Assignment Based Subjective Questions

1. Among all categorical variables, working day and year plays a significant role for determining target variable.
2. We use the above said command to drop the redundant variables from the data.
3. According to the pairplot, temp and atemp variables had the highest correlation with target variable.
4.Correlation in pair plots and heatmap shows that it is perfect to use linear regression model. Also, the histogram of error terms showed that errors are centrally distributed and hence our model is successful.
5. Based on model, temp, working day and year played the most significant role.

General Subjective Questions

1. The mathematical equation can be given as:

Y = β0 + β1*x
Where
Y is the response or the target variable
x is the independent feature
β1 is the coefficient of x
β0 is the intercept
β0 and β1 are the model coefficients (or weights). To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the target variable such as Sales!
NOTE: The main aim of the regression is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the points to the regression line.

2. 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. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

4. The goal of applying feature scaling is to make sure features are on almost the same scale so that each feature is equally important and make it easier to process by most machine-learning algorithms.

Standardization

The result of standardization (or Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation are 0 and 1, respectively.

This technique to rescale features value with the distribution value between 0 and 1 is useful for the optimization algorithms, such as gradient descent, that are used within machine-learning algorithms that weight inputs (e.g., regression and neural networks). Rescaling is also used for algorithms that use distance measurements, for example, K-nearest-neighbours (KNN).

Max/Min Normalization

Another common approach is the so-called max/min normalization (min/max scaling). This technique is to re-cales features with a distribution value between 0 and 1. For every feature, the minimum value of that feature gets transformed into 0 and the maximum value gets transformed into 1.

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior