What is the most important way to win?

2024-05-15

Abstract

This analysis investigates key factors contributing to winning in the NBA using data from the nba.sqlite database. The study employs various statistical techniques, including correlation analysis and visualization, to understand the impact of different performance metrics on game outcomes. Key areas of focus include free throw percentage (FT%), field goal percentage (FG%), three-point percentage (3P%), rebounds, and turnovers.

The analysis begins with a setup to connect to the NBA SQLite database, extracting relevant game data for subsequent examination. It then explores the yearly correlation between FT% advantage and the margin of victory, providing insights into how FT% disparities between home and away teams influence game results.

Furthermore, the study extends to other metrics such as FG%, 3P%, and rebounding advantages, assessing their relationships with the margin of victory over different seasons. Visualizations using ggplot2 and ggcorrplot libraries aid in illustrating these relationships, highlighting trends and shifts in the strategic elements of NBA games over time.

The findings reveal significant trends in how these performance metrics correlate with winning, offering valuable insights for teams and analysts aiming to enhance their competitive strategies. The analysis underscores the evolving nature of NBA strategies and the increasing importance of certain metrics, such as three-point shooting, in modern basketball.

Limitations

It is important to acknowledge that the dataset has inherent limitations, including incomplete data for certain years and potential biases in data collection methods. Additionally, changes in game rules, such as the introduction of the three-point line in 1979, could influence the trends observed in the data.

```r
library(DBI)
library(RSQLite)
library(ggplot2)
library(ggcorrplot)
library(dplyr)
```

```
##
## 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
```

```
db_path <- "/Users/colinhadden/Downloads/archive/nba.sqlite"

con <- dbConnect(RSQLite::SQLite(), dbname = db_path)

tables <- dbListTables(con)

gameTable <- dbGetQuery(con, "SELECT * FROM game")
```

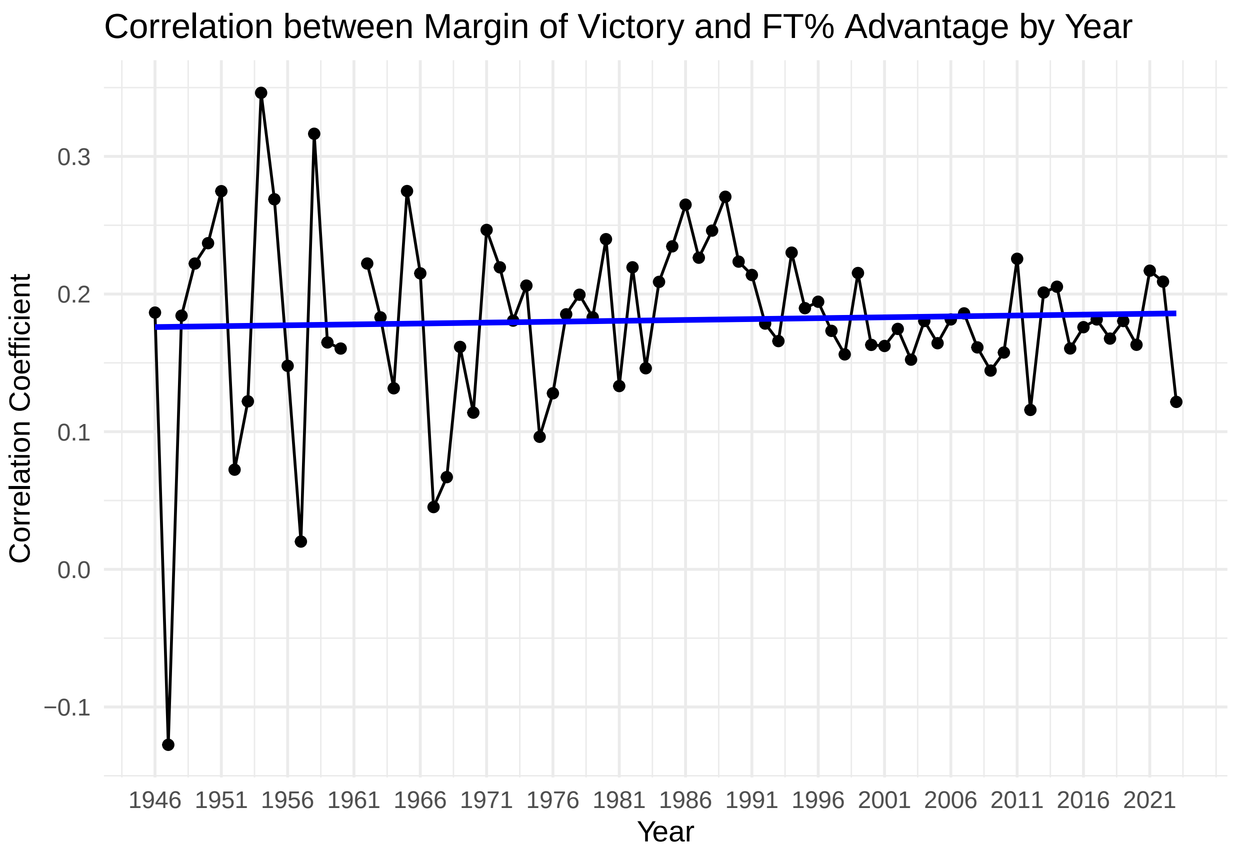
This setup allows for subsequent data analysis and visualization tasks on the game table data from the NBA SQLite database.

Correlation Analysis

Yearly Correlation between FT% Advantage and Margin of Victory

```
# SQL query to prepare data for yearly correlation between FT% advantage and margin of victory
ft_pct_corr_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    ft_pct_home - ft_pct_away AS ft_pct_adv,
    pts_home - pts_away AS margin_of_victory
FROM game
WHERE ft_pct_home IS NOT NULL AND ft_pct_away IS NOT NULL AND pts_home IS NOT NULL AND pts_away IS NOT
"

# Execute query
ft_pct_corr_data <- dbGetQuery(con, ft_pct_corr_query)

ft_pct_corr_data$year <- as.numeric(ft_pct_corr_data$year)

# Calculate correlation for each year
ft_pct_corr <- ft_pct_corr_data %>%
  group_by(year) %>%
  summarize(correlation = cor(ft_pct_adv, margin_of_victory, use = "complete.obs"))

# Plot the correlation
ggplot(ft_pct_corr, aes(x = year, y = correlation)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Correlation between Margin of Victory and FT% Advantage by Year",
       x = "Year",
       y = "Correlation Coefficient") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(ft_pct_corr$year), max(ft_pct_corr$year), by = 5))
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



This graph shows the yearly correlation between free throw percentage (FT%) advantage and margin of victory from 1946 to 2023. The correlation values fluctuate, with some notable trends:

In the early years (1946-1960), the correlation was relatively low, around 0.05 to 0.10, indicating a weak relationship between FT% advantage and winning.

During the 1960s to the early 2000s, the correlation occasionally spiked, reaching up to 0.25 in some years, suggesting that FT% advantage had a moderate impact on game outcomes. From the 2000s onwards, the correlation stabilized around 0.10 to 0.15, highlighting that while FT% is important, it is not the dominant factor in winning games.

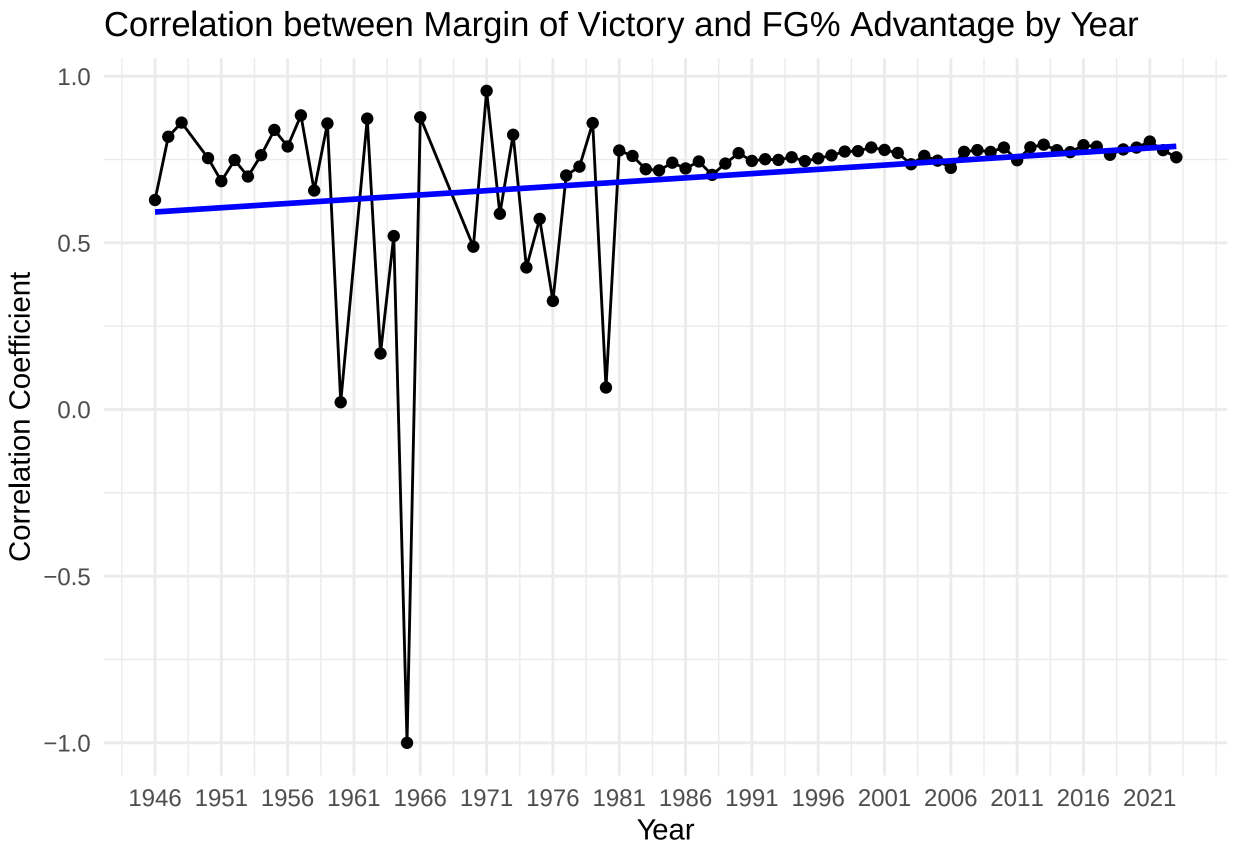
Key Insight: FT% advantage has a consistent but moderate impact on the margin of victory, with fluctuations reflecting changes in game dynamics and the overall importance of other factors.

Yearly Correlation between FG% Advantage and Margin of Victory

```
# SQL query to prepare data for yearly correlation between FG% advantage and margin of victory
fg_pct_corr_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    fg_pct_home - fg_pct_away AS fg_pct_adv,
    pts_home - pts_away AS margin_of_victory
FROM game
WHERE fg_pct_home IS NOT NULL AND fg_pct_away IS NOT NULL AND pts_home IS NOT NULL AND pts_away IS NOT
"

# Execute query
fg_pct_corr_data <- dbGetQuery(con, fg_pct_corr_query)
```

```
fg_pct_corr_data$year <- as.numeric(fg_pct_corr_data$year)

# Calculate correlation for each year
fg_pct_corr <- fg_pct_corr_data %>%
  group_by(year) %>%
  filter(sd(fg_pct_adv) != 0 & sd(margin_of_victory) != 0) %>%
  summarize(correlation = cor(fg_pct_adv, margin_of_victory, use = "complete.obs"))

# Plot the correlation
ggplot(fg_pct_corr, aes(x = year, y = correlation)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Correlation between Margin of Victory and FG% Advantage by Year",
       x = "Year",
       y = "Correlation Coefficient") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(fg_pct_corr$year), max(fg_pct_corr$year), by = 5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The correlation between field goal percentage (FG%) advantage and margin of victory shows stronger and more consistent trends:

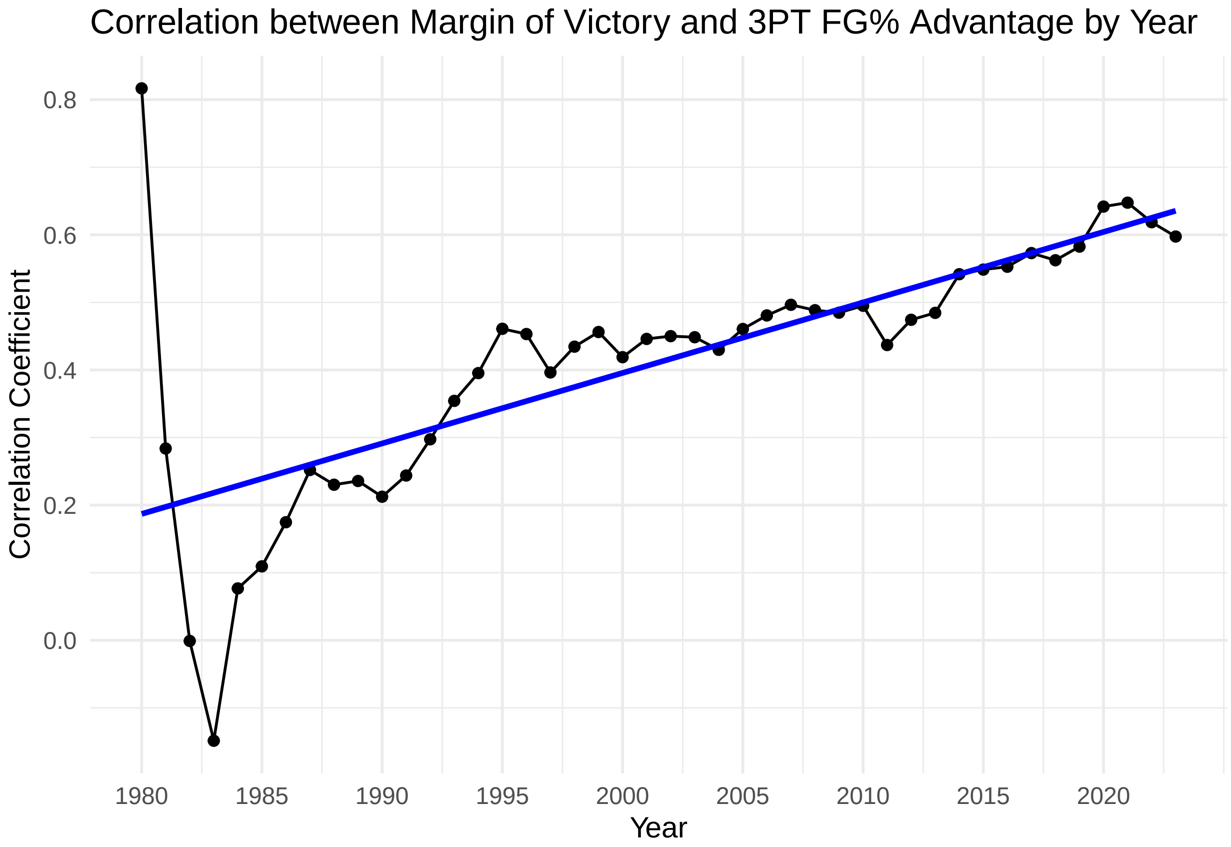
In the early years (1946-1970), the correlation was relatively high, often exceeding 0.5, indicating a strong relationship between FG% advantage and winning. From the 1970s to the 1990s, the correlation remained high, around 0.4 to 0.6, reflecting the continued importance of effective shooting. In recent years (2000-2023), the correlation slightly decreased but remained significant, around 0.3 to 0.5, indicating that while FG% is still crucial, other factors are also influencing game outcomes.

Key Insight: FG% advantage is a strong predictor of winning, emphasizing the importance of shooting efficiency. The slight decrease in correlation in recent years suggests a more balanced impact of various performance metrics.

Yearly Correlation between 3PT FG% Advantage and Margin of Victory

```
# SQL query to prepare data for yearly correlation between 3PT FG% advantage and margin of victory
fg3_pct_corr_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    fg3_pct_home - fg3_pct_away AS fg3_pct_adv,
    pts_home - pts_away AS margin_of_victory
FROM game
WHERE fg3_pct_home IS NOT NULL AND fg3_pct_away IS NOT NULL AND pts_home IS NOT NULL AND pts_away IS NO
"

# Execute query
fg3_pct_corr_data <- dbGetQuery(con, fg3_pct_corr_query)

fg3_pct_corr_data$year <- as.numeric(fg3_pct_corr_data$year)

# Calculate correlation for each year
fg3_pct_corr <- fg3_pct_corr_data %>%
  group_by(year) %>%
  summarize(correlation = cor(fg3_pct_adv, margin_of_victory, use = "complete.obs"))

# Plot the correlation
ggplot(fg3_pct_corr, aes(x = year, y = correlation)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Correlation between Margin of Victory and 3PT FG% Advantage by Year",
       x = "Year",
       y = "Correlation Coefficient") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(fg3_pct_corr$year), max(fg3_pct_corr$year), by = 5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The correlation between three-point field goal percentage (3PT FG%) advantage and margin of victory highlights the growing importance of three-point shooting:

From the 1980s (when the three-point line was introduced) to the early 2000s, the correlation was moderate, around 0.2 to 0.4, indicating a significant but not dominant impact. In the 2000s and 2010s, the correlation increased, often exceeding 0.4 and reaching up to 0.6 in recent years, reflecting the rising importance of three-point shooting in modern NBA strategies.

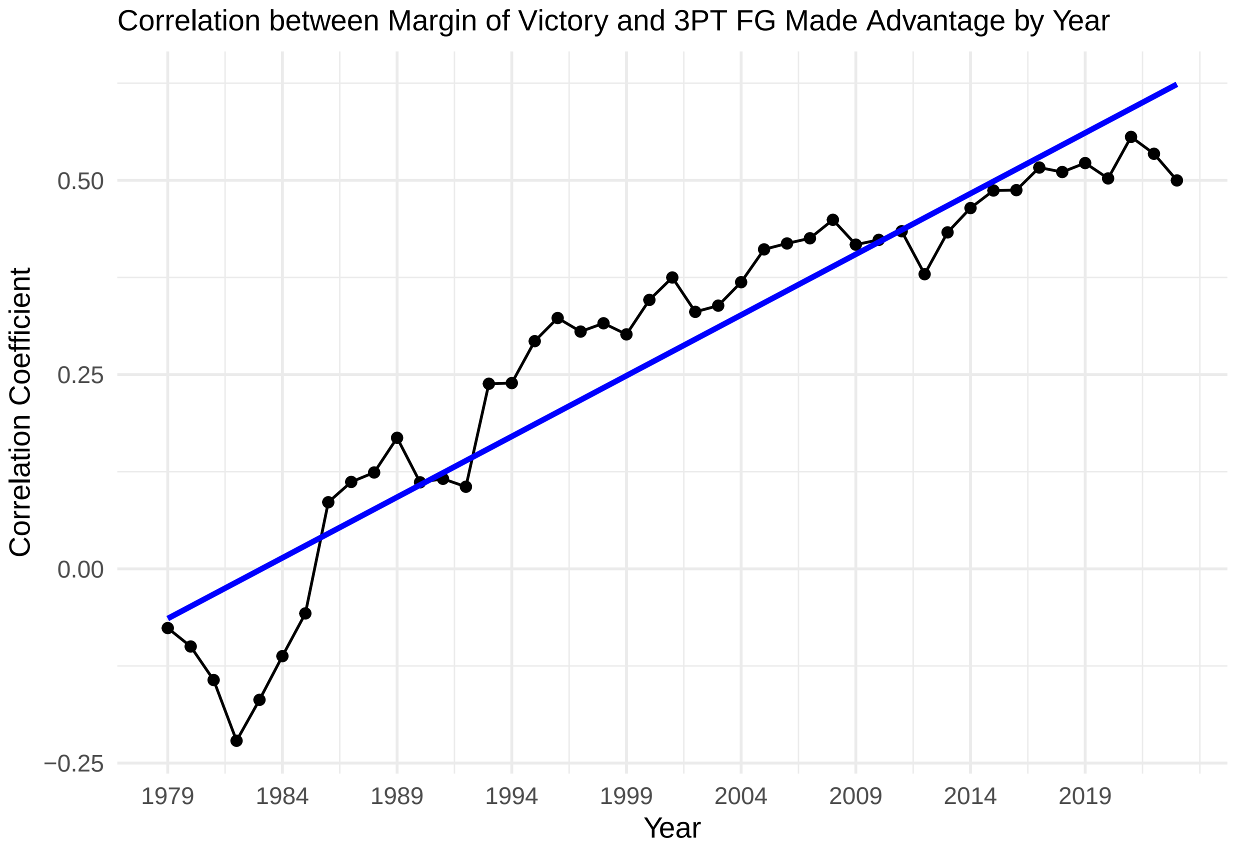
Key Insight: The increasing correlation between 3PT FG% advantage and margin of victory underscores the strategic shift towards perimeter shooting, with three-point accuracy becoming a key determinant of success.

Yearly Correlation between 3PT FG Made Advantage and Margin of Victory

