How Does Height Affect Rebounds per Game in the NBA?

Tim Damasco, Aidan Murphy, Tyler Gilbert

2025-12-17

Purpose and Motivation

Professional basketball rebounding is often assumed to be dominated by height. Taller players are expected to secure more rebounds simply due to physical reach and proximity to the basket. However, rebounding is also influenced by player role, positioning, and team strategy. These are factors that vary substantially by on the court position. This project investigates how player height relates to rebounds per game during the 2008–2009 NBA season, and critically examines whether the importance of height differs by position.

Rather than assuming a uniform relationship across all players, this analysis explores position specific patterns, distinguishing between offensive and defensive rebounds. This approach allows us to move beyond the simplistic "taller is better" assumption and toward a more nuanced understanding of rebounding dynamics in professional basketball.

Data Sources

This analysis combines two publicly available datasets:
• NBA Rebounding Statistics (2008–2009)-Per-game rebounding statistics (ORB, DRB, TRB, games played, position) scraped from Basketball Reference.
• NBA Player Heights Dataset-Height measurements sourced from the OpenIntro NBA Heights dataset.

Inclusion Criteria:
To ensure meaningful and comparable statistics, players were filtered to:
- Have played at least 65 games.
- Average at least 1 rebound per game.

After filtering and matching, the final dataset includes 178 NBA players across all five standard positions.

Data Tidying, Wrangling, and Cleaning

Prior to analysis, both datasets required preparation to ensure consistency and accurate player matching. Player names across the rebounding statistics and height datasets were formatted differently, including the presence of accents, abbreviated first names, and alternative spellings. To address this, all player names were standardized by removing accents and normalizing text formatting. A two stage matching procedure was then implemented. First, an exact name match was performed for players with identical standardized names across both datasets. For remaining unmatched players, a secondary soft matching approach was used, which relied on a shortened identifier composed of the player's first initial and last name. This approach allowed for successful matching of players with minor naming discrepancies while minimizing the risk of incorrect merges.

Additional cleaning steps were taken to ensure the integrity of the dataset. Players with insufficient participation were excluded by filtering out those who appeared in fewer than 65 games, as their per-game statistics may not reliably reflect typical performance. Players averaging fewer than one rebound per game were also removed to eliminate cases where rebounding was not a meaningful component of the player's role. Height values were converted into numeric form to support regression analysis, and redundant helper variables created during the matching process were removed to produce a clean, dataset. These steps resulted in a final analytical dataset consisting of 178 players with complete and reliable information suitable for exploratory and inferential analysis.

Principle Evaluation

The datasets used in this project largely satisfy the FAIR data principles. The data are findable and accessible, as both sources are publicly available and clearly documented within the project repository. Accessibility is further supported by the use of open file formats, such as CSV, which can be read by a wide range of statistical and computational tools. Interoperability is achieved through the use of standardized variable structures and open source software, enablingintegration and reuse. Reusability is supported by thorough documentation of the data cleaning and matching procedures, allowing other researchers to replicate or extend the analysis.

With respect to the CARE principles, this project aligns appropriately given the nature of the data. The analysis contributes to collective benefit by improving understanding of how physical attributes interact with positional roles in professional basketball. Authority to control is not a concern, as the data consist solely of aggregated, publicly reported performance statistics from professional athletes. Responsibility is demonstrated through careful contextualization of results, avoiding overgeneralization or misrepresentation of individual players. Ethical considerations are minimal, as the data contain no sensitive or personally identifiable information beyond what is already publicly available.

Attributes Used in Analysis

This analysis focuses on a combination of original and derived attributes that together capture both individual performance and positional context. The primary attributes sourced directly from the datasets include player height, offensive rebounds per game, defensive rebounds per game, total rebounds per game, position, and games played. These variables form the foundation of the analysis, allowing direct examination of rebounding performance and its relationship to physical stature.

In addition to the original variables, several derived attributes were created to support deeper insight. Position averages for offensive, defensive, and total rebounds were computed to summarize typical rebounding behavior across roles. Regression slopes representing rebounds gained per additional inch of height were calculated separately for each position, along with associated confidence intervals. These derived measures enable a comparison of the marginal importance of height across positions and facilitate interpretation beyond simple correlations. Together, the original and derived attributes allow the analysis to address both broad trends and nuanced positional differences.

Exploratory Data Analysis

Exploratory data analysis revealed clear structural patterns in rebounding behavior across positions. Average rebounds per game increase consistently from guards to centers, reflecting the positional responsibilities inherent in professional basketball. Defensive rebounds exceed offensive rebounds across all positions, highlighting the strategic emphasis on securing possession following opponent misses. Height distributions vary substantially by position, with centers exhibiting less variation and guards showing the widest range, further justifying an analytical approach focused on position.

The exploratory analysis also indicated meaningful variability in rebounding performance within positions, suggesting that height alone does not fully explain rebounding success. Some shorter players outperform taller peers within the same role, pointing to the influence of factors such as positioning, effort, and team scheme. These observations motivated the decision to analyze the relationship between height and rebounding separately by position rather than relying on a single pooled model. By doing so, the analysis better captures the contextual role that height plays within different on the court responsibilities and sets the foundation for the inferential results.

Findings

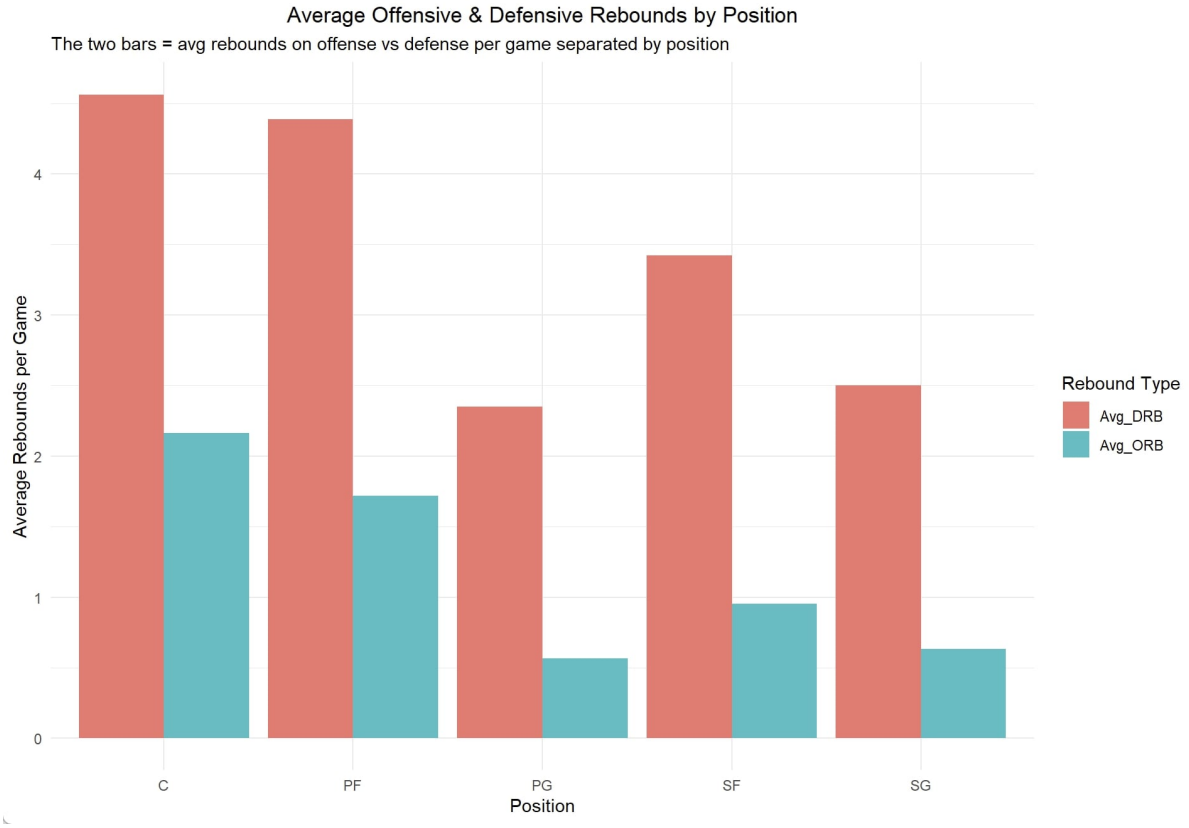

Figure 1: Bar chart showing average offensive and defensive rebounds per game for NBA players during the 2008–2009 season, grouped by position (Center, Power Forward, Small Forward, Shooting Guard, and Point Guard).

Figure 1 presents average offensive and defensive rebounds per game by position, providing an initial overview of how rebounding output varies across on-court roles. A clear positional gradient emerges, with centers averaging the highest number of rebounds per game on both offense and defense, followed by power forwards, small forwards, shooting guards, and point guards. Defensive rebounds exceed offensive rebounds for every position, reflecting the structural reality that defensive possessions create more rebounding opportunities and that rebounding responsibilities are heavily emphasized on the defensive end. This figure establishes an important baseline: rebounding volume is strongly tied to position, and positions traditionally associated with interior play are responsible for the majority of rebounds. In this sense, centers dominate rebounding not only because of height but also because their role places them closest to the basket and most frequently involved in rebounding situations.

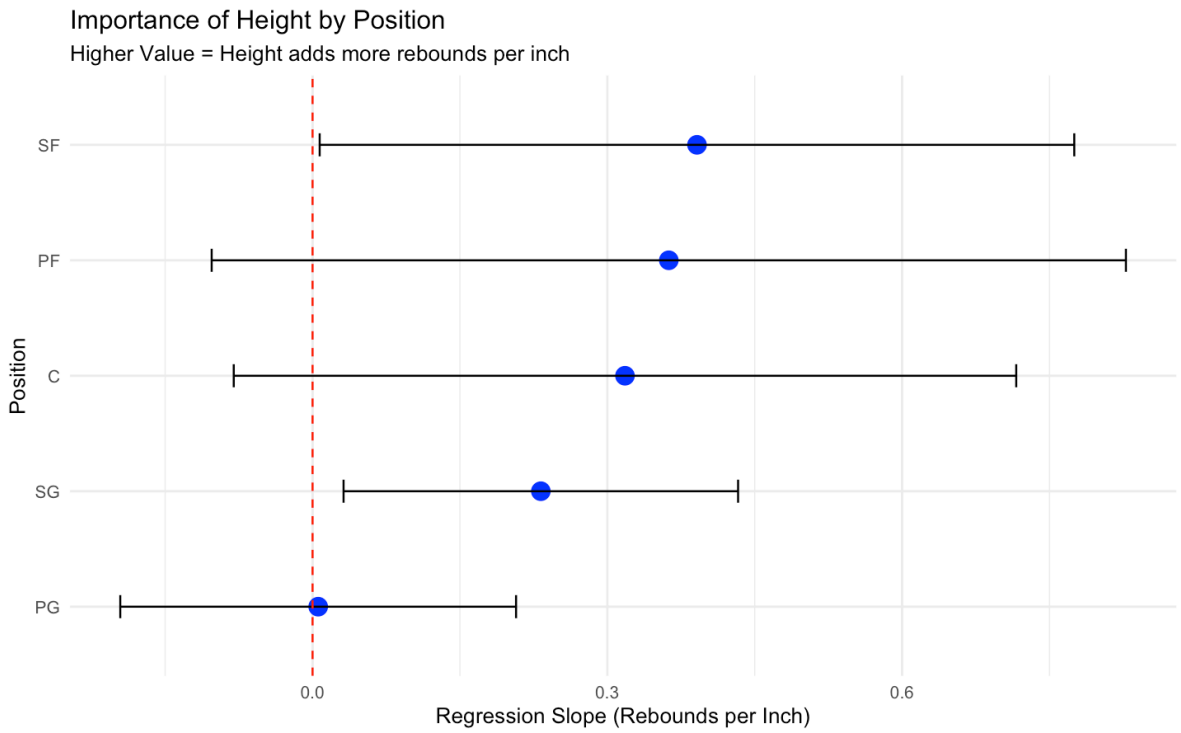

Figure 2: Horizontal dot-and-interval plot showing the relationship between player height and rebounds per game by position. The vertical axis lists positions, and the horizontal axis shows regression slopes representing additional rebounds gained per inch of height.

While Figure 1 demonstrates who rebounds the most, it does not explain how height influences rebounding within each position. This distinction is addressed by Figure 2, which examines the importance of height by position using regression slopes that represent the increase in rebounds per game associated with each additional inch of height. In this figure, position is displayed on the vertical axis and the regression slope (rebounds per inch) on the horizontal axis, allowing direct comparison of how strongly height contributes to rebounding across roles. The results show that although centers average the most rebounds overall, the marginal effect of height is not strongest at the center position. Instead, small forwards exhibit the largest positive relationship between height and rebounds per game, indicating that additional height provides a greater competitive advantage for players in this role.

This finding highlights a central insight of the project: height matters differently depending on position. For centers, height is relatively uniform across the position and rebounding is already an expected component of their role, which limits the marginal benefit of being taller than peers. In contrast, small forwards operate in a more versatile and variable role, balancing perimeter play with interior responsibilities. As a result, additional height can substantially influence

how effectively a small forward rebounds, allowing taller players at this position to secure rebounds more like front-court players. Guards show the weakest relationship between height and rebounding, reinforcing the idea that speed, ball handling, and perimeter responsibilities are more important than size for those roles.

Taken together, these two visualizations directly answer the project's research question. Height is clearly associated with rebounding success, but its impact is not uniform across the NBA. While centers lead the league in total rebounding output, height plays its most influential role among small forwards, where differences in stature meaningfully shape rebounding performance. This distinction underscores the importance of considering positional context when evaluating the relationship between physical attributes and on court performance.

Code Appendix

Code for Creating the Dataframe

```
library(tidyverse)
library(stringi)

# 1. Read website and CSV data
rebounding_df <- read_csv("~/Downloads/Stat184_Rebounding_Data.csv")
heights_df <- read_tsv("https://www.openintro.org/data/tab-delimited/nba_
                       heights.txt")

# 2. Create a "Short ID" Function. Nessary step because the names between
#the website and CSV file are different. For example, Lou Williams vs
#Louis Williams
create_short_id <- function(name) {
  clean_name <- stri_trans_general(name, "Latin-ASCII") # Remove accents
  first_initial <- substr(clean_name, 1, 1)             # Get first letter
  last_name <- word(clean_name, -1)                     # Get last word

  # Combine and lowercase. For example "l williams", to standardize by
  #the similarities between datasets
  tolower(paste(first_initial, last_name))
}

# 3. Prepare Both Datasets
# Added a 'clean_name' for exact matching and 'short_id' for soft matching
heights_prep <- heights_df %>%
  unite(col = "Player", first_name, last_name, sep = " ", remove = FALSE) %>%
```

```
  mutate(
    clean_name = stri_trans_general(Player, "Latin-ASCII"),
    short_id = create_short_id(Player)
  )

rebounding_prep <- rebounding_df %>%
  mutate(
    clean_name = stri_trans_general(Player, "Latin-ASCII"),
    short_id = create_short_id(Player)
  )

# --- This is where the matching process begins ---

# Step 4A: The Exact Match
matches_exact <- rebounding_prep %>%
  inner_join(heights_prep, by = "clean_name") %>%
  select(-ends_with(".y"), -short_id.y) %>%
  rename(Player = Player.x, short_id = short_id.x)

# Step 4B: Identify the "Leftovers". So, who is in CSV and website but
#not an exact match
leftover_players <- rebounding_prep %>%
  anti_join(matches_exact, by = "Player")

available_heights <- heights_prep %>%
  anti_join(matches_exact, by = "clean_name")

# Step 4C: The "Soft" Match. This spefically pairs the leftovers between the
#data sets. For example, some players on the website only had first names
matches_soft <- leftover_players %>%
  inner_join(available_heights, by = "short_id") %>%
  select(-ends_with(".y"), -clean_name.y, -clean_name.x) %>%
  rename(Player = Player.x)

# 5. Combine and clean up the DF for presentation purposes
final_df <- bind_rows(matches_exact, matches_soft) %>%
  select(
    -first_name,
    -last_name,
    -starts_with("..."), # This automatically removes ...8, ...9, ...10
    -clean_name,         # Removes our helper column
    -short_id            # Removes our helper column
```

```
  )

# 6. View Final Result
head(final_df)
```

Baseline Stats Creation

```
# 1) Prepare the Data
table_data <- final_df %>%
  group_by(Pos) %>%
  summarize(
    Count = n(),
    Height = round(mean(h_in), 1),
    TRB = round(mean(TRB), 1),
    ORB = round(mean(ORB), 1),
    DRB = round(mean(DRB), 1)
  ) %>%
  arrange(desc(TRB)) %>% # Sort by Rebounds so the table has an order
  mutate(Pos = factor(Pos, levels = rev(Pos))) %>% # Fix the row order
  #for the plot
  pivot_longer(cols = -Pos, names_to = "Column", values_to = "Value") %>%
  mutate(Column = factor(Column, levels = c("Count", "Height", "TRB", "ORB",
                                            "DRB")))

# 2) Create the Table Plot
ggplot(table_data, aes(x = Column, y = Pos)) +

  geom_tile(fill = "white", color = "black", linewidth = 0.5) +
  geom_text(aes(label = Value), size = 5, color = "black") +
  scale_x_discrete(position = "top") +
  labs(
    title = "Position Baseline Stats",
    x = NULL,
    y = NULL
  ) +
  theme_minimal() +
  theme(
    panel.grid = element_blank(),
    axis.text.x = element_text(face = "bold", size = 12, color = "black"),
    axis.text.y = element_text(face = "bold", size = 12, color = "black"),
```

```
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold", margin =
                                margin(b = 10))
  )
```

Regression Creation

```
# 1) Create Visualization
ggplot(slopes_by_pos, aes(x = reorder(Pos, estimate), y = estimate)) +
  geom_point(size = 4, color = "blue") +
  geom_errorbar(aes(ymin = estimate - 1.96 * std.error,
                    ymax = estimate + 1.96 * std.error),
                width = 0.2) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Importance of Height by Position",
    subtitle = "Higher Value = Height adds more rebounds per inch",
    x = "Position",
    y = "Regression Slope"
  ) +
  coord_flip() +
  theme_minimal()
```

Final Project Table and Visualization

```
library(dplyr)
library(ggplot2)
library(tidyverse)

# Read player rebounding data
reb_data <- read.csv("C:/Users/trock/Downloads/Stat184_Reb_Data.csv")

# Read all players height data
height_data <- read.csv("C:/Users/trock/Downloads/nba_heights.csv")

# Used to view how it read into R
# glimpse(reb_data)
# glimpse(height_data)
```

```r
# Selected only the needed attributes
reb_data <- reb_data %>%
  select(Player, Pos, G, GS, ORB, DRB, TRB)


reb_pos_summary <- reb_data %>%
  group_by(Pos) %>%     # Group players into 5 positions: PG, SG, SF, PF, C
  summarise(
    # Averages of all statistics by position
    Avg_ORB = round(mean(ORB), 2),
    Avg_DRB = round(mean(DRB), 2),
    Avg_TRB = round(mean(TRB), 2),
    Players = n()
  ) %>%
  arrange(desc(Avg_TRB))

view(reb_pos_summary)

# Reshaping data
reb_long <- reb_pos_summary %>%
  select(Pos, Avg_ORB, Avg_DRB) %>%
  pivot_longer(
    cols = c(Avg_ORB, Avg_DRB),
    names_to = "Rebound_Type",
    values_to = "Average_Rebounds"
  )


# Create the visualization
ggplot(reb_long, aes(x = Pos, y = Average_Rebounds, fill = Rebound_Type)) +
  geom_col(position = "dodge") +
  labs(
    # Characteristics of the visualization
    title = "Average Offensive & Defensive Rebounds by Position",
    subtitle = "The two bars = avg rebounds on offense vs defense per game
    separated by position",
    x = "Position",
    y = "Average Rebounds per Game",
    # Key to identify what each bar represents
    fill = "Rebound Type"
  ) +
  theme_minimal() +
```

```
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```