General Subjective Questions

1.Explain the linear regression algorithm in detail.

*Ans-*Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used.

Linear regression is one of the most commonly used ML algorithms for predictive analysis.

The overall idea of regression is to examine two things:

o Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

o Which variables, in particular, are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

impact the outcome variable? These regression estimates are in turn used to explain the relationship between one dependent variable with one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = b0 +

b1*x, where y = estimated dependent variable score, b0 = constant, b1 = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are:

o determining the strength of predictors

o forecast an effect, and

o trend forecasting

. First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of the relationship between dose and effect, sales and marketing spending, or age and income

Second, it can be used to forecast the effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, “what will the price of gold be in 6 months?”

There are several types of linear regression analyses available to researchers.

Simple linear regression

1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression

1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

Logistic regression

1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression

1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

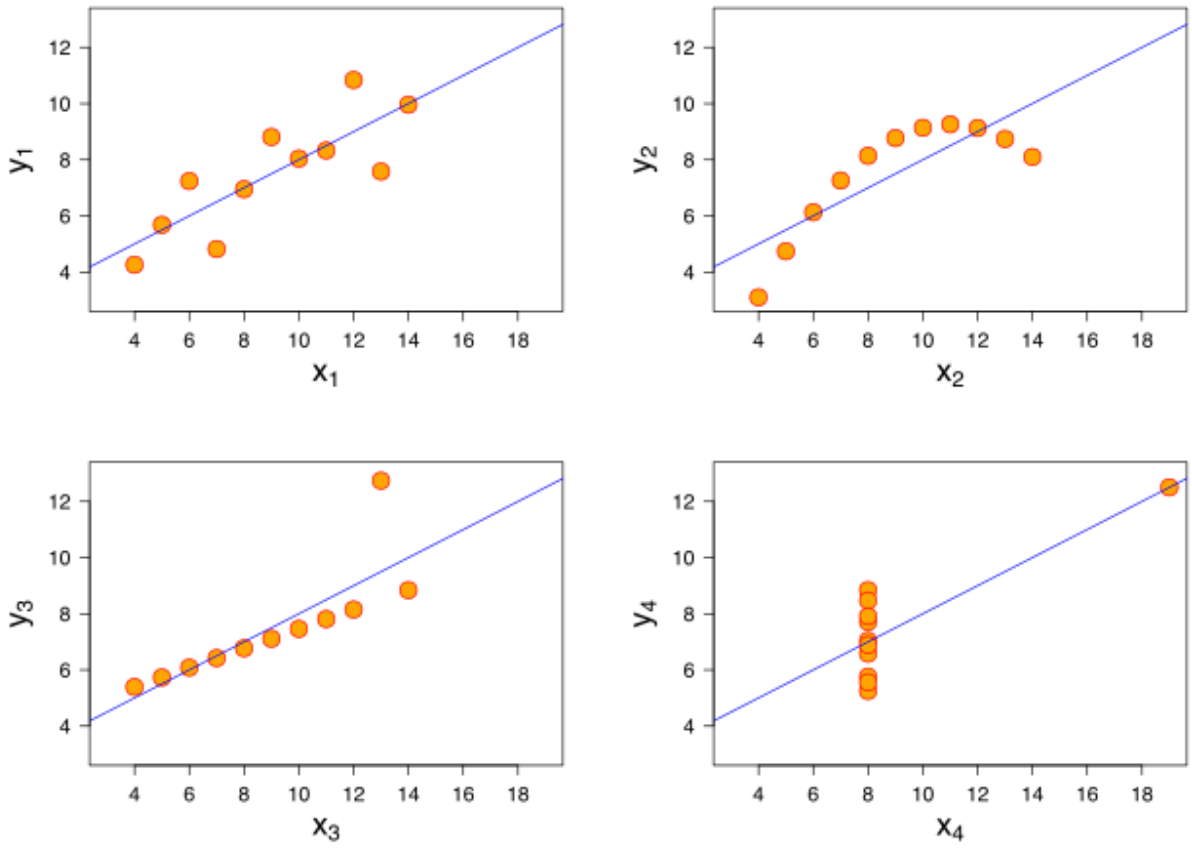
Discriminant analysis

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

2.Explain the Anscombe's quartet in detail.

Ans- **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



The first scatter plot (top left) appears to be a simple linear regression, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson's Correlation Coefficient is not relevant. In the third graph (bottom left), the distribution is linear, but with a different regression line,

which is offset by the one outlier, which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

For all four datasets:

Property	Value
Mean of x in each case:	9 (exact)
Variance of x in each case:	11 (exact)
Mean of y in each case:	7.50 (to 2 decimal places)
Variance of y in each case:	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case:	0.816 (to 3 decimal places)
Linear regression line in each case:	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The x values are the same for the first three datasets

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

3.What is Pearson's R?

*Ans-*It is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

In a sample it is denoted by r and is by design constrained as follows:

Furthermore,

1. Positive values denote positive linear correlation.

2. Negative values denote negative linear correlation.

3. 0 denotes no linear correlation.

4. The closer the value is to 1 or -1, the stronger the linear correlation. It Is given by,

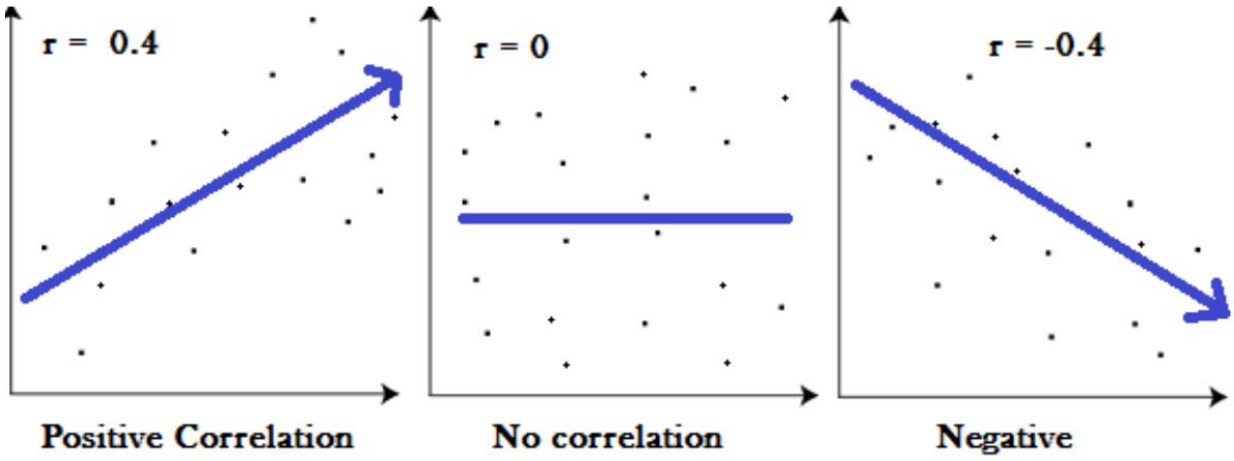
There are several types of correlation coefficient formulas.

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula. If you're taking a basic stats class, this is the one you'll probably use:

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - [\mathrm{E}[X]]^2}\,\sqrt{\mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2}}.$$

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



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

*Ans-*Scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In scaling (also called min-max scaling), you transform the data such that the features are within a specific range e.g. [0, 1]

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x' is the normalized value.

It is mainly performed because most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem. If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To supress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Normalized Scaling:

Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

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

where x is an original value, x' is the normalized value. For example, suppose that we have the students' weight data, and the students' weights span [160 pounds, 200 pounds]. To rescale this data, we first subtract 160 from each student's weight and divide the result by 40 (the difference between the maximum and minimum weights).

To rescale a range between an arbitrary set of values [a, b], the formula becomes:

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

where a, and b are the min-max values.

Mean normalization

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

where x is an original value, x' is the normalized value. There is another form of the mean normalization which is when we divide by the standard deviation which is also called standardization.

Standardized Scaling:

In machine learning, we can handle various types of data, e.g. audio signals and pixel values for image data, and this data can include multiple dimensions. Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and artificial neural networks). The general method of calculation is to determine the distribution mean and standard deviation for each feature. Next we subtract the mean from each feature. Then we divide the values

(mean is already subtracted) of each feature by its standard deviation.

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

Where x is the original feature vector, $\bar{x}$ = average (x) is the mean of that feature vector, and σ is its standard deviation. Scaling to unit length.

Another option that is widely used in machine-learning is to scale the components of a feature vector such that the complete vector has length one. This usually means dividing each component by the Euclidean length of the vector:

$$x' = \frac{x}{\|x\|}$$

In some applications (e.g. Histogram features) it can be more practical to use the L1 norm (i.e. Manhattan Distance, City-Block Length or Taxicab Geometry) of the feature vector. This is especially important if in the following learning steps the Scalar Metric is used as a distance measure.

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

*Ans-*The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the R-squared statistic of the regression where the predictor of interest is predicted by all the other predictor variables. The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

When R-squared reaches 1, VIF reaches infinity. When R-squared reaches 1 then it means multicollinearity exists. Different variables are highly correlated with each other.

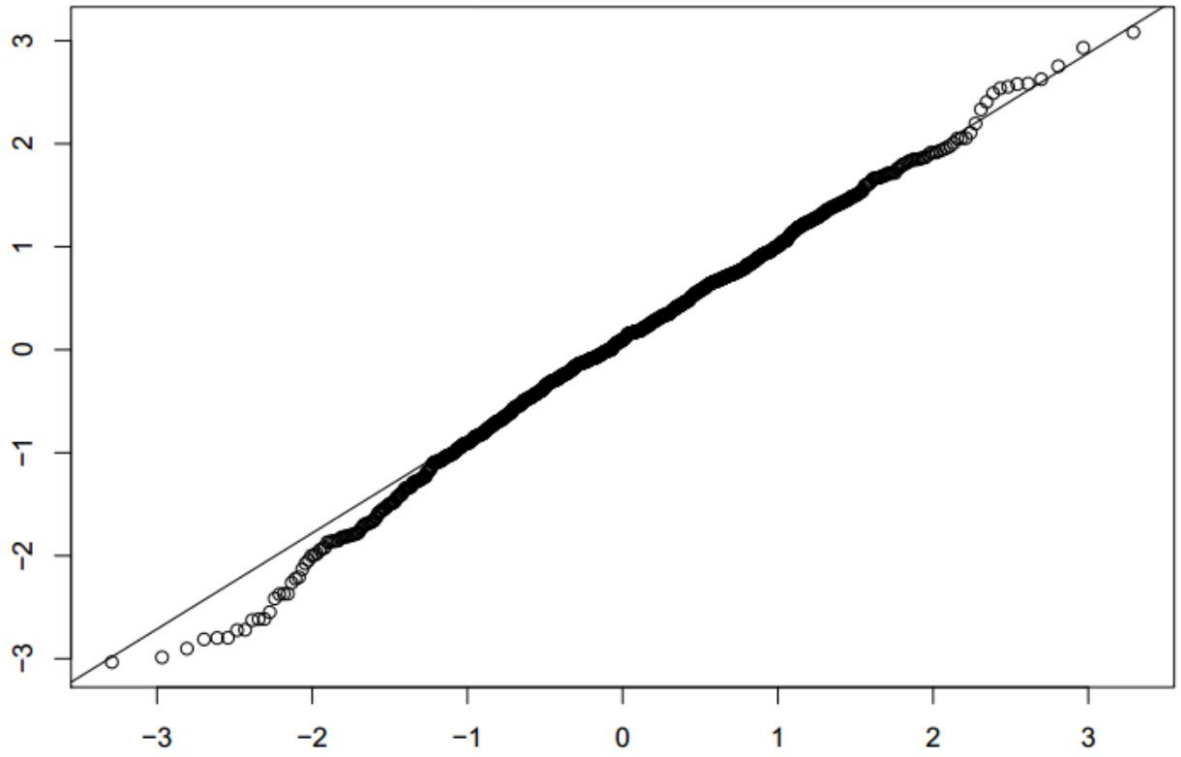
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

*Ans-*The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

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.

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.



Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.

QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small

sample sizes, you can't assume your estimator is Gaussian either, so the standard confidence intervals and significance tests are invalid.

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- a) COUNT OF BIKE RENTALS INCREASED AND BECAME POPULAR IN YEAR 2019 THAN 2018

b)COUNT OF BIKE RENTALS IS MORE DURING CLEAR WEATHER

c)FALL AND SUMMER ARE MORE FAVOURABLE FOR BIKE RENTALS THAN SPRING

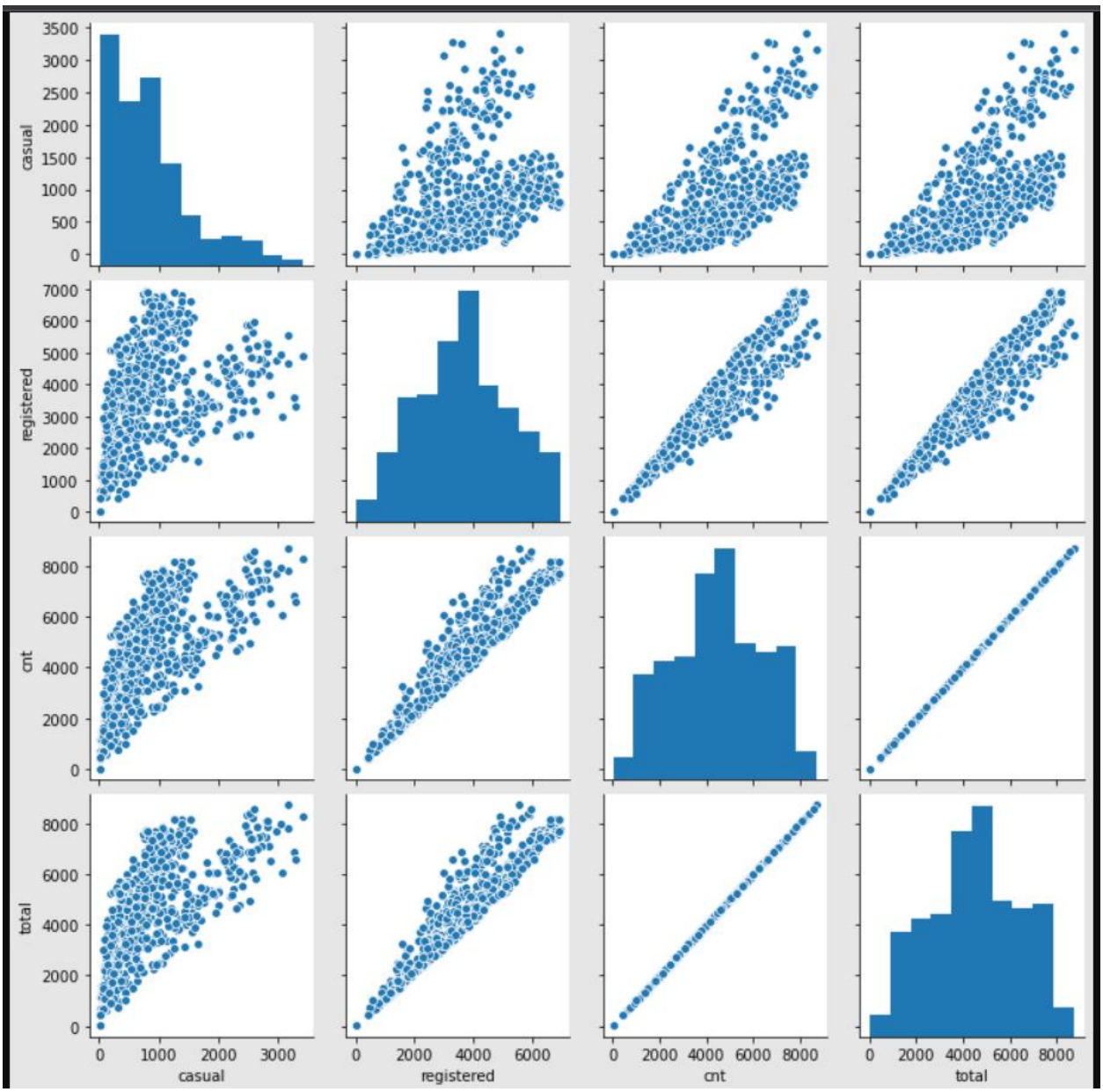
2. Why is it important to use drop_first=True during dummy variable creation?

Ans- a) TO AVOID MULTICOLLINEARITY (IF WE DON'T DROP ,DUMMY VARIABLES WILL BE CORRELATED) AND AFFECTS THE MODEL ADVERSELY

b)TO AVOID REDUNDANT FEATURES

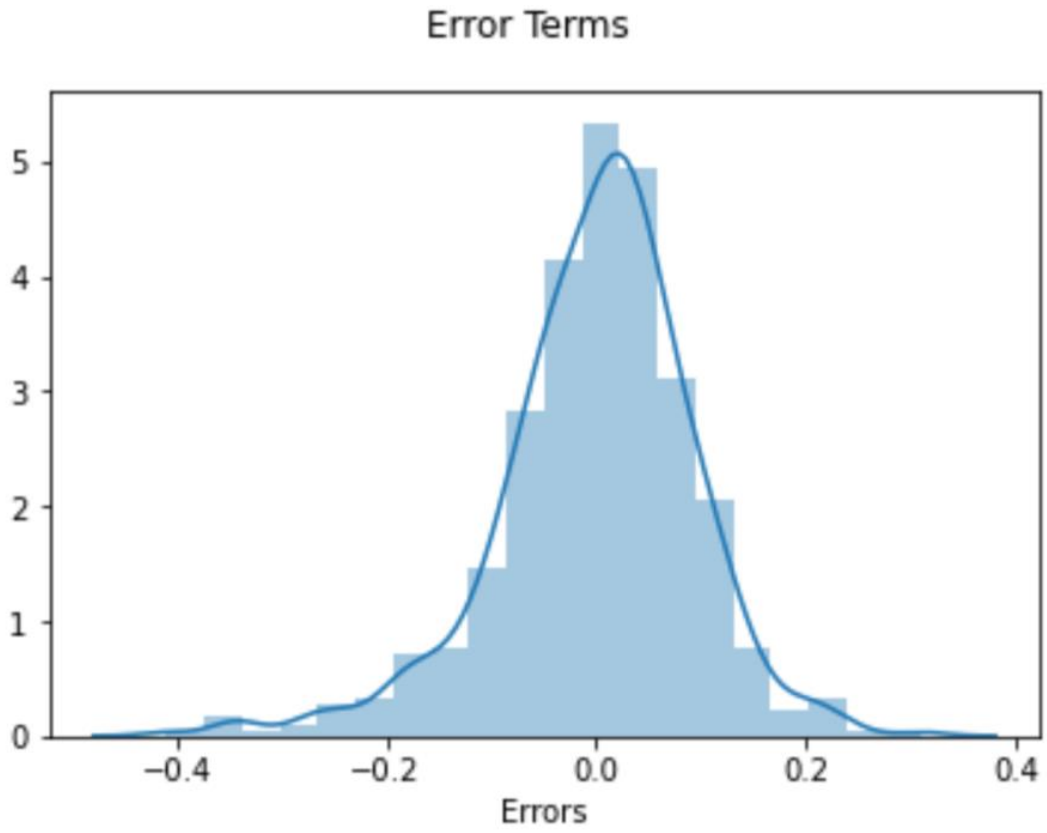
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans-COUNT (TARGET VARIABLE)HAS SIGNIFICANTLY HIGH CORRELATION WITH TEMPERATURE (TEMP)

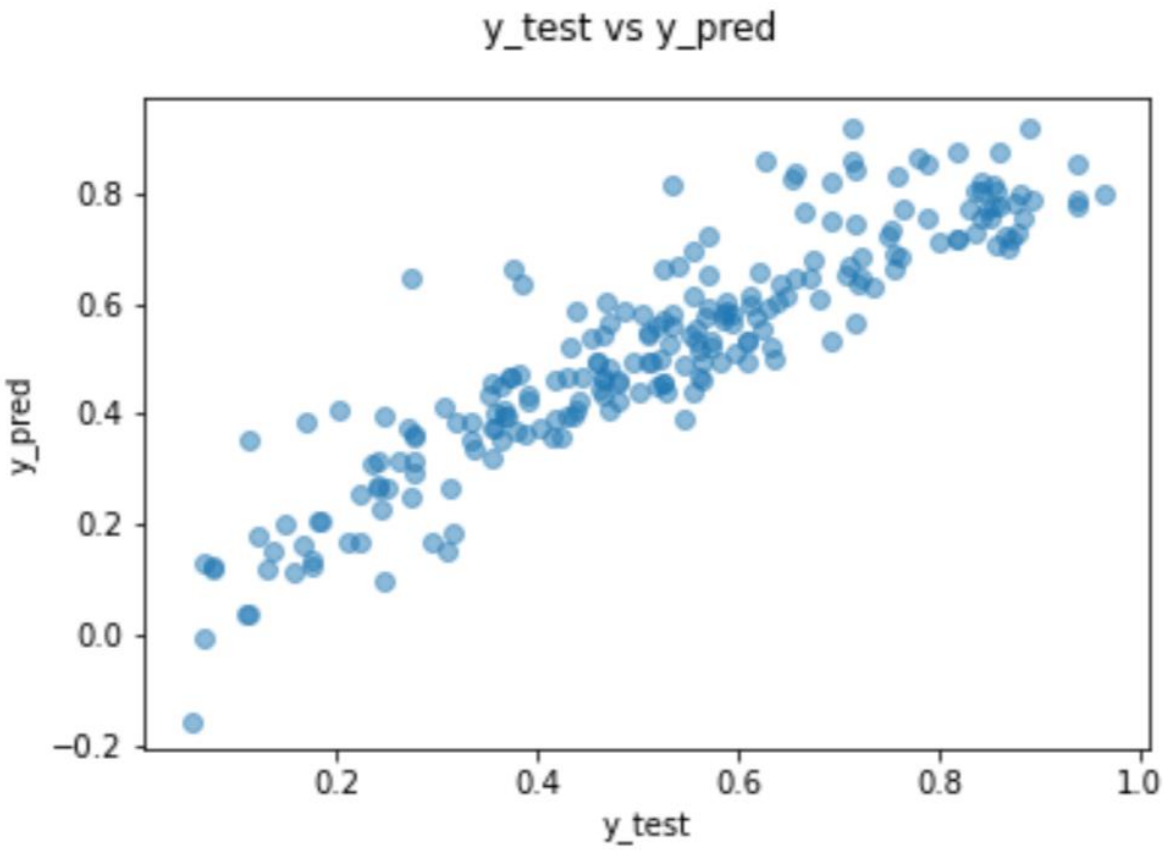


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans- a) RESIDUAL ERRORS FOLLOW NORMAL DISTRIBUTION



b)MAINTAINS LINEAR RELATION BETWEEN DEPENDANT VARIABLE



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

Ans-a) TEMPERATURE(0.4280)

b)WEATHER SITUATION LIGHT AND SNOWY(-0.3063)

c) YEAR(0.2345)