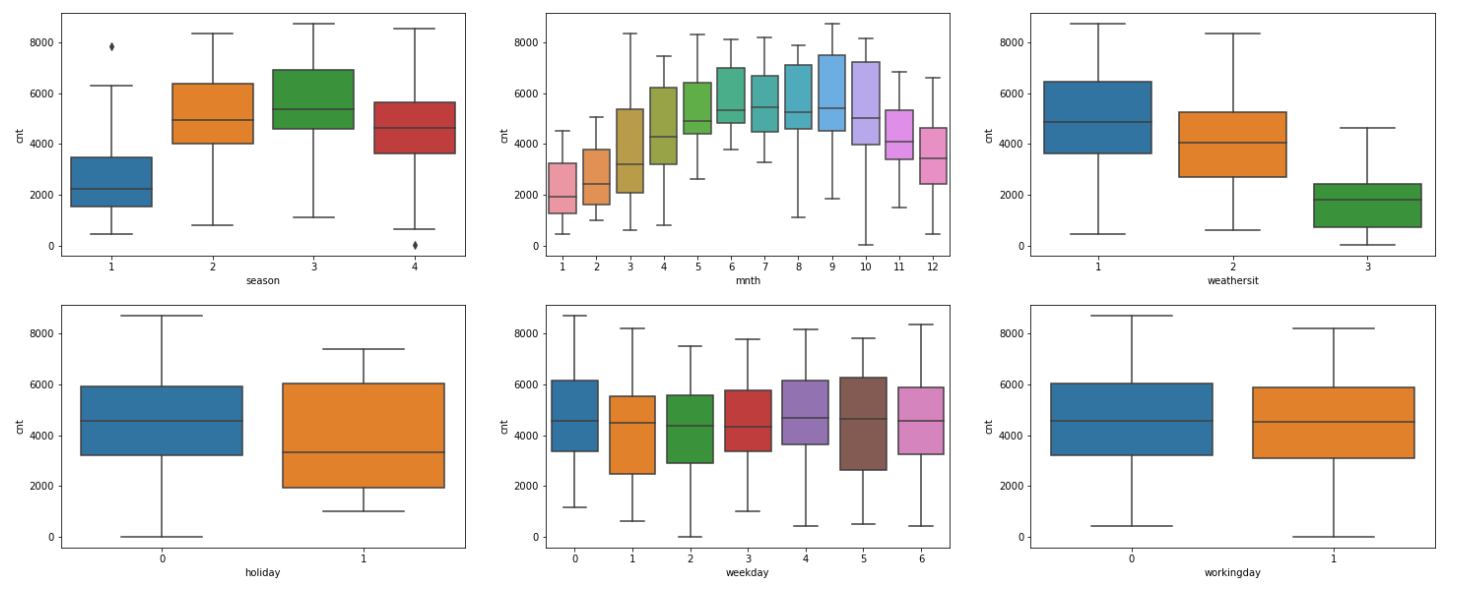
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable?

ans:-



As above the boxplot is show :

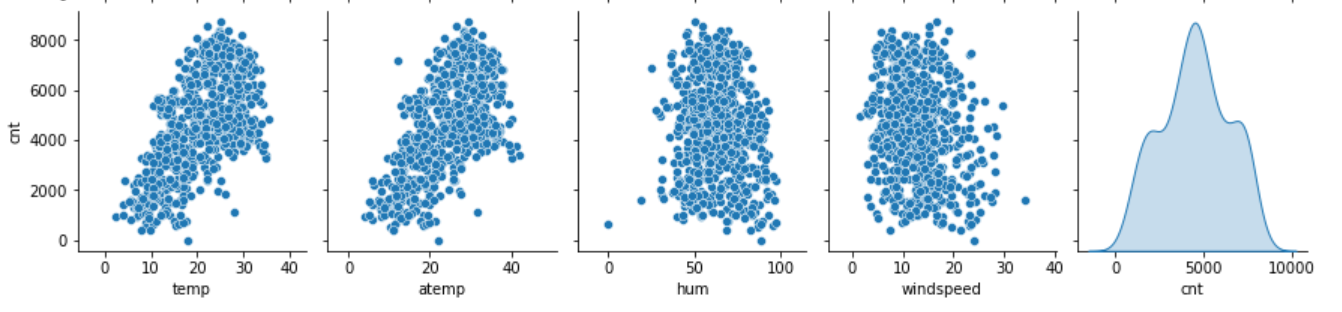
1. First boxplot is cnt and season
   1. As we know that season divided into 4 variable (1:spring, 2:summer, 3:fall, 4:winter) so, we can see that at spring the cnt is low as compare to other variable .
   2. Summer and fall have same number of cnt and winter slightly lower then summer and fall.
2. Second boxplot is cnt and month(mnth)
   1. In this plot we can see that cnt is increasing from month 1 to month 6 .
   2. But , month 7,8,9 cnt remain same but afterward it start decreasing.
3. Third boxplot is cnt and weathersit
   1. As we know that weathersit divide into 3 variable ( 1: Clear, Few clouds, Partly cloudy, Partly cloudy, 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)
   2. For the plot we can see that at clear weather the count is more as compare to other two variable
4. Fourth boxplot is cnt and holiday
   1. As holiday is give as 1 and 0 . we know that 1 means it is holiday and 0 mans it is not -holiday.
   2. we can say that if it is holiday or not count remain same . it is not affecting Business if it is holiday or not
5. Fiveth boxplot is cnt and weekday
   1. All day the count are same there not major difference but at weekday (0) as Sunday the count are high.

2. Why is it important to use drop\_first=True during dummy variable creation?

Ans:- it is important to use it help to reduce number of columns form the data and it also help to reduce the correlations created among dummy variable.

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

Ans:-



From the pair plot with target variable we , can say that temp is more correlated with target variable and another variable is atemp iwhich is also correlated to target variable . hum and windspeed is to more related to target variable .

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

Ans:

We can validate the assumptions of liner Regression after the model on the training set by residual analysis and draw plot and whether the error is normal distributed or not.

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

Ans: according to final temp , season\_winter , year are top 3 features contributing significantly toward demand shared bikes

* 1. Temperature (temp) - A coefficient value of ‘0.570606’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.570606.
  2. Weather Situation 3 (weathersit\_Light-Snow) - A coefficient value of ‘-0.236675’ indicated that, w.r.t weathersit\_Mist-Cloudy, a unit increase in weathersit\_Light-Snow variable decreases the bike hire numbers by 0.2366 units.
  3. Year (yr) - A coefficient value of ‘0.228914 indicated that a unit increase in yr variable increases the bike hire numbers by 0.228914 units.

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

y= a0+a1x+ ε

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

1. Explain the Anscombe’s quartet in detail.

Ans:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

1. What is Pearson’s R?

Ans:

In statistics, the Pearson correlation coefficient ― also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation or colloquially simply as the correlation coefficient ― is a measure of linear correlation between two sets of data. It is the ratio between the covariance[3] of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

1. What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling?

Ans:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

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

Ans:

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.