NBA Scoring Analysis

Varun Gullanki, Anirudh Ganesan, Zezhou Zhuang

Introduction

In professional basketball, a player's scoring performance is influenced by numerous factors such as their physical attributes, playing experience, role within the team, and overall skillset. Among these attributes, height and age are often considered fundamental components of a player's potential and performance ceiling. In this project, we conduct an exploratory data analysis to investigate the extent to which a player's height and age affect their ability to score points in the NBA.

Specifically, we seek to answer the following research questions: First, does height significantly impact scoring output in the NBA? Second, at what age do NBA players typically reach their peak scoring performance? Lastly, what combination of height and age tends to yield the highest average scoring? By exploring these questions, we aim to derive meaningful insights that could inform player scouting, training strategies, and performance forecasting models in professional basketball.

Data Provenance

The dataset analyzed in this project was compiled and uploaded by Kaggle contributor Justinas. It includes player-level statistics from multiple NBA seasons, with each row representing a single player's performance during a specific season. The dataset captures key variables such as the player's age, height (recorded in centimeters), and average points scored per game, among others. This structure allows us to conduct a detailed temporal and demographic analysis of player scoring trends.

Citation: Justinas. "NBA Players Data." *Kaggle*, 2023, https://www.kaggle.com/datasets/justinas/nba-players-data. Accessed May 2025.

FAIR/CARE Principles

The dataset complies with the FAIR principles of responsible data usage. It is **Findable**, as it is hosted on a well-known and accessible platform, Kaggle. It is **Accessible**, provided in a clean CSV format that facilitates seamless integration into various data analysis tools. The data is also **Interoperable**, compatible with programming environments such as R and Python, and **Reusable**, with clearly labeled variables and consistent formatting that support a wide range of analytical objectives. These qualities make the dataset an appropriate and ethical choice for our research.

Summary Statistics

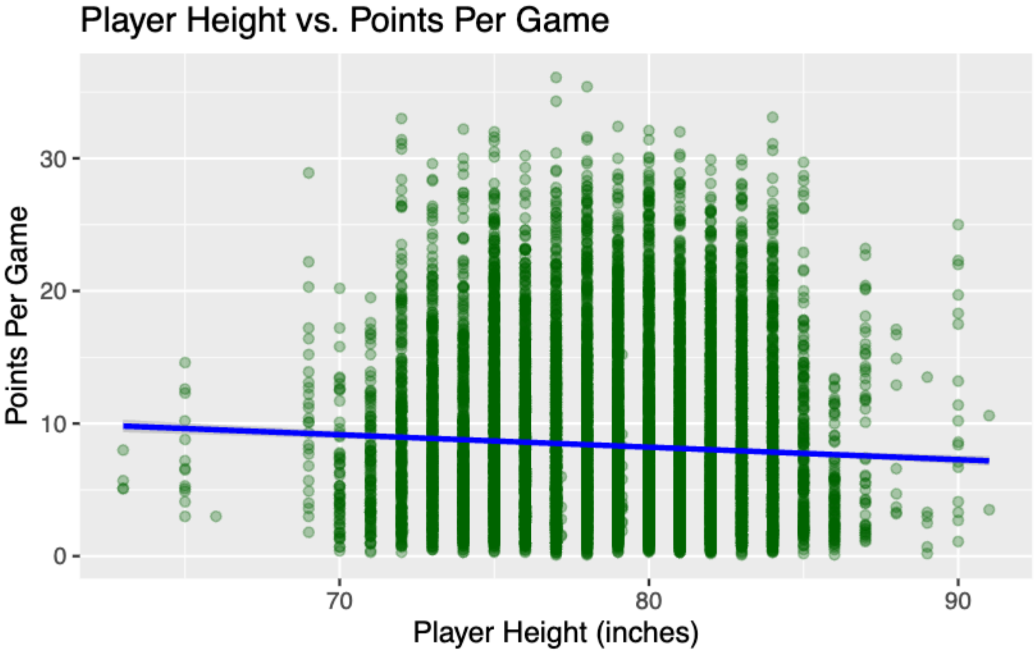
Before beginning our analysis, we carried out several important preprocessing steps to ensure the integrity and clarity of our data. First, we converted the `player_height` variable from centimeters to inches, a more familiar unit for interpreting player stature in basketball contexts. Next, we filtered out rows where key variables such as `pts` (points per game), `age`, or `player_height` were missing or zero. This step was necessary to maintain the validity of our statistical models and visualizations.

We narrowed our focus to three variables central to our analysis: age, height, and points per game. The following table provides a summary of the distributions of these variables in the cleaned dataset:

Table 1: Summary Statistics for Age, Height (in), and Points Per Game

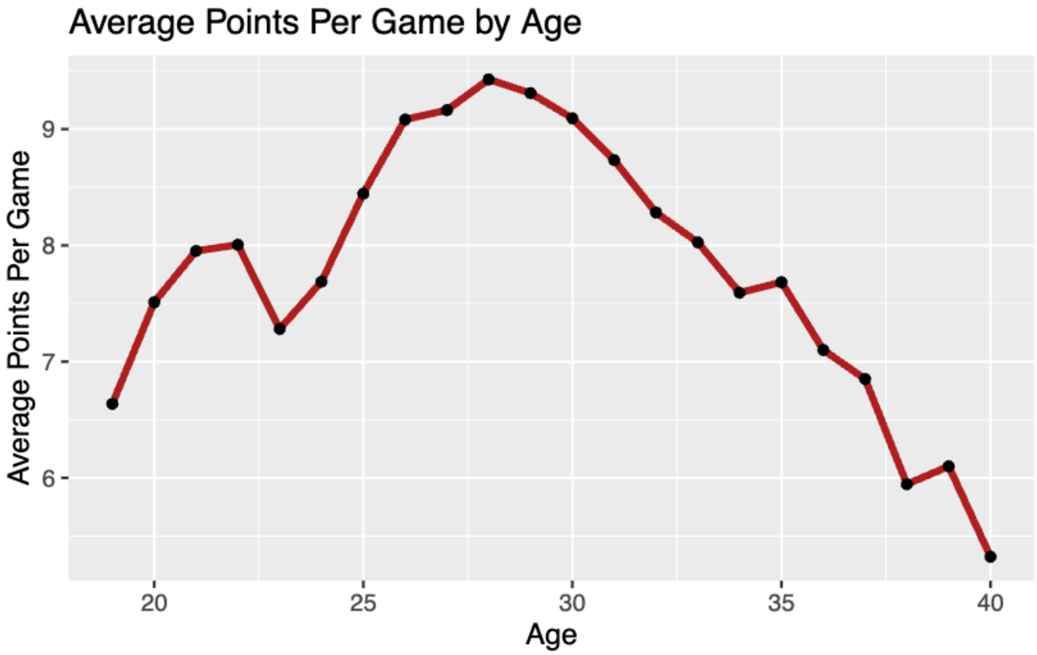
age	player_height	pts
Min. :18.00	Min. :63.00	Min. : 0.100
1st Qu.:24.00	1st Qu.:76.00	1st Qu.: 3.700
Median :26.00	Median :79.00	Median : 6.800
Mean :27.06	Mean :78.96	Mean : 8.317
3rd Qu.:30.00	3rd Qu.:82.00	3rd Qu.:11.600
Max. :44.00	Max. :91.00	Max. :36.100

Q1: Does Player Height Affect Scoring?



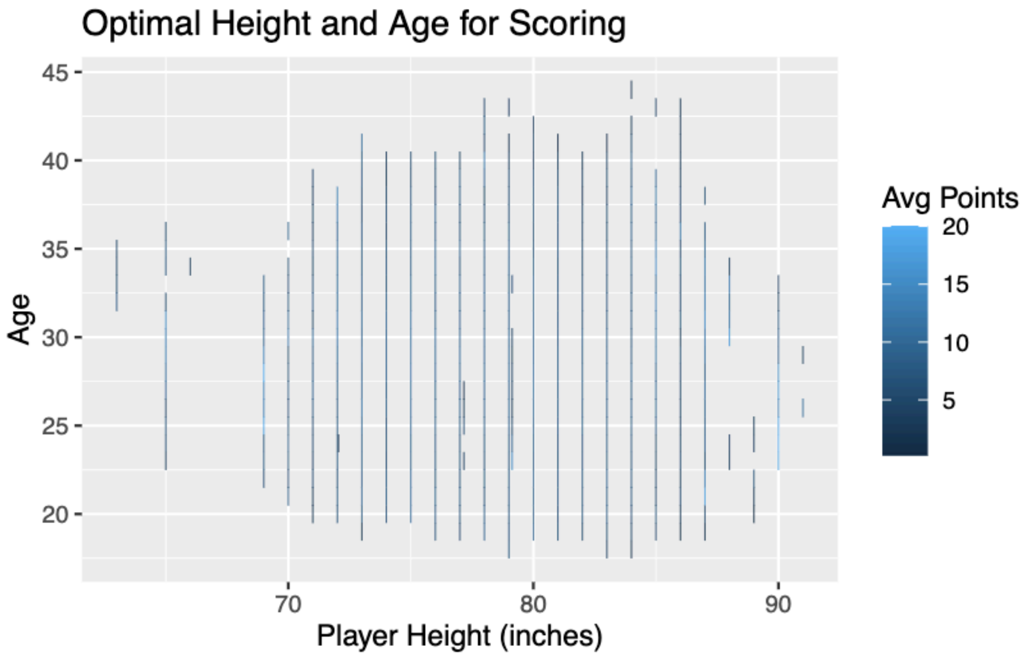
Interpretation:
The scatterplot above shows a wide distribution of scoring performance across the height spectrum. While the linear regression line does suggest a marginal increase in scoring as height increases, the overall relationship is weak. This implies that player height alone is not a strong predictor of scoring performance in the NBA. In fact, the presence of high-scoring guards, who are typically shorter than forwards and centers, underscores the importance of factors such as player role, play style, and usage rate. Our findings suggest that taller players do not necessarily score more, and that scoring potential is more nuanced and position-dependent.

Q2: At What Age Do NBA Players Peak in Scoring?



Interpretation:
The plot of average points per game by age shows a clear pattern. Scoring ability generally increases from age 20, peaking between ages 26 and 28, and then gradually declines. This trend reflects the physical and mental development of players, by their mid-20s, most have developed a strong understanding of the game while still maintaining peak physical condition. The decline after age 30 may be attributed to reduced minutes, physical wear, or shifts in team role. Our analysis reinforces the notion that a player's **prime scoring years** typically occur in their **late 20s**.

Q3: What Is the Optimal Height/Age for Scoring?



Interpretation:
The heatmap reveals a distinct region where scoring output is consistently high: players who are between **75 and 79 inches tall** (roughly 6'3" to 6'7") and between **25 and 28 years old**. This intersection likely reflects elite guards and wings who are central to their team's offense and who have accumulated enough experience to maximize their scoring opportunities. These findings suggest that the **ideal profile for scoring in the NBA combines athletic maturity with skill development**, placing players in their late 20s with moderate-to-tall height as the most prolific scorers.

Data Visualization Summary

To answer our research questions, we employed three main visualizations: a scatterplot, a line graph, and a heatmap. Each plot served a specific analytical purpose. The scatterplot demonstrated the weak and noisy relationship between player height and scoring. The line graph offered insight into the progression of scoring over a player's career lifespan, highlighting a peak in the late 20s. Finally, the heatmap allowed us to visualize the joint effect of height and age, clearly identifying the demographic profile of the league's top scorers. Together, these visual tools paint a robust picture of what drives scoring in the NBA.

Conclusion

This exploratory analysis yields several noteworthy conclusions. First, **height is not a major factor in predicting scoring output**. However, shorter players, particularly guards, often score more due to their offensive responsibilities. Second, **age is a more reliable indicator of scoring success**, with most players achieving their highest scoring averages between 26 and 28 years old. Finally, the **combination of 75–79 inches tall (6'3" to 6'7") and aged 25–28** emerges as the optimal range for scoring performance, reflecting the synergy between physical prime and professional experience.

For future studies, it would be valuable to control for additional variables such as player position, minutes played, team pace, and advanced metrics like usage rate or true shooting percentage. These additions would enable a more granular understanding of what influences scoring and provide even more actionable insights for coaches, analysts, and front offices.

Code Appendix

```r
library(tidyverse)
nba <- read_csv("/Users/varun/Downloads/all_seasons.csv")
nba_clean <- nba %>%
  mutate(player_height = player_height * 0.393701) %>%
  filter(!is.na(pts), !is.na(age), !is.na(player_height), pts > 0)

summary_data <- nba_clean %>%
  select(age, player_height, pts)

knitr::kable(summary(summary_data), caption = "Summary Statistics for Age, Height (in), and
ggplot(nba_clean, aes(x = player_height, y = pts)) +
  geom_point(alpha = 0.3, color = "darkgreen") +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(
    title = "Player Height vs. Points Per Game",
    x = "Player Height (inches)",
    y = "Points Per Game"
  )
age_avg_pts <- nba_clean %>%
  group_by(age) %>%
  summarize(mean_pts = mean(pts), count = n()) %>%
  filter(count > 20)

ggplot(age_avg_pts, aes(x = age, y = mean_pts)) +
  geom_line(color = "firebrick", size = 1.3) +
  geom_point(color = "black") +
  labs(
    title = "Average Points Per Game by Age",
    x = "Age",
    y = "Average Points Per Game"
  )
nba_grid <- nba_clean %>%
  group_by(age, player_height) %>%
  summarize(mean_pts = mean(pts), .groups = "drop")

ggplot(nba_grid, aes(x = player_height, y = age, fill = mean_pts)) +
  geom_tile() +
  labs(
    title = "Optimal Height and Age for Scoring",
    x = "Player Height (inches)",
    y = "Age",
    fill = "Avg Points"
  )
```