Basketball Players Analysis

Background Information

Basketball is a team sport in which two teams, most commonly of five players each, opposing one another on a rectangular court, compete with the primary objective of shooting a basketball through the defender's hoop, while preventing the opposing team from shooting through their own hoop. (Wikipedia)

Objectives

The main objective of this analysis is to visualise the capabilities and similarities of each player using graphs, and to use clustering methods to group players into differentiable categories. The target stakeholders of this analysis are:

- Basketball enthusiasts who wish to gain a deeper understanding of the international basketball landscape.
- Basketball officials who wish to know and visualise the capabilities of each player.
- Basketball team owners who want to gain insight into which players they should draft and potentially purchase.

Potential Hindrances to the analysis

While the dataset used in this analysis is comprehensive, it only contains information on 100 NBA (National Basketball Association) players, while there are approximately 500 total players in the NBA. This may give us a limited view into the distribution of players.

Dataset

The dataset used for this analysis is called Social Power NBA and can be found on Kaggle. In its primary form it contains 100 rows for 100 players, and 63 attributes for each player (63 columns).

Feature	Description
PLAYER_NAME	Player's name.
AGE	Player age.
GP	Games played.
W	Games played where the team won.
L	Games played where the team lost.
W_PCT	Percentage of games played won.
MIN	Minutes played.
OFF_RATING	Player offensive rating
DEF_RATING	Player defensive rating
NET_RATING	Average of the offensive/defensive rating.
AST_PCT	Assist percentage.
AST_TO	Assists-to-turnovers.
AST_RATIO	Assists-to-turnovers ratio.
OREB_PCT	Offensive rebounds.
DREB_PCT	Defensive rebounds.
REB_PCT	Total rebounds.
TM_TOV_PCT	Team turnover rate.
EFG_PCT	Effective field goal percentage.
TS_PCT	True Shooting Percentage.
USG_PCT	Usage percentage, an estimate of how often a player makes team plays.
PACE	Pace factor, an estimate of the number of possessions.
PIE	Player impact factor, a statistic roughly measuring a player's impact on the games that they play that's used by `nba.com`.
FGM	Field goals made.

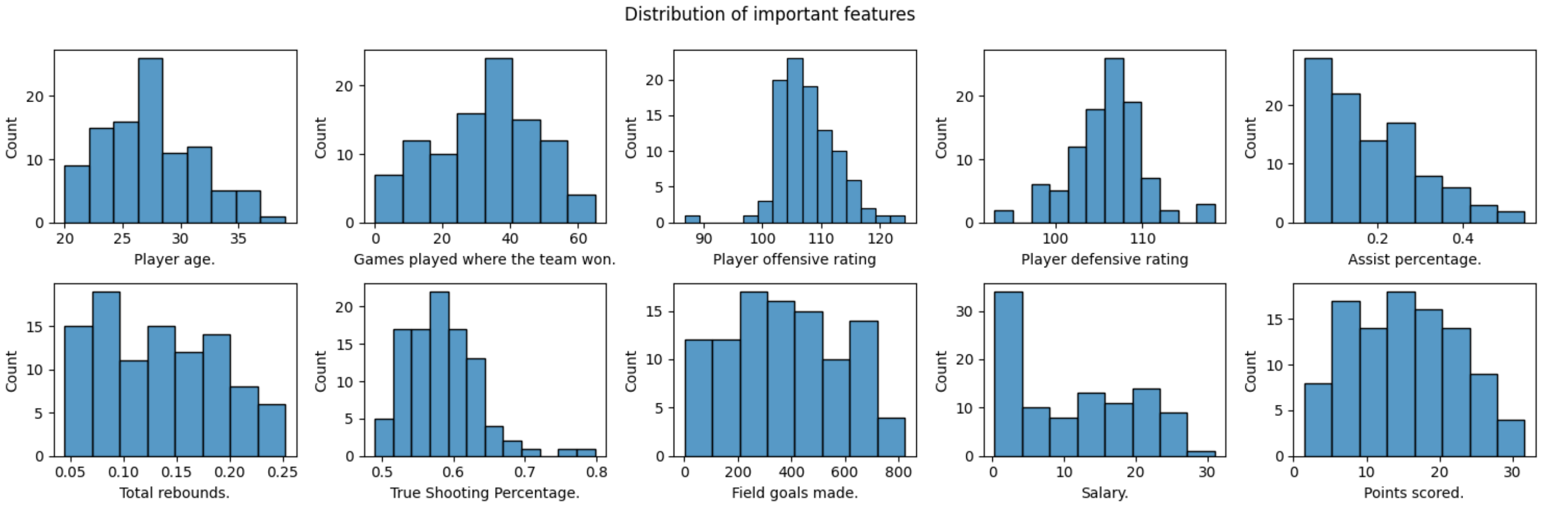
FGA	Field goals attempted.
FGM_PG	Field goals made percentage.
FGA_PG	Field goals attempted percentage.
FG_PCT	Field goals total percentage.
GP_RANK	Games played, league rank.
W_RANK	Wins, league rank.
L_RANK	Losses, league rank.
W_PCT_RANK	Win percentage, league rank.
MIN_RANK	Minutes played, league rank.
OFF_RATING_RANK	Offensive rating, league rank.
DEF_RATING_RANK	Defensive rating, league rank.
NET_RATING_RANK	Net rating, league rank.
AST_PCT_RANK	Assists percentage, league rank.
AST_TO_RANK	Assists-to-turnovers, league rank.
AST_RATIO_RANK	Assist ratio, league rank.
OREB_PCT_RANK	Offensive rebounds percentage, league rank.
DREB_PCT_RANK	Defensive rebounds percentage, league rank.
REB_PCT_RANK	Rebounds percentage, league rank.
TM_TOV_PCT_RANK	Team turnover, league rank.
EFG_PCT_RANK	Effective field goal percentage, league rank.
TS_PCT_RANK	True shooting percentage, league rank.

USG_PCT_RANK	Usage percentage, league rank.
PACE_RANK	Pace score, league rank.
PIE_RANK	Player impact, league rank.
FGM_RANK	Field goals made, league rank.
FGA_RANK	Field goals attempted, league rank.
FGM_PG_RANK	Field goals made percentage, league rank.
FGA_PG_RANK	Field goal attempted percentage, league rank.
FG_PCT_RANK	Field goal percentage, league rank.
SALARY_MILLIONS	Salary.
PTS	Points scored.
TWITTER_FOLLOWER_COUNT_MILLIONS	Number of Twitter followers.

As you can see this is an exhaustive dataset with a wide range of features covering all aspects of playstyle, rankings, and other statistics.

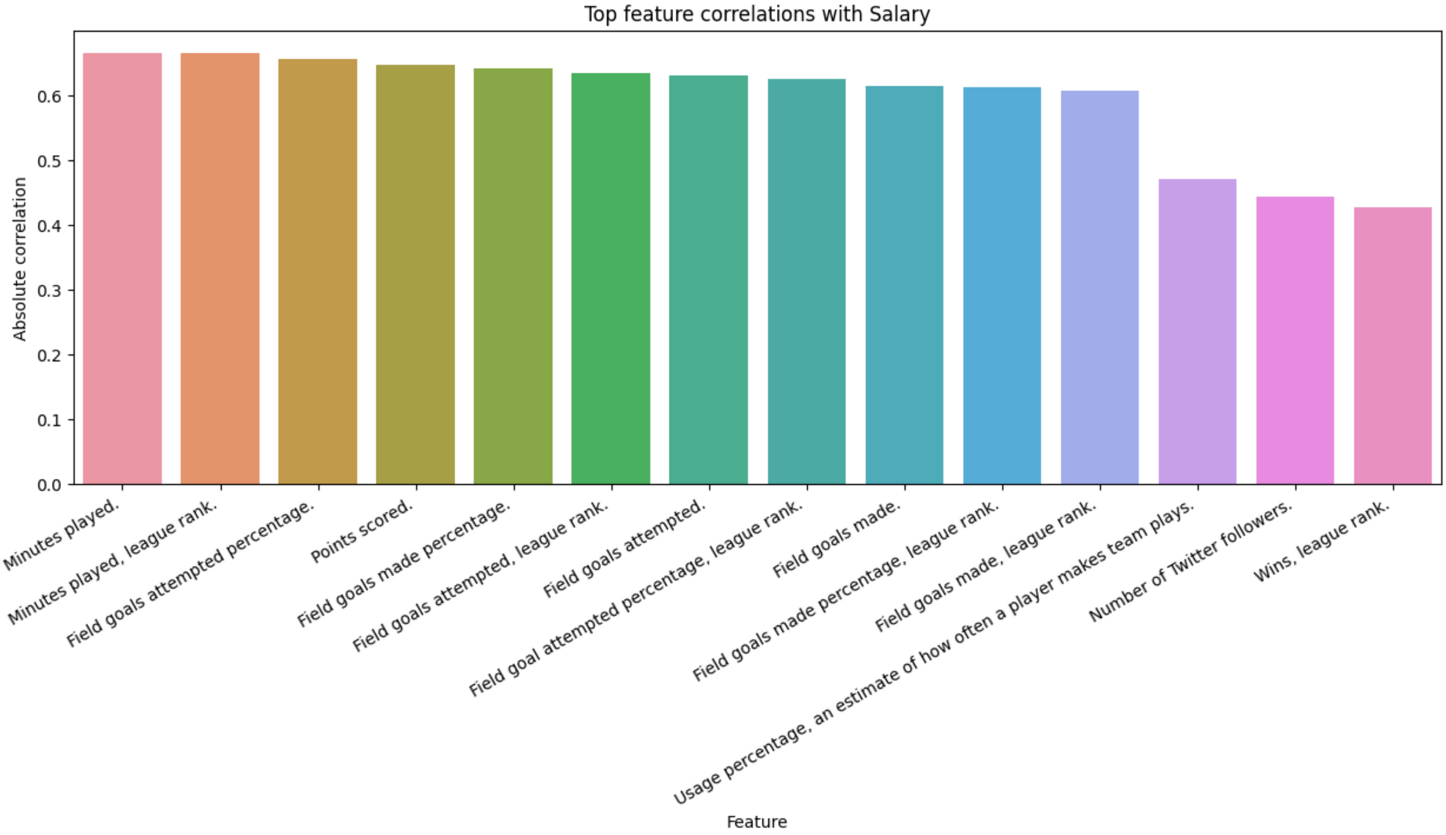
Feature distributions

Since it's not feasible to show a distribution of all 63 features, I have handpicked the features which give the most insight into the capability of a player.



There are evidently outliers in this dataset, as can be observed in *Player offensive rating*, *True Shooting Percentage*, and *Salary*. This indicates that some players have significantly higher capabilities than other players. We would expect these players to be grouped into their own cluster.

Relationships between features and Salary

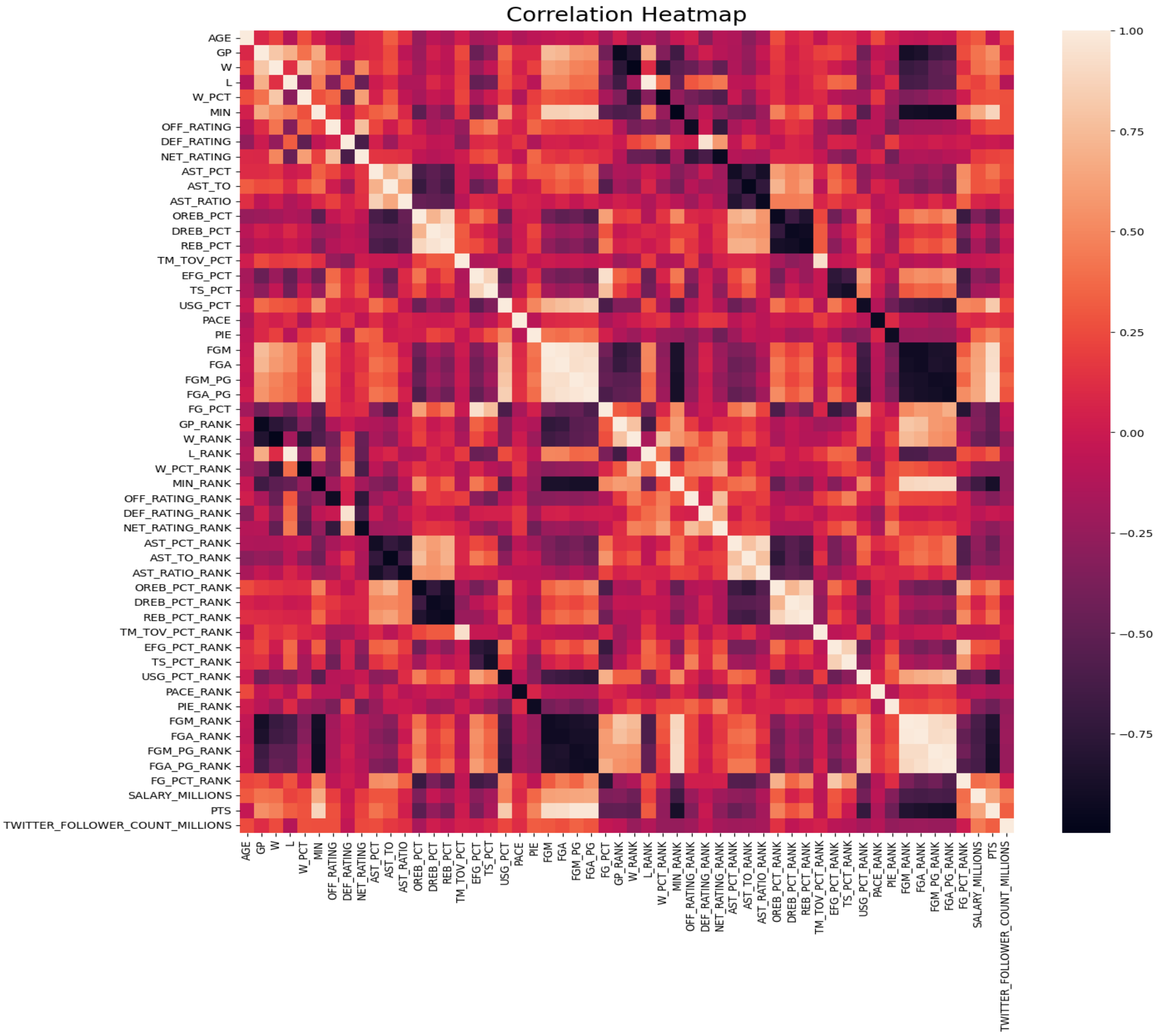


The above chart details the features which are most correlated with players' *Salary*.

Importantly, features such as *Minutes played* and *Points scored* are highly correlated with *Salary,* indicating the more time played and the more points scored, the higher the player's salary will be.

Correlation heatmap

The heatmap below shows the correlations between each feature.



Important: this dataset is divided into 2 main categories: ratings and ranks.

The ratings (first half of the dataset) refer to individual scores for each player, such as Offensive Rating and Defensive Rating.

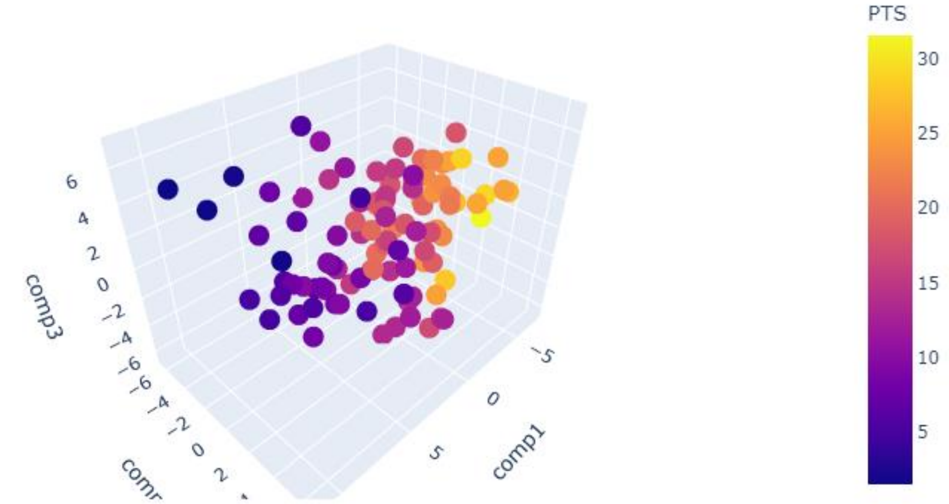
The ranks (second half of the dataset) refer to relative scores for each player, such as Offensive Rating Rank and Defensive Rating Rank.

Thus, a player with a higher Offensive Rating will have a lower Offensive Rating Rank.

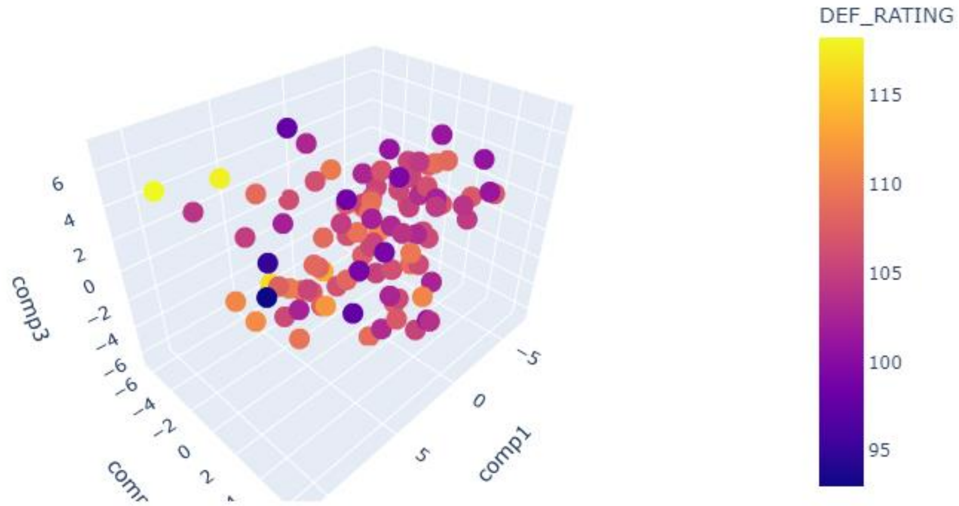
PCA Visualisation

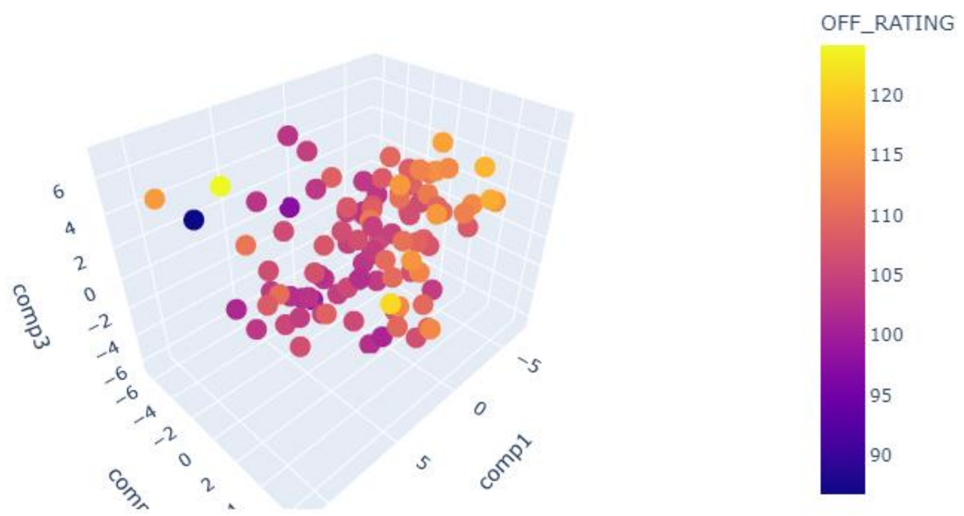
Using PCA, I reduced the dimensionality of the dataset from 63 features to 3 features, so we can visualise various features of the dataset in a 3-dimensional scatter plot.

Visualisation of Players: Points scored.

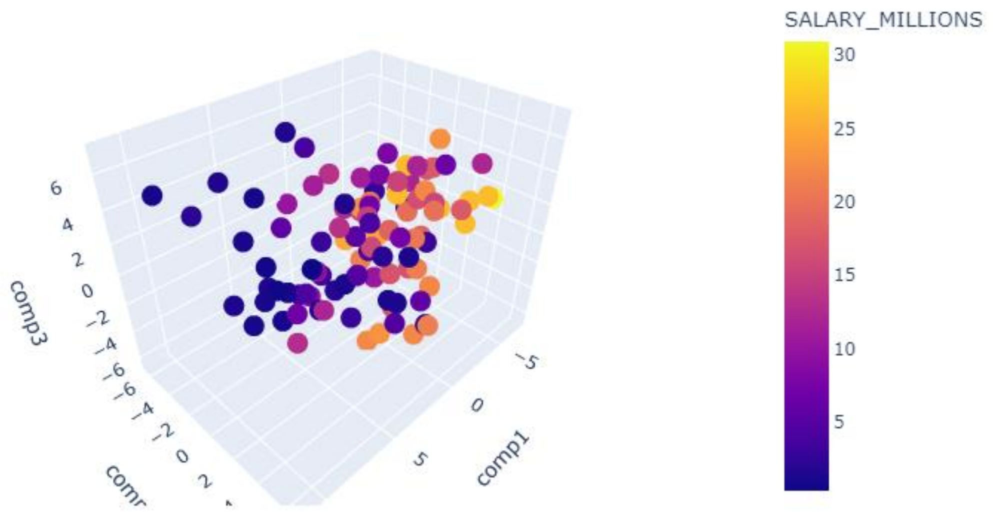


Visualisation of Players: Player defensive rating





Visualisation of Players: Player offensive rating



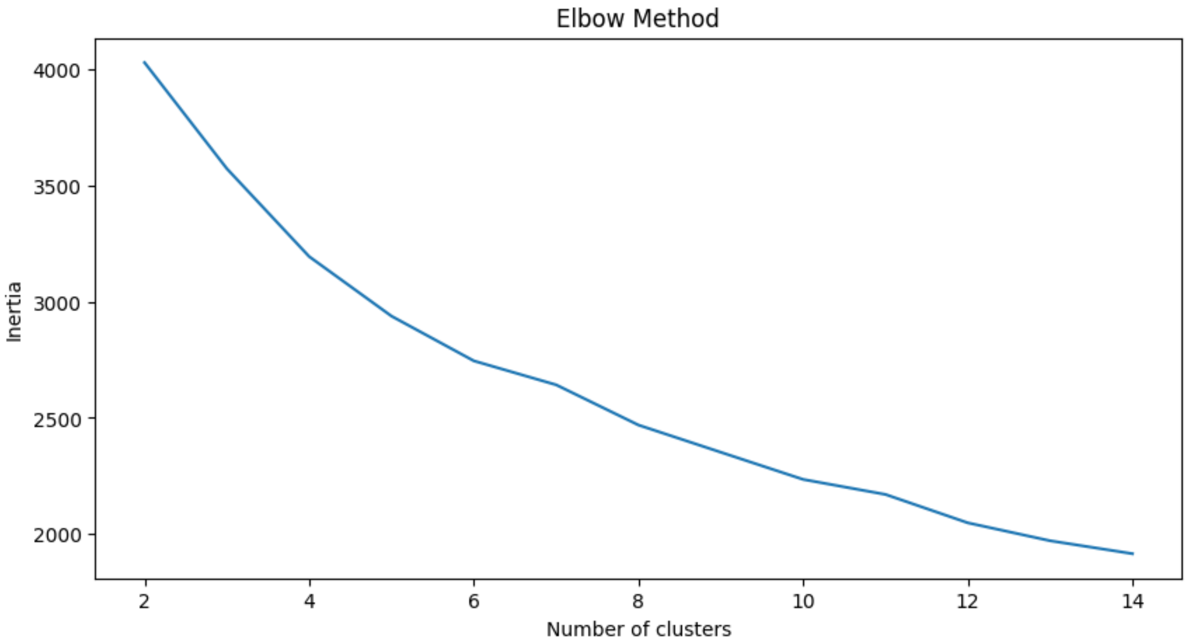
Visualisation of Players: Salary.

There is a clear trend in the dataset as some players seem to possess superior capabilities compared to the rest. Specifically, as *comp1* decreases, *Salary, Offensive Rating,* and *Points Scored* generally increase, demonstrating a clear trend in the dataset.

Specifically in *Salary,* there is clear distinction between high-earners and low-earners in the NBA. We would expect this distinction to be demonstrated in the clusters formed later.

Clustering: K-Means Elbow Method

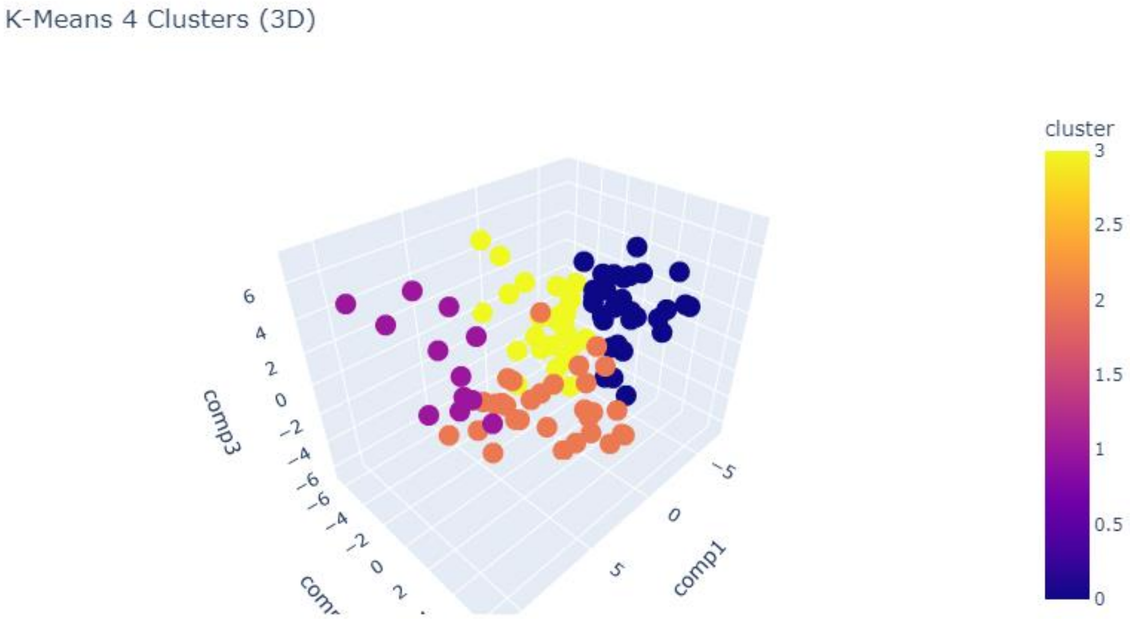
The elbow-method is non-definitive way of determining the optimal number of clusters to use for the K-Means clustering algorithm.



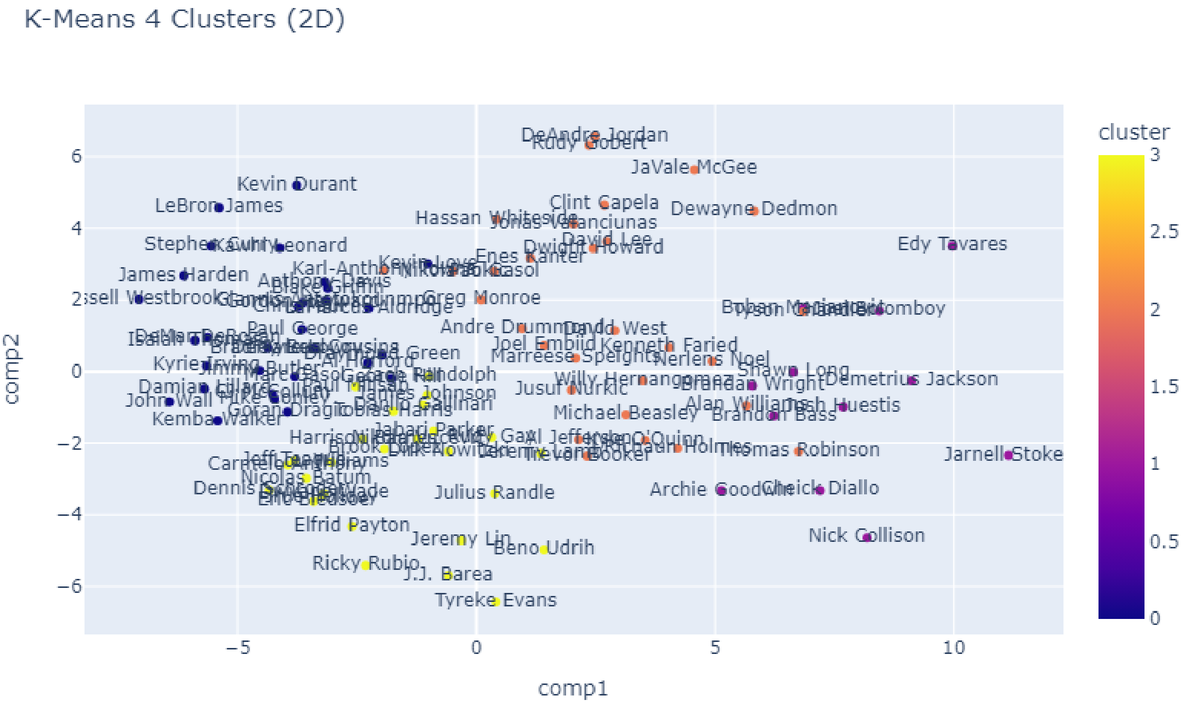
For this analysis, let’s use 4 clusters and 6 clusters.

Clustering: K-Means (4 Clusters)

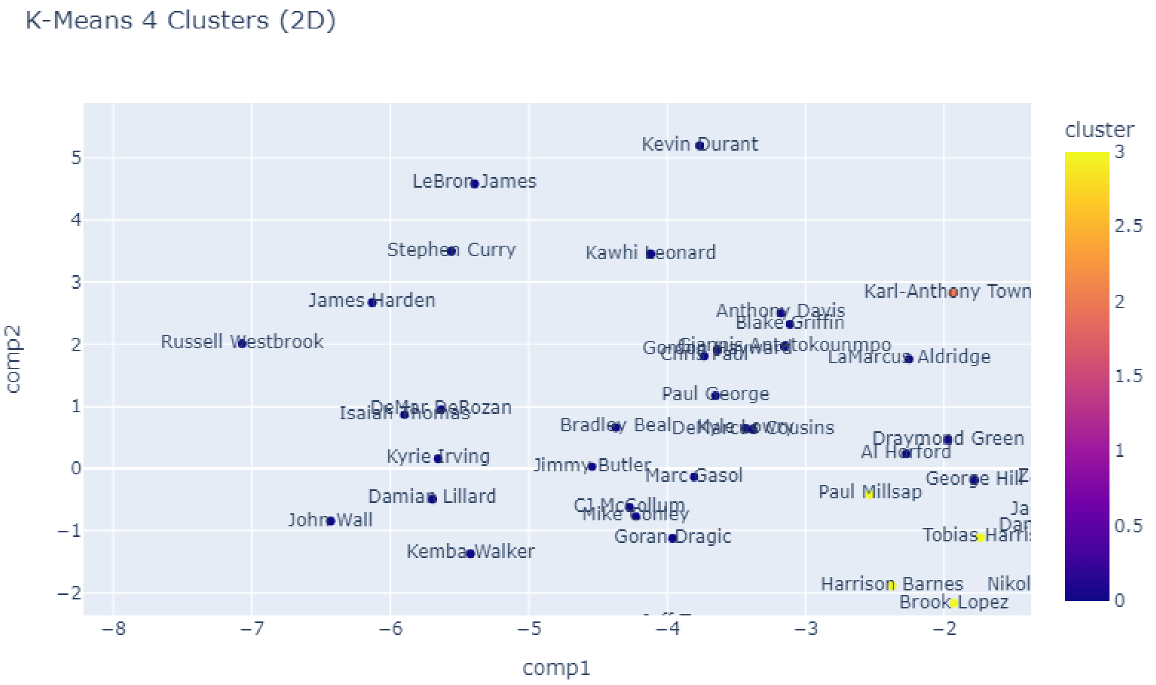
Using K-Means and PCA, we can visualise the clusters in 3-dimensional and 2-dimensional graphs.



The 3-dimensional graph gives us a pleasant overview of the clusters. However, it would be nice to see the names of the players so that we can understand how different players were organised and clustered. Unfortunately plotting the names on the 3-d graph makes it indecipherable, thus a 2-d plot must be used instead.



Let’s zoom in on the **blue** cluster.

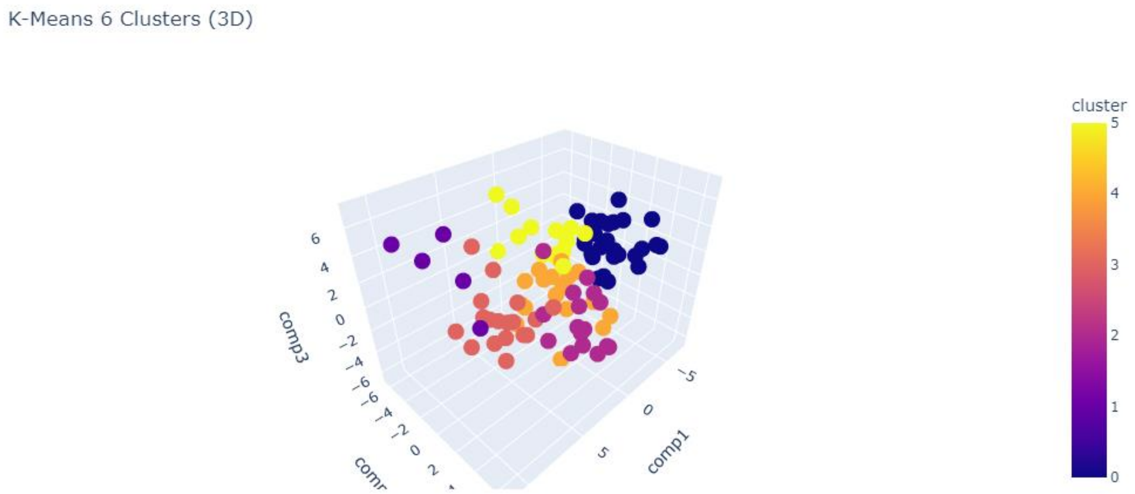


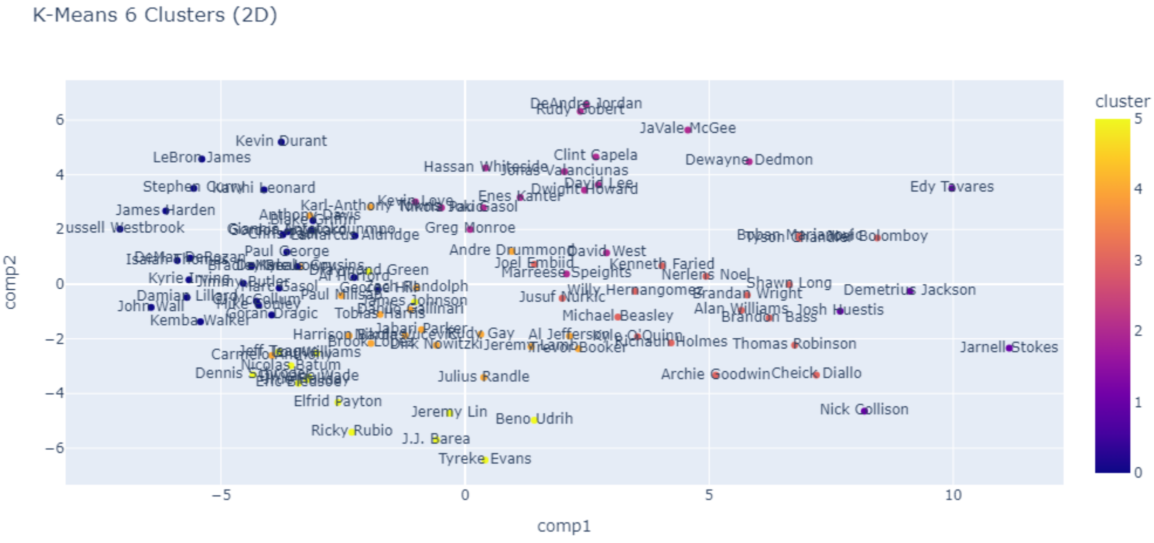
We can see that top players such as LeBron James, Keven Durant, Stephen Curry, James Harden, Russel Westbrook, and Kyrie Irving have all been assigned to the same cluster.

This suggests 2 things:

- PCA was successful in maintaining the relative distances between each player, as strong players are close to each other.
- K-Means was able to identify the strongest players and put them into a cluster. This probably goes for the weaker players as well.

Clustering: K-Means (6 Clusters)

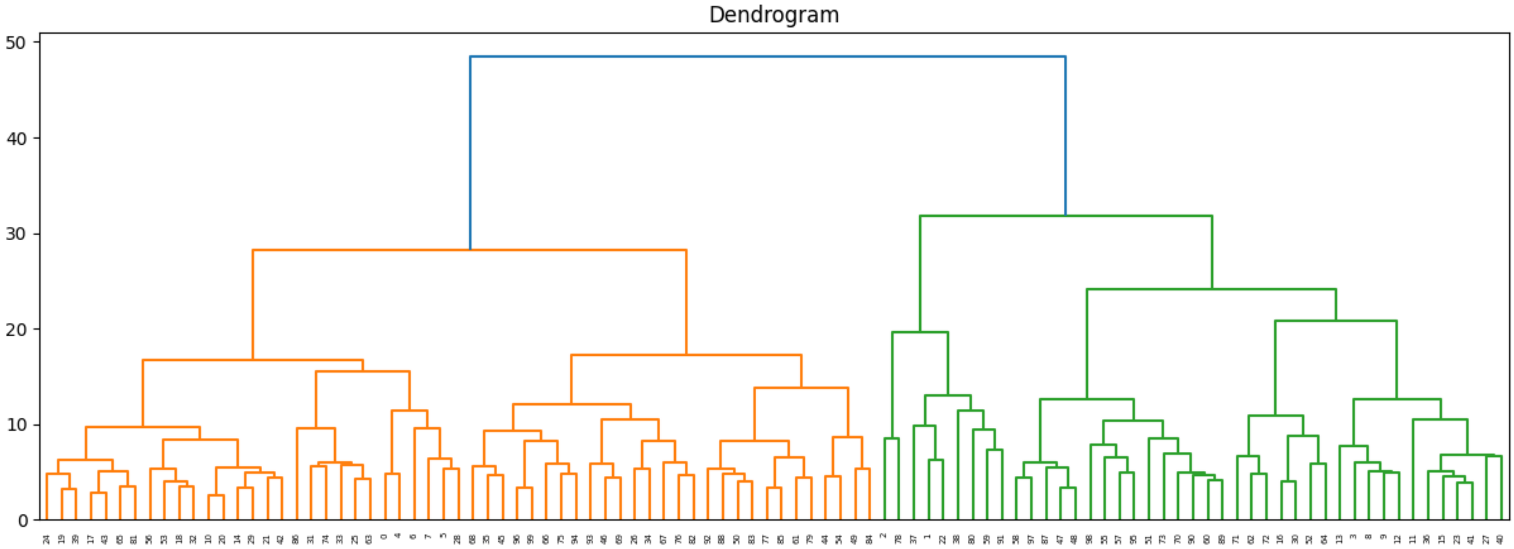




Hierarchical Agglomerative Clustering: Dendrogram

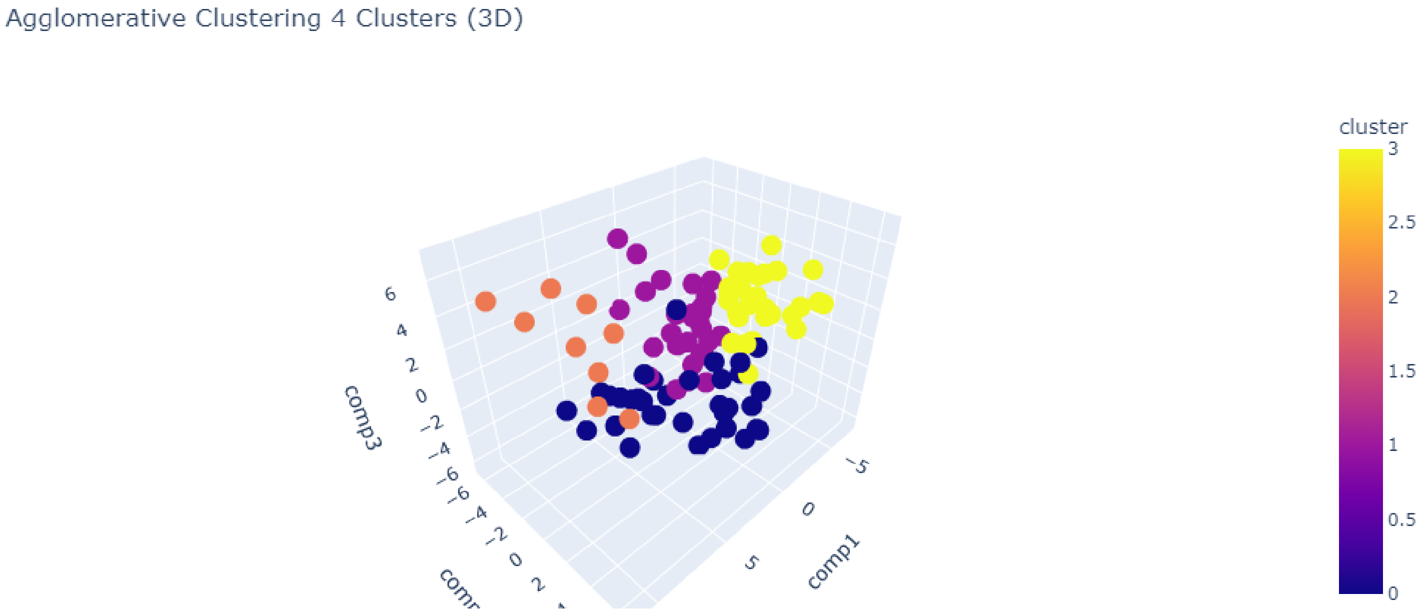
Hierarchical Agglomerative Clustering is a clustering technique which iteratively builds up clusters based on a similarity metric.

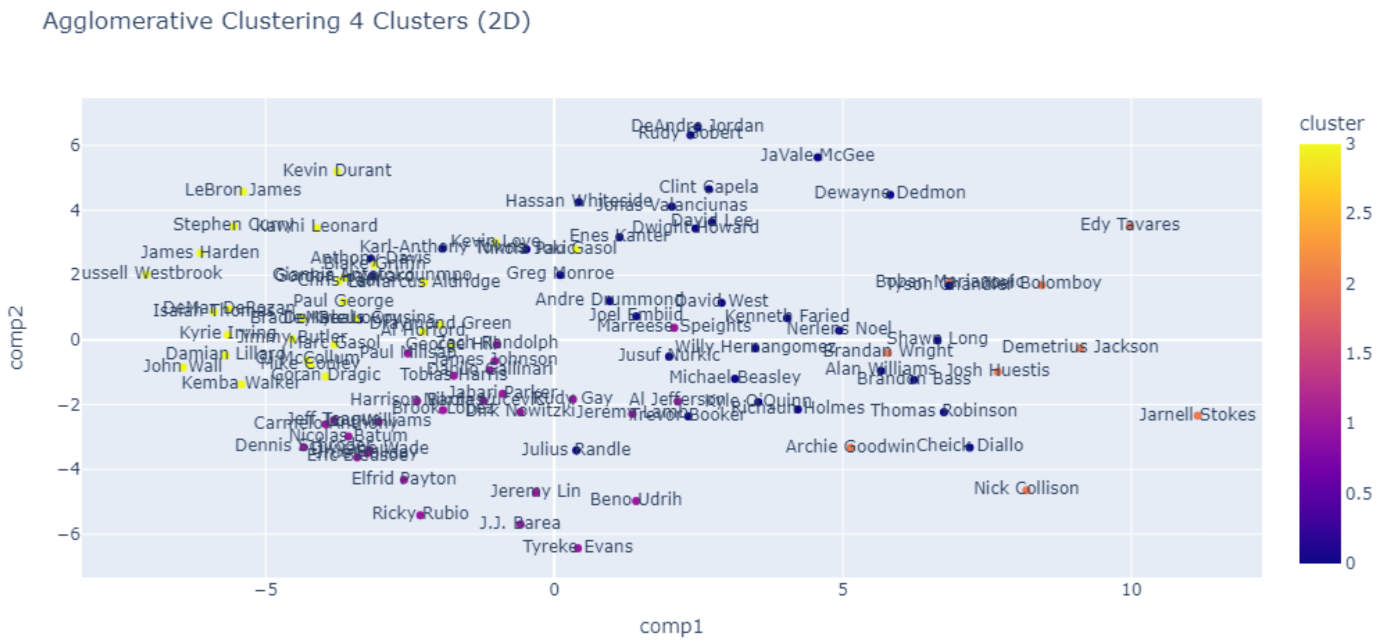
Using a dendrogram, we can visualise the clusters created by the HAC method. For this example, the *ward* linkage was used.



Based on this diagram, there seems to be 4 main clusters (this is a subjective conclusion).

Hierarchical Agglomerative Clustering: 4 Clusters





We can observe that HAC produced similar clusters to K-Means.

Comparison of clustering methods

In this analysis we used 3 clustering methods:

1. K-Means with 4 clusters
2. K-Means with 6 clusters
3. Hierarchical Agglomerative Clustering with 4 clusters

Advantages of K-Means:

- K-Means is fast.
- K-Means tends to find even-sized clusters.

Disadvantages of K-Means:

- Have to try different k-values to find the optimal number of clusters.
- Bad with non-standard clusters shapes

Advantages of HAC:

- Produces a full hierarchy tree, which is useful for interpretation.
- Can find uneven cluster sizes.

Disadvantages of HAC:

- Have to try different k-values to find the optimal number of clusters.
- There are a lot of distance metric and linkage options.
- Can be slow to calculate (complexity is proportional to the squared number of observations).

Final recommendation

Each clustering method provided valuable insight into the similarities and groupings of various NBA basketball players.

Since K-Means is fast and produced similar results to HAC, K-Means with 4 clusters is recommended as it effectively segmented strong players from weaker ones.

Summary of Key Findings:

- Objective: The main objective of the analysis was to visualize player capabilities and similarities using graphs and clustering methods. The target stakeholders included basketball enthusiasts, officials, and team owners looking for insights into player performance.
- Feature Distributions: Key features such as minutes played and points scored were highly correlated with player salaries, indicating that more playing time and higher scoring ability led to higher salaries.
- PCA Visualization: Principal Component Analysis (PCA) was used to reduce dimensionality, allowing for a 3-dimensional scatter plot that showed a clear trend. Some players had superior capabilities, especially in terms of salary, offensive rating, and points scored.
- Clustering with K-Means: K-Means clustering was applied with both 4 and 6 clusters. The 3-dimensional plot provided an overview of the clusters, but a 2-dimensional plot was necessary for player identification. Notably, top players like LeBron James, Kevin Durant, Stephen Curry, and others were grouped together, indicating successful clustering.
- Hierarchical Agglomerative Clustering (HAC): HAC with a dendrogram revealed four main clusters, which aligned with the results obtained from K-Means clustering.
- Comparison of Clustering Methods: K-Means was favoured for its speed and tendency to find even-sized clusters, while HAC offered interpretability and the ability to find uneven cluster sizes. Both methods produced similar results, but K-Means with 4 clusters was recommended due to its speed and effectiveness.

Possible Improvements

- Data Quantity: Expand the dataset to include a larger sample of NBA players, as 100 players out of approximately 500 may not fully represent the player population.
- Feature Engineering: Explore the possibility of creating new features that could capture player performance more effectively, such as advanced player statistics or performance in specific game situations (e.g., clutch moments).
- Dimensionality Reduction: Experiment with different dimensionality reduction techniques besides PCA to capture more nuanced relationships between features.
- Validation: Validate the results of clustering by comparing them to external benchmarks or expert opinions to ensure the clusters have real-world significance.
- Model Selection: Consider other clustering algorithms besides K-Means and hierarchical clustering, such as DBSCAN or Gaussian Mixture Models, to identify the most appropriate method for the data.
- Predictive Modelling: Extend the analysis to include predictive modelling, such as predicting player performance in future seasons or estimating player salaries based on historical data and performance metrics.

Conclusion

In conclusion, this analysis provided valuable insights into player capabilities and groupings, aiding basketball enthusiasts, officials, and team owners in understanding the landscape of NBA players and making informed decisions related to drafting and player acquisition.