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 – Season, mnth, weathersit and weekday show significant relationship with the target variable for certain categories, by transforming these variables into dummy variables will be helpful for model training.

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

Ans – To remove the redundant variable, we must add drop_first=True.

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

Ans – temp and atemp variables show highest correlation with dependent variable.

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

Ans – By performing residual analysis and making predictions on test dataset, after estimating the r2_score of train dataset and test dataset which were close to each other.

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

Ans – temp, light rain with scattered clouds weather and year, these features are most significant in explaining demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans - Linear Regression is a supervised machine learning algorithm used for predicting a continuous numeric output variable based on one or more input features. It establishes a linear relationship between the input features and the target variable. The main goal of linear regression is to find the best-fit line (or hyperplane in higher dimensions) that minimizes the difference between the predicted values and the actual target values. It assumes that the relationship between input features and target variable can be represented by linear equation.

2. Explain the Anscombe's quartet in detail.

Ans - The datasets have identical statistical properties, including means, variances, correlations, and linear regression coefficients, yet they exhibit very different patterns when visualized. The purpose of Anscombe's quartet is to highlight the importance of data visualization in understanding and interpreting statistical relationships.

3. What is Pearson's R?

Ans - Pearson's correlation coefficient, often denoted as r or Pearson's r, is a statistical measure that quantifies the linear relationship between two continuous variables. It is used to determine how strongly two variables are related and the direction of their relationship. The coefficient ranges from -1 to 1, Pearson's correlation is widely used in various fields, such as statistics, data analysis, machine learning, and social sciences, to analyze the relationship between two continuous variables and to assess the strength and direction of that relationship.

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

Ans - Scaling is a preprocessing step in data preparation for machine learning models. It involves transforming the features of a dataset to a specific range or distribution to ensure that all features contribute equally to the model's training process. Scaling is performed to avoid biasing the model towards features with larger magnitudes, as some machine learning algorithms are sensitive to the scale of the input features.

Normalized Scaling (Min-Max Scaling):

In normalized scaling, the data is transformed to a specific range, usually between 0 and 1. The formula to perform min-max scaling on a feature x is:

$x_{scaled} = (x – min(x))/(max(x)-min(x))$

where x_{scaled} is the scaled value, x is the original value of the feature, min(x) is the minimum value of the feature, max(x) is the maximum value of the feature.

Standardized Scaling (Z-score Scaling):

In standardized scaling, the data is transformed to have a mean of 0 and a standard deviation of 1. The formula for z-score scaling on a feature x is:

$x_{scaled} = (x – mean(x))/std(x)$

where x_{scaled} is the scaled value, x is the original value of the feature, mean(x) is the mean value of the feature, std(x) is the standard deviation of the feature.

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

Ans - This occurs when there is perfect multicollinearity among the predictor variables in a multiple regression model. Perfect multicollinearity means that one or more of the predictor variables can be perfectly predicted by a linear combination of other predictor variables.

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

Ans - A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the data with the quantiles of the specified theoretical distribution. The main purpose of a Q-Q plot is to visually check if the data deviates significantly from the assumed distribution.