NBA DATA ANALYSIS PROJECT

John Olajire

502 Final Semester Project

Table of Contents

Dataset

Dataset Description

The dataset used was gotten form the nbastatR package. The nbastatR is a package for professional basketball data in R. The package is an open source package and was created and is being maintained by Alex Bresler the R evangelist . According to the nbastatR website, https://rdocumentation.org/packages/nbastatR/versions/0.1.110202031, NBA Stats API, Basketball Insiders, Basketball-Reference, HoopsHype, RealGM, and nbadraft.net are some of the data sources in the package. A description of each of the data set variables used for analysis is provided in Table 1. The dataset was loaded into R by installing the nbastatR package along with packages that will aid in our analysis using the “install.packages” and the “library” functions. From the nbastatR, the NBA game IDS and gamelog data of the NBA seasons 2010 – 2022 was extracted.

````
```{r Get Game IDs and Gamelogs data}
selectedSeasons <- c(2010:2022)
# Get game IDs for Regular Season and Playoffs
gameIds_RegSea <- suppressWarnings(seasons_schedule(seasons = selectedSeasons, season_types = "Regular Season") %>%
select(idGame, slugMatchup))
gameIds_PlOfs <- suppressWarnings(seasons_schedule(seasons = selectedSeasons, season_types = "Playoffs") %>%
select(idGame, slugMatchup))
gameIds_all <- rbind(gameIds_RegSea, gameIds_PlOfs)
# Peek at the game IDs
head(gameIds_all)
tail(gameIds_all)

## Extract game log data for players and teams

# Get player game logs
P_gamelog_regSea <- suppressWarnings(game_logs(seasons = selectedSeasons, league = "NBA", result_types = "player",
season_types = "Regular Season"))
P_gamelog_po <- suppressWarnings(game_logs(seasons = selectedSeasons, league = "NBA", result_types = "player",
season_types = "Playoffs"))
P_gamelog_all <- rbind(P_gamelog_regSea, P_gamelog_po)
View(head(P_gamelog_all))
View(tail(P_gamelog_all))
# Get team game logs
T_gamelog_regSea <- suppressWarnings(game_logs(seasons = selectedSeasons, league = "NBA", result_types = "team",
season_types = "Regular Season"))
T_gamelog_po <- suppressWarnings(game_logs(seasons = selectedSeasons, league = "NBA", result_types = "team",
season_types = "Playoffs"))
T_gamelog_all <- rbind(T_gamelog_regSea, T_gamelog_po)
view(head(T_gamelog_all))
View(tail(T_gamelog_all))
````

A view at the Player and Team game logs using the “view” function.
````

Player gamelog below

	yearSeason	slugSeason	slugLeague	typeSeason	dateGame	idGame	numberGameTeamSeason	nameTeam	idTeam
1	2010	2009-10	NBA	Regular Season	2009-10-27	20900002	1	Washington Wizards	1610612764
2	2010	2009-10	NBA	Regular Season	2009-10-27	20900003	1	Portland Trail Blazers	1610612757
3	2010	2009-10	NBA	Regular Season	2009-10-27	20900004	1	Los Angeles Lakers	1610612747
4	2010	2009-10	NBA	Regular Season	2009-10-27	20900004	1	Los Angeles Clippers	1610612746
5	2010	2009-10	NBA	Regular Season	2009-10-27	20900003	1	Portland Trail Blazers	1610612757
6	2010	2009-10	NBA	Regular Season	2009-10-27	20900003	1	Portland Trail Blazers	1610612757

Team gamelog below

	yearSeason	slugSeason	slugLeague	typeSeason	dateGame	idGame	numberGameTeamSeason	nameTeam	idTeam
1	2022	2021-22	NBA	Playoffs	2022-06-10	42100404	22	Boston Celtics	1610612738
2	2022	2021-22	NBA	Playoffs	2022-06-10	42100404	20	Golden State Warriors	1610612744
3	2022	2021-22	NBA	Playoffs	2022-06-13	42100405	23	Boston Celtics	1610612738
4	2022	2021-22	NBA	Playoffs	2022-06-13	42100405	21	Golden State Warriors	1610612744
5	2022	2021-22	NBA	Playoffs	2022-06-16	42100406	24	Boston Celtics	1610612738
6	2022	2021-22	NBA	Playoffs	2022-06-16	42100406	22	Golden State Warriors	1610612744

Variable Descriptions

The table below describes the variables used in the analysis of NBA data.

Column Name	Description	Mode	N/As
yearSeason	NBA year season	Numeric	N
slugSeason	NBA full season	Character	N
slugLeague	Professional League	Character	N
typeSeason	Type of season	Character	N
dateGame	Date of Game	Date	N
idGame	Game Unique ID	Numeric	N
numberGameTeamSeason	Number of games per season	Integer	N

nameTeam	NBA team Name	Character	N
idTeam	NBA team ID	Numeric	N
Player	Player Name	Character	N
MIN	Minutes Played	Numeric	N
PTS	Points Made per game	Numeric	N
W	Won games	Boolean	N
L	Lost games	Boolean	N
P2M	Two points Field goals made	Numeric	N
P2A	Two points field goal attempted	Numeric	N
P2p	Two points field goal %	Numeric	N
P3M	Three points Field goals made	Numeric	N
P3A	Three points Field goals attempted	Numeric	N
P3p	Three points Field goals %	Numeric	N
FTM	Free throws made	Numeric	N
FTA	Free throws attempted	Numeric	N

FTp	Free throws %	Numeric	N
OREB	Offensive Rebound(s) per game	Numeric	N
DREB	Defensive Rebound(s) per game	Numeric	N
AST	Assist(s) per game	Numeric	N
TOV	Turnover(s) per game	Numeric	N
STL	Steal(s) per game	Numeric	N
BLK	Block(s) per game	Numeric	N
PF	Personal foul(s)	Numeric	N
plusminusTeam	Plus/minus each team	Numeric	N

Data Wrangling

Data wrangling, the act of acquiring, choosing, and converting data, is one of the most critical and tedious components of data analysis. Unfortunately, if you're working with raw data, data wrangling will almost certainly take up more than half of your time on a data analysis project. Fortunately, *nbastatR* imports fairly clean data for us! In most circumstances, we can skip the data cleansing step and go right to the interesting things, such as data visualization and modeling. However, there are numerous scenarios in which we still need to alter and filter data in order to gain better insights and make better metrics. In this section we are going to be using the logs and IDS extracted and we are going to be creating box scores for the NBA seasons and converting the box scores into data frame with the “as.data.frame()” function.

Teambox scores: The instances (rows) in this data frame, termed Tbox, are the examined teams, and the variables (columns) are the team achievements in the considered games.

Opponentbox scores: The instances (rows) in this data frame, known as Obox, are the examined teams, and the variables (columns) are the achievements of each team's opponents in the games under consideration.

Playerbox scores: The instances (rows) in this data frame, known as Pbox, are the examined players, and the variables (columns) are the individual achievements in the games under consideration.

For each of the boxscores, I selected and grouped the table by the variables in focus “Season” and “Team”.

Source Visual

````
### Create player and team box scores
########
# Create Tbox (Team box score) per season
Tbox_all <- T_gamelog_all %>%
  group_by("Season"=yearSeason, "Team"=slugTeam) %>%
  dplyr::summarise(GP=n(), MIN=sum(round(minutesTeam/5)),
                   PTS=sum(ptsTeam),
                   W=sum(outcomeGame=="W"), L=sum(outcomeGame=="L"),
                   P2M=sum(fg2mTeam), P2A=sum(fg2aTeam), P2p=P2M/P2A,
                   P3M=sum(fg3mTeam), P3A=sum(fg3aTeam), P3p=P3M/P3A,
                   FTM=sum(ftmTeam), FTA=sum(ftaTeam), FTp=FTM/FTA,
                   OREB=sum(orebTeam), DREB=sum(drebTeam), AST=sum(astTeam),
                   TOV=sum(tovTeam), STL=sum(stlTeam), BLK=sum(blkTeam),
                   PF=sum(pfTeam), PM=sum(plusminusTeam)) %>%
  as.data.frame()
# Create Obox (Opponent Team box score) per season
Obox_all <- T_gamelog_all %>%
  group_by("Season"=yearSeason, "Team"=slugOpponent) %>%
  dplyr::summarise(GP=n(), MIN=sum(round(minutesTeam/5)),
                   PTS=sum(ptsTeam),
                   W=sum(outcomeGame=="L"), L=sum(outcomeGame=="W"),
                   P2M=sum(fg2mTeam), P2A=sum(fg2aTeam), P2p=P2M/P2A,
                   P3M=sum(fg3mTeam), P3A=sum(fg3aTeam), P3p=P3M/P3A,
                   FTM=sum(ftmTeam), FTA=sum(ftaTeam), FTp=FTM/FTA,
                   OREB=sum(orebTeam), DREB=sum(drebTeam), AST=sum(astTeam),
                   TOV=sum(tovTeam), STL=sum(stlTeam), BLK=sum(blkTeam),
                   PF=sum(pfTeam), PM=sum(plusminusTeam)) %>%
  as.data.frame()
````

````
# Create Pbox (Player box score) per season
Pbox_all <- P_gamelog_all %>%
  group_by("Season"=yearSeason, "Team"=slugTeam, "Player"=namePlayer) %>%
  dplyr::summarise(GP=n(), MIN=sum(minutes), PTS=sum(pts),
                   P2M=sum(fg2m), P2A=sum(fg2a), P2p=100*P2M/P2A,
                   P3M=sum(fg3m), P3A=sum(fg3a), P3p=100*P3M/P3A,
                   FTM=sum(ftm), FTA=sum(fta), FTp=100*FTM/FTA,
                   OREB=sum(oreb), DREB=sum(dreb), AST=sum(ast),
                   TOV=sum(tov), STL=sum(stl), BLK=sum(blk),
                   PF=sum(pf)) %>%
  as.data.frame()


```
````

Let’s take a look at the boxscores of the NBA champion of the 2021 – 2022 season Golden State Warriors and also the star player Stephen Curry through the “view” function.
````

	Season	Team	GP	MIN	PTS	W	L	P2M	P2A	P2p	P3M	P3A	P3p	FTM	FTA	FTp	OR
9	2010	GSW	82	3946	8922	26	56	2696	5407	0.4986129	633	1687	0.3752223	1631	2085	0.7822542	
39	2011	GSW	82	3966	8477	36	46	2566	5298	0.4843337	685	1749	0.3916524	1290	1695	0.7610619	
69	2012	GSW	66	3183	6453	23	43	1965	4092	0.4802053	524	1351	0.3878608	951	1235	0.7700405	
100	2013	GSW	94	4552	9528	53	41	2836	5953	0.4763985	762	1900	0.4010526	1570	2005	0.7830424	
130	2014	GSW	89	4302	9294	54	35	2663	5366	0.4962728	841	2221	0.3786583	1445	1926	0.7502596	
160	2015	GSW	103	4969	11185	83	20	3089	6048	0.5107474	1123	2858	0.3929321	1638	2174	0.7534499	
190	2016	GSW	106	5133	12006	88	18	3040	5844	0.5201916	1383	3370	0.4103858	1777	2349	0.7564921	
220	2017	GSW	99	4772	11531	83	16	3064	5498	0.5572936	1198	3121	0.3838513	1809	2282	0.7927257	
250	2018	GSW	103	4959	11623	74	29	3214	5757	0.5582769	1161	3028	0.3834214	1712	2096	0.8167939	
280	2019	GSW	104	5027	12161	71	33	3161	5712	0.5533964	1355	3544	0.3823363	1774	2207	0.8038061	
310	2020	GSW	65	3145	6912	15	50	1832	3698	0.4954029	678	2032	0.3336614	1214	1511	0.8034414	
340	2021	GSW	72	3461	8187	39	33	1925	3558	0.5410343	1048	2789	0.3757619	1193	1520	0.7848684	
370	2022	GSW	104	5002	11563	69	35	2749	4930	0.5576065	1484	4052	0.3662389	1613	2100	0.7680952	

	Season	Team	GP	MIN	PTS	W	L	P2M	P2A	P2p	P3M	P3A	P3p	FTM	FTA	FTp	OR
9	2010	GSW	82	3946	9217	26	56	2842	5519	0.5149484	567	1514	0.3745046	1832	2390	0.7665272	
39	2011	GSW	82	3966	8668	36	46	2606	5202	0.5009612	567	1586	0.3575032	1755	2262	0.7758621	
69	2012	GSW	66	3183	6678	23	43	1993	4150	0.4802410	468	1281	0.3653396	1288	1725	0.7466667	
100	2013	GSW	94	4552	9453	53	41	2754	5814	0.4736842	747	2165	0.3450346	1704	2286	0.7454068	
130	2014	GSW	89	4302	8931	54	35	2654	5656	0.4692362	636	1840	0.3456522	1715	2247	0.7632399	
160	2015	GSW	103	4969	10193	83	20	3027	6608	0.4580811	735	2227	0.3300404	1934	2559	0.7557640	
190	2016	GSW	106	5133	11019	88	18	3222	6839	0.4711215	860	2577	0.3337214	1995	2661	0.7497182	
220	2017	GSW	99	4772	10347	83	16	2921	6040	0.4836093	911	2787	0.3268748	1772	2325	0.7621505	
250	2018	GSW	103	4959	10923	74	29	3006	6141	0.4894968	1067	3075	0.3469919	1710	2249	0.7603379	
280	2019	GSW	104	5027	11557	71	33	2916	5814	0.5015480	1243	3551	0.3500422	1996	2556	0.7809077	
310	2020	GSW	65	3145	7478	15	50	1864	3499	0.5327236	879	2261	0.3887660	1113	1422	0.7827004	
340	2021	GSW	72	3461	8111	39	33	2006	3927	0.5108225	891	2481	0.3591294	1426	1834	0.7775354	
370	2022	GSW	104	5002	10998	69	35	2625	5199	0.5049048	1312	3779	0.3471818	1812	2387	0.7591119	

	Season	Team	Player	GP	MIN	PTS	P2M	P2A	P2p	P3M	P3A	P3p	FTM	FTA	FTp	OR
151	2010	GSW	Stephen Curry	80	2895	1399	362	763	47.44430	166	380	43.68421	177	200	88.50000	
669	2011	GSW	Stephen Curry	74	2491	1373	354	711	49.78903	151	342	44.15205	212	227	93.39207	
1205	2012	GSW	Stephen Curry	26	728	383	90	175	51.42857	55	121	45.45455	38	47	80.85106	
1748	2013	GSW	Stephen Curry	90	3484	2067	414	917	45.14722	314	706	44.47592	297	329	90.27356	
2273	2014	GSW	Stephen Curry	85	3139	2034	420	827	50.78597	283	672	42.11310	345	390	88.46154	
2825	2015	GSW	Stephen Curry	101	3441	2494	469	902	51.99557	384	878	43.73576	404	452	89.38053	
3388	2016	GSW	Stephen Curry	97	3320	2827	471	852	55.28169	482	1084	44.46494	439	483	90.89027	
3936	2017	GSW	Stephen Curry	96	3238	2476	430	794	54.15617	396	961	41.20708	428	476	89.91597	
4493	2018	GSW	Stephen Curry	66	2184	1729	289	505	57.22772	276	663	41.62896	323	349	92.55014	
5096	2019	GSW	Stephen Curry	91	3173	2501	376	717	52.44073	446	1054	42.31499	411	444	92.56757	
5716	2020	GSW	Stephen Curry	5	139	104	21	33	63.63636	12	49	24.48980	26	26	100.00000	
6320	2021	GSW	Stephen Curry	63	2152	2015	321	564	56.91489	337	801	42.07241	362	395	91.64557	
6972	2022	GSW	Stephen Curry	86	2976	2232	361	685	52.70073	376	979	38.40654	382	427	89.46136	

I proceeded to modify the code to select the regular season from the logs extracted.

Purpose

Data Science applied to sports data is increasing in popularity as coaches, players, scouts, sport management, and sports fans understand its worth as a decision support tool and a quantitative approach. In this research, I want to learn about the players, playing patterns, and factors that impact NBA teams performance.

Analysis

Bar Plots

"A barplot is used to depict the connection between a numeric and a categorical variable," according to the R Graph Gallery. What we want to accomplish is to show the association between the points and shot types (numerical variables) of the Golden State Warriors players throughout the 2022 season (categorical variables).

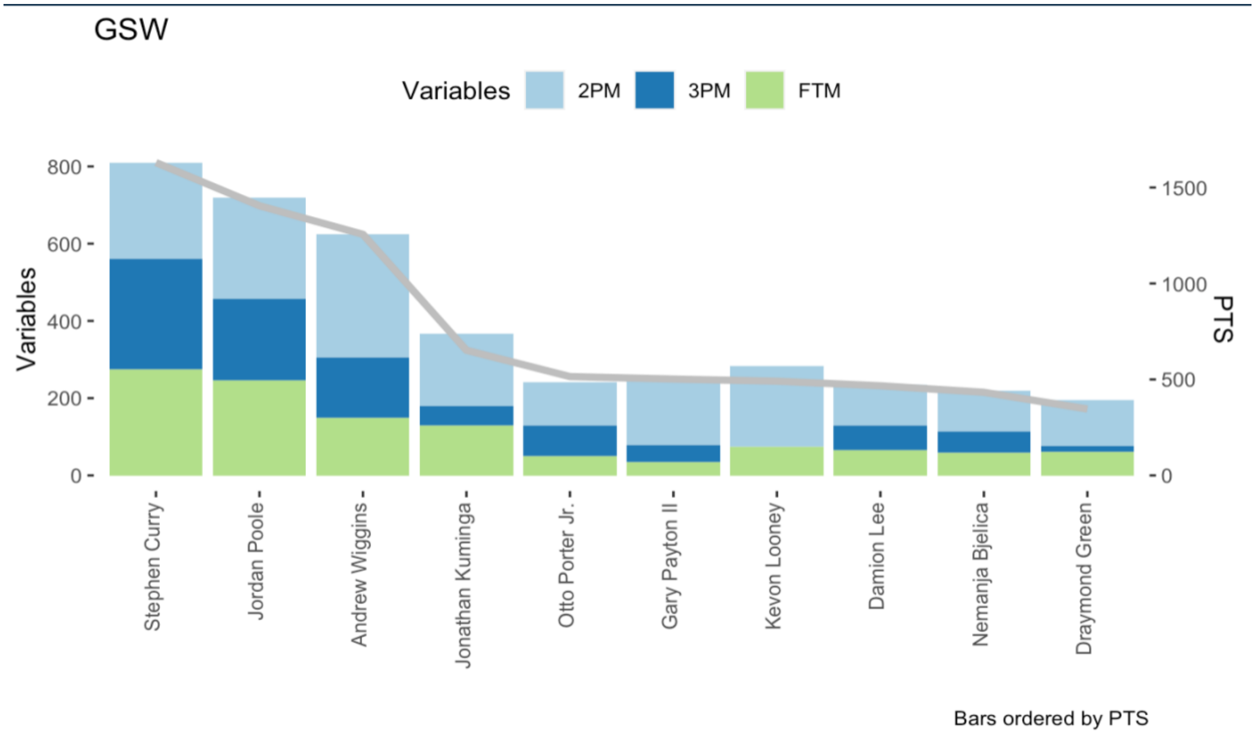

Figure 1 Barline plots of the NBA champion Golden State Warriors

This is a visual representation of the playerbox scores for the Golden State Warriors players with over 1000 minutes played in a season. From this graph we see why Stephen Curry is the most important player on the team. He had the highest three points field goals made, highest free throws made and the third highest two points field goals made with over 1500 points in the 2022 season. Andrew Wiggins made the highest two points shots with an above average three points shots made.

Scatter Plots

"A Scatterplot depicts the relationship between two numeric variables," according to the R Graph Gallery. Each dot indicates a different observation. The values of the two variables are represented by their location on the X (horizontal) and Y (vertical) axes." What we want to accomplish is to show the correlation between assists and turnovers (our numeric variables). I've introduced a new number variable, the points, which are represented by a color. Each dot symbolizes a Golden State Warriors player from the 2010 - 2022 season (our categorical variable) using the “scatterplot” function.

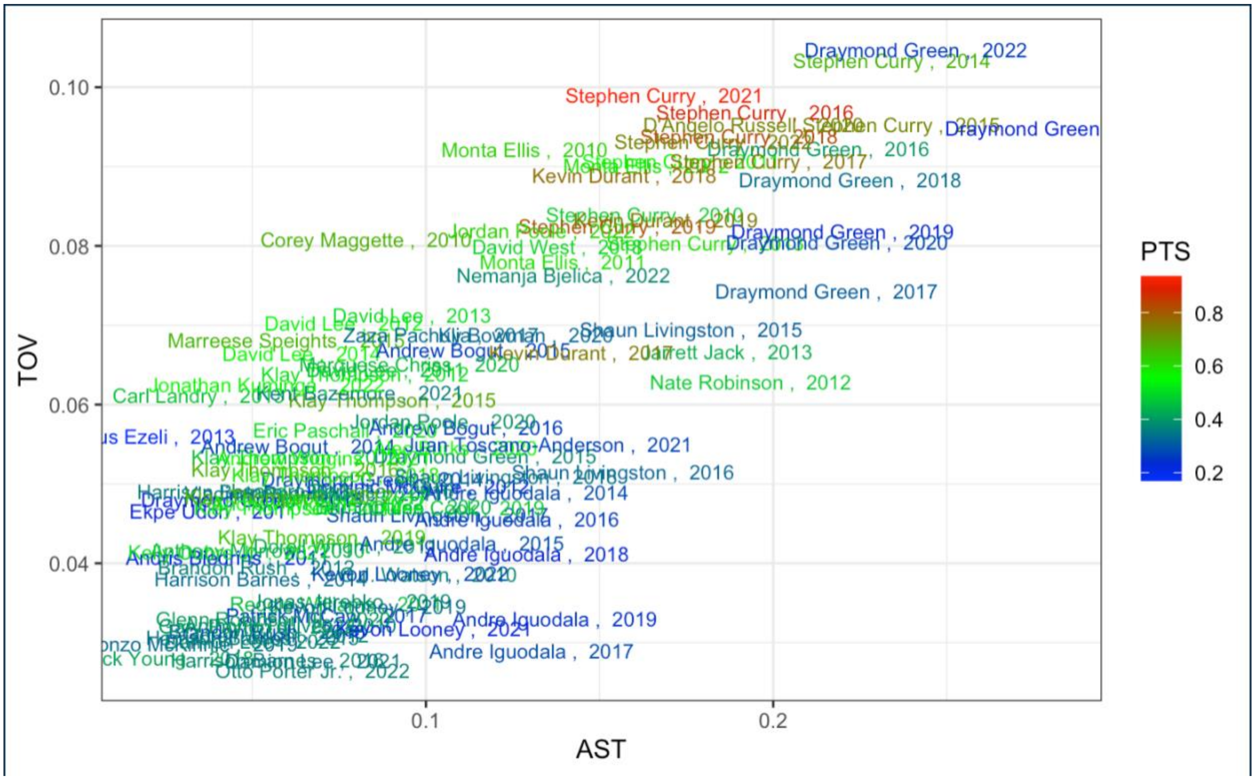

Figure 2.0 Scatterplots of assists vs turnovers per Min with Points color coding of “GSW”

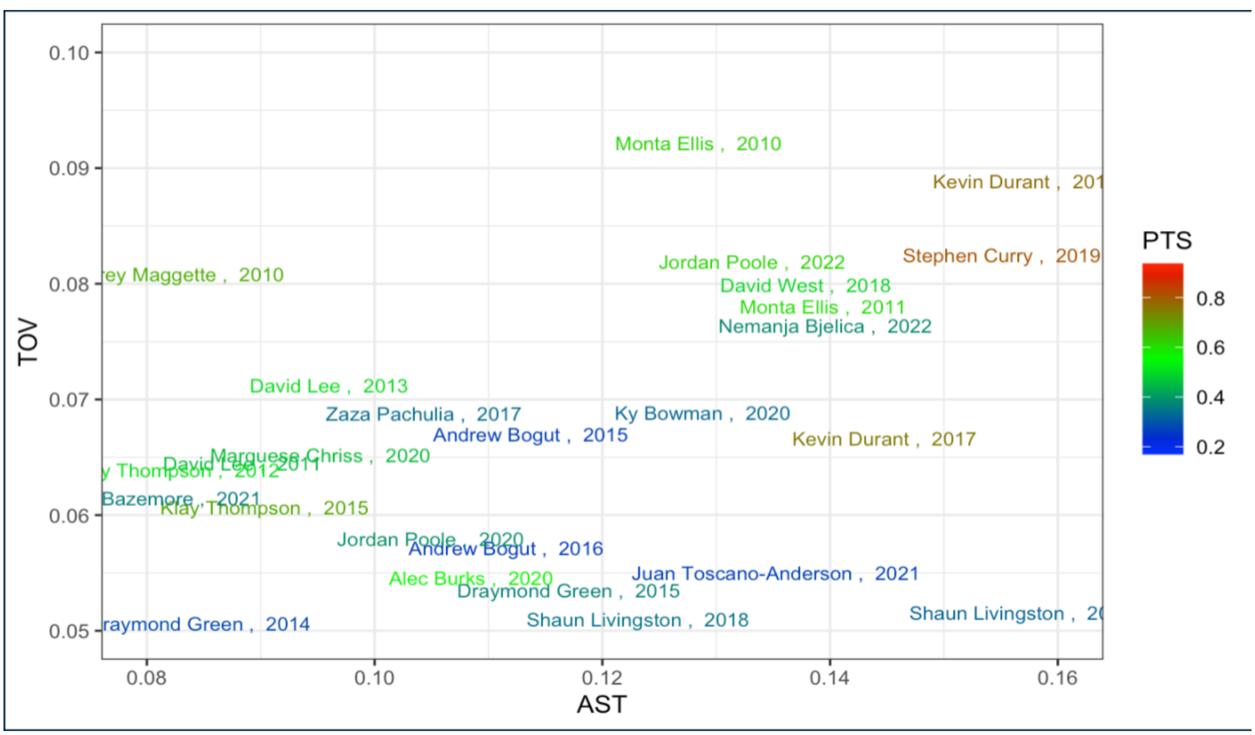

Figure 2.1 Scatterplots of selected players assists vs turnovers per Min with Points color coding of “GSW”

This is a visual representation of the playerbox scores for the Golden State Warriors players with over 1000 minutes played in a season. With the color coding of Points, we can easily notice the players with the highest and lowest Turnovers and Assists per season. In the graph we see that Stephen Curry has been the most important player for the Golden State Warriors as he has brought the most points for 7 seasons with 2021 - 2022 being his best as he led his team to win the NBA title. One notable player is “Kevin Durant”, during his spell at Golden Sate Warrior, he had an above average turnovers and assists which contributed to the team winning the NBA title for the 2017 and 2018 seasons.

Cluster Analysis

Cluster analysis, often known as clustering, is the process of classifying a collection of things (data instances) together so that items in the same cluster are similar while objects in other clusters are unlike. The labels "similar" and "dissimilar" are determined by the variables in the dataset and the application domain. The distance between two related attributes can be used to determine how similar or different the two objects are.

K-means clustering of NBA teams

Next, I wanted to see the similarities between the 2022 NBA teams. I looked at the free throw rates, the turnover ratio, the rebound % and the effective field goal %.

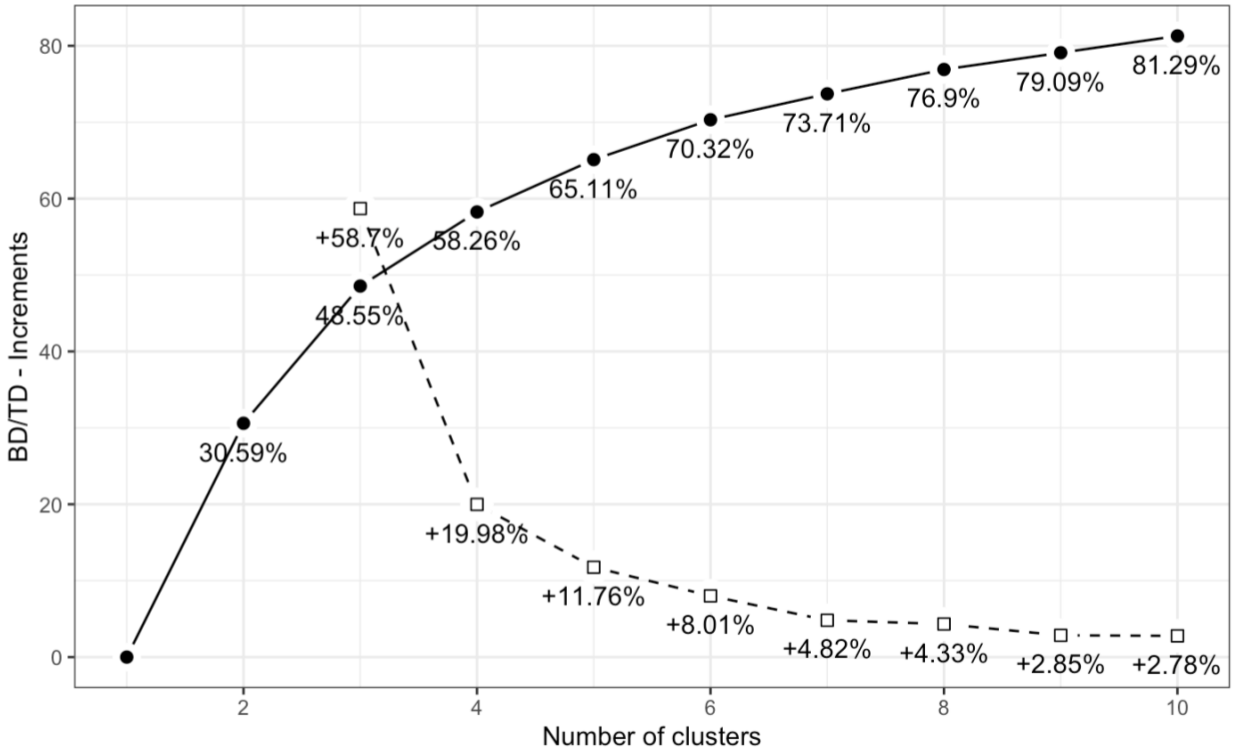

Figure 3 Graph showing the Clusterization quality pattern as a function of cluster number

The solid line depicts the (average across all variables) ratio of the Between to Total Deviance BD/TD, which improves as the number of clusters grows. What I want to achieve is to reduce the number of clusters while maintaining the highest level of consistency and information. The graph suggests that 7 clusters are the best option as values above 50 percent is considered satisfactory.

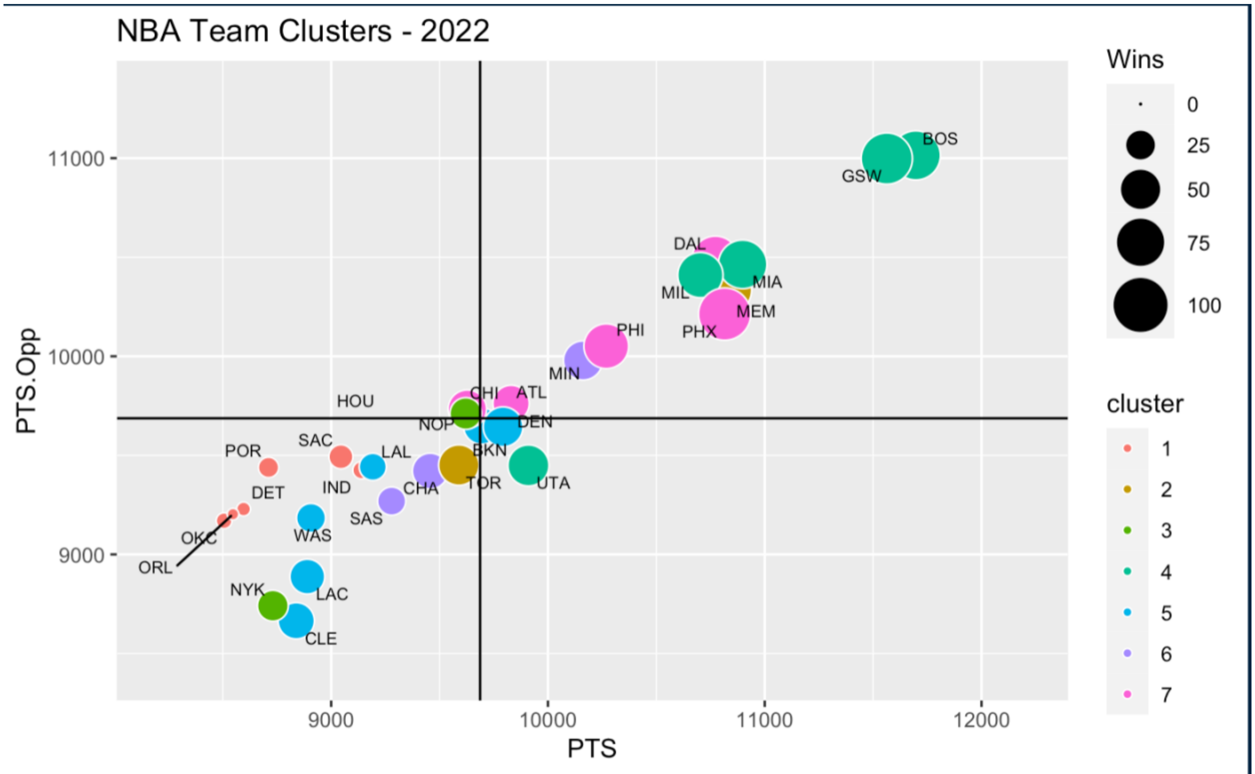

Figure 4 Bubble plots of the 2022 NBA teams

The following bubble plot represents the 2022 NBA teams, with the x-axis representing points scored and the y-axis representing points against. The colors represent the cluster which a team is placed, and the size of the bubble represents the number of victories.

From the above graph we see that cluster 4 contains 3(“GSW”,”BOS”,”MIA”) out of the 4 finalist of each conferences (western and eastern). This can be explained through the radial plot below. We see that they have a high ratio of P3M.ff three pointers made compared to the other clusters. Cluster 4 has the second highest F3.Def Rebound percentage compared to other clusters. This shows how these factors are important in taking a team to the final and wining an NBA title. Also based on the clusters, each team can improve the factor ratios which are below average to help the team become more successful.

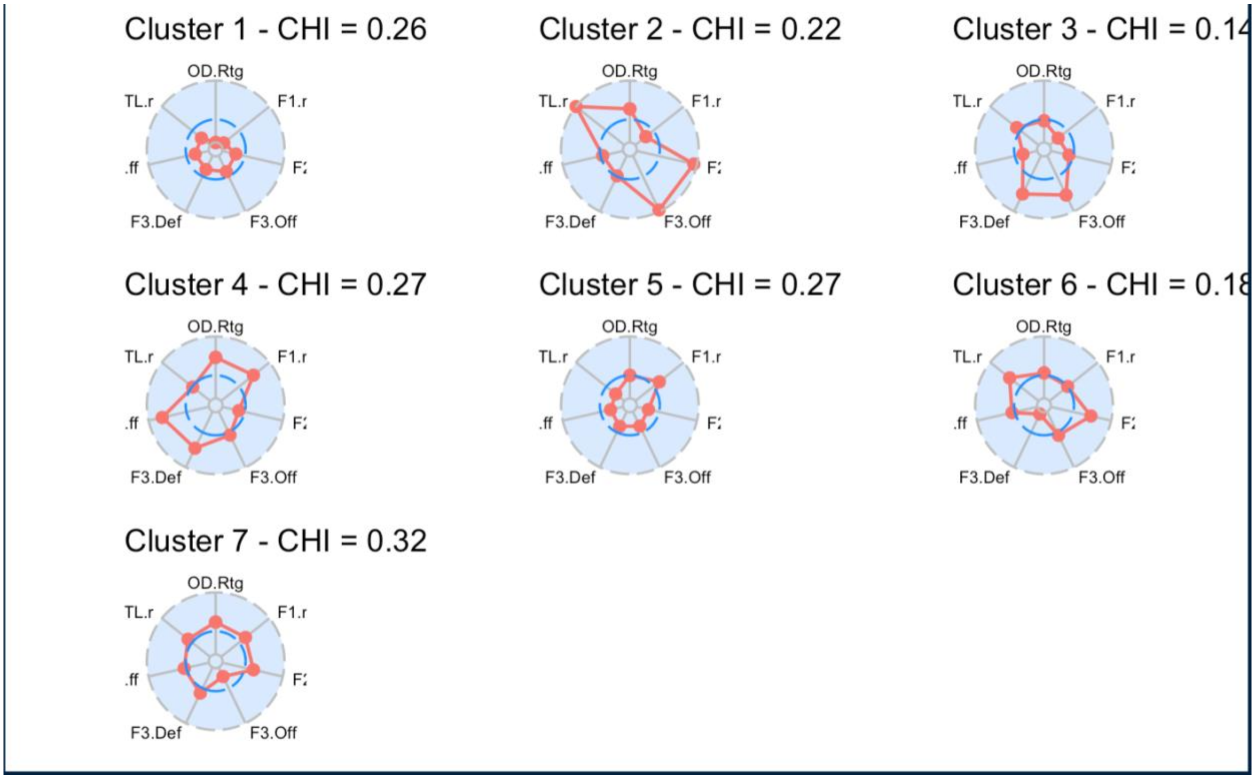

Figure 5 Radial graphs of (average profiles) NBA teams cluster

Hierarchical Clustering of NBA players in 2020, 2021, and 2022

Hierarchical clustering is an alternate approach to partitional clustering for grouping things based on their similarity. Unlike the k-means clustering technique, hierarchical clustering does not need a pre-specified number of groups.

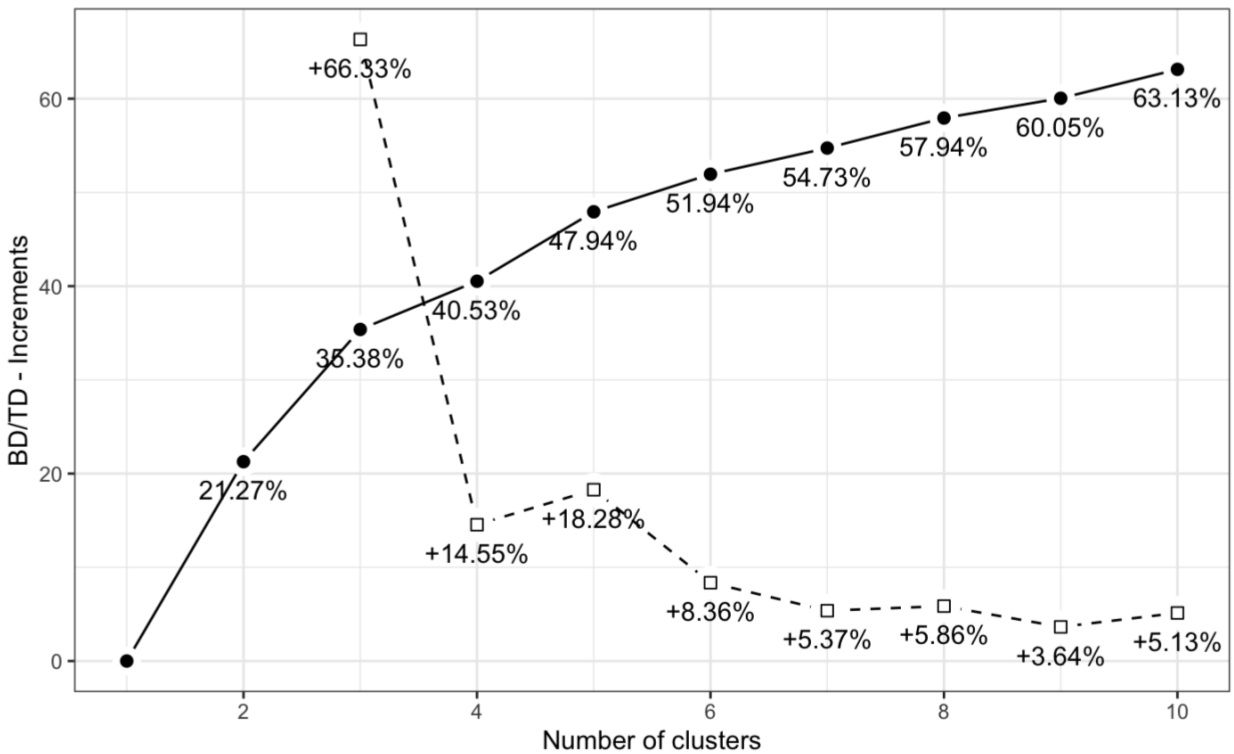

Figure 6 Graph showing the Clusterization quality pattern as a function of cluster number

The graph suggests that 5 clusters are the best option as values above 50 percent is considered satisfactory.

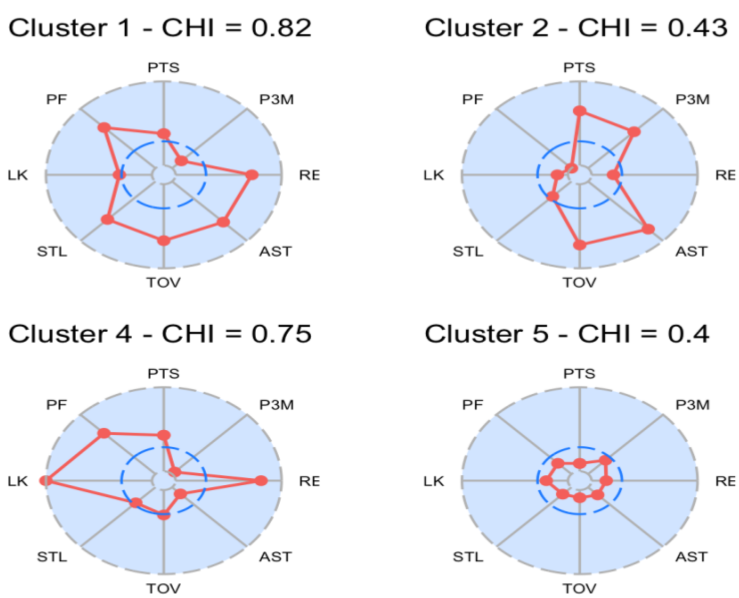

Figure 7 Radial plots of NBA teams average profiles cluster

The graph above depicts the radial plots of the average cluster profiles. It hints at what the clusters signify. Cluster 1 has a high Cluster Heterogeneity Index & features players with a high number of points scored but average stats in all other categories. Cluster 4 has a high percentage of Blocks(BLK) and Rebound(REB) per game. Outliers in the Blocks % have been identified in Cluster 4. Cluster 5 has below the average performance in all categories for the 3 seasons under analysis.

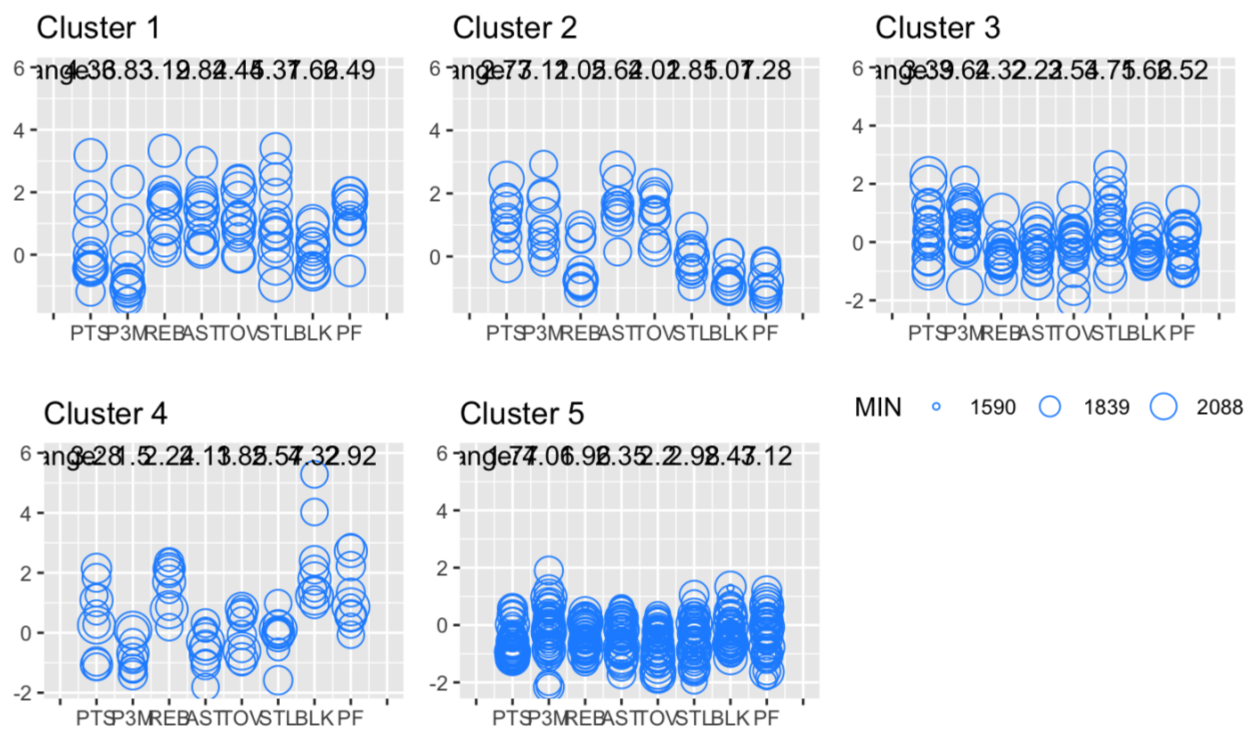

Figure 8 Variability graphs within clusters.

The above variability diagrams reflect all of the players' season performance in each cluster. We can see that there are several obvious outliers in terms of points scored in Cluster 4. Clusters 4 and 1 also have very impressive block stats.

Future research

In future work, it would be interesting to perform a time series analysis to see the repeated measurements of NBA teams over time and to create a regression model to predict the outcome of an NBA game.