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 variables can have various impacts on the dependent variable, often introducing non-linear relationships that significantly influence predictions. However their impact cannot be fully indentified with linear regression model.

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

Using `drop_first=True` avoids multicollinearity by removing one category from each categorical variable to avoid the redundancy and may affect the regression result.

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

Temperature in Celsius (temp) and Feeling temperature in Celsius (atemp)

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

a) Rsquare is 0.843 which means 84.3% of the variance in Sales is explained by the model which is a good number
b) Adjusted Rsquare (0.838) is also a good number
c) Prob F-statistic is close to zero
d) All the p-values of the model is less than 0.05 which means all of them are statistically significant
e) VIF for the input variables is less than 5 which shows less Multicollinearity

Also after running the model on Test set and found out the following performance indicators that validated the model is good:

a) Rsquare is 0.818 which has almost the same accuracy of Rsquare and Adjusted Rsquare.
b) Residuals are normally distributed.

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

- Temperature (temp) : 0.48. Sales tend to rise with increased temperature
- Year (2019) : 0.23. Shows a growing demand from the previous year
- Light_snow : -0.25 Weather factor that is negatively affecting the sales

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression aims to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The algorithm minimizes the sum of squared differences between observed and predicted values to determine the best-fit line.
When it comes to multi variable model, the approach involves either adding variables one by one to the model or adding all variables and then removing the least significant variables while observing the Rsquare of the model and p-value & VIF values of input variables for each iterations. General practice is to the latter method and REF method is used to select the variables for a faster approach.
The data is basically split into train sets and test sets and the iterations are done on the train set and the final model is run on the test set. The final model Rsquare value should be closer to the train set values and the residuals should be normally distributed.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of datasets that have the same mean, standard deviation and regression line but vastly different distributions and appearances. It illustrates the importance of visualizing data rather than relying solely on summary statistics.

3. What is Pearson’s R?

Pearson’s R is a measure of the linear correlation between two variables, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 a perfect negative correlation, and 0 no correlation.

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

Scaling transforms data to a specific range or distribution. It's performed to ensure that variables contribute equally to the analysis. Normalized scaling (min-max scaling) rescales data to a [0,1] range, while standardized scaling (z-score normalization) transforms data to have a mean of 0 and standard deviation of 1. One benefit of Min-max scaling is that it generally takes care of the Outliers.

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

VIF (Variance Inflation Factor) can be infinite when there is perfect multicollinearity, meaning one predictor variable is an exact linear combination of other predictor variables. This makes the model's matrix non-invertible, causing VIF to spike. The formula VIF = $1/(1-R^2)$ to becomes infinite when R^2 becomes 1 or say that 100% of variance is explained by this variable which will corrupt the model.

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

This plot compares the quantiles of the residuals with the quantiles of a standard normal distribution. In linear regression, it helps assess whether residuals follow a normal distribution, a key assumption for valid hypothesis testing and confidence intervals.