Vaulting into Victory: Optimizing for US Gymnastics Medal Count at the Paris 2024 Olympics

Benjamin Thorpe

12-16-2023

```
library(readr)
library(dplyr)
library(tidyr)
library(stringr)
library(knitr)
library(lme4)
library(ggplot2)
library(kableExtra)
```

```
nba_data <- read_csv("NBA_Play_Types_16_23.csv")

# Sort the data
nba_data <- nba_data |>
  arrange(PLAYER_ID, PLAY_TYPE, SEASON) |>
  filter(PLAY_TYPE %in% c("Spotup", "PnR Ball-Handler", "Isolation"))

# Creating the next season's PPP and PPP_PCTL columns
nba_data <- nba_data |>
  group_by(PLAYER_ID, PLAY_TYPE) |>
  mutate(PPP_next = lead(PPP),
         PPP_PCTL_next = lead(PPP_PCTL)) |>
  ungroup()

# Handling the 2022-23 season by replacing with NA
nba_data_clean <- nba_data |>
  mutate(PPP_next = ifelse(SEASON == "2022-23", NA, PPP_next),
         PPP_PCTL_next = ifelse(SEASON == "2022-23", NA, PPP_PCTL_next),
         PPP_change = PPP_next - PPP) |>
  filter(POSS > 19,
         !is.na(PPP_next)) |>
  select(c(SEASON, PLAYER_NAME, PLAY_TYPE, TEAM_ABB, FREQ, PPP, PPP_PCTL,  GP, EFG_PCT, FTA_FREQ, TOV_FREQ
```

Steps: 1. Decide on case study idea - What teams best develop spotup shooters, effective cutters, and efficient PnR ball-handlers 2. Write introduction on why this is important + background info about the topic 3. Describe data source in detail (where I found it, website it's from, companies involved, what key variables mean, etc.) 4. Do 3-5 visualizations explaining some things about the data - use past project as inspiration 5. Write goals of the study + any expectations I have 6. Come up with model structure I will use to accomplish the goals - multilevel linear model where level one is player's attributes (num possessions, age, PPP, frequency) and level two is the team they are on that year, response will be PPP_PCTL_next to account for league-wide efficiency changes and will examine team-level coefficients 7. Write methodology section using the above 8. Run the model and visualize results nicely in a table 9. Write a couple results paragraphs 10. Write conclusion, limitations and future study, and overall summary of the case

Introduction

In the contemporary landscape of the National Basketball Association (NBA), the ability of teams to develop players' offensive skills has become a focal point of team strategy and success. This emphasis stems from the evolving nature of the game, where player's offensive versatility and efficiency have become paramount to earning playing time. As the pace of play has increased in every season but one going back to 2012-2013 (Scaletta), having players who can maintain strong efficiency numbers while the volume of shots taken continues to increase year over year is essential for teams looking to be successful. Additionally, as the use of analytics has become widely accepted in the league over the past 15 years, the number of three point shot attempts taken every year has went up in each season since 2010-2011 (Wal). This reliance on shooting has meant a transition to more guard-heavy and "small-ball" lineups and a decline in teams building around a dominant big man, which is demonstrated by guards "using" almost 50% of offensive possesions in the NBA as of 2018 (Thinking). Here, this usage rate is defined as "the proportion of possessions used by a player by either shooting, winning free throws, or committing a turnover" (Thinking).

With these trends in mind, I want to determine if any teams stand out, either positively or negatively, in developing their players' offensive game. I will be focusing on the development of more guard-related attributes in this study, however as the league continues to see an influx of taller athletes with guard-like skills, it should be noted that these abilities can be improved upon by all NBA players. Also, I am interested in how a set of player-level variables affect the development of each of these skills, and this will be an additional question I answer. In regards to which skills I will focus on, I will be examining the play types of isolation, pick and roll ball handler, and spot-up shooter to try and identify which teams are the best at developing these specific skills, as each is an important part of offense in the NBA today. Isolation plays, which rely on a player's ability to score one-on-one, demand high skill levels in ball-handling, shooting, and decision-making. Similarly, proficiency in the pick and roll as a ball handler necessitates a blend of vision, timing, and scoring ability, making it a cornerstone of modern offensive systems. Meanwhile, the spot-up shooter role, critical for spacing the floor and capitalizing on defensive lapses, hinges on precise shooting and quick decision-making. The teams that excel in cultivating these skills not only enhance their offensive firepower but also gain a tactical edge. This ability to develop offensive prowess in players aligns with the league's shift towards a more dynamic, fast-paced, and perimeter-oriented style of play, as discussed above. Consequently, identifying which NBA teams are best at developing these specific skills provides insights into their potential for long-term success and adaptability in an ever-changing basketball environment. This analysis not only reflects on the teams' coaching and training methodologies but also on their talent scouting and player utilization strategies, making it a multifaceted and critical aspect of understanding team dynamics and future prospects in the league.

