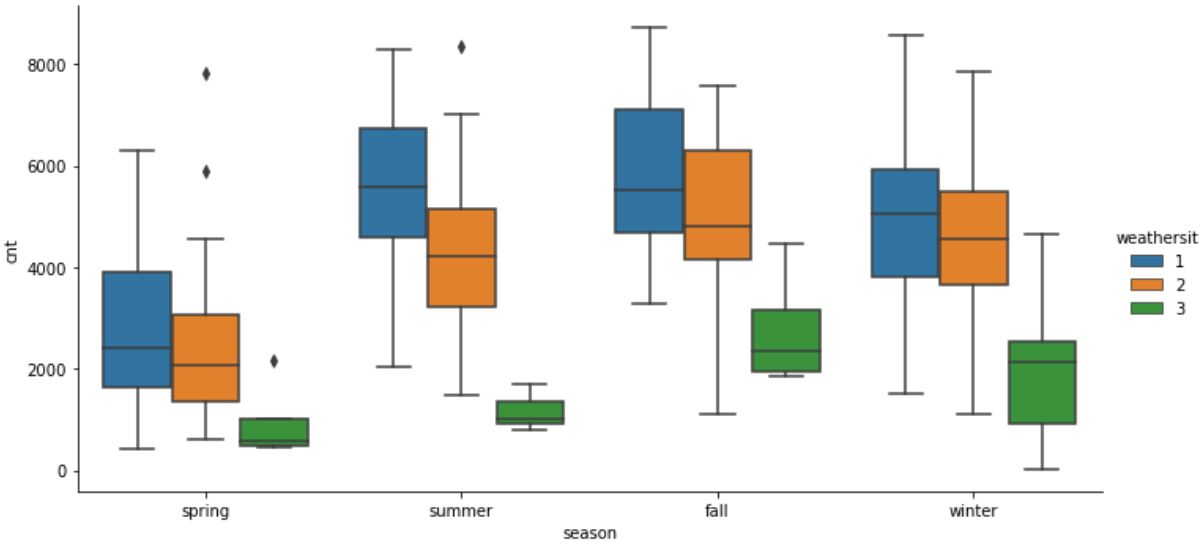
Assignment-based Subjective Questions

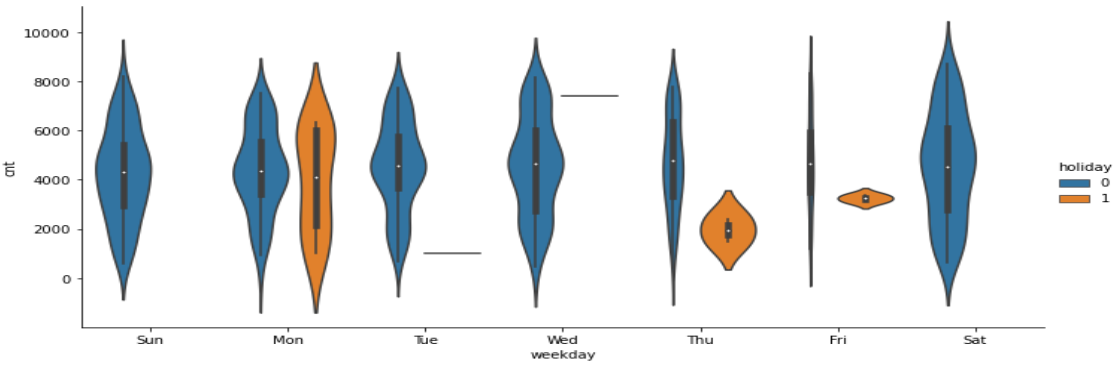
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: From our analysis we identified below categorical features from Data preparation.
Categorical features = season, yr, holiday, weekday, workingday, weathersit
Here 'cnt' is the dependent/target variable which indicates total number of bike rentals
Below are the assumptions/inferences from our EDA on these categorical variables

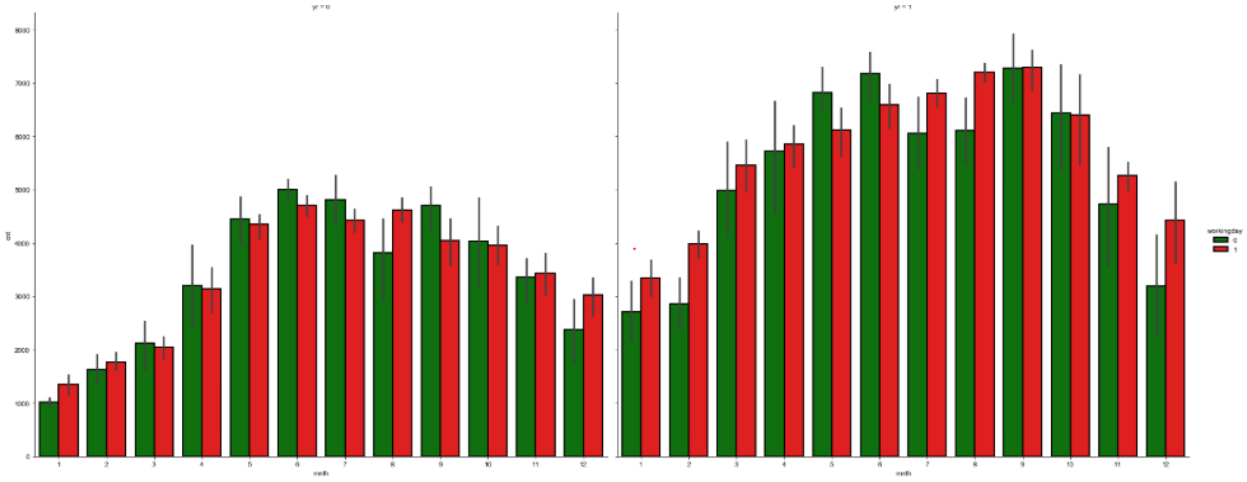
a. The box plot describes count of total rental bikes (cnt) across various seasons and weathers. We can see that bike rentals are more in fall season for weathersit: 1(weathersit: - 1: Clear, Few clouds, partly cloudy) as compared with other seasons.



b. The violin plot depicts total rental bikes across various days of week also on holiday vs no holiday. We can observe that during holiday and on weekends (Saturday, Sunday) count is more compared to no-holiday. But for Wednesday which is a weekday where rental count is more.



c. Below is the total rental bikes count across various months and year during working day or non-working day. As per the below bar graph Count of rental bikes are more for categorical variable yr 1 (yr: year (0: 2018, 1:2019)) for workingday in the month (9: September) as compared to other months



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

Answer:

A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels.

It is preferable to use drop_first=True. This will give n-1 dummy variables. There is no loss of information.E.g. For 4 seasons 3 variables can represent 4 seasons. 4th season can be expressed by 3 variables being zero.

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

Answer:

From the Calculation of correlation coefficient and pairplot we can observe that the output/target variable (cnt—total number of bike rentals) is having high correlation with 'atemp' variable which is 0.631.

atemp—0. 631066
temp ---0. 6274
yr ---0.566710
season 0.406100

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

Answer:

After building the model on training set based on OLS Regression results we will assess the final model.

a. After determining the highest significant coefficient, we will use p- value metrics to verify overall model fit is significant the p-value should be less (0.05)

b. After dropping the variables which are having high p-value we need to focus on VIF factor (Variance Inflation Factor) it should be less than 5.

c. We need to train the model until both VIF and p-values are within acceptable range. So we go ahead and make our predictions using the final model.

d. Perform the residual analysis to check if the error terms are also normally distributed.

e. After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

Below are the expectations to build a significant model.

1. The Two Variables should be in a Linear Relationship.

2. All the Variables Should be Multivariate Normal.

3. There should be No Multicollinearity in the Data.

4. There should be No Autocorrelation in the Data.

5. There should be Homoscedasticity among the Data

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

Answer:

From our final model below are top3 features

a. Temp (β =0.557) min=2.424346 and max=35.328347 So as the temp increases count in rental bike also increases. As temp around 20-30 is considered favourable for customers for riding bike

b. Yr (β =0.236) this feature denotes 2019 year (after first_drop during dummy variable creation. In this year i.e. 2019, there was a drastic increase in count of rental bike. It might be because of gradual popularity gain over years, health consciousness, environment or pollution related consciousness, increasing traffic where 2 wheeler is a better option than 4 wheeler.

c. winter (β =0.1374) This feature denotes in winter there was a drastic increase in count of rental bike.

General Subjective Questions

1. **Explain the Linear Regression Algorithm in detail**

Answer:

Machine Learning is means the machine learns from the data as humans learn from their activities. Which is the algorithm and used in AI(Artificial Intelligence) to predict future/target/dependent variables.
ML models are classified in 3 types:
1. Regression: The Dependent variable to be predicted is a real or continuous variable, Ex: Salary
2. Classification: The dependent variable to be predicted is categorical variable, ex: mails (spam or Ham)
3. Clustering: There will not be a labeling for group\clusters in this, ex: In shopping Customer segmentation to grant discount.
Regression is a commonly used predictive analysis model. It is mainly estimating the relationship between variables

There are 2 types of linear regressions

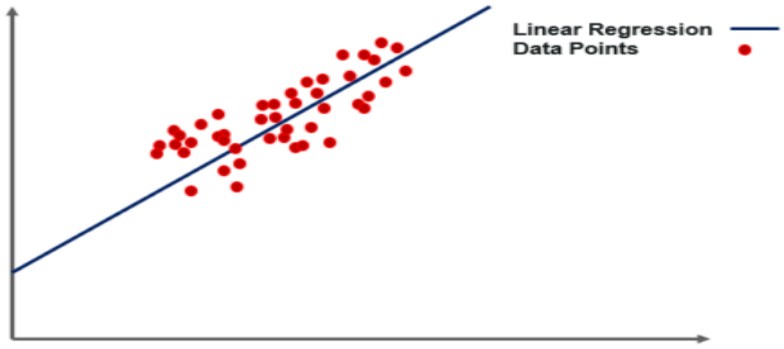
a. Simple Linear Regression:

The independent variable has linear relationship with dependent variable. (i.e. Explains change in dependent variable with change in the value of predictors). Here the number of independent variables =1

Equation of the best fit regression line $Y = \beta_o + \beta_1 X$

β_o --- Y intercept
β_1 --- Slope
Y --- Dependent variable
X --- Independent variable

The below straight line image shows the best fit line. The main aim of Simple Linear Regression is to consider the data points and to plot the best fit line to fit in a model in the best possible way.



b. Multiple Linear Regression:

The MLR is an extension of simple linear regression model in which Multiple independent variables/features that are used to predict the dependent variable. Here the number of independent variables =1

Equation of the best fit regression line $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n$

β_o	--- Y intercept
$\beta_1, \beta_2, \beta_3$---- $\beta n X n$	--- Slopes (co-efficients)
Y	--- Dependent variable
X1, X2, X3------X_n	--- Independent variable

Cost function:

The Cost function helps to identify the best possible values for β_o and β_1 which would provide the best fit line for data points. The cost function takes two values, i.e. (β1, β0), where β1 is the coefficient and c is the intercept. The cost function iterates through each point in the given dataset and then computes the sum of the square distance between each point and the line. It is also known as MSE (Mean Squared Error)

$$J(\beta_1, \beta_0) = \sum^{N}_{i=1} (y_i - y_i(p))^2$$

Our aim 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.

Gradient descent:

Gradient descent is an optimization algorithm used to find the values of the parameters (coefficients) of a function (f) that minimizes a cost function.

The strength of linear regression model is explained by

1. R^2 (Coefficient of Determination): It takes always value between 0 & 1, the higher R^2 the better model fits your data.

$R^2 = 1 - (RSS/TSS)$.

RSS: Residual sum of squares

TSS: Total sum of squares

2. Explain the Anscombe's quarter in detail

Answer:

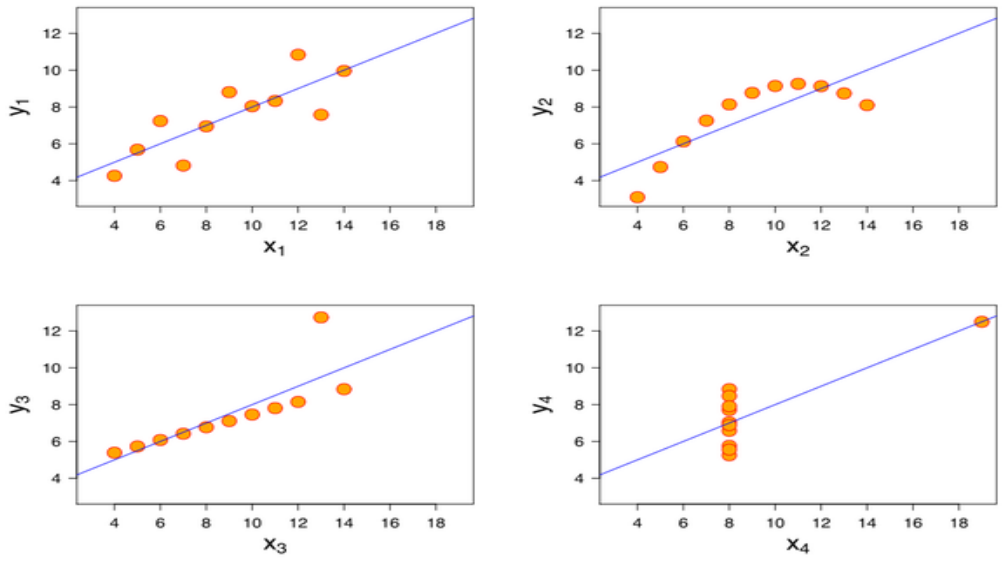
"Numerical calculations are exact, but graphs are rough." Below represents a table of numbers. It contains four distinct datasets (hence the name Anscombe's Quartet).

1. Each with statistical properties that is essentially identical:

2. The mean of the x values is 9.0, mean of y values is 7.5, they all have nearly identical variances, correlations, and regression lines (to at least two decimal places).

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

But when plotted, they suddenly appear very different as below.



While dataset I appear like many well-behaved datasets that have clean and well-fitting linear models, the others are not served nearly as well.

Dataset II does not have a linear correlation;
Dataset III does, but the linear regression is thrown off by an outlier. It would be easy to fit a correct linear model, if only the outlier were spotted and removed before doing so.
Dataset IV, finally, does not fit any kind of linear model, but the single outlier makes keeps the alarm from going off.

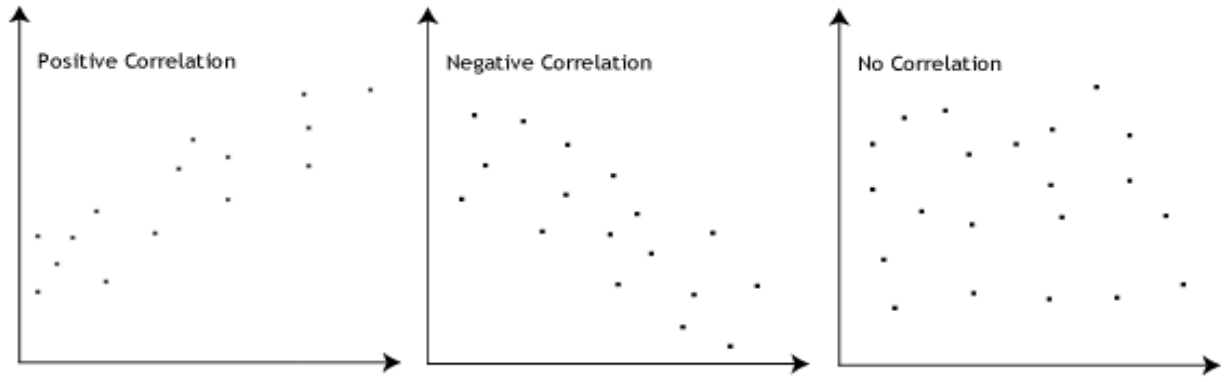
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

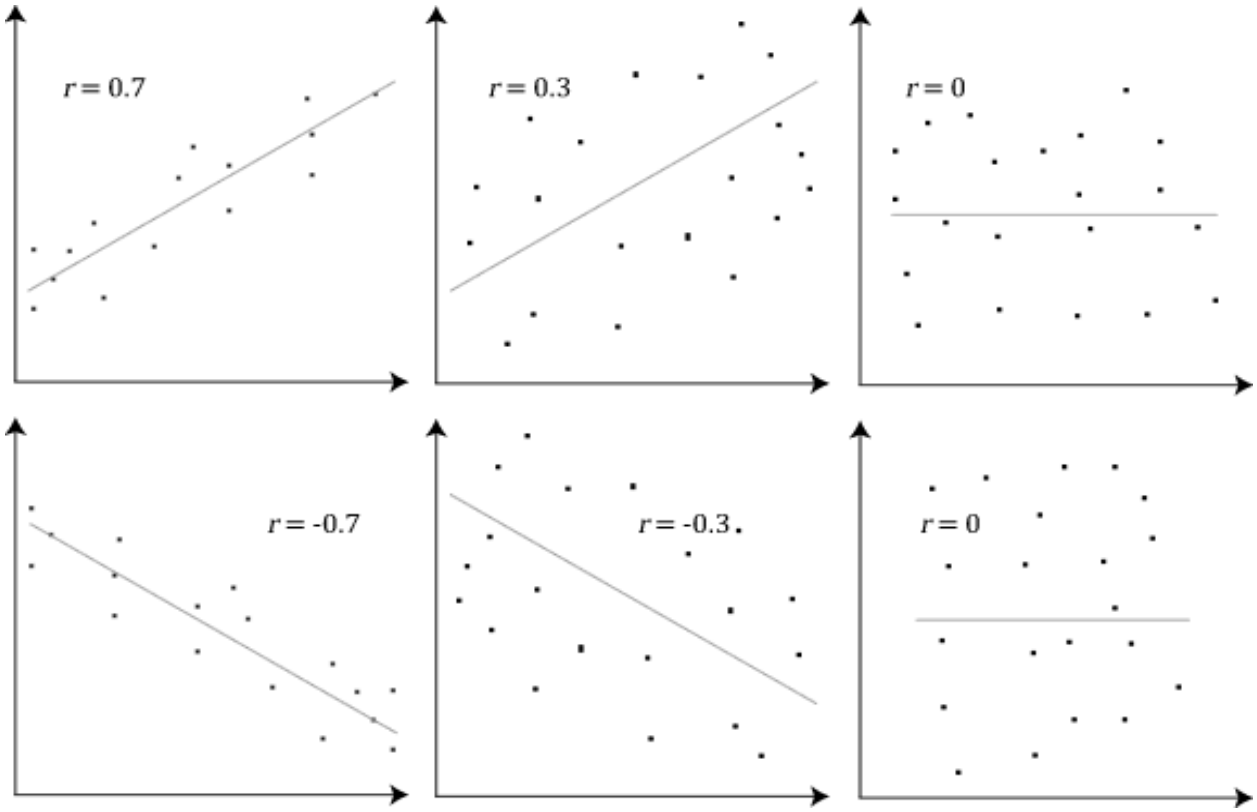
Answer:

The Pearson correlation coefficient or Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson correlation coefficient attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of r to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Answer:

Scaling data is the process of increasing or decreasing the magnitude according to a fixed ratio, in simpler words you change the size but not the shape of the data.

It helps handling disparities in units. During long processes it definitely helps reduce computational expenses.

In the machine learning eco space, it helps improve the performance of the model and reducing the values/models from varying widely.

Normalized Scaling:

Normalization often also simply called Min-Max scaling basically shrinks the range of the data such that the range is fixed between 0 and 1 (or -1 to 1 if there are negative values). It works better for cases in which the standardization might not work so well. If the distribution is not Gaussian or the standard deviation is very small, the min-max scaler works better. Normalization is typically done via the following equation:

$Xi1 = (Xi – \min(X))/(\max(X) – \min(X))$

Standardized scaling:

Standardization (or Z-score normalization) is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with $\mu = 0$ and $\sigma = 1$

Where μ is the mean and σ is the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows:

$Z = (X - \mu)/ \sigma$

Normalizing the data is sensitive to outliers, so if there are outliers in the data set it is a bad practice. Standardization creates a new data not bounded (unlike normalization).

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

Answer:

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$$Y_1 = C + \beta_2 X_2 + \beta_3 X_3 + \cdots$$
$$VIF_1 = 1/ (1-R_1^2)$$

Next, we fit the model between Y2 and the other independent variables to estimate the coefficient of

Determination R2:

$$Y_2 = C + \beta_1 X_1 + \beta_3 X_3 + \cdots$$
$$VIF_2 = 1/ (1-R_2^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with

high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

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

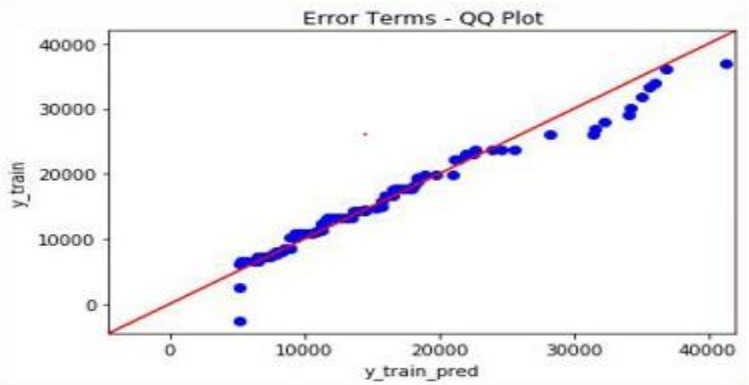
Answer:

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

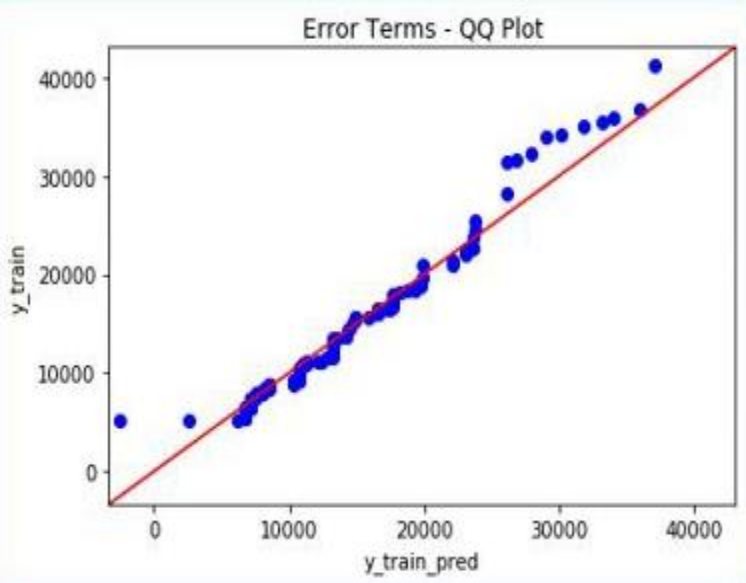
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Use and importance of a Q-Q plot in linear regression:

a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
If we have 2 data sets, it can be used to check some scenarios like,
i. whether from populations with a common distribution
ii. Have common location and scale
iii. Have similar distributional shapes
iv. Have similar tail behavior
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis.