manipalglobal
Academy of Data Science

Statistical Techniques for Data Science

Correlation

Dr. Subhabaha Pal
Manipal Global Academy of Data Science

Objective

manipalglobal
Academy of Data Science

After attending this session, you will be able to –

- **Explain what is Correlation**
- **Describe Correlation Co-efficient**
- **Explain effect of Outliers on Correlation**
- **Explain what is Multi-collinearity**

Correlation

manipalglobal
Academy of Data Science

- **Correlation**: The degree of relationship between the variables under consideration is measured through the correlation analysis.
- The measure of correlation is called the correlation coefficient
- The degree of relationship is expressed by coefficient which range from correlation **(-1 ≤ r ≥ +1)**
- The direction of change is indicated by a sign
- The correlation analysis enable us to have an idea about the degree & direction of the relationship between the two variables under study.

Correlation

manipalglobal
Academy of Data Science

- Correlation is a statistical tool that helps to measure and analyze the degree of relationship between two variables.
- Correlation analysis deals with the association between two or more variables.

Correlation & Causation

manipalglobal
Academy of Data Science

- Causation means cause & effect relation.
- Correlation denotes the interdependency among the variables for correlating two phenomenon, it is essential that the two phenomenon should have cause-effect relationship,& if such relationship does not exist then the two phenomenon can not be correlated.
- If two variables vary in such a way that movement in one are accompanied by movement in other, these variables are called cause and effect relationship.
- Causation always implies correlation but correlation does not necessarily implies causation.

Methods of Studying Correlation

manipalglobal
Academy of Data Science

- Karl Pearson's Coefficient of Correlation
- Scatter Diagram Method

Karl Pearson's Correlation Co-efficient

manipalglobal
Academy of Data Science

- Covariance explains the relationship between 2 variables, but units of covariance is attached to the unit of covariance
- Expression of Covariance for Population –

$$\text{Covariance}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

- Expression of Covariance for Sample –

$$\text{Covariance}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Karl Pearson's Correlation Co-efficient

manipalglobal
Academy of Data Science

- Expression for Karl Pearson's Correlation Co-efficient – (it has no unit) –

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{Var}(x)}\sqrt{\text{Var}(Y)}}$$

Or,

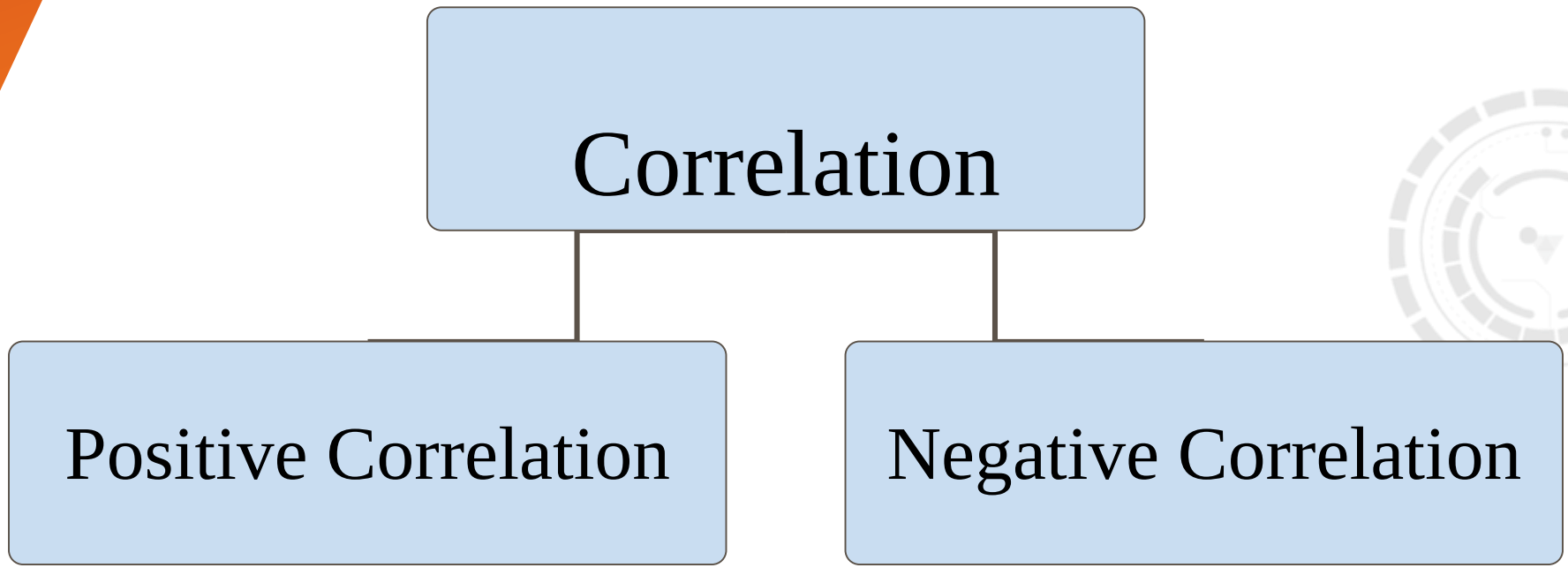
$$r = \frac{\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- In general for computation purpose, the following formula is used -

$$r = \frac{n\sum_{i=1}^{n}X_iY_i - \left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2\right)}\sqrt{\left(n\sum_{i=1}^{n}Y_i^2 - \left(\sum_{i=1}^{n}Y_i\right)^2\right)}}$$

Types of Correlation
Type I



Types of Correlation Type I

manipalglobal
Academy of Data Science

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Pub. Exp. & sales, Height & weight.

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Price & qty. demanded.

Direction of the Correlation

- **Positive relationship** – Variables change in the same direction.
  - As X is increasing, Y is increasing
  - As X is decreasing, Y is decreasing
  - E.g., As height increases, so does weight.
- **Negative relationship** – Variables change in opposite directions.
  - As X is increasing, Y is decreasing
  - As X is decreasing, Y is increasing
  - E.g., As TV time increases, grades decrease

Indicated by sign; (+) or (-).

More examples

Positive relationships

water consumption and temperature.

study time and grades.

Negative relationships:

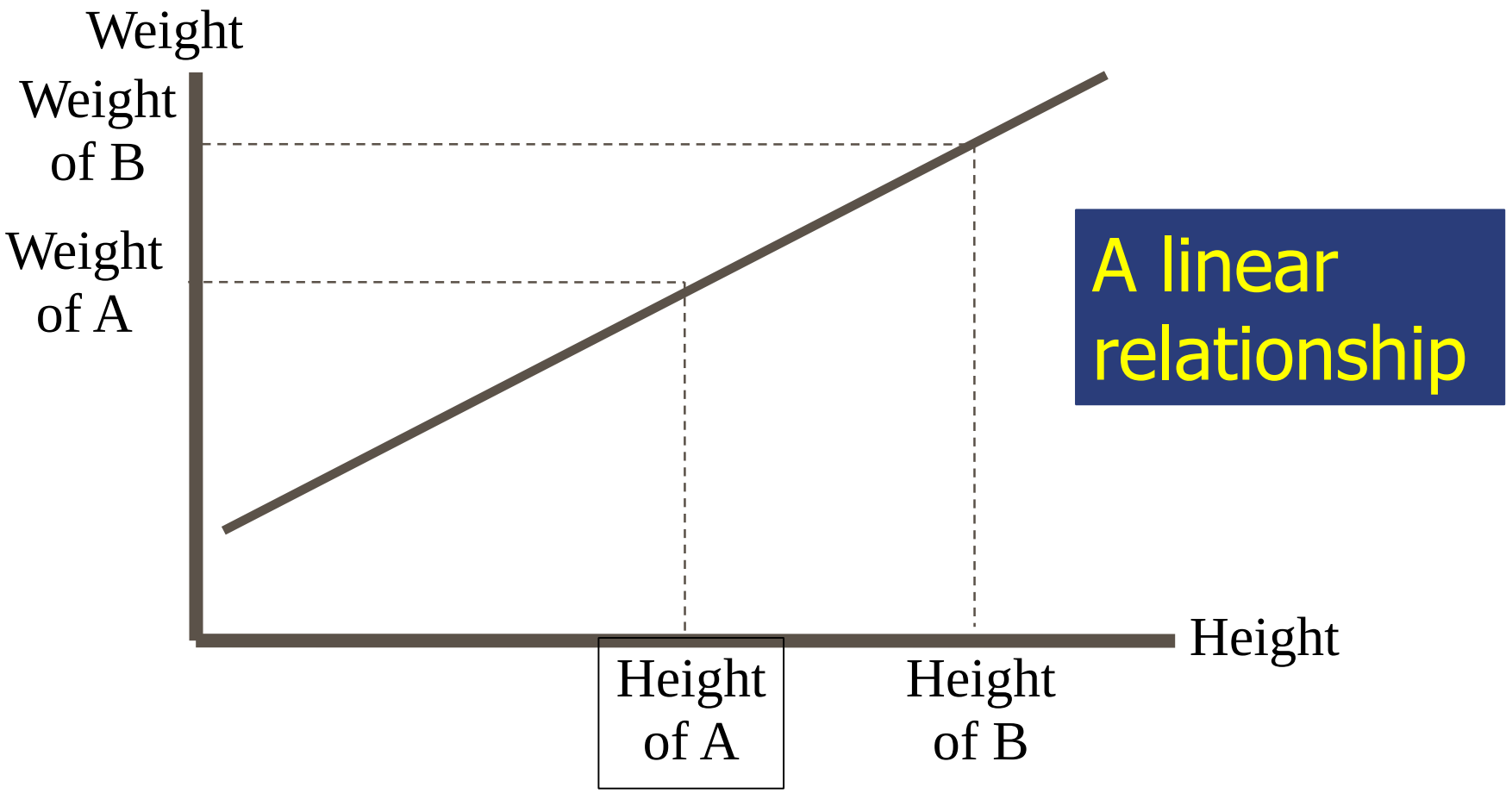
alcohol consumption and driving ability.