Data Description

For my analysis, I am using the "NBA_Play_Types_16_23" dataset available on Dominic Samangy's github page (DomSamangy). Dominic is currently an analyst for the New Orleans Pelicans and has over time created a great set of sports analytics resources on a google sheet called "Guide to Sports Analytics", and I found this set of data, which he scraped and uploaded into a csv, from the "Data Resources" tab.

The data itself comes from NBA.com, which has an immense database of all different kinds of basketball statistics from the NBA. The play type data I will be utilizing here is via Synergy Sports (though publicly available on NBA.com), and at a high level contains data describing how well players perform when they are involved in different play types over the course of a season, and comprises data from the 2015-2016 to 2022-2023 seasons. A row will contain a player's identifying information, the season which the data is from, a certain play type, their points per possession (PPP, explained in more detail below) and its associated percentile in that season, and multiple other metrics such as games played, effective field goal percentage (shooting percentage that adjusts for 3-point shots being worth 50% more), percentage of times free throws were attempted, and percentage of times a turnover was committed, which I will use as variables I hold constant in my model. From domain knowledge I believe each of the above statistics when combined can give a good overview of the offensive level a player is at, and thus are important to account for when judging team's player development abilities. There are also a few other statistics that I will not be using in my analysis. There are 11 different play type categories, however I will only be focusing on three as mentioned in the introduction to specifically look at a few key areas of offense. How efficient a player is at a given play type is defined here by their PPP, which is the average points per possession of that player, in that play type, in that year. So if on all possessions in a season where a player takes a catch and shoot shot their team scores an average of 1.2 points, this would be their PPP for their "spotup" row. The change in PPP from one year to the next will be my response variable.

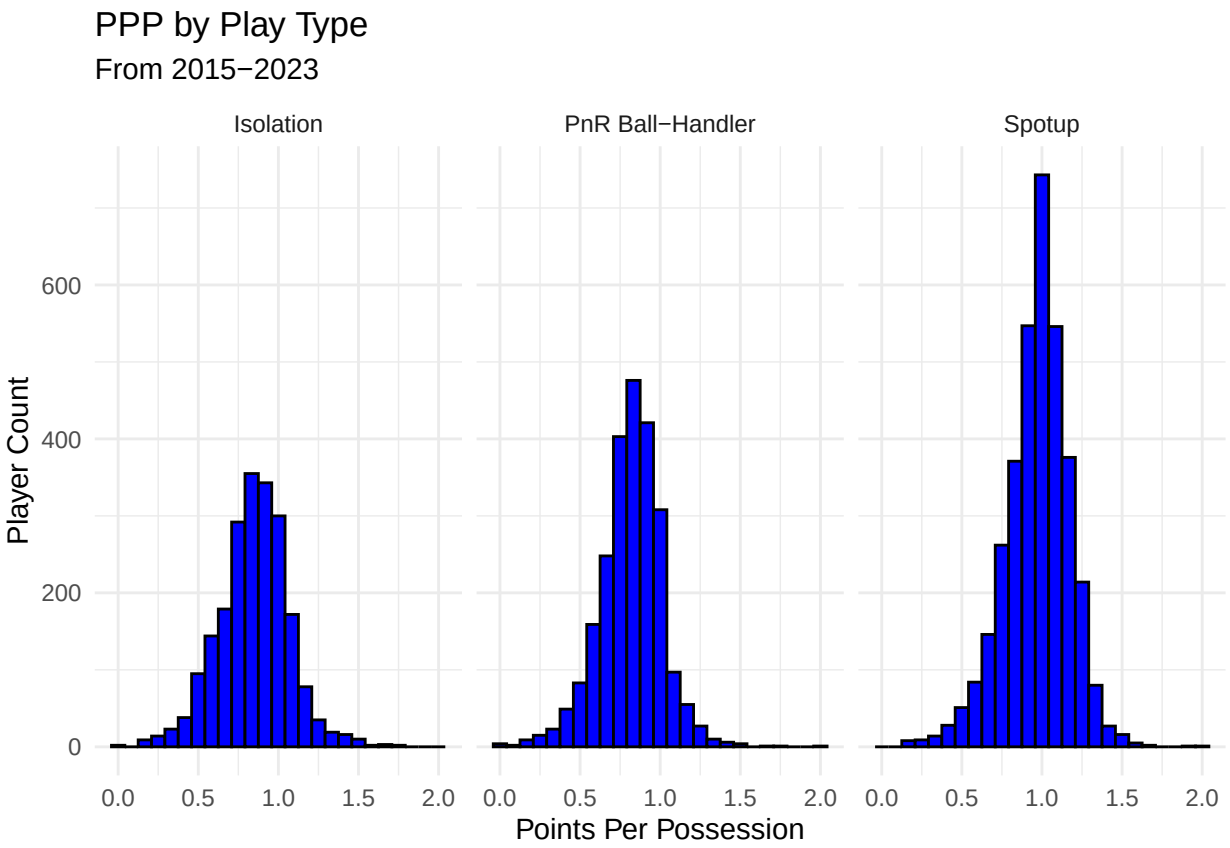
As far as data cleaning, I mostly kept the dataset as is besides filtering for my three play types of interest. The only additional change I made was excluding rows where a player had fewer than 20 possessions of a certain play type in

a given season, as below this mark there were many outliers that may have skewed my findings (though this was chosen arbitrarily). My goal here is to determine which teams have the best offensive development program, and including data on extremely small sample sizes from players would not be beneficial as my response variable is an average over a season, and thus does not take frequency into account since I want to maintain a focus on efficiency. This resulted in a final dataframe with 5,193 observations which I used to complete my analysis.

Exploratory Data Analysis

First, let's visualize the PPP distirbution for each play type:

```
ggplot(nba_data, aes(x = PPP)) +
  geom_histogram(bins = 25, fill = "blue", color = "black") +
  facet_wrap(~ PLAY_TYPE) +
  theme_minimal() +
  labs(title = "PPP by Play Type",
       subtitle = "From 2015-2023",
       x = "Points Per Possession",
       y = "Player Count")
```

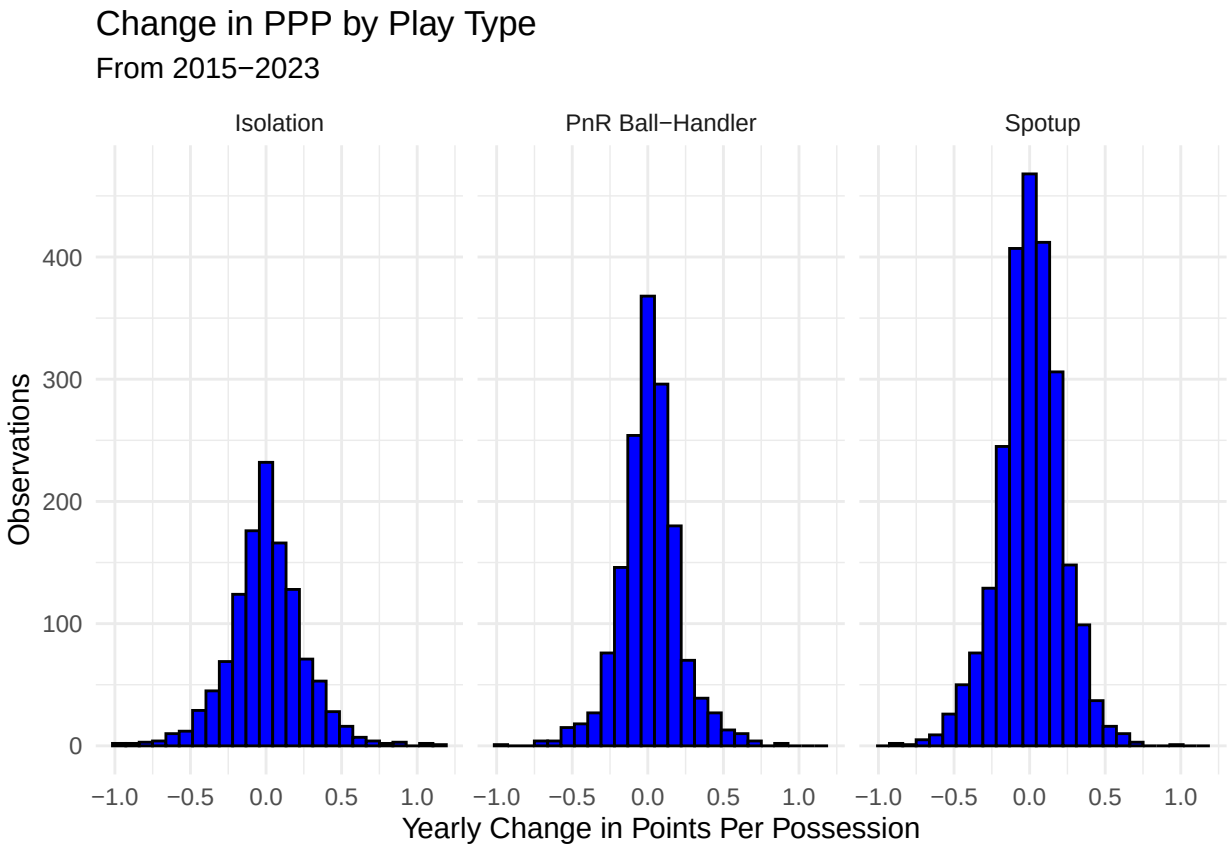


Here, we can see that each play type approximately follows a normal distribution, with the most players qualifying for the spotup 20 possession cutoff, followed by the pick and roll ball handlers, and slightly below that in terms of sample size comes the isolation observations. Additionally, we can see that catch and shoot opportunities (spotups) have the highest average PPP of around 0.95, with the other two play types of interest having an average PPP closer to 0.75. This finding follows conventional wisdom that jump shots off the catch are one of the best shots you can get on offense.

Now, let's look at how a player's skill in each of these three categoried changes, on average, from year to year:

```
ggplot(nba_data_clean, aes(x = PPP_change)) +
  geom_histogram(bins = 25, fill = "blue", color = "black") +
  facet_wrap(~ PLAY_TYPE) +
  theme_minimal() +
```

```
  labs(title = "Change in PPP by Play Type",
       subtitle = "From 2015-2023",
       x = "Yearly Change in Points Per Possession",
       y = "Observations")
```



Somewhat surprisingly, these histograms exhibit the finding that the majority of NBA players do not improve on these skills year-to-year. I would have expected the center of the distributions to be sightly higher than zero with players working year-round to improve their game, however clearly this is not the case. This makes the priority of my research even more important, as consistently having players get better over time will give teams able to achieve this a substantial leg up over their opponents.

As the final part of my EDA, I wanted to take a look at an example of what good player development looks like.

```
# Assuming reshaped_data is already created as before
reshaped_data <- nba_data %>%
  filter(PLAYER_NAME == "Tyrese Haliburton",
         SEASON %in% c("2020-21", "2022-23")) %>%
  select(SEASON, PLAY_TYPE, PPP, POSS) %>%
  pivot_longer(cols = c(PPP, POSS), names_to = "Metric", values_to = "Value") %>%
  unite("Season_Metric", SEASON, Metric, sep = "_") %>%
  pivot_wider(names_from = Season_Metric, values_from = Value, names_sort = TRUE)

# Rearrange the columns so that PPP and POSS for each season are adjacent
column_order <- c("PLAY_TYPE",
                  "2020-21_PPP", "2020-21_POSS",
                  "2022-23_PPP", "2022-23_POSS")
reshaped_data <- reshaped_data %>% select(all_of(column_order))

# Create a kable table
kable_table <- kable(reshaped_data, caption = "<span style='font-size: 24px; color: black;'>Tyrese Haliburt
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F) %>%
  column_spec(1, bold = T)
```

Table 1: Tyrese Haliburton

PLAY_TYPE	2020-21_PPP	2020-21_POSS	2022-23_PPP	2022-23_POSS
Isolation	0.70	27	0.91	105
PnR Ball-Handler	0.95	283	1.04	495
Spotup	1.13	187	1.20	128

```
# Print the kable table
kable_table
```

The player I chose to look at here is Tyrese Haliburton, a point guard for the Indiana Pacers who was drafted 12th in 2020. Above, we can see how both the efficiency and number of possessions in each of our three play types has increased as he has gained more experience in the NBA between his rookie season and year #3. There are always going to be other factors in play as I will discuss in my limitations section, however this still reflects positively on the Pacers coaching staff, especially considering he was traded to the team in the middle of the 2021-2022 season, so they only have had one off-season to work with him prior to the start of the 2022-2023 season. Additionally, although the dataset does not include any numbers from the current season, Haliburton has been one of the most improved players in the league and one of its best offensive players this year. This is why I wanted to look at his development as an instance of a strong success, as when he was drafted noone anticipated his ability to make an impact of this magnitude and thus this demonstrates the importance of a team's ability to develop talent.

Methodology

I am employing a random intercepts linear model to analyze the influence of team-level factors on individual player development in offensive play among three play types: isolations, the ball handler in pick and rolls, and spotups. This modeling choice is driven by the hierarchical nature of the data, where player-level statistics (level one predictors) are nested within teams (level two random intercepts). The response variable in this analysis is a player's Points Per Possession (PPP) for a specified play type, which, as established in the exploratory data analysis (EDA), follows a normal distribution.

At the first level of the model, the predictor variables are drawn from a player's statistics related to their involvement in a particular play type within a single season. These variables are chosen for their relevance in quantifying a player's current skill level and stage in their career, offering a comprehensive baseline for assessing their offensive capabilities.

PPP (Points Per Possession): This metric is a direct measure of a player's offensive efficiency. Including it as a predictor allows for assessing how past performance in terms of scoring efficiency influences future performance, under the premise that a player's past efficiency is indicative of their skill level.

GP (Games Played): This variable indicates the extent of a player's involvement in the game. It serves as a proxy for experience and endurance, assuming that more game time translates to more opportunities for development and contribution.

EFG_PCT (Effective Field Goal Percentage): This statistic adjusts the standard field goal percentage to account for the fact that three-point field goals are worth more than two-point field goals. It's a crucial measure of shooting efficiency, reflecting a player's ability to score points more effectively.

FTA_FREQ (Free Throw Attempt Frequency): This represents the frequency of free throw attempts relative to the player's overall play. It's an important aspect of scoring efficiency, highlighting a player's ability to draw fouls and capitalize on free-throw opportunities.

TOV_FREQ (Turnover Frequency): This is the frequency of turnovers per play. It's essential to consider this in evaluating a player's offensive development, as minimizing turnovers is key to maximizing scoring opportunities and maintaining offensive efficiency.

FREQ (Frequency): This measures how often a player is involved in the specified play type. It indicates a player's role and involvement in the team's offensive strategies, providing context to their PPP.

At the second level, the random intercept for each team (TEAM_ABB) accounts for the team's overarching influence on a player's development. This aspect of the model captures the effect of team-specific factors such as coaching and

team dynamics that are pivotal in player development but are not directly measured in the dataset.

By incorporating these player-level variables and a team-level random intercept, the model aims to isolate the impact of a team's ability to enhance offensive skills, controlling for individual player attributes and career stage factors. This approach allows for a nuanced understanding of how different teams contribute to or detract from their players' offensive development in the NBA.

The full model in mathematical notation is displayed below:

$$PPP_{next_{ij}} = \beta_0 + \beta_1 \times PPP_{ij} + \beta_2 \times GP_{ij} + \beta_3 \times EFG\_PCT_{ij} + \beta_4 \times FTA\_FREQ_{ij} + \beta_5 \times TOV\_FREQ_{ij} + \beta_6 \times FREQ_{ij} + u_{0j} + \varepsilon_{ij}$$

Results

```
# Fit the random intercept model
model <- lmer(PPP_next ~ PPP + GP + EFG_PCT + FTA_FREQ + TOV_FREQ + FREQ + (1 | TEAM_ABB), data = nba_data_

# View the summary of the model
summary(model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: PPP_next ~ PPP + GP + EFG_PCT + FTA_FREQ + TOV_FREQ + FREQ +
##     (1 | TEAM_ABB)
##    Data: filter(nba_data_clean, PLAY_TYPE == "Isolation")
##
## REML criterion at convergence: -444.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2775 -0.5564  0.0354  0.6325  4.5822
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  TEAM_ABB (Intercept) 0.000316 0.01778
##  Residual             0.039055 0.19762
## Number of obs: 1189, groups:  TEAM_ABB, 30
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)  0.7134006  0.0455248  15.671
## PPP          0.3146628  0.1452124   2.167
## GP          -0.0001748  0.0003529  -0.496
## EFG_PCT     -0.3516309  0.2428799  -1.448
## FTA_FREQ    -0.1185222  0.1412637  -0.839
## TOV_FREQ     0.1065693  0.1758904   0.606
## FREQ         0.5678768  0.1168308   4.861
##
## Correlation of Fixed Effects:
##           (Intr) PPP    GP     EFG_PC FTA_FR TOV_FR
## PPP       -0.522
## GP        -0.500 -0.027
## EFG_PCT    0.364 -0.969  0.045
## FTA_FREQ   0.244 -0.719 -0.008  0.695
## TOV_FREQ -0.555  0.756 -0.027 -0.734 -0.556
## FREQ     -0.083 -0.092  0.007  0.053 -0.025  0.004
```

```
random_effects <- ranef(model)$TEAM_ABB

# Convert to a data frame
```

```r
random_effects_df <- as.data.frame(random_effects)

# Reset the row names to create a column with the team abbreviations
random_effects_df$TEAM_ABB <- rownames(random_effects_df)

# Rename the intercept column for clarity
colnames(random_effects_df)[1] <- "Random_Intercept"

# Rearrange the columns for readability
random_effects_df <- random_effects_df[, c("TEAM_ABB", "Random_Intercept")]

# Sort the dataframe by the random intercept values
random_effects_df <- random_effects_df[order(random_effects_df$Random_Intercept, decreasing = TRUE), ]

# View the table
random_effects_df
```

```
##     TEAM_ABB Random_Intercept
## PHI      PHI     0.0148840404
## GSW      GSW     0.0117671881
## ATL      ATL     0.0113956610
## CHA      CHA     0.0111879513
## HOU      HOU     0.0098843475
## DEN      DEN     0.0067506732
## NYK      NYK     0.0065554687
## OKC      OKC     0.0064214293
## CHI      CHI     0.0053635482
## MIN      MIN     0.0053578698
## WAS      WAS     0.0048876388
## SAC      SAC     0.0036669128
## POR      POR     0.0036295272
## MIL      MIL     0.0030247894
## BKN      BKN     0.0023355236
## IND      IND     0.0021846279
## CLE      CLE    -0.0002180384
## UTA      UTA    -0.0021875785
## MIA      MIA    -0.0021904462
## TOR      TOR    -0.0040963189
## LAC      LAC    -0.0041796082
## ORL      ORL    -0.0043970026
## NOP      NOP    -0.0062053311
## SAS      SAS    -0.0071727021
## BOS      BOS    -0.0090431768
## LAL      LAL    -0.0102831081
## PHX      PHX    -0.0133315982
## MEM      MEM    -0.0137600037
## DAL      DAL    -0.0142735919
## DET      DET    -0.0179586925
```

Citations

Scaletta - https://www.lineups.com/articles/why-nba-game-pace-is-at-historic-high/ Wal - https://medium.com/@gwal325/how-the-nba-has-changed-in-the-past-20-years-and-insights-to-win-23f8e9f17643 Thinking Machines Data Science - https://stories.thinkingmachin.es/nba-in-30-years/