Subjective Questions

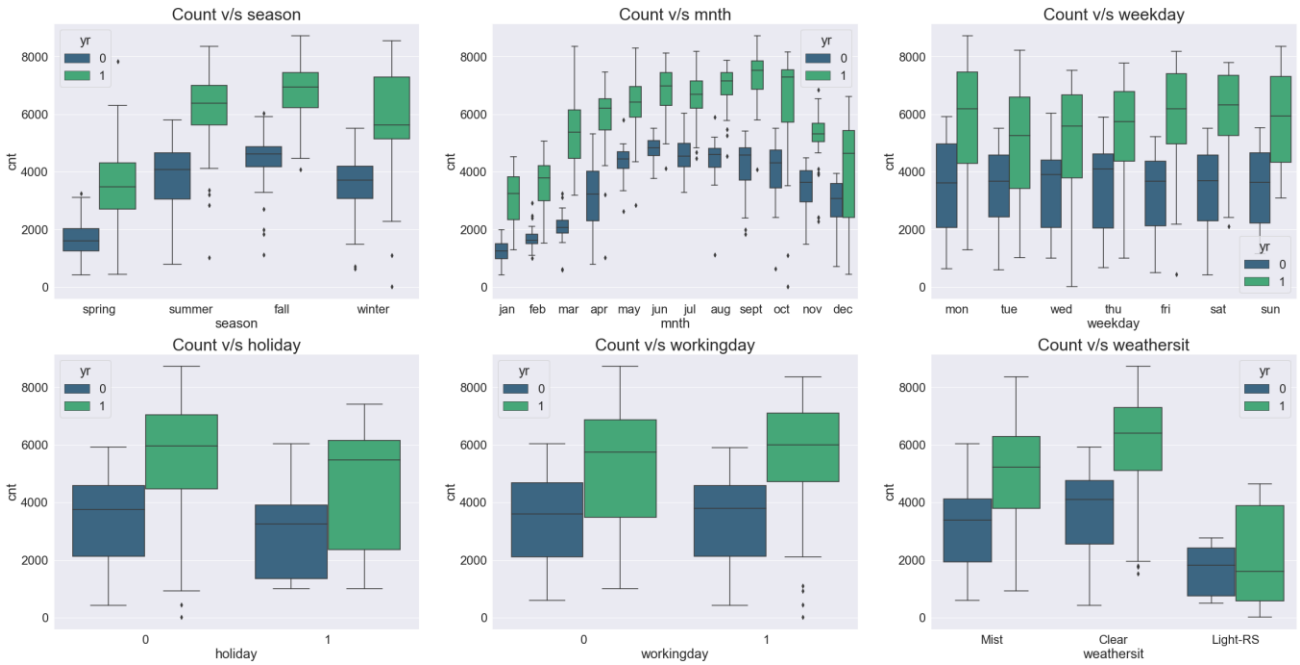
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:

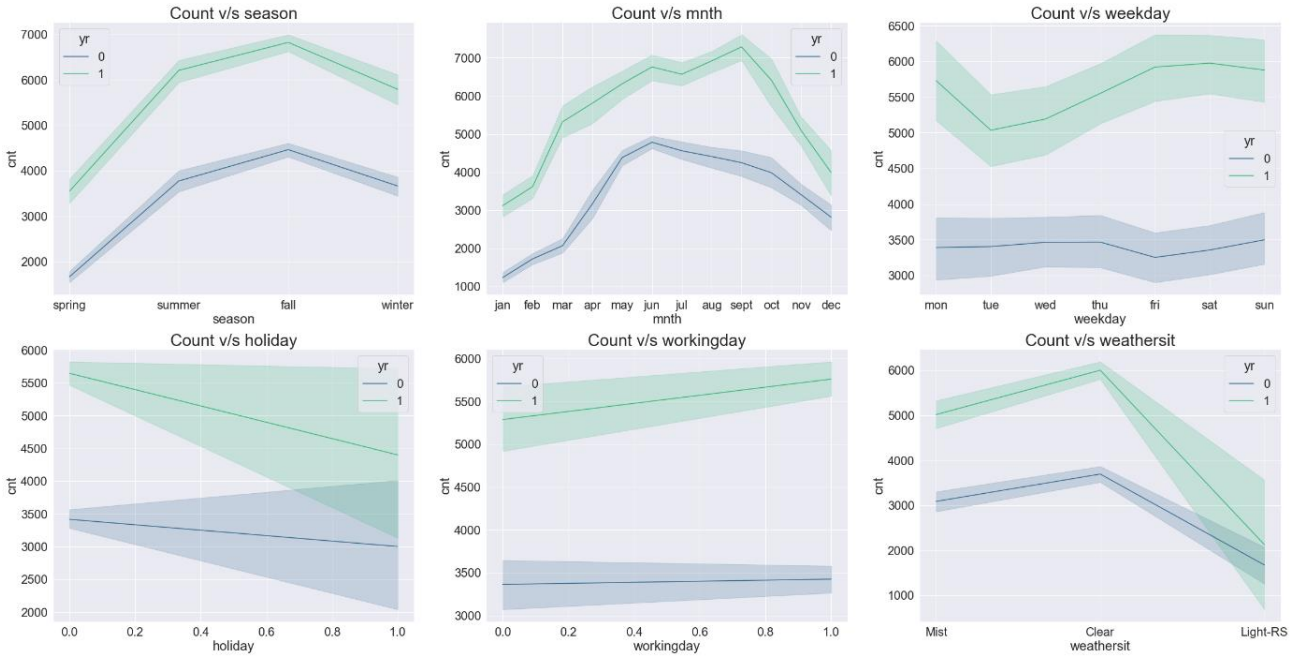
The following observations can be noted from the analysis of categorical variables in the dataset:

- The hiring for bikes have increased for 2019 than 2018.
- Temperature is having a positive correlation with demand. And higher temp attracts more bike hiring.
- The features **temp** and **atemp** are highly correlated and essentially one is derived from the other.
- Fall shows a higher demand followed by summer season.
- May to October period also shows a higher demand then other parts of the year.
- Working days are having a higher demand then holidays and weekends.
- Clear weather days attract a hire demand then other days.
- Wind-speed shows somewhat negative relation with count.

Box-plot for Categorical Variables.



Line-plot for Categorical Variables.



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

Ans:

The **drop_first=True** helps to drop the redundant column created during dummy variable creation. As seen in the example code below:

At index 3 where A is 1, B & C are 0. Thus if both columns indicate 0 then it automatically infers that A is 1, for that we do not need to keep the extra variable.

Code:

Input:

```
import pandas as pd
data = {'Trial': ["A","B","C","A","B"]}
pd.DataFrame.from_dict(data)
data1 = pd.get_dummies(data["Trial"])
print(f"Without drop_first : \n {data1}")
data2 = pd.get_dummies(data["Trial"], drop_first = True)
print(f"With drop_first : \n {data2}")
```

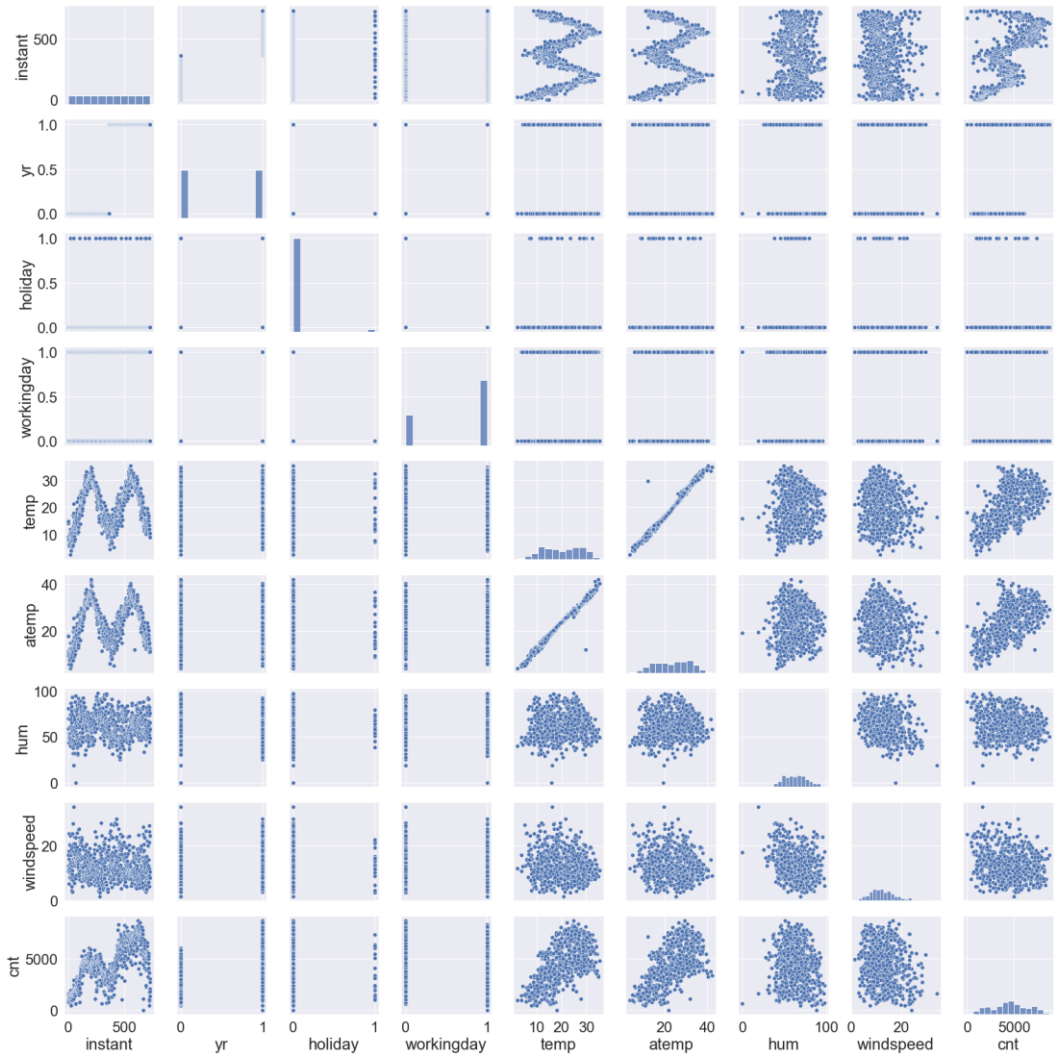
Output:

```
Without drop_first :
    A  B  C
0   1  0  0
1   0  1  0
2   0  0  1
3   1  0  0
4   0  1  0
With drop_first :
    B  C
0   0  0
1   1  0
2   0  1
3   0  0
4   1  0
```

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

Ans:

Pair-plot for the features:



The categorical columns **temp** & **atemp** have the highest correlation with the target variable.

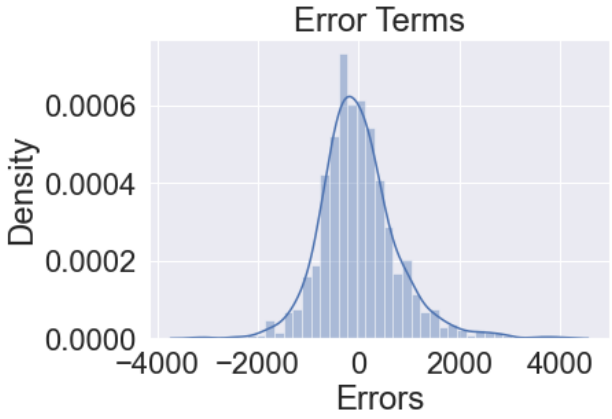
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

There are four assumptions associated with a linear regression model:

- **Linearity:** The relationship between X and the mean of Y is linear.
- **Homoscedasticity:** The variance of residual is the same for any value of X.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of X, Y is normally distributed.

From the study a linear relationship between the terms have been seen. The variance of residual is same and as normally distributed. The error terms are also normally distributed around mean 0 and no visible pattern is seen in the distribution of the error points.



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

Ans:

From the analysis it can be seen that the most significant features are:

- Temperature (Positively related).
- Year (Positively related).
- Windspeed (Negatively related).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

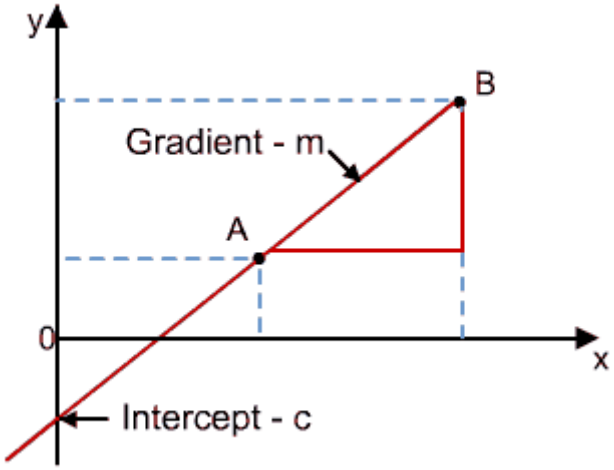
To get the best fit line the linear regression model follows the best fit line concept.

i.e., **y = mx +c**

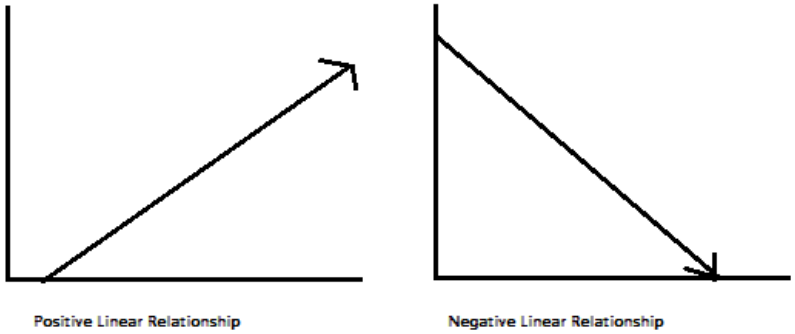
where y is the dependent Variable we are trying to predict.

x is the independent variable

m is the slope and c is the intercept.



Put simply, a linear relationship implies some constant straight line relationship. Also the linear relationship can be positive or negative in nature.



Linear Regression can also be categorised into the following types:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression

There are four assumptions associated with a linear regression model:

- **Linearity:** The relationship between X and the mean of Y is linear.
- **Homoscedasticity:** The variance of residual is the same for any value of X.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of X, Y is normally distributed.

The three most common evaluation metrics for regression problems:

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

It is given as:

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

Mean Squared Error (MSE) is the mean of the squared errors:

It is given as:

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

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

It is given as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Comparing these metrics:

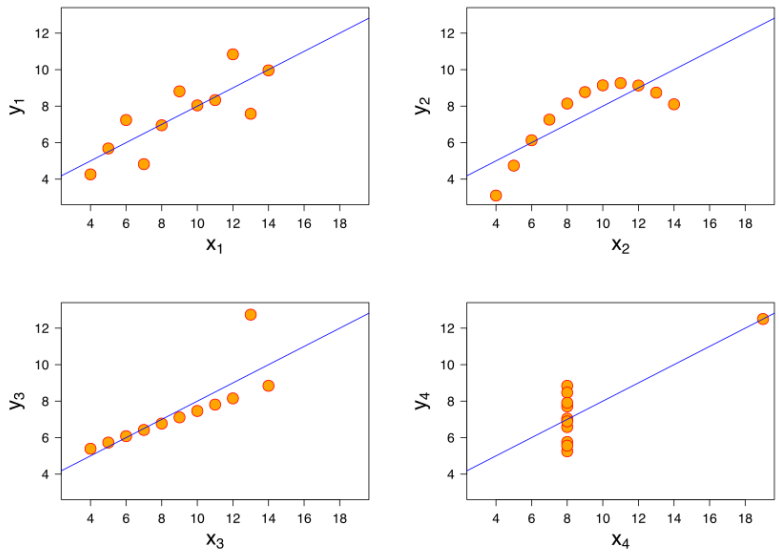
- **MAE** is the easiest to understand, because it's the average error.
- **MSE** is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.
- **RMSE** is even more popular than MSE, because RMSE is interpretable in the "y" units.

All of these are **loss functions**, because we want to minimize them.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x :	11	exact
Mean of *y*	7.5	to 2 decimal places
Sample variance of *y* :	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

The datasets are as follows. The *x* values are the same for the first three datasets.

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50

12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.

3. What is Pearson's R?

Ans:

In statistics, **Pearson's r —** also known as the Pearson correlation coefficient (PCC), the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient— is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

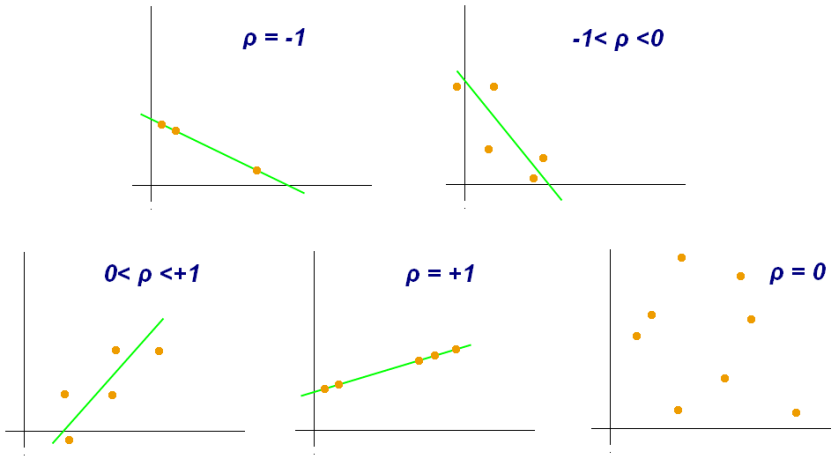


Fig: Examples of scatter diagrams with different values of correlation coefficient (ρ).

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

Ans:

Feature scaling:

It is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

The following are the types of scaling used:

Min-Max Normalization:

This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization:

It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example:

If an algorithm is not using the feature scaling method then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

Differences between Standard Scaler and Min-Max Scaler:

Sl. No.	Normalized Scaling	Standard Scaling
1	Minimum & Maximum values of the Features are used.	Mean & Standard Deviation is used.
2	It is highly affected by outliers.	Not much affected by outliers.
3	Scaled data lies between [0,1] or [-1,1].	Not bounded to a certain range.
4	MinMaxScaler transformer fron Scikit Learn is used.	StandardScaler is used

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

Ans:

VIF is calculated as:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

In case of Perfect Correlation R-squared value becomes 1. Thus the 1-R^2 value becomes 0 and the factor VIF becomes infinity.
To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans:

Q-Q Plot:

A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two distributions being compared, some of the quantiles may not be defined, so an interpolated quantile may be plotted. If the Q–Q plot is based on data, there are multiple

quantile estimators in use. Rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions.

A simple case is where one has two data sets of the same size. In that case, to make the Q–Q plot, one orders each set in increasing order, then pairs off and plots the corresponding values. A more complicated construction is the case where two data sets of different sizes are being compared. To construct the Q–Q plot in this case, it is necessary to use an interpolated quantile estimate so that quantiles corresponding to the same underlying probability can be constructed.

Use and Importance of Q-Q Plot:

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.