**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?**

## Bike Rentals are more during the Fall (Monsoon) season.

### Bikes seem to be rented more in Partly cloudy weather

### Bikes seem to be rented more on working days.

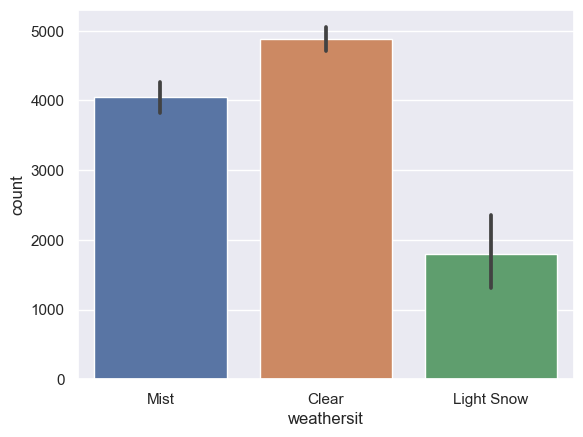
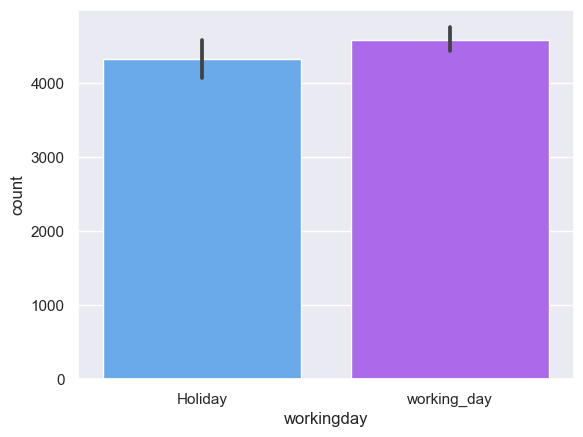
### Bike Rental popularity has increased in 2019 when compared to 2018.

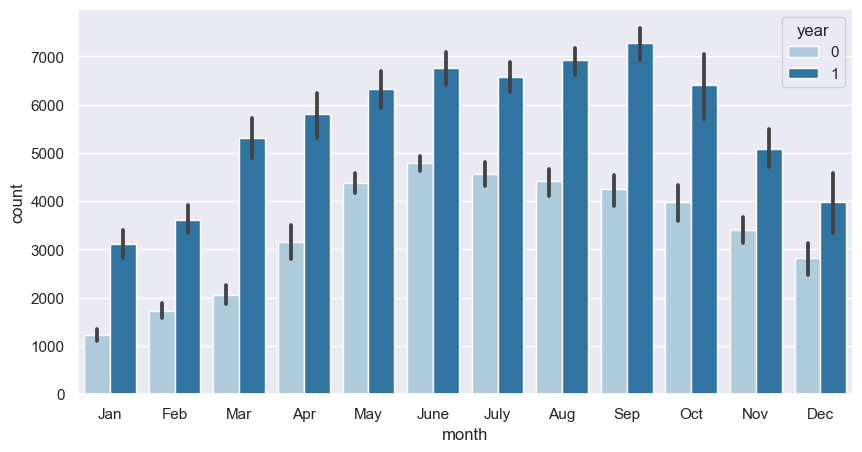
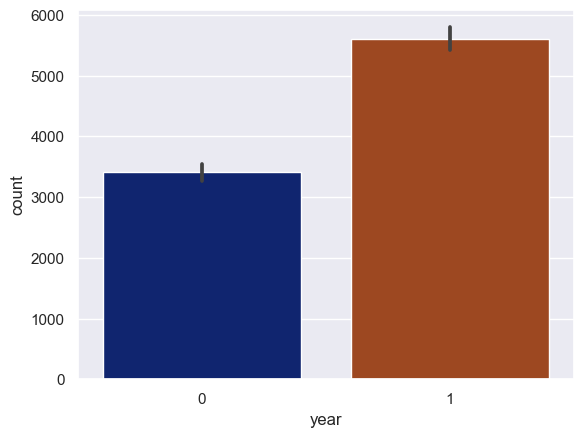
### Bike Rentals are maximum on Sunday and Monday

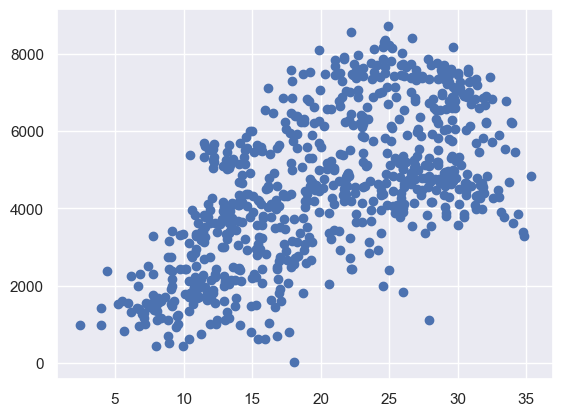
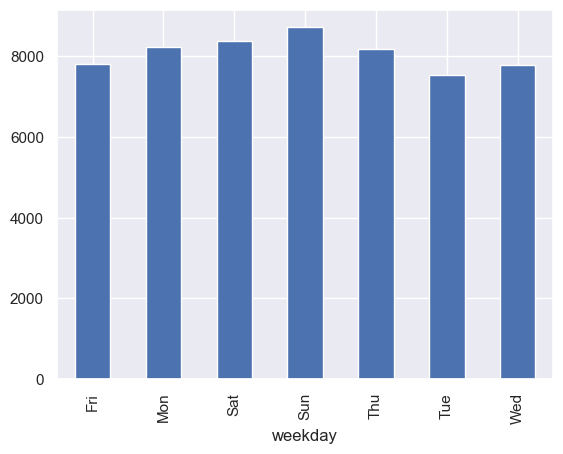
### Bike Rentals are observed at higher temperatures.

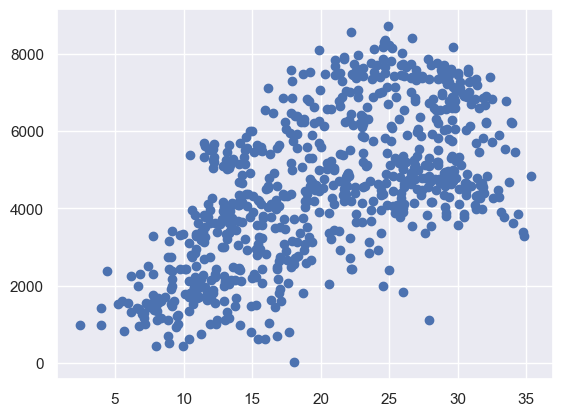
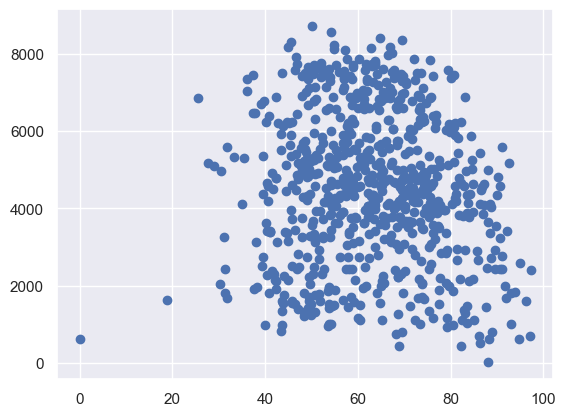
### Bike Rentals are observed at higher "feel-like" temperatures.

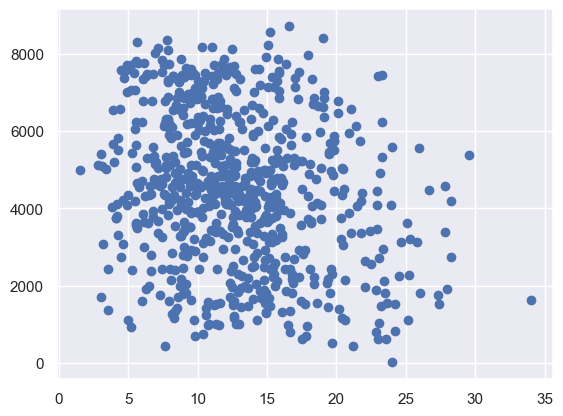
### Temperature being directly proportional to Humidity, Bike Rentals are making during high humidity.











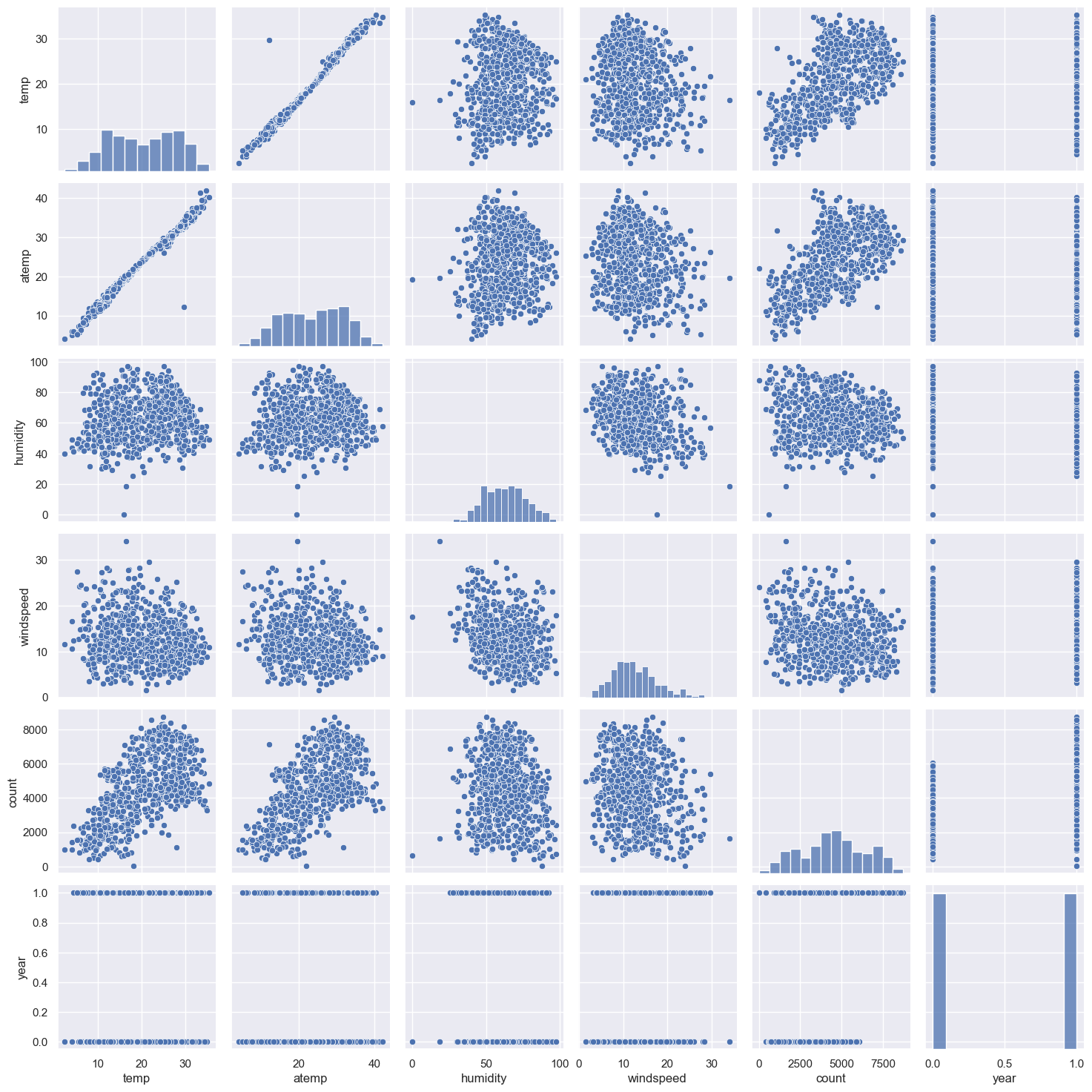
1. **Why is it important to use drop\_first=True during dummy variable creation?**

If we don't drop the first column then the dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted.

It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

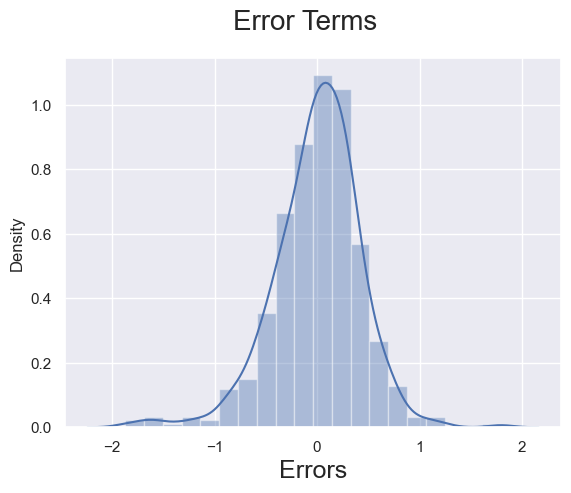
1. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

* atemp and temp are highly correlated among all numerical variables.

****

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

The following tests were done to validate the assumptions of linear regression: 1. First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pair plot to see if the variables are linearly related or not. Refer to the notebook for more details. 2. Secondly, Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a dist plot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.

****

3. Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

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

* Sun = 0. 2749
* Working\_day = 0.2327
* Temperature = 0.3999

**General Subjective Questions**

1. **Explain the linear regression algorithm in detail.**

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

y = a + bx

Where a and b given by the formulas:

b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^{2}-(\sum x)^{2}}

a(intercept) = \frac{n\sum y - b(\sum x)}{n}

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

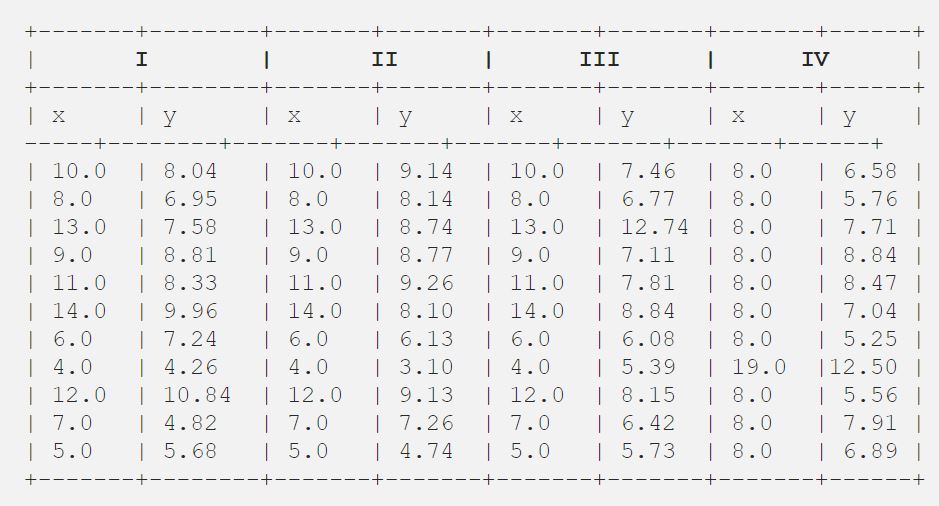
y = Dependent variable from dataset.

1. **Explain the Anscombe’s quartet in detail.**

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 before analysing it and the effect of outliers on statistical properties.

**Simple understanding:**

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

**3. What is Pearson’s R?**

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

* Pearson’s *r*
* Bivariate correlation
* Pearson product-moment correlation coefficient (PPMCC)
* The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

| **Pearson correlation coefficient (*r*) value** | **Strength** | **Direction** |
| --- | --- | --- |
| Greater than .5 | Strong | Positive |
| Between .3 and .5 | Moderate | Positive |
| Between 0 and .3 | Weak | Positive |
| 0 | None | None |
| Between 0 and –.3 | Weak | Negative |
| Between –.3 and –.5 | Moderate | Negative |
| Less than –.5 | Strong | Negative |

The Pearson correlation coefficient is also an inferential statistics, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

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

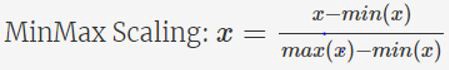
It is a step of data Pre-Processing which is applied to independent variables to normalise the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

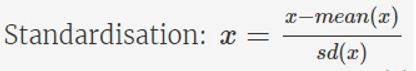
**Normalization/ Min-Max Scaling:**

It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler**helps to implement normalization in python.



**Standardization Scaling:**

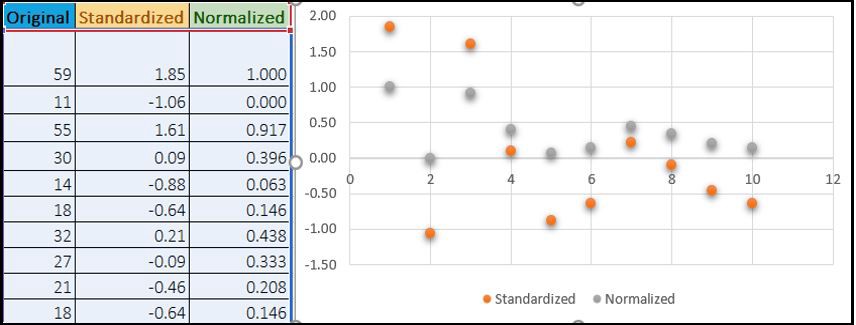
* Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).



* **sklearn.preprocessing.scale** helps to implement standardization in python.
* One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

**Example:**

Below shows example of Standardized and Normalized scaling on original values.



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

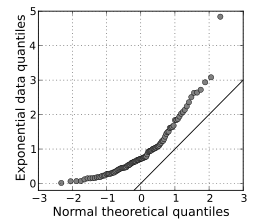
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



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.