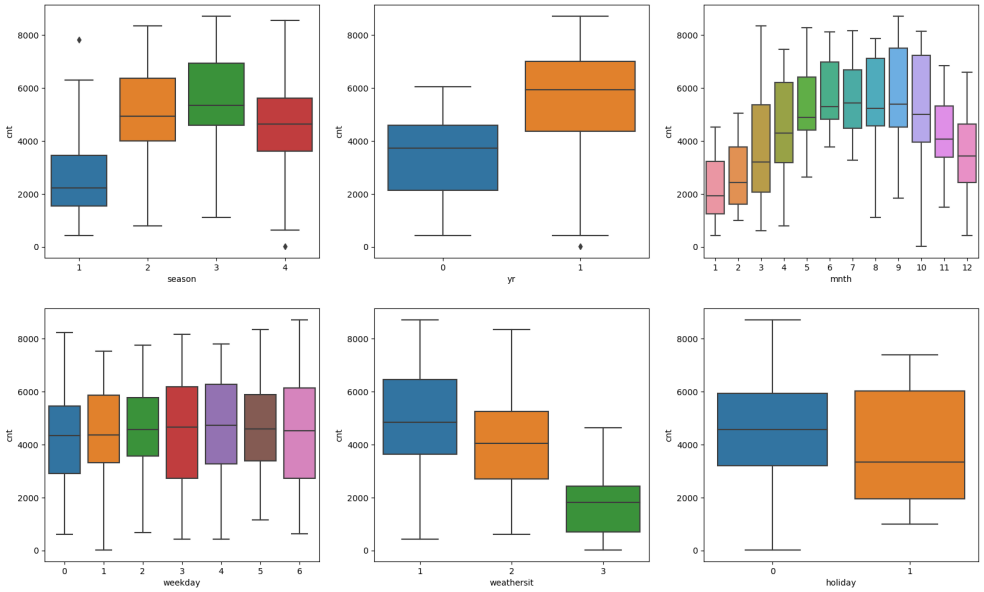
Assignment-based Subjective Questions

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



- The graph clearly shows the qualitative distributions of the data, now if the model suggests the important predictors, using these graphs we can be more confident about the predictions of the model.
- For the variable season, we can clearly see that the category 3 : Fall, has the highest median, which shows that the demand was high during this season. It is least for 1: spring.
- The year 2019 had a higher count of users as compared to the year 2018.
- The count of rentals is almost even throughout the week.
- There are no users when there is heavy rain/ snow indicating that this weather is quite adverse. Highest count was seen when the weather situation was Clear, Partly Cloudy.
- The number of rentals peaked in July and September. This observation is consistent with the observations made

regarding the weather. As a result of the typical substantial snowfall in December, rentals may have declined.

- The count of users is less during the holidays.
- From the "Workingday" boxplot we can see those maximum bookings happening between 3000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.

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

Using drop_first=True when creating dummy variables is important for correctly interpreting the coefficients in regression models and avoiding multicollinearity. It ensures that the model is properly specified and that each coefficient reflects the effect of a particular category relative to a specified baseline category (typically the first category in alphabetical order or the first encountered in the data). This practice enhances the reliability and interpretability of the regression analysis or machine learning model..

Dummy Variable Creation:

When dealing with categorical variables in regression analysis or machine learning models, it's common practice to convert categorical variables into numerical format using dummy variables. Each category of the categorical variable is represented by a binary (0 or 1) indicator variable.

In Bike Sharing case study, we have a categorical variable "season" with four categories: spring, summer, fall and winter, we would create three dummy variables:

- season _ summer: 1 if the observation is summer, 0 otherwise.
- season _ fall: 1 if the observation is fall, 0 otherwise.
- season _winter: 1 if the observation is winter, 0 otherwise.

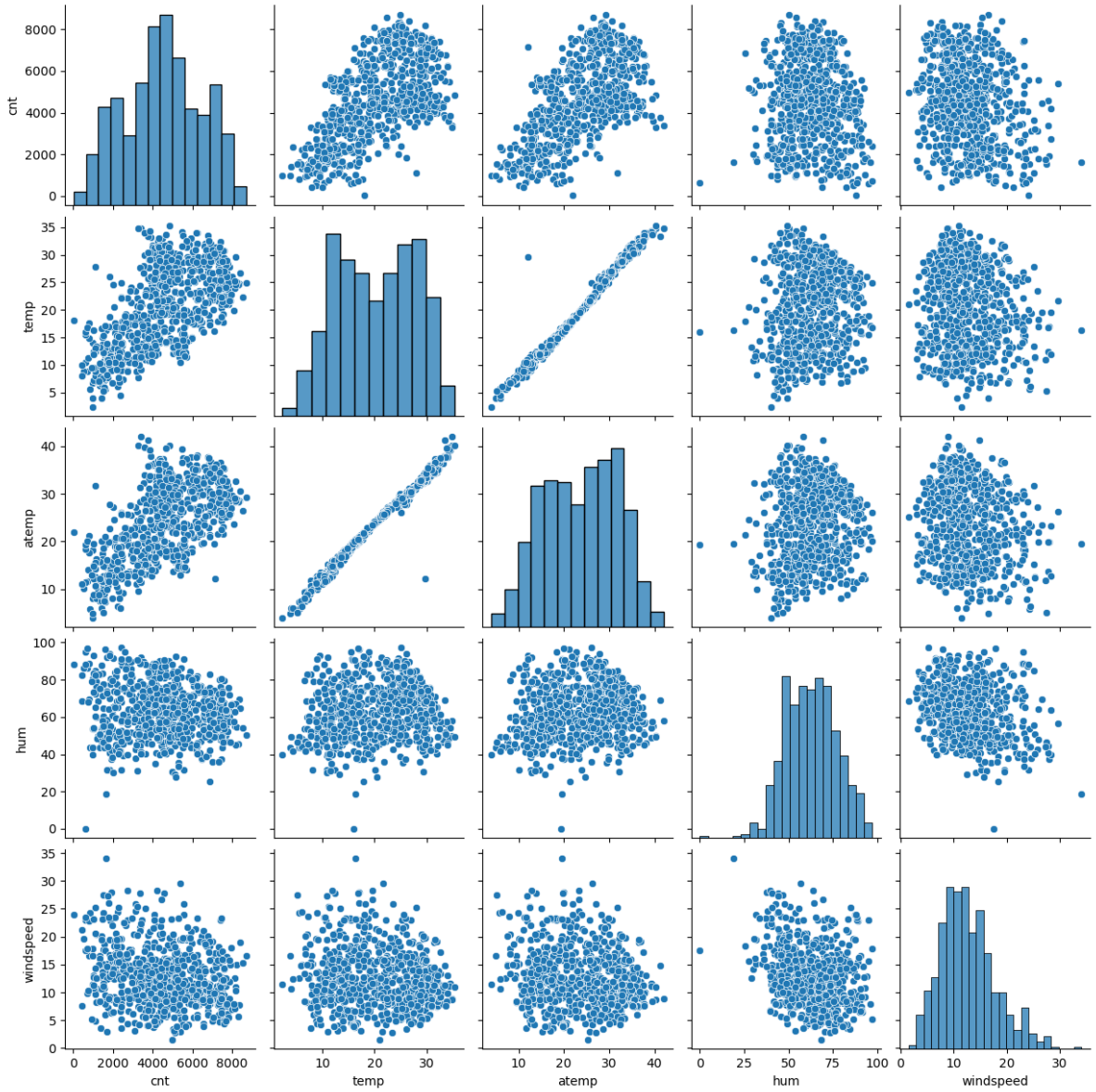
If season _ summer is 0, season _fall is 0 and season _ winter is also 0, then the observation is spring (baseline category).

Importance of drop_first=True:

1. **Avoiding Multicollinearity**:
    - If you include dummy variables for all categories (e.g., season _spring, season _summer, season _fall, season _winter), one category becomes redundant. This is because the presence of the information about any three categories is enough to predict the presence of the fourth (due to the constant sum constraint).
    - Including all dummy variables can lead to multicollinearity, where one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.
2. **Interpreting Coefficients**:
    - In regression models (like linear regression), coefficients of dummy variables represent the change in the dependent variable associated with a particular category relative to the baseline category (the category omitted).
    - When drop_first=True, one dummy variable is dropped (typically the one representing the baseline category). This ensures that each coefficient represents the effect of being in that category compared to the baseline category.
3. **Improving Model Performance**:

- By reducing multicollinearity, models can perform better because they avoid the issues of unstable coefficients and inflated standard errors that can arise when predictors are highly correlated.

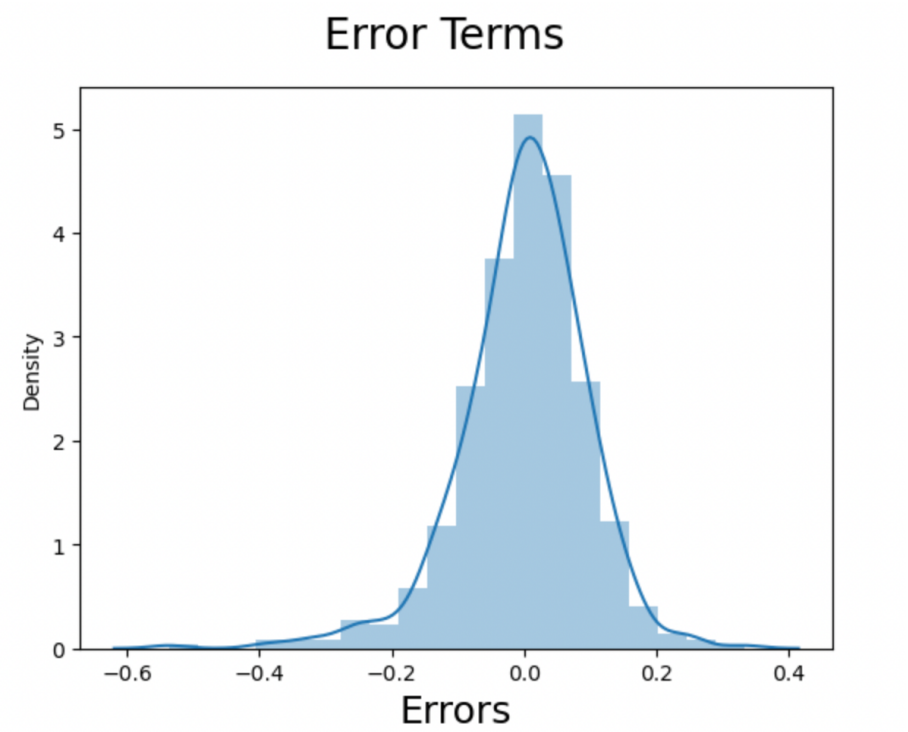
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



After looking at above pairplot graph, we can say that temp and atemp have the highest correlation with the target variable cnt.

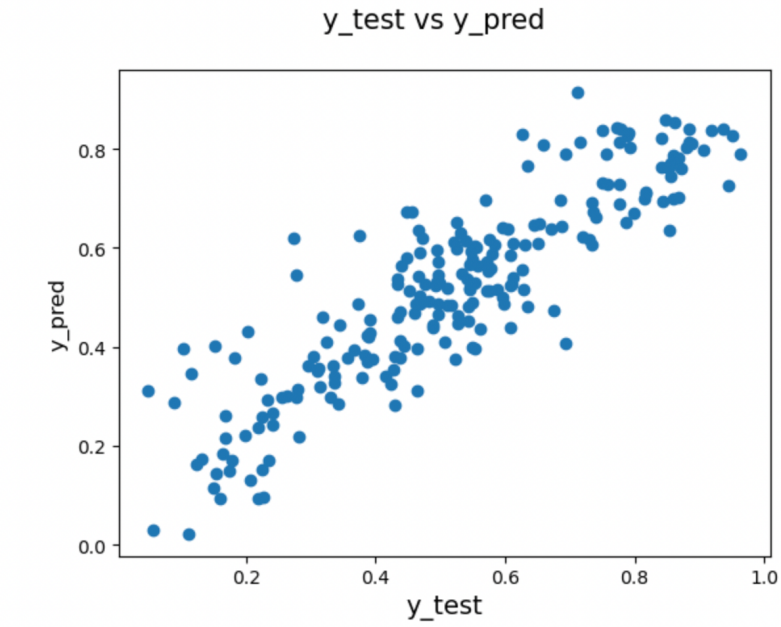
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Residual Analysis of the train data

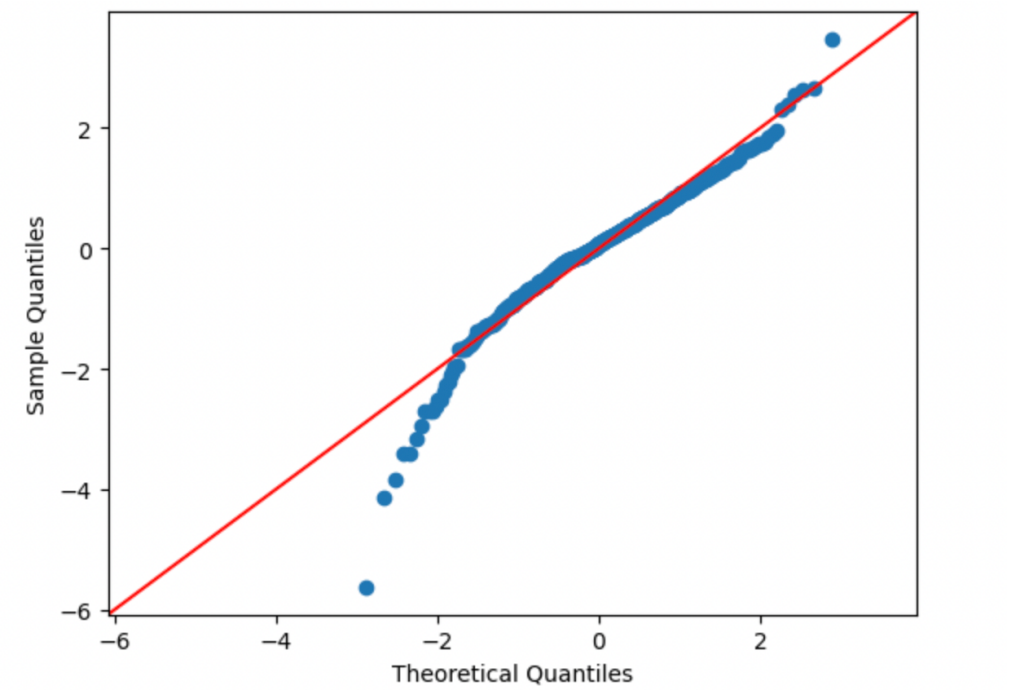


Error terms are normally distributed as per above histogram.

Test vs Predicted:



Cross-verifying the above conclusion using a qq-plot as well:



Here we see that most of the data points lie on the straight line , which indicates that the error terms are normally distributed .

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

Based on final model top three features contributing significantly towards explaining the demand are:

Temperature (0.488858)

windspeed : (-0.180140)

year (0.237253)

Hence, it can be clearly concluded that the variables temperature , windspeed and month are significant in predicting the demand for shared bikes .

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical and machine learning technique used for predicting a continuous dependent variable (outcome) based on one or more independent variables (features). It models the relationship between the dependent variable y and independent variables X by fitting a linear equation to observed data.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

The linear regression model assumes a linear relationship between the independent variables X and the dependent variable y. Mathematically, it can be expressed as:

$y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n+\epsilon$

Where:

1. y is the dependent variable (the variable we want to predict),
2. $x_1,x_2,...,x_n$ are the independent variables (features),
3. β_0 is the intercept (constant term),
4. $\beta_1,\beta_2,...,\beta_n$ are the coefficients of the independent variables $x_1,x_2,...,x_n$ respectively,
5. ϵ is the error term which represents the noise or error in the model.

The assumptions of linear regression are:

1. The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the ‘linearity assumption’.

2. Assumptions about the residuals:

1. Normality assumption: It is assumed that the error terms, ε, are normally distributed.
2. Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
3. Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ^2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
4. Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

3. Assumptions about the estimators:

1. The independent variables are measured without error.
2. The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

Extensions and Variants:

1. **Regularized Regression**: Techniques like Ridge Regression (L2 regularization) and Lasso Regression (L1 regularization) add a penalty term to the loss function to prevent overfitting.
2. **Multiple Linear Regression**: Extension of simple linear regression to multiple independent variables.
3. **Polynomial Regression**: Uses higher-degree polynomial terms to capture nonlinear relationships.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a famous dataset in statistics and data visualization that illustrates the importance of graphing data before analyzing it and highlights the pitfalls of relying solely on summary statistics. It consists of four small datasets that have nearly identical statistical properties when examined using simple summary statistics (mean, variance, correlation, etc.), but appear very different when plotted.

The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the impact of outliers and the effect of different data distributions on statistical properties. Despite having similar statistical properties, the datasets exhibit distinct patterns and relationships when visualized.

Description of the Datasets:

Each dataset in Anscombe's quartet consists of 11 data points:

1. **Dataset I**:
    - **x**: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
    - **y**: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82
2. **Dataset II**:
    - **x**: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
    - **y**: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26
3. **Dataset III**:
    - **x**: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
    - **y**: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42
4. **Dataset IV**:
    - **x**: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8
    - **y**: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 5.56, 7.91, 6.89

Statistical Properties:

When summarized with basic statistics (mean, variance, correlation, etc.), all four datasets appear very similar:

- **Mean of x**: 9.0
- **Variance of x**: 10.0
- **Mean of y**: Approximately 7.5
- **Variance of y**: Approximately 3.75
- **Correlation coefficient (r)**: Approximately 0.82
- **Linear regression line**: y=0.5x+3

Key Observations and Lessons:

1. **Visual Differences**: Despite identical summary statistics, each dataset exhibits different patterns when plotted. For instance:
    - **Dataset I**: Roughly linear relationship with a slight upward trend.
    - **Dataset II**: Appears as a curved pattern.
    - **Dataset III**: Dominated by one outlier, influencing the linear fit.
    - **Dataset IV**: Perfectly linear relationship except for one outlier.
2. **Impact of Outliers**: Dataset III demonstrates how an outlier can significantly affect the correlation coefficient and linear fit, despite other statistical measures being similar.
3. **Importance of Data Visualization**: Anscombe's quartet underscores the importance of visualizing data to gain insights beyond summary statistics. Graphical exploration can reveal patterns, outliers, and relationships that summary statistics might obscure.
4. **Statistical Fallacy**: Relying solely on summary statistics (mean, variance, correlation) can lead to misleading interpretations, as seen in datasets with vastly different distributions but similar summary statistics.

Implications for Data Analysis

Anscombe's quartet serves as a powerful reminder for data analysts to:

- Always visualize data before performing statistical analysis.
- Be cautious about relying solely on numerical summaries.
- Consider the potential impact of outliers on statistical results.
- Explore different types of relationships between variables.

3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as r, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between paired continuous data.

Pearson's correlation coefficient r is defined as the covariance of two variables X and Y divided by the product of their standard deviations. Mathematically, it is expressed as:

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

Where:

- x_i and y_i are the individual data points of variables X and Y,
- $\bar{x}$ and $\bar{y}$ are the mean values of X and Y respectively,
- n is the number of data points.

Properties:

1. **Range of r**: Pearson's r ranges from -1 to 1.

