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

Winter season is the season where there is most demand for the shared bikes.demand for shared bikes is increasing YoY, the demand is higher on working days. September is the month where the demand is higher.

On the other hand, demand for shared bikes decreases when there is light rain, snow, and thunderstorm and with increase in the windspeed. In the spring season the demand for the shared bikes decreases.

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

Using 'drop_first=True' in the dummy variable prevents in creating one additional column during the dummy variable creation for categorical variables. This would reduce the correlations among dummy variables.

For categorical variables, the value of nth type of value can be inferred from the values present in the rest of n-1 type value columns. An additional dummy variable just to identify the value for nth type element seems redundant.

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

'atemp' column has highest correlation with the target variable 'cnt'.

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

Once the model is built using the training dataset, we use the initially sliced test dataset to make predictions on this model. The test dataset will undergo the same pre-processing steps which the training dataset has undergone. We transform the test dataset based on the train dataset values. We predict y values using this model.

To evaluate the model, we plot the scatter plot to between y predicted values to the y actual values of the of test set to see it follows a liner regression line.

We also compare the computed R2 square value for the trained and the test dataset of the model, which should be comparable.

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

The top three features are :

atemp: feeling temperature in Celsius – with the increase in temperature the demand for bikes increases.
Yr - demand for shared bikes is increasing YoY.
weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds – will effectively negatively affect the demand of the shared bikes.

Explain the linear regression algorithm in detail:

Linear regression is a basic predictive analysis algorithm based on supervised learning.it is the most used predictive analysis model to find out the relationship between the independent variables and the derived variable. the relationship between the variables is defined using the best fitted line of the linear equation.

Linear regression model can provide the understanding on the below:

- The affect of the independent variables on the derived variables.
- To find out the change in the value of the derived variables w.r.t to the one or more independent variables.
- To predict and forecast of the derived variable based on the linear regression model.

Linear regression is further classified into two types:

Simple linear regression:

the simple linear regression provides the relationship between a dependent variable and single independent variable with a straight line. The straight line is plotted on the scatter plot on the dependent variable and one independent variable.

Multiple linear regression:

the multiple linear regression provides the relationship between a dependent variable with more than one independent variables. A linear equation is derived based using the observed data of the independent variables to compute the derived variable.

Explain the Anscombe's quartet in detail

Anscombe's Quartet are a group of four data sets which have nearly identical simple descriptive statistics, but the distributions of the data are very different. The visualisation of the graphs are completely different when plotted on the scatter plots. Each of the datasets consists of 11 data points on the (x,y) plane. The statistical observations of mean, sample variance, linear regression line, correlation between x and y, R squared values are either exact or close to the decimal points for these four groups of datasets.

What is Pearson's R?

Pearson correlation coefficient (Pearson's R) is a measure to determine the strength of the relationship between two sets of data.

The range of the Pearson's correlation coefficient is from -1 to 1.where negative indicates inverse corelation and positive indicates direct corelation between the variables.

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

Scaling is an pre-processing technique applied to independent variables to normalize the data to a particular range. It helps to improve the performance and speed up the calculations in an algorithm.

Scaling can be of two types:

Normalized scaling: in Normalized scaling one or more attributes are rescaled to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

Standardised scaling: In Standardised scaling, the data is rescaled to have the mean of 0 and standard deviation of 1.

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

VIF values are shown as infinite when there is absolute corelation between two independent variables. During this case, the value of R squared is equal to 1. which makes the computed VIF value to infinity ($1/(1-R2)$).

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

Q – Q plot is scatter plot in which two quantiles are plotted against one another. If both the quantiles have same distribution, we would see a single straight line.it helps us visualize the similarity of two distributions such as , shifts in the locations, scales and symmetry can be identified from the Q-Q plot.

In linear regression, q-q plot helps in a scenario when we receive training and test data sets separately. Q-Q plot can help us understand and confirm both datasets have identical populations and data distribution.