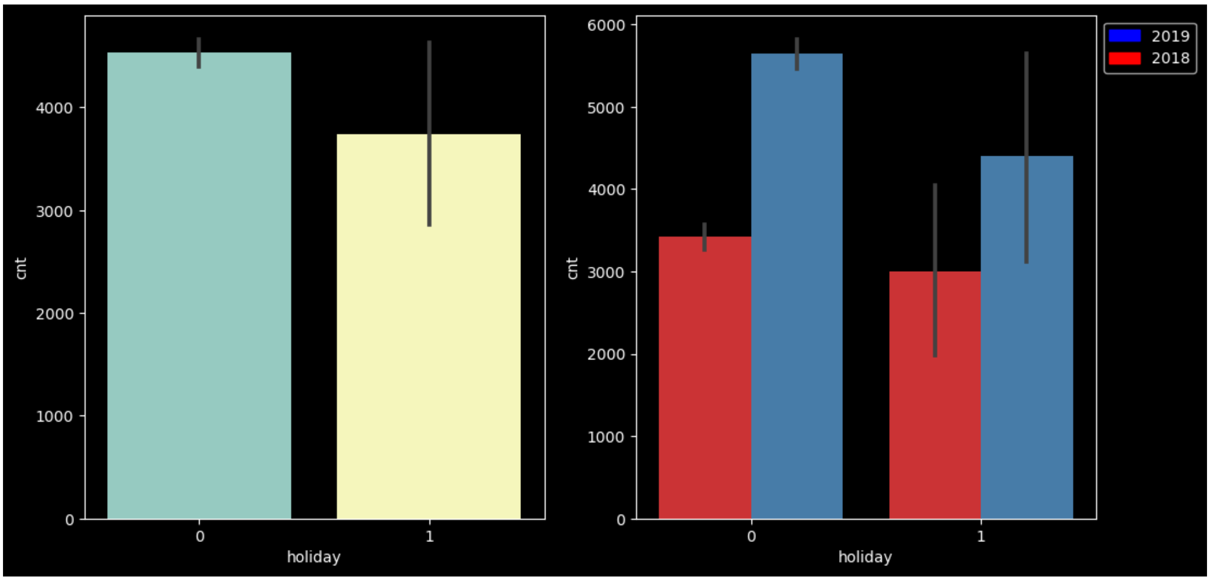
Assignment-based Subjective Questions

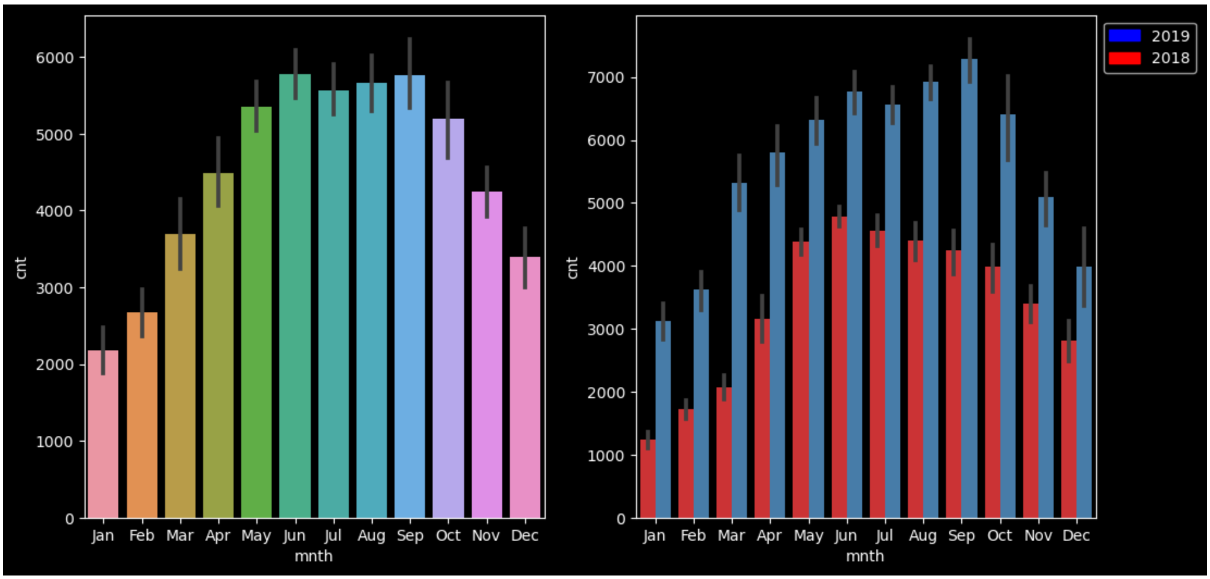
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

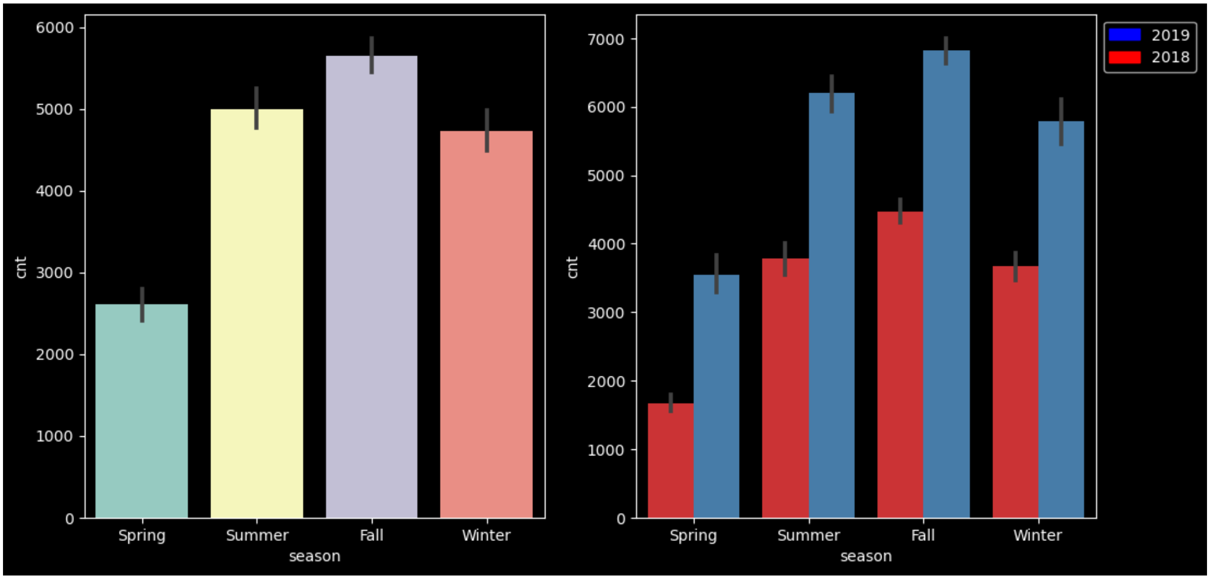
Answer:

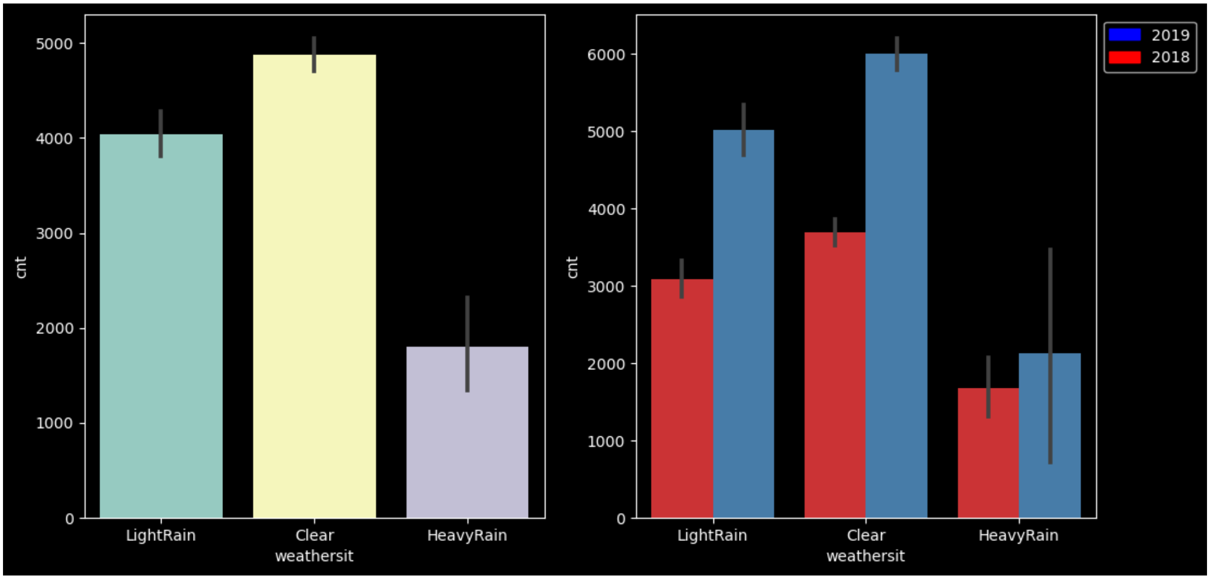
Based on the data understanding we got 6 different categorical variables (holiday, month, season, weathersit, weekday, working day) , as shown in the below graphs mostly all the categorical variables having the positive co-relation with the used bike count, and compared to 2018 , usage of boom bikes has increased in 2019 with respect to all the categorical variables,

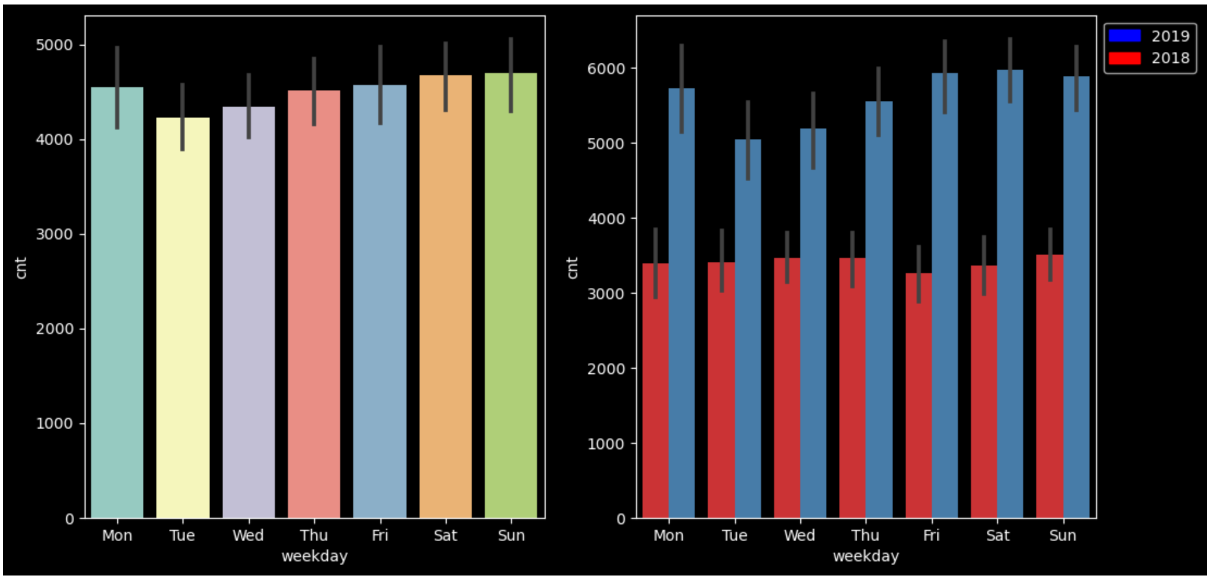
- Mostly June to August months having highest count of usage.
- Mostly in summer and fall seasons people having more interest to use the bikes.
- Mostly clear and light rain also having more count.
- People are also interested in both holidays and working days

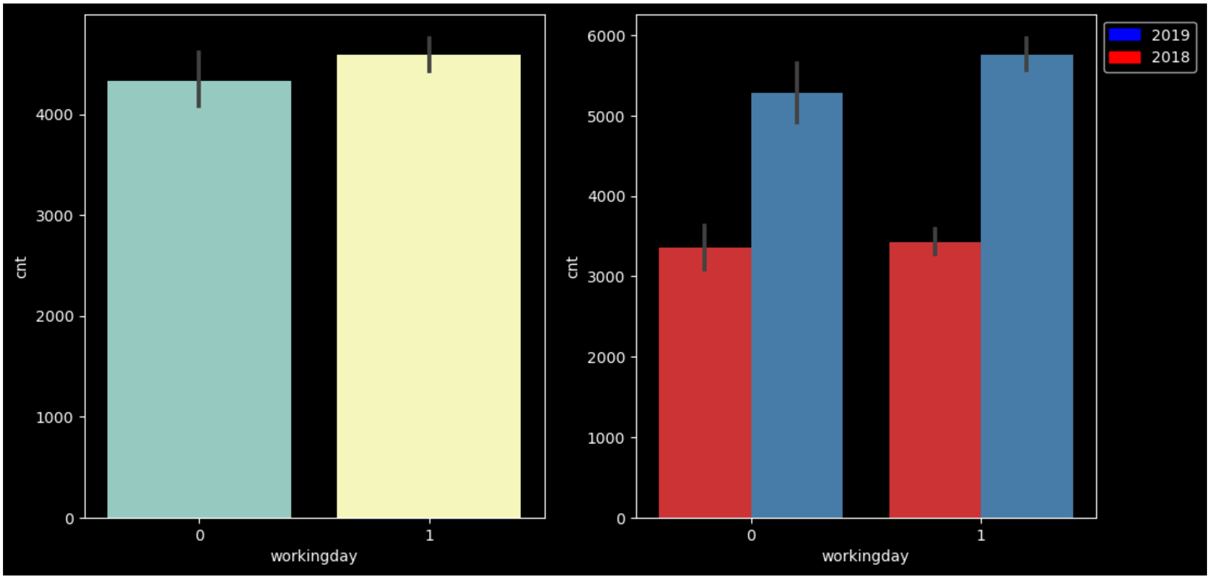



cnt
0
1000
2000
3000
4000
5000
6000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
mnth
cnt
7000
2019
2018
mnth


cnt
0
1000
2000
3000
4000
5000
6000
Spring
Summer
Fall
Winter
season
cnt
7000
2019
2018
season


cnt
0
1000
2000
3000
4000
5000
LightRain
Clear
HeavyRain
weathersit
cnt
6000
2019
2018
weathersit





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

Answer:

Using drop_first=True when creating dummy variables (one-hot encoding) is important because it helps to avoid the "dummy variable trap," which occurs when there is perfect multicollinearity in the dataset. Perfect multicollinearity happens when one predictor variable in a regression model can be perfectly predicted by the other predictors. Here's why drop_first=True is important:

→**Avoiding Multicollinearity**:

- **Perfect Multicollinearity**: When all categories of a categorical variable are included as dummy variables, one of the dummy variables can be perfectly predicted by the others. This leads to perfect multicollinearity, which can make the regression coefficients unstable and their interpretation difficult.

- **Example**: If you have a categorical variable "Color" with three categories: Red, Blue, and Green, creating dummy variables for all three will result in perfect multicollinearity because if an observation is not Red and not Blue, it must be Green.

→**Model Interpretation**:

- Dropping the first category (or any one category) sets a reference level. The coefficients of the remaining dummy variables represent the effect of each category relative to this reference level.
- **Example**: If you drop the "Red" category, the coefficients of the "Blue" and "Green" dummy variables will show how the response variable differs for Blue and Green compared to Red.

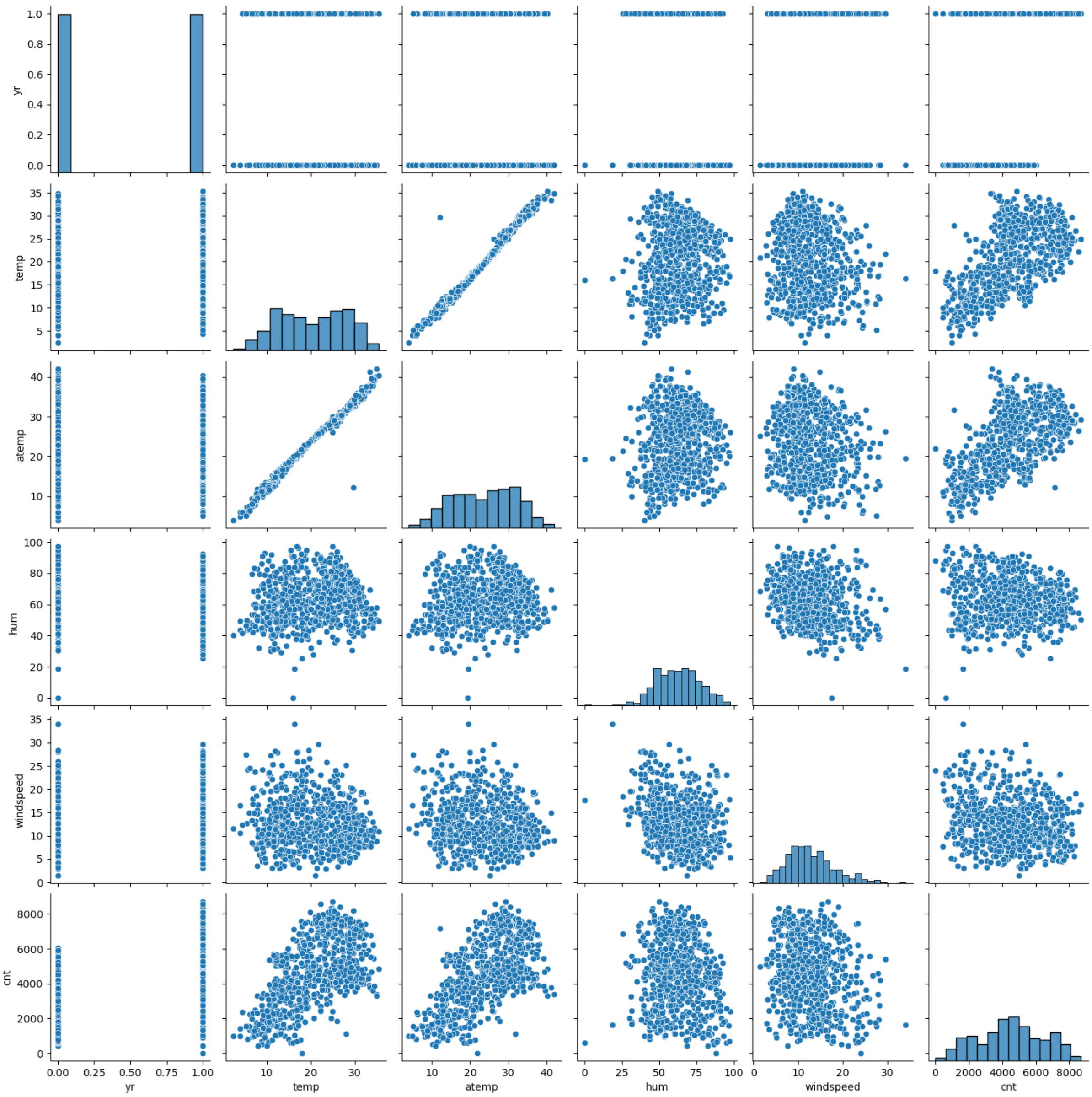
→**Degrees of Freedom**:

- Including all categories reduces the degrees of freedom available for the model, which can affect the statistical significance of the predictors.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

Based on the below graph temp and atemp numerical variable having highest correlation with the target variable.

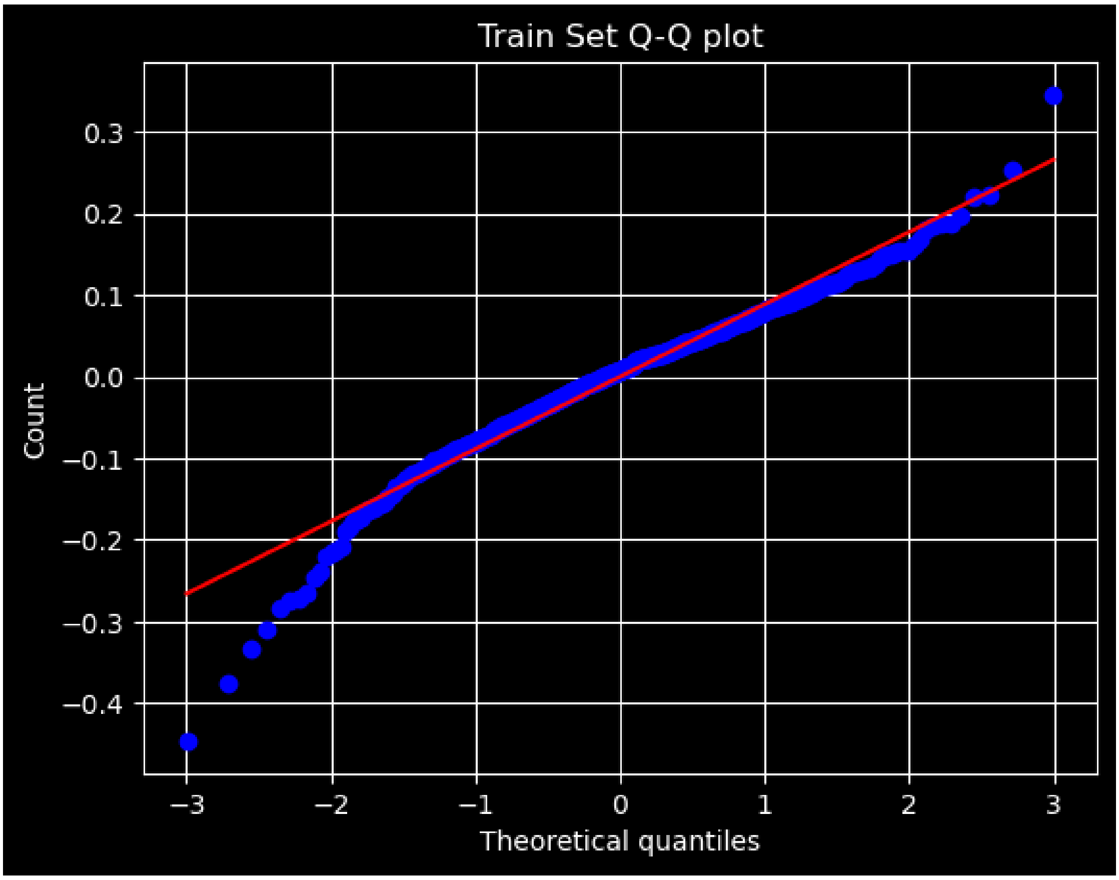


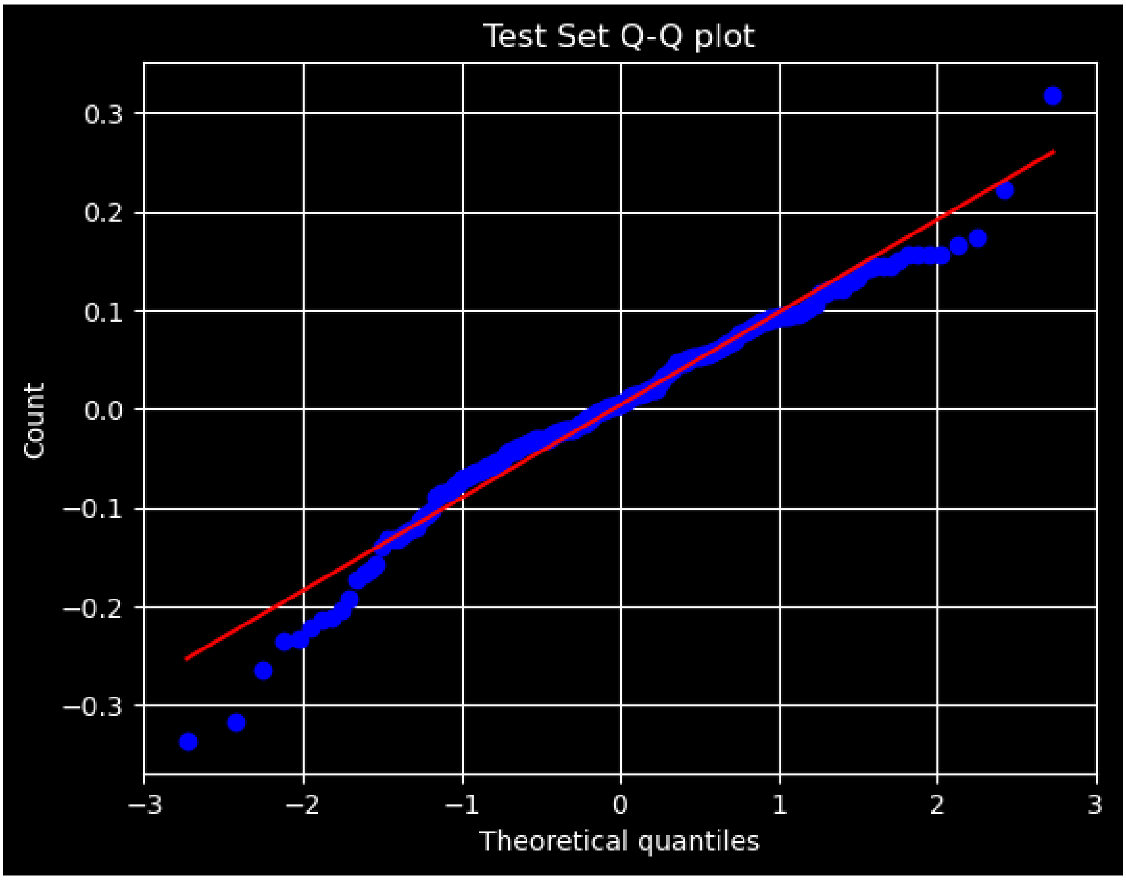
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

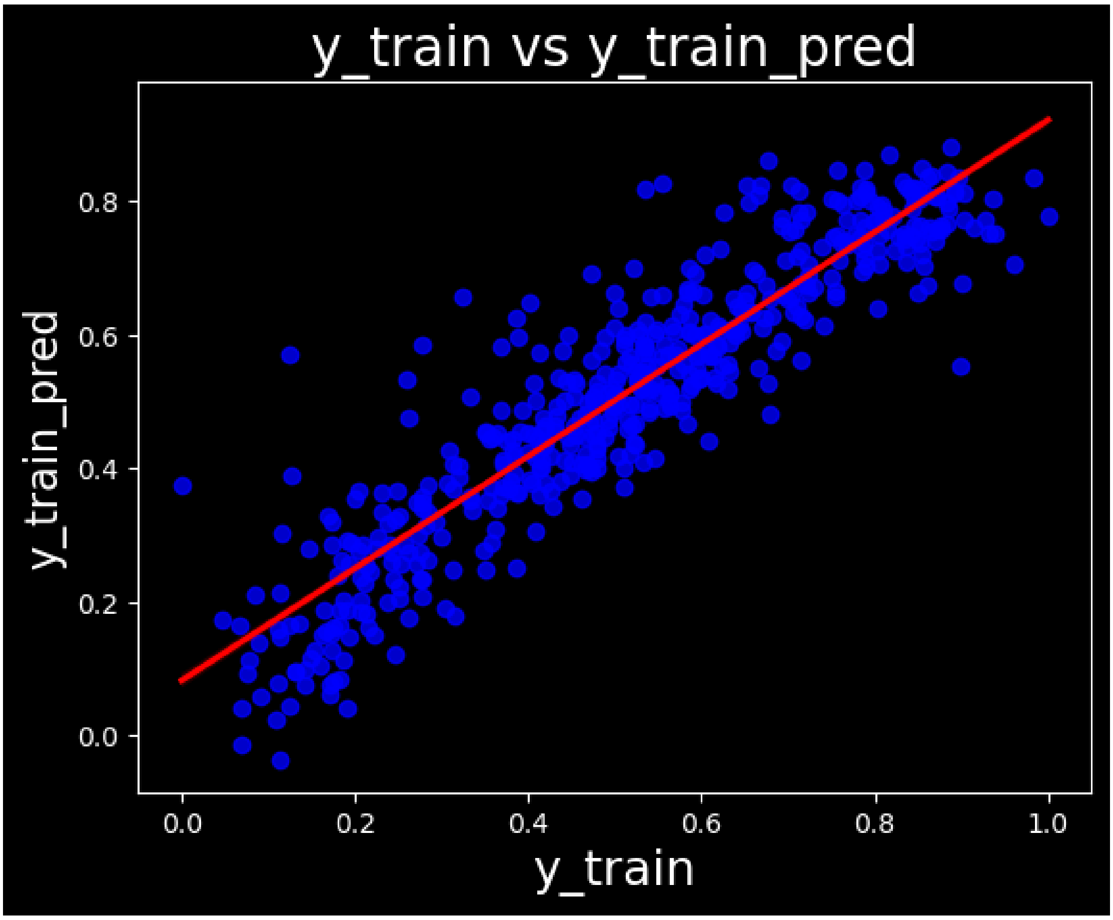
Based on the Below graph for QQ plot:

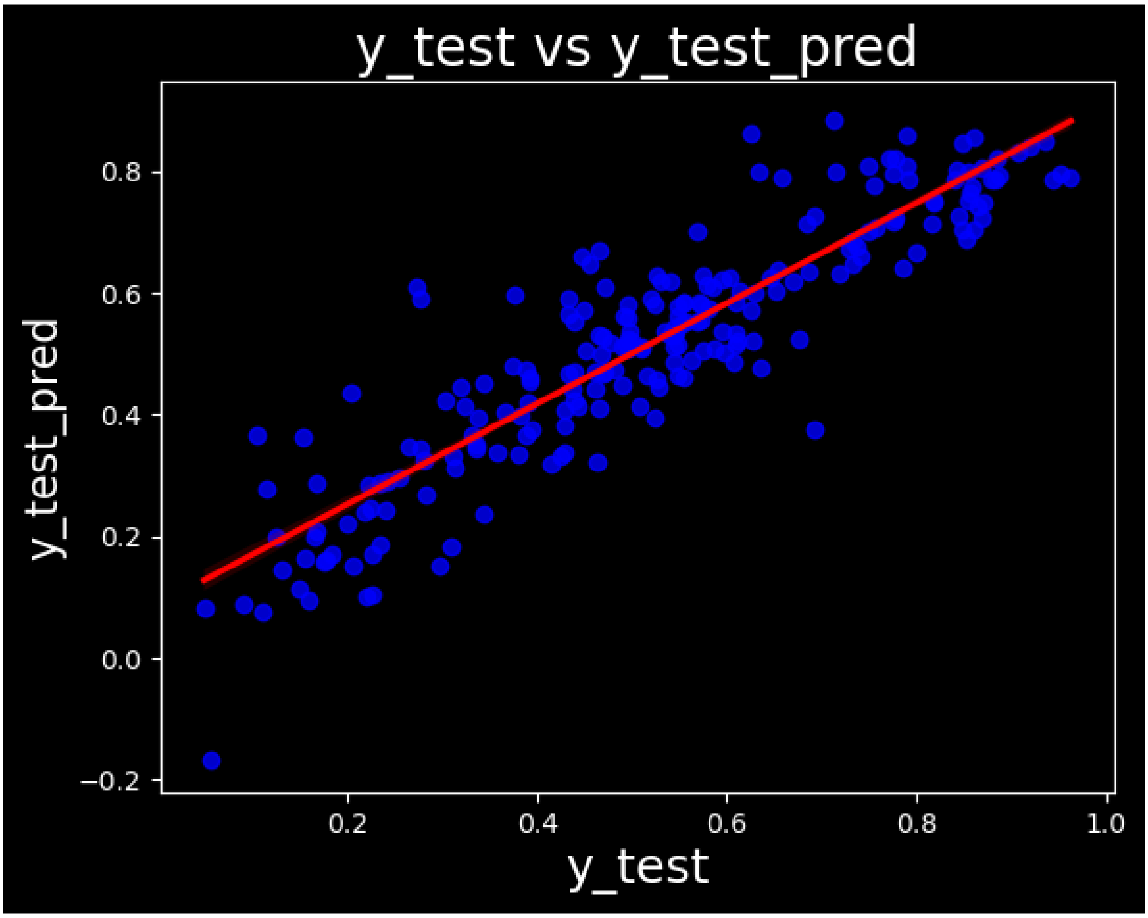
Multivariate Normality: The differences between observed and predicted values are normally distributed across all the data points.


Train Set Q-Q plot
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
Count
−3
−2
−1
0
1
2
3
Theoretical quantiles

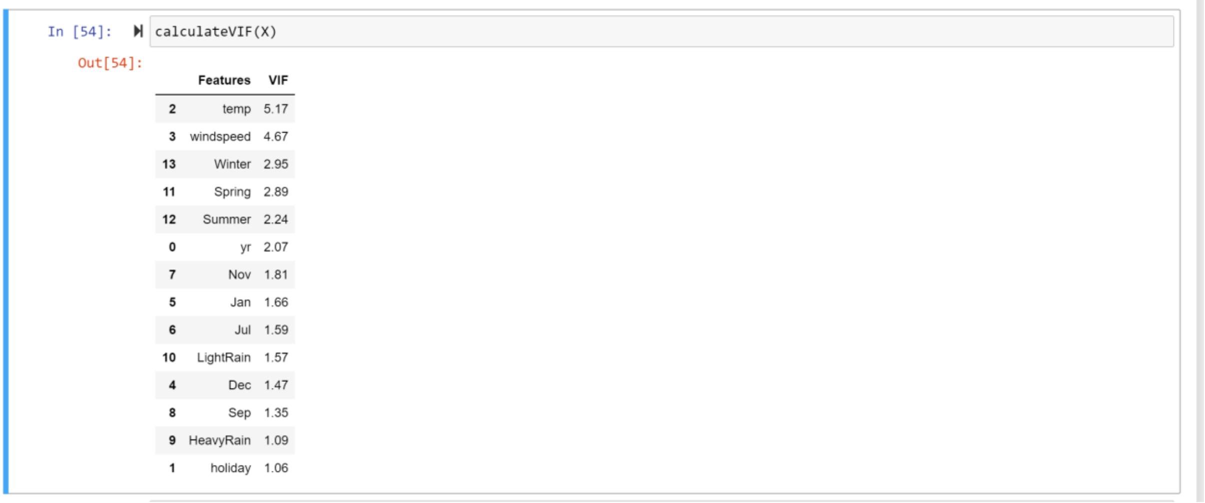

Test Set Q-Q plot
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
Count
−3
−2
−1
0
1
2
3
Theoretical quantiles

<u>Linear Relationship:</u> Based on the below graphs multiple linear regression is the existence of a linear relationship between the dependent (outcome) variable and the independent variables having straight-line relationship not a curvilinear one.





Variance Inflation Factor (VIF): with VIF values above 5 indicating problematic multicollinearity. Based on the below image we kept all the VIF scores below 5 except the temp variable, because temp factor is important in the domain of these boom bikes in-terms of usage.

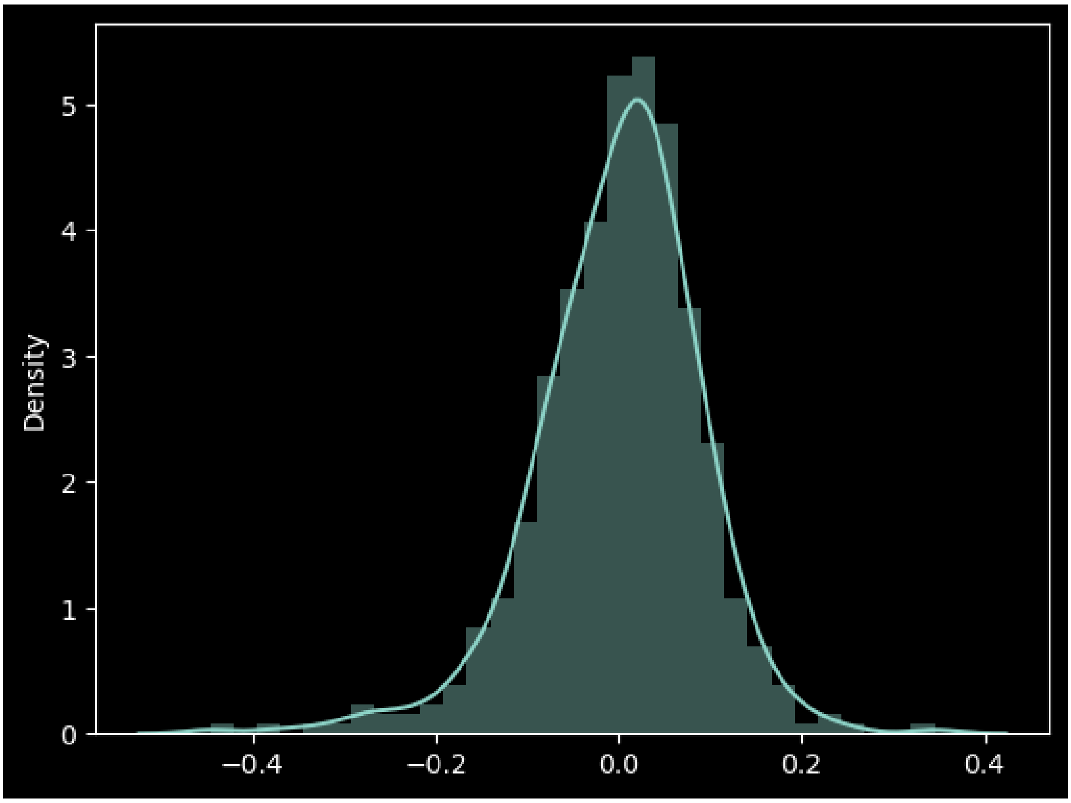


```
In [54]:  calculateVIF(X)
Out[54]:
```

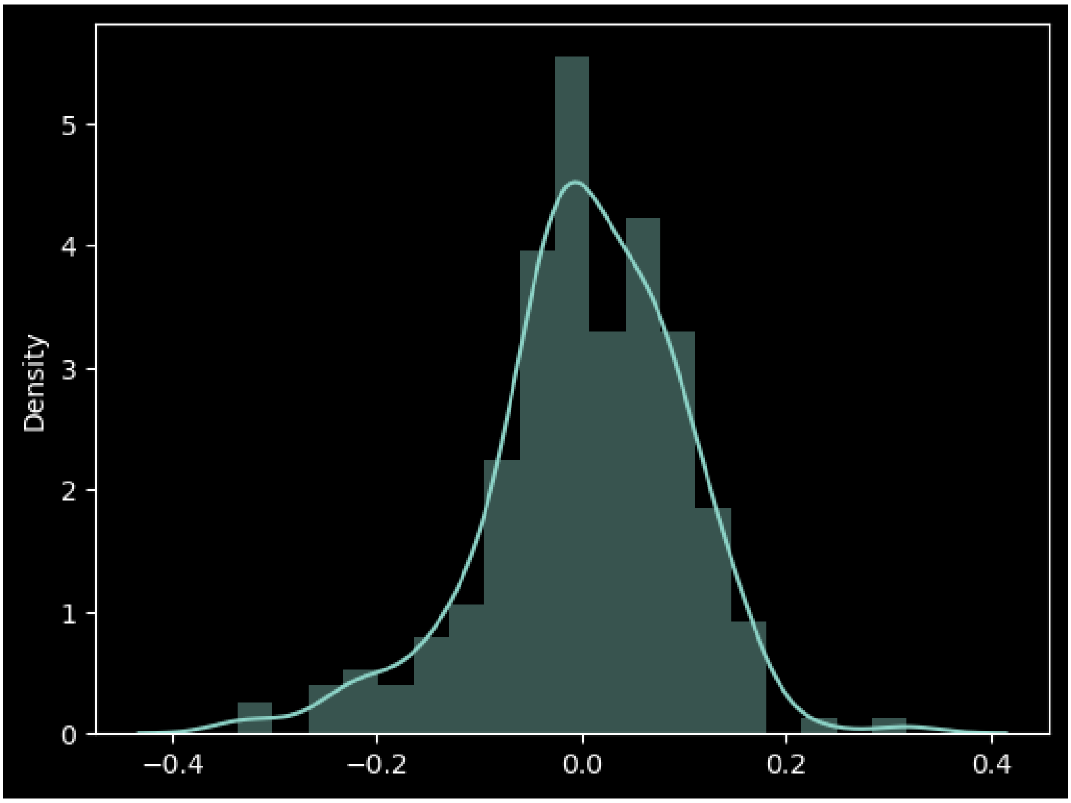
	Features	VIF
2	temp	5.17
3	windspeed	4.67
13	Winter	2.95
11	Spring	2.89
12	Summer	2.24
0	yr	2.07
7	Nov	1.81
5	Jan	1.66
6	Jul	1.59
10	LightRain	1.57
4	Dec	1.47
8	Sep	1.35
9	HeavyRain	1.09
1	holiday	1.06

Homoscedasticity: Based on the below graphs,The variance of error terms (residuals) should be consistent across all levels of the independent variables.

Train Set Residuals:



Test Set Residuals:



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:

Based on the final model: temp, holiday,months(Jun,Dec,jul,sep)

General Subjective Questions

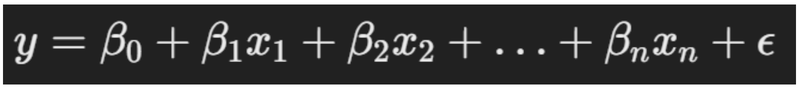
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

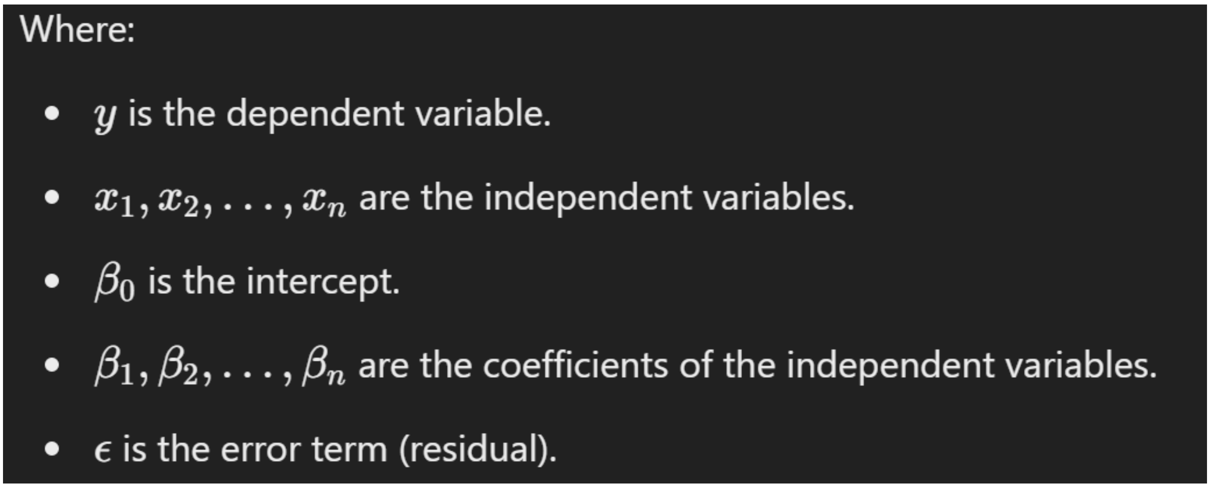
Linear regression is a statistical method for modelling the relationship between a dependent variable (response) and one or more independent variables (predictors). It is one of the simplest and most widely used regression techniques. Here's a detailed explanation of the linear regression algorithm:

→ Basic Concept

Linear regression aims to find the best-fitting straight line (in the case of simple linear regression) or a hyperplane (in the case of multiple linear regression) that describes the relationship between the dependent and independent variables. The equation of a linear regression model can be written as:



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$



Where:

- y is the dependent variable.
- $x_1, x_2, \ldots, x_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables.
- ϵ is the error term (residual).

→Assumptions of Linear Regression:

To apply linear regression effectively, certain assumptions should be met:

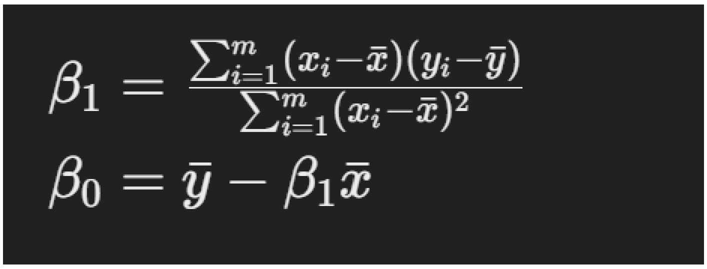
- **Linearity**: The relationship between the dependent and independent variables should be linear.
- **Independence**: The residuals (errors) should be independent.
- **Homoscedasticity**: The residuals should have constant variance at every level of xxx.
- **Normality**: The residuals should be normally distributed (especially important for small sample sizes).

→Estimating Coefficients

The coefficients (β\betaβ) are estimated using the method of least squares, which minimizes the sum of the squared differences between the observed values and the values predicted by the model.

→Solving for Coefficients

For a simple linear regression (one predictor), the formulas for the coefficients are:



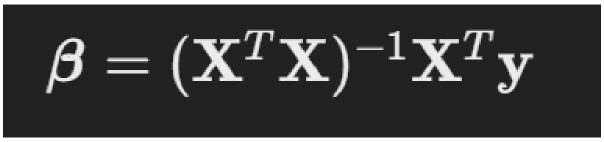
$$\beta_1 = \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{m}(x_i-\bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

For multiple linear regression, the coefficients can be solved using matrix operations. Let:

- X be the matrix of input features (with a column of ones for the intercept).
- Y be the vector of observed values.
- Beta be the vector of coefficients.

The solution is:

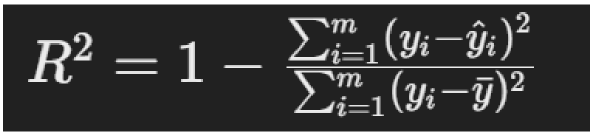


$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

→Model Evaluation

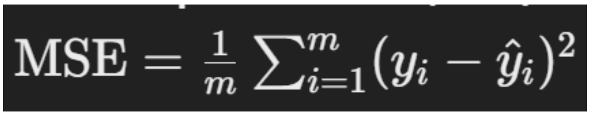
After fitting the model, it is important to evaluate its performance using various metrics:

- **R-squared (R^2)**: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.



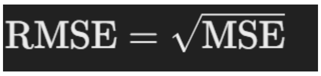
$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

- **Mean Squared Error (MSE)**: Measures the average of the squares of the errors.



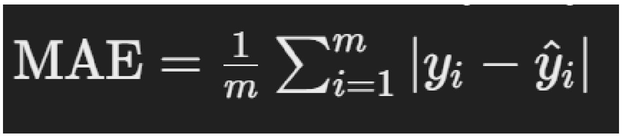
$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

- **Root Mean Squared Error (RMSE)**: The square root of MSE, providing a measure of error in the same units as the dependent variable.



$$\text{RMSE} = \sqrt{\text{MSE}}$$

- **Mean Absolute Error (MAE)**: Measures the average of the absolute errors.



$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

→Making Predictions:

Once the model is trained, it can be used to make predictions on new data. Given a new set of independent

→Regularization Techniques

To handle overfitting and multicollinearity, regularization techniques like Ridge Regression (L2 regularization) and Lasso Regression (L1 regularization) can be used. These techniques add a penalty term to the loss function:

- **Ridge Regression**
- **Lasso Regression**

Conclusion

Linear regression is a fundamental and straightforward technique for modeling relationships between variables. Understanding its assumptions, how to estimate coefficients, and how to evaluate the model's performance is crucial for effectively applying it to real-world problems. Regularization techniques can further enhance its utility in the presence of multicollinearity or when dealing with high-dimensional data.

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

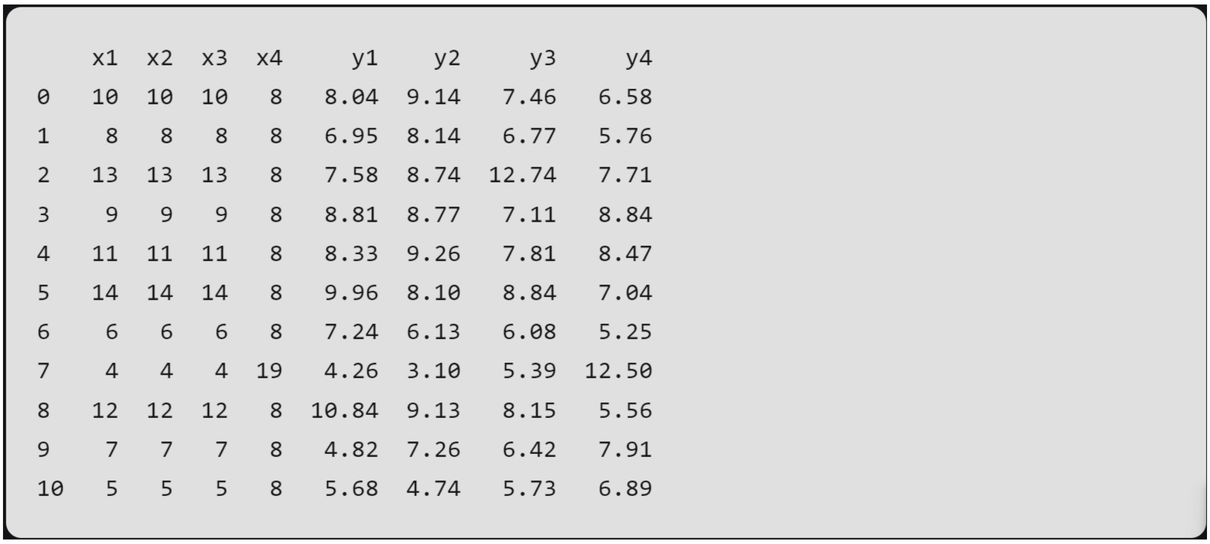
Answer:

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. The quartet, created by the British statistician Francis Anscombe in 1973, is used to illustrate the importance of graphical analysis of data alongside traditional statistical analysis.

Anscombe's quartet illustrates several key lessons for data analysis:

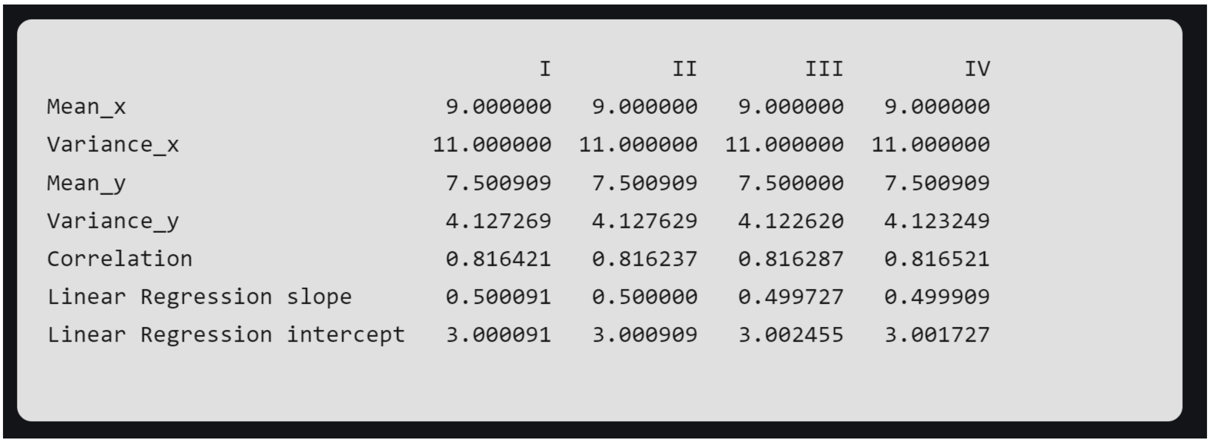
- Graphical Analysis: Visualizing data is crucial. Summary statistics can be misleading without graphical representations.
- Outliers and Leverage Points: Outliers can have a significant impact on statistical measures and models. Identifying and understanding these points is important.
- Model Fit: A good fit according to summary statistics does not guarantee that the model is appropriate for the data. Checking residuals and the model fit visually can help.
- Data Context: Always consider the context of the data and the specific characteristics that might not be captured by simple statistics.

Data Sets:



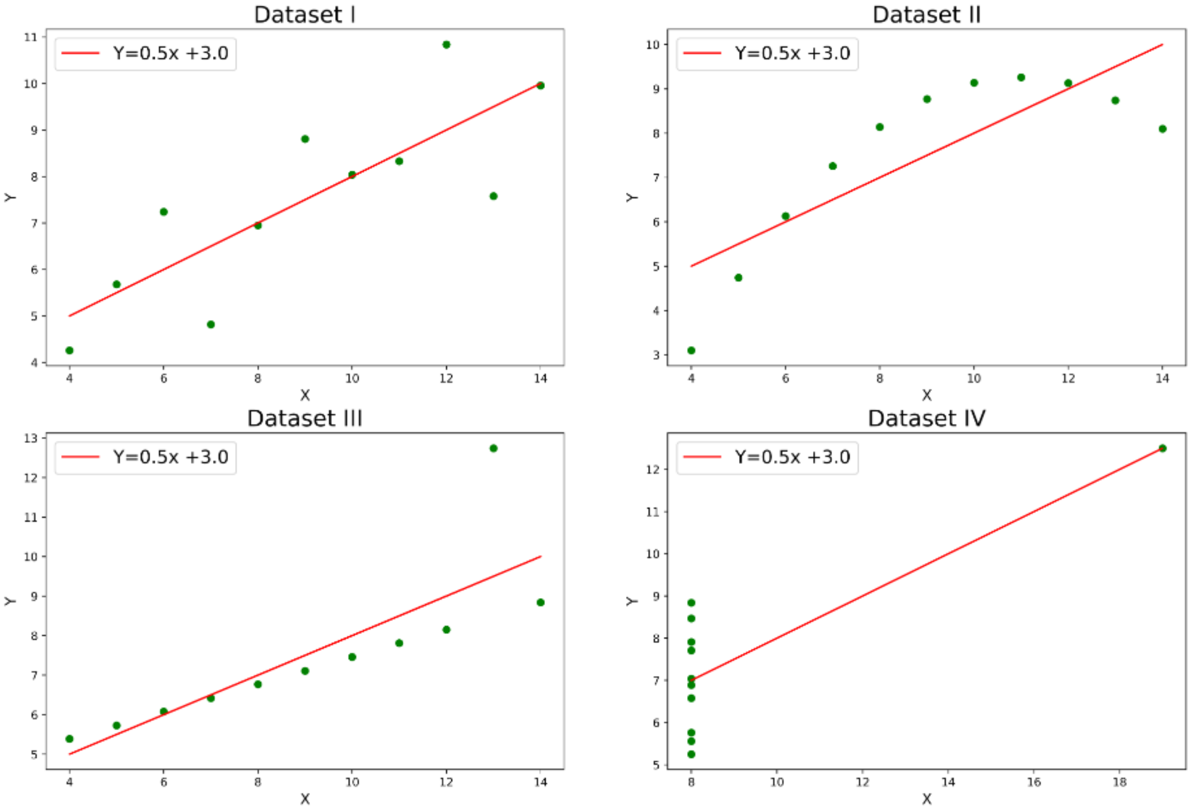
```
    x1  x2  x3  x4     y1    y2     y3     y4
0   10  10  10   8   8.04  9.14   7.46   6.58
1    8   8   8   8   6.95  8.14   6.77   5.76
2   13  13  13   8   7.58  8.74  12.74   7.71
3    9   9   9   8   8.81  8.77   7.11   8.84
4   11  11  11   8   8.33  9.26   7.81   8.47
5   14  14  14   8   9.96  8.10   8.84   7.04
6    6   6   6   8   7.24  6.13   6.08   5.25
7    4   4   4  19   4.26  3.10   5.39  12.50
8   12  12  12   8  10.84  9.13   8.15   5.56
9    7   7   7   8   4.82  7.26   6.42   7.91
10   5   5   5   8   5.68  4.74   5.73   6.89
```

Data Set Statistical Representation:



```
                                    I          II         III         IV
Mean_x                       9.000000    9.000000    9.000000    9.000000
Variance_x                  11.000000   11.000000   11.000000   11.000000
Mean_y                       7.500909    7.500909    7.500000    7.500909
Variance_y                   4.127269    4.127629    4.122620    4.123249
Correlation                  0.816421    0.816237    0.816287    0.816521
Linear Regression slope      0.500091    0.500000    0.499727    0.499909
Linear Regression intercept  3.000091    3.000909    3.002455    3.001727
```

Graphical Representation:



Dataset I

Dataset II

Dataset III

Dataset IV

Conclusion

Anscombe's quartet serves as a powerful reminder of the importance of thorough data analysis, including both statistical measures and graphical representations. It highlights the potential pitfalls of relying solely on summary statistics and underscores the necessity of a comprehensive approach to data exploration and model evaluation.

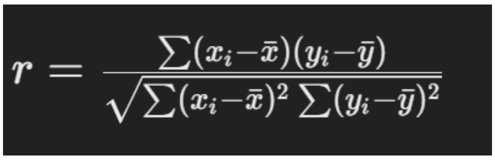
3. What is Pearson’s R? (3 marks)

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which two variables are linearly related, providing both the strength and direction of the relationship. Here’s a detailed explanation:

Formula

The Pearson correlation coefficient rrr is calculated using the following formula:



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

Where:

- X and y are individual data points.
- X bar and y bar are the means of the xxx and yyy variables, respectively.
- The numerator is the covariance of xxx and yyy.
- The denominator is the product of the standard deviations of xxx and yyy.

Interpretation

The value of r ranges from -1 to 1:

- **r=1**: Perfect positive linear relationship. As one variable increases, the other variable increases proportionally.
- **r=−1**: Perfect negative linear relationship. As one variable increases, the other variable decreases proportionally.
- **r=0**: No linear relationship between the variables. However, this does not necessarily mean there is no relationship at all; it just indicates no linear relationship.

Properties

- **Symmetry**: r(x,y)=r(y,x)
- **Unit-free**: Pearson’s R is a dimensionless measure, meaning it does not depend on the units of the variables.
- **Linearity**: It only measures the strength of a linear relationship. Non-linear relationships are not accurately captured by Pearson’s R.
- **Sensitivity to Outliers**: Pearson’s R can be heavily influenced by outliers, which can distort the perceived strength and direction of the relationship.

Calculation Steps

- **Compute the means** of the xxx and yyy variables.

- **Subtract the mean** from each data point to get the deviations from the mean.
- **Compute the product** of the deviations for corresponding xxx and yyy values.
- **Sum these products** to get the covariance.
- **Compute the squared deviations** for each variable, sum them, and then take the square root to get the standard deviations.
- **Divide the covariance** by the product of the standard deviations.

Conclusion:

Pearson’s R is a valuable tool for determining the strength and direction of a linear relationship between two variables. It is widely used in statistics, data analysis, and various scientific disciplines to assess correlations. However, it is essential to complement Pearson's R with graphical analysis and consider potential non-linear relationships and outliers that can affect the interpretation.

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

Answer:

What is Scaling?

Scaling is a process used in data preprocessing to adjust the range and distribution of numerical features. This helps to ensure that different features contribute equally to the analysis or machine learning model. Without scaling, features with larger ranges can dominate those with smaller ranges, leading to biased models.

Why is Scaling Performed?

Scaling is performed for several reasons:

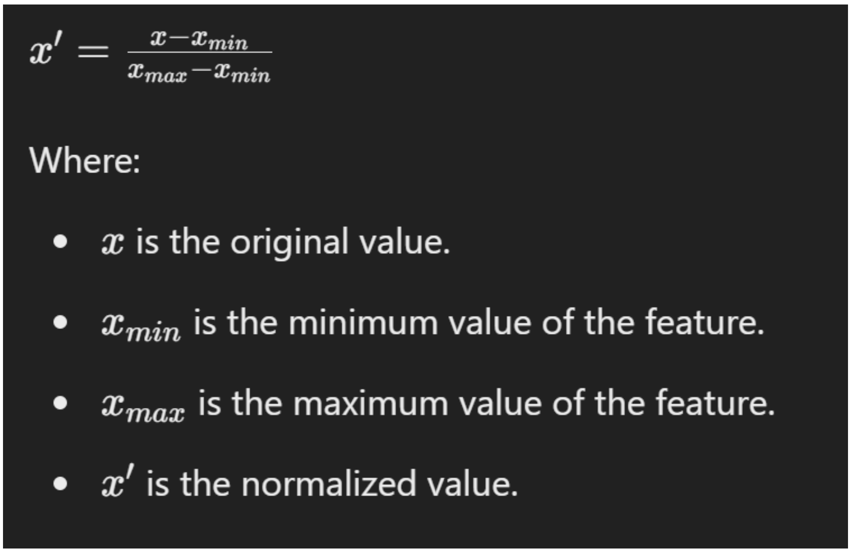
- **Model Performance**: Many machine learning algorithms perform better when the numerical features are on a similar scale. Algorithms like gradient descent converge faster when features are scaled.
- **Distance-Based Algorithms**: Algorithms that rely on distance metrics (e.g., K-Nearest Neighbors, Support Vector Machines, and clustering algorithms) require features to be scaled to ensure that each feature contributes equally to the distance calculation.
- **Improved Interpretability**: Scaling can make the model coefficients more interpretable, especially in linear models.
- **Handling Regularization**: Regularization techniques (e.g., Lasso and Ridge regression) are affected by the scale of the data. Scaling ensures that the regularization term penalizes each feature equally.

Types of Scaling

The two most common types of scaling are normalization and standardization.

→Normalization (Min-Max Scaling)

Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. The formula for min-max normalization is:



$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where:

- x is the original value.
- x_{min} is the minimum value of the feature.
- x_{max} is the maximum value of the feature.
- x' is the normalized value.

Advantages:

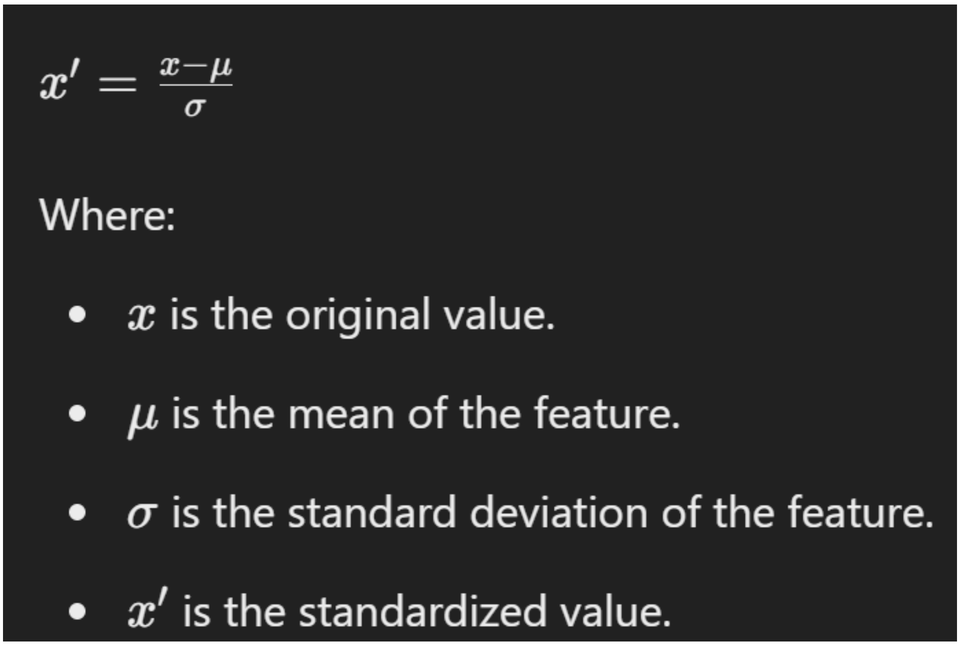
- Preserves the original distribution of the data.
- Ensures all features are within the same range.

Disadvantages:

- Sensitive to outliers, as they can skew the min and max values.

→Standardization (Z-Score Scaling)

Standardization scales the data to have a mean of 0 and a standard deviation of 1. The formula for standardization is:



$$x' = \frac{x-\mu}{\sigma}$$

Where:

- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.
- x' is the standardized value.

Advantages:

- Centres the data around 0, which can be beneficial for algorithms assuming data is centered.
- Less sensitive to outliers compared to normalization, though extreme outliers can still have an impact.

Disadvantages:

- Does not bound the values to a fixed range.

→**Comparison**

Normalization (Min-Max Scaling):

- Range: Typically [0, 1] or [-1, 1].
- Sensitive to outliers.
- Useful when you want to bound values within a specific range.

Standardization (Z-Score Scaling):

- Mean: 0.
- Standard deviation: 1.
- Less sensitive to outliers.
- Useful for algorithms that assume data is normally distributed or centered around zero.

When to Use Which

- **Normalization**: Use when you know the data has a bounded range and you want to preserve this range. It's often used when the algorithm requires features to be on a common scale but the exact distribution doesn't matter (e.g., neural networks).

- **Standardization**: Use when the data has a Gaussian distribution or when the algorithm assumes the data is centred around zero with unit variance (e.g., linear regression, logistic regression, and principal component analysis).

Conclusion

Scaling is a crucial preprocessing step that can significantly impact the performance and interpretability of machine learning models. Understanding the difference between normalization and standardization helps in choosing the appropriate method based on the specific requirements of the algorithm and the nature of the data.

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

Answer:

Infinite VIF: Causes and Implications

Perfect Multicollinearity

An infinite VIF value occurs when there is perfect multicollinearity among the predictors. Perfect multicollinearity means that one predictor variable can be expressed as an exact linear combination of other predictor variables. In such cases, the R-square value becomes 1. VIF to become infinite:

Formula: VIF=1/1-R-square

Why Does Perfect Multicollinearity Occur?

- **Duplicate Variables**: If a dataset contains duplicate columns or variables that are linear transformations of each other, perfect multicollinearity will be present.

  Example: If X2=2X1, then the VIF for X1 and X2 will be infinite.

- **Dummy Variable Trap**: In the case of categorical variables encoded as dummy variables, including all dummy variables without dropping one category leads to perfect multicollinearity.

  Example: For a categorical variable with three categories, creating dummy variables for all three will result in perfect multicollinearity because the sum of the dummy variables is always 1.

- **Exact Linear Relationships**: If one predictor variable can be perfectly predicted by a combination of other predictors, perfect multicollinearity occurs.

  Example: If X3=X1+X2 , then the VIF for X3 will be infinite.

Implications of Infinite VIF

- **Unstable Coefficients**: Regression coefficients become highly unstable and sensitive to changes in the data. Small changes in the data can lead to large changes in the estimated coefficients.
- **Inflated Standard Errors**: The standard errors of the coefficients increase, making it difficult to determine the significance of predictors.
- **Reduced Model Interpretability**: It becomes challenging to assess the individual impact of collinear predictors on the dependent variable.

Addressing Infinite VIF

To resolve issues of perfect multicollinearity and infinite VIF values, consider the following strategies:

- **Remove Redundant Predictors**: Identify and remove duplicate or highly correlated predictors.

  Example: If two predictors are perfectly correlated, drop one of them.
- **Avoid Dummy Variable Trap**: When using dummy variables, always drop one category to avoid perfect multicollinearity.

  Example: For a categorical variable with three categories (A, B, C), create dummy variables for two categories (A and B) and drop the third (C).
- **Combine Predictors**: If two or more predictors are highly correlated, consider combining them into a single predictor (e.g., by taking their average).
- **Regularization**: Use regularization techniques like Ridge Regression (L2 regularization) to handle multicollinearity. Regularization adds a penalty term to the regression, reducing the impact of collinear predictors.

Conclusion: Infinite VIF values indicate perfect multicollinearity among predictor variables, which can severely affect the reliability and interpretability of a regression model. Identifying and addressing multicollinearity is crucial for building robust and interpretable models. Common solutions include removing redundant predictors, avoiding the dummy variable trap, combining correlated predictors, and applying regularization techniques.

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

Answer:

What is a Q-Q Plot?

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, such as the normal distribution. The plot displays the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution closely, the points on the Q-Q plot will lie approximately on a straight line.

Use of a Q-Q Plot in Linear Regression

In the context of linear regression, Q-Q plots are used to check the assumption that the residuals (errors) of the regression model are normally distributed. This assumption is crucial for the validity of various statistical tests and confidence intervals in linear regression.

How to Construct a Q-Q Plot

1. **Obtain Residuals**: Calculate the residuals from the fitted linear regression model.
2. **Sort Residuals**: Sort the residuals in ascending order.
3. **Calculate Theoretical Quantiles**: Calculate the theoretical quantiles of the normal distribution that correspond to the sorted residuals.
4. **Plot Residuals vs. Theoretical Quantiles**: Create a scatter plot with the sorted residuals on the y-axis and the theoretical quantiles on the x-axis.

Interpreting a Q-Q Plot

- **Straight Line**: If the points lie on or near a straight line, it indicates that the residuals are approximately normally distributed.
- **S-shaped Curve**: If the points form an S-shaped curve, it suggests heavy tails or a distribution that is more spread out than a normal distribution (e.g., a distribution with high kurtosis).
- **Inverted S-shaped Curve**: If the points form an inverted S-shaped curve, it indicates light tails or a distribution that is less spread out than a normal distribution (e.g., a distribution with low kurtosis).
- **Other Deviations**: Systematic deviations from the straight line may indicate other issues with the distribution of the residuals, such as skewness or the presence of outliers.

Importance of a Q-Q Plot in Linear Regression

1. **Assessing Normality of Residuals:** One of the key assumptions in linear regression is that the residuals are normally distributed. The Q-Q plot helps in visually assessing whether this assumption holds.
   - **Validity of Statistical Tests:** Many inferential statistics and hypothesis tests, such as t-tests for coefficients and F-tests for overall significance, assume normally distributed residuals. Violations of this assumption can lead to inaccurate p-values and confidence intervals.
2. **Identifying Outliers:** Q-Q plots can help identify outliers or extreme values in the residuals, which can influence the regression model significantly.
   - **Model Robustness:** Identifying and addressing outliers can lead to a more robust and reliable regression model.
3. **Model Diagnostics:** Q-Q plots are a part of model diagnostics to evaluate the goodness-of-fit of the linear regression model.
   - **Improving Model Fit:** If the residuals are not normally distributed, it may indicate that the model is not capturing all the underlying patterns in the data. This can prompt the analyst to consider alternative models or transformations of the data.

Conclusion:

Q-Q plots are an essential diagnostic tool in linear regression analysis. They provide a visual method to check the normality of residuals, identify outliers, and assess the overall fit of the regression model. Ensuring that residuals are normally distributed helps validate the results of the regression analysis and supports the reliability of statistical inferences made from the model.