    - r = 1: Perfect positive linear correlation (as X increases, Y also increases proportionally).
    - r = −1: Perfect negative linear correlation (as X increases, Y decreases proportionally).
    - r = 0: No linear correlation between X and Y.
2. **Strength of Correlation**:
    - The closer r is to 1 or -1, the stronger the linear relationship between X and Y.
    - The closer r is to 0, the weaker the linear relationship (or no linear relationship).
3. **Direction**:
    - If r is positive, X and Y have a positive linear relationship (both variables increase or decrease together).
    - If r is negative, X and Y have a negative linear relationship (as one variable increases, the other decreases).
4. **Assumption**: Pearson's r assumes that the relationship between the variables is linear and that both variables are normally distributed.
5. **Sensitive to Outliers**: Pearson's r can be sensitive to outliers, especially when they significantly influence the covariance between X and Y.

Limitations:

- Pearson's r measures only linear relationships. It may not capture nonlinear relationships.
- It assumes that both variables are normally distributed and can be influenced by outliers.
- It does not imply causation; a strong correlation does not necessarily mean that changes in one variable cause changes in the other.

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

Scaling is a preprocessing step in data analysis and machine learning where numerical features are transformed to a standard scale, typically to make the data comparable and improve the performance and stability of certain algorithms. Here's a detailed explanation addressing your questions:

What is Scaling?

Scaling refers to the process of transforming the range of variables (features) to a similar scale. It is essential because many machine learning algorithms perform better or converge faster when features are on a relatively similar scale and close to normally distributed.

Why is Scaling Performed?

1. **Algorithm Performance**: Many machine learning algorithms, such as linear and logistic regression, K-nearest neighbors, support vector machines (SVMs), and neural networks, are sensitive to the scale of input features. Features with larger scales can dominate those with smaller scales, affecting the algorithm's ability to learn effectively.
2. **Convergence**: Gradient descent-based optimization algorithms, used in many machine learning models, converge faster when features are scaled. This is because the steps taken in each iteration are more consistent when the scales are similar.
3. **Distance-based Algorithms**: Algorithms that rely on distances between data points, such as K-nearest neighbors and clustering algorithms (e.g., K-means), are sensitive to the scale of features. Features with larger scales will have a greater impact on the distance calculations.

Types of Scaling:

1. **Normalized Scaling (Min-Max Scaling)**:
    - **Formula**: $X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$
    - **Range**: Transforms data to a scale between 0 and 1.
    - **Purpose**: Ensures all features have the same scale, preserving the shape of the original distribution. Useful when the data needs to be bound within a specific range.
2. **Standardized Scaling (Z-score Normalization)**:
    - **Formula**: $X_{std} = (X - \mu)/\sigma$
    - **Properties**: Transforms data to have a mean of 0 and a standard deviation of 1.
    - **Purpose**: Centres the data around zero and adjusts the variance, making it suitable for algorithms that assume normally distributed data, such as linear regression and linear discriminant analysis.

Differences between Normalized Scaling and Standardized Scaling:

- **Range of Values**:
    - **Normalized Scaling**: Values are scaled to a fixed range (e.g., 0 to 1).
    - **Standardized Scaling**: Values are cantered around 0 with a standard deviation of 1.
- **Impact on Distribution**:
    - **Normalized Scaling**: Preserves the shape of the original distribution, but may not handle outliers well.
    - **Standardized Scaling**: Does not bound values to a specific range, but is less sensitive to outliers and makes the data more suitable for algorithms that assume normally distributed features.
- **Algorithm Suitability**:

- **Normalized Scaling**: Often used when the algorithm (e.g., neural networks) requires inputs to be within a specific range.
- **Standardized Scaling**: Generally preferred for algorithms that assume normally distributed data, or when interpreting coefficients in linear models is important.

Considerations:

- **Choice of Scaling**: The choice between normalized and standardized scaling depends on the specific requirements of the machine learning algorithm and the characteristics of the data.
- **Impact on Interpretation**: Standardized scaling is often preferred when the interpretation of coefficients or feature importance is important, as the scaling does not change the relationship between variables, only their scales.

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

Variance Inflation Factor (VIF) is a measure used to quantify how much the variance of a regression coefficient is inflated due to collinearity with other predictor variables in a linear regression model. It assesses how much the variance of the estimated regression coefficients increases if your predictors are correlated.

