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)

-Less in spring compared to other season , September can be at its best with public advertisements

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

-To Remove extra column created during creation

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

-atemp & temp with value = .99

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

- Scatter plot between target and feature variable

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

Features	***VIF***
temp	*8.28*

windspeed 5.67
workingday 5.06
season_summer 2.62

General Subjective Questions

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

-Linear regression is a machine learning model , using 2 variables having linear co -relation
. Can be used to predict the behaviour
Mathematical equation is
Y=a+bx
it is a supervised learning, Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)

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

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistic

3. What is Pearson's R? (3 marks)
Pearson corelaation co-efficient -
The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation

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

-It is a data pre-processing , applied on independent variables to normalize the data in a range , it helps ML algorithm to do speed calculations

Collected data will contains features which highly varies in range, if scaling is not done, algorithm speeds up calculation on scaling which helps to bring variables in range, as scaling will not affect the parameters like P-values, R-squared etc. .

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

-If there is a perfect co-relation VIF will be infinite

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

-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. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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.