Assignment-based Subjective Questions

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

Answer:-

Seasonal Trends:
Bike demand experiences its peak in the fall, with higher levels of demand from **May to October**. This indicates that warmer months and possibly better weather conditions contribute significantly to bike usage. Conversely, there is a noticeable dip in demand during the spring.

Yearly Comparison:
In 2019, **bike demand was higher** compared to 2018, suggesting an upward trend or an increase in interest during that year. This could be attributed to various factors such as improved marketing, new infrastructure, or shifts in consumer behavior.

Weather Impact:
Clear or misty, cloudy weather generally sees higher bike demand, as cyclists are more likely to ride in these conditions. However, demand significantly drops when the weather is characterized by **light rain** or **light snow**, likely due to the inconvenience or discomfort these conditions cause for cyclists.

Weekday Consistency:
Bike demand remains **relatively constant throughout the weekdays**. There is no significant variation in demand on working days versus weekends, which could indicate that cycling for transportation, exercise, or leisure is consistently popular regardless of the day of the week.

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

Answer:-

Drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

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

Answer:-

The temp variable has the highest correlation with the target variable.

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

Answer:-

The predictor and response variables have a linear relationship.
• The error distribution's normality (the error terms' normal distribution).
• Homoscedasticity, or constant variance of the mistakes.
• Reduced feature multi-collinearity (low VIF)

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

Answer: -
The Top 3 features are temperature, the year and the holiday variables

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression is a supervised machine learning algorithm used to predict a dependent (target) variable based on one or more independent variables. It establishes a linear relationship between the target variable and the predictors. There are two types of linear regression:

1. Simple Linear Regression: Predicts the target variable using a single independent variable.
2. Multiple Linear Regression: Predicts the target variable using multiple independent variables.

The relationship between the variables is represented by a regression line. A positive linear relationship occurs when both variables increase together, while a negative linear relationship occurs when the dependent variable decreases as the independent variable increases.

2. **Explain the Anscombe's quartet in detail.**

   The four data sets in Anscombe's quartet have essentially the same basic descriptive statistics, but their distributions and visual representations differ greatly. Eleven points make up each dataset. Anscombe's quartet primarily aims to demonstrate the value of examining a set of data graphically before starting the research process because statistics alone cannot accurately depict the two datasets under comparison.

3. **What is Pearson's R?**

   The Pearson's Correlation A linear relationship between two values is established using the coefficient. It indicates how strongly two variables are related, and the coefficient's value can range from -1 to +1.

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

   Scaling is a pre-processing technique used to standardise the dataset's independent feature variables within a predetermined range while developing a machine learning model.
   There may be a number of features in the dataset that range widely in magnitude and unit. The units of all the features included in the model will not match if scaling is not done on this data, which results in inaccurate modelling. The distinction between standardisation and normalisation is that the former replaces the values with their Z scores, whilst the latter sets all of the data points in a range between 0 and 1.

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

   When the two independent variables have a perfect correlation, the VIF value is infinite. In this instance, the R-squared value is 1. Since VIF is equal to 1/(1-R2), this results in VIF infinity. According to this theory, in order to create a functional regression model, one of these variables must be eliminated because multi-collinearity is an issue.

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

The quantile-quantile (Q-Q) plot compares the quantiles of a sample distribution to a theoretical distribution to assess if the dataset in question follows a normal, uniform, or exponential distribution. It helps us establish whether two datasets have the same distribution. It also helps to determine whether or not the errors in the dataset are normal.