John K. Hancock
jkhancock@gmail.com
P2: Investigate a Data Set – Project Submission

The Immediate Impact of Integration on MLB -- A Data Analysis

Prior to 1947, Major League Baseball had a "gentlemen's agreement", wherein owners agreed to segregate the game based on the race of the players. In 1947, the Brooklyn Dodgers and the Cleveland Indians broke that agreement when they fielded two African American players, Jackie Robinson and Larry Doby. These moves ushered in integration into major league baseball. From 1947 onward, the game would no longer be restricted based on race. Ending discrimination meant that MLB would become a true representation of the best players in the world. However, not every team embraced integration early on like the Dodgers and the Indians. It would take 12 years after Jackie Robinson for every team to be integrated with at least one African American player. The table[1] below lists the racial demographics of MLB for the first 20 years after integration:

Year	White	African-Americans	Latino	Asian
1947	98.30%	0.90%	0.70%	0.00%
1948	98.50%	0.70%	0.70%	0.00%
1949	96.60%	1.50%	1.90%	0.00%
1950	95.30%	1.70%	3.00%	0.00%
1951	94.30%	2.90%	2.80%	0.00%
1952	94.40%	2.90%	2.70%	0.00%
1953	93.30%	3.70%	3.00%	0.00%
1954	90.70%	5.60%	3.70%	0.00%
1955	89.80%	5.20%	5.00%	0.00%
1956	88.20%	6.70%	5.10%	0.00%
1957	88.10%	6.70%	5.20%	0.00%
1958	86.70%	7.40%	5.90%	0.00%
1959	84.80%	8.80%	6.50%	0.00%
1960	82.30%	8.90%	8.90%	0.00%
1961	82.60%	9.70%	7.70%	0.00%
1962	81.90%	10.10%	8.00%	0.00%
1963	80.10%	11.70%	8.20%	0.00%
1964	79.30%	11.70%	8.90%	0.10%
1965	78.30%	12.70%	8.80%	0.10%
1966	76.90%	13.40%	9.70%	0.00%
1967	75.60%	13.60%	10.70%	0.00%

[1] "Baseball Demographics, 1947-2012", by Mark Armour and Daniel R. Levitt. (SABR Baseball Biography Project, **http://sabr.org/bioproj/topic/baseball-demographics-1947-2012**)

So, what was the impact of racial de-segregation on the game? Were there any real differences in key offensive and defensive statistics? How did the winning percentage of the team who integrated early compare with the team that was the last to integrate? Finally, how did the African American players perform in MLB?

This data analyst project will provide the following:

1. Compare hitting summary statistics for the 20 years prior to integration to the 20 years after integration
2. Looks at the correlation between the increase in HRs to the number of games
3. Look at Slugging Percentage over time
4. Compare pitching summary statistics for the 20 years prior to integration to the 20 years after integration
5. Look at WHIP (Walks + Hits /Innings Pitched) over time
6. Calculates correlations between WHIP/Strikeouts and HRs and Strikeouts
7. Compares the winning percentages between the Dodgers and Red Sox
8. Looks at the offensive performance of a sample of the first African American players.

Note (1): This project provides observation-data analysis, not a statistical analysis, The project will look at the data, but it does not create a null hypotheses to test for the statistical significance of the data.

Analysis
Setting up the Data Environment

A. I read the batting and pitching csv files from the Lahman data files into two pandas DataFrames.

B. From there, I created two DataFrames for the batting stats, one for the 20 years prior to integration and one for the 20 years post integration, and I did the same for the pitching stats.

C. Next, I verified that each DataFrame has the correct time frames.

John K. Hancock
jkhancock@gmail.com

Analysis

1. Summary statistics for hitting

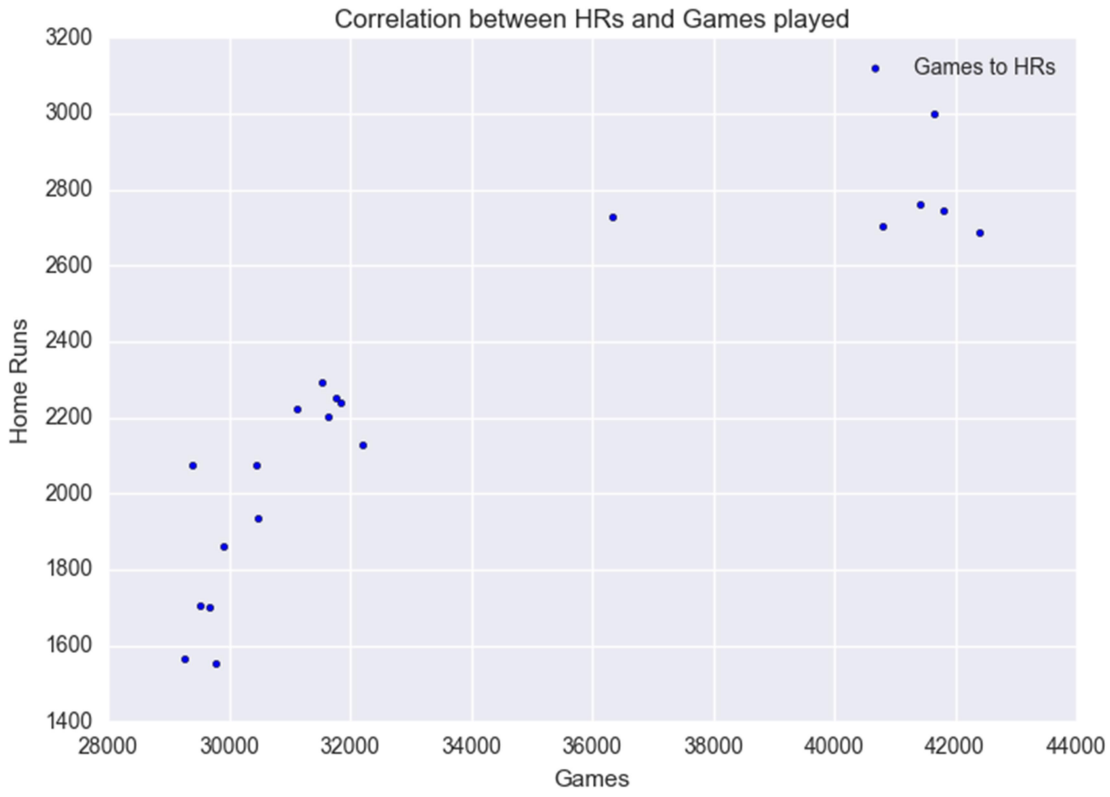
(*See Tables 1.1 to 1.4 in Python*)

Home Runs

In Table 1.4, the biggest difference between the two time periods were the percentage increases in the total number of Home Runs. From 1961 through 1966, HRs increased, 105%, 180%, 199%, 167%, 167%, and 126%. Could this be a result of integration? There is not enough information to directly attribute the increases in HRs to integration. There are other variables outside of race which could affect these differences. One major factor is that the 20 years before these years, 1941 through 1946, were the years of WWII when MLB players went off to war.

2. Correlation between HRs and the Number of Games

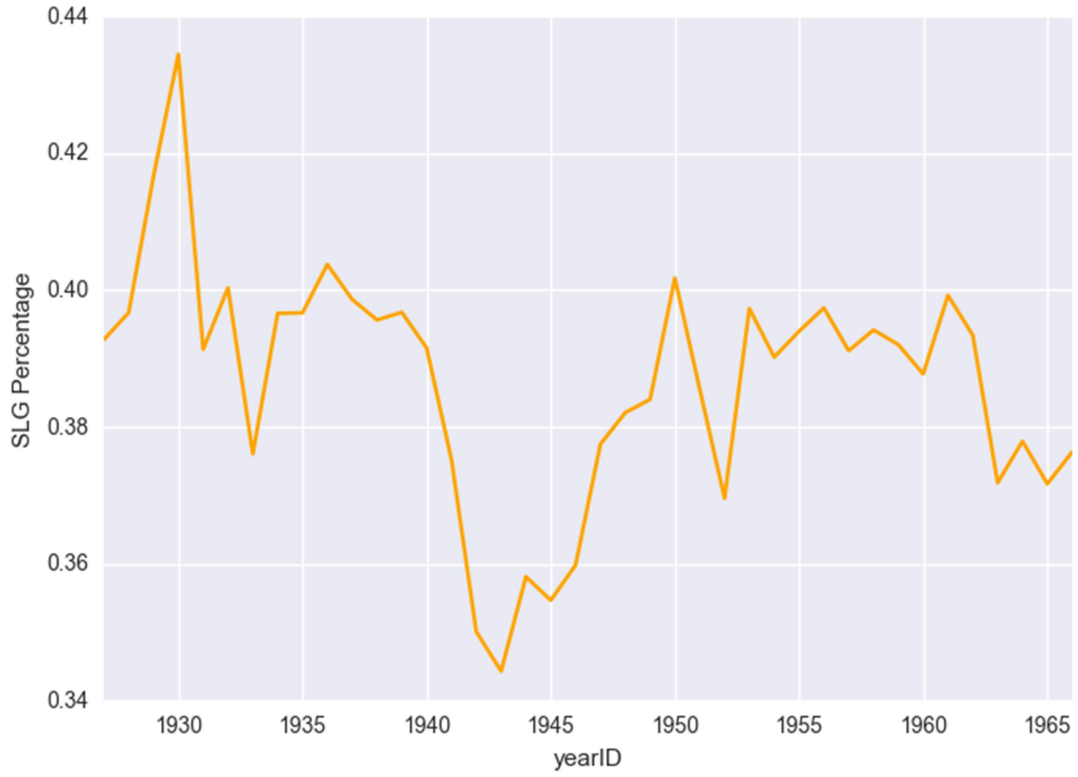
Another factor is that MLB expanded during this period. There were more teams and more games.



In the scatter plot above, you can see a positive correlation between the total number of games played to the increases in HRs. A Pearson R test shows the correlation coefficient to be .89322. Here, we can definitely say that there was a strong relationship between the number of games played and HRs.

3. Advanced Statistic for Hitting: Slugging Percentage (SLG)

The advanced hitting statistic is the slugging percentage which is the total number of bases per at-bats. Thus, a single counts as one base. A double counts as 2 bases. A triple is three bases, and a homerun is 4 bases. The function, def calc_SLG(Single, Double, Triple, HR, AB), calculates the SLG percentage. The highest possible SLG percentage is 4. That would mean that a player hit a homerun for every at-bat, which is not really possible. The Lahman data did not have a column for singles. Thus, I create a new column, '1B', which is Hits-(2B+3B+HR).



This time series chart shows the SLG percentage for both the pre- and post-integration eras. What the data shows is that there was rapid spike in the total bases per at-bat in 1930, a big drop off during the war years, an uptick after the war, and leading to a tight range of values post-integration. In addition, in the post-integration era, HRs became the primary source of power, as hits, doubles, and triples all declined from the pre-integration era. (See Python Table 1.4)

So, did racial integration of MLB cause the increase in HRs? This data analysis does not show that. This analysis does show that there was an increase in the total number of games played and that were was a strong correlation between the number of games and the increase in HRs. In the next section, this project will look into the pitching.

4. **Summary statistics for pitching**

(*See Tables 4.1 to 4.4 in Python*)

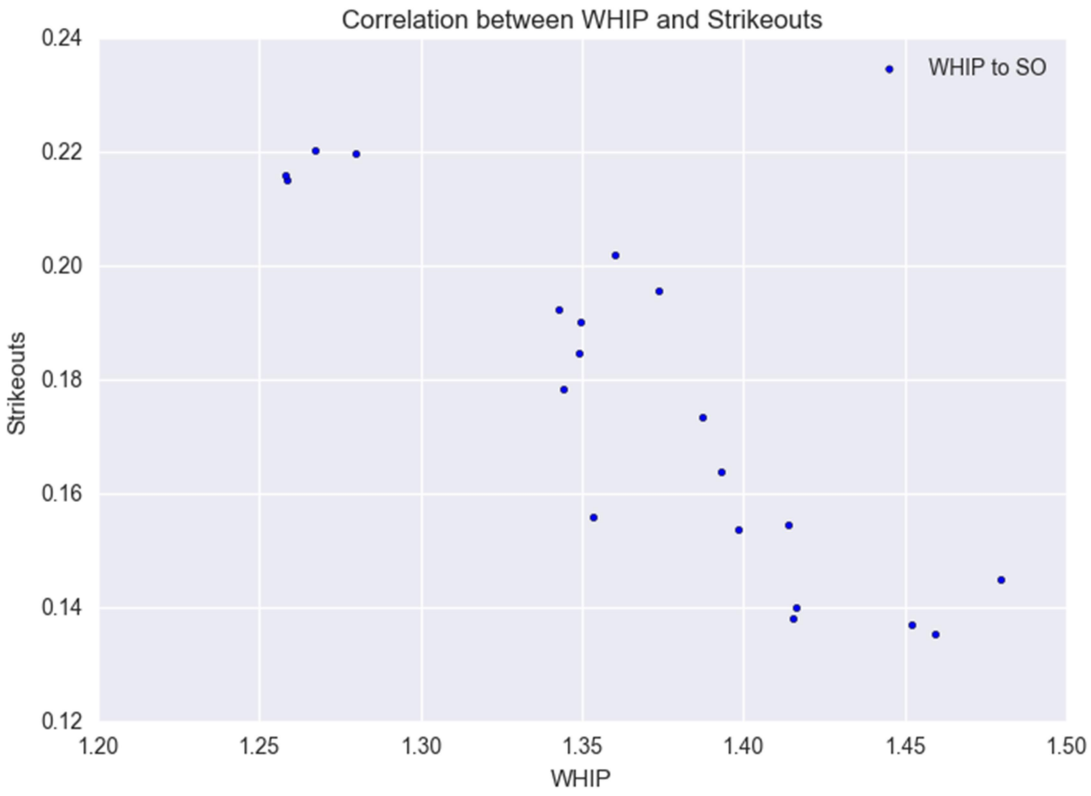
Table 4.4 shows the percentage differences between the two eras for Wins, IPouts, Earned Runs, Walks(BB), Hits(H), Strikeouts(SO), ERA, and Opppents Batting Avg. Against(BAOpp). The most glaring data difference is the strike outs which went up an average 120% from 1962 through 1966.

5. **Advanced Pitching statistic - Walks + Hits to Innings Pitched (WHIP)**

Walks + Hits / Innings Pitched (WHIP) is a measurement of the number of baserunners a pitcher has allowed per inning pitched. WHIP reflects a pitcher's propensity for allowing batters to reach base, therefore a lower WHIP indicates better performance.[2]

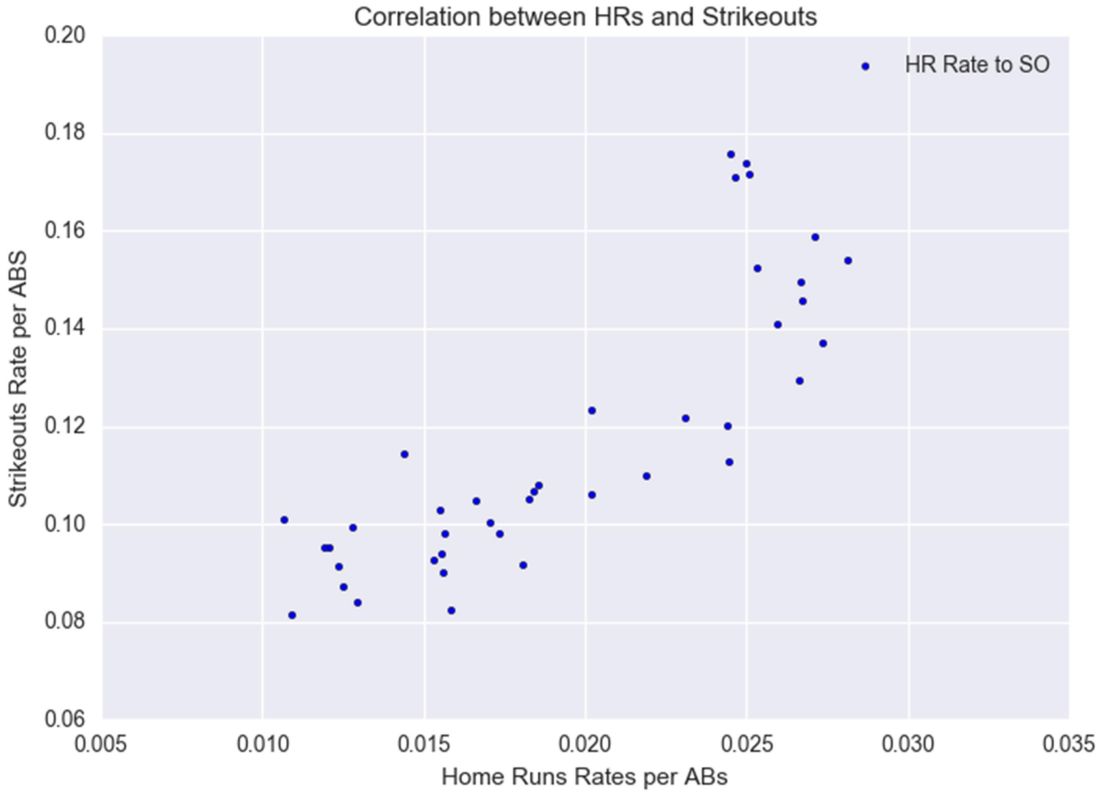
The function, def calc_WHIP(BB, H, IP), calculates this statistic. The average WHIP prior to integration was:1.656, but the average WHIP post integration was:1.585. The increase in strikeouts correlates to a lower WHIP. As you strike out more batters, you give up less walks and hits. The scatter plots below show the strong negative correlation between higher strike outs per inning pitched and a lowered WHIP. Pearson's R test shows the negative correlation at -.91.

[2] https://en.wikipedia.org/wiki/Walks_plus_hits_per_inning_pitched


Correlation between WHIP and Strikeouts
WHIP to SO
Strikeouts
WHIP
0.24
0.22
0.20
0.18
0.16
0.14
0.12
1.20
1.25
1.30
1.35
1.40
1.45
1.50

6. **Correlation between HRs and Strikeouts**

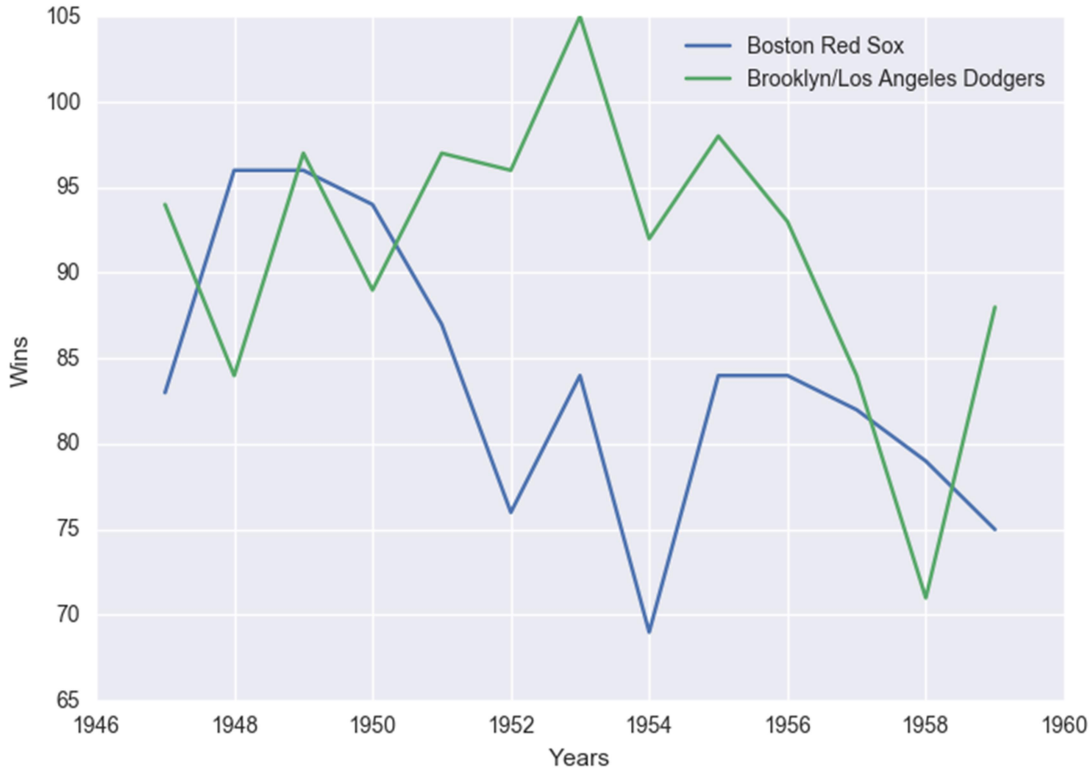
The figure below shows a strong correlation between strike outs per at-bats and home runs per at-bats.



As the Home Runs per at-bats increased, so did the strike outs which is also confirmed by the lowering of the WHIP statistic. Thus, the data analysis shows that the difference between the two eras was an increase in home run power hitting and power pitching. Was this difference due to integration? This data analysis does not show a direct relationship between integration and the increase in home runs and strikeouts. In addition, this project does not show that the differences are statistically significant. It does not test a hypothesis and use a test statistic to accept or reject the hypothesis.

7. Dodgers vs. Red Sox -- Early Adopters vs. Late Adopters

As previously stated, the Dodgers were the earliest adopters of racial integration. The last team to integrate was the Boston Red Sox until 1959. The time series chart below compares the number of wins between the two teams from 1947 to 1959.



The average wins and standard deviation for each team from 1947 through 1959 was:

Average wins:
BOS 83.769231
LAD 91.384615

Standard Deviation:
BOS 7.826521
LAD 8.157474

Additionally, the Dodgers won 4 NL pennants and 2 World Series titles during this period, and the Red Sox won neither. See Python Table 7.1.

This analysis does not prove that early adoption of racial integration contributed to the Dodgers' success. Nor does it prove that the Red Sox's performance was hampered by being late adopters to racial integration. The analysis does shows that the team that was the earliest adopter enjoyed a higher average number of wins from 1947 through 1959 along with post-season success. Also, it does show that racial integration did not hamper the Dodgers' success.

8. **A Sample of African American Players vs. A Sample of Non African American Players**

See Python Section - 8. Early African American Players

Runs created (RC) is a baseball statistic invented by Bill James to estimate the number of runs a hitter contributes to his team. [3] It is used to evaluate an individual's contribution to a team's total number of runs. Below is the formula for Runs Created:

A: $H + BB - CS + HBP - GIDP$
B: $(1.125 \times \text{Singles}) + (1.69 \times \text{Doubles}) + (3.02 \times \text{Triples}) + (3.73 \times \text{HR}) + .29 \times (\text{BB} - \text{IBB} + \text{HBP}) + .492 \times (\text{SH} + \text{SF} + \text{SB}) - (.04 \times \text{K})$
C: $AB + BB + HBP + SH + SF$

where K is strikeout.

The initial individual runs created estimate is then:

$$RC = \left(\frac{(2.4C + A)(3C + B)}{9C} \right) - .9C$$

https://en.wikipedia.org/wiki/Runs_created

Using a sample of 22 African American and Non African-American hitters, we see that even though the Non African-American sample had 27,015 less at-bats, the difference in the total runs created was only 1,756.17. The difference in the means is 109

Total At-Bats
African American Players: 67807.0
Non African-American: 94822.0

Total Runs Created
African American Players: 11668.712417940715
Non African-American Players: 13424.879473025956

Average Runs Created
African American Players: 530.3960189973052
Non African-American Players: 639.279974905998

[3] https://en.wikipedia.org/wiki/Runs_created

Standard Deviation Runs Created

African American Players: 540.8595108983942
Non African-American Players: 338.2169227409541

The data shows that with 27,015 less at-bats, the difference between the mean Runs Created (109) was less than half of the standard deviation for the Non African American players (338). Again, this is a data analysis, To see if the difference between the two means is statistically significant requires a two sample t-test.

Conclusion

The data shows that the impact of de-segregation of MLB was increases in Home Run hitting and Strike Outs. Also, the Dodgers, an early adopter of integration, out-performed the Red Sox, the last team to integrate for the years 1947-59. Finally, the data shows that a sample of African American players contributed to their teams' runs total.

References

1. Baseball Demographics, 1947-2012", by Mark Armour and Daniel R. Levitt. (SABR **Baseball Biography Project,** http://sabr.org/bioproj/topic/baseball-demographics-1947-2012**)**

2. https://en.wikipedia.org/wiki/Walks_plus_hits_per_inning_pitched

3. https://en.wikipedia.org/wiki/Runs_created

4. The Lahman Baseball Database