Linear Regression Subjective Questions

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?

Ans: From the analysis of categorical variables such as season, day of the week, month, and weathersit, it is evident that these variables have a significant impact on cnt. The derived dummy variables from these categories contribute to explaining approximately 83% of the variation in cnt. This high R-squared value underscores the importance of these categorical predictors in capturing and presenting the fluctuations in bike rental demand observed in the dataset.

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

Ans: When we create dummy variables from categorical data in a regression model, using drop_first=True means we drop one of the categories to avoid redundancy. This helps ensure clarity in the model about the relationships between different categories, making it easier to understand how each category affects the outcome we're predicting, like bike rental demand. It also ensures our model runs smoothly and gives us accurate predictions without issues.

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

Ans: Looking at the pair-plot among the numerical variables, temp 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?

Ans: After building a linear regression model, we checked if the residuals (the differences between predicted and actual values) were random and normally distributed. We also examined whether the variance of residuals remained consistent across all predicted values, ensuring the model's reliability. Additionally, we verified that the relationship between each predictor and the bike rental count was linear. Finally, we looked for any signs of multicollinearity among predictors and assessed if any data points had undue influence on the model's predictions. These steps helped validate the model's assumptions and ensure its accuracy in predicting bike rental demand

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

Ans: Based on the final model, the top three features contributing significantly to explaining the demand for shared bikes are temp, whether it is a workingday, and the season. These variables show strong correlations with bike rental counts, indicating their importance in predicting demand.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a simple yet powerful algorithm for predicting numeric values based on input features. Here's how it works in simple terms:
i) **Understanding the Relationship**: Linear regression assumes a linear relationship between the input variables (features) and the output variable (target). It tries to find the best-fitting straight line (or hyperplane in higher dimensions) that explains how feature changes affect the target.
ii)**Model Representation**: The model is represented as y = mx + c where y is the predicted value (target), x are the input features, m is the slope (weights or coefficients), and c is the intercept (bias).
iii)**Training the Model**: During training, the algorithm adjusts m and c to minimize the difference between predicted values and actual values (known as residuals or errors). This process uses a method called Ordinary Least Squares (OLS) to find the optimal values of m and c.
iv)**Making Predictions**: Once trained, the model can predict the target variable for new input data by applying the learned coefficients to the feature values.
v)**Assumptions**: Linear regression assumes that the relationship between variables is linear, residuals are normally distributed, and there is minimal multicollinearity (high correlation between predictors). Violations of these assumptions can affect model accuracy.
vi)**Evaluation**: The model's performance is evaluated using metrics like R2 (coefficient of determination), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) to assess how well it fits the data.

linear regression provides a straightforward method to understand and predict outcomes based on input, making it a fundamental tool in statistical modeling and machine learning.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a famous example in statistics that consists of four datasets with very similar simple statistical properties—like mean, variance, and correlation—yet they look very different when plotted. This demonstrates that datasets can appear similar in terms of summary statistics but exhibit entirely different patterns when graphed. It underscores the importance of visualizing data to understand its relationships and variability, as relying solely on summary statistics may overlook important nuances and trends in the data.

3. What is Pearson's R?

Ans: Pearson's R, or Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. It is widely used to assess how closely two variables move together and to what extent changes in one variable are associated with changes in another.

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

Ans: Scaling in the context of data preprocessing refers to transforming the numerical values of variables to a standardized range.
It's done to ensure that variables with different scales and units contribute equally to the analysis and model training process.
Normalized scaling typically refers to scaling variables to a range of [0, 1], maintaining their original distribution but adjusting the range. Standardized scaling (or z-score scaling) transforms variables to have a mean of 0 and a standard deviation of 1, making them comparable by standard units of deviation from the mean. This ensures that variables are on a common scale for better interpretability and model performance.

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

Ans: Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to collinearity with other predictors. When VIF is infinite, it indicates perfect multicollinearity, where one predictor variable can be perfectly predicted from others. This situation arises when two or more variables are nearly perfectly correlated, meaning their information content is redundant and cannot be separated in the regression model, leading to instability in estimating their coefficients.

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

Ans: A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess whether a dataset follows a particular distribution, such as the normal distribution.
It compares the quantiles of the dataset against those of a theoretical distribution. In linear regression, Q-Q plots are crucial for checking the assumption of normality in residuals. If the residuals (the differences between observed and predicted values) are normally distributed, the points on the Q-Q plot will fall approximately along a straight line. Deviations from this line indicate departures from normality, which can affect the reliability and interpretation of regression results. Therefore, Q-Q plots help ensure that the assumptions underlying linear regression are valid and that the model is appropriate for the data at hand.