Seraj Mehrabkhani

Fidan Allahverdiyeva

Data Analytics Final Project

0.0 Introduction

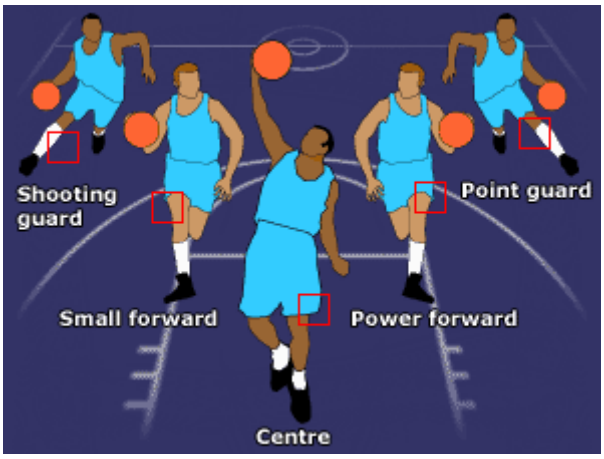
The field of sports analytics has grown in popularity in the recent history as sport organizations have understood that they can use this field to get better insights for their decision-making processes. Since our team has always been interested in the pro basketball and inspired by its players, we decided to focus our project in the world of NBA (National Basketball Association).

0.1 The world of NBA

NBA has been the dominant league of basketball in the US and is widely accepted as the pinnacle of the game of basketball that attracts the best talent from every corner of the world.

0.2 Positions in NBA

In basketball there are positions that each player play:



Point guard (PG): The point guard is typically the team's primary ball-handler and playmaker. They are responsible for bringing the ball up the court and initiating the team's offensive plays. Point guards are usually smaller and quicker than other players on the court, and they need to have good ball handling skills and the ability to pass the ball well.

Shooting guard (SG): The shooting guard is typically the team's best perimeter shooter. They are often responsible for scoring points from the outside, either by shooting three-pointers or mid-

range jumpers. Shooting guards are typically taller than point guards and are generally good athletes.

Small forward (SF): The small forward is a versatile player who can play both on the perimeter (3point line) and in the paint (close to the basket). They are usually taller and more physical than guards, but smaller and quicker than power forwards. Small forwards need to have good shooting skills, ball handling skills, and the ability to rebound.

Power forward (PF): The power forward is typically the team's primary post player. They are usually taller and more physical than other players on the court, and they are responsible for scoring points in the paint and rebounding the basketball. Power forwards need to have good footwork, good hands, and the ability to finish around the basket.

Center (C): The center is typically the team's tallest player and primary rebounder (defender). They are responsible for playing defense in the paint, blocking shots, and scoring points in the paint. Centers need to have good size, strength, and the ability to finish around the basket.

Historically the positions have dictated the way that a team plays. Based on traditions, center players don’t need to be good at shooting, and point guards are not good at dunking (a role reserved only for the biggest of players, especially centers). However, developments in the game in recent years has led to the development of a new type of basketball that blurs the lines between positions and roles and allows the game to be so much more complicated, unpredictable and exciting.

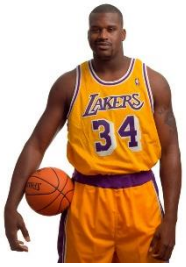
Britney Griner, Center for Phoenix Mercury Dunking

0.3 Position-less Basketball

Position-less basketball is a style of play that emphasizes versatility and skill over traditional positions. In this style of play, players are not limited by their height or traditional positions, and instead are expected to have the ability to play multiple positions and perform a variety of roles

on the court. This allows teams to be more adaptable on defense and offense and creates matchup problems for the opponent, making it harder to defend a specific player. One of the key characteristics of position-less basketball is the emphasis on three point shots regardless of the position and size of the player.

The following is a demonstration of how the game of basketball has changed especially for the big players of the league. Shaquille O'Neal represents the traditional sense of playing basketball while Kevin Durant is a clear representation of how position-less basketball is played.



Shaquille O'Neal

drafted to NBA in 1992

2.16 m

23.7 points per game

1 three pointers in career

Kevin Durant

drafted to NBA in 2007

2.08 m

27.3 points per game

1841 three pointers in career

0.4 The Goal of Our Project

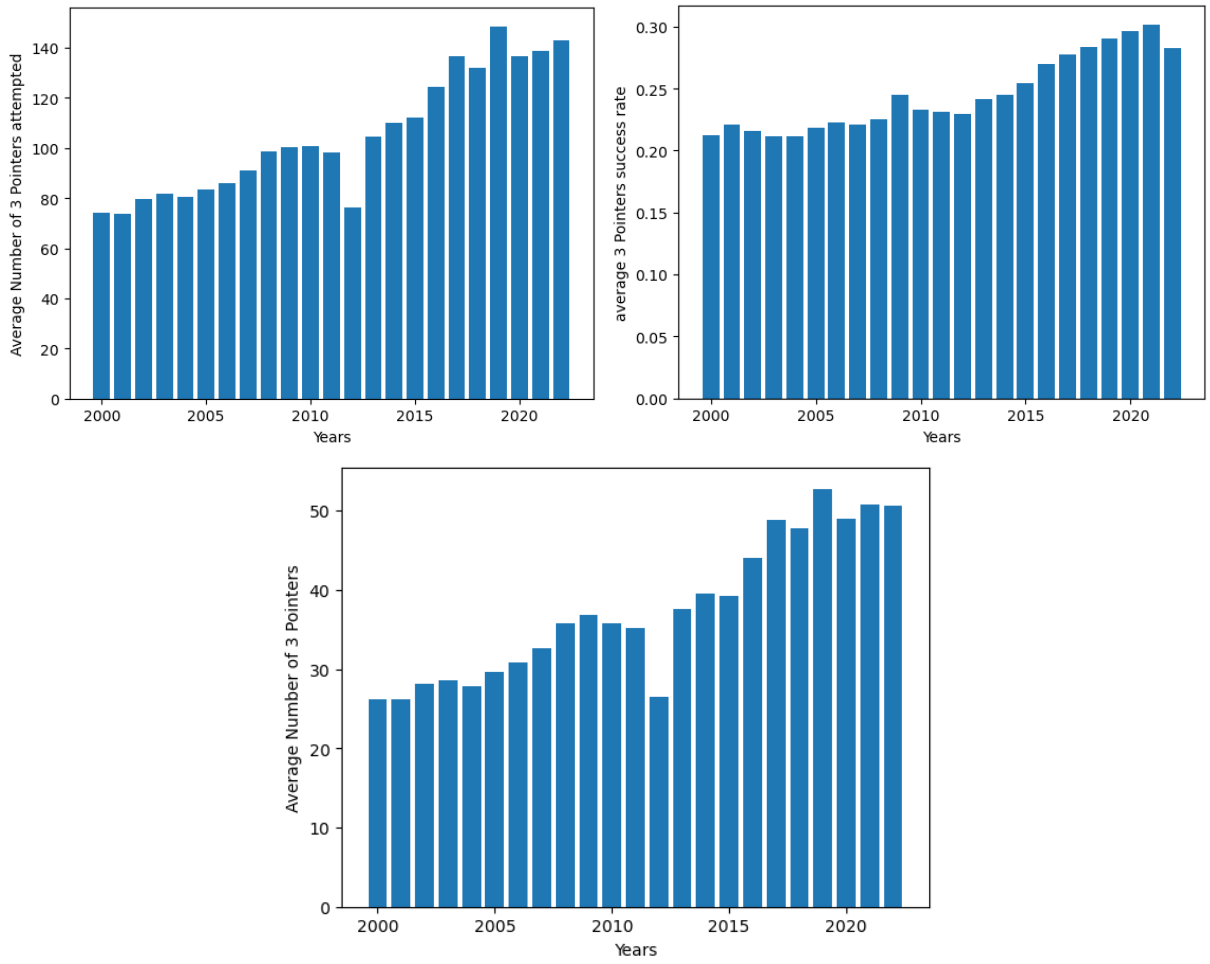
The change that the game of basketball has seen begs the question that can we still predict it. Can we still make models using the traditional methods to understand where wins come from, or the game has gone to complicated that is beyond our current modeling skills.

To answer this question, we decided to focus our effort on predicting one of the most important aspects of the new basketball, Three-point shots. Any shot in the basket that is shot outside the three-point arches is a three-point shot.

Steve Kerr Shooting a Three pointer (1996). Kerr later became one of the pioneers of position-less basketball as the head coach of Golden State Warriors

1.0 Problem 1/Descriptive Analysis

Given our focus on the three-point shots, here we are trying to describe how over time the league's tendency to make these shots have changed.



These three graphs clearly portray the upward trend in the number of three-point shots in the NBA over the last two decades.

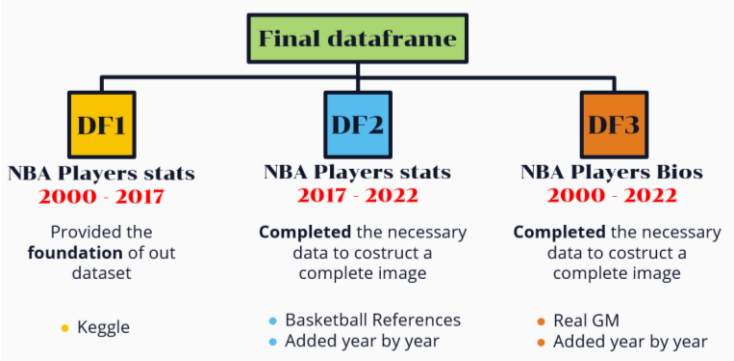
This increase is happening since: **1.** On average NBA players are attempting more three-point shots and **2.** Players are shooting with a higher accuracy over time which gives them a higher success rate. These two factors together have led to an increase overtime in three-point shots scored by each player in a year. From the average of below 30 successful shots a year in 2000 to 50 successful shots a year in 2022.

2.0 Problem 2: predicting the three point shots

Now that we know the popularity of three-point shots are on the rise we would like to see if we can predict a player's ability to make these shots given multiple variables.

2.1 Data and Preprocessing

The data has been collected from multiple sources and had been cleaned and matched in a quite extensive process. Here we were lucky to have a friend more experienced in the world of coding who helped us with the pre-processing. **We must emphasize that our friend's contribution is only limited to the preprocessing, and the main body of the work is done by us.**



The sources of our data and the content of each

There were many issues that we faced during the pre-processing, which we mention the important ones in our presentation. **Our preprocessing codes is uploaded with this report.**

2.2 columns and descriptions

Rk : Rank

Year: Year of the season

Player : Player's name

Pos : Position

Age : Player's age

HT: Height in cm

WT: Weight in kg

YOS: Years of service/experience

Draft Year: The year that the player was recruited in NBA

Final Pick: In which rank was he drafted (the lower the better, 1 the best)

Tm : Team

G : Games played

GS : Games started

MP : Minutes played per game

FG : Field goals per game

FGA : Field goal attempts per game

FG% : Field goal percentage

3P : 3-point field goals per game

3PA : 3-point field goal attempts per game

3P% : 3-point field goal percentage

2P : 2-point field goals per game

2PA : 2-point field goal attempts per game

2P% : 2-point field goal percentage

eFG% : Effective field goal percentage

FT : Free throws per game

FTA : Free throw attempts per game

FT% : Free throw percentage

ORB : Offensive rebounds per game

DRB : Defensive rebounds per game

TRB : Total rebounds per game

AST : Assists per game

STL : Steals per game

BLK : Blocks per game

TOV : Turnovers per game

PF : Personal fouls per game

PTS : Points per game

2.3 Process

Given the continuous nature of most of our parameters we decided to do a linear regression to see the effects of different career aspects on the three pointers scored by a player.

Originally, we decided to have all the 35 columns in our dataframe in our model and then omit the unnecessary ones, but that turned out to give us too many parameters after getting the dummies. So we decided to omit the ones that we are sure are unnecessary from the beginning. Given that we started with: Pos, HT, WT, YOS, Draft Year, Final Pick, G, FT%, TRB, AST, STL, BLK.

To remove Multicollinearity, we omitted STL, TRB, and WT.

We made our train and test sets and trained a regression model. We omitted the HT since its insignificance was quite high. Final Pick was also insignificance at the end but its insignificance dropped below 10% and we decided to keep it since its removal didn't really affect the performance of the model.

2.4 Parameters and interpretation

```
const        -3067.739172
HT               7.105032
```

```
YOS              1.946779
Draft Year       1.500460
Final Pick      -0.037769
G                0.643086
FT%             21.028414
AST              0.139856
BLK             -0.207915
Pos_PF          11.481943
Pos_PG           9.012602
Pos_SF          27.401321
Pos_SG          33.560419
```

A negative constant emphasizes the difficulty of making these types of shots.

Height and Years of Service have a positive effect understandably.

