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
```

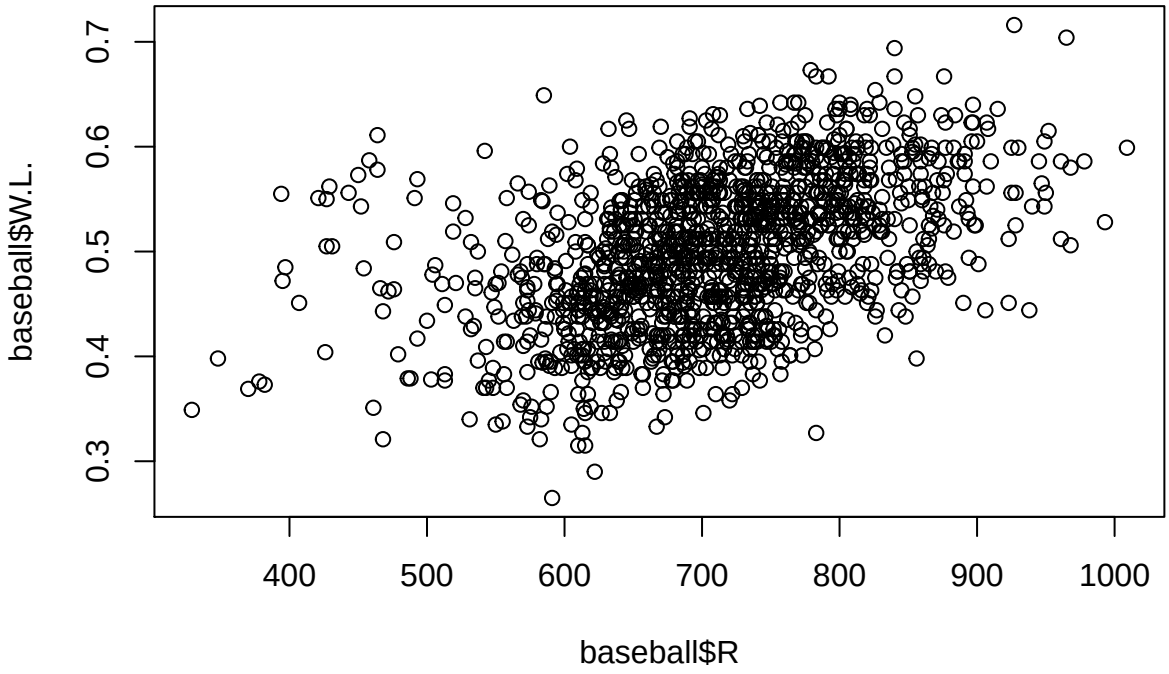
Part II: Data Cleaning

```
# Use data from 1969-2018 because data before 1969 have different leagues from now
baseball <- baseball %>% filter(Year %in% 1969:2018)
baseball <- subset(baseball, select = -Tm)
```

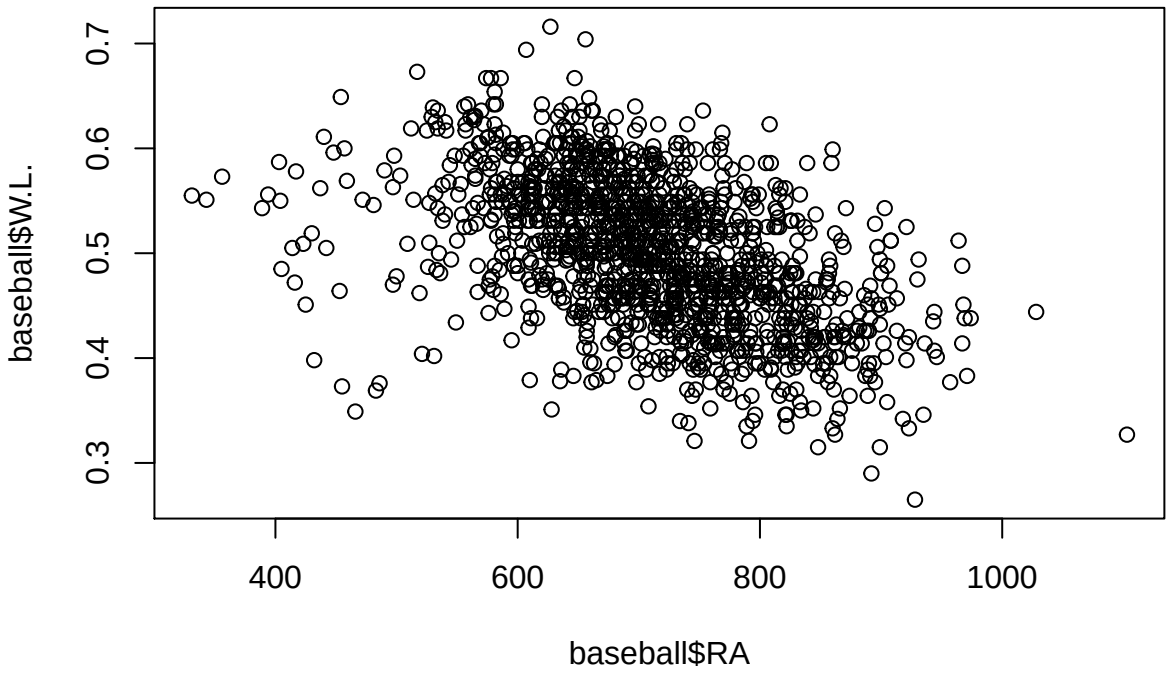
```
baseball$GB[which(baseball$GB == "--")] <- 0
baseball$GB <- as.integer(baseball$GB)
```

Part III: Exploratory Data Analysis

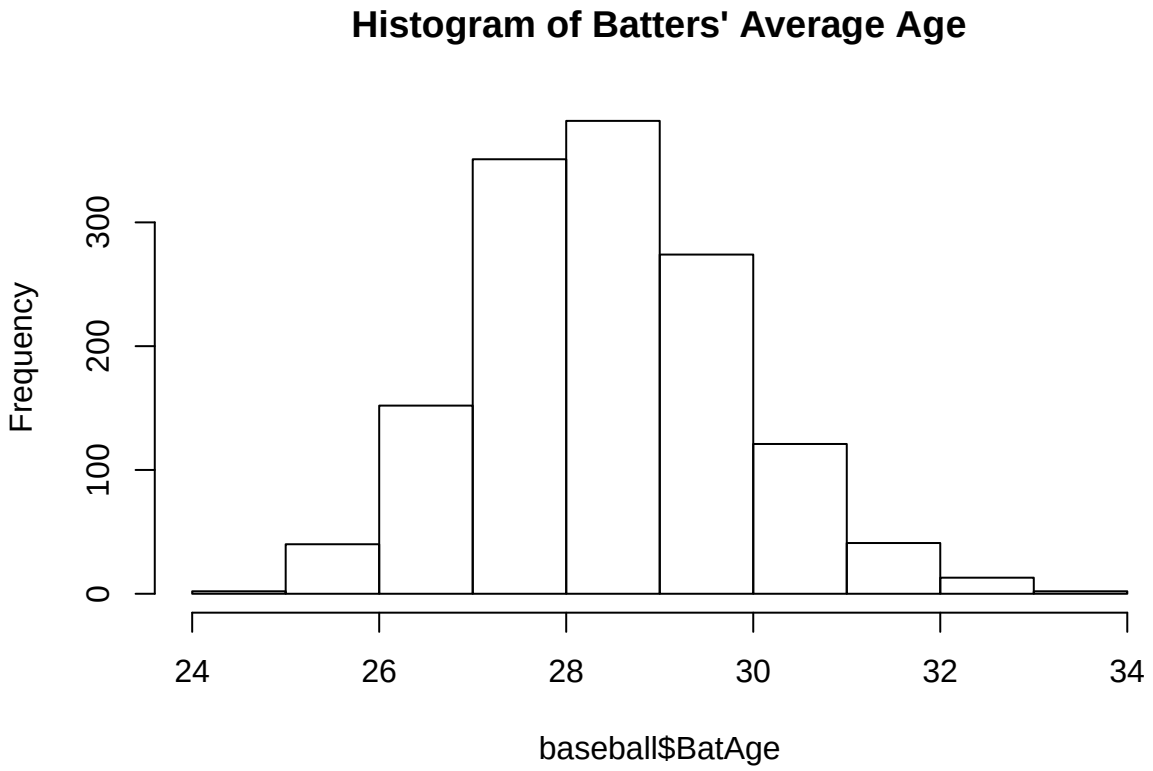
```
plot(baseball$R, baseball$W.L.)
```



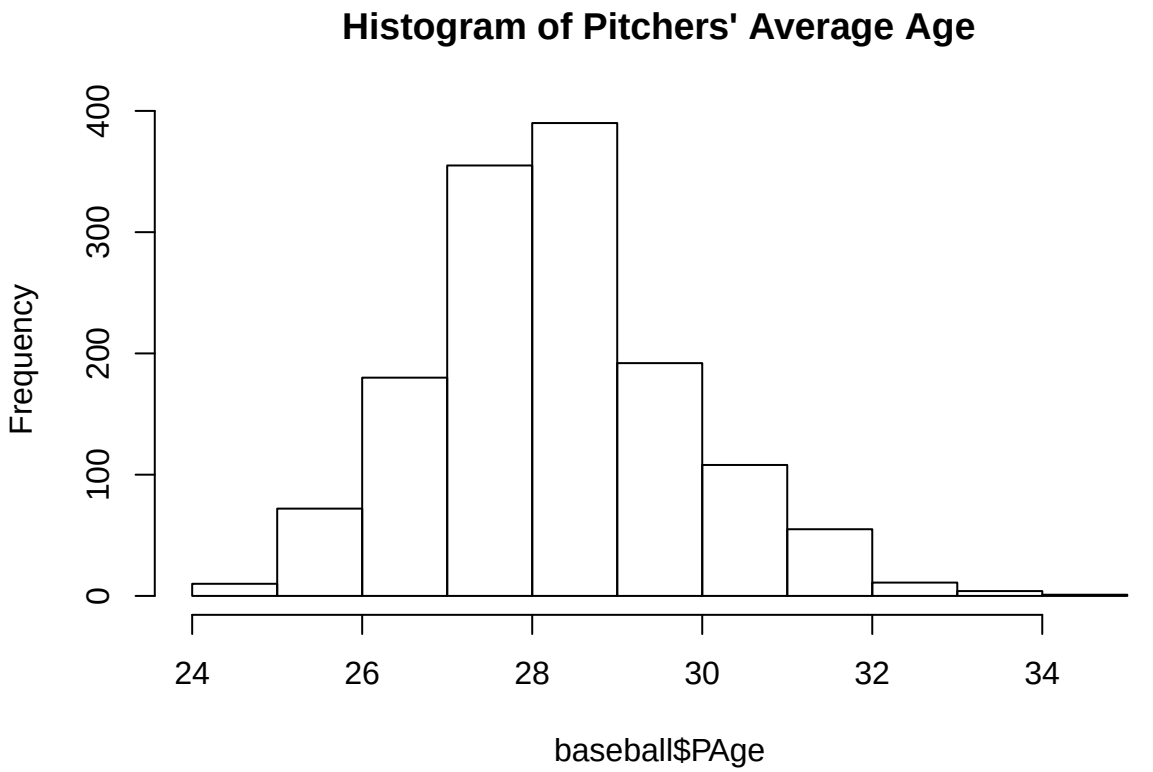

```
plot(baseball$RA, baseball$W.L.)
```



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

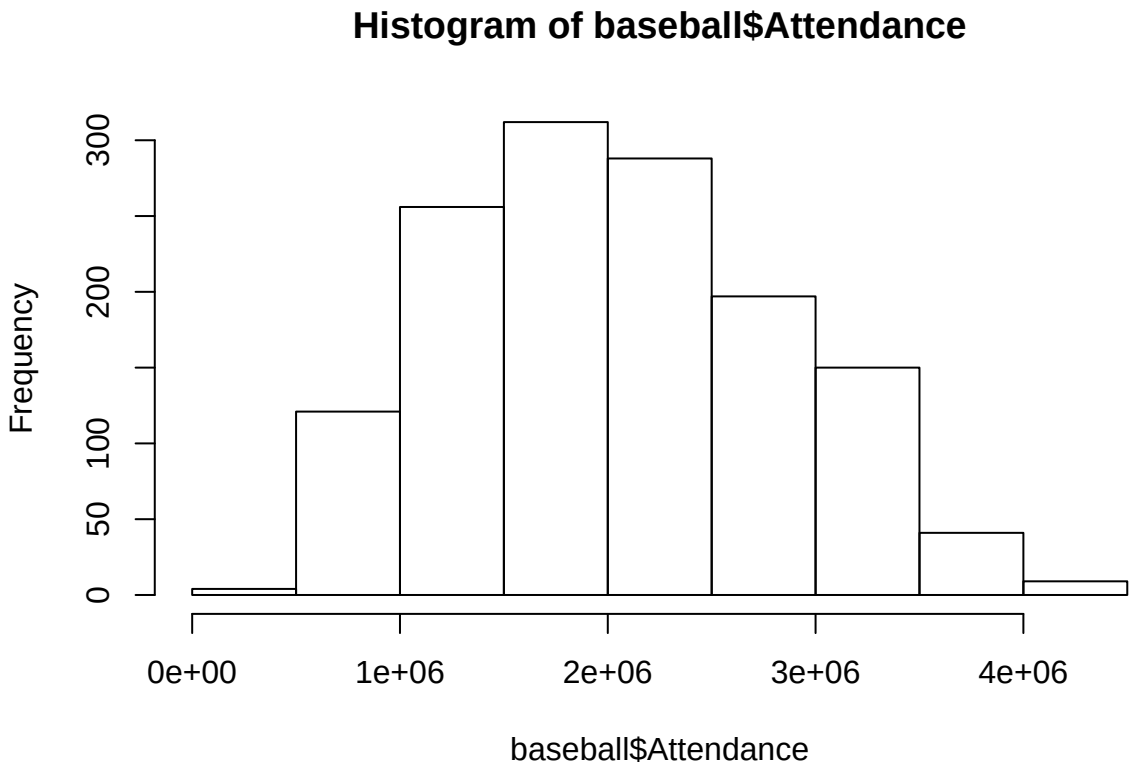



```r
hist(baseball$PAge, main = "Histogram of Pitchers' Average Age")
```



Part IV: Feature Engineering

```
# Categorize attendance to indicate popularity
baseball$Attendance <- as.numeric(gsub("," ,"", baseball$Attendance))
# Draw histogram to decide cutoff points
hist(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)
baseball$W.L. <- baseball$W.L.*100
```

Part V: Building Linear Models

```
set.seed(418)
obs <- sample(1:nrow(baseball), nrow(baseball)*0.7)
train <- baseball[obs,]
test <- baseball[-obs,]

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

```
##
## Call:
## lm(formula = W.L. ~ Lg + GB + R + RA + Popularity + BatAge +
```

```
##     PAge + X.Bat + X.P, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5623 -1.3915  0.0614  1.3862  6.6888
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            55.760129   1.957266  28.489  < 2e-16 ***
## LgAL East               0.644601   0.280756   2.296  0.02190 *
## LgAL West               0.228146   0.283342   0.805  0.42091
## LgNL Central           -0.103211   0.316912  -0.326  0.74474
## LgNL East               0.263466   0.284606   0.926  0.35483
## LgNL West               0.211968   0.288738   0.734  0.46306
## GB                     -0.215005   0.010610 -20.264  < 2e-16 ***
## R                       0.042448   0.001297  32.724  < 2e-16 ***
## RA                     -0.042714   0.001342 -31.830  < 2e-16 ***
## PopularityUnpopular     0.591215   0.274777   2.152  0.03168 *
## PopularityPopular       0.876946   0.308056   2.847  0.00451 **
## PopularityVery popular  0.979468   0.370973   2.640  0.00842 **
## BatAge                  0.017739   0.062399   0.284  0.77626
## PAge                   -0.094488   0.055066  -1.716  0.08651 .
## X.Bat                  -0.046355   0.025679  -1.805  0.07136 .
## X.P                     0.031143   0.031981   0.974  0.33040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.093 on 948 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9088
## F-statistic: 640.6 on 15 and 948 DF,  p-value: < 2.2e-16
```

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

```
##
## Call:
## lm(formula = W.L. ~ Lg + GB + R + RA + Popularity + PAge + X.Bat,
##     data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.459 -1.370  0.039  1.396  6.663
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            55.706785   1.552354  35.885  < 2e-16 ***
## LgAL East               0.606835   0.276243   2.197  0.02828 *
## LgAL West               0.177490   0.277934   0.639  0.52323
## LgNL Central           -0.106884   0.316499  -0.338  0.73566
## LgNL East               0.230553   0.282472   0.816  0.41459
## LgNL West               0.161129   0.283900   0.568  0.57047
## GB                     -0.214902   0.010571 -20.329  < 2e-16 ***
## R                       0.042541   0.001292  32.929  < 2e-16 ***
## RA                     -0.042642   0.001339 -31.841  < 2e-16 ***
```

```
## PopularityUnpopular     0.602296   0.272902   2.207  0.02755 *
## PopularityPopular       0.903296   0.301138   3.000  0.00277 **
## PopularityVery popular  1.019119   0.358246   2.845  0.00454 **
## PAge                   -0.087964   0.050701  -1.735  0.08307 .
## X.Bat                  -0.025171   0.013480  -1.867  0.06216 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.092 on 950 degrees of freedom
## Multiple R-squared:  0.9101, Adjusted R-squared:  0.9089
## F-statistic: 739.9 on 13 and 950 DF,  p-value: < 2.2e-16
```

```
pred <- predict(m2, subset(test, select = -W.L.))
cor(pred, test$W.L.)
```

```
## [1] 0.9546554
```