Levi Nickerson
109340569
CSCI 4830

Baseball Lineup Projector

Introduction:

The world of sports analytics has become growingly complex over the last couple decades. Professional and collegiate teams have begun to care about the numbers behind the sports. Along with the growth of sport analytics, daily fantasy has emerged. This program is intended to work as a bridge between the two worlds. This program tackles the difficult task of selecting an appropriate daily fantasy lineup for MLB games.

Visual analysis is crucial for this program for two reasons. First, the size of the statistical data that is produced from the MLB is too large to understand by only looking at numbers. The program compiles all of this data and computes a single "effectiveness value" for each player. Second, visualization shows more than numbers can. When looking at a large table it is impossible to correctly determine who will be successful. Visualization allows the user to view selected graphs and determine effective players.

Dataset:

I will be integrating two datasets. The first dataset is the MLB statistical data for the current season. This dataset consists of every single athlete that has made an appearance in the current MLB season. It consists of hundreds of players and 19 recorded statistics for each batter (Figure 1) and 17 recorded statistics for pitchers (Figure 3). The second dataset is the salary projections from the daily fantasy league. This dataset consists of all of the players that are playing on a selected day and their respective salaries. There is limited integration between the two sets, but both are crucial. The entirety of the MLB statistical data will be used to compute an "effectiveness value" for each player. This value is a derived attribute that will be computed from the statistical data. For batters this value is dependent on their batting average, home runs hit, on-base percentage, and many other values. For pitchers it is a simpler calculation depending primarily on their win percentage, strikeout rate, and average number of innings pitched per start.

This value for each player will then be coupled with the respected player's salary from the second dataset to complete the visualization.

The data quality of the utilized datasets is exceptional. The data should contain very few if any errors or missing points. There is not uncertainty in the data, but there is uncertainty in what the data can mean. What I mean by this is that certain values in the data can be misleading. For example, when looking at a batter it is very important to pay attention to his number of at bats. If a batter has very few at bats his statistics can be skewed to be above or below his average. A pitchers number of innings pitched is equally important. Athletes, especially MLB players, have cold and hot streaks. It is important to have a large sample size when determining many of the attributes in the data. The only cleaning of the data that is required is to parse through the data in order to limit what players are evaluated. For a batter to be evaluated they must have more than 50 at bats. A pitcher must have started more than 5 games in order to be evaluated. The two datasets are currently retrieved as csv Excel files. Prior to operation of the program the datasets have to be opened up in excel and resaved as CSV files. This is done in order for python to correctly load the datasets.

First Nam	Last Name	Team	Pos	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	SH	SF	HBP	AVG	SLG	OBP	OPS
Ozzie	Albies	ATL	2B	84	367	66	102	27	2	18	49	8	3	19	65	1	2	4	0.278	0.51	0.319	0.829
Marcus	Semien	OAK	SS	84	349	48	87	16	1	7	31	7	4	29	73	0	5	0	0.249	0.361	0.303	0.664
Francisco	Lindor	CLE	SS	84	347	75	103	27	0	23	55	11	3	39	69	3	2	5	0.297	0.573	0.374	0.947

Figure 1 - MLB Batters Data

The first four columns of the data set in Figure 1 display categorical attributes. These attributes are: first name, last name, team, and position. The remaining attributes in the data set are all quantitative. They are as follows: games, at bats, runs, hits, doubles, triples, home runs, runs batted in, stolen bases, caught stealing, bases on balls, strikeouts, sac hits, sac flies, hit by pitches, average, slugging percent, on-base percent, on-base plus slugging. These are the values that will factor into every batter's effectiveness value. The effectiveness value is computed by combining a player's statistics with the scoring breakdown from the daily fantasy site. The scoring breakdown for batters is as follows (Figure 2).

Action	Points
Single	+3
Double	+5
Triple	+8
Home Run	+10
Run Batted In	+2
Run	+2
Base on Balls	+2
Hit By Pitch	+2
Stolen Base	+5

Figure 2 - Scoring Summary for Batters

First Nam	Last Name	Team	G	GS	CG	SH	IP	H	ER	K	BB	HR	W	L	SV	BS	HLD	ERA	WHIP
Trevor	Bauer	CLE	18	18	0	0	121.3	95	33	156	37	5	8	6	0	0	0	2.45	1.09
Max	Scherzer	WAS	18	18	1	1	120.7	75	29	174	30	10	10	5	0	0	0	2.16	0.87
Corey	Kluber	CLE	18	18	1	0	119.3	91	35	120	13	17	12	4	0	0	0	2.64	0.87

Figure 3 - MLB Pitchers Data

The data for pitchers in the MLB is very similar and can be seen in Figure 3 above. The first three categorical attributes are first name, last name, and team. The remaining numerical attributes are: games, games started, games closed, shutouts, innings pitched, hits allowed, earned runs, strikeouts, bases on balls, home runs, wins, losses, saves, blown saves, holds, earned run average, and walks plus hits per inning pitched. These values will be factored into each pitcher's effectiveness value. The effectiveness value is computed by combining a player's statistics with the scoring breakdown from the daily fantasy site. The scoring breakdown for pitchers is as follows (Figure 4).

Action	Points
Inning Pitche	+2.25
Strikeout	+2
Win	+4
Earned Run Allowed	-2
Hit Against	-0.6
Base on Balls Against	-0.6
Hit Batsman	-0.6
Complete Game	+2.5
Complete Game Shutout	+2.5
No Hitter	5

Figure 4 - Scoring Summary for Pitchers

Position	Name + ID	Name	ID	Roster Po	Salary	Game Info	TeamAbbr	AvgPoints
SP	Max Scherzer (10897509)	Max Scherzer	10897509	P	14200	MIA@WAS 07/06/2018 07:05PM ET	WAS	26.96
SP	Chris Sale (10896660)	Chris Sale	10896660	P	14000	BOS@KC 07/06/2018 08:15PM ET	BOS	27.32
SP	Trevor Bauer (10897510)	Trevor Bauer	10897510	P	13600	OAK@CLE 07/06/2018 07:10PM ET	CLE	25.98

Figure 5 - MLB Players Salary Data

The data for the salary information is very simple in comparison (Figure 5). The two attributes of interest are name and their corresponding salary. The name of the athlete is used for Boolean comparison in the code to correctly determine who is playing on a selected day. The salary for each player will be coupled with the player's effectiveness value to complete the visualization.

Tasks:

The overall task of the project is to develop a program that aids in the selection of a contending lineup for daily fantasy. In terms of what we have learned so far the primary task I am concerned with is analysis. Within the analyze action I am venturing into the subsets of consume and produce. I am producing a derived attribute and then consuming the information that the attribute displays. I am deriving my own attribute from a dataset for analysis. Once this attribute has been derived I am presenting it in a way that allows the user to look for outliers. Identifying outliers is at the core of what this program does. It is intended to highlight athletes who have a low salary in comparison to their effectiveness value. This effectiveness value is the derived attribute that is dependent on many different factors. The effectiveness value for batters is computed using the following equation:

$$Effectiveness\ Value$$
$$P(single) * 3 + P(double) * 5 + P(triple) * 8 + P(home\ run) * 10 + P(RBI) * 2$$
$$+ P(run) * 2 + P(BB) * 2 + P(hbp) * 2 + P(SB) * 5$$

For pitchers the effectiveness value is computed using the equation below:

$$Effectiveness\ Value$$
$$P(innings) * 2.25 + P(strikeout) * 2 + P(win) * 4 + P(era) * (-2) + P(hit) * (-0.6)$$
$$+ P(base\ on\ balls) * (-0.6) + P(hit\ baseman) * (-0.6) + P(CG) * 2.5$$

The probabilities for certain events are computed and then multiplied by their given point value from the daily fantasy site. These values are then summed up as the effectiveness value.

The secondary task that is most important to me is to enjoy the visualization. I am an avid sports fan and have always been fascinated by the numbers behind sports. This project is an enjoyable way for me to dive deeper into baseball and see how well I can predict outcomes. I also believe that there are many sports fans that are interested in a project of this realm.

The program will initially ask the user for an input. This input should be the position of interest. The input options are the positions in baseball: catcher, first base, second base, shortstop, third base, outfielders, and pitchers. Once a position has been selected the program will determine the athletes at that position that are playing on the current day. The visualization will then be generated and the user will be able to highlight athletes that have a high effectiveness value and a low salary. In terms of tasks, this visualization is concerned with producing and consuming. It is also the visualization that is used to identify outliers. Finding outliers is the core of the program and this visualization provides enough information to correctly determine these outliers.

The second visualization will be determined by user inputs. This visualization is intended to put the decision making into the user's hands and let them determine the experience. It will produce lineups that are projected to be successful according to a desired trait. For batters the traits are: Average Heavy, Home Run Heavy, On Base Heavy, and Run Heavy. For pitchers the traits are: Innings Heavy, Strikeout Heavy, and Win Heavy. These selections allow the user to tailor to the style of baseball they would potentially like to watch.

The user of this system will be able to generate graphics that will allow them to pick an ideal lineup. The fun of daily fantasy is having the opportunity to flex your baseball knowledge, but it is impossible to know every lineup and matchup. The program is intended to step in and perform the heavy lifting. The user can input the two data sets and the program will output various things. First, the program will compute an effectiveness value for each player. This is a value that is dependent on a players batting or pitching stats. Second, this value coupled with a

player's salary will be used to create a 3D visualization. This 3D visualization will allow the user to delve into the data and select outliers that may be potential good picks for the day. Finally, the program will produce is theoretical lineups based on user inputted traits.

Visual Designs:

The primary visualization used is a 2D graph of athletes Effectiveness Values versus their Salary. This visualization can be seen below in Figures 6 and 7.

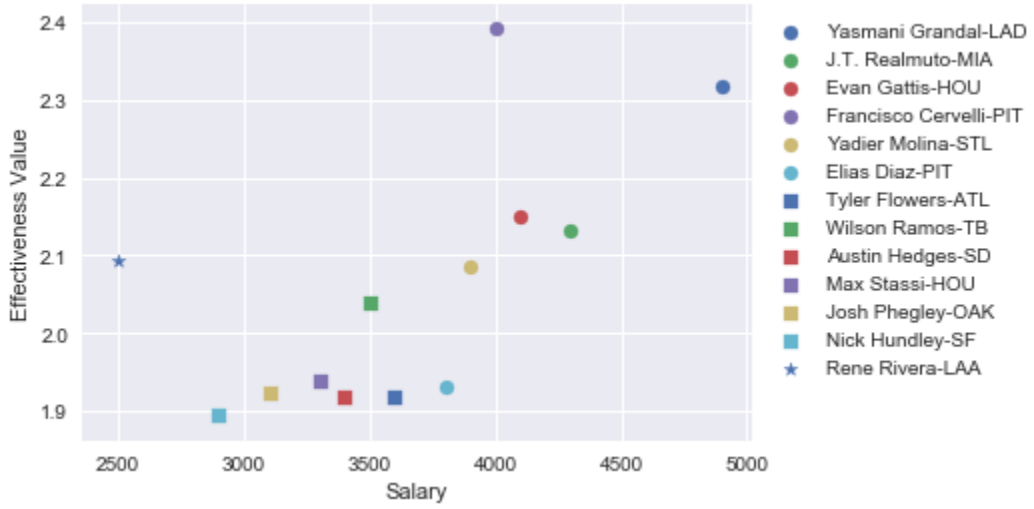

Figure 6 - Catcher Visualization

This figure shows the effectiveness value for catchers that were playing on the selected day. The visualization does not show a selected number of athletes, but rather it selects athletes based on a percentile. Currently the program selects the 75^{th} percentile and plots the results. Graphs for every position can be generated according to user inputs. This is the visualization that is used to identify outliers. Figure 7 below shows the graph for the third baseman for the selected day.

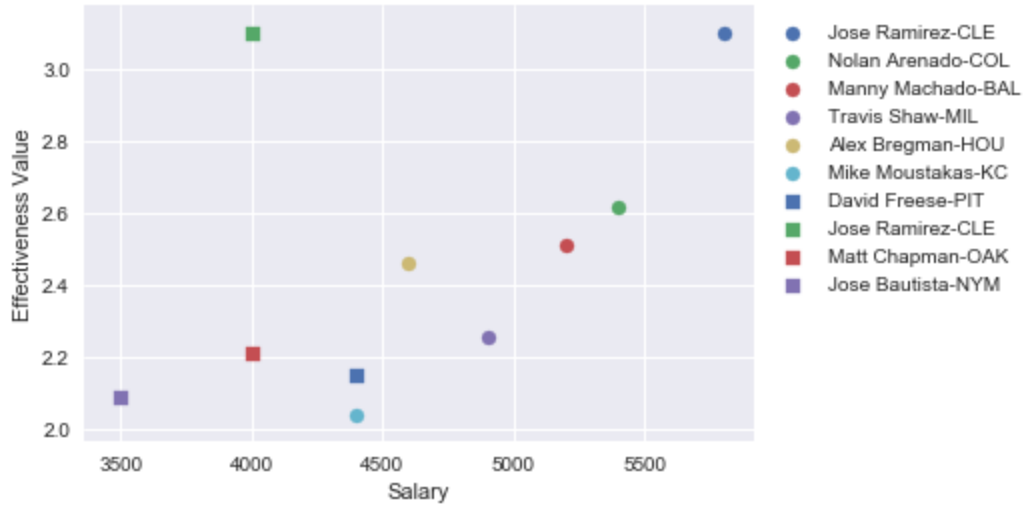

Figure 7 - Third Baseman Visualization

These figures allow the user to quickly determine effective athletes. For example, when looking at Figure 6 it can be identified that Rene Rivera is a potential candidate. The same can be said for Jose Ramirez when looking at Figure7. These athletes show a high effectiveness value for a limited salary. This is the purpose of the program. Daily fantasy comes down to selecting effective players at every position.

In Figures 8 and 9 below, the secondary visualization can be seen. This visualization goes hand in hand with the secondary task of the program. This visualization is intended to bring enjoyment to the user's lineup. This visualization is produced from the user's input.

```
Home Run Heavy Lineup
Catcher:  Evan Gattis Salary:  4100
First Base:  Matt Carpenter Salary:  5200
Second Base:  Max Muncy Salary:  5000
Short Stop:  Francisco Lindor Salary:  5800
Third Base:  Jose Ramirez Salary:  4000
Outfield:  J.D. Martinez Salary:  5800
```

Figure 8 - Batter Lineup Projector

```
What kind of Pitcher would you like to see? (Innings, Strikeout, Win, or exit) Innings
Innings Heavy Pitchers:
Pitcher:  Max Scherzer Salary:  13000
Pitcher:  Corey Kluber Salary:  10900

What kind of Pitcher would you like to see? (Innings, Strikeout, Win, or exit) Strikeout
Strikeout Heavy Pitchers:
Pitcher:  Max Scherzer Salary:  13000
Pitcher:  Gerrit Cole Salary:  12500

What kind of Pitcher would you like to see? (Innings, Strikeout, Win, or exit) Win
Win Heavy Pitchers:
Pitcher:  Max Scherzer Salary:  13000
Pitcher:  Luis Severino Salary:  11900
```

Figure 9 - Pitcher Lineup Projector

This visualization allows the user to tailor their lineup to the viewing experience they are seeking. If a user wants to watch some games with their friends and want to have an exciting lineup, they can select batters from the Home Run Heavy projection. This visualization is intended to produce a wider spectrum of information for the user to view.

Results:

The program runs and correctly produces both visualizations. Prior to running the program the datasets need to be downloaded, opened in excel, and resaved. This resave procedure allows python to correctly load in the data. Currently the program takes about 20 seconds to correctly parse through the data and compute the effectiveness value for each athlete. After that the user can input what positions they would like to view. The visualizations produce the 75^{th} percentile of athletes at their respective positions. The user can then view the graphs and identify outliers or potential candidates for their fantasy lineup. After the graphs the user can select various lineups to choose from. The user is given further information concerning athletes that are more prone to certain actions. The program effectively aids the user in producing a baseball lineup that can compete and bringing enjoyment to the user.