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

   A. Categorical varaibles are mostly needed to explain the significance of the model. In bike sharing dataset we have few categorical variables like
      yr,mnth,season,weathersit,holiday,weekday,workingday.

      yr: The users are increased from 2018 to 2019.
      mnth : during the month of september there is increase in bike booking demand.
      weekday : Almost all days has same pattern.
      workingday: Almost all days has same patterns.

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

    A. During dummy varaible creation drop_first = True drops the extra column that was created. As a result it reduces the correlation created among dummy variables.
       To represent a dummy variable we need to represent by n-1 levels of categorical variables with n levels.

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

    A. Temp and atemp numeric feature are highly correlated to one another with the dependent variable 'cnt'

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

    A. we have plotted a distplot between target trained variable and target predicited variable, which gives us a residual errors.
       The distribution of erros should be nomal and centred at 0.

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

    A. temp
       yr
       windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

    A. Linear Regression is a statistical model which analyses the linear relationship between dependent and independent varaibles.
       Mathematical representation of relationship between variables:
       y = mX + c
       where y is the dpendent variable we are trying to predict.
             m is the slop of the regression line.
             c is constant, known as y-intercept.
      Linear Relationship can be categorised as positive and negative Linear Relationship.

Positive Linear Relationship:
-> A linear Relationship is called as positive if both dependent and independent varaibles increase along with another.

Negative Linear Relationship:
-> A linear Relationship is called as negative if independent variables increase and dependent variables decreases.

Linear Regression is classified into two types:
-> Simple Linear Regression --> Prediciton of model based on one dependent and one independent varaible
-> Multiple Linear Regression --> Prediciton of model based on one dependent and multi independent varaibles

2.Explain the Anscombe’s quartet in detail?

A. Basically it has four datasets where they have identical statistical properties.
But when you plot them in a graph it visually appears in different shape. Each dataset has 11 x,y points.
It is constructed mainly to view the importance of graphing the data before analysing it and the effect of outliers on the properties.

3. What is Pearson’s R?

A. It is a numerical summary of the linear associations between the variables. if the variable tend to go up and down together, the correlation coefficient will be positive.
If the varaible tend to go up and down in opposite direction with low value of one associated
with high value of another, the correlation coefficient will be negative.
The pearson correlation coefficient, r can take a range of value from -1 to 1.
A value 0 indicates that there is no association between the two variables.

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

A. Scaling a technique to standardize independent features in a data in fixed range.
It is performed during data preprocessingto handle data with very high magnitudes or units.

Normalized Scaling --> Min and Max values of features are used for scaling. It is affected by outliers. it is used when features are of different scales.
The values for scaling a variable lies between [0,1] [-1,1]

standardized Scaling --> Mean and Standard Deviation are used for Scaling. It is much less affected by the outliers.It is used when we want zero mean and unit Standard Deviation.

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

A. This happens when there is a possibilty of perfect correlation, then VIF =

infinity.
In case of perfect coreelation we get R2 value is 1, which makes 1/(1-R2) infinity.
if these kind of scenario happens, then we can drop a variable which causing perfect Multicollinearity.

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

A. The quantile- quantile(q-q) plot is a graphical technique for determining if two data set come from
population with a common distribution.
A q-q plot is a plot of quantities of first data set against the quantiles of second dataset.
A 45-degree reference line is plotted. If the two sets come from a population with the same
distribution, the points should fall approximately along the reference line.