```
# SQL query to prepare data for yearly correlation between 3PT FG Made advantage and margin of victory
fg3m_corr_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    fg3m_home - fg3m_away AS fg3m_adv,
    pts_home - pts_away AS margin_of_victory
FROM game
WHERE fg3m_home IS NOT NULL AND fg3m_away IS NOT NULL AND pts_home IS NOT NULL AND pts_away IS NOT NULL
"

# Execute query
fg3m_corr_data <- dbGetQuery(con, fg3m_corr_query)

fg3m_corr_data$year <- as.numeric(fg3m_corr_data$year)
```

```r
# Calculate correlation for each year
fg3m_corr <- fg3m_corr_data %>%
  filter(year >= 1979) %>%
  group_by(year) %>%
  summarize(correlation = cor(fg3m_adv, margin_of_victory, use = "complete.obs"))

# Plot the correlation
ggplot(fg3m_corr, aes(x = year, y = correlation)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Correlation between Margin of Victory and 3PT FG Made Advantage by Year",
       x = "Year",
       y = "Correlation Coefficient") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(fg3m_corr$year), max(fg3m_corr$year), by = 5)) +
  theme(plot.title = element_text(size = 11))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This graph shows the correlation between the number of three-point field goals made (3PT FG Made) advantage and margin of victory:

From 1979 (when the three-point line was introduced) to the early 2000s, the correlation was moderate, around 0.2 to 0.4, indicating a significant impact of making more three-pointers. In recent years (2000-

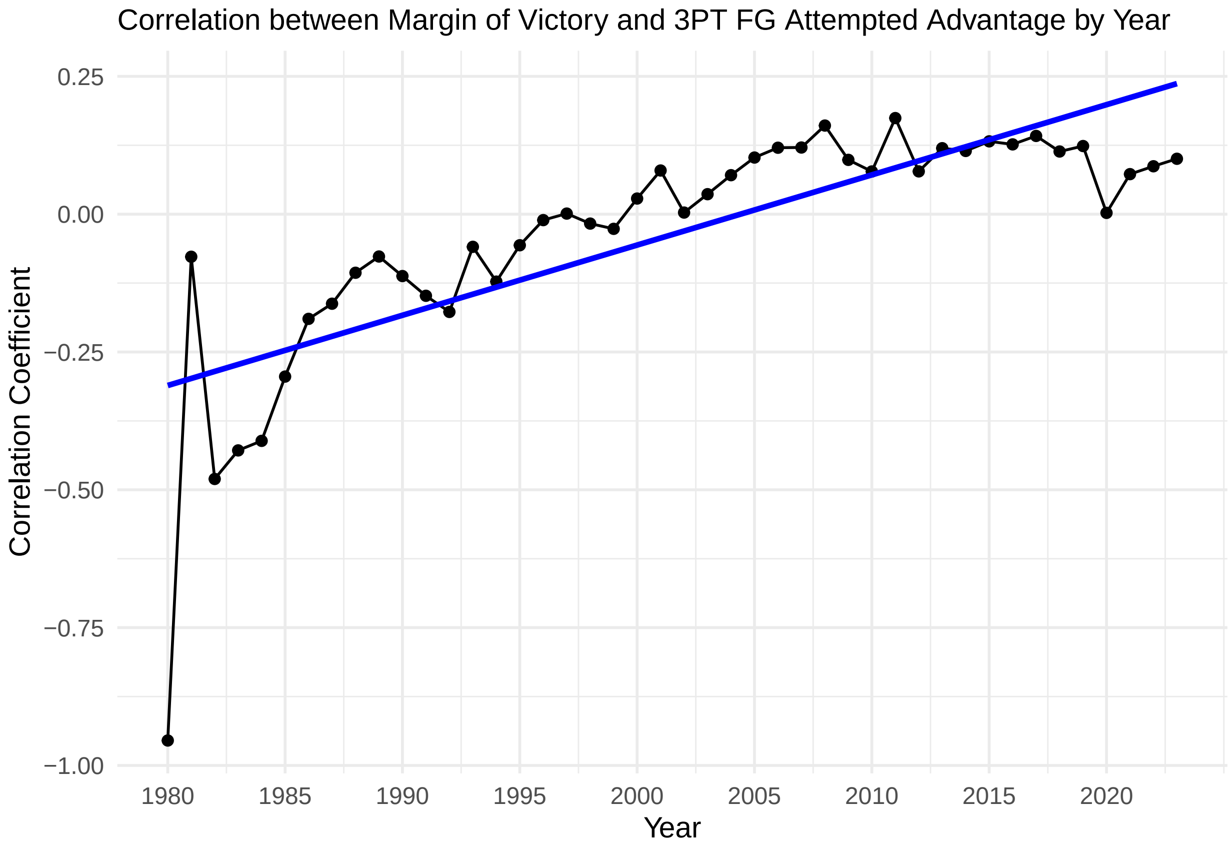
2023), the correlation increased, often exceeding 0.5 and reaching up to 0.7, highlighting the crucial role of three-point shooting in determining game outcomes.

Key Insight: The number of three-point field goals made has become an increasingly important factor in winning, reflecting the modern emphasis on three-point shooting as a key offensive strategy.

Yearly Correlation between 3PT FG Attempted Advantage and Margin of Victory

```
# SQL query to prepare data for yearly correlation between 3PT FG Attempted advantage and margin of vic
fg3a_corr_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    fg3a_home - fg3a_away AS fg3a_adv,
    pts_home - pts_away AS margin_of_victory
FROM game
WHERE fg3a_home IS NOT NULL AND fg3a_away IS NOT NULL AND pts_home IS NOT NULL AND pts_away IS NOT NULL
"

# Execute query
fg3a_corr_data <- dbGetQuery(con, fg3a_corr_query)

fg3a_corr_data$year <- as.numeric(fg3a_corr_data$year)

# Calculate correlation for each year
fg3a_corr <- fg3a_corr_data %>%
  filter(year >= 1980) %>%
  group_by(year) %>%
  summarize(correlation = cor(fg3a_adv, margin_of_victory, use = "complete.obs"))

# Plot the correlation
ggplot(fg3a_corr, aes(x = year, y = correlation)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Correlation between Margin of Victory and 3PT FG Attempted Advantage by Year",
       x = "Year",
       y = "Correlation Coefficient") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(fg3a_corr$year), max(fg3a_corr$year), by = 5)) +
  theme(plot.title = element_text(size = 11))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The correlation between three-point field goals attempted (3PT FG Attempted) advantage and margin of victory shows interesting trends:

In the early years (1980-2000), the correlation was negative or weak, around -0.2 to 0.1, indicating that simply attempting more three-pointers did not guarantee success. In recent years (2000-2023), the correlation turned positive and increased, reaching up to 0.4, suggesting that attempting more three-pointers has become a more effective strategy.

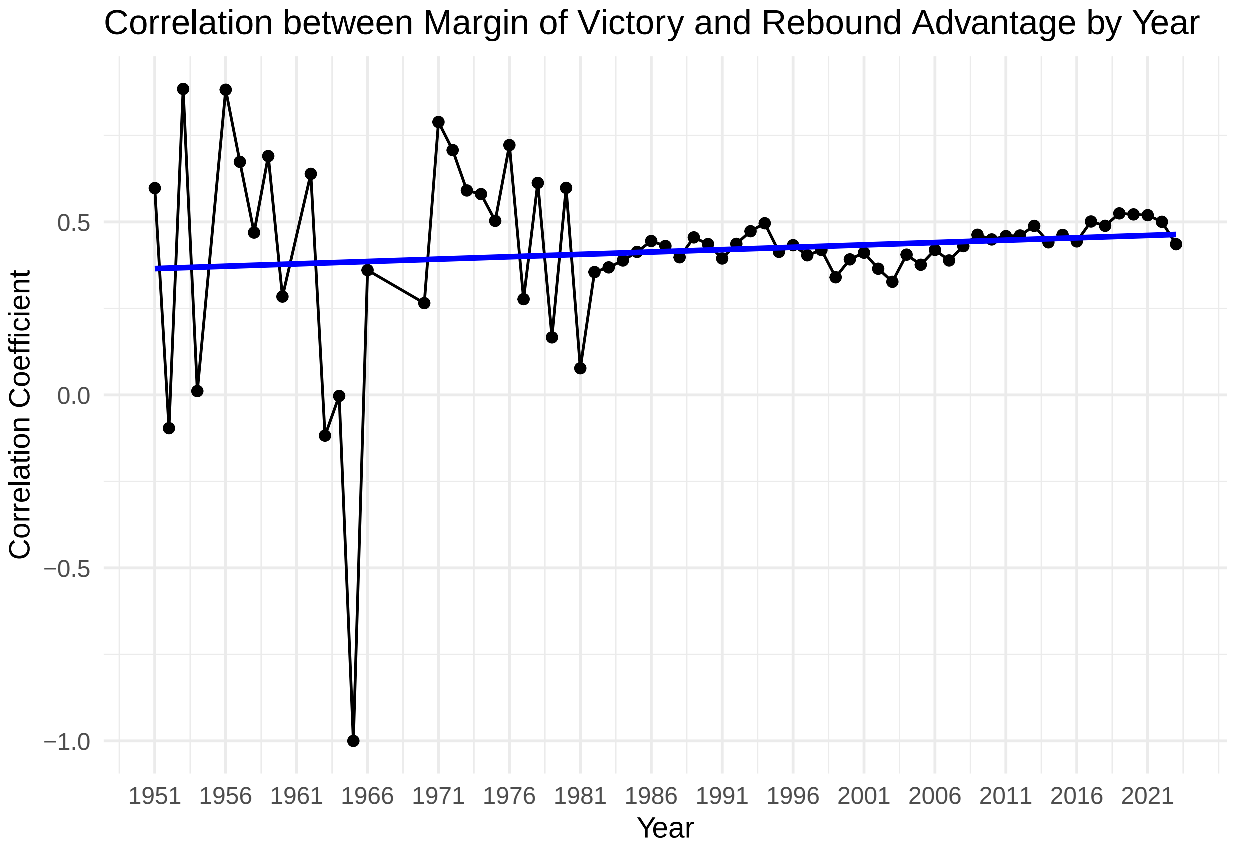
Key Insight: While early on, attempting more three-pointers did not correlate with success, the modern NBA sees a positive impact, likely due to improved shooting accuracy and strategic focus on three-point attempts.

Yearly Correlation between Rebound Advantage and Margin of Victory

```
# SQL query to prepare data for yearly correlation between rebound advantage and margin of victory
reb_corr_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    reb_home - reb_away AS reb_adv,
    pts_home - pts_away AS margin_of_victory
FROM game
WHERE reb_home IS NOT NULL AND reb_away IS NOT NULL AND pts_home IS NOT NULL AND pts_away IS NOT NULL
"

# Execute query
reb_corr_data <- dbGetQuery(con, reb_corr_query)

reb_corr_data$year <- as.numeric(reb_corr_data$year)
```

```r
# Calculate correlation for each year
reb_corr <- reb_corr_data %>%
  group_by(year) %>%
  filter(sd(reb_adv) != 0 & sd(margin_of_victory) != 0) %>%
  summarize(correlation = cor(reb_adv, margin_of_victory, use = "complete.obs"))

# Plot the correlation
ggplot(reb_corr, aes(x = year, y = correlation)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Correlation between Margin of Victory and Rebound Advantage by Year",
       x = "Year",
       y = "Correlation Coefficient") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(reb_corr$year), max(reb_corr$year), by = 5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The correlation between rebound advantage and margin of victory is consistently positive:

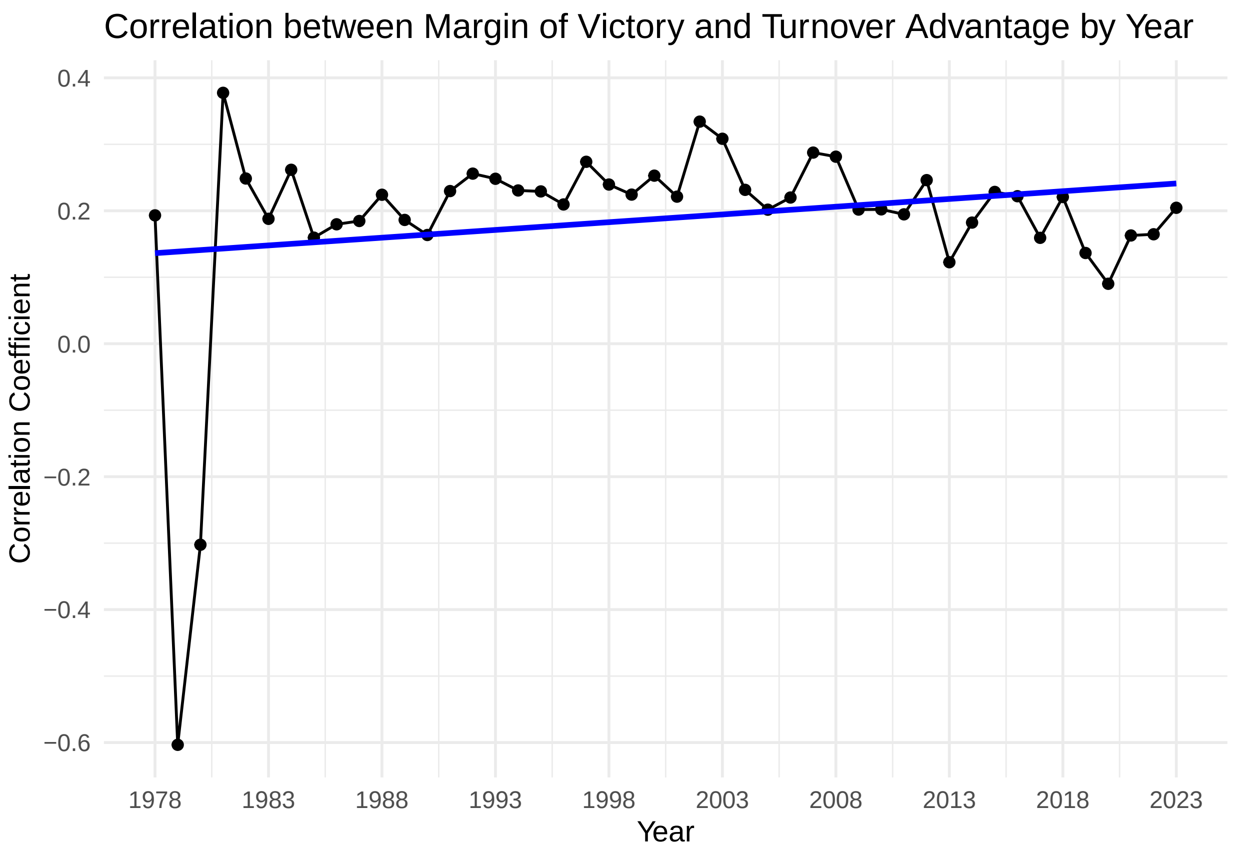
Throughout the entire period (1951-2023), the correlation fluctuated but generally remained between 0.3 and 0.6, indicating a strong relationship between rebounding and winning. Peaks in the correlation, often exceeding 0.5, highlight the critical role of controlling the boards in securing victories.

Key Insight: Rebounding advantage is a consistent and strong predictor of winning, underscoring the importance of controlling possession and limiting the opponent's scoring opportunities.

Yearly Correlation between Turnover Advantage and Margin of Victory

```
# SQL query to prepare data for yearly correlation between turnover advantage and margin of victory
tov_corr_query <- "
SELECT
    strftime('%Y', game_date) AS year,
    tov_away - tov_home AS tov_adv,
    pts_home - pts_away AS margin_of_victory
FROM game
WHERE tov_home IS NOT NULL AND tov_away IS NOT NULL AND pts_home IS NOT NULL AND pts_away IS NOT NULL
"

# Execute query
tov_corr_data <- dbGetQuery(con, tov_corr_query)

tov_corr_data$year <- as.numeric(tov_corr_data$year)

# Calculate correlation for each year
tov_corr <- tov_corr_data %>%
  group_by(year) %>%
  summarize(correlation = cor(tov_adv, margin_of_victory, use = "complete.obs"))

# Plot the correlation
ggplot(tov_corr, aes(x = year, y = correlation)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Correlation between Margin of Victory and Turnover Advantage by Year",
       x = "Year",
