Assignment-based Subjective Questions

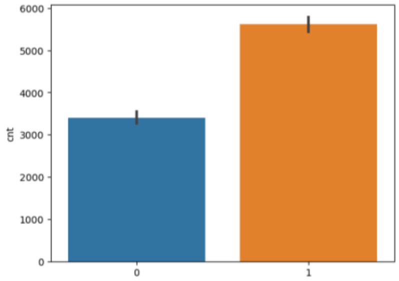
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:

Multiple categorical variables have affected dependent variable “cnt”, i.e., count of bike rentals.

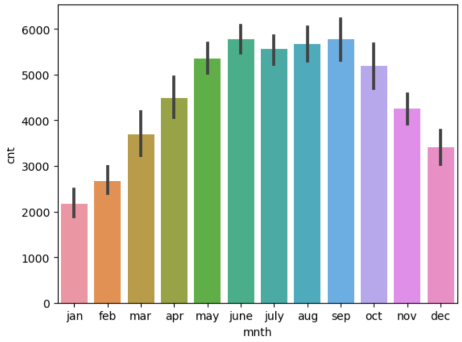
- Year

In 2019 there are more bike rentals as compared to 2018, which we can say it is increasing with year.



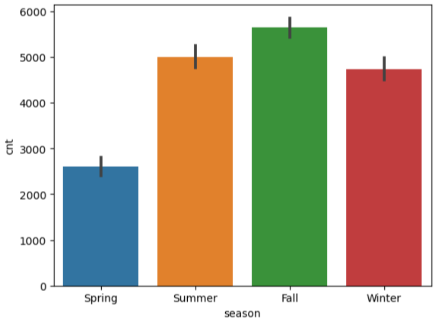
- Month

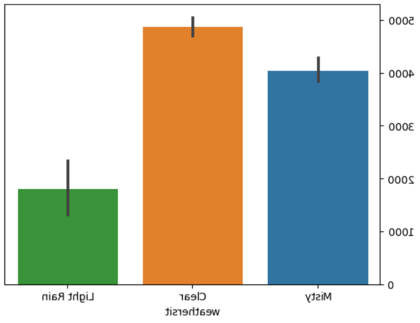
From May to September there are usually high rentals in comparison to other months.



- Season and Weather

All the seasons apart from spring have high rental bookings also when weather is Misty or Clear there are more bookings.





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

Ans:

By dropping one of the dummy columns from each categorical feature, we ensure there are no reference column means the remaining columns become linearly independent.

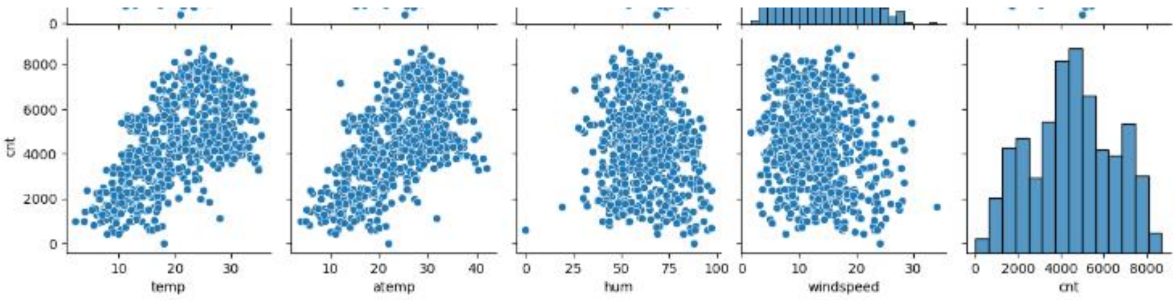
If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

Dropping your first categorical variable is possible because if every other dummy column is 0, then this means your first value would have been 1.

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

Ans:

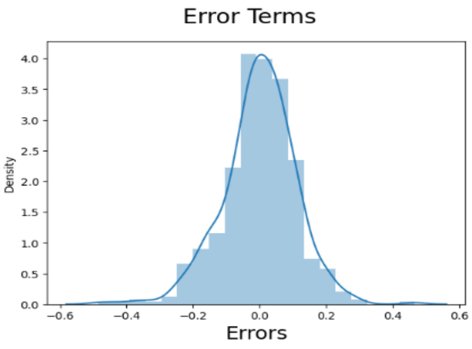
temp or atemp (Both are same as they have 0.99 correlation value) looks collinear   with target variable cnt



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

- Normal distribution of error terms:

The fourth assumption is that the error(residuals) follows a normal distribution.



- Error is normally distributed and looks concentrated to zero.

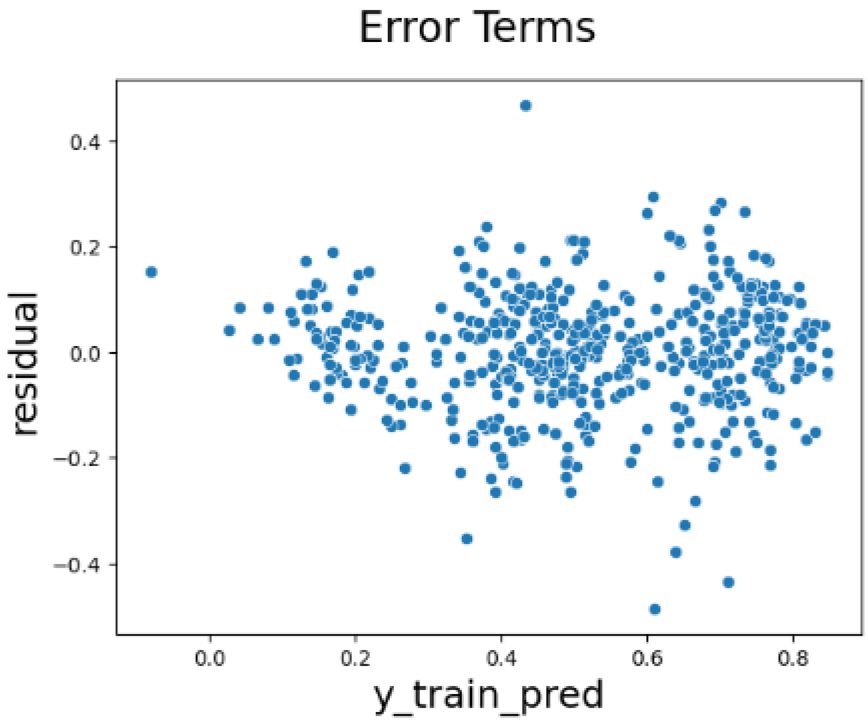
- No Multicollinearity:

We have calculated Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. And all independent variables of model were having VIF lower than 5 which is below the threshold.

- Homoscedasticity Assumption:

  Homoscedasticity means the residuals have constant variance at every level of x.

  Create a scatter plot that shows residual vs fitted value. If the data points are spread across equally without a prominent pattern, it means the residuals have constant variance (homoscedasticity). Otherwise, if a funnel-shaped pattern is seen, it means the residuals are not distributed equally and depicts a non-constant variance (heteroscedasticity).



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

Ans:

Top Features explaining the demands of bike rental are

Weather condition (weathersit):

If weather is having Light/Heavy rain or Snow there will be low demand. Else for Clear and Misty weather demand will be high.

Season:

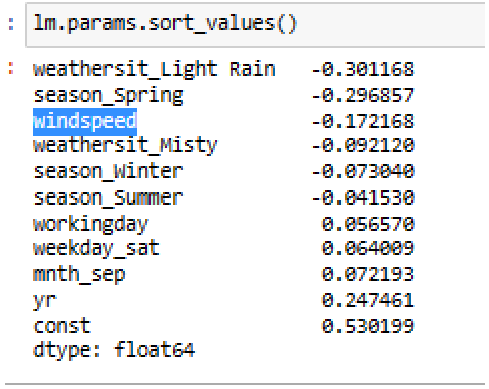
In case of Spring season demand will be low. Else it will be higher.

Year (yr):

Bike rental is increasing with year.

Windspeed:

As per our model in case of higher windspeed there will be low demand.



```
: lm.params.sort_values()

: weathersit_Light Rain   -0.301168
  season_Spring           -0.296857
  windspeed               -0.172168
  weathersit_Misty        -0.092120
  season_Winter           -0.073040
  season_Summer           -0.041530
  workingday               0.056570
  weekday_sat              0.064009
  mnth_sep                 0.072193
  yr                       0.247461
  const                    0.530199
  dtype: float64
```

General Subjective Questions

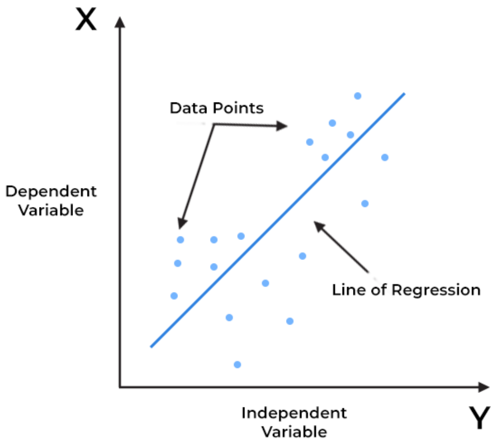
1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

Linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

Line of regression = Best fit line for a model



Best Fit Line for a Linear Regression Model

The cost function of linear regression is the root mean squared error or mean squared error (MSE).

Fundamentally, MSE measures the average squared difference between the observation's actual and predicted values. The output is the cost or score associated with the current set of weights and is generally a single number. The objective here is to minimize MSE to boost the accuracy of the regression model.

Along with the cost function, a 'Gradient Descent' algorithm is used to minimize MSE and find the best-fit line for a given training dataset in fewer iterations, thereby improving the overall efficiency of the regression model.

- **Simple Linear Regression**:

Simple linear regression uses traditional slope-intercept form, where m and b are the variables, our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

Mathematically a line is represented as:

Y = m*X + b

However, machine learning has a different notation to the above slope-line equation,

y(x) = b0 + b1 * x

y = output/dependent variable. Variable y represents the continuous value that the model tries to predict.

x = input variable or independent variable in statistics

b0 = y-axis intercept.

b1 = the regression coefficient or scale factor. b1 is the equivalent of the slope of the best-fit straight line of the linear regression model. The equation for multiple linear regression is similar to the equation for a

- Multiple linear regression

Multiple linear regression establishes the relationship between independent variables (two or more) and the corresponding dependent variable. Here, the independent variables can be either continuous or categorical. This regression type helps foresee trends, determine future values, and predict the impacts of changes.

simple linear equation, i.e., y(x) = b0 + b1x1 plus the additional weights and inputs for the different features which are represented by b(n)x(n). The formula for multiple linear regression would look like,

y(x) = b0 + b1x1 + b2x2 + … + b(n)x(n)

The machine learning model uses the above formula and different weight values to draw lines to fit. Moreover, to determine the line best fits the data, the model evaluates different weight combinations that best fit the data and establishes a strong relationship between the variables.

Assumptions of Linear Regression

1. Linear relationship: There exists a linear relationship between each predictor variable and the response variable.

2. No Multicollinearity: None of the predictor variables are highly correlated with each other.

3. Independence: The observations are independent.

4. Homoscedasticity: The residuals have constant variance at every point in the linear model.

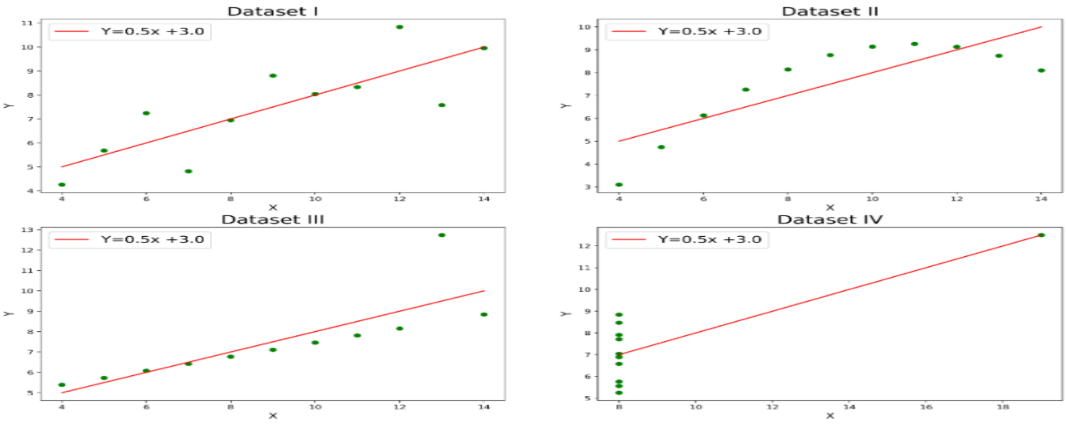
5. Multivariate Normality: The residuals of the model are normally distributed.

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

Ans:

Anscombe’s quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

It is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

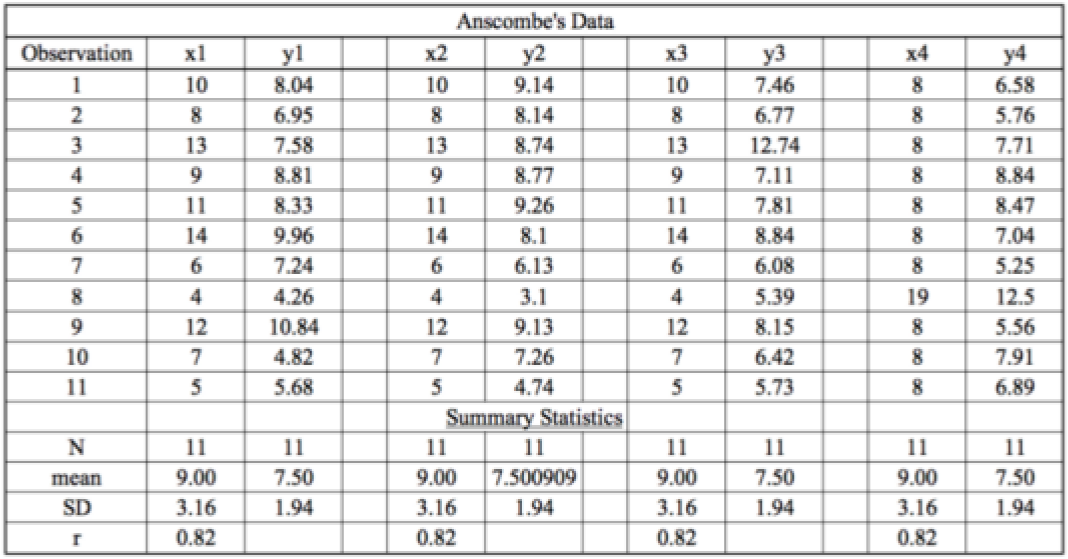


Anscombe's quartet Plot

1. Dataset 1 fits linear regression model as it seems to be linear relationship between X and y

2. Dataset 2 does not show a linear relationship between X and Y, which means it does not fit the linear regression model.
3. Dataset 3 shows some outliers present in the dataset which can’t be handled by a linear regression model.
4. Dataset 3 has a high leverage point means it produces a high correlation

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.



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	

3. What is Pearson’s R?

Ans:

The Pearson correlation method is the most common method used for measuring a linear correlation for numerical variables. It assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product- moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.

Pearson’s R Formula is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

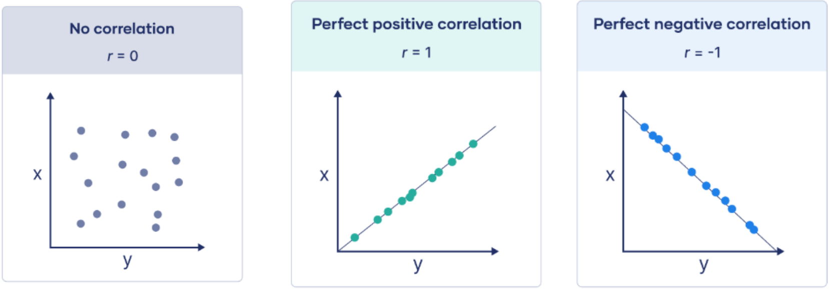
r = correlation coefficient

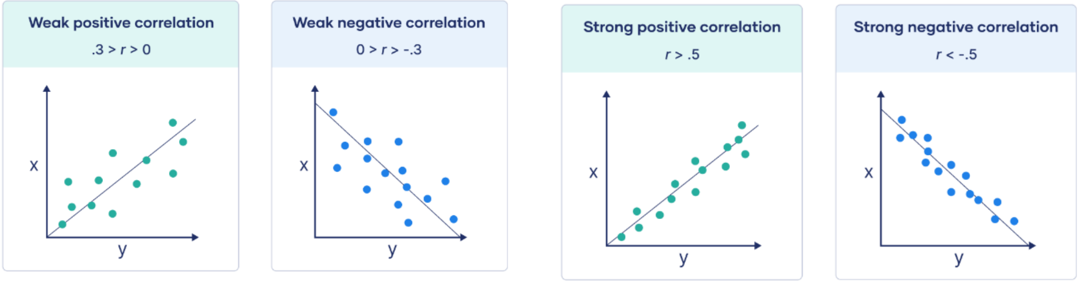
xi = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

yi = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable





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

Ans:

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

Feature scaling is a data preprocessing technique that involves transforming the values of features or variables in a dataset to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model. Feature scaling is essential when working with datasets where the features have different ranges, units of measurement, or orders of magnitude. Common feature scaling techniques include standardization, normalization, and min-max scaling. By applying feature scaling, the data can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models.

Why to do scaling:

- Ease of interpretation.
- Having features on a similar scale can help the gradient descent converge more quickly towards the minima.
- we scale our data before employing a distance-based algorithm so that all the features contribute equally to the result.
- 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.

Normalized scaling

Normalization is a data preprocessing technique used to adjust the values of features in a dataset to a common scale. This is done to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Xmax and Xmin are the maximum and the minimum values of the feature, respectively.

Standardization

Standardization is another scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Formula:

$$X' = \frac{X - \mu}{\sigma}$$

µ = is the mean of the feature values

sigma = standard deviation of the feature values

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

Ans:

Variance Inflation Factor (VIF): VIF determines the strength of the correlation between the independent variables. VIF score of an independent variable represents how well the variable is explained by other independent variables

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

- So, the closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

Variance Inflation Factor (VIF) can be used in solving multicollinearity in a regression analysis. If multicollinearity is detected among predictor variables, VIF can help identify which variables are contributing to the multicollinearity. The higher the VIF value for a variable, the more it contributes to multicollinearity. Removing variables with high VIF values can help reduce multicollinearity and improve the accuracy and stability of the regression model. A rule of thumb is to remove variables with VIF values greater than 5 or 10, depending on the specific context. VIF values have standard errors, and the confidence intervals for the VIF can be calculated and used to assess the significance of the collinearity.

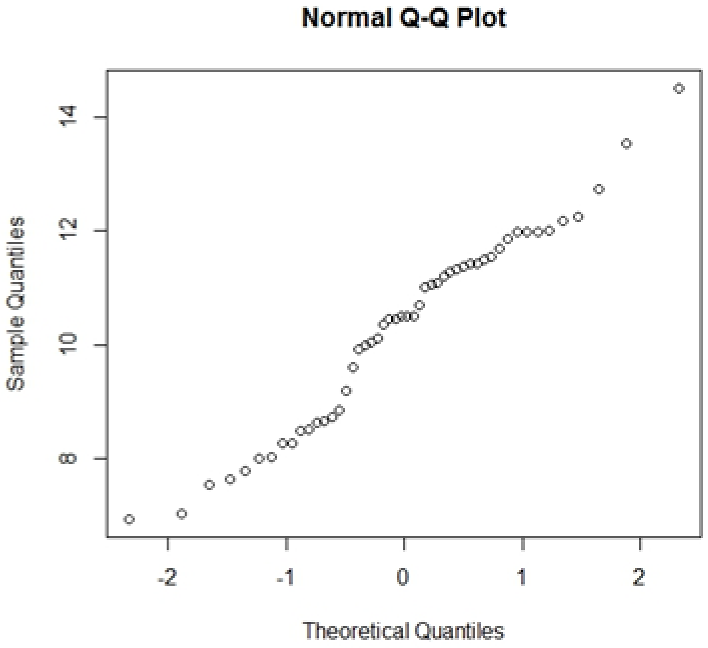
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Ans:

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.