

Moneyball Report

Introduction

From the hobby of fans keeping scorecards at the ballpark to the modern implementation of sabermetrics, baseball has always held a special relationship with statistics. However, are the historical data collected an accurate measure of a team's performance? This unsupervised study looks at standard MLB team records over the past century to attempt to identify latent traits through Factor Analysis that explain larger groups of variables. For example, I hypothesize that there are three underlying factors that correlate with the statistics collected in the Moneyball dataset. An offensive factor may be described by the typical hitting records: hits, doubles, triples, homeruns, walks, and strikeouts. A pitching factor may be described by the typical pitching records: strikeouts, walks, hits, and homeruns allowed. A defensive factor may be described by typical fielding records: errors and double plays. Finally, once the factors have been defined, I want to apply them in a linear regression model to measure their ability to predict team wins.

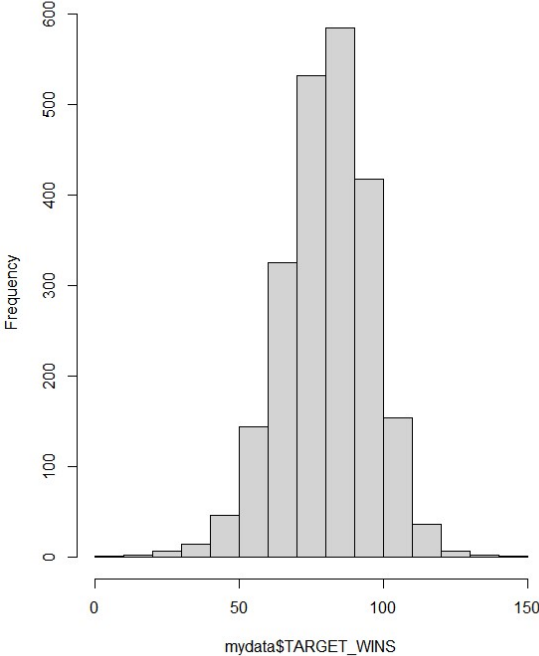
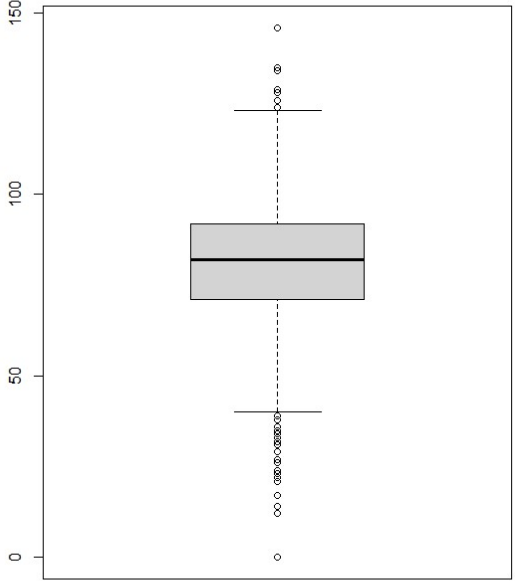
EDA

The Moneyball dataset contains professional baseball records from 1871-2006. There are 2,276 regular season team statistics. Because the regular season length has varied over time, all statistics are adjusted for the modern 162-game season. Perhaps more important than what is included in the data, is what has been left out. The records are not dated, and there is no denotation marking which result have been altered. The table below lists variables measured in this dataset along with their descriptions and theoretical impacts on win totals.

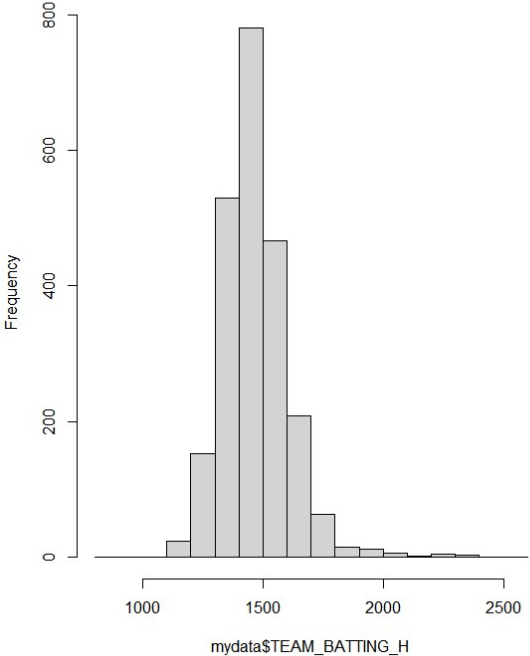
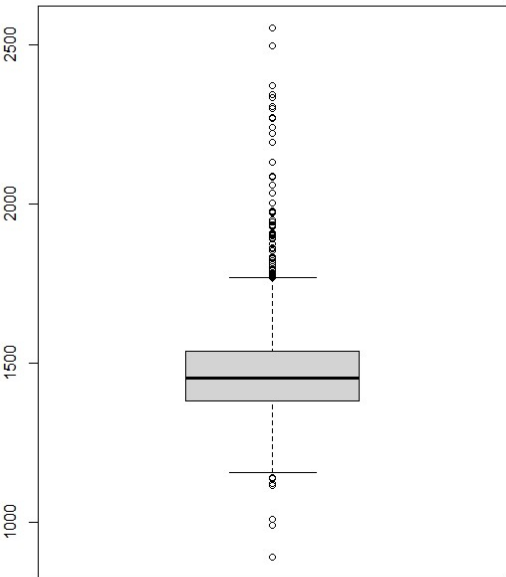
Moneyball Data Dictionary

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
TEAM_BATTING_H	Base Hits by batters	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

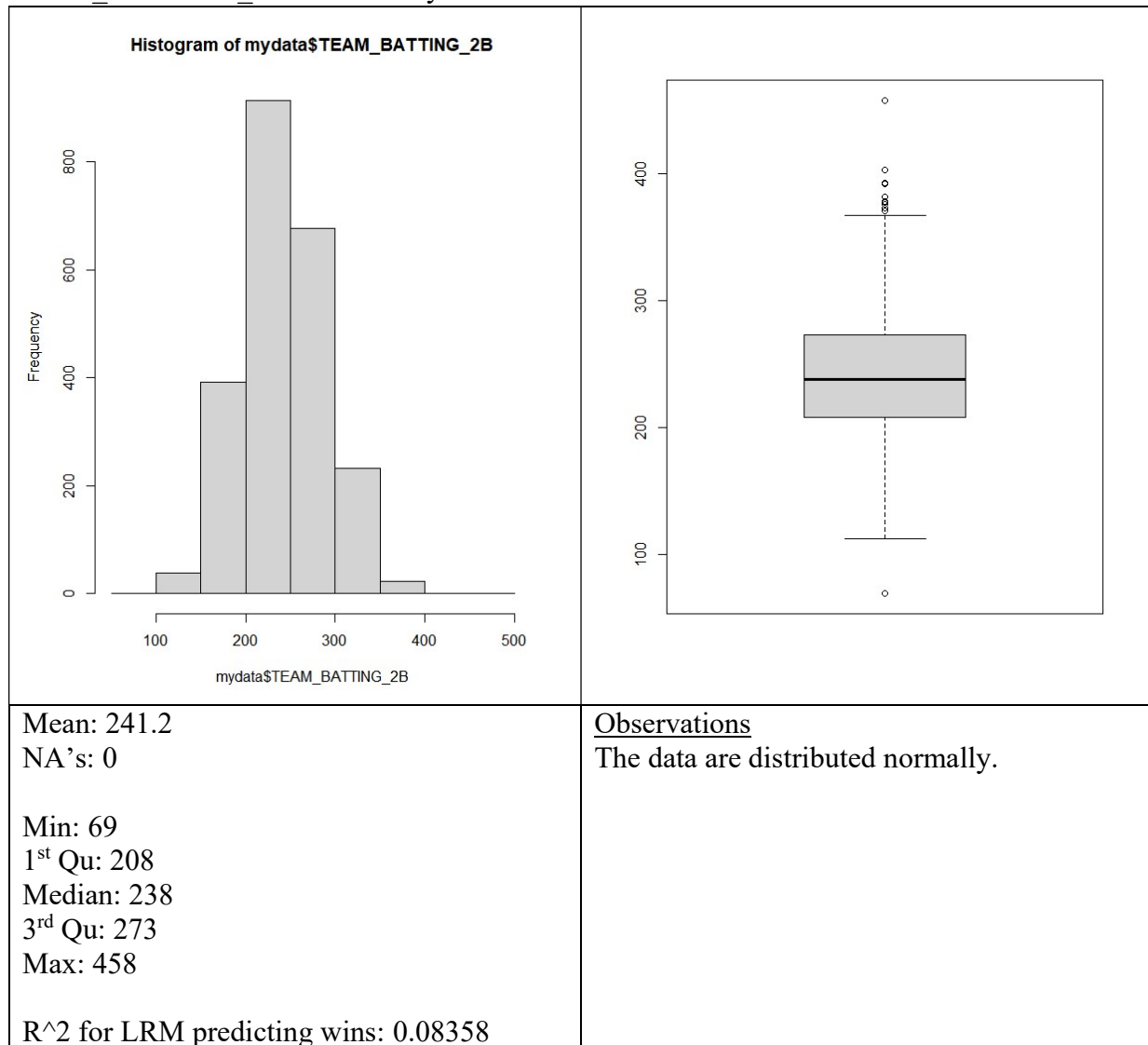
TARGET_WINS- Wins over 162 game season

<p>Histogram of mydata\$TARGET_WINS</p> 	
<p>Mean: 80.79 NA's: 0 Min: 0 1st Qu: 71 Median: 82 3rd Qu: 92 Max: 146</p>	<p><u>Observations</u> The maximum and minimum are severe outliers. Winning 0 or 146 games over a season both appears unreasonable.</p>

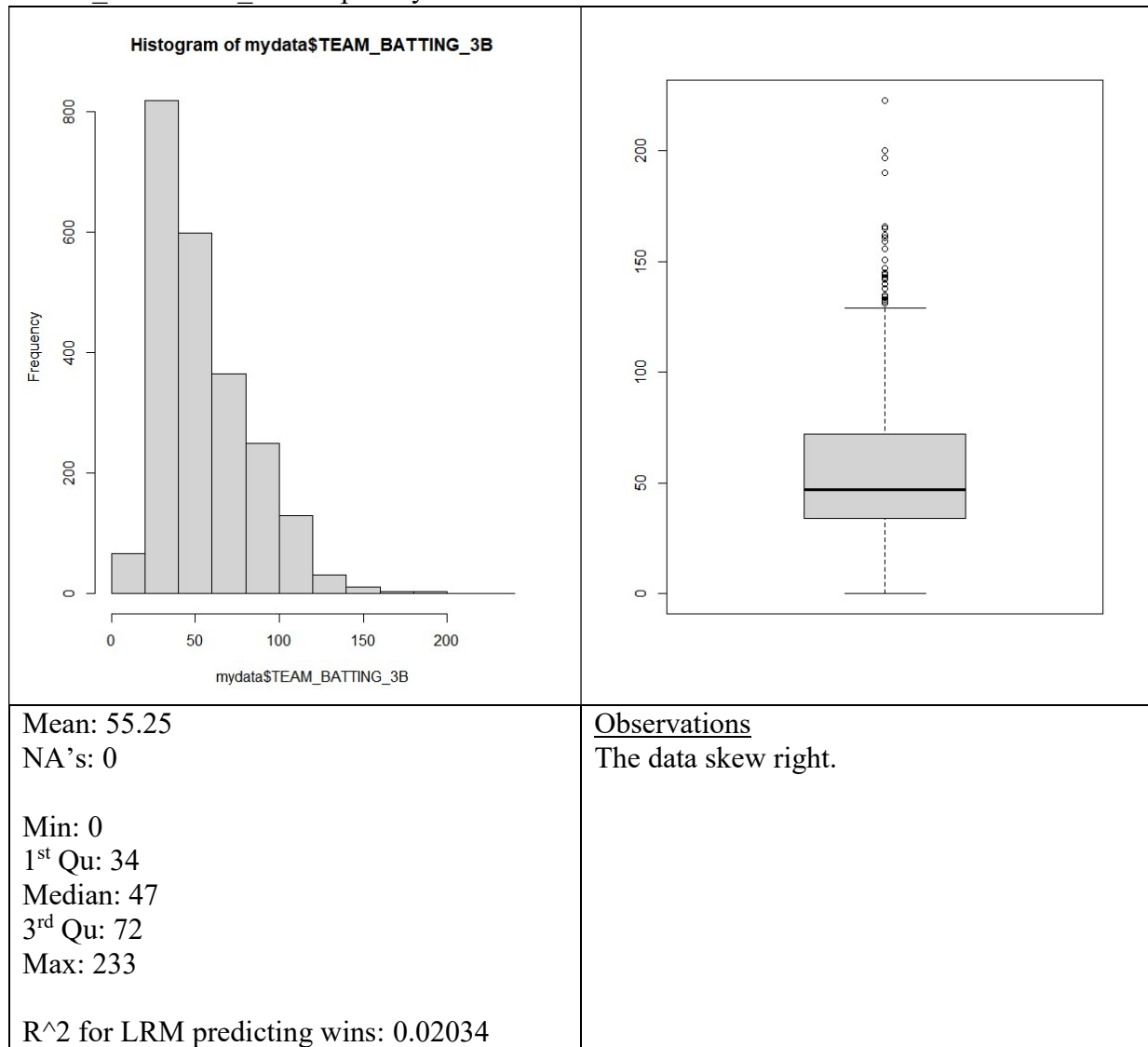
TEAM_BATTING_H- Base Hits by batters

<p>Histogram of mydata\$TEAM_BATTING_H</p> 	
<p>Mean: 1469 NA's: 0 Min: 891 1st Qu: 1383 Median: 1454 3rd Qu: 1537 Max: 2554 R² for LRM predicting wins: 0.1511</p>	<p><u>Observations</u> The distribution right with many positive outliers.</p>

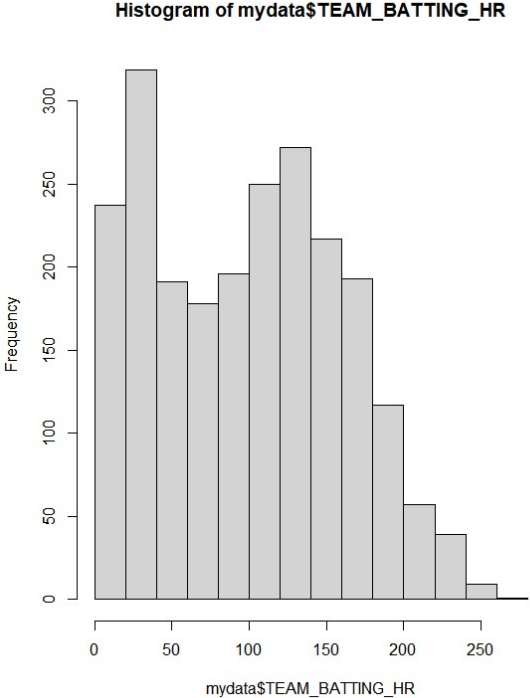
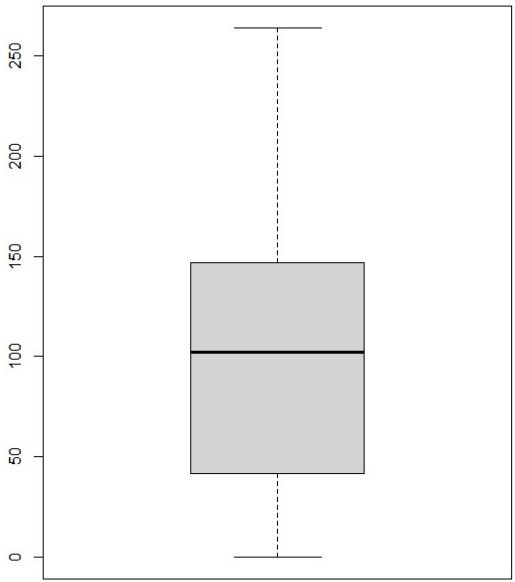
TEAM_BATTING_2B- Doubles by batters



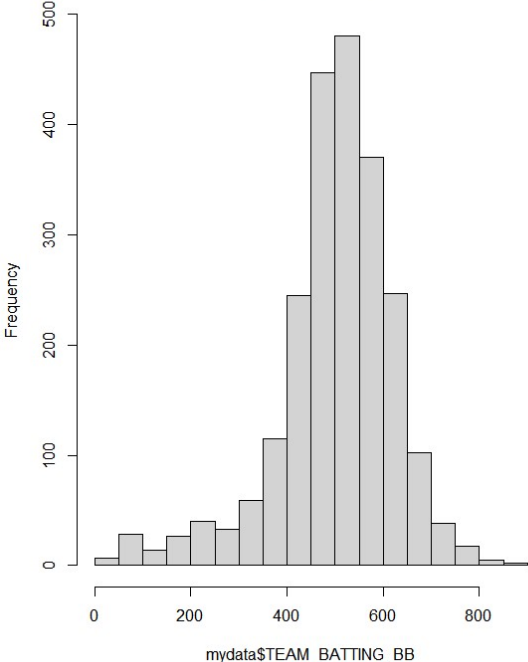
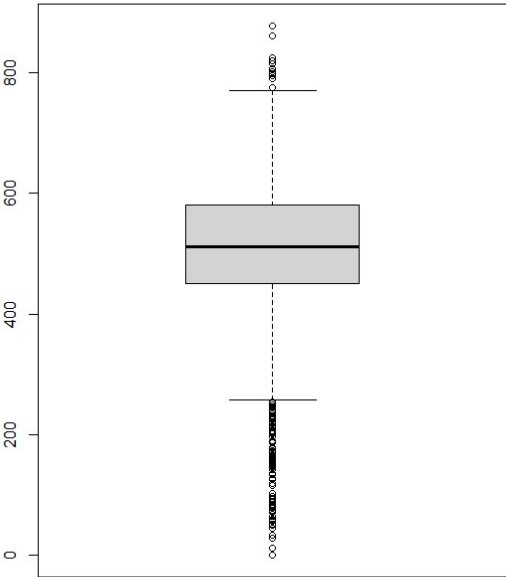
TEAM_BATTING_3B- Triples by batters



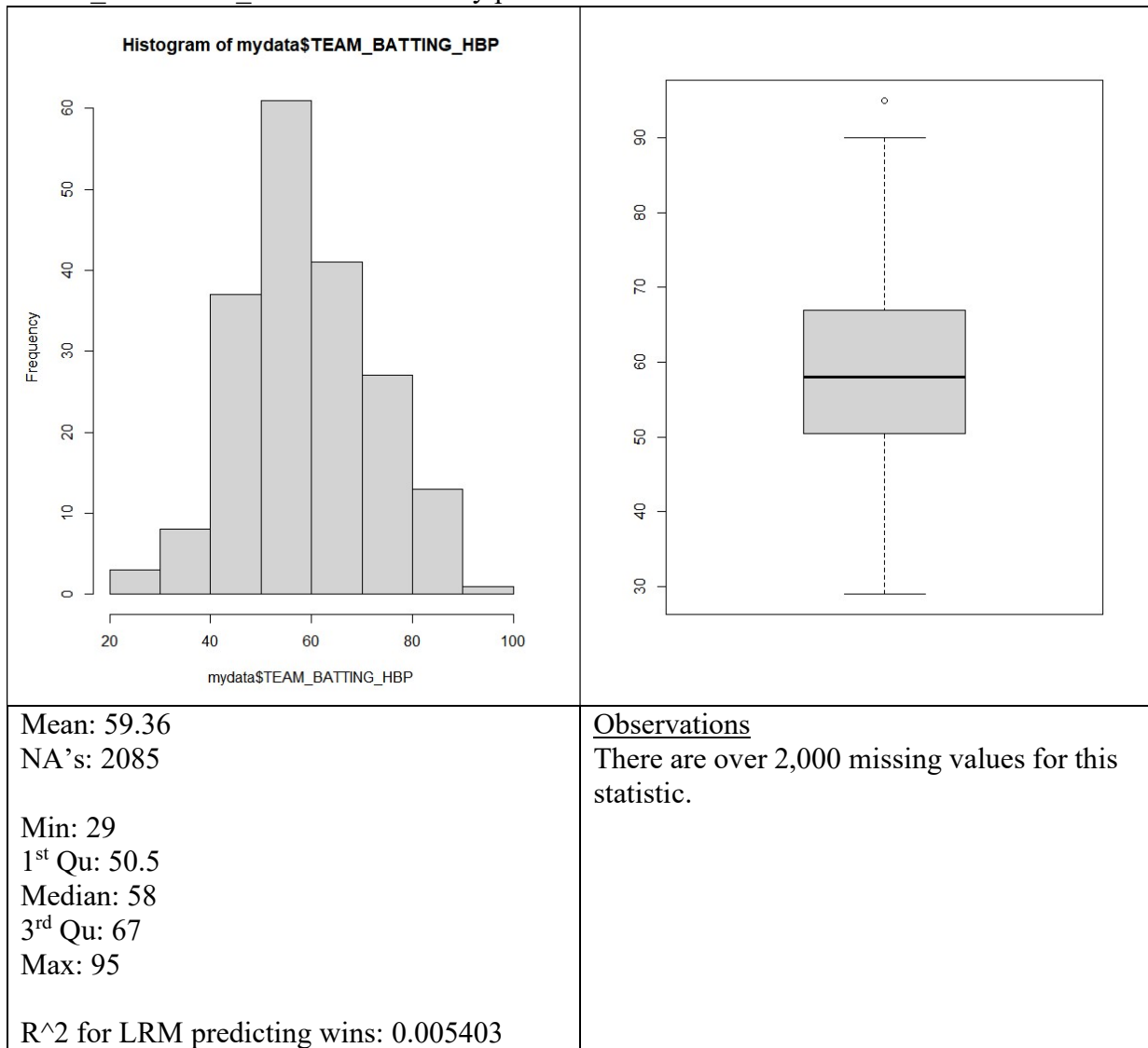
TEAM_BATTING_HR- Homeruns by batters

<p>Histogram of mydata\$TEAM_BATTING_HR</p> 	
<p>Mean: 99.61 NA's: 0</p> <p>Min: 0 1st Qu: 42 Median: 102 3rd Qu: 147 Max: 264</p> <p>R² for LRM predicting wins: 0.03103</p>	<p><u>Observations</u></p> <p>That data do not have a normal distribution.</p> <p>In today's game, it would be unheard of to go an entire season with zero homeruns.</p>

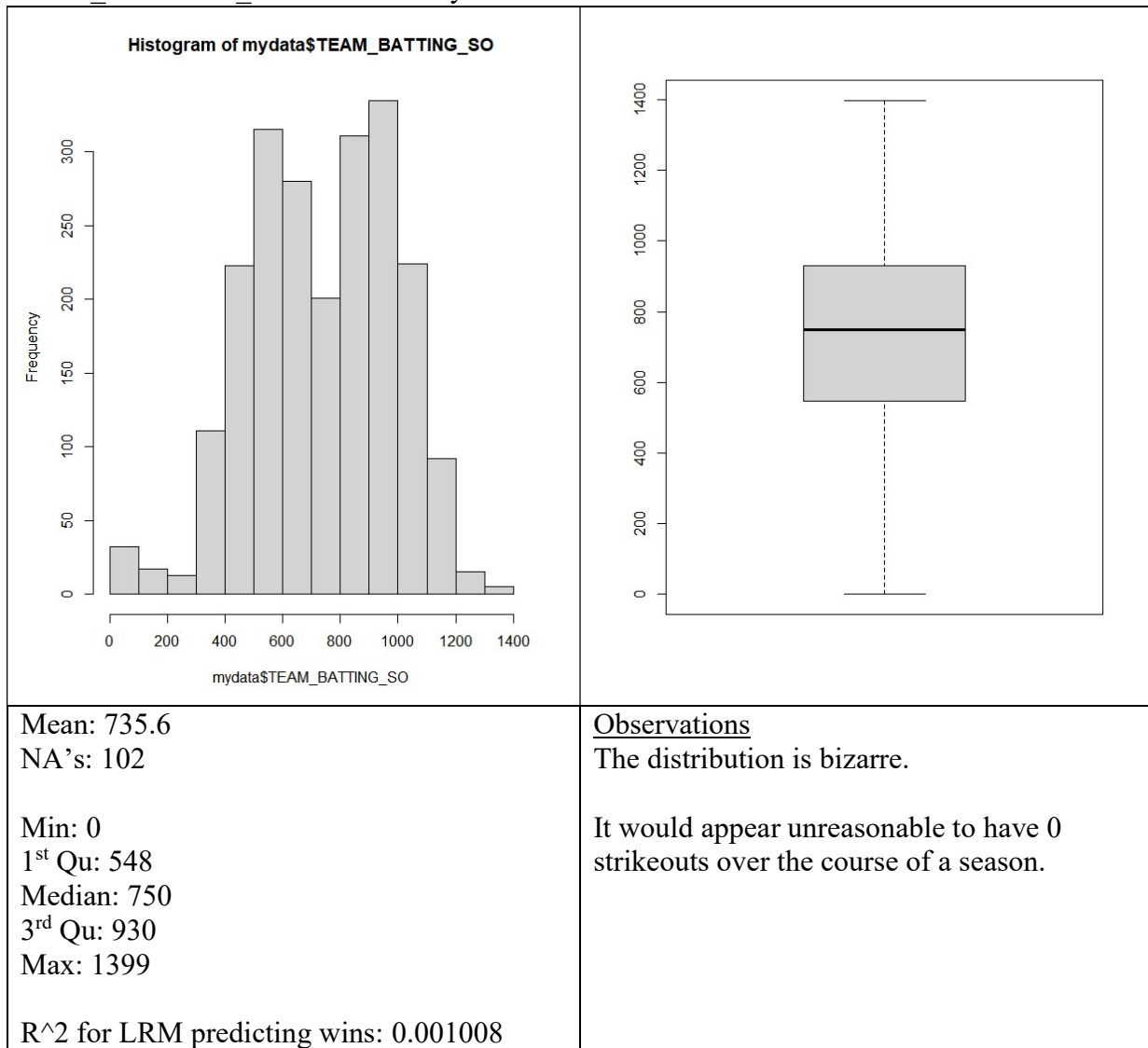
TEAM_BATTING_BB- Walks by batters

<p>Histogram of mydata\$TEAM_BATTING_BB</p> 	
<p>Mean: 501.6 NA's: 0 Min: 0 1st Qu: 451 Median: 512 3rd Qu: 580 Max: 878 R² for LRM predicting wins: 0.05408</p>	<p><u>Observations</u> The distribution skews left. It would appear unreasonable to have 0 walks over the course of a season.</p>

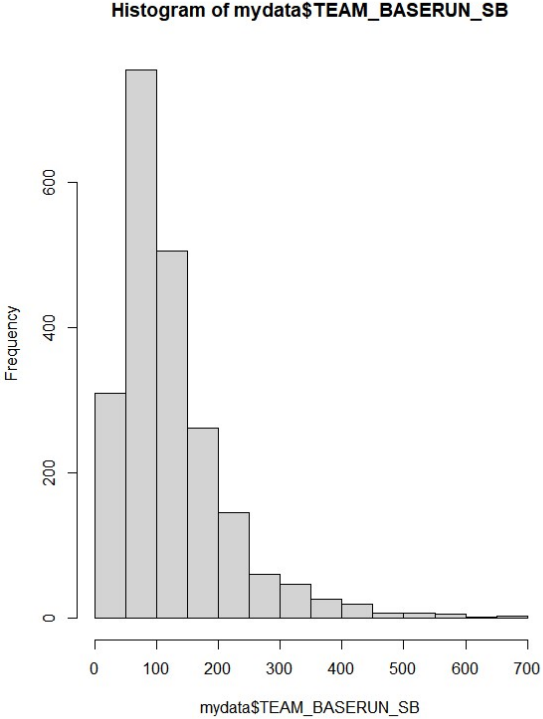
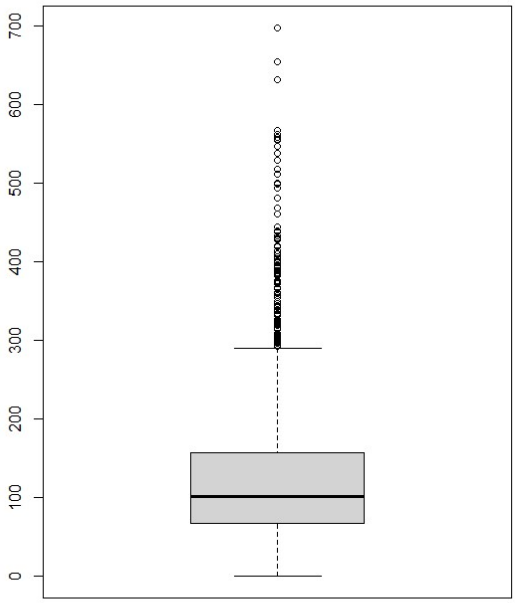
TEAM_BATTING_HBP- Batters hit by pitch



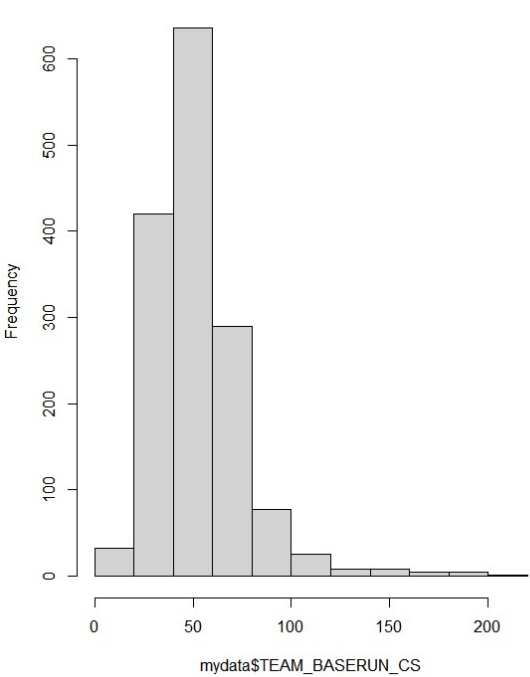
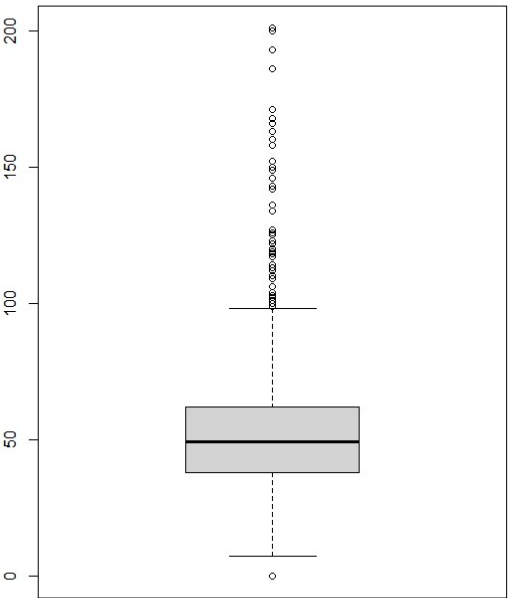
TEAM_BATTING_SO- Strikeouts by batters



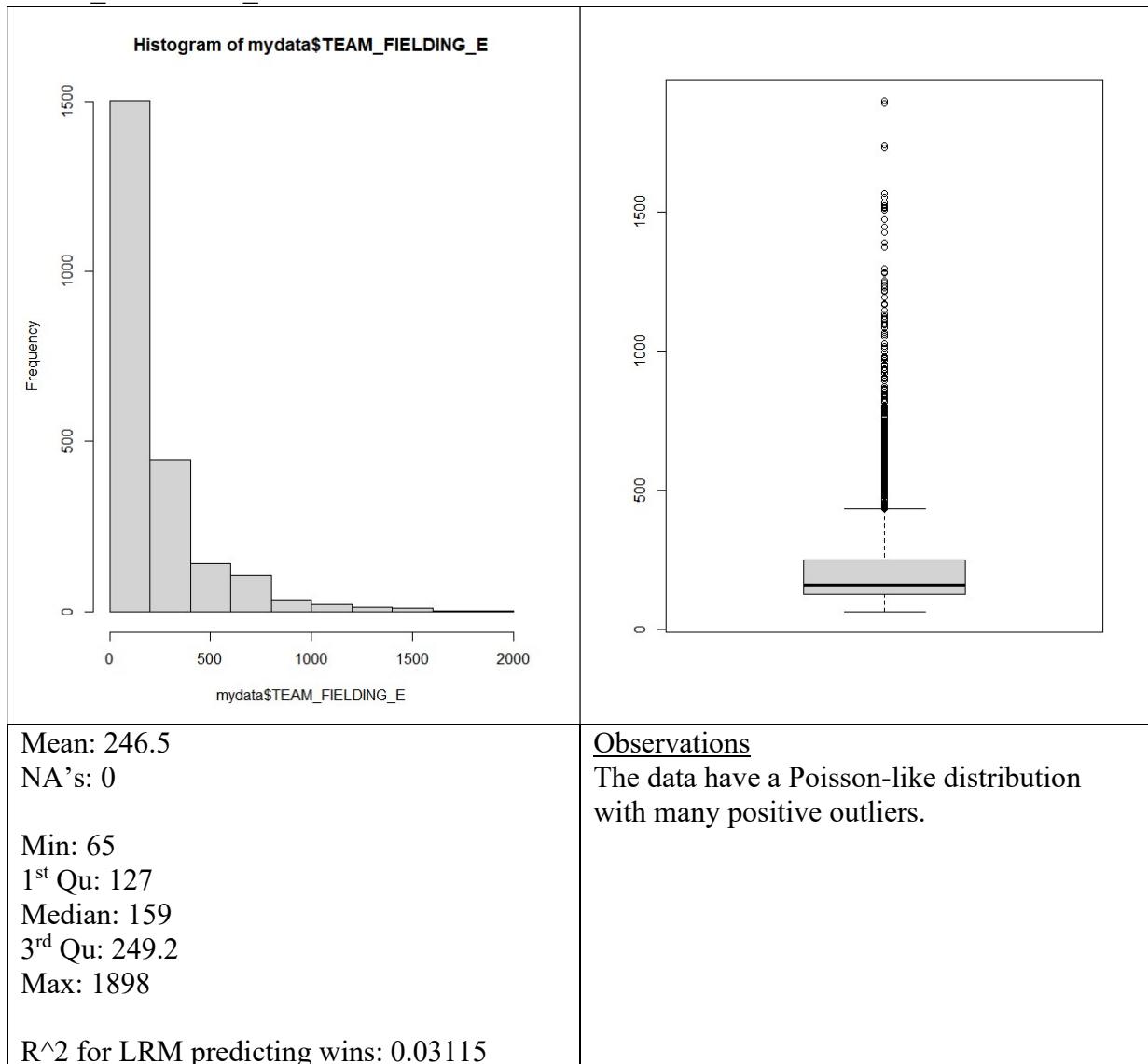
TEAM_BASERUN_SB- Stolen bases

<p>Histogram of mydata\$TEAM_BASERUN_SB</p> 	
<p>Mean: 124.8 NA's: 131 Min: 0 1st Qu: 66 Median: 101 3rd Qu: 156 Max: 697 R² for LRM predicting wins: 0.01826</p>	<p><u>Observations</u> The distribution is skewed right. It would appear unreasonable to have 0 stolen bases over the course of a season.</p>

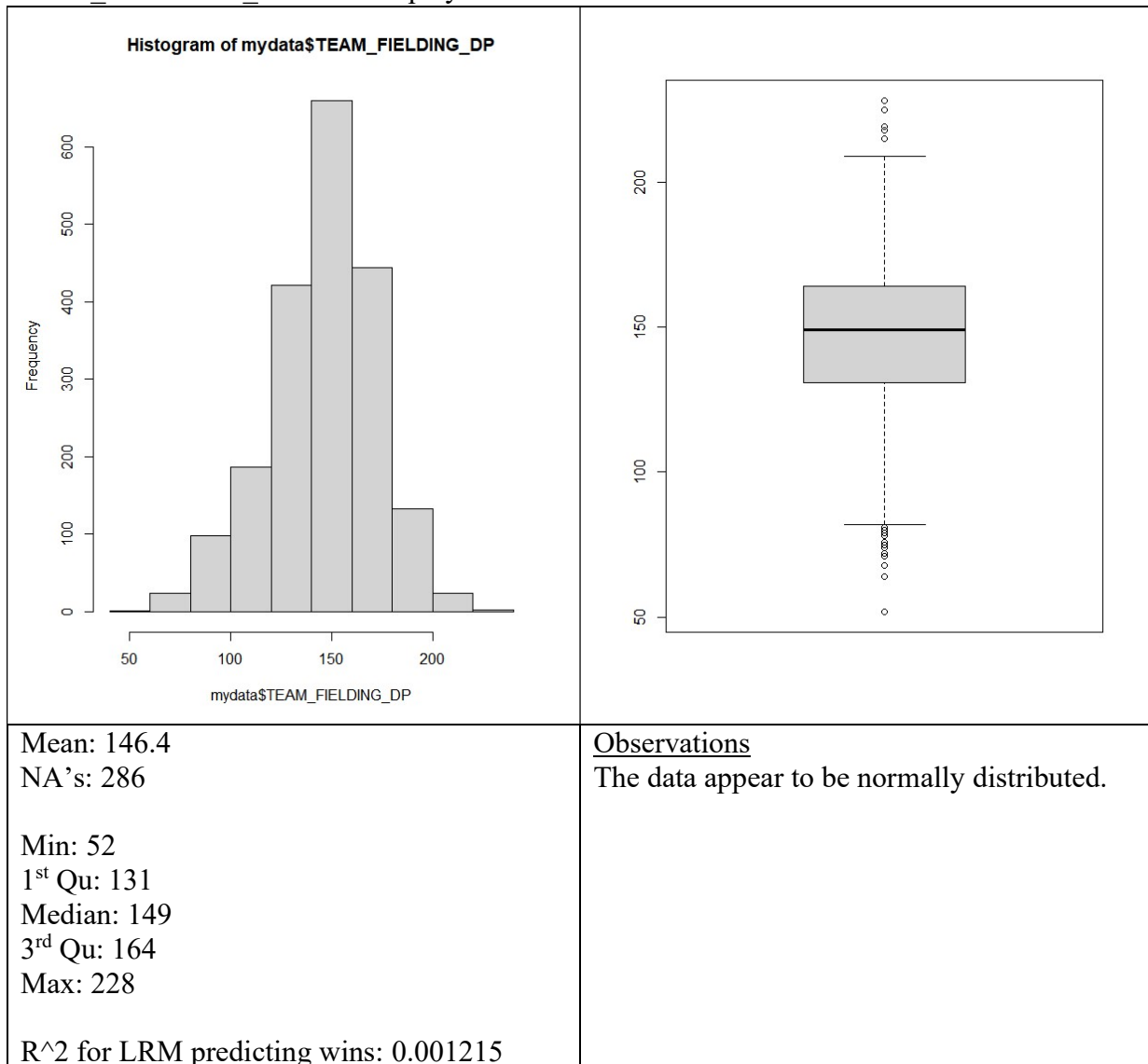
TEAM_BASERUN_CS- Caught stealing

<p>Histogram of mydata\$TEAM_BASERUN_CS</p> 	
<p>Mean: 52.8 NA's: 772 Min: 0 1st Qu: 38 Median: 49 3rd Qu: 62 Max: 201 R² for LRM predicting wins: 0.0005019</p>	<p><u>Observations</u> It would appear unreasonable to never get caught stealing over the course of a season. There are over 700 missing values.</p>

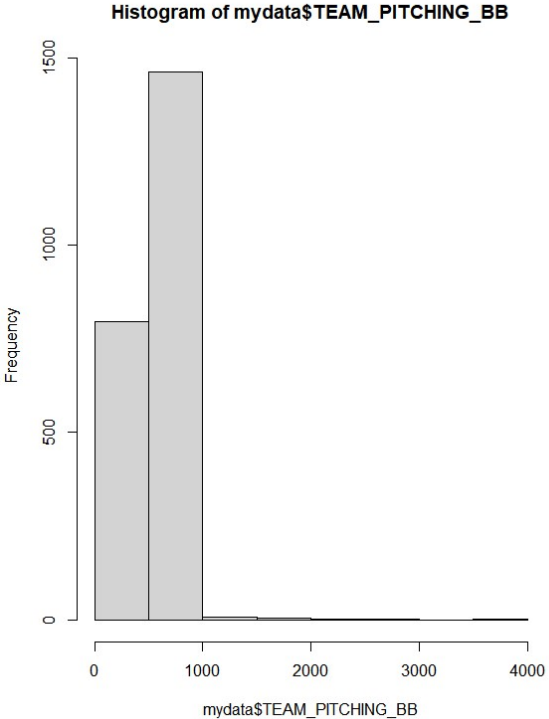
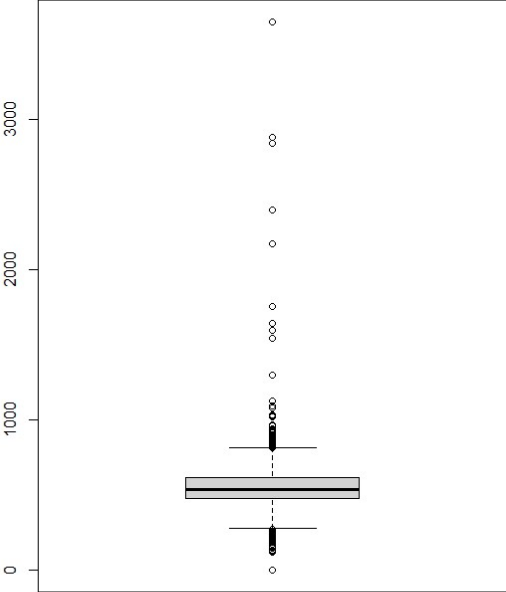
TEAM_FIELDING_E- Errors



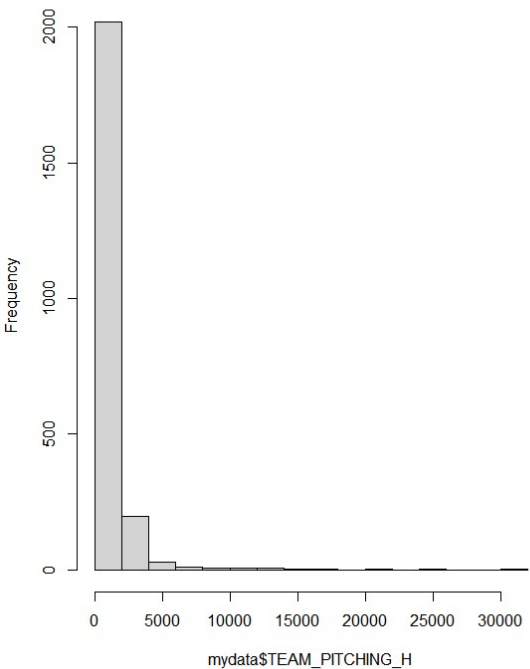
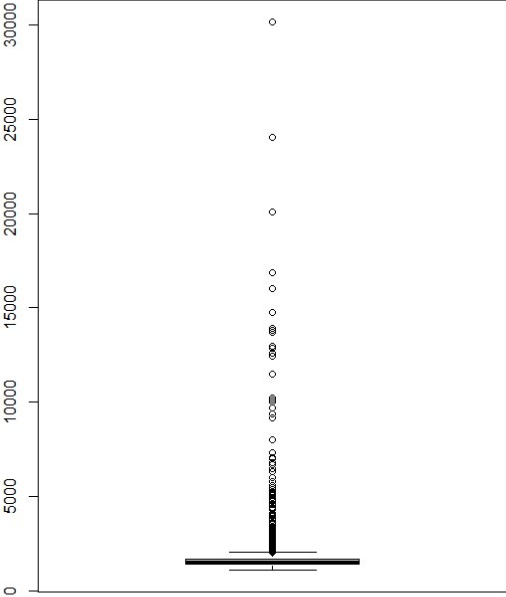
TEAM_FIELDING_DP- Double plays



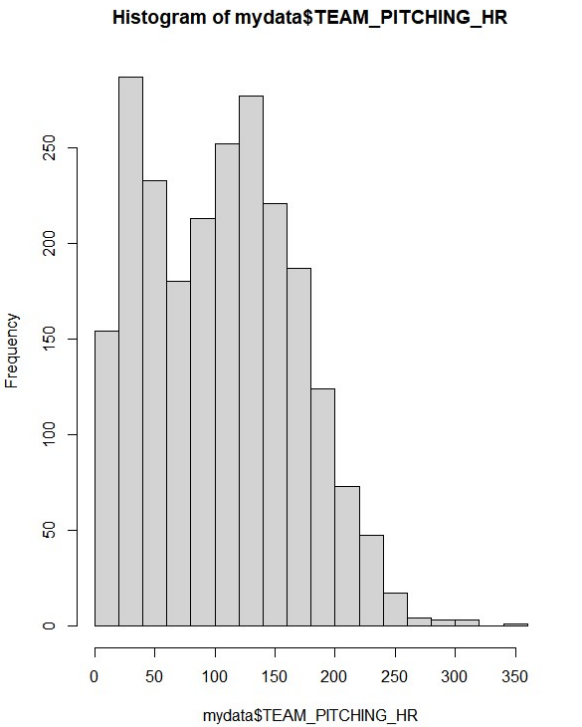
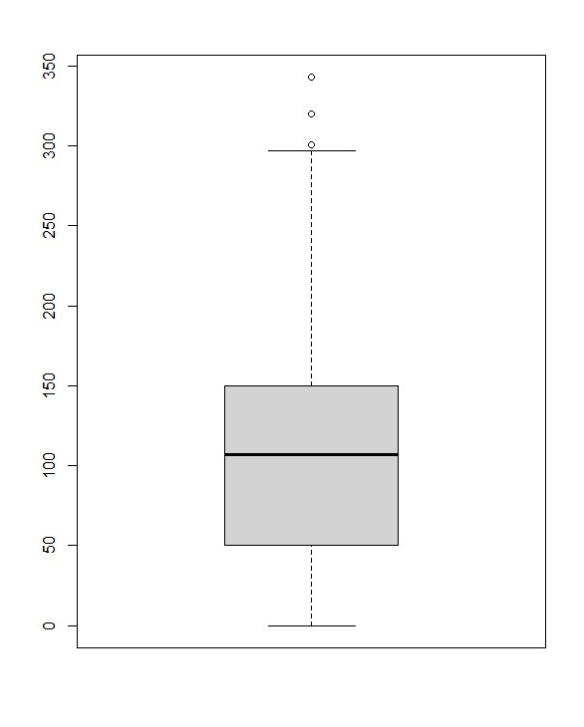
TEAM_PITCHING_BB- Walks allowed

	
<p>Mean: 553</p> <p>NA's: 0</p> <p>Min: 0</p> <p>1st Qu: 476</p> <p>Median: 536.5</p> <p>3rd Qu: 611</p> <p>Max: 3645</p> <p>R² for LRM predicting wins: 0.01542</p>	<p><u>Observations</u></p> <p>It would appear unreasonable to allow 0 walks over the course of a season.</p> <p>The data have a bizarre distribution.</p>

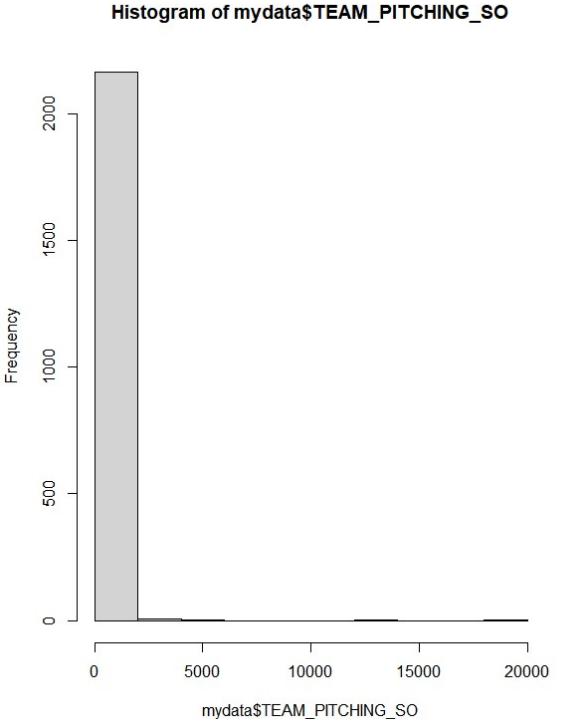
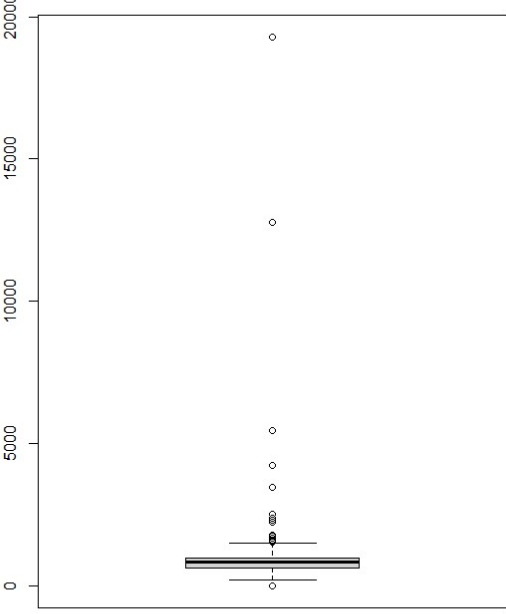
TEAM_PITCHING_H- Hits allowed

<p>Histogram of mydata\$TEAM_PITCHING_H</p> 	
<p>Mean: 1779 NA's: 0 Min: 1137 1st Qu: 1419 Median: 1518 3rd Qu: 1682 Max: 30132 R² for LRM predicting wins: 0.01209</p>	<p><u>Observations</u></p> <p>It would appear unreasonable to allow 0 hits over the course of a season.</p> <p>The data take a Poisson distribution and appear to be zero-inflated.</p>

TEAM_PITCHING_HR- Homeruns allowed

<p>Histogram of mydata\$TEAM_PITCHING_HR</p> 	
<p>Mean: 105.7 NA's: 0 Min: 0 1st Qu: 50 Median: 107 3rd Qu: 150 Max: 343 R² for LRM predicting wins: 0.03573</p>	<p><u>Observations</u> The distribution appears bizarre. It would appear unreasonable to allow 0 homeruns over the course of a season.</p>

TEAM_PITCHING_SO- Strikeouts by pitchers

 <p>Histogram of mydata\$TEAM_PITCHING_SO</p>	
<p>Mean: 817.7 NA's: 0</p> <p>Min: 0 1st Qu: 615 Median: 813.5 3rd Qu: 968 Max: 19278</p> <p>R² for LRM predicting wins: 0.006152</p>	<p><u>Observations</u> The data have a bizarre distribution.</p> <p>It would appear unreasonable to strikeout 0 batters over the course of a season.</p>

Data Cleaning

A couple issues are apparent in the Moneyball dataset. First, several variables have high numbers of missing values. These variables with their corresponding counts of NA values are listed below:

- Batting Strikeouts: 102
- Pitching Strikeouts: 102
- Stolen Bases: 131
- Double Plays: 286
- Caught Stealing: 772
- Batting Hit by Pitch: 2085

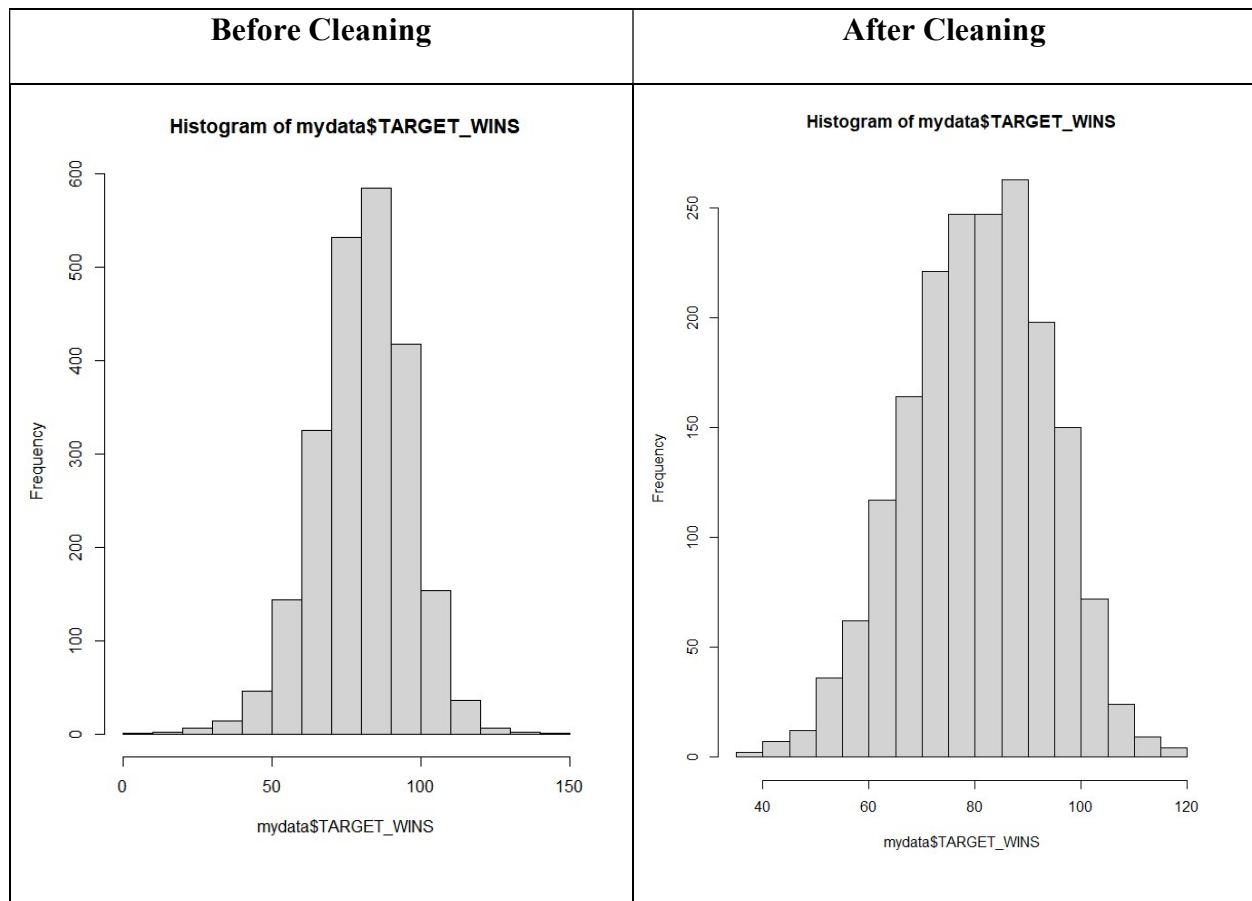
In addition, almost all variables that do not contain missing values appear to have incorrectly used a zero in place of an NA. This created some bizarre distributions for the impacted variables. For example, some teams record zero hits allowed by their pitching staff course of an entire season. Unless there are historical records of a team pitching 162 consecutive no-hitters, action needs to be taken to remove erroneous values.

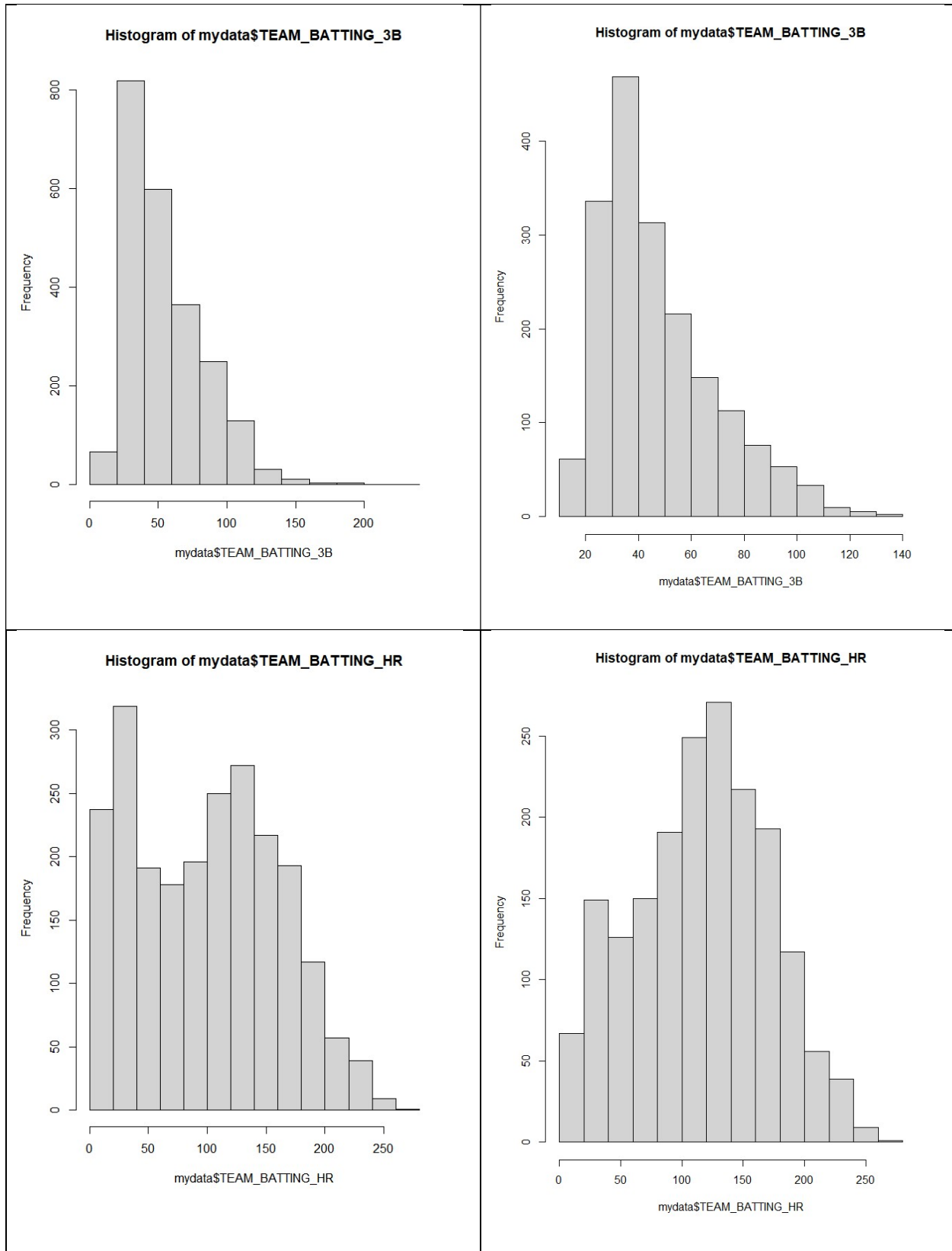
To begin my data cleaning, I simply remove the caught stealing and hit by pitch variables. The missing values here represent a huge portion of the 2,276 total records, and the initial EDA demonstrates that they have little impact on predicting wins.

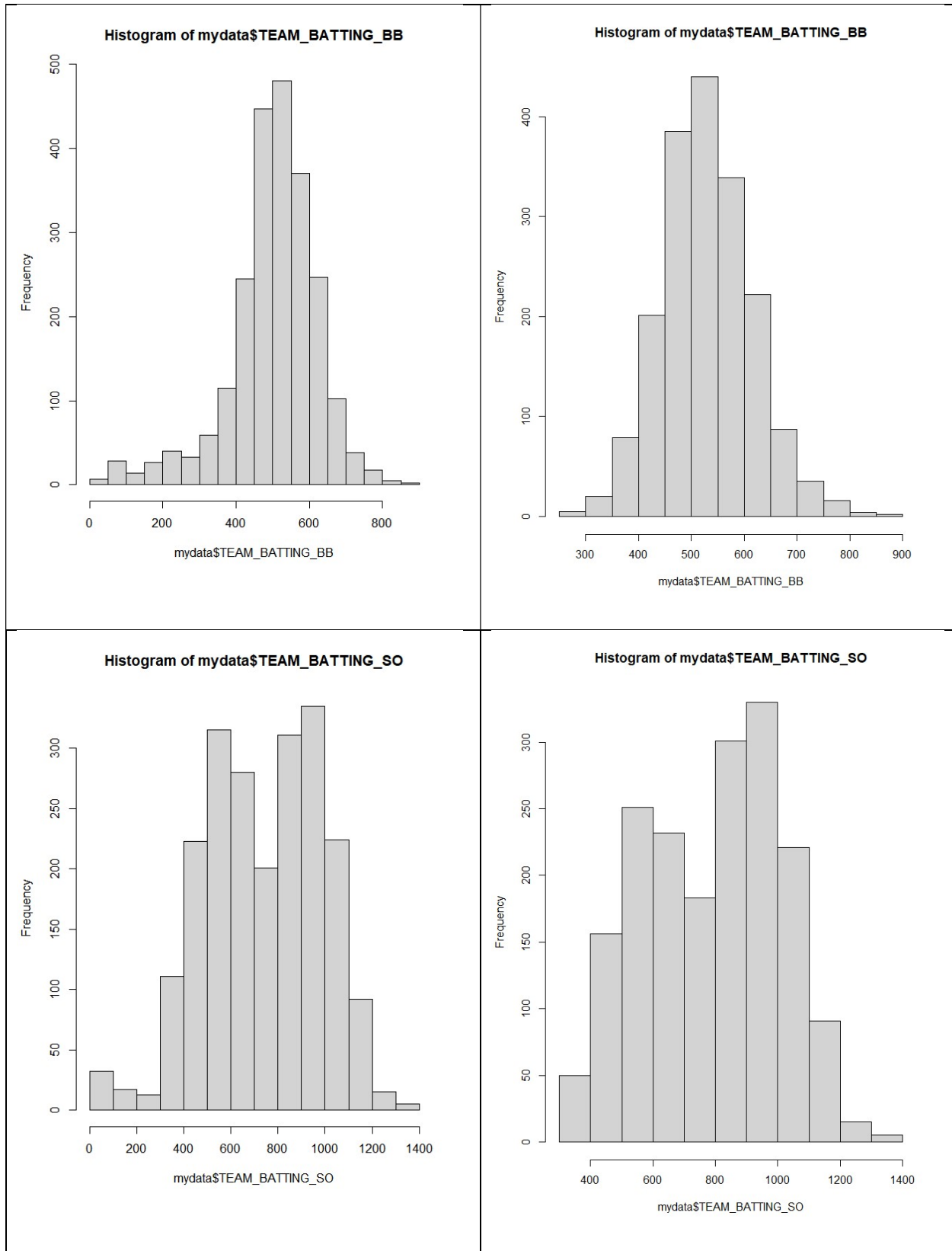
From here, I make a critical assumption with the Moneyball records that date back to 1871. The statistics recorded for this dataset are standard measures that have been consistently collected over the past century and placed on the back of baseball cards. If records like strikeouts or stolen bases are not present, I assume that the data must be significantly dated and perhaps not

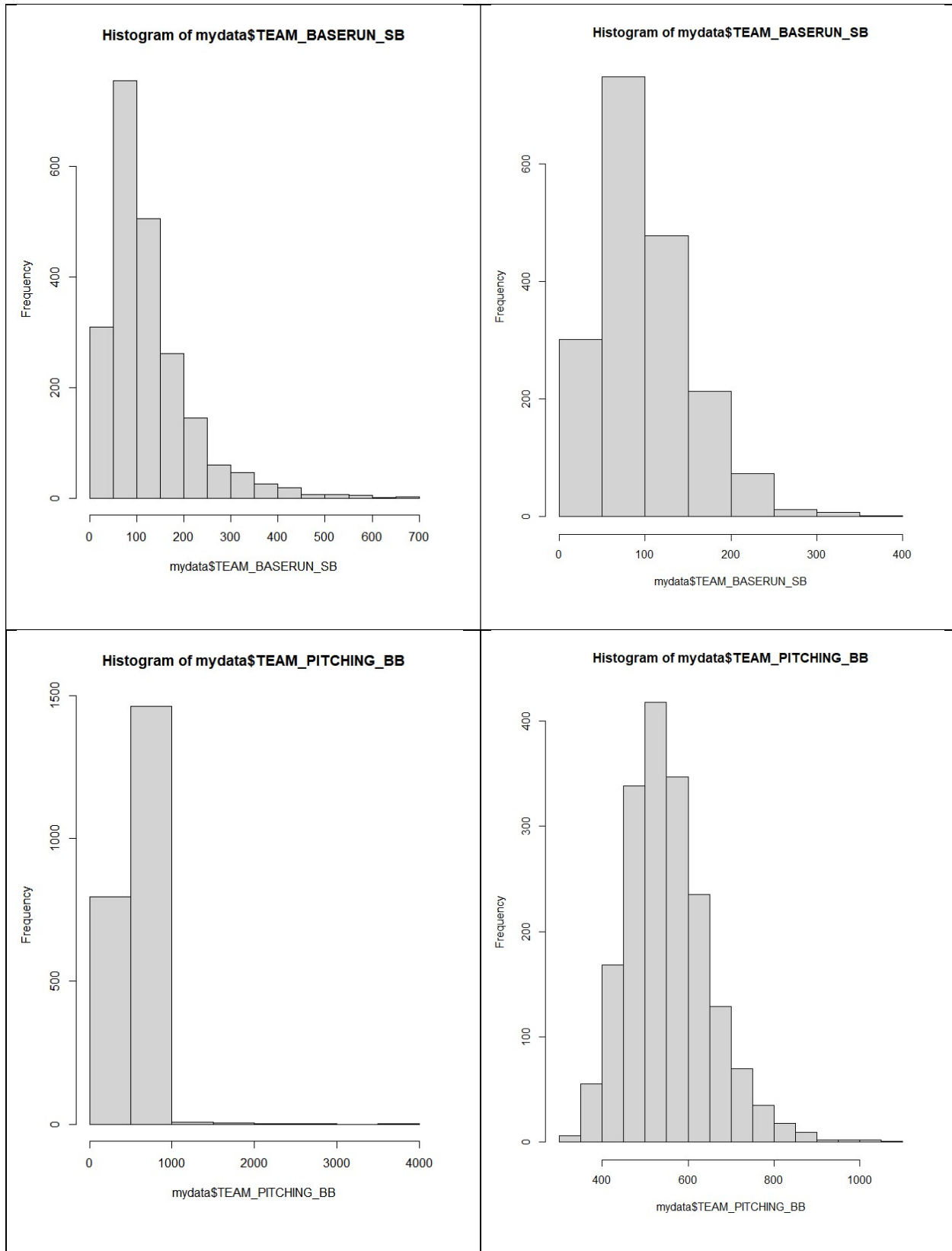
as relevant to the modern game and my analysis. I delete any remaining rows with missing values.

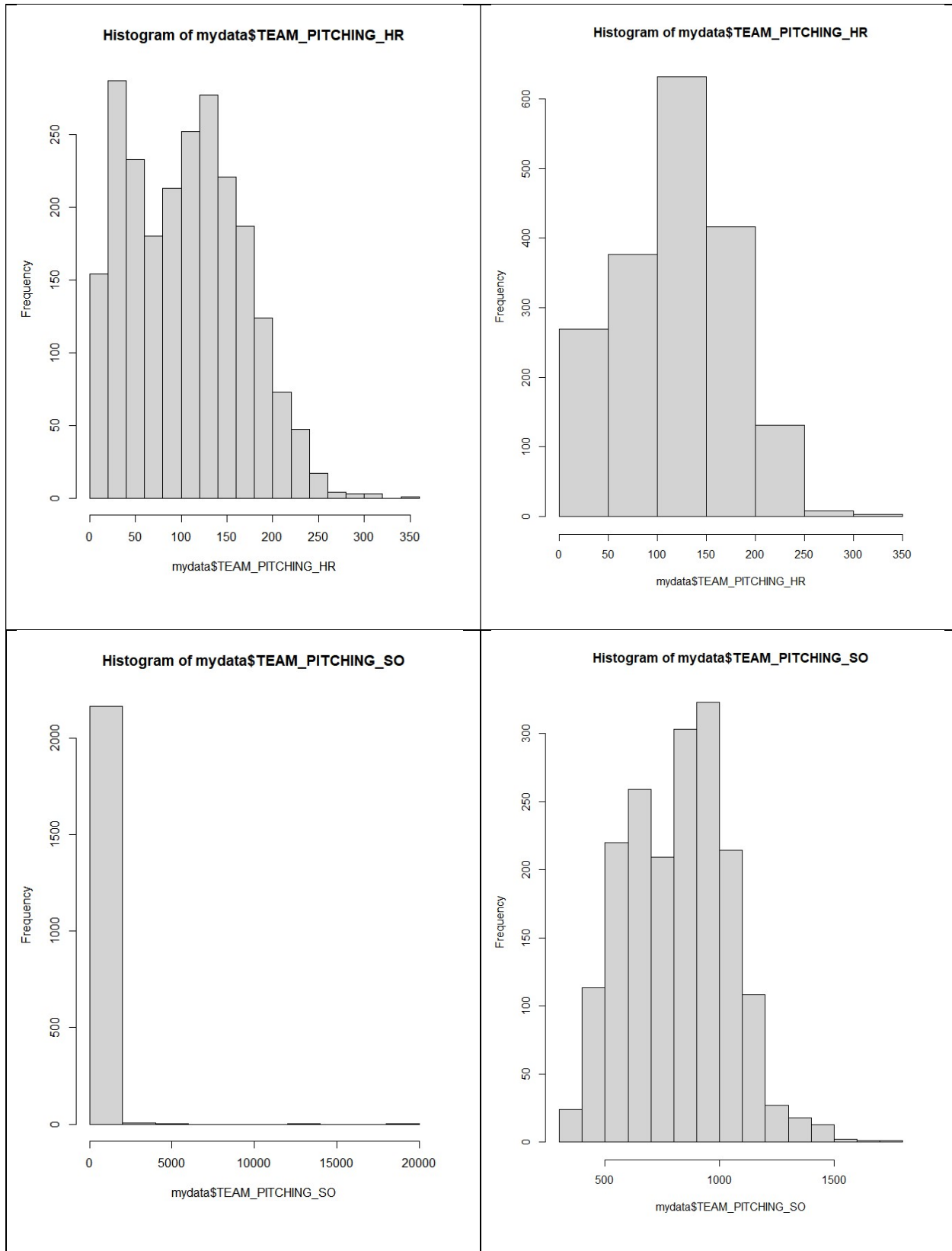
The actions taken for data preprocessing result in a reduced dataset of 1,835 records. In a bit of serendipity, removing the missing rows also corresponded in the removal all the NA records incorrectly listed as zero. Statistics that originally made no sense, like teams obtaining 0 or 146 wins, were pretty much completely removed. Bizarre distributions for many of the variables were also improved through data cleaning. Of note are the pitching records at the bottom of the table below, which morph into near normal distributions.





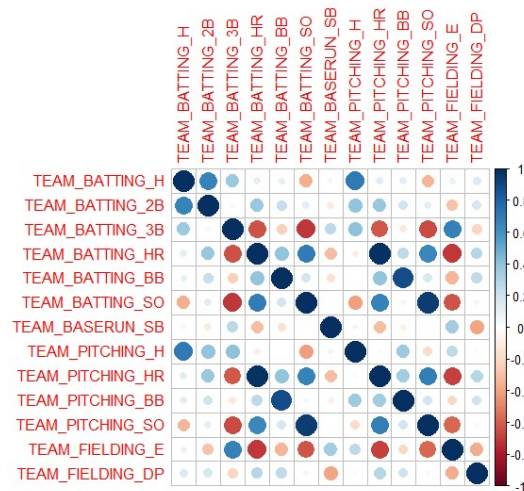




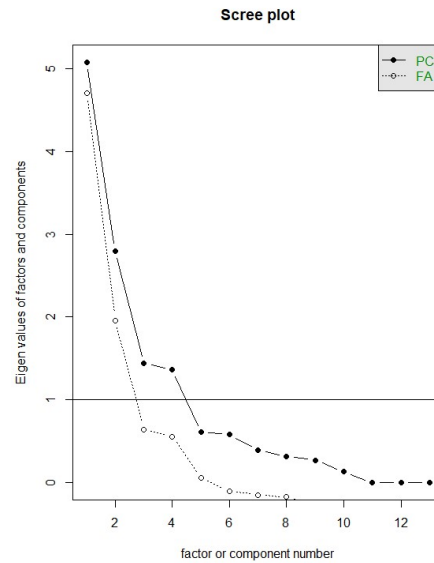


Analysis Methods

To begin, I created correlation matrix from the cleaned dataset. Factor analysis uses both the positive and negative correlations between the variables (shown in blue and red, respectively, in the matrix below) to identify latent factors.



Next, I needed to select the number of factors that I would retain by using the eigenvalues from the correlation matrix. Eigenvalues, in short, represent the number of original variables encapsulated by the factor. Because I am looking to reduce data dimensionality, I only choose to retain the four factors that have an eigenvalue above one, meaning they all explain one or more variables. The scree plot below shows eigenvalues for all thirteen potential factors. However, it is necessary to note that the eigenvalues from the correlation matrix are calculated slightly differently from the eigenvalues for the factor analysis on the scree plot. The eigenvalues for PC were the actual values considered for this study.



Finally, a varimax factor rotation was applied to maintain orthogonality. A loading cutoff of 0.5 or more was implemented to identify the factors that correlated significantly with each factor.

Final Model Presentation

Factor	1	2	3	4
% Variance	39.07	21.51	11.09	10.49
Loadings	Pitching Strikeouts: 0.980	Pitching Hits: 0.935	Pitching Walks: 0.963	Batting Homeruns: 0.617
	Batting Strikeouts: 0.963	Batting Hits: 0.748	Batting Walks: 0.901	Pitching Homeruns: 0.531
	Pitching Homeruns: 0.767			
	Batting Homeruns: 0.754			
	Batting Triples: -0.692			
	Fielding Errors: -0.630			

Factor 1: “Strikeouts+”

The first factor most strongly correlated with pitching and batting strikeouts. I called this factor “Strikeouts+” because in addition to strikeouts, it correlated positively to pitching homeruns and batting homeruns. On the other hand, the first factor correlated negatively to batting triples and fielding errors.

Factor 2: “Hits”

The second factor correlated positively with pitching hits and batting hits.

Factor 3: “Walks”

The third factor correlated positively with pitching walks and batting walks.

Factor 4: “Homeruns”

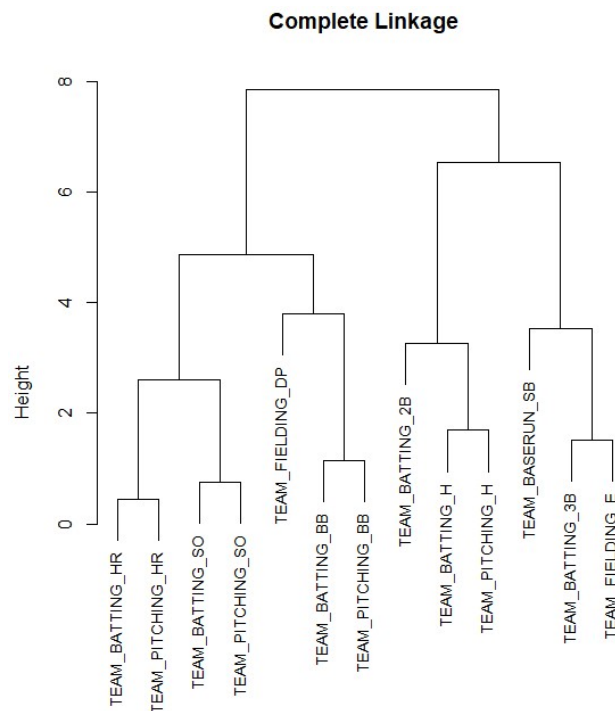
The fourth factor correlated positively with batting homeruns and pitching homeruns.

Complete Model Evaluation

While the factors can easily be named due to their correlation with opposing offensive and defensive statistics, they are difficult to interpret for this same reason. For example, Factor 2 is correlated positively with pitching and batting hits. This would mean that a team that gets lots of hits offensively is also allowing numerous hits from the mound. The same interpretation would be true for Factor 3, where hitting and pitching walks have high positive loadings. A team that receives many walks on the plate also allows many walks while pitching. Factor 4 describes a scenario where teams that hit high numbers of home runs are also susceptible to the long ball on defense. Even the more nuanced Factor 1 contains positive loading for pitching and hitting strikeouts along with batting and pitching homeruns.

To confirm that this was indeed the scenario described by the model, I took an unorthodox approach by constructing a dendrogram with the scaled distances from the

correlation matrix used earlier for Factor Analysis. This reverse dendrogram treats the variables as objects and uses their correlations to make clusters. Similar to the Factor Analysis results, offensive and defensive variations of the same variable cluster close together. Batting homeruns clusters tightly with homeruns allowed. Batting strikeouts clusters with pitching strikeouts. Batting walks pairs with pitching walks allowed, and batting hits clusters with pitching hits allowed. Though this is not a typical application of a dendrogram, it provides further confirmation that many of the offensive and defensive matching variables are correlated positively together.



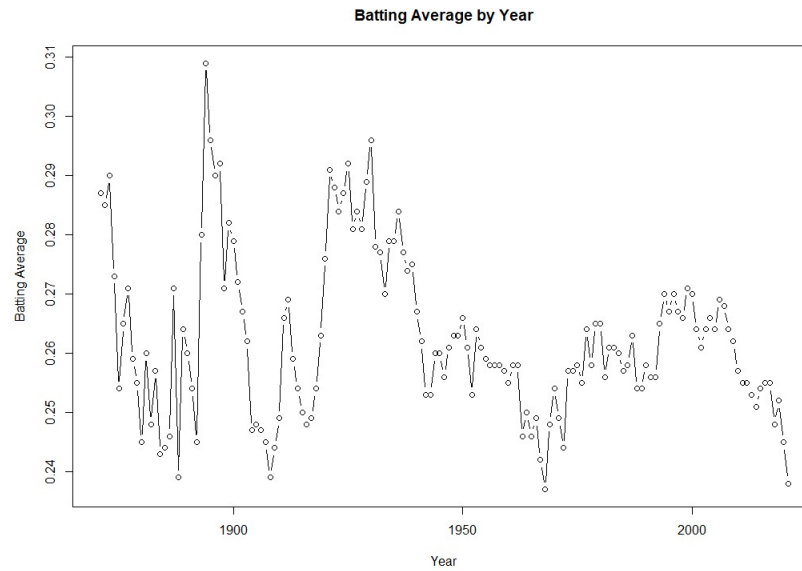
However, none of this evidence explains why these opposing records correlate positively with one another. The FA model describes a scenario where strong offensive teams tend to struggle on defense. The opposite would also be true as teams with a weak offense tend to have success on the mound. Yet, none of this makes intuitive sense because hitting and pitching are specialized roles. For the most part, position players do not pitch, and pitchers do not hit.

There are many intriguing theories that could explain this unexpected pairing of offensive and defensive records. Quite simply, it could reflect that every baseball team is going to have a weakness. Due to financial constraints perhaps teams that overspend for offensive talent end up with pitching liabilities and vice-versa.

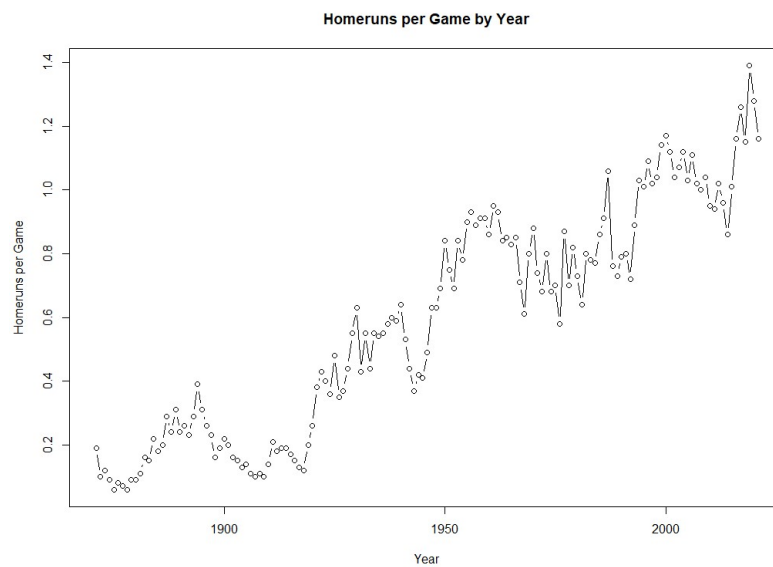
Another possibility is that baseball parks themselves are driving the correlations. Outfields in baseball are not uniform, and every field park has its own dimensions and quirks. This gives certain stadiums the reputation of being hitting friendly or pitching friendly. In turn, a team placed in a hitter's ballpark would have strong batting statistics but would also surrender high numbers of hits and homeruns while pitching. The opposite would be true in a pitcher's ballpark.

An additional scenario could be that the designated-hitter rule is behind this statistical pattern. Major League Baseball teams are split evenly between the National and American Leagues. The National League requires pitchers to bat just like every other position player. Yet, because they are notoriously poor hitters, the American League allows the pitcher's spot in the batting order to be covered by a designated hitter. This slight variation in rules creates a scenario where batting statistics are typically higher in the American League, making it a more difficult for pitchers. Perhaps this quirk in the rules is contributing the unexpected correlations discovered in Factor Analysis.

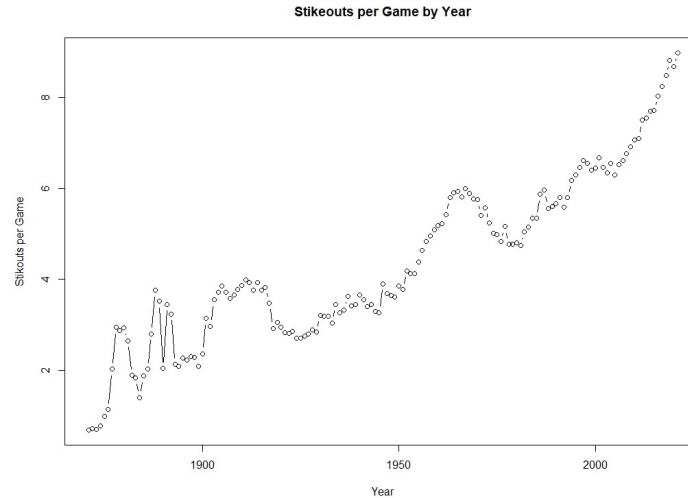
Personally, I hold a favorite theory that could also be working in conjunction with the other scenarios described above. Through its deep history, Major League Baseball has been game of eras. There are decades when pitchers have thrived at the expense of batters, and then there are decades where batters have beat up on pitchers. The plots below come from Baseball-Reference.com, and they show how several lead-wide statistics have evolved since 1871.



Batting average, which is just hits per at-bat, has fluctuated over the based century. Around 1900, teams were batting roughly 0.300, while this season they are struggling at 0.238.



Homeruns have become much more frequent in today's game.



Strikeouts, too, have steadily risen over the past century.

These figures show that baseball statistics vary greatly across decades and are perhaps what drives the composition of the factors. Teams that play in a strikeout era will rack up numerous strikeouts both on offense and defense. Teams that play in a homerun era will hit lots of long balls but will also surrender many homeruns while pitching. In this interpretation, factor analysis could probably best predict the era that a team played in, rather than estimate the number of wins it obtained.

Finally, this leads back to the supervised component of this study where I did examine the factors' ability to predict team wins. The table below shows the two linear regression models used to predict wins. The first was fitted on the 4 factors retained during FA. The second uses a kitchen-sink approach with all 13 variables that were not cut-out of this study.

Linear Regression Model	R-Squared	p-Value
4 Factors	0.1863	0.0000
13 Variables	0.4059	0.0000

As further evidence that the factors are not an effective predictor of wins, the factor model was only able to explain 19% of the variance in win totals. This looks especially weak compared to the variable model which obtain an r-squared value of 41%.

Conclusion

The Factor Analysis carried out on the Moneyball dataset identified four factors. They did not reflect expected latent traits like hitting, pitching, and fielding. Instead, the factors incorporated sister variables like hitting strikeouts and pitching strikeouts. Though there were many theories explaining why opposing offensive and defensive variables clustered together, the most likely was that team statistics mostly reflect their baseball era. For this reason, factors were not insightful when applied in a linear regression model to predict wins.

The next step to furthering this investigation would be obtaining the dates for the team records. This information would allow me to put my baseball era theory to scrutiny. Specifically, I could employ k-means clustering with the expectation that baseball teams from certain eras would form a cluster. A confirmatory result would show that strong and weak baseball teams of the same era are more alike than strong or weak teams across eras.