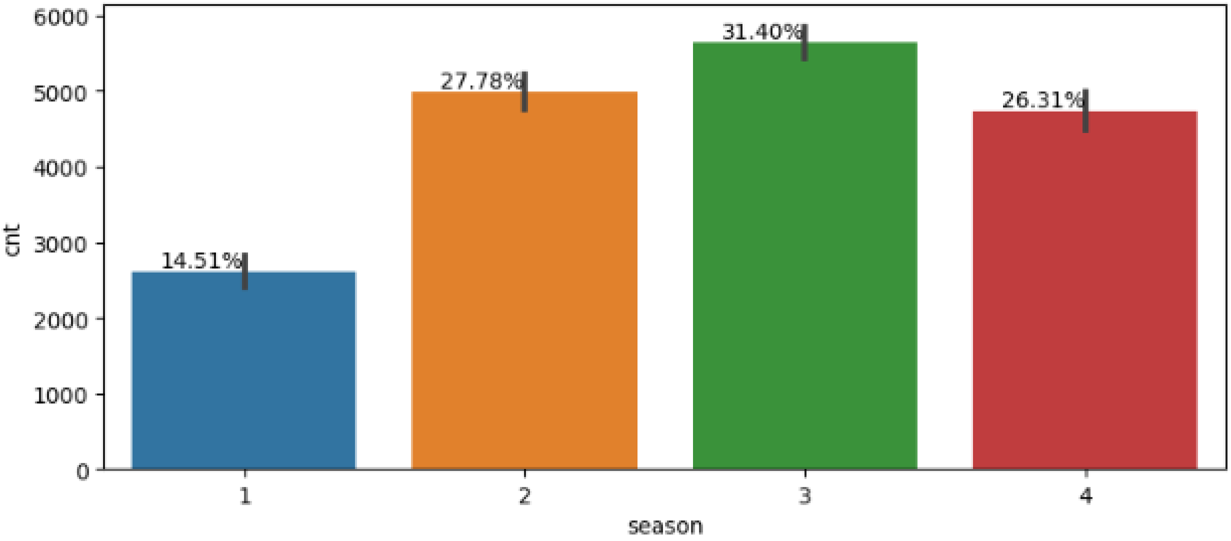
Assignment-based Subjective Questions

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

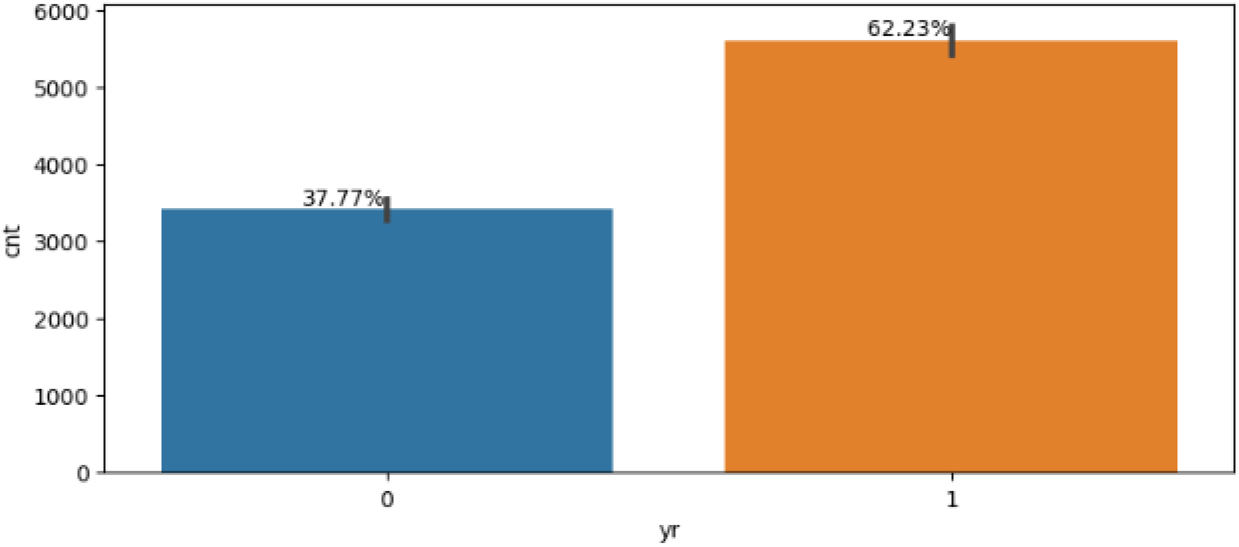
We have 7 categorical variables namely _'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit'_

- **season : season (1:spring, 2:summer, 3:fall, 4:winter)**



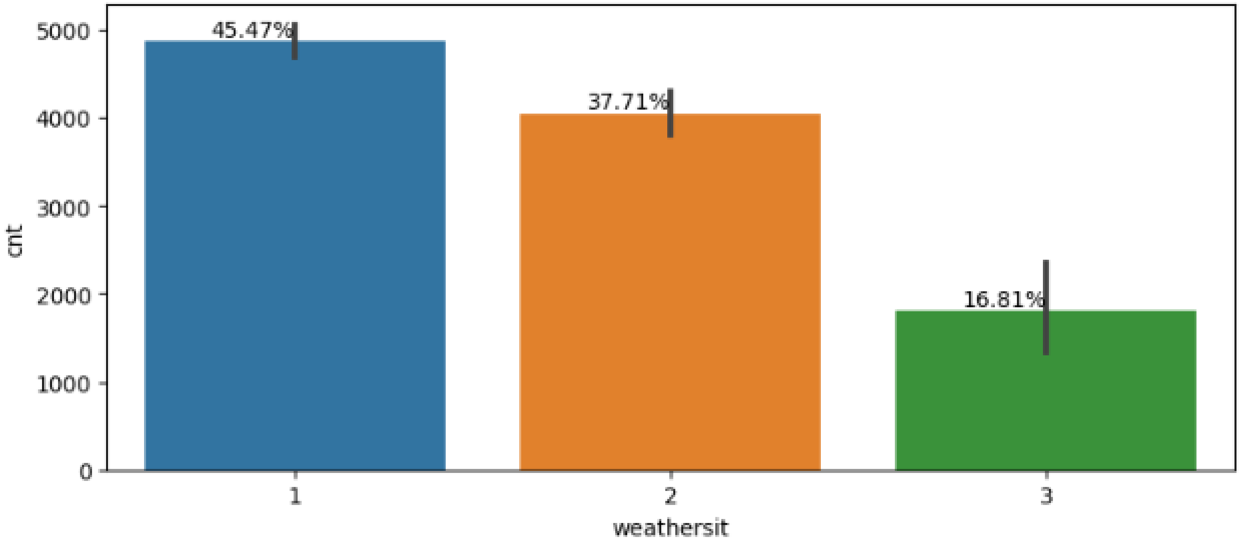
As we can see there is high demand in summer and fall than that of spring and winter
Spring has lowest demand among all seasons

- **yr : year (0: 2018, 1:2019)**



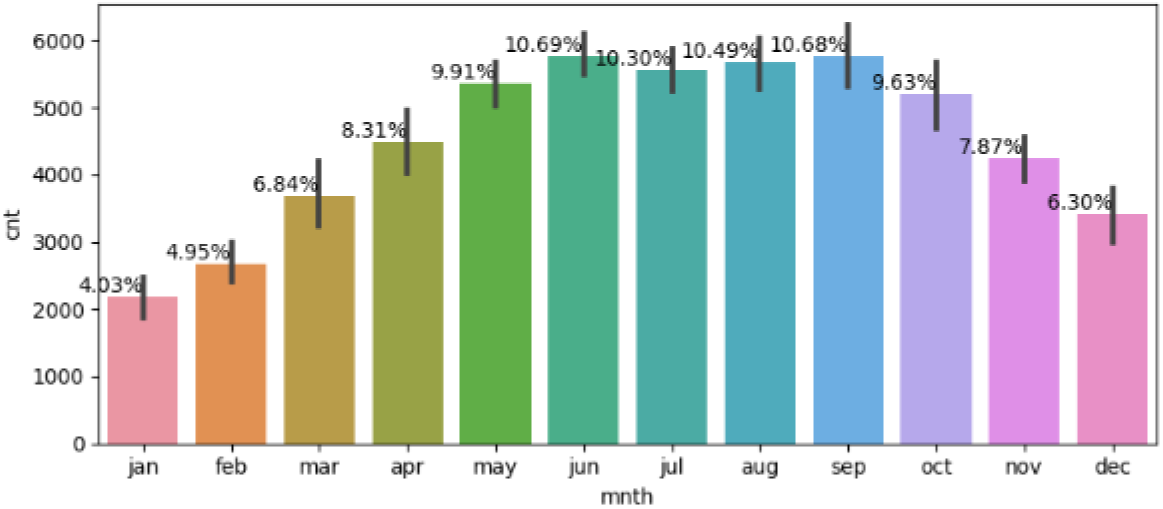
As we can see in year 2019 there is increase in demand

- **weathersit :**
  1: Clear, Few clouds, Partly cloudy, Partly cloudy
  2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog



As we can see the demand increases as weather gets clear

- **mnth :**



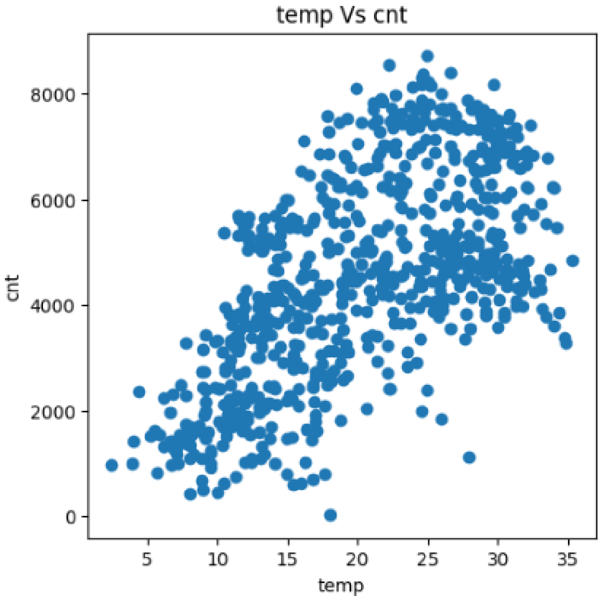
As we can see Bike demand is more in Jun , Jul , Aug and Sep

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

1. Multicollinearity : Creating binary dummy variables for all categories can lead to multicollinearity (high correlation between variables), causing unstable model coefficients. Dropping the first helps avoid this issue
2. Interpretability : Including all dummies makes it harder to interpret coefficients. With drop_first=True, coefficients show how each category differs from a reference category, enhancing interpretability.
3. Dimensionality: Using all dummies increases dataset complexity. drop_first=True reduces the number of variables, easing computational burden without losing information.

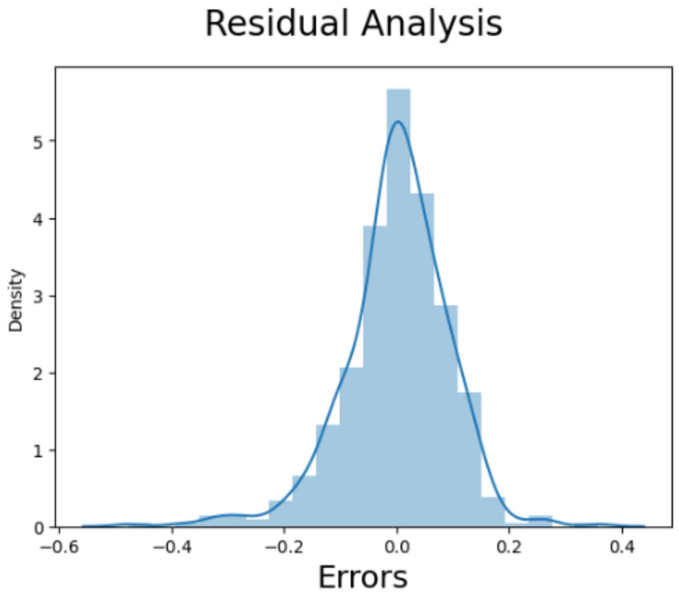
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

‘temp’ and ‘atemp’ has highest correlation with ‘cnt’ (target variable)


temp Vs cnt

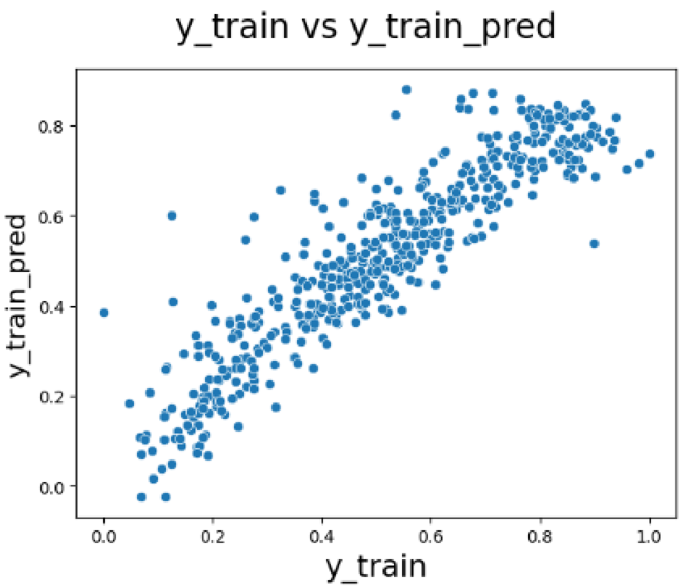
How did you validate the assumptions of Linear Regression after building the model on the training set?

- ➔ Residual Analysis:
  Calculate the residuals for the training data.



As we can see residuals approximately follow a normal distribution.

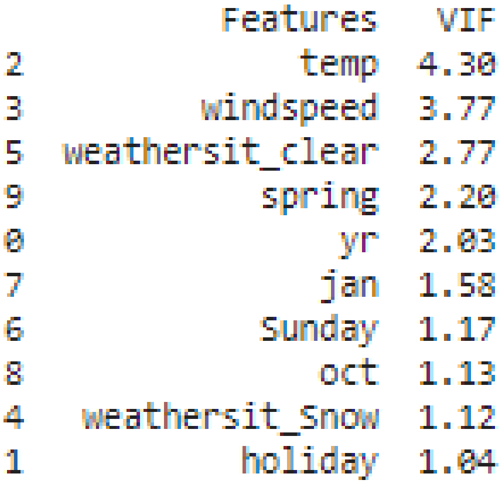
- ➔ Linearity of Relationships:



As we can see the relation between y_Train Vs y_Train_Predicted is approximately linear

➔ Variance Inflation Factors (VIF):

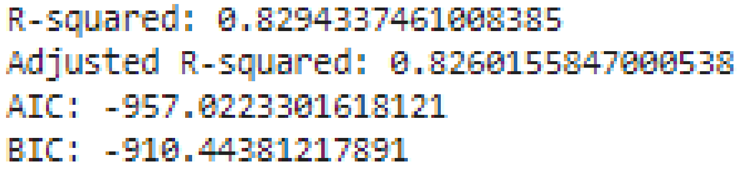
This measures the Multicollinearity



```
            Features   VIF
2               temp  4.30
3          windspeed  3.77
5  weathersit_clear  2.77
9             spring  2.20
0                 yr  2.03
7                jan  1.58
6             Sunday  1.17
8                oct  1.13
4    weathersit_Snow  1.12
1            holiday  1.04
```

As we can see the VIF values for variables are less than 5

➔ R-squared and R-squared Adjusted:



```
R-squared: 0.8294337461008385
Adjusted R-squared: 0.8260155847000538
AIC: -957.0223301618121
BIC: -910.44381217891
```

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

Top 3 features are

1. yr : Coefficient => 0.2368   P-value=> 0.000 VIF => 2.03
2. holiday : Coefficient => -0.0913   P-value=> 0.001 VIF => 1.04
3. temp :   Coefficient => 0.3551   P-value=> 0.000 VIF => 4.30

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data. It is often used for tasks such as predicting numerical values, making it a fundamental algorithm in regression analysis.

1. Linear Equation: The core idea behind linear regression is to model the relationship between the dependent variable (Y) and one or more independent variables (X) using a linear equation:

Simple Linear Regression: Y = b0 + b1*X

Multiple Linear Regression (with 'n' independent variables): Y = b0 + b1X1 + b2X2 + ... + bn*Xn

Here, 'Y' represents the target variable, 'X' represents the independent variables, 'b0' is the intercept, and 'b1', 'b2', ..., 'bn' are the coefficients that need to be estimated from the training data.

2. Assumptions: Linear regression makes several key assumptions:

Linearity: The relationship between the independent and dependent variables is assumed to be linear.

Independence: Observations are assumed to be independent of each other.

Homoscedasticity: The variance of the residuals (the differences between predicted and actual values) should be constant across all levels of the independent variables.

Normality: The residuals should follow a normal distribution.

No or Little Multicollinearity: Independent variables should not be highly correlated with each other.

Violations of these assumptions can affect the accuracy of linear regression results.

3. Steps Involved:

a. Data Collection: Gather a dataset containing the dependent variable and one or more independent variables.

b. Data Preprocessing: Clean and preprocess the data by handling missing values, encoding categorical variables, and scaling features if necessary.

c. Model Building: Choose the appropriate type of linear regression (simple or multiple) based on the number of independent variables. Then, use a mathematical algorithm to estimate the coefficients 'b0', 'b1', 'b2', ..., 'bn' that minimize the error between predicted and actual values.

d. Model Evaluation: Assess the model's performance using various metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared (R2) to measure how well the model fits the data.

4. Optimization: The optimization technique used in linear regression aims to find the optimal values of the coefficients that minimize the cost function, typically Mean Squared Error (MSE). Gradient Descent is a common optimization algorithm used to iteratively adjust the coefficients to reach the minimum MSE.

5. Accuracy Testing: To evaluate the accuracy of the linear regression model, you can use techniques such as cross-validation, where the dataset is divided into training and testing sets. The model is trained on the training set and then tested on the testing set to assess its

generalization performance. You can also use metrics like R-squared to quantify the goodness of fit and make predictions on new, unseen data.

In summary, linear regression is a simple yet powerful algorithm used for modeling the relationship between variables. It relies on assumptions about the data and uses optimization techniques to estimate the model coefficients. Accurate evaluation is crucial to ensure the model's reliability and predictive capability.

Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous statistical dataset consisting of four sets of paired data. Each set appears quite different when plotted on a graph, yet they share nearly identical summary statistics. This dataset was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before drawing conclusions and to highlight the limitations of relying solely on summary statistics.

Here are the details of Anscombe's quartet:

1. Set I:

x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82

Summary statistics: Mean(x) ≈ 9, Mean(y) ≈ 7.5, Variance(x) ≈ 11, Variance(y) ≈ 4.12, Correlation ≈ 0.82, Linear regression: y = 3 + 0.5x

2. Set II:

x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26

Summary statistics: Mean(x) ≈ 9, Mean(y) ≈ 7.5, Variance(x) ≈ 11, Variance(y) ≈ 4.12, Correlation ≈ 0.82, Linear regression: y = 3 + 0.5x

3. Set III:

x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42

Summary statistics: Mean(x) ≈ 9, Mean(y) ≈ 7.5, Variance(x) ≈ 11, Variance(y) ≈ 4.12, Correlation ≈ 0.82, Linear regression: y = 3 + 0.5x

4. Set IV:

x-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8

y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91

Summary statistics: Mean(x) ≈ 9, Mean(y) ≈ 7.5, Variance(x) ≈ 11, Variance(y) ≈ 4.12, Correlation ≈ -0.03, Linear regression: y = 3 + 0.5x

Key Takeaways:

Despite having identical means, variances, and correlations, each set exhibits a different data distribution when graphed.

Set I and Set II represent a linear relationship, but the data points are scattered differently.

Set III shows a clear non-linear relationship, and Set IV has an outlier that significantly affects the linear regression line.

Anscombe's quartet underscores the importance of visualizing data to gain insights and highlights the limitations of relying solely on summary statistics.

This dataset serves as a powerful reminder of the value of data visualization in statistical analysis and decision-making. It demonstrates how summary statistics alone can be insufficient for understanding the underlying structure and patterns in data.

What is Pearson’s R?

Pearson's R is a measure of the linear correlation between two variables. It is a statistical measure that indicates the extent to which two variables are linearly related. The value of Pearson's R can range from -1 to 1. A value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates no correlation.
Pearson's R is calculated using the following formula:
$r = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{(\sum(x_i - \bar{x})^2)}(\sum(y_i - \bar{y})^2)$

where:
r is the Pearson's R correlation coefficient
x_i is the value of the independent variable for the ith data point
$\bar{x}$ is the mean of the independent variable
y_i is the value of the dependent variable for the ith data point
$\bar{y}$ is the mean of the dependent variable

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

Scaling is the process of transforming data so that it has a common scale. This is done to improve the performance of machine learning algorithms.
Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
There are two main types of scaling:
Normalized scaling:
This involves transforming the data so that it has a mean of 0 and a standard deviation of 1.

$$X_{scaled} = (X - \min(X)) / (\max(X) - \min(X))$$

Standardized scaling:
Standardized scaling, also known as Z-score scaling or standardization, transforms the data to have a mean of 0 and a standard deviation of 1.

$$X_{standardized} = (X - \text{mean}(X)) / (\text{standard deviation})$$

Key Differences:
Range: Normalized scaling scales data to a specific range (usually 0 to 1), while standardized scaling centers the data around 0 with a standard deviation of 1.
Impact on Distribution: Normalized scaling preserves the original distribution of the data, whereas standardized scaling transforms the data into a standard normal distribution with a mean of 0 and a standard deviation of 1.
Outliers: Standardized scaling is more robust to outliers because it relies on the mean and standard deviation, which are less affected by extreme values. Normalized scaling can be sensitive to outliers as it depends on the range of the data.

Interpretability: Normalized scaling maintains the interpretability of the original data, as it preserves the original range and units. Standardized scaling transforms the data into z-scores, which can be more challenging to interpret directly.

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 assess multicollinearity in a multiple linear regression model. Multicollinearity occurs when two or more independent variables in the regression model are highly correlated, making it challenging to distinguish their individual effects on the dependent variable. VIF quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity.
VIF = 1 / (1 - R^2)

- VIF is the variance inflation factor
- R^2 is the coefficient of determination

The coefficient of determination is a measure of how well the independent variables explain the dependent variable. A perfect correlation between two independent variables will result in a coefficient of determination of 1, which will give a VIF of infinity.

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

A Q-Q plot, short for "Quantile-Quantile plot," is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. It helps in visualizing and comparing the quantiles (ordered values) of the observed data against the quantiles of a specified theoretical distribution. Q-Q plots are widely used in statistics to check if the assumption of normality holds for a dataset or to identify deviations from other theoretical distributions.

Here's how a Q-Q plot works:
Sorting: First, you sort the values of your dataset in ascending order.
Calculating Theoretical Quantiles: For each observed data point, you calculate the expected quantile that it would have if it followed the specified theoretical distribution. These expected quantiles are based on the cumulative distribution function (CDF) of the theoretical distribution.
Plotting: You then plot the observed data quantiles (y-axis) against the expected theoretical quantiles (x-axis).
In a perfect Q-Q plot, the points would fall along a straight line, indicating that the data closely follows the theoretical distribution. Deviations from the straight line suggest departures from the assumed distribution.

Use and Importance of Q-Q Plots in Linear Regression:

Q-Q plots are valuable in linear regression and statistics in general for several reasons:

Normality Assumption Check: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots help you assess whether this assumption holds. If the points in the Q-Q plot closely follow a straight line, it suggests that the residuals are approximately normally distributed, which is important for valid hypothesis testing and confidence intervals in linear regression.

Detecting Outliers: Q-Q plots can reveal outliers or data points that deviate from the assumed distribution. Outliers can have a significant impact on regression results, and identifying them is crucial for model accuracy.

Distributional Checks: Beyond normality, Q-Q plots can be used to check whether the residuals follow other theoretical distributions, such as the t-distribution or the Cauchy distribution. This is useful when you suspect that your data might not conform to the normal distribution assumption.

Model Diagnostics: Q-Q plots are a part of a suite of diagnostic tools used to assess the overall fit and validity of a regression model. By examining the Q-Q plot alongside other diagnostic plots, you can gain insights into the quality of your linear regression model.

Q-Q plots are a practical and intuitive way to assess the goodness of fit of your data to a theoretical distribution, especially the normal distribution. In linear regression, they play a crucial role in validating assumptions, identifying outliers, and ensuring the reliability of regression results.