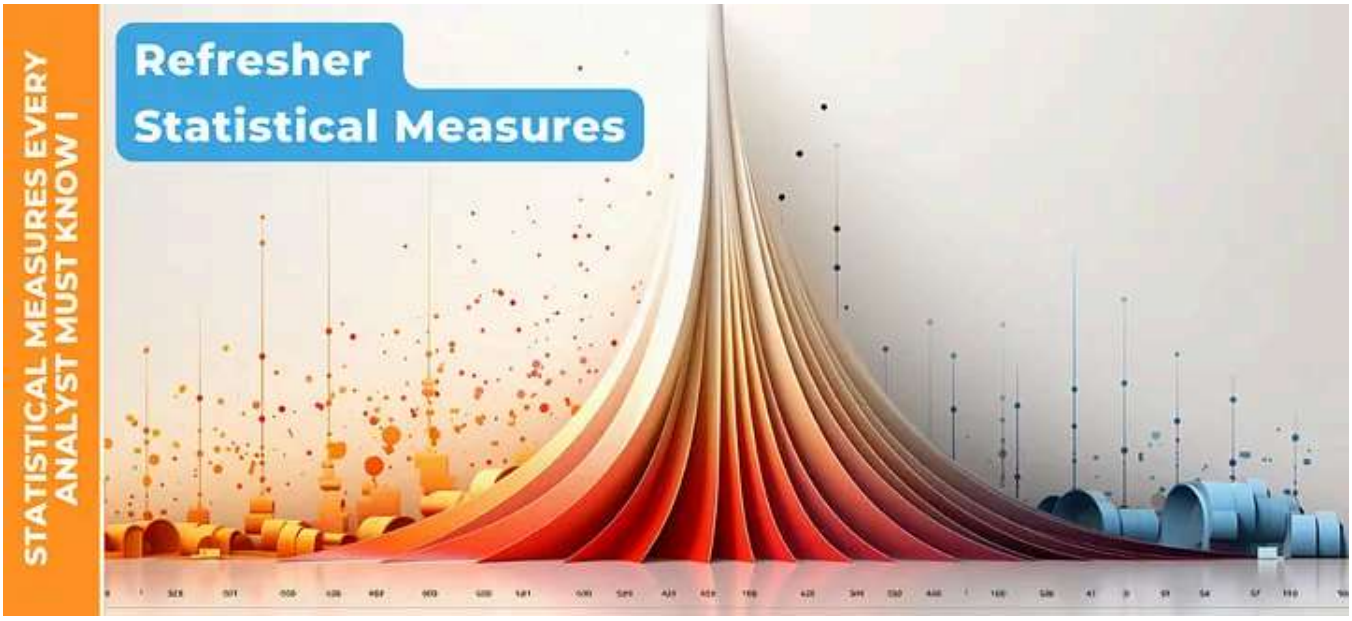
Prof. Frenzel · Follow
11 min read · Feb 5, 2024

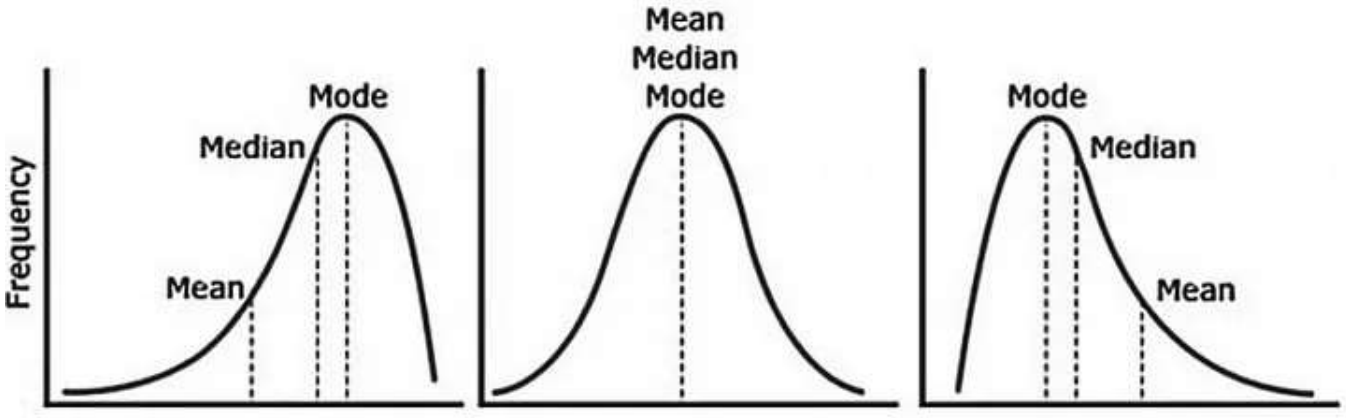


Statistical Measures Every Analyst Must Know — Part1

Have you ever considered how decisions are formulated under uncertainty, or how to condense large volumes of data into comprehensible formats? Imagine being a healthcare analyst tasked with spotting trends in patient recovery times across different hospitals. With hundreds of thousands of data points to work with, the challenge is not only to analyze this data but also to interpret it in a way that aids decision-making. Statistical measures are key when making comparisons, spotting trends, and forecasting future events. Without them, the data would remain a mere collection of numbers without significance.

Measures of Central Tendency

Grasping the measures of central tendency is essential for anyone engaging in data analysis. These measures — mean, median, and mode — identify a central point that data values revolve around, providing a concise summary of the dataset.



Measures of Central Tendency

📌**Mean** is the arithmetic average of a set of numbers, calculated by summing all the values and then dividing by the count of values.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where xi represents each value in the dataset, and n is the number of values

For example, in a dataset containing the values [2, 3, 5, 7, 11], the mean would be (2 + 3 + 5 + 7 + 11) / 5 = 5.6. The mean provides a useful overall measure but is sensitive to extreme values or outliers.

📌**Median** is the middle value in a dataset when the values are arranged in ascending or descending order. If the dataset has an even number of observations, the median is the average of the two middle numbers.

$$\text{Median} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

where x represents each value in the dataset arranged in ascending order and n is the number of values

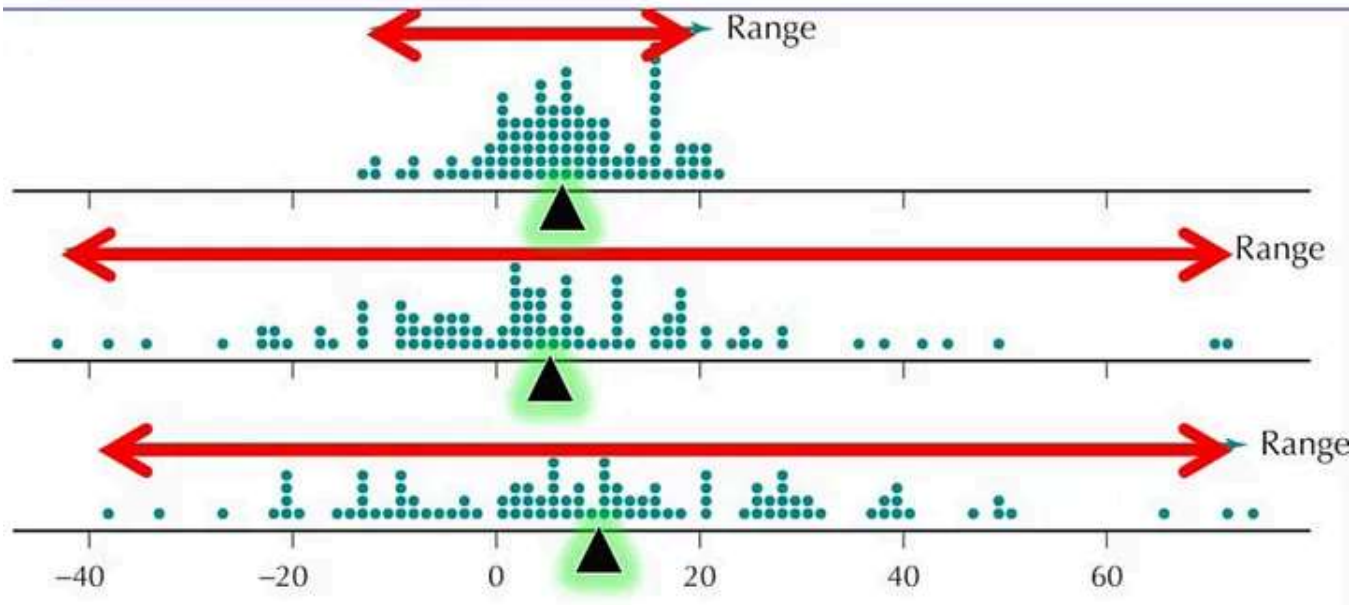
For the sorted dataset [2, 3, 5, 7, 11], the median is 5, directly indicating the center of the dataset. If we had an additional value, say 13, making the dataset [2, 3, 5, 7, 11, 13], the median would be (5 + 7) / 2 = 6. The median is particularly useful for datasets with outliers, as it is not affected by them.

📌**Mode** refers to the most frequently occurring value(s) in a dataset. It's the value that appears the most number of times. For a dataset [1, 2, 2, 3, 4], the mode is 2

since it appears more frequently than the other numbers. In datasets with no repeating values, there might be no mode, and in datasets with multiple values repeating the same number of times, there can be multiple modes. The mode is especially valuable for categorical data where mean and median cannot be defined.

Measures of Variability

While measures of central tendency give us a sense of the dataset's center, measures of variability tell us about the spread of the data. Understanding the range, variance, and standard deviation is crucial for comprehensively analyzing the behavior of a dataset.



Measures of Variability

📌**Range** is the simplest measure of variability, calculated as the difference between the highest and lowest values in the dataset. The formula is represented as

$$\text{Range} = \text{Max}(x_i) - \text{Min}(x_i)$$

where xi represents each value in the dataset

For instance, in a dataset with values [1,3,4,8,10], the range would be 10−1=9. The range gives a quick sense of the overall spread of the data but does not account for how the values are distributed between the extremes.

📌**Variance** provides a more detailed measure of spread by calculating the average squared deviation from the mean. The formula for the sample variance is

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

where xi are the values in the dataset, x̄ is the mean of the dataset, and n is the number of values in the sample

This formula captures how each data point differs from the mean, offering insight into data dispersion. For the same dataset [1,3,4,8,10], with a mean of 5.2, the variance would involve calculating each value's deviation from 5.2, squaring those deviations, summing them, and then dividing by n−1 for a sample variance.

📌**Standard Deviation** is the square root of the variance and provides a measure of dispersion in the same units as the data. Its formula is

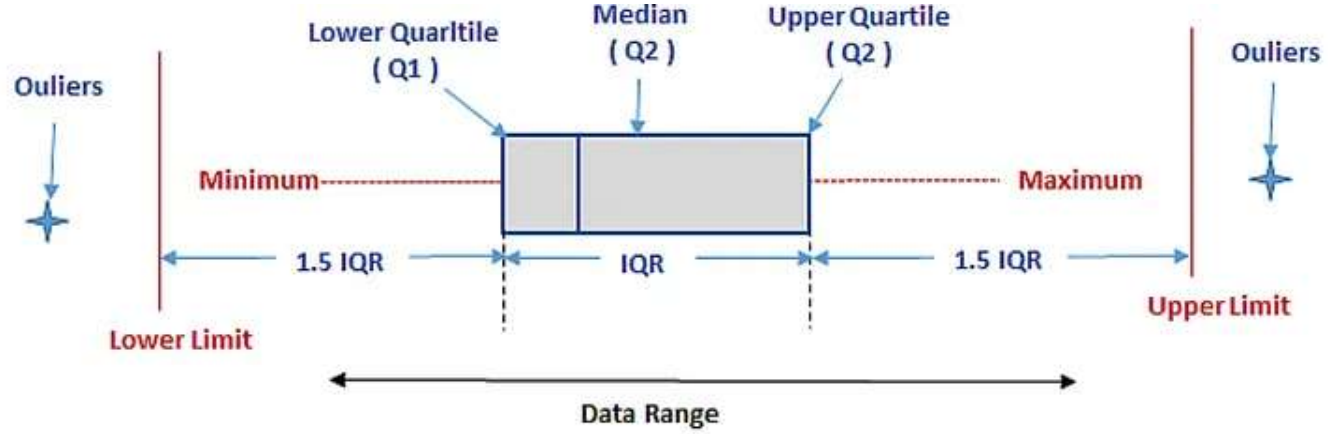
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where xi are the values in the dataset, x̄ is the mean of the dataset, and n is the number of values in the sample

Standard deviation is particularly useful because it gives a sense of how spread out the data points are from the mean. A low standard deviation means data points are clustered close to the mean, while a high standard deviation indicates data points are spread out over a wider range of values.

Quartiles and Percentiles

Quartiles and percentiles are powerful statistical tools that divide a dataset into distinct parts, providing a deeper understanding of its distribution. These measures are essential for describing the spread and shape of the data beyond the average or middle value.



Quartiles and Interquartile Range

📌**Quartiles** split the data into four equal parts. The first quartile (Q1) is the median of the lower half of the dataset, not including the median if the number of observations is odd. The second quartile (Q2) is essentially the median of the dataset, and the third quartile (Q3) is the median of the upper half of the dataset. The Interquartile Range (IQR) measures the middle 50% of the data and is crucial for identifying outliers.

$$Q_1 = x_{\left(\frac{n+1}{4}\right)}, \quad Q_3 = x_{\left(\frac{3(n+1)}{4}\right)}, \quad \text{IQR} = Q_3 - Q_1$$

Quartiles

For example, in a dataset sorted in ascending order [2,3,4,8,9,12,14], Q1 is 3, Q2 (the median) is 8, and Q3 is 12. The IQR is 12−3=9, indicating the range within which the central half of the data points lie.

📌**Percentiles** indicate the position of a value in relation to the entire dataset, showing what percentage of the data lies below a given value. The 25th, 50th, and 75th percentiles correspond to the first, second, and third quartiles, respectively.

$$P_k = x_{\left(\frac{k(n+1)}{100}\right)}$$

Percentiles

Calculating a specific percentile, say the 90th percentile, involves sorting the data and finding the value below which 90% of the data points fall. Percentiles are particularly useful in performance and capacity analysis, such as understanding test scores or income distribution, where distinguishing between different levels of data distribution is key. By providing detailed insights into the distribution of data,

quartiles and percentiles help in identifying trends, understanding variability, and detecting outliers, thereby enabling more nuanced data analysis and decision-making processes.

Z-Scores and Standardization

Z-scores, a cornerstone of statistical analysis, offer a method for comparing data points from different distributions or scales by standardizing them. A Z-score represents how many standard deviations a given data point is from the mean of its dataset.

The formula for calculating a Z-score is:

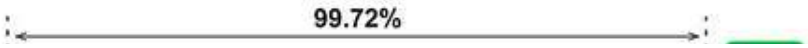
$$z = \frac{x - \mu}{\sigma}$$

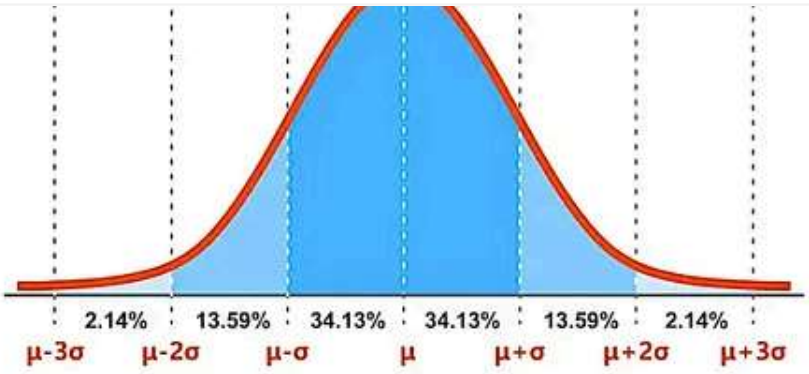
where x is the data point, μ is the mean of the dataset, and σ is the standard deviation of the dataset

This formula transforms the data into a distribution with a mean of 0 and a standard deviation of 1, known as the standard normal distribution.

For instance, if a student scored 85 on a test with a mean score of 75 and a standard deviation of 5, the Z-score would be 2. This means the student's score is 2 standard deviations above the mean score.

Z-scores are particularly useful for identifying outliers, comparing scores from different distributions, and conducting hypothesis testing. **When utilizing Z-scores for data cleaning, particularly in identifying outliers, it's necessary to assume that the data follows a normal (Gaussian) distribution.** This assumption underlies the interpretation and effectiveness of Z-scores since their values correspond to the number of standard deviations a data point is from the mean, which directly relates to probabilities in a normal distribution.





The Standard Normal Distribution and the Z-Score

- Z = 1 signifies that a data point is one standard deviation away from the mean. In a normal distribution, approximately 68% of the data lies within one standard deviation of the mean (between Z = -1 and Z = 1), indicating that a Z-score of ±1 is within the range of common variation.
- Z = 2 indicates that a data point is two standard deviations away from the mean. About 95% of the data falls within two standard deviations (between Z = -2 and Z = 2), making a Z-score of ±2 less common and potentially an outlier.
- Z = 3 means that a data point is three standard deviations away from the mean. In practice, nearly 99.7% of the data is within three standard deviations (between Z = -3 and Z = 3). A Z-score of ±3 is widely regarded as a threshold for identifying outliers because it signifies that the data point is in the extreme 0.3% of the distribution, making it highly unusual under the assumption of normality.

The use of Z = 3 as a common threshold for outlier detection is based on this statistical property of the normal distribution. It helps in minimizing the risk of mistakenly identifying normal variations within the data as outliers, thus ensuring that only the most extreme values, which could indicate errors or significant anomalies, are flagged for further investigation or removal.

This approach, however, requires careful consideration of the data's distribution. If the data is not normally distributed, the interpretation of Z-scores may not be as straightforward, and alternative methods or transformations might be necessary to accurately identify outliers.

Correlation and Causation

In data analysis, discerning the relationship between variables is fundamental. The concepts of correlation and causation are central to understanding these relationships. Correlation measures the degree to which two variables move in relation to each other, whereas causation indicates that a change in one variable is responsible for a change in another.

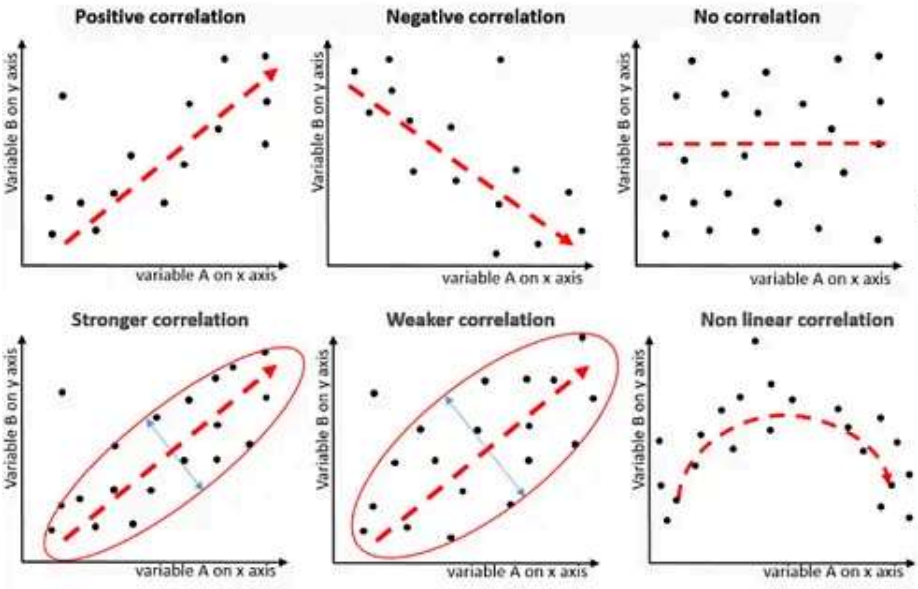
📌**Correlation** is quantified by correlation coefficients, with Pearson and Spearman being the most commonly used. The Pearson correlation coefficient measures the linear relationship between two variables, providing a value between -1 and 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 signifies no linear relationship.

The formula for the Pearson correlation coefficient is:

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

where n is the number of data points, xi and yi are the individual data points, and x¯ and y¯ are the means of the x and y data sets, respectively.

When assessing correlation visually, scatter plots are a basic tool that gives an immediate idea of the relationship between two variables. The closeness of the scatter to a line — upward for a positive correlation or downward for a negative one — provides a straightforward understanding of how strong the relationship is. But this method has its limits, especially for non-linear relationships that might take the shape of a curve or a more intricate pattern. These types of relationships, despite potentially being strong, cannot be accurately depicted by a linear model or a correlation matrix.



Grading Standards	Correlation Degree
$\rho = 0$	no correlation
$0 < \|\rho\| \leq 0.19$	very week
$0.20 \leq \|\rho\| \leq 0.39$	weak
$0.40 \leq \|\rho\| \leq 0.59$	moderate
$0.60 \leq \|\rho\| \leq 0.79$	strong
$0.80 \leq \|\rho\| \leq 1.00$	very strong
1.00	monotonic correlation

Analysts should interpret these coefficients by understanding their mathematical implications and by creating a narrative that explains what these figures represent in practical terms. For example, a coefficient of 0.32 could indicate a preliminary link between variables, suggesting a trend but not a definite prediction. The qualifications of the outcomes are especially important when talking to a non-technical audience who is not familiar with the concept of correlation and its limitations.

