library(tidymodels)

library(randomForest)

library(dplyr)

library(ggplot2)

# Read the datasets

player\_awards <- read.csv("Player Award Shares.csv")

advanced\_stats <- read.csv("Advanced.csv")

# Check the unique values in the key columns

print(unique(advanced\_stats$player))

print(unique(player\_awards$player))

print(unique(advanced\_stats$season))

print(unique(player\_awards$season))

# Merge the datasets

combined\_data <- merge(advanced\_stats, player\_awards, by = c("season", "player"))

# Rename 'tm.x' to 'team' in combined\_data

combined\_data <- combined\_data %>%

rename(team = tm.x)

team\_abbrev <- read.csv("Team Abbrev.csv")

team\_abbrev <- team\_abbrev %>%

rename(team\_name = team, team = abbreviation)

# Merge combined\_data with team\_abbrev

combined\_data <- merge(combined\_data, team\_abbrev, by = c("season", "team"))

combined\_data <- combined\_data %>%

select(-c(age.y, tm.y, seas\_id.y, player\_id.y, lg.y))

colnames(combined\_data) <- toupper(colnames(combined\_data))

# Inspect the first few rows of the merged data

head(combined\_data)

seasons\_2015to2022 <- final\_combined\_data %>%

filter(season >= 2015, season <= 2022)

seasons\_2015to2022$dpoy\_winner <- with(seasons\_2015to2022, as.factor(award == "dpoy" & winner == TRUE))

head(seasons\_2015to2022)

seasons\_2015to2022 <- seasons\_2015to2022 %>%

filter(!is.na(season) & !is.na(award) & !is.na(winner) &

!is.na(dws) & !is.na(blk\_percent) & !is.na(stl\_percent) &

!is.na(trb\_percent) & !is.na(dbpm) & !is.na(playoffs))

seasons\_2015to2022 <- seasons\_2015to2022 %>%

select(-birth\_year)

colnames(seasons\_2015to2022) <- toupper(colnames(seasons\_2015to2022))

# Check the updated dataset

head(seasons\_2015to2022)