

Аналіз даних та статистичне виведення

Тиждень

1



аналіз даних

0.5 TB 

під час одного польоту генерує Boeing 787

100,000 перельотів

в середньому за день



IN AN
**INTERNET
MINUTE**

50,200
MOBILE APPS
DOWNLOADED¹



94
TWITTER
ACCOUNTS
CREATED²



2.4 MILLION
GOOGLE
SEARCHES³



1,389
UBER
RIDES⁴



30
IDENTITY
THEFTS⁵



2,083,333
MINUTES
USED ON
SKYPE CALLS⁶



142,361,111
EMAILS SENT
AND RECEIVED⁷



347,222
TWEETS⁸



120
LINKEDIN
ACCOUNTS
CREATED⁹



300
HOURS OF VIDEO
UPLOADED
ON YOUTUBE¹⁰



216,000
PHOTOS
POSTED TO
INSTAGRAM¹²

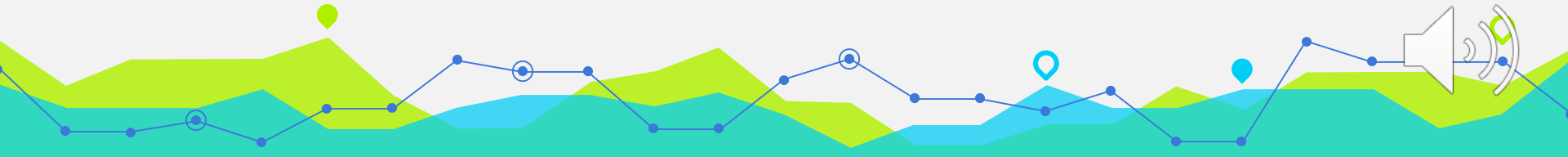


\$203,579
IN AMAZON
SALES¹¹

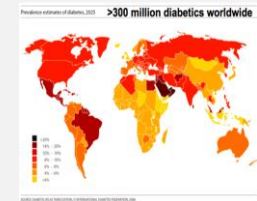
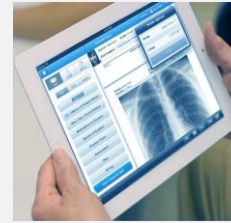


обмін даними

- **Uber** – найбільша в світі служба таксі, не володіє жодним авто
- **Facebook** - найпопулярніший власник медіа, не створює жодного контенту
- **Alibaba** - retailer, нічого не виробляє
- **Airbnb** – сервіс короткострокової оренди житла, не володіє жодним помешканням



коли дані мають сенс



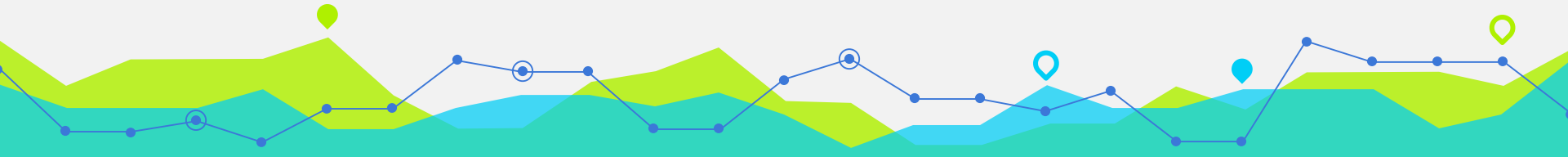
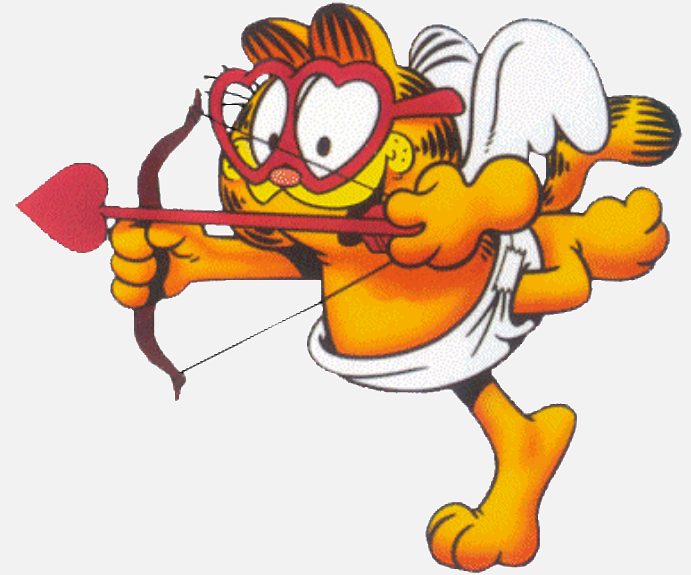
коли дані мають сенс

- Оцінка ризику трансгенних продуктів харчування або нових вакцин.
- Передбачення кількості захворювань на грип чи СНІД по регіонах.
- Передбачення результату наступних виборів
- Голосові асистенти для смартфонів
- Самокеровані автомобілі



eHarmony

- сервіс для пошуку пари для шлюбу
- сукупний дохід > 1 мільярда доларів
- близько 4% шлюбів у США в 2012
- більше 33 млн користувачів у понад 150 країнах





процес аналізу даних

90% відсотків часу



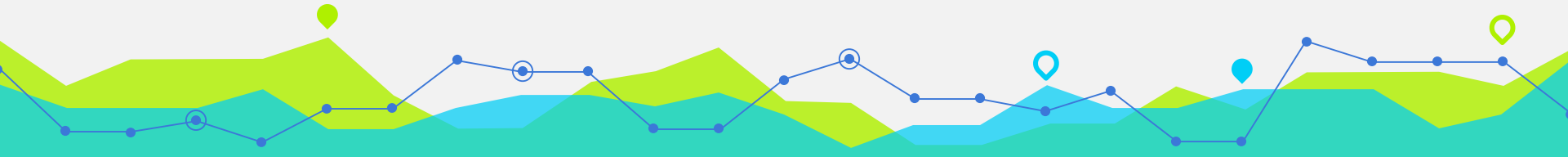
Підготовка до побудови
моделі
(збір даних, очищення,
трансформація)

Побудова та
валідація моделі

решта 90% відсотків
часу



Трактування та
презентація
результатів





статистика



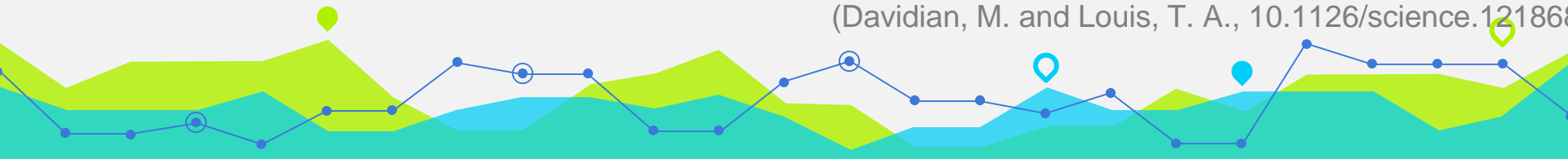
варіативність та невизначеність





Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances

(Davidian, M. and Louis, T. A., 10.1126/science.1218661)



статистика – наука про збір, організацію, аналіз та трактування даних

Описова статистика –
вивчає властивості
спостережуваних даних

Вивідна статистика –
виводимо припущення
про властивості розподілу
даних з яких походять
спостережувані дані.



- чи є залежність між кількістю злочинів та фазою Місяця?
- яка ймовірність викликати Uber в Києві?
- побудувати довірчий інтервал часу, за який ви потрапляєте на роботу
- проводити опитування та трактувати їх результати



Чи вважаєте Ви себе європейцем?



	Безумовно, так	Скоріше, так	Скоріше, ні	Безумовно, ні	Важко відповісти
Тра.13	10	24.3	29.1	25.9	10.7
Тра.14	10	27.6	28	24.2	10.3

Загальнонаціональне дослідження громадської думки населення України було проведене Фондом «Демократичні ініціативи ім.Ілька Кучеріва» разом із Центром Разумкова з 14 по 18 травня 2014 р. Опитування проводилося в усіх регіонах України, за винятком Криму. Усього було опитано 2011 респондентів за вибіркою, репрезентативною для дорослого населення України (старше 18 років). Теоретична похибка вибірки не перевищує 2.3%.



дані

типи даних

кількісні

дискретні

- кількість дітей у сім'ї
- кількість медалей олімпійської збірної

неперервні

- ріст
- вага
- заробітна плата

категоріальні

невпорядковані

- імена
- назви міст
- група крові

бінарні

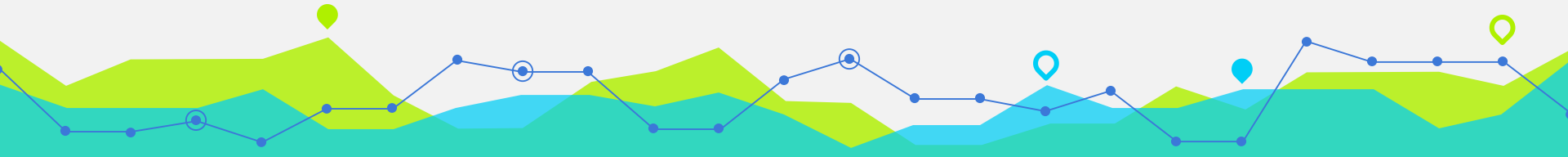
- так, ні
- 0, 1

впорядковані

- бакалавр, магістр, доктор
- погоджуюсь, частково погоджуюсь, не погоджуюсь, важко відповісти

рівні виміру

- точка: 5, 7.5
- інтервал: $[0, 5]$, $[5, 7]$
- відношення: 95%



матриця даних

місто	кількість кімнат	площа	ціна(грн)
Київ	3	80	2 162 162
Київ	1	58	874 000
Київ	3	124	2 837 838
Київ	2	67	1 616 216
Київ	2	67	535 135
Івано-Франківськ	3	77	754 600
Івано-Франківськ	1	44	445 946
Запоріжжя	3	60	540 541
Одеса	1	38	500 000
Одеса	2	58	837 838



частотні таблиці

місто	кількість спостережень
Київ	5
Івано-Франківськ	2
Запоріжжя	1
Одеса	2

місто	кількість спостережень
Київ	50%
Івано-Франківськ	20%
Запоріжжя	10%
Одеса	20%







центральна
тенденція

середнє значення

3, 2, 5, 6, 7, 2, 3, 3

$$\bar{x} = \frac{3+2+5+6+7+2+3+3}{8} = \frac{31}{8} = 3.875$$

$$\bar{x} = \frac{\sum x}{N}$$

Місце серед членів ООН	Місце серед усіх територій	Країна	Загальна очікувана тривалість життя при народженні	Чоловіча очікувана тривалість життя при народженні	Жіноча очікувана тривалість життя при народженні
	1	 Макао	84,36	81,39	87,47
1	2	 Андорра	82,51	80,33	84,84
2	3	 Японія	82,12	78,8	85,62
3	4	 Сінгапур	81,98	79,37	84,78
4	5	 Сан-Марино	81,97	78,53	85,72
	6	 Гонконг	81,86	79,16	84,79
5	7	 Австралія	81,63	79,25	84,14
6	8	 Канада	81,23	78,69	83,91
7	9	 Франція	80,98	77,79	84,33
8	10	 Швеція	80,86	78,59	83,26

медіана

3, 2, 5, 6,	7, 2, 4, 3
2, 2, 3, 3,	4, 5, 6, 7

$$\frac{3+4}{2} = 3.5$$

значення, яке ділить вибірку навпіл

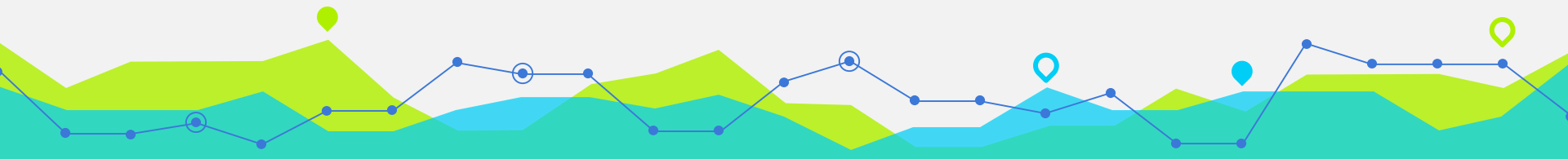


мода

3, 2, 5, 6, 7, 2, 3, 3

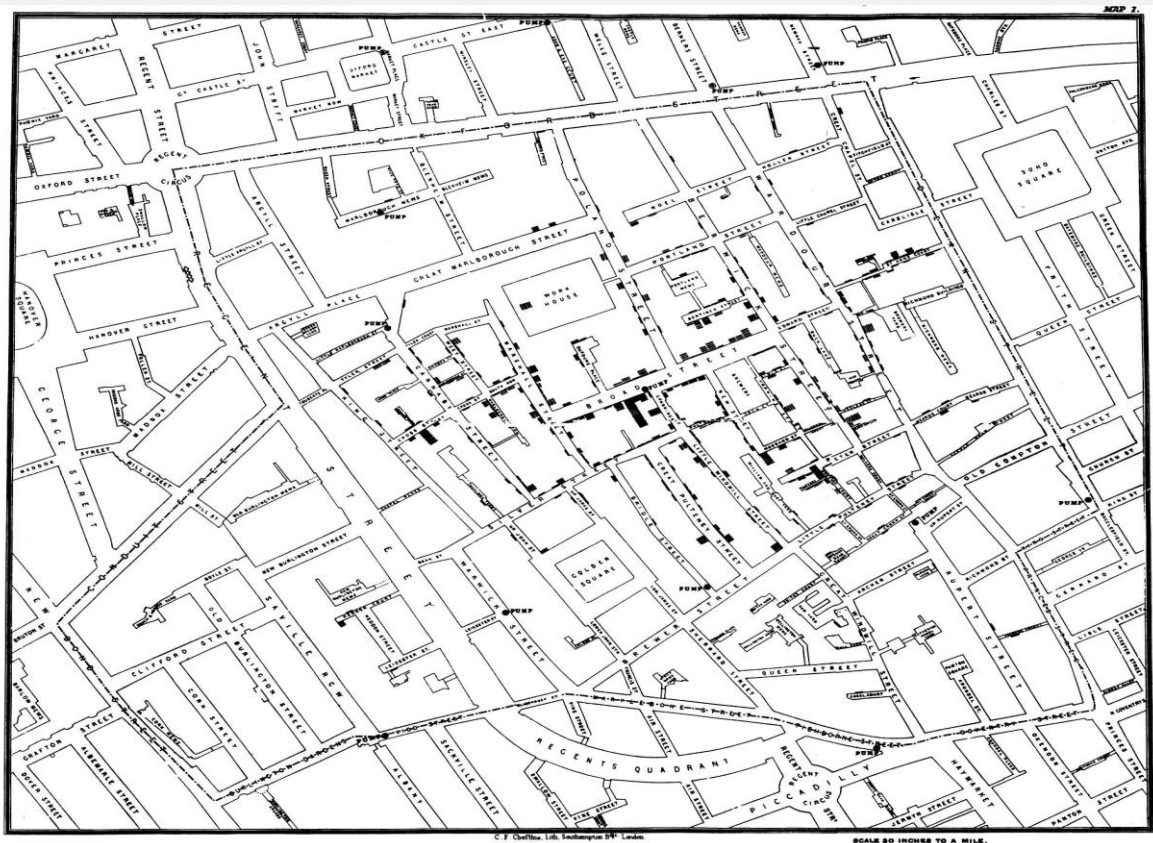
2, 2, 3, 3, 3, 5, 6, 7

значення, яке найчастіше трапляється

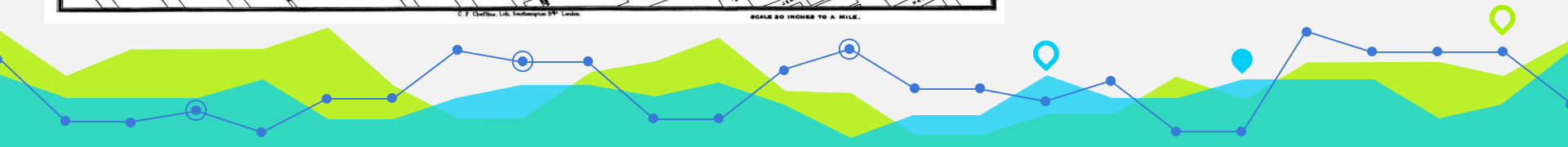


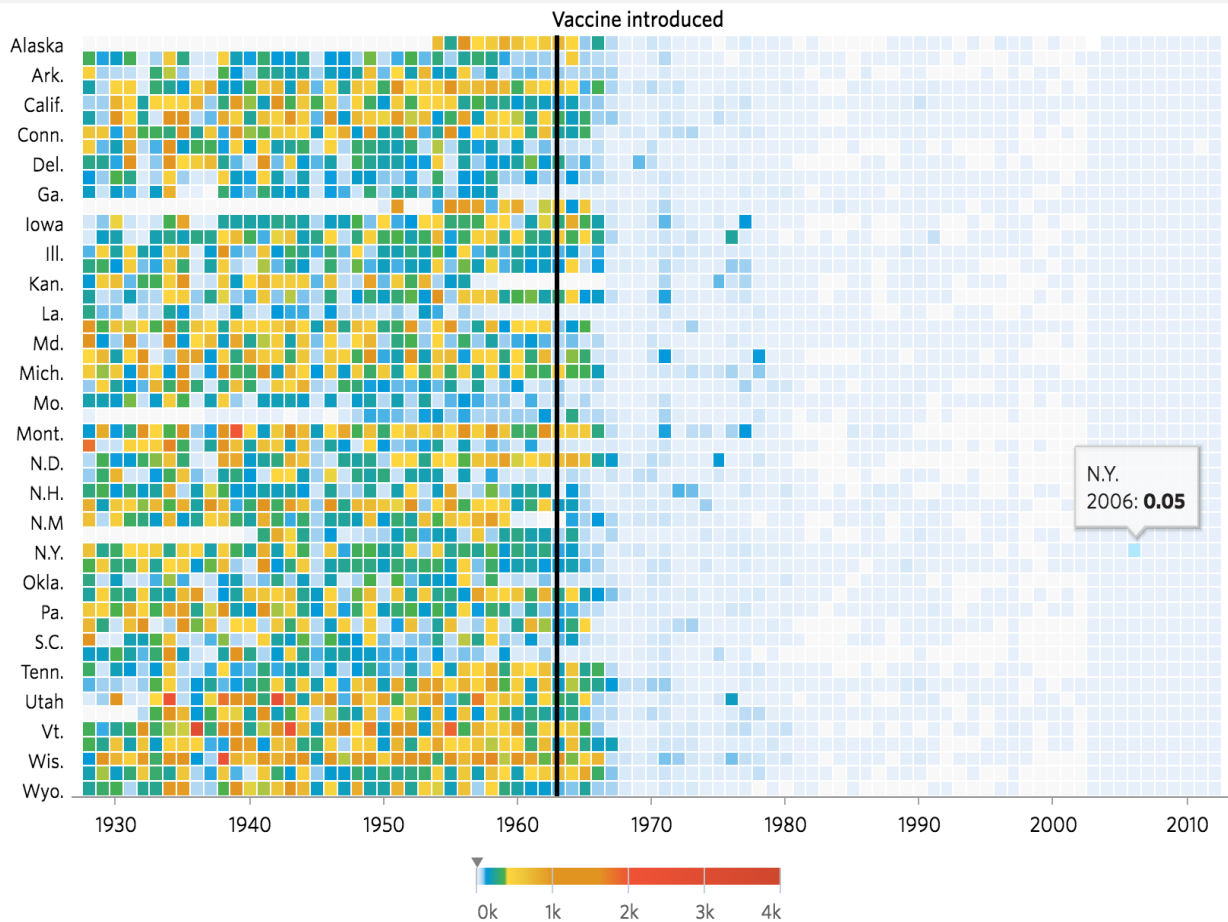


візуалізація



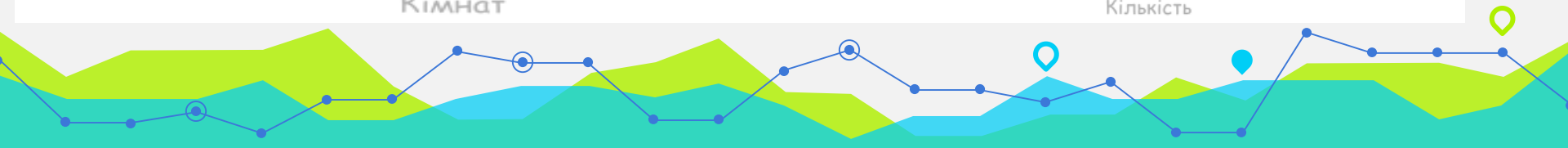
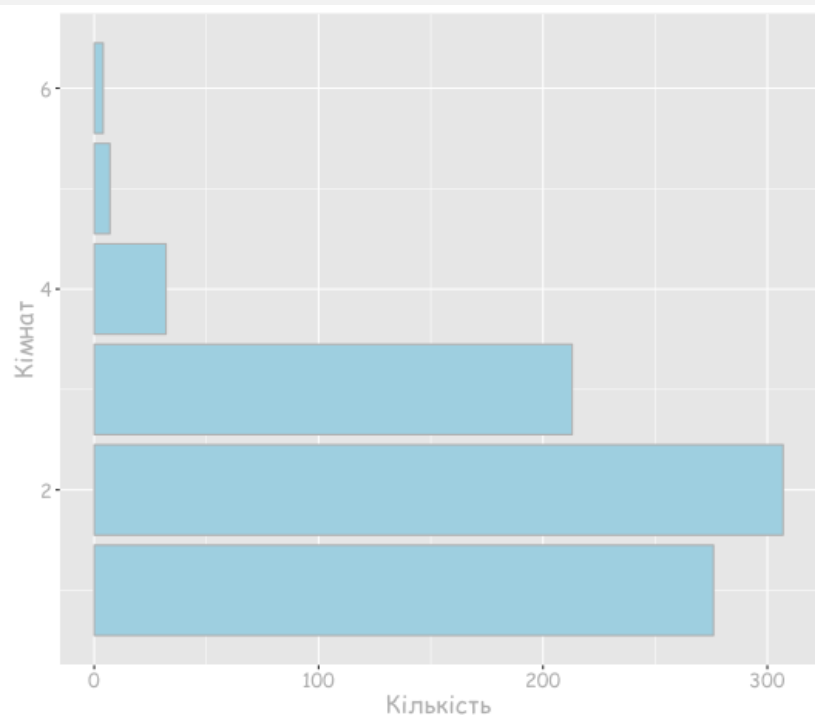
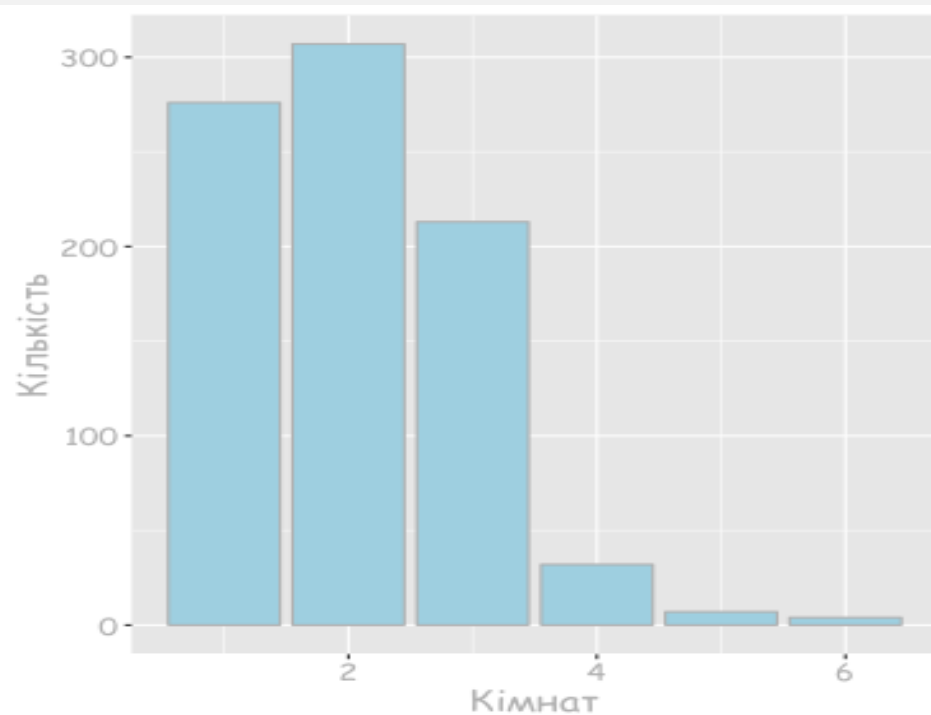
В 1848 в лондонському районі Сохо було зафіксовано спалах холери під час якого загинуло 616 жителів. Під час цього спалаху Лікар епідеміолог Джон Сноу на основі візуального аналізу даних зробив припущення, що джерелом зараження є вода.

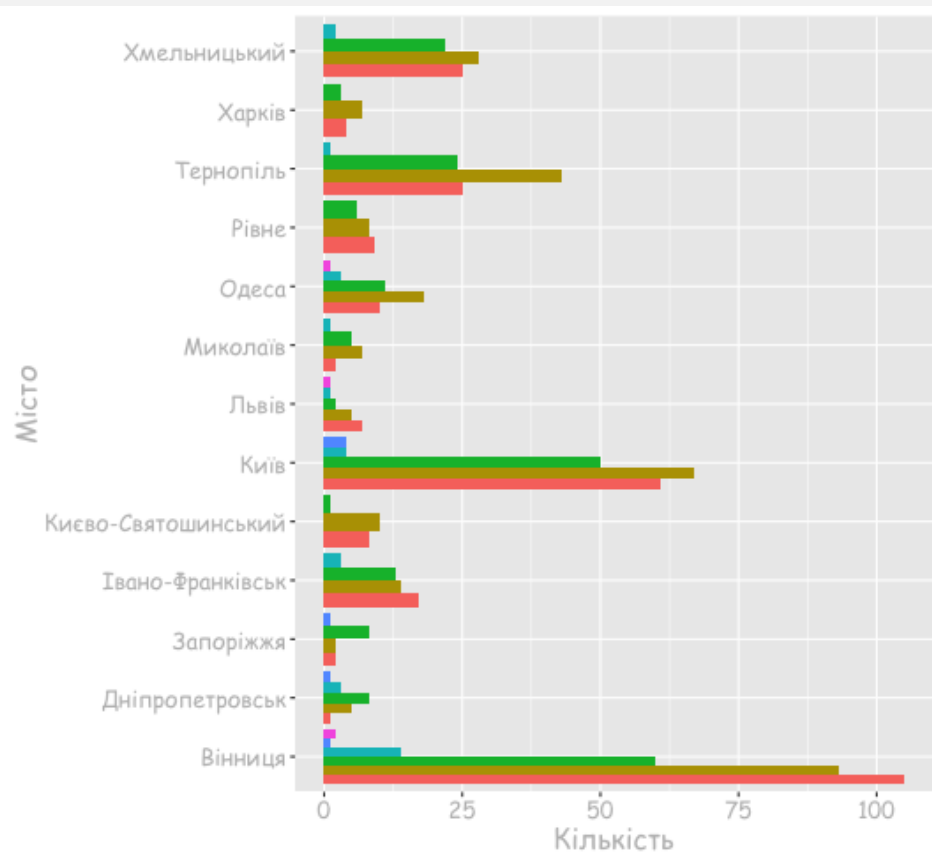
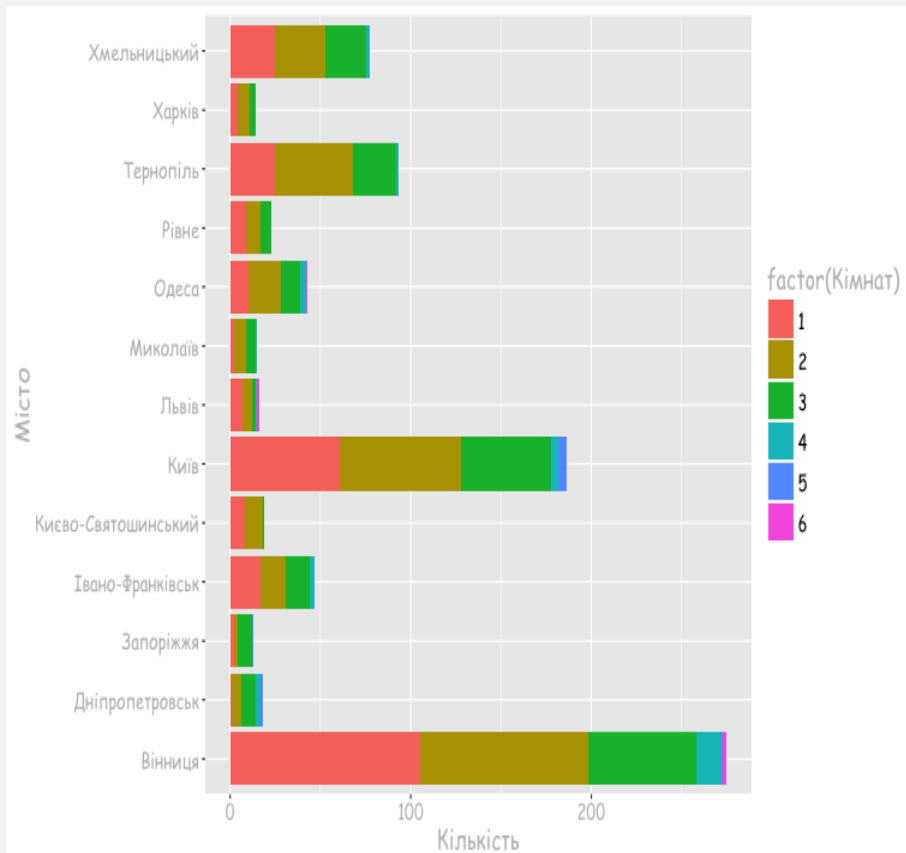




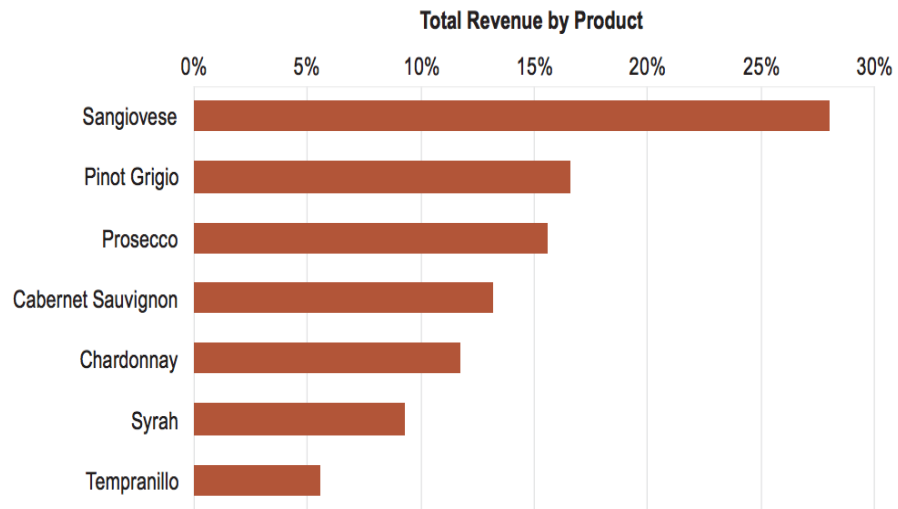
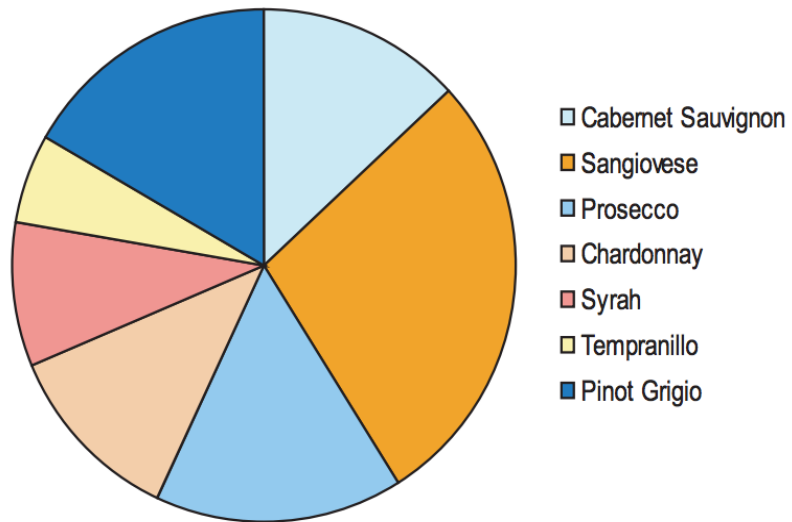
Видання WSJ підготувало інтерактивні візуалізації, які відображають рівень захворюваності на кір, поліомієліт, кашлюк та інші хвороби до і після запровадження вакцини. Дані показують рівень захворюваності у США протягом 80 років

Стовпчикова діаграма

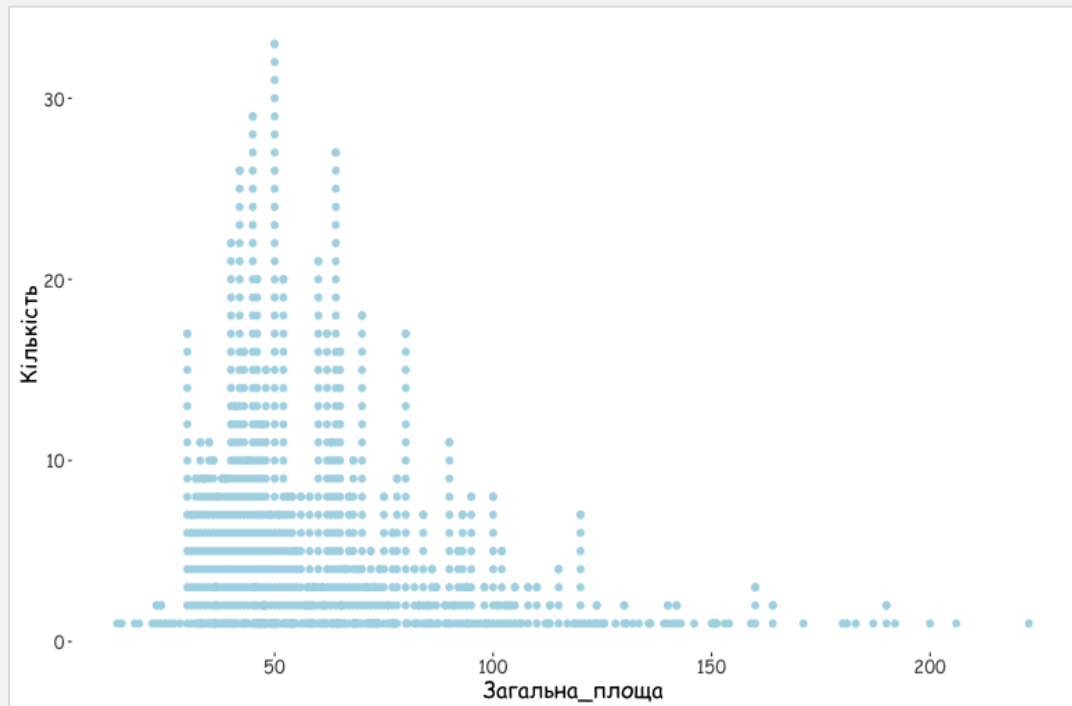




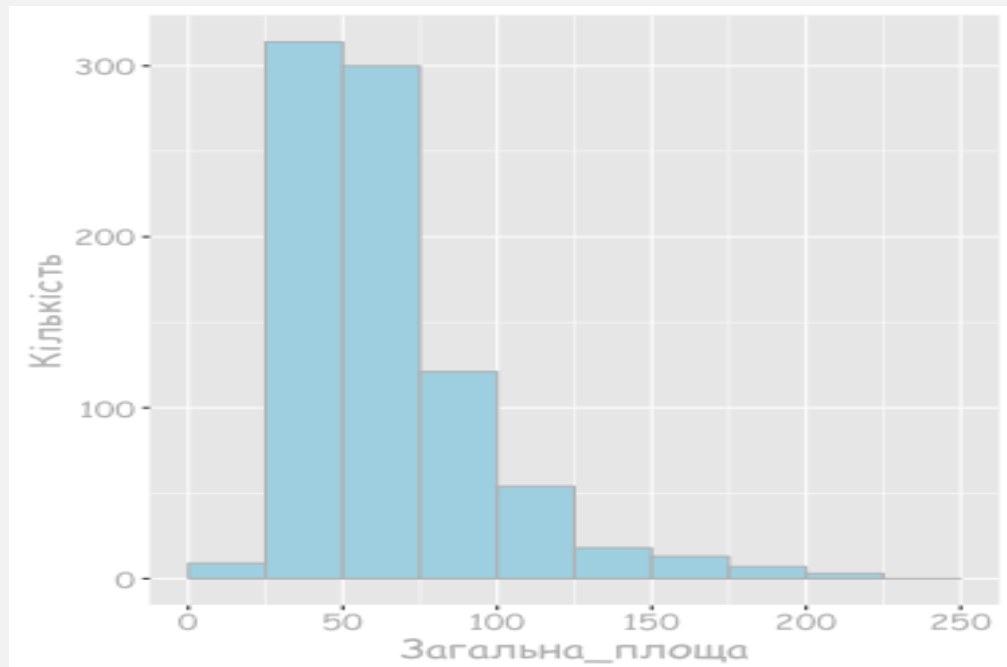
кругова діаграма



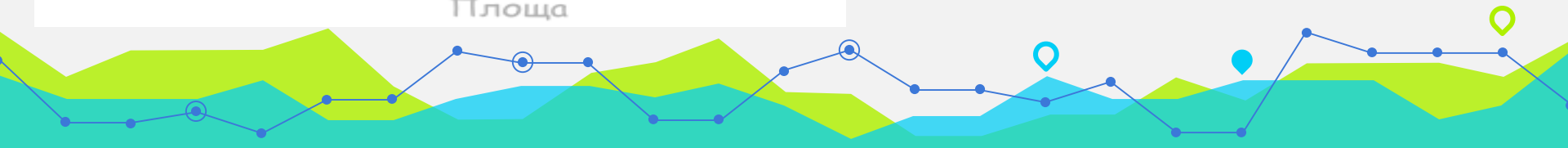
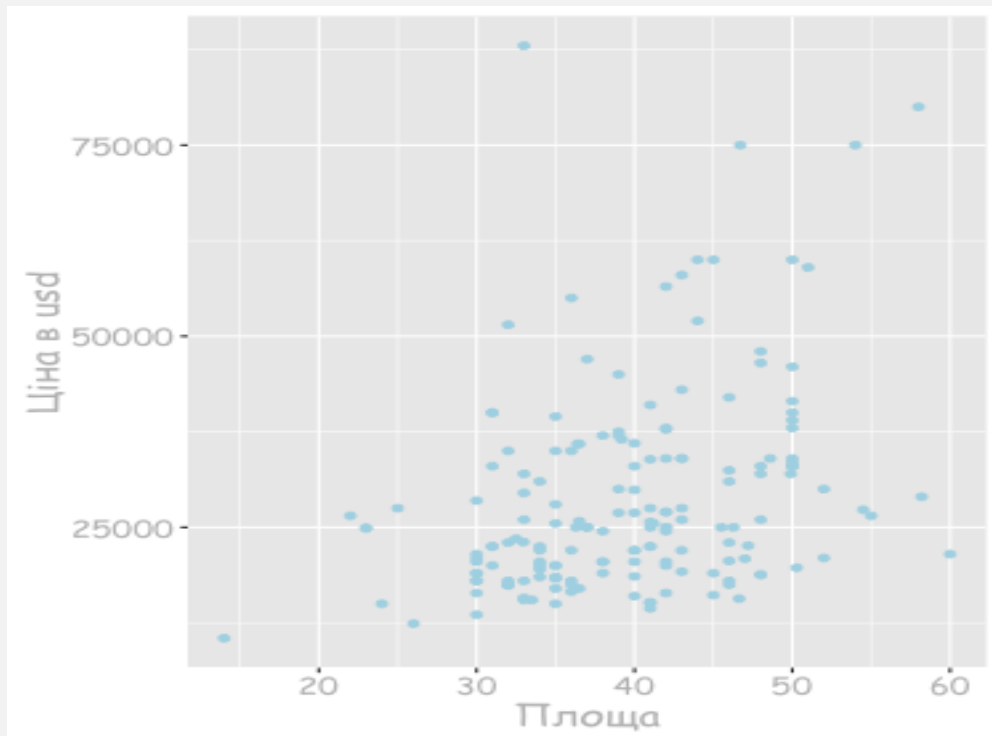
точкові графіки



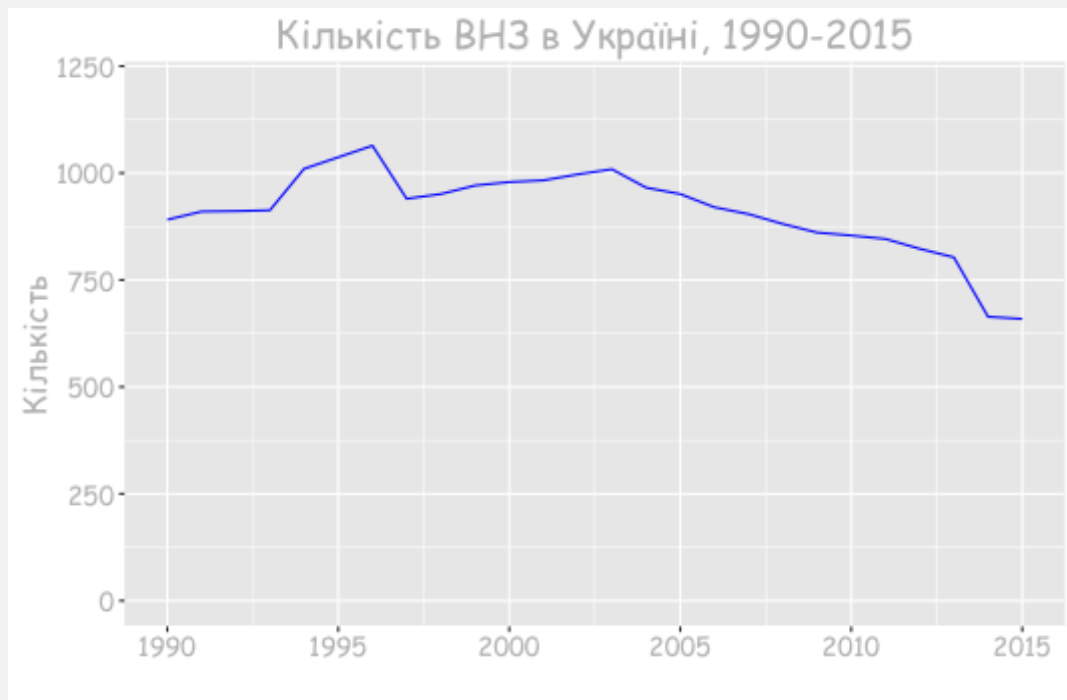
гістограма



діаграма розсіювання



лінійний графік



який тип діаграми краще застосовувати?

- Порівнювати значення: стовпчикова діаграма, лінійний графік, графік розсіювання
- Зрозуміти композицію: (виділити складові) - стовпчикова діаграма, кругова діаграма
- Оцінити розподіл даних: лінійний графік, графік розсіювання, стовпчикова діаграма, гістограма
- Зрозуміти тренд: лінійний графік, стовпчикова діаграма
- Зрозуміти відношення між даними: лінійний графік, графік розсіювання





трактування
результатів

парадокс Сімпсона

Факультет А

	Подало заяв	Прийнято	Відсоток прийнятих
Чоловіки	900	450	50%
Жінки	100	80	80%

Факультет Б

Чоловіки	100	10	10%
Жінки	900	180	20%



парадокс Сімпсона

Факультет А

	Подало заяв	Прийнято	Відсоток прийнятих
Чоловіки	900	450	50%
Жінки	100	80	80%

Факультет Б

Чоловіки	100	10	10%
Жінки	900	180	20%

Обидва

Чоловіки	1000	460	46%
Жінки	1000	260	26%

