**HR Analytics Project- Understanding the Attrition in HR**



**Switching** **to** **new** **company(Attrition)**

**or text here**

**Problem Statement**

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the compnaies for their existing employees as well.

The objective of the model to increase the effectiveness of their employees and reduce the time and money investing in employees.

**HR Analytics:**

Human resource analytics (HR analytics) is an area in the field of analtyics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analtyics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees

How does Attrition affect companies? and how does HR Analytics help in analysing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**Attrition affecting Companies**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

**Importing the Libraries**

import pandas as pd

import numpy as np

import seaborn as sns

from scipy import stats

import warnings

warnings.filterwarnings('ignore')

from matplotlib import pyplot as plt

from scipy.stats import zscore

#data preprocessing

from sklearn import preprocessing

from sklearn.preprocessing import MinMaxScaler

#Over Sampling the data using SMOTE

from imblearn.over\_sampling import SMOTE

#modelling

from sklearn.metrics import classification\_report, confusion\_matrix, accuracy\_score

from sklearn.linear\_model import LogisticRegression

from sklearn .ensemble import RandomForestClassifier

from sklearn.ensemble import BaggingClassifier

from sklearn.model\_selection import train\_test\_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import roc\_curve

from sklearn.metrics import roc\_auc\_score

from sklearn.model\_selection import cross\_val\_score

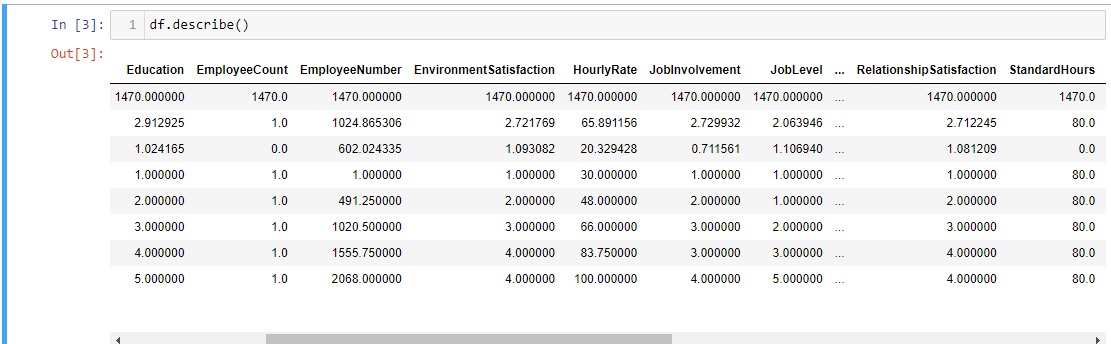
from matplotlib import pyplot

from sklearn.svm import SVC

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.tree import DecisionTreeClassifier

**EXPLORATORY DATA ANALYSIS**

****

Employee's average number of years at company is 7.

Mean is not equal to median stating that the data is not normally distributed. Most normally distributes column is Daily rate where mean is almost equal to median



**Numeric variables:**

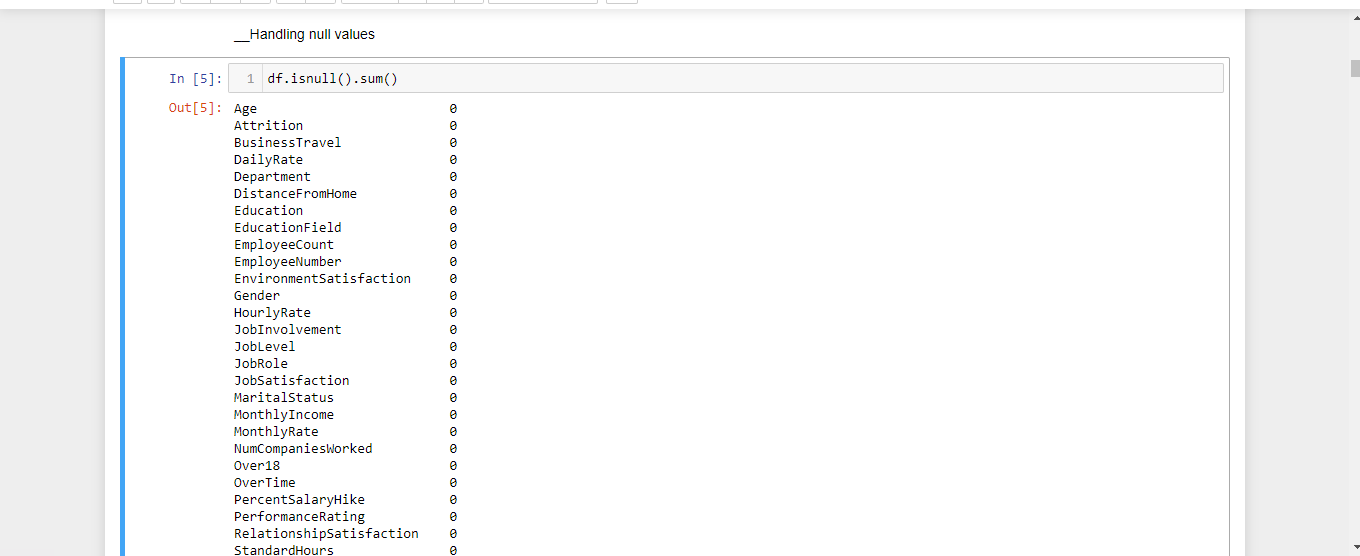
* Related to personal information: age, distance\_from\_home, employee\_number
* Related to income: hourly\_rate, daily\_rate, monthly\_rate, monthly\_income, percent\_salary\_hike

Related to duration in company: years\_at\_company, years\_in\_current\_role, years\_since\_last\_promotion, years\_with\_curr\_manager, total\_working\_years

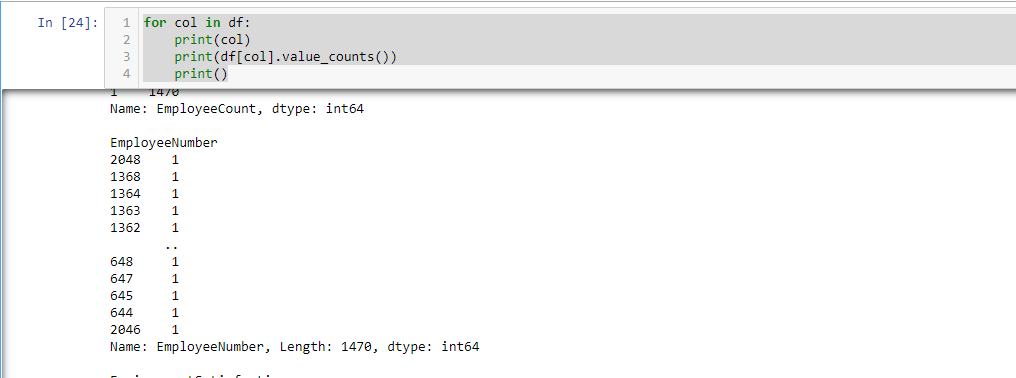
num\_companies\_worked,standard\_hourstraining\_times\_last\_year, employee\_count

**Categorical variables:**

* Binary variables: attrition(target variable), gender, over18, over\_time
* Nominal variables: department, education\_field, job\_role, marital\_status
* Ordinal variables:
* Ordinal regarding satisfaction and performance:environment\_satisfaction,job\_satisfaction, relationship\_satisfaction,work\_life\_balance,job\_involvement,performance\_rating
* Other ordinal: business\_travel, education, job\_level, stock\_option\_level

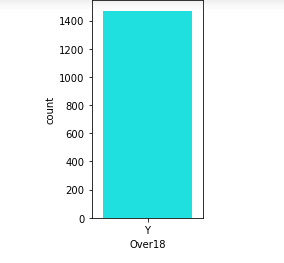
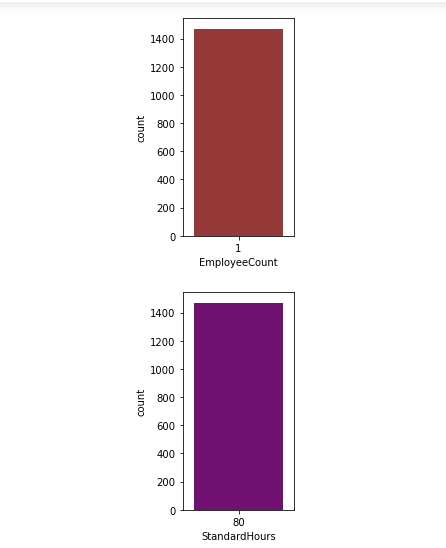


This dataset has no null values



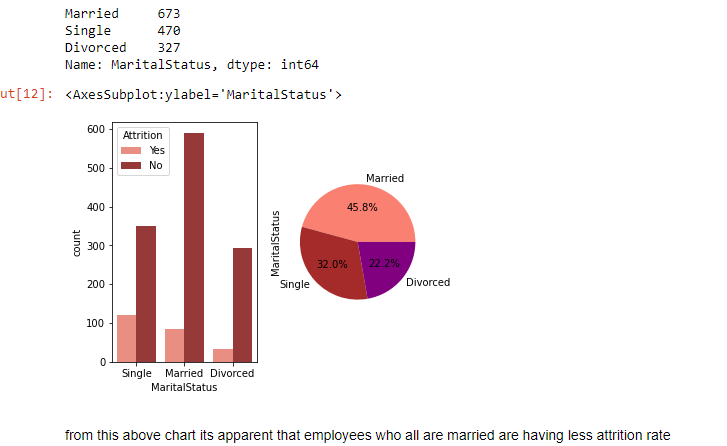
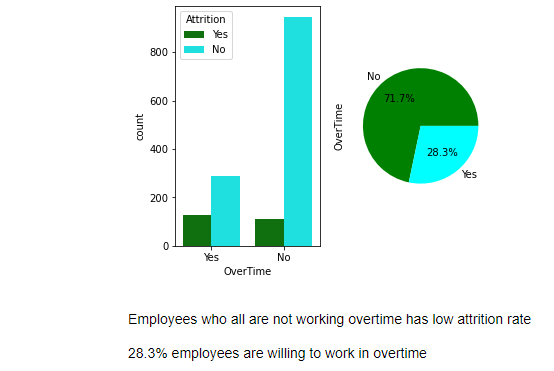
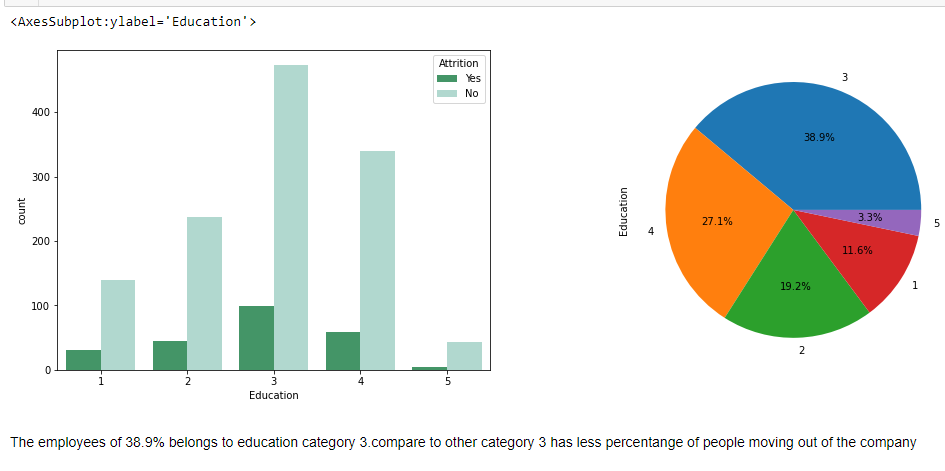
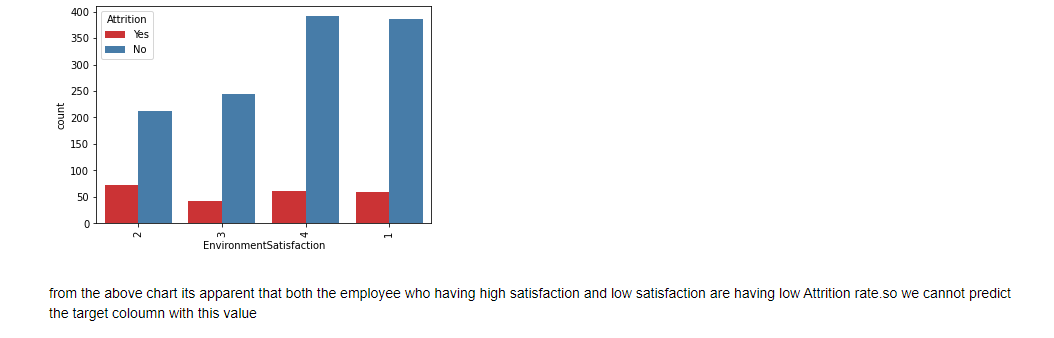
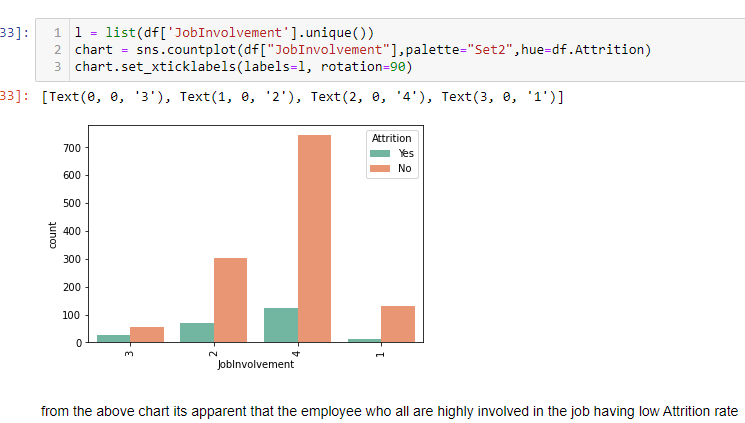
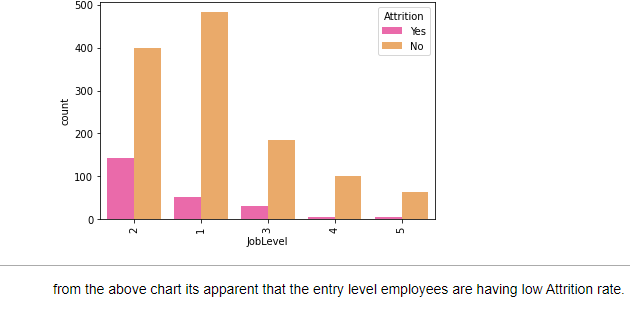
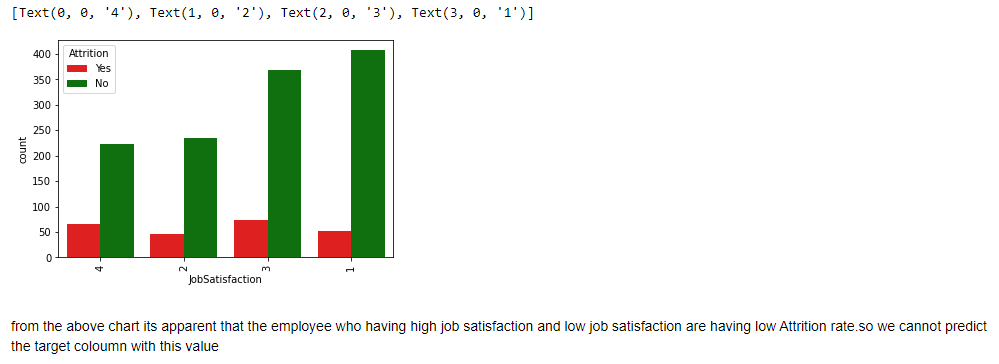
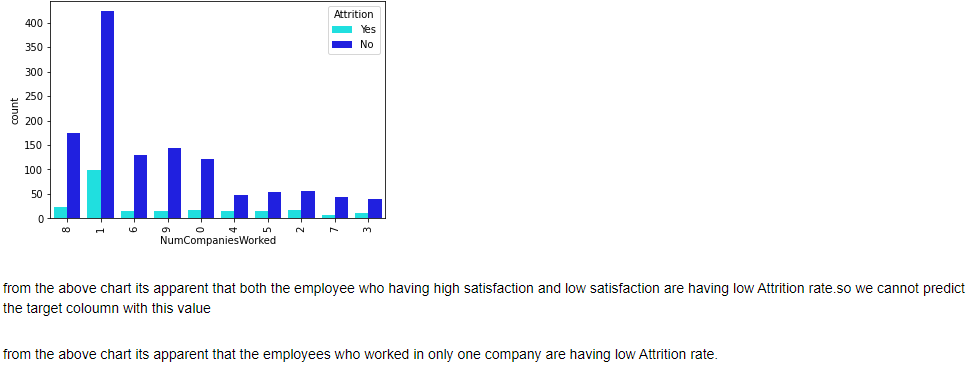
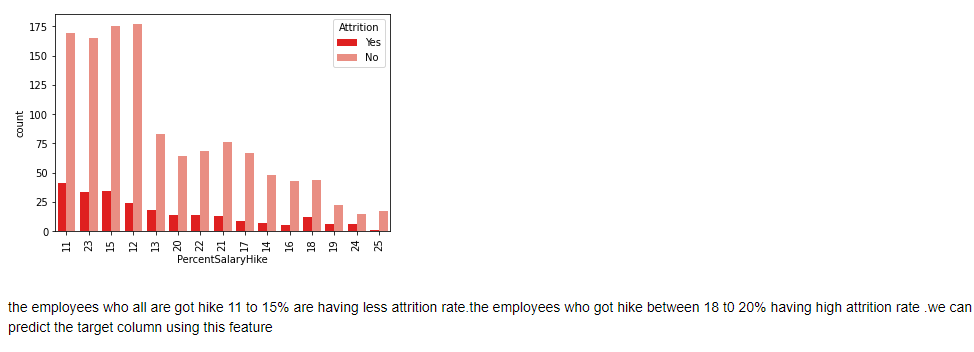
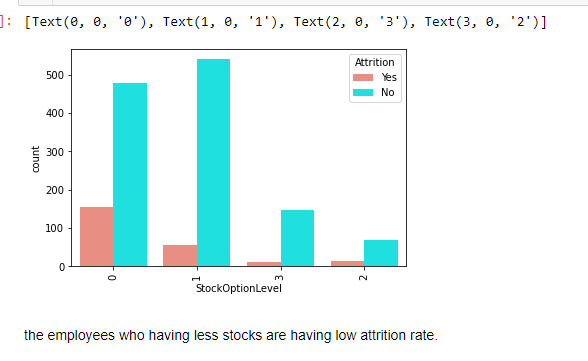
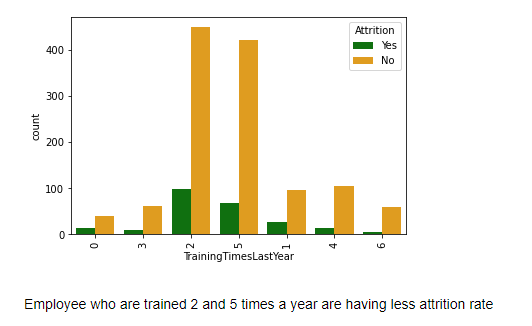
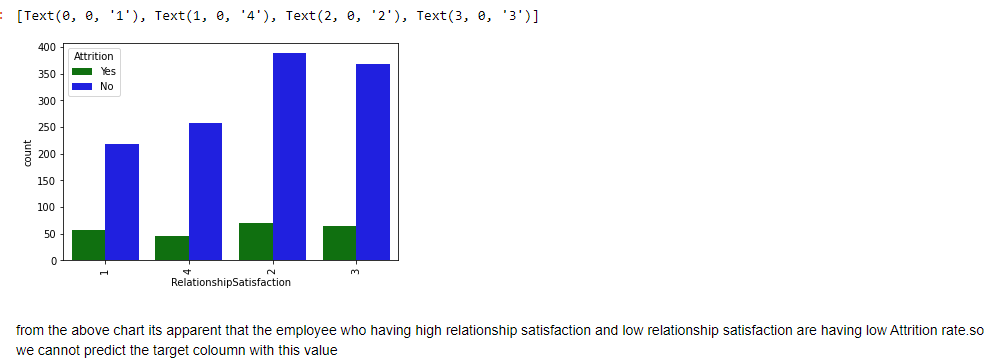
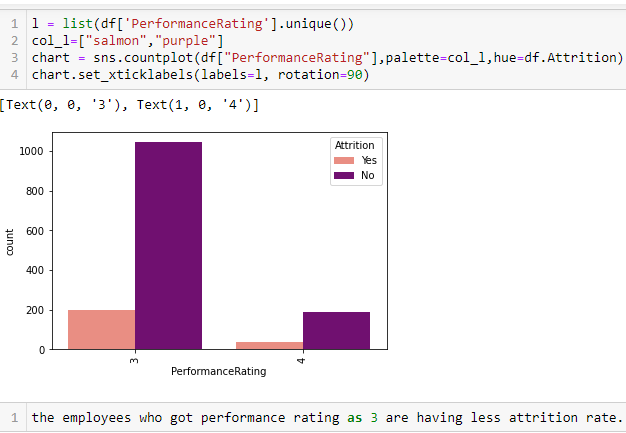
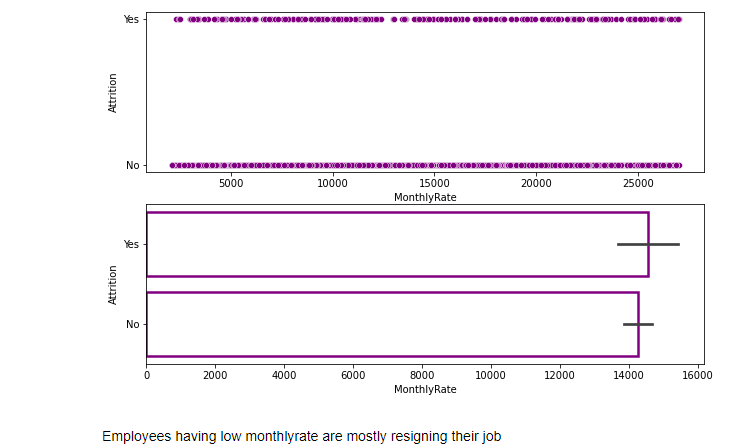
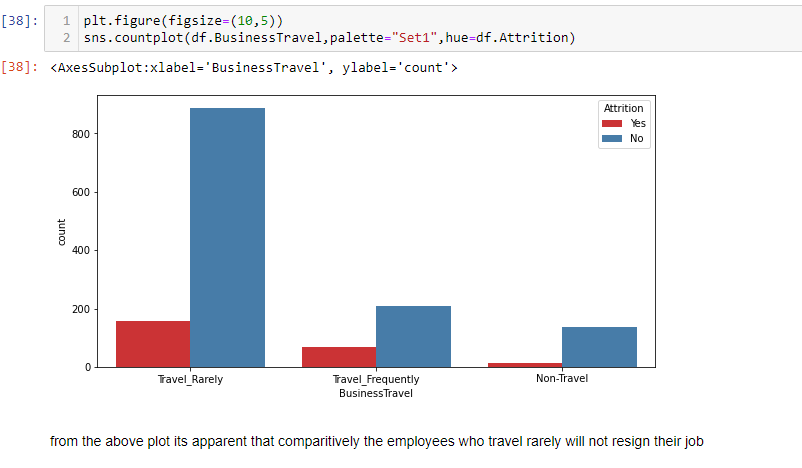
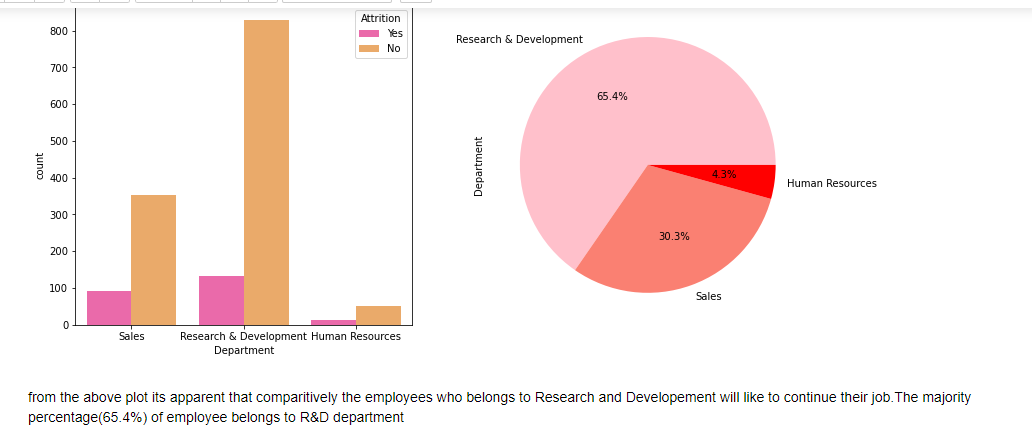
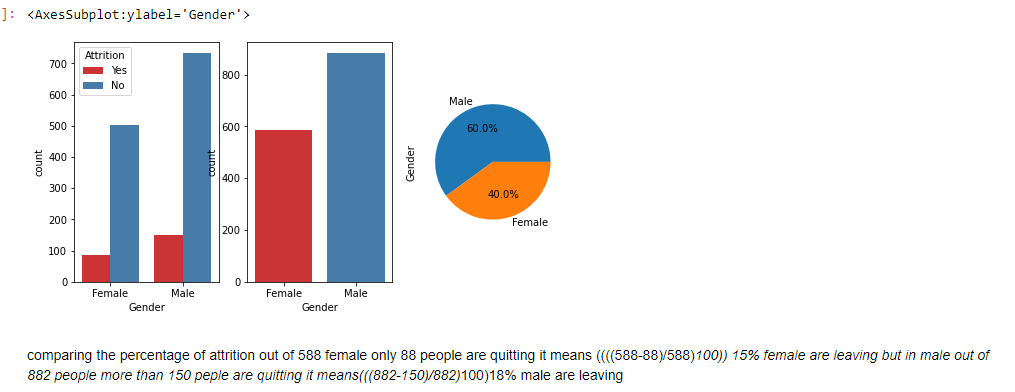
Displaying value count of unique value in each feature. To identify column having single unique value

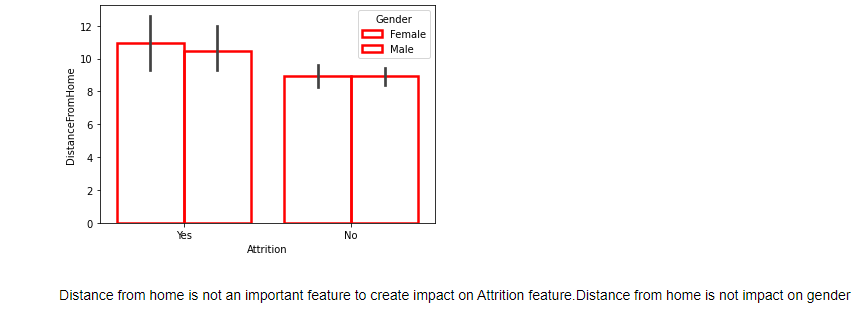
**UNI VARIATE ANALYSIS**

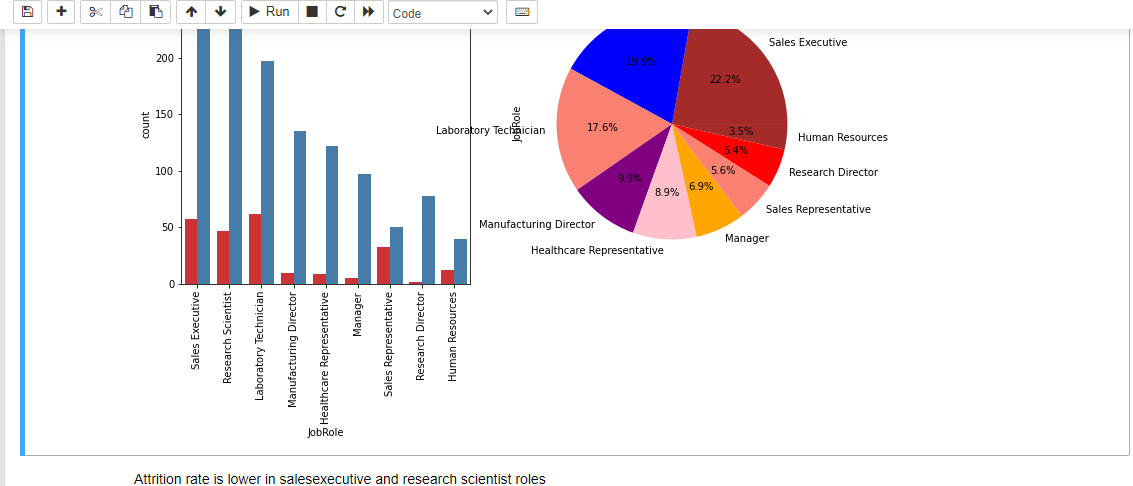


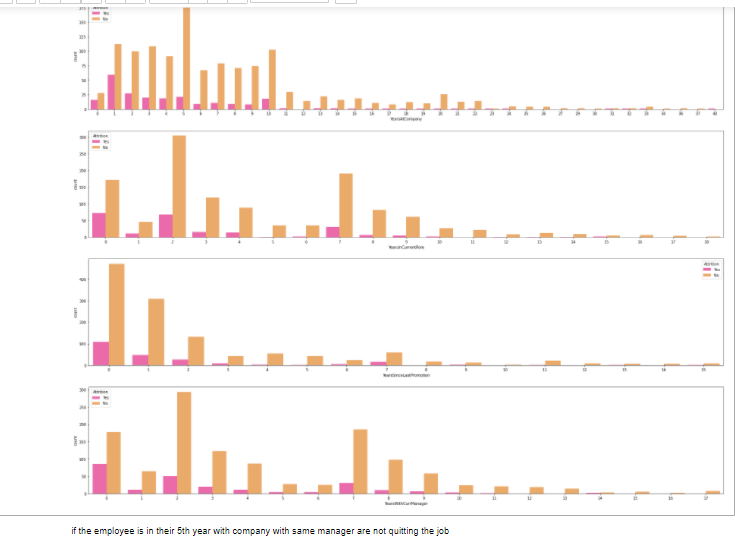
The features standard hours, over18 and employee count has only one value so it won’t create any impact on the target feature Attrition.

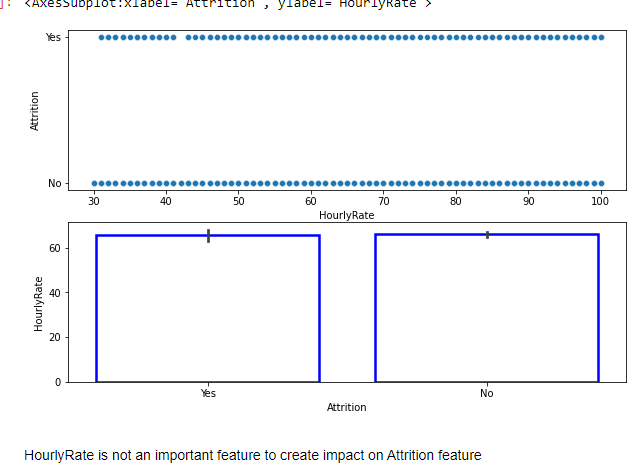
All the three columns having single value so I'm going to dropping it from the given dataset

**BI VARIATE ANALYSIS (**Categorical columns vs Target**)** 







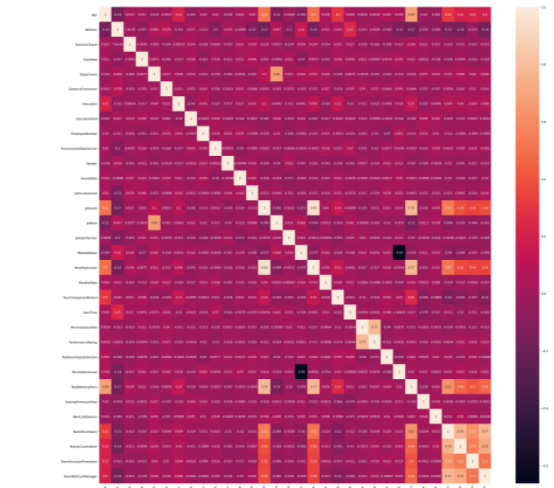
  
**EDA CONCLUSION:**

Employees who belongs to below category having less attritionon rate

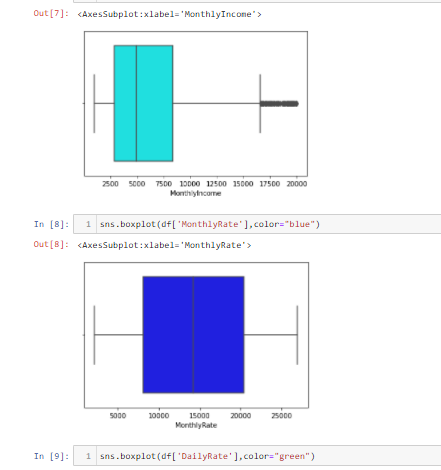
* travel rarely
* who belongs to R&D department
* who belongs to life science and medical field
* female
* working as sales executive and research scientists
* unmarried
* not working over time
* moderate work life balance
* high job involvement
* working in single company
* performance rating:3
* employees who having 0 stocks
* low monthly income.

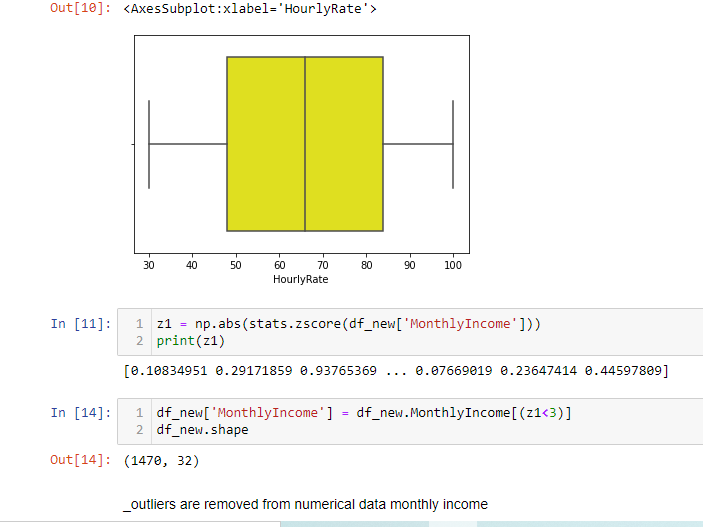
we cannot predict using relationship satisfaction,job satisfaction,Environment satisfaction features

**Correlation**



Over time feature is highly correlated with attrition

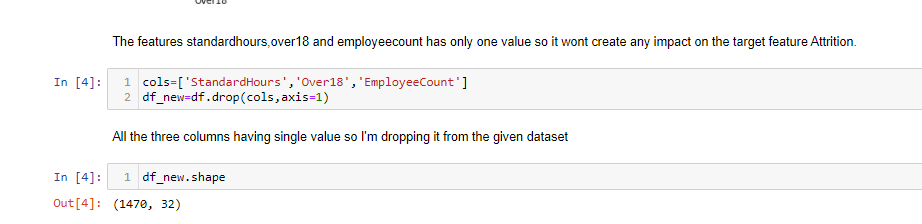
**Handling outliers in numerical column:**



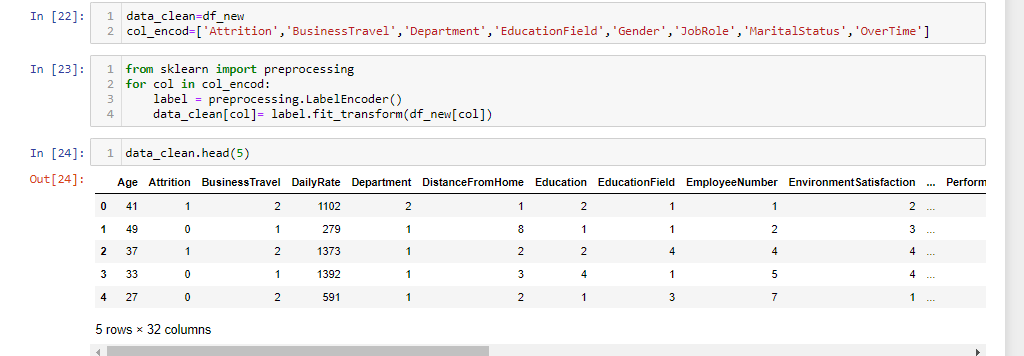
**Data pre-processing:**

The features standard hours, over18 and employee count has only single value so it won’t create any impact on the target feature Attrition.

Employee Number feature is just an identifier and it's not required for modelling either. So I’m dropping these features

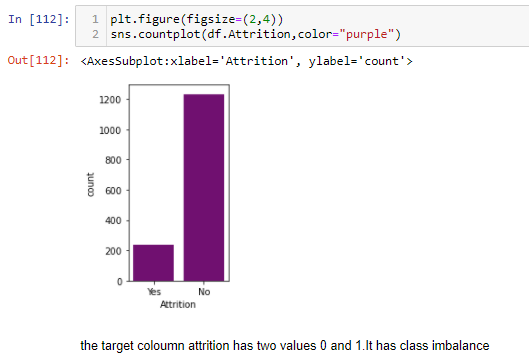


Encoding all categorical column into numerical column using label encoding technique

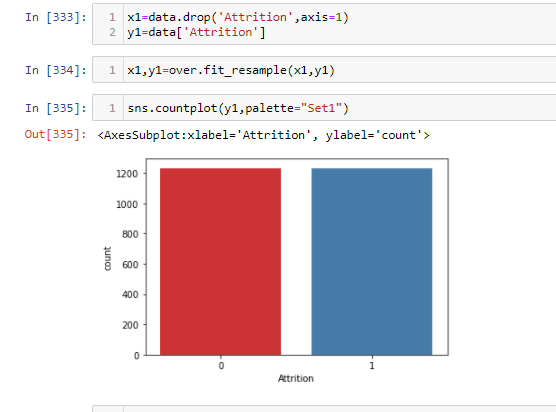


**HANDLING CLASS IMBALANCE**

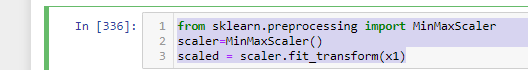
Classification problem where the distribution of examples across the known classes is biased or skewed. To avoid this we are using SMOTE technique



SMOTE synthetic over-sampling works to cause the classifier to build larger decision regions that contain nearby minority class points. This will in turn avoid data loss

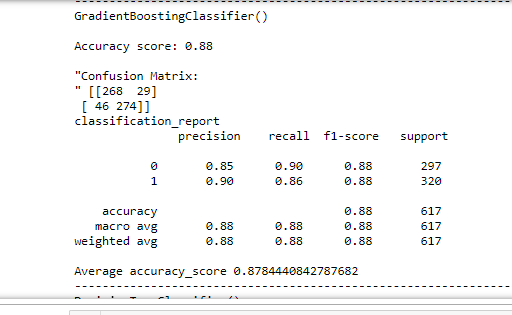
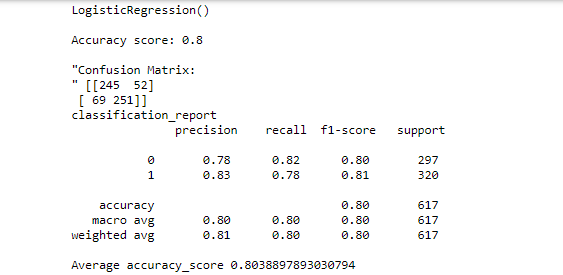
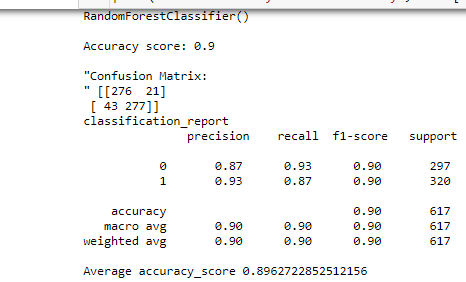
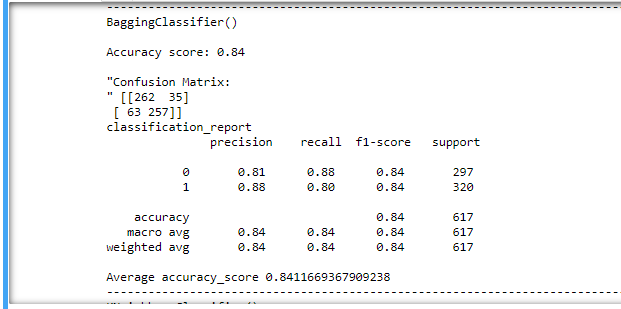
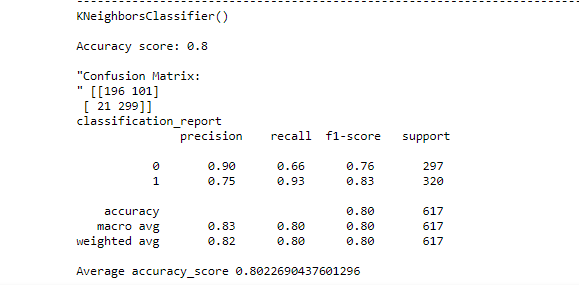
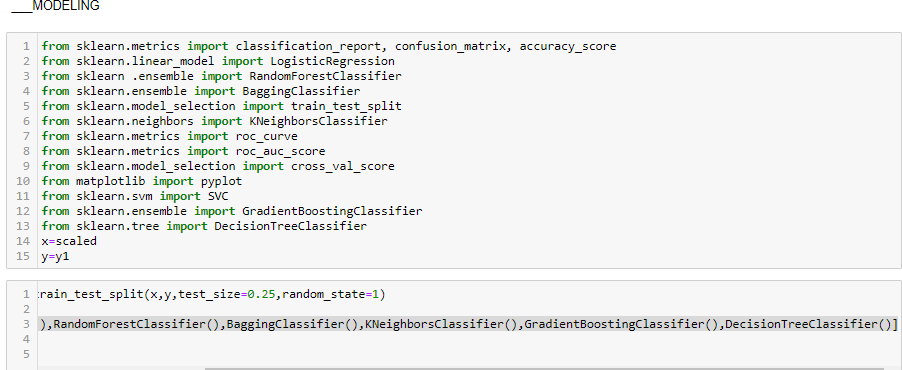


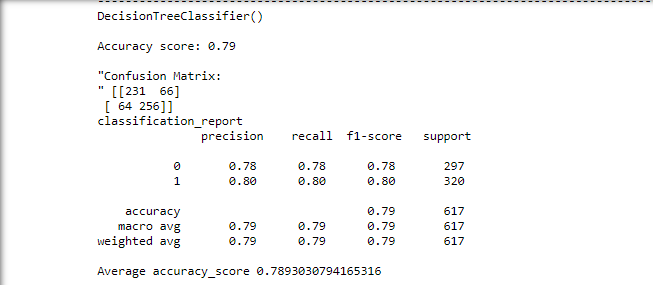
### Scaling using min max scaler



**Modelling**

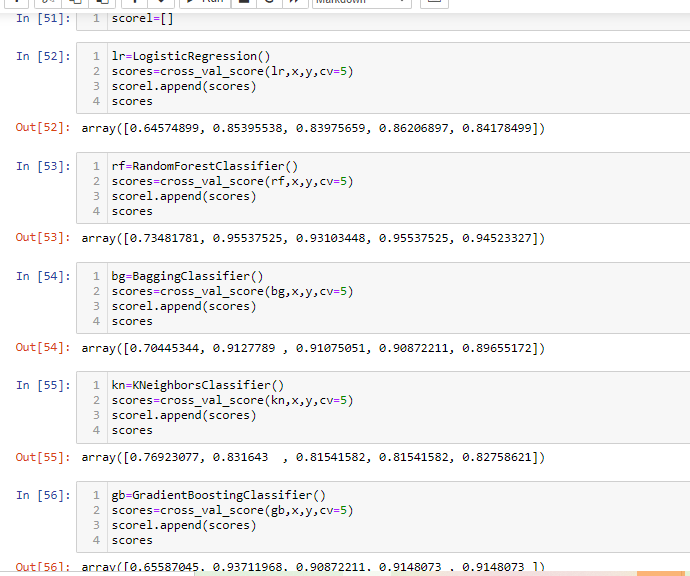
It is a binary classification problem so I have modelled using logistic regression and other classification models



 Random forest classifier has highest accuracy is **0.896272**

**Cross Validation:**

In order to avoid over fitting, Cross-validation is used  to estimate the skill of a machine learning model on unseen data.



**Difference of predicted model and crossvalidation score:**

* LogisticRegression() difference is0.0378952
* RandomForestClassifier() difference is 0.05058173
* BaggingClassifier() difference is 0.06024702
* KNeighborsClassifier() difference is 0.02531716
* GradientBoostingClassifier() difference is 0.03636322
* DecisionTreeClassifier() difference is 0.05733428

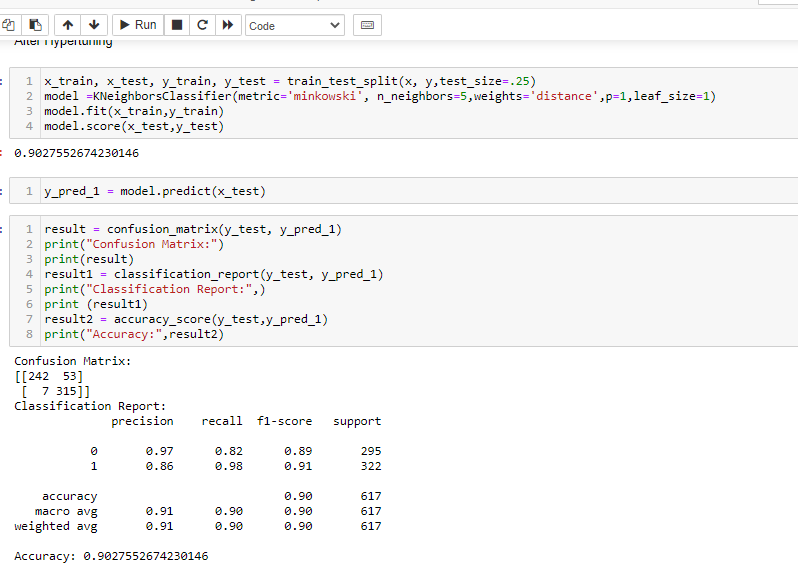
from the observation KNeighborsClassifier model has least difference so I'm selecting KNeighborsClassifier as best model

**Hyper Tuning:**

****

**Best parameters**: {'leaf\_size': 1, 'metric': 'minkowski', 'n\_neighbors': 5, 'p': 1, 'weights': 'distance'}

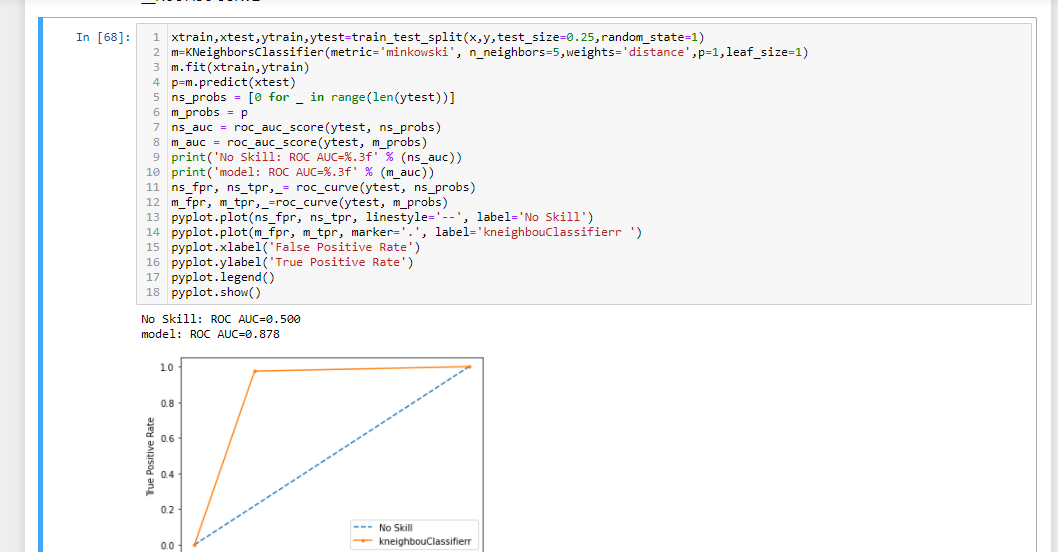
**Modelling using best parameter and best model:**

****

Final model after hyper tuning with accuracy **0.9027552674230146**

Best model:KNeighbourClassifier Best param: {'leaf\_size': 1, 'metric': 'minkowski', 'n\_neighbors': 5, 'p': 1, 'weights': 'distance'}

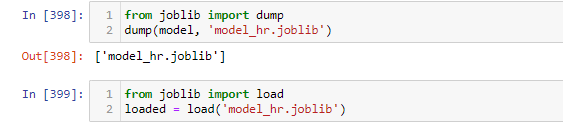
**ROC AUC CURVE**:

****

**Conclusion:**

I have developed a model to predict attrition of an employee with 90.2% accuracy

**Saving the model**

****

**Baseball Case Study**



**Problem Statement:**

This dataset utilizes data from 2014 Major League Baseball seasons in order to develop an algoirthm that predicts the number of wins for a given team in the 2015 season based on several different indicators of success. This model is used to select best team based on best input feature. .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:**

* R: Runs-times reached home plate legally and safely
* AB: At Bats-plate appearances, not including bases on balls, being hit by pitch, sacrifices, interference, or obstruction
* H: Hits- reaching base because of a batted, fair ball without error by the defence
* 2B: Doubles-hits on which the batter reaches second base safely without the contribution of a fielding error
* 3B: Triples- hits on which the batter reaches third base safely without the contribution of a fielding error
* HR: Homeruns
* BB: Walks-times pitching four balls, allowing the batter to take first base
* SO: Strikeouts
* SB: Stolen Bases- number of bases advanced by the runner while the ball is in the possession of the defence
* RA: Runs Allowed
* ER: Earned Runs- number of runs that did not occur as a result of errors or passed balls
* ERA: Earned Run Average (ERA)- total number of earned runs (see "ER" above), multiplied by 9, divided by innings pitched
* SO: Shutouts- number of complete games pitched with no runs allowed
* SV: Saves- number of games where the pitcher enters a game led by the pitcher's team, finishes the game without surrendering the lead, is not the winning pitcher, and either (a) the lead was three runs or fewer when the pitcher entered the game; (b) the potential tying run was on base, at bat, or on deck; or (c) the pitcher pitched three or more innings
* CG: Complete Games-number of games where player was the only pitcher for their team
* E: Errors- the judgment of the official scorer, of a fielder misplaying a ball in a manner that allows a batter or base runner to advance one or more bases

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

**Importing the Libraries**

import pandas as pd

import numpy as np

import seaborn as sns

from scipy import stats

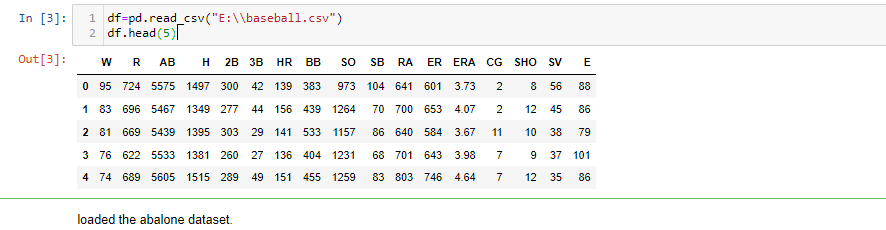
import warnings

warnings.filterwarnings('ignore')

from matplotlib import pyplot as plt

from scipy.stats import zscore

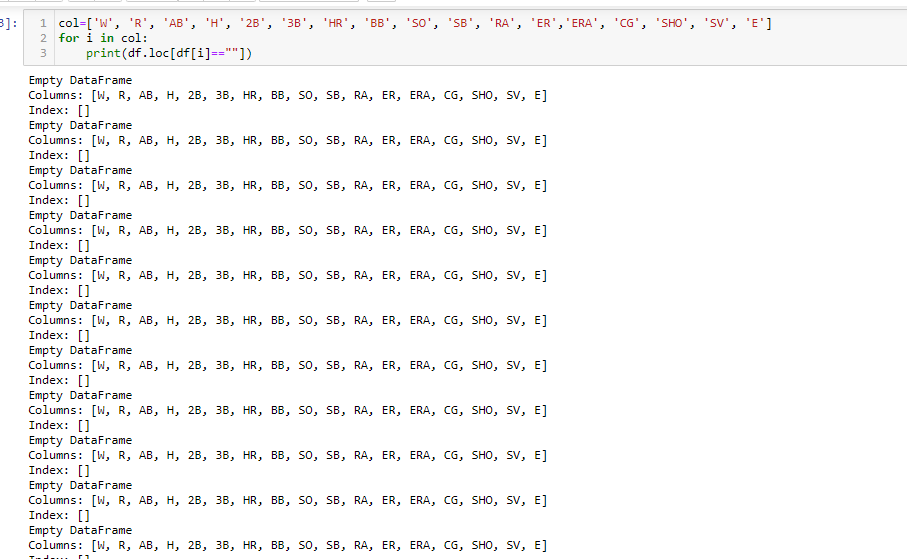
**Loading Dataset:**

****



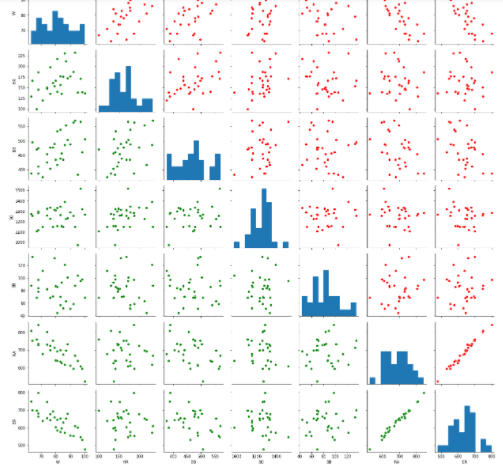
This dataset has no null values and it has 30 rows and 17 columns

CHECKING FOR EMPTY SPACE IN DATASET



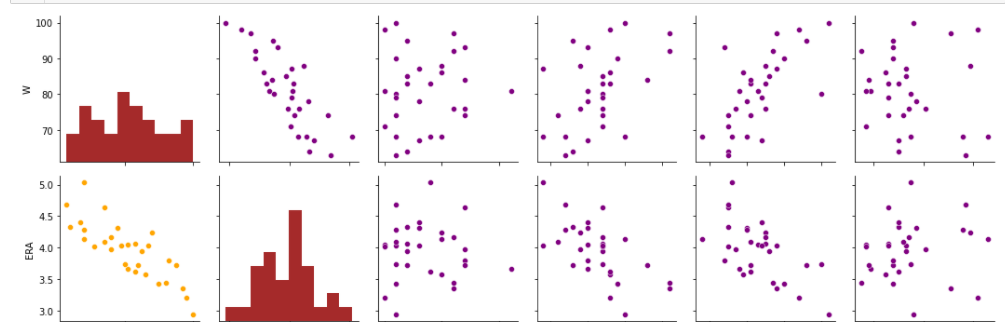
This dataset has no empty space as value

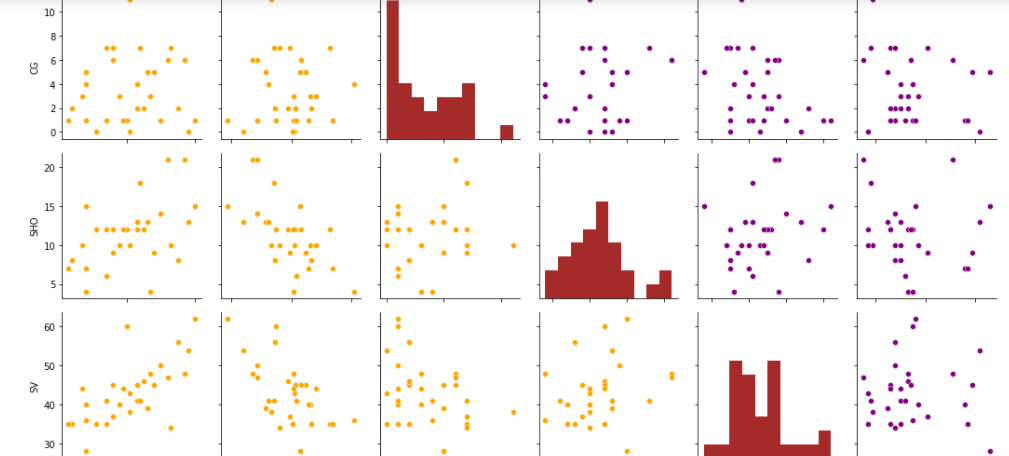
The data points of the features R and AB showing an uphill pattern as you move from left to right so it has positive relationship. The data points of the features R and w showing an uphill pattern as you move from left to right so it has positive relationship.

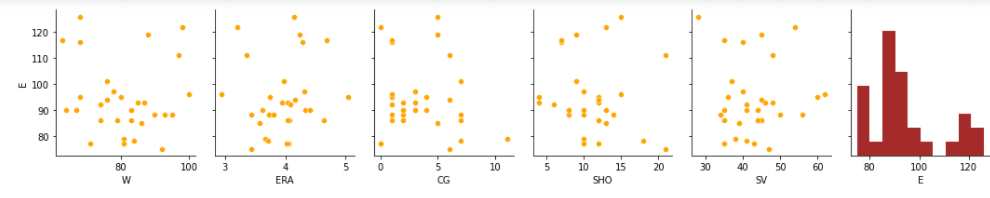
  
The data points of ER and RA showing an uphill pattern as you move from left to right so it has positive relationship.

The data points of ER and w showing an downhill pattern as you move from left to right so it has negative relationship.

The data points of RA and w showing an downhill pattern as you move from left to right so it has negative relationship.



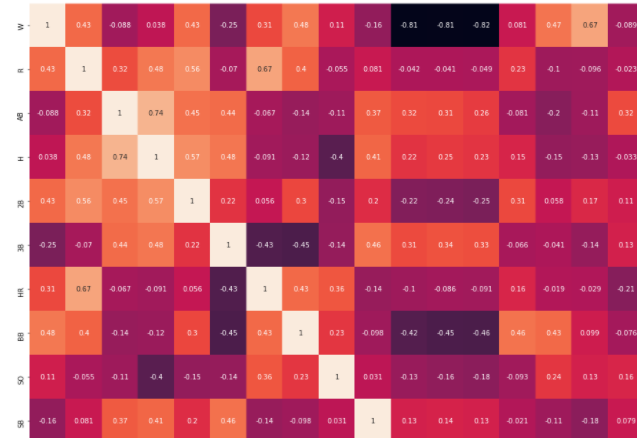
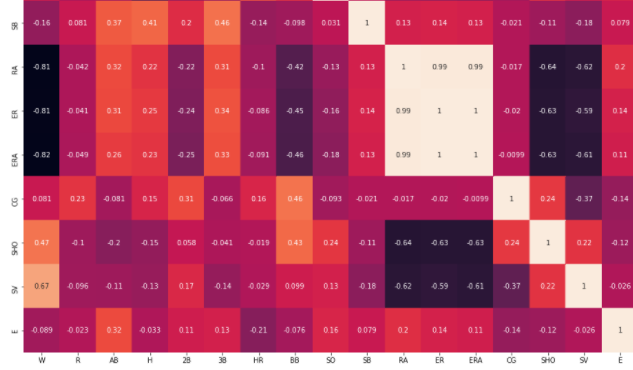




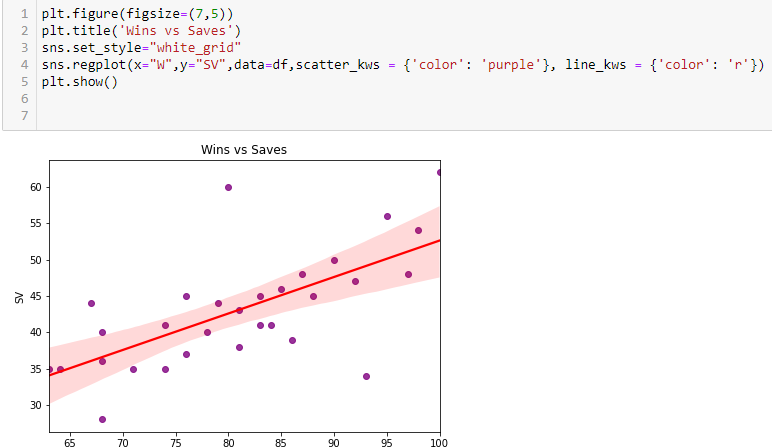
The data points of the features ERA and w showing an downhill pattern as you move from left to right so it has negative relationship.

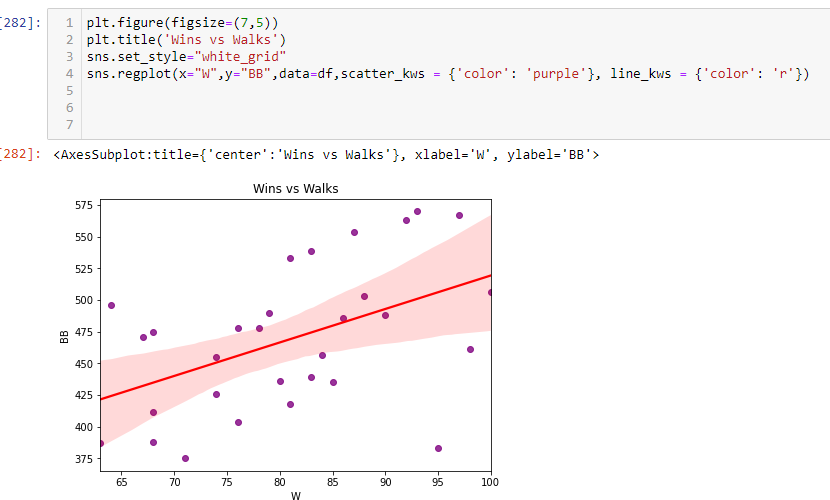
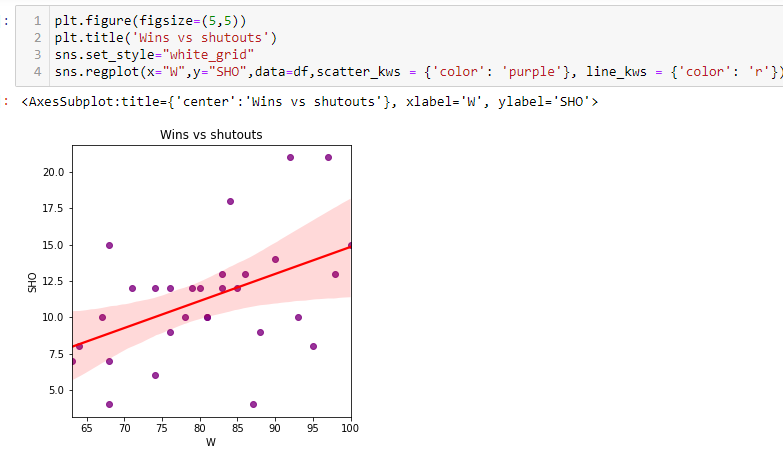
The data points of the features SV and W showing an uphill pattern as you move from left to right so it has positive relationship

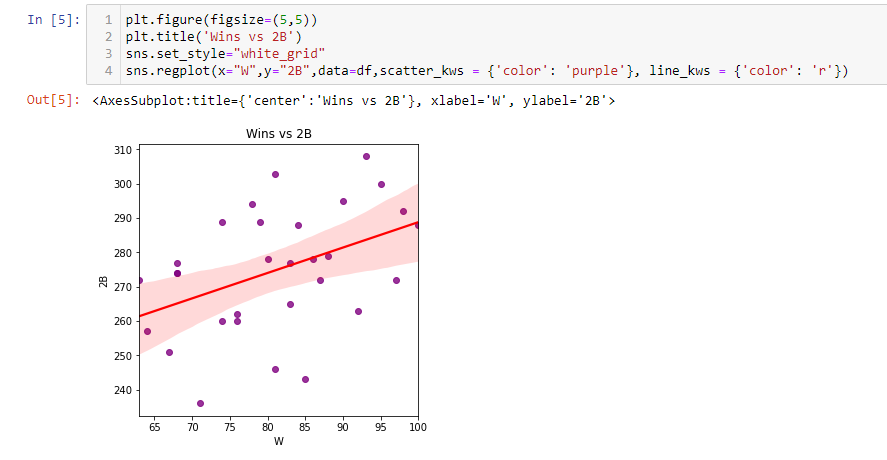
**CORRELATION MATRIX**

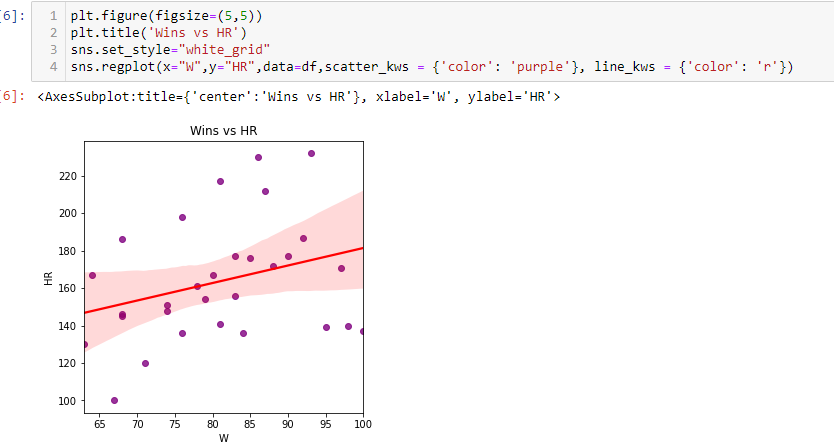
****

* RA,ER,ERA features are highly negatively correlated with wins.
* RA,ER,ERA features are highly negatively correlated with BB.
* RA,ER,ERA features are highly negatively correlated with sv and sho.
* Sho,bb,2b,SV,HR are positively correlated with win
* ER andERA are posively correlated with RA
* wins feature is highly and positively correlated with saves
* RA,ER,ERA features are highly negatively correlated with saves.
* h,cg having very less correlation with winsRA,ER,ERA features are highly negatively correlated with wins.
* RA,ER,ERA features are highly negatively correlated with BB.
* RA,ER,ERA features are highly negatively correlated with sv and sho.
* Sho,bb,2b,SV,HR are positively correlated with win
* ER andERA are posively correlated with RA
* wins feature is highly and positively correlated with saves
* RA,ER,ERA features are highly negatively correlated with saves.
* h,cg having very less correlation with wins

**Finding whether all features which are positively correlated with Wins feature have linear relationship:** In this reg plot it is apparent that SV is positively correlated with W data.most of the datapoints are almost close to best fit line it means that the feature SV and W having linear relationship

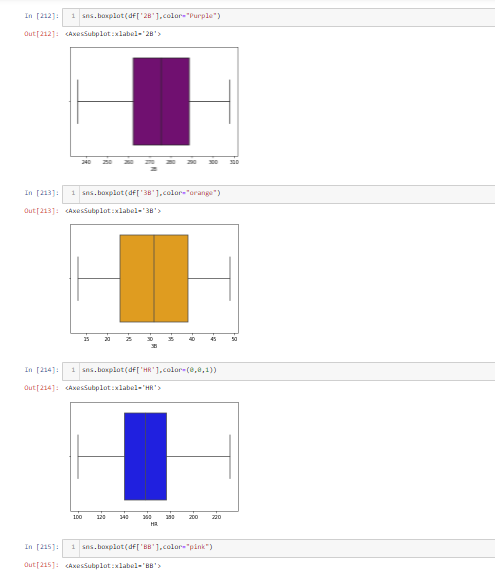
In this reg plot it is apparent that BB is positively correlated with W data. Only few datapoints are close to best fit line it means that the feature BB and W are not having linear relationship. If the value of bb is between 400 to 450 there is more chance to winIn this reg plot it is apparent that SHO is positively correlated with W data.only few datapoints are close to best fit line it means that the feature SHO and W are not having linear relationship

In this reg plot it is apparent that 2B is positively correlated with W data. Only few data points are close to best fit line it means that the feature 2B and W are not having linear relationship

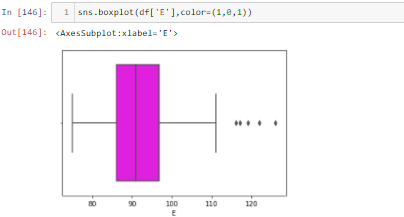
In this reg plot it is apparent that HR is positively correlated with W data. Only few datapoints are close to best fit line it means that the feature HR and W are not having linear relationship

**IDENTIFYING OUTLIERS:**

****

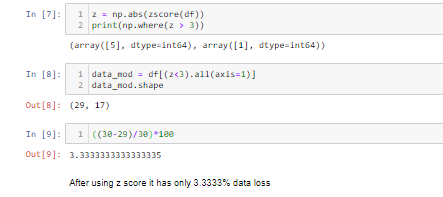
****

****

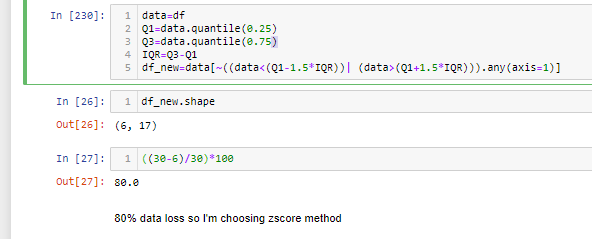


The features R,ERA,SHO,SV,E are having outliers

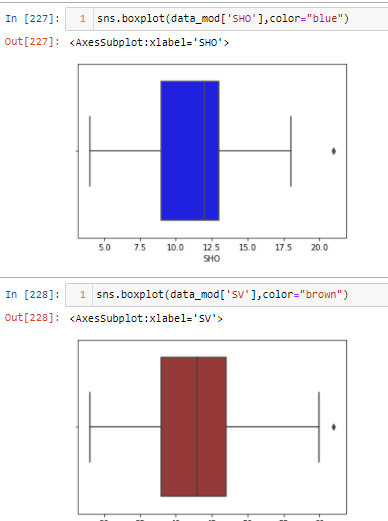
**REMOVING OUTLIERS**

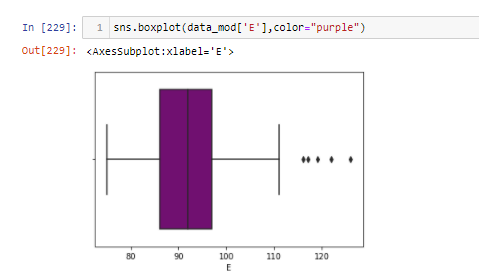


IQR TO REMOVE OUTLIERS:



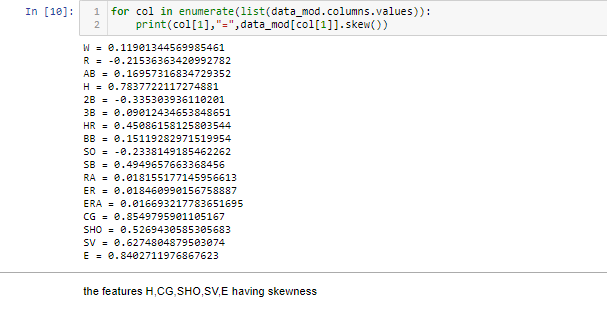
AFTER USING ZSCORE:

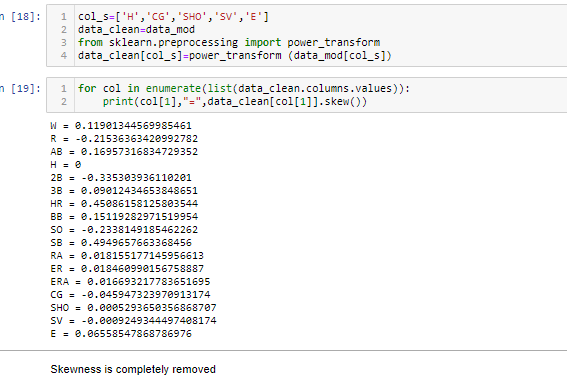


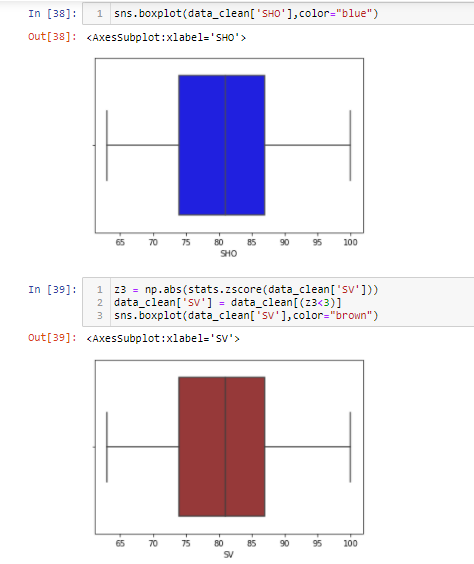
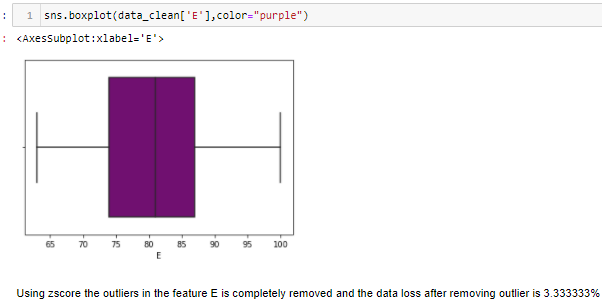


After using z\_score the outliers in the features R and ERA are removed but the features SV,SHO,E are still having outliers

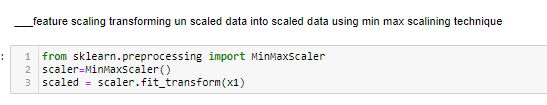
**HANDLING SKEWNESS**:

 To stabilize variance, make the data more normal distribution like, improve the validity of measures of association and to remove skewness I have used power transformation

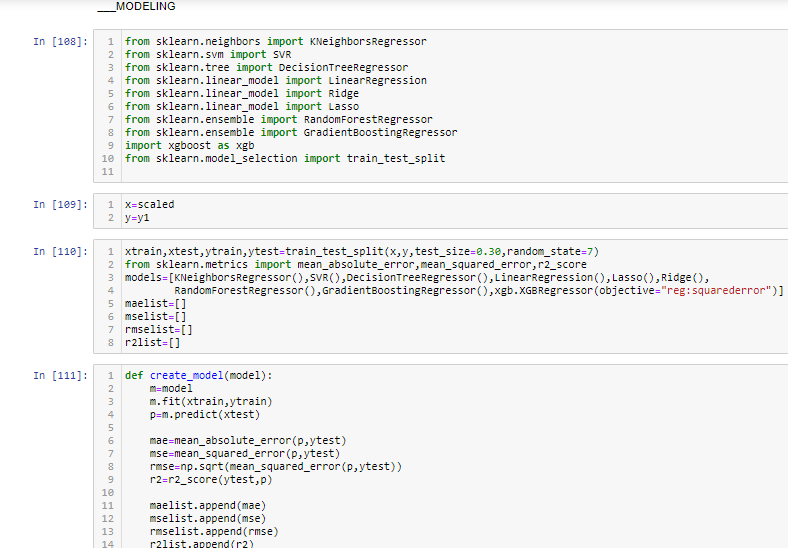


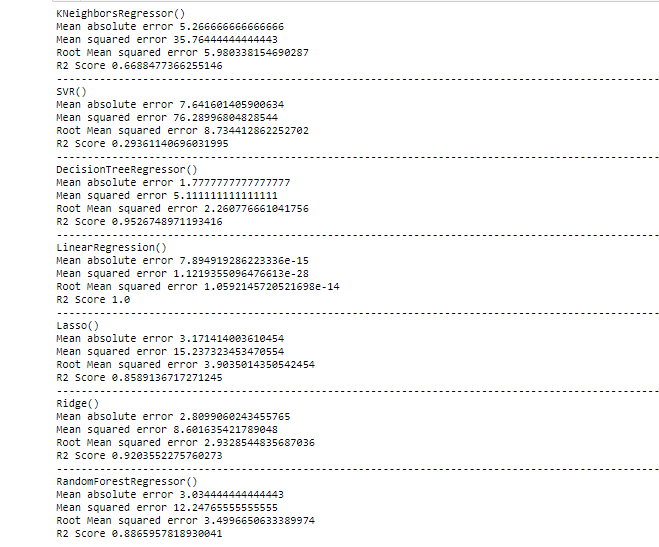
I have used Power transform on skewed data to make it symmetric, and then fit it to a symmetric distributionI have used zscore to remove outliers in the features SV, SHO and E.

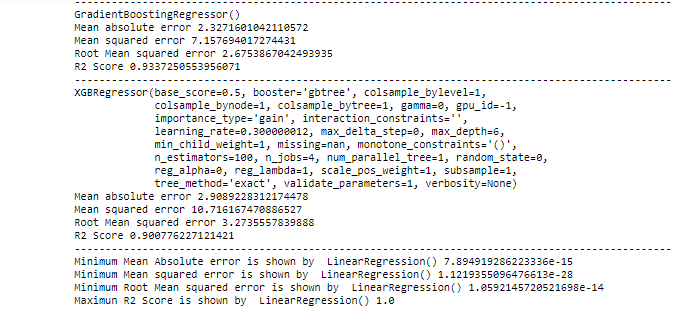
**SCALING:**



**MODELLING**:



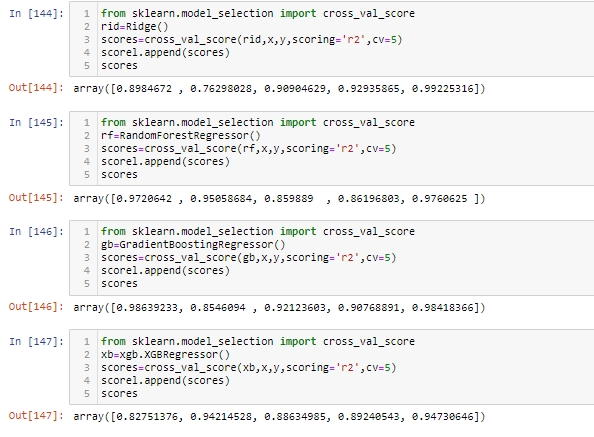




**Cross Validation:**

In order to avoid over fitting, Cross-validation is used to estimate the skill of a machine learning model on unseen data.



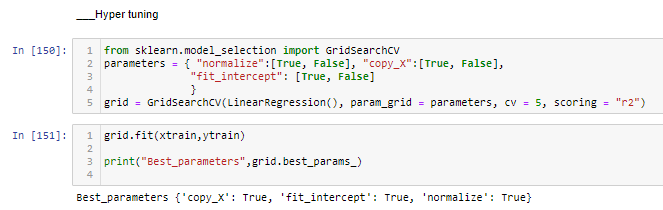


**Difference of predicted model and cross validation score:**

* KNeighborsRegressor - 0.2337466
* SVR()-0.12155981
* DecisionTreeRegressor()-0.05291075
* LinearRegression() -0
* Lasso - 0.02748291
* Ridge()-0.07189794
* RandomForestRegressor()-0.08946672
* GradientBoostingRegressor()-0.0504586
* XGBRegressor-0.046530

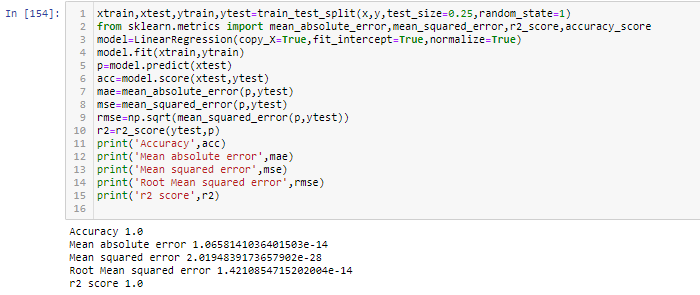
from the observation Linear regression model model has least difference so I'm selecting Linear regression as best model

**Hyper Tuning:**

****

**Best parameters**: **{**'copy\_X': True, 'fit\_intercept': True, 'normalize': True}

**Modelling using best parameter and best model:**

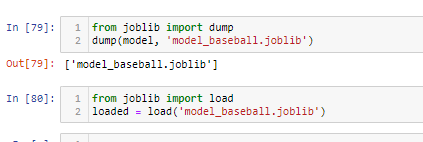
****

Final model after hyper tuning its retaining 100%accuracy and error values got reduced

**Conclusion:**

I have developed a model to predict number of wins with 100% accuracy

**Saving the model**

****