project-1

February 21, 2024

[2]:
```
## LINEAR REGRESSION based projects

## Salary Analysis based on Linear Regression

import pandas as pd
df = pd.read_csv("Linear_regr_Salary_dataset.csv")
df.head()
```

[2]:

	Unnamed: 0	YearsExperience	Salary
0	0	1.2	39344.0
1	1	1.4	46206.0
2	2	1.6	37732.0
3	3	2.1	43526.0
4	4	2.3	39892.0

[3]:
```
df.shape
```

[3]:
```
(30, 3)
```

[4]:
```
df.isnull().sum()
```

[4]:
```
Unnamed: 0         0
YearsExperience    0
Salary             0
dtype: int64
```

[5]:
```
x = df[['YearsExperience']]
x
y = df[['Salary']]
y
```

[5]:

	Salary
0	39344.0
1	46206.0
2	37732.0
3	43526.0
4	39892.0
5	56643.0

```
6    60151.0
7    54446.0
8    64446.0
9    57190.0
10   63219.0
11   55795.0
12   56958.0
13   57082.0
14   61112.0
15   67939.0
16   66030.0
17   83089.0
18   81364.0
19   93941.0
20   91739.0
21   98274.0
22  101303.0
23  113813.0
24  109432.0
25  105583.0
26  116970.0
27  112636.0
28  122392.0
29  121873.0
```

[6]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3,␣
 ↪random_state=101)
```

[7]:
```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model
```

[7]:
```
LinearRegression()
```

[8]:
```
model.fit(x_train, y_train)
```

[8]:
```
LinearRegression()
```

[9]:
```
y_pred = model.predict(x_test)
y_pred
```

[9]:
```
array([[ 91101.58255782],
       [109298.20888234],
       [ 56623.76425873],
       [ 82482.12798305],
       [ 40342.57228416],
```

```
       [117917.66345711],
       [116959.94628213],
       [ 74820.39058325],
       [112171.36040726]])
```

[10]:
```
import numpy as np
y_test
```

[10]:
```
      Salary
20   91739.0
24  109432.0
7    54446.0
18   81364.0
2    37732.0
27  112636.0
26  116970.0
16   66030.0
25  105583.0
```

[11]:
```
inputdata = [[14.5]]
prediction = model.predict(inputdata)
prediction
```

```
C:\ProgramData\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X
does not have valid feature names, but LinearRegression was fitted with feature
names
  warnings.warn(
```

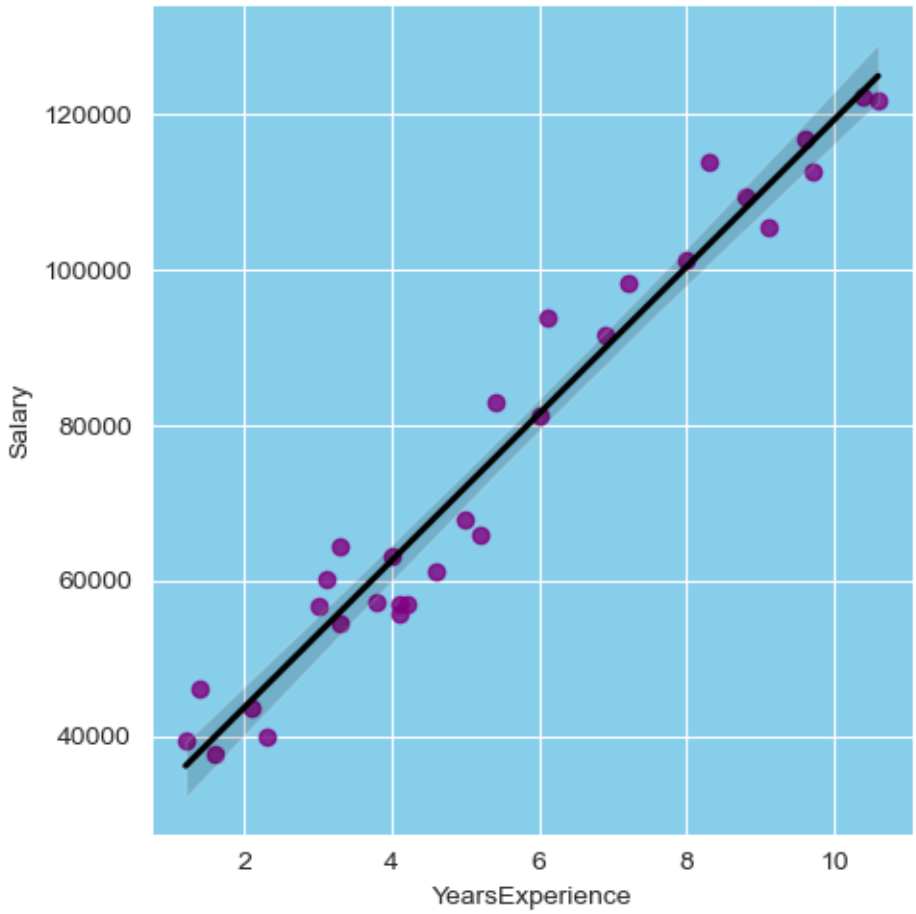
[11]:
```
array([[163888.08785589]])
```

[12]:
```
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, y_pred)
mse
```

[12]:
```
17978409.497344103
```

[49]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.lmplot(x="YearsExperience", y="Salary", data=df, scatter_kws={'color':
 ↪'purple'}, line_kws={'color':'black'})
sns.set_style('darkgrid')
ax = plt.gca()
plt.gca().set_facecolor('skyblue')
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning:
The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

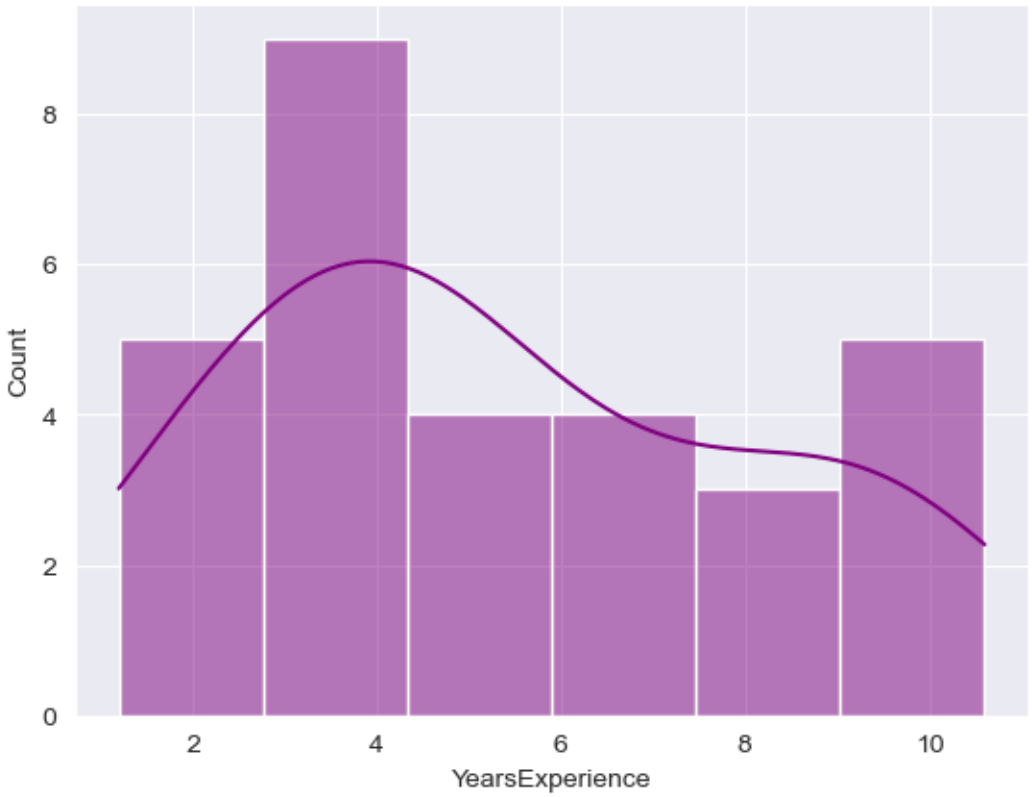


[50]:
```
# Based on index value try to check the performance
response = df["YearsExperience"]
response.dtype
```

[50]:
```
dtype('float64')
```

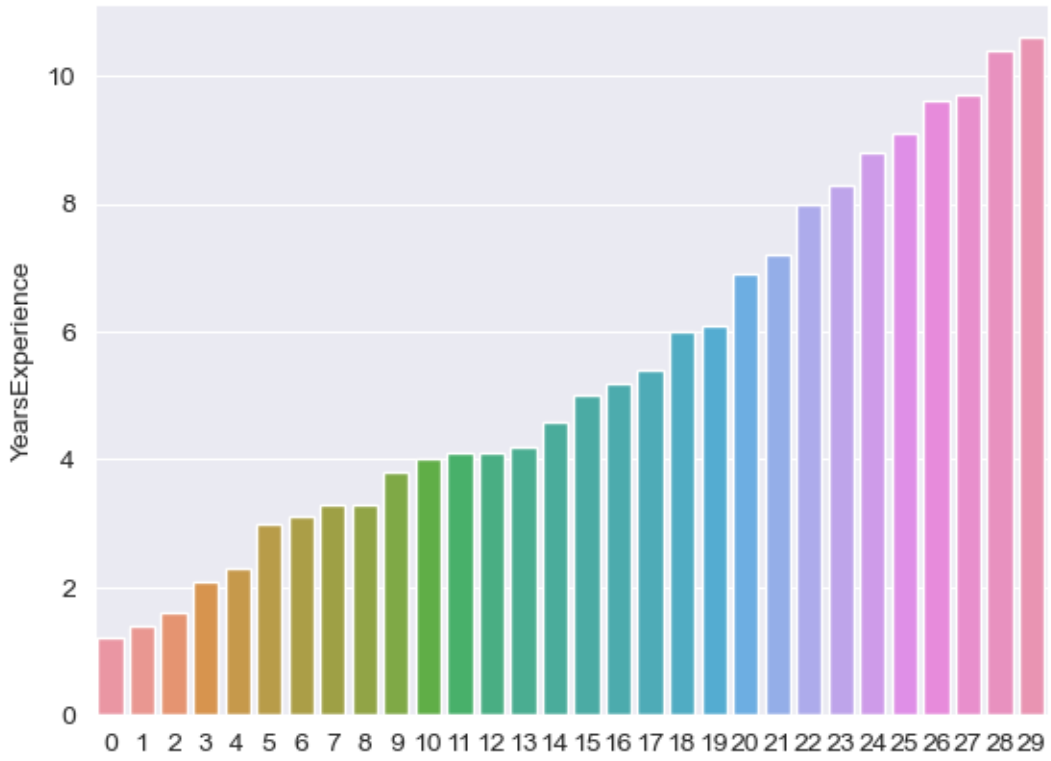
[51]:
```
