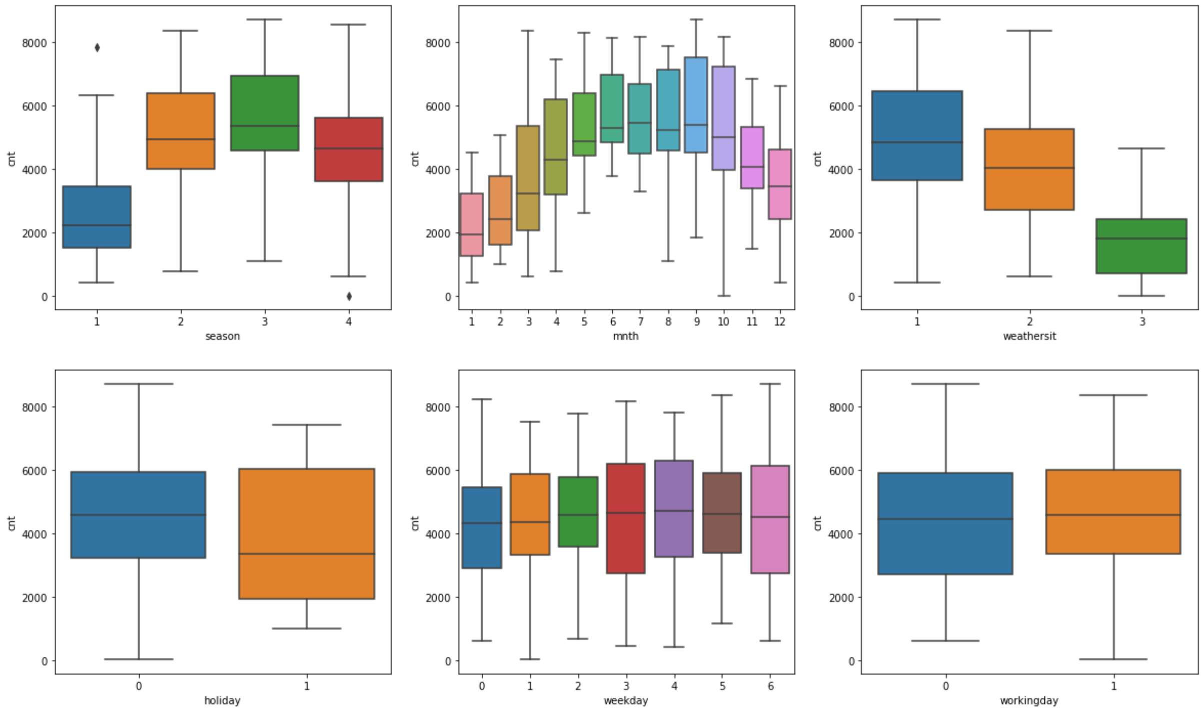
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   Ans: Below is the analysis for each of the categorical variable on the dependent variable "cnt"
   Season - There are more number of bookings in season3(fall) almost nearly 32%, than season2(summer) 27% and season4(winter) 25%, all have a median above 5000.
   Mnth – There is a increase in booking between the months 5 to 10
   Weathersit – huge amount of bookings(nearly 67%) occur in the weathersit1 which falls under Clear, Few clouds, Partly cloudy, Partly cloudy
   Holiday - Most of the bikes are booked when there is no holiday, here the variable is biased with "cnt". The variable holiday is not a good predictor for cnt
   Weekday- all the weekdays maintain similar trend with median between 4000 to 5000.
   Workingday - Most of the bikes are booked in working day nearly 69% when compared to holiday,this is also a good predictor for "cnt"



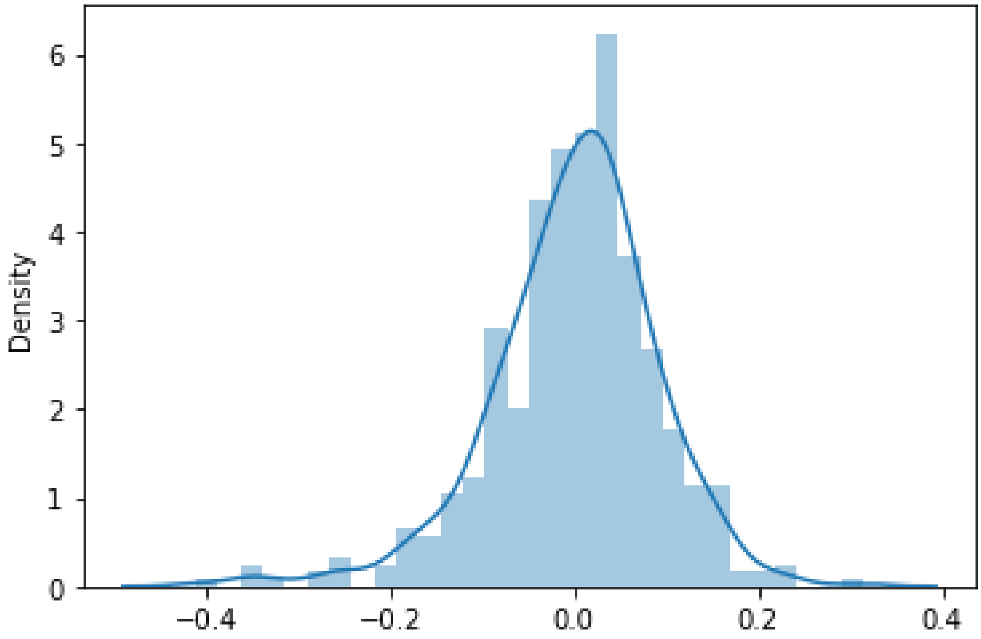
2. Why is it important to use **drop_first=True** during dummy variable creation?
   Ans: The drop_first=True will remove one column which is redundant, we can analyse the data even we don't have that column. If there are "n" columns being created using get_dummies we can use n-1 columns to analyse the data and predict the other element with the values present in the created columns
   Eg: if male = 0 and female = 1, while creating dummy variables we don't need to create two variables we can create only one variable female, if the variable has value 1 then it is female and if it is 0 then it is not female and can be taken as male. drop_first=True serves the purpose.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: atemp and temp has high correlation with the target variable.
   atemp has 63% and temp has 62%

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans: Verified whether the error terms or residuals are normally distributed or not. For this took the difference between the actual result(y_train) and predicted result (y_train_pred). Used the difference in the distplot which shows the normal distribution with mean 0

```
#validating the Assumptions
#error terms are normally distributed with mean zero

y_train_pred = lr6_model.predict(X_train_sm)
res = y_train - y_train_pred
sns.distplot(res)
```



Validated the Multicollinearity between the predictor variables using the VIF values. All the variables have VIF values below 5

	Features	VIF
2	temp	4.76
1	workingday	4.04

Features	VIF	
3	windspeed	3.44
0	yr	2.02
7	weekday_6	1.69
4	season_2	1.57
8	weathersit_2	1.53
5	season_4	1.40
6	mnth_9	1.20
9	weathersit_3	1.08

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: The 3 features contributing significantly towards explaining the demand of the shared bikes are **workingday** - Most of the bikes are booked in working day nearly 69%, **weathersit** - huge amount of bookings(nearly 67%) occur in the weathersit1 which falls under Clear, Few clouds, Partly cloudy, Partly cloudy and **temp**

General Subjective Questions

1. Explain the linear regression algorithm in detail
   Ans: Linear Regression is one of the Supervised Learning methodologies in Machine Learning. This algorithm helps in predicting the target values by using the predictor variables which are independent to each other. Linear Regression is the most widely used model in real time. The relationship between the independent and the target variables is linear positively or negatively.
   There are two types of Linear Regression Models
   1. Simple Linear Regression – in this there will be only one target and one independent variable
   2. Multiple Linear Regression – in this there will be only one target and multiple independent variables.

   The algorithm finds the best fitted line from the linearly ordered data which is the predicted target value. To calculate the best fitted line the algorithm uses the equation
   Y = mx + c
   Where Y is the predicted target value, X is the independent variable or value, m is the slope (intercept)of the line which is generally measured in **tan.** C is the constant value(coefficient) which measures the line from origin when X=0 through Y axis.
   For multiple independent variables the equation is
   Y = m1x1+m2x2+…..+mnxn+c

   This show for a unit change in x1, Y would change by m1 when all the other variables are constant. This applies to all the input or predictor variables which makes them independent.

   Linear Regression is used in scenarios like predicting the weather, predicting the score of the batsman or team etc.

   The difference between the actual and the predicted values are called residuals, the line which has the minimum Residual square sum is the best fitted line.
   To find the best fitted line Linear Regression uses the Gradient Descent method.

2. Explain the Anscombe's quartet in detail
   Ans: Anscombe's Quartet is defined as the group of four data sets which are nearly identical in simple descriptive statistics. 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 analysing 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. What is Pearson's R?
   Ans: Pearsons R or the Pearsons correlation coefficient is the formula used to find the relationship between the data or independent and target variables. The R explains what portion of the given data variation is explained by the developed regression model.
   r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
   r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
   r = 0 means there is no linear association
   r > 0 < 5 means there is a weak association
   r > 5 < 8 means there is a moderate association
   r > 8 means there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans: Scaling is the process of normalizing the range of predictor or independent variables. The range is between 0 and 1. Generally we do this for numeric variables.
   Scaling would help all the variables both numeric and categorical in the same range of 0 to 1, which helps in analysing the data easily.
   There are two types of scaling Normalization and Standardization scaling.
   Normalization is also known as min max scaling, with this the independent variables will be rescaled so that the data will fall in range 0 to 1
   The formula for normalization scaling is
   X – min(X)/max(X) – min(X)

   Standardization Scaling makes the values of each column or variable in the data have zero mean and unit variance. The formula to calculate Standardization Scaling is
   X – avg(X)/ standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: The VIF will be infinite if there is perfect correlation between independent variables. This happens when R- square is 1 and the formula for VIF = 1/1-Rsuare = 1/1-1 = infinity.
   To resolve this issue we need to drop the columns which cause perfect multicollinearilty.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?
   Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.