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)

Ans) Count (dependent variable) is very much dependent on all the categorical variables present in the dataset as inferred from the equation for the best fit line.

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

Ans) drop_first=True prevents the creation of extra column during dummy variable creation, thereby reducing the possibility of correlation between dummy variables.

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

Ans) Temperature(temp)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - The error terms must have constant variance - Validated using error terms histogram and scatter plots.
   - The independent variables should not be correlated – Validated using heat map and pair plot
   - There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s) - This was validated using predictions on test set and using various plots.
   - There should be no correlation between the residual (error) terms – This was validated using error terms scatter plot. No relation was observed between error terms.

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) Temperature, Weather and Year

General Subjective Questions

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

   Ans) Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression method which models a target variable based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

   It is used for predicting numerical values and modeling is done by obtaining a best fit line corresponding to the data and predicting dependent variable based on this information.

   It is one of the commonly used algorithms in machine learning.

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

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? (3 marks)

Ans) The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

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

Ans) Scaling helps to preprocess the data by normalizing it within a range. It is used to handle highly varying magnitude of data thereby making the process speed up.

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

Ans) Infinite VIF means that the independent variable is highly perfectly correlated to other independent variables and can be easily predicted from them

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

Ans) Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.