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?

Answer: Following are the inferences we could derive after using a box plot on the categorical variables:

1. Season falls have the highest number of bookings for the period of two years with a median of over 5000 bookings. It could be significant in predicting the demand for shared bikes.
2. Year has increasing bookings from 2018 to 2019. It could be significant in predicting the demand for shared bikes.
3. Month variable shows a higher trend in the middle months of 5-9 with a median of over 4000 bookings per month. It could be significant in predicting the demand for shared bikes.
4. The holiday variable doesn't look helpful as most of the bookings are happening during the holiday season.
5. Weekday variable also doesn't show any clear trend, it could have little or no significance.
6. Working day has around 70-30 split for the period of 2 years.
7. Clear weather definitely has more number of bookings, it could be a good predictor.

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

Answer: drop_first=True drops the first column while dummy variable creation. It is essential to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
For eg:- If there are only 3 categories ‘YES’, ‘NO’, ‘MAYBE’ then we know that it is going to be 3rd option automatically if it is not the first two options. It helps us to reduce the redundancy.

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

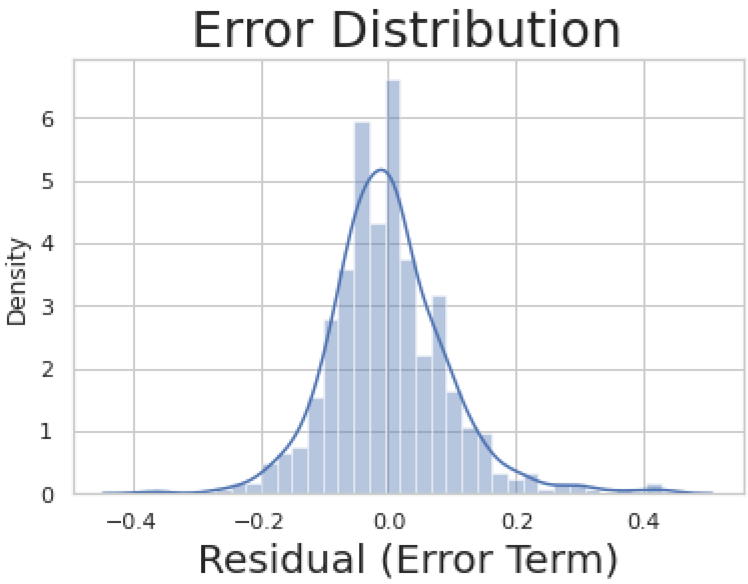
Answer: temp variable has the highest correlation with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the**

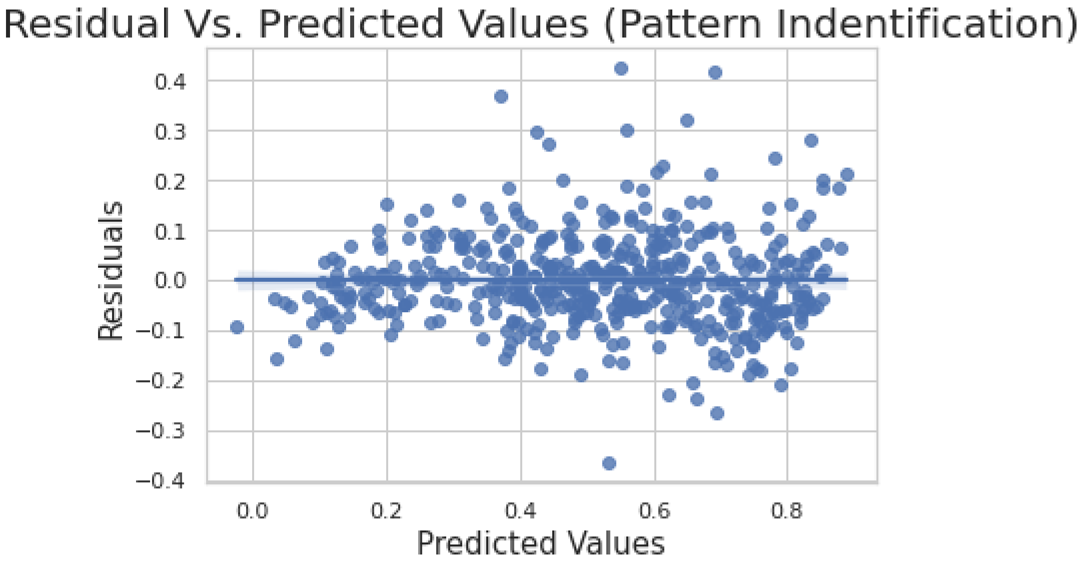
training set?

