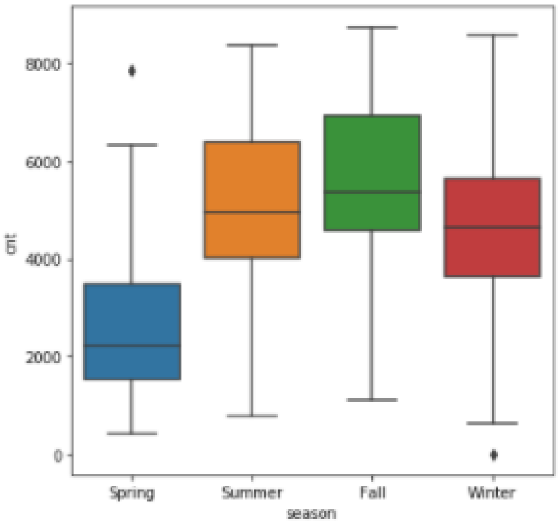
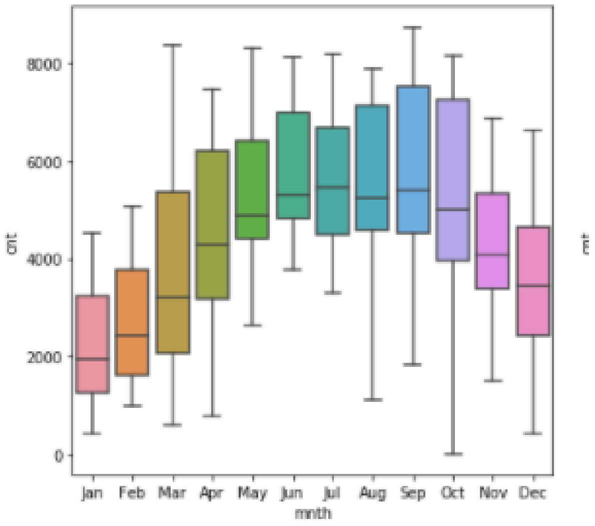
Assignment-based Subjective Questions

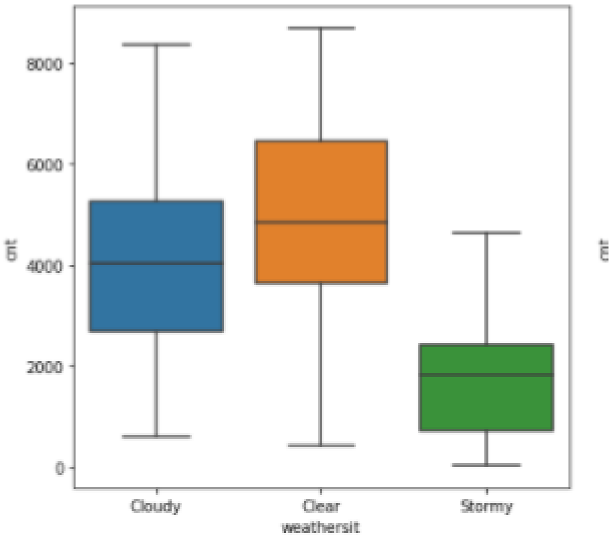
Q1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

1. From The boxplot we can see that most number of bikes were rented in fall and summer season.
2. There is a lesser demand to rent bikes in the months of Jan-Feb-March. The bike demand increases in the middle of the year (June-July) and again sees a dropping trend in the end of the year.
3. People hire bikes when the skies are clear (Clear, Few clouds, Partly cloudy, Partly cloudy)







Q2 Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: Consider the variable ‘season’ in the dataset. It has 4 possible values: Spring, Summer, Fall, Winter. As our machine can’t understand categorical values, we encode the season values in 0’s and 1’s

	Summer	Fall	Winter	Spring
Summer	1	0	0	0
Fall	0	1	0	0

Winter	0	0	1	0
Spring	0	0	0	1

We have created 4 variables to represent 4 seasons. A combination of ‘000’ can also be used to identify 4 variables.

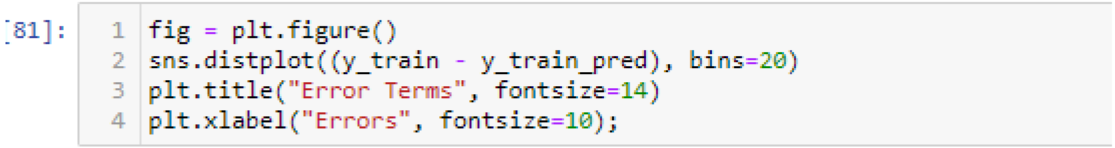
	Summer	Fall	Winter
Summer	1	0	0
Fall	0	1	0
Winter	0	0	1
Spring	0	0	0

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: ‘registered’ has the highest correlation with ‘cnt’

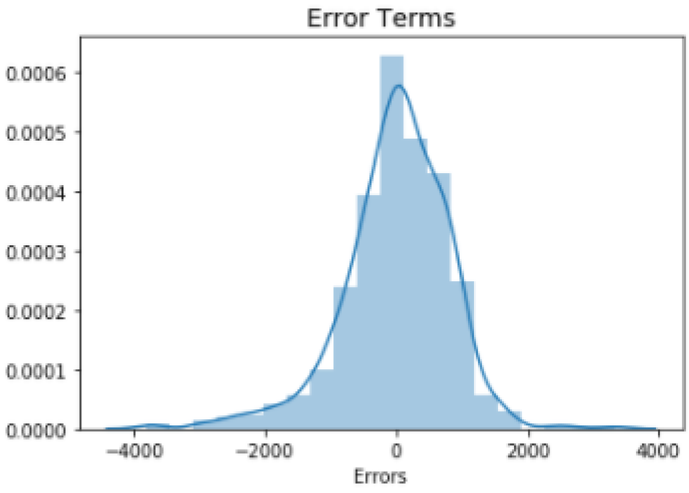
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: 1. We see that the error terms are normally distributed.

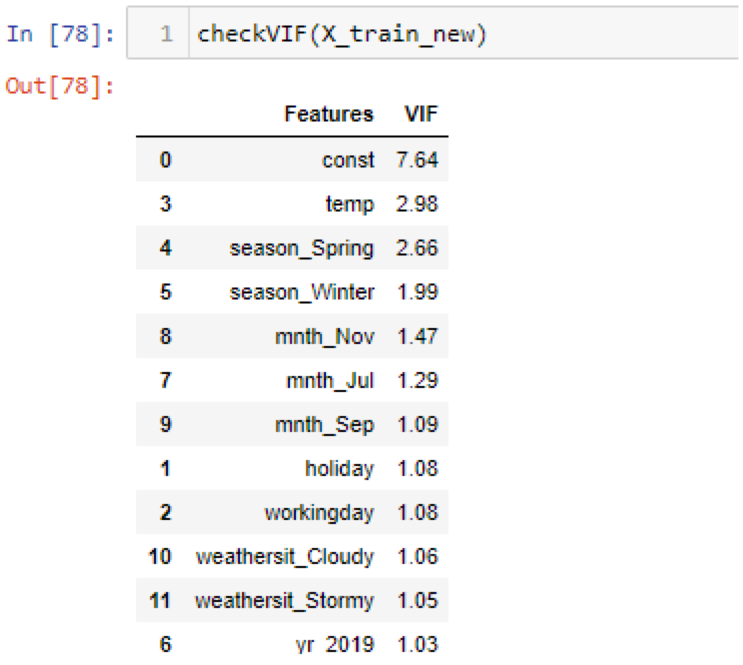


[81]:

```python
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins=20)
plt.title("Error Terms", fontsize=14)
plt.xlabel("Errors", fontsize=10);
```



2. The independent variables are not correlated. All independent variables have VIF<5. Thus eliminated multicollinearity.



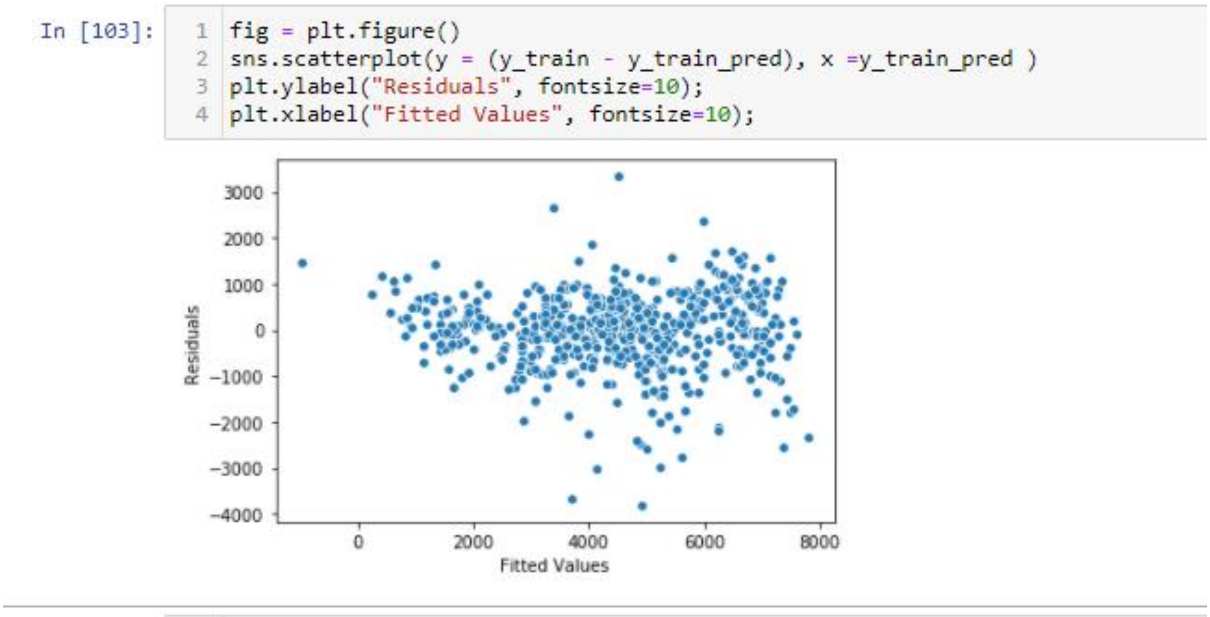
In [78]:

```
checkVIF(X_train_new)
```

Out[78]:

	Features	VIF
0	const	7.64
3	temp	2.98
4	season_Spring	2.66
5	season_Winter	1.99
8	mnth_Nov	1.47
7	mnth_Jul	1.29
9	mnth_Sep	1.09
1	holiday	1.08
2	workingday	1.08
10	weathersit_Cloudy	1.06
11	weathersit_Stormy	1.05
6	yr_2019	1.03

3. Homo**skedasticity**



In [103]:

```
fig = plt.figure()
sns.scatterplot(y = (y_train - y_train_pred), x =y_train_pred )
plt.ylabel("Residuals", fontsize=10);
plt.xlabel("Fitted Values", fontsize=10);
```

We see that the variation in error terms is constant and mean is around 0.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

```
Ans: temp, season(season_Spring, season_Winter)
, weathersit(weathersit_Cloudy, weathersit_Stormy)
)
```

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

1. Linear Regression is a type of Supervised Learning method. It attempts to explain the relationship between dependent and independent variables.
2. Equation of best fit line $Y = \beta_0 + \beta_1 X$ is found by minimising the residual sum of squares (RSS).
3. Gradient descent algorithm is used to optimise the objective function to reach the optimal solution.
    a. It starts by taking random values for each coefficient. The sum of squared errors is calculated for each pair of input and output values. Learning rate is used as a scale factor and coefficient is updated in direction towards minimising the error. The process is repeated until minimum sum of squared errors is reached.

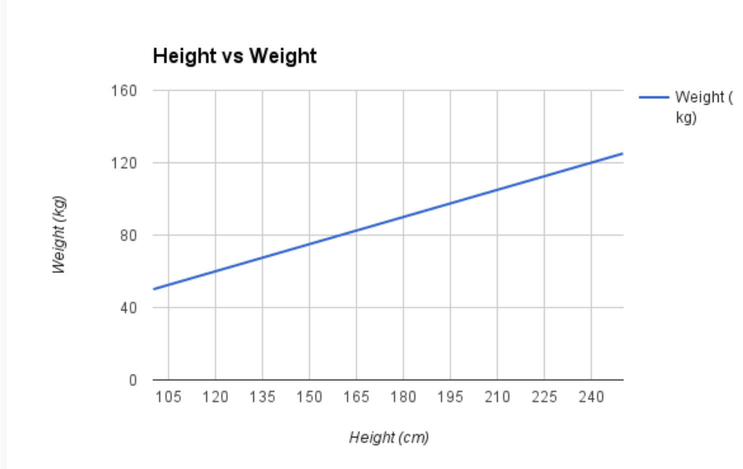
Making Predictions with Linear regression:

Imagine we are predicting weight(y) from height(x). Linear regression problem would be:

y = B0 + B1 * x1 or,

weight = B0+B1*height

We use a learning technique to find a good set of coefficient values. Once found we can give different height values to get the weight.



Sample Height vs Weight Linear Regression

4. Linear regression can be classified into two types
    a. Simple Linear Regression: Number of independent variables is 1
    b. Multiple Linear Regression: Number of independent variables is more than 1.
5. R^2 measures the goodness of linear regression model. It represents the percentage of variance denoted by the model.
6. Aspects to consider in Multiple Linear Regression:
    a. Overfitting: Model is said to overfit when training accuracy is high and test accuracy is low.
    b. Multicolinearity: Associations between dependent variables.
    c. Feature Selection: Selecting optimal features from a given set of features.

7. Multicollinearity

Multicollinearity can be dealt in the following ways:

a. Looking at pairwise correlations
b. Checking Variance Inflation factor: when instead of one variable, independent variable is dependent on the combination of other variables.
    i. If VIF > 10: The variable is eliminated

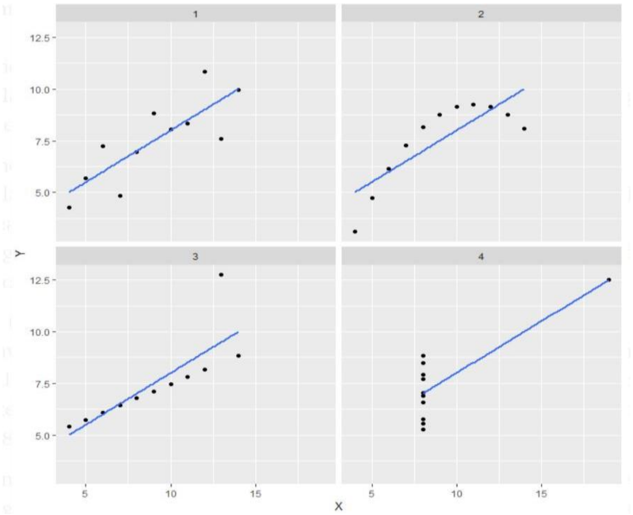
ii. VIF > 5: can be okay

iii. VIF < 5 Good VIF value. No need to eliminate this variable.

8. **Dummy Variables**: Non numeric variables can’t be fed directly into the model. If we have a categorical variable with n levels, the idea of dummy variable creation is to create n-1 variables.

Q2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe’s quartlet refers to four datasets that have nearly identical statistic properties and appear very different when graphed.



Top Left: There seems to be a linear relationship between x and
Top Right: Non-linear relationship between x and y
Bottom Left: Perfect linear relationship between x and y except one which seems to be an outlier.
Bottom Right: A high leverage point Is enough to produce a high correlation
Coefficient.
It emphasises on the need of analysing the dataset graphically before going on to describe the statistical relationship

Q3. What is Pearson’s R?

Ans: Pearson’s R coefficient determines strength and direction of a linear relationship
The correlation coefficient ranges from -1 to +1.
1 implies perfect linear relationship while 0 implies no relationship
If both increase or decrease together, correlation is positive, upward slope
If one increases and other decreases, correlation is negative, downward slope

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

1. Feature Scaling standardises features into a fixed scale. It is used to handle highly varying magnitudes or scales.
2. Scaling helps in faster convergence of algorithm in case of Gradient Descents.

Scaling is used when the result change depends on the units of data, when distance based metrics are involved as in Support vector machines (SVMs) or K nearest neighbours (k nn). Choosing not to scale data in either of the cases may lead to drastically misleading results.

Scaling can be done in two ways:

Min-Max Normalization: Rescales a feature with distribution value between 0 and 1

$$Xsc=X-Xmin/Xmax-Xmin.$$

Standardization: Rescales a feature such that mean is 0 and standard deviation is 1.

It is completed by taking each value of column, subtracting mean of column and dividing by standard deviation of column

$$Z = (x-mu)/sigma$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF is a measure of how much variance of estimated regression coefficient increases due to collinearity

We fit a model between one independent variable with other independent variables to determine R1 and use this value to estimate VIF.

$X1 = c+ a2\ X2 + a3\ X3 +a4\ X4 \ldots$

$VIF1=1/(1-R1^2)$

$X2 = c+ a1\ X1 + a3\ X3 +a4\ X4 \ldots$

$VIF2=1/(1-R1^2)$

Infinite VIF means perfect collinearity between independent variables.

If VIF is too large, we need to take corrective measures before we use multiple regression.

We can do this by identifying terms that are duplicates and not adding value to explain variation in the model.

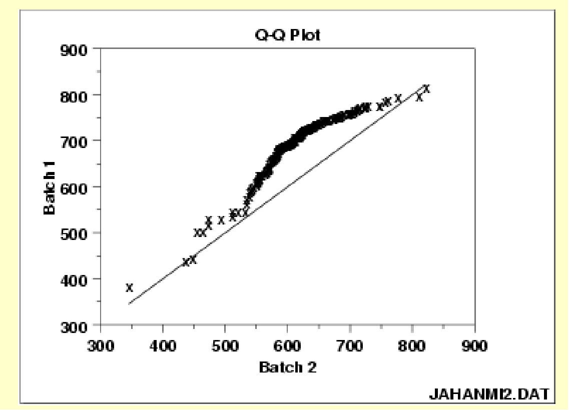
Dropping a variable with large VIF will hopefully fix the VIF of other columns to be under threshold limits. If not so, we may need to drop other columns as required.

Q6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Ans: QQ plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

If both sets of quantiles come from same distribution, we see that the points form a line that is roughly straight.

It helps us check that the data meets assumptions of normality. It compares the distribution of data to a normal distribution by plotting quartiles of data against quartiles of normal distribution. If data is normally distributed then it should form an approximately straight line.



Consider the above Q-Q plot of two different batches in a Lab test.

1. It could be seen that the two different batches do not come from a distribution of common population (as they do not fall on 45 degree reference line)
2. Batch 1 values are significantly higher than Batch2 values.