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?**

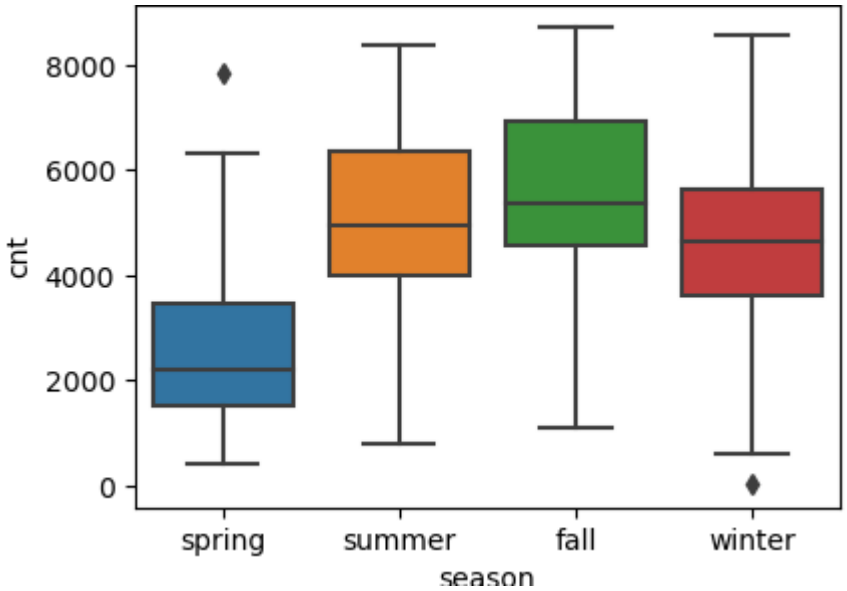
Ans : From our analysis we found below observation:

Dependent variable : cnt

categorical_features : ['season','yr','mnth','holiday',
'weekday','workingday','weathersit']

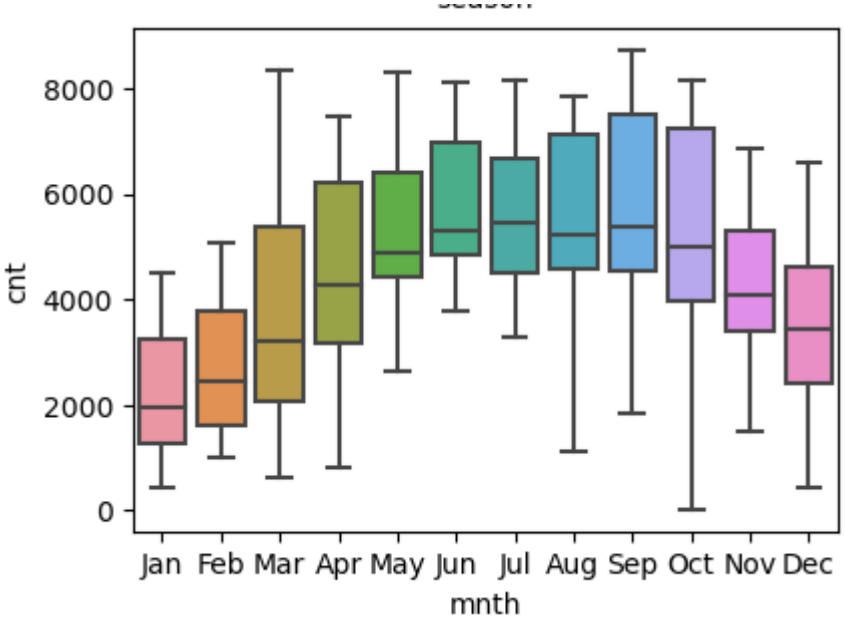
1.Season impact on cnt feature :

1. From the box plot drawn between season and cnt column we can see that on an average during fall season around 5700 user has used boom bikes followed by Summer season with 5500 user.
2. Lowest use of boombike is done during spring season



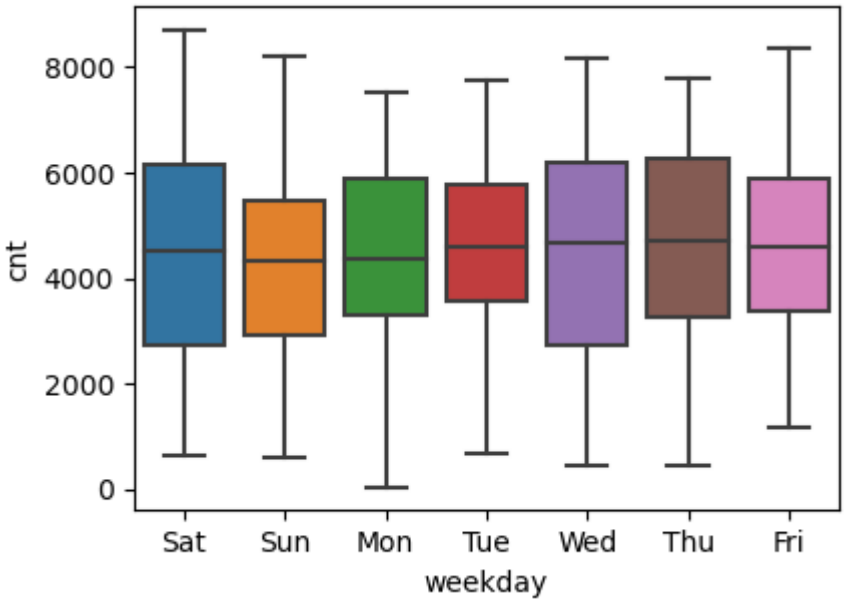
2.Mnth impact on cnt feature:

We can see that most use of the boombike comes in between the month of may to October.Least use of boombike is in the month of Jan.



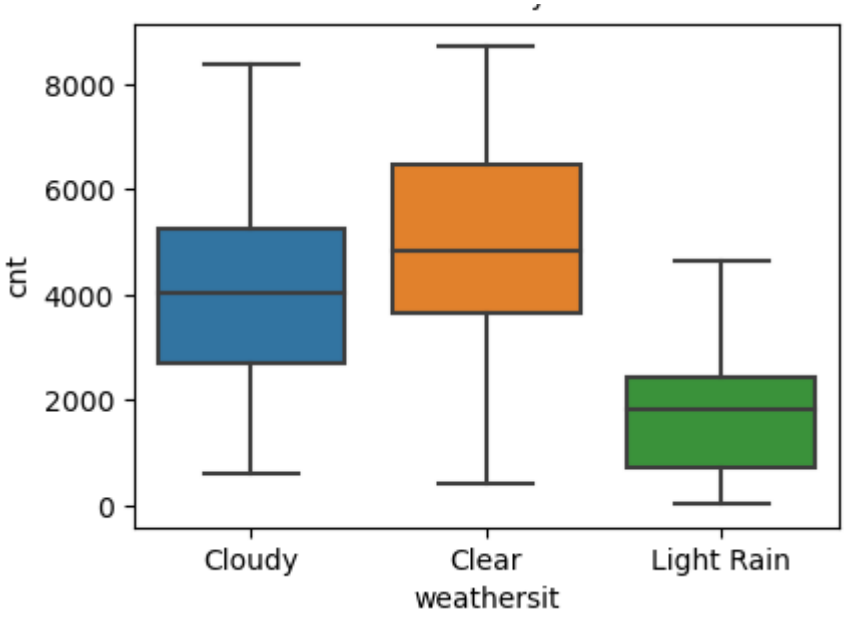
3.Weekday impact on cnt feature:

On an average we see similar use of the boombike throughout the week.



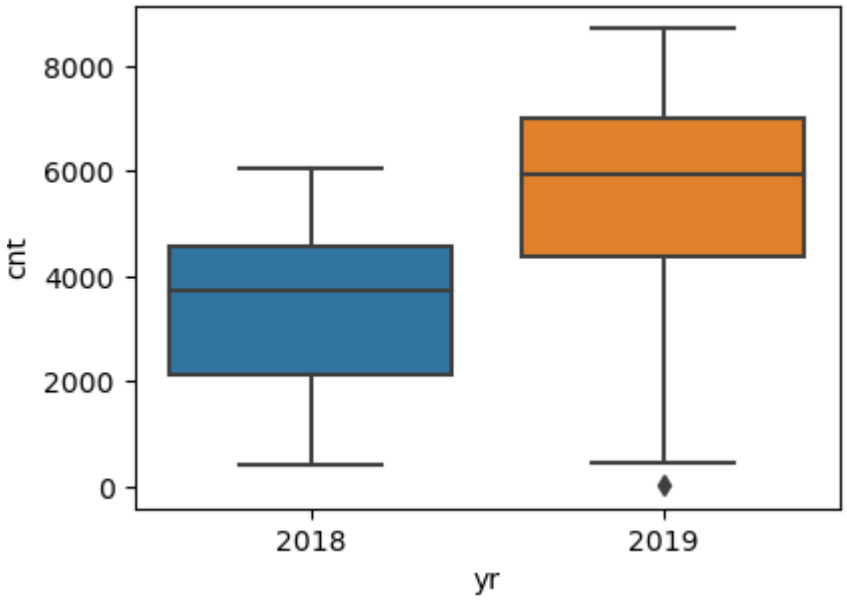
4 weathersit impact on cnt feature:

Maximum uses of the boombike is used during clear weather and then in cloudy season and very less use during rainy season.



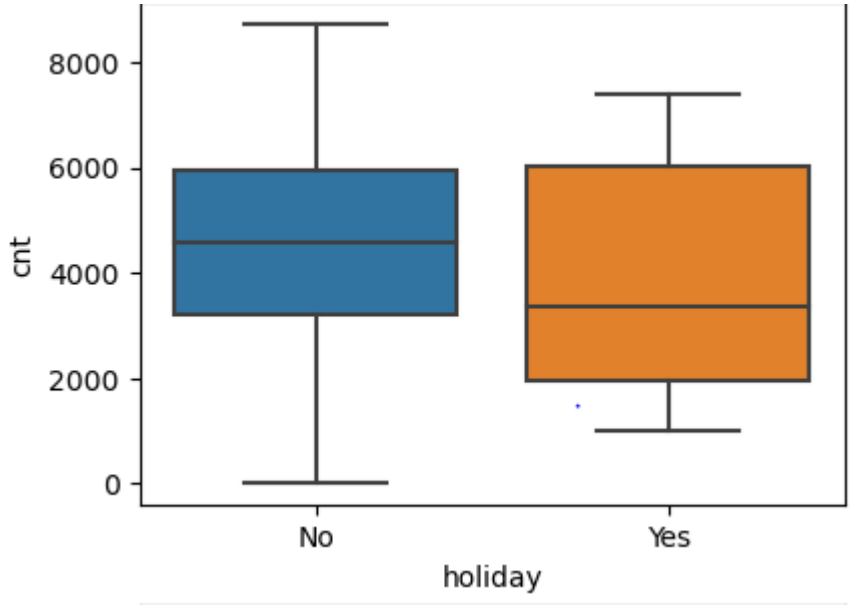
5. Year impact on cnt variable:

We see increase in the usage of the boombike from 2018 to 2019 to almost 70 to 80 %.



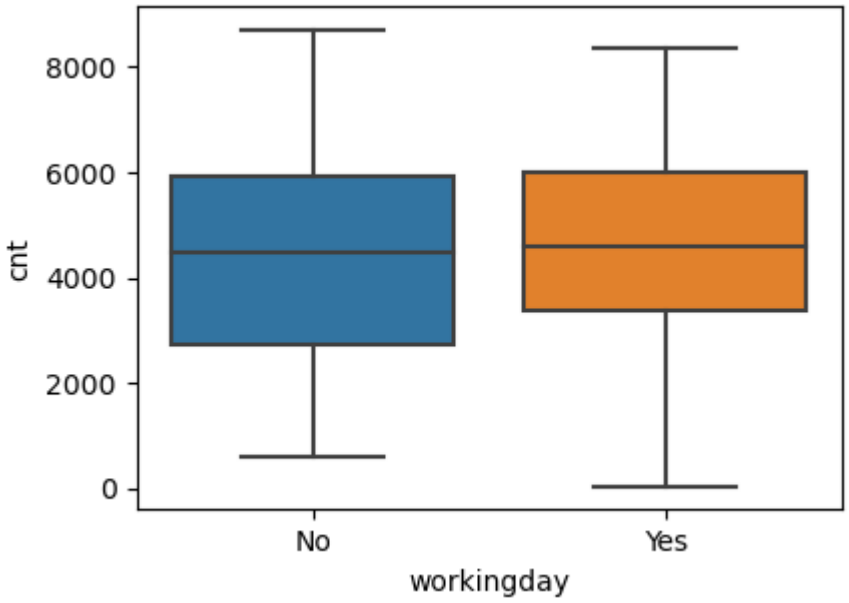
6. Holiday impact on cnt Feature;

During Holiday we see that there is not much usage of boombike in comparison to non-holiday :



7.Weekday impact on cnt feature:

On an average we see similar use of the boombike throughout the working and non working day.



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

Ans: Using drop_first=True during dummy variable creation, is a technique to avoid multicollinearity and improve the interpretability of the model. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can create problems in model estimation and interpretation. Dropping one of the dummy variable categories when creating dummies from a categorical variable helps address this issue.

Advantage of using drop_first=True are mentioned below:

Avoiding Perfect Multicollinearity: In linear regression, if you include all dummy variables representing the categories of a categorical variable, you may end up with perfect multicollinearity. This happens because the sum of the dummy variables will always be equal to 1 for each data point. This means that one of the dummy variables can be perfectly predicted from the others. Including all dummy variables can result in unstable coefficient estimates and inflated standard errors.

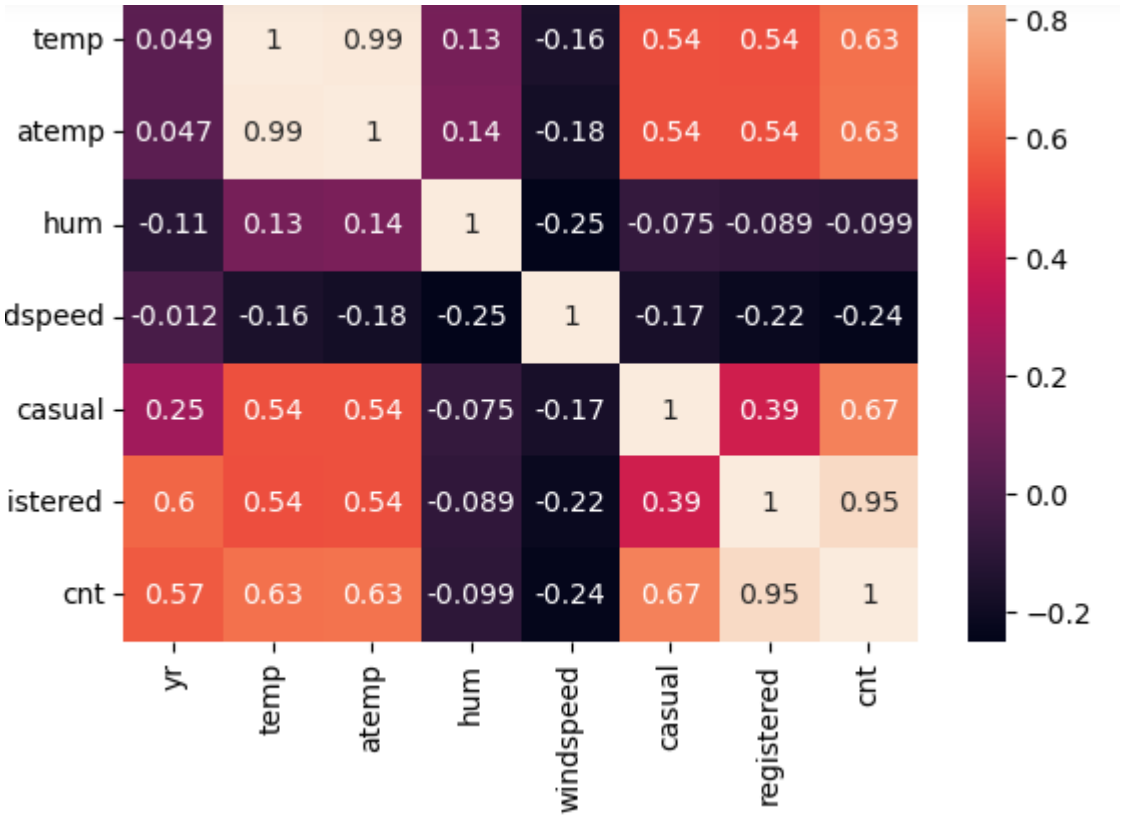
Interpretability: When using dummy variables, you are essentially splitting a categorical variable into binary indicators (0 or 1) for each category. If you include all of them, it becomes challenging to interpret the coefficients of each dummy variable. By dropping one category , you establish a clear reference point for comparison.

Efficiency: Including fewer dummy variables is more efficient in terms of model complexity and computation. It reduces the number of predictors in the model, which can be especially important when dealing with large datasets.

