





Submitted by:- Ayush Pathak

**ACKNOWLEDGMENT**

I would like to express my gratitude towards Datatrained for giving me this opportunity to show case my talent and also for their constant support and guidance. Also, I would like to thank all data-trained support member’s for helping me all the time whenever needed. I express my deepest thanks to **Deepika Mam**, for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make my project easier. I choose this moment to acknowledge her contribution gratefully. I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Thanks & Regards

Ayush Pathak

**INTRODUCTION**

* Business Problem Framing

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

* Conceptual Background of the Domain Problem

We are going to test multiple regresssion algorithms with multiple evaluation metrics and select the best model based on proper metric, perform GridSearchCV for best parameter settings.

* Review of Literature

First of all data is loaded and then it can be observed that, this is a regression based problem. It is clearly visible that the dataset needs not that much of cleaning as there are no garbage & outliers values present in our dataset, luckily there’s no missing value(no null values present).

* Motivation for the Problem Undertaken

As, after looking into the dataset, it is clearly visible that this pretty straigh forward but handling such datasets brings your perception, judgemental ideas and curiosity to the next level.And, after working on this assignment many challenges I have faced, that boosten up my confidence too and a lot more things I have to learn also.

**Analytical Problem Framing**

The sample dataset is provided to us from our project database.

**Data Analysis:-**

Source: <https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/baseball.csv>

The data set have 16 independent variables and 1 target variable, i.e. Baseball dataset.We have 30 rows and 17 columns in the dataset. The column names are as follows:-

W- Number of wins (Target variable)

R- Runs

AB- At Bats

H- Hits

2B- Doubles

3B- Triples

HR- Homeruns

BB- Walks

SO- Strikeouts

SB- Stolen Bases

RA- Runs Allowed

ER- Earned Runs

ERA-Earned Run Average

CG- Complete Games

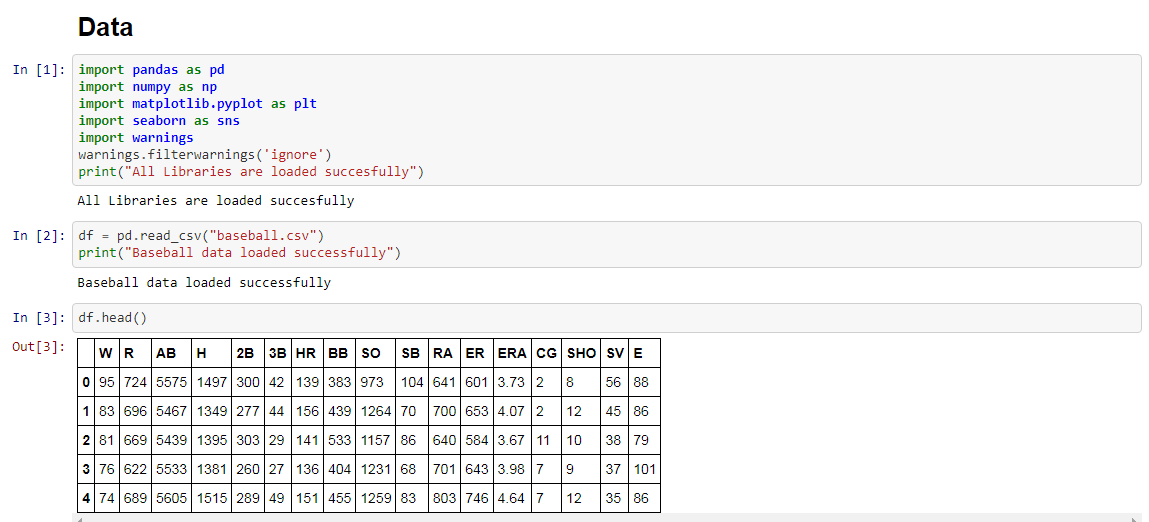
SHO-Shutouts

SV- Saves

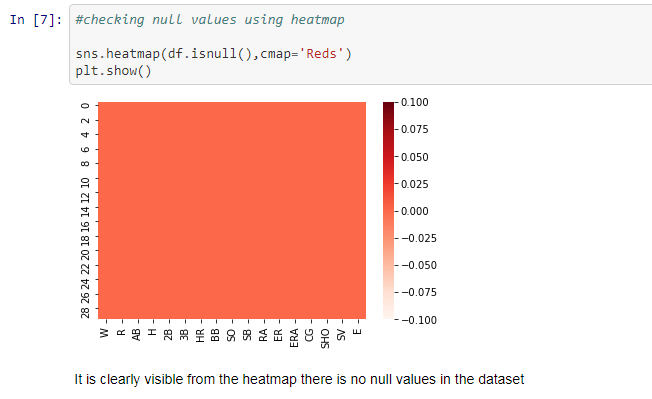
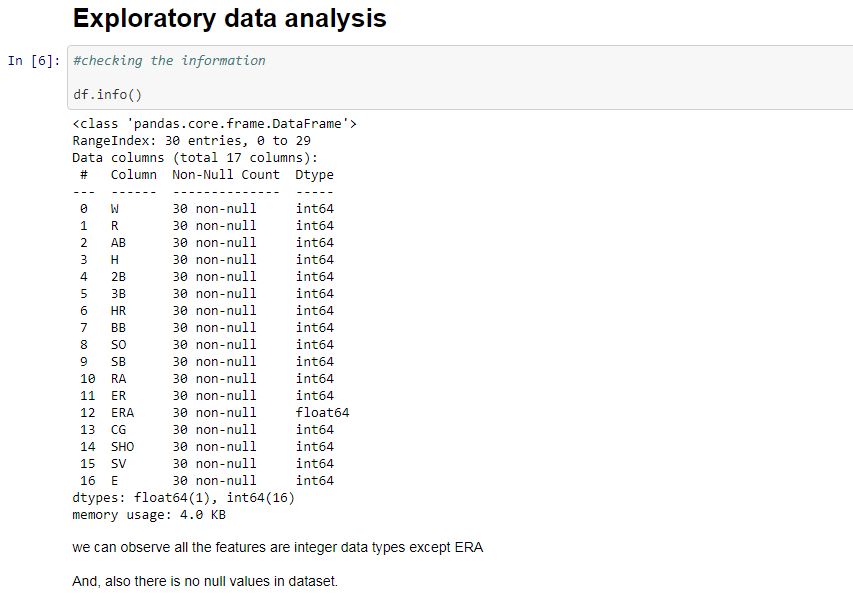
E- Errors

**Steps-**

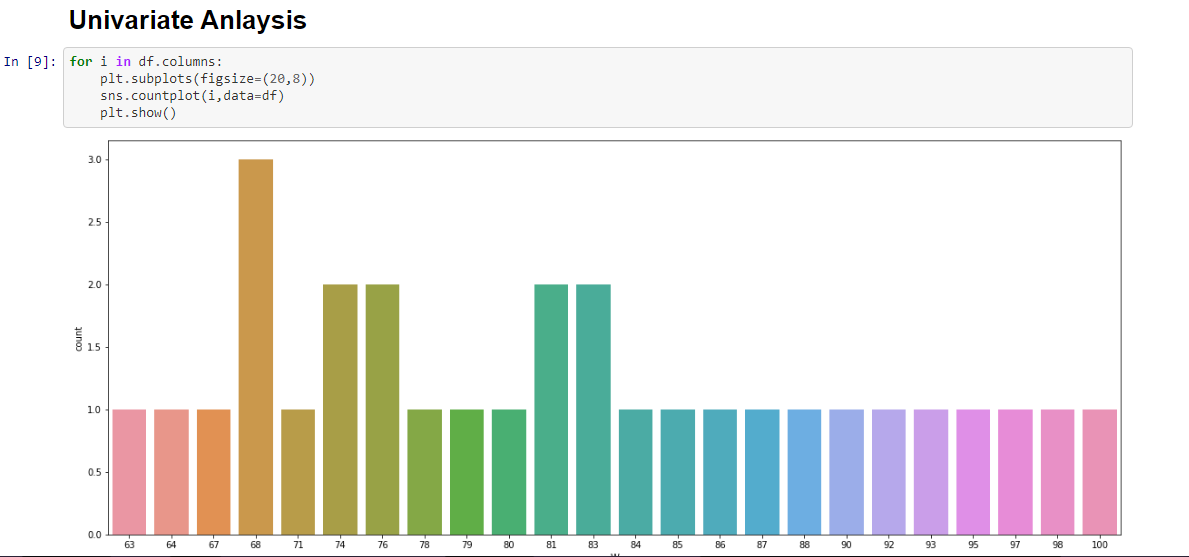
1) Firstly data is being loaded and previewing the first 5 rows.



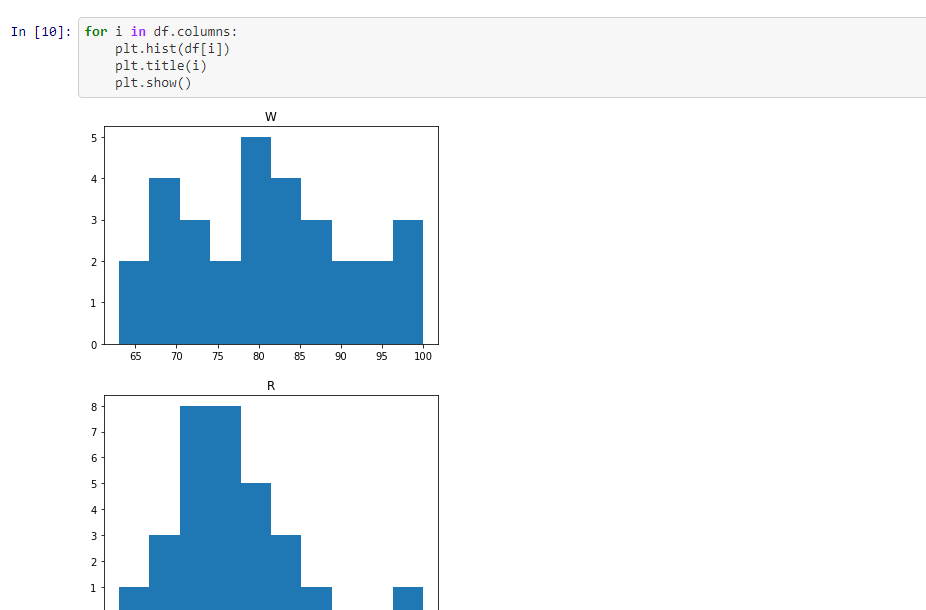
2) Analysing the columns present in the dataset and checking the basic information(if null values are present or not).



3) Now we are going for data visualization, firstly we identify all the columns using “countplot” & “histogram”.We are working here on univariate analysis of dataset.



Like similarly for all the columns, visualizing.



Observations we got after this as follow-

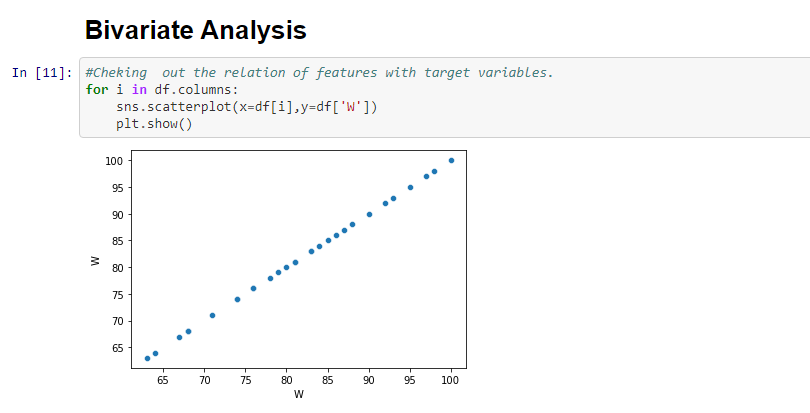
a) number of wins ranges from 63 to 100, and most number of wins at 68,74,76,81& 83.

b) runs ranges from 570 to 891.

c) at bats ranges from 5380 to 5650.

d) there is gap in between ranges in theese columns - R,H,2B,BB,SO,RA,ER,CG,SHO & E.

4) Now we are again going for data visualization using scatter plotting,we are working here on bivariate analysis of dataset.

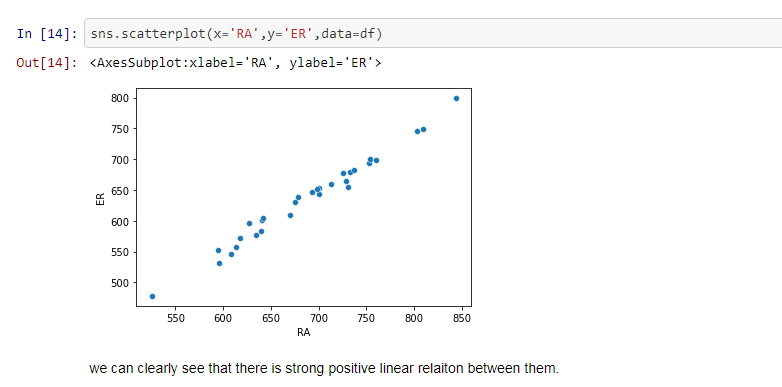


Similarly for all the columns plotted with target variable “W”.

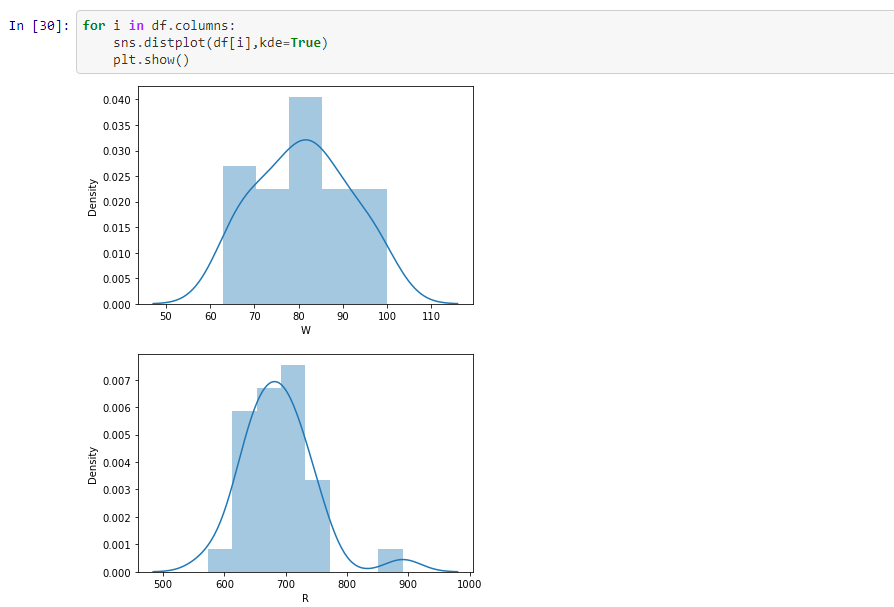
Observation we got after this plot-

a)There is positive linear relation between W with R,SV,BB & SV.

b)There is a negative linear relation between W with RA, ER & ERA.



5) Now we are plotting column using seaborn as distplot for checking the distribution of the each column whether normal distributed or not? Skewness is present or not!!

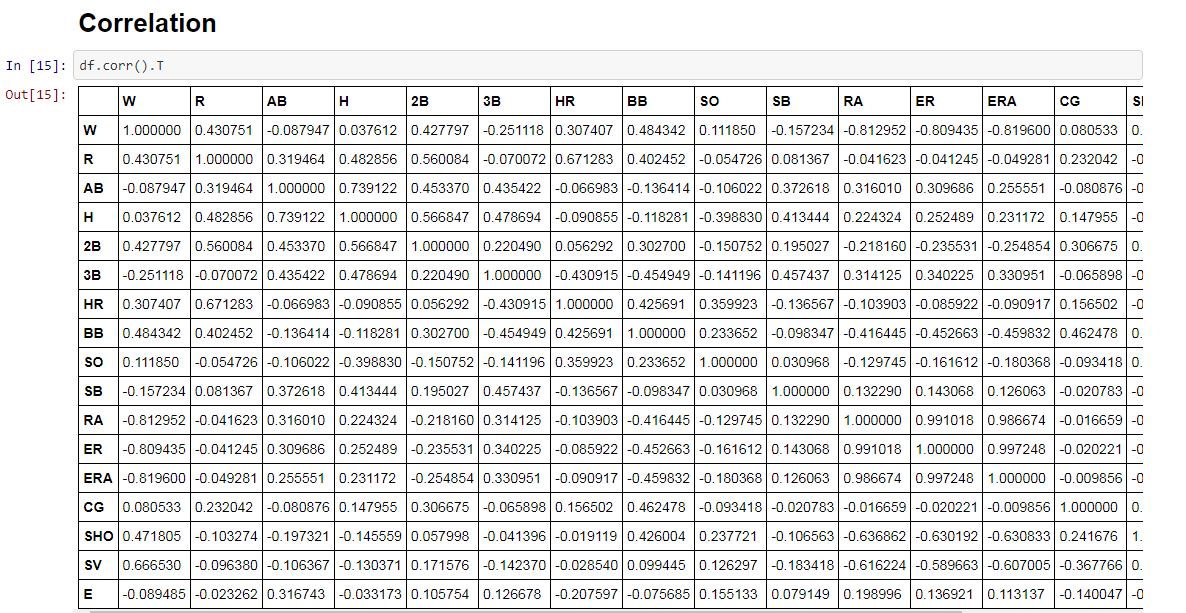


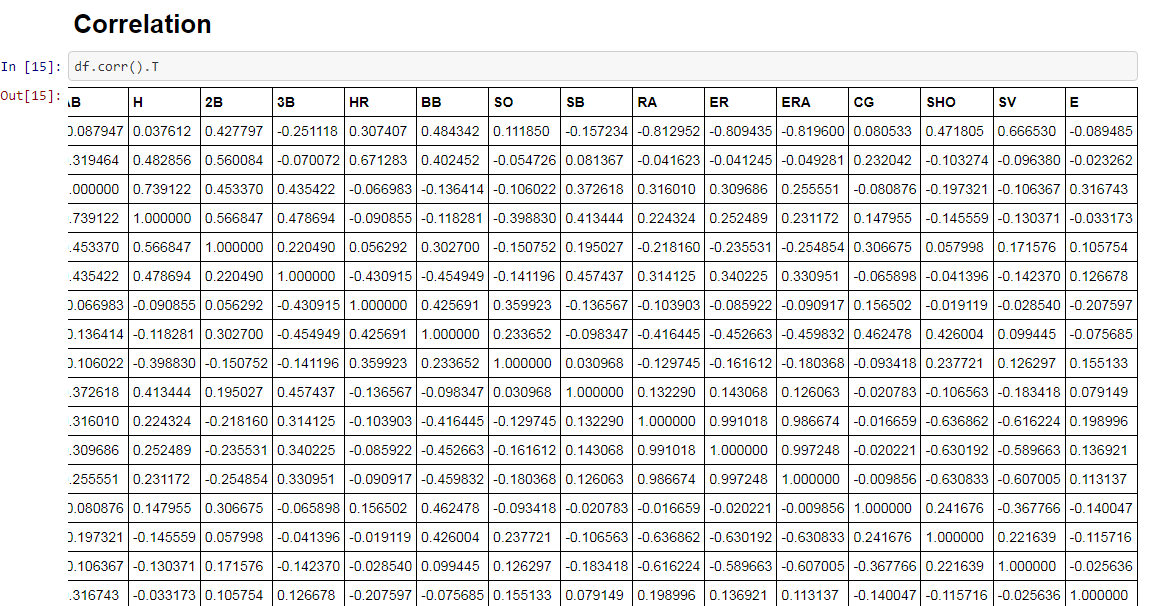
Similarly for other columns too.

Observation we got after this –

Most of the columns are having normal distribution, few columns have skewness we can observe that from the above plotting.

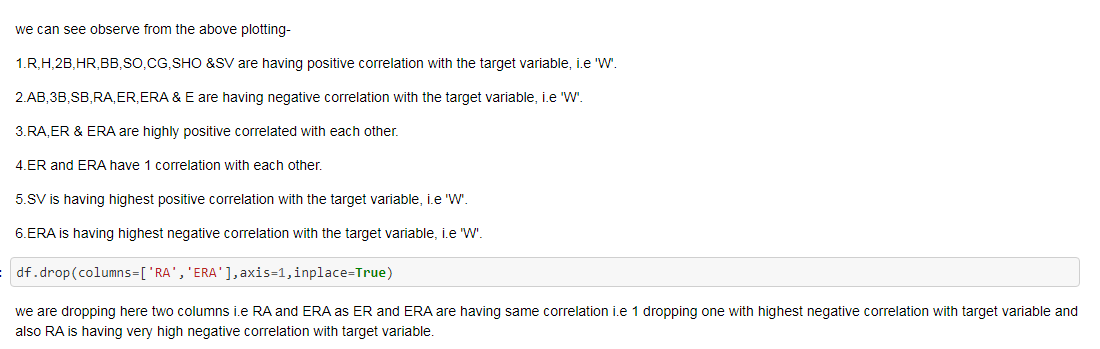
6)Checking the statistical description of the dataset. This report’s helps us to decide which transformation will improve the model’s accuracy,outlier and range of particular column.



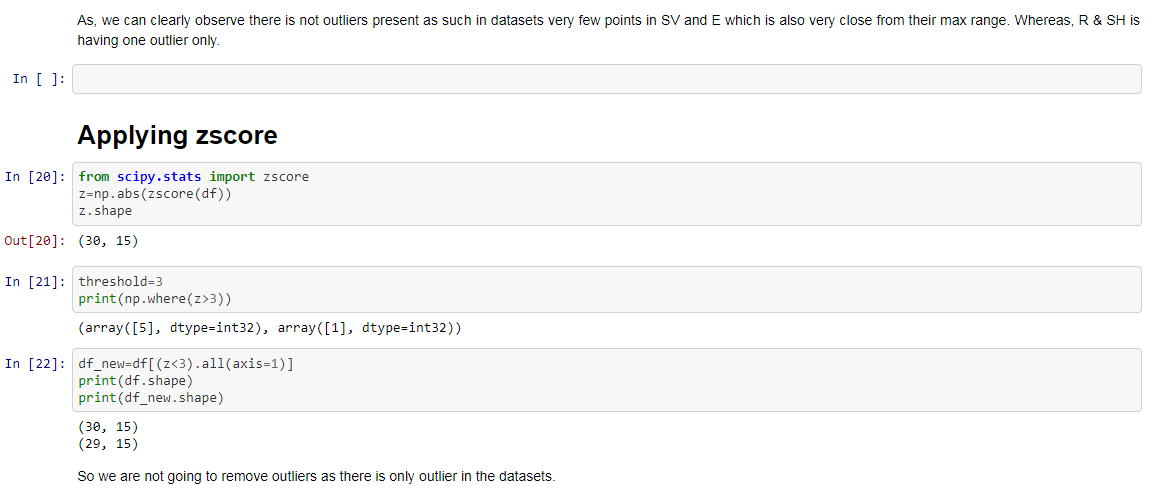
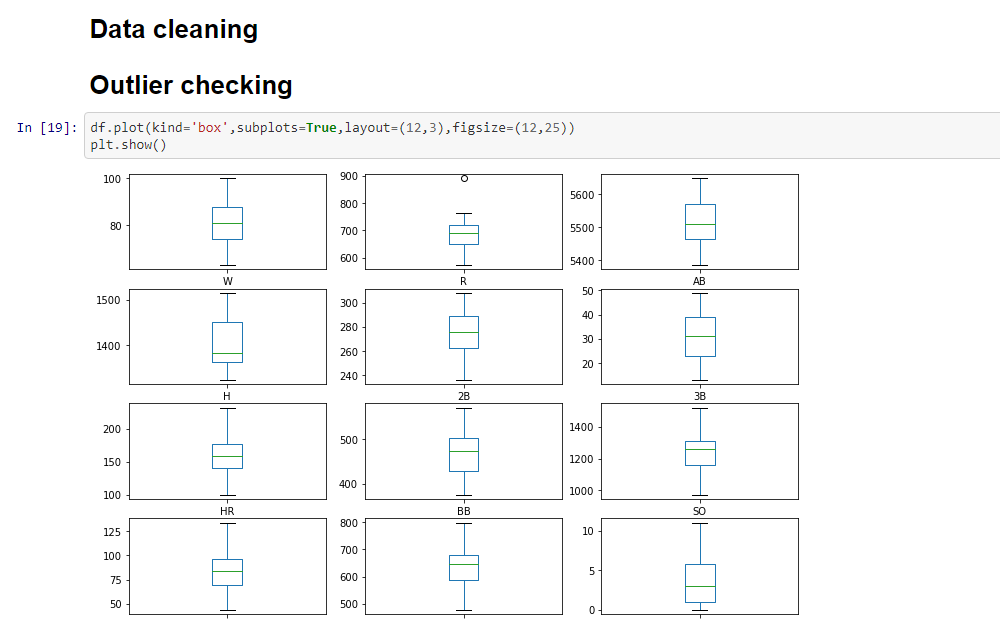


7) Now we are checking the correlation by using heatmap from which we can figure out some columns that we need to drop on the basis of correlation factor.



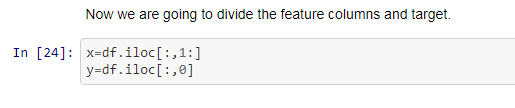


8)We’ll go for data cleaning, firstly going to check outliers through plots and then by applying z-score(factor3) on the dataset.

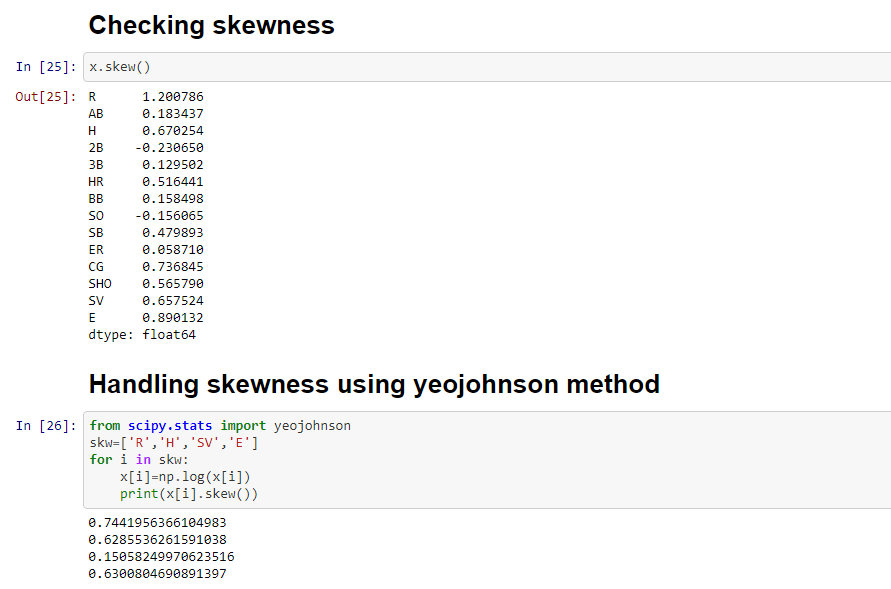


No outliers as such present in the dataset.Very few that can be neglected.

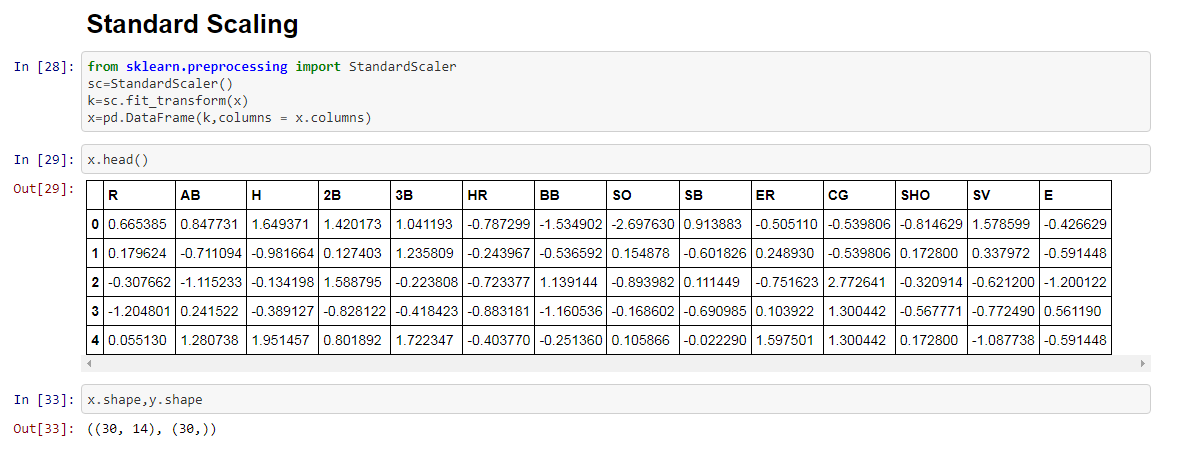
9) Let’s go for dividing the feature and target, separating the input and output variable.



10)While checking skewness, few columns are skewed so,handling skewness using yeo-johnsoon method to minimise the skewness present in the datatset.

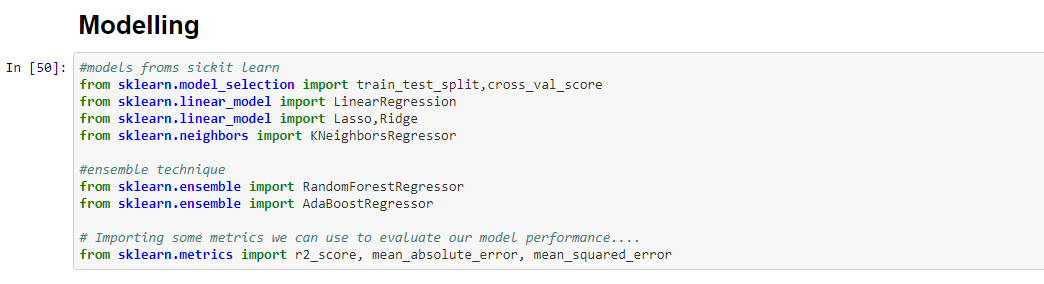


11)Standardizing the input dataset(Standard scaler makes the mean of the distribution 0).

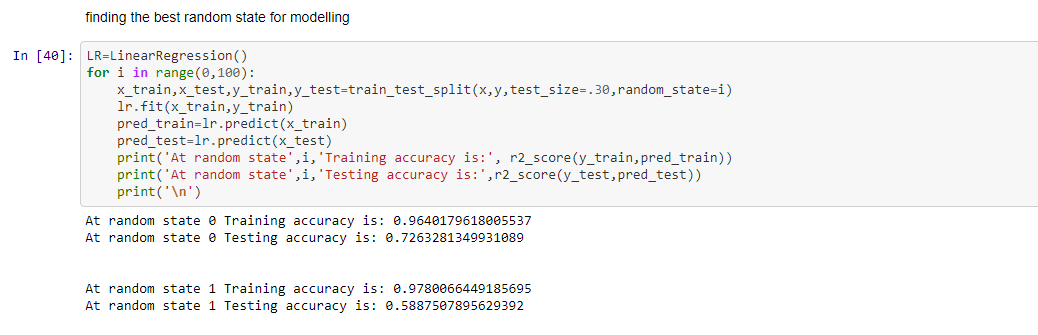


**Model/s Development and Evaluation**

1) Loading all the libraries required for modeling as well as evaluation metrics.



2) Find the best random state for modelling.



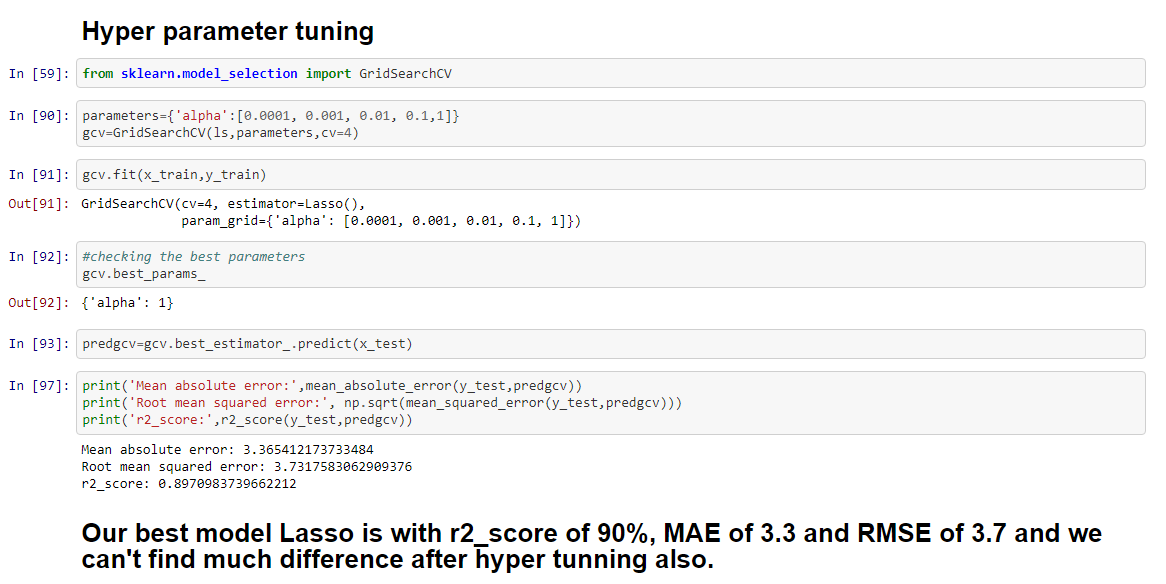
We are getting best result at 36 & 39. So our random state will be 36 or 39.

3)Basic models were chosen to check the best performing model and also calculating the evaluation metrics score(mae,rmse,r2 score & accuracy score).



4) After performing modelling,Lasso is giving the best results with testing r2 score- 90% , training r2 score- 88%, RMSE- 3.4 and MAE-3.7

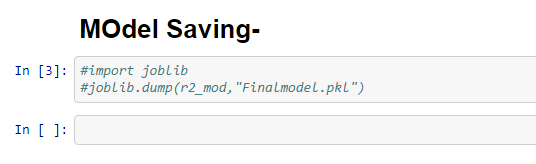
5)Therefore , best model I have chosen for hyper tuning i.e Lasso is having highest accuracy score,r2\_score & cross\_val\_score, So, I am choosing this for hyper tuning.



**CONCLUSION**

From the r2\_score observations, mae & rmse this model is performing very well with the data’s. Note- there is not so variations in evaluation metrics after hyper tuning the model.

**Model Saving-**

****

* Limitations of this work and Scope for Future Work

As, off now, I don’t find any such limitation in this project but apart from that, more learning , alayzing & exploration is needed in Data science field that I am continuing. Because, **The Journey to Data Science is a Marathon, not a Sprint.**

For more information, please visit:-