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

---> 1. we get to know that most people use bike summer, fall and winter
2. in 2019 the number of bike user has increased
3. from 4th month to 10th month bikes users are most
4. bike users uses bike wheather it's working day or not
5. bike users mostly use bikes on Clear, Few clouds, Partly cloudy, Partly cloudy and least use on Light Snow, Light Rain + Thunderstorm +
Scattered clouds, Light Rain + Scattered clouds
6. most people use bikes when i's not holiday

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

---> Setting drop_first=True during dummy variable creation is important to avoid the problem of multicollinearity in the model, which occurs
when two or more predictor variables are highly correlated. By dropping one of the dummy variables, we avoid perfect multicollinearity and
make it easier to interpret the effects of the predictor variables.

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

---> The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

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

---> Linear Regression models are validated based on Linearity,No auto-correlation,Normality of error,Homoscedasticity, Multicollinearity.

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

---> Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature,year and season

General Subjective Questions

1. Explain the linear regression algorithm in detail ?

---> Linear regression is a supervised learning algorithm used for predicting a continuous output variable (also called dependent variable) based on
one or more input variables (also called independent variables). The algorithm assumes that there is a linear relationship between the independent and
dependent variables.

Here are the steps involved in the linear regression algorithm:

Data preparation: The first step in linear regression is to prepare the data. This includes cleaning the data,

handling missing values, and removing
outliers. The data should be split into training and testing sets.

Model building: The next step is to build the linear regression model. This involves choosing the independ ent variables that are most relevant to

the dependent variable, and fitting a line through the data that best describes the relationship between th em. The line is described by the equation
Y = b0 + b1X1 + b2X2 + ... + bn*Xn, where Y is the dependent variable, X1, X2, ..., Xn are the independe nt variables, b0 is the intercept, and b1, b2,
\ ..., bn are the coefficients of the independent variables.

Model training: Once the model is built, it is trained on the training data. This involves finding the values of the coefficients b0, b1, b2, ..., bn
that minimize the sum of the squared errors between the predicted values and the actual values in the trai ning data. This is done using a technique
called least squares regression.

Model evaluation: After the model is trained, it is evaluated on the test data to see how well it performs. T his involves calculating the mean squared
error (MSE), which is the average of the squared differences between the predicted values and the actua l values in the test data.

Model improvement: If the model does not perform well, it can be improved by changing the independent variables, adding more data, or using a
different type of regression algorithm.

2. Explain the Anscombe’s quartet in detail?

---> Anscombe’s quartet is a set of four datasets with nearly identical statistical properties that were creat ed by the statistician Francis Anscombe
in 1973. The quartet is often used to illustrate the importance of visualizing data in addition to calculating summary statistics.

Each dataset in the quartet consists of 11 (x,y) pairs, and all four datasets have the same mean, variance, correlation coefficient, and linear
regression line. However, the datasets are visually distinct and have different patterns of relationships bet ween the variables.

Here are the characteristics of each dataset in the quartet:

Dataset I: This dataset has a linear relationship between the variables and no outliers. It is a simple, straig htforward dataset that is easy to
analyze.

Dataset II: This dataset has a non-linear relationship between the variables and a strong outlier. The outli er has a large influence on the
summary statistics, and the linear regression line is a poor fit for the data.

Dataset III: This dataset has a linear relationship between the variables, but with one outlier that has a lar ge effect on the regression line.
Removing the outlier would result in a completely different regression line.

Dataset IV: This dataset has a non-linear relationship between the variables, but no outliers. The summar y statistics are the same as the other

datasets, but a linear regression line is not an appropriate fit for the data.

3. What is Pearson's R?

---> quantifies the linear relationship between two variables. It is used to determine the strength and direction of the association between two
variables, where a value of +1 indicates a perfect positive linear relationship, 0 indicates no linear relationship, and -1 indicates a perfect
negative linear relationship.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It ranges from -1 to +1,
and its value indicates the degree and direction of the linear relationship between the variables. Pearson's R is widely used in fields such as
psychology, social sciences, finance, and engineering to explore relationships between variables and to predict outcomes based on these relationships.

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

----> Scaling refers to the process of transforming the range of data into a standard range, typically between 0 and 1 or -1 and 1. It is performed
to ensure that all features in a dataset are on a similar scale, which is important for certain machine learning algorithms to function properly.

The need for scaling arises because different features may have different units, scales, and ranges. For example, a dataset may include features
such as age (in years), income (in dollars), and temperature (in Celsius), and these features may have vastly different scales. When using machine
learning algorithms such as K-nearest neighbors, support vector machines, or clustering algorithms, this difference in scales may cause some
features to dominate over others, leading to biased results.

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

--->The Variance Inflation Factor (VIF) is a measure of multicollinearity in a regression model. It indicates how much the variance of the estimated
regression coefficient is increased due to multicollinearity among the predictor variables. A high VIF value indicates a high degree of correlation
between the independent variables, which can lead to instability in the estimated coefficients and difficulty in interpreting the model.

In some cases, the VIF value may be reported as infinity. This happens when there is a perfect linear relationship between one or more pairs of
predictor variables. Perfect linear relationships occur when two or more variables are identical or when one variable can be expressed as a linear
combination of the other variables. In these cases, the VIF value cannot be computed because the variance of the estimated regression coefficient is
undefined due to the perfect collinearity among the variables.

To address this issue, it is important to identify the variables that are causing the perfect collinearity and either remove them from the model or
find a way to transform them so that they are no longer collinear. Common techniques for dealing with per

fect collinearity include combining the
collinear variables into a single variable, dropping one of the collinear variables from the model, or using a regularization technique such as ridge
regression or Lasso regression.

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

---->A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess whether a set of data follows a specific probability distribution, such
as the normal distribution. The Q-Q plot compares the quantiles of the data to the quantiles of a theoretical distribution, and any deviations from
a straight line indicate departures from the theoretical distribution.

In linear regression, Q-Q plots are used to check the assumption of normality of the residuals, which is one of the key assumptions of linear
regression. The residuals are the differences between the observed values of the dependent variable and the values predicted by the regression model.
If the residuals are normally distributed, the Q-Q plot will show the residuals as a straight line, with deviations from the line indicating departures
from normality.

The use and importance of a Q-Q plot in linear regression can be summarized as follows:

Checking normality assumption: A Q-Q plot is a quick and easy way to check whether the residuals from a linear regression model are normally
distributed. If the residuals are not normally distributed, it can affect the validity of the regression results.

Identifying non-linear relationships: A Q-Q plot can also reveal non-linear relationships between the dependent variable and the independent
variables that may not be apparent from the scatterplot or other diagnostic plots.

Outlier detection: A Q-Q plot can also help in identifying outliers or influential observations in the data. Outliers are observations that do not
follow the pattern of the rest of the data and can have a large impact on the regression results.