ASSIGNMENT-1(Statistics)

Qus 1.Explain the different types of data (qualitative and quantitative) and provide examples of each. Discuss nominal, ordinal, interval, and ratio scales.

Ans: Data can be broadly classified into two basic types: **qualitative** and **quantitative**. Each category serves diverse goals in research and data analysis.

Qualitative Data

Qualitative data is non-numerical and describes qualities or characteristics. It can provide insights into people's thoughts, feelings, and behaviors. Qualitative data is often collected through interviews, open-ended surveys, observations, and focus groups.

Quantitative Data

Quantitative data is numerical and can be measured and analyzed statistically. It provides information that can be quantified and is often collected through surveys with closed-ended questions, experiments, and structured observations.

Scales of Measurement

Quantitative data can be further categorized based on the scale of measurement: nominal, ordinal, interval, and ratio.

1. **Nominal Scale:** Categorizes data without a specific order. Each category is mutually exclusive.

Examples:

- Types of fruit: apples, oranges, bananas.
- Gender: male, female, non-binary.

2. **Ordinal Scale:**Categorizes data with a defined order but without consistent intervals between categories. The order matters, but the differences are not measurable.

- **Examples:**
    - Customer satisfaction ratings: unsatisfied, neutral, satisfied.
    - Educational levels: high school, bachelor's degree, master's degree.

3. **Interval Scale:** Similar to ordinal data but with equal intervals between values. However, there is no true zero point, which means ratios are not meaningful.

- **Examples:**
    - Temperature in Celsius or Fahrenheit (20°C is not twice as hot as 10°C).

4. **Ratio Scale**

- **Definition:** Like the interval scale, but it has a true zero point, allowing for meaningful ratios. This makes it possible to say that one value is so many times larger than another.
- **Examples:**
    - Weight: 0 kg means no weight; 10 kg is twice as heavy as 5 kg.
    - Height: 0 cm indicates no height; 180 cm is 1.5 times taller than 120 c

Qus 2. What are the measures of central tendency, and when should you use each? Discuss the mean, median, and mode with examples and situations where each is appropriate.

Ans: A single value that represents the distribution's centre is used to summarise a collection of data points is measures of central tendency. The three main metrics—**mean, median,** and **mode**—each appropriate in a particular context.

Mean: The average of all data points, calculated by summing them and dividing by the number of values.

- **Use:** Best for normally distributed data without outliers.
- **Example:** For test scores of 80, 85, and 90, the mean is (80 + 85 + 90) / 3 = 85.
- **Formula for calculating:**

Population Mean

The formula for the population mean (μ) is:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where:

- μ = population mean
- N = total number of observations in the population
- x_i= each individual observation in the population

Sample Mean

The formula for the sample mean (x̅)is:

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

where:

- x̅= sample mean
- n = total number of observations in the sample
- x_i = each individual observation in the sample

Median: The middle value when data points are ordered from least to greatest. If there's an even number of values, the median is the average of the two middle numbers.

- **Use:** Effective for skewed distributions or when outliers are present. Useful in income data where a few high earners might skew the mean
- **Example:** For the data set 3, 7, 8, 12, 13, the median is 8. For 3, 7, 8, 12, 13, 14, the median is (8 + 12) / 2 = 10.

Mode: The value that appears most frequently in a data set. A set can have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode.

- **Use:** Helpful for categorical data or to identify the most common item. Useful in market research to identify the most popular product.
- **Example:** In the set 2, 3, 4, 4, 5, the mode is 4.

Therefore we choose the mean for normally distributed data, the median for skewed data or outliers, and the mode for identifying the most frequent item

Qus 3. Explain the concept of dispersion. How do variance and standard deviation measure the spread of data?

Ans: Dispersion refers to the extent to which data points in a dataset spread out or vary from the central tendency (mean, median, or mode). Understanding dispersion is crucial because it provides insight into the variability and reliability of the data. A dataset with low dispersion indicates that the values are close to the central value, while high dispersion suggests that the values are more spread out.

Key Measures of Dispersion

1. **Variance**
    - **Definition:** Variance measures the average squared deviation of each data point from the mean. It quantifies how much the data points differ from the mean on average.
    - **Formula:**
        - For a population:

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

        - For a sample:

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

      where xix_ixi are the individual data points, μ\muμ is the population mean, x¯\bar{x}x¯ is the sample mean, NNN is the population size, and nnn is the sample size.

    - **Interpretation:** A higher variance indicates a greater spread of data points. Since variance is expressed in squared units, it can sometimes be less intuitive to interpret directly.
2. **Standard Deviation**
    - **Definition:** Standard deviation is the square root of the variance. It provides a measure of dispersion in the same units as the original data, making it easier to interpret.
    - **Formula:**
        - For a population:
        - For a sample:
    - **Interpretation:** A larger standard deviation indicates that the data points are more spread out from the mean. A smaller standard deviation suggests that the data points are closer to the mean.

Importance of Variance and Standard Deviation

- **Comparison of Datasets:** These measures allow for the comparison of variability between different datasets. For instance, two classes may have the same average test score (mean), but the class with the higher standard deviation may have a wider range of scores, indicating more variability in student performance.
- **Understanding Risk and Uncertainty:** In fields like finance, standard deviation is often used to measure the risk associated with investment returns. A higher standard deviation in investment returns indicates more volatility.

Qus 4. What is a box plot, and what can it tell you about the distribution of data?

Ans: A box plot, also known as a whisker plot, is a graphical representation of a dataset that summarizes its distribution through five key summary statistics: the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. Box plots are useful for visualizing the spread and skewness of data, as well as identifying potential outliers.

Components of a Box Plot

1. **Box**: The central box represents the interquartile range (IQR), which contains the middle 50% of the data. It extends from the first quartile (Q1) to the third quartile (Q3).
2. **Median Line**: A line inside the box indicates the median (Q2) of the dataset.
3. **Whiskers**: Lines extending from the box to the minimum and maximum values that are not considered outliers. Whiskers typically extend to 1.5 times the IQR from the quartiles.
4. **Outliers**: Data points that fall outside the range defined by the whiskers are plotted as individual points and are considered potential outliers.

Box Plot Tell About :

1. **Central Tendency**: The median line inside the box gives a quick visual indication of the central value of the dataset.
2. **Spread of Data**: The length of the box (IQR) shows how much variation exists within the middle 50% of the data. A longer box indicates greater variability.
3. **Skewness**: The position of the median line within the box can indicate skewness. If the median is closer to Q1, the data may be right-skewed; if closer to Q3, it may be left-skewed.
4. **Presence of Outliers**: Individual points outside the whiskers indicate potential outliers, which may warrant further investigation. This can highlight data points that differ significantly from the rest of the dataset.
5. **Comparison Between Groups**: Box plots can be used to compare distributions across different groups or categories. Multiple box plots side by side can help visualize differences in central tendency and variability between groups.

Qus 5. Discuss the role of random sampling in making inferences about populations.

Ans: Random sampling is a fundamental technique in statistics that plays a crucial role in making inferences about populations. Here's a detailed look at its significance and impact:

Random sampling involves selecting a subset of individuals from a larger population in such a way that every individual has an equal chance of being chosen. This method minimizes bias and helps ensure that the sample accurately represents the population.

Role of Random Sampling in Making Inferences

1. **Representativeness**:
   - **Objective**: The primary goal of random sampling is to obtain a sample that accurately reflects the characteristics of the population.
   - **Benefit**: When a sample is representative, the results obtained from it can be generalized to the entire population with greater confidence.

2. **Reduction of Bias**:
   - **Bias Introduction**: Non-random sampling methods, like convenience sampling or voluntary response sampling, can introduce systematic bias that skews the results.
   - **Mitigation**: Random sampling helps to eliminate selection bias, making the sample more likely to include a diverse range of individuals.
3. **Statistical Inference**:
   - **Foundation of Inference**: Random samples allow researchers to use statistical methods to estimate population parameters (like means, proportions) and test hypotheses.
   - **Example**: If a researcher randomly samples 200 voters to estimate the proportion of a population that supports a particular candidate, the results can be extrapolated to the entire voting population.
4. **Margin of Error and Confidence Intervals**:
   - **Calculation**: Random sampling enables the calculation of margins of error and confidence intervals, which provide a range within which the true population parameter is likely to fall.
   - **Interpretation**: For example, a survey might find that 60% of a random sample supports a policy, with a margin of error of ±5%. This indicates that the true support in the population is likely between 55% and 65%.
5. **Facilitates Hypothesis Testing**:
   - **Basis for Tests**: Many statistical tests rely on the assumption that the data were collected through random sampling. This ensures that the test statistics are valid and the conclusions drawn from them are sound.
   - **Example**: In a clinical trial, random sampling helps to determine if a new drug is more effective than a placebo by minimizing confounding variables.
6. **Generalizability**:
   - **Expanding Findings**: Random samples allow researchers to generalize their findings beyond the sample to the broader population, enhancing the external validity of the study.
   - **Practical Applications**: For instance, a medical study using random sampling can lead to findings that are applicable to the entire patient population, not just the study participants.

Qus 6. Explain the concept of skewness and its types. How does skewness affect the interpretation of data?

Ans: Skewness is a statistical measure that describes the asymmetry of the distribution of data points in a dataset around its mean. It indicates whether the data tend to lean more to one side (left or right) of the distribution. Understanding skewness is important for interpreting data accurately, as it can affect statistical analyses and the conclusions drawn from the data.

Types of Skewness

1. **Positive Skewness (Right Skewness)**
   - **Description**: In a positively skewed distribution, the tail on the right side (higher values) is longer or fatter than the left side. This means that there are a significant number of higher values that pull the mean to the right.
   - **Characteristics**:
     - Mean > Median > Mode
     - The majority of data points are concentrated on the left side.
   - **Example**: Income distribution is often positively skewed, as a small number of individuals earn significantly higher incomes than the majority.
2. **Negative Skewness (Left Skewness)**
   - **Description**: In a negatively skewed distribution, the tail on the left side (lower values) is longer or fatter than the right side. This means that there are a significant number of lower values that pull the mean to the left.
   - **Characteristics**:
     - Mean < Median < Mode
     - The majority of data points are concentrated on the right side.

- **Example**: Age at retirement can be negatively skewed, where most people retire around the same age, but a few retire significantly earlier.

3. **Zero Skewness (Symmetric Distribution)**
   - **Description**: In a perfectly symmetrical distribution, the left and right tails are balanced and equal. Commonly represented by a normal distribution.
   - **Characteristics**:
     - Mean = Median = Mode
   - **Example**: The heights of adult males in a population often approximate a normal distribution.

Skewness Affects Interpretation of Data

1. **Impact on Measures of Central Tendency**:
   - In skewed distributions, the mean can be heavily influenced by extreme values (outliers). As a result, the mean may not accurately represent the "typical" value in the dataset.
   - In positively skewed data, the mean is higher than the median, potentially giving a misleading impression of the data's central location.
2. **Effect on Statistical Analysis**:
   - Many statistical tests assume normality (zero skewness) in the data. Skewed data can violate these assumptions, leading to incorrect conclusions.
   - When conducting regression analysis, skewness can affect the relationship between variables, leading to biased estimates.
3. **Interpretation of Spread**:
   - Understanding the skewness helps in interpreting the spread and variability in the data. For instance, in a positively skewed distribution, the presence of high outliers may warrant further investigation into their causes.
4. **Guiding Data Transformations**:
   - If skewness is significant, researchers might choose to apply transformations (e.g., log transformation) to make the data more symmetric, allowing for better applicability of statistical techniques.
5. **Visualization**:
   - Recognizing skewness can influence how data is visualized. Box plots and histograms can clearly show the degree and direction of skewness, aiding in data interpretation.

Qus 7. What is the interquartile range (IQR), and how is it used to detect outliers?

Ans: The interquartile range (IQR) is a measure of statistical dispersion that describes the range within which the central 50% of a dataset lies. It is calculated by taking the difference between the third quartile (Q3) and the first quartile (Q1):

IQR=Q3−Q1

Components of IQR

- **First Quartile (Q1)**: The value below which 25% of the data falls. It represents the 25th percentile.
- **Third Quartile (Q3)**: The value below which 75% of the data falls. It represents the 75th percentile.

Purpose of IQR

The IQR provides a measure of the spread of the middle half of the data, which helps to understand the variability without being affected by extreme values or outliers.

Detecting Outliers Using IQR

Outliers are data points that fall significantly outside the expected range of values. The IQR is commonly used to identify these outliers through the following steps:

1. **Calculate the IQR**:
   - Determine Q1 and Q3 and then compute the IQR using the formula $IQR=Q3-Q1$
2. **Determine the Lower and Upper Bounds**:
   - **Lower Bound**: $Q1-1.5\times IQR$
   - **Upper Bound**: $Q3+1.5\times IQR$
3. **Identify Outliers**:
   - Any data point below the lower bound or above the upper bound is considered an outlier.

Eg: Dataset: 3,7,8,12,13,14,18,20,22,24

- Q1 (25th percentile): 12.5 (the average of 12 and 13)
- Q3 (75th percentile): 20 (the average of 20 and 22)

Calculate IQR:

$IQR=Q3-Q1=20-12.5=7.5$

Determine Lower and Upper Bounds:

- Lower Bound: $Q1-1.5\times IQR=12.5-1.5\times 7.5=12.5-11.25=1.25$
- Upper Bound: $Q3+1.5\times IQR=20+1.5\times 7.5=20+11.25=31.25$

Identify Outliers:

- Any data point below 1.25 or above 31.25 is an outlier. In this case, the dataset has no outliers.

Qus 8. Discuss the conditions under which the binomial distribution is used.

Ans: The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials (experiments with two possible outcomes: success or failure). For a random variable to follow a binomial distribution, certain conditions must be met:

Conditions for Using the Binomial Distribution

1. **Fixed Number of Trials (n)**:
   - The number of trials (n) must be predetermined and fixed in advance. Each trial is considered a single event where a success or failure can occur.
2. **Two Possible Outcomes**:
   - Each trial must have exactly two possible outcomes: success (often denoted as "1") and failure (denoted as "0"). These outcomes should be mutually exclusive.
3. **Constant Probability of Success (p)**:
   - The probability of success (p) must remain constant across all trials. This means that the likelihood of achieving a success does not change from one trial to another.
4. **Independence of Trials**:
   - The trials must be independent of one another, meaning that the outcome of one trial does not affect the outcomes of others. For example, flipping a coin multiple times should not influence the result of previous flips.

When these conditions are satisfied, the random variable X (representing the number of successes in n trials) can be described by the binomial distribution, denoted as X∼Binomial(n,p). The probability of obtaining exactly k successes in n trials is given by the formula:

$$P(X = k) = \overbrace{\binom{n}{k} p^k (1-p)^{n-k}}^{\text{Probability Mass Function for a Binomial}}$$

Probability that our variable takes on the value k

Where:

- (nk) the binomial coefficient, representing the number of ways to choose kkk successes from nnn trials.
- p^k is the probability of success raised to the number of successes.
- (1−p)n−k is the probability of failure raised to the number of failures.

Examples of Binomial Distribution Use Cases

- **Quality Control**: Testing a batch of products where each product can either be defective or non-defective.
- **Surveys**: Conducting a survey to determine the proportion of a population that supports a particular candidate (success) versus those who do not (failure).
- **Clinical Trials**: Evaluating the effectiveness of a new drug, where each patient can either respond positively to the treatment (success) or not (failure).

Qus 9. Explain the properties of the normal distribution and the empirical rule (68-95-99.7 rule).

Ans: The normal distribution, often referred to as the Gaussian distribution, is a continuous probability distribution that is symmetrical and bell-shaped. It plays a crucial role in statistics and natural phenomena due to its unique properties. Here are the key properties and the empirical rule associated with the normal distribution:

Properties of the Normal Distribution

1. **Symmetry**:
   - The normal distribution is perfectly symmetrical around its mean. This means that the left half of the distribution is a mirror image of the right half.
2. **Mean, Median, and Mode**:
   - In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.
3. **Bell-Shaped Curve**:
   - The distribution has a bell shape, with most of the data points clustering around the mean and fewer observations as you move away from the center.
4. **Defined by Two Parameters**:
   - The normal distribution is fully characterized by its mean (μ) and standard deviation (σ). The mean determines the center of the distribution, while the standard deviation measures the spread or width of the distribution.
5. **Asymptotic**:
   - The tails of the normal distribution approach the horizontal axis but never actually touch it. This means that there is always a non-zero probability of observing extreme values far from the mean.
6. **Area Under the Curve**:

- The total area under the curve of a normal distribution is equal to 1, representing the total probability of all outcomes.

The Empirical Rule (68-95-99.7 Rule)

The empirical rule provides a quick way to understand the distribution of data in a normal distribution:

1. **68% Rule**:
   - Approximately 68% of the data falls within one standard deviation (σ) of the mean (μ). This means that if you calculate μ ± σ, about 68% of the observations will lie within this range.
2. **95% Rule**:
   - Approximately 95% of the data falls within two standard deviations of the mean. So, if you calculate μ ± 2σ, about 95% of the observations will lie within this interval.
3. **99.7% Rule**:
   - Approximately 99.7% of the data falls within three standard deviations of the mean. Hence, calculating μ ± 3σ will encompass about 99.7% of the data points.

Visual Representation

A normal distribution curve will typically look like this:

diff
Copy code
```
    |        *
    |      *   *
    |     *      *
    |    *         *
    |   *           *
    |  *             *
    | *               *
----|------------------------|----
   μ - 3σ   μ - 2σ  μ - σ  μ  μ + σ  μ + 2σ  μ + 3σ
```

Importance of the Normal Distribution and Empirical Rule

- **Statistical Inference**: Many statistical tests assume that data follows a normal distribution, making it essential for hypothesis testing and confidence interval estimation.
- **Real-World Applications**: Numerous natural phenomena (e.g., heights, test scores, measurement errors) approximate a normal distribution, allowing the empirical rule to be used for practical decision-making and predictions.

The normal distribution is characterized by its symmetry, bell shape, and defined by its mean and standard deviation. The empirical rule provides a valuable guideline for understanding how data is distributed around the mean, highlighting the percentages of data that fall within one, two, and three standard deviations from the mean. These concepts are foundational in statistics and are widely used in various fields for analysis and inference.

Qus 10. Provide a real-life example of a Poisson process and calculate the probability for a specific event.

Ans: A Poisson process is a statistical model that describes events occurring randomly and independently over a specified interval of time or space. It is characterized by a constant average rate of occurrence. A common real-life example of a Poisson process is the number of customer arrivals at a bank during a specific hour.

Real-Life Example: Customer Arrivals at a Bank

Scenario: Suppose a bank receives an average of 6 customers per hour. We can model the arrivals of customers as a Poisson process.

Poisson Distribution Formula

The probability of observing k events (customer arrivals) in a given interval is given by the Poisson probability mass function:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where
x = 0, 1, 2, 3, …
λ = mean number of occurrences in the interval
e = Euler's constant $\approx$ 2.71828

Eg: calculate the probability of exactly 5 customers arriving in one hour (k=5).

1. Substitute the values into the formula:

- $\lambda=10$
- k=5 $P(X=5)=10^5 e^{-10}/5!$

2. Calculate 10^5: 10^5=100000
3. Calculate 5!: 5!=120
4. Calculate e^{-10} (using a calculator): $e^{-10}\approx 0.0000453999$
5. Combine all parts: $P(X=5)=100000\times 0.0000453999/120$
6. Perform the calculation: $P(X=5) \approx 4.53999/120 \approx 0.03783$

Qus 11. Explain what a random variable is and differentiate between discrete and continuous random variables.

Ans: A **random variable** is a numerical outcome of a random phenomenon or experiment. It assigns a real number to each possible outcome of that experiment, allowing for the quantification of randomness. Random variables are crucial in statistics and probability theory as they help in modeling and analyzing uncertainties.

Types of Random Variables

Random variables are typically classified into two main types: **discrete** and **continuous**.

1. Discrete Random Variables

- **Definition**: A discrete random variable takes on a countable number of distinct values. This means that the values can be enumerated or listed.
- **Characteristics**:
  - Can only take specific values (often integers).
  - Often arises from counting processes (e.g., the number of students in a classroom, the number of heads in a series of coin flips).
- **Example**:
  - The number of customers arriving at a store in one hour.

