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?
   Answer: Following are the analysis points on the effect of categorical variables:
   - Summer and fall seasons have a greater number of users than other two seasons
   - User's count was more in year 2019
   - User count decreases as the weather situation becomes worse
   - Most number of user's count was from May to October month

2. Why is it important to use drop_first=True during dummy variable creation?
   Answer: By dropping one variable, it reduces the number of final variables. The value for the dropped variable can be found out based on the other variables hence it reduces the complexity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer: "atemp" variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Answer: By two plots:
   - First, I plotted a histplot/distplot for residuals/error terms which was normally distributed, and mean was at 0
   - Second, I plotted a scatter plot of residuals/error terms along with y_train data which showed that error terms were independent of the y_train

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer: "atemp", "year", and "weekday_sat"

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Answer: Linear regression algorithm is a machine learning algorithm. It comes under supervised learning where target variable is known/labelled.

   It predicts continuous/regression value/variable which is called dependent variable. Prediction happens based on independent variables. As the name suggests, prediction happens by a best fit line between independent and dependent variables.

   There are two categories of Linear Regression:
   - Simple Linear Regression
   - Multiple Linear Regression

   If there is one independent variable, then it is simple linear regression. In case of multiple independent variables, it is called multiple linear regression. In mathematical terms, simple

linear regression is represented by y = mx + c and multiple linear regression is represented by y = m1x1 + m2x2 + m3x3 + …… + mnxn + c.

Target variable in linear regression is always continuous whereas independent variables can be continuous or categorical.

Best fit line is found by minimizing the error between actual values and predicted values of target variable thus by finding the best values of slope and constant which can explain the relationship between independent and dependent variables. To find the best fit line and to minimize the error, Root Mean Squared Error function is minimized which is the cost function in case of linear regression. Most of the model used gradient descent to find the optimal values of slope and constant.

2. Explain the Anscombe's quartet in detail.
   Answer: Anscombe's quartet defines that only the basic statistic properties of a data set is not sufficient to explain it but plotting it graphically is more important for better and correct analysis of a data set.

   Anscombe explained it by having four data sets having 11 values in each of them. X values for three of these four data sets were identical. These data sets were having almost similar basic statistic properties. Still if these data sets were plotted graphically, then they were entirely different.
   One of them showed linear relationship between X and Y variables.
   Second one showed nonlinear relationship.
   Third one showed almost perfect linear relationship except for one outlier.
   Fourth one showed that only one point is enough to show high correlation coefficient between variables whereas other data points showed no relationship.

3. What is Pearson's R?
   Answer: Pearson's R is Pearson's Correlation Coefficient which measure the linear correlation between two data sets. It ignores other type of relationship between the two data sets and its value ranges between -1 to 1. A positive value means that with an increase in x, y will also increase and vice versa. A negative value means that with an increase in x, y will decrease and vice versa.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer: Scaling is done to get all the independent variables at the same range and normalize them. In linear regression it helps the gradient descent to find the minima faster. Normalized scaling is also called as Min Max scaling, and it scale the data between a range normally between 0 to 1. It is the simplest method. Whereas Standardized scaling distributes the data such that mean becomes 0 and standard deviation is 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer: As per VIF formula this will happen when R-squared value will be 1. R-squared value 1 will mean that variance in the variable having VIF value as infinite can be 100% explained by other variables. Basically, the correlation is perfect among the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Answer: Q-Q plots are quantile-quantile plots. Quantile of a data set can be obtained by sorting the values and then check how many values fall under a particular quantile value. Quantile of a sample distribution is plotted against the quantile of theoretical distribution to determine if the two data sets follow the same type of distribution or not. For example, if the data sets follow a normal distribution or uniform distribution.
   In linear regression one of the assumptions is that error terms follow a normal distribution so Q-Q plot can be utilised to prove the same.