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?

For weather, Light Snow was most impactful, since it negatively affected demand. For season, similarly, winter was most impactful, but in a slightly positive direction.

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

Using drop_first=True created n-1 variables given n categories in the original variable. The n-th category, then, can be inferred when all n-1 variable are False. Doing this reduces redundancy, which is important because redundant variables introduce multicollinearity in the model.

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

temp is the numerical variable with the highest correlation (0.65)

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

Throughout the model building process, I checked VIF to eliminate multicollinearity from the model. After training, I ran residual analysis on the model. I created a histogram of the residuals, i.e., the error terms and verified that it followed a normal distribution with mean 0. I also created a scatterplot of the predicted y-values vs the residuals and verified that there was no discernible pattern in the distribution of residuals.

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

The top 3 features are *temp*, *yr*, and *Light Snow*. *temp* and *yr* are both positively correlated, while *Light Snow* negatively affects demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Given independent variables $x_1, \dots, x_n$ and a dependent variable y, and assuming that the relationship between the independent and dependent variables is linear, the linear regression algorithm seeks to find the line of best fit for the given data points.
The line of best fit would be of the form $\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$
Here β_0 is the y-intercept, and ϵ is the error term, i.e., the difference between the line and the data point.
So to find the coefficients $\beta_0, \dots, \beta_n$, we use a method called Ordinary Least Squares. We minimize the square of the difference between the actual y-value and the predicted y-value $\hat{y}$. So the cost function to be minimised becomes:

$$\sum_{i=1}^{m} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}))^2$$

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet is a set of four distinct datasets created by the statistician Francis Anscombe. Each dataset consists of 11 (x, y) pairs. The datasets have near-identical summary statistics, i.e., mean, median, variance, even the equation of the line of best fit is identical. Yet the four datasets have drastically different distributions when graphed. Anscombe created his quartet to demonstrate the importance of graphing data before analyzing it.

3. What is Pearson's R? (3 marks)

Pearson's R is a measure of correlation, i.e. how strongly two variables are linearly related. The value ranges between 1 and -1. A value of 1 implies that the two variables are perfectly linearly related with a positive slope; when one value increases, so does the other. A value of -1 implies a linear relationship with negative slope. A value of 0 implies no linear relationship.

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

Scaling is performed to bring all values of variables into the same range. This is done because variables can have wildly different values, and that can affect performance of the algorithm, as well as interpretability of the resultant coefficients. Normalised scaling, also called min-max, scales all values into the range of 0 to 1. Standardised scaling standardises the values into having mean 0 and standard deviation 1. Standardised scaling can thus change the values of dummy variables, while min-max scaling does not.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF = $\frac{1}{1-R^2}$. For the VIF is infinite, R^2 must be 1. This means that all the variation of that variable is explained by other variables in the model. The variable is perfectly multicollinear with other variables, and is thus redundant.

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

In linear regression, we assume that the residuals (the difference between the actual values and the predicted values) follow a normal distribution.

The Q-Q plot allows us to visually inspect whether the residuals are normally distributed. If the residuals follow a normal distribution, the points on the Q-Q plot should approximately form a straight line.