Linear Regression Assignment and General Subjective Questions

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?

From the analysis of the categorical variables, we can infer the following effects on the dependent variable (*cnt*):

- **Season:** The highest bike rentals are seen during the summer and fall seasons, while the lowest rentals are during winter.
- **Year:** There is an increase in bike rentals in 2019 compared to 2018, indicating growing popularity.
- **Month:** Bike rentals vary by month, with peaks in the summer months.
- **Holiday:** There are slightly more bike rentals on non-holidays compared to holidays.
- **Weekday:** Bike rentals vary across weekdays, with a noticeable pattern of higher rentals on weekends.
- **Working Day:** There are more bike rentals on non-working days compared to working days.
- **Weather Situation:** Clear weather conditions have the highest bike rentals, while heavy rain/snow conditions have the lowest.

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

Using `drop_first=True` during dummy variable creation is important because it helps to avoid the dummy variable trap, which occurs when the dummy variables are highly collinear. By dropping the first category, we ensure that the dummy variables are independent of each other, which helps in creating a more stable and interpretable model.

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

From the pair-plot analysis, the numerical variable with the highest correlation with the target variable (*cnt*) is the *temp* (temperature), showing a positive correlation.

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

To validate the assumptions of Linear Regression, the following steps were taken:

- **Linearity:** Scatter plots of the residuals vs. predicted values were checked to ensure there is no clear pattern, indicating a linear relationship.
- **Normality:** A Q-Q plot of the residuals was examined to check if the residuals are normally distributed.
- **Homoscedasticity:** Plots of residuals vs. predicted values were checked for constant variance.
- **Multicollinearity:** Variance Inflation Factor (VIF) values were calculated to ensure there is no multicollinearity among the independent variables.

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

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. Working Day
2. Year
3. Temperature

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The algorithm assumes that the relationship between the variables is linear. The linear regression equation is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Where y is the dependent variable, $x_1, x_2, \ldots, x_n$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and ϵ is the error term. The goal is to find the best-fitting line by minimizing the sum of the squared differences between the observed and predicted values.

2. Explain Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. Each dataset has the same mean, variance, correlation, and linear regression line. Anscombe's quartet demonstrates the importance of graphing data before analyzing it and the limitations of statistical properties alone. It underscores the need for visual data analysis to detect patterns, outliers, and differences that simple statistical measures may not reveal.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, 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.
- 0 indicates no linear relationship.

Pearson's R is calculated as the covariance of the variables divided by the product of their standard deviations.

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

Scaling is the process of adjusting the range of feature values to a standard range. It is performed to ensure that all features contribute equally to the model, especially in algorithms that are sensitive to the scale of the data, such as linear regression and k-nearest neighbors.

- **Normalized Scaling:** Rescales the data to a range of [0, 1]. It is useful when the data needs to be bounded.
- **Standardized Scaling:** Rescales the data to have a mean of 0 and a standard deviation of 1. It is useful when the data follows a Gaussian distribution.

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

The value of the Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity in the data, meaning that one predictor variable is an exact linear combination of other predictor variables. This perfect correlation leads to a division by zero in the VIF calculation, resulting in an infinite value.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a particular distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of a theoretical distribution. In linear regression, a Q-Q plot is used to check the normality of residuals. If the residuals are normally distributed, the points on the Q-Q plot will lie approximately along a straight line. This validation is important because normality of residuals is an assumption of linear regression, affecting hypothesis tests and confidence intervals.