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

Categorical variables are : season, weekday, workingday,
Depended variables are: weathersit, temp, humidity, windspeed

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

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

4.How did you validate the assumptions of Linear Regression after building the model on the training set?
Here we have calculated r2_score.

```
from sklearn.metrics import r2_score
r2_score(y_test, y_test_pred)
```

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

- Spring season : -0.5186
- Temperature : 0.4524
- Mist : -0.3666

General Subjective Questions

1.Explain the linear regression algorithm in detail.
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as 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.

$$y= a_0+a_1x+ \varepsilon$$

2.Explain the Anscombe's quartet in detail.
Anscombe's quartet data is a type of data with same summary stats but data looks different when plot on graph that's why its imp not to just rely on Summary stats but also look into data visualization.

3.What is Pearson's R?
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Feature scaling is one of the most important data pre-processing step in machine learning.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

$$\boxed{X\_new = (X - X\_min)/(X\_max - X\_min)}$$

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$\boxed{X\_new = (X - mean)/Std}$$

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
VIF stands for Variance Inflation Factor.
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. If there is perfect correlation, then VIF is infinite.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.