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?
    a. Demand for bike is less during the spring season.
    b. Demand for bike is maximum during fall.
    c. Demand for bikes increases during April – October.
    d. Demand for bikes is low when the weather situation is rainy or snow.
    e. Demand for bikes is similar during the weekdays.

2. Why is it important to use **drop_first=True** during dummy variable creation?
    a. Dropping first column helps in reducing collinearity between the dummy variables.
    b. To achieve k-1 dummy variables which can be used to delete the extra column created during generating dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable.
    a. The two fields atemp and temp have same correlation with the target variable 0.63, which is the highest.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. By plotting the residuals.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes.
    a. The top 3 features contributing significantly are:
        i. atemp
        ii. yr
        iii. weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. A linear regression algorithm defines the relationship between independent and dependent variables using a straight line. All the variables used to perform linear regression are numeric only.
        i. We have divided the data into training & testing data.
        ii. Training data is further divided into features and target datasets.
        iii. A linear model is fitted by using training data, the gradient decent algorithm uses the coefficient to draw a straight line which can be considered as the best fit line.

2. Explain the Anscombe's quartet in detail.

   a. Anscombe’s quartet consists of four datasets that have same statistics but different distribution, the statistics contains variance of x and y, linear regression line, R-square, mean and correlation coefficient, the quartet showcases that even if the statistical differences between the dataset are similar the graphical representation can be different.

3. What is Pearson’s R?
   a. Pearson’s R measures the strength of association between two variables.
   b. The standard deviation values rely between -1 to +1
      i. +1 means positive linear correlation.
      ii. -1 means total negative correlation.
      iii. 0 means there is no correlation between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   a. Scaling is a pre-processing step in linear regression, we scale a variable to improve computation of gradient decent faster.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   a. VIF becomes infinite if the R-square is 1, which reflects that there is a perfect correlation between the features.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   a. Q-Q plot is a scatter plot of two sets of quantiles against each other, if the data is from same source the visual plot will appear as a line, the main purpose is to check if the two sets of data come from the same distribution.