Calculation of VIF:

For each predictor variable X_j, the VIF is calculated as:

$$\text{VIF}_j = \frac{1}{1-R_j^2}$$

Where:

- R_j^2 is the R^2 value obtained by regressing X_j against all other predictor variables X_{-j}.

Occurrence of Infinite VIF:

1. **Perfect Multicollinearity**: If a predictor variable X_j is perfectly linearly dependent on other predictor variables (i.e., X_j can be expressed as a perfect linear combination of other predictors), then R_j^2 will be 1.
2. **Calculation of VIF**: When $R_j^2=1$, the denominator $1-R_j^2$ becomes 0. This leads to:

   $VIF_j = 1/0 = \infty$

   Hence, the VIF for the predictor variable X_j becomes infinite.

Practical Implications:

- **Model Fitting**: Infinite VIF values indicate that one or more predictor variables are perfectly collinear with others. In such cases, the regression coefficients cannot be estimated uniquely because the model matrix is rank-deficient.
- **Interpretation**: Infinite VIFs make it impossible to interpret the regression coefficients properly because the model is overparameterized due to perfect multicollinearity.

Handling Infinite VIF:

To address infinite VIF values, consider the following steps:

1. **Identify and Remove Variables**: Identify variables that are causing perfect multicollinearity and consider removing one of them from the model.
2. **Combine Variables**: Sometimes, variables may be transformed or combined to remove multicollinearity. For

example, instead of using both height in centimeters and height in inches, use only one.

3. **Regularization Techniques**: Techniques like Ridge Regression or Lasso Regression can help in reducing multicollinearity by penalizing large coefficients.
4. **Principal Component Analysis (PCA)**: PCA can be used to reduce dimensionality and address multicollinearity by creating new orthogonal variables (principal components) that are uncorrelated.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical technique used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution.

Understanding Q-Q Plot:

1. **Purpose**:
    - **Distribution Assessment**: Q-Q plots are used to visually inspect if a dataset follows a specific distribution (e.g., normal distribution).
    - **Residual Analysis**: In linear regression, Q-Q plots of residuals (the differences between observed and predicted values) are used to check if the residuals are normally distributed.
2. **Construction**:
    - The Q-Q plot is constructed by:
        - Sorting the data points in ascending order.
        - Computing the quantiles for both the dataset and the theoretical distribution (e.g., normal distribution).

        - Plotting the quantiles of the dataset against the quantiles of the theoretical distribution.
3. **Interpretation**:
    - If the dataset follows the theoretical distribution (e.g., normal distribution), the points in the Q-Q plot should roughly follow a straight line.
    - Deviations from the straight line indicate departures from the assumed distribution.

Importance of Q-Q Plot in Linear Regression:

1. **Assumption Checking**:
    - **Normality of Residuals**: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed with mean 0 and constant variance (homoscedasticity). Q-Q plots of residuals help assess this assumption.
    - **Diagnostic Tool**: Q-Q plots provide a visual check to see if residuals exhibit patterns or deviations from normality, which can indicate potential issues with the regression model.
2. **Model Validity**:
    - A linear regression model assumes that residuals are normally distributed. Deviations from normality can affect the validity of statistical inference, such as hypothesis testing and confidence intervals.
    - Q-Q plots help identify situations where non-normality of residuals might affect the reliability of model predictions and interpretations.
3. **Decision Making**:
    - Based on the Q-Q plot:
        - If residuals closely follow a straight line, it suggests that the assumption of normality is reasonable, and the linear regression model is appropriate for inference.

    - If residuals deviate significantly from the straight line, it may indicate that the linear regression assumptions are violated, and further investigation or model refinement is necessary.

Practical Use:

- **Model Refinement**: If Q-Q plots reveal non-normality in residuals, transformations (e.g., logarithmic transformation) or alternative modeling approaches (e.g., generalized linear models) may be considered to address the issue.
- **Diagnostic Tool**: Q-Q plots complement other diagnostic tools in regression analysis, such as residual plots and tests for homoscedasticity, providing a comprehensive assessment of the model's assumptions.