Assignment-based Subjective Questions

1. From your anaxlysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Few of the categorical variables have a significant effect on the dependable variable. They are:

   - **7Jul**: A coefficient value of '-0.047814' indicated that a unit increase in 7Jul variable, decrease the bike hire numbers by 0.047814 units.
   - **9Sep**: A coefficient value of '0.096174' indicated that, a unit increase in 9Sep variable increases the bike hire numbers by 0.096174 units.
   - **Light_rainsnow**: A coefficient value of '-0.231830' indicated that a unit increase in Light_rainsnow variable, decreases the bike hire numbers by 0.231830 units.
   - **Misty**: A coefficient value of '-0.050192' indicated that, a unit increase in Misty variable decreases the bike hire numbers by 0.050192 units.
   - **Summer**: A coefficient value of '0.081529' indicated that a unit increase in Summer variable increase the bike hire numbers by 0.081529 units.
   - **Winter**: A coefficient value of '0.134695' indicated that a unit increase in Winter variable increases the bike hire numbers by 0.134695 units.

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)

   Temperature

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

   Steps to validate assumptions of linear regression model are:

   - **Linear Relationship:** A scatter plot was plotted between one independent and one dependent variable, a straight line passing through the points could be observed.
   - **Homoscedasticity:** Variance of error terms was observed and found that the variance of error terms is constant.
   - **Absence of Multicollinearity:** Heatmap and VIF was used.
   - **Independence of residuals:** Durbin Watson test was conducted.
   - **Normality of Errors:** Histogram and QQ plots were used.

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

   A. Temperature
   B. Year
   C. Light_rainsnow (Weatherit = 3)

General Subjective Questions

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

   As per the definition given by Yale University, Linear regression attempts to model the best relationship between variables by fitting a linear equation to observed data. One variable is considered to be a dependent variable, and the others are considered to be explanatory variable.

   The linear regression algorithm consists of following steps:

   1. **Analysis and conversion of variables:** Variables must be converted to required format, ie, Conversion of Categorical variables. Analysis of variables, to understand correlation and directionality of the data.
   2. **Dividing the model into test and train dataset:** The data set must be divided ideally in 70-30 proportion. This is done to check the predictive capacity of final regression model.
   3. **Estimating the model, i.e., fitting the line:** A final model is estimated which has the best representation of maximum points in a linear line. After developing the model, we check the assumptions of linear regression model to determine usefulness of the model.
   4. **Evaluating the validity and accuracy of the model:** The model is run of the test dataset to obtain the R^2 and other factors.

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

   Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions when graphed.
   There are four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This explains about the importance of visualising the data before applying various algorithms to build models.
   One out of Anscombe's quartet suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient is known as best method of measuring the association between variables of interest because it is based on the method of covariance. It is the test statistics that measures the statistical relationship, or association, between two continuous variables. It gives a multidimensional information - the magnitude of the association, correlation and the direction of the relationship.

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

Scaling is a part 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.
This is done when the collected data set contains variables with highly varying magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. This brings all the variables to the same level of magnitude.
Normalization scales a variable to have a value between 0 and 1, While standardization transforms data to have a mean of zero and a standard deviation of 1.

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
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.