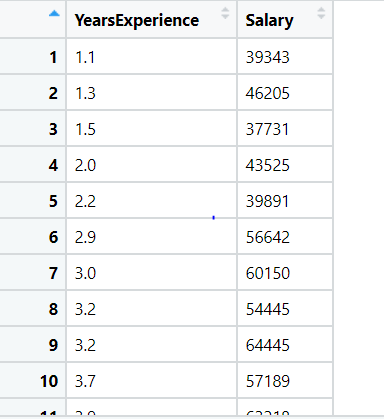
**Problem Statement: -**

The head of HR of a certain organization wants to automate their salary hike estimation. The organization consulted an analytics service provider and asked them to build a basic prediction model by providing them with a dataset that contains the data about the number of years of experience and the salary hike given accordingly. Build a Simple Linear Regression model with salary as the target variable. Apply necessary transformations and record the RMSE and correlation coefficient values for different models.



* **Business Objective:**

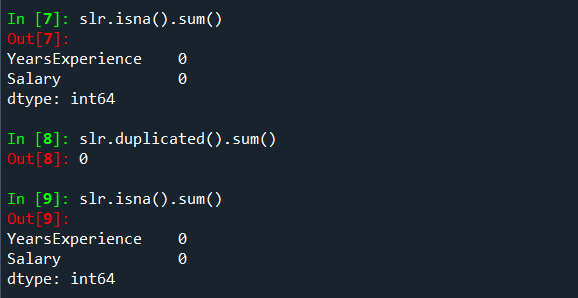
It's important to give employees raise on a regular basis because it shows that you value them and their contributions to the company. Hiking the salary of the employees not only increases their salary but also results in increasing organization’s income, so usually hiking the salary for the employees differs from organization to organization, where one might do it every 2 to 5 years, some might do it based on employee performance regularly or years of experience.

This system will evaluate each employee’s performance depending on attributes mentioned by an organization and then after evaluating, the results are used in calculation of salary hike for each employee.

* **Data Preprocessing:**

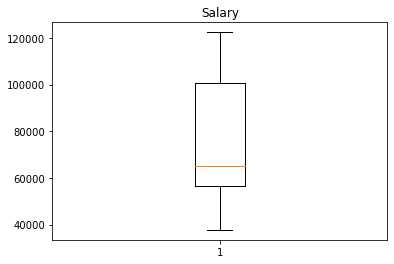
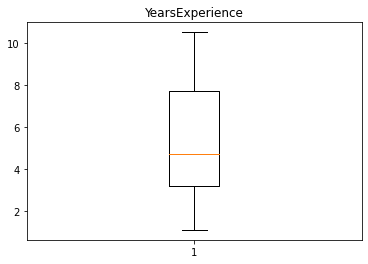
We have a dataset of employees of a company having a year of experience and salary of employee. By using this dataset, we are going to make a salary prediction model.

Before making the prediction model we need to preprocess the data. We have described the data and checked for null values and duplicate values.



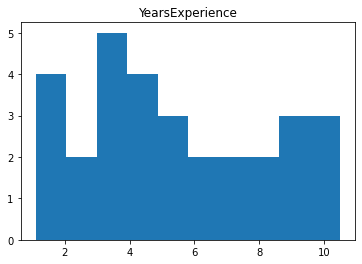
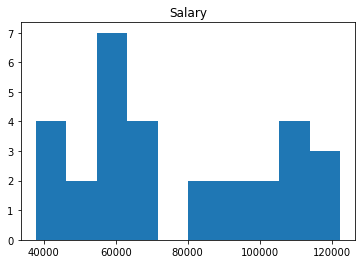
Here we found that there are no null values and no duplicate values in the data set.

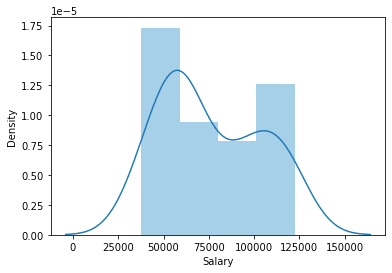
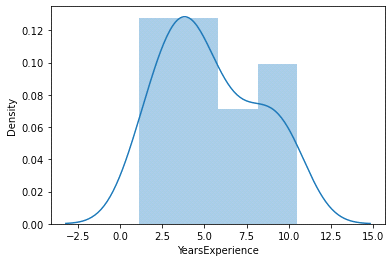
Then we checked for outliers and plotted a boxplot. But here also we found that there are not any outliers in the data.

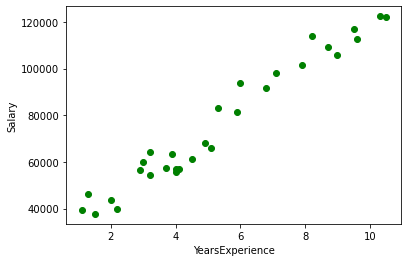
* **Exploratory Data Analysis (EDA):**

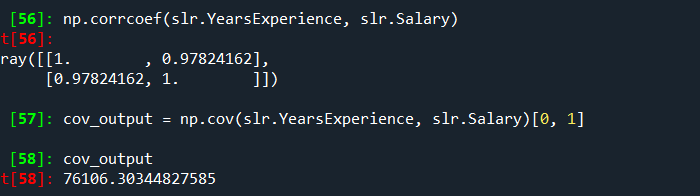
Now we have plotted graphs for understanding the data and gather insights from it.

From this histogram we can analyze that the data is distributed and the maximum years of experience of employee is about 10 years and maximum salary is 122000.

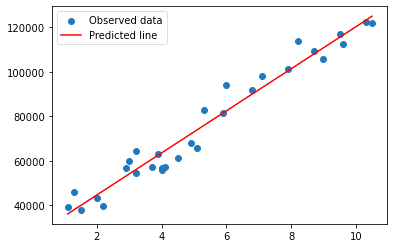




From the above scatter plot and summary, we can analyze that the data i.e. Years of experience and Salary is Strongly Corelated to each other. So, we can say that as the years of experience increases salary of employee also increases. As the number of years of experience affects the salary of data we can make a prediction model from this dataset.

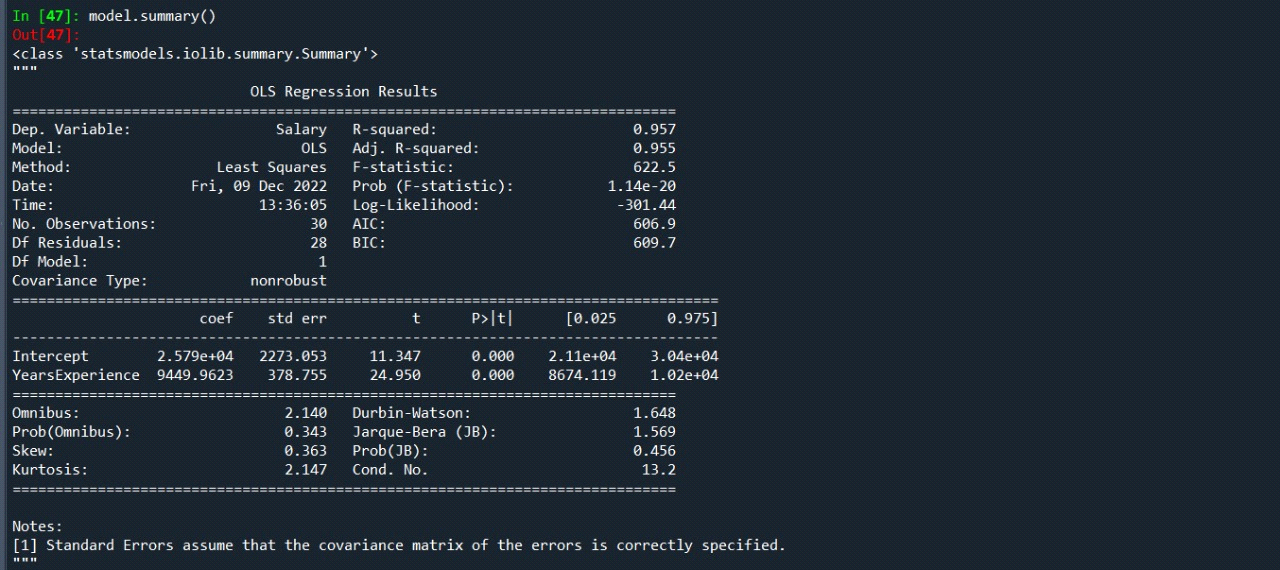
* **Model Building:**

For prediction we have applied the simple linear regression on the dataset and plotted a scatter plot and the plotted the regression line for prediction.



Here the red line indicates the predicted output and blue dots are original data. Now we calculate the error to check whether the model is the best fit for prediction or not.

Summary table 1

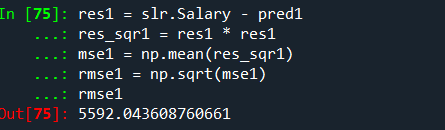


The summary table describes the overall data characteristics of data. With the help of the summary table, we can identify the data of different types

For simple linear regression,

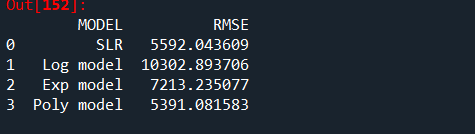
we consider the R-squared value which determines the nature of data Also consider coefficient of correlation (coef), standard error (std\_err) and probability of confidence interval (p>|t|)

By referring the above summary, we conclude the data is not useful as it contains more errors.

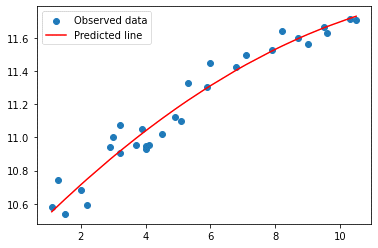


Here we calculated the error i.e., difference between Observed value and predicted value but we got the error as 5592.04 which is high.

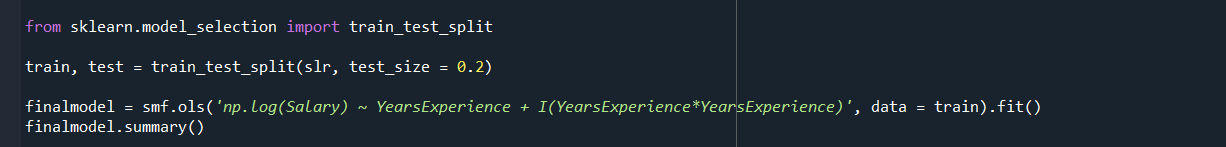
So to minimize the error and to make a best model we do different transformations like log transformation, exponential transformation and polynomial transformation on the data . The transformation which will give less error will be suitable for the prediction model.



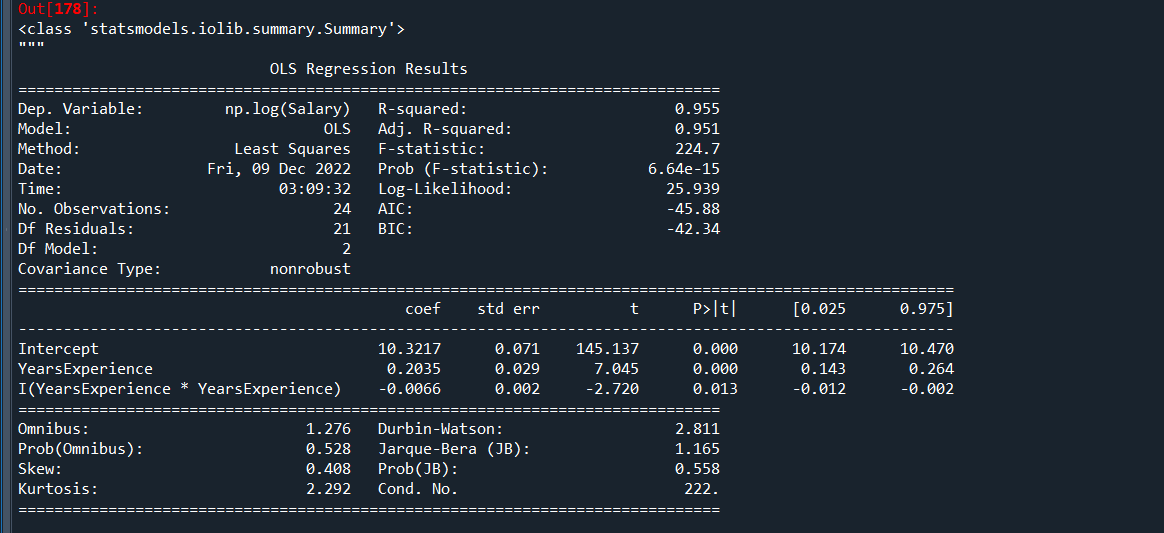
From the above table we got the less error after polynomial transformation. The scatter plot for the same is as below:



From above graph we can see that the as the data was in curve the polynomial model is best fit for prediction. Hence by using polynomial model we make the prediction model for salary hike estimation.



Here we have splitted out dataset into training and testing data randomly in 80% training data and 20% testing data. And then we train our model. And test the data is it gives the correct result or not.

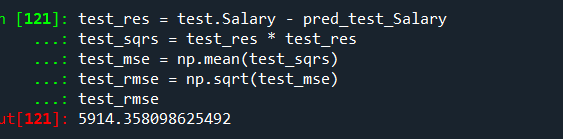
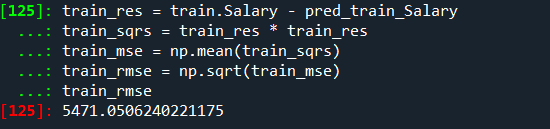


* **Summary table**

the value of R-squared is 0.955 which is less, that means we can consider this output and the value of (p>|t|) is less than 0.005.

we can consider this as the best fit model as it follows the required parameters.

Then after training and testing of the data we calculate the error value to check is the built model is correct or not.

the result of test data is (5914.3580)

the result of train data is (5471.05062)

The model we have made for prediction by applying the polynomial transformation on the data as the scatter plot of the data was a curve. So to predict data and make a good model we need to minimize the error and to minimize the error we have applied different transformations on the data and calculated the error value. We splitted the same data into train and test data randomly to test the accuracy of model. We have trained the data and tested it also and calculated error value. After testing wee got that the model we have built is having less error so we can use it for prediction of salary hike.

* **conclusion**

we can conclude that the given data now shows the prediction values more accurate. this can be used to predict the new / future values.

Most of companies increase the salary of employee based on their years of experience. When the employee work in a company for long time then as a compensation of their work company increases their salary by percent.

So, by using this model you can automatically hike in salary of employee. You do not need to do any calculations. You can hike salary of employee just by using their work of experience.

If you increase salary of employee, then employees work harder and consistently. And this will help in growth of company. So this model will help company and employee to grow.

The employee having experience greater than 4 years should get salary hike.

The employees having experience more then 4 years should get salary hike.