PART IV: Cinematography & Mathematics

AGE RANGE: 16 – 18

Statistics (Source: @Lukas from Pexels.com)

TOOL 38: PROBABILITY AND STATISTICS THROUGH THE MOVIE "MONEYBALL"

SPEL – Sociedade Promotora de Estabelecimentos de Ensino

TheArtOf Maths

Co-funded by the Erasmus+ Programme of the European Union

Educator’s Guide

Title: Probability and Statistics through movie ”Moneyball”

Age range: 16 – 18 years old

Duration: 2 hours

Mathematical concepts: Statistics, Probabilities

Artistic concepts: Sabermetrics

General objectives: To provide students with concept of probability theory and statistics.

Instructions and Methodologies: Show the excerpt of the movie Moneyball in which the sabermetrics concept is shown (cf. link on “Learn More…” section) and suggest students to watch the full movie at home;

Resources: A pen and a calculator.

Tips for the educator: For a smoother understanding by the students, have them understand some basic Baseball rules and positions beforehand.

Learning Outcomes and Competences: At the end of this tool, the student will be able to:

- Assess information and use it to solve equations involving probabilities;
- Understand how statistics can be used to predict an outcome of an event.

Debriefing and Evaluation:

Write 3 aspects you liked about this activity:	1. 2. 3.
Write 2 aspects that you have learned	1. 2.
Write 1 aspect for improvement	1.

Introduction

Sometimes we find aspects related to Mathematics in television series or movies. In such cases, sometimes these Mathematical concepts are not given much importance, because they do not influence the story itself. However, there are a few cases in which they do.

Some examples include: "21" (USA, 2008), by Robert Luketic; "Proof" (USA, 2005), by John Madden; "A Beautiful Mind" (USA, 2001), by Ron Howard; "Enigma" (USA, 2001), by Michael Apted; "Pi" (USA, 1988), by Darren Aronofsky; "Good Will Hunting" (USA, 1997), by Gus Van Sant and "Cube" (Canada, 1997), by Vincenzo Natali.

In this tool, the movie “Moneyball” (USA, 2011), by Bennet Miller, will be discussed and its mathematical concepts, such as probabilities and statistics, will be covered.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Moneyball

Moneyball (2011) is an American sports film in which an account and general manager of Oakland Athletics baseball team tries to assemble a competitive team for the 2002 season with a very limited budget.

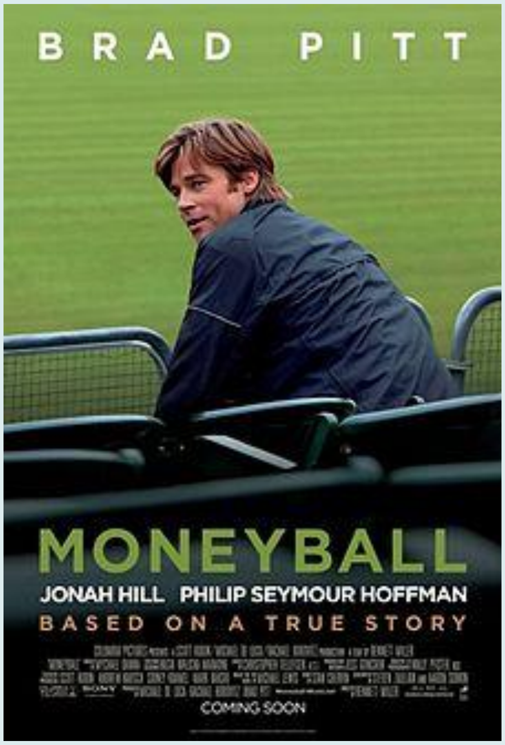


Fig. 1 – Moneyball (2011) movie poster
(Source:https://pt.wikipedia.org/wiki/ Moneyball)

In the film, after losing 3 key players, Oakland's general manager Billy Beane (played by Brad Pitt) and his assistant Peter Brand (played by Jonah Hill) resort to an unorthodox sabermetric approach in order to scout underrated baseball players. Underestimated due to biased reasons (such as age, appearance and personality), these players are overlooked by big teams, which makes them affordable for the low-budget Oakland Athletics to invest on.

Facing heavy resistance by the original, old-fashioned Oakland scouts, who lessen this approach arguing that their experience and knowledge in baseball has far more value than any statistics, Beane ignores their objections and forms a team following Peter Brand's sabermetric data statistics.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Glossary

Sabermetrics – A term coined by baseball statistician and writer Bill James. Derives from Society for American Baseball Research. It is a method that collects and analyses relevant in-game baseball stats in order to evaluate players' and teams' performance in every aspect of the game, such as:

Standard Batting/Fielding – In baseball, two opposing teams take turns in batting and fielding.

→ **Batting** - the act of hitting the ball when thrown by an opponent's pitcher. The player occupying this position is known as Hitter (H);

→ **Fielding** - the positions in which a team is spread around a baseball field. There are 9 positions: the pitcher (P) and the catcher (C), which occupy fixed positions, and the first baseman (1B), second baseman (2B), third baseman (3B), shortstop (SS), left fielder (LF), center fielder (CF) and right fielder (RF), which may move around freely.

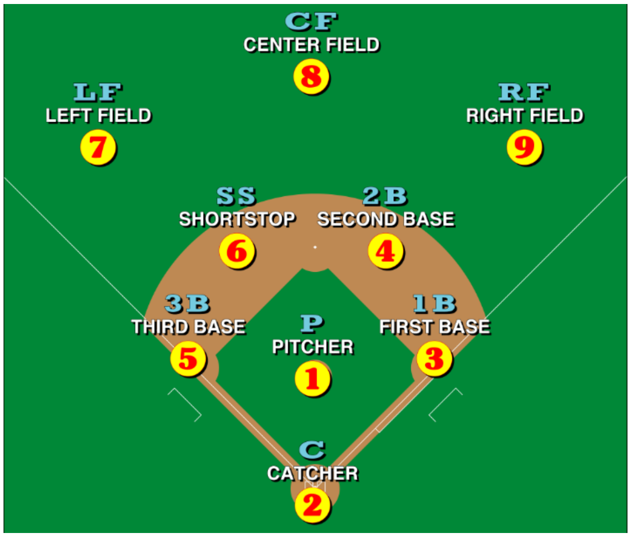


Fig. 2 – Baseball positions (Source: By Michael J, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=40095322)

Standard/Advanced stats – Statistics from player/team's performance in any sport. Below you can find the baseball statistic terms covered in this tool:

→ **Assist (A)** – Occurs when a defensive player touches the ball before a putout is recorded by another fielder;

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

→ **At Bats (AB)** – Occurs when a batter reaches the base by the fielder's choice, hit or error;

→ **Base on Ball (BB)** – Also known as Walks. Occurs when a hitter receives four pitches called out as "Ball" by the umpire (baseball referee);

→ **Caught Stealing (CS)** – A foul that occurs when a baserunner attempts to advance from one base to the other before the ball is hit and then is tagged out by a fielder while making that attempt;

→ **Games played (G)** – Total of games played by a single player. A player is credited with a Game played if he appears at any point in a game. If a player has 162 G, it means he played in the whole season's game;

→ **Hit (H)** – When a hitter (also known as batter) strikes the ball into a fair territory and reaches base. Hits can be singles, doubles, triples and home runs;

→ **Plate Appearances (PA)** – Occurs every time a player completes a turn at batting with a hit, walk, out or reaching base on an error;

→ **Putout (PO)** – The act of physically completing an out, whether by stepping on the base, tagging a runner, catching a batted ball, or catching a third strike;

→ **Run (R)** – Occurs when a player crosses the plate;

→ **Stolen Bases (SB)** – Occurs when a baserunner successfully reaches the next base when the pitcher is throwing a pitch;

→ **Total Bases (TB)** – The total number of bases a player has gained with hit.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

The Math behind Moneyball

In the movie Moneyball, at a certain point, when making the team's projections for the 2002 season, Peter Brand states that for the team to make it to the playoffs it must win at least 99 out of 162 games. To reach to this value, he projects the minimum number of runs that needs to be scored and the maximum of runs that can be allowed.

To come up with these results, he uses one of the equation originally developed by sports statistician Bill James known as "Pythagorean Winning Percentage", which results in a team's approximate winning ratio based on the runs scored and runs allowed. The equation goes as follows:

$$\mathbf{Winning\ Ratio} = \frac{\mathbf{Runs\ scored}^2}{\mathbf{Runs\ scored}^2 + \mathbf{Runs\ allowed}^2}$$

For the 2002 season, Peter projects that the team should score at least 814 runs and allow no more than 645, which results in the following:

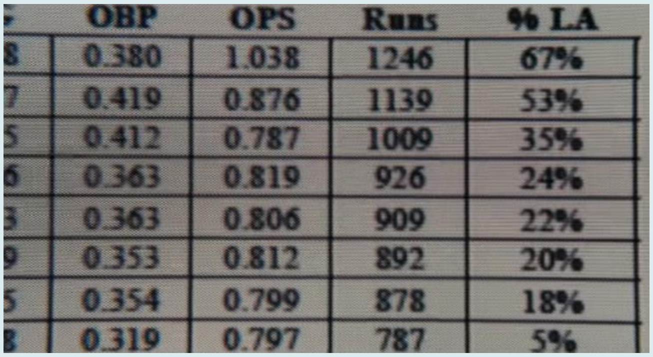
$$\mathbf{Winning\ Ratio} = \frac{\mathbf{814^2}}{\mathbf{814^2 + 645^2}} = \frac{\mathbf{662596}}{\mathbf{1078621}} = \mathbf{0.614299\%}$$

The win ratio is given in percentage and, when multiplied by the number of games in a baseball season (162), results in an approximate number of games that the team would have to win in order to make it to the playoffs.

0.614299% x 162 = 99.516438 games

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Peter then shows a database that he compiled with information on individual players on their average gaming stats (Fig. 3), in which they will be working out in order to find the most cost-efficient players.



	OBP	OPS	Runs	% LA
8	0.380	1.038	1246	67%
7	0.419	0.876	1139	53%
5	0.412	0.787	1009	35%
6	0.363	0.819	926	24%
3	0.363	0.806	909	22%
9	0.353	0.812	892	20%
5	0.354	0.799	878	18%
8	0.319	0.797	787	5%

Fig. 3 – Baseball player's sabermetric stats database shown by Peter (Source: Movie "Moneyball")

Despite only having 1/3 of the payroll of big market teams such as the New York Yankees, who were champions in the American League East division, The Oakland A's tied with 103 wins in the regular season, clinched the record of 20 consecutive victories in the American League and were champions in their division (American League West).

Even though they were eliminated in the postseason, this sabermetric approach changed an entire industry forever by using maths and statistics.

Who is Bill James?

George William James (born 1949), is an American baseball writer, historian and statistician mostly known for introducing the sabermetrics statistics method.

Besides the before mentioned "Pythagorean Winning Percentage", other statistical innovations that Bill James introduced include Runs Created (RC), Range Factor (RF) and Secondary Average (SecA):

Fig. 4 – Bill James, in 2010
(Source: https://en.wikipedia.org/wiki/Bill_James)

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Runs Created: a statistic that estimates an offensive contribution of a team/player to the runs scored in-game. This method can also be used as a means to obtain an approximate number of runs that a team will score when it is batting. The formula is as followed:

$$\mathbf{RC} = \frac{\mathbf{TB} * (\mathbf{H} + \mathbf{BB})}{\mathbf{PA}}$$

Where:

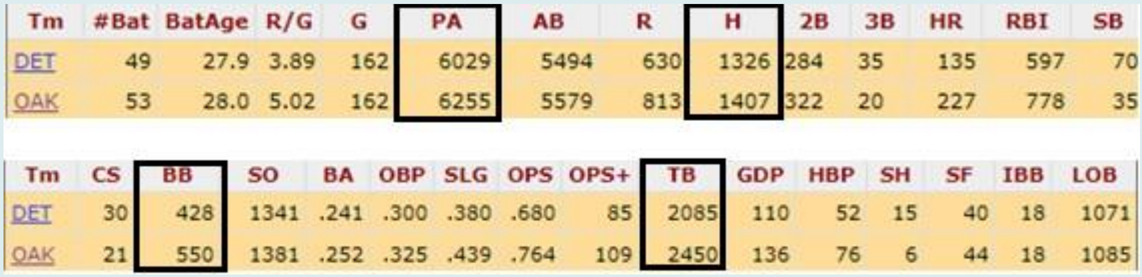
TB = Total Bases;

H = Hits;

BB = Base on Balls / Walks;

PA = Plate Appearances.

Consider the following 2018 MLB Statistics from Detroit Tigers (DET) and Oakland Athletics (OAK) (from baseball-reference.com):



Tm	#Bat	BatAge	R/G	G	PA	AB	R	H	2B	3B	HR	RBI	SB
DET	49	27.9	3.89	162	6029	5494	630	1326	284	35	135	597	70
OAK	53	28.0	5.02	162	6255	5579	813	1407	322	20	227	778	35

Tm	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	LOB
DET	30	428	1341	.241	.300	.380	.680	85	2085	110	52	15	40	18	1071
OAK	21	550	1381	.252	.325	.439	.764	109	2450	136	76	6	44	18	1085

Fig. 5 – Statistics from Detroit Tigers and Oakland Athletics from the 2018 MLB season

(Source: https://www.baseball-reference.com/leagues/MLB/2018.shtml)

Calculating the Runs Created by the OAKs:

$$\mathbf{RC\ (OAK)} = \frac{\mathbf{2450} * (\mathbf{1407} + \mathbf{550})}{\mathbf{6255}} = \mathbf{766.53}$$

According to the calculations, the Oakland Athletics should have created around 767 runs. In fact, Oakland Athletics actually scored 813 runs. By

If 813 runs correspond to 100%, then 767 runs correspond to 94,34%, which means there is a minimal 5,6% deviation.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Let us do the same for the DETs:

$$\mathbf{RC\ (DET)} = \frac{\mathbf{2085 * (1326 + 428)}}{\mathbf{6029}} = \mathbf{606.58}$$

Based on the statistics provided, there should have been created around 607 runs by the team in the past '18 season. Detroit Tigers ended the season with 630 runs. Once again, there is a just minimal percentage off (3,6%).

This calculation can also be applied to individual players and is useful to verify how well a Hitter has performed his job: the creation of runs.

Secondary Average: an improved version of the Batting Average equation. Whilst still operating under the Batting Average principles, Secondary Average also covers a player's power (extra bases), eye (walks) and speed (stolen bases). Its formula attempts to measure an overall offensive effectiveness of a player/team and is represented as follows:

$$\mathbf{SecA} = \frac{\mathbf{BB + (TB - H) + (SB + CS)}}{\mathbf{AB}}$$

Where:

BB = Base on Balls/Walks;

TB = Total Base;

H = Hits;

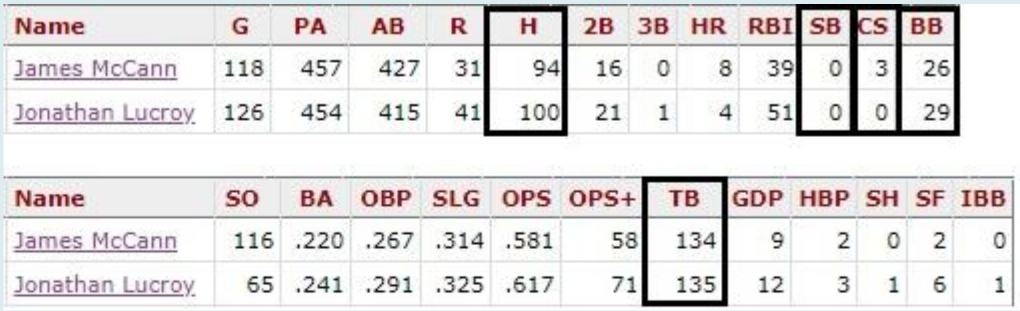
SB = Stolen Bases;

CS = Caught Stealing;

AB = At Bats.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Observe the Standard Batting stats from Catchers James McCann, from the DETs, and Jonathan Lucroy, from the OAKs.



Name	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB
James McCann	118	457	427	31	94	16	0	8	39	0	3	26
Jonathan Lucroy	126	454	415	41	100	21	1	4	51	0	0	29

Name	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB
James McCann	116	.220	.267	.314	.581	58	134	9	2	0	2	0
Jonathan Lucroy	65	.241	.291	.325	.617	71	135	12	3	1	6	1

Fig. 6 – Statistics from the 2018 MLB season
(Source: https://www.baseball-reference.com/players/m/mccanja02.shtml
and https://www.baseball-reference.com/players/l/lucrojo01.shtml)

Using the Secondary average equation, we have:

$$\mathbf{SecA\ (James\ McCann)} = \frac{\mathbf{26 + (134 - 94) + (0 + 3)}}{\mathbf{427}} = \mathbf{0.161}$$

$$\mathbf{SecA\ (Jonathan\ Lucroy)} = \frac{\mathbf{29 + (135 - 100) + (0 + 0)}}{\mathbf{454}} = \mathbf{0.140}$$

The resulting number rounded to the thousandth place represents a player's secondary average. In this case, James McCann has a better overall effectiveness, which, theoretically, means that, in the long term, he is more effective offensively.

Range Factor: a statistic that quantifies the contribution of a player at a given defensive position. The equation is as follows:

$$\mathbf{RF} = \frac{\mathbf{A + PO}}{\mathbf{G}}$$

Where:

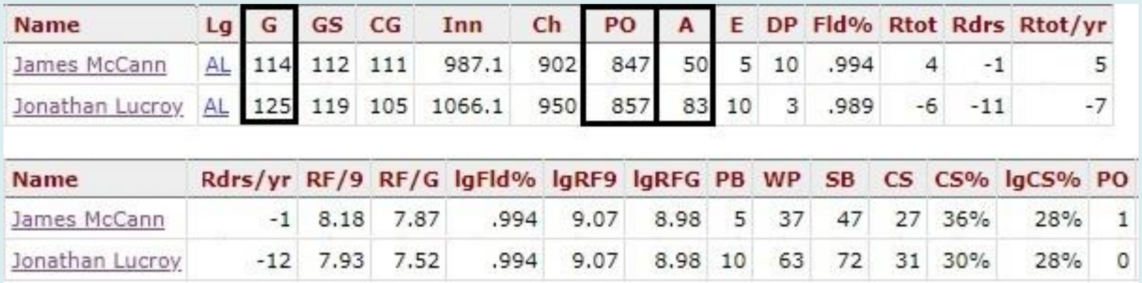
A =Assists;

PO = Putouts;

G = Games played.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Consider the Standard Fielding stats from the two same players:



Name	Lg	G	GS	CG	Inn	Ch	PO	A	E	DP	Fld%	Rtot	Rdrs	Rtot/yr
James McCann	AL	114	112	111	987.1	902	847	50	5	10	.994	4	-1	5
Jonathan Lucroy	AL	125	119	105	1066.1	950	857	83	10	3	.989	-6	-11	-7

Name	Rdrs/yr	RF/9	RF/G	lgFld%	lgRF9	lgRFG	PB	WP	SB	CS	CS%	lgCS%	PO
James McCann	-1	8.18	7.87	.994	9.07	8.98	5	37	47	27	36%	28%	1
Jonathan Lucroy	-12	7.93	7.52	.994	9.07	8.98	10	63	72	31	30%	28%	0

Fig. 7 – Statistics from the 2018 MLB season
(Source: https://www.baseball-reference.com/teams/DET/2018.shtml
and https://www.baseball-reference.com/teams/OAK/2018.shtml)

The same players have produced the following results while in a fielding position:

$$\textbf{RF (James McCann)} = \frac{\mathbf{50+847}}{\mathbf{114}} = \mathbf{7,86}$$

$$\textbf{RF (Jonathan Lucroy)} = \frac{\mathbf{83+857}}{\mathbf{125}} = \mathbf{7,52}$$

James McCann's Range Factor is higher than Jonathan Lucroy's. In other words, James McCann has a significantly more relevant defensive play.

Like in every factor analysis, it is important to understand that the greater the sample size/data used, the more accurate and precise the results will be.

Many other formulas were developed by Bill James that account for many other standard and advanced stats; along the time, some of them were refined and others created by other statisticians. Whilst these were initially conceived for baseball games, they have since been developed and adapted in order to be able to produce equivalent results in other sports.

In 2006, American weekly news magazine Time nominated Bill James as one of the Top 100 most influential people in the world.

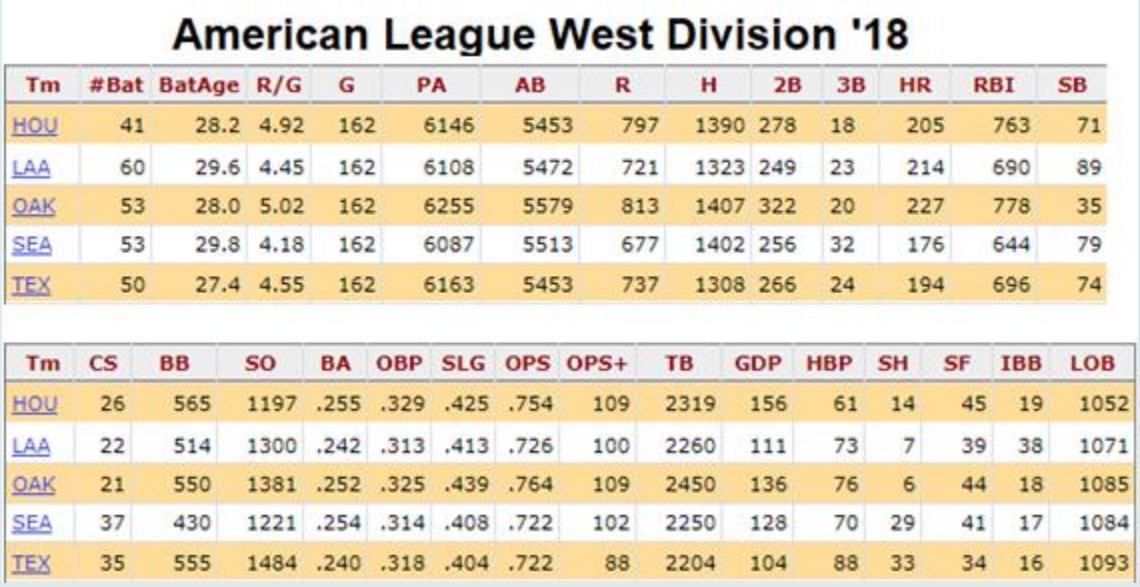
This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

TASKS

TASK1

1. **The MLF American League West division is composed by 5 teams: Houston Astros (HOU), Los Angeles Angels (LAA), Oakland Athletics (OAK), Seattle Mariners (SEA) and the Texas Rangers (TEX).**

   Observe the table below with 73738 the ALW division from season '18 and solve the questions using sabermetric approaches mentioned in this tool**.**



American League West Division '18

Tm	#Bat	BatAge	R/G	G	PA	AB	R	H	2B	3B	HR	RBI	SB
HOU	41	28.2	4.92	162	6146	5453	797	1390	278	18	205	763	71
LAA	60	29.6	4.45	162	6108	5472	721	1323	249	23	214	690	89
OAK	53	28.0	5.02	162	6255	5579	813	1407	322	20	227	778	35
SEA	53	29.8	4.18	162	6087	5513	677	1402	256	32	176	644	79
TEX	50	27.4	4.55	162	6163	5453	737	1308	266	24	194	696	74

Tm	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	LOB
HOU	26	565	1197	.255	.329	.425	.754	109	2319	156	61	14	45	19	1052
LAA	22	514	1300	.242	.313	.413	.726	100	2260	111	73	7	39	38	1071
OAK	21	550	1381	.252	.325	.439	.764	109	2450	136	76	6	44	18	1085
SEA	37	430	1221	.254	.314	.408	.722	102	2250	128	70	29	41	17	1084
TEX	35	555	1484	.240	.318	.404	.722	88	2204	104	88	33	34	16	1093

Fig. 8 – Statistics from American League West Division from the 2018 MLB Division
(Source: https://www.baseball-reference.com/leagues/MLB/2018.shtml)

1.1 Calculate the approximate number of Runs Created by all 5 teams.

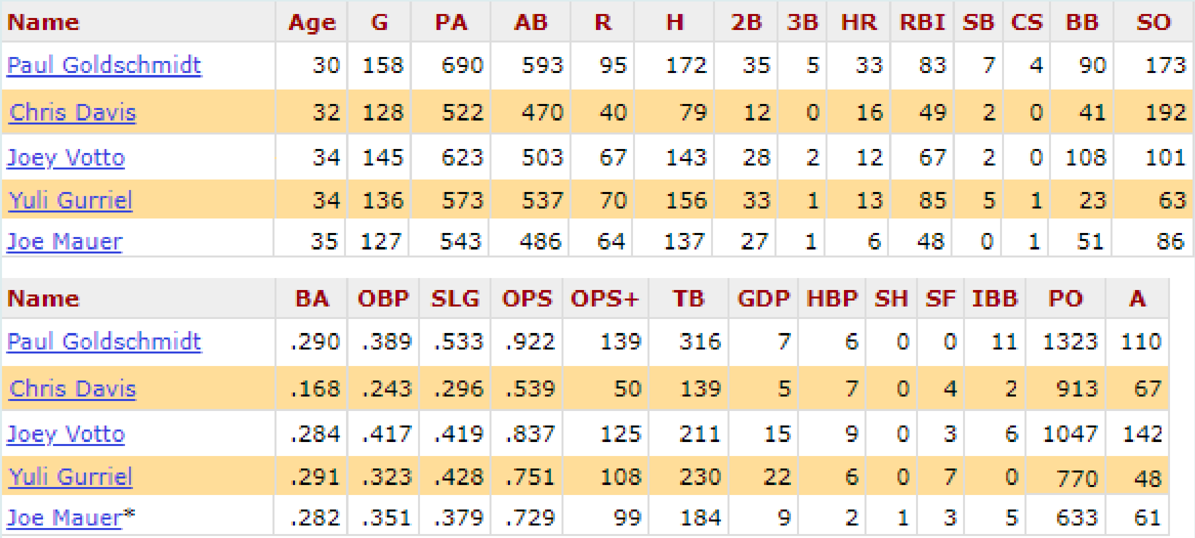
1.2 Compare the results obtained with the numbers from the table.
How big was the deviation?

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

TASK 2

Consider the following scenario:

The Oakland Athletics just had their top First Base player drafted to another team. In order to replace his position in the field, they have searched on the available First Base players in the market that had performed well in the '18 season. They have concluded that the players in the table below are fit for the job, but can only hire one of them.



Name	Age	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO
Paul Goldschmidt	30	158	690	593	95	172	35	5	33	83	7	4	90	173
Chris Davis	32	128	522	470	40	79	12	0	16	49	2	0	41	192
Joey Votto	34	145	623	503	67	143	28	2	12	67	2	0	108	101
Yuli Gurriel	34	136	573	537	70	156	33	1	13	85	5	1	23	63
Joe Mauer	35	127	543	486	64	137	27	1	6	48	0	1	51	86

Name	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	PO	A
Paul Goldschmidt	.290	.389	.533	.922	139	316	7	6	0	0	11	1323	110
Chris Davis	.168	.243	.296	.539	50	139	5	7	0	4	2	913	67
Joey Votto	.284	.417	.419	.837	125	211	15	9	0	3	6	1047	142
Yuli Gurriel	.291	.323	.428	.751	108	230	22	6	0	7	0	770	48
Joe Mauer*	.282	.351	.379	.729	99	184	9	2	1	3	5	633	61

Fig. 9 – Player Statistics from American League West Division from the 2018 MLB Division
(Source: https://www.baseball-reference.com/players/)

2.1 The Oakland Athletics wants an effective 1B. According to the statistics above, which one is likely to be more overall effective? Find it out using the Secondary Average equation.

2.2 Calculate the Range Factor of the player with the best Secondary Average statistics.

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

LEARN MORE…

Moneyball (2011) movie plot

https://www.imdb.com/title/tt1210166/?ref_=nv_sr_1

Sabermetrics on Moneyball

https://www.youtube.com/watch?v=KWPhV6PUr9o

Inside the stats that created 'Moneyball'

http://www.espn.com/espnw/news-commentary/article/7577771/stats-created-moneyball

Standard metric stats

http://m.mlb.com/glossary/standard-stats

Advanced metric stats

http://m.mlb.com/glossary/advanced-stats

Baseball positions

https://en.wikipedia.org/wiki/Baseball_positions

Baseball rules

http://www.rulesofsport.com/sports/baseball.html

Database with all-time baseball players, teams, scores and leaders.

https://www.baseball-reference.com/

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.