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.: The following is my inference based on the analysis of the categorical variables:

- The demand for Spring is considerably low compared to the other seasons.
- There is a significant rise in demand in the year 2019.
- Demand across the year seems to follow a normal distribution with June-August showing peak performance.
- Even though the median of demand is high on regular days, holidays seem to have a wider spread in demand.
- The median across all weekdays is very similar, but Saturday seems to have the highest spread of demand.
- Days with a clear and/or cloudy weather have relatively higher demand compared to days with snow.

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

Ans.: drop_first=True is used to remove one level from a multi-level categorical variable during dummy variable creation. This reduces the number of features/columns in the dataframe for the model, by removing a redundant feature which can be easily denoted through the other dummy variables, by flagging all of them as '0'.

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

Ans.: From the plotted pair-plot among the numerical variables, it is evident that the variables ‘temp’ and ‘atemp’ have the highest linear correlation with the target variable. Also, it is to be noted that ‘temp’ and ‘atemp’ show high collinearity with ‘atemp’ being a derived variable of ‘temp’.

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

Ans.: The following assumptions of Linear Regression were made before building the model:

i. Normality
ii. No multicollinearity
iii. Homoscedasticity
iv. Linearity

The following assumptions were validated:

i. Normality: The residual analysis shows the residuals were normally distributed with mean at or around 0
ii. No multicollinearity: presence of multicollinearity was checked thoroughly during the model building process using Variance Inflation Factor(VIF)
iii. Homoscedasticity: Homoscedasticity was validated by plotting the actual values against the predicted values which show similar spread across all levels.
iv. Linearity: Linearity was validated by plotting pair-plot of independent variables against the target variable which showed linear correlation.

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

Ans.: Based on the model, the top 3 features the strongly contribute towards the demand of shared bikes are:

i. ‘temp’ showing highest positive contribution.
ii. ‘weathersit_bad’ or rainy days showing highest negative contribution.
iii. Winter showing strong positive contribution

General Subjective Question:

1. Explain the linear regression algorithm in detail.

Ans.: Linear Regression is a type of supervised machine algorithm. Linear regression model works when the dependent variable is linearly correlated to one or more of the independent variables. Linear regression model can be simple(target variable depends on one independent variable) or multiple(target variable depends on more than one independent variable). The main objective of the model is to fit a straight line that is best fit for the data. However, certain assumptions are made when working with linear regression model. They are:

- The dependent/target variable is linearly correlated with the independent variable/s
- The independent variables are not correlated with each other, i.e. no multicollinearity
- The variance of the errors is constant, i.e. homoscedasticity across all independent variables

2. Explain the Anscombe’s quartet in detail.

Ans.: Anscombe’s quartet is a group of four datasets. The four datasets, in spite of having very similar statistical structure, looks very different when plotted. It is often used to show the importance of analysing and visualising the data before building a model.

3. What is Pearson's R?

Ans.: Pearson's R or more commonly known as Pearson's Coefficient is the most commonly used formula to find a linear correlation between two sets of data.

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

Ans.: Scaling is a data preprocessing technique where the values of the independent variables are transformed to a similar scale.

Scaling is performed to transform the values of all the numerical independent variables to a similar scale. This in turn helps speed up the process of gradient descent, providing us with a model with coefficients which are similar in scale and understandable from a layman's perspective.

Standardised scaling is used to transform all the values of the independent variables into one standard deviation of mean. Normalised scaling on the other hand is used to transform the values of the independent variables within the range of 0 – 1.

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

Ans.: The formula for VIF is as follows:

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

As is evident from the formula, VIF can be infinite only when there is a perfect correlation, i.e. there is perfect multicollinearity between the independent variables, making R^2 = 1, which results in division by 0

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

Ans.: Q-Q plot, or Quantile-Quantile plot is a graphical tool used to compare two probability distributions by plotting against their quantiles.

Q-Q plot is can be used to detect shifts in scale, location, symmetry and even the presence of outliers. It also helps in comparing if the two datasets have similar distributional shapes and/or tail behaviours.