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. There is improvement in bike sharing in 2019 vs 2018
   b. People are less likely to opt for Bike sharing on Holidays
   c. Fall season seems to have more subscriptions.
2. Why is it important to use drop_first=True during dummy variable creation?
   a. The get dummies will create n variables if there are n different levels in a variable. But this adds multicollinearity in the data as any newly added level can be obtain using the other variables (1 – sum of the rest). This inflates the coefficients in Linear models. Hence, it is better to remove of the levels using the drop_first parameter.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   a. 'atemp'
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   a. The linearity can be validated by plotting the residuals and checking is there is no non-linear pattern
   b. The normality assumption can be validated by plotting the distplot
   c. The other two assumptions can be validated using looking at the variance of the residual plots where there should not be a pattern and variance should not increase/decrease based on residue.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   a. Year, Month and Weather

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   a. Linear regression is a statistical technique of finding the relationships between independent and dependant variables(as coefficients) assuming that the relationshil between the independent and dependant variable in linear in nature. It uses Ordinary Least squares to attain the best possible coefficients.
2. Explain the Anscombe's quartet in detail.
   a. A combination of four X-Y combinations where all the statistical properties will be same between them but the actual distributions are quite different.
3. What is Pearson's R?
   a. Pearson's R represents how strong is a linear relation between to variables
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   a. Scaling brings the numeric variables in a dataset to the same range so that the model treats them with same priority. Normalized Scaling will bring values between the range of 0 and 1. Standardized Scaling will bring the values to a mean of 0 and standard deviation of 1

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

   a. This refers to perfect correlation between two variables.

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

   a. This is a graphical way of validation the ‘normal’ assumption of Linear regression. To validate the assumption, the plot has to be linear especially at the tails.