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

Answer:
We find that there are multiple caterogical variable in dataset. like season, weather, holiday, month, weekday, working day.

When weather is clear the overall number of rides are high compared with disturbing weather conditions like "Light snow / Light Rain", high "Mist". During the "Light snow / Light Rain" conditions, the number of rides are very less. However, impact on the number of rides when the weather is Clear vs Mist is less. Spring season has lesser rides compared to all other three seasons. While the number of rides are increasing year on year, Fall season has highest rides compared to other seasons in this category across a single year. Holidays affect the active count which drops.

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

Answer: If there are n dummy variables of one category, n-1 columns would also give the same information as n columns. So, in-order to be more efficient
drop_first=True removes the column that has the first True value which essentially has no impact on the overall information we get from the data set.

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

Answer: While "temp" and "atemp" columns have a great correlation, next comes "registered" and "cnt" variables

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

Answer: below is the assumption of linear Regression
No Multicollinearity: The correlation matrix between all the independent variables is calculated. Correlations of close to 1 or -1 indicate a strong collinearity.

Linear Relationship: The initial scatter plot of the independent and dependant variables is almost in line with the relationship between the
predicted and the actual values at the end. Confirming the assumption to be true.

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

Answer: The top 3 features directly influencing the count are the features with highest coefficients. These are: Temp, Year (positively influencing) and snowy and rainy weather (negatively influencing).

General subjective question:
Explain the linear regression algorithm in detail.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression. Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression.

Types of Linear Regression
There are two main types of linear regression:

Simple Linear Regression
This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$y = \beta 0 + \beta 1X$

where:

Y is the dependent variable
X is the independent variable
β0 is the intercept
β1 is the slope

Multiple Linear Regression
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

y = β0 + β1X + β2X ..... βnX
where:

Y is the dependent variable
X1, X2, …, Xp are the independent variables
β0 is the intercept
β1, β2, …, βn are the slopes
The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.
In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

2. Explain the Anscombe’s quartet in detail.

Answer: Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
Purpose of Anscombe’s Quartet
Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3.What is Pearson’s R?

Answer: The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.
To find the Pearson coefficient, also referred to as the Pearson correlation coefficient or the Pearson product-moment correlation coefficient, the two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association. Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

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

Scaling is necessary for a model to be functional with the appropriate range of coefficients. For e.g., if there were two independent variables named price and months on which the sale of car depended, the price range would be far too high because there are only 12 months in a year. In that case, scaling the variable price appropriately won't allow decimal errors to happen in the model. There are two types of scaling:

Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly.
Standardized scaling: The example given above is of standardized scaling. Here, the values of variable(s) is/are compressed into a specific range to suit the model.

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

Answer: Infinite VIF is an indicator of a perfect correlation between the variables. This means one independent variable can be expressed entirely as a linear combination of the other independent variables. Reasons for infinite VIF is explained as follows: VIF Calculation: VIF is calculated using this formula: 1 / (1 - R^2), where R^2 is the coefficient of determination between a specific independent variable and all the other independent variables combined. Perfect Multicollinearity: In this case, R^2 between the variable and the others becomes exactly 1. This signifies a perfect linear relationship, where one variable can be perfectly predicted by the others.
Division by Zero: Plugging R^2 of 1 into the VIF formula results in a denominator of zero (1 - 1 = 0). Dividing by zero is mathematically undefined, hence the infinite VIF.
Essentially, an infinite VIF indicates that the variable you're looking at is redundant because it carries no unique information.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Q-Q plot is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model.