**BASEBALL CASE STUDY ARTICLE**



**Project Description:-**

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 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.

**Input Features:-**

**1.** W- This indicates the number of wins credited to a pitcher: number of games where pitcher was pitching while their team took the lead and went on to win, also the starter needs to pitch at least 5 innings of work.

**2.** R- This indicates Runs scored. A run is scored when a player advances around first, second and third base and returns safely to home plate, touching the bases in that order, before three outs are recorded and all obligations to reach base safely on batted balls are met or assured: number of times a player crosses home plate.

**3.** AB- This means At bat or time at bat. It is a batter’s turn batting against a pitcher: plate appearances, not including bases on balls, being hit by pitch, sacrifices, interference, or obstruction.

**4.** H- This means Hit. It’s also called a “Base hit”, is credited to a batter when the batter safely reaches or passes first base after hitting the ball into fair territory, without the benefit of either an error or a fielder’s choice: reaching base because of a batted, fair ball without error by the defense.

**5.** 2B- This means the act of a batter striking the pitched ball and safely reaching second base without being called out by the umpire, without the benefit of a fielder’s misplay (see error) or another runner being put out on a fielder’s choice. A double is a type of hit (the others being the single, triple and home run) and is sometimes called a “two-bagger” or “two-base hit”: hits on which the batter reaches second base safely without the contribution of a fielding error.

**6.** 3B- This means a Triple. It is the act of a batter safely reaching third base after hitting the ball, with neither the benefit of a fielder’s misplay nor another runner being put out on a fielder’s choice. A Triple is sometimes called a ”three-bagger” or “three base-hit”: hits on which the batter reaches third base safely without the contribution of a fielder’s error.

**7.** HR- This means Home runs. It is scored when the ball is hit in such a way that the batter is able to circle the bases and reach home plate safely in one play without any errors being committed by the defensive team. A home run is usually achieved by hitting the ball over the outfield fence between the foul poles (or hitting either foul pole) without the ball touching the field: hits on which the batter successfully touched all four bases, without the contribution of a fielding error.

**8.** BB- This means Base on balls(also called a “walk”). It occurs in baseball when a batter receives four pitches that the umpire call balls, and is in turn awarded first base without the possibility of being called out: hitter not swinging at four pitches called out of the strike zone and awarded first base.

**9.** SO- Also denoted as “K” means strikeout. It occurs when a batter accumulates three strikes during a time at bat. It usually means the batter is OUT: number of batters who received strike three.

**10.** SB- This means stolen base. It occurs when a runner advances to a base to which they are not entitled and the official scorer rules that the advance should be credited to the action of runner: number of bases advanced by the runner while the ball is in the possession of defense.

**11.** RA- This means run average. It refers to measures of the rate at which runs are allowed or scored.

**# Problem Definition:-**

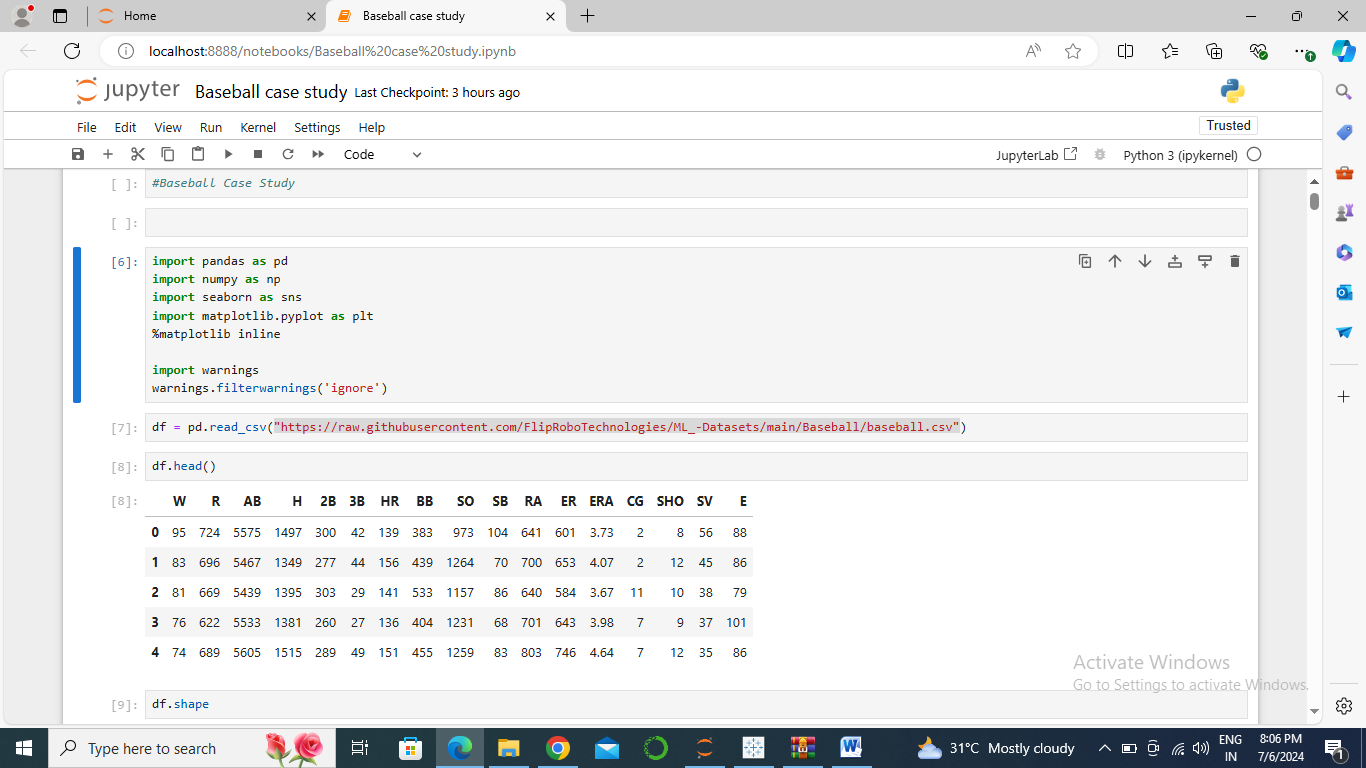
It is a case study of very famous sport i.e. Baseball. In this we have used a data from 2014 Major League Baseball seasons in order to develop an algorithm that predicts the number of wins for a given team in the 2015 season based on several different indicators of success. It will help the management of the team to make changes in the team as per the predicted results.

**# Data Analysis and EDA concluding remarks:-**

Importing raw data for our model.

<https://raw.githubusercontent.com/FlipRoboTechnologies/ML_-Datasets/main/Baseball/baseball.csv>.

Now Import important libraries:-



In the above data, we have following attributes. In this we have to predict Wins (W) based on 16 different attributes:

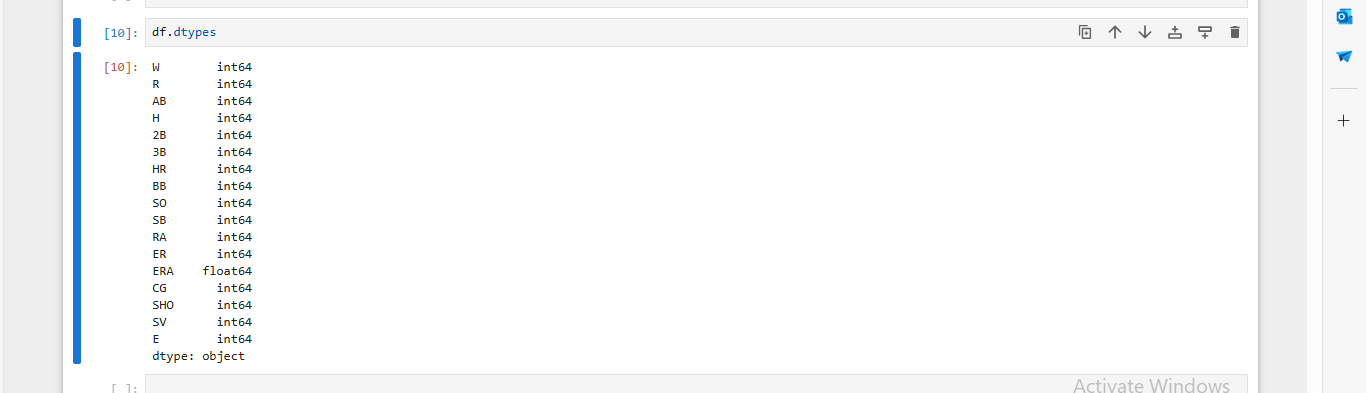
Here our output is W i.e. win and rest are inputs. R:Runs AB:At Bats H:Hits 2B:Doubles 3B:Triples HR:Homeruns BB:Walks SO:Strikeouts SB:Stolen Bases RA:Runs Allowed ER:Earned Runs ERA:Earned Run Average (ERA) CG:Shutouts SV:Saves, SV:Complete Games E:Errors W:Win

Data Frame Shape



Dataset has 30 rows and 17 columns.

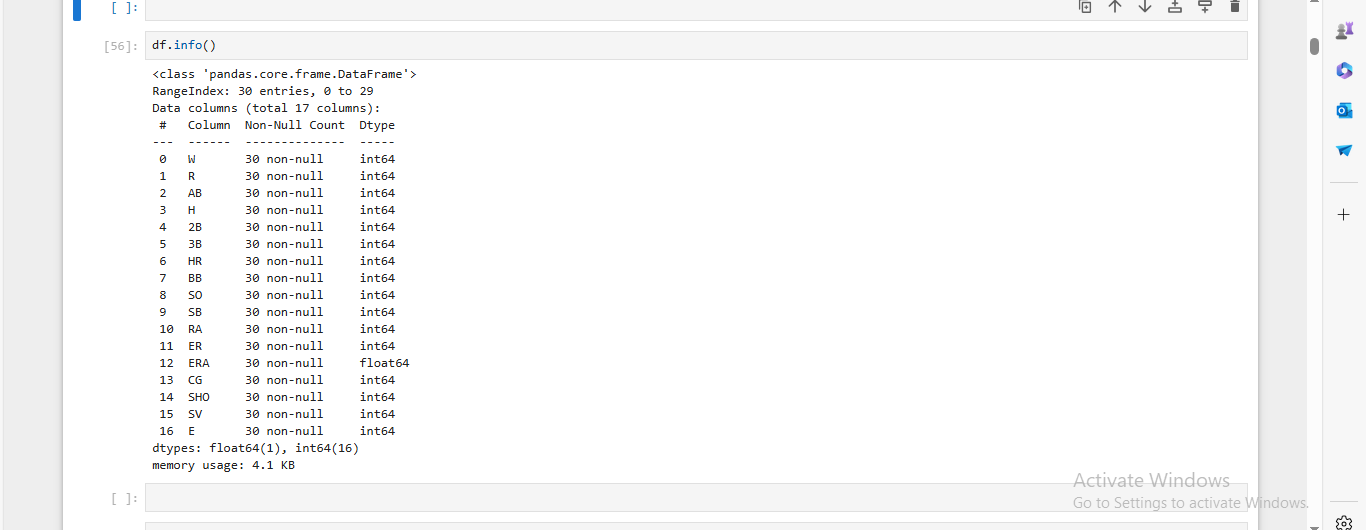
Now, checking for data types



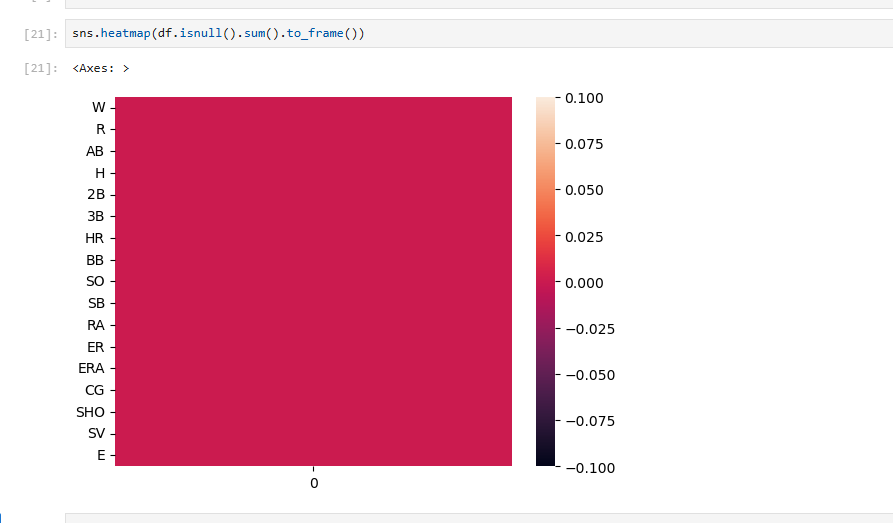
In the above inputs and target, have integer and float values. So, we can proceed easily as we don’t have any object data type.

Now, Checking for Null value in data.

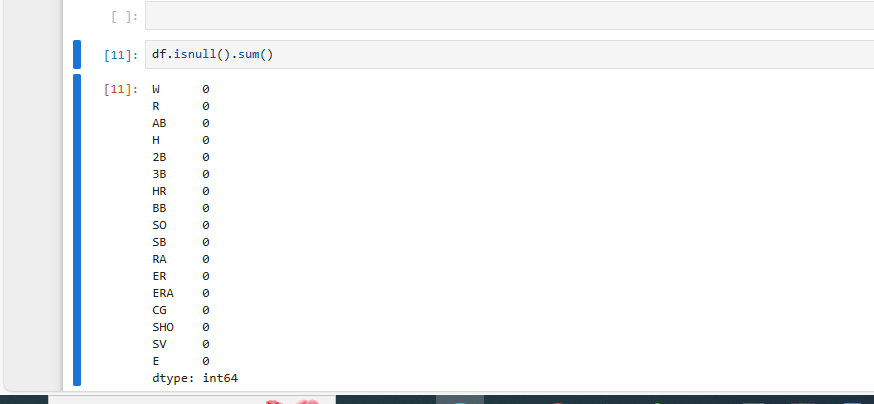
First using info logic:



Now using heat map for visualizing null values:

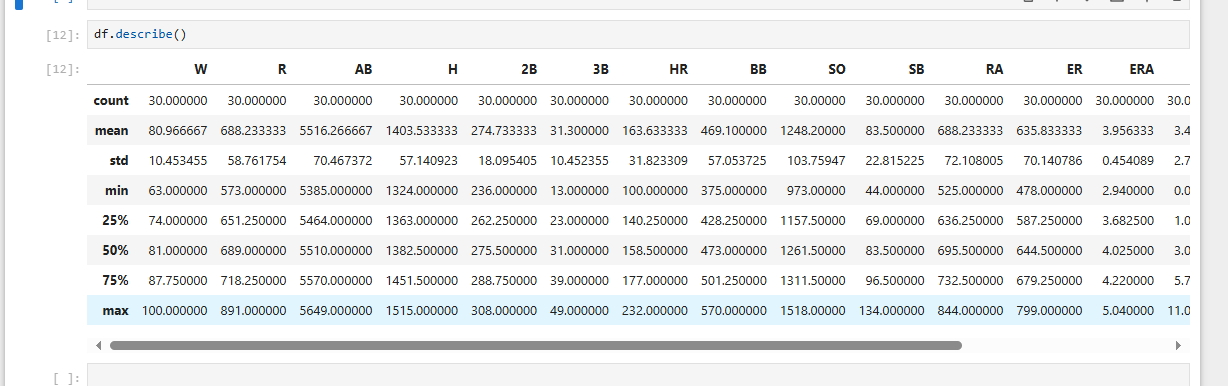


We have seen null values using heat map. So, no null values are present in the data. We will check it using sum also.



So, finally we can conclude that no null values are present.

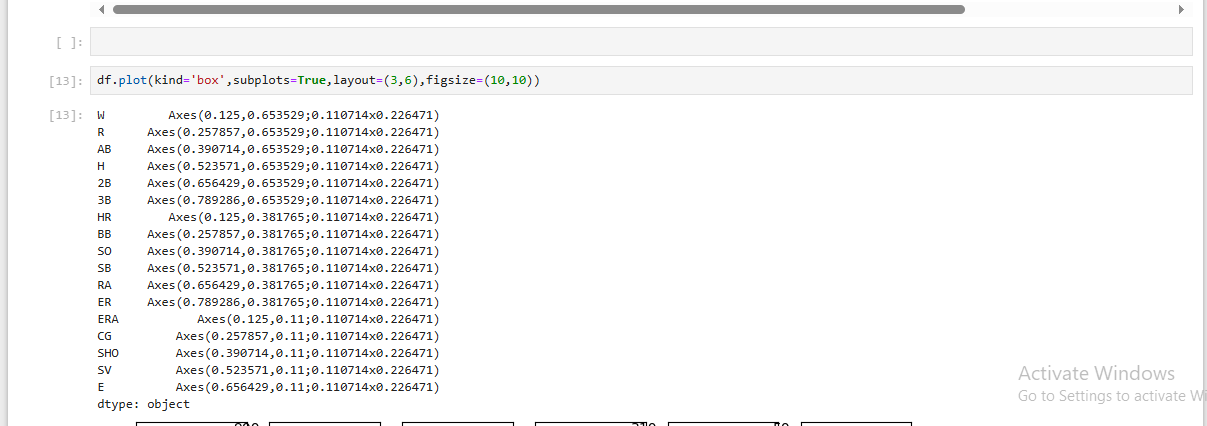
**Description of Data**

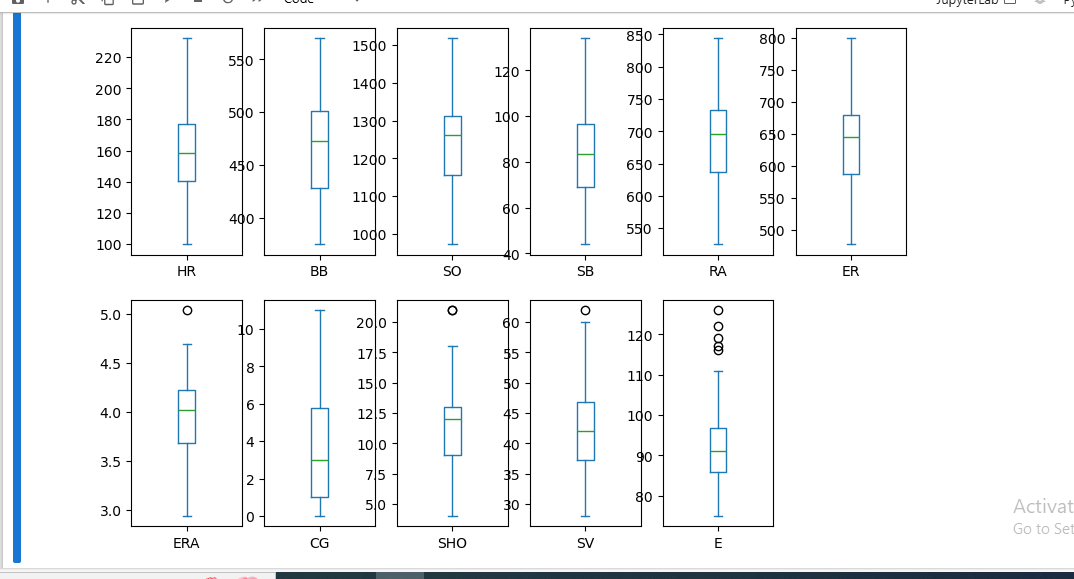


As we can see from the above data, that count is equal in all columns, hence no missing data is present. In columns AB,H,HR and E mean is higher than median. It means data is right skewed in these columns. In columns HR,BB,SO,SB,RA,ER there is large gap between 75th and max, there are chances that some outliers may present. High Standard Deviation (SD) in column R,AB,H,BB,SO,RA,ER. It means data is spread.

**Outliers Visualization:**

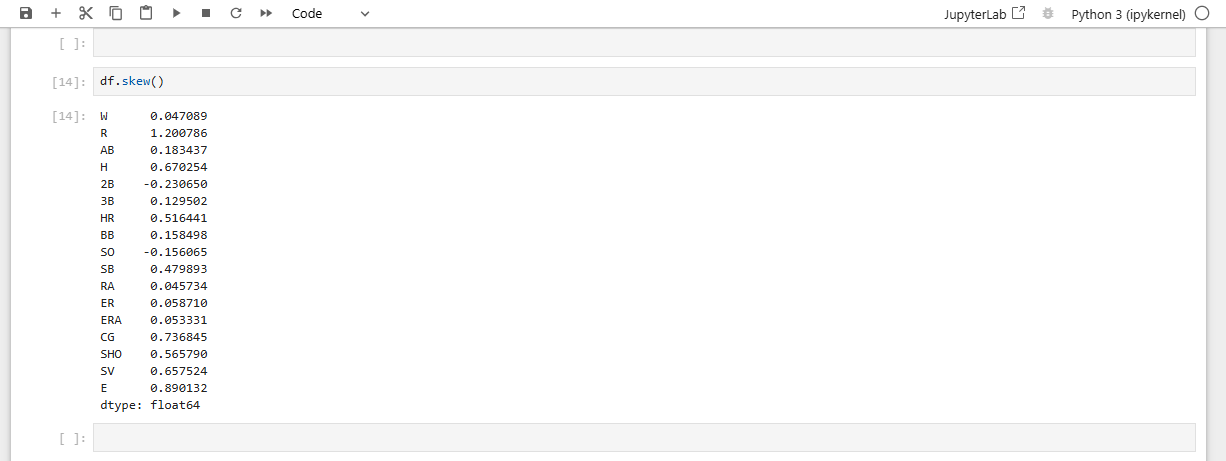
We will use boxplot type of visualization to see the outliers.

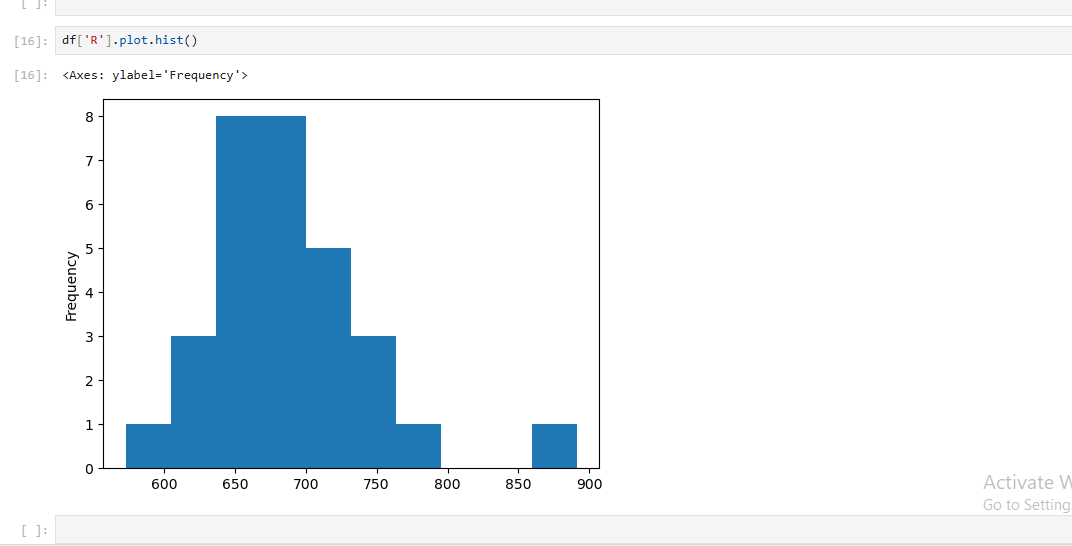


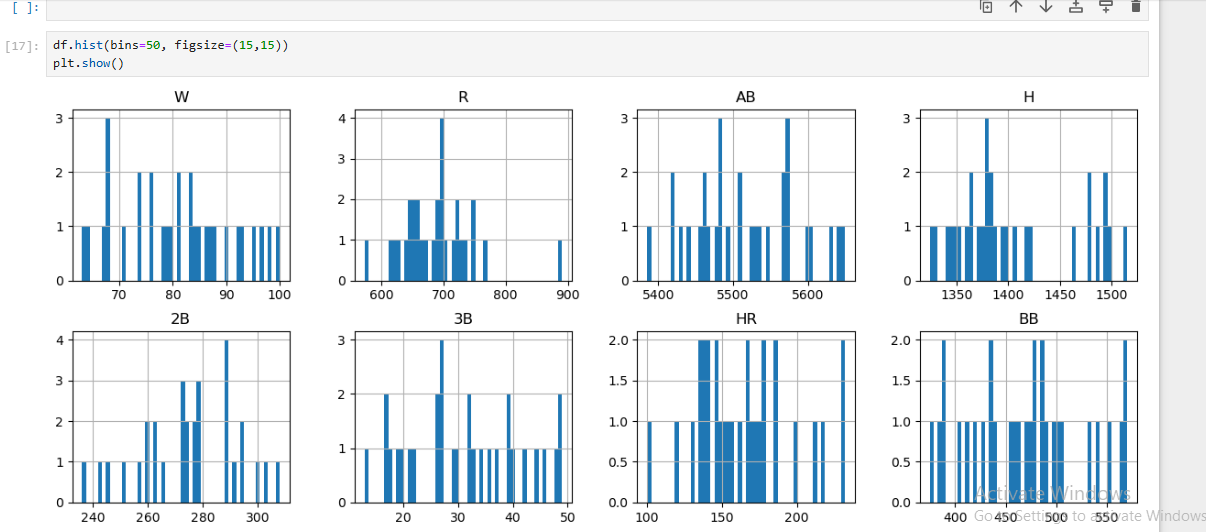


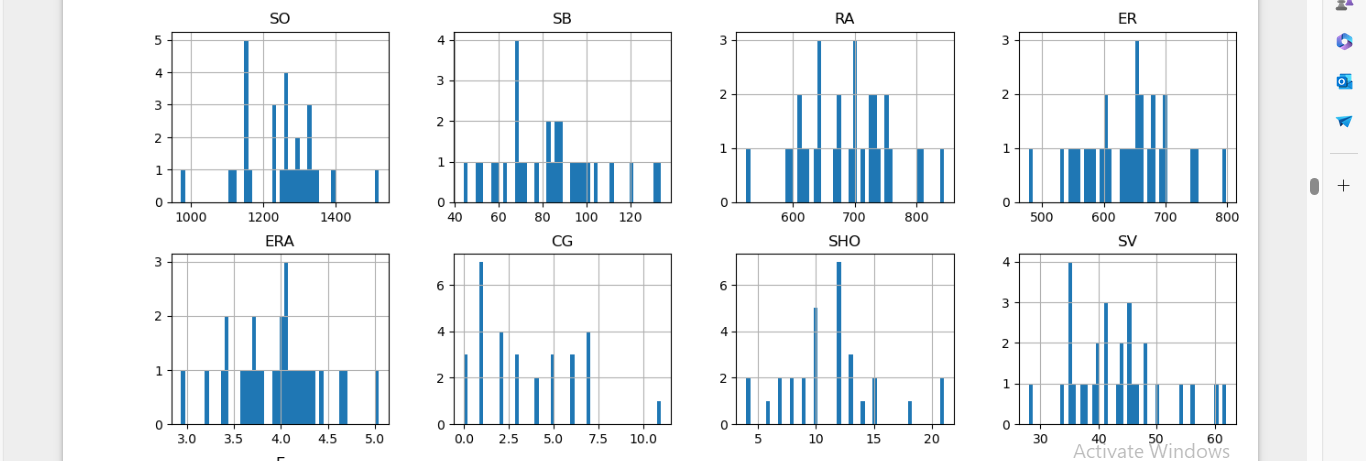
we can see from above Boxplots that some outliers are present in the data.

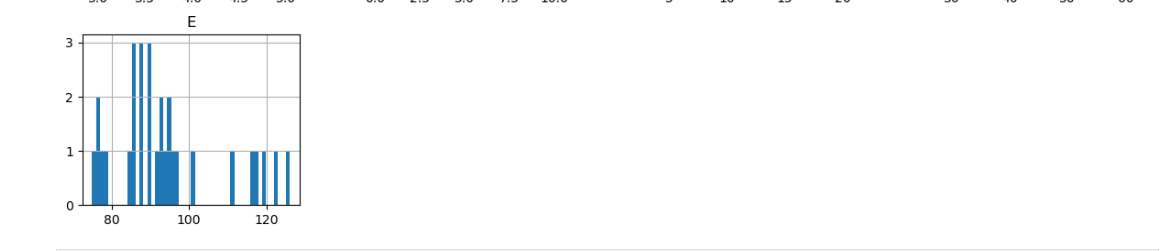
**Skewness Visualization:**



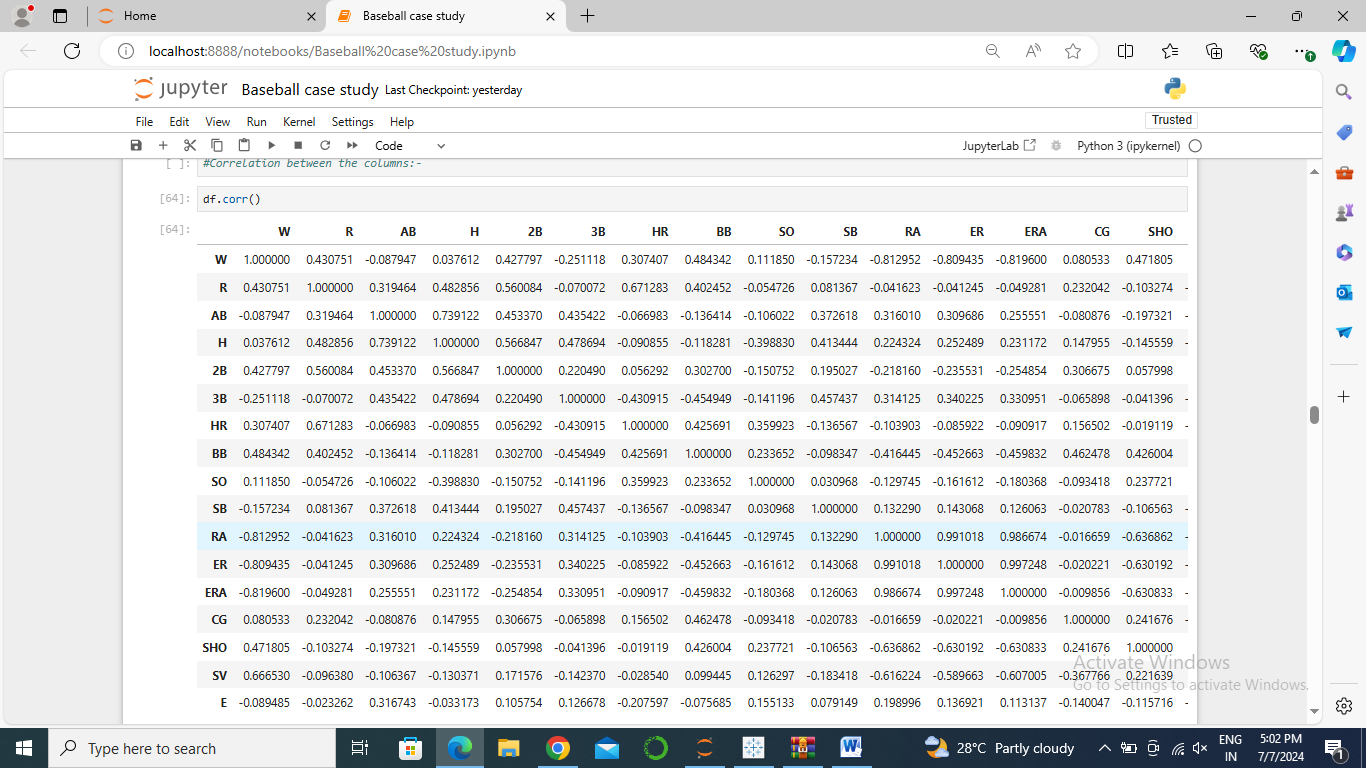


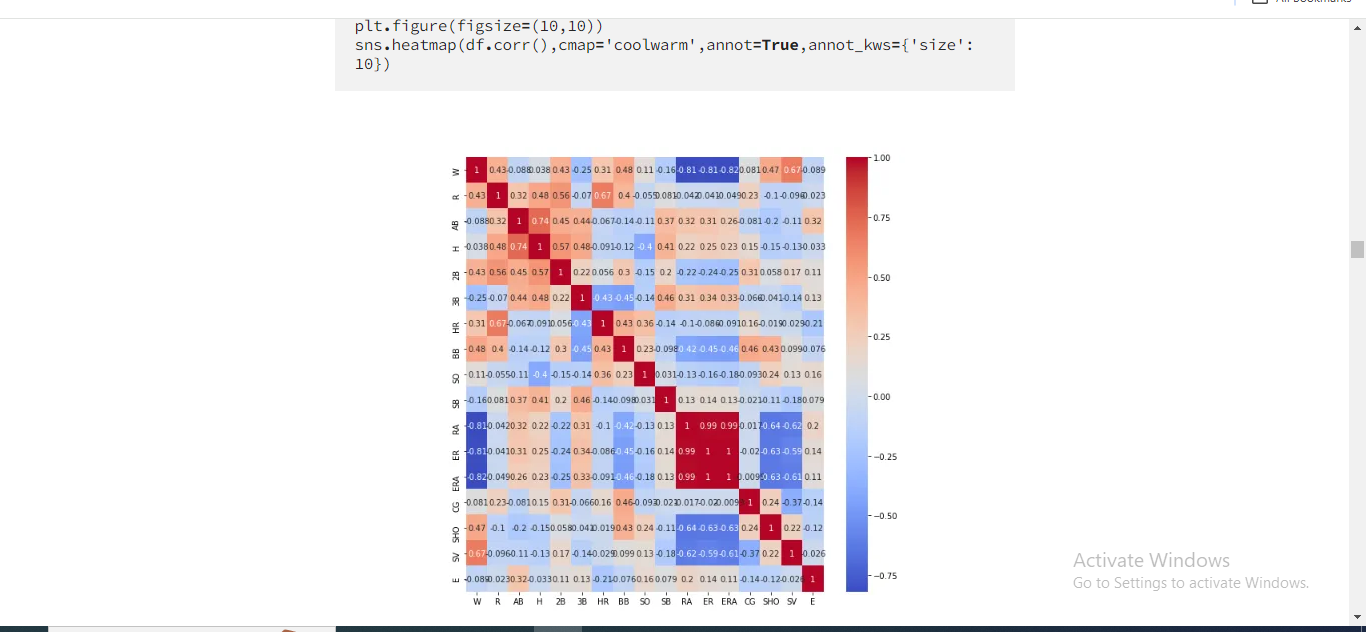






Correlation between the column:-





From above heat map, we can see that output is in good correlation with many inputs but for some inputs it is not in good correlation. The inputs such as Shutouts, Walks, runs, Saves, Shutouts, home runs and doubles are positively correlated. Earned Runs, Stolen Bases and Runs Allowed are negatively correlated. The dataset has lot of randomness present.

**# Pre-Processing Pipeline:-**

As we have already checked that all the data present the dataset is of continuous or integer nature, so we can proceed to the pre processing of the data.

Pre-Processing of data consists of steps below steps:-

**Multicollinearity using VIF**- When there is a correlation and linear relationship b/w one or more independent inputs. It creates a trouble in the multiple regression as all the inputs are persuading each other. So, we will remove it using Variation Inflation Factor (VIF).

**Removing Outliers using Z-Score:-**

**Outliers**These are the unusual values in the data set, these are far from the mean. They can cause tests to either miss important findings or mislead real results. Here we will use Z-score method to remove outliers.

**Z-score method**- it changes the dataset into z-score and then verify if absolute of z-score is larger than 3, then it will remove all those data having value more than 3. It tries to normalized the data as in a normal distribution, data lies within 3 standard deviations.

**Standardization-**It is a method to transform the dataset into a normal distribution. There are several methods to do it.

1. Log-transformation- It can be used if data is highly right skewed.
2. Square root transformation- It can be used if data is little bit right-skewed
3. Cube root transformation
4. Reciprocal transformation
5. Box-cox transformation

**Normalization- It**is a mathematical technique which scales the data into the range which is required for many ML algorithms. It tries to minimize the difference b/w the low and high value, which helps in better estimation of the desired output.

Here we will use Standard scaling technique to convert the data set as mean=0 and standard deviation =1.

The whole step-by-step process is packed in a set that can be created in function, class, or using pipeline from sklearn library.

preprocessing helps in reducing the noise or randomness from data. It consists of the following elements

1. **Data Cleaning:-**removing the Null value
2. **Data Integration:-** integrates data from a multitude of sources into a single data warehouse.
3. **Data Transformation:-**Transforming data such as standardization, normalization, encoding, etc
4. **Data Reduction:-**removing redundant data.

Enable function step by step prepossessing functions are called to make a pipeline first function segregates numerical and categorical data, the second function plots the numerical data for the analysis, The third function standardizes the numerical data, the Forth function normalizes the numerical data, the fifth function to remove the outlier from the numerical data set and the last function is used for dimension reduction This function is the pipeline for preprocessing and visualization.

**# Building Machine Learning Model:-**

As output is of continuous type, so we will use the ML models as per the output values.

Firstly we will use linear regression model and see the outcome.

Machine learning is as good as preprocessing. Now let’s take a look into some algorithms of machine learning for classification.

1. Linear regression — it is an algorithm that minimizes the loss function i.e summation of square error.
2. Decision Tree — It is a tree-based algorithm that uses the gini or entropy method to generate the tree.
3. Random forest- It is a multiple trees algorithm that uses bootstrap sampling and bagging for classification.
4. Extreme Gradient boosting — it is a boosting algorithm that uses a multi-decision tree.
5. Support vector machine- it constructs Hyperplane and marginal plane and distance between them is optimized.

There are multiple machine learning algorithms to select from. Based on the quite best method to select the algorithm having the minimum difference between predicted score and cross-validation score.

**Lasso Model:**Lasso (least absolute shrinkage and selection operator) is a regression analysis method which accomplishes both regularization and variable selection to increase the prediction correctness and interpretability of the resulting statistical model.

**Ridge Regression Model:**It is a method used for evaluating multiple regression where the data undergoes multi-col linearity. It evades over fitting, it works just like the linear regression, but it just add up an additional term (α) which assists in the reduce the over fitting.

**Cross Validation:**

We will use the Cross Validation technique to find the scores and match it with the already calculated scores of all above models. It is to see that our model is working fine, it not over fitting or under fitting.

**Hyper parameter tuning:**

Her we will select our 2 best model i.e. Lasso and Ridge and try to improve its outcome.

Now after selecting the best algorithm now we should do Hyper parameter tuning for extreme gradient boosting.

1. Random search cv:- It randomly Search the grid for the best scoring. Processing time is less but the accuracy is not that great
2. Grid search cv:- it searches the grid one by one for the best scoring. Processing time is more but the accuracy is great

**# Concluding Remarks:-**

We have started our baseball case study project by importing several libraries and imported the dataset from GitHub. Studying the various important details like what is the problem type and observing the data types of all the columns i.e. how many columns having integer, float and object type values.

As per statistic observations we found that all our columns are of numeric type. We have also noticed that there were outliers and skewness present in the data. We have applied standard scalar technique to scale the variables. We have also noticed that there were no Null Value (NaN) present in our data. During this procedure we have used matplotlib as well as seaborn for various visualizations and heatmaps. Subsequently we have started training of distinct machine learning models. We have utilized cross validation on all the regression models and then try to tune the model by using Hyperparameter Tuning.

We can also use other ML models such as Decision Tree Regressor, Random Forest Regressor, Support vector machine and Extreme Gradient boosting to predict the result.

This is the first part of the ‘Baseball’ project. I imported important libraries in Jupyter Notebook, cleaned it, explored it to visualize, engineered some new features. Then, I applied K-Means clustering model, linear regression model qazx1 and ridge regression model to predict wins. Finally, I used mean absolute error to evaluate these models.

I will use classification models to make some other interesting predictions in the second part of project.

As discussed in EDA and storytelling the randomness of the dataset should be reduced by reducing the number of variables. Which was done by R-square dimension reduction.

In this project I had used machine learning models to predict Major-League Baseball team wins per season based on Lahman’s Baseball data.

Agenda of using it:-

* Importing the data
* Cleaning and preparing the data
* Exploring and visualizing the data
* Adding new features
* Adding one more feature using K-means clustering
* Splitting dataset into train\_set and test\_set
* Selecting error metric:mean absolute error vs mean squared error
* Linear regression model
* Ridge regression model
* Conclusion