Answer 1

Categorical variables have a high effect on the count of the bikes taken. Year, month and weather situation are highly correlated to count variable.

Answer 2

Drop_first=True is used so that no extra columns are created while making dummy variables.

Answer 3

Temperature has the highest correlation amongst the numerical variables with count

Answer 4

I checked for heteroscedasticity, autocorrelation, linearity, normality of the error terms.

Finally found that autocorrelation exists here so we should use some other model instead of linear regression model

Answer 5

The top 3 variables are month, year and temperature

General question

1. First, we import the dataset and clean the dataset. We remove all the unnecessary variables. Then we encode the categorical variables. After that we divide the data into training and testing sets. Now we train the data on the training set and check for the assumptions of multiple linear regression. After that we test the model on the testing data and check if the model is performing well. After that we find the slope and the intercepts to know which variable has the highest impact.
2. Anscombe's quartet can be defined as a set of four data sets that are nearly identical in simple descriptive statistics, but the data sets have some idiosyncrasies that fool the regression model during construction. They have very different distributions and look different when plotted on a scatterplot.
3. Pearson's correlation coefficient correlation or colloquially simply correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio of the covariance of two variables and the product of their standard deviations. So, it's essentially a normalized measure of covariance, so the result will always be a value between -1 and 1. Like the covariance itself, the measure only reflects the linear correlation of the variables and ignores many other types of relationships or correlations.
4. A data preprocessing step applied to the independent variables to normalize the data within a certain range. It also helps in speeding up the computation of algorithms.
   In most cases, collected datasets contain features that vary widely in size, units, and extent.

Without scaling, the algorithm only considers the size and not the unit, which is incorrect modeling. To solve this problem, all variables should be scaled to have the same size. Normalisation brings all of the data in the range of 0 and 1. Whereas Standardization replaces the values by their Z scores.

5. If there is a perfect correlation then VIF = infinity
6. The q-q plot is a graphical technique for determining if two datasets come from population with common distribution. The advantage of a qq plot is that sample size need not be equal and many distribution aspects can be simultaneously checked.