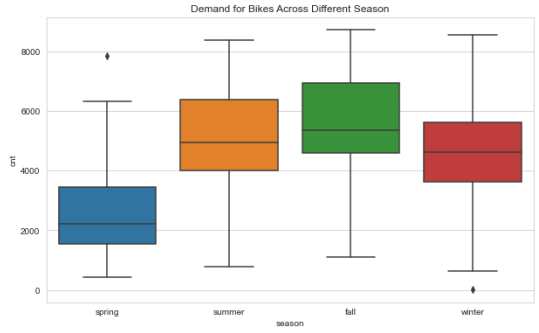
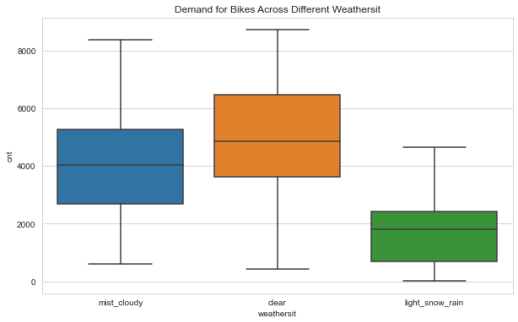
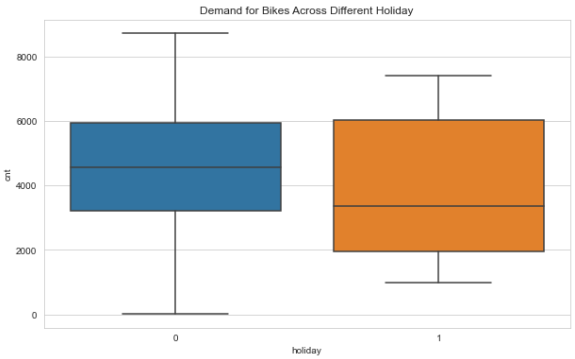
# Assignment-based Subjective Question

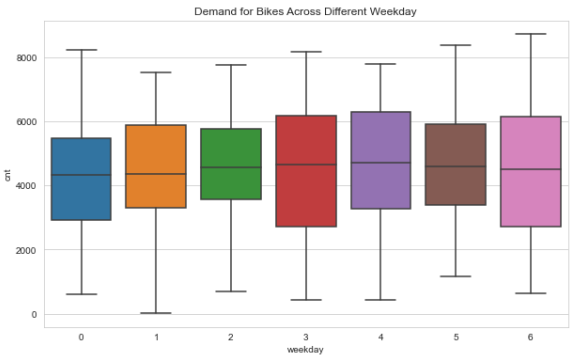
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

**Ans:**




**Annova table for reason:**

```
ANOVA Table for season:
                sum_sq     df           F        PR(>F)
C(season)  9.440517e+08    3.0  127.749824  1.894920e-66
Residual   1.788343e+09  726.0         NaN           NaN


ANOVA Table for weathersit:
                   sum_sq     df          F        PR(>F)
C(weathersit)  2.693719e+08    2.0  39.754686  4.123203e-17
Residual       2.463023e+09  727.0        NaN           NaN


ANOVA Table for holiday:
                 sum_sq     df         F    PR(>F)
C(holiday)  1.292000e+07    1.0  3.458668  0.063324
Residual    2.719475e+09  728.0       NaN       NaN


ANOVA Table for weekday:
                 sum_sq     df        F   PR(>F)
C(weekday)  1.795875e+07    6.0  0.79723  0.57222
Residual    2.714436e+09  723.0      NaN      NaN
```

The categorical variables from the dataset suggests that season and weather conditions have significant effects on the demand for bikes. The highest demand is seen in fall and summer, as well as on clear weather days, which indicates favorable conditions for bike rentals. Conversely, spring and adverse weather conditions see a lower demand. Holidays do not show a statistically significant difference in demand, and the day of the week appears to have the least effect on bike rental

numbers. These insights could guide BoomBikes in optimizing their inventory and marketing strategies according to seasonal and weather-related trends.

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

### Ans:

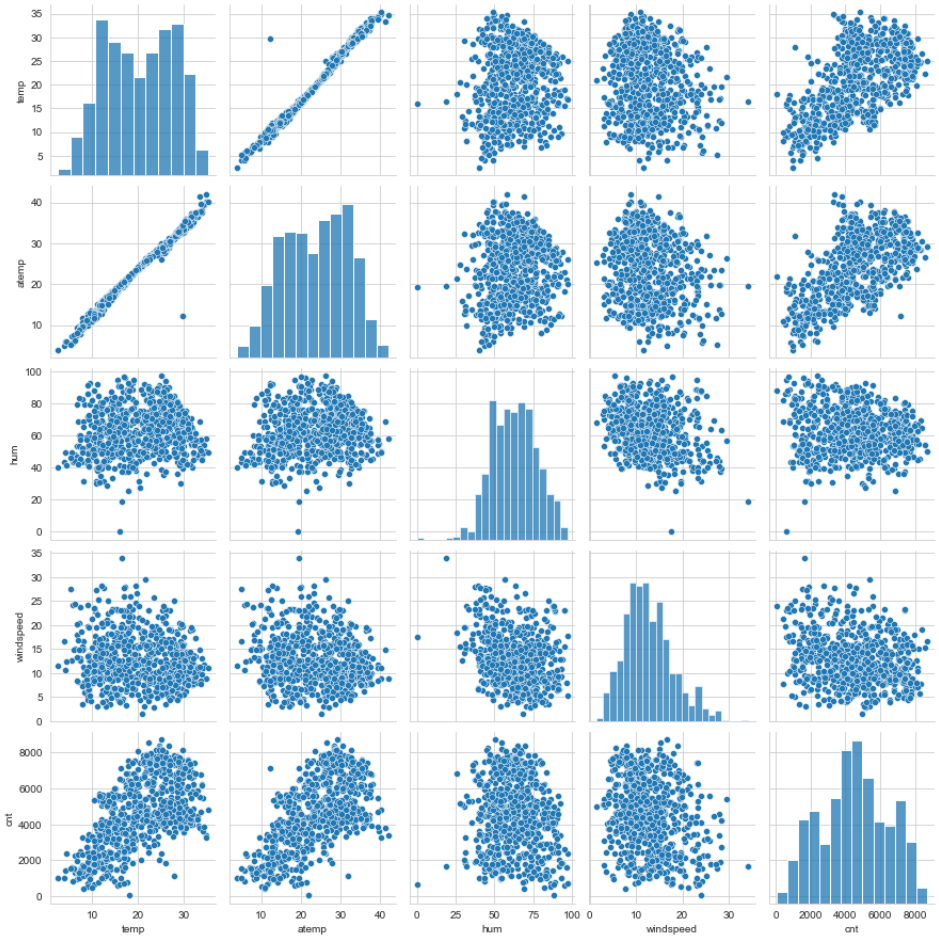
When you convert categorical variables into dummy/indicator variables (also known as one-hot encoding), you typically create a new binary (0/1) variable for each level of the categorical variable. However, this can lead to multicollinearity, which is a problem in regression models where two or more variables are highly correlated.

To avoid multicollinearity, it's important to use the drop_first=True parameter in functions like pd.get_dummies(). This parameter drops the first category level and creates dummy variables for the remaining levels. Here's why this is important:

1. **Redundancy Removal:** In the case of categorical variables, one level can always be inferred from the others.
2. **Avoiding Multicollinearity:** Including all dummy variables for a categorical variable in a regression model will result in perfect multicollinearity. This happens because the sum of all dummy variables equals one for each observation, leading to a situation where one variable can be perfectly predicted from the others. Dropping the first dummy variable helps in removing this perfect multicollinearity.
3. **Interpretation and Simplicity:** Dropping one category simplifies the model without losing interpretative power. The coefficients of the included dummy variables represent the change in the response variable relative to the dropped category. This makes it easier to interpret the coefficients in terms of differences from a baseline category.

**In summary,** using drop_first=True is a best practice in dummy variable creation as it helps in avoiding multicollinearity, simplifies the model, and retains the interpretability of the regression coefficients. This practice is especially crucial in linear regression models, where multicollinearity can significantly impact the estimates of the coefficients.

