Reclassifying NBA Players using Clustering

By: Christopher Le

UNDERSTANDING THE PROBLEM

- **Fantasy NBA Basketball**
  - 13+ million players in 2022[1]
  - $3.8 billion industry in 2022[2]
  - Fantasy teams dependent on NBA player statistical production

- **Traditional NBA Positions:**

  Point Guard, Shooting Guard, Small Forward, Power Forward, Center
  - NBA players no longer can be defined by the positions they play in.
  - Not representative of player's on-the-court statistical production
  - Need new way to group players into certain classifications beyond positioning

- **Goal:** use unsupervised learning clustering to create tiers to classify NBA players that better represents how they contribute to the team on the court.

DATASET

The NBA player stats dataset was retrieved from [Kaggle](). This dataset contains 2021-2022 regular season NBA player stats per game.

Data Wrangling

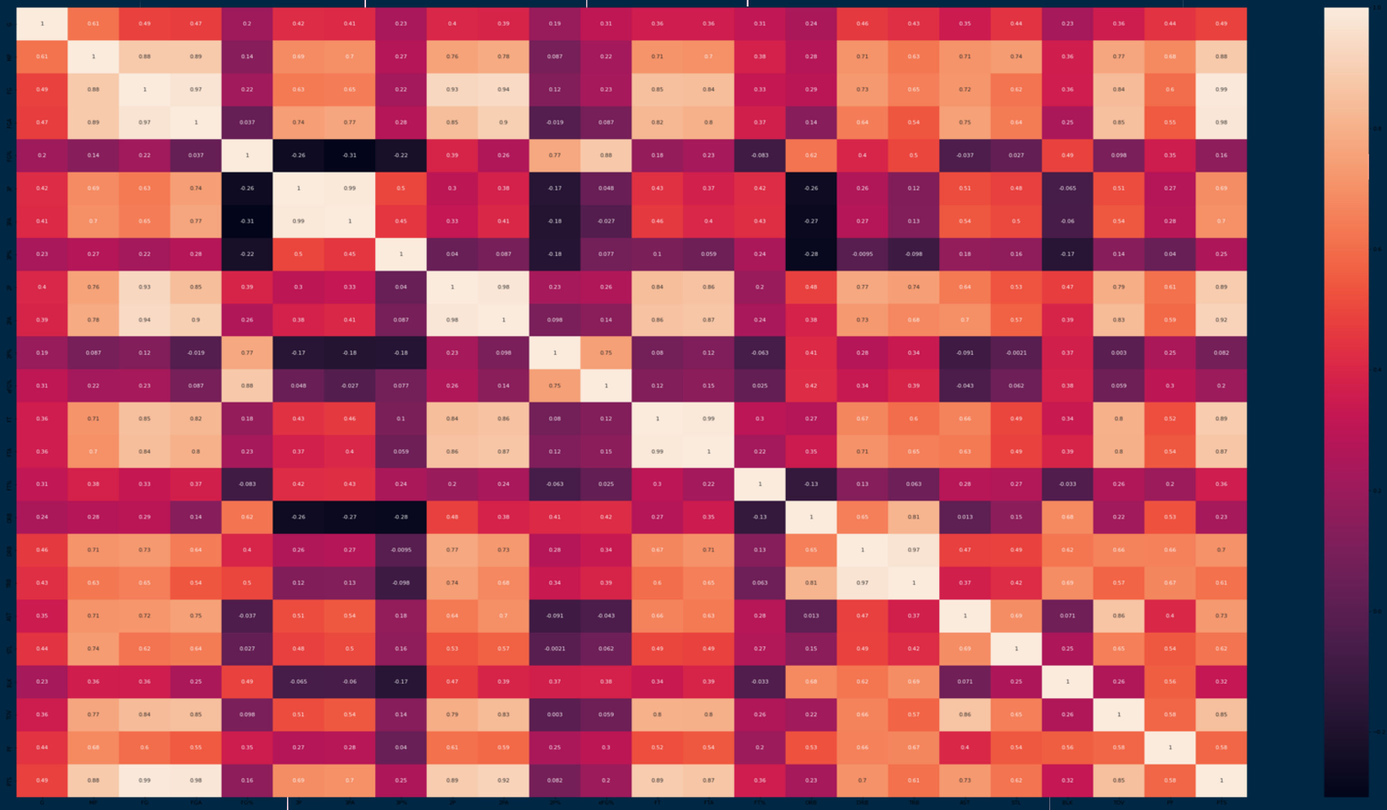
01

DATA WRANGLING

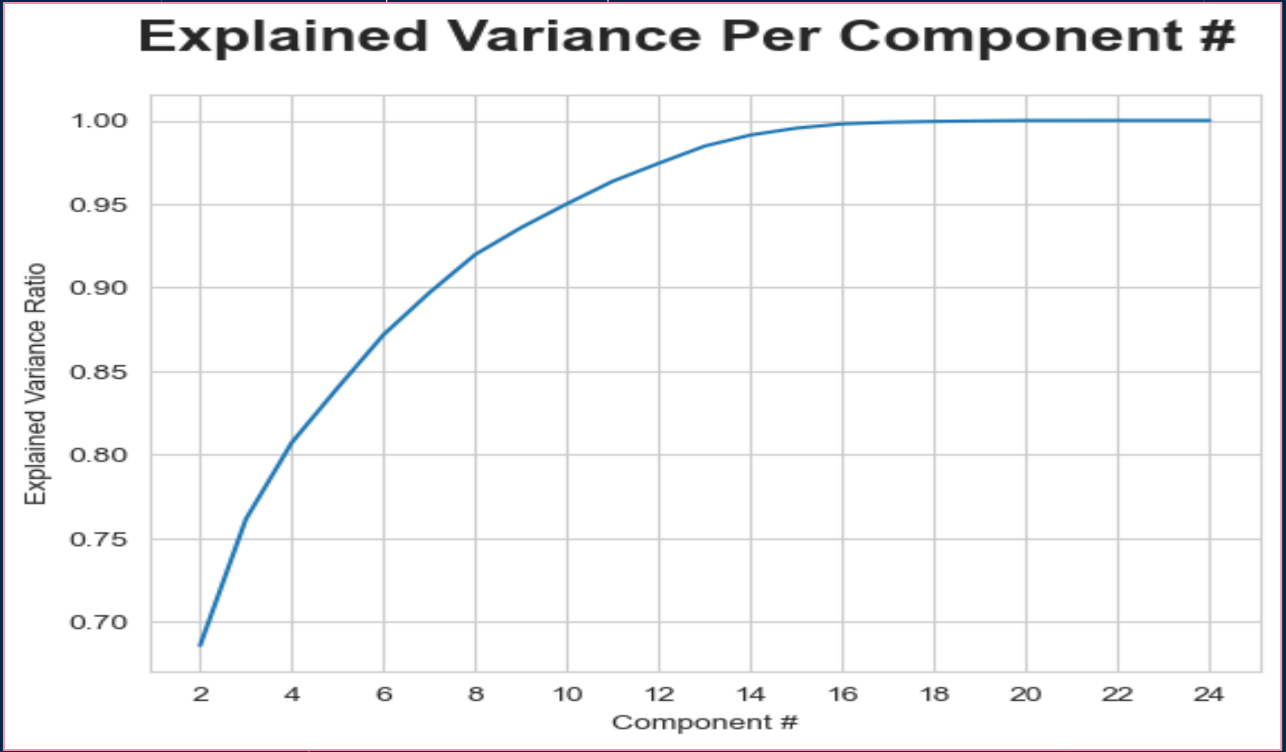
- **Initial Shape:** 812 rows x 30 columns
- **Issue 1** - Duplicate Player Entries
  - Dropped 207 rows contained names for players who changed teams
- **Issue 2** – Columns Unrelated to Game Statistics
  - Dropped 4 columns: rank (RK), age (AGE), team (Tm), & games started (GS)
- **Issue 3** – Inconsequential NBA Players
  - Filtered out players playing < 10 games.
  - Dropped 105 entries
- **Final shape:** 500 rows x 26 columns

Exploratory Data Analysis

02



A heatmap representing the correlations between different features.
Many features were highly correlated including: "FG"/"FGA", "3P"/"3PA",
"2P"/"2PA", "FT"/"FTA", and "DRB"/"TRB".

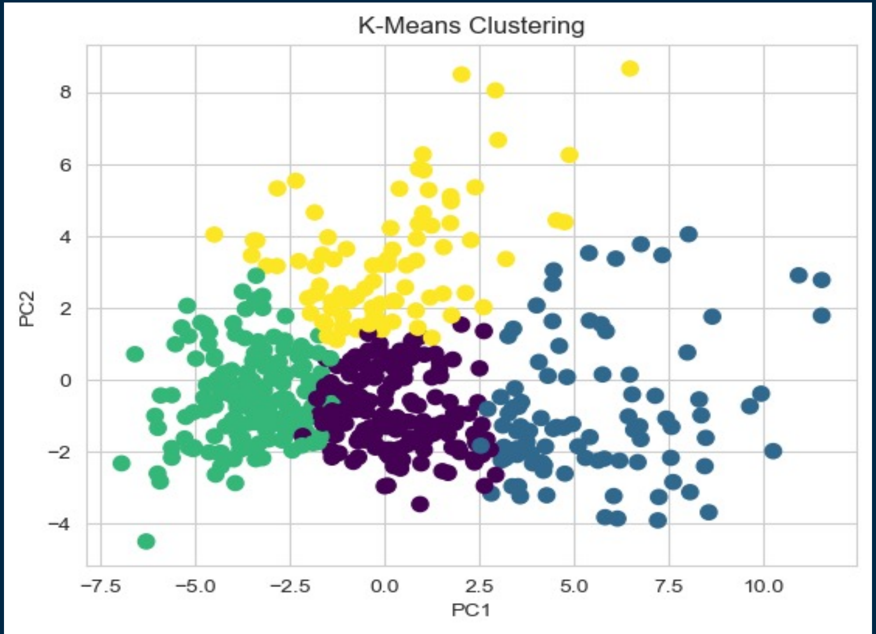


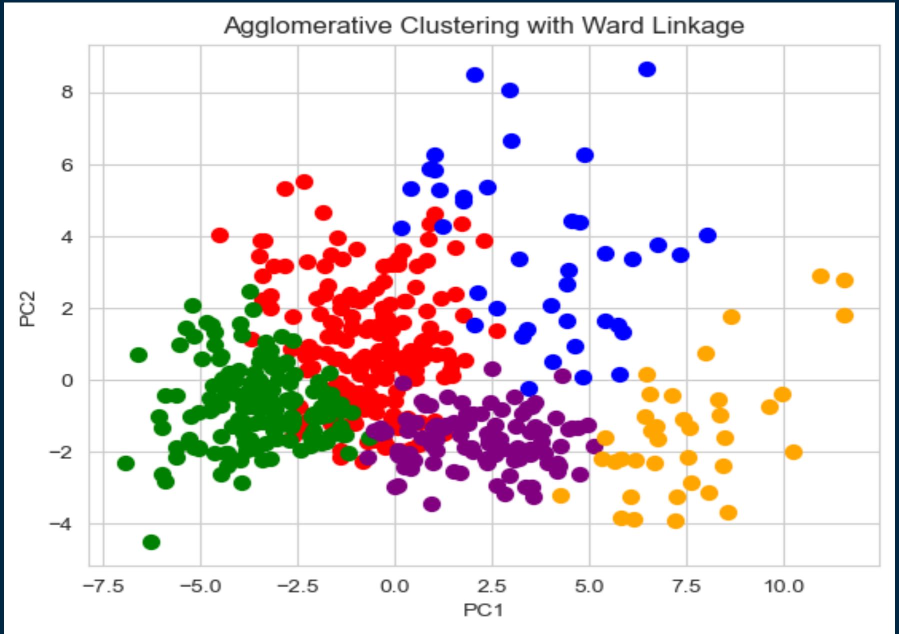
The first 5 components account for 84% of the variance and the first 10 components for 95% of the variance

Modeling

03

CLUSTERING MODELS

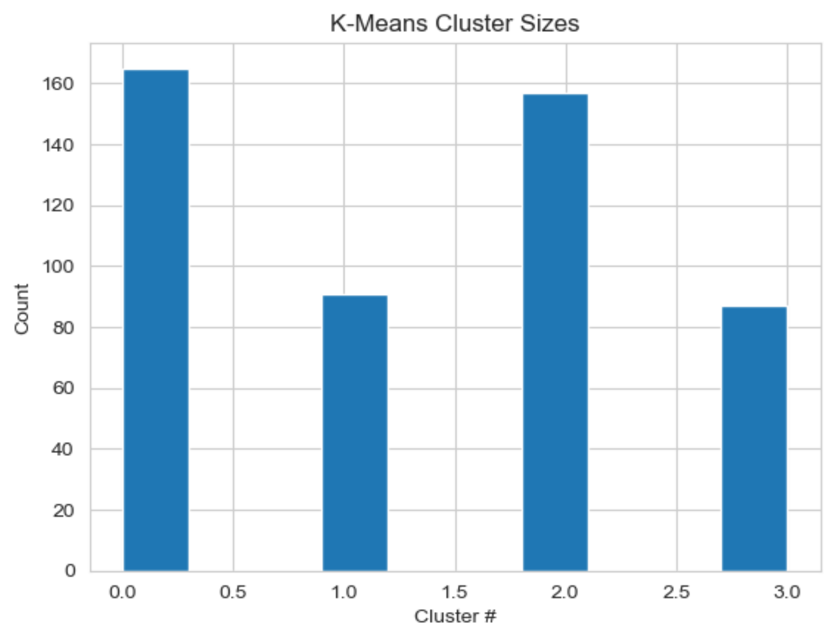


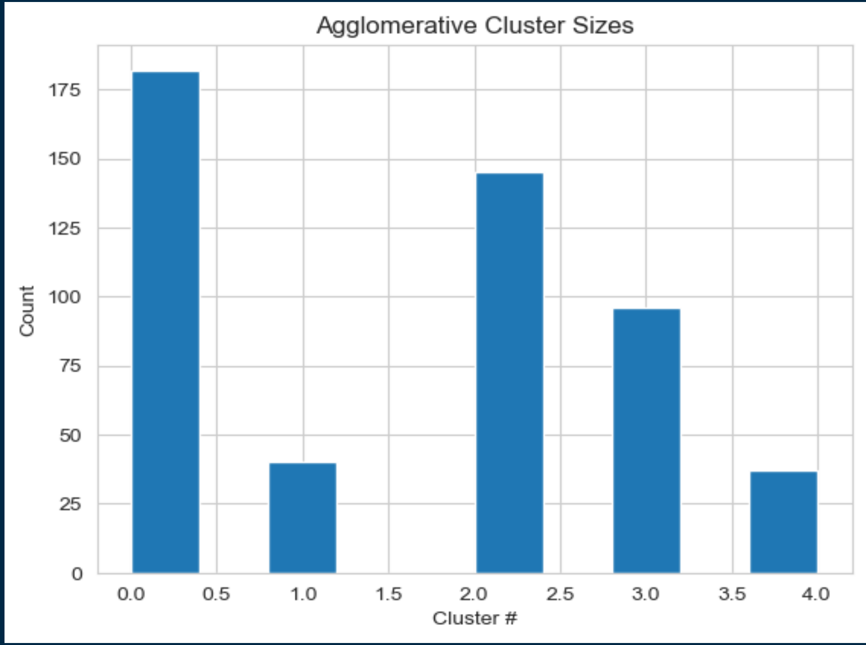


- K-Means Clustering (k=4)
- Optimal K based on: Silhouette Method
- Average silhouette scores: 0.222

- Agglomerative Hierarchal Clustering with Ward linkage (k=5).
- Optimal K based on: Dendrogram
- Average silhouette scores: 0.184

CLUSTER SIZES

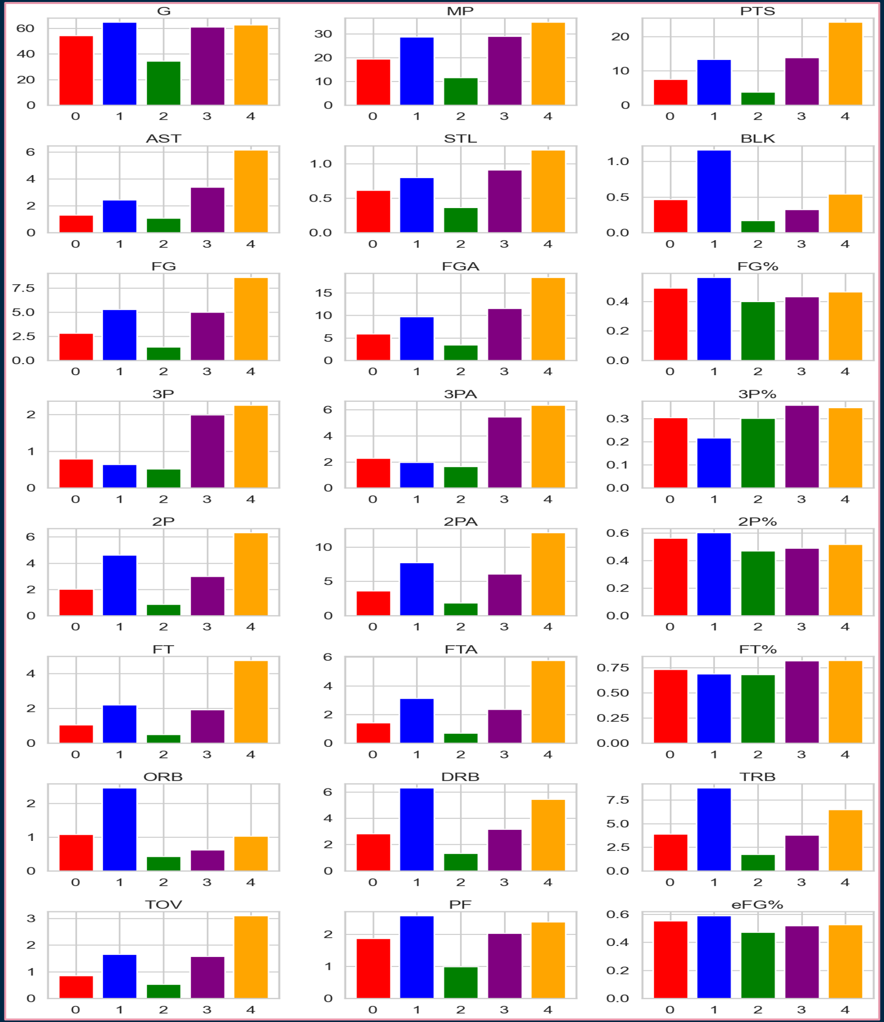




- Comparison of the cluster size distributions.
- **Final Model Selection:** Agglomerative Clustering

CLUSTERING ANALYSIS

- Cluster 0 (Red): Role Players
- Cluster 1 (Blue): Traditional Big Men
- Cluster 2 (Green): Bench Players
- Cluster 3 (Purple): 3-and D- Players
- Cluster 4 (Orange): Versatile All-Around Players



TAKEAWAYS

- Through Agglomerative Hierarchal Clustering, I was able to classify players into five different tiers that had a blend of players from traditional positions.
- Using these redefined clusters, teams can now assemble rosters with NBA players based on their on-the-court statistical contributions rather than what positions they played.
- Tiers provide an efficient way for NBA fantasy team managers to shore up their teams by picking players based on their statistical category needs.

FUTURE RESEARCH

- The dataset used for these models only contained data for the 2021-2022 NBA season. The clustering models can be improved by incorporating data from multiple NBA seasons.
- This clustering model provide 5 clusters. In the future I want to classify players into even more clusters to see if there is a way to further differentiate players.
- I would like to expand the dataset to incorporate advanced gameplay statistics that may illustrate playstyles such as spot-up shots, screen assists, isolated field goal attempts, transition shots made, deflections, charges drawn, or loose balls recovered. How do individual playstyle characteristics influence statistical contributions?

Do you have any questions?

THANKS

Christopher Le
Email: lecm130@gmail.com
https://github.com/chrismle/NBA-Clustering