**Introduction**

In this report I take publicly available data from basketball-reference.com to explore some interesting questions about NBA basketball. First, I look at the relationship between the box score statistics a player accumulates in an average game and how valuable they are towards winning the game. Then, I do a preliminary exploration of how the NBA has changed over the past few decades and whether player types have evolved and shifted over the years. Traditionally there are five classic player types each with their own traditional role in the game. (Point Guard, Shooting Guard, Small Forward, Power Forward, Center) Do players still play the same now in 2020 as they did in the 1990s, and has there been an evolutionary shift in playing style over the years?

**Description of the Dataset**

I use three different datasets in this analysis all taken from basketball-reference.com. I took per game statistics for active players in the 1997-98, 2013-14 and current 2019-20 season. In order to have a more accurate reflection and understanding of our data I filtered our observations to players who played meaningful minutes and a minimum number of games. I filtered out players who played less than 10 games in the season and less than 14 minutes a game (in a 48-minute game, 82 game season). This leaves around 300 players for the 2019 dataset and ~180,240 respectively for 1997 and 2013 which in a 30-league team with 15 roster spots keeps observations in line with a standard 8-10 player rotation.

The following variables are used in the analysis and relate to per game averages:

(X3PA% was removed because there were a few missing values, but the essence of the variable is in X3P and X3PA)

MP- minutes played FT% - percent of free throws made

FG- field goals made ORB – offensive rebounds

FGA – field goals attempted DRB – defensive rebounds

FG% - percent of field goals made TRB – total rebounds

X3P – three pointers made AST - assists

X3PA – three pointers attempted STL - steals

FT – number of free throws made BLK -blocks

FTA – number of free throws attempted TOV- turnovers

PTS – points scored WS.48 – win shares per 48 minutes

Win shares per 48 is an advanced metric that estimates how many wins a player is worth over the span of a game and is the evaluation we use to estimate how valuable a player is and whether they contribute positively or negatively to a win with a higher win share meaning they contribute more and are therefore more valuable.

**Choice of Methodology and Motivation**

To answer the first question about the relationship between box score statistics and player value I choose to do a PCA analysis where I label the points by categorizing them on their win share value. PCA lets us maximally explain the variance in our dataset and reduces the dimensionality so we can reduce the 18 variables down to 2. This lets us visualize our data and we can look at the PCA loadings to see what variables are most important to these principal components.

To look at player evolution over time and whether certain players can be grouped by playstyle I use clustering with principal components on the axes. I filtered the variables even further to ones that are directly related to play style. This includes only actual contributions to the game such as shooting the ball or grabbing a rebound and removes points scored and efficiency values. By clustering we can see whether players are grouped together based on their in-game actions. I set the axes to be the first two principal components for the 2019 year and used these loadings to calculate the scores for the 1997 and 2013 years. These can then be graphed to see if they cluster at different geometric points corresponding to different play styles. I used five clusters to emulate the idea of the five main positions.

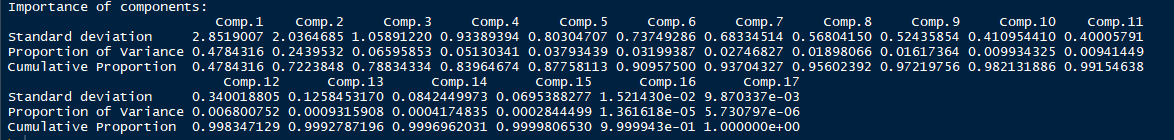
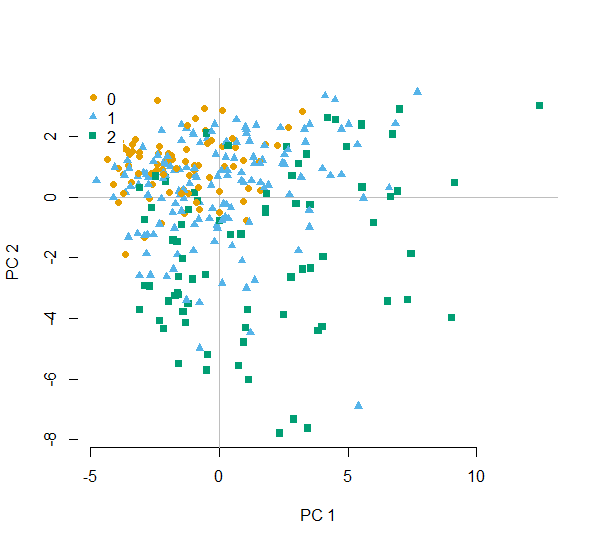
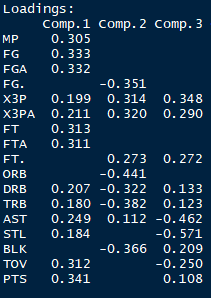
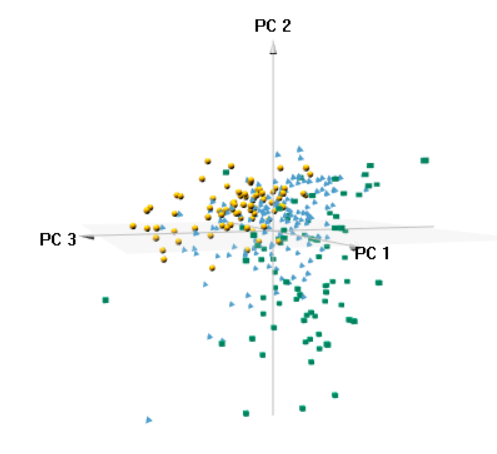
**R-output (Only the most important figures)** 

Figure 1: Variance explained by the principal components for the 2019 season box stats statistics



**PC1 vs. PC2 Box scores 2019 Season (2D)**





**PC1 vs. PC2 vs PC3 Box scores 2019 Season (3D)**

Figure 3: Loadings scores for first 3 PC based on box score statistics.

Figure 4: Principal Component 1 vs. Principal Component 2 vs. Principal Component 3 Labeled by WinShares per 48 minutes

Figure 2: Principal Component 1 vs. Principal Component 2 Labeled by Winshare per 48 minutes

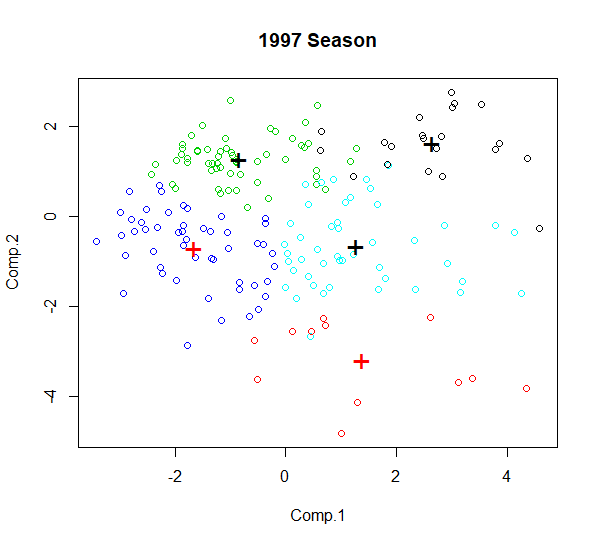
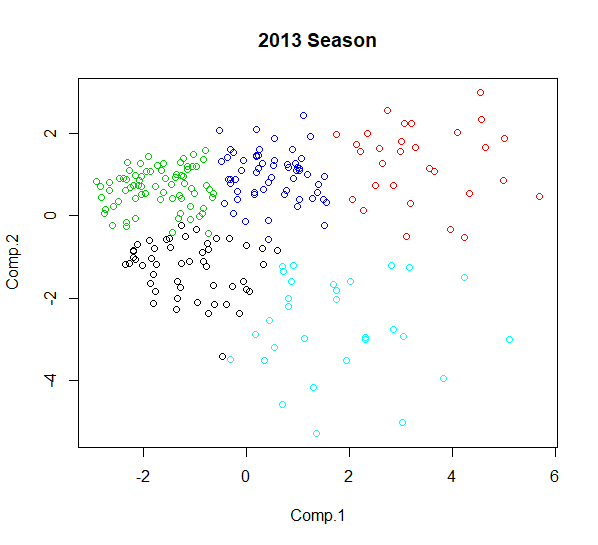
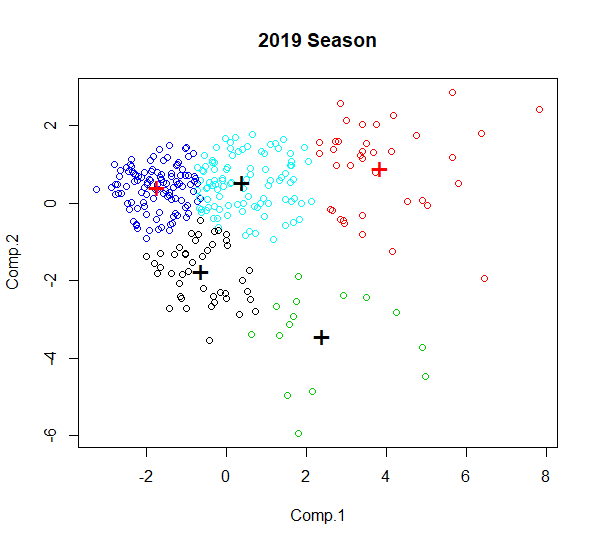


Figure 6: Clustering of players based on 2013 Season scores using Principal component loadings from the 2019 Season data

Figure 5: Clustering of players based on 1997 Season scores using Principal component loadings from the 2019 Season data



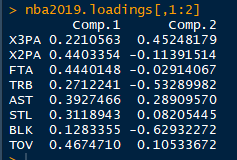


Figure 8: Loading scores for Cluster analysis based on 2019 Season data

Figure 7: Clustering of 2019 players using PC1 and PC2

**Interpretation**

Our first exploration was on whether there is a relationship between the box score statistics a player puts up and how much they contribute to a win. I created a 2D and 3D representation that visualizes the first 2/3 PC components labeled by the contribution a player has towards winning (win shares).In Figure 2, our 2D plot we can see that low win share players tend to group towards the top left with high PC2 scores and low PC1 scores, while the most valuable players shift down and to the right with increasing PC1 scores and lowering PC2 scores. Figure 4 on the other hand is a 3 dimensional showing and while it is difficult to view without the flexibility of rotating it, we can notice a similar pattern of low value players being in the top left and higher value going towards the bottom right. Looking at the loading scores in Figure 3 we can identify that high value players put up a lot of positive stat totals which leads to a high PC1 score and a low PC2 score while low value players are the reverse leading to high PC2 scores and low PC1 scores. This confirms what would we expect since players who contribute positively to the game will lead their team towards winning.

The cluster analysis did not do a great job in my mind of showing a convincing relationship and trend of new ways of playing the game of basketball. Currently over the last few years there has been a shift in how the professional game is being played where significantly more three pointers are being shot every game. When players used to shoot maybe one or two per game now some are taking over 10. I expected to see a big difference in clustering based on this, but I do not see a clear difference in the 1997 to 2019 representations. I think this is related to three point shooting not having a high enough weight on the Principal component loadings and therefore it does not do a good job of separating players based on that variable alone. Possibly other techniques and applications could do a better job of showing the evolution of the game that I believe to be there. The biggest difference I noticed is the extension of the PC1 axis from a max score of around 4 to 6 then 8 over the years. Looking at the increase in PC1 scores over time this is an increase in total actions in the game (shots, rebounds etc.) which seems to be more closely related to the increased speed and pace of the game rather than individual playing style I originally posited.

**Reference**

https://www.basketball-reference.com/

**R-code**

setwd("C:/Users/samue/Desktop/Fall 19 Semester/Stats/Assignment")

install.packages('ggfortify')

library(ggfortify)

# Load in three different per game statistics,we will filter by games played and minutes played to eliinate players without enough data

data.2019 = read.csv("2019-20 data NBA")

data.2019advanced = read.csv("advancedstats2019")

data.2013 = read.csv("2013-14 NBA")

data.1997 = read.csv("1997-98 NBA")

# Remove some variables that are not necessary for our analysis

# (Rank, Team, games played and games started,Personal fouls,age and 3point% to account for dividing by zero)

data.2019advancedupdated= data.2019advanced[,c(2,24)]

data.2019updated = data.2019[,c(2,3,6,8:13,15:28,30)]

final2019 = merge(data.2019updated,data.2019advancedupdated, by = "Player")

# remove name, games played and position and Filtering by minutes played and games played

final2019 = final2019[which(final2019$G >10 & final2019$MP >14),]

predictors2019 = final2019[,c(4:9,14:24)]

# PCA

pca2019 = princomp(predictors2019,cor = T)

plot(pca2019, type = 'l')

summary(pca2019)

# 3 Principal components explain ~78% of the variance

final2019$wsfactor = ifelse(final2019$WS > .135, "2", ifelse(final2019$WS >.059, "1", "0"))

summary(final2019$WS.48)

# Used 1st Quartile and 3rd quartile for partioning

install.packages("pca3d")

library(pca3d)

pca3d(pca2019, group =final2019$wsfactor)

pca2d(pca2019, group = final2019$wsfactor, legend = 'topleft')

pca2019$loadings

# Clustering with 9 variables based on player type

# first we want to remove all the variables except for ones that contribute to playing style

cl2019 = final2019clustering[,c(9,11,15,19:23)]

dim(cl2019)

apply(cl2019,2,min)

apply(cl2019,2,max)

plot(c(0,10),c(0,20),type="n",xlab=var,ylab="Value",main="Profile Plot")

for (k in (1:306))

{

points(1:8,cl2019[k,],type="l")

}

#standardize data

scaled2019<-scale(cl2019,center=TRUE,scale=TRUE)

var(scaled2019)

apply(scaled2019,2,min)

apply(scaled2019,2,max)

plot(c(0,10),c(-2,6),type="n",xlab=var,ylab="Value",main="Profile Plot")

for (k in (1:306))

{

points(1:8,scaled2019[k,],type="l")

}

# Andrews plot

install.packages("andrews")

library(andrews)

andrews(scaled2019,type = 1,clr =1, ymax = 5)

# Trim filter and scale 1997 and 2013 data

data1997.updated = data.1997[which(data.1997$G >10 & data.1997$MP >14),]

dim(data1997.updated)

data1997.trim = data1997.updated[,c(13,16,20,24:28)]

data1997.scaled<-scale(data1997.trim,center=TRUE,scale =TRUE)

data2013.updated = data.2013[which(data.2013$G >10 & data.2013$MP >14),]

dim(data2013.updated)

data2013.trim = data2013.updated[,c(13,16,20,24:28)]

data2013.scaled<-scale(data2013.trim,center=TRUE,scale =TRUE)

# Do clustering on 2 principal components of the scaled data

library(cluster)

nba2019.pca <- princomp(scaled2019)

nba2019.pkcl <- kmeans(nba2019.pca$scores,5,30)

par(mfrow=c(1,1))

plot(nba2019.pca$scores[,1:2], col = nba2019.pkcl$cluster, main = “2019 Season”)

points(nba2019.pkcl$centers[,c(1,2)], col = 1:2, pch ="+", cex=2)

#create scores now based on 2019 loadings for 2013 and 1997 data

nba2019.loadings = loadings(nba2019.pca)[]

nba2013.scores = data2013.scaled%\*%nba2019.loadings

nba1997.scores = data1997.scaled%\*%nba2019.loadings

#Use first two PC components of 2019 for 1997 and 2013 data

par(mfrow=c(1,1))

nba2013.pkcl <- kmeans(nba2013.scores,5,30)

plot(nba2013.scores[,1:2], col = nba2013.pkcl$cluster, main = “2013 Season”)

points(nba2013.pkcl$centers[,c(1,2)], col = 1:2, pch ="+", cex=2)

par(mfrow=c(1,1))

nba1997.pkcl <- kmeans(nba1997.scores,5,30)

plot(nba1997.scores[,1:2], col = nba1997.pkcl$cluster, main = “1997 Season”)

points(nba1997.pkcl$centers[,c(1,2)], col = 1:2, pch ="+", cex=2)