Exploratory Data Analysis

Springboard Capstone 3

Patrick Long

4/20/2022

Introduction

Conduct Exploratory Data Analysis (EDA) on the Figshare NBA data set to gather insights through visualizations to answer how best to predict an NBA game. The data was cleaned, merged, and presented in a final data frame that had both Player data and Team data from the years 2009-2015.

Approach

-Find Patters in the data

-Determine relationships in the data

-Checking assumptions

-Drawing Conclusions

Which Questions are we trying to solve?

1. How to best predict, before a game even starts which team will win.
2. Find the data in both the player data sets and team data sets that help best predict how a team will do.
3. Insights in the top how the top 25 teams in the League (based on W/L over the 2009-2015 timeline) performed.
4. Insights that will assist in recommendations later, by understanding the NBA data that helps teams win.

Data Cleaning and Wrangling

The data was initially gathered from Figshare's NBA stats. Initial data set gathered was the Team data set:

	Season	Lg	Team	W	L	W/L%	Finish	SRS	Pace	Rel_Pace	ORtg	Rel_ORtg	DRtg	Rel_DRtg	Playoffs	Coaches	Top WS
0	2015-16	NBA	Atlanta Hawks	48.0	34.0	0.585	2.0	3.49	97.1	1.3	105.1	-1.3	101.4	-5.0	Lost E. Conf. Semis	M. Budenholzer (48-34)	P. Millsap (10.1)
1	2015-16	NBA	Boston Celtics	48.0	34.0	0.585	2.0	2.84	98.5	2.7	106.8	0.4	103.6	-2.8	Lost E. Conf. 1st Rnd.	B. Stevens (48-34)	I. Thomas (9.7)
2	2015-16	NBA	Brooklyn Nets	21.0	61.0	0.256	4.0	-7.12	95.2	-0.6	103.2	-3.2	110.9	4.5	NaN	L. Hollins (10-27), T. Brown (11-34)	B. Lopez (6.2)
3	2015-16	NBA	Charlotte Hornets	21.0	61.0	0.256	4.0	-7.12	95.2	-0.6	103.2	-3.2	110.9	4.5	NaN	L. Hollins (10-27), T. Brown (11-34)	B. Lopez (6.2)
4	2015-16	NBA	Chicago Bulls	42.0	40.0	0.512	4.0	-1.46	95.7	-0.1	105.0	-1.4	106.5	0.1	NaN	F. Hoiberg (42-40)	J. Butler (9.1)

Initially we needed to clean up some of the NaN values, change the how the Season column was presented, and get rid of some of the redundant data like Lg (League). Below is the cleaned data set. We changed the Playoffs to just mark if the team made the playoffs or not, 1 being yes, 0 being no.

	Season	Team	W	L	W/L%	Finish	SRS	Pace	Rel_Pace	ORtg	Rel_ORtg	DRtg	Rel_DRtg	Playoffs	Coaches	Top WS
0	2015	Atlanta Hawks	48.0	34.0	58.5	2.0	3.49	97.1	1.3	105.1	-1.3	101.4	-5.0	1	M. Budenholzer (48-34)	P. Millsap (10.1)
1	2015	Boston Celtics	48.0	34.0	58.5	2.0	2.84	98.5	2.7	106.8	0.4	103.6	-2.8	1	B. Stevens (48-34)	I. Thomas (9.7)
2	2015	Brooklyn Nets	21.0	61.0	25.6	4.0	-7.12	95.2	-0.6	103.2	-3.2	110.9	4.5	0	L. Hollins (10-27), T. Brown (11-34)	B. Lopez (6.2)
3	2015	Charlotte Hornets	21.0	61.0	25.6	4.0	-7.12	95.2	-0.6	103.2	-3.2	110.9	4.5	0	L. Hollins (10-27), T. Brown (11-34)	B. Lopez (6.2)
4	2015	Chicago Bulls	42.0	40.0	51.2	4.0	-1.46	95.7	-0.1	105.0	-1.4	106.5	0.1	0	F. Hoiberg (42-40)	J. Butler (9.1)

Next I gathered the Player data set:

	Player	Season	Age	Tm	Lg	G	GS	MP	PER	3PAr	...	ORtg	DRtg	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP
0	Thanasis Antetokounmpo	2015-16	23.0	NYK	NBA	2	0.0	6.0	32.9	0.250	...	140.0	110.0	0.0	0.0	0.0	0.291	6.3	-10.5	-4.1	0.0
1	Rakeem Christmas	2015-16	24.0	IND	NBA	1	0.0	6.0	32.0	0.000	...	204.0	112.0	0.0	0.0	0.0	0.343	9.7	-6.6	3.2	0.0
2	Stephen Curry	2015-16	27.0	GSW	NBA	79	79.0	2700.0	31.5	0.554	...	125.0	103.0	13.8	4.1	17.9	0.318	12.4	0.1	12.5	9.8
3	Kevin Durant	2015-16	27.0	OKC	NBA	72	72.0	2578.0	28.2	0.348	...	122.0	104.0	11.0	3.5	14.5	0.270	7.0	0.9	7.9	6.4
4	Boban Marjanovic	2015-16	27.0	SAS	NBA	54	4.0	508.0	27.7	0.000	...	130.0	96.0	2.3	1.2	3.4	0.325	2.7	0.9	3.6	0.7

Again, I changed the Season column to be more presentable and easier to work with, cleaned any NaN values. I then merged the two data sets into one, on the one column Season.

Featured attributes:

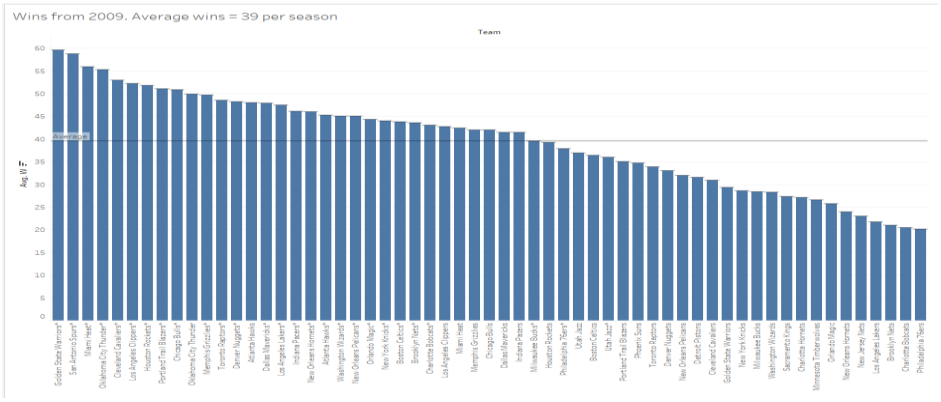
Minutes played - 3 Point shot attempts - 3 Point shot attempts – Pace - Offensive and Defensive Ratings - Steals/Assists/Turnovers - VORP (Value over Replacement Player)

Data Stories and Visualizations via Tableau

In this EDA section we will try to highlight the approach that will be taken to find the stats in both player and team that lead to NBA teams winning.

This will give us insight into which stats will predict games. For this section of the project we used Tableau to create our visualizations and analytics.

The first thing we need to look at is overall wins. This will gives us a picture of the top 25 teams, and the average wins, by season.



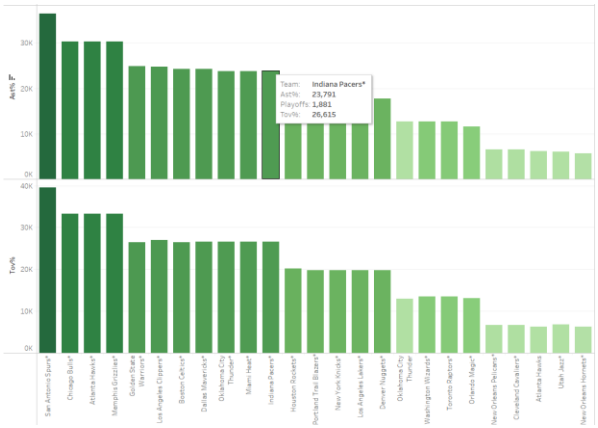
Here we see the league as a whole. Golden State is #1 with nearly 60 average wins per season, where Philadelphia is last with just over 20.

The average win total on average is 39. The top 5 are; Golden State, San Antonio, Miami Heat, Oklahoma city, and Cleveland.

Now that we've seen the team win totals, lets take a look at the players that have the highest win shares.

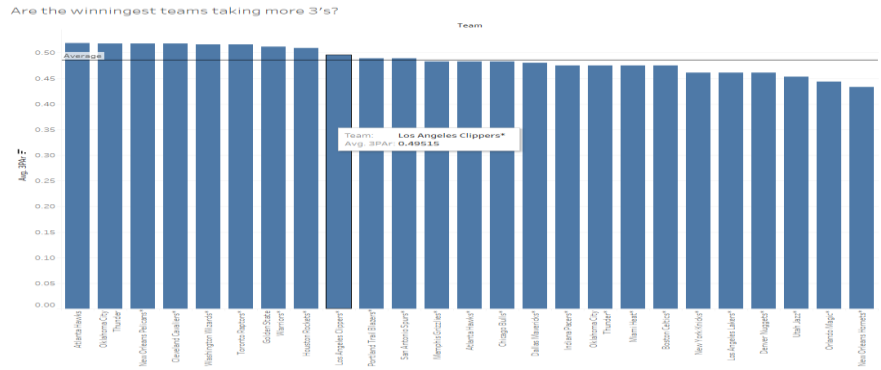
Here we see Lebron James, Kevin Durant, Chris Paul, James Harden, and Stephen Curry make up the top 5 win shares per players. This means that these players are responsible for more wins than any other players.

Let's take a look at the top 25 teams by Assist percentage and turnover percentage. We will shade them with their records of making the playoffs (the darker the bar, the more playoff appearances the team had).



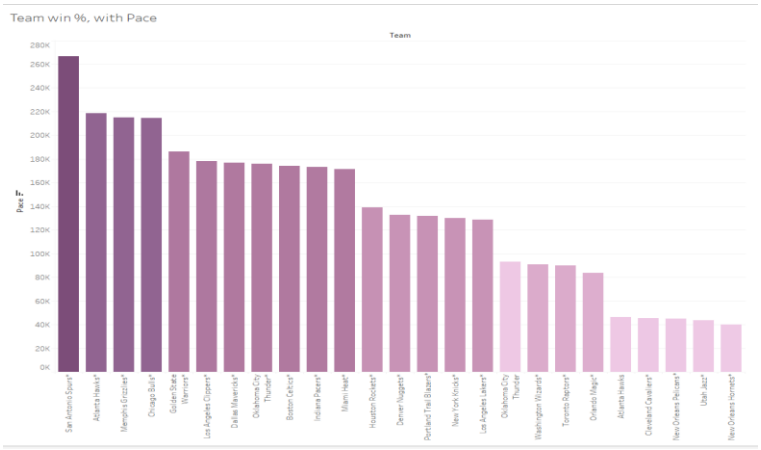
We can see that the teams that are most likely to make the playoffs will have a high assist percentage, and also a high turnover percentage. They are more likely to make the playoffs as well.

Now, we can dive into the team statistics that we think will best show winning teams. The first stat we will look at is 3 point attempts. Do winning teams take more 3's per game?



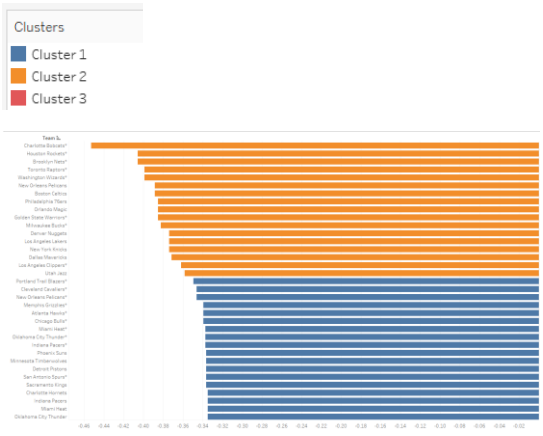
The top 5 teams with the most 3 point shot attempts per game are; Atlanta, Oklahoma City, New Orleans, Washington, and Toronto.

Now, lets look at Pace. That is the number of possessions a team uses per game.



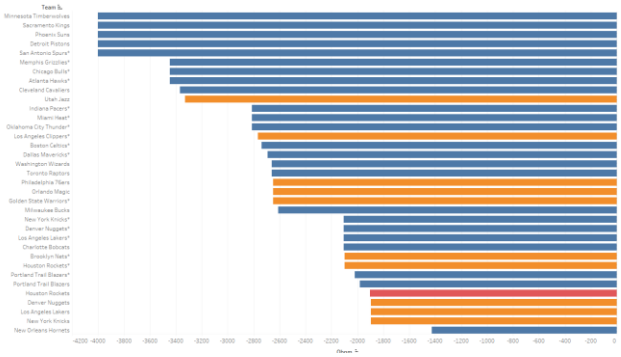
The top 5 teams with the most Pace are San Antonio, Atlanta, Memphis, Chicago, and Golden State. This graph has also been shaded to show, the darker the purple, teams that have made the playoffs most. Which means, more wins. The average Pace is 135. We can see its clearly shown that teams with the highest Pace are winning and going to the playoffs much more than the teams with less Pace.

Now, lets look at Defensive and Offensive ratings of the teams.



Here we can see the top defensive teams are Charlotte, Houston, Brooklyn, Toronto, Washington.

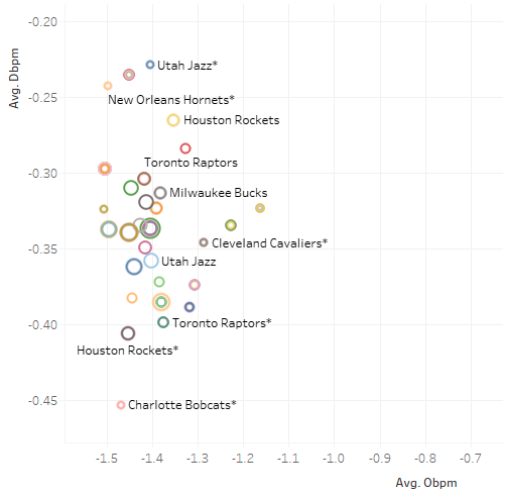
Now, lets look at the best offensive teams



The top offensive teams New Orleans, New York, Denver, Los Angeles Lakers, Charlotte.

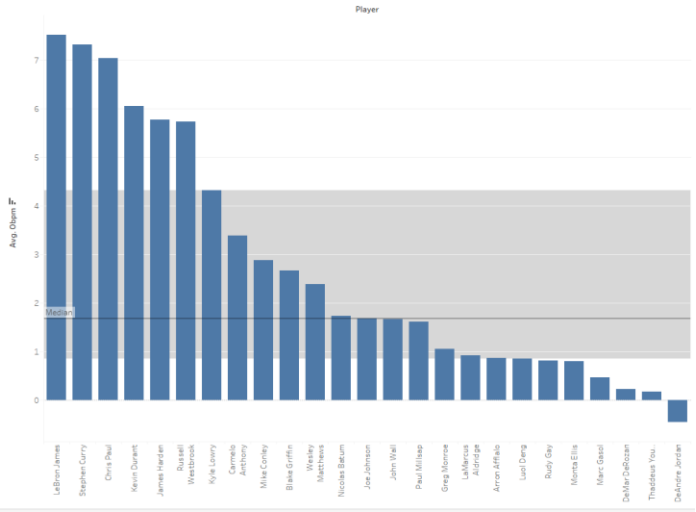
These have been clustered. Cluster 1 is the top offensive and defensive teams, cluster 2 is the middle teams, and cluster 3 are the worst offensive and defensive teams.

Lets see if we can make more sense of this, and see if the combined ratings means more wins.

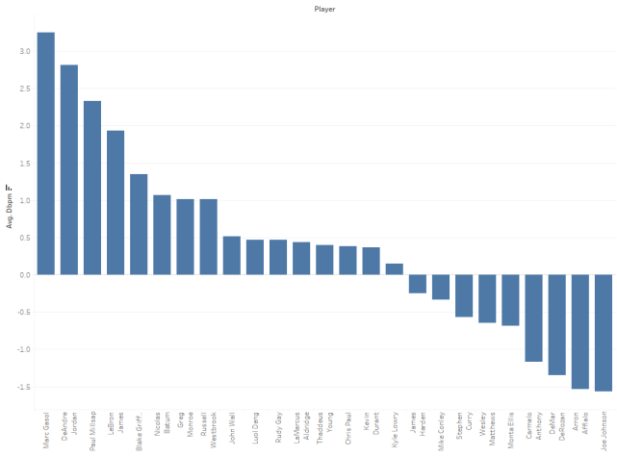


We can see the top teams are Milwaukee, Cleveland, Utah, New Orleans, Raptors, and Rockets. The worst teams look like Charlotte and Houston.

Lets see if offensive players means more wins. Here I have the best offensive players with their total wins per 48 minutes played.



Here we can see that the top offensive players are again the stop 5 players with the most win shares. Lets see if Defensive players win games.



This is a much different list than the top win share players. The top defensive players are Marc Gasol, DeAndre Jordan (who we saw had a negative offensive rating in the last viz.) Paul Millsap, Lebron James (the only repeat) and Blake Griffin. We can also see that the defensive metrics go negative after 16 players.

Summary and Findings

This EDA has given us some great insights into what makes a team and a player 'winning'. We can see that Players who have a higher VORP are more likely to be apart of winning teams, meaning that they are going to have a higher wins per 48 minutes played.

The average wins per team per year is 39. The top 5 teams, by wins, are Golden State, San Antonio, Miami Heat, Oklahoma City, and Cleveland.

When looking at defensive metrics, we can see that there isn't much correlation between them and winning. The top defensive team that is also part of the top 5 teams is Golden State. The top offensive teams have no correlation between the top 5 winning teams. That begs the question, very few top teams aren't on the top offensive or defensive list, there must be something else that provides them wins.

Just because offensive or defensive teams don't necessarily mean wins, we can see that offensive players do. The top offensive players are the top VORP players, which have the most wins per 48 minutes played. Those players are also on the teams with the most wins. Lebron James = Cleveland, Stephen Curry = Golden State, Kevin Durant = Oklahoma, James Harden = Huston.

We can see that being a top Defensive and Offensive team don't necessarily mean wins for an NBA team. We do have a metric that seems to point to it exactly, Pace. The top 5 Pace teams are San Antonio, Atlanta, Memphis, Chicago, and Golden State. That includes 2 of the 5 top teams in terms of wins. The visualization also clearly shows that the higher the pace, the more playoff appearances your team will have.

The biggest take away, for us, from this EDA is the better the players the more wins. The NBA seems to be a very heavy player league. As in, the teams with the best players will win and go to the playoffs. Unless, the team is able to push the pace and even out the uneven player balance.