Analysis of NBA Players
Rory Gardner

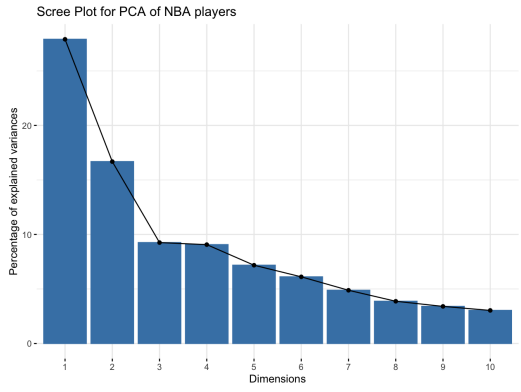
Data-driven statistical analysis is an essential component of breaking down factors of success in professional basketball. Advanced methods and exploratory analysis techniques help to reveal vital metrics in player performance. Throughout this analysis, my primary goal is to identify the key metrics that define player success and how these metrics relate to each other. I will evaluate if the most important advanced metrics of success can be predicted in previous seasons in which they were not recorded. Finally, I will explore the variation of primary performance factors by player position.

The dataset used in this analysis is "NBA Stats (1947-present)" by Sumitro Datta on Kaggle. The dataset originally contained 32 columns of raw data about players and their performance and advanced statistical metrics of player success. The dataset was last updated on December 10, 2024, and it is updated through the latest season that is currently in progress.

The original dataset contained observations in adjacent leagues to the NBA, such as the ABA. I narrowed the observations to only the NBA. The dataset also contained a large number of missing values, and I found 99% of these to exist in observations before 1990. I chose to primarily work with data recorded after 1990 to reduce skewness. Finally, I removed several columns that were irrelevant to my analysis, such as season id and birth year.

Many of the variables in this dataset are common in basketball statistics. However, for further information, a glossary is included before the appendices to explain the purpose of each statistic.

Principal component analysis is a dimension reduction technique that reveals the key metrics that explain variance in the dataset. It highlights the most important variables regarding player success, and may uncover advanced measures of performance through relationships between dimensions.



Scree plots help determine the optimal number of dimensions to retain by displaying the percentage of variance explained by each principal component using the eigenvalues. The goal is to capture most of the dataset's variance using the PC's. Typically, the point at which the percentage of explained variance drops significantly indicates the number of PC's that should be retained. In this case, the drop occurs after PC2, indicating that the first two PC's should be retained to best explain the variance in the dataset without including redundant or unnecessary information.

Biplots help visualize the variables within the PC space, illustrating each original dimension's contribution and representation of the principal components.

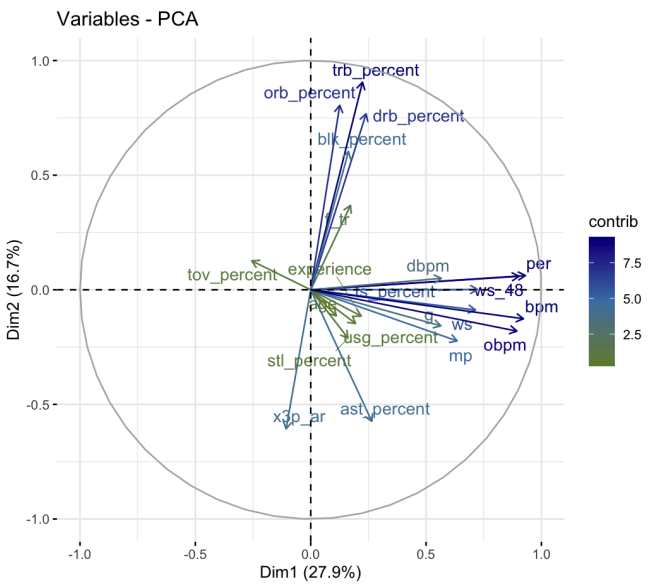


Figure 2 displays the contribution of each variable relative to the first two PC's. The direction of the loading vectors relative to the x and y axes indicates the direction of their contribution. The length and color of the vectors represents the strength of their contribution. This is indicated by a color gradient legend to the right of the plot. Dimensions such as player efficiency rating, box plus/minus, and win shares contribute strongly to PC1. Dimensions such as total rebound percent, blocking percent, and average three point percent contribute to PC2.

Figure 2 (above) and figure 3 (below)

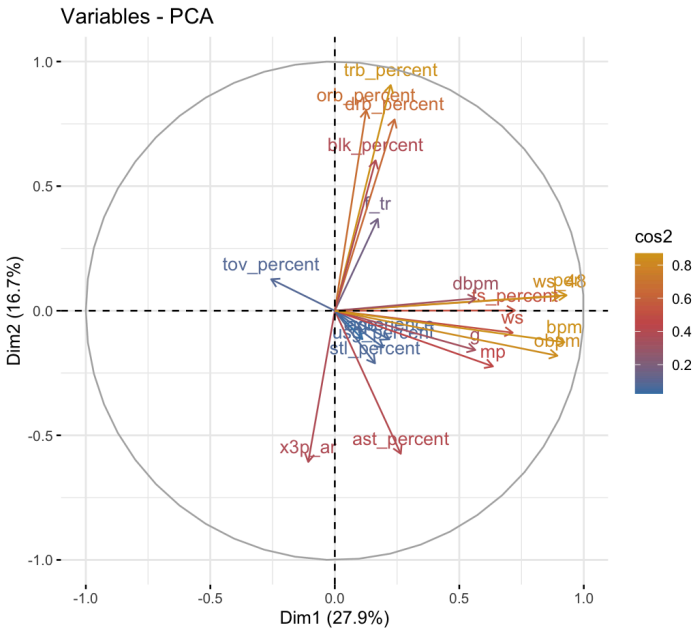


Figure 3 displays the cosine squared value of each original dimension in the PC space. The cosine squared value indicates the quality of the contributions of each variable to the PC's. A higher cosine squared value suggests that the variable contributes more to the variance explained by the principal component. This plot displays similar directions of vectors and strengths of dimensions to the biplot containing contribution values. The player efficiency rating, box plus/minus, and win share variables are the strongest contributors to PC1. The rebound and block percent variables are the strongest contributors to PC2. This may suggest that PC1 is a metric of primarily offensive performance and PC2 best represents defensive performance. This can be further explored by clustering individual players and projecting them onto the PC space. Then, the skillsets of individual players can be evaluated and compared to other known metrics.

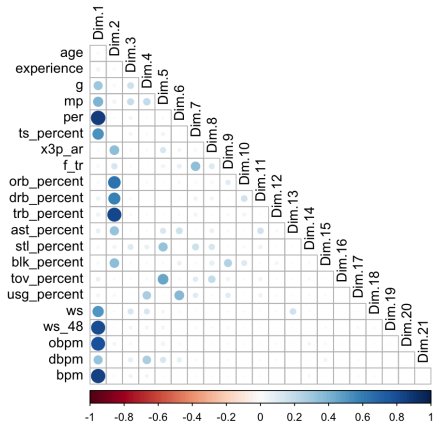


Figure 4 (left)

The first correlation plot (figure 4) displays the quality of contribution, the cosine squared values, of the original variables and each PC through dot size and color. This aligns with the findings of the biplots. Player efficiency rating and box plus/minus are the biggest contributors to PC1, and total rebound and blockage percent are the biggest contributors to PC2.

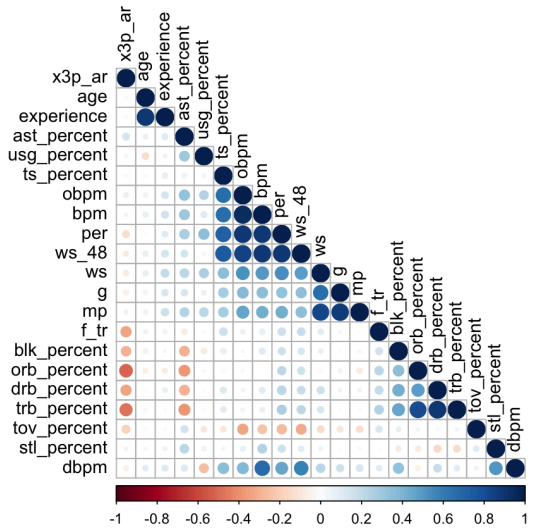
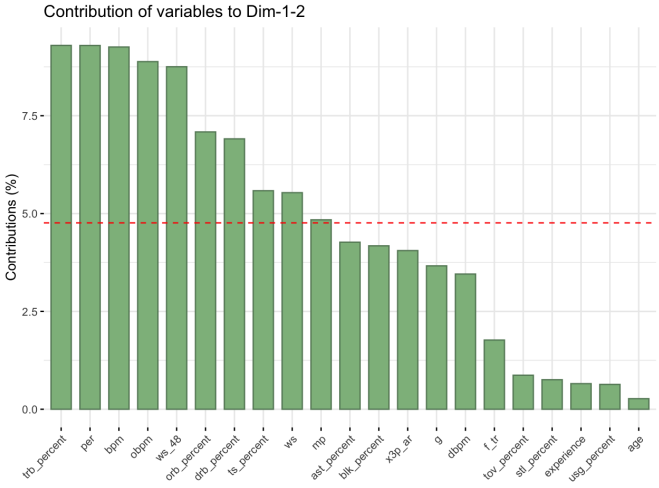


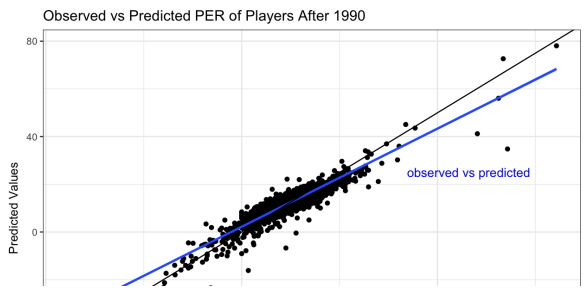
Figure 5 represents the correlations between the original dimensions using dot size and color. This
plot is ordered by their hierarchical cluster, so similar dimensions are located near each other in the plot. The clusters in the plot align with the associations between the variables and the PC's. The variables that contribute to PC1 are located close together, and the variables that contribute to PC2 are located in another cluster.
Figure 5 (left)

Figure 6 (below)

A barplot of the contributions of each variable (figure 6) is another way to visualize the relationship between the original variables and the PC's. However, in this plot, the contribution represents the variables' representation across both PC's. This highlights the key variables in the original data. The dimensions with high percent contributions to the principal components are total rebound percent, player efficiency rating, box plus/minus, and win shares.

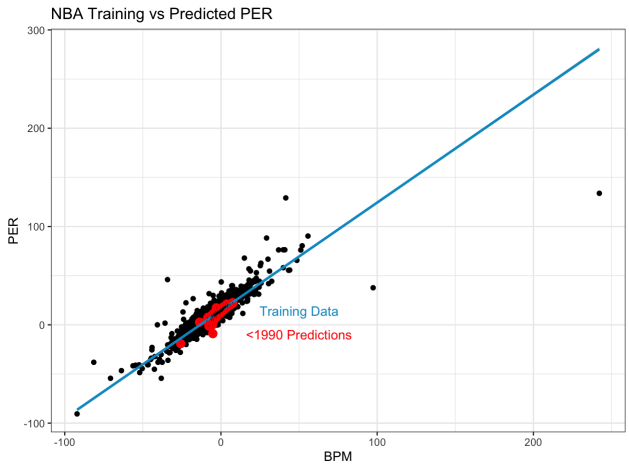


Many advanced or modern metrics of player success were not recorded in the early days of the NBA. One key performance metric, revealed through PCA, is player efficiency rating (PER). Before 1990, this variable was rarely recorded or calculated. A linear model was generated to predict the values of player efficiency rating for players in seasons before 1990. The model was trained on a random sample of values of PER for 70% of the observations after

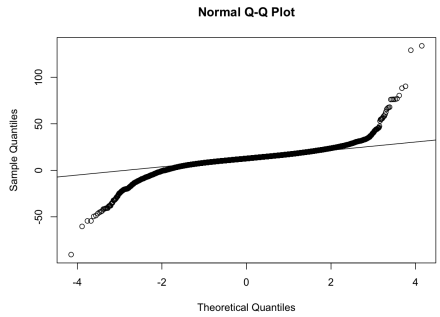


1990. Metrics with the strongest associations with PER, BPM and win shares, are the predictor variables. The model was tested on a testing dataset containing the remaining observations of PER after 1990 after the training data was removed. The root mean square percentage error of the model is 3.5%, and the R2 value is 0.8227. The model explains a high amount of the variance in PER, and the RMSPE is relatively low.

A plot of the observed vs predicted values of PER after 1990 (figure 7, above) displays a 1:1 line and the line representing predictions by the model. The model does not follow the 1:1 line, but it is relatively close to the 1:1 line.



The model was used to predict values of PER given predictor variables from a dataset before 1990. This is represented on a plot of bpm (one predictor variable) and PER. The red dots represent predictions, and the black dots are observations after 1990. The values of the predictor values are not outside of the range of the training predictor variables. ***Figure 8 (left)***



One assumption of linear regression is the normality of the y-variable. A QQplot represents the theoretical quantiles of a normal distribution against the sample quantiles of the response variable. Ideally, the observations form a diagonal line. The observations in this sample are not normal, as they form a distinct, nonlinear pattern on the QQplot. This model may benefit from a transformation of the y variable to fulfill the assumption that the y-variable is normally distributed.

Figure 9 (above), Figure 10 (below)

The key performance metrics that contribute the most to the variance in the dataset, found by using PCA, are PER, win shares, box plus/minus, and total shot percentage. To explore the variation in these variables, along with usage percent, across different positions, the average of each metric by position was plotted on the y-axis with

the most common positions on the x-axis (figure 10). A dot plot allows for easy comparison between the different positions, as each metric operates on a unique scale and the values do not overlap.

The win shares and true shooting percentages are relatively evenly distributed about the different positions. The highest player efficiency rating exists for center positions, which aligns with the role of a center on the court. They are closer to the basket, making it easier to score efficiently. Point guards have the highest average box plus/minus, which aligns with their larger understanding of the game and their advanced control of offense and read of defense.

The use of PCA in this analysis allowed key performance metrics to be uncovered by portraying their contributions to each primary component. The PC's can be used to cluster players by skillset and evaluate performance and expertise using the key metrics. It also provides a deep, efficient analysis using the most important variables. Generating a linear model would not utilize the dimensions in order of contribution to variance, and it would be far more tedious to combine linear models to reach an equivalent level of accuracy.

The PCA of the NBA database provided key performance metrics, such as player efficiency rating, box plus/minus, and win shares. The generated linear model was not successful in predicting values of PER before 1990, as the assumption of normality of the response variable was not met. The PCA provided metrics on which to compare performance of different positions, and the generated insights aligned with real-world observations of the positions.

The abundance of data about professional basketball invites further statistical analyses to be explored. Data-driven analytics in sports is vital to understanding and improving performance and maximizing success.

Glossary

Box Plus/Minus (BPM) - estimate of points per 100 possessions

Player Efficiency Rating (PER) - A calculation intended to capture all of the measures of a player's accomplishments as a per-minute rating of performance (developed by ESPN)

Win Shares (WS) - estimate of number of wins a player helped create

Appendix

Data Cleaning:

```r
##cleaning=========================================================================

#subset for only nba
nba <- subset(advanced, advanced$lg == "NBA")
View(nba)
nrow(nba)


#missing values
sum(is.na(nba))
nrow(nba)

#see missing vals proportion first half vs last half (season)

#find 1/2 year (~1990)
is.numeric(nba$season)
min(nba$season) + ((max(nba$season)-min(nba$season))/2)

#na before and after
sum(is.na(nba[nba$season <= 1990,]))
sum(is.na(nba[nba$season > 1990,]))

#proportion
sum(is.na(nba[nba$season < 1990,]))/sum(is.na(nba))
sum(is.na(nba[nba$season >= 1990,]))/sum(is.na(nba))

### 99% before 1990 - work with only more recent than 1990
nba <- nba[nba$season >= 1990,]

#na by column
colSums(is.na(nba))
max(colSums(is.na(nba)))

#columns
colnames(nba)

#remove: seas_id, birth_year, lg
nba <- nba[, -c(1, 5, 9, 25, 26)]
colnames(nba)
```

PCA:

```r
#numeric cols
colnames(nba)
sapply(nba, class)
nba.num <- na.omit(nba[, c(5, 6, 9:27)])
sum(is.na(nba.num))
colnames(nba.num)

nba.pca <- prcomp(nba.num, scale. = T)

scores <- nba.pca$x
loading <- nba.pca$rotation
```

Figure 1 (Scree)

```
#scree
fviz_eig(nba.pca, main = "Scree Plot for PCA of NBA players")

nba.eigval <- get_eigenvalue(nba.pca)
```

Figure 2 (Biplot - Contributions)

```
#Biplots

#contribution
fviz_pca_var(nba.pca,
             col.var = "contrib", repel = T,
             gradient.cols = c("darkolivegreen4", "steelblue", "darkblue"))
)
```

Figure 3 (Biplot - Cos^2)

```
#cos2
fviz_pca_var(nba.pca,
             col.var = "cos2",
             gradient.cols = c("steelblue", "indianred", "goldenrod"))
```

Figure 4 (correlation plot - cos^2)

```
#get pca variables
nba.var <- get_pca_var(nba.pca)

#corrplot of cos2 vals
corrplot(nba.var$cos2,
         type = "lower",
         tl.col = "black")
```

Figure 5 (correlation plot - dimensions)

```
corrplot(cor(nba.num), type = 'lower', order = 'hclust',
         tl.col = 'black', tl.srt = 90)
```

Figure 6 (bar plot - dimension contribution)

```r
#barplot - contrib of each variable

nba.bar1 <- fviz_contrib(nba.pca, choice = "var",
                         axes = 1:2,
                         fill = "darkseagreen",
                         color = "darkseagreen4")
nba.bar1

```

Linear model generation:

```r
#nba2: dataframe for players before 1990
nba2 <- nba1[nba1$season <= 1990,]
nba2 <- nba2[rowSums(is.na(nba2[, -10])) == 0, ]

set.seed(987)
training.size <- floor(0.8*nrow(nba))
train.individuals <- sample(seq_len(nrow(nba)), size = training.size)

training.data <- nba[train.individuals, ]
testing.data <- nba[-train.individuals, ]

lm1 <- lm(data = training.data, per ~ bpm + ws_48)
summary(lm1)

testing.data$prediction <- round( predict(lm1,
                                          newdata = testing.data), 2)

nba2$predper <- round(predict(lm1,
                              newdata = nba2), 2)

rmspe.nba <- sqrt(mean(testing.data$per - testing.data$prediction)^2)
rmspe.nba
```

Figure 7 (Observed vs predicted)

```
#pred v observed
p2.obs_v_pred <- (
  ggplot(data = testing.data, aes(x = per, y = prediction)) +
    geom_point() +
    geom_smooth(method = "lm") +
    geom_abline(slope = 1, intercept = 0) +
    labs(x = "Observed Values", y = "Predicted Values",
         title = "Observed vs Predicted PER of Players After 1990",
         caption = "Figure 2") +
    annotate("text", -20, -40, label = "1:1 line", size = 5) +
    annotate("text", 58, 25, label = "observed vs predicted",
             col = "blue", size = 4) +
    theme_bw())
plot(p2.obs_v_pred)
```

Figure 8 (training vs predictions)

```
p1.train_v_pred <- (
  ggplot() +
    geom_point(data = nba, aes(x = bpm, y = per)) +
    geom_point(data = nba2, aes(x = bpm, y = predper),
               col = "red", size = 3) +
    geom_smooth(data = nba, aes(x = bpm, y = per),
                method = "lm", col = "deepskyblue3") +
    geom_errorbar(data = nba2, aes(x = bpm,
                                   ymin = predper - rmspe.nba,
                                   ymax = predper + rmspe.nba)) +
    labs(x = "BPM", y = "PER",
         title = "NBA Training vs Predicted PER") +
    annotate("text", 50, 14,label = "Training Data", color = "deepskyblue3") +
    annotate("text", 50, -10, label = "<1990 Predictions", color = "red") +
    theme_bw())
plot(p1.train_v_pred)

```

Figure 9 (normality of PER)

```
#test normality of y
qqnorm(nba1$per)
qqline(nba1$per)

```

Figure 10 (Performance by Position)

```r
ggplot(data = pm.by.pos, aes(x = pos)) +
  geom_point(aes(y = ws, color = "Win Shares"),
             size = 4) +
  geom_line(aes(y = ws, group = 1, color = "Win Shares"),
            lty = 3) +
  geom_point(aes(y = per, color = "PER"),
             size = 4) +
  geom_line(aes(y = per, group = 1, color = "PER"),
            lty = 3) +
  geom_point(aes(y = usg, color = "Usage Percent"),
             size = 4) +
  geom_line(aes(y = usg, group = 1, color = "Usage Percent"),
            lty = 3) +
  geom_point(aes(y = bpm, color = "Box Plus/Minus"),
             size = 4) +
  geom_line(aes(y = bpm, group = 1, color = "Box Plus/Minus"),
            lty = 3) +
  geom_point(aes(y = tsp, color = "True Shooting Percent"),
             size = 4) +
  geom_line(aes(y = tsp, group = 1, color = "True Shooting Percent"),
            lty = 3) +
  scale_color_manual(values =
                       c("Win Shares" = "firebrick1",
                         "PER" = "dodgerblue",
                         "Usage Percent" = "orange",
                         "Box Plus/Minus" = "darkolivegreen4",
                         "True Shooting Percent" = "purple")) +
  labs(title = "Measures of Performance by Position",
       x = "Position",
       y = "Value",
       color = "Performance Metric") +
  theme_minimal()
```

Creation of pm.by.pos

```r
#all together!
avg.ws <- aggregate(nba.perf, ws ~ pos, mean)
avg.per <- aggregate(nba.perf, per ~ pos, mean)
avg.usg <- aggregate(nba.perf, usg_percent ~ pos, mean)
avg.bpm <- aggregate(nba.perf, bpm ~ pos, mean)
pm.by.pos <- avg.ws
pm.by.pos$per <- avg.per$per
pm.by.pos$usg <- avg.usg$usg_percent
pm.by.pos$bpm <- avg.bpm$bpm
is.data.frame(pm.by.pos)
str(pm.by.pos)

#add ts%
avg.tsp <- aggregate(nba.perf, ts_percent ~ pos, mean)
pm.by.pos$tsp <- avg.tsp$ts_percent

```

Works Cited

Glossary. Basketball. (n.d.). https://www.basketball-reference.com/about/glossary.html

Sumitro Datta. December 2024. NBA Stats (1947-present), Version 40. Retrieved December 11, 2024 from https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats.