General Subjective Questions:

1.Explain the linear regression algorithm in detail. ?

Answer: It is one of the machine learning techniques that fall under supervised learning.

1. Autocorrelation:

This assumption made by linear regression indicates little to no autocorrelation in data. Autocorrelation takes place when residual errors are dependent on each other in one or the other way.

2. Multi-collinearity:

This assumption says that data multi-collinearity either doesn't exist at all or is present scarcely. Multi-collinearity happens when independent features or variables show some dependency.

3. Variable relationship:

The model has an assumption that there is a linear relationship between feature and response variables.

A few instances where you can use linear regression include the estimation of the price of a house depending on the number of rooms it has, determining how well a plant will grow depending on how frequently it is watered, and so on. For all these instances, you would already have an idea about the type of relationship that exists between different variables.

When you use linear regression analysis, you back your idea or hypothesis with data. When you develop a better understanding of the relationship between different variables, you are in a better position to make powerful predictions. If you don't already know, let us tell you that linear regression is a supervised machine learning technique as well as a statistical model.

In machine learning terms, the regression model is your machine, and learning relates to this model being trained on a data set, which helps it learn the relationship between variables and enables it to make data-backed predictions.

How does linear regression work?
Before we run the analysis, let us assume that we have two types of teams – those that perform their jobs well and those that don't. There are could several reasons why a team isn't good at what it is doing. It could be because it doesn't have the right skill set or it doesn't have the experience required to perform certain duties at work. But, you can never be certain of what it is.

We can use linear regression to find out candidates that have all that's required to be the best fit for a particular team that is involved in a particular line of work. This will help us in selecting candidates that are highly likely to be good at their jobs.

The objective that regression analysis serves is creating a trend curve or line that is suitable for the data in question. This helps us in finding out how one parameter (independent variables) is related to the other parameter (dependent variables).

Before anything else, we need to first have a closer look at all the attributes of different candidates and find out whether they are correlated in some way or the other. If we find some correlations, we can go ahead start making predictions based on these attributes.

Relationship exploration in the data is done by using a trend curve or line and plotting the data. The curve or line will show us if there is any correlation. We can now use linear regression to refute or accept relationships. When the relationship is confirmed, we can use the regression algorithm to learn his relationship. This will enable us to make the right predictions. We will be able to more accurately predict whether a candidate is right for the job or not.

Importance of training a model
The process involved in training a linear regression model is similar in many ways to how other machine learning models are trained. We need to work on a training data set and model the relationship of its variables in a way that doesn't impact the ability of the model to predict new data samples. Model is trained to improve your prediction equation continuously.

It is done by iteratively looping through the given dataset. Every time you repeat this action, you simultaneously update the bias and weight value in the direction that the gradient or cost function indicates. The stage of the completion of training is reached when an error threshold is touched or when there is no reduction in cost with the training iterations that follow.

Before we start training the model, there are a few things that we need to prepare. We need to set the number of iteration required as well as the rate of learning. Apart from this, we also have to set default values for our weights. Also, record the progress that we are able to achieve with every repeat.

What is regularisation?
If we talk about the linear regression variants that are preferred over others, then we will have to mention those that have added regularisation. Regularisation involves penalizing those weights in a model that have larger absolute values than others.

Regularisation is done to limit overfitting, which is what a model often does as it reproduces the training data relationships too closely. It doesn't allow the model to generalize never seen before samples as it is supposed to.

When do we use linear regression?
The power of linear regression lies in how simple it is. It means that it can be used to find answers to almost every question. Before using a linear regression algorithm, you must ensure that your data set meets the required conditions that it works on.

The most important of these conditions is the existence of a linear relationship between the variables of your data set. This allows them to be easily plotted. You need to see the difference that exists between the predicted values and achieved value in real are constant. The predicted values should still be independent, and the correlation between predictors should be too close for comfort.

You can simply plot your data along a line and then study its structure thoroughly to see whether your data set meets the desired conditions or not.

Linear regression uses
The simplicity by which linear aggression makes interpretations at the molecular level easier is one of its biggest advantages. Linear regression can be applied to all those data sets where variables have a linear relationship.

Businesses can use the linear regression algorithm is their sales data. Suppose you are a business that is planning to launch a new product. But, you are not really sure at what price you should sell this product. You can check how your customers are responding to your product by selling it at a few well thought of price points. This will allow you to generalize the relationship between your product sales and price. With linear regression, you will be able to determine a price point that customers are more likely to accept.

Linear regression can also be used at different stages of the sourcing and production of a product. These models are widely used in academic, scientific, and medical fields. For instance, farmers can model a system that allows them to use environmental conditions to their benefit. This will help them in working with the elements in such a way that they cause the minimum damage to their crop yield and profit.

In addition to these, it can be used in healthcare, archaeology, and labour amongst other areas. is how the interpretation on a linear model

Conclusion
Regression analysis is a widely adopted tool that uses mathematics to sort out variables that can have a direct or indirect impact on the final data. It is important to keep it in mind while analysis is in play! Linear regression is one of the most common algorithms used by data scientists to establish linear relationships between the dataset's variables, and its mathematical model is necessary for predictive analysis.

2.Explain the Anscombe's quartet in detail. ?

Answer: According to the definition given in Wikipedia, 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.

Simple understanding:
Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Python program to find mean, standard deviation, and the correlation between x and y.

```
# Import the required libraries
import pandas as pd
import statistics
from scipy.stats import pearsonr

# Import the csv file
df = pd.read_csv("anscombe.csv")

# Convert pandas dataframe into pandas series
list1 = df['x1']
list2 = df['y1']

# Calculating mean for x1
print('%.1f' % statistics.mean(list1))

# Calculating standard deviation for x1
print('%.2f' % statistics.stdev(list1))

# Calculating mean for y1
print('%.1f' % statistics.mean(list2))

# Calculating standard deviation for y1
print('%.2f' % statistics.stdev(list2))

# Calculating pearson correlation
corr, _ = pearsonr(list1, list2)
print('%.3f' % corr)

# Similarly calculate for the other 3 samples
```

Output:
9.0
3.32
7.5
2.03
0.816
So let me show you the result in a tabular fashion for better understanding.

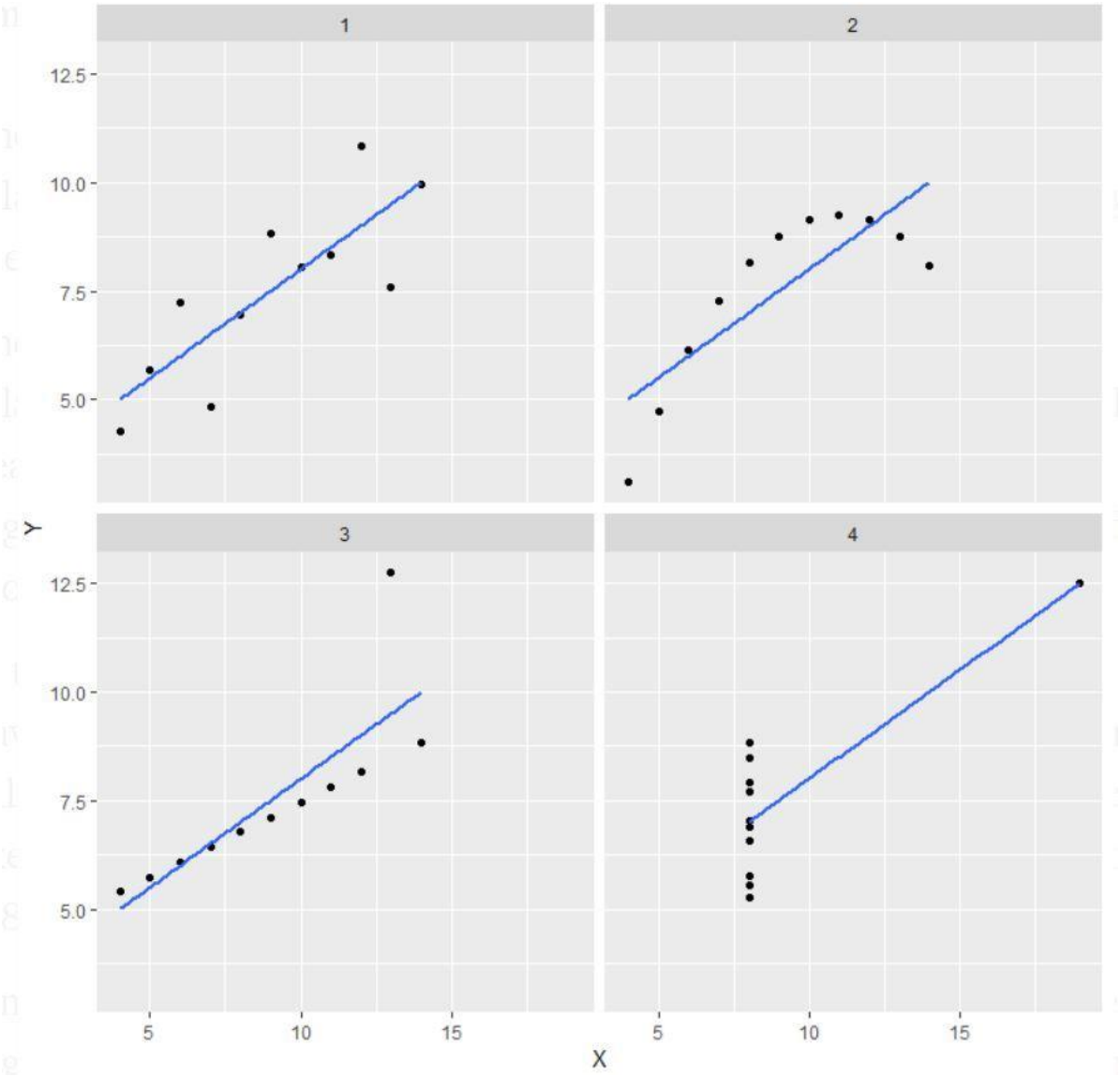
Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Python program to plot scatter plot

```
# Import the required libraries
from matplotlib import pyplot as plt
import pandas as pd

# Import the csv file
df = pd.read_csv("anscombe.csv")

# Convert pandas dataframe into pandas series
list1 = df['x1']
list2 = df['y1']

# Function to plot scatter
plt.scatter(list1, list2)

# Function to show the plot
plt.show()
```

Similarly plot scatter plot for other 3 data sets
For regression line refer this.
Output:



Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Explanation of this output:

In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:
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.

3.What is Pearson's R?

Answer: 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

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

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

Some steps are needed to be followed:

Step 1: Make a Pearson correlation coefficient table. Make a data chart using the two variables and name them as X and Y. Add three additional columns for the values of XY, X^2, and Y^2. Refer to this table.

Person	Age (X)	Income (Y)	XY	X^2	Y^2
1					
2					
3					
4					

Step 2: Use basic multiplications 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 3: 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 4: Use these values in the formula to obtain the value of r.

r = [4 * 695000 - 140 * 17000] / √{4 * 5400 - (140)^2} {4 * 92500000 - (17000)^2}

= [2780000 - 2380000] / √{21600 - 19600} {370000000 - 289000000}

= 400000 / √{2000} {81000000}

= 400000 / √162000000000

= 400000 / 402492.24

= 0.99

The positive value of Pearson's correlation coefficient implies that if we change either of these variables, there will be a positive effect on the other. For example, if we increase the age there will be an increase in the income.

