Using Text Analysis to evaluate Softball Run expectancy

By: Sean Jin, Zachary Siegel, Anna Tan
Project Supervisor: Peter Freeman *Project Advisor: Rebecca Nugent*

Carnegie Mellon University
Statistics & Data Science

Introduction & Background

Initiatives and Goals

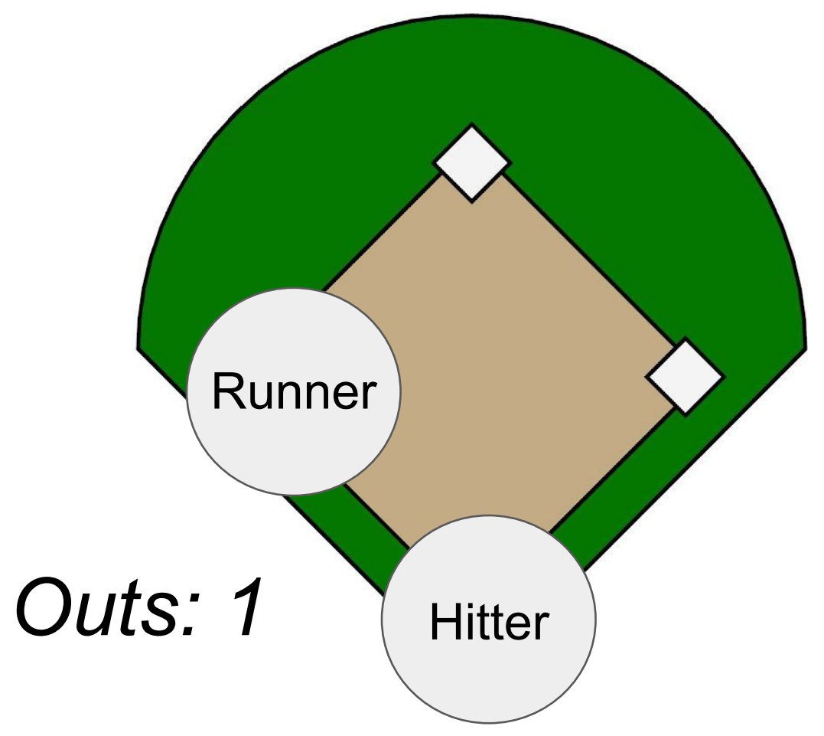
Statistical analysis has proved to be very effective in improving sport performance. For this project specifically, our team focus on analyzing the softball run expectancy.
After communicating with coach Harrison and coach Maples from Carnegie Mellon Softball team, we decide it will be helpful to analyze run expectancies from . Our main goals are :

- Create text analysis of previous play-by-play games in record;
- Exploring the run expectancy of the teams of UAA conference;
- Exploring run expectancy for any Division-III games.

Our data cleansing and text analysis are all conducted in Rstudio. As a final product, we also created an interactive R-shiny app.

Softball Rules Overview

In a game of softball, teams switch between hitting and fielding, and only the hitting team may score points, which are represented as runs. Each player from the hitting team tries to hit the ball and run around four bases without getting out in order to score a run. A hitter cannot get out if they are on one of the first three bases, and two runners cannot be on the same base. The teams switch once three outs occur. Therefore, there are 24 possible combinations of runners on the basepath and outs, which we will call a state.



Pictured on the left is an example of a state, in which there is a runner on third base with one out.

Data Retrieval & Data cleansing

Our data retrieval and cleansing consists of three steps:

- Retrieve our data from the NCAA statistics website stats.ncaa.org , which contains the softball play by play data for any Divisions up to 2020.
- Analyze the data of the complete 2018-19 season for Division III, which consists of 8043 games, and included 603,236 plays across 46 different conferences.
- Read each web page that corresponds into R studio, identified the tables in the webpage, and extracted the texts from the tables. We then recreated a new table with game ID and play number assigned to each action:

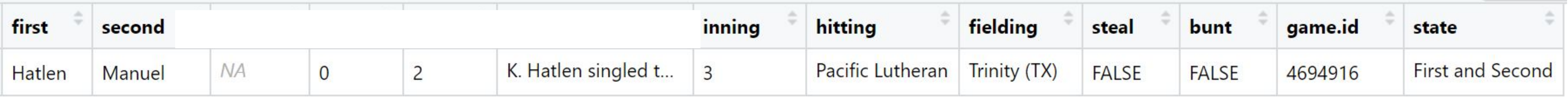
Plays	Away team	Home Team	Game ID	Play Number
D.Reyes singled to left field	St.Katherine	La Verne	4707355	1
Y.Sanchez reached first on an error by 2b	St.Katherine	La Verne	4707355	2
Mitcher Singled to right field	St.Katherine	La Verne	4707355	3
Y.Sanchez reached first on an error by 2b	St.Katherine	La Verne	4707355	4
...	...	...	..	...

Text Anlaysis & Run expectancy

Given the processed play by play tables, we use the following method to do the text analysis:

- Search for keywords and extract the "state" of the game.
- Look at the name and the action of each hitter to determine their location on the basepath as well as keep track of which teams are currently playing, the number of outs, the number of runs scored per inning, and whether there are stealing or bunting occuring.
- Keep track of inning number: if an inning is finished, we reset the states so that there is nobody on base with zero outs.

Post Processed data :

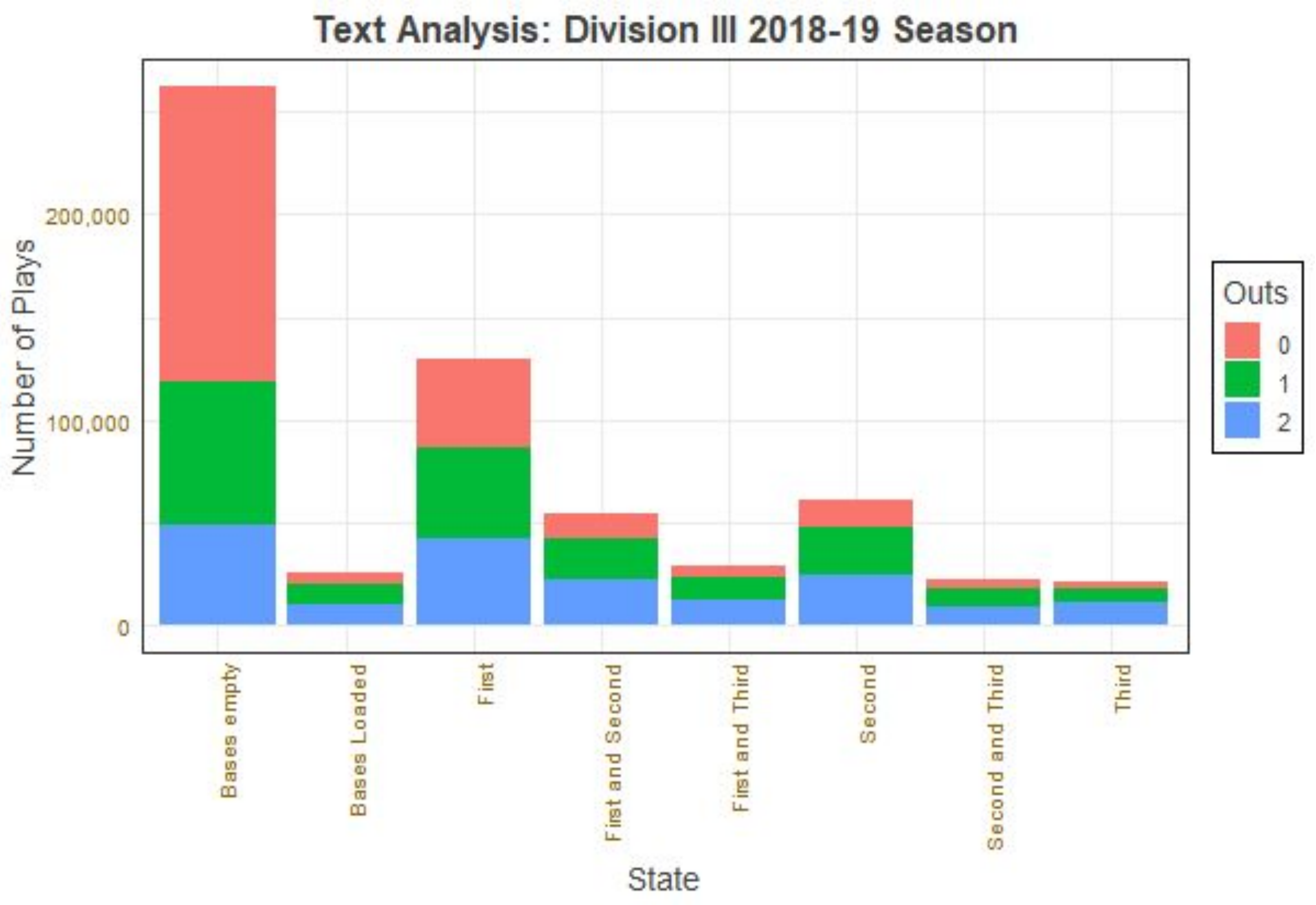


first	second					inning	hitting	fielding	steal	bunt	game.id	state
Hatlen	Manuel	NA	0	2	K. Hatlen singled t...	3	Pacific Lutheran	Trinity (TX)	FALSE	FALSE	4694916	First and Second

Text Analysis

With the post-processed data, we calculate the expected number of runs during an average inning.

Division III 2018-19 Season Sample



E[Runs per Inning] = 1.47

We are interested in the expected number of runs given any state in the game, which we expect to be higher than this value for states with more base runners and less outs, and lower for states with fewer base runners and more outs.

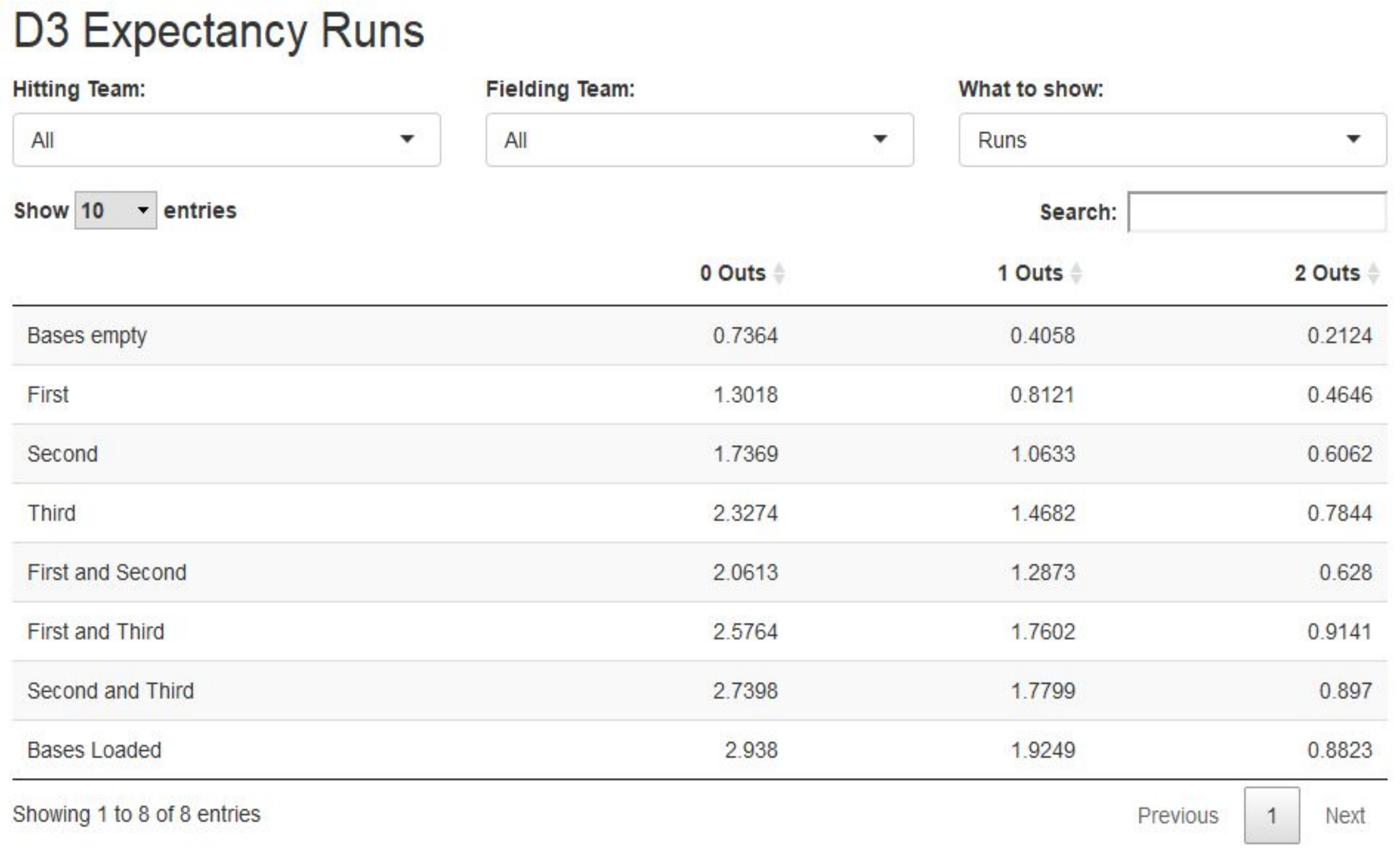
Markov Chains

We will be using Markov Chains , which contains the probability to go from state by state, to calculate the run expectancy.

We did the following steps; :

- Identified 24 different states that are based on the binary output of the 3 bases and the number of outs
- For every half inning, took the average number of runs scored from the current state until the end of the half inning, which results in the expected number runs for every state
- Created an R shiny app such that our clients are able to customize their input by selecting hitting team, fielding team, and the option to look at bunting and stealing probabilities instead.
Here is an example:

Division III Run Expectancy as shown in R Shiny App



D3 Expectancy Runs

Hitting Team: All
Fielding Team: All
What to show: Runs

Show 10 entries
Search:

	0 Outs	1 Outs	2 Outs
Bases empty	0.7364	0.4058	0.2124
First	1.3018	0.8121	0.4646
Second	1.7369	1.0633	0.6062
Third	2.3274	1.4682	0.7844
First and Second	2.0613	1.2873	0.628
First and Third	2.5764	1.7602	0.9141
Second and Third	2.7398	1.7799	0.897
Bases Loaded	2.938	1.9249	0.8823

Showing 1 to 8 of 8 entries

Conclusions

We obtained an overall expected runs per inning of 1.47. From the run expectancy chart , we can see that the expected number of runs decrease as the number of outs increase, and also tends to increase in the order of first, second, first and second, third, first and third, second and third base. Client could also go to our R shiny app to explore its features.