Assignment Based Subjective Questions:

1. Influential categorical variables are:
i. Season: The mean of count increased significantly in the second quarter, in the third quarter there was marginal increase. Finally in the forth quarter the mean count dropped again. Although the range of the quarter count was more or less the same.
ii. Month (mnth): The mean of count increased steadily from January to June, then the peak maintained till October, and then dropped in the last two months.
iii. Season and Month are two variables that showcase the same pattern. Which is to be expected because they are inter-dependent.
iv. Weathersit : is a variable that indicates the weather situation of the day. Customers came out the most when its value was one, the count dropped a little for its second value. Then the count decreased rapidly for its third value.
3 Binary variables are:
i.Workingday: The count was more or less unaffected for this variable.
ii. Yr: The business grew significantly in the second year
iii.Holiday: The mean is higher when it is not a holiday. Holiday mean is the same as non-holiday 25% mark.

2. By dropping dummies, we signify the absent column as the referent column. Meaning the model initially has the value 0 for all the dummies of that categorical variable. And that state represents the reference state where it is true for the dropped dummy. And only when any other dummy column has a tick, the dropped dummy becomes insignificant for that data point, and the ticked column represents the attribute.

3.Temp and atemp, or the temperature and air-temperature of are the two most correlated variables, they are in turn closely correlated themselves, having almost a linear relationship in themselves.

4. At the beginning of the task I reserved 30% of the data for testing. After the model was built I used the data frame of independent variables, of the test set to predict the output. Difference between the R-square of the output of train and test data-frame was within 5%. So I concluded that the model is not over-fitting.

5. The top three variables:
1. Temp: the count has a high positive correlation with the temperature variable.
2. Windspeed: Windspeed contributes significantly
3. Yr: the count was higher in the second year
4. S2: whether the count was made in the second season is a contributing factor

General Subjective Questions:

1.Linear Regression Algorithm

It uses linear equation as means to explaining the dependence of one variable on the other. Simple linear regression has one independent variable, and multiple linear regression has multiple independent variables. Target or the dependent variable is represented in mathematical terms like this:

$y = c + a1x1 + a2x2 + \ldots + anXn$

There are as many coefficients as there are independent variable. Aim of this algorithm is to find the coefficients of this equation. The way it is done is by the concept of best fit line in case of simple linear regression, and best fit plane in case of multiple linear regression. When drawing this best fit line or plane some considerations are made, and they are as follows:

1. The fit has to be acquired in such a way that the error terms (predicted y - actual y) are distributed normally.
2. The fit must dissect the data points efficiently, so that there will not be outliers.
3. The independent variables are not correlated with each other, to preserve the readability of the model.
4. The fit must be a good generalised estimation, or the model should not memorise the data points to the extent that its efficiency drops drastically in case of unseen data.

Steps to linear regression:

1. First step is to load and read the data, and perform sanity check on the data. If there are any missing value, whether the variables are assigned to correct data type etc.
2. Then the categorical variables are processed by getting dummies for them. So they are turned into a bunch of binary variables, which together represent the status of the categorical variable for that data point.
3. EDA is performed to find out which independent variables have a high correlation with the target variable and with each other.
4. Data set is split into training data and testing data.
5. Numeric variables are scaled within the range (0,1), this process is necessary for getting proportionate coefficients for the variable.
6. Feature selection: in this section our target is to find the minimum set of independent variables that can explain the target data.
    i. RFE analysis helps us to minimise the number of independent variables to a

reasonable number.

ii. Manual analysis: is the processes where we handpick the remaining variables based on their p-value and VIF score. P-value indicated whether the variable is statistically significant, and the VIF score is an indicator of multicollinearity.

7. Going back and forth we come down to the minimum set of variables. For which the R-squared value stagnates.

8. The model is then tested on the test data, and if the R-squared value of the test data is reasonably closer to the train data, we accept the model, or conclude the model is not over-fitting.

9. After that the residual analysis is performed. In this section we want to check whether the error terms are following the normal distribution. It should be true for both train and test data.

2.Anscombe's quartet:

Is a graphical representation of data, that aims to show that data sets can have different kinds of graphical relationships among themselves, even when they display similar statistical properties.

This outlines the importance of data visualisation as a means to understanding geometric relationships.

Anscombe used four similar datasets to demonstrate this fact, hence the name. The four datasets had similar values along the y axis for the values of x, and they show similar statistical properties.

But,

1. The first datasets plot a linear relationship, and seems like a data where we can fit a line through linear regression.
2. The second plot is Gaussian
3. Third plot is a clean linear relationship, but the correlation coefficient is hampered by the presence of an outlier.
4. And the fourth is a high density plot along a small span of X axis, with the presence of an outlier.

It is not clear how the datasets were created, and they seem a tad manipulated to illustrate the point. But they exemplify the importance of EDA, because statistical information of the data is not enough to explain the relationships it exhibits.

3. Pearsons' R

Is used as a measure of linear correlation between two variables. Or it can be understood as the slope of the line that is fitted between the data points, given a somewhat-linear relationship exists.

Covariance is a measure for linear relationship, if the relationship is direct it has a high value, if it's inverse the value is negative.

Standard deviation is the measure of how far the values are distributed from the mean of the data-set.

Petersons' R is calculated by dividing the product of the standard deviation of the variables by their covariance.

4.Scaling: Scaling is a process of compressing the numeric variables into a small range of values. The MinMax scaling fo example equates the maximum values with 1, and minimum values with 0, and every other value is mapped in between. This is important because it keeps the coefficients of the linear model in proportion. If this process is omitted then the coefficients derived from the model will be very different from each other. For example some will be in the range of tens, and some in the range of thousands. Needless to say this will hamper the readability of the model. But when they are scaled we can understand from the coefficients exactly how much weightage is assigned to these variables, and how these weights compare among each other.

5.VIF: is represented by the equation 1/(1-R2)
For a perfect correlation among two independent variables the R2 would be 1. Which will result in a denominator that is 0.
So for any case of perfect multi-collinearity the VIF would be infinite.

6.Q-Q plot:
Or quantile Quantile plot is between the statistically observed value and their quantile value. This plot is used to verify whether a distribution is normal. For n set data points the quantile axis is divided into n+1 regions. It would be assumed that the smallest observed value will be on the first quantile, and the highest observed value will be on the last quantile. For a normal distribution the curve should be almost a straight line. Depending on the tail of the normal distribution the curve would show certain patterns, if one of the tails is longer than the other, QQ plot would show a discontinuity or erratic behaviours at that end of the plot. If the distribution has distant values both at maximum and minimum, the dispersal from the straight line at both ends would be significant.