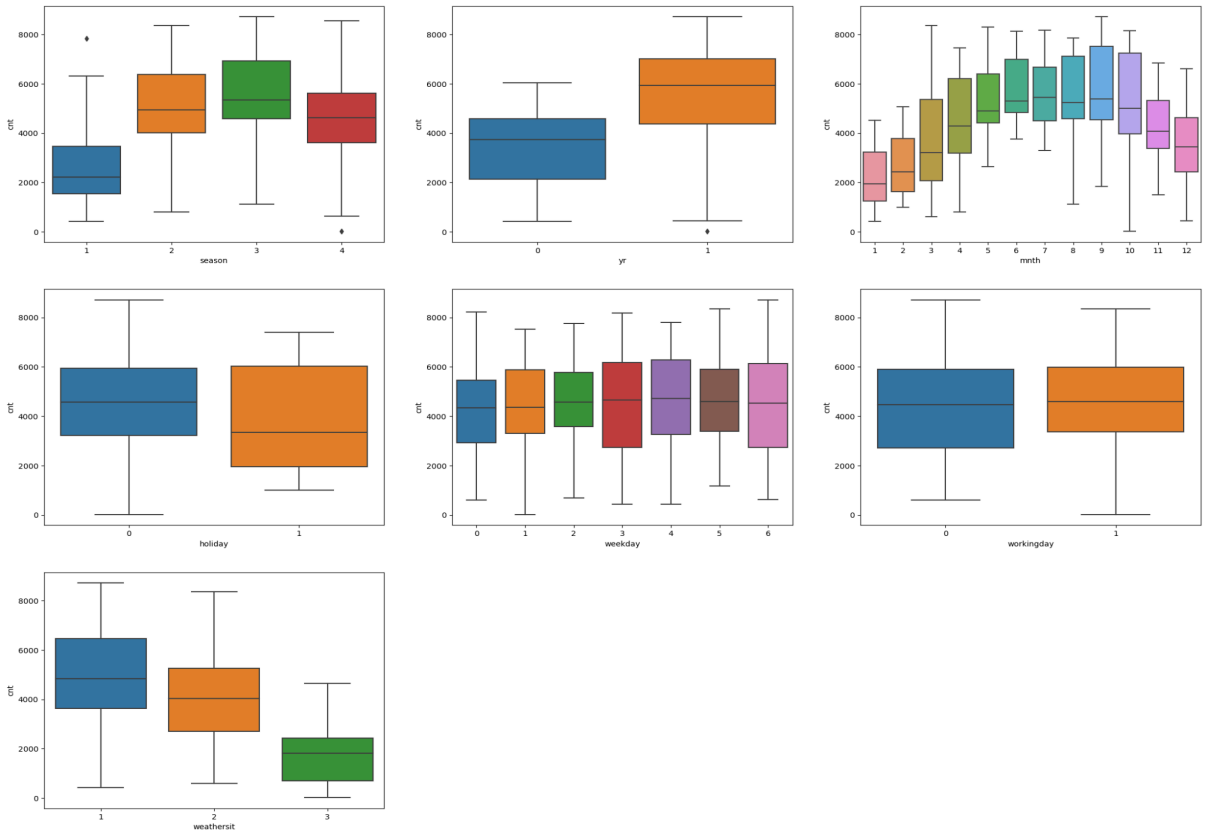
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**



The categorical variables in the dataset were season, yr, holiday, weekday, workingday, and weathersit and mnth. These were visualized using a boxplot (Fig. attached) .

These variables had the following effect on our dependent variable:-

- Season - The boxplot showed that the spring season had least value of cnt whereas the fall had the maximum value of cnt. Summer and winter had intermediate values of cnt.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable. The highest count was seen when the weathersit was' Clear, Partly Cloudy'.
- Yr - The number of rentals in 2019 was more than in 2018.
- Holiday - rentals are reduced during the holiday.
- Mnth - September saw the highest no of rentals while December saw least. This observation is in accordance with the observation made in weathersit. The weather situation in December is usually heavy snow due to which the rentals might have dropped.
- Weekday - The count of rentals is almost even throughout the week.
- Workingday – The median count of users is constant almost throughout the week.

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

Setting `drop_first=True` during dummy variable creation is important to avoid multicollinearity issues in regression analysis. When creating dummy variables for categorical variables, if we include all of the dummy variables, it introduces perfect multicollinearity because one category can be perfectly predicted from the others.

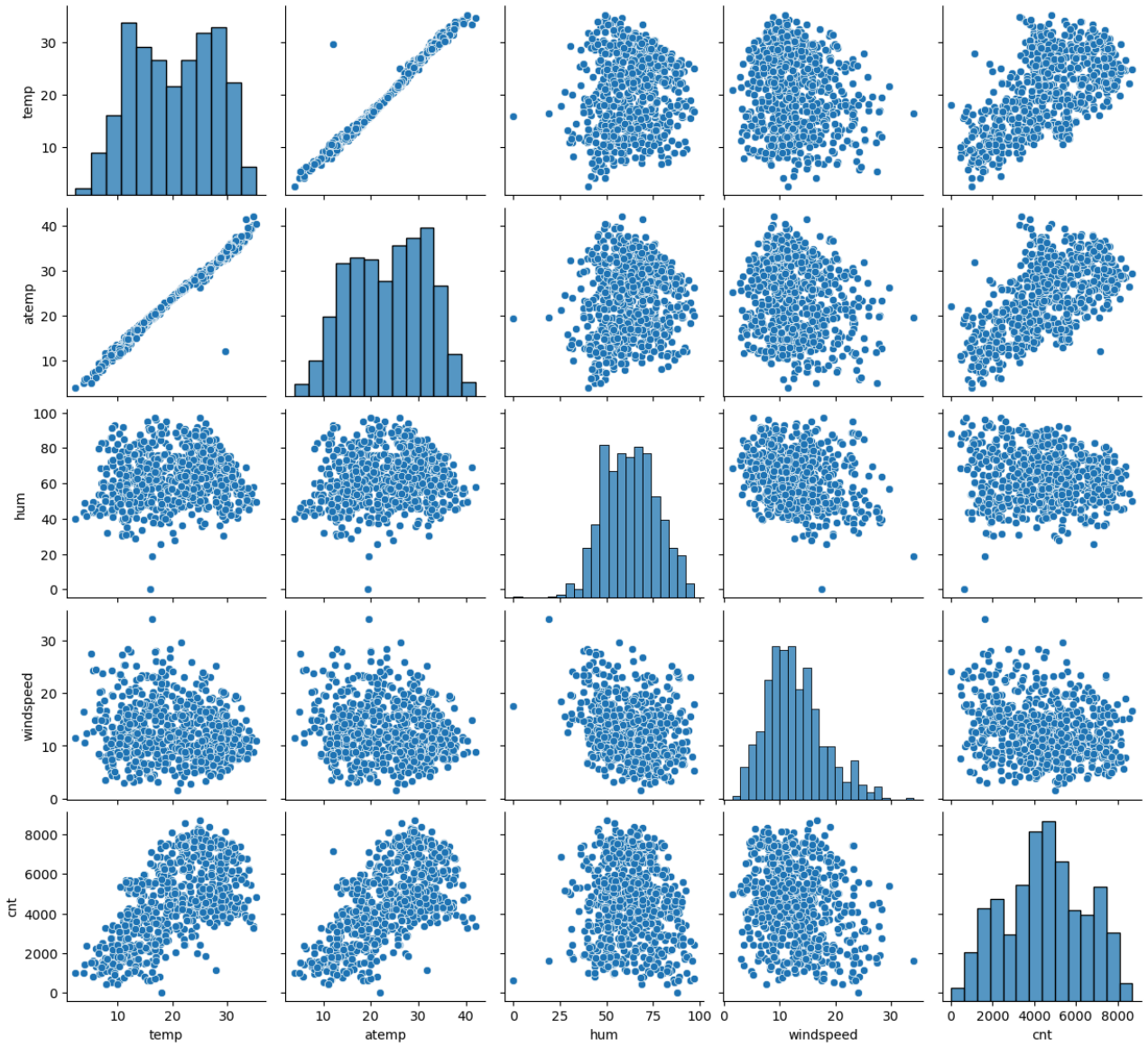
For example, consider a categorical variable "color" with three categories: red, green, and blue. If we create dummy variables for each category without dropping the first one, we would end up with two dummy variables: "green" and "blue". However, having both "green" and "blue" as predictors in the model inherently implies the presence of "red" (since the sum of probabilities for all categories must be 1). This creates perfect multicollinearity, where one dummy variable's value can be perfectly predicted from the others.

By dropping the first dummy variable (i.e., using `drop_first=True`), we eliminate this multicollinearity issue. It effectively makes one category the reference category, and the model includes dummy variables for all other categories except the reference one. This way, each category's effect is measured relative to the reference category, avoiding redundancy and multicollinearity.

Overall, using `drop_first=True` helps to ensure the model's stability and interpretability by mitigating multicollinearity and improving the efficiency of the regression analysis.

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

Using the below pairplot it can be seen that, “temp” and “atemp” are the two numerical variables that are highly correlated with the target variable (cnt)

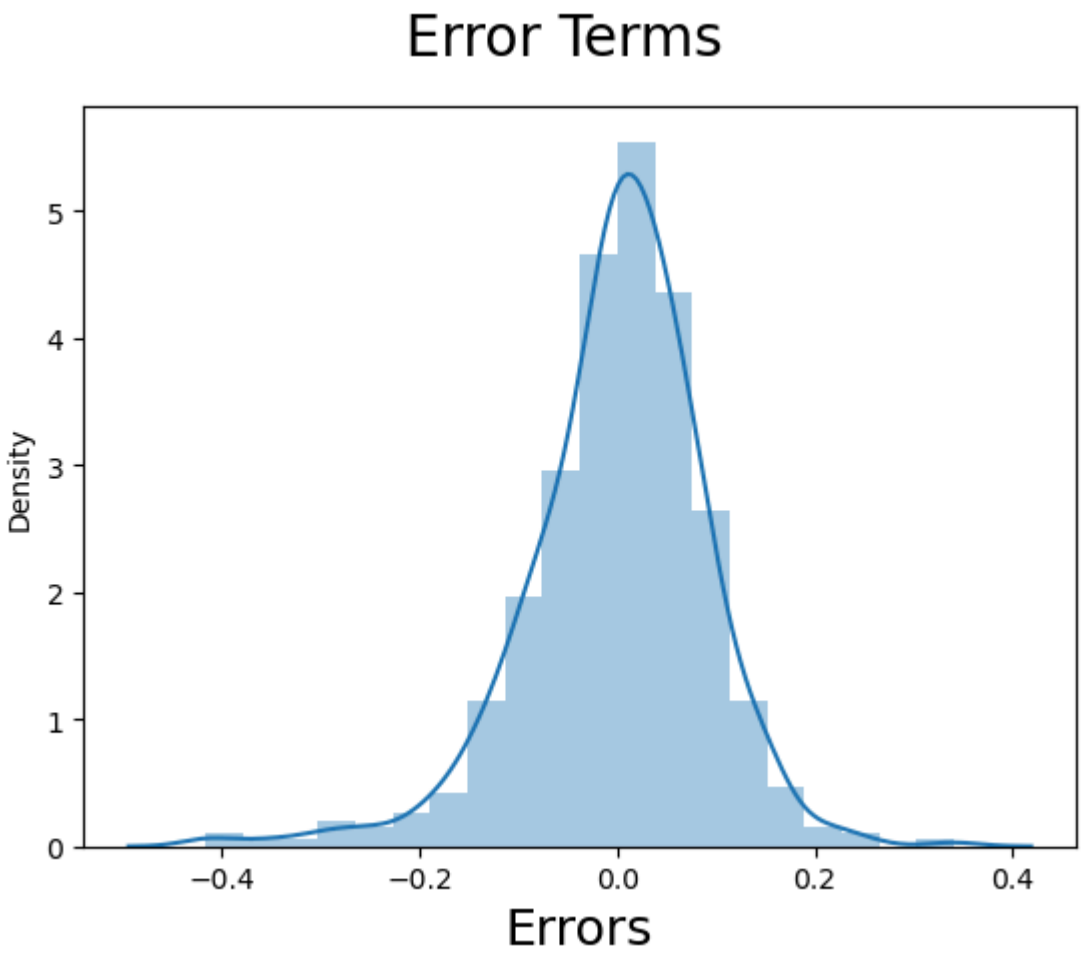


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The following tests were done to validate the assumptions of linear regression:

1. First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. Refer to the notebook for more details.

2. Secondly, Residuals distribution should follow a normal distribution and centered around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.



3. Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get a quantitative idea of how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

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

The top 3 features are:
1. temp - coefficient : 0.493363
2. yr - coefficient : 0.234126
3. weathersit_Light Snow & Rain - coefficient : -0.285761

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The basic idea is to find the best-fitting straight line through the data points.

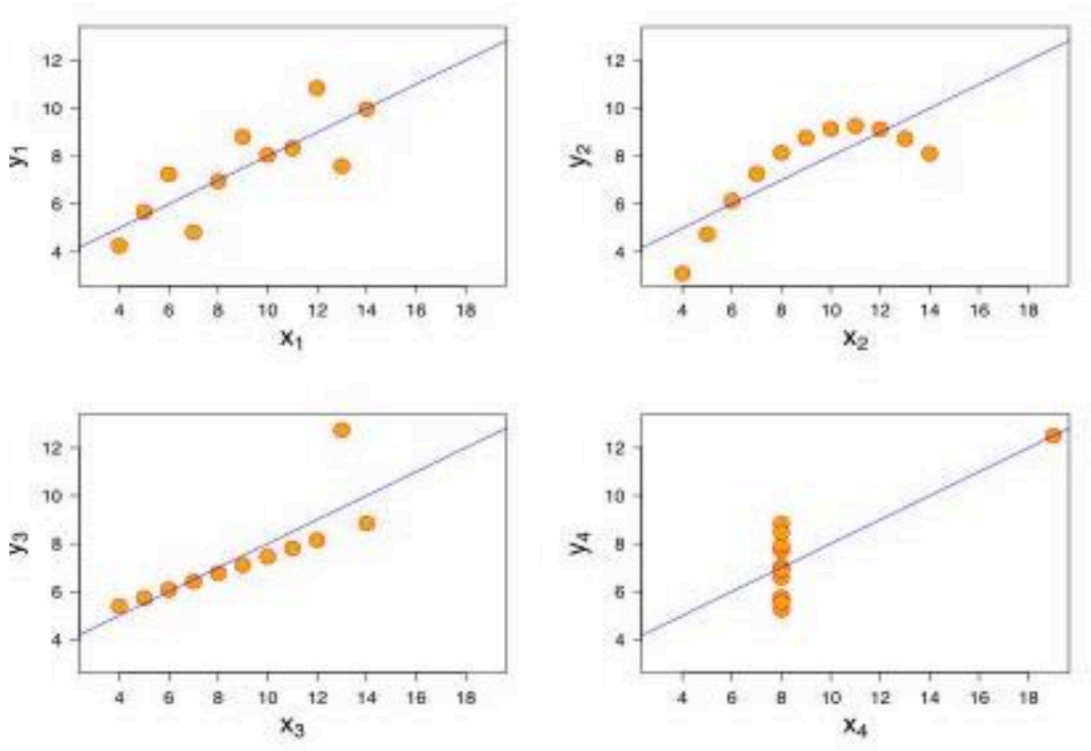
Here's a step-by-step explanation of how the linear regression algorithm works:

1. **Model Representation**: In linear regression, we model the expected value of the dependent variable (y) as a linear combination of the independent variables (x_1, x_2, …, x_n), often with a constant term (b) added (known as the intercept). The linear equation can be written as:
   y = b + m_1x_1 + m_2x_2 + ... + m_nx_n
2. **Best Fit Line**: The algorithm tries to find the values of the coefficients (m_1, m_2, …, m_n) and the intercept (b) that minimize the difference between the predicted values and the actual data points. This line is called the "best fit" or regression line.
3. **Cost Function**: To find the best coefficients, linear regression uses a cost function known as Mean Squared Error (MSE), which calculates the average of the squares of the errors (the differences between the predicted and actual values).
4. **Gradient Descent**: This is an optimization algorithm used to minimize the cost function. It iteratively adjusts the coefficients (m_1, m_2, …, m_n) and the intercept (b) to find the values that result in the smallest possible MSE.
5. **Prediction**: Once the model has been trained (i.e., the coefficients have been found), it can be used to make predictions. You simply plug in the values of the independent variables into the regression equation to get the predicted value of the dependent variable.
6. **Evaluation**: The performance of the regression model can be evaluated using metrics such as R-squared, which measures how well the observed outcomes are replicated by the model.
7. **Assumptions**: Linear regression assumes that there is a linear relationship between the independent and dependent variables, the residuals (prediction errors) are normally distributed, and there is no multicollinearity among the independent variables.
8. **Implementation**: Linear regression can be implemented from scratch using mathematical equations or with the help of libraries like scikit-learn in Python, which provide built-in functions for fitting the model to data.

Linear regression is widely used because of its simplicity and interpretability. It's a good starting point for regression tasks and can provide insights into the importance of different variables for prediction. However, it has limitations when dealing with non-linear relationships or when the assumptions mentioned above are not met.

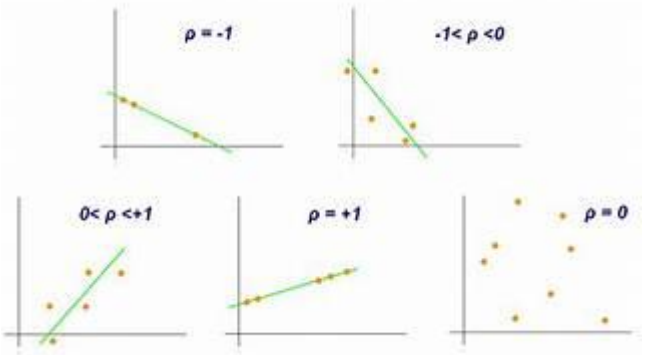
2. Explain the Anscombe's quartet in detail.

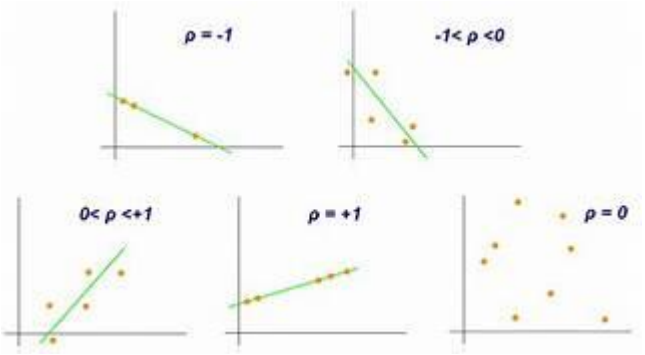
Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough
- influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?





The **Pearson correlation coefficient ®** is a widely used statistical measure that quantifies the **strength and direction** of the **linear relationship** between **two quantitative variables**. Let's dive into the details:

1. **Definition**:
    - The Pearson correlation coefficient is also known as:
        - **Pearson's r**
        - **Bivariate correlation**
        - **Pearson product-moment correlation coefficient (PPMCC)**
    - It ranges between **–1 and 1**.
    - The sign indicates the direction of the relationship:
        - **Positive correlation**: When one variable increases, the other tends to increase as well.
        - **Negative correlation**: When one variable increases, the other tends to decrease.
        - **No correlation**: When there's no clear relationship between the variables.
2. **Interpretation**:
    - The following general rules of thumb apply to the magnitude of the correlation coefficient:
        - **Strong positive correlation**: ($r > 0.5$)
        - **Moderate positive correlation**: ($0.3 < r \leq 0.5$)
        - **Weak positive correlation**: ($0 < r \leq 0.3$)
        - **No correlation**: ($r = 0$)
        - **Weak negative correlation**: ($-0.3 \leq r < 0$)
        - **Moderate negative correlation**: ($-0.5 \leq r < -0.3$)
        - **Strong negative correlation**: ($r < -0.5$)
3. **Inference**:
    - The Pearson correlation coefficient is not only descriptive but also inferential.
    - It helps test whether there's a **significant relationship** between the two variables.
4. **Visual Representation**:
    - Imagine a scatter plot of data points.
    - The closer the points align to a **straight line**, the higher the absolute value of (r).
    - The slope of the line indicates the direction of the relationship.

Remember, the Pearson correlation coefficient provides valuable insights into how variables move together, but it assumes a **linear relationship**. If the relationship is nonlinear, other correlation measures may be more appropriate.

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

Scaling is a crucial data preprocessing step in machine learning. It ensures that features are on a similar scale, making algorithms perform better and converge faster. Let's explore the concepts of scaling, normalized scaling, and standardized scaling:

Scaling:

- **Definition**: Scaling refers to adjusting the range or spread of values in a dataset.
- **Purpose**: It makes data more manageable and comparable, especially when dealing with variables of different units or vastly different ranges.
- **Importance**:
    - Algorithms computing distances between features are biased toward numerically larger values if data is not scaled.
    - Tree-based algorithms are less sensitive to feature scale.
    - Scaling helps machine learning and deep learning algorithms train effectively.
- **Methods**: Two common scaling techniques are **Normalization** and **Standardization**.

Normalized Scaling (Min-Max Scaling):

- **Formula**:
    - Given a feature value (X), the normalized value (X_{\text{new}}) is calculated as: [X_{\text{new}} = \frac{{X - X_{\text{min}}}}{{X_{\text{max}} - X_{\text{min}}}}]
- **Range**: Scales values between ([0, 1]) or sometimes ([-1, 1]).
- **Geometric Interpretation**:
    - Transforms n-dimensional data into an n-dimensional unit hypercube.
- **Use Case**:
    - Useful when there are no outliers.
    - Example: Scaling age (uniform distribution) but not incomes (few high values).

Standardized Scaling (Z-Score Normalization):

- **Formula**:
    - Given a feature value (X), the standardized value (X_{\text{new}}) is calculated as: [X_{\text{new}} = \frac{{X - \text{mean}}}{{\text{Std}}}]
- **Properties**:
    - Quantifies the difference between values.
    - Translates data to the mean vector of the original data (mean = 0) and adjusts for standard deviation.
- **Use Case**:

- Helpful when data follows a Gaussian distribution (but not strictly necessary).
- Robust to outliers due to no predefined range of transformed features.

Differences:

- **Normalization**:
  - Uses minimum and maximum values for scaling.
  - Suitable for features with different scales.
  - Scales values between ([0, 1]) or ([-1, 1]).
  - Highly affected by outliers.
  - Scikit-Learn provides MinMaxScaler.
- **Standardization**:
  - Uses mean and standard deviation for scaling.
  - Ensures zero mean and unit standard deviation.
  - Not bounded to a specific range.
  - Less affected by outliers.
  - Scikit-Learn provides StandardScaler.

Choose the appropriate scaling method based on your data distribution and the requirements of your machine learning model.

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

VIF - Variance Inflation Factor

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity".

The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e.the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

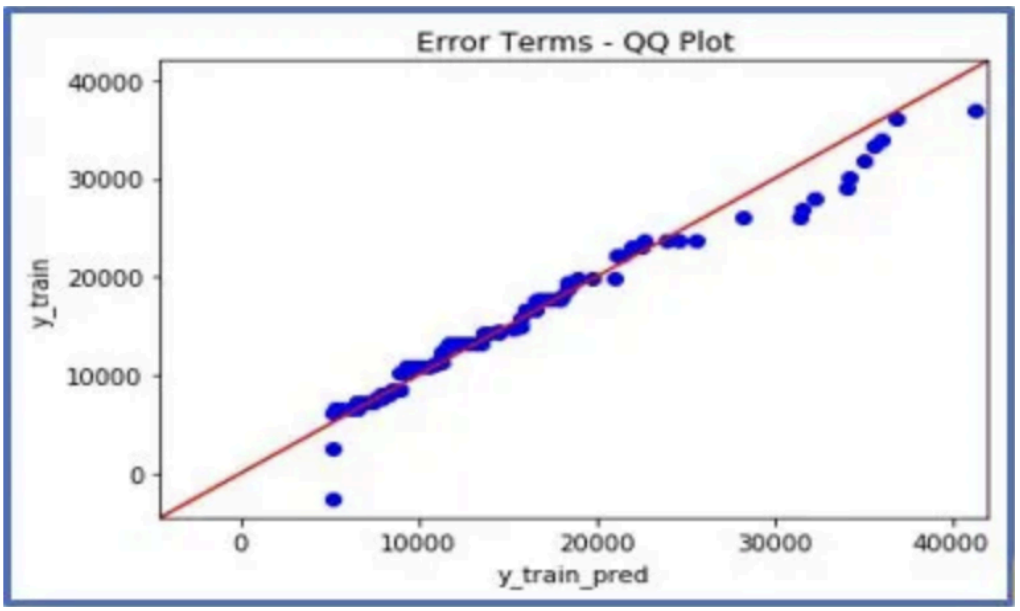
Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

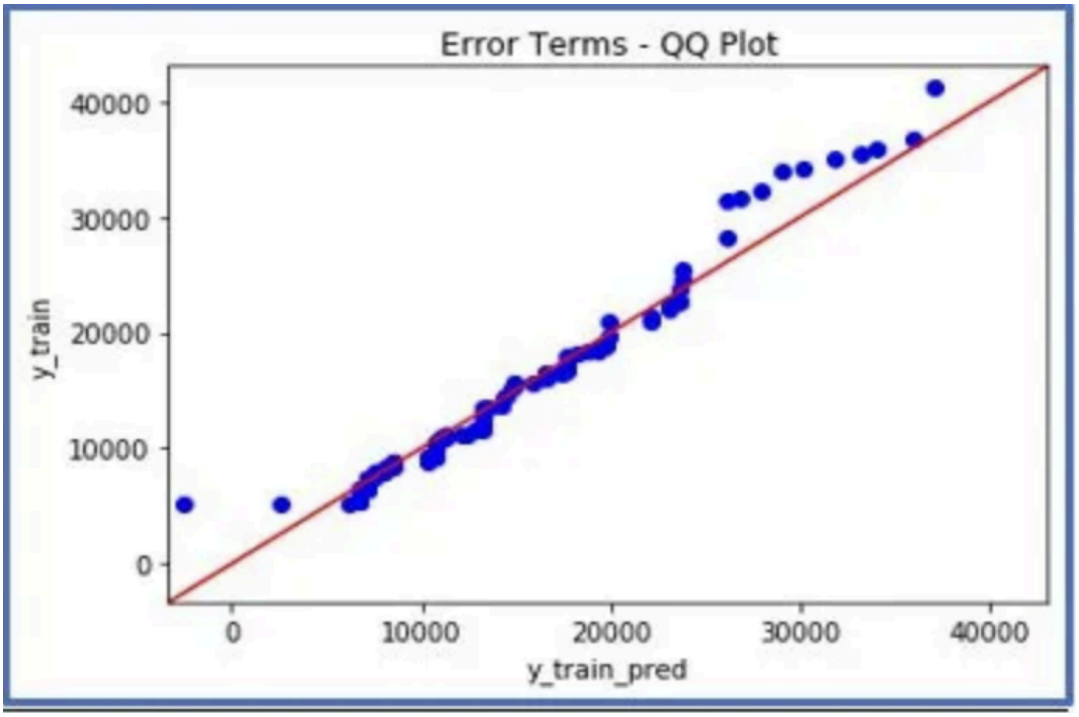
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

a) ***Similar distribution****: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis*

b) ***Y-values < X-values:*** *If y-quantiles are lower than the x-quantiles.*



c) X-values < Y-values: *If x-quantiles are lower than the y-quantiles.*



d) ***Different distribution:*** *If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*