Levi Nickerson
109340569
CSCI 4830
Project Dataset

Project Dataset

Data Quality and Cleaning:

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.

Deriving Attributes and/or Integrating Multiple Datasets:

I will be integrating two datasets. The first dataset is the MLB statistical data for the current season. The second dataset is the salary projections from the daily fantasy league. 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.

Transformation in Proper Format:

Transforming the data set format is not something that I am considering at this time. The two datasets are currently retrieved as csv Excel files which is the desired type. Unless somewhere down the line a better format is determined I will not have to transform the data sets.

Data Abstraction:

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

The first four columns of the data set 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.

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

The data for pitchers in the MLB is very similar. 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.

Position	Name + ID	Name	ID	Roster Po	Salary	Game Info	TeamAbb	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

The data for the salary information is very simple in comparison. The two attributes of interest are name and their corresponding salary. The salary for each player will be coupled with the players effectiveness value to complete the visualization.