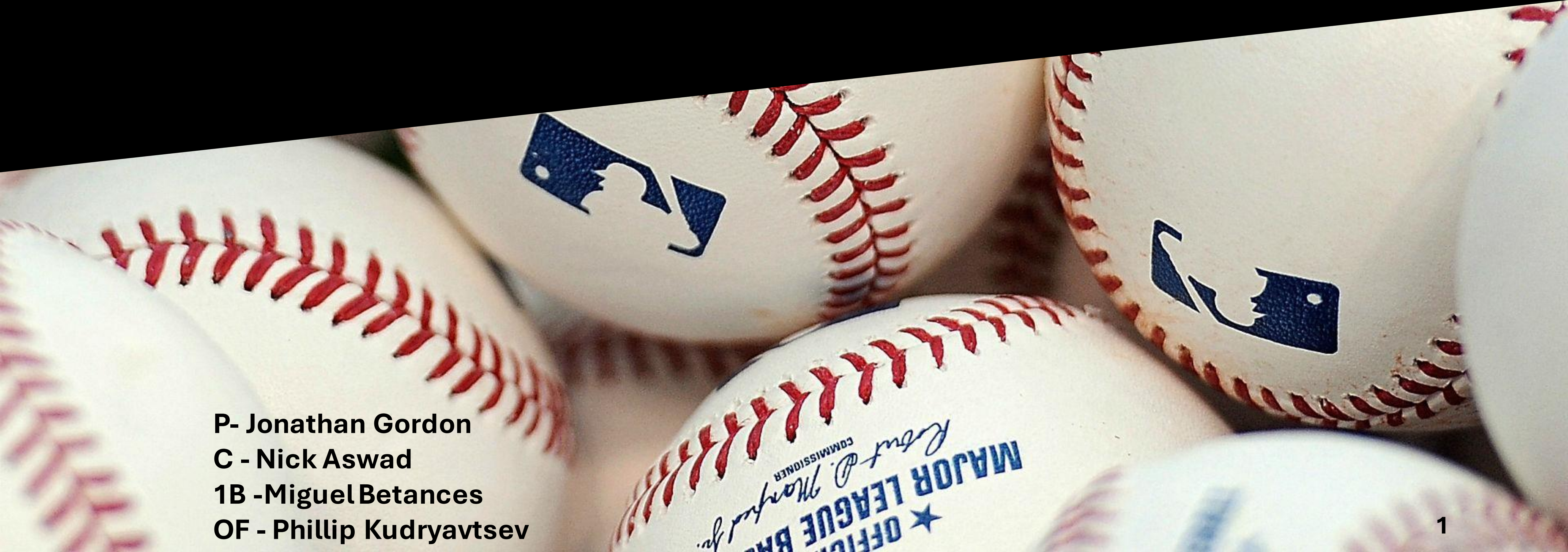
Are Arbitration Eligible Baseball Players Paid Fairly?



P- Jonathan Gordon
C - Nick Aswad
1B -Miguel Betances
OF - Phillip Kudryavtsev

Agenda

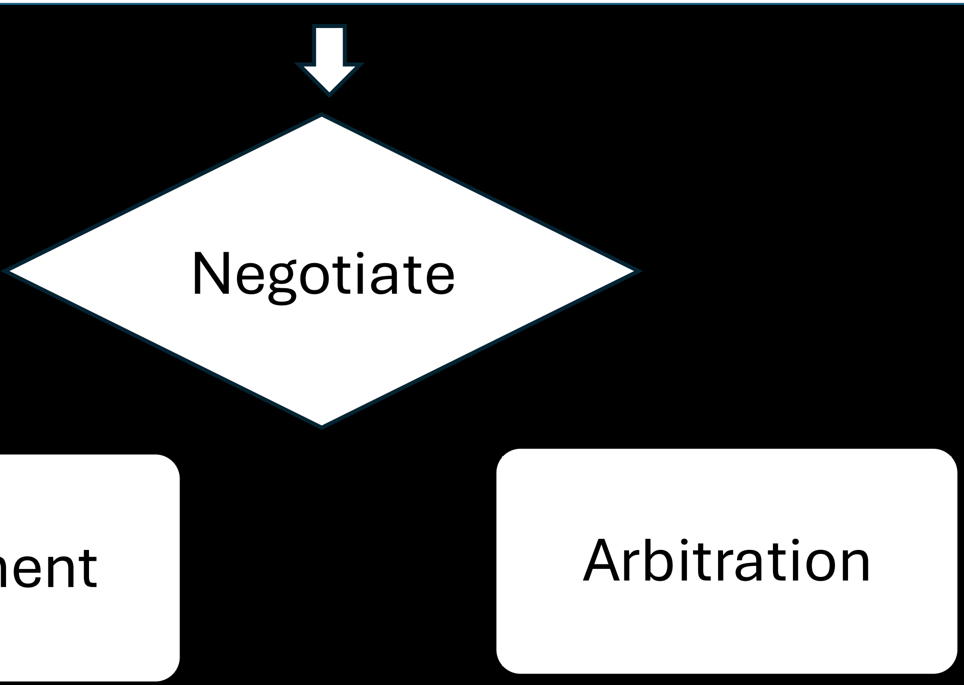
1. Introduction and background information
2. Data Prep and Cleaning
3. Methodology
4. Results
5. Conclusion

You may be wondering …

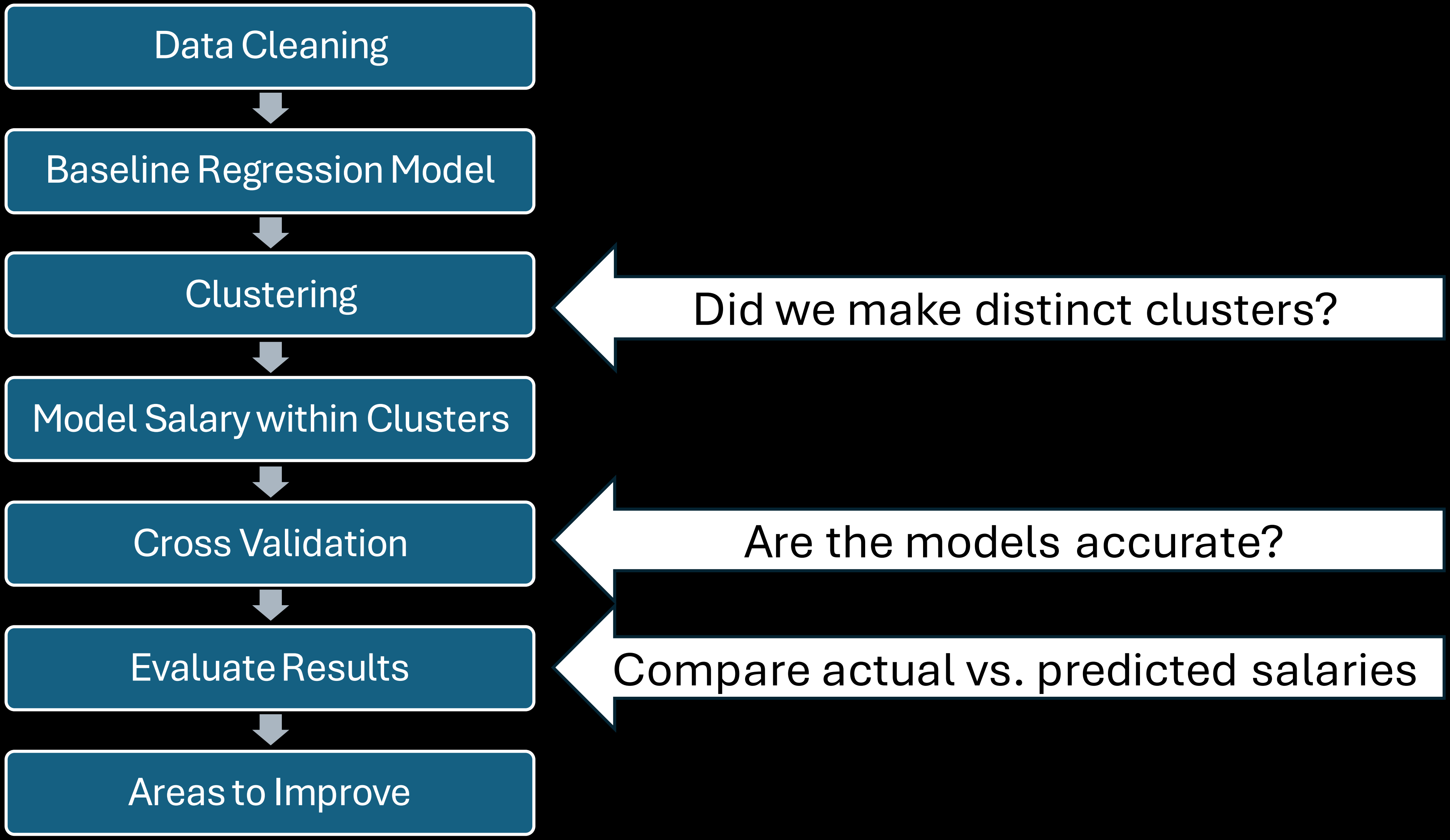
- What does arbitration eligible mean?
- Is there evidence players are not being paid fairly?
- How do we propose to determine ”fair” value?

Arbitration Process

```
Drafted                      0 years                     3 years                          6 years
----|--------------------------|---------------------------|---------------------------------|---------------->
Rookie                        Enters                     Arbitration                        Free
Contract                       MLB                        Eligible                          Agency
```



Methodology



Data Cleaning

Baseline Regression Model

Clustering

Model Salary within Clusters

Cross Validation

Evaluate Results

Areas to Improve

Did we make distinct clusters?

Are the models accurate?

Compare actual vs. predicted salaries

Data Set

- 500+ Position Players
- Excluded Pitchers
- Stats include Hits, RBIs, HRs, SBs, AVG., OBP., etc.
- 2023 Season Only
- Spotrac.com

Data Details

- Salaries ranged from $200K - $40M
- Player Ages ranged from 20-43 years old
- Many players in late 20’s made league minimum
- Excluded players who played less than 60 games
- Having only 1 season of data could limit usability
- Performance vs. Salary not always consistent

Baseline Predictions*, no clustering (first 10 players)

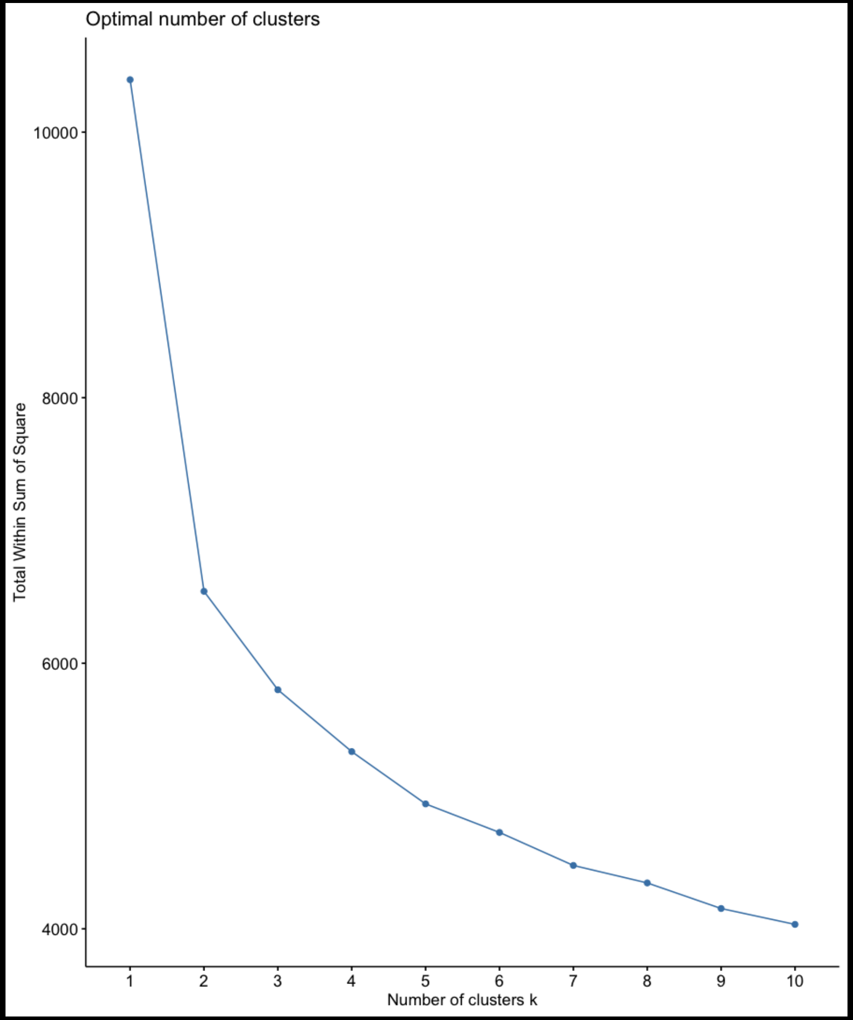
Player First	Player Last	Actual Salary	Predicted Salary	Delta
Juan	Soto	$23,000,000	$14,302,020	38%
Pete	Alonso	$14,500,000	$13,371,610	8%
Gleyber	Torres	$9,950,000	$5,942,636	40%
Anthony	Santander	$7,400,000	$10,775,858	-46%
Christian	Walker	$6,500,000	$14,342,036	-121%
Alex	Verdugo	$6,300,000	$3,272,654	48%
Luis	Arraez	$6,100,000	$3,189,310	48%
Mike	Yastrzemski	$6,100,000	$5,881,519	4%
Kyle	Farmer	$5,585,000	$9,572,123	-71%
Kyle	Tucker	$5,000,000	$11,467,566	-129%

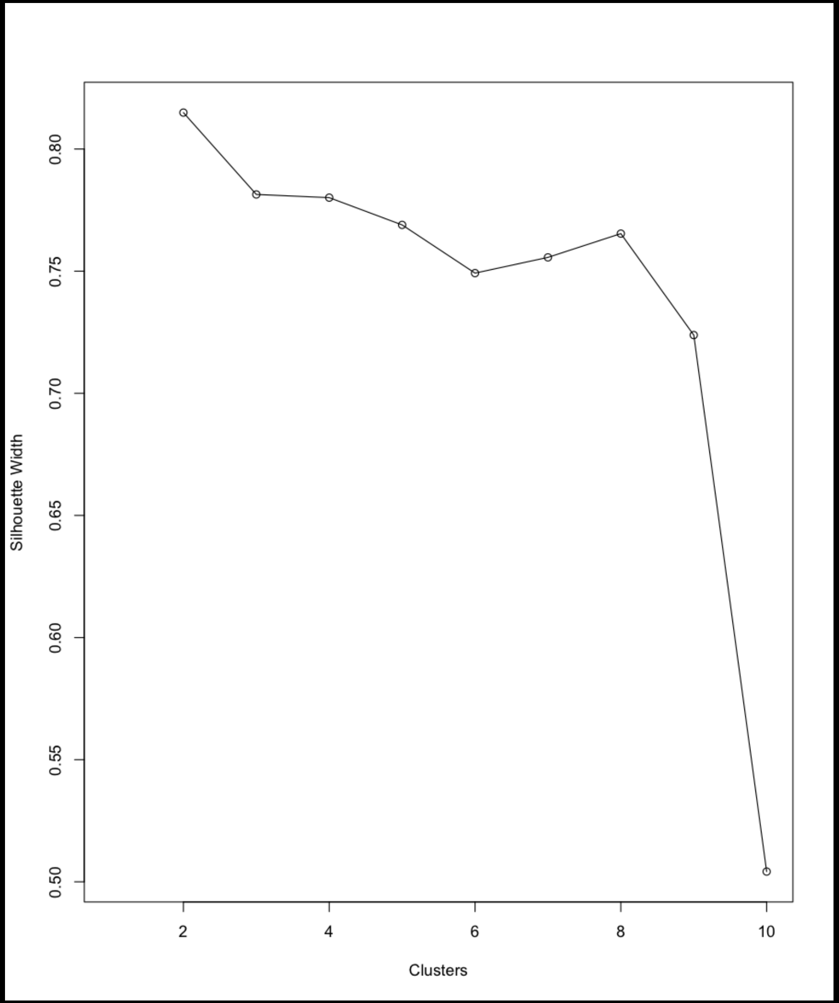
High variance

mean	-55%
sd	97%

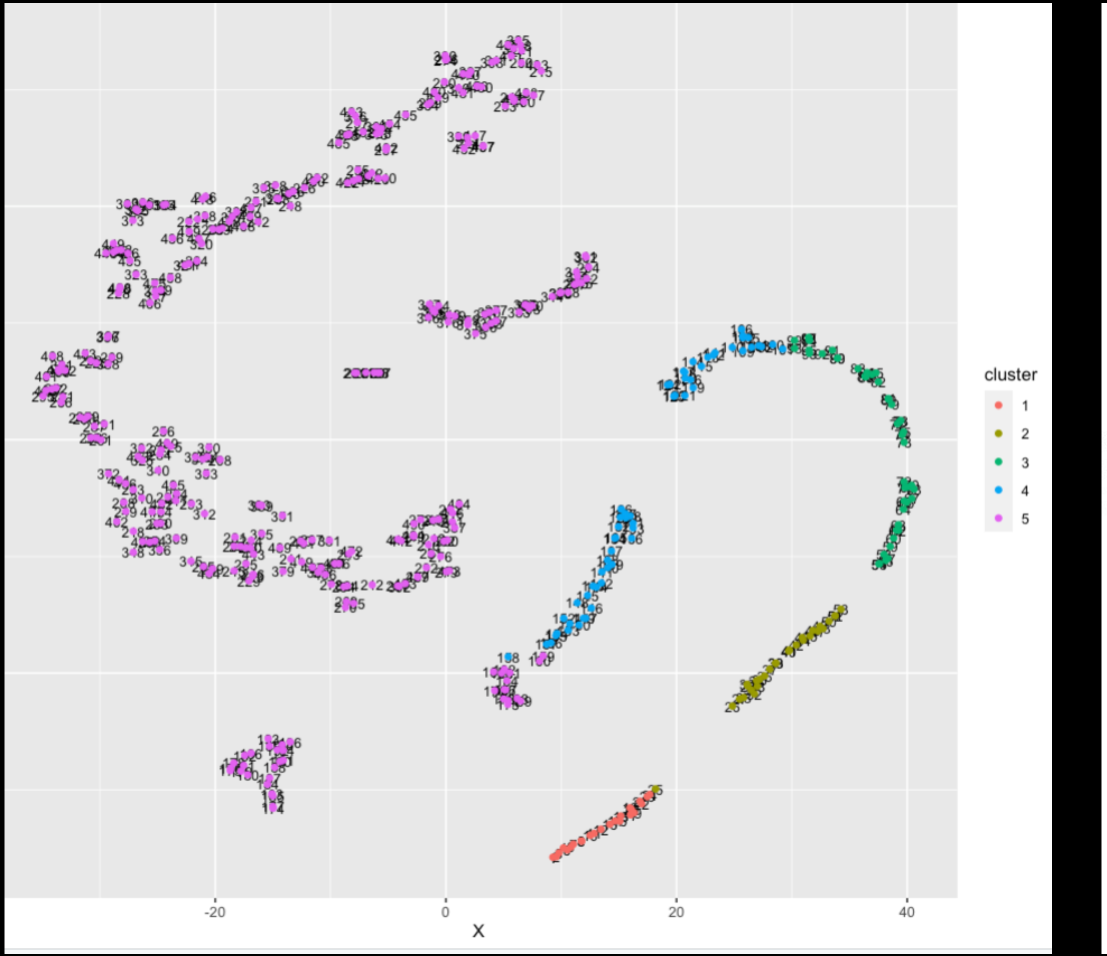
