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? (3 marks)

- Season: When season is warm there are more booking as compare to colder season.
- Yr: there are more booking for 2019 then 2018 which show increase in customer base
- Mnth: same as season as month become warmer booking increase and as colder month approaches per-day booking starts decreasing gradually
- Holiday: days which has holiday has more number of per-day booking
- Weekdays: Weekday has almost similar pattern for booking and there median is almost same.
- Weathersit: same as season when weather is clear or partially clouded mean per-day booking is more then 75% of per-day booking count for cloudy days

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

Ans: drop_first will drop the 1st column so that we can k-1 dummy variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: atemp has the highest correlation with temp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: After finalizing the features. I have calculated the residual between the actual and predicted value of y_test and created a histogram of it. From the histogram, I can see that the mean is almost near 0(approx -1), and the error is normally distributed. Also, the VIF of all features is less than 5, and adj. R-square is quite high. From here we can stop and conclude that this is the final model.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: windspeed, month, and weathersit these 3 columns contribute significantly towards explaining the demand for the shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

ANS: Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting line through the data. The algorithm minimizes the sum of squared differences between the observed and predicted values. The equation of a simple linear regression model is often written as y=mx+b, where
y is the dependent variable,

x is the independent variable,
m is the slope, and
b is the intercept.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics but differ greatly when graphed. It highlights the importance of visualizing data and the limitations of relying solely on summary statistics. The quartet was created by Francis Anscombe to emphasize the need for graphical exploration in addition to numerical summaries when analyzing data.

3. What is Pearson's R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling? (3 marks)

ANS: Scaling is the process of transforming the numerical features of a dataset to a standard range. It is performed to ensure that no variable dominates the others due to differences in their scales. Normalized scaling brings the values within a specific range, typically [0, 1]. Standardized scaling (z-score normalization) transforms the data to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms sensitive to the scale of variables.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. VIF becomes infinite when perfect multicollinearity exists, meaning one predictor can be perfectly predicted by a linear combination of others. In such cases, the model cannot distinguish the individual impact of correlated predictors, leading to instability in coefficient estimation.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANS: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as a normal distribution. It compares the quantiles of the observed data to the quantiles of the expected distribution. In linear regression, Q-Q plots help to check the assumption of normality in the residuals. If the points in the plot fall approximately along a straight line, it suggests that the residuals are normally distributed, validating a key assumption of linear regression