Here are some recommended packages, not all are required and depends on your solution.

In [2]:
```
# imports
import pandas as pd
import seaborn as sns
import statsmodels.formula.api as smf
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.model_selection import train_test_split
import numpy as np

# allow plots to appear directly in the notebook
%matplotlib inline
```

Questions

You are a consultant for a company that sells widgets. They have historical data on their sales on their investments in advertising in various media outlets, including TV, radio, and newspapers. On the basis of this data, how should they be spending their advertising money in the future?

Your analysis should answer the following questions:

Is there a relationship between ads and sales?

How strong is that relationship?

Which ad types contribute to sales?

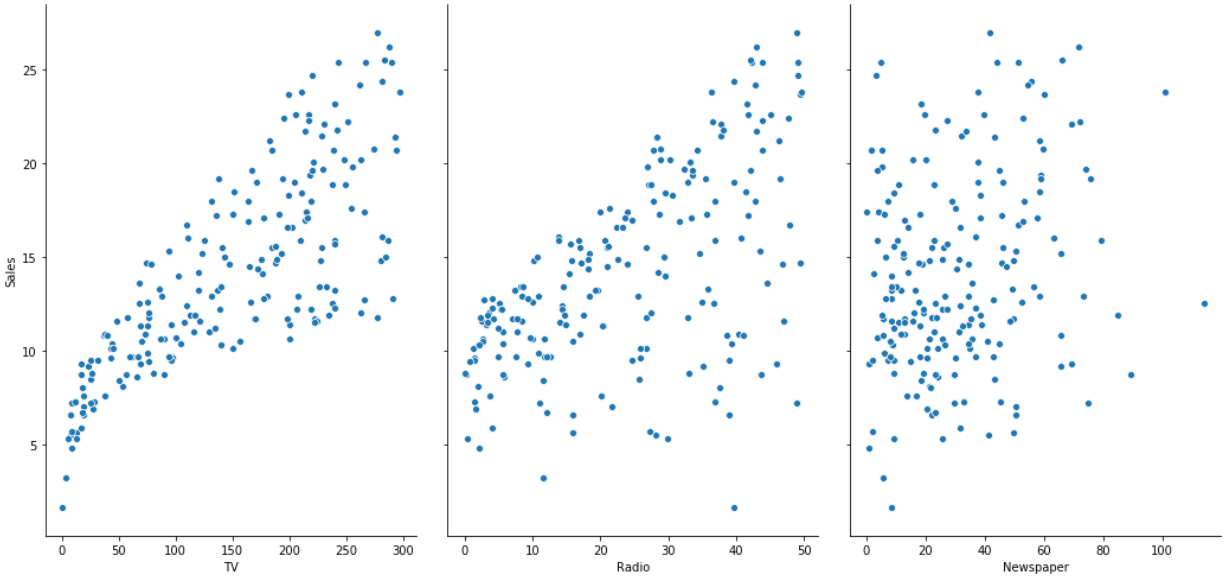
What is the effect of each ad type of sales?

Given ad spending in a particular market, can sales be predicted?

In [3]:
```
# read data into a DataFrame, this is money spent on different medias
data = pd.read_csv('https://raw.githubusercontent.com/lneisenman/isl/master/data/Advertising.csv', index_col=0)
print(data.head())
# visualize the relationship between the features and the response using scatterplots
sns.pairplot(data, x_vars=['TV','Radio','Newspaper'], y_vars='Sales', height=7, aspect=0.7)
```

```
      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
```

Out[3]: <seaborn.axisgrid.PairGrid at 0x2a957e0e8d0>



In the lecture, we covered how to perform a linear regression model. We did not however explore how "good" this model is. The task below will have you identifying ways to evaluate a linear regression model.

Machine learning focuses on what the model predicts. If you would like to dive into the meaning of fit parameters within the model, other tools are available, including the Statsmodels Python package. Take some time to look at this package (https://www.statsmodels.org/stable/regression.html) and also an example of evaluating a linear regression (https://www.statsmodels.org/stable/examples/notebooks/generated/gls.html).

Similar to Scikit-learn, one can calculate the intercept and coefficient for a linear fit for a set of data.

In [4]:
```
list_of_comparisons = ['TV ~ Sales', 'Radio ~ Sales', 'Newspaper ~ Sales']
for x in list_of_comparisons:
    print(x)
    result = smf.ols(formula=x, data=data).fit()
    print(result.params)
```

```
TV ~ Sales
Intercept   -33.450228
Sales        12.871651
dtype: float64
Radio ~ Sales
Intercept    0.271298
Sales        1.639701
dtype: float64
Newspaper ~ Sales
Intercept    17.191090
Sales         0.952962
dtype: float64
```

A confidence interval can be used to describe a linear model. How would you calculate the confidence interval of this model and what does this confidence interval mean?

In [10]:
```
list_of_comparisons = ['TV ~ Sales', 'Radio ~ Sales', 'Newspaper ~ Sales']

for x in list_of_comparisons:
    print(x)
    result = smf.ols(formula=x, data=data).fit()
    print(result.conf_int(alpha=0.05))
```

```
TV ~ Sales
                   0          1
Intercept -54.939198 -11.961258
Sales      11.434949  14.308354
Radio ~ Sales
                  0         1
Intercept -4.603750  5.146346
Sales      1.313766  1.965635
Newspaper ~ Sales
                  0          1
Intercept  8.672356  25.709824
Sales      0.383419   1.522505
```

Other metrics that are used to describe the appropriateness of a model is a p-value. How would you calculate the p-value and r-squared values of the model? What do these values mean?

In [19]:
```
list_of_comparisons = ['TV ~ Sales', 'Radio ~ Sales', 'Newspaper ~ Sales']

for x in list_of_comparisons:
    print(x)
    result = smf.ols(formula=x, data=data).fit()
    print(result.summary())
```

```
TV ~ Sales
                            OLS Regression Results
==============================================================================
Dep. Variable:                     TV   R-squared:                       0.612
Model:                            OLS   Adj. R-squared:                  0.610
Method:                 Least Squares   F-statistic:                     312.1
Date:                Thu, 03 Oct 2019   Prob (F-statistic):           1.47e-42
Time:                        09:38:59   Log-Likelihood:                -1079.2
No. Observations:                 200   AIC:                             2162.
Df Residuals:                     198   BIC:                             2169.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -33.4502     10.897     -3.070      0.002     -54.939     -11.961
Sales         12.8717      0.729     17.668      0.000      11.435      14.308
==============================================================================
Omnibus:                       21.952   Durbin-Watson:                   1.973
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.224
Skew:                           0.882   Prob(JB):                     2.02e-06
Kurtosis:                       3.193   Cond. No.                         43.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie
d.
Radio ~ Sales
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Radio   R-squared:                       0.332
Model:                            OLS   Adj. R-squared:                  0.329
Method:                 Least Squares   F-statistic:                     98.42
Date:                Thu, 03 Oct 2019   Prob (F-statistic):           4.35e-19
Time:                        09:38:59   Log-Likelihood:                -782.49
No. Observations:                 200   AIC:                             1569.
Df Residuals:                     198   BIC:                             1576.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.2713      2.472      0.110      0.913      -4.604       5.146
Sales          1.6397      0.165      9.921      0.000       1.314       1.966
==============================================================================
Omnibus:                       15.769   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.838
Skew:                           0.732   Prob(JB):                     0.000134
Kurtosis:                       2.991   Cond. No.                         43.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie
d.
Newspaper ~ Sales
                            OLS Regression Results
==============================================================================
Dep. Variable:              Newspaper   R-squared:                       0.052
Model:                            OLS   Adj. R-squared:                  0.047
Method:                 Least Squares   F-statistic:                     10.89
Date:                Thu, 03 Oct 2019   Prob (F-statistic):            0.00115
Time:                        09:38:59   Log-Likelihood:                -894.12
No. Observations:                 200   AIC:                             1792.
Df Residuals:                     198   BIC:                             1799.
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     17.1911      4.320      3.980      0.000       8.672      25.710
Sales          0.9530      0.289      3.300      0.001       0.383       1.523
==============================================================================
Omnibus:                       24.387   Durbin-Watson:                   1.882
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.955
Skew:                           0.834   Prob(JB):                     3.13e-07
Kurtosis:                       3.902   Cond. No.                         43.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie
d.
```

Is there a relationship between ads and sales?
Based on the pairplot visualization, Sales appeared to increased with TV and Radio ads, but not newspaper ad.

How strong is that relationship?
Based on R2 values, there was a strong relationship between TV ad and sales. A slighlty weaker relationship between Radio ad and sales was also observed. There was no relationship between newspaper ad and sales.

Which ad types contribute to sales?
Based on the slopes of the models, TV and Radio ads had the strongest contribution to sales.

What is the effect of each ad type on sales?
Positive relationships were found between TV and Radio ads on Sales. Newspaper ad did not have much effect on sales

Given ad spending in a particular market, can sales be predicted?
Based on P>|t| values and alpha = 0.05, all models can be used to predict the market sales.
The model equation of the prediction for each ad is:
1) TV: Sales = -33.45 + (12.87 * TV_ad) (note that the negative intercept doesn't mean much, because sales can't go below 0)
2) Radio: Sales = 0.27 + (1.64 * Radio_ad) (note that the y-intercept is not significant, so use with caution)
3) Newspaper: Sales = 17.19 + (0.95 * Newspaper_ad)

In [ ]: