In [1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
```

```
Linear regression model steps -
1) Import labrayies (p,n,m,s)
2) For checking all values on histogram.(df2.hist(bins=40, figsize=(20,15),color ='c')
   plt.show())
2) Visulise the data using (%matplotlib inline)plt.scatter.
3) Split data using drop into 2 different dataset ie.X and Y (X = data_set.drop(['Salary'],axis=1)
   Y = data_set['Salary'])
4) Import sklearn (from sklearn.model_selection import train_test_split
   X_train, x_test, Y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=0))
5) Check X_train, x_test, Y_train, y_test
6) Import sklearn(from sklearn.linear_model import LinearRegression
   model = LinearRegression())
7) Fitting the model (model.fit(X_train,Y_train))
8) Checking the score (model.score(x_test, y_test))
9) Creat prediction variable (y_pred = model.predict([[1]]))
```

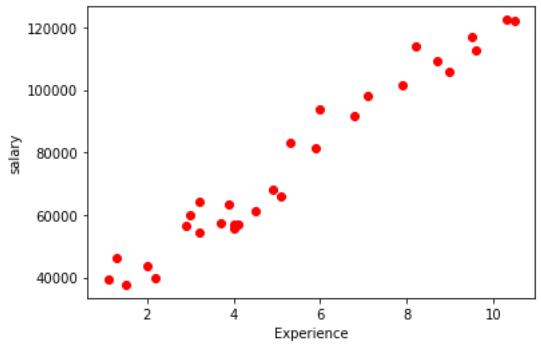
In [2]:
```
data_set = pd.read_csv('Salary_Data.csv')
```

In [3]:
```
data_set
```

Out[3]:

	YearsExperience	Salary
0	1.1	39343
1	1.3	46205
2	1.5	37731
3	2.0	43525
4	2.2	39891
5	2.9	56642
6	3.0	60150
7	3.2	54445
8	3.2	64445
9	3.7	57189
10	3.9	63218
11	4.0	55794
12	4.0	56957
13	4.1	57081
14	4.5	61111
15	4.9	67938
16	5.1	66029
17	5.3	83088
18	5.9	81363
19	6.0	93940
20	6.8	91738
21	7.1	98273
22	7.9	101302
23	8.2	113812
24	8.7	109431
25	9.0	105582
26	9.5	116969
27	9.6	112635
28	10.3	122391
29	10.5	121872

In [4]:
```
%matplotlib inline
plt.xlabel('Experience')
plt.ylabel('salary')
plt.scatter(data_set.YearsExperience, data_set.Salary, color='red')
plt.show()
```



In [5]:
```
X = data_set.drop(['Salary'],axis=1)
Y = data_set['Salary']
```

In [6]:
```
X = data_set.drop(['Salary'],axis=1)
```

In [8]:
```
X
```

Out[8]:

	YearsExperience
0	1.1
1	1.3
2	1.5
3	2.0
4	2.2
5	2.9
6	3.0
7	3.2
8	3.2
9	3.7
10	3.9
11	4.0
12	4.0
13	4.1
14	4.5
15	4.9
16	5.1
17	5.3
18	5.9
19	6.0
20	6.8
21	7.1
22	7.9
23	8.2
24	8.7
25	9.0
26	9.5
27	9.6
28	10.3
29	10.5

In [9]:
```
Y
```

Out[9]:
```
0      39343
1      46205
2      37731
3      43525
4      39891
5      56642
6      60150
7      54445
8      64445
9      57189
10     63218
11     55794
12     56957
13     57081
14     61111
15     67938
16     66029
17     83088
18     81363
19     93940
20     91738
21     98273
22    101302
23    113812
24    109431
25    105582
26    116969
27    112635
28    122391
29    121872
Name: Salary, dtype: int64
```

In [10]:
```
from sklearn.model_selection import train_test_split

X_train, x_test, Y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=0)
```

In [11]:
```
X_train
```

Out[11]:

	YearsExperience
27	9.6
11	4.0
17	5.3
22	7.9
5	2.9
16	5.1
8	3.2
14	4.5
23	8.2
20	6.8
1	1.3
29	10.5
6	3.0
4	2.2
18	5.9
19	6.0
9	3.7
7	3.2
25	9.0
3	2.0
0	1.1
21	7.1
15	4.9
12	4.0

In [12]:
```
Y_train
```

Out[12]:
```
27    112635
11     55794
17     83088
22    101302
5      56642
16     66029
8      64445
14     61111
23    113812
20     91738
1      46205
29    121872
6      60150
4      39891
18     81363
19     93940
9      57189
7      54445
25    105582
3      43525
0      39343
21     98273
15     67938
12     56957
Name: Salary, dtype: int64
```

In [13]:
```
x_test
```

Out[13]:

	YearsExperience
2	1.5
28	10.3
13	4.1
10	3.9
26	9.5
24	8.7

In [14]:
```
y_test
```

Out[14]:
```
2      37731
28    122391
13     57081
10     63218
26    116969
24    109431
Name: Salary, dtype: int64
```

In [15]:
```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
```

In [16]:
```
model.fit(X_train,Y_train)
```

Out[16]:
```
LinearRegression()
```

In [17]:
```
model.score(x_test, y_test)
```

Out[17]:
```
0.988169515729126
```

In [18]:
```
y_pred = model.predict([[1]])
```

In [19]:
```
y_pred
```

Out[19]:
```
array([36092.67427736])
```