Project Design Writeup

# Problem Statement & Hypothesis

The aim of this project is to identify whether statistics on basketballers throughout a season can be used to identify and group similar players. I will be collecting data points for every basketballer in the NBA for the 2015/16 season. That’s just under 450 players, who played 10 games or more last season. I will be testing K-means and Heirarchical clustering methods to cluster the players. I will also be using PCA to visualise the cluster of players. My hypothesis is that the player stats will be in some way useful of grouping players together. The benchmark is set in an article by Neil Charles who used football stats from the European football leagues to group footballers into distinct sets of players.

# Assumptions & Questions

Some of the assumptions that I am making:

* The specific positions of the players on the court will not affect their stats
* A players’ performance over a whole season is the best indication of their ability
* As well as accounting for how many times a player does something, it is important to consider how well they do it (i.e. success rate)

# Motivation & Solution Use

Similar problems have been solved using football data. One of the motivations for revisiting this problem is to try and replicate the same problem for the NBA.

The initial output will be an iPython notebook detailing the process to get the results. If I find some interesting groups of players in the clusters, I will work on operationalising the solution, and building an interactive dashboard in Tableau which may help basketball coaches looking to scout new players.

# Step by Step Methodology

1. Collect player stats data from stats.nba.com
2. Do some exploratory analysis to determine how many bins to create for each stat
3. If there is no clear number of bins, create quintiles of all the data points
4. Try using K-means to create clusters of players
5. Also, try using Hierarchical clustering to create clusters of players
6. Change the length of the trees till there is an optimal number of players in each group
7. Visualise the player data using PCA

# Alternate Hypothesis

Can we use the stats of the NBA players in the previous season to predict their salary the next season?

# Data Dictionary

GP – Games Played

W – Wins

L – Losses

MIN – Minutes Played

FGM - Field Goals Made

FGA - Field Goals Attempted

FG% - Field Goal Percentage

3PM - 3 Point Field Goals Made

3PA - 3 Point Field Goals Attempted

3P% - 3 Point Field Goals Percentage

FTM - Free Throws Made

FTA - Free Throws Attempted

FT% - Free Throw Percentage

OREB - Offensive Rebounds

DREB - Defensive Rebounds

REB – Rebounds

AST – Assists

TOV – Turnovers

STL – Steals

BLK – Blocks

PF - Personal Fouls

DD2 - Double doubles

TD3 - Triple doubles

PTS – Points

+/- Plus Minus