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?**

   After analyzing the categorical variables, we can infer the following:

   - **Season:** The demand for shared bikes is higher during the summer and fall seasons, indicating a positive correlation with these seasons compared to spring and winter. People are more likely to use bikes in warmer weather.
   - **Weather Situation (Weathersit):** Demand is highest in clear or partly cloudy conditions (weathersit = 1) and significantly drops in adverse weather conditions such as heavy rain or snow (weathersit = 4).
   - **Year (yr):** Demand has increased from 2018 (yr = 0) to 2019 (yr = 1), suggesting growing popularity and adoption of the bike-sharing service.
   - **Month (mnth) and Weekday (weekday):** Certain months, like June to September, show higher demand. The demand also tends to fluctuate during weekdays, potentially peaking on weekends when people have more leisure time.

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

   Using drop_first=True when creating dummy variables is important to avoid the "dummy variable trap," which can lead to multicollinearity in the regression model. Multicollinearity occurs when one or more independent variables are highly correlated, causing redundancy and making it difficult for the model to interpret the impact of individual predictors. By dropping the first dummy variable, we prevent the introduction of perfect multicollinearity and ensure 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?**

   Among the numerical variables, the variable 'temp' (temperature in Celsius) shows the highest positive correlation with the target variable 'cnt' (count of total bike rentals). This indicates that as the temperature increases, the demand for bike rentals also tends to increase.

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

   After building the model, the following methods were used to validate the assumptions of linear regression:

- **Linearity:** Checked using scatter plots between the predicted values and residuals. The residuals should be randomly scattered without any clear pattern.
- **Normality of Residuals:** Verified by plotting a Q-Q (quantile-quantile) plot of the residuals to ensure they follow a normal distribution.
- **Homoscedasticity:** Ensured that the residuals have constant variance across all levels of the independent variables by analyzing a residual vs. fitted values plot.
- **Multicollinearity:** Evaluated using the Variance Inflation Factor (VIF) to check for high correlations between predictor variables.

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

   The top three features contributing significantly to explaining bike demand, as per the final model, are:

   1. **Temperature (temp):** Higher temperatures are associated with increased bike rentals.
   2. **Year (yr):** Demand increased in 2019 compared to 2018, indicating growth in the popularity of bike sharing.
   3. **Season (season_summer and season_fall):** The demand is significantly higher in the summer and fall seasons compared to spring, suggesting a seasonal impact.

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 objective is to find the best-fitting linear equation ($y = mx + c$) that predicts the dependent variable (y) as a linear combination of the independent variables (x).

   The algorithm works by minimizing the sum of the squared differences (residuals) between the observed values and the predicted values. The coefficients of the linear equation are estimated using methods like Ordinary Least Squares (OLS), which seeks to minimize the error terms. Key assumptions include linearity, independence of errors, homoscedasticity, and normality of residuals.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, such as mean, variance, and correlation, yet appear very different when graphed. It was created by statistician Francis Anscombe to demonstrate the importance of visualizing data before analyzing it. The quartet illustrates that relying solely on statistical measures without graphical representation can lead to incorrect conclusions, as the underlying data patterns (e.g., linear, non-linear, outliers) are crucial for proper analysis.

3. **What is Pearson's R?**

   Pearson's R, or Pearson correlation coefficient, measures the linear relationship 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.

   It is calculated by dividing the covariance of the two variables 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 involves adjusting the range of the data so that different features contribute equally to the model. It is crucial when using algorithms sensitive to the scale of data, such as linear regression.

   - **Normalized Scaling** rescales the data to a fixed range, typically [0, 1].
   - **Standardized Scaling** centers the data around the mean (0) and standard deviation (1). It is more appropriate when the data has outliers or is normally distributed.

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

   The Variance Inflation Factor (VIF) can become infinite when perfect multicollinearity exists among independent variables, i.e., when one predictor is an exact linear combination of other predictors. In such cases, the denominator of the VIF formula becomes zero, leading to 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 plot, or quantile-quantile plot, is a graphical tool to assess whether a dataset follows a specific distribution, typically normal. It plots the quantiles of the dataset against the quantiles of a normal distribution. If the points roughly align along a straight line, the residuals are normally distributed, satisfying one of the key assumptions of linear regression. Deviations from the line suggest non-normality, which can impact the reliability of model estimates.

Contact Information

- Name – Pushkar Mishra
- Email – p2k3m_2002@yahoo.com