дякую за увагу!