Draft Year has a positive effect which totally make sense given that as we go forward in NBA new players have always been better at making threes in comparison to their predecessors.

The higher the Final Pick the lower is the player's demand by organizations when he is drafted. So it is understandable for it to have a negative effect.

Higher the Games Played higher the capability of that player which understandably has a positive effect.

Free Throw % is a proxy for accuracy of shooting here and the higher it is the higher is the three point shots.

Assists are historically the job of Point Guards and Shooting guards who are also good at three pointers, so it makes sense for it to have a positive effect.

Blocks are usually the job of big centers who are not very much responsible for shooting so it make sense for it to have a negative effect.

The dropped position is Center which is the lowest in three points.

All the other positions have a positive effect on a players ability to shoot threes. Among which the highest is the Shooting Guard which historically has been the most involved player with shootings.

2.5 Performance

Unfortunately, our model does not have the best performance. In the final version of it, we reached a 50.2% in r squared which is slightly better than guessing.

On test set our performance goes even lower to 49.8% which really questions our ability to predict a high performing three-point player.

Our mean squared error is at 34 which might sound low but given the fact that players in the test set on average scored 36.7 three pointers, it is very high.

In an attempt to increase the performance, we decided to limit our time scope only to years between 2015 to 2020 however that also resulted in a low performance **(code available).**

Overall, we understand that our model has not been strong in making a good prediction. **Possible reasons for our model's inaccuracy are explained in Conclusion.**

3.0 Problem 3: Clustering players

To go beyond our original scope of analysis and to perhaps do more worthy work, we decided to cluster players based on their attributes. Data preprocessing and columns are the same as problem 2.

3.1 Process

We again limited our scope initially on purpose to manage the search and landed on these parameters to do the clustering: [HT, WT, Final Pick, 2P, 3P, AST, BLK]

We first did Hierarchical clustering with Dendrogram method using Euclidean as our metric.

We also did an analysis using K-means while using the Elbow figure to pick out the ideal number of clusters. We decided to have 4 clusters of players which turned out to give quite a meaningful result.

3.4 Results

Overall, the clusters seem to be differentiated based on two different factors: size and performance.

	clusterid	HT	WT	Final Pick	2P	3P	AST	BLK
		mean	mean	mean	mean	mean	mean	mean
0	0	1.927881	90.016462	49.651611	70.172840	22.622403	62.574526	9.856971
1	1	2.073988	109.808891	18.193297	359.257989	19.556508	135.175370	91.993765
2	2	1.960633	94.707804	22.926610	277.128321	107.808645	291.745610	22.859973
3	3	2.068955	108.858902	29.011773	83.069198	18.805863	46.512254	16.427679

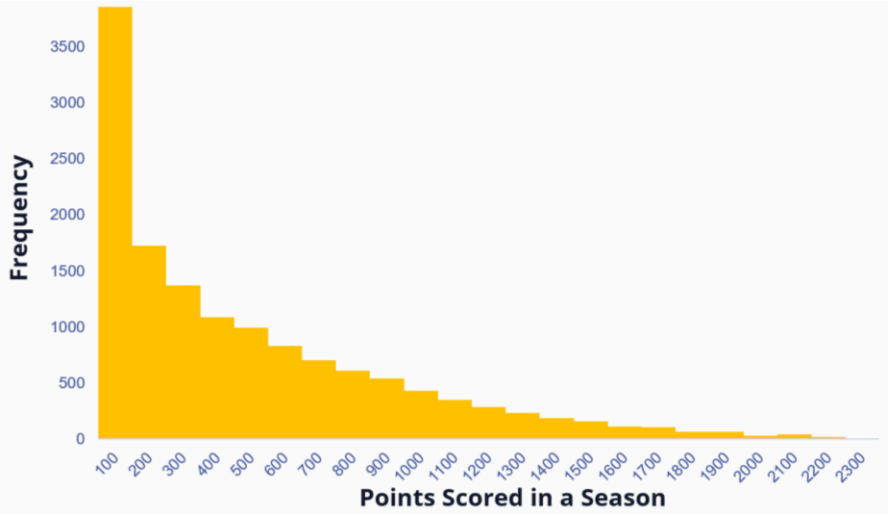
	High Performance	Low Performance
Big Players	Cluster 1	Cluster 3
Smaller Players	Cluster 2	Cluster 0

There are interesting things that we can see here. Overall big players are better in 2P and BLK in comparison to the small counterparts. And Small Players are better in 3P and AST in comparison.

Our clusterings seem to have a satisfying result. Perhaps a follow up on this analysis can be the study of player salaries and incorporating it in these clusters.

4.0 Problem 4: Points Distributions

Throughout our initial analysis of our data frames, we found that the distributions of points scored among players have a shape that we did not anticipate. We anticipated a normal distribution, meaning that most players score an average and the higher and lower we go the number of players decrease. In reality we faced a distribution like the following:



This ignited our curiosity to go deeper into this distribution which we understood is called Pareto Distribution.

4.1 Pareto Distributions

Pareto distributions are often used in economics and business to model the distribution of wealth or income, where a small percentage of the population holds a large portion of the wealth or income.

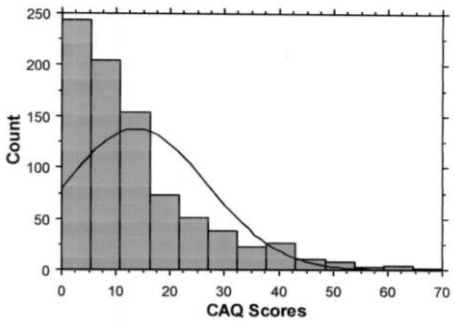
This distribution is named after the Italian civil engineer, economist, and sociologist Vilfredo Pareto. Pareto Observed that a minority of the population (20%) control the majority of the wealth (80%). That is why this distribution is called the 80-20 distribution as well.

In the context of professional sports, a Pareto distribution might be used to model the distribution of performance or skill among the players in a league.

Economist Mark J. Perry in 2008 demonstrated that among 450 NBA players, the top 20% of the players have scored 79.5% of the total points proving the existence of a Pareto distribution in the context of pro basketball.

It's worth noting that Pareto distribution are quite common when it comes to analysis of skills and performance.

The Creative Achievement Questionare is another example. CAQ was developed in 1992 in an effort to measure the creative ability of people through asking them a list of questions and ranking their answers from 0 (no creativity) to 10 (reputation for creatvity).



The distribution also follows a Pareto Distribution which further demonstrate the fact that when we put individuals in a distribution based on their performance a minority of them tend to amass the majority of the scores.

5.0 Conclusion

The hardships of predicting a player's performance are indeed baffling for many reasons. As we went over the Kaggle notebooks trying to find inspirations for our project we found that almost all of the notebook's available focus solely on descriptive analysis. Even though their findings are mesmerizing, **they never try to put in test those findings through predicting**. Pro basketball has

become incredibly more complex especially in the past two decades. This new form (Position-less) puts away the traditional notion of the game that big players should be in the center and smaller players should make passes or do shooting.

It seems that the position-less basketball has made the job of analysts much harder since the simple methods of prediction no longer works.

5.1 Future Possible Work

As mentioned earlier, before the 21st century the game of basketball has been played in a much more traditional sense. Given that, as researchers perhaps what we can do in the future is to check our methods for the data before 2000 to see if the simplicity of the game in the past allow us to do better predictions.

References

BBC. (2004, November 17). *BBC Sport Academy | Basketball | rules | players | getting to know basketball positions*. BBC News. Retrieved January 10, 2023, from http://news.bbc.co.uk/sportacademy/hi/sa/basketball/rules/players/newsid_3954000/3954283.stm

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and Factor Structure of the creative achievement questionnaire. *Creativity Research Journal*, *17*(1), 37–50. https://doi.org/10.1207/s15326934crj1701_4

Perry, M. (n.d.). *Top 20% of NBA players scored 80% of total points*. Retrieved January 10, 2023, from https://www.aei.org/carpe-diem/top-20-of-nba-players-scored-80-of-total-points/

Stuartwleung. (2020, September 16). *The Inventor(s) of the three-point shot*. Interbasket. Retrieved January 10, 2023, from https://www.interbasket.net/news/the-inventors-of-the-three-point-shot/278/

Weinfuss, J. (2021, September 23). *2021 WNBA playoffs: Brittney Griner owns the WNBA dunking record -- and is coming for more*. ESPN. Retrieved January 10, 2023, from https://www.espn.com/wnba/story/_/id/32258450/2021-wnba-playoffs-brittney-griner-owns-wnba-dunking-record-coming-more