Group 12

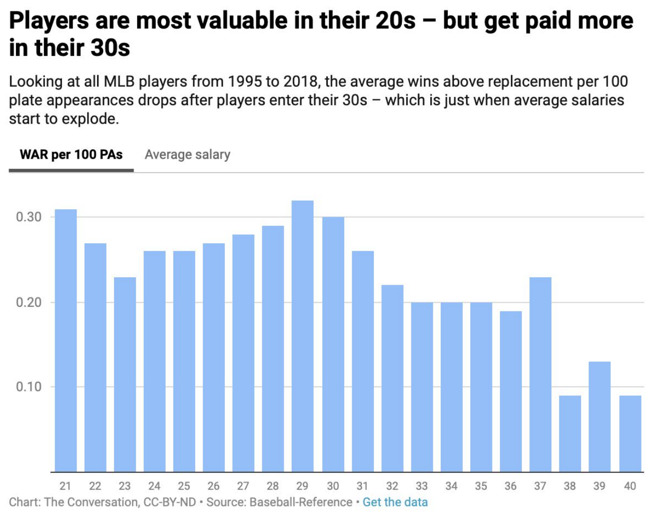
Matthew Caruso, Jon Saksvig, Conner Snavely, Chris Weigand

Zoom Link

Processing Steps and Data Sources

- Plan:
    - What factors bring in most money for a baseball team?
        - Ticket sales/attendance, attendance per game
            -
        - Playoff appearances, playoff wins
        - All-Star players on roster
        - Merchandise sales
            - Correlated to having star players?
    - Analyzing how both payroll and franchise value correlates to team wins. What aspects should teams pursue to win games?
        - Create 2 scatter plots with payroll/franchise value on x-axis and win percentage on y-axis. Analyze correlation.
        - Create map visualizing these statistics
    - Considering if a player's payroll is consistent with his performance?
        - Developing a model that uses various in game stats to determine how much a player should be paid
        - Adjusting predicted payroll for price level of city and comparing it to actual payroll
        - Analyze correlation between predicted payroll and actual payroll
            - Correlation strength and visual via scatterplot

Players more valuable in 20s but paid more in 30s



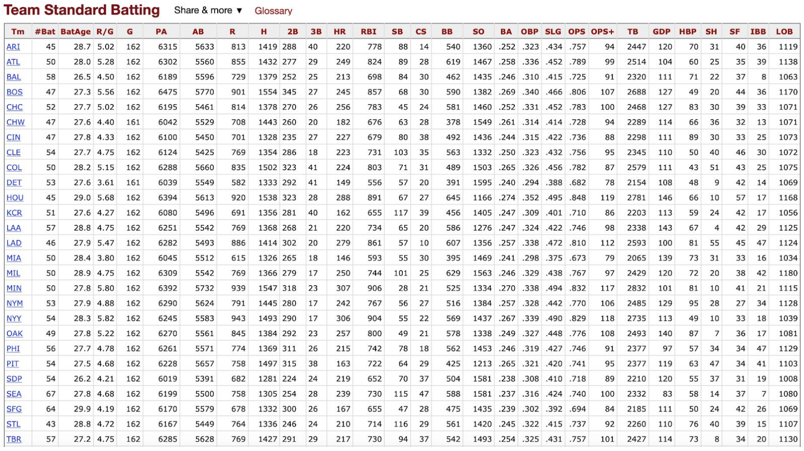
Players are most valuable in their 20s – but get paid more in their 30s

Looking at all MLB players from 1995 to 2018, the average wins above replacement per 100 plate appearances drops after players enter their 30s – which is just when average salaries start to explode.

Chart: The Conversation, CC-BY-ND • Source: Baseball-Reference • Get the data

- What data sources could be useful?
    - Lahman Database: Player Analytics (hitting, fielding, pitching) for every team and every player since 1871 (already in R) Lahman
        - Documentation
        - More player/team on-field stats, not much on payroll

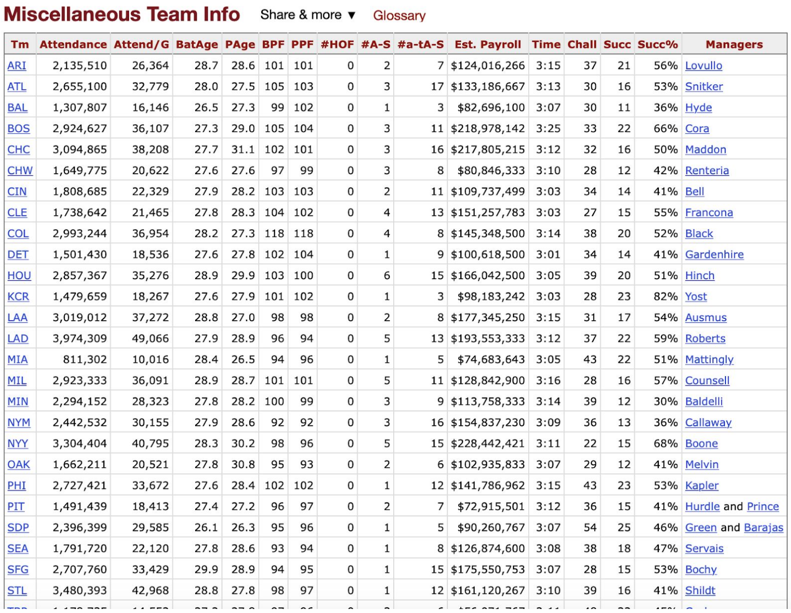
        - Player ID: first five letters of last name + first two of first + ID number
        - Ex: Clayton Kershaw = kershcl01
    - Baseball-Reference.com: Massive trove of data on all teams and players since 1871. All databases below can be accessed for any year of interest
        - MLB 2019



Team Standard Batting Share & more ▾ Glossary

Tm	#Bat	BatAge	R/G	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	LOB
ARI	45	28.7	5.02	162	6315	5633	813	1419	288	40	220	778	88	14	540	1360	.252	.323	.434	.757	94	2447	120	70	31	40	36	1119
ATL	50	28.0	5.28	162	6302	5560	855	1432	277	29	249	824	89	28	619	1467	.258	.336	.452	.789	99	2514	104	60	25	35	39	1138
BAL	58	26.5	4.50	162	6189	5596	729	1379	252	25	213	698	84	30	462	1435	.246	.310	.415	.725	91	2320	111	71	22	37	8	1063
BOS	47	27.3	5.56	162	6475	5770	901	1554	345	27	245	857	68	30	590	1382	.269	.340	.466	.806	107	2688	127	49	20	44	36	1170
CHC	52	27.7	5.02	162	6195	5461	814	1378	270	26	256	783	45	24	581	1460	.252	.331	.452	.783	100	2468	127	83	30	39	33	1071
CHW	47	27.6	4.40	161	6042	5529	708	1443	260	20	182	676	63	28	378	1549	.261	.314	.414	.728	94	2289	114	66	36	32	13	1071
CIN	47	27.8	4.33	162	6100	5450	701	1328	235	27	227	679	80	38	492	1436	.244	.315	.422	.736	88	2298	111	89	30	33	25	1073
CLE	54	27.7	4.75	162	6124	5425	769	1354	286	18	223	731	103	35	563	1332	.250	.323	.432	.756	95	2345	110	50	40	46	30	1072
COL	50	28.2	5.15	162	6288	5660	835	1502	323	41	224	803	71	31	489	1503	.265	.326	.456	.782	87	2579	111	43	51	43	25	1075
DET	53	27.6	3.61	161	6039	5549	582	1333	292	41	149	556	57	20	391	1595	.240	.294	.388	.682	78	2154	108	48	9	42	14	1069
HOU	45	29.0	5.68	162	6394	5613	920	1538	323	28	288	891	67	27	645	1166	.274	.352	.495	.848	119	2781	146	66	10	57	17	1168
KCR	51	27.6	4.27	162	6080	5496	691	1356	281	40	162	655	117	39	456	1405	.247	.309	.401	.710	86	2203	113	59	24	42	17	1056
LAA	57	28.8	4.75	162	6251	5542	769	1368	268	21	220	734	65	20	586	1276	.247	.324	.422	.746	98	2338	143	67	4	42	29	1125
LAD	46	27.9	5.47	162	6282	5493	886	1414	302	20	279	861	57	10	607	1356	.257	.338	.472	.810	112	2593	100	81	55	45	47	1124
MIA	50	28.4	3.80	162	6045	5512	615	1326	265	18	146	593	55	30	395	1469	.241	.298	.375	.673	79	2065	139	73	31	33	16	1034
MIL	50	28.9	4.75	162	6309	5542	769	1366	279	17	250	744	101	25	629	1563	.246	.329	.438	.767	97	2429	120	72	20	38	42	1180
MIN	50	27.8	5.80	162	6392	5732	939	1547	318	23	307	906	28	21	525	1334	.270	.338	.494	.832	117	2832	101	81	10	41	21	1115
NYM	53	27.9	4.88	162	6290	5624	791	1445	280	17	242	767	56	27	516	1384	.257	.328	.442	.770	106	2485	129	95	28	27	34	1128
NYY	54	28.3	5.82	162	6245	5583	943	1493	290	17	306	904	55	22	569	1437	.267	.339	.490	.829	118	2735	113	49	10	33	18	1039
OAK	49	27.8	5.22	162	6270	5561	845	1384	292	23	257	800	49	21	578	1338	.249	.327	.448	.776	108	2493	140	87	7	36	17	1081
PHI	56	27.7	4.78	162	6261	5571	774	1369	311	26	215	742	78	18	562	1453	.246	.319	.427	.746	91	2377	97	57	34	34	47	1129
PIT	54	27.5	4.68	162	6228	5657	758	1497	315	38	163	722	64	29	425	1213	.265	.321	.420	.741	95	2377	119	63	47	34	41	1103
SDP	54	26.2	4.21	162	6019	5391	682	1281	224	24	219	652	70	37	504	1581	.238	.308	.410	.718	89	2210	120	55	37	31	19	1008
SEA	67	27.8	4.68	162	6199	5500	758	1305	254	28	239	730	115	47	588	1581	.237	.316	.424	.740	100	2332	83	58	14	37	7	1080
SFG	64	29.9	4.19	162	6170	5579	678	1332	300	26	167	655	47	28	475	1435	.239	.302	.392	.694	84	2185	111	50	24	42	26	1069
STL	43	28.8	4.72	162	6167	5449	764	1336	246	24	210	714	116	29	561	1420	.245	.322	.415	.737	92	2260	110	76	40	39	15	1107
TBR	57	27.2	4.75	162	6285	5628	769	1427	291	29	217	730	94	37	542	1493	.254	.325	.431	.757	101	2427	114	73	8	34	20	1130

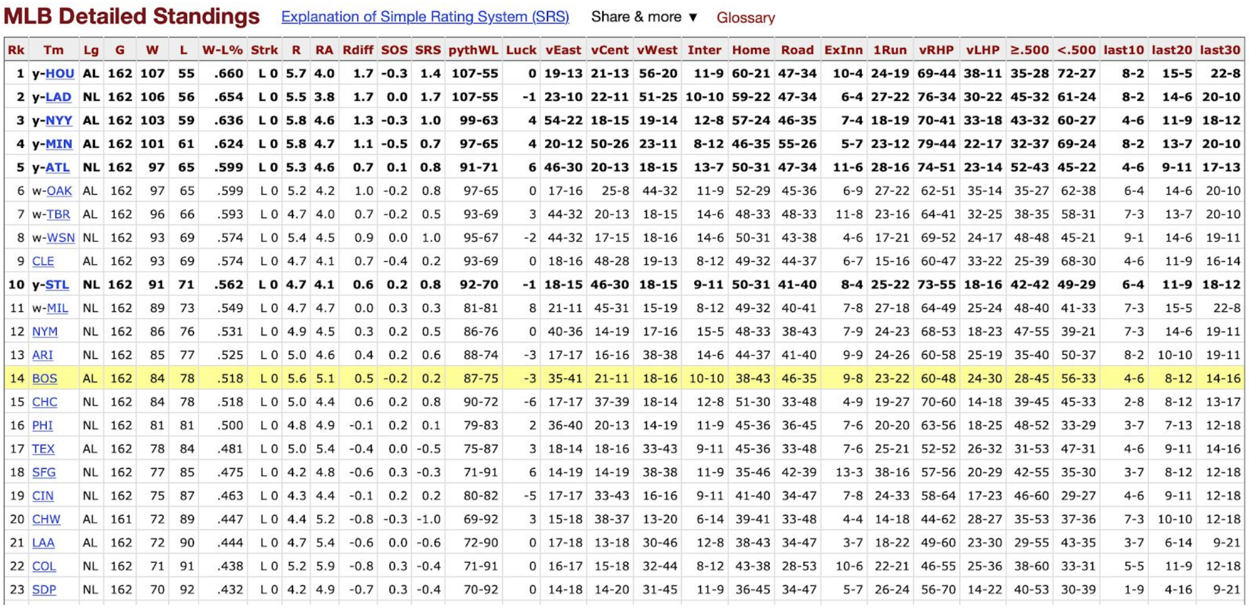
        - Misc. Franchise Information 2019
            - Money, attendance, etc



Miscellaneous Team Info Share & more ▾ Glossary

Tm	Attendance	Attend/G	BatAge	PAge	BPF	PPF	#HOF	#A-S	#a-tA-S	Est. Payroll	Time	Chall	Succ	Succ%	Managers
ARI	2,135,510	26,364	28.7	28.6	101	101	0	2	7	$124,016,266	3:15	37	21	56%	Lovullo
ATL	2,655,100	32,779	28.0	27.5	105	103	0	3	17	$133,186,667	3:13	30	16	53%	Snitker
BAL	1,307,807	16,146	26.5	27.3	99	102	0	1	3	$82,696,100	3:07	30	11	36%	Hyde
BOS	2,924,627	36,107	27.3	29.0	105	104	0	3	11	$218,978,142	3:25	33	22	66%	Cora
CHC	3,094,865	38,208	27.7	31.1	102	101	0	3	16	$217,805,215	3:12	32	16	50%	Maddon
CHW	1,649,775	20,622	27.6	27.6	97	99	0	3	8	$80,846,333	3:10	28	12	42%	Renteria
CIN	1,808,685	22,329	27.9	28.2	103	103	0	2	11	$109,737,499	3:03	34	14	41%	Bell
CLE	1,738,642	21,465	27.8	28.3	104	102	0	4	13	$151,257,783	3:03	27	15	55%	Francona
COL	2,993,244	36,954	28.2	27.3	118	118	0	4	8	$145,348,500	3:14	38	20	52%	Black
DET	1,501,430	18,536	27.6	27.8	102	104	0	1	9	$100,618,500	3:01	34	14	41%	Gardenhire
HOU	2,857,367	35,276	28.9	29.9	103	100	0	6	15	$166,042,500	3:05	39	20	51%	Hinch
KCR	1,479,659	18,267	27.6	27.9	101	102	0	1	3	$98,183,242	3:03	28	23	82%	Yost
LAA	3,019,012	37,272	28.8	27.0	98	98	0	2	8	$177,345,250	3:15	31	17	54%	Ausmus
LAD	3,974,309	49,066	27.9	28.9	96	94	0	5	13	$193,553,333	3:12	37	22	59%	Roberts
MIA	811,302	10,016	28.4	26.5	94	96	0	1	5	$74,683,643	3:05	43	22	51%	Mattingly
MIL	2,923,333	36,091	28.9	28.7	101	101	0	5	11	$128,842,900	3:16	28	16	57%	Counsell
MIN	2,294,152	28,323	27.8	28.2	100	99	0	3	9	$113,758,333	3:14	39	12	30%	Baldelli
NYM	2,442,532	30,155	27.9	28.6	92	92	0	3	16	$154,837,230	3:09	36	13	36%	Callaway
NYY	3,304,404	40,795	28.3	30.2	98	96	0	5	15	$228,442,421	3:11	22	15	68%	Boone
OAK	1,662,211	20,521	27.8	30.8	95	93	0	2	6	$102,935,833	3:07	29	12	41%	Melvin
PHI	2,727,421	33,672	27.6	28.4	102	102	0	1	12	$141,786,962	3:15	43	23	53%	Kapler
PIT	1,491,439	18,413	27.4	27.2	96	97	0	2	7	$72,915,501	3:12	36	15	41%	Hurdle and Prince
SDP	2,396,399	29,585	26.1	26.3	95	96	0	1	5	$90,260,767	3:07	54	25	46%	Green and Barajas
SEA	1,791,720	22,120	27.8	28.6	93	94	0	1	8	$126,874,600	3:08	38	18	47%	Servais
SFG	2,707,760	33,429	29.9	28.9	94	95	0	1	15	$175,550,753	3:07	28	15	53%	Bochy
STL	3,480,393	42,968	28.8	27.8	98	97	0	1	12	$161,120,267	3:10	39	16	41%	Shildt

        - 2019 Standings

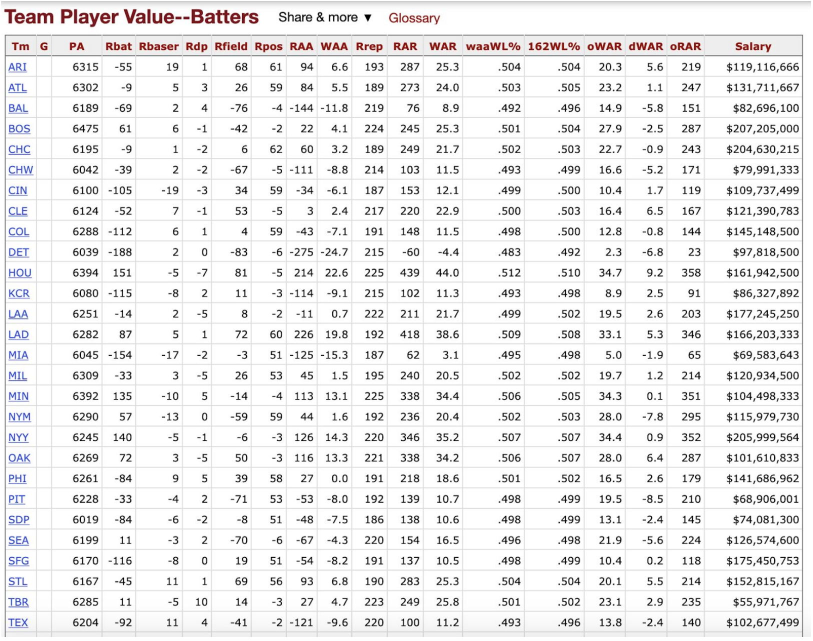


MLB Detailed Standings

Explanation of Simple Rating System (SRS) Share & more ▼ Glossary

Rk	Tm	Lg	G	W	L	W-L%	Strk	R	RA	Rdiff	SOS	SRS	pythWL	Luck	vEast	vCent	vWest	Inter	Home	Road	ExInn	1Run	vRHP	vLHP	≥.500	<.500	last10	last20	last30
1	y-HOU	AL	162	107	55	.660	L 0	5.7	4.0	1.7	-0.3	1.4	107-55	0	19-13	21-13	56-20	11-9	60-21	47-34	10-4	24-19	69-44	38-11	35-28	72-27	8-2	15-5	22-8
2	y-LAD	NL	162	106	56	.654	L 0	5.5	3.8	1.7	0.0	1.7	107-55	-1	23-10	22-11	51-25	10-10	59-22	47-34	6-4	27-22	76-34	30-22	45-32	61-24	8-2	14-6	20-10
3	y-NYY	AL	162	103	59	.636	L 0	5.8	4.6	1.3	-0.3	1.0	99-63	4	54-22	18-15	19-14	12-8	57-24	46-35	7-4	18-19	70-41	33-18	43-32	60-27	4-6	11-9	18-12
4	y-MIN	AL	162	101	61	.624	L 0	5.8	4.7	1.1	-0.5	0.7	97-65	4	20-12	50-26	23-11	8-12	46-35	55-26	5-7	23-12	79-44	22-17	32-37	69-24	8-2	13-7	20-10
5	y-ATL	NL	162	97	65	.599	L 0	5.3	4.6	0.7	0.1	0.8	91-71	6	46-30	20-13	18-15	13-7	50-31	47-34	11-6	28-16	74-51	23-14	52-43	45-22	4-6	9-11	17-13
6	w-OAK	AL	162	97	65	.599	L 0	5.2	4.2	1.0	-0.2	0.8	97-65	0	17-16	25-8	44-32	11-9	52-29	45-36	6-9	27-22	62-51	35-14	35-27	62-38	6-4	14-6	20-10
7	w-TBR	AL	162	96	66	.593	L 0	4.7	4.0	0.7	-0.2	0.5	93-69	3	44-32	20-13	18-15	14-6	48-33	48-33	11-8	23-16	64-41	32-25	38-35	58-31	7-3	13-7	20-10
8	w-WSN	NL	162	93	69	.574	L 0	5.4	4.5	0.9	0.0	1.0	95-67	-2	44-32	17-15	18-16	14-6	50-31	43-38	4-6	17-21	69-52	24-17	48-48	45-21	9-1	14-6	19-11
9	CLE	AL	162	93	69	.574	L 0	4.7	4.1	0.7	-0.4	0.2	93-69	0	18-16	48-28	19-13	8-12	49-32	44-37	6-7	15-16	60-47	33-22	25-39	68-30	4-6	11-9	16-14
10	y-STL	NL	162	91	71	.562	L 0	4.7	4.1	0.6	0.2	0.8	92-70	-1	18-15	46-30	18-15	9-11	50-31	41-40	8-4	25-22	73-55	18-16	42-42	49-29	6-4	11-9	18-12
11	w-MIL	NL	162	89	73	.549	L 0	4.7	4.7	0.0	0.3	0.3	81-81	8	21-11	45-31	15-19	8-12	49-32	40-41	7-8	27-18	64-49	25-24	48-40	41-33	7-3	15-5	22-8
12	NYM	NL	162	86	76	.531	L 0	4.9	4.5	0.3	0.2	0.5	86-76	0	40-36	14-19	17-16	15-5	48-33	38-43	7-9	24-23	68-53	18-23	47-55	39-21	7-3	14-6	19-11
13	ARI	NL	162	85	77	.525	L 0	5.0	4.6	0.4	0.2	0.6	88-74	-3	17-17	16-16	38-38	14-6	44-37	41-40	9-9	24-26	60-58	25-19	35-40	50-37	8-2	10-10	19-11
14	BOS	AL	162	84	78	.518	L 0	5.6	5.1	0.5	-0.2	0.2	87-75	-3	35-41	21-11	18-16	10-10	38-43	46-35	9-8	23-22	60-48	24-30	28-45	56-33	4-6	8-12	14-16
15	CHC	NL	162	84	78	.518	L 0	5.0	4.4	0.6	0.2	0.8	90-72	-6	17-17	37-39	18-14	12-8	51-30	33-48	4-9	19-27	70-60	14-18	39-45	45-33	2-8	8-12	13-17
16	PHI	NL	162	81	81	.500	L 0	4.8	4.9	-0.1	0.2	0.1	79-83	2	36-40	20-13	14-19	11-9	45-36	36-45	7-6	20-20	63-56	18-25	48-52	33-29	3-7	7-13	12-18
17	TEX	AL	162	78	84	.481	L 0	5.0	5.4	-0.4	0.0	-0.5	75-87	3	18-14	18-16	33-43	9-11	45-36	33-48	7-6	25-21	52-52	26-32	31-53	47-31	4-6	9-11	14-16
18	SFG	NL	162	77	85	.475	L 0	4.2	4.8	-0.6	0.3	-0.3	71-91	6	14-19	14-19	38-38	11-9	35-46	42-39	13-3	38-16	57-56	20-29	42-55	35-30	3-7	8-12	12-18
19	CIN	NL	162	75	87	.463	L 0	4.3	4.4	-0.1	0.2	0.2	80-82	-5	17-17	33-43	16-16	9-11	41-40	34-47	7-8	24-33	58-64	17-23	46-60	29-27	4-6	9-11	12-18
20	CHW	AL	161	72	89	.447	L 0	4.4	5.2	-0.8	-0.3	-1.0	69-92	3	15-18	38-37	13-20	6-14	39-41	33-48	4-4	14-18	44-62	28-27	35-53	37-36	7-3	10-10	12-18
21	LAA	AL	162	72	90	.444	L 0	4.7	5.4	-0.6	0.0	-0.6	72-90	0	17-18	13-18	30-46	12-8	38-43	34-47	3-7	18-22	49-60	23-30	29-55	43-35	3-7	6-14	9-21
22	COL	NL	162	71	91	.438	L 0	5.2	5.9	-0.8	0.3	-0.4	71-91	0	16-17	15-18	32-44	8-12	43-38	28-53	10-6	22-21	46-55	25-36	38-60	33-31	5-5	11-9	12-18
23	SDP	NL	162	70	92	.432	L 0	4.2	4.9	-0.7	0.3	-0.4	70-92	0	14-18	14-20	31-45	11-9	36-45	34-47	5-7	26-24	56-70	14-22	40-53	30-39	1-9	4-16	9-21

Batter Value



Team Player Value--Batters

Share & more ▼ Glossary

Tm	G	PA	Rbat	Rbaser	Rdp	Rfield	Rpos	RAA	WAA	Rrep	RAR	WAR	waaWL%	162WL%	oWAR	dWAR	oRAR	Salary
ARI		6315	-55	19	1	68	61	94	6.6	193	287	25.3	.504	.504	20.3	5.6	219	$119,116,666
ATL		6302	-9	5	3	26	59	84	5.5	189	273	24.0	.503	.505	23.2	1.1	247	$131,711,667
BAL		6189	-69	2	4	-76	-4	-144	-11.8	219	76	8.9	.492	.496	14.9	-5.8	151	$82,696,100
BOS		6475	61	6	-1	-42	-2	22	4.1	224	245	25.3	.501	.504	27.9	-2.5	287	$207,205,000
CHC		6195	-9	1	-2	6	62	60	3.2	189	249	21.7	.502	.503	22.7	-0.9	243	$204,630,215
CHW		6042	-39	2	-2	-67	-5	-111	-8.8	214	103	11.5	.493	.499	16.6	-5.2	171	$79,991,333
CIN		6100	-105	-19	-3	34	59	-34	-6.1	187	153	12.1	.499	.500	10.4	1.7	119	$109,737,499
CLE		6124	-52	7	-1	53	-5	3	2.4	217	220	22.9	.500	.503	16.4	6.5	167	$121,390,783
COL		6288	-112	6	1	4	59	-43	-7.1	191	148	11.5	.498	.500	12.8	-0.8	144	$145,148,500
DET		6039	-188	2	0	-83	-6	-275	-24.7	215	-60	-4.4	.483	.492	2.3	-6.8	23	$97,818,500
HOU		6394	151	-5	-7	81	-5	214	22.6	225	439	44.0	.512	.510	34.7	9.2	358	$161,942,500
KCR		6080	-115	-8	2	11	-3	-114	-9.1	215	102	11.3	.493	.498	8.9	2.5	91	$86,327,892
LAA		6251	-14	2	-5	8	-2	-11	0.7	222	211	21.7	.499	.502	19.5	2.6	203	$177,245,250
LAD		6282	87	5	1	72	60	226	19.8	192	418	38.6	.509	.508	33.1	5.3	346	$166,203,333
MIA		6045	-154	-17	-2	-3	51	-125	-15.3	187	62	3.1	.495	.498	5.0	-1.9	65	$69,583,643
MIL		6309	-33	3	-5	26	53	45	1.5	195	240	20.5	.502	.502	19.7	1.2	214	$120,934,500
MIN		6392	135	-10	5	-14	-4	113	13.1	225	338	34.4	.506	.505	34.3	0.1	351	$104,498,333
NYM		6290	57	-13	0	-59	59	44	1.6	192	236	20.4	.502	.503	28.0	-7.8	295	$115,979,730
NYY		6245	140	-5	-1	-6	-3	126	14.3	220	346	35.2	.507	.507	34.4	0.9	352	$205,999,564
OAK		6269	72	3	-5	50	-3	116	13.3	221	338	34.2	.506	.507	28.0	6.4	287	$101,610,833
PHI		6261	-84	9	5	39	58	27	0.0	191	218	18.6	.501	.502	16.5	2.6	179	$141,686,962
PIT		6228	-33	-4	2	-71	53	-53	-8.0	192	139	10.7	.498	.499	19.5	-8.5	210	$68,906,001
SDP		6019	-84	-6	-2	-8	51	-48	-7.5	186	138	10.6	.498	.499	13.1	-2.4	145	$74,081,300
SEA		6199	11	-3	2	-70	-6	-67	-4.3	220	154	16.5	.496	.498	21.9	-5.6	224	$126,574,600
SFG		6170	-116	-8	0	19	51	-54	-8.2	191	137	10.5	.498	.499	10.4	0.2	118	$175,450,753
STL		6167	-45	11	1	69	56	93	6.8	190	283	25.3	.504	.504	20.1	5.5	214	$152,815,167
TBR		6285	11	-5	10	14	-3	27	4.7	223	249	25.8	.501	.502	23.1	2.9	235	$55,971,767
TEX		6204	-92	11	4	-41	-2	-121	-9.6	220	100	11.2	.493	.496	13.8	-2.4	140	$102,677,499

        - Pitcher Value
        - ESPN Attendance
    - Forbes franchise valuation
        - Forbes valuation of 2019 team
- Do we have all the data needed to observe the outcome we are looking for?
    - Yes. Using the Lahman database which is already loaded in R, as well as the extensive data sources from BaseballReference, we will have a wealth of data with which to complete our analysis. Lahman only has data up to 2016 for some datasets and up to 2018 for others, so we will use baseball reference to supplement the datasets.

- How does each data source contribute to the analysis?
    - The Lahman database for player data is similar to the BaseballReference data, but the BaseballReference website also contains data on team payroll and player contracts, which will also be vital. We will need to join both on-field and contractual information to complete the analysis we are looking for.
- What other data sources could possibly be useful? How can we get them?
    - We may continue looking for other databases from ESPN or other sources, but at the moment it seems that the Lahman and BaseballReference databases will offer us more than enough data.
- Add the schema of each data source. What does it look like?
    - The Lahman database is included in R. It has 24 different dataframes. We plan to use AwardsPlayers(Player awards), Appearances(Data on the number of and type of game appearances of a player), Teams(General team data), SeriesPost(Postseason data on how a team did in the postseason),
    - The BaseballReference database is organized with rows of players or franchises, and columns representing statistics like batting averages, contract value, team attendance, etc. A screenshot is in the main folder.
- What cleaning and smoothing methods do we think could be appropriate for each data source and for each data column (attribute/feature)?
    - Many of the Lahman and BaseballReference datasets have a lot of statistics which are not necessarily relevant to our project. One step of cleaning will be to eliminate those unnecessary columns
- Do we need to bin/discretize the data samples?
    - What time interval could be useful for aggregating/binning data and extracting features
        - We will bin most of the data by year as we compare year to year values for revenue, statistics, etc. and what affects those values year to year.
- What transformation technique could be more appropriate
    - Normalizing the data?
        - Player payroll should be normalized using the price level of the city
    - Discretizing the data?
    - What data columns need to be normalized or discretized?
- What features would be useful to extract from each data source and data column
    - It will be useful to extract data regarding both player performance and salary as well as data about the value and income structure of franchises as a whole.
- What visualization techniques could be useful for analysis and presentation of findings?
    - Currently we are not set on a particular visualization technique. Although it is likely that we will use some type of scatter plot, we will use the visualization techniques which best displays our findings as the project progresses.

- What other processing steps do we need to consider?
    - We may need to look into merging aspects of the Lahman and Baseball Reference datasets to cover years 2017 through 2019, as the Lahman database only contains financial information through 2016, and player metrics through the 2018 season, while the BaseballReference database has data through the end of the 2019 season for both player performance and compensation.