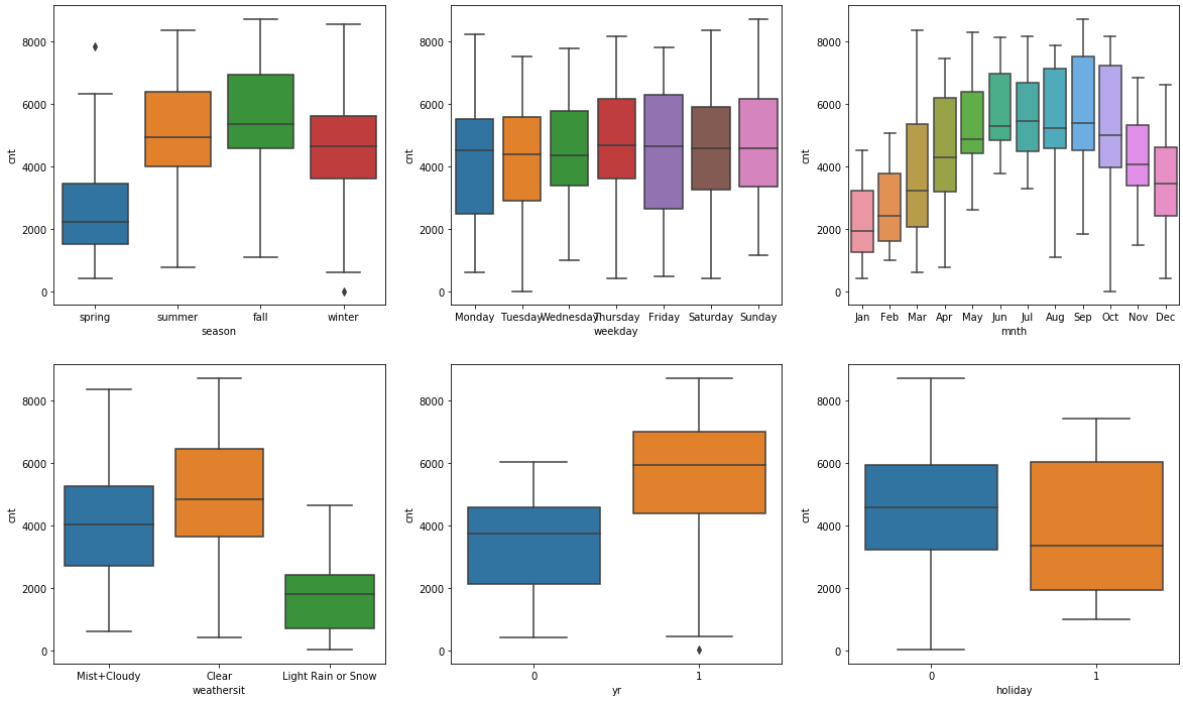
Linear Regression Subjective Questions

Assignment Based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

In [4]:

```
plt.figure(figsize=(20, 12))
plt.subplot(2,3,1)
sns.boxplot(x = 'season', y = 'cnt', data = df)
plt.subplot(2,3,2)
sns.boxplot(x = 'weekday', y = 'cnt', data = df)
plt.subplot(2,3,3)
sns.boxplot(x = 'mnth', y = 'cnt', data = df)
plt.subplot(2,3,4)
sns.boxplot(x = 'weathersit', y = 'cnt', data = df)
plt.subplot(2,3,5)
sns.boxplot(x = 'yr', y = 'cnt', data = df)
plt.subplot(2,3,6)
sns.boxplot(x = 'holiday', y = 'cnt', data = df)
plt.show()
```



Answer-1

Following are the inferences can be drawn from above analysis:

1. Customers prefer to rent bikes in Clear weather as opposed to Light Rain or Snow.
2. Count of Rentals have significantly incresed from 2018 to 2019
3. Around the year, significantly more number of bikes are rented in Summer and Fall Season, specifiacally from April to October

Question 2. Why is it important to use drop_first=True during dummy variable creation?

Answer-2

It's neccessary to use drop_first=True in dummy variable creation to drop one variable out of all categories in a variable because, if we do not drop it then there will one dummy variable with **High VIF due to high correlation with other dummy variables** in that bucket. Let's understand with an example,

Suppose there is a categorical variable in our dataset called Season which has 4 categories namely, Spring, Summer, Fall and Winter. After we create dummy variables, let's look at the dataset below

```
Spring | Summer | Fall | Winter
     0 |      0 |  0   |   1
     1 |      0 |  0   |   0
     0 |      1 |  0   |   0
     0 |      0 |  1   |   0
```

Now even if I remove any of the above columns my model would still be able to decide which category a case falls into. For example if I remove column 2 i.e. Summer,

```
Spring | Fall | Winter
     0 |    0 |     1
     1 |    0 |     0
     0 |    0 |     0
     0 |    1 |     0
```

Look at row number 3, all have 0 value which means it does not falls into either of 3 categories which simply means it falls in 4th category.

Looking at this from linear regression perspective, value of coeffient of column 'Summer' would be adjusted by coeffients of other 3 categories as well as our constant value.

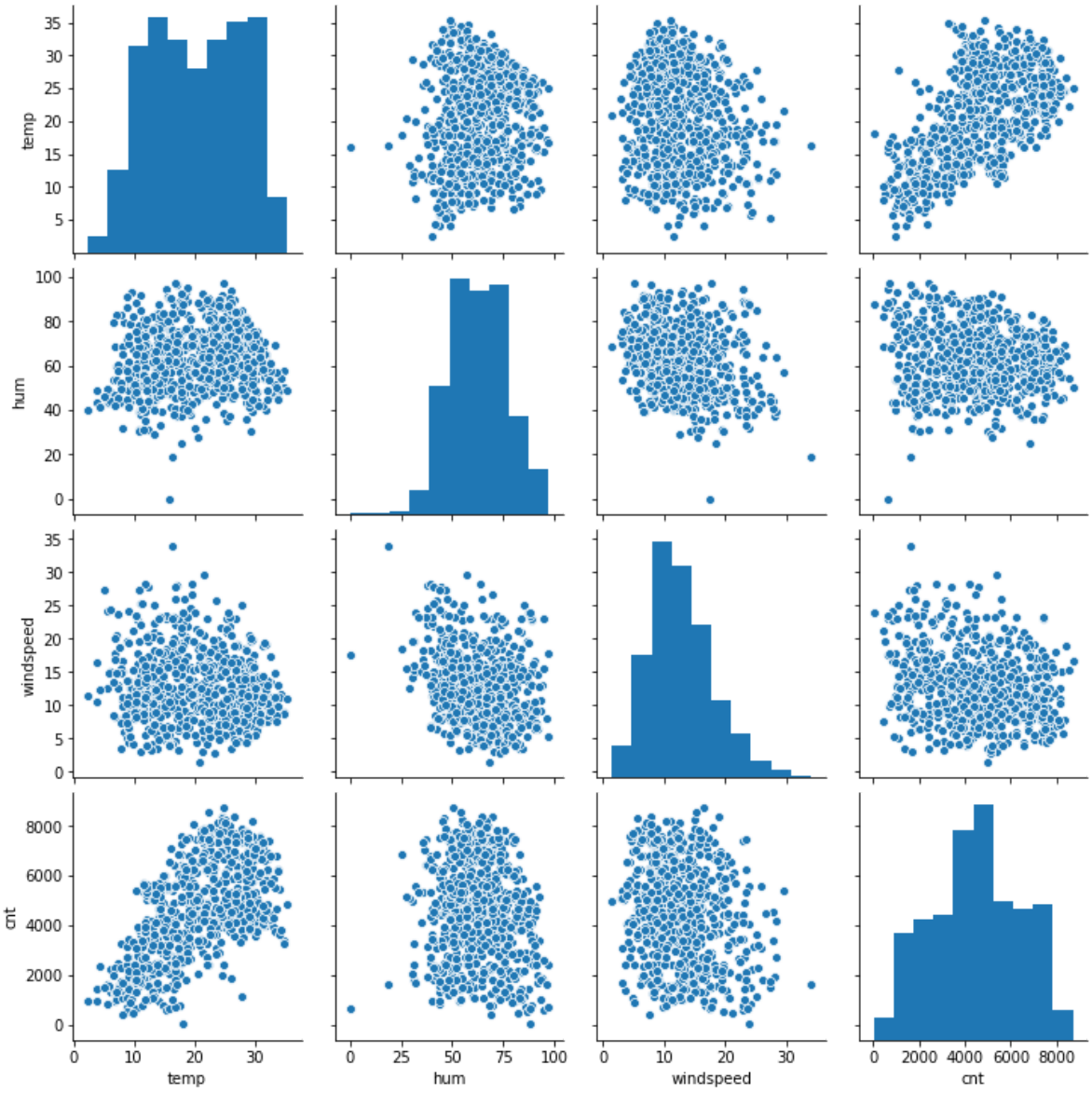
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

In [2]:

```
sns.pairplot(df[['temp','hum','windspeed','cnt']])
```

Out[2]:

```
<seaborn.axisgrid.PairGrid at 0x1d9e0326748>
```



Answer-3

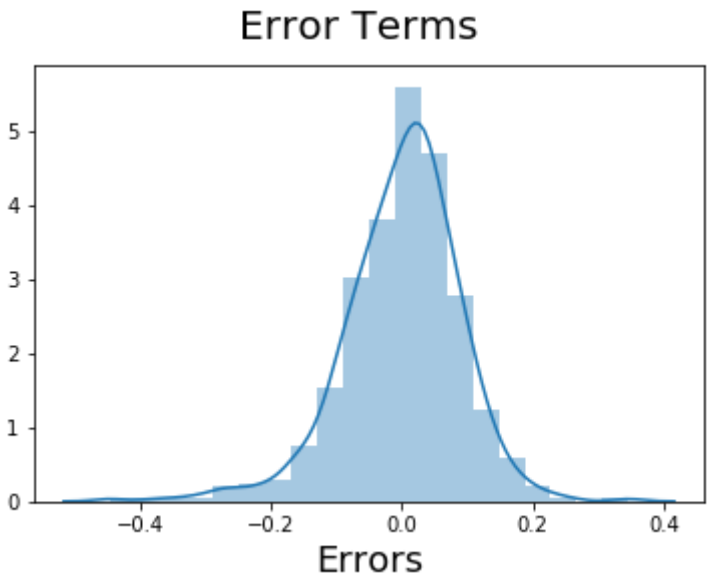
Looking at the pairplot above, Temperatue clearly has higher correlation with cnt (Target Variable). Cnt increases with increasing temperature as opposed to disstorted plot of other two variables (hum and windspped).

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

We find out the error terms by Subtracting (y_train-y_train_pred) and then plotting a distribution plot of these error terms.

In [ ]:

```
# Plotting the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)
```



Here as we can see, Error terms are normally distributed with mean as 0. Therefore assumptions are validated.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Top Three feature with p value less than 0.000 and contributing towards prediction are :

1. yr
2. temp
3. windspeed

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Answer-1

A linear regression algorithm finds the best fit of a straight line with independent variables as predictors(x1, x2, x3, etc.) and dependent variable as target(y).

There are two types of linear regression algorithms:

- Simple Linear Regression
- Multiple Linear Regression

In Simple Linear Regression we have only one independent variable whereas in multiple, as the name suggests we have more than one predictors of Target Variable.

$$y = B0 + B1 * x$$

Above is the basic equation of linear regression, in multiple-linear regression we have higher dimension (more than 1 x) and that is called a **hyper-plane**.

The two major tasks of a linear regression algorithm are:

- Tune Values of Beta-coefficients (B0 and B1 in above equation)
- To find most significant factors for prediction

For first major task we have approches like gradient descent and Ordinary Least Square and for second major task we can use methods like Recursive Feature Elimination (RFE), Vif and p-values.

Gradient Descent works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

In OLS, we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This quantity is then minimized with iterations. Statsmodel has OLS method implementation which is majorly used for linear regression algorithms.

Question 2. Explain the Anscombe’s quartet in detail.

Answer-2

Anscombe's quartet is an experiment conducted in 1973 by Francis Anscombe to explain the importance of plotting graphs for analysing the data and how viewing only statistical summary can result in poor analysis.

It consists of 4 datasets which have identical statistics summary as shown below:

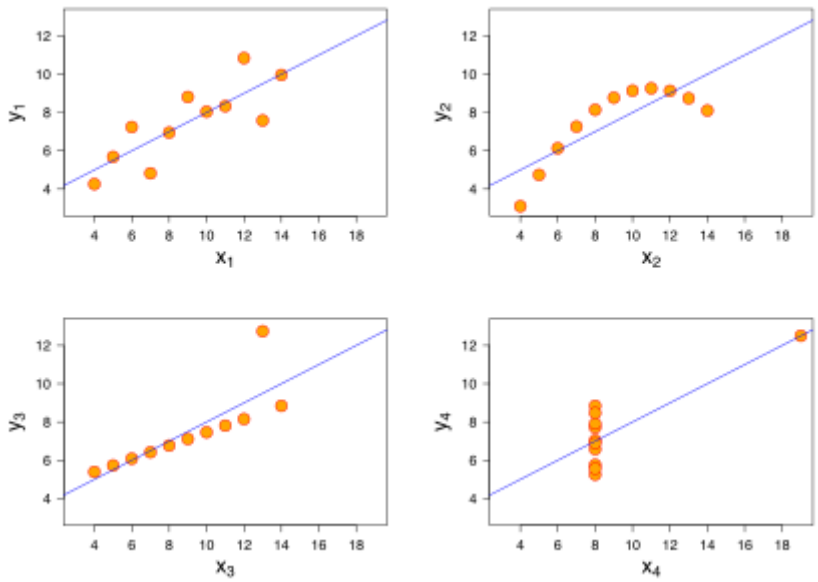
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary of this dataset is shown below and as we can see it's same for all 4.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

But when we plot x vs y for all 4 datasets we will get to see the difference in distribution of datapoints between these 4 datasets.



Note - Image Sources: Geeks-For-Geeks and Wikipedia

We can clearly see that dataset I and III have slight linear relationships and dataset II has a non-linear relationship. Whereas in dataset IV all datapoints except one have same value of x=8.0 and only a single datapoint is enough to generate similar correlations as other three

Question 3. What is Pearson's R?

Answer-3

Pearson's coefficient is a measure of linear correlation between 2 variables. Pearson coefficient can take any value between -1 to +1, where 0 indicated no correlation between the two variables. The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively.

Pearson's coefficient is calculated using the below formula

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

here, n indicates total number of data points and x and y are the values on both axis.

Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer-4

Scaling is the process of bringing feature values to a specific range we want without effecting there distribution.

There are two most common techniques of scaling, Standardization and Normalization.

It is done so that all features lie in the same range and it's better to understand the model coefficients.

In Normalized Scaling, we bring all the datapoints between 0 and 1 such that minimum value if mapped to zero, maximum value is mapped to 1 and all other values lie between 0 and 1. Below is the formula followed for Normalization.

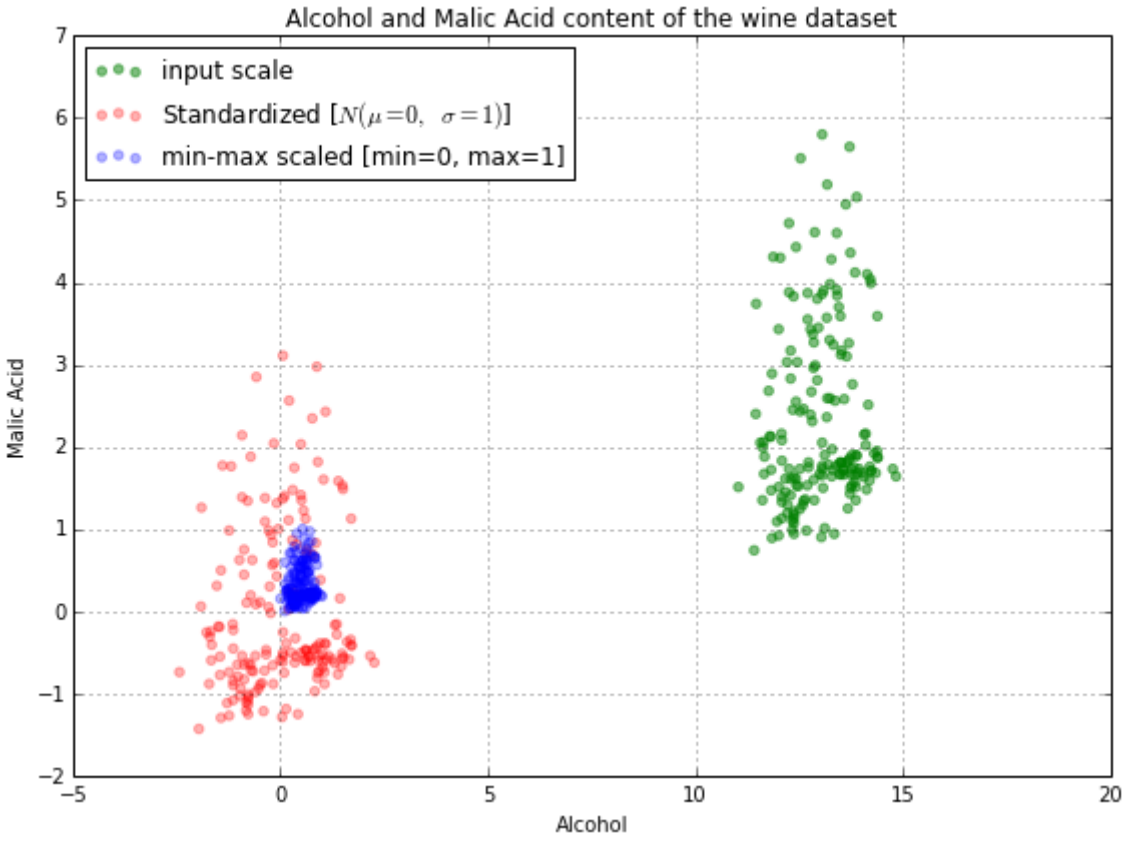
$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In Standardized scaling, we map the datapoints into a normal distribution such that mean is 0 and standard deviation is 1. The formula used for standardization is as shown below.

$$x_{new} = \frac{x - \mu}{\sigma}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers. Below diagram taken from an answer on *stacksoverflow* explains difference between normalization and standardization.

Ignore the Heading and axis values, purpose is to understand the difference.



Question 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer-5

Vif is a measure of multicollinearity between 2 variables. Variance inflation factor is calculated using the below formula, where R-square is the regression coefficient between variable i and the dependent variable.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

The reason why Vif value reaches infinity is when **R-square is equal to 1**, indicating that both variables are highly correlated. When we plot datapoints of these 2 variables on x and y axis, we'll get a stright line.

Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

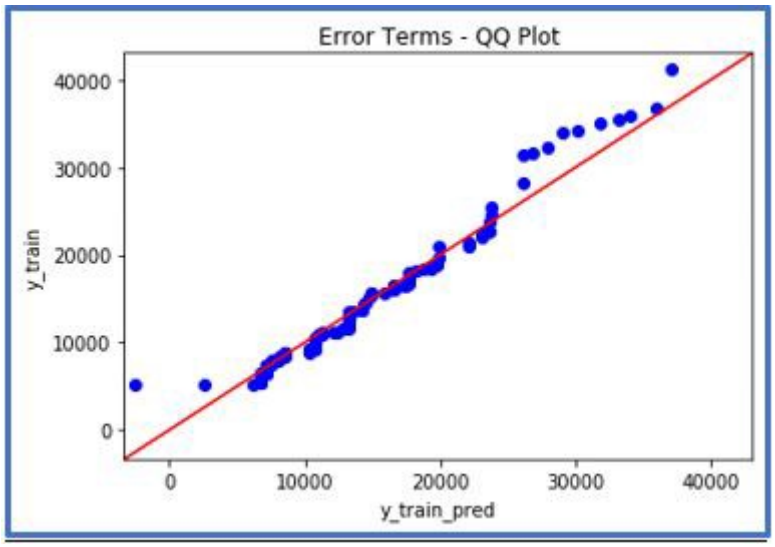
Answer-6

Q-Q plots are plotted between 2 variable's quantile values. Q-Q plots are plotted to find out if two sets of data come from the same distribution. If the two data sets come from a common distribution, the points will fall on reference line (45 degree line)

Advantages:

```
- Many distributional aspects like shifts in location and scale, changes in symmet
ry presence of outliers can be detected         from this plot.
- Q-Q plots can also be used with sample sizes.
```

Below is an example of a q-q plot.



Possible interpretations of two variables:

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x - axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

In python, statsmodels.api can be used to plot q-q plots.