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?

    - There is an increase in bike daily rental count during summer and fall, followed by a decrease in winter and spring.
    - The daily rental count witnesses growth from 2018 to 2019.
    - The bike rental count grows from January to June followed by a drop in July and an increase from August to October after which it begins to fall in next two months gradually.
    - The median for holidays is higher compared to non-holidays, indicating that there are significantly more no of bike rentals during holidays.
    - The median shows consistent usage during all weekdays.
    - The demand for bike rental count is highest in mild weather condition, followed by moderate and extremely moderate.

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

The drop_first =True can reduce multicollinearity in data by removing the first level of the categorical variable.

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

The temp and a_temp variables are highly correlated with the target variable cnt.

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

I validated the assumptions of Linear Regression based on:

- Finding normality of error terms and ensuring they are normally distributed.
- By plotting the observed values against the predicted values (I found a linear relationship between both)
- By checking multicollinearity between variables and ensuring the VIF values are less than 5.

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

temp, Season_Winter, and sep are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is crucial algorithm used for predicting the value of a target variable based on one or more independent variables. In case of our model, the algorithm assumes a linear relationship between the dependent variable cnt: count of total rental bikes and independent variables such as temp, windspeed, season, etc. It models the relationship using the equation:

Y = a + bx

Where x and Y are two variables on the line (independent and dependent), a is the intercept of the line and b is the slope of the line.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets whose purpose is to demonstrate the importance of data visualization. Anscombe's quartet consists of 11 x-values and their corresponding 11 y-values. It provides nearly identical summary such as mean, variance, correlation coefficient, and regression line.

3. What is Pearson's R?

Pearson's R, or Pearson correlation coefficient, is used to measure the strength and direction of the relationship between two continuous variables. It ranges from -1 to +1 where -1 indicates perfect negative correlation, +1 indicates perfect positive correlation, and 0 states no correlation.
Its formula is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where xi and yi are individual data points,
$\bar{x}$, $\bar{y}$ are means of x and y,
and n is no. of data points.

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

Scaling is a preprocessing technique in machine learning that adjusts the range of independent variables or features so that all features contribute equally and are on comparable scales in our model. Machine learning algorithms assume features are on the same scale so features with larger scales can disrupt our model, leading to demand for scaling.
There are two major types of scaling:

- Normalized Scaling (Min-Max Scaling)

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

It is useful when we want all features strictly between 0 and 1.

- Standardized Scaling

$$X_{scaled} = \frac{X - X_{mean}}{X_{stddev}}$$

In this mean and standard deviation is used for scaling.

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

An infinite VIF occurs when there is a perfect correlation between two or more independent variables. This leads the r squared value in the VIF formula to become 1, which leads to 1/(1- r squared) to be equal to infinity. To fix this, one of these variables is usually removed.

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) are plots of two quantiles against each other and it is used to find if two samples of data came from the same population or not. In linear regression, it is used to check if residuals are normally distributed. If points lie roughly on the diagonal line, the normality assumption holds.