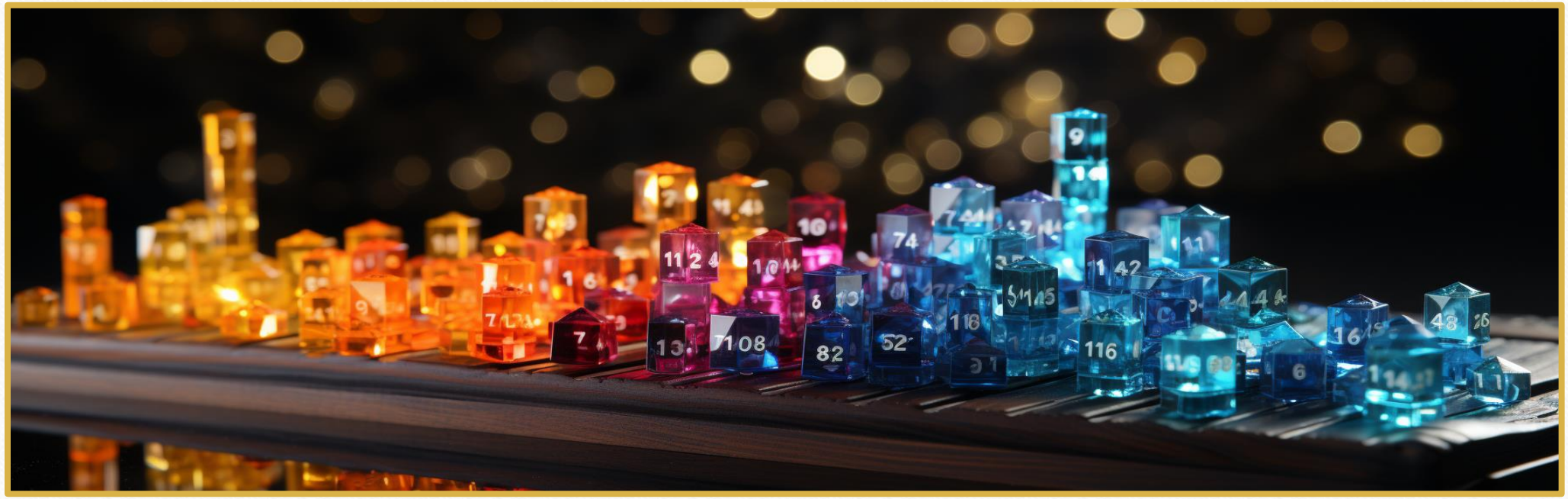
UXG1205 Lecture
14. Scatter Plots and Correlation Coefficient

LIN QINJIE

Outline

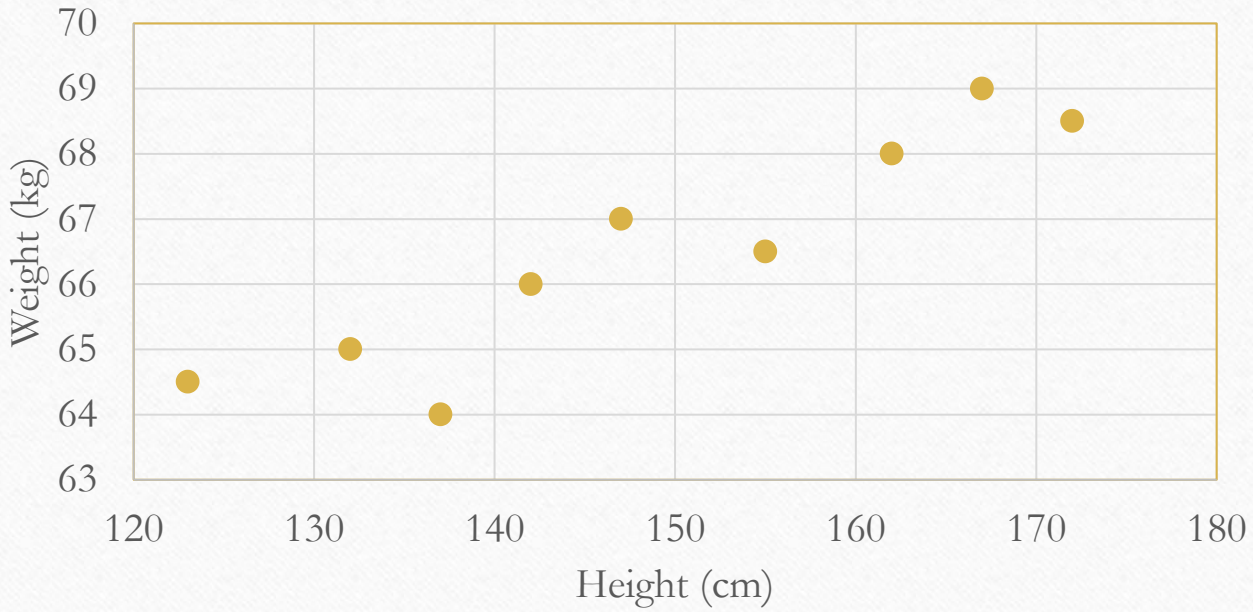
- Scatter plots
- Correlation coefficient



Scatter Plots

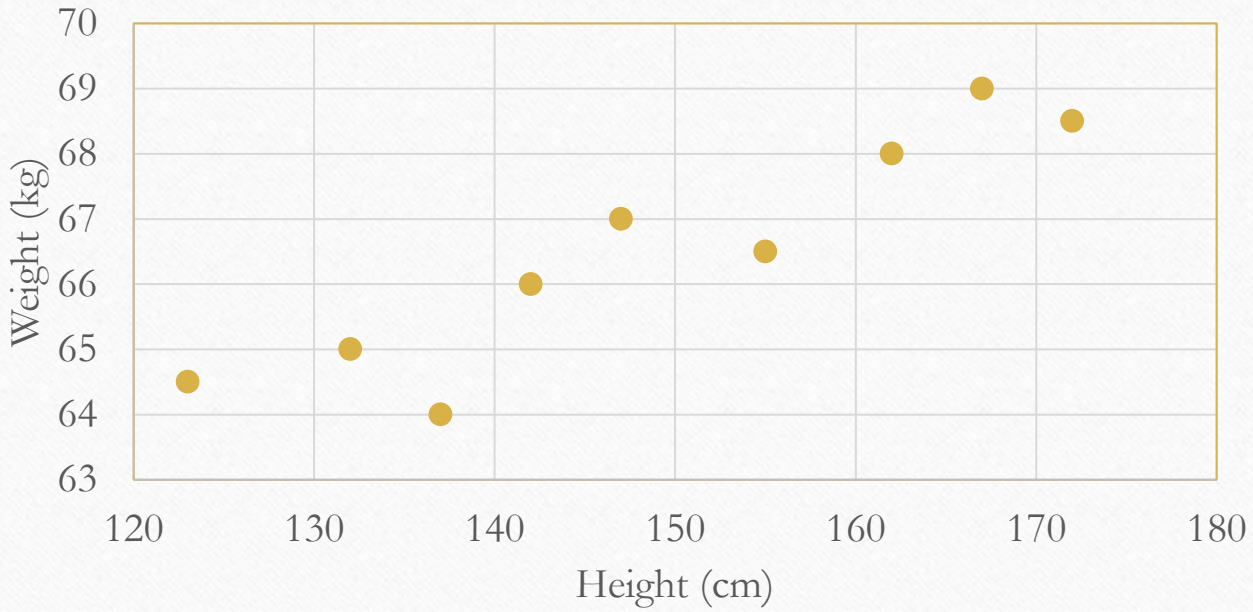
- Correlation in scatter plots
- Positive and negative relationship

Scatter Plots



- A **scatter plot** is a graph that showcases the connection between two data sets. Basically, we use scatter plots to determine if there's a relationship between two variables.

Scatter Plots



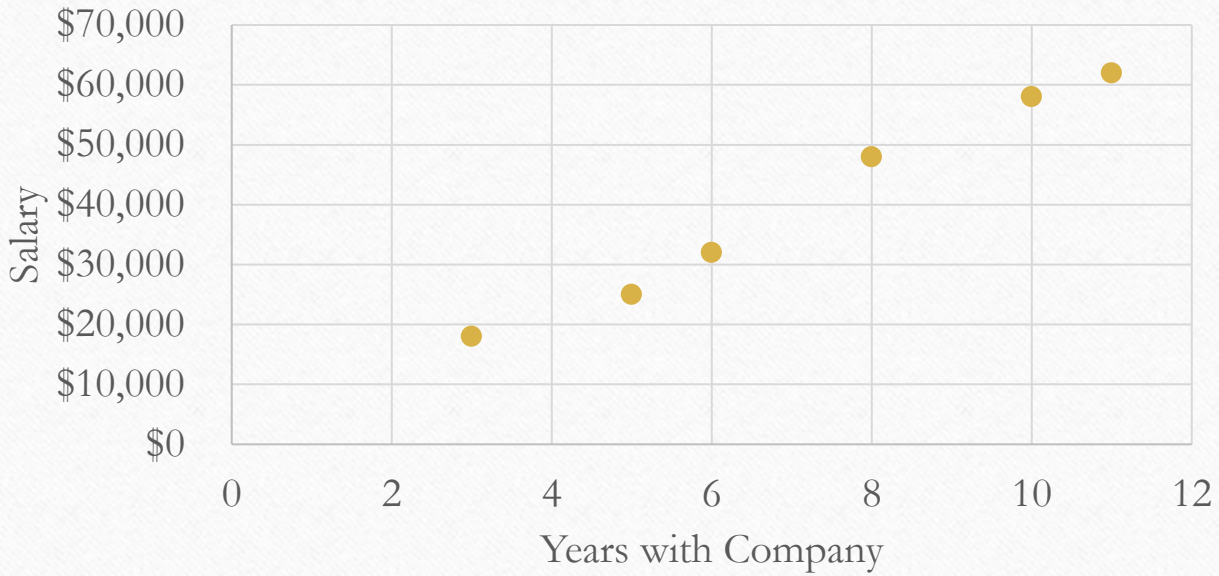
- The horizontal line of the scatter plot is known as the **x-axis**, while the vertical line is referred to as the **y-axis**.
- In the given illustration, height is represented on the x-axis and weight on the y-axis.

Scatter Plots

Employee	Years with Company	Salary
Kaya	5	$25,000
Liam	8	$48,000
Tara	6	$32,000
Ethan	3	$18,000
Aria	11	$62,000
Oscar	10	$58,000

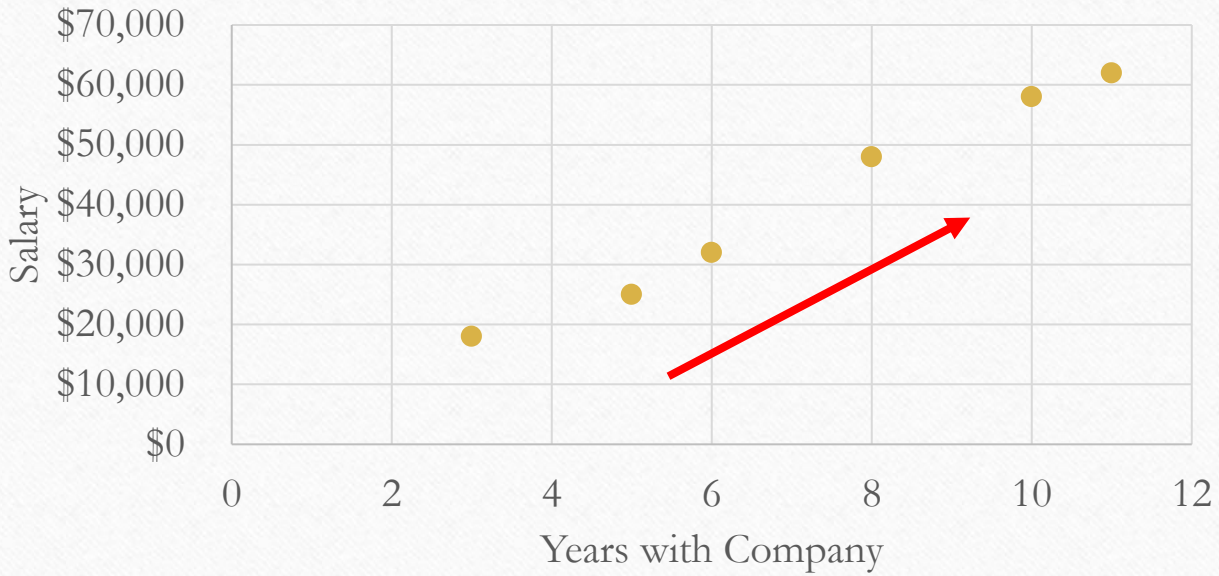
- On a **scatter plot**, variables must be **quantitative**.
  - Both variables must have numeric values.
- Therefore, from the chart mentioned above, only "years with company" and "salary" can be plotted; "employee" cannot be included.

Scatter Plots



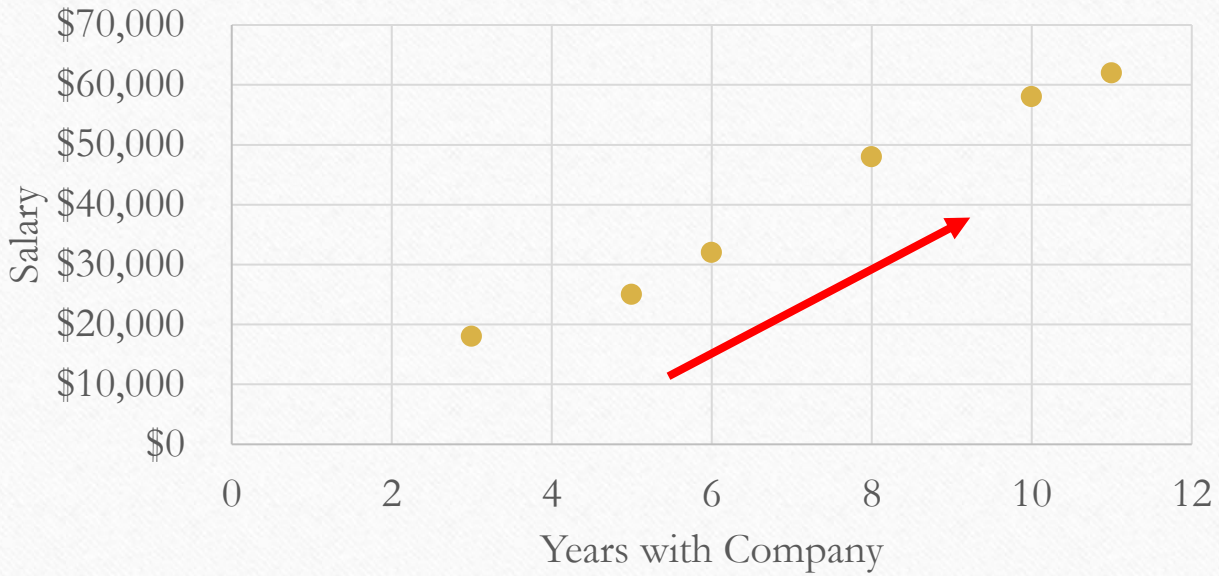
- The above shows the **scatter plot** from the same data in the previous table.
- This **scatter plot** displays "years with company" on the x-axis and "salary" on the y-axis.
- Each point symbolizes an employee at the company. Which one is the data point for Oscar?

Scatter Plots



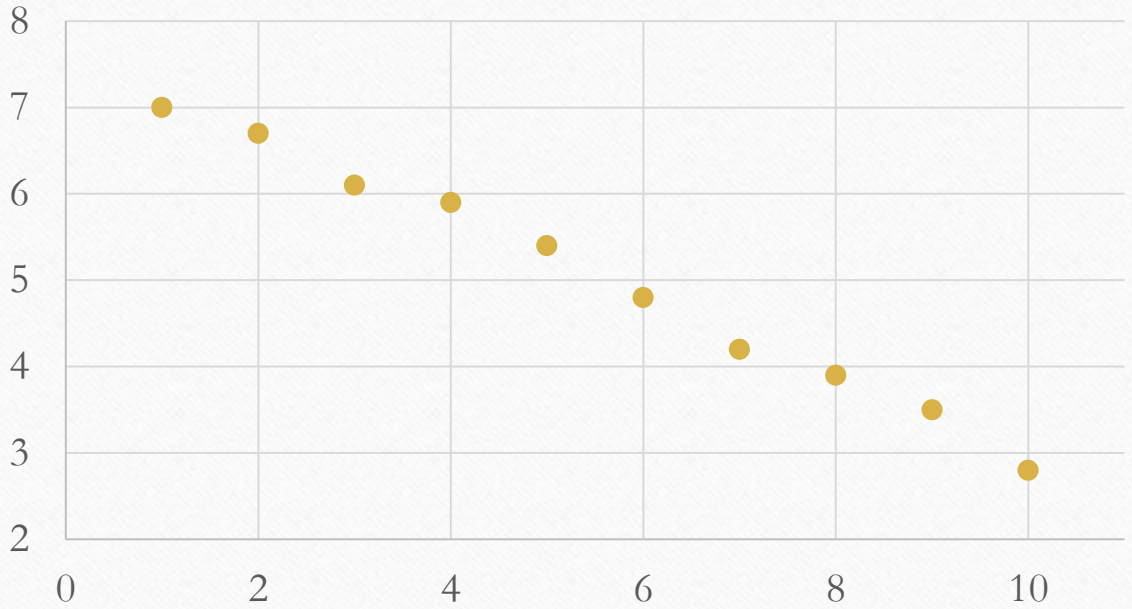
- In this example, the points form a line, indicating a **linear relationship** between the duration of employment at the company and the salary received.
- Even if the points aren't perfectly aligned in a straight line, it's still regarded as a linear relationship.

Scatter Plots



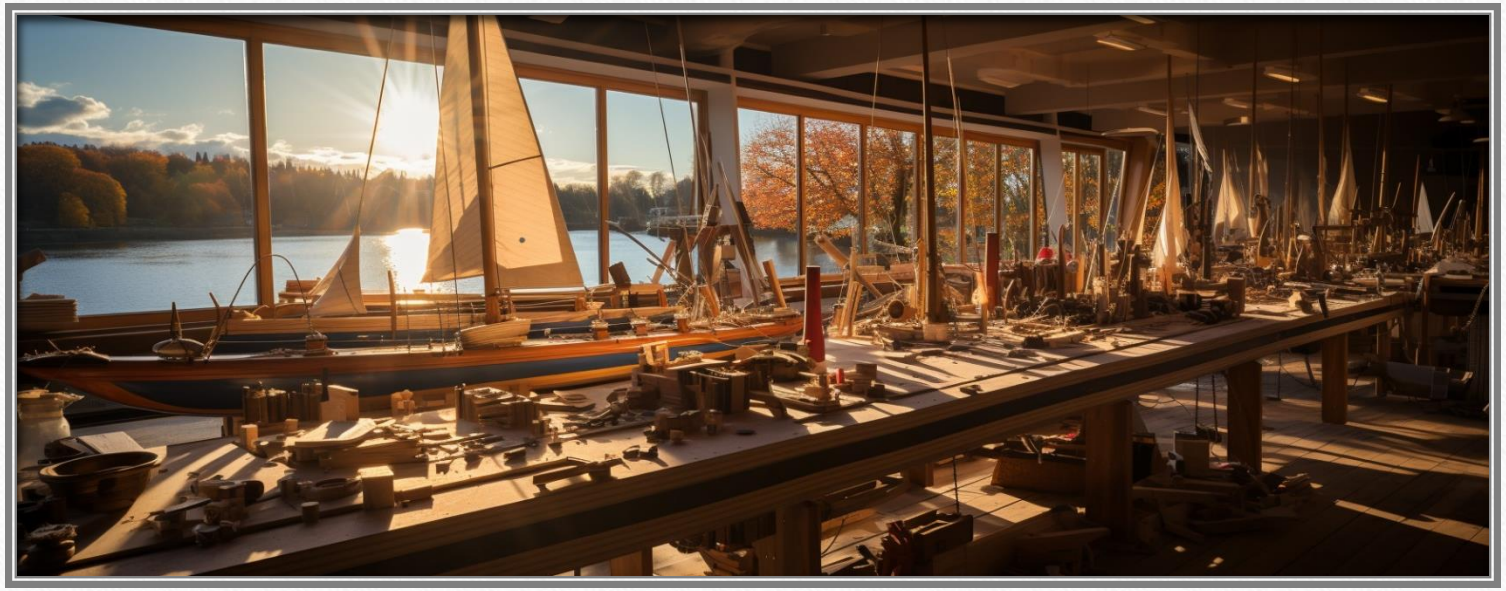
- The line trends upwards from left to right, indicating that as one value rises, the other does too. This is termed a **direct relationship**.
- As mentioned in previous lessons, when two variables increase together, what kind of correlation do they have?

Scatter Plots



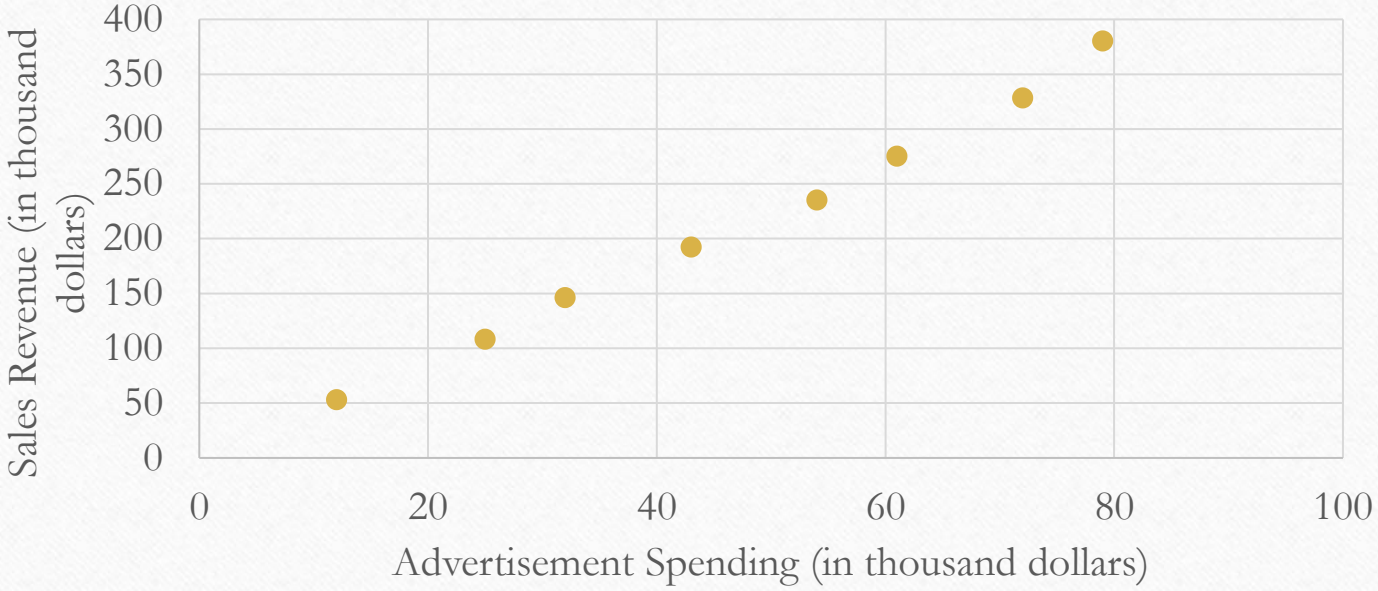
- In contrast, the plot above indicates an **inverse relationship**, as the line tracing the points moves downward to the right. As the x-values rise, the y-values decrease.
- What kind of correlation is it for this one?

Scatter Plots



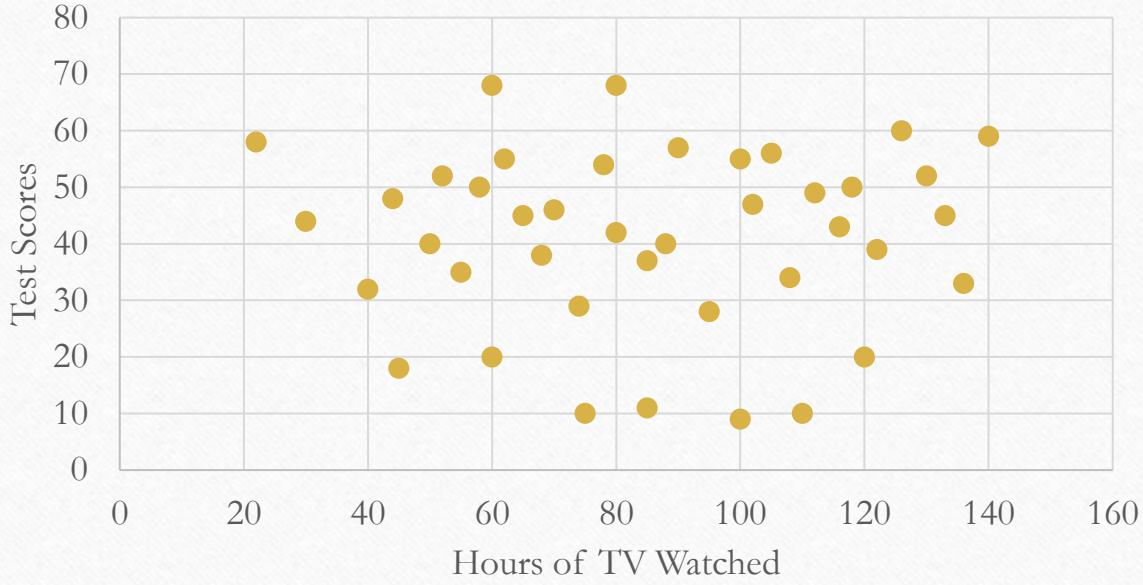
- Lee manages Bay Sailcrafts, a boutique boat-making venture.
- He wishes to investigate the correlation between his advertisement spending and sales revenue using data gathered over the last 8 years.

Scatter Plots



- In the scatter plot for the boat-building company, advertisement spending is on the x-axis, while sales revenue is on the y-axis.
- The data points create a relatively straight line, suggesting a **linear relationship**. The upward movement of the line indicates a **positive** or **direct relationship** between the two variables.

Scatter Plots



- Conversely, this plot showcases what is termed as a **zero correlation**, resembling a random scatter.
- Such diagrams suggest that there's no apparent association between the data sets being compared, indicating no relationship between the x and y variables.

- Question:
  - A scatter plot comparing the hours spent studying and test scores of students shows a line moving up and to the right. What type of correlation does this represent?

  a) Positive correlation

  b) Negative correlation

  c) Zero correlation

  d) Undefined correlation

- Question:
  - You're examining a scatter plot comparing monthly rainfall and crop yield. The plot shows a strong positive correlation. However, upon further analysis, you realize that during months with festivals, there's increased human activity, which could also influence crop yield. This additional factor represents:

a) A confounding variable

b) An outlier

c) A negative correlation

d) An independent variable

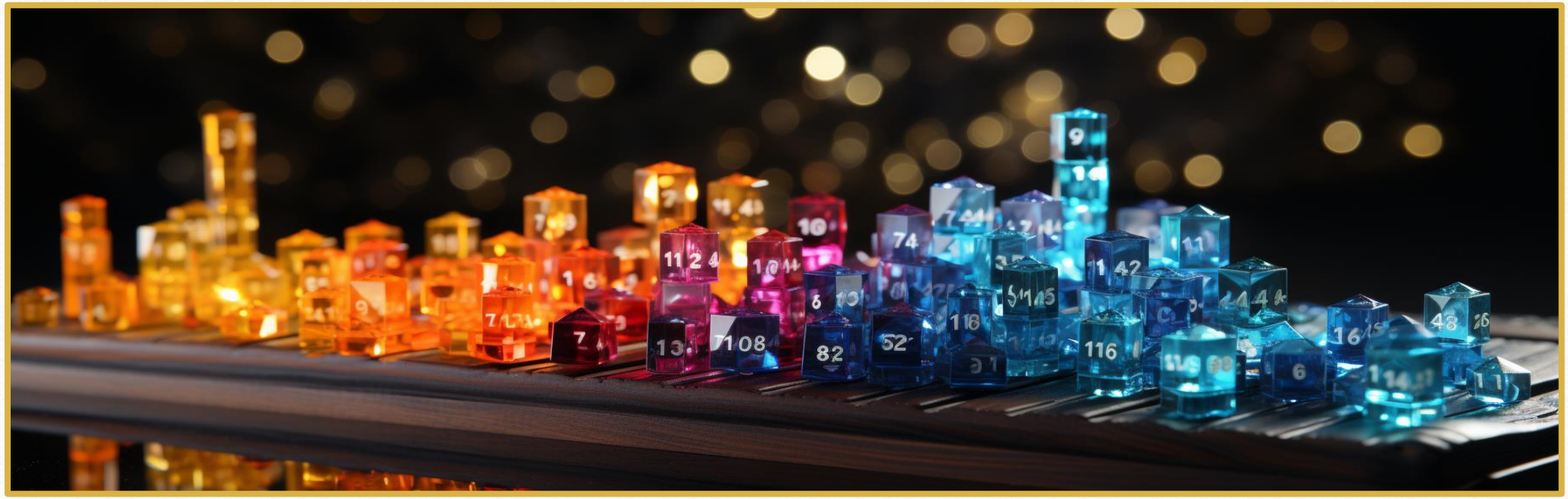
- Question:
  - You're presented with a scatter plot comparing the speed of cars (in km/h) on the x-axis and their fuel efficiency (in km/litre) on the y-axis. If the points on the plot are moving down and to the right, which of the following can be inferred?

a) Faster cars are generally more fuel-efficient.

b) Slower cars are generally more fuel-efficient.

c) Car speed has no relation to its fuel efficiency.

d) All cars have the same fuel efficiency.



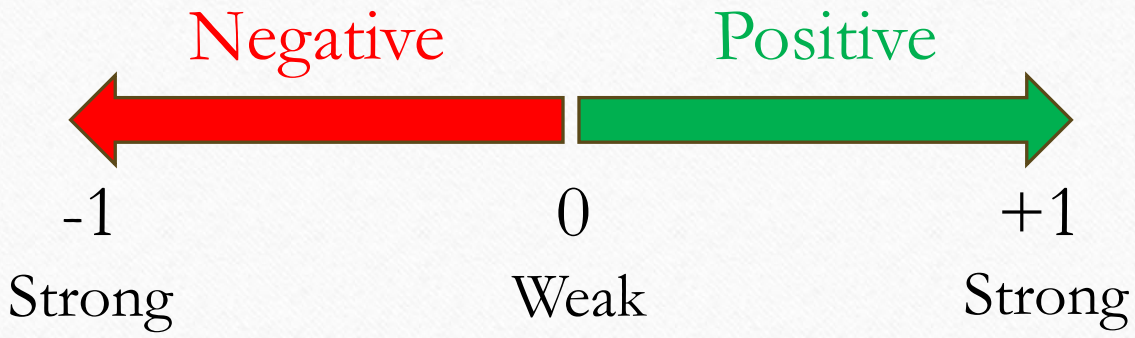
Correlation Coefficient

- Relationship between variables

Correlation Coefficient

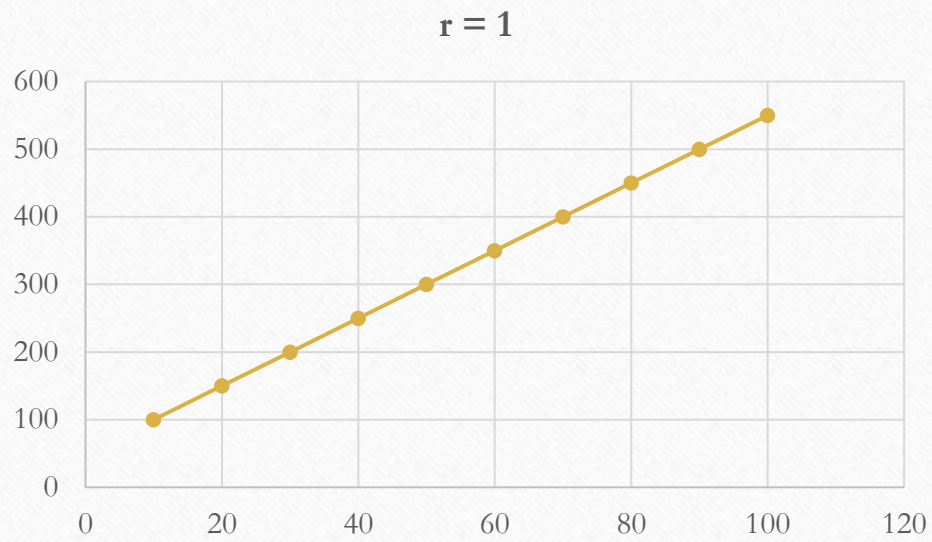
- In this lesson, we'll explore how to assess the relationship between two variables by computing the **correlation coefficient**, often represented by *r*.
- This coefficient indicates the strength and direction of the association between the two variables.

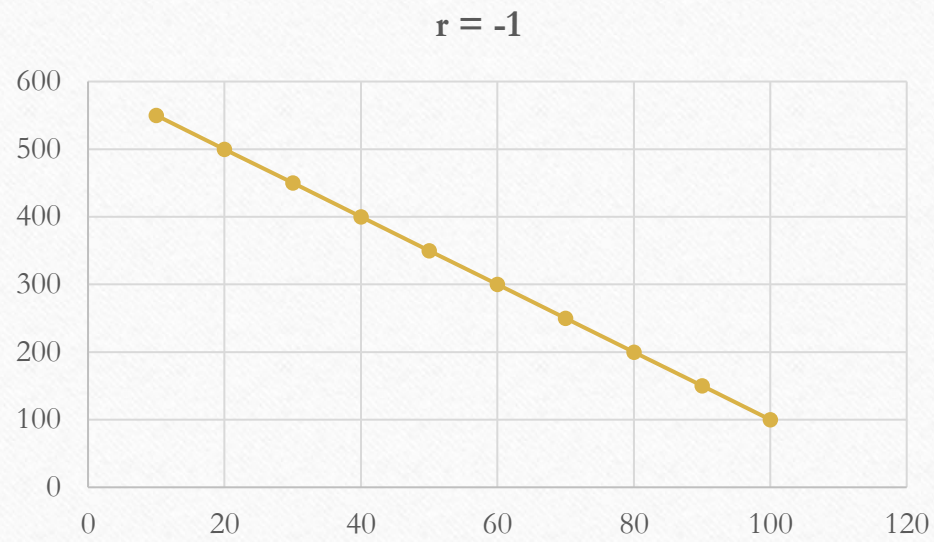
Correlation Coefficient



- *r* always falls between -1 and +1.
- The closer *r* is to either -1 or +1, the higher the strength of the correlation.
  - Therefore, both -1 and +1 signify the strongest correlations.

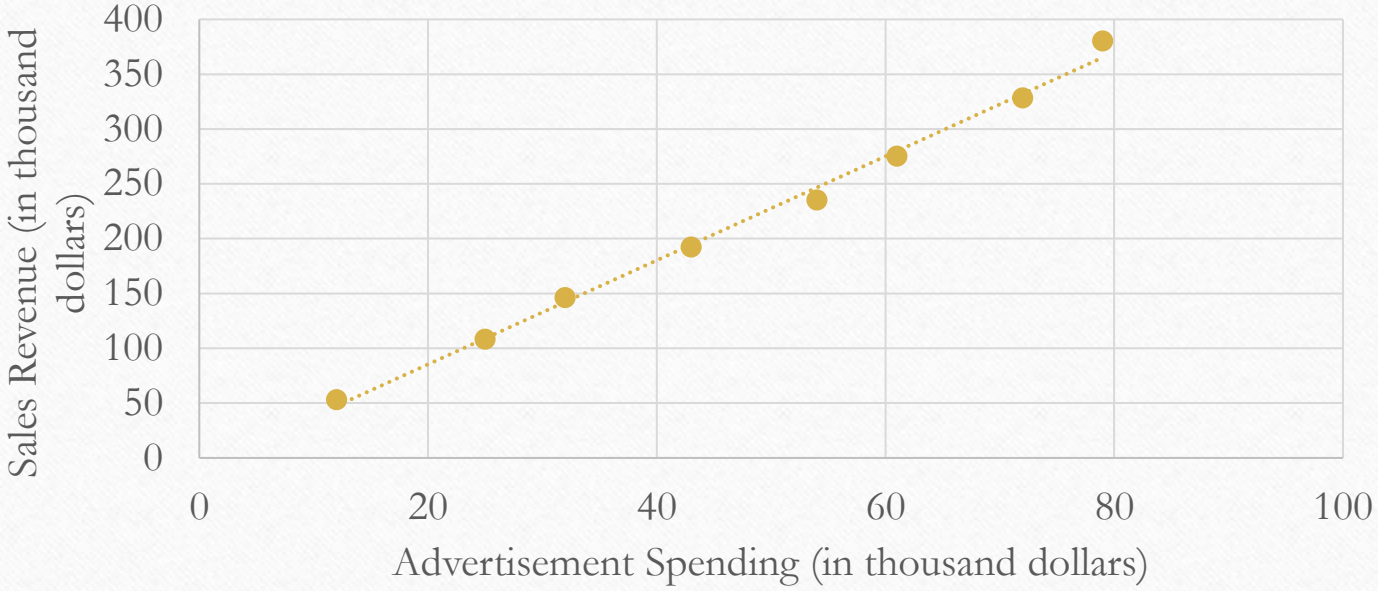
Correlation Coefficient





- Points that align perfectly on an upward diagonal line indicate a correlation of +1.0.
- Those on a downward diagonal line represent a correlation of -1.0.

Correlation Coefficient



- If the alignment is anything other than a perfect line, the correlation coefficient will fall between -1 and +1.
- Above graph is from Bay Sailcrafts's scatter plot. Since it's nearly perfect (but not), the *r* value is 0.9974 (very close to 1 but not 1).

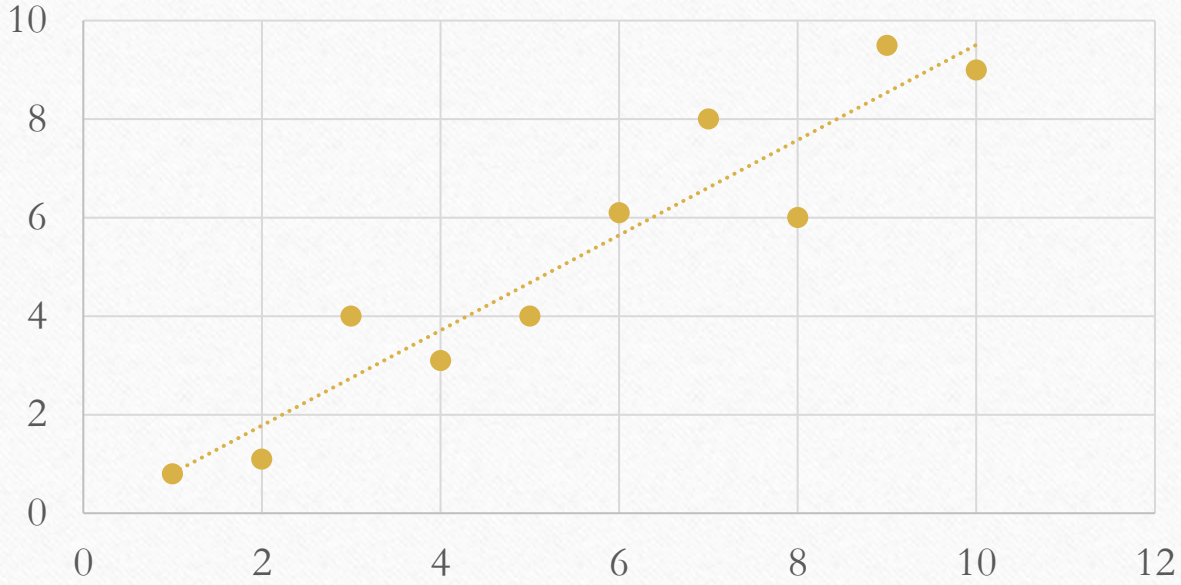
- If you're wondering how to compute the (Pearson) correlation coefficient:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- Where:
  - x_i and y_i are the individual data points.
  - $\bar{x}$ and $\bar{y}$ are the means of the x-values and y-values, respectively.

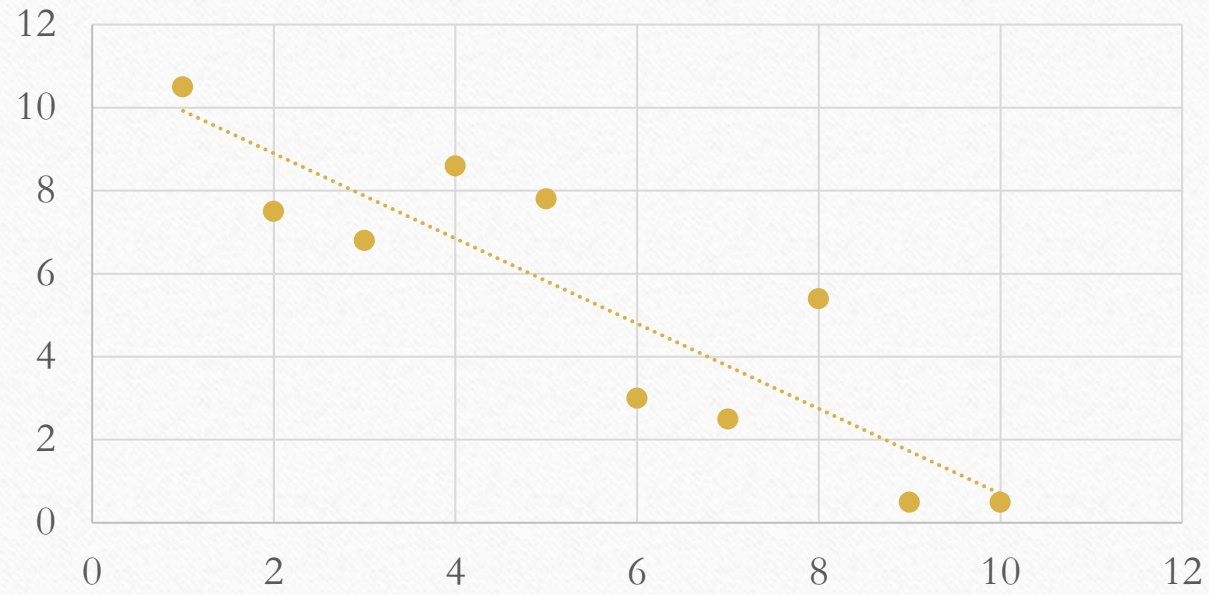
- Anyway, most calculators or spreadsheet tools can handle the computation for you. Let's focus on understanding *r* rather than manually calculating it.

Correlation Coefficient



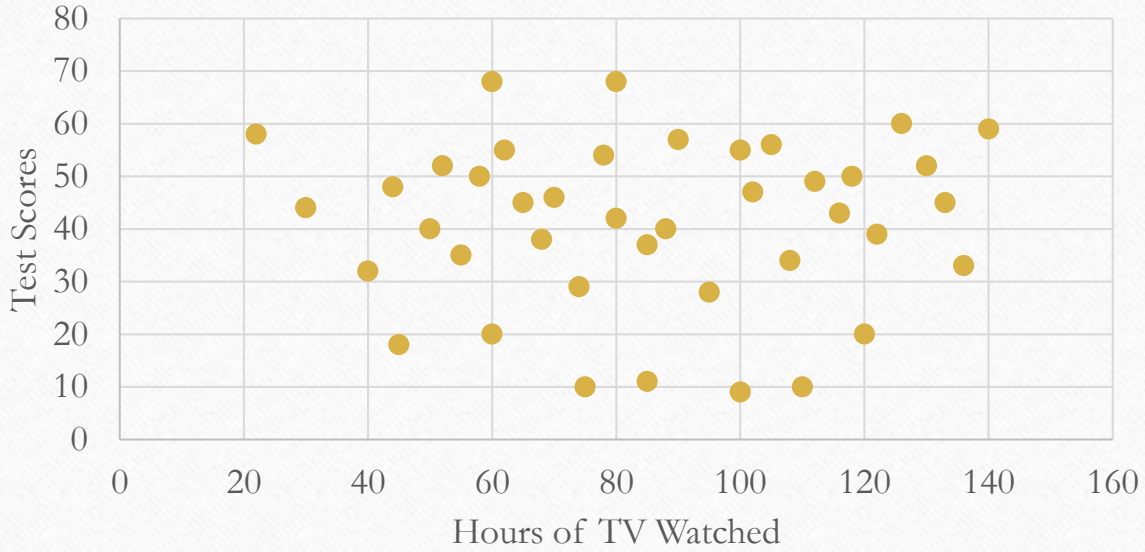
- A positive *r* indicates a **direct relationship**, with the scatter plot displaying an upward trend from left to right.
- $r = 0.948$ for the above graph.

Correlation Coefficient



- A negative *r* signifies an **inverse relationship**, with the scatter plot showing a downward trend from left to right.
- $r = -0.886$ for the above graph.

Correlation Coefficient



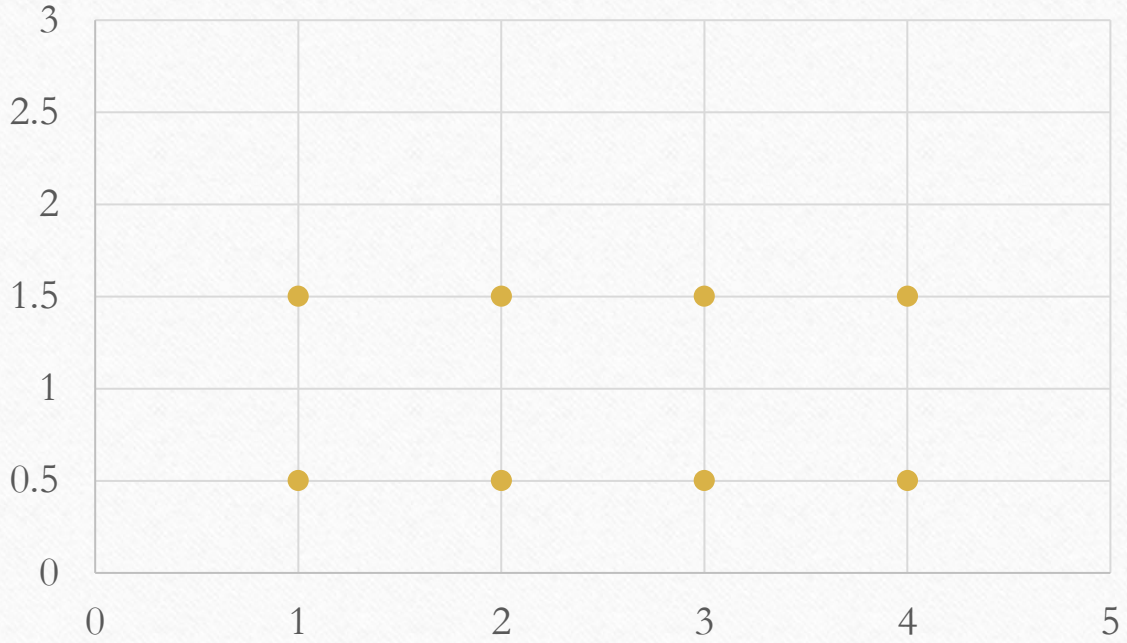
- For variables that have little or no correlation, r is almost 0.
- $r = 0.00013$ for the above graph.

- So what values of *r* indicate a strong correlation? And at what point can *r* be considered moderate?
- Some statisticians might still engage in fights at conferences debating this.

- In general, a correlation coefficient greater than 0.8 (or smaller than -0.8) indicates a strong correlation.

Range of *r*	**Description of Correlation**
$0.80 \leq r \leq 1.00$	Strong correlation
$0.50 < r \leq 0.80$	Moderately strong correlation
$0.30 < r \leq 0.50$	Moderately weak correlation
$0.00 \leq r \leq 0.30$	Weak correlation

Correlation Coefficient



- Observe this set of data that will give a zero-sloped (horizontal) line. What do you estimate its *r* value to be?

A. 1
B. 0
C. 0.5
D. 0.2
E. 0.8

- Question:
  - Which of the following statements best describes a correlation coefficient of $r=-0.9$?

a) Strong positive correlation

b) Moderate negative correlation

c) No correlation

d) Strong negative correlation

- Question:
  - A researcher finds that as hours of study increase, grades on a particular exam decrease. This type of relationship is:

a) Positive correlation

b) Negative correlation

c) Zero correlation

d) Perfect correlation

- Question:
  - If two sets of data have an *r* value of 0.75, this means:

a) 75% of the data is similar.

b) There is a 75% chance of prediction accuracy.

c) 75% of the variation in one variable is explained by the other variable.

d) There is a moderately strong positive correlation between the two sets of data.

- Question:
  - Upon examining a scatter plot, you see no evident pattern, and the points are widely dispersed. You calculate the correlation coefficient and get a value very close to 0. What can you infer?

a) There is a strong positive correlation.

b) There is a strong negative correlation.

c) There is a very weak or no correlation between the two variables.

d) The variables are inversely proportional.

The End