Data Analysis Project

Baseball's Hidden Hall of Famers

Is the Hall of Fame Missing Negro League Candidates?

I. Purpose of Analysis:

Historically, the Baseball Hall of Fame induction process has disenfranchised non-Major League Baseball players from candidacy. Because they were barred from playing on a national stage, Negro League players have not always appealed to the voters based on the candidacy of their leagues and associated statistics rather than candidacy of the players. Today, numerous researchers (The Negro Leagues are Major Leagues) as well as the Hall of Fame voters themselves have made strides to reconsider Negro League statistics as Major League statistics. The goal of this research is to use the continually improving Negro League data (Ashwell) to determine if there is statistical support for other players' historical induction into the Baseball Hall of Fame. Despite the improvement in historical records, Negro League game by game data is still very scarce and must be considered alongside subjective reasoning.

To assess the historical evidence, this analysis will focus on offensive production. It is the easiest part of the game to measure quantitatively, and thus any obvious candidates are likely to rise to the top in statistical analysis. The analysis will begin by combining the Negro League data with Major League data (Lahman) and categorizing players by the identifying characteristics: League, Hall of Fame status (Hall of Fame Batting Register), and career length.

The first lens that Hall of Fame voters and this analysis will use to evaluate players is career production. The analysis will look at career totals for games played and at bats as supplementary information to career length. Then it will consider on-field production through runs, hits, homeruns, and runs batted in. Career leaders (or players near the record) in any major statistical category are almost assuredly guaranteed induction because sustained excellence for long enough to approach existing records (generally held by Hall inductees) generally indicates a player had success against a wide range of teams, pitchers, league rule changes, stadiums, etc.

The second lens is rate statistics: how good is a player on average every time they step up to the plate? The analysis will consider batting average, on-base plus slugging percentage (OPS), and wins above replacement (WAR).[1] Using comparison tests and regression analysis, the analysis identifies lower thresholds for these batting statistics, which will be useful in isolating Negro League players who demonstrated Hall of Fame-worthy production without being recognized. The analysis contextualizes these thresholds by acknowledging that Hall of Fame voters may be biased. Is there evidence that voters are stricter on Negro League candidates?

However, no statistic(s) can perfectly isolate player performance within the confines of baseball's rules. This is why Hall of Famers use the third lens: narrative and impact on the sport. Impact is predominantly immeasurable, especially because we are considering players decades or centuries after they played, and without footage. In response, this statistical analysis concludes by comparing what was found to be the most promising candidate to another Negro League player who received the exact retroactive induction proposed through this analysis. This is the foundation for the historical narrative evidence presented in conclusion.

[1] WAR is not a rate statistic, but could easily be converted to one by dividing by any appropriate measure of longevity: games, at bats, plate attempts, seasons, etc. This is not included because the goal of this analysis is to examine voter behavior and create a justification for player candidacy based on their own patterns. It remains in this section as it uses findings from other rate statistics.

II. Data Collection & Description

General player data was collected from the Lahman Database. The list of Hall of Famers was taken from Baseball Reference's register. Negro League player data was taken from Ashwell's database.

298 Hall of Famers were considered. The Lahman Baseball Database considers 16,363 MLB players through 2006 because statistics are still being corrected and added for 2007 on. Unfortunately, Negro League data was only available by statistic and capped at 500 players. Thus, the top 500 players in both OPS and WAR were taken, but these sets do not necessarily overlap. As a consequence, data was only considered for 448 Negro League players.

General player data includes games played, at bats, career length, runs, hits (including how many bases), homeruns, runs batted in, batting average, and OPS. WAR data was taken from the Hall of Fame batting register and Ashwell database.

An at-bat occurs when a batter reaches base via a fielder's choice[1], hit, an error (not including catcher's interference), or when a batter is put out on a non-sacrifice.

A hit is a single, double, triple, or homerun achieved without a fielder's error or choice.

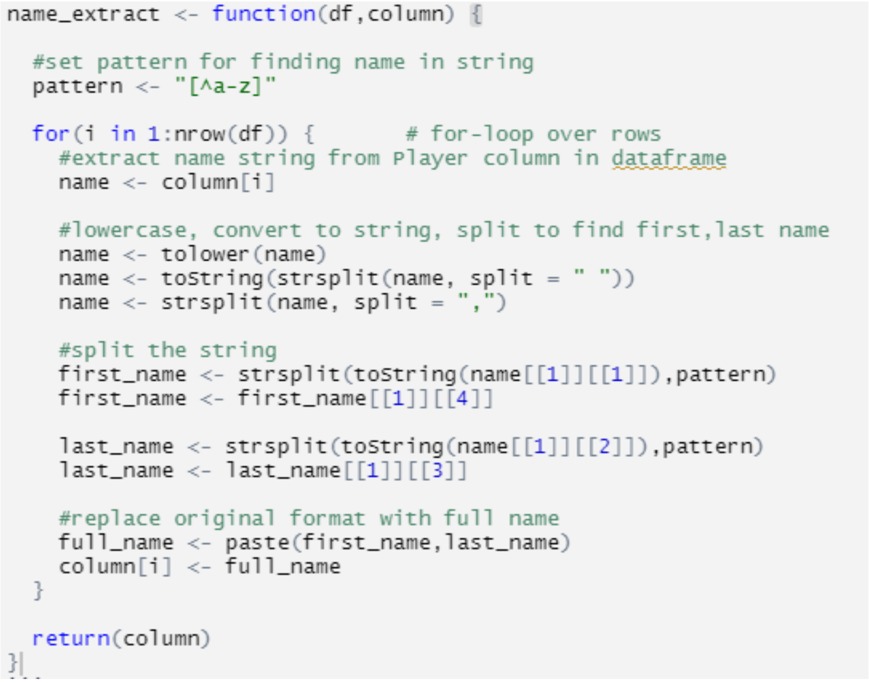
There are some features to note regarding the statistics constructed from these totals. OPS = On-base + Slugging. OBP = (Hits + Walks + Hit by Pitch) / (At Bats + Walks + Hit by Pitch + Sacrifice Flies). Slugging = (1B + 2Bx2 + 3Bx3 + HRx4) / AB. Clearly these denominators are not the same, so adding them together is mathematically problematic. What is valuable despite its problems is that the average is around .7-.8 (importance will become evident in this analysis), and that an OPS of at least 1 means a player generally contributes something offensively every time they come up to the plate.

WAR holistically evaluates batting, baserunning, fielding, and pitching. It summarizes how many runs a player produces on offense and prevents on defense and translates both into team wins (standard: 1 team win = 10 net runs created). Those wins are calculated in comparison to a replacement level player at the same position making minimum salary for that year. This standardizes WAR against league or era changes.

[1] when a fielder can elect to throw out one of multiple baserunners

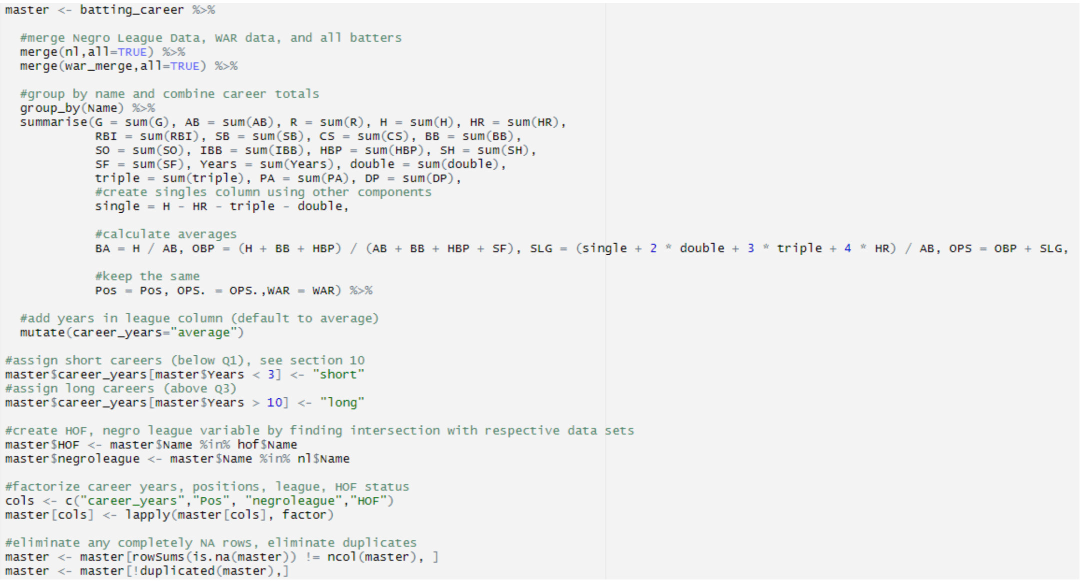
III. Data Wrangling

Cleaning the data started with the Negro League dataset because this it was the least standardized. A separate function was created and passed to pull the first and last name for each player.



```r
name_extract <- function(df,column) {

  #set pattern for finding name in string
  pattern <- "[^a-z]"

  for(i in 1:nrow(df)) {        # for-loop over rows
    #extract name string from Player column in dataframe
    name <- column[i]

    #lowercase, convert to string, split to find first,last name
    name <- tolower(name)
    name <- toString(strsplit(name, split = " "))
    name <- strsplit(name, split = ",")

    #split the string
    first_name <- strsplit(toString(name[[1]][[1]]),pattern)
    first_name <- first_name[[1]][[4]]

    last_name <- strsplit(toString(name[[1]][[2]]),pattern)
    last_name <- last_name[[1]][[3]]

    #replace original format with full name
    full_name <- paste(first_name,last_name)
    column[i] <- full_name
  }

  return(column)
}
```

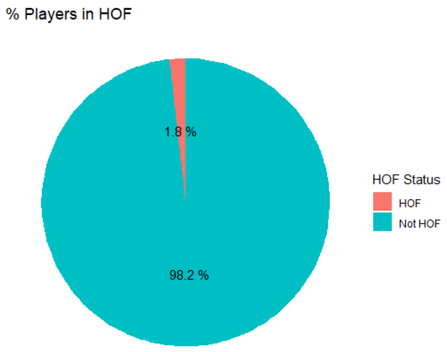
After some minor standardizations to full name, the Negro League and general player data were merged, using the other datasets to add classifiers for league, career length (discussed in next section) and Hall of Fame status. Because batting average was not given for all players, it was calculated from the available components. Finally, duplicates and NAs were removed.

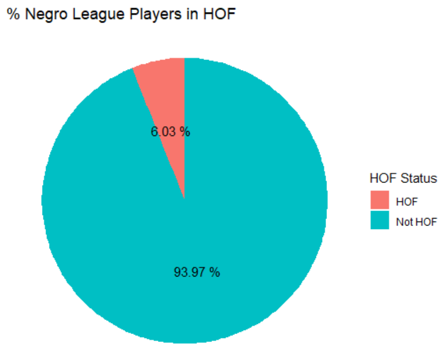


```r
master <- batting_career %>%

  #merge Negro League Data, WAR data, and all batters
  merge(nl,all=TRUE) %>%
  merge(war_merge,all=TRUE) %>%

  #group by name and combine career totals
  group_by(Name) %>%
  summarise(G = sum(G), AB = sum(AB), R = sum(R), H = sum(H), HR = sum(HR),
            RBI = sum(RBI), SB = sum(SB), CS = sum(CS), BB = sum(BB),
            SO = sum(SO), IBB = sum(IBB), HBP = sum(HBP), SH = sum(SH),
            SF = sum(SF), Years = sum(Years), double = sum(double),
            triple = sum(triple), PA = sum(PA), DP = sum(DP),
            #create singles column using other components
            single = H - HR - triple - double,

            #calculate averages
            BA = H / AB, OBP = (H + BB + HBP) / (AB + BB + HBP + SF), SLG = (single + 2 * double + 3 * triple + 4 * HR) / AB, OPS = OBP + SLG,

            #keep the same
            Pos = Pos, OPS. = OPS.,WAR = WAR) %>%

  #add years in league column (default to average)
  mutate(career_years="average")

#assign short careers (below Q1), see section 10
master$career_years[master$Years < 3] <- "short"
#assign long careers (above Q3)
master$career_years[master$Years > 10] <- "long"

#create HOF, negro league variable by finding intersection with respective data sets
master$HOF <- master$Name %in% hof$Name
master$negroleague <- master$Name %in% nl$Name

#factorize career years, positions, league, HOF status
cols <- c("career_years","Pos", "negroleague","HOF")
master[cols] <- lapply(master[cols], factor)

#eliminate any completely NA rows, eliminate duplicates
master <- master[rowSums(is.na(master)) != ncol(master), ]
master <- master[!duplicated(master),]
```

IV. Exploratory Data Analysis

HOF Status





Players in HOF

	Not HOF	HOF
Freq	16391	300[2]

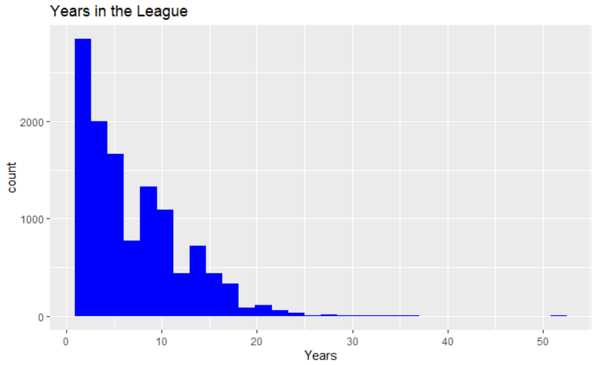
Negro League Players in HOF

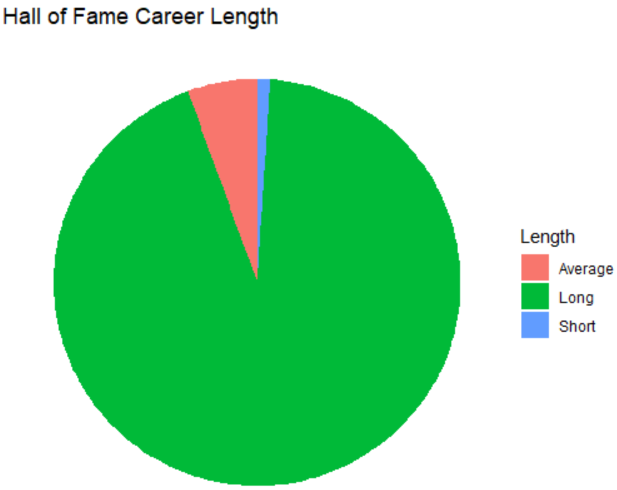
Not HOF	HOF
421	27

The first pie chart shows the proportion of all players in the Hall of Fame, while the second pie chart shows the proportion of Negro League Players in the Hall of Fame. We immediately see the impact of not having all Negro League data, as the real proportion of Negro League Hall of Famers will be lower if all players were recorded. Hence, these proportions will not be ideal for determining bias in induction voting. However, the 1.8% induction rate is valuable in noting that all Hall of Famers by definition should be outliers in some manner.

[2] This number is higher than 298 because some players played in both leagues, and thus are considered as two different players

Career Length





All Players' Career Length

	Short	Average	Long
Freq	2843	6303	2792
Percent	23.81%	52.80%	23.39%

Hall of Famers' Career Length

Short	Average	Long
3	16	263
1.06%	5.67%	93.26%

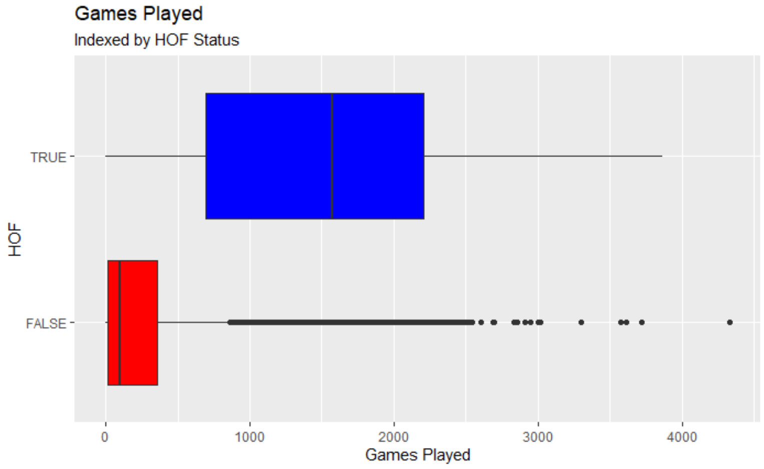
The histogram shows how many years each player played Major League baseball. Generally, fewer players play at each increase in career length. The pie chart shows how many Hall of Famers had short, average, or long careers. Because players' years in the league are inconsistently recorded, particularly if they switched to the Major Leagues with desegregation, only careers that spanned at least a season were tabulated[3]. This also applies to the following variables.

"Short" careers were defined as less than 3 years (1st quartile of career length), "long" careers were defined as greater than 10 years (3rd quartile of career length), and "average" careers were defined in between. The first justification for converting this quantitative variable into a categorical one was the aforementioned inconsistencies with 0 year careers[3]. The second reason is that players may have extended careers for different reasons. Ideally, we want to compare players with others we would have had we known how long each player was regularly in the lineup. Categorizing years enables cross year comparisons by range.

The second pie chart confirms what should be true: Hall of Famers tended to have long careers justified by sustained success. 98.93% of Hall of Famers had at least average careers, most of which were long. Thus, analysis will focus on Negro league players who played at least average.

[3] career lengths of 0 generally indicate NA as opposed to a professional player who never actually played

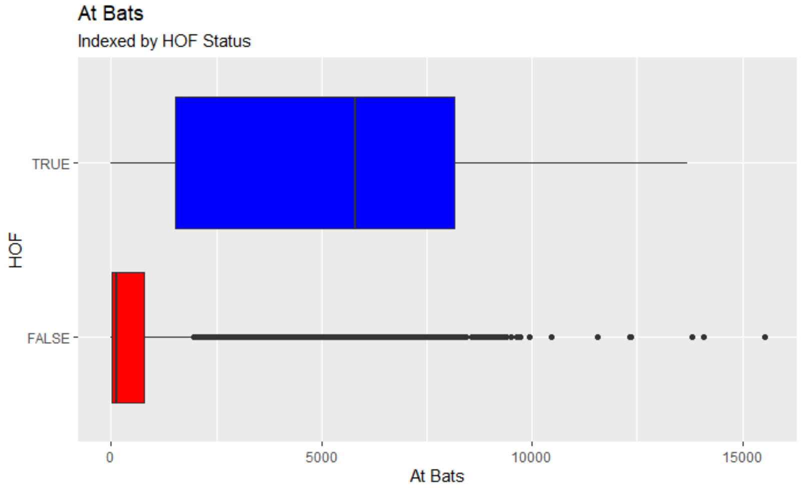
Games Played (G)



Summary	Not Hall of Fame	Hall of Fame
Min	1	1
Q1	17	694
Median	97	1588
Mean	286	1519
Q3	345.2	2210
Max	4331	3862

The boxplots show the number of games played by Hall of Famers vs non-Hall of Famers. Interestingly, the upper quartile of non-Hall of Famers starting (block of individual points starting around 800) corresponds very closely to the upper 75% of Hall of Famers (starting around 700). Analysis will focus on Negro league players in this overlap. This is supported by the fact that the mean is lower than the median for Hall of Famers, while the opposite is true for non-inductees. This suggests that the non-inductee mean is more heavily influenced by players in the upper quartile, those who in theory should be good candidates.

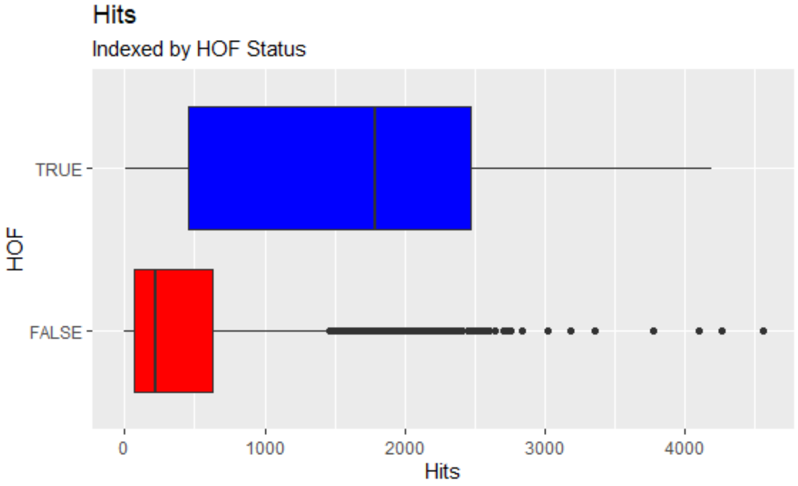
At Bats (AB)



Summary	Not Hall of Fame	Hall of Fame
Min	1	1
Q1	16	1532
Median	130	5836
Mean	811.4	5279
Q3	789	8154
Max	15527	13707

Comparing the boxplots to one another tells a similar story to games played: the upper quartile of non-Hall of Fame players is heavily concentrated above the lower quartile of Hall of Famers (around 1500). Analysis will focus on Negro league players above this lower quartile threshold.

Hits (H)



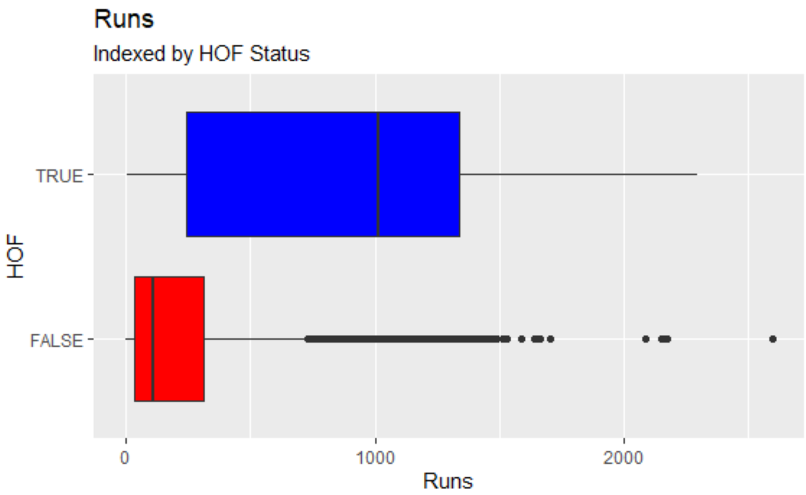
Summary	Not Hall of Fame	Hall of Fame
Min	1	3
Q1	67	471
Median	214	1813
Mean	436.9	1657
Q3	622	2467
Max	4555	4189

The boxplots show the number of hits by Hall of Famers vs non-Hall of Famers.

Here, the “overlap” observed in games played and at-bats is more apparent, as the middle 50% of non-Hall of Famers crosses the first quartile of Hall of Famers. The chunk of individual points in the upper quartile of non-Hall of Famers (beginning around 1500) more closely corresponds to the upper half of Hall of Famers (median is about 1800). Additionally, the middle 50% of non-Hall of Famers has a wider range. This indicates hits has a relatively large gray area as a measure of Hall of Fame candidacy because Hall of Famers are closer to non-Hall of Famers in hits than games or at-bats. Consideration of fielder’s errors and choice may account for this: a player’s ability to reach base for a hit is affected by defensive players, whereas at-bats include balls in play regardless of the defense’s response, and games played obviously do not depend on defense. Defense creates a larger variance in hits.

High hit totals are still an indication of sustained success; however, caution must be taken with the lower bound of potential Hall of Fame candidates so as to not exclude players who still exceed the 1st quartile of inductee hits.

Runs (R)



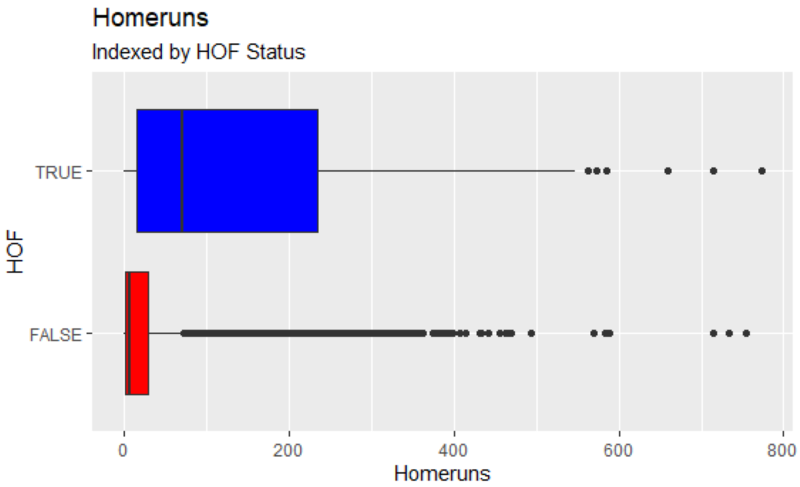
Summary	Not Hall of Fame	Hall of Fame
Min	1	2
Q1	31	248
Median	105	1024
Mean	221	912.7
Q3	308	1338
Max	2596	2295

The boxplots show the number of runs by Hall of Famers vs non-Hall of Famers. A run is scored whenever a player crosses home plate.

A similar overlap to hits can be seen with runs, which makes sense considering a player with more hits is more likely to cross home plate. Again, caution must be taken when considering runs for Hall of Fame candidacy. The overlap here is not only subject to defensive performance, but also team performance. Unlike hits, runs also depend on teammates because any hit other than a homerun only converts to a run if batters after a baserunner are able to hit them home. Runs make no differentiation between a player who singles and is brought home by a teammate versus a player who brings themselves home by hitting a homerun.

The lower quartile of Hall of Famers is close to the mean of non-Hall of Famers, providing a potential lower threshold for Negro League Hall of Fame candidates.

Homeruns (HR)



Summary	Not Hall of Fame	Hall of Fame
Min	1	1
Q1	2	16
Median	7	71
Mean	30.6	145.7
Q3	30	234
Max	755	773

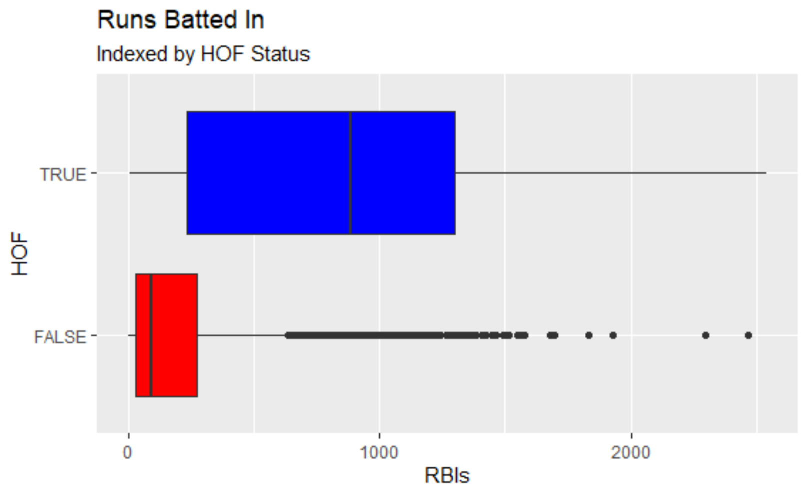
The boxplots show the number of homeruns by Hall of Famers vs non-Hall of Famers.

Compared to hits and runs, the career totals of the middle 50% for both is significantly more narrow and the distribution is skewed right. This suggests homeruns is a better indication of a desired quality in Hall of Fame candidates: hitting for power. Standout players are more obvious in this category, as career totals are more likely to be small regardless of Hall of Fame status unless a player is exceptional. This may be because homeruns are independent of defense and teammates: no one can influence a homerun except pitchers and hitters.

This reduces variance, which is reflected in the narrower middle 50% of both boxplots. Of the career totals analyzed so far, homeruns is the best indicator of performance because 75% of all players have a relatively small amount of homeruns (fewer than 30), which draws top performers significantly further away from the median and mean.

Additionally, the mean of homeruns for non-Hall of Famers is higher than the upper quartile, suggesting the high mean is due to outliers rather than everyone averaging around 30 homeruns. In contrast, the upper quartile of Hall of Famers is nearly a hundred homeruns larger than the mean. This indicates homeruns is also a good indicator of induction many Hall of Famers hit a lot of homeruns, not just the outliers.

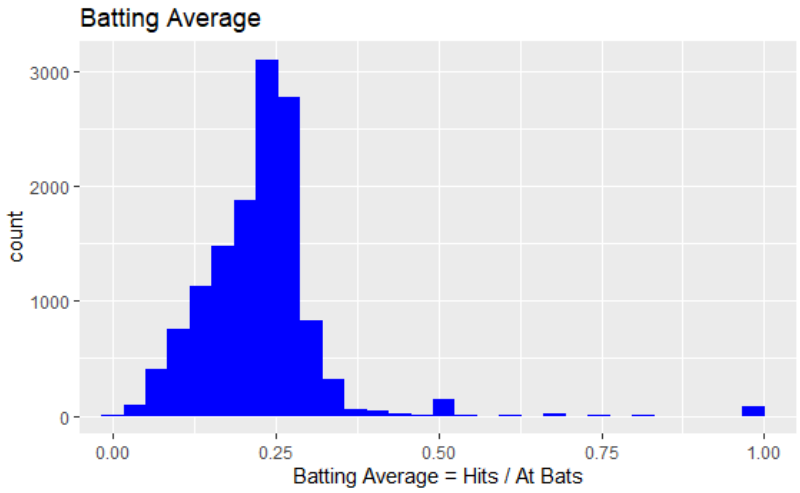
Runs Batted in (RBI)



Summary	Not Hall of Fame	Hall of Fame
Min	1	2
Q1	27	230
Median	91	883
Mean	197.8	844
Q3	270	1299
Max	2467	2541

The comparative boxplots for RBIs looks almost identical to that of runs and hits, and is subject to some of the same issues of dependence. A player who bats after teammates who get on base frequently is more likely to score RBIs regardless of their own ability, and the same issue with defense also comes into play. A cautious lower threshold must be considered for this variable.

Batting Average (BA)

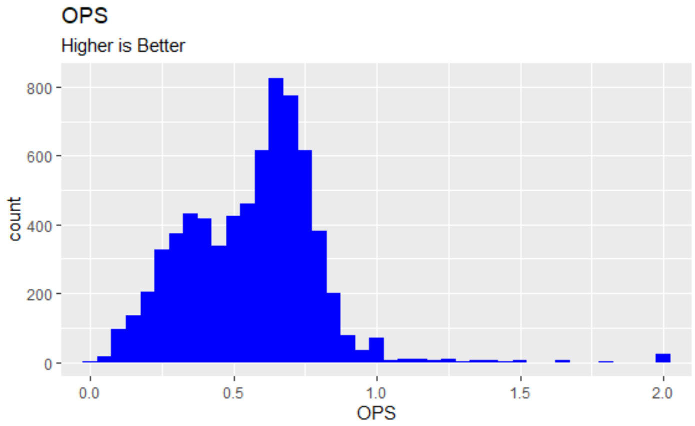


Summary	All Players	Hall of Fame
Min	.01587	.03448
Q1	.17241	.22644
Median	.23156	.28535
Mean	.22528	.26680
Q3	.26431	.31254
Max	1	.36636

The histogram shows the batting average of all players. Considering the non-Hall of Fame data set includes numerous players who played less than a full season of baseball and the distribution is heavily skewed right, the summary statistics of non-Hall of Famers and Hall of Famers are remarkably similar. Most notably, both means and medians lie around the histogram's peak of .25.

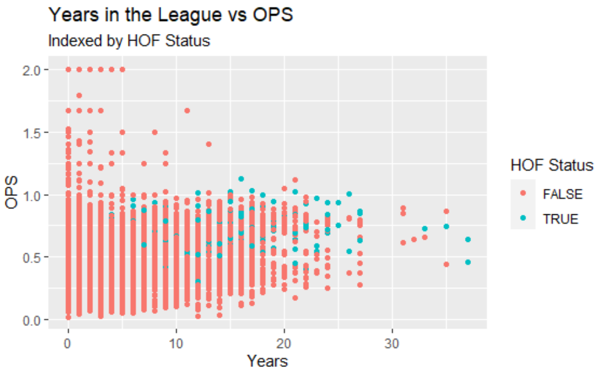
While the career averages of Hall of Famers are certainly higher, batting average appears to be one of the more problematic variables considered thus far. This may be due to batting average being dependent on hits and at-bats, which also depend on hits. Therefore, batting average is subject to the variance of both hits and at-bats, making it one of the noisiest variables. One exception that could be useful would be a player whose role on their team is specifically to accumulate hits and get on base so that a power hitter can get hit them home. However, in this case career totals might be just as good, if not a better measure of that prototype's performance.

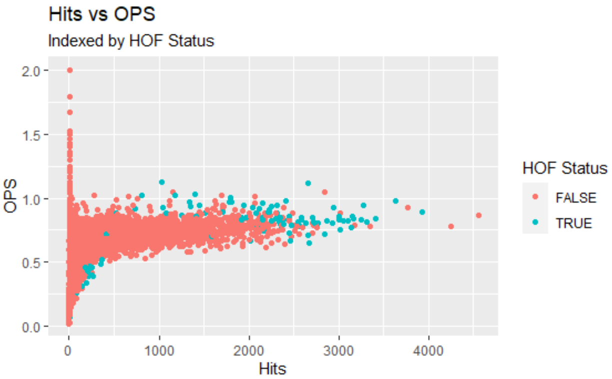
OPS



Summary	Not Hall of Fame	Hall of Fame
Min	.02381	.06897
Q1	.39231	.52121
Median	.59422	.80444
Mean	.56514	.71224
Q3	.70406	.87039
Max	4	1.12197

OPS has been restricted to 2 max, but all OPS values will be considered in final analysis. However these outliers may have been recorded due to a low number of games. These outliers were omitted in the end because they were achieved on a number of games that does not reach the minimum threshold (Hall of Fame lower quartile).

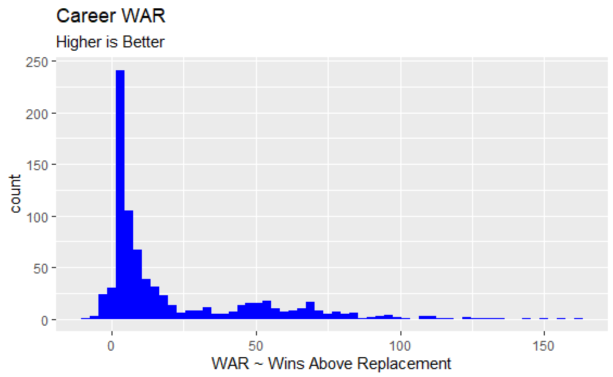


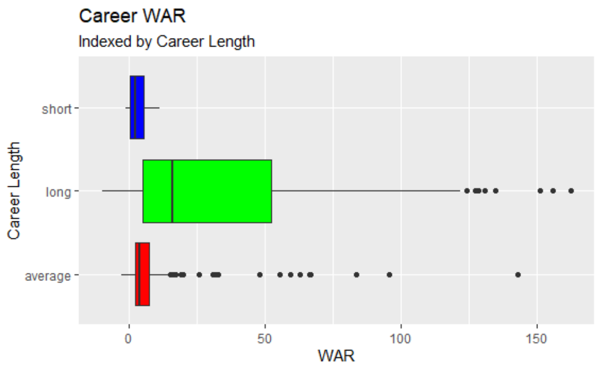


Unlike batting average, there is a marked difference between elite and non-elite OPS. The first scatterplot shows years in the league vs OPS. Hall of Famers (blue dots) tend to appear at the top in terms of OPS for average and above average careers (>3 years). Hall of Famers tend to hover around the .800 mark. It is important to note that further to the right, Hall of Famers standout based on sustaining a high OPS: a high OPS alone is relatively common.

Similarly, the second scatterplot shows long term production though hits vs OPS. Hall of Famers still hover around the .800 mark, but their separation in terms of longevity is even more apparent. This makes sense since players who accumulate many hits are in the league longer on average than those with fewer hits. Hall of Famers sit distinctly towards the graph’s upper right, strengthening the connection between longevity and high performance of Hall of Famers.

WAR



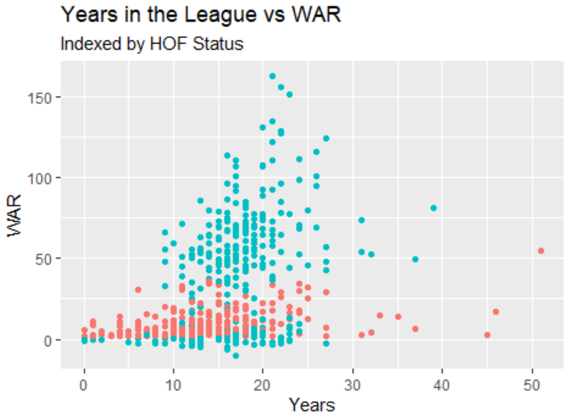


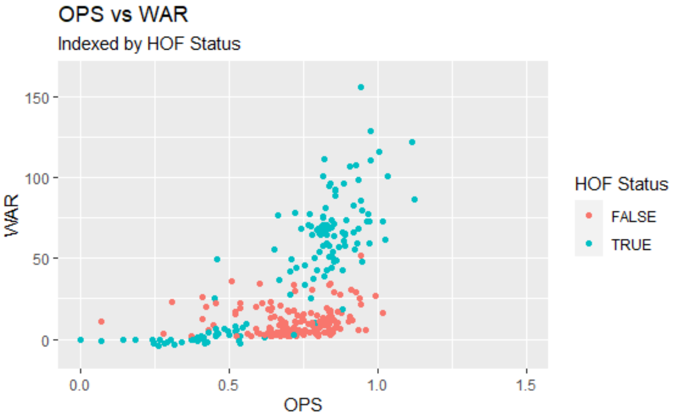
	Min	Q1	Median	Mean	Q3	Max
Not Hall of Fame	1.6	2.9	5.3	7.647	10.05	54.6
Hall of Fame	-9.8	3.8	46.8	43.1	67.9	162.7

The histogram shows wins above replacement for Hall of Famers and Negro League players.

The boxplot suggests that WAR is heavily affected by career length, as the middle 50% of long careers have a very wide range of WAR. No players with short careers reached the median WAR of long careers. The middle 50% of short and average careers are quite similar and narrow, making elite performance standout as outliers far from the median.

The outliers of average career length are the most interesting cases. They correspond with the upper 50% of long careers, suggesting these players greatly contributed to winning in a shorter amount of time.



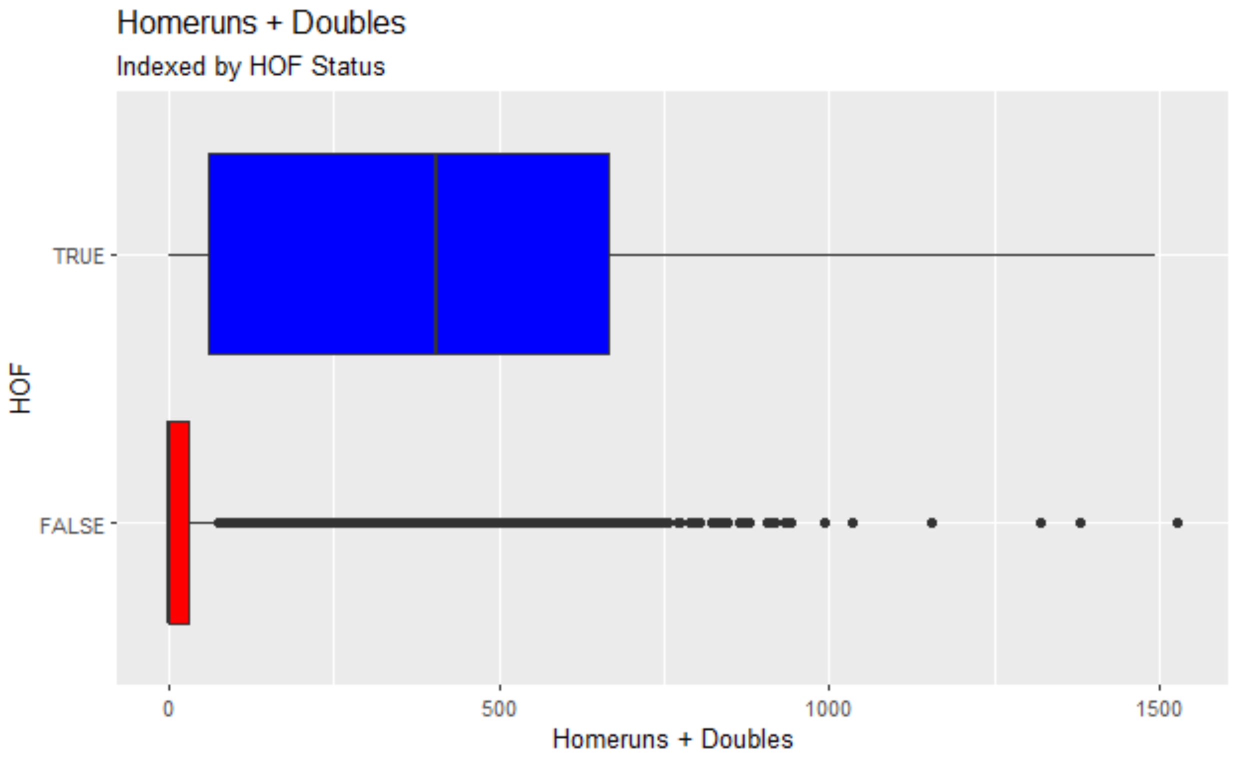


The first scatterplot shows years in the league vs WAR. The abundance of non-Hall of Famers at the bottom and its linear shape demonstrates that an average player will not accrue WAR even if they stay in the league for a long time. A linear relationship is less apparent in Hall of Famers. As career length increases, players have a wider variation in WAR, suggesting it is a stronger differentiator between elite players than average or below average players.

The second scatterplot shows OPS vs WAR may have an exponential relationship. This means that a small increase in OPS leads to a large increase in WAR. Again, there is a clear separation between Hall of Famers and non-Hall of Famers.

Compared to OPS, Hall of Famers are even more distinct from non-Hall of Famers in WAR, suggesting it may be the strongest indicator of voter preferences. For every career length, the top of the first scatterplot is dominated by Hall of Famers. For high OPS (>.75), Hall of Famers have the highest WAR.

Power (= Homeruns + Doubles): Comparison of Means between HOF & Non-HOF Players



Power was considered after initial exploration on how homeruns and doubles differentiated Hall of Famers from non-inductees. The boxplot supports this aggregation, as the upper quartile of non-inductees almost perfectly corresponds to the upper 75% of inductees. This provides a clearer lower threshold for candidates.

Based on all plots, career length, power, OPS, and WAR appear to be the best differentiators between whether or not someone gets inducted into the Hall of Fame.

V. Statistical Models

The restriction of candidates to at least average career length was suggested by the pie chart in exploratory data analysis. This was confirmed using a chi-square test, which determined whether or not the extremely high proportion of at least average career Hall of Famers is noteworthy when compared to the careers of non-inductees.

Evaluation of career success started with power, as it consolidates most of the information available in hits, at bats, homeruns, and doubles. A t-test to compare means between inductees and non-inductees was used to determine if voting patterns expect higher power.

Three linear regression models were used for OPS and WAR against games played, years, and power: Hall of Famers, non-Hall of Famers, and all players. Power was omitted from OPS because OPS is linearly dependent on doubles and homeruns. Years was left in the rmd solely to demonstrate its relative redundancy after regressing on games played. Since we are chiefly concerned in the differences between inductees and non-inductees, breaking the linear models down provides further insight into why Hall of Famers stood out in visual analysis. The estimated parameters can be compared and interpreted beyond the fit of the model.

Finally t-tests for comparisons of means were used between Negro League and non-Negro League players to evaluate the presence of voter bias.

VI. Results

Career Length: Comparison of Proportion of Short Careers by HOF

	Not in HOF	HOF
Short	7403 (9116)	5 (167)
At Least Average	8988 (7275)	295 (133)

X-squared = 224.06, df = 1, p-value < 2.2e-16

Considering assumptions: all expected counts are > 5. We disregard independence condition as we are specifically interested in voting bias towards longer careers.

The p-value well below conventional values suggests we reject the null hypothesis that induction is independent of career length. We are unlikely to observe such a low proportion of Hall of Famers with short careers (or such a high proportion of at least average careers) if it were not a factor when considering induction. Voters either may be actively considering this variable or weighing career production heavily, which increases with career length. Either way, this supports focus on Negro League players with at least average career lengths.

Power: Comparison of Means between HOF & Non-HOF Players

	Min	Q1	Median	Mean	Q3	Max
Not Hall of Fame	0	0	2	44.62	30	1528
Hall of Fame	0	60.25	405.5	399.88	665.75	1493

t = -18.1, df = 284.02, p-value < 2.2e-16
alternative hypothesis: true difference in means between HOF and non-HOF is not 0
95 percent confidence interval: [-395.1012, -317.9789]
mean non-HOF: 44.33319, mean HOF: 400.87324

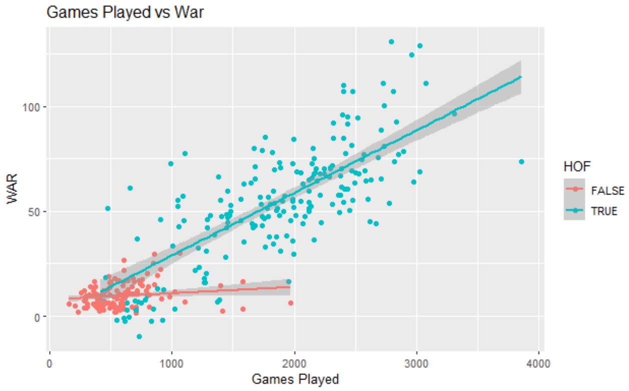
--

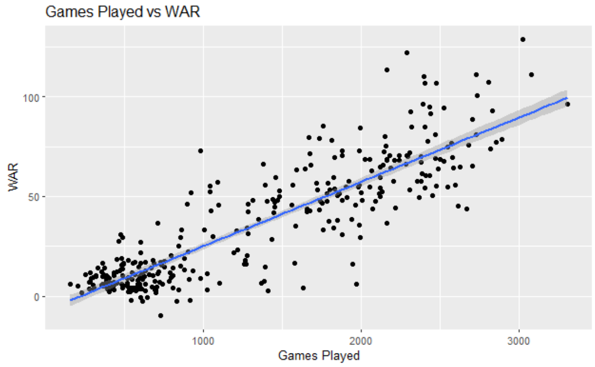
It was found in comparing this boxplot to that of just homeruns that the difference between Hall of Famers and non-Hall of Famers is more pronounced, evidenced by the lack of overlap between the middle 50% of each power boxplot. This is perhaps due to the fact that the difference between a home run and double is often trivial and independent of the player (e.g. park dimensions, outfield wall height, wind, etc.). Players who hit with elite power will hit both frequently and relatively indiscriminately, which justifies its use for comparison here.

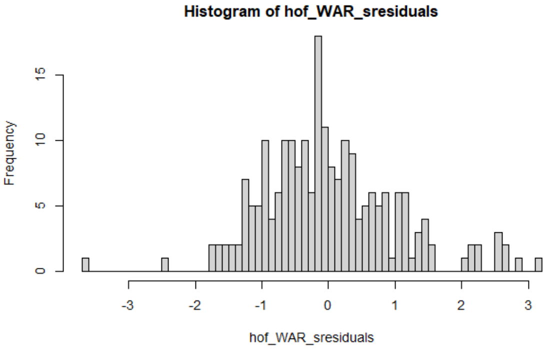
Considering assumptions: both samples are well above 40, so normality condition does not need to be satisfied. We disregard independence for the sake of analysis.

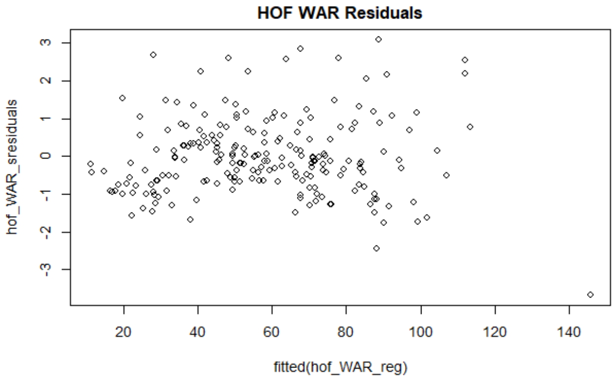
The p-value well below conventional values (e.g. alpha = .05) suggests we reject the null hypothesis that there is no difference between mean power between inductees and non-inductees. Such a difference in average power between Hall of Famers and non-Hall of Famers is unlikely if it were not a factor when considering induction. All outliers of non-inductees lie in upper 75% of inductees. This and the relatively wide spread of inductee power suggests power alone does not guarantee induction. Analysis will focus on these non-inducted outliers whose power lands in the upper 75% of inductees (power = 60.25).

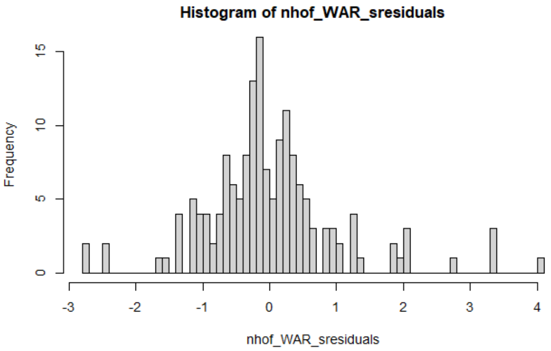
WAR: Multiple Linear Regression

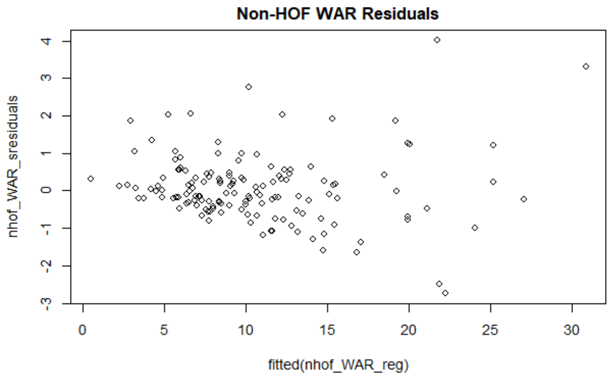


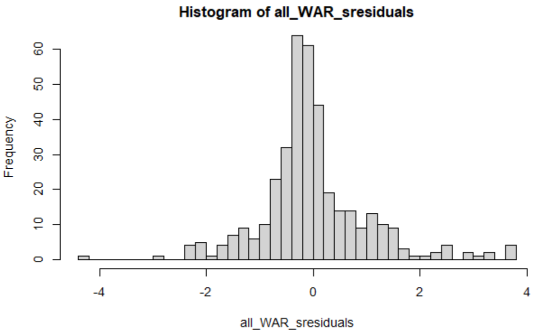


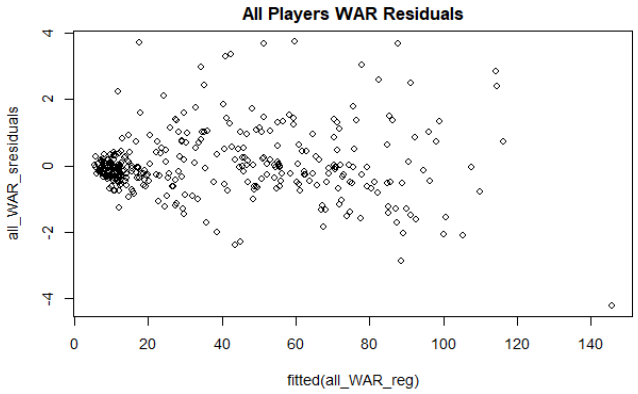












HOF:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-3.317143	4.867993	-0.681	0.49642
G	0.008493	0.003120	2.722	0.00708
Years	0.522150	0.270892	1.928	0.05538 .
power	0.062646	0.007198	8.703	1.39e-15

R^2 = . 6788, p-value < 2.2e-16

Non-HOF:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	2.886952	1.180482	2.446	0.015811
G	-0.011209	0.002594	-4.322	3.06e-05
Years	0.300581	0.077787	3.864	0.000176
power	0.073358	0.010856	6.757	4.36e-10

R^2 = . 3066, p-value = 2.849e-10

All Players:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-3.408821	2.085696	-1.634	0.103
G	0.012268	0.001973	6.217	1.55e-09
Years	0.008283	0.128816	0.064	0.949
power	0.065261	0.005367	12.159	< 2e-16

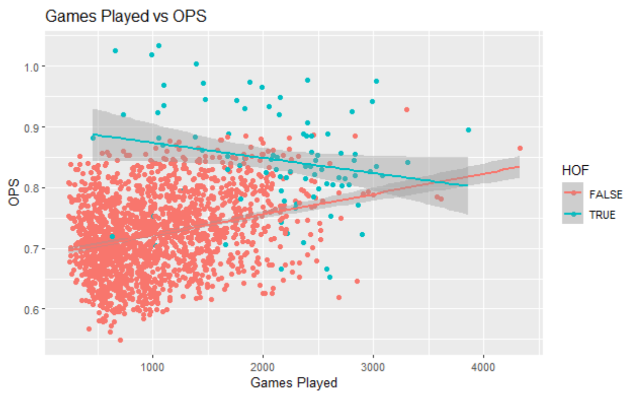
R^2 = . 8436, p-value < 2.2e-16

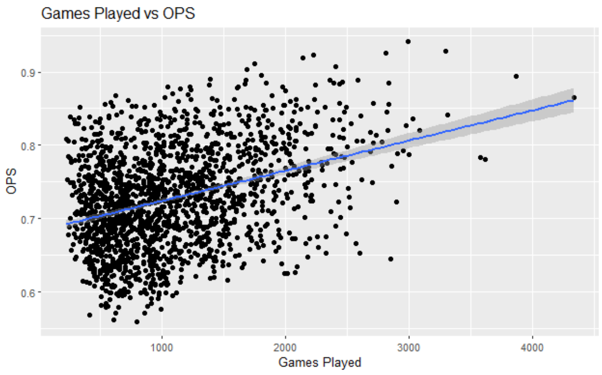
Considering the linearity assumption, this is questionable. It was certainly shown that there is a positive correlation between WAR and games played and power. However, WAR tends to increase rapidly for high volumes of years, particularly for Hall of Famers. In terms of the independence assumption, the residual plot for inductees, non-inductees, and all players is relatively random. As far as equal variances, all the residual plots are relatively homoscedastic. The all players plot is a little bunched up on the left. The residual histograms are approximately normal, satisfying the normality condition.

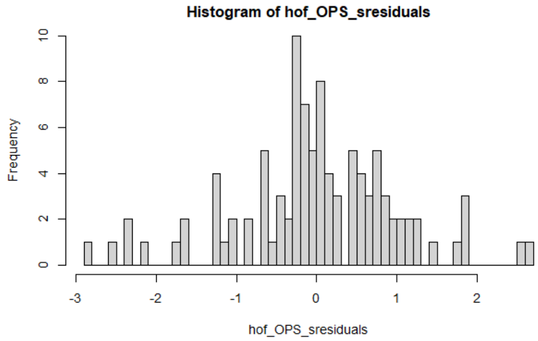
Comparing the R^2s, the explanatory variables account for significantly more of inductee data (68% vs 31%). The model also describes a significant portion (84%) of all player data. Comparing this to the low proportion of non-Hall of Famers explained by the model, this suggests either that WAR is a better measure of elite players' contributions to winning than average players or that voters rely on some form of the metric even if it is a poor indicator of contribution.

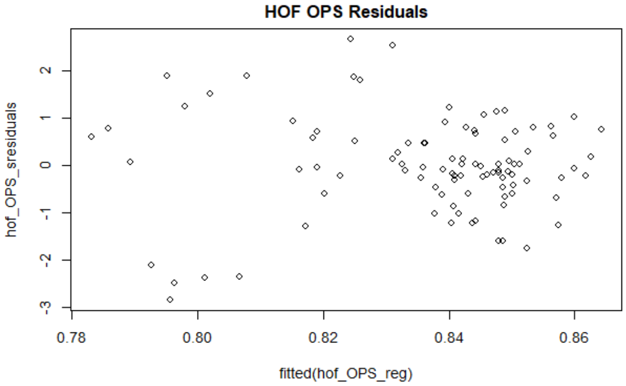
P-values for games and power are relatively low, suggesting we reject the null hypothesis that they are not associated with WAR. It is unlikely to observe such an association with WAR due to random variation. By comparing probabilities, years is generally the weakest linear indicator (especially with all players), whereas power was the strongest.

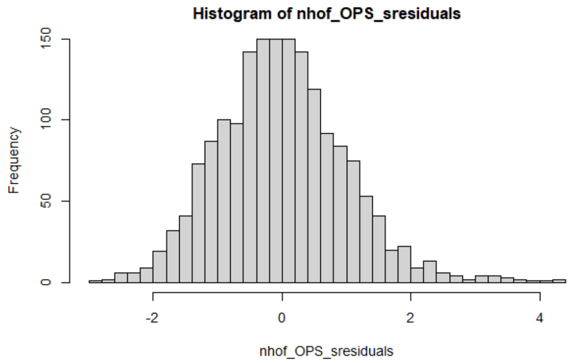
OPS: Multiple Linear Regression

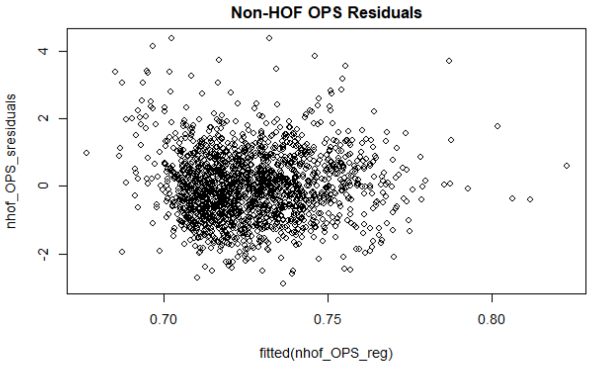

Games Played vs OPS

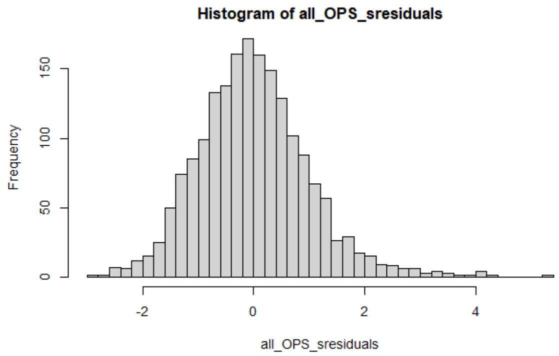

Games Played vs OPS

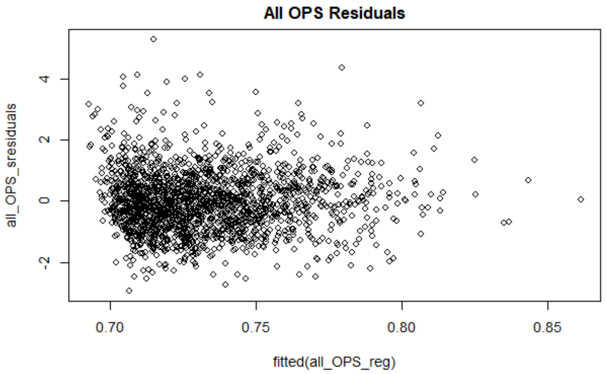

Histogram of hof_OPS_sresiduals


HOF OPS Residuals


Histogram of nhof_OPS_sresiduals


Non-HOF OPS Residuals


Histogram of all_OPS_sresiduals


All OPS Residuals

```
HOF:          Estimate   Std. Error    t value   Pr(>|t|)

(Intercept)   8.719e-01  3.909e-02     22.307    <2e-16 ***
G            -3.273e-05  1.299e-05     -2.519    0.0136 *
Years         2.388e-03  2.093e-03      1.141    0.2571
```

R^2 = . 08285, p-value = . 04516

```
Non-HOF:      Estimate     Std. Error    t value       Pr(>|t|)

(Intercept)   7.033e-01    4.011e-03     175.324       < 2e-16
G             5.392e-05    4.054e-06      13.300       < 2e-16
Years        -3.328e-03    5.014e-04      -6.638       4.27e-11
```

R^2 = . 1051, p-value < 2.2e-16

```
All Players:    Estimate     Std. Error   t value      Pr(>|t|)

(Intercept)     6.926e-01    3.951e-03    175.296      < 2e-16
G               5.348e-05    3.823e-06     13.992      < 2e-16
Years          -2.070e-03    4.903e-04     -4.223      2.54e-05
```

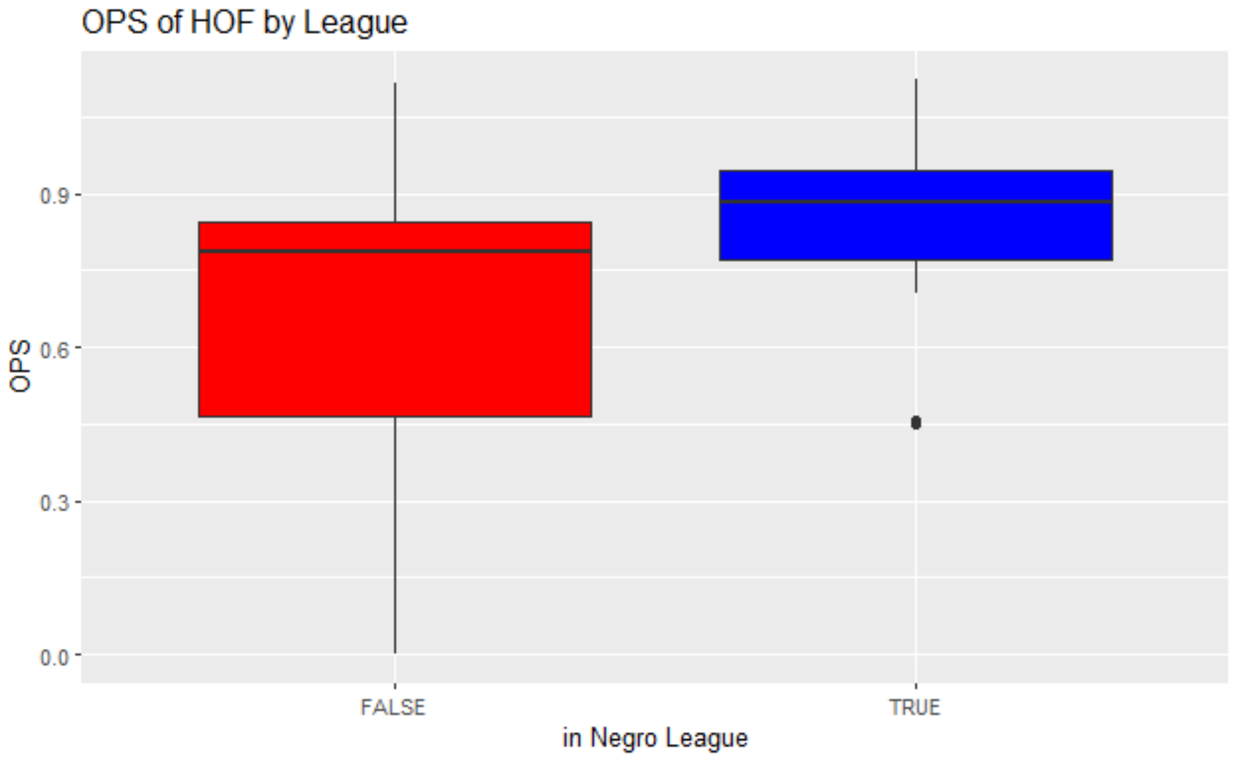
R^2 = .1379, p-value < 2.2e-16

Considering the linearity assumption, this is questionable for all variables involved. The aforementioned flatness suggests OPS is consistent regardless of play. All variables are quantitative. In terms of the independence assumption, the residual plots for non-inductees, and all players are relatively random, however there are distinct left and right bunches for inductees. As far as equal variances, the residual plots of non-inductees and all players are relatively homoscedastic, whereas the thickness varies for Hall of Famers. The residual histograms are approximately normal, satisfying the normality condition.

Comparing the R^2s to that of WAR, it is clear that a linear model does not do a good job of predicting OPS versus games or years. This is consistent with the aforementioned flatness. A player's OPS generally is higher if they played more games. This does not imply playing games increases OPS, but rather one must be a consistently good hitter in order to be played in many games. This is especially obvious when we index the model by Hall of Fame induction, wherein Hall of Famers who played more games actually average a lower OPS. This is perhaps because most (batting) Hall of Famers should be standout hitters with high OPS and longer careers. Thus, they will play more games and have more at bats. As their career lengthens, their at bats (denominator of OPS) will increase faster than their hits because performance drops with age (numerator of OPS), thereby decreasing what started as a high OPS.

It is interesting to note that linear predictions intersect around .8, which is roughly where inductee OPS converges to as career length increases

OPS of HOF: Comparison of Means of Negro League vs Non-NL Players



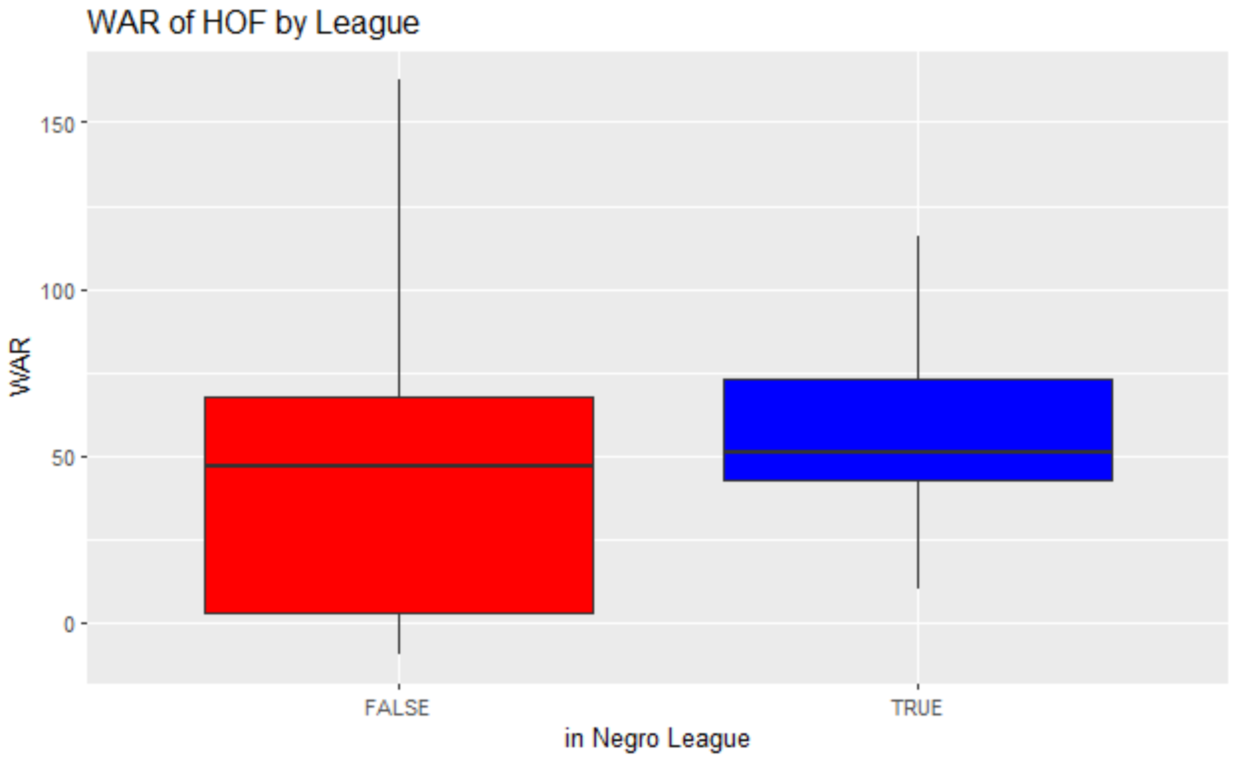
t = -4.473, df = 39.511, p-value = 6.363e-05
alternative hypothesis: true difference in means between non-NL and NL in HOF is not 0
95 percent confidence interval: [-0.2743062, -0.1035214]
mean OPS non-NL HOF = 0.6732187 , mean OPS NL HOF = 0. 8621325

The boxplot shows OPS is generally higher for Negro League inductees.

Considering assumptions: both samples are well above 40, so normality condition does not need to be satisfied. We disregard independence as we are explicitly interested in voting bias.

The p-value well below conventional values suggests we reject the null hypothesis that there is no difference between mean OPS of Negro League and non-Negro League Hall of Famers. It is unlikely that we would observe such a difference in average OPS between Negro League and non-Negro League inductees if it were not a factor when considering induction. This may suggest voters want to err on the side of caution when including players with standout numbers based on their competition. It was noted in part I that other researchers have sought to justify that Negro League statistics are equivalent to other Major Leagues. While voters' caution should not downplay the achievements of Negro League players, we will consider this for our research. Observing the upper 75% of Negro League players is near the median of all Hall of Famers (.7802 versus .7268), the analysis will use the latter as a conservative lower threshold for OPS.

WAR of HOF: Comparison of Means of Negro League vs Non-NL Players



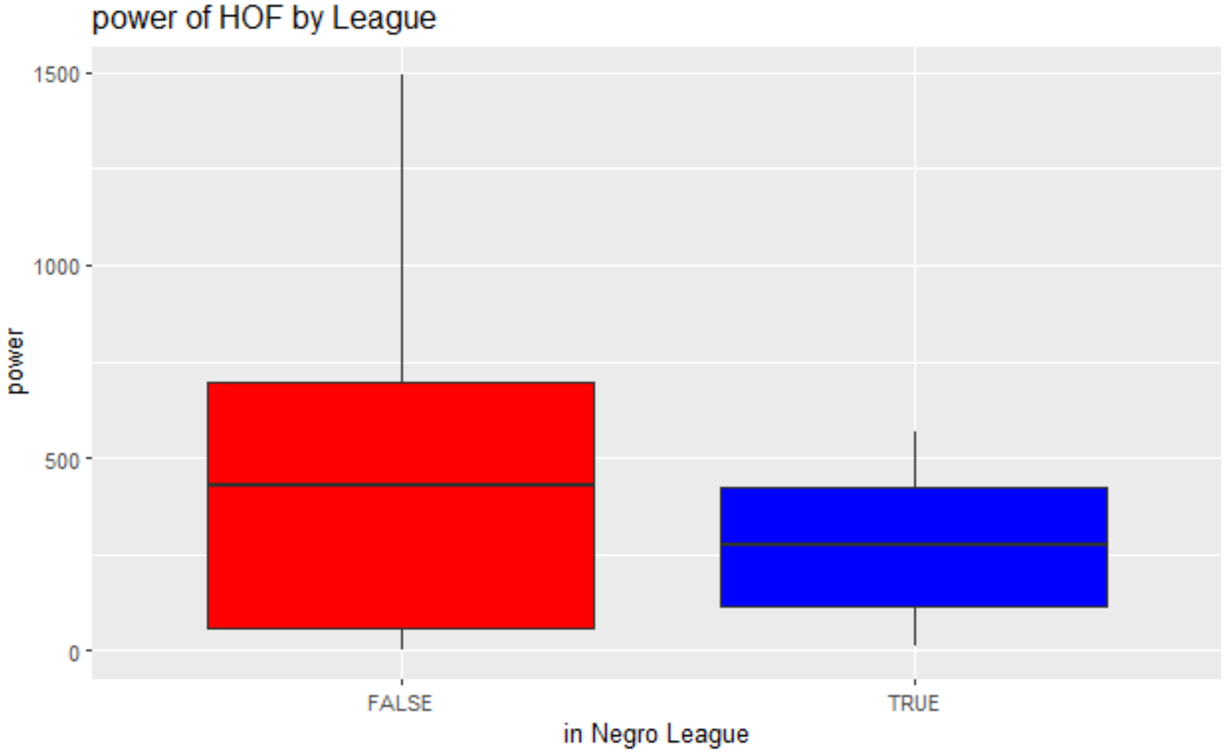
t = -2.3744, df = 36.836, p-value = 0.0229
alternative hypothesis: true difference in means between group FALSE and group TRUE is not 0
95 percent confidence interval: [-24.878993, -1.966672]
mean WAR non-NL HOF = 41.88828, mean WAR NL HOF = 55.31111

The boxplot shows WAR is generally higher for Negro League inductees. All Negro League players have WAR that would land in the upper 75% of non-Negro League players.

Considering assumptions: both samples are well above 40, so normality condition does not need to be satisfied. We disregard independence as we are explicitly interested in voting bias.

The p-value well below conventional values suggests we reject the null hypothesis that there is no difference in mean WAR between Negro League and non Negro League inductees. It is unlikely that we would observe such a difference in average WAR between Negro League and non-Negro League inductees if it were not a factor when considering induction. This may suggest voters err on the side of caution when considering WAR as reasoning for induction. However, all Negro League players’ WARs fit within the range of non-Negro League players. While voters might want Negro League players to have higher WAR on average, they do not expect them to have higher WAR overall. Thus, the linear model from analysis 3 provides better information on inductees’ WAR than just comparing the means. We will focus on Negro League players who outperform their predicted WAR rather than those who just have a high WAR by itself.

Power of HOF: Comparison of Means of Negro League vs Non-NL Players



t = 4.0354, df = 50.367, p-value = 0.0001854
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval: [79.92992, 238.29230]
mean power non-NL HOF= 416.000, mean Power NL HOF = 256.8889

In contrast to the previous two analyses, the boxplot shows power is generally higher for non-Negro League players. This is surprising considering WAR is also an accumulated statistic in which voters tend to expect higher WAR.

Considering assumptions: both samples are well above 40, so normality condition does not need to be satisfied. We disregard independence as we are explicitly interested in voting bias.

The p-value well below conventional values suggests we reject the null hypothesis that there is no difference between the mean power of Negro League and non-Negro League inductees. It is unlikely that we would see such a difference in average WAR between Negro League and non-Negro League inductees if it were not a factor in voting. However, here the story is flipped from the last two analyses. This may be because counting statistics like homeruns + doubles are very inconsistent across eras or leagues based on talent, rules, ballpark sizes, etc. While high power is a good indicator of induction, it may be used less by voters when scoring environments are wildly different (as is the case between Negro and non-Negro Leagues). Voters also may be more lenient with power acknowledging the lack of data of Negro Leagues primarily affects career totals.

VII. The Dick & Dobie Case

A data frame was created for all Negro League players who fit the criteria above except for WAR. The expected WAR was calculated for each player based on games played, years in the league, and power. Then candidates who met their expected WAR but were not inducted were isolated. They are Bill Hoskins, Bill Pettus, Dick Lundy, Dobie Moore, Frank Austin, Henry Kimbro, Howard Easterling, Pythias Russ, Steel Arm, and Ted Strong.

Of these, Dick Lundy and Dobie Moore was chosen as the candidate of interest primarily because even having the highest WAR among potential candidates places Lundy & Moore in the lower quartile of Hall of Famers, thus even they are fringe candidates. Dobie Moore, having nearly half the games and at bats as Lundy, arguably has arguably the more impressive WAR, but at the number of games played this analysis compared him directly to Mart Dihigo. Meanwhile, the analysis compared Lundy to Judy Johnson. While the Hall of Fame would certainly acknowledge that many Negro League Hall of Famers exceeded many of their Major League counterparts, this analysis used Negro League comparisons to meet the voters where they already are at with retroactive induction.

These comparisons were chosen primarily because in both cases, their WAR are or are nearly identical. While not identical, at bats were close in both comparisons. Lundy had a higher OPS than Johnson, and while Dihigo best's Moore in this category, Moore's .908 would make him only the third Hall of Famer above .900 when restricting to under 500 games played. It was then determined how many more at bats each non-inductee would need to match their respective comparison. Both were negligible, and in the case of Moore, he actually had higher career totals than Dihigo. To quantify this insignificance for Lundy's comparison, he would need 148 games and 586 at bats (approximately one season) over the course of his 20+ year career of criminally awful gameplay just to match Judy Johnson's statistics: resulting in a batting average = 0.167, on-base percentage = 0.167, slugging = 0.234, OPS = 0.401 (essentially half his career average). It is important to note that one cannot simply add performance that didn't happen to a finished career, this analysis simply shows that the differences in their actual statistics are negligible given that if Lundy had an extra horrible year's worth of games in his career, he would still be on par with another Negro League inductee.

With this established basis that statistics cannot be the hold up for Dick & Dobie's induction, we turn to narrative. Judy Johnson, a 1975 inductee, played 18 seasons of elite third base, and made additional contributions as a manager, scout, coach. He won 3 pennants, the 1925 Colored World Series, the 1929 Chicago Defender and Pittsburgh Courier Negro Leagues MVP, and served on the Hall of Fame Committee (Baseball Hall). It is worth noting that he was inducted the year he stepped down from the committee, suggesting there could be an unfair burden of extracurricular demands placed on a similar Lundy candidacy. Johnson may also have received a boost in vote favor from his participation in the 1932 Pittsburgh Crawfords, a team that saw five members inducted (MiLB).

Dick Lundy was a player manager like his inducted counterpart, winning 2 pennants as the team's premier shortstop, played in but did not win a World Series, and made manager and coach contributions after his playing career (NLB emuseum). Both Lundy and Johnson were arguably the best of their era at premium defensive positions. While Lundy does not have the team success of Johnson, he also never had four future Hall of Famers playing alongside him. Without the narrative boost of working with the Hall of Fame, Lundy and Johnson offer extremely similar individual contributions to the history of baseball. Speaking of individual ability, the Hall of Fame recognizes Johnson was not a power hitter on its own website, whereas Lundy hit for contact and power reliably throughout his career. This is supported by the career average comparisons in the previous part. On the field, Lundy provided every bit the body of work demonstrated of a deserving inductee.

Martin Dihigo, a 1977 inductee, played 20+ seasons of standout second base (Baseball Hall). He is most strongly recognized for his prowess as a two way player and his influence on the sport in numerous Latin American countries (SABR). Dobie Moore was a dynamic hitting shortstop, led his team to 3 pennants and a World Series win (NLB emuseum). Most notably, his career was cut short due to an off-field shooting injury. It is with this detail that despite the similarity in statistics, Moore's case actually works better in comparison to Johnson. While the difference in games played complicates how we view rate statistics, Dihigo's Hall of Fame story cannot be viewed without incorporating his pitching statistics. While Moore's rate statistics obviously compare favorably to that of Johnson, further research should focus on comparison of these players' rate statistics and WAR to further solidify the comparison. Nevertheless, Dick Lundy and Dobie Moore offer compelling examples of possibly overlooked Hall of Fame candidates. Their impact on their leagues cannot be understated, particularly when compared to the likes of other inductees. It is with these findings that I propose the induction committee strongly consider thorough reexamination of the historical precedents established by the lack of induction of certain qualified, historical candidates.

VIII. Sources

Ashwell, Gary. "Top 500 Negro League Players by GWAR." https://www.seamheads.com/NegroLgs/history.php?tab=metrics_at&first=1886&last=1948&lgID=All&lgType=N&bats=All&pos=All&HOF=All&results=500&sort=Tot_a. Accessed 20 Mar. 2022.

Ashwell, Gary. "Top 500 Negro League Players by OPS." *Negro Leagues Database*, Seamheads, https://www.seamheads.com/NegroLgs/history.php?tab=bat_basic_at&first=1886&last=1948&lgID=All&lgType=N&HOF=All&pos=All&bats=All&minPA=500&results=500&sort=OPS2_a. Accessed 21 Mar. 2022.

"Hall of Fame Batting Register." *Baseball Reference*, https://www.baseball-reference.com/awards/hof_batting.shtml. Accessed 14 Mar. 2022.

"Judy Johnson." *Baseball Hall of Fame*, Baseball Hall of Fame, https://baseballhall.org/hall-of-famers/johnson-judy.

Lahman, Sean. "2006." *Lahman's Baseball Database*, www.seanlahman.com, https://www.seanlahman.com/baseball-archive/statistics/. Accessed 17 Mar. 2022.

"Martin Dihigo." *Society for American Baseball Research*, Admin /Wp-Content/Uploads/2020/02/sabr_logo.Png, 22 Aug. 2022, https://sabr.org/bioproj/person/martin-dihigo/.

Negro Leagues Baseball Emuseum: Personal Profiles: Dick Lundy, Negro Leagues Baseball Emuseum, https://nlbemuseum.com/history/players/lundy.html.

Negro Leagues Baseball Emuseum: Personal Profiles: Dobie Moore, https://nlbemuseum.com/history/players/moorew.html.

"The Negro Leagues Are Major Leagues." *Baseball Reference*, Baseball Reference, https://www.baseball-reference.com/negro-leagues-are-major-leagues.shtml.

"Wilmington Blue Rocks Legacy of Judy Johnson." *MiLB.com*, MiLB, https://www.milb.com/wilmington/community/legacy-of-judy-johnson/.