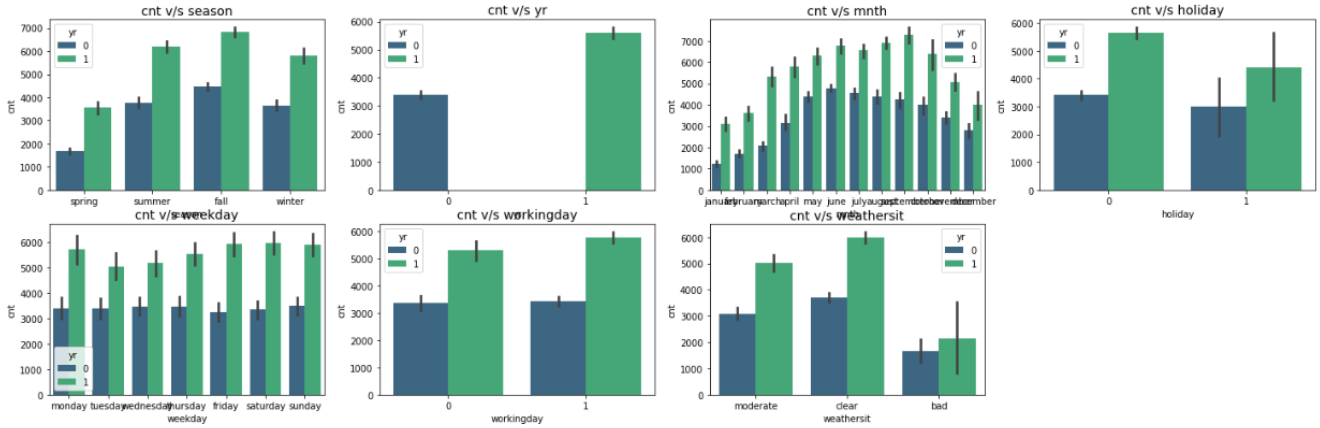
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

A:
display bar plot for each categorical variable with 'cnt' as target variable, display trend year wise



Effect of categorical variables on target variable:

- The Hiring rate is more for 2019 than 2018.
- 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 than holidays and weekends.
- Clear weather days attract a hire demand then other days.
- Temperature is having a positive correlation(0.63) with demand. And higher temp attracts more bike hiring.

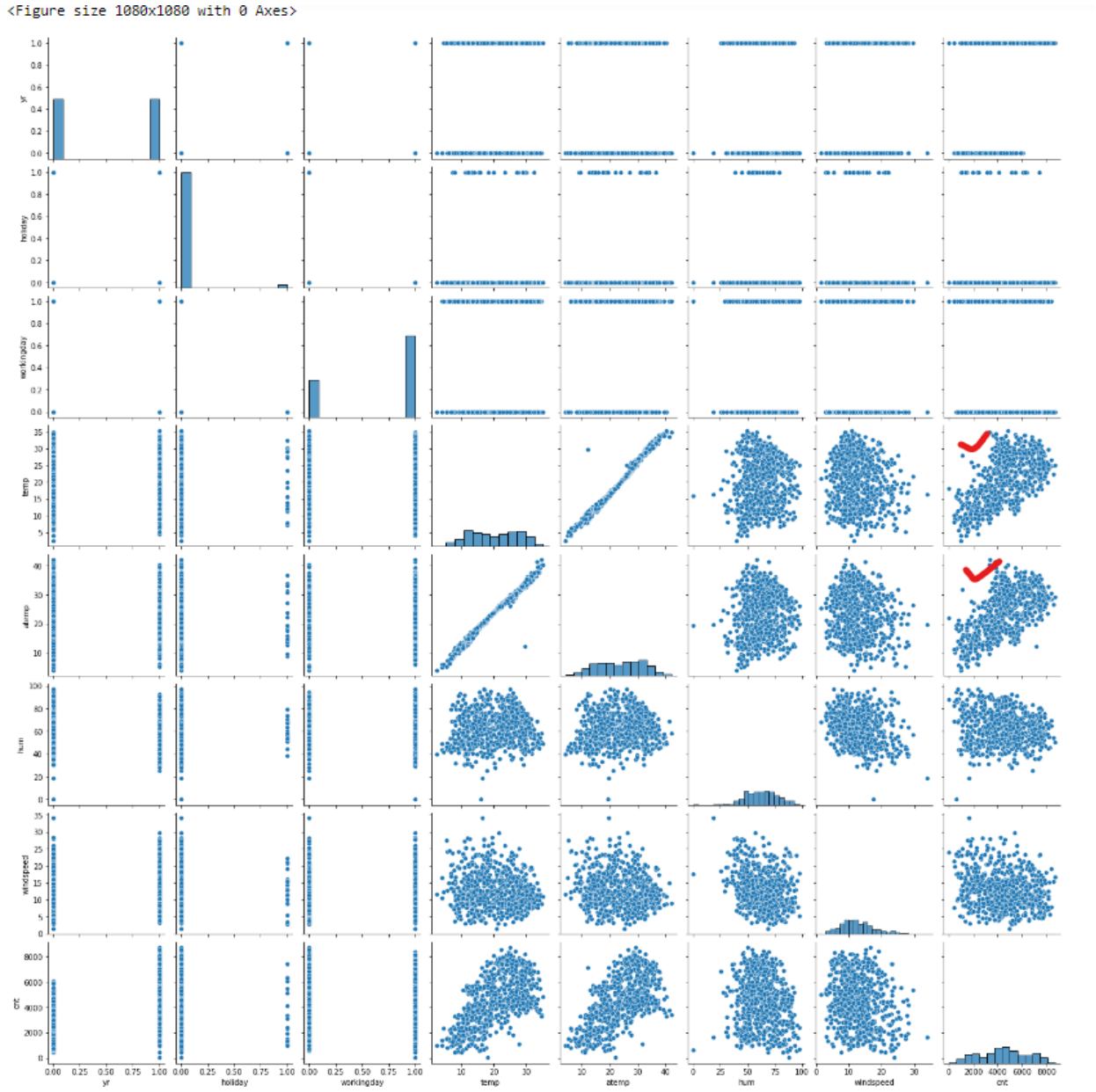
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

A:

- n-1 columns required to represent n categorical variables, in order to reduce correlation created among dummy variables, we need to remove any one dummy variable such that all dummy variables will be independent.
- Pandas has option to remove one dummy variable by passing drop_first=True argument during dummy variables creation.

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

A: Temperature is having a highest correlation (0.63) with target variable.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A:

Assumptions of Linear Regression are:

1. Linear relationship between X and Y

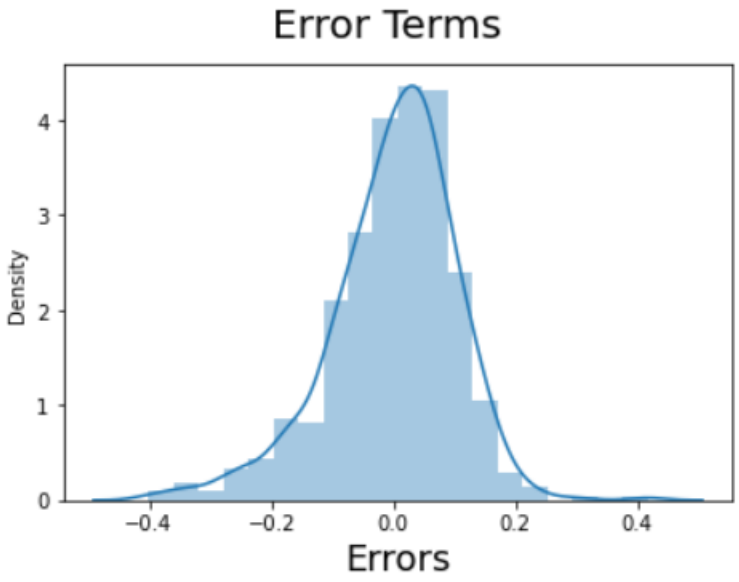
2. Error terms are normally distributed (not X, Y)

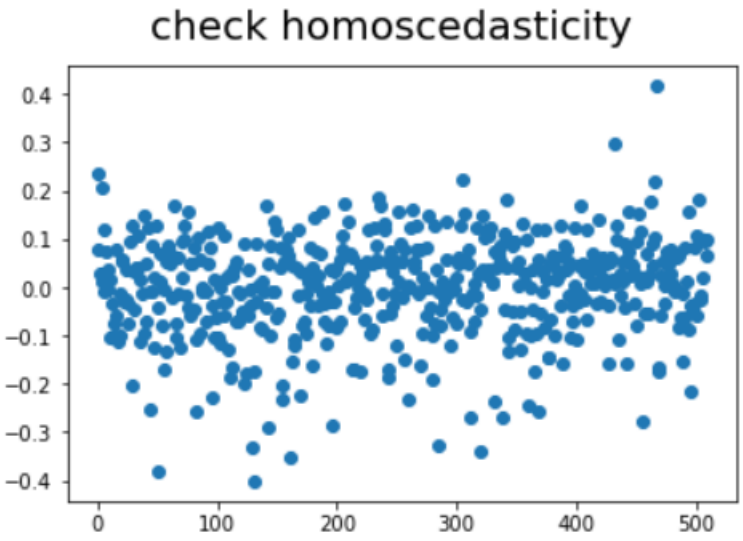
3. Error terms are independent of each other

4. Error terms have constant variance (homoscedasticity)

There is linear relation between X and Y.

From the distribution plot for error term, it clearly shows that mean is centered around zero and normally distributed.





From above scatter plot, it clearly shows that error is having constant variance and no visible patterns are there. It satisfies condition of homoscedasticity.

All the assumptions of Linear Regression are satisfied.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A: Based on trained model and our analysis, top three features contributing to shared bikes demand are yr, weather situation, September month.

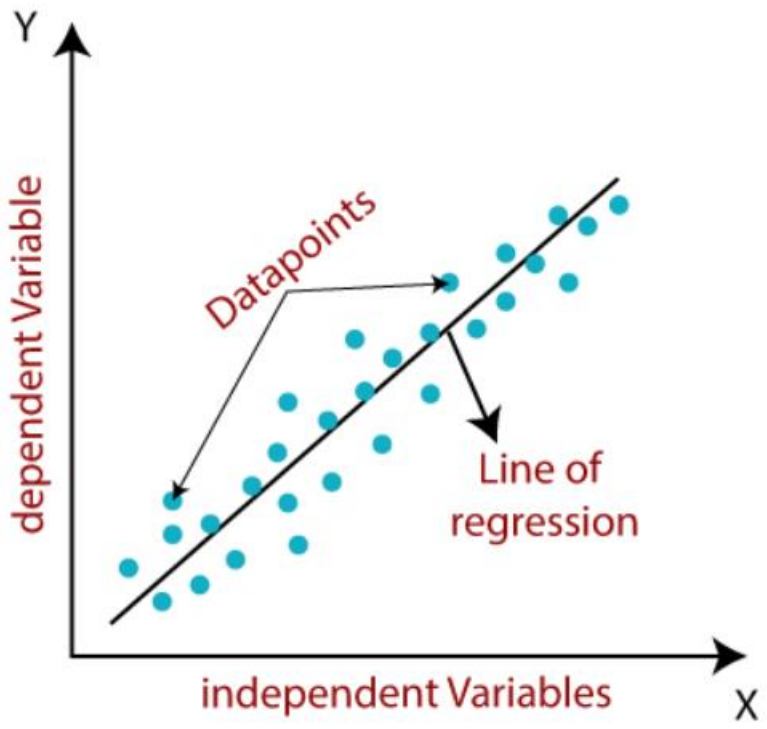
General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

A:
Linear Regression is a supervised machine learning algorithm. It performs a regression task that means predicting numerical value based on independent variables. Linear regression algorithm has linear relationship between a dependent variable (y) and one or more independent (x) variables. Since it is linear relationship, the dependent variable is changing according to the value of the independent variables.

The linear regression model fits a straight line for representing the relationship between the dependent and independent variables.

Please check below image for better understanding:



Equation of Simple Linear Regression:

$$y = b_o + b_1 x$$

bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

Equation of Multiple Linear Regression:

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

Assumptions of Linear Regression are:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

2. Explain the Anscombe's quartet in detail. (3 marks)

A:

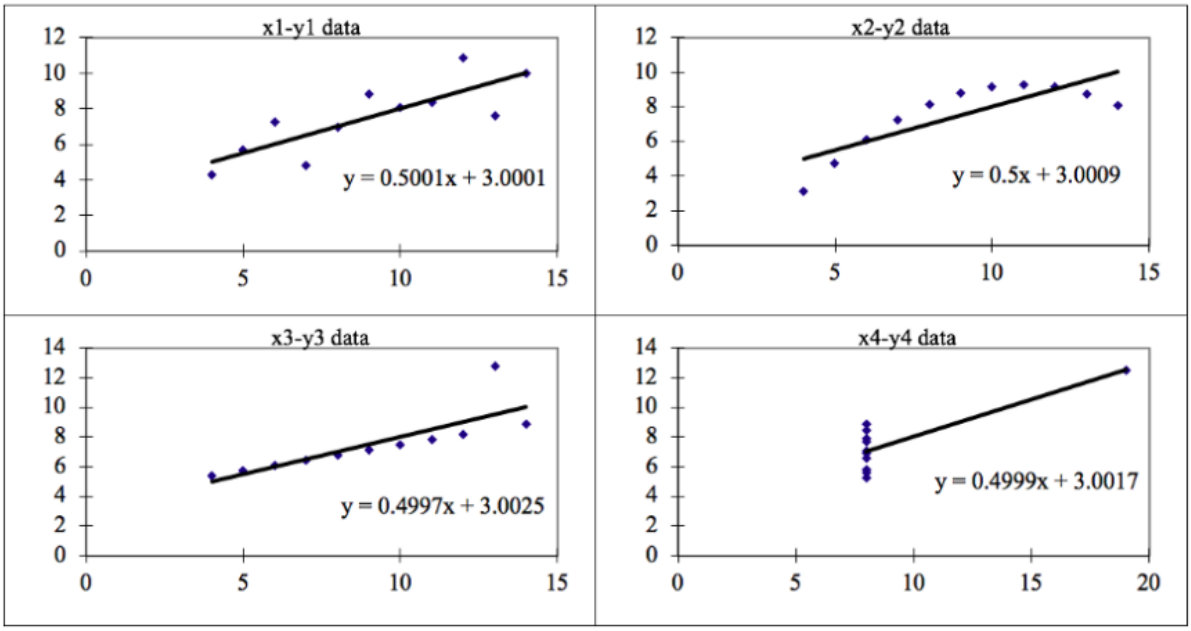
Anscombe's quartet is constructed in 1973 by statistician Francis Anscombe to illustrate the importance of data visualization before constructing the machine learning model. There are four datasets which have nearly same statistical information like mean and variance. when we try to fit linear regression model, all the four datasets having almost same linear regression model parameters but failed to generalize the distribution of the data, so this dataset can easily fool the linear regression model.

So it is very important to analyze the data by visualization before building any model on any data. Data visualization helps us to understand the distribution of the data, outliers in the data, linear separability of the data.

Anscombe's Data and Statistics shown below:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When we check below plots, all the data set showing same regression line but having different distribution data.



The four datasets can be described as:

Dataset 1: Linear regression fits pretty well.
Dataset 2: non linear data, linear regression doesn't fit well.
Dataset 3: outliers present in the data, linear regression doesn't fit well.
model
Dataset 4: outliers present in the data, linear regression doesn't fit well.
model

The summary is, Anscombe's quartet helps us to understand the importance of data visualization and how someone can fool regression model with the data. So before building any model, first try to visualize the data and understand the different properties and then build the model.

3. What is Pearson's R? (3 marks)

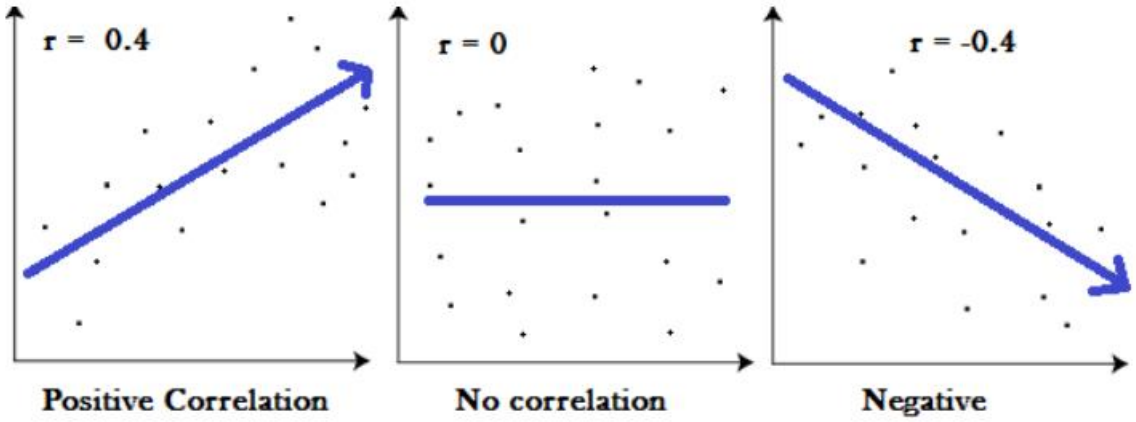
A:

Pearson's R is also called Pearson's correlation is a correlation coefficient used in linear regression. Basically, correlation measures the relation between two variables.

Pearson's correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Correlation coefficient values range between -1 and 1, where:



- A correlation coefficient of 1 means positive correlation, increase in one variable result same proportion of increase in another variable.
- A correlation coefficient of -1 means negative correlation, increase in one variable result same proportion of decrease in another variable.
- Zero correlation means these two variables are not related to each other.
-

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

A:

Machine learning algorithms are biased towards large numerical data values if the data is not scaled. So, we need to scale all the data values into common range and the Machine learning model can fit on that data. Also, it helps to converge Machine learning algorithms faster.

There are variety of scaling techniques in machine learning, primarily used methods are normalized scaling and standardized scaling.

Normalized Scaling:

This method is used to transform features into similar scale. The new value is calculated by using below equation.

X_new = (X - X_min)/(X_max - X_min)

Output data range is [0, 1] or [-1, 1].
This method is useful when there are no outliers and user don't have information about distribution of the data. This transformation squishes the n -dimensional data into n dimensional hyper cube.

standardized scaling:
This method is used to transform the data into a standard normal distribution, all the features are subtracted by mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std
This method is useful when user knows about data distribution. It will also affect dummy variables (Ex: 0 &1). Standardized scaling does not get affected by outliers.

Differences between normalized scaling and standardized scaling:

Normalization	**Standardization**
Scaling uses Minimum and maximum value of features	Standardization uses Mean and standard deviation of features
This method is used when the features are of different scales.	This method is used to ensure zero mean and unit standard deviation.
Output values range between [0, 1] or [-1, 1].	There is no output range criteria.
This method is affected by outliers.	This method is less affected by outliers.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

A:

- VIF gives a basic idea about how much feature variables are correlated to each other, that means how well a predictor variable is correlated with all other variables, excluding target variable.
- Formula for VIF is 1/(1-R2).
- If there is perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity, then VIF = infinity, the corresponding variable may be expressed exactly by a linear combination of other variables.
- we need to drop the variable from the dataset which is causing this perfect multicollinearity.

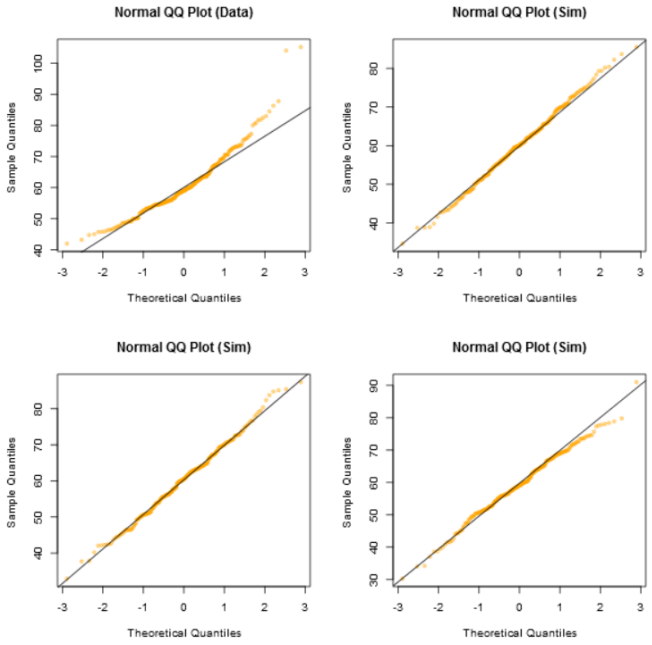
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

A:

Quantile-Quantile (Q-Q) plot is used to compare two data sets from population having same distribution or not. For example, it can compare Uniform, Normal and Exponential distributions of the two datasets. It compares the properties like location, scale, and skewness.

A 45 degree reference line on Q Q plot:



The median is a quantile where 50% of the data fall below that point and 50% lie above it, A 45 Degrees line is plotted on Q Q plot, if two datasets are having same distribution then the points exactly or approximately lie on the 45 degrees line. If two distributions are different then points may lie above or below the 45 degrees line.