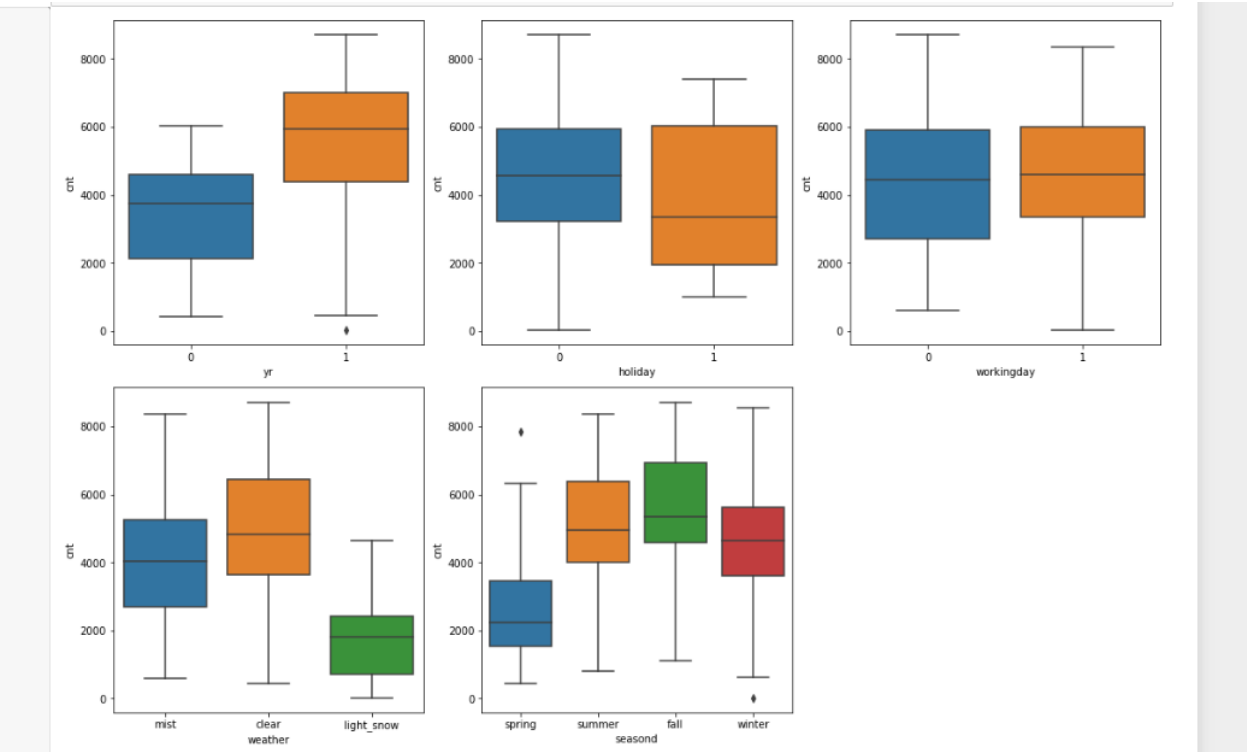
Assignment-based Subjective Questions

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

A1: The below insights can be derived from the dataset regarding the effect of categorical variables on dependent variable:

- Demand of bike is more in 2019 as compared to 2018
- Demand of bikes are high for clear weather and comparatively similar for misty weather. However, it is considerably low for light snow.
- Demand of bike is similar for working day or weekend.
- Demand of bike is higher in fall and summer. Less in Spring as compared to the other seasons.



Q2: Why is it important to use drop_first=True during dummy variable creation?

A2: Using drop_first= True helps us to remove the redundant dummy variable.

For example, without using *drop_first=True*, we get the below four variables. However, the same information can be described from the 3 variables – spring, summer, and winter. Variable “fall” is therefore redundant.

In [231]:
```
season_status_redundant=pd.get_dummies(df_data['seasond'])
#season_status=pd.get_dummies(df_data['seasond'], drop_first='True')
season_status_redundant.head()
#season_status.head()
```

Out[231]:

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

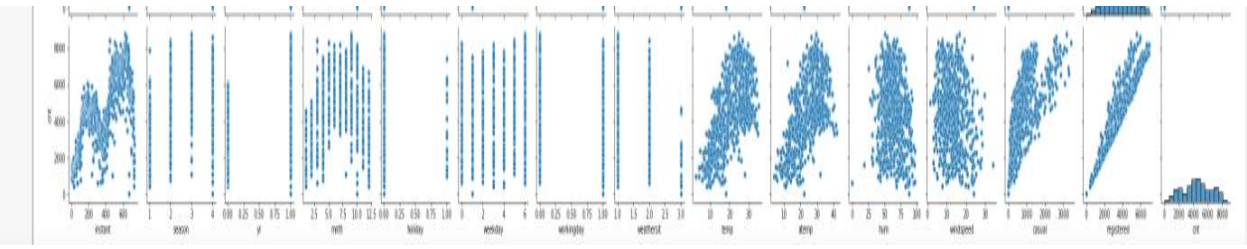
In [232]:
```
season_status=pd.get_dummies(df_data['seasond'], drop_first='True')
season_status.head()
```

Out[232]:

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

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

A3: Looking at the paid-plot as below, target variable “cnt” has the highest correlation with “atemp” i.e. 0.630685

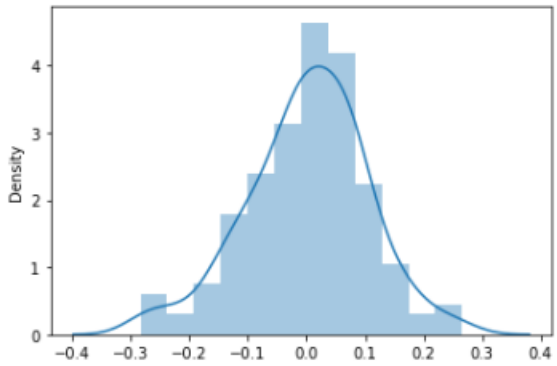


Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

A4: The assumptions of Linear Regression were validated using the following steps:

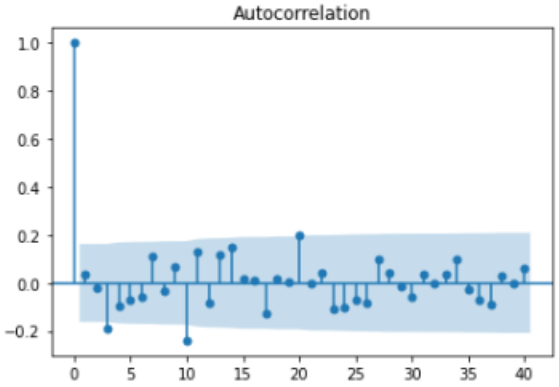
- Residual Plot is Normal distribution

```
In [424]: residual = y_test - y_test_pred
          sns.distplot(residual)
Out[424]: <AxesSubplot:ylabel='Density'>
```



- Correlation

```
In [429]: acf = smt.graphics.plot_acf(residual, lags=40 , alpha=0.05)
          acf.show()
```



Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5: The top three features are:

1. Temperature
2. Year
3. Weather – Light Snow (Negative)

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

A1: Linear Regression model explains the relationship between dependent and independent variables using a straight line. The standard equation that explains the model is as below:

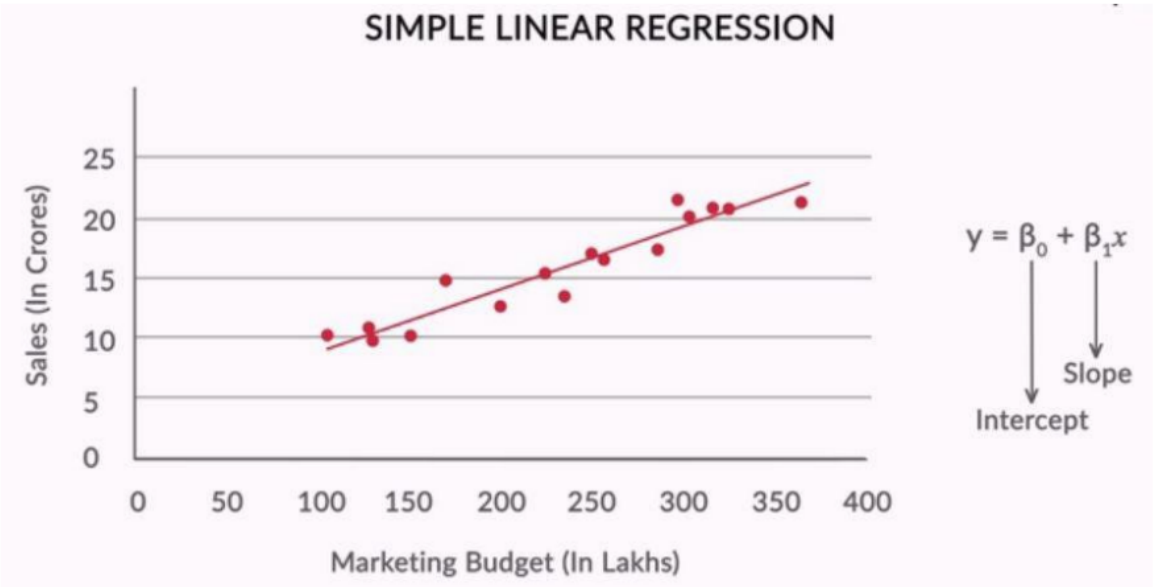


Figure 3 - Regression Line

Assumptions:

- Y and X are linearly related
- Errors/Residuals are normally distributed and independent of each other
- Errors should have constant variance (homoscedasticity)

The idea is to fit the line on the datapoints and understand if the line fits the dataset "significantly". This is achieved used NULL Hypothesis on the beta coefficient. Usually P-value method is used to find the significance of the fit and therefore the variables.

Features are also scaled using Min-Max Scaling to ensure that coefficients of multiple dependent variables are comparable and portray the right information.

Q2: Explain the Anscombe's quartet in detail

A2: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset comprises of 11 datasets which have the same

- Mean
- Standard Deviation
- Correlation

However, when plotted on a graph, the datasets are completely different in nature and therefore used to illustrate the importance of understanding the data visually instead of directly calculating the stats variables –

Dataset:

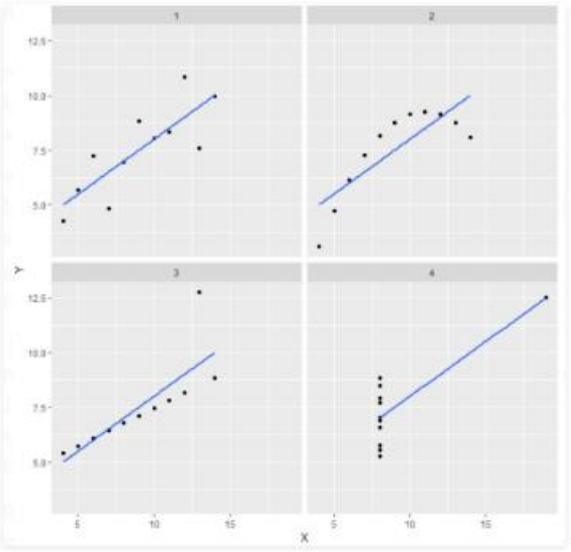
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Stats:

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

Graphical Representation of Data:



Observations:

- Dataset 1 is linear Relationship

- Dataset 2 is non-linear
- Dataset 3 is linear except outlier values
- Dataset 4 shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Q3: What is Pearson's R?

A3: Pearson correlation coefficient known as Pearson's r, the Pearson product-moment correlation coefficient, is a measure of linear correlation between two sets of data.

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

A4: Scaling is the process of resetting the variables on a comparable scale especially when lot of independent variables are involved.

Scaling is performed mainly for the below reasons:

- Ease of interpretation
- Faster convergence for gradient descent method

Standardized Scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.

Normalized Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

Q5: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A5: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.