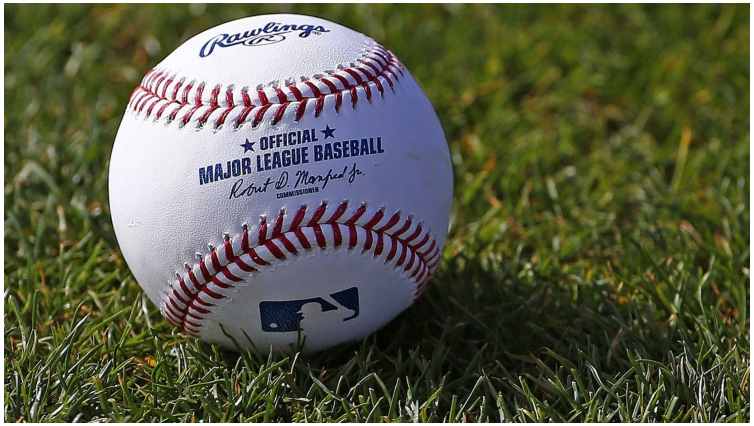
Prediction of MLB Player Performance for the 2025 Season

Marcus Saffold, Jack Fritschel

Motivation



- We both have an interest in sports analytics
- The data we are working with is very recent
- This topic does not have much prior research online

Literature Review

Conforti, Christian & Crotin, Ryan & Oseguera, Jordan. (2021). An Analysis of Playoff Performance Declines in Major League Baseball. Journal of Strength and Conditioning Research. Publish Ahead of Print. 10.1519/JSC.0000000000004140.

- This article compared an individual's performance in the MLB postseason and compared it to their regular season performance
- 1477 players were tested, ranging from the 1994-2019 seasons and divided each player into 4 categories: below average, average, above average, and excellent
- Procedure included creating 3 models for different positions:
    - Pitchers: Fielding Independent Pitching (FIP)
    - Batters: Weighted Runs Created Plus (WRP+)
    - Fielders: Errors per Inning Out (EpIO)

Research Questions

- How does regular season performance on continuous metrics correlate with postseason performance on those same metrics? What metrics improve/stay the same/decrease?

- Can we accurately predict individual on base percentage in the regular season using regular season player statistics and identify which variables are the most statistically significant?

- Do walks predict regular season on base percentage better than regular season batting average alone, achieving at least 10% less error in mean absolute error?

Postseason Data

- Used espn.com to find batting stats for all players in the 2025 postseason
- Created a python script to create and write a csv file
  - Used LLMs to fill in rows for each players, and then manually scanned and cleaned any errors

RK	NAME	POS	GP	AB	R	H	AVG	2B	3B	HR	RBI	TB	BB	K	SB	OBP	SLG	OPS	WAR
1	Aaron Judge NYY	RF	7	26	5	13	.500	2	0	1	7	18	4	5	0	.581	.692	1.273	-
3	Nico Hoerner CHC	2B	8	31	4	13	.419	1	0	1	2	17	0	2	1	.424	.548	.973	-
4	Ernie Clement TOR	3B	18	73	13	30	.411	6	1	1	9	41	1	5	1	.416	.562	.977	-
5	Vladimir Guerrero Jr. TOR	1B	18	73	18	29	.397	5	0	8	15	58	14	7	0	.494	.795	1.289	-
6	Trevor Story BOS	SS	3	13	2	5	.385	0	0	1	3	8	0	4	1	.385	.615	1.000	-
7	Spencer Steer CIN	1B	2	8	0	3	.375	0	0	0	1	3	0	2	0	.444	.375	.819	-
8	Addison Barger TOR	3B	17	60	8	22	.367	4	0	3	9	35	8	12	0	.441	.583	1.025	-
9	Freddy Fermin SD	C	3	11	0	4	.364	2	0	0	0	6	0	1	0	.364	.545	.909	-
10	J.T. Realmuto PHI	C	4	17	3	6	.353	2	1	1	3	13	0	3	0	.353	.765	1.118	-
11	Josh Naylor SEA	1B	12	47	7	16	.340	2	0	3	5	27	4	6	2	.392	.574	.967	-
12	Xander Bogaerts SD	SS	3	12	0	4	.333	1	0	0	1	5	0	3	1	.333	.417	.750	-
12	Alec Bohm PHI	3B	4	12	3	4	.333	0	0	0	0	4	6	1	0	.556	.333	.889	-
12	Jackson Merrill SD	CF	3	9	2	3	.333	2	0	1	2	8	1	2	0	.364	.889	1.253	-
15	Javier Baez DET	CF	8	32	3	10	.313	2	0	1	5	15	1	3	1	.333	.469	.802	-

```
data = [
    [1, "Aaron Judge", "NYY", "RF", 7, 26, 5, 13, .500, 2, 0, 1, 7, 18, 4, 5, 0, .581, .692, 1.273, ""],
    [3, "Nico Hoerner", "CHC", "2B", 8, 31, 4, 13, .419, 1, 0, 1, 2, 17, 0, 2, 1, .424, .548, .973, ""],
    [4, "Ernie Clement", "TOR", "3B", 18, 73, 13, 30, .411, 6, 1, 1, 9, 41, 1, 5, 1, .416, .562, .977, ""],
    [5, "Vladimir Guerrero Jr.", "TOR", "1B", 18, 73, 18, 29, .397, 5, 0, 8, 15, 58, 14, 7, 0, .494, .795, 1.289, ""],
    [6, "Trevor Story", "BOS", "SS", 3, 13, 2, 5, .385, 0, 0, 1, 3, 8, 0, 4, 1, .385, .615, 1.000, ""],
    [7, "Spencer Steer", "CIN", "1B", 2, 8, 0, 3, .375, 0, 0, 0, 1, 3, 0, 2, 0, .444, .375, .819, ""],
    [8, "Addison Barger", "TOR", "3B", 17, 60, 8, 22, .367, 4, 0, 3, 9, 35, 8, 12, 0, .441, .583, 1.025, ""],
    [9, "Freddy Fermin", "SD", "C", 3, 11, 0, 4, .364, 2, 0, 0, 0, 6, 0, 1, 0, .364, .545, .909, ""],
    [10, "J.T. Realmuto", "PHI", "C", 4, 17, 3, 6, .353, 2, 1, 1, 3, 13, 0, 3, 0, .353, .765, 1.118, ""],
    [11, "Josh Naylor", "SEA", "1B", 12, 47, 7, 16, .340, 2, 0, 3, 5, 27, 4, 6, 2, .392, .574, .967, ""],
    [12, "Xander Bogaerts", "SD", "SS", 3, 12, 0, 4, .333, 1, 0, 0, 1, 5, 0, 3, 1, .333, .417, .750, ""],
    [12, "Alec Bohm", "PHI", "3B", 4, 12, 3, 4, .333, 0, 0, 0, 0, 4, 6, 1, 0, .556, .333, .889, ""],
    [12, "Jackson Merrill", "SD", "CF", 3, 9, 2, 3, .333, 2, 0, 1, 2, 8, 1, 2, 0, .364, .889, 1.253, ""],
    [15, "Javier Baez", "DET", "CF", 8, 32, 3, 10, .313, 2, 0, 1, 5, 15, 1, 3, 1, .333, .469, .802, ""],
```

Postseason Final Dataset

	player	team	pos	gp	ab	runs	hits	batting_average	doubles	triples	hr	runs_batted_in	total_bases	walks	strikeouts	stolen_bases	on_base_percent	slg_percent	on_base_plus_slg
1	Aaron Judge	NYY	RF	7	26	5	13	0.500	2	0	1	7	18	4	5	0	0.581	0.692	1.273
2	Nico Hoerner	CHC	2B	8	31	4	13	0.419	1	0	1	2	17	0	2	1	0.424	0.548	0.973
3	Ernie Clement	TOR	3B	18	73	13	30	0.411	6	1	1	9	41	1	5	1	0.416	0.562	0.977
4	Vladimir Guerrero Jr.	TOR	1B	18	73	18	29	0.397	5	0	8	15	58	14	7	0	0.494	0.795	1.289
5	Trevor Story	BOS	SS	3	13	2	5	0.385	0	0	1	3	8	0	4	1	0.385	0.615	1.000
6	Spencer Steer	CIN	1B	2	8	0	3	0.375	0	0	0	1	3	0	2	0	0.444	0.375	0.819
7	Addison Barger	TOR	3B	17	60	8	22	0.367	4	0	3	9	35	8	12	0	0.441	0.583	1.025
8	Freddy Fermin	SD	C	3	11	0	4	0.364	2	0	0	0	6	0	1	0	0.364	0.545	0.909
9	J.T. Realmuto	PHI	C	4	17	3	6	0.353	2	1	1	3	13	0	3	0	0.353	0.765	1.118
10	Josh Naylor	SEA	1B	12	47	7	16	0.340	2	0	3	5	27	4	6	2	0.392	0.574	0.967
11	Xander Bogaerts	SD	SS	3	12	0	4	0.333	1	0	0	1	5	0	3	1	0.333	0.417	0.750
12	Alec Bohm	PHI	3B	4	12	3	4	0.333	0	0	0	0	4	6	1	0	0.556	0.333	0.889
13	Jackson Merrill	SD	CF	3	9	2	3	0.333	2	0	1	2	8	1	2	0	0.364	0.889	1.253
14	Javier Baez	DET	CF	8	32	3	10	0.313	2	0	1	5	15	1	3	1	0.333	0.469	0.802
15	Cal Raleigh	SEA	C	12	46	8	14	0.304	2	0	5	8	31	8	17	0	0.407	0.674	1.081
16	Jackson Chourio	MIL	CF	9	33	3	10	0.303	3	0	2	8	19	1	9	0	0.314	0.576	0.890
17	Alex Bregman	HOU	3B	3	10	0	3	0.300	1	0	0	1	4	3	2	0	0.462	0.400	0.862
18	Michael Busch	CHC	1B	8	27	4	8	0.296	0	0	4	4	20	3	4	0	0.387	0.741	1.128
19	Matt Chapman	TOR	3B	18	71	7	21	0.296	3	0	4	11	37	7	18	0	0.368	0.521	0.889
20	George Springer	TOR	RF	16	67	14	19	0.284	6	0	4	10	37	5	17	0	0.347	0.552	0.899
21	Kerry Carpenter	DET	RF	8	32	4	9	0.281	1	0	2	6	16	7	11	0	0.410	0.500	0.910
22	Will Smith	LAD	C	15	58	8	16	0.276	2	0	2	8	24	6	18	0	0.364	0.414	0.777
23	Caleb Durian	MIL	3B	9	29	3	8	0.276	2	1	0	2	12	3	8	3	0.364	0.414	0.777
24	Nathan Lukes	TOR	RF	17	62	5	17	0.274	4	0	0	8	21	5	10	0	0.328	0.339	0.667
25	Shohei Ohtani	LAD	DH	17	68	13	18	0.265	3	1	8	14	47	16	23	1	0.405	0.691	1.096
26	Kyle Tucker	CHC	RF	8	27	5	7	0.259	0	0	1	1	10	5	5	0	0.375	0.370	0.745
27	Teoscar Hernandez	LAD	RF	17	70	8	18	0.257	1	0	5	13	34	5	21	0	0.303	0.486	0.788
28	Alejandro Kirk	TOR	C	18	71	11	18	0.254	2	0	5	13	35	9	13	0	0.349	0.493	0.842
29	Enrique Hernandez	LAD	1B	17	64	9	16	0.250	4	0	1	7	23	4	24	0	0.290	0.359	0.649
30	Jose Ramirez	CLE	3B	3	8	1	2	0.250	0	0	0	1	2	4	2	0	0.500	0.250	0.750

Regular Season Data

- Used baseballsavant.com to find batting stats for each player
- Filtered columns so that they match columns of postseason dataset
- Downloaded csv of created table
- Edited csv via excel and R to ensure rows and columns match each other

Rk.	Player	Year	AB	PA	H	2B	3B	HR	BB	K%	AVG	SLG	OBP	OPS	RBI	TB	SB	R
1	Lindor, Francisco	2025	644	732	172	35	0	31	65	17.9	.267	.466	.346	.812	86	300	31	117
2	Devers, Rafael	2025	607	729	153	33	0	35	112	26.3	.252	.479	.372	.851	109	291	1	99
3	Ohtani, Shohei	2025	611	727	172	25	9	55	109	25.7	.282	.622	.392	1.014	102	380	20	146
4	Olson, Matt	2025	624	724	170	41	2	29	91	24.3	.272	.484	.366	.850	95	302	1	98
5	Schwarber, Kyle	2025	604	724	145	23	2	56	108	27.2	.240	.563	.365	.928	132	340	10	111
6	Perdomo, Geraldo	2025	597	720	173	33	5	20	94	11.5	.290	.462	.389	.851	100	276	27	98
7	Soto, Juan	2025	577	715	152	20	1	43	127	19.2	.263	.525	.396	.921	105	303	38	120
8	Rodríguez, Julio	2025	652	710	174	31	4	32	44	21.4	.267	.474	.324	.798	95	309	30	106
9	Arozarena, Randy	2025	613	709	146	32	1	27	64	26.9	.238	.426	.334	.760	76	261	31	95

Regular Season Final Dataset

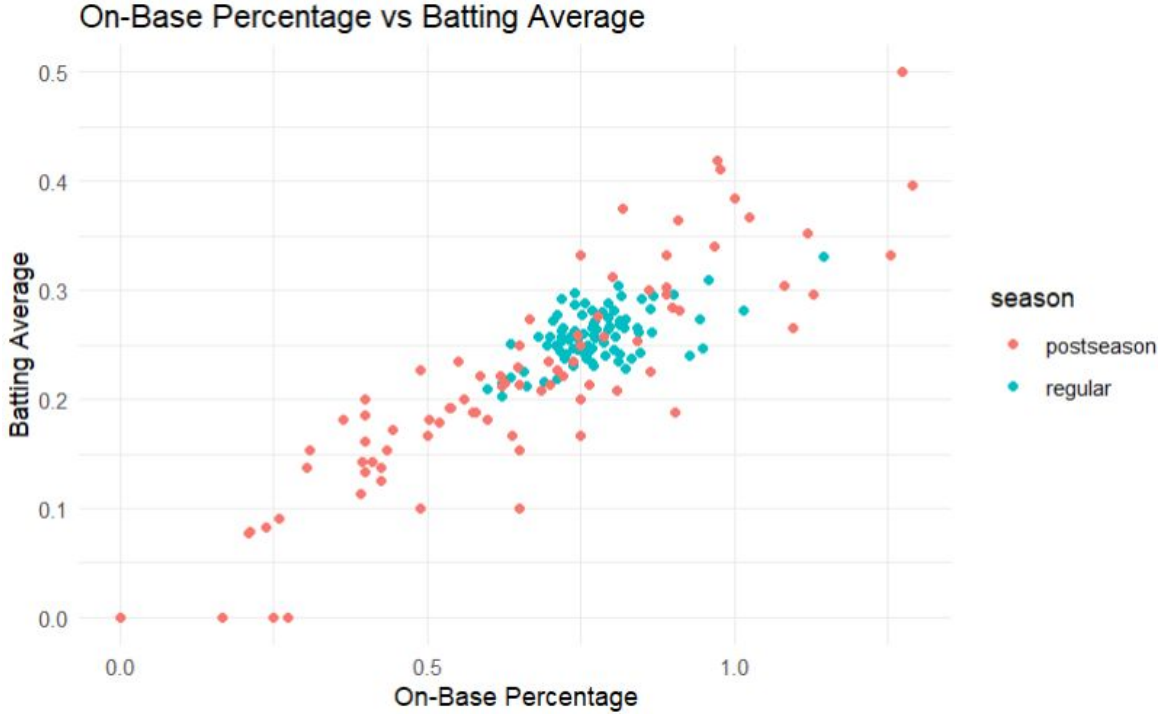
	player	team	pos	pa	ab	runs	hits	batting_average	doubles	triples	hr	runs_batted_in	total_bases	walks	strikeouts	stolen_bases	on_base_percent	slg_percent	on_base_plus_slg
1	Daulton Varsho	TOR	CF	271	248	43	59	0.238	13	2	20	55	136	17	77	2	0.284	0.548	0.832
2	Ceddanne Rafaela	BOS	CF	587	546	84	136	0.249	34	4	16	63	226	28	117	20	0.295	0.414	0.709
3	Josh Naylor	SEA	1B	604	543	81	160	0.295	29	1	20	92	251	48	83	30	0.353	0.462	0.815
4	Ryan O'Hearn	SD	1B	544	474	67	133	0.281	21	1	17	63	207	58	109	3	0.366	0.437	0.803
5	Alec Bohm	PHI	3B	504	464	53	133	0.287	18	3	11	59	190	29	82	2	0.331	0.409	0.740
6	Nico Hoerner	CHC	2B	649	599	89	178	0.297	29	4	7	61	236	39	49	29	0.345	0.394	0.739
7	Fernando Tatis Jr.	SD	RF	691	594	111	159	0.268	27	2	25	71	265	89	129	32	0.368	0.446	0.814
8	Austin Wells	NYY	C	448	401	51	88	0.219	22	1	21	71	175	30	118	5	0.275	0.436	0.711
9	Matt Chapman	TOR	3B	535	454	76	105	0.231	23	2	21	61	195	71	126	9	0.340	0.430	0.770
10	Anthony Volpe	NYY	SS	596	539	65	114	0.212	32	4	19	72	211	43	150	18	0.272	0.391	0.663
11	Bryce Harper	PHI	1B	580	501	72	131	0.261	32	0	27	75	244	70	121	12	0.357	0.487	0.844
12	Kerry Carpenter	DET	RF	464	433	66	109	0.252	18	5	26	62	215	18	106	1	0.291	0.497	0.788
13	Trea Turner	PHI	SS	639	589	94	179	0.304	31	7	15	69	269	43	107	36	0.355	0.457	0.812
14	Mookie Betts	LAD	SS	663	589	95	152	0.258	23	2	20	82	239	61	68	8	0.326	0.406	0.732
15	Gleyber Torres	DET	2B	628	532	79	136	0.256	22	0	16	74	206	85	101	4	0.358	0.387	0.745
16	Parker Meadows	DET	CF	213	191	22	41	0.215	6	2	4	16	63	21	56	4	0.291	0.330	0.621
17	Jackson Chourio	MIL	CF	589	549	88	148	0.270	35	4	21	78	254	30	121	21	0.308	0.463	0.771
18	Randy Arozarena	SD	LF	709	613	95	146	0.238	32	1	27	76	261	64	191	31	0.334	0.426	0.760
19	Ian Happ	CHC	LF	663	569	87	138	0.243	32	0	23	79	239	87	151	6	0.342	0.420	0.762
20	Cody Bellinger	NYY	LF	656	588	89	160	0.272	25	5	29	98	282	57	90	13	0.334	0.480	0.814
21	William Contreras	MIL	C	659	566	89	147	0.260	28	0	17	76	226	84	120	6	0.355	0.399	0.754
22	Brandon Marsh	PHI	CF	425	379	59	106	0.280	25	2	11	43	168	38	110	7	0.342	0.443	0.785
23	Aaron Judge	NYY	RF	679	541	137	179	0.331	30	2	53	114	372	124	160	12	0.457	0.688	1.145
24	Nathan Lukes	TOR	RF	438	388	55	99	0.255	19	2	12	65	158	38	60	2	0.323	0.407	0.730
25	Manny Machado	SD	3B	678	615	91	169	0.275	33	0	27	95	283	55	131	14	0.335	0.460	0.795
26	Dansby Swanson	CHC	SS	645	590	84	144	0.244	24	3	24	77	246	47	168	20	0.300	0.417	0.717
27	Tommy Edman	LAD	2B	377	346	49	78	0.225	13	1	13	49	132	19	61	3	0.274	0.382	0.656
28	Shohei Ohtani	LAD	DH	727	611	146	172	0.282	25	9	55	102	380	109	187	20	0.392	0.622	1.014
29	Elly De La Cruz	CIN	SS	699	629	102	166	0.264	31	7	22	86	277	67	181	37	0.336	0.440	0.776
30	Max Muncy	LAD	3B	388	313	48	76	0.243	10	2	19	67	147	64	83	4	0.376	0.470	0.846

Combined Dataset

- Combined the columns from both tables to create one big table
- Works better for the statistical models we used

	player	team	pos	post_gp	pa	ab	post_ab	runs	post_runs	hits	post_hits	batting_average	post_batting_average	doubles	post_doubles	triples	post_triples	hr	post_hr
23	Aaron Judge	NYY	RF	7	679	541	26	137	5	179	13	0.331	0.500	30	2	2	0	53	1
63	Addison Barger	TOR	3B	17	502	460	60	61	8	112	22	0.243	0.367	32	4	1	0	21	3
5	Alec Bohm	PHI	3B	4	504	464	12	53	3	133	4	0.287	0.333	18	0	3	0	11	0
43	Alejandro Kirk	TOR	C	18	506	451	71	45	11	127	18	0.282	0.254	18	2	0	0	15	5
68	Alex Bregman	HOU	3B	3	495	433	10	64	0	118	3	0.273	0.300	28	1	0	0	18	0
77	Andres Gimenez	TOR	2B	18	369	329	65	39	10	69	14	0.210	0.215	11	2	1	0	7	2
74	Andrew Vaughn	MIL	1B	9	447	406	26	35	4	103	4	0.254	0.154	22	0	0	0	14	2
57	Andy Pages	LAD	CF	16	624	581	51	74	2	158	4	0.272	0.078	27	1	1	0	27	0
10	Anthony Volpe	NYY	SS	7	596	539	26	65	2	114	5	0.212	0.192	32	1	4	0	19	1
54	Austin Hays	CIN	LF	7	416	380	28	60	3	101	4	0.266	0.143	16	2	5	0	15	1
8	Austin Wells	NYY	C	7	448	401	22	51	1	88	5	0.219	0.227	22	0	1	0	21	0
22	Brandon Marsh	PHI	CF	4	425	379	13	59	1	106	1	0.280	0.077	25	0	2	0	11	0
34	Brice Turang	MIL	2B	9	659	584	35	97	2	168	4	0.288	0.114	28	1	2	0	18	1
11	Bryce Harper	PHI	1B	4	580	501	15	72	1	131	3	0.261	0.200	32	1	0	0	27	0
73	Bryson Stott	PHI	2B	4	560	499	13	66	0	128	2	0.257	0.154	22	0	3	0	13	0
64	Cal Raleigh	SEA	C	12	705	596	46	110	8	147	14	0.247	0.304	24	2	0	0	60	5
76	Carlos Narvaez	BOS	C	3	446	403	8	51	0	97	0	0.241	0.000	27	0	0	0	15	0
71	Carson Kelly	CHC	C	8	421	369	28	48	2	92	5	0.249	0.179	13	0	1	0	17	1
2	Ceddanne Rafaela	BOS	CF	3	587	546	10	84	2	136	0	0.249	0.000	34	0	4	0	16	0
31	Christian Yelich	MIL	LF	9	644	573	33	88	3	151	6	0.264	0.182	21	1	0	0	29	0
20	Cody Bellinger	NYY	LF	7	656	588	28	89	4	160	6	0.272	0.214	25	2	5	0	29	1
26	Dansby Swanson	CHC	SS	8	645	590	26	84	1	144	4	0.244	0.154	24	1	3	0	24	0
1	Daulton Varsho	TOR	CF	18	271	248	75	43	12	59	17	0.238	0.227	13	4	2	1	20	3
38	Dillon Dingler	DET	C	8	469	435	30	54	3	121	5	0.278	0.167	21	2	2	0	13	1
29	Elly De La Cruz	CIN	SS	2	699	629	6	102	1	166	0	0.264	0.000	31	0	7	0	22	0
85	Enrique Hernandez	LAD	1B	17	256	232	64	30	9	47	16	0.203	0.250	8	4	0	0	10	1
49	Ernie Clement	TOR	3B	18	588	545	73	83	13	151	30	0.277	0.411	35	6	2	1	9	1
78	Eugenio Suarez	SEA	3B	12	657	558	47	91	4	134	10	0.228	0.213	28	1	0	0	49	3
7	Fernando Tatis Jr.	SD	RF	3	691	594	12	111	2	159	1	0.268	0.083	27	0	2	0	25	0

Models: Scatterplot



Correlation

```
> print(cor_matrix)
            reg_ba   reg_obp  reg_slug   reg_ops   post_ba  post_obp post_slug  post_ops
reg_ba    1.0000000 0.7009783 0.3915409 0.5665530 0.2867827 0.3550440 0.1881866 0.2678891
reg_obp   0.7009783 1.0000000 0.5144588 0.7757772 0.1469787 0.3088923 0.1660389 0.2347458
reg_slug  0.3915409 0.5144588 1.0000000 0.9402034 0.2083691 0.3424380 0.3569983 0.3832597
reg_ops   0.5665530 0.7757772 0.9402034 1.0000000 0.2117109 0.3746799 0.3286514 0.3752670
post_ba   0.2867827 0.1469787 0.2083691 0.2117109 1.0000000 0.8193942 0.8032473 0.8809828
post_obp  0.3550440 0.3088923 0.3424380 0.3746799 0.8193942 1.0000000 0.6532018 0.8422880
post_slug 0.1881866 0.1660389 0.3569983 0.3286514 0.8032473 0.6532018 1.0000000 0.9583255
post_ops  0.2678891 0.2347458 0.3832597 0.3752670 0.8809828 0.8422880 0.9583255 1.0000000
```

- Not very much correlation between regular season and postseason metrics
- Some of the regular season metrics are a bit more highly correlated (ex. Reg_ba & Reg_obp)

Correlation continued

```
> summary(mlb_diff %>% select(diff_OBP, diff_SLG, diff_OPS, diff_BA))
    diff_OBP           diff_SLG           diff_OPS           diff_BA
 Min.   :-0.36700   Min.   :-0.44000   Min.   :-0.7440   Min.   :-0.26400
 1st Qu.:-0.08500   1st Qu.:-0.21100   1st Qu.:-0.2940   1st Qu.:-0.09700
 Median :-0.03600   Median :-0.08300   Median :-0.1240   Median :-0.04900
 Mean   :-0.03673   Mean   :-0.08465   Mean   :-0.1214   Mean   :-0.04162
 3rd Qu.: 0.01900   3rd Qu.: 0.02600   3rd Qu.: 0.0590   3rd Qu.: 0.01900
 Max.   : 0.22500   Max.   : 0.43200   Max.   : 0.4790   Max.   : 0.16900
```

- Our difference statistics are calculated by taking the post metric - regular season metric
- This shows a reduction in average mlb player metrics across the mlb reflecting in the mean reduction in postseason vs regular season performance

Linear regression

```
> summary(lm_model)

Call:
lm(formula = y_train ~ ., data = X_train)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0116335 -0.0048552 -0.0008455  0.0040044  0.0253167

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.118e-02  1.622e-02   3.156 0.002661 **
reg_hits       -4.077e-04  1.083e-04  -3.764 0.000427 ***
reg_hr          5.300e-05  3.398e-04   0.156 0.876647
reg_walks       1.168e-03  5.985e-05  19.518  < 2e-16 ***
reg_strikeouts -2.115e-06  3.927e-05  -0.054 0.957267
reg_slug       -2.212e-02  5.036e-02  -0.439 0.662327
reg_ba          1.097e+00  1.105e-01   9.925 1.34e-13 ***
reg_runs       -4.815e-05  1.311e-04  -0.367 0.714971
reg_rbi         1.468e-05  1.232e-04   0.119 0.905622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007499 on 52 degrees of freedom
Multiple R-squared:  0.9513,    Adjusted R-squared:  0.9438
F-statistic: 126.9 on 8 and 52 DF,  p-value: < 2.2e-16
```

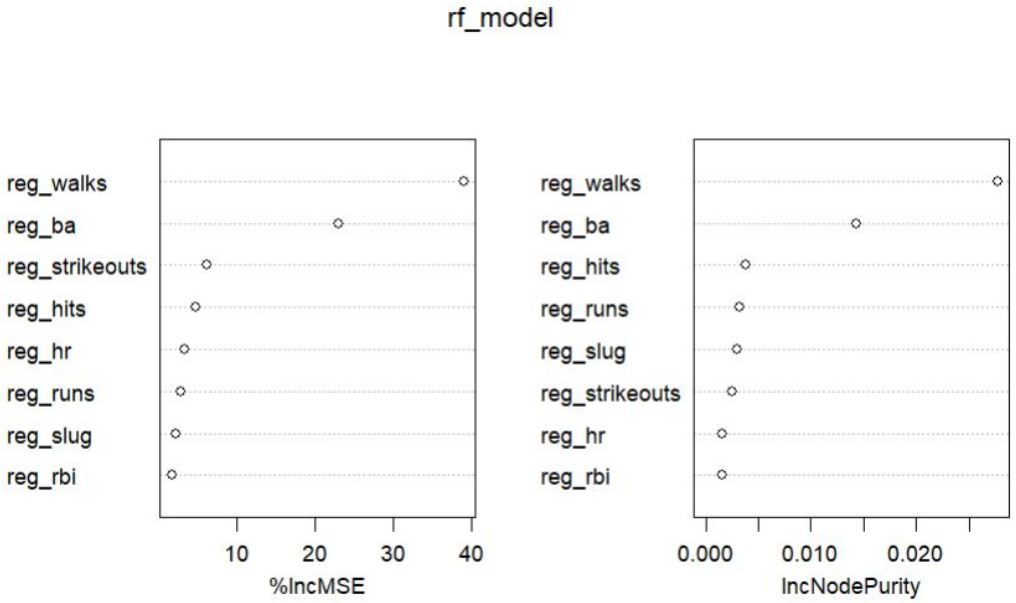
```
> eval_metrics(lm_pred, y_test)
$RMSE
[1] 0.007013435

$MAE
[1] 0.005741512

$R2
[1] 0.9692142
```

- Very effective model for predicting OBP and isn’t overfitted due to data partitioning
- Recognizes walks, hits, and batting average as the most significant predictor variables

Random Forest




```
> eval_metrics(rf_pred, y_test)
$RMSE
[1] 0.02221509

$MAE
[1] 0.0145694

$R2
[1] 0.7771266
```

- Less effective than linear regression model at predicting OBP
- Changes the significance of variables such as strikeouts

Predicting OBP using walks vs batting avg.

```
> pred_BA <- predict(lm_BA, newdata = X_test)
> MAE_BA <- mean(abs(pred_BA - y_test))
> # Percent improvement
> improvement <- ((MAE_BA - MAE_walks) / MAE_walks) * 100
> print(paste("MAE using walks:", round(MAE_walks, 4)))
[1] "MAE using walks: 0.018"
> print(paste("MAE using batting average:", round(MAE_BA, 4)))
[1] "MAE using batting average: 0.0224"
> print(paste("Percent Improvement:", round(improvement, 2), "%"))
[1] "Percent Improvement: 24.91 %"
```

- Generated 2 linear regression models predicting OBP, one with only walks and one with only batting average
- Results show walks is a much better predictor in a linear regression model than average
- Walks are linearly related to OBP which is why it is better than avg

Results/Conclusion

- The postseason data is more scattered compared to the regular season
- Were we able to predict regular season OBP with regular season predictor variables
  - Yes! These are the most impactful in order of significance:
    1. Walks
    2. Batting average
    3. Hits
    4. Strikeouts
- Walks is an incredibly powerful predictor for OBP easily clearing the second most significant predictor (Batting Average) with 24.91% less MAE.

Reproducibility

- All models run off a singular R script, making it easy to reproduce
- If you are interested in running the script or working with the datasets yourself, send us an email @saffo026@d.umn.edu or @frits079@d.umn.edu and we will send you the files needed

Questions?

The End

Thanks for listening!