Baseball Case Study

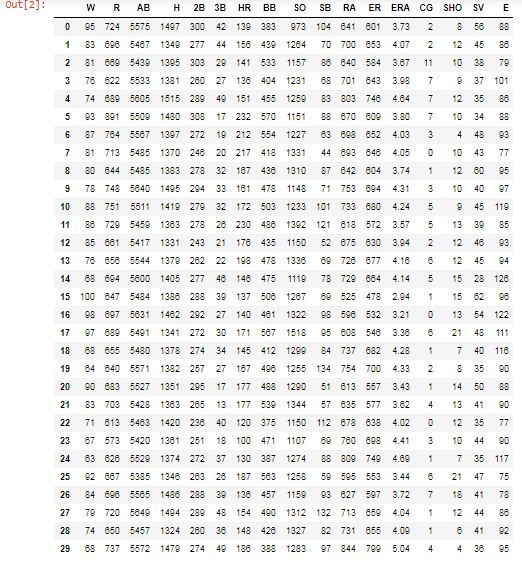
This dataset utilizes data from 2014 Major League Baseball seasons in order to develop an algorithm that predicts the number of wins for a given team in the 2015 season based on several different indicators of success. There are 16 different features that will be:-

**-- Input features:** Runs, At Bats, Hits, Doubles, Triples, Homeruns, Walks, Strikeouts, Stolen Bases, Runs Allowed, Earned Runs, Earned Run Average (ERA), Shutouts, Saves, and Errors.

These features are used as the inputs to the machine learning and the output will be a value that represents the number of wins.

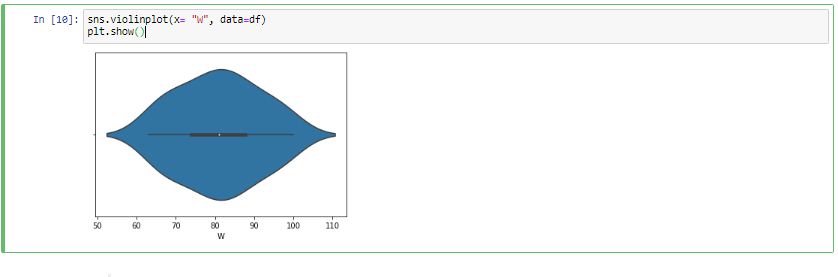
**-- Output:** Number of predicted wins (W).

1. This is the dataset that we have :-



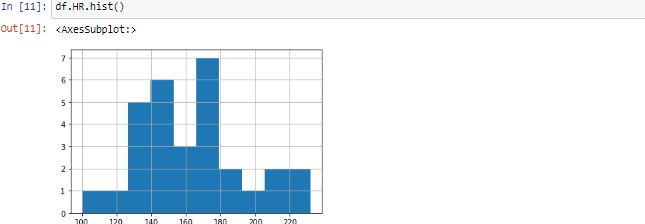
**Visualization of the data:-**

1. In this line code we are plotting the number of wins in violinplot.



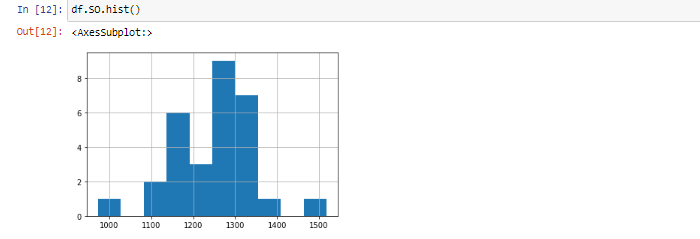
* Through the above image we can observe that majority of the teams have number of wins in the range of 65 to 100.

1. In this line of code we are making a Histogram for ‘Home Runs’.



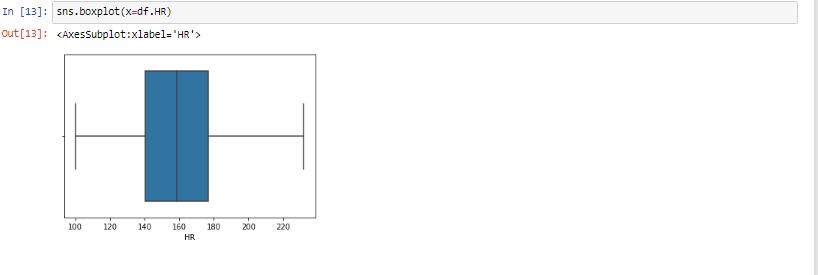
* As we can see through the above graph the runs scored by home runs by the teams are majority lie in the range of 127 to 180.

1. In this line of code we are making the Histogram for ‘Strike Out’.



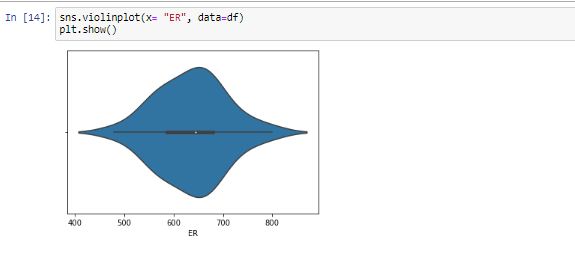
* Through the above graph we can see the majority of players that strike out in a team in this season(2014) are lie in the range of 1100 to 1350.

1. In this line of code we are making boxplot for ‘Home Runs’.



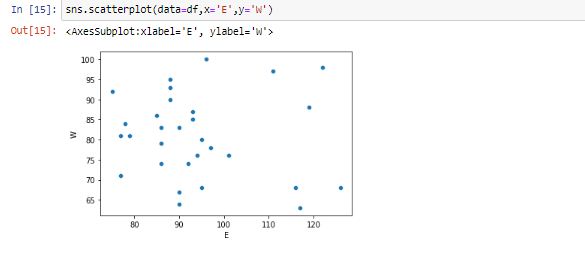
* We Can see from the above graph that the minimum 100 Home runs is scored by a team and maximum 230 Home Runs are scored by a team in this season(2014). The range in which most of the teams scored Home Runs is between 140 to 178.

1. In this line of code we are plotting a violinplot for ‘Earned Runs’.



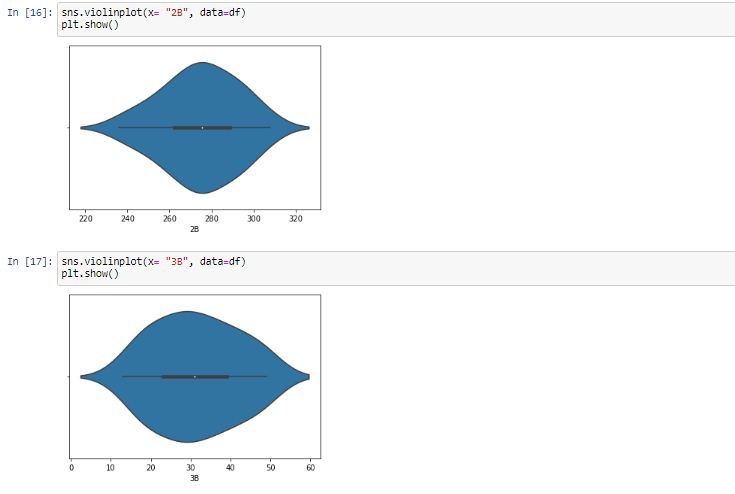
* We can see through above graph that majority of the team scored 550 to 750 runs.

1. In this line of code we are plotting a scatterplot for ‘Errors’ and ‘wins’.



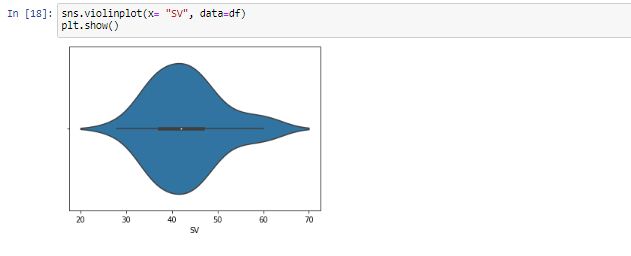
* As we can see through the above graph that as the errors are increasing, the number of wins are becoming less.

1. In the below two line of code we are plotting a violinplot for 2bases and 3bases.



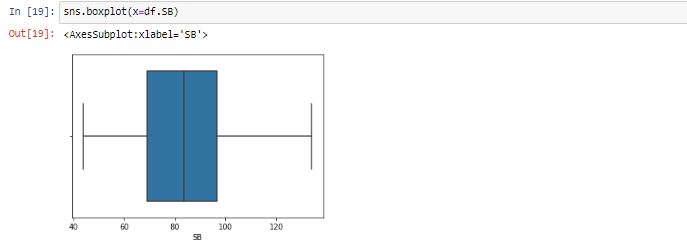
* Through the above 2 graphs, we can observe that Majority of the times the players reach the Second base as compare to the third base.

1. In this line of code we are plotting a violinplot for number of saves.



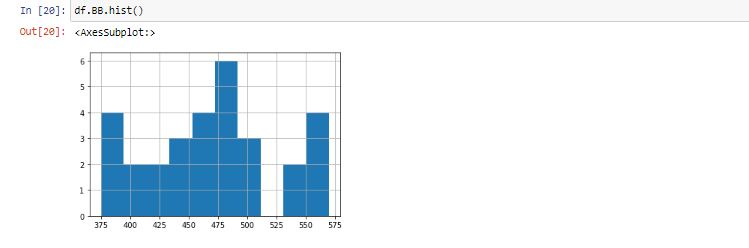
* Through the above graph we can see that between 30 to 55 times the players are saved in the 2014 season.

1. In this line of code we are plotting a boxplot for the stolen bases by the battar.



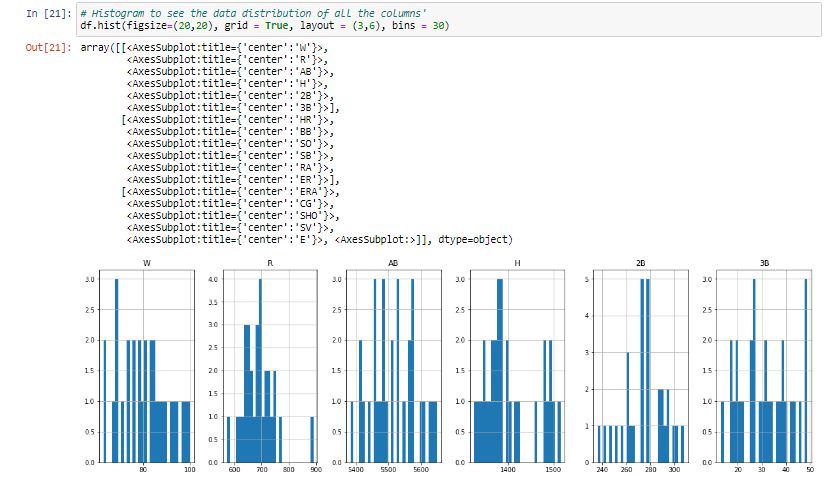
* Through the above graph , we can see that 70 to 95 times the bases were stolen by the batter and the pitcher was not able to out that batter.

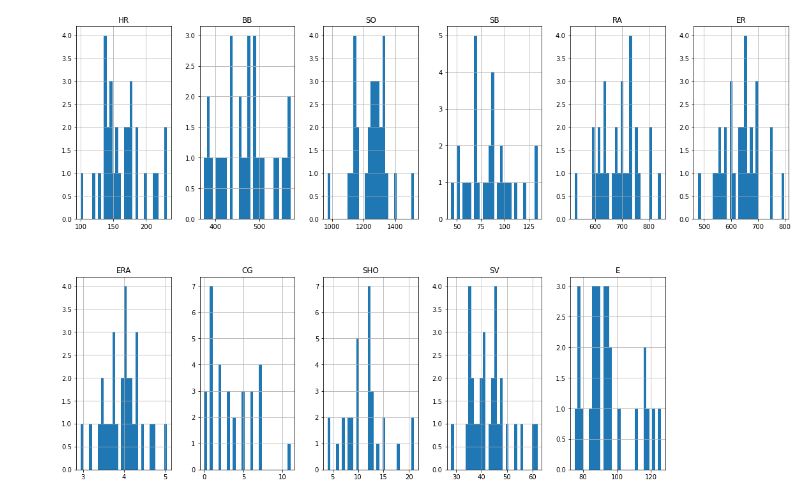
1. In this line of code we are plotting a histogram for the number of walks battar are getting by the umpire.



* From the above graph we can see that the batter is getting the walks by the umpire in the range of 375 to 570.

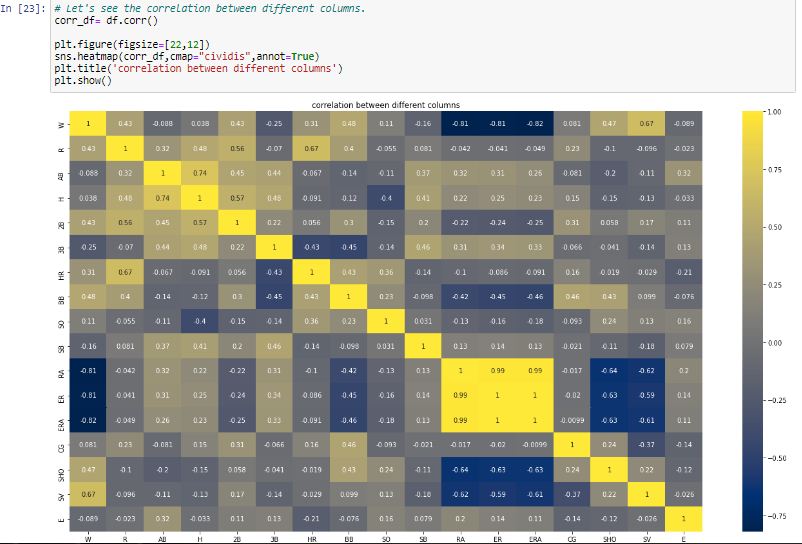
1. In this line of code we are plotting the histograms for every column in the dataset.





* Through the above graphs we can see that data is highly distributed in almost all the columns.

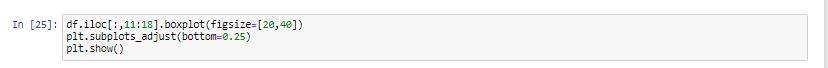
1. In this line of code we are checking the correlation between different column of the dataset.

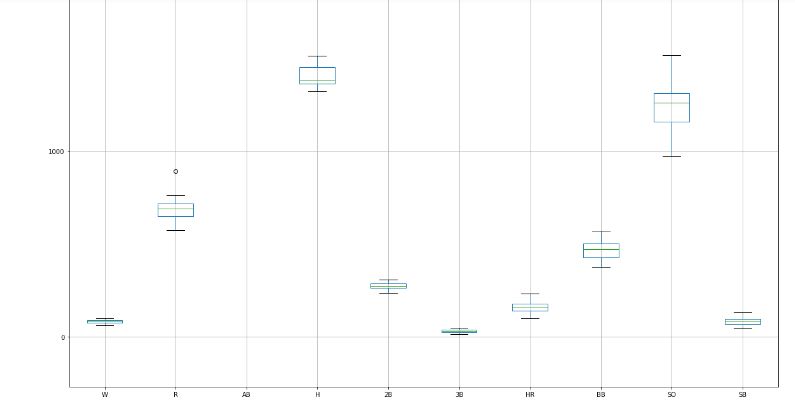
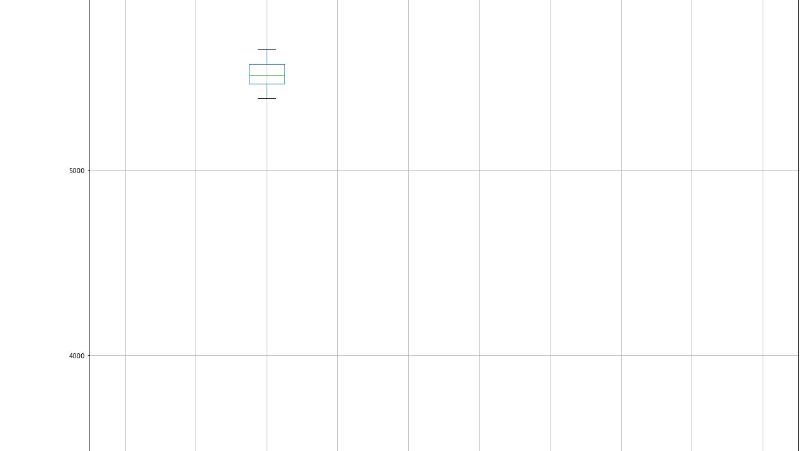


* Through the above correlation heatmap following are the findings:-
* Runs allowed, Earned runs, Earned runs average and saves are the most positive correlated columns with the wins columns.
* Error are the least correlated column with the wins column, as if errors are increasing the no of wins will be decreased.
* We can observe that all the columns which have runs in it are moderately positive correlated to the wins column as runs help the team to win the match

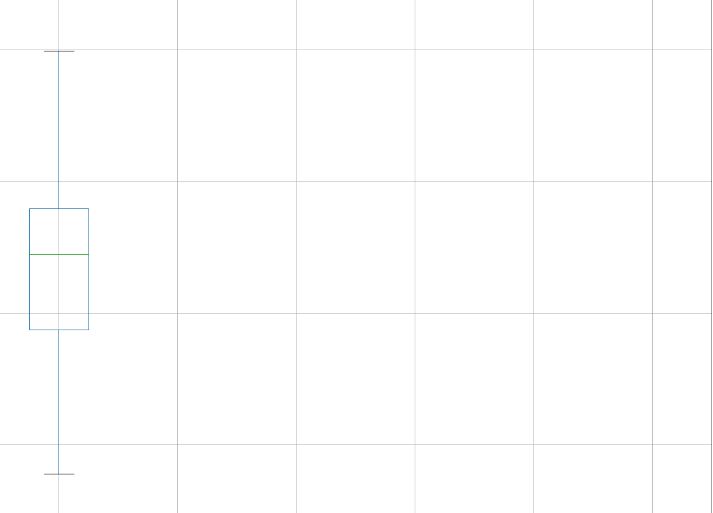
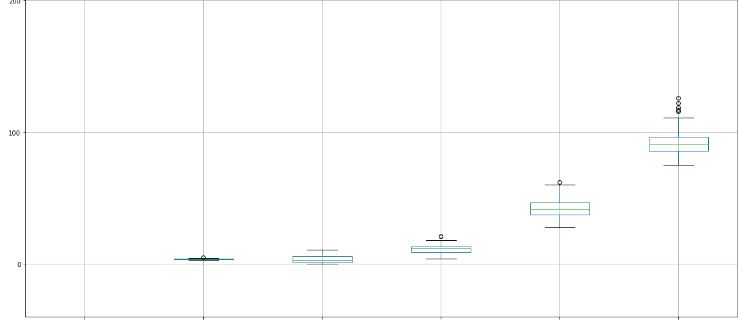
**Data Pre-Processing:-**

1. In this line of code we are checking the outliers in the dataset. We are taking 10 column in one code and remaining column in next code as it will help us to understand more quickly and effectively. By using ‘iloc’ method, we are taking 10 columns first.

VD 13(3).JPG

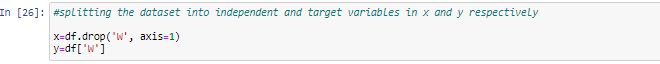


* As we can see through the above 2 plots that their are no considerable outliers present in the dataset.

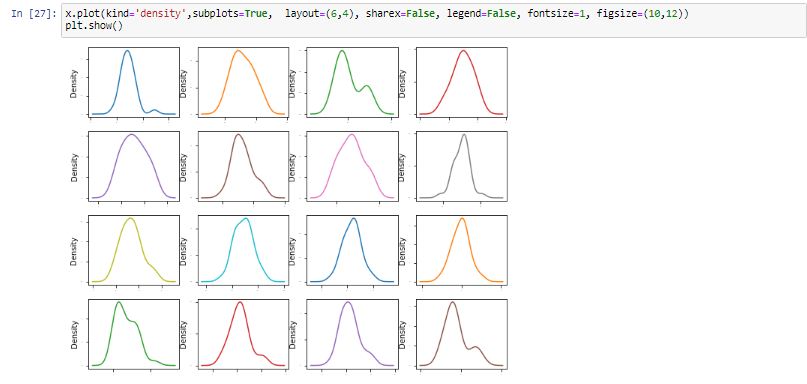
.

* As we can see through the above 2 plots that their are no considerable outliers present in the dataset.

1. In this line of code we are splitting the dataset into ‘x’ and ‘y’ variable by using the ‘drop’ function. In ‘x’ variable all the independent columns are there and in the ‘y’ variable the target column is added.



1. In this line of code we are using a plot in which we can see the skewness of different columns in the input variable ‘x’.

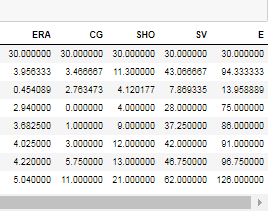
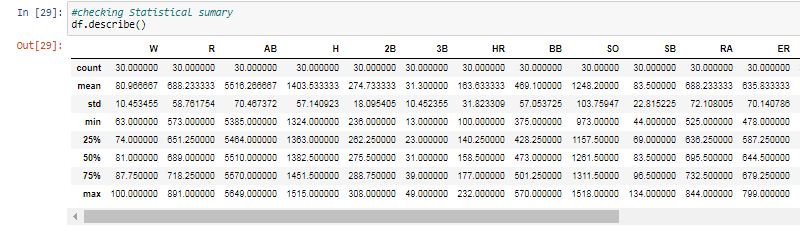


1. In this line of code we are checking the skewness values of different columns in the ‘x’ variable by using the ‘skew’ function.



* As we see from the above desription that their is some skewness present in the 'Runs' column but it might be possible that some teams had scored more runs and some teams had scored less runs, so we don't have to remove skewness from the runs column.

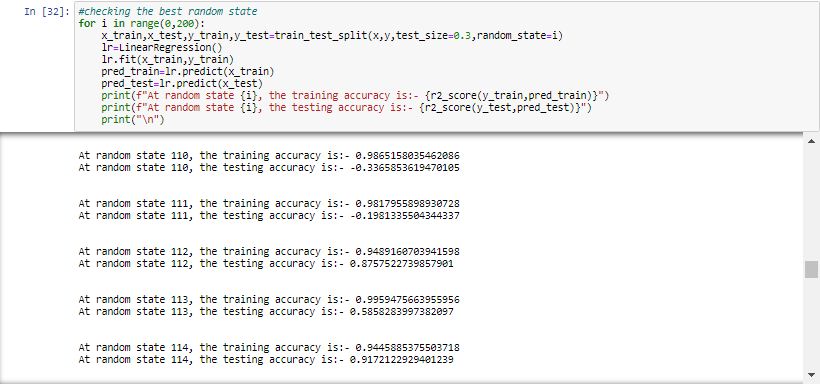
1. In this line of code we are describing the various statistics of the dataset.



* We can see through the above description that the standard deviation are high in 'R', 'AB', 'H', 'BB', 'SO', 'RA', 'ER'. So that means the data is spread too much.
* Range is high in these columns.

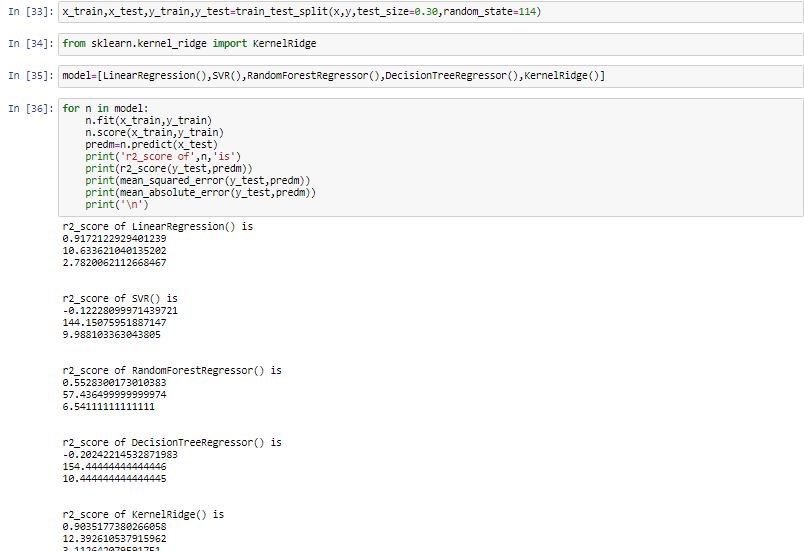
**Model Building**:-

1. In this line of code we are checking the best random state for our algorithms that we are going to use later.



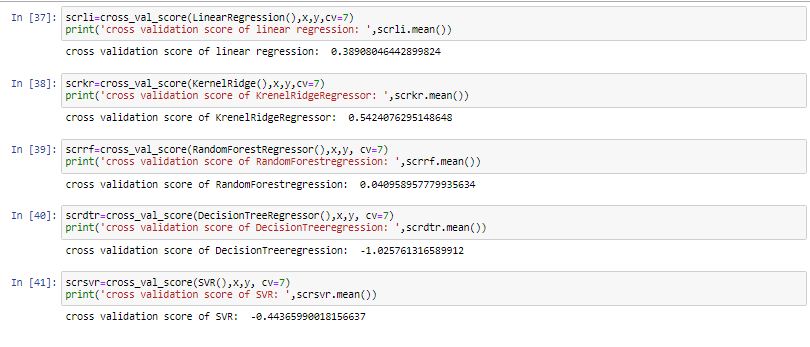
* From the above stats we can see that at random state 114, the testing accuracy and the training accuracy is highest.
* Now we have to test the best random state in different algorithms

1. In this line of code we using splitting the data into training and testing to add the data in different algorithms. In this case study we are using Linear Regression, SVR, RandomForestRegressor, DecisionTreeRegressor, Kernel Ridge.



* We can see that the best score is for Linear Regression.

1. Now checking the Cross Validation Score to check whether our models are overfitted or underfitted.



* We can see that Minimum difference in r2\_score and cross validation score is for KernalRidge. So it is our best model.
* Now we have to do hyperparameter tuning to increase our model score.

1. In this line of code we are doing hyperparameter tuning for our best model that is Kernel Ridge. For Hyperparameter tuning we are using GridSearch CV. It helps us to find the best parameters for kernel Ridge Regression.



* In the first line of code we make a dictionary for the parameters that we are using in our model. I am using ‘alpha’ , ‘degree’ , ‘kernel’.
* After I pass the dictionary and our best model in the GridSearchCV and fit it in our train dataset.
* After the parameters are fitted in the train dataset , we get the best parameters by using ‘best\_params\_’ function available in GridSearchCV.
* Now we pass these parameters to our independent columns of the test dataset.
* After passing the best parameters, we can see that the score of the kernel ridge model is increased.
* At last we save our model by using joblib function. We just have to call this model name to predict the wins for the baseball team.

***Case Study:-***

* + This dataset utilizes data from 2014 Major League Baseball seasons in order to develop an algorithm that predicts the number of wins for a given team in the 2015 season based on several different indicators of success. There are 16 different features which are Runs, At Bats, Hits, Doubles, Triples, Homeruns, Walks, Strikeouts, Stolen Bases, Runs Allowed, Earned Runs, Earned Run Average (ERA), Shutouts, Saves, and Errors that will be used as the inputs to the machine learning and the output will be a value that represents the number of wins. With these inputs we have to predict the number of wins for a given team.

***Findings:-***

* + the columns in the dataset have only numeric values.
  + there are no null values present in the dataset.
  + majority of the teams have no of wins in the range of 65 to 100.
  + the runs scored by home runs by the teams are majority lie in the range of 127 to 180.
  + the majority of players that strike out in a team in this season(2014) are lie in the range of 1100 to 1350.
  + the minimum 100 Home runs is scored by a team and maximum 230 Home Runs are scored by a team in this season(2014).The range in which most of the teams scored Home Runs is between 140 to 178.
  + majority of the team scored 550 to 750 runs.
  + we can observe that Majority of the times the players reach the Second base as compare to the third base.
  + we can see that between 30 to 55 times the players are saved in the 2014 season.
  + we can see that 70 to 95 times the bases were stolen by the batter and the pitcher was not able to out that batter.
  + Runs allowed, Earned runs, Earned runs average and saves are the most positive correlated columns with the wins columns.
  + Error are the least correlated column with the wins column, as if errors are increasing the no of wins will be decreased.
  + We can observe that all the columns which have runs in it are moderately positive correlated to the wins column as runs help the team to win the match.

***Model building:-***

* + we can see that at random state 114, the testing accuracy and the training accuracy is highest.
  + The best r2\_score we get are from Logistic Regression and kernel Ridge.
  + After doing Cross Validation. Minimum difference in r2\_score and cross validation score is for KernalRidge. So it is our bestmodel.
  + After doing hyperparameter tuning. we increase the score of our model from 54% to 92%.