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

It is observed that categorical variable like Wathersit_3 decreases the bike hires where as spring season and winter season increases bike hires

The demand of bike is almost similar throughout the weekdays

Bike demand doesn't change whether day is working day or not

Bike demand is high in the months from May to October.

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

drop_first=True helps in reducing the extra column created during dummy variable creation,If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

atemp and temp both have same correlation with target variable of 0.63 which isthe highest among all numerical variables.

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

Validated Assumption by:

1.Verifying if Error terms are normally distributed

2.Carrying Residual analysis to prove residuals are normally distributed

3.Eliminating variables to remove multi collinearity

4.Also assuming that there is relationship between x and y

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

Temperature (temp)

Year (yr)

weathersit

1.Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares.

The following is an example of a resulting linear regression equation:

y=b0+b1x1+b2x2...

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is.

R-squared values range between 0 and 1

linear regression consists of 3 stages –

(1) analyzing the correlation and directionality of the data,

(2) estimating the model, i.e., fitting the line, and

(3) evaluating the validity and usefulness of the model.

2.Explain the Anscombe's quartet in detail

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. The

y have very different distributions and appear differently when plotted on scatter plots.
The four datasets of Anscombe’s quartet.

3.What is Pearson’s R?
In Statistics, the Pearson’s Correlation Coefficient is also referred to as Pearson’s r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
Whenever we discuss correlation in statistics, it is generally Pearson’s correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
Pearson’s correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
What?
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Why?
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean ( ) zero and standard deviation one (σ).sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?
When the VIF is infinite, it typically means that there is perfect multicollinearity between one or more predictor variables in the regression model. Perfect multicollinearity occurs when one or more variables can be perfectly predicted from a linear combination of other variables in the model.
In other words, there is a perfect linear relationship among the predictor variables.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let’s discuss some of the probability distributions.
The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine:
1.If two populations are of the same distribution
2.If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
3.Skewness of distribution