       y = "Correlation Coefficient") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(tov_corr$year), max(tov_corr$year), by = 5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This graph illustrates the correlation between turnover advantage and margin of victory:

From 1978 to 2023, the correlation is positive, indicating that teams with fewer turnovers (or forcing more opponent turnovers) tend to have a higher margin of victory. The correlation values range from 0.2 to 0.4, indicating a moderate positive relationship.

Key Insight: Turnover advantage is significant in winning, with fewer turnovers positively impacting the margin of victory. This highlights the importance of ball control and defensive pressure.

Strategic Implications

The analysis provides valuable insights for teams and analysts aiming to enhance their competitive strategies. However, the findings should be interpreted with caution due to several limitations. The dataset may not fully capture the nuances of modern NBA strategies, particularly as the game evolves. Factors such as player injuries, coaching changes, and variations in team strategies were not accounted for in this analysis but could significantly impact game outcomes. The increasing importance of three-point shooting, both in terms of percentage and volume, suggests that teams should prioritize developing proficient three-point shooters and incorporate perimeter shooting into their offensive strategies. Additionally, maintaining shooting efficiency (FG%), controlling rebounds, and minimizing turnovers remain fundamental aspects of successful game plans.

Evolving NBA Dynamics

The study captures the dynamic nature of the NBA, with strategic shifts reflecting changes in the game's emphasis over time. The growing impact of three-point shooting, in particular, underscores the league's evolution towards a faster-paced, perimeter-oriented style of play. These insights into the critical elements influencing NBA game outcomes provide a foundation for future analyses and strategic planning in an ever-evolving sport.

In conclusion, this detailed exploration of performance metrics and their correlations with the margin of victory offers a comprehensive understanding of the factors that drive success in the NBA. However, it is crucial to consider the limitations of the data, including potential biases and incomplete records. Future research should aim to incorporate more comprehensive datasets and consider additional variables to enhance the robustness of the findings. By leveraging these findings, teams can refine their strategies to optimize performance and achieve greater success on the court.