DATA606_FALL_2018_FINAL_PROJECT_PROPOSAL

John K. Hancock

October 29, 2018

Data Preparation

```
mlb_df <- read.csv('DATA/MLB_PITCHING_STATS_1998_to_2017.csv')
head(mlb_df, 10)
```

```
##         Team   W  L  ERA   SO WHIP  FIP
## 1     Braves 106 56 3.25 1232 1.22 3.53
## 2     Astros 102 60 3.50 1187 1.29 3.86
## 3     Padres  98 64 3.63 1217 1.30 3.84
## 4       Mets  88 74 3.77 1129 1.31 4.18
## 5    Dodgers  83 79 3.81 1178 1.33 4.06
## 6    Yankees 114 48 3.82 1080 1.25 4.15
## 7    Pirates  69 93 3.91 1112 1.35 4.08
## 8     Giants  89 74 4.19 1089 1.37 4.42
## 9    Red Sox  92 70 4.19 1025 1.33 4.40
## 10 Blue Jays  88 74 4.29 1154 1.39 4.36
```

```
mlb_df[mlb_df$W >= 100, "Record"] <- "100 wins or Above"
mlb_df[mlb_df$W >= 82  & mlb_df$W < 100, "Record"] <- "Above .500 and Less than 100 Wins"
mlb_df[mlb_df$W < 82 & mlb_df$W > 50, "Record"] <- "Below .500 but above 50 wins"
mlb_df[mlb_df$W <= 50, "Record"] <- "Less than 50 wins"
mlb_df$Record <- as.factor(mlb_df$Record)
```

```
mlb_df$League <- case_when(mlb_df$Team == 'Braves'~'National League',
 mlb_df$Team == 'Marlins'~'National League',
 mlb_df$Team == 'Mets'~'National League',
 mlb_df$Team == 'Phillies'~'National League',
 mlb_df$Team == 'Nationals'~'National League',
 mlb_df$Team == 'Cubs'~'National League',
 mlb_df$Team == 'Reds'~'National League',
 mlb_df$Team == 'Brewers'~'National League',
 mlb_df$Team == 'Pirates'~'National League',
 mlb_df$Team == 'Cardinals'~'National League',
 mlb_df$Team == 'iamondbacks'~'National League',
 mlb_df$Team == 'Rockies'~'National League',
 mlb_df$Team == 'Dodgers'~'National League',
 mlb_df$Team == 'Padres'~'National League',
 mlb_df$Team == 'Giants'~'National League',
 TRUE ~ as.character('American League'))
```

```
rows <- as.numeric(row.names.data.frame(mlb_df[mlb_df$Team == 'Astros',]))
rows <- rows[rows < 480]
mlb_df$League[ rows] <- "National League"
```

```
mlb_df[mlb_df$Team == 'Astros', ]
```

```
##       Team   W   L  ERA   SO WHIP  FIP                          Record
## 2   Astros 102  60 3.50 1187 1.29 3.86                 100 wins or Above
## 33  Astros  97  65 3.84 1204 1.35 3.69 Above .500 and Less than 100 Wins
## 88  Astros  72  90 5.42 1064 1.53 5.14      Below .500 but above 50 wins
## 106 Astros  93  69 4.39 1228 1.33 4.48 Above .500 and Less than 100 Wins
## 132 Astros  84  78 4.00 1219 1.36 3.88 Above .500 and Less than 100 Wins
## 157 Astros  87  75 3.87 1139 1.32 4.23 Above .500 and Less than 100 Wins
## 187 Astros  92  70 4.05 1282 1.35 4.07 Above .500 and Less than 100 Wins
## 212 Astros  89  73 3.52 1164 1.23 3.84 Above .500 and Less than 100 Wins
## 245 Astros  82  80 4.09 1160 1.30 4.27 Above .500 and Less than 100 Wins
## 291 Astros  73  89 4.70 1109 1.42 4.73      Below .500 but above 50 wins
## 317 Astros  86  75 4.39 1095 1.36 4.53 Above .500 and Less than 100 Wins
## 354 Astros  74  88 4.54 1144 1.45 4.35      Below .500 but above 50 wins
## 376 Astros  76  86 4.09 1210 1.39 3.89      Below .500 but above 50 wins
## 418 Astros  56 106 4.51 1191 1.42 4.35      Below .500 but above 50 wins
## 445 Astros  55 107 4.57 1170 1.43 4.27      Below .500 but above 50 wins
## 480 Astros  51 111 4.79 1084 1.49 4.67      Below .500 but above 50 wins
## 505 Astros  70  92 4.14 1137 1.34 3.93      Below .500 but above 50 wins
## 516 Astros  86  76 3.57 1280 1.20 3.66 Above .500 and Less than 100 Wins
## 551 Astros  84  78 4.06 1396 1.29 3.85 Above .500 and Less than 100 Wins
## 581 Astros 101  61 4.12 1593 1.27 3.91                 100 wins or Above
##              League
## 2   National League
## 33  National League
## 88  National League
## 106 National League
## 132 National League
## 157 National League
## 187 National League
## 212 National League
## 245 National League
## 291 National League
## 317 National League
## 354 National League
## 376 National League
## 418 National League
## 445 National League
## 480 American League
## 505 American League
## 516 American League
## 551 American League
## 581 American League
```

Research question

Does a mlb team's key pitching statistics predict their win totals? If so, which pitching metric, ERA, Strikeouts, WHIP, or FIP provide the best predictor for wins?

Cases

The data above consists of each MLB team's wins, losses, ERA, SO, WHIP, and FIP for the past 20 MLB seasons from 1998 to 2017. The cutoff year of 1998 was chosen since that was the last year of MLB expansion. Thus, there are 600 cases.

Data Collection

Data was collected from the website FanGraphs.com. For the purpose of this academic exercise, I purchased a year's membership to the site and I was able to download the datasets needed for this project.

Type of study

The data is observational. Statistics are compiled from 30 teams over 20 years–1998 to 2018

Data Source

Fangraphs.com is website operated by FanGraphs, Inc. Fangraphs compiles historical statistical data for the entire history of Major League Baseball. In addition, it creates and records advanced baseball metrics outside of the established statistics. FanGraphs is well established as a chronicler and compiler of baseball statistics. It has parternership deals with ESPN and SB Nation. Link to website: FanGraphs.com (https://www.fangraphs.com/) Wikipedia (https://en.wikipedia.org/wiki/Fangraphs)

Dependent Variable

There are two response variables. One is Wins which is numeric (quantative variable), and the other is Record which is qualitative.

Independent Variable

The independent variables are ERA, Strikeouts("SO"), Walks, Hits to Innings Pitched ("WHIP), Fielded Independent Pitching ("FIP") which are all numeric. The qualitative variable is League.

```
describe(mlb_df$ERA)
```

```
## mlb_df$ERA
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      600        0      205        1     4.28   0.6167     3.43     3.58
##      .25      .50      .75      .90      .95
##     3.89     4.24     4.66     5.01     5.20
##
## lowest : 2.94 3.02 3.03 3.13 3.15, highest: 5.56 5.67 5.69 5.71 6.03
```

```
describe(mlb_df$SO)
```

```
## mlb_df$SO
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      600        0      363        1     1129      157      926      958
##      .25      .50      .75      .90      .95
##     1025     1120     1221     1311     1372
##
## lowest :  764  794  831  846  859, highest: 1549 1560 1580 1593 1614
```

```
describe(mlb_df$WHIP)
```

```
## mlb_df$WHIP
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      600        0       51    0.999    1.368   0.1096     1.22     1.24
##      .25      .50      .75      .90      .95
##     1.30     1.36     1.44     1.50     1.53
##
## lowest : 1.11 1.14 1.15 1.16 1.17, highest: 1.60 1.61 1.62 1.64 1.71
```

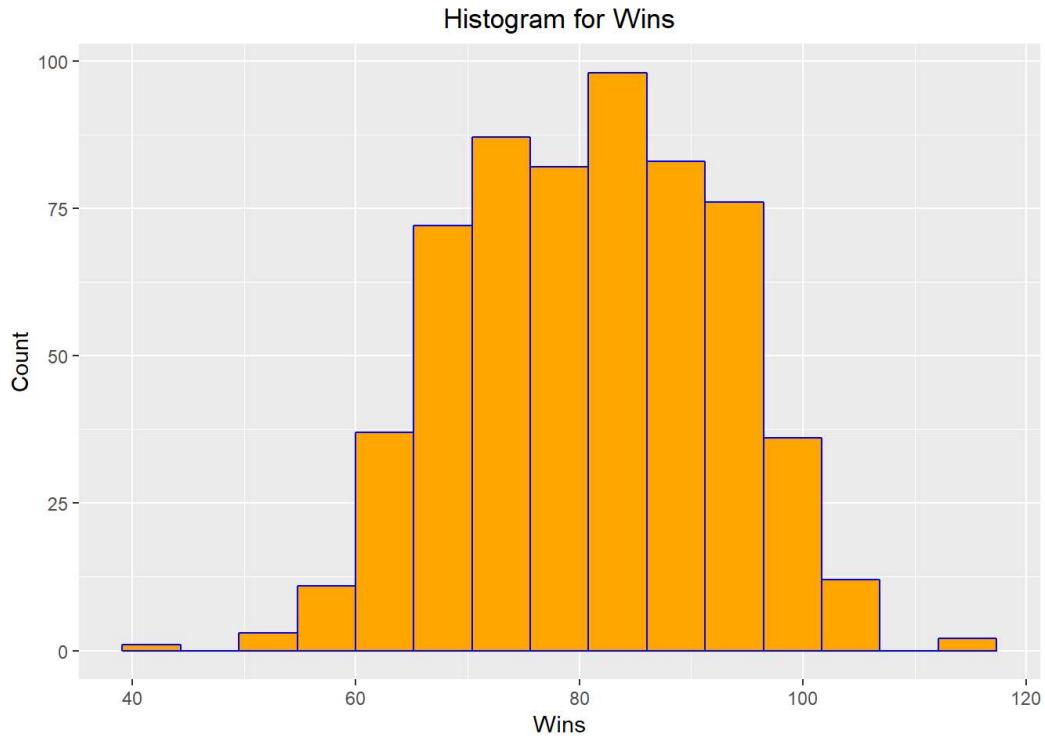
```
describe(mlb_df$FIP)
```

```
## mlb_df$FIP
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      600        0      179        1    4.279   0.4915    3.580    3.729
##      .25      .50      .75      .90      .95
##    3.970    4.280    4.570    4.841    5.010
##
## lowest : 3.18 3.24 3.27 3.30 3.33, highest: 5.29 5.31 5.46 5.52 5.54
```

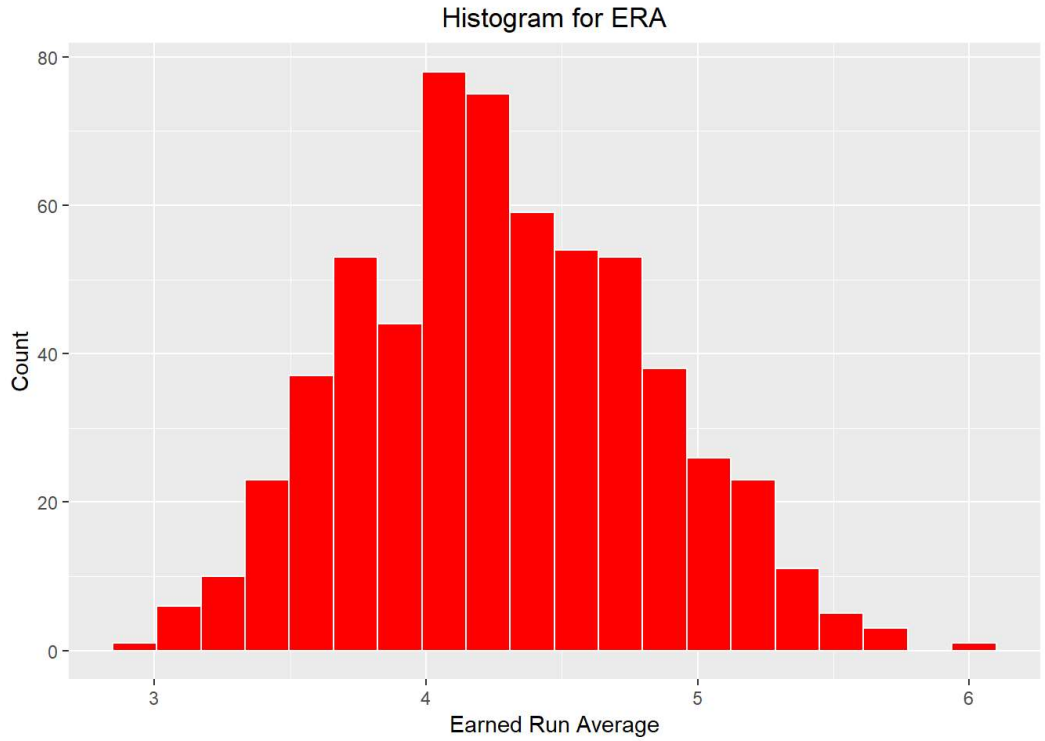
```
table(mlb_df$Record, useNA='ifany')
```

```
##
##                 100 wins or Above Above .500 and Less than 100 Wins
##                                  24                              270
##      Below .500 but above 50 wins                 Less than 50 wins
##                                 305                                1
```

```
ggplot(mlb_df, aes(x=W)) + geom_histogram(bins=15,
                col="blue",
                fill="orange") + labs(title="Histogram for Wins", align="center") +
                                 labs(x="Wins", y="Count")
```




```
ggplot(mlb_df, aes(x=ERA)) + geom_histogram(bins=20,
                col="white",
                fill="red") + labs(title="Histogram for ERA") +
                                 labs(x="Earned Run Average", y="Count")
```



```
ggplot(mlb_df, aes(x=FIP)) + geom_histogram(bins=15,
                 col="blue",
                 fill="orange") + labs(title="Histogram for Fielding Independent Pitching") +
                                    labs(x="FIP", y="Count")
```