San Francisco Giants Batting Performance Analysis

The GMs

Amie Touray, Pavan Ketankumar Patel, Yangyang Li, Yen Chun Lin

American University

ITEC 620 Business Insights/Analytics

Professor Shawn Janzen

Teaching Assistant Kritika Goyal

April 28th, 2024

Section I: Report Summary

The application of data analysis in sports offers significant benefits for athletes aiming to optimize their performance. The analysis yields a range of statistical insights, including the visualization of player performance and the identification of standout performers. By leveraging these analysis results, athletes can identify the most effective training methods to improve their performance. Coaches can develop better game plans for the team to achieve a higher winning percentage in a season, and team managers can rely on this information to identify and target suitable players to enhance team composition and strategic competitiveness.

This report presents the analysis of the San Francisco Giants' batting performance, utilizing data spanning from 1958 to 2023. Our project employed data analysis techniques, including clustering and linear regression, to explore and enhance the team's performance. Through clustering, we categorized players based on their batting abilities, revealing specific strengths and weaknesses. We cluster players' batting skills in two different ways: one by players and another by their defensive position. Our linear regression analysis investigated the relationship between various batting metrics, such as how hits (H) and home runs (HR) influence runs batted in (RBI), to propose effective game strategies for higher team scores. The findings suggest targeted improvements for players, strategic game planning for coaches, and informed decision-making for team managers regarding player recruitment and trading. These strategies are aimed at enhancing the team's competitiveness and achieving a higher season win rate.

Section II: Project Introduction

Our project focuses on sports analytics in Major League Baseball, specifically analyzing the San Francisco Giants' batting performance. This analysis will examine the team's batting performance. Based on the analysis, the manager could identify strategies for future trades or drafts and seek good players in free agency, while coaches can devise better game plans for the games. Additionally, it could serve as a valuable reference for both players and coaches to visualize the strengths and weaknesses of each player and develop tailored training programs for improvement.

1. **Problem Statement**:

The common goal for a baseball team is to build the strongest team possible to win the league title. There are several methods that a baseball team can use to improve its performance. For example, players can focus on training to enhance their performance at the plate, coaches can devise better game plans to help the team score more runs in a game, and general managers can engage in player trading, select promising rookies in the draft, or recruit players from the free agency system to strengthen the team. The problem we aim to address is identifying potential weaknesses in the team and developing strategies to improve team performance. Our goal is to help the team achieve the ultimate goal of winning the World Series title.

2. **Data Sources**:

This dataset from Kaggle provides offensive information for San Francisco Giants players from 1958 to 2023. It comprises 2722 observations and 31 variables. Among these variables, there are 3 categorical variables and 25 numerical variables. The categorical variables are Position, Dominant Hand, and Switch Hitter. The numerical variables include Runs, Hits, Stolen Bases, etc.

3. **Variables:**

To analyze the team's batting performance, we have identified 12 variables related to batting information that we would like to include in the following analysis. These variables can be divided into two categories: traditional batting metrics and advanced batting metrics. Traditional batting metrics include Hits (H), Doubles (2B), Triples (3B), Home Runs (HR), Runs Batted In (RBI), Bases on Balls (BB), Strikeouts (SO or K), and Double Plays Grounded Into (GIDP). These metrics provide a single numerical value for each batting metric a player has. Advanced metrics include Batting Average (AVG), Slugging Percentage (SLG), On-base Percentage (OBP), and On-base Plus Slugging (OPS). These batting metrics can provide more comprehensive and objective batting information for a player." (For a detailed explanation of variables, please see the data dictionary)

4. **Objective**:

We hope that the results will reveal the strengths and weaknesses within the team. Based on these findings, we can identify strategies to improve both individual players and team performance.

Additionally, the analysis results can provide valuable information for the general manager to make decisions regarding player trading.

5. **Research Question**:

In this section, we focus on two key research questions:

RQ 1: What methods are available for developing distinct training schedules aimed at improving the batting skills of various players?

RQ 2: Based on basic baseball knowledge, which suggests that more hits can result in more runs batted in (RBIs), what strategies can coaches employ to score more runs in a game?

6. **Analytics Techniques**

To address these research questions, we employ two main analytics techniques:

- Clustering: Using k-means clustering to group players and their defensive positions based on their batting performance, and then analyzing their strengths and weaknesses.
- Linear Regression: To determine the relationship between Runs Batted In (RBI), Hits (H), Doubles (2B), Triples (3B), and Home Runs (HR).

7. **Model Challenges:**

This dataset contains 2722 observations spanning from 1958 to 2023. It appears that there are multiple rows for some players. Before conducting the data analysis, we should undergo the data cleaning process to ensure that each row in the dataset represents information for only one player, and that each player's information appears only once in the dataset.

Additionally, there may be outstanding players within the team who could be considered outliers in the dataset. We intend to use statistical measures such as DFFITS, DFBETAS, and Cook's Distance to identify whether these outliers are influential observations. Based on this analysis, we will decide whether to remove these outliers.

Section III: Data Preparation and Methodology

Given the large dataset with 2722 observations and 31 variables, it's essential to preprocess and clean the data to extract the most important variables relevant to our research question. Here's a general outline of the data-cleaning process:

Filtering Data:

We began by filtering out players who had been with the San Francisco Giants for a minimum of five years. This step can ensure that our analysis focuses on players with tenure and contributions.

Variable Selection:

We narrowed our focus to the player's name, position, games played, and 14 batting skills variables as outlined in our data dictionary. This selection was directed by our research question which prioritizes batting performance.

Threshold Setting:

We set a threshold of at least 235 plate appearances to be included in the dataset. This method was used to maintain a standard of reliability in the batting data and avoid statistical anomalies for players who do not play often.

Data Summarization:

The dataset was summarized into two groups: one by individual players, and the values represent the statistics for each player per season. The other by position, providing insights into the contributions of different positions within the team.

Outliers Checking:

After analyzing the data using boxplots, it indicates that there are outliers within the dataset. However, the results of statistical measures such as DFFITS, DFBETAS, and Cook's Distance indicate that these outliers in the dataset are not influential cases. Therefore, we have decided not to remove those outliers from the dataset.

Dataset Outcome:

After completing the data cleaning process, for the data set with players, there are 51 observations and 17 variables, and there 8 observations and 17 variables for the data set with positions, these are the two datasets for our subsequent analysis.

Section IV: Analytical Techniques and Results

1. K-Means Clustering Analysis

The objective of Research Question 1 is to uncover the strengths and weaknesses of the San Francisco Giants, aiming to tailor specific training programs that address these areas and devise strategies for player trading. In this research question, we plan to create cluster plots to identify players' performance. We will present two different cluster plots: one for the different batting clusters based on players' batting performance, and another showing batting clusters based on players' defensive positions. (Appendix: the cluster analysis result)

This analysis employs the k-means clustering technique on key batting performance metrics. We chose the k-means clustering technique because it groups players based on their performance, allowing us to identify distinct areas for improvement.

The k-means clustering technique is particularly suited for our analysis as it is efficient with large datasets and provides clear, distinct clusters, which is essential for identifying specific performance metrics to focus on. Unlike other clustering methods, k-means is sensitive to the mean structure of data, making it ideal for recognizing patterns in batting performance where the mean is an important indicator.

1) Dataset Used

For this analysis, 12 variables were selected from the whole dataset to represent crucial aspects of batting performance. These variables include Hits, Home Runs, Strikeouts, Doubles, Triples, Runs Batted In, Base On Balls, Batting Average, Double Plays Grounded Into, Slugging Percentage, and On Base Percentage, and On Base Plus Slugging Percentage. These metrics serve as the foundation for our clustering analysis, providing a comprehensive view of each player's batting abilities and outcomes.

We did not omit outliers in cluster analysis, players with exceptionally top or low performance metrics were both included. Since our goal is to develop strategic training programs that benefit the entire team, it is essential to include all players, regardless of performance level. By analyzing every player, our clustering technique divides the team into groups based on their performance. This allows us to design specific strategies for each group, helping both top players and those who need more basic skills, which improves the whole team.

2) Analysis Results

Clustering By Players:

Optimal k Value:

When determining the optimal k of clusters, the silhouette method is used, which assesses the average silhouette width for various values of k. The silhouette width is a metric that gauges the similarity of an object to its cluster in contrast to other clusters, with a higher value signifying better cluster fit. For the dataset in this research question, the number of clusters was optimally set at 2, 3 and 4, as this value yielded the highest average silhouette width, indicating distinct and well-separated clusters. However, a k of 2 resulted in clusters of 24 and 27 players with considerable performance variance within each group, making targeted coaching challenging. A k of 4 produced the smallest cluster with only one player, which fails to represent a generalizable group. Therefore, the optimal k value is 3, as it provides balanced cluster sizes and shows a similarity within clusters and clear distinction between clusters. These three clusters are composed of 20, 25, and 6 players respectively, indicating varied group sizes within the team. (Appendix: the cluster analysis result)

Cluster 1 (Low Performers):

This cluster consists of 20 players who show below-average performance in all selected metrics. Players in this group have lower Runs, Hits, Home Runs, and significantly reduced batting and slugging percentages. These metrics suggest that these players struggle more consistently at the plate compared to their teammates.

Cluster 2 (Average Performers):

This cluster is the largest group containing 25 players, who represents the average performance level within the team. Players in this cluster exhibit moderate scores in metrics such as Batting Average and On Base Plus Slugging Percentage, slightly better than those in cluster 2 but still below the top performers. Their performance suggests a balanced, albeit unremarkable, contribution to the team's overall batting efforts.

Cluster 3 (Top Performers):

This cluster has 6 players, and players in this cluster have the highest performance metrics such as Hits, Home Runs, and Strikeouts. Despite their excellent performance in generating runs and hitting

home runs, their relatively high strikeout rates suggest an area where targeted training could reduce vulnerabilities and enhance their decision-making at the plate. This group's significant statistical impact across key metrics such as Slugging Percentage and On Base Plus Slugging Percentage indicates their crucial role on the team.

Clustering By Positions

Optimal k Value:

We are also using the silhouette method to determine the optimal value of K. This dataset focuses on batting performance across different positions in a baseball team, with only 8 observations available. The number of clusters was determined to be optimal at 2, indicating that the data points are best grouped into two clusters. This optimal K value provides the best balance of similarity within clusters and a clear distinction between clusters. Each of these two clusters consists of 4 different positions, indicating varied group sizes within the team. (Appendix: the cluster analysis result)

Cluster 1 (Top Performers):

The players who are first basemen, left outfielders, center outfielders, and right outfielders are in Cluster 1, exhibiting better batting performance. They usually provide more hits and home runs for the team than the players in cluster 2.

Cluster 2 (Low Performers):

The players within Cluster 2 are catchers, second basemen, third basemen, and shortstops. Compared to those in Cluster 1, these players typically exhibit lower batting performance for the team.

3) Strategic Development of Training Programs

Clustering By Players:

For Cluster 1 (Low Performers):

A custom training approach will be developed to specifically address each player’s deficits. This includes focused exercises on fundamental batting techniques, increased one-on-one coaching sessions, and regular performance evaluations to monitor progress. If significant improvement is not observed, the strategy includes considering potential trades or replacements. This approach ensures that the team maintains high performance and competitiveness, by either developing or restructuring the team composition.

For Cluster 2 (Average Performers):

This cluster will focus on improving accuracy and pitch selection to enhance contact quality and minimize strikeouts. Training will include extensive practice on recognizing pitches and choosing the right moments to swing, combined with technical adjustments to improve batting precision. Coaches will use data analytics to customize training sessions based on each player's specific weaknesses. This will help improve the overall performance of the group, making the players more consistent and effective at batting. Additionally, these players could come off the bench or be part of the bottom section of the batting lineup.

For Cluster 3 (Top Performers):

The strategy aims to refine existing skills to reduce the frequency of strikeouts and double plays. This involves enhancing decision-making at the plate, improving swing mechanics, and incorporating advanced video analysis to identify and correct specific batting flaws. Additionally, personalized coaching will focus on situational batting practice and mental conditioning to help players make better choices in high-pressure situations, thus maximizing their potential to contribute positively during games. Moreover, these players could be candidates for the cleanup hitter position in the batting lineup.

Clustering By Positions:

For Cluster 1 (Top Performers):

The players in Cluster 1 are outfielders and first basemen. They have better performance in hits, home runs, and RBIs compared to players in Cluster 2. Regarding the batting lineup strategy, these players in Cluster 1 could be considered as leadoff batters and typically would be among the first four players in the starting batting lineup. (Appendix: Descriptive Statistics for Position Analysis- Cluster 1)

For Cluster 2 (Low Performers):

Players in Cluster 2 are infielders, and their performance in terms of hits, home runs, and RBIs falls into the lower category within the team. These players could be considered for the bottom section of the batting lineup. (Appendix: Descriptive Statistics for Position Analysis - Cluster 2)

Conclusion:

For the cluster analysis of players, those in Cluster 3 (Top Performers) represent the potential strength within the team. They are the most powerful players on the San Francisco Giants. Coaches can

develop specific game plans to position these players at the top of the batting lineup or consider them as candidates for the cleanup hitter role. General managers could focus on catchers, second basemen, third basemen, and shortstops who are in the first cluster (Low Performers) during free agency for player trading to improve the overall team strength.

Continuous Evaluation:

We will monitor the outcomes of training and adjust our methods as player performance evolves. What's more, our dataset includes performance metrics from past seasons, with some players still on the San Francisco Giants and others who have left. For players currently on the team, we can directly apply the strategies outlined above. For new players, we will compare their performance metrics to those of our three clusters to determine where they fit. Based on this comparison, we will provide a targeted coaching program tailored to their specific needs.

2. Linear Regression Analysis

The objective of research question 2 is to determine the attributes that can influence the number of Runs Batted In (RBI) in a baseball team. The basic baseball concept illustrates a strong association between hits (H) and RBI. If a baseball team aims to score more runs in a game, they need to provide more hits. Based on this concept, we will perform linear regression, treating RBI as the dependent variable and hits, doubles, triples, and home runs as independent variables, to test which attributes can help a team score more runs in a game. Additionally, we will combine the clustering results from research question 1 to examine the performance of RBI differences between the players in those three clusters. This can provide advice for the coach to set up a better batting lineup and develop a better game plan for the team to win the game.

1) Data Used

For the data used in research question 2, 5 variables are selected: hits, doubles, triples, home runs, and runs batted in. Moreover, we also include clusters from the previous research question as dummy variables in the data to test the difference in runs batted in performance between the three clusters.

2) Analysis Result:

Our analysis investigates the relationship between the Runs batted in (RBI) and hits for San Francisco Giants baseball players using four regression models. Based on the dataset, we selected runs batted in (RBI), hits, doubles, triples, and home runs, which can have a major impact on the result analysis for the research question.

$$\widehat{RBI} = -27.961 + 0.72682(Hits)$$

First, we used a linear regression model where we first observed a significant relationship between hits (1.84e-13) and RBI ($p<0.05$). For each additional hit, the model predicts the RBI increase by 0.73 runs. This model gave us a baseline understanding of the relationship between RBI and hits.

$$\widehat{RBI} = 3.869814 + 0.178433(Hits) + 0.002262(Hits)^2$$

Second, we add the quadratic term of hits to the nonlinear model, which shows no significant improvement in our predictive analysis as compared to the simple linear model. The quadratic term of hits (hits^2) does not show statistical significance based on the p-value (<0.05). Thus, the simple linear regression model seems more sufficient.

$$\widehat{RBI} = 0.41287 + 0.21482(Hits) + 0.60505(Doubles) - 1.02953(Triples) + 1.60651(HomeRuns)$$

Third, we created a multiple linear regression model of RBI as a predictor and hits, doubles, triples, and home runs as a response variable. Overall, the models show a significant relationship between RBI and the other response variables, where hits (0.000763), doubles (0.006029), and home runs(2e-16) show significant predictors of RBI. While triples (0.07) is a marginally significant predictor of RBI.

$$\widehat{RBI} = 10.3051 + 1.1429(Doubles) - 0.2693(Triples) + 1.7336(HomeRuns)$$

Fourth, we excluded the hits and kept the doubles, triples, and home runs in the model. It shows that the doubles (1.19e-08) and home runs (2e-16) are still statistically significant, whereas triples (0.65010) are not significant to the RBI predictor. By excluding the hits from the model, we cannot see any drastic changes in the predictive model. It may partially mediate the effect of doubles and home runs on the RBI.

We also tried using the previous analysis result on clustering to test the difference in the estimated mean value of RBI between the clusters. Using the clustering result by players, we set cluster 2 as the baseline group, as it represents the group of average performers in the analysis. In the cluster by position, we set the first baseman as the baseline group.

Setting clusters based on players' batting performance as a dummy variable:

$$\widehat{RBI} = -3.18 + 0.2274(Hits) + 0.6409(Doubles) - 0.9606(Triples) + 1.6838(HomeRuns) + 1.1397(Cluster1) - 2.9429(Cluster3)$$

Players in cluster 1, compared to players in cluster 2, have an average of 1.14 more RBI, all else equal, but it is not statistically significant. Similarly, players in cluster 3, compared to players in cluster 2, have an average of 2.94 less RBI, all else equal, but it is not statistically significant.

Setting clusters based on positions' batting performance as a dummy variable:

$$\widehat{RBI} = 1.5803 + 0.1835(Hits) + 0.7116(Doubles) - 0.2858(Triples) + 1.6078(HomeRuns) - 4.1562(SecondBasemen) - 1.1323(ThirdBasemen) + 0.0434(Catchers) - 3.9763(CenterOutfielders) - 3.4430(LeftOutfielders) - 4.3825(RightOutfielders) - 0.2211(ShortStops)$$

Second basemen, compared to first basemen, have an average of 4.16 less RBI, all else equal, but it is not statistically significant. Similarly, third basemen, catchers, center outfielders, left outfielders, right outfielders, and shortstops, compared to first basemen, have average RBI differences of 1.13, 0.04, 3.98, 3.44, 4.38, and 0.22, respectively, all else equal, but none are statistically significant. (Appendix: Regression analysis results)

3) Conclusion

Overall, the simple linear regression model shows a significant relationship between hits and RBI. Furthermore, offensive statistics such as doubles, home runs, and hits remains significant to the RBI predictor in the model. While the quadratic term of hits in the nonlinear model does not show any significance, which does not enhance our predictive accuracy. Based on the analytics, we can suggest that hits are indeed important for more Run Batted In(RBI) but we can also suggest that the coaches can make a comprehensive strategy that takes into account different offensive parameters that offer more thorough knowledge regarding the player's performance. Additionally, setting clusters as dummy

variables also provides some information for coaches to make better decisions regarding batting lineup; for instance, first basemen in cluster 1 could be considered as candidates for the cleanup hitter position.

Section V: Conclusion and Recommendations

In conclusion, our analysis of Major League Baseball data, focusing on the batting performance of the San Francisco Giants, has provided valuable insights for coaches, players, and the general manager. Through clustering analysis, we identified clusters in batting performance and highlighted areas for improvement. These insights can inform training programs, player trading decisions, and game strategies.

Moving forward, it's essential to consider not only batting performance but also defensive ratings when making strategic decisions. Incorporating defensive abilities into player evaluations can lead to more comprehensive and effective team strategies.

Overall, this project demonstrates the power of data analytics in enhancing performance and decision-making in professional sports. By leveraging data-driven insights, teams can optimize their strategies and maximize their chances of success on the field.

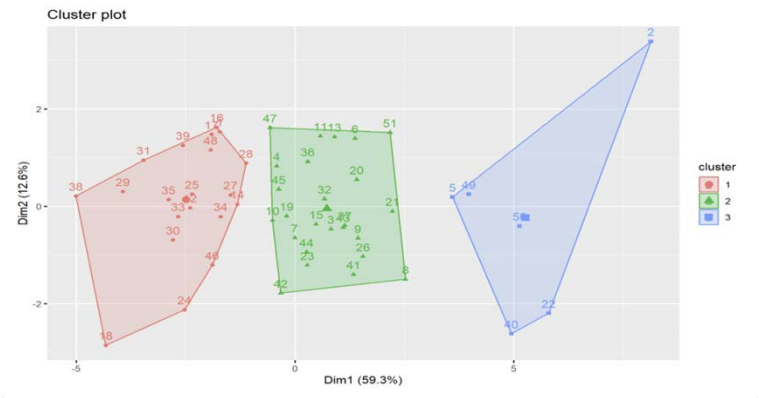
Appendix

Data Dictionary

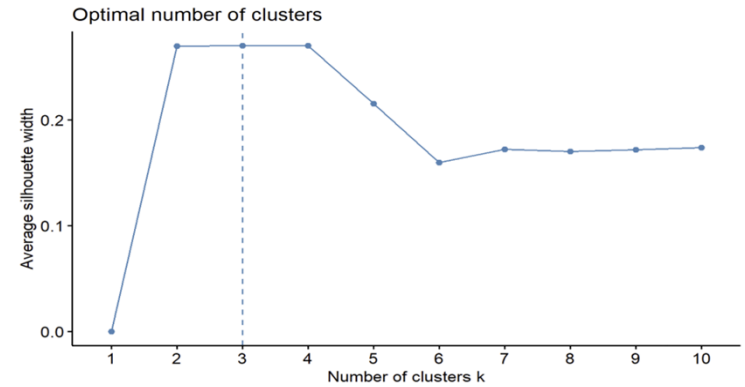
- Name: Player’s name
- Position: Player’s defense position
- Games: How many games the player played
- Plate Appearance: A plate appearance refers to a batter's turn at the plate. Each completed turn batting is one plate appearance.
- At Bat: An official at-bat comes when a batter reaches base via a fielder's choice, hit or an error (not including catcher's interference) or when a batter is put out on a non-sacrifice.
- Hits (H): A hit occurs when a batter strikes the baseball into fair territory and reaches base without doing so via an error or a fielder's choice.
- Doubles(2B): A batter is credited with a double when he hits the ball into play and reaches second base without the help of an intervening error or attempt to put out another baserunner.
- Triples(3B): A triple occurs when a batter hits the ball into play and reaches third base without the help of an intervening error or attempt to put out another baserunner.
- Home Runs (HR): A home run occurs when a batter hits a fair ball and scores on the play without being put out or without the benefit of an error.
- Runs Batted In (RBI): A batter is credited with an RBI in most cases where the result of his plate appearance is a run being scored.
- Base on ball (BB): A walk occurs when a pitcher throws four pitches out of the strike zone, none of which are swung at by the hitter.
- Strikeout (SO, K): A strikeout occurs when a pitcher throws any combination of three swinging or looking strikes to a hitter.
- Batting average (AVG): One of the oldest and most universal tools to measure a hitter's success at the plate, batting average is determined by dividing a player's hits by his total at-bats for a number between zero (shown as .000) and one (1.000). (H/AB)

- Slugging Percentage (SLG): SLG represents the total number of bases a player records per at-bat. Unlike on-base percentage, slugging percentage deals only with hits and does not include walks and hit-by-pitches in its equation. (1B) + (2*2B) + (3*3B) + (4*4B)/AB
- On-base Percentage: OBP refers to how frequently a batter reaches base per plate appearance. Times on base include hits, walks and hit-by-pitches, but do not include errors, times reached on a fielder's choice or a dropped third strike.
- On-base Plus Slugging (OPS): OPS adds on-base percentage and slugging percentage to get one number that unites the two. It's meant to combine how well a hitter can reach base, with how well he can hit for average and for power. OBP+SLG
- Double Plays Grounded Into (GIDP): A GIDP occurs when a player hits a ground ball that results in multiple outs on the bases.

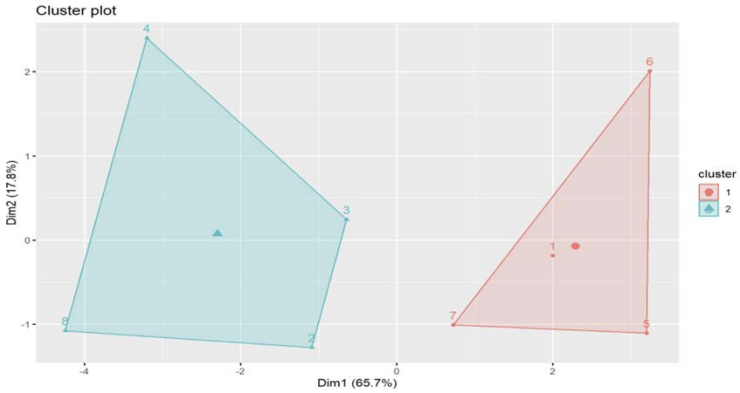
Cluster result by players



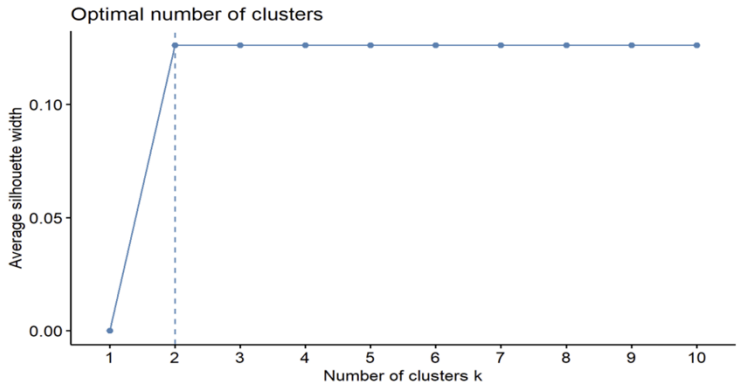
Output of optimal K for cluster by players:



Cluster result by position:



Output of optimal K for cluster by players



Descriptive Statistics for Players Analysis - Cluster 1

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
Hits	20	96.0	16.4	69.5	133.0
Doubles	20	14.7	2.9	8.0	21.0
Triples	20	2.8	1.6	0.5	7.2
Home_Runs	20	5.3	2.7	1.0	10.3
Runs_Batted_In	20	35.6	7.3	16.0	50.5
Base_On_Balls	20	29.3	9.2	11.6	42.6
Strikeouts	20	54.8	18.3	25.0	89.0
Batting_Average	20	0.3	0.01	0.2	0.3
Slugging_Percentage	20	0.4	0.04	0.3	0.4
On_Base_Percentage	20	0.3	0.02	0.3	0.3

Descriptive Statistics for Players Analysis - Cluster 2

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
Hits	25	120.5	11.4	96.3	145.2
Doubles	25	22.5	3.7	13.5	28.8
Triples	25	2.8	1.0	0.8	4.0
Home_Runs	25	15.8	4.6	7.0	26.8
Runs_Batted_In	25	62.6	9.1	45.1	79.2
Base_On_Balls	25	49.8	15.0	23.8	75.6
Strikeouts	25	78.0	15.7	54.8	116.4
Batting_Average	25	0.3	0.01	0.2	0.3
Slugging_Percentage	25	0.4	0.03	0.4	0.5
On_Base_Percentage	25	0.3	0.02	0.3	0.4

Descriptive Statistics for Players Analysis - Cluster 3

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
Hits	6	161.6	14.1	138.5	180.9
Doubles	6	30.8	5.6	26.7	41.2
Triples	6	4.3	1.3	2.9	6.0
Home_Runs	6	30.6	6.6	22.0	41.5
Runs_Batted_In	6	97.9	12.9	78.9	114.8
Base_On_Balls	6	73.7	34.3	36.0	138.4
Strikeouts	6	97.1	27.5	67.4	145.1
Batting_Average	6	0.3	0.01	0.3	0.3
Slugging_Percentage	6	0.5	0.1	0.5	0.7
On_Base_Percentage	6	0.4	0.05	0.4	0.5

Descriptive Statistics for Position Analysis - Cluster 1

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
At_Bats	4	450.1	29.8	405.8	468.1
Hits	4	119.4	8.8	110.2	127.9
Doubles	4	21.5	3.0	18.9	25.5
Triples	4	2.8	1.0	1.4	3.7
Home_Runs	4	12.4	3.8	8.0	17.3
Runs_Batted_In	4	56.0	7.9	47.2	66.2
Base_On_Balls	4	43.0	4.4	38.6	49.0
Strikeouts	4	70.6	7.5	59.4	75.2
Batting_Average	4	0.3	0.01	0.2	0.3
Slugging_Percentage	4	0.4	0.04	0.3	0.4
On_Base_Percentage	4	0.3	0.02	0.3	0.3

Descriptive Statistics for Position Analysis - Cluster 2

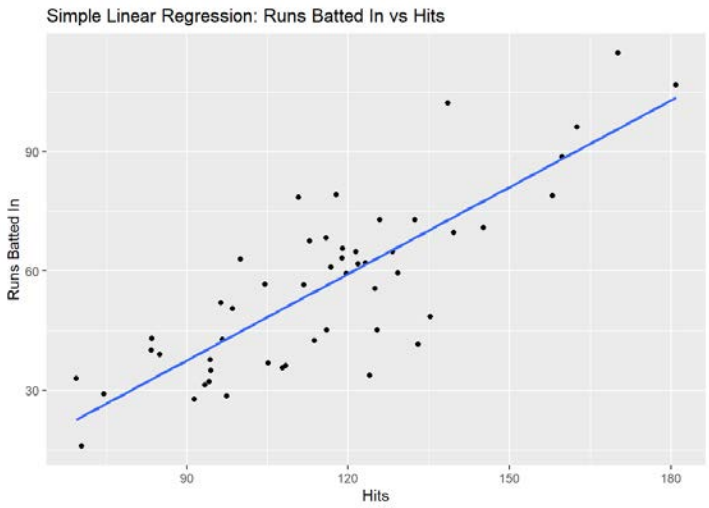
Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
At_Bats	4	464.6	27.1	439.1	500.7
Hits	4	131.5	7.9	126.3	142.9
Doubles	4	23.5	0.7	22.7	24.3
Triples	4	3.6	0.8	2.9	4.7
Home_Runs	4	21.9	3.8	17.2	26.1
Runs_Batted_In	4	73.1	5.7	64.8	77.8
Base_On_Balls	4	62.3	13.2	47.8	79.9
Strikeouts	4	81.8	6.6	72.7	88.3
Batting_Average	4	0.3	0.005	0.3	0.3
Slugging_Percentage	4	0.5	0.03	0.4	0.5
On_Base_Percentage	4	0.4	0.02	0.3	0.4

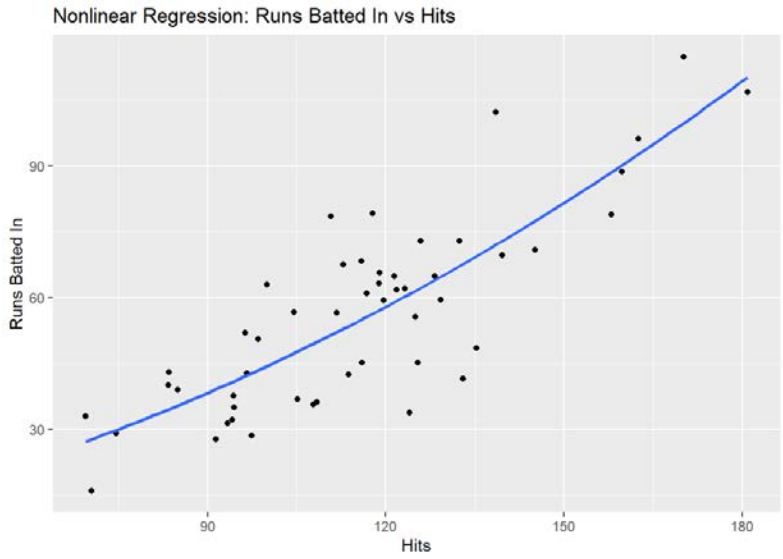
Regression Analysis Results:

	Dependent variable: RBI					
	(1)	(2)	(3)	(4)	(5)	(6)
Hits	0.727*** (0.072)	0.178 (0.521)	0.215*** (0.060)		0.227*** (0.064)	0.183*** (0.064)
Hits 2		0.002 (0.002)				
Doubles			0.605*** (0.210)	1.143*** (0.166)	0.641** (0.241)	0.712*** (0.236)
Triples			-1.030* (0.567)	-0.269 (0.590)	-0.941 (0.594)	-0.286 (0.744)
Home Runs			1.607*** (0.109)	1.734*** (0.116)	1.684*** (0.185)	1.608*** (0.126)
Cluster:cluster1					1.140 (3.320)	
Cluster:cluster3					-2.943 (4.765)	
Second Basemen						-4.156 (3.362)
Third Basemen						-1.132 (2.743)
Catchers						0.043 (3.083)
Center Outfielders						-3.976 (3.163)
Left Outfielders						-3.443 (3.173)
Right Outfielders						-4.382 (3.022)
Shortstops						-0.221 (3.099)
Constant	-27.961*** (8.573)	3.870 (31.129)	0.413 (3.807)	10.305*** (2.957)	-3.180 (8.166)	1.580 (4.716)
Observations	51	51	51	51	51	51
R2	0.672	0.680	0.952	0.939	0.953	0.958
Adjusted R2	0.666	0.667	0.948	0.935	0.946	0.946
Residual Std. Error	12.612 (df = 49)	12.596 (df = 48)	4.962 (df = 46)	5.560 (df = 47)	5.051 (df = 44)	5.079 (df = 39)
F Statistic	100.527*** (df = 1; 49)	50.964*** (df = 2; 48)	229.992*** (df = 4; 46)	240.823*** (df = 3; 47)	148.043*** (df = 6; 44)	80.269*** (df = 11; 39)
Note:						*p<0.1; **p<0.05; ***p<0.01

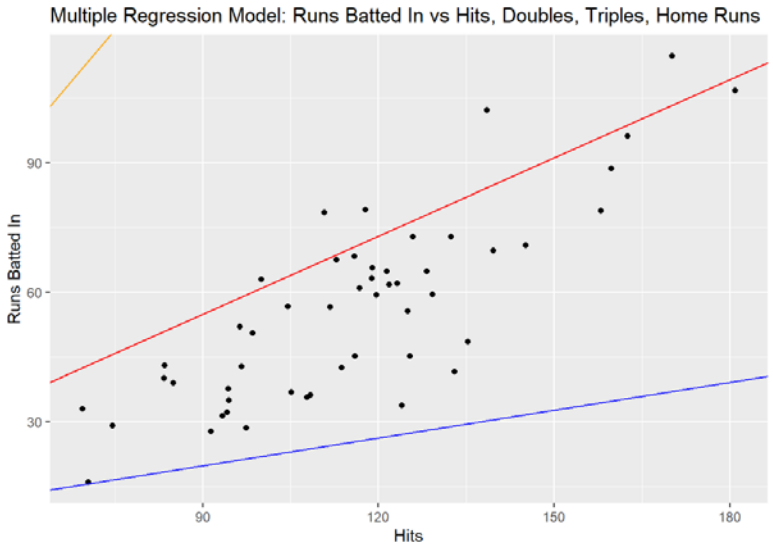
		Dependent Variable Run_Batted_In (RBI)					
		Hit.RBI	Nonlinear_model	Model	Model2	Model 3	Model 4
Hits	Estimates	0.72682	0.178433	0.21482		0.22741	0.1835
	P-Value	1.84E-13	0.733	0.000763		0.000912	0.00677
I(Hits^2)	Estimates		0.002262				
	P-Value		0.293				
Doubles	Estimates			0.60505	1.1429	0.64087	0.71159
	P-Value			0.006029	1.19E-08	0.01098	0.00444
Triples	Estimates			-1.02953	-0.2693	-0.94061	-0.28584
	P-Value			0.075955	0.6501	0.120719	0.703
Home_Runs	Estimates			1.60651	1.7336	1.68376	1.60782
	P-Value			2.00E-16	2.00E-16	1.19E-11	1.83E-15
base21 (Cluster 1)	Estimates					1.13965	
	P-Value					0.73305	
base23 (Cluster 3)	Estimates					-2.94285	
	P-Value					0.539983	
Position2B	Estimates						-4.15623
	P-Value						0.22374
Position3B	Estimates						-1.13226
	P-Value						0.68206
PositionC	Estimates						0.0434
	P-Value						0.98884
	Estimates						
PositionCF	P-Value						-3.97631
							0.21612
	Estimates						
	P-Value						
PositionLF							-3.44301
	Estimates						0.2845
	P-Value						
PositionRF							-4.38247
	Estimates						0.155
	P-Value						
PositionSS							-0.22112
	Estimates						0.94349
	P-Value						
Constant		-27.96072	3.869814	0.41287	10.3051	-3.17997	1.58033
		0.00202	0.902	0.914104	0.00108	0.698837	0.73932



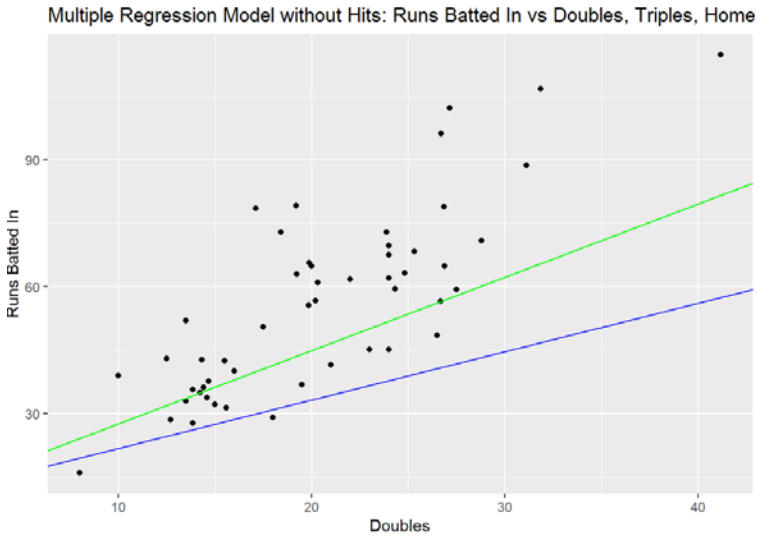
The graph shows the positive relationship between hits and RBI for the San Francisco Giants. As the number of hits increases, the expected number of RBI also increases. Which suggests that hits is a significant predictor of RBI.



We can see that the fitted curve does not significantly deviate from a straight line. Which suggests that the relationship between hits and RBI can be understood as a simple linear relationship.



The graph shows the relationship between hits, doubles, triples, home runs, and RBI. Each regression line in the graph represents each predictor while holding the other constant. The blue is hits, the red line is doubles, and the orange line is home runs. All three variables show a positive relation with the RBI, which suggests that higher RBI counts are related to more hits, doubles, and home runs.



The graph shows the relationship between doubles, triples, home runs, and RBI while excluding the hits. The blue is doubles, and the green line is home runs. Both the variables show a positive relationship with the RBI, which suggests that higher RBI counts are related to more doubles and home runs even when hits are not considered.

References:

ChatGPT, personal communication, April. 13, 2024. https://chat.openai.com/share/67f2ef91-5874-4c6d-87ef-cc8cf5b46218

Goro. (2024, 3 23). 3-22-2024-Nike-St-George-cross. Retrieved from Super Torch Ritual: https://www.supertorchritual.com/wp-content/uploads/2024/02/Super-Bowl-LVIII-2024-Vegas-b.png

OP, M. (2023, December 27). San Francisco Giants batting & pitching data. Kaggle. https://www.kaggle.com/datasets/mattop/san-francisco-giants-batting-and-pitching-data?resource=download

Statcast leaderboard. baseballsavant.com. (n.d.). https://baseballsavant.mlb.com/leaderboard/statcast

We acknowledge that the following resources were used to complete this presentation:

Wordtune

Grammarly