Source: http://www. baseball-reference.com

Project member :

1) Tae Kyun Kim: SAS data output analyst
2) Hyun Jung Kim: project editing
3) Hwan Suk Lee: project design & mentor

Project Design:

Based on the record of 2008 Major league season, we assumed that good batters are batters who has AVG (batting average) higher than league average. Also, we assumed that good pitchers are pitchers who has ERA (earned runs average) lower than league average. From the record, we set AVG and ERA as our two variables. From these two variables, we compare each team and counts hitters who are above the entire season of AVG, and pitchers who are below the entire season of ERA. Then, we find out the relationship between wins and both AVG and ERA. We only considered batters who attended more than 100 times at bat. Also, we only considered pitchers who attended equal or more than 40 innings for same reason. Batters and hitters who have not been satisfied above these conditions are excluded due to the possibility to be outliers.

We analyze 2 correlations: one is batting AVG and WINS, and the other is between ERA and WINS Figure out each correlation to know which factor (ERA, AVG) affect more on the numbers of WINS.

i) Present scatter plot

ii) Present correlation coefficient

Prediction!

In the baseball game, at least the team is going to win at the game when the pitcher played well in the game. However, no matter how well hitters hit the ball; there are still possibilities that the team might lose the game. For those reasons, some people are saying that pitcher has more influence on the number of wins and teams reputation than the hitters.

In order to prove this hypothesis, we did our project following this prediction:

Higher correlation coefficient is more effective to number of WINS. According to our research, we got too low correlations that between number of wins and either our factors AVG or ERA have too weak relationships. So we infer there must be some other lurking variables that lower our results. To get more clear correlation, we added one more variable to each pitcher and hitter variable, and find out another correlation coefficient by putting in these two variables in our study: WHIP (Walks and Hits per Innings Pitched) that is related to pitchers' ability, and OPS (On-Base Percentage + Slugging Percentage) that is related to hitters' ability.

Pitcher who has lower WHIP tends not to load bases, so make fielders less pressure of defense. Thus, we assumed that the pitcher who has lower WHIP contribute team's win as same as

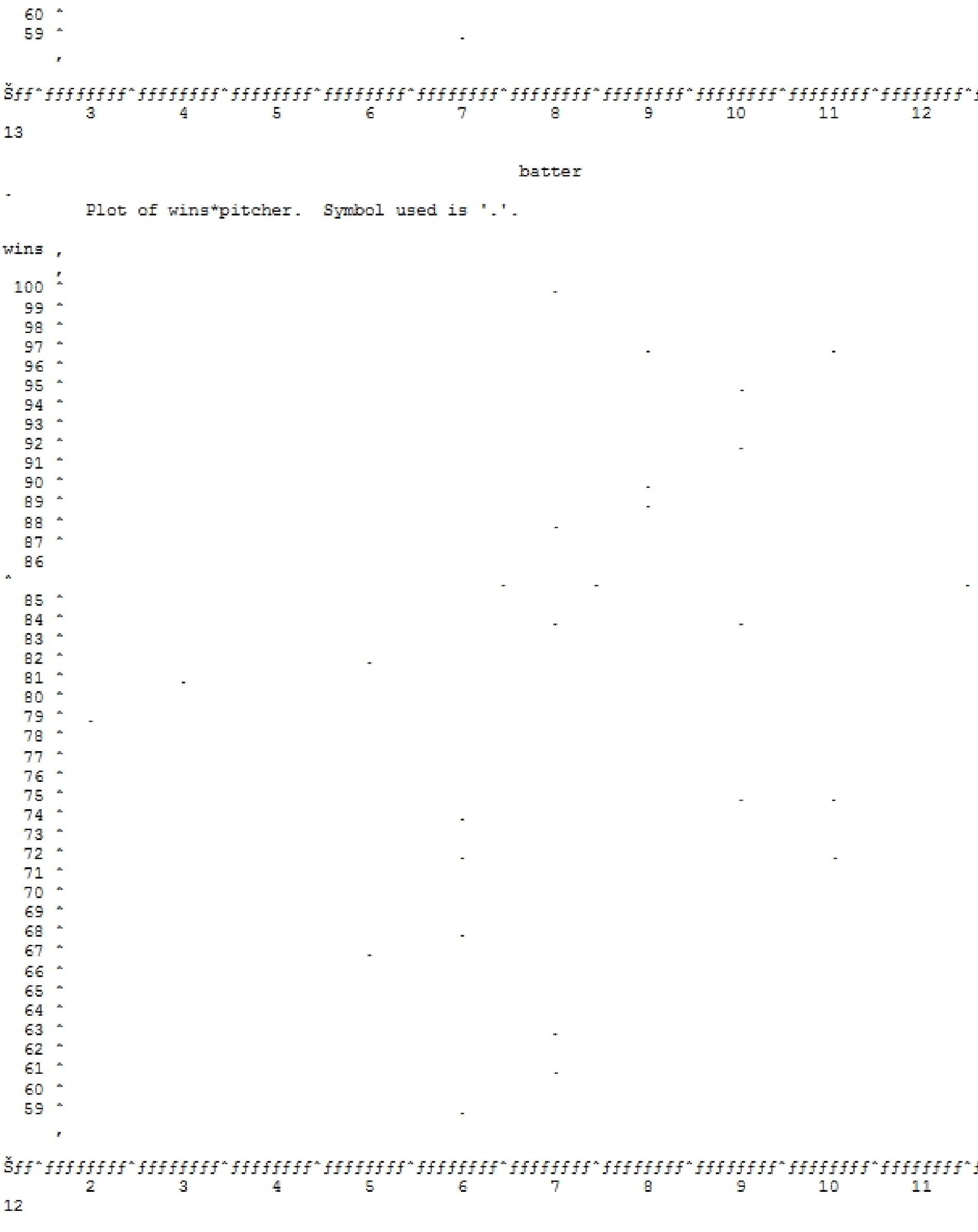
the pitcher who has lower ERA. On the other hand, OPS is generally used for defining power hitter. Of course, power hitter tends to make more extra base hits than just contact hitter and extra base hit has more possibility that runner makes score than just single base hit. Hence, power hitters have high slugging percentage as well. Also, for preventing extra base hit and power hitters usually hit so hard and lose their contact, pitchers threw more balls to outside of strike zone against power hitter than contact hitter, power hitters usually have more four-ball than contact hitters. Thus, we assumed that OPS, which is the sum of hitter's On Base Percentage and Slugging Percentage contributes for team's win. We only included WHIP that is above the season average and OPS that is below the season average.

First, we found out correlation between Average at bat (AVG) and number of Wins and also about Earned Runs Average (ERA) and Wins

SAS output indicate that two variables (AVG and ERA) do not show strong relationship with number of Wins.

```
The SAS System            08:54 Monday, April 27, 2009   1

                              Plot of wins*batter.  Symbol used is '.'.
wins ,
     ,
 100 ^                                                                .
  99 ^
  98 ^
  97 ^                                    .                   .
  96 ^
  95 ^                                                                                .
  94 ^
  93 ^
  92 ^                          .
  91 ^
  90 ^                          .
  89 ^                  .                       .         .
  88 ^                                                                .
  87 ^
  86 ^                          .                       .                         .
  85 ^
  84 ^                                    .                   .
  83 ^
  82 ^             .
  81 ^                                    .
  80 ^
  79
^                                                                                       .
  78 ^
  77 ^
  76 ^
  75 ^    .                       .
  74 ^                                .                   .
  73 ^
  72 ^                                                    .         .
  71 ^
  70 ^
  69 ^
  68 ^                                        .
  67 ^                                                    .
  66 ^
  65 ^
  64 ^
  63 ^                                                    .
  62 ^
  61 ^                                .
```



Batters and wins: point estimate r is 0.18041

Pitchers and wins: point estimate r is 0.33924