Answer: Following are the assumptions of Linear Regression which I validated:

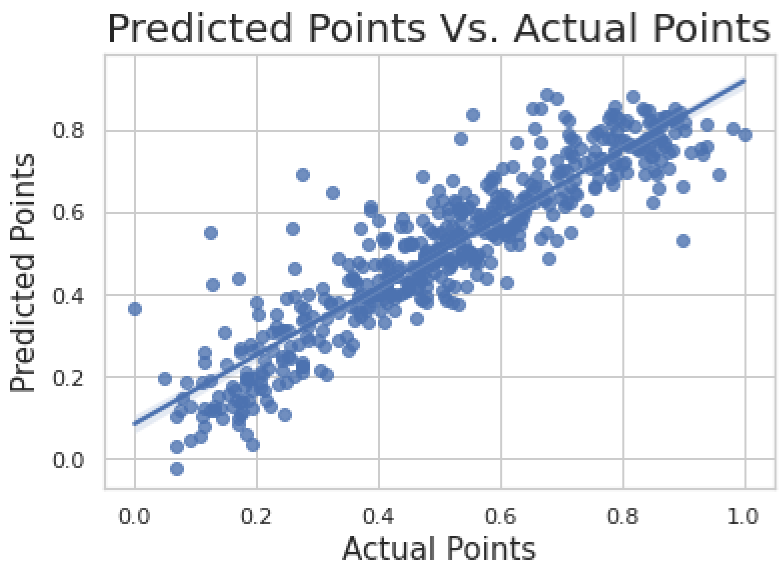
1. **Normally Distributed Error Terms**: I plotted the error terms and checked that if they are normally distributed around zero.



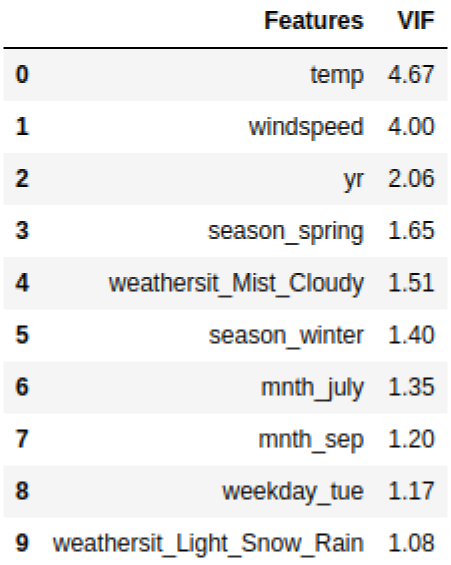
2. **Error Terms Being Independent:** I verified that the residual are not having any pattern and are independent.



3. **Homoscedasticity:** I verified that residuals are equally distributed across predicted value. The below graph shows equal variance and do not observe high/low concentration of data points in any region.



4. **Multicorrelation:** Indivual features correlation co-efficients which are impacting target variables. As VIF is low than 5 for all the variables, this assumption holds true.



	Features	VIF
0	temp	4.67
1	windspeed	4.00
2	yr	2.06
3	season_spring	1.65
4	weathersit_Mist_Cloudy	1.51
5	season_winter	1.40
6	mnth_july	1.35
7	mnth_sep	1.20
8	weekday_tue	1.17
9	weathersit_Light_Snow_Rain	1.08

5. **Based on the final model, which are the top 3 features contributing significantly towards**

explaining the demand of the shared bikes?

Answer: Following three features contribute significantly towards explaining the demand of the shared bikes:

- Temparature: 0.64
- Year: 0.59

- Spring Season: -0.55

General Subjective Questions

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

Answer: Linear Regression is the basic form of regression analysis. It assumes that there is a linear relationship between the dependent variable and the predictor(s). In regression, we try to calculate the best fit line, which describes the relationship between the predictors and predictive/dependent variables.

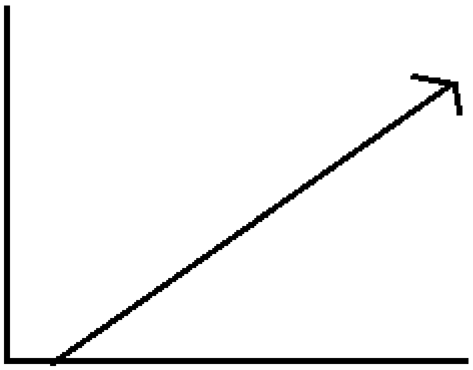
Mathematically the relationship can be represented with the help of following equation −

$$Y = mX + c$$

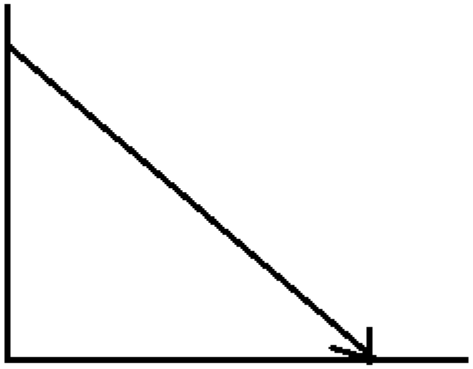
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept.
If X = 0, Y would be equal to c.
Furthermore, the linear relationship can be positive or negative in nature as explained below−

• Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph

• Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph



Positive Linear Relationship



Negative Linear Relationship

Types of Linear regression:

• Simple Linear Regression: It explains the relationship between a dependent variable and one independent variable using a straight line.

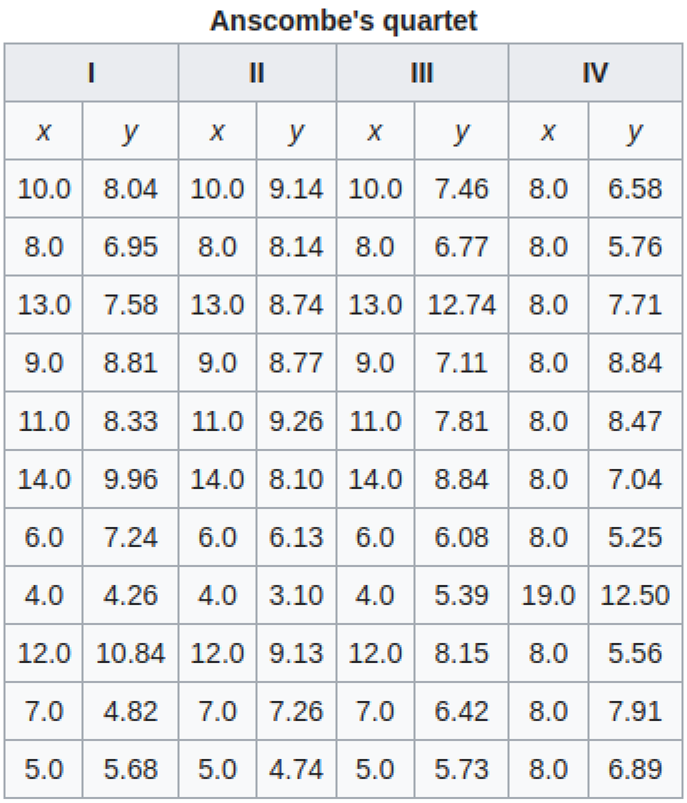
• Multiple Linear Regression: It explains the relationship between one dependent variable and several independent variables.

Assumptions: The following are some assumptions about dataset that is made by Linear Regression model

1. Multi-collinearity: Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
2. Auto-correlation: Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
3. Relationship between variables: Linear regression model assumes that the relationship between response and feature variables must be linear.
4. Normality of error terms: Error terms should be normally distributed
5. Homoscedasticity: There should be no visible pattern in residual values.

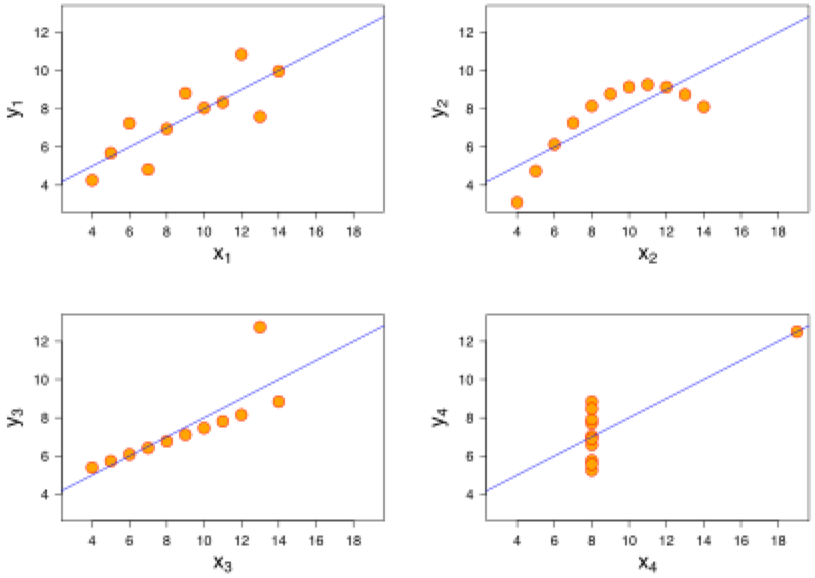
2. **Explain the Anscombe's quartet in detail.**

Answer: 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.



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



This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R?

Answer: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

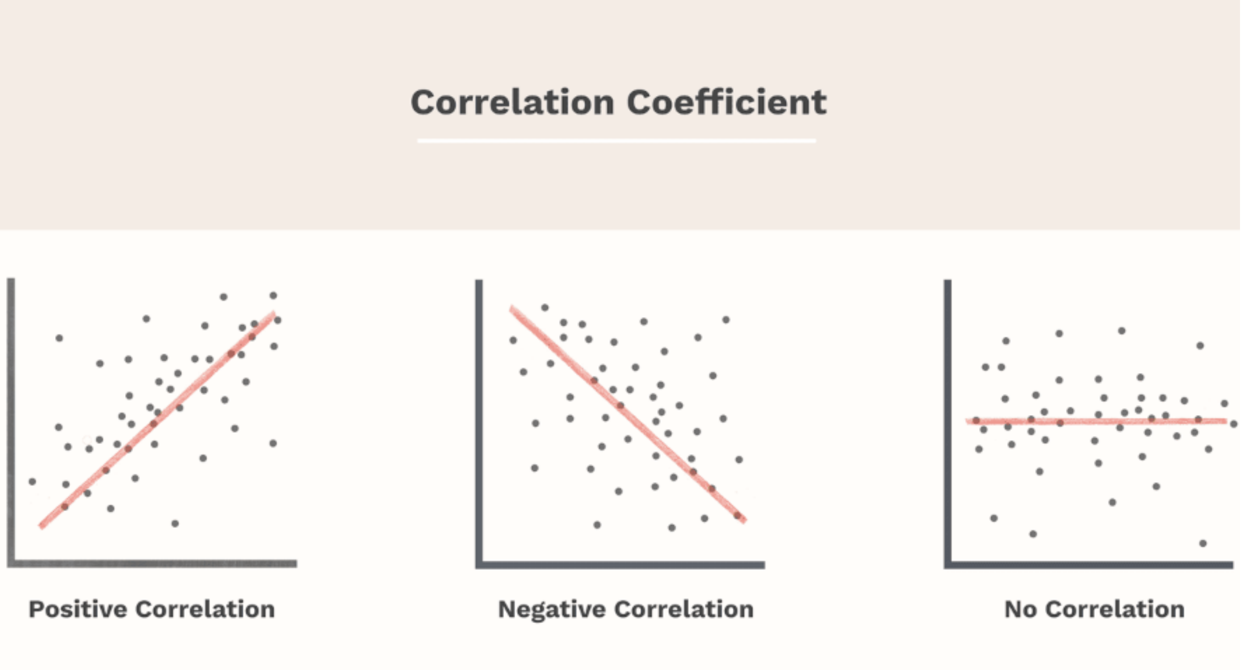
Properties:

1. Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..
2. Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
3. Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

1. Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
2. High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

3. Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
4. Low degree: When the value lies below + .29, then it is said to be a small correlation.
5. No correlation: When the value is zero.


Correlation Coefficient

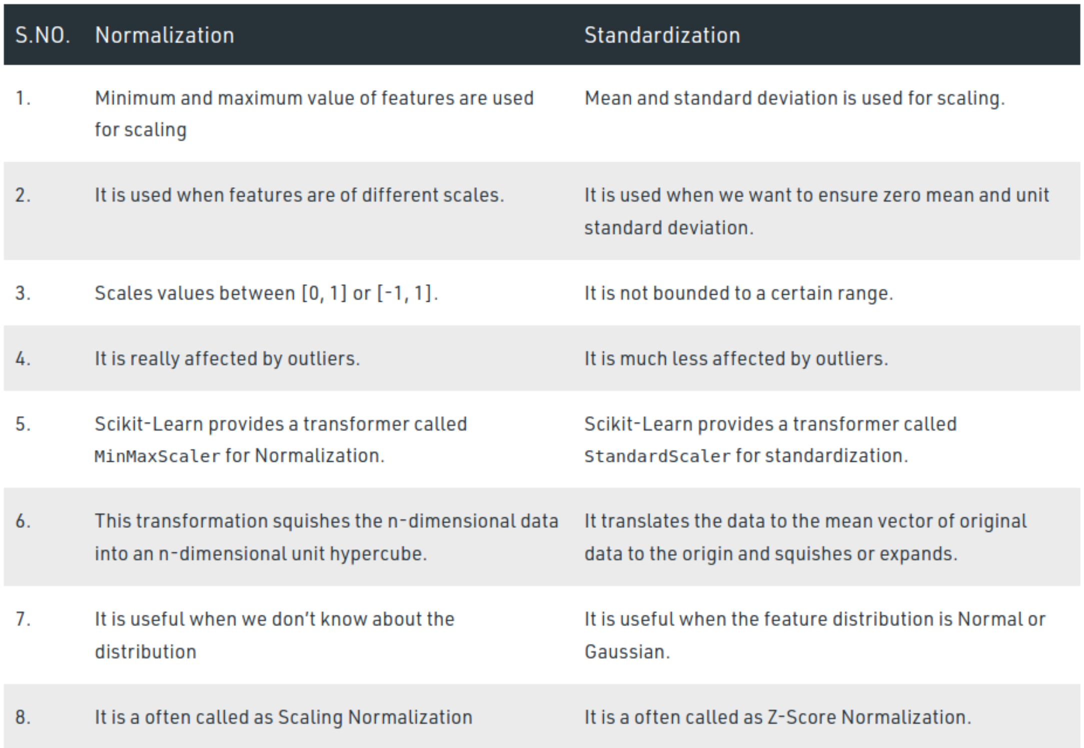
Positive Correlation

Negative Correlation

No Correlation

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

Answer: Scaling is very important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.
Also, scaling helps machine learning algorithms train and converge faster. In regression, it is often recommended to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means.



S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Answer: It happens due to perfect correlation, in that case VIF tends to infinity. As, R2 =1, 1/(1-R2) becomes infinity. This problem is solved by dropping one variable from the dataset solving multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Answer: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.