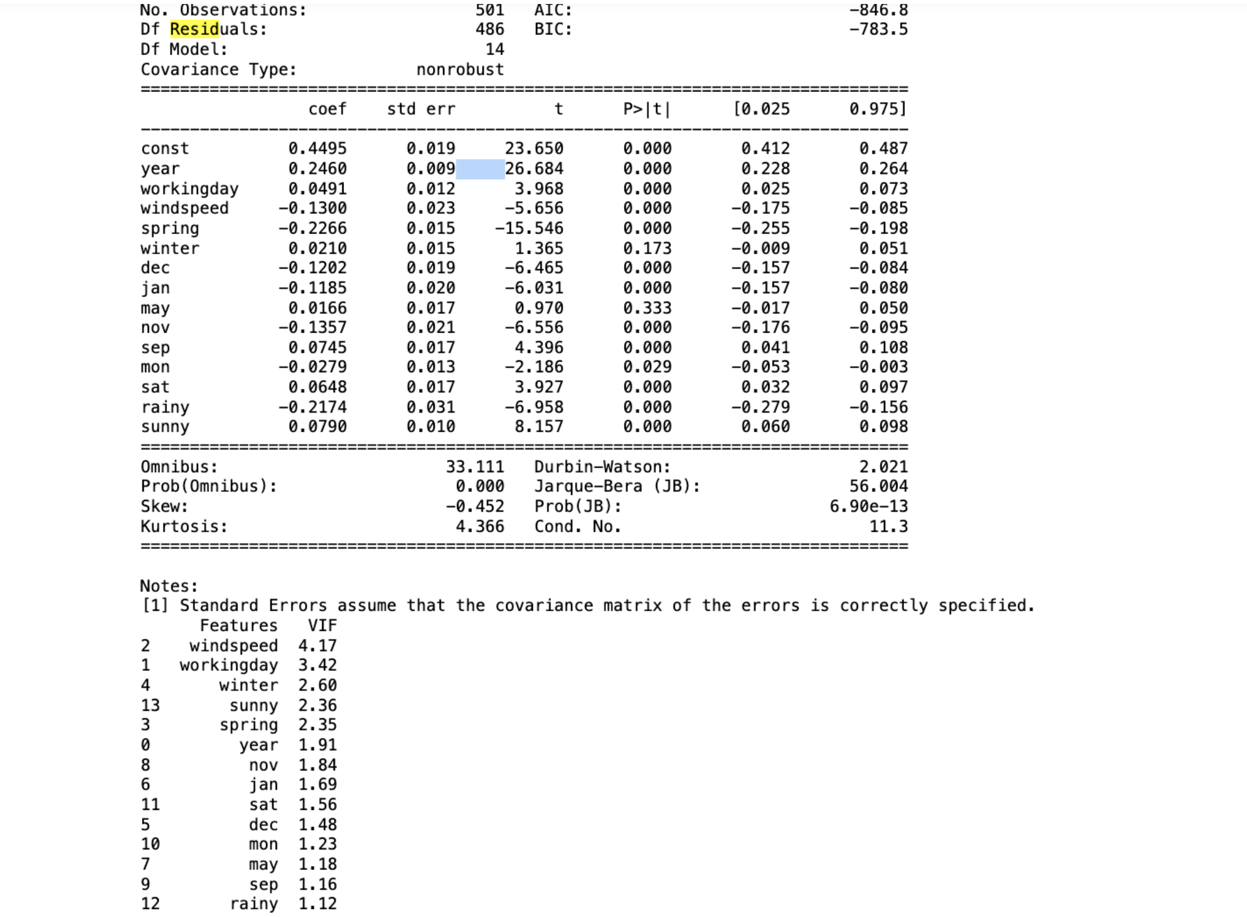
Assignment-based Subjective Questions

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

The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables



```
No. Observations:                 501   AIC:                            -846.8
Df Residuals:                     486   BIC:                            -783.5
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4495      0.019     23.650      0.000       0.412       0.487
year           0.2460      0.009     26.684      0.000       0.228       0.264
workingday     0.0491      0.012      3.968      0.000       0.025       0.073
windspeed     -0.1300      0.023     -5.656      0.000      -0.175      -0.085
spring        -0.2266      0.015    -15.546      0.000      -0.255      -0.198
winter         0.0210      0.015      1.365      0.173      -0.009       0.051
dec           -0.1202      0.019     -6.465      0.000      -0.157      -0.084
jan           -0.1185      0.020     -6.031      0.000      -0.157      -0.080
may            0.0166      0.017      0.970      0.333      -0.017       0.050
nov           -0.1357      0.021     -6.556      0.000      -0.176      -0.095
sep            0.0745      0.017      4.396      0.000       0.041       0.108
mon           -0.0279      0.013     -2.186      0.029      -0.053      -0.003
sat            0.0648      0.017      3.927      0.000       0.032       0.097
rainy         -0.2174      0.031     -6.958      0.000      -0.279      -0.156
sunny          0.0790      0.010      8.157      0.000       0.060       0.098
==============================================================================
Omnibus:                       33.111   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.004
Skew:                          -0.452   Prob(JB):                     6.90e-13
Kurtosis:                       4.366   Cond. No.                         11.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
      Features   VIF
2    windspeed  4.17
1   workingday  3.42
4       winter  2.60
13       sunny  2.36
3       spring  2.35
0         year  1.91
8          nov  1.84
6          jan  1.69
11         sat  1.56
5          dec  1.48
10         mon  1.23
7          may  1.18
9          sep  1.16
12       rainy  1.12
```

Categorical Variable list: spring , winter , dec , jan , may , nov **,** sep , mon , sat , rainy , sunny

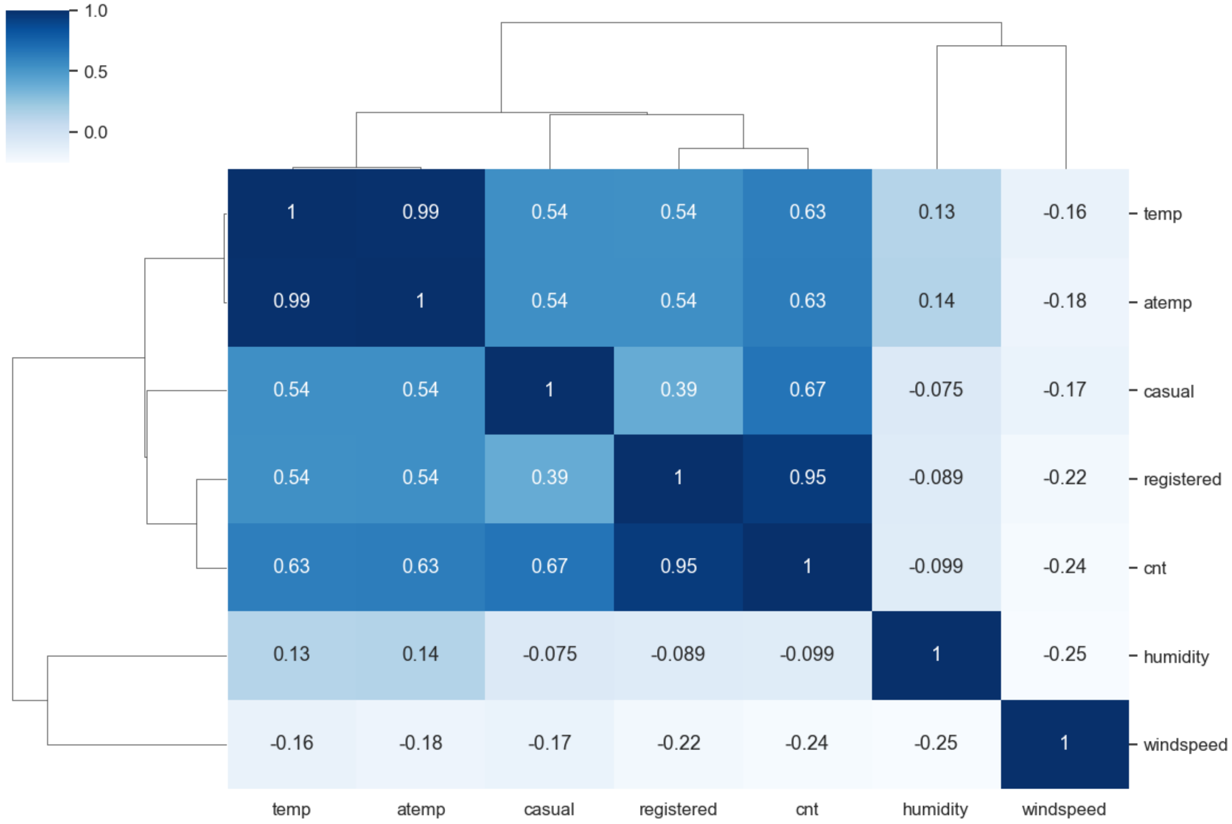
2. Why is it important to use drop_first=True during dummy variable creation?

Setting drop_first=True, you're dropping the first dummy variable, which helps to:

- ***Avoid multicollinearity****: By removing one of the highly correlated dummy variables, you reduce the risk of multicollinearity.*
- ***Avoid the dummy variable trap****: By dropping the first dummy variable, you avoid the perfect collinearity with the intercept term.*
- ***Reduce redundancy****: You remove redundant information, which can help improve model performance and generalization.*

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

Before model building and training, the pair plot shows highest correlation for registered variable having correlation 0.95(Before Clean Up process). But we are not using casual and registered in our pre-processed training data for model training. casual + registered = cnt. This might leak out the crucial information and model might get overfit. So, excluding these two variables atemp is having highest correlation with target variable cnt which is followed by temp. As per the correlation heatmap, correlation coefficient between atemp and cnt is 0.63. And correlation coefficient between temp and cnt is 0.63.

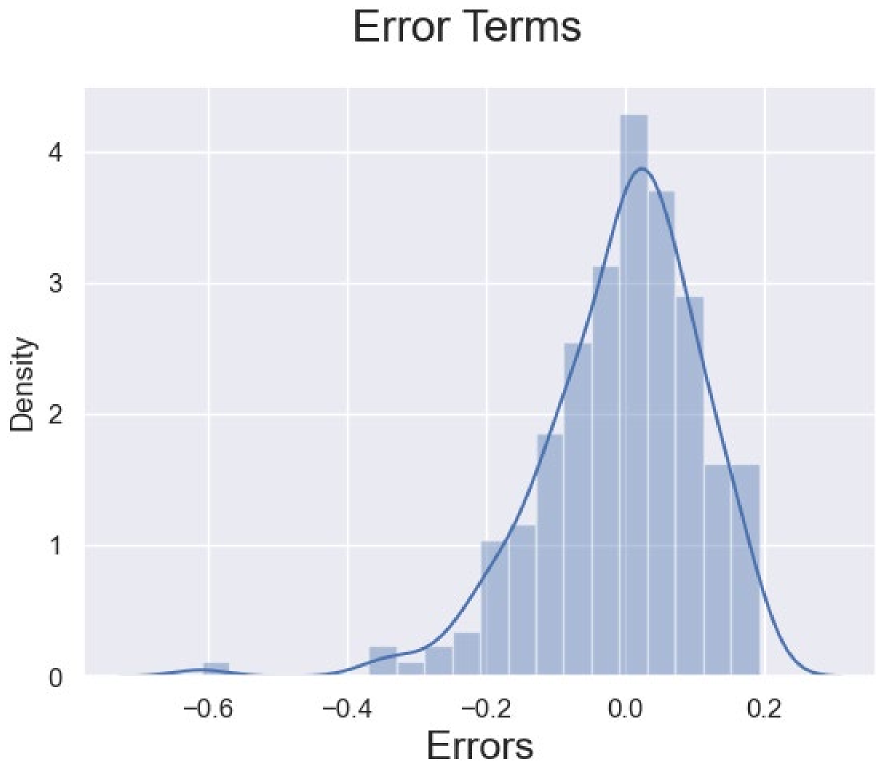


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

Residual Analysis:

We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms, and this is what it looks like:



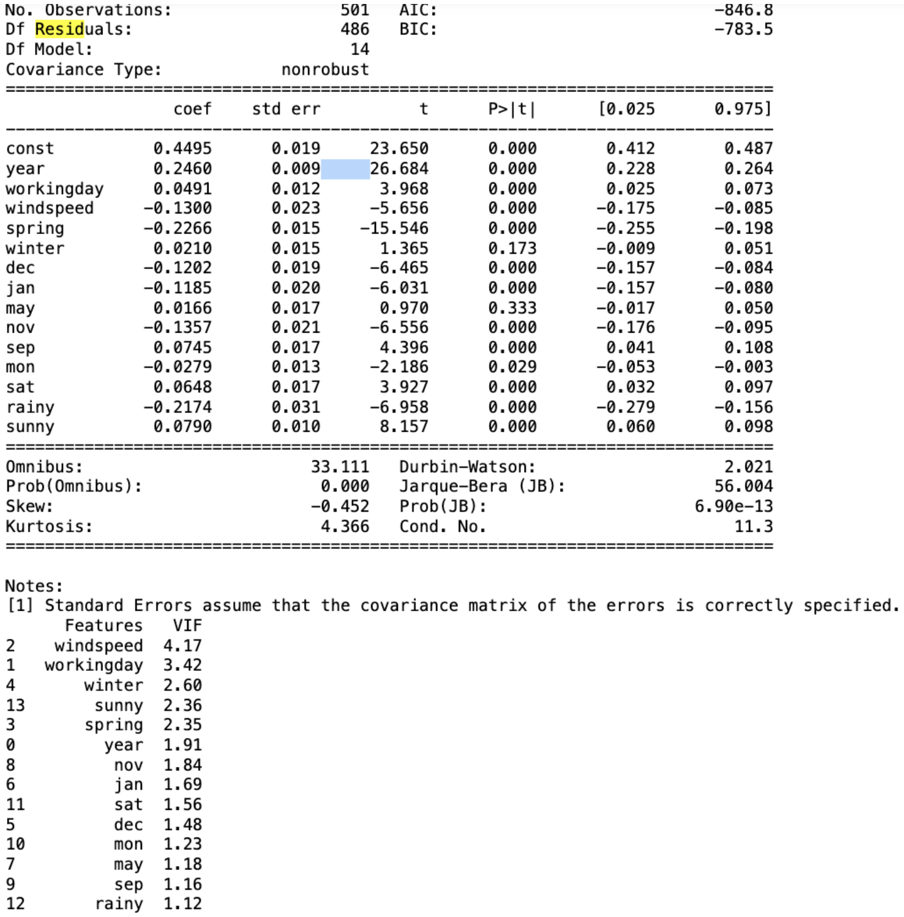
The residuals are following the normally distribution with a mean 0. All good! Linear relationship between predictor variables and target variables:

This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0.801 and adjusted R-Squared value on training set is 0.795. This means that variance in data is being explained by all these predictor variables.

Error terms are independent of each other: Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5 and P-Value is less than the 0.05

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

Top 3 features significantly contributing towards demand of shared bikes are: 1) ***year*** *2)* ***sept*** 3) ***rainy***



```
No. Observations:                 501   AIC:                            -846.8
Df Residuals:                     486   BIC:                            -783.5
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4495      0.019     23.650      0.000       0.412       0.487
year           0.2460      0.009     26.684      0.000       0.228       0.264
workingday     0.0491      0.012      3.968      0.000       0.025       0.073
windspeed     -0.1300      0.023     -5.656      0.000      -0.175      -0.085
spring        -0.2266      0.015    -15.546      0.000      -0.255      -0.198
winter         0.0210      0.015      1.365      0.173      -0.009       0.051
dec           -0.1202      0.019     -6.465      0.000      -0.157      -0.084
jan           -0.1185      0.020     -6.031      0.000      -0.157      -0.080
may            0.0166      0.017      0.970      0.333      -0.017       0.050
nov           -0.1357      0.021     -6.556      0.000      -0.176      -0.095
sep            0.0745      0.017      4.396      0.000       0.041       0.108
mon           -0.0279      0.013     -2.186      0.029      -0.053      -0.003
sat            0.0648      0.017      3.927      0.000       0.032       0.097
rainy         -0.2174      0.031     -6.958      0.000      -0.279      -0.156
sunny          0.0790      0.010      8.157      0.000       0.060       0.098
==============================================================================
Omnibus:                       33.111   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.004
Skew:                          -0.452   Prob(JB):                     6.90e-13
Kurtosis:                       4.366   Cond. No.                         11.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
      Features   VIF
2    windspeed  4.17
1   workingday  3.42
4       winter  2.60
13       sunny  2.36
3       spring  2.35
0         year  1.91
8          nov  1.84
6          jan  1.69
11         sat  1.56
5          dec  1.48
10         mon  1.23
7          may  1.18
9          sep  1.16
12       rainy  1.12
```

General Subjective Questions

1. Explain the linear regression algorithm in detail.

***Ans**: Linear Regression aims to establish a linear relationship between a dependent variable (target) and one or more independent variables (features). The goal is to create a model that predicts the target variable based on the features.*

Assumptions:

1. **Linearity:** The relationship between the target and features should be linear.
2. **Independence**: Each data point should be independent of the others.
3. **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variable.
4. **Normality**: The residuals should be normally distributed.
5. **No or little multicollinearity**: The features should not be highly correlated with each other.

Algorithm:

1. Model Formulation:

- Define the linear regression model: y = β0 + β1x + ε (simple linear regression) or y = β0 + β1x1 + β2x2 + ... + ε (multiple linear regression).

- y is the target variable.
- x (or x1, x2, ⋯) are the features.
- β0 is the intercept or constant term.
- β1 (or β1, β2, ⋯) are the coefficients or slopes.
- ε is the error term or residual.

2. Cost Function:

- Define the cost function to minimize: Sum of Squared Errors (SSE) = Σ(yi - (β0 + β1xi))^2 (simple linear regression) or SSE = Σ(yi - (β0 + β1x1i + β2x2i + ...))^2 (multiple linear regression).

3. Optimization:

- Use an optimization algorithm (e.g., Ordinary Least Squares (OLS)) tofind the best-fitting line by minimizing the SSE.
- The algorithm adjusts the coefficients (β0, β1, ⋯) to reduce the SSE.

4. Model Evaluation:

- Calculate metrics like Mean Squared Error (MSE), R-Squared (R^2), and Coefficient of Determination (CD) to evaluate the model's performance.

5. Prediction:

- Use the trained model to predict the target variable for new, unseen data.

Interpretation:

Linear Regression provides insights into the relationships between the target variable and features. The coefficients (**β0, β1, ⋯) represent the change in the target variable for a one-unit change in the corresponding feature, while controlling for other features.**

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet underscores the importance of not solely relying on summary statistics and emphasizes the need for visual exploration to understand the true nature of data relationships. It serves as a cautionary example against drawing conclusions solely based on numerical summaries. Anscombe's quartet is a famous example that highlights the importance of visualizing data before drawing conclusions. It consists of four datasets that have identical descriptive statistics (mean, median, variance, and correlation coefficient), but when plotted, they exhibit distinct relationships.

1. **Dataset 1**: This dataset shows a clear linear relationship between the x and y variables. The points are evenly distributed around the regression line, and the correlation coefficient is 0.816.

2. **Dataset 2**: This dataset also has a correlation coefficient of 0.816, but the relationship between the x and y variables is non-linear. The points are scattered around the regression line in a parabolic shape.

3. **Dataset 3**: This dataset has the same mean, median, variance, and correlation coefficient as Datasets 1 and 2, but it contains a single outlier. The outlier is a point that

is much higher than the rest of the data points, and it distorts the regression line. The correlation coefficient is still 0.816, but it is no longer a good measure of the relationship between the x and y variables.

4. **Dataset 4**: This dataset also has the same mean, median, variance, and correlation coefficient as Datasets 1, 2, and 3, but it contains a pair of outliers. The outliers are two points that are much higher than the rest of the data points, and they influence the regression line. The correlation coefficient is still 0.816, but it is not a good measure of the relationship between the x and y variables. Anscombe's quartet demonstrates that summary statistics can be misleading and that it is important to visualize data before drawing conclusions. By plotting the data, we can see the actual relationships between the variables and identify any outliers that may be distorting the results.

3. What is Pearson's R?

The Pearson's correlation coefficient, denoted as r, takes values between -1 and 1, with 0 indicating no linear relationship, 1 indicating a perfect positive linear relationship, and -1 indicating a perfect negative linear relationship. A positive value of r suggests that as one variable increases, the other tends to increase as well, while a negative value indicates an inverse relationship, where an increase in one variable corresponds to a decrease in the other. The formula for calculating Pearson's correlation coefficient is: $r = (\Sigma (x - \bar{x})(y - \bar{y})) / (\sqrt{\ } \Sigma (x - \bar{x})^2 \ \Sigma (y - \bar{y})^2)$

where x and y represent the data points of the two variables, $\bar{x}$ and ȳ are their respective means, and Σ denotes summation. The Pearson's correlation coefficient assumes that the data follows a bivariate normal distribution, meaning that the relationship between the variables is linear and the data points are normally distributed around the regression line. If these assumptions are not met, the correlation coefficient may not accurately reflect the strength and direction of the relationship between the variables. Despite its limitations, the Pearson's correlation coefficient remains a valuable tool for exploring relationships between continuous variables. It provides a concise summary of the linear association between two variables, allowing researchers to assess the degree of dependence or independence between them. However, it is important to interpret the correlation coefficient cautiously, considering the underlying assumptions and potential confounding factors.

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

Scaling is a process of transforming data to a common range to prevent differences in scales from affecting analysis or modeling results. It's performed to:

1. Prevent feature dominance: Ensure that no single feature dominates the analysis or modeling due to its large range.

2. Improve model performance: Scaling can improve the performance of algorithms that assume similar scales, such as clustering, neural networks, and principal component analysis (PCA).

3. Enhance interpretability: Scaled data can lead to more interpretable results, as all features are on the same scale.

There are two main types of scaling:

Normalized Scaling (Min-Max Scaling):

Rescales data between a minimum and maximum value, usually between 0 and 1. The formula is: X_normalized = (X - X_min) / (X_max - X_min)

Standardized Scaling (Z-Score Scaling):

Rescales data to have a mean of 0 and a standard deviation of 1. The formula is:

X_standardized = (X - μ) / σ

where μ is the mean and σ is the standard deviation.

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

Perfect multicollinearity can arise from various sources, including:

1. **Data entry errors:** Duplicating a column or entering identical values for multiple columns.

2. **Functional relationships:** Including variables that are mathematical functions of each other (e.g., x and x^2).

3. **Redundant variables:** Adding a variable that is a linear combination of existing variables. To address this issue, you can:

1. **Remove redundant variables**: Identify and drop one of the perfectly correlated variables.

2. **Transform variables:** Use techniques like orthogonalization or principal component analysis (PCA) to reduce correlation.

3. **Use regularization techniques:** Apply methods like ridge regression or the lasso to reduce the impact of multicollinearity.

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

A Q-Q (quantile-quantile) plot is a graphical method used to assess the normality of a dataset.

- It compares the quantiles of the observed data with the quantiles of a theoretical distribution, usually a normal distribution.
- If the data is normally distributed, the points on the Q-Q plot will fall along a straight line.
- Deviations from linearity indicate departures from normality.
- Q-Q plots are useful in linear regression because they can help identify non-linear relationships between the independent and dependent variables.
- They can also be used to assess the adequacy of the linear regression model and to identify outliers.

The use and importance of a Q-Q plot in linear regression are:

1. **Checking normality of residuals:** Q-Q plots help verify if the residuals follow a normal distribution, which is essential for the validity of linear regression assumptions.

2. **Identifying outliers and skewness:** Q-Q plots can reveal outliers, skewness, or other departures from normality, which may indicate issues with the data or the model.

3. **Evaluating model fit:** By comparing the distribution of residuals to a normal distribution, Q-Q plots provide a visual assessment of the model's goodness of fit.

4. **Selecting appropriate transformations:** If the residuals are not normally distributed, Q-Q plots can guide the selection of appropriate transformations to normalize the data.

5. **Validating assumptions:** Q-Q plots are a visual check for the normality assumption of linear regression, complementing statistical tests like the Shapiro-Wilk test. In summary, Q-Q plots are a valuable tool for assessing the normality of residuals in linear regression, enabling the identification of potential issues and guiding model improvement.