Dominance, Day in and Day out

An analysis of MLB team strength throughout baseball's many eras, focused on long-term winning rather than postseason accolades

By Ashling Scott

December 10, 2023

Introduction

Major League Baseball has always been plagued by an overemphasis on winning in the postseason. The perception of an entire season, long-fought over the grueling course of 162 games, can disappear in a blink at a simple loss in a 3 game series. That's all it takes for a winning season to turn from a source of pride to an embarrassing reminder that in baseball nothing is guaranteed.

This is particularly cruel in baseball, because baseball has ultra-high variance. In a given baseball season, the best team in the league by win-loss record will often win less than 65% of their games. A full third of the time, the best team will simply lose. Compare this to winrates of top teams in leagues like the NBA or the NFL, which can easily reach 80% or 90%, and you begin to see the problem. When the best team's win-rate is uncomfortably close to a coin flip, upsets are going to happen, and happen often.

But there's always another way to look at success. In this report, I will attempt to determine who were the strongest baseball teams throughout baseball's long history, and do it primarily using games from the regular season, without being swayed by championships or playoff berths. This question can be better answered through data, without the lens of human bias to distort the results.

We're going to look at 5 out of the 7 major eras of baseball history. We're looking at the Deadball Era, the Liveball Era, the Expansion Era, the Steroid Era, and the Contemporary Era. I left the Integration Era and the Free Agency Era out of my analysis because it was running long already.

Our dataset is mlb_elo, courtesy of 538.com. The idea behind the data was to construct an elo rating system, and apply it to teams in the MLB as an objective metric of gauging team strength and predicting the outcome of games. While the focus of the project was on making predictions for games yet to be played, they applied their methods to past seasons data as well, giving us ratings from over 200,000 MLB games spanning back to the 1800s. The dataset contains several ratings which relate to elo rating for a team, or for a player. Elo rating is a rating system borrowed from chess, which after every game recalculates a player's current rating based on the rating of their opponent and whether they won or lost. With these elo ratings, 1500 is the average rating, and will be our baseline for comparison

Computational Methods

The data required significant wrangling to get into a suitable place for my purposes. By default, the data-set is designed to analyze head-to-head match-ups between two teams. Each observation is a single game, and most of the variables have 2 versions, one corresponding to each team. This makes sense for making predictions on a match-up between two specific teams and pitchers, but for analyzing the variables to determine long-term past success, this data is untidy. We want each observation to focus on a single team, not multiple teams.

First I separated the values for team 1 and team 2 into separate dataframes, then re-named the variable names to match up with each other, before combining them together with bind_rows(). With the data-set pivoted to be longer, I next wanted to add a new variable. I wanted to find a model that uses the data present to predict whether or not a team will win a game. While no win/loss data was present, fortunately there was a variable for scores. Using transform(), I was able to create a new column in the data, won, recording whether the team won or lost the game. This will be used as our response variable.

To separate the data-set into the different eras of MLB history, I used the filter() command to filter by season. I also had to filter by team, after finding out the top teams from each era, to find games from specific teams for use in plotting and predicting. Finally, in order to draw correlations and build the models, i used na.rm() to build a "cleaned" version of the full reworked data-set, to get rid of some missing values from older games.

Now that we have a column for wins, we can check to see which of our ratings corresponds best to winning games. We have 4 variables to look at: elo_pre, elo_post, elo_prob, and pitcher_rgs. elo_pre is the team elo rating before the game starts, while elo_post is the updated team rating after the game ends. elo_prob is an estimated probability that the team will win the game based on their elo rating and the rating of their opponent. The last variable to look at is pitcher_rgs, which is a rolling game score. This is a measure of how successful that pitcher has been over their last 5 games.

Let's check the correlations between each of these variables and whether or not that team won the game.

```
cor(mlb_elo_cleaned$elo_pre, mlb_elo_cleaned$won)
```

```
## [1] 0.1152594
```

```
cor(mlb_elo_cleaned$elo_post, mlb_elo_cleaned$won)
```

```
## [1] 0.1773214
```

```
cor(mlb_elo_cleaned$elo_prob, mlb_elo_cleaned$won)
```

```
## [1] 0.1762107
```

```
cor(mlb_elo_cleaned$pitcher_rgs, mlb_elo_cleaned$won)
```

```
## [1] 0.08502205
```

These correlations may seem low, but as previously established, baseball is a high variance game. This is part of why baseball was one of the first major sports to fully embrace the sports analytics movement, stemming from a desire to shed light on the chaos. If anyone was able to predict game outcomes with correlation near 1, they would be making a fortune in a team's front office or for sports betting companies.

Predictably, our highest correlation is for elo_post, our team elo rating that gets updated after the result of the game. This makes it always correlate to wins better than elo_pre, which is the team's elo rating before the game begins. The correlation for elo_prob comes in at a very similar value to elo_post. This makes sense, because this value is simply calculated from the elo rating of one team compared to the team they are playing. Over the span of the entire dataset, the opposing team rating will wash out to simply be an average rating, leaving the observed teams elo rating to explain almost the entirety of the correlation. The correlations from pitcher_rgs lags behind, showing us that while the starting pitcher is important, the team rating is a more powerful predictor of team strength.

What about a combination of these variables? Let's take a look at some models that might best model the probabiliy to win games, and compare their R-Squared values. We'll start with our strongest correlation, elo_post and add variables from there. We won't be using elo_pre, as elo_post is the same metric but with a stronger relationship.

```
model1 <- lm(won~elo_post, data = mlb_elo_cleaned)
summary(model1)$r.squared
```

```
## [1] 0.03144288
```

```
model2 <- lm(won~elo_post + elo_prob, data = mlb_elo_cleaned)
summary(model2)$r.squared
```

```
## [1] 0.0375837
```

```
model3 <- lm(won~elo_post + elo_prob + pitcher_rgs, data = mlb_elo_cleaned)
summary(model3)$r.squared
```

```
## [1] 0.03813441
```

It seems like the more variables we add, the stronger the model becomes. Our model 3 has the highest R-Squared value, of 0.038. However, the jump between model2 and model3, where I added the weakest correlation, seems minor. We don't want to use a more complex model without assurance that it's predictive power makes it worth doing so, so we'll do an Analysis of Variance to test model2 and model3.

```
anova(model2, model3)
```

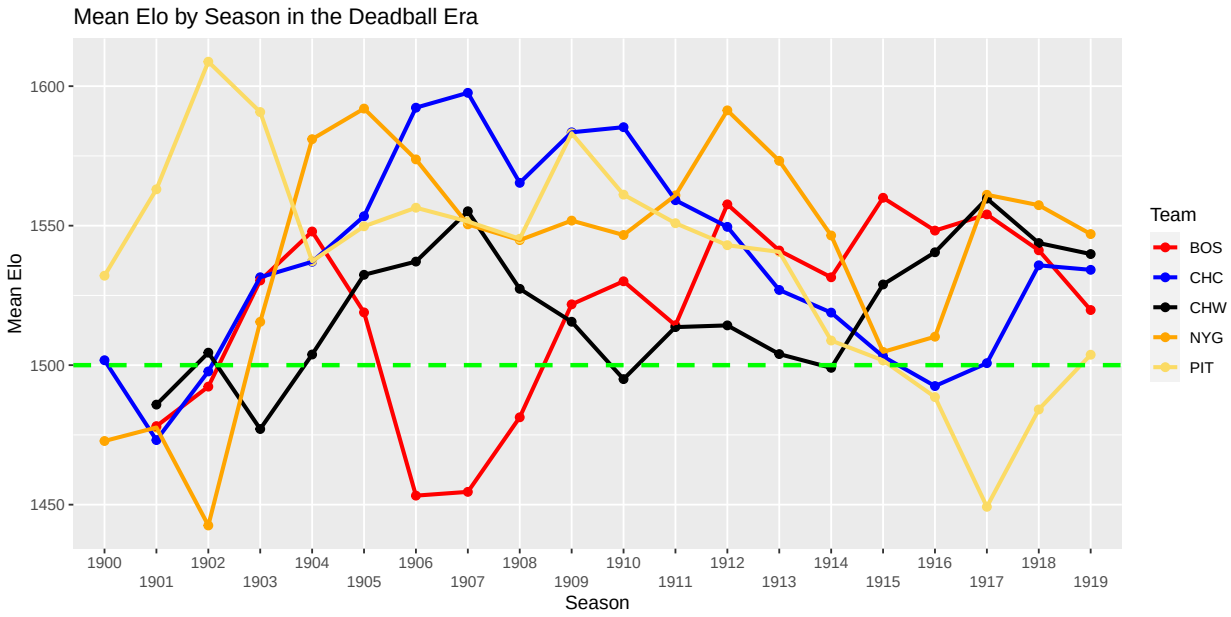
```
## Analysis of Variance Table
##
## Model 1: won ~ elo_post + elo_prob
## Model 2: won ~ elo_post + elo_prob + pitcher_rgs
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1 385146 92668
## 2 385145 92615  1    53.026 220.51 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p value of < 2.2e-16, we have confirmation that model3 is definitely worth using over model2, despite the extra complexity.

Now that we have our model, we can begin the analysis. Going forward, we will be using elo_post for data visualization purposes, and using model3 to make win-rate predictions to help us decide who our most dominant teams are for each era.

Data Analysis and Results

Deadball Era



To start, I'll explain the above plot, which will serve as visualization for team strength over the decade. I have constructed this plot by taking all the data from the deadball era and finding the teams with the highest average elo_post ratings. The average elo_post rating for each of these teams in each of the seasons of the era is displayed here, to visualize the trend in rating over time.

The dotted green line is the average league elo, which is always set at 1500. Sometimes these plots can tell the whole story of an era.

Now its time to use our model to generate win-rate predictions. For this era, some of our data is incomplete. pitcher_rgs, which keeps track of how well a starting pitcher has performed over the last five games, is not present until the 1913 season. For this particular era, we will have to use a less sophisticated model, that does not use pitcher_rgs.

```
predict(model2, data.frame(elo_post = mean(NYG$elo_post), elo_prob = mean(NYG$elo_prob)))
```

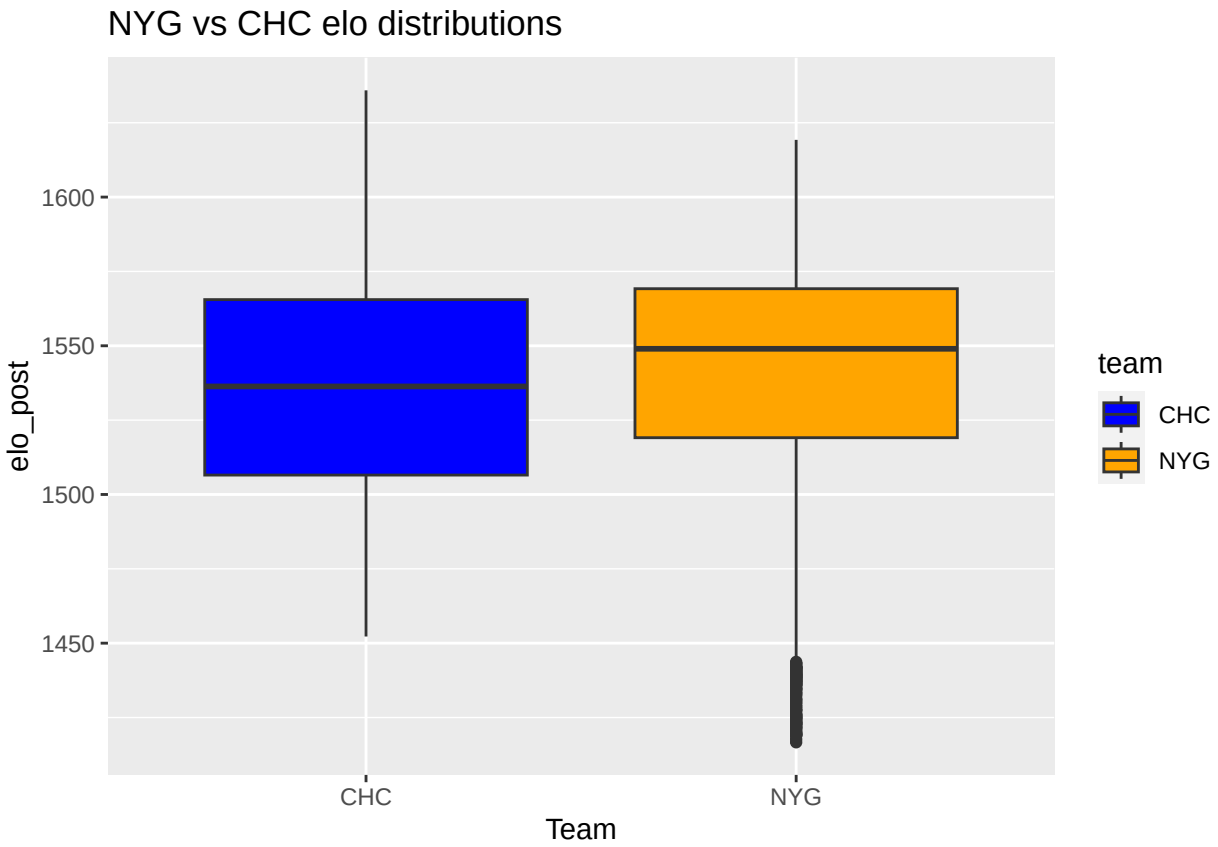
```
##         1
## 0.5885824
```

The resulting value is a predicted winrate for the New York Giants (Now known as the San Francisco Giants) over the course of the Deadball Era. Our predicted win-rate is 0.588. We can compare that to the actual historical winrate for the Giants over that time period, which is 58.0%, and we see that this prediction is quite reasonable, within 1% of the true winrate. Lets see how it compares to the prediction for the other four teams:

Team	Prediction
New York Giants	0.5885824
Chicago Cubs	0.5812465
Pittsburgh Pirates	0.5807293
Boston Red Sox	0.5454622
Chicago White Sox	0.5464404

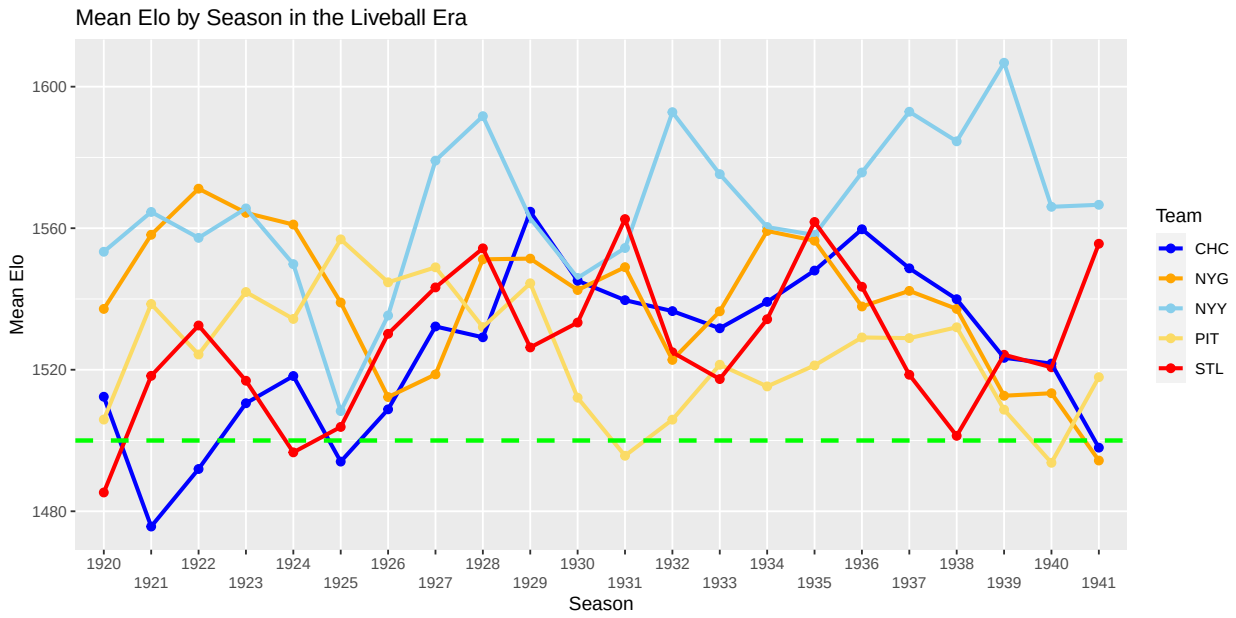
The deadball era is often considered the beginning of the modern era of baseball. Its named after the suppression of offensive power that was the hallmark of the era. Baseballs were not cleaned or replaced as often midgame, and pitchers got away with using now-banned pitches like the Spitball, which marred the ball to influence aerodynamics. We have several teams here which we will be seeing a lot of in the future, with San Francisco and Boston, and this is also the only era where the New York Yankees are not present.

Our model likes the New York Giants the best, but it's by less than a percentage point. In order to server as a tie breaker, I took a look at a box plot of the teams' distributions of elo_post over the deadball era.



The results paint the picture of consistency. The median and both quartiles of the NYG plot are clarly above the corresponding values for the Chicago Cubs. The only blemishes on NYG are the significantly lower tail at the bottom, and the slightly lower peak. Given that this project is aimed at identifying consistency over the long haul, and that our prediction already favored the New York Giants, I will declare NYG the most dominant team of the era.

Liveball Era



Now that we have pitcher_rgs data, we can go back to using our full model. This is our prediction for NYY:

```
predict(model3, data.frame(elo_post = mean(NYY$elo_post), elo_prob = mean(NYY$elo_prob),
                           pitcher_rgs = mean(NYY$pitcher_rgs)))
```

```
##         1
## 0.6524249
```

The full predictions are:

Team	Prediction
New York Yankees	0.6524249
New York Giants	0.5874363
St. Loius Cardinals	0.5575061
Chicago Cubs	0.5551590
Pittsburgh Pirates	0.5509214

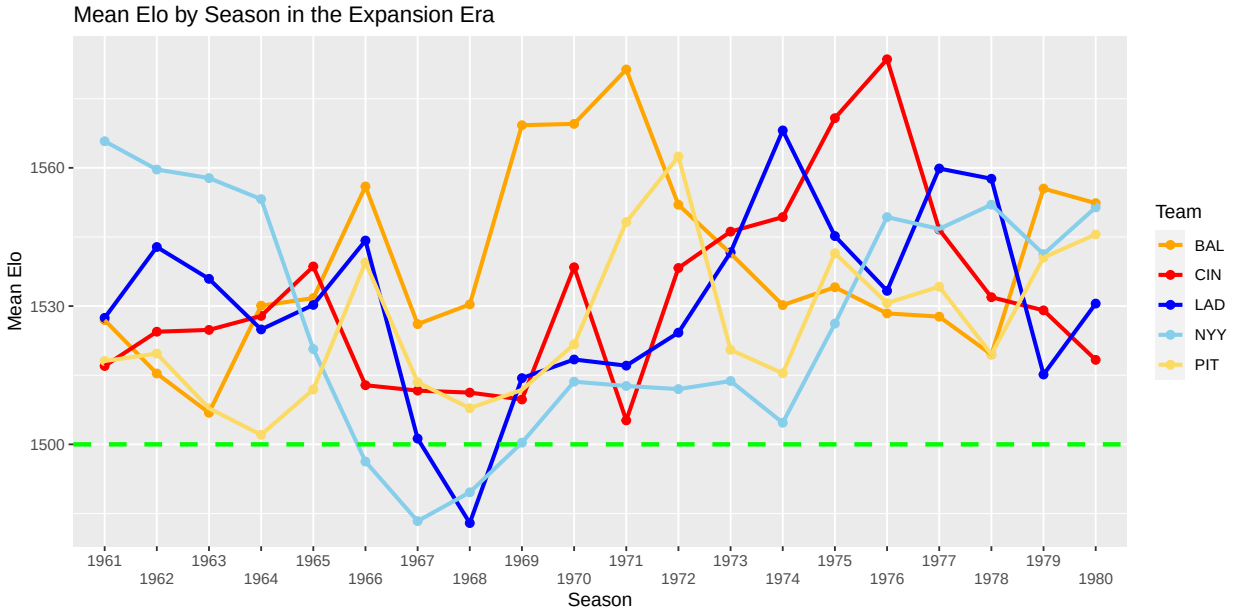
The Liveball era began with rules changes to keep pitchers from tampering with balls. It brought in an explosion of offense, and an explosion of popularity to go with it. As you can see by the plot, this story of this era is the rise of the New York Yankees. Lead by legendary player Babe Ruth, the Liveball Yankees took the game to new heights and new crowds.

According to our data-set, the most lopsided matchup of all time occured in 1939. The match was between the Yankees and the Baltimore Orioles, two games played on the same day, September 17th. The elo_prob win probability prediction for the Yankees that game was 80%, the highest

it would ever reach, followed in the night game by the 2nd highest recorded value, 79%. The Yankees lost both games. That's baseball.

The winner for this era is clear. With what may be the most dominant era for a single team ever, the New York Yankees get the nod.

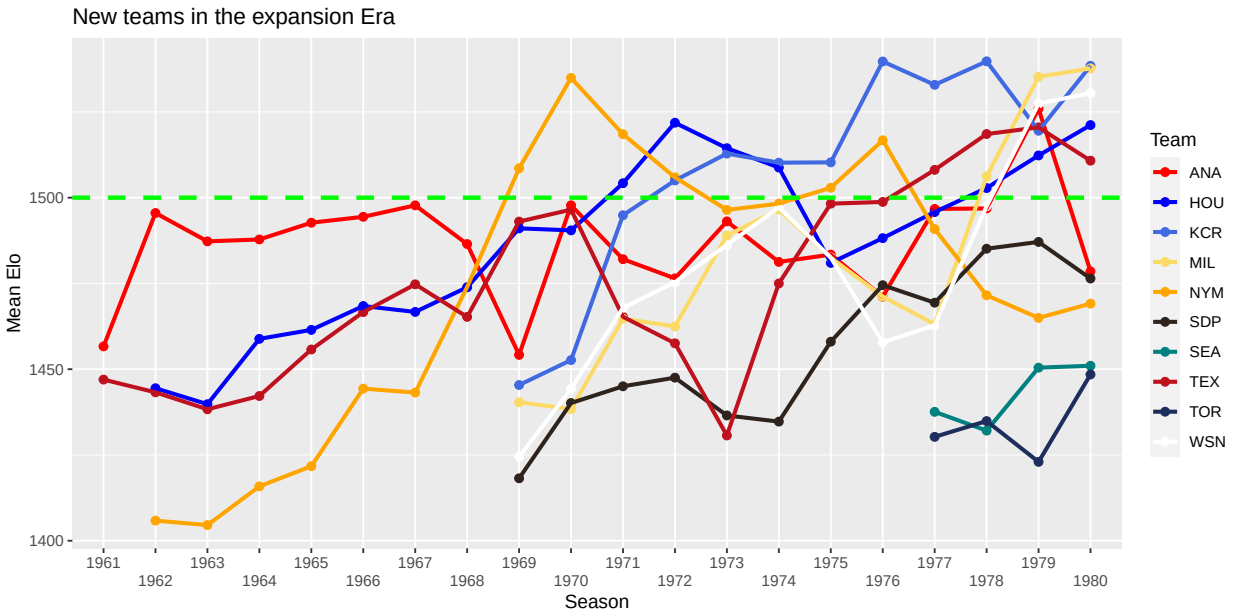
Expansion Era



Team	Prediction
Baltimore Orioles	0.5883944
Cincinatti Reds	0.5654821
Los Angeles Dodgers	0.5689387
New York Yankees	0.5600208
Pittsburgh Pirates	0.5520309

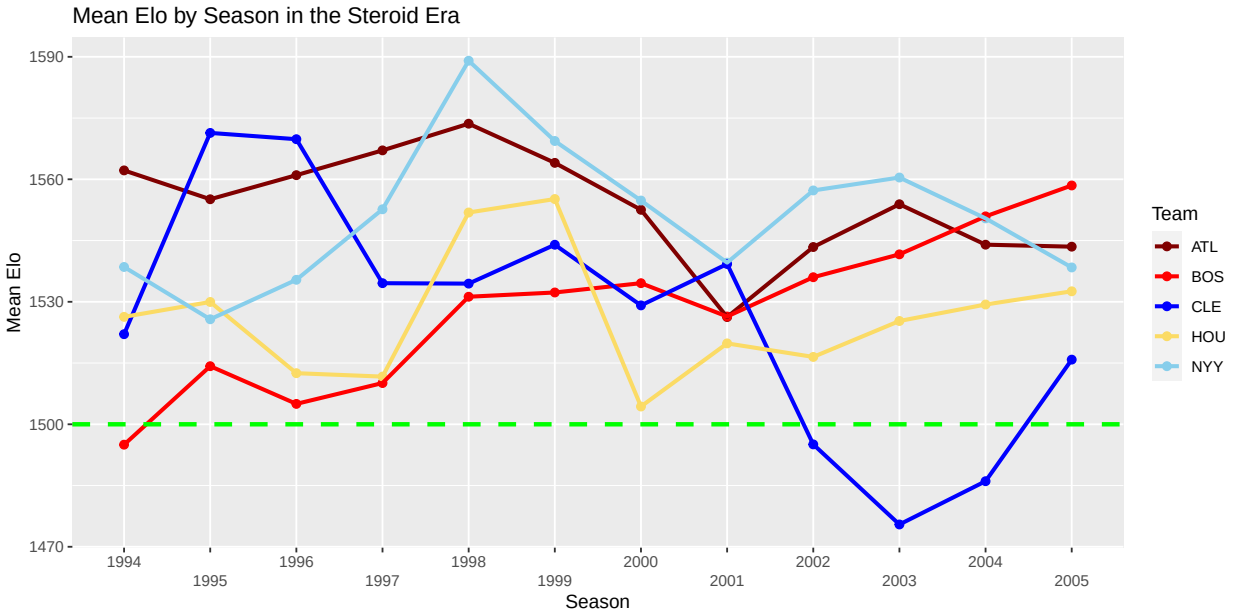
The expansion era was a time of unmatched parity. New contenders rose to prominence, such as the Baltimore Orioles and the Cincinatti Reds. This era is an excellent time-period to apply our model to, since determining the best team through championship count alone is uninformative. Several teams managed to get multiple titles in this time, including the Yankees, Dodgers, Cardinals, Orioles, Reds, Pirates, and Athletics. Fortunately for us, we don't have to agonize over these counts, since the model gives a sizable win-rate predictive advantage to the Orioles, and we will be giving them the nod for this era.

The expansion Era was named after the explosion of teams that were added to the league. A total of 10 teams were added to MLB during this 20 year span, changing the shape of the league forever. Success can be difficult for a new team, who are forced to get their players from other teams via an expansion draft. Lets take a look at how thew new teams of this era performed in their first seasons.



While most of the new expansion teams languished below the mean elo as we would expect, some teams found quick success. Specifically, the Kansas City Royals (KCR) had a fast and meteoric rise into prominence, stringing together season after season of high mean elo after only 3 seasons getting their footing. This incredible surge from an expansion team would only be matched by the Miama Marlins in the late 90s.

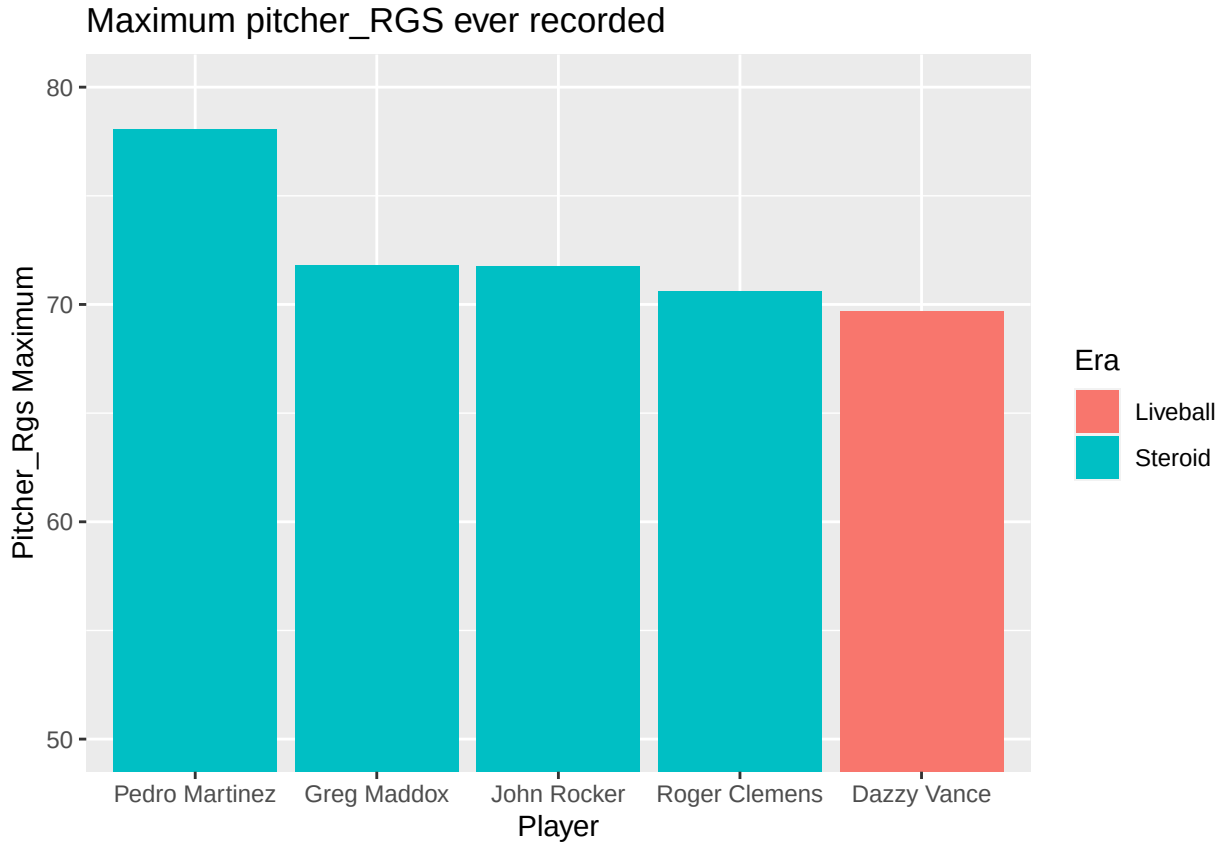
Steroid Era



Team	Prediction
Atlanta Braves	0.6217202
New York Yankees	0.6173569
Boston Red Sox	0.5647631
Cleveland Indians	0.5541991
Houston Astros	0.5543915

The Steroid era is the most infamous time in baseball's long history. It was defined by steroids and other performance enhancing drugs, which many players were taking to drive their game to the next level. Hitters like Mark McGwire, Sammy Sosa, and Barry Bonds set new home run records and put up gaudy offensive numbers year in and year out, lasting for around a decade before the house of cards all came crumbling down. The league admonished many of its biggest stars for cheating, and added rules to crack down on performance enhancing drugs in the future. Despite all that, it remains a golden age of baseball in the memory of many.

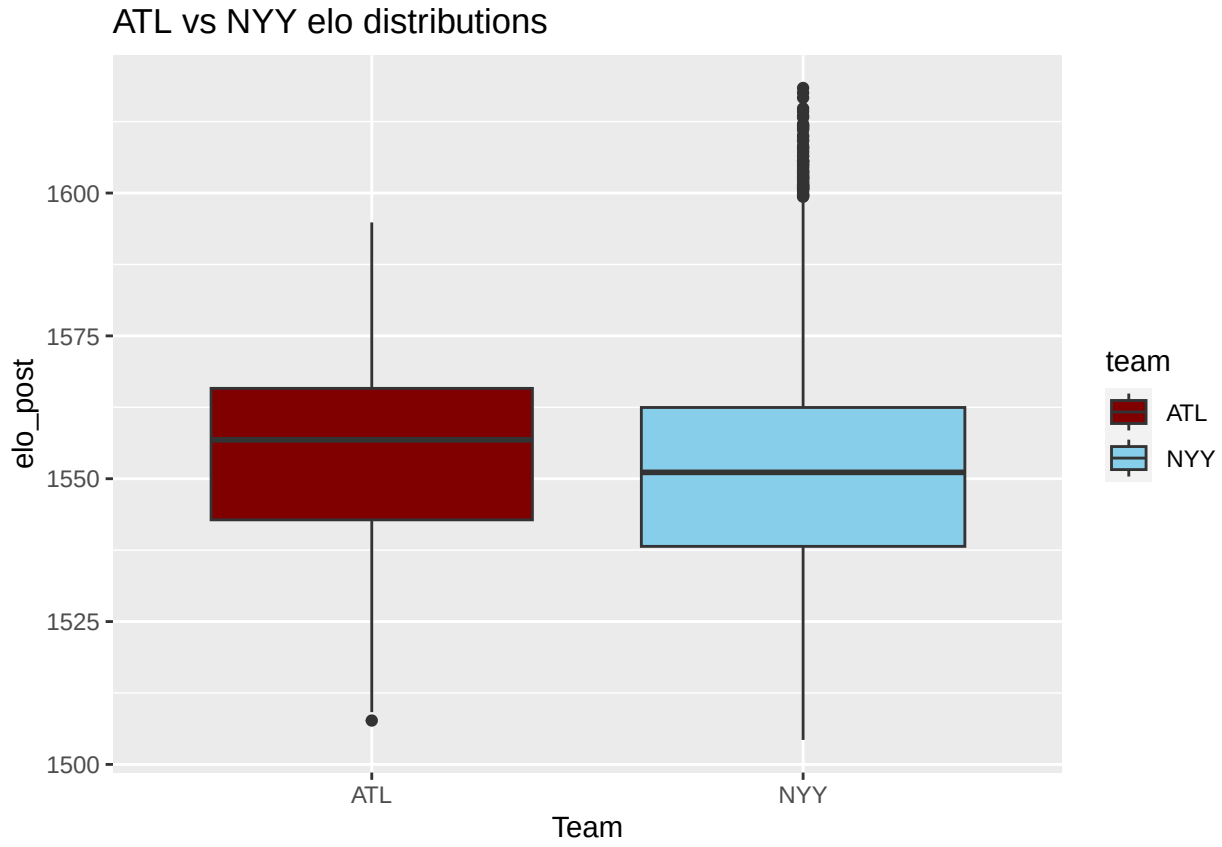
One interesting note about the Steroid era, is that even thought he perception lingers that it was an offense favored period, pitchers were also using performance enhancing drugs during this time. It's been argued that the era was actually favored toward pitching, rather than hitting. We can see that reflected in our pitcher_rgs rating. Let's take a look at the highest pitcher_rgs scores ever recorded by a unique player.



As we can see, four out of the five highest pitcher_rgs scores ever recorded came from the steroid

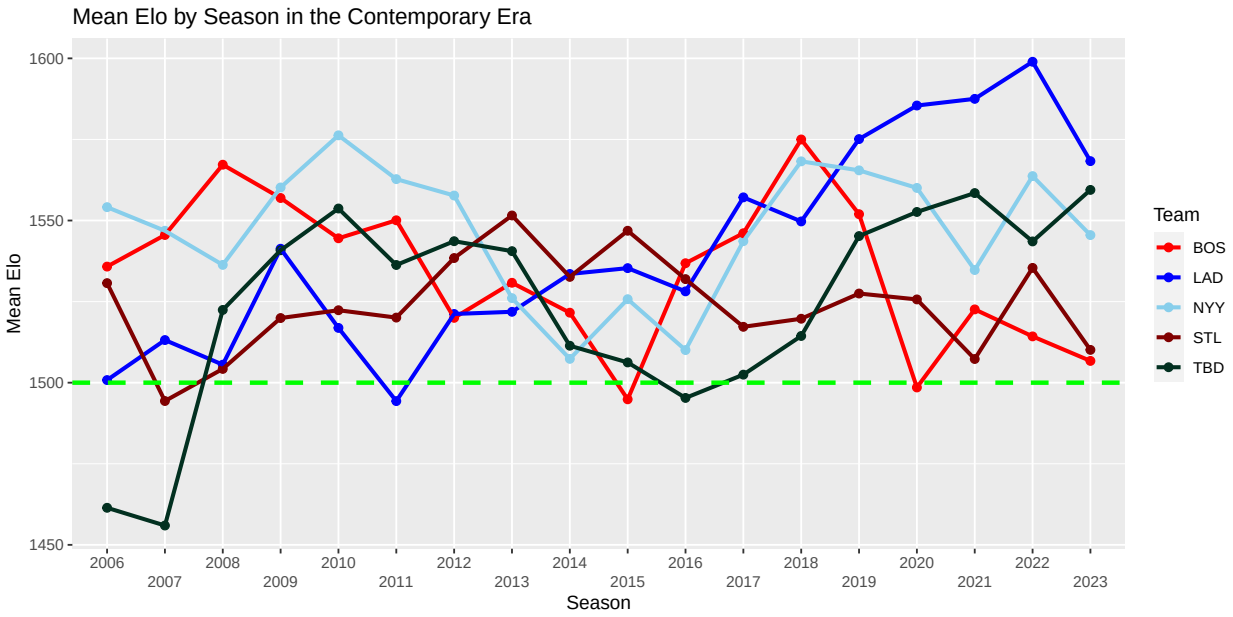
era, with the next highest coming in all the way in the liveball era. This shows that even in an era of unmatched offensive power, pitchers were still able to put together 5 game streaks that can be considered the greatest individual pitching spans in MLB history.

This era ends up being the closest gap in predicted winrate between the top two teams, the Atlanta Braves and the New York Yankees, with a gap of only 0.004. Since we have a close race again, lets again take a look at a bot plot to see the overall shape of the distribution of their elo means.



The Yankees have a higher peak, and a lower trough, while the Braves more a more consistent performance overall, including a higher median and higher quartiles. Given this consistency and the higher prediction value of our model, I'll declare the most dominant team of this era the Atlanta Braves, who crushed the the early half of the era with their historically strong pitching.

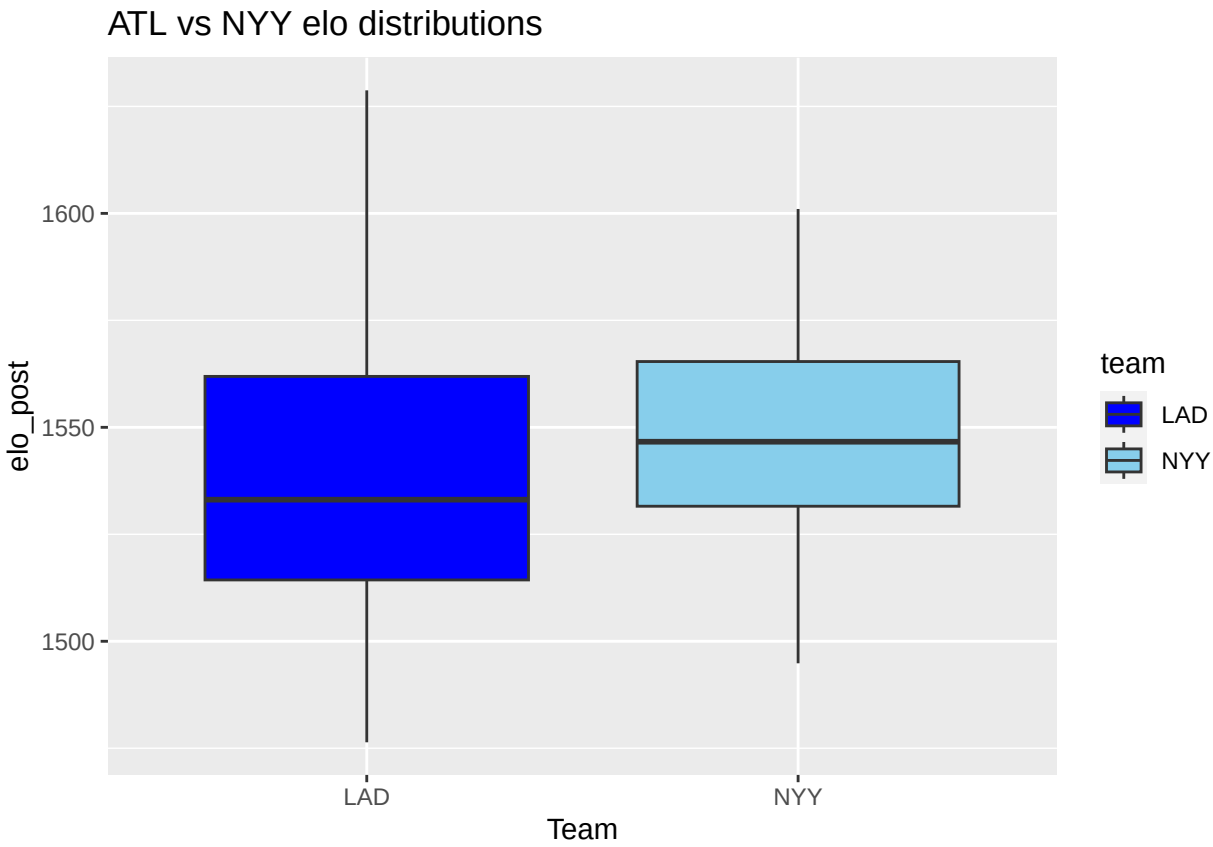
Contemporary Era



Team	Prediction
New York Yankees	0.5969500
Los Angeles Dodgers	0.5871034
Boston Red Sox	0.5713158
St. Louis Cardinals	0.5522986
Tampa Bay Rays	0.5371991

Finally we come to the modern era. In the contemporary era, steroids and other performance enhancing drugs have been cracked down on, reducing some of the biggest outliers in terms of player production. It has been another era of parity, with the top winrate predictions falling below 60% once again. We see a lot of familiar faces here, with one newcomer, the Tampa Bay Rays. Established during the Steroid Era, the Rays have become one of the most successful expansion stories of all time, having only 3 seasons in the entire era with a below-average mean elo.

Since our top team winrates are the same, lets take one last look at a bot plot for the New York Yankees and the LA Dodgers.



As is the case with the other boxplot tie-breakers, one team has higher median and quartiles, as well as a more condensed overall spread of elo. The graph once again confirms what the predictions suggested, and we'll give the title of dominance to the New York Yankees.

Final Results

At the end of our analysis, the strongest teams from the eras we looked at are:

Deadball: New York Giants (San Francisco Giants) Liveball: New York Yankees Expansion: Baltimore Orioles Steroid: Atlanta Braves Contemporary: New York Yankees

Conclusion

In conclusion, I believe my model did a good job of estimating consistent team strength, era to era. It's been interesting to approach this question from an entirely different angle. Instead of relying on post-season success, we are simply using game data from every game, as if all are equal. This gives us some interesting answers, such as the Baltimore Orioles from the Expansion Era. Most people, if asked who the top team from this era was, might have overlooked the Orioles, despite them competing year in and year out and having several world championships in this span. This may be due to bias, as the Baltimore Orioles are not a major market team, nor are they a historically important or successful team. It's easy to overlook their success as a blip and give the award for

best team of the era to someone flashier, but the data keeps us grounded. The elo rating and win-rate predictions have vindicated the Orioles as the team most likely to win games in this era.

Equally important, our model and graphs were also able to confirm the most obvious of results, such as the dominance of the Yankees from the Liveball era. Had our model decided on another team being superior during this time period, it would have been a massive red flag that something was wrong. Instead, the model and the Yankees were vindicated with the largest predicted win-rate gap of all the eras. Overall, I am quite happy with how well the results came out.

The scope of my project is massive, nearly the entirety of MLB history. Technically, the start of baseball history was the mid to late 1800s, but while we do have some data from that era, it's spotty and wildly variant. The rules of the game at the time had yet to be solidified, and so teams from many leagues were playing different versions of the game, with wildly different outcomes. Not to mention, many of the teams from that era were dissolved and did not go on to become teams in the modern era. Overall, choosing 1900 as the starting point for our analysis was a solid decision, and I believe represents a good cutoff for the scope of the project. As for generalizability, the elo stat is designed to be compared across eras, and I think it does a good job of doing so. Elos from wildly different eras can be compared within a similar range, without one era getting some kind of bias over another.

One limitation of this data-set was the lack of quantitative performance statistics, such as a pitcher's Earned Run Average, or hitters Batting Average, or the typical stats that baseball players are judged upon. Had I wanted to incorporate some of these metrics into our model, or used for visualization purposes, I would have had to access another supplementary data-set. It was interesting tackling this problem from within the confines of just this data-set, but more could have been done with additional data.

A major area of improvement that I could think of would be expanding upon the concept of pitcher_rgs and adding its methodology to other metrics. I think that pitcher_rgs is the most interesting variable in the data-set, given that it fluctuates depending on a span of recent performance. While all the other ratings are more stable overall ratings, pitcher_rgs is a dynamic metric that measures momentum, how well a pitcher has done in a recent period. I think that had there been a rolling game score for team elo, it would have been interesting to see which teams had the most dominant 5 game stretches in history. Assembling a team_rgs rating would probably have required an outside data-set of performance stats, as it would be difficult to build it from the ratings present in this data-set alone.