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?

[Naresh]

I. Fall has the highest median, which is expected as weather conditions are most optimal to ride bike followed by summer
II. Clear weather is most favourable for bike renting as temperature is decent, humidity is less, and temperature is less.
III. Total spread for the month plot reflects that of season plot as fall months have higher median
IV. Median bike rents are increasing year on year (i.e., 2019 has a higher median than 2018)

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

[Naresh] This is an important step during the creation of dummy variables as it reduces the additional columns generated in the process as well 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?

[Naresh] The temp variable has good correlation with cnt variable, and the scatter plot demonstrates When the temp variable tends to increase as the cnt variable increases, there is a positive correlation between the variables.

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

[Naresh] The four principal assumptions which justify the use of linear regression models for purposes of inference or prediction:

i. linearity and additivity of the relationship between dependent and independent variables:

   (a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.
   (b) The slope of that line does not depend on the values of the other variables.
   (c) The effects of different independent variables on the expected value of the dependent variable are additive.

ii. statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)
iii. homoscedasticity (constant variance) of the errors

   (a) versus time (in the case of time series data)
   (b) versus the predictions
   (c) versus any independent variable

iv. normality of the error distribution.

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

[Naresh] Based on final model top three features contributing significantly towards explaining the demand are:

I. Temperature (temp) - A coefficient value of '0.5209' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5209 units.
II. Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2292' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2292 units.
III. Year (yr) - A coefficient value of '0.2250' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2250 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the most popular algorithms in Machine learning landscape and a simple one to begin. It's a statistical method used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

Basically, this algorithm demonstrates a linear relationship between a dependent (y) and one or more independent (y) variables. In simple terms it finds how the value of the dependent variable is changing according to the value of the independent variable.

2. Explain the Anscombe's quartet in detail.

[Naresh] Anscombe's quartet are a group of four data sets which look similar in simple descriptive statistics. However, there are some differences in the dataset that deceives the regression model if built. They appear different when these datasets are scatter plotted and they have very different distributions.

3. What is Pearson's R?

[Naresh] It's a numerical summary of the strength of the linear association between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

[Naresh] Scaling is a data pre-processing step applied on independent variables to normalize the data within a certain range. During collection of data the quantities and units of the variables are not same scale. This scaling operation standardize the variables to common that can be used for modelling.

The scaling doesn't impact the metrics (i.e., t-statistic, F-statistic, p-values, R-squared, etc) used to study the data and model driven results.

The two types are scalers are

1. MinMaxScaler: It brings all the data in the range of 0 and 1.
2. StandardScaler: It brings all the data into a standard normal distribution which has mean zero and standard deviation one.

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

[Naresh] VIF is infinite only If there is perfect correlation i.e., a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

[Naresh] It's a graphical representation of data from the distributions such as Normal, exponential or uniform. The Q-Q plot helps to determine the similarity of data sets.