1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    a. The season variable in the dataset have a significant and positive influence on the dependent variable cnt the changes in the seasons are drastically changes the value of the cnt
    b. The coefficient for the 'holiday' variable is negative, but the p-value is relatively high (0.077). This suggests that there may be a slight decrease in bike rental count on holidays compared to non-holidays
    c. The pvalue for the weekdays is significantly high this means that there is a slight increase in the rental count on weekdays compare to the weekends

2. Why is it important to use drop_first=True during dummy variable creation?
    a. The main aim of dropping the first dummy variable is to prevent the multicollinearity
    b. Adding more dummy variables will effect the model complexity adding a dummy variable without any meaning or where the variable can be predicted from the other variable is just adding more complexity to the model
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    a. The registered variable is having highest correlation with the target variable also the temperature variable is having correlation
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. Validating the linearity of the variables with the targeting variable
    b. Calculate VIF and check and prevent the multicollinearity
    c. Scale all the outliers variables
5. Explain the linear regression algorithm in detail.
    a. Linear regression algorithm is a method to find the relationship between one dependent variable and one or more independent variable by creating a linear relationship between the dependent and the independent variable
        i. The linear regression is required to find the coefficients of the linear regression, in order to find the best coefficient the algorithm minimises the sum of squared differences between the observed and predicted values. This is mostly done using the method of least squares method. The algorithm adjusts the coefficients to minimise the residual sum of squares (RSS) or the mean squared error (MSE) between the predicted and actual values.
        ii. There are multiple steps involved in order to build and train the model as follows

Data preparation

- Data quality checks
- Categorical variables
- Creation of Dummy variables
- Data Cleaning
- Deriving new matrics

Model Building

- Parameter tuning
- Variable selection

Model evaluation

- Residual analysis
- Model evaluation

iii. Once the model is build and trained successfully this model can be used to make predictions on the new data by inputting the independent variable

6. Explain the Anscombe's quartet in detail
    a. **Anscombe's quartet** comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*, *y*) points. Source: (Wikipedia)
7. What is Pearson's R?
    a. Pearson's R is a measurement of the linear relationship between two variable it indicates the strength and direction of the association between the variables
    b. 1 indicates positive relation which means the variables changes proportionally with one other
    c. -1 indicates negative relation
    d. 0 indicate no relation
8. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling is the process of transforming data to a common scale without distorting differences in the ranges of values. It is a crucial preprocessing step in data analysis and machine learning for interpretability
    b. Scaling helps to remove the outliers and ensure all variables are scaled equally by applying similar scale
    c. Normalised Scaling: Normalized scaling rescales the data to a fixed range, typically between 0 and 1. MinMaxScaling is a normalised scaling
    d. Standardized scaling: Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1. Standaridization is an example for this
9. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. The Variance Inflation Factor is used to assess multicollinearity in regression model. It indicates the multicollinearity between the variables in the model. A VIF value greater than 10 is often indicates high multicollinearity. When the value of VIF is infinite, it indicates perfect multicollinearity among the predictor variables. Perfect multicollinearity occurs when one predictor variable can be exactly predicted by a linear combination of other predictor variables in the model.
    b. It can be occurred due to the overfitting of the model to the training data
    c. Or if the dummy variables are not properly added like if one dummy variable is predicted from the other one
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

a. Q-Q plot is an important tool in analysing the relationship between the variable in a linear regression model
b. By using a Q-Q plot we can access whether a dataset follows any specific probability distribution
c. A Q-Q plot can be used for
d. **Normality Assumption**: Linear regression models often assume that the residuals are normally distributed.
e. **Identification of Outliers**: Outliers in the data may cause deviations from the expected straight line pattern in the Q-Q plot.