Creating a Frontier League Player Projection System

Lance Brady

Due Date: 2024-12-16

Abstract

This project introduces a comprehensive analytic framework for projecting the professional trajectories of Frontier League baseball players by integrating advanced performance metrics, predictive modeling techniques, and comparative analysis with affiliated Minor League Baseball (MiLB) data. Leveraging high-resolution data acquired via Yakkertech—an optical tracking system that captures pitch characteristics, batted-ball metrics, and player performance—this study computes expected statistics (xBA, xSLG, xwOBA) and constructs wOBA-based run value models adapted for an independent league context. These results are then aligned with analogous metrics from the Florida State League, a Single-A league equipped with Statcast technology, enabling direct calibration of Frontier League performance against the affiliated development pipeline. Multiple multinomial logistic regression models are employed to predict the likelihood of a Frontier League player's advancement to higher MiLB levels or even Major League Baseball, incorporating traditional, expected, and hit-quality metrics. The models highlight that while conventional statistics currently best predict promotions within affiliated systems, advanced analytics and hit-quality measures (e.g., exit velocity, launch angle) offer robust supplementary insights. Additionally, the study finds that certain Frontier League standouts possess statistical profiles commensurate with their affiliated Single-A counterparts, underscoring a potentially underutilized talent pool for MLB organizations. This work underscores the feasibility and value of integrating independent league data into established talent evaluation pipelines. By providing a data-driven roadmap for comparing and projecting player outcomes across unaffiliated and affiliated baseball ecosystems, the study lays foundational groundwork for more informed scouting, player development strategies, and future research leveraging multi-source performance data.

All work is contained on the GitHub Repository on the following public website: https://github.com/lancebrady15/425-capstone-project

Introduction

Independent Baseball

Independent baseball occupies a unique space within the broader ecosystem of professional baseball. Unlike teams affiliated with Major League Baseball (MLB) organizations, independent leagues operate outside the traditional farm system structure. This arrangement provides players who might have slipped through the cracks—or are looking to resurrect their careers—an opportunity to showcase their abilities and potentially gain the attention of MLB scouts. The quality of play in these leagues has improved markedly in recent years, with some organizations rivaling the talent found in the lower levels of affiliated systems. In 2020, Major League drastically cut down on the MiLB farm system, making independent baseball an even more important part of the baseball ecosystem, and increasing their talent pool.

Frontier League

The Frontier League was established in 1993 as an independent professional baseball league, initially serving smaller markets without access to affiliated minor league teams. Over the past three decades, the league

has grown in both size and stature, now consisting of 16 teams across the United States and Canada. The league has a long history of sending players to MLB-affiliated teams, highlighting its role as a developmental pipeline for overlooked or underutilized talent. The recent partnership with Major League Baseball, granting it the status of an MLB Partner League, underscores its growing legitimacy and importance.

Relations with Affiliated Baseball

In 2020, the Frontier League became one of MLB's Partner Leagues, signifying closer ties between independent baseball and the affiliated minor league system. This designation allows the Frontier League to share resources and data with MLB while maintaining its autonomy. For players, this relationship enhances visibility and opportunities to advance into affiliated baseball. The league has adopted various MLB-inspired rules and technologies, further integrating itself into the broader baseball landscape. This collaboration represents a mutual benefit: MLB gains a developmental resource for evaluating players outside its traditional system, while the Frontier League solidifies its reputation as a legitimate stepping stone to higher levels of professional baseball.

Yakkertech and BaseballCloud

A major development in the Frontier League's evolution has been its adoption of advanced player performance technologies like Yakkertech. Yakkertech is a cutting-edge data collection system capable of tracking granular player metrics such as pitch velocity, spin rate, exit velocity, and launch angle. These statistics provide teams with valuable insights into player performance and potential, allowing for more informed decision-making in player development and scouting. The collected data is stored and analyzed using BaseballCloud, a software platform that integrates and visualizes performance metrics for players, coaches, and analysts. This technology has revolutionized the evaluation process in independent baseball, offering Frontier League teams the same level of analytical depth previously reserved for MLB-affiliated organizations. Yakkertech data from every game was available to teams starting in 2023.

Statcast Minor League Baseball

Statcast, Major League Baseball's premier data-tracking technology, has become a cornerstone of modern baseball analytics. Originally introduced at the MLB level in 2015, Statcast has gradually expanded its reach into affiliated minor league baseball, providing a wealth of data on player performance across all levels of the minor league system. Since 2021, Statcast has been implemented in the Florida State League (FSL), a Single-A league comparable in talent level to the Frontier League. This expansion has made advanced metrics, such as pitch velocity, spin rate, exit velocity, and defensive positioning, readily available for analysis at lower levels of professional baseball.

For this study, Statcast data from the Florida State League serves as a critical benchmark for evaluating Frontier League players. By leveraging this dataset, I can identify similarities and differences in player performance metrics across the two leagues. These insights enable the development of a multinomial logistic regression model that predicts the likelihood of Frontier League players advancing to higher levels of the affiliated minor league system or even Major League Baseball. The availability of Statcast data at the minor league level not only enhances the scope of this analysis but also highlights the growing importance of advanced analytics in evaluating and developing baseball talent at all levels.

This Study

This study explores the potential trajectories of Frontier League players by leveraging advanced player performance data collected through Yakkertech and publicly available Statcast data from the Florida State League. The primary objective is to determine the likelihood of Frontier League players reaching higher levels of affiliated professional baseball, such as Single-A, Double-A, Triple-A, or even Major League Baseball.

Using multiple multinomial logistic regression models, this analysis integrates performance metrics such as exit velocity, launch angle, among others. These metrics are compared between Frontier League players and their counterparts in the Florida State League, a Single-A league with a comparable talent level. By aligning

the two datasets, this study bridges the gap between independent and affiliated baseball, providing insights into how performance in the Frontier League translates to affiliated baseball.

I use another multinomial logistic regression model to compare Frontier League players in 2023 and Frontier League players in 2024 to predict where 2024's players will end up next year.

Data Exploration and Visualization with 2023 Data

Section 1: Data Cleaning

To begin, I downloaded the CSV files from the OnePoint Shared Folder, which contained pitch-by-pitch data for the 2023 and 2024 Frontier League seasons. I then wrote a function `process_season` to process the season's worth of data, which I had downloaded locally into folders titled "2023" and "2024". `process_season` was able to connect to MySQL and write the data row by row into a SQL table while recognizing the headers. R is able to handle the large amounts of data I have (over 300,000 rows per season of over 100 columns), but it took a while to load, so I thought that working with SQL to select columns from the database would work best. Due to the fact that I are taking at-bats as independent events, the goal was to take the game-by-game pitch data and turn it into season-by-season pitch data. Thus, I do not care when a pitch occurred or who threw the pitch. I just care about each batter's outcomes. Thus, I have two SQL Tables, one for Frontier League 2023 pitches, and one for 2024 pitches.

Note about notation: - ExitSpeed = Exit Velocity of a batted ball - Angle = Launch Angle of a batted ball - xBA = Expected Batting Average of a batted ball - xSLG = Expected Slugging Percentage of a batted ball - xWOBA = Expected Weighted On-Base Average of a batted ball - xOBP = Expected On-Base Percentage of a batted ball - xOPS = Expected On-Base Plus Slugging Percentage of a batted ball - K% = Strikeout Percentage of a batter - BB% = Walk Percentage of a batter

Section 2: Is Direction a Good Predictor of Hits?

The first thing I did was query to only have pitches that resulted in batted balls, but only batted balls that would be considered for expected statistics. So, fouls were not included.

Now that I had all batted balls, I wanted to create models to predict both whether a particular batted ball will be a hit and what type of hit for xBA and xSLG. While some research includes Direction (otherwise known as spray angle) as a feature, I will first use Leave-one-out cross validation to see if it is a good predictor of getting a single, double, triple, or home run in our model, a logistic regression model with a binary outcome (single or not, double or not, etc.)

I performed five-fold cross validation to see if log loss was significant decreased by adding Direction to the model.

model_set <chr>	outcome <chr>	log_loss <dbl>
NoDirection	is_double	0.24362930
WithDirection	is_double	0.24326272
NoDirection	is_home_run	0.11617948
WithDirection	is_home_run	0.11618128
NoDirection	is_single	0.51681230
WithDirection	is_single	0.51688101
NoDirection	is_triple	0.03909420
WithDirection	is_triple	0.03839405

Figure 1: Log Loss Direction Table

Using this model, there was virtually no difference in log loss between models with and without Direction. This suggests that Direction was not a good predictor of hits. I will proceed with the models without Direction.

Section 3: Checking our Model

I then realized that because each logistic regression is operating separately, this could lead to a batted ball with probabilities of being a single, double, triple, and home run that sum to more than 1. I will fit a new multinomial logistic regression model to predict the type of hit for each batted ball. Our model converged as seen in the image.

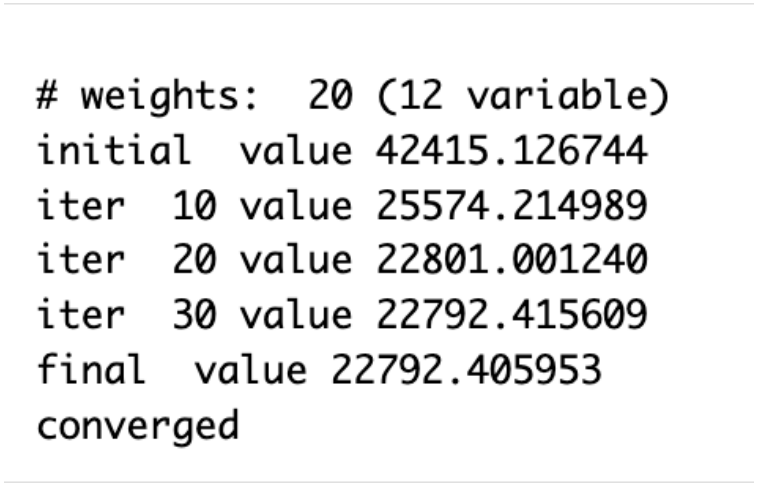

Figure 2: Multinomial Model Convergence

Then, I predicted the probability that a batted ball would be a single, double, triple, or home run, and merged those batted balls back into our original dataset.

I created graphs to check our model, of ExitSpeed versus xBA and xSLG, and Angle versus xBA and xSLG. I found that my model was a good fit for the data, as these graphs make sense.

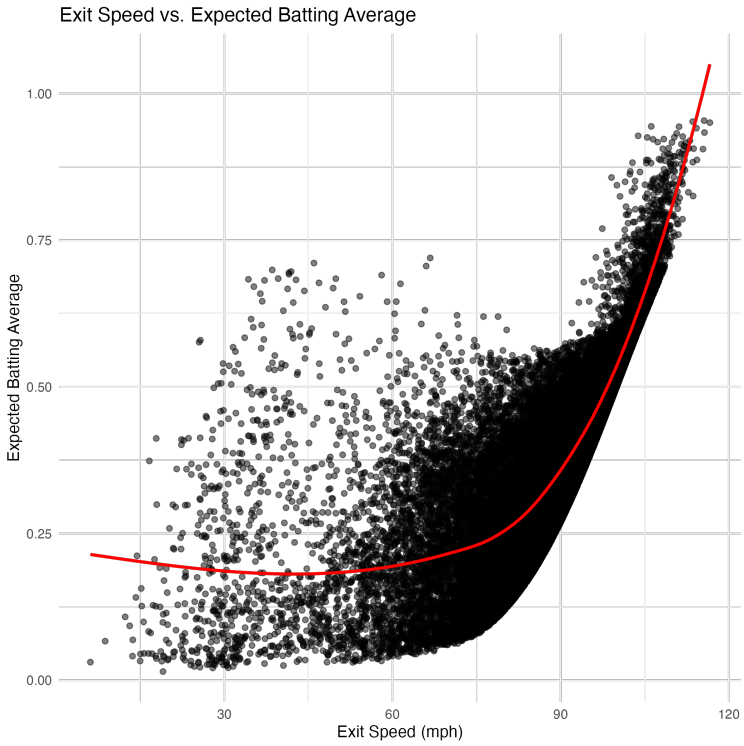

Figure 3: 2023 Expected Batting Average vs. Exit Velocity

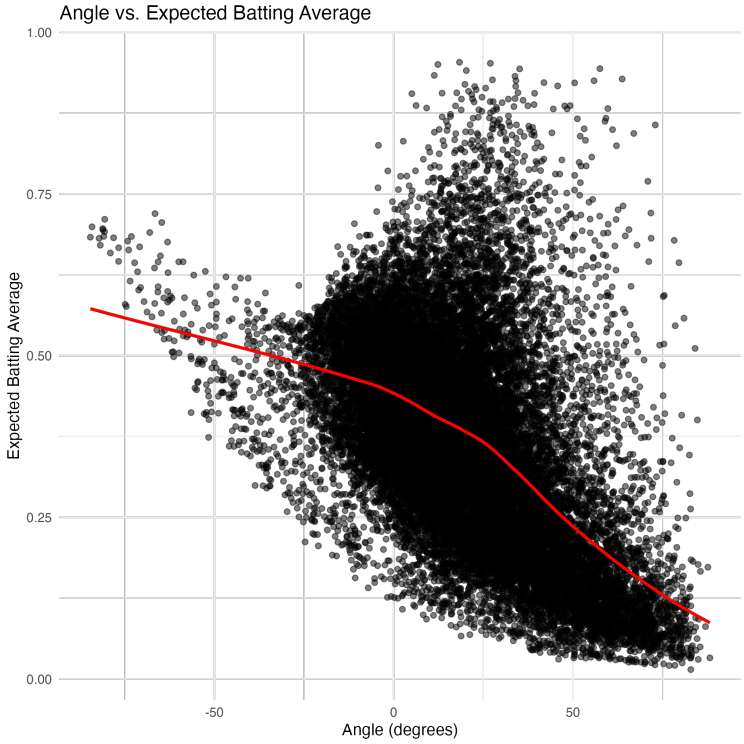

Figure 4: 2023 Expected Batting Average vs. Launch Angle

In general, batting average increases as Exit Speed increases, with lots of variation below 90 MPH, which is when the ball is hit hard enough for it to matter; this is a definition that MLB Statcast uses, as OPS is significantly higher for hard hit balls than non-hard hit balls. However, because I are talking about all kinds of hits, there is significant variation below 90mph. A slow dribbler along the third base line could have a high probability for a hit, and so could a 110 MPH exit velocity rocket with an optimal launch angle. As for Angle, the expected batting averages are most variable between 0 and 50 degrees, which makes sense as those will generate the most diverse outcomes (usually to the outfield). All other angles have a more consistent expected batting average because on the low end of Angle, they will likely just be ground balls (single or not), and on the high end, they are almost all pop ups.

Let's graph ExitSpeed and Angle vs. xSLG, because this will be an even better check on if our model is working. xSLG will be a better angle not just if our model is predicting hits, but if it is predicting offesnive output weighted by the number of bases each hit is worth.

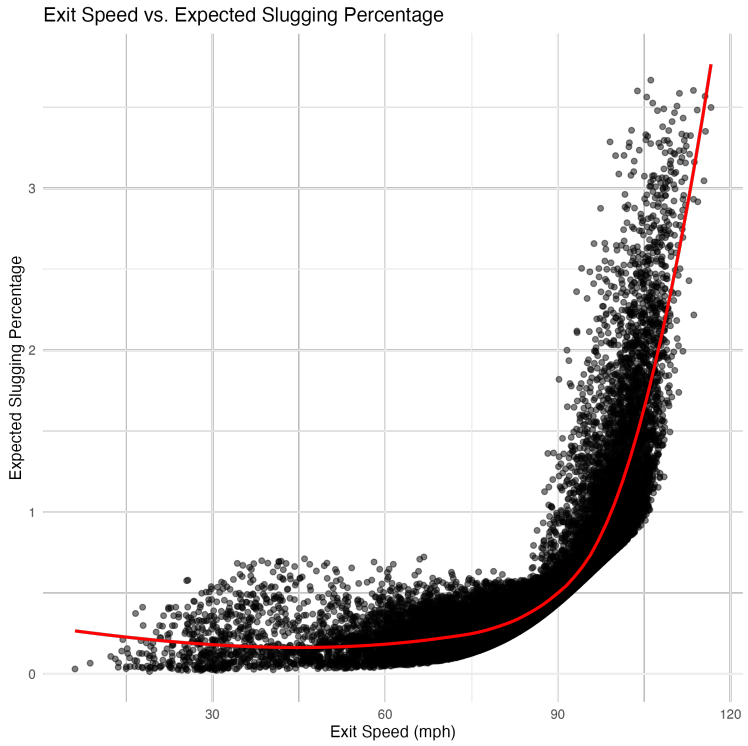

Figure 5: 2023 Expected Slugging Percentage vs. Exit Velocity

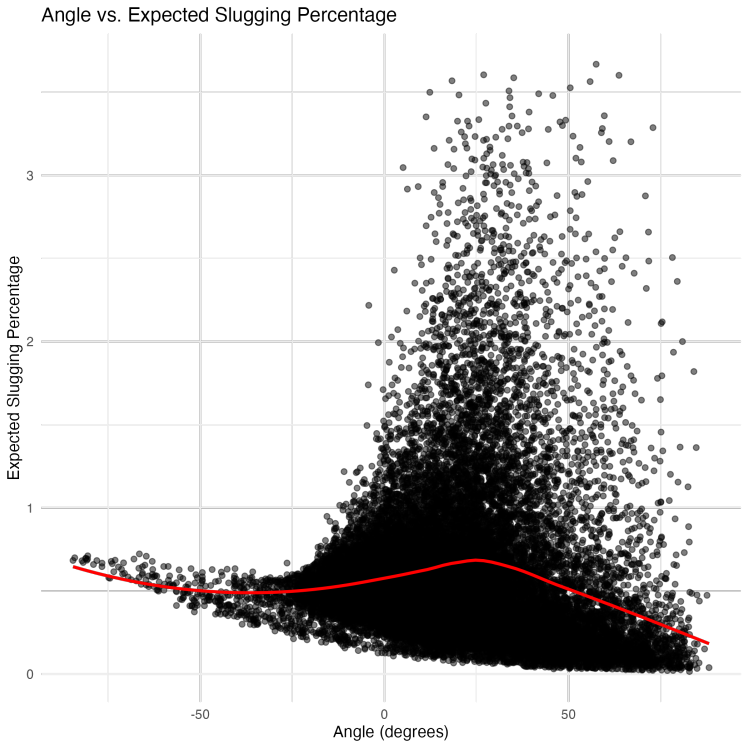

Figure 6: 2023 Expected Slugging Percentage vs. Launch Angle

Both graphs are easier to understand than the xBA graphs. For Exit Speed, the expected slugging percentage increases exponentially with Exit Speed, which makes sense because higher Exit Speed in baseball is generally associated with extra base hits. The reason there is less variation is because in this graph, all of the weakly hit balls that had a chance to be singles in our xBA graphs are now just singles and not super significant in improving xSLG, where extra base hits are weighed more.

For Angle, the expected slugging percentage is highest between 20 and 40 degrees, which makes sense because those are the angles that are most likely to result in line drives, and thus, extra base hits. Statisticians say that a Launch angle is in the sweet spot between 8 and 32 degrees in the MLB, which aligns with this graph.

Section 4: Getting Qualified Hitters

There are multiple ways to qualify hitters based on the number of plate appearances they have, including the standard 3.1 plate appearances per game. However, due to the nature of the minor/independent leagues, where players are moved around a lot and often don't do a full season in one place, I used a more flexible approach. I defined a qualified hitter as someone who has at least 130 plate appearances in a season, which is the definition used for rookie status in the MLB.

I calculated the number of plate appearances for each batter in the 2023 season.

Batter <chr>	total_PA <int>
Juremi Profar	377
Andrew Penner	373
Justin Gideon	372
Nate Scantlin	370
Cito Culver	369
Tj White	362
Chase Dawson	355
Pavin Parks	342
Tucker Nathans	338
Edwin Mateo	337

1–10 of 147 rows

Figure 7: Qualified Hitters by Plate Appearances in 2023

Defining when the end of a PA occurred using this pitch by pitch data was not easy, but I decided to define it by the occurrence of a PlayResult, which means that something happened on the field. It is not exact, but with 130 PAs of all of these players, it should be enough PAs to get a good idea of their expected statistics. Based on checks with Baseball Reference, each player was missing about 30 PAs of data, which shouldn't matter much. Some stadiums were worse than others in keeping track of players.

I now have a leaderboard of 147 players with at least 130 plate appearances in the 2023 Frontier League season. I used this list to filter out the data for our qualified hitters.

With a dataframe of pitches only with qualified hitters, I found each of those qualified hitter's xBA and xSLG. First, I only wanted to keep rows where there is a PlayResult. Then, for each batter, I used a combination of their xBA's, their strikeouts, walks, errors, and sacrifices to get their xBA. Remember that a sacrifice is not considered an at-bat, so I will need to filter those out.

Section 5: Calculating Expected Statistics

I then calculated expected statistics for each player in the 2023 season. For xBA, I found it by summing the probabilities they get a hit on each batted ball event and dividing by at-bats. I did this for xSLG as well, but I weighted each hit by the number of bases it was worth. I then calculated xOBP (by incoporating walks and dividing by Plate Appearances), xOPS (xOBP + xSLG).

Batter <chr>	xOPS <dbl>	xHomeRuns <dbl>
Carson Mccusker	1.1332435	15.0286830
Matthew Warkentin	1.0744562	24.9104132
Josh Rehwaldt	1.0658544	17.6990968
Anthony Brocato	0.9857075	24.1566572
Andrew Czech	0.9459027	17.4058661
Trey Hair	0.9405566	10.1264930
Pat Kivlehan	0.9309176	23.2692622
Aaron Altherr	0.9297727	21.3316787
Brendon Dadson	0.9141681	6.9203566
James Nelson	0.9131971	9.6907570

1–10 of 147 rows

Figure 8: Qualified Hitters Leaderboard by xOPS in 2023

This leaderboard by xOPS gives us a good indication of the overall top offensive performers in the Frontier League, as OPS is considered a fairly predictvie all-encompassing offensive statistic, and using expected values increases its predictability of offensive output.

Section 6: 2023 Run Expectancy Matrix

In order to obtain one of our most important offensive statistics, wOBA, which was a statistic that the Statcast data included for their minor league players, I needed to calculate the run value of each play. To do that, I went back to our pitch-by-pitch data. While Run Expectancy Matrices were typically based on base-out states (number of outs and number of men on), I did not have runner data for the Frontier League, and thus needed to attempt to calculate run value in different ways.

Run Value Method #1: Outs

In our first method of run value determination, Method #1, instead of base-out states, I treated each play as having only an out state. I used the expected number of runs scored until the end of the inning from each out state to see how each play (Single, Double, Triple, Home Run, Walk, Strikeout) changed the expected number of runs scored. I then used this to calculate the average run value for each type of play.

Using this method, I got the following run values for each type of play:

Outs <dbl>	AvgRunsToEnd <dbl>
0	0.6591166
1	0.4039484
2	0.1169742

Figure 9: 2023 Out States for Run Values (Method 1)

PlayResult <chr>	AvgRunValue <dbl>
Double	0.50480802
Error	0.24554893
FieldersChoice	-0.09624484
HomeRun	1.63976480
Out	-0.21039798
Sacrifice	0.41741492
Single	0.24064880
StrikeoutLooking	-0.21873337
StrikeoutSwinging	-0.21670589
Triple	0.67172557

Figure 10: 2023 Run Values by Play (Excluding Walks) from Method 1

Walks had zero run value in this model, as rarely do walks act as batting runs in. Otherwise, the trends looked good. Home runs were worth the most, and outs the least. However, I will want to try a new model to make sure I account for the fact that walks generate runs in the long-term.

Run Value Method #2: Out-PlayResult States

I then tried a revised approach where I calculate the average number of runs scored from that point until the end of the inning, grouped by Outs and the PlayResult.

In this method, I grouped all plays by Outs and PlayResult, and saw how many runs were scored, on average, until the end of the inning, plus the number of runs scored on that play. This allowed us to give walks some positive run value, as they generate runs by the end of the inning. I then used this to calculate the average run value for each type of play.

PlayResult <chr>	AvgRunValue <dbl>
Double	1.3709830
Error	1.1801347
FieldersChoice	0.5417277
HomeRun	2.0287648
Out	0.2232049
Sacrifice	1.3451493
Single	1.0167729
StrikeoutLooking	0.2352212
StrikeoutSwinging	0.2186906
Triple	1.5663265

Figure 11: 2023 Run Values by Play (Excluding Walks) from Method 2

I now had, for each PlayResult, a number that encapsulated both how many runs were scored on that play and how many runs were scored until the end of the inning. This is a much better way to calculate run value, as it gives walks a positive value, and also gives more weight to plays that score more runs.

The only issue with this method is that Strikeouts still had a positive run value, as even after strikeouts, runs could be scored.

To remedy this, I will try Method #3.

Method #3: Outs-PlayResult with Run Expectancy

I then saw that the run value of a play is how many runs are scored on that play, plus the difference in run expectancy by our base-out state before and after the play, plus the number of runs scored until the end of the inning. This method should give us the most accurate run value for each play in theory.

PlayResult <chr>	AvgRunsValueByPlay <dbl>
Double	1.3698797875
Error	1.1760574068
FieldersChoice	0.3096375096
HomeRun	2.0273617011
Out	-0.0047199416
Sacrifice	1.0821983445
Single	1.0080449437
StrikeoutLooking	0.0152210847
StrikeoutSwinging	0.0009515766
Triple	1.5822518823

Figure 12: 2023 Run Values by Play (Excluding Walks) from Method 3

This made the most sense, and Strikeouts now had zero run value (should be negative, but I needed to zero it out anyway later), so I moved forward using these run values. Walks had a positive value slightly less than singles, as expected.

Section 7: Calculating wOBA

Now that I had the run value for each play, I could calculate the weights for wOBA. The first step to obtaining linear weights was to adjust all weights so that the run value of an Out/StrikeoutLooking/StrikeoutSwinging became 0. So, I substracted the value of an out from each play's run value.

These linear weights were contained within the `df2_run_values` and `df3_run_values` dataframes.

In order to obtain the actual weights for our wOBA equation, I needed the average wOBA to equal the average OBP. I had to return to our Pitches dataframe to find the average OBP for all Frontier League batters, not just those who were qualified in 2023.

According to the code, the average OBP for all Frontier League batters in 2023 was 0.3079113. I used our Pitch-by-Pitch data to determine wOBA weights according to our two methods of linear weights. I needed the total number of Walks, Singles, Doubles, Triples, and Home Runs, divided by the total number of plate appearances, to derive the linear weight for each play type.

Using Method #2, I obtained an average wOBA, based on unadjusted linear weights, of 0.2814665; and using Method #3, I obtained an average wOBA of 0.3501425.

I then took our run value dataframes and scaled them so that wOBA equaled OBP. Finally, I applied these weights to calculate wOBA for each qualified hitter in 2023.

PlayResult <chr>	AvgRunValue <dbl>
Double	1.255615995
Error	1.046836796
FieldersChoice	0.348449127
HomeRun	1.975198809
Out	0.000000000
Sacrifice	1.227355089
Single	0.868126529
StrikeoutLooking	0.013145219
StrikeoutSwinging	-0.004938483
Triple	1.469312776

Figure 13: Final wOBA Weights Using Method 2

PlayResult <chr>	AvgRunsValueByPlay <dbl>
Double	1.208807312
Error	1.038362123
FieldersChoice	0.276442355
HomeRun	1.786989402
Out	0.000000000
Sacrifice	0.955823535
Single	0.890613880
StrikeoutLooking	0.017535911
StrikeoutSwinging	0.004987468
Triple	1.395564908

Figure 14: Final wOBA Weights Using Method 3

Sanity Check: Our wOBA weights were about the same for the two methods. Because method #3 incorporated more information, I used those weights to calculate wOBA and xWOBA for each qualified hitter in 2023.

I then calculated wOBA and xWOBA for each qualified hitter in 2023. An output of wOBA+ is included later in this paper.

This was the wOBA formula. Note that I assumed that no walks were intentional and that HBPs were the same as walks (in reality, their run values should be slightly different).

$$\text{wOBA} = \frac{0.8(\text{BB} + \text{HBP}) + 0.89(\text{1B}) + 1.21(\text{2B}) + 1.40(\text{3B}) + 1.79(\text{HR})}{\text{AB} + \text{BB} + \text{HBP} + \text{SF}}$$

Figure 15: wOBA Formula

Section 8: Plus Statistics

In order for our statistics to be more comparable across different leagues and years of the Frontier League, I calculated plus statistics for each qualified hitter. I calculated xwOBA+ and wOBA+ for each hitter, which will be the ratio of their wOBA or xwOBA to the league average wOBA or xwOBA. I did the same for xBA, xOBP, xSLG, xOPS, K% and BB%. I made a couple of assumptions here that should not hurt the study too badly and employed two methods, which I used for different models.

The first method: I disregarded the fact that Plus Statistics are compared against the league-average hitter and just used qualified hitters to get our "league average." I did this because I did not have Pitch by Pitch data for our Single A Statcast data, so I could not calculate league average statistics.

The second method: I used the average statistics of all qualified hitters in 2023 to calculate league average statistics. This was to be used in our comparison to 2024 Statcast hitters.

In both methods, I did not take into account Park Factor because it was rather hard to calculate for this particular pitch by pitch data. Most Plus Statistics take Park Factor into account on a game-by-game basis.

Method #1: Using Qualified Hitters

First, I found the average xBA, xOBP, xOPS, K%, BB%, xWOBA, and xWOBA for all qualified hitters in 2023.

Avg_xBA <dbl>	Avg_xOBP <dbl>	Avg_xSLG <dbl>	Avg_xOPS <dbl>	Avg_K_percentage <dbl>	Avg_BB_percentage <dbl>	Avg_wOBA <dbl>	Avg_xWO... <dbl>
0.2420609	0.2903381	0.3890893	0.6794275	0.2195594	0.06714392	0.2963974	0.2927477

Figure 16: Average Statistics of Qualified Hitters in 2023

Then, I added add the ratio of each qualified hitter's statistics to the league average statistics to the qualified hitters dataframe.

Batter <chr>	xBA_plus <dbl>	xOBP_plus <dbl>	xSLG_plus <dbl>	xOPS_plus <dbl>	K_percentage_plus <dbl>
Carson Mccusker	143.08979	129.27368	194.79144	166.79389	74.86973
Matthew Warkentin	131.89845	118.47783	187.73834	158.14141	129.85385
Josh Rehwaldt	124.64909	119.32626	184.89450	156.87538	121.45534
Brendon Dadson	123.95761	138.68262	131.46585	134.54977	66.42089
Andrew Czech	111.22956	129.35227	146.58426	139.22055	122.95621
Anthony Brocato	122.04860	115.50350	167.14846	145.07914	122.55948
Trey Hair	121.23345	115.48071	155.56121	138.43370	102.31292
Brennan Price	120.34797	120.22199	144.11176	133.90300	120.30953
Aaron Altherr	115.93612	113.29292	154.42216	136.84650	133.57941
James Nelson	120.67766	118.60613	146.19735	134.40686	104.43612

Figure 17: Plus Statistics of Qualified Hitters in 2023, Ranked by xWOBA+

Our ranking of hitters did not change, but now we had numbers that compare hitters to league average, making it easier to compare hitters across different leagues and years.

Method #2: Using All Hitters in 2023

Then, I found the average xBA, xOBP, xSLG, xOPS, K%, BB%, xWOBA, and xWOBA for all batters in 2023.

Total_PA <int>	Total_AB <int>	Walks <int>	xOBP <dbl>	Singles <int>	Doubles <int>	Triples <int>	HomeRuns <int>	K_percentage <dbl>
45669	42256	2877	0.2737911	7691	2116	196	1182	0.2141715

Figure 18: Average Statistics of All Hitters in 2023

We added the ratio of each qualified hitter's statistics to the league average statistics to the qualified hitters dataframe.

Batter <chr>	xWOBA_plus <dbl>
Carson Mccusker	161.75244
Matthew Warkentin	152.48437
Josh Rehwaldt	151.50471
Brendon Dadson	143.82230
Andrew Czech	143.43939
Anthony Brocato	142.36296
Trey Hair	137.09500
Brennan Price	136.47464
Aaron Altherr	136.26095
James Nelson	136.03486

Figure 19: xWOBA+ of All Hitters in 2023

Methods 1 and 2 here got the same order of hitters, but the magnitude to which they were above average ws different. This was because the league average statistics were different in each method. I used Method 2 for our comparison to 2024 Statcast data, as it is more accurate, and this is the way Statcast calculated their Plus Statistics. wOBA+ was much higher when comparing to the average hitter than to hitters who spent played many games in the league this past season.

Section 8: Adding Hit Quality Variables

I then added hit quality variables to our qualified hitters dataframe. In the end, I added columns for Average Exit Velocity, Hard Hit Percentage (Percentage of batted balls above 95MPH Exit Velocity), and Average Launch Angle to all of our dataframes and used them as predictors. Within the affiliated leagues, it is commonplace for poorly performing players to receive more attention because their exit velocity and launch angle are good, even if they struggle in games. So, I wanted to see if this was a valid predictor of promotions to higher leagues.

Section 9: 2024 Frontier League Data

This entire processing of 2023 was then repeated for 2024.

Some notable parts were: -Exit Speed vs. xBA, Exit Speed vs. xSLG, Angle vs. xBA, Angle vs. xSLG graphs were all very similar to the 2023 graphs, which is a good sign that our model is working well.

These graphs were just like 2023, and seem to have even less spread around the trend line. This is a good sign that our model is working well.

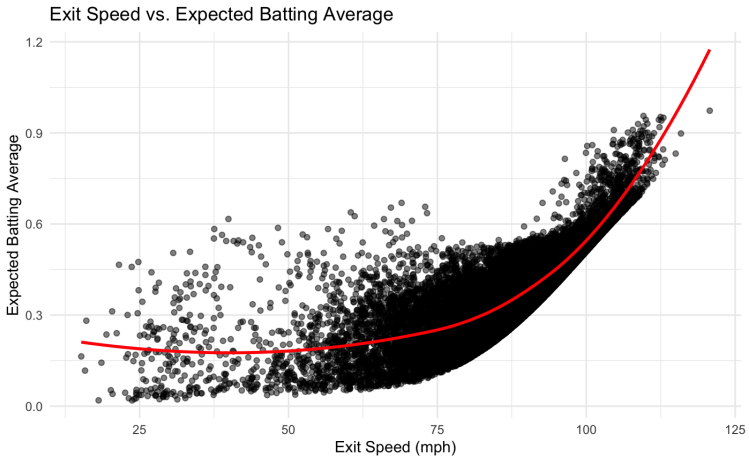

Figure 20: 2024 Expected Batting Average vs. Exit Velocity

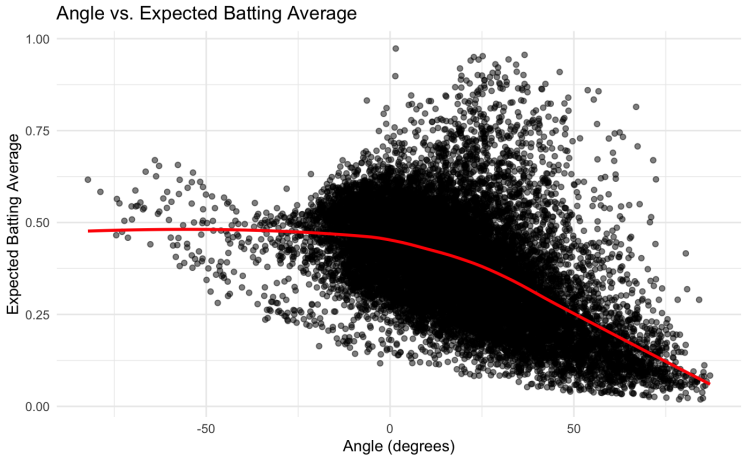

Figure 21: 2024 Expected Batting Average vs. Launch Angle

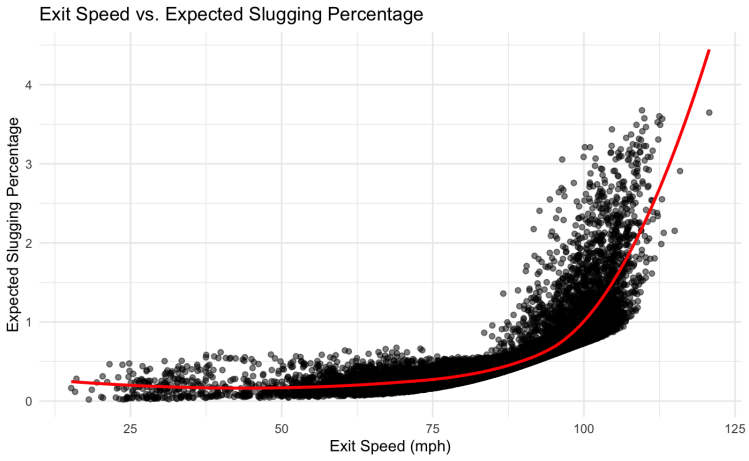

Figure 22: 2024 Expected Slugging Percentage vs. Exit Velocity

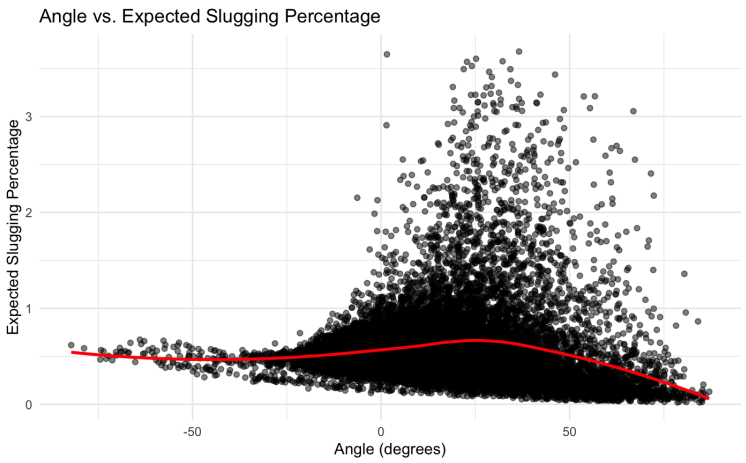

Figure 23: 2024 Expected Slugging Percentage vs. Launch Angle

- I had just 91 qualified hitters from 2024, as there was a lack of data.

Batter <chr>	total_PA <int>
Austin Dennis	307
Chris Kwitzer	296
Jaxon Hallmark	291
David Vinsky	288
Matthew Warkentin	282
Austin White	280
Ian Walters	275
Liam Mcarthur	274
Chase Dawson	273
Jake Boone	262

1–10 of 91 rows

Figure 24: 2024 Qualified Hitters

- There was a new group of top players in 2024.

Batter <chr>	xOPS <dbl>
Matthew Warkentin	1.0208000
Tyreque Reed	0.9505270
Isaac Bellony	0.9330405
Joe Deluca	0.9044970
Andrew Czech	0.8740657
Kyle Fitzgerald	0.8719435
Brendon Dadson	0.8658203
Caleb Mcneely	0.8621559
Aj Wright	0.8397448
Tommy Seidl	0.8320149

Figure 25: 2024 Qualified Hitters Leaderboard

- Run Expectancy Matrices for 2024 Out states were about the same in 2024, which is a good sign that my code was correct. I just used Method 3 for our wOBA weights.

Outs <int>	AvgRunsToEnd <dbl>
0	0.6506227
1	0.3986752
2	0.1108009

Figure 26: 2024 Out State Run Values

Runs until the end of the inning for each out state were about the same.

PlayResult <chr>	AvgRunsValueByPlay <dbl>
Double	1.3698797875
Error	1.1760574068
FieldersChoice	0.3096375096
HomeRun	2.0273617011
Out	-0.0047199416
Sacrifice	1.0821983445
Single	1.0080449437
StrikeoutLooking	0.0152210847
StrikeoutSwinging	0.0009515766
Triple	1.5822518823

Figure 27: 2024 Run Values by Play (Excluding Walks) from Method 3

- Calculating wOBA and xwOBA procedure the same, but new players. The average OBP for all Frontier League batters in 2024 was 0.3212346. I needed to get the total number of Walks, Singles, Doubles, Triples, and Home Runs, and divide by the total number of plate appearances to get the linear weight for each play type.

Using Method #3, I got an average unadjusted wOBA of 0.3515049.

I then took our run value dataframes and scale so that wOBA is equal to OBP. I used these weights to calculate wOBA for each qualified hitter in 2023.

PlayResult <chr>	AvgRunsValueByPlay <dbl>
Double	1.24277090
Error	1.06753675
FieldersChoice	0.28420949
HomeRun	1.83719804
Out	0.00000000
Sacrifice	0.98267909
Single	0.91563726
StrikeoutLooking	0.01802861
StrikeoutSwinging	0.00512760
Triple	1.43477578

Figure 28: 2024 wOBA Weights

These, combined with the value of a walk being 0.81072066, gave me a formula for wOBA:

$$\text{wOBA} = \frac{0.81 \cdot \text{BB} + 0.9156 \cdot \text{1B} + 1.2428 \cdot \text{2B} + 1.4348 \cdot \text{3B} + 1.8372 \cdot \text{HR}}{\text{PA}}$$

Figure 29: 2024 wOBA Formula

And I got wOBA rankings of the following:

Batter <chr>	wOBA <dbl>	xWOBA <dbl>
Matthew Warkentin	0.3594149	0.4201109
Tyreque Reed	0.3922100	0.3984848
Joe Deluca	0.3724609	0.3920346
Isaac Bellony	0.3605793	0.3916729
Aj Wright	0.4041954	0.3882477
Andrew Czech	0.3919323	0.3837028
Kyle Fitzgerald	0.3565158	0.3825859
Brendon Dadson	0.3876057	0.3804442
Caleb Mcneely	0.3718359	0.3661727
Hank Zeisler	0.3813222	0.3619815

Figure 30: 2024 wOBA Rankings

- New Plus Statistics for our 2024 players I calculated new average and plus statistics for our 2024 hitters as well.

Avg_xBA <dbl>	Avg_xOBP <dbl>	Avg_xSLG <dbl>	Avg_xOPS <dbl>	Avg_K_percentage <dbl>	Avg_BB_percentage <dbl>	Avg_wOBA <dbl>	Avg_xWO... <dbl>
0.2410801	0.3066507	0.3675287	0.6741793	0.2082181	0.08977257	0.3103049	0.3067045

Figure 31: Average Statistics of Qualified Hitters in 2024

Batter <chr>	xWOBA_plus <dbl>
Matthew Warkentin	143.50611
Tyreque Reed	136.11885
Joe Deluca	133.91550
Isaac Bellony	133.79196
Aj Wright	132.62192
Andrew Czech	131.06944
Kyle Fitzgerald	130.68792
Brendon Dadson	129.95632
Caleb Mcneely	125.08130
Hank Zeisler	123.64963

Figure 32: xWOBA+ of Qualified Hitters in 2024

BB_percentage <dbl>	xBA <dbl>	xSLG <dbl>	xOPS <dbl>	x1B <dbl>	x2B <dbl>	x3B <dbl>	xHR <dbl>	wOBA <dbl>	xwOBA <dbl>
0.08458789	0.2212319	0.3351998	0.6194165	4002	1146	133.9999	526	0.3212346	0.2833026

Figure 33: Average Statistics of All Hitters in 2024

Batter <chr>	xWOBA_plus <dbl>
Matthew Warkentin	148.29052
Tyreque Reed	140.65697
Joe Deluca	138.38017
Isaac Bellony	138.25251
Aj Wright	137.04346
Andrew Czech	135.43922
Kyle Fitzgerald	135.04498
Brendon Dadson	134.28899
Caleb Mcneely	129.25144
Hank Zeisler	127.77204

Figure 34: xWOBA+ of All Hitters in 2024

Our tables look around the same as our data for 2023, which was a good sign.

Section 10: Other Statistics to Include

Because more and more of the data-driven baseball scene is not necessarily focused on results, but tapping into potential, I also wanted to include a few other statistics that are not included in the above calculations. These include hard hit percentage and launch angle. I got this data from our `pitches_data_2024_m1` dataframe.

Here is an example dataframe of hard hit percentage and launch angle from 2024:

Batter <chr>	Hard_Hit_Percentage <dbl>	Avg_Angle <dbl>
Tyreque Reed	0.49038462	20.40984
Matthew Warkentin	0.45291480	17.76659
Brendon Dadson	0.45045045	14.54444
Joe Deluca	0.43046358	21.01343
Tommy Seidl	0.43010753	22.25989
Andrew Czech	0.42666667	23.57092
Caleb Mcneely	0.41573034	22.85528
Isaac Bellony	0.39639640	16.67522
David Vinsky	0.39520958	18.97361
John Cristino	0.39010989	19.03632

Figure 35: 2024 Hit Quality Dataframe

Section 11: 2021 Statcast Data

After obtaining Statcast Minor League data directly from the Statcast website (screenshot below), I was able to select the columns I wanted to use for Analysis1.R.

Search Results

Rk.	Player	Pitches	Total	Pitch %	K%	BB%	BA	xBA	OBP	xOBP	SLG	xSLG	wOBA	xwOBA	EV (MPH)	LA (°)	
41	Volpe, Anthony	1069	1069	100.0	16.8	19.9	.303	.261	.457	.431	.626	.460	.453	.395	91.5	18	Graphs
54	Julien, Edouard	928	928	100.0	26.6	24.1	.299	.248	.488	.456	.456	.366	.420	.381	90.4	14	Graphs
51	Jiménez, Leo	952	952	100.0	14.5	20.7	.315	.244	.515	.471	.381	.310	.416	.378	86.4	9	Graphs
16	Hauver, Trevor	1355	1355	100.0	26.2	21.5	.289	.236	.446	.410	.500	.414	.410	.369	90.7	18	Graphs
124	Bell, Chad	479	479	100.0	26.4	14.0	.279	.262	.380	.367	.538	.478	.390	.367	91.2	5	Graphs
112	Dunham, Elijah	508	508	100.0	18.1	19.7	.276	.244	.441	.421	.500	.377	.409	.365	89.7	5	Graphs
151	Rumfield, T.J.	414	414	100.0	11.0	20.0	.250	.264	.420	.435	.263	.349	.331	.364	87.2	8	Graphs
131	Hinds, Rece	454	454	100.0	23.1	7.4	.299	.262	.372	.344	.654	.502	.424	.363	88.8	20	Graphs
31	Mack, Charles	1158	1158	100.0	26.0	18.6	.239	.243	.379	.387	.381	.410	.341	.356	87.8	8	Graphs
21	Wells, Austin	1306	1307	99.9	20.9	16.9	.256	.234	.395	.383	.479	.411	.377	.354	88.3	15	Graphs

Figure 36: Minor League Statcast Leaderboard

First, I changed column names to match the Frontier League data.

Once I had all of the statistics, I moved on to the comparison with the 2024 Statcast data. First, I saved all relevant dataframes as CSVs for use in subsequent projects.

Because some teams stopped employing interns to set up their Yakkertech systems for each game, there was less data this season, and thus fewer hitters qualified for our leaderboard. I ended up with only 91 players who reached 130 PAs in our dataset.

Section 12: Manual Data Entry

I then imported my four CSV files (2023 Frontier League using Method 1 of Plus Statistics, 2024 Frontier League using Method 1 of Plus Statistics, 2021 Minor League Statcast using Method 2 of Plus Statistics, and 2024 Frontier League using Method 2 of Plus Statistics) into an Excel file and manually entered some data.

Here's how I edited each file: - 2023 Frontier League using Method 1 of Plus Statistics: Added columns for "Age", "2024.Level" - 2024 Frontier League using Method 1 of Plus Statistics: Unchanged. - 2021 Minor League Statcast using Method 2 of Plus Statistics: Added columns for "Age", "2022.Level", "2023.Level", "2024.Level" - 2024 Frontier League using Method 2 of Plus Statistics: Unchanged.

"Level" was determined by multiple criteria: - Highest Level reached by a player during the season with minimum 130 PAs - If the player did not reach 130 PAs in any level, the level at which they had the most PAs was chosen. - If a player did not have any statistics for a season, they were considered "DNP" = "Did Not Play." This was common for Frontier League players, who often retire upon seeing the realities of independent league baseball. - Winter Leagues were not considered - For the 2023 Frontier League data, Independent Leagues were considered separately, but not for Statcast, as there was enough data in affiliated leagues.

Modeling/Analysis

Analysis1.R

Predicting Leagues Reached by Frontier League Players

In this section, I predicted probabilities that every qualified Frontier League hitter would reach higher levels based on 2024 Statcast data.

I defined three predictor sets:

- Traditional Statistics: OBP, SLG, K%, BB%
- Expected Statistics: Hard Hit%, Average Launch Angle

- Hit Quality: xWOBA+

A Summary of Analysis1.R:

I began by loading the necessary libraries and reading data from Excel spreadsheets. I then standardized column names to ensure consistent formatting and calculated an average xwOBA value from Statcast data, using it to create xwOBA+. Next, I consolidated multiple player level categories (such as A+, AA, AAA, MLB, etc.) into a simplified set of categories, treating this as a target variable for modeling and converting it into a factor for classification.

After preparing the data, I defined a custom function to train a multinomial logistic regression model given a set of predictors, calculate predicted probabilities, and compute a log loss value to assess model quality. I tested three distinct predictor sets—Traditional, Hit Quality, and Expected—by training a model for each and comparing their log losses.

Having identified the best predictor set in each year (which, for all three years, turned out to be Traditional Statistics), I then used that model to predict probabilities for future seasons (2025, 2026, and 2027) for Frontier League players. I added these probabilities as new columns in the FL_2024 dataframe, thus providing insights into the players' likelihood of advancing to particular levels (A+, AA, AAA, MLB, or Other) in subsequent seasons. Throughout this process, I maintained a consistent workflow of reading, cleaning, modeling, evaluating, and predicting, culminating in a dataset enriched with probability-based forecasts for player advancement.

The Traditional statistics served as the best predictor set of future success for Statcast players, so I used that to map probabilities for the Frontier League players in 2024. I made seven graphs to visualize this.

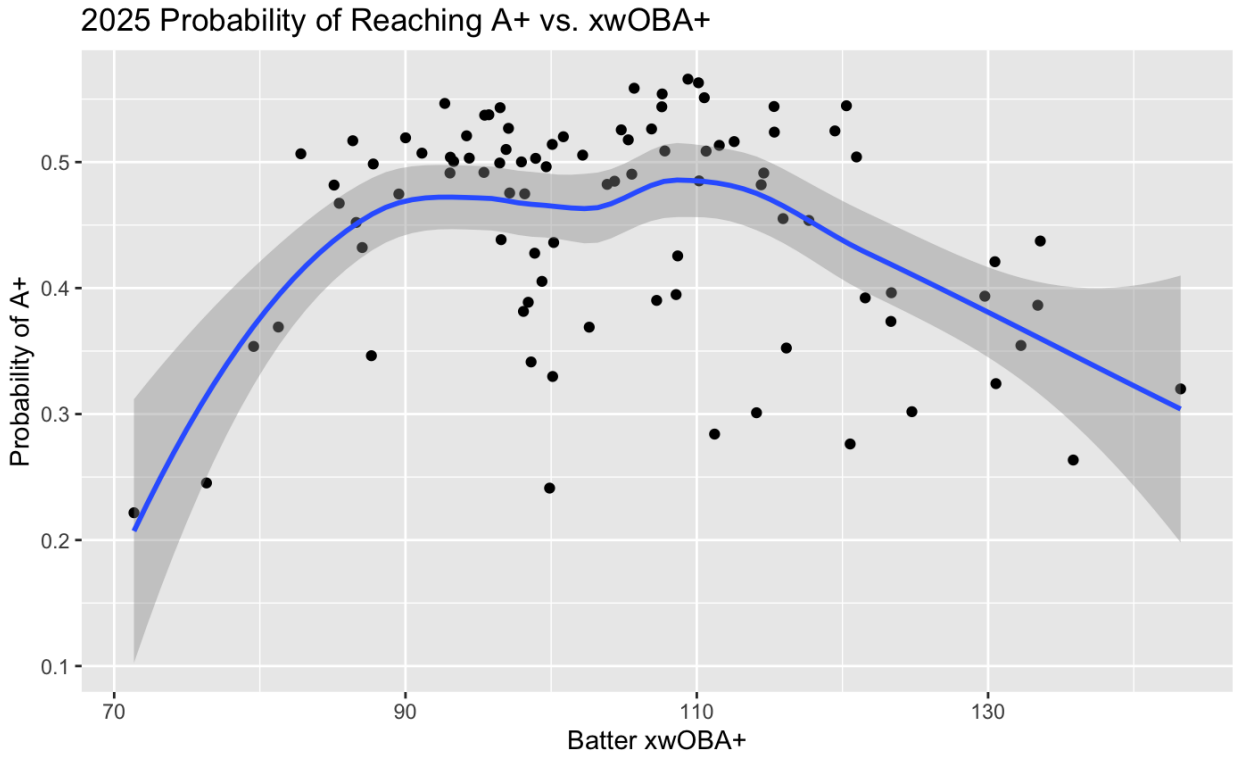

Figure 37: 2025 Probability of A+ with xWOBA+

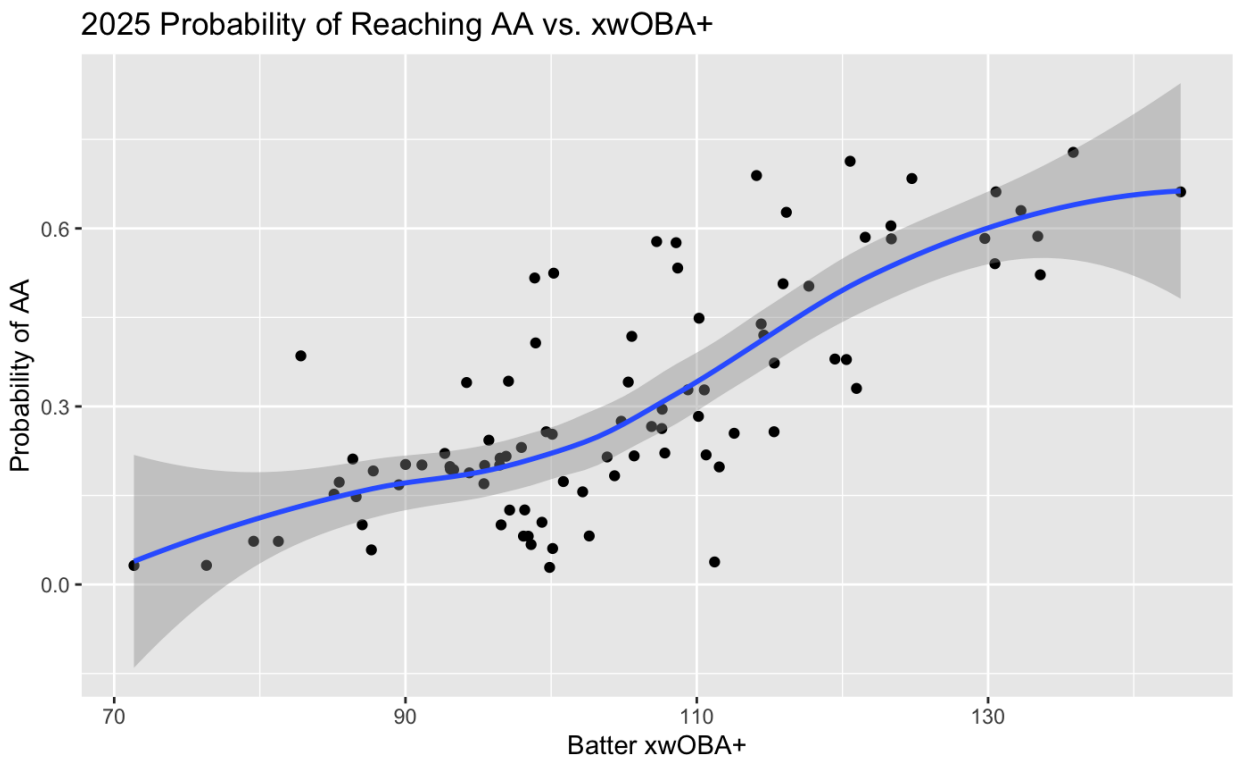

Figure 38: 2025 Probability of AA with xWOBA+

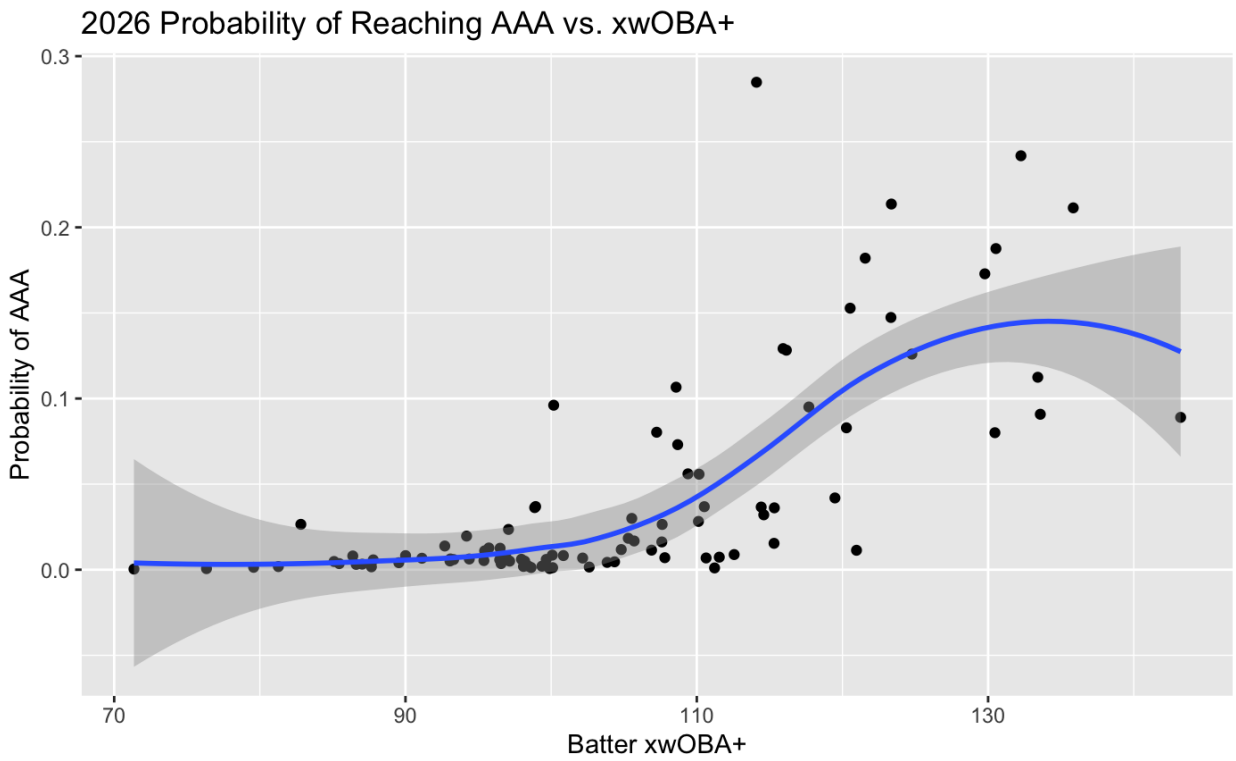

Figure 39: 2026 Probability of AAA with xWOBA+

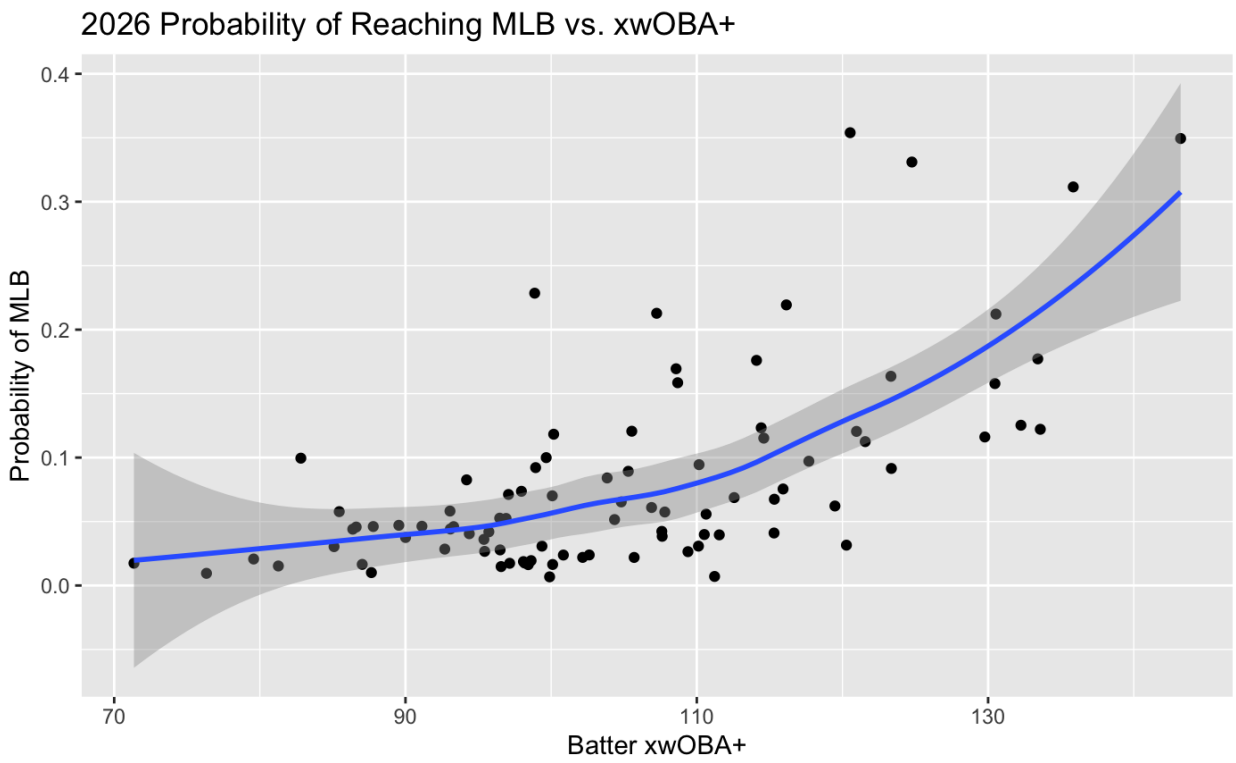

Figure 40: 2026 Probability of MLB with xWOBA+

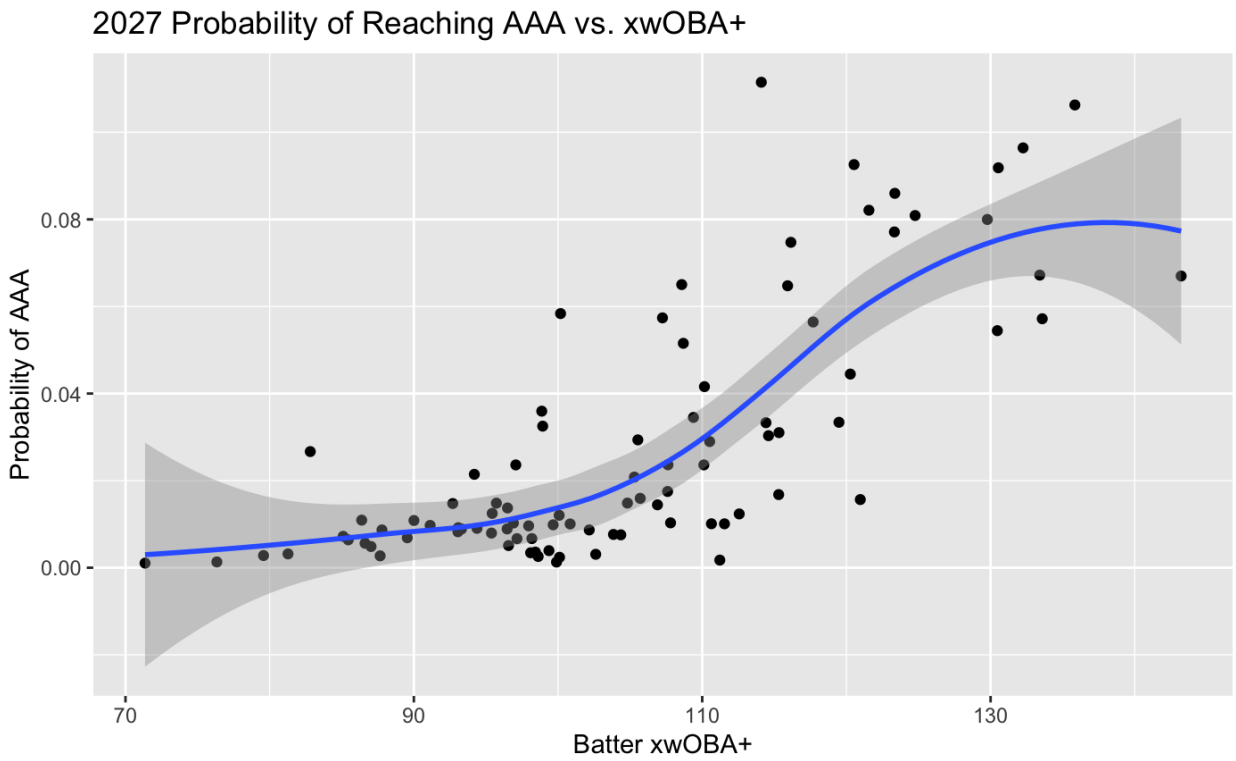

Figure 41: 2027 Probability of AAA with xWOBA+

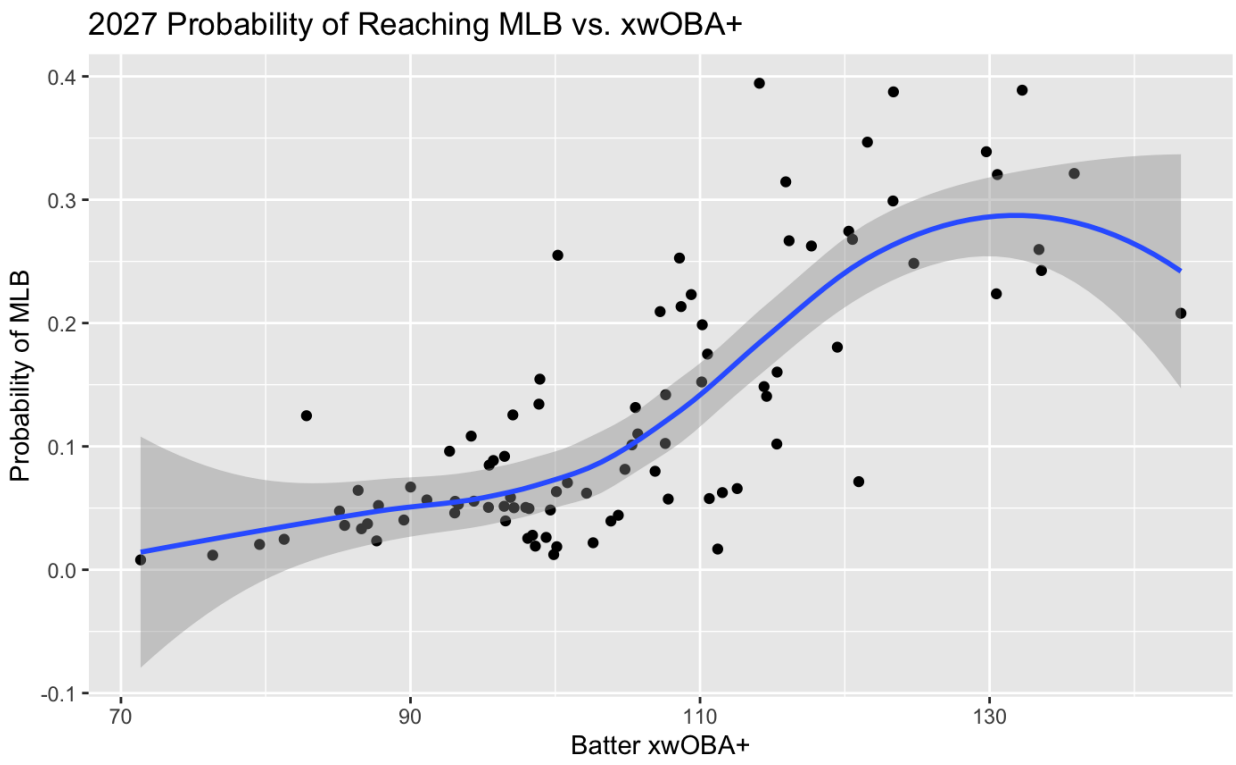

Figure 42: 2027 Probability of MLB with xWOBA+

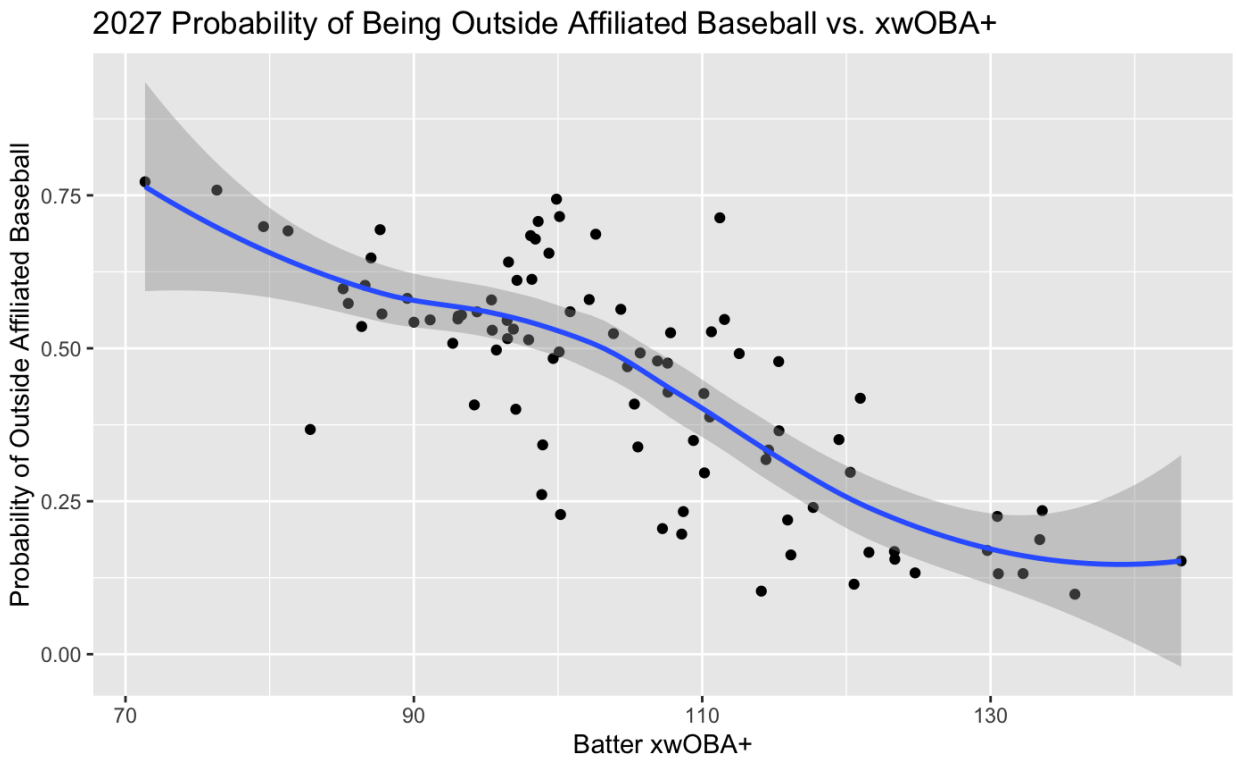

Figure 43: 2027 Probability of Not Playing Affiliated Baseball with xWOBA+

I see pretty clear relationships worth noting. Though it is not exactly linear, players with higher xWOBA+ (and thus, better offensive players) tend to make it to higher levels of affiliated baseball, even if the traditional predictor set was the most predictive. It is obviously harder for players to get to AAA and MLB, but it is not impossible, as there are plenty of players with projections higher than 10% to make it to those levels. The chances of players not playing affiliated baseball are also higher for players with lower xWOBA+, which makes sense. Our model is likely a good one, as it is not exactly related to talent level (as represented by xWOBA+), but contains variation likely representing the multitude of factors that are in play with a player moving up the ranks.

Analysis2.R

Goal: This analysis focused on predicting outcomes of 2024 Frontier League players based on similar players in the Frontier League in 2023.

The first thing I did was manually create an age column of these player's ages.

I then created the following table to see the highest level that the 2023 Frontier League players reached in 2024.

A	A+	AA	Amer	Atl	DNP	FL	Foreign	Mex	PL
1	1	1	3	13	40	76	2	4	6

Figure 44: Highest Levels Reached by 2023 Frontier League Players in 2024

As I can see, most players stayed in the Frontier League, although some made it to equivalent or higher levels of affiliated baseball. 40 players did not play professionally in 2024, and although a couple of those were season-ending injuries, it is not a good sign to see so many players just quit baseball after playing in the Frontier League. Age in our model should come in handy for this, because that should help to explain why older players tend to give up on playing professionally.

Next, I made a graph of xWOBA+ and 2024 Level for our 2023 Frontier League Players.

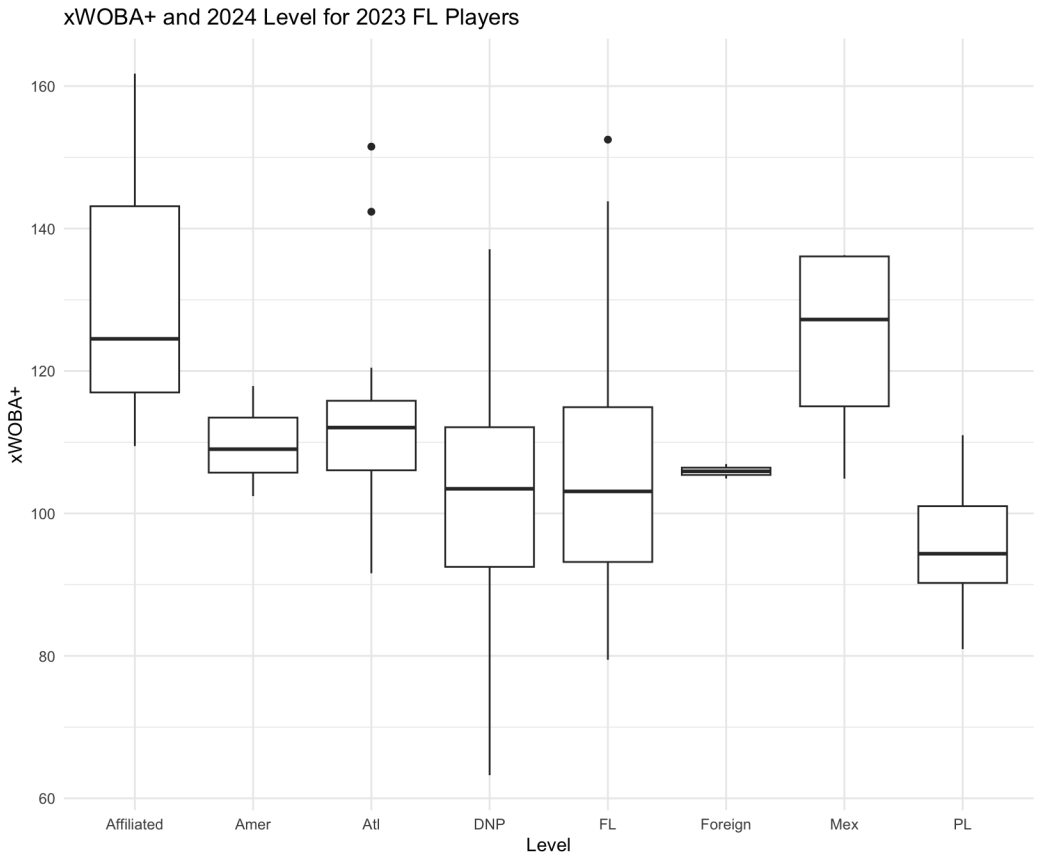

Figure 45: xWOBA+ vs. 2024 Level for 2023 Frontier League Players

I saw a couple of noticeable trends from this graph.

- Our one player to make it to AA in 2024 after playing in the Frontier League in 2023 had the highest xWOBA+. This makes sense, as the jump from independent league baseball to AA is a great one, so a team taking the chance on an independent league player and putting them into AA requires great talent.
- While the range of players is high, the players who did not play in 2024 are slightly lower in talent level than those who played in other leagues. This is likely because players who are not good enough to play in higher leagues are more likely to quit baseball.
- Players who stayed in the Frontier League are about average, though there is a wide range and a outlier on the high xWOBA side, signalling that there are players who are good enough to play in higher leagues but choose to stay in the Frontier League.
- The Pioneer League seems to get worse players than other leagues, and the Mexican League tends to get better players on average versus other leagues.

Later in my analysis, I decided that "Hit Quality" could also be a useful set of predictors for these players. I added Average ExitSpeed and Angle (exit velocity and launch angle, respectively) to my model.

A Summary of Analysis2.R:

I began by loading the necessary libraries and reading the Frontier League (FL) hitters' data from Excel files. After I examined the distribution of levels reached by 2023 players, I consolidated rare categories (A, A+, AA) into a single "Affiliated" category and plotted both the distribution of levels and their relationship

to xWOBA+. I also merged additional hitting quality metrics—such as average exit velocity and launch angle—into both the 2023 and 2024 dataframes by linking them with pitch-level data from CSV files.

Once I had prepared the data, I converted the 2024.Level variable to a factor, created training and testing splits for the 2023 data, and defined sets of predictor variables: traditional stats, rate stats, expected stats, hit quality stats, and various combinations of these. I then fit multiple multinomial logistic regression models, each using a different predictor set, and evaluated their performance by predicting 2024 levels for the 2023 players. I assessed the performance of these models through confusion matrices.

I chose to model with multinomial regression because the outcome variable was categorical with more than two levels. During this analysis, I built six distinct multinomial logistic regression models, each leveraging a different set of predictor variables. Model 1 focused solely on Age, providing a baseline for how a player's age might influence their future level. Model 2 incorporated expected stats such as xwOBA+, aiming to capture underlying skill. Model 3 relied on traditional stats like on-base percentage and slugging percentage, gauging how conventional metrics might predict advancement. Model 4 used hit quality measures, including average exit velocity and launch angle, to determine whether the physical quality of contact offered a predictive edge. Model 5 examined rate stats, such as strikeout and walk percentages, to see if plate discipline metrics enhanced prediction. Finally, Model 6 combined traditional and rate stats, as those were the only two models that outperformed the no-information rate statistic. I trained, evaluated, and compared all of these models, which allowed me to make an informed decision on which predictor set best explained and predicted future player levels. I found that Model 6 had the best accuracy. However, none of the models were significantly better than the no-information rate, so I tried a seventh model.

Next, I consolidated the outcomes further, grouping various affiliated and independent leagues into three simpler categories: "DNP," "Independent," and "Affiliated." I plotted the resulting distributions and xWOBA+ patterns again. Then I retrained and evaluated another model (Model 7, using Model 6's variables) to check its predictive performance. Model 7 performed worse than Model 6, as this further grouping made the no-information rate model much stronger by comparison.

In the end, Model 6 had the highest accuracy and lowest p-value for being better than the no-information rate, so I decided to use it for predicting where our 2024 Frontier League players would be in 2025.

These are the outputs from those final dataframes:

	Batter	Affiliated
14	John Cristino	0.31000320
6	Andrew Czech	0.10421702
9	Caleb Mcneely	0.09195066
24	Aaron Simmons	0.06208829
38	Tj Reeves	0.04798488

Figure 46: Highest Probabilities of Making it to Affiliated Leagues in 2025

	Batter	DNP
91	Carson Clowers	0.7653331
85	Elvis Peralta	0.6207194
77	Chris Burgess	0.5844244
83	Kingston Liniak	0.5771829
79	Joe Johnson	0.5599653

Figure 47: Highest Probabilities of Not Playing in 2025

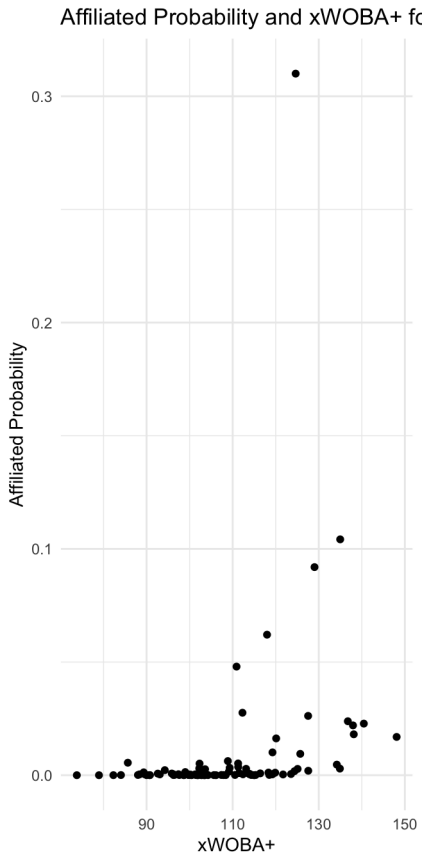

Figure 48: Affiliated Probabilities for 2025 Frontier League Players

There are no significant trends based on this graph, although it seems like John Cristino has a good chance of going to the affiliated leagues in 2025 based on his traditional statistics. There are just simply not enough players making it out of the Frontier League, so it is hard to predict otherwise, even with extensive statistics on each player. There are plenty of players with a higher than 50% chance of leaving baseball in 2025, signifying the difficult nature of the Frontier League.

Visualization and Interpretation of the Results

There is no specific visualizations section, as visualizations were included throughout the analysis.

Conclusions and Recommendations

Takeaway 1: While Single A and the Frontier League may have similar talent levels, Frontier League players do not get the same opportunities as their affiliated counterparts.

As I can see by comparing probabilities from Analysis1.R and Analysis2.R, probabilities of reaching higher affiliated leagues are much higher for Single A players, even though talent levels are considered about equivalent.

One possible reason behind this: Often, choosing independent baseball is not because the affiliated leagues wouldn't be interested, but because equivalent minor league teams do not offer the same relaxation and, as has been described to me, "chill factor" as the Frontier League and other independent leagues. Likely stemming from the fact that the Frontier League has much less at stake in terms of "investing" in these players than affiliated teams do, this fact allows very good players to stay in the Frontier League or other independent leagues while their Single A equivalents rise through affiliated ranks towards the MLB. It should also be noted that often, Frontier League players were "blocked" by a glut of organizational talent at their position in the affiliated leagues, and the Frontier League is a way to continue to get professional reps and show off their talents.

Another possible reason: This could also be due to a lack of data, or a lack of a desire to use the available data. Anecdotally, I have met people in the industry who have expressed disinterest with Frontier League Yakkertech data, simply because it not as as clean and requires much more manual validation than other data sources.

Takeaway 2: There is a wealth of talent in the Frontier League, and Major League Baseball clubs could optimize values by exploring their options.

As I can see in our analysis of similarly talented players in Single A and the Frontier League, there are Frontier League hitters who dominate their league and likely have the skills to have in higher affiliated baseball. I find plenty of players who have a 50% or higher chance of making it to the affiliated leagues in 2025 from our Statcast data comparison, so it would be smart of MLB clubs to look for talent in the Frontier League.

Takeaway 3: Minor League promotions are still based on archaic principles that traditional statistics matter.

As I saw with our log loss on the predictions for higher levels for the 2021 Statcast data and with the 2023 Frontier League data, traditional statistics (batting average, on-base percentage, slugging percentage) are still the best predictors of whether a player gets called up in future years. While there are many more factors at play on whether a player gets called up (monetary investment from a signing bonus, defense, sprint speed, no one "blocking" them at higher levels), this should not be the case with more predictive expected statistics, and undeniable hit quality statistics (exit velocity, launch angle) available for the Single A level.

One possible reason: This data is relatively new, and the usage of it by Major League ballclubs is likely going to take some time.

Another possible reason: Some analytics departments are behind in their usage of advanced analytics, and they will continue to incorrectly evaluate talent at the lower levels.

Takeaway 4: More data is needed.

As seen in Analysis2.R there is a need for more data to see what the true probabilities of Frontier League players are for making it to higher levels. This data is new and unstudied, and I will need years of MLB clubs evaluating Frontier League players to get a true gauge. Also, many Frontier League players do not move out of independent baseball, making it difficult to project the players that could using multinomial logistic regression.

Acknowledgements

I'd like to acknowledge my mentors at the New York Boulders, who encouraged me to use this data for this project.