SIMPLE LINEAR REGRESSION USING STATSMODELS AND SKLEARN

The purpose of this analysis is to predict Salary based on one's years of experience

The analysis aimed at comparing R_squared values from statsmodels and sklearn using linear regression

In [1]:
```
#Importing relevant libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
sns.set()
```

In [2]:
```
#loading the data
data=pd.read_csv("Desktop/Salary_dataset.csv")
```

In [3]:
```
#first five rows of data
data.head()
```

Out[3]:

	YearsExperience	Salary
0	1.2	39344
1	1.4	46206
2	1.6	37732
3	2.1	43526
4	2.3	39892

In [4]:
```
#information on data
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   YearsExperience  30 non-null     float64
 1   Salary           30 non-null     int64
dtypes: float64(1), int64(1)
memory usage: 612.0 bytes
```

In [5]:
```
#descriptive statistics
data.describe()
```

Out[5]:

	YearsExperience	Salary
count	30.000000	30.000000
mean	5.413333	76004.000000
std	2.837888	27414.429785
min	1.200000	37732.000000
25%	3.300000	56721.750000
50%	4.800000	65238.000000
75%	7.800000	100545.750000
max	10.600000	122392.000000

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

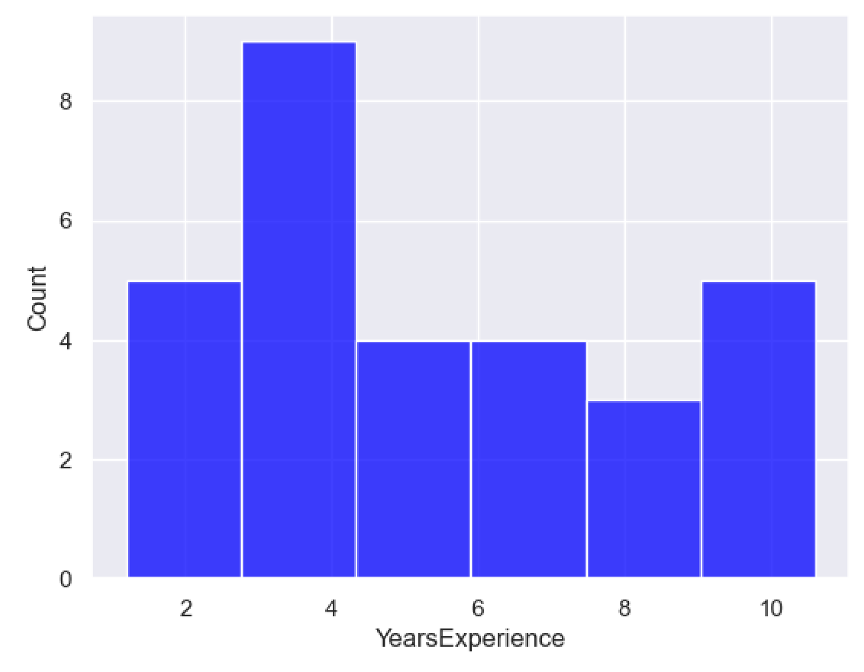
Out[6]:
```
(30, 2)
```

In [7]:
```
#checking null values
data.isnull().sum()
```

Out[7]:
```
YearsExperience    0
Salary             0
dtype: int64
```

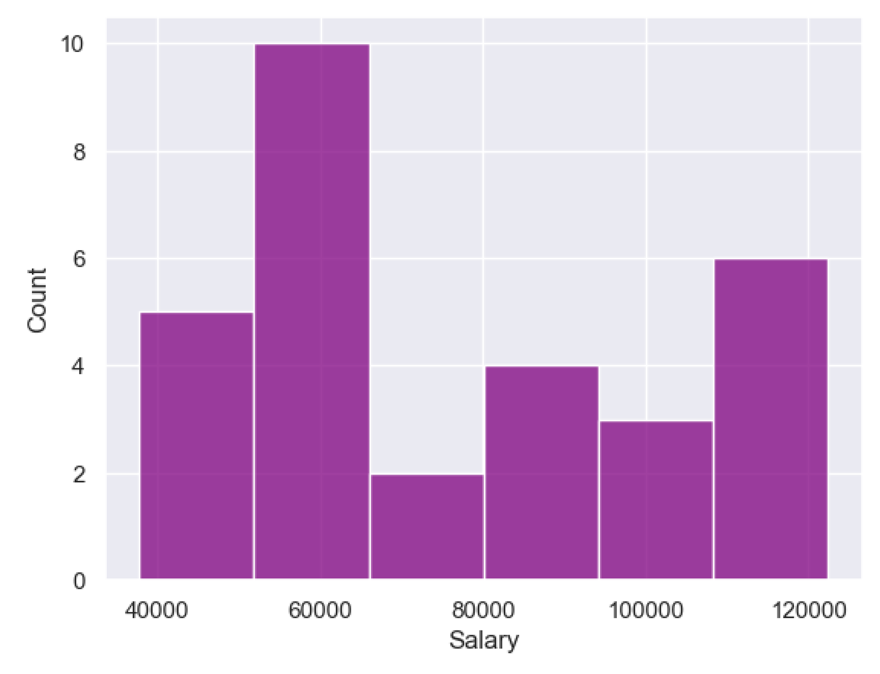
In [8]:
```
#histogram on years of experience
sns.histplot(x='YearsExperience',data=data,color='blue')
```

Out[8]:
```
<Axes: xlabel='YearsExperience', ylabel='Count'>
```



In [9]:
```
#histogram on Salary
sns.histplot(x='Salary',data=data,color='purple')
```

Out[9]:
```
<Axes: xlabel='Salary', ylabel='Count'>
```



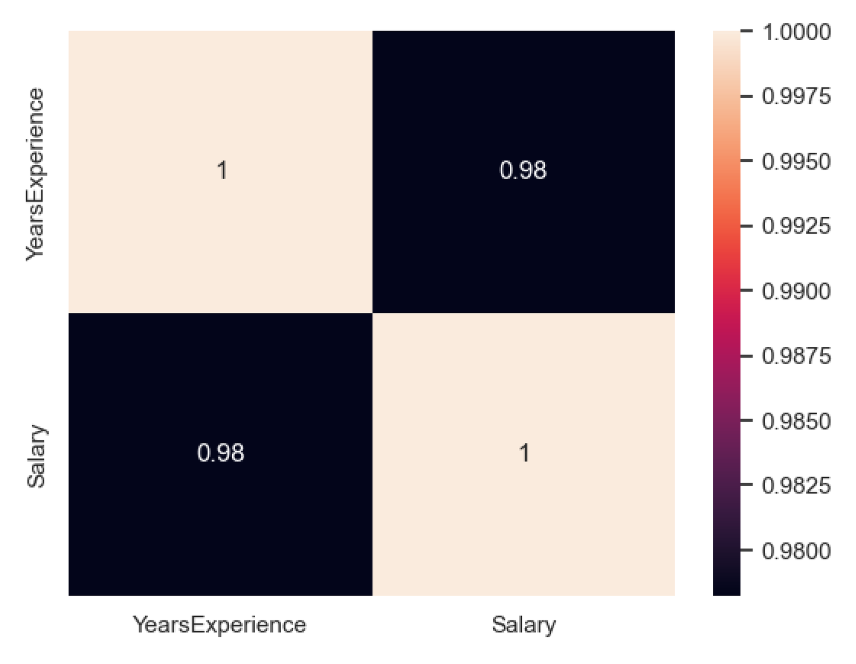
In [10]:
```
#correlation
data.corr()
```

Out[10]:

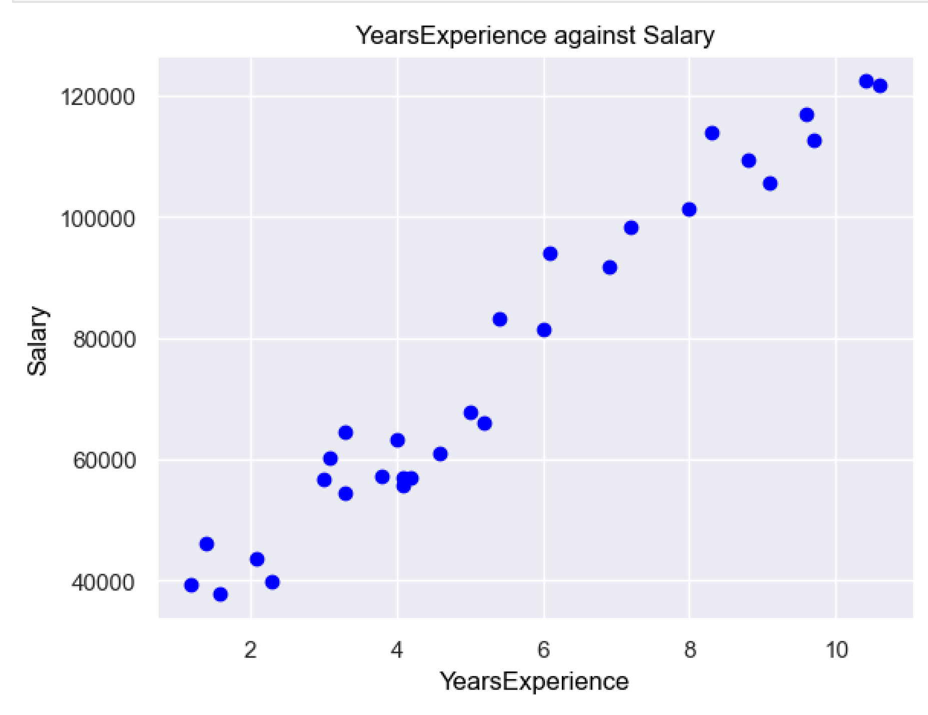
	YearsExperience	Salary
YearsExperience	1.000000	0.978242
Salary	0.978242	1.000000

In [11]:
```
#heatmap on correlation
sns.heatmap(data.corr(),annot=True)
```

Out[11]:
```
<Axes: >
```



In [12]:
```
#scatter plot on YearsExperience against Salary
plt.scatter(data['YearsExperience'],data['Salary'],c='blue')
plt.title('YearsExperience against Salary',c='black')
plt.xlabel('YearsExperience',c='black')
plt.ylabel('Salary',c='black')
plt.show()
```



USING STATS MODELS

In [13]:
```
x=x=data['YearsExperience']
y=data['Salary']
```

In [14]:
```
# adding y_intercept to the regression equation
X=sm.add_constant(x)
results=sm.OLS(y,X).fit() #Using ordinary least square regression on x and y
results.summary()   #obtaining summary statistics
```

Out[14]:

OLS Regression Results

Dep. Variable:	Salary	**R-squared:**	0.957
Model:	OLS	**Adj. R-squared:**	0.955
Method:	Least Squares	**F-statistic:**	622.5
Date:	Thu, 14 Dec 2023	**Prob (F-statistic):**	1.14e-20
Time:	15:24:00	**Log-Likelihood:**	-301.44
No. Observations:	30	**AIC:**	606.9
Df Residuals:	28	**BIC:**	609.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.485e+04	2306.654	10.772	0.000	2.01e+04	2.96e+04
YearsExperience	9449.9623	378.755	24.950	0.000	8674.119	1.02e+04

Omnibus:	2.140	**Durbin-Watson:**	1.648
Prob(Omnibus):	0.343	**Jarque-Bera (JB):**	1.569
Skew:	0.363	**Prob(JB):**	0.456
Kurtosis:	2.147	**Cond. No.**	13.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [15]:
```
# Fitting a regression line
plt.scatter(x,y,c='GREEN')
yhat = 9449.96*x +24850
fig=plt.plot(x,yhat,lw=4,c='BLUE',label='regression line')
plt.title('Salary against Years of Experience')
plt.xlabel('YearsExperience')
plt.ylabel('Salary')
plt.show()
```

RESULTS INTEPRETATION

The regression equation is yhat=b0+b1x

Salary=24850+9449.96*YearsExperience

This shows that a 1 increase in Years of Experience leads to increase in Salary by 9450

The adjusted R2 value is 0.955.This is equal to 95.5%.It means that 95.5% of the variability in Salary is caused by by Years of Experience.

The remainining 4.5% is caused by other factors not captured.

Test of hypothesis,Ho:Beta=0 i.e coefficient=0

The p value of Years of experience is 0.000 which is less than 0.05. This means that independent variable is significant for predicting .i.e beta is not equal to zero

USING SKLEARN

In [16]:
```
from sklearn.linear_model import LinearRegression
```

In [17]:
```
#changing the order of matrix x
x_matrix=x.values.reshape(-1, 1)
x_matrix.shape
```

Out[17]:
```
(30, 1)
```

In [18]:
```
#fitting regression on x and y
reg=LinearRegression()
reg.fit(x_matrix,y)
```

Out[18]:
```
▾ LinearRegression
LinearRegression()
```

In [19]:
```
#R_squared value
r_squared=reg.score(x_matrix,y)
r_squared
```

Out[19]:
```
0.9569566641435086
```

In [20]:
```
k=x_matrix.shape[0]
n=x_matrix.shape[1]
```

In [21]:
```
r_squared_adjusted=1-(1-r_squared)*(k-1)/(k-n-1)
r_squared_adjusted
```

Out[21]:
```
0.9554194021486339
```

In [22]:
```
#coefficient
reg.coef_
```

Out[22]:
```
array([9449.96232146])
```

In [23]:
```
#Intercept
reg.intercept_
```

Out[23]:
```
24848.203966523208
```

In [24]:
```
# making predictions
reg.predict([[35]])
```

Out[24]:
```
array([355596.88521745])
```

In [25]:
```
predicted=reg.predict(x_matrix)
```

In [26]:
```
# f_regression
```

In [27]:
```
from sklearn.feature_selection import f_regression
```

In [28]:
```
F_statistics=f_regression(x_matrix,y)[0]
F_statistics
```

Out[28]:
```
array([622.50720263])
```

In [29]:
```
P_values=f_regression(x_matrix,y)[1]
P_values
```

Out[29]:
```
array([1.14306811e-20])
```

In [30]:
```
#summary _table
```

In [31]:
```
summary_table=pd.DataFrame(data=['YearsExperience'],columns=['feature'])
```

In [32]:
```
summary_table['coefficient']=reg.coef_
summary_table['intercept']=reg.intercept_
summary_table['r_squared']=reg.score(x_matrix,y)
summary_table['adjusted_r_squared']=1-(1-r_squared)*(k-1)/(k-n-1)
summary_table['F_statistics']=f_regression(x_matrix,y)[0]
summary_table['P_value']=f_regression(x_matrix,y)[1]
```

In [33]:
```
summary_table
```

Out[33]:

	feature	coefficient	intercept	r_squared	adjusted_r_squared	F_statistics	P_value
0	YearsExperience	9449.962321	24848.203967	0.956957	0.955419	622.507203	1.143068e-20

CONCLUSION

Both the two models had r_squared value of 95% hence it shows that 95.5% of the variability in Salary is caused by by Years of Experience for this given dataset.