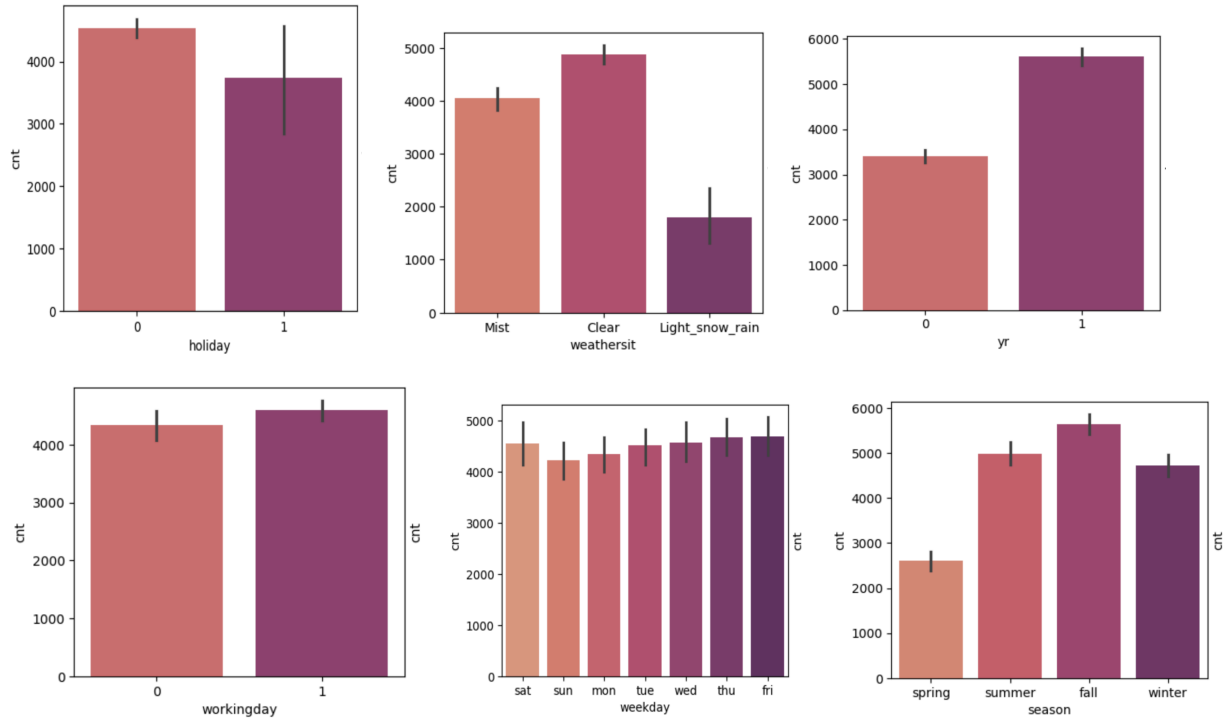
Assignment-based Subjective Questions

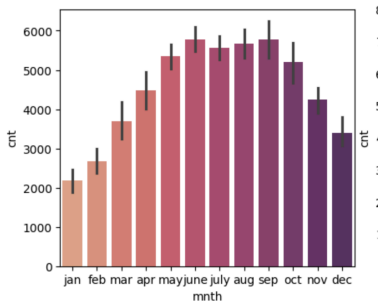
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

The categorical variables in the dataset were season , yr , holiday, weekday ,workingday, weathersit and mnth. These were visualized using a barplot(Fig. attached) . These variables had the following effect on our dependant variable total bike count (cnt):-

- Season - The plot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate values of cnt.
- Weathersit - Highest count was seen when the weathersit was' Clear'.
  There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable for bike rentals.
- Yr - The number of rentals has increased in 2019 as compared to 2018.
- Holiday - Bike Rentals are reduced during holiday.
- Mnth - September saw highest no of rentals while December saw least.
- Weekday - The count of rentals is almost even throughout the week
- Workingday – The median count of users is constant almost throughout the week

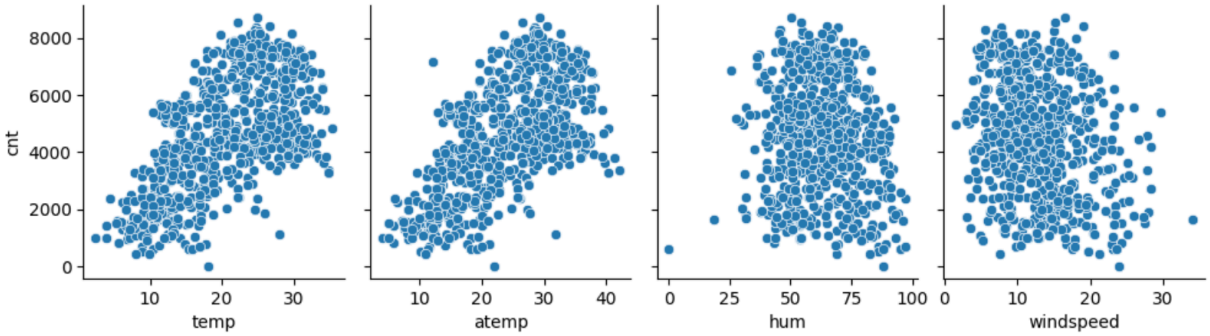




2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer:
The drop_first =True is important as it helps in reducing the extra column created during creation of the dummy variable. Dropping the column is important because the importance or value of that left over variable can be found by remaining variables. So to avoid redundancy we are dropping a column. This helps the column to become linearly independent.

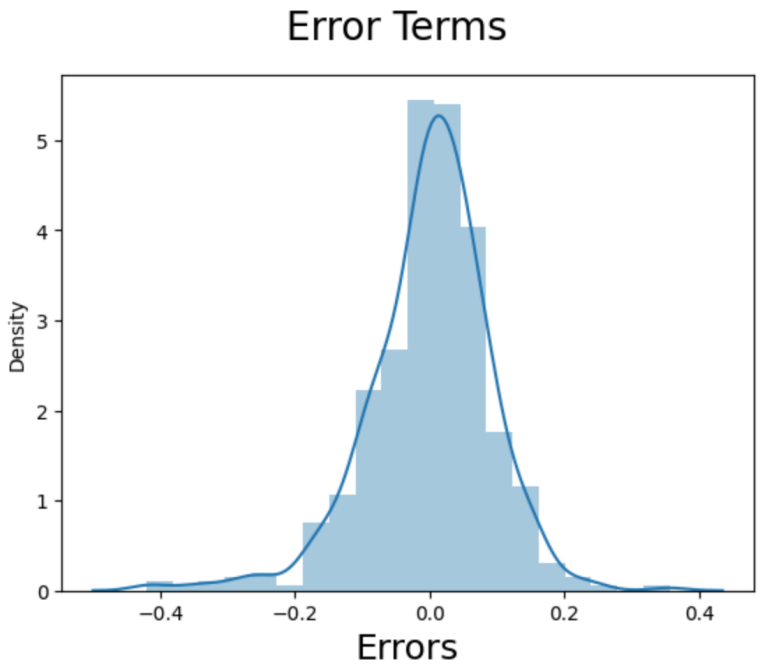
Example: Let's say we have 3 types of values in the categorical column 'Color' with values: Red , Blue and Green. We want to create a dummy variable for that column. If one variable is not Red and Blue, then It is obviously Green. So we do not need the 3rd variable to identify the column as Green.
For a categorical variable with n categories, we typically create n − 1 dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer:
Using the below pairplot it can be seen that , "temp" is the numerical variables which is highly correlated with the target variable (cnt).
Given that 'atemp' and 'temp' are highly correlated variables, so only one of them is selected during the determination of the best fit line



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
The following tests were done to validate the assumptions of linear regression:

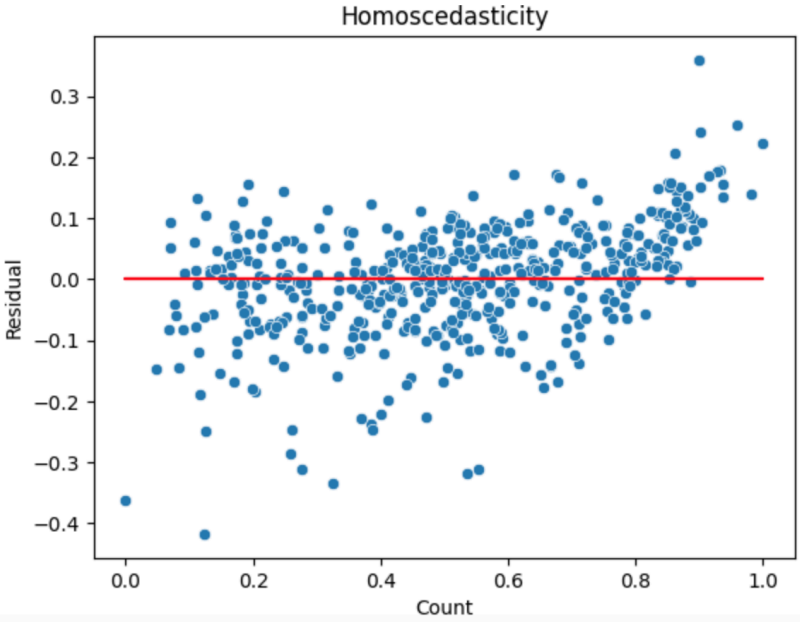
• **Residual Analysis**- Error term should be normally distributed.



• **Multicollinearity check**- There should be no multicollinearity among variables.We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model.VIF values should be below a certain threshold (5) to ensure no multicollinearity.

• **Homoscedasticity**- There should be no visible pattern in residual values.

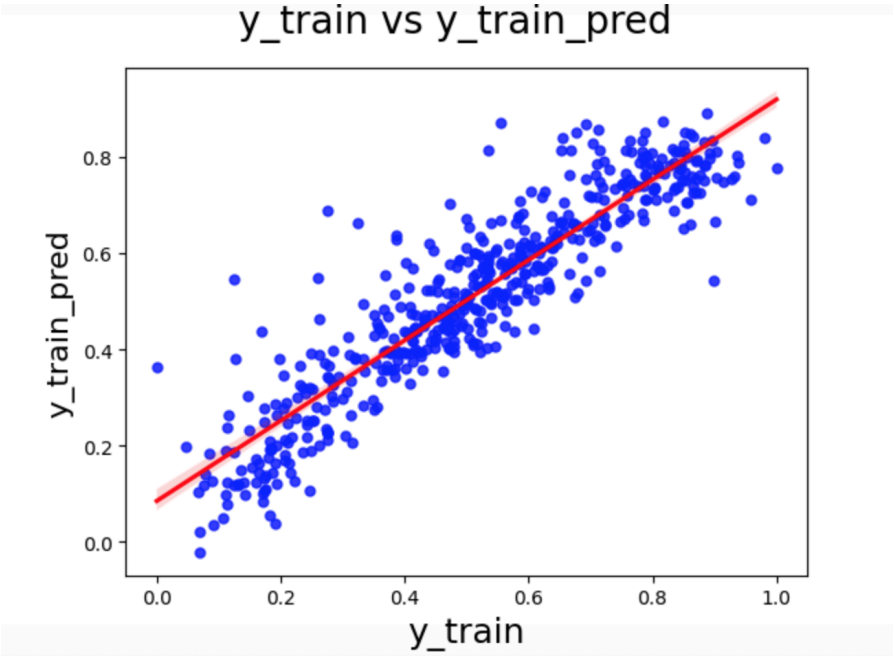
The spread of residuals should be roughly constant across all levels of the predicted values.



• **Independence of residuals** - The observations are independent.

• **Linear relationship validation-** Linearity should be visible among variables.

The points should fall approximately along a diagonal line, indicating a linear relationship.



- **<u>Cross-Validation:</u>** Validated the model on a test dataset and assess the model's performance on new data to ensure generalizability and consistency. Evaluate model performance on a test set to ensure that the model generalizes well to new, unseen data without overfitting the training set.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:
The equation of the best fit line is given by:
cnt = 0.1910 + 0.2343 * yr + 0.4799 * temp - 0.1499 * windspeed + 0.0613 * season_summer + 0.0951 * season_winter + 0.0853 * month_sep -0.0462 * weekday_sun -0.0570 * season_spring - 0.2865* weathersit_Light_snow_rain - 0.0803 * weathersit_Mist

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes are:
- Temperature (temp)
- year (year)
- weathersit_Light_snow_rain

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It is widely used for predicting the value of the dependent variable based on the values of one or more independent variables. The basic idea is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that minimizes the sum of the squared differences between the observed and predicted values of the dependent variable.

Regression is broadly divided into simple linear regression and multiple linear regression:
1. **Simple Linear Regression** : SLR is used when the dependent variable is predicted using only one independent variable. The equation for SLR will be:

$y= a_0+a_1x+ \varepsilon$

Where,
a0= It is the intercept of the Regression line (can be obtained putting x=0)
a1= It is the slope of the regression line, which tells whether the line is increasing or decreasing.
ε = The error term. (For a good model it will be negligible)
y is the dependent variable we are trying to predict

2. **Multiple Linear Regression** :MLR is used when the dependent variable is predicted using multiple independent variables. The equation for MLR will be:

$y= a_0+a_1x+ a_2x_2+.....a_kx_k + \varepsilon$

where: - x1,x2,.....xk are the independent variables, and
a_0,a_1,a_2, a_k are the coefficients.
ε = The error term
y is the dependent variable we are trying to predict

Assumptions -
The following are some assumptions about dataset that is made by Linear Regression model

- Multicollinearity – Linear regression model assumes that there is very little or no multicollinearity in the data. Basically, multi- collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation – Linear regression model assumes that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms – Error terms should be normally distributed.
- Homoscedasticity -There should be no visible pattern in residual values.

Linear regression relies on the assumption of a linear relationship between independent and dependent variables, normally distributed errors, constant error variance (homoscedasticity),

and the absence of perfect multicollinearity, ensuring that there is no perfect linear relationship among the predictors.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer

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.

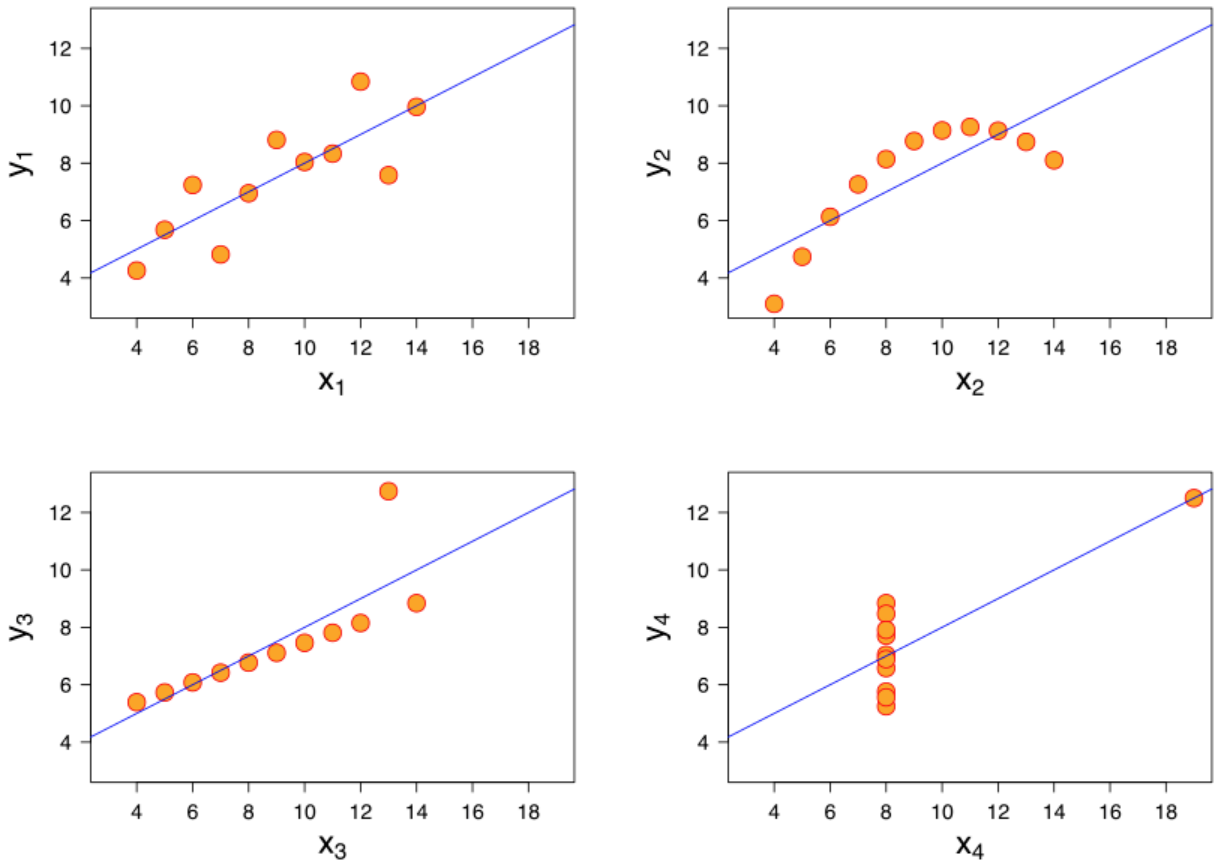
The summary statistics show that the means and the variances were identical for x and y across the groups:

1- Mean of x is 9 and mean of y is 7.50 for each dataset.

2- The variance of x is 11 and variance of y is 4.13 for each dataset

3- The correlation coefficient between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



Dataset I appears to have clean and well-fitting linear models.

Dataset II appears to have a non-linear relationship between x and y.
Dataset III the distribution is perfectly linear except for one large outlier.
Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R? (3 marks)
Answer:

Pearson's correlation coefficient is a measure of the linear relationship between two variables. It quantifies the strength and direction of a linear association between two continuous variables.

Formula

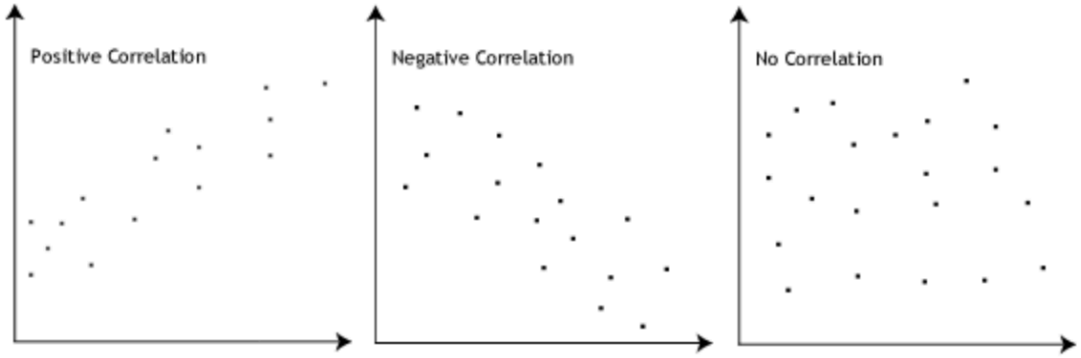
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

The coefficient takes values between -1 and 1, where:

- r = 1: Perfect positive linear correlation.
- r = −1: Perfect negative linear correlation
- r = 0: No linear correlation.

A value greater than 0 indicates a positive association: as the value of one variable increases, so does the value of the other variable.
A value less than 0 indicates a negative association: as the value of one variable increases, the value of the other variable decreases.
This is shown in the diagram below:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.
Example: If an algorithm is not using feature scaling method then it can consider the value 300 grams to be greater than 5 kg but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes with the help of scaling.

1. **Normalized Scaling (Min Max Scaling)**: - Scales the values of a variable to a specific range, usually [0, 1]. -
   Formula:

$$X_{\text{normalized}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

   Advantages: Useful when the distribution of the variable is unknown or not Gaussian.
   Disadvantages: Sensitive to outliers.
   Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

2. **Standardized Scaling:**
   Scales the values to have a mean of 0 and a standard deviation of 1.
   Formula:

$$X_{\text{standardized}} = \frac{X - \text{mean}\,(X)}{\text{std}(X)}$$

   Advantages: Less sensitive to outliers and preserves the shape of the distribution.

Disadvantages: Assumes that the variable follows a Gaussian distribution
Scikit-Learn provides a transformer called StandardScaler for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer

VIF - Variance Inflation Factor gives how much the variance of the coefficient estimate is being increased 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.

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

A large value of VIF indicates that there is a correlation between the variables. When the value of VIF is infinite it shows a perfect correlation between two independent variables.
In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/ (1-R2) = infinity. Thus if there is perfect correlation, then VIF = infinity.

To address this issue, it's crucial to identify and handle multicollinearity in the dataset. This can involve removing one of the perfectly correlated variables, combining them or using dimensionality reduction techniques. Addressing multicollinearity not only resolves the infinite VIF problem but also improves the overall regression model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions .A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Use of QQ plot:
It is used to check whether 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

Importance of QQ plot:

Identifying Outliers:Outliers in the residuals can be detected by examining points that deviate from the expected straight line in the Q-Q plot.

Model Fit Assessment: Q-Q plots provide a visual assessment of how well the residuals follow a normal distribution.

Normality Assessment: Q-Q plots are valuable for checking that residuals are normally distributed.

Comparing Distributions: Q-Q plots can be used to compare two datasets to see if they belong to same distribution

Q-Q plots are powerful diagnosis tools in linear regression for assessing the normality of residuals, identifying outliers, and ensuring the validity of statistical inferences. They provide a visual and intuitive way to check the assumptions underlying the regression model.