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. During Spring and winter seasons the demand is less
    b. During summer and fall the demand is more
    c. The demand is more in working day than in holiday
    d. If weather sit is (Clear, few clouds, Partly cloudy, Partly cloudy) the demand is more and Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds the demand is less
    e. The Demand is increasing every year (from 2018 to 2019)
2. Why is it important to use drop first=True during dummy variable creation?
    a. It helps in reducing the extra column created during dummy variable creation. And so 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?
    a. "temp" column has a highest correlation with the target variable
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. Plot the error or residuals (y train – y predict) and check whether it is normally distributed and its mean is zero
    b. Plot the error and variable X or Y there should not be any pattern in scatter plot . The data should be scattered.
    c. From R2 value we will understand how much variance in dependent variable is addressed by the independed variable
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. "temp" - if the temp is very low as less than 12deg people don't prefer to come out so demand is less and vice versa
    b. "yr" - The demand seems to be increasing every year
    c. "weathersit" – if the weather is rainy and snow people don't prefer to come out instead if the weather is clear.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Linear regression analysis is the process of predicting the depended variable trend or values using independent variables
    b. It fits a straight line that minimizes the predicted variable and actual output values.
    c. There are two types
        i. Simple linear regression - The number of independent variable is one
        ii. Multiple linear regression – The number of independent variables is more than one
    d. Simple Linear regression –
        i. Assumptions:
            1. Linear relationship between X and y.

2. Normal distribution of error terms.
3. Independence of error terms.
4. Constant variance of error term

ii. Build a linear model using least squares
iii. Check the p value(should be less than 5%) and r2 values (more than 75%) for a best fit
iv. Residual analysis – check whether it is normal from histplot and is there any independence in x and y

e. Multiple linear regression – MLR helps us to understand how much will the dependent variable change when we change the independent variables.
   i. Handle the categorical variables to a meaning full insights
   ii. Remove the multicollinear variables
   iii. Use a scaling for numeric variable and dummy variable for categorical variable
   iv. Build the base line model with OLS for all independent variables
   v. Now do a trial using RFE for feature selection
   vi. Compare the R2 value , IVF(less than5) and P value(less than 5%)
   vii. Drop the variable which has IVF greater than 5 and re-model again
   viii. Plot the error and check whether it is normally distributed
   ix. Perform prediction using test data set and compare the R2 with the final model of trained set

2. Explain the Anscombe's quartet in detail.
   a. Even though the statistical summary (mean , median , variance and so on) for the data sets are same if we visualize the same in scatter plot there is a possibility of having completely four different distribution
   b. Anscombe's quartet is to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties

3. What is Pearson's R?
   a. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations
   b. The values will be between -1 to +1 (-1 & +1 means best fit, 0 means not a good fit)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   a. Scaling and why it is performed - scaling is a technique to make them closer to each other or we can say that the scaling is used for making data points generalized so that the distance between them will be lower.
   b. Normalization or min max – Is used to transform features to be on a similar scale.
      i. Normalization is useful when there are no outliers
      ii. After scaling all the data in respective columns will be between 0 and 1
   c. Standardized scaling - is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.
      i. It is used when we want to ensure zero mean and unit standard deviation.
      ii. It is useful when the feature distribution is Normal or Gaussian.

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

a. If we have perfect correlated feature (the correlated between the feature to the other is 100%) means R2 = 1.
b. VIF = 1/1-R2 , if R2 =1 then VIF will be infinite**.**

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 similar kind of scatter plot
b. These are plots of two quantiles against each other
c. The purpose of Q Q plots is to find out if two sets of data come from the same distribution
d. In LR it is used to ensure the distribution of the error terms or prediction error using a Q-Q plot. The distribution should be a normal distribution