Final Project

Haorui Zhang

5/1/2019

```
library(rvest)
library(dplyr)
library(stringr)
```

Part I: Data Web Scraping

```
teampage <- read_html("http://www.baseball-reference.com/teams/")
fran_name <- teampage %>% html_nodes(".left") %>% html_text()
fran_name <- fran_name[2:31]
s <- html_session("http://www.baseball-reference.com/teams/")
baseball <- data.frame()
```

```
for(i in 1:length(fran_name)) {  #length(fran_name)
  hist <- s %>% follow_link(fran_name[i]) %>% read_html()
  sub_tb <- as.data.frame(hist %>% html_nodes("#franchise_years") %>% html_table())
  sub_tb['Team'] <- fran_name[i]
  baseball <- rbind(baseball, sub_tb)
}
```

```
all.equal(charToRaw(baseball$Tm[1]), charToRaw("Arizona Diamondbacks"))
```

```
## [1] "Lengths (21, 20) differ (comparison on first 20 components)"
## [2] "13 element mismatches"
```

```
char_cols <- which(lapply(baseball, typeof) == "character")

for(i in char_cols){
    baseball[[i]] <- str_conv(baseball[[i]], "UTF-8")
    baseball[[i]] <- str_replace_all(baseball[[i]],"\\s"," ")
}

all.equal(charToRaw(baseball$Tm[1]), charToRaw("Arizona Diamondbacks"))
```

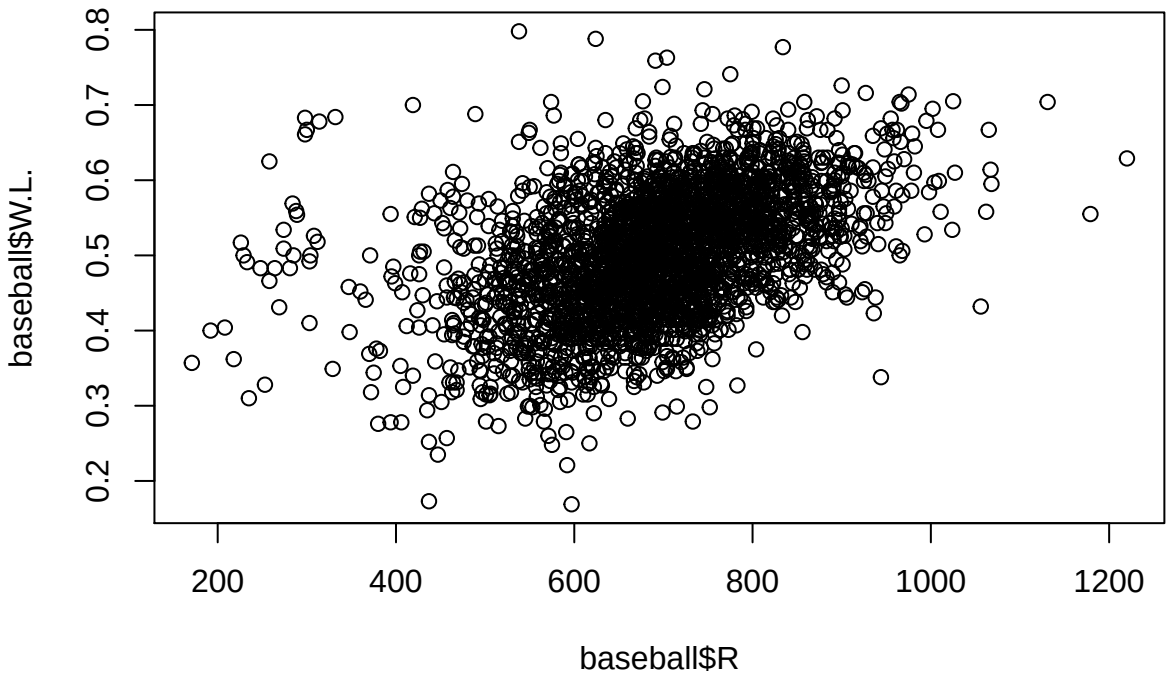
```
## [1] TRUE
```

```
dim(baseball)
```

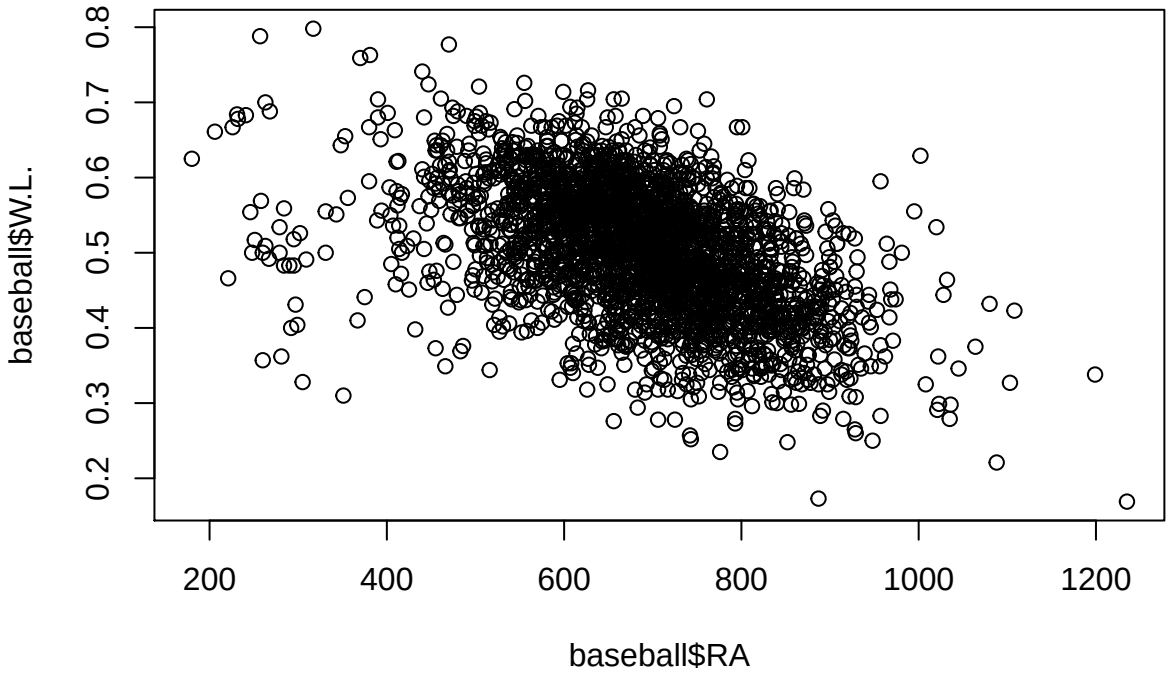
```
## [1] 2684   22
```

Part II: Exploratory Data Analysis

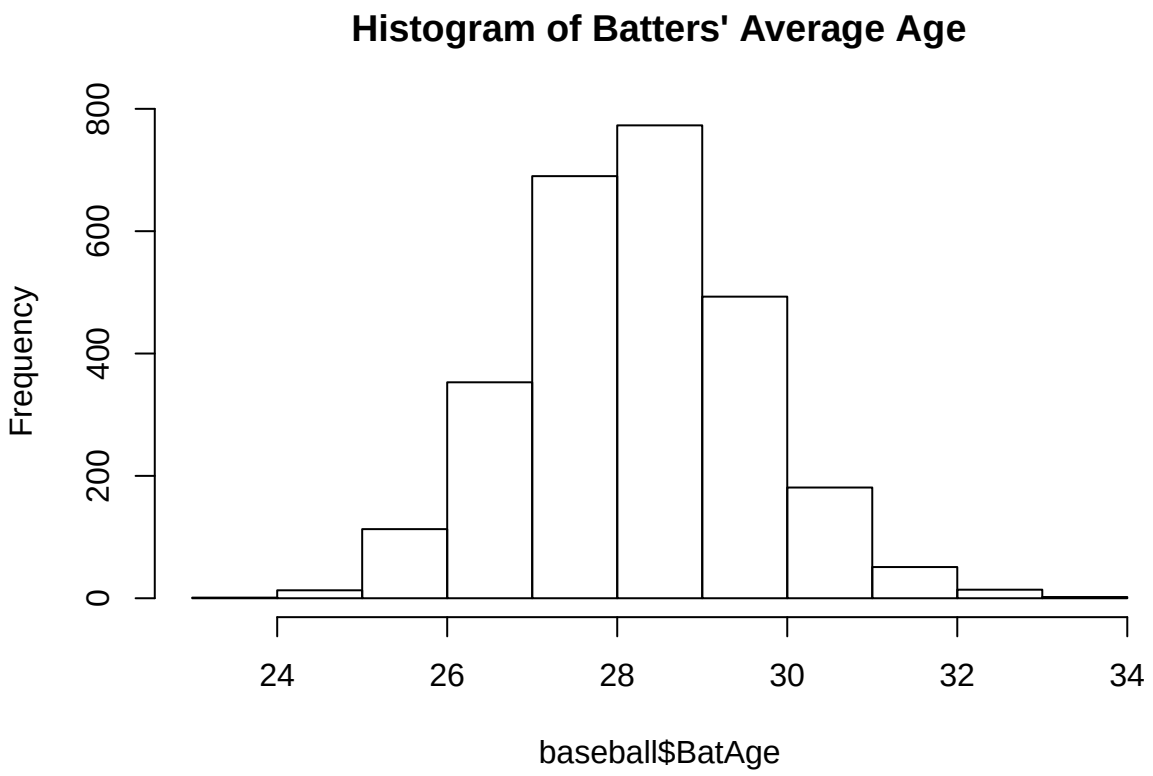
```
plot(baseball$R, baseball$W.L.)
```



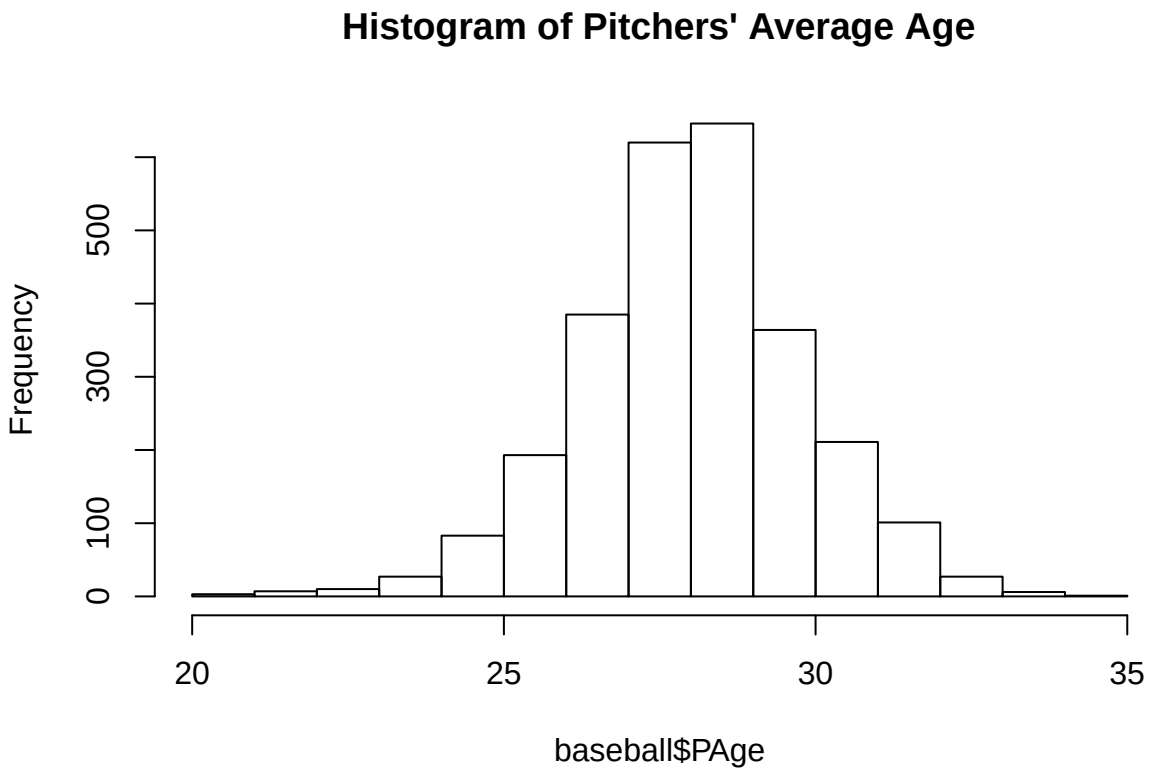

```
plot(baseball$RA, baseball$W.L.)
```




```
hist(baseball$BatAge, main = "Histogram of Batters' Average Age")
```




```
hist(baseball$PAge, main = "Histogram of Pitchers' Average Age")
```



Part III: Feature Engineering

```
# Use data from 1969-2018
baseball <- baseball %>% filter(Year %in% 1969:2018)
baseball <- subset(baseball, select = -Tm)
baseball$GB[which(baseball$GB == "--")] <- 0
baseball$GB <- as.integer(baseball$GB)
# Categorize attendance to indicate popularity
baseball$Attendance <- as.numeric(gsub("," ,"", baseball$Attendance))
baseball$Popularity <- cut(baseball$Attendance, breaks=c(-Inf, 1000000, 2000000, 3000000, Inf),
                           labels=c("Very unpopular", "Unpopular", "Popular", "Very popular"))
baseball$Lg <- as.factor(baseball$Lg)
baseball$Popularity <- as.factor(baseball$Popularity)
```

Part IV: Building Linear Models

```
# Start with a full model
m1 <- lm(W.L. ~ Lg + GB + R + RA + Popularity + BatAge + PAge + X.Bat + X.P, data = baseball)
summary(m1)
```

```
##
## Call:
## lm(formula = W.L. ~ Lg + GB + R + RA + Popularity + BatAge +
##     PAge + X.Bat + X.P, data = baseball)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.072677 -0.014001  0.000527  0.013472  0.070660
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.614e-01  1.651e-02  34.006  < 2e-16 ***
## LgAL East                6.393e-03  2.347e-03   2.724 0.006531 **
## LgAL West                1.404e-03  2.367e-03   0.593 0.553197
## LgNL Central            -9.083e-04  2.646e-03  -0.343 0.731406
## LgNL East                3.482e-03  2.374e-03   1.467 0.142695
## LgNL West                1.091e-03  2.414e-03   0.452 0.651476
## GB                      -2.139e-03  8.716e-05 -24.540  < 2e-16 ***
## R                        4.266e-04  1.078e-05  39.580  < 2e-16 ***
## RA                      -4.204e-04  1.115e-05 -37.719  < 2e-16 ***
## PopularityUnpopular      7.091e-03  2.262e-03   3.135 0.001755 **
## PopularityPopular        1.068e-02  2.544e-03   4.198 2.87e-05 ***
## PopularityVery popular   1.133e-02  3.085e-03   3.674 0.000248 ***
## BatAge                  -2.460e-04  5.250e-04  -0.469 0.639362
## PAge                    -7.817e-04  4.673e-04  -1.673 0.094620 .
## X.Bat                   -5.241e-04  2.220e-04  -2.361 0.018389 *
## X.P                      2.475e-04  2.689e-04   0.920 0.357511
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02105 on 1362 degrees of freedom
```

```
## Multiple R-squared:  0.9108, Adjusted R-squared:  0.9098
## F-statistic: 927.3 on 15 and 1362 DF,  p-value: < 2.2e-16
```

```
# Remove insignificant predictors
m2 <- lm(W.L. ~ Lg + GB + R + RA + Popularity + PAge + X.Bat, data = baseball)
summary(m2)
```

```
##
## Call:
## lm(formula = W.L. ~ Lg + GB + R + RA + Popularity + PAge + X.Bat,
##     data = baseball)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.072038 -0.014100  0.000491  0.013538  0.070582
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.537e-01  1.309e-02  42.313  < 2e-16 ***
## LgAL East                6.015e-03  2.318e-03   2.595 0.009554 **
## LgAL West                1.062e-03  2.333e-03   0.455 0.648891
## LgNL Central            -8.649e-04  2.644e-03  -0.327 0.743672
## LgNL East                3.327e-03  2.358e-03   1.411 0.158513
## LgNL West                8.351e-04  2.387e-03   0.350 0.726560
## GB                      -2.134e-03  8.686e-05 -24.565  < 2e-16 ***
## R                        4.276e-04  1.073e-05  39.846  < 2e-16 ***
## RA                      -4.202e-04  1.112e-05 -37.786  < 2e-16 ***
## PopularityUnpopular      7.001e-03  2.245e-03   3.119 0.001850 **
## PopularityPopular        1.053e-02  2.487e-03   4.232 2.47e-05 ***
## PopularityVery popular   1.103e-02  2.965e-03   3.720 0.000207 ***
## PAge                    -8.642e-04  4.315e-04  -2.003 0.045422 *
## X.Bat                   -3.472e-04  1.143e-04  -3.038 0.002428 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02105 on 1364 degrees of freedom
## Multiple R-squared:  0.9107, Adjusted R-squared:  0.9099
## F-statistic:  1071 on 13 and 1364 DF,  p-value: < 2.2e-16
```