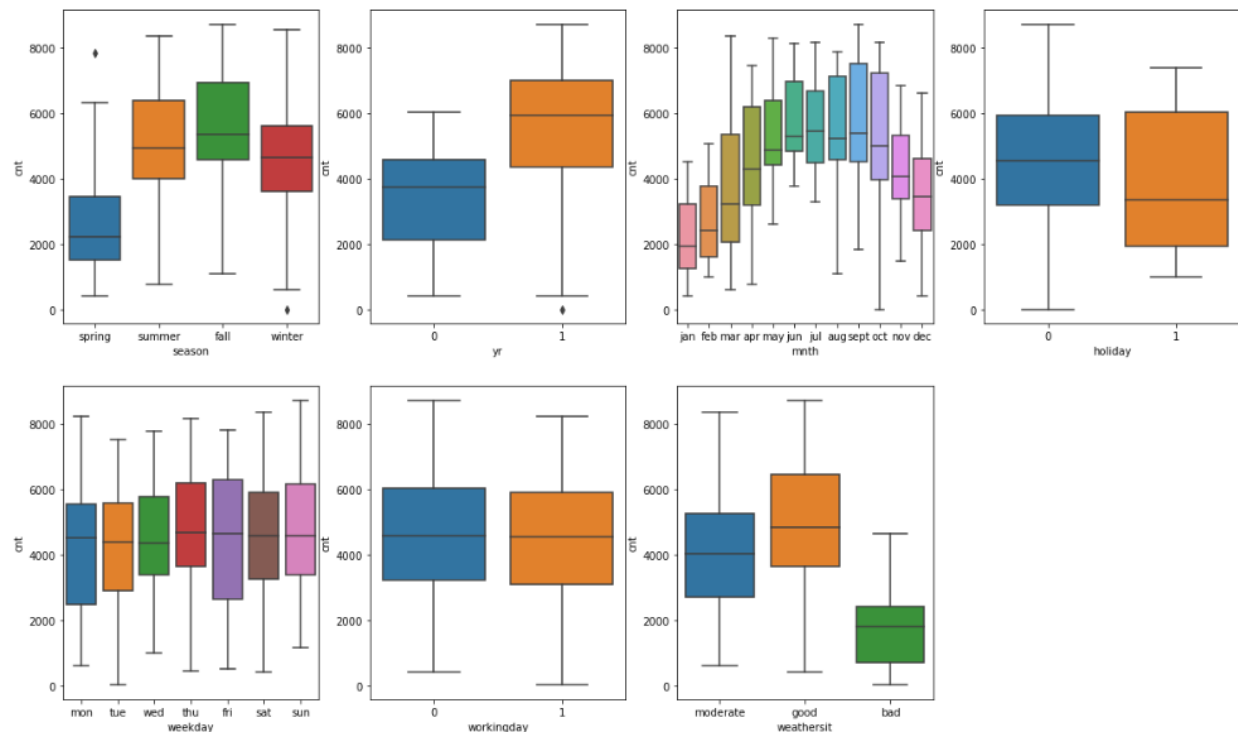


Assignment-based Subjective Questions

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

Answer:

- I. The Demand of Bike in spring is the lowest and highest in the fall season.
- II. The Demand of Bikes was significantly more in year 2019 compared to year 2018.
- III. The Demand of Bikes starts to rise from the month of Feb onwards and reaches to the highest in Sept and then drops in Oct, Nov, Dec.
- IV. When weather is good demand is more for the bikes.



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

Answer:

While we create the dummy variables for categorical variables, it will create p number of dummy variables, so we will have to deal with more variables and correlations,

If we use `drop_first = True`, it reduces the one variable (only create p-1 dummy variables) column so we have deal with less columns, It does not affect the final interpretation.

Example : a categorical variable that can take on three different values ("Single", "Married", or "Divorced"), we need to create $p-1 = 3-1 = 2$ dummy variables.

To create this dummy variable, we can let "Single" be our baseline value since it occurs most often. Thus, here's how we would convert marital status into dummy variables:

Income	Age	Marital Status
\$45,000	23	Single
\$48,000	25	Single
\$54,000	24	Single
\$57,000	29	Single
\$65,000	38	Married
\$69,000	36	Single
\$78,000	40	Married
\$83,000	59	Divorced
\$98,000	56	Divorced
\$104,000	64	Married
\$107,000	53	Married



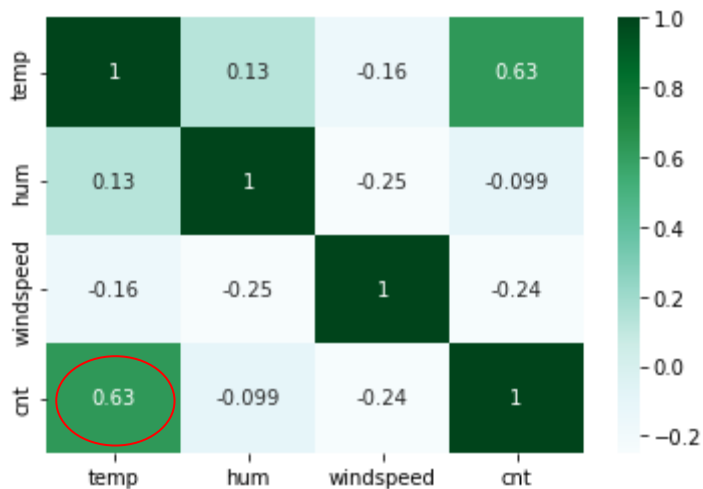
Income	Age	Married	Divorced
\$45,000	23	0	0
\$48,000	25	0	0
\$54,000	24	0	0
\$57,000	29	0	0
\$65,000	38	1	0
\$69,000	36	0	0
\$78,000	40	1	0
\$83,000	59	0	1
\$98,000	56	0	1
\$104,000	64	1	0
\$107,000	53	1	0

0 0 – Single, 0 1 – Divorced, 1 0- Married

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

Answer:

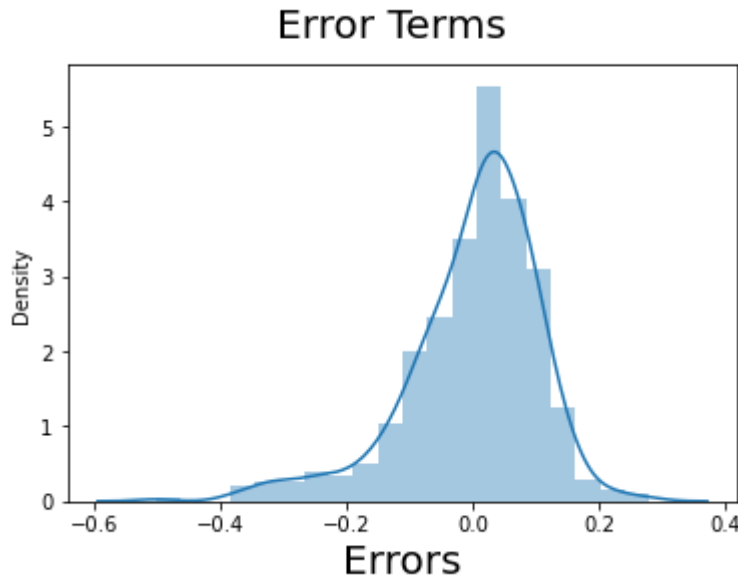
By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'



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

Answer:

- i) Residual Analysis of Training data (draw Error terms distribution which is normally distributed)
- ii) From the below histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid



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

Answer:

- yr 2114.9333
- holiday -583.8529
- temp 2994.9679
- wind speed -987.2657
- season spring -1345.8000
- mnth_oct 539.6068
- mnth_sept 578.3473
- weathersit_moderate -537.9063

General Subjective Questions

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

Answer:

- Linear Regression is a supervised learning technique which supports to find out the correlation between variables. (variables can be categorical or continuous values). As the name suggested linear, which means the two variable which are on x-axis and y-axis should be linear correlated.
- Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where b is slope, a is the intercept (constant), y = dependent variable, x = Independent variable

- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.
- Hypothesis function for Linear Regression

$$y = \theta_1 + \theta_2 * X$$

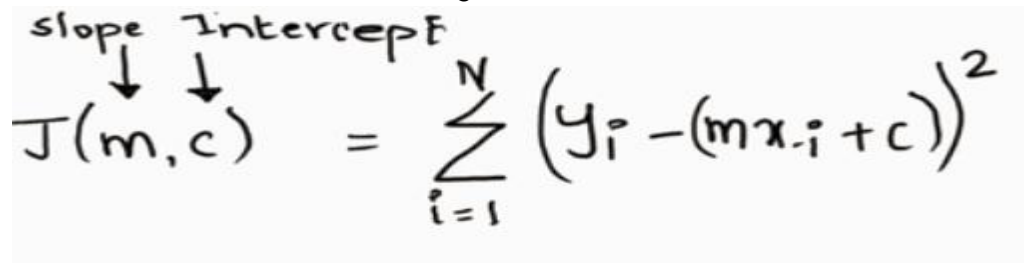
- **How to update θ_1 and θ_2 values to get the best fit line**

Cost Function (J):

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y)

$$J(\theta_0, \theta_1) = \sum_{i=1}^N (y_i - y_i(p))^2$$

- Our goal is to minimize the cost function, which will result in lower error values. If we minimize the cost function, we will get the best fit line to our data.



The image shows a handwritten formula for the cost function: $J(m, c) = \sum_{i=1}^N (y_i - (mx_i + c))^2$. Above the formula, the word "slope" has an arrow pointing down to the variable m , and the word "Intercept" has an arrow pointing down to the variable c .

(m,c), where m is the coefficient and c is the intercept

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

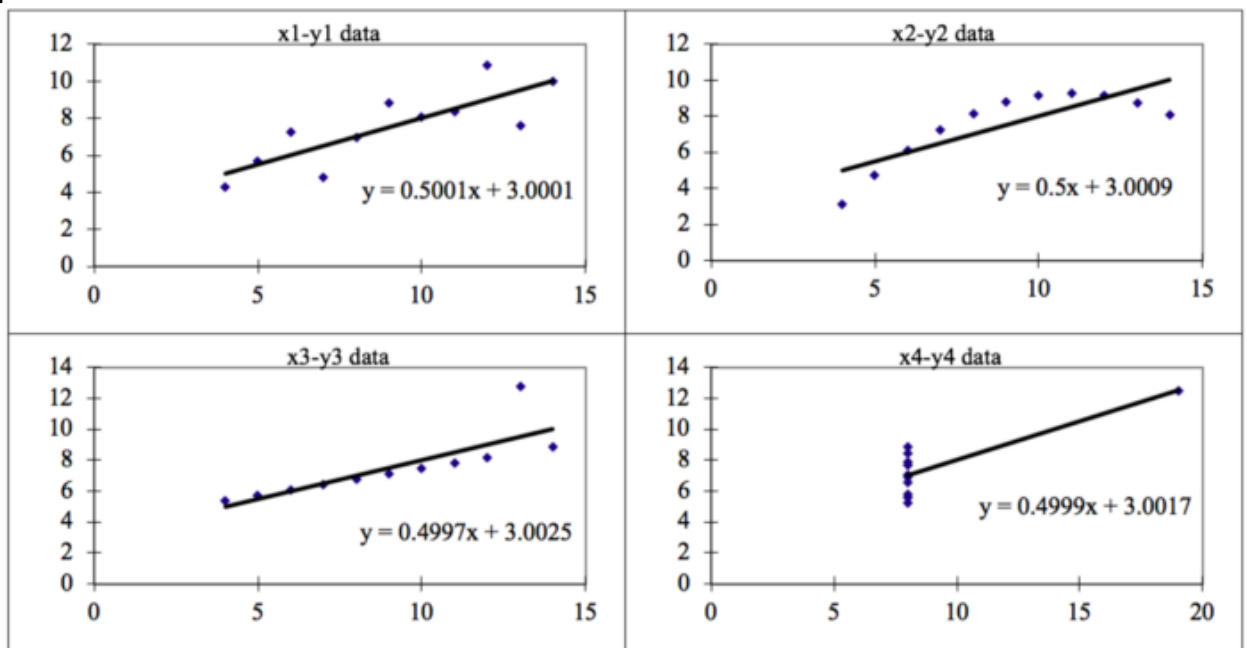
Answer:

- **Anscombe's Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.
- There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets. So it is important to plot and visualize the data .
- These four plots can be defined as follows:

All four data sets have almost same statistics data as 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 these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



- The Above data sets can be interpreted as below:
 1. **1st Data set:** In this the linear regression line fits nicely.
 2. **2nd Data set:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
 3. **3rd Data Set :** this has the **outliers** involved in the dataset which **cannot be handled** by linear regression model
 4. **4th Data Set :** this shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
- Conclusion:
It is important to draw the plot and visualize the data before applying the machine learning algorithms.

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

Answer:

Pearson's Correlation Coefficient (R):

- In the terms of Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bi-variate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
- There are certain requirements for Pearson's Correlation Coefficient:
 - Scale of measurement should be interval or ratio
 - Variables should be approximately normally distributed
 - The association should be linear
 - There should be no outliers in the data

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

Here, above terms are meant as following:

N = Number of pairs of scores, $\sum xy$ = the sum of the products of paired scores

$\sum x$ = the sum of x scores, $\sum y$ = the sum of y scores, $\sum x^2$ = the sum of squared x scores

$\sum y^2$ = the sum of squared y scores

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

Answer:

- Scaling is a technique of bringing down the values of all the independent features of our dataset on the same scale. Feature selection helps to do calculations in algorithms very quickly. It is the Important stage of data preprocessing
- If we do not perform the scaling the machine learning model will give the higher weightage to higher values and lower weightage to the lower values.
- As these weightage cause model to take more time to train the model.

Reason to perform the Scaling:

- Some machine learning algorithms are sensitive, they work on distance formulas and use **gradient descent** as an optimizer. Having values on the same scales helps gradient descent to reach **global minima** smoothly.

Normalized scaling vs. standardized scaling

Normalized scaling	Standardized scaling
Normalization is a scaling technique in which the values are rescaled between the ranges 0 to 1 .	Standardization is another scaling technique in which the mean will be equal to zero and the standard deviation equal to one .
To perform the Normalized scaling , we need to import the MinMaxScalar from Sci-Kit learn Library.	To perform the Standardized scaling, we need to import the StandardScalar from Sci-Kit learn Library.
$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$	$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$

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

(3 marks)

Answer:

➤ If there is perfect correlation, then $VIF = \text{infinity}$.

This shows a perfect correlation between two independent variables.

In the case of perfect correlation,

We get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multi-collinearity.

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)

$$Vif = \frac{1}{1 - R_j^2}$$

if $R_j = 1$

$$vif = \frac{1}{1-1} = \infty$$

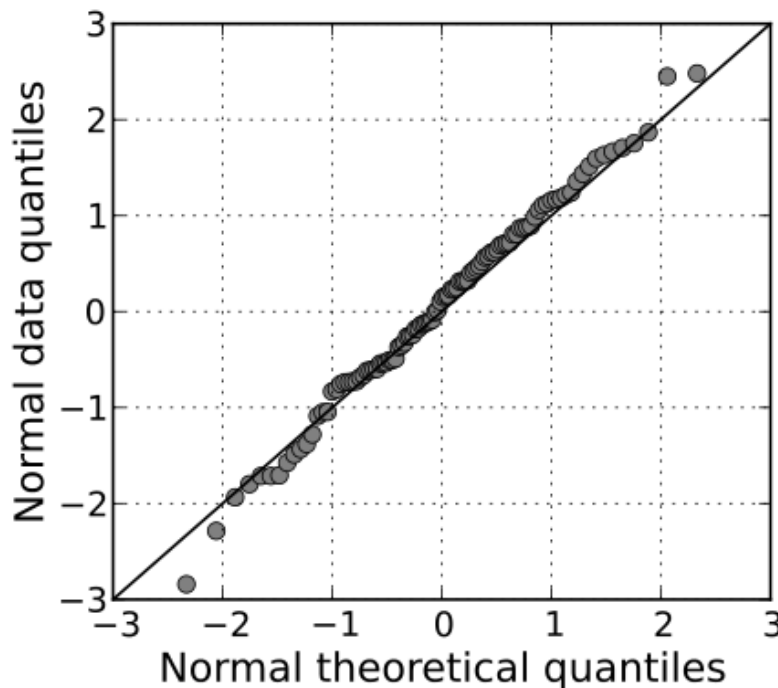
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Answer:

Q-Q (quantile – quantile) Plots:

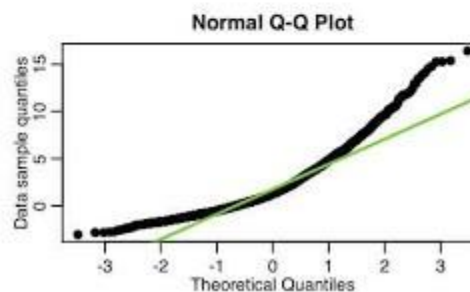
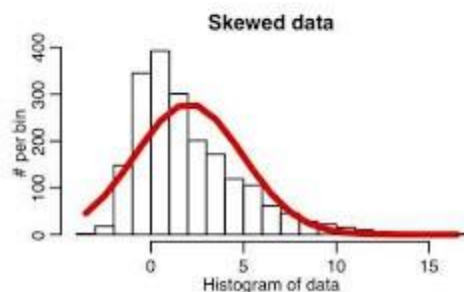
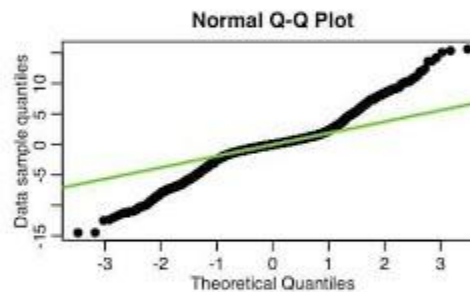
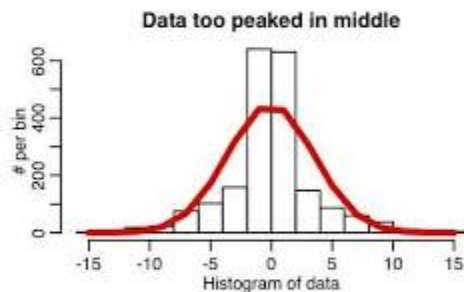
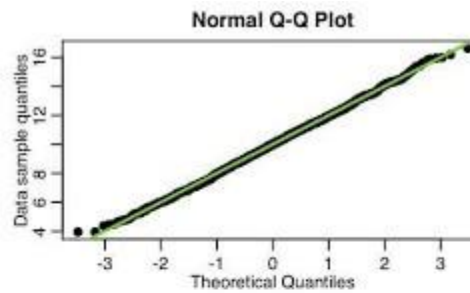
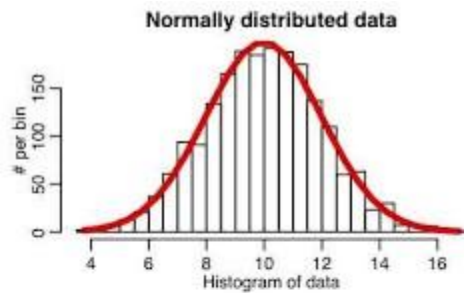
- It is the plot between quantiles of two probability distributions, it plays a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If two distributions which we are trying to compare,
- If there are two exactly equal distributions which we are trying to compare then the points on a Q-Q plot will perfectly lie on a straight line, $y = x$.
- Q-Q plot answers the most fundamental question, it tells whether the curve is normally distributed or not.



The Reason we need Normal Distribution:

- Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. In general, we are talking about **Normal distributions** only because we have a very beautiful concept of **68–95–99.7 rule** which perfectly fits into the normal distribution so we know how much of the data lies in the range of first standard deviation, second standard deviation and third standard deviation from the mean. So knowing if a distribution is Normal opens up new doors for us to experiment with the data easily.
- We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1) on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph?

- Now we have to focus on the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but indeed are scattered significantly from the positions then we cannot conclude a relationship between the x and y axes which clearly signifies that our ordered values which we wanted to calculate are not Normally distributed.
- If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normal distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.



Note: references , [google.com](https://www.google.com), [geeksforgeeks](https://www.geeksforgeeks.com), towardsdatascience.com etc.