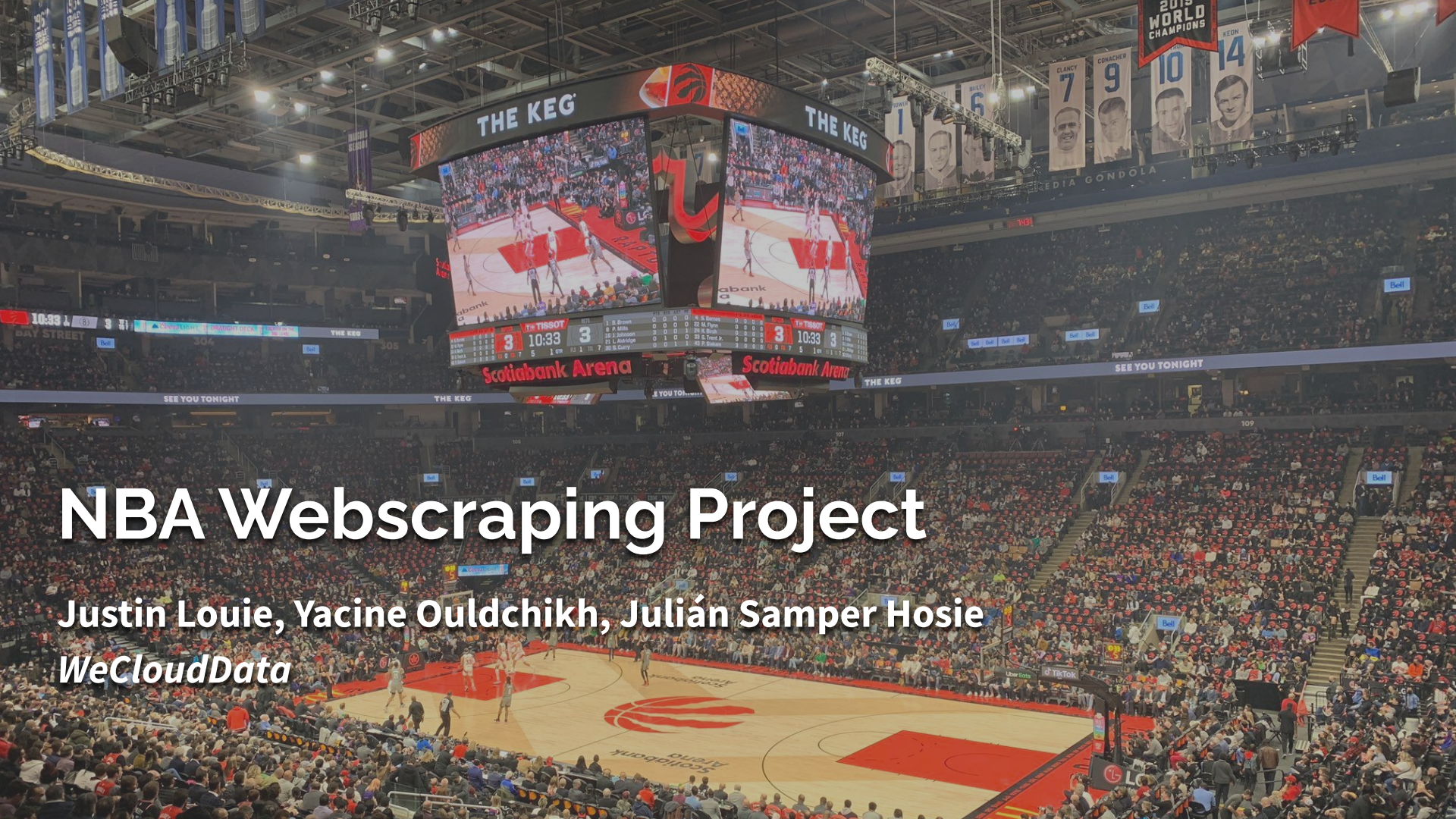


NBA Webscraping Project

Justin Louie, Yacine Ouldchikh, Julián Samper Hosie

WeCloudData

Project Structure

1. Introduction

2. Webscraping and Data Analysis
   - Data Scraping and Filtering
   - Search for the best players
   - Search for most improved players
   - Search for over- and undervalued players
   - Optimized starting lineup

3. Conclusion

Introduction

- The National Basketball Association (NBA) is a multi-billion-dollar business which features 30 teams
- The size of the market and the competitiveness of each team are two of the most important factors when it comes to generating revenue.
- Many teams cannot control the size of their market but the one thing they can control is how good the team is.
- The best way to measure a player's effectiveness is by analyzing their statistics.
    - Did a player do better in the current season vs past seasons?
    - What criteria is important when determining whether they did better or worse?
    - How much is each player truly worth?
    - How can you compare two players with different playing times?
    - Is the player's salary well spent or could the money be put to better use?

"Analytics are part and parcel of virtually everything we do now" — NBA Commissioner Adam Silver

Introduction

- Every member of this group is a NBA fan that regularly follows games as well as developments surrounding the trade market.
    - We are very interested in digging in the data to extract actionable insights.
- Companies like Second Spectrum and Genius Sports are capitalizing on the importance of integrating the use of technology in order to make data-driven decisions.
- Using data analytics in sports has been proven to be quite an effective strategy in order to turn a mediocre team into championship contenders.
- We are looking to find the most valuable players across the last 2 NBA regular seasons (2020-21 and 2021-22)

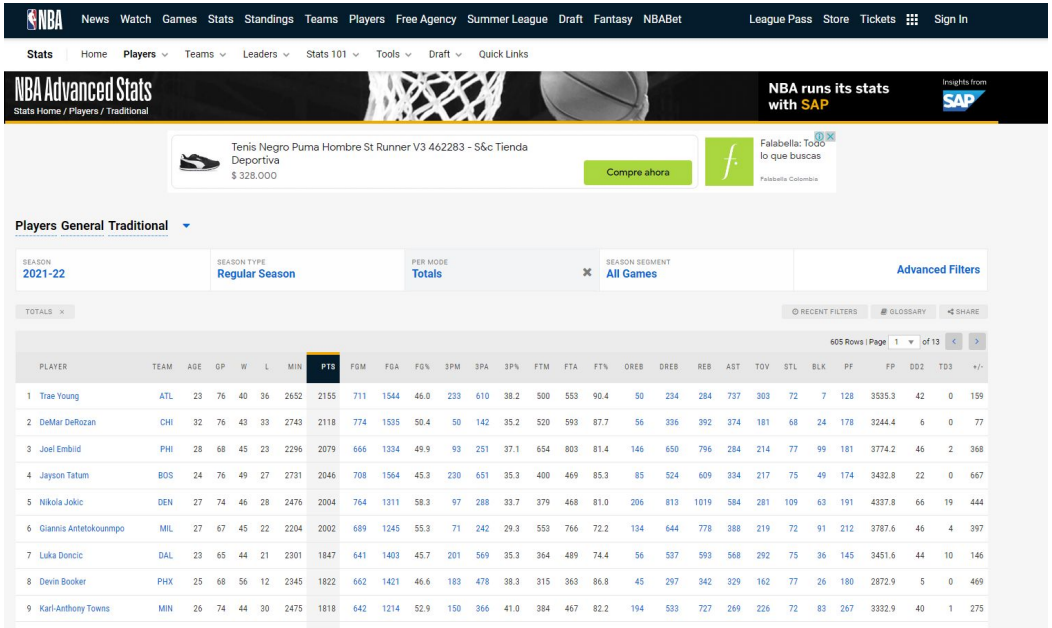
The ultimate championship goal

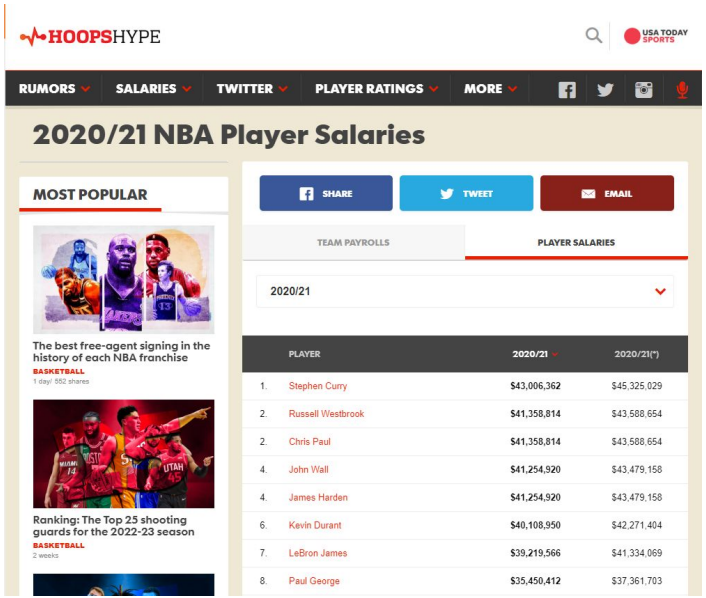
↓ ↓ ↓

Build an optimized starting lineup with the average budget of all starting lineups in the league

Data Scraping and Filtering

Screenshots of the scraped websites





First dataframe

	PLAYER	TEAM	AGE	GP	W	L	MIN	PTS	FGM	FGA	...	PF	FP	DD2	TD3	+/-	SEASON	SALARY	POSITION	HEIGHT	WEIGHT (lbs)
0	Udonis Haslem	MIA	41	1	1	0	3	53.9	26.9	26.9	...	0.0	70.1	0	0	40.4	2020-21	2564753.0	F	6-8	235.0
1	Stephen Curry	GSW	33	63	37	26	2152	33.7	11.0	22.8	...	2.0	50.5	8	0	4.1	2020-21	43006362.0	G	6-2	185.0
2	Joel Embiid	PHI	27	51	39	12	1585	33.0	10.5	20.4	...	2.8	57.1	31	0	9.2	2020-21	29542010.0	C-F	7-0	280.0
3	Bradley Beal	WAS	28	60	32	28	2147	31.5	11.2	23.2	...	2.3	45.3	4	0	-0.1	2020-21	28751774.0	G	6-4	207.0
4	Giannis Antetokounmpo	MIL	26	61	40	21	2013	30.7	11.2	19.7	...	3.0	58.8	41	7	7.3	2020-21	27528088.0	F	6-11	242.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1140	Nate Hinton	IND	23	2	0	2	2	0.0	0.0	14.8	...	0.0	-14.8	0	0	59.3	2021-22	96213.0	NaN	NaN	NaN
1141	Sam Dekker	TOR	28	1	1	0	1	0.0	0.0	0.0	...	0.0	0.0	0	0	-198.5	2021-22	350000.0	NaN	NaN	NaN
1142	Trayvon Palmer	DET	27	1	0	1	17	0.0	0.0	2.1	...	4.3	3.0	0	0	-25.8	2021-22	53176.0	NaN	NaN	NaN
1143	Tyler Hall	NYK	25	1	1	0	2	0.0	0.0	0.0	...	0.0	0.0	0	0	-93.1	2021-22	53176.0	NaN	NaN	NaN
1144	Zylan Cheatham	UTA	26	1	0	1	5	0.0	0.0	20.8	...	0.0	0.0	0	0	-90.3	2021-22	256734.0	NaN	NaN	NaN

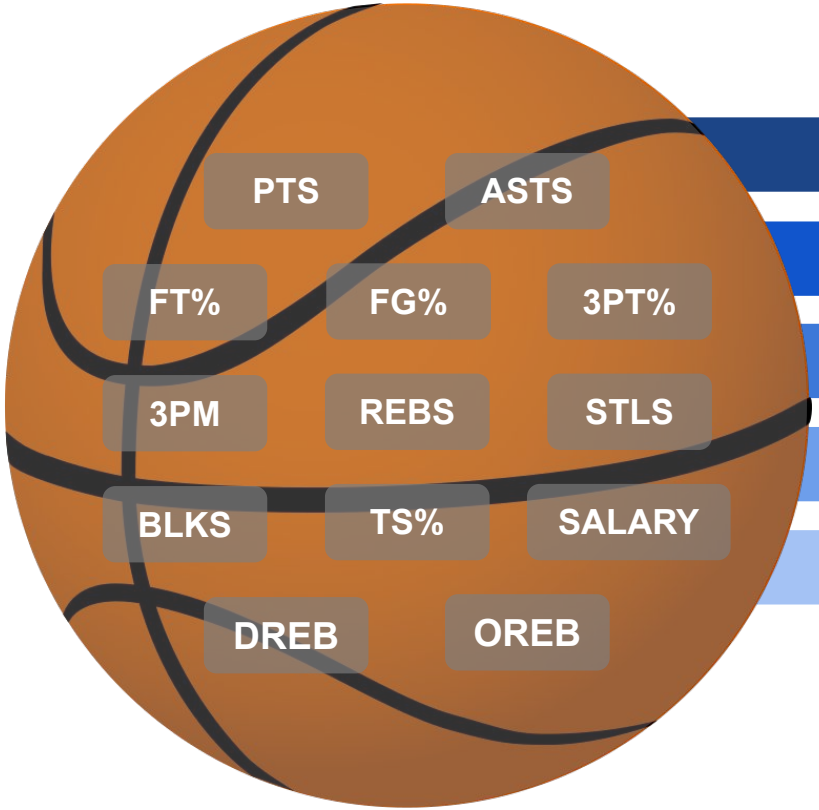
1145 rows × 34 columns

Cleaned dataframe

	PLAYER	TEAM	AGE	GP	W	MIN	PTS	FGM	FGA	FG%	...	PTS/FGA	AST/TOV	STL/PF	STL+BLK	PTS+AST	PTS+AST / FGA+TOV	FTR	AOREB	A
1	Stephen Curry	GSW	33	63	37	2152	33.7	11.0	22.8	48.2	...	1.478070	1.694444	0.650000	1.4	5.524590	0.209265	28.95	8.62	
2	Joel Embiid	PHI	27	51	39	1585	33.0	10.5	20.4	51.3	...	1.617647	0.916667	0.392857	2.7	10.000000	0.416667	60.78	21.31	
3	Bradley Beal	WAS	28	60	32	2147	31.5	11.2	23.2	48.5	...	1.357759	1.419355	0.521739	1.6	7.159091	0.272209	33.19	25.53	
4	Giannis Antetokounmpo	MIL	26	61	40	2013	30.7	11.2	19.7	56.9	...	1.558376	1.729730	0.433333	2.6	4.796875	0.204995	52.79	14.17	
5	Kevin Durant	BKN	32	35	23	1157	29.3	10.1	18.8	53.7	...	1.558511	1.648649	0.363636	2.2	4.803279	0.213479	39.36	5.19	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1075	Matisse Thybulle	PHI	25	66	43	1685	8.1	3.2	6.4	50.0	...	1.265625	2.000000	0.735294	4.0	5.062500	0.703125	14.06	28.12	
1076	Tomas Satoransky	WAS	30	55	21	906	7.9	2.9	7.6	37.7	...	1.039474	4.055556	0.428571	1.4	1.082192	0.115127	27.63	22.00	
1080	Andre Iguodala	GSW	38	31	20	603	7.4	2.7	7.2	38.0	...	1.027778	4.000000	0.800000	2.9	1.088235	0.122274	16.67	21.67	
1093	Chris Chiozza	GSW	26	34	21	372	6.5	2.3	7.8	29.6	...	0.833333	2.032258	0.419355	1.3	1.031746	0.094656	3.85	25.71	
1105	Jericho Sims	NYK	23	41	19	555	5.8	2.5	3.5	72.2	...	1.657143	1.076923	0.162791	2.1	4.142857	0.863095	54.29	36.36	

771 rows × 44 columns

Metrics Used



KPI's

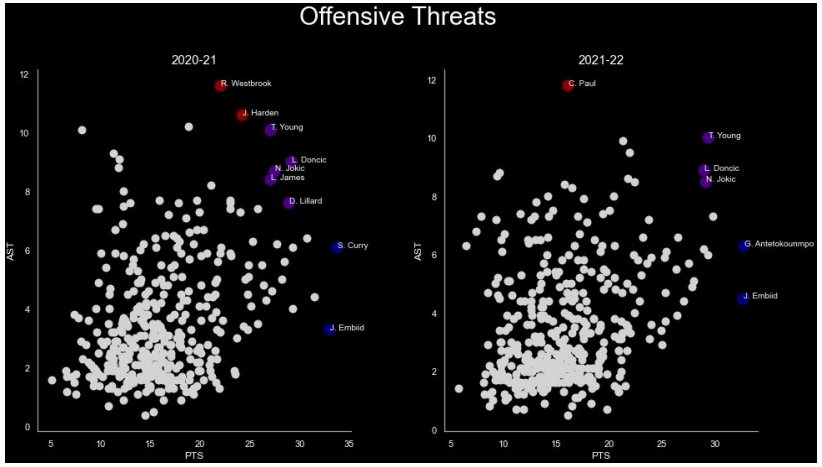
Rate of Change Analysis

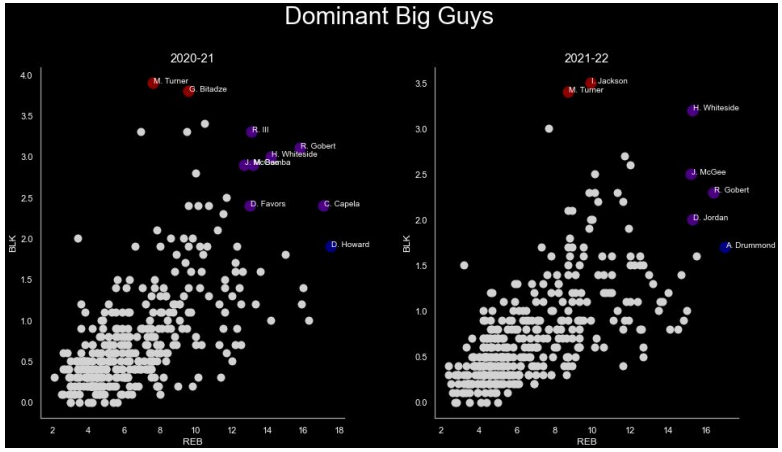
Percentile Analysis

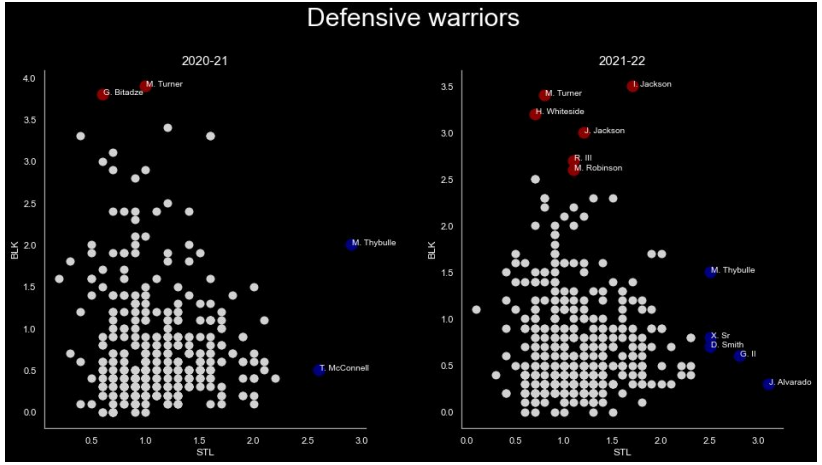
Relative Performance Analysis

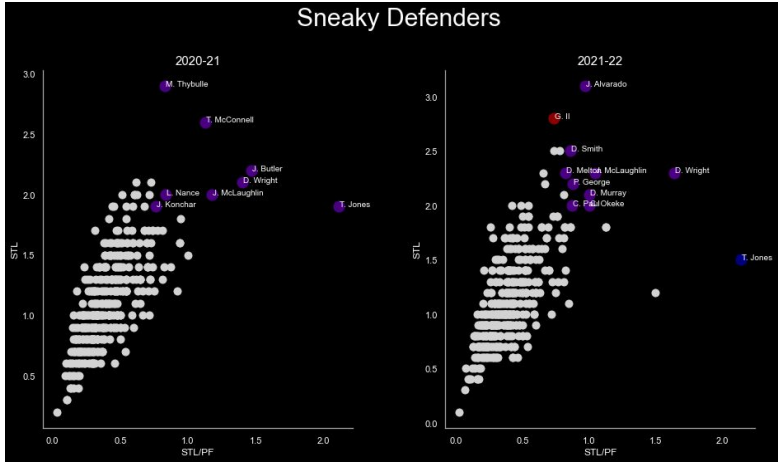
Salary vs Performance

Search for best players





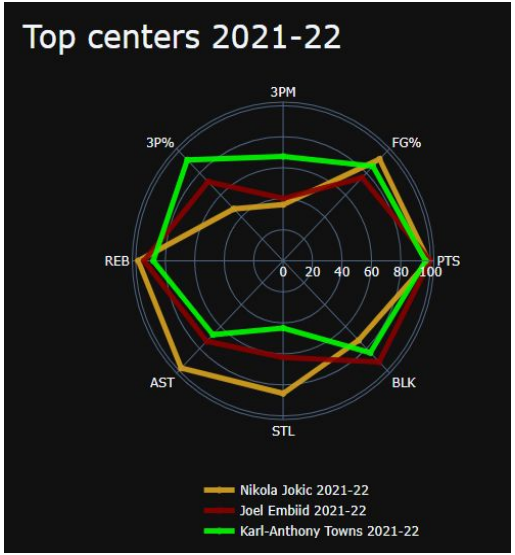




Top Centers

```
statsForCenters = [('MIN', 1), ('PTS', 1), ('AST', 0.5), ('PTS+AST / FGA+TOV', 0.7),
                   ('FG%', 0.6), ('3PM', 0.5), ('3P%', 0.5), ('TS%', 0.6), ('FT%', 0.6), ('PTS/FGA', 0.8),
                   ('STL', 0.5), ('REB', 0.8), ('OREB', 0.8), ('DREB', 0.7),
                   ('STL/PF', 0.5), ('BLK', 0.6), ('STL+BLK', 0.5)]
KPIs_df=customKPI(KPIs_df, 'centers_KPI', 'CENTER', statsForCenters)
KPIs_df[KPIs_df['SEASON']=='2021-22'][['PLAYER', 'SEASON',  'TEAM', 'PTS', 'FG%', 'BLK', 'REB', 'AST', 'centers_KPI'
```

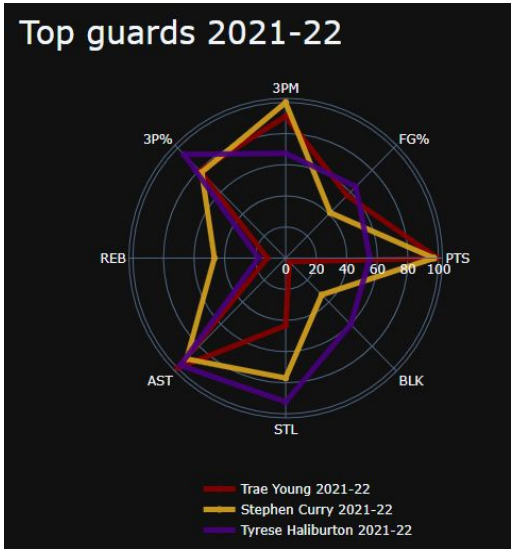
	PLAYER	SEASON	TEAM	PTS	FG%	BLK	REB	AST	centers_KPI
551	Nikola Jokic	2021-22	DEN	98.51	91.83	72.18	98.18	97.60	80.49
544	Joel Embiid	2021-22	PHI	99.61	75.29	90.99	94.94	73.80	79.28
560	Karl-Anthony Towns	2021-22	MIN	96.17	85.21	82.81	88.78	67.51	78.15
668	Jarrett Allen	2021-22	CLE	72.11	98.70	89.36	93.19	20.43	70.81
688	Rudy Gobert	2021-22	UTA	68.09	99.35	96.50	99.61	3.76	70.77
597	Bam Adebayo	2021-22	MIA	86.45	89.04	72.18	90.01	64.40	70.69
595	Jonas Valanciunas	2021-22	NOP	86.45	87.48	76.39	96.89	54.41	70.44
719	Isaiah Hartenstein	2021-22	LAC	62.45	95.59	96.50	85.41	75.81	70.35
567	Kristaps Porzingis	2021-22	WAS	94.68	55.25	94.94	86.32	50.52	68.48
599	Deandre Ayton	2021-22	PHX	85.93	96.37	66.67	94.94	20.43	68.17
571	Anthony Davis	2021-22	LAL	93.13	85.60	96.50	86.90	54.41	68.11
601	Mike Muscala	2021-22	OKC	85.41	53.76	90.99	70.62	5.58	67.98
933	Robert Williams III	2021-22	BOS	21.73	99.74	98.05	92.15	39.36	67.87



Top Guards

```
statsForGuards = [('MIN', 1), ('PTS', 1), ('AST', 0.8), ('FG%', 0.8), ('3PM', 0.7), ('3P%', 0.5),
                  ('STL', 0.5), ('REB', 0.3), ('PTS/FGA', 0.8), ('AST/TOV', 0.5),
                   ('STL/PF', 0.4), ('TS%', 0.8), ('FT%', 0.6)]
KPIs_df = customKPI(percentiles_df, 'guards_KPI', 'GUARD', statsForGuards)
KPIs_df[KPIs_df['SEASON']=='2021-22'][['PLAYER', 'SEASON', 'TEAM', 'PTS', 'AST', 'FG%', '3PM', '3P%', 'STL', 'POSITI
```

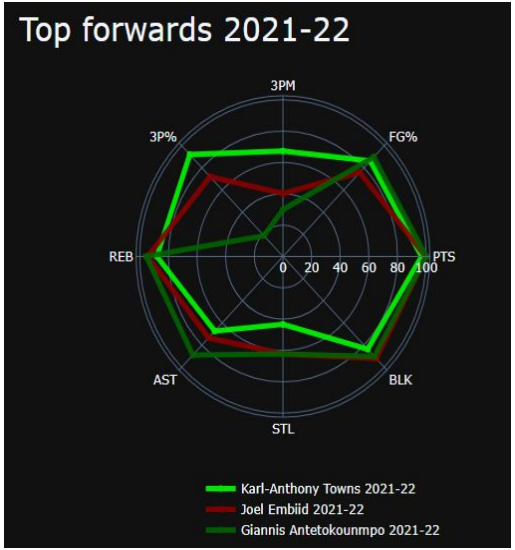
	PLAYER	SEASON	TEAM	PTS	AST	FG%	3PM	3P%	STL	POSITION	GUARD	guards_KPI
550	Trae Young	2021-22	ATL	98.90	99.22	55.97	89.49	73.99	44.03	G	True	81.42
559	Stephen Curry	2021-22	GSW	96.37	91.37	40.01	99.87	72.37	77.17	G	True	79.94
752	Tyrese Haliburton	2021-22	IND	54.99	97.34	63.55	64.85	92.41	93.26	G	True	77.39
618	Jrue Holiday	2021-22	MIL	82.56	95.20	75.94	64.85	90.73	93.26	G	True	76.99
585	Darius Garland	2021-22	CLE	89.49	97.80	57.13	76.72	74.58	70.10	G	True	75.75
583	Desmond Bane	2021-22	MEM	90.21	57.98	56.61	95.59	96.76	82.43	G	True	75.30
554	Devin Booker	2021-22	PHX	97.67	80.09	59.27	82.17	74.58	61.80	G	True	74.67
579	LaMelo Ball	2021-22	CHA	91.12	97.60	33.72	91.44	78.47	93.26	G	True	74.57
561	Kyrie Irving	2021-22	BKN	95.98	83.20	61.35	91.44	93.64	77.17	G	True	73.86
591	James Harden	2021-22	PHI	87.55	99.09	21.40	64.85	38.98	61.80	G	True	73.65
552	Luka Doncic	2021-22	DAL	98.18	98.44	54.28	87.94	56.10	61.80	F-G	True	73.50



Top Forwards

```
statsForForwards = [('MIN', 1), ('PTS', 1), ('AST', 0.7), ('PTS+AST / FGA+TOV', 0.7),
                    ('FG%', 0.8), ('3PM', 0.3), ('3P%', 0.3), ('TS%', 0.8), ('FT%', 0.6), ('PTS/FGA', 0.8),
                    ('REB', 0.8), ('OREB', 0.8), ('DREB', 0.7), ('BLK', 0.5), ('STL+BLK', 0.5)]
KPIs_df = customKPI(KPIs_df, 'forwards_KPI', 'FORWARD', statsForForwards)
KPIs_df[KPIs_df['SEASON']=='2021-22'][['PLAYER', 'SEASON',  'TEAM', 'PTS', 'AST', 'FG%', 'REB', 'BLK', 'forwards_KPI
```

	PLAYER	SEASON	TEAM	PTS	AST	FG%	REB	BLK	forwards_KPI
560	Karl-Anthony Towns	2021-22	MIN	96.17	67.51	85.21	88.78	82.81	82.78
544	Joel Embiid	2021-22	PHI	99.61	73.80	75.29	94.94	90.99	81.81
543	Giannis Antetokounmpo	2021-22	MIL	99.74	89.23	88.46	95.78	89.36	79.28
602	JaVale McGee	2021-22	PHX	85.41	5.58	95.98	98.57	97.67	74.17
548	LeBron James	2021-22	LAL	98.90	86.90	84.11	71.66	76.39	73.83
624	Domantas Sabonis	2021-22	SAC	80.87	82.23	91.44	95.78	45.07	73.13
597	Bam Adebayo	2021-22	MIA	86.45	64.40	89.04	90.01	72.18	73.03
719	Isaiah Hartenstein	2021-22	LAC	62.45	75.81	95.59	85.41	96.50	72.93
600	Christian Wood	2021-22	HOU	85.41	46.24	75.94	92.48	79.44	72.81
567	Kristaps Porzingis	2021-22	WAS	94.68	50.52	55.25	86.32	94.94	72.41
553	Kevin Durant	2021-22	BKN	98.18	88.52	82.04	64.53	72.18	72.23



Search for most improved players

We calculated the rate of change for our custom KPIs

$$ROC = 100 * \left(\frac{KPI_{21\text{-}22}}{KPI_{20\text{-}21}} - 1 \right)$$

Guards KPI	
Player	**ROC**
Josh Green	**94.3**
Matisse Thybulle	**66.7**
Cam Reddish	**62.4**
Troy Brown	**59.7**
Gary Harris	**55.8**
Isaac Okoro	**55.8**
Justice Winslow	**54.5**
Ben McLemore	**52.3**
Gabe Vincent	**48.9**
Tyrese Maxey	**45.0**

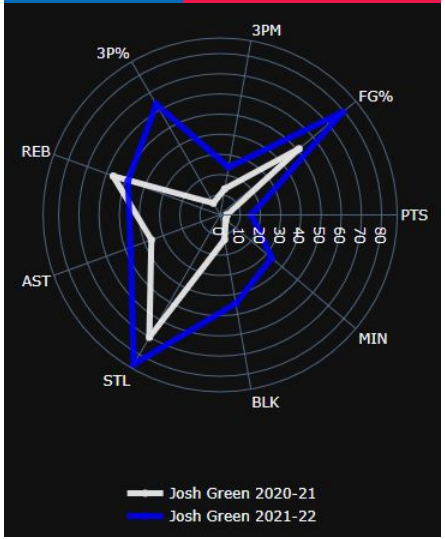
Forwards KPI	
Player	**ROC**
Caleb Martin	**93.8**
PJ Tucker	**66.9**
Mathisse Thybulle	**59.3**
Grant Williams	**54.6**
Cody Martin	**48.4**
Isaac Okoro	**45.5**
Max Strus	**44.8**
Deni Avidja	**43.8**
Cameron Johnson	**37.8**
Obi Toppin	**34.9**

Centers KPI	
Player	**ROC**
Mike Muscala	**26.5**
Goga Bitadze	**23.7**
JaVale McGee	**22.7**
LaMarcus Aldridge	**19.6**
Isaiah Hartenstein	**19.0**
Damian Jones	**17.9**
Bismack Biyombo	**13.7**
Jalen McDaniels	**13.6**
Hassan Whiteside	**13.4**
Mo Bamba	**13.3**

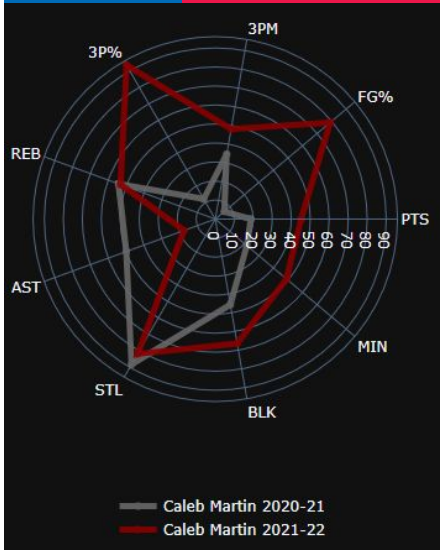
Most improved players

Radar chart visualization of the evolution of the MIPs by position:

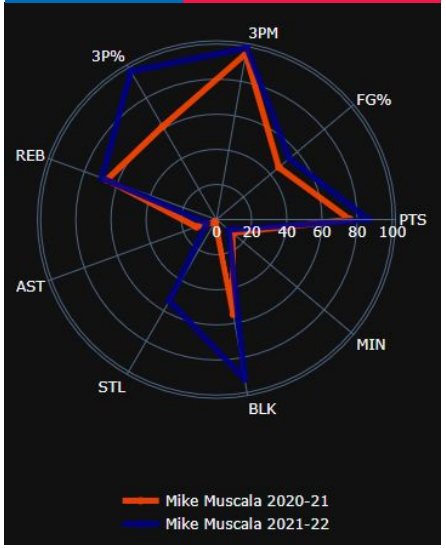
Guard Josh Green (DAL)



Forward Caleb Martin (MIA)



Center Mike Muscala (OKC)



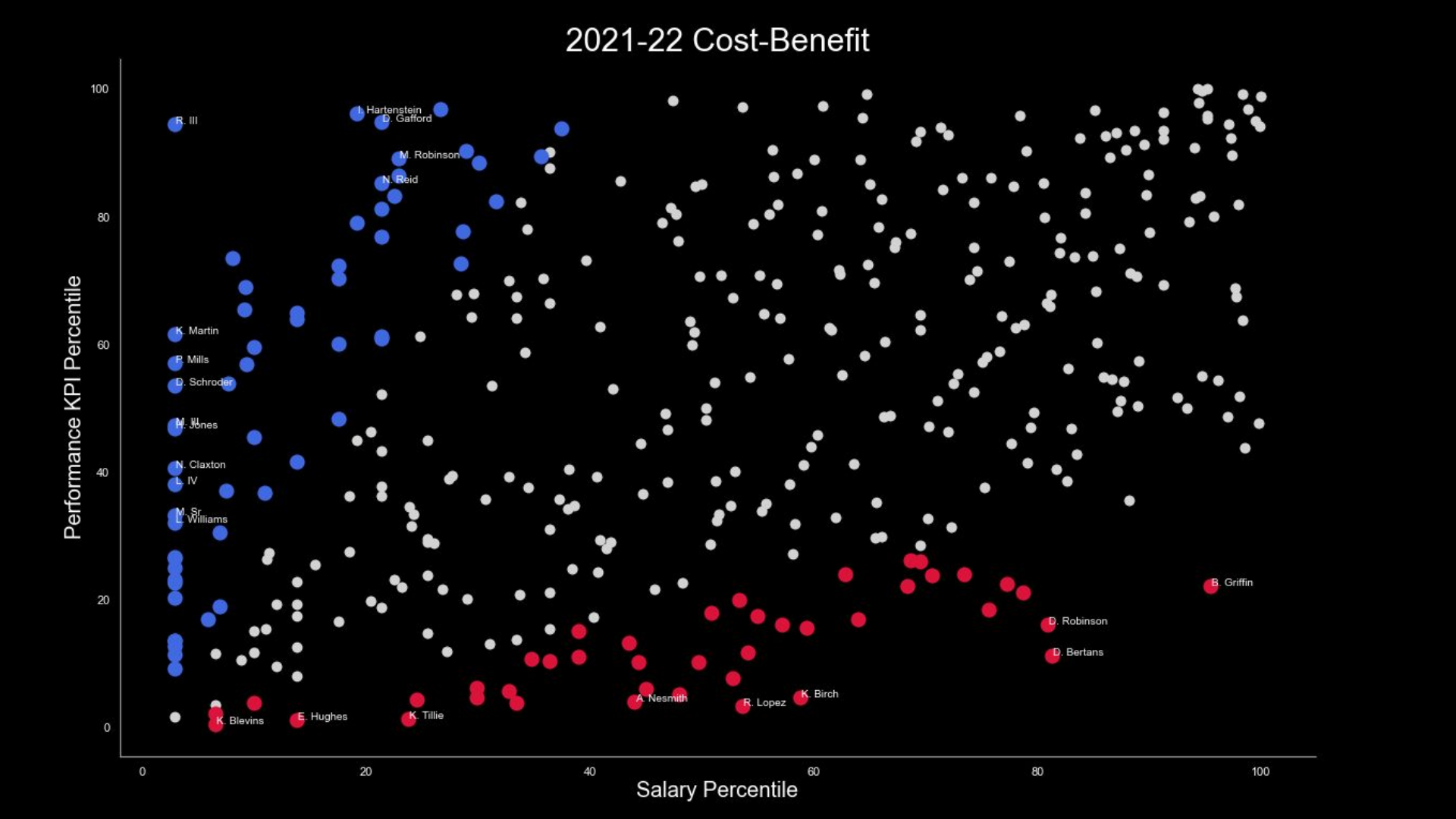
Search for over- and undervalued players

Having an approximation of each players' performance, and the salary data, we can check if the players that get paid the most are also the most valuables on the court

Salary vs Performance

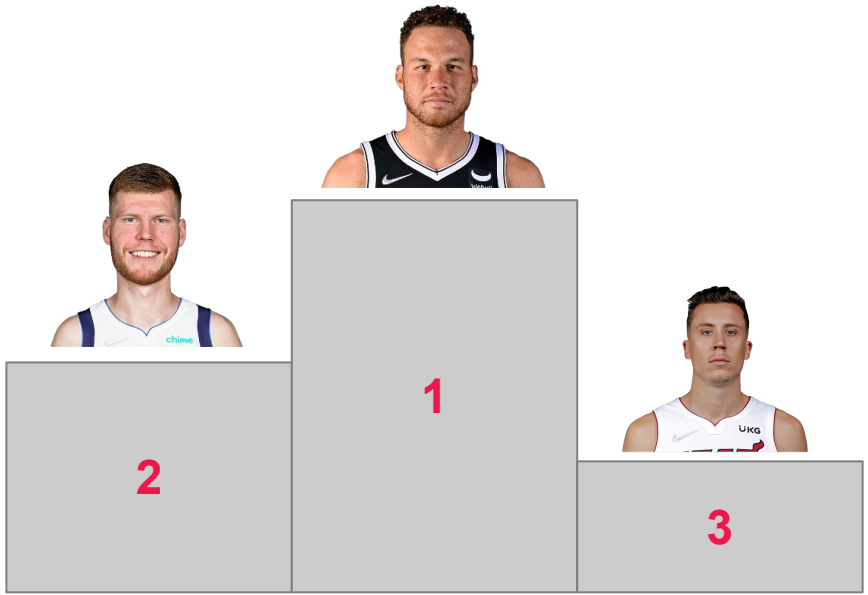
In order to have a numeric approach to this idea, we can compare each player's performance KPI percentile with his salary percentile

- A combination of high salary percentile and low KPI percentile could mean a player is being overpaid
- Similarly, a player with a high KPI percentile and low salary percentile could be undervalued


2021-22 Cost-Benefit
Performance KPI Percentile
Salary Percentile
0
20
40
60
80
100
R. III
I. Hartenstein
D. Gafford
M. Robinson
N. Reid
K. Martin
P. Mills
D. Schroder
M. III
H. Jones
N. Claxton
L. IV
M. Sr.
L. Williams
B. Griffin
D. Robinson
D. Bertans
A. Nesmith
R. Lopez
K. Birch
K. Blevins
E. Hughes
K. Tillie

Top overpaid players

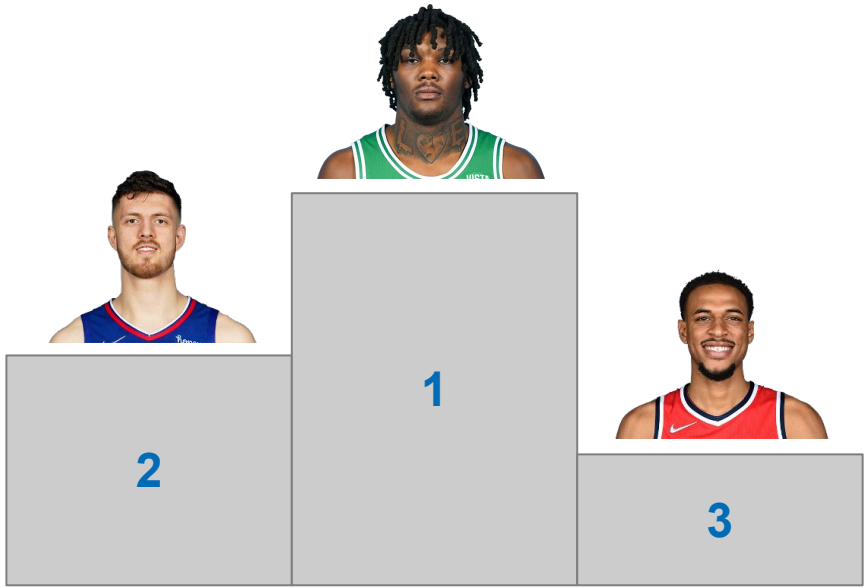
Overpaid			
Player	KPI PCTL	Salary PCTL	Salary vs KPI
Blake Griffin	22	95	**-73**
Davis Bertrans	11	81	**-70**
Duncan Robinson	16	81	**-65**
Troy Doug McDermott	21	78	**-58**
Terrence Ross	18	75	**-57**
Joe Ingles	23	77	**-55**
Paul George	44	98	**-55**
Khem Birch	5	59	**-54**
Tim Hardaway	36	88	**-53**
Russell Westbrook	48	99	**-52**



Images: ESPN

Top underpaid players

Underpaid			
Player	KPI PCTL	Salary PCTL	Salary vs KPI
Robert Williams III	95	3	**-73**
Isaiah Hartenstein	96	19	**-70**
Daniel Gafford	95	21	**-65**
Desmond Bain	97	27	**-58**
Mitchell Robinson	89	23	**-57**
Amir Coffey	74	8	**-55**
Naz Reid	85	21	**-55**
Jalen Brunson	86	23	**-54**
Jordan Poole	90	29	**-53**
Malik Monk	83	23	**-52**



Images: ESPN

Optimized starting lineup

Salary budget selection

- Function to get the average starting lineup salary for a given season.
    - Average starting lineup in 2020-21 season was $89m
    - Average starting lineup in 2021-22 season was $92m
- Assumed that the average starting lineup season in 2022-23 would be closer to $95m
    - Appx $20m for each player in a starting lineup

Optimized starting lineup

Explain selection logic

- MERGE → KPIs table + Player/Salary table
- FILTER → recent season + salary < $25m
- FILTER → top 10 by position (Guard - Forward - Center)
- We then hand picked a selection of great players by combining KPI and salary →
    - 1st and 4th best guards
    - 1st and 4th best centers
    - 4th best forward

```
KPIs_df2=pd.merge(KPIs_df2, final_df[['PLAYER', 'SEASON', 'SALARY']], on=['PLAYER', 'SEASON'],
```

```
KPIs_df2[(KPIs_df2['SEASON']=='2021-22') & (KPIs_df2['SALARY_VALUE']<25000000)]
```

	PLAYER	SEASON	POSITION	KPI_bestPosition	SALARY_VALUE
382	Trae Young	2021-22	G	99.214660	8326471.0
555	Tyrese Haliburton	2021-22	G	98.167539	4023600.0
414	Darius Garland	2021-22	G	97.382199	7040880.0
431	JaVale McGee	2021-22	C-F	97.164948	5000000.0
412	Desmond Bane	2021-22	G	96.858639	2033160.0
446	Domantas Sabonis	2021-22	F-C	96.649485	18500000.0
525	Isaiah Hartenstein	2021-22	C-F	96.134021	1729217.0
429	Christian Wood	2021-22	F	95.876289	13666667.0
408	LaMelo Ball	2021-22	G	95.549738	8231760.0
519	Daniel Gafford	2021-22	F-C	94.845361	1782621.0

Optimized starting lineup

Jrue Holiday (G)

Trae Young (G)

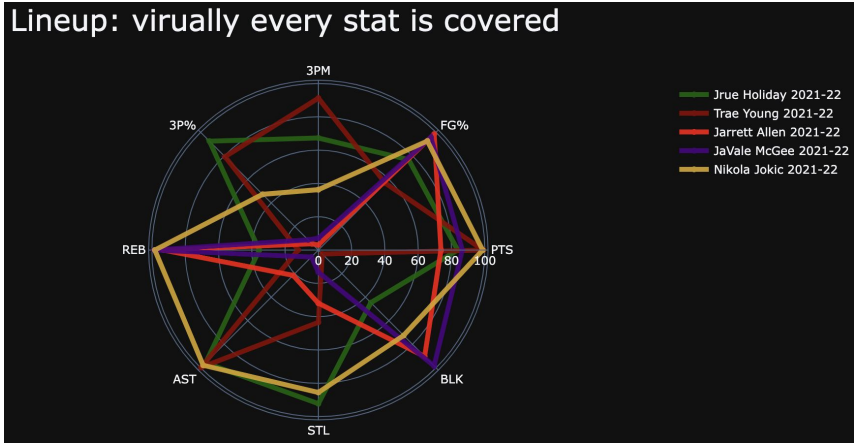
Jarrett Allen (C)

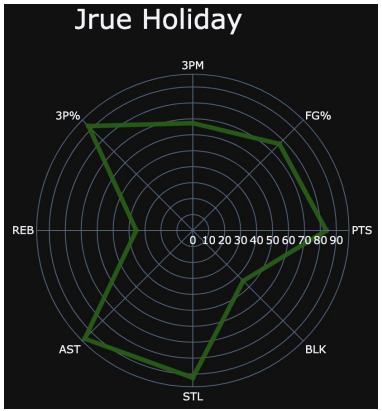
JaVale McGee (C-F)

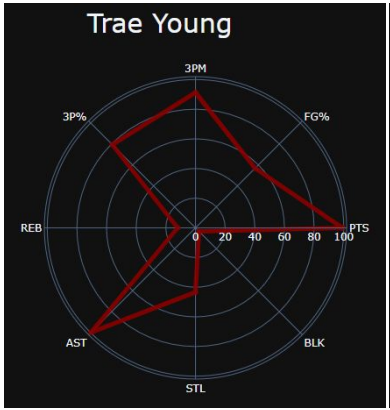
Nikola Jokic (C)

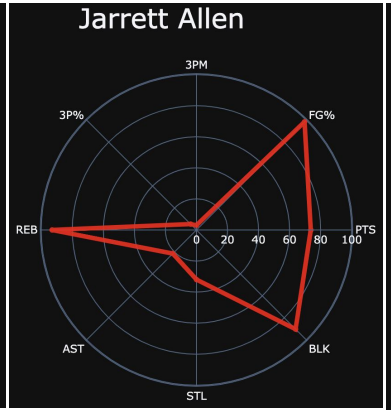
Optimized starting lineup

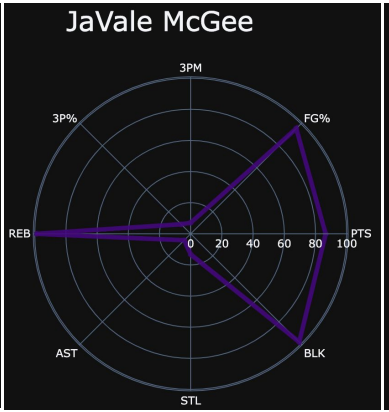
Radar charts of starting lineup

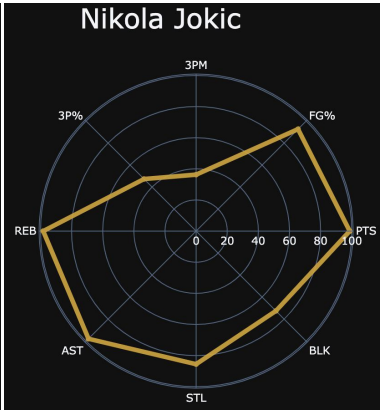












Conclusions

- Excited to finalize a web scraping project that involves analyzing raw data with pandas
  - We are satisfied with our results given our level of knowledge

- Impressive to see what valuable insights you can extract by manipulating the data
  - Benefits of open-source libraries
  - Even more capabilities with advanced data visualization libraries

- This project could have been much more interesting with the use of machine learning
  - Advanced regression analysis with scatter plots
  - Deeper analysis regarding under- and over-valued players.
  - Objective selection of starting lineup

Thank You