P Manohar Rao

Roll No:197158

Data Science Lab Assignment 3.1.1

```
import pandas as pd
path="/content/drive/MyDrive/Salary_Data.csv"
df=pd.read_csv(path)
```

```
import numpy as np
```

```
df.head()
```

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

```
x=df.loc[:,['YearsExperience']]
x.head()
```

	YearsExperience
0	1.1
1	1.3
2	1.5
3	2.0
4	2.2

```
y=df.loc[:,['Salary']]
y.head()
```

	Salary
0	39343.0

1	46205.0
2	37731.0
3	43525.0

```
from pandas.core.common import random_state
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.8,random_state=1)


x_train
```

	YearsExperience
26	9.5

```
x_test
```

	YearsExperience
17	5.3
21	7.1
10	3.9
19	6.0
14	4.5
20	6.8

```
y_train
```

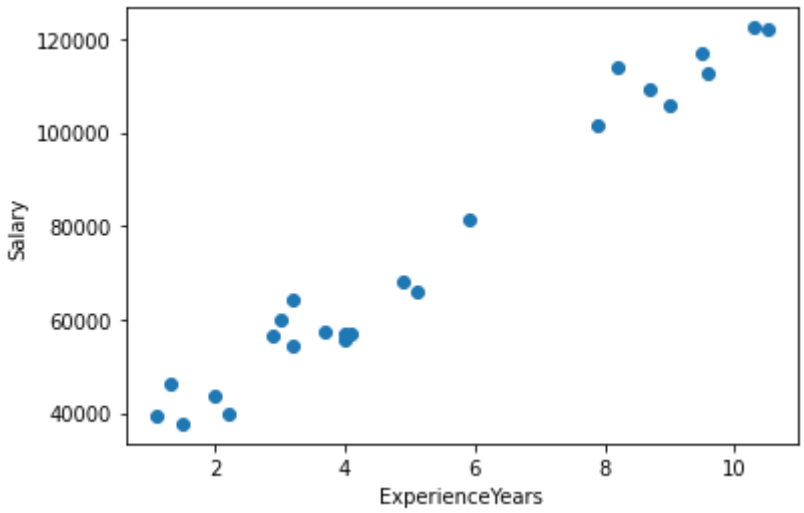
	Salary
26	116969.0

3	43525.0
24	109431.0
22	101302.0

```
y_test
```

	Salary
17	83088.0
21	98273.0
10	63218.0
19	93940.0
14	61111.0
20	91738.0

```
import matplotlib.pyplot as plt
plt.scatter(x_train,y_train)
plt.xlabel('ExperienceYears')
plt.ylabel('Salary')
plt.show()
```




```
from sklearn import linear_model


lm=linear_model.LinearRegression()
model1=lm.fit(x_train,y_train)
```

```
LinearRegression()
```

```
model1.coef_
```

```
array([[9332.94473799]])
```

```
model1.intercept_
```

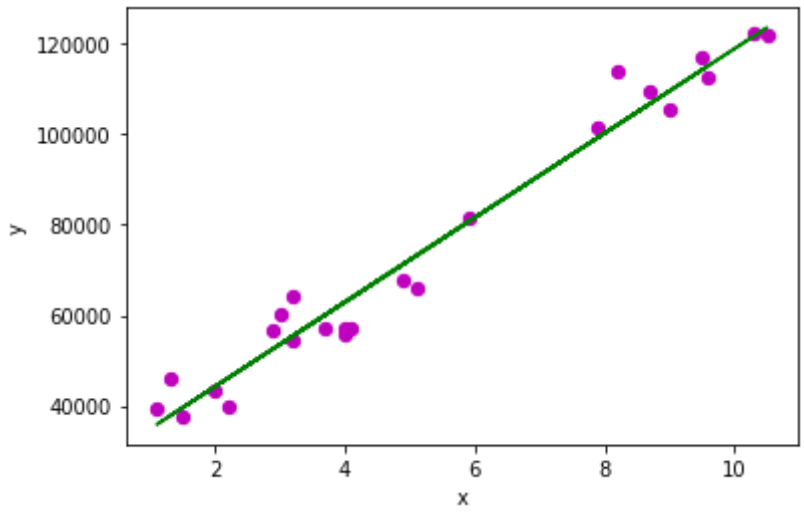
```
array([25609.89799835])
```

```
type(model1)
```

```
sklearn.linear_model._base.LinearRegression
```

```
y_pred=model1.coef_*x_train + model1.intercept_
```

```
plt.plot(x_train,y_pred,color="g")
plt.xlabel('x')
plt.ylabel('y')
plt.scatter(x_train,y_train,color='m',marker='o',s=40)
plt.show()
```




```
test_pred=model1.predict(x_test)
test_pred
```

```
array([[75074.50510972],
       [91873.8056381 ],
       [62008.38247653],
       [81607.56642631],
       [67608.14931932],
       [89073.92221671]])
```

```
cost=np.sqrt(np.mean(np.sum(np.square(y_test-test_pred))))
cost
```

```
17550.73049629247
```

✓ 0s completed at 9:02 PM

Assignment 3.1.2

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
path="/content/drive/MyDrive/housing.csv"
df=pd.read_csv(path,delim_whitespace=True,header=None)
```

```
from sklearn import linear_model
```

```
from sklearn import preprocessing
X=np.array(df.loc[:, df.columns != 13])
Y=np.array(df.loc[:, df.columns == 13])
X
```

```
array([[6.3200e-03, 1.8000e+01, 2.3100e+00, ..., 1.5300e+01, 3.9690e+02,
        4.9800e+00],
       [2.7310e-02, 0.0000e+00, 7.0700e+00, ..., 1.7800e+01, 3.9690e+02,
        9.1400e+00],
       [2.7290e-02, 0.0000e+00, 7.0700e+00, ..., 1.7800e+01, 3.9283e+02,
        4.0300e+00],
       ...,
       [6.0760e-02, 0.0000e+00, 1.1930e+01, ..., 2.1000e+01, 3.9690e+02,
        5.6400e+00],
       [1.0959e-01, 0.0000e+00, 1.1930e+01, ..., 2.1000e+01, 3.9345e+02,
        6.4800e+00],
       [4.7410e-02, 0.0000e+00, 1.1930e+01, ..., 2.1000e+01, 3.9690e+02,
        7.8800e+00]])
```

```
X=preprocessing.scale(X)
Y=preprocessing.scale(Y)
Y
```

```
X=np.hstack((np.ones((Y.size,1 )),X))
print(X.shape)
print(Y.shape)
```

```
(506, 18)
(506, 1)
```

```
def fit(X,Y):
    coeff = []

    coeff = np.linalg.inv(X.transpose().dot(X)).dot(X.transpose()).dot(Y)
    return coeff
```

```
from sklearn.model_selection import train_test_split
train_X,test_X,train_Y,test_Y =train_test_split(X,Y,test_size=1/5)
ml=linear_model.LinearRegression()
ml.fit(train_X,train_Y)
```

```
LinearRegression()
```

```
test_pred=ml.predict(test_X)
```

```
test_Y.shape
```

```
(102, 1)
```

```
import math
import sklearn
a = []
for i in range(0,test_Y.shape[0]):
    a.append(test_Y[i][0])
mse = sklearn.metrics.mean_squared_error(a, test_pred)
rmse = math.sqrt(mse)
rmse
```

```
0.5513969227100752
```

✓ 0s completed at 10:29 PM