Lab05

Question 1

1) Predictive question 1: Which bastmen will have the highest strike rate against fastballs in the next baseball game between Team A vs Team B

2) Predictive question 2: What type of ball will have the highest % of taking wickets in the next championship.

3) Descriptive 1: What is the propotion of left hand bats vs right hand

4) Descriptive 2: What is mean speed a pitcher dilivers at.

The predictive questions are more intrested in the factor that predicts the behavior where as there is no interpretation of the descriptive questions as they are a fact.

Question 2

Each team is a unit of observation and Each player for the batting data set identified through the player ID.

Question 3

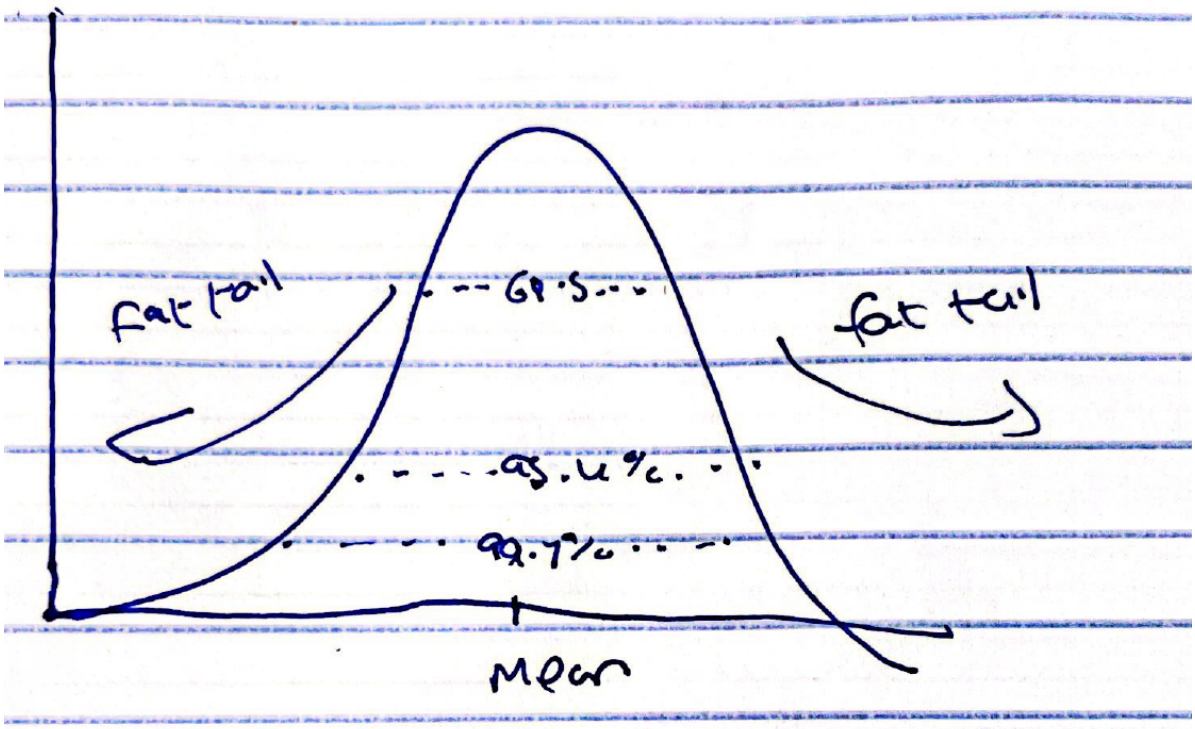
Teams data set answers how many games each team has won/lost but batting doesn;t, whereas batting answers how many each player has played for a team and scored but teams doesn't.

Question 4

We would need more granual data to answer questions along the lines of what % of the runs scored were in a winning cause by a player or Against which team does a player have the highest avg against. For this we need to have specific data about how many runs a player scored against a specific team or the details and a full breakdown for each match with the runs scored and how impactful inings details.

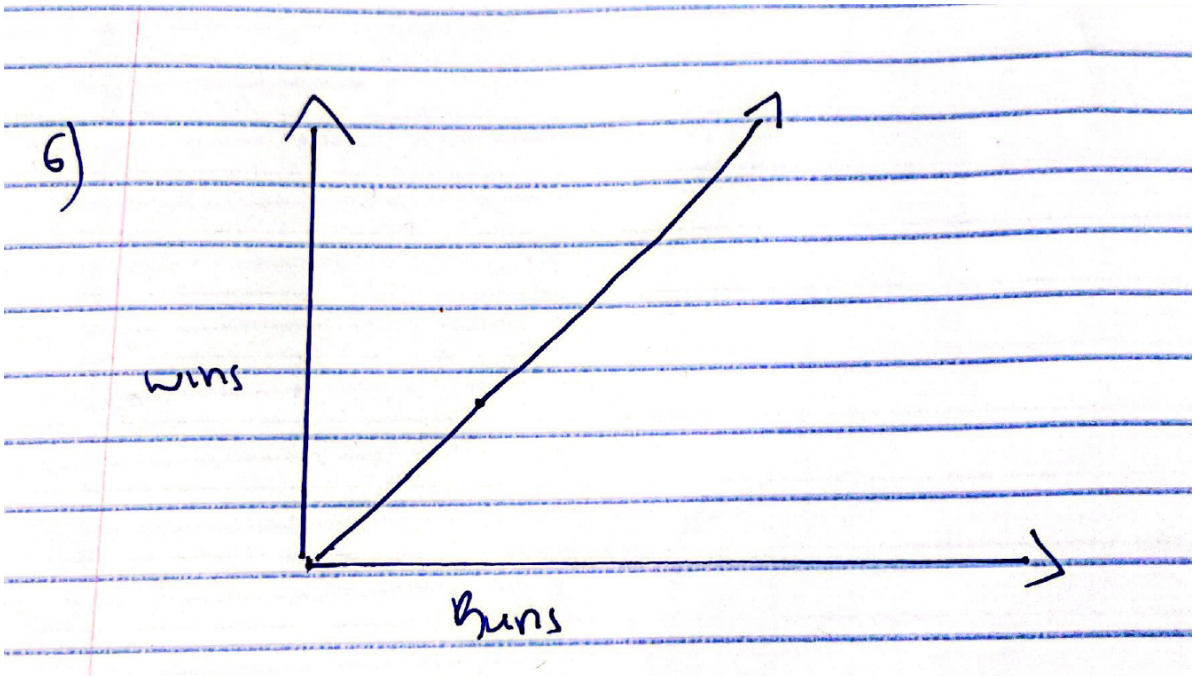
Question 5

The distrubtion of wins depends on the team deduced from the data on the table but on average its close to a 50/50 distrubution to the number of wins and loses in a team faces shown in the distrubution below.



Question 6

As the number of runs are scored the number of wins are also more, this is possitive linear relationship beceause it becomes harder for the team to chase the runs scored.



Question 7

In my opinion it is for the better, maybe I am biased because I have seen money ball, but I believe it is a tool that is used for the better of the game because one can use statistics to analyze a player's potential and how effective they can be in a certain teams setup, such as in the movie where they were on looking at players that could get off base.

Question 8

660x48 is the dimensions the filtered data frame

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

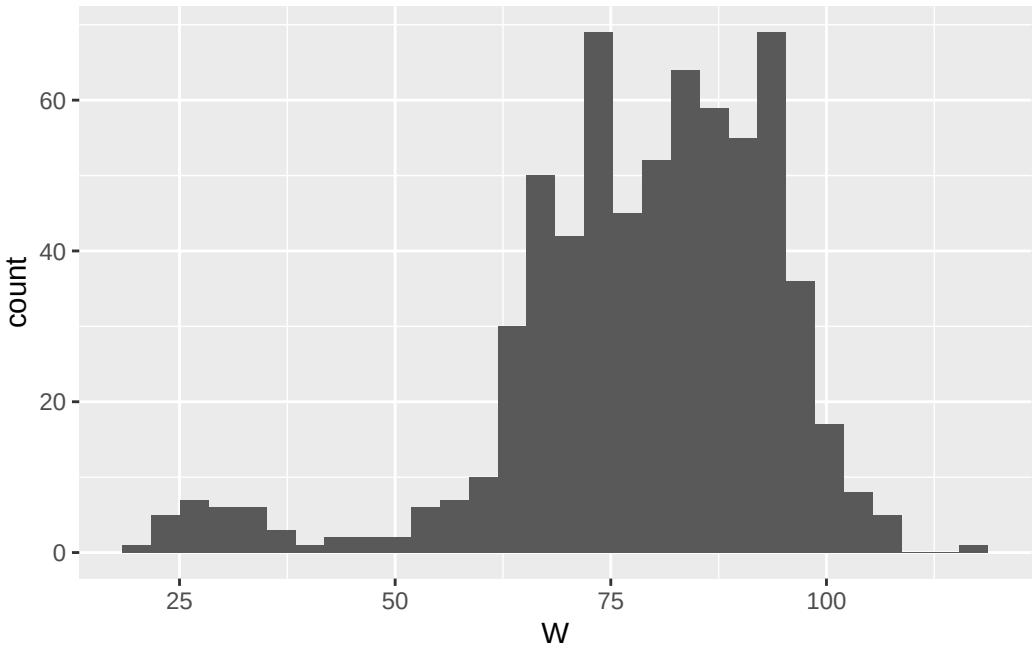
```r
library(stat20data)
library(Lahman)
data(Teams)

f1<-Teams %>%
  filter(yearID >= 2000)

dim(f1)
```

```
[1] 660  48
```

Question 9

```r
library(ggplot2)
library(tidyverse)
library(stat20data)
library(Lahman)
data(Teams)
f1 %>%
  ggplot(aes(x=W)) + geom_histogram()
```

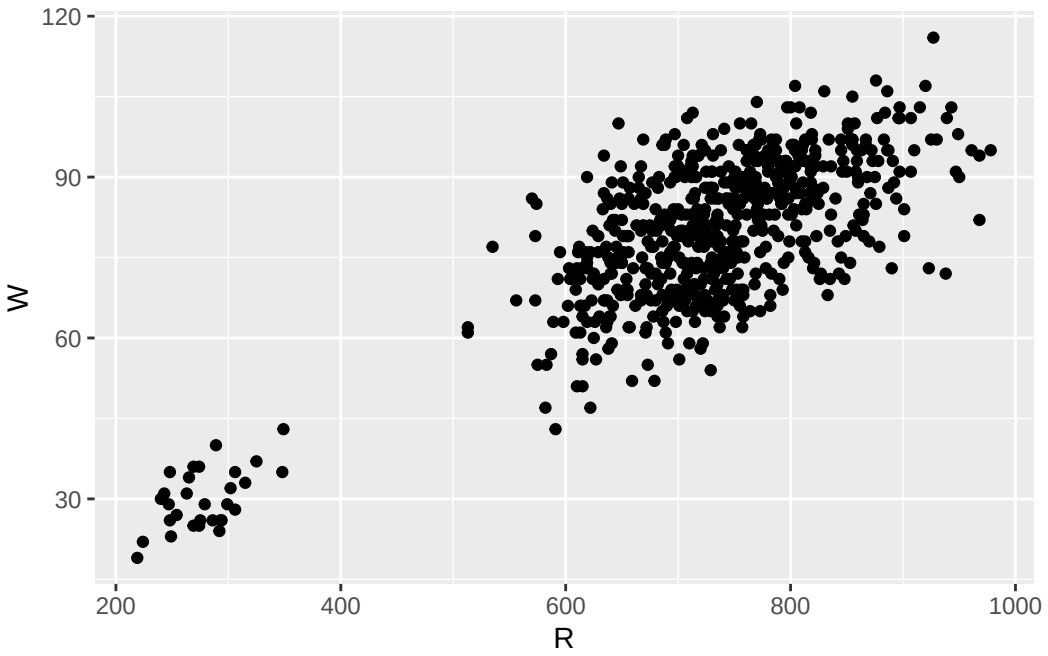
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



it's showing the distribution of the count of wins and the on the x-axis we can see that Wins ranges from 41(approx) to 118(approx) and the count(frequency) of each wins count has been plotted. The highest count lies between 75-80 and the frequency is +300. The distribution of the wins is sort of normally distributed as from the graph we can see it's bell curved graph but it is left skewed which means more values are concentrated on the right side (tail) of the distribution graph while the left tail of the distribution graph is longer.

Question 10

```
f1 %>%
  ggplot(aes(x=R, y=W)) + geom_point()
```

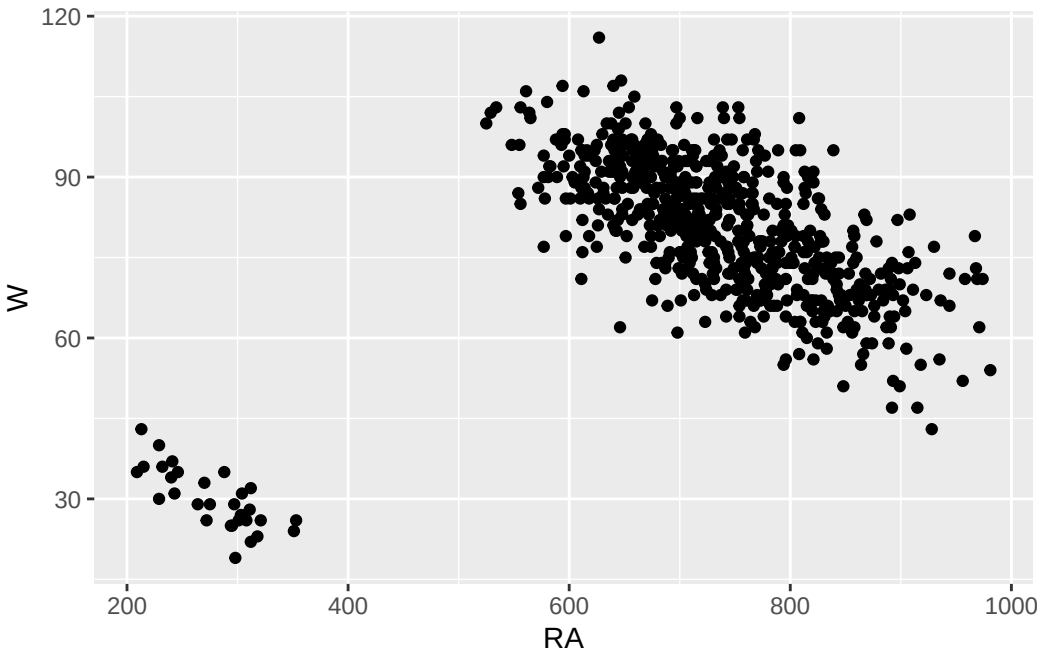


The graph shows a possitive linear relationship as the graph is slopping up in a straight/linear manner showing a strong and possitive correlation between the amount of runs scored and wins. Since the points are closely packed the relationship is very strong. There are a few outliers such as 1250 and 1100 or near 750 which are not direclty in line with the general trend of the graph however there is a strong general trend upwards, where the points are most concentrated. The gap exists because many data points have been filtered out but if one is to compare the orginal data set and this there is a clear possitive relationship.

Question 11

The graph has a weak positive correlation compared to the graph we have seen above which was very strong, the form is linear to an extent and has a high concetnration of points from 500-900 runs allowed. The rest of the points are sparesly plotted around it with a few outliers like the Runs after 900 which have lower wins, but generally the graph can show a possitive relationship of more wins as more runs are scored. The gap exists because many data points have been filtered out but if one is to compare the orginal data set and this there is a clear possitive relationship but weak strenght.

```
f1 %>%
  ggplot(aes(x=RA, y=W)) + geom_point()
```



Question 12

R^2 =

0.6119

Equation = 0.098271x + 7.960933

Two different ways to plot below:

```
f1 %>%
  lm(W~R,.) %>%
  summary()
```

```
Call:
lm(formula = W ~ R, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-28.139  -7.160   0.584   6.590  28.457
```

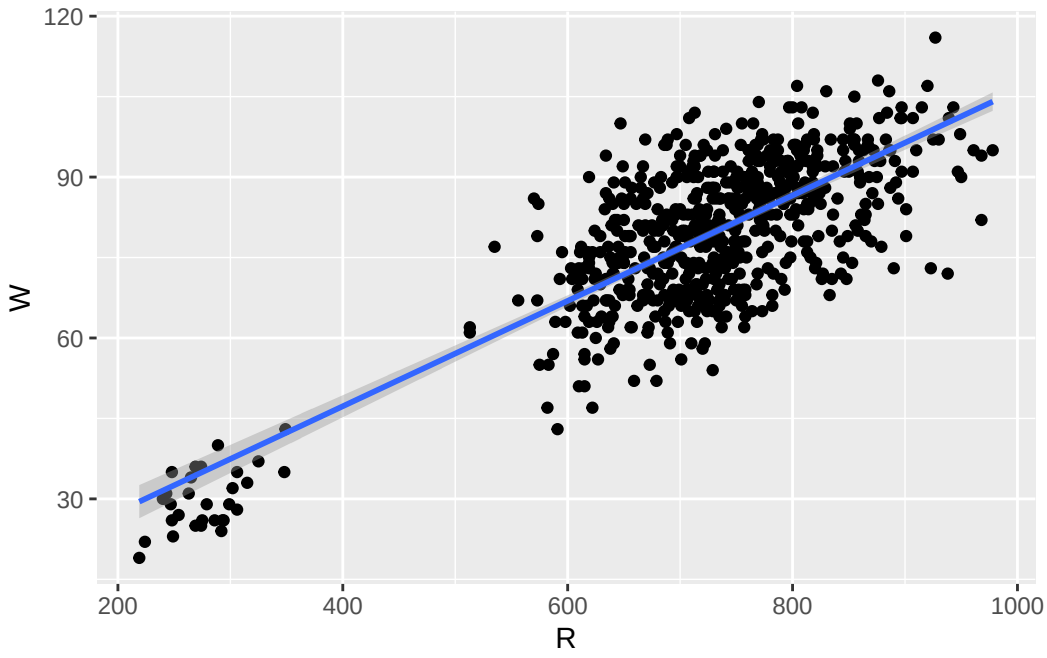
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.960933   2.228150   3.573 0.000379 ***
R           0.098271   0.003051  32.207  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.859 on 658 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6113
F-statistic:  1037 on 1 and 658 DF,  p-value: < 2.2e-16
```

```
l<-f1 %>%
  lm(W~R,.)
l
```

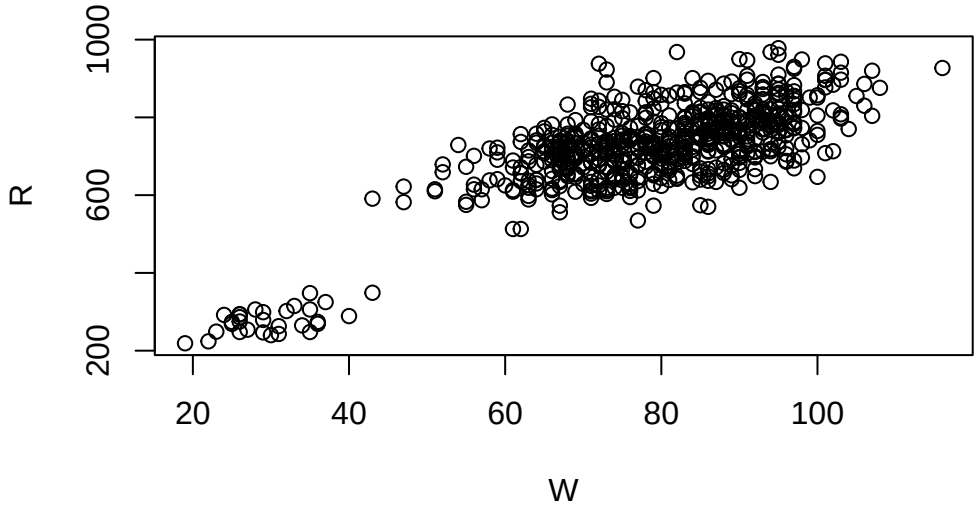
```
Call:
lm(formula = W ~ R, data = .)

Coefficients:
(Intercept)            R
    7.96093      0.09827
```

```
f1 %>%
  ggplot(aes(x=R, y=W)) + geom_point() + geom_smooth(method = "lm")
```

```
`geom_smooth()` using formula 'y ~ x'
```




```
plot(R ~ W, data = f1) + abline(l)
```




```
integer(0)
```

Question 13

Average number of season runs =

```
681.0308
```

Average number of season wins =

```
74.61106
```

using the equation where x is the runs

0.098271x + 7.960933

substituting avg num of runs

0.098271(681.0308) + 7.960933 = wins with avg runs according to the formula is 75

0.098271(500) + 7.960933 = 57 with 500 runs

0.098271(800) + 7.96093 = 86.5= 87 with 800 runs

```
library(ggplot2)
library(tidyverse)
library(stat20data)
library(Lahman)
data(Teams)
Teams %>%
  summary()
```

```
    yearID     lgID          teamID        franchID       divID
Min.   :1871   AA:  85   CHN    : 146   ATL    : 146   Length:2985
1st Qu.:1922   AL:1295   PHI    : 139   CHC    : 146   Class :character
Median :1967   FL:  16   PIT    : 135   CIN    : 140   Mode  :character
Mean   :1959   NA:  50   CIN    : 132   PIT    : 140
3rd Qu.:1997   NL:1519   SLN    : 130   STL    : 140
Max.   :2021   PL:   8   BOS    : 121   PHI    : 139
               UA:  12   (Other):2182   (Other):2134
     Rank              G           Ghome             W
Min.   : 1.000   Min.   :  6   Min.   :24.00   Min.   :  0.00
1st Qu.: 2.000   1st Qu.:154   1st Qu.:77.00   1st Qu.: 66.00
Median : 4.000   Median :159   Median :81.00   Median : 77.00
Mean   : 4.039   Mean   :150   Mean   :78.05   Mean   : 74.61
3rd Qu.: 6.000   3rd Qu.:162   3rd Qu.:81.00   3rd Qu.: 87.00
Max.   :13.000   Max.   :165   Max.   :84.00   Max.   :116.00
                               NA's   :399
      L             DivWin             WCWin              LgWin
```

```
Min.   :  4.00   Length:2985        Length:2985        Length:2985
1st Qu.: 65.00   Class :character   Class :character   Class :character
Median : 76.00   Mode  :character   Mode  :character   Mode  :character
Mean   : 74.61
3rd Qu.: 87.00
Max.   :134.00

   WSWin                R              AB             H
Length:2985        Min.   :  24   Min.   : 211   Min.   :  33
Class :character   1st Qu.: 614   1st Qu.:5135   1st Qu.:1299
Mode  :character   Median : 691   Median :5402   Median :1390
                   Mean   : 681   Mean   :5129   Mean   :1339
                   3rd Qu.: 764   3rd Qu.:5519   3rd Qu.:1465
                   Max.   :1220   Max.   :5781   Max.   :1783

     X2B             X3B              HR             BB
Min.   :  1.0   Min.   :  0.00   Min.   :  0.0   Min.   :  1.0
1st Qu.:194.0   1st Qu.: 29.00   1st Qu.: 45.0   1st Qu.:425.8
Median :234.0   Median : 40.00   Median :110.0   Median :494.0
Mean   :228.7   Mean   : 45.67   Mean   :105.9   Mean   :473.6
3rd Qu.:272.0   3rd Qu.: 59.00   3rd Qu.:155.0   3rd Qu.:554.2
Max.   :376.0   Max.   :150.00   Max.   :307.0   Max.   :835.0
                                                 NA's   :1
      SO              SB              CS             HBP
Min.   :   3.0   Min.   :  1.0   Min.   :  3.00   Min.   :  7.00
1st Qu.: 516.0   1st Qu.: 62.5   1st Qu.: 33.00   1st Qu.: 32.00
Median : 761.0   Median : 93.0   Median : 44.00   Median : 43.00
Mean   : 762.1   Mean   :109.4   Mean   : 46.55   Mean   : 45.82
3rd Qu.: 990.0   3rd Qu.:137.0   3rd Qu.: 56.00   3rd Qu.: 57.00
Max.   :1596.0   Max.   :581.0   Max.   :191.00   Max.   :160.00
NA's   :16       NA's   :126     NA's   :832      NA's   :1158
      SF              RA             ER              ERA
Min.   : 7.00   Min.   :  34   Min.   :  23.0   Min.   :1.220
1st Qu.:38.00   1st Qu.: 610   1st Qu.: 503.0   1st Qu.:3.370
Median :44.00   Median : 689   Median : 594.0   Median :3.840
Mean   :44.11   Mean   : 681   Mean   : 573.4   Mean   :3.841
3rd Qu.:50.00   3rd Qu.: 766   3rd Qu.: 671.0   3rd Qu.:4.330
Max.   :77.00   Max.   :1252   Max.   :1023.0   Max.   :8.000
NA's   :1541
      CG              SHO              SV            IPouts
Min.   :  0.00   Min.   : 0.000   Min.   : 0.00   Min.   : 162
1st Qu.:  9.00   1st Qu.: 6.000   1st Qu.:10.00   1st Qu.:4080
Median : 41.00   Median : 9.000   Median :25.00   Median :4252
```

```
Mean   : 47.55   Mean   : 9.588   Mean   :24.42   Mean   :4013
3rd Qu.: 76.00   3rd Qu.:12.000   3rd Qu.:39.00   3rd Qu.:4341
Max.   :148.00   Max.   :32.000   Max.   :68.00   Max.   :4518

      HA             HRA             BBA             SOA
Min.   :  49   Min.   :  0.0   Min.   :  1.0   Min.   :   0.0
1st Qu.:1287   1st Qu.: 51.0   1st Qu.:429.0   1st Qu.: 511.0
Median :1389   Median :113.0   Median :495.0   Median : 762.0
Mean   :1339   Mean   :105.9   Mean   :473.7   Mean   : 761.6
3rd Qu.:1468   3rd Qu.:153.0   3rd Qu.:554.0   3rd Qu.: 997.0
Max.   :1993   Max.   :305.0   Max.   :827.0   Max.   :1687.0

      E               DP              FP             name
Min.   : 20.0   Min.   :  0.0   Min.   :0.7610   Length:2985
1st Qu.:111.0   1st Qu.:116.0   1st Qu.:0.9660   Class :character
Median :141.0   Median :140.0   Median :0.9770   Mode  :character
Mean   :180.8   Mean   :132.6   Mean   :0.9664
3rd Qu.:207.0   3rd Qu.:157.0   3rd Qu.:0.9810
Max.   :639.0   Max.   :217.0   Max.   :0.9910

    park             attendance          BPF             PPF
Length:2985        Min.   :      0   Min.   : 60.0   Min.   : 60.0
Class :character   1st Qu.: 538461   1st Qu.: 97.0   1st Qu.: 97.0
Mode  :character   Median :1190886   Median :100.0   Median :100.0
                   Mean   :1376599   Mean   :100.2   Mean   :100.2
                   3rd Qu.:2066598   3rd Qu.:103.0   3rd Qu.:103.0
                   Max.   :4483350   Max.   :129.0   Max.   :141.0
                   NA's   :279
  teamIDBR         teamIDlahman45     teamIDretro
Length:2985        Length:2985        Length:2985
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

```
mean(Teams$R, na.rm=TRUE)
```

```
[1] 681.0308
```

```r
mean(Teams$W, na.rm=TRUE)
```

```
[1] 74.61106
```

Question 14

r^2 =

0.7851

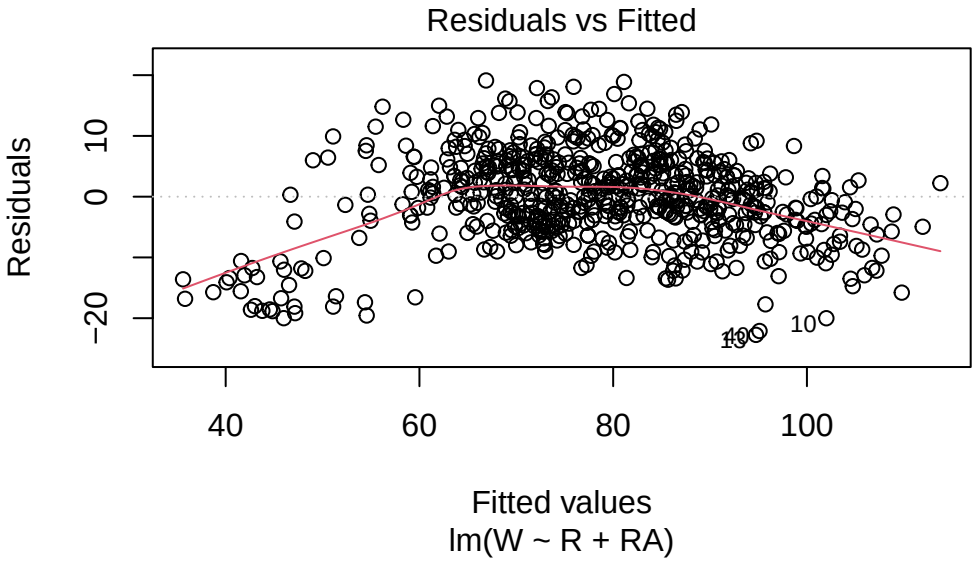
equation = 0.140288r -0.064911ra + 24.429271

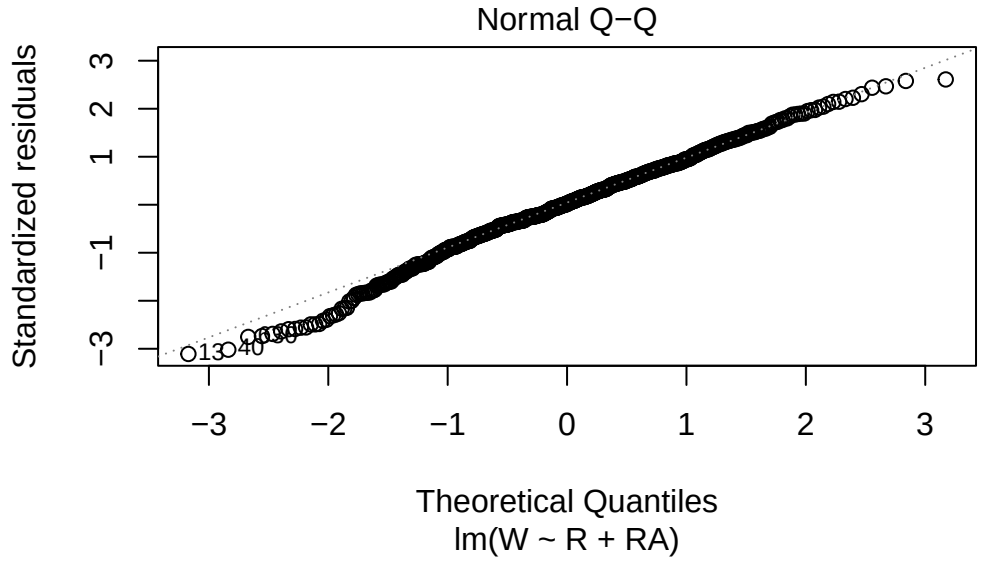
This equation compared to the last one has a higher r sqaured meaning it has smaller differences between the observed data and the fitted values thus it is a better linear regression model for our data.

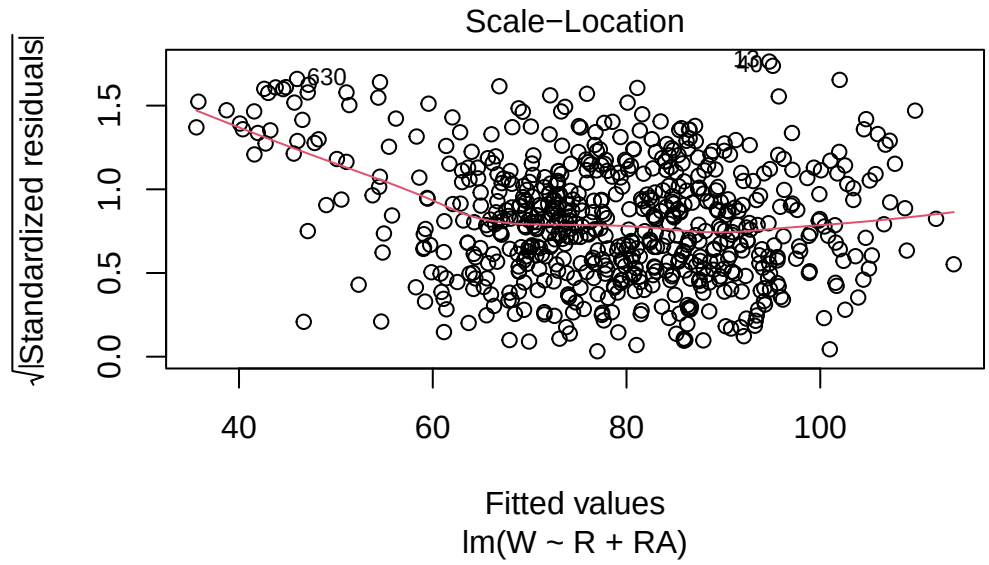
```r
library(ggplot2)

l<-f1 %>%
  lm(W~R + RA,.)
summary(l)
```

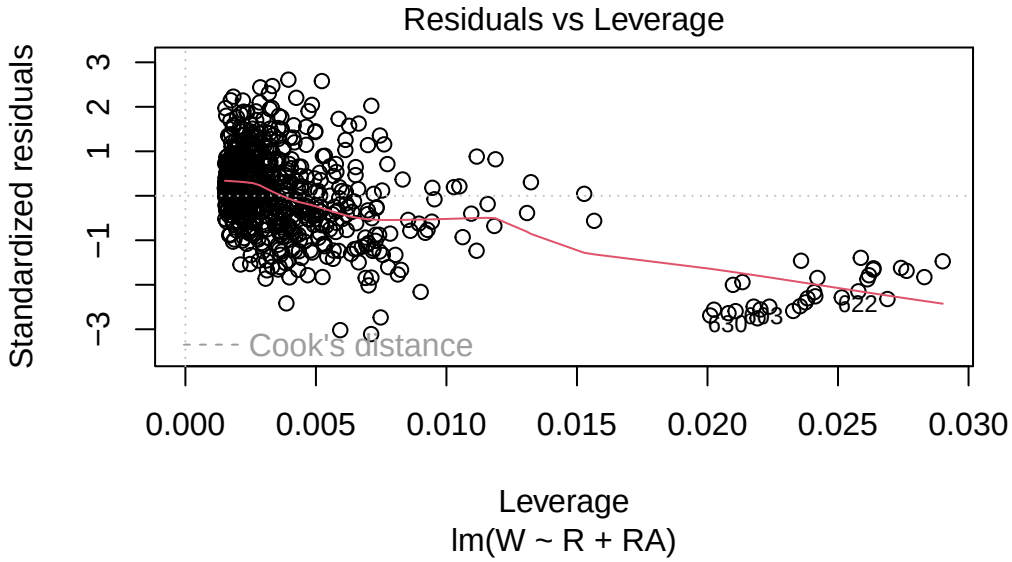
```
Call:
lm(formula = W ~ R + RA, data = .)

Residuals:
     Min       1Q   Median       3Q      Max
-22.7436  -4.2996   0.1466   4.9615  19.1250

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.429271   1.807148   13.52   <2e-16 ***
R            0.140288   0.002915   48.12   <2e-16 ***
RA          -0.064911   0.002821  -23.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.342 on 657 degrees of freedom
Multiple R-squared:  0.7851,    Adjusted R-squared:  0.7844
F-statistic:  1200 on 2 and 657 DF,  p-value: < 2.2e-16
```

```
plot(l)
```


Residuals vs Fitted
Residuals
10
0
−20
40
60
80
100
Fitted values
lm(W ~ R + RA)


Normal Q−Q
Standardized residuals
3
2
1
−1
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(W ~ R + RA)





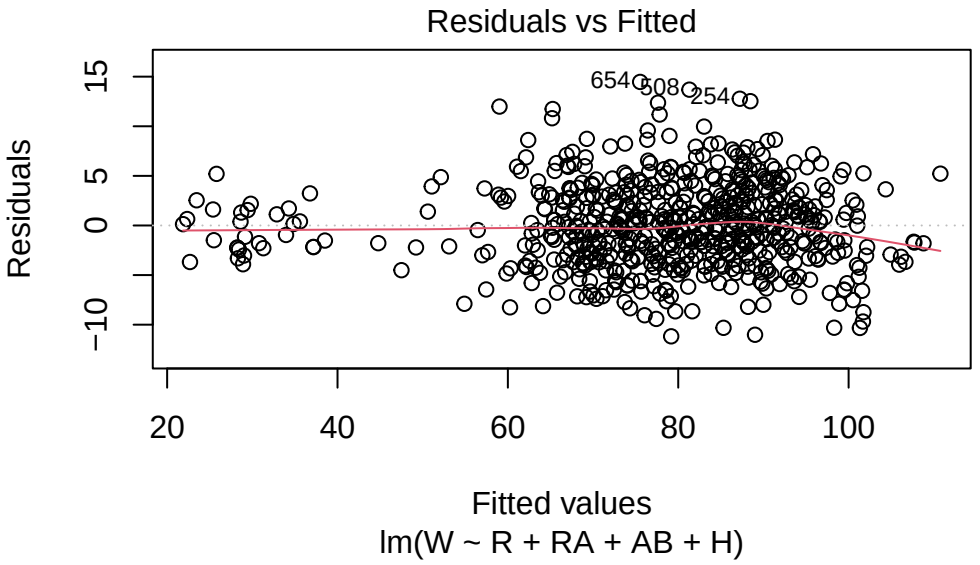
Question 15

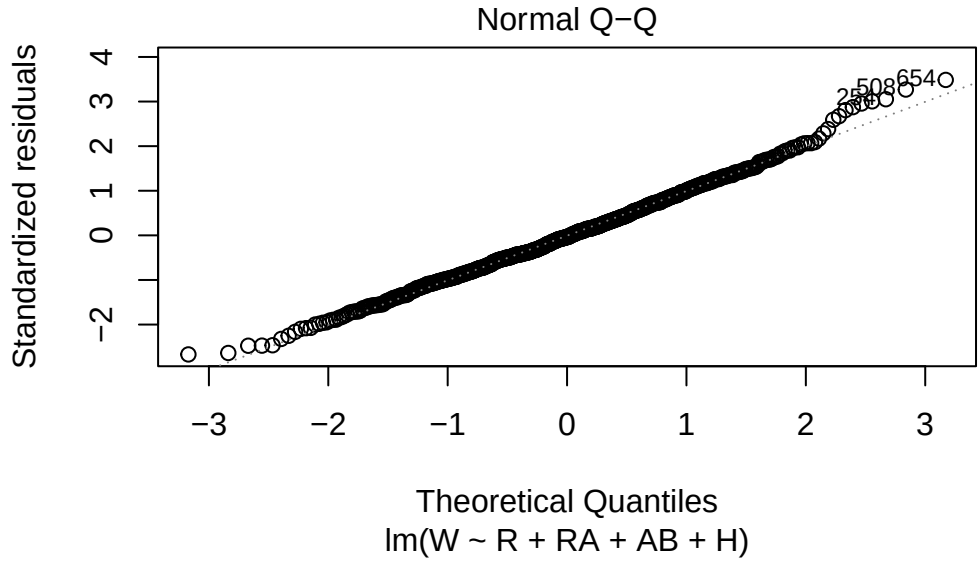
r^ value = 0.9302 becomes almost close to one which means the predictions are almost identical to the observed value, thus I think this model predicts wins better due to the higher r ^2 value.

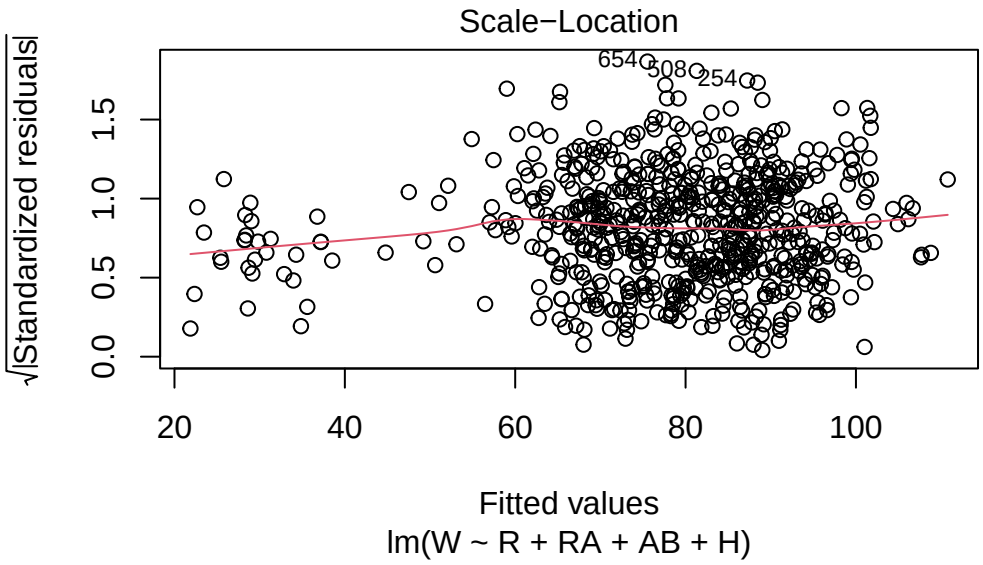
```r
library(ggplot2)

l<-f1 %>%
  lm(W~R + RA + AB + H,.)
summary(l)
```

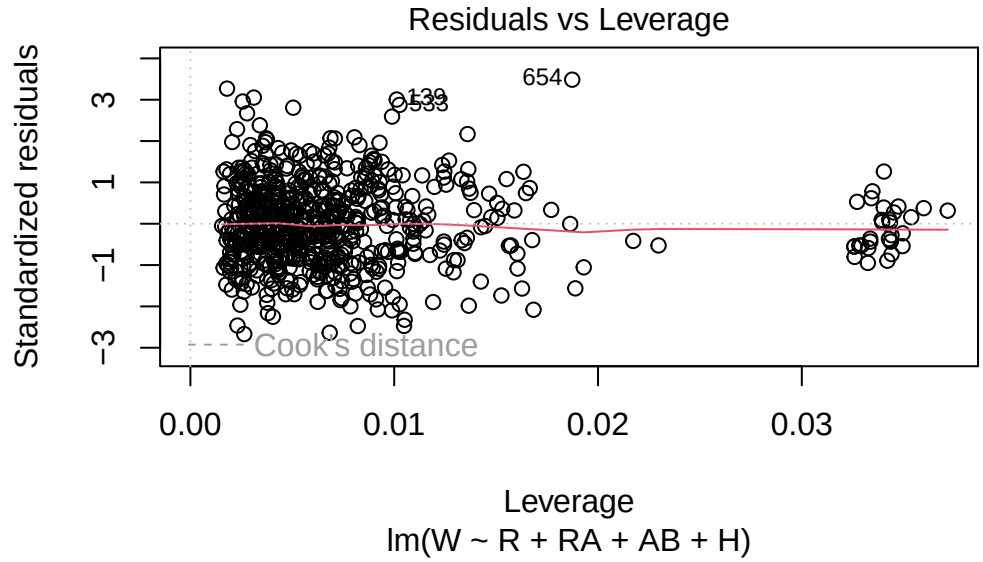
```
Call:
lm(formula = W ~ R + RA + AB + H, data = .)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1741  -2.8465  -0.1601   2.7844  14.4693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.075190   1.212979   1.711 0.087588 .
R            0.097121   0.002949  32.935  < 2e-16 ***
RA          -0.102488   0.001920 -53.384  < 2e-16 ***
AB           0.018254   0.000756  24.146  < 2e-16 ***
H           -0.012677   0.003312  -3.828 0.000142 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.191 on 655 degrees of freedom
Multiple R-squared:  0.9302,    Adjusted R-squared:  0.9298
F-statistic:  2182 on 4 and 655 DF,  p-value: < 2.2e-16
```

```r
plot(l)
```


Residuals vs Fitted
Residuals
15
5
0
−10
654
508
254
20
40
60
80
100
Fitted values
lm(W ~ R + RA + AB + H)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−2
654
508
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(W ~ R + RA + AB + H)


Scale−Location
654
508
254
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
40
60
80
100
Fitted values
lm(W ~ R + RA + AB + H)


Residuals vs Leverage
654
139
533
Standardized residuals
−3
−1
1
3
Cook's distance
0.00
0.01
0.02
0.03
Leverage
lm(W ~ R + RA + AB + H)