Cluster analysis of players from rugby world cup 2015

# Introduction

The Rugby World Cup was one of the biggest sporting events of the 2015 calendar year. The tournament saw records broken on and off the pitch. Bryan Habana matched Jonah Lomu’s Rugby World Cup try scoring record with 15 tries while on the stands over 2.47 million tickets were sold. Television numbers were also outstanding; the final was estimated to have been enjoyed by audiences of 120 million worldwide.

This study used player statistics from the 2015 Rugby World Cup to try to group (or cluster) player into meaningful groups. Cluster Analysis is used in many other areas including social network analysis to recognize communities within large groups of people. I first chose cluster analysis was to investigate if the groups created by the cluster analysis would place players into groups with other players that play in the same positions, this will be explained further in the next sections.

# methods used

Data collection

For this analysis there were two data sets used:

1. RWC 2015 player statistics dataset provided by Kitman Labs (<https://gist.github.com/itsakettle/7f62833a149a0b09424d>)
2. RWC 2015 player sizes dataset (<http://www.rugbyhow.com/2015RWC-all-player-sizes.html>)

The player sizes dataset is only used in this analysis to provide the position of each of the players that took part in the competition.

Clustering players

“ I’m going to try to use K-means clustering to see if we can group players by positions by using just the 10 fields provided in the data. With K-means you have to specify the number of groups to use for clustering I’m going to start with 9, one for each position: prop, hooker, lock, back row, scrum half, fly half, center, wing and full back. “

The above quote shows my initial intention of using clustering algorithm, having some knowledge of Rugby I knew that k-means was unlikely to be able to create 9 separate clusters not due to a fault in the algorithm but because of the nature of the sport. Many players can interchange between multiple positions for example a player like Australia’s Kurtley Beale can basically play any position in the back line. As well as players being interchangeable in modern day rugby the role of players in different positions can be very similar for example back-row forwards and centers can end up doing a lot of the same jobs and when looking at their stats it can be difficult to differentiate between position.

After a lot of experimentation with the number of clusters I decided that 3 clusters made the most sense, while this didn’t split the players into each individual position it did create a pretty interesting division of the player.