Assignment-based Subjective Questions Answers

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

   Ans: season, yr, mnth, holiday, weekday, weathersit are categorical variables in the dataset. From the analysis, it can be inferred that

   - Fall is the season to get maximum customers (September being the month). 2019 observed more sale than 2018.
   - Holidays drop the total rental bikes count.
   - Count of users decreases if weather condition is not good(Like Misty cloud and Light Rain)

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

   Ans: the use of drop_first=True helps remove **multicollinearity** and **redundancy** by eliminating one of the dummy variables for each categorical feature.

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

   Ans: *atemp* and *temp* has the highest correlation with the target variable.

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

   Ans: The assumptions of Linear Regression are validated using:

   - Q-Q plot and distribution plot of residuals to check if the errors are normally distributed.
     Scatter plot of residuals to verify constant variance (homoscedasticity) of the residuals.

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

   Ans: The top 3 features directly influencing the count are the features with highest coefficients:

   - Temp, Year (positively influencing)
   - Windspeed and Light Rain (negatively influencing).

General Subjective Questions Answers

1. Explain the linear regression algorithm in detail.

   Ans: Linear Regression is a fundamental algorithm in machine learning used for predicting a dependent variable (target) based on one or more independent variables (features). Its goal is to model the relationship between the dependent and independent variables by fitting a straight line (in simple linear regression) or a hyperplane (in multiple linear regression) to the data.

   The objective of linear regression is to find optimal β coefficient in such a way that the predicted values of y are as close as possible to the actual values. This is achieved by minimizing the **residual sum of squares** (RSS), which is the sum of squared differences between predicted and actual values.

   To find optimal value of β coefficient linear regression uses **Ordinary Least Squares (OLS)** method

   Assumptions of Linear Regression:

   - The relationship between the independent and dependent variables should be linear.
   - The residuals should have constant variance (Homoscedasticity).
   - The residuals should be normally distributed.
   - In multiple linear regression, the independent variables should not be highly correlated with each other.

2. Explain the Anscombe's quartet in detail.

   Ans: Anscombe's quartet is a collection of four different datasets that have nearly identical simple descriptive statistics (such as mean, variance, and correlation) and also same R-squar, but they have very different distribution when plotted.

   Anscombe's quartet demonstrates that while building model relying solely on summary statistics can hide important patterns in the data. Visualization allows you to detect relationships, outliers, and nonlinearities that descriptive statistics alone would miss.

3. What is Pearson's R?

   Ans: Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables. It is commonly used in linear regression. The value of Pearson's R always lies between -1 and +1. -1 represent perfectly negative corelation, +1 represent perfectly positive corelation and 0 represent no corelation.

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

Ans: Scaling refers to the process of transforming the features of a dataset to ensure that all features have a common scale or range. This is often necessary when different features have different units or ranges, which can affect the performance of many machine learning algorithms. Scaling is important because some machine learning algorithm such as Gradient Descent are sensitive to the scale of input data. There are two types of scaling:

- **Normalization (or min-max scaling)** rescales the values of features to a specific range, usually [0, 1] and it is sensitive to outliers.
- **Standardization** transforms the data such that the mean becomes 0 and the standard deviation becomes 1. Standardization centres the data and adjusts the scale of features. It doesn't have predefined range and is less sensitive to outliers.

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

Ans: VIF can become infinite when the predictor is perfectly correlated with one or more of the other independent variables mean one predictor is a linear combination of another predictor.

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

Ans: A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the **normal distribution.** It plots the quantiles of the observed data against the quantiles of a specified theoretical distribution.
In linear regression, one of the key assumptions is that the residuals (errors) should follow a normal distribution. A Q-Q plot helps to visually assess whether the residuals are normally distributed.