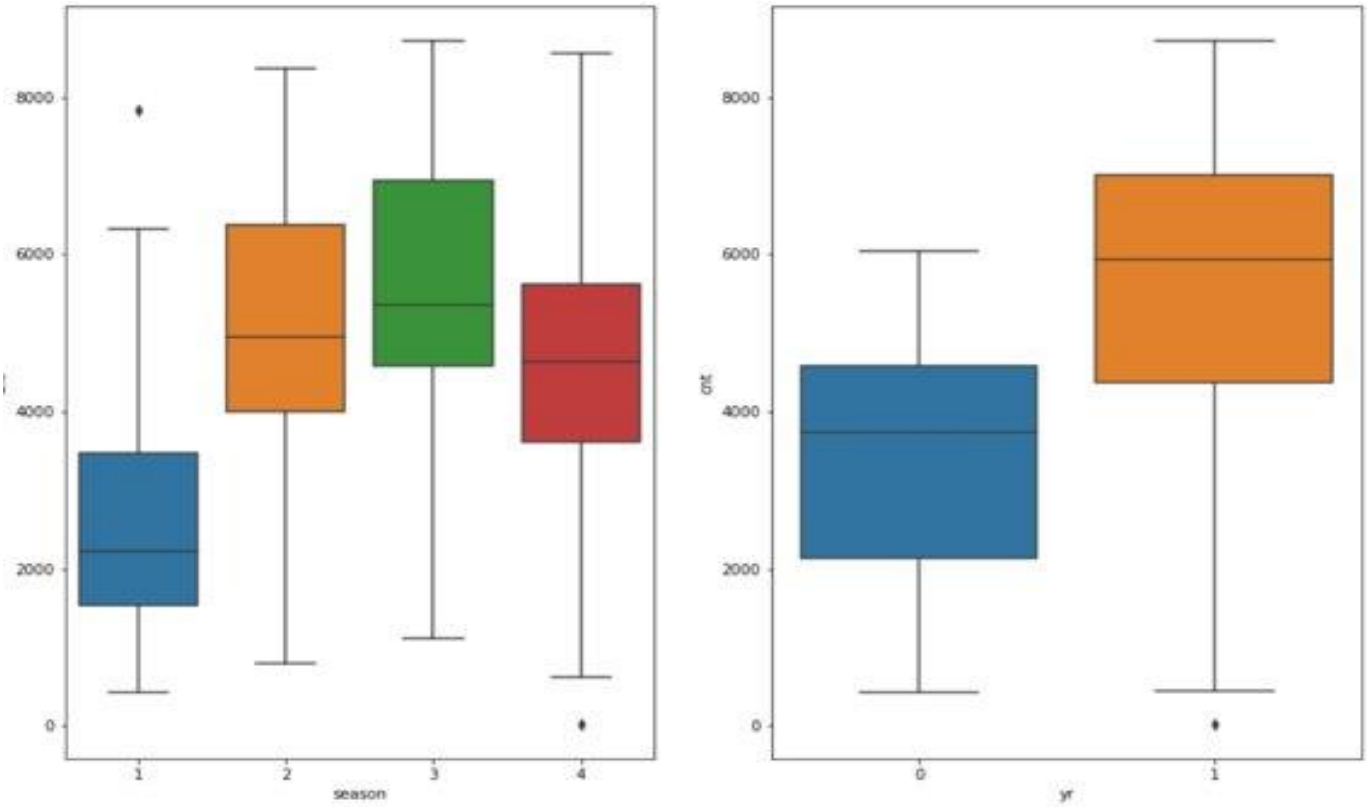
Assignment-based Subjective Questions

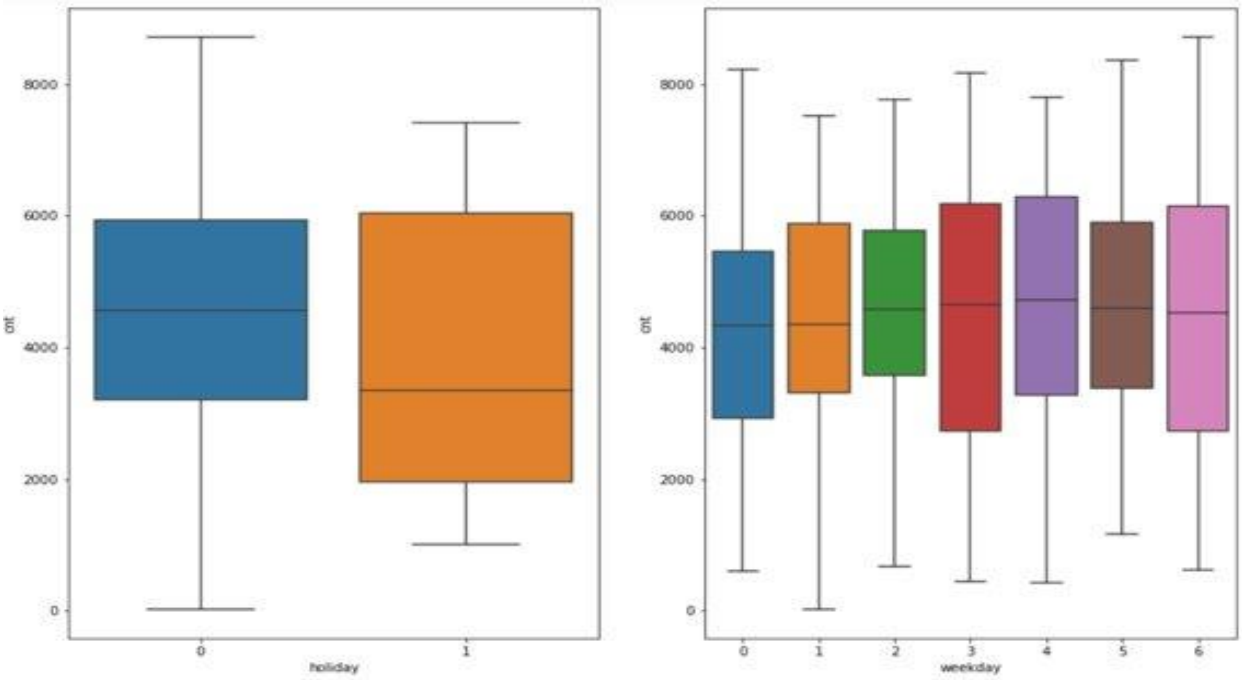
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:



Year(yr):- We can see more demand of bike booking in 2019 as compared to 2018 because of its more popularity and advertisement.

season (1: spring, 2: summer, 3: fall, 4: winter): - Fall has the maximum median, as weather conditions are too good in fall compared to spring and winter

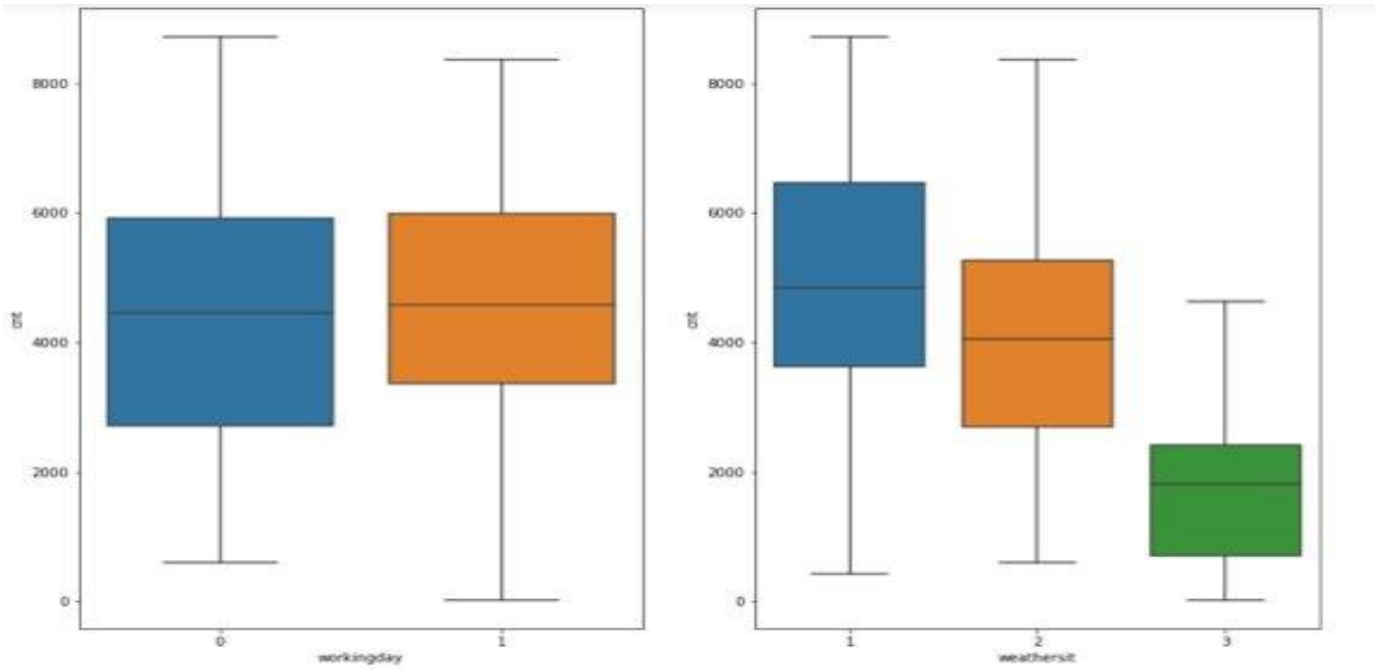


Holiday: - Median of bike booking is more on not a holiday while spread is more on holiday.

More bike booking on non-holiday as compared to holiday.

On holiday spent time with family and book cabs instead of bike.

Weekday: - Medain is same across all days, but spread is more on 3rd and 6th day.



Workingday: - Median is same for working and non-working day. But spread is more on non-working day as they have to plans for non-working day

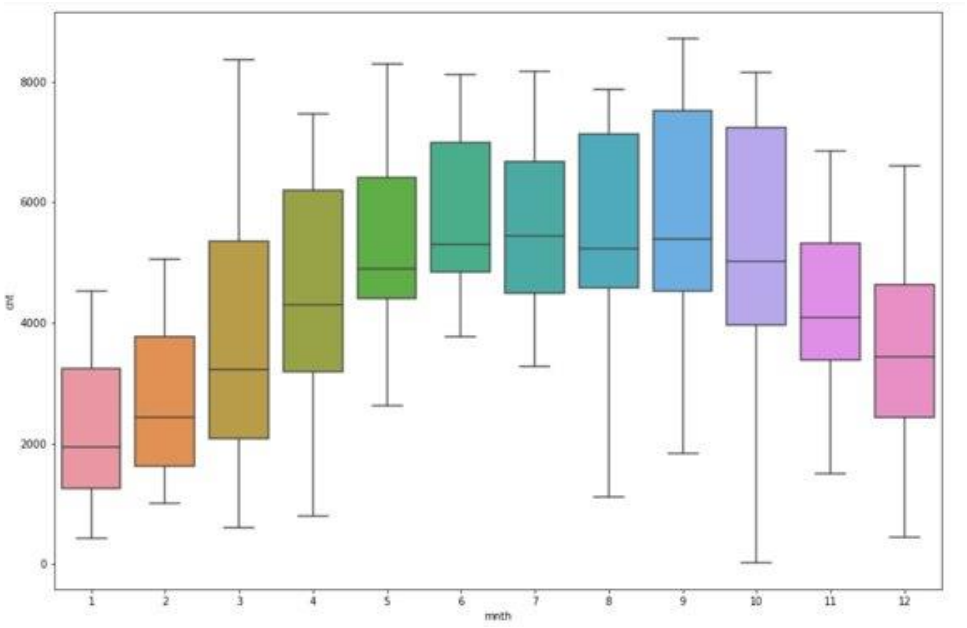
weathersit : - 1: Clear, Few clouds, Partly cloudy, Partly cloudy

- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

It's clearly visible that if weather is clear, few clouds its good for a ride. So most demand in this season. When its mist and cloudy demand of bike will reduce.



Mnth: As weather varies with month , so it's much clear from above. That demand will also vary with mnth

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

It's important to drop first dummy variable to avoid redundant feature.

A variable with n values can be represented by n-1 values.

If you don't use "drop_first" you will get a redundant feature, let's see an example.

If you have a feature "Is_male", you use "get_dummies" you will get two features "Is_male_0" and "Is_male_1", but if you look carefully, they are redundant actually you just need one of them, the other one will the exact opposite of the other.

Example :-

row	is_male
0	1
1	0

after applying get_dummies

row	is_male_0	is_male_1
0	0	1
1	1	0

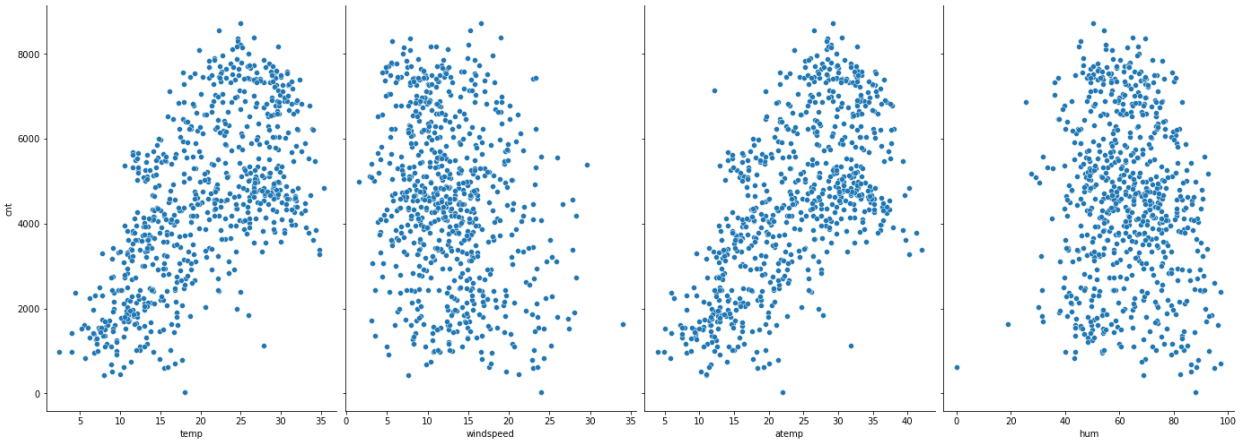
So, we don't want to keep redundant features.

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

Answer : Temperature(temp) and atemp has highest correlation with value of 0.65

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

Assumption: Linear relationship



Clearly observed that temp and temp are positively linear related to cnt and windspeed and hum ae negatively related.

Assumption **No Perfect Multicollinearity**

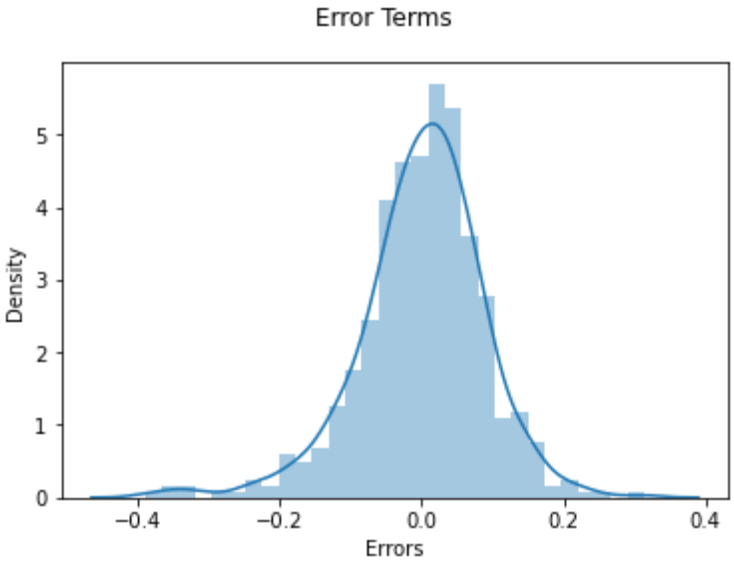
VIF calculation.

In case of very less variables, one could use heatmap, but that isn't so feasible in case of large number of columns.
Another common way to check would be by calculating VIF (Variance Inflation Factor) values.
If VIF=1, Very Less Multicollinearity
VIF<5, Moderate Multicollinearity
VIF>5 , Extreme Multicollinearity (This is what we have to avoid)

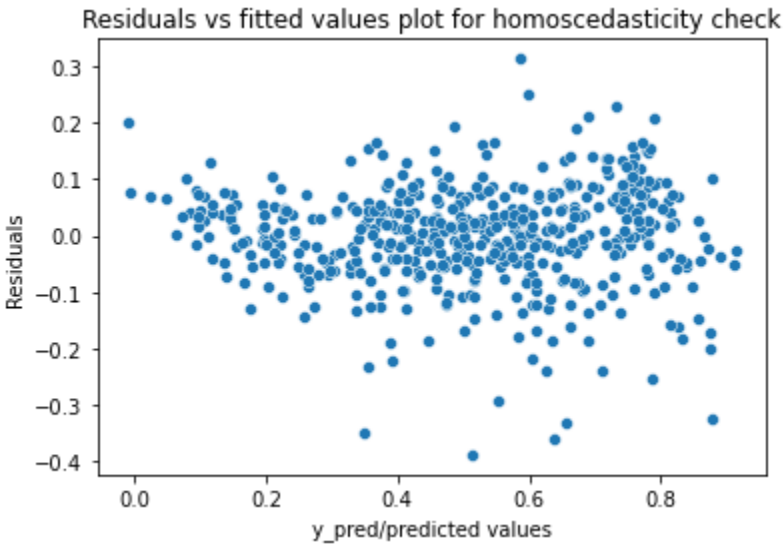
In our model all independent features have VIF<5

	features	VIF
0	const	51.11
4	hum	1.88
2	workingday	1.65
8	Sat	1.64
3	temp	1.60
10	Mist	1.56
11	July	1.43
6	summer	1.33
7	winter	1.29
9	Light Snow	1.24
12	September	1.19
5	windspeed	1.18
1	yr	1.03

Assumption: ***Residuals must be normally distributed.***



- Assumption: **Homoscedasticity:** Homoscedasticity means that the residuals have equal or almost equal variance across the regression line. By plotting the error terms with predicted terms, we can check that there should not be any pattern in the error terms.



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

- **Temperature** (0.595988) -: A coefficient value of '0.595988' indicated that a unit increase in temp variable increases the bike hire numbers by 0.595988 units

- **weathersit** (Light Snow: (-0.239137) Mist : (-0.053623) =0.29276): - A coefficient value of '-0.29276' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.29276 units
- **Year** (yr) - A coefficient value of '0.228581' indicated that a unit increase in yr variable increases the bike hire numbers by 0.228581 units.

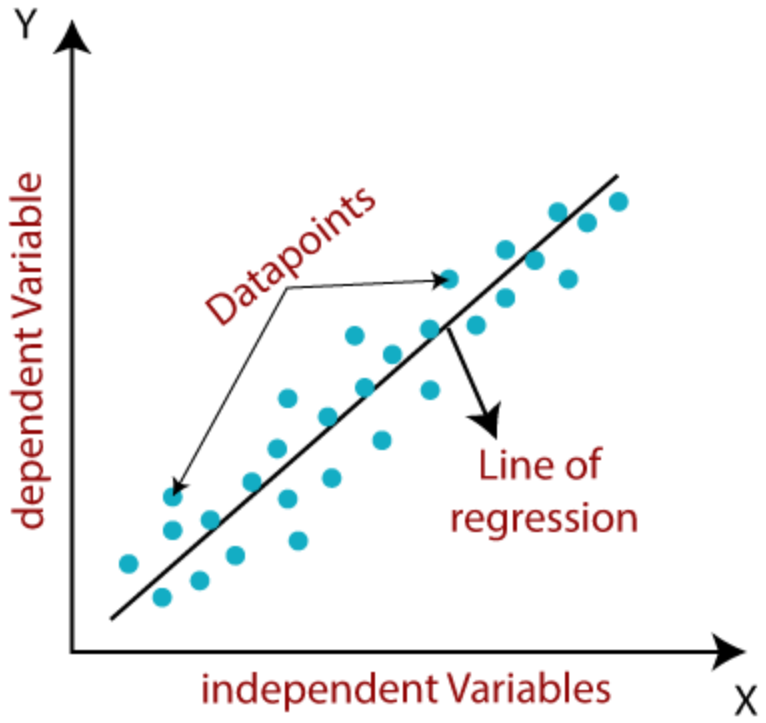
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

$y = a_0 + a_1 x + \varepsilon$

Here,

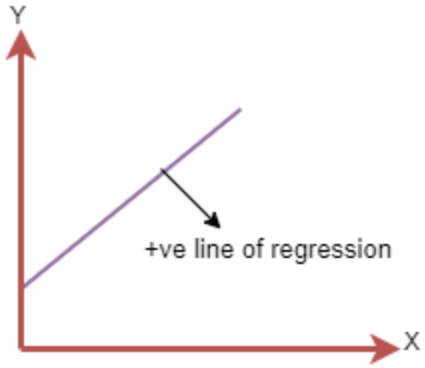
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Linear Regression Line

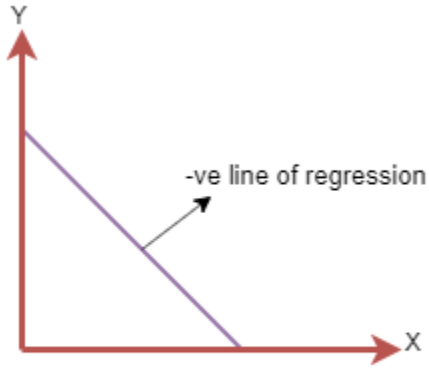
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $\mathbf{Y= a_0+a_1x}$

- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by the following method:

1. R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- **Linear relationship between the features and target:** Linear regression assumes the linear relationship between the dependent and independent variables.
- **Small or no multicollinearity between the features:** Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may be difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.
- **Homoscedasticity Assumption:** Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot**.**
- **Normal distribution of error terms:** Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.

2. Explain the Anscombe's quartet in detail.

Anscombe quarter is **a group of four data sets that are nearly identical in simple descriptive statistics, which provide the same information (involving variance and mean) for each x and y point in all four data sets, but there are peculiarities that fool the regression model once you plot each data set. However, when you plot these data sets, they look very different from one another.**

Anscombe's quartet states: **Before analyzing your data and building your model, you must first plot the data set.** Illustrates the i**mportance of plotting data before you analyze it and build your model.**

This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

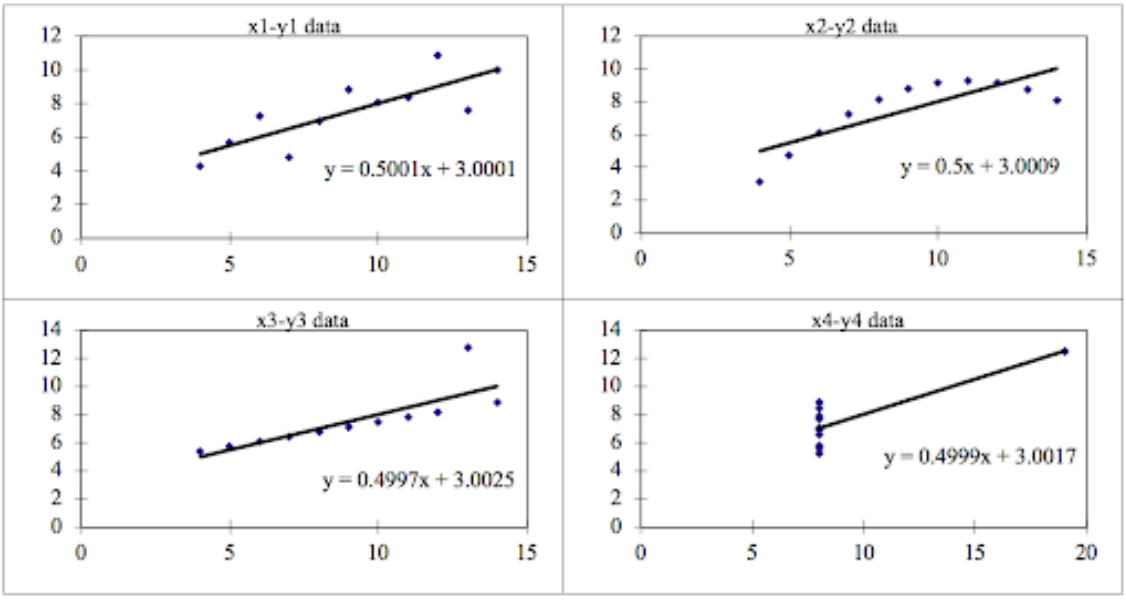
We can define these four plots as follows:

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

The statistical information for these four data sets are approximately similar. We can compute them as follows:

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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



Data Set 1: fits the linear regression model pretty well

Data Set 2: cannot fit the linear regression model because the data is non-linear

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model

Conclusion: Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other and is denoted by r.

Basically, a Pearson product-moment correlation 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).

In simple words, Pearson's correlation coefficient **calculates the effect of chang**e in one variable when the other variable changes.

For example: Up till a certain age, (in most cases) a child's height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc.

The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

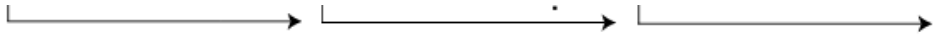
For example:

- **Positive linear relationship:** In most cases, universally, the income of a person increases as his/her age increases.
- **Negative linear relationship:** If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

From the example above, it is evident that the Pearson correlation coefficient, r, tries to find out two things – the strength and the direction of the relationship from the given sample sizes.

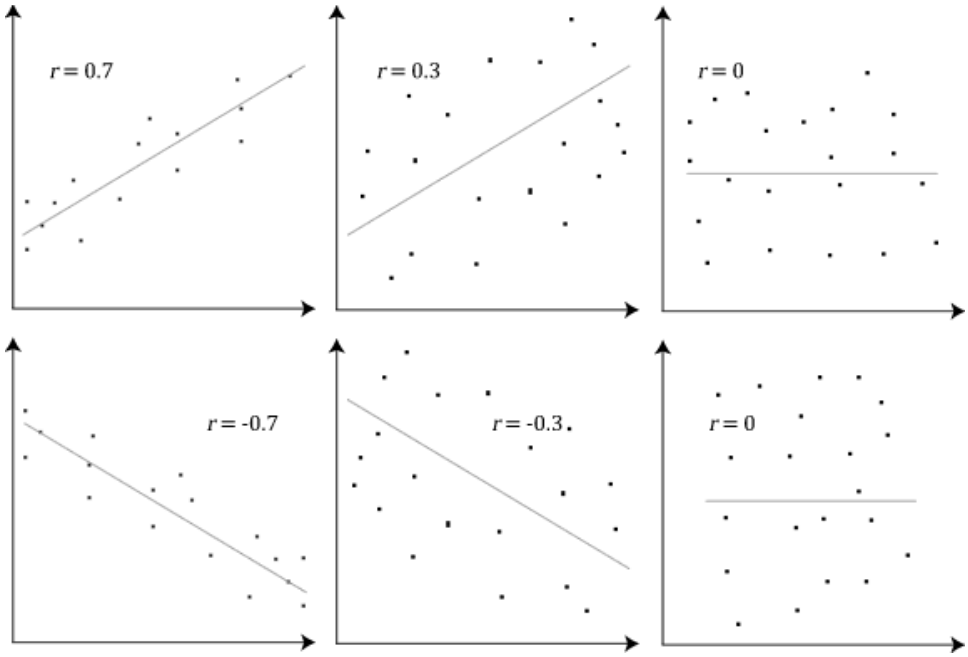
Pearson correlation coefficient values: -

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:



Strength of association based on the Pearson correlation coefficient?

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:



Pearson correlation coefficient formula

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1. Use the below Pearson coefficient correlation calculator to measure the strength of two variables.

***Requirements**:*

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

Pearson correlation coefficient formula:

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

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Pearson correlation coefficient calculator

Here is a step-by-step guide to calculating Pearson's correlation coefficient:

Step one: Create a Pearson correlation coefficient table. Make a data chart, including both the variables. Label these variables 'x' and 'y.' Add three additional columns – (xy), (x^2), and (y^2). Refer to this simple data chart.

Person	Age (x)	Income (y)	(xy)	(x^2)	(y^2)
1					
2					
3					
4					

Step two: Use basic multiplication to complete the table.

Person	Age (x)	Income (y)	(xy)	(x^2)	(y^2)
1	20	1500	30000	400	2250000
2	30	3000	90000	900	9000000
3	40	5000	200000	1600	25000000
4	50	7500	375000	2500	56250000

Step three: Add up all the columns from bottom to top.

Person	Age (x)	Income (y)	(xy)	(x^2)	(y^2)
1	20	1500	30000	400	2250000
2	30	3000	90000	900	9000000
3	40	5000	200000	1600	25000000
4	50	7500	375000	2500	56250000
Total	140	17000	695000	5400	92500000

Step four: Use the correlation formula to plug in the values.

Conclusion: If the result is negative, there is a negative correlation relationship between the two variables. If the result is positive, there is a positive correlation relationship between the variables. Results can also define the strength of a linear relationship i.e., strong positive relationship, strong negative relationship, medium positive relationship, and so on.

Pearson correlation coefficient guidelines:

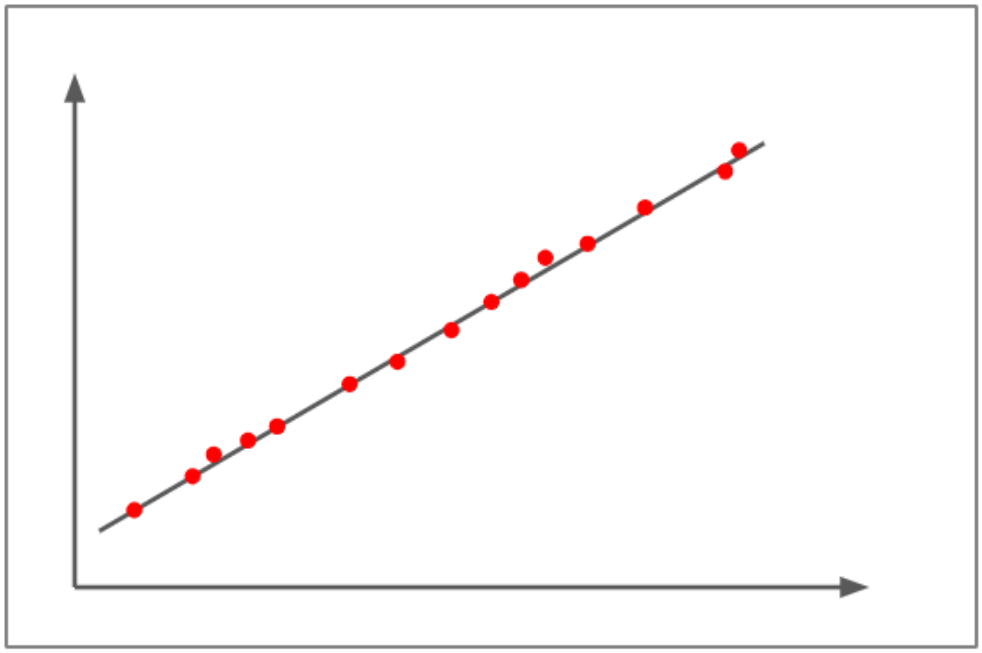
	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Remember that these values are guidelines and whether an association is strong or not will also depend on what you are measuring.

Examples of Pearson's correlation coefficient

Let's look at some visual examples to help you interpret a Pearson correlation coefficient table:

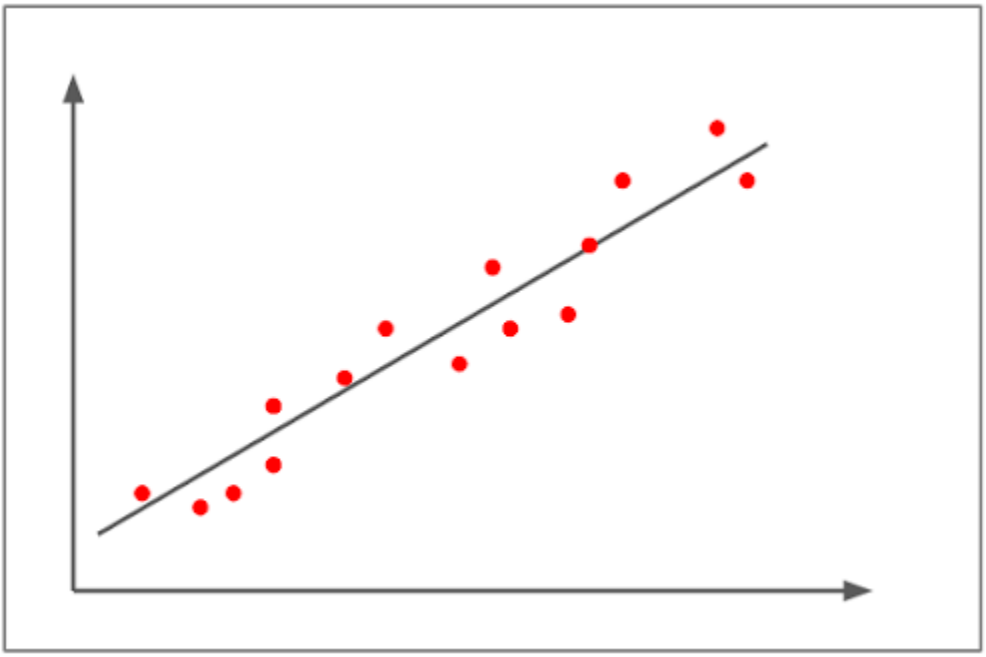
- **Large positive correlation:**



The above figure depicts a correlation of almost +1.
The scatterplots are nearly plotted on the straight line.
The slope is positive, which means that if one variable increases, the other variable also increases, showing a positive linear line.
This denotes that a change in one variable is directly proportional to the change in the other variable.
An example of a large positive correlation would be – As children grow, so do their clothes and shoe sizes.

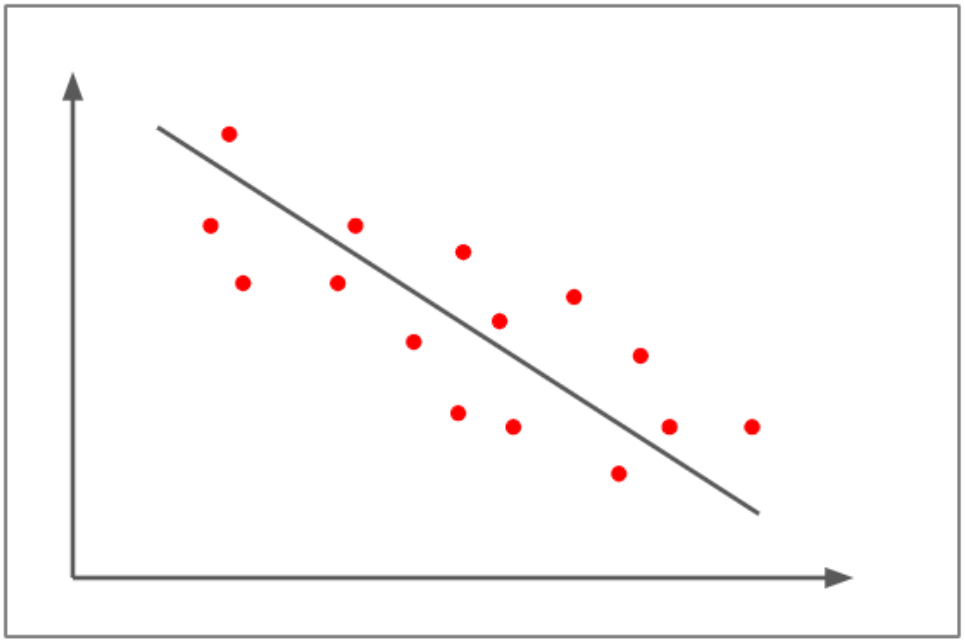
Let's look at some visual examples to help you interpret a Pearson correlation coefficient table:

- **Medium positive correlation:**



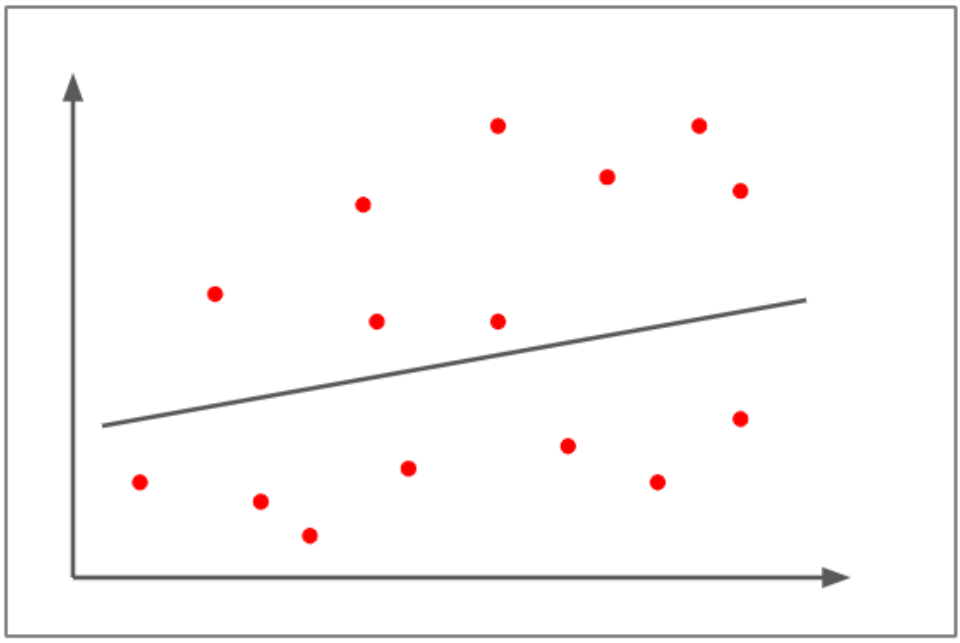
The figure above depicts a positive correlation.
The correlation is above than +0.8 but below than 1+.
It shows a pretty strong linear uphill pattern.
An example of a medium positive correlation would be – As the number of automobiles increases, so does the demand in the fuel variable increases.

- **Small negative correlation**



In the figure above, the scatter plots are not as close to the straight line compared to the earlier examples
It shows a negative linear correlation of approximately -0.5
The change in one variable is inversely proportional to the change of the other variable as the slope is negative.
An example of a small negative correlation would be – The more somebody eats, the less hungry they get.

- **Weak / no correlation**



The scatterplots are far away from the line.
It is tough to practically draw a line.
The correlation is approximately +0.15
It can't be judged that the change in one variable is directly proportional or inversely proportional to the other variable.
An example of a weak/no correlation would be – An increase in fuel prices leads to lesser people adopting pets.

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

What is scaling?

It is a data Pre-Processing step 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.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using the feature scaling method, then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

Why is scaling performed?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. 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.

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-squared, etc

Techniques for feature scaling:

Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

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

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

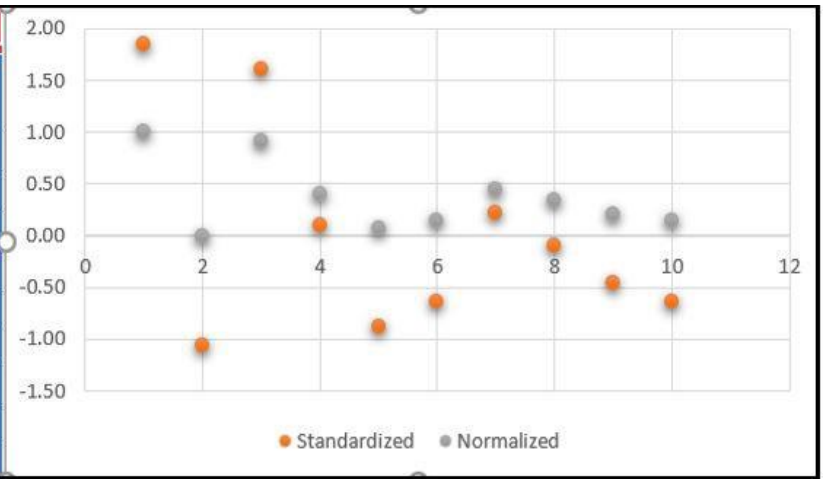
  ```
  X_new = (X - mean)/Std
  ```

It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Example:

Below shows example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

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.

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

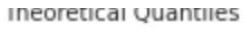
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

Normal Distribution

Uniform distribution



Exponential Distribution

Check for skewness of distribution:

Left-skewed distribution

Right-skewed distribution

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.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

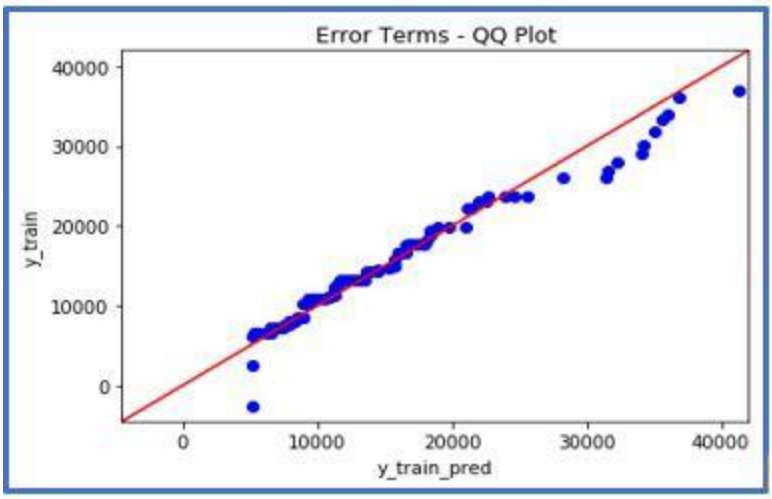
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

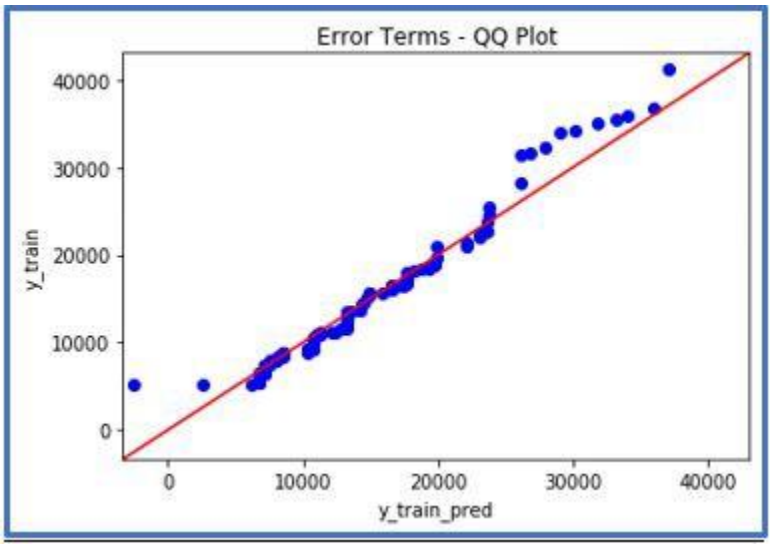
Below are the possible interpretations for two data sets.

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*