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

The demand for bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

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

By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
a. There should be a linear and additive relationship between the dependent (response) variable and the independent (predictor) variable(s).

b. A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.

c. There should be no correlation between the residual (error) terms. The absence of this phenomenon is known as Autocorrelation.

d. The independent variables should not be correlated. The absence of this phenomenon is known as multicollinearity.

e. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to as heteroskedasticity.

f. The error terms must be normally distributed.

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

The Top 3 features contributing significantly towards the demands of share bikes are:
1. weathersit_Light_Snow(negative correlation).
2. yr_2019(Positive correlation).
3. temp(Positive correlation).

1. Explain the linear regression algorithm in detail

Linear Regression is a machine learning algorithm that is based on a supervised learning category. It finds the best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses the Sum of Squared Residuals Method.
Linear regression is of 2 types:
i. Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.
The formula for the Simple Linear Regression:
$Y=\beta0+\beta1X1 +\epsilon$
ii. Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.
The formula for the Multiple Linear Regression:
$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$
The equation of the best fit regression line $Y = \beta_0 + \beta_1X$ can be found by the following two methods:
· Differentiation
· Gradient descent
We can use statsmodels or SKLearn libraries in python for the linear regression.

2. Explain the Anscombe's quartet in detail

Anscombe's Quartet was developed by statistician Francis Anscombe. This is a method that keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.
Additionally, Anscombe’s Quartet warns of the dangers of outliers in data sets.

3. What is Pearson’s R?\

Pearson’s R was developed by Karl Pearson and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by ‘r’. it has a value between +1 and −1, where 1 is a total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.
Mathematically, Pearson’s correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a “product moment”, that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Formula
r =\frac{\sum\left(x_{i}-\bar{x}\right)\left(y_{i}-\bar{y}\right)}{\sqrt{\sum\left(x_{i}-\bar{x}\right)^{2} \sum\left(y_{i}-\bar{y}\right)^{2}}}
r = correlation coefficient
x_{i} = values of the x-variable in a sample
\bar{x} = mean of the values of the x-variable
y_{i} = values of the y-variable in a sample
\bar{y} = mean of the values of the y-variable

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

Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.
The two most discussed scaling methods are Normalization and Standardization. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

The value of VIF is calculated by the below formula:
1 / (1 − Ri2)

Where ‘i’ refers to the ith variable.

If the R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value becomes infinite. It denotes perfect correlation in variables.

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

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.
Importance of Q-Q plot: Below are the points:
I. The sample sizes do not need to be equal.
II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
III. The q-q plot can provide more insight into the nature of the difference than analytical methods.