BasketballWorksheet

# Introduction

This worksheet guides you through analyzing how NBA player roles and scoring have changed from the 1990s to the 2020s. You’ll explore trends in points, assists, rebounds, and 3-point shooting using real per-game stats from four NBA seasons: 1995–96, 2004–05, 2014–15, and 2023–24.

Your final goal is to answer this question:

### How have NBA player roles and scoring styles evolved across the decades, and how can we use data and visual tools to understand that shift?

# Part 1: Data Setup and Warm-Up

library(readr)

Warning: package 'readr' was built under R version 4.4.3

library(dplyr)

Warning: package 'dplyr' was built under R version 4.4.3

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)

Warning: package 'ggplot2' was built under R version 4.4.3

nba\_all <- read\_csv("../Data/nba\_all.csv")

Rows: 2520 Columns: 34

── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (5): Player, Team, Pos, Awards, Player-additional  
dbl (29): Rk, Age, G, GS, MP, FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, eFG%...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Check structure  
glimpse(nba\_all)

Rows: 2,520  
Columns: 34  
$ Rk <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,…  
$ Player <chr> "Michael Jordan", "Hakeem Olajuwon", "Shaquille O'…  
$ Age <dbl> 32, 33, 23, 32, 30, 22, 23, 32, 25, 30, 33, 22, 24…  
$ Team <chr> "CHI", "HOU", "ORL", "UTA", "SAS", "WSB", "DAL", "…  
$ Pos <chr> "SG", "C", "C", "PF", "C", "PF", "SF", "PF", "C", …  
$ G <dbl> 82, 72, 54, 82, 82, 15, 18, 71, 70, 81, 76, 81, 82…  
$ GS <dbl> 82, 72, 52, 82, 82, 15, 18, 71, 70, 81, 76, 81, 82…  
$ MP <dbl> 37.7, 38.8, 36.0, 38.0, 36.8, 37.2, 37.2, 37.1, 38…  
$ FG <dbl> 11.2, 10.7, 11.0, 9.6, 8.7, 10.0, 8.1, 8.2, 8.0, 7…  
$ FGA <dbl> 22.6, 20.8, 19.1, 18.5, 16.8, 18.4, 21.3, 16.3, 15…  
$ `FG%` <dbl> 0.495, 0.514, 0.573, 0.519, 0.516, 0.543, 0.379, 0…  
$ `3P` <dbl> 1.4, 0.0, 0.0, 0.2, 0.0, 1.0, 1.9, 0.7, 0.1, 2.8, …  
$ `3PA` <dbl> 3.2, 0.2, 0.0, 0.5, 0.1, 2.3, 5.7, 2.5, 0.4, 6.4, …  
$ `3P%` <dbl> 0.427, 0.214, 0.500, 0.400, 0.333, 0.441, 0.343, 0…  
$ `2P` <dbl> 9.8, 10.6, 10.9, 9.4, 8.6, 9.0, 6.1, 7.5, 7.9, 4.8…  
$ `2PA` <dbl> 19.4, 20.6, 19.1, 18.0, 16.7, 16.1, 15.6, 13.9, 14…  
$ `2P%` <dbl> 0.506, 0.517, 0.573, 0.522, 0.517, 0.558, 0.391, 0…  
$ `eFG%` <dbl> 0.525, 0.515, 0.574, 0.524, 0.517, 0.571, 0.424, 0…  
$ FT <dbl> 6.7, 5.5, 4.6, 6.2, 7.6, 2.7, 5.4, 6.2, 7.0, 5.2, …  
$ FTA <dbl> 8.0, 7.6, 9.5, 8.6, 10.0, 4.6, 7.4, 8.0, 10.2, 6.1…  
$ `FT%` <dbl> 0.834, 0.724, 0.487, 0.723, 0.761, 0.594, 0.729, 0…  
$ ORB <dbl> 1.8, 2.4, 3.4, 2.1, 3.9, 2.5, 2.1, 3.4, 3.1, 0.7, …  
$ DRB <dbl> 4.8, 8.4, 7.7, 7.7, 8.3, 5.1, 3.3, 8.1, 7.3, 2.7, …  
$ TRB <dbl> 6.6, 10.9, 11.0, 9.8, 12.2, 7.6, 5.4, 11.6, 10.4, …  
$ AST <dbl> 4.3, 3.6, 2.9, 4.2, 3.0, 5.0, 2.8, 3.7, 2.3, 3.1, …  
$ STL <dbl> 2.2, 1.6, 0.6, 1.7, 1.4, 1.8, 0.8, 1.6, 1.0, 1.5, …  
$ BLK <dbl> 0.5, 2.9, 2.1, 0.7, 3.3, 0.6, 0.2, 0.8, 2.7, 0.2, …  
$ TOV <dbl> 2.4, 3.4, 2.9, 2.4, 2.3, 3.3, 3.1, 3.1, 3.7, 2.7, …  
$ PF <dbl> 2.4, 3.4, 3.6, 3.0, 3.2, 3.4, 2.2, 2.9, 3.5, 2.9, …  
$ PTS <dbl> 30.4, 26.9, 26.6, 25.7, 25.0, 23.7, 23.4, 23.2, 23…  
$ Awards <chr> "MVP-1DPOY-6ASNBA1DEF1", "MVP-4DPOY-5ASNBA2DEF2", …  
$ `Player-additional` <chr> "jordami01", "olajuha01", "onealsh01", "malonka01"…  
$ season <dbl> 1996, 1996, 1996, 1996, 1996, 1996, 1996, 1996, 19…  
$ Era <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 19…

# Warm up Questions

1. What columns are available in this data set?
2. What does the column Era represent, and how was it created?
3. How many total players are in the data set?

# Use tools like nrow(), unique(), table(), names(), etc.

# Summarize Key stats by Era

nba\_all |>   
 group\_by(Era) |>   
 summarise(  
 avg\_pts = mean(as.numeric(PTS), na.rm = TRUE),  
 avg\_ast = mean(as.numeric(AST), na.rm = TRUE),  
 avg\_trb = mean(as.numeric(TRB), na.rm = TRUE),  
 avg\_3p = mean(as.numeric(`3P`), na.rm = TRUE),  
 avg\_3pa = mean(as.numeric(`3PA`), na.rm = TRUE),  
 avg\_fg\_pct = mean(as.numeric(`FG%`), na.rm = TRUE)  
 )

# A tibble: 4 × 7  
 Era avg\_pts avg\_ast avg\_trb avg\_3p avg\_3pa avg\_fg\_pct  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 1990 7.98 1.86 3.43 0.453 1.27 0.444  
2 2000 7.78 1.74 3.43 0.463 1.32 0.427  
3 2010 7.80 1.81 3.35 0.632 1.87 0.427  
4 2020 8.02 1.93 3.24 0.909 2.58 0.448

#What Do You Notice 1. In which era was 3-point shooting highest on average?

1. Has assist or rebound volume changed significantly over time?
2. What does this suggest about how NBA offense has evolved?
3. Based on these averages, how would you describe the shift in player roles?

# Part 2: Visualizing Player Trends

Now you’ll create three visualizations to explore how NBA player performance has shifted across eras.

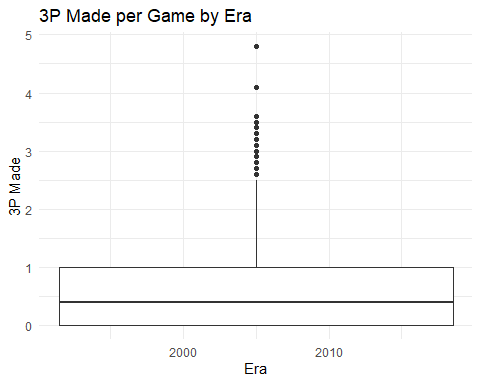
## Plot 1: Boxplot of 3P Made by Era

ggplot(nba\_all, aes(x = Era, y = as.numeric(`3P`), fill = Era)) +  
 geom\_boxplot() +  
 labs(title = "3P Made per Game by Era", y = "3P Made", x = "Era") +  
 theme\_minimal()

Warning: Continuous x aesthetic  
ℹ did you forget `aes(group = ...)`?

Warning: Removed 4 rows containing non-finite outside the scale range  
(`stat\_boxplot()`).

Warning: The following aesthetics were dropped during statistical transformation: fill.  
ℹ This can happen when ggplot fails to infer the correct grouping structure in  
 the data.  
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
 variable into a factor?



Which era has the highest median number of 3-pointers made?

Which era has the most variability?

How do these findings compare to your summary table from Part 1?

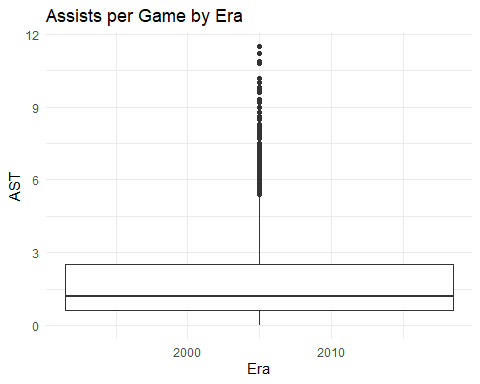
Plot 2: Boxplot of Assists by Era

ggplot(nba\_all, aes(x = Era, y = as.numeric(AST), fill = Era)) +  
 geom\_boxplot() +  
 labs(title = "Assists per Game by Era", y = "AST", x = "Era") +  
 theme\_minimal()

Warning: Continuous x aesthetic  
ℹ did you forget `aes(group = ...)`?

Warning: Removed 4 rows containing non-finite outside the scale range  
(`stat\_boxplot()`).

Warning: The following aesthetics were dropped during statistical transformation: fill.  
ℹ This can happen when ggplot fails to infer the correct grouping structure in  
 the data.  
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
 variable into a factor?



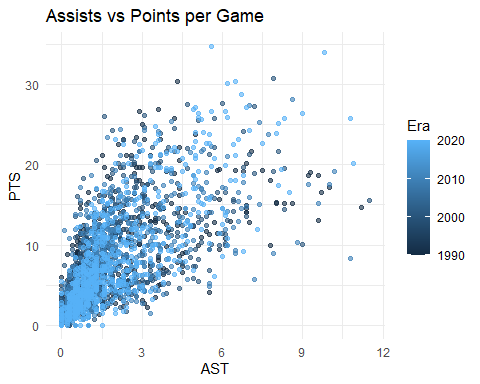
How have assist trends changed across eras?

What might this say about team strategy and ball movement?

Plot 3: Scatterplot of Assists vs Points

ggplot(nba\_all, aes(x = as.numeric(AST), y = as.numeric(PTS), color = Era)) +  
 geom\_point(alpha = 0.6) +  
 labs(title = "Assists vs Points per Game", x = "AST", y = "PTS") +  
 theme\_minimal()

Warning: Removed 4 rows containing missing values or values outside the scale range  
(`geom\_point()`).



Is there a visible relationship between assists and points per game?

Do you see differences in this relationship across eras?

# Part 3: Modeling Scoring with Regression

Now let’s explore what factors explain a player’s points per game using a linear regression model.

We’ll use assists, rebounds, and 3-pointers made as predictors.

## Build a Linear Model

model <- lm(PTS ~ AST + TRB + `3P`, data = nba\_all)  
summary(model)

Call:  
lm(formula = PTS ~ AST + TRB + `3P`, data = nba\_all)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-16.1241 -1.3918 -0.1122 1.0055 13.1701   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.48980 0.10115 -4.843 1.36e-06 \*\*\*  
AST 1.08721 0.03574 30.418 < 2e-16 \*\*\*  
TRB 1.30174 0.02304 56.509 < 2e-16 \*\*\*  
`3P` 3.18089 0.08554 37.186 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2.651 on 2512 degrees of freedom  
 (4 observations deleted due to missingness)  
Multiple R-squared: 0.7997, Adjusted R-squared: 0.7994   
F-statistic: 3343 on 3 and 2512 DF, p-value: < 2.2e-16

Which variables are statistically significant predictors of points per game?

What does the sign of each coefficient (positive or negative) suggest about its relationship to scoring?

How strong is the model overall? Look at the R-squared value.

What kinds of players might this model not describe well?

# Part 4: Clustering Player Roles

We’ll now group players into clusters based on similar stat profiles. This will help us understand different player “roles” across eras.

## Prepare the Data and Run K-Means

nba\_cluster <- nba\_all |>   
 mutate(  
 PTS = as.numeric(PTS),  
 AST = as.numeric(AST),  
 TRB = as.numeric(TRB),  
 `3P` = as.numeric(`3P`)  
 ) |>   
 filter(!is.na(PTS), !is.na(AST), !is.na(TRB), !is.na(`3P`))  
  
set.seed(123)  
k\_result <- kmeans(scale(nba\_cluster |> select(PTS, AST, TRB, `3P`)), centers = 4, nstart = 20)  
  
nba\_cluster$Cluster <- as.factor(k\_result$cluster)

How many players are in each cluster? Use table(nba\_cluster$Cluster).

Describe each cluster based on stat profiles:

Which cluster scores the most?

Which cluster has the most assists or rebounds?

What types of roles might these clusters represent? (e.g., scorer, playmaker, rebounder)

# Part 5: Build an Interactive Shiny App

In this section, you’ll help build a fully featured Shiny app that explores NBA player roles by era, position, and player clusters. The app uses filtered data to create plots and a table just like the one used in your own analysis.

## Step 1: Create a New File

Open a new R Script and save it as app.R.

## Step 2: Setup and Data

Paste this at the top of app.R:

library(shiny)  
  
# Load cleaned dataset with cluster assignments  
nba\_data <- read\_csv("../Data/nba\_all.csv")|>   
 mutate(  
 PTS = as.numeric(PTS),  
 AST = as.numeric(AST),  
 TRB = as.numeric(TRB),  
 `3P` = as.numeric(`3P`)  
 ) |>   
 filter(!is.na(PTS), !is.na(AST), !is.na(TRB), !is.na(`3P`))   
  
# Add clusters  
set.seed(123)  
k\_result <- kmeans(scale(nba\_data |> select(PTS, AST, TRB, `3P`)), centers = 4, nstart = 20)  
nba\_data$Cluster <- as.factor(k\_result$cluster)

## Step 3: Build the App UI \* A lot of this is unformatted for the sake of rendering refer to basketball app.r for the full shiny code\*

Paste the full layout below. This creates dropdown menus and areas for the plots and table.

# UI and server code commented out just for the sake of renering- include this in your code without it commented out!-…

ui <- fluidPage(

# App title titlePanel(“NBA Role Explorer”),

# Sidebar layout with input and output #sidebarLayout(

# Inputs on the left  
sidebarPanel(  
 # Dropdown to select era  
 selectInput("era", "Select Era:", choices = unique(nba\_data$Era)),  
   
 # Checkboxes to select positions (G, F, C, etc.)  
 checkboxGroupInput("pos", "Select Position(s):", choices = unique(nba\_data$Pos)),  
   
 # Optional text input to search for a player  
# textInput("search", "Search Player (optional):", "")

# ),

# Output area on the right  
mainPanel(  
 plotOutput("scatterPlot"),  
 plotOutput("boxPlot"),  
 tableOutput("playerTable")  
)

) #)

Step 4: Write the Server Logic This controls how your inputs filter the data and how the results appear.

# Server

server <- function(input, output) {  
 filtered\_data <- reactive({  
 data <- read\_csv("../Data/nba\_all.csv")|> filter(Era == input$era)  
 if (!is.null(input$pos)) data <- data |> filter(Pos %in% input$pos)  
 if (input$search != "") data <- data |> filter(grepl(input$search, Player, ignore.case = TRUE))  
 data  
 })  
  
 output$scatterPlot <- renderPlot({  
 ggplot(filtered\_data(), aes(x = AST, y = PTS, color = Cluster)) +  
 geom\_point(alpha = 0.6) +  
 labs(title = "AST vs. PTS by Cluster", x = "Assists", y = "Points")  
 })  
  
 output$boxPlot <- renderPlot({  
 ggplot(filtered\_data(), aes(x = Cluster, y = `3P`, fill = Cluster)) +  
 geom\_boxplot() +  
 labs(title = "3P by Cluster", y = "3P Made", x = "Cluster")  
 })  
  
 output$playerTable <- renderTable({  
 filtered\_data() |> select(Player, Pos, Era, Cluster, PTS, AST, TRB, `3P`)  
 })  
}  
  
# UI and server code commented out just for the sake of renering- include this in your code without it commented out!-... shinyApp(ui = ui, server = server)

Reflection Questions Try filtering by different positions and clusters — what do you notice?

How would you improve the app for future users?

What roles appear most common in recent eras?

# Part 6: Wrap-Up and Final Reflection

In this module, you’ve used real NBA data to explore how player roles and performance have evolved over time. You visualized key trends, fit a regression model to understand scoring, used clustering to define role types, and built an interactive Shiny app to explore those roles across eras.

Return to the big question:

### How have NBA player roles and scoring styles evolved across the decades, and how can we use data and visual tools to understand that shift?

Use evidence from: - Your summary statistics - The three visualizations - The regression model - Your cluster analysis - The Shiny app filters and results

Write 4–6 sentences summarizing what you discovered.

## Optional

Try changing the number of clusters in your k-means model. How does it change the interpretation of roles? Would 3 clusters make more sense? What about 5?

set.seed(123) k\_result\_alt <- kmeans(scale(nba\_data |> select(PTS, AST, TRB, 3P)), centers = 3, nstart = 20) table(k\_result\_alt$cluster)