The Sparks Foundation

Data Science & Business Analytics - June 2021

Task 1- Prediction using supervised ML

Problem Statement: **What will be predicted score if a student studies for 9.25 hrs/ day?**

Submitted by:-Awitijhya Chakraborty

Importing necessary libraries

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

%matplotlib inline
```

In [3]:
```
url="http://bit.ly/w-data"
data=pd.read_csv(url)
print('Data sucessfully loaded')
```

```
Data sucessfully loaded
```

In [4]:
```
data.head(10)
```

Out[4]:

	Hours	Scores
0	2.5	21
1	5.1	47
2	3.2	27
3	8.5	75
4	3.5	30
5	1.5	20
6	9.2	88
7	5.5	60
8	8.3	81
9	2.7	25

Understanding the data

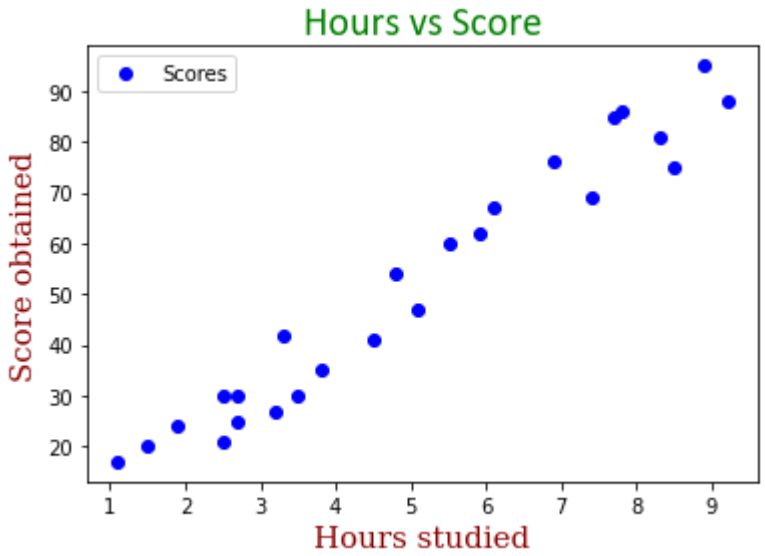
In [6]:
```
data.shape
```

Out[6]:
```
(25, 2)
```

In [8]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25 entries, 0 to 24
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Hours   25 non-null     float64
 1   Scores  25 non-null     int64  
dtypes: float64(1), int64(1)
memory usage: 528.0 bytes
```

In [12]:
```
font1={'family':'calibri','color':'green','size':20}
font2={'family':'serif','color':'darkred','size':15}
data.plot(x='Hours',y='Scores',style='o',c='blue')
plt.title('Hours vs Score',fontdict=font1)
plt.xlabel('Hours studied',fontdict=font2)
plt.ylabel('Score obtained',fontdict=font2)
plt.show()
```



In [13]:
```
data.corr()
```

Out[13]:

	Hours	Scores
Hours	1.000000	0.976191
Scores	0.976191	1.000000

In [14]:
```
data.isnull().sum()
```

Out[14]:
```
Hours     0
Scores    0
dtype: int64
```

In [15]:
```
x=(data['Hours'].values).reshape(-1,1)
y=data['Scores'].values
```

In [16]:
```
x
```

Out[16]:
```
array([[2.5],
       [5.1],
       [3.2],
       [8.5],
       [3.5],
       [1.5],
       [9.2],
       [5.5],
       [8.3],
       [2.7],
       [7.7],
       [5.9],
       [4.5],
       [3.3],
       [1.1],
       [8.9],
       [2.5],
       [1.9],
       [6.1],
       [7.4],
       [2.7],
       [4.8],
       [3.8],
       [6.9],
       [7.8]])
```

In [17]:
```
y
```

Out[17]:
```
array([21, 47, 27, 75, 30, 20, 88, 60, 81, 25, 85, 62, 41, 42, 17, 95, 30,
       24, 67, 69, 30, 54, 35, 76, 86], dtype=int64)
```

In [18]:
```
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)
print('splitting is done')
```

```
splitting is done
```

In [19]:
```
from sklearn.linear_model import LinearRegression
regn = LinearRegression()
regn.fit(x_train,y_train)
print('tranning is done')
```

```
tranning is done
```

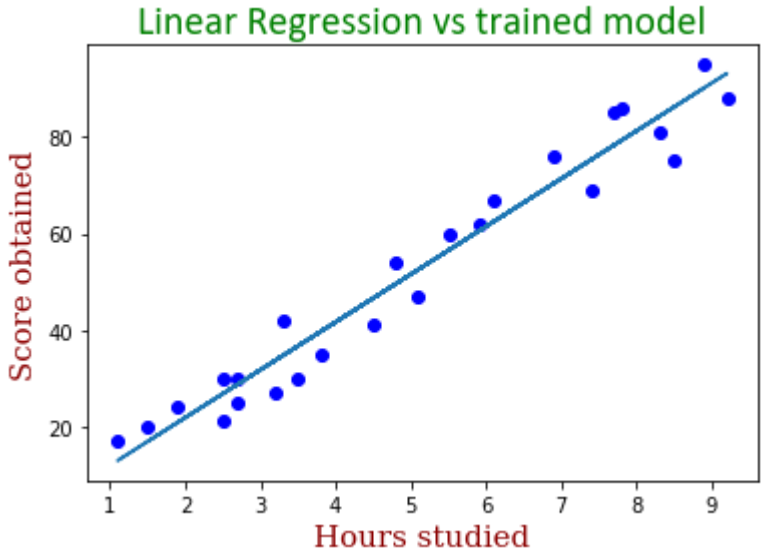
In [20]:
```
print('Intercept value is :',regn.intercept_)
print('Linear coefficient is:',regn.coef_)
```

```
Intercept value is : 2.018160041434683
Linear coefficient is: [9.91065648]
```

In [24]:
```
# plotting the regression line
line = regn.coef_*x+regn.intercept_

#plotting for the the test data
plt.scatter(x,y,c='blue')
plt.title('Linear Regression vs trained model',fontdict=font1)
plt.xlabel('Hours studied',fontdict=font2)
```

```
plt.ylabel('Score obtained',fontdict=font2)
plt.plot(x, line);
plt.show()
```



In [25]:
```
#to predict scores of testing data
y_pred = regn.predict(x_test)
```

In [26]:
```
y_pred
```

Out[26]:
```
array([16.88414476, 33.73226078, 75.357018  , 26.79480124, 60.49103328])
```

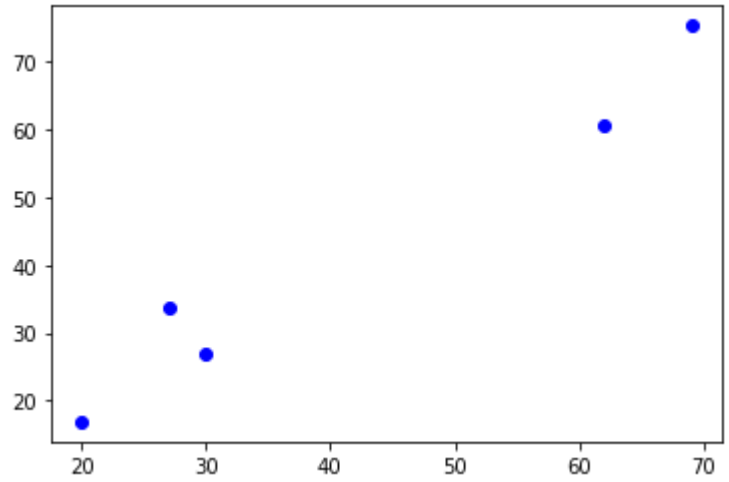
In [28]:
```
df=pd.DataFrame({'Actual':y_test,'predicted':y_pred})
```

In [29]:
```
df
```

Out[29]:

	Actual	predicted
0	20	16.884145
1	27	33.732261
2	69	75.357018
3	30	26.794801
4	62	60.491033

In [37]:
```
plt.scatter(y_test,y_pred,c='blue')
plt.show()
```



In [ ]:
```
what will be predicted score if a student studies for 9.25 hrs/day?
```

In [40]:
```
hours=9.25
pred_score=regn.predict([[hours]])
print("Number of hours={}".format(hours))
print("Predicted score ={}".format(pred_score[0]))
```

```
Number of hours=9.25
Predicted score =93.69173248737538
```

evaluating the model

In [43]:
```
from sklearn import metrics
print('Mean Absolute Error:',
      metrics.mean_absolute_error(y_test,y_pred))
```

```
Mean Absolute Error: 4.183859899002975
```

CONCLUSION for a student stuyding 9.25Hrsa day, the modl predicts his score as 93.6917