URL to Kernel:

https://www.kaggle.com/chen2870/nba-players-clustering-model

Project Explanation:

This NBA Social Power Project is to utilize NBA team, player and their social media data to explore NBA on the court performance and players' social power. My analysis mainly focuses on NBA players' individual abilities. I also clustered these NBA players into three groups based on their individual abilities.

Data Processing

- **Merge Tables**

To collect all the on-the-court performance information and salary information to evaluate NBA players' individual abilities, I created new table *players* by merging four tables together.

- **Fill Null Values with Mean Value**

Two variables, "3P%" and "FT%", included in the table player contain null values, and I filled null values with mean values of these two variables.

Data Analysis

- **Normalize Players' Information**

The point of normalization is to make variables comparable to each other. Normalization is the process of reducing measurements to a "neutral" or "standard" scale. (1)

- **Compute Euclidean Distance between Players Based on Individual Abilities**

Euclidean distance is most often used to compare profiles of respondents across variables. The distance between vectors X and Y is defined as follows (2): $d(x,y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}$

- **Pair NBA Players with their Most Similar Player**

The following table shows partial result on pairing NBA Players with their most similar player:

	PLAYER	CLOSE_PLAYER_ABILITY
0	Russell Westbrook	[James Harden]
1	James Harden	[Russell Westbrook]
2	Isaiah Thomas	[Kyrie Irving]
3	Anthony Davis	[DeMarcus Cousins]
4	DeMarcus Cousins	[Anthony Davis]
5	Damian Lillard	[Paul George]
6	LeBron James	[James Harden]
7	Kawhi Leonard	[Stephen Curry]
8	Stephen Curry	[Kawhi Leonard]
9	Kyrie Irving	[Bradley Beal]

This table is a good resource for coaches and basketball fans to compare basketball players.

- **Visualize NBA Player's Pair-Up Using Hierarchical Clustering Dendrogram**

The following Hierarchical Clustering Dendrogram is to visualize the players' pair-up result:
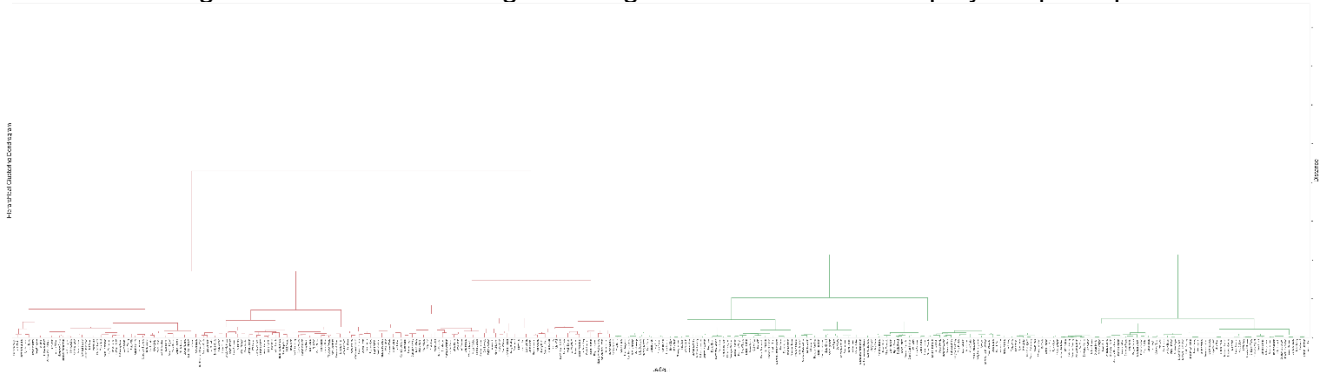
(Larger-size picture can be viewed and downloaded from Kernel.)

Data Analysis

- **Determine the Number of Clusters**

I used the Elbow Method to determine the number of clusters. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters. (3) The following graph indicates k = 3.
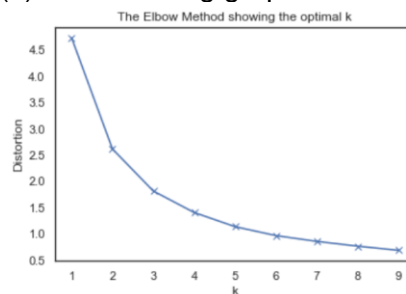
- **K-Nearest Neighbor (KNN) Clustering Result**

The KNN clustering model is displayed below. With the clustered groups, we will be able to identify the belonging group of the new basketball players added to the dataset.
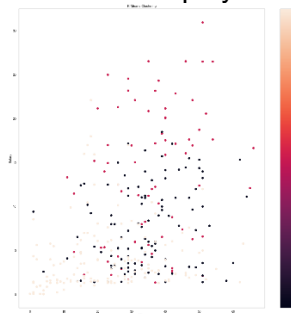
Reference

1. Borgatti, S. (n.d.). Distance and Correlation. Retrieved from
   http://www.analytictech.com/ba762/handouts/normalization.htm
2. Borgatti, S. (n.d.). Distance and Correlation. Retrieved from
   http://www.analytictech.com/mb876/handouts/distance_and_correlation.htm
3. Dangeti, P. (n.d.). Statistics for Machine Learning. Retrieved from https://learning.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml