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

   Ans. As we have seen in dataset year, holiday, season, weathersit are more correlated categorical variables with dependent variable in dataset. We have converted categorical variables into one hot encoding.

   As per my understanding, p-value for year and weathersit is more than 0.05 so we should remove it. But after applying VIF, removed other variables with more than 5 value. Year and weathersit's p-value automatically became less than 0.05. So not need to remove them.

2. **Why It is important to remove drop_first = True during dummy variable creation?**

   Ans: We should always use **drop_first = True** during dummy variable creation. Because whenever we apply dummy variable creation, All the categories convert to columns. If we remove any one from them. Remaining columns with all zero values in row indicate dropped columns. So no need to keep all columns. For machine learning it is better to have less columns.

3. **Looking at pair plot among the numerical variable which one has highest correlation. With the target variable.**
   Ans: registered variable has more correlation with dependent variable

4. **How did you validate assumptions of linear regression after building the model on training data set?**

   Ans. Validation of linear regression model. We should take care of below steps.

   A. Adjusted R-Squared value should be close to 1
   B. No input variable should be with greater than 0.05 p-value
   C. Correlated input variable wit dependent variable.
   D. Before checking summary, first check VIF values of all variables. If anyone of then has mode then 5 then remove it and train and check summary again.

5. **Based on final model which are top 3 features contributing significantly towards explaining demand of bikes.**
   Ans: causal, year and season has more coefficient with less or zero values. So looks these three are most contributor in bike demand prediction.

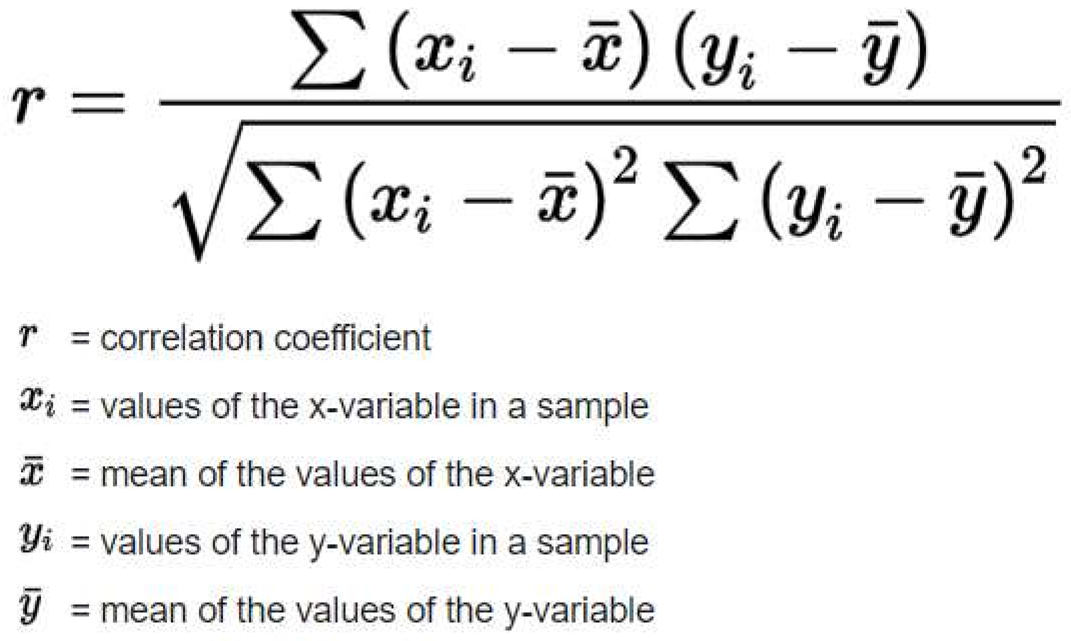
6. **Linear regression algorithms:** It is a supervised machine learning algorithm used for solving regression problems. This algorithm fit input data with its output data. So that after completing training on history data, we can get prediction for new data.

   Y = b1x1 + b2x2 …….. + b0

B1, b2…. Are coefficient and b0 is intercept. In training of linear regression model, using cost function it find out best values for coefficient and intercept.

7. **Anscombe's quartet:** 4 datasets of 11 point. When council analysis this dataset they found mean, standers deviation and correlation between them.

8. Pearson's R: It is also called Pearson's correlation coefficient. It is used to find correlation between two variables. We can find negative and positive correlation using it. Below is formula.



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

9. **What is scaling why is scaling performed. What is difference between normalization and standardisation scaling?**
   In linear regression technique, we find out coefficient and intercept values. Before finding coefficient values, data should be in same scaling. So that if we multiply a coefficient value it impact equally to the output variables.
   Normalisation: it fix a range between 0 and 1.
   Standardisation: range of values are standardised, how many standard deviation far from mean. Normally mean is zero and standers deviation is 1

10. **You might have observed that some time VIF values are infinite. What does it mean?**
    And: VIF is used to find multicollinearity in the independent variables. High value of VIF mean more correlation two variables. Infinite means 100% correlation between two variables.

11. **Quantile-Quantile plot,** is a plotting tool to find if a set of data came from some theoretical distribution such as a Normal, exponential or uniform distribution.
    Also, it helps to determine if two data sets come from populations with a common distribution.
    This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.