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?

It was observed that the variables `weathersit`, `yr`, `season`, `mnth` and `holiday` had an association with the demand for bikes. In particular,

- The median demand for bikes was lower during light snow and rain as compared to clearer and slightly misty weather.
- The median demand for bikes was higher in the year 2019 as compared to the year 2018.
- The median demand for bikes was the lower during spring season as compared to fall, summer and winter seasons.
- The median demand for bikes was higher during the third quarter of the year as compared to the other quarters.
- The median demand for bikes was lower on holidays as compared to working days.

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

This ensures that dummy variables are only created for n-1 levels if a categorical variable has overall n levels. This ensures brevity as redundant information is not fed to the model, given that all the information about the categorical variable is captured in n-1 levels.

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

The `temp` variable 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?

- A histogram of the error terms was plotted to check for the normality of the error terms.
- A scatter plot of the error terms against the dependent variable was plotted to check for constant variance and independence of error terms.

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

Temperature, light snow weather conditions and year of operation are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a popular statistical method used to model the relationship between a dependent variable and one or more independent variables. The relationship between the dependent and the independent variables are modelled as a straight line, which can be expressed as:

$Y = \beta 0 + \beta 1X + \varepsilon$

where Y is the dependent variable, X is the independent variable, β0 is the intercept, β1 is the slope of the line, and ε is the error term.

The model is trained using the least squares method such that the values of the intercept and coefficients that minimize the sum of squares between the predicted and actual values of the dependent variable are found.

The model is evaluated using the R-squared or coefficient of determination score which determined the goodness of fit of the model, that is, how much of the variance in the dependent variable can be explained by the fitted model.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that have the same statistical properties but are visually distinct. Each dataset in the quartet contains 11 (x, y) pairs.

Dataset I: This dataset has a linear relationship between x and y. The relationship is strong and positive.

Dataset II: This dataset has a non-linear relationship between x and y. The relationship is also strong and positive, but it is best described by a quadratic function.

Dataset III: This dataset has a linear relationship between x and y, but with an outlier that greatly affects the correlation coefficient. The outlier is located at the far right of the dataset.

Dataset IV: This dataset has no linear relationship between x and y, but with an influential point that greatly affects the slope of the regression line.

3. What is Pearson's R?

Pearson's correlation coefficient is a measure of the strength and direction of the linear relationship between two continuous variables. Its values range between +1 and -1 where =1 indicates a perfect positive correlation and -1 indicates a perfect negative correlation and 0 indicates no linear relationship.

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

Scaling is the process of transforming values of features such that they lie within a similar range. When variables in a model have different scales, it can lead to unreasonable coefficients that might be difficult to interpret. Scaling is performed for ease of interpretation of model coefficients and for faster convergence of gradient descent methods.

Normalized scaling method rescales the feature values to be within a range of 0 to 1. It is done by subtracting the minimum value of the feature and dividing by the range (i.e., the difference between the maximum and minimum values). Standardized scaling method transforms the feature values to have a mean of 0 and a standard deviation of 1. It is done by subtracting the mean of the feature and dividing by the standard deviation.

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

VIF of a variable can be infinite when the R-squared of a model that predicts the variable from the remaining predictor variables is 1, that is, there is perfect multicollinearity and the variable under question can be fully explained by all other independent variables in the dataset.

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

A Q-Q plot (quantile-quantile plot) is a graphical technique used to compare the distribution of a sample to a theoretical distribution. The Q-Q plot plots the quantiles of the sample against the corresponding quantiles of the theoretical distribution. If the sample follows the theoretical distribution, the points in the Q-Q plot will fall along a straight line.

In linear regression, a Q-Q plot is used to check the normality assumption of the error terms. By using a Q-Q plot, we can visually check whether the residuals of a linear regression model follow a normal distribution. If the normality condition is satisfied, the points in the Q-Q plot should fall along a straight line. However, if the residuals are not normally distributed, the points in the Q-Q plot will deviate from the straight line.