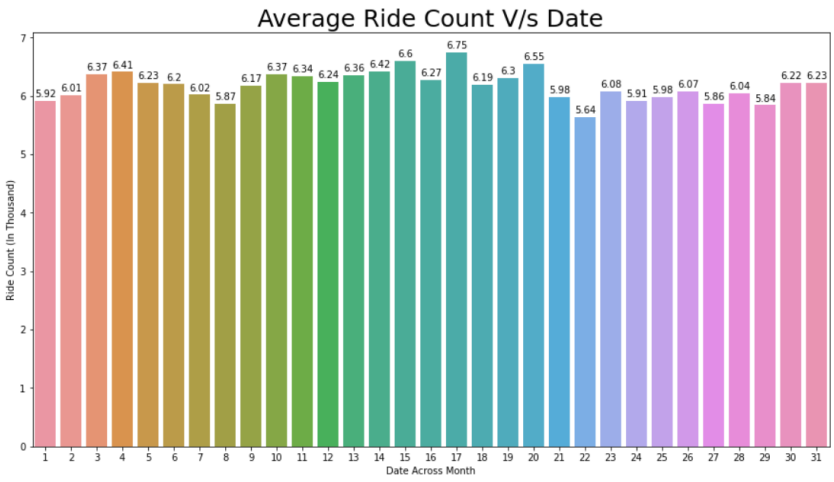
Assignment-based Subjective Questions

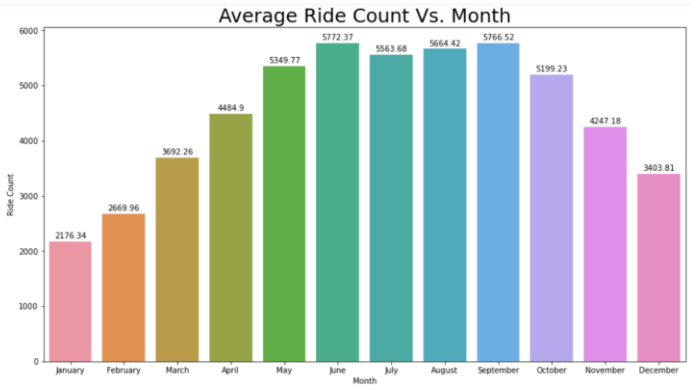
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

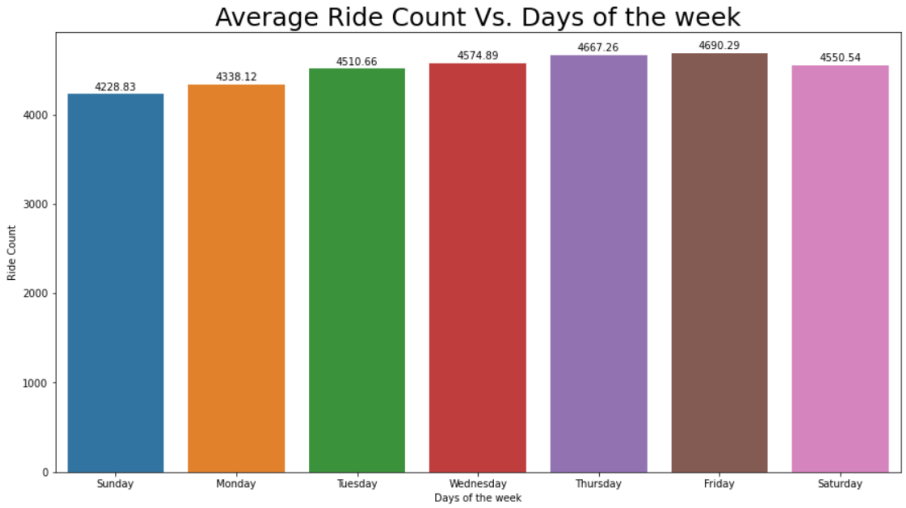
- Analysis of average ride count across all the days didn't specify any significant insight while visualizing the data.



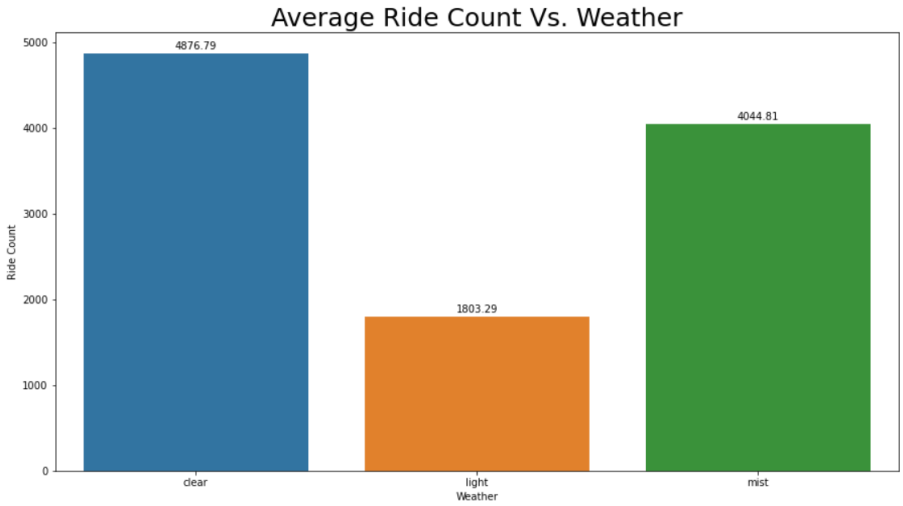
- Analysis of average ride count against month indicates that number of Ride count drastically increased between May to October



- Analysis of average ride count against days of week indicates that number of Ride Count is highest between Wednesday to Friday during middle of working weeks which shows most of the people rented bikes for work as the number is quite low on Sundays.



- Analysis of average ride count against weather indicates high number of bikes were reneted when the weather was either clear, few clouds, partly cloudy



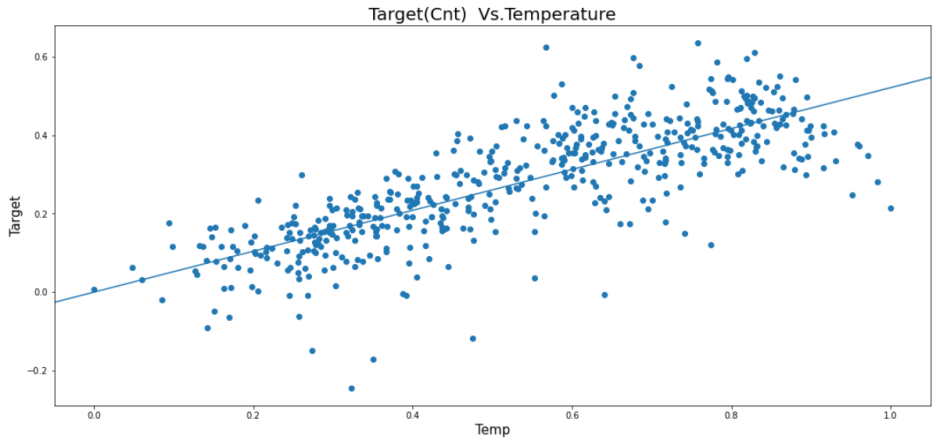
2. Why is it important to use drop_first=True during dummy variable creation?

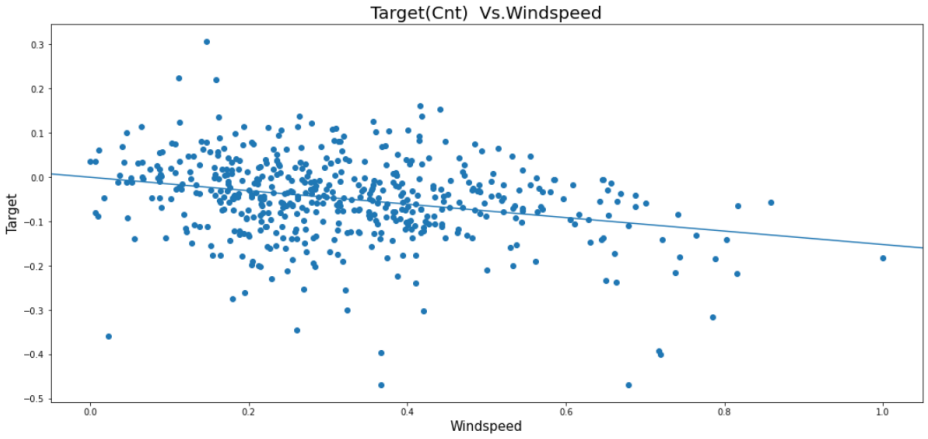
   - **Answer:** drop_first=True is important to use as it helps in reducing extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

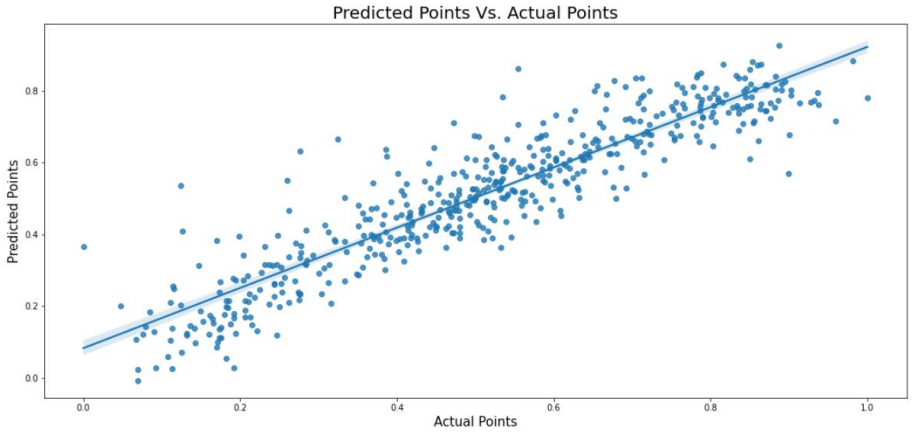
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - **Answer:** temp/atemp column has highest correlation with the target variable as per pairplot. AS the pair-plot for these two variables shows linear kind of relationship with **cnt** target variable.
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - **Answer:**
     i. **Checking for Linear Relationship**: There exists a linear relationship between independent variable, x, and the dependent variable, y

Equation of best fitted linear model is given by:

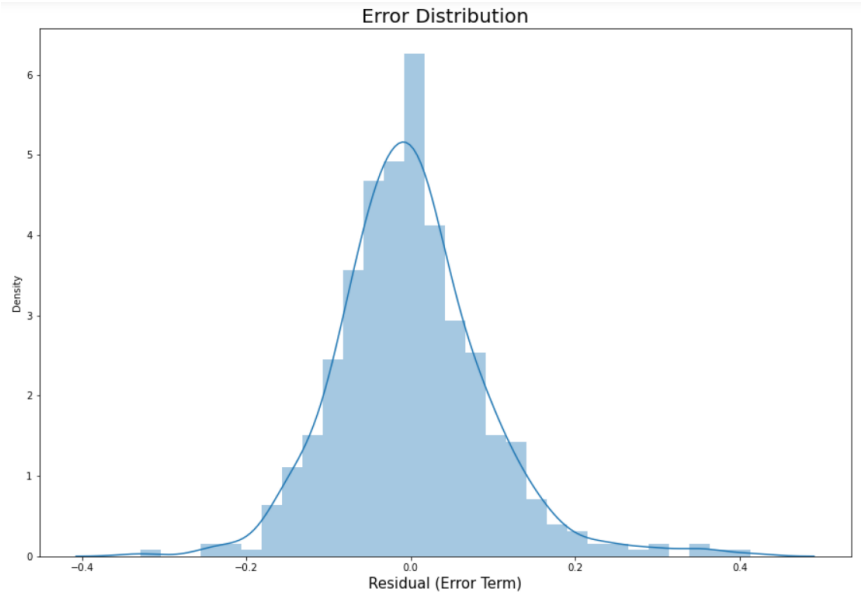
CNT(Target Varibale) = 0.080941 + (yr × 0.232844) + (workingday * 0.054605) + (temp * 0.520838) - (windspeed * 0.151781) + (season_2 * 0.101039) + (season_4 * 0.137942) + (mnth_8 * 0.051225) + (mnth_9 * 0.111729) + (weekday_6 * 0.065324) - (weathersit_light * 0.285973) - (weathersit_mist * 0.082579)



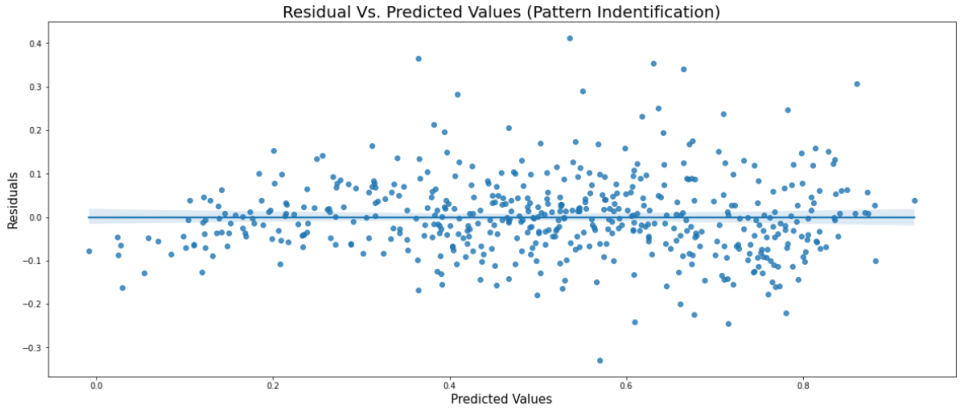




ii. The residuals of model are normally distributed around mean, which indicates that model has handled the assumption of Error Normal Distribution properly

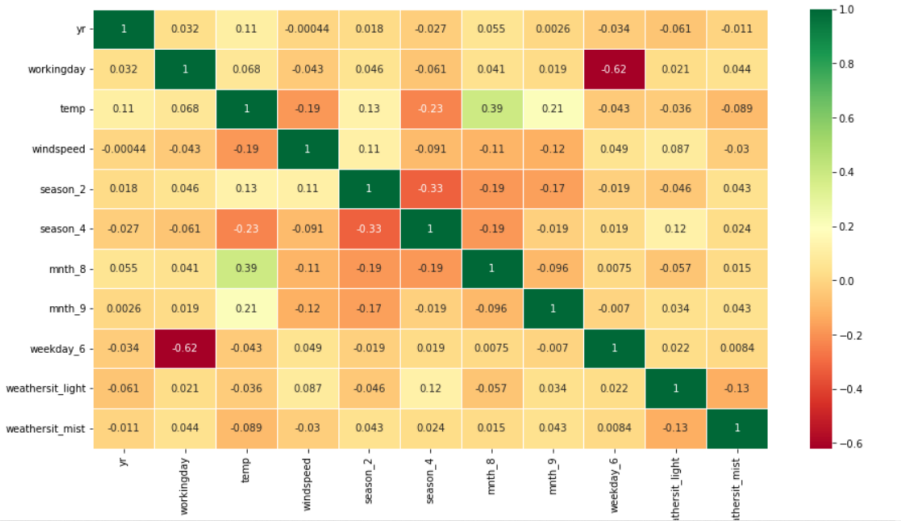


iii. **Homoscedasticity:** The residuals have constant variance at every level of x as indicated by below graph.
Above graph indicates, there is no relation between Residual & Predicted Value. This is what is expected from the model to not have specific pattern. This describes homoscedasticity.



iv. **Checking for independence** i.e., residuals are independent. In particular, there is no correlation between consecutive residuals in time series data. The Durbin-Watson value for Final Model lr 5 is 2.05 which indicates no autocorrelation.

v. **Validating absence of multicollinearity:**



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

   Answer: Top 3 features contributing significantly towards explaining the demand of shared bikes are **temp, yr and mnth_8**

   - A unit increase in temp(Temperature) variable increase the bike hire numbers by 0.520838 units.

   - A unit increase in mnth_8 variable increase the bike hire numbers by 0.051225 units.

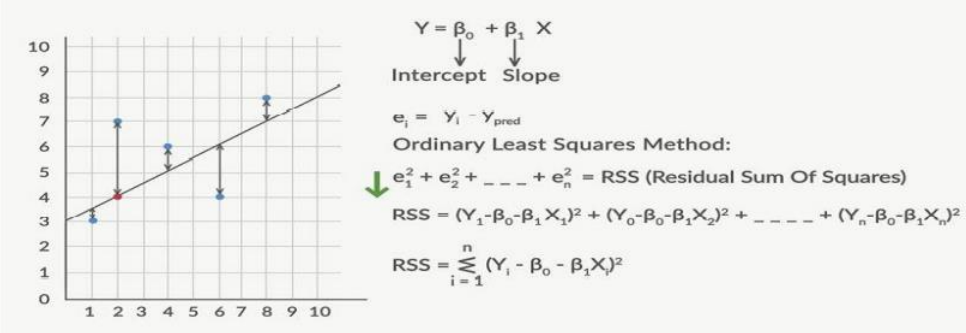
   - A unit increase in yr(Year) variable increase the bike hire numbers by 0.232844 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer: The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

   The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE).

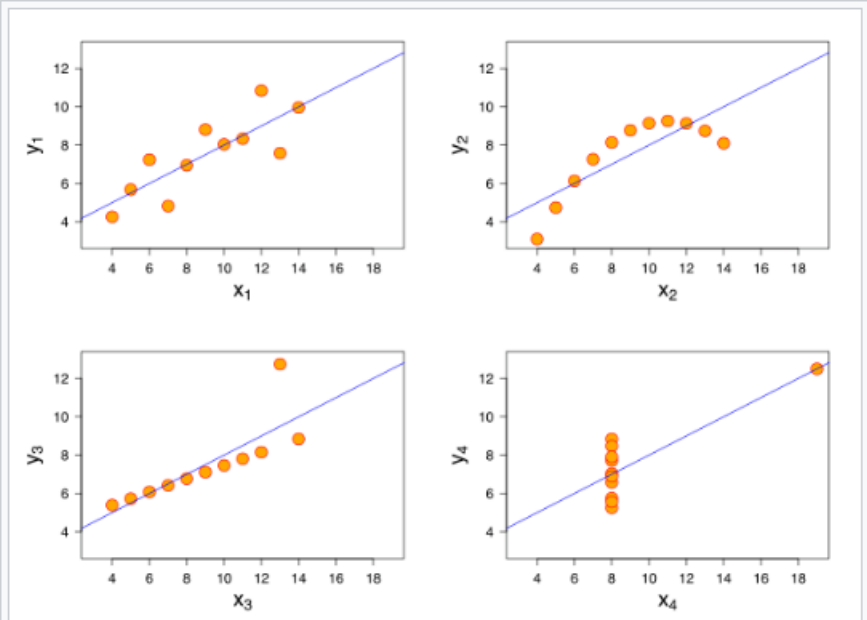
Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$ Fig 5 - R-squared RSS(ResidualSumofSquares): Instatistics,itisdefinedas thetotal sumoferroracross the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows: TSS(Total sumofsquares): Itis thesumoferrorsofthedatapoints from meanof response variable. Mathematically, TSS is: Importance of RSS/TSS: Think aboutitfor a second. If you know nothing aboutlinear regression and still have to draw a line to represent those points, the least you can do is have a line pass through the mean of all the points as shown below.

Assumptions of simple linear regression were:

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) With these assumptions we can go ahead and make inferences about the model which, otherwise, we wouldn't have been able to. Also note that, there is NO assumption on the distribution of X and Y, just that the error terms have to have a normal distribution.

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. 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 analysing 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



- 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 analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3. What is Pearson's R?
   Answer: The Pearson correlation coefficient, also called Pearson's R, is a statistical calculation of the strength of two variables' relationships. In other words, it's a measurement of how dependent two variables are on one another.The value of Person r can only take values ranging from +1 to -1 (both values inclusive). If the value of r is zero, there is no correlation between the variables.
   If the value of r is greater than zero, there is a positive or direct correlation between the variables. Thus, a decrease in first variable will result in a decrease in the second variable.
   If the value of r is less than zero, there is a negative or inverse correlation. Thus, a decrease in the first variable will result in an increase in the second variable.
   When plotted on a diagram, a positive correlation will see a line which slopes downwards from left to right and a negative correlation will see a line which slopes downwards from right to left.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer: When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.
   So we need to scale features because of two reasons:
   1. Ease of interpretation
   2. Faster convergence for gradient descent methods

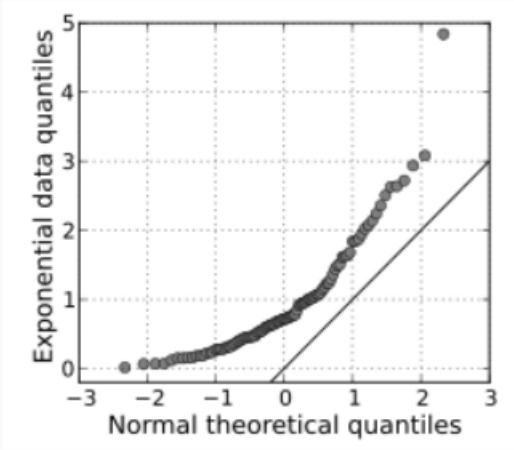
   We can scale the features using two very popular method:
   1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.
   2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. 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-square, etc.

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



   q-q plot in linear regression

   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.