sns.histplot(df["YearsExperience"], kde=True, color="purple")
```

[51]:
```
<Axes: xlabel='YearsExperience', ylabel='Count'>
```



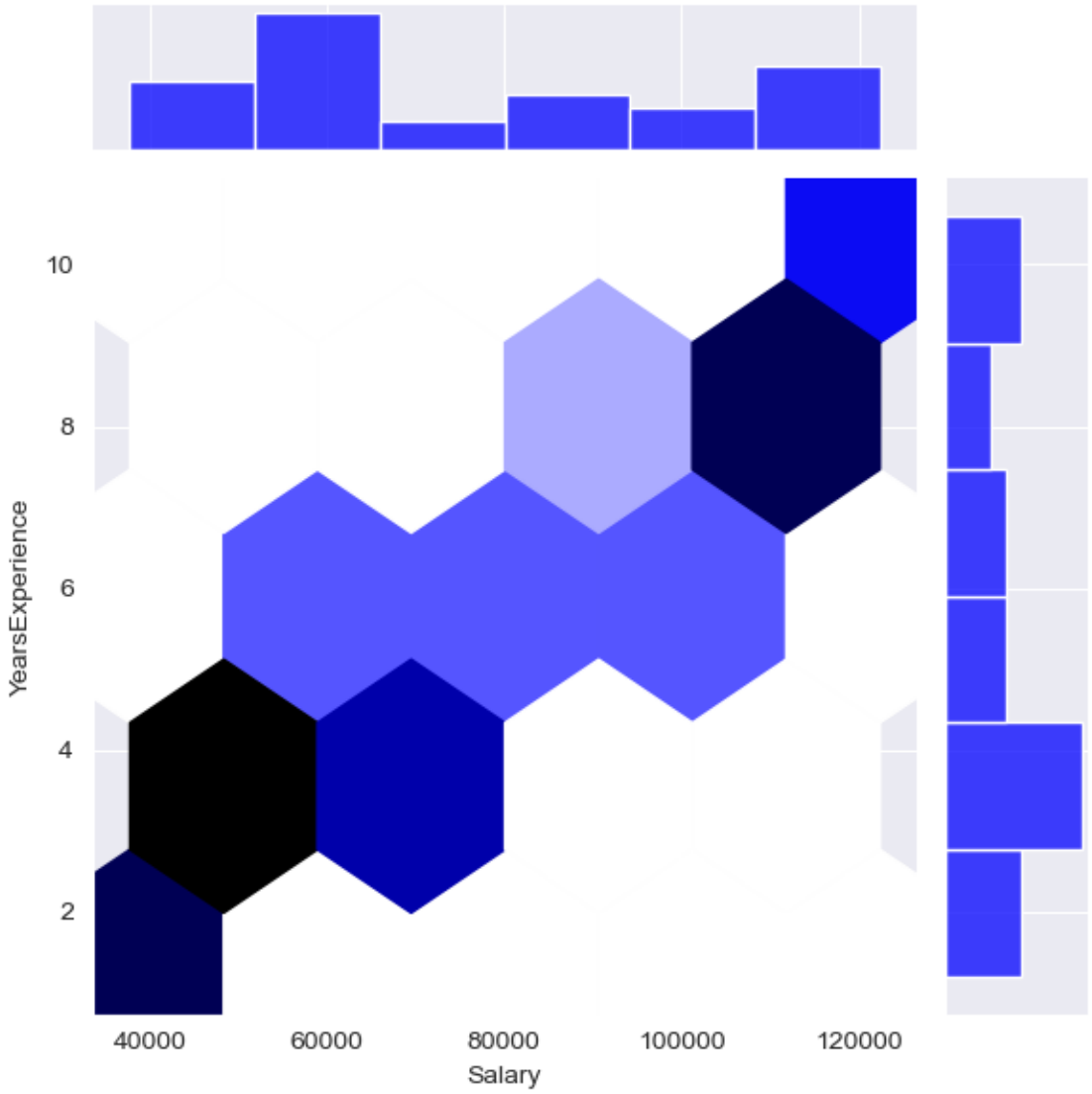
[57]:
```
sns.barplot(y="YearsExperience",x=response.index,data=df)
```

[57]: <Axes: ylabel='YearsExperience'>



[30]:
```
sns.jointplot(x="Salary", y="YearsExperience", data=df, kind="hex",␣
↪color="blue")
```

[30]:
```
<seaborn.axisgrid.JointGrid at 0x1f145605ed0>
```


10
8
6
4
2
YearsExperience
40000
60000
80000
100000
120000
Salary