Assignment-based Subjective Questions –

-Aishwarya Alex John

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

• Bike rentals increase in spring and summer but decline in fall and winter.

• Higher demand in 2019 compared to 2018.

• Peak demand from June to September; lowest in January.

• Rentals are lower on holidays.

• Demand remains consistent across weekdays.

• No major difference in rentals on working vs. non-working days.

• Highest rentals in clear/partly cloudy weather, followed by misty/cloudy, and lowest in light snow/rain.

Q2.Why is it important to use drop\_first=True during dummy variable creation?

* drop\_first=True
* It prevents unnecessary extra columns, reducing multicollinearity among dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?  
• temp and atemp are highly correlated, carrying similar information.

• total\_count, casual, and registered show high positive correlation.

Q4. . How did you validate the assumptions of Linear Regression after building the model on the training set?

* R-squared (0.81) was used, indicating the model explains 81% of the variance in the target variable.

Q5. **What are the top 3 significant features influencing bike demand?** (2 marks)

• Temperature

• Weather situation (weathersit)

• Year

*General Questions and Answers*

1. Linear regression algorithm. (4 marks)

• Linear regression computes the value of the dependent variable while using an independent variable and a linear relationship between the two.

• The formula:  
y = a + bx  
Where a is constant also called intercept, b is rate of change of x also called slope, x is IV also called independent variable, and y is dv.

1. State what Anscombe’s Quartet is. (3 marks)

• A collection of four datasets with almost the same summary statistics but different distributions - This demonstrates the need to visualize data prior to analysis.

1. Pearson’s R. (3 marks)

• It refers to the measure of linear correlation between two variables also known as Pearson correlation coefficient.

1. What is the meaning of scaling? Why is it necessary? Discuss the difference between normalized and standardized scaled. (3 marks)

• Scaling Adjusts data to appropriate levels preventing data bias.

• Normalization (min-max scaling): Rescale values into a range of 0 to 1.

• Standardization (Z-score Scaling): Transform the values into a normal distribution (mean=0, standard deviation=1).

1. value of VIF is infinite. Why does this happen? (3 marks

Infinite VIF occurs when independent variables are perfectly correlated (multicollinearity). This happens when  R^2 = 1 , leading to  VIF = \infty . Dropping one of the correlated variables resolves the issue.

6. **What is a Q-Q plot, and why is it important in linear regression?** (3 marks)

* A Quantile-Quantile (Q-Q) plot is used to compare two distributions. It helps to determine if the data follows a specific distribution, such as the normal distribution.