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. Many Categorical variables had varying levels of significance but close to 10 have higher such as Spring, working day , sun, temp etc.

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

ANS. Because using drop_first=True ensures that unnecessary dummy features make it harder for algorithm to fit or make it easier to overfit.

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

ANS. temp variable

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

ANS. Assumptions of Linear Regression: 1.The error terms are normally distributed. 2.The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation. 3.The predicted values have linear relationship with the actual values

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
ANS. The top variables that are seen affecting and benefitting the Bike rental Count are :- 1.Spring Season 2.Working day 3.Sun

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   ANS. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

   While training the model we are given :

   x: input training data (univariate – one input variable(parameter))

   y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

θ1: intercept

θ2: coefficient of x

Hypothesis function for Linear Regression : y = $\theta_1 + \theta_2 . x$

2. Explain the Anscombe's quartet in detail.

ANS. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R?

ANS. Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1.

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

ANS. Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

Standardization (also called z-score normalization) transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

The point of normalization is to change your observations so that they can be described as a normal distribution. Normal distribution (Gaussian distribution), also known as the bell curve, is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.

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

ANS. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model

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

ANS. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. If the two data sets come from a common distribution, the points will fall on that reference line. The assumption of normality is an important assumption for many statistical tests; you assume you are sampling from a normally distributed population. The normal Q Q plot is one way to assess normality.