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

   >>

- Bike demand is less during spring
- Bike demand is highest during summer
- Bike demand is less during heavy rain/thunderstorm
- Bike demand is growing year on year.

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

   >> Because values in rest of the dummy variables can be used to describe first variable.

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

   >> Temperature has highest correlation with cnt

4. How did you validate the assumption of Linear regression after building model on the training set

   >> We tested it against the test data to see if training model is able to predict accurately. We plotted test data against prediction to validate if spread is linear in nature, and also calculated R-square value and compared it with R-square calculated in linear model of training data.

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

   >> Season (less in winter, more in summer)

   >> Weather condition (more in clearer sky day, and less in heavy rain/show etc)

   >> Year. It is growing each year

General Questions

1. Explain linear regression in details.
   >> Linear regression is a way to build a prediction model for a dataset that has linear relations ship between dependent variable and independent variables. E.g Sales could grow somewhat linearly with good advertising spend.
   Linear regression tries to exploit this linear relationship to build a model that can extrapolate values dependent variable for any given independent variable.
   In this we establish a mathematical relationship with variables in dataset, try to draw a line that best fits and then use it for prediction.

2. Explain Anscombe's quartet in detail:
   >> Anscombe's quartet consists of four different data sets each containing 11 points and 2 variables (namely x and y) with identical statistics. But when viewed graphically, it shows big difference although being identical in statistical values.
   It stresses the importance of looking at the dataset graphically before analysis

3. What is pearson's R?
   >> It (Pearson's Correlation coefficient) is way of measuring linear correlation between 2 variables.

4. What is scaling and why is it performed? What is difference between normalized scaling and standardized scaling?
   >>Scaling is a process of fitting values of variables into same range. E.g. Temperature of 5 cities can vary between 0 to 45 degrees, and population of these 5 cities can vary from thousand to millions, but scaling can fit all these values between 0 and 1 (0 corresponding to lowest and 1 corresponding to highest value). This is used for improved model performance.

   In normalized mapping we map lowest value to 0 and highest value to 1, and in standardized we don't enforce this defined range. In standardized, we transform to have a mean of zero with a std deviation of 1.

5. You might have observed that sometimes VIF is infinite, Why does this happen?
   >> This situation occurs when independent variables are orthogonal to each other, i.e. perfectly correlated to each other.

6. What is a Q-Q plot? Explain the use and importance of it in linear regression?
   >>A Q-Q plot in linear regression is used to validate if residuals of model are normally distributed or not. It's a scatter plot that shows relation between sample data and corresponding percentile of an independent variable.