In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

Load the data.

In [2]:
```
data = pd.read_csv('http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv', index_col=0)
data.head()
```

Out[2]:

	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

Amount (in throusand dollars) spent on different types of media advertising. Response variable is sales of items.

In [3]:
```
%matplotlib inline
```

In [4]:
```
data.shape #shape of the dataframe
```

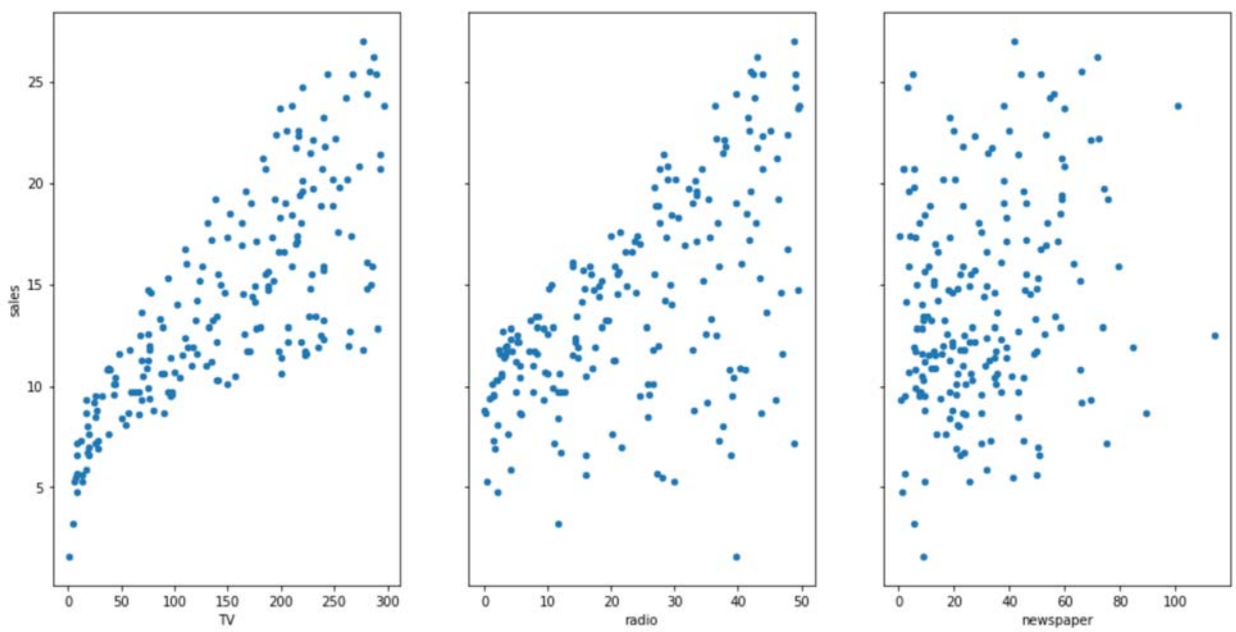
Out[4]: (200, 4)

There are 200 observations and 4 variables in the dataset.

Create scatterplots to visualize the relationship between each independent variable and dependent (reponse) variable.

In [7]:
```
fig, axs = plt.subplots(1, 3, sharey=True)
data.plot(kind='scatter', x='TV', y='sales', ax=axs[0], figsize=(16, 8))
data.plot(kind='scatter', x='radio', y='sales', ax=axs[1])
data.plot(kind='scatter', x='newspaper', y='sales', ax=axs[2])
```

Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x21edb46bbe0>



In [8]:
```
import statsmodels.formula.api as smf
```

Create a linear regression model

In [10]:
```
lm = smf.ols(formula='sales ~ TV', data=data).fit()
```

In [11]:
```
lm.params #print coefficients
```

Out[11]:
```
Intercept    7.032594
TV           0.047537
dtype: float64
```

One unit of TV spending increases 0.047 units of sales.

predicted value of y = a + bx = 7.032594 + 0.047537x

How much sales can we expect if we spend $100,000 on TV ads based on the regression equation? You can calculate manually.

In [14]:
```
7.032594 + 0.047537*100
```

Out[14]:
```
11.786294
```

11.7 thousand units

This time predict using pandas.

In [17]:
```
X_new = pd.DataFrame({'TV': [100]})
X_new.head()
```

Out[17]:

	TV
0	100

In [18]:
```
lm.predict(X_new)
```

Out[18]:
```
0    11.786258
dtype: float64
```

Plot a regression line using the OLS - least squares.

In [19]:
```
X_new = pd.DataFrame({'TV': [data.TV.min(), data.TV.max()]})
X_new.head()
```

Out[19]:

	TV
0	0.7
1	296.4

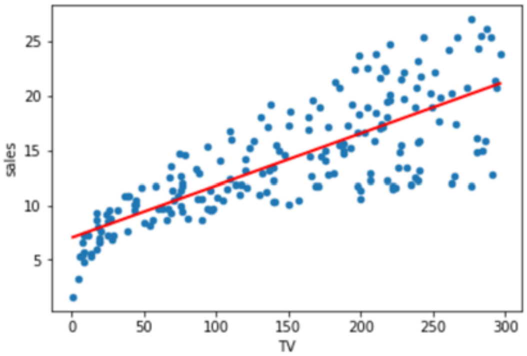
In [20]:
```
preds = lm.predict(X_new)
preds
```

Out[20]:
```
0     7.065869
1    21.122454
dtype: float64
```

In [22]:
```
# first, plot the observed data
data.plot(kind='scatter', x='TV', y='sales')

# then, plot the least squares line
plt.plot(X_new, preds, c='red', linewidth=2)
```

Out[22]: [<matplotlib.lines.Line2D at 0x21edd43b208>]



In [24]:
```
lm.conf_int() #confidence intervals - 95% confidence intervals
```

Out[24]:

	0	1
Intercept	6.129719	7.935468
TV	0.042231	0.052843

In [25]:
```
lm.pvalues #check for p-values
```

Out[25]:
```
Intercept    1.406300e-35
TV           1.467390e-42
dtype: float64
```

p-value for TV is far less than 0.05

In [26]:
```
lm.rsquared #calculate r squared
```

Out[26]: 0.611875050850071

The R squared value is fairly good. You can use r squared to comapre different models.

In [27]:
```
lm = smf.ols(formula='sales ~ TV + radio + newspaper', data=data).fit() #Create a mutiple regression model
```

In [28]:
```
lm.params
```

Out[28]:
```
Intercept    2.938889
TV           0.045765
radio        0.188530
newspaper   -0.001037
dtype: float64
```

In [29]:
```
lm.summary() #summary
```

Out[29]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Mon, 26 Mar 2018	**Prob (F-statistic):**	1.58e-96
Time:	20:31:55	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	**Durbin-Watson:**	2.084
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	151.241
Skew:	-1.327	**Prob(JB):**	1.44e-33
Kurtosis:	6.332	**Cond. No.**	454.

The above model has higher r squared compared to the previous model. So this model is a better fit.

TV and Radio have higher p-values (~0.05) thus we can reject the null hypothesis for TV and Radio that there is no association between them and sales. The p-value for newspaper is low so we fail to reject the null hypothesis for newspaper. TV and Radio ad spending are both positively associated with sales, while newspaper ad spending is slightly negatively associated with sales.

You may try different models, and only keep predictors in the model if they have small p-values. It should also increase r squared.

Since regression is prone to overfitting, you should cross-validate your model. You can use scikit learn for this.