NBA “Moneyball”: An Analytical Approach to Cost-Effective Team Formation Using Hierarchical Clustering and Regression

Calvin Yu, Daniel Lisko, Toan Vang, Ruchi Bhavsar

Why this matters?

- Advanced statistics in sports drive the selection of the best players, which lead to an increased number of wins
- Achieving a well-rounded team is often easy for already successful and “rich” teams given their power and resources.
    - Less successful teams limited salary caps and it’s a challenging journey to get out of the bottom and middle tier

- Solving this problem has both economic and social implications. In general, we can infer that having better players leads to more wins and success as a team. As a result, fan culture and viewership increases, allowing for economic growth in cities.

Unsupervised Clustering Analysis of NBA Players

Overview of Clustering Model

Beautiful Soup:

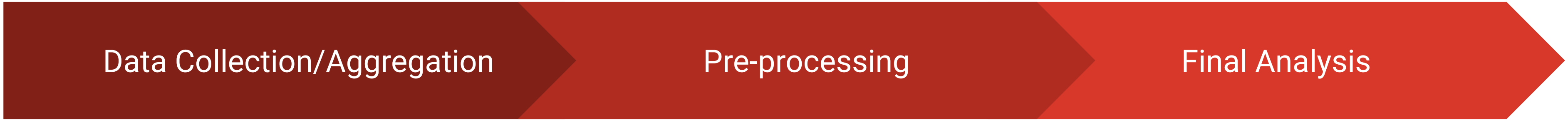
- Player Stats (2010 -2022)
- All Star Appearances

Calculated Median Statistics
Data Dimensionality Reduction
Measure Clustering Method Performances:

- Ward
- Single Linkage
- Complete Linkage
- Simple Average

Compare advanced stats for each cluster.

Investigate the top players of each cluster.



Data Collection/Aggregation

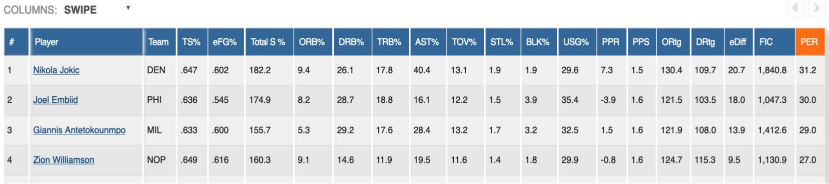
Pre-processing

Final Analysis

Data Collection

Beautiful Soup

NBA Advanced Stats: 2010 - 2022

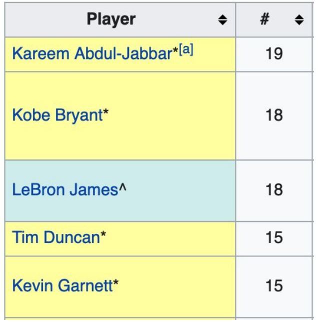


COLUMNS: SWIPE ▾

#	Player	Team	TS%	eFG%	Total S %	ORB%	DRB%	TRB%	AST%	TOV%	STL%	BLK%	USG%	PPR	PPS	ORtg	DRtg	eDiff	FIC	PER
1	Nikola Jokic	DEN	.647	.602	182.2	9.4	26.1	17.8	40.4	13.1	1.9	1.9	29.6	7.3	1.5	130.4	109.7	20.7	1,840.8	31.2
2	Joel Embiid	PHI	.636	.545	174.9	8.2	28.7	18.8	16.1	12.2	1.5	3.9	35.4	-3.9	1.6	121.5	103.5	18.0	1,047.3	30.0
3	Giannis Antetokounmpo	MIL	.633	.600	155.7	5.3	29.2	17.6	28.4	13.2	1.7	3.2	32.5	1.5	1.6	121.9	108.0	13.9	1,412.6	29.0
4	Zion Williamson	NOP	.649	.616	160.3	9.1	14.6	11.9	19.5	11.6	1.4	1.8	29.9	-0.8	1.6	124.7	115.3	9.5	1,130.9	27.0

basketball.realgm.com

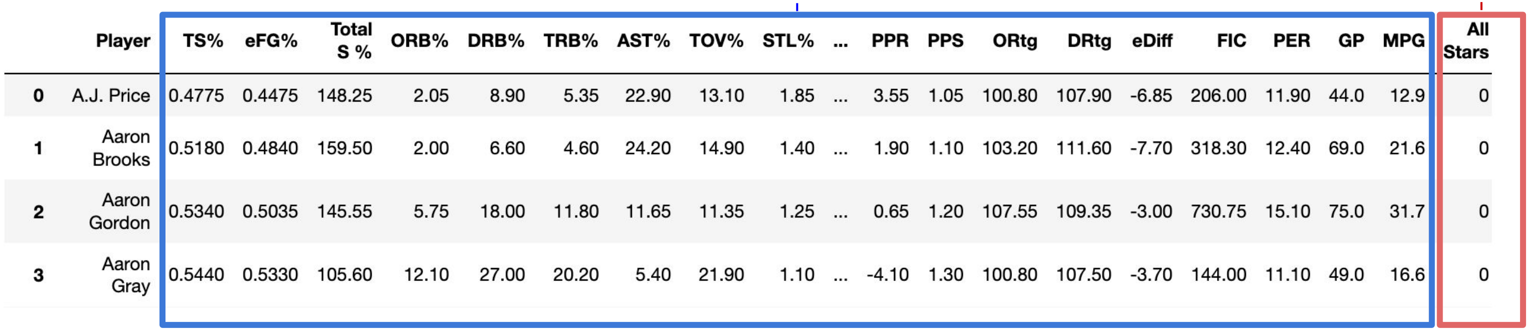
# All Star Games



Player ⬍	# ⬍
Kareem Abdul-Jabbar*[a]	19
Kobe Bryant*	18
LeBron James^	18
Tim Duncan*	15
Kevin Garnett*	15

https://en.wikipedia.org/wiki/List_of_NBA_All-Stars

Data Aggregation



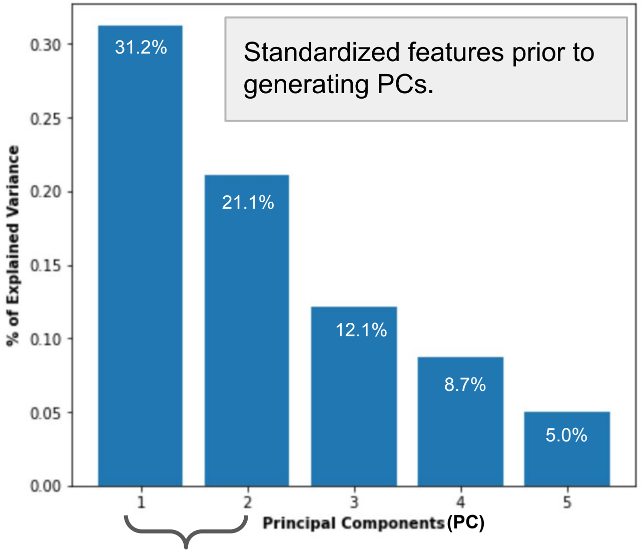
	Player	TS%	eFG%	Total S %	ORB%	DRB%	TRB%	AST%	TOV%	STL%	...	PPR	PPS	ORtg	DRtg	eDiff	FIC	PER	GP	MPG	All Stars
0	A.J. Price	0.4775	0.4475	148.25	2.05	8.90	5.35	22.90	13.10	1.85	...	3.55	1.05	100.80	107.90	-6.85	206.00	11.90	44.0	12.9	0
1	Aaron Brooks	0.5180	0.4840	159.50	2.00	6.60	4.60	24.20	14.90	1.40	...	1.90	1.10	103.20	111.60	-7.70	318.30	12.40	69.0	21.6	0
2	Aaron Gordon	0.5340	0.5035	145.55	5.75	18.00	11.80	11.65	11.35	1.25	...	0.65	1.20	107.55	109.35	-3.00	730.75	15.10	75.0	31.7	0
3	Aaron Gray	0.5440	0.5330	105.60	12.10	27.00	20.20	5.40	21.90	1.10	...	-4.10	1.30	100.80	107.50	-3.70	144.00	11.10	49.0	16.6	0

Median Advanced Stats (2010-2022 Seasons)

Preprocessing: Choosing PCs and Clustering Method

PC1 and PC2 of PCA chosen for clustering because they explain the most variance.

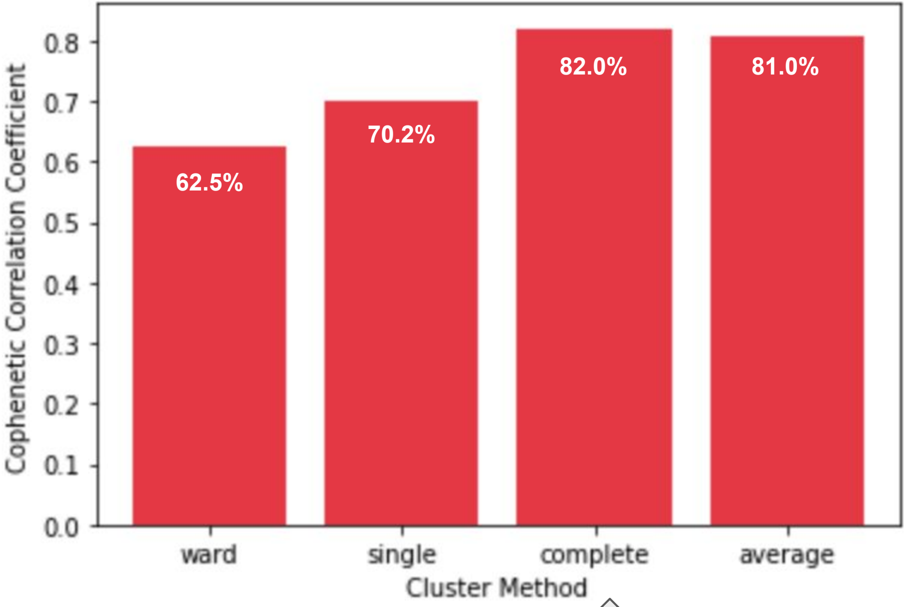
Scree Plots for PCA



Standardized features prior to generating PCs.

54.4% Total Explained Variance

Comparing Cluster Methods

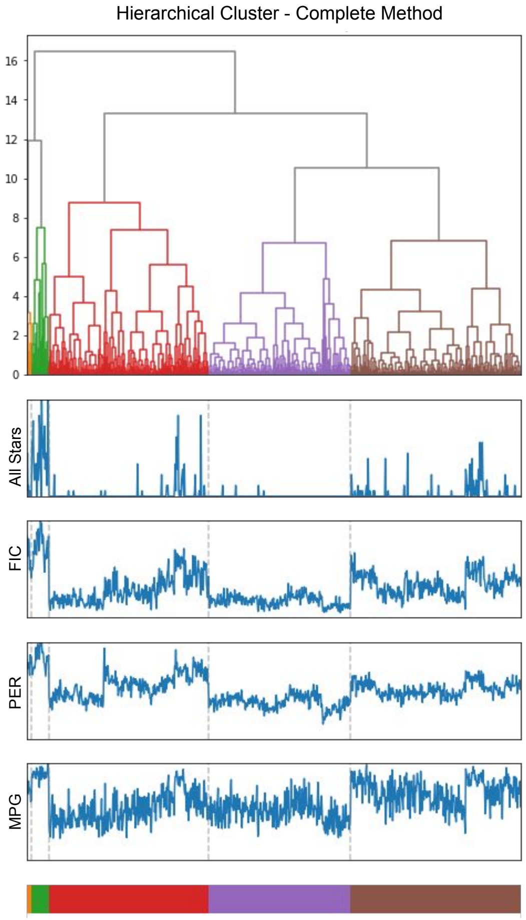


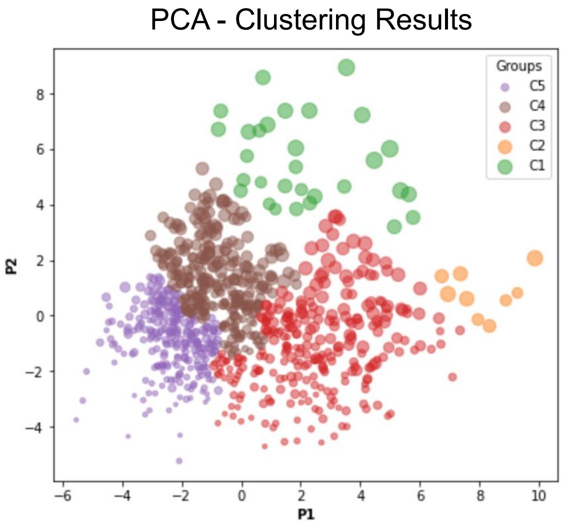
Chosen Clustering Method

Final Clustering Analysis of NBA Players

Player’s Cluster into 5 Groups

FIC: Floor Impact Factor
PER: Player Efficiency Rating
MPG: Minutes Per Game

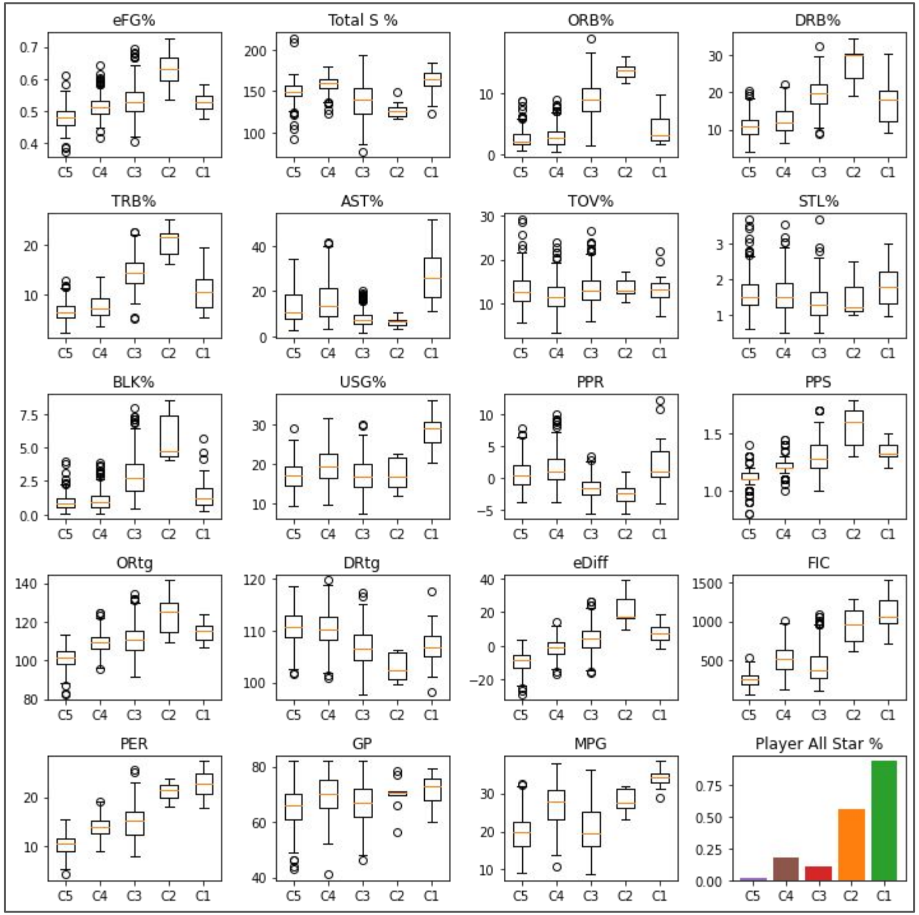

Hierarchical Cluster - Complete Method


PCA - Clustering Results

- 5 types of players in the NBA.
  - C1 & C2: Most likely high caliber players
  - C3 & C4: Most likely role players.
  - C5: Most likely bench players.

- Tight clustering within groups.
  - High separation of cluster C1 and C2.

Differences in Advanced Stats Between Groups



Group C1:

- Involved in high number of offensive plays (USG%)
- Play highest number of minutes
- Highly efficient players (PER)
- Have high impact when on the court (FIC)
- Play high number of minutes (MPG)

Group C2:

- Highly efficient players (PER)
- Have high impact when on the court (FIC).
- Elite rebounders (ORB%, DRB%, TRB%)
- Highest blocks in a game (BLK%)

High % of All-Star Players

Group C3:

- Have high impact when on the court (FIC).
- Effective rebounders (ORB%, DRB%, TRB%)
- High number blocks in a game (BLK%)
- Efficient players (PER)

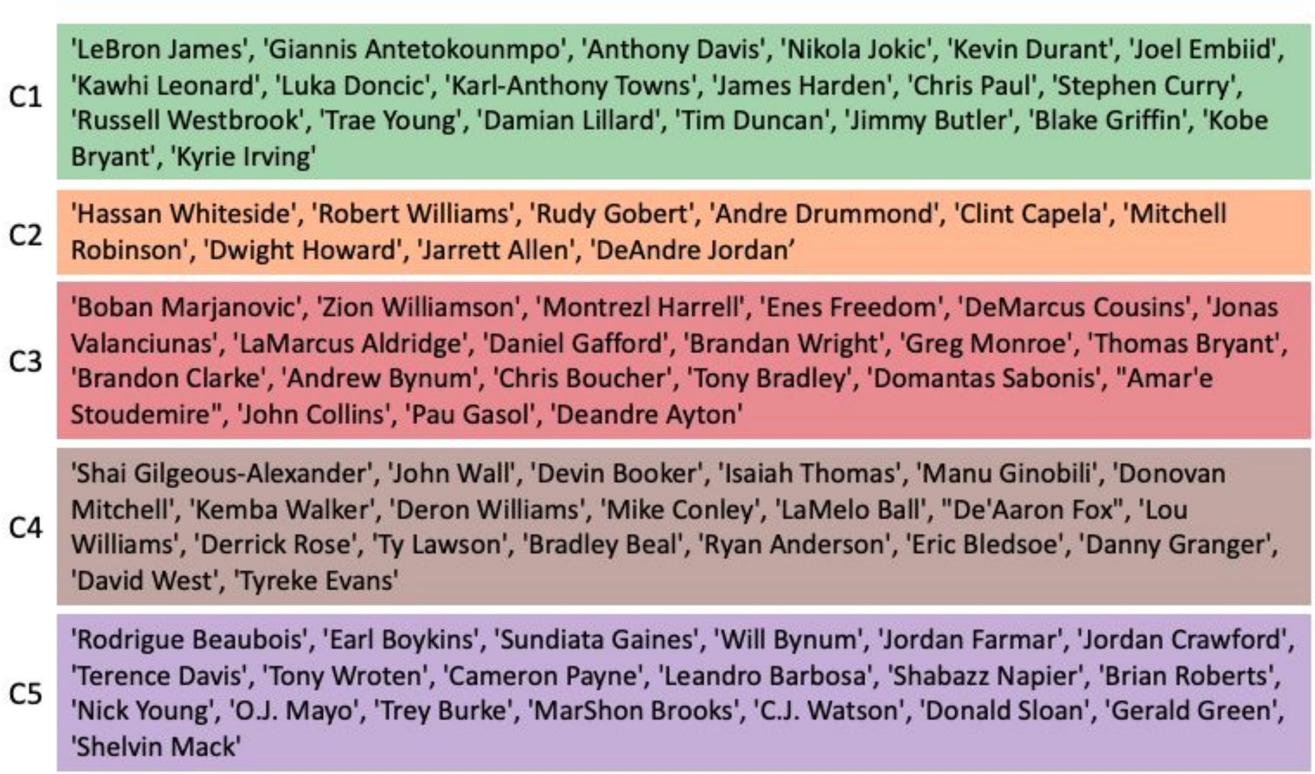
Group C4:

- Efficient players (PER)
- Have an impact when on the court (FIC)
- Play high number of minutes (MPG)

Group C5:

- Rank the lowest in many of the above many categories.
- Play lowest number of minutes.

Top 20 Players of Each Cluster - based on PER



C1	'LeBron James', 'Giannis Antetokounmpo', 'Anthony Davis', 'Nikola Jokic', 'Kevin Durant', 'Joel Embiid', 'Kawhi Leonard', 'Luka Doncic', 'Karl-Anthony Towns', 'James Harden', 'Chris Paul', 'Stephen Curry', 'Russell Westbrook', 'Trae Young', 'Damian Lillard', 'Tim Duncan', 'Jimmy Butler', 'Blake Griffin', 'Kobe Bryant', 'Kyrie Irving'	Many consider future hall of farmers.
C2	'Hassan Whiteside', 'Robert Williams', 'Rudy Gobert', 'Andre Drummond', 'Clint Capela', 'Mitchell Robinson', 'Dwight Howard', 'Jarrett Allen', 'DeAndre Jordan'	Recognized as top “Big Men” in the NBA.
C3	'Boban Marjanovic', 'Zion Williamson', 'Montrezl Harrell', 'Enes Freedom', 'DeMarcus Cousins', 'Jonas Valanciunas', 'LaMarcus Aldridge', 'Daniel Gafford', 'Brandan Wright', 'Greg Monroe', 'Thomas Bryant', 'Brandon Clarke', 'Andrew Bynum', 'Chris Boucher', 'Tony Bradley', 'Domantas Sabonis', "Amar'e Stoudemire", 'John Collins', 'Pau Gasol', 'Deandre Ayton'	Mostly “Big Men” of the NBA. ● Some All-Stars
C4	'Shai Gilgeous-Alexander', 'John Wall', 'Devin Booker', 'Isaiah Thomas', 'Manu Ginobili', 'Donovan Mitchell', 'Kemba Walker', 'Deron Williams', 'Mike Conley', 'LaMelo Ball', "De'Aaron Fox", 'Lou Williams', 'Derrick Rose', 'Ty Lawson', 'Bradley Beal', 'Ryan Anderson', 'Eric Bledsoe', 'Danny Granger', 'David West', 'Tyreke Evans'	Key Offensive Players: ● Some have All-Star appearances ● Some considered contributors to a teams success.
C5	'Rodrigue Beaubois', 'Earl Boykins', 'Sundiata Gaines', 'Will Bynum', 'Jordan Farmar', 'Jordan Crawford', 'Terence Davis', 'Tony Wroten', 'Cameron Payne', 'Leandro Barbosa', 'Shabazz Napier', 'Brian Roberts', 'Nick Young', 'O.J. Mayo', 'Trey Burke', 'MarShon Brooks', 'C.J. Watson', 'Donald Sloan', 'Gerald Green', 'Shelvin Mack'	Mainly bench players.

Conclusion on Hierarchical Clustering Model

- Hierarchical clustering showed there are five distinct clusters of players in the NBA.
- Investigation in advanced stats shows distinct players characteristics in each group.
- Differences in advanced statistics between clusters validates its use for downstream regression analysis
- Finally, investing top players of each cluster, based on PER, provided a final confirmation of the types of players that are found in each cluster.

Regression Model for Team Predictions

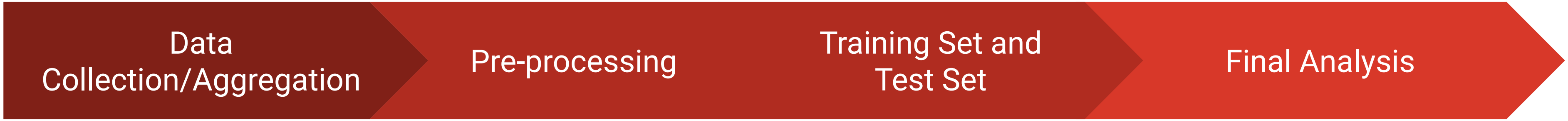
Overview of Regression Model

- Player and team data downloaded from basketball-reference.com for 2021-22 Season

- Players parsed according certain thresholds
- Unnecessary features removed

- Training set generated from season standings and advanced statistics
- Test set generated from random combinations of players

- Linear regression model is applied to project wins
- Teams compared against currently successful teams



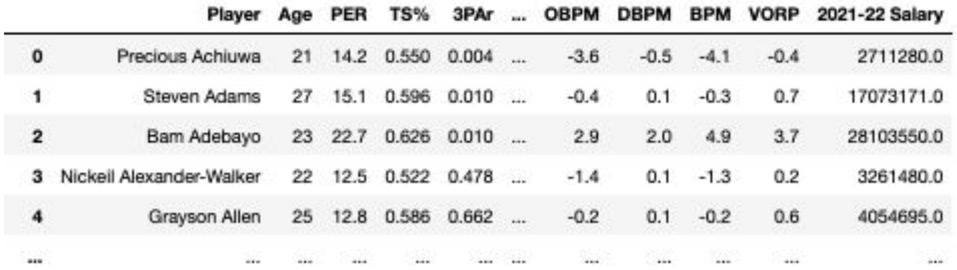
Data Collection/Aggregation

Pre-processing

Training Set and Test Set

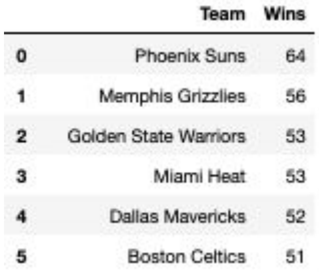
Final Analysis

Data Collection and Aggregation



	Player	Age	PER	TS%	3PAr	...	OBPM	DBPM	BPM	VORP	2021-22 Salary
0	Precious Achiuwa	21	14.2	0.550	0.004	...	-3.6	-0.5	-4.1	-0.4	2711280.0
1	Steven Adams	27	15.1	0.596	0.010	...	-0.4	0.1	-0.3	0.7	17073171.0
2	Bam Adebayo	23	22.7	0.626	0.010	...	2.9	2.0	4.9	3.7	28103550.0
3	Nickeil Alexander-Walker	22	12.5	0.522	0.478	...	-1.4	0.1	-1.3	0.2	3261480.0
4	Grayson Allen	25	12.8	0.586	0.662	...	-0.2	0.1	-0.2	0.6	4054695.0
...	...	...	...	...	...	...	...	...	...	...	...

Player Advanced Statistics and 2021-22 Salary



	Team	Wins
0	Phoenix Suns	64
1	Memphis Grizzlies	56
2	Golden State Warriors	53
3	Miami Heat	53
4	Dallas Mavericks	52
5	Boston Celtics	51

Season Standings for 2021-22 Season

Data Pre-processing

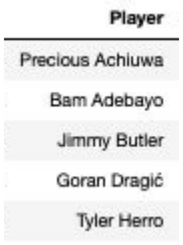
- Players filtered if minimum number of games and minutes played is not met

- Unnecessary features removed

- Players who played on multiple teams in the season removed

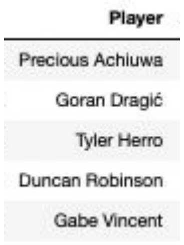
Training Set Generation

- For each NBA team, create combinations of roster players
  - Team size = 5
  - Players meeting minimum thresholds
- Average stats for each team
- Assign team wins
- Example
  - Miami Heat had 53 wins
  - Example instances would be the averaged statistics of the below teams



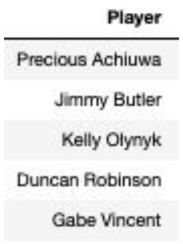
Player
Precious Achiuwa
Bam Adebayo
Jimmy Butler
Goran Dragić
Tyler Herro

,



Player
Precious Achiuwa
Goran Dragić
Tyler Herro
Duncan Robinson
Gabe Vincent

,



Player
Precious Achiuwa
Jimmy Butler
Kelly Olynyk
Duncan Robinson
Gabe Vincent

, …

“Test” Set Generation

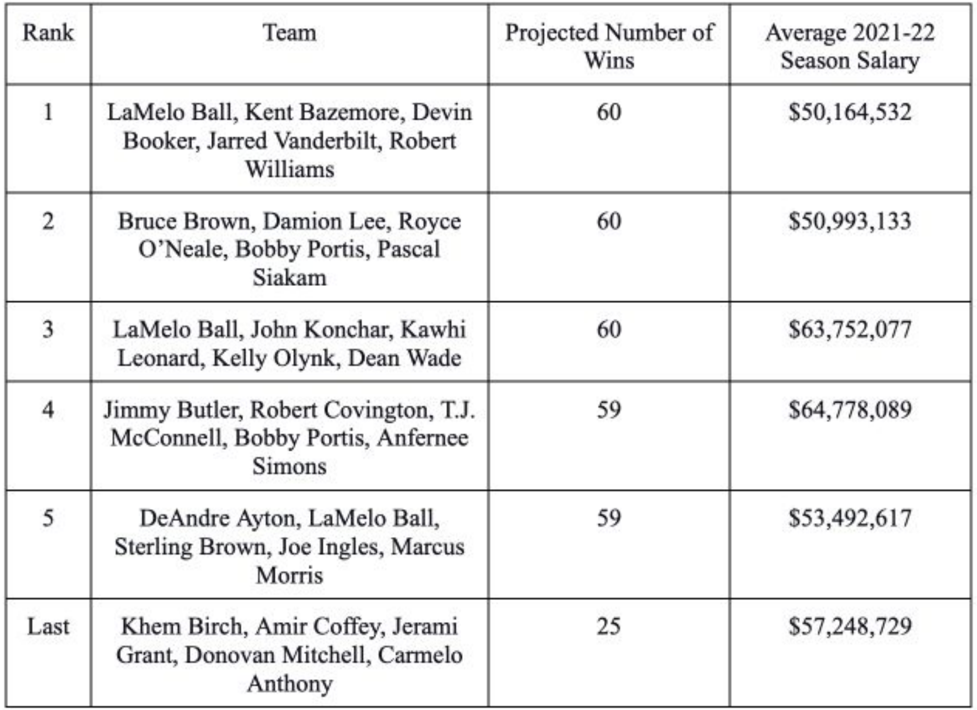
- Generate a sets of 5 person teams within combined salary range

- Randomly choose indices to reduce total number of combos

- Average team statistics an assign as test instance

Regression Model

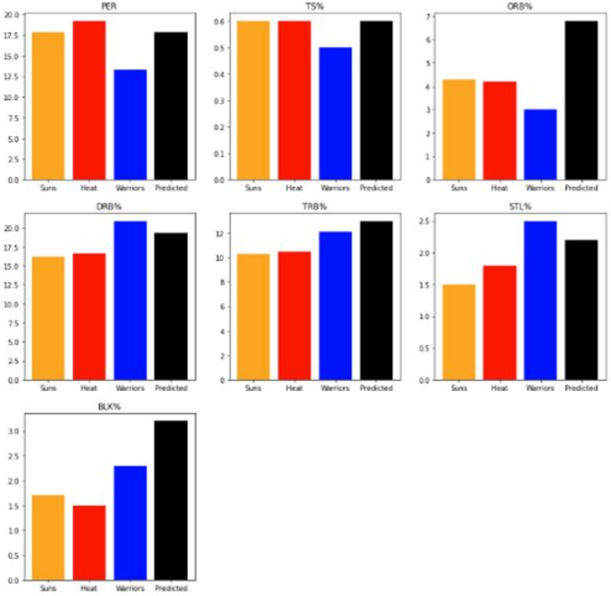
- Standardize instances
- Train linear regression model
- Predict wins for “test” set teams
- Rank according to wins



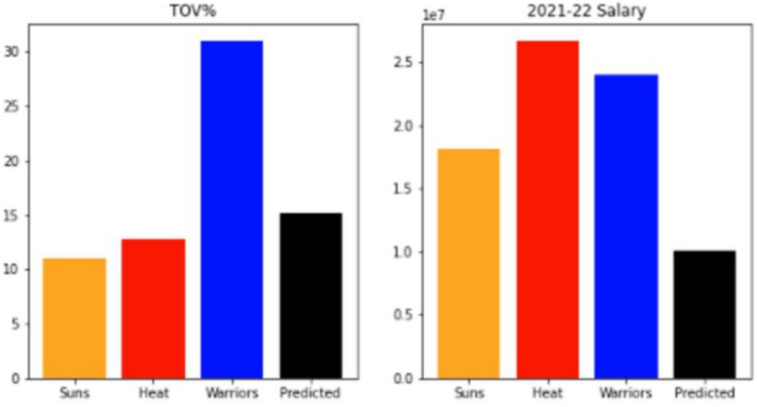
Rank	Team	Projected Number of Wins	Average 2021-22 Season Salary
1	LaMelo Ball, Kent Bazemore, Devin Booker, Jarred Vanderbilt, Robert Williams	60	$50,164,532
2	Bruce Brown, Damion Lee, Royce O’Neale, Bobby Portis, Pascal Siakam	60	$50,993,133
3	LaMelo Ball, John Konchar, Kawhi Leonard, Kelly Olynk, Dean Wade	60	$63,752,077
4	Jimmy Butler, Robert Covington, T.J. McConnell, Bobby Portis, Anfernee Simons	59	$64,778,089
5	DeAndre Ayton, LaMelo Ball, Sterling Brown, Joe Ingles, Marcus Morris	59	$53,492,617
Last	Khem Birch, Amir Coffey, Jerami Grant, Donovan Mitchell, Carmelo Anthony	25	$57,248,729

Comparison of Current Successful Teams vs Predicted Rank 1

A)



B)



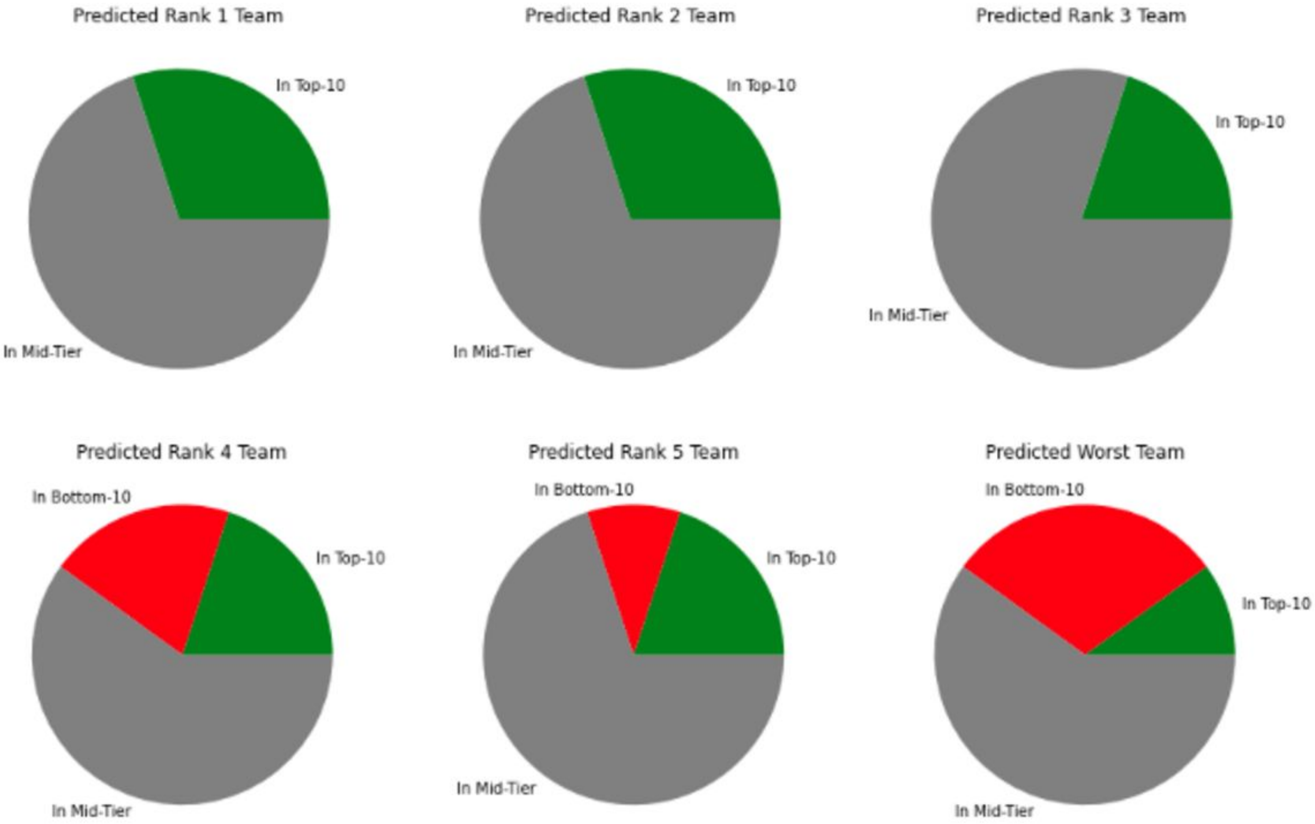
Higher Value is Better

- PER, steal percentage, etc.

Lower Value is Better

- Turnover percentage, average salaries

Portion of Players in Top-10 and Bottom-10 Teams



Conclusion and Future Work

Conclusion

- The hierarchical clustering model indicated that there are 5 distinct clusters of players and provided information on types of players found in each cluster.

- The algorithm helped us verify that player’s performance is credited to advanced statistics to some extent. We were therefore, able to find viable teams given the data.

- Observed an inconsistency with salaries; some players being paid less than others with similar advanced statistics.

Future Considerations

- Need to generate larger teams to evaluate whether the teams can perform better
- Another feature such as elo rating for the players can be included to further indicate their performance
- Update the dataset with every season for better evaluation
- With large dataset, using PCA would become difficult. This can be improved by swapping to PCA with LDA.
- Explore more algorithms that can be used instead of linear regression.