📌**Spearman correlation** assesses the monotonic relationship between two variables, suitable for ordinal data or when the relationship is not linear. Unlike Pearson, Spearman's rank correlation does not require the assumption of linear relationships and is less sensitive to outliers.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

di is the difference between the ranks of corresponding variables xi and yi, and n is the number of data points.

Imagine a study comparing the amount of time students spend on social media (X) with their level of happiness (Y). The data for 5 students is as follows:

Student	Hours on Social Media (X)	Happiness Score (Y)
A	1	2
B	2	3
C	3	1
D	4	4
E	5	5

*First, we **rank** both variables:*

Student	X	Rank X	Y	Rank Y
A	1	1	2	2
B	2	2	3	3
C	3	3	1	1
D	4	4	4	4
E	5	5	5	5

Next, calculate the difference in ranks (di) and square these differences (d^2t):

Student	d_i	d_i^2
A	0	0
B	0	0
C	2	4
D	0	0
E	0	0

The Spearman correlation coefficient formula and calculation:

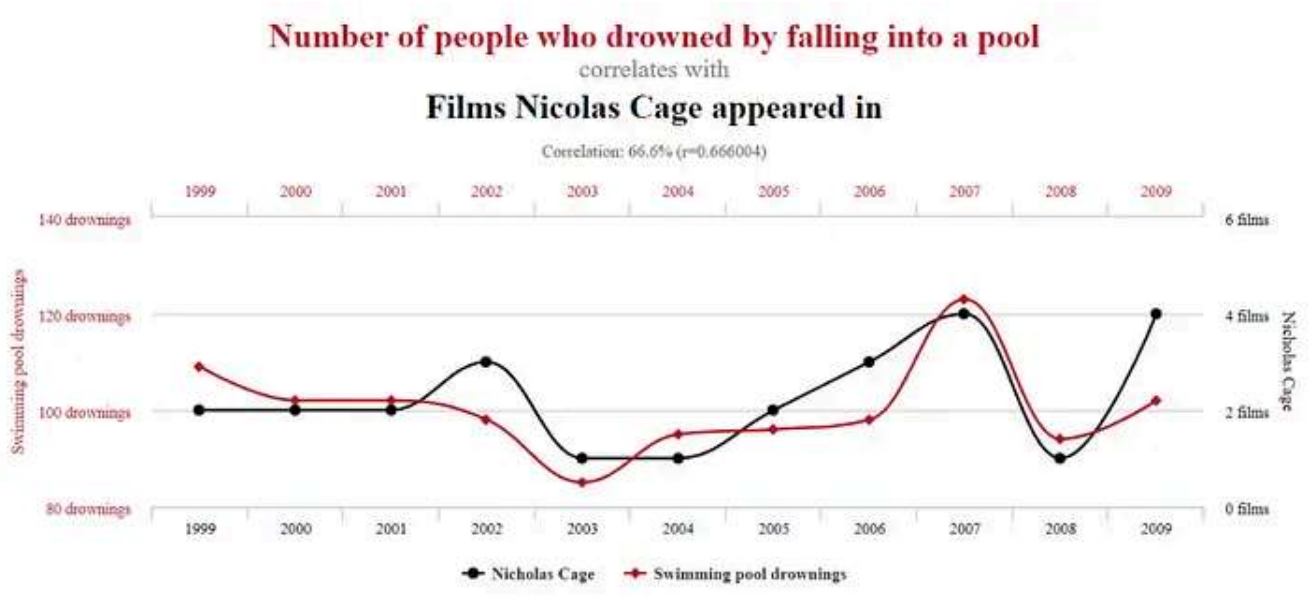
$$\rho = 1 - \frac{6 \times 4}{5(5^2 - 1)} = 1 - \frac{24}{120} = 0.8$$

The Spearman correlation coefficient of 0.8 indicates a strong positive monotonic relationship between time spent on social media and happiness score.

Why Not Pearson? *In this scenario, using Pearson correlation could be misleading if the relationship between social media usage and happiness is not linear. For example, the happiness score does not increase at a constant rate with more hours spent on social media.* ***Pearson correlation measures linear relationships and might not capture the true nature of the association as effectively as Spearman****, which can handle monotonic relationships even if they are not linear. This makes Spearman more suitable for this example, where the underlying relationship may not strictly follow a straight line.*

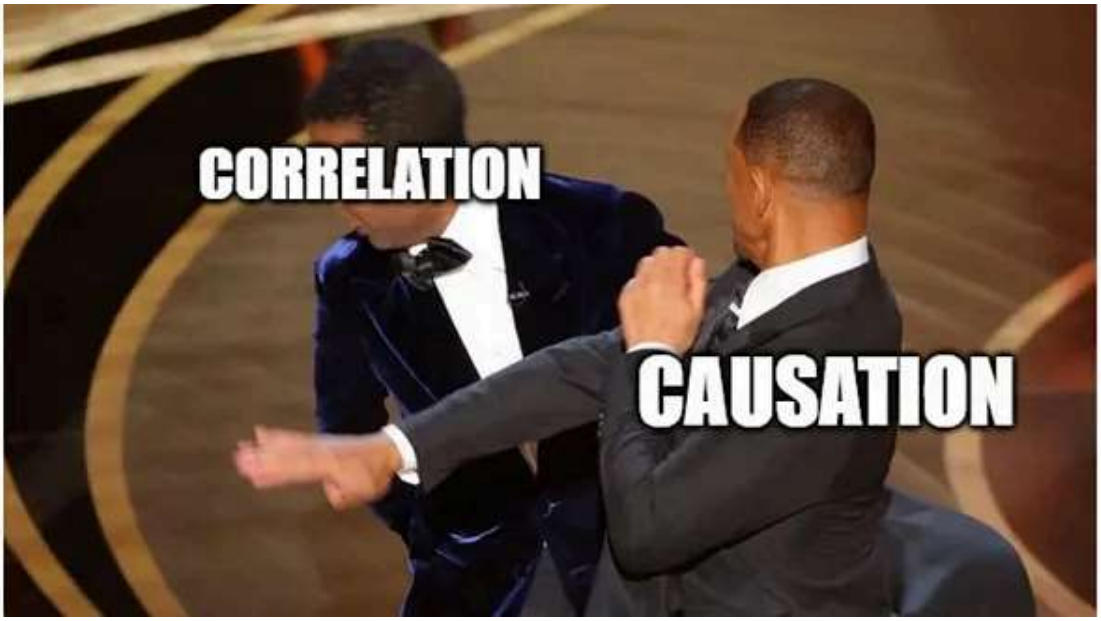
⚠ Correlation does not imply Causation!

The statement "Correlation does not imply causation" highlights a common misunderstanding in data analysis and statistical interpretation. This confusion stems from the fact that both correlation and causation indicate relationships between variables, but they do so in fundamentally different ways. Correlation identifies when two variables move together in some manner, without specifying whether one variable's movement causes the other's. **Causation, on the other hand, indicates a direct relationship where changes in one variable directly result in changes in another.**



Nicolas Cage certainly does not produce good movies lately, but drowning is a little extreme, don't you think?

The error often happens because observing a correlation can intuitively lead to the assumption that one variable may influence the other. **This jump to causation fails to consider other factors, such as third variables that affect both correlated variables or the possibility that the correlation is merely coincidental.**



Without thorough experimental control to rule out other explanations, claiming causation based solely on correlation can result in incorrect decisions and theories. Understanding this distinction is important for accurate data interpretation and the appropriate use of statistical findings.

In the field of economics, a classic example is the correlation between GDP growth and education levels. While there is a strong correlation, implying that countries with higher education levels also have higher GDP growth, it is not solely education that causes GDP growth. Numerous other factors, including political stability, infrastructure, and access to natural resources, also play significant roles.

In marketing, a firm may observe a correlation between social media ad spend and increased sales. While tempting to attribute the rise in sales directly to the ad spend, this correlation does not confirm causation. Sales could also be influenced by seasonality, changes in consumer preferences, or a concurrent email marketing campaign. To investigate causation, analysts might use A/B testing to compare the sales performance between a control group and a group exposed to the ad campaign, or multivariate regression analysis to isolate the effect of ad spend while controlling for other variables.

In sports, a team might notice a connection between practice hours and victories. But there is a good chance that the number of practice hours alone does not determine the increase in victories. Factors such as coaching quality, player health, and the intensity of competition also influence outcomes. Analysts interested in identifying the true causal factors might use tools like predictive modeling to predict victories based on various inputs, or engage in time series analysis to investigate the relationship over different periods, taking into account other influential factors.