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)

   The demand of bike is less in the month of spring when compared with other seasons.

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

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

   According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

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

   Temperature (0.552); weathersit: Light Snow, Light Rain + Thunderstorm + Scattered clouds; Humidity: Constant

General Subjective Questions

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

   Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

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 statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson’s R? (3 marks)

   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)

   It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

   It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

   Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution. Normalization scales in a range of [0,1] or [-1,1]. Standardization is not bounded by range. Normalization is highly affected by outliers.

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

   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.

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

   The Q-Q plot is **a graphical plotting of the quantiles of two distributions with respect to each other**. In other words, we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line.
   Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this **helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential**.