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

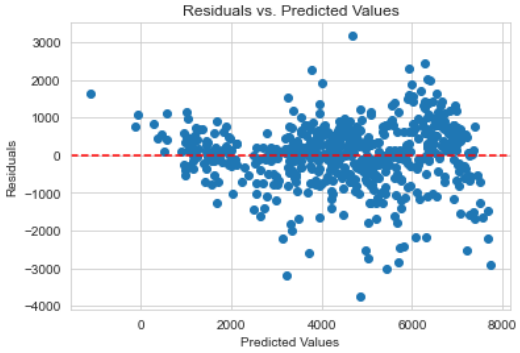



```
cnt           1.000000
registered    0.945411
casual        0.672123
atemp         0.630685
temp          0.627044
yr            0.569728
mnth          0.278191
weekday       0.067534
workingday    0.062542
holiday      -0.068764
hum          -0.098543
windspeed    -0.235132
```

"In examining the pair plot and the correlation coefficients among the numerical variables, 'atemp' and 'temp' show the highest correlation with the target variable 'cnt', which represents the total count of bike rentals. Specifically, 'atemp' has a correlation coefficient of approximately 0.63, followed closely by 'temp' at 0.63, both indicating a strong positive relationship with bike demand. This suggests that as the perceived temperature increases, the number of bike rentals also tends to increase, likely due to more favorable biking conditions during warmer weather. Conversely, 'hum' (humidity) and 'windspeed' are negatively correlated with bike demand, hinting that less comfortable weather conditions could deter bike rentals."

## Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

**Linearity:**




Visualizing the relationship between predicted values and residuals

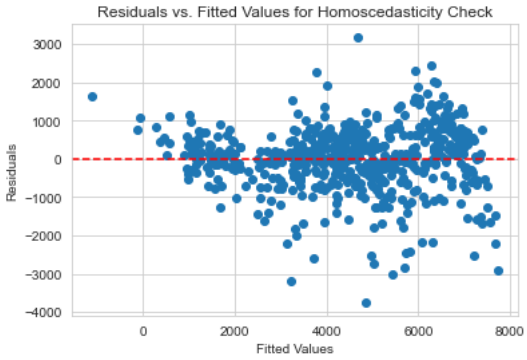
The scatter plot of residuals vs. predicted values does not exhibit any distinct patterns or systematic deviations, suggesting that the linearity assumption is reasonable for our model

**Independence:**

```
Durbin-Watson test: 2.03850248473114
```

The Durbin-Watson test yields a value of approximately 2.038, which is close to the ideal value of 2, indicating a lack of autocorrelation among the residuals and supporting the independence assumption

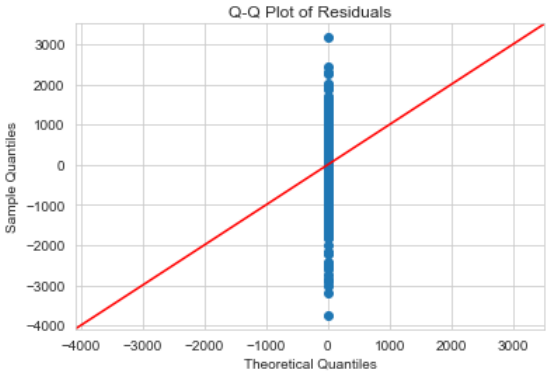
**Homoscedasticity:**




Visualizing the residuals vs. fitted values for constant variance

The residuals vs. fitted values plot shows a random scatter of points without a discernible pattern, which suggests that the homoscedasticity assumption holds and the residuals have constant variance across predictions

**Normality of Residuals**




**Q-Q plot**

The Q-Q plot of the residuals aligns closely with the theoretical quantiles line, with only minor deviations at the tails. This indicates that the residuals are approximately normally distributed, satisfying the normality assumption.

**`Multicollinearity`**

The VIF values for all features included in the model are well below the threshold of 5 or 10, suggesting that there is no concerning level of multicollinearity within our predictors. This is positive as it implies that our model coefficients can be estimated with a higher degree of reliability.

**`VIF calculation`**

| | Feature | VIF |
|---|---|---|
| 0 | const | 60.211542 |
| 1 | yr | 1.012265 |
| 2 | holiday | 1.088015 |
| 3 | weekday | 1.014587 |
| 4 | workingday | 1.089879 |
| 5 | temp | 3.439644 |
| 6 | windspeed | 1.093747 |
| 7 | season_spring | 4.548809 |
| 8 | season_summer | 1.963401 |
| 9 | season_winter | 2.961192 |
| 10 | weathersit_light_snow_rain | 1.064431 |
| 11 | weathersit_mist_cloudy | 1.033469 |

**Summary:**

Our analysis indicates that the regression model fulfills the necessary assumptions for linear regression. The linearity and independence of the residuals are affirmed by visual inspection and the Durbin-Watson statistic, respectively. Homoscedasticity is supported by the lack of patterns in the residuals vs. fitted values plot, and the normal distribution of residuals is confirmed by the Q-Q plot. Lastly, multicollinearity does not appear to be a concern based on the VIF values, suggesting that our model's predictors are sufficiently independent. These validations give us confidence in the reliability of our model's findings.

**Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | cnt | R-squared: | 0.817 |
| Model: | OLS | Adj. R-squared: | 0.814 |
| Method: | Least Squares | F-statistic: | 232.7 |
| Date: | Tue, 21 Nov 2023 | Prob (F-statistic): | 5.91e-203 |
| Time: | 00:57:57 | Log-Likelihood: | -4757.3 |
| No. Observations: | 584 | AIC: | 9539. |
| Df Residuals: | 572 | BIC: | 9591. |
| Df Model: | 11 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 1663.6320 | 270.843 | 6.142 | 0.000 | 1131.665 | 2195.599 |
| yr | 2017.3534 | 70.255 | 28.715 | 0.000 | 1879.364 | 2155.343 |
| holiday | -504.8612 | 205.214 | -2.460 | 0.014 | -907.926 | -101.797 |
| weekday | 72.4077 | 17.452 | 4.149 | 0.000 | 38.130 | 106.686 |
| workingday | 142.9892 | 77.672 | 1.841 | 0.066 | -9.569 | 295.547 |
| temp | 3963.7122 | 285.104 | 13.903 | 0.000 | 3403.733 | 4523.691 |
| windspeed | -958.1061 | 195.114 | -4.910 | 0.000 | -1341.334 | -574.879 |
| season_spring | -864.3853 | 177.044 | -4.882 | 0.000 | -1212.120 | -516.650 |
| season_summer | 322.9471 | 112.949 | 2.859 | 0.004 | 101.102 | 544.792 |
| season_winter | 667.6139 | 137.780 | 4.846 | 0.000 | 396.998 | 938.230 |
| weathersit_light_snow_rain | -2417.6481 | 202.977 | -11.911 | 0.000 | -2816.320 | -2018.976 |
| weathersit_mist_cloudy | -614.3891 | 74.686 | -8.226 | 0.000 | -761.082 | -467.696 |

| | | | |
|---|---|---|---|
| Omnibus: | 70.112 | Durbin-Watson: | 2.039 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 161.641 |
| Skew: | -0.653 | Prob(JB): | 7.94e-36 |
| Kurtosis: | 5.221 | Cond. No. | 47.2 |

**Summary:**

1. **Temperature (temp)**: With a coefficient of approximately 3963.71, temperature has the most substantial positive impact on bike demand. This suggests that as the temperature increases, so does the number of bike rentals, likely due to more comfortable riding conditions.
2. **Weather Situation (weathersit_light_snow_rain)**: This variable has a coefficient of approximately -2417.65, indicating a significant negative impact on bike demand. The negative sign implies that light snow or rain reduces the number of bike rentals, which can be attributed to less favorable or unsafe riding conditions.
3. **Year (yr)**: The year feature has a coefficient of approximately 2017.35, reflecting the increasing trend in bike rentals from year to year. This positive coefficient highlights the growth of the bike-sharing system's popularity over time.

# General Subjective Question

## Q1. Explain the linear regression algorithm in detail.

**Ans:** Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables. The "linear" part means that the algorithm assumes a straight-line relationship between the variables. Here's how it works step by step:

1. **Model Specification**: First, we decide which variable is the outcome we're interested in (the dependent variable) and which variables we think might influence that outcome (the independent variables).
2. **Best-Fit Line**: Linear regression aims to find a line that best fits the data points. This line is described by the equation **y = mx + c**, where **y** is the dependent variable, **x** is the independent variable, **m** is the slope of the line, and **c** is the y-intercept.
3. **Estimating Coefficients**: We then calculate the values of **m** (slope) and **c** (intercept) that minimize the difference between the actual data points and the predictions made by our line. These differences are called residuals.
4. **Least Squares Criterion**: The specific method we use to find the best line is called the "least squares" method. It minimizes the sum of the squares of the residuals. Squaring is used because it penalizes larger errors more severely than smaller ones, which tends to give us a line that fits the majority of points well, rather than fitting a few points extremely closely.
5. **Computing the Fit**: Mathematically, finding the slope and intercept that minimize the sum of squared residuals usually involves solving two equations simultaneously. These are derived from calculus and linear algebra and give us the "least squares estimates."
6. **Interpreting the Model**: Once we have our estimates of **m** and **c**, we can interpret them. The slope **m** tells us how much **y** changes for a one-unit change in **x**. The intercept **c** tells us the value of **y** when **x** is zero.
7. **Making Predictions**: With the final model, we can predict the value of **y** for any given value of **x** by simply plugging **x** into our equation and calculating the corresponding **y**.
8. **Model Assumptions**: Linear regression comes with several assumptions, such as the linearity of the relationship, homoscedasticity (constant variance of residuals), independence of residuals, and normality of the error distribution. We check these assumptions through various diagnostic plots and tests to ensure our model is valid.
9. **Goodness of Fit**: We use metrics like R-squared to understand how well our line fits the data. R-squared tells us the proportion of variance in the dependent variable that can be predicted from the independent variable(s).

In essence, linear regression is a foundational tool that helps us understand how one variable affects another and allows us to make predictions based on that understanding. It's widely used across different fields for its simplicity and interpretability.

## Q2. Explain the Anscombe’s quartet in detail.

**Ans:**

Anscombe's quartet refers to four different datasets constructed by the statistician Francis Anscombe. The quartet is famous because all four datasets have almost identical simple statistical properties, yet they look very different when graphed. Each dataset consists of eleven points and was designed to have the same mean, variance, correlation, and regression line (when a linear regression is performed) for both x and y variables.

Anscombe created these datasets to demonstrate the importance of not just relying on statistical properties when analyzing data. He wanted to show that similar summary statistics could lead to vastly different distributions and relationships when you actually visualize the data.

The quartet teaches us a crucial lesson: always graph your data before interpreting it. This is because the underlying relationship between variables can be lost if we only look at summary statistics. The datasets show scenarios like linear relationships, non-linear relationships, and even one dataset where one outlier has a huge effect on the statistical properties.

In practice, this means that when we analyze data and find that certain statistics fit a particular narrative, it's essential to plot the data to ensure there isn't a different story hiding behind the numbers. Anscombe's quartet is a cornerstone in data analysis, emphasizing the role of data visualization to complement statistical analysis.

## Q3. What is Pearson’s R?

**Ans:**

Pearson's R, also known as the Pearson correlation coefficient, is a measure that quantifies the strength and direction of a linear relationship between two continuous variables. It's a statistic that you would use when you want to know how closely two things vary together. Here's how it works and what it means:

1. **Value Range:** Pearson's R can range from -1 to +1. The value tells you about the direction and strength of the relationship between the two variables.
2. **Interpreting Values:**
    - A value of +1 indicates a perfect positive linear relationship. This means if one variable increases, the other variable increases in a perfectly predictable way.
    - A value of -1 indicates a perfect negative linear relationship. Here, if one variable increases, the other decreases in a perfectly predictable way.
    - A value of 0 means there is no linear relationship between the variables.
3. **Calculation:** Mathematically, Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. In simpler terms, it measures how much the variables change together compared to how much they vary individually.
4. **Usage:** You use Pearson's R when you want to find out if there's a linear relationship between things like height and weight, study time and test scores, or temperature and ice cream sales.
5. **Limitations:**
    - It only measures linear relationships. Non-linear relationships are not well captured by this statistic.
    - It is sensitive to outliers. A single outlier can significantly affect the value of the correlation coefficient.
    - It doesn’t imply causation. Just because two variables are correlated doesn’t mean one causes the other.

In summary, Pearson's R is a handy tool for exploring relationships between variables. It gives you a quick sense of whether and how strongly variables are related in a linear way. However, it's always important to remember that correlation does not imply causation and to consider the context and other potential variables involved.

## Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a technique used in data preprocessing, especially in the context of machine learning and statistics, to standardize the range of independent variables or features of data. Here's a detailed explanation:

1. **What is Scaling?**
    - Scaling transforms the values of numerical features to a common scale, without distorting differences in the ranges of values or losing information.
2. **Why is Scaling Performed?**
    - Consistency: Many algorithms, particularly those that use distance measurements (like K-means clustering or K-nearest neighbors), perform better when the features are on the same scale, as it ensures that one feature doesn't dominate others simply due to its scale.
    - Speed: Algorithms converge faster when features are on a similar scale, which is essential for gradient descent algorithms or algorithms that use optimization.
    - Preventing Misleading Interpretations: In regression models, unscaled features can lead to misleading interpretations of feature importance.
3. **Difference Between Normalized Scaling and Standardized Scaling:**
    - **Normalized Scaling (Min-Max Scaling):**
        - Normalization scales the values into a range of [0, 1] or [-1, 1].
        - It's useful when you need to scale the data to fit into a particular range.
        - The formula is (X - X_min) / (X_max - X_min), where X_min and X_max are the minimum and maximum values of the feature respectively.
    - **Standardized Scaling (Z-score Normalization):**
        - Standardization scales data to have a mean of 0 and a standard deviation of 1 (unit variance).
        - It's useful when you want to compare the relative importance of features.
        - The formula is (X - X_mean) / X_std, where X_mean and X_std are the mean and standard deviation of the feature respectively.

## Q5. Why Might the Value of VIF (Variance Inflation Factor) Sometimes Be Infinite?

**Ans:**

The Variance Inflation Factor (VIF) is a measure that quantifies the extent of multicollinearity in an ordinary least squares regression analysis. It assesses how much the variance of a regression coefficient is inflated due to multicollinearity among the predictor variables. Sometimes, the VIF can be infinite, and here's why:

1. **Perfect Multicollinearity:** The primary reason for an infinite VIF is perfect multicollinearity. This situation occurs when one independent variable in the regression model is an exact linear combination of another. In simpler terms, one variable can be perfectly predicted from the others.
2. **Math Behind Infinite VIF:** Mathematically, VIF is calculated as $1/(1-R^2)$, where $R^2$ is the coefficient of determination of a regression of one independent variable on the other independent variables. If there's perfect multicollinearity, $R^2$ becomes 1 (as one variable can perfectly predict another), leading to division by zero (1/(1-1)), which results in an infinite VIF.
3. **Implications:** An infinite VIF indicates that the model needs revision since perfect multicollinearity violates the assumption of no multicollinearity in linear regression. It means that the affected variables do not provide unique or independent information in the context of the model.
4. **Dealing with Infinite VIF:** To address this issue, you would typically remove one of the perfectly collinear variables from your regression model. The choice of which variable to remove depends on your understanding of the variables, their importance, and the context of your study or analysis.

In summary, an infinite VIF is a clear indicator of a serious issue in your regression model, specifically pointing to the presence of perfect multicollinearity. It's a signal to revisit your model's variables and make necessary adjustments for a more accurate and reliable analysis.

## Q6. What is a Q-Q Plot? Explain the Use and Importance of a Q-Q Plot in Linear Regression.

**Ans:**

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a certain distribution, such as the normal distribution. In the context of linear regression, it's primarily used to check the normality of residuals. Here's a detailed explanation:

1. **What is a Q-Q Plot?**
    - A Q-Q plot compares the quantiles of a dataset to the quantiles of a theoretical distribution, allowing us to see if the two distributions are similar.
    - It plots the observed values against the expected values of a distribution, typically the normal distribution.

2. **Use in Linear Regression:**
    - One of the key assumptions in linear regression is that the residuals (the differences between the observed values and the values predicted by the model) are normally distributed.
    - To validate this assumption, we plot the residuals of the regression model on a Q-Q plot against the quantiles of a standard normal distribution.
3. **Importance of Q-Q Plot in Linear Regression:**
    - Assumption Validation: It helps in validating the assumption of normality, which is crucial for the reliability of various statistical tests used in linear regression, such as tests for coefficients' significance.
    - Identifying Deviations: A Q-Q plot can highlight deviations from normality, like skewness or kurtosis. If the residuals lie along the reference line in the plot, they are normally distributed. Deviations from this line indicate non-normality.
    - Informing Model Adjustments: If the Q-Q plot reveals that the residuals are not normally distributed, it may prompt a reevaluation of the model. This could include data transformation, using a different set of variables, or even considering non-linear models.

In conclusion, a Q-Q plot is an essential diagnostic tool in linear regression analysis. It visually assesses whether the residuals of the model adhere to the normal distribution, which is a fundamental assumption for making valid inferences about the significance of predictors and the overall model fit.