Kaelyn Hughes
Stockton Smith
CS 5830
Project 1

Introduction:

For the last 20 years, Major League Baseball has sponsored a free-to-enter competition known as “Beat The Streak” in which participants predict which baseball players will get at least one hit in their next game. The end goal of the competition is to make 57 correct picks in a row, which will earn the winner a grand prize of $5.6 million. Our analyses will provide relevant information to Beat The Streak participants that will assist them in making more accurate predictions, which will allow them the best chance at winning the grand prize. The significance of this accomplishment is illustrated in the following scatterplot.

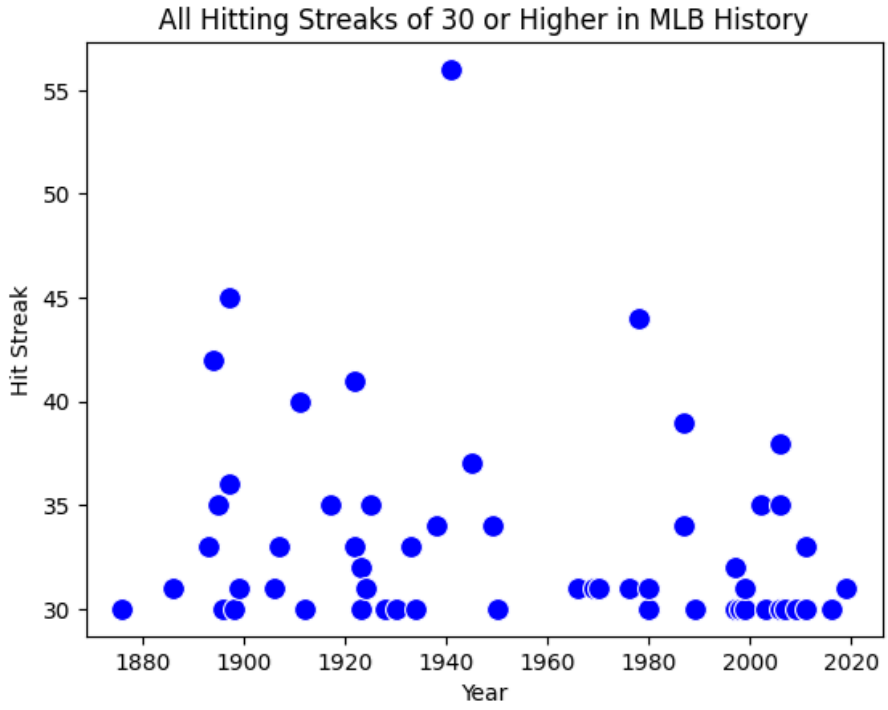

Fig. 1: Scatterplot of all notable MLB hitting streak records, showing the significance of the 1941 record of 56

The findings outlined in this report were obtained using a Major League Baseball dataset containing nearly 150 years’ worth of baseball statistics. We used this data to determine key factors in predicting the best potential hitting streaks, and our findings are displayed in the plots and figures below. The results from our analyses suggest that hitting streaks are closely tied to players’ batting average, nationality, and batting hand. These topics will be explored in more depth in the remaining sections of this report. Further information on our findings is given in our slide presentation and in our github repository.

Dataset:

The baseball dataset we used in our analyses contains extensive records on the batting, fielding, and pitching statistics of Major League Baseball players, as well as important personal and biographical information. These records span from the years 1871 to 2017. Given that our analyses are concerned with hitting streaks, we elected to primarily focus our research on the batting and biographical portions of the dataset. Finally, to ensure that the dataset was suitable for analysis, we threw out all data for any players with fewer than 50 at-bats in a season. This allowed us to only analyze professional batters (rather than be misled by data on professional pitchers, for example). We also elected to only use data from the year 2010 and onwards, as relying on data from the 1800's and early 1900's would likely be misleading when attempting to make predictions about 21st century players.

Analysis Techniques:

Once we collected our data, we performed our analyses on them using various methods, including means, scatterplots, bar charts, and probability density diagrams. Means, bar charts, and probability density diagrams were especially useful when analyzing players by category, as we were able to compute statistics using only data from specialized groups. Scatterplots were also useful when analyzing the relationship between two numerical factors. These were especially appropriate when used upon our baseball dataset; thanks to the meticulous record-keeping of sports statisticians, there was almost always data recorded for every player in each of the provided categories.

Results:

Participants of the Beat The Streak game are shown a few pieces of information about MLB players when making their selections, including a profile picture, given name, batting average, and batting hand. Given that these are the most readily accessible pieces of information for Beat The Streak participants, we have chosen them as our main points of interest in determining factors that influence hitting streaks. The results of our analyses are illustrated in the figures below.

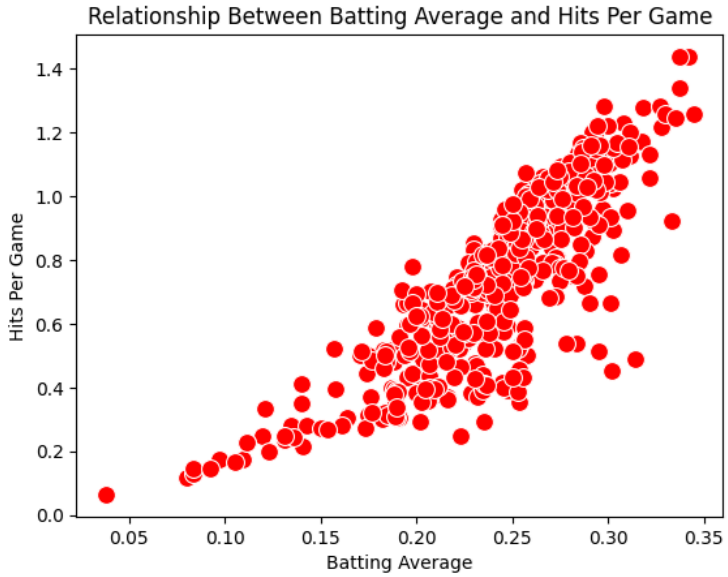

Fig. 2: Scatterplot showing a positive relationship between a player's batting average and average number of hits per game

The above relationship is to be expected, as both statistics are defined to increase with each successful hit. However, it is worth noting that a higher batting average does not guarantee a higher average of hits per game. The underlying explanation comes from the fact that batting average is also dependent on the number of at-bats; the fewer at-bats a player is given, the less likely the player becomes to get a hit. This suggests to Beat The Streak participants that, while batting average serves as a strong indicator of a potential hitting streak, it is worth researching how many at-bats players are expected to have in their next games before making a selection.

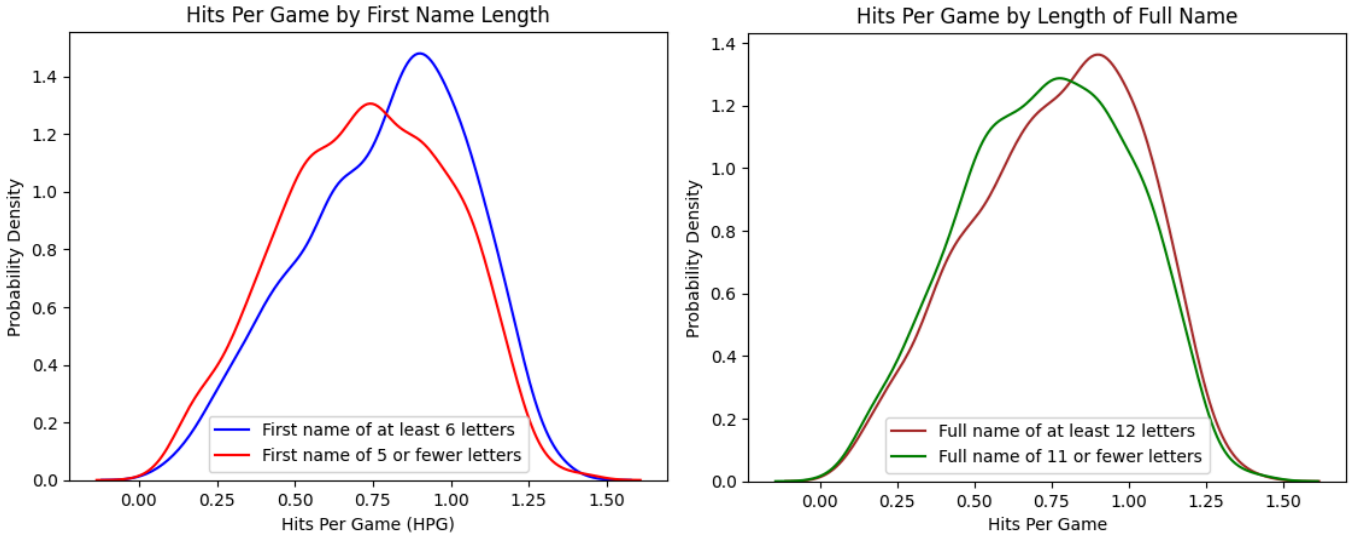

Fig. 3: A possible positive correlation exists between the length of a player's name and their average number of hits per game.

Continuing with our analyses on how information shown on the Beat The Streak app relates to hitting streaks, our observations on player names yielded some surprising results: players with longer names seem to hit slightly more hits per game than those with shorter names. When considering first names only, players with at least 6 letters averaged about 0.79 hits per game, while players with shorter names averaged about 0.73 hits. Additionally, players with combined name lengths of at least 12 letters averaged about 0.77 hits per game, while the players with fewer than 12 letters averaged about 0.73 hits.

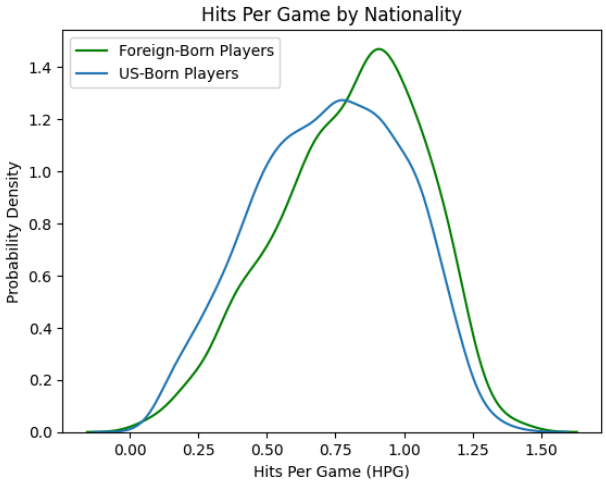

Fig. 4: Hits per game based on country of birth

In order to potentially explain the results from the name-length analysis, we performed the same test on players by nationality. We observed that players born in the United States averaged about 0.73 hits per game (which is the same average found for players with short names) while the average was about 0.8 hits per game for players born anywhere else. This suggests to Beat The Streak participants that, when attempting to decide between players with similarly high batting averages, a foreign-born player may be the advantageous selection.

Finally, we performed an analysis to determine the relationship between hitting streaks and preferred batting hand. The results are not quite as significant as those from earlier tests, but it was observed that left-handed and right-handed hitters were equally successful (about 0.75 hits per game), but seemed to be overshadowed by switch-hitters, who averaged about 0.77 hits per game. This phenomenon is demonstrated in the figure below.

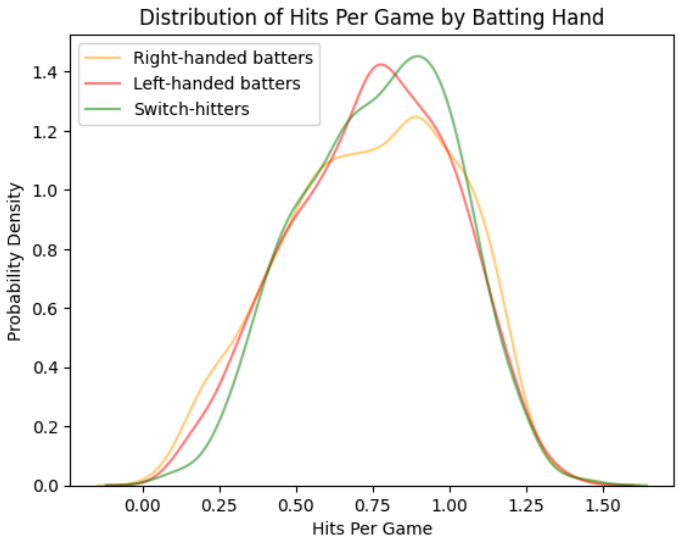

Fig. 5: How hits per game varies by which hand is used to bat

Overall, our findings suggest that certain factors may have stronger implications of a potential hitting streak than others. For participants of Beat The Streak, this means that optimal selections are closely tied to high batting averages, more at-bats, and nationalities based outside of the United States. Other factors, such as being a switch-hitter, may also be helpful when making a decision. Ultimately, there is a reason that Beat The Streak has never been won, even after more than 20 years of the competition's existence, but perhaps our findings will be useful to those who are interested in the challenge of trying to take down one of the most unbreakable records in human history.

Technical Section:

Some external research was made in order to prepare our dataset for analysis. The original baseball dataset did not provide data for hitting streaks, so all the records displayed in Fig. 1 had to be added to the batting table manually. We then merged the provided player table with the batting table, allowing all the information relevant to our analyses to exist in one location. As mentioned above, we also filtered out any data from before the year 2010, as well as any data for players who had fewer than 50 at-bats in a season. Doing this allowed our plots and figures to display accurate data entirely generated from the statistics of current professional batters in the MLB. Extensive sports recordkeeping and data collection over the course of 150 years allowed for our scatterplots and categorical charts to make effective comparisons.

When performing our analyses, we elected to observe our data through the lens of individual seasons of baseball per player rather than by individual players. Therefore, an alternative approach for collecting our findings would have been to condense our dataset so that duplicate entries for each player were combined into a single entry. This would have given less influence to players who have played for many seasons of baseball, and we might have come to different conclusions as a result. This would be a worthwhile approach to anyone seeking to replicate our experiment.