Subjective Questions Answers

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

   From all of the categorical variables, the following variables has the most effect on the dependent variable – cnt

   a. Season
   b. Yr
   c. Weather condition
      i. Clouds
      ii. Light_rain
   d. Holiday – whether that day is a holiday or not

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

   To eliminate the redundancy which will be there if all the three variables are being used.
   Two variables are sufficient to define the information completely and the third (nth) variable does not add any new information.

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

   As per the pairplot the strongest correlation was of registered with the target variable total count
   At second number the correlation of the felt temperature (atemp) was second.

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

   The assumptions of the linear model was fielded against the test data set and then the comparison between the y_test and Y-pred was done. It was showing a great accuracy

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

   Season

Year
Holiday

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear Regression is one of the simplest type of machine learning algorithm which tries to establish a linear relationship between the predictor variables (variables whose effect we are studying) and the predicted variable (variable on which the effect we are studying).

If the effect of only one dependent variable is being studied it comes under linear regression and can be plotted on 2-D graph.

However with increased number of variables it becomes difficult to visualize the predicted geometry.

The prediction is done by the algorithm in such a way that the difference from the predicted values and the actual values referred to as residual / cost-function is minimum.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a famous set of four datasets, each with four variables: x and y, as well as x1 and y1. Despite having nearly identical summary statistics (mean, median, variance, etc.), these datasets look dramatically different when visualized. This highlights the importance of visualizing data before drawing conclusions, even when numerical summaries seem to tell a consistent story

3. What is Pearson's R?

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1:

r = 1: Perfect positive correlation, meaning the variables increase or decrease together perfectly.
r = -1: Perfect negative correlation, meaning one variable increases as the other decreases perfectly.
r = 0: No correlation between the variables.

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

Scaling is a process of transforming data to a common scale, usually between 0 and 1 or -1 and 1. → min max / mean s.d. scaling

Scaling is performed to ensure that the coeffecients values are in the similar range and there are no wild variations when testing the predicted values

Normalized scaling scales data between 0 and 1 called as min max scaling
Standardized scaling scales the data such tha mean is 0 and standard deviation is 1.