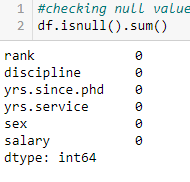
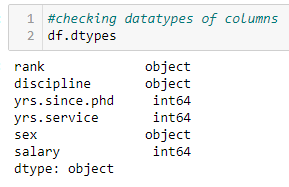
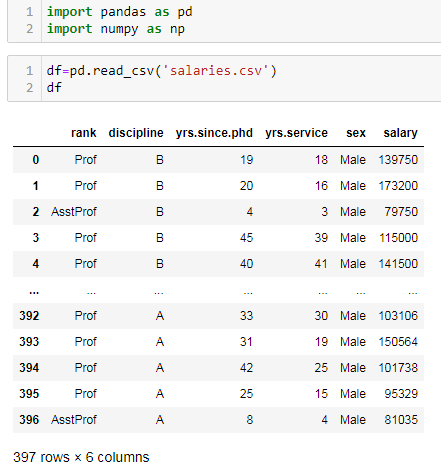
**Salary Project**

**Problem Statement:**

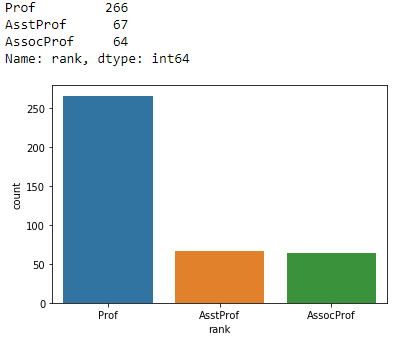
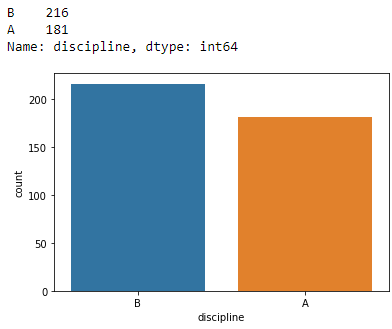
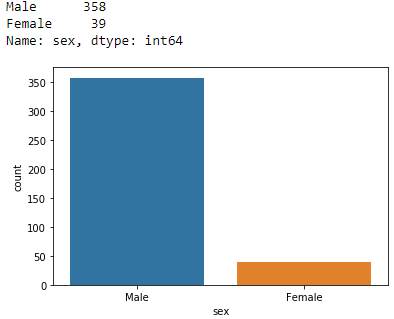
We have to predict the salary from the given information of the employee provided in the dataset.

**Data Analysis:**



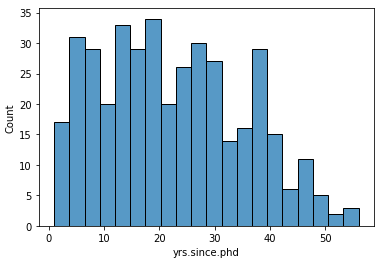
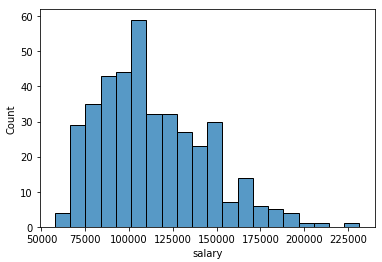
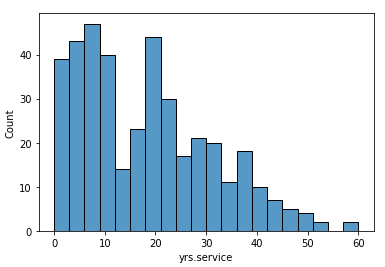
* We are importing the libraries initially important to check the dataset.
* After loading the dataset, we can see that dataset has 397 rows and 6columns.
* ‘rank’, ‘discipline’, ‘sex’ are object datatype whereas ‘yrs.since.phd’, ‘yrs.service’ and ‘salary’ are of int datatype.
* There are no null values or any sort of a missing data in the dataset.

**Data Visualization:**



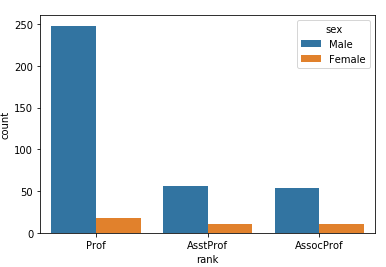
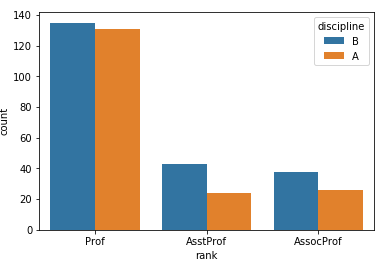
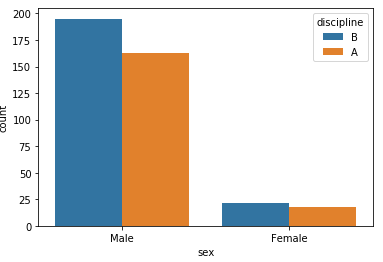
We have plotted the graph for the categorical data in the dataset. Observations for the graphs are as follows:

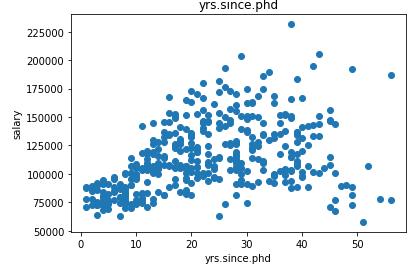
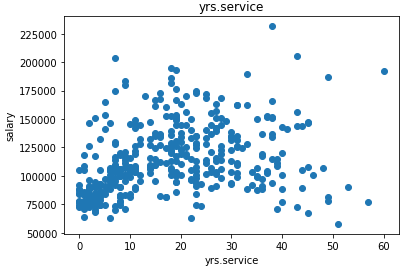
1. ‘sex’ column has two values i.e male and female. The count for male is 358 and for female is 39. We can clearly see that the male employees are more than the female employees working in the company.
2. ‘discipline’ column has two values i.e A and B. The count for rankA is given to 181 employees. The rankB is given to 216 employees. Maximum employees has remarked as B in discipline.
3. ‘rank’ column has three values which are professor, assistant professor and associate professor. The count of professor is 266. The count of assistant professor is 67. The count of associate professor is 64. Maximum employees are professor in the company.

We have plotted the graph for the continuous data in the dataset. Observations for the graphs are as follows:

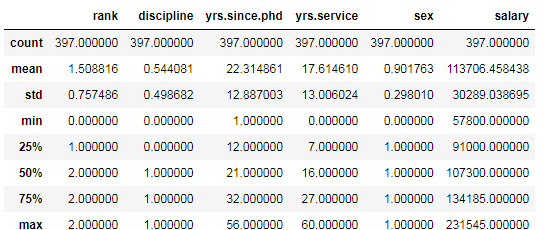
1. ‘yrs.since.phd’ column has a continuous data. We can see that maximum employees has the phd from 15-20years. Very less employees hold the phd degree since 45-50years.
2. ‘salary’ column has a continuous data. We can see that maximum employee has the salary between 100000 to 125000. Very less employees has the salary of 225000.
3. ‘yrs.service’ column has a continuous data. We can see that maximum employees are in the range of 1-5years of service, followed by employees having 5-10years of service and least employees are in the range of 50-60years of services.

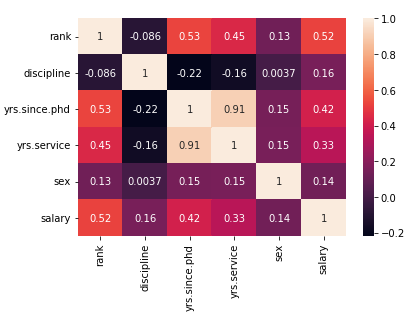
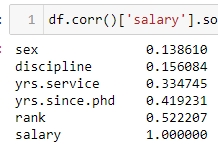
1. Maximum male employees are professor and very less are female professor.
2. Male assistant professor and associate professor are more as compare to females and less as compare to male professor.
3. Maximum professors has ranked has B followed by assistant professor and associate professor. Rank A is marked maximum to professors.
4. Maximum male employees are ranked as B as compared to female employees.
5. Employee with less years of services gets less salary. Employee with more years of services get maximum paid.
6. The more years since phd holders employees get maximum salary as compared to employees holding phd for less years.

**Checking statistical information:**



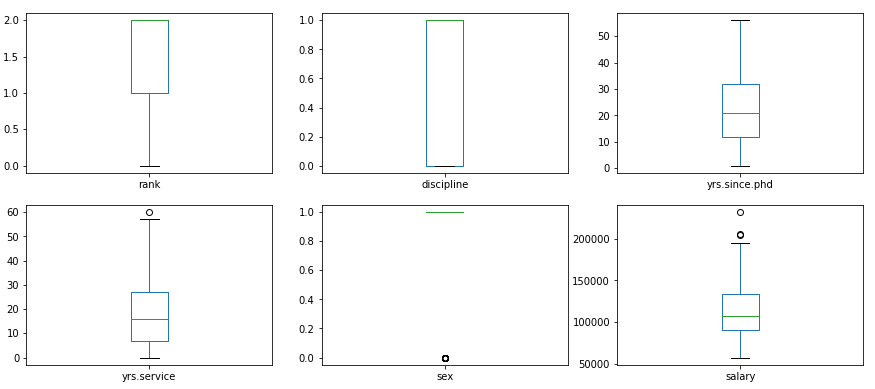
1. Count for all the columns are same.
2. Mean is less than 50% for all the columns.
3. Only for yrs.service and salary column mean is greater than 50%.
4. Max and 75% are same rank, discipline and sex columns.
5. There is a huge difference between max and 75% in salary, yrs.since.phd and yrs.service columns.

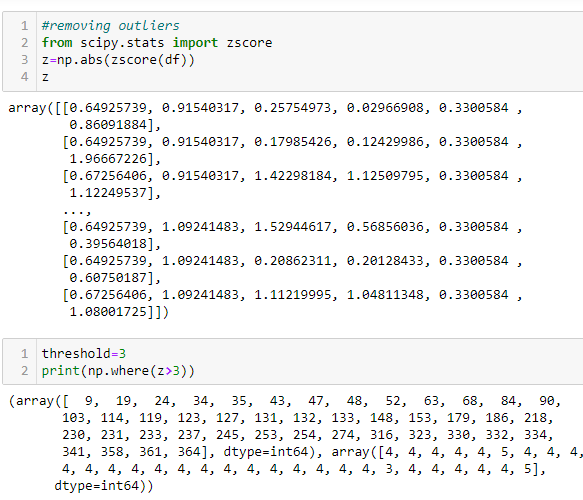
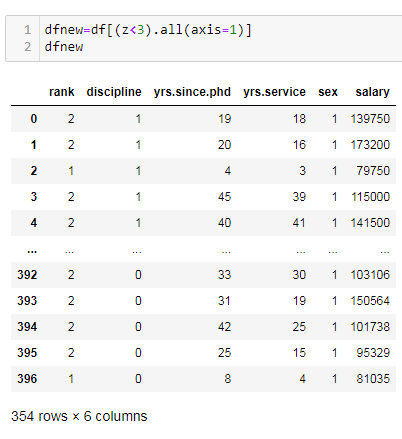
**Checking the correlation with target column:**

1. salary has maximum correlation with rank then with yrs.since.phd and then with yrs.services.
2. salary column is least positively correlated with sex.

