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:-Some of the variables such as hum,windspeed,weathersit,holiday has decreasing trend w.r.t cnt whereas some variables like season,mnth,yr,temp,atemp are having increasing trend w.r.t cnt

2. Why is it important to use drop_first=True during dummy variable creation?
   Ans :- drop_first=True helps in creating a more efficient and interpretable model by reducing multicollinearity and redundancy.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans:- Variable such as temp and atemp has highest corelation with target variable cnt

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans:- Validating the assumptions of Linear Regression is crucial to ensure the reliability and accuracy of the model.
   Linearity: The relationship between the independent variables and the dependent variable should be linear.
   Independence: The residuals (errors) should be independent.
   Homoscedasticity: The residuals should have constant variance at every level of the independent variables.
   Normality: The residuals should be normally distributed.
   No Multicollinearity: The independent variables should not be highly correlated with each other.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans:- Features lie temp,atemp,yr are the 3 features contributing towards the demand of the shared bikes.

General Subjective Questions

1. Explain the linear algorithm in detail.
   Ans:-Linear regression is a fundamental algorithm in statistics and machine learning used to model the relationship between a dependent variable and one or more independent variables

2. Explain the Anscombe's quartet in detail.

Ans:-Anscombe’s quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. These datasets are designed to illustrate the importance of graphing data before analyzing it and to show how different datasets can have identical statistical properties but very different distributions and appearances when plotted.

3. What is Pearson’s R?
   Ans:-Pearson’s R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol (r) and ranges from -1 to 1:

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans:- Scaling is a preprocessing technique used in machine learning to adjust the range of feature values in a dataset. This ensures that all features contribute equally to the model, especially when they have different units or magnitudes. Scaling is crucial for algorithms that rely on distance calculations or gradient descent optimization.
   Why is Scaling Performed?

Improves Model Performance: Many machine learning algorithms, such as k-nearest neighbors (KNN), support vector machines (SVM), and neural networks, perform better when features are on a similar scale.

Faster Convergence: For gradient descent-based algorithms, scaling helps in faster convergence by ensuring that the steps taken towards the minimum are consistent for all features.

Prevents Bias: Features with larger ranges can dominate the learning process, leading to biased models. Scaling ensures that each feature contributes equally.

Reduces Impact of Outliers: Scaling can help mitigate the impact of outliers by bringing all features to a comparable range.

Normalized Scaling vs. Standardized Scaling

Normalized Scaling (Min-Max Scaling)

- **Definition**: Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1].
- **Formula**: [x’ = \frac{x - \min(x)}{\max(x) - \min(x)}]

Standardized Scaling (Z-score Standardization)

- **Definition**: Standardization scales the data to have a mean of 0 and a standard deviation of 1.
- **Formula**: [x’ = \frac{x - \mu}{\sigma}] where (\mu) is the mean and (\sigma) 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:-An infinite value for the Variance Inflation Factor (VIF) indicates perfect multicollinearity among the predictor variables in a regression model. This situation arises when one predictor variable is a perfect linear combination of one or more 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 plot, or Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. It plots the quantiles of the sample data against the quantiles of a specified theoretical distribution.

Use and Importance in Linear Regression

In the context of linear regression, a Q-Q plot is primarily used to check the normality of residuals .importance as below

Assumption of Normality:.

Model Diagnostics: **Identifying Deviations**: