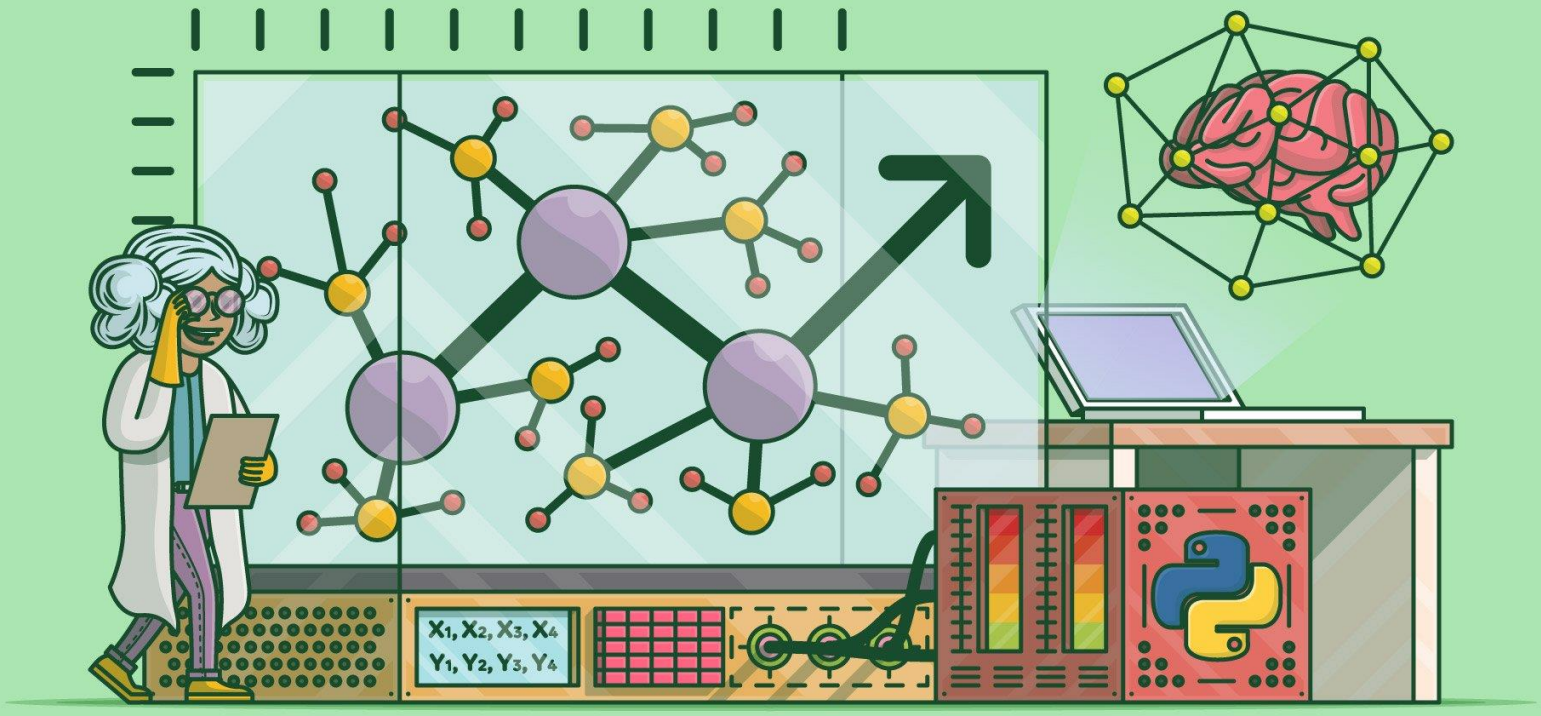
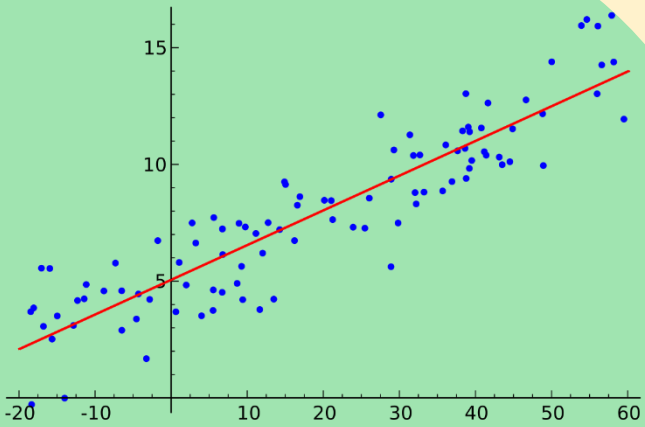
python Pandas



Mini Project Simple Linear Regression



-by Shivam Kumar Giri

Mini Project Simple Linear Regression

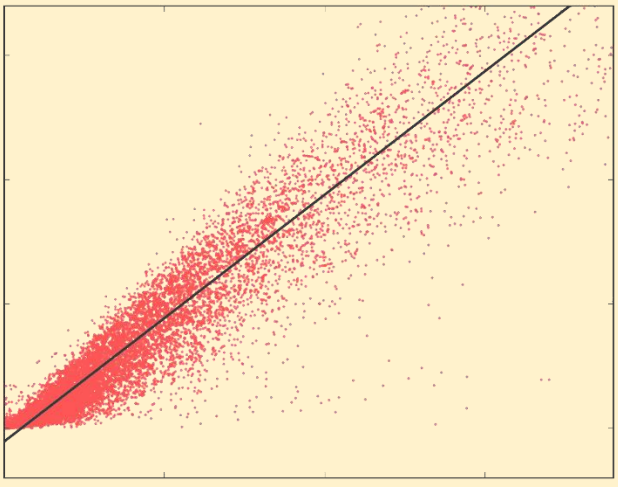
PROBLEM STATEMENTS

The dataset consists of salary information with years of experience. Write a report covering following points: -

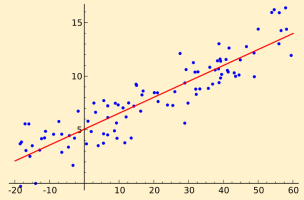
1. Find summary statistics of both variables and find out is there any anomaly or not.
2. Draw a scatter plot and explain if there is a linear relationship between salary and years of experience.
3. Randomly split the into train dataset (70%) and test dataset (30%).
4. If you find if there is a linear relationship, then fit a simple linear model and explain: -
    a. Goodness of fit
    b. Estimated values of parameters
    c. p-value of each parameter.
    d. Validate the result with test dataset

ABSTRACT

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. Here we are analyzing Linear Regression for predicting the salary of the employee as per there year of Experience.



A sample graph with the regression to predict future value of the dependent variable



METHODOLOGY

1. Find summary statistics of both variables and find out is there any anomaly or not.

Step 1: Import the libraries

Libraries such as pandas, matplotlib, numpy, tabulate, sklearn and satsmodels.api. Matplotlib and tabulate for data visualization, padas and numpy for data exploration, sklearn and statsmodels.api for importing linear regression and various parameters regarding Linear Regression.

Step 2: Import and the data

Here we have salary_data.csv file to get the data regarding the salary and the years of experience. The corresponding head is printed below:

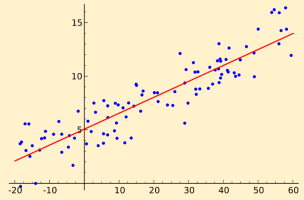
	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Step 3: Explore the data

```
>> df.shape
```

(30, 2)

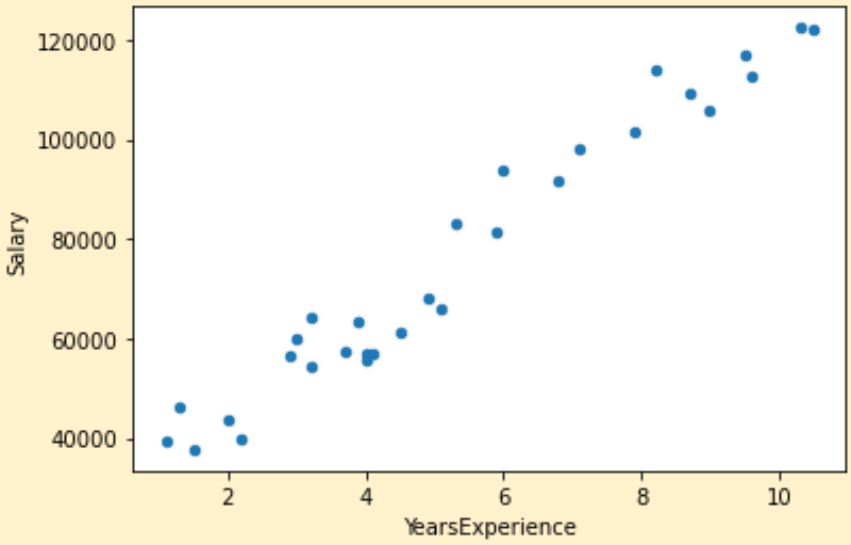
```
>> df.describe()
```



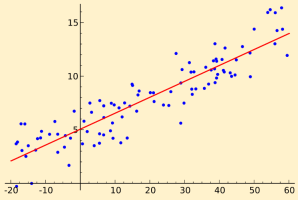
	YearsExperience	Salary
count	30.000000	30.000000
mean	5.313333	76003.000000
std	2.837888	27414.429785
min	1.100000	37731.000000
25%	3.200000	56720.750000
50%	4.700000	65237.000000
75%	7.700000	100544.750000
max	10.500000	122391.000000

2. Draw a scatter plot and explain if there is a linear relationship between salary and years of experience.

Step 4: Plot the salary v/s Year of Experience



We can see that there **exists a direct relationship between** Salary (dependent variable) and YearsExperience (explanatory variable)



3. Randomly split the into train dataset (70%) and test dataset (30%).

Step 5: Split the data for testing and training

Here Data is splited into set of training data and test data, the training data is trained by decision tree while test data is validated for accuracy.

Here We used 70% of data for training ad rest 30% for testing, which is recommended ratio of splitting.

4b. Estimated values of parameters

Step 6: Use Linear Regression and train the inputs find regression coefficient

```
reg = LinearRegression()
reg.fit(x_train,y_train);
```

Step 7: Predict the output of validation class

The Predicted value of validation classes are here:

[**40817.78327049**

123188.08258899

65154.46261459

63282.41035735

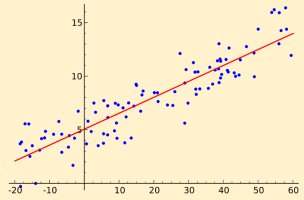
115699.87356004

108211.66453108

116635.89968866

64218.43648597

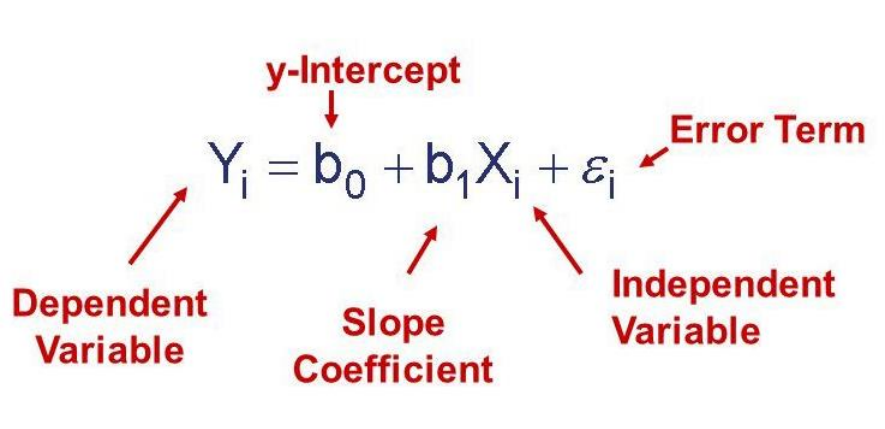
76386.77615802]



Step 8: Find regression coefficient and intercept

Intercept: 26777.391341197625

Coefficient: 9360.26128619

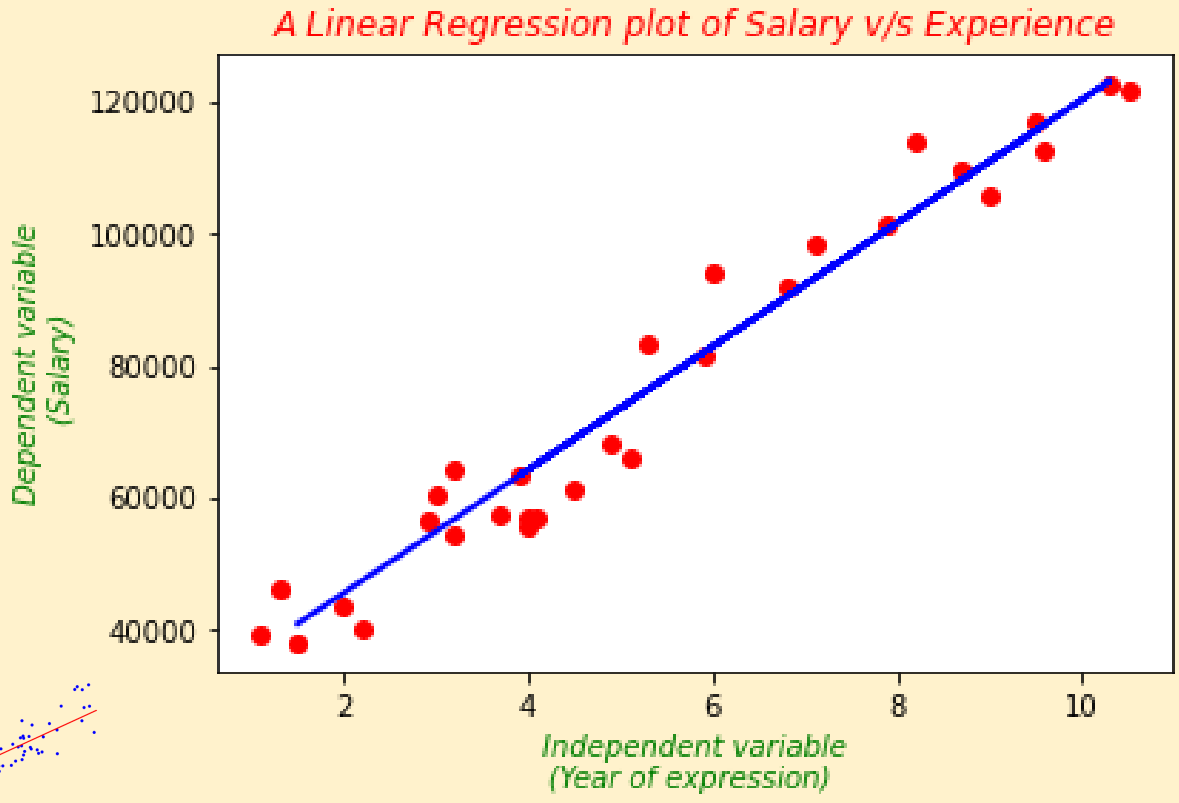


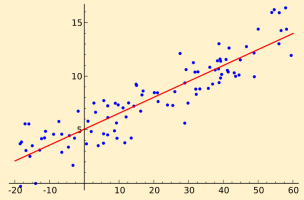
Here we have Y= Intercept + Coefficient(X)

Hence Equation become **Salary= 26777.39 + 9360.26 (YearsExperience)**

4d. Validate the result with test dataset

Step 9: Plot the graph of Linear regression





Step 10: Tabulate Predicted v/s Actual Test Data

Predicted Data v/s Actual Data

Predicted Data	Actual Data	Difference
40817.78	37731.00	-3086.78
123188.08	122391.00	-797.08
65154.46	57081.00	-8073.46
63282.41	63218.00	-64.41
115699.87	116969.00	1269.13
108211.66	109431.00	1219.34
116635.90	112635.00	-4000.90
64218.44	55794.00	-8424.44
76386.78	83088.00	6701.22

Hence, we see all the predicted values are nearby actual values and some of them are very closely related to Actual data.

4c. p-value of each parameter.

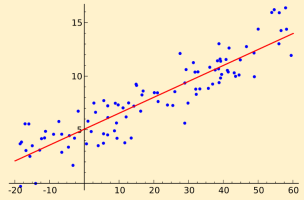
Step 11: Find p-value and R^2 Value

OLS Regression Results

Dep. Variable:	y	R-squared:	0.957
Model:	OLS	Adj. R-squared:	0.955
Method:	Least Squares	F-statistic:	622.5
Date:	Wed, 01 Jul 2020	Prob (F-statistic):	1.14e-20
Time:	17:19:05	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	2.579e+04	2273.053	11.347	0.000	2.11e+04	3.04e+04
x1	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.	13.2



OBSERVATION

We have R^2 value as 0.957 and adjusted R^2 as 0.955, hence the model is well trained and had a greater accuracy while predicting result.

Also, the p-value is 0.000 which is great for the regression model, hence our model is good for predicting salary based on Year of experience

Salary= 26777.39 + 9360.26 (YearsExperience) ± 2273.053

This is the final Equation for deriving new salary for a candidate with some year of Experience along with mean std. error of ± 2273.053.

The End