Correlation Concept

Correlation

Definition: Correlation measures the strength and direction of a linear relationship between two variables.

Formulas:

1. **Pearson Correlation Coefficient (r)**:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where n is the number of pairs, x and y are the individual sample points indexed with i.

Concepts:

- **Strength of Relationship**: The correlation coefficient ranges from -1 to +1. A value close to +1 implies a strong positive relationship, close to -1 implies a strong negative relationship, and around 0 implies no linear relationship.
- **Direction of Relationship**: Indicates whether the variables increase together (positive) or as one increases, the other decreases (negative).

Real-life Use Case:

- Determining the relationship between consumer income and expenditure on luxury items.
- In finance, correlating different stock movements to diversify the portfolio.

Correlation Matrix

Types of Correlation Coefficients

1. **Pearson Correlation Coefficient**: Measures the linear relationship between two continuous variables. It assumes that the variables follow a normal distribution.
2. **Spearman's Rank Correlation Coefficient**: A non-parametric measure of rank correlation, assessing how well the relationship between two variables can be described using a monotonic function.
3. **Kendall's Tau**: Another rank-based correlation coefficient, often used for data with a natural ordinal classification.

dataset →

employee

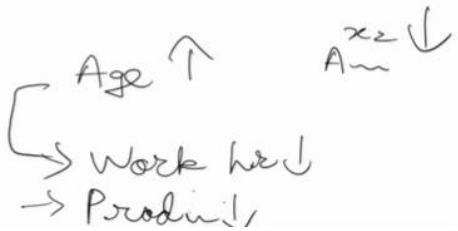
	Age	Annual Income (K$)	Spending Score (1-100)	Years with Company
0	34	52	48	6
1	27	48	57	4
2	33	62	65	6
3	21	37	25	2
4	38	73	88	8
...	...	...	...	...

	Age	Annual Income (K$)	Spending Score (1-100)	Years with Company
Age	1.00	-0.10	-0.20	0.95
Annual Income (K$)	-0.10	1.00	0.60	-0.05
Spending Score (1-100)	-0.20	0.60	1.00	-0.15
Years with Company	0.95	-0.05	-0.15	1.00

Understanding a Correlation Matrix

A correlation matrix displays correlation coefficients between sets of variables. Each cell in the matrix represents the correlation between two variables. The value ranges from -1 to +1, where:

- **+1** indicates a perfect positive linear relationship,
- **-1** indicates a perfect negative linear relationship,
- **0** indicates no linear relationship.



Real-Life Use Cases

1. **Finance**: Correlation matrices are used to understand the relationships between different stocks or financial instruments, aiding in portfolio diversification.
2. **Marketing**: Analyzing customer data to find correlations between different buying behaviors.
3. **Healthcare**: Studying correlations between different lifestyle factors and health outcomes.