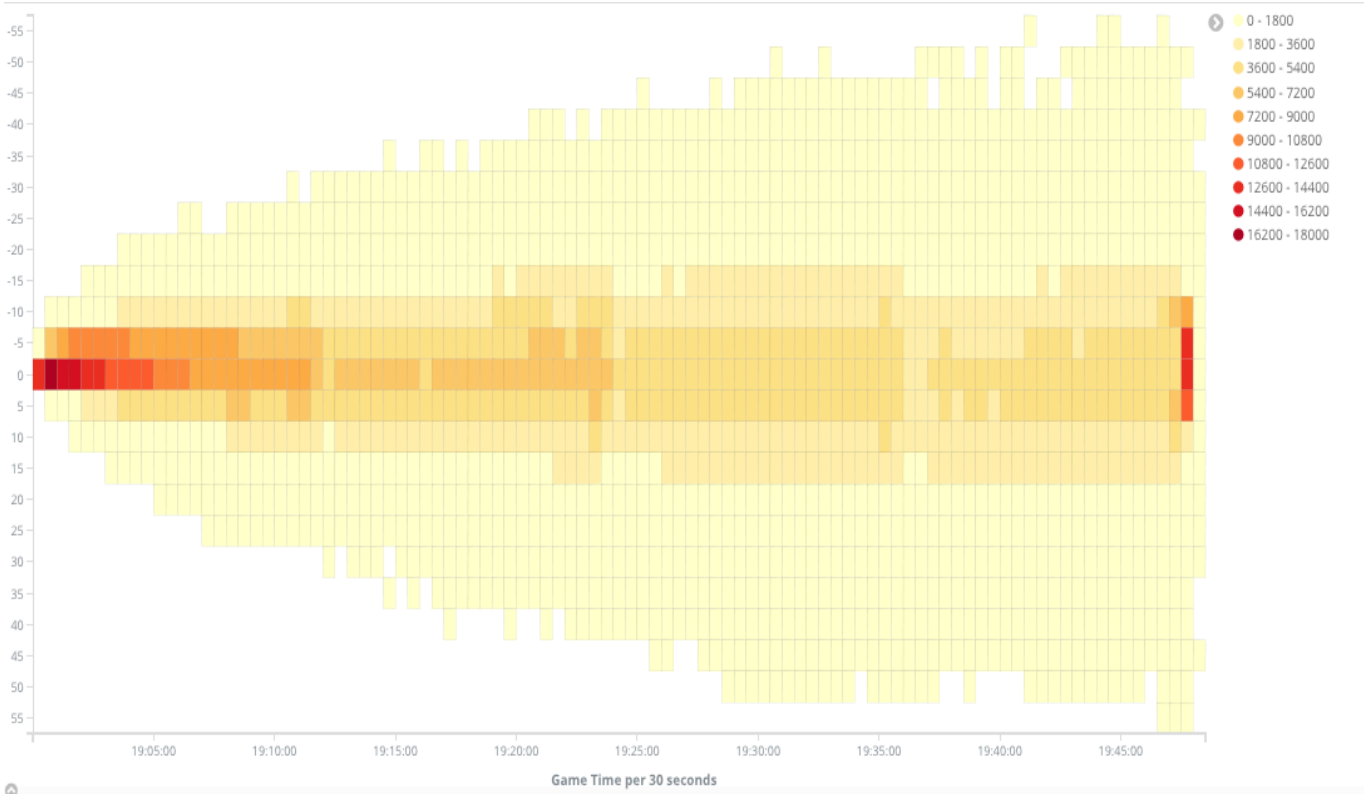
Player Classification and Clutch-ness - Clustering Analysis of NBA Players based on Play-By-Play Data

Xudong Liang (Brandon)

In basketball, everything is relative: almost every accolade in the NBA is based on your performance compared to your closest peers/teams. The nature of competition dictates it is all about comparison. With comparison, comes similarity: who score alike? Who rebound alike? Who assist alike? In short, who play alike? Who are clutch and consistent? This research clusters players based on their on-court contributions over game-time and dive into the trends behind numbers. I scraped play-by-play data of each NBA regular season game from 04-05 season to 16-17 season, thanks to NBA website's stats API (stats.nba.com), consolidated the raw data into a normalized distribution of each player's on-court activities over time by event type (scoring, rebounding, assisting, turnover, etc.) and used frequencies under each time interval (every 30 seconds) as his features for unsupervised clustering (K-Nearest Neighbors). The challenge and thrill of this methodology is that it drills down to the play-by-play level over the span of 48-minute games, something that is rarely conducted and studied before. I used 100 as the number of clusters. Preliminary results somewhat align with expectation. For example, in scoring distribution over time, there are several large clusters containing superstar players like Kobe, LeBron, Dirk, Wade, Durant, Curry, Harden (not all in the same cluster), etc. However, there are also players who are not as high profile that get clustered with those star players. This is because the distribution over time is normalized for each player, meaning we are not comparing total number of occurrences against other players, but rather, the proportion/likelihood of a player scoring in each time interval against others. Thus, this makes the clusters more inclusive for players in all tiers, enabling us to reveal more uncommon yet also important patterns. We can locate these trends at player level as well as through a macroscopic view. Below is a heat map detailing a scoring distribution over game-time for all plays, spanned vertically by the score difference. As expected, the most concentrated scoring takes place in the start of games and toward the end of close games. Moreover, by clustering players based on scoring/rebounding/assisting in the final minutes of close games, we can see which players are indeed "clutch" and consistent by comparing clusters from the down the stretch to clusters that span 48 minutes.



These analyses spark an endless stream of fascinating discussions: which clusters consist of the most "clutch" players? What are the puzzles (clusters) for a quality team? Can we predict teams' performances based on current rosters? Which young players have the potential to become "clutch" based on their clusters? Who are outliers and where are they now? Which are the most common clusters for each position? What is the distribution of positions within each cluster? Are there anomalies and why? I believe a mature clustering analysis will offer revisionist insights on players' value and development in the NBA.