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

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

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,

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/(1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

The command Tools > VIF of Descriptors... supports the detection of multicollinearities by means of the variance inflation factor (VIF). The user has to select the variables to be included by ticking off the corresponding check boxes. In general one starts with the selection of all variables, and proceeds by repeatedly deselecting variables showing a high VIF. Ideally, the VIF values should be below 10.

As the calculation of the VIF can be quite time consuming, you may choose to use only a random sample of 1000 pixels to calculate the VIF. This increases the speed of calculation considerably, even though the accuracy of the VIF values is degraded. However this should be sufficient to get a rough overview.

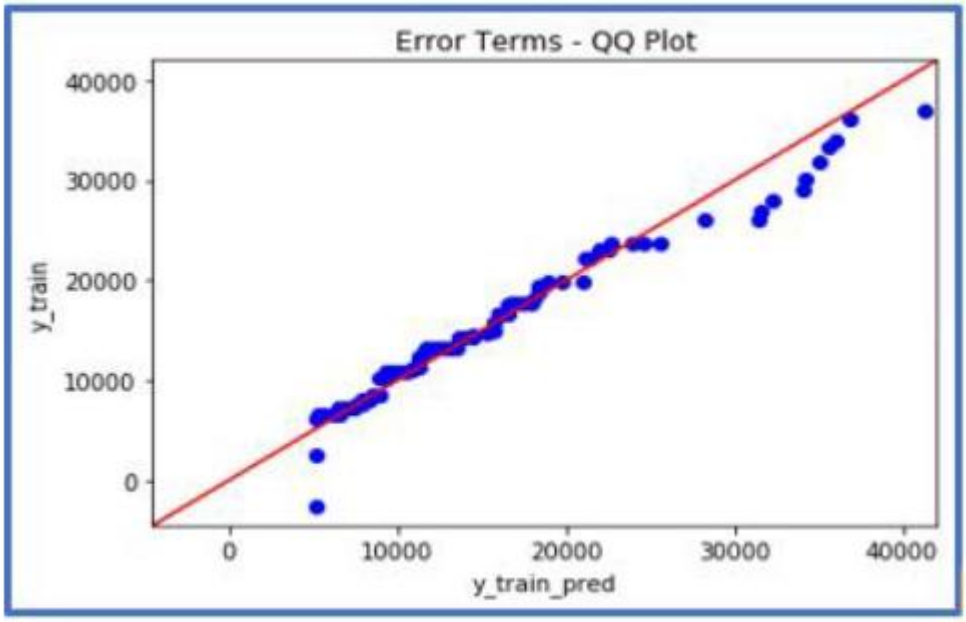
or descriptor sets with less than 25 spectral descriptors the calculation of the VIF values is performed automatically upon following any change of the selected descriptors. If a set contains a higher number of descriptors the user can decide when to calculate the VIF values by clicking the "start the calculation".

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).

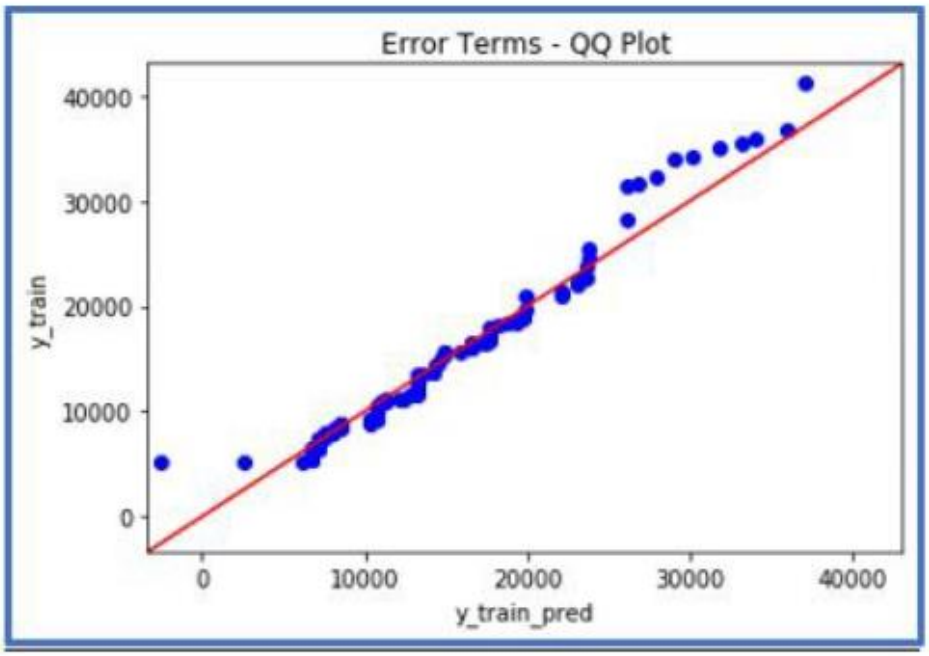
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.?

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.

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 behaviour
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

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: for Season : During Fall no of people tend to use the boom bikes more and during spring the no of users are less compared to other seasons.
For Year : 2019: Many people have preferred to use boom bikes compared to year 2018.
For the month of September many user have preferred the boom bikes.
During Holidays many users have preferred boom bikes compared to when it was not a holidays.
During clear weather condition the count of users is the highest and less in case of weather condition is light conditions.
During heavy rains no users have preferred boom bikes.

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

Answer: drop first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. Example

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Also it depends on the model. If you don't drop the first column then your dummy variables will be correlated .This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted.

If you have a small number of dummies, remove the first dummy. For example, if you have a variable gender, you don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female. However if you have a category with hundreds of values, It would be better not dropping the first column. That will make it easier for the model to "see" all the categories quickly during learning .

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

Answer: Numeric variables temp, atemp, Registered, Casual have a very good correlation. But Variable Registered has ahigh correlation of 0.95 with the target variable count.

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

Answer: 1. The Two Variables Should be in a Linear Relationship

The first assumption of simple linear regression is that the two variables in question should have a linear relationship.

The example of Sarah plotting the number of hours a student put in and the amount of marks the student got is a classic example of a linear relationship.

2. All the Variables Should be Multivariate Normal

The first assumption of linear regression talks about being ina linear relationship. The second assumption of linear regression is that all the variables in the data set should be multivariate normal.

In other words, it suggests that the linear combination of the random variables should have a normal distribution. The same example discussed above holds good here, as well.

3. There Should be No Multicollinearity in the Data

Another critical assumption of multiple linear regression is that there should not be much multicollinearity in the data. Such a situation can arise when the independent variables are too highly correlated with each other.

example, the variable data has a relationship, but they do not have much collinearity. There could be students who would have secured higher marks in spite of engaging in social media for a longer duration than the others.

4. There Should be No Multicollinearity in the Data

Another critical assumption of multiple linear regression is that there should not be much multicollinearity in the data. Such a situation can arise when the independent variables are too highly correlated with each other.

In our example, the variable data has a relationship, but they do not have much collinearity. There could be students who would have secured higher marks in spite of engaging in social media for a longer duration than the others.

5.There Should be Homoscedasticity Among the Data

Finally, the fifth assumption of a classical linear regression model is that there should be homoscedasticity among the data. The scatterplot graph is again the ideal way to determine the homoscedasticity. The data is said to homoscedastic when the residuals are equal across the line of regression. In other words, the variance is equal.

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

Answer: From the Final model we can conclude that the following 3 variables have an major impact on demand for the shared bikes.

1.Year 2019 : has a higher positive coefficient meaning the number of boom bike users demand in 2019 is high.
2.Weather sit light has a higher negative coefficient : meaning the number of boom bike user counts is drastically less when weather conditions are light.
3. Temperature : has a higher positive coefficient meaning the number of boom bike users demand is high when temperature conditions are good .