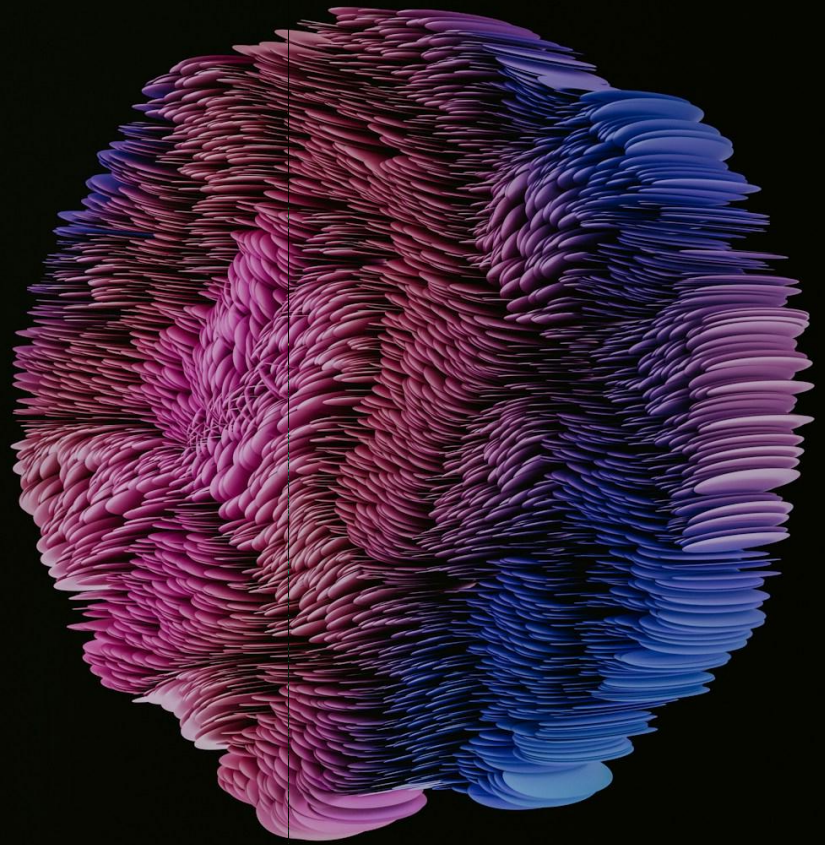
Культура роботи з відкритими даними: вимоги до даних та їх машиночитаності, формати та структура



Дані з точки зору аналізу

Три характеристики:

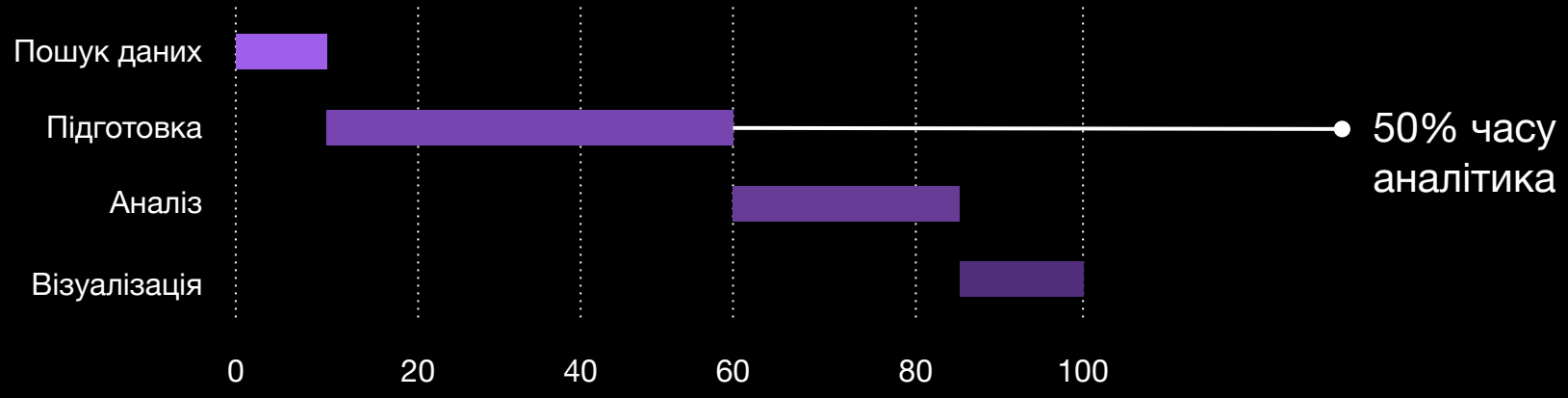
01 Категоризованість

02 Вимірюваність

03 Порівнюваність

Підготовка даних

Станом на серпень 2021 року:



Машиночитаний формат

Формат, що дозволяє автоматизоване оброблення даних електронними засобами. Це означає, що не тільки людина, але й комп'ютер має розпізнавати дані.

Дані: Дія. Відкриті дані

Людиночитаний формат

Формат, який можна відкрити за допомогою простого текстового редактора (наприклад, Блокнота) і щось там розібрати або відредагувати. Тобто цілком читабельний текст у зображенні не є ані машиночитаним ані людиночитаним.

Дані: Відкритий посібник з відкритих даних для громадських організацій, журналістів, і не тільки…

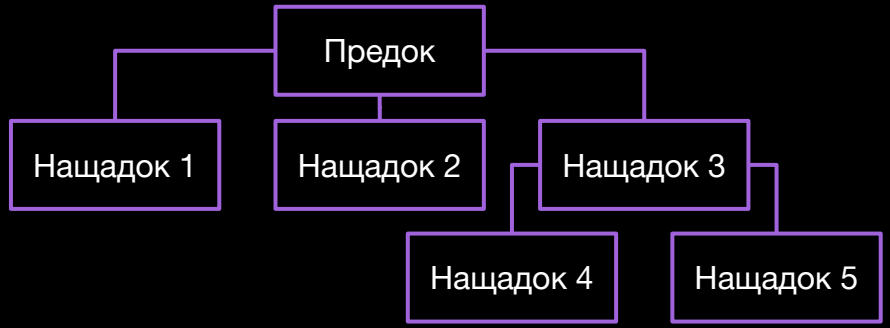
Існує два основні типи структурованих даних

Табличні дані

Стовпчик 1	Стовпчик 2	Стовпчик 3	Стовпчик 4
Рядок 2	1	2	3
Рядок 3	4	5	6
Рядок 4	7	8	9

Формати: XLS, XLSX, ODS, CSV

Ієрархічні



Формати: JSON, XML

Приклади табличних й ієрархічних даних

```
gender,age,birthday_date,percent,votes
F,46.0,1973-11-04,10.85,8802.0
M,41.0,1978-06-27,11.12,9018.0
M,35.0,1984-03-31,4.81,3905.0
F,36.0,1983-05-24,4.29,3479.0
M,54.0,1965-05-31,9.58,7765.0
M,42.0,1977-01-24,6.9,5595.0
M,44.0,1975-10-13,15.11,12249.0
F,42.0,1977-10-03,1.32,1070.0
M,36.0,1983-03-24,26.62,21582.0
F,32.0,1987-04-21,3.75,3046.0
M,44.0,1975-03-06,2.74,2228.0
M,28.0,1991-04-09,0.95,774.0
M,53.0,1966-08-24,1.89,1537.0
```

Фрагмент CSV-файлу

```
{
  "documents": [
    {
      "id": 1376029791,
      "edrpou": "41274732",
      "documentNumber": "8",
      "documentDate": "2018-10-11",
      "signDate": "2018-10-12",
      "amount": 10000,
      "currency": "UAH",
      "currencyAmountUAH": 0,
      "contractors": [
        {
          "contractorType": 1,
          "name": "Чернявський С. О.",
          "firstName": "Сергій"
           ...
```

Фрагмент JSON-файлу

```
<DECLAR
xmlns:xsi="http://www.w3.org/2001/XMLSch
ema-instance"
xsi:noNamespaceSchemaLocation="S0100113.
XSD">
  <DECLARHEAD>
    <TIN>39804651</TIN>
    <C_DOC>S01</C_DOC>
    <C_DOC_CNT>458</C_DOC_CNT>
    <C_REG>462</C_REG>
    <C_RAJ>227</C_RAJ>
    <PERIOD_MONTH>6</PERIOD_MONTH>
    <PERIOD_TYPE>3</PERIOD_TYPE>
    <PERIOD_YEAR>2019</PERIOD_YEAR>
    <D_FILL>26072019</D_FILL>
    <SOFTWARE>MEDOC</SOFTWARE>
  </DECLARHEAD>
</DECLAR>
```

Фрагмент XML-файлу

csv — «comma-separated values»

Але значення можуть розділятися не лише комами, а по суті будь-яким символом.

1. Countries GDP.csv

```
Country,Agricolture,Industry,Services
United States,209027,3327015,13882883
China,944615,4422042,5013724
Japan,55396,1269492,3296063
Germany,30876,1084533,2744138
United Kingdom,20616,618481,2306049
France,54091,520981,2271817
Brazil,127063,644729,1581233
Italy,42959,519804,1585189
India,356319,528335,1165204
Russia,72441,668686,1116334
```

Так виглядає csv у блокноті

	A	B	C	D
1	Country	Agricolture	Industry	Services
2	United States	209027	3327015	13882883
3	China	944615	4422042	5013724
4	Japan	55396	1269492	3296063
5	Germany	30876	1084533	2744138
6	United Kingdom	20616	618481	2306049
7	France	54091	520981	2271817
8	Brazil	127063	644729	1581233
9	Italy	42959	519804	1585189
10	India	356319	528335	1165204
11	Russia	72441	668686	1116334

А це csv, відкритий в Excel

Приклад JSON файлу


```
{
  "name": "ПРИВАТНЕ АКЦІОНЕРНЕ ТОВАРИСТВО \"ЗАВОД \"КУЗНЯ НА РИБАЛЬСЬКОМУ\"",
  "shortName": "ПРАТ \"ЗАВОД \"КУЗНЯ НА РИБАЛЬСЬКОМУ\"",
  "address": "04176, м.Київ, Подільський район, ВУЛИЦЯ ЕЛЕКТРИКІВ, будинок 26",
  "director": "ШАНДРА ВАЛЕРІЙ ОЛЕКСАНДРОВИЧ",
  "status": "Не перебуває в процесі припинення",
  "economicActivity": {
    "code": "30.11",
    "description": "Будування суден і плавучих конструкцій"
  },
  "founders": [
    "АКЦІОНЕРИ ЗГІДНО РЕЄСТРУ, розмір частки - 168087346,00 грн.",
    "КІНЦЕВИЙ БЕНЕФІЦІАРНИЙ ВЛАСНИК - ТІГІПКО СЕРГІЙ ЛЕОНІДОВИЧ, 13.02.1960 Р.Н., УКРАЇНА, , М.КИЇВ, ВУЛ. ГОРОДЕЦЬКОГО, БУД. 12, КВ. 69, ІПН. (БАЙЛИКАН ЛІМІТЕД, Т.А.С.ОВЕРСІАС ІНВЕСТМЕНТС ЛІМІТЕД, ЕВІНЗ ЛІМІТЕД)"
  ],
  "actualDate": "2021-05-24T22:44:30+03:00",
  "registrationDate": "1996-12-25T00:00:00+02:00"
}
```

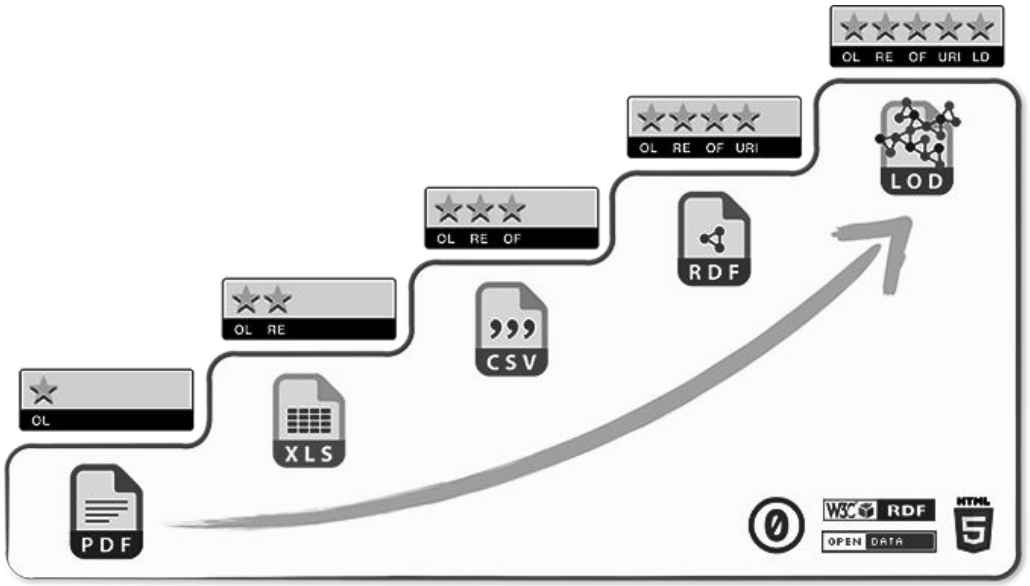
Приклад XML файлу

```
<DECLARBODY>
  <REP_NYEAR>2021</REP_NYEAR>
  <REP_NMONTH>12</REP_NMONTH>
  <FIRM_NAME>Комунальне виробниче підприємство "Краматорська тепломережа"  Краматорської міської ради</FIRM_NAME>
  <FIRM_EDRPOU>00131133</FIRM_EDRPOU>
  <FIRM_TERR>ДОНЕЦЬКА</FIRM_TERR>
  <FIRM_OGU xsi:nil="true" />
  <FIRM_SPODU xsi:nil="true" />
  <FIRM_KVEDNM>Постачання пари, гарячої води та кондиційованого повітря</FIRM_KVEDNM>
  <FIRM_KVED>35.30</FIRM_KVED>
  <REP_KSKZ xsi:nil="true" />
  <FIRM_RUK>Безкоровайний Григорій Ілліч</FIRM_RUK>
  <FIRM_BUH>Шаванова Оксана Іванівна</FIRM_BUH>
  <FIRM_OPFCD>150</FIRM_OPFCD>
  <FIRM_OPFNM>Комунальне підприємство</FIRM_OPFNM>
```

Словник форматів структурованих даних

CSV	comma-separated values
TSV	tab-separated values
XML	eXtensible Markup Language
XLS, XLSX	документ Excel
JSON	JavaScript Object Notation
KML	Keyhole Markup Language

П’ятизіркова концепція машиночитаних даних та “Інтернет даних”



Структуровані VS неструктуровані дані

Вивчаючи статистику нашого опитування, ми виявили цікаві факти про учасників. Серед 10 респондентів було 5 жінок та 5 чоловіків. При цьому середній вік учасників становив приблизно 34 роки. Ці дані допомагають нам отримати уявлення про вікову структуру опитаних осіб.

Особливий інтерес представляє сімейний статус респондентів. З 5 жінок лише 2 вказали, що вони заміжні. Щодо чоловіків, 3 з 5 відповіли, що вони одружені. Цікаво, що більшість учасників опитування має дітей.

Поглиблюючись у деталі, ми розглядали також рік народження учасників. Середній рік народження виявився 1990 роком, що підтверджує наші спостереження з віковою структурою. Незалежно від статі та віку, опитані показали велику активність і готовність взяти участь у нашому дослідженні, що вказує на високий рівень їхнього зацікавлення.

Питання до тексту:

1. Скільки всього жінок та чоловіків?
2. Скільки років наймолодшому та найстаршому учаснику?
3. Скільки років наймолодшому одруженому чоловіку та наймолодшій одруженій жінці?
4. Який середній вік жінок та середній вік чоловіків?

Неструктуровані дані

Ім'я	Стать	Вік	Наявність дітей	Заміжня/Одружений	Рік народження
Олексій	Ч	34	Так	Ні	1989
Ірина	Ж	28	Ні	Так	1995
Петро	Ч	42	Так	Так	1981
Олена	Ж	38	Так	Ні	1985
Дмитро	Ч	25	Ні	Ні	1998
Анна	Ж	30	Ні	Ні	1993
Віталій	Ч	45	Так	Так	1978
Наталія	Ж	29	Так	Так	1994
Сергій	Ч	36	Ні	Так	1987
Марія	Ж	33	Так	Ні	1990

- Середній вік: 34 роки
- Кількість жінок: 5
- Кількість чоловіків: 5
- Кількість заміжніх жінок: 2
- Кількість одружених чоловіків: 3

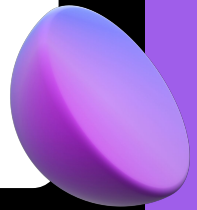
https://docs.google.com/spreadsheets/d/12IDlH4S3nMV6nmk46XlesWhWl_DwViro5k284oHhvq0/edit?usp=sharing

OpenData Academy

Структуровані дані

“Охайні дані”

Спосіб організації даних (вимірів та показників) в окремому файлі.



Концепція “охайних даних”

Загальні принципи акуратних даних викладені Гедлі Вікгемом (Hadley Wickham) у його статті **«Tidy Data»**, опублікованій у 2014 році:

- Кожна **змінна** (**variable**), яку ви вимірюєте, повинна бути в **одному стовпці**.
- Кожне окреме **спостереження** (**observation**) цієї змінної — в **окремому рядку**.
- Для **кожного «виду»** змінної має бути одна таблиця.
- Якщо у вас є **декілька таблиць,** вони повинні включати стовпець (**ідентифікатор, код**) у таблиці, завдяки якому їх можна поєднати.

“Погані дані”

Дата	Місто	Показник
2015-08-01	Київ	1 678 910
01-07-2015	Киев	1,567,890
1 черв 2015	Кийів	1 млн 456 тис 789
2015.06.01	Львів	1.345.678
2015/07/01	Львов	1 234 567

Дата	Місто	Показник
2015-08-01	Київ	1678910
2015-07-01	Київ	1567890
2015-06-01	Київ	1456789
2015-06-01	Львів	1345678
2015-07-01	Львів	1234567

“Погані дані”

Область	2012	2013	2014
Вінницька	35441	37323	39184
Волинська	19546	20609	21971
Дніпропетровська	95349	99995	109545
Донецька	128767	135362	114135

Область	Рік	Дохід населення, млн грн
Вінницька	2012	35441
Волинська	2012	19546
Дніпропетровська	2012	95349
Донецька	2012	128767
Вінницька	2013	37323
Волинська	2013	20609
Дніпропетровська	2013	99995
Донецька	2013	135362
Вінницька	2014	39184
Волинська	2014	21971
Дніпропетровська	2014	109545
Донецька	2014	114135

“Погані дані”

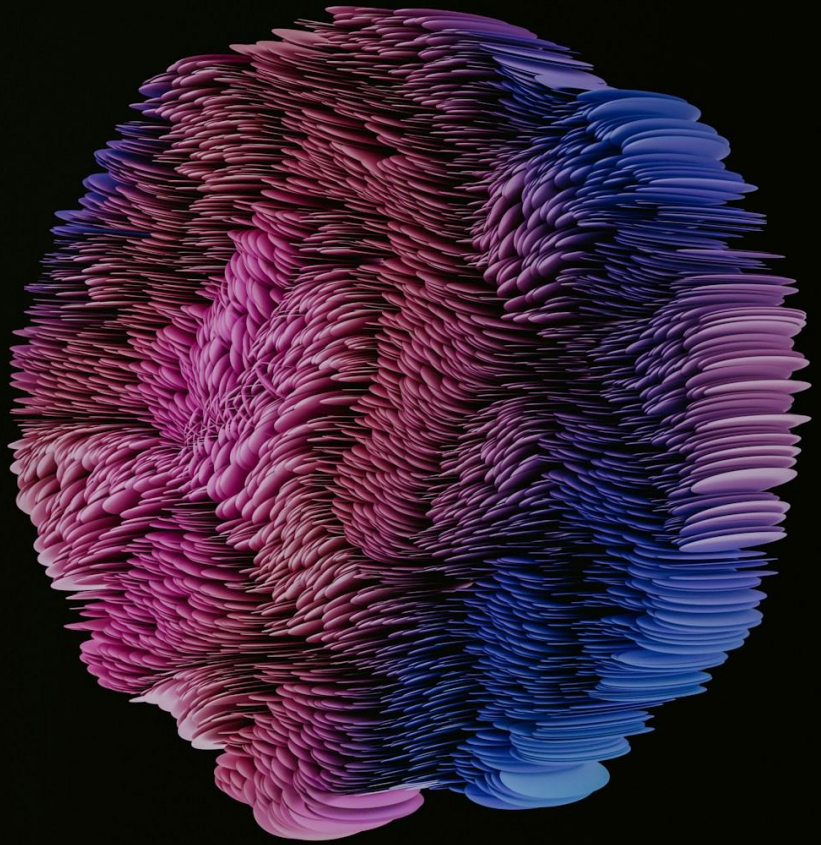
	Чисельність безробітного населення в середньому за І півріччя 2017р.			
	усього, у віці 15-70 років		з нього працездатного віку	
	тис. осіб	у % до економічно активного населення відповідного віку	тис. осіб	у % до економічно активного населення відповідного віку
Україна	**1709,7**	**9,6**	**1709,4**	**10,0**
Вінницька	79,1	10,9	79,1	11,3
Волинська	53,0	12,7	53,0	12,8
Дніпропетровська	128,0	8,4	128,0	8,7

Погано структуровані дані

Область	тис. осіб	%	Вік
Україна	1709,7	9,6	15-70 років
Україна	1709,4	10	працездатного віку
Вінницька	79,1	10,9	15-70 років
Вінницька	79,1	11,3	працездатного віку
Волинська	53	12,7	15-70 років
Волинська	53	12,8	працездатного віку
Дніпропетровська	128	8,4	15-70 років
Дніпропетровська	128	8,7	працездатного віку

Ті ж самі дані у гарній структурі

Табличні дані



Яка структура неправильна?

❶ Ідентифікатор, найменування	Дата реєстрації	Електронна пошта
❷ (1)	(2)	(3)
ТОВ «АБВ», 01234567* ❶	2010-02-03	sales@abv.com
ГО «Актив», 12345678 ❶	2011-04-05	contact@active.org
КП «Поліклініка №1», 23456789 ❶	2012-06-07	mail@clinic1.com
Луцька міська рада, 34567890 ❶	1995-09-10	pr@lutsk.gov.ua
❸❹ **Фізичні особи-підприємці**		
ФОП Ткач Гліб Львович, 098765432100 ❶	2013-07-08	h.tkach@mail.com
*юридична особа у стані припинення ❺		

Помилки: ❶ декілька типів даних в одній колонці, ❷ нумерація колонок, ❸ об’єднані комірки, ❹ заголовки, ❺ примітки.

Як некоректно заповнити таблицю?

Ідентифікатор	Найменування	Дата реєстрації	Електронна пошта
6 12345678	ТОВ “АБВ”	2010-02-03	sales@abv.com, 044-321-45-67 1
78923389	6 ГО «Актив»	2011-04-05 1	contact@active.org
23456789 3	КП “Поліклініка №1”	07.06.2012 1	mail[at]clinic1.com
9.8E+10 2	ФОП Ткач Г.Л.	08/07/2013 1	h.tkach@mail.com
3I56Ш89o 4	ЛМР	– 5	null 5

Помилки: ❶ різні формати даних в одній колонці, ❷ неправильно визначені формати даних в колонці, ❸ кодування інформації форматуванням, ❹ одруківки, помилки, ❺ нестандартизований запис у випадку відсутності даних, ❻ пробіли перед записом.

Описова статистика

Що таке змінна

Змінна — будь-яка характеристика об’єкта, що вимірюється чи досліджується. Змінна може набувати різних значень для різних об’єктів.

Типи змінних

Тип	Опис	Операції
Категорійні (номінальні)	Позначають унікальну категорію, до якої належить або не належить спостереження. Наприклад, стать, країна, форми 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

Підрахунок частот

Категорія	Кількість готелів
Без зірки	1
1 зірка	2
2 зірки	3
3 зірки	4
4 зірки	3

Відносні частоти

Категорія	Кількість готелів (частота)	Відносна частота
Без зірки	1	0,06252
1 зірка	2	0,125
2 зірки	3	0,1875
3 зірки	4	0,25
4 зірки	3	0,1875
5 зірок	3	0,1875

Аналітичні методи

01 Розуміння повноти даних

- Кількість
- Сума
- Відсоток
- Частота
- Порожні або відсутні значення

02 Розуміння центру та розподілу даних

- Середнє арифметичне
- Медіана
- Мода

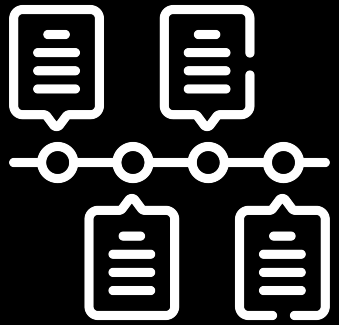
03 Розуміння меж та варіативності даних

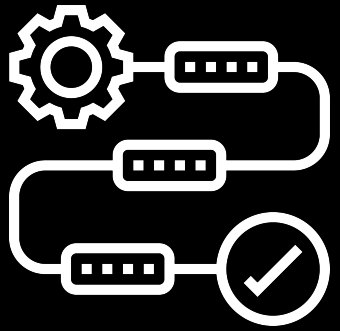
- Мінімальне/максимальне
- Дисперсія/нормальний розподіл
- Стандартне відхилення

04 Розуміння позиції в датасеті

- Процентилі
- Квантилі

Вплив факторів на розуміння даних при аналізі





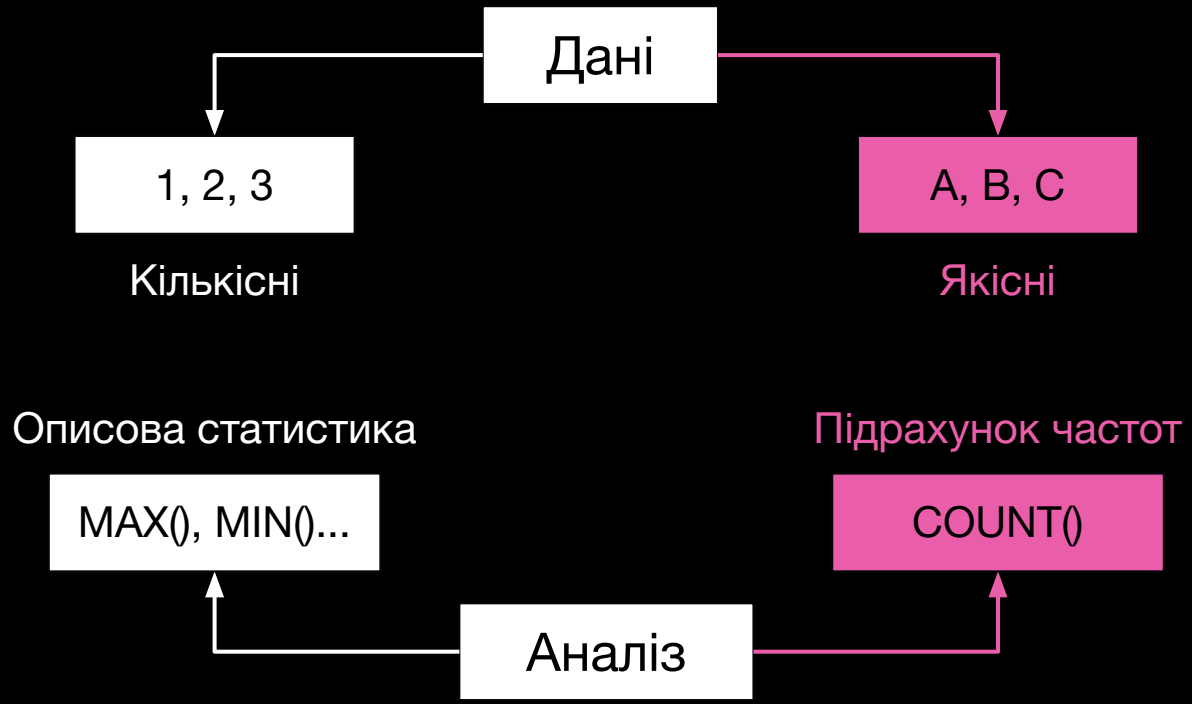
Хронологічний контекст

Залежно від проміжку часу, який ви берете для аналізу, ви отримуватимете різні результати.

Методологія збору інформації

Дані прямо залежать від методології збору інформації.

Види даних та підходи до їх аналізу

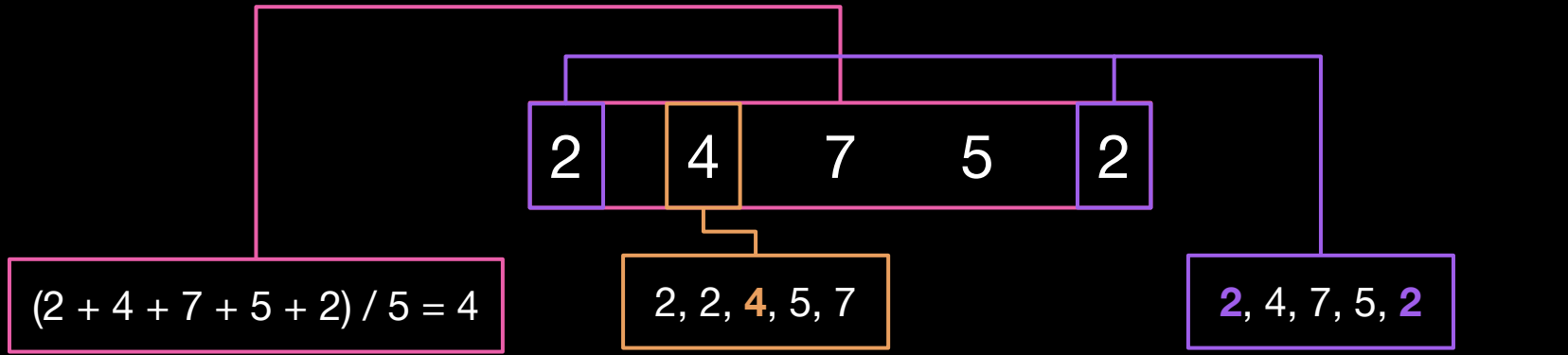


Корисні команди для якісних даних

Назва	Формула в Microsoft Excel (англійською)	Формула в Microsoft Excel (російською)	Опис
Частота	FREQUENCY()	ЧАСТОТА()	Частота знаходження певних значень у діапазоні
Підрахунок частот	COUNTA()	СЧЕТЗ()	Підрахунок кількості значень
Підрахунок за умовою	COUNTIF	СЧЕТЕСЛИ()	Підрахунок кількості значень, які відповідають умові
Унікальні значення	UNIQUE()	УНИК()	Пошук унікальних значень із діапазону

Описова статистика

Які показники є типовими або узагальнюючими для всієї сукупності?



AVERAGE() або СРЗНАЧ()
середнє арифметичне

MEDIAN() або МЕДИАНА()
медіана

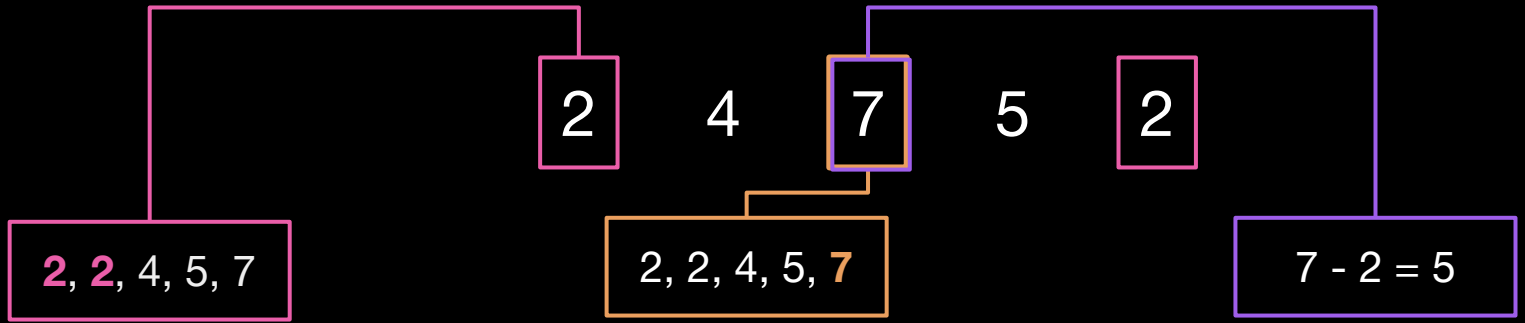
MODE() або МОДА()
мода

Корисні команди для якісних даних

Назва	Формула в Microsoft Excel (англійською)	Формула в Microsoft Excel (російською)	Опис
Середнє арифметичне	AVERAGE()	СРЗНАЧ()	Сума всіх показників сукупності, що поділена на їх кількість.
Медіана	MEDIAN()	МЕДИАНА()	Число, що ділить упорядкований ряд значень (від найбільшого до найменшого або навпаки) навпіл.
Мода	MODE()	МОДА()	Значення, що найчастіше зустрічається в сукупності.

Описова статистика

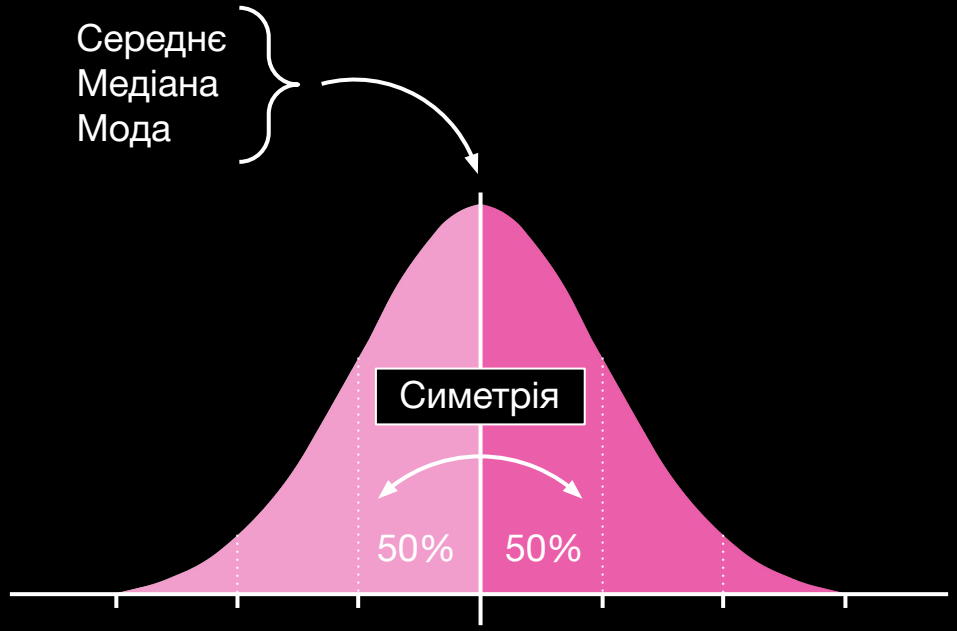
Які відмінності наявні в сукупності показників? В яких межах вони знаходяться і на скільки відрізняються одне від одного?



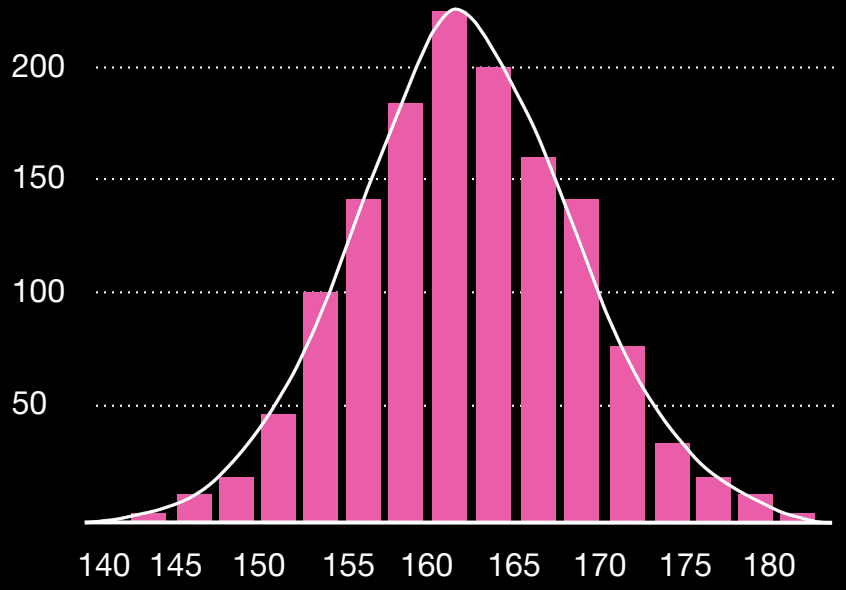
Корисні команди

Назва	Формула в Microsoft Excel (англійською)	Формула в Microsoft Excel (російською)	Опис
Мінімальне значення	MIN()	МИН()	Найменше значення, що наявне в сукупності показників
Максимальне значення	MAX()	МАКС()	Найбільше значення, що наявне в сукупності показників
Розмах	MAX()-MIN()	МАКС()-МИН()	Різниця між найбільшим та найменшим значенням в сукупності показників

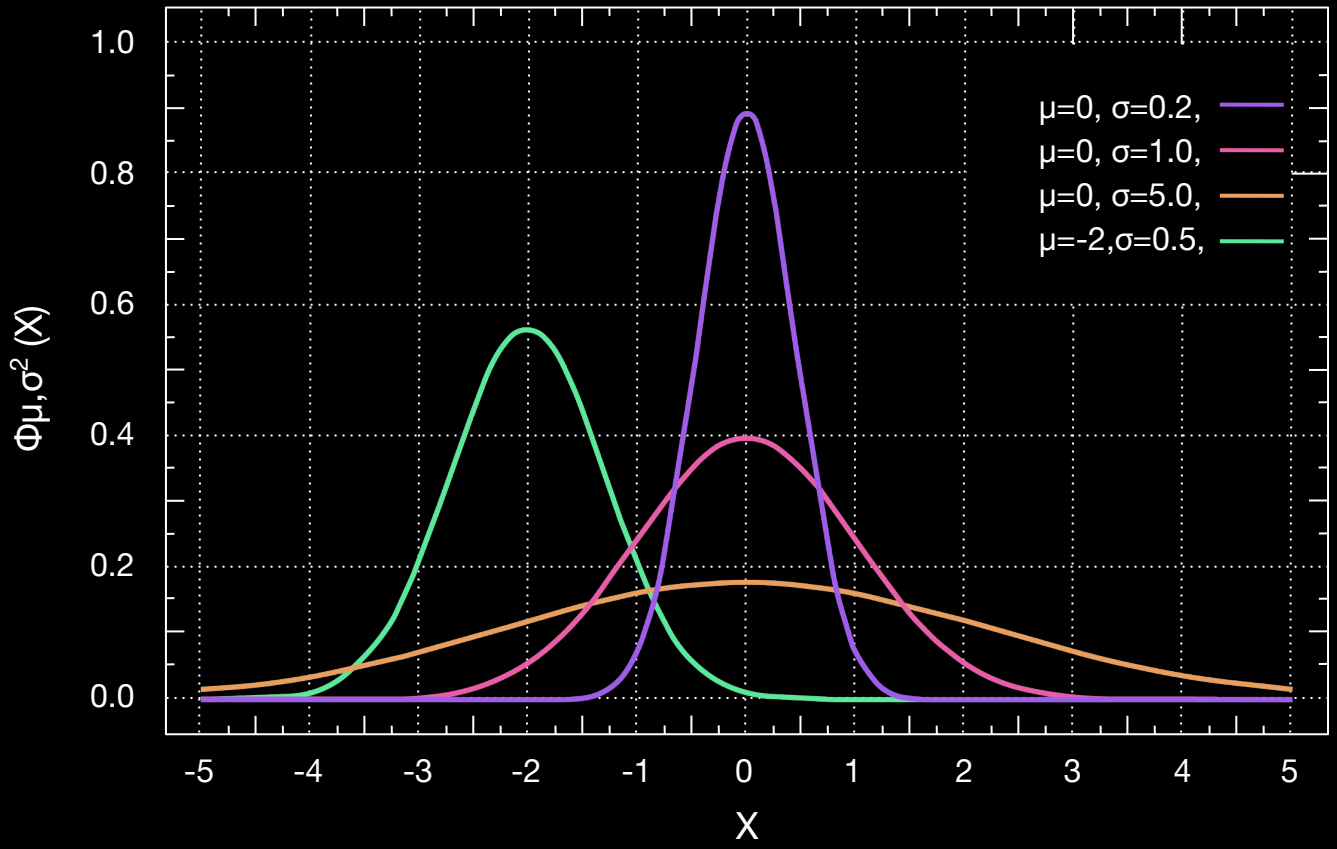
Приклад нормального розподілу



Приклад нормального розподілу



Приклад нормального розподілу



симетричний відносно точки X яка одночасно є модою, медіаною і середнім значенням розподілу.

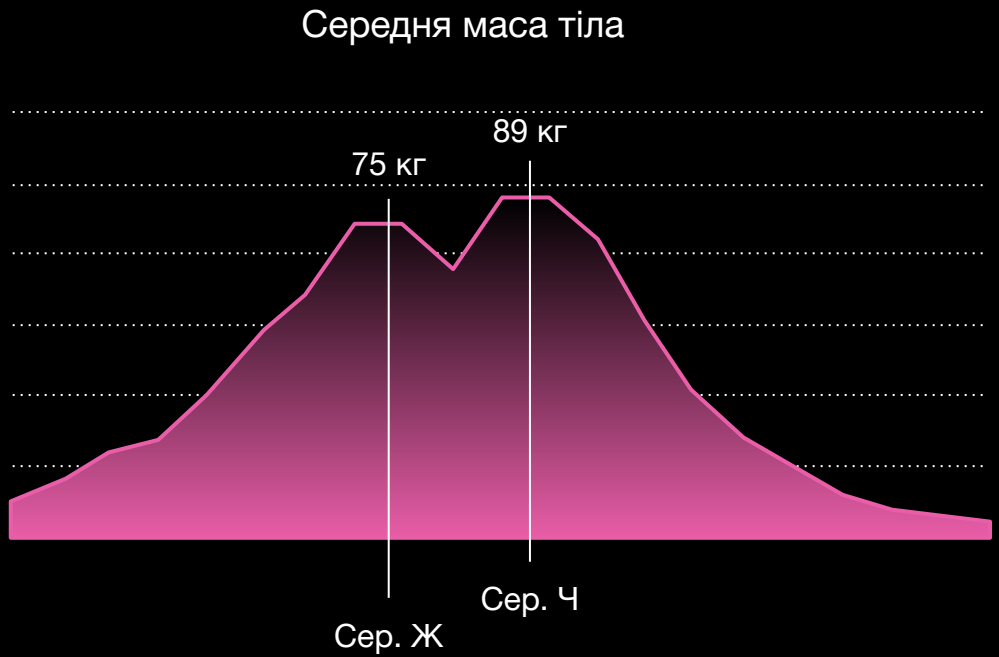
Чому модель нормального розподілу є корисною:

- Багато речей у світі є «нормально розподіленими», або ж дуже близькими до нормального розподілу. Окрім зросту, про який ми говорили, похибки вимірювання також мають нормальний розподіл.

- З нормальним розподілом легко працювати математичними засобами. У багатьох практичних випадках, методи, розроблені з використанням теорії нормального розподілу, працюють досить добре, якщо розподіл і не є нормальним.

- Застосування нормального розподілу дозволяє виявляти різні аномалії і в суспільному житті – наприклад фальсифікації на виборах

Бімодальний розподіл



API

Відкриті доступи до державних систем

Відкриті доступи до державних систем

Openprocurement API

https://prozorro.gov.ua/ *документація*

Spending API

https://spending.gov.ua/ *документація*

Агентство з розвитку інфраструктури фондового ринку України https://smida.gov.ua/

Документація

Портал відкритих даних Верховної Ради України

Документація

OPENDATA. СТАТИСТИЧНІ ДАНІ НМТ/ОСНОВНОЇ СЕСІЇ ЗНО

Документація

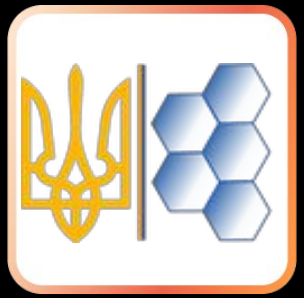
Відкриті доступи до державних систем

Реєстр декларацій

Документація

Декларації політичних партій

Документація



ЄДЕБО

Документація

Сервіси НБУ

Документація

Реєстр корупціонерів

Документація

Відкриті доступи до державних систем

Prozorro.Продажі

Документація

УкрНОІВІ

Документація

DREAM

API

Документація

OpenBudget

API

Інструкція користувача

Відкриті доступи до державних систем

Prozorro.Продажі

Документація

УкрНОІВІ

Документація

DREAM

API

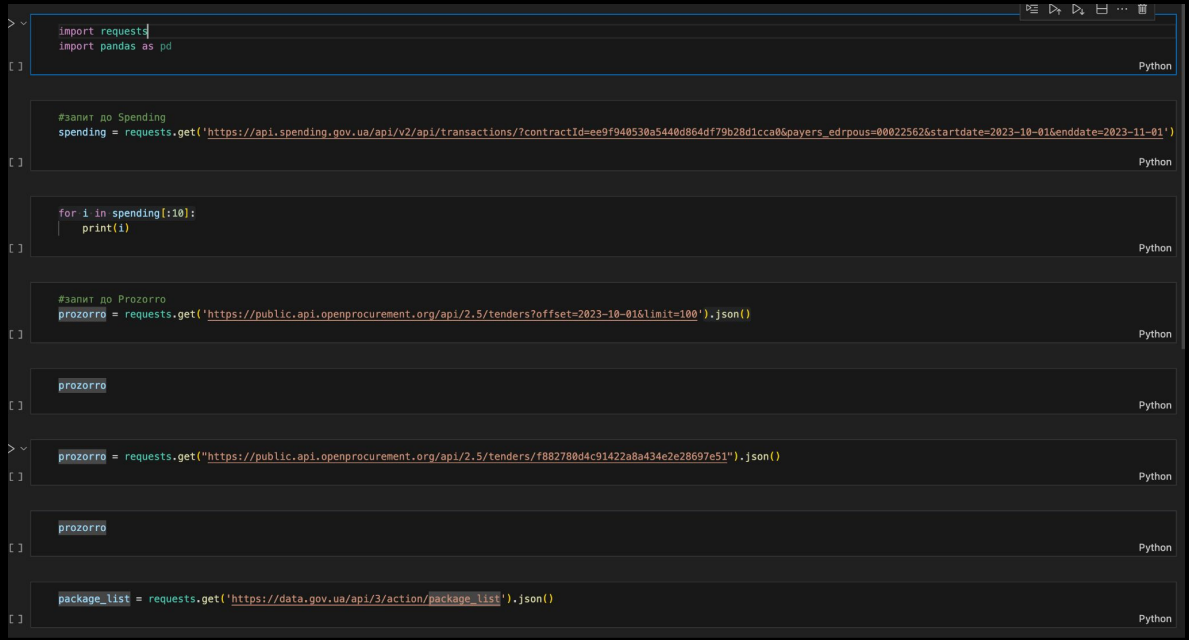
Документація

OpenBudget

API

Інструкція користувача

Код



```
import requests
import pandas as pd
```

Python

```
#запит до Spending
spending = requests.get('https://api.spending.gov.ua/api/v2/api/transactions/?contractId=ee9f940530a5440d864df79b28d1cca0&payers_edrpous=00022562&startdate=2023-10-01&enddate=2023-11-01')
```

Python

```
for i in spending[:10]:
    print(i)
```

Python

```
#запит до Prozorro
prozorro = requests.get('https://public.api.openprocurement.org/api/2.5/tenders?offset=2023-10-01&limit=100').json()
```

Python

```
prozorro
```

Python

```
prozorro = requests.get("https://public.api.openprocurement.org/api/2.5/tenders/f882780d4c91422a8a434e2e28697e51").json()
```

Python

```
prozorro
```

Python

```
package_list = requests.get('https://data.gov.ua/api/3/action/package_list').json()
```

Python