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? (3 marks)**
   Ans> Please find the categorical variables and their effect on the dependent variable(cnt) below:
   - Season – season 2 and season 4 were found to have significant effect on cnt.
   - Mnth – mnth_3 and mnth_9 have the most effect on the cnt
   - Weekday – weekday_6 is the only day to have most impact on the cnt
   - Weathersit – weathersit_2 and weathersit_3 have significant impact on the cnt

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Ans> The drop_first=True removes the first dummy variable. For example, if there is a categorical variable called colours with values white , black and red. The get_dummies by default will create 3 dummy variables representing each of the 3 colours. Now, if we specify drop_first=True, it will create 2 dummy variables (1 for black and 1 for red). The understanding here is that if both the dummy variables are "0" then the first one should be "1", and in order to represent that, we do not need to create 1 more variable.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Ans> The "temp" variable has the highest correlation.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Ans>
   The residual analysis was done on the train data to confirm that the error terms are normally distributed and have constant variance
   Using VIF, it was confirmed that the variables did not have multicollinearity
   Using pairplots for numeric variables, it was confirmed that there is linear relationship between X and y.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Ans> As per the model, the top 3 predictor variables are :
   - temp
   - weathersit_3
   - year

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Ans> Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear

regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s). The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process. A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

Ans> Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. **What is Pearson's R? (3 marks)**

Ans> 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.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and -.3	Weak	Negative
Between -.3 and -.5	Moderate	Negative

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

   Ans> Scaling is a technique to standardize the independent features present in the data in a fixed range. Independent variables are of different types. The numerical ones can be of type age(0-100 years), salary(in the range of 1000s), dimensions(decimal points), and many more. We don't want our machine learning model to confuse a feature with a larger magnitude as a better one. Feature scaling in Machine Learning would help all the independent variables to be in the same range, for example- centered around a particular number(0) or in the range (0,1), depending on the scaling technique.

   Standardization and normalization are two primary ways to apply feature scaling in Machine Learning. The differences between the two are as follows:

Normalisation	**Standardisation**
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers
It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

   Ans> An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

   The formula of VIF for a variable X1 is VIF1 = 1/(1-R2),

   If there is collinearity between X1 and the other independent variables suppose X2, X3, then R2 will tend to 1 and thus the formula of VIF will evaluate to infinity.

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

   Ans> Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

   Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s).

   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.