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

Ans : from the heatmap we see that registered user has +0.95 but it also part of the target variable so we can ignore it.

Next highest correlation is with temp +0.63.



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

Ans: The primary assumptions of linear regression include linearity, independence of Residuals, homoscedasticity, and normality of residuals.

Assumption 1: Linearity:

- To check the linearity assumption, create a scatterplot of the residuals against the predicted values .
- Look for any discernible patterns or curvature in the plot. Ideally, the residuals should be randomly scattered around zero without any noticeable trends.

Assumption 2: Independence of Residuals:

- Examine the residuals for any temporal or spatial patterns. If your data is time-series data or spatial data, autocorrelation or spatial autocorrelation could indicate a violation of the independence assumption.

Assumption 3: Homoscedasticity:

- Check for homoscedasticity by creating a scatterplot of the residuals against the predicted values.
- Look for a consistent spread of residuals across all predicted values. If the spread widens or narrows as predictions change, it indicates heteroscedasticity.

Assumption 4: Normality of Residuals:

- Draw distplot to virtually examine the normal distribution of the residuals

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

Ans : based on the corelation value we see that temp,yr,spring value is the 3 top feature contributing significantly towards explaining the demand of the shared bikes

```
temp          0.627044
yr            0.569728
spring        0.561702
```

.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Ans: **Linear Regression** is the supervised Machine Learning model in which the **model finds the best fit linear line between the independent(X) and dependent(Y) variable** i.e it finds the linear relationship between the dependent and independent variable.

Linear Regression is of two types:

1.**Simple Linear Regression**

2.**Multiple Linear Regression**

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable.

equation: y = b0+b1*x

where y = dependent/target variable

x = independent/input variable

b0 is the intercept

b1 is coefficient/slope

Multiple Linear Regression is where more than one independent variables is present for the model to find the relationship.

equation: y = b0+b1*x1+b2*x2+.....+bn*xn

where y = dependent/target variable

x1,x2.....xn = independent/input variable

b0 is the intercept

b1,b2....bn is coefficient/slopes

The basic assumptions of Linear Regression are as follows:

Linearity: The relationship between input features and the target variable is linear.

Independence: The errors (residuals) are independent of each other.

Homoscedasticity: The variance of the errors is constant across all levels of the input features.

Normality: The errors are normally distributed.

2: **Explain the Anscombe’s quartet in detail**

Ans: Anscombe's quartet highlights the importance of data visualization. Descriptive statistics like means, variances, and correlations can be misleading, as different datasets can have similar summary statistics but exhibit distinct patterns.

It emphasizes the concept that a simple linear regression analysis or the reliance on summary statistics alone may not be appropriate for understanding complex data relationships.

In practice, when working with data, it is crucial to explore the data visually through plots and charts to gain insights into its underlying structure and relationships.

This quartet serves as a cautionary example in statistics and data analysis, reminding researchers and analysts to not make assumptions about their data solely based on summary statistics and to be aware of the limitations of such statistics when interpreting results.

The four datasets within Anscombe's quartet are labeled I, II, III, and IV. Here is a brief overview of each dataset:

Dataset I:

Simple linear relationship between x and y.

A scatterplot of the data points shows a clear linear trend.

The equation for the linear relationship is approximately $y = 3x + 2$.

Dataset II:

Non-linear relationship with a strong outlier.

The data forms an inverted U-shape, but a single outlier significantly affects the regression line.

The equation for the linear relationship is also approximately y = 3x + 2 when the outlier is ignored.

Dataset III:

Perfect linear relationship except for one outlier.

All points line up on a straight line except for one data point, which is far from the line.

This single outlier can significantly affect the regression analysis.

Dataset IV:

No linear relationship, but simple summary statistics are similar.

The data points do not exhibit any linear relationship, yet the mean, variance, and correlation coefficient for this dataset are quite similar to the other datasets.

3. **What is Pearson's R?**

Ans: Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure of the linear relationship or correlation between two continuous variables. It quantifies the strength and direction of the linear association between the two variables. Pearson's r has a value between -1 and 1, with the following interpretations:

r = 1: Perfect positive linear correlation. This means that as one variable increases, the other increases in a linear fashion.

r = -1: Perfect negative linear correlation. As one variable increases, the other decreases in a linear fashion.

r ≈ 0: Little to no linear correlation. There is no strong linear relationship between the variables.

