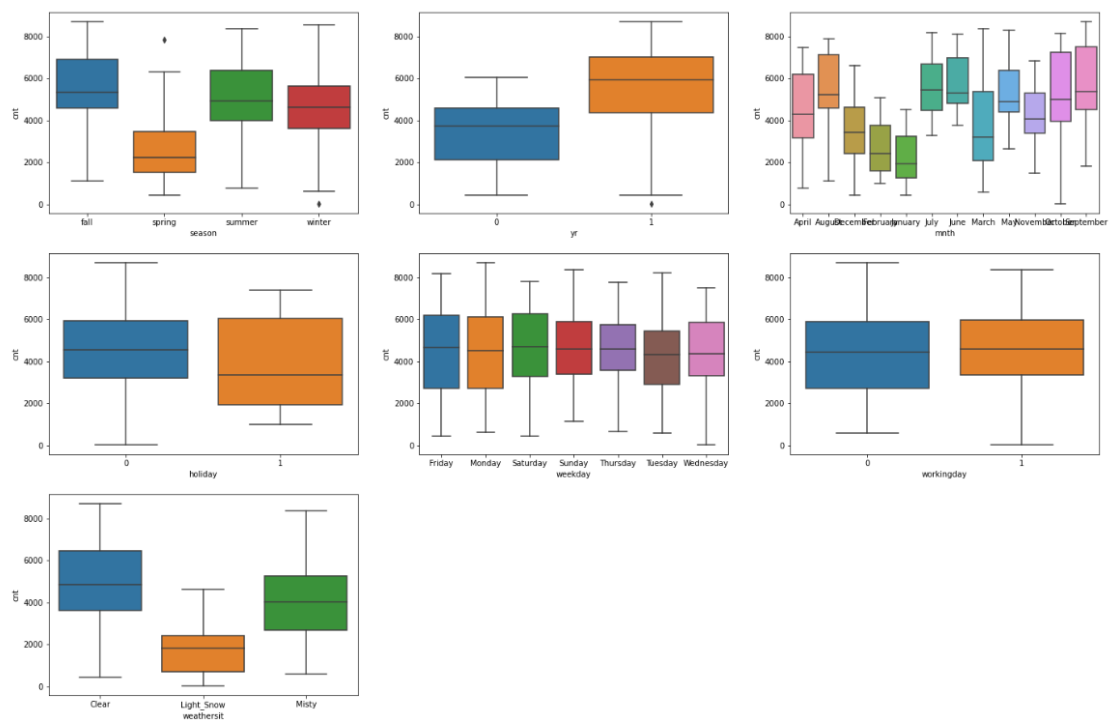
Assignment-based Subjective Question:

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



Answer: Based on the above plots these are the inferences that we can take:

1. Fall has the maximum rentals followed by summer and winter seasons. There is some trend here and could be used for predictor for dependent variable
2. Months July, June and November, December have the most rentals
3. Maximum bookings take place during clear weather situation and it falls rapidly during light snow. This shows some trend of weather situation with cnt
4. Most of bookings take place when there is no holiday
5. The rental for each week day remains the same, similar with working day
6. The bike rental increases rapidly from 2018 to 2019 meaning there could be a trend to predict the dependent variable

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

Answer: This is because drop_first = True removes redundant variables keeping the model lean and improve performance. For a categorical variable with n ordinal values, only n-1 binary columns are required to identify the correct ordinal value.

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

Answer: temp (and atemp) has the highest correlation with the target variable cnt

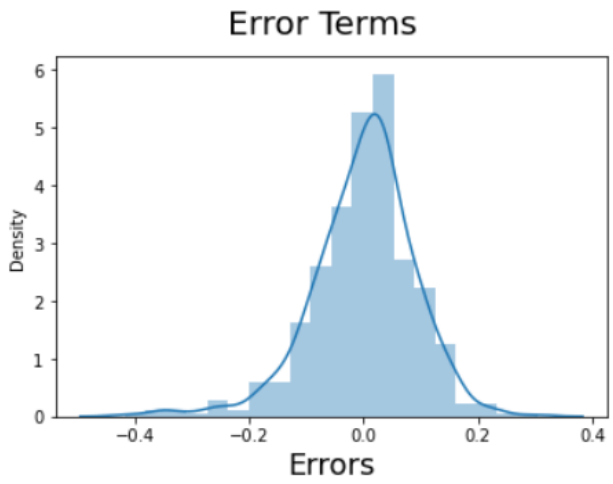
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

Assumption - Error terms are normally distributed with mean zero (not X,Y)

```
In [219]:
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
```

```
Out[219]: Text(0.5, 0, 'Errors')
```



As shown above the mean is zero and are normally distributed

Assumption - No multicollinearity between the predictor variables:

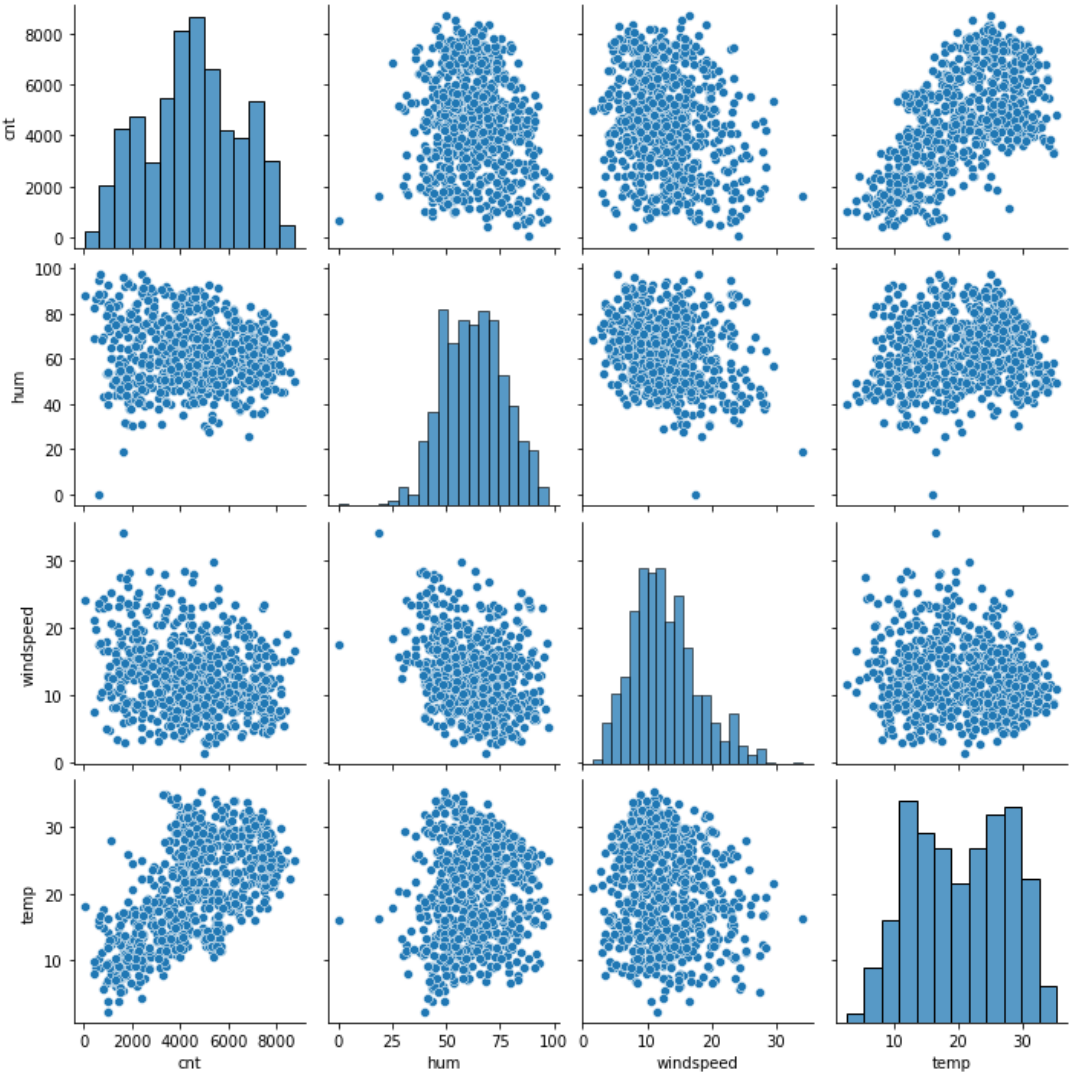
In [322]:
```
# Create a dataframe that will contain the names of all the featur
vif = pd.DataFrame()
vif['Features'] = X_train_modified_latest.columns
vif['VIF'] = [variance_inflation_factor(X_train_modified_latest.va
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

Out[322]:

	Features	VIF
2	temp	5.10
1	workingday	4.28
3	windspeed	3.57
0	yr	2.05
8	weekday_Monday	1.75
4	season_summer	1.63
10	weathersit_Misty	1.55
5	season_winter	1.47
6	mnth_January	1.29
7	mnth_September	1.20
9	weathersit_Light_Snow	1.08

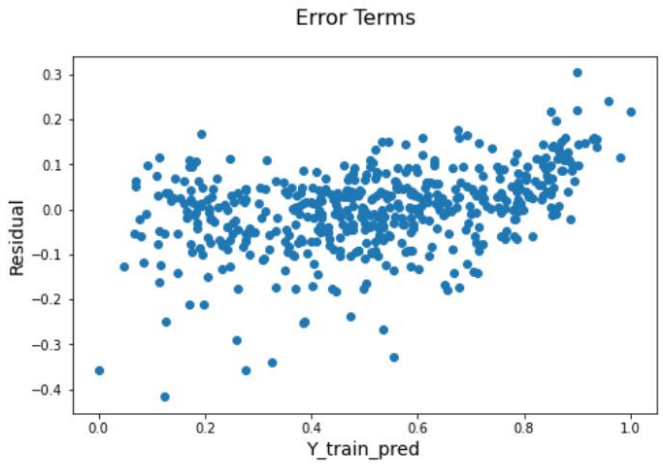
VIF values are close to 5 and below for all predictor variables

Assumption – There is a linear relationship between X and Y



There is a linear relationship between temp and cnt

Assumption - Error terms have constant variance (homoscedasticity)



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

Answer:

Following are the coefficients:

Final Coefficients

```
In [336]: lr_5.params

Out[336]: const                   0.102228
          yr                      0.234049
          workingday              0.056333
          temp                    0.519743
          windspeed              -0.163135
          season_summer           0.080151
          season_winter           0.120092
          mnth_January           -0.046352
          mnth_September          0.095030
          weekday_Monday          0.067922
          weathersit_Light_Snow  -0.289776
          weathersit_Misty       -0.080389
          dtype: float64
```

From here we can see that temp = 0.519, yr = 0.234 and weathersit_Light_snow = -0.289 have the highest contribution. While temp and yr influence it positively, weathersit_Light_snow contributes negatively

General Subjective Questions:

Question1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression performs a task for to predict a dependant variable value based on one or multiple independent variable.

y = beta1 + beta2*x

Where beta1 is known as the intercept while beta2 is known as the coefficient of x.

When training the model – it fits the best line to predict the value of y for a given value of x.

Once we get the value of beta1 and beta2, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Now the dependent variable y can be explained by more than 1 independent variable (which is the case in most cases), in which case the formula becomes:

y = beta + beta_1*x1 + beta_2*x2 + ….. + beta_n*xn

where beta_1…beta_n are the coefficients which explain the variation in y

There are also some assumption of simple linear regression:

a. Linear relationship between x and y
b. Error terms are normally distributed (and not x,y)
c. Error terms are independent of each other
d. Error terms have constant variance (homoscedasticity)

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

Answer: 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.

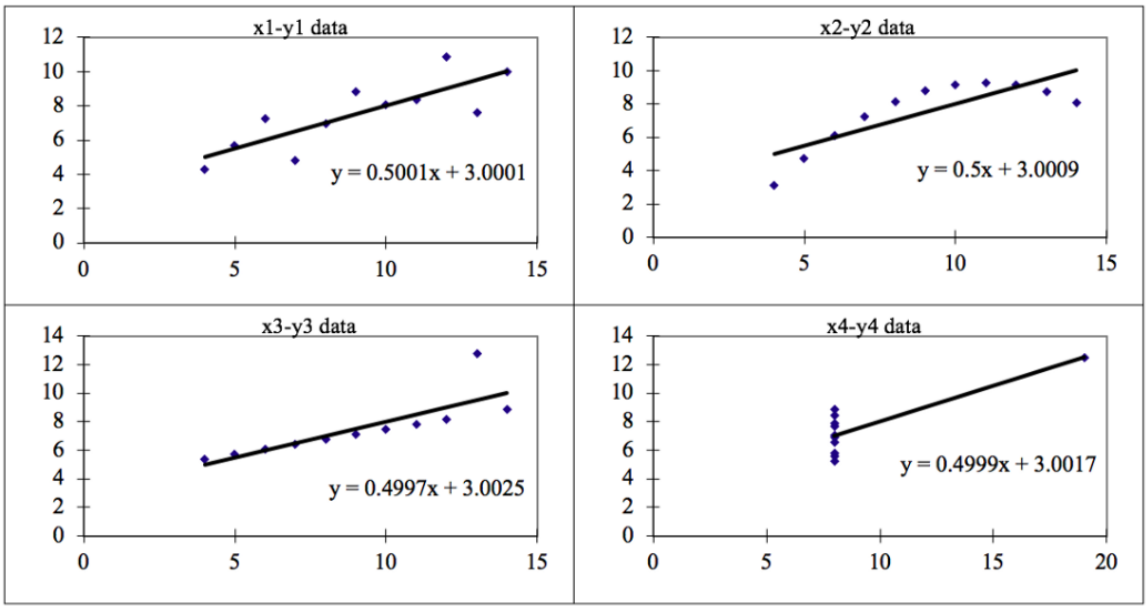
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data 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.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are similar and can be computed as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm



Here, only 1 can fit linear regression model very well.

Question 3. What is Pearson's R?

Answer: Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their

association with each other. In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

Example:

Positive linear relationship – Height of a child increases with age

Negative linear relationship – Faster we run lesser is the time to complete certain distance

Formula:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Following is the way the coefficient is interpreted:

Small – 0.1<r<0.3 OR -0.3<r<-0.1

Medium - 0.3<r<0.5 OR -0.5<r<-0.3

Large - 0.5<r<=1.0 OR -1.0<=r<-0.5

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

Answer: Scaling is a step of data pre-processing which is applied to independent variables to limit the data within a particular range.

Scaling is performed since collected data set more often or not contain features with varying magnitudes (0,1 to 1000 or 1000000 etc.). If scaling is not performed then algorithm only takes magnitude into account and not units and hence the coefficients come out to be wrong. If there is a requirement that the business requires not only the predictor model but the coefficients of the independent variables as well, scaling is required. It has an added benefit that it speeds up calculation in the algorithm being applied.

Note that *scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.*

Normalized Scaling – It brings all the data in the range of 0 and 1

Formula – x = (x-Min(x))/max(x)-min(x)

Standardised Scaling – Standardization replaces the values by their z scores. It brings all of the data into a standard normal distribution which has mean = 0 and standard deviation = 1

Formula – x= (x-mean(x))/sd(x)

Standardization maintains useful information about outliers and makes the algorithm less sensitive to them in contrast to min-max scaling

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

Answer: VIF is defined by 1/(1-R2) where R2 is the coefficient of determination from linear regression model
Now, if R2 = 1 then VIF is infinite
R2 comes from the following linear regression equation for an independent variable:

X_1 = beta_0 + beta_1*X_2 + beta_2*X_3 + error terms

Here, when R2 is 1 this means that the independent variables X_1, X_3 are exactly predictive of X_1 (or they are perfectly linearly related). Dropping X_1 from the dataset will not have any issues since X_2 and X_3 can take its place for prediction. If we don't drop it the contribution of X_1 would be like doubling the combined coefficients for X_1, X_2 and X_3.

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

Answer: Quantile-Quantile (Q-Q) plot are plots of two quantiles with each other. 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.

It is used to check following scenarios:

If two data sets —

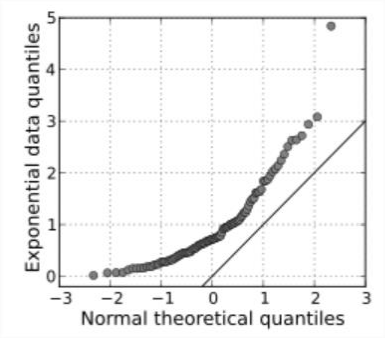
a. come from populations with a common distribution

b. have common location and scale

c. have similar distributional shapes

d. have similar tail behavior

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the 45 degree line. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line but not exactly on the line.

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 60 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be useful for checking if the split between train and test datasets distributions is the same because if it is not, it will not provide the expected predictions.