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?

The categorical variables in the dataset, such as season, holiday, and working day, have a significant impact on the demand for shared bikes. The model's coefficients indicate that summer and spring seasons have a positive impact on the demand, while holidays and non-working days have a negative impact.

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

Using `drop_first=True` during dummy variable creation is important to avoid multicollinearity in the model. When creating dummy variables, one category is used as the reference category, and the coefficients of the other categories are calculated relative to this reference. Without dropping the first category, the model becomes over-specified, resulting in non-unique coefficients.

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

The pair-plot shows that the temp variable has the highest correlation with the target variable cnt.

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

The assumptions of Linear Regression were validated as follows:

- Linearity: Checked using pair-plot.
- Homoscedasticity: Assessed through the residual plot.
- Normality of residuals: Verified using a Q-Q plot.
- Multicollinearity: Checked using VIF (Variance Inflation Factor) values.

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 contributing significantly to explaining the demand for shared bikes are:

1. temp (Temperature)
2. workingday
3. season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised learning algorithm used for predicting the value of a continuous output variable based on one or more input features. The goal of linear regression is to find the best-fitting linear line that minimizes the sum of the squared errors between the observed responses and the predicted responses.

The linear regression algorithm works as follows:

- The algorithm starts by assuming a linear relationship between the input features and the output variable.
- The algorithm then uses a cost function, typically the mean squared error (MSE), to measure the difference between the observed responses and the predicted responses.
- The algorithm uses an optimization technique, such as gradient descent, to minimize the cost function and find the best-fitting linear line.
- The algorithm outputs the coefficients of the linear line, which can be used to make predictions on new data.

The linear regression algorithm can be represented mathematically as follows:

$y = β0 + β1x + ε$

where y is the output variable, x is the input feature, β0 is the intercept, β1 is the slope, and ε is the error term.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that were created by Francis Anscombe in 1973 to illustrate the importance of visualizing data before analyzing it. Each dataset consists of 11 data points and has the same mean, variance, and correlation coefficient, but they have very different distributions.

The four datasets are:

- Dataset A: A simple linear relationship between the input feature and the output variable.
- Dataset B: A non-linear relationship between the input feature and the output variable.
- Dataset C: A linear relationship between the input feature and the output variable, but with one outlier.
- Dataset D: No relationship between the input feature and the output variable.

Anscombe's quartet highlights the importance of visualizing data before analyzing it, as the same statistical measures can be obtained from very different datasets.

3. What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that calculates the strength and direction of the linear relationship between two continuous variables. It is a value between -1 and 1, where:

- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

Pearson's R is calculated using the following formula:

$R = \Sigma[(x_i - \bar{x})(y_i - \bar{y})] / \mathrm{sqrt}[\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2]$

where xi and yi are the individual data points, $\bar{x}$ and $\bar{y}$ are the means of the two variables, and Σ denotes the sum.

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

Scaling is a technique used to transform the data into a common range, usually between 0 and 1, to prevent features with large ranges from dominating the model.

Scaling is performed to:

- Improve the stability and speed of the algorithm.
- Prevent features with large ranges from dominating the model.
- Improve the interpretability of the results.

There are two types of scaling:

- Normalized scaling: This method scales the data to a common range, usually between 0 and 1, using the following formula:

$x' = (x - \min(x)) / (\max(x) - \min(x))$

- Standardized scaling: This method scales the data to have a mean of 0 and a standard deviation of 1 using the following formula:

$x' = (x - \mu) / \sigma$

where μ is the mean and σ is the standard deviation

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

The Variance Inflation Factor (VIF) is a measure of multicollinearity between two or more features. It is calculated using the following formula:

$VIF = 1 / (1 - R^2)$

where R^2 is the coefficient of determination.

The VIF can be infinite when:

- Two or more features are perfectly correlated, resulting in a R^2 of 1.
- The data is highly multicollinear, resulting in a R^2 close to 1.

In such cases, the VIF is infinite, indicating that the features are highly correlated and should be removed or transformed.

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

A Q-Q plot, also known as a quantile-quantile plot, is a graphical method used to compare the distribution of two datasets. In linear regression, a Q-Q plot is used to check the normality assumption of the residuals.

A Q-Q plot plots the quantiles of the observed residuals against the quantiles of a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot should form a straight line. If the points deviate from the straight line, it indicates that the residuals are not normally distributed.

The use and importance of a Q-Q plot in linear regression are:

- To check the normality assumption of the residuals.
- To identify outliers and non-normality in the residuals.
- To ensure that the linear regression model is valid and reliable