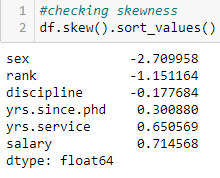
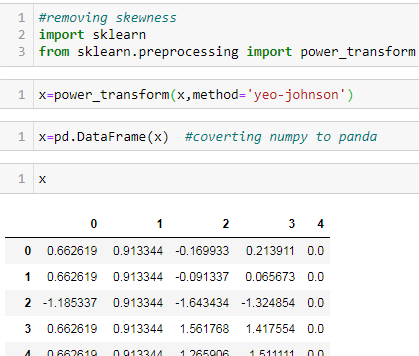
**Checking and handling outliers:**



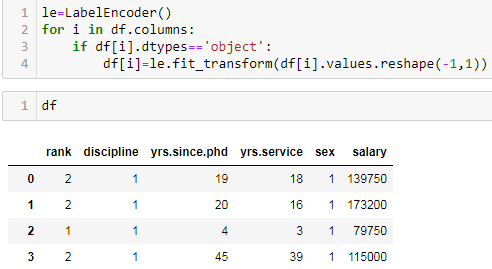
1. An outlier is a data point in a data set which lies outside the overall distribution which is available in the dataset.
2. A box plot is a method or a process for graphically representing groups of numerical data through their quartiles. Outliers may also be plotted as an individual point. If there is an outlier it will plotted as point in box plot but other numerical data will be grouped together and displayed as boxes in the diagram. In most cases a threshold of 3 or -3 is used i.e., if the Z-score value is higher than or less than 3 or -3 respectively, that particular data point will be identified as outlier.
3. As the outliers were present in the dataset, we treated by zscore.
4. After treating outliers, 43 rows were dropped, i.e the present rows are 354 in the dataset.

**Checking and handling skewness:**

Skewness refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution in a set of data. Besides positive and negative skew, distributions can also be said to have zero or undefined skew. The skewness value can be positive, zero, negative, or undefined. We have two method 1. yeo-johnson' which is default one and 2. Boxcox.

**Pre-Processing Pipeline:**

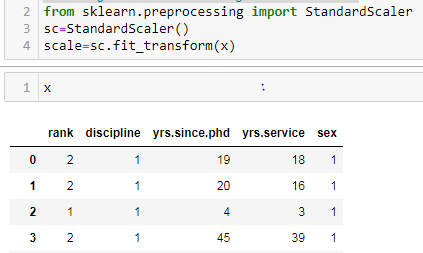


Label encoding, machine learning algorithm takes numerical input for the learning, thus it becomes important that data should be converted into number. We have a coverted ‘sex’, ‘rank’ and ‘discipline’ column into integer datatype.

**Building Machine Learning Models**

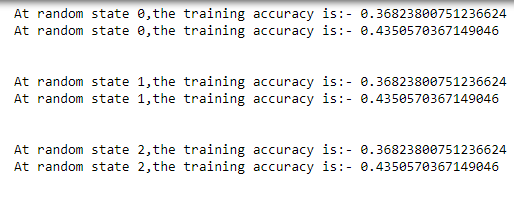
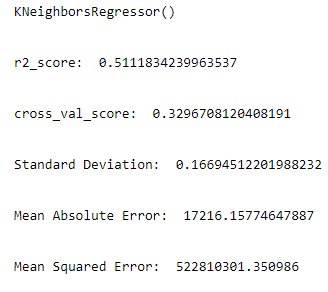
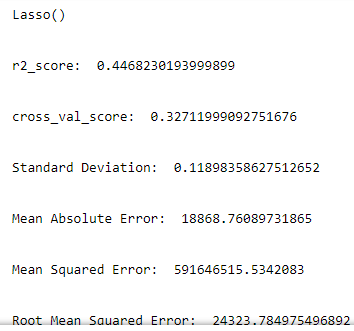
As target column which is ‘salary’ has the continuous data, we are going to use linear regression.

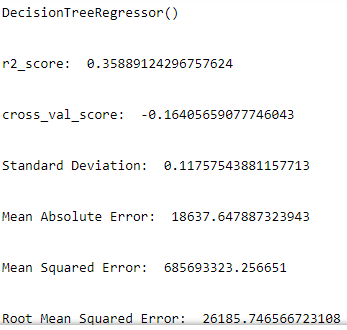
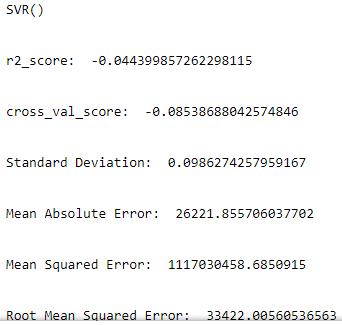
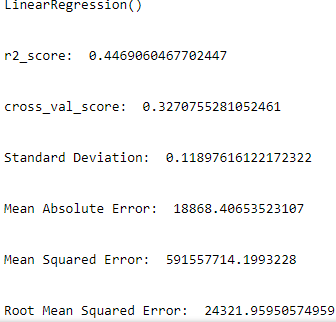
**Scaling the dataset using StandardScaler:**



For each value in a feature, StandardScaler subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum. StandardScaler preserves the shape of the original distribution.

**Run and evaluate selected models:**

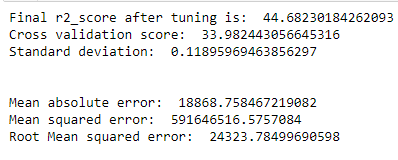
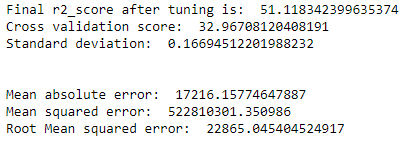
  

1. linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regressions.
2. lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.
3. Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
4. KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood*.*
5. As all the random state gave us the same value, we selected the random state as 4.
6. After applying the models, knr gave the maximum r2score, cross validation score, mean squared error and mean absolute error.

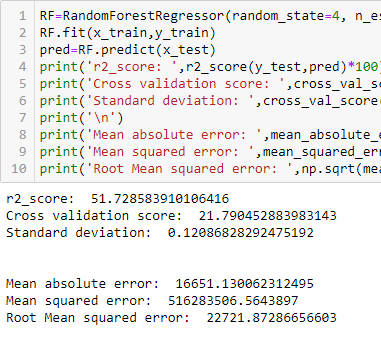
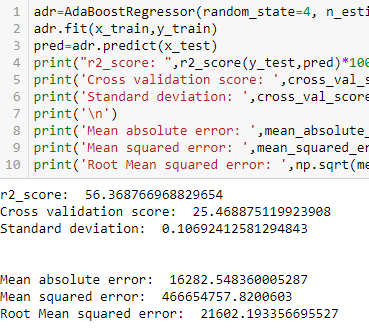
**Hyperparameter Tuning:**

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

**GridSearchCV:**

GridSearchCV is a function that comes in Scikit-learn (or SK-learn) model selection package. An important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

After applying Ensemble Techniques, we can see that AdaBoostRegressor is the best performing algorithm among all other algorithms as it is giving a r2\_score of 56.36 and cross validation score of 25.46. It has also the less amount of error values obtained. Lesser the RMSE score, the better the model. Now we will finalize the model.

**Saving The Final Model:**

Pickle is a useful Python tool that allows you to save your ML models, to minimise lengthy re-training and allow you to share, commit, and re-load pre-trained machine learning models.



**Conclusion:**

Dataset was very short and simple. It was the mixture of categorical and continuous data. We did all the important and possible data analysis. Build the best possible model of machine learning.