- The outcome of rolling a die (1, 2, 3, 4, 5, or 6).

2. Continuous Random Variables

- **Definition**: A continuous random variable can take on an infinite number of values within a given range. These values can include fractions and decimals.
- **Characteristics**:
  - The possible values are uncountable and can be represented on a continuum.
  - Often arise from measurement processes (e.g., height, weight, temperature).
- **Example**:
  - The height of individuals in a population (can take any value within a range).
  - The time it takes to complete a task (can be measured in seconds, minutes, etc., and can include decimal values).

Key Differences

Feature	Discrete Random Variables	Continuous Random Variables
Values	Countable (e.g., integers)	Uncountable (e.g., real numbers)
Nature	Often results from counting	Often results from measuring
Probability Distribution	Described using a probability mass function (PMF)	Described using a probability density function (PDF)
Examples	Number of students, number of cars	Height, weight, time

Qus 12. Provide an example dataset, calculate both covariance and correlation, and interpret the results

Ans: with two variables: **X** and **Y**.

Dataset

X	Y
2	3
4	5
6	7
8	9
10	11

Step 1: Calculate the Covariance

Covariance measures how much two random variables vary together. The formula for covariance Cov(X,Y)\text{Cov}(X, Y)Cov(X,Y) is:

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

Where:

- n = number of data points
- X̅= mean of X
- Y̅= mean of Y

Calculate the means:

- X̅=2+4+6+8+105/5=6
- Y̅=3+5+7+9+115/5=7

Calculate Cov(X, Y):

Cov(X,Y)=(2−6)(3−7)+(4−6)(5−7)+(6−6)(7−7)+(8−6)(9−7)+(10−6)(11−7)/(5-1)

Calculating each term:

- For (2−6)(3−7)=(−4)(−4)=16
- For (4−6)(5−7)=(−2)(−2)=4
- For (6−6)(7−7)=(0)(0)=0
- For (8−6)(9−7)=(2)(2)=4
- For (10−6)(11−7)=(4)(4)=16

Now sum them up:

16+4+0+4+16=40

Thus,

Cov(X,Y)=40/4=10

Step 2: Calculate the Correlation

Correlation is a standardized measure of the relationship between two variables. The formula for correlation (r) is:

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Where

$$\sigma_x = \frac{\sqrt{\Sigma(x_i - \bar{x})^2}}{\sqrt{n-1}}$$

$$\sigma_y = \frac{\sqrt{\Sigma(y_i - \bar{y})^2}}{\sqrt{n-1}}$$

Calculating σX :

$\sigma X = \sqrt{\{[(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2]/4\}}$

Calculating each term:

- $(2-6)^2=16$
- $(4-6)^2=4$
- $(6-6)^2=0$
- $(8-6)^2=4$
- $(10-6)^2=16$

Now sum them up:

16+4+0+4+16=40

Thus,

$\sigma X=\sqrt{(40/4)}= \sqrt{10}\approx 3.16$

Calculating σY:

$\sigma Y = \sqrt{\{[(3-7)^2+(5-7)^2+(7-7)^2+(9-7)^2+(11-7)^2]/4\}}$

Calculating each term:

- $(3-7)^2 = 16$
- $(5-7)^2 = 4$
- $(7-7)^2 = 0$
- $(9-7)^2 = 4$
- $(11-7)^2 = 16$

Now sum them up:

16+4+0+4+16=40

Thus,

$\sigma Y=\sqrt{(40/4)}= \sqrt{10}\approx 3.16$

Calculate Correlation r:

$r=Cov(X,Y)/(\sigma X*\sigma Y)$
$=10/(3.16\times 3.16)$
$=10/10$
$=1$

Interpretation of Results

- **Covariance of 10**: Indicates that as X increases, Y tends to also increase, but the magnitude alone doesn't provide a standardized comparison.
- **Correlation of 1**: Suggests a perfect positive linear relationship between X and Y. This means that as X increases, Y increases in a perfectly linear manner.

Therefore, this dataset shows a strong positive correlation, indicating that the two variables are closely related and move together in a predictable way.