Jacob Gavel

MA 346

Professor Carter

April 27, 2022

Project 2 Report

https://share.streamlit.io/jakegavel/ma346\_project2/main/Project-2.py

The goal of this application was to provide more representation for an underrepresented period in the sport of baseball. As an avid baseball fan, Jackie Robinson is one of the most focused upon black athletes in the sport as he was the first to break the color barrier of the MLB. However, little is shown on those who paved the way for Robinson in the Negro Baseball League. The talent of these players proved that baseball is meant for everyone and ushered in the era of integration of the leagues. Despite their historic accomplishment, not enough is known about the athletes of the Negro Baseball League. This app is aimed to change that as fans of the MLB will be able to find athletes of the Negro League who they would’ve been fans of. The app is able to accomplish this task by taking in the favorite player of a user, either past of present, and based on that player, the app will output the player most like the MLB player in the Negro League.

Luckily, there was already a data set that has all the necessary information to create this application. The site fivethirtyeight had a baseball dataset with statistics on both MLB and NLB players. The dataset included information on the players position as well as their play style with scores for speed, power, patience, defense, and hitting. Little was needed to be done in terms of dating cleaning and prep for this project. All columns that had statistics for pitchers were dropped along with any column containing a pitcher, as only batters will be focused on for this application. If both pitchers and batters were used, there would be a lot of zero values in the data set that would skew the normalization of the statistics in a way that would affect the strength of the comparison. The data types of start and end year columns were changed to be strings as no analysis besides printing the years will be done. The player label column was also made numeric so it’s easier to split the data set based on league. Finally, generic value was substituted for all NA values. With the information now cleaned and ready on the play styles for both NLB and MLB players in a numeric format, a scoring system was able to be made. The scores were all normalized for every athlete and the data set was subseted three ways so there was a dataset for past MLB player, present MLB players, and NLB players. From this point the user will select a player from whichever subset of the data set they select, and the scores can be computed for each Negro League player that shares the same position as the selected player. The score value that was used was the distance formula for each athlete, the square root of the sum of the difference of each performance statistic squared. The row that had the lowest score value would be the closest match and the final application page showed some basic information on the player. The page showed the name of player, when they played, the skills they were best at, and a fun fact as well as a link to their Wikipedia page for further research on the player. When run by a user, this application will provide users with basic information on Negro League players that might be unknown as well as an outlet for the user to continue to further their knowledge on the league.