**Employee Salary Prediction in Machine Learning**

We are going to implement our salary prediction using the machine learning algorithm. This model predicts the salary of the employee based on the year of experience of employee. This is a regression problem which is solved using a Linear Regression. This model implemented by using the following steps such as :-

1. Import all required libraries
2. load data
3. perform EDA
4. visualize data
5. prepare data
6. split data into training and testing
7. define model
8. test model
9. check accuracy
10. save the model

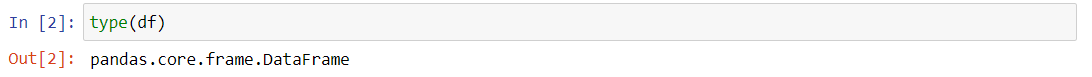
**Problem Statement**

          To build a machine learning model and predict the salary of the employees based on year of experience.

First of all we will import all the required libraries:-



Now we will check the type of the data



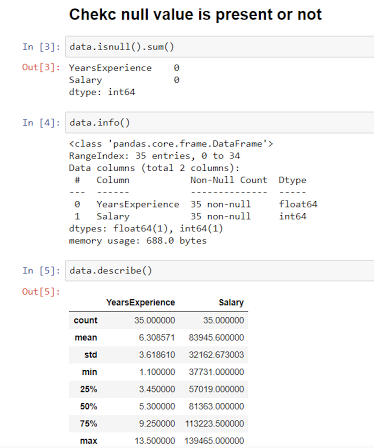
After this we will print df



After printing df we come to know that there are 397 rows and 6columns. The columns are described as rank, discipline, yrs. since phd., yrs. service, sex and salary. The rank is categorized into professor and assistant professor. Discipline is categorized as A,B. The minimum years since phd. is 4 years and the maximum years is 45 years and there are few falling between these also.

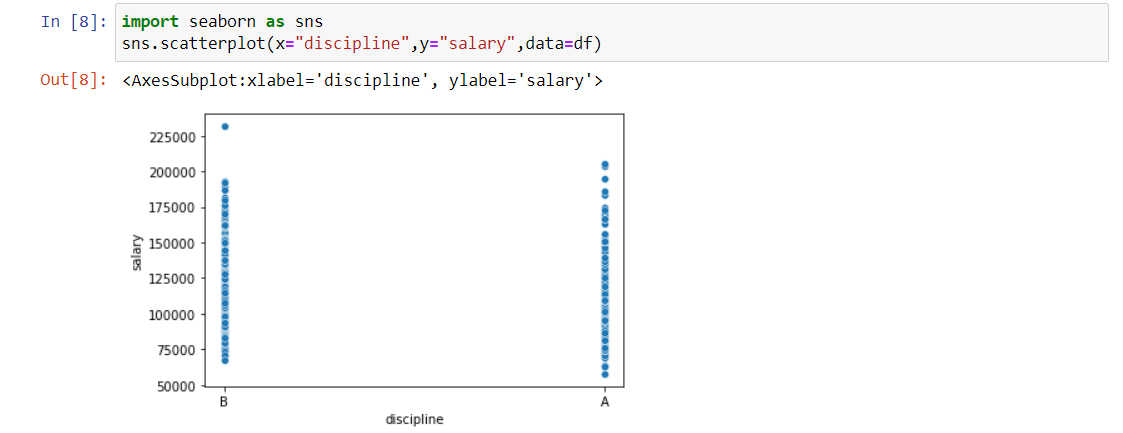
As per above chart we can see that 3 years is the minimum service experience and 30 years is the maximum. The salary ranges as per rank, work experience and on the phd. degree also.

Now perform an Exploratory Data Analysis. In Exploratory Data Analysis, firstly we check that there are Null values present or not,  then check the information of the data, then describe the data which shows the mean value, standard deviation value, minimum value, Maximum value etc.

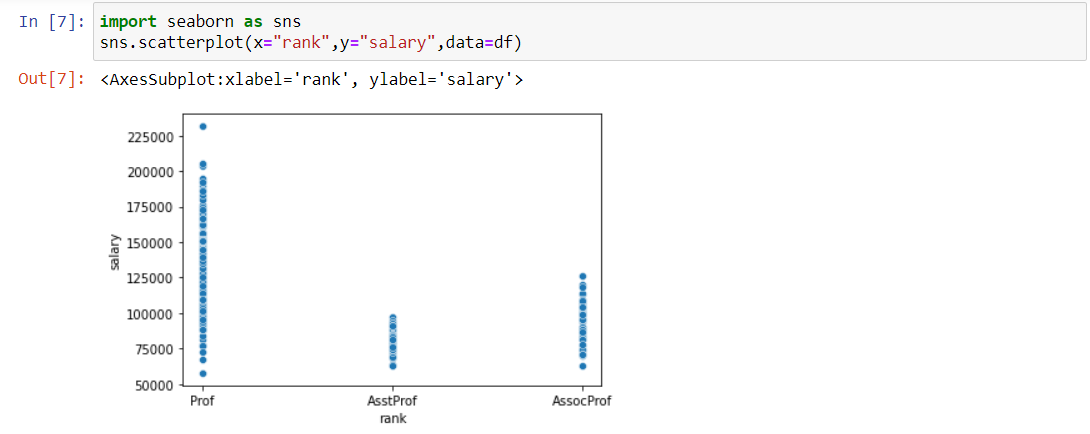


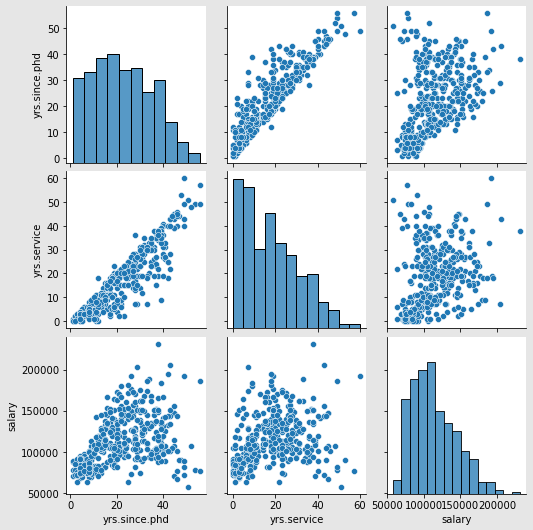
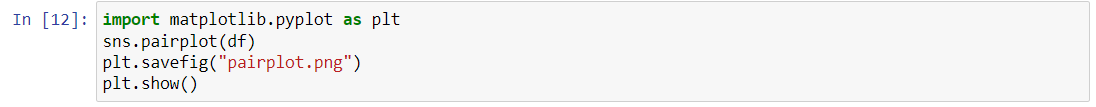
**DATA VISUALIZATION**

Now visualize the data discipline and Salary using the scatterplot function.



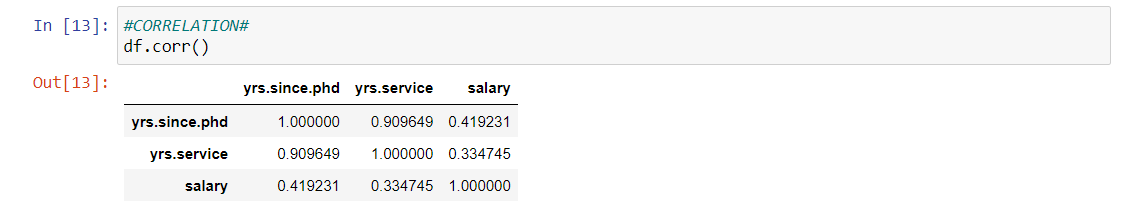
Visualize the data Rank and Salary using the scatterplot function.

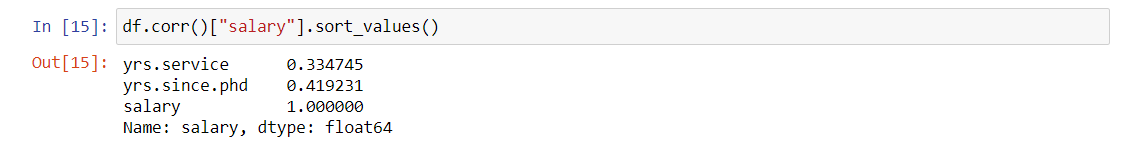




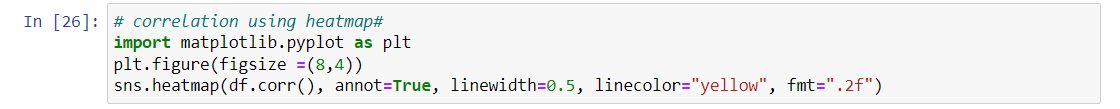
After pair plotting we can see all the columns in one chart. This show us how each and every columns is related to each other. For eg. Salary varies as per the work experience and how many years old phd. degree candidate is holding.

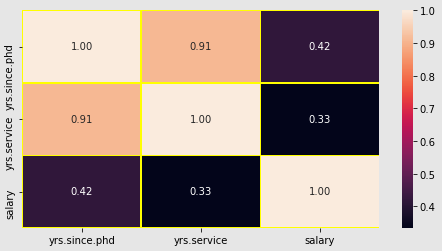
**Now we will check the correlation of the data.**



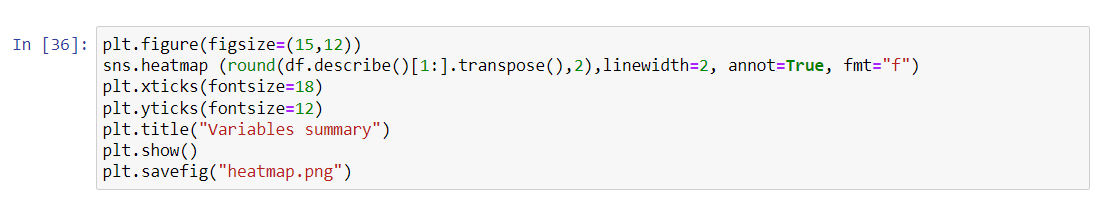


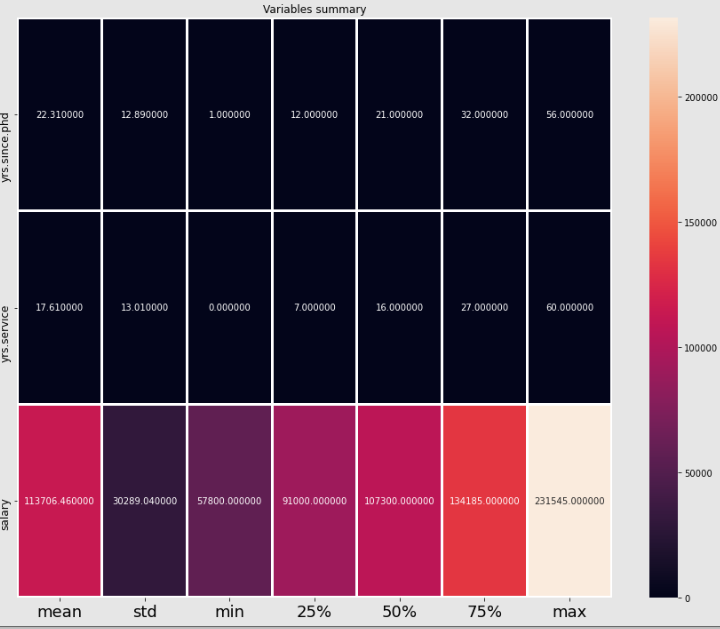
**## Correlation using heat map.**





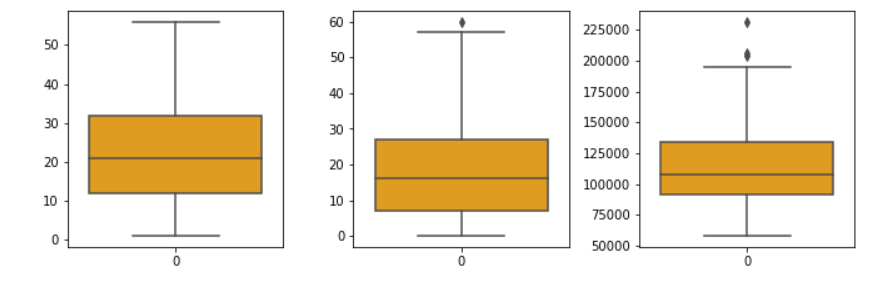
Now after checking the correlation using heat map. We will check the variable summary.



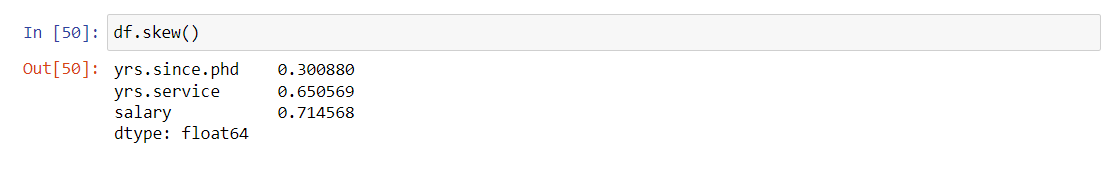


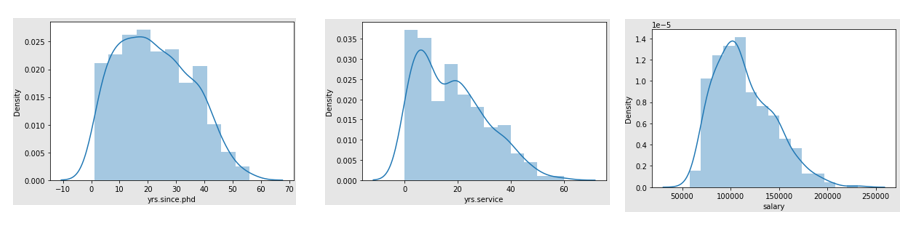
**## Next we are going to check the outlier.**





## After checking the outlier in the data. We are going to see the skewness.

**Normal Distribution Curves:-**



Normal distribution curves have different curve levels as per years since phd., years in service and the salary drawn by the candidate.

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

# Removing Outliers:-

# 

# 

# Before calculating the percentage and making the model we have to remove the outliers. So that we can get the accurate percentage and the correct model.

# #Percentage Data Loss:

# Now define the LinearRegression model with by default parameters and trained LinearRegression model with training data ( X\_train and Y\_train ). And test the model using the testing data (X\_test). and display the predicted and actual data

# 

# 

# 

# 

# Now the final data we received after removing all the unwanted rows and columns. There are 366 rows and 2 columns.

## **Conclusion:-**

1. In this project, We saw how we can build a machine learning model ie., Regression model and predict the salary of the employees based on years of experience.