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

Ans: **Scaling** is a preprocessing technique in data analysis and machine learning that involves transforming the features (variables) of a dataset to a common scale. The primary goal of scaling is to ensure that the variables have similar magnitudes, which can be important for various algorithms and models. Scaling is typically performed on numerical features and is especially essential in cases where features have significantly different ranges or units of measurement.

Here are some reasons why scaling is performed:

1. **Improved Model Performance**: Many machine learning algorithms are sensitive to the scale of the input features. Scaling can help improve the model's convergence speed and overall performance.
2. **Equal Weighting**: Scaling ensures that each feature contributes equally to the analysis. Without scaling, features with larger ranges can dominate and overshadow the impact of other features.
3. **Interpretability**: Scaling makes it easier to interpret the coefficients in linear models. It ensures that the coefficients represent the effect of a one-unit change in the corresponding feature.
4. **Regularization**: Regularization techniques like L1 and L2 regularization assume that features are on a similar scale. Scaling ensures that regularization operates uniformly across all features

There are two common scaling techniques: **normalized scaling** and **standardized scaling**.

1.Normalized Scaling (Min-Max Scaling):Also known as Min-Max scaling, this method scales the features to a specific range, usually between 0 and 1. Normalized scaling is useful when you want to preserve the relationships between the data points but ensure that they all fall within a common range.

2 Standardized Scaling (Z-score Standardization):Also known as Z-score standardization, this method scales the features to have a mean of 0 and a standard deviation of 1. Standardized scaling transforms the data to have a Gaussian distribution with a mean of 0 and a standard deviation of 1. This is helpful for many statistical and machine learning algorithms that assume normally distributed data.

The choice between normalized and standardized scaling depends on the specific requirements of your analysis or the machine learning algorithm you plan to use. In practice, both methods are widely employed, and it is essential to consider the characteristics of your data and the specific needs of your modeling task when deciding which scaling method to use.

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

Ans: The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in a multiple linear regression model. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it challenging to determine their individual effects on the dependent variable. A high VIF for a variable indicates that it has a strong linear relationship with one or more of the other independent variables in the model.

In the context of VIF calculations, a VIF value can become infinite (or very close to infinity) due to perfect multicollinearity. This happens when one or more independent variables can be perfectly predicted by a linear combination of the other independent variables in the model.

Formula : VIF(X) = 1/(1-R2) Where R2 is the coefficient of determination (R-squared) for the regression of X against all the other independent variables in the model.

If R2 is equal to 1, it means that X can be perfectly predicted from the other independent variables, resulting in a denominator of $1-1=0$. This causes the VIF to be infinite.

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

Ans: A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution, such as the normal distribution. It is a visual method for comparing the quantiles (ordered values) of dataset to the quantiles of a theoretical distribution, often the normal distribution. The Q-Q plot helps you determine whether your data fits the expected distribution or if there are deviations from it.

Construction of a Q-Q Plot:

1. Start by sorting the data in dataset in ascending order.
2. Calculate the expected quantiles for a theoretical distribution (e.g., the normal distribution) corresponding to the same dataset size.
3. Plot the sorted data points against the expected quantiles on a scatterplot. The x-axis represents the quantiles from the theoretical distribution, while the y-axis represents the quantiles from your dataset.
4. If the points closely follow a diagonal line it suggests that data follows the theoretical distribution. Deviations from the line indicate departures from the expected distribution.

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

1. **Distribution Assessment**: Q-Q plots are valuable for assessing whether the residuals in a linear regression model follow a normal distribution. In linear regression, the assumption of normally distributed residuals is important for the validity of statistical tests and confidence intervals.
2. **Identifying Departures from Normality**: If the Q-Q plot shows significant deviations from the diagonal line (i.e., the quantiles of the residuals do not match the expected quantiles of the normal distribution), it indicates that the residuals are not normally distributed. This can alert you to potential issues with the linear regression model.
3. **Model Validity**: The Q-Q plot can be a diagnostic tool to check the validity of the linear regression model. Deviations from the expected distribution may suggest that the linear regression assumptions are violated, which could impact the reliability of the model's results.
4. **Outlier Detection**: In the context of linear regression, Q-Q plots can also help identify outliers or extreme values in the residuals. Outliers can have a disproportionate influence on the model, so it's essential to detect and address them.

5. **Model Improvement**: If a Q-Q plot reveals non-normality in the residuals, you may consider transformations or alternative modeling techniques to improve the model's performance and reliability.