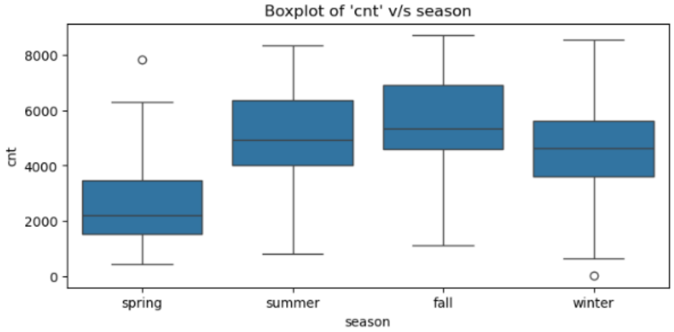
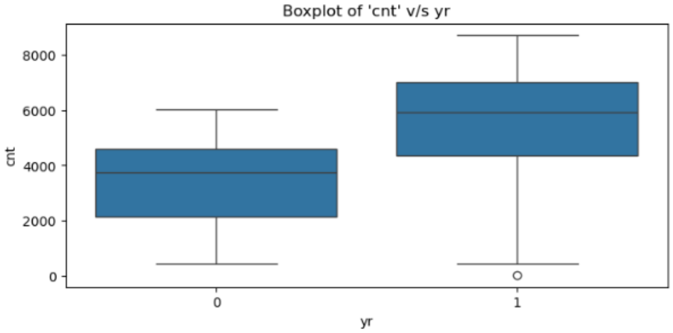
Linear Regression Assignment

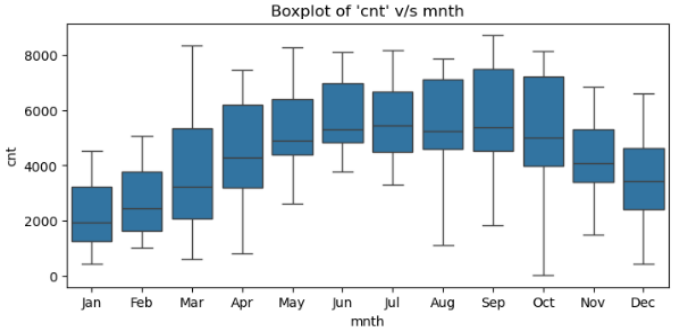
Assignment-based Subjective Questions

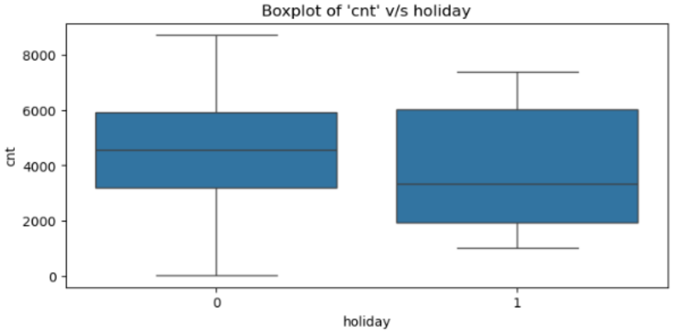
Submitted By Anuradha Bharti

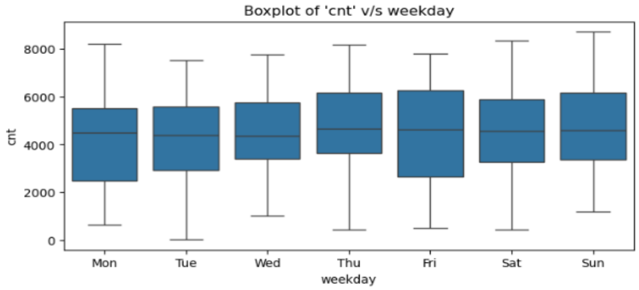
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

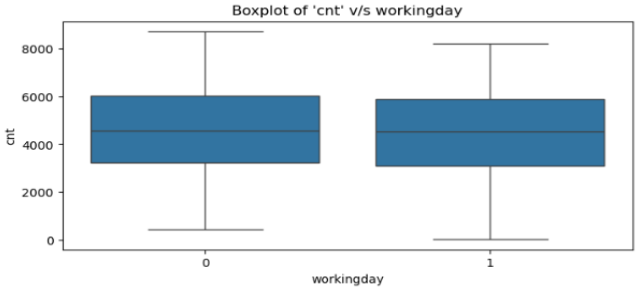


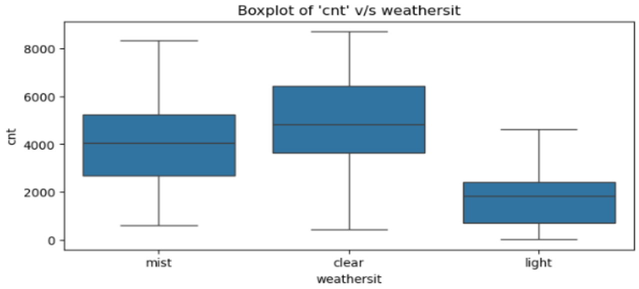












The dataset contains these categories: -

season, **year**, **holiday**, **weekday**, **working day**, **weather situation** & **month**.
I have plotted these using boxplots to see how they affect bike rentals. Here's what I found:

- **Season**: Fall (season 3) had the highest bike rentals, while Spring (season 1) had the lowest.
- **Year**: More people rented bikes in 2019 than in 2018.
- **Holiday**: Bike rentals dropped during holidays.
- **Weekday**: Bike rentals were pretty steady throughout the week.
- **Working Day**: Rentals were mostly between 4000 and 6000, and there wasn't much difference in bookings whether it was a working day or not.
- **Weather Situation**: There were no rentals during heavy rain or snow. The most rentals happened when the weather was clear or partly cloudy.
- **Month**: Rentals were highest in September. However, rentals might have dropped in December due to snowfall.

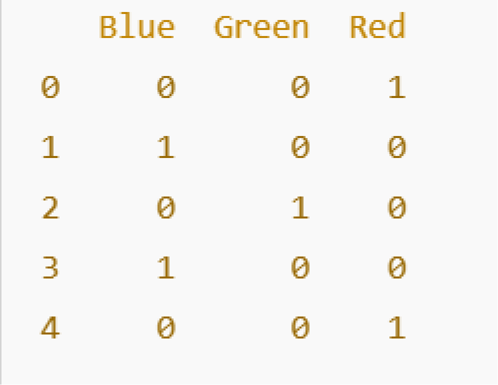
2. Why is it important to use drop_first=True during dummy variable creation?

Using drop_first=True during dummy variable creation is important to avoid **multicollinearity**. When you create dummy variables for categorical data, it generates a new binary column for each category. However, one of these columns will be redundant. By using drop_first=True, we remove one of the columns to prevent this redundancy, ensuring that the model doesn't treat the columns as highly correlated, which could distort the results.

Example: Consider a column "Color" with three categories: Red, Blue, and Green.

Without drop_first=True:

When we create dummy variables without dropping the first category:



```
   Blue  Green  Red
0     0      0    1
1     1      0    0
2     0      1    0
3     1      0    0
4     0      0    1
```

Output without drop_first=True

In this case, we have three columns: Blue, Green, and Red, and each row gets a binary value (0 or 1) for these colors.

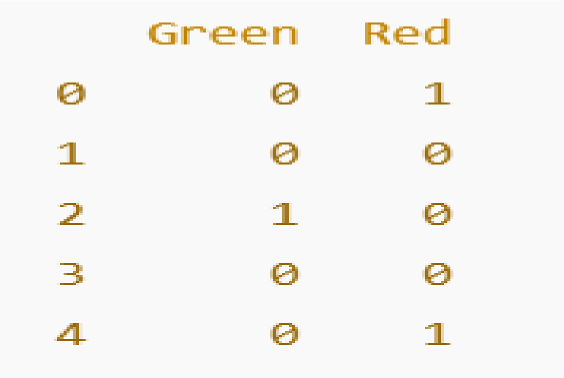
Multicollinearity Issue:

If we know the values of two of the columns (e.g., Blue and Green), we can always determine the value of the third column (e.g., Red). For example:

If Blue = 1 and Green = 0, then Red must be 0.

If Blue = 0 and Green = 0, then Red must be 1.

This redundancy creates multicollinearity, which can cause problems in linear regression.



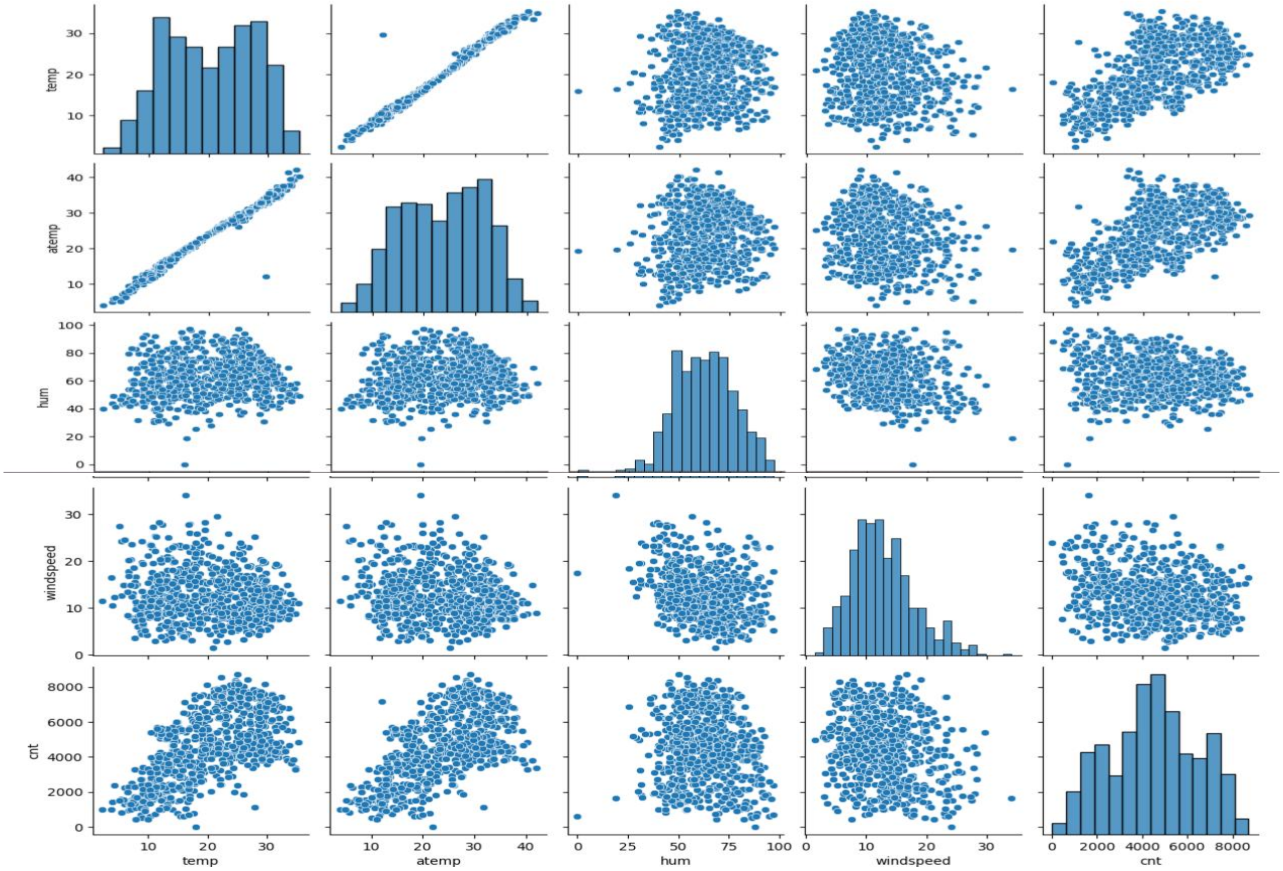
```
   Green  Red
0      0    1
1      0    0
2      1    0
3      0    0
4      0    1
```

Output with drop_first=True

By dropping the first column (Blue), we remove the redundancy and prevent multicollinearity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

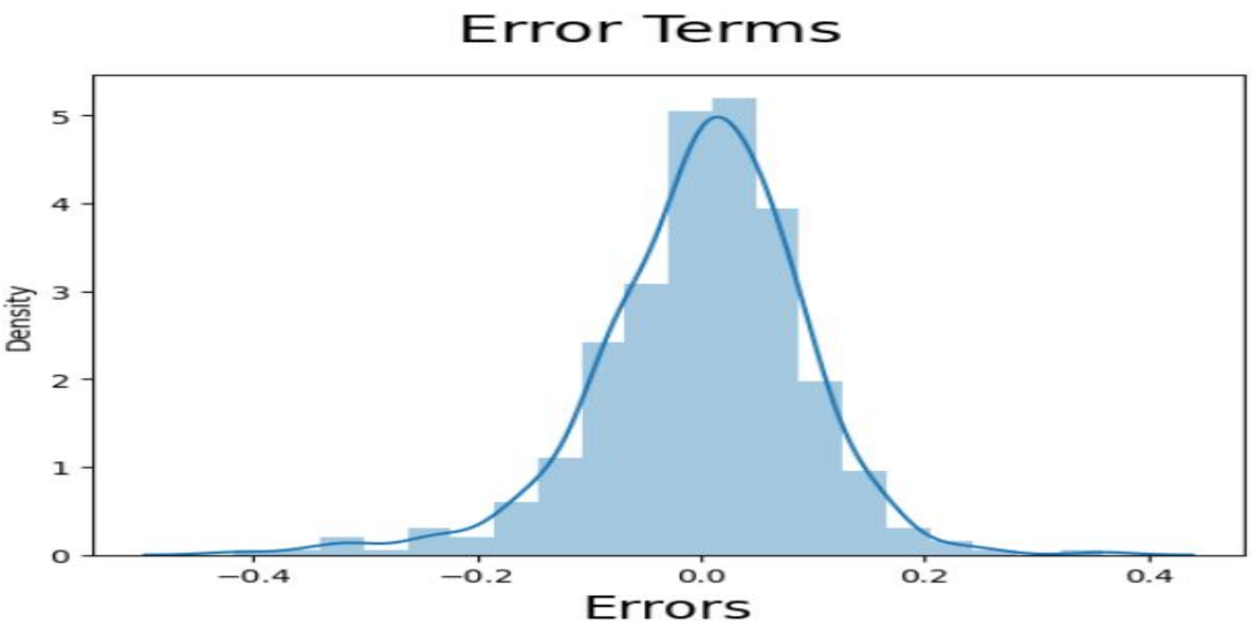
"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).



4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Linear Relationship: We checked if there is a straight-line relationship between the independent (predictor) and dependent (outcome) variables. To do this, we used a pair plot to visualize the data and see if the variables seem to be related in a linear way.

- Residuals Distribution: We looked at the residuals (the difference between the predicted and actual values) to check if they are normally distributed and centered around 0(mean=0). We plotted a distribution of the residuals to see if they follow a normal pattern.



- No Multicollinearity: We checked if the independent variables are not too strongly related to each other, as this could cause problems in the model. We used the VIF (Variance Inflation Factor) to measure how much each variable is correlated with the others in the model.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Temp (Temperature): 0.4782
  This has the highest positive contribution, suggesting that temperature has a strong impact on bike demand. Higher temperatures likely increase bike rentals.
- Weathersit_light: -0.286002
  This feature indicates a negative relationship with the demand, meaning that light weather conditions (which may indicate good weather) tend to reduce bike demand compared to other weather conditions.

- Yr (Year): 0.234060
  The year variable shows a positive relationship, suggesting that over time, the demand for shared bikes has increased.
  These features have the most noticeable influence on predicting bike demand in the model.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a type of **supervised machine learning** that learns from **labeled data** (data with known outcomes). It tries to find the best straight-line relationship between the input data points and uses that relationship to make predictions on new data. The algorithm looks at how different factors (independent variables) are connected to the target variable (dependent variable) and tries to fit a straight line to these data points, known as **best fit line**.

Example:

- For example, if we want to predict the price of a house, we look at various factors like the house's age, distance from the main road, location, size, and number of rooms. Linear regression considers these factors and assumes there's a straight-line relationship between them and the house price. By doing so, it can predict the price of a house based on the values of these factors.
- For **simple linear regression** the line can be
- represented by an equation = c + mx
- where c is the intercept (the value of y when x is zero) and m is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of m and c, we need to define a **cost function** that measures how well the line fits the data. We can choose **the mean squared error** (MSE), which is the average of the squared differences between the actual values of y and the predicted y values:

MSE = (1/n) * Σ (y - y') ^2

where n is the number of data points, y is the actual value, and y' is the predicted value.

Our goal is to minimize the MSE by adjusting m and c. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

For **multiple linear regression** the equation of line is given by:

y = c + m1x1 + m2x2 + … + mnxn

Model Evaluation:

After finding the optimal parameters, the model's performance can be evaluated using several metrics, one metric is R-squared.

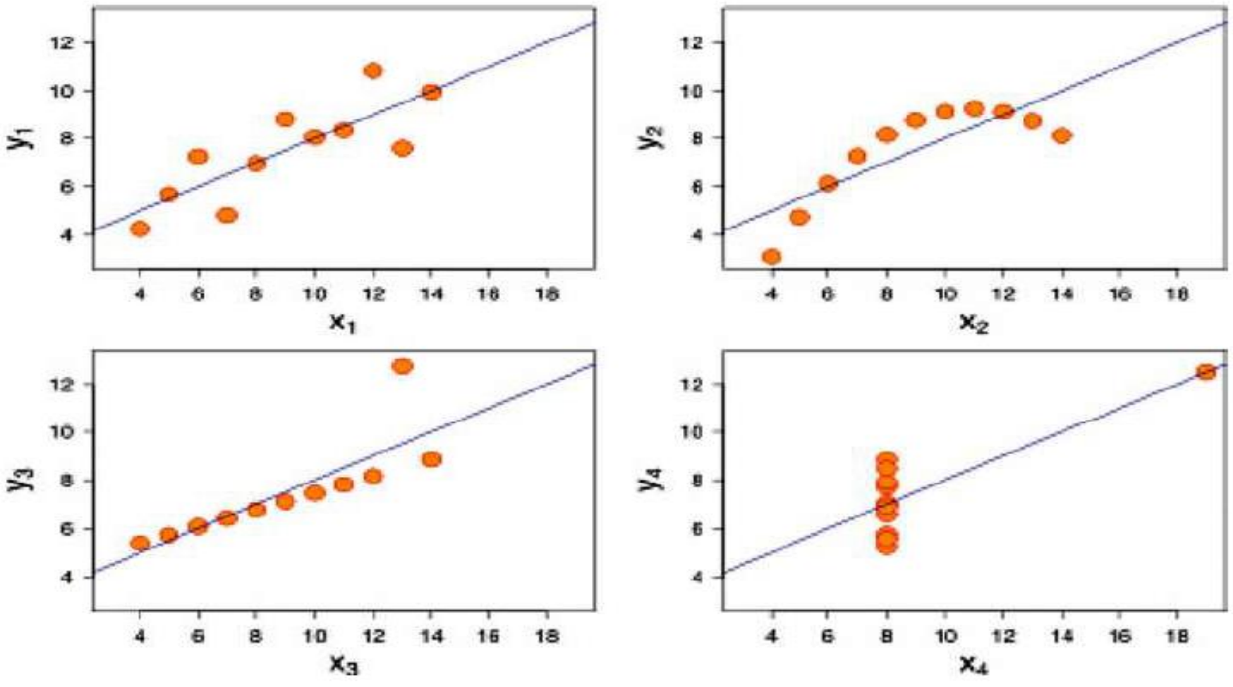
- **R-squared:** This metric measures the proportion of the variance in the target variable that is explained by the model. The value ranges from 0 to 1, where a higher value indicates a better fit.

Limitations:

- It assumes a linear relationship between variables.
- It is sensitive to outliers, which can significantly affect the model.
- It assumes independence between predictors (no multicollinearity).
- It can't capture complex relationships or non-linear patterns.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was created by statistician Francis Anscombe to show how four datasets, despite having nearly identical statistical summaries, can look completely different when plotted on a graph. The goal is to highlight the importance of graphing data before analyzing it and how outliers (unusual data points) can affect statistical results.



- **First Graph (Top Left)**: This graph shows a clear linear relationship. It looks like a straight line, which means that as one variable increases, the other variable also increases in a predictable way.
- **Second Graph (Top Right)**: This graph doesn’t follow a normal pattern. Even though there's a relationship between the variables, it's not linear (not in a straight line), so it doesn't follow the same trend as the first graph.
- **Third Graph (Bottom Left)**: This graph looks linear at first, but there is an outlier (a point that’s far from the rest). This outlier affects the regression line (the line that fits the data) and lowers the correlation from 1 (perfect relationship) to 0.816. The outlier has a big impact on the results.
- **Fourth Graph (Bottom Right)**: This graph shows that even if most of the data points don't follow a clear pattern, one high-leverage point (an outlier with a big influence) can create a high correlation. This means that the data may appear to have a strong relationship, even when most of the points do not show any real connection.

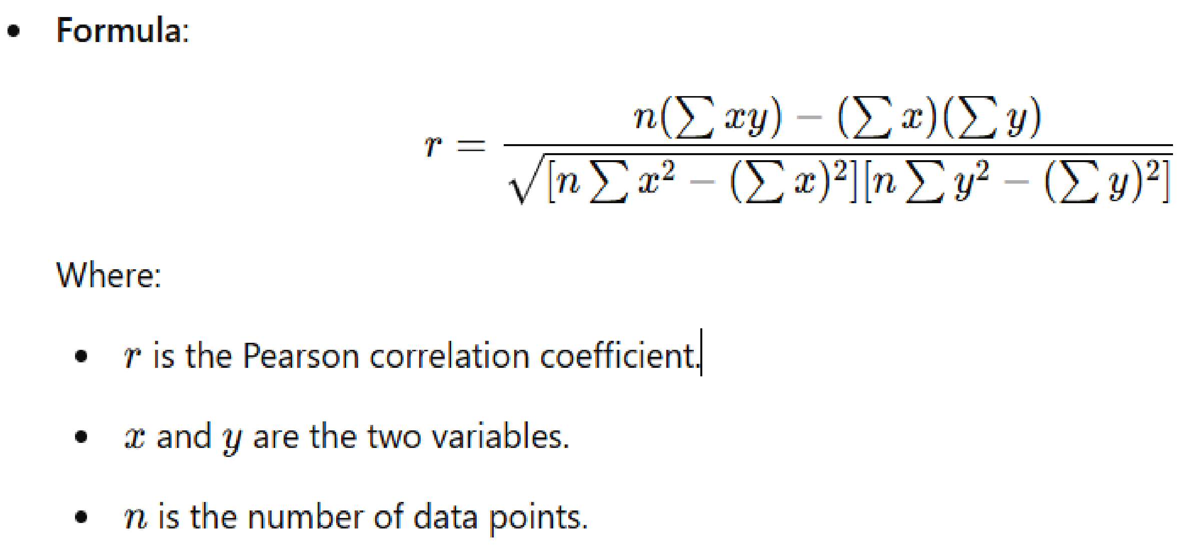
In summary, Anscombe's Quartet teaches us that looking at data visually is important because it can reveal things that numbers alone cannot. Even if the summary statistics are the same, the data can behave very differently when plotted, especially when outliers are present.

3. What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that describes the strength and direction of the linear relationship between two variables. It ranges from -1 to 1 and is used to determine how closely the two variables are related in a straight-line fashion.

Key Points:

- Value Range: Pearson's R ranges from -1 to 1.
- +1 means a perfect positive linear relationship: As one variable increases, the other variable also increases in a perfectly straight line.
- -1 means a perfect negative linear relationship: As one variable increases, the other variable decreases in a perfectly straight line.
- 0 means no linear relationship between the variables.
- Strength of the Relationship:
- 0.7 to 1 (or -0.7 to -1) indicates a strong linear relationship.
- 0.3 to 0.7 (or -0.3 to -0.7) indicates a moderate linear relationship.
- 0 to 0.3 (or 0 to -0.3) indicates a weak linear relationship.



- **Formula**:

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

Where:

- r is the Pearson correlation coefficient.
- x and y are the two variables.
- n is the number of data points.

If I am studying the relationship between hours studied and test scores, a high positive Pearson's R (e.g., +0.85) would suggest that as the number of hours studied increases, the test scores tend to increase as well, and the relationship is relatively strong.

Example-

On the other hand, if the Pearson's R is close to 0, it means that the number of hours studied doesn't strongly predict the test scores, and the relationship may be weak or non-linear.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a process of transforming the features (variables) of dataset to a specific range or distribution. This is done to ensure that the different features in the dataset contribute equally to the analysis, particularly when using algorithms that rely on distances or require features to be on a similar scale.

Scaling is performed for several reasons:

- Improves Model Performance: Some machine learning algorithms, like K-nearest neighbors (KNN), Support Vector Machines (SVM), and Gradient Descent, are sensitive to the scale of the data. If features are on different scales (e.g., one is in the range of 1-10 and another in the range of 1000-10000), the model might place more importance on the features with larger values, leading to biased results.
- Accelerates Convergence in Gradient Descent: For algorithms like linear regression and logistic regression that use gradient descent, scaling ensures faster convergence, as the algorithm can move more evenly across all dimensions of the feature space.

- Better Interpretation: Scaling helps to better interpret the results of the model, especially when comparing the impact of different features.
- Makes Features Comparable: Scaling ensures that all features are treated equally and allows algorithms to analyze all features in a fair manner.

Difference Between Normalized Scaling and Standardized Scaling:

Scaling can be done in two common ways: Normalization and Standardization.

1. Normalized Scaling (Min-Max Scaling)

Normalization or Min-Max scaling transforms the data to fit within a specific range, typically 0 to 1. This is done by subtracting the minimum value and then dividing by the range of the data (i.e., the difference between the maximum and minimum values).

The formula for Normalization is:

Where:

- x is the original value,
- min(x) is the minimum value of the feature,
- max(x) is the maximum value of the feature.

Normalization is useful when we know that our data is bounded within a specific range or when we want to compress all feature values to fall within a known range (like 0-1).

It works best when the data does not contain outliers, as they can distort the scaling process.

2. Standardized Scaling (Z-Score Scaling)

Standardization (or Z-score scaling) transforms the data by subtracting the mean and dividing by the standard deviation. The result is a distribution with a mean of 0 and a standard deviation of 1. This method is more robust to outliers compared to normalization.

The formula for Standardization is:

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma}$$

Where:

- xis the original value,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

- Standardization is useful when data follows a **normal distribution** or when our model assumes that data is normally distributed (like in many linear models or algorithms using distance metrics).
- It is also useful when the data contains **outliers**, as standardization is less sensitive to outliers than normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models. Multicollinearity occurs when two or more predictor variables are highly correlated with each other, which can make it difficult to determine the individual effect of each predictor variable on the dependent variable. The VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.

The formula for VIF for a predictor variable Xi is:

$$VIF(X_i) = \frac{1}{1 - R_i^2}$$

The value of VIF can become **infinite** when there is **perfect multicollinearity** between the predictor variables. This happens when one predictor variable is a perfect linear combination of other predictor(s) in the model.

In other words, if one predictor variable can be perfectly predicted by a linear equation of other predictors, the **R squared value** for that predictor will be **1**, and as a result, the VIF becomes:

$$VIF = \frac{1}{1-1} = \frac{1}{0} = \infty$$

Reasons for Infinite VIF

1. **Perfect Linear Relationship**:

- If two or more independent variables are perfectly correlated with each other (e.g., one variable is a multiple of another), the regression model cannot distinguish their individual contributions to the dependent variable. In this case, **multicollinearity** is perfect, and the **VIF** for at least one of the variables becomes infinite.

Example: If you have two predictor variables, X1 and X2, where X2 is exactly 2 times X1 (i.e., X2=X1) then the VIF for X1 or X2 will be infinite because they are perfectly linearly related.

2.Inclusion of Redundant Variables:

- Including **redundant variables** that represent the same underlying concept can also cause infinite VIF. This can happen if there is an overlap in the information conveyed by the predictor variables.

Consequences of Infinite VIF

- **Unstable Coefficients**: When VIF is infinite, the model's regression coefficients become **unstable** and may fluctuate wildly with small changes in the data.
- **Inability to Estimate Effects**: Perfect multicollinearity makes it impossible to distinguish the individual effect of highly correlated predictors on the dependent variable.
- **Model Interpretation Issues**: It becomes difficult to interpret the individual effects of the predictor variables because their contributions to the model are perfectly overlapping.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, most commonly the normal distribution. It helps to visually assess whether a set of data follows a particular distribution.

In a Q-Q plot:

The x-axis represents the theoretical quantiles (from a normal distribution or any other chosen distribution).

The y-axis represents the quantiles of the data being compared.

Use and Importance of a Q-Q Plot in Linear Regression:

- **Checking Normality of Residuals**:

    1. In **linear regression**, one of the assumptions is that the **residuals** (errors) of the model are **normally distributed**. A Q-Q plot helps in verifying this assumption.
    2. If the residuals deviate from the straight line in a Q-Q plot, it suggests that the normality assumption may not hold, which could affect the reliability of the model’s statistical tests (like t-tests or F-tests).

- **Detecting non-normality**:

    1. A Q-Q plot can reveal issues like:

        1. **Skewness**: If the points curve upward or downward, it suggests the data may be skewed.

2. **Heavy tails (Kurtosis)**: If the points deviate from the line at the ends, this may indicate that the data has heavy or light tails compared to a normal distribution.

- **Assessing Model Fit**:

1. For **linear regression**, the assumption of normality is important because many inferential statistics (such as confidence intervals and hypothesis tests) rely on it. A Q-Q plot helps assess whether this assumption holds and whether the model's predictions are trustworthy.

Importance:

- **Validating Assumptions**: A Q-Q plot is useful for validating the assumption of normality in residuals. If the assumption is violated, it may suggest the need for data transformation or the use of non-parametric models.
- **Model Diagnostics**: By identifying any deviations from normality, it allows you to take corrective actions (e.g., transformation of data or using robust regression models) to improve the model's performance and accuracy.