Introduction to pandas

• Measures of centeral Tendency :

1) mean (Arithmetic Average)

- The sum of all values divided by the total number of values
- It's the common measure but very sensetive to Outliers

. المتوسط : المجموع على العدد

. It's used when the data is normal distribution
. there is no "outliers"

. حساس جدًا للقيم الشاذة

In pandas : df['column'].mean()

2) median (middle value)

- The middle point of the data when it is arranged in ascending order.
- It's not affected by outliers
- If you have a few people with massive salaries, the median gives a more "honest" picture of the average person's income.

الوسيط : العنصر المتوسط بعد ترتيب الداتا.
لا يتأثر بالقيم الشاذة

In pandas : df['column'].median()

3) mode (most frequent)

- The value that appears most often in the dataset.
- Mainly used for Categorical Data (words / labels)

الأكثر تكرارًا
يُفضل استخدامه مع البيانات النصية

In pandas : df['column'].mode()

Note : Strategy : Skewness, Distribution :
before you choose which one to use for Imputation (filling missing values)
You look at the distribution of your data :

1- Symmetrical (Normal distribution) عندما تكون البيانات منظمة "لا يوجد بها أرقام شاذة"
use the Mean

2- Skewed (Asymmetrical Distribution) الأرقام عالية جدًا
The mean is pulled toward the "tail" Use the Median. "بها قيم شاذة"

3- Data Categorical:

عندما تكون البيانات "نصية"

Use the Mode

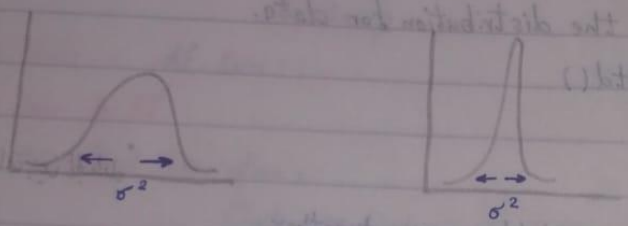
- df['Column'].describe()

↳ Calculate all of them.

4) Variance (σ^2)

أنتشار الأرقام حول "mean"

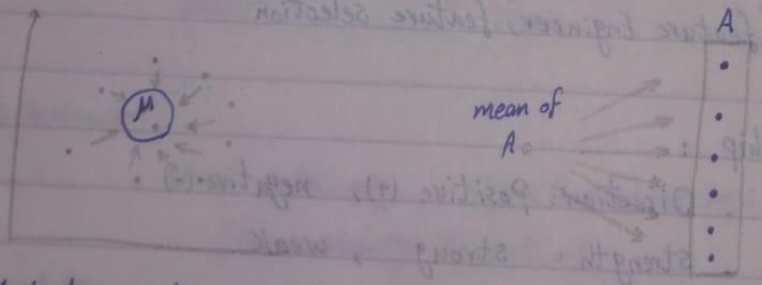
- It measure how far each number of the set from the mean.
- It is average of the squared difference from the mean.
- High variance means the data is spread out
- Low variance means the data points are clustered closely around the average.



$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

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

In AI High variance often refers to overfitting (where the model is too sensetive to small fluctuation).



if μ is large, it means there is variance.

if μ is small, it means there is no variance.

σ^2 . means here the difference between elements of A and there mean (μ)

In pandas: df['Column'].var()

5) Standard Deviation (σ) الانحراف المعياري.

- The square root of the variance. $\sigma = \sqrt{\text{variance}}$
- we use it to identify Outliers

The difference :

Variance (σ^2)	Standard deviation (σ)
• The average of squared difference from the mean.	• The square root of the variance
• squared units (it makes it hard to visulize	• original Units (This is intuitive)
• Used in mathematical optimization and loss functions.	• Use to detect Outliers and for data scaling.

There are used for Calculate the distribution for data.

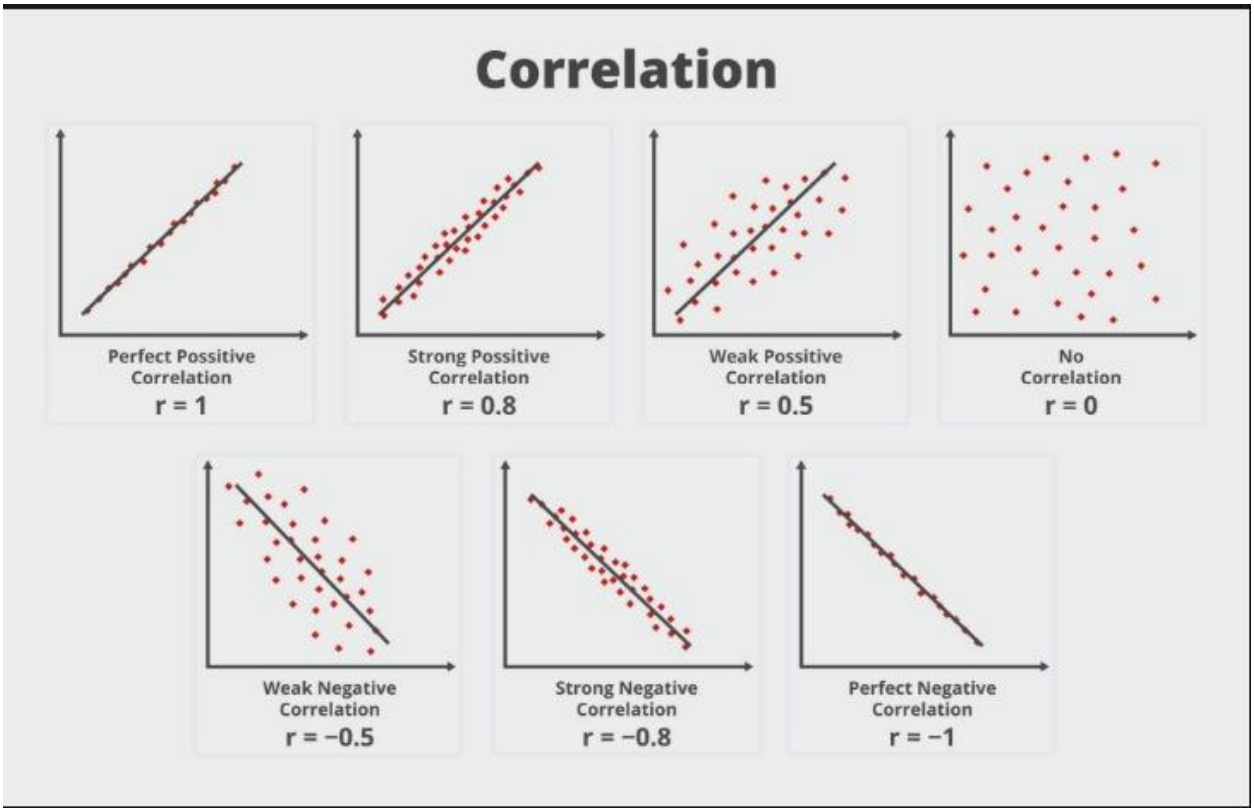
In pandas: `df['Column'].std()`

6) Covariance "العلاقات بين الجداول"

- It means how two variables move together.
  - positive Covariance : Both variables move together
  - negative Covariance : One increases, the other decreases.
- Used for "feature Engineer,"
- "PCA"
- Relation ship :
  - Direction: positive (+), negative (-)
  - strength : strong , weak

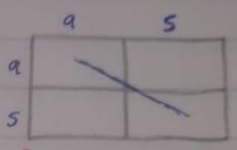
Four Quadrants :

strong positive : when Column A goes up, Column B goes up significantly.
strong negative : when Column A goes up, Column B goes down significantly.
weak positive : There is a slight upward trend, but it is 'noisy'
weak negative : There is a slight downword trend,


Correlation
Perfect Possitive Correlation
r = 1
Strong Possitive Correlation
r = 0.8
Weak Possitive Correlation
r = 0.5
No Correlation
r = 0
Weak Negative Correlation
r = −0.5
Strong Negative Correlation
r = −0.8
Perfect Negative Correlation
r = −1

7) Correlation

- A standardized version of Covariance . It ranges from -1 to +1
- +1 perfect positive relationship.
- -1 perfect negative relation ship.
- 0 No relationship at all.
- It's vital for "Feature selection". if two columns are 99% correlated You can drop one of them because they are provide the same information for AI model.
- Used in
  - Confusion matrix.
  - Correlation matrix
  - heatmap



"main diagonal" means relationship with "itself."

In Pandas : df.corr()

Use Case In AI:

Standard Deviation :

Data scaling and Outlier detection.

Variance :

Understanding model error

Co Variance :

Used internally in algorithms like PCA (Dimensionality Reduction)

Correlation :

Selecting which features (columns) are most important for the model.

Feature Engineering: Creating new information for existing columns . If we have "length", "width", You might create a new feature like 'Area'.

Feature selection: The process of picking the most important variable. if two columns are (highly related), we can drop one of them to make model faster and more accurate.