1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ans: cnt is more in fall season
cnt is much more in month of june, july, august, september
there is no much effect in cnt in weekdays
when weather is clear there is more cnt
cnt is more in 2019
cnt is more when its not holiday

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

ans : When creating dummy variables from categorical features, the drop_first=True parameter is used to address the issue of multicollinearity and ensure numerical stability in statistical models.

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

ans: temp variable is highly correleated

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

ans: by checking r2, low vif and p2

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

ans: temperature , year and holiday variables

General subjective questons

1. Explain the linear regression algorithm in detail.

ans: Linear regression is a widely used statistical algorithm for modeling the relationship between a dependent variable and one or more independent variables. It aims to find the best linear relationship that fits the data.

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 to demonstrate the importance of visual exploration and statistical analysis. Despite having similar summary statistics, each dataset has distinct characteristics, highlighting the limitations of relying solely on summary statistics to understand the underlying data.

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 between -1 and 1.

Pearson's R is widely used to assess the degree of association or correlation between variables in many fields, including statistics, social sciences, finance, and epidemiology

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 the process of transforming variables to a specific range or distribution. It is performed to bring all the variables to a similar scale or level of magnitude. Scaling is particularly important when dealing with features that have different units, ranges, or variances, as it helps to avoid biased or misleading results in data analysis and modeling.

Normalization rescales the variable to a range of 0 to 1, while standardization transforms the variable to have zero mean and unit variance. The choice between normalization and standardization depends on the nature of the data and the requirements of the specific analysis or modeling technique.

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

ans: The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a regression analysis. It quantifies the extent to which the variance of the estimated regression coefficients is inflated due to multicollinearity among the independent 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 (Quantile-Quantile plot) is a graphical tool used to assess the distributional similarity between a dataset and a theoretical distribution. It compares the quantiles of the observed data against the quantiles of a specified theoretical distribution, typically the standard normal distribution (mean 0, standard deviation 1).