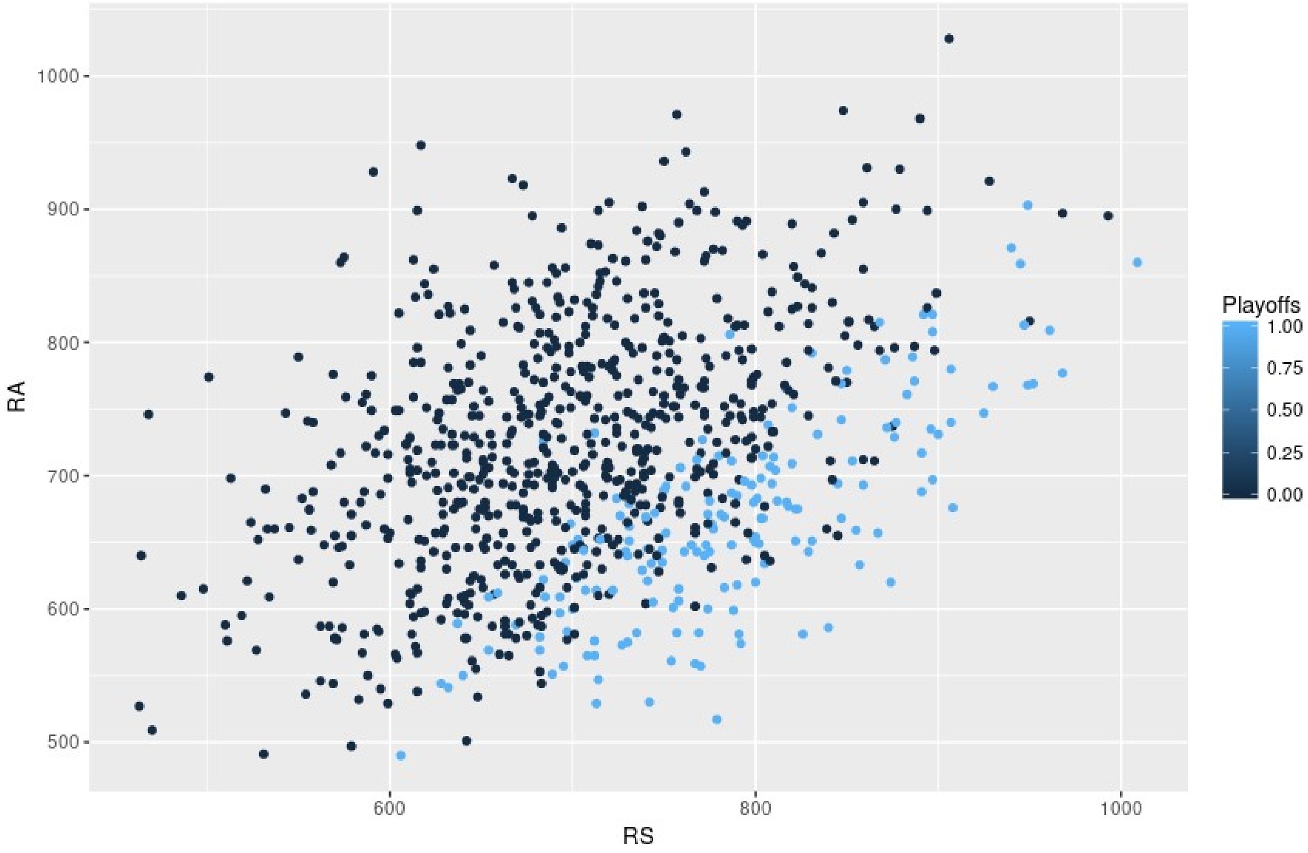
REPORT

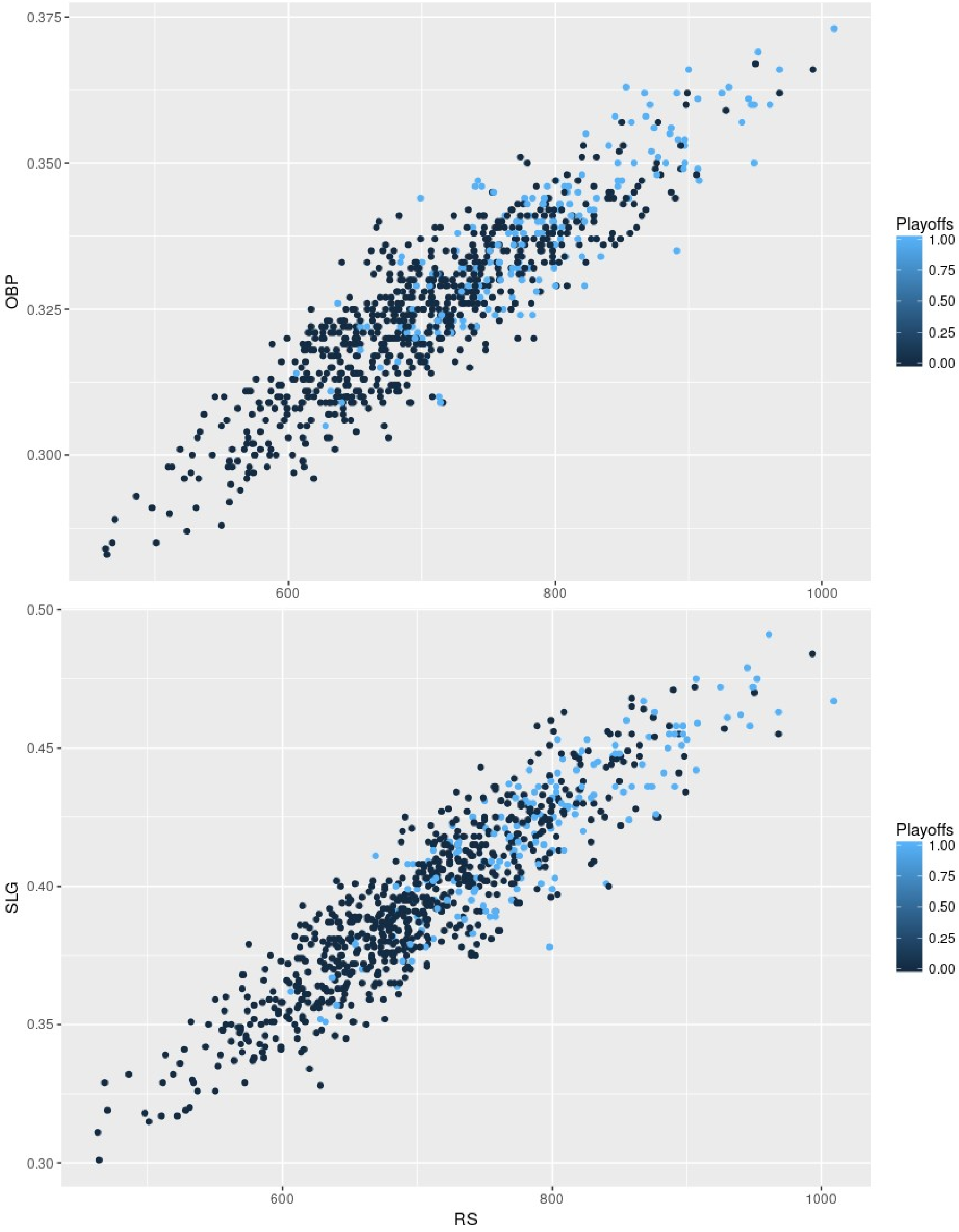
1. Problem :

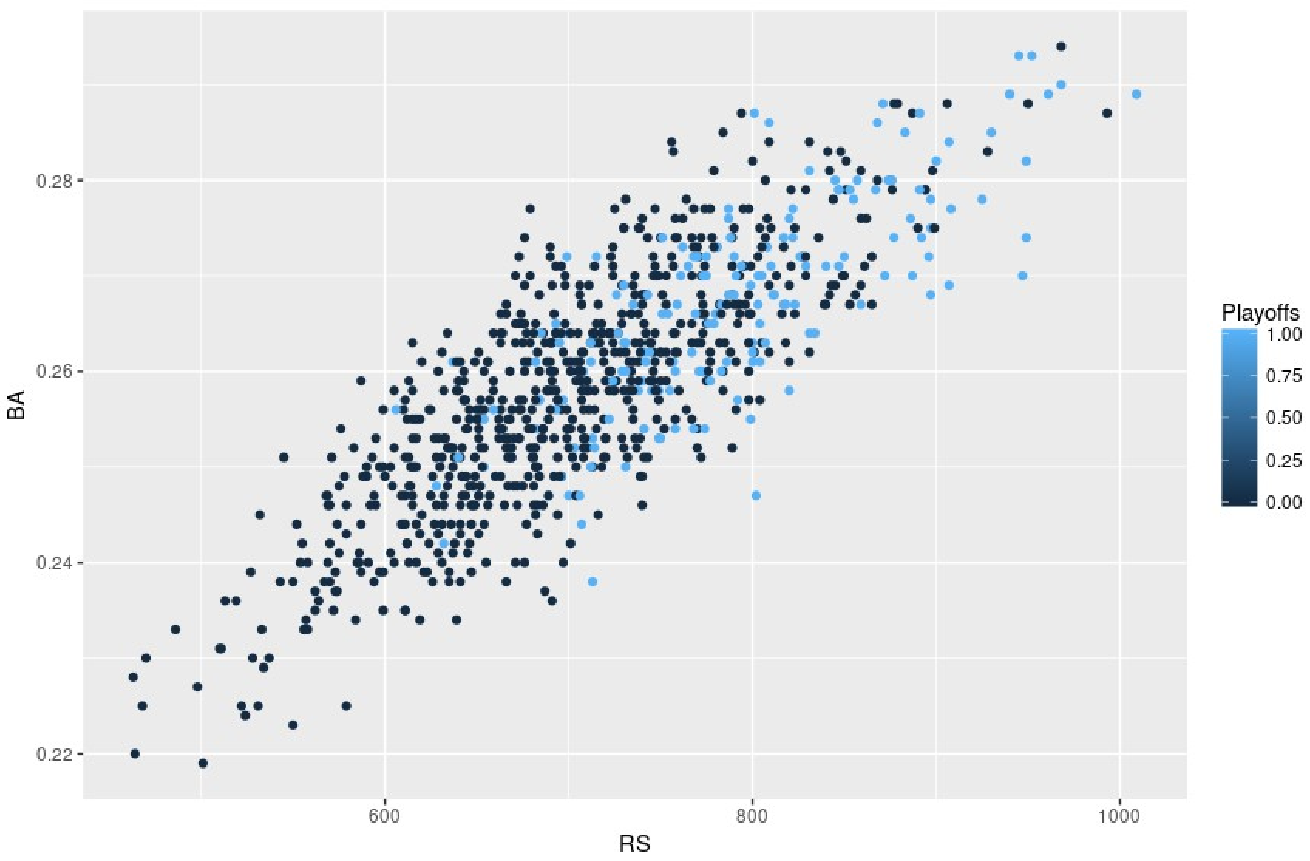
Predict whether the baseball team will qualify for the Playoffs based on the features given.

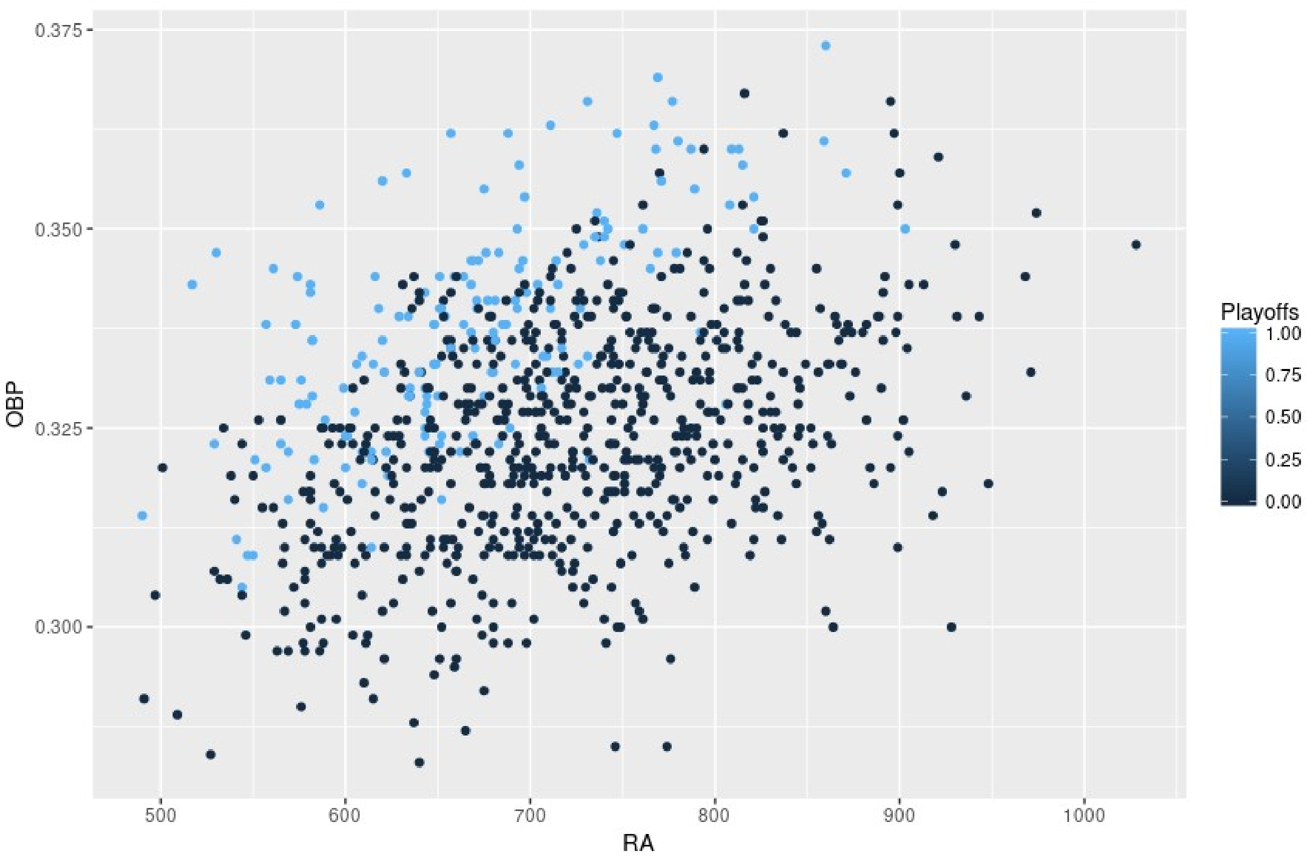
2. Visualization

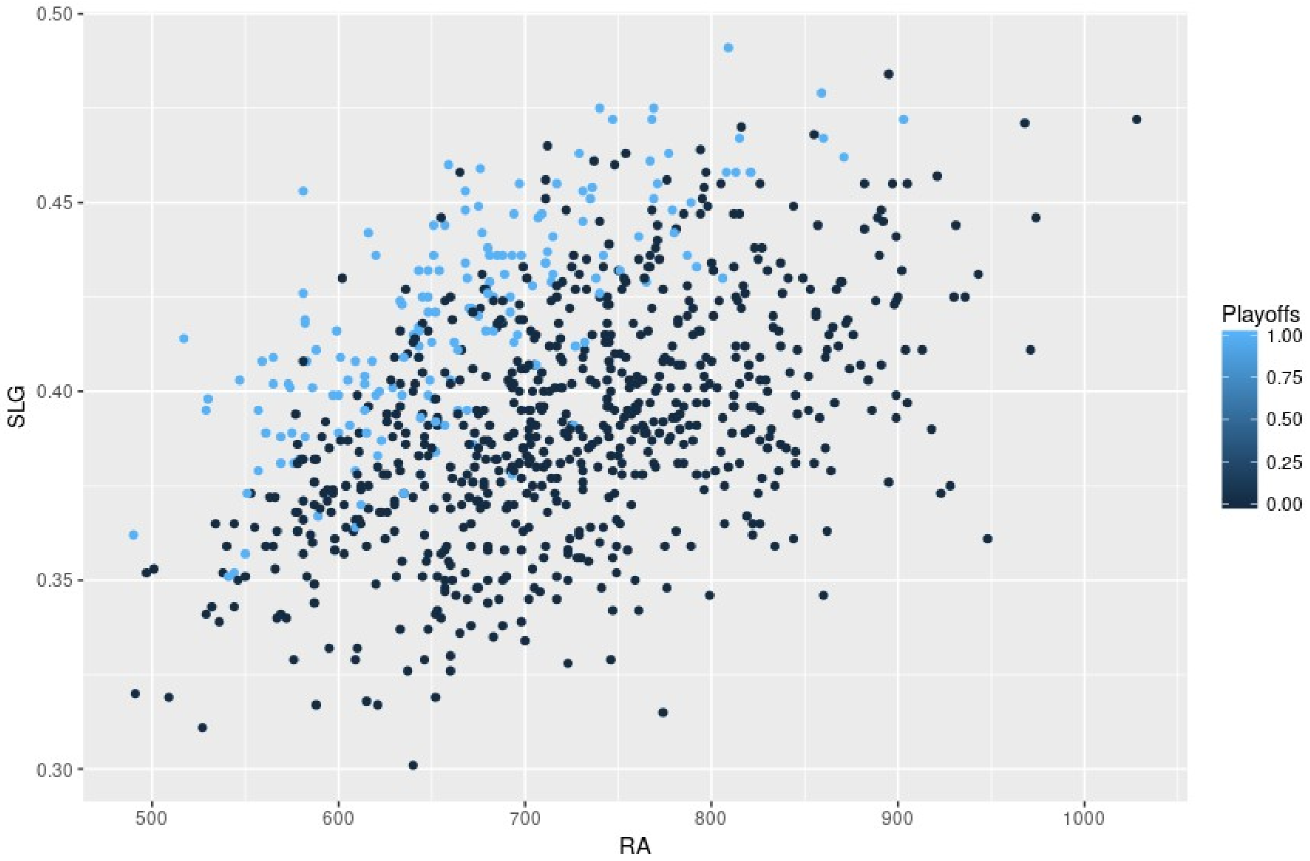
The follwing list of plots describe how each and every feature is effecting the Playoffs. It represent the visualization of the readings and the resultant inference(qualify for playoffs(1) or not(0)?) made from the data.

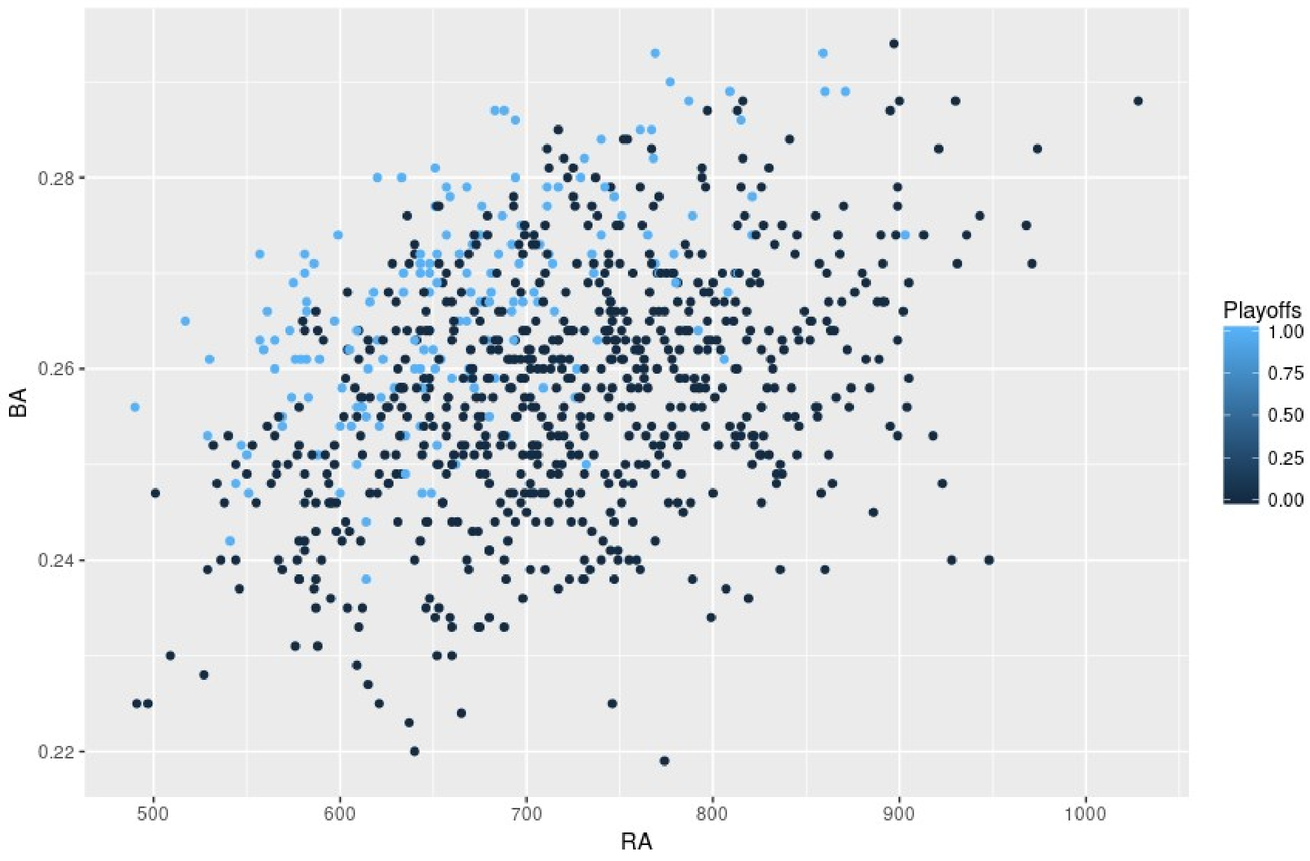


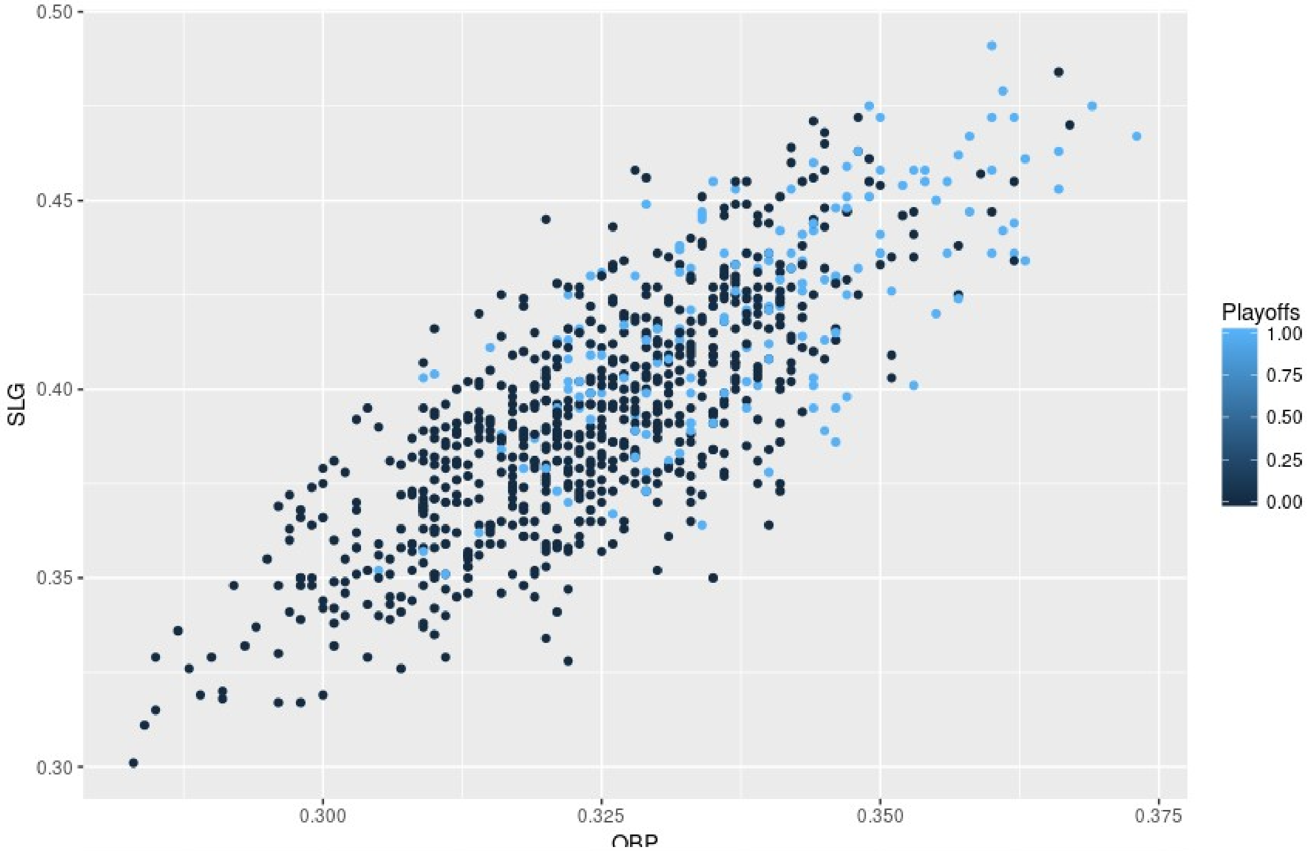

OBP
0.375
0.350
0.325
0.300
600
800
1000
Playoffs
1.00
0.75
0.50
0.25
0.00
SLG
0.50
0.45
0.40
0.35
0.30
600
800
1000
RS
Playoffs
1.00
0.75
0.50
0.25
0.00

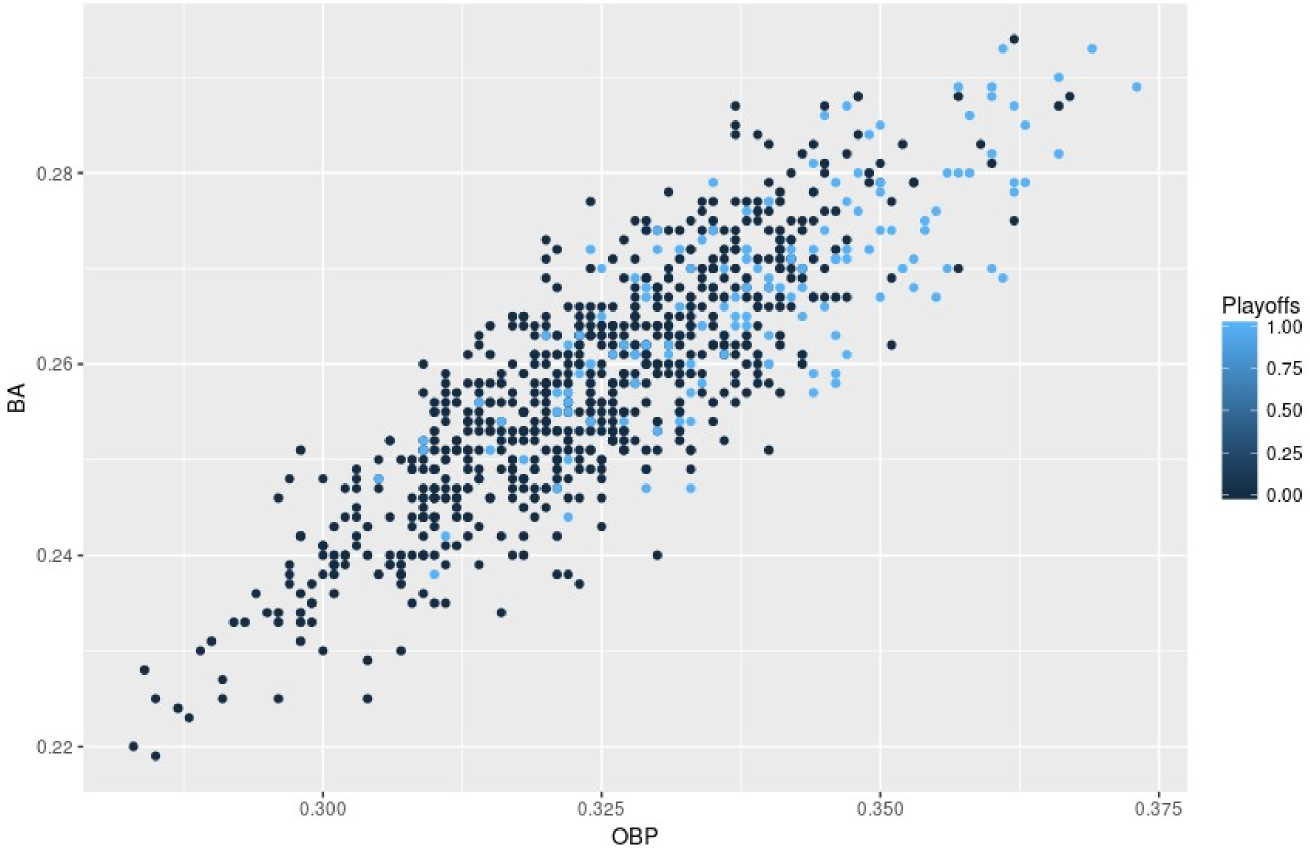

BA
0.28
0.26
0.24
0.22
600
800
1000
RS
Playoffs
1.00
0.75
0.50
0.25
0.00

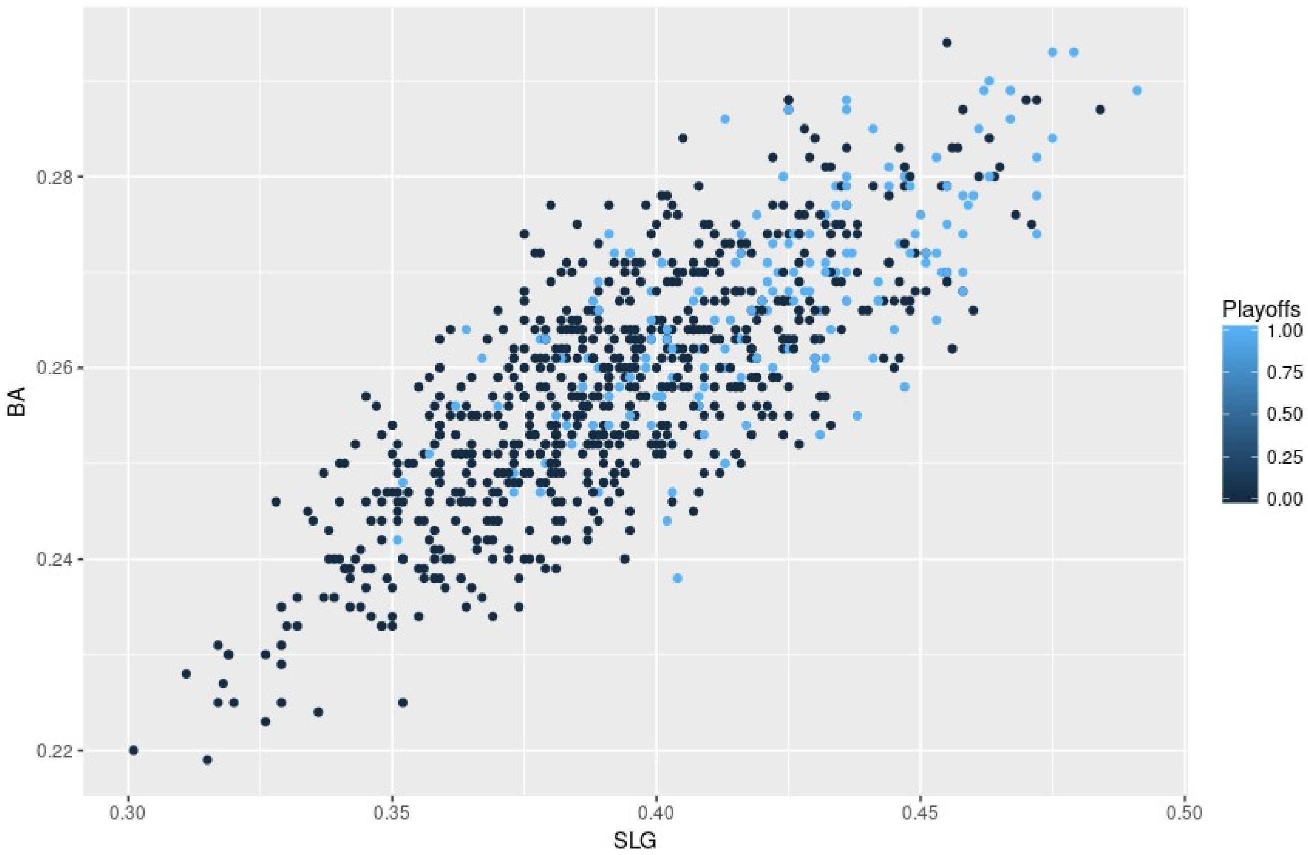

OBP
0.375
0.350
0.325
0.300
500
600
700
800
900
1000
RA
Playoffs
1.00
0.75
0.50
0.25
0.00


0.50
0.45
0.40
0.35
0.30
SLG
500
600
700
800
900
1000
RA
Playoffs
1.00
0.75
0.50
0.25
0.00


0.28
0.26
0.24
0.22
BA
500
600
700
800
900
1000
RA
Playoffs
1.00
0.75
0.50
0.25
0.00


SLG
0.50
0.45
0.40
0.35
0.30
0.300
0.325
0.350
0.375
OBP
Playoffs
1.00
0.75
0.50
0.25
0.00


BA
0.28
0.26
0.24
0.22
0.300
0.325
0.350
0.375
OBP
Playoffs
1.00
0.75
0.50
0.25
0.00



By observation one can see that the feature are have postive correlation with each other . Hence we can use line fitting to predict the unknown values.

3. Algorithms

3.1 Linear Regression

Since there is linear dependecy between the feature values we can model the given data using a linear model by using the followin R code.

```
predictM <- lm(Playoffs ~  RS + RA + OBP + SLG + BA, data = train)
summary(predictM)
```

3.1.1 Approach

We use the following list of steps to build the model.

- Split the given training data to train and test data sets so that we can perform the testing on the known set of training samples
- set a seed so that the algorithm return identical result for each execution.
- Apply the linear regression.
- For the values predicted below a particular threshold(here, we take mean of the values predicted) assign as not qualify to playoffs(0) those values below mean and others as to qualify for playoffs(1).
- Now, apply the model on unknown set of values.

4. *Inferences*

Using Linear Regression (ie. Line fitting) we were able to arrive at an accuracy of 72.69% on the random sample taken from the training set. The so created model is used for predicting the unknown labels in the test data file.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 103   2
         1  57  54

               Accuracy : 0.7269
                 95% CI : (0.6623, 0.7851)
    No Information Rate : 0.7407
    P-Value [Acc > NIR] : 0.7095
```

From the model summary we were able to infer that the features RS, RA, OBP, BA are more favorable for predict the unknow value Playoffs. Sinc they are found significant.

```
> summary(predictM)

Call:
lm(formula = Playoffs ~ RS + RA + OBP + SLG + BA, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59149 -0.22398 -0.06967  0.13715  0.80095 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.9955377  0.4929242  -2.020  0.04383 *  
RS           0.0012733  0.0004821   2.641  0.00846 ** 
RA          -0.0019396  0.0001461 -13.274  < 2e-16 ***
OBP          6.9022954  2.2403381   3.081  0.00215 ** 
SLG          2.0603652  0.9872029   2.087  0.03728 *  
BA          -5.4351615  1.9370299  -2.806  0.00517 ** 
```

5. *Packages used*

caret, ggplot2