Title
Clustering Professional Basketball Players by Performance

Author
Patel, Riki

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Clustering Professional Basketball Players by Performance

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Riki Patel

2017

ABSTRACT OF THE THESIS

Clustering Professional Basketball Players by Performance

by

Riki Patel
Master of Science in Statistics
University of California, Los Angeles, 2017
Professor Frederic R Paik Schoenberg, Chair

Basketball players are traditionally grouped into five distinct positions, but these designations are quickly becoming outdated. We attempt to reclassify players into new groups based on personal performance in the 2016-2017 NBA regular season. Two dimensionality reduction techniques, t-Distributed Stochastic Neighbor Embedding (t-SNE) and principal component analysis (PCA), were employed to reduce 18 classic metrics down to two dimensions for visualization. k-means clustering discovered four groups of players with similar playing styles. Player representation in each of the four clusters is similar across the 30 NBA teams, but better teams have players located further away from cluster centroids on the scatterplot. The results indicate that strong teams have players whose success cannot be attributed to fundamentals alone, meaning these players have advanced or intangible factors that supplement their performance.

The thesis of Riki Patel is approved.

Maryam M Esfandiari

Chad J Hazlett

Frederic R Paik Schoenberg, Committee Chair

University of California, Los Angeles

2017

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my mother, brother, and late father for their earnest support throughout my life. I would also like to thank the Statistics Department at UCLA for being extremely friendly and for providing a great M.S. program.

CHAPTER 1

Introduction

In basketball, players are grouped into five distinct positions: point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C).

Point Guard: Initiates offensive plays with great passing and dribbling skills

Shooting Guard: Positions on the wing with strong 3-point shooting ability

Small Forward: Versatile shooter who can drive to the basket

Power Forward: Strong inside presence; scores close and mid-range shots

Center: Scores near the basket with strong rebounding and blocking

It has been increasingly apparent, however, that modern National Basketball Association (NBA) players adopt playstyles that do not quite fit into one of these five positions. For example, many point guards and shooting guards have strong driving skills, and many taller power forwards and centers have great long-range shooting ability.

Is it possible that a new player designation can be determined based on the actual performance of modern basketball players? If such a designation is found, could it be used to build a strong basketball team?

We attempt to answer these questions using player performance metrics for the 2016-2017 NBA season. First, two dimensionality reduction techniques, t-Distributed Stochastic Neighbor Embedding (t-SNE) and principal component analysis (PCA), will be employed to visualize player performance in two dimensions. Next, k-means clustering will be employed to find similarities among groups of players on the scatterplot. Finally, the 30 NBA teams will be analyzed to see how their players are represented within this clustering scheme.

CHAPTER 2

Player Data

The dataset for this analysis comes from www.basketball-reference.com [1]. There are 486 players in the dataset and 18 metrics associated with each. All metrics represent per-100-possessions statistics. In contrast to per-game statistics, per-100-possessions statistics tend to better approximate player ability because some teams have a much faster offensive pace than others. The data represents player performance in the regular 2016-2017 NBA season, so each player has the same maximum number of games played. The variables are described below:

1) Points Scored

2) Field Goals Made

3) Field Goals Attempted

4) Field Goal %

5) 2-Point Field Goals Made

6) 2-Point Field Goals Attempted

7) 2-Point Field Goal %

8) 3-Point Field Goals Made

9) 3-Point Field Goals Attempted

10) 3-Point Field Goal %

11) Free Throws Made

12) Free Throws Attempted

13) Free Throw %

14) Offensive Rebounds

15) Defensive Rebounds

16) Assists

17) Steals

18) Blocks

Variables that have been removed from the full dataset are turnovers, fouls, and total rebounds. Turnovers and fouls are variables that do not reflect positive player performance, and total rebounds are encapsulated by adding offensive and defensive rebounds together.

In addition, a player's classic position (PG, SG, SF, PF, or C) and team membership have been excluded from our data. Player position is a classification we are trying to redefine, so we must exclude it from our dataset in order to minimize its influence on clustering. The team each player belongs to is similarly excluded to prevent natural grouping of teammates.

Table 2.1 below shows average statistics for players in each of the five classic positions.

Position	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%
PG	7.37	17.74	0.41	2.10	6.08	0.33	5.27	11.66	0.45
SG	6.75	16.80	0.39	2.43	7.10	0.33	4.32	9.70	0.43
SF	6.87	15.99	0.43	2.13	6.41	0.32	4.73	9.58	0.49
PF	7.33	16.55	0.45	1.52	4.77	0.27	5.81	11.77	0.49
C	7.77	15.06	0.53	0.50	1.44	0.13	7.26	13.62	0.54
FT	FTA	FT%	ORB	DRB	AST	STL	BLK	PTS	
3.73	4.67	0.77	0.92	4.59	8.18	1.90	0.41	20.56	
2.96	3.73	0.76	1.07	5.38	3.96	1.41	0.46	18.90	
3.08	3.99	0.72	1.60	6.26	3.30	1.63	0.77	18.95	
2.79	3.86	0.69	3.15	8.18	2.86	1.40	1.24	18.97	
3.57	5.31	0.65	4.67	9.81	2.79	1.40	2.15	19.61	

Table 2.1: Average Performance of NBA Players in Classic Positions

CHAPTER 3

t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a modern machine learning algorithm used for dimensionality reduction. The algorithm is extremely useful for visualizing high-dimensional data on a 2- or 3-dimensional plot.

In contrast to principal component analysis (PCA), a linear dimensionality reduction technique, t-SNE is a non-linear dimensionality reduction technique. t-SNE is able to take high-dimensional data points that lie on or near a non-linear manifold and preserve the local structure when mapping onto a low-dimensional space, which is not possible with any linear technique [2].

Panels A and B of Figure 3.1 display the famous "Swiss Roll" dataset, which is a non-linear manifold in three dimensions [3]. Panel C shows the application of PCA to the data and subsequent plotting in two dimensions. The data is essentially flattened down without any regard to the underlying spiral manifold relationship among points because it is a linear mapping. The blue and red data points are neighbors in two dimensions despite being far apart on the original manifold. When the non-linear dimensionality reduction method Isomap is used, however, we see in Panel D that the spiral manifold actually unrolls itself, preserving local structure of the original data.

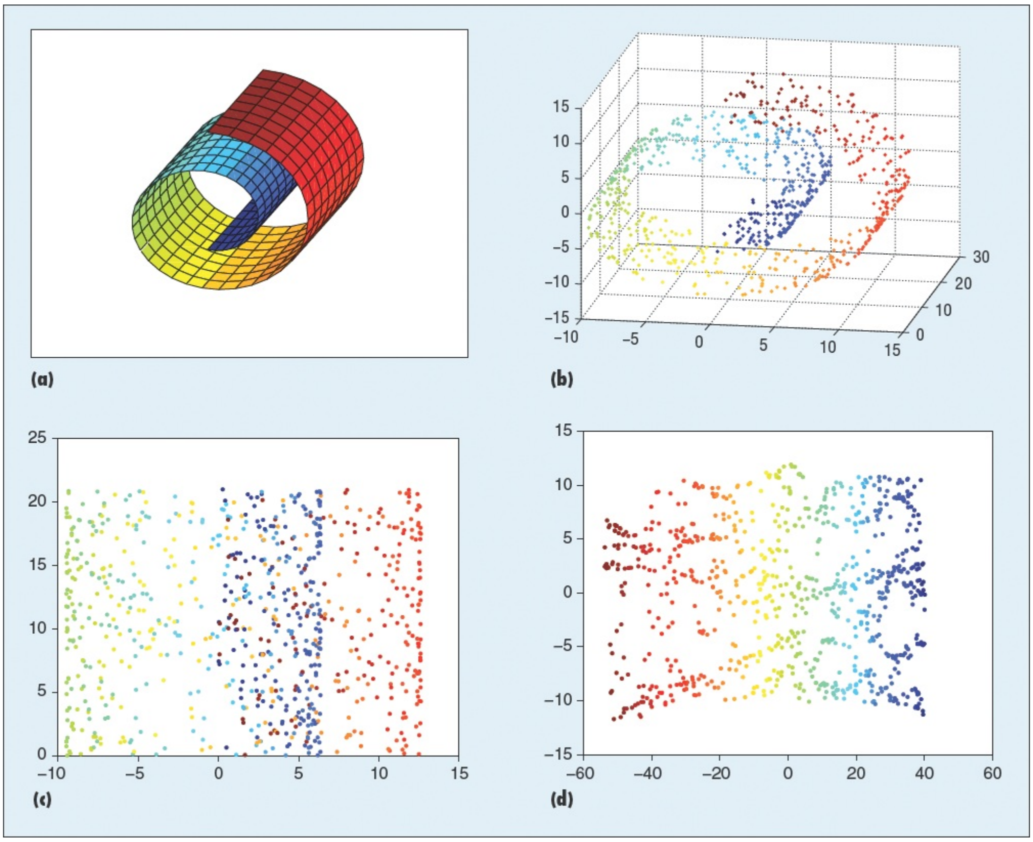

Figure 3.1: Application of PCA (c) and Isomap (d) to Swiss Roll Data

The full t-SNE algorithm is described below:

Data: $X = x_1, x_2 ... x_n$

Cost Function Parameter: Perplexity $Perp$;

Optimization Parameters: Number of iterations T, Learning rate η, Momentum α(t);

Result: Low-dimensional data representation $Y^{(T)} = y_1, y_2 ... y_n$

begin

Compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ using Equation 1;

Set $p_{ij} = \frac{p_{ji}+p_{ij}}{2n}$;

Sample initial solution $Y^{(0)} = y_1, y_2 ... y_n$ from $N(0, 10^{-4}I)$;

for *t=1 to T* **do**

Compute low-dimensional affinities q_{ij} using Equation 2;

Compute gradient $\frac{\delta C}{\delta Y}$ using Equation 3;

Set $Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$;

end

end

Equation 1: $p_{j|i} = \frac{exp(-||x_i - x_j||^2)/2\sigma_i^2}{\Sigma_{k \neq i} exp(-||x_i - x_k||^2)/2\sigma_i^2}$

Equation 2: $q_{ij} = \frac{(1+||y_i - y_j||^2)^{-1}}{\Sigma_{k \neq l}(1+||y_k - y_l||^2)^{-1}}$

Equation 3: $\frac{\delta C}{\delta y_i} = 4\Sigma_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i + y_j||^2)^{-1}$

CHAPTER 4

Application of t-SNE and PCA to Player Data

t-SNE has one cost function parameter, perplexity, and three optimization parameters: number of iterations, learning rate, and momentum. The optimal set of parameters varies for each dataset, so we have chosen common values that perform well based on existing literature in order to minimize bias [4]. The values are described in Table 4.1

Parameter	Value
Perplexity ($Perp$)	30
Number of Iterations (T)	1000
Learning Rate (η)	200
Momentum ($\alpha(t)$)	0.5

Table 4.1: Selected Values of t-SNE Parameters

Figure 4.1 below displays the application of t-SNE to our 486 x 18 player data. The 486 players are now represented as coordinates on a 2-dimensional scatterplot. The legend shows the classic basketball position (PG, SG, SF, PF, C) of each player to help us understand whether natural grouping by these positions is occurring. Evidently, basketball power forwards and centers are highly grouped together while the other three positions are relatively mixed among each other. Also, there appears to be some grouping of point guards near the center of the scatterplot.

Figure 4.2 is a plot of the first two principal components after performing PCA on the dataset. We are interested in seeing whether PCA, the standard for dimensionality reduction, performs similarly to t-SNE. Evidently, PCA produces a 2-dimensional plot that has much more crowded points. There is still a nice grouping of NBA centers toward the top of the main group, but everything else is quite fuzzy. The most important variables determined by PCA are field goals made/attempted, 2-point field goals made/attempted, 3-point field goals made/attempted, and offensive/defensive rebounds.

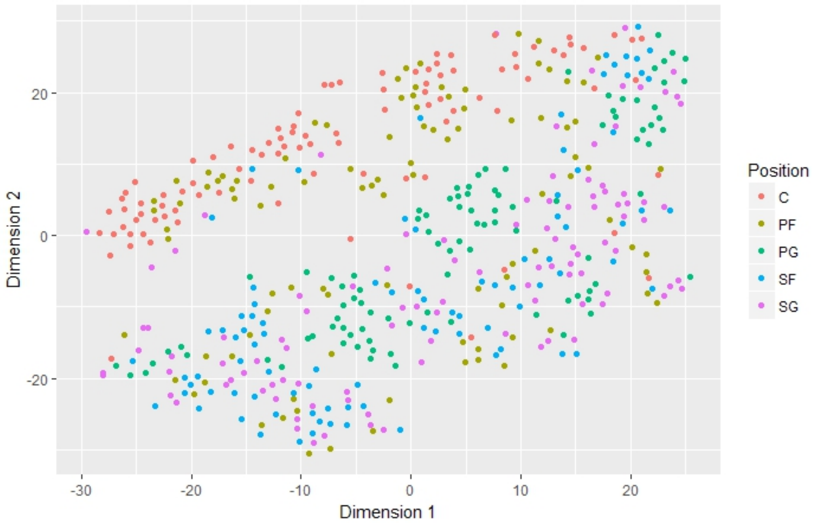

Figure 4.1: Application of t-SNE to 2016-2017 NBA Player Data

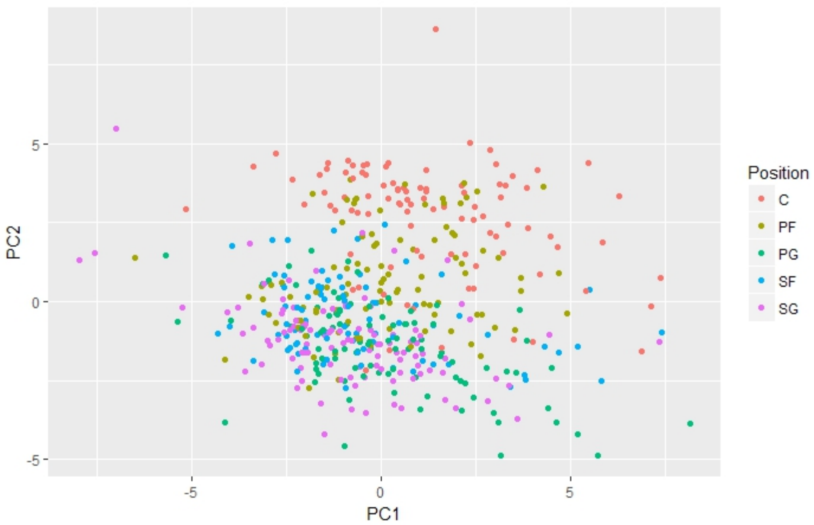

Figure 4.2: Application of PCA to 2016-2017 NBA Player Data

CHAPTER 5

k-Means Clustering and Characterization of Clusters

k-means clustering is one of the most widely used clustering techniques because of its inherent simplicity. Given a set of observations, we are interested in finding a set of k points, called centroids, that minimize the mean-squared distance from each data point to its nearest centroid [5]. The value of k represents how many clusters are present. We are interested in seeing how many general clusters that NBA players fall into.

Three validation methods will be used to select the best number of clusters: within-sum-of-squares (WSS) [6], the silhouette score [7], and the gap statistic [8].

$$WSS : \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

where k = number of clusters, n = number of points,and c = center for cluster j

$$Silhouette\ score : s(i) = \frac{b(i) - a(i)}{max\left\{a(i), b(i)\right\}}$$

where $a(i)$ is the average dissimilarity of i to other points in the same cluster and $b(i)$ is the lowest average dissimilarity of i to other clusters

$$Gap\ Statistic : Gap_n(k) = E_n^*\left\{log(W_k)\right\} - log(W_k)$$

where E_n^* denotes expectation from a simulated reference distribution for sample size n and W_k is the pooled within-cluster sum of squares

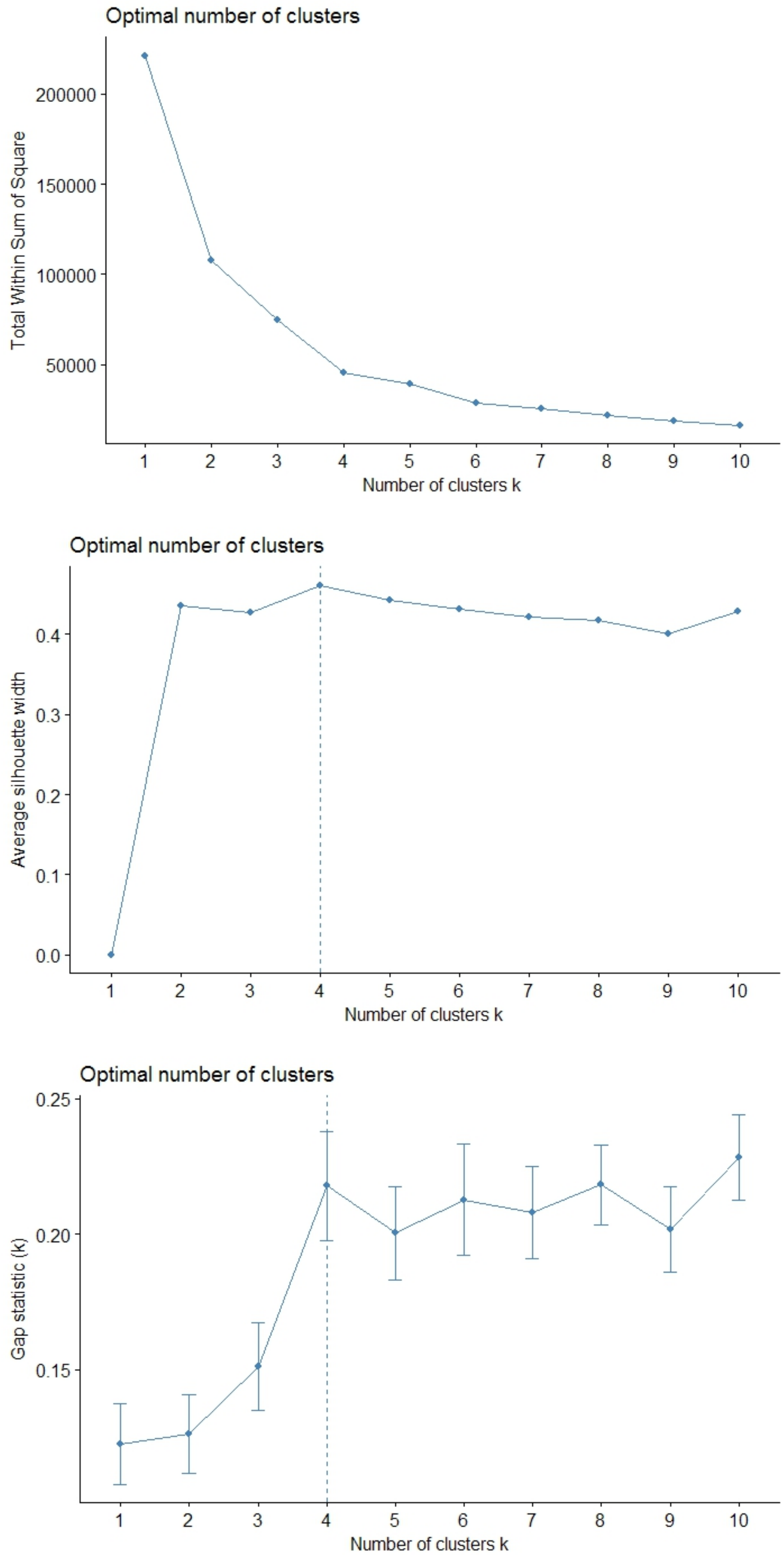

Figure 5.1: Cluster Validation Measures for t-SNE Scatterplot

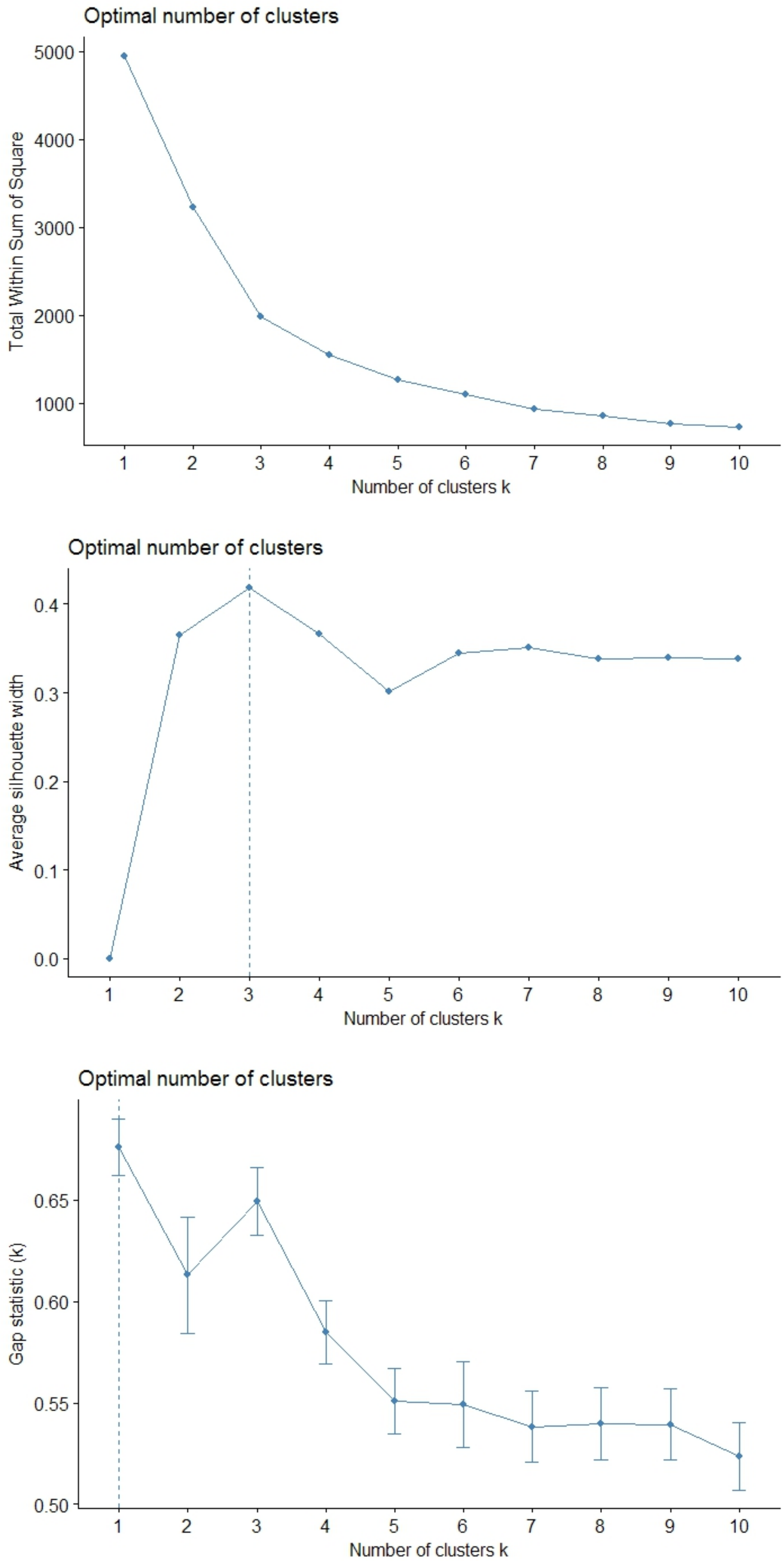

Figure 5.2: Cluster Validation Measures for PCA Scatterplot

Figure 5.1 suggests that all cluster validation measures select four clusters to be optimal for our t-SNE scatterplot. The within-sum-of-squares metric is based on an "elbow", or the number of clusters where reduction in total WSS begins to level off. This is seen at two and four clusters in the first panel of Figure 5.1, but the elbow at four clusters is slightly more pronounced. Since all three measures lead to the same result, we can confidently say that the t-SNE scatterplot can be broken into four groups.

In Figure 5.2, the WSS and silhouette score point to three clusters for our PCA scatterplot, while the gap statistic suggests a single cluster. Moving forward, we will disregard this PCA scatter plot because, visually, the data is much more clumped together than our t-SNE scatterplot, and our three measures could not reach a consensus on number of clusters.

Figure 5.3 below represents our final four-group clustering scheme based on the t-SNE scatterplot.

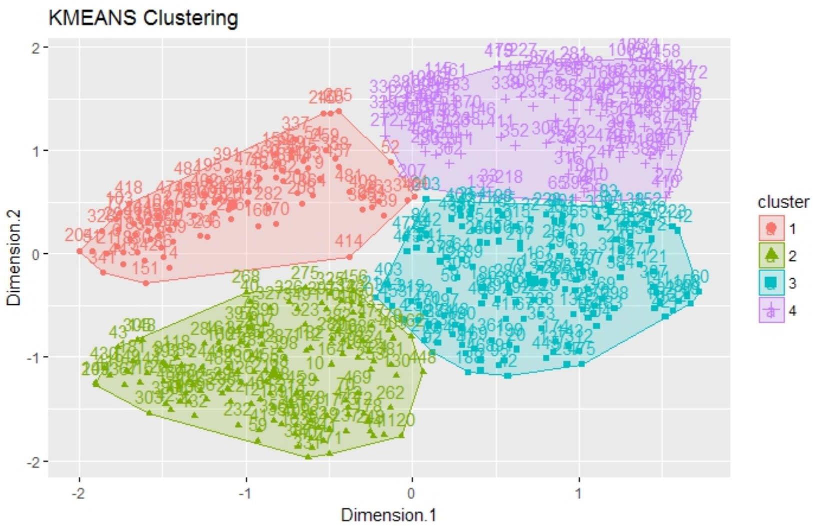

Figure 5.3: Four-Group Clustering Scheme of NBA Players Based on 2016-2017 Performance

The next step is to characterize our clusters; what do players in each cluster excel at? Table 5.1 shows the average performance of players in each of the four clusters.

Cluster	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	AST	STL	BLK	PTS
1 (Red)	6.07	11.82	0.52	0.30	0.96	0.13	5.78	10.86	0.53	2.91	4.45	0.64	4.50	9.59	2.54	1.46	1.87	15.35
2 (Green)	4.78	12.86	0.38	1.85	5.98	0.29	2.93	6.87	0.43	1.89	2.50	0.69	1.26	5.37	4.40	1.65	0.60	13.30
3 (Blue)	7.50	17.76	0.42	2.58	7.32	0.35	4.92	10.44	0.47	3.01	3.89	0.76	1.30	5.38	4.52	1.50	0.60	20.57
4 (Purple)	10.45	22.29	0.48	1.69	4.78	0.28	8.77	17.50	0.51	5.22	6.73	0.76	2.95	8.19	4.92	1.55	1.30	27.82

Table 5.1: Average Performance for Each Cluster

Cluster 1 (Red): ***The Paint Protectors***, moderate scorers with great 2-point shots, rebounds, and blocks but poor 3-point shooting ability

Cluster 2 (Green): ***The Supporters***, relatively low scorers with good assists, stealing ability, and decent 3-point shooting ability

Cluster 3 (Blue): ***The Shooters***, moderate scorers with great free throw, 2-point, and 3-point shooting ability

Cluster 4 (Purple): ***The Insiders***, high scorers who excel at free throws, 2-point shots, rebounds, and blocks

CHAPTER 6

Determining the Relationship Between Clusters and NBA Team Success

Our clustering scheme is only useful if it can be linked to the strength of basketball teams as a whole.

Two scenarios will be investigated with regards to our clustering scheme. One scenario is that better teams could have players with either balanced or imbalanced cluster representation. The second scenario is that better teams may have players who are closer or farther from the centroid of each cluster.

Figures 6.1 and 6.2 show how players from the best basketball team (as measured by win percentage) in the 2016-2017 NBA season, the Golden State Warriors, compare to the worst team, the Brooklyn Nets.

Based on visual inspection, it appears that there are differences in cluster representation; Golden State has a high number of players in clusters 1 and 4 while Brooklyn has strong representation in clusters 2 and 3. In addition, Golden State players seemed to be more tightly grouped together and further away from cluster centroids.

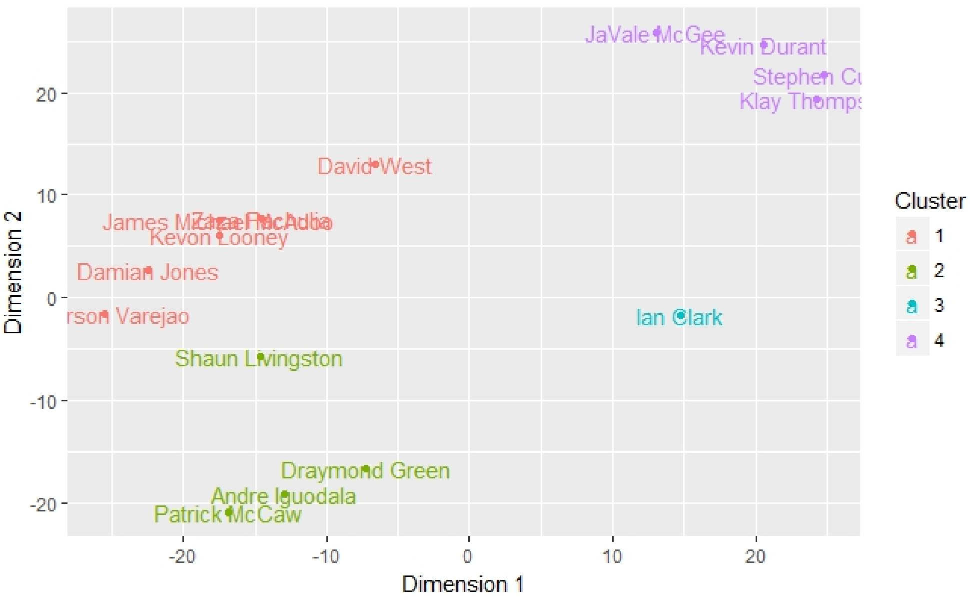

Figure 6.1: Representation of the Golden State Warriors in the Four Cluster Scheme

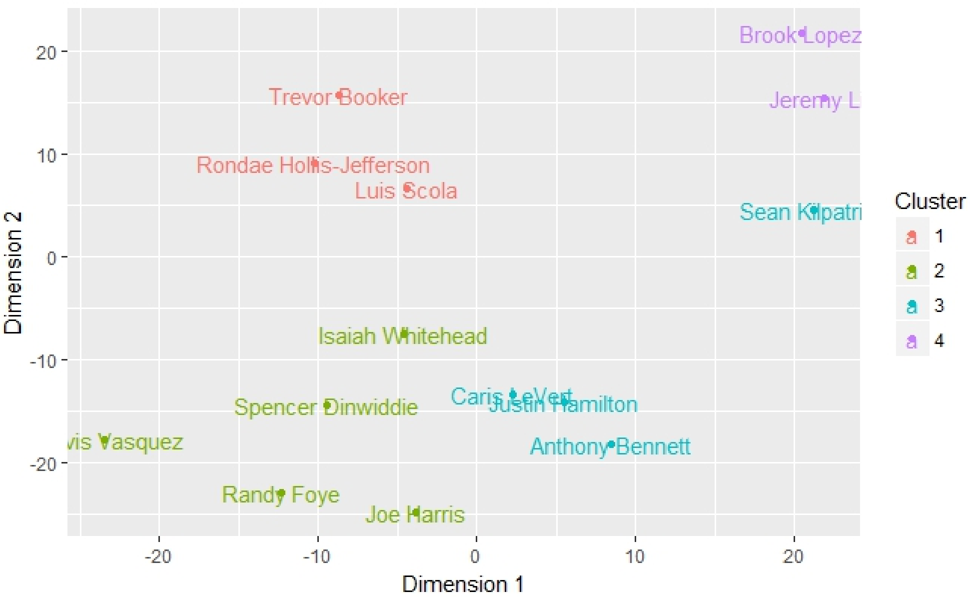

Figure 6.2: Representation of the Brooklyn Nets in the Four Cluster Scheme

The first scenario, player representation in clusters, will be investigated by finding what fraction of players reside in each cluster for all 30 NBA teams. Linear regression will be applied to plots of team rank versus cluster fraction to see if there is a trend. For example, we notice that Golden State has many players in cluster 1. It could be possible that better teams tend to favor players in cluster 1, and this will be revealed through linear regression.

Figure 6.3 shows team rank versus fraction of players in each cluster. There is no significance when applying linear regression to all four plots. This suggests that there is no relationship between how good a team is and membership in a particular cluster. All clusters are equally important, suggesting there is no one type of player that dominates the NBA.

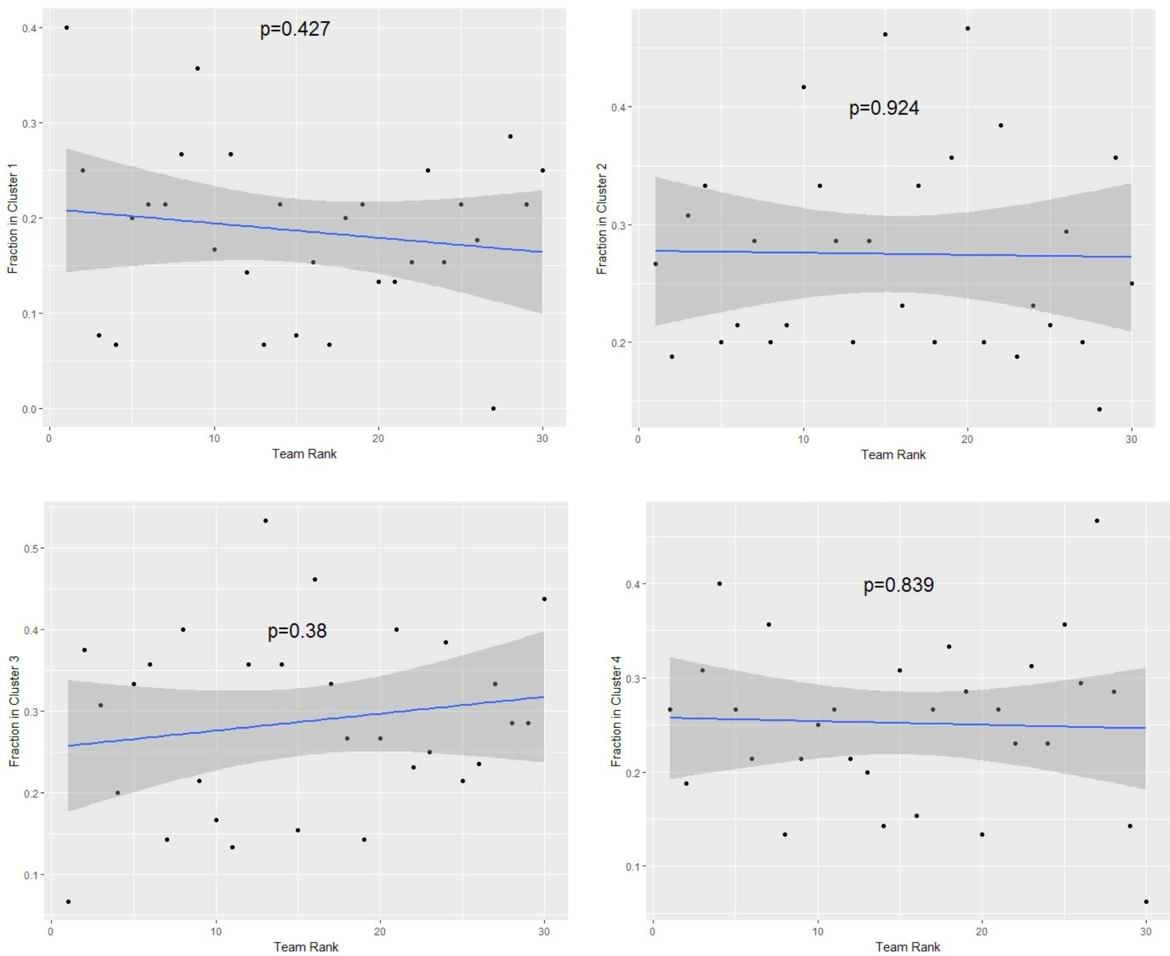

Figure 6.3: Team Rank Versus Membership in Each Cluster

The second scenario, distance of each player to his nearest cluster centroid, will be analyzed by finding the simple Euclidean distance from a player's location on the scatterplot to the closest centroid within his respective cluster. Each team will have a grand total from summing all of the players' distances. To account for teams with varying number of players, 12 players will be randomly selected from each team before calculating a total. This process will be repeated 20 times and results will be averaged for consistency.

Figure 6.4 shows a remarkable result. Better teams have players *further* away from their nearest centroids, while worse teams have more grouping toward the centroid. This result is revealed through a p-value of 0.02545 for the slope in linear regression.

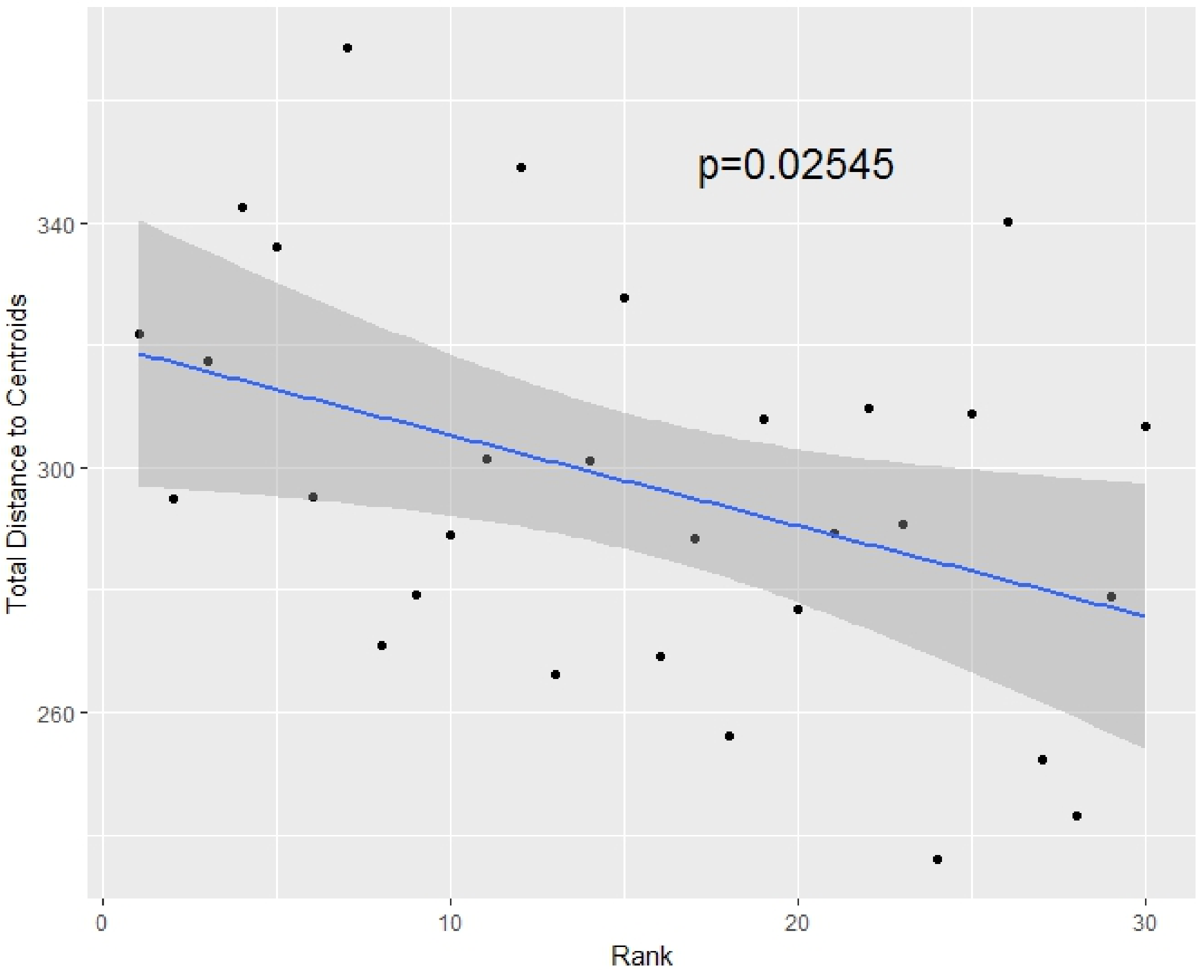

Figure 6.4: Team Rank Versus Total Player Distance to Nearest Cluster Centroid

It is possible that poorer players are grouped toward the center of clusters while better players are grouped toward the periphery of clusters. This would explain why better teams have players with significantly more distance from centroids on our cluster scatterplot.

Table 6.1 displays 3 players from each cluster who are near the centroids.

Player	Cluster	Pos	Tm	FG	FGA	F%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	AST	STL	BLK	PTS
Gorgui Dieng	1	PF	MIN	6.30	12.60	0.50	0.30	0.80	0.37	6.00	11.80	0.51	2.60	3.20	0.81	3.60	8.80	3.00	1.70	1.80	15.60
Amir Johnson	1	PF	BOS	6.60	11.40	0.58	0.80	2.00	0.41	5.70	9.40	0.61	2.10	3.10	0.67	3.60	7.60	4.30	1.60	1.90	16.00
Larry Nance Jr.	1	PF	LAL	6.40	12.20	0.53	0.30	1.20	0.28	6.10	11.00	0.55	2.00	2.70	0.74	4.10	8.40	3.20	2.80	1.30	15.20
Pat Connaughton	2	SG	POR	5.80	11.30	0.51	2.70	5.20	0.52	3.10	6.10	0.51	1.10	1.40	0.78	1.60	6.60	4.40	0.90	0.30	15.40
Andre Iguodala	2	SF	GSW	5.30	10.00	0.53	1.50	4.30	0.36	3.70	5.70	0.65	1.70	2.50	0.71	1.20	6.10	6.30	1.80	0.90	13.80
C.J. Watson	2	PG	ORL	4.70	12.10	0.39	1.60	5.10	0.30	3.10	7.00	0.45	2.80	3.20	0.86	0.80	3.60	5.60	2.10	0.10	13.80
Marquese Chriss	3	PF	PHO	7.80	17.40	0.45	2.00	6.20	0.32	5.80	11.20	0.52	3.10	5.00	0.62	2.60	6.90	1.60	1.80	1.90	20.70
Jeff Green	3	PF	ORL	7.10	18.00	0.39	1.70	6.20	0.28	5.40	11.80	0.46	4.70	5.40	0.86	1.30	5.70	2.60	1.20	0.40	20.60
Malachi Richardson	3	SG	SAC	7.20	17.40	0.41	2.00	7.20	0.29	5.10	10.20	0.50	3.80	4.90	0.79	0.80	5.10	2.80	1.30	0.30	20.20
Harrison Barnes	4	PF	DAL	11.10	23.80	0.47	1.40	4.10	0.35	9.70	19.60	0.49	4.50	5.20	0.86	1.70	5.60	2.20	1.20	0.30	28.20
Rudy Gay	4	SF	SAC	10.00	22.10	0.46	2.10	5.60	0.37	7.90	16.40	0.48	5.90	6.90	0.85	1.70	7.70	4.10	2.20	1.30	28.10
Nikola Jokic	4	C	DEN	11.80	20.50	0.58	1.10	3.30	0.32	10.80	17.10	0.63	4.50	5.50	0.82	5.10	12.10	8.60	1.50	1.30	29.30

Table 6.1: Players Near the Centroid of Each Cluster

Arguably, these players who are grouped near the centroid of each cluster are great performers. One of the players, Andre Iguodala, was instrumental in helping the Golden State Warriors win the championship. All of these players are great scorers for their respective positions and would not be considered weak performers in the 2016-2017 NBA season.

Thus, the implication of this final result is that better teams have players who exhibit special characteristics that cannot be captured by the basic metrics used in this analysis. There are countless advanced statistics that can be measured for a player, such as loose balls recovered, second-point opportunity success, scoring based on distance from the basket, and offensive/defensive rating, but our dataset included only 18 variables that represent the core metrics of basketball so we can easily characterize classic NBA positions and our new clusters.

Many other scholars have attempted to reclassify NBA positions in a similar fashion. Muthu Alagappan of Stanford University found that there are actually 13 groups of players [9], and Alex Cheng of Cornell University found 8 groups of players [10]. So, it is clear that strong basketball teams do more than excel at the fundamentals; there are higher-level metrics beyond those used in this thesis that contribute to player performance.

CHAPTER 7

Conclusion

We have identified a novel clustering scheme based on player performance in the 2016-2017 NBA regular season. 18 performance variables were reduced down to 2 using the dimensionality reduction technique t-SNE, and four groups of similar players were identified using k-means clustering.

All NBA teams have equal player representation in the four clusters, but better NBA teams had players who resided further away from cluster centroids than worser NBA teams. It was determined that strong and weak players are homogeneously distributed throughout our clusters, so better players residing on the periphery of clusters must have additional elements that account for their success beyond the 18 basic metrics used in this thesis.

Further exploration will seek to cluster players using as many variables as possible and with a variety of methods. Exactly what these advanced metrics are that account for NBA team success must be determined. Is it plausible that team chemistry, an unmeasurable factor, largely influences whether or not a championship is won, but the science of sports is certainly heading in a quantitative direction, so all available data must be analyzed.

References

1) https://www.basketball-reference.com/leagues/NBA_2017_per_poss.html, 2017.

2) L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 9(Nov):2579-2605, 2008.

3) Junping Zhang, Jue Wang, Hua Huang. Manifold Learning for Visualizing and Analyzing High-Dimensional Data. IEEE Intelligent Systems, Vol. 25, pp. 54-61, July/August 2010.

4) L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research, 15(Oct):3221-3245, 2014.

5) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (7): 881892, 2002.

6) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society, Series C, 28 (1): 100108, 1979.

7) Peter J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, 20: 5365, 1987.

8) Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63.2: 411-423, 2001.

9) Alagappan, Muthu. From 5 to 13: Redefining the Positions in Basketball. http://www.sloansportsconference.com/people/muthu-alagappan, 2012.

10) Cheng, Alex. Using Machine Learning to Find the 8 Types of Players in the NBA. https://fastbreakdata.com/classifying-the-modern-nba-player-with-machine-learning-539da03bb824, 2017.