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)
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Assignment-based Subjective Questions

Answers:

1. Categorical variables in the dataset, such as season, weather situation, and month, significantly impact the dependent variable, which in this case is the demand for shared bikes. For instance, bike rentals might be higher in summer and lower in winter. Weather conditions like clear skies or light rain also influence people's willingness to use bikes. Each categorical variable adds a unique aspect to understanding the demand, showing how different conditions and times of the year affect bike rental patterns.

2. Using drop_first=True in dummy variable creation helps to avoid multicollinearity, a situation where two or more variables in a regression model highly correlate. By dropping the first category, we reduce the redundancy in the model. This helps in making the model more interpretable and efficient

3. Based on the pair-plot analysis among numerical variables, temperature (temp or atemp) typically shows the highest correlation with the target variable (bike rentals). Higher temperatures usually correlate with increased bike rental demand.

4. To validate the assumptions of Linear Regression, the following checks were performed:

> Verified that the relationship between independent variables and the dependent variable is linear.

> Checked for constant variance of errors (residuals).

> Confirmed that residuals are independent across observations.

>Assessed using Variance Inflation Factor (VIF) to ensure that independent variables are not highly correlated with each other.

5.

>Temperature (Temp/Atemp): Represents the weather conditions influencing bike usage.

>Season (e.g., Summer, Spring): Indicates the time of year affecting rental patterns.

>Year (Yr): Reflects the trend or growth in bike rentals over time.

General Subjective Questions

1. Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The algorithm assumes a linear relationship between these variables. It aims to find the best-fitting line (regression line) that minimizes the sum of squared differences (errors) between the observed values and the values predicted by the model. The basic form of a linear regression model with one independent variable is $Y = \beta0 + \beta1X + \varepsilon$, where Y is the dependent variable, X is the independent variable, β0 is the y-intercept, β1 is the slope, and ε is the error term. The algorithm uses methods like Ordinary Least Squares (OLS) to estimate the coefficients (β values) that best describe the relationship between the variables.

2. Anscombe's quartet comprises four different datasets that have nearly identical simple statistical properties (mean, variance, correlation, regression line), yet appear very different when graphed. Each dataset consists of eleven (x, y) points. This quartet illustrates the importance of visualizing data before analyzing it and the effect of outliers and unusual distributions on statistical properties. It demonstrates that different datasets can have the same statistical properties, so the same statistical metrics do not always imply similar data structures.

3. Pearson’s R, also known as Pearson’s correlation coefficient, is a statistic that measures the linear correlation between two variables. It ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 implies no linear correlation.

4. Scaling is a processing technique used to standardize the range of independent variables in data. It ensures that variables with larger scales don't dominate those with smaller scales, leading to a more balanced and effective model training process. Normalized scaling transforms data to fit within a [0, 1] range, maintaining the proportion of original data values. Standardized scaling, in contrast, transforms data to have a mean of 0 and a standard deviation of 1, effectively aligning the data distribution with the standard normal distribution.

5. The Variance Inflation Factor (VIF) may reach infinity when there is perfect multicollinearity in the dataset, indicating that one predictor variable in the regression model can be perfectly predicted from the others with no error. This typically occurs when two or more variables are highly correlated or when a variable is a linear combination of others, making it impossible to assess the individual contribution of the correlated variables to the model.

6. A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. In the context of linear regression, a Q-Q plot is commonly used to assess the normality of residuals (errors). It plots the observed residuals against the expected normal quantiles. If the residuals are normally distributed, the points on the Q-Q plot will approximately lie on a straight line. Deviations from this line indicate deviations from normality, which is a key assumption in linear regression models.