Re-classifying NBA Players Using Unsupervised Clustering

Law Tsz Kin
School of Electronic Engineering and Computer Science, Queen Mary University of London

(Dated: March 19, 2022)

In modern basketball game, only five positions are not enough to accurately describe players' diverse playstyles. This research paper aims at redefining classes for NBA players using two unsupervised learning methods: K-Means Clustering and Gaussian Mixture Models. Players' metrics were crawled and then further cleansed. Concerning the dimensionality of the dataset, feature selection using correlation matrix and dimensionality reduction via Principal Component Analysis were both applied. The processed matrix was fit into both clustering method and Gaussian Mixture Models yielded desirable results with 8 clusters being successfully grouped.

I. INTRODUCTION

As basketball game has become more dynamic, players' playstyles must also follow to adapt the change. Traditionally, basketball game has five positions. Point guards usually organize team offense. Shooting guards focusing on perimeter scoring. Small forwards drive and attack the rim frequently. Power forwards and centers mainly take responsibility of rebounding, rim protection and inside scoring. However, the above descriptions might not fit to present current NBA players with respective positions. For example, Nikola Jokic, the center of Denver Nuggets, averaged 8.2 assists and 4.2 three-pointers attempts this season. Wolves' power forward, Karl-Anthony Towns has just won the three-point contest in the all-star week with the record-breaking 29 points. There is an increasing number of players that are considered as "abnormal" when we apply the old-fashioned five position categories to examine their performance.

Stepping into 21st century, people often claim that "data is the new oil". Through appropriate process, data could be transformed into valuable information. In the field of sports a complete set of player's statistics could accurately describe his playstyle and potentially help improve his skillset. Nowadays, NBA teams are hiring basketball data scientists to help coaches to analyze the trend of the game and team's performance and devise smarter strategies to win [8]. Inspired from that, the objective of this data analytic report is to re-classify NBA players through machine learning algorithms and investigate how the new classes shape today's basketball game.

II. LITERATURE REVIEW

To obtain insights for this research's objective, relevant academic journals and discussions are studied and analyzed in this part. The ideas and core concepts would also contribute to the latter sections.

Sports analytics – Evaluation of basketball players and team performance, 2020

This report serves as an overview of the sports analytics applications in basketball industry. Data specialists in sports organizations are combining data mining techniques and machine learning models to generate new insights on team composition, player evaluation, coaching and on-site tactics. This research provided popular data analysis methods used in sports, such as Neural Networks, Decision Trees, Linear and Logistic Regression and clustering. Useful metrics were listed with descriptions, followed by formula of converting them into advanced rating key performance indicators (KPI). Case studies were conducted on the top five NBA players in the 2018-19 season, setting as examples of how the KPI could generate a comprehensive and predictive basketball analysis [6]. This research displayed how the captured complex players' metrics could be useful parameters in machine learning models.

Application of K-Means Clustering Algorithm for Classification of NBA Guards, 2016

This report studied the implementation of K-Means Clustering method on classifying NBA guard players. The dataset included only scores, rebounds and assists of players in that season as the features for clustering analysis. The researchers used Euclidean Distance between centroids and data points, along with mean square error function to calculate the optimal number of clusters. Eventually, NBA guards were separated into six categories [15]. This research can serve as an approach guideline and an introduction to K-Means Clustering techniques.

Defensive Player Classification in the National Basketball Association, 2016

This report studied the implementation of Gaussian Mixture Models (GMM) clustering method on classifying NBA players in a defensive perspective. The dataset used

was the players' metrics in that season and features selected were mainly defensive attributes. After cleansing and normalizing the dataset, GMM was introduced to perform clustering analysis. The number of clusters was optimized by calculating Bayesian Information Criterion scores. The GMM model eventually produced five clusters [13]. This research provides GMM as an alternative approach of studying our unsupervised learning classification problem.

NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball & modeling line up efficiency of soft lineup aggregates, 2020

This report was published during 14th MIT SLOAN Sports Analytics Conference. The research supported the idea that basketball has already become a "position-less" sport as the tradition five positions could not define modern NBA players' skillsets and preferences accurately. The researcher scraped NBA player statistics from 2009 to 2018. Most variables were first filtered and the final 23 selected variables were further compressed to a lower dimensional space through Principal Component Analysis (PCA). K-Means Clustering and GMM were tried out and K-Means Clustering performed poorly due to the unfavorable silhouette score. Meanwhile, GMM produced satisfactory results and nine distinctive clusters were created [4]. This report provides strong confidence to our research topic and an insight of conducting feature selection with dimensionality reduction could yield a better model. In addition, this study previewed the potential failure of K-Means Clustering and GMM could be a superior algorithm for our objective.

III. METHODOLOGY

Referencing from the above section, K-Means Clustering and GMM would be used for this research topic. This part will briefly introduce their basic principles and provide rationale for choosing them.

K-Means Clustering

K-Means Clustering is a classic clustering technique for unsupervised learning. To initialize the model, we input the unlabeled dataset and a number k to determine the number of clusters. At first the algorithm will initialize k centroids in random positions. Then data points are assigned to the closest centroid. The grouping of data points under the proximity of each centroid forms a cluster. After that, centroid positions will be re-calculated to minimize the Euclidean distances between the centroid and data points within each cluster. Data points are then re-assigned to the nearest centroid. The model learns clustering via repeating the above process until the centroid positions are optimized so that their distances with the data points within their own cluster are shortest.

$$E_{KM}\left(D, c_{1:N}, \mu_{1:K}\right) = \sum_{i=1}^{N} \left(N_i - \mu_{c_i}\right)^2$$

[1]. Elbow method is commonly used to find the optimal k via plotting the changes of the Within Cluster Sum of Square (WCSS) along with increasing number of clusters. The graph should be in the shape of an elbow and the elbow point is usually the optimal [2].

K-Means Clustering Justification

This algorithm aims at finding groups and patterns in dataset where data are not explicitly labeled. The objective of K-Means Clustering aligns with the business logic of our research topic [14]. Moreover, K-Means Clustering is simple to implement, easy to interpret and lower in computational cost [16]. Therefore, it is the clustering of choice for unsupervised learning in usual terms.

Gaussian Mixture Models (GMM)

GMM, as implied by the name, is a superposition of multiple Gaussian distributions [10].

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$$

It often serves as an alternative clustering method for unsupervised learning problems. The model addresses the issue that a data point could fit into multiple clusters but inclined to a particular one. GMM is a probabilistic algorithm returning confidence of a data point belonging to a certain cluster. To set up the model, we input the unlabeled dataset and a number k to determine the number of clusters. k Gaussians will be initialized and then learnt by density estimation. With the learnt Gaussians distributions, we could compute the probabilities of each point belonging to each cluster. As the distributions are not evenly proportionate, the weights for them are also considered [1][9]. The above processes are repeated until all probabilities are maximized. Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are two model functions typically used to find k for GMM. Their implementations are similar with the Elbow method.

GMM Justification

When reviewing our dataset, it should be reminded that each player is not limited to only one playstyle. The difference in terms of performance between players could

create great variance within the same group which might cause the shape of the cluster to be unpredictable. Meanwhile, GMM as a probabilistic clustering algorithm, applies soft assignment to data points and enables non-linear decision boundaries and non-spherical shape for clusters. Due to the weight of Gaussian distributions, clusters could have different sizes and volumes [1][5]. These characteristics provide flexibility and match the traits of our dataset. It is reasonable to believe that GMM has the power to yield a robust and accurate model.

IV. DATA MANAGEMENT

Data Pipeline

1. Use for-loop to collect and store all necessary sets of NBA players' statistics
2. Use for-loop to transform all static datasets to dataframes
3. Remove noise and redundant features
4. Apply correlation matrix to select distinctive features and rename them
5. Repeat step 3 – 4 for each dataframe
6. Merge all dataframe to create a player dashboard
7. Apply dimensionality reduction via PCA on the player dashboard
8. Apply both K-Means Clustering and GMM to the selected components

Data Collection

As I intended to use recent (season 2020-21) and customizable NBA player statistics, I decided to crawl the data by myself. Each page of players' metrics were scraped to an Excel workbook. As this research aimed at re-classifying NBA players, in order to precisely illustrate the diversities, 20 sets of features were crawled.

Data Cleansing

The scraped datasets were raw. Thus, before conducting feature selection, we should firstly remove the noise or redundant data in it. Example code as follow:

```
general_traditional.drop(['Unnamed:␣0', '
    ↪ PLAYER_ID', 'NICKNAME', 'TEAM_ID' # ...
    ↪ more redundant features], axis=1,
    ↪ inplace=True)
```

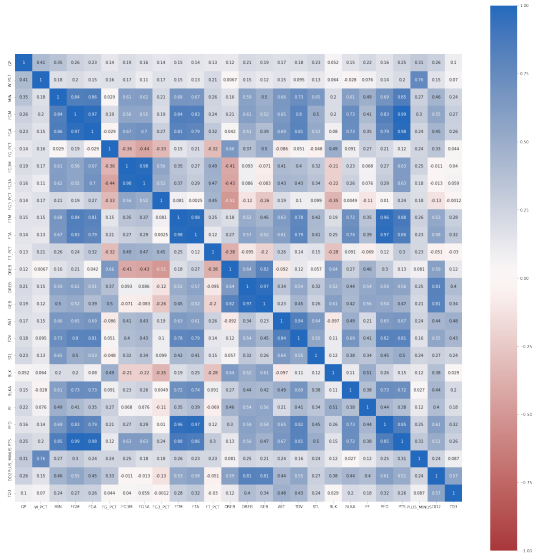

FIG. 1. This was the heatmap for dataframe "general_traditional". For example, field goal made (FGM), field goal attempted (FGA), points (PTS) obviously had strong positive relationship. FGM and FGA were dropped due to the reason that a player's scoring ability and accuracy could be reflected by PTS and FG_PCT respectively. They were more useful when acting as discrete performance indicators.

Feature Selection

After cleaning the dataset, correlation matrix was applied to the remaining variables for each dataframe. The aim of the evaluation was to test the similarity or likelihood between variables. Features with strong positive correlation could be considered as duplicate and then dropped. Personal domain knowledge was also contributed to the feature selection process apart from the statistical method. Eventually, all dataframes would only consist of distinctive features and then were then merged into one giant matrix (396 samples x 89 features).

```
general_traditional_corr = general_traditional
    ↪ .corr()
plt.figure(figsize=(25,25))
sns.heatmap(general_traditional_corr, vmin=-1,
    ↪  vmax=1,center=0, cmap='vlag_r', square=
    ↪ True, annot=True)
```

Dimensionality Reduction

Although only 89 features remained in the dataset, it was still considered as high dimensional data. The "Curse of Dimensionality" causes the existing data to be dispersed. As a result, it aggravates the computation loading and makes the model harder to interpret [3]. Principal Component Analysis (PCA) mitigates this problem by diminishing the model complexity while retaining as much in-

TABLE I. Top 10 Principal Components Variance Explained

	Variance	**% variance**	Cum. % variance
PC1	19.926538	0.231119	0.231119
PC2	17.214583	0.199664	0.430783
PC3	5.223597	0.060586	0.491369
PC4	3.293922	0.038205	0.529574
PC5	3.045287	0.035321	0.564895
PC6	2.494112	0.028928	0.593823
PC7	2.182078	0.025309	0.619132
PC8	1.833290	0.021263	0.640395
PC9	1.642621	0.019052	0.659447
PC10	1.590081	0.018443	0.677890

formation as possible [11].

Firstly, the player dashboard matrix was standardized to prevent large scale data dominating the dataset and deviating the result as PCA includes calculations of variance [11]. The next step was fitting standardized metrics into the PCA algorithm. Afterwards, the explained variance of each Principal Component (PC) was computed. It is a figure reflecting the amount of variation each PC accounted for. The first four PCs had explained over a half of the total variance of the original dataset. I believed slightly above 60% could be the threshold which balanced the complexity and interpretability of the model. Eventually, we selected the first seven PCs to compose the new feature subspace (396x7).

V. ANALYSIS

Interpreting Principal Components

Each PC owns an eigenvector. Each eigenvector is a linear combination of original dataset features. The proportions of each feature are the loading scores [12]. They reflect the association of each variable with a PC. Table I records the top 20 features with highest absolute loading score magnitude for each component. From that, we observed:

PC1 It rewards players who get many rebounds, often create screens for teammates, contest 2-point shots, always cut or roll towards the basket and attempt putbacks. It penalizes players who are the ball handlers during pick-and-rolls, do handoffs, have good passing skills and score from the perimeter.

PC2 It generally rewards players who have great scoring abilities, can play isolation and post-up games, make passes to facilitate offense, make defense impact at almost all areas, have high player-impact-efficiency and have high usage.

PC3 It rewards players who have many assists or potential assists, have high assist-to-pass ratio and are good at passing the ball around. It penalizes players who play transitions, make spot-up shots, use off-screens and cut towards the basketball.

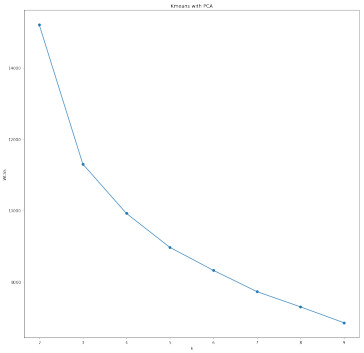

FIG. 2. Elbow Method

PC4 It rewards players who defend and contest 3-point shots, make deflections, have many steals, are good at fighting defensive loose balls, commit personal fouls and make spot-up shots. It penalizes players who have other scoring abilities.

PC5 It rewards players who defend the perimeter and the paint area, contest both 2-point and 3-point shots, make long distance shots and have good accuracy. It penalizes players who play transitions, cut towards the basket and make putbacks.

PC6 It rewards players who have high usage, have midrange and inside scoring ability, are good at playing post-up games and use off-screens. It penalizes players who play transitions, make spot-up shots, make deflections, get steals and fight defensive loose balls.

PC7 It rewards players who use off-screens, do handoffs, have good field goal percentage, have high true shooting percentage. It penalizes players who make spot-up shots, play post-up games and play transitions.

K-Means Clustering Model

Before fitting the model to K-Means Clustering algorithm. We had to find the optimal number of clusters k by observing the change of WCSS and select the k using Elbow method. From Fig 2, the elbow point showed up when the number of clusters was 3. After that, we fit the model with k = 3 and generated a heatmap to further investigate the correlation between PCs and each cluster.

```
km = KMeans(n_clusters=3, init='k-means++')
km.fit(x)
```

From that we observed:

Cluster 0 - Big Men This cluster was heavily biased towards PC1 which described tendencies of reward-

TABLE II. Principal Components with their top 20 features with highest absolute loading score magnitude

PC1		PC2		PC3		PC4		PC5		PC6		PC7	
OREB%	0.191	PTS	0.199	AST%	0.263	3PT_DFGM	0.235	BT10_15FT_DFGM	0.267	USG%	0.249	SPOT_UP_FG%	-0.292
PNR_HANDLER_PPP	-0.172	PIE	0.197	TRANSITION_PPP	-0.224	CONT_SHOTS_3PT	0.234	BT16FT_3PT_DFGM	0.245	TRANSITION_FG%	-0.240	SPOT_UP_PPP	-0.257
SCREEN_AST	0.170	DREB	0.196	SPOT_UP_POSS	-0.220	DEFLECTIONS	0.230	TRANSITION_FG%	-0.239	BT6_9FT_FREQ	-0.238	OFF_SCREEN_POSS	0.253
PNR_HANDLER_FG%	-0.167	LT_5FT_FGM	0.195	OFF_SCREEN_PPP	-0.209	BT5_9FT_FG%	-0.227	BT6_9FT_DFGM	0.237	PLUS_MINUS	-0.228	HANDOFF_POSS	0.246
3PT_FREQ	-0.166	LT6FT_DFGM	0.177	AST	0.206	BT10_15FT_FREQ	-0.204	BT10_15FT_FREQ	0.202	AST_TO	-0.221	OFF_SCREEN_PPP	0.216
PNR_HANDLER_POSS	-0.163	OFF_LOOSE_BALLS	0.172	POTENTIAL_AST	0.203	STL	0.203	FG%	-0.190	TRANSITION_PPP	-0.218	OFF_SCREEN_FG%	0.210
AST_TO_PASS%	-0.161	BT5_9FT_FGM	0.170	OFF_SCREEN_FG%	-0.195	BT10_14FT_FG%	-0.198	BT20_24FT_FGM	0.188	BT15_19FT_FGM	0.211	POST_UP_POSS	-0.202
OREB	0.159	PASSES_MADE	0.168	SPOT_UP_PPP	-0.188	PF	0.192	LT6FT_DFGM	0.183	DEFLECTIONS	-0.190	POST_UP_FG%	-0.199
HANDOFF_FG%	-0.155	USG%	0.155	BT20_24FT_FGM	-0.184	DEF_LOOSE_BALLS	0.192	CONT_SHOTS_2PT	0.170	POST_UP_PPP	0.186	POST_UP_PPP	-0.198
HANDOFF_PPP	-0.153	POST_UP_POSS	0.153	CUT_PPP	-0.180	BT6_9FT_FREQ	-0.182	TRANSITION_PPP	-0.167	TS%	-0.183	TS%	0.192
HANDOFF_POSS	-0.152	BT16FT_3PT_DFGM	0.147	CUT_FG%	-0.178	TS%	-0.179	BT16FT_3PT_FREQ	0.167	POST_UP_POSS	0.176	FG3%	-0.181
CONT_SHOTS_2PT	0.152	PF	0.145	AST_TO_PASS%	0.176	PIE	-0.178	BT25_29FT_FG%	0.163	LT6FT_FREQ	0.171	FG%	0.173
BT25_29FT_FGM	-0.150	ISO_POSS	0.144	OFF_SCREEN_POSS	-0.173	PLUS_MINUS	-0.168	BT25_29FT_FGM	0.161	POST_UP_FG%	0.171	LT_5FT_FG%	0.167
PUTBACK_POSS	0.148	POST_UP_FG%	0.144	SPOT_UP_FG%	-0.166	BT25_29FT_FG%	-0.159	CUT_FG%	-0.160	BT10_15FT_FREQ	-0.169	TRANSITION_PPP	-0.164
PNR_ROLLMAN_POSS	0.147	POST_UP_PPP	0.144	PASSES_MADE	0.165	FG3%	-0.157	CUT_PPP	-0.152	STL	-0.169	TRANSITION_FG%	-0.161
CUT_POSS	0.145	BT6_9FT_DFGM	0.143	AST_TO	0.164	BT10_14FT_FGM	-0.156	3PT_DFGM	0.150	BT6_9FT_DFGM	-0.161	BT25_29FT_FG%	-0.152
BT20_24FT_FGM	-0.145	BT10_15FT_DFGM	0.142	TS%	-0.160	BT20_24FT_FG%	-0.153	LT_5FT_FG%	-0.146	BT10_14FT_FGM	0.154	HANDOFF_PPP	0.131
DREB%	0.143	3PT_DFGM	0.139	FG3%	-0.159	LT6FT_DFGM	0.143	CONT_SHOTS_3PT	0.145	SPOT_UP_FG%	-0.151	BT6_9FT_FREQ	0.129
PNR_ROLLMAN_FG%	0.143	BT10_14FT_FGM	0.138	TM_TOV%	0.154	BT15_19FT_FGM	-0.139	TM_TOV%	-0.144	OFF_SCREEN_POSS	0.150	BT25_29FT_FGM	0.128
SECONDARY_AST	-0.143	PUTBACK_POSS	0.137	TRANSITION_FG%	-0.143	SPOT_UP_POSS	0.131	PUTBACK_FG%	-0.143	DEF_LOOSE_BALLS	-0.144	SPOT_UP_POSS	-0.127

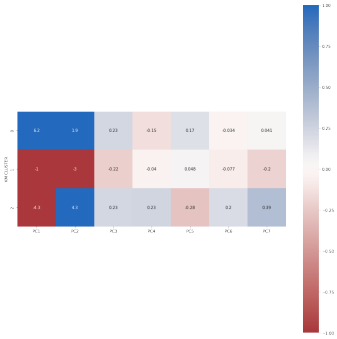

FIG. 3. The Mean values of PCs for each cluster

ing players who provided rim and paint area protection, grabbed rebounds and set screens for teammates to facilitate offence. As this cluster also leaned towards PC2 moderately, these players could have high usage and impact on the game. They also had certain scoring abilities, usually being the rollman of pick-and-rolls, cutting the basketball and putbacks. These traits indicated players who belonged to this group were mostly power forwards and centers.
Notable Players: Rudy Gobert, Clint Capela, Enes Freedom

Cluster 1 - 3D Wingman This cluster only had a positive mean value with PC5 which described tendencies of rewarding players who had good outside scoring ability, especially shooting long distance shots while favoring defensive attributes at the same time. These traits indicated players who belonged to this group were mostly guards or small forwards whose tasks on court were shooting 3-pointers and defending opposite teams' star players.
Notable Players: P.J. Tucker, Danny Green, Robert Covington

Cluster 2 - Elite All-Stars This cluster was heavily biased towards PC2 which described tendencies of rewarding players who impacted the game massively by scoring, organizing offense and defense. These traits indicated players who belonged to this group were superstars with dominant personal performance at both ends, capable of leading a team by themselves.
Notable Players: Lebron James, Kawhi Leonard, Kevin Durant

The performance of K-Means Clustering was not very satisfactory. At first the number of clusters advised was only three, while traditional basketball positions had five. As basketball games evolve, players' classifications should be parallelly diverse. Grouping players in fewer categories than the conventional positioning does not support our research topic. Secondly, although Cluster 1 is positively related to PC5, the mean value was extremely low, approaching zero. Meanwhile, the cluster had negative means with all the other PCs. It was difficult to interpret what kinds of players were in that cluster. Thirdly, Cluster 0 was an ambiguous grouping and could not distinctively separate different types of Big Mans. The accuracy of this model was questioned.

GMM Model

Switching to this probabilistic algorithm, we found the optimal k via the BIC and AIC scores. From Fig x, BIC and AIC scores seemed to be contradictory to each other. As BIC tends to penalize complex dimensionality and AIC presumably fits the data better, we picked number 8 when its AIC score was the lowest and no typical elbow point was spotted. After that, we fit the model with k = 8 and generated Fig 5 to further investigate the correlation between PCs and each cluster.

```
gm = GaussianMixture(n_components=8, n_init
    ↪ =10)
gm_labels = gm.fit_predict(x)
```

From that we observed:

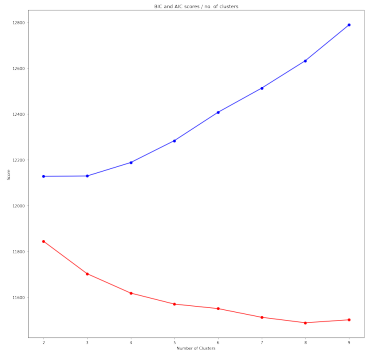

FIG. 4. BIC & AIC scores

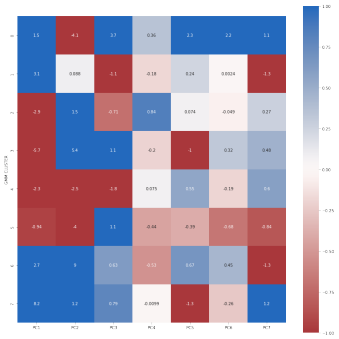

FIG. 5. The Mean values of PCs for each cluster

Cluster 0 – Role Players This cluster had fairly balanced mean values of almost every PC. It exceptionally punished PC2 which rewarded players who scored many points and had high usage. These players usually had a limited amount of play time and were not assigned as the main scorer. These traits indicated players who belonged to this group were role players because they could either do a little bit of everything or specialized in a particular skillset, which explained why the mean values were quite leveled.
Notable Players: Matthew Dellavedova, Al-Farouq Aminu, Otto Porter Jr.

Cluster 1 – Low Usage Big Men This cluster was biased towards PC1 and moderately leaned towards PC5. Players showed tendencies of protecting the rim, the paint area and the rebounds. Some of them might occasionally stretch beyond the 3-point line to help defense or take long-distance shots. Their negative mean values of PC7 and PC3 revealed they might not be able to consistently score or facilitate team offense, which explained their low usage rate as exposed by the low PC2 mean. These traits indicated players who belonged to this group were mostly Big Men reserves who were usually benched and substituted starters to consolidate defense.
Notable Players: Dwight Howard, Mo Bamba, Serge Ibaka

Cluster 2 – Two-Way Perimeter Players / Wingman This cluster was moderately biased towards PC2 and PC4 which altogether described tendencies of rewarding players who could score and play hustle defense at the same time. They had a decent amount of usage and impacted the game by being highly efficient at both ends. These traits indicated players who belonged to this group were usually the startup swingman.
Notable Players: Andrew Wiggins, Jerami Grant, Harrison Barnes

Cluster 3 - Elite All-Stars This cluster was heavily biased towards PC2 and moderately leaned towards PC3 which altogether described tendencies of rewarding players who had very high usage and were efficient in scoring many points through isolation, post-up or mid-range shots. Players also delivered assists, or their passes had high potential converting to assists. These traits indicated players who belonged to this group were the elite offensive players.
Notable Players: Luka Doncic, LeBron James, James Harden

Cluster 4 - Sharpshooters This cluster was moderately biased towards PC7 and PC5 which altogether described tendencies of rewarding players who efficiently scored from long-distance shots after doing handoffs or using off-screens. These traits indicated players who belonged to this group were the 3-point specialists and due to their positions, they had to defend the perimeter.
Notable Players: Doug McDermott, Duncan Robinson, Eric Gordon

Cluster 5 - Playmakers This cluster only had positive mean values with PC3 which described tendencies of rewarding players who made plenty of assists and were good at passing. These traits indicated players who belonged to this group were the floor generals with key responsibilities of playmaking.
Notable Players: Rajon Rondo, Mike James, Tyus Jones

Cluster 6 - Elite Modern Big Men This cluster was heavily biased towards PC2 and PC1. It also had fairly positive mean values with PC5, PC3 and PC6. The statistics described a group of players who had high usage, great impact on the game, excellent scoring abilities through post-up games and isolations while being effective at rebounding and rim protection. They could also make plays and stretch beyond the 3-point line to make long-distance shots. These traits indicated players who belonged to this group were mostly top-tier modern big men with complete and versatile skillsets.
Notable Players: Nikola Jokic, Joel Embiid, Giannis Antetokounmpo

Cluster 7 - High Usage Traditional Big Men This cluster was heavily biased towards PC1. It also had fairly positive mean values with PC2, PC7 and PC3. The statistics described a group of players who were excellent at setting screens, rebounding and providing rim and paint area protection. They had a decent amount of usage and could reliably score via rolling towards the

basket and putbacks. The negative mean value of PC5 disclosed players' incompetence to defend the perimeter or score beyond it. These traits indicated players who belonged to this group were mostly traditional centers with no flexibility to stretch.
Notable Players: Rudy Gobert, Clint Capela, Enes Freedom

The performance of GMM was satisfactory. The number of clusters was reasonable. Each cluster was distinguishable, and their features were aligned with the dynamic of modern basketball games. The improved performance could possibly be related to the characteristics of GMM's algorithm which used soft assignment and allowed weighting on distributions. It was more favorable to our dataset as players' metrics were diverse and unevenly distributed. Hence, GMM yielded a significant better result.

VI. CONCLUSION

Through applying PCA and GMM, we successfully reclassified NBA players into 8 new clusters. The hardest part of this project would be pre-processing dataset and interpreting PCs. We could improve this project by setting up a more efficient data pipeline. In future, we could also include players metrics in different seasons such that we would be able to keep track of their playstyles. We hope this research could benefit both basketball fans and sports analysts. For fans it helps the general public to understand the current trend of NBA games better. For the specialists, this project could possibly provide insights on team composition. For example, data scientists might compare the clusters with the composition of the past champion squads to investigate what kinds of players a title-chasing team seeks the most and then advise coaches to take the right action in draft picks and trades.

VII. REFERENCE

[1] Constantinou, A. (2022). Unsupervised Learning. *School of Electronic Engineering and Computer Science, Queen Mary University of London.*

[2] Saji, B. (2021). In-depth Intuition of K-Means Clustering Algorithm in Machine Learning. *Data Science Blogathon 4.* https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

[3] Abhigyan. (2020). Importance of Dimensionality Reduction. https://medium.com/analytics-vidhya/importance-of-dimensionality-reduction-d6a4c7289b92

[4] Kalman, S., Bosch, J. (2020). NBA Lineup Analysis on Clustered Player Tendencies:A new approach to the positions of basketball modeling lineup efficiency of soft lineup aggregates. *14th Annual MIT Sloan Sports Analytics Conference.*

[5] Kubara, K. (2020). Gaussian Mixture Models vs K-Means. Which One to Choose?. https://towardsdatascience.com/gaussian-mixture-models-vs-k-means-which-one-to-choose-62f2736025f0

[6] Sarlis, V., Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems, 93, 101562.* https://doi.org/10.1016/j.is.2020.101562

[7] Sinaga, K. P., Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access, 8,* 80716–80727. https://doi.org/10.1109/access.2020.2988796

[8] Bloomberg Quicktake. (2019). The NBA Data Scientist. https://www.youtube.com/watch?v=MpLHMKTolVwt=294s.

[9] Carrasco, O. C. (2019). Gaussian Mixture Models Explained. https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

[10] Maklin, C. (2019). Gaussian Mixture Models Clustering Algorithm Explained. https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

[11] Tripathi, A. (2019). A Complete Guide to Principal Component Analysis — PCA in Machine Learning. https://towardsdatascience.com/a-complete-guide-to-principal-component-analysis-pca-in-machine-learning-664f34fc3e5a

[12] StatQuest. (2018). StatQuest: Principal Component Analysis (PCA), Step-by-Step. https://www.youtube.com/watch?v=FgakZw6K1QQt=512s

[13] Seward, N. (2016). Defensive Player Classification in the National Basketball Association. *arXiv preprint arXiv:1612.05502.*

[14] Trevino, A. (2016). Introduction to K-means Clustering. *Oracle AI Data Science Blog.* https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering

[15] Zhang, L., Lu, F., Liu, A., Guo, P., Liu, C. (2016). Application of K-means clustering algorithm for classification of NBA guards. *International Journal of Science and Engineering Applications, 5(1), 1-6.*

[16] Google. k-Means Advantages and Disadvantages. *Clustering in Machine Learning.* https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages