INTRODUCTION

In this analysis, I compare two measurement systems namely, System A and System B which measure baseball Exit Velocity and Launch Angle metrics, where System A is more accurate in measurement than System B.

The goal is to calibrate System B to closely match System A, ensuring accuracy across both systems. Additionally, this analysis also projects future True Speed performance metrics for batters, using this calibrated data alongside with the Exit Velocity of System A to offer reliable insights

DATA OVERVIEW

In this section, I discussed the comprehensive overview of the data used in the analysis, including quality checks, distribution patterns, descriptive statistics, and insights drawn from preliminary visualizations.

Data Quality Check

The initial quality assessment revealed some missing values and variations in data coverage between the different systems and variables.

	Total NaN	Percent of NaN	Nunique	Dtype
speed_A	7572	10.319591	65803	float64
vangle_A	7572	10.319591	65803	float64
speed_B	1402	1.910733	71973	float64
vangle_B	1402	1.910733	71973	float64
batter	0	0.000000	816	int64
pitcher	0	0.000000	645	int64
hittype	0	0.000000	5	object

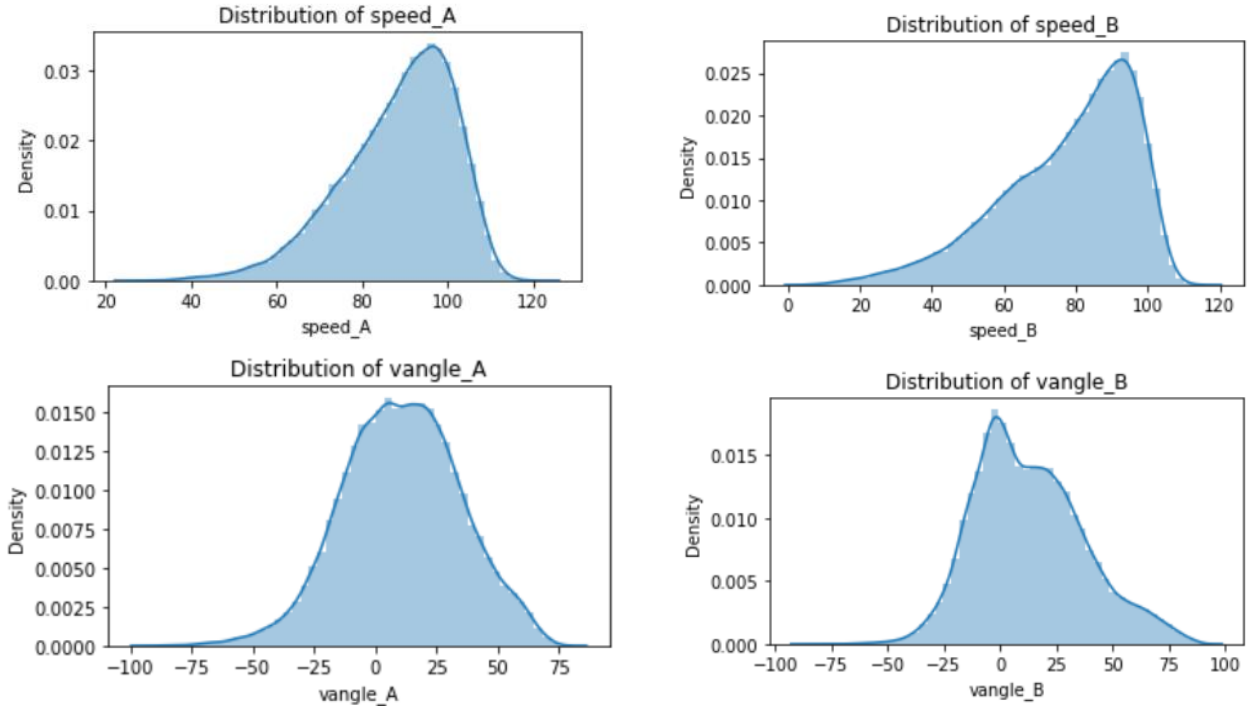
This analysis shows that there is a significant proportion of missing values in Exit Velocity (speed_A) and Launch Angle (vangle_A) of 10.32% and a smaller percentage in Exit Velocity (speed_B) and Launch Angle (vangle_B) of 1.91%. There is a unique count of values for batter, pitcher, and hit type which shows that there is diversity of players and hit types in the dataset, which will be important for assessing individual performance patterns.

Distribution of Key Variables

I further performed a distribution analysis to understand the behavior and distribution of the variables (speed_A, speed_B, vangle_A, and vangle_B) such that it can reveal distinct patterns and potential measurement discrepancies. From the visualizations, speed measurements (speed_A and speed_B) show a right-skewed distribution which suggests that lower speeds are more common, with some high-speed outliers. In contrast, angle measurements (vangle_A and

vangle_B) are more symmetrically distributed, implying a balanced capture of upward and downward hit angles.

The discrepancies in mean values and spreads between the systems highlight potential calibration needs where System B records a broader range of values at generally lower speeds and higher angles, which may introduce inconsistency without adjustment.



Descriptive Statistics Summary

To further understand the dataset, a descriptive statistical summary was performed as shown below

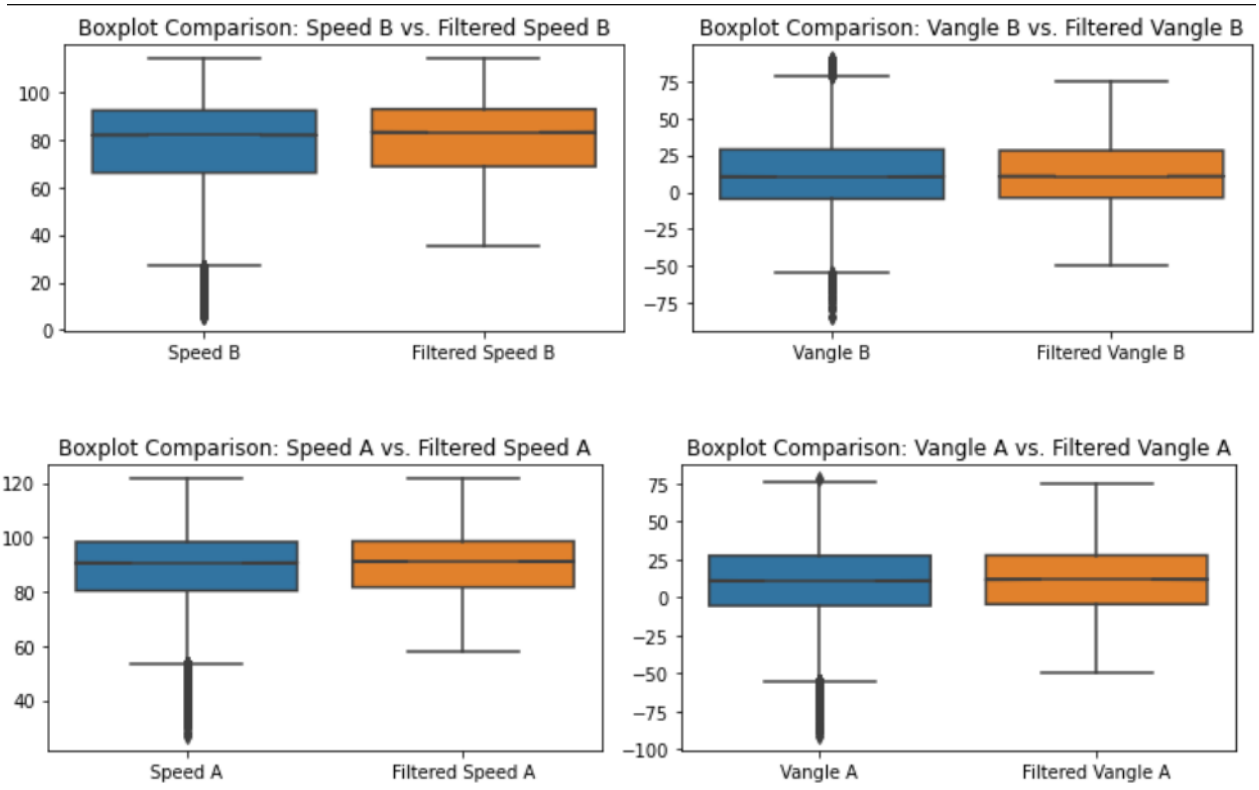
	batter	pitcher	speed_A	vangle_A	speed_B	vangle_B
count	73375.000000	73375.000000	65803.000000	65803.000000	71973.000000	71973.000000
mean	365.135005	289.957547	88.425429	10.853558	77.748243	13.094305
std	229.816539	186.406565	13.192938	24.024058	18.621538	24.429536
min	1.000000	1.000000	26.461824	-91.898629	5.152318	-85.090929
25%	170.000000	121.000000	80.362652	-5.558918	66.074054	-4.634815
50%	341.000000	283.000000	90.638747	11.005251	81.886750	10.546230
75%	550.000000	447.000000	98.317818	27.404976	92.326944	28.880589
max	816.000000	645.000000	121.847456	78.460978	114.403356	90.900819

The descriptive analysis gave insights that, while both systems are capturing the essential characteristics of hit speed and angle, System B records slightly lower speeds of mean of 77.75. and slightly higher launch angles of mean of 13.09 than System A having a speed mean of 88.43 and launch angle mean of 10.85. What this means is that System B needs certain adjustments that will ensure that the True Speed projection in the following season is accurate. This foundational analysis is pointing towards the need for calibration or alignment, particularly for System B, to ensure consistency and accuracy when comparing data from both systems.

DATA PROCESSING

Outlier Detection and Treatment

There is need to as well identify potential outliers that might affect the projections, and this was done by plotting Boxplots for vangle_A, speed_B, vangle_B, and speed_A



Observations:

- **The Outliers** identified are the data points that lie significantly outside the interquartile range (IQR) which indicates values that deviate from the main data pattern.
- **Action Taken**: Outliers were removed from the dataset to create a cleaner, more reliable dataset for analysis. Removing these extreme values helps in achieving more accurate and

stable projections which will further reduce the noise that could impact model calibration/adjustment.

Key Observations from Boxplots:

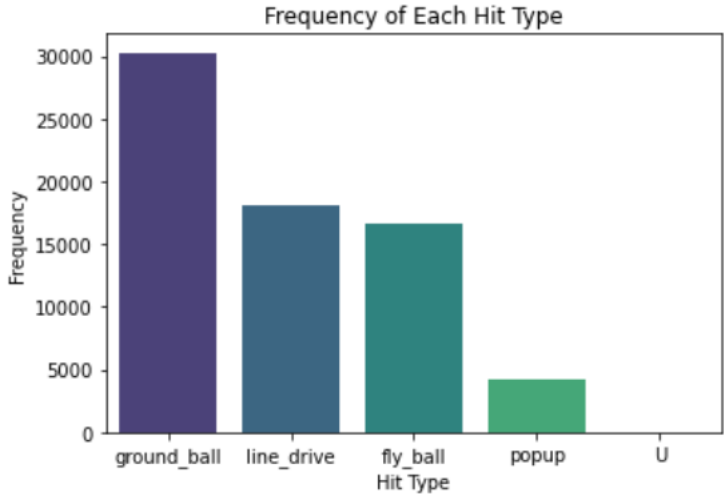
- **System B** tends to record lower speeds than System A, as evident in:

  Median Speed: The median for System B is consistently lower than System A.

  Range of Values: System B includes lower values with some extremely low outliers, suggesting that it may capture slower speeds more often, potentially due to errors measurement

  Filtered Comparison: After removing outliers, System B's median speed gets closer to System A, though it still leans slightly lower, supporting the need for measurement alignment. This can be seen in the plots above where yellow color boxplots are the filtered system without outliers.

Hit Type Distribution: A frequency count of each hit type was performed to further understand the prevalence of hit types and this is because it will provide further analysis or model adjustments specific to each type.



This analysis shows that ground balls are the most frequent hit type, followed by line drives, fly balls, and popups, with two unidentified cases.

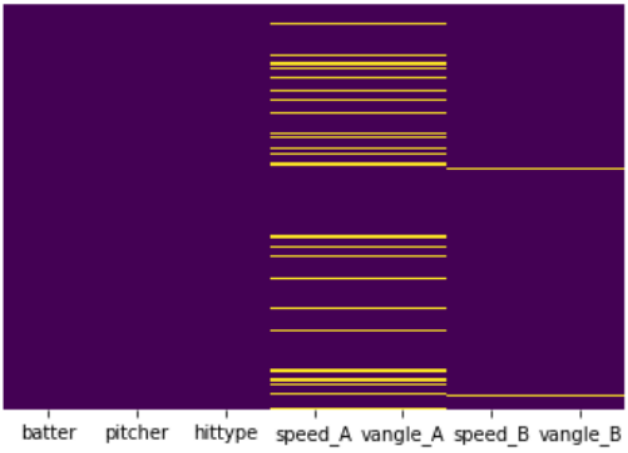
Treating Missing Values

In the data cleaning process, I used a strategic approach to fill the missing values for exit velocity (speed_A and speed_B) and launch angle (vangle_A and vangle_B) based on hit type (hit type) and batter specific patterns from the data. To ensure only relevant records were considered, I first removed rows where the hit type was labeled "U," as these were not part of the hit types of interest and would not contribute meaningfully to the analysis, which reduces the size of the data from 73375 to 69282 rows/data points.

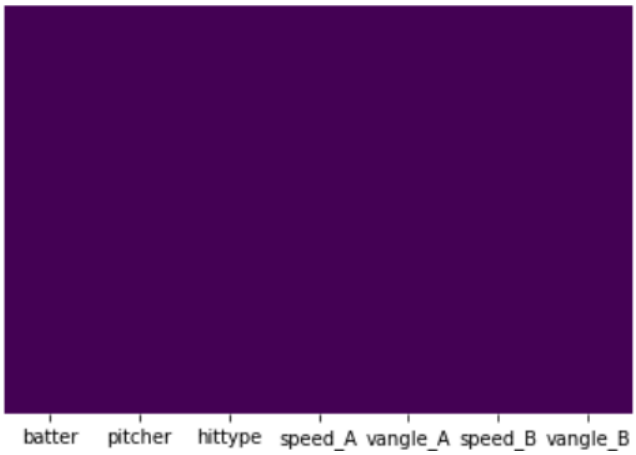
For the filling strategy, I calculated median values for each combination of hit type and batter. This hit type-batter median was used for the replacement of missing values with representative averages

specific to each batter and hit type combination. This approach is important because batters exhibit distinct performance tendencies for different hit types coupled with the fact, they have different hit-types frequency, therefore using their specific median values captured these individualized patterns. However, for some batters, that might not have insufficient records for certain hit types, making median calculation unreliable due to limited data. For these cases, I created a fallback by calculating median values for each hit type alone, which allowed filling based on the broader hit type group when individualized data was unavailable.

The approach of filing these missing values maintained the data's inherent structure and variability. If a batter-specific median was unavailable, the general median for that hit type provided a reasonable proxy which is effectively balancing detail of individual batter. It is a good strategy because it reduces the bias that might occur if only batter-wide averages were used, thereby preserving each batter's unique impact on the data. This ensures that any missing values are filled meaningfully while reducing potential skew which will make subsequent analyses more accurate and reliable. Finally, I verified the effectiveness of this approach using a quality check to ensure no remaining missing values, confirming the completeness and consistency of the dataset for further analysis.

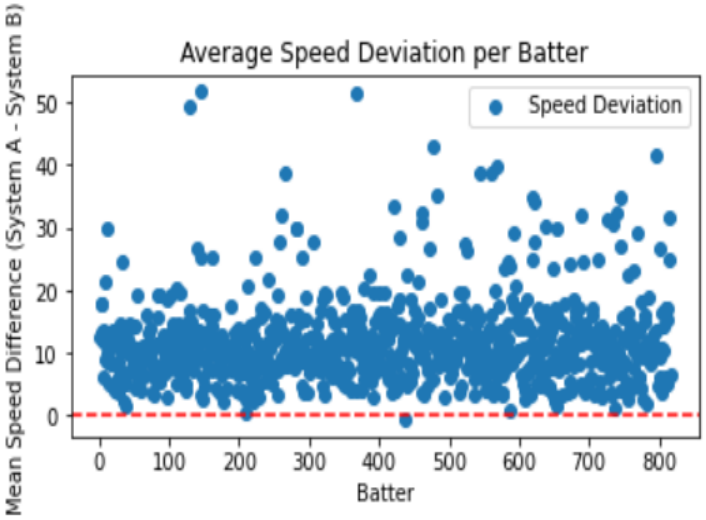


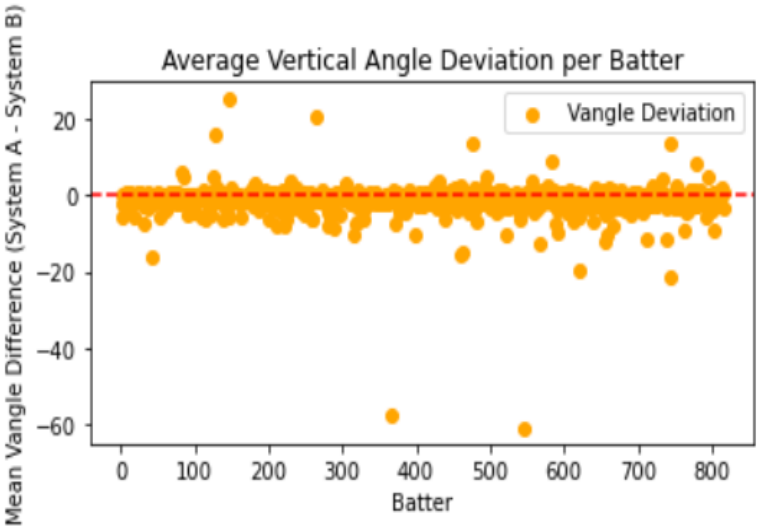
Data with Missing Values



Data after treating Missing Values

Exploring the Deviations Between System A and System B After Data Cleaning

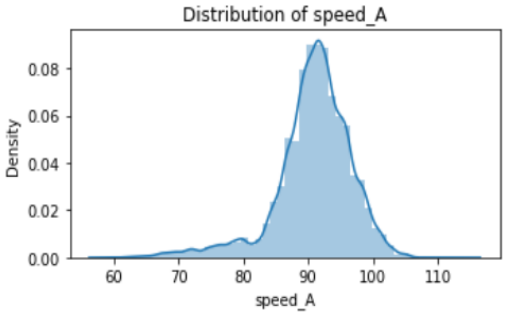


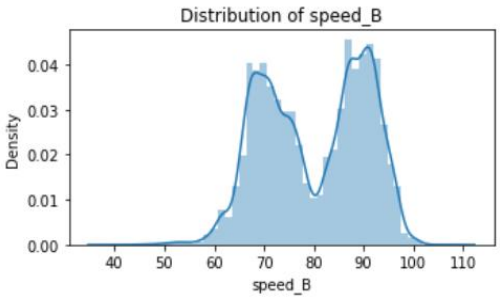


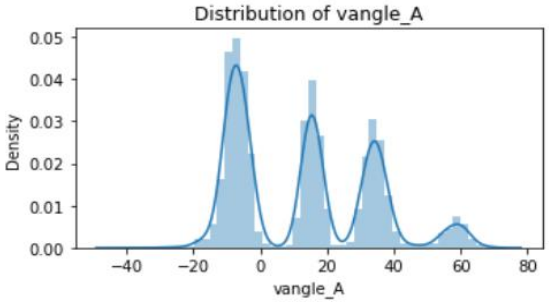
Having observed those visuals, it shows there is significant difference in speed deviations, Since System B shows larger and more varied speed deviations compared to System A, it confirms that System B's raw measurements are less consistent with System A, as indicated by the variability in the scatter plot. These larger deviations justify a stronger calibration or measurement adjustment approach to bring System B closer to System A, particularly for exit velocity (speed) values.

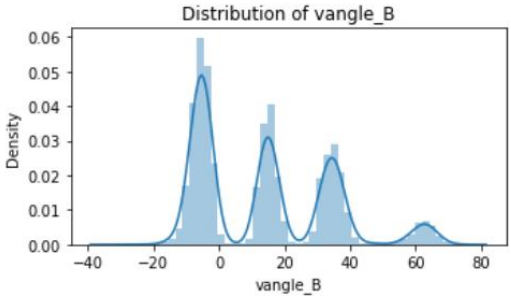
For the launch angle, it deviates with smaller spread out, suggesting that System B's launch angle measurements are closer to System A in terms of distribution. However, some consistent biases remain, as seen from the clustering around zero with occasional outliers.

Rechecking The Distribution of The Data After Data Cleaning









Compared to the initial distributions before data was cleaned, the data cleaning process brought significant improvements to the clarity and interpretability of the dataset making it to be more representative of the typical speed and angle values observed. By removing outliers and carefully imputing missing values, distinct clusters emerged in the vangle_B, vangle_A, and speed_B distributions which provides hint at possible patterns tied to specific hit types or batter-pitcher dynamics.

By removing those outlier results to a normalizing effect, especiually for Exit Velocity (speed_A), which now aligns more closely with the expected distribution for batters in System A. Additionally, the multiple peaks observed in launch angle and exit velocity (speed_B) offer opportunities to segment the data further, potentially revealing insights by hit type or player tendencies.

Descriptive Analysis After Cleaning

	batter	pitcher	speed_A	vangle_A	speed_B	vangle_B
count	69282.000000	69282.000000	69282.000000	69282.000000	69282.000000	69282.000000
mean	364.763055	290.164170	90.830590	12.661297	80.166625	13.595005
std	229.850070	186.416579	5.956201	20.667606	10.436417	20.562277
min	1.000000	1.000000	58.187713	-42.474370	38.031542	-32.641852
25%	170.000000	121.000000	88.301596	-6.494889	70.753729	-4.869675
50%	341.000000	283.000000	91.394238	13.712355	81.323933	13.115106
75%	550.000000	447.000000	94.591727	31.729070	89.622198	31.931567
max	816.000000	645.000000	114.583564	71.349293	108.923314	74.862364

After cleaning the data, key insights into central tendencies and variability were seen which will further enhance the dataset's interpretability.

For **Speed (speed_A vs. speed_B)**, System A records a higher average speed of 90.83 mph compared to System B's 80.17 mph, pointing to a potential bias or offset in System B's measurements. Additionally, System A's lower standard deviation (5.96) suggests it captures speed more consistently than System B, which has a higher variability (10.44), indicating potential precision issues in System B's data.

In **Vertical Angle (vangle_A vs. vangle_B)**, both systems show similar averages (12.66 for vangle_A and 13.60 for vangle_B), reflecting minimal bias in angle measurements. The standard deviations are also closely aligned (20.67 for vangle_A and 20.56 for vangle_B), indicating that both systems capture vertical angles with similar variability.

For **Batter and Pitcher IDs**, unique identifiers for 816 batters and 645 pitchers remain consistent post-cleaning, preserving the dataset's structure for subsequent analysis.

Understanding Typical Speed and Angle Patterns by Hit Type

After this I analysis the different speed and angle patterns for each hit type which reveals that fly balls and popups have distinctly higher average launch angles compared to ground balls, with popups exhibiting the sharpest upward angles, while line drives demonstrate moderate angles and slightly higher speeds, indicating a flatter and faster trajectory among hit types, which helps in avoiding biases linked to common outliers in specific hit types.

General Insights from Exploratory Data Analysis

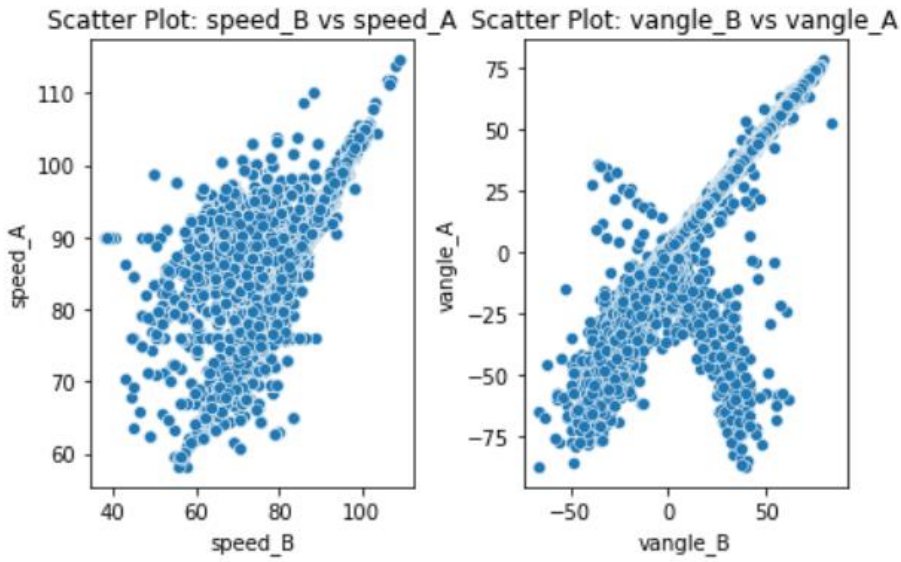
On analyzing both System A and System B measurements, it was very clear that there are discrepancies in both Exit Velocity and Launch Angle. System B consistently records lower speeds than System A, with an average difference of around 10 mph and greater variability, suggesting reduced precision and a potential bias in System B's speed measurements. This variability makes System B's speed data less reliable on its own, especially as System A captures speed with tighter consistency.

For vertical angles (vangle), the two systems show closer averages and similar variability, though System B displays a slight upward bias.

Given these systematic differences, the calibration of System B to align with System A, particularly in exit velocity measurements, is essential such that the calibration will reduce discrepancies, allowing both systems to converge in scale and central tendency, which will result in a more accurate, combined dataset for projections and further analyses.

MODELING A MEASUREMENT ADJUSTMENT OF SYSTEM B TO ALIGN WITH SYTEM A

To determine the most suitable approach for adjusting System B to align with System A, I examined the correlation between their respective speed and angle measurements. The analysis reveals a moderate correlation of 0.576 between Exit Velocity (speed_A and speed_B) which suggests that a linear or spread adjustment might be appropriate for aligning speed values. In contrast, launch angle (vangle_A and vangle_B) exhibit a very high correlation of 0.995, indicating near-perfect alignment in vertical angles. Given this strong relationship, a mean adjustment using linear regression should be effective for aligning launch angle (vangle) values between the two systems, while speed may benefit from a better adjustment.



I used an adjustment process that involved building separate regression models for each hit type as the batters' Exit Velocity and Launch Angles vary across hit types.This way it ensures that the alignment is customized based on the unique characteristics of ground balls, line drives, fly balls, and popups.

In the modeling approach, I iterated over each unique hit type of the batters which allows for distinct adjustments based on the unique characteristics of each hit type (ground balls, line drives, fly balls, and popups), where exit velocity and launch angle behaviors differ significantly. For each hit type, the data is filtered to create a subset specific to that type which is further divided by batter to capture the individual tendencies of each player within each hit type.

Within each batter-hit type combination, two linear regression models are then applied: one for speed (with speed_B as the independent variable and speed_A as the dependent variable) and one for launch angle (with vangle_B as the independent variable and vangle_A as the dependent variable). This dual-model approach captures the relationship between System B and System A's measurements for each specific batter and hit type, ensuring that adjustment captures both individual batter tendencies and hit-specific characteristics.

The intercepts and coefficients for each batter-hit type model are stored in a dictionary, making them easy to reference. These coefficients are then applied to calibrate each value of Exit Velocity (speed_B) and Launch angle (vangle_B) in the dataset with the following transformations:

Exit Velocity Calibration: calibrated_speed_B = intercept + coefficient * speed_B

Launch Angle Calibration: calibrated_vangle_B = intercept + coefficient * vangle_B

This per-batter hit type approach provides a precise, individualized adjustments that captures each batter's unique tendencies across various hit types, resulting in a good and reliable adjustments across diverse playing styles and contexts.

Interpretations

1. Ground Ball:

   Speed Calibration: I observed a large intercept of (69.37) and a lower coefficient (0.29) which shows that System B underestimates ground ball speeds relative to System A. What this low coefficient implies is that there is limited sensitivity to incresing exit velocity, requiring a significant baseline adjustment to align with System A.

   Vangle Calibration: An intercept of -1.56 and a coefficient near 1 (1.02) suggest only minor adjustments are needed for vertical angle, as System B's measurements are already fairly accurate for ground ball angles.

2. Line Drive:

   Speed Calibration: It has a near-zero intercept of (2.20) and a coefficient close to 1 (1.02), System B's speed measurements for line drives align closely with System A, requiring minimal adjustment.

   Vangle Calibration: Minor adjustments with an intercept of 1.15 and a coefficient of 0.96 indicate that System B captures line drive angles with reasonable accuracy.

3. Fly Ball:

   Speed Calibration: An intercept of 5.65 and a coefficient of 0.98 suggest that System B slightly under-measures fly ball speeds, but only a small adjustment is needed.

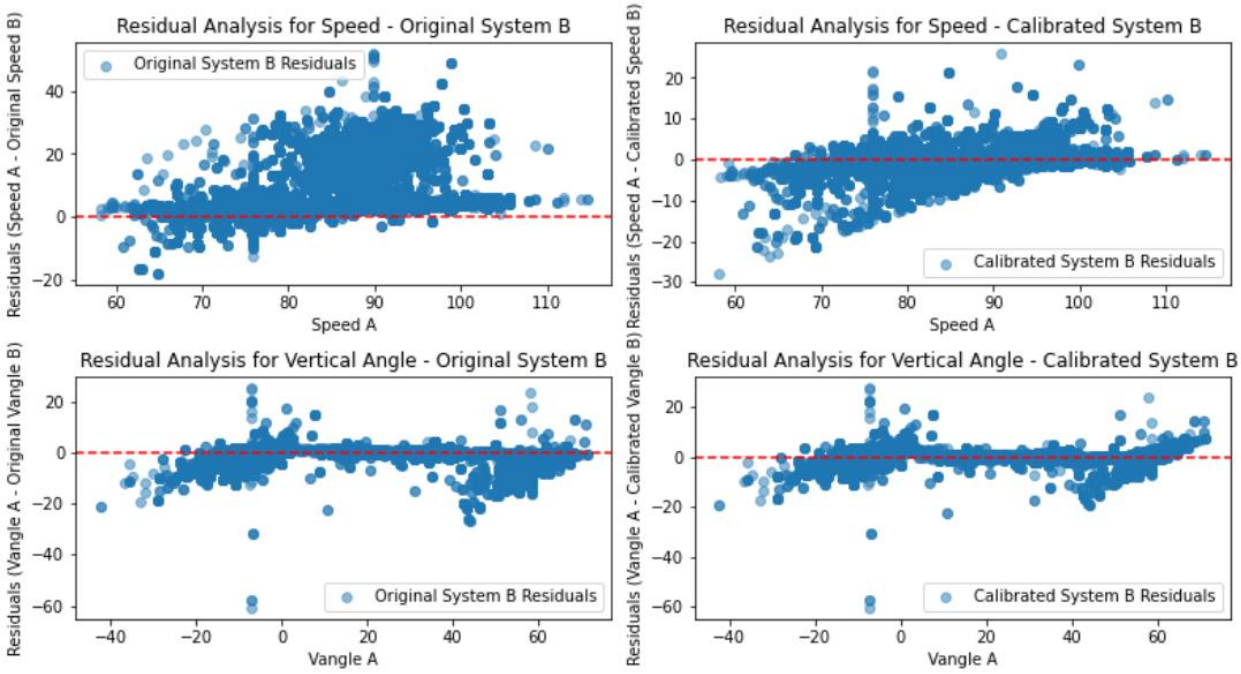
   Vangle Calibration: With an intercept of 0.85 and a coefficient of 0.97, System B's angle measurements for fly balls are well-aligned with System A.

4. Popup:

   Speed Calibration: There is a high intercept (20.87) and a lower coefficient (0.75) which reveals that System B records popups at much lower speeds which necessitate the need for significant adjustment.

   Vangle Calibration: An intercept of 20.28 and a coefficient of 0.61 indicate that System B under-measures popup angles, requiring a notable upward correction.

Residual Analysis of System B Before and After Calibration



The residual analysis compares the differences between **System A** and **System B** before and after calibration, focusing on how well System B's measurements align with System A.

Original System B Residuals (Before Calibration)

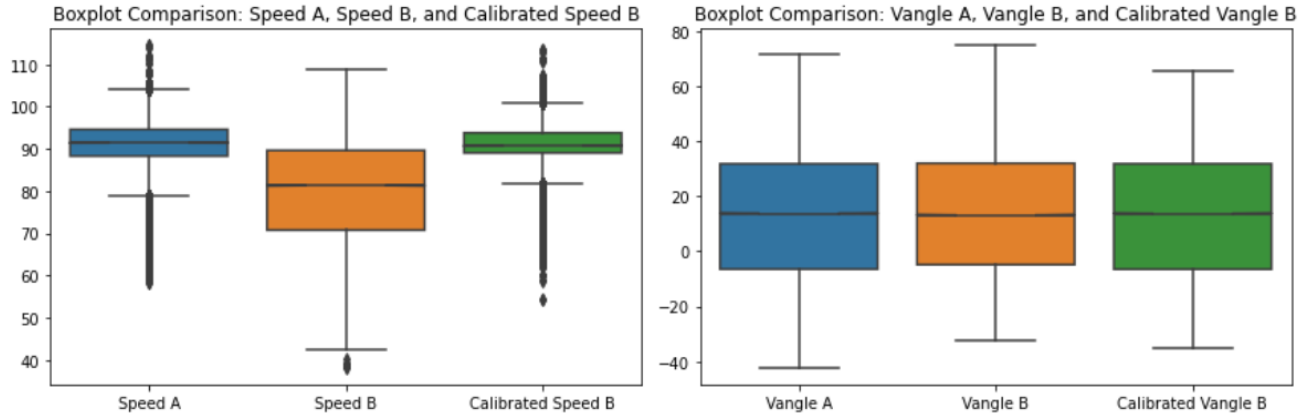
In the initial residual plot, **System B** shows a substantial spread, particularly with residuals skewing. This clustering highlights a consistent underestimation by System B. The wide range of residuals reflects the systematic bias in System B's measurements, with several data points significantly deviating from System A.

Calibrated System B Residuals (After Calibration)

After calibration, the residual plot for **Calibrated System B** shows a marked improvement which has the residuals tightly clustering around the zero line. This tighter alignment suggests that the calibration has effectively reduced the discrepancies, particularly within the central. Although a few outliers persist, the overall reduction in residual spread demonstrates a successful adjustment, bringing System B's measurements closer to System A.

This shows that calibration process has successfully corrected much of the systematic bias in System B. The improved clustering around zero confirms that the calibration has reduced errors across most data points, enhancing alignment with System A, although minor residual discrepancies remain.

Boxplot Comparison of Speed and Vertical Angle Before and After Calibration



The box plots comparison reveals that the median of Calibrated Exit Velocity (Speed B) aligns closely with Exit Velocity (Speed A) unlike the Original Exit Velocity (Speed B), indicating that the calibration successfully adjusted System B's exit velocity measurements to better match System A. While some outliers remain in both Calibrated Speed B and Speed A which was caused as results of the calibration, the central tendency shows strong alignment, suggesting that the calibration effectively reduced the bias in speed.

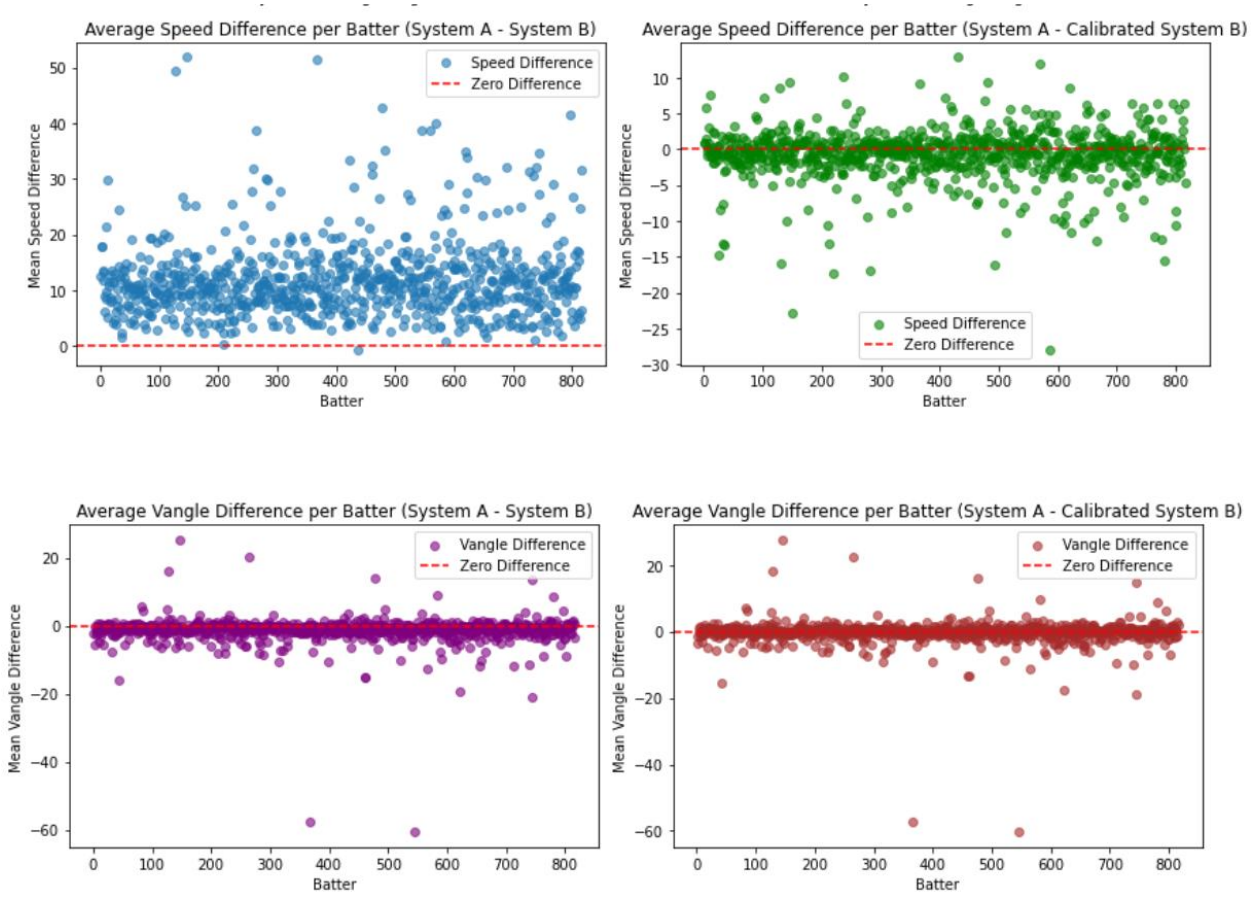
For Launch Angle, the median of Calibrated Launch Angle B is also much closer to Launch Angle A which further tells a successful calibration. The calibrated values demonstrate improved consistency with System A's measurements, validating the effectiveness of the adjustments.

Calibration Impact on Speed and Vertical Angle Measurements: Analysis of Deviation

The calibration process was evaluated using two types of plots to assess its effectiveness in aligning System B measurements with System A. These plots include:

1. batter-specific average deviations and

2. overall distribution of deviations for speed and vertical angle



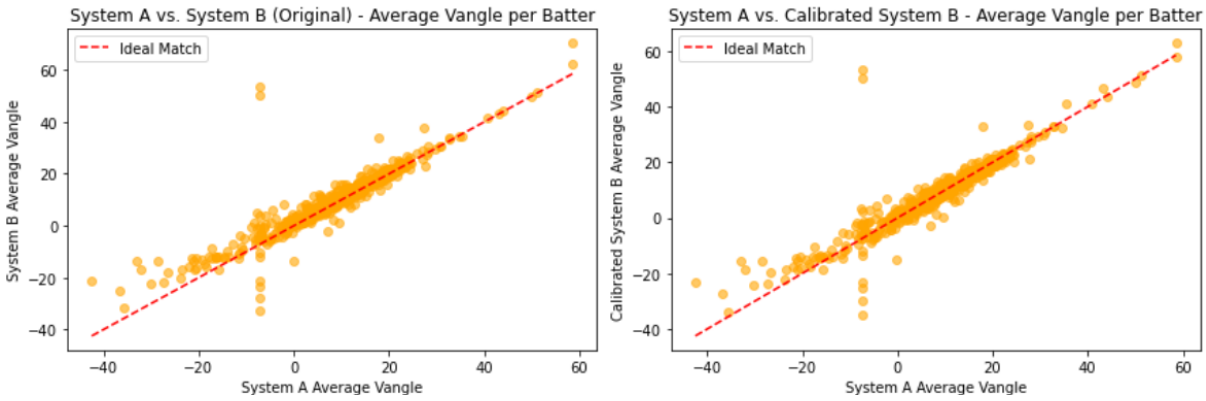
Exit Velocity Deviation per Batter:

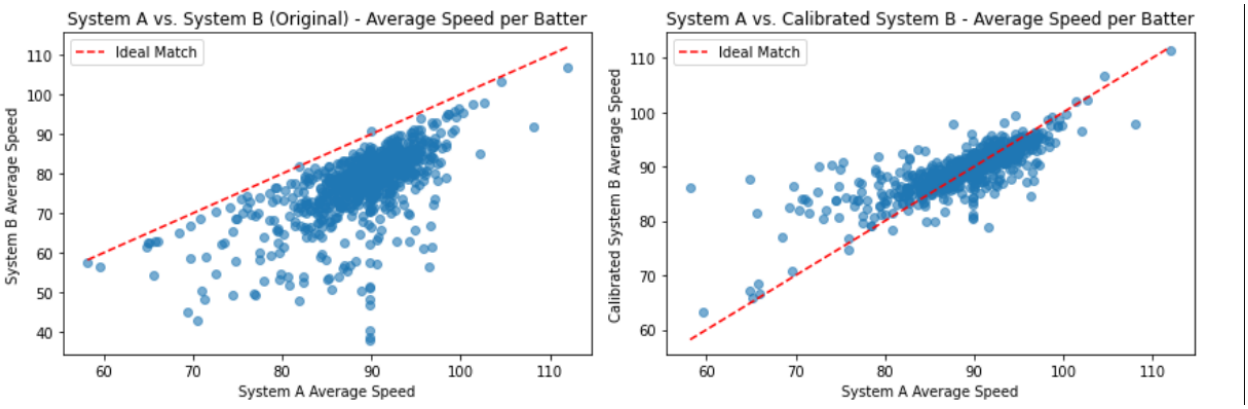
The scatter plots compare the average speed deviations per batter (System A - System B) before and after calibration which therefore reveal a significant improvement. In the pre-calibration plot, deviations are widely scattered around the zero line, showing inconsistent speed measurements in System B relative to System A across different batters. After calibration, it was evident that there are deviations which were tightly clustered around zero for most batters which indicates that calibration has successfully aligned System B's speed measurements with System A's standards. Nonetheless, certain batters still show noticeable residual differences, suggesting areas where additional fine-tuning could further reduce individual variations.

Launch Angle Deviation per Batter:

For the launch angle, the batter specific deviation plots exhibit similar improvements in post-calibration**.**

Comparison of Average Speed per Batter: System A vs. System B (Original and Calibrated)





The two scatter plots compare average speed values recorded by System B (both original and calibrated) against System A for each batter, with the red dashed line representing an ideal match (y = x), where System B's measurements would align perfectly with System A.

System A vs. Original System B (Left Plot):

In the original (uncalibrated) data, a broad scatter below the ideal match line was observed which indicates that System B consistently recorded lower average speeds compared to System A across batters, with numerous points significantly deviating from the line. The widespread suggests systematic underestimation by System B and highlights the variability in measurement misalignment.

System A vs. Calibrated System B (Right Plot):

After calibration, the plot also shows a much tighter clustering around the ideal match line showing that calibration effectively corrected the average speed discrepancies, bringing System B's measurements in closer alignment with System A. The reduced spread around the line shows that

most batters' average speeds in System B are now comparable to System A's values, with only minor deviations for some batters. This shows the success of the calibration.

Evaluating Calibration Accuracy Through Error Metrics

To assess the effectiveness of the calibration, we performed error metrics calculations by comparing System B's data against System A's values for both exit velocity and launch angle. By measuring Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) before and after calibration, we quantified the alignment accuracy.

These results show a significant reduction in both MAE and RMSE post-calibration, particularly for Exit Velocity which indicates that the calibration model effectively minimized discrepancies and also enhanced the precision of System B's measurements. This improvement suggests that the calibrated System B values now more reliably represent true measurements, as defined by System A.

Error Metrics Before and After Calibration:

	MAE	RMSE
Pre-Calibration Speed	10.734128	13.657307
Post-Calibration Speed	1.577954	2.542065
Pre-Calibration Vangle	1.370506	2.292017
Post-Calibration Vangle	0.895780	1.634807

After verifying calibration accuracy through the alignment checks, I conducted a spread comparison to ensure that the variability (measured by standard deviation) of calibrated System B exit velocity aligns well with System A across different hit types. This additional check confirms how well the calibration preserved the natural spread of values, crucial for accurately reflecting each hit type's behavior.

For Fly Balls, the standard deviations are very close; System A has a std of 3.86, and calibrated System B has 3.79 which indicates that the calibration successfully maintained the natural variability of exit velocity thereby aligning closely with System A. This suggests that fly ball exit velocity in calibrated System B reliably reflects the original range observed in System A.

For Ground Balls, however, there's a noticeable difference: System A's std is 3.80, while calibrated System B shows a reduced std of 1.55. This substantial narrowing suggests the calibration might have over-corrected which might constrain the natural variability for ground balls in System B, which may not fully capture System A's broader range.

Line Drives show a well-aligned calibration which is closely matched with a standard deviation (System A std of 4.11 and calibrated System B at 4.04), which means that line drive speeds in System B maintain similar variability to System A, preserving their natural spread effectively.

Popups present a modest reduction in variability (System A std of 5.71 vs. calibrated System B at 4.17), suggesting a slight narrowing in calibrated System B's spread. While the reduction isn't as pronounced as with ground balls, it does indicate that popups in System B might be less variable than in System A although still reflective of the general exit velocity patterns.

Overall, this analysis shows effective calibration for Fly Balls and Line Drives, as the standard deviations in calibrated System B closely match those of System A. But the calibration for both Ground Balls and Popups seems to reduce their natural spread with the potential of limiting System B's ability to fully capture the range of speeds observed in System A. While preserving natural variability is essential, I'll continue with the current calibration model without further adjustments, acknowledging that future adjustment may be needed to better represent System A's spread for these hit types

WEIGHTED CALIBRATION TO INTEGRATE SYSTEM A AND CALIBRATED SYSTEM B FOR ENHANCED METRIC RELIABILITY

Since I know that preserving the natural variability of hit types is very crucial, particularly for ground balls and popups. I proceeded with a weighted calibration approach to integrate System A and calibrated System B metrics to form a single metric. This decision aims to leverage the strengths of both systems while optimizing for accuracy. This method of applying the weights based on RMSE values is an indicator of each system's reliability such that the calibration strategy effectively combines System A's stable representation with the enhanced precision of calibrated System B, particularly for exit velocity measurements.

To determine each system's contribution to the final metric, I calculated weights using the RMSE values as indicators of accuracy. Specifically, I extracted the RMSE values for exit velocity and launch angle (vangle) from both systems, pre-calibration for System A and post-calibration for System B.

The weights were then calculated inversely to RMSE such that the system with a lower RMSE received a higher weight, indicating greater reliability. For exit velocity, calibrated System B received a higher weight (0.843) due to its lower RMSE, while System A contributed less (0.157), reflecting its relatively larger error. Similarly, for launch angle, calibrated System B still held a higher weight (0.592) than System A (0.408), although both systems contributed significantly.

All in all, using these weights, I applied a weighted average to combine the exit velocity and launch angle metrics from System A and calibrated System B, producing a projected combined exit velocity and combined launch angle. This way I effectively balances the accuracy of System A and calibrated System B, optimizing the metric's reliability by enhancing alignment with observed batted ball behavior forming a single projected combine metrics as seen below.

Weights Calculation for Exit Velocity:

$$Weight\_(speedA) = RMSE\_(speedB) / (RMSE\_(speedA) + RMSE\_(speedB))$$

$$Weight\_(speedB) = RMSE\_(speedA) / (RMSE\_(speedA) + RMSE\_(speedB))$$

Projected Combined Metrics

The projected combined metrics are calculated using these weights to blend the values from System A and calibrated System B:

$$ProjectedCombinedSpeed = (speed_A \times Weight_{speed\_A}) + (calibratedspeed_B \times Weight_{speed\_B})$$

$$ProjectedCombinedVangle = (vangle_A \times Weight_{vangle\_A}) + (calibratedvangle_B \times Weight_{vangle\_B})$$

These formulas provide a weighted average of each system's values, creating a unified metric that leverages the strengths of both System A and calibrated System B.

	speed_A	calibrated_speed_B	projected_combined_speed	vangle_A	calibrated_vangle_B	projected_combined_vangle
0	95.668364	88.838321	89.910116	-12.841748	-10.133916	-11.261235
2	95.328485	94.996419	95.048528	16.732373	16.814607	16.780372
3	87.742940	87.496935	87.535539	33.475045	33.644328	33.573853
4	114.583564	113.541337	113.704887	21.329170	22.055295	21.752996
5	92.172531	89.812619	90.182945	-6.976002	-7.262919	-7.143470

Analysis and Visualization of Error Metrics for Combined Metric Projections vs. System A

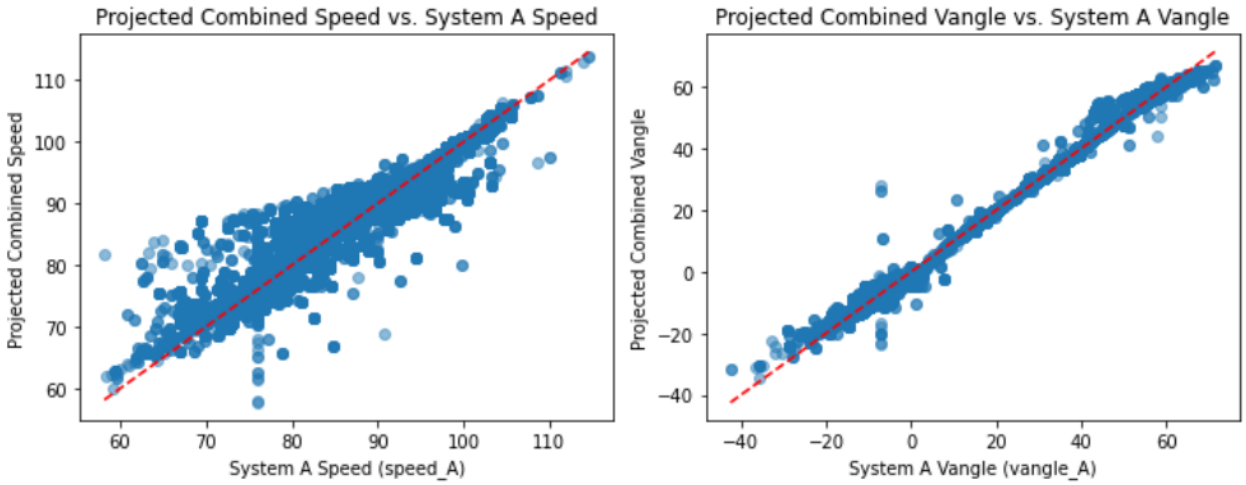
Then I assessed the accuracy of the projected combined single metrics called projected_combined_speed and projected_combined_vangle by calculating the performance metrics (MAE and RMSE) to measure alignment with System A's values. These error metrics help me in evaluating the accuracy of the combined projections gotten from exit velocity A and the Calibrated exit velocity B.

Error Metrics Results:

- **Combined Exit Velocity:** MAE = 1.33, RMSE = 2.14

- **Combined Launch Angle:** MAE = 0.53, RMSE = 0.94

The low error values indicate that the projected combined metrics closely match the actual System A values, demonstrating effective calibration and combination for both exit velocity and launch angle.



"TRUE" AVERAGE SPEED-OFF-BAT PROJECTION FOR NEXT SEASON

To create accurate speed-off-bat projections for next season, I implemented a Per-Batter Hit Type Data Averages approach. This approach involved grouping the data by both batter and hit type (e.g., ground balls, line drives, fly balls, popups) and calculating averages within each group. This way it ensures that the projections capture the unique performance patterns of individual batters and the characteristics of each hit type

How the Projection Was Calculated:

Data Grouping: I grouped the dataset by each batter and hit type, calculating the average speed-off-bat using the newly created (projected combined speed) for each combination and this grouping captures both the batter's individual tendencies and the unique exit velocity characteristics of different hit types.

Calculation of "True Speed" Benchmarks: After combining each of the batter hit type combination, I calculated the average of the combined exit velocity values, treating this as the 'True Speed.' This approach assumes that batters maintain a certain level of consistency across seasons, which allows these averaged values to serve as a stable performance baseline.

Use of Hit Type-Specific Grouping: The averages of each hit type were determined separately in such a way that the model considers the natural variability in speeds for different hit types. For example, ground balls generally have lower speeds, while fly balls and line drives show more variation. The batted ball grouping by hit type ensures that the projections align with real-life patterns thereby making the projections more accurate.

Reliability Through Aggregation: Grouping by batter and hit type smooths out random noise, reducing the impact of outliers. I also calculated the count of data points for each projection, using

higher counts as indicators of reliability. Batters with higher data counts have more stable projections, while lower counts suggest the need for careful interpretation.

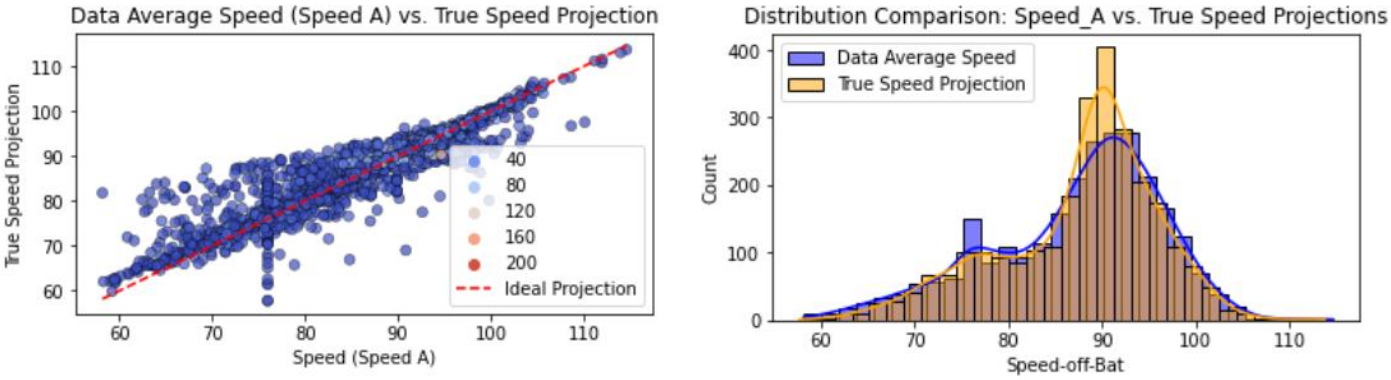
Sample heading of the True Exit Velocity below:

Batter Hit Type-Specific Projections with Initial Speed A (Sample):

	batter	hittype	speed_A	True_speed	data_count
0	1	fly_ball	87.734149	87.568091	18
1	1	ground_ball	91.639142	90.313399	40
2	1	line_drive	93.504552	93.574648	18
3	1	popup	82.783712	81.262984	3
4	2	fly_ball	79.907608	80.282937	2

After this I converted the True Exit Velocity with the batter's data into a data frame and save it as csv file for easy referencing for next season projection.

Visualization of Exit Velocity (Speed_A) Data Averages vs. Projections (True Speed)



To evaluate how well the projected True Speed aligns with Exit Velocity – Speed A (System A's speed-off-bat measurements), I plotted two visual comparisons: a distribution overlay and a scatter plot.

Distribution Comparison:

The overlay of Exit Velocity (Speed_A) and the (True Speed) shows a close alignment in which both distributions peak in similar regions. This indicates that the True Speed projection captures the typical range and central tendencies of Exit Velocity (Speed_A) well.

There is, however, some minor divergence, which could account for variations in individual batter performances (improved or under-performance) and hit types.

Scatter Plot Comparison:

The projected True Speed values is seen to have align with Exit Velocity (Speed_A) along the ideal projection line (dashed red line) in the scatter plot, which suggests that the projections accurately reflect each batter's performance tendencies.

Most points appear to fall near this ideal line therefore demonstrating a reliable calibration, though there are some deviations that may indicate larger discrepancies for specific batters or hit types.

Evaluation of Projection Accuracy

To measure the alignment, I first calculated two error metrics:

Mean Absolute Error (MAE): 1.9094 which is low value which suggests that, on average, the projections are very close to Exit Velocity (Speed_A), affirming the accuracy of the projection model in capturing typical batter performance.

Root Mean Squared Error (RMSE): 3.3707, the slightly higher RMSE indicates that, while most projections are close, there are occasional larger discrepancies. These discrepancies likely arise from certain batters or hit types that show higher variability in Exit Velocity, which can account for batter either getting better or less good in performance for the coming season.

These findings indicate that the projection approach is effective, accurately reflecting the natural performance patterns captured in Exit Velocity (Speed_A).

CONCLUSION

The projection True speed looks effective with each batter's historical performance (speed_A) using the scatter plot alignment and low deviations that capture individual tendencies effectively. The stability in absolute and squared error distributions confirms that most projections deviate only slightly from historical averages, ensuring consistency. Yes, the True speed has few larger deviations present in it which shows that there are different variations in batter performance which highlight the model's reliability in capturing batter-specific trends while providing stable, adaptable projections.

Curious About Batter that Improved or remain steady or declined slightly in Performance?

Categorizing Batter's Performance: I further analyzed batter performance by grouping the batters into three performance groups—Above Performance, Maintain Performance, and Underperform, by comparing each batter's projected speed-off-bat ('True Speed') to their historical Exit Velocity average speed (speed_A) for each hit type. A threshold of ±1.0 unit was used: if the projected speed was more than 1.0 unit above their historical average, the batter was classified as **Above Performance**; if within the threshold, they were grouped as **Maintain Performance**; and if below the threshold, as **Underperform**.

The results show the top batter in each performance category for each hit type (e.g., fly ball, ground ball). The **Above Performance** batters showed improvements over their Exit Velocity (speed_A) averages, the **Maintain Performance** batters remained consistent with their past performance, and the **Underperform** batters exhibited slight declines. This way I was able to effectively capture individual batting trends within specific hit types, providing insights into batters who have excelled, maintained stability, or underperformed**.**

The batters with Above Performance (good performance) in each hit type, based on the previous results, are as follows:

Fly Ball: Batter 737

True Speed: 106.45

Historical Exit Velocity (speed_A): 104.53

Performance Category: Above Performance

Line Drive: Batter 380

True Speed: 105.54

Historical Exit Velocity (speed_A): 104.29

Performance Category: Above Performance

Ground Ball: Batter 438

True Speed: 97.89

Historical Exit Velocity (speed_A): 96.70

Performance Category: Above Performance

Popup: Batter 765

True Speed: 85.84

Historical Exit Velocity (speed_A): 75.97

Performance Category: Above Performance