1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ANS: - the demand for bikes was least in the season of spring and high during summer and fall.
- 2019 saw considerably high demand as compared to 2018.
- the demand for bikes increases gradually from jan till september and then decreases again.
- demand on holidays is considerably lesser.
- least demand of bikes in Light rain/snow weather condition.

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

ANS: categorical variables can be worked with n-1 dummy variables if there are n categories.

E.g. for the category variable weathersit

`000` will correspond to fall

`001` will correspond to winter

`010` will correspond to summer

100 will correspond to spring

Thus 4 categories can be identified with just the last three columns where. Hence the first column is dropped

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

ANS: The variable temp has the highest correlation with cnt.

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

ANS: by plotting the distplot of error terms and confirming that the mean of the error is 0.

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

ANS: Temp, Mist + Cloudy and windspped

General Subjective Questions

1. Explain the linear regression algorithm in detail.

ANS: Linear Regrssion algorithm is centred around the equation of a line given by y=B0+B1X. Y being the dependent and X as the independent variables.

The idea is to find the best fitted line such that the squares of the difference between the Predicted value of Y is and original Y (Mean Squared Error) is minimum. This is achieved by obtaining the values of coefficients B0 and B1 using gradient dessent algorithm.

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

ANS: Anscombe's quartet is a group of datasets that have the same mean, standard deviation, and regression line, but which are qualitatively different.

3. What is Pearson’s R?

ANS: The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It measures the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
ANS: Scaling is performed since some variable is on a different scale with respect to all other numerical variables which take very small values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.Differenence between normalization and standardization is that normalization brings the value between 1 and 0 while with standardization then mean of the column becomes 0 and starndard deviation 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
ANS: The value of VIF is infinite when there is perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
ANS: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. In linear Regression it is used in model evaluation by plotting the y_test against y_test_pred.