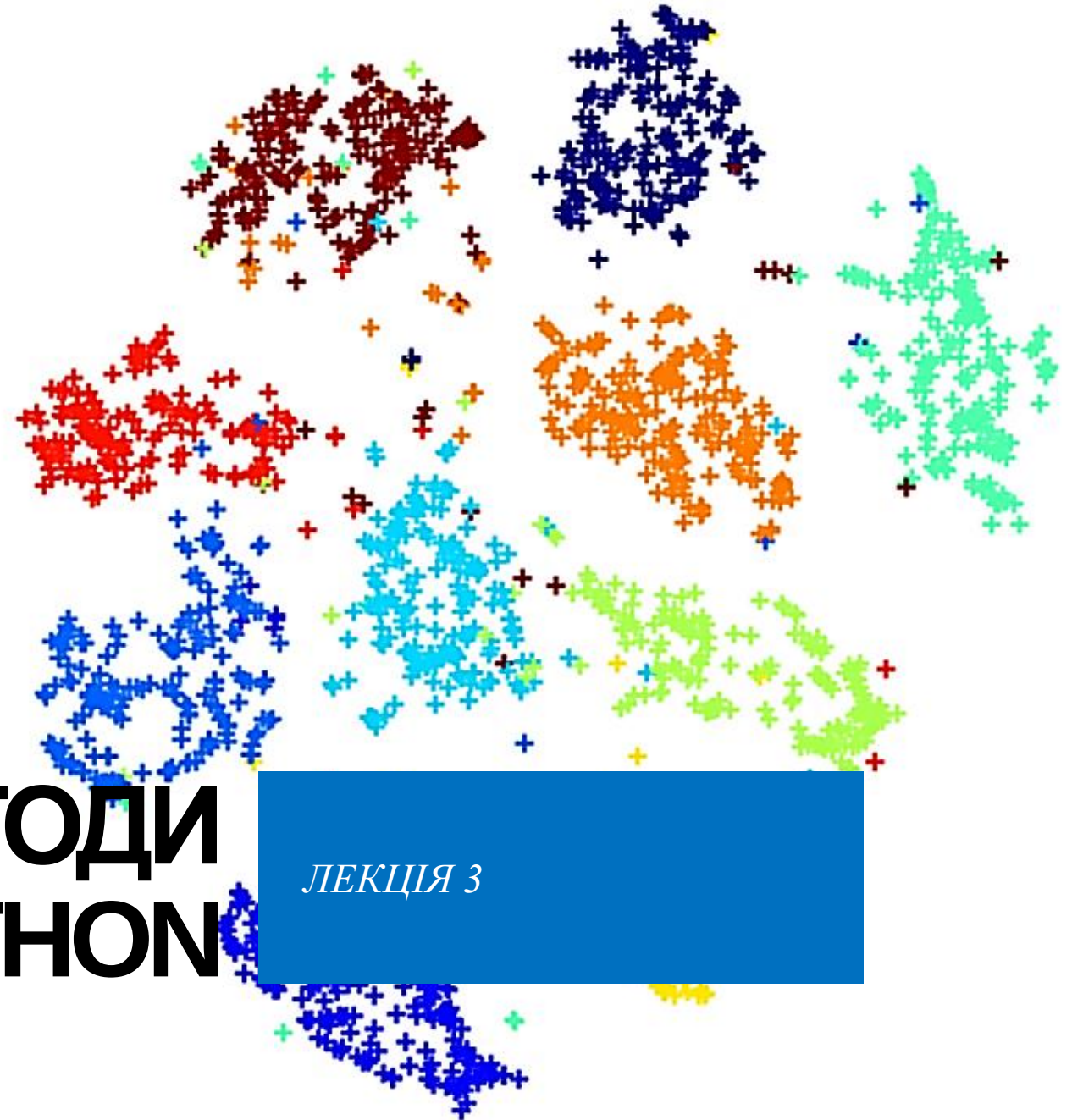




СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ В PYTHON

ЛЕКЦІЯ 3

План

1. Що таке статистика?
2. Описова статистика
3. Міри центральної тенденції
4. Нормальний розподіл
5. Дисперсія і стандартне відхилення
6. Квартилі та інтерквартильний розмах.



ЩО ТАКЕ СТАТИСТИКА?



- **Статистика** - потужний інструмент в Data Science. Вона дозволяє отримати інформацію з даних, дізнатися їх структуру і на основі отриманої інформації провести подальший аналіз.
- **Статистичні характеристики** - найбільш часто використовувана концепція в Data Science. Зазвичай це перше, що застосовують при дослідженні набору даних. У цю концепцію входять такі поняття як *відхилення, дисперсія, середнє значення, медіана, процентилі* і багато іншого.

ВИДИ СТАТИСТИКИ?

Статистика допомагає оцінити варіативність та зменшити невизначеність. Розрізняють описову та вивідну статистики.

- **Описова** - вивчає властивості спостережуваних даних.
- **Вивідна статистика** — виводимо припущення про властивості розподілу даних з яких походять спостережувані дані .

З допомогою статистики можна дати відповідь на питання:

- Чи є залежність між кількістю злочинів та фазами Місяця?
- Яка ймовірність вчасно потрапити на роботу в час пік?
- Побудувати довірчий інтервал часу, за який ви потрапляєте на роботу.
- Проводити опитування та трактувати їх результати

ПОПУЛЯЦІЯ, ВИБІРКА, СПОСТЕРЕЖЕННЯ

Генеральна сукупність - це повна сукупність об'єктів дослідження, *вибірка* - її частина, яка сформована певним науково обґрунтованим способом

Термін "генеральна сукупність" використовується тоді, коли йдеться про велику, але кінцеву сукупність досліджуваних об'єктів. Наприклад, про сукупність абітурієнтів України у 2022 році або сукупність дітей дошкільного віку міста Житомир.



Матриця даних – стартовий елемент для аналізу даних. Зазвичай йому передуює етап збору, очищення та представлення у табличному вигляді.

По рядках – респонденти, суб'єкти, учасники, спостереження
По стовпцях - характеристики кожного запису(змінні).

Для узагальнення категоріальних даних використовують **частотні таблиці**.

ПРИКЛАД. МАТРИЦІ ДАНИХ

Ця таблиця включає в себе 4 стовпця, однак зібрані дані мають майже 800 спостережень (спостереження зібрані з ресурсу <https://dom.ria.com/> і містять інформацію про квартири, які продаються).

Для того, щоб описати вміст цієї таблиці в більш зрозумілій формі використовують узагальнення та опис типових чи середніх значень.

Місто	Кімнат	Загальна площа	Ціна
Вінниця	3	120	1875000
Вінниця	3	66	975000
Вінниця	2	66	1375000
Вінниця	2	44	637500
Вінниця	3	63	853000
Вінниця	1	31	562500

ПРИКЛАД . ЧАСТОТНІ ТАБЛИЦІ

Для дослідження опитали 25 студентів: наскільки їх провал на екзамені залежав від причин, які вони ніяк не могли контролювати. Відповіді даються за шкалою від 1 до 7 (1 - зовсім не залежав, 7 - повністю залежав).

Дані опитування: 3,5,6,5,2,3,6,4,6,7,6,4,5,5,1,2,5,4,4,5,5,7,3,3,4

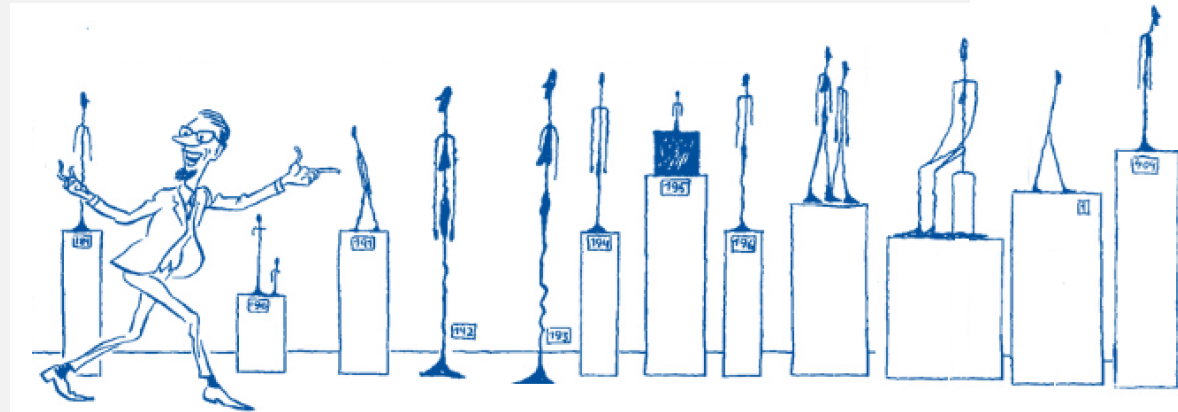
*Впорядковуємо вибірку у напрямку зростання і одержуємо **варіаційний ряд**: 1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,6,6,6,6,7,7*

З даного ряду можна побудувати частотну таблицю:

Варіант	1	2	3	4	5	6	7
Частота	1	2	4	5	7	4	2
Відносна частота, %	1/25 4%	2/25 8%	4/25 16%	1/5 20%	7/25 28%	4/25 16%	2/25 8%

ОПИСОВА СТАТИСТИКА

Описова статистика або **дескриптивна статистика** (англ. *descriptive statistics*) — розділ статистики, який займається обробкою емпіричних даних, їх систематизацією, наочним представленням у вигляді графіків та таблиць, а також їх кількісним описом через основні статистичні показники.



МІРИ ЦЕНТРАЛЬНОЇ ТЕНДЕНЦІЇ

Залежно від даних, в якості центральної тенденції використовується *середнє значення, медіана або мода*

Міри центральної тенденції потрібні з наступних міркувань:

Щоб отримати загальну картину розподілу. Ми не можемо запам'ятати кожен факт, що стосується сфери дослідження.

Щоб отримати чітку картину щодо досліджуваної сфери для розуміння та отримання потрібних висновків.

Щоб отримати чіткий опис групи в цілому та мати змогу порівнювати дві або більше груп у термінах типової «поведінки».

СЕРЕДНЄ (Mean)

Середнє значення підходить для узагальнення кількісних даних(як дискретних, так і неперервних). Формула обрахування:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Наприклад, якщо є група з 5 учнів, оцінки яких 12, 3, 5, 10, 5.

Середнє значення оцінок = 7.

Однак із використанням середнього значення в якості опису центральної тенденції є невелика проблема. Якщо є нетипово великі чи малі для даного набору значення – вони роблять великий внесок у значення середнього.

Наприклад, є певне невелике підприємство, яке має 5 працівників.

Заробітні плати працівників в гривнях: 5000, 7000, 2000, 4000, 50 000. Середнє значення заробітної плати 13600 грн. Однак, якщо відкинути екстремальне значення 50 000, то середнє значення буде = 4500.

СЕРЕДНЄ (Mean)

Середнє значення більш стабільне, ніж медіана чи мода. Тому, коли потрібно знайти найбільш стабільну міру центральної тенденції, використовують середнє.

Переваги середнього:

- Середнє визначене дуже жорстко.
- Це найбільш поширена міра центральної тенденції, оскільки її легко зрозуміти.
- Середнє легко підрахувати.
- Враховує всі значення розподілу.

Обмеження чи недоліки середнього:

- На значення середнього впливають екстремальні значення (жарт про «середню температуру по лікарні»).
- Часом середнім є значення, що не присутнє в розподілі.
- Часом результатом можуть бути абсурдні значення.

МЕДІАНА (Median)

Медіана – це значення, яке ділить вибірку навпіл, тобто 50% є меншими за це значення, 50% більшими.

Основна перевага використання медіани - менша чутливість до екстремальних значень.

Приклад із заробітною платою: 2000, 4000, **5000**, 7000, 50 000.

Медіаною є число, що стоїть посередині - 5000.

Коли значень парна кількість, то беруть два центральні значення, і знаходять їхнє середнє.

Приклад із заробітною платою: 2000, 4000, **5000**, **7000**, 10000, 50 000.

Медіаною є середнє значення – $(5000+7000)/2= 6000$.

МЕДІАНА (Median)

Обмеження медіани:

- Не враховує всі спостереження.
- З медіаною не можна робити алгебраїчні перетворення так, як із середнім.
- Потребує впорядкування значень або класів інтервалів у висхідному чи спадному порядку.
- Часом медіаною може бути значення, не присутнє у самому розподілі.

Для чого використовують медіану?

- Коли потрібно знайти точну середню точку.
- Коли екстремальні значення впливають на середнє — медіана є найкращою мірою центральної тенденції.

МОДА (Mode)

Мода використовується для визначення центральної тенденції категоріальних або кількісних дискретних даних.

Мода - це значення, яке найчастіше трапляється.

Наприклад, за інформацією міністерства юстиції України, у 2014 року хлопчиків найчастіше називали Дмитром, Артемом, Максимом та Іваном, дівчаток – Анею, Анастасією, Софією та Дар'єю.

Найпопулярніші імена в Україні у 2025 році (перші півріччя):
Хлопчики: *Олександр, Марк, Тимофій, Матвій, Максим, Артем, Богдан, Дамір, Данило, Дмитро, Макар.*
Дівчатка: *Софія, Емілія, Єва, Мілана, Соломія, Вікторія, Поліна, Анна, Марія, Вероніка*

МОДА (Mode)

Переваги моди:

- Мода показує найбільш поширене значення в розподілі.
- На моду не впливають екстремальні значення – так як на середнє.
- Моду можна визначити для відкритих інтервалів / категорій.
- Моду можна виявити просто побудувавши графік розподілу чи стовпчасту діаграму.

Обмеження:

Не включає до визначення / розрахунку всі спостереження розподілу, а лише концентрацію частот.

Подальші алгебраїчні перетворення неможливі – на відміну від середнього.

Буває важко визначити моду у випадку багатомодального чи бімодального розподілу

ПРИКЛАД. МІРИ ЦЕНТРАЛЬНОЇ ТЕНДЕНЦІЇ

Знайдіть міри центральної тенденції вибірки:

6, 6, 8, 10, 11, 13, 14, 14, 15, 23.

M - мода вибірки, m - медіана, S - середнє значення вибірки.

$M = \{6, 14\}$, бо значення 6 і 14 трапляються найчастіше

$m = 12$, бо воно ділить вибірку навпіл

$S = (6*2+8+10+11+13+14*2+15+23)/10 = 120 / 10 = 12$

Дано вибірку: 4, 5, 5, 6, 7, 7, 8, 10, 11.

Знайдіть міри центральної тенденції цієї вибірки

НОРМАЛЬНИЙ РОЗПОДІЛ (Normal distribution)

Розподіл у якому всі три міри центральної тенденції збігаються

$$\text{середнє} = \text{медіана} = \text{моді}$$

називається **нормальним**.

Його також називають розподілом Гауса або «дзвоноподібним» (bell-shaped curve) – адже графік нормального розподілу подібний на форму дзвона у профіль.

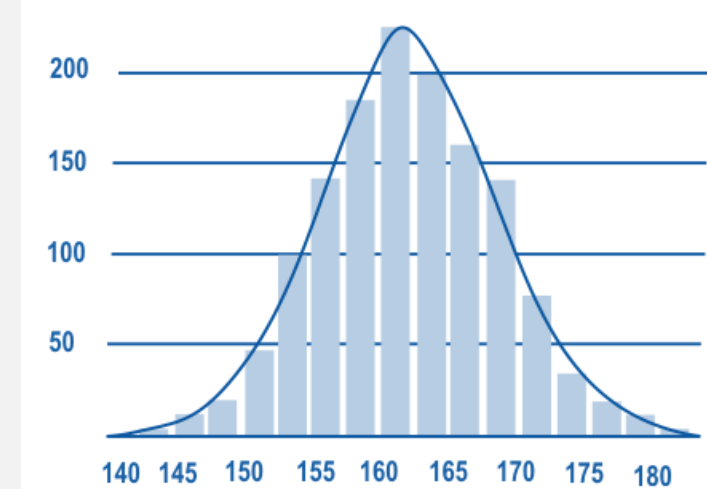


Застосування нормального розподілу дозволяє виявляти різні аномалії і в суспільному житті – наприклад фальсифікації на виборах

(news.liga.net/articles/politics/767097-gauss_protiv_falsifikatsiy_anomalii_na_vyborakh_2012.htm)

ПРИКЛАД. НОРМАЛЬНИЙ РОЗПОДІЛ

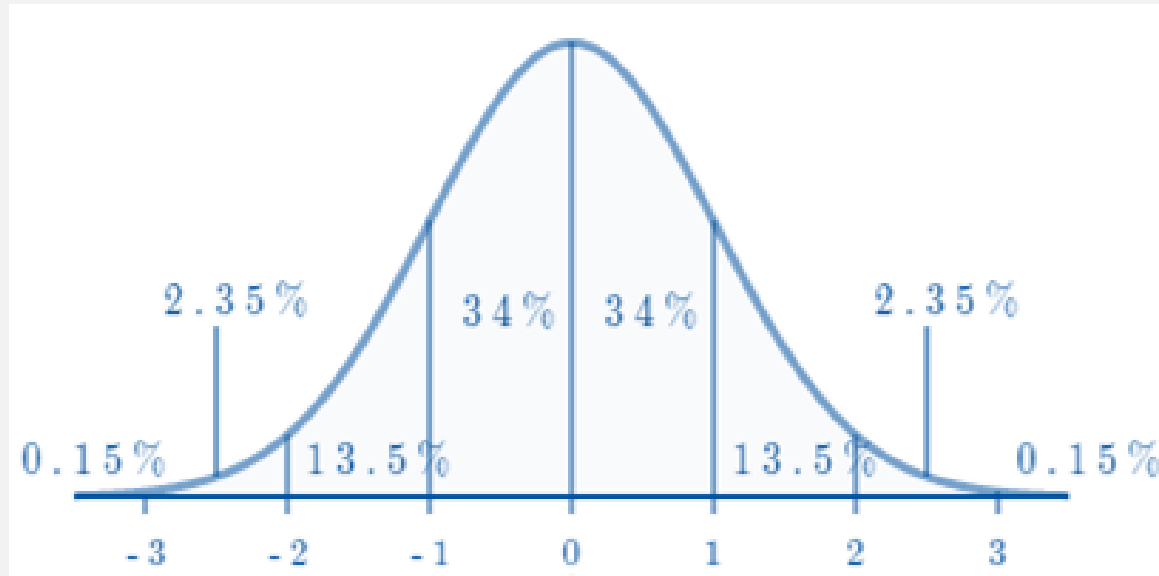
Для прикладу, візьмемо розподіл зросту жінок певного віку в певній країні. Найбільш популярними будуть значення, що відповідають «середньому зросту», а екстремальних значень – дуже низький зріст і дуже високий – буде дуже мало.



Зріст, артеріальний тиск, помилка вимірювання і показники IQ підкоряються нормальному розподілу.

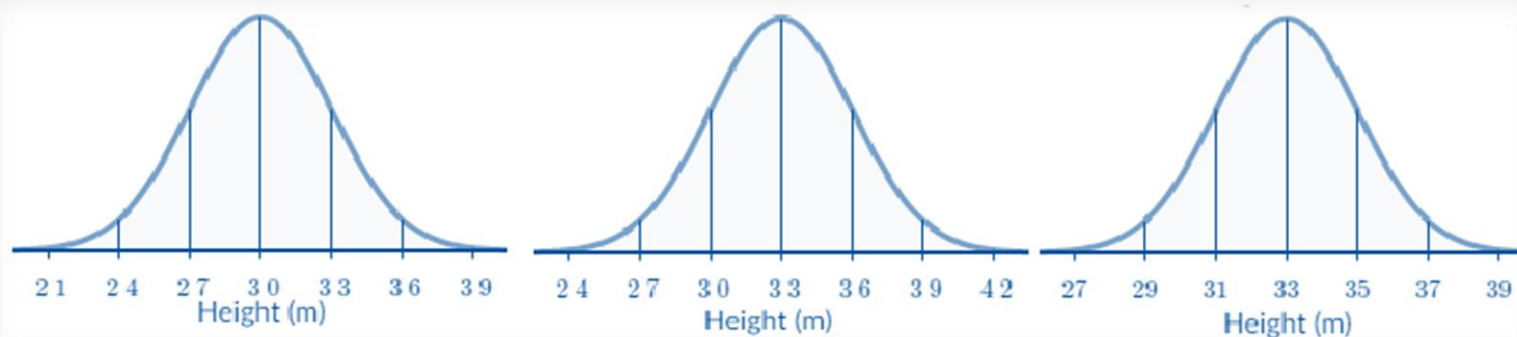
ОСОБЛИВОСТІ НОРМАЛЬНОГО РОЗПОДІЛУ:

- симетрична форма дзвону;
- середнє і медіана рівні (розташовані в центрі);
- $\approx 68\%$ даних потрапляє в 1 стандартне відхилення середнього;
- $\approx 95\%$ даних потрапляє в 2 стандартні відхилення середнього;
- $\approx 99.7\%$ даних потрапляє в 3 стандартні відхилення середнього.



ПРИКЛАД. ЯКИЙ З НАВЕДЕНИХ НИЖЧЕ НОРМАЛЬНИХ РОЗПОДІЛІВ НАЙКРАЩЕ УЗАГАЛЬНЮЄ ДАНІ?

Висота одного з різновидів сосни також нормально розподілена.
Середня висота = 33 м, і стандартне відхилення = 3.



ДИСПЕРСІЯ І СТАНДАРТНЕ ВІДХИЛЕННЯ

Середньоквадратичне відхилення (**standard deviation**) дає розуміння, наскільки далеко знаходиться типове спостереження від середнього значення.

Дисперсія (variance) - обчислюється як середнє значення відстаней від всіх спостережень до середнього значення в квадраті.

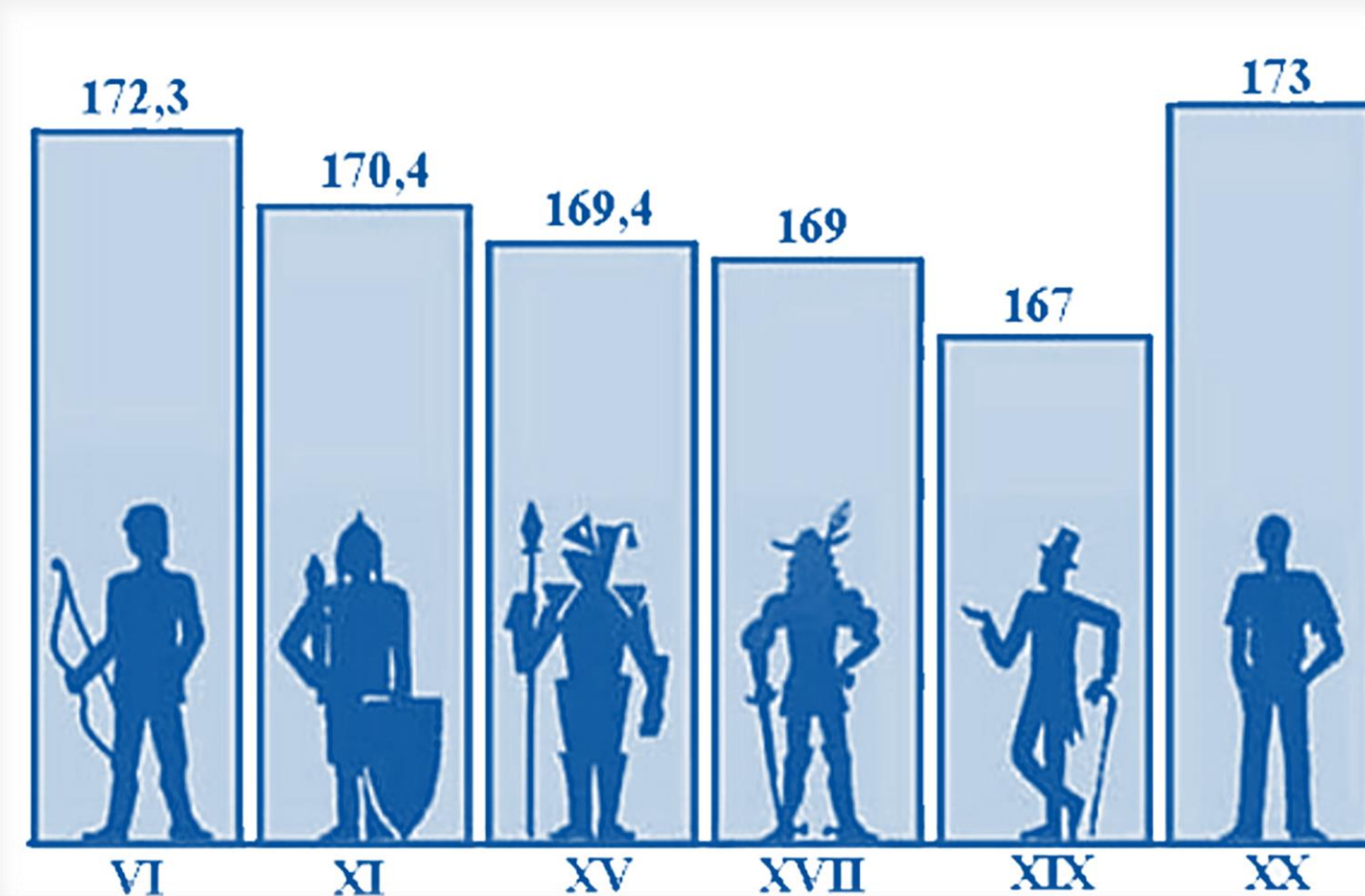
$$var = \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2$$

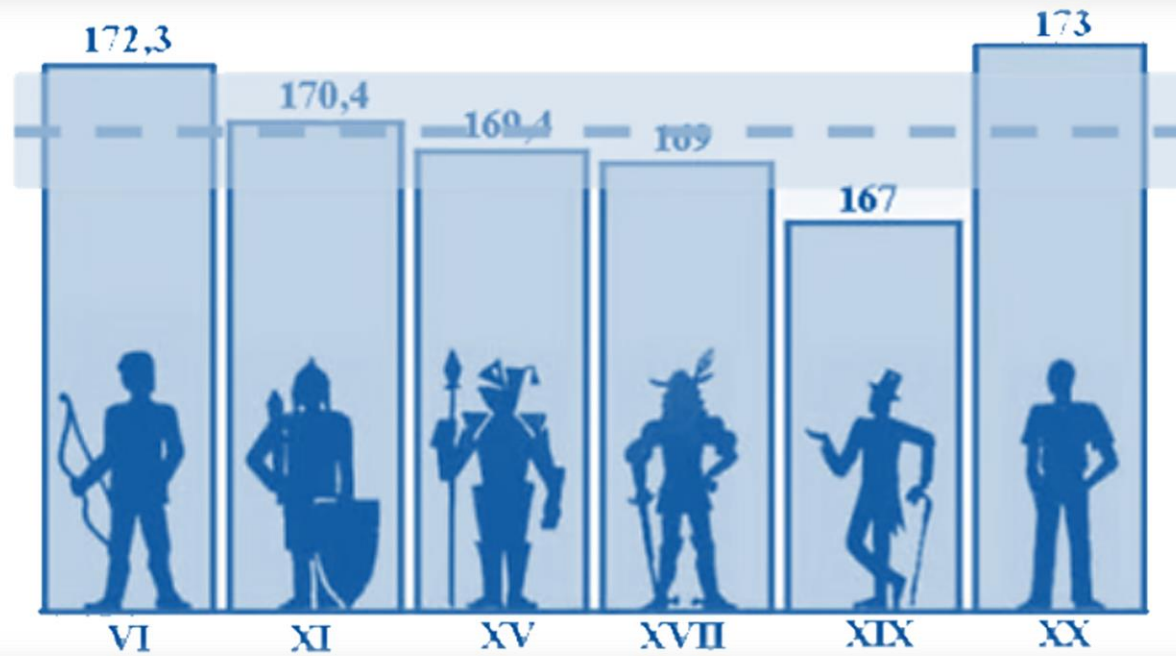
Середньоквадратичне відхилення (standard deviation) σ - обчислюється як корінь квадратний з дисперсії.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{N}}$$

ПРИКЛАД. СЕРЕДНЬОКВАДРАТИЧНЕ ВІДХИЛЕННЯ

Люди в різні століття мали різний зріст.





$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = 170.18 \text{ cm}$$

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}} = 2.03$$

ПРИКЛАД. ОБРАХУВАТИ ДИСПЕРСІЮ ТА СЕРЕДНЬОКВАДРАТИЧНЕ ВІДХИЛЕННЯ.

Нехай маємо ряд 5, 3, 2, 8, 2

Середнє значення =4.

Обчислимо дисперсію та середньоквадратичне відхилення:

Дисперсія:

$$\begin{aligned}\text{var} &= ((5-4)^2 + (3-4)^2 + (2-4)^2 + (8-4)^2 + (2-4)^2) / 5 = \\ &= (1^2 + (-1)^2 + (-2)^2 + 4^2 + (-2)^2) / 5 = (1 + 1 + 4 + 16 + 4) / 5 = 5.2\end{aligned}$$

Середньоквадратичне відхилення: $\sigma = 2.28$

Для ряду 3, 5, 5, 3, 4, середнє значення = 4, $\text{var} = 1$, $\sigma = 1$

Якщо поглянути на обидва ряди, бачимо що значення другого більш тісно розташовані навколо середнього. Значення дисперсії та середньоквадратичного відхилення дозволяють це виразити чисельно.

КВАРТИЛІ ТА ІНТЕРКВАРТИЛЬНИЙ РОЗМАХ

Якщо медіана ділить дані порівну, то квартилі ділять їх на чотири частини. Вони позначаються Q1, Q2, Q3, Q4.

- Q1 - 25%
- Q2 - 50% (співпадає з медіаною)
- Q3 - 75%
- Q4 - 100%

Інтерквартильний розмах(IQR) = Q3 - Q1

Нехай маємо ряд 62, 81, 63, 77, 64, 81, 64, 70, 72, 76.

Спочатку відсортуємо дані в зростаючому порядку:

62, 63, 64, 64, 70, 72, 76, 77, 81, 81.

Медіана(та Q2)): $(70+72)/2 = 71$

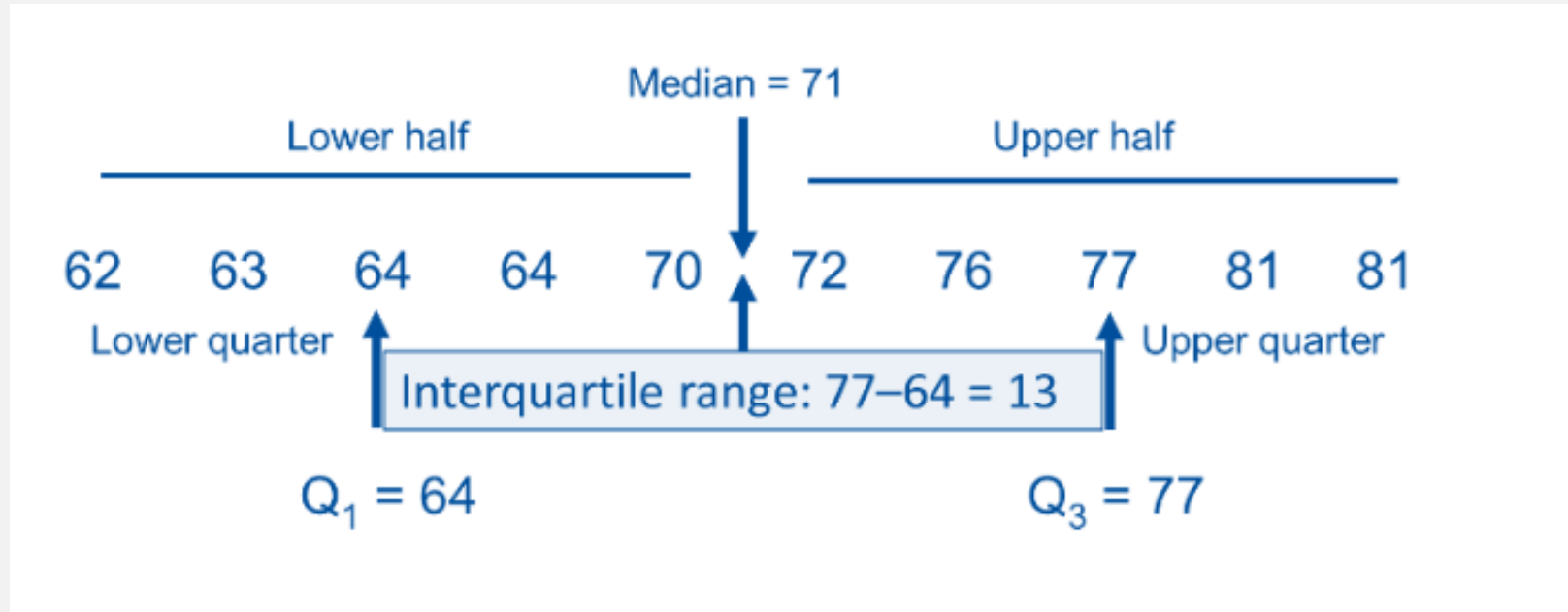
Для обрахунку Q1 та Q3 значення медіани включаються до інтервалу.

$$Q1 = 64$$

$$Q3 = 77$$

Інтерквартильний розмах: $Q3 - Q1 = 77 - 64 = 13$

КВАРТИЛІ ТА ІНТЕРКВАРТИЛЬНИЙ РОЗМА



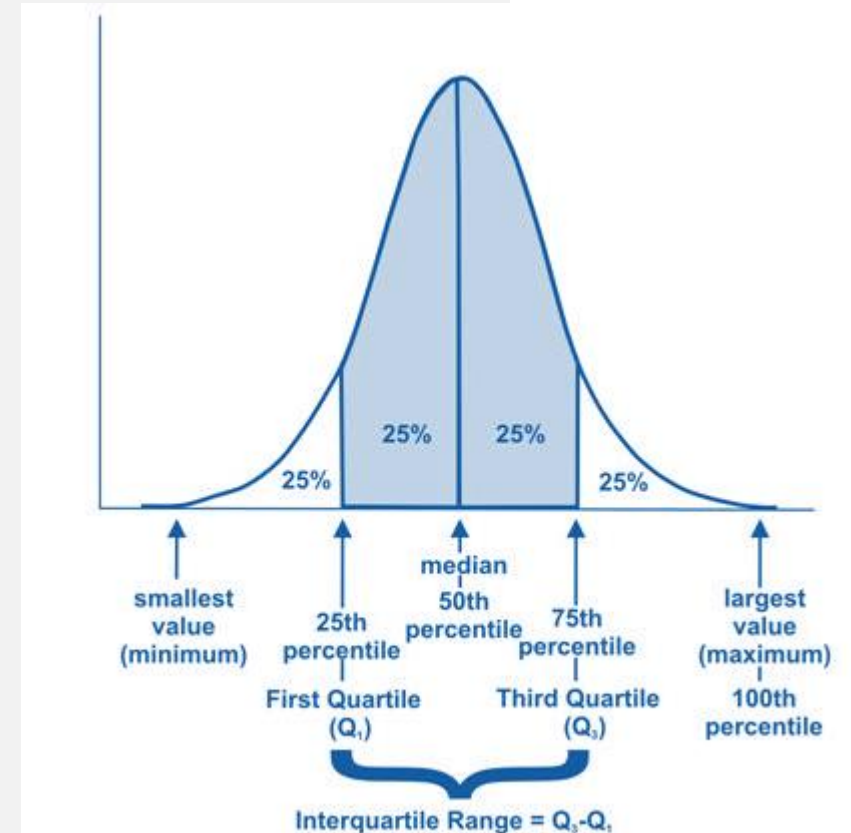
Квартилі – це корисний інструмент виміру розкиду, оскільки вони набагато менше залежні від екстремальних значень чи перекосу в наборі даних, аніж середнє та стандартне відхилення. Тому інформацію про квартилі часто подають разом із медіаною (як міри розкиду та центральної тенденції) – у випадку роботи з масивами даних, де є суттєві перекоси або екстремальні значення.

ПІДСУМОК 3 П'ЯТИ ЗНАЧЕНЬ

Середнє, мода та медіана є описовими статистиками. Також виділяють узагальнення п'яти чисел (five numbers summary). Воно включає в себе:

- Найменше значення
- Перший (нижній) кuartиль
- Медіана (другий кuartиль)
- Третій (верхній) кuartиль
- Найбільше значення

П'ять показників дають інформацію про центральну тенденцію (медіана), розмах (кuartилі) і діапазон (мінімальне та максимальне значення).



Метод `info()` виводить інформацію про `DataFrame`.

`df.info()`

flats.csv X

1 to 10 of 839 entries Filter

Місто	Кімнат	Загальна_площа	Ціна
Вінниця	3	120	1875000
Вінниця	3	66	975000
Вінниця	2	66	1375000
Вінниця	2	44	637500
Вінниця	3	63	835000
Вінниця	1	31	562500
Вінниця	3	46	1150000
Вінниця	3	64	8e+05
Вінниця	1	35	424975
Вінниця	6	200	12500

Show 10 per page 1 2 10 80 84

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 839 entries, 0 to 838
Data columns (total 4 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Місто           839 non-null    object
1   Кімнат          839 non-null    int64
2   Загальна_площа  839 non-null    float64
3   Ціна            839 non-null    float64
dtypes: float64(2), int64(1), object(1)
memory usage: 26.3+ KB
```

Міри центральної тенденції

```
df.groupby('Кімнат')['Загальна_площа'].mean()
```

```
Кімнат
1      40.015471
2      60.783094
3      86.299202
4     114.978125
5     169.000000
6     200.650000
Name: Загальна_площа, dtype: float64
```

```
df.groupby('Кімнат')['Ціна'].mean()
```

```
Кімнат
1      6.471109e+05
2      9.928795e+05
3      1.387506e+06
4      1.925796e+06
5      4.094643e+06
6      1.397500e+06
Name: Ціна, dtype: float64
```

```
df.groupby('Кімнат')['Ціна'].median()
```

```
Кімнат
1      575000.0
2      800000.0
3     1125000.0
4     1512487.5
5     2875000.0
6     1163750.0
Name: Ціна, dtype: float64
```

```
df.describe()
```



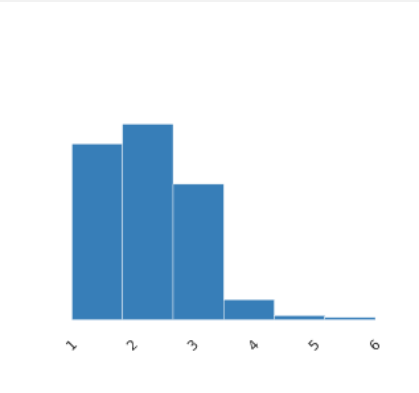
	Кімнат	Загальна_площа	Ціна
count	839.000000	839.000000	8.390000e+02
mean	2.045292	64.065924	1.042710e+06
std	0.941156	30.400174	1.057237e+06
min	1.000000	14.000000	1.020000e+04
25%	1.000000	43.750000	5.375000e+05
50%	2.000000	56.000000	7.750000e+05
75%	3.000000	75.000000	1.200000e+06
max	6.000000	222.600000	1.225000e+07



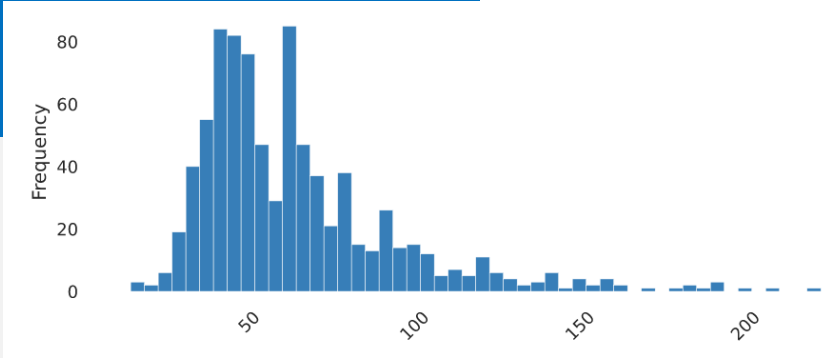
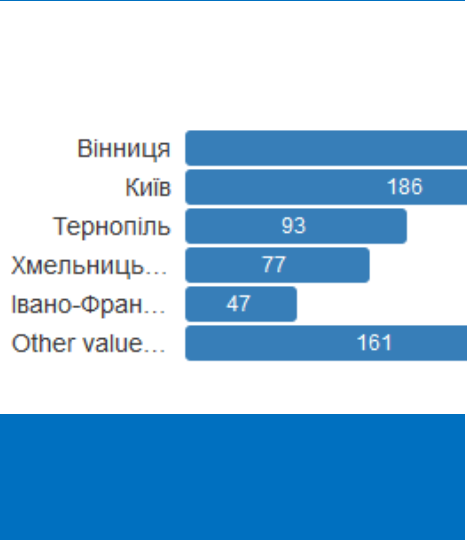
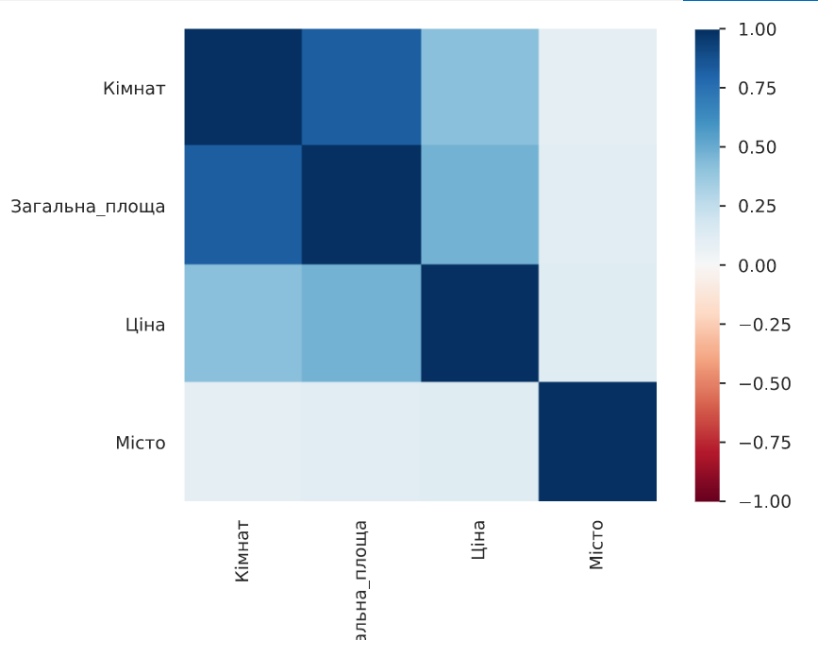
Для числових даних результат буде включати: count, mean, std, min, max і нижній 50 і верхній квартилі.

За замовчуванням нижній квартиль дорівнює 25, а верхній дорівнює 75. Квартиль 50 збігається з медіаною.

Описова статистика

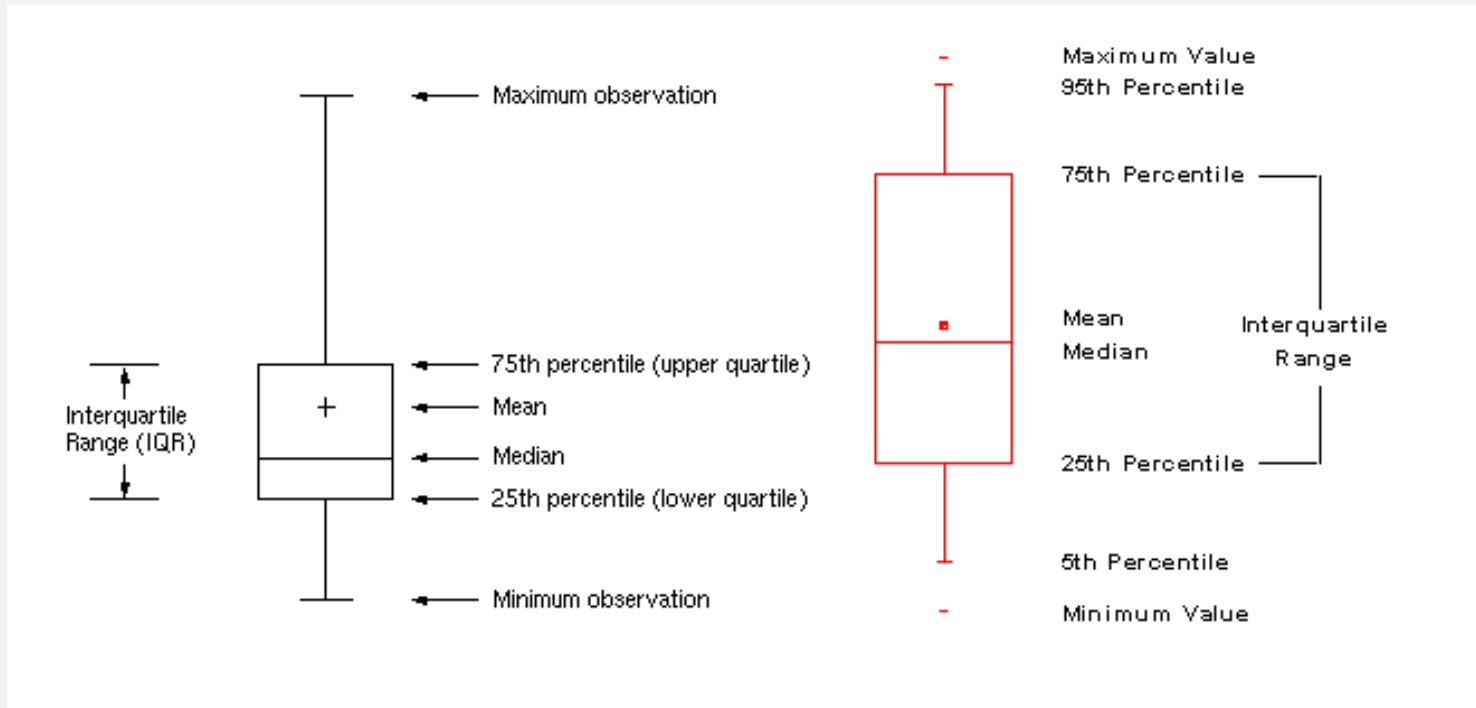


pandas_profiling.ProfileReport(df)



Count	Frequency (%)
275	32.8%
186	22.2%
93	11.1%
77	9.2%
47	5.6%
43	5.1%
23	2.7%
19	2.3%
18	2.1%
16	1.9%
42	5.0%

ВАРІАНТИ ВІЗУАЛІЗАЦІЇ «ПІДСУМКУ П'ЯТИ ЗНАЧЕНЬ»

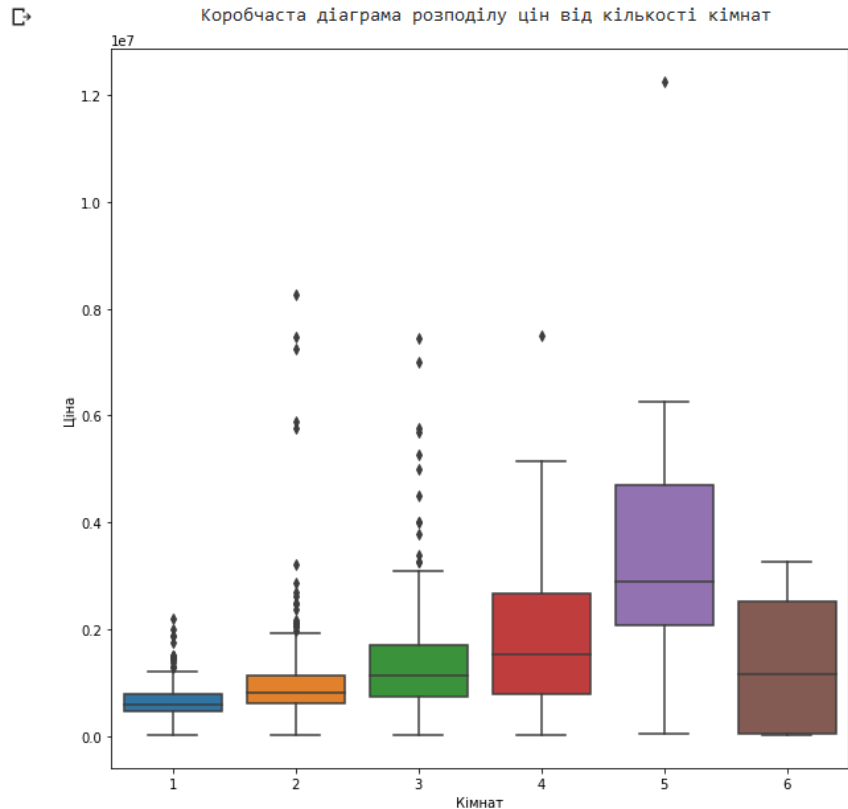


Коробчаста діаграма (або діаграма «шухляда з вусами») показує розподіл кількісних даних таким чином, щоб полегшити порівняння між змінними або між рівнями категоріальної змінної. У рамці показані квартилі набору даних, в той час як вуса розширюються, щоб показати решту розподілу, за винятком точок, які визначені як викиди з використанням методу, який є функцією міжквартильного діапазону.

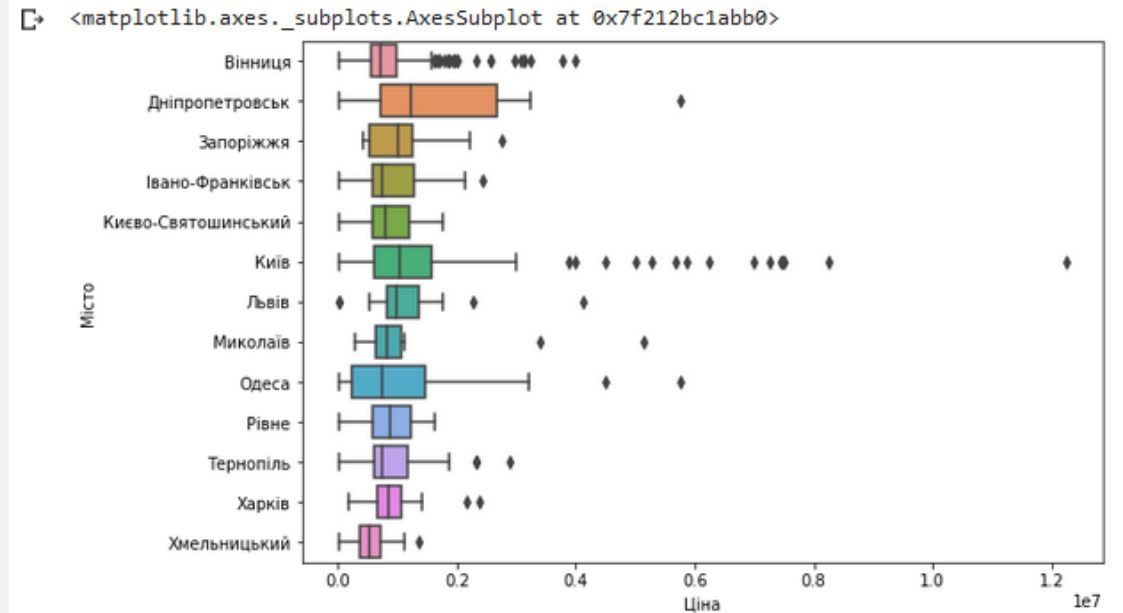
ВАРІАНТИ ВІЗУАЛІЗАЦІЇ «ПІДСУМКУ П'ЯТИ ЗНАЧЕНЬ»

seaborn.boxplot

```
figure = plt.figure(1, figsize=(10, 10))  
sns.boxplot(x='Кімнат', y='Ціна', data=df)
```



```
figure = plt.figure(1, figsize=(9, 6))  
sns.boxplot(x='Ціна', y='Місто', data=df)
```



САМОСТІЙНА РОБОТА

Розрахувати статистичні характеристики ряду (середнє, дисперсію, середнє квадратичне відхилення, моду, медіану) використовував Python модуль статистики, метод **describe()**.

Прискорений розвідувальний аналіз даних з використанням бібліотеки **pandas-profiling**

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html>