Price & quantity demanded

Scatter Diagram Method

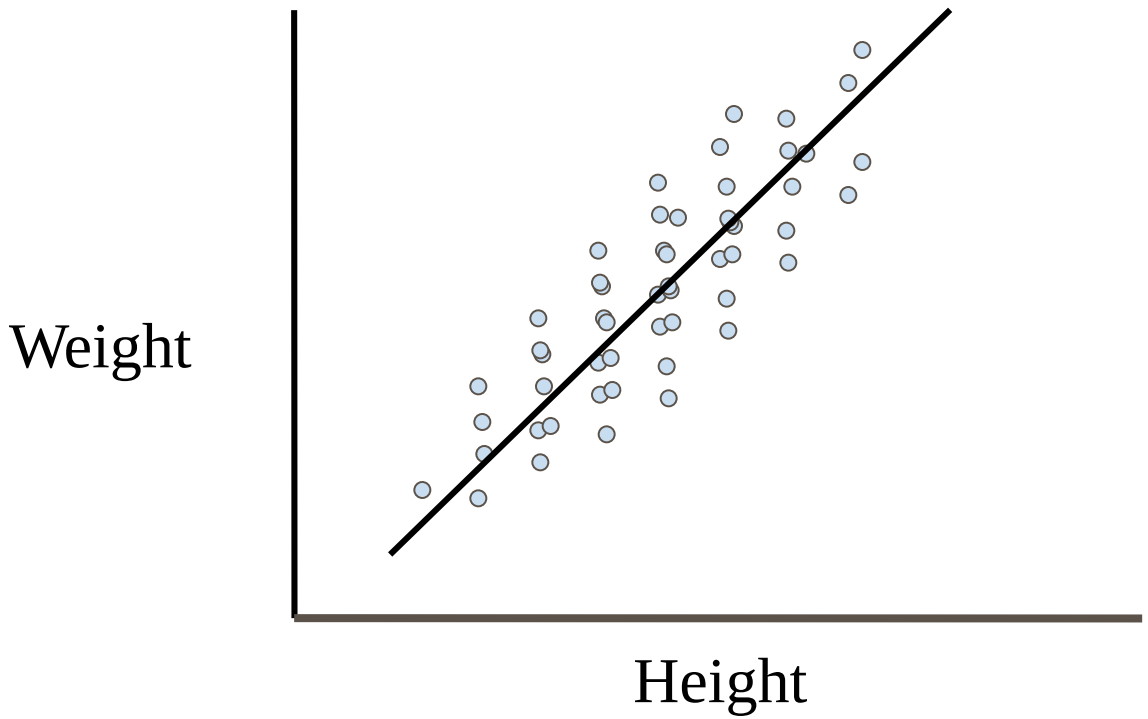
- Scatter Diagram is a graph of observed plotted points where each points represents the values of X & Y as a coordinate. It portrays the relationship between these two variables graphically.

A perfect positive correlation



High Degree of positive correlation

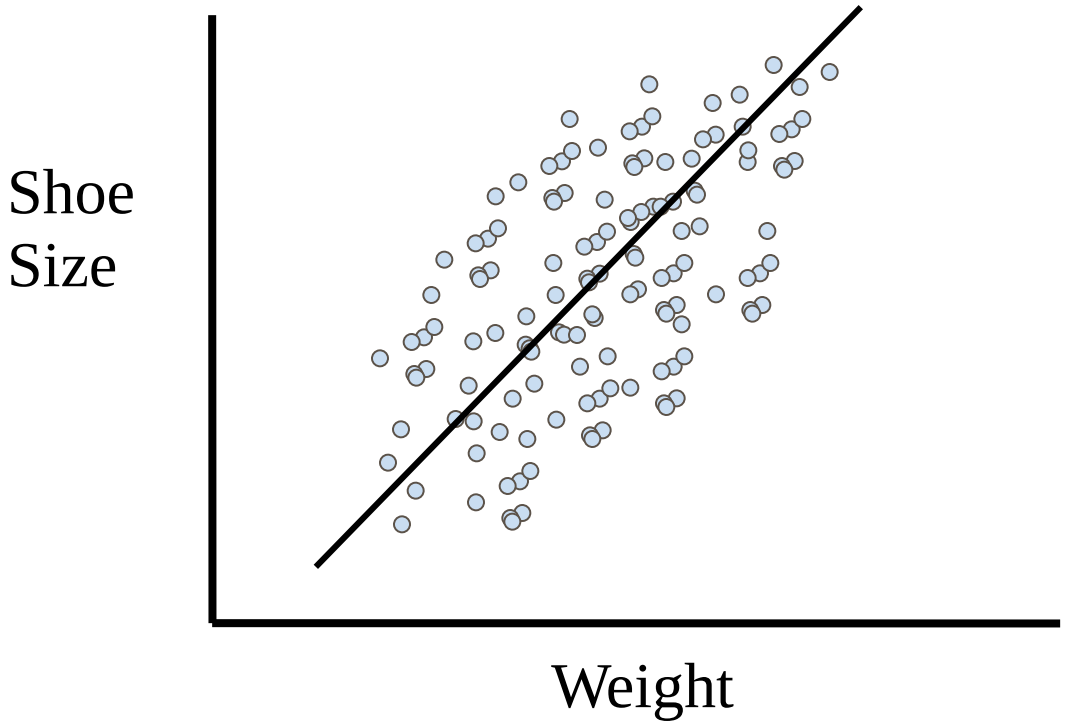
- Positive relationship



r = +.80

Degree of correlation

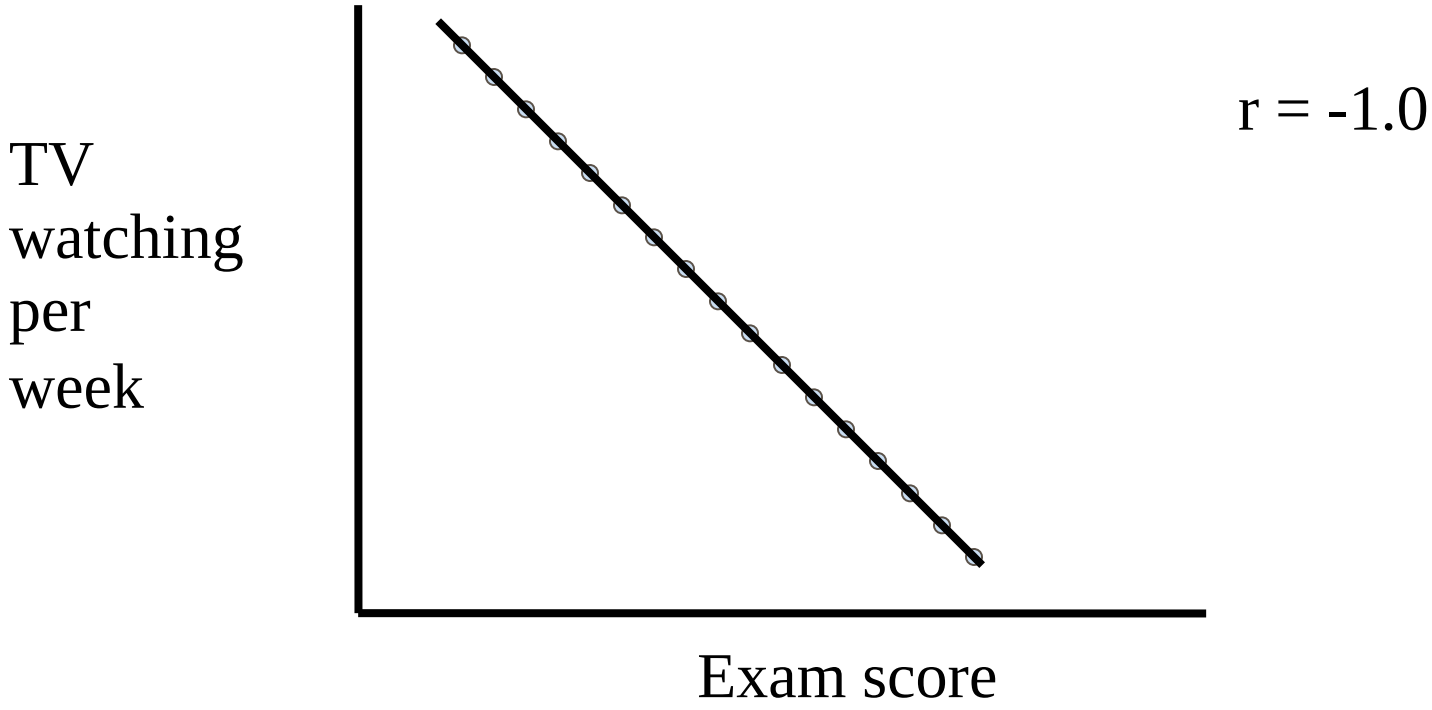
- **Moderate Positive Correlation**



r = + 0.4

Degree of correlation

- **Perfect Negative Correlation**

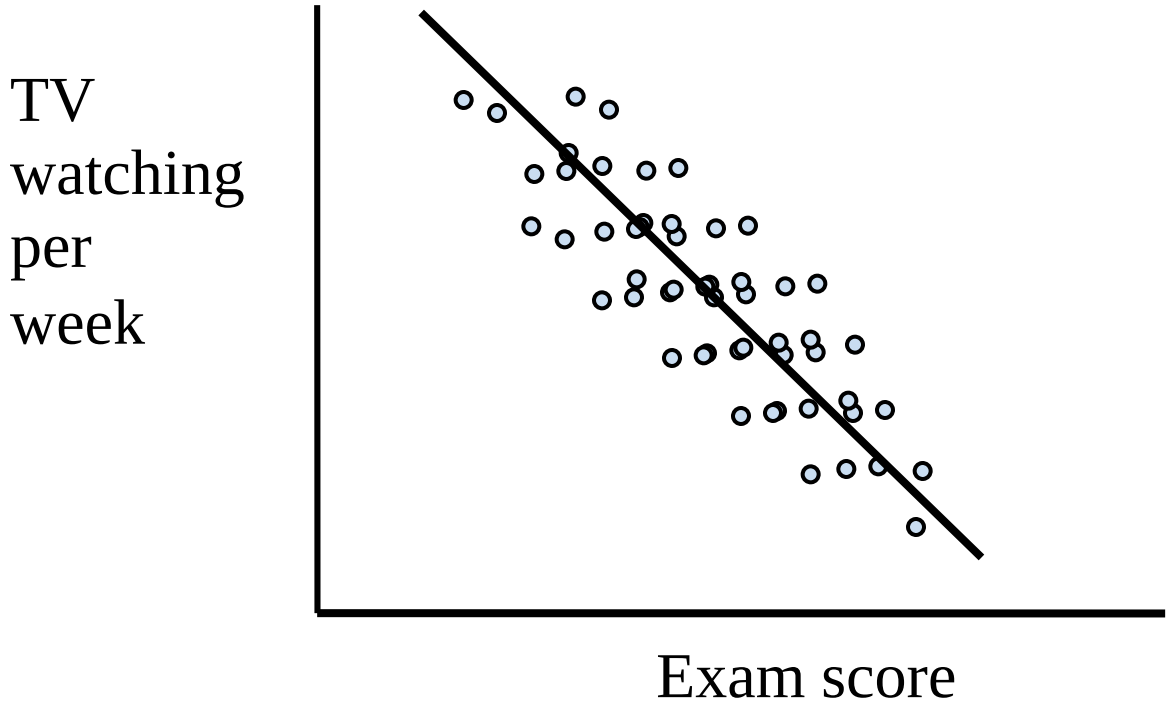


r = -1.0

Degree of correlation

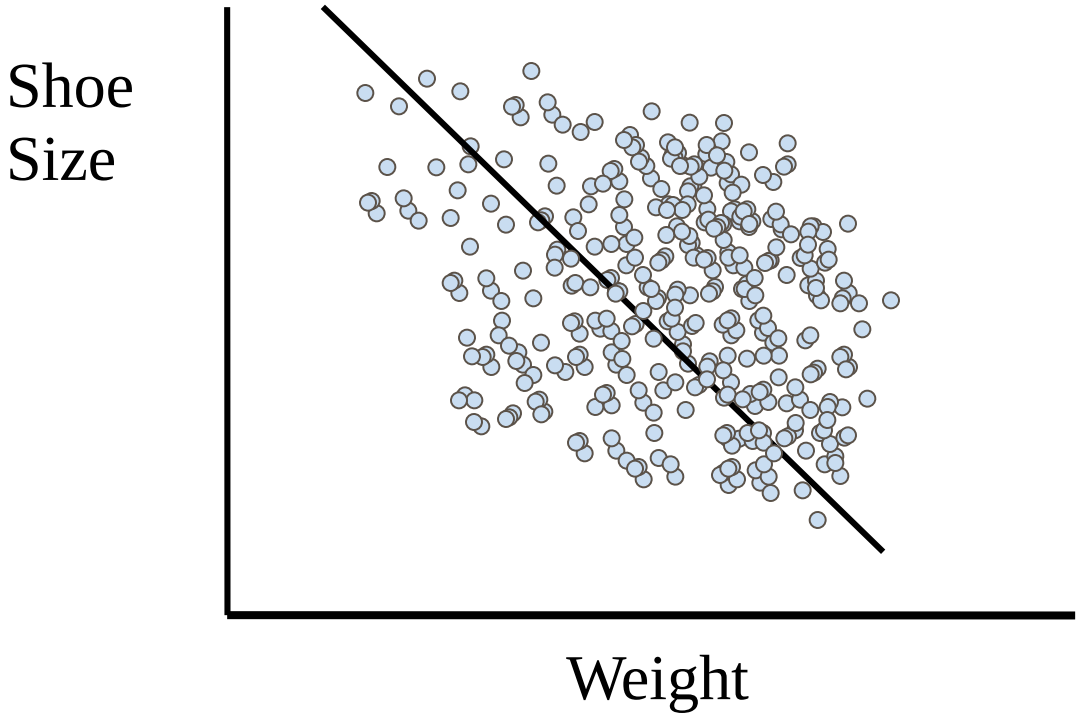
- **Moderate Negative Correlation**

r = -.80



Degree of correlation

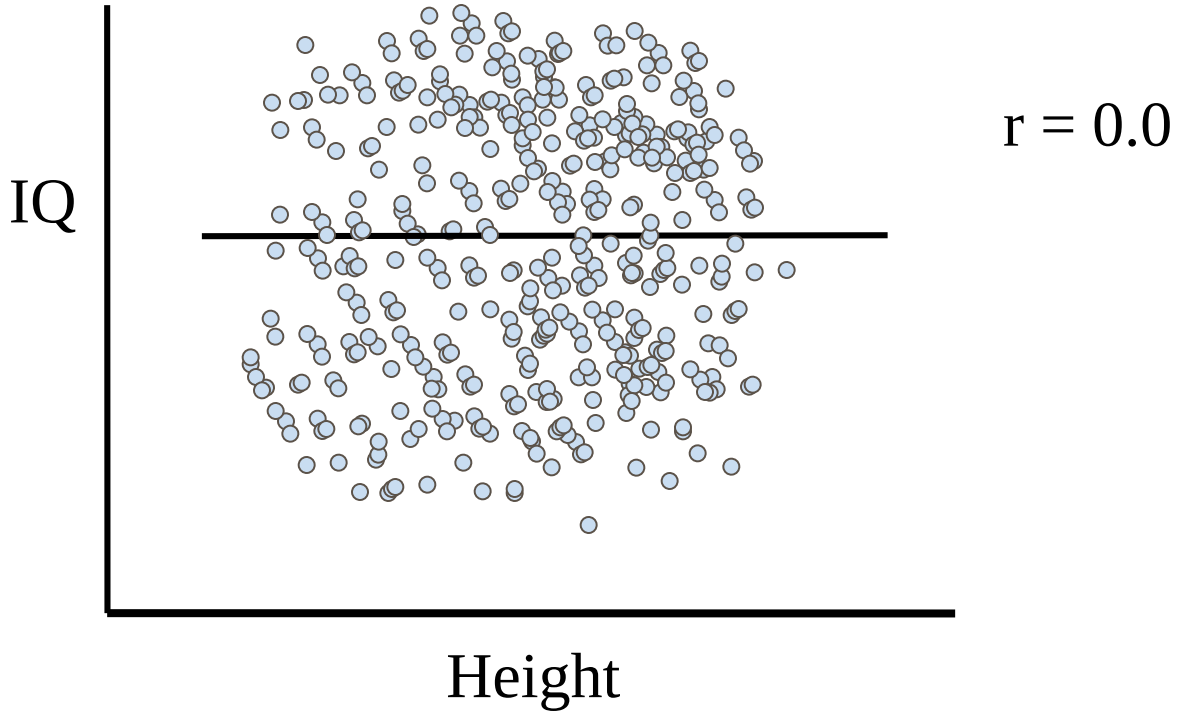
- **Weak negative Correlation**



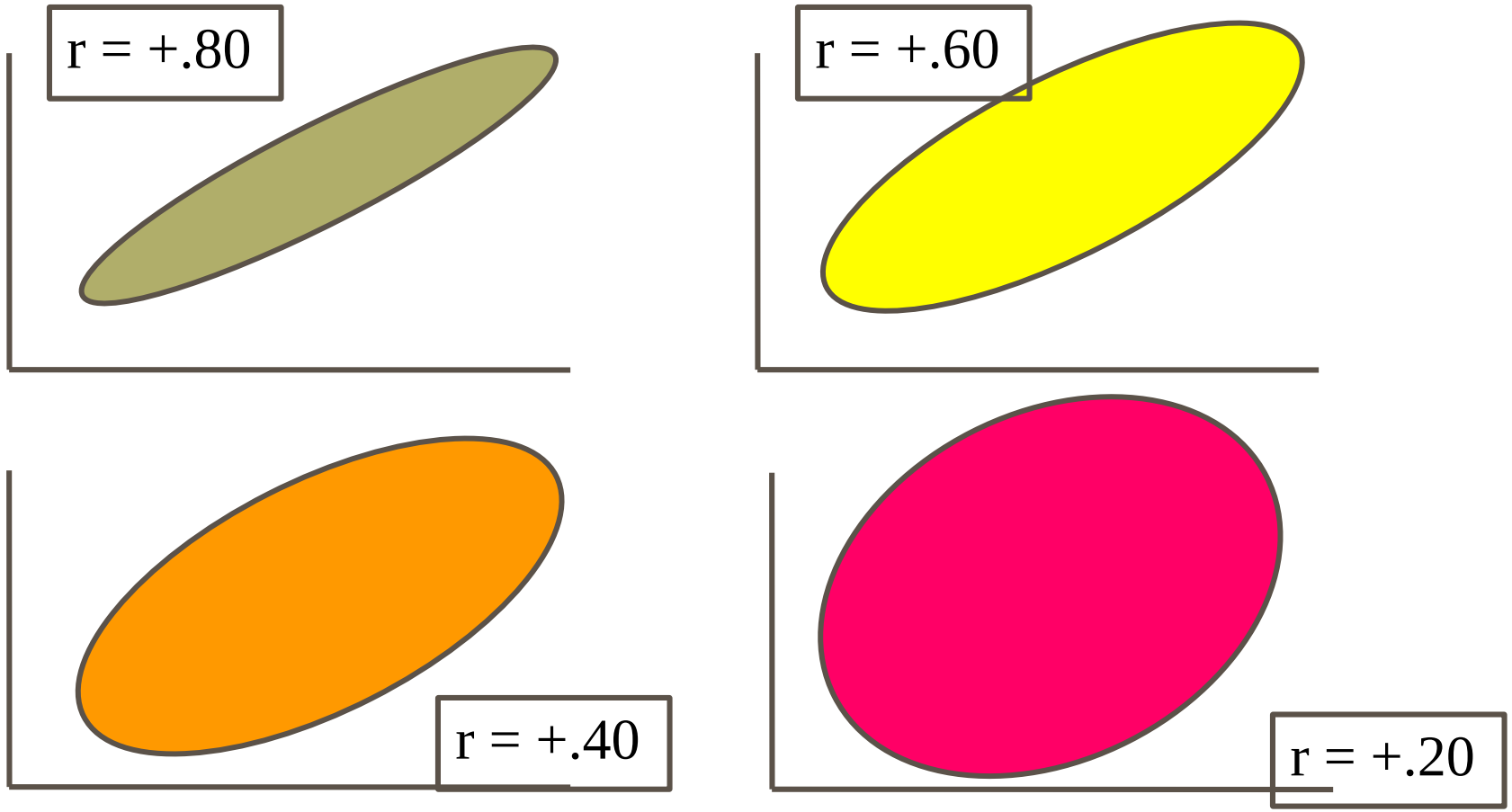
r = - 0.2

Degree of correlation

- **No Correlation (horizontal line)**



Degree of correlation (r)



Advantages of Scatter Diagram

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

Disadvantage of scatter diagram

Can not adopt the an exact degree of correlation

Assumptions of Pearson's Correlation Coefficient

manipalglobal
Academy of Data Science

- There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- Cause and effect relation exists between different forces operating on the item of the two variable series.

Advantages of Pearson's Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

Limitation of Pearson's Coefficient

manipalglobal
Academy of Data Science

- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming methods

Karl Pearson's correlation coefficient (r) Example

manipalglobal
Academy of Data Science

Serial No	Age (yrs)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Karl Pearson's correlation coefficient (r)

manipalglobal
Academy of Data Science

Sl no.	Age (yrs) (X)	Wt (Kg) (Y)	XY	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	$\sum X=41$	$\sum Y=66$	$\sum XY= 461$	$\sum X^2= 291$	$\sum Y^2= 742$

Karl Pearson's correlation coefficient (r)

$$r = \frac{6 \text{ x } 461 - 41 \text{ x } 66}{\sqrt{\left[6 \text{ x } 291 - (41)^2\right]} . \sqrt{\left[6 \text{ x } 742 - (66)^2\right]}}$$

r = 0.759 (Positive (Direct) strong correlation)

Example: Relationship between anxiety and test Scores

manipalglobal
Academy of Data Science

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY = 129$

Karl Pearson's correlation coefficient (r)

manipalglobal
Academy of Data Science

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{\left(6(230) - 32^2\right)\left(6(204) - 32^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -0.94$$

r = - 0.94 (Negative (Indirect) strong correlation)

Effect of Outlier on Correlation

manipalglobal
Academy of Data Science

- In most practical circumstances an outlier decreases the value of a correlation coefficient
- If both the variables X and Y have outlier, then in certain case, the correlation coefficient may increase also

manipalglobal
Academy of Data Science