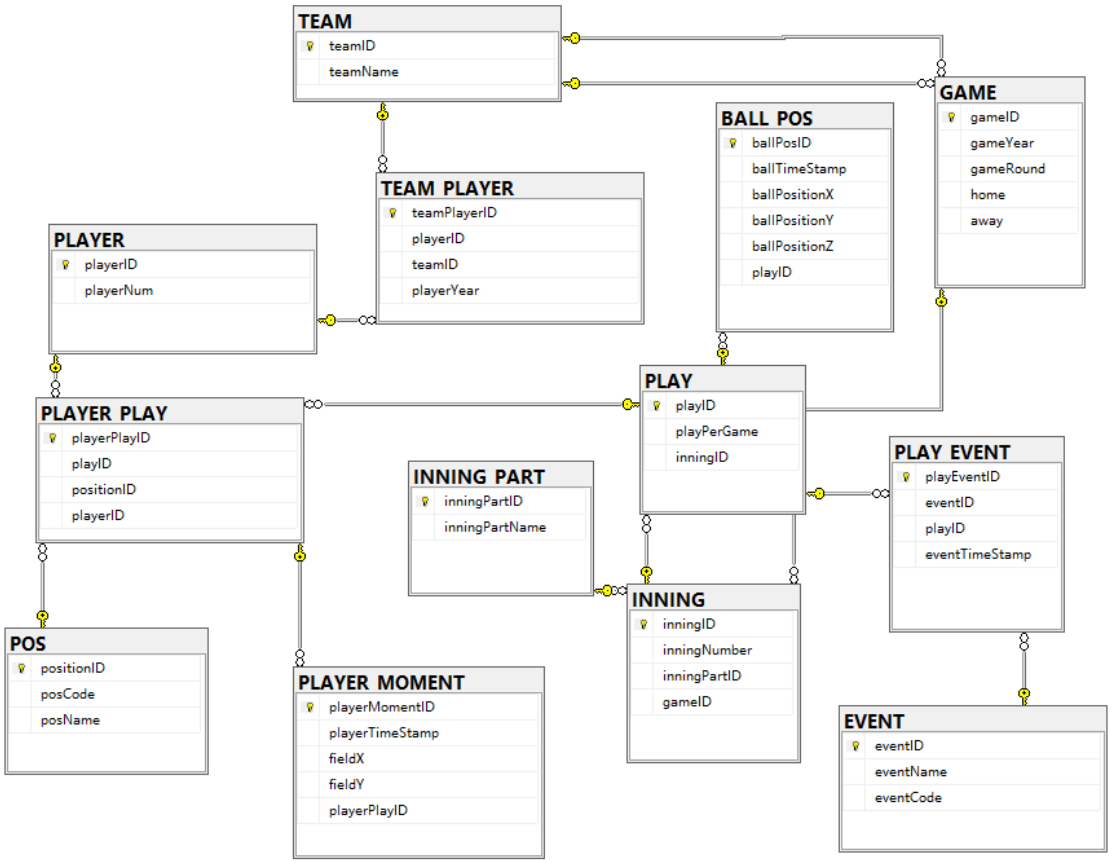
Abstract

This paper delves into an in-depth analysis of the batter's behavior in the context of baseball, focusing on key metrics and dynamics that characterize the trajectory of the ball during play events. The study employs rigorous methodologies to compute metrics such as flying distance, straight-line distance, route efficiency, ball travel speed, ball direction, and the likelihood of a batter hitting on base. By leveraging a comprehensive dataset and advanced statistical techniques, the paper seeks to unravel the intricate spatial attributes and performance correlations that influence a batter's success in the game.

Introduction

Baseball is a very popular national sport in the U.S., and its seemingly simple rules often have a mystery behind them. Behind a simple hit to get on base is a game between the entire offensive team and the defensive team. The role of the Batter in that round of play is crucial throughout the game. In this article, we will focus on the role of Batter, and how the various factors related to him, such as the direction of the bat, the speed of the bat, etc., affect the outcome of the game.

Data



In order to better deal with our data, we design a whole new relational database to help us build the csv we want.

To sustain the modularity of the implementation process, we splitted our SQL queries into 5 stages: schema initialization, get-ID procedure creation, single-row insert procedure creation, raw data bulk-insert process, and wrapper-insert transaction.
Schema initialization includes classifying variables into entities, recognizing one-many relationships among entities, creating tables for entities, and declaring key references. Through careful inspection of the raw datasets, we designed multiple combinations of variables that each could define the uniqueness of an entity (Appendix D).

Measuring the extensive utility of extracting the primary keys, we implemented a Get-ID stored procedure for each entity so that the repetitive process of extraction can be supplanted by a few parameters. Such modularization was also applied to INSERT procedures with Get-ID procedures nested in, where executing with a few parameters could complete one row of insertion (Appendix E).

When it comes to bulk-inserting the raw-data, the structure of the raw csv did not fit into our schema because it columnizes the “player position” data. To address that, we used Python Pandas to convert all position columns into the value series of one single column from game_info.csv file, and each player-position pair is classified into separate rows so that the many-to-many relationship between these two entities can be easily normalized into the database (Appendix F).

Our final step of normalization was to design procedures that each include a massive population process, embedded with the single-row inserting procedures, from its corresponding raw data entity, matching the foreign keys’ conditions based on the other variables from the entity (Appendix G). Such a process repetitively executes the insert procedure for each row of the raw datasets, allowing us more than 10,000,000 data strips into the database in less than an hour (Appendix H).

Methodology

In this section, we describe the detailed process used to calculate essential metrics related to ball trajectory in the dataset. These metrics include factors like flight distance, direct distance, efficiency of the route taken, travel speed, ball direction. To establish connections between these metrics, we formulated specific methodologies. These methods enable us to quantify the ball's path within the dataset and assess the batter's success in reaching a base.

1. **Ball Flying Distance:** We determine the flying distance by measuring the distance the ball travels from the moment the batter hits it until it's caught by a player or results in a home run in each playID. We implement the calculation of flying distance of the three-dimensional Euclidean distance formula: travel_distance = sqrt((x2-x1)^2+(y2-y1)^2+(z2-z1)^2). A loop iterates over each play event, and the relevant positions and timestamps are extracted for calculation. (Appendix C)
2. **Straight Line Distance:** To determine the aerial distance covered by each ball, we computed the spatial gap between the initial impact point and the final point. The

trajectory starts the moment the batter connects with the ball, marked by event code 4. The culmination point is indicated by event codes 2 or 11, reflecting ball retrieval or a home run. This framework utilizes coordinates at these moments to compute aerial distance via: d = √[(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2]. (Appendix B)

3. **Route Efficiency:** Route Efficiency = Straight Line Distance / Ball Flying Distance. This metric effectively captures the effectiveness of the trajectory undertaken by the ball in reaching its ultimate destination. (Appendix B)
4. **Ball Average Travel Speed (feet/second):** The travel speed of the ball would be helpful for us to comprehend the swiftness of its movement. It is computed by analyzing the time intervals between consecutive positions and the associated distances covered from the ball getting hit to the ball acquired or homerun (Ball Flying Distance/Total Timestamp) in each play event. (Appendix C)
    a. In order to explore more possibilities on ball travel speed, we also include the initial acceleration of the ball when the batter hits the ball. By calculating the first two initial speed and timestamp, we use calculation formula: a = Δv/Δt (Appendix C)
5. **Ball Direction:**For ball direction, by calculating the x,y from the first two points after the ball is hitted by the batter, we can get the ball's direction. (Appendix A)
6. **Batters go through bases**: We go through all the games and plays, through tracking playerNum in the next playPerGame, we can find out whether the batter reaches a base and which base he reaches. (Appendix A)
7. **Regression Model:** Logistic regression allows us to model binary outcomes, which aligns perfectly with our interest in understanding the factors that affect the probability of a batter getting on base or not. By employing logistic regression, we can quantitatively assess the impact of different predictor variables on the probability of success, thus providing valuable insights into the intricate dynamics of baseball performance.

Discussion

Ball Direction

In fact when we are calculating the angle we may be ignoring the effect of the wind speed on the direction of the ball's flight because we are taking the initial two points after the ball has been struck in our calculations. This may have an effect on the analysis (when the ball is hit at an angle of around 90, the wind may cause its direction to shift from the first quadrant to the second and vice versa), but given the sheer size of the population, this is not so much of a factor as to have any decisive effect.

Route Efficiency

1. To ensure the precision of our analysis, meticulous attention was devoted to exclusively considering data depicting instances where the ball was struck by the batter. Our concentration remained steadfastly directed toward elucidating the interconnection between the route efficiency of balls and the batters' success in reaching base.

2. The outcome variable encompasses distinct categories signifying the base onto which the batter successfully advances. In order to holistically assess the potentiality of batters reaching base, an auxiliary binary variable termed "onBase" was conceived. This auxiliary variable is assigned the value of 1 when the outcome variable is equal to or greater than 1, and is assigned 0 otherwise.

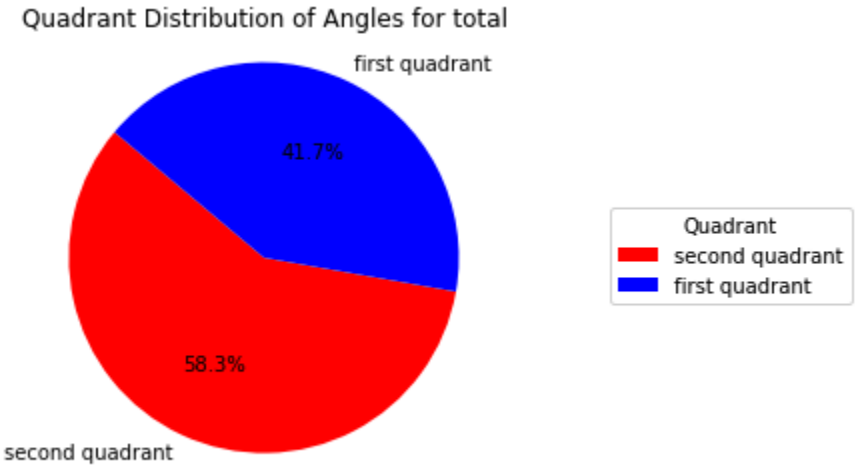
Ball Flying Distance

We organize the ball's positional information based on event codes during each play event. We determine the beginning of ball position data using eventCode = 4, and its conclusion is marked by either eventCode = 2 or 11. We opt to begin with eventCode 4 because we aim to explore how the batter's performance relates to the ball being hit. However, our choice of where the ball's data ends has limitations concerning various ball landing scenarios with differing endpoints. To reinforce our assumptions about the connection between ball flying distance and batter performance, we only consider situations where the ball is acquired by a player or results in a home run. These situations provide clearer evaluations of the landing point and total travel distance of the ball.

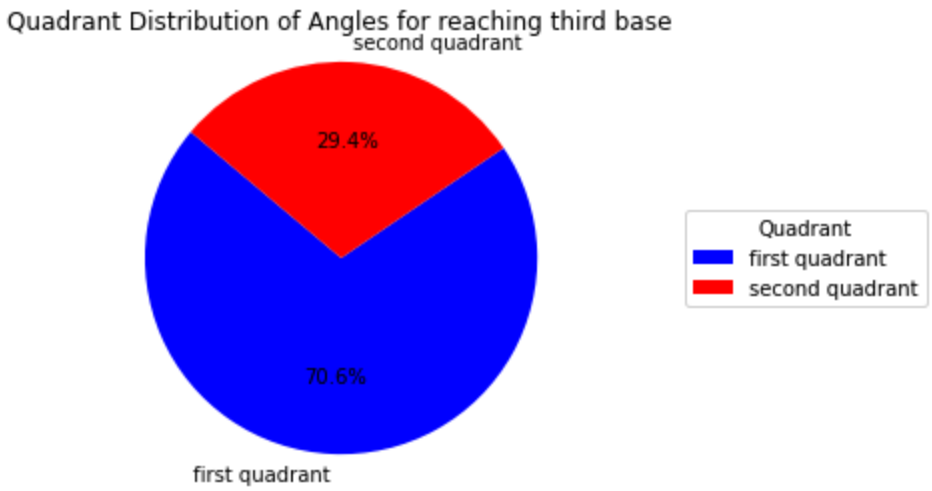
Conclusion

Ball Direction

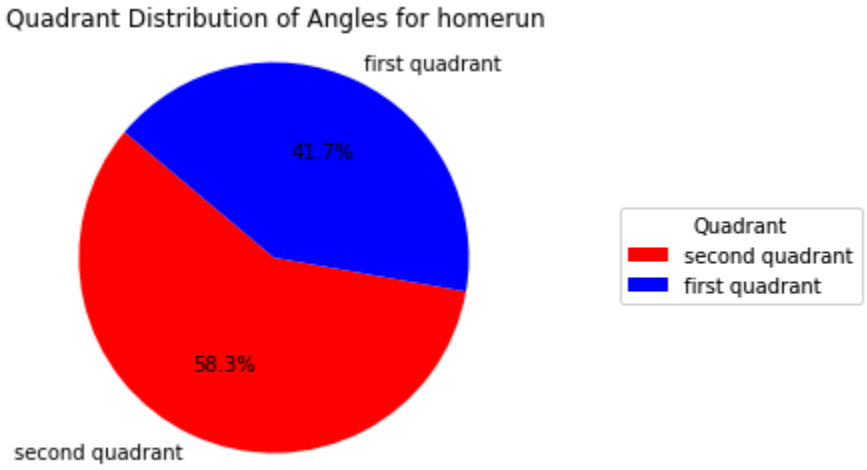
In total, if the batter hits the ball to the second quadrant, then he has a better chance to reach any base in this round (play) than the first quadrant.

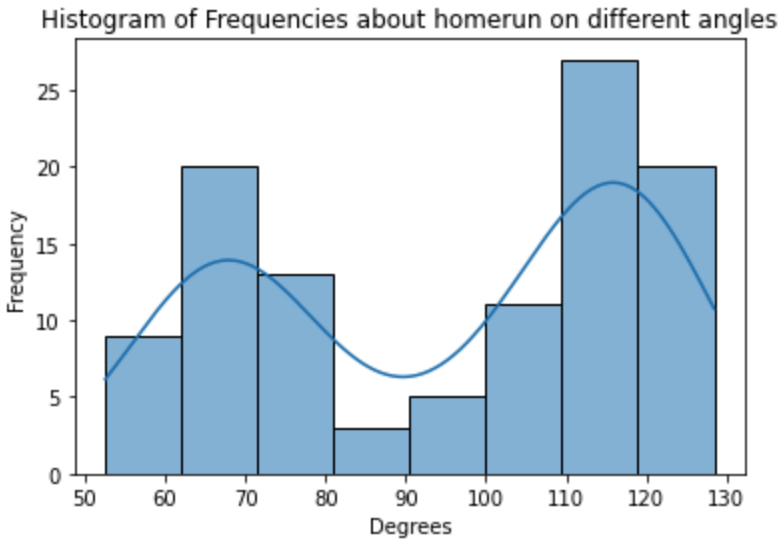


However, for the case when the batter runs directly to third base. Hitting to the first quadrant will give nearly three times more chances than hitting to the second quadrant.



This is most likely due to the fact that after flying a longer distance out towards the first quadrant, it takes longer to be shaped back to the first or second base defender, giving them no chance to block or touch Batter, who then has a chance to get to third base without incident. As for the home run, which is the main concern of the audience and the coaching staff, the data shows that hitting the ball towards the second phenomenon has a greater possibility of accomplishing a home run. Angle-wise, the angle is centered around 110 degrees or 70 degrees.





All the graphs above about ball direction are screenshot from Ball_direction_analysis.ipynb

By controlling the direction in which the ball is hit, better teamwork can be done. For example, if the current batter needs to advance to third base to ensure that the current third base attacker scores successfully, he should hit the ball towards the first quadrant. Or if the current defense is making it difficult to get to the other bases and is only able to get to first base, he should hit the ball in the second quadrant.

Route Efficiency

After processing and analyzing our data, it turns out that there is a positive correlation between the ball's route efficiency and the possibility of the batter getting on base. According to our logistic regression model, the coefficient is 2.19 that indicates the increase in route efficiency would lead to a higher likelihood of batter getting on base. However, there are some concerns about our findings. The Pseudo R-square is extremely small while the p-value is approximately 0.(Fig 1.0) A low Pseudo R-squared suggests that our model might not capture a large portion of the variability in the outcome. This could be very likely due to the presence of unobserved factors, since we only have one factor in the model.

Logit Regression Results

Dep. Variable:	onBase	No. Observations:	4488
Model:	Logit	Df Residuals:	4486
Method:	MLE	Df Model:	1
Date:	Thu, 31 Aug 2023	Pseudo R-squ.:	0.01280
Time:	21:49:47	Log-Likelihood:	-2072.2
converged:	True	LL-Null:	-2099.1
Covariance Type:	nonrobust	LLR p-value:	2.301e-13

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.5186	0.310	-11.339	0.000	-4.127	-2.910
routeEfficiency	2.1927	0.335	6.552	0.000	1.537	2.849

Figure 1.0

Looking at the scatter plot, we don't see a very obvious pattern. However, almost every point whoes result isn't 0 has route efficiency larger than 0.2. (Fig 1.1)

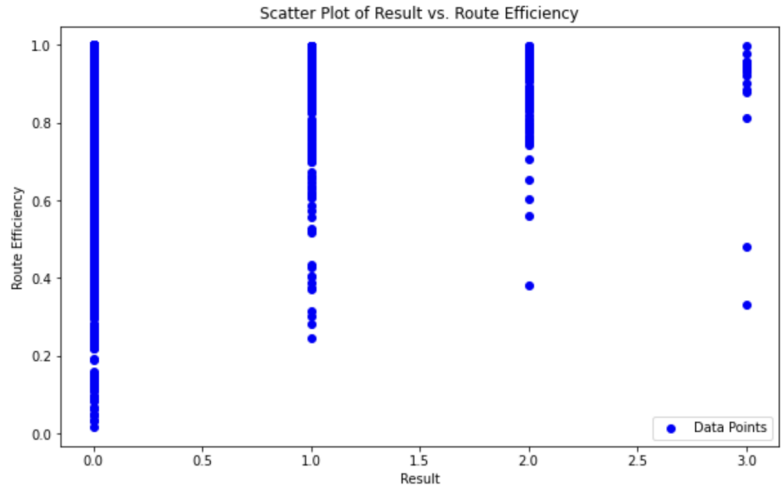


Figure 1.1

Adding another factor(average ball speed) into our logistic regression model, the Pseudo R-square becomes a little bit higher but still considerably small. (Fig 1.2)

Logit Regression Results

Dep. Variable:	onBase		No. Observations:		4488	
Model:	Logit		Df Residuals:		4485	
Method:	MLE		Df Model:		2	
Date:	Thu, 31 Aug 2023		Pseudo R-squ.:		0.03007	
Time:	21:49:47		Log-Likelihood:		-2036.0	
converged:	True		LL-Null:		-2099.1	
Covariance Type:	nonrobust		LLR p-value:		3.848e-28	
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.0505	0.332	-9.201	0.000	-3.700	-2.401
routeEfficiency	2.9348	0.366	8.013	0.000	2.217	3.653
averageBallSpeed	-0.0163	0.002	-8.474	0.000	-0.020	-0.013

Figure 1.2

Ball Travel Speed

Correlation Analysis

Dep. Variable:	result	**No. Observations:**	4488
Model:	Logit	**Df Residuals:**	4484
Method:	MLE	**Df Model:**	3
Date:	Fri, 01 Sep 2023	**Pseudo R-squ.:**	0.003772
Time:	02:50:08	**Log-Likelihood:**	-2091.2
converged:	True	**LL-Null:**	-2099.1
Covariance Type:	nonrobust	**LLR p-value:**	0.001226

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.9335	0.113	-17.154	0.000	-2.154	-1.713
TotalDistance	0.0010	0.000	3.132	0.002	0.000	0.002
AverageBallSpeed	-0.0002	0.002	-0.093	0.926	-0.004	0.004
InitialBallAcceleration	6.104e-05	8.75e-05	0.698	0.485	-0.000	0.000

Figure 2.0

Dep. Variable:	result	**No. Observations:**	4488
Model:	Logit	**Df Residuals:**	4486
Method:	MLE	**Df Model:**	1
Date:	Fri, 01 Sep 2023	**Pseudo R-squ.:**	0.001372
Time:	02:51:13	**Log-Likelihood:**	-2096.2
converged:	True	**LL-Null:**	-2099.1
Covariance Type:	nonrobust	**LLR p-value:**	0.01641

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.7275	0.091	-19.015	0.000	-1.906	-1.549
AverageBallSpeed	0.0014	0.001	2.408	0.016	0.000	0.003

Figure 2.1

By using a logistic regression model, we can conclude that, with a single feature of average ball speed, there is relatively low positive correlation between ball speed and the batter hitting base (Fig 2.1). However, as long as we include more features in our analysis, including total distance, average ball speed, and initial ball acceleration, only the total travel distance and ball acceleration has relatively low positive correlation with batter hitting base (Fig 2.0).

Binary Classification Models

```
Logistic Regression Results:
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90       733
         1.0       0.00      0.00      0.00       165

    accuracy                           0.82       898
   macro avg       0.41      0.50      0.45       898
weighted avg       0.67      0.82      0.73       898

Random Forest Results:
              precision    recall  f1-score   support

         0.0       0.82      0.93      0.87       733
         1.0       0.26      0.10      0.15       165

    accuracy                           0.78       898
   macro avg       0.54      0.52      0.51       898
weighted avg       0.72      0.78      0.74       898
```

Figure 2.3

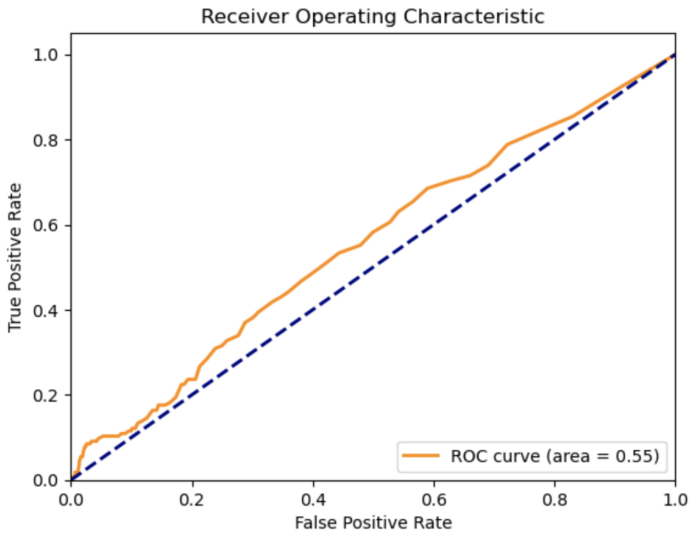


Figure 2.4

For the overall performance by using two models above, we can conclude the following conclusions (Fig 2.3).

1. Logistic Regression: The model has high precision and recall for class 0, indicating that it is effective at identifying instances of class 0. However, its performance for class 1 is poor, as indicated by the low precision, recall, and F1-score. The model has a higher overall accuracy of 0.82.
2. Random Forest: Similar to the logistic regression model, the random forest model also performs well for class 0 with relatively high precision and recall. However, it performs even worse for class 1, with a low F1-score, precision, and recall. The random forest model's overall accuracy is 0.78.

For both models, the macro average F1-score is around 0.45 to 0.51. This indicates that the models struggle to perform well on both classes simultaneously. It's important to consider the context of your problem and the trade-off between false positives and false negatives when interpreting these results.

An ROC curve area (AUC) of 0.55, indicates a model with slightly better performance than random guessing, where the True Positive Rate and False Positive Rate are balanced, but still relatively low discriminatory power (Fig 2.4).

Appendix

We reference all the code in our repo here.

A. File: /Ball_Direction_Analysis/Ball_direction_analysis.ipynb
B. File: /RouteEfficiencyAnalysis/Route_efficiency_analysis.ipynb
C. File: /Ball_Speed_Analysis/ball_pos_total_distance.ipynb
D. File: /Database/Database Implementation/CreateSchemaQuery.sql
E. File: /Database/Database Implementation/GetIDProcedureQuery.sql
F. File: /Database/Database Implementation/RawDataInsertQuery.sql
G. File: /Database/Database Implementation/InsertProcedureQuery.sql
H. File: /Database/Database Implementation/PopulateTableQuery.sql