35_ml_101

March 23, 2022

1 Machine Learning 101

[ ]:
```
# importing libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
plt.style.use({'figure.facecolor':'white'})
```

[ ]:
```
data = pd.read_csv('../../data/salary_data.csv')
data.head()
```

[ ]:
```
   YearsExperience  Salary
0              1.1   39343
1              1.3   46205
2              1.5   37731
3              2.0   43525
4              2.2   39891
```

[ ]:
```
# X = data.iloc[:, :-1].values
# X
# type(X)
```

[ ]:
```
X = data.loc[:, ["YearsExperience"]].values
X
```

[ ]:
```
array([[ 1.1],
       [ 1.3],
       [ 1.5],
       [ 2. ],
       [ 2.2],
       [ 2.9],
       [ 3. ],
       [ 3.2],
       [ 3.2],
       [ 3.7],
       [ 3.9],
           [ 4. ],
           [ 4. ],
           [ 4.1],
           [ 4.5],
           [ 4.9],
           [ 5.1],
           [ 5.3],
           [ 5.9],
           [ 6. ],
           [ 6.8],
           [ 7.1],
           [ 7.9],
           [ 8.2],
           [ 8.7],
           [ 9. ],
           [ 9.5],
           [ 9.6],
           [10.3],
           [10.5]])
```

```
[ ]: y = data.loc[:, ["Salary"]].values
     y
```

```
[ ]: array([[ 39343],
            [ 46205],
            [ 37731],
            [ 43525],
            [ 39891],
            [ 56642],
            [ 60150],
            [ 54445],
            [ 64445],
            [ 57189],
            [ 63218],
            [ 55794],
            [ 56957],
            [ 57081],
            [ 61111],
            [ 67938],
            [ 66029],
            [ 83088],
            [ 81363],
            [ 93940],
            [ 91738],
            [ 98273],
            [101302],
            [113812],
            [109431],
            [105582],
            [116969],
            [112635],
            [122391],
            [121872]], dtype=int64)
```

```
[ ]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,␣
      ↪random_state=1)
```

1.1 Exploring train and test data

```
[ ]: X_train
```

```
[ ]: array([[ 8.2],
            [ 2.2],
            [ 1.5],
            [ 9. ],
            [ 3. ],
            [ 5.9],
            [ 4.1],
            [ 3.2],
            [ 9.6],
            [ 1.3],
            [ 5.1],
            [ 1.1],
            [ 4.9],
            [10.5],
            [10.3],
            [ 3.7],
            [ 3.2],
            [ 4. ],
            [ 4. ],
            [ 2.9]])
```

```
[ ]: X_test
```

```
[ ]: array([[5.3],
            [7.1],
            [3.9],
            [6. ],
            [4.5],
            [6.8],
            [9.5],
            [2. ],
            [8.7],
            [7.9]])
```

```
[ ]: y_train
```

```
[ ]: array([[113812],
            [ 39891],
            [ 37731],
            [105582],
            [ 60150],
            [ 81363],
            [ 57081],
            [ 54445],
            [112635],
            [ 46205],
            [ 66029],
            [ 39343],
            [ 67938],
            [121872],
            [122391],
            [ 57189],
            [ 64445],
            [ 56957],
            [ 55794],
            [ 56642]], dtype=int64)
```

```
[ ]: y_test
```

```
[ ]: array([[ 83088],
            [ 98273],
            [ 63218],
            [ 93940],
            [ 61111],
            [ 91738],
            [116969],
            [ 43525],
            [109431],
            [101302]], dtype=int64)
```

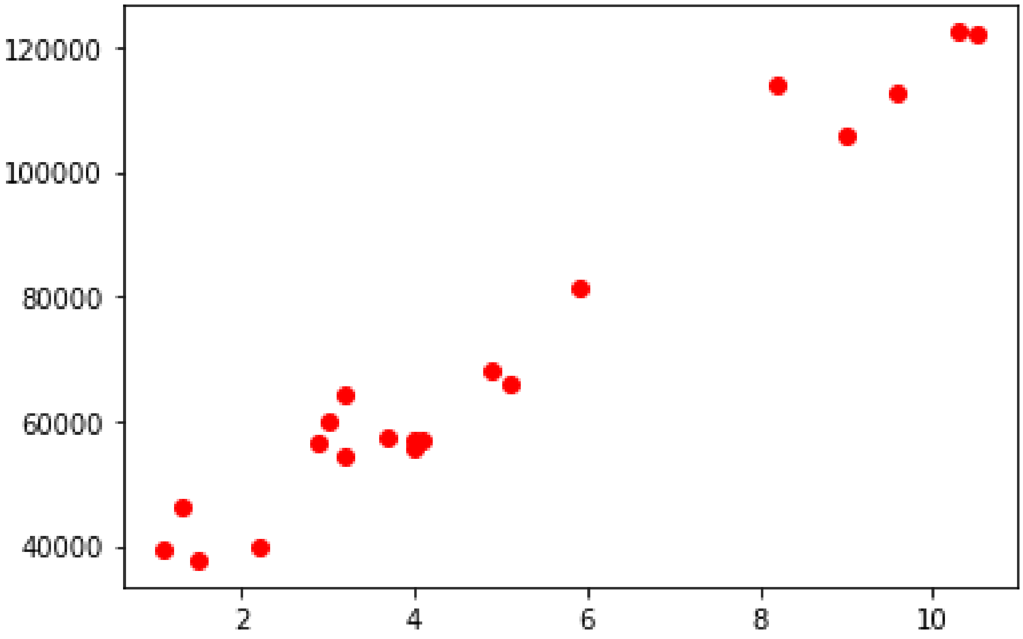
```
[ ]: regressor = LinearRegression()
     regressor.fit(X=X_train, y=y_train)
```

```
[ ]: LinearRegression()
```

```
[ ]: help(regressor.fit)
```

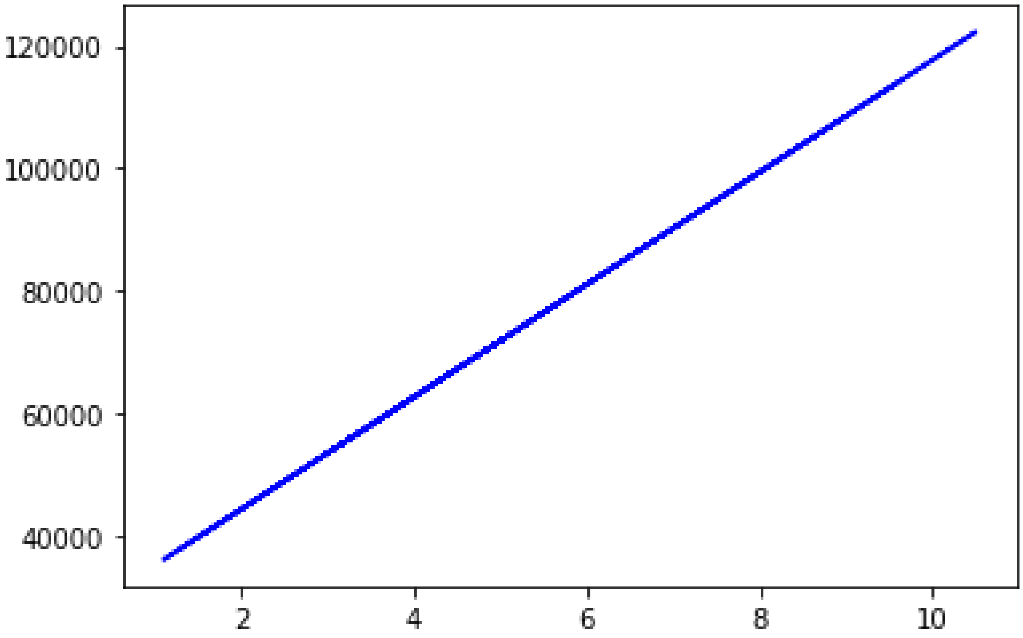
```
[ ]: # viz_train = plt
     plt.scatter(X_train, y_train, color='red')
```

```
[ ]: <matplotlib.collections.PathCollection at 0x2b1f54b9f60>
```



```
[ ]: plt.plot(X_train, regressor.predict(X_train), color="blue")
```

```
[ ]: [<matplotlib.lines.Line2D at 0x2b1f55d6230>]
```

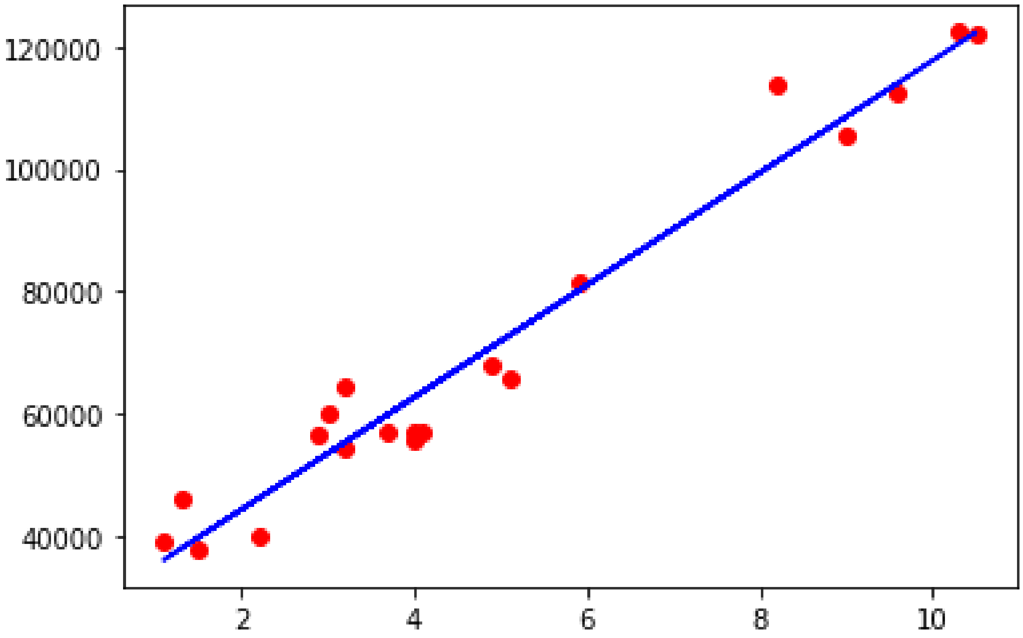


[ ]:

```
plt.scatter(X_train, y_train, color='red')
plt.plot(X_train, regressor.predict(X_train), color="blue")
```

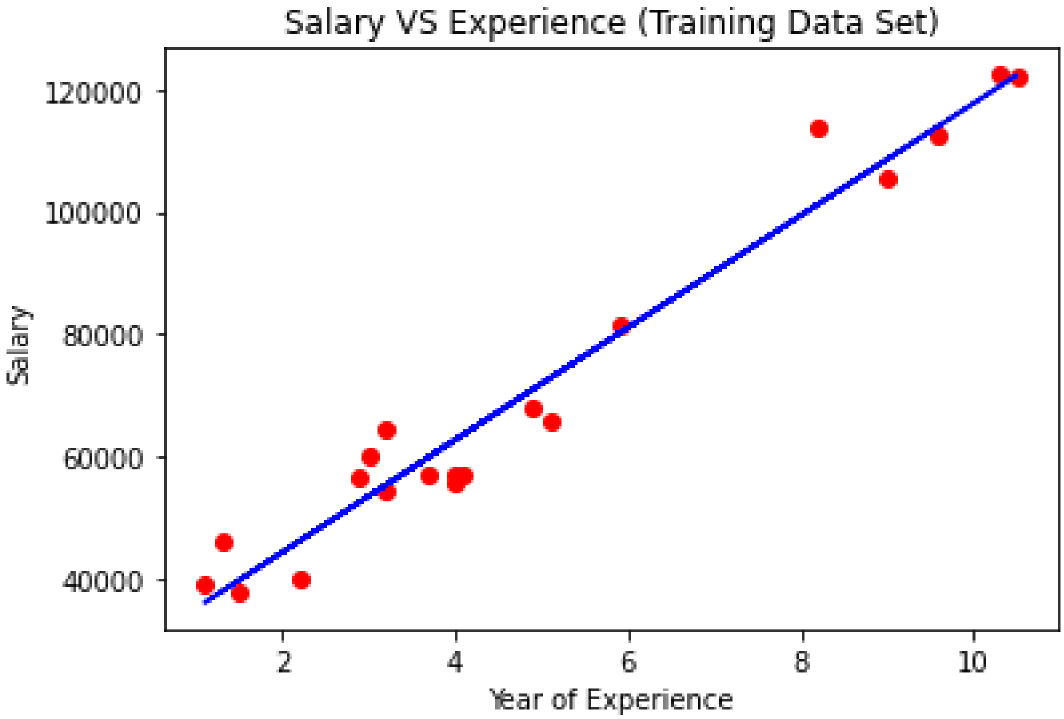
[ ]:

```
[<matplotlib.lines.Line2D at 0x2b1f56303a0>]
```



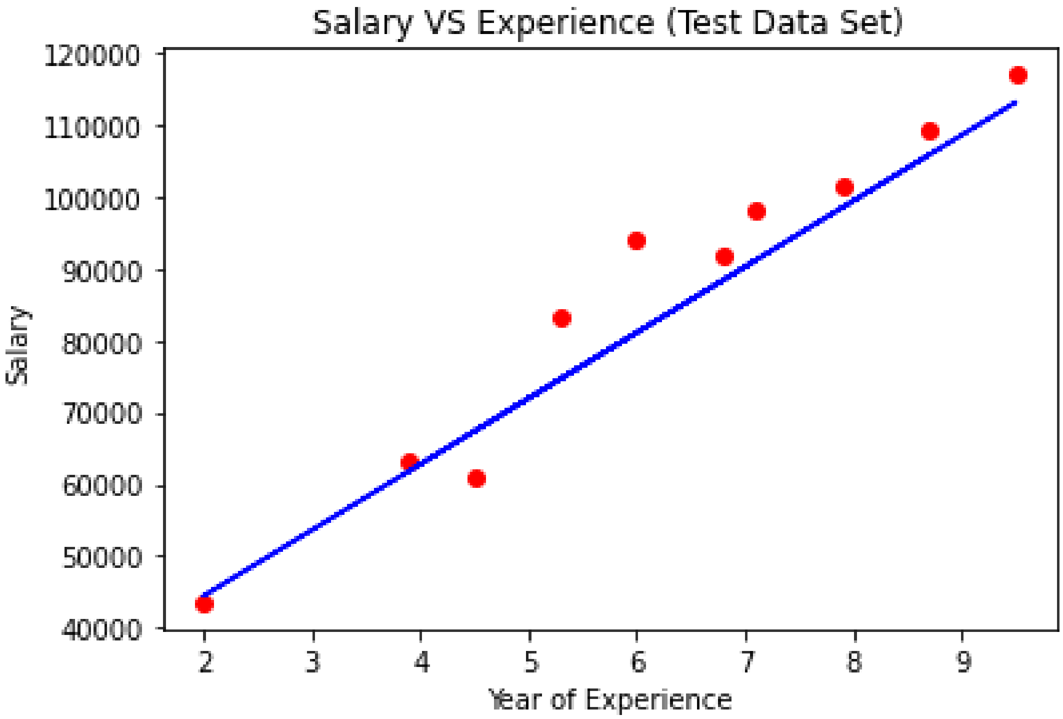
[ ]:

```
plt.scatter(X_train, y_train, color='red')
plt.plot(X_train, regressor.predict(X_train), color="blue")
plt.title('Salary VS Experience (Training Data Set)')
plt.xlabel('Year of Experience')
plt.ylabel('Salary')
plt.show()
```



[ ]:

```python
plt.scatter(X_test, y_test, color='red')
plt.plot(X_test, regressor.predict(X_test), color="blue")
plt.title('Salary VS Experience (Test Data Set)')
plt.xlabel('Year of Experience')
plt.ylabel('Salary')
plt.show()
```



```
[ ]: # predict 5 Years of experience's salary
     y_pred_arr = regressor.predict(X_test)
     y_pred_arr
```

```
[ ]: array([[ 74675.37776747],
            [ 91160.02832519],
            [ 61853.98288925],
            [ 81086.07520659],
            [ 67348.86640849],
            [ 88412.58656557],
            [113139.56240215],
            [ 44453.51841166],
            [105813.05104316],
            [ 98486.53968418]])
```

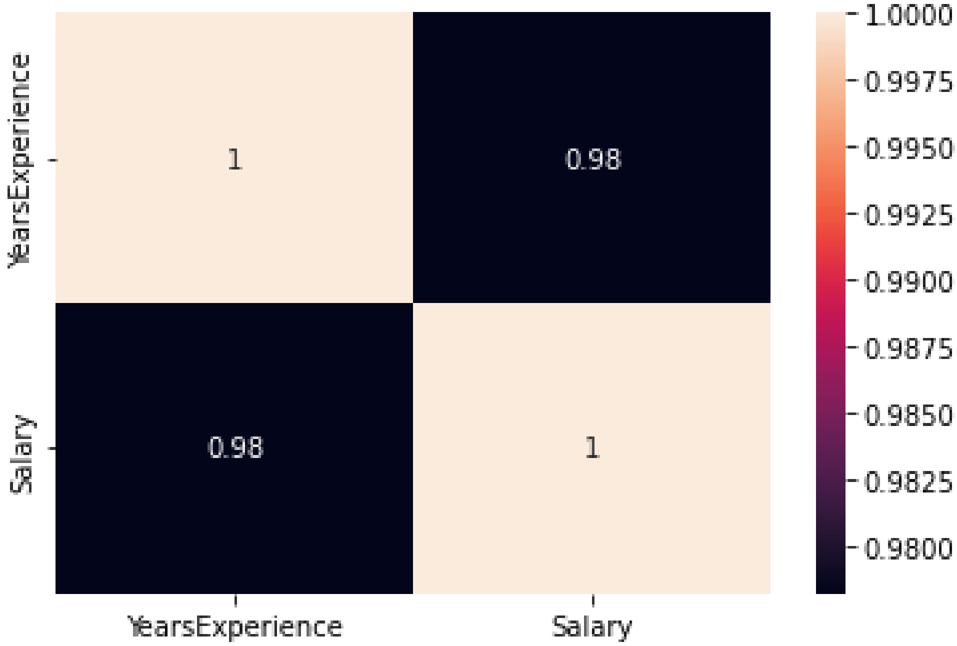
```
[ ]: y_pred = regressor.predict([[5]])
     y_pred
```

```
[ ]: array([[71927.93600785]])
```

```
[ ]: data.corr()
[ ]:                  YearsExperience    Salary
     YearsExperience         1.000000  0.978242
     Salary                  0.978242  1.000000
```

```
[ ]: sns.heatmap(data.corr(), annot=True)
```

```
[ ]: <AxesSubplot:>
```


YearsExperience
Salary
YearsExperience
Salary