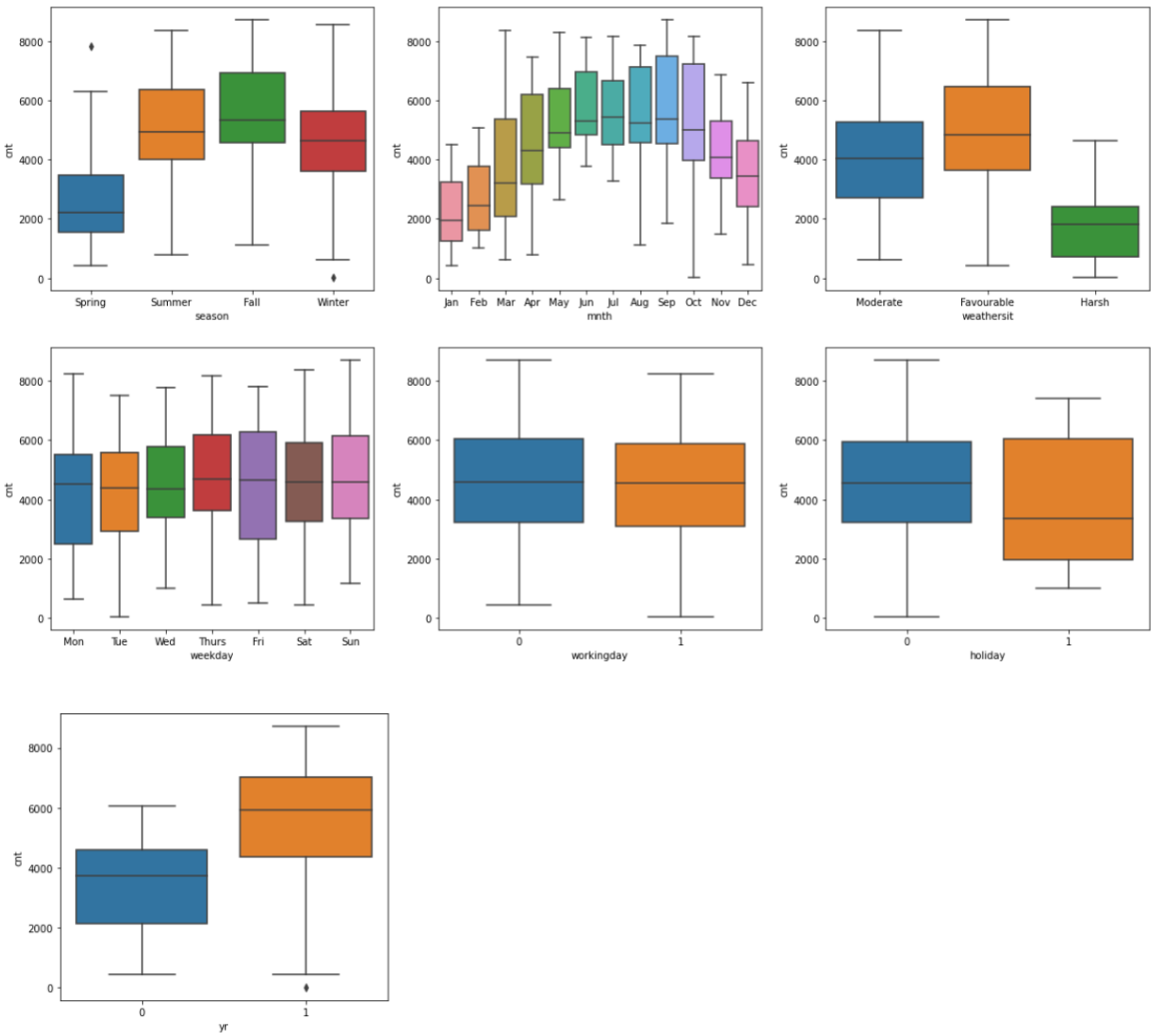
Bike Sharing Assignment

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.



All the categorical variables were plotted against the target variable to investigate the effect. Following insights have been found:

1. The count seems to have increased from Spring to Summer. Another significant increase can be seen from Summer to Fall. However, when it comes to Winter the count has decreased.
2. The months also follow a similar trend with a more granular level of information. But the trend remains the same, increasing at the start of the year and then dipping by the end.
3. According to the weather conditions, moderate and favorable weather conditions seem to bring the best business. Harsh weather conditions see a dip.
4. Surprisingly, the weekdays do not seem to have any difference in the business done, when it comes to the median part. Business has had a lower range some days but the overall picture looks to be a uniform one.

5. Business has also been the same irrespective of the fact if it is a working day or not.
6. However, on holidays we have seen a lower count and median value which makes sense. People like to have some time off.
7. Another interesting insight is, the business has seen an increase on the year level. The upper tail of 2018 corresponds to the lower tail of 2019 which means the business has significantly attracted more people over the course of time.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. Dummy variables are used to represent categorical variables in the process of model building. The function pd.get_dummies() has a very important parameter called drop_first. It is an important practice to set drop_first = True. Let's say we have n possible values for a categorical variable, of course the categorical variable can be represented by n dummy variables but if we dive deeper into the understanding of this concept, the same amount of information can be represented by n-1 dummy variables as well. For example, let's say Marital_Status is a categorical variable with three possible values, Single, Married and Divorced.

Representing Single with 3 dummy variables:

Index	Marital_Status_Single	Marital_Status_Married	Marital_Status_Divorced
1	1	0	0

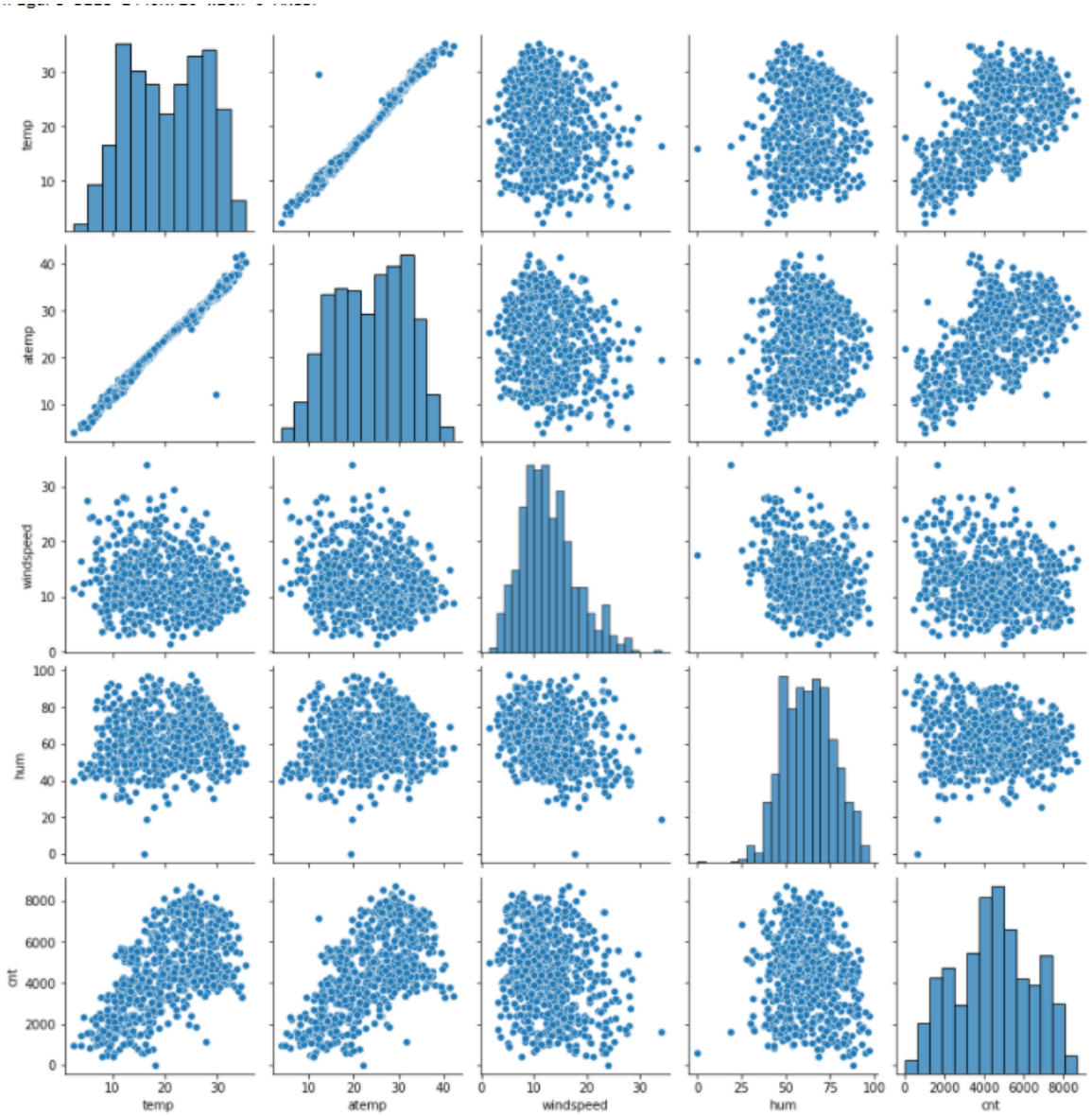
Representing Single with 2 dummy variables:

Index	Marital_Status_Married	Marital_Status_Divorced
1	0	0

Also, setting drop_first = True helps in reducing the correlations created among the dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans.



Looking at the pairplot, it seems that variables temp and atemp have a high correlation with the target variable cnt. Also, temp and atemp are highly correlated with each other(almost 1), hence we can use any one of these and drop the other.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

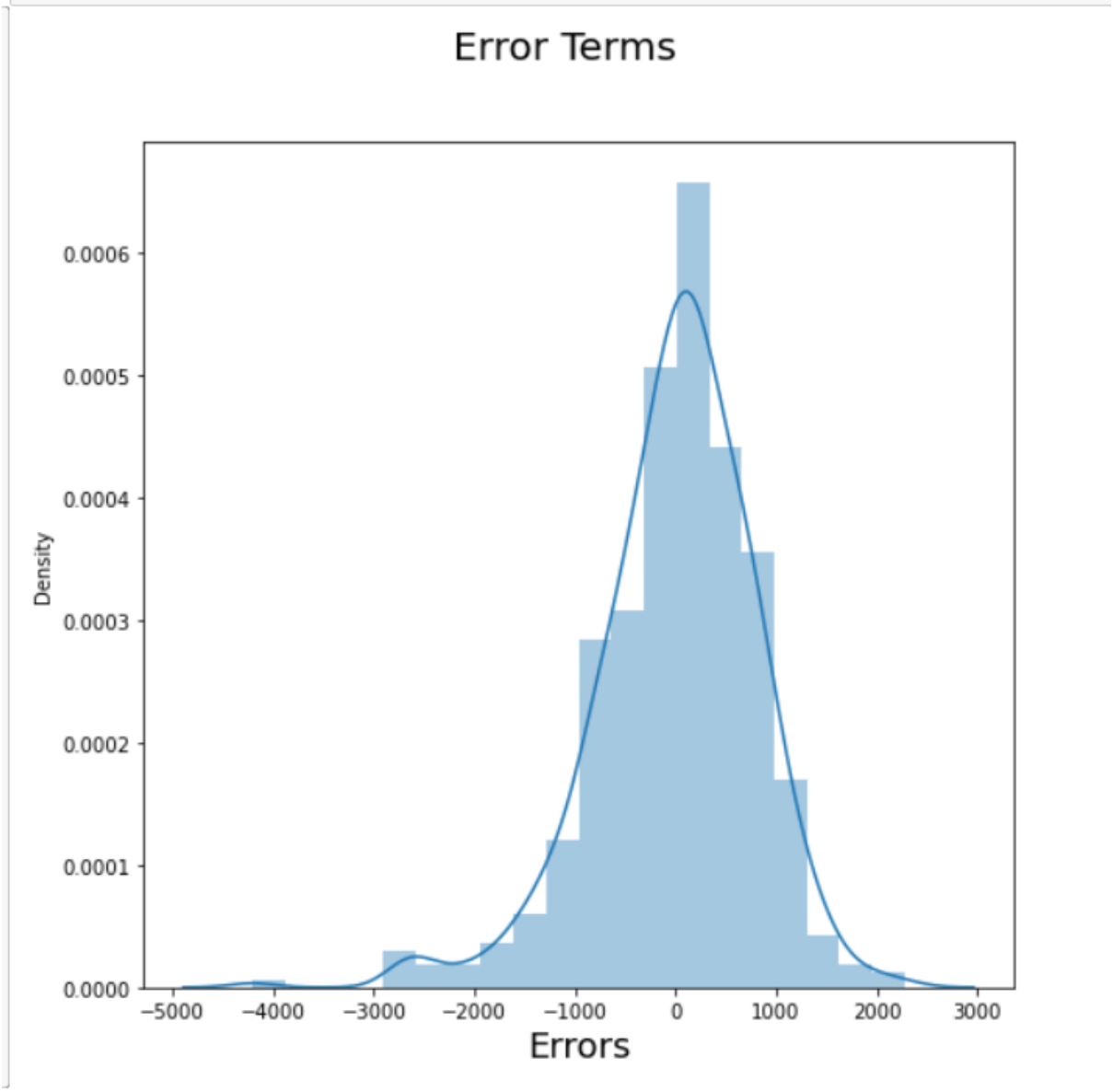
Ans. The assumptions of Linear Regression are as follows:

1. There is a linear relationship between X and Y.

Validation: It is quite evident from the pairplot that the variables have a linear relationship (not perfect) with the target variable.

2. Error terms are normally distributed with mean zero.

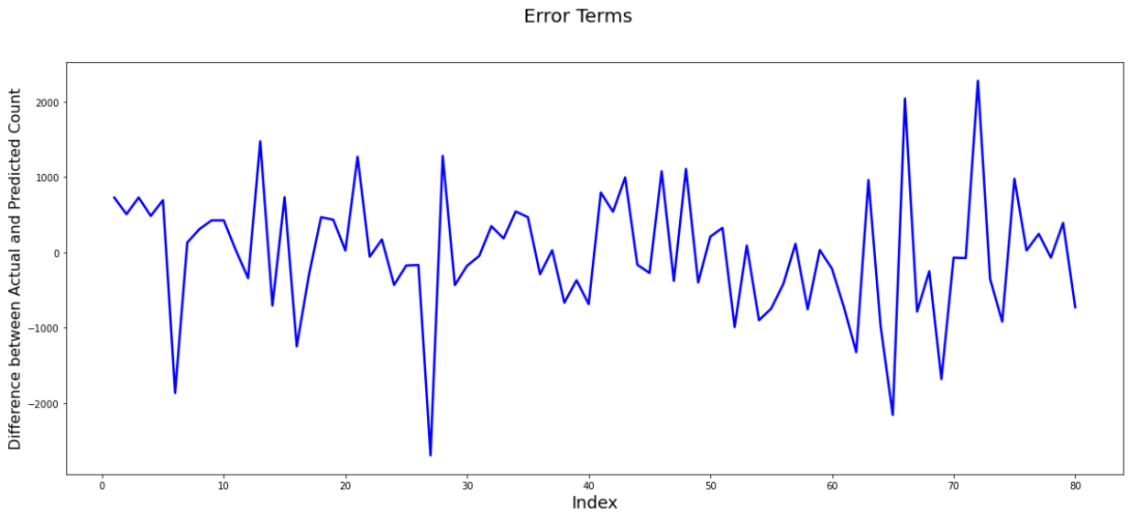
   Validation:



   It is quite evident that the error terms follow a normal distribution with zero mean.

3. Error terms are independent of each other.

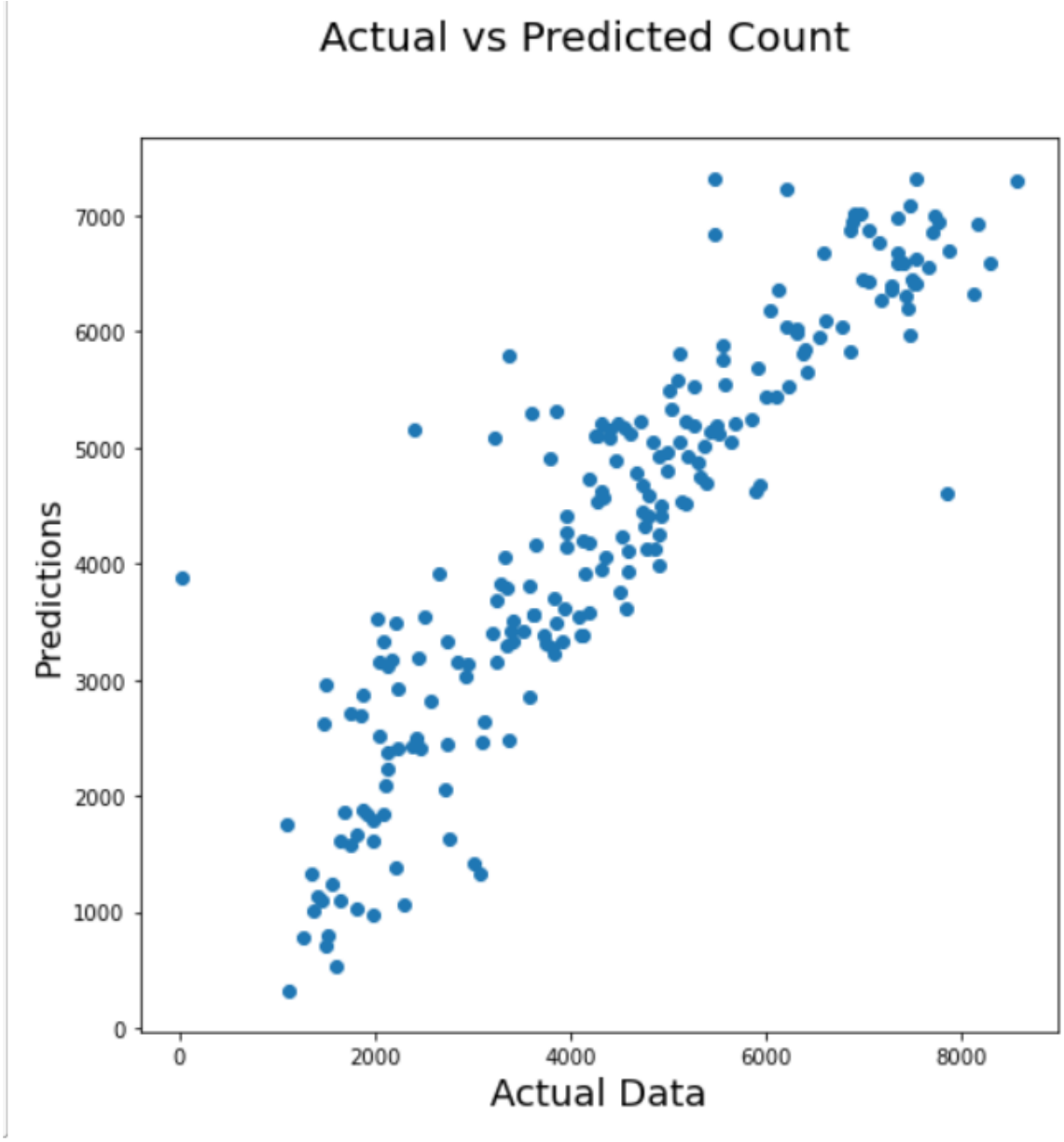
Validation:



As you can see, there's no repeated pattern that can be found in the error terms. It suggests that they are independent of each other.

4. Error terms have constant variance (homoscedasticity).

Validation:



The variance of error terms does not seem to follow any pattern or show any kind of change. It seems to be following a constant trajectory which confirms homoscedasticity.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.831
Model:	OLS	Adj. R-squared:	0.827
Method:	Least Squares	F-statistic:	204.3
Date:	Tue, 08 Feb 2022	Prob (F-statistic):	1.51e-183
Time:	00:16:10	Log-Likelihood:	-4138.2
No. Observations:	511	AIC:	8302.
Df Residuals:	498	BIC:	8357.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2641.7527	232.649	11.355	0.000	2184.658	3098.847
yr	2000.0848	71.805	27.854	0.000	1859.006	2141.163
holiday	-623.6019	220.235	-2.832	0.005	-1056.306	-190.898
temp	3062.4606	269.862	11.348	0.000	2532.251	3592.670
windspeed	-771.1493	202.048	-3.817	0.000	-1168.120	-374.179
season_Spring	-1394.2980	142.564	-9.780	0.000	-1674.399	-1114.197
season_Winter	726.5275	121.394	5.985	0.000	488.021	965.034
mnth_Dec	-575.5195	153.703	-3.744	0.000	-877.505	-273.534
mnth_Mar	476.3625	154.339	3.086	0.002	173.127	779.598
mnth_Nov	-712.6947	161.703	-4.407	0.000	-1030.400	-394.990
mnth_Sep	480.6005	129.339	3.716	0.000	226.483	734.718
weathersit_Harsh	-2286.3359	218.700	-10.454	0.000	-2716.024	-1856.648
weathersit_Moderate	-685.9703	76.789	-8.933	0.000	-836.840	-535.100

Omnibus:	78.024	Durbin-Watson:	2.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	167.173
Skew:	-0.832	Prob(JB):	5.00e-37
Kurtosis:	5.255	Cond. No.	15.3

According to the best model, the top three features contributing significantly towards explaining the demand of shared bikes are:

1. Temperature
2. Weathersit_Harsh
3. Year

Each selected feature is viable and makes a lot of sense.

Temperature decides how warm or cold a particular day will be and does affect people's choice of mode of transport that they are likely to use for their commute.

If the weather is harsh (Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog

), it is very unlikely that someone would rent a bike for their commute. People will generally stay where they are or use safer modes of transport when the weather is harsh.

Year is also a significant predictor. Year basically can be thought of as the age of a business. Conventionally, businesses that have been in the game for longer are expected to have better revenues as compared to the new ones. Moreover, it is expected that if things fall in place, business grows over the course of time.

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a

two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

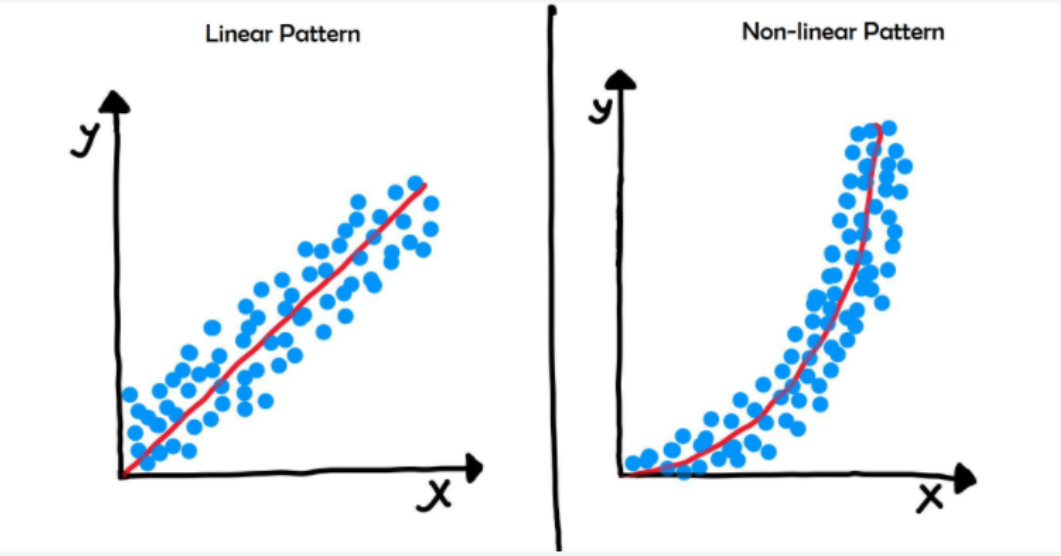
In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

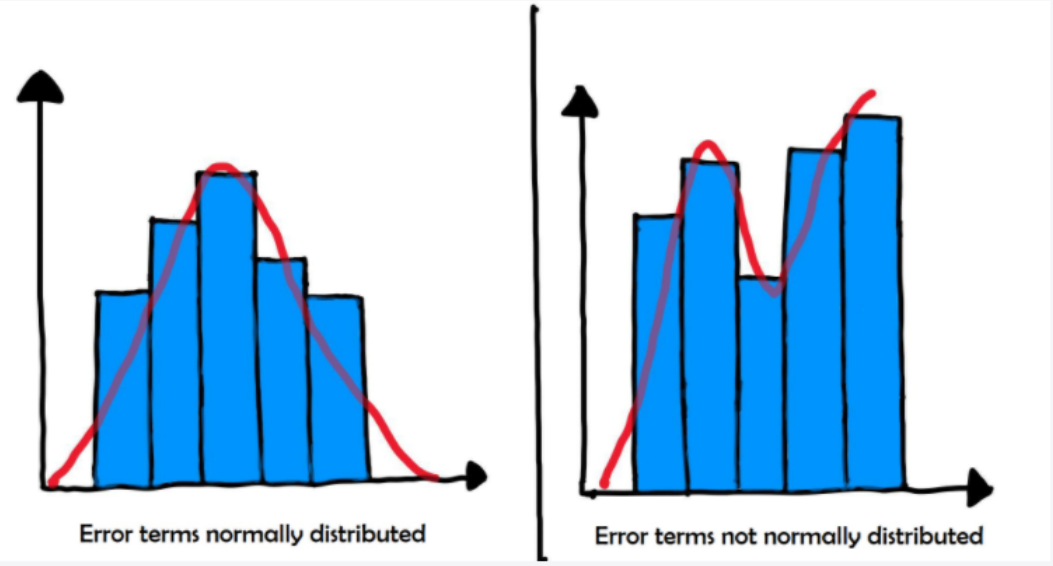
When it comes to linear regression, we also need to consider certain assumptions:

1. There is a linear relationship between X and Y. X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.



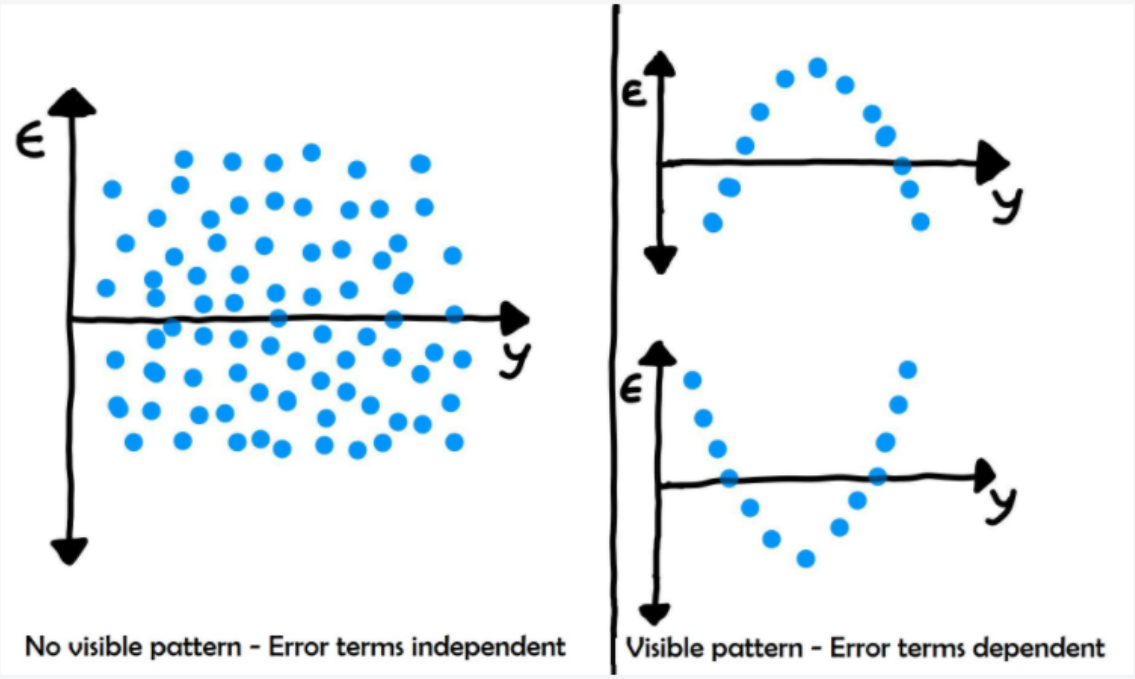
2. Error terms are normally distributed with mean zero(not X, Y):

- There is no problem if the error terms are not normally distributed if you just wish to fit a line and not make any further interpretations.
- But if you are willing to make some inferences on the model that you have built (you will see this in the coming segments), you need to have a notion of the distribution of the error terms. One particular repercussion of the error terms not being normally distributed is that the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable. (You'll see this in a later segment)
- The assumption of normality is made, as it has been observed that the error terms generally follow a normal distribution with mean equal to zero in most cases.



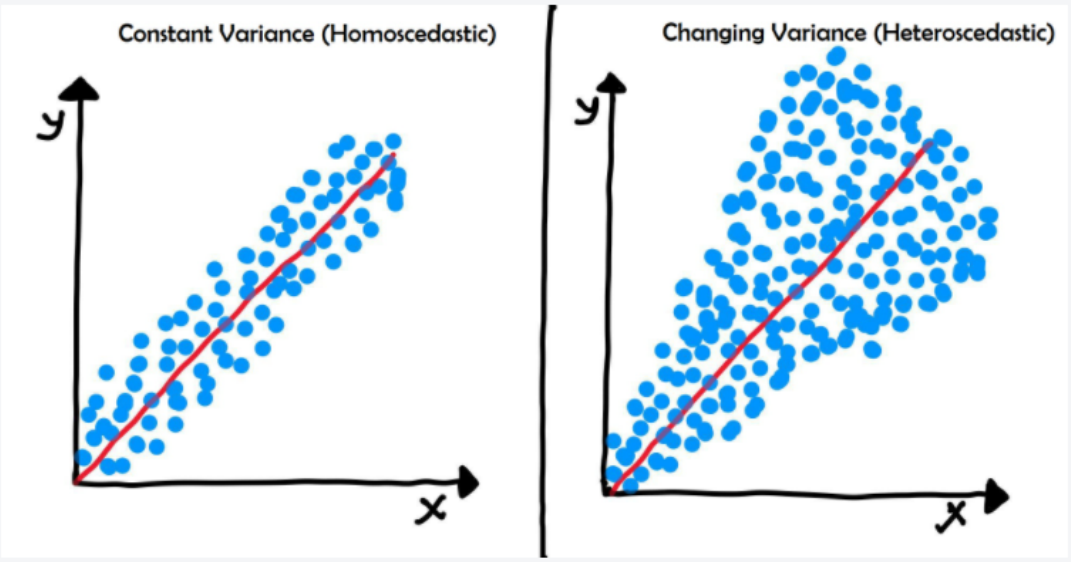
3. Error terms are independent of each other:

The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one).



4. Error terms have constant variance (homoscedasticity):

- The variance should not increase (or decrease) as the error values change.
- Also, the variance should not follow any pattern as the error terms change.



Q2. Explain the Anscombe's quartet in detail.

Ans. **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.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.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.

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests

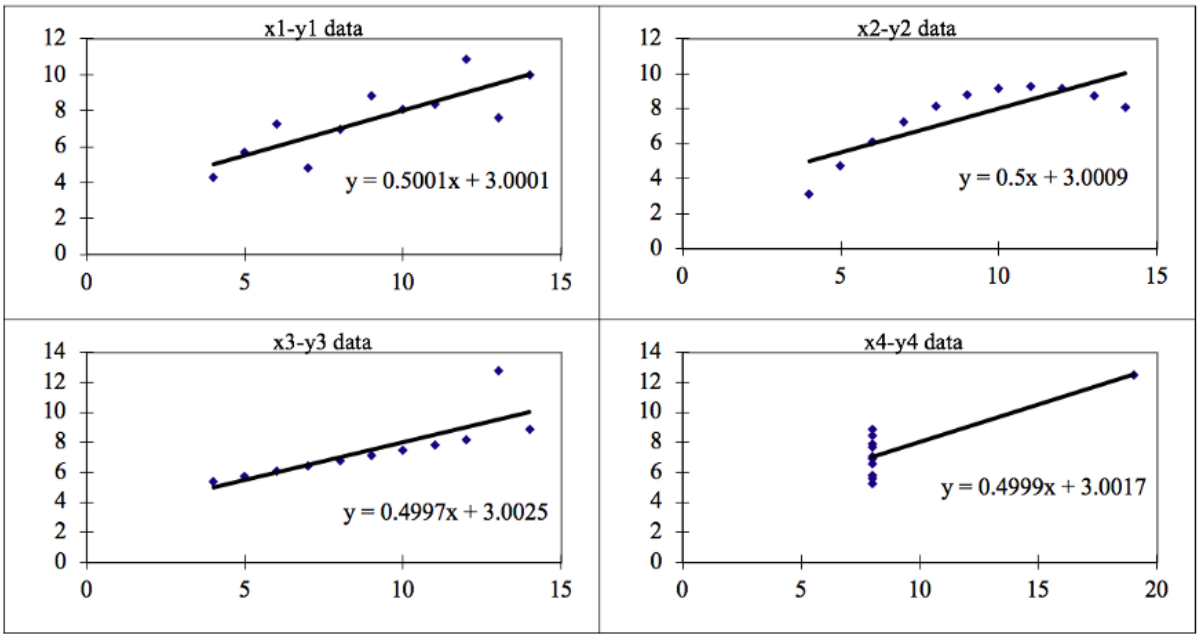
that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined 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 all these four datasets are approximately similar and can be computed 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	

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 four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression mode

Conclusion:

It can be very well concluded that the four datasets that were intentionally created to describe the importance of data visualization and

how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

Q3. What is Pearson's R?

Ans. Before understanding what Pearson's R is, it is important to know the underlying concept of correlation.

Correlation is a statistic that measures the relationship between two variables in the finance and investment industries. It shows the strength of the relationship between the two variables as well as the direction and is represented numerically by the correlation coefficient. The numerical values of the correlation coefficient lies between -1.0 and +1.0.

A negative value of the correlation coefficient means that when there is a change in one variable, the other changes in a proportion but in the opposite direction, and if the value of the correlation coefficient is positive, both the variables change in a proportion and the same direction.

When the value of the correlation coefficient is exactly 1.0, it is said to be a perfect positive correlation. This situation means that when there is a

change in one variable, either negative or positive, the second variable changes in lockstep, in the same direction.

A perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no linear relationship at all. We can determine the strength of the relationship between two variables by finding the absolute value of the correlation coefficient.

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate 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.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

How is the Correlation coefficient calculated?

Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. 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

The formula given is:

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

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

Determining the strength of the Pearson product-moment correlation coefficient

The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.

	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

A notable point is that the strength of association of the variables depends on the sample size and what you measure.

What do the terms strength and direction mean in Statistics?

We have been mentioning the two terms ‘strength’ and ‘direction’, throughout the blog. These terms have a great statistical significance. Let us discuss them in detail.

Strength: Strength implies the relationship connection between the two given factors. It implies how reliably one variable will change because of the adjustment in the other. Qualities that are near +1 or - 1 show a solid relationship. These qualities are achieved if the information focuses fall on or near the line. The further the information focuses move away, the more vulnerable the strength of the direct relationship. When there is no useful method to draw a straight line because the information focuses

are dissipated, the strength of the direct relationship is the most vulnerable.

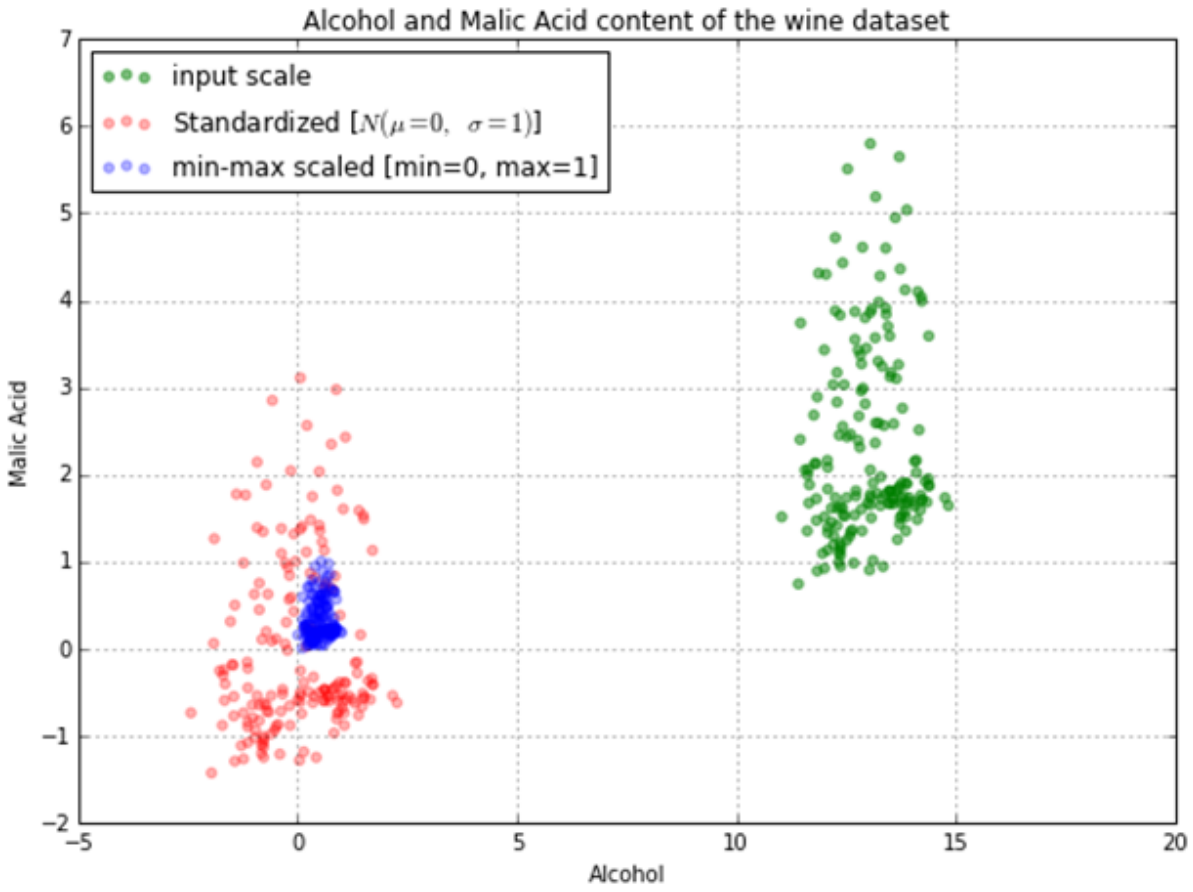
Direction: The direction of the line demonstrates a positive direct or negative straight connection between factors. On the off chance that the line has an upward slant, the factors have a positive relationship. This implies an expansion in the estimation of one variable will prompt an increment in the estimation of the other variable. A negative relationship portrays a descending slant. This implies an expansion in the measure of one variable prompts a lessening in the estimation of another variable.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example-

centered around 0 or in the range (0,1) depending on the scaling technique.

In order to visualize the above, let us take an example of the independent variables of alcohol and Malic Acid content in the wine dataset from the "Wine Dataset" that is deposited on the UCI machine learning repository. Below you can see the impact of the two most common scaling techniques (Normalization and Standardization) on the dataset.



Scaling is performed to bring all the parameters on a common and comparable scale which is very important for the following reasons:

Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. Take a look at the formula for gradient descent below:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

Distance-Based Algorithms

Distance algorithms like KNN, K-means, and SVM are most affected by the range of features. This is because behind the scenes they are using distances between data points to determine their similarity.

For example, let's say we have data containing high school CGPA scores of students (ranging from 0 to 5) and their future incomes (in thousands Rupees):

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

Since both the features have different scales, there is a chance that higher weightage is given to features with higher magnitude. This will impact the performance of the machine learning algorithm and obviously, we do not want our algorithm to be biased towards one feature.

Therefore, we scale our data before employing a distance based algorithm so that all the features contribute equally to the result.

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

The effect of scaling is conspicuous when we compare the Euclidean distance between data points for students A and B, and between B and C, before and after scaling as shown below:

- Distance AB before scaling => $\sqrt{(40-60)^2+(3-3)^2}=20$
- Distance BC before scaling => $\sqrt{(40-40)^2+(4-3)^2}=1$
- Distance AB after scaling => $\sqrt{(1.1+1.5)^2+(1.18-1.18)^2}=2.6$
- Distance BC after scaling => $\sqrt{(1.1-1.1)^2+(0.41+1.18)^2}=1.59$

Scaling has brought both the features into the picture and the distances are now more comparable than they were before we applied scaling.

Tree-Based Algorithms

Tree-based algorithms, on the other hand, are fairly insensitive to the scale of the features. Think about it, a decision tree is only splitting a node based on a single feature. The decision tree splits a node on a feature that increases the homogeneity of the node. This split on a feature is not influenced by other features.

So, there is virtually no effect of the remaining features on the split. This is what makes them invariant to the scale of the features!

Normalization and Standardization are explained below:

Normalization

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

We can also do a normalization over different intervals, e.g. choosing to have the variable lying in any [a, b] interval, a and b being real numbers. To rescale a range between an arbitrary set of values [a, b], the formula becomes:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

Standardization

Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. 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:

$$X_1=C+ \alpha_2 X_2+\alpha_3 X_3+\cdots$$

$$〖VIF〗_1=1/(1-R_1^2)$$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+\cdots$$

$$〖VIF〗_2=1/(1-R_2^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. 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 the 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.

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).

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. 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 the same distributions.

Few advantages:

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.

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

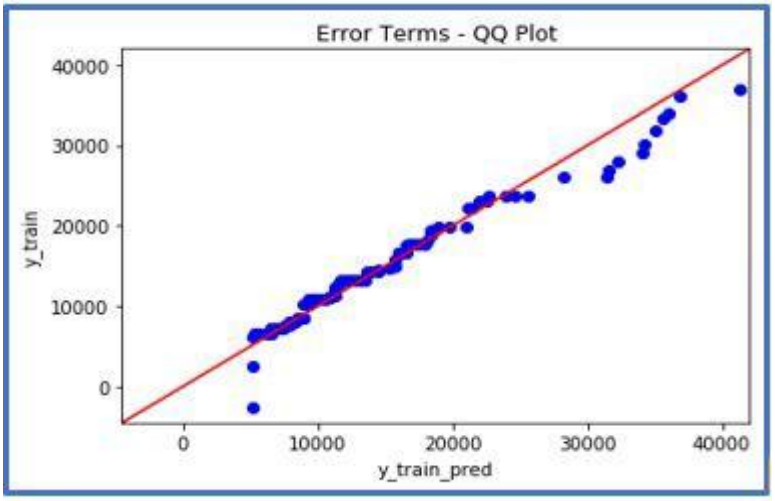
Interpretation:

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

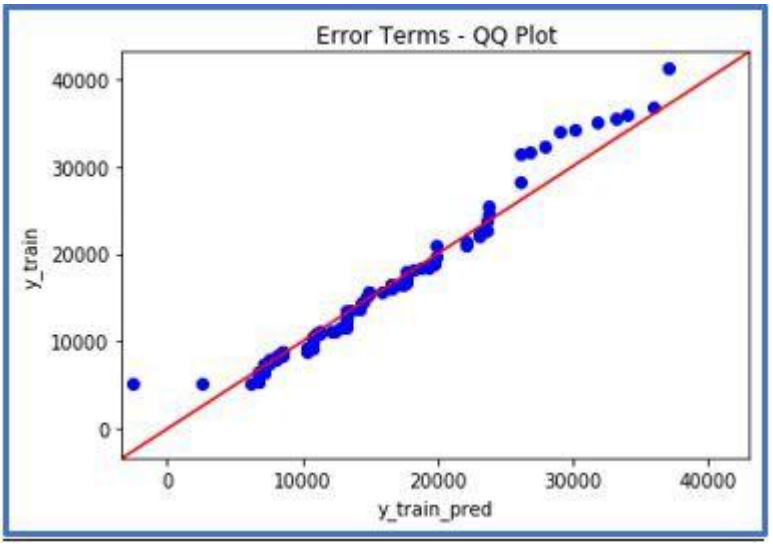
Below are the possible interpretations for two data sets.

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