* Used Random Forest

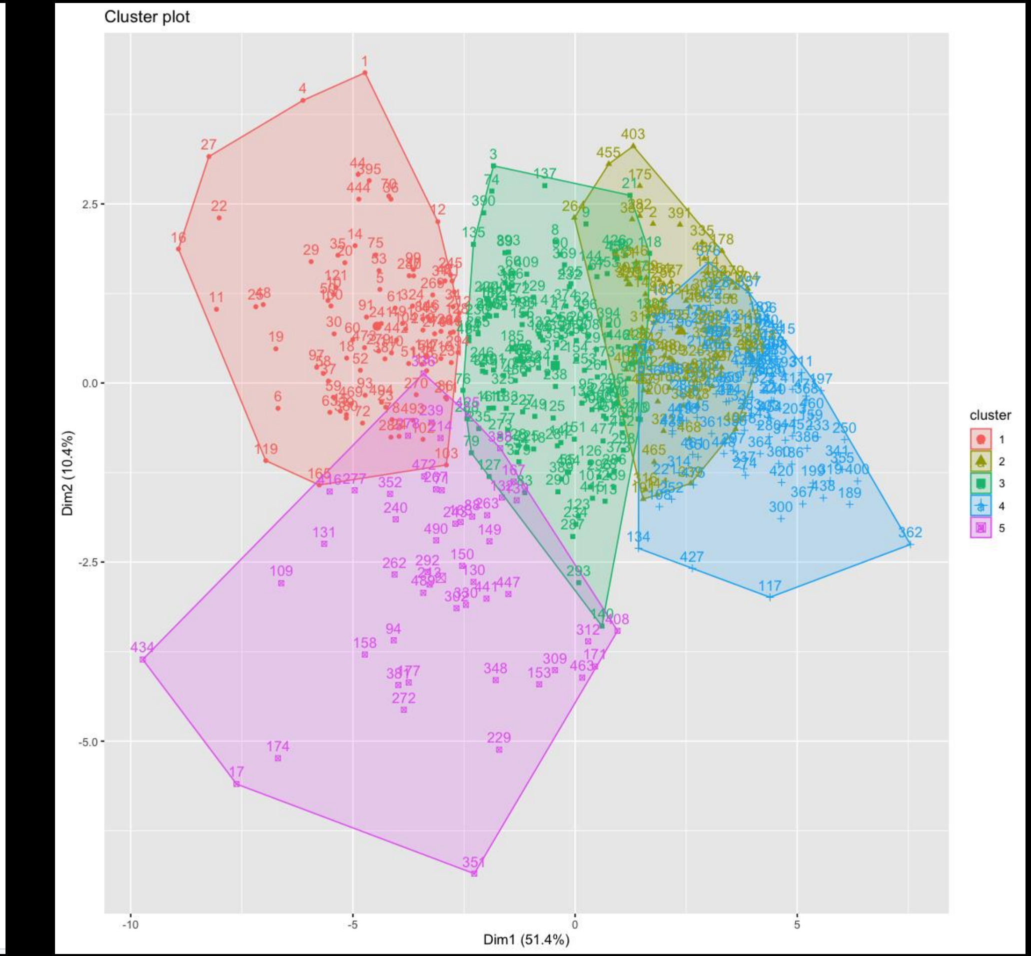
Cluster Analysis: Elbow Curves, Silhouette Width



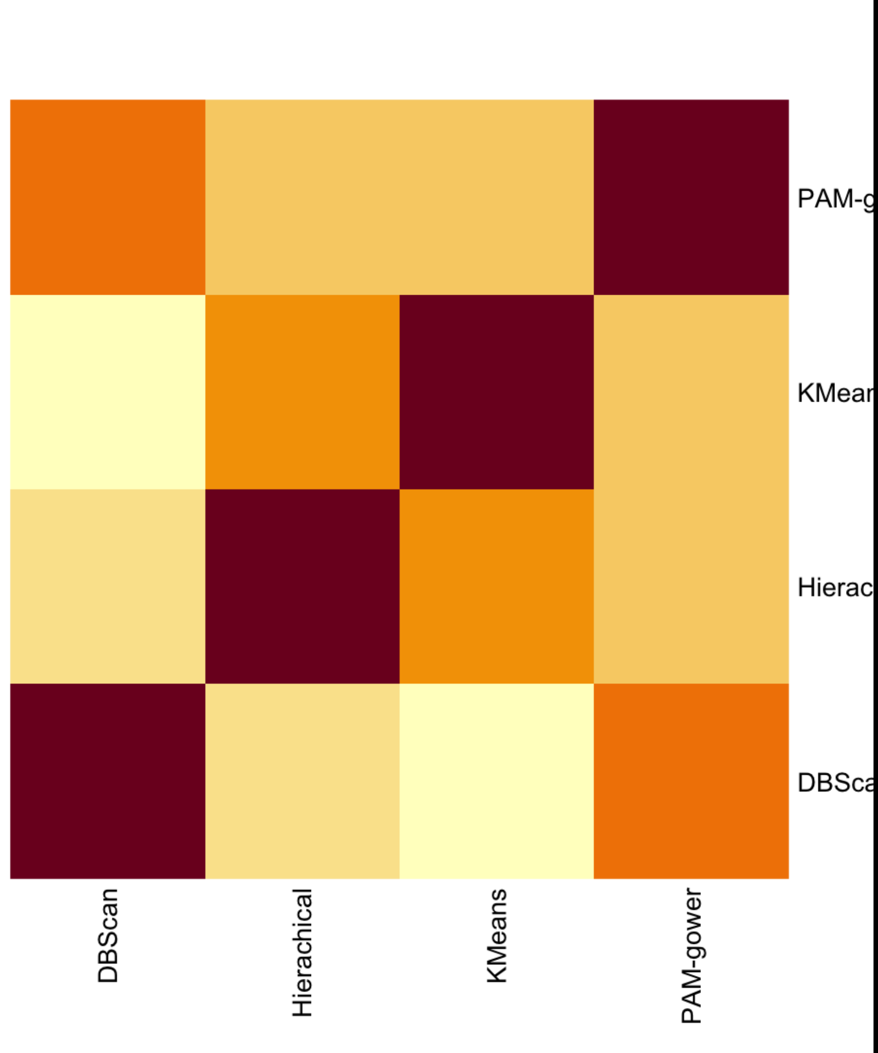


PAM vs. K-Means Clusters, k = 5





Similarity Matrix + Heatmap

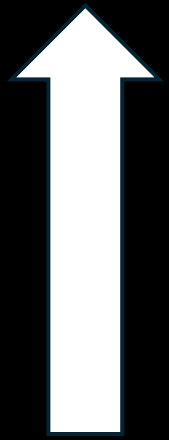


	DBScan	Hierarchical	K-Means	PAM
DBScan	1	0.4884	0.3433	0.685
Hierarchical	0.4884	1	0.6341	0.5365
K-Means	0.3433	0.6341	1	0.5231
PAM	0.685	0.5365	0.5231	1

PAM-TSNE Centroids – Distinct Separation

#	Age	G	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB	SO	SH	SF	HBP	AVG	OBP	SLG	OPS	Salaries ($K)
12	33	90	360	76	112	21	2	17	51	14	2	44	71	0	1	5	0.311	0.393	0.522	0.915	29000
39	29	138	515	75	147	31	6	20	74	6	7	59	130	0	7	2	0.285	0.357	0.485	0.842	18000
76	32	142	478	64	122	24	3	21	74	3	4	34	122	1	2	1	0.255	0.305	0.45	0.755	9000
129	33	123	422	42	106	18	1	17	61	0	0	31	132	0	2	2	0.251	0.304	0.419	0.724	4200
385	28	72	238	27	55	14	0	9	21	5	0	28	67	0	0	0	0.231	0.312	0.403	0.715	770

Superstar



Parttime

1	2	3	4	5
24	29	46	59	338

Fit Models for each Cluster

cluster	knn-rmse	tree-rmse	bag-rmse	rf-rmse
1	4237.6168	3775.797	4371.8455	4214.2594
2	2603.0117	2666.247	2807.4025	2665.0321
3	1792.3729	1622.864	1754.5978	1735.6482
4	1154.3705	1209.274	1177.8082	1116.3821
5	323.8267	317.239	321.2326	316.1857

Choose Model for each cluster based on lowest RMSE

Prediction Results with Clustering (first 10 players)

Player First	Player Last	Cluster	Actual Salary	Predicted Salary	Delta
Juan	Soto	2	$23,000,000	$17,989,899	22%
Pete	Alonso	2	$14,500,000	$16,590,909	-14%
Gleyber	Torres	3	$9,950,000	$9,226,795	7%
Anthony	Santander	3	$7,400,000	$9,260,281	-25%
Christian	Walker	4	$6,500,000	$4,700,289	28%
Alex	Verdugo	4	$6,300,000	$4,572,668	27%
Luis	Arraez	4	$6,100,000	$4,606,039	24%
Mike	Yastrzemski	4	$6,100,000	$4,483,754	26%
Kyle	Farmer	4	$5,585,000	$4,466,519	20%
Kyle	Tucker	4	$5,000,000	$4,554,702	9%

Salary Delta per Cluster (PAM-TSNE)

Cluster	min	Q1	median	Q3	max	mean	sd	n
2	-0.16	-0.0725	0.015	0.1025	0.19	1.5%	24.7%	2
3	-0.28	-0.1975	-0.115	-0.0325	0.05	-11.5%	23.3%	2
4	-0.82	-0.4025	-0.14	0.055	0.26	-18.8%	30.9%	28
5	-0.29	0.0575	0.275	0.4625	0.59	23.3%	25.5%	28

Smaller variance

- There were no arbitration players in cluster 1
- The mean delta ranged from 1.5% to 23%
- Standard deviation between 23% and 30%

Conclusion and Future Work

- Clustering made a big difference, reducing delta mean, variance
- Lowest tier predicted overpaid, but on small average salaries
- Mid tiers predicted players were underpaid
- Only 2 higher tier players, predictions were spot on
- Add more years of data (only have 2023 data here)
- Identify which players used arbitration vs settled with club