Final Project:
Cricket and Baseball: Evaluation of Leagues, Players and Play

Andrew Curtis

April 23, 2023

1 Introduction

An American audience is likely familiar with the much beloved sport of baseball. However there is another sport garnering an astonishingly larger and similarly dedicated fan base, cricket! Estimates of fans and participants of these sports comes in at 500 million for baseball and over 2.5 billion for cricket [1]. The purpose of this investigation is to elucidate similarities and differences between these sports. In order to accomplish this, the sports are evaluated from the viewpoint of a newly-hired quantitative analyst from a team in Major League Baseball (MLB) and from a team in the Indian Premier League (IPL).

Baseball is often analyzed from an analytical perspective by both fans and professionals involved with the sport. As a result, top-notch data is available. However even though cricket is by far the more popular sport, records and data are messy and less available. Novel data collection for cricket was a core component and challenge to this project.

The setup for investigations into baseball are as follows. The New York Yankees fans have been booing calling for a change of quantitative analyst. The manager, Aaron Boone, has relented and hired you for your strong data science background. He wants to identify some key performance metrics and baselines for players. In order to analyze performance, he also wants you to look at some mechanical and physical data for batters and pitchers. Finally, he wants to know if he should be planning to give contracts to aging star players or invest in new talent; specifically should he continue to pay Aaron Judge (A star positional player) and Gerrit Cole (a star pitcher), players with large contracts. The objectives of the baseball investigation are summarized below.

- What are important performance metrics for batters and pitchers and what are statistics and baselines on each?
- How do tracked mechanics of batters and pitchers relate to the identified performance metrics and what can we learn from these relations?
- What is the effect of age on a players performance? Should aging star players be retained?

The IPL is the most popular cricket league in the world. It is sixth-highest for attendance of any sports league and in 2014 was evaluated with to have a brand value of $11 billion dollars. There are currently 10 teams and each team must have 25 players on their roster, with a maximum of 8 import players; although, during a game only 11 players are fielded. One of the novel analysis conducted by this project is amending a feature to the players that identifies them as a batter, bowler or all-arounder. Finally, we investigated the effect of home-field advantage for the rival teams of Chennai Super Kings (CSK) and Royal Challengers

Bangalore (RCB). These two teams are led by star players and often finish first and second in the IPL. The objectives of the cricket investigation are summarized below.

- What are important performance metrics for batters and bowlers and what are statistics on each?
- What is the importance of the new feature of player type (batter, bowler, all-arounder)?
- What is the effect of toss result and home field advantage for our featured match-up of CSK vs. RCB?

2 Dataset

The data collection and cleaning process was a key challenge for this project, primarily with the cricket datasets. Initially, two datasets were found for cricket from Kaggle. Although, these were limited with the amount of information contained and with inconsistent and messy data. For this project, team members worked to web-scrape from Wikipedia additional season, match, and player data for use. Additionally, a great deal of manual entry was required. Manual entry added the player's dates of birth as well as a new feature for the type of player (batter, bowler, or all-arounder). Once this process was completed, data pre-processing was the next step. This involved matching different teams and stadiums in the dataset. The 2020 and 2021 seasons were removed as covid affected the league tremendously during this time period. UAE grounds matches were also removed to have the data just contain the most popular league, IPL. Through dedicated feature engineering of group members, the following attributes were developed: average runs, strike rates, bowling economy per-100-balls, toss winners and match winners for teams and players.

In 2015, baseball moved into what is called the Statcast-era. This was a league-wide (over 30 stadiums) adoption of new camera and radar technology to collect additional data on batters and pitchers. New attributes being tracked include exit velocity and launch angle off the bat for batters and pitch velocities and spins for pitchers. This was also enabled by an update to the system in 2019 called, Hawkeye. Hawkeye increased the amount of tracked balls from 89% to 99%. Due to the better tracking and to keep analysis more relevant to the current era, the baseball dataset analysis was conducted over the years of 2019-2022. Data was collected from the MLB and Google Statcast partnership website, baseballsavant.mlb.com. Many attributes were available, but some key performance statistics and predictors were collected.

Performance metrics that were focused for batters include: Batting average and on-base percentage, and through feature engineering runs-per-at-bat and home-runs-per-at-bat were introduced. In this investigation we wanted to try to deduce the cause and relation of performance for the two sports. So for batters, player age, exit velocity and launch angle were examined for their relations with the scoring performance metrics.

For pitchers, the main performance metric was ERA (expected-runs-allowed), this is a measure of the overall defensive capability of a pitcher. Various analysis were conducted to find which attributes were significant for pitching. The main statistically-significant factors explored include pitch velocity and spin rates for pitch types as well as player age.

3 Analysis Technique

In order to answer the questions presented, the statistical analysis often involved testing for correlations between different numeric variables. This was done via pearson correlation

tests. Results of analysis are also given by standard measures such as mean and quantiles. Visualization for comparison is given by a variety of plots including,: scatter, regression, histogram and bar plots. Linear regression models were tested on the data, although due to the high variance for some statistics in the datasets, are not as useful as other analysis methods employed.

4 Results

4.1 Baseball

Focusing first on baseball batting, many different statistics can be tracked that capture various aspects of scoring. The first two chosen to focus on are popular within analysis of the sport, batting average and on-base-percentage. Batting average is calculated by the number of hits per the number of times at bat. This stat has some bias towards batters that are generating a lot of hits, but due to walks there are other ways to get on base. This is why on-base-percentage is also considered; it is the number of times a batter makes it on base per at bat. In this way we can access hitting performance but also some hidden factors that effect making it on base such as taking walks.

The other two performance metrics were created to capture more information about batting. In order to create a fair statistic, runs scored and home runs hit are both considered in by dividing by at bats for each player. So our final two metrics are runs-per-at-bat and homeruns-per-at-bat.

Batting mechanics stats are interesting but only a few are tracked and available. The two considered are average exit velocity off the bat and average launch angle off the bat of each player. Summary statistics of each performance metric and batting mechanics are shown in Table 1, along with the our identified star players.

Using pearson correlation tests, exit velocity significantly predicts for all of our identified performance metrics. However, launch angle is only significant for home runs per at bat. This is an interesting result. Harder batted balls with higher exit velocity have a better chance of resulting in the player getting on base. A lower launch angle could indicate the player is hitting more ground balls or less fly balls. This overall results in hitting balls that do not carry as far, but this isn't necessarily a negative result. For hitting many home runs, it is important to have both a higher exit velocity and higher launch angle. Figure 1 shows the exit velocity and launch angle and their relation to batting average (left) and to home runs per-at-bat (right). Home runs per-at-bat increases with launch angle whereas batting average does not. It should be noted that for both batting and pitching statistics variance is high within the data set and while results are significant, players should be viewed individually for their performance when making final managerial decisions.

For batters, the distribution of player ages has a gaussian distribution with a spike around

	Bottom 25%	Median	Top 25%	Aaron Judge
Batting Average	0.235	0.261	0.286	0.300
On Base Percentage	0.309	0.337	0.364	0.389
Runs per At Bat	0.124	0.147	0.171	0.207
Home Runs per At Bat	0.027	0.042	0.056	0.091
Exit Velocity (mph)	87.6	89.3	90.8	95.3
Launch Angle (deg)	9.35	12.9	16.4	16.0

Table 1: Batting performance and tracked mechanics summary statistics for players and star player Aaron Judge.

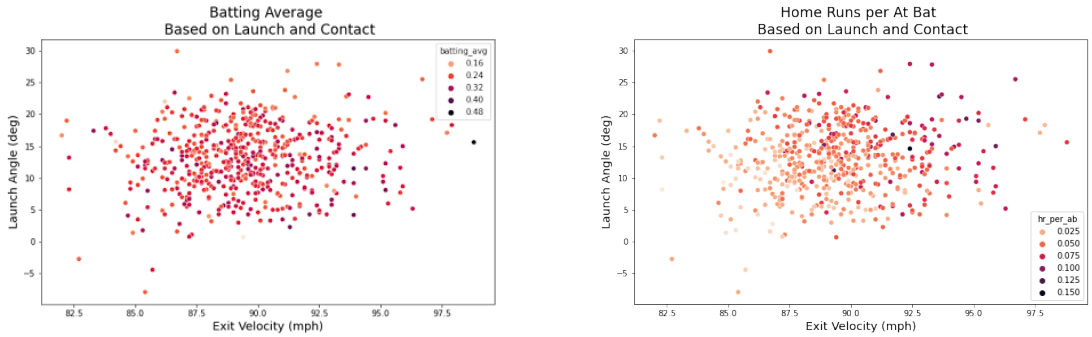

Figure 1: Average Exit Velocity (mph) and average Launch Angle (deg) of players, colored by Batting Average (left) and Home Runs Per-At-Bat (right).

27 years of age, seen in Figure 2. This spike may specifically deal with the second round of contract years for players. Player age is not significantly correlated with batting average, on-base percentage, home runs per-at-bat or exit velocity. The scatter plot of exit velocity and player age can be seen in Figure 3 (left). Launch angle increases with player age, seen in Figure 3 (right). However, the most surprising result was a negative correlation of player age with runs per-at-bat, also seen in Figure 3 (bottom). These relations indicate that for getting on base and hitting aging batters don't get worse on average and may even hit more home runs. Where they do suffer is overall runs scored. This could be because they become slower at base running and thus score less runs. Or perhaps they are placed at less opportune positions in the batting lineup. Overall, it is the conclusion of these analysis that it is safe to give contracts to aging positional players as long as their summary statistics in identified performance metrics have not decreased significantly.

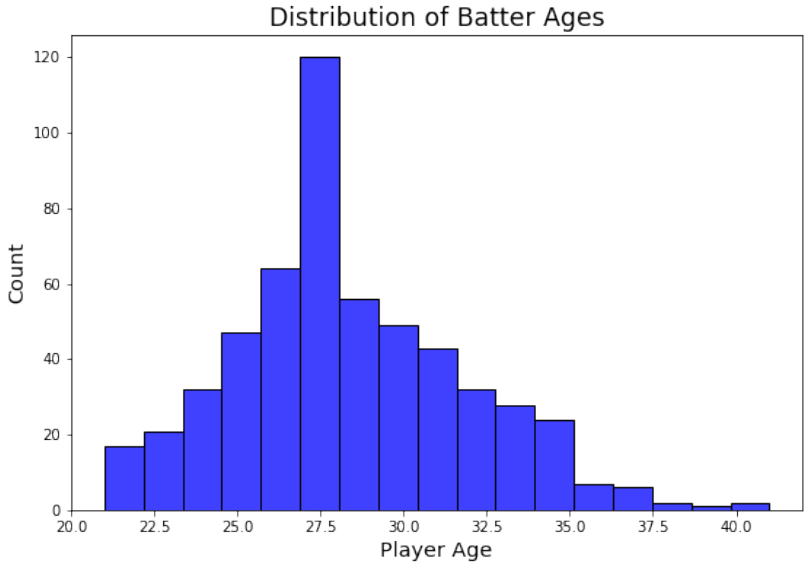

Figure 2: Distribution of batter player ages.

Only one performance metric is presented for pitchers, a common measurement called ERA or expected runs allowed. ERA is calculated by taking the average of earned runs allowed by a pitcher divided by nine innings pitched. It is essentially a holistic measure of the scoring defense ability of a pitcher. This metric is able to capture many factors of a game. Through

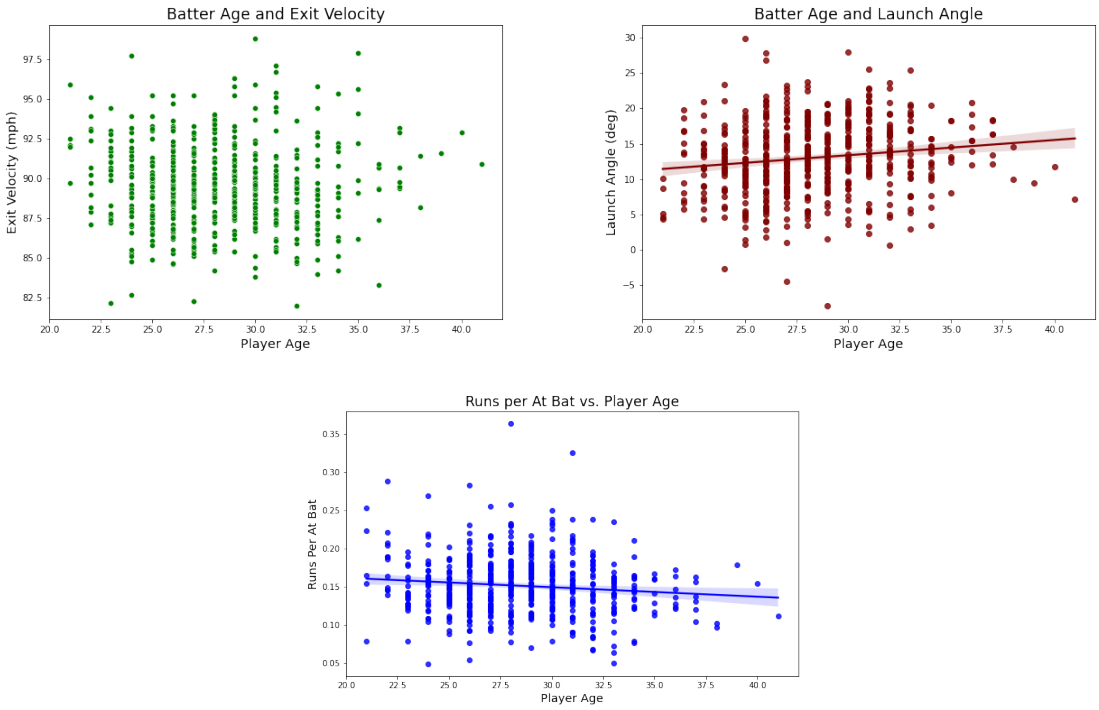

Figure 3: Batter player age compared to average Exit Velocity (mph) (left), average Launch Angle (deg) (right) and Runs Per-At-Bat (bottom).

	Bottom 25%	Median	Top 25%	Gerrit Cole
ERA	4.40	3.36	0.270	0.274
Fastball Velocity (mph)	92.0	93.4	95.0	97.8
Fastball Spin (rpm)	2200	2288	2398	2428
Slider Velocity (mph)	83.5	85.7	87.2	88.7
Slider Spin (rpm)	2289	2412	2557	2569

Table 2: Pitching performance and tracked mechanics summary statistics for players and star player Gerrit Cole.

analysis, it was found that the tracked mechanics that were both significant and most highly correlated with ERA are fastball velocity and spin, and slider velocity and spin. Fastballs and sliders are by far the most common types of pitches thrown. Interestingly, it was found that the variety or count of pitch types thrown by pitchers was not significantly correlated with ERA. Summary statistics for the selected performance metric and tracked mechanics are presented in Table 2, along with values for our star pitcher Gerrit Cole.

The distribution of pitcher player ages is given in Figure 4. Unlike batters, the count of pitchers by age falls off consistently. This is likely due to players becoming worse at some of the tracked mechanics as well as more significant injuries experienced when compared positional players. Figure 5 shows the correlations between player age and increasing ERA (left) and decreasing strikeout percentage (right).

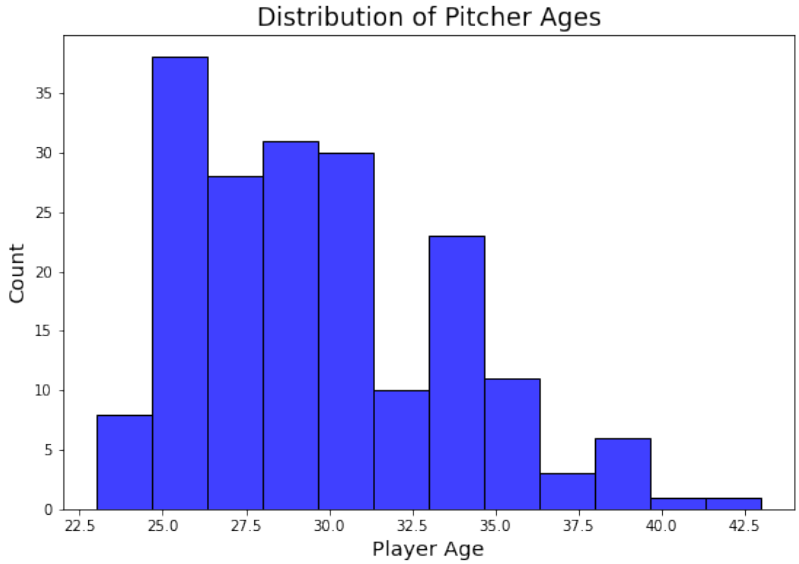

Figure 4: Distribution of pitcher player ages.

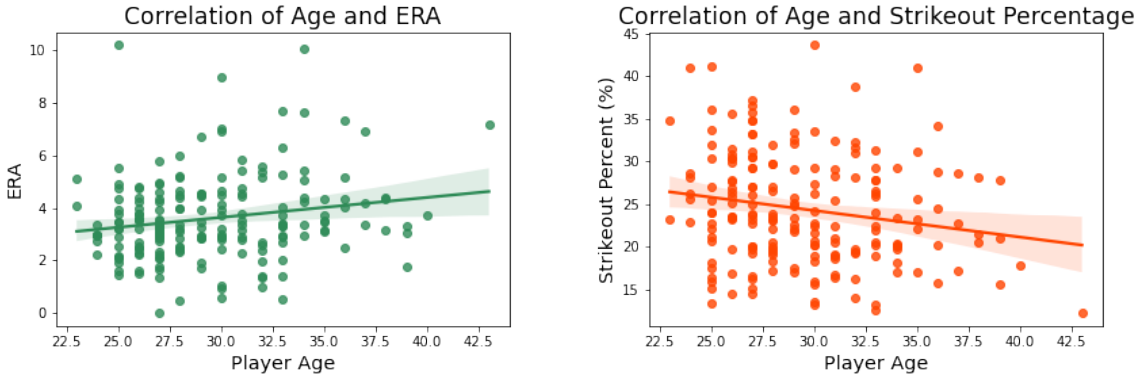

Figure 5: Pitcher age compared with ERA and Strikeout Percentage.

Figure 6 shows pitcher fastball velocity and spin rates compared to ERA and is colored by player age. Figure 7 shows the same for sliders. The takeaway from this analysis is that these four tracked mechanics have significant correlations with ERA. Older pitchers lose velocity on their fastballs and sliders leading to higher ERA. However, their spin rates do not significantly decrease with age. Overall, as pitchers get older they lose velocity and perform worse as a result. There is a high amount of variance in the data, so a manager would want to look player to player when making managerial decisions. It is much more likely that a pitcher will decrease in value over time when compared to a batter. This is reflected by the distribution of player ages of pitchers falling off much more consistently than the gaussian distribution of batter ages.

4.2 Cricket

Important performance metrics for cricket were developed from the produced datasets. For batters, strike rate is presented. This measure is caluclated by the number of runs scored for every 100 balls, where higher is better and above 120 would be considered a good player. Economy is presented for bowlers. Economy is the average number of runs conceded per over

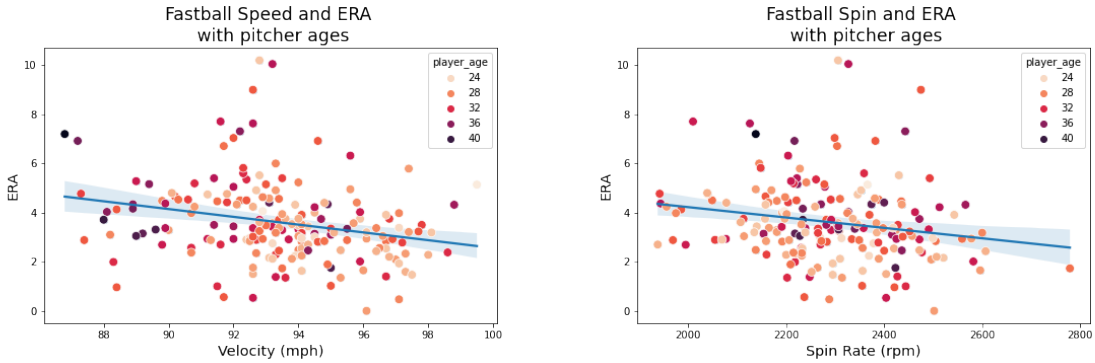

Figure 6: Fastball Velocity (left) and Spin (right) compared to ERA and colored by Pitcher ages.

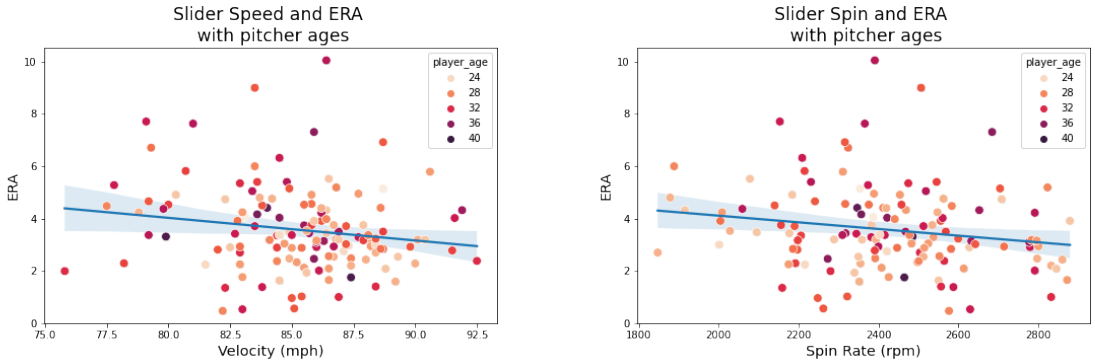

Figure 7: Slider Velocity (left) and Spin (right) compared to ERA and colored by Pitcher ages.

bowled. Strike rate can be thought of as similar to batting average and economy would be similar to ERA for baseball. In Figure 8, the distributions of strike rate (left) and economy (right) can be seen.

Cricket players can generally be broken down into 3 groups: batters, bowlers and all-arounders. In cricket, all players that take the field will have to perform both batting and bowling, but it is often advantageous to match up good bowlers against the opposing team's good batters. Thus the lineup order is important to a game's outcome in cricket. Although, the more modern trend is to field players that are good at both batting and bowling, essentially all -aroudner star players. The count of these three groups can be seen in Figure 9.

Finally, we compare the win rates of Royal Challengers Bangalore and Chennai Super Kings at their respective stadiums. Homefield advantage for RCB at M Chinnaswamy Stadium can be seen in Figure 10 (left). Homefield advantage for CSK at MA Chidambaram Stadium, Chepauk is seen in Figure 10 (right).

Figure 10 (right) helps elucidate that CSK tends to win almost all of their matches at home regardless of toss result. When RCB has homefield advantage, the toss has still often gone in favor of CSK. Despite this result, RCB wins about half of the games, seen in Figure 10 (left). This indicates that toss result has less of an effect when compared to homefield advantage for the teams.

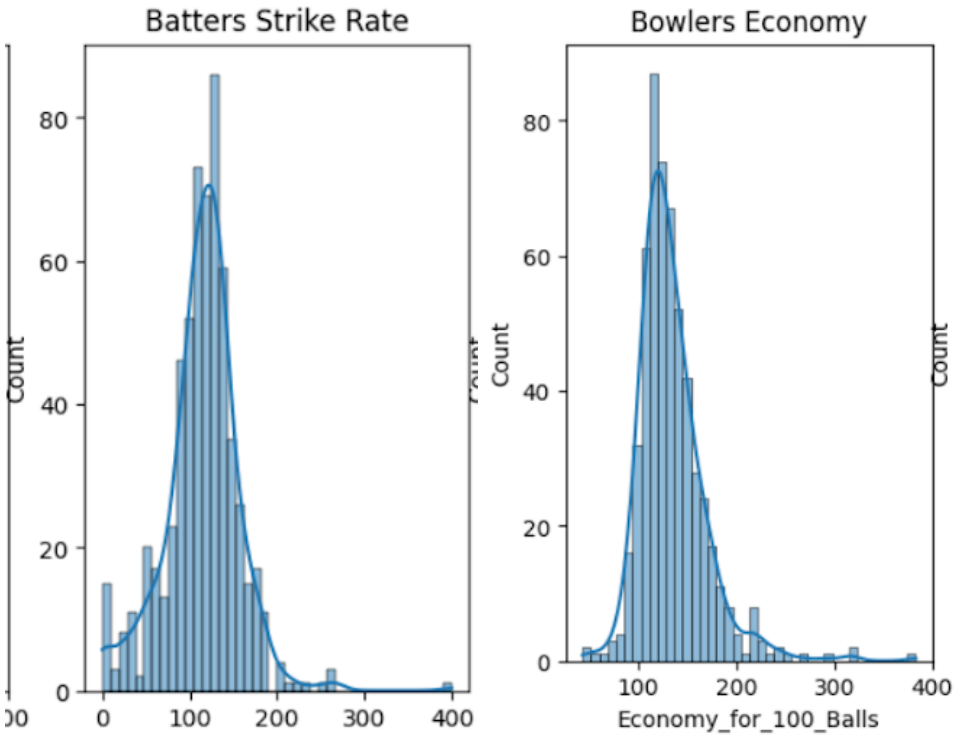

Figure 8: Distributions of strike rate (left) and economy (right) for cricket players

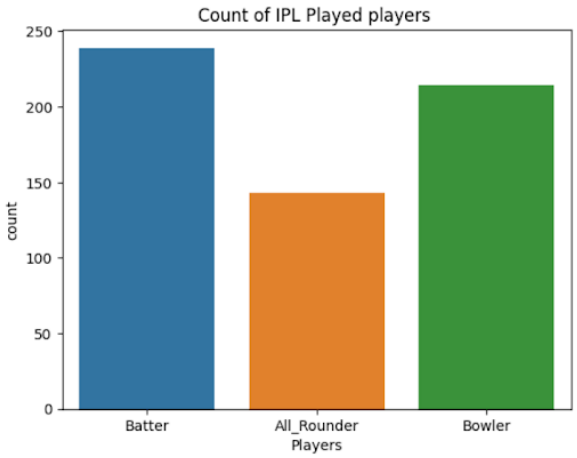

Figure 9: Count of cricket players by type.

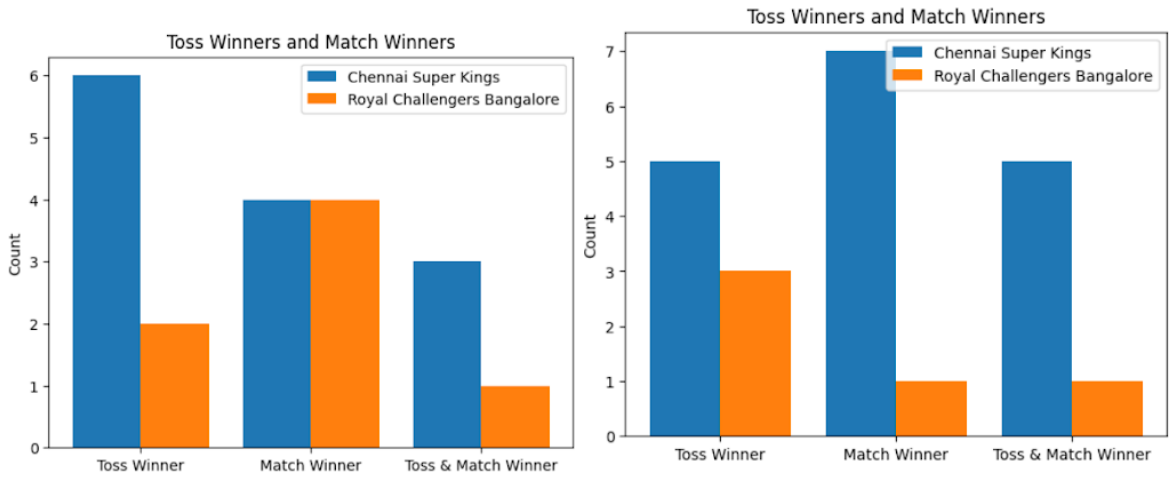

Figure 10: Toss and match winners between Royal Challengers Bangalore and Chennai Super Kings split by homefield advantage (for RCB left and CSK right).

5 Technical

The most time consuming and difficult steps with this project involved the novel data collection for the cricket datasets. While cricket is a much more popular sport than baseball, the availability of data on players and play is much harder to come by. Web-scraping and manual entry was employed to develop the datasets needed for cricket. After the data collection process, was a good deal of pre-processing necessary to clean the data and develop the features for analysis. Some of these pre-processing measures are discussed in the datasets section. Linear regression models were explored for the baseball and cricket datasets. However, especially in the baseball dataset, variance for the performance metrics was high. It was determined that these predictive models would not be useful to the targeted "customer", which in this case would be a team manager. Trends can be found but a manager would ultimately want to use a player-by-player basis for determining lineups and contracts. For our defined investigation questions, summary statistics and pearson correlation tests were especially helpful and informative. These analysis directions represent the main divergence from our proposal.

6 Conclusions

Comparing baseball and cricket was an interesting and more complex process than was initially anticipated. While many similarities exist, performance metrics, the ways in which they are calculated, and what they say about the players differ between the two sports. Collecting, processing and interpreting cricket data was a difficult challenge but the created cricket datasets could be useful for further analysis in future projects.

For baseball, exit velocity was found to be an important tracked attribute for batters when checking for correlations with scoring performance metrics. Launch angle was unimportant for performance metrics like batting average but did correlate with home run rates. The distribution of batter ages was gaussian and it was shown that increasing age does not lower the tracked mechanics of exit velocity and launch angle. Player age also did not significantly correlate with the identified performance metrics. The only metric which decreases with player age was runs per-at-bat, perhaps indicating that older players are not as good at base running. Overall, older players are not performing worse than their younger counterparts.

For pitching in baseball, ERA was the main performance metric analyzed. It was found that the values of velocity and spin rates of the two most popular pitch types, fastballs and sliders, had the biggest effect on ERA. Player ages of pitchers consistently decreases. This is perhaps because older pitchers consistently lose velocity on their pitches leading to higher ERA and lower strikeout rates. Older players don't drop off for spin rates but it appears that the lost velocity is enough to decrease their effectiveness. In general, managers should be hesitant when offering new contracts to aging pitchers.

A new feature was added to group and analyze cricket players, placing them in categories called batters, bowlers and all-arounders. It was found that teams are shifting to more all-arounders over time. These players have good strike rates and economy, key metrics for batting and bowling. Two dominant IPL teams, the Royal Challengers Bangalore and Chennai Super Kings, were compared to determine the effect of toss result and homefield advantage. It was found that homefield advantage plays a large role in determining the outcome of games. Future exploration of the created cricket datasets can help to increase knowledge of the sport and enhance analysis for related professional investigations.

Link to Github
Link to Slides

7 References

[1] "The World's Most Watched Sports". Jul 17, 2017. Media, News Freeview, News Members. https://sportforbusiness.com/the-worlds-most-watched-sports/