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?

Categorical variable in dataset is having good effect on dependent variable because if we see the box-plots of categorical variables like months, year, weathersit, season then we find that the rental bikes count are much variating according to these columns. So, these columns can tell or explain much variance about dependent variable.

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

It removes the first column of the dummy variable to reduce the number of columns created during the dummy variable creation process.

As a result, it reduces the correlations formed between dummy variables.

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

"temp" & "atemp" has the highest correlation with the target variable.

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

1. Linear relationship between X and Y
2. By plotting a histogram of residuals and checked if error terms are normally distributed or not
3. If Error terms are independent of each other or not
4. By plotting a scatter plot of residuals & fitted values(predicted values) and check If Error terms have constant variance (random distribution of data points) (homoscedasticity) or not

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

Based on the final model the top 3 features contributing significantly towards explaining the demand are:

A) temp = 0.451

B) yr_2019 = 0.234

C) weathersit_Light_rainsnow = -0.286

General Subjective Questions

1. **Explain the linear regression algorithm in detail**.
Linear Regression is a supervised learning-based machine learning algorithm this method is used when target is continuous, it works on the strength of linear relation between independent variables and dependent variable.
Mathematically, we can write a linear regression equation as:
$y = mx+c$
Here, x and y are two variables on the regression line. m= Slope of the line
c = y-intercept of the line
x = Independent variable from dataset y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail.
Anscombe's Quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but have some peculiarities that fool the regression model if built. They have very different distributions and show up differently on scatter plots.

3. What is Pearson's R?
Pearson's Correlation Coefficient is also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation in statistics. It's a statistic that calculates the linear relationship between two variables. It, like all correlations, has a numerical value between -1.0 and +1.0.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
scaling refers to putting the feature values into the same range.

It is common practise in regression to scale the features so that the predictors have a mean of 0. When the predictor values are set to their means, it is easier to interpret the intercept term as the expected value of Y.

Normalization means rescales the values into a range of [0,1]. Standardization means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

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

When the VIF is infinity then this shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

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

Plots (Quantile-Quantile Plots) are comparisons of two quantiles. A quantile is a fraction of values that fall below that quantile. The median, for example, is a quantile where 50% of the data falls below it and 50% fall above it. The goal of Q Q plots is to determine whether two sets of data are from the same distribution. On the Q Q plot, a 45-degree angle is plotted; if the two data sets are from the same distribution, the points will fall on that reference line.