National Basketball Analysis Progress Report

Background

For our project, we've decided to use data from the National Basketball Association (NBA) that keeps track of statistics for professional basketball players, like the average number of points a player scores per game, as well as advanced statistics (think Moneyball) such as a player's value over an estimated average player at their position, number of wins contributed by that player, and efficiency rating. There are hundreds of players in the NBA, and there are a great number different statistics that have been tracked for each player for many years. We hope to find similarities and difference between players, which can be used in scouting and personnel decisions. We begin by reducing the data's dimension for visualization, will continue with non-linear dimensionality reduction and clustering players by certain data subsets, and perhaps execute regression on similar players and salary data.

Data Collection and Cleaning

The first step in the project was collecting the data. Although all of the statistics we are using are publicly available, it was necessary to collect the data into a usable format. To achieve this, we used Scrapy[1], a web crawling and scraping framework for Python, to scrape data from the Basketball Reference website.[2] The statistics were also pruned to remove some irrelevant features and to remove incomplete data (e.g. data from this current season).

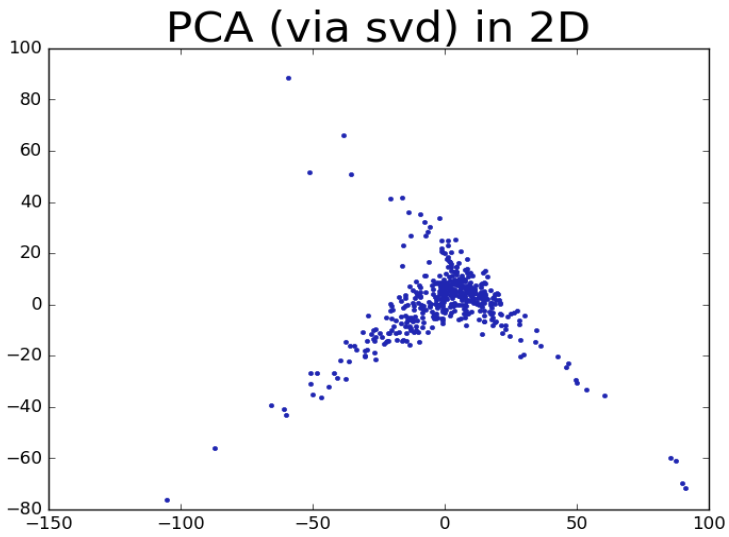
A good portion of time had to be spend on "data-cleaning". For example, given the large scraped dataset, we had to remove 'season-rows' to focus on career statistics, fill null values with appropriate type (different for absolute counts, averages, percentages, etc.) and select the columns we desired. Also finding which variables are absolute, normalized, averages, etc. Finally, we removed variables which would not give us good information about subtle player differences because you would expect huge differences *(relying on domain specific knowledge for this analysis)*, like 3-Point attempts for centers, who (usually) never take such attempts. A lot of the process was learned through trial-and-error, by looking at data column printouts, and observing PCA/MDS plots showing obvious weight to one variable (which led us to remove that variable).

[1] https://scrapy.org/
[2] A basketball statistics reference site (http://www.basketball-reference.com/).

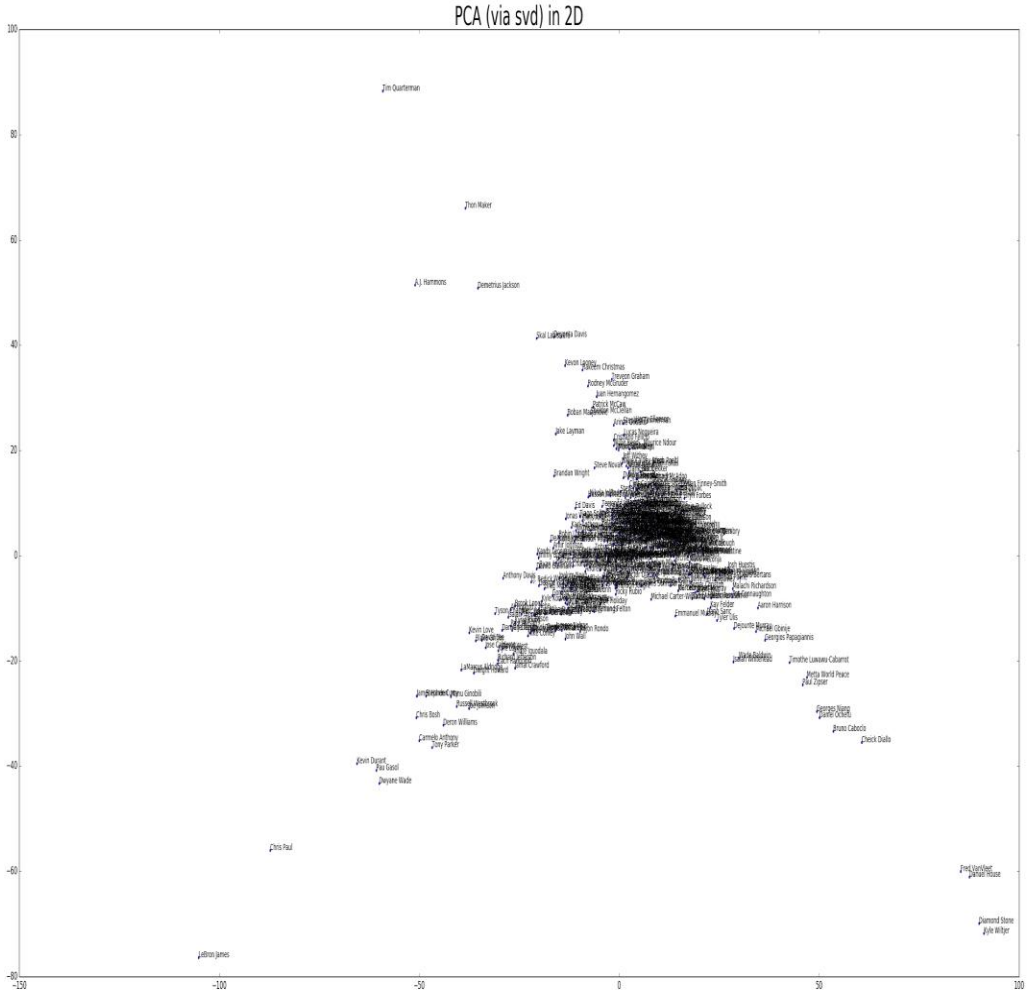
Early Results

For testing our initial analysis and code in general, we decided to focus on Offensive statistics. Our real interest is Defense, so we thought it good to work out the kinks with something better understood - also there are far more offensive statistics since shooting data is more thoroughly tracked and contains obvious, discrete events. Shooting data is about a player's identifiable actions, and shooting can be done in isolation, whereas defensive statistics are about how one player's presence influences others, and therefore contains more complexity.

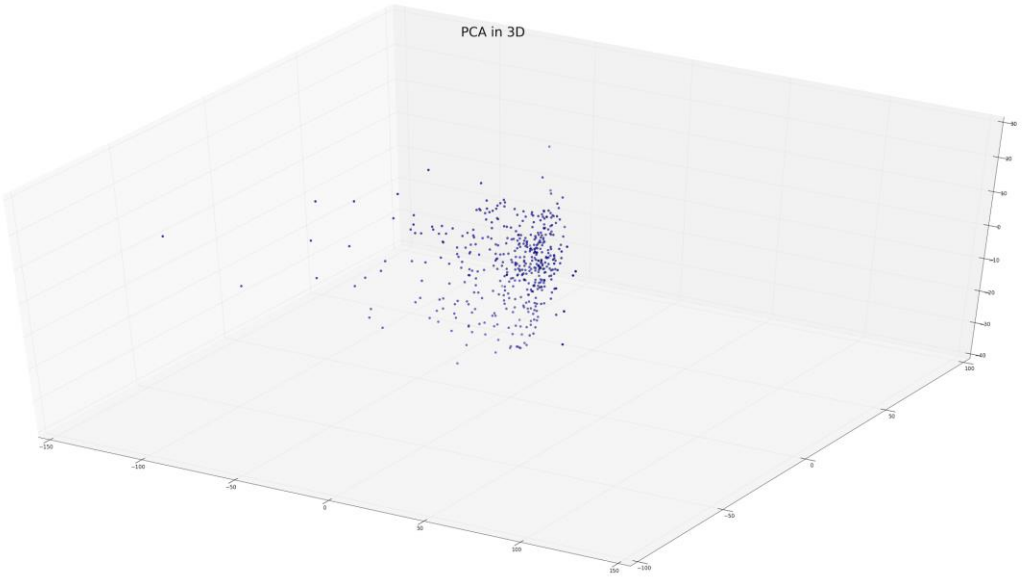
We chose the columns we labeled as "offensive" statistics, and then we removed those that were not normalized and had an overwhelming effect in our first plots. While potentially there is more data cleaning to do, as it seems that maybe 2 or 3 variables dominate, it does not appear dominated by 1 variable as was the difficulty without original plots. Below are the standard dimensionality-reduction plots. We have used cMDS and PCA. PCA shown.



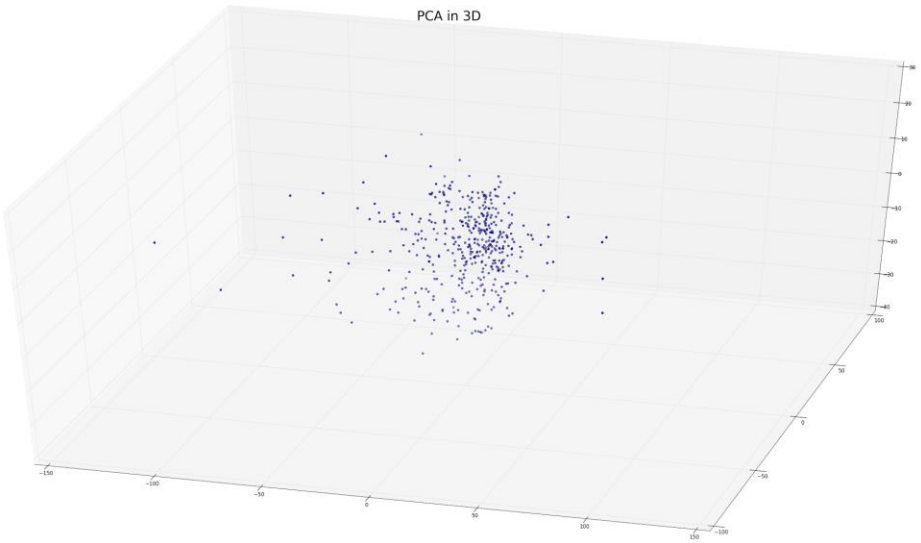
Below is the PCA plot with names of players mapped to point locations:



Below are some 3D views of the PCA technique applied to Offensive data.



Altering the viewing angles, we also obtain:



It seems, from qualitative analysis, that a considerable amount of information is present in the 3rd dimension, and further comparisons (like metrics used on eigen/singular values corresponding to principal axes) will be done to determine "natural" dimensionality of the data, or what we can reasonably approximate with reduced dimensions.

Future Steps

A good portion of our work has been spent on setup, that it: planning, scraping data, loading, writing helper functions, reviewing "pandas" documentation, writing helper functions for plotting, formatting, and reducing dimension, statistical and domain-specific analysis and decisions about which variable to include, hypothesis generation. Now that setup is complete, further analysis has a much lower fixed cost (in terms of time), so that simple things like **pd.DataFrame.query("<expr>")** can be used to sort the columns and values easily, to choose different statistics and reduce dimension/cluster on different metrics. That is, we believe the annoying set-up work to be done.

To further our project, more complex forms of analysis will be performed. For example, non-linear methods of dimensionality reduction and non-Euclidean distance metrics will be tested. In addition, we hope to be able to use our results to draw previously unseen conclusions about NBA players' playing styles and effectiveness, particularly as it relates to defensive ability. Further decisions will be made as how to capture this information as distinct from offensive performance.