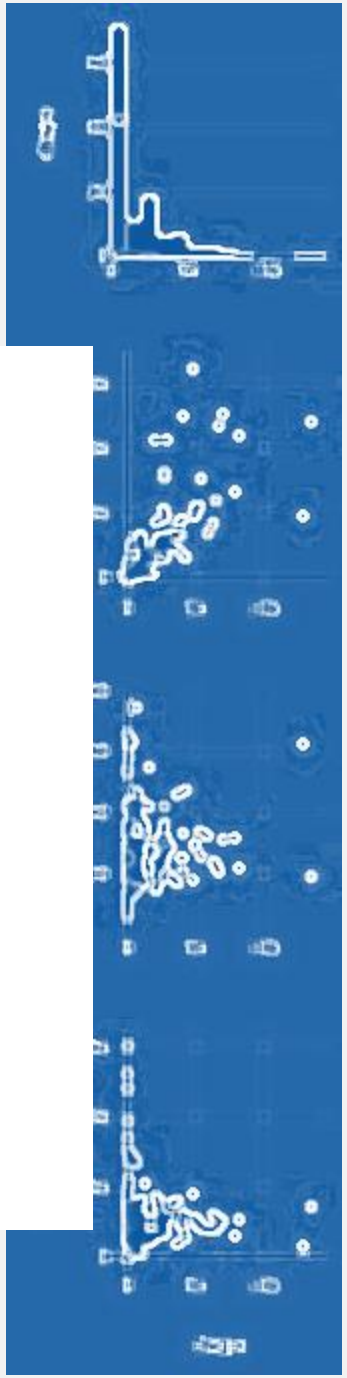
ЖИТОМИРСЬКА ПОЛІТЕХНІКА
ДЕРЖАВНИЙ УНІВЕРСИТЕТ

СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ В PYTHON

ЛЕКЦІЯ 5

План

1. Кореляційний аналіз
2. Кореляція і коваріація
3. Коефіцієнти кореляції



КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Кореляція і регресія – це статистичні виміри, які використовуються для встановлення взаємозв'язку між двома змінними.

Наприклад, припустимо, що людина водить дорогу машину, тоді передбачається, що вона має бути забезпечена у фінансовому відношенні. Для кількісної оцінки цього взаємозв'язку використовуються кореляція і регресія.

Кореляція - це статистична міра, що визначає взаємозв'язок або асоціацію двох змінних.

Регресія визначає, як незалежна змінна чисельно пов'язана із залежною змінною.

КОРЕЛЯЦІЙНИЙ АНАЛІЗ

При вивченні кореляцій намагаються встановити, чи існує якийсь зв'язок між двома показниками в одній вибірці (*наприклад, між зростом і вагою або між рівнем IQ і успішністю)* або між двома різними вибірками (наприклад, при порівнянні пар близнюків), і якщо цей зв'язок існує, то чи супроводжується збільшення одного показника зростанням (позитивна кореляція) або зменшенням (негативна кореляція) іншого.

Іншими словами, кореляційний аналіз допомагає встановити, чи можна прогнозувати можливі значення одного показника, знаючи величину іншого.

КОРЕЛЯЦІЯ І КОВАРІАЦІЯ

Коваріація - це показник, який вказує на те, наскільки дві випадкові величини змінюються в тандемі.

Кореляція - це статистичний показник, який вказує, наскільки сильно пов'язані дві змінні.

КОВАРИАЦІЯ

Ковариація - міра спільної варіативності (лінійної залежності) двох випадкових величин, іншими словами - це міра взаємодії двох випадкових змінних. Для випадкових величин X і Y ковариація обчислюється за формулою:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_I-\bar{y})}{n-1}$$

де $\bar{x}$ і $\bar{y}$– вибіркові средні X і Y;
n – число спостережень.
Середнє значення вибірок обчислюється за формулами:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

КОРЕЛЯЦІЯ

Коефіцієнт кореляції (ще має назву коефіцієнт Пірсона або лінійний коефіцієнт кореляції) між двома змінними дорівнює коваріації двох змінних, або сумі добутків відхилень, поділеній на добуток їх стандартних відхилень.

Нехай є дві вибірки $x^n = (x_1, \dots x_n), y^n = (y_1, \dots y_n)$, коефіцієнт кореляції Пірсона розраховується за формулою:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

де $\bar{x}$ і $\bar{y}$– вибіркові средні x^n *і* y^n ,
σ_x *і* σ_y - стандартне відхилення X і Y відповідно, $r_{xy} \in [-1, 1]$.

КОРЕЛЯЦІЯ І КОВАРІАЦІЯ

Основа для порівняння	Коваріація	Кореляція
Значення	Коваріація - це показник, який вказує на те, наскільки дві випадкові величини змінюються в тандемі.	Кореляція - це статистичний показник, який вказує, наскільки сильно пов'язані дві змінні.
Відношення	Кореляцію можна отримати з коваріації	Кореляція дає значення коваріації за стандартною шкалою
Значення лежать в межах	$-\infty$ і $+\infty$	-1 і +1
Зміна масштабу	Впливає на коваріацію	Не впливає на кореляцію

КОЕФІЦІЄНТИ КОРЕЛЯЦІЇ

Коефіцієнт	Коли застосовувати
Pearson	Лінійний зв’язок, нормальний розподіл
Spearman	Монотонний зв’язок, ранги
Kendall	Стійкий до викидів, малі вибірки

КОРЕЛЯЦІЯ

Кореляція — це статистична міра залежності між двома змінними, яка показує напрямок і силу зв’язку.

!!! кореляція ≠ причинність

шкала Чеддока

r_{xy}	тіснота зв'язку
0,1-0,3	слабка
0,3-0,5	помірна
0,5-0,7	помітна
0,7-0,9	висока
0,9-1,0	дуже висока

ІНСТРУМЕНТАРІЙ PYTHON

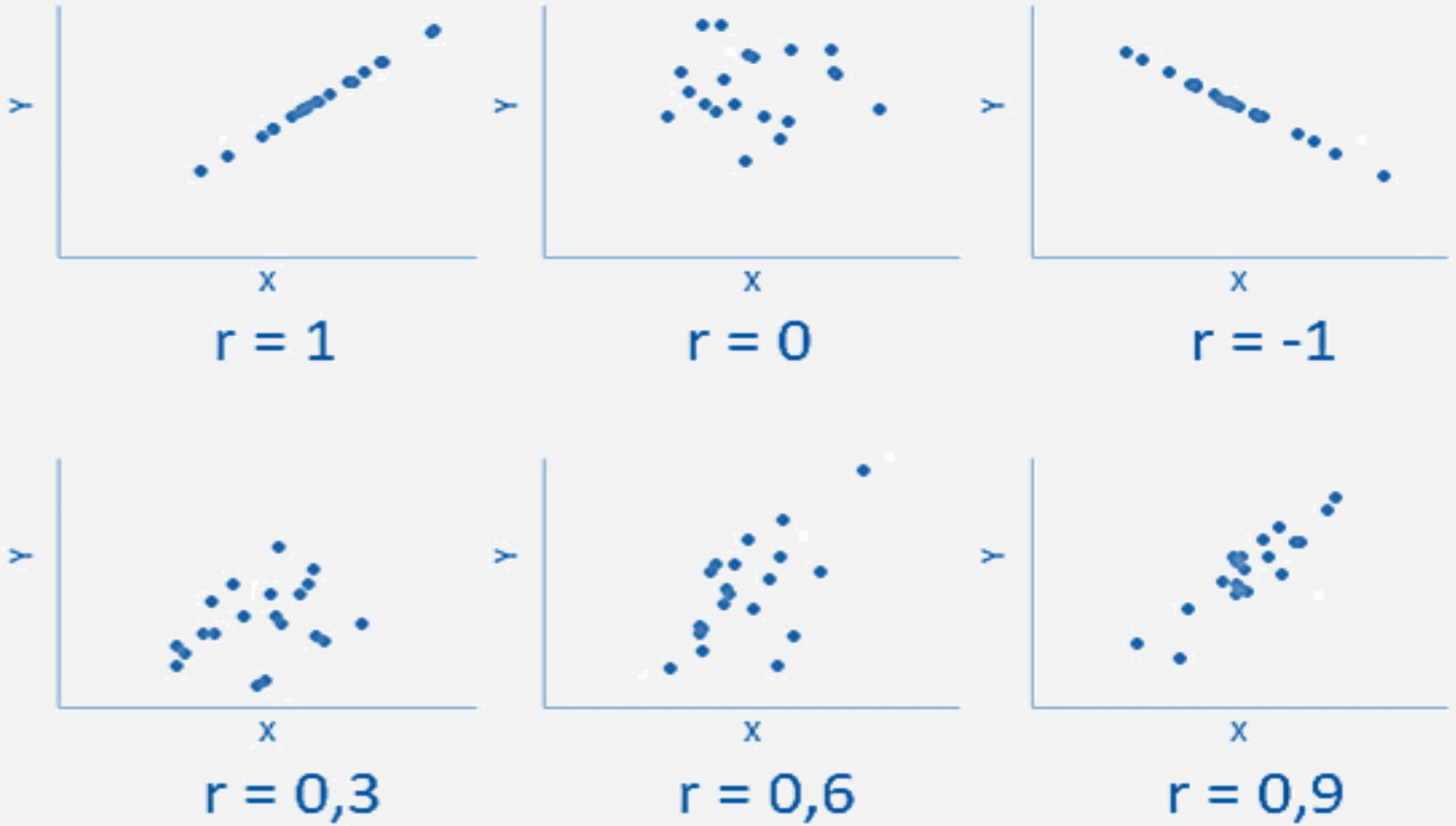
Кореляція вважається сильною, якщо її коефіцієнт вище 0.6, якщо ж він перевищує 0.9, то кореляція вважається дуже сильною.

Однак для того, щоб можна було робити висновки про зв'язки між змінними, велике значення має обсяг вибірки: чим вибірка більше, тим вірогідніше величина отриманого коефіцієнта кореляції.

Існують таблиці з критичними значеннями коефіцієнта кореляції Пірсона та Спірмена для різного числа ступенів свободи (воно дорівнює числу пар за відніманням 2, т. п. n-2).

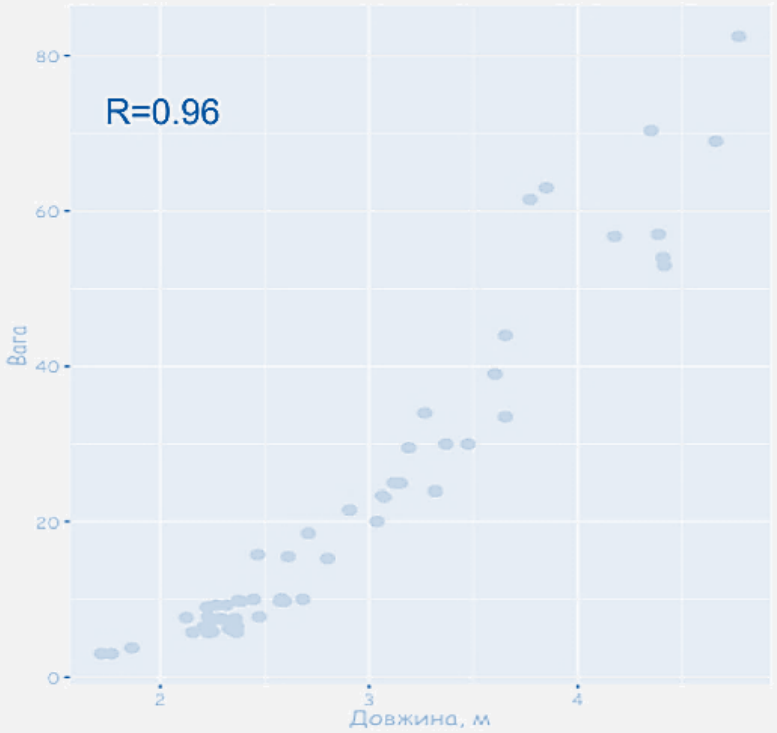
Лише в тому випадку, якщо коефіцієнти кореляції більше цих критичних значень, вони можуть вважатися достовірними. Так, для того щоб коефіцієнт кореляції 0.7 був достовірним, в аналізі має бути не менше 8 пар даних ($\eta = n-2 = 6$) при обчисленні r , 7 пар даних ($\eta = n-2 = 5$) при обчисленні r .

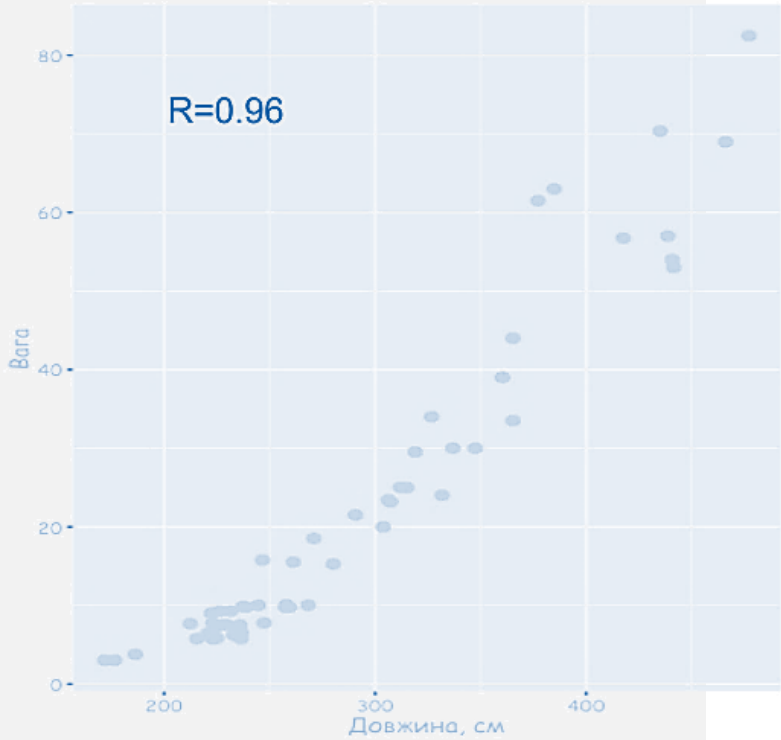
Приклад розсіювання і відповідних коефіцієнтів кореляції



ВЛАСТИВОСТІ КОЕФІЦІЄНТА КОРЕЛЯЦІЇ

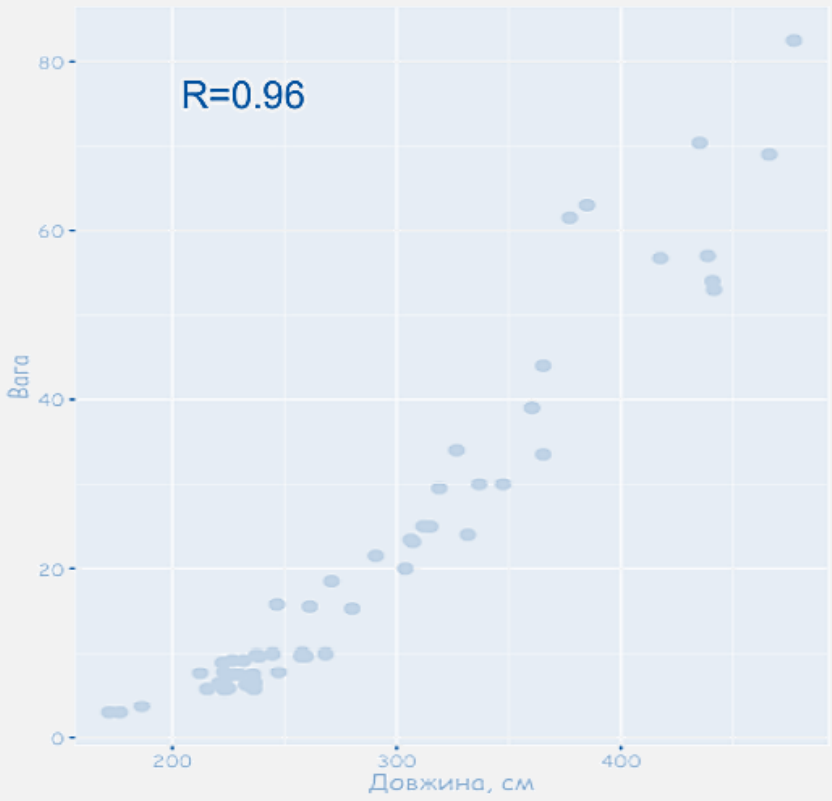
- коефіцієнт кореляції не змінюється при зміні одиниць виміру(наприклад від кілограм до грам)

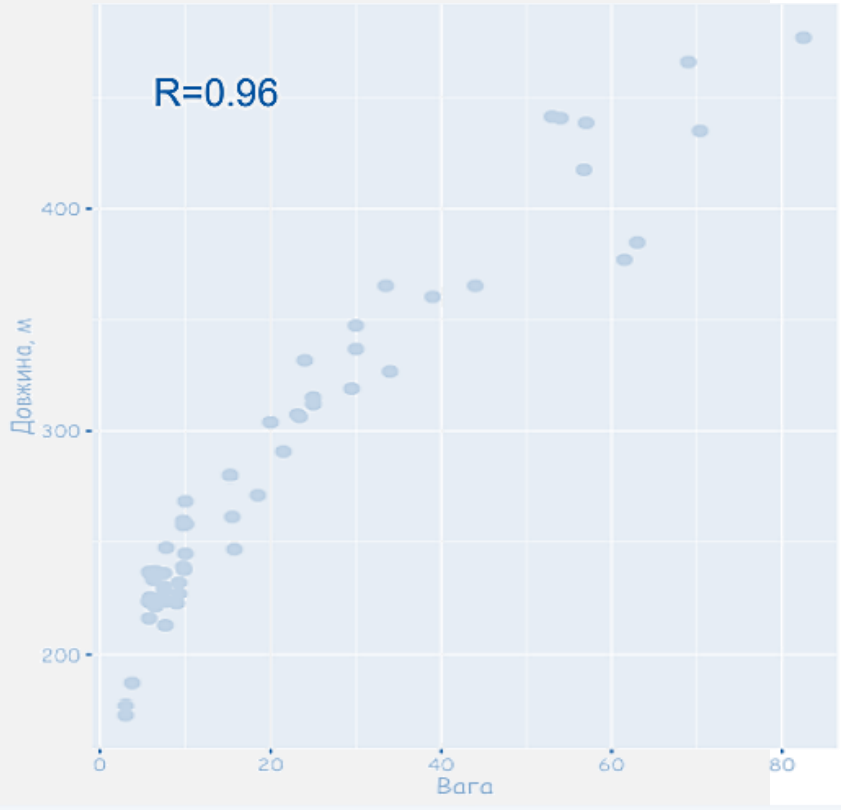




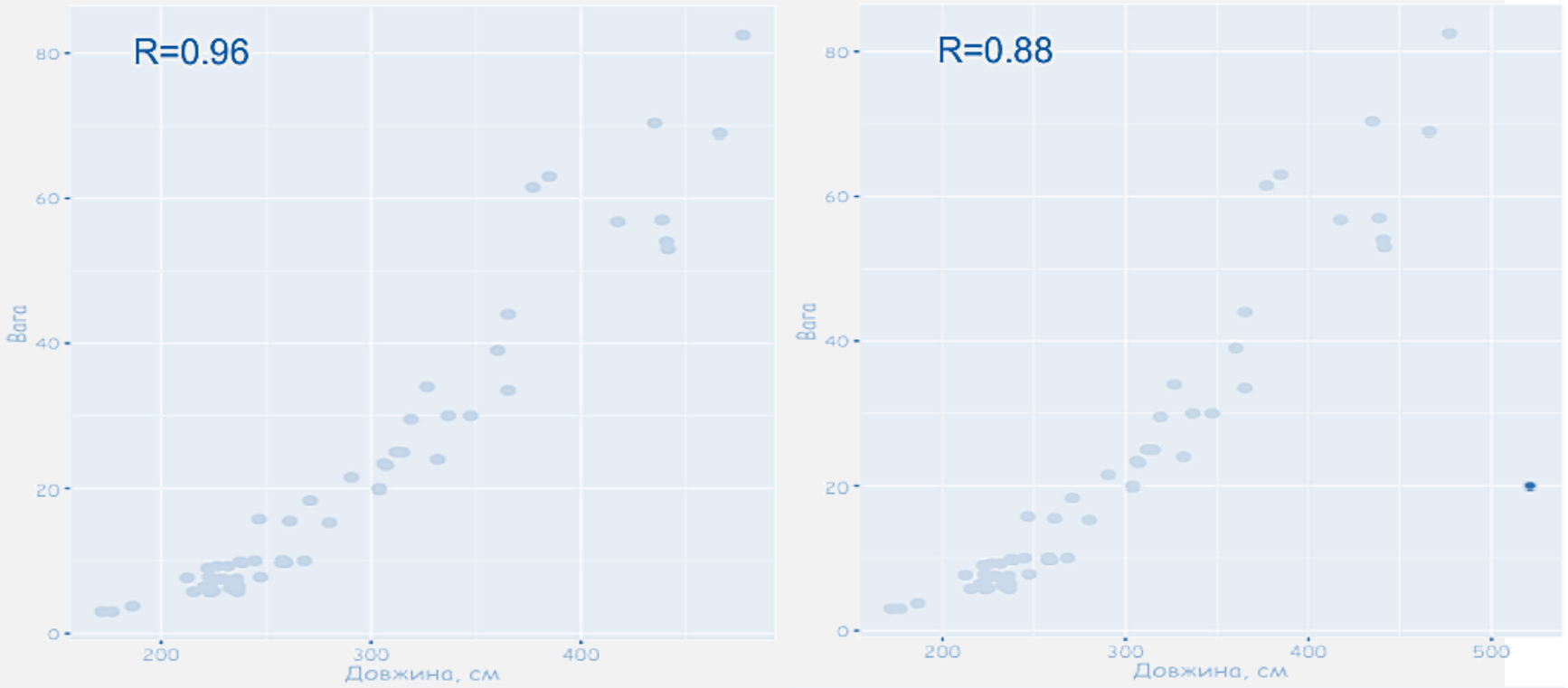
- коефіцієнт кореляції є симетричним

$$r(x, y) = r(y, x)$$





• коефіцієнт кореляції чутливий до викидів



Приклад

Коефіцієнт кореляції Пірсона (r) - це параметричний показник.
Приклад:

	X	Y	ср.X	ср.Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
A	6	82	3,166667	72,16667	2,833333	9,833333	27,86111111	8,027778	96,69444
B	2	63			-1,16667	-9,16667	10,69444444	1,361111	84,02778
C	1	57			-2,16667	-15,1667	32,86111111	4,694444	230,0278
D	5	88			1,833333	15,83333	29,02777778	3,361111	250,6944
E	2	68			-1,16667	-4,16667	4,861111111	1,361111	17,36111
F	3	75			-0,16667	2,833333	-0,472222222	0,027778	8,027778
						∑	104,8333333	18,83333	686,8333
								113,7337	
							r=	0,921743569	

ВАЖЛИВО!!!

Коефіцієнт кореляції r оцінює тільки лінійний зв'язок змінних. Нелінійний зв'язок даний коефіцієнт виявити не може.

ЛІНІЙНА І РАНГОВА КОРЕЛЯЦІЯ

Метод лінійної кореляції (кореляції Пірсона) застосовується для визначення міри відповідності двох ознак, виражених кількісно, іншими словами, - для численних величин.

Метод рангової кореляції (кореляція Спірмена) можна застосувати до будь яких кількісно виміряним або ранжируваним даним. Цей метод здатний, на відміну від інших, вимірювати узгодженість зміни різних ознак у одного випробуваного або виявляти збіги індивідуальних рангових показників у двох досліджуваних; або будь-які показники в порівнянні двох груп.

РАНГОВА КОРЕЛЯЦІЯ

У багатьох випадках результати спостережень подаються не у вигляді кількісних вимірювань, а у вигляді бальних оцінок (рангів).

Наприклад, студенти у групі можуть бути впорядковані по номерам за середнім балом в сесії, країни – за кількістю населення, учасники конкурсу – за зайнятим місцем тощо. При цьому інколи виникає можливість упорядкувати об’єкти дослідження за двома або більше показниками. У зв’язку з цим виникає задача дослідження кореляції цих показників.

РАНГОВА КОРЕЛЯЦІЯ

Коефіцієнт кореляції рангів Спірмена (r_s) - це непараметричний показник, за допомогою якого намагаються виявити зв'язок між рангами відповідних величин в двох рядах вимірів.

Цей коефіцієнт розраховувати простіше, проте результати виходять менш точними, ніж при використанні r. Це пов'язано з тим, що при обчисленні коефіцієнта Спірмена використовують порядок проходження даних, а не їх кількісні характеристики та інтервали між класами.

РАНГОВА КОРЕЛЯЦІЯ

Коефіцієнт r_s обчислюють за формулою

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

де d - різниця між рангами сполучених значень ознак (незалежно від знаку), $d=x_i-y_i$

n - число пар.

Зазвичай цей непараметричний аналіз використовується в тих випадках, коли потрібно зробити якісь висновки не стільки про інтервали між даними, скільки про їх ранги, а також тоді, коли криві розподілу занадто асиметричні і не дозволяють використовувати такі параметричні критерії, як коефіцієнт r (в цих випадках буває необхідно перетворити кількісні дані в порядкові).

ПРИКЛАД

У таблиці наведено дані про місця, що займають 8 провідних компаній галузі за собівартістю продукції (фактор x) та часткою ринку (фактор y). Обчислити коефіцієнт рангової кореляції Спірмена.

Підприємство	A	B	C	D	E	F	G	H
Фактор x	8	3	1	4	2	7	5	6
Фактор y	3	5	6	7	8	4	1	2
d=x-y	5	-2	-5	-3	-6	3	4	4

$$\sum_{i=1}^{8} d_i^2 = 25+4+25+9+36+9+16+16 = 140.$$

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)} = 1 - \frac{6\cdot 140}{8\cdot 63} = -0{,}67.$$

Отримане значення коефіцієнта рангової кореляції Спірмена свідчить про наявність взаємозв'язку між собівартістю продукції компанії та часткою ринку. При цьому спостерігається обернена залежність: зі зростанням собівартості продукції компанії її частка ринку зменшується

ІНСТРУМЕНТАРІЙ PYTHON

В Python для обчислення коефіцієнта кореляції використовується функція corr(), яка по замовчуванню рахує коефіцієнт кореляції Пірсона.

У Pandas:

```
df.corr(method='pearson')
df.corr(method='spearman')
```

У SciPy:

```
pearsonr(x, y)
spearmanr(x, y)
```

ВІЗУАЛІЗАЦІЯ КОРЕЛЯЦІЇ

Для інтерпретації кореляцій застосовують:

- кореляційні матриці
- теплові карти (heatmap)

Практична цінність:

- виявлення мультиколінеарності
- відбір ознак перед побудовою моделей
- аналіз взаємозв’язків між показниками здоров’я або ризику

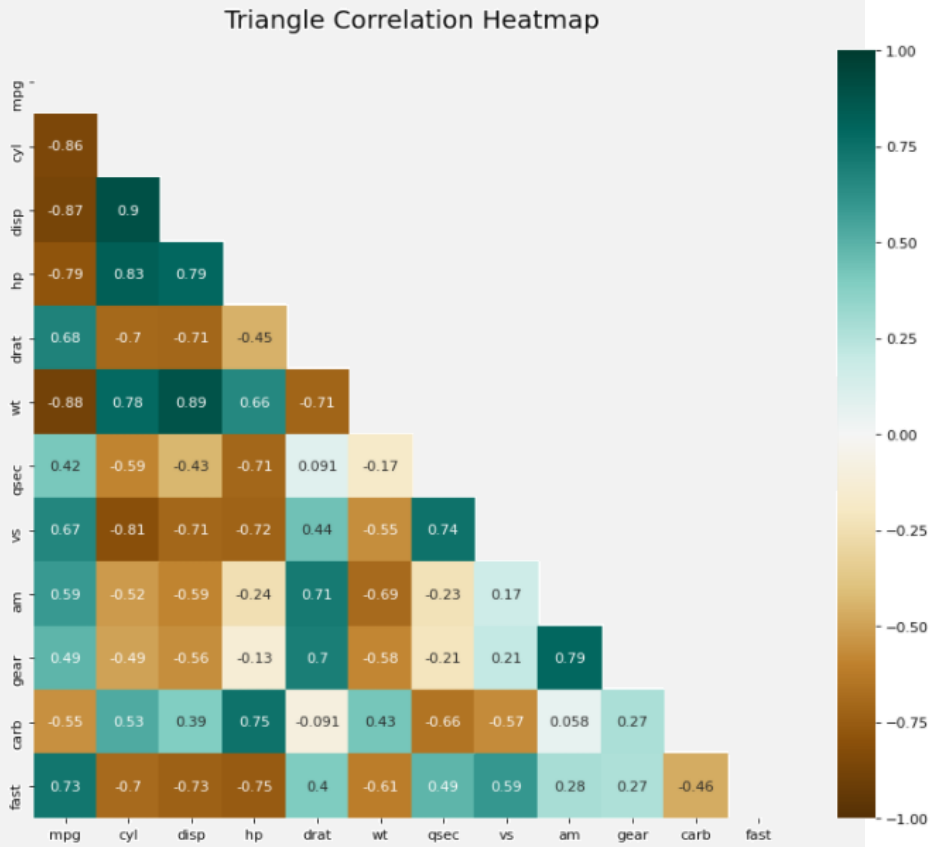
ІНСТРУМЕНТАРІЙ PYTHON

```
plt.figure(figsize=(12,10), dpi= 80)
sns.heatmap(df.corr(), xticklabels=df.corr().columns, yticklabels=df.corr().
columns, cmap='RdYlGn', center=0, annot=True)
```

Діаграма кореляції використовується для візуального перегляду метрики кореляції між усіма можливими парами числових змінних в даному наборі даних (або двовимірному масиві).

ІНСТРУМЕНТАРІЙ PYTHON

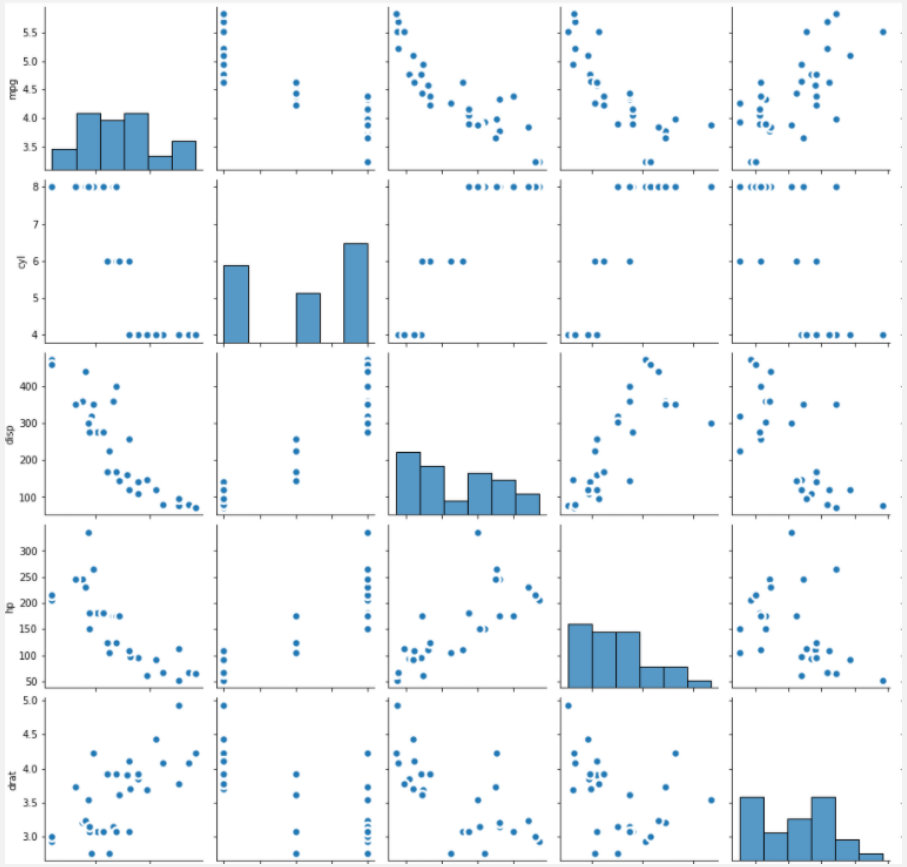
```
plt.figure(figsize=(12,10), dpi= 80)
mask = np.triu(np.ones_like(df.corr(
), dtype=np.bool))
heatmap = sns.heatmap(df.corr(), mas
k=mask, vmin=-
1, vmax=1, annot=True, cmap='BrBG')
heatmap.set_title('Triangle Correlat
ion Heatmap', fontdict={'fontsize':1
8}, pad=16)
```



ІНСТРУМЕНТАРІЙ PYTHON

Парний графік часто використовується в дослідницькому аналізі, щоб зрозуміти взаємозв'язок між усіма можливими парами числових змінних.

Це обов'язковий інструмент для двовимірного аналізу.




```
sns.pairplot(df,plot_kws=dict(s=80, edgecolor="white",
 linewidth=2.5))
plt.show()
```