

# Final Paper

Noah Edwards-Thro

1/23/2022

## Schedule

2/14 - Results Copied from Sloan Paper

2/21 - New Results

2/28 - Discussion

3/14 - Clean Up Visuals

3/21 - Open Week

























3/28 - Introduction/Final Edits Part 1

4/4 - Final Edits Part 2

## 2. Methods

### 2.1 Reproduction and Replication

To recreate the Sloan model as close as possible, I endeavored to use as close to the same data and methods that Kalman and Bosch used as I could.

	Kalman and Bosch	Replication	Expansion
Population			
Question			
Hypothesis			
Data	01100 10110 11110	01100 10110 11110	01100 10110 11110
Variables			
Analyst			
Code			
Results			
Claim			

## 2.2 Reproduction of Sloan Paper

### 2.2.1 Data

Kalman and Bosch manually scraped 10 years (ranging from the 2009-2018 seasons) of player statistics covering advanced aggregation statistics, per possession statistics, and shot distribution statistics. Their data consisted of 5512 observations with 73 variables where each observation was a single season of a single player (if a player played all ten seasons during this range, he would show up 10 times).

While I desired to stay as close to the process that Kalman and Bosch used as possible, I had knowledge of a package in R called `nbastatr` that pulled data directly from Sports Reference LLC (the same website that Kalman and Bosch manually scraped their data from). In an effort to get more experience working with this package, I decided to pull the data from this package using the `bref_player_stats` function. Additionally, the shot distribution statistics were unable to download via the `nbastatr` package so I manually scraped them from Sports Reference LLC. Finally, I used the `player_profiles` function in the `nbastatr` package to download the heights (in inches) for all of the players.

```
## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.5    v purrr  0.3.4
## v tibble  3.1.6    v dplyr  1.0.8
## v tidyr   1.1.4    v stringr 1.4.0
## v readr   2.0.1    v forcats 0.5.1

## Warning: package 'dplyr' was built under R version 4.1.2
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()
```

My data consisted of 4760 observations and 136 variables. Through some exploration, I found that the discrepancy in observations has to do with players who are traded. In my dataset, any player who is traded mid-season will still only show up as one row, with that players' team being "TOT" (signaling that the player played for multiple teams). Looking at the data on the Sports Reference LLC website, it seems that Kalman and Bosch likely had a separate row for each team that a player played on in a single season (so if a player played games for three separate teams, he would have three separate rows for that season in Kalman and Bosch's data while he would only have 1 row in mine).

In Kalman and Bosch's analysis, they filtered the data so that only observations with more than 30 games in a season are counted, ending with 3,608 observations. After using the same filter, I ended with 3,676 observations, likely due to the difference in data format with traded players (a traded player who plays 20 games for two different teams doesn't show up in their analysis but will show up as playing 40 games in mine).

## 2.2.2 Variable Selection

I used the same variables that Kalman and Bosch used in their model (with the exception of using per 36 minute statistics instead of per 100 possession statistics for points and field goal attempts). The variables were a combination of offensive, defensive, and aggregate statistics and all statistics were calculated as rates except for player height. The drawback of these rate statistics is that they does not take into account how much a player is on the court. For instance, a player who plays 10 minutes a night could have the same rate statistics as a player who plays 35 minutes a night, but we would consider these players very different in their abilities (largely due to the 35 minute player being able to play efficiently for 35 minutes).

Table 1: Recreation from Kalman and Bosch

Variable	Description
Height	Player height, in inches
Offensive Rebound Rate	% of available offensive rebounds a player gets while on the floor
Defensive Rebound Rate	% of available defensive rebounds a player gets while on the floor
Assist Rate	% of teammate field goals that a player assisted while on the floor
Steal Rate	% of opponent possessions that end with a steal b the player while on the floor
Block Rate	% of opponent field goal attempted blocked by the player while on the floor
Turnover Rate	Turnovers committed per 100 offensive possessions
Points <sup>1</sup>	Points scored per 100 offensive possessions
Usage Rate	% of offensive team possessions used by the player while on the floor
Player Efficiency Rating	Per-minute production standardized such that the league average is 15
Free Throw Rate	Number of free throws made per field goals attempted
Free Throw Percentage	Number of free throws made per free throw attempt
Field Goals Attempted	Number of field goals attempted per 100 possessions
2FG%	Number of two-point field goals made per attempt
3FG%	Number of three-point field goals made per attempt
2FG Assist Rate	% of two-point field goals that are assisted
3FGA%	% of field goal attempts that are three-point attempts
Corner 3FGA%	% of three point-field goal attempts that are from the corner
3FG Assist Rate	% of three-point field goals that are assisted
Dunk Attempt Rate	% of all field goal attempts that are dunks
0-3 ft FGA%	% of all field goal attempts between zero and three feet from the basket
3-10 ft FGA%	% of all field goal attempts between three and ten feet from the basket

Variable	Description
10ft-3p FGA%	% of all field goal attempts between ten feet from the basket and the three-point line

### 2.2.3 Gaussian Mixture Clustering

After attempting K-means clustering and being unsatisfied with the results, Kalman and Bosch pivoted to model-based clustering to cluster the players in their dataset. Model based clustering, specifically finite Gaussian mixture modeling, uses an expectation-maximization (EM) algorithm to fit observations into clusters. An advantage of model-based clustering is that it assigns “soft clusters”, showing the probability that each observation will be in each cluster. The figure below (INSERT FIGURE), displays a graphical representation as to the clustering distributions produced by Gaussian mixture modeling.

As done by Kalman and Bosch, I used the “mclust” package in R to implement the Gaussian mixture clustering.

## 2.3 Expansion of Kalman and Bosch’s Paper

### 2.3.1 Data Expansion

While Kalman and Bosch’s original paper only used data from the 2009-2018 seasons, I wanted to see what predictions the model would make on more recent seasons (2019-2021 seasons) given the great degree to which the NBA has changed over the past half decade alone. As before, I used the `nbastatr` package to download the majority of the data and I still had to manually scrape the shooting statistics necessary.

My three new seasons of data consisted of 1600 observations and 136 variables. To keep in line in the original analysis, I filtered the data to players with greater than 30 games, resulting in 1125 observations remaining.

### 2.3.2 Variable Reduction

With the new seasons of data, I first wanted to see what the predictions looked like from the previous model. I used the same 23 variables and ran the initial model on the new dataset.

Following this, I decided to change the input variables to the model and reduce the number of clusters that it output. I believed that for the most part, steal and block rates were mostly random and not necessarily indicative of position (or had overlapping value with other variables - e.g. height and block rates both being a strong predictor for traditional center). I also cut out the PER variable as I believed that it overlapped with other variables. Furthermore, while Kalman and Bosch used solely rate statistics in their analysis, I sought to investigate how game time (in minutes per game - MPG) could influence the model and player predictions. In addition to adding this MPG variable to the model, I combined some of the shooting variables into single variables (10-16 and 16-3P became Midrange) in an effort to reduce the number of input variables.

### 2.3.3 Cluster Reduction

Kalman and Bosch never outlined in their paper why they chose 9 clusters as the optimal number, but I hypothesized that a slight reduction in clusters (to 6 clusters) would make more intuitive sense. I believed that some clusters should be grouped together instead of separated because they were so similar.

## 3. Results

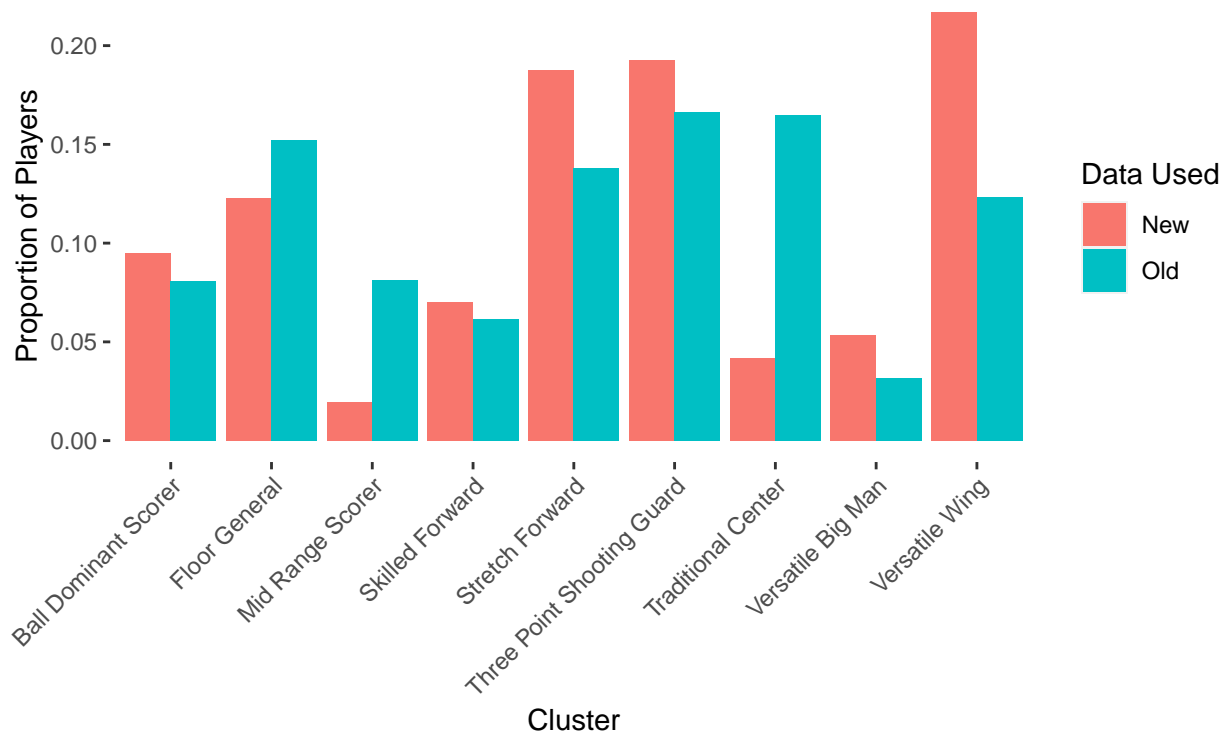
### 3.1 Reproduction of Kalman and Bosch's Model

In their analysis, Kalman and Bosch found 9 clusters and assigned them labels based on player types within those clusters. While my clusters were fairly close those of Kalman and Bosch's analysis, they were different in a few clusters. My Mid Range cluster was more diverse than just big men so I named it Mid Range Scorer. I did not have a High Usage Guard cluster and found that the players Kalman and Bosch listed in their High Usage Guard cluster were clustered under Ball Dominant Scorer.

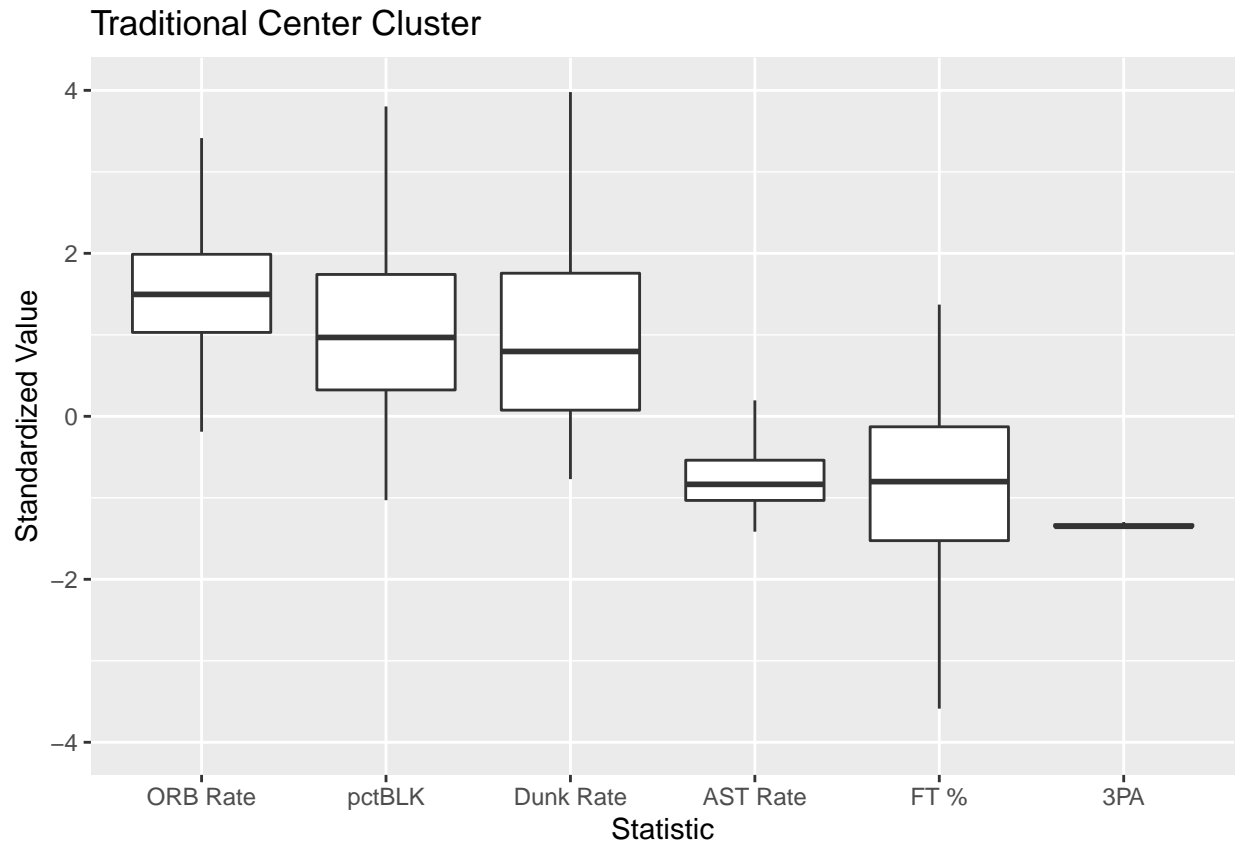
In my analysis, Three Point Shooting Guard was the highest cluster, and while it would have been the highest cluster with just the seasons that Kalman and Bosch used, the three new seasons that I added had an even higher proportion of players clustered into the Three Point Shooting Guard cluster.

```
ggplot(data = scaled_combined_preds,
       aes(
         x = First_Prediction,
         group = data_time,
         fill = data_time,
         y = ..prop..
       )) +
  geom_bar(position = "dodge", stat = "count") +
  labs(
    x = "Cluster",
    y = "Proportion of Players",
    title = "Figure 3: Distribution of Players Across 9 Clusters",
    subtitle = "Recreation of Kalman and Bosch Graph",
    fill = "Data Used"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_blank())
```

Figure 3: Distribution of Players Across 9 Clusters  
Recreation of Kalman and Bosch Graph



Below is a graphical representation of the Traditional Center cluster. The Traditional Center was very high in Offensive Rebounding, Block Rate, and Dunk Rate. Additionally, the Traditional Center was low in Assist Rate, Free Throw Percentage, Percentage of Shots Taken from 3.



One of the challenging things about clustering players is that there is not necessarily a way to tell if the clustering model is effective or not. I chose to go about this problem by using the same clustering method over all 13 seasons (instead of just the original 10) and seeing how many players remained in the same cluster. When I did this I found that 67.3817955 percent of the players remained in the same cluster. This number is largely drawn down by just 22.5% of the Versatile Wings and 26% of the Versatile Bigs remaining in the same cluster.

### 3.2 Expansion of Kalman and Bosch's Analysis

While working with the data from the Kalman and Bosch analysis, I hypothesized that the data would fit better into a smaller number of clusters. Additionally, I believed that a smaller number of variables should be used as there were a few variables in Kalman and Bosch's analysis (specifically variables related to defense such as block rate and steal rate) that are not meaningful in establishing the clusters.

```
ggplot(data = reduced_preds,
  aes(
    x = First_Prediction,
    group = 1,
    y = ..prop..
  )) +
geom_bar( stat = "count", fill = "Blue") +
labs(
  x = "Cluster",
  y = "Proportion of Players",
  title = "Distribution of Players Across 6 Clusters",
```

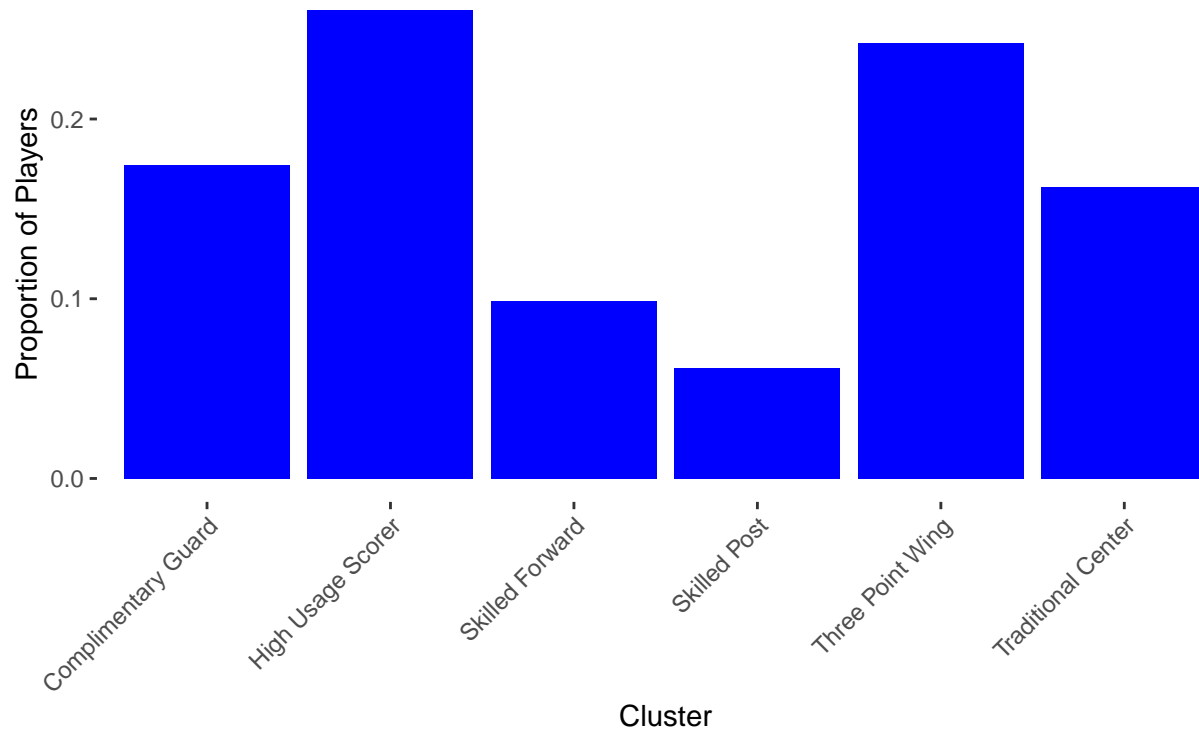
```

    subtitle = "Model Based off of Data from 2009-2018"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_blank())

```

## Distribution of Players Across 6 Clusters

Model Based off of Data from 2009–2018



```

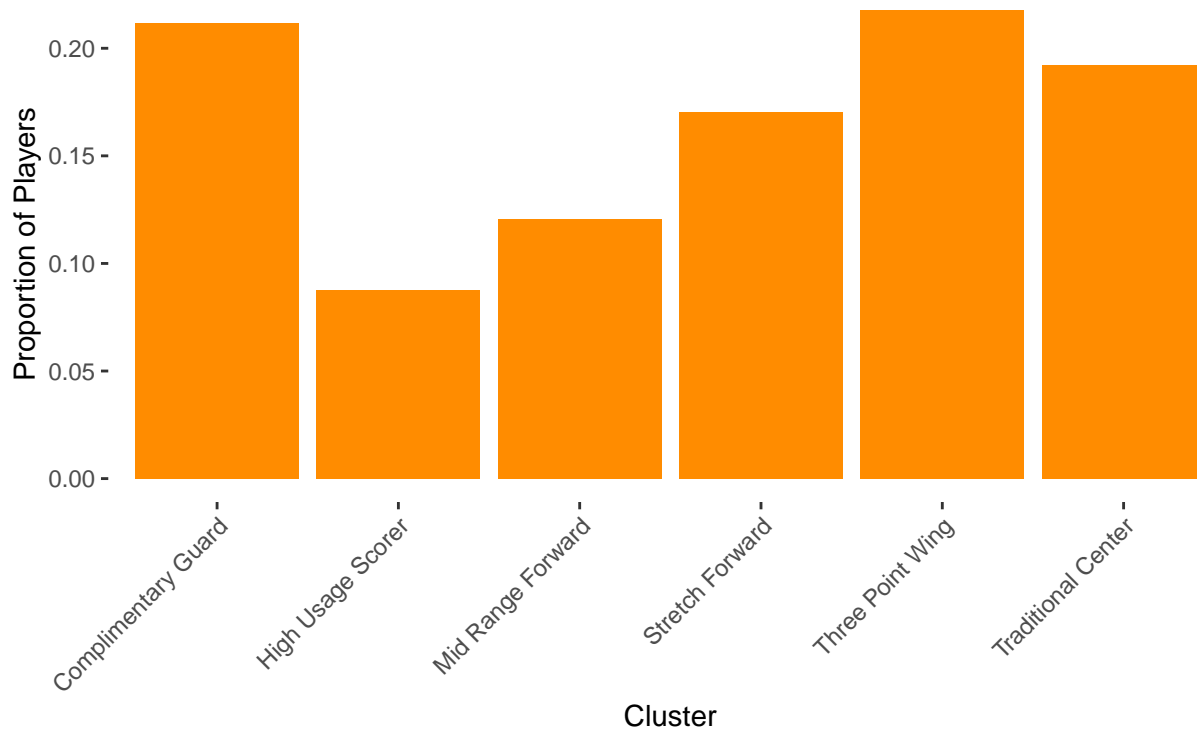
ggplot(data = reduced_preds,
       aes(
         x = New_Prediction,
         group = 1,
         y = ..prop..
       )) +
  geom_bar(stat = "count", fill = "Dark Orange") +
  labs(
    x = "Cluster",
    y = "Proportion of Players",
    title = "Distribution of Players Across 6 Clusters",
    subtitle = "Model Based off of Data from 2009-2021"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.background = element_blank())

```



## Distribution of Players Across 6 Clusters

Model Based off of Data from 2009–2021



I initially created a model using 6 clusters based on the original 10 seasons worth of data that Kalman and Bosch used. My initial model created clusters of Complimentary Guard, High Usage Scorer, Skilled Forward, Skilled Post, Three Point Wing, and Traditional Center. The proportion of players in the High Usage Scorer and Three Point Wing clusters were particularly high while the proportion of players in the Skilled Forward and Skilled Post clusters are relatively low compared to the other clusters.

After this, I created a new model based on all 13 seasons worth of data. The clusters in this model were slightly different than the clusters in the previous model. The Complimentary Guard, High Usage Scorer, Three Point Wing, and Traditional Center remained the same but the Skilled Forward and Skilled Post clusters became Mid Range Forward and Stretch Forward clusters. This was particularly interesting to me because many of the players that were clustered into the Skilled Post cluster moved into the Traditional Center cluster with the new model. It appears that the model with all 13 seasons saw it more important to divide up the wings and forwards into different clusters rather than divide up the interior post players. I hypothesize that this is because the first 10 seasons are heavy upon post players while adding the three most recent seasons shows a shift away from a game centered around the post towards a game dominated on the perimeter.

Cluster		
	Description	High Stats
Complimentary Guard	Assist Rate	Height
	Usage Rate	Def. Reb %
High Usage Scorer	Points Per Minute	Off. Reb %
	Usage Rate	FGA (0-10 FT)
Mid Range Forward	FGA (16FT-3P)	3P FGA
	Off. Reb %	Points Per Minute

<sup>1</sup>The points here are points per game while Kalman and Bosch used point per 100 possessions

Cluster			
<sup>1</sup> The points here are points per possession while Kalman and ghost stats point per 100 possession			
	Description	High Stats	Low Stats
Stretch Forward		3P FGA	Usage Rate
		3P FG%	Points Per Minute
Three Point Wing		3P FG%	Off. Reb %
		3P FGA	FT Rate
Traditional Center		Off. Reb %	3p FG%
		FT Rate	Assist Rate