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?

To analyze the effect of categorical variables on the dependent variable, we can refer to the coefficients of the respective categorical variables in the regression model.
Below mentioned categorical variables from the dataset:

- **Season (season_2, season_4):**
  - The positive coefficients for season_2 (summer) and season_4 (winter) suggest a significant positive impact on bike demand.
  - Bike demand is higher in both summer and winter seasons compared to other seasons.
- **Months (month_8, month_9):**
  - The positive coefficients for month_8 (August) and month_9 (September) indicate a significant positive impact on bike demand during these months.
  - There is an increase in bike demand during the late summer months.
- **Weather Conditions (weather_2, weather_3):**
  - The negative coefficients for weather_2 (partly cloudy) and weather_3 (light rain, thunderstorm) suggest a significant negative impact on bike demand.
  - Bike demand decreases during partly cloudy weather and light rain or thunderstorms.
- **Holiday (holiday):**
  - The negative coefficient for the holiday variable indicates a significant negative impact on bike demand during holidays.
  - Bike demand tends to decrease on holidays.

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

When creating dummy variables from categorical variables, the parameter **drop_first=True** is important to avoid multicollinearity issues in regression models.
Below mentioned in details:

- **Avoiding Multicollinearity:**
  - When you create dummy variables without dropping the first category (i.e., drop_first=False), the dummy variables become perfectly correlated because knowing the values of all but one dummy variable allows you to infer the value of the dropped one.
  - This perfect correlation leads to multicollinearity in the model, which can cause problems in regression analysis.
- **Multicollinearity Issues:**
  - Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to determine the individual effect of each variable on the dependent variable.
  - This can lead to unstable coefficient estimates, inflated standard errors, and difficulties in interpreting the model.
- **Drop First to Resolve Multicollinearity:**
  - Setting **drop_first=True** addresses multicollinearity by creating n-1 dummy

variables for a categorical variable with n categories.
    - It removes the perfect correlation between dummy variables and ensures that each dummy variable is independent and provides unique information to the model.
- **Interpretability:**
    - Dropping the first category makes the interpretation of coefficients more straightforward. The coefficients for the remaining dummy variables represent the change in the dependent variable compared to the dropped category.

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

- 'atemp' has a positive correlation of approximately 0.65.
- temp' has a positive correlation of approximately 0.627.

These correlations suggest that as the value of the variable increases, the bike rental count tends to increase.

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

Validation of the assumptions of Linear Regression are done by:

- **Normality of residuals:**
    - This was done by using histogram plot.
    - The residuals are normally distributed and the mean of the residuals are centered around zero(-6.661338e-17)
- **Homoscedasticity:**
    - This was done by plotted a scatter plot of residuals against predicted values which helped check for homoscedasticity.
    - The residuals have constant variance across all levels of the independent variables**.**

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

From the final model, the top three features contributing significantly towards explaining the demand for shared bikes are:

- **Temperature (temp):** The coefficient for temperature is 0.5174, indicating a positive relationship. As temperature increases, the demand for shared bikes tends to increase.
- **Windspeed:** The coefficient for windspeed is -0.1497, indicating a negative relationship. As windspeed increases, the demand for shared bikes tends to decrease.
- **Year:** The coefficient for the year is 0.2325, indicating a positive relationship. As the year increases, the demand for shared bikes tends to increase

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal is to find the best-fitting line that minimizes the sum of the squared differences between the observed values and the values predicted by the linear model.

Below mentioned explanation of the linear regression algorithm:

1) **Model Representation:** Linear regression represents the relationship between the independent variable X and the dependent variable Y using a linear equation: $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\epsilon$
   Y is the dependent variable (the variable we want to predict).
   X2,…,Xn are the independent variables (features or predictors).
   β0 is the intercept term (the value of Y when all X values are zero).
   β1,β2,…,βn are the coefficients (slope) representing the change in Y with a one-unit change in X.
   ϵ is the error term, representing the unexplained variation in Y.
2) **Cost Function:** The cost function is often used to measure the error. In linear regression, the most common cost function is the Mean Squared Error (MSE), which is the average of the squared errors.
   $MSE=n1\sum i=1n(Yi-Y\hat{}i)2$
3) **Training the Model:** The model is trained by adjusting the coefficients to minimize the objective function. This is often done using optimization algorithms like gradient descent.
4) **Gradient Descent:** Gradient descent is an iterative optimization algorithm used to minimize the objective function. It works by iteratively adjusting the coefficients in the negative direction of the gradient (partial derivatives) of the objective function.
   $\beta j:=\beta j-\alpha\partial(\beta j/\partial J)$
   Here, α is the learning rate, controlling the size of each step in the optimization process.
5) **Assumptions of Linear Regression:** Linear regression makes several assumptions, and it performs well when these assumptions are met:
   - **Linearity:** The relationship between the independent and dependent variables is linear.
   - **Independence:** The residuals (errors) are independent of each other.
   - **Homoscedasticity:** The variance of the residuals is constant across all levels of the independent variables.
   - **Normality of Residuals:** The residuals are normally distributed.
   - **No Multicollinearity:** The independent variables are not highly correlated.
6) **Evaluation:** Once the model is trained, it can be evaluated using metrics like Mean Squared Error (MSE), R-squared, or others, depending on the problem.
7) **Prediction:** The trained model can be used to make predictions on new, unseen data.

2. What is Pearson's R?

Pearson's correlation coefficient, often denoted as r or Pearson's r, is a statistical measure of the strength and direction of a linear relationship between two continuous variables. It quantifies how well a straight line can describe the relationship between two variables. The coefficient ranges from -1 to 1, where:

- r=1: Perfect positive linear correlation
- r=−1: Perfect negative linear correlation
- r=0: No linear correlation

The formula for Pearson's correlation coefficient (r) between variables X and Y in a dataset is given by:

$r=\sum i=1n(Xi-X^-)\ (Yi-Y^-)/\sum i=1n(Xi-X^-)^2(Yi-Y^-)^2$

Where:

- Xi and Yi are the individual data points.
- X^- and Y^- are the means of X and Y respectively.
- n is the number of data points.

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

Scaling:

Scaling is the process of transforming numerical variables to a standard range or distribution. The purpose of scaling is to bring all features to the same level of magnitude, making it easier to compare and interpret the data. This is particularly important when using machine learning algorithms that are sensitive to the scale of input features.

Why Scaling is Performed:

- **Uniform Magnitude:** Scaling ensures that all features have a similar scale, preventing some features from dominating others.
- **Convergence:** It helps optimization algorithms converge faster by providing a more spherical shape to the cost function.
- **Distance-Based Algorithms:** Scaling is crucial for distance-based algorithms, as it ensures that distances are calculated accurately.

Normalized Scaling vs. Standardized Scaling:

1) **Normalized Scaling (Min-Max Scaling):**
    - **Formula:** $X_{NORMALIZED} = X\text{max}-X\text{min} / X-X\text{min}$
    - **Range:** Transforms values to a range between 0 and 1.
    - **Advantages:** Simple and intuitive; preserves the shape of the original distribution.
    - **Disadvantages:** Sensitive to outliers; may not handle extreme values well.
2) **Standardized Scaling (Z-score Normalization):**
    - **Formula:** $X_{STANDARDIZED} = (X-\mu) / \sigma$
    - **Range:** Transforms values to have a mean (μ) of 0 and a standard deviation (σ) of 1.
    - **Advantages:** Less sensitive to outliers; useful for algorithms that assume

normally distributed features.

- **Disadvantages:** Alters the original distribution; may not be suitable for algorithms that do not assume normality.

4. 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 the predictor variables. Perfect multicollinearity means that one or more independent variables in a regression model can be perfectly predicted by a linear combination of other variables.

5. 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 used to assess whether a given dataset follows a particular theoretical distribution, such as a normal distribution. In the context of linear regression, Q-Q plots are often used to check the normality of the residuals or errors, which is one of the assumptions of linear regression.

Use and Importance in Linear Regression:

- **Normality Check:** One of the assumptions of linear regression is that the residuals (the differences between the observed and predicted values) are normally distributed. A Q-Q plot helps in visually assessing whether the residuals follow a normal distribution.
- **Identification of Outliers:** Outliers in the residuals can be detected as points that deviate from the expected straight-line pattern in the Q-Q plot.
- **Model Assumptions:** Checking the normality of residuals is crucial for ensuring the validity of statistical inferences